

LEVELS OF EXPLAINABILITY **FOR HUMAN-AI INTERACTION** **IN VISUAL TEXT ANALYTICS**

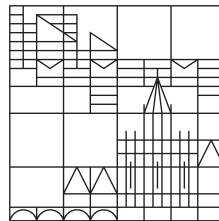
Dissertation for obtaining the academic degree
of *Doctor of Natural Sciences* (Dr.rer.nat.)

submitted by

MENNATALLAH EL-ASSADY

at the

Universität
Konstanz



FACULTY OF SCIENCES
DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE

Doctoral Thesis of the University of Konstanz
Day of the oral examination: January 23rd, 2023

1. Referee: Prof. Dr. Daniel A. Keim, University of Konstanz
2. Referee: Prof. Dr. Christopher Collins, OntarioTech University
3. Referee: Prof. Dr. Oliver Deussen, University of Konstanz

DISSERTATION

**Levels of Explainability
for Human-AI Interaction
in Visual Text Analytics**

Mennatallah El-Assady

In loving memory of my grandma,
who believed in me even when I did not believe in myself.
Thank you for waiting for me.

Dedicated to Georges Grinstein.
Thank you for teaching me about Synchronicity.

Losing you both in the same week has been the most trying time of my life.

Acknowledgement

” *Always remember people who have helped you along the way, and don’t forget to lift someone up.*

— **Roy T. Bennett**

The Light in the Heart [Ben20]

This dissertation marks the end of a multi-year journey full of learning opportunities, travel experiences, collaborations, and many exciting ideas. The years of my doctoral studies have been exhilarating ones in my life, full of change and many ups and downs. During this time, I met some of the most talented researchers from around the world, who I now have the privilege to consider dear colleagues and friends. I also found a home in a supportive research community that inspires me again and again through intellectual exchange and discussions. I am grateful and humbled to have had this opportunity, and I look forward to repaying this debt to the research community and future students.

This dissertation would not have been possible without the help and continuous support of *many* people, all of whom contributed in their own unique way to get it and me past the finish line! First and foremost, my acknowledgment must go to my co-supervisors, Profs. Daniel Keim and Christopher Collins.

Daniel has been a mentor and role model to me since my undergraduate studies. His open and welcoming attitude shaped the data analysis and visualization group at the University of Konstanz to be a robust support system and a second family for me. Daniel has shown me how entrusting responsibilities to individuals while being there as support can accelerate their personal growth. I am grateful for having had this learning opportunity in such a demanding environment. Daniel, thank you for trusting me, pushing me, and being there when I needed support. Thank you for your encouraging discussions, great advice, and constructive feedback. Thank you for your willingness and your patience in motivating me. And thank you for giving me a home away from home. Without your constant support, I wouldn’t be where I am now. I am looking forward to our future collaborations.

I met Chris for the first time at a research conference during my master's degree. We had a good conversation and exchanged a few ideas, not knowing that this interaction would mark the start of a fruitful long-term collaboration with many trips between Canada and Germany for the two of us. I now consider Chris not just a mentor and supervisor but also a dear friend. Chris, thank you for hosting me in Canada. Thank you for teaching me how to write scientific papers, especially in a deadline crunch. Thank you for encouraging me to keep going over the years and across time zones. Thank you for your support, encouragement, enthusiastic supervision, and trust. I am looking forward to future visits and collaborations.

I also gratefully acknowledge the unwavering and steadfast support of many other mentors, who helped me with personal and professional advice at different stages of my studies. Most notably, my sincere gratitude goes to Profs. Miriam Butt, Harald Reiterer, Oliver Deussen, Katharina Holzinger, and Min Chen. Thank you for being there for me!

The work in this dissertation and throughout my doctoral studies would not have been possible without the constant support, exchange, and collaboration with my colleagues in Germany and Canada. In particular, I am grateful to Michael Blumenschein, Rita Sevastjanova, Fabian Sperrle, Wolfgang Jentner, Rebecca Kehlbeck, Adam Bradley, Annette Hautli-Janisz, Valentin Gold, Thilo Spinner, Udo Schlegel, Hanna Hauptmann, Matthias Miller, Juri Buchmüller, and Christian Rohrdantz. Moreover, I would like to thank Sabine Kuhr for taking care of the whole group and especially for her administrative support.

I want to thank my friends who supported me during the writing of my dissertation through countless discussions, constant motivation, as well as great feedback and proofreading. Thank you, Mariam, Barb, and Alex, for being my support system over the past few years. I am blessed and grateful to have you all in my life.

Last but not least, it is important to me to express my gratitude to my family for their love, support, trust, and encouragement. I want to thank my mom for always believing in me and being a reliable source of comfort whenever I need it. A special thank you goes to my brother, Hassan, who went above and beyond to push me throughout the thesis writing process. Thank you for always being there for me when I need you. To my dad, I will always remain grateful for your support and love, and treasure your faith in me. And to my brother, Yassin, thank you for being the positive energy you are in my life and for continuously reminding me that I work too much. After this dissertation, it might become even more.

Abstract

The increasing demand for accountable decision-making through complex artificial intelligence systems has accelerated the need for *understanding*, *diagnosing*, and *refining* their underlying machine learning models. Hence, these three tasks have become the focus of explainable artificial intelligence research. This dissertation examines how we can make the refinement and optimization of content and topic modeling accessible to different stakeholders. Based on a theoretical framework for the explainability of artificial intelligence, we derive three *Explainability Spaces* that are tailored to four user groups. The Explainability Spaces build a continuous spectrum and inform design decisions for interactive and explainable visual analytics systems. We place **machine learning experts** and **domain experts** on opposite ends of the spectrum, in the **Model Space** and the **Semantic Space**, respectively, and position **model analysts** and **domain analysts** in the middle-ground, in the **Input-Output Space**.

The first main content chapter of this dissertation contributes an encompassing **conceptual framework** to frame the central challenge of designing tailored mixed-initiative approaches for interactive and explainable machine learning. This framework combines aspects from the analysis domain and the machine learning model to introduce interaction and explainability processes and strategies. In addition, we propose and discuss the *Explainability Spaces* as a targeted design rationale for tailoring the tasks of *understanding*, *diagnosis*, and *refinement* to the main stakeholders of explainable artificial intelligence. Based on the proposed design space, we derive four instantiations, demonstrating the applicability of this framework. Hence, this dissertation's remaining content chapters present four techniques, each utilizing novel *intelligence augmentation* paradigms to explain, diagnose, and refine comparable models.

For the **domain experts**, we present a technique designed in the **Semantic Space** that allows the experts to externalize their domain knowledge while remaining model-agnostic. This approach maintains two content-representation hierarchies that operate on a shared vector space, enabling experts to perform **guided machine teaching** for topic modeling refinement. Using word-embedding projections, experts can refine *concept regions* independent of a particular document collection or topic model. Their interactions directly affect the semantic relations of the underlying vectors, which, in turn, induce changes in the topic modeling.

For the **domain analysts**, we present a technique designed in the **Input-Output Space** (with a focus on *semantics*). It tackles the refinement of topic models through a user-driven **progressive reinforcement learning** process that does not require a deep understanding of the underlying algorithms. The system initializes two model configurations based on a parameter space analysis that enhances document separability, and lets the models compete for the analyst's satisfaction. Using automatic topic matching, topic summaries, and parameter distribution views, analysts can investigate the modeling results before providing *document-based relevance feedback*. This feedback is used to distill a user-endorsed topic distribution, allowing the system to train new model instances, and restart the feedback process to iteratively converge on a refinement.

For the **model analysts**, we present a technique designed in the **Input-Output Space** (with a focus on the *model*). It untangles reply chains by combining supervised and unsupervised machine learning models to model thematic content relations. This approach enables analysts to create and compare various reconstruction models, enriching them with user-defined queries and rule-based heuristics. To investigate the models' inner workings and performance, we visualize the **model decision spaces**, including all considered candidate relations. The proposed system allows model analysts to understand and diagnose models in order to fine-tune them based on a rich set of computed and user-derived features.

Lastly, for the **model experts**, we present an intelligible topic modeling technique designed in the **Model Space**. It relies on an incremental hierarchical topic modeling algorithm to visualize the algorithmic decision-making process. Model experts can use this system to understand the inner-workings of the model based on different parameters. To effectively assess the potential consequences of human interventions, we introduce **speculative execution** as a paradigm for creating user-steerable preview mechanisms. Whenever the measured model quality deteriorates, the system automatically triggers a speculative execution of various optimization strategies, requesting external intervention. Experts compare the proposed optimizations to the current model state, previewing their effect on the next model iterations before applying one.

Overall this dissertation presents four different techniques contributing novel visual analytics workflows that follow tailored interaction and explainability methods. They are instantiations of the proposed *Explainability Spaces*. All techniques are designed to target the same tasks of understanding, diagnosing, and refining topic and thematic content models. We have conducted comparable evaluations for all techniques to measure their effectiveness for different stakeholders. Our evaluations confirm that we can enhance problem-solving and decision-making support for mixed-initiative systems through tailored interfaces using the proposed interaction and explainability workflows.

Zusammenfassung

Die zunehmende Nachfrage nach verantwortungsvollen Entscheidungen durch komplexe künstliche Intelligenzsysteme hat den Bedarf für das Verständnis, die Diagnose und die Optimierung der zugrunde liegenden maschinellen Lernmodelle erhöht. Daher sind diese drei Aspekte zum Schwerpunkt der erklärbaren künstlichen Intelligenzforschung geworden. In dieser Dissertation wird untersucht, wie wir die Modellierung von Inhalten und Themen (*Topic Modeling*) verbessern und optimieren können, um sie für die verschiedenen Anwendergruppen zugänglich zu machen. Basierend auf einem theoretischen Framework für die Erklärbarkeit der künstlichen Intelligenz leiten wir drei *Explainability Spaces* ab, die auf vier Anwendergruppen zugeschnitten sind. Die *Explainability Spaces* bilden ein kontinuierliches Spektrum und unterstützen die Designentscheidungen für interaktive und erklärbare visuelle Analysesysteme. Um dieses Spektrum aufzuspannen, platzieren wir *Machine-Learning-Experten* und *Domain-Experten* an den entgegengesetzten Enden, im *Model Space* bzw. im *Semantic Space*, und positionieren *Modell-Analysten* und *Domain-Analysten* in der Mitte, im *Input-Output Space*.

Im ersten Inhaltskapitel dieser Dissertation wird ein umfangreiches konzeptionelles Framework vorgestellt, das die zentrale Designherausforderung von maßgeschneiderten, *mixed-initiative* Methoden zum interaktiven und erklärbaren maschinellen Lernen umreißt. Das Framework kombiniert Aspekte aus den Bereichen der Analysedomäne und des maschinellen Lernens, um Interaktions- und Erklärungsprozesse vorzustellen. Zusätzlich schlagen wir vor, die *Explainability Spaces* als eine gezielte Gestaltungsgrundlage für die Verständnis-, Diagnose- und Optimierungsaufgaben der Nutzergruppen zu verwenden. Basierend auf dem vorgeschlagenen Designraum leiten wir vier Instanziierungen ab, um die Anwendbarkeit dieses Frameworks zu demonstrieren. In den nachfolgenden Kapiteln dieser Dissertation werden daher vier Methoden vorgestellt, die jeweils neuartige Paradigmen zur Intelligenzanreicherung nutzen, um vergleichbare Modelle zu erklären, zu diagnostizieren und zu optimieren.

Für die *Domain-Experten* stellen wir eine Methode vor, die im *Semantic Space* konzipiert wurde. Sie ermöglicht es den Experten, ihr Domänenwissen zu externalisieren und dabei modellunabhängig zu bleiben. Bei diesem Ansatz werden zwei Hierarchien für Inhaltsdarstellungen aufrechterhalten, die auf einem gemeinsamen Vektorraum operieren. Dies ermöglicht es den Experten, das *Topic Modeling* mithilfe von *Guided Machine Teaching* zu verfeinern. Durch die Verwendung von *Word-Embedding*-Projektionen können Experten Konzeptbereiche unabhängig von spezifischen Dokumentensammlung oder Themenmodellen verfeinern. Ihre Interaktionen wirken sich direkt auf die semantischen Beziehungen der zugrunde liegenden Vektoren aus, die wiederum Anpassungen in der Themenmodellierung hervorrufen.

Für die *Domain-Analysten* stellen wir eine Methode vor, die im *Input-Output Space* konzipiert wurde (mit Schwerpunkt auf die Semantik). Sie ermöglicht die Optimierung von *Topic Models* durch einen nutzergesteuerten, progressiven *Reinforcement Learning* Prozess, der kein tiefgreifendes Verständnis der zugrunde liegenden Algorithmen erfordert. Das System initialisiert zwei verschiedene Modellkonfigurationen basierend auf einer *Parameter-Space*-Analyse, die die Trennbarkeit der Dokumente verbessert. Danach treten die zwei Modelle gegeneinander an. Mittels Ansichten auf den automatischen *Topic Matching*, *Topic Summaries* und Parameterverteilungen können Analysten die Modellierungsergebnisse untersuchen, bevor sie ein Feedback zur Relevanz der Dokumente geben. Dieses Feedback wird verwendet, um eine vom Benutzer befürwortete Themenverteilung zu ermitteln. Dadurch kann das System neue Modell-Instanzen trainieren und den Feedback-Prozess neu anstoßen, um iterativ zu optimieren.

Für die *Modell-Analysten* präsentieren wir eine Methode, die im *Input-Output Space* konzipiert wurde (mit Schwerpunkt auf dem Modell). Sie kombiniert supervised und unsupervised Methoden, um Antworten und Kommentare zu entflechten. Dieser Ansatz ermöglicht es Analysten, verschiedene *Thread-Reconstruction*-Modelle zu erzeugen und zu vergleichen, indem sie die Modelle mit benutzerdefinierten Abfragen und regelbasierten Heuristiken anreichern. Um das Innenleben und die Leistung der Modelle zu untersuchen, werden die Entscheidungsräume der Modelle visualisiert, inklusive aller in Erwägung gezogenen Relationen. Unser System ermöglicht es *Modell-Analysten*, ein Verständnis für die Modelle zu entwickeln, um sie anhand einer Vielzahl von berechneten und vom Benutzer abgeleiteten *Features* zu optimieren.

Schließlich präsentieren wir für *Modell-Experten* eine Methode für *Topic Modeling*, die im *Model Space* konzipiert wurde. Sie beruht auf einem inkrementellen hierarchischen Algorithmus zur Themenmodellierung. Dieser visualisiert den algorithmischen Entscheidungsprozess. *Modell-Experten* können mithilfe dieses Systems die innere Funktionsweise des Modells unter Berücksichtigung verschiedener Parameter analysieren. Um die potenziellen Konsequenzen menschlicher Eingriffe wirksam zu beurteilen, führen wir den Konzept der *Spekulative Execution* ein. Verschlechtert sich die gemessene Modellqualität, löst das System automatisch eine *Speculative Execution* verschiedener Optimierungsstrategien aus und fordert ein externes Eingreifen der Experten an. Diese vergleichen die vorgeschlagenen Optimierungen mit dem aktuellen Zustand des Modells und prüfen deren Auswirkungen auf die nächsten Modelliterationen, bevor sie diese anwenden.

Insgesamt werden in dieser Doktorarbeit vier verschiedene Methoden vorgestellt, die zu neuartigen visuellen Analyse-Workflows beitragen. Diese beruhen auf maßgeschneiderten Interaktions- und Erklärbarkeitsverfahren und sind Instanzen der oben genannten *Explainability Spaces*. All diese Methoden verfolgen das gleiche Ziel; das Verstehen, Diagnostizieren und Optimieren von *Content* und *Topic Modelle*. Wir haben mit allen Methoden eine vergleichbare Evaluierung durchgeführt, um ihre Wirksamkeit für verschiedene Anwender zu messen. Unsere Auswertungen bestätigen, dass wir durch maßgeschneiderte Schnittstellen unter Verwendung der vorgeschlagenen Interaktions- und Erklärbarkeits-*Workflows* die Problemlösungs- und Entscheidungsunterstützung für *Mixed-Initiative*-Systeme fördern können.

Contents

Abstract/Zusammenfassung	ix
---	-----------

Motivation and Background

1 Introduction	3
1.1 Human-AI Interaction	4
1.2 Visual Text Analytics	6
1.3 Levels of Explainability	8
1.4 Problem Statement and Research Questions	10
1.5 Contributions and Structure of the Dissertation	12
1.6 Publications, Citations, and Terminology	14
2 Related Work	19
2.1 Interactive and Explainable Artificial Intelligence	20
2.1.1 Algorithm-Centered Interactive and Explainable AI	21
2.1.2 Human-Centered Interactive and Explainable AI	23
2.1.3 Explainability Considerations for Effective Human-AI Interaction	25
2.2 Visual Text Analytics	27
2.2.1 Task and Technique Overview	28
2.2.2 Analysis and Visualization of Conversational Data	30
2.2.3 The LingVis.io Framework	31
2.3 Content and Topic Model Refinement	33
2.3.1 Content Analysis and Topic Modeling	34
2.3.2 Visualization, Interpretation, & Refinement of Content & Topic Models	36
2.3.3 Topic Modeling Quality Metrics	38

Conceptual Framework

3 Levels of Explainability for Human – AI Interaction	43
3.1 Introduction	46
3.2 The Design Space of Interactive and Explainable Artificial Intelligence	47
3.2.1 Users and Stakeholders	48
3.2.2 Analytical Tasks and Application Domains	49

3.2.3	Data Characteristics	51
3.2.4	Types and Classes of Machine Learning Models	52
3.2.5	Stages in the KDD Process	52
3.2.6	Explainability Goals and Values	53
3.2.7	Explainability Tasks and Processes	55
3.2.8	Methods, Mediums, and Strategies	55
3.3	Operationalizing Interactive and Explainable Machine Learning	57
3.3.1	The <i>explAIner</i> Pipeline	57
3.3.2	Single- and Multi-Model <i>explAIners</i>	60
3.3.3	Global Monitoring and Steering Mechanisms	61
3.4	Explainability Spaces: A Conceptual Framework	62
3.4.1	Semantic Space	66
3.4.2	Input-Output Space	68
3.4.3	Model Space	70
3.5	Explainability Spaces for Topic Model Refinement	73

Semantic Space

4	Guided Topic Model Refinement using Word-Embedding Projections	77
4.1	Introduction	80
4.1.1	Research Question and Challenges	81
4.1.2	Contributions	82
4.2	Designing Visual Analytics Approaches for Domain Experts	83
4.2.1	Application Design Space	83
4.2.2	I&XAI Workflow and Visual Analytics Pipeline	84
4.3	Modeling the Semantic Concept Space	85
4.3.1	Interactive Concept Generation	87
4.3.2	Concept Neighborhood Computation	88
4.3.3	Concept Hierarchy Building	89
4.3.4	Layered Canvas Mapping	91
4.4	Visual Analytics Workspace	92
4.4.1	Concept View	93
4.4.2	Topic View	94
4.4.3	Concept Space Interactions	96
4.5	Interactive Learning of the User's Semantics	96
4.5.1	Concept Refinement	97
4.5.2	Topic Modeling Adaptation	99
4.6	Evaluation	100
4.6.1	Qualitative Results: Expert Feedback	101
4.6.2	Quantitative Results: Quality Assessments	103
4.7	Discussion: Investigating I&XAI Design Patterns	104

Input-Output Space: Semantic

5	Progressive Learning of Topic Modeling Parameters	109
5.1	Introduction	112
5.1.1	Research Questions and Challenges	113
5.1.2	Contributions	114
5.2	Designing Visual Analytics Approaches for Domain Analysts	114
5.2.1	Application Design Space	115
5.2.2	I&XAI Workflow and Visual Analytics Pipeline	116
5.3	Parameter Space Analysis	116
5.3.1	Data-Driven Feature Selection and Weighting	117
5.3.2	Document Descriptor Vector Generation	118
5.3.3	Initializing the Workspace	119
5.4	Visual Analytics Workspace	119
5.4.1	Topic Matching	120
5.4.2	Topic Summarization	121
5.4.3	Parameter Distribution Analysis	123
5.4.4	Document Relevance Feedback	124
5.5	Task-Driven Topic Convergence	127
5.5.1	Reinforcement Learning	127
5.5.2	Iterative Topic Evolution	128
5.5.3	Global Learning of Parameter Scales	129
5.6	Evaluation	130
5.6.1	Quantitative Results: Model Improvement	132
5.6.2	Qualitative Results: Expert Feedback	133
5.7	Discussion: Investigating I&XAI Design Patterns	134

Input-Output Space: Model

6	Modeling Reply Chains to Untangle Conversational Text	139
6.1	Introduction	142
6.1.1	Research Questions and Challenges	143
6.1.2	Contributions	145
6.2	Designing Visual Analytics Approaches for Model Analysts	145
6.2.1	Application Design Space	145
6.2.2	I&XAI Workflow and Visual Analytics Pipeline	146
6.3	Background: Thread Reconstruction	147
6.4	Data-Driven Feature Extraction and Weighting	149
6.4.1	Reply Relation Features	149
6.4.2	Message Categorization	150
6.4.3	Data-Driven Feature Weighting	150
6.5	User-Driven Model Generation	151

6.5.1	Classifier Model	151
6.5.2	User-Defined Queries	151
6.5.3	Rule-Based Heuristics	152
6.6	Visual Analytics Workspace	152
6.6.1	Text-Level View	153
6.6.2	Reply Relation View	154
6.6.3	Thematic Forest View	155
6.7	Visual Model Diagnostics	156
6.7.1	Parent-Candidate View	156
6.7.2	Model Optimization and Transfer	158
6.8	Evaluation	158
6.8.1	Scalability and Model Reliability	158
6.8.2	Expert User Studies	160
6.8.3	Case Studies	161
6.8.4	Feedback	165
6.8.5	Lessons Learned	167
6.9	Discussion: Investigating I&XAI Design Patterns	168

Model Space

7	Topic Model Optimization based on User-Steerable Speculative Execution	173
7.1	Introduction	176
7.1.1	Research Questions and Challenges	177
7.1.2	Contributions	178
7.2	Designing Visual Analytics Approaches for Model Experts	178
7.2.1	Application Design Space	179
7.2.2	I&XAI Workflow and Visual Analytics Pipeline	180
7.3	Progressive and Explainable Model Building	180
7.3.1	Similarity Computation	181
7.3.2	Quality Measurement	182
7.3.3	IHTM Evaluation	183
7.4	Visual Analytics Workspace	184
7.4.1	IHTM-Backbone Priors	186
7.4.2	Model-Space View	187
7.4.3	Model Quality Monitoring	190
7.5	Speculative Execution	190
7.5.1	Automatic Optimization Strategies	191
7.5.2	Comparative Speculation Mode	192
7.6	Evaluation	194
7.6.1	Qualitative Results: Expert Feedback	195
7.6.2	Quantitative Results: Model Quality Assessment	197

7.6.3 Lessons Learned	197
7.7 Discussion: Investigating I&XAI Design Patterns	198

Closing

8 Summary	203
8.1 The Design Space of I&XAI Applications	204
8.1.1 I&XAI Design Dimensions	204
8.1.2 <i>explAIner</i> Properties	204
8.1.3 Explainability Considerations	205
8.2 The Explainability Spaces Framework	205
8.2.1 The Semantic Space	205
8.2.2 The Input-Output Space	205
8.2.3 The Model Space	206
8.3 I&XAI Techniques for Content & Topic Modeling	206
8.3.1 Designing for Domain Experts: <i>Semantic Concept Spaces</i>	206
8.3.2 Designing for Domain Analysts: <i>Progressive Learning</i>	207
8.3.3 Designing for Model Analysts: <i>ThreadReconstructor</i>	208
8.3.4 Designing for Model Experts: <i>Speculative Execution</i>	208
8.4 I&XAI Design Patterns	208
9 Discussion	211
9.1 Limitations	212
9.1.1 I&XAI Framework and Design Space Methodology	212
9.1.2 Evaluation and Generalizability	212
9.1.3 Targeted Human Modeling for Personalization and Guidance	213
9.2 Lessons Learned	213
9.2.1 I&XAI Workflows for Intelligence Augmentation	213
9.2.2 Co-Adaptive Analytics: I&XAI as Dynamic Systems	214
9.2.3 Reasoning and Knowledge Representation in Human-AI Communication	216
9.3 Open Research Questions and Challenges	216
9.3.1 Detecting and Mitigating Human-AI Miscommunication Issues	216
9.3.2 Interaction Intent Identification to Bridge Knowledge Gaps	217
9.3.3 Trustworthy Decision-Making Support	217
10 Conclusion	219
10.1 Take-Home Messages	220
10.1.1 Challenging Problems Require Shared Human-AI Agency	220
10.1.2 Stakeholders and Application Contexts Determine I&XAI Workflows	220
10.1.3 Tailored Visual Analytics System Design Empowers Users	221
10.2 Future Research Directions	221
10.2.1 Distilling Best Practices, Design Patterns, and Guidelines	221

10.2.2 Developing a Structured I&XAI Methodology	221
10.2.3 Supporting Seamless Human-AI Collaborations	221
Bibliography	223

List of Figures

1.1 Intelligence Augmentation.	5
1.2 The Spectrum of Explainability Spaces.	9
1.3 Traditional Workflow of Interactive Machine Learning.	10
1.4 Overview of the Dissertation Structure.	12
2.1 Explainability Considerations for Effective Human-AI Interaction.	26
2.2 The Text Visualization Browser by Kucher and Kerren [KK15].	27
2.3 The LingVis.io Framework.	31
2.4 LingVis.io Processing Pipeline.	32
2.5 Settings Menu for the LingVis.io Document Descriptor Extractor.	32
3.1 The <i>explAIner</i> Pipeline.	57
3.2 The Spectrum of Explainability Spaces in Machine Learning.	63
3.3 Explainability Spaces.	65
3.4 The Semantic Space.	67
3.5 The Input-Output Space.	69
3.6 The Model Space.	71
4.1 Semantic Concept Spaces enabling knowledge externalization for <i>domain experts</i>	81
4.2 Deductive I&XAI Workflow.	84
4.3 Interactive Workflow of the Semantic Concept Space Approach.	85
4.4 Semantic Abstraction Levels for Concepts.	89
4.5 Duality of Concept and Topic Views.	93
4.6 Two Options for Concept Refinement.	97
4.7 Guided Relevance Feedback for Refining the <i>Semantic Concept Space</i>	99
5.1 Progressive Learning based on relevance feedback by <i>domain analysts</i>	113
5.2 Contrastive I&XAI Workflow.	115
5.3 Interactive Workflow of the Progressive Learning Approach.	116
5.4 Topic Matching View.	120
5.5 Topic Summarization View	121
5.6 Parameter Distribution View	123
5.7 Document Relevance Feedback View.	125

5.8	The Automatic Model Uncertainty.	130
6.1	Model Decision Space revealing considered decisions to <i>model analysts</i>	143
6.2	Contrastive I&XAI Workflow.	144
6.3	Interactive Workflow of the ThreadReconstructor Approach.	146
6.4	Text-Level View.	154
6.5	Thematic Forest of Untangled Reply Chains from a full Conversation.	155
6.6	Close-up on the Parent-Candidate View.	156
6.7	Comparison of two Classifier Models in Parent-Candidate View.	157
6.8	Thematic Forest View showing Connected Components from Figure 6.5b.	162
7.1	Speculative Execution of topic model optimizations for <i>model experts</i>	177
7.2	Inductive I&XAI Workflow.	179
7.3	Interactive Workflow of the SpecEx Approach.	180
7.4	IHTM Workspace Overview.	185
7.5	Topic Modeling Backbone Interface.	187
7.6	IHTM Result Comparison.	189
7.7	Tree-Speculation View.	193
7.8	SpexEx-VA Pipeline.	194
8.1	The Design Space for I&XAI Applications.	204
8.2	Summary of the Explainability Spaces Framework.	206
8.3	Overview of the Four I&XAI Techniques of the Dissertation.	207
8.4	Overview of I&XAI Design Patterns.	209
9.1	I&XAI Workflow Objectives and Desired Outcomes.	213
9.2	Co-Adaptive Analytics.	214
9.3	Adaptive Guidance Agents for Topic Model Optimization.	215
9.4	Contrastive I&XAI Workflow.	215
10.1	I&XAI Interfaces for Content and Topic Model Refinement.	220

List of Tables

4.1	Ranked Output of the Concept Space.	104
5.1	User Study Results Indicating a clear Improvement for all Participants and Topics.	131
7.1	Comparison of IHTM against State-of-the-Art Topic Models.	184
7.2	Result Quality for Six Different Models.	196
7.3	Quality Metric Change (in %) when adding a Backbone and using SpecEx.	198

Preface

” *I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.*

— **Alan Turing**

Computing Machinery and Intelligence [Tur50]

From the early hunter-gatherers to the modern sapiens, we humans have an intricate history of development and change. We have evolved from simple social structures to an ever more interconnected and sophisticated coexistence. Throughout our shared history, we have been learning from experiences and *interactions*; from information passed down through the generations; from our elders’ teachings and *explanations*. Humans’ path to reflective and critical *thinking* is paved by sharing *knowledge* through *teaching* and *learning*. We experiment, gather information, and grow our complex network of knowledge. This accumulated knowledge has allowed our species to evolve from prehistoric nomads to reaching the moon and beyond.

A Series of Revolutions | From the agricultural revolution to the industrial revolution, making *informed decisions* based on knowledge and evidence drove human innovation. Through these intellectual and technological leaps, we have learned to understand our surroundings, use resources, and create new *tools*. As a response to various events and challenges around us, we evolved our thinking and tool-making, shaping culture and sometimes profoundly transforming society. Such transformational changes come in waves; they build up from evolutionary to revolutionary, catalyzing fundamental paradigm shifts. These breakthroughs open up novel markets and economies, requiring new actors and toolkits.

The Thinking Machine | In every step in human evolution, our tools have influenced how we define ourselves as individuals and our society as a collective. When tools were primitive stones, humans were hunters, gatherers, and fishermen. When they evolved to metal, humans became farmers, sailors, and soldiers. With the steam engines of the industrial age, humans became engineers, manufacturers, and explorers. By generating and using electricity, a new digital age allowed humans to talk on telephones, fly airplanes, and program computers. Finally, with the technological leaps in computation, a newly found information age brought about a revolutionary paradigm, a *non-human* decision-maker; a ‘*thinking*’ machine?

Thinking and Intelligence | With increasing computational power and algorithmic advancements, came a surge in automatic decision-making. The brilliance of automation led us to ask ourselves *whose intelligence we’re observing?* Are our machines thinking or merely *reflecting back at us our scaled, collective problem-solving capabilities?* It made us ask critical questions about our intelligence. *What is thinking anyway? Can a machine think unselfishly? Creatively? Realistically? Strategically? Can it be observant or self-critical?* Now, the tools we have developed are holding a mirror in our faces, confronting us with questions about intelligence and reasoning, making us rethink what makes us human and what makes our societies function.

Artificial Intelligence | To study and evolve our evermore complex automated decision partners, we started researching this *artificial* intelligence. This research accelerated innovations in every scientific field through tailored automated problem-solving. Now, our writers and thinkers started predicting futures in which exponential innovations would lead to ‘*General Artificial Intelligence*’ or other bleaker outcomes where humans will be ruled by a singular ‘*Superintelligence*.’ However, current advances in this field of research put us first on the road to figuring out how to *explain* and *justify* the automatic decision-making.

Opening the Black-Box | To *understand* what machines have *learned*, we need to observe their decision-making behaviors, especially when these decisions are critical or costly. Hence, we started ‘*opening the black box of artificial intelligence*’ to explain how machines think. But what is a good explanation? Does it have to be *faithful* to the decision-making behaviors or only *comprehensive* to humans? And after constructing the first explanations, we asked ourselves how we can *correct, adapt, or refine* the automated decision-making. *Can we have bi-directional communication between a human and an artificial intelligence?*

Tailoring Human-AI Interactions | Our need to *interact with AIs* has led us to develop interfaces that allow experts to diagnose issues and communicate their solutions. However, to enable non-AI-experts to have the same level of communication, we need to tailor our *interaction paradigms and methods*. This led us to ask ourselves, what makes communication successful? What can we learn from human-human interaction? And how different or similar should *human-AI interaction* be? Addressing these challenges opened up new research directions.

Collaborative Problem-Solving | With the development of more and more explanation and interaction possibilities between human and artificial intelligence, we started to distinguish between the strengths and weaknesses of each agent when it comes to problem-solving. AIs can outperform when it comes to scalable computing and structured analysis, while humans are much better at connecting and deriving knowledge through their semantic understanding and common sense. So we ask ourselves, *how do we make these two agents collaborate?*

Intelligence Augmentation | We are currently on the frontier to develop tools that allow for bi-directional human-AI communication for explanations, adaptations, and informed decision-making. Such human-AI interaction facilitates teaming up both agents for joint problem-solving and decision-making. This *augments* both intelligences, giving humans computing superpowers and AIs human superpowers. With this research, we can facilitate a spectrum of *intelligence augmentation* through varying the agency of the two actors. For some challenges, we might want both humans and AIs to equally contribute to the problem-solving, so-called *mixed-initiative analysis*. In other situations, we might want to put the *human in the loop* of automated analysis or the *AI in the loop* of a manual one.

Towards Seamless Human-AI Communication | This dissertation does not claim to answer all of these questions, but it contributes to our understanding of effective human-AI interaction for problem-solving. It investigates tailoring interactions to different user backgrounds and experience profiles. We present techniques that vary their interaction and explainability design to cater to various stakeholder groups. For example, we explore attributing *agency* and control of the analysis process to AIs, a capability traditionally reserved for human agents.

To echo Alan Turing's thought, welcome to the future of *thinking machines*...

Motivation and Background

” *Do not go where the path may lead,
go instead where there is no path and leave a trail.*

— **Ralph Waldo Emerson**
(attributed)

Introduction

“AI has by now succeeded in doing essentially everything that requires ‘thinking’ but has failed to do most of what people and animals do ‘without thinking’ – that, somehow, is much harder!

— Donald Knuth

(as quoted by Bostrom [Bos17])

Summary

This dissertation investigates *how to empower different stakeholders to interact with machine learning models for collaborative problem-solving and decision-making*. In particular, it addresses the question of *how to tailor explainability design to target the stakeholders’ tasks and expertise levels*. To study these issues, we chose the problem of mixed-initiative content and topic model refinement, as it is a relevant challenge for many domain scientists, but more importantly, since this problem is subject to a *multi-objective* optimization where the automatic quality metrics available are not aligned with human intuition. Hence, in practice, to refine such models for a specific task and corpus, users rely on semi-supervised approaches that enable them to *externalize their domain knowledge*. This dissertation presents four visual text analytics techniques that enable such human-AI interaction with the focus on facilitating the *understanding, diagnosis, and refinement of content and topic models*.

Contents

1.1	Human-AI Interaction	4
1.2	Visual Text Analytics	6
1.3	Levels of Explainability	8
1.4	Problem Statement and Research Questions	10
1.5	Contributions and Structure of the Dissertation	12
1.6	Publications, Citations, and Terminology	14

1.1 | Human-AI Interaction

Artificial intelligence (AI), in particular through the rapid advancement of machine learning (ML), is becoming more prominent and essential in every part of our lives. Our dependence and reliance on automated problem-solving and decision-making are growing and becoming more intricate. We augment human judgments with artificial intelligence to tackle many challenges in our modern societies. On the two sides of this fruitful collaboration, each of the agents (humans and AIs) boasts particular strengths. Humans can tackle challenging problems by relying on their common sense, connected general knowledge, creativity, and perception. On the other hand, AIs can leverage their computing power, scalable data storage, and systematic search capability. Through teaming up humans and AIs, we are thus able to solve more challenging problems than each can tackle by themselves.

To allow for such a collaboration, we need both agents to communicate. Hence, one of the goals of this research is to design effective and empowering communication interfaces for human-AI interaction. To illustrate why this line of research is needed, let's start with a thought experiment. Imagine you are trying to make a decision that requires input from multiple people based on their expertise in particular domains. You invite all these stakeholders to a meeting room and discuss all aspects of the decision with them. Some might explain their rationales verbally; others might use a visual aid such as a presentation or drawing. Each of the people in the room can ideally “read the room” and thus sense if the others do not understand their point of view and maybe resume their explanations using a different strategy or medium. Such a human-human interaction is the foundation of our shared capacity for joint problem-solving and decision-making. It is a capability that people are taught to acquire throughout their lives and requires many social mechanisms to be effective, such as active listening and understanding, mentalizing and perspective-taking, etc.

Now let's consider two AI agents collaborating; they do not require any of the social communication essential for humans. However, they do require clearly defined interfaces, specific problem statements, a shared understanding of the data dimensions, etc. Now, let's come back to human-AI teams; one of the open questions about their collaboration is how to facilitate effective communication. The current state-of-the-art attributes the agency of the joint problem-solving squarely to either of the two agents. We are surrounded by automated systems where humans are asked for specific inputs by an AI, leaving it in control of the decision-making process. Or, on the flip side, environments where humans have integrated AIs as tools in their workflows. We are still lacking prominent examples where both agents see eye-to-eye.

To conclude our thought experiment, imagine that a few years from now, you are again trying to make a decision that requires input from multiple domains. You invite domain experts consisting of specialized AI and human agents to a meeting. They each start deliberating and explaining their viewpoints and data to each other, each understanding the viewpoints others

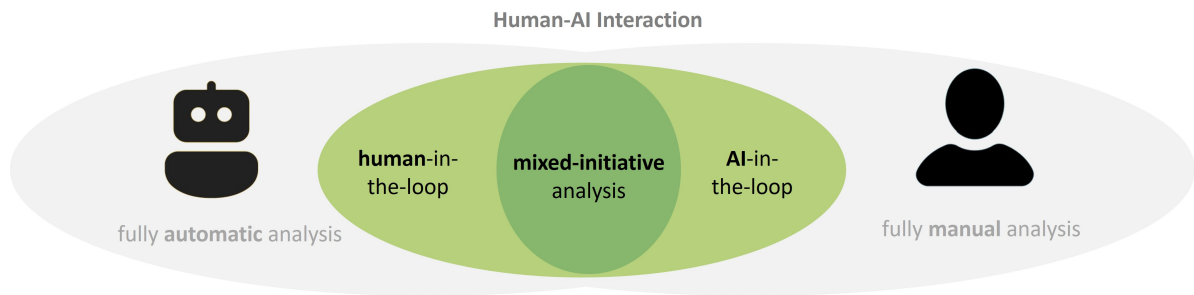


Fig. 1.1: The spectrum of **Intelligence Augmentation** through human-AI interaction from fully automatic to fully manual analysis, encompassing automatic analysis with humans in the loop, mixed-initiative analysis, and manual analysis with an AI in the loop.

are making, able to sense agreement or confusion in the room, effortlessly switching between explanation mediums and strategies, and articulating reliable information in different contexts to reach a consensus and solve the given problem. Such seamless human-AI communication is still far away from the current state of the art. It requires research advancements in many fields to fulfill. This dissertation takes one step towards studying human-AI interaction by investigating different levels of explainability and tailored visual analytics workflows.

Intelligence Augmentation One of the research directions on the rise to tackle some of the challenges of human-AI collaboration is *intelligence augmentation* [Zho+21]. This line of work frames human-AI interactions across a spectrum ranging from fully manual to fully automatic analysis. As illustrated in Figure 1.1, human-AI interactions encompass integrating humans into an automatic analysis loop, e.g., in semi-supervised learning; mixed-initiative systems where both agents share agency, e.g., in co-adaptive analytics; and integrating AIs in manual analysis settings. The analytical tasks and applications suitable for each interaction paradigm are discussed in Section 3.2.2. Overall, this dissertation contributes process models to design various intelligence augmentation workflows.

Visual Analytics Visual analytics, as defined by Keim et al. [Kei+08], argues for the tight integration of visual and automatic data analysis methods for information exploration and scalable decision support. Hence, it promotes the design of visual interfaces that facilitate human-AI interaction. Interactive visualizations as a communication medium are a well-established technique for exploratory, confirmatory, and task-driven data analysis. By combining interactive visual interfaces with machine learning models, we can design effective workflows for human-AI collaboration, facilitating data-driven problem-solving and decision-making. Visual analytics techniques can also be applied to provide a comprehensive understanding of algorithms, support the exploration of applying AI models on specific datasets, show the effect of changing parameters, and enable users to tune the models to their tasks and data. This dissertation uses visual analytics methods to design interactive and explainable interfaces that enable different stakeholders to understand, diagnose, and refine machine learning models, or more precisely, content and topic models, as described in the following section.

1.2 | Visual Text Analytics

With the increasing availability of large digitized text corpora, as well as powerful computational linguistics methods, scholars and practitioners have been striving to gain insights into various domain problems through a systematic automated analysis of document collections. To make such an analysis accessible, the field of visual text analytics has contributed numerous methods to aid the exploration of data, the targeted analysis of text for various applications, or the inspection of computational models to understand their processing methods. For a detailed overview of the state-of-the-art in visual text analytics, please refer to Section 2.2.

Kucher and Kerren [KK15] provide a comprehensive survey on the tasks, domains, data, and visualization techniques developed for text visualization. Generally, there is an observed dichotomy between general-purpose tool design and specialized applications tailored to a specific set of tasks and data. A particular challenge for analyzing text data is that it varies in its structure and content, the linguistic information varies in the quantity of information it encodes, and analysts vary in their background, tasks, and expertise. Hence, there are always multiple objectives to consider when modeling language and text data, as surveyed by Liu et al. [Liu+18c]. As these objectives might not completely align with human intuition and understanding of the data, a recent focus of visual text analytics research has been the study and development of approaches for interacting with linguistic modeling techniques. For example, approaches for the exploratory understanding of the models' decision-making, the interactive creation and refinement of models, and others.

Designing model-driven visualization approaches to enhance the interpretability and trust for automatic text analysis techniques has proven helpful [Chu+12]. Visual text analytics enables data- and task-centric model creation through a human-in-the-loop design. Hence, an effective model visualization with an iterative feedback cycle is a promising approach for a user-steerable and interpretable machine learning process [Spe+21a]. For example, consider as stakeholders, humanities and social science scholars, who often use topic and content models to gain insight into the thematic structure of document collections without the need for time-consuming classification and close-reading [Jän+15]. Despite their convenience and broad applicability, these models typically remain black boxes, not readily understood by these end users [Chu+12]. However, understanding these algorithms' basic principles is essential to configure and utilize them properly. Hence, there is a need to *understand* how the results of these models are created and to *adapt* the models to given data and tasks in order to enhance their provenance and reliability [Ble12b]. In their study on how non-experts perceive, interpret, and fix topic models, Lee et al. [Lee+17b] recommend that “more active, mixed-initiative support may be useful during the refinement process, such as pointing users to topics with high refinement potential and providing immediate feedback”.

Content and Topic Modeling In the context of visual text analytics, content and topic modeling algorithms are widely used as a processing step to efficiently segment document collections into thematically related groups [BHM17]. More precisely, they are a class of unsupervised machine learning algorithms that categorize text corpora based on the distribution of topics discovered within; a process that typically remains mostly concealed from end-users to disguise its algorithmic complexity [Müh+14; Liu+17b]. However, treating content and topic models as typical black-box machine learning components limits the trustworthiness of their results and, in turn, of the whole system they inform [Chu+12]. Tackling the issue of such a one-way process, several approaches have emerged, promoting trust through linking the inputs, outputs, and parameter spaces of topic models into visual analytics frameworks [Wei+10]. These techniques present a step toward truly interactive machine learning by enabling a human-in-the-loop process that allows users to adapt the models to their data and tasks. In summary, content and topic models typically operate in a high-dimensional vector space defined by the keywords of a corpus, with documents represented as frequency distributions in that space. The most commonly used class of algorithms are generative, probabilistic models that create topic-document assignments in multiple optimization iterations. A comprehensive overview of these methods and how to interact with and explain them is provided in Section 2.3.1.

Model Explainability and Refinement Content and topic models are notoriously difficult to work with [Cha+09]. Typical algorithms, such as latent Dirichlet allocation (LDA, [BNJ03]), attribute keywords to topics, not documents, which is not intuitive for non-machine-learning-experts to understand, let alone optimize [Sch12]. The main obstacle to designing effective visualizations for such models is the discontinuity between their *parameter space* and their *model space*. It is challenging to get an intuition of the effect of changing algorithmic parameters or directly manipulating the model space on the output of these models. As the recent paper investigating how non-experts perceive, interpret, and fix topic models puts it, “*with an LDA-based approach [...], seemingly small changes on the user side could have unpredictable and nonsensical cascading side effects*” [Lee+17b]. Yet, Blei [Ble12b] argues that their power will be realized best when used in the service of history, sociology, linguistics, and other social sciences and humanities fields. This is usually accomplished by teaming computer scientists with non-computer scientists to create topic models. However, with the popularity of toolkits such as MALLET [MM02], employing topic models is becoming more common for people at all levels of expertise. Consequently, it is critical to design tailored human-AI interaction techniques that enhance the understandability and adaptability of the parameters to refine the models regardless of the stakeholders’ backgrounds and expertise. This dissertation presents four interactive techniques for content and topic model explainability and refinement, each tailored to a specific stakeholder profile, as discussed in Section 3.2.1. Note that we chose this research area as these models need to consider multiple objectives that, despite all being relevant, still do not fully capture human intuition and understanding. In addition, there are multiple plausible solutions to thematic analysis and no single ground truth. Hence we usually cannot optimize the models unless we consider their domain-, task-, and data-contexts.

1.3 | Levels of Explainability

As motivated in the previous sections, there is a pressing need to understand machine learning models' problem-solving and decision-making processes in order to analyze and justify their results, and adapt them to different contexts. With artificial intelligence becoming more and more intertwined in various application domains, the advancement of techniques to understand, diagnose, and refine models has formed the research subareas of interactive and explainable AI (I&XAI). Especially methods to explain neural networks and other black-box models are in high demand. They have attracted significant interest from funding agencies [GA19] and, due to their increasingly consequential decision-making, become subject to regulation deliberations [Smu19], prompting an even more fast-paced development of this young field.

This gold rush has led to a scattered research landscape where different methods have been developed and tested to meet the need of the various domains and stakeholders. Numerous approaches from diverse research areas now successfully explain how models achieve their decisions, as described in Section 2.1. However, the design space of I&XAI has still not been systematically captured or investigated. Several attempts to structure this field have chosen different foci, prioritizing either specific types of models or exact application domains. This dissertation contributes to systematically structuring this space, distilling arguments from previous works, and incorporating algorithm-centered and human-centered I&XAI design considerations (see Section 2.1.3). The resulting main building blocks of I&XAI are presented as different levels of explainability. These include a general design space spanning eight dimensions in Section 3.2; a formalization of the explainability aspects from the design space into the *explAIner* pipeline with five properties in Section 3.3; as well as a conceptual framework encompassing three explainability spaces in Section 3.4.

All these explainability levels inform the work conducted for this dissertation and, in turn, its structure. In particular, one of the aspects that we are considering in the presented approaches is to answer the question of *how to tailor human-AI interfaces to specific stakeholders to achieve the goals of I&XAI*. To answer this question in the context of visual text analytics, we investigate ways to perform the same I&XAI tasks of understanding, diagnosis, and refinement of content and topic models for four different stakeholders.

Explainability Tasks and Processes One of the dimensions in our design space is the process of I&XAI consisting of the three tasks of model understanding, diagnosis, and refinement, as described in Section 3.2.7. These tasks applied to the content and topic models are the main ones considered for each of the four techniques presented in this dissertation.

Explainability Stakeholders The second important dimension is the four stakeholders we identify in Section 3.2.1: domain experts, domain analysts, model analysts, and model experts. These inform the design of our four techniques as we tailor each approach to one of them.

Explainability Spaces One of the main contributions of this dissertation is the introduction of the explainability spaces framework. It structures the design of I&XAI techniques across a spectrum based on the expertise and demands of the stakeholders. As depicted in Figure 1.2, the three anchor stones of this framework are (1) the *Semantic Space*—designed for domain knowledge generation and externalization; (2) the *Input-Output Space*—enabling concrete problem-solving; and (3) the *Model Space*—tailored to expert model building.

Semantic Space This space is tailored to domain experts. It enables interactions on the domain semantics level, facilitating the *machine teaching* interaction paradigm. This space allows for *theory-driven* learning, as well as *top-down*, *deductive reasoning*. A detailed description of this space is introduced in Section 3.4.1 and an instantiation is presented in Chapter 4.

Input-Output Space This is the most common design space, targeting both domain and model analysts. It enables interactions on the level of input and output objects, thus allowing for concrete problem-solving, using a concrete domain theory and AI model for *task-dependent* learning. Consequently, it facilitates *contrastive reasoning*. A detailed description of this space is introduced in Section 3.4.2 and an instantiation with a focus on semantics for the domain analyst is presented in Chapter 5, followed by another instantiation with a focus on the model decisions for the model analyst in Chapter 6.

Model Space This space is tailored to model experts. It opens up the model’s inner workings, allowing for model-aware explanations and interactions. This space facilitates *bottom-up*, *inductive reasoning*. A detailed description of this space is introduced in Section 3.4.3 and an instantiation is presented in Chapter 7.

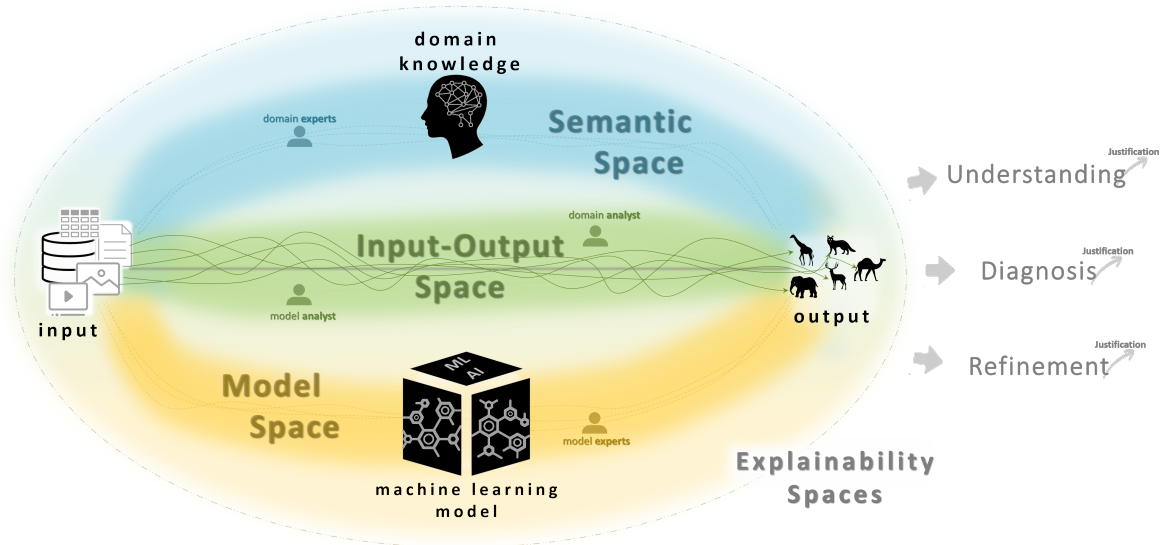


Fig. 1.2: The continuous spectrum of *Explainability Spaces* in machine learning, enabling the understanding, diagnosis, and refinement of machine learning models.

1.4 | Problem Statement and Research Questions

Combining the challenges described in the previous sections, we derive a set of open research gaps to address in this dissertation. Overall, we aim to tackle the challenge of designing effective I&XAI interfaces for content and topic model refinement. In particular, we strive to reveal the decision-making processes of the machine learning models, such that the explanations provided are *faithful* and *representative* to the model behavior, as well as *comprehensible* for the target audience, exemplified by our four stakeholder groups.

Hence, one of the primary considerations for our research is to investigate how we can adapt the traditional interactive machine learning workflow, as depicted in Figure 1.3, to novel interaction paradigms informed by our conceptual framework. More specifically, we would like to investigate how the derived mental models, knowledge representations and interaction paradigms, as well as the different learning and reasoning processes in our *explainability spaces* framework, affect the design of I&XAI workflows. In addition, we are interested in figuring out whether adapting the I&XAI frameworks to our four stakeholder groups will result in generalizable *design patterns and rationales* that would inform guidelines and best practices for other applications, domains, and tasks.

Therefore, to constrain this dissertation’s design space, we selected one class of machine learning models, i.e., content and topic models; and fixed the tasks and data types of the techniques we developed. We chose, however, to vary the stakeholders according to our four groups. Doing so allows us to set an appropriate focus and control on the I&XAI workflow designs developed for this dissertation. In other words, by narrowing the application design space, we can investigate and compare targeted I&XAI design patterns for human-AI interaction that support collaborative problem-solving and decision-making.

Lastly, analyzing and evaluating the developed techniques enable us to assess the applicability of their design considerations. This furthers our understanding of effective *intelligence augmentation* practices and allows us to work towards establishing a structured methodology for I&XAI processes to support human-AI collaboration in data analytics.

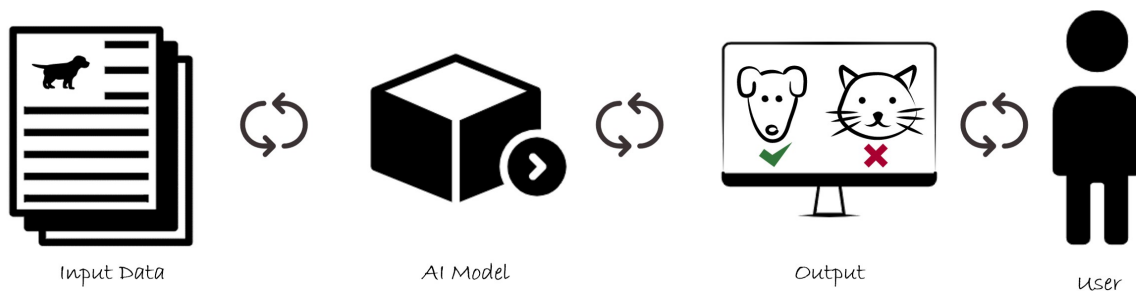


Fig. 1.3: Traditional workflow of interactive machine learning with a *linear, bi-directional* process.

Overall, this dissertation tackles the problem of mixed-initiative model refinement through I&XAI techniques. In particular, we address the following general research question:

How can we design effective and empowering communication interfaces for Human-AI collaboration?

To answer this broad research question, we break down our research into four more concrete ones, building the research agenda of the dissertation. In the following, we describe the four research questions alongside the research challenges they address:

RESEARCH QUESTION 1

What is the **Design Space** of I&XAI Applications?

I&XAI Design Space The first research challenge this dissertation aims to tackle is to define a design space of I&XAI applications. A summary of the contributions made towards addressing this research question is presented in Section 8.1.

RESEARCH QUESTION 2

Can we **tailor I&XAI Workflows** to empower Stakeholders with varying Expertise?

I&XAI Workflows The second research challenge is to identify factors to enable a targeted I&XAI workflow design. A summary of the contributions made towards addressing this research question is presented in Section 8.2.

RESEARCH QUESTION 3

How do we team up Humans & AIs for Mixed-Initiative **Content & Topic Model Refinement**?

I&XAI Applications The third challenge is to design I&XAI techniques for a concrete application, i.e., content and topic model refinement. A summary of the contributions to this research question is presented in Section 8.3.

RESEARCH QUESTION 4

Are there generalizable **Design Elements & Patterns** to distil and potentially transfer?

I&XAI Design Patterns The last challenge is to identify generalizable design patterns as possibly transferable best practices. A summary of the contributions made towards addressing this research question is presented in Section 8.4.

By tackling these four research challenges, in this dissertation, we contribute (1) a comprehensive description of the design space of I&XAI; (2) the conceptual explainability spaces framework; (3) four distinct visual analytics techniques for content and topic model refinement; and (4) twelve generalizable design patterns for I&XAI derived from the best practices of our proposed applications. With these contributions, we aim to make a step towards enabling seamless human-AI interaction for collaborative problem-solving and decision-making.

In addition to these overall contributions, every content chapter (i.e., the four techniques) and the framework chapter list their respective detailed contributions at the beginning of the chapters. In the following section, we also situate the research conducted and our contributions within the overall structure of the dissertation.

1.5 | Contributions and Structure of the Dissertation

As depicted in Figure 1.4, this dissertation is structured as follows; the first part describes the motivation and background, consisting of this introduction and the related work summary in the next chapter. The introduction portrays the three main areas related to this dissertation, i.e., Human-AI Interaction, Visual Text Analytics, and Levels of Explainability. This also serves as an explanation of the title of the dissertation. The introduction also gives an overview of the dissertation’s research challenges, contributions, and structure. Chapter 2 situates this dissertation in the current state-of-the-art literature. The three main related work areas considered are I&XAI, visual text analytics, and topic model refinement.

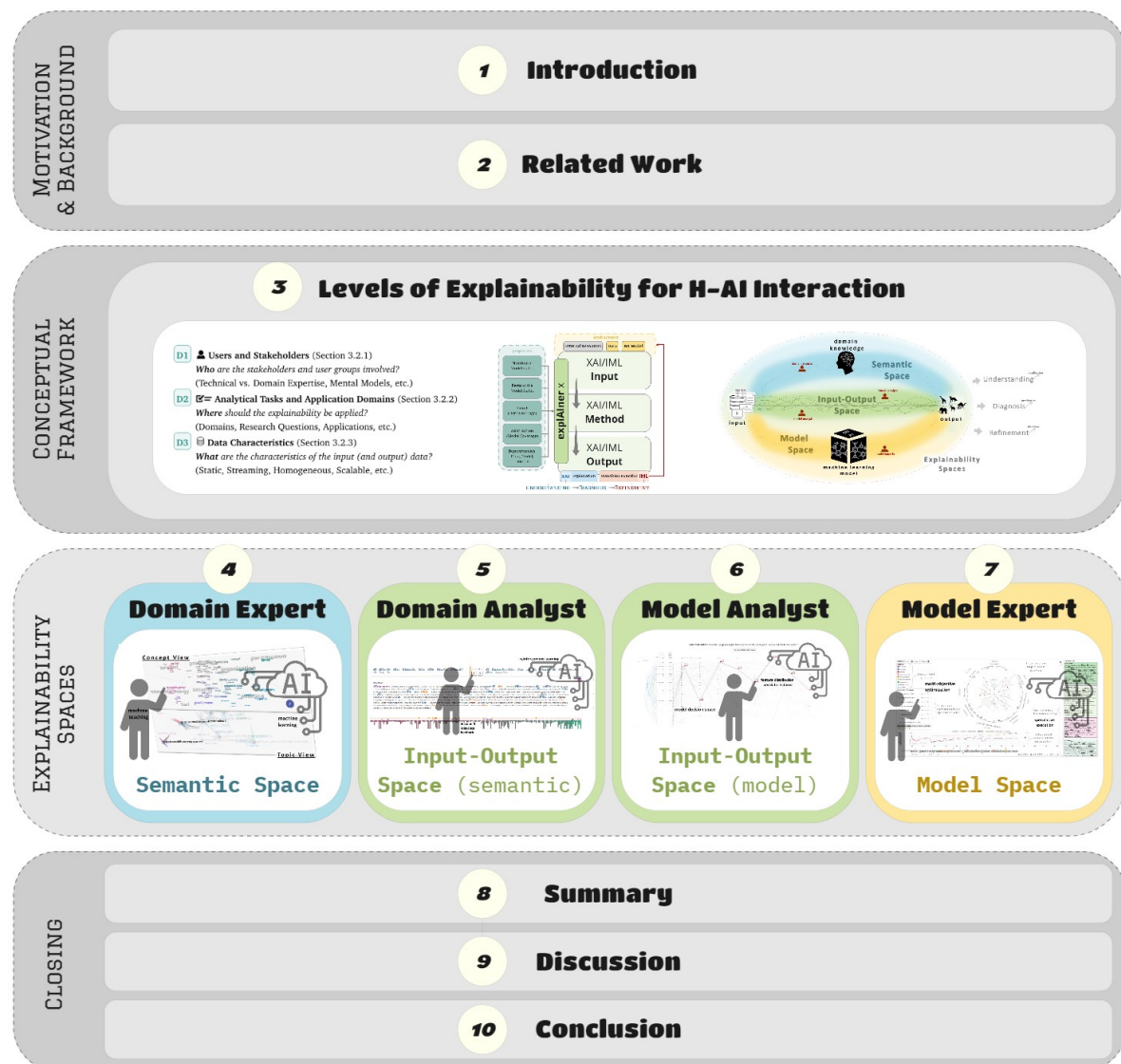
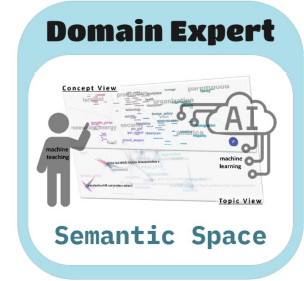


Fig. 1.4: Overview of the dissertation structure, divided into four parts spanning ten chapters.

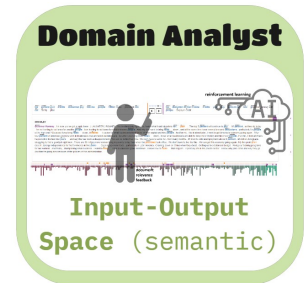
The second part of the dissertation presents the conceptual framework, one of the main contributions. Chapter 3 describes the levels of explainability for human-AI interaction, including a general design space spanning eight dimensions in Section 3.2; a formalization of the ex-

plainability aspects from the design space into the *explAIner* pipeline with five properties in Section 3.3; as well as the deduced conceptual framework encompassing three explainability spaces in Section 3.4. The introduced explainability spaces are instantiated for the use case of topic and content model refinement by the four main content chapters that follow.

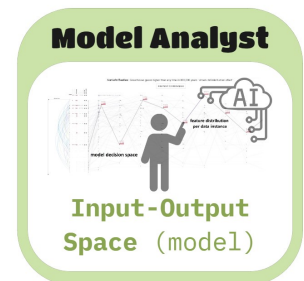
The first technique, introduced in Chapter 4, targets domain experts. It is an instance of the **Semantic Space** and demonstrates how we can enable *theory-driven* learning and *top-down, deductive* reasoning. The presented technique, *Semantic Concept Spaces*, allows the experts to externalize their domain knowledge without directly interacting with the topic model. The main contribution of this technique is facilitating the model refinement on the level of the user’s semantics through a *guided machine teaching* paradigm.



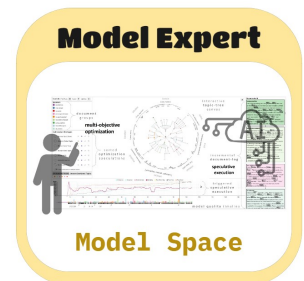
The second technique, introduced in Chapter 5, targets domain analysts. It is an instance of the **Input-Output Space** with a focus on the domain semantics, facilitating *task-dependent* learning and *contrastive* reasoning. The *progressive parameter learning* technique enables analysts to provide document-based relevance feedback by comparing the results of two different models. Through iterative refinement using *reinforcement learning*, the system trains new model instances aligned to the user-endorsed topic distributions.



The third technique, introduced in Chapter 6, targets model analysts. It is an instance of the **Input-Output Space** with a focus on models, facilitating *task-dependent* learning and *contrastive* reasoning. The presented technique, *ThreadReconstructor*, allows analysts to build and refine content models. In addition, it enables model exploration and comparison, as well as explaining the inner workings of the models through visualizing the *decision spaces* of their reconstruction.



The fourth technique, introduced in Chapter 7, targets model experts. It is an instance of the **Model Space**, facilitating *bottom-up* learning and *inductive* reasoning. The presented technique is based on an intelligible topic modeling algorithm and enables experts to view the model’s inner workings and algorithmic decision-making processes. To allow for targeted refinement, we introduce *speculative execution* as a paradigm for creating user-steerable previews, triggered whenever the measured model quality deteriorates.



The last part of the dissertation is the closing, with a summary in Chapter 8, a discussion in Chapter 9, and a conclusion in Chapter 10.

1.6 | Publications, Citations, and Terminology

This section clarifies the terminology used throughout the dissertation. In addition, it lists the citation rules and the publications the dissertation builds upon and further relevant ones. Overall, in the dissertation, I follow the academic convention of writing in the third person using the editorial ‘we.’ I only switch to the first person and use personal pronouns in this section and the contribution clarification boxes at the beginning of the content chapters. This is to make clarifying contributions, naming conventions, and citation rules easier to understand.

Terminology Clarification This dissertation uses the established terminology of the scientific communities of visual analytics, computational linguistics, and machine learning. In the remainder of this dissertation, I interchangeably use the acronyms ML and AI to refer to the field of machine learning (ML) and the broad research area of artificial intelligence (AI) when describing their research outcomes and models. To avoid confusion, however, I define the combined research area of interactive and explainable ML/AI as I&XAI.

Citation Rules Following the accepted scientific practices in the field of computer science, all technical contributions of this dissertation have been previously published in scientific journals. All publication parts that appear in this dissertation have been written by me with only slight editing from my co-authors. Hence, based on the established scientific community guidelines on self-plagiarism, I only rephrase parts of the technical descriptions from these publications to make them fit into the overall dissertation, leaving the remainder of the technical descriptions as verbatim excerpts from the papers. Beyond the five publications listed in the next paragraph, I explicitly quote and cite any other contribution that is based on my own research. In addition, I adhere to the accepted scientific practices for citation of any further works beyond my own. These citation rules allowed me to author a well-rounded dissertation that integrates my existing work into a comprehensive thesis.

Dissertation Publications and Contribution Clarification This dissertation is based on five main publications, as listed below. In the following, I briefly summarize each publication’s contributions. The papers are ordered by their appearance in the dissertation’s chapters.

Chapter 3 is partly based on the following publication:

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and **Mennatallah El-Assady**. “explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 1064–1074 [Spi+20]

This publication resulted from a collaboration aiming to structure and operationalize I&XAI pipelines. The paper consists of two distinct contributions, a conceptual framework, and a system implementation. The conceptual framework was devised and written by me, while the system implementation was developed and written by the first two co-authors. Only the former is part of this dissertation, constituting the *explAIner* pipeline described in Section 3.3.

Chapter 4 is based on the following publication:

Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. “Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 1001–1011 [El-+20]

This publication is based on a system implementation by Rebecca Kehlbeck as part of her master’s project, which I closely supervised. The implementation was integrated into the lingvis.io framework. After the completion of her project, I extended and refined the approach, conducted a user study, analyzed its outcomes, and wrote the paper based on the findings.

Chapter 5 is based on the following publication:

Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. “Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 382–391 [El-+18b]

This publication is based on a system implementation by me with the help of Rita Sevastjanova and Fabian Sperrle as research assistants. The implementation was integrated into the lingvis.io framework. Based on the system implementation, I conducted a user study, analyzed its outcomes, and wrote the paper. This publication won the *honorable mention for best paper* at IEEE VAST in 2017 (one of the top three papers for the year).

Chapter 6 is based on the following publication:

Mennatallah El-Assady, Rita Sevastjanova, Daniel Keim, and Christopher Collins. “Thread-Reconstructor: Modeling Reply-Chains to Untangle Conversational Text through Visual Analytics”. In: *Computer Graphics Forum* 37.3 (2018), pp. 351–365 [El-+18a]

This publication is based on a system implementation by Rita Sevastjanova as part of her master’s project, which I closely supervised. The implementation was integrated into the lingvis.io framework. After the completion of her project, I extended and refined the approach, conducted a user study, analyzed its outcomes, and wrote the paper based on the findings.

Chapter 7 is based on the following publication:

Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. “Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 374–384 [El-+19c]

This publication is based on a system implementation by Fabian Sperrle as part of his master’s project, which I closely supervised. The implementation was integrated into the lingvis.io framework. After the completion of his project, I extended and refined the approach, conducted a user study, analyzed its outcomes, and wrote the paper based on the findings.

Other Relevant Publications during Doctoral Studies In addition to the publications presented in this dissertation, during my doctoral studies, I co-authored over 65 peer-reviewed publications, including seven TVCG and eight CGF papers, as well as publications at venues ranging from the digital humanities to computational linguistics and machine learning. Most notably, the following seven publications highlight some of the research I contributed during this time.

Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. “ConToVi: Multi-Party Conversation Exploration using Topic-Space Views”. In: *Computer Graphics Forum* 35.3 (2016), pp. 431–440 [El-+16]

Mennatallah El-Assady, Annette Hautli-Janisz, Valentin Gold, Miriam Butt, Katharina Holzinger, and Daniel Keim. “Interactive Visual Analysis of Transcribed Multi-Party Discourse”. In: *Proceedings of the Association for Computational Linguistics, System Demonstrations*. ACL, 2017, pp. 49–54 [El-+17a]

Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. “NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations”. In: *Computer Graphics Forum*. Vol. 36. 3. 2017, pp. 213–225 [El-+17b]

Michael Behrisch, Michael Blumenschein, Nam Wook Kim, Lin Shao, **Mennatallah El-Assady**, Johannes Fuchs, Daniel Seebacher, Alexandra Diehl, Ulrik Brandes, Hanspeter Pfister, et al. “Quality Metrics for Information Visualization”. In: *Computer Graphics Forum*. Vol. 37. 3. 2018, pp. 625–662 [Beh+18]

Adam James Bradley, **Mennatallah El-Assady**, Katharine Coles, Eric Alexander, Min Chen, Christopher Collins, Stefan Jänicke, and David Joseph Wrisley. “Visualization and the Digital Humanities”. In: *IEEE Computer Graphics and Applications* 38.6 (2018), pp. 26–38 [Bra+18]

Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou, Fangxin Ouyang, **Mennatallah El-Assady**, Liu Jiang, and Daniel A Keim. “Bridging Text Visualization and Mining: A Task-driven Survey”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.7 (2018), pp. 2482–2504 [Liu+18c]

Mennatallah El-Assady, Wolfgang Jentner, Fabian Sperrle, Rita Sevastjanova, Annette Hautli-Janisz, Miriam Butt, and Daniel Keim. “lingvis.io - A Linguistic Visual Analytics Framework”. In: *Proceedings of the Association for Computational Linguistics, System Demonstrations*. ACL, 2019, pp. 13–18 [El-+19b]

Related Work

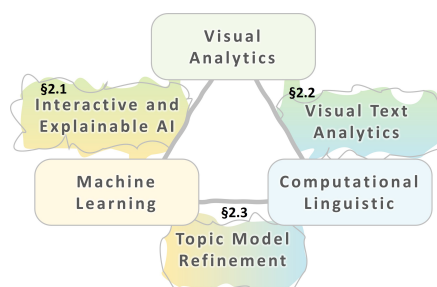
“ If I have seen further
it is by standing on the shoulders of Giants.

— Isaac Newton

Letter to Robert Hooke, 1675

Summary

This chapter situates the dissertation in the context of related research from different areas. In general, the techniques presented in this dissertation use methods from three fields: machine learning, visual analytics, and computational linguistics. In particular, the related work chapter gives an overview of the following: interactive and explainable AI techniques in Section 2.1 (at the intersection of ML and visual analytics); visual text analytics in Section 2.2.3 (at the intersection of visual analytics and computational linguistics); and topic modeling refinement in Section 2.3 (at the intersection of computational linguistics and ML).



Contents

2.1	Interactive and Explainable Artificial Intelligence	20
2.1.1	<u>Algorithm-Centered</u> Interactive and Explainable AI	21
2.1.2	<u>Human-Centered</u> Interactive and Explainable AI	23
2.1.3	Explainability Considerations for Effective Human-AI Interaction . . .	25
2.2	Visual Text Analytics	27
2.2.1	Task and Technique Overview	28
2.2.2	Analysis and Visualization of Conversational Data	30
2.2.3	The LingVis.io Framework	31
2.3	Content and Topic Model Refinement	33
2.3.1	Content Analysis and Topic Modeling	34
2.3.2	Visualization, Interpretation, & Refinement of Content & Topic Models	36
2.3.3	Topic Modeling Quality Metrics	38

2.1 | Interactive and Explainable Artificial Intelligence

Motivated by the ubiquitous presence of AI-based systems in society and their potential applicability to various use cases, several regulators and funding agencies around the world have launched recent programs to incentivize the research and development of explainable artificial intelligence systems, as well as ethical guidelines for regulating AI systems through explainability. For instance, in the United States, the Defense Advanced Research Projects Agency (DARPA) launched its explainable artificial intelligence program in May 2017, intending to create “more human-understandable AI systems through the use of effective explanations” [Gun16; GA19]. In an effort to address the ethical and legal concerns related to AI-based techniques, the European Parliament adopted the General Data Protection Regulation (GDPR) in May 2018 with the goal of providing “meaningful explanations of the logic involved” [GF17] in decision-making. Subsequently, the European Commission established an independent High-Level Expert Group on Artificial Intelligence (AI HLEG) in June 2018, which was “tasked to draft two deliverables: AI Ethics Guidelines and Policy and Investment Recommendations” [Smu19].

In April 2019, AI HLEG presented their Ethics Guidelines [HLE19] for Trustworthy Artificial Intelligence, outlining that trustworthy AI should be lawful, ethical, and robust. The guidelines name four groups of stakeholders: Developers, deployers, end-users, and the broader society. It identifies the following requirements for trustworthy AI systems: (1) **Human agency and oversight**, including fundamental rights, human agency, and human oversight. (2) **Technical robustness and safety**, including resilience to attack and security, fallback plans and general safety, accuracy, reliability, and reproducibility. (3) **Privacy and data governance**, including respect for privacy, quality and integrity of data, and access to data. (4) **Transparency**, including traceability, explainability, and communication. (5) **Diversity, non-discrimination and fairness**, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation. (6) **Societal and environmental wellbeing**, including sustainability and environmental friendliness, social impact, society and democracy. (7) **Accountability**, including auditability, minimization and reporting of negative impact, trade-offs, and redress.

This need for explainable artificial intelligence has sparked a surge in novel techniques, accompanied by an array of related publications. Nowadays, the field of interactive and explainable AI (I&XAI) is a diverse area of research with varying goals. These include, achieving *intelligence augmentation* [Hor99], *auditing of models* [Kra+17], and *model steering* [Cai+19]. Additionally, I&XAI techniques are multidisciplinary and address a diverse range of applications [Tan+12; TM11; Lee+15]. This section summarizes the main directions and arguments related to this research area of I&XAI. First, in Section 2.1.1, we give an overview of existing I&XAI techniques, most of which are **algorithm-centered**. Second, in Section 2.1.2, we highlight the need for **human-centered** I&XAI techniques and showcase some applications. Lastly, in Section 2.1.3, we deduce eight *explainability considerations* for effective human-AI interaction.

2.1.1 | Algorithm-Centered Interactive and Explainable AI

In literature, there is no conclusive definition of I&XAI. Often three aspects are named; explainability, intelligibility, and interpretability. According to Guidotti et al. [Gui+18], “*explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.*” Intelligibility refers to inherently transparent models [WB19], such as generalized additive models (GAMs) [Hoh+19a]. Lastly, according to Molnar [Mol18], “*interpretable machine learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans.*” In this section, we summarize the main advancements toward these goals, highlighting I&XAI techniques, methods, and frameworks.

Explainable Machine Learning Explainable Machine Learning (or Explainable Artificial Intelligence, XAI) refers to the research area of explaining ML model behaviors, decisions, and outputs [AB18]. According to Mohseni et al. [MZR20], we can distinguish between multiple types of explanations, including “*How*” that demonstrate how the model works [LBL16]; “*Why*” that describe why a prediction is made [RSG16]; “*Why-not*” that enable users to comprehend the absence of outputs [Ver+10]; “*What-If*” that demonstrate the effects of changing the models’ inputs [CJH19]; “*How-to*” that illustrate how to achieve a specific output [Lim+19]; and “*What-else*” that present similar inputs to test for similarity in the outputs [Cai+19]. Given the breadth of this research area and its methods, Arya et al. [Ary+19] argue for the need for multiple levels of explanations, e.g., model agnostic multi-level explanations [Ram+20]. However, countering this trend, Rudin [Rud19] argues to favor inherently interpretable rather than posthoc explainable models for high-stake decisions, as the latter usually uses a non-intelligible explanation method (black-box) to interpret a non-intelligible machine learning method (another black-box). In this dissertation, we follow the call by Rudin [Rud19] and present an inherently explainable topic modeling algorithm in Chapter 7.

Interactive Machine Learning Interactive machine learning (IML), as coined by Fails and Jr [FJ03], describes supporting users to cooperate with a machine learning model, for example, for parameter adjustment or model building. An early example of a process for the interactive building of classifiers was presented by Ware et al. [War+01]. Such processes enable humans to be in the loop and accelerate model development and scientific progress [Aod+14]. According to Dudley and Kristensson [DK18], the main tasks for interactive machine learning are feature selection, model selection, quality assessment, termination assessment, model steering, and transfer. In their survey on human-AI interaction, Maadi et al. [MAA21] review the reasons for humans to interact with AIs, listing three aspects, namely: (1) task complexity, (2) model complexity, transparency, and explainability, and (3) output dissatisfaction. They give examples from the medical domain for interactive machine learning, such as the approaches by Girardi et al. [GKH15] and Holzinger [Hol16]. In this dissertation, we introduce techniques to interact with machine learning models that pursue the goals described by Maadi et al. [MAA21].

Surveys and Frameworks In the following, we review the surveys in the field of I&XAI, starting with the ones from the visualization community. Endert et al. [End+17] and Liu et al. [Liu+17b] survey visual methods for interacting with machine learning, while Sacha et al. [Sac+19] present an ontology for visual analytics assisted ML. On the machine learning side, several surveys attempt to structure the field, formalizing their findings into specialized frameworks. Most prominently, Guidotti et al. [Gui+18], as well as Adadi and Berrada [AB18] provide exhaustive overviews of algorithm-centered XAI methods. Arrieta et al. [Arr+20] review opportunities for I&XAI in different application domains. Jiang et al. [JLC19] review advances in interactive machine learning. Wang et al. [WWL19] focus on what they refer to as crowd-assisted machine learning. Dudley and Kristensson [DK18] review and identify solution principles for effective explainability interface building. Finally, Mohseni et al. [MZR20] present a nested model for the design and evaluation of explanation methods, inspired by Munzner [Mun09]. However, each of these surveys focuses on specific aspects of XAI or IML, and none provides a comprehensive design space for I&XAI [Dha+21]. Thus, in Section 3.2 we propose our I&XAI design dimensions, distilling the lessons learned from these surveys.

I&XAI Methods As mentioned above, multiple surveys structured the field of I&XAI, in the following, we summarize some notable methods based on four categories. The first category consists of *inherently interpretable methods*, these include, for example, decision trees or support vector machines. We can utilize such methods to train simplified models [TKC17], interactively build models, e.g., *Baobabview* [VW11], or use them as surrogate models, e.g., *RuleMatrix* [MQB19]. Such methods also encourage interactive model building, e.g., *Gamut* [Hoh+19a]. The second category consists of methods for *model probing and sensitivity analysis*, these include, for example, *adversarial testing* methods [NQ17; RD18], *counterfactuals* [WMR18] to test the models' decision boundaries, as well as contrastive explanations, e.g., *What-If analysis* [Wex+20]. The third category consists of attribution methods, these include occlusion- or perturbation-based methods, e.g., *LIME* [RSG16], feature-based attribution, e.g., *SHAP* [LL17], gradient-based methods, e.g., *GradCAM* [Sel+17], pixel-based attribution, e.g., *Saliency Maps* [SVZ14], and many others. The last category consists of methods for activation and feature visualizations [OMS17], e.g., *SUMMIT* [Hoh+19b]. We classified these methods based on their explanation approach. However, we can also group them by their properties, as described in Section 3.3.1. Note, that many of the existing algorithm-centered I&XAI methods can be misleading [Lip18], error-prone [Dom+19], or unrobust [GAZ19].

I&XAI Goals and Challenges According to Guidotti et al. [Gui+18], interpretable machine learning models have to fulfill certain criteria, the *desiderata of interpretable models*, including comprehensibility, accuracy, fidelity, fairness, privacy, monotonicity, usability, reliability, robustness, causality, scalability, and generality. For more details on the different goals and values of I&XAI, please refer to Section 3.2.6. In addition, I&XAI approaches face a number of challenges, for example, producing *faithful* explanations to the underlying AI model (high fidelity) while still remaining *comprehensible* [WB19]. They also have to ensure they use suitable objectives and adequate features, as well as provide *trustworthy* explanations [Spe+20a].

2.1.2 | Human-Centered Interactive and Explainable AI

Based on the findings of the previously-mentioned reviews, most existing I&XAI approaches are algorithm-centered, expert-level systems. This is problematic because with a majority of users being non-experts, such interfaces are neither inclusive nor empowering [JSG06]. In addition, I&XAI techniques do not differentiate between the contexts in which explanations are used [ER20]. Hence, motivated by these issues and recognizing that not all humans are experts in machine learning, there was a shift to give “*power to the people*” [Ame+14]. For example, Doshi-Velez and Kim [DK17] describe high-level design goals for I&XAI, including, safety, ethics, user reliance, and scientific understanding. In this section, we describe the advances in human-centered I&XAI, emphasizing visual and interactive techniques.

Human-Centered Machine Learning According to Sperrle et al. [Spe+21a], “*human-centered machine learning is a field of research that considers humans and machines as equally important actors in the design, training, and evaluation of co-adaptive machine learning scenarios.*” Adding to that, Riedl [Rie19] defines human-centered AI as “*a perspective on AI and ML that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity.*” In addition, Shneiderman [Shn20] argues for the advancement of human-centered AI to *empower*, rather than replace humans. To achieve this, we need to understand the user mindsets and cater to them through human-centered design [EBH21]. We have to consider how explanations resonate with the users’ beliefs and understanding of the models [Kul+13; Kau+20]. Furthermore, we need to investigate whether the users’ *mental model* [Nor83] aligns with the systems’ conceptual model. Lastly, we need to balance human control and *machine agency* [Sun20]. To tackle these challenges, in Section 3.4, we introduce our explainability spaces framework consisting of three design spaces that are tailored to specific stakeholder needs, backgrounds, and expertise.

I&XAI Interface Design Visual, verbal, and interactive interfaces are the gateway of human-AI communication; they showcase the AI’s results and explanations while allowing humans to provide feedback for refinement. Dudley and Kristensson [DK18] survey user interface design for interactive machine learning and structure the main components of an interactive ML pipeline to be the user, AI model, data, and interface. This is the same definition we adhere to for the typical workflows, e.g., as depicted in Figure 1.3. However, effective human-AI interaction is considerably challenging to design, Yang et al. [Yan+20] present a framework that examines these design challenges. They illustrate that the challenges that designers face extend to, e.g., understanding AI capabilities, mapping out the right user stories, and collaborating with AI engineers. Hence, there is a rising demand for human-centered AI design practices [Aue20], in particular through combining AI and HCI methods [Xu+21]. These two fields complement each other to enable human-AI interaction [Gru09]. Tailoring I&XAI interface design to different stakeholders and *sociotechnical* application contexts [Ehs+21] reduces the communication friction between humans and AIs and facilitates effective collaboration [LGM20].

Visual Analytics for I&XAI The role of visual analytics in human-centered machine learning gained significant attention in recent years. Visualizations are typically applied to provide a comprehensive understanding of algorithms, support the exploration of applying ML models on specific datasets, show the effect of changing parameters, and enable users to tune the models to their tasks and data. Especially methods to explain neural network models, and other black-box models, are in high demand. Hohman et al. [Hoh+18] surveyed visualization and visual analytics methods in deep learning research. They derived six directions of open research problems; namely, (1) furthering interpretability; (2) system and visual scalability; (3) design studies for evaluation: utility and usability; (4) the human role in interpretability; (5) social good and bias detection; and (6) protecting against adversarial attacks. In this dissertation, we address the first and fourth challenges identified by Hohman et al. [Hoh+18] by introducing a conceptual framework describing the levels of explainability and interaction in human-centered machine learning. In particular, we investigate the factors influencing human-AI interactions to enable effective I&XAI for problem-solving and decision-making.

Visual I&XAI Techniques According to Liu et al. [Liu+17b] and Spinner et al. [Spi+20], I&XAI encompasses a process comprising the tasks of understanding, diagnosing, and refining AI models, their behavior, outcomes, and decision-making. Visual I&XAI techniques for *understanding* include, for example, understanding model decisions [MQB19; Als+18], understanding model architectures and activities [Rau17; Min+17; Won+18; Kah+19], and understanding model behaviors [Str+18b; Str+19]. Visual I&XAI techniques for *diagnosis* include, for example, diagnosing model outputs [Kra+17; Zha+19; Wex+20], diagnosing model robustness [Liu+18a; Ma+20], diagnosing neural activations [Kah+18; Wan+19b; Hoh+19b], diagnosing AI processes [Liu+18b], and diagnosing fairness [AL19; Cab+19]. Lastly, visual I&XAI techniques for *refinement* include, for example, refining single models [Liu+17a; Pez+18; Kwo+19], multimodel refinement [Mur+18], iterative refinement [Cai+19], and multilevel refinement [Ary+19; Ram+20]. In this dissertation, we introduce four visual I&XAI techniques all targeting the three tasks of understanding, diagnosis, and refinement.

Going Beyond Visualization As described above, there is a multitude of works that propose interactive visual analytics techniques and novel workflow designs to tackle the challenges of human-centered I&XAI. These range from understanding a machine learning model’s input data distributions and feature compositions; to optimizing the model’s learning process and comparing its results to other models; deploying a refined model on various data collections; and finally, reporting the model’s outcomes [Mit+19]. While these methods rely primarily on visual explanation, various other explanation mediums and strategies can be effective [El+19a], e.g., narrative storytelling [SH10]. An overview of these mediums and strategies is provided in Section 3.2.8. In addition, we need to consider appropriate metaphors [Jen+18] and contextualizations [Bré99; BB05; Bré07] for adequate explanations. Expanding these arguments, in the following section, we introduce eight explainability considerations derived from arguments in the literature and utilize them in our I&XAI design space.

2.1.3 | Explainability Considerations for Effective Human-AI Interaction

Following the call for *concrete Human-AI assemblage* [Ehs+21], we derive eight *explainability considerations* for effective human-AI interaction, as depicted in Figure 2.1. These build upon the algorithm-centered desiderata for interpretations by Murdoch et al. [Mur+19], and the human-centered guidelines for human-AI interaction by Amershi et al. [Ame+19].

- (C1) **Stakeholder & Context Dependency:** We know from psychology research that humans contrast explanations with their prior knowledge and current circumstances [Hil90], thus **explanations depend on the audience and context** [FZB11]. Hence, we need to consider what the background and expertise of the stakeholders are; what their *knowledge gap* is, for which they require an explanation [Sac+16]; and in which *socio-organizational context* they utilize the AI [Ehs+18; Ehs+21].
- (C2) **Rationalization:** Humans tend to search for probable connections to attribute and understand phenomena based on their existing knowledge [HSK95; Hil04], oftentimes they tend to be *overconfident* [Bia+05; FK12]. Due to this bias, we need to consider that **explanations can lead to harmful rationalizations** [Esl+15]. On the one hand, they can lead I&XAI algorithms learning from past data and interactions to “*fair wash*” decisions [Aiv+19]. On the other hand, they can lead humans to *rationalize* an action or decision, connecting them to their beliefs, values, and mental models regardless of actual causes and effects [Cus19]. To avoid harmful rationalizations, systems can, for example, have a tailored rationale generation [ER20].
- (C3) **Conditionality:** We need to consider that **explanations are conditional**; they depend on a set of factors that may not be obvious to the audience and can thus confuse or mislead [BLL12]. Model probing and counterfactual explanation can allow for *contrastive explanations* [Mil19] that distinguish true causes and probabilities from likely side effects [HJ07]. In addition, communicating the precise constraints of an explanation can foster a more trustworthy interaction [Ame+19].
- (C4) **Simplification:** Many AI algorithms have complex architectures and thus complex decision-making processes. We need to consider that **explanations can over-simplify complex models**. In contrast, some simpler models might have straightforward or even obvious decision-making processes; here, **explanations have the potential to over-complicate these simple models** [End+17]. We thus have to consider the abstraction level and the simplification level of an explanation for effective human-AI interaction [CW20].

- (C5) **Narratives & Metaphors:** Explanations build narratives, as they are a form of storytelling [Che+18]. Hence, to ensure effective communication, **explanations need to use the right metaphors**, i.e., respect mental models, value systems, and beliefs [Ban+19]. Introducing new and objective *metaphorical narratives* [Jen+18] can facilitate the calibration of trust between humans and AIs.
- (C6) **Mutual Empowerment:** Humans and AIs have different strengths and weaknesses. Through targeted explanation and interaction design, they can mutually empower each other. Hence, **explanations have the potential of bestowing superpowers on humans through AIs and vice versa** [Wil+21].
- (C7) **Co-Adaptation & Trust-Calibration:** *Explanations are part of a dynamic process*, hence to support sequential decision-making [Edw54], both, humans and AIs need to co-adapt [Spe+21b]. Furthermore, **explanations allow for a mindful calibration of trust**, avoiding, both, over- and under-trust. This is important to know when humans can trust an AI [Spe+20a], but also when an AI can trust human interaction [Spe+19].
- (C8) **Problem-Solving & Decision-Making:** The ultimate goal of an **explanation is to close knowledge gaps** [Sac+16]. This happens in two directions, users can learn about the AI or validate their knowledge; while AIs can be taught by users [Sim+17], e.g., through learning from their interactions [Ame+19]. This knowledge generation process underpins collaborative problem-solving and decision-making [Sac+14].

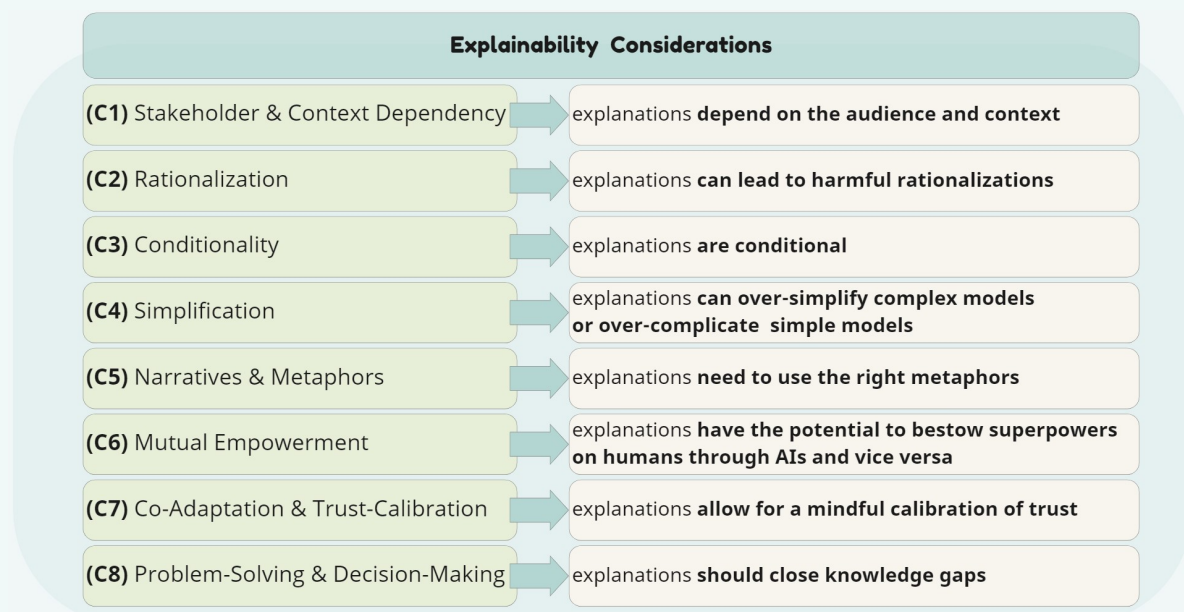


Fig. 2.1: Explainability considerations for effective human-AI interaction.

2.2 | Visual Text Analytics

The rapid advancement of text digitization and the move to online communication led to a substantial increase in available text data. Simultaneously, the emergence of robust and scalable natural language processing (NLP) methods has enabled large-scale and widespread automated text analysis. With the prevalence of such abundant data sources and processing methods came the demand for inspecting their results in different application contexts, comparing their performance on various tasks and parameter configurations, as well as applying such techniques in low- to no-code environments. Hence, to facilitate these tasks, the research field of *visual text analytics* emerged to combine the power of visual and interactive data analysis with the effectiveness of computational linguistics. As Liu et al. [Liu+18c] put it, “*The ultimate goal of visual text analytics is to enable human understanding and reasoning about large amounts of textual information in order to derive insights and knowledge.*” To date, this bustling research area has produced hundreds of techniques. Figure 2.2 shows some of the approaches that are categorized in the *text visualization browser* by Kucher and Kerren [KK15].

In this section, we give an overview of visual text analytics by reviewing different techniques and tasks in Section 2.2.1. In addition, in Section 2.2.2, we focus on conversational data analysis and visualization, a task that is relevant to the approaches presented in this dissertation, in particular, the *ThreadReconstructor* technique presented in Chapter 6. Lastly, we present the LingVis.io framework [El-+19b] in Section 2.2.3. It provides a modular, web-based platform for the development of visual text analytics applications, combining state-of-the-art computational linguistics methods with customizable, user-driven pipelines to produce visual analytics and I&XAI applications. This framework builds the basis for all of the techniques presented in the content chapters of this dissertation.

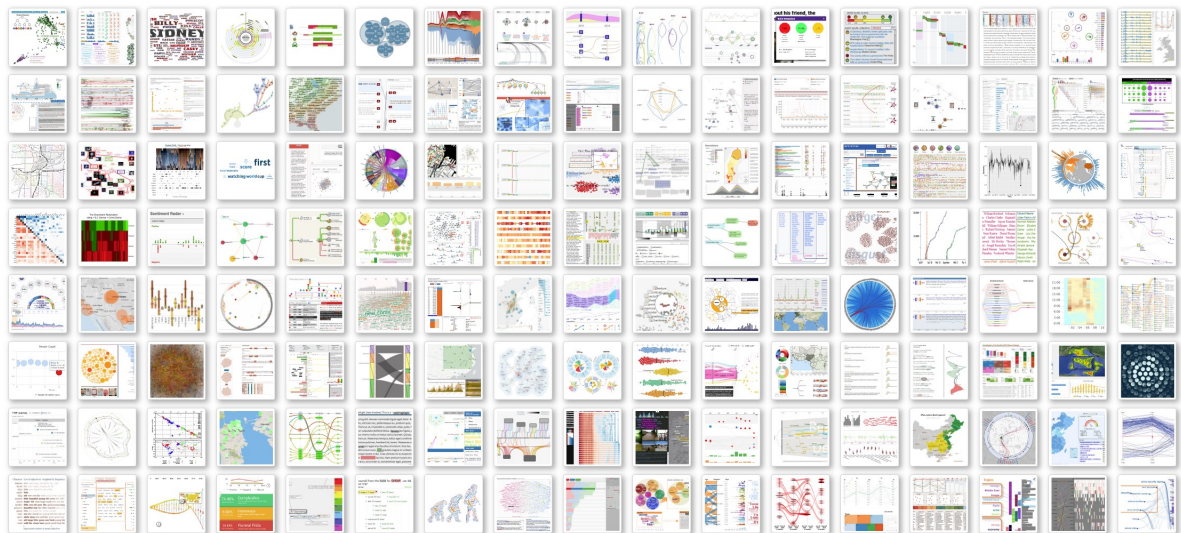


Fig. 2.2: A screenshot of the *text visualization browser* by Kucher and Kerren [KK15], providing a visual survey of over 400 text visualization and visual text analytics techniques.

2.2.1 | Task and Technique Overview

As introduced above, the field of visual text analytics has experienced rapid advancements in both computational linguistics and visual analytics. Hence, there have been several attempts to structure and categorize the methods at this research intersection. Notably, Alencar et al. [AOP12] contributed an early survey on visual text analytics reviewing around thirty methods. This was followed by a survey by Gan et al. [Gan+14] that focused on document visualizations. Later, Kucher and Kerren [KK15] contributed a visual survey of text visualization approaches, providing an online repository, the *text visualization browser*, where authors can include their new techniques. Their website currently catalogs over 400 techniques for visual text analytics, as depicted in Figure 2.2. In addition to these general surveys, Jänicke et al. [Jän+17] provide an overview of visual text analysis in digital humanities, while Federico et al. [Fed+17a] review approaches for analyzing scientific literature and patents. This section provides an overview of text analysis tasks, mining techniques, and visualization approaches. We rely primarily on the comprehensive task-driven survey by Liu et al. [Liu+18c] for our categorization.

Text Analysis Tasks Liu et al. [Liu+18c] identify five high-level analysis tasks, each consisting of multiple subtasks. The first high-level task is **information retrieval**, it describes the analysis aiming to obtain and structure information, such as through *entity ranking* [Bas+07; Agr+09; JBS17], *search* [AM01; Gom+14], or *filtering* [ZCM02; Lu+15; DXT17]. Other subtasks in this category include *structure analysis* [Ram+04; Oel+10; Pro+17], *recommendation* [Bro+07; Lv+11; Abb+13], *query analysis* [Hav02; Zha+11; SBN14], and *XML retrieval* [FG01; KRS04; Chu+06]. Furthermore, we can distinguish between structured retrieval, such as for *user activity tracking* [AWS06; BCB10], *unstructured retrieval* [Yu+03; Del+08; Fan+16], and *interactive retrieval* [Hav+01; SSS04; Hei+12]. The second high-level task is **content analysis and clustering**. As this is the most relevant for this dissertation, it is discussed separately in Section 2.3.1. The next high-level task is **natural language processing**, it describes linguistic analysis operations, such as *discourse analysis* [ASW12; Zha+12; Kwo+16], *sentiment analysis* [Oel+09; Wu+10; Wan+16a], *event analysis* [Fis+08; Luo+12; Cha+12], and *geotagging* [Ami+04; Yin+12; KHY15]. In addition, it covers subtasks such as *question answering* [Wan+09; Zha+07; She+17b], and *text summarization* [CNP06; Koc+14; Chi+15]. The fourth high-level task is **text classification**, describing different types of methods, such as *binary classification* [RW11; Sap+16; MDB17], *hierarchical classification* [DC00; Per+11; Ren+14], and *sub-document classification* [JKM12; EOS13]. It also covers more specialized subtasks, for instance, *cross language text classification* [GX12; XG13]. Lastly, the fifth high-level task is **exploratory analysis**, with subtasks ranging from *navigation* [Mad+14; BMS17], to *comparison* [CVW09; Dou+15], and *monitoring* [Als+11; Bos+13; SGA14]. These five high-level tasks and their subtasks represent the breadth of this research field. Note that the methods listed in this paragraph encompass visualization approaches, as well as linguistic analysis techniques. For a complete and detailed review of the tasks identified in the related work, please refer to the survey by Liu et al. [Liu+18c].

Text Mining Techniques Similarly to the task categorization, Liu et al. [Liu+18c] identify three high-level mining technique classes with subclasses each. The first high-level class is **data processing** methods. They operate on different levels of the text, i.e., on a *word, phrase, or entity-level* [Sch93; Sen12; ALA17], on a *sentence-level* [Wan+08; Li+15; Wan+16c], or on a *document-level* [WC06; RS08; OC03]. Furthermore, there are also *hybrid* methods [Bro+17; Hsi+17; Xia+17], combining different levels. The second high-level class is **modelling** approaches, including *language models* [Ben+03; Hua+10; Dau+17], *graphical models* [Pin+03; RW16; ZP17], *neural networks* [JZ15; Lai+15; Ma+16], and *mixture models* [Teh+06; BF11; ISU17]. Note that the underlying techniques for these approaches can be combined to form new models. These modeling approaches are also used on a variety of tasks, such as *classification* [Lod+02; DCP08; Cam+17], *clustering* [Son+10; Wan+13; Son+17], *dimensionality reduction* [SS97; Zha+15; Luo+17], and *regression* [SC02; KM16; HHL17]. However, the most relevant modeling subclass for this dissertation is *content and topic models* [BL06b; Kat+11; He+17]. These are described in more detail in Section 2.3.1. Lastly, the third high-level class is **model inference** techniques these include techniques for *probabilistic inference* [Liu+15; YN16; Lee+17a] and for *non-probabilistic inference* [Yua+10; DYY16; Du+17].

Text Visualization Techniques Lastly, analogously to the previous categorizations, the text visualization techniques are grouped into nine high-level classes by Liu et al. [Liu+18c]. These are further subdivided into various subclasses. The first high-level class consists of techniques that utilize **typographic visualizations**, for example, through *text highlighting* [McC+16; Liu+16; BMS17] or using *word clouds* [Lee+10; Str+12; Chi+15]. Moreover, *hybrid* approaches [CVW09; Cao+15; Sun+17] combine several typographic variables in one visualization technique. The second high-level class comprises techniques that utilize **chart visualizations**, including *scatterplots* [CPC09; Zha+14; Kim+17], *line charts* [Fis+08; Lee+10; HDA13], *tables* [Dou+15; Sza+16; JW17], and *bar charts* [CHF10; Hei+16; CCL16]. The next high-level visualization technique class is **graph visualizations**, such as, *node-link diagrams* [Sta+07; EFN12; El+17b], *trees* [Cua+07; WV08; Abd+13], and *matrices* [CCP09; ASW12; OKK13]. The fourth high-level class consists of techniques that utilize **timeline visualizations**, for example, by using *stream graphs* [Hav+02; KBK11; Wu+14] or *flow graphs* [WG12; Dou+13; FBM16]. The next high-level class comprises techniques that utilize **spatial projections**, including *galaxies* [Fuj+08; Kwo+16; Ise+17] and *voronoi tessellations* [And+02; SC04; Che+17]. The sixth high-level visualization technique class consists of **high-dimensional visualizations**, including *glyphs* [CMH12; OKK13; BKW16] and *parallel coordinate plots* [Koc+11; Lee+12a; Cho+16]. The last three high-level classes are less common for visual text analytics and, thus, do not have any subclasses. These are **topological maps** [Woo+07; Mac+11; Wan+15], **radial visualizations** [Chu+12; Wan+15; She+17a], and **3D visualizations** [Har+02; Tho+12; CC13]. Note that some of these visualization techniques are static, and others are interactive. In addition, some of the visualization techniques are directly applied to the text data, and others are applied to transformations of the data, for example, through linguistic modeling. This dissertation’s visualization techniques utilize this entire design space to derive their visual encoding for the various components.

2.2.2 | Analysis and Visualization of Conversational Data

Following the general overview of visual text analytics techniques and tasks, in this section we specifically focus on conversational data visualization. This is a task that is particularly relevant to the *ThreadReconstructor* technique presented in Chapter 6.

Challenges for Analyzing Conversational Data Transcripts of multi-party discourse, such as in forums and debates, are based on the interaction between speakers and, thus, comprise a structure of implicit reply relations. Some forum interfaces show such a structure in the form of threaded discussions. However, figuring out a possible reply-relation structure in other domains is a more challenging task. Moreover, analyzing such lengthy conversation transcripts often indicates multiple valid threaded structures, depending on the considered textual and semantic features. Therefore, our *ThreadReconstructor* technique is designed to support users in creating and refining reconstruction models to tailor them to their respective tasks and data.

Visual Analytics for Conversational Data Conversational data has been visualized using an approach called *thread arcs* [Ker03], which represents a discussion as a tree, with messages as nodes and reply relations as edges. Hubmann-Haidvogel [Hub08] created *ThreadVis*, a similar arc-diagram-based visualization of email threads. The chronology of messages is combined with the branching tree structure of a conversational thread to create a form of temporal arc diagram. Fu et al. [Fu+17] extend the basic thread arcs technique to present *thread river*, which can illustrate temporal and structural information of lengthy threaded discussions. In their tool, *iForum*, they offer a set of visualization designs for presenting the main interleaving aspects of MOOC forums at three different scales (posts, users, and threads). *ForumReader* [DWM04] is a tool combining visualization techniques with automatic topic extraction algorithms to help users explore *flash forums* (representing shallow threads). *ForAVis* [WRK11] allows exploring online forums using author-level, post-level, and discussion-level features, including sentiment analysis. Hoque and Carenini [HC14] have developed *ConVis* to support the multi-faceted exploration of blog conversations, which contains multiple views to provide thread information at different granularities. Our *ThreadReconstructor* approach borrows inspiration from these works, which visualize the thread structure embedded in the data. We extend this to construct a threaded relation over non-threaded data. In contrast to works tailored to the content-exploration of threaded forum data [HSS10; Che15], the *ThreadReconstructor* is designed to enable the generation and comparison of thread reconstruction models.

Content and Topic Modeling for Conversational Data Some techniques depict topic changes within discussions, e.g., *ConToVi* [El+16] showing the temporal speaker dynamics across topics; others use theme-river-based visualizations to represent temporal content evolutions [TG10; Liu+16]. Trampuš and Grobelnik [TG10] use a timeline to display a forum's topic evolution, and a semantic “*atlas*” to show a thematic overview. Liu et al. [Liu+16] provide an overview of the evolving hierarchical topics by connecting the corresponding topics at different times. More details on general content and topic visualizations are discussed in Section 2.3.2.

2.2.3 | The LingVis.io Framework

To facilitate interdisciplinary research at the intersection of computational linguistics, visual analytics, and explainable machine learning, we developed the LingVis.io framework [El-+19b]. The aim behind its development was to provide a platform that tightly integrates state-of-the-art linguistic computation with visual analytics to generate new insights for model and domain experts and analysts, such as political scientists working with large text corpora. This helped establish the research subarea of *Linguistic Visualization* (short: *LingVis*). Besides enabling visual linguistic tool development, we were particularly interested in explaining how natural language processing models transform and manipulate the input data to unveil patterns and detect linguistic structures.

As depicted in Figure 2.3, the LingVis.io framework combines several computational linguistic techniques and provides a multitude of interactive visual interfaces. It powers several mixed-initiative visual analytics approaches, ranging from generating high-level corpus overviews using *Lexical Episode Plots* [GRE15] to sophisticated I&XAI techniques for content and topic refinement, as described in this dissertation. Other applications include, content analysis, e.g., *NEREx* [El-+17b]; discourse analysis, e.g., *DiscourseMaps* [EH19]; argumentation analysis, e.g., *ConToVi* [El-+16]; verbalization and active learning [Sev+18a; Sev+18b]; and multi-corpora analysis, e.g., *AlignmentVis* [Jen+17]. Another prominent set of applications are the multi-party discourse analysis approaches in the context of the *VisArgue* project [El-+17a].

To support the development of such applications, we implemented the LingVis.io framework to be a modular web-based platform. In particular, the framework’s architecture accommodates the need for rapid prototyping, disguising the involved engineering complexity for application developers. Furthermore, the framework’s compartmentalization facilitates the rebuilding, sharing, and reusing of implementation components. As this framework is the foundation of the techniques presented in the content chapters of this dissertation, in the remainder of this section, we describe some of its technical details.

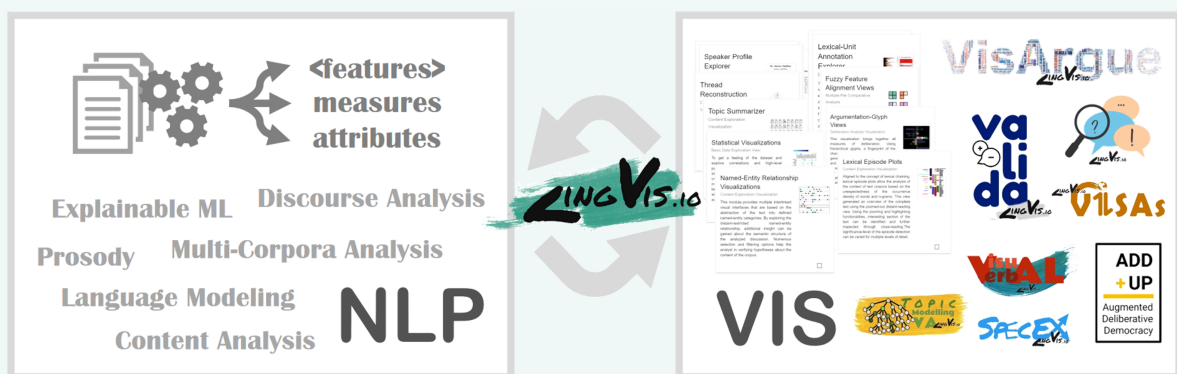


Fig. 2.3: LingVis.io powers many visual text analytics applications by modularly connecting state-of-the-art natural language processing methods with tailored interactive visualization interfaces.

Auto-Configured Processing Pipeline Our framework relies on defined dependencies between processing steps to enable application developers to request natural language processing steps for their visualizations. If a step is requested, the framework automatically assesses all its downstream dependencies and resolves them into a directed acyclic graph, i.e., it auto-configures the processing pipeline. Furthermore, each processing step has a visual representation in the interface to enable users to configure and adjust parameters interactively. Hence, our framework architecture strictly separates and modularizes tasks into atomic components that are compartmentalized in subdomains (i.e., auto-scaling cluster environments). This facilitates tailored interface design, rapid prototyping, and processing flexibility, enabling developers to focus their design on their novel contributions. To illustrate this concept, Figure 2.4 depicts an auto-configured processing pipeline for topic modeling.

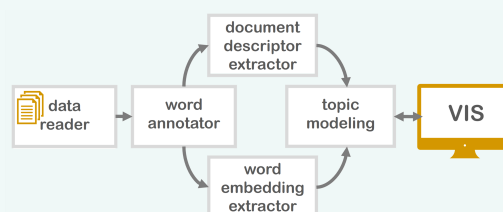


Fig. 2.4: LingVis.io Processing Pipeline.

The Document Descriptor Extractor A considerable advantage of relying on an auto-configured processing pipeline with modular components is that it enables a flexible, user-driven definition of parameters and controls. Notably, this led to the development of the *document descriptor extractor* component, as depicted in Figure 2.5. It allows users to configure the descriptor vectors used to represent the extracted sentences from the analyzed corpus. Users can opt to include certain word classes and exclude others. In addition, they can use computed features to augment the embedding vectors, for example, using n-grams, lexical episodes [GRE15], or named entities. In addition to this flexible *parametrization*, users can use the configuration interface to select scoring functions for document vectors, e.g., *tf-idf* [Jon72]. Lastly, they can also configure the internal weighting functions used in the content and topic modeling, e.g., *log-likelihood ratio* [MS99]. All techniques presented in this dissertation rely on these components to provide flexible model configurations.

Part-of-Speech Tags			Other Features and Settings	
Words with the selected POS tags will be included in the document vectors.			The selected feature words will be included in the document vectors.	
Open class words	Closed class words	Others	Named Entities	Bigrams
☒ Nouns	☐ Coordinating Conjunctions	☐ Symbols	☒ Episodes	☒ Trigrams
☒ Proper Nouns	☐ Adpositions (Pre, Post)	☐ Punctuations	☐ Stopwords	☒ Lemmas
☒ Adjectives	☐ Cardinal Numbers	☐ Others		
☒ Adverbs	☐ Determiners			
☒ Verbs	☐ Particles			
☐ Interjections				

Fig. 2.5: Settings menu for the *document descriptor extractor* component of the LingVis.io framework.

2.3 | Content and Topic Model Refinement

Topic modeling is a special case of document clustering that adds label generation [AZ12]. Most commonly, topic modeling is understood as a soft clustering problem [XX13], i.e., with probabilistic membership assignment of documents to topics [GS06]. However, other classes of algorithms are also applicable to achieve a meaningful partitioning of corpora. More specifically, four different classes of clustering algorithms can be distinguished: **(1) Generative clustering** algorithms are probabilistic and include the popular LDA topic model [BNJ03] and its diverse extensions (hierarchical [BGJ10], temporal [BL06b], with metadata [Ram+09a], for spoken text [Pur+06]) as well as many other models (Pachinko allocation [MLM07], Author Topic Model [Ros+04], HDP [Teh+06], pLSI [DLP08]). **(2) Spectral clustering approaches** like NMF [XLG03; KCP15], divide and merge clustering [Che+06], or fuzzy co-clustering [KDK03], perform a segmentation-based clustering on a dimensionality reduction of the data. **(3) Discriminative distance-based clustering** works on concept-, document- or keyword-distances [Fis87; GLF89]. **(4) Finally, word- and phrase-based clustering** incorporates text semantics or linguistic knowledge into the clustering process [BEX02; HK04]. Independent of the clustering algorithm used, labels for any topic can be created by specialized algorithms [Lau+10; Lau+11].

Topic Modeling is used to understand large corpora of text and summarize the knowledge contained in them. The basic premise of topic modeling is to cluster groups of documents and label them, obtaining topics. Probabilistic models often produce results of higher quality at the expense of reproducibility and the understandability of the algorithm. They often require the specification of the expected number of topics before the start of the modeling process. This makes these models harder to use in visual analytics systems promoting understanding and interactivity. Blei [Ble12b] provides an overview of probabilistic topic modeling algorithms. Several works explicitly address the embedding of domain knowledge into the topic space. Andrzejewski et al. [AZC09; And+11] use Dirichlet forest priors to split and merge concepts using domain knowledge, improving topic descriptors. Chen et al. developed the MDK-LDA variant on LDA which takes into account domain knowledge directly to provide better topic descriptors [Che+13]. Furthermore, approaches that combine word embeddings with topic modeling can be beneficial for learning both models jointly [Moo16], as well as improving topic model representations for short texts through word embeddings [Ngu+15; ZDB17; Li+17], or creating improved word embeddings using LDA [RWJ16].

In addition to clustering documents into topics, we can utilize such models to incorporate metadata beyond the thematic composition of a given corpus. In this dissertation, we mainly address the challenge of topic model refinement. However, in Chapter 6, we also extend this task to include content models for reconstructing threaded conversations. To give an overview of this research area, we introduce content analysis and topic modeling techniques in Section 2.3.1. Next, we discuss interpretation, visualization, and refinement approaches in Section 2.3.2. Lastly, we introduce eight topic model quality metrics in Section 2.3.3.

2.3.1 | Content Analysis and Topic Modeling

This section gives an overview of content analysis and topic modeling methods, their evaluation, optimization, and the applications they empower. In addition, we briefly describe content and topic modeling in the context of the LingVis.io framework.

Content Analysis According to Liu et al. [Liu+18c], content analysis is one of the central tasks in text mining with five main subtasks. First, *contextual text mining*, analyzing the thematic contexts of text corpora, e.g., using mixture models [MZ06] or network regularization [Mei+08]. The second subtask is *text segmentation* [Cho00], e.g., based on sequence modeling [Wan+17], anisotropic diffusion and dynamic programming [JZ03], genetic algorithms [Lam+07], or by exploiting hybrid contexts [Li+13]. The third subtask is *community discovery*, e.g., for social influence analysis [Tan+09], author community analysis [LNG09], or participation analysis [Zhe+11]. The fourth subtask is *document ranking and clustering* [LA99], e.g., using non-negative matrix factorization [XLG03; KP13] or covariance analysis [IPT16]. Clustering can also put documents into a hierarchical relation [TSS15]. The last subtask is *topic analysis and discovery*, this mainly relies on topic modeling algorithm to investigate different aspects [Bic+17], for example, hierarchical relationships of concepts [CSS08], conversation clusters [BK09], thematic summaries [EC08; Wu+17], or the topics' temporal evolution [Cui+14]. In this dissertation, we focus on refining such topic modeling algorithms to support the last subtask. More details on these algorithms are described in the next paragraph.

Topic Modeling Algorithms As described at the beginning of this section, there are four general classes of topic models. However, the two most prominent ones are probabilistic models [AGM12], including the popular Latent Dirichlet Allocation (LDA) [BNJ03] approach, and non-probabilistic approaches, notably, non-negative matrix factorization (NMF) [XLG03]. Probabilistic models (e.g., [Ros+04; Teh+06; BL06b; MLM07; Rei+10]) are the most prominent and are based on the assumption of the existence of a latent space [LD97] in which relations between objects are determined. Generally, probabilistic approaches can produce higher quality results [CSS06], but at the price of determinism and iterative stability. Both types of models have several factors in common, including *input parameters*, which specify the model characteristics, such as the number of topics or weightings on classes of words in the input documents, *keyword vectors*, which are ranked lists of words which represent extracted topics [KP02], and *document descriptors*, which are vectors of scores relating each document to each topic. A comprehensive survey of different probabilistic topic modeling approaches is provided by Blei [Ble12b]. Note that as topic modeling algorithm attempt to reduce the dimensionality of the document descriptors [Cra+12; SC16] to find patterns and correlations [BL06a], various optimizations can influence their output quality. These include, for example, using lexical priors [JDU12], reducing the number of parameters [Ngu+14], modeling characteristic linguistic properties [FY99; McF+13], or relying on specific data structures to support the modeling, e.g., directed acyclic graphs [LM06]. The methods used in this dissertation integrate some of these optimizations to enhance the modeling outputs.

Model Output Evaluation Content and topic modeling approaches usually attempt to satisfy multiple quality criteria in their modeling. Hence, due to their multi-objective optimization, various solutions and outputs can be equally plausible. Additionally, some objectives are subjective or do not align with human intuition, making them exceptionally tricky to capture [LNB14]. Therefore, evaluating the outcomes of topic models is a challenging task [Wal+09]. In addition, the subjectivity of the outcomes implies a lack of ground truth data to train on [Hei+12]. Tackling these issues, researchers have proposed multiple evaluation metrics for topic modeling outcomes, for example, topic coherence [NBB11; Ste+12]. We discuss the eight most prominent quality metrics in Section 2.3.3. In addition, a recent research trend has focused on evaluating effective visual representations of topic modeling outputs [Smi+17]. In this dissertation, we follow their recommendations for our proposed visual designs.

Knowledge Integration To meaningfully capture the thematic relations between documents, we need to model their vectors to capture semantic associations [GST07], moving beyond simple bag-of-words models [Wal06]. Hence, in addition to relying on embedding vectors [Dev+19], a fruitful approach to optimizing content and topic models is to enable knowledge integration [XYX15]. On the one hand, we can enable users to integrate their knowledge as priors, e.g., through defining anchor words before modeling [Vos14; Lun+17]. In Section 7.4.1, we describe a similar approach in our proposed workspace. On the other hand, we can introduce interventions during the modeling, e.g., by using active learning [Yan+14] or reinforcement learning [TC16]. Such approaches can support gradual personalization and refinement, as discussed in Section 5.5.1. Overall, integrating the users' knowledge can boost the modeling by introducing missing semantic connections [Che+08]. Lastly, as discussed in the following section, interactive visual interfaces can facilitate knowledge integration for model refinement [Cañ+05].

Content and Topic Modeling Applications As discussed above, content and topic models facilitate various tasks. They can operate on different types of data [Lin+08]. The volume of data can also vary, ranging from large text corpora [Gre+12] to short texts [Li+17], with specialized models, e.g., the *biterm* topic model [Yan+13] or *BTM* [Che+14]. Hence, such models can support a multitude of applications, especially in the humanities and social sciences [Sch12], for example, for education [DD06], conversation analysis [NBR12], or political agenda exploration [GC17]. However, content and topic models can also be useful beyond these traditional fields, e.g., for analyzing source code [Ehr17], online language [Der+14], or even in a medical context for disease subtyping [Glu+18].

Content and Topic Modeling in LingVis.io As described in Section 2.2.3, the LingVis.io framework allows for modular workflow generation through its auto-configured pipeline and through incorporating a flexible, user-driven parametrization of models in the visual interface. This has led to a constant integration of new content and topic models into the framework that support various downstream visualization applications. The approaches proposed in this dissertation allow for a targeted refinement of these models to support all downstream applications.

2.3.2 | Visualization, Interpretation, and Refinement of Content and Topic Models

In addition to developing novel content and topic modeling approaches, a growing research trend has focused on visualizing their outcomes, interpreting their results, and refining the models to suit different data and application contexts [Ble12a]. Thus, in this section, we describe these advancements, listing the most prominent techniques in this research area. In particular, we focus on approaches that use interactive visual interfaces to explain the models' results and facilitate user feedback integration.

Visual Content and Topic Analysis In recent years, various interactive visualization approaches have been developed for the content analysis and exploration of document collections [BHM17]. Largely, these are based on the LDA model [BNJ03], e.g., *LDAExplore* [GPC15]. Most commonly, approaches utilize ThemeRiver-based [Hav+02] visualizations to highlight temporal trends in topics, e.g., *TextFlow* [Cui+11], *RoseRiver* [Cui+14], *Visual Backchannel* [Doe+10], and *TIARA* [Wei+10]. Other approaches go beyond exploring the temporal dynamics and evolution, e.g., *SolarMap* [Cao+11], *TopicViz* [Eis+12], *Hiérarchie* [SHM14], *Termite* [CMH12], *Serendip* [Ale+14], and *VISTopic* [YYQ17]. Another set of systems does not only interactively visualize the modeling results but also allows users to change or regenerate the underlying model. Such interactions can add new model constraints to group keywords in a single topic [Hu+14], produce more detailed subtopics [Kim+17], manually merge or split topics [HC15; Dou+13], or modify weights for individual keywords [Cho+13]. Note that these systems recompute at least parts of the model after every interaction.

Visual Analytics for Model Refinement Exploratory visualizations for understanding topic spaces include *ParallelTopics* [Dou+11] for exploring single and multi-topic documents using parallel coordinates. However, in this dissertation, our work focuses not on viewing the topic modeling itself but on finding an intuitive way for the users to guide the modeling process. Visual topic modeling approaches often include interactive mechanisms for users to modify the modeling output. For example, *TopicPanorama* [Wan+16b] creates a graph of topic relations extracted from multiple sources. Interactive tools are embedded to allow users to modify the graph matching to suit their needs. *HierarchicalTopics* [Dou+13] is a visualization for understanding a large dataset at different levels of granularity. Interactions allow users to adjust the hierarchy. *UTOPIAN* employs a semi-supervised iterative feedback loop for users to steer the modeling process [Cho+13]. *ConceptVector* [Par+18] enables users to embed domain knowledge interactively by guiding the building of concepts that are then used to analyze documents. This idea of an interactive loop for refining topics also appears in the work of Hu et al. [Hu+14], in which users guide the modeling process through constraints on the topic descriptors. Lastly, Hoque and Carenini [HC15] embed similar feedback into a visualization system in the *ConVisIt* project. In this dissertation, we follow parallel directions to enable users to refine their content and topic modeling outputs.

I&XAI Processes for Content and Topic Modeling For an effective visual analytics process we enable active steering of the modeling process while it is making decisions, as proposed by Mühlbacher et al. [Müh+14], allowing our proposed techniques to quickly provide feedback on user interactions and avoid waiting times, or, in some cases, even avoid restarting the modeling process. Lee et al. [Lee+17b] specify a set of refinement operations for topic modeling that are intuitive for non-expert users, including splitting and merging topics or adding and removing keywords that have been implemented in some of our techniques. Furthermore, one of the most relevant visual analysis approaches to the techniques presented in this dissertation is the task-driven comparison of topic models by Alexander and Gleicher [AG16]. Using their so-called “*Buddy Plots*”, they highlight the differences in the modeling results between two different models by fixing one or multiple topics. In addition, their paper categorizes three topic modeling tasks: Understanding topics, understanding similarities, and understanding change. They perform the comparison of topic models using three techniques: topic alignment, distance comparison, and timeline comparison. While this paper paves the way for comparative topic analysis [Cro+11], the techniques proposed in this dissertation go beyond single comparisons and extends the analysis to iterative optimization cycles. As described in Section 2.1.2, Liu et al. [Liu+17b] note that “*most users often treat a machine learning model as a black box*” and have thus defined *understanding*, *diagnosis* and *refinement* as three important tasks for visual analytics systems. In this dissertation, our proposed visual workspaces address these three tasks with tailored visualizations and interactions for different stages of the I&XAI process.

Interpretation of Topic Modeling Outcomes A more recent trend around the analysis of topic models is the enhancement of the comprehension and interpretation of their results [LNB14]. This is motivated by the evidence that most currently used automated, likelihood-based quality measures for topic models do not capture their quality correctly. In fact, Chang et al. [Cha+09] found that they are negatively correlated with perceived quality. To improve this situation, we have to consider interpretability when designing models and tools, a fact that has often been overlooked in the past, as emphasized by Chuang et al. [Chu+12; Chu+13]. They contributed a set of guidelines that should be employed when developing new models. Trustworthy and reproducible topic models are crucial for social sciences, where Ramage et al. [Ram+09b] find there is strong and growing demand. In order to make systems more understandable for non-experts, Choo et al. [Cho+13] propose to give users the option to add individual weights to single keywords in order to reach a more understandable topic model. As illustrated in Section 2.2.3, our LingVis.io framework enables such interactions before and during model building. In addition, as described in the previous paragraph, Lee et al. [Lee+17b] isolated a few primitive interactions that were intuitive to non-experts, such as adding and removing keywords from topics. Both of these propositions are present in our proposed techniques. In some of our proposed techniques, we reveal the underlying internal quality metrics used in the multi-objective optimization during the modeling process to enhance the interpretability further. More details on these topic modeling quality metrics are described in the next section.

2.3.3 | Topic Modeling Quality Metrics

Assessing the quality of a topic model can be achieved using one of several available metrics or a combination thereof. While each of the metrics studied by Chang et al. [Cha+09] addresses some aspect of quality, they found the measures did not capture the human intuition of quality well, thus these measures alone cannot be used to optimize a model for human expectations.

The lingvis.io framework [El+19b] implements different topic model quality metrics. Some of them measure the structural properties of topic trees, while others operate on the keyword level within the topic descriptor keyword list. The eight most notable ones are:

- (M1) **Pointwise Mutual Information** [Bou09] measures the co-occurrence of the most frequent keywords in a topic.
- (M2) **Coherence** [Mim+11] measures the conditional probability of keywords given more frequent ones from the same topic. It performs especially well for very small and specific topics as well as large, generic ones [NKK17].
- (M3) **Separation** measures the inter-topic separability of keywords of a topic or subtopic, complementing coherence.
- (M4) **Distinctiveness** is the ratio of the previous two metrics, allowing us to optimize for coherent yet separable topics.
- (M5) **Certainty** is a metric that asserts the uniqueness of keywords for their respective topics or subtopics.
- (M6) **Variance** measures the similarities between subtopics of one parent topic as an indicator for topics that only contain equally similar subtopics.
- (M7) The **tree branching factor** measures the ‘fan-out’ of topic tree nodes. This is a structural graph property of a topic tree.
- (M8) **Compactness**, defined as the ratio between leaf nodes and inner nodes, is helpful to avoid document chaining artifacts that introduce sparse and deep trees.

Additional metrics include the **document entropy** [MM02], the **word mover’s distance** [Kus+15], and the **topic distribution size**. We also use the **number of topics**, the average **topic size** [MM02] and, if available, the **number of authors or speakers**. Topic modeling algorithms typically perform multi-objective optimization based on these and other metrics.

Conceptual Framework

” *Science is a way of thinking
much more than it is a body of knowledge.*

— **Carl Sagan**
(attributed)

Levels of Explainability for Human – AI Interaction

“ *The key to good decision making is not knowledge.
It is understanding. We are swimming in the former.
We are desperately lacking in the latter.*

— **Malcolm Gladwell**

Blink: The Power of Thinking

Without Thinking [Gla06]

Summary

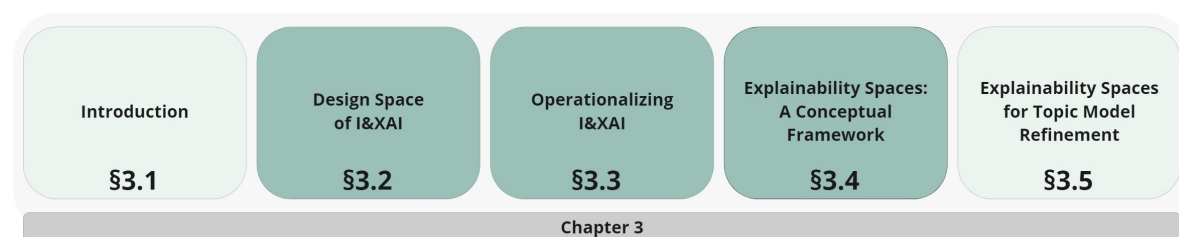
This chapter frames the entire dissertation by introducing all components and aspects influencing human-AI interaction, in particular for explaining the inner workings of machine learning models. Besides the introduction and discussion, this chapter comprises three main content sections. First, we span the design space of interactive and explainable machine learning (I&XAI) by introducing eight design dimensions. This is followed by a model for operationalizing I&XAI. Lastly, a novel conceptual framework for Human-AI interaction is introduced, consisting of three **Explainability Spaces** for interaction and explainability. This framework builds the foundation of all the following content chapters of the dissertation.

Contributions

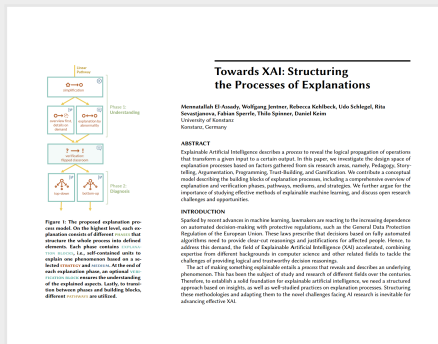
- (i) A comprehensive description of the *design space* of interactive and explainable artificial intelligence (I&XAI) in Section 3.2.
- (ii) The *explAIner* pipeline for operationalizing I&XAI applications in Section 3.3.
- (iii) A novel conceptual framework, called *Explainability Spaces*, for the design of Human-AI interfaces for I&XAI in Section 3.4.

Contents

3.1	Introduction	46
3.2	The Design Space of Interactive and Explainable Artificial Intelligence	47
3.2.1	Users and Stakeholders	48
3.2.2	Analytical Tasks and Application Domains	49
3.2.3	Data Characteristics	51
3.2.4	Types and Classes of Machine Learning Models	52
3.2.5	Stages in the KDD Process	52
3.2.6	Explainability Goals and Values	53
3.2.7	Explainability Tasks and Processes	55
3.2.8	Methods, Mediums, and Strategies	55
3.3	Operationalizing Interactive and Explainable Machine Learning	57
3.3.1	The <i>explAI</i> ner Pipeline	57
3.3.2	Single- and Multi-Model <i>explAI</i> ners	60
3.3.3	Global Monitoring and Steering Mechanisms	61
3.4	Explainability Spaces: A Conceptual Framework	62
3.4.1	Semantic Space	66
3.4.2	Input-Output Space	68
3.4.3	Model Space	70
3.5	Explainability Spaces for Topic Model Refinement	73



This chapter is partly based on the following publications:

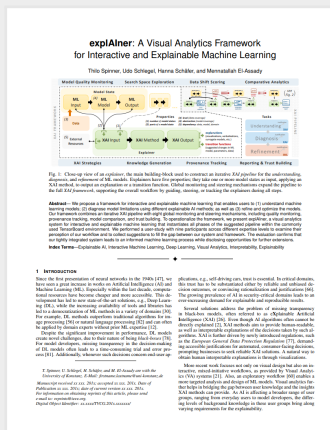


Towards Explainable Artificial Intelligence: Structuring the Processes of Explanations

Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, Daniel Keim

ACM CHI 2019 Workshop: Human-Centered Machine Learning Perspectives, 2019.

El-Assady et al. [El+19a]



explainer: A Visual Analytics Framework for Interactive and Explainable Machine Learning

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady

IEEE Transactions on Visualization and Computer Graphics, 26(1), 2020.

Spinner et al. [Spi+20]

Contribution Clarification

Most of the content of this chapter is a novel contribution, specifically developed for this dissertation. The only two parts that were previously published are the *Building Blocks* framework [El+19a], described in Section 3.2.8; and the *explainer* pipeline [Spi+20], described in Section 3.3. The former publication results from a collaborative group effort of the co-authors, led by me. The latter publication has two distinct contributions, a conceptual framework, and a system implementation. The conceptual framework is devised and written by me, while the system implementation is devised and written by the first two co-authors. Only the conceptual framework is part of this dissertation.

3.1 | Introduction

To push our problem-solving and decision-making capabilities beyond the bounds of constrained tasks and well-specified problems, we need to team up human and artificial intelligence. This requires designing effective human-AI interfaces that mitigate between the two actors, enabling collaborative decision-making and symbiotic problem-solving. Tailoring the interaction design of such *mixed-initiative* systems to different users, tasks, and data has been the subject of various research approaches; in particular, in the emerging research fields of interactive machine learning (IML) [FJ03; End+17] and explainable artificial intelligence (XAI) [AB18; Arr+20; MZR20]. Together, they form the research area of interactive and explainable AI/ML (henceforth, I&XAI) [Spi+20; Liu+17b].

Beside problem-solving and decision-making, human-AI interfaces can be used for explainability (i.e., understanding, diagnosing, and refining AI models) [Liu+17b]; guidance (i.e., closing a knowledge gap) [Cen+18]; interaction dynamics, such as, learning and teaching [Spe+20b]; co-adaptive analysis [Spe+21b]; and trust-building [Spe+20a].

At their core, all these research directions deal with the process of knowledge discovery in data (KDD) [FPS96]. In particular, how data is transformed from a given input to an output using a machine learning model. This transformation usually consists of a complex, multi-step process. Understanding, diagnosing, and refining this process allows for humans to investigate data patterns, connect this information to their understanding of the world, and finally, externalize their newly generated knowledge to refine the AI models [Sac+14]. Beyond changing a model's data, parameters, or architecture, models can adapt using active, reinforcement, or imitation learning [Hus+17]. Additionally, humans can point out missing relations, performing machine teaching [Sim+17].



There are different paradigms for interacting between humans and AI-based systems. Amershi et al. [Ame+19] have investigated the differences between pairing only human or AI actors among themselves to pairing a human with an AI. They propose guidelines for human-AI interaction, focusing on empowering the human stakeholders [Ame+14]. In this dissertation, we ask ourselves; *how to effectively team-up humans and AIs for I&XAI?* The answer to this question is partly covered by investigating the algorithm-centered desiderata for interpretations [Mur+19]. However, such requirements are missing a human-centered perspective [ER20]; hence, in Section 2.1.3, we proposed eight *explainability considerations* for effective human-AI interaction.

Next, we ask ourselves; *what is the design space of I&XAI?* In particular, what are the aspects influencing I&XAI techniques? These aspects can be specific to the application domain, the machine learning model, or the explainability. We, thus, build on existing research to derive eight comprehensive design dimensions, presented in Section 3.2. In addition to the general

design space, we present the *explAIner* pipeline, in Section 3.3, to answer the question of *how to achieve the goals of explainable AI?*

The derived design space forms the basis for our conceptual framework, introduced in Section 3.4. Based on the need for human-centered design rationales, we consider the various stakeholders, their backgrounds, expertise, and mental models; asking ourselves; *how should different stakeholders interact with an AI?* To answer this question, we propose *Explainability Spaces*, a novel interaction and explainability design space [WY04], forming a continuous spectrum of possible designs. We exemplify possible design considerations based on three spaces; the **Semantic Space** (targeting *Domain Experts*), the **Input-Output Space** (targeting *Domain and Model Analysts*), and the **Model Space** (targeting *Model Experts*). Lastly, we discuss how different interface, interaction, and explainability designs can empower these user groups in the use case of topic modeling refinement. Section 3.5 describes the application design spaces of the four main content chapters of the dissertation.

3.2 | The Design Space of Interactive and Explainable Artificial Intelligence

This section introduces the design space of interactive and explainable artificial intelligence (I&XAI), spanning eight dimensions. These dimensions describe specific aspects regarding the application domain, the machine learning model, and the explanation method, respectively. The following paragraphs give an overview of these aspects, followed by detailed descriptions of each individual dimension in the subsections.

First, we introduce **Domain-Specific Aspects**, describing the target audience of an I&XAI application, their specific tasks, and their data. The first three dimensions of the design space capture aspects related to the application domain, thus enabling us to compare applications and reason about their similarities and differences.

- D1**  **Users and Stakeholders** (Section 3.2.1)
Who are the stakeholders and user groups involved?
(Technical vs. Domain Expertise, Mental Models, etc.)
- D2**  **Analytical Tasks and Application Domains** (Section 3.2.2)
Where should the explainability be applied?
(Domains, Research Questions, Applications, etc.)
- D3**  **Data Characteristics** (Section 3.2.3)
What are the characteristics of the input (and output) data?
(Static, Streaming, Homogeneous, Scalable, etc.)

Second, we introduce **Model-Specific Aspects**, describing types of machine learning models and different stages in the data processing pipeline. The two model-specific dimensions enable a targeted discussion of interaction and explainability strategies for varying ML models and development stages.

- D4** ✨ **Types and Classes of ML Models** (Section 3.2.4)
Which types and classes of ML models can be explained?
(Supervised, Unsupervised, Interactive, etc.)
- D5** ⚙️ **Stages in the KDD Process** (Section 3.2.5)
When in the data processing and ML pipeline is explainability useful?
(Training, Testing, Parameter Exploration, etc.)

Lastly, we introduce **Explainability Aspects**, describing the different goals and tasks for I&XAI, as well as methods, strategies, and mediums to achieve these goals. The last three dimensions specify aspects of explainability.

- D6** 🎯 **Explainability Goals and Values** (Section 3.2.6)
Why are we seeking explainability and interpretability?
(Accountability, Responsibility, Trustworthiness, etc.)
- D7** ➡️ **Explainability Methods and Processes** (Section 3.2.7)
How do we achieve the goals of explainable AI? – Part I
(Understanding, Diagnosis, Refinement, etc.)
- D8** 📄 **Explainability Methods, Strategies, and Mediums** (Section 3.2.8)
How do we achieve the goals of explainable AI? – Part II
(Visualization, Verbalization, Gamification, etc.)

In the following, each of the eight dimensions is described in more detail.

3.2.1 | Users and Stakeholders

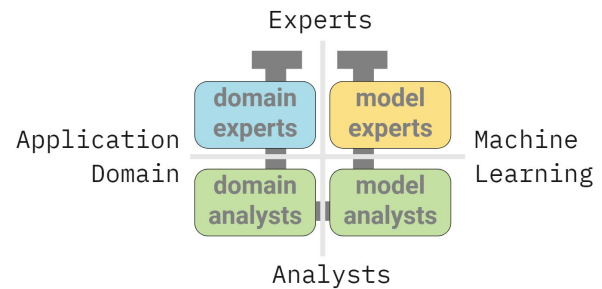
There are various categorizations of user groups related to I&XAI applications. Most categorizations differentiate between model developers and end users [Hoh+18].

👤 **Who are the stakeholders and user groups involved?**

Some categorizations focus on the role of users in the applications, e.g., distinguishing between data scientists, decision-makers, and subjects [CW20]. Others focus on the expertise of the users, e.g., machine learning novices, operators, analysts, or experts [Spi+20].

By combining the arguments of the different categorizations, we can derive two important aspects when distinguishing user groups. First, the **background** of the users, i.e., whether their

qualification is in machine learning or the application domain. Second, the **expertise** of the users, i.e., whether they are experts (builders & developers) or analysts (operators & novices). Taken together, we can derive the following four user groups:



- **Domain Experts:** This user group describes stakeholders who are experts in the application domain at hand and can thus contribute to general structural knowledge and high-level research questions. These experts might be interested in solving a concrete problem, but more often, they want to derive a broad understanding of multiple phenomena, e.g., the causal relation between semantic concepts. Typical representatives of this group are senior researchers in the various application domains.
- **Domain Analysts:** This user group describes stakeholders who are analysts of the application domain. These are operators interested in solving one concrete application problem based on a specified set of data and tasks. Typical representatives of this group are practitioners and industrial data scientists. They have a very detailed understanding of their specific problem, but are typically not interested in deriving broad and abstract conclusions for the research domain at hand.
- **Model Analysts:** Similar to the domain analysts, model analysts are interested in solving concrete problems. Rather than being operators and practitioners with a domain background, they have a more substantial machine learning background. However, both of these roles are in practice very similar and thus target a related set of people.
- **Model Experts:** This user group describes experts in machine learning, interested in gathering knowledge about the broader application of machine learning techniques, to contribute to developing and building new models. These experts solve multiple concrete problems in different domains to enhance their understanding of model behaviors. Typical representatives of this group are senior researchers in machine learning.

Based on assumptions about their different *mental models* [MZR20], these user groups describe general roles that different stakeholders can take. The roles can overlap in different applications, and the same person can perform multiple roles. Additionally, one person can change their role, for example, by gaining expertise. We did not include institutions, collectives, or society groups as stakeholders in this dimension [Ehs+21], as they are not performing direct AI-interactions.

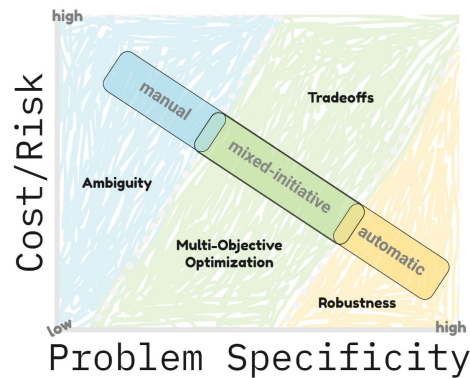
3.2.2 | Analytical Tasks and Application Domains

Interactive and explainable machine learning can be applied in many different application domains, particularly, in situations where human-involvement is needed. Application areas from biomedical research to social science employ machine learning models for data-driven problem-

☑= Where should the explainability be applied?

solving and decision-making. In the context of human-centered machine learning, tasks range from diagnosing, comparing, and refining ML models; through exploring and understanding their behavior; to using their results to verify hypothesis and justify decision-making [Spe+21a]. For every domain, specific tasks are of interest, for example, investigating model architectures for deep learning [Gar+18], or analyzing semantic models in text mining [Liu+18c].

While some of these tasks can be completely automated, others have to rely on a human in the loop. Interactive and explainable machine learning is especially useful for problems that lay in the center field of the spectrum of human-machine intelligence. This spectrum can be, among other things, effectively spanned by two factors related to the task at hand. First, the **specificity** of the problem, i.e., can precise metrics be defined to capture the target objective [SMM12]. Second, the involved **cost and risk** of the problem, i.e., how much is at stake if the solution was incorrect. Based on these two criteria, analytical tasks are performed in one of three ways:



- **Manual:** For ambiguous tasks with a low problem specificity [Bec+20], it is infeasible to look for automated solutions. These tasks are usually performed manually, as each individual solution is dependent on non-measurable success criteria. In addition to tasks too irregular for automation, some manual tasks are not automated due to them being too infrequent or not laborious.
- **Mixed-Initiative:** A wide range of tasks is either under-specified, lacks metrics to optimize for, or has no clear favored solution due to multiple optimization objectives. In these circumstances, the parts of the tasks that can be solved accurately enough through automation are performed automatically by the system. However, some of the sub-tasks need human involvement to steer the models, provide missing information, decide based on general knowledge or domain expertise, etc. To engage in these trade-offs, interaction and explainability are essential communication mechanisms between humans and ML models. These mixed-initiative tasks can have equal involvement from a human agent and an ML agent, but can range from automatic tasks with a human-in-loop to manual ones with a machine-in-the-loop, so-called, “*AI-infused systems*” [Ame+19].
- **Automatic:** When problems are clearly specified, they can be automated, even in high-risk situations. To train accurate and robust models, the tasks have to be specific and supported by precise metrics, as well as a large and diverse data sample.

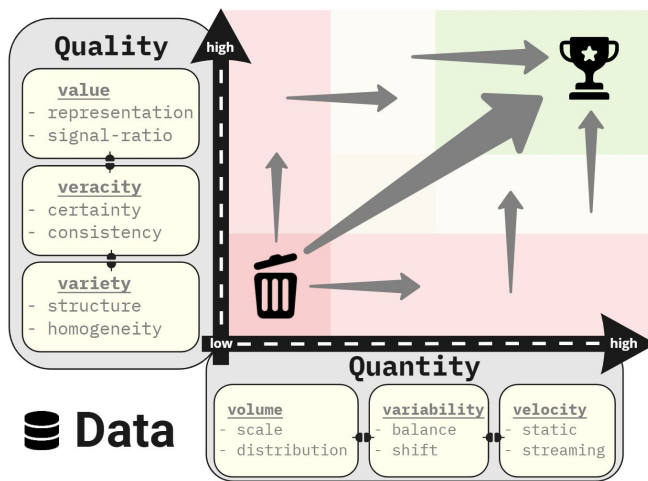
While explainability is undoubtedly helpful for pure manual or automatic tasks to communicate and justify decision-making processes, it becomes inevitable for mixed-initiative tasks, as both the human and the ML model need to explain their decisions to enable effective interactions. A discussion of the goals of explainable and interactive ML is presented in Section 3.2.6.

3.2.3 | Data Characteristics

Data-driven problem-solving and decision-making rely on the quantity and quality of the data to produce meaningful outputs. Data refers to a collection of raw facts, observations, or statistics. These are inherently meaningless until processed and analyzed. We can put data in context through a task-driven analysis, thus recognizing patterns, producing information, and ultimately generating knowledge [Sac+ 14].

 **What** are the characteristics of the input (and output) data?

The types of data processed (e.g., textual, geospatial, temporal, etc.) depend on the task at



hand. However, there are several vital characteristics beyond data type to consider. These can be grouped into characteristics related to **data quality** and others related to **data quantity**. Overall, an interactive and explainable machine learning approach's success depends on the underlying data characteristics. Hence, we strive for high-quality data with a balanced distribution and quantity. In the following, these two characteristics are discussed in more detail based on three aspects each. These aspects are based on dimensions outlined by Khan et al. [KUG14] and have no inherent order.

- **Data Quality:** In a data-driven analysis workflow, a substantial amount of time is spent on data cleaning, preprocessing, and structuring. High-quality data is essential for all upstream tasks. Hence, the data should contain enough discriminative signals to learn the desired task, i.e., be relevant, precise, and consistent. More specifically, three aspects define data quality; (1) **Variety** referring to the structure and homogeneity of the data; (2) **Veracity** referring to the certainty and consistency of the data; and (3) **Value** referring to the representation of the data and its signal-ratio. To train robust and reliable ML models, we need to check the relevance and quality of the data for a given task, as false or low-quality data is worse than no data at all.
- **Data Quantity:** Training ML models requires enough data samples with sufficient variety across phenomena. Besides the amount of data required for learning, we need to balance the data to avoid data shifts and out-of-distribution problems [HG16]. More specifically, three aspects define data quantity; (1) **Volume** referring to the scale and distribution of the data; (2) **Variability** referring to the balance of the data and possible data shifts; and (3) **Velocity** referring to the temporal dynamics of the data, i.e., static vs. streaming data. These issues enable us to define suitable ML models, for example, using deep learning

models only when we have sufficiently large datasets. In addition, it allows us to judge the scalability and transferability of the trained models.

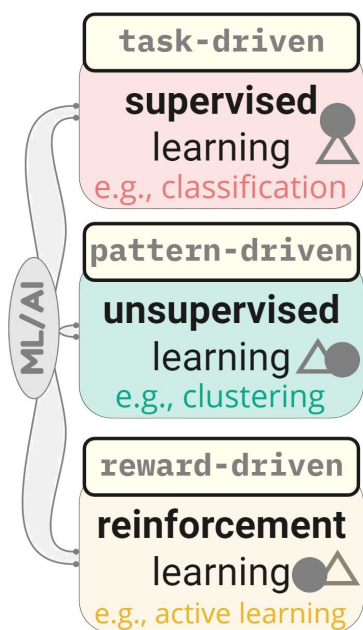
The quantity and quality of data directly affect the explainability of ML models, as some of the most common approaches for XAI are using data samples as examples of classes or as counterfactuals. These strategies are discussed further in Section 3.2.8.

3.2.4 | Types and Classes of Machine Learning Models

Machine learning models learn based on different criteria [CMM83]. In addition, they can be categorized based on their learning into symbolic, sub-symbolic, and hybrid models [CCO20]. However, for this dimension, we group them based on their method into three classes:

- **Supervised:** “Supervised learning algorithms experience a dataset containing features, but each example is also associated with a label or target” [GBC16]. I&XAI techniques are typically designed to explain the data labeling and discriminative features.
- **Unsupervised:** “Unsupervised learning algorithms experience a dataset containing many features, then learn useful properties of the structure of this dataset” [GBC16]. I&XAI techniques are typically designed to explain the data distribution and similarity, and discriminative features.
- **Reinforcement:** “In Reinforcement learning, an autonomous agent must learn to perform a task by trial and error, without any guidance from the human operator” [GBC16]. I&XAI techniques are typically designed to explain the environment, learning policy, reward function, and discriminative features.

✦ Which types and classes of ML models can be explained?



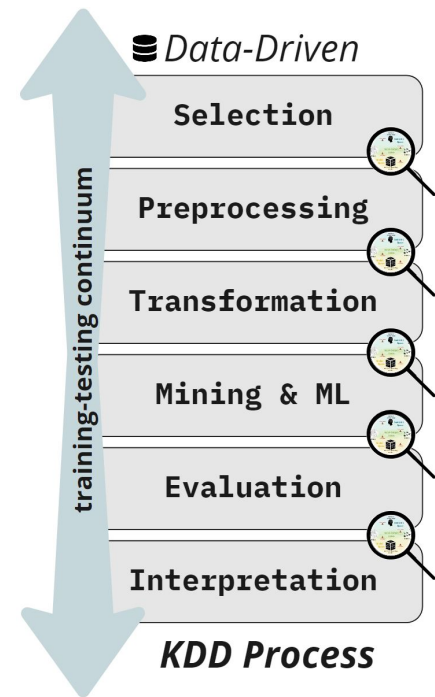
3.2.5 | Stages in the KDD Process

The knowledge discovery in databases (KDD) [FPS96] process describes the *training–testing continuum* of machine learning. Within this process, explanations can help in understanding, diagnosing, and refining the different processing steps. Explanations can emerge directly from the processing step, e.g., if it is intelligible (self-explaining), otherwise, steps can be explained by an additional method, which requires post-processing (post-hoc explainability) [Arr+20]. More details on explainability methods are discussed in Section 3.3.

⚙ When in the KDD process is explainability useful?

In the following, we describe the usefulness of explanations on the *training–testing continuum*:

- **Selection:** During data selection, its underlying distribution needs to be explained. This allows for representative training and testing data generation, as well as avoiding data-shift pitfalls [SKE20].
- **Preprocessing:** During preprocessing, different steps for data cleaning, normalization, and missing value treatment are performed. Here, explanations help avoid the introduction of biases, ambiguities, or mistakes during this process [Bec+20].
- **Transformation:** Data transformations are often-times necessary to use machine learning models. However, these transformations can skew the data distributions or use non-bijective mappings. To understand the effects of such transformations, I&XAI techniques are crucial, e.g., for investigating noise robustness [Liu+18a].
- **Mining & ML:** The most studied step in the KDD process with respect to I&XAI techniques is the machine learning step. Here, explanations can help understand, diagnose, and refine models [Liu+17b].
- **Evaluation:** Evaluations allow for the examination of the complex and dynamic systems in which models operate, ensuring that they fulfill their goals. Especially, explanations in human-centered evaluations of I&XAI are becoming imperative [Spe+21a].
- **Interpretation:** Explanations can support users in probing models to calibrate their beliefs about the model’s behavior, outputs, or patterns; avoiding harmful rationalization [Ehs+19].



In addition to interactions and explanations at the different stages of the *training–testing continuum* [MAA21], three types of explanations can be distinguished; **a priori explanations**, i.e., before training and testing, inspecting data patterns and characteristics to determine what the ML models could learn; **ad-hoc explanations**, i.e., during training and testing, inspecting what the models are learning; and **post-hoc explanations**, investigating the model’s decision-making.

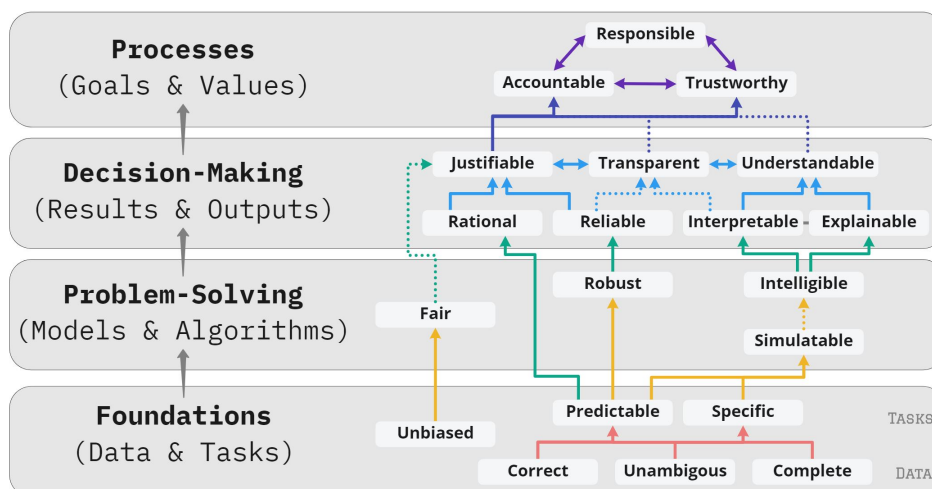
3.2.6 | Explainability Goals and Values

According to *Desiderata for Interpretations* proposed by Murdoch et al. [Mur+19], there are various aims for pursuing I&XAI. However, there is no clear consensus on the goals and values of I&XAI [Arr+20]; some of the most prominent ones are comprehensibility [MSM18; Fer+19], interpretability [MCM83; Gui+18], explainability [Gle16; Cra96],

🔍 *Why are we seeking explainability and interpretability?*

and transparency [Lip18]. We structure the goals and values according to their associated rank from foundations through problem-solving and decision-making to high-level processes, as described in the following:

- **Foundations (Data & Tasks):** On the lowest rank, we have to ensure that the data is *correct*, *unambiguous* [Bec+20], and *complete*. This allows us to define tasks that are *predictable* [WB19] and *specific* [Jon72]. In addition, we need both the data and tasks to be *unbiased* [San+14].
- **Problem-Solving (Models & Algorithms):** On the next rank, we build upon the foundations to start solving concrete problems. To design *robust* [RD18] models, we rely on the predictability of the tasks. Similarly, for models to be *intelligible* [Abd+18; BE01; WB19] they have to be *simulatable* [RSG18; RSG16], and in turn, the task they address has to be predictable and specific. Lastly, to guarantee *fair* [Woo+18; Lep+18] models, we have to build on unbiased tasks and data.
- **Decision-Making (Results & Outputs):** The third rank describes the decision-making in form of results and outputs. To achieve *justifiable* [Mur+19] decisions, they have to be *rational* [Gui+18], and in turn, their underlying tasks have to be predictable. To get to *transparent* [CH19; Lip18] outcomes, we have to build on top of *interpretable* [Mol18; Arr+20; Lip18] and *reliable* [Hof+13; Hof+18] decisions, these, in turn, depend on intelligible and robust models, respectively. Lastly, to achieve *understandable* [Arr+20; Lim11; MSM18] outputs, we rely on interpretable and *explainable* [Wan+19a; Arr+20; Lip18; Gui+18] decisions, which require intelligible models.
- **Processes (Goals & Values):** The last rank describes the overall processes in the form of high-level goals and values. The first goal is for the processes to be *accountable* [Dos+17; Lep+18] which requires justifiable outcomes. The second goal is for the processes to be *trustworthy* [Hof+13; Hof+18] which also requires justifiable decisions. Lastly, the processes can become *responsible* [Mur+19] when all the other conditions are fulfilled.



These goals and values have ambiguous definitions throughout the literature. In addition, some of their relations have to still be established based on future research. Hence, this dimension is application- and task-dependent.

3.2.7 | Explainability Tasks and Processes

Interactive and explainable AI applications tackle challenges, ranging from understanding the input data distributions and feature compositions of a machine learning model; through optimizing the model's learning process and comparing its results to other models; and finally, to deploying a refined model on various data collections. In particular, human-AI interaction techniques have been proposed to achieve the goals of explainable and interactive machine learning. As introduced in Section 2.1.1, according to Liu et al. [Liu+17b] and Spinner et al. [Spi+20], these can be broadly defined to consist of three tasks:

- **Understanding:** Investigating why a model behaves the way it does and how different models compare to understand and justify the model decisions and behavior.
- **Diagnosis:** Inspecting why the model fails to converge or does not achieve an acceptable performance to diagnose and justify the model's performance and applicability.
- **Refinement:** Optimizing and improving the performance and robustness of a model through refining the model to the given users, tasks, and data.

➡ *How do we achieve the goals of explainable AI?*



Liu et al. [Liu+17b] conclude that a promising research direction for interacting with machine learning models and explaining their behavior is through tightly integrating them into visual analytics workflows. They further highlight four promising research opportunities for visual analytics in this context; namely, creating explainable models; analyzing online training processes; mixed-initiative guidance; and uncertainty analysis. These three steps form the process of I&XAI, and in Section 3.2.8, they are used as phases in an explanation strategy. Additionally, they are operationalized in the *explAIner* pipeline as tasks, as described in Section 3.3.

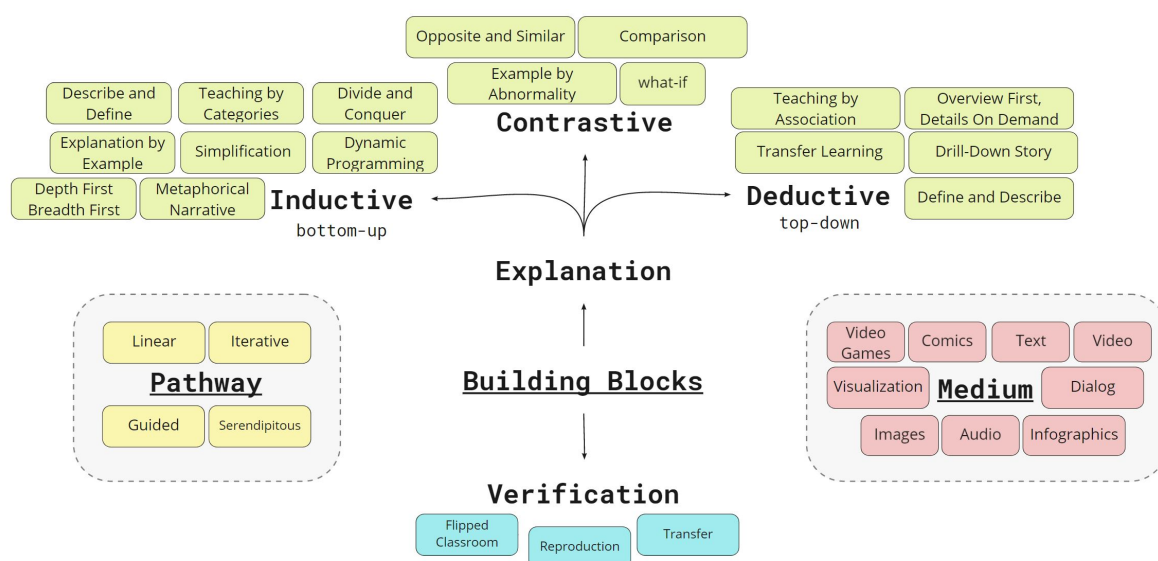
3.2.8 | Methods, Mediums, and Strategies

Based on the guidelines for human-AI interaction [Ame+19], we propose considering global I&XAI strategies based on the *Building Blocks* framework [El+19a].

These strategies allow for the construction of a multistep explanation process consisting of different **phases**, such as understanding, diagnosis, and refinement. Within these phases, different **building blocks** can be used for explanation and verification. These building blocks are connected using iterative vs. linear **pathways**, defining guided vs. serendipitous connections.

📖 *How do we achieve the goals of explainable AI?*

Diverse **mediums** allow for a variety of explanation expressions, however, the most commonly used are visualizations or verbalizations [Sev+18b].



The act of making something explainable entails a *process* that reveals and describes an underlying phenomenon. Deciding on an effective strategy for the user group at hand is essential. This also involves configuring the amount of user guidance needed, which, in turn, can be informed through tracking the user behavior and through monitoring the model quality over time. *explAIners* (Section 3.3.2) can be regarded as building blocks that use a certain explanation strategy and medium to explain an aspect. In general, we can distinguish between three *explainability paradigms* [Ron14; SS16], as follows:

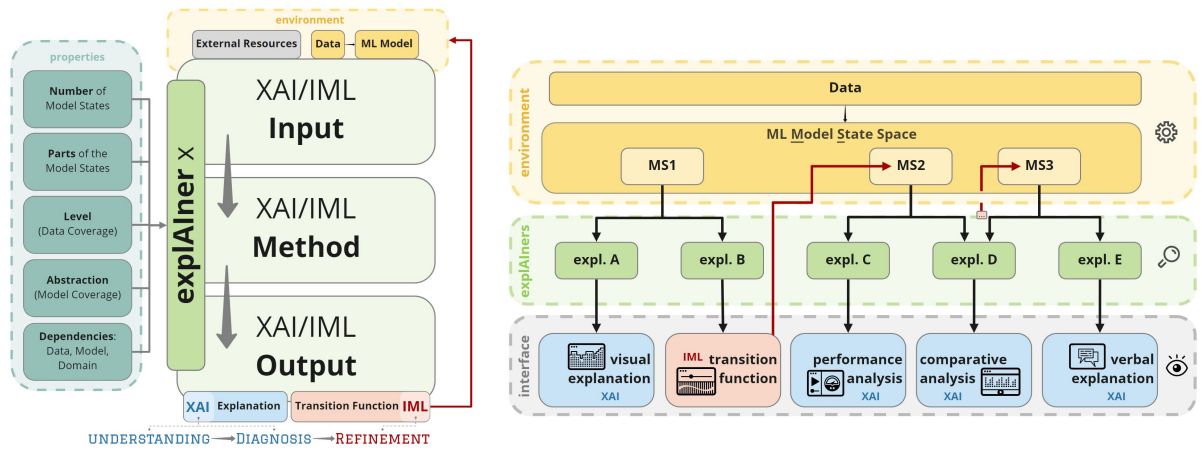
- **Inductive:** Bottom-up or inductive explanations are data-driven and start with specific samples or patterns. They are the closest explanations to the typical learning paradigm of AI models. Inductive explanations include, *explanations by example*, *simplifications*, or *divide-and-conquer* strategies.
- **Contrastive:** *Counterfactuals*, *comparisons*, and *probing-based explanations* are typical contrastive explanation strategies. They rely on opposing phenomena to distinguish similarities and differences.
- **Deductive:** Top-down or deductive explanations are theory-driven and start with high-level overviews of relations before exposing the details. They are the closest explanations to human procedural knowledge and thus lend themselves to domain knowledge externalization. Examples include *teaching by association* and *drill-down-stories*.

3.3 | Operationalizing Interactive and Explainable Machine Learning

Building on the general design space proposed in Section 3.2, this section formalizes the **Explainability Aspects** into a tailored conceptual model that *operationalizes* interactive and explainable machine learning processes. First, to address the explainability tasks (Section 3.2.7), we introduce the *explAIner* pipeline in Section 3.3.1 and two categories of *explAIners* in Section 3.3.2. Second, to support the strategies (Section 3.2.8) in orchestrating the components of the pipeline, we present global monitoring and steering mechanisms in Section 3.3.3.

3.3.1 | The *explAIner* Pipeline

To structure the space of interactive and explainable machine learning, we propose a modular pipeline that captures the iterative process of model understanding, diagnosis, and refinement [Liu+17b]. The *explAIner pipeline* provides explanations for a given ML model state or suggests a transition from one model state to another. The main modules of the pipelines are the, so-called, *explAIners*. These *explAIner* modules interact with the machine learning model to derive (1) *explanations* in the form of visualizations, verbalizations, or surrogate models, as well as (2) *transition functions* for model refinement. Global monitoring and steering mechanisms support the overall workflow through guiding, steering, or tracking the *explAIners* and can be accessed and applied during all steps, as described in Section 3.3.3.



- (a) *explAIners* are modular wrappers for I&XAI methods. They take data, ML models, and external resources as inputs to output explanations or transition functions.
- (b) The unrolled *explAIner pipeline* shows the iterative process of understanding, diagnosis, and refinement. In the depicted example, different *explAIners* are applied to model states, resulting in various explanations or transitions to new model states.

Fig. 3.1: The *explAIner pipeline* consists of a number of *explAIner* modules that operate in a given environment to support the process of model understanding, diagnosis, and refinement.

Explainable and interactive ML is a *mixed-initiative process* where a machine learning model is extended with an *I&XAI* method. As depicted in Figure 3.1, this can be modeled in the proposed

explAIner pipeline using *explAIners* as wrapper modules for the I&XAI method. Figure 3.1a shows a single *explAIner* including its inputs, outputs, and properties.

Combining a series of *explAIners* forms a pipeline, as depicted in Figure 3.1b. This example shows a pipeline of five *explAIners* and three ML model states. The process starts with *explAIner A* that takes model state 1 as an input. It produces a visual explanation that might be used for model understanding or diagnosis. On the other hand, *explAIner B* is used for refinement as it suggests a transition function based on the same model state. The pipeline continues with *explAIner C*, which utilized the newly created models state 2 to run a performance analysis. *explAIner D* is the first multi-model *explAIner*. It analyzes the model states 2 and 3 to perform a comparative analysis. Such a comparative analysis can enable a model selection task or inform a new model state. Lastly, *explAIner E* generates a verbal explanation based on the third model state. This could, for example, be a provenance report.

In the following subsections, the inputs, outputs, properties, and methods of *explAIners* will be described in more detail.

***explAIner* Inputs and Outputs**

The input of an *explAIner* comes from the environment in which it operates. As depicted in Figure 3.1a, this environment consists of data (for training and testing), the ML model (sub-structured into different model states), as well as external resources (i.e., data that is not part of the training and testing of the ML model but that is used for explanation). Based on this environment, different *explAIners* can be designed, for example, ones operating only on the data level, so-called, *model-agnostic explAIners*; or ones considering multiple model states as input for comparison, so-called, *multi-model explAIners*.

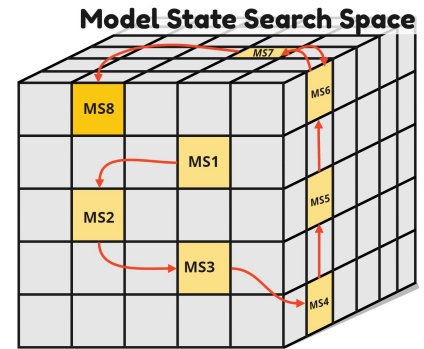
The output of *explAIners* are either (1) **explanations** – (interactive) visualizations, verbalizations, model surrogates, etc; or (2) **transition functions** to a new model state. Hence to achieve the goals of **explainable (XAI)** and **interactive (IML)** machine learning, **explanations** are used to **understand** and **diagnose (XAI)** a model, and **transition functions** are used for model **refinement (IML)**.

***explAIner* Methods and Properties**

As described in Chapter 2, there is an enormous variety of methods for explainable and interactive machine learning. The *explAIner* pipeline is a descriptive model to make existing I&XAI methods comparable. It proposes the use of *explAIners* as wrapper objects to modularize explainable and interactive machine learning processes. Hence, the proposed structure can represent any method that operates in the context of an ML environment to produce expla-

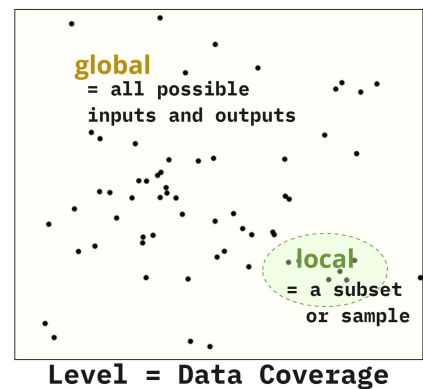
nations or transition functions. This approach then enables us to compare I&XAI methods by examining their properties. These consist of the following main properties:

[P1] Number of Model States As depicted in Figure 3.1b, our proposed workflow constitutes multiple *model states* and multiple *explainers*. A model state is one configuration of a trained ML model with a given set of parameters, weights, etc. Changing such parameters or weights transitions the model to a different state. Thus, to find an ‘optimal’ model for a given data and task, we consider the search space that spans all possible model states, i.e., all different configurations of a model. Note, that changing the model architecture would result in another model with its own search space, while retraining the same model only transitions the model state. Using this notion, we can argue that the goal of interactive machine learning is to enable model refinement such that we transition to model states that are more suitable for a given problem characteristic. More formally, we can define all operations that change a property of a model state as a transition function $f : MS_x \rightarrow MS_y$, with MS_x and MS_y describing two model states (could be equivalent) within the search space. Hence, a refinement process can be seen as a traversal through multiple model states. Our proposed *explainers* are components that take into account inputs from a single model state (referred to as **single-model explainers**) or from multiple model states (referred to as **multi-model explainers**). More details about this categorization are discussed in Section 3.3.2.

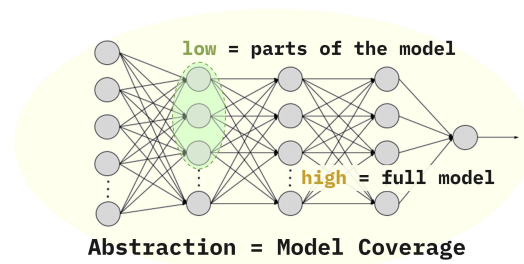


[P2] Parts of the Model States Besides the number of model states considered, we categorize *explainers* based on the parts of the model state considered as input (ML input, ML model, and/or ML output). As described in Section 3.3.2, *explainers* can either be used for understanding the ML input, model, or output by themselves; or used to put these parts in relation, i.e., considering the whole pipeline (referred to as **model-ware explainers**) or only the data level (referred to as **model-agnostic explainers**).

[P3] Level (Data Coverage) *explainers* typically use data objects to showcase the model behavior. Hence, we define the *explainer* level as the data coverage of the *explainer*. It is *global*, when all possible data inputs and outputs of an ML model are considered for the explanation. On the other hand, a *local* level refers to the *explainer* only considering a subset or sample of the data (e.g., for explaining decision boundaries). Note, that *explainers* can use a disjunct data set from the one used for training or testing the model.



[P4] Abstraction (Model Coverage) In addition to the data coverage, we consider the coverage of the model in the explanation. Thus, we define the *explAIner* abstraction as the model coverage. A *low* abstraction only considers a part of the model, while a *high* abstraction considers the whole model. This is particularly useful for debugging model architectures, for example, for deep neural networks.

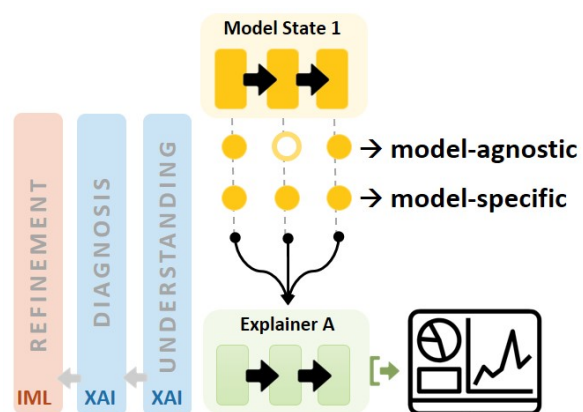


[P5] Dependencies: Data, Model, Domain Lastly, each *explAIner* can have dependencies on the data, model, and/or domain knowledge, resulting in explanations dependent on some of these factors. For example, an ML model might use longitude and latitude data for a prediction. An *explAIner* could add the country associated with each longitude and latitude pair as a piece of additional information to contextualize the provided explanation. Note that adding external resources can lead to biases in the explanation.

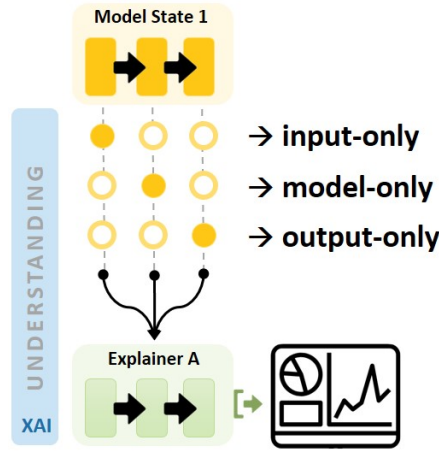
3.3.2 | Single- and Multi-Model *explAIners*

In the following, we discuss two categories of *explAIners*, i.e., single- and multi-model *explAIners*. The two described categories are the ones commonly used, but other categorizations might also be applicable, e.g., *explAIners* that consider a different combination of inputs, including *model-external resources*.

Single-Model *explAIners* The most straightforward type of *explAIners* are the ones that consider the input and output of a machine learning model, as well as its inner workings. These **model-specific** *explAIners* are particularly useful for model developers as they can help in diagnosing the internal structure of a model and refining it based on the interplay of the inputs and outputs, with respect to the given architecture. Examples include: [Bac+15; Str+18b]. In contrast, **model-agnostic** *explAIners* operate on the data level. They consider the model to be a black box that performs a transition from input to output, and, thus, attempt to explain or approximate this transition. These *explAIners* are particularly useful for non-experts and model users in machine learning who are not interested in understanding the underlying model but rather in applying it to their data and tasks. Examples of such *explAIners* include: [RSG16; RSG18].



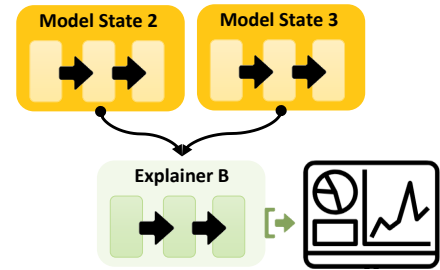
Our proposed *explAIner* pipeline can be used in different stages of the *training–testing continuum* (of the KDD pipeline, see Section 3.2.5) of a model by stopping or extracting a state from the current training process. However, for different user groups, some states and *explAIners*



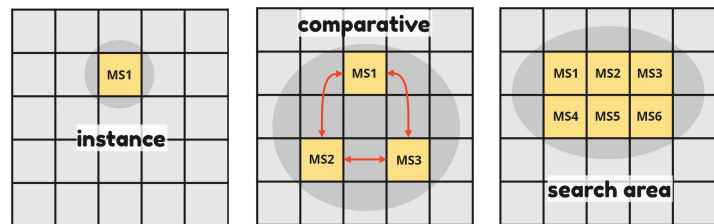
might be more favorable. In addition to the two types mentioned above, we particularly would like to highlight *explAIners* used for understanding the ML input, model, and output. Understanding the model without considering the input and output of the data is a useful task for inspecting the model architecture and weights, as well as, for educational purposes. Similarly, inspecting the input or output data distribution and characteristics is a common task. Hence, these types of *explAIners* are particularly useful for understanding but not as much for diagnosis or refinement. One example of such a **model-only**

explAIner are *look-up explAIner* showing wiki-style entries to enable the understanding of specialized nodes in neural networks. Examples include: [Won+18; Har15].

Multi-Model *explAIners* Complementary to single-model *explAIners*, multi-model ones are primarily used for the *comparative analysis* of model states. These take as an input two (or more) model states to compare and their output is tailored to such a task, i.e., comparative visualizations, model selection components, or a transition function based on all input states. Examples include: [Mur+18; Zha+19].



Multi-model *explAIners* help in exploring the model state search space beyond single instances or points in the space. The comparative analysis, thus, enables a **point-wise inspection** of similarities and differences between a selection of different model states. In addition, we can subdivide the search space by defining a **search area** in which model states are compared.



3.3.3 | Global Monitoring and Steering Mechanisms

To guide, steer, and track the *explAIner* pipeline, we propose a number of global mechanisms, as observed in the related work. These mechanisms ensure that users are supported in their goals during all phases of the pipeline. In the following, we describe the most important ones:

- **Model Quality Monitoring:** Internal **performance metrics**, such as accuracy, precision, and recall [Pow08], as well as measures of *bias* [DK95] and *uncertainty* [SS09] can support the *explAIner* pipeline in pointing the users to potential model improvement possibilities. We propose including a global monitoring component to constantly track and assess the internal quality of the different model states.
- **Data Shift Scoring:** In analogy with the continuous monitoring of the model quality, if the *explAIner* pipeline is targeting the analysis of changing data sources, we propose the monitoring of potential *data shifts* [Qui+09] along the *training–testing continuum*. Especially for deciding when to stop the training and retraining [GBD92], a measurement of the data fitness to the model is of immense importance.
- **Search Space Exploration:** For a *targeted refinement and optimization*, efficient navigation of both the model input and output spaces is vital. Examples to achieve *minimum feedback for maximum gain* include, for example, *Speculative Execution* [Spe+18], where different potential optimizations are performed and presented to the user *before* applying them to the model.
- **Comparative Analytics:** Another important task is comparing and selecting models. *explAIners* can be designed to compare different model states on multiple levels. This, in turn, facilitates tasks, such as *model selection* [FJ03], automated *model recommendation* [MSG16], or automated *model configuration search (AutoML)* [BYC13; EMH19], i.e., transitions from one model state to another.
- **Provenance Tracking:** Model refinement and optimization is a “*garden of forking paths*” [GL13]. To track the *temporal evolution* of the user’s workflow, we propose a provenance tracking component. An *interaction tree* [Sti+16], for example, could reveal the sequence of decisions users undertook in the *explAIner* pipeline.
- **Reporting and Trust Building:** To enable a reasoned *justification* of the user’s decision-making, as well as allow for communicating the results of a workflow, we propose the implementation of reporting components. These can be used in *educational* settings, for example, by designing them as *Storytelling* [Che+18] components. Such mechanisms facilitate the calibration of trust between the users and the ML model [Jen+18].
- **Knowledge Generation:** Lastly, the ultimate goal of such a visual analytics framework is *knowledge generation* [Sac+14]. This can go in two directions, users can *learn* something about the machine learning model and process, or *validate* their knowledge; while models, can be *taught by users* [Sim+17], e.g., through *learning from their interactions* [Ame+19].

3.4 | Explainability Spaces: A Conceptual Framework

To empower (C6) the different user groups and stakeholders (described in Section 3.2.1) to interact with an AI or with AI-based systems, we need to design explanation strategies and interaction interfaces that are tailored to their mental models (C1). This will allow us to customize the explanations and interactions, avoiding rationalization (C2) and simplification

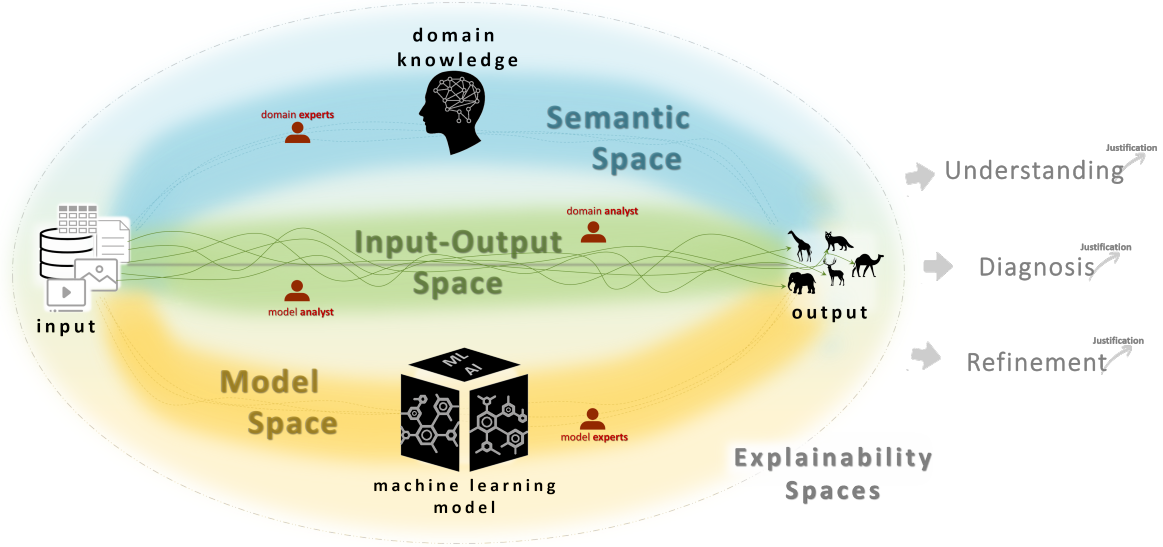


Fig. 3.2: The continuous spectrum of *Explainability Spaces* in machine learning, enabling the understanding, diagnosis, and refinement of different processes in the KDD pipeline. (1) The *Model Space* is tailored to expert model building. (2) The *Input-Output Space* enables concrete problem-solving. (3) The *Semantic Space* is designed for domain knowledge generation.

issues (C4). Furthermore, based on the expected tasks and mental models, we can decide on an appropriate level of abstraction (C3) to mold the used narratives and metaphors (C5). Hence, following the human-centered I&XAI goals (described in Section 2.1.2), we put the human first and ask ourselves: *How should different stakeholders interact with an AI?* In particular, we are interested in deriving a targeted design rationale to tailor the explainability process (D7) of *understanding*, *diagnosis*, and *refinement* to the main stakeholders (D1), i.e., *domain experts*, *domain analysts*, *model analysts*, and *model experts*. Thus, we propose a conceptual framework for designing tailored mixed-initiative approaches.

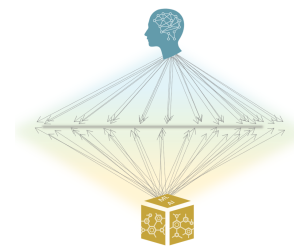
Our conceptual framework builds upon the idea that different types of users have different backgrounds and expertise, so they also have different reasoning processes and knowledge gaps, thus they can provide varying types of inputs to the AI. Therefore, we need to design interaction and explanation techniques that respect these mental models and also consider the main tasks these stakeholders are interested in addressing, as described in Section 3.2.1. To tackle this, we present *Explainability Spaces* as interaction design rationales for human-AI interfaces. The explainability spaces form a continuous spectrum and inform design decisions for I&XAI systems. We place **domain experts** and **model experts** on opposite ends of the spectrum, in the **Semantic Space** and the **Model Space**, respectively; and position **domain analysts** and **model analysts** in the middle-ground, in the **Input-Output Space**.

Figure 3.2 depicts the spectrum anchored by the four stakeholders. The resulting explainability spaces represent a human-AI interaction paradigm that can be used to understand, diagnose, or refine AI models. The most common and most straightforward space is the **Input-Output Space** (Section 3.4.2). Here, analysts can investigate how a model transforms an input into

an output by investigating how they relate to each other. Analysts are typically interested in solving concrete problems and have specific tasks. They are usually not interested in receiving a generalizable explanation, nor are they interested in how their interactions might generalize beyond their data and tasks, or how they might be transferred to other domains or models. In contrast, experts on both ends of the spectrum are interested in exactly these aspects. On the one hand, domain experts are interested in abstract domain tasks and theories. Hence, the **Semantic Space** (Section 3.4.1) describes an interaction centered around specific domain knowledge, for example, investigating the applicability of a domain theory or externalizing knowledge based on specific domain understanding. This requires expertise in the application domain, but does not require deeper familiarity with AI principles or model behaviors. On the other hand, model experts are typically interested in how their models generalize beyond a specific dataset and task. Thus, the **Model Space** (Section 3.4.3) is designed to allow for the investigation of specific model behaviors under different conditions. This requires familiarity with the metrics and parameters of the AI model.

As described, the spectrum of the explainability spaces depicts *concrete* tasks and problems in the middle, and *abstract* tasks and problems on the outside. More specifically, we can describe tasks and problems in relation to their associated knowledge dependencies, as depicted in Figure 3.3. With respect to the **application domain knowledge**, we can identify a gradient starting from the **Semantic Space**, where users target an abstract domain problem within the scope of a known domain theory; through the **Input-Output Space**, where users target a concrete domain problem; to the **Model Space**, where users are agnostic of domain theories and aim for general-purpose domain representations. On the other hand, when it comes to the **AI model knowledge**, we observe a reverse gradient starting from the **Model Space**, where users target an abstract machine learning problem within the scope of a specific model (transparent box); through the **Input-Output Space**, where users target a concrete machine learning problem; to the **Semantic Space**, where models are interchangeable (black boxes) and users are agnostic to specific models.

Explainability spaces depict different knowledge representations, as well as varying reasoning and learning approaches. The **Semantic Space** is informed by domain assumptions, hypotheses, questions, and theories. Thus, it allows for *theory-driven* learning, as well as *top-down, deductive reasoning*. On the other end, in the **Model Space**, AI models learn in *data-driven* fashion based on given distributions, attributes, parameters, and features. This allows for *bottom-up, inductive reasoning*. The **Input-Output Space** creates a middle ground for concrete problem-solving, using a concrete domain theory and AI model for *task-dependent* learning. Hence, it allows for *contrastive reasoning*.



Besides knowledge representations, these spaces allow for different types of interaction paradigms. For example, the **Semantic Space** allows for **concept-level** interactions, i.e., interaction with data abstractions into concepts and symbolic reasoning units. Thus, an ef-

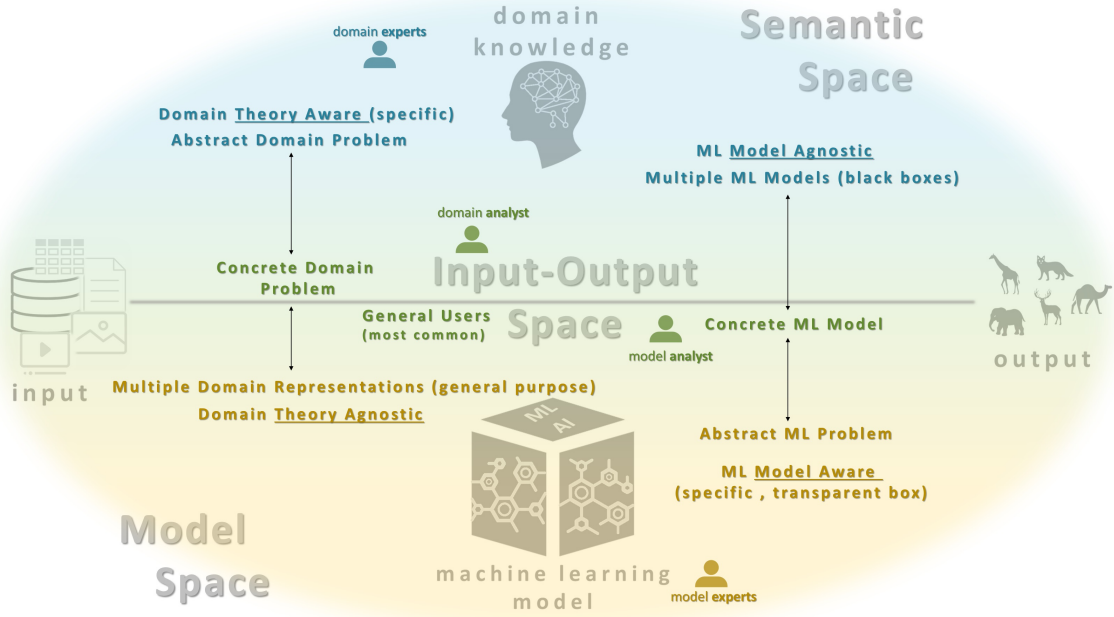
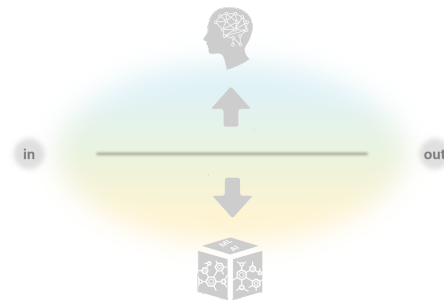


Fig. 3.3: Explainability Spaces form a continuous spectrum from the **Semantic Space** through the **Input-Output Space** to the **Model Space**. Experts are situated on either end of the spectrum, as they are interested in abstract tasks related to their area of expertise. In contrast, analysts are interested in concrete problem-solving and are situated in the middle, inspecting the direct relations of inputs and outputs.

ffective human-AI interface in this space would allow for interactions and explanations based on *concepts*, *ontologies*, and *relations*. In contrast, interactions and explanations in the **Input-Output Space** happen on a **data-level**, allowing users to inspect the transition process from input to output based on *data objects*, *instances*, and *groups*. Lastly, in the **Model Space**, interactions and explanations happen on a **feature-level** based on a data vectorization into feature spaces and sub-symbolic distributions. Thus, these interactions use *features*, *vectors*, and *distributions*.

In the following, we describe the three spaces mentioned above to exemplify prototypical design rationales for mixed-initiative systems based on the described stakeholders. However, every application has unique characteristics and thus will comprise a mixture of the described characteristics. The aim of identifying these spaces is not to argue for separability, but rather to allow for informed design decisions based on the aspects discussed in the following subsections. As mentioned in Section 3.2.1, it is worth noting that individual users can change their knowledge over time. For example, they can become an expert in a specific area over time. To counter these learning effects, we argue for designing co-adaptive interfaces [Spe+21b], allowing for a gradual transition between spaces (C7) based on the user’s knowledge gaps.



3.4.1 | Semantic Space

Humans are very skilled at making connections and deducing patterns. This general intelligence relies on our context-dependent semantic interpretation of information, as well as our procedural thinking and symbolic understanding. Humans are capable of distilling intricate abstract phenomena from information and experience, structuring them into recallable knowledge. Hence, to support capturing and externalizing this knowledge, we have to design systems that can interact on the level of abstract concepts and relations.

The **Semantic Space** can also be described as “*Domain Knowledge Space*,” it is where explainability and interaction are designed on the level of the user’s *real-world understanding* of an application’s semantics. This space can be used to perform *machine teaching* [Sim+17], i.e., to allow experts to externalize their understanding of a topic through defining, modeling, and relating concepts. It also allows for *domain intelligence augmentation*, i.e., using the human’s “*soft knowledge*” to enhance the learning of the AI [KAC17].

This space does not require the user to understand the inner workings of the AI, nor to rely on a concrete AI model. As depicted in Figure 3.4, multiple models can be used interchangeably or simultaneously. This model independence facilitates model comparison; however, it also leads to an imprecise interaction mapping, i.e., when interacting with high-level concepts that don’t reflect the model’s internal knowledge representation, it becomes challenging to map the user’s feedback to actionable refinement steps for the model.

Stakeholders – The **Semantic Space** is conceptualized mainly for *Domain Experts*. These specialized users are typically very familiar with their application domain’s knowledge and semantics. Thus, they can generalize concepts and relations to give high-level feedback to the AI model. Depending on the domain, experts might not be very common. However, for some tasks that require readily available knowledge, a broader user base can be identified. Domain experts are typically interested in building, testing, and verifying domain theories; inferring knowledge based on high-level data patterns; as well as generating ground truth data through labeling and annotation. This type of users usually works *theory-driven*; i.e., top-down.

Mental Model – Domain experts strive to build, expand, and externalize their knowledge. Thus, they are interested in “*seeing the big picture*.” This mental model enables them to search for *semantic knowledge gaps* and enhance the “real-world” understanding of phenomena based on their knowledge. This implies that interaction and explainability have to be designed on the level of the user’s semantics, using the appropriate metaphors (C5) and controls to support top-down feedback. This also presents a challenge if the user’s mental models diverge from the AI’s data models and patterns. Here, a delicate interaction design is needed to ensure the appropriate user support and guidance, avoiding rationalization (C2), conditionality (C3), and

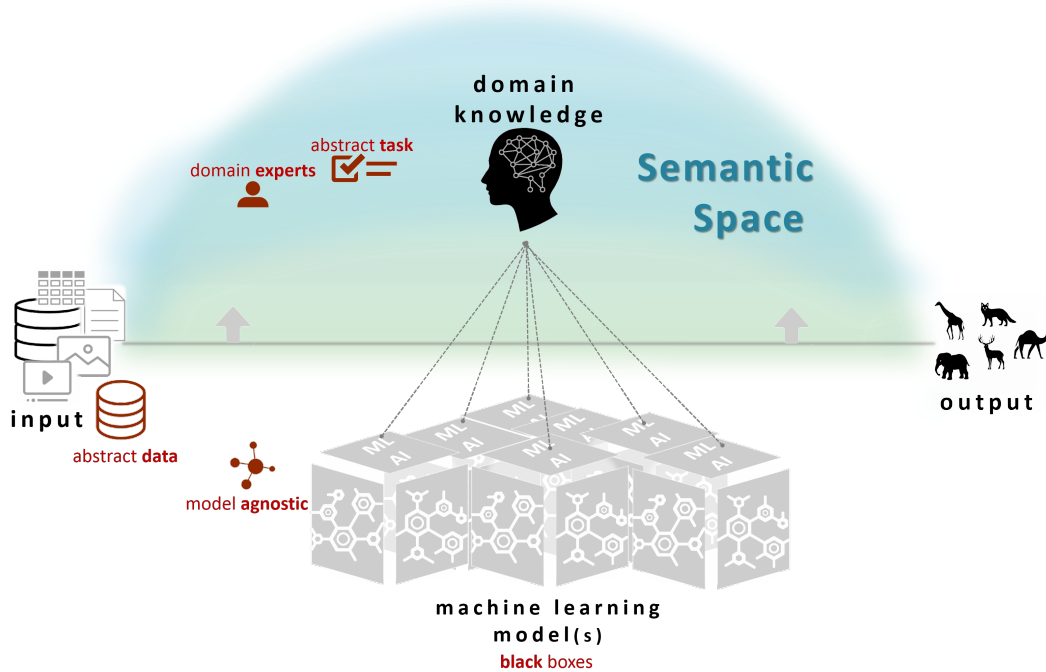
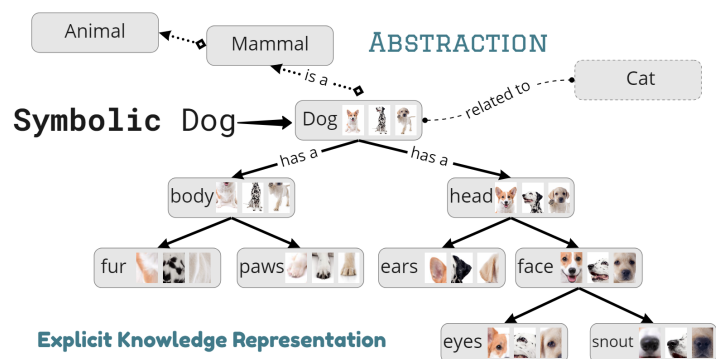


Fig. 3.4: The **Semantic Space** enables domain experts to externalize their domain knowledge, test theories and hypotheses, and compare different AI models. In this space, interactions and explanations are AI model agnostic, thus models are interchangeable. Within their domain, experts are typically interested in abstract tasks and data.

simplification (C4) pitfalls. In particular, a method for modeling a possible “semantic gap” has to be derived.

Knowledge Representation – In the **Semantic Space**, interactions and explanations are performed on a **concept-level**. Hence, users inspect and manipulate *concepts*, *ontologies*, and *relations*. This allows for an *explicit knowledge representation* (similar to Kahneman’s *System 2* [Kah11]), and thus, enables users to externalize their procedural knowledge and logical reasoning, e.g., through feature engineering. Such “*human-like*” knowledge representations are closely related to symbolic AI [Smo87], e.g., rule-based models. They allow for the mapping of high-level semantic variables and the identification of causal chains. However, maintaining and generalizing such knowledge is costly and can introduce biases [SCK18].



Reasoning & Explainability – To enable the understanding and diagnosis of the AI model’s decision-making, explanations in the **Semantic Space** need to use semantic variables. Hence, the behavior and output of the models need to be mapped onto semantic concepts and re-

lations. If this mapping is faithful to the model's inner workings (i.e., high fidelity), then it enables deductive explanations. This allows users to reason about causation and intentions, pick correct contexts to refine their understanding [Hea06], and recognize required interventions [HF96]. Note, that factorizing knowledge into concepts and relations requires distributional representations to allow for generalizability.

Learning & Interaction Paradigms – Domain experts can refine AI models based on their domain knowledge externalization. In such a case, the AI model learns from the human, for example, using a *machine teaching* paradigm [Sim+17]. Moreover, the **Semantic Space** allows for **Semantic Interactions** [End16], i.e., high-level manipulations of concepts and relations. Thus, it also supports *object-centric learning* [Loc+20]. However, if the underlying AI model does not support an object-centric representation, such an interface design might cause a “*semantic gap*” between the human and the AI, leading to a *lossy* mapping of the feedback. In other words, if the human interacts with highly abstract objects, but the underlying data is represented through low-level feature distributions, mapping human-AI interactions back-and-forth requires a translation between the two learning paradigms, which might not be accurate due to ambiguities in the mapping.

3.4.2 | Input-Output Space

Most analysts and practitioners are interested in applying models and domain knowledge to solve specific problems. In particular, they do not care about transferring AI models across different application domains or generalizing the task requirements to fit multiple types of models. They are interested in seeing the model's or application's details only when it is relevant for the task at hand.

Hence, in contrast to the expert-level spaces, the **Input-Output Space** can be seen as a “*Practical Problem-Solving Space*,” it is where explainability and interaction are designed on the level of a *concrete task and dataset* (see Figure 3.5). This space serves as an interface to justify the model's outputs and to inject *contextual-knowledge* into the model decision-making (C1), cf., *sociotechnical challenges* [Ehs+21].

In the **Input-Output Space**, the main interactions are on the level of the input and output data, inspecting how they and their relation changes when data or model parameters vary. Hence, a contrastive model comparison should be facilitated in this space, as non-experts are more comfortable making *relative* judgments (“*this model is better than this*”) than making *absolute* judgments (“*this model's quality/suitability is x%*”). To tailor, personalize, and contextualize the interfaces of the **Input-Output Space**, designing them can benefit from user-centered rapid prototyping and scenario-based design practices [RC09].

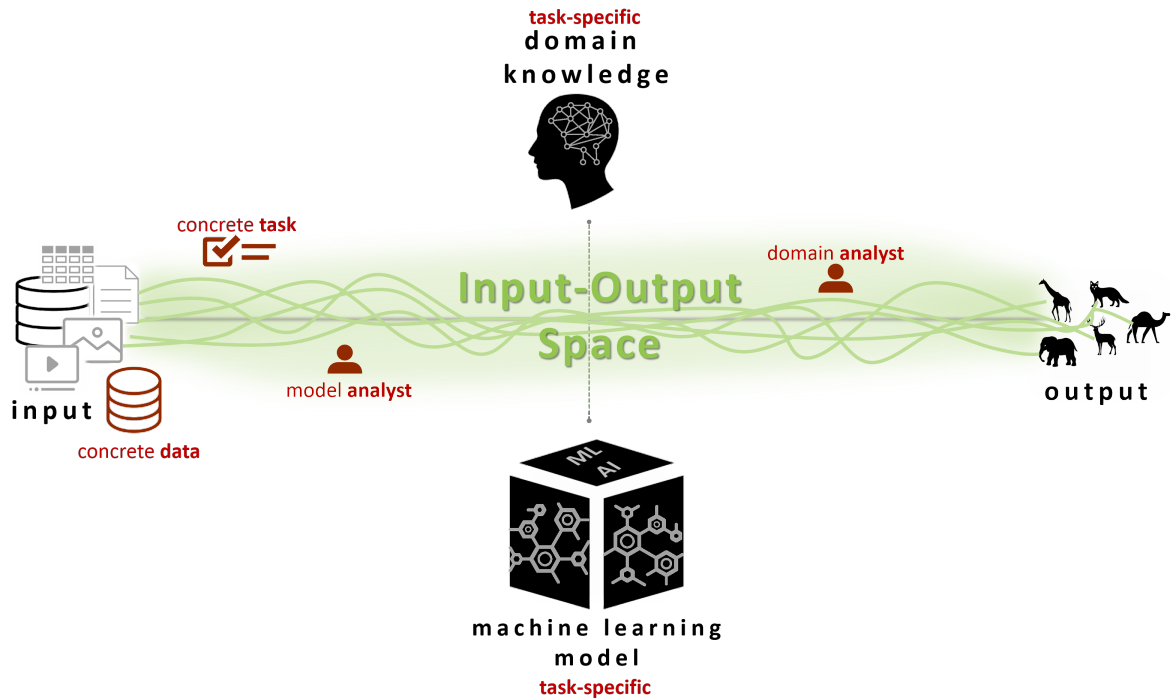


Fig. 3.5: The **Input-Output Space** enables domain and model analysts to solve concrete problems based on specific datasets and task requirements. In this space, interactions and explanations are typically contrastive, comparing the impact of changes in the model, data, or task on the input-output transformation.

Overall, this space does not require extensive domain knowledge, nor does it require extensive machine learning background. However, in practice, most analysts are trained in either the domain or the models' side. Hence, in this space, we distinguish between domain and model analysts, while recognizing that the mixture of expertise is individual and subjective.

Stakeholders – The **Input-Output Space** is targeting general users in the form of *Domain Analysts* and *Model Analysts*. These are by far the most common types of stakeholders and have varying degrees of expertise in AI and the application domain. Their common characteristic is that they are typically interested in concrete problem-solving and decision-making (C8). To achieve these goals in a mixed-initiative collaboration, they rely on tailored system designs ranging from human-in-the-loop (minimal intervention from the human); through a symbiotic intelligence augmentation; to AI-in-the-loop (minimal intervention from the AI). The design rationale is to optimize for solving the task with *minimal feedback for maximum gain*. Hence, these types of users usually work in a *task-driven* manner, providing contrastive feedback.

Mental Model – Analysts concentrate on solving concrete problems. Thus, they are interested in “*focusing on the task- and data-relevant issues.*” This mental model enables them to cut through irrelevant information and make informed decision for the set scope of their problem. This implies that interaction and explainability have to be designed on the level of the user's tasks and data. Deciding on the relevance of information and on the appropriate abstraction

level, simplifications (C4), narratives, and metaphors (C5) is the main challenge of the **Input-Output Space**. Note, that due to their partial knowledge, non-expert users are especially vulnerable to rationalization (C2) and conditionality (C3) pitfalls. Therefore, it is crucial to empower users (C6) to give appropriate feedback through implementing tailored incentives and guidance strategies.

Knowledge Representation – in the **Input-Output Space**, interactions and explanations are performed on an **instance-level**. Hence, users inspect and manipulate *data objects*, *instances*, and *groups*. This allows for a *problem-specific knowledge representation* and, thus, enables users to map the model decisions to their concrete problem specification. In particular, this knowledge representation supports comparing models and understanding the impact of changes in their parameters and input data, on the output.

Reasoning & Explainability – To understand and diagnose a model’s performance for a concrete problem, we need to explain input-output relations and contrast the current model decision with alternative decisions, for example, by changing the model’s parameters or by comparing it to a competitor. This can reveal model sensitivities and decision boundaries, allowing the analysts to make informed decisions based on the application context. Note, using contextualized and personalized explanations increases trust in the model [Kun+19].

Learning & Interaction Paradigms – The **Input-Output Space** allows for contrastive learning based on task-driven **Data Interactions**. We can thus enable users to use the results of counterfactual explanations [WMR18] in order to refine the models. Similar interactions use interventions [CMQ20], interactive probing [LGM20], and what-if questioning [Wex+20]. Here, the interaction mapping to actionable model steps depends on the chosen abstraction, but is generally equally challenging as described in Section 3.4.1 if the data objects do not correspond to internal model representations. However, in this space hybrid systems can be designed by borrowing design elements from both the Semantic and the Model Spaces.

3.4.3 | Model Space

AI model developers and experts are specialized in designing machine learning models; they know the best practices for model parameters and architectures based on their experience. They are typically interested in understanding the inner workings of models to relate the causes of different model behaviors to observable effects in the model’s output. Their knowledge is usually not tied to a very specific application domain, but rather they strive for their models to serve the most justifiable *general purpose*.

The **Model Space** can be described as a “*Model Building Space*,” it is where explainability and interaction are designed on the level of the model expert’s understanding of the *parameters*, *functions*, *objectives*, and *attributes* of a model. In addition to detecting unexpected model

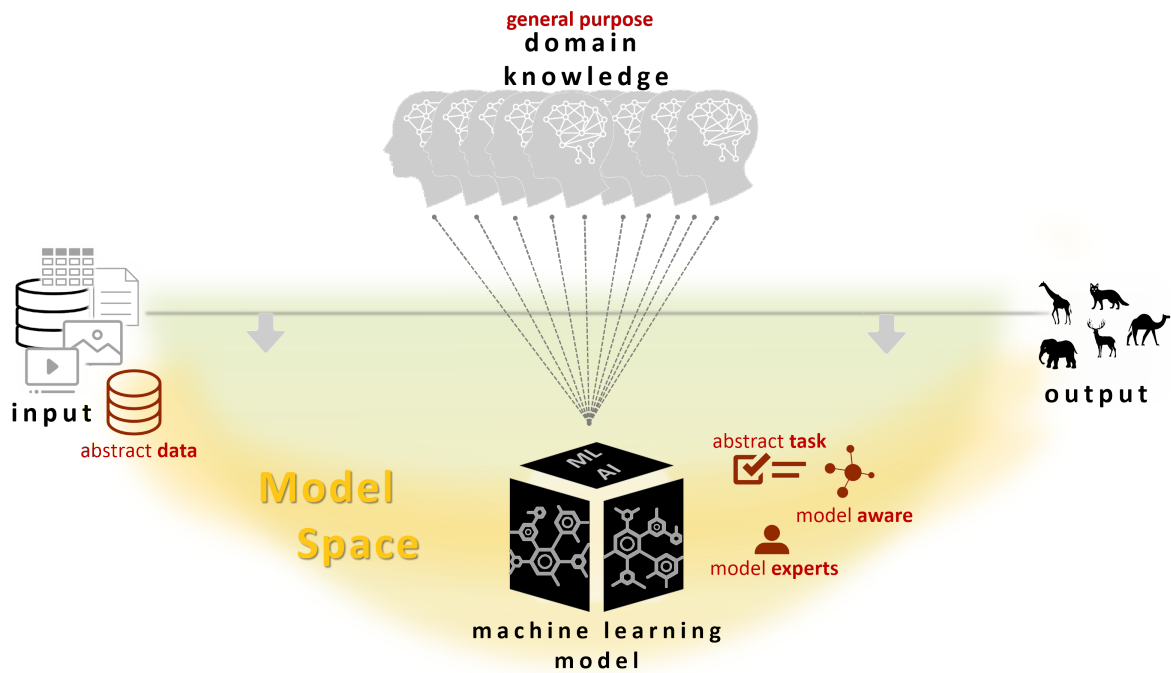


Fig. 3.6: The **Model Space** enables model experts to externalize their machine learning knowledge. They can observe the model behavior and intervene in the model-building process. In this space, interactions and explanations are agnostic of specific domain theories, but rather represent application requirements as abstract and interchangeable.

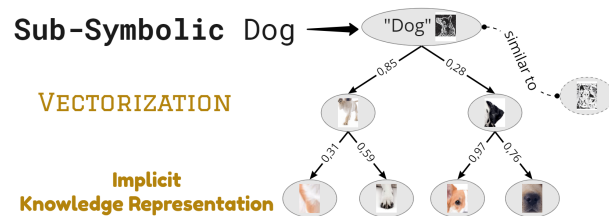
behavior and blind spots, this space serves as an interface to inject the expert’s *practical knowledge* into the model refinement and optimization.

This space does not require the users to understand the intricate details of a specific application domain, but allows for representing application requirements on a more abstract level. As depicted in Figure 3.6, multiple application domains can be served interchangeably or simultaneously. This allows for the abstraction and generalizability of the learned AI model, thus contributing to making it more robust across varying tasks and data. In the **Model Space**, users can investigate a model across the training–testing continuum, for example, to inspect the impact of data shifts. Here, the model’s inner workings are shown, opening up the black box.

Stakeholders – The **Model Space** is designed mainly for *Model Experts*. These specialized users are familiar with the details of the AI models at hand, and have extensive knowledge of machine learning. Thus, they can generalize their feedback and make targeted adjustments to the model’s parameters and architecture. Due to the level of expertise required, these users are the least common. However, as described in Section 2.1, a majority of available I&XAI interfaces are expert-level systems, catered to this specialized user group. Model experts are typically interested in investigating model behaviors to intervene in the model-building process. This requires them to think in an “AI-like” manner, understanding the impact of architectures, parameters, and data distributions. Hence, these types of users usually work in a *data-driven* manner, i.e, bottom-up.

Mental Model – Model experts strive to diagnose model behaviors to intervene in the model-building process. Thus, they are interested in “*looking under the hood*” of AI models. This mental model enables them to *refine their understanding of machine learning* and enhances the models with *practical knowledge* based on their experience. This implies that interactions and explanations designed on this level have to be as close to the underlying model as possible to avoid rationalization (C2), conditionality (C3), and simplification (C4) pitfalls. In addition, to allow for interventions on the “*model-decision level*,” the I&XAI methods deployed have to guarantee high fidelity explanations to aid the decision-making of the users (C8). Here, a *direct interaction mapping* is possible, as the feedback given by the users is targeting precise model actions and functions.

Knowledge Representation – In the **Model Space**, interactions and explanations are performed on a **feature-level**. Hence, users inspect and manipulate *features*, *vectors*, and *distributions*. This implies an *implicit knowledge representation* (similar to Kahneman’s *System 1* [Kah11]), and thus, enables users to externalize their intuitive and habitual knowledge, e.g., through best practices or distributional inferences. Such “*AI-like*” knowledge representations are closely related to sub-symbolic AI [Smo87], e.g., neural models. They allow for the modeling of inferred knowledge through low-level actions. Such distribution-dependent representations can be scaled and transferred across domains. However, they *only* explain observed data patterns and correlations based on a narrow context.



Reasoning & Explainability – To enable the explainability of model behavior and outputs on the **Model Space**, we need to represent the internal model metrics and parameters. This allows the experts to learn from the model [Arr+20]. If the I&XAI methods produce high-fidelity explanations, then through bottom-up reasoning, inductive explanations are possible. These, however, are highly dependent on data distributions, making them vulnerable to data shift and out-of-distribution issues [SKE20].

Learning & Interaction Paradigms – The **Model Space** allows for, both, direct and indirect *Model Interactions*. Direct interactions are low-level, context-dependent actions that users can perform during or after training to target specific model parameters. These are effective to test hypotheses about the model behavior, but can lead to biased or non-representative models. One solution to tackle such issues is for the system to evaluate the quality of the user’s input, modeling the system’s trust in the human feedback [Spe+19]. On the other hand, indirect interactions are suggestions by the system that the user can evaluate and respond to. These interactions are typically closer to the model’s learning and decision-making approach and, thus, can generalize and transfer to different tasks and data. Note, that the learning in this space does not have to have a semantic grounding in a specific application domain.

3.5 | Explainability Spaces for Topic Model Refinement

Topic modeling is the task of segmenting a text corpus into thematically related clusters or topics, each described with a set of topic descriptors. This task is particularly interesting as an I&XAI application, as it is highly subjective and domain-dependent, i.e., more often than not, there is *no established ground-truth* [Bec+20] to train models across domains. Additionally, the metrics used for a topic model’s internal multi-objective optimization do not always capture or align with human intuition [Lee+17b; Cha+09]. Despite these challenges, topic modeling algorithms are quite popular in different disciplines [Ble12a]. Hence, it is imperative to develop I&XAI techniques to enable users with different backgrounds to interact with these models.

In the following chapters, we present a multi-year journey to tackle the research question: *How to enable different stakeholders to understand, diagnose, and refine topic and content models using tailored I&XAI techniques?* We present an overview of four novel visual analytics workflows that introduce custom interaction and explainability methods for the task of topic modeling optimization. We have conducted comparable evaluations for all techniques to measure their effectiveness for different stakeholders. Our evaluations confirm that we can enhance problem-solving and decision-making support for mixed-initiative systems through tailored interfaces using the proposed levels of machine learning interaction and explainability. All techniques are part of the lingvis.io framework [El+19b] and tightly integrate state-of-the-art linguistic computation with visual analytics to generate new insights. More specifically, they span the following **Application Design Spaces**:

		Chapter 4	Chapter 5	Chapter 6	Chapter 7
Explainability Aspect	D1	Domain Experts, e.g., Linguists	Domain Analysts, e.g., Linguists and Political Scientists	Model Analysts, e.g., Political Scientists and Computer Scientists	Model Experts, e.g., Comp. Scientists
	D2	Mixed-Initiative Topic Modeling Refinement		Content Model Refinement for Thread Reconstruction	Mixed-Initiative Topic Modeling Ref.
	D3	Transcript of Second 2012 US Presidential Debate		Reddit Threads	2nd 2012 US Presidential Debate
	D4	Semi-Supervised Topic and Content Modeling			
	D5	Interaction and Explainability after every Training Cycle			I&XAI during Training
	D6	Justifiable Results and Outputs based on Independent Annotations			
	D7	Understanding, Diagnosis, and Refinement			
	D8	Mainly Visual, Deductive Expl.	Mainly Visual, Contrastive Explanations		Mainly Visual, Inductive Expl.
explAIner Propert.	P1	Multi-Model <i>explAIner</i>			
	P2	Model-Agnostic <i>explAIner</i>			Model-Aware <i>explAIner</i>
	P3	Global and Local Levels			
	P4	High Abstraction			Low Abstr.
	P5	None			

Semantic Space

” *All models are wrong but some are useful.*

— **George E. P. Box**
in [Box79]

Guided Topic Model Refinement using Word-Embedding Projections

” *If meaning lies even partially in usage, then you subtly alter the language every time you use it. You couldn’t leave it intact if you tried.*

— Brian Christian

The Most Human Human [Chr11] p. 207

Summary

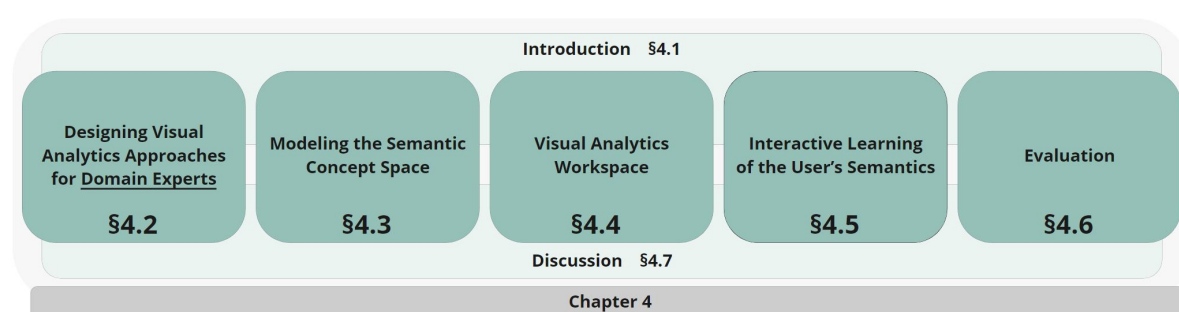
For the **domain experts**, we present *Semantic Concept Spaces*, a technique designed in the **Semantic Space**. It allows the experts to externalize their domain knowledge without directly interacting with the topic model. They remain model-agnostic. This approach maintains two content-representation hierarchies that operate on a shared vector space, enabling experts to perform *guided machine teaching* for topic modeling refinement. The two hierarchies are represented as separate layers on the interface, i.e., the **concept view** and the **topic view**. Both are anchored using a shared word-embedding projection. Domain experts can refine *concept regions* independent of a particular document collection or topic model. Their interactions directly affect the semantic relations of the underlying vectors, which, in turn, induces changes in the topic modeling. The topic view updates in response, showing changes in the topic modeling results.

Contributions

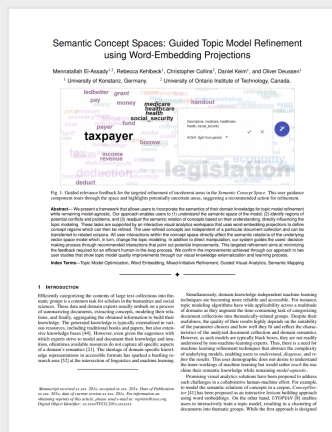
- (i) An I&XAI workflow and visual analytics pipeline for topic model refinement tailored to domain experts in Section 4.2.
- (ii) A visual analytics approach that captures user *semantic knowledge* to inform machine learning systems in Section 4.3 and Section 4.4.
- (iii) An iterative *machine teaching* technique incorporating user guidance and *relevance feedback* in Section 4.5.
- (iv) An expert user study confirming the applicability of our approach in Section 4.6.

Contents

4.1	Introduction	80
4.1.1	Research Question and Challenges	81
4.1.2	Contributions	82
4.2	Designing Visual Analytics Approaches for Domain Experts	83
4.2.1	Application Design Space	83
4.2.2	I&XAI Workflow and Visual Analytics Pipeline	84
4.3	Modeling the Semantic Concept Space	85
4.3.1	Interactive Concept Generation	87
4.3.2	Concept Neighborhood Computation	88
4.3.3	Concept Hierarchy Building	89
4.3.4	Layered Canvas Mapping	91
4.4	Visual Analytics Workspace	92
4.4.1	Concept View	93
4.4.2	Topic View	94
4.4.3	Concept Space Interactions	96
4.5	Interactive Learning of the User's Semantics	96
4.5.1	Concept Refinement	97
4.5.2	Topic Modeling Adaptation	99
4.6	Evaluation	100
4.6.1	Qualitative Results: Expert Feedback	101
4.6.2	Quantitative Results: Quality Assessments	103
4.7	Discussion: Investigating I&XAI Design Patterns	104



This chapter is based on the following publication:



Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections
Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen
 IEEE Transactions on Visualization and Computer Graphics, 26(1), 1001-1011, 2020.

El-Assady et al. [El-+20]

Contribution Clarification

This chapter is based on the publication entitled “*Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections*” [El-+20]. The system presented in this paper is an extension of Rebecca Kehlbeck’s master’s project, which I closely supervised. The user study was conducted and analyzed by me, and the paper writing was conducted by me with only slight edits from my co-authors.

4.1 | Introduction

Efficiently categorizing the contents of large text collections into thematic groups is a common task for scholars in the humanities and social sciences. These data and domain experts usually embark on a process of summarizing documents, extracting concepts, modeling their relations, and finally, aggregating the obtained information to build their knowledge. The generated knowledge is typically externalized in various resources, including traditional books and papers, but also extensive knowledge bases [Sad+17]. However, even given the eagerness with which experts strive to model and document their knowledge and intuition, oftentimes available resources do not capture all specific aspects of a domain’s semantics [Fed+17b]. The shortage of domain-specific knowledge representations in accessible formats has sparked a bustling research area [Wan+14] at the intersection of linguistics and machine learning.

Simultaneously, domain-knowledge-independent machine learning techniques are becoming more reliable and accessible. For instance, topic modeling algorithms have wide applicability across a multitude of domains as they augment the time-consuming task of categorizing document collections into thematically-related groups. Despite their usefulness, the quality of their results highly depends on the suitability of the parameter choices and how well they fit and reflect the characteristics of the analyzed document collection and domain semantics. However, as such models are typically black boxes, they are not readily understood by non-machine-learning experts. Thus, there is a need for machine learning refinement techniques that abstract the complexity of underlying models, enabling **domain experts** to *understand*, *diagnose*, and *refine* the results. As discussed in Section 3.4.1, this user demographic does not desire to understand the inner workings of machine learning but would rather *teach* the machine their semantic knowledge while remaining *model-agnostic*.

As discussed in Section 2.3.1, promising visual analytics solutions have been proposed to address such challenges in a collaborative human-machine effort. For example, to model the semantic relations of concepts in a corpus, *ConceptVector* [Par+18] has been proposed as an interactive lexicon-building approach using word embeddings. On the other hand, *UTOPIAN* [Cho+13] enables users to interactively train a topic model, resulting in a clustering of documents into thematic groups. While the first approach is designed to consider the user’s knowledge for *top-down* concept generation, the second one is data-driven, generating topics *bottom-up*. Techniques combining high-level analysis concepts with low-level model interaction, e.g. through *bidirectional* semantic interaction [Dow+18], have proven effective since “*the power of the computational models can be leveraged without their complexity.*” [End+15]

In this chapter, we present a technique that allows domain experts to incorporate the semantics of their domain knowledge for topic model refinement while remaining model-agnostic. Our approach enables users to (1) *understand* the semantic space of the model, (2) *identify* regions of potential conflicts and problems, and (3) *readjust* the semantic relation of concepts based on

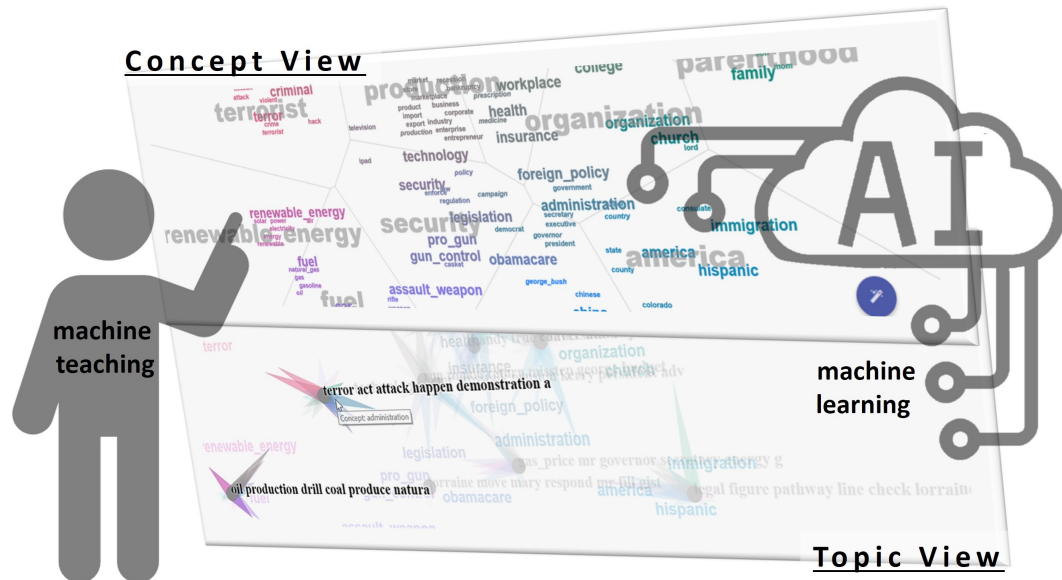


Fig. 4.1: Semantic Concept Spaces enabling knowledge externalization for *domain experts*. The **concept view** shows the *machine teaching* layer in the interface. The **topic view** shows the corresponding *machine learning* layer. When users rearrange concepts or relations in the first layer, the latter responds by recalibrating its underlying embedding space for topic modeling. Topics are represented using *topic glyphs* that use spikes to point to corresponding concepts. The length of the spikes encodes the topic’s similarity to its surrounding concepts, indicating the alignment between the model results and the user’s domain understanding.

their understanding, directly influencing the topic modeling. As depicted in Figure 4.1, these tasks are supported by an interactive visual analytics workspace that uses word-embedding projections to define *concept regions* which can then be refined. The user-refined concepts are independent of a particular document collection and can be transferred to related corpora. All user interactions within the concept space directly affect the semantic relations of the underlying vector space model, which, in turn, change the topic modeling. In addition to direct manipulation, our system guides the users’ decision-making process through recommended interactions that point out potential improvements. This targeted refinement aims at minimizing the feedback required for an efficient human-in-the-loop process. We confirm the improvements achieved through our approach in two user studies that show topic model quality improvements through our visual knowledge externalization and learning process.

4.1.1 | Research Question and Challenges

We present a visual analytics technique that tightly links *top-down* and *bottom-up* perspectives to allow domain experts to externalize their knowledge for topic model refinement without understanding the inner workings of such models. Thus, our main research question is:

How can we enable domain experts to refine topic modeling algorithms without understanding their inner workings?

To address this question, we have to tackle two challenges: First, defining accurate mappings from the user interactions back to *actionable* instructions for the topic model optimization. Second, on the visualization side, the challenge of finding an accurate and faithful *spatialization* of concepts and topics on a canvas, while reducing clutter and retaining semantic neighborhoods. We address these challenges by employing the following two design choices:

(1) Machine Teaching Our main motivation for the iterative refinement process is to enable users to *teach* [Sim+17] the machine learning model (through *concept refinement*), and in turn, the model to respond by *learning* a new refined representation (through a *topic model update*) that is presented to the users to show them the effects of their interactions. Hence, our technique relies on two independent hierarchical structures, (1) the **concept hierarchy**, representing the user’s semantics (*top-down*), and (2) the **topic hierarchy** that is based on the automatically computed results of a topic model (*bottom-up*). Both hierarchies operate on the same vector space but are presented in two separate views. The *concept view* is used as the *interactive* view for domain knowledge externalization, while the *topic view* is a *reactive* component for inspecting the topic model updates caused by refining semantic relations in the first view. This *duality* is captured in the topic and concept representations as two superimposed canvases, facilitating the analysis of associations [Gle18].

(2) Semantic Interactions and User Guidance We are inspired by the call by Endert et al. [EBN13] for interaction “*beyond control panels*,” where data is spatialized and the interaction with that *spatialization* is the primary mechanism for manipulating the data space. Semantic interaction is a type of direct feedback users can provide to embed their semantic understanding into a visual workspace [End+15]. In this paradigm, user interactions are used to feed forward into model refinements. For example, the global layout in *ForceSPIRE* [EFN12] is adjusted based on users moving words and documents to externalize their knowledge in the workspace. Cavallo and Demiralp [CD18] employ both forward and backward-projection interactions to enable user interaction with the dimension reduction algorithms. Forward projection enables users to change the high dimensional vector and see how the projection is changed, and backward projection uses direct manipulation to move nodes and see how the input vector changes. For our presented technique, we use semantic interaction to enable users to modify a *word2vec* word embedding space [Mik+13] by modifying the groupings of words into concepts. As there may be many ways to adjust the semantic space, we provide suggested interactions as a form of user guidance [Col+18].

4.1.2 | Contributions

This chapter contributes an iterative visual analytics approach that captures the semantic knowledge of domain experts for topic model refinement. The presented technique provides user guidance and relevance feedback for overcoming the “looseness” of the interaction

mapping, see Section 3.4.1. We evaluated our technique in a three-stage study. First, we conducted a mixed-method expert study, where six participants used the Semantic Concept Spaces for a topic model refinement task. Second, in a quantitative evaluation, we computed the model improvement achieved by experts, across eight model quality measures. Finally, four independent annotators rated the quality of the concept spaces and topic model results.

4.2 | Designing Visual Analytics Approaches for Domain Experts

Domain experts deeply understand their subject area and can map the relations between concepts by externalizing their understanding. We need to design visual analytics workspaces that enable deductive reasoning and interactions to allow them to represent and communicate their knowledge. Hence, the primary consideration when designing visual analytics approaches for domain experts is to allow for *machine teaching* through enabling conceptual representations and semantic interactions. This chapter instantiates the **Semantic Space** as described in Section 3.4.1. The following subsections describe the application design space for a topic model refinement approach tailored to domain experts, followed by a concrete I&XAI process model depicting a machine teaching workflow and the corresponding visual analytics pipeline.

4.2.1 | Application Design Space

The design space of this chapter is spanned by the aspects described in Section 3.5. The most important distinction is choosing the stakeholders to be domain experts (D1). This decision informs some of the other explainability aspects, for example, opting for a model-agnostic (P2), deductive explanation strategy (D8), with a high abstraction (P4). These aspects are motivated by the design rationales outlined for the **Semantic Space** in Section 3.4.1. Our approach for designing an I&XAI technique for domain experts is based on the deductive interaction paradigm of *machine teaching* [Sim+17], see Figure 4.2. Here, the experts teach the AI model based on their domain understanding, while the AI responds by learning a representation of their knowledge. Hence, the I&XAI interface updates after every model training cycle (D5) to showcase what the model has learned.

The application described in this chapter presents a topic modeling refinement technique (D2), using a semi-supervised hierarchical model (D4). The main tasks of understanding, diagnosis, and refinement (D7) are performed on specific corpora (D3), for example, the second 2012 US presidential debate between Obama and Romney, as exemplified in Section 4.6. To justify the outputs and refinement results, we rely on independent annotators (D6).

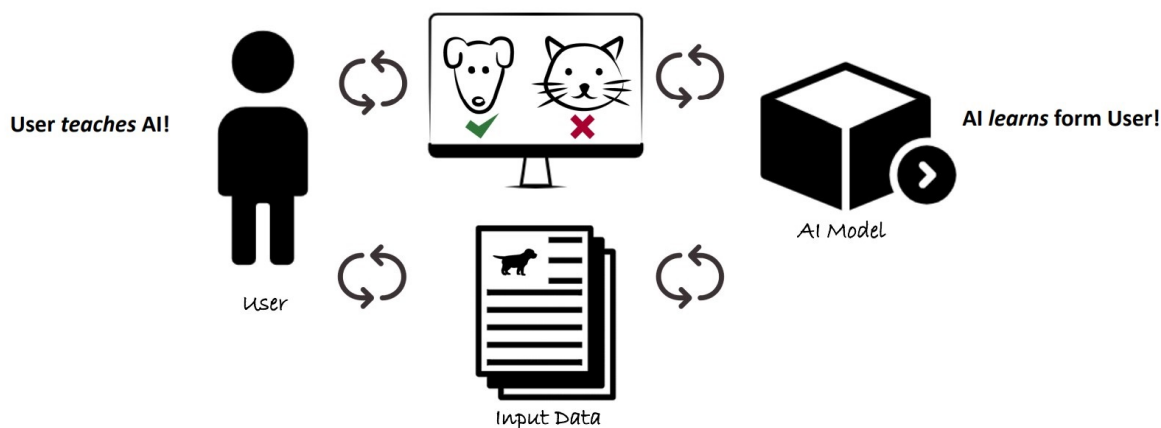


Fig. 4.2: The **deductive** I&XAI workflow starts with the user’s interaction with the data to *teach* the AI model. The model, in turn, *learns* from the user and updates its internal representations.

Tailoring the design further to domain experts, this I&XAI technique hides all the complexity of the model in a separate view and simplifies the representation (C4) using a consistent visual narrative (C5). The central metaphor for this interaction and explanation design is the notion of a co-adaptive *data dialog* (C7) in which the experts “*tidy up the concept space*” by manipulating concept hierarchies and relations; in response, the AI relearns its topic representation and, in turn, “*tidies up the topic space.*” This duality empowers users (C6) to try out different theories without having to worry about implications on the AI side (C8). However, to control for rationalization (C2) and conditionality (C3) pitfalls, the system supports *guided tours* of the space in which the AI presents prioritized regiment suggestions to the user for consideration. Furthermore, to enable the user to control the application context (C1), the system allows for different levels of data abstractions as entry points to the analysis.

Overall, our approach provides different entry points and abstraction layers for the users’ analysis; integrates users in every step of the semantic concept creation; allows for *targeted refinement* through guided relevance feedback, as well as, *concept discovery* through serendipitous exploration; enables cross-corpus and model-agnostic learning to allow the transferability of the learned concepts to other topic models and similar document collections; and abstracts from the refined semantics to update domain-specific concepts, avoiding future *cold starts* [Sch+02].

4.2.2 | I&XAI Workflow and Visual Analytics Pipeline

We designed the *Semantic Concept Spaces* as a mixed-initiative technique tailored to support users in modeling their domain knowledge through defining semantic relations between concepts. Figure 4.3 depicts the visual analytics pipeline of our approach. It starts with processing a document collection to extract relevant keywords and embeddings [LG14]. These build the basis for the semantic similarity that generates scored keyword vectors as input for topic modeling, they also initialize the interactive concept generation. This step extracts

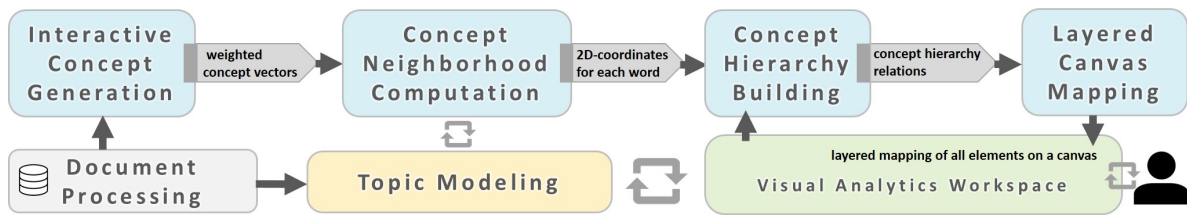


Fig. 4.3: The visual analytics pipeline shows a process model of a deductive human-AI interaction for topic model refinement in the **Semantic Space** tailored to domain experts. To inform the visual analytics workspace, the pipeline relies on four processing steps to generate the concept hierarchy, in addition to the topic model to generate the topic hierarchy.

seed words for the concept generation, optionally including user-defined structures. To define a meaningful spatialization and object positionings, concept neighborhoods are calculated using t-SNE [HR03]. After the building of the initial concept hierarchy, all elements of the visualization are projected [MH08] onto a canvas in layers. The visual analytics interface is the main workspace for the user’s interaction, this enables users to inspect concepts and topics to **[T1] understand** their relationships, **[T2] diagnose** potential conflicts, **[T3] refine** the concept space based on their domain understanding, and update the topic modeling based on the refined concept space. A continuous quality monitoring and refinement recommendation supports these tasks and enables targeted user guidance.

4.3 | Modeling the Semantic Concept Space

This section describes the four-step process of modeling the underlying data structure of the *semantic concept space*, see Figure 4.3. This includes building the concept hierarchy, as well as deriving a spatialization of all objects in the concept and topic model views based on the relations of the underlying vector space model. This spatialization is used to initialize the two views, as described in Section 4.4. To facilitate the readability of this section, the output of each of the four steps is described using a simplified example of two generic agenda items from recent US presidential debates, namely, *healthcare* and *taxes*.

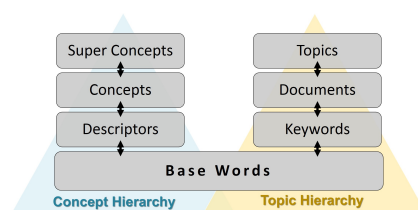
More concretely, after a brief description of the terminology, in the following, we discuss the creation of the **concept hierarchy**. However, since our approach is independent of concrete topic modeling techniques, we do not discuss the topic modeling process in detail. The **topic hierarchy** is subject to the concrete topic modeling algorithm used, see Section 2.3.1. Rather, we focus on the *model-agnostic* optimization of topic modeling through concept space refinement. Section 4.5 discusses this *iterative* refinement process and the interplay between the concept and topic hierarchies, in more detail. Both hierarchies operate on the same underlying word vectors. Changes in the concept hierarchy, therefore, influence the scoring of words and, in turn, affect the topic modeling.

Terminology: The Concept and Topic Hierarchies

To model the semantic concept space of a corpus, we consider all the words it contains and all their embeddings as a foundation (these are a subset of all words in a language’s vocabulary). Based on this set of words, we build two separate, parallel hierarchies; the concept and the topic hierarchies. Both contain four abstraction levels, sharing the lowest level of all *base words*. These two structures inform the global importance and weights of the words but are kept strictly separate, to guarantee a detachment between the user-defined concepts and the concrete topic modeling approaches. This, in turn, ensures model transferability and cross-corpus learning.

We generically refer to all words (also n-grams) in the corpus, as well as words transitively contained in their embedding vectors, as “words.”

Base Words are all words that are neither part of the higher levels of the concept nor of the topic hierarchies. They can be promoted to become keywords and/or descriptors through user interaction. On the other hand, demoted keywords and/or descriptors traverse down the hierarchy to become base words. As suggested by their name, these form the basis of the two data hierarchies.



The **Concept Hierarchy** is user-driven and reflects the semantic relation between the words based on the domain knowledge externalization of users. **Descriptors** build the lowest level above the base word and are all the words that describe a concept (one level up) but that are not concepts or super concepts themselves. Descriptors have a strict parent-child relation (1:n) to concepts. **Concepts** define the users’ semantics and are used as the main level of interaction. They are the link between the descriptors (their children) and the super concepts (their parents). Although a word can only be either a descriptor or a concept (exclusive relation), super concepts can *include* concept words. The reason for this decision is that in some corpora there are multiple super concepts that only contain one concept each. Hence, **Super Concepts** are automatically computed as a summary of the underlying region. Users can define the level of abstraction, i.e., the number of super concepts interactively. However, in contrast to concepts, the parent-child relationship between concepts and super concepts cannot be manually adjusted but is computed to give a faithful overview of the current state of the concept hierarchy at the time of viewing.

In contrast, the **Topic Hierarchy** is data-driven. It reflects the structure of the underlying corpus based on the selected topic modeling approach. **Keywords** are all words contained in all corpus documents, *including* all descriptive document keywords. **Documents** are the given unit of analysis in a corpus and are each represented by their top n -keywords. **Topics** are computed using a topic modeling algorithm and are each represented by their top m -keywords. Note, that the number of top keywords n, m for document and topics, respectively, can be adjusted by the user. By default, both parameters are set to 15 keywords.

All words used in this approach are processed through the lingvis.io pipeline [El-+19b], which includes stemming, POS tagging, n-gram extraction, stop-word removal, and scoring. We treat each word as a weighted vector initialized using a user-selected scoring function [MS99], e.g., tf-idf [Jon72], log-likelihood ratio [MS99], or G^2 [RG00].

4.3.1 | Interactive Concept Generation

The first step in this modeling pipeline is the generation of *weighted concept vectors*. Assuming that the users' domain understanding can effectively guide the automatic computation in this initial step, we allow users to *optionally* intervene and interactively edit suggested concept keywords which are used as priors for the further computation. This initial concept generation is described in the following four-step process:

(1) Seed Concept Extraction After pre-processing and annotating all the words in the document collection to be analyzed, we extract seed words. We rely on (a) *Latent Dirichlet Allocation* [BNJ03], as well as a (b) *Document Descriptor Extractor* [El-+18b] to extract the most descriptive keywords in a corpus based on word frequencies, tf-idf [Jon72], log-likelihood ratio [MS99], and G^2 [RG00] metrics. Note, that these two methods are only used as a heuristic for an initial fast separation of the overall corpus space. We do not apply LDA for topic modeling. The extracted seed words are considered the first concepts and are expanded in the next step to concept vectors. In our example, this step might return two keywords like *medical* and *taxes*.

(2) Concept Vector Expansion In this step, the initial concept words are enriched with semantically similar words using the word embedding service *ConceptNet* [SCH17] to create *concept vectors*. Words that are not part of the corpus, but contained in an enriched vector are discarded to focus the vector space and avoid skewness. Note, that we extract word embedding vectors for all words in the corpus but only vectors associated with *concept* words are called *concept vectors*. All words in a concept vector are regarded as *descriptors* for their respective concept. In our example, the concept vectors might contain the following descriptors: $\vec{medical}: \langle system, health, relief, care \rangle$ and $\vec{taxes}: \langle deduction, money, cuts, relief \rangle$.

(3) Interactive Editing and Enrichment After the first two unsupervised steps, we involve the user in the concept generation. Similar to our proposed *topic backbone* [El-+19c], users have the option to adjust the seed concepts and their vectors as they see fit. They can as well introduce new concepts or remove descriptors to adapt the generated concepts to their understanding. However, as we cannot always assume that users have existing knowledge about the corpus before exploring the visualization, this processing step is optional. If skipped, the concept vectors from the previous step will remain unchanged. A user might, for instance, choose to add the descriptor *healthcare* to $\vec{medical}$.

(4) Scoring and Ranking After the generation of the *concept vectors*, in this step, we use the scoring functions from the *Document Descriptor Extractor* (1b) to rank the descriptors of each concept. The ranking and scores of each concept is used for weighting them later on. For instance, the words *system* and *relief* in our example concepts could be ranked low as these words are, in the one case, too generic and, in the other, too undescriptive (i.e., occurring in both concepts).

4.3.2 | Concept Neighborhood Computation

Based on the weighted concept vectors, this step computes semantic concept neighborhoods to determine the spatialization of all the words in the analyzed corpus. The output of this step is, therefore, a set of *2D-coordinates* $\{x,y\}$ for each word, anchored by concept neighborhoods. We rely on *t-distributed Stochastic Neighbor Embedding* (t-SNE) [HR03] for the computation of the concept neighborhoods based on the word embedding vectors. To guarantee a more stable projection result we use the *concept vectors* as anchors throughout this work. Furthermore, we configure the t-SNE calculation with the following parameters; a perplexity of 5, a theta of 0.5 and 5000 learning iterations. These were determined based on trials with different corpora using a projection inspection approach [Str+18a]. As the perplexity parameter describes the expected minimum number of neighbors each point should have, to ensure a convergence with few errors (i.e., separable while preserving object distances), it is essential to maintain a partially overlapping set of descriptors in the enriched concept vectors. In the following, we describe the three-step process for computing semantic concept neighborhoods.

(1) Corpus and Topic Keyword Insertion In order to consider all relevant words in the projection, in this first step, we combine all corpus and topic keywords (each represented by their word embedding vector) with all extracted concepts (represented by their respective concept vectors). We use all word embedding vectors in the second step to determine the initial positioning of the concept vectors. These positions, in turn, are used as anchors in the third step. To ensure that the concept space is representative of the analyzed corpus, in this step we additionally assign the top twenty keywords from each document to their closest concept vector as descriptors. In our example, we might add keywords like *company* or *spending* to *taxes*, as well as *affordable* to *medical*. Note that this technique is independent of the concrete topic modeling approach, as long as each topic is represented by a keyword vector and each document is assigned to a topic. In this chapter, we use the *Incremental Hierarchical Topic Model* (IHTM) [El+19c] throughout, as it is deterministic and provides the required topic-document-keyword hierarchy.

(2) Initial Concept-Anchor Setting To meaningfully initialize the t-SNE projection, in this step we compute $\{x,y\}$ -coordinates for all extracted concepts and set these as anchors for the projection in the next step. We determine these coordinates based on a run of t-SNE on the

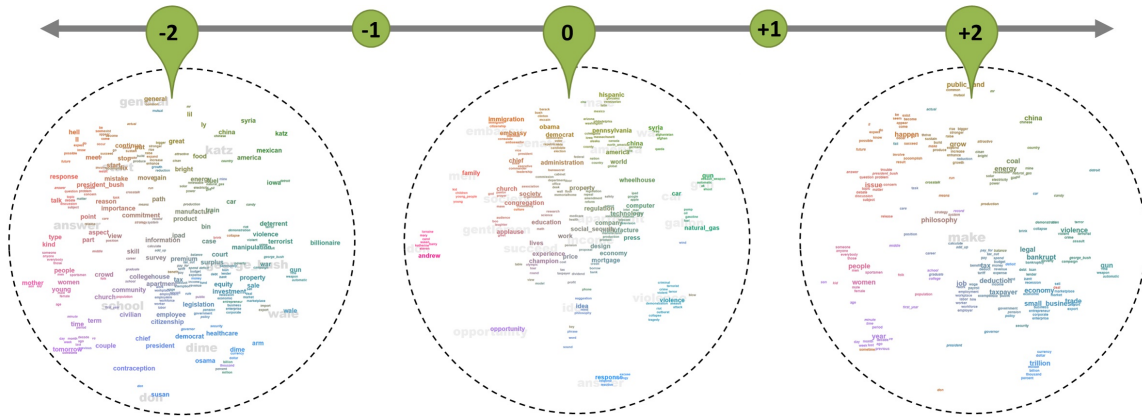


Fig. 4.4: Semantic abstraction levels for concepts. By default, the entry point of the visualization (0) shows all major concepts. Users can opt to start at a lower abstraction level (-2), revealing more concepts, or choose a higher abstraction (+2), resulting in fewer initially visible concepts.

complete set of word vectors in the corpus. This first run uses random initial positions for the words as it is only employed to determine a meaningful spatialization for the semantic concepts. Therefore, other than the $\{x,y\}$ -coordinates for the concepts, the word positions of this run are discarded and recalculated in the next step.

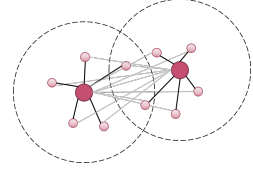
(3) t-SNE Reduction To retain a stable projection after t-SNE convergence, in this step, we use the previously determined concept positions as anchors. We then run t-SNE a second time to determine the $\{x,y\}$ -coordinates for all word vectors in the space. In later steps, when users edit and change the concept hierarchy, we re-run this step on-demand to update the concept space. In our example, each of the two concept vectors, as well as their associated descriptors have a determined position as $\{x,y\}$ -coordinates in the 2D space.

4.3.3 | Concept Hierarchy Building

Based on the neighborhoods determined by the word embedding projection, in this step, we build the *concept hierarchy relations*, getting rid of descriptor overlap by assigning each descriptor to only one concept. This is achieved based on the following four-step process:

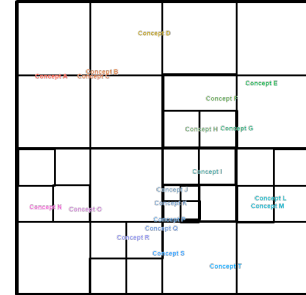
(1) Parameter and Constraint Setting The abstraction level of the concept space has a considerable impact on the visual analysis and refinement process. We, therefore, present users with a choice of different entry points in the visual analytics interface. We provide non-overlapping level-of-abstraction sliders to adjust the *semantic abstraction levels* for concepts and super concepts. To disguise the complexity of this multi-level abstraction, we present the users with five and seven discrete abstraction steps for concepts and super concepts, respectively. For example, Figure 4.4 shows three out of five abstraction levels for concepts.

The parameters chosen to determine these abstraction levels are two-fold: The minimum semantic cosine similarity threshold $\varepsilon_{similarity}$, and the minimum number of descriptors or concepts in a neighborhood $\varepsilon_{neighborhood}$. By default the similarity threshold is set to $\varepsilon_{similarity} = 0.4$, and the neighborhood parameter is set to $\varepsilon_{neighborhood} = 6$ for concepts and to $1.5 \times \varepsilon_{neighborhood}$ for super concepts. Changing the abstraction slider adjusts $\varepsilon_{neighborhood}$, directly resulting in a higher or lower level of abstraction. Based on these parameters we perform a hierarchical, density-based clustering to obtain the concept hierarchy, as described in the next steps.

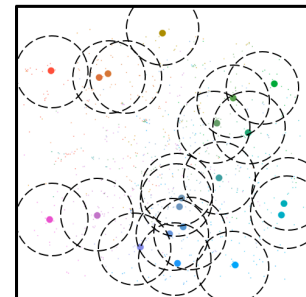


(2) Semantic Similarity Update Beside the word positioning, to perform the hierarchical clustering, we use the above-mentioned semantic similarity. We keep words inside the same cluster (i.e., concept) if they are similar with respect to their *cosine similarity*. We check this at two points during clustering. First, when deciding which words could initially form a cluster using the $\varepsilon_{similarity}$ threshold. Secondly, when clusters overlap, only clusters that have a high word-embedding coherence are merged. Otherwise, all overlapped members are redistributed to their most similar cluster. The word-embedding coherence defines the threshold for the minimal acceptable inter-cluster coherence on the current abstraction level and is dependent on the $\varepsilon_{similarity}$ and the current concept abstraction level. Hence, updating the semantic similarity based on the $\varepsilon_{similarity}$ threshold is essential to ensure a coherent semantic concept hierarchy. In our example, the word *system* and the concept *medical* might not meet the $\varepsilon_{similarity}$ threshold.

(3) Quadtree Mesh Generation The second criterion used in the clustering is neighborhood preservation. Based on the $\{x,y\}$ -coordinates previously obtained for each word, we generate a *quadtree* [FB74] mesh, such that every word is positioned in its own quadrant. The quadtree recursively partitions the 2D space into squares, where each non-empty square is further divided into four equal-sized squares. Hence, each point (i.e., word) has its own leaf node. Coincident points are stored as a linked list. The quadtree is used in later steps as an index for collision detection and neighborhood search.



(4) Hierarchical Density-Based Clustering Based on the $\varepsilon_{similarity}$ and $\varepsilon_{neighborhood}$ thresholds, we calculate the concept hierarchy in two separate clustering iterations, one for concepts and another for super concepts. We perform an *agglomerative, density-based clustering* [ZKF05] that assigns each word in the space to a concept, and, in turn, each concept to a super concept. Based on the quadtree, we extract the $\varepsilon_{neighborhood}$ nearest neighbors of each concept word and form initial concept clusters. If a concept does not have enough neighbors ($\varepsilon_{neighborhood}$) or these neighbors do not satisfy $\varepsilon_{similarity}$, we do not create a cluster. Once two clus-



ters overlap, their overlapping children are either split up or the two clusters are merged (depending on their pairwise cosine similarity). After the initial clustering is formed, all descriptors not belonging to a cluster are assigned to their most similar concept. The process is repeated to group concepts into super-concepts using their respective parameters. After clustering, our example concepts become: *taxes*: $\langle cuts, deductions, spending, company \rangle$ and *medical*: $\langle healthcare, health, care, affordable \rangle$. Note, that for *medical* the word *health* is added, as it is among its nearest neighbors.

4.3.4 | Layered Canvas Mapping

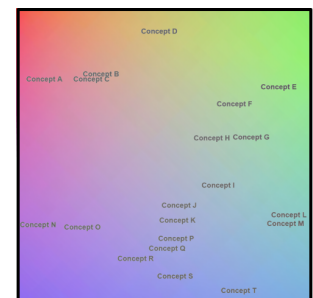
Based on the spatialization of all words and the generated concept hierarchy, the last step in modeling the semantic concept space is the *layered mapping of all elements on a canvas*. In the following, we describe the five-step layout process.

(1) Transformation and Rescaling To maximize available screen space, all data points are transformed and rescaled to the boundary of the rectangular viewport. The canvas on which the points are mapped can then be interactively panned and zoomed by users during analysis.

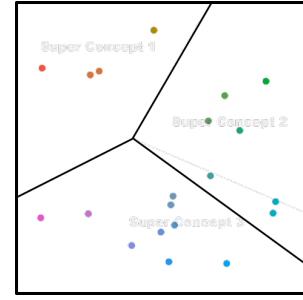
(2) Concept-Anchored Projection Based on the rescaled canvas and the adjusted concept hierarchy, we recompute a concept-anchored t-SNE reduction and project the result onto the canvas. In addition to the words being projected as points based on their respective $\{x,y\}$ -coordinates, we include a *bounding box* for every word based on its length and size (i.e., concepts are shown larger than descriptors, etc.) The quadtree index is updated in this step.

(3) Overlap Reduction Since some of the word vectors might be projected onto close coordinates on the canvas, in this step, we reduce the potential overlap. Iterating over the quadtree index, we detect areas of potential occlusion based on the object's position, as well as its bounding box. The overlapping objects are moved away from each other until the process has reached a local optimum.

(4) Color Mapping The layout process results in $\{x,y\}$ -coordinates which reflect semantically similar neighborhoods. In addition to this spatial encoding, we use the *LAB Color Space* [WS68] to assign each concept a color based on its respective 2D-position in that space. Descriptors are assigned the colors of their parent concept. This double encoding of similarity reveals descriptors that are projected in a neighborhood of a different color, indicating that their underlying word embedding is in conflict with the concept hierarchy.



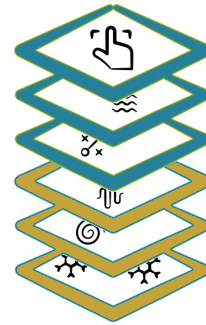
(5) Voronoi Tessellation To structure the concept space, we rely on the positioning of the extracted concept hierarchy. To enhance the visual association of the words in the space to super concepts (creating a high-level overview), we partition the space based on the extracted super concepts. We employ a *Sweepline Voronoi Algorithm* [For87] to determine super concept boundaries which can be visualized on the concept space canvas on demand.



4.4 | Visual Analytics Workspace

The generated data structures and spatialization, build the foundation of the visual analytics workspace. The *Semantic Concept Space* is designed as a layered, interactive canvas that consists of two stacks of layers; (1) the **Concept View** and (2) the **Topic View**.

Design Rationale Building on the basis of the word positioning, we designed the two views to be separate, super-positioned canvases. Users can interact with one view at a time, while the other is toggled inactive. To facilitate comparison between views, the inactive view is shown with a low opacity in the background of the active one, making its elements *shine through* the canvas. Each view is composed of three layers, representing its hierarchy levels above the base words. The layers of the **Concept View** are initialized with the extracted concept hierarchy and the **Topic View** layers with the topic modeling result. A concept refinement process enables users to directly adapt the concept hierarchy to their semantic knowledge. They can promote words up the hierarchy (base word → descriptor → concept) or demote them. Only the super concept layer is not interactively adjustable, as it is supposed to reflect a high-level view of the complete space. At any point, of this iterative process can users trigger a **recomputation of the t-SNE projection** to adjust the word spatialization to the new concept hierarchy. On the other hand, the topic modeling view can *not* be adjusted directly but is used for inspecting and analyzing the topic modeling result. Only through **recomputing the topic modeling algorithm** (on-demand) do the layers of the topic modeling change to adapt to the concept refinements. This *duality* of views enables users to *teach* the machine learning model their domain knowledge, as well as the machine learning model to respond through *learning* the new semantics.



Visual Encoding Our visual workspace is designed to support (1) finding different elements on the canvas, as well as (2) the spatial association of words. As a second-level task users are expected to (3) decode the type of word object at hand. To design an appropriate visual encoding we consulted a study of the design space of keyword summaries [FFB18] that indicates that there is a trade-off in the effectiveness of typography versus marks with respect to our tasks; search speed (1&2) and value judgement (3). According to their findings font

colored dot and/or a small label. The color of a descriptor is based on the position of its parent concept, while its position is based on its weighted word embedding vector. This enhances the detection of outliers, i.e., as colors directly reflect the user-refined concept hierarchy.

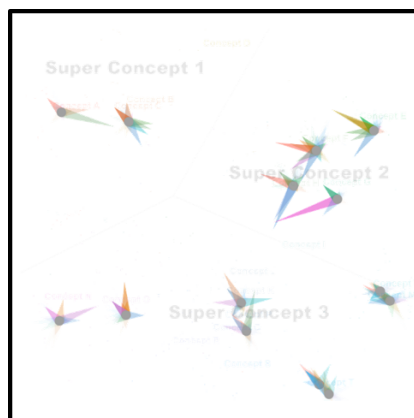
A single descriptor can be (de-)selected through a toggle-click. Selections can be (1) *deleted* from concept view, i.e., demoted to become a word; (2) *promoted* to become a concept; or (3) *(re-)assigned* to an existing concept. A group of descriptors can also be used to create a concept, promoting a selected one to a concept and all others as its new descriptors. Additionally, users can *add a word* to the descriptor layer, effectively promoting it to become a descriptor.

Concept Layer This is the central layer for the refinement of the concept hierarchy. Concepts are represented by a colored dot and/or a medium-sized label. In contrast to descriptors, the color and position of concepts are synchronized in order to anchor the concept space. When selected, a concept and all its descriptors are highlighted, while the remaining elements in the space get faded out. A single concept can be (1) *demoted* to become a descriptor, redistributing its descriptors; (2) *deleted* together with all its associated descriptors (becoming base words); or (3) *swapped* by one of its descriptors. A selected group of concepts can be (4) *merged* to form one joint concept.

Super Concept Layer The highest level abstraction in the concept hierarchy is formed by super concepts. These are represented by large, faded-gray labels positioned in the background. Super concepts are generated automatically and are non-interactive. However, users can vary the super concept abstraction level. In addition to the labels, as described in Figure 4.3.3, super concepts structure the space into subdivisions. These are represented by a Voronoi tessellation.

4.4.2 | Topic View

To build this view, Figure 4.5b, as a counterpart to the concept view, we rely on a spatialization of the keyword vectors of documents and topics. These vectors consist of a weighted set of the most descriptive keywords extracted by the underlying topic model [El-+19c]. These weights represent the importance of a keyword to their respective document or topic. In addition, every keyword has its own word embedding vector and weights, corresponding to its global importance in the corpus. The former weight is the *learned* weight by the topic modeling algorithm, while the latter is influenced by the concept hierarchy manipulation to *teach* the model. This section describes all three layers of the topic view, as well as, the *topic glyph* design.

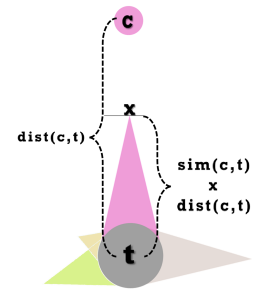


Keyword Layer This layer corresponds to the descriptor layer of the concept view. Keywords are all descriptive words extracted by the topic modeling. Some of them are at the same time descriptors in the concept hierarchy. They are represented by a black label and/or a circle. Keywords are assigned to documents. However, in contrast to the strict descriptor–concept assignment, in topic hierarchy, more than one document can share the same keyword. Other than showing or hiding them from the canvas, keywords are not interactively adjustable.

Document Layer The main unit of analysis in a corpus is the documents. These are represented by their most descriptive keywords. In addition to circles and labels, documents can also be depicted using the *topic glyph*. This indicates all related concept regions of a document, as described at the end of this section. Selecting a document reveals all its corresponding keywords. If one of its keywords is also a descriptor it gets highlighted in color, otherwise, document keywords are shown in gray. Moreover, hovering over a document object shows the underlying text for *close-reading*.

Topic Layer The top layer in this view is the topic layer. Similar to documents, topics are depicted by their top keywords. They are also represented by labels, circles, as well as topic glyphs. Selecting a topic shows all the documents assigned to it.

Topic Glyphs To facilitate the association of topics and documents to concepts, the topic glyphs relate both through spikes that point to their most related concepts in the embedding space. For each concept in the concept hierarchy, we include one spike such that the *percentage of how similar* a given topic t and the concept c is proportional to the *length* of the spike. We calculate the Euclidean distance between the two objects as $dist(c, t)$, and the normalized cosine similarity of their respective word embedding vectors as $sim(c, t)$. The distance marking the end point of the spike x is thus the normalized product of the two factors: $dist(t, x) = sim(c, t) \times dist(c, t)$.



In addition to being the scaling factor for each spike's length, the cosine similarity $sim(c, t)$ is also mapped to the opacity of the spike, making the ones pointing to *similar but distant* concepts more *visually prominent*. To facilitate finding the locations of the concepts associated with a glyph spike, we map the color of each spike with the color of the corresponding concept. We further orient the spikes to point to their associated concepts. Hence, using this representation, we can reveal topics and documents that are mixtures of different concepts. Some of which, are intentionally bringing together aspects of two concepts, e.g., the topic on the *attack on Libya* in Figure 4.5b, bringing together the concepts *terrorism* and *oil production*. In other cases, topics are only associated with one concept and are correctly placed atop that concept, showing almost no visible spikes. Overall, these spikes can be seen as the directions in which topics or documents are *pulled*, based on the semantics of the concept space.



4.4.3 | Concept Space Interactions

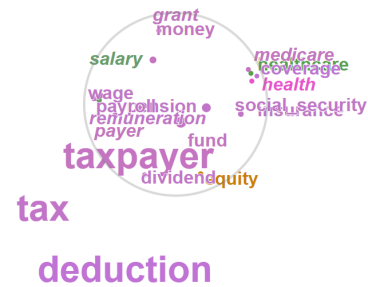
Users are supported in their exploration and analysis through a number of *instruments*. The most noteworthy interactions are:



Lasso Selection To facilitate the selection of multiple objects in the space with one interaction, we implemented a lasso selection. By holding down the ALT-key users can flexibly select multiple objects across large regions of the semantic space. Single objects can be added to, or removed from the selection using the CTRL-key.

Navigation through Word Search We provide a search query interface. Base words not found in the current hierarchy can be added as new descriptors. In addition, to visual search, users can rely on a query interface to search through the data hierarchy based on the quadtree index. Furthermore, words can be added as new descriptors if not found in the current hierarchy.

X-Ray Lens The semantic concept space is designed in multiple layers in order to avoid visual clutter and support targeted analysis. Users might want to inspect a neighborhood in more detail, for example, to look for specific words, or to understand why an area might be empty in a specific layer. We, therefore, enable them to peek through all the layers at once using a distortion lens. When activated in one of the views, the lens can be used to reveal all objects in a particular position throughout the hierarchy. The lens only operates on the active view.



Guided Tours Lastly, users can toggle the ‘magic wand’ icon  to start a guided refinement tour [Meh+17] through the semantic space. This targeted refinement zooms the canvas on the most uncertain part of the concept space and suggests a refinement that the user can accept or reject to go to the next suggestion. More details on this guided relevance feedback are discussed in Section 4.5.1.

4.5 | Interactive Learning of the User’s Semantics

The foundations of our mixed-initiative technique are the user guidance and learning components. These constantly monitor the quality of the concept space to tailor the suggestions for refinement of the most uncertain areas. However, as *serendipitous* exploration has been deemed useful for content exploration [Ale+14], our approach is designed to encourage exploratory refinement. When needed, users can request *guided refinement suggestions* on demand. Both

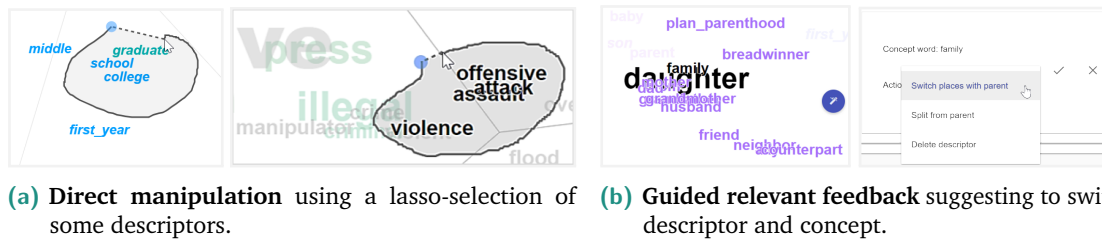


Fig. 4.6: Two options for Concept Refinement. The direct manipulation enables exploratory refinement, while the guided relevance feedback is designed for targeted refinement. Both options can be used anytime throughout the visual analytics process to adjust the concept hierarchy.

of these options are part of the *concept refinement*, designed to teach the system the users' semantics. The counterpart to this knowledge externalization consists of the topic model learning the new semantics, as well as the corresponding adaption of the topics to the learned word associations.

The main challenge of such an approach is the lack of specificity in the user interactions, i.e., the performed semantic interactions are not directly linked to actionable steps for topic model refinement. We, therefore, rely on learning the 'importance' of words for the given corpus, as well as their relations. Hence, the weighted word vectors are the common ground used for learning. In addition to corpus-specific concept refinements, we learn global word associations to enable knowledge transfer across comparable document collections. For example, if users refine a concept space for a specific presidential debate, they can reuse that space to initialize the concept extraction for another presidential debate, avoiding a cold start to the second analysis. Overall, for every word in the system we keep track of several scores, including its relevance for every concept, topic, and document, as well as for the corpus and globally. To learn the importance of a word, its level in the concept hierarchy is weighed-in, with super concepts having the largest impact.

During the refinement process, users have two controls to start a new cycle, on-demand. They can update the spatialization of objects by clicking the '*update t-SNE*'-button. On the other hand, users can retrain the topic modeling by clicking the '*update TM*'-button. To avoid confusion, the objects' positioning on the screen only changes when triggered through these controls.

4.5.1 | Concept Refinement

As shown in Figure 4.6, we offer users two ways to refine the concept space; (1) *direct manipulation* and (2) *guided relevance feedback*. Tasks performed during concept refinement include: adjusting the concept hierarchy based on the users' domain knowledge; *cleaning up* potential projection errors; resolving *word chaining* issues (i.e., two words linked through their association with a third, polymorphic word); as well as, finding '*hidden*' concepts based on the topic modeling result.

Actions can be carried out on selected objects on the canvas. All available commands fall under three primitive types:

- (1) **change of hierarchy level:** (a) promoting, (b) demoting;
- (2) **change of a parent-child relationship:** (a) reassign children, (b) reassign parent;
- (3) **splitting or merging siblings.**

Within the concept hierarchy, every level supports certain interactions: *super concepts* are not interactive; *concepts* support 1b, 2a, 3; *descriptors* support 1, 2, 3; while *base words* support 1a.

Direct Manipulation

As described in Section 4.4.1, users can directly interact with words in the concept view. The interactions available in the context menu change based on the selected object types. Users are typically offered sophisticated interactions that combine more than one of the three primitive actions. For example, as depicted in Figure 4.6a, for a selection of descriptors, users can ‘*create a concept from the selection.*’ This is equivalent to (1a) promoting the descriptor clicked (in this case: ‘*college*’) to become a concept, then (2) assigning all other descriptors in that selection to the newly created concept.

Guided Relevance Feedback

Actions can also be recommended to the user. The intuition behind this targeted refinement is that the system offers users a *guided tour* through the data space, pointing them to potential problems, intending to achieve *maximal gain for minimal feedback*. To start a tour, users click the ‘*magic wand*’-button  which opens up a suggestion window, displaying the first refinement recommendation, as shown in Figure 4.6b. Simultaneously, the concept map zooms to the region of refinement, centering and highlighting the objects concerned. Users can then accept or reject the suggestion, or choose a different interaction to perform. Internally, this action prompts the *refinement recommender* to reevaluate the semantic space based on information collected through the constant *quality monitoring*, filling up the recommendation queue.

Quality Monitoring In order to make more informed decisions for user guidance, the system tracks several quality criteria across actions: the scatter of word clusters (words in the same concept cluster or neighborhood) based on *cluster-density*, *intra-cluster variance*, and *inter-cluster variance*. Here, we rely on several cluster validity measurement techniques [IBL05], including the *root-mean-square standard deviation* [HBV02], as well as the *S_Dbw validity index* [HBV02]. In addition, for *every word*, we keep track of its *neighborhood count*, *semantic similarity to its children and/or parent*, as well as its *spatial distance to children and/or parent*. Furthermore, the quality monitoring component evaluates the *internal quality of the topic modeling* based on the criteria outlined in Section 2.3.3.

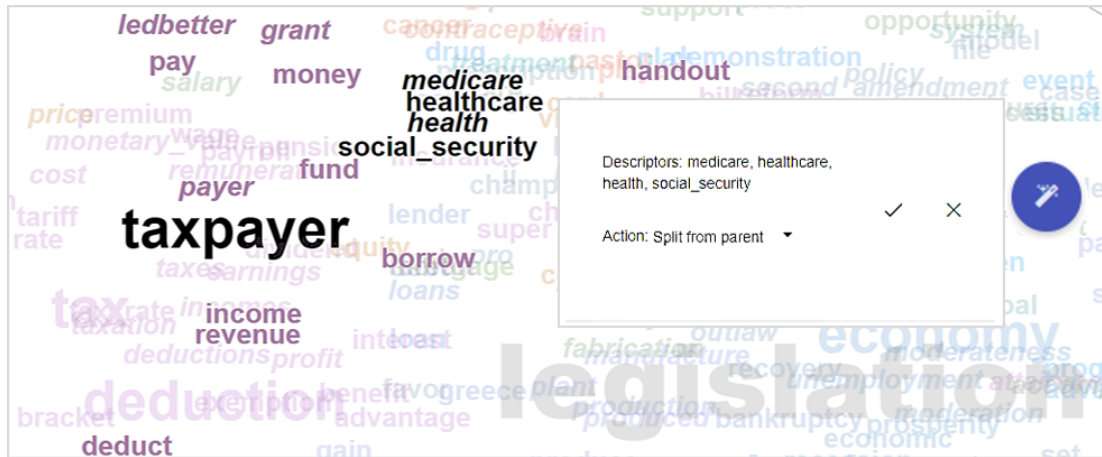


Fig. 4.7: Guided relevance feedback for the targeted refinement of incoherent areas in the *Semantic Concept Space*. This user guidance component tours through the space and highlights potentially uncertain areas, suggesting a recommended action for refinement.

Refinement Recommendation Based on the results of the quality monitoring, the recommender keeps a constantly-updated queue of words and their suggested actions. This queue is formed from words that are chosen based on their importance to the corpus using *tf-idf* scoring [Jon72]. The intuition is that users should be presented with refinements affecting the *worst-performing*, high-impact words, in order to achieve substantial improvements and give *minimum feedback for maximum gain*. Words important to the whole corpus (high *tf-idf*) should be concepts, while those important for single documents (low *tf-idf*) should be descriptors. After retrieving the top 50 high-impact words using *tf-idf*, the recommender loop starts by (1) *ranking* them based on the quality metrics; (2) *sorting* possible refinement actions for each word, as well as word clusters, using a decision tree; and finally, (3) *adding* the words along with their most suitable recommendation to the queue. The queue is reevaluated if the concept space changes. The recommended actions come from the three primitive interaction types listed above. An example of a recommendation is depicted in Figure 4.7.

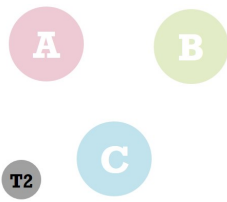
4.5.2 | Topic Modeling Adaptation

As described in Section 4.4.2, we depict the associations (semantic similarity) of topics and documents to concepts through topic glyphs. During refinement, four different cases of topics (and documents) can be observed.

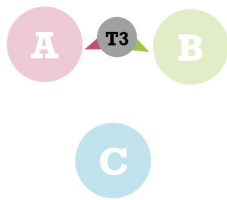
- A

C

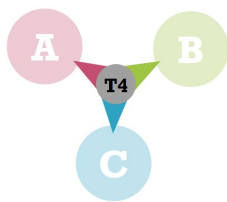
(1) Single-Concept Topics These topics are related to only one concept and their glyph is typically placed atop that concept. In the side figure, for example, topic T1 is placed on top of concept B (green). Hence, the corresponding glyph has no large spikes. Such topics are coherent with the concept structure and do not require further refinement.



(2) Unrepresented Topics These are topics that are not related to any concept. They are placed atop an empty region in the concept space,  with no large spikes visible in their glyphs. If relevant, users can introduce a respective new concept to the empty area, otherwise, the top topic keywords have to be inspected further.



(3) Multi-Concept Topics These topics are related to a close neighborhood of concepts. They are placed in-between the related concepts,  with small spikes to these concepts in the corresponding glyph. If some of these concepts are closely related, users can merge them, otherwise, no further refinement is needed.



(4) Concept-Incoherent Topics These topics are related to multiple concepts *across* the space. They are typically placed in-between the related concepts according to similarity,  with large spikes pointing to these concepts in the corresponding glyph. This is the most critical case, it has to be resolved through targeted concept refinement and a closer inspection of the topic hierarchy.

In the refinement, users can investigate words that are deemed important to a topic but are not relevant for concept distinction. These words are typically less descriptive (in terms of their unique semantic contribution) to a topic than those modeled by the algorithm. For example, in a topic model, based on a bag-of-words-representation, (frequent) verbs, adjectives, and adverbs can cause documents to seem similar even though they are not. To avoid such chaining effects, we use the learned weights and scores from the concept refinement to readjust the keyword weighting for the topic model training. These act as “must-link” and “cannot-link” constraints [AZC09] to introduce the user-defined notion of relevance to the topic modeling. Ideally, a stable and deterministic topic model [El-+19c] should be used in this process. However, probabilistic models are also applicable but take longer to converge due to *inconsistencies* between their individual runs [Ble12b].

4.6 | Evaluation

Methodology We evaluated our approach in three stages. (1) To assess the **usefulness and usability** of our technique, we conducted an expert mixed-initiative study [Ise+13] with six participants, involving two phases of semi-structured interviews, as well as a pair-analytics session [KF14]. (2) Based on the concept space refinements of the study results, we automatically computed the **topic modeling improvements** across eight quality metrics, as described in Section 2.3.3. (3) To assess the **perceived quality difference**, we asked four independent annotators to rank the quality of five different concept spaces and their

associated topic modeling results. This section reports the results of all three stages, grouped into *quantitative* and *qualitative* insights.

Dataset and Controls For our evaluation, we sought a dataset with broadly familiar content, where the expected topic distribution is known. To ensure comparability across all chapters, we chose to use the second US Presidential Debate between Romney and Obama in 2012, as our corpus for all studies. In this dataset, we treat every speaker utterance as a document.

Participants and Tasks After conducting a pilot study, we designed the expert user study for a target group of people generally interested in politics. Based on our experience in previous works, we envisioned that scholars in the social sciences and the humanities would fit this profile, we, therefore, invited two political scientists $Pol_{\{1,2\}}$ and two linguists $Ling_{\{1,2\}}$ to take part in the study. As a control group, we recruited two computer scientists $CS_{\{1,2\}}$ with no prior knowledge of debate analysis. For the annotation task, we invited two political scientists $Pol_{\{3,4\}}$ and two linguists $Ling_{\{3,4\}}$. Across the three stages, our goal was to assess the technique’s support for the three tasks of [T1] understanding, [T2] diagnosis, and [T3] refinement of the concept space to progressively update the topic modeling.

4.6.1 | Qualitative Results: Expert Feedback

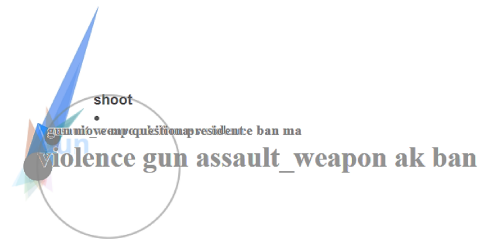
The six sessions of the expert study, 1.5h each, were structured into three parts. We started with a semi-structured interview (40 mins) in which the explanation of the approach was interwoven; we checked the expectations of the participants before introducing new concepts. Second, we gave the participants full control over the tool and asked them to refine the concept space. In this pair-analytics session (30 mins), we encouraged participants to think aloud. Lastly, we ended the study with another semi-structured interview (20 mins) that incorporated the participants’ expectation statements from the first part, as well as a reflection of the analysis process. All sessions were screen-captured and audio-recorded for further analysis.

Initial Feedback Regardless of their prior experience with topic modeling or familiarity with the data, all experts saw benefits in our technique and potential application areas. Some of them were more familiar with automatic content analysis, like $Ling_1$, who stated that she “[had] a mixed experience with using topic models, [as] they sometimes extract useful concept but often contain nonsense words.” On presenting her the idea of our approach she commented: “I find it a very helpful concept to be able to include prior knowledge, we often do that after modeling, but it is also good to do before.” On the other hand, CS_1 declared that he had no previous experience in using topic modeling but has developed similar machine learning techniques before. He observed that “the idea of relevance feedback is good because you adapt what you see, but with a topic modeling black-box, you don’t know what the machine learns, which [he] would like to be able to evaluate.”

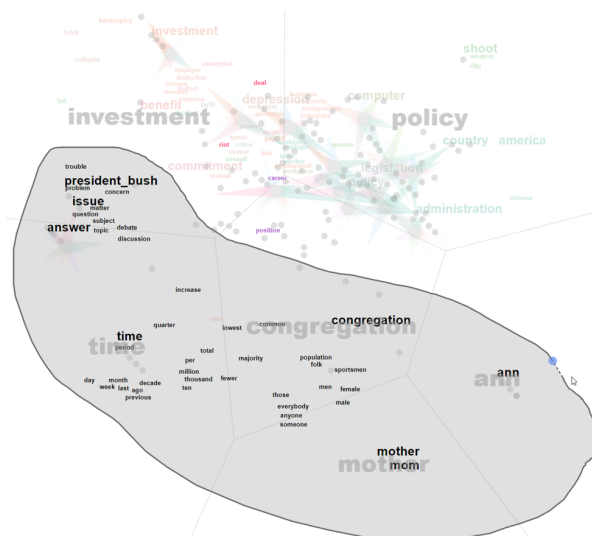
When asked about his expected workflow, Pol_2 was quite certain how he wanted to proceed in the refinement, stating that he “*would try to differentiate between situational and general concepts.*” However, when questioned about the guided refinement, he responded that “[*he is*] *not sure about the guidance component, it might speed up the process, but it’s not transparent.*” CS_2 , on the other hand, described his expected workflow as follows: “*My workflow would be to start with looking at concepts, to figure out outliers, then group concept descriptors.*”

Observations During the Refinement Process Before they started their interactive sessions, we asked participants to rate different concept space abstractions. We got mixed feedback on the favored entry point to the analysis from the different participants. Pol_1 , for example, stated that she would be interested in “*the system to suggest a full space and [that she] would clean up the unsuitable details*” As opposed to this bottom-up refinement, $Ling_1$ said: “*A full space is too cluttered for me, I need to understand the problems, then build up my solution.*” $Ling_2$ also favored a top-down refinement stating that “*starting with fewer concepts is helpful for the exploration.*” All participants agreed that interactively choosing an entry point to the analysis is a desirable functionality that they would make use of, depending on their tasks and data, different abstraction levels would make sense for the refinement.

The participants then generally continued with exploring the semantic space first, to get a better understanding of the corpus. At this stage, many used the x-ray lens to explore empty regions in the concept and topic views, or find related words to an object. For instance, Pol_1 was interested in exploring a topic on *gun violence* and used the lens to find related keywords. After some refinements, she remarked that “*it is satisfying to clean this mess and see the model respond,*” describing the interface as a “*neat combination of ecstatically pleasing components.*”



During the exploration of the space, $Ling_2$ pointed out that “*the tool is good at identifying communities.*” After observing the concept regions, $Ling_1$ found a region in the space that she deemed incoherent, commenting: “*This is a fuzzy area.*” The contained descriptors included a *moderation cluster*, *temporal keywords*, and *person names*, she pointed out that “[*she*] *wouldn’t use these words.*” She continued by selecting the entire region using the lasso tool and deleting its content. She then stated building up new concepts that, in her opinion, described the underlying phenomena more accurately than the previous space. She then updated the projection and topic



modeling, observing a better semantic representation. The effects of direct manipulation were also praised by *Pol₂*; he stated *“I can now build my own semantics and theories to test out.”* Similarly, after observing some positive changes in the semantic space, *Ling₂* said: *“It’s like adding my intuition to a stupid machine.”* She also commented on the guided refinement, finding the suggested operations useful, only disagreeing with one proposed action.

Overall, the refinement process was well-received by all users. Most went through several iteration cycles (up to eight in 30 mins), often trying out the effects a refinement would have on the topic modeling and claiming that they had a better intuition of the expected model after the first few cycles. When asked about their final goal for refinement (or stopping criterion), most participants stated that it would be a trade-off between the importance of the result (e.g., when used for further analysis) and their familiarity with the domain semantics. *Pol₂* observed that *“not every topic has to be coherent to be helpful,”* meaning that his final goal would not be to make each topic perfectly fit only one concept but rather to make them meaningful.

Usability and General Assessment All experts enjoyed the interactive refinement session. In her general feedback, *Ling₁* immediately cautioned that *“it’s so easy to use topic modeling results in a wrong way, I find it good to explore the space and understand the reasons for the results.”* The same sentiment was shared by *Ling₂*, who stated: *“I like the idea that I can use my knowledge to put things in order, that’s really useful and very satisfying.”* On the other hand, *Pol₂* proposed to use our approach for communication, saying: *“I would like to use this tool for presentation, it would be a nice feature to animate though the regions and create a storyline”* However, he also requested that *“[he] would like to be able to track the changes happening in the topic modeling over the different refinement cycles.”*

Reflecting his workflow *CS₂* commented: *“I was adjusting concepts to reduce topic spikes, but then I started asking why a topic is [placed] there and how the system understands my interactions.”* He continued that *“the color encoding of the words was useful to find outliers.”* However, *CS₁* commented that *“since [he is] not an expert in with these data, [he] would like to verify the performance of the topic modeling automatically.”* Nevertheless, *“[he] like[d] the design of the interactions, [that he] can insert words directly on the canvas and not in a side panel.”*

The most notable additional features suggested by experts include; a feature to start typing to autocomplete a concept and jump anywhere in the space (*CS₂*); to add a space-out or blow-up button for a selected area (*Ling₁*); to “deep-search” for similar concepts (*Ling₁*), and finally to enable zoom-dependant resizing of text labels (*Pol₂*).

4.6.2 | Quantitative Results: Quality Assessments

Based on the logging results of the expert studies, we can compare the automatic quality metrics for the topic modeling. We did not observe any significant difference in the refinement

Compared Output	Semantic Concept Space	Topic Modeling Result
Manual Refinement	1.00 (0.00)	2.00 (0.70)
Guided Refinement	2.25 (0.43)	1.50 (0.51)
Default Model	3.25 (0.82)	2.75 (1.09)
High-Abstraction	4.00 (0.71)	4.25 (0.82)
Low-Abstraction	4.50 (0.87)	4.50 (0.50)

Table 4.1: Ranked output of the concept space and the corresponding topic modeling result (scale: 1–best, 5–worst) according to the annotators’ perception of quality, the standard deviation is shown in parentheses.

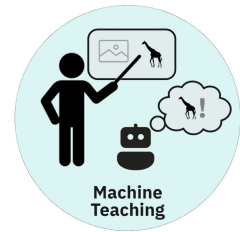
results across the three user groups. Experts and non-experts, alike, were able to enhance the topic modeling results through our technique. The average relative change, from the initial model to the refined model, based on the eight observed quality metrics was as follows: **Coherence** (-5.49%); **Separation** (-12.09%); **Distinctiveness** (331.31%); **Point-wise Mutual Information** (4.32%); **Certainty** (0.66%); **Branching Factor** (-26.47%); **Compactness** (-11.77%); and **Topic Size** (1.45%). Hence, while on average the topics became slightly less coherent and separated, they became *significantly more distinct* during the refinement.

The last stage of our evaluation is the assessment and ranking of five models by four annotators. The manual refinement model was generated by a participant in the first study. The other models were created following the guided refinement suggestions, the initial model at the default level of abstraction, and the initial model at high and low abstraction levels. Annotators were given the agenda of the debate and asked to rank the concept view, as well as a keyword list of the corresponding top topic descriptors. In our annotation guidelines, we asked them to base their ranking on four criteria: completeness, coherence, separability, and descriptiveness. The results, as shown in Table 4.1, confirm that the manual refinement of the concept space yields the most well-perceived concept view, while the guided topic refinement leads to the highest-ranking topic modeling result. This might be due to overlooked, uncertain regions during the manual refinement. In their annotations, they pointed out that incoherent areas are mainly comprised of clusters of names or general words, like *tomorrow*, *country*, etc. This suggests a potential improvement through domain-specific filters for non-informative words.

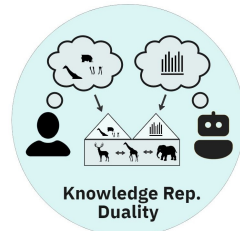
4.7 | Discussion: Investigating I&XAI Design Patterns

Based on the presented technique, in this section, we discuss three potentially generalizable I&XAI design patterns. These are tailored to interactive and explainable applications for the *domain experts* in the [Semantic Space](#). An overview of all design patterns discussed in this dissertation can be found in Section 8.4

(1a) Machine Teaching To allow experts to *externalize* their domain knowledge based on their theories and semantic understanding, we need to facilitate a *top-down* interaction approach. Hence, we propose enabling users to perform machine teaching, i.e., providing their domain knowledge to the system using semantic interactions. The concept relations derived from the user can then be used to train an aligned representation on the AI side, communicating the induced changes back to the user.



(1b) Duality of Knowledge Representation To facilitate the interaction mapping between explicit knowledge relations provided by the user and implicit feature distributions learned by the AI, we need to model both types separately. Here, a dual knowledge representation design can provide a tailored solution that facilitates the effective mapping between the human and the AI sides.



(1c) User-Requested Guidance Experts usually prefer to have control over the analysis. Hence, we can enable user agency in an *AI-in-the-loop* workspace. Here users can request guidance and the AI points them to the areas in the analysis where it's most uncertain, like asking a *teacher* for help when not knowing or unsure about answers.



Input-Output Space

Semantic Decisions

” *Simplicity is the Ultimate Sophistication.*

— **Clare Boothe Luce**
Stuffed Shirts, 1931

Progressive Learning of Topic Modeling Parameters

” Judgment can [...] be described as measurement
in which the instrument is a human mind.

— Kahneman, Sibony, and Sunstein
Noise: A Flaw in Human Judgment [KSS21]

Summary

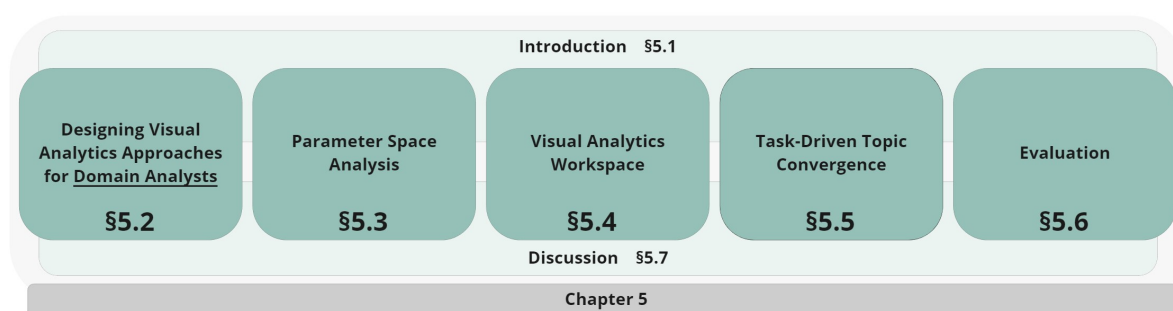
For the **domain analysts**, we present a technique designed in the **Input-Output Space** (with a focus on *semantics*). It tackles the refinement of topic models through a user-driven **progressive learning** process that does not require a deep understanding of the underlying algorithms. The system initializes two model configurations based on a parameter space analysis that enhances document separability and lets the models *compete* for the analyst’s satisfaction. Using automatic topic matching, topic summaries, and parameter distribution views, analysts can investigate the modeling results before providing *document-based relevance feedback*. This feedback is used to distill a user-endorsed topic distribution, allowing the system to train new model instances while keeping the preferred model instances as an anchor for the interaction, and restarting the feedback process to iteratively converge on a refinement.

Contributions

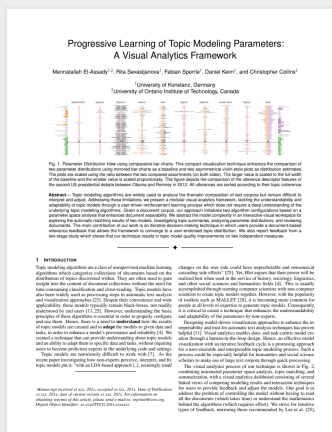
- (i) An I&XAI workflow and visual analytics pipeline for topic model refinement tailored to domain analysts in Section 5.2.
- (ii) A *parameter space analysis* counteracting potential *cold-starts* in Section 5.3.
- (iii) A task-oriented visual analytics workspace for comparative understanding and *relevance feedback* in Section 5.4.
- (iv) A *progressive learning* technique for topic refinement in Section 5.5.
- (v) An expert user study confirming the applicability of our approach in Section 5.6.

Contents

5.1	Introduction	112
5.1.1	Research Questions and Challenges	113
5.1.2	Contributions	114
5.2	Designing Visual Analytics Approaches for Domain Analysts	114
5.2.1	Application Design Space	115
5.2.2	I&XAI Workflow and Visual Analytics Pipeline	116
5.3	Parameter Space Analysis	116
5.3.1	Data-Driven Feature Selection and Weighting	117
5.3.2	Document Descriptor Vector Generation	118
5.3.3	Initializing the Workspace	119
5.4	Visual Analytics Workspace	119
5.4.1	Topic Matching	120
5.4.2	Topic Summarization	121
5.4.3	Parameter Distribution Analysis	123
5.4.4	Document Relevance Feedback	124
5.5	Task-Driven Topic Convergence	127
5.5.1	Reinforcement Learning	127
5.5.2	Iterative Topic Evolution	128
5.5.3	Global Learning of Parameter Scales	129
5.6	Evaluation	130
5.6.1	Quantitative Results: Model Improvement	132
5.6.2	Qualitative Results: Expert Feedback	133
5.7	Discussion: Investigating I&XAI Design Patterns	134



This chapter is based on the following publication:



Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework

Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle,
Daniel Keim, and Christopher Collins

IEEE Transactions on Visualization and Computer
Graphics, 24(1):382–391, 2018.

El-Assady et al. [El-+ 18b]

Contribution Clarification

This chapter is based on the publication entitled “*Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework*” [El-+ 18b], which was awarded an honorable mention for best paper at IEEE VAST 2017 (one of the top three papers for the year). The system was developed by me with the help of Rita Sevastjanova and Fabian Sperrle as research assistants. The user study was conducted and analyzed by me, and the paper writing was conducted by me with only slight edits from my co-authors.

5.1 | Introduction

Domain analysts also face the task of categorizing document collections into topics. However, as described in Section 3.2.1, they typically refine topic models for a concrete task and text corpus. For example, through our long-standing collaboration with political scientists, we got to know that their analysis workflow consisted of deploying multiple topic modeling algorithms, tweaking their parameters, and comparing their results until satisfied. To facilitate this targeted topic model refinement task, in this chapter, we introduce a progressive learning technique that encapsulates this targeted refinement as a visual analytics process.

The proposed process combines automated parameter space analysis, topic matching [El+18c], and summarization, with a visual analytics dashboard consisting of several linked views of competing modeling results and interaction techniques for users to provide feedback and adjust the models. As depicted in Figure 5.1, our goal is to address the problem of controlling the model without having to read all the documents (which takes time) or understand the mathematics behind the algorithms (which requires effort). We strive for intuitive types of feedback, mirroring those recommended by Lee et al. [Lee+17b], such as *which topic better suits a document* or *which word does not belong in a keyword set*.

The user feedback is used to generate new candidate models, which can be further refined. The process supports users in comparing, evaluating, and optimizing topic models in order to achieve an output that more coherently describes the document collection. The visualization step of the process is designed with four linked views, each to support a task: exploring the automatic matching results of two models (**Topic Matching**), investigating topic summaries (**Topic Summarization**), analyzing parameter distributions (**Parameter Distribution Analysis**), and reviewing documents (**Document Relevance Feedback**). Our tasks follow the process of understanding, diagnosis, and refinement; and conform with the model understanding and comparison tasks proposed by Alexander and Gleicher [AG16]. However, where they choose to be *parameter agnostic*, we choose to reveal the parameter space and how the values affect the model while still remaining model-agnostic.

The amount of feedback to provide in each iteration is up to the user. The more documents get rated, the closer the next model will be to the ideal topic composition. However, at some point, the cost of the refinement process outweighs the benefits of an unsupervised algorithm. The balance would be providing minimum feedback for maximum improvement. Thus our visualizations are also targeted at guiding users to those ambiguous documents and topics for which feedback would have the most impact on the next iteration of learning. Through enhancing feature distributions and descriptor keywords we enable users to analyze the effects of parameters on topic models, understand the impact of document descriptors on the topic keyword vectors through topic summarization, and, ultimately, optimize the topic modeling results in an iterative loop using reinforcement learning.

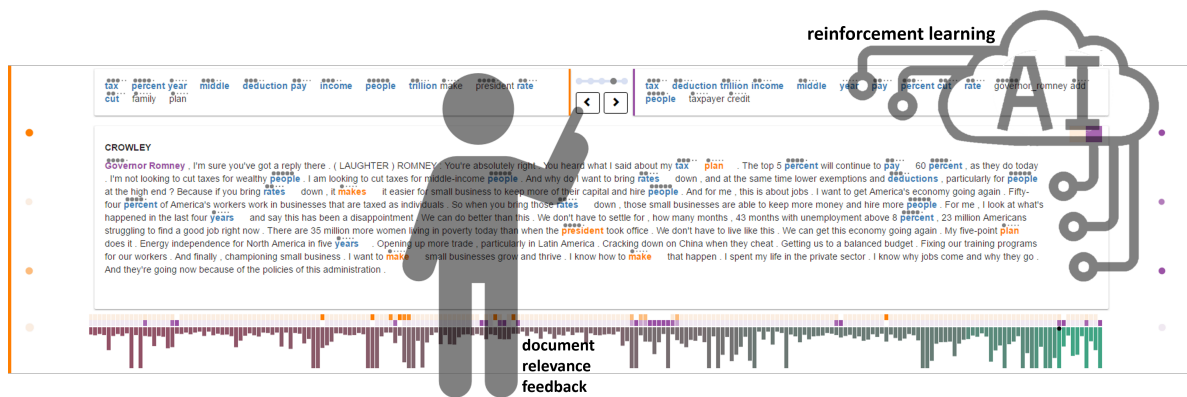


Fig. 5.1: Progressive Learning based on relevance feedback by *domain analysts*. Users interact with the central decision slider to choose *which topic modeling result (left or right) better represents the shown document*. This feedback is in turn used to refine the under-performing model.

We evaluated our technique with a mixed-methods study using empirical quality metrics alongside human-expert judgments [Ise+13]. Six participants carried out model refinement using our technique, and the results were analyzed both with quantitative metrics of topic coherence and uncertainty, as well as manual quality coding carried out by three experts in the domain of the data. All measures showed improvements in topic model quality after several learning iterations.

5.1.1 | Research Questions and Challenges

We present a visual analytics technique tailored to *contrastive* topic model refinement by comparing multiple model configurations' results and parameter impacts. This targeted refinement is designed to enable domain analysts to achieve their goals while remaining model-agnostic. As opposed to the approach presented in the next chapter, this technique aims to facilitate a comparison of model outputs on a semantic level. Thus, we provide users with a direct link between model inputs and outputs, showing semantic contexts for explainability, e.g., through summarization. Our main research question is:

How can we enable domain analysts to refine topic modeling algorithms without understanding their inner workings?

To address this question, we have to tackle two main challenges: The first challenge is to facilitate the *comparability* of the topic modeling results of different model configurations, i.e., the same model with different parameters or different models. Second, on the visualization side, the challenge is to provide the users with *just-in-time* feedback on the impact of their interactions. In addition to these challenges, we also need to map the users' feedback to actionable steps for model refinement, similar to what is discussed in Section 4.1.1. We address these challenges by employing the following two design choices:

(1) Layered Topic Matching Comparing the results of different model configurations is not a straightforward task. Topic models typically produce a list of topics, each with several topic descriptors. Documents are then assigned probabilistically or deterministically to the generated topics. Hence, to compare their results, we need to match the two layers of outputs, looking for common topic descriptors as well as high document overlap. To address this issue, we rely on the Layered Topic Matching Algorithm (LTMA) [El-+18c], which produces two possible matching outputs. Each topic pair can either be a complete match (i.e., high keyword and document overlap) or a similarity-only match (i.e., high keyword overlap but low document overlap). More details on topic matching are described in Section 5.4.1.

(2) Visual Confirmation To enable domain analysts to understand the impact of their document relevance feedback, we implemented some mechanisms for *just-in-time* feedback. After each interaction in the refinement step, the interface provides users with an estimation of the impact of their feedback on the newly calculated modeling output. The visual confirmation is designed as a decision aid and does not recompute the models but instead performs linear boosts and penalties on the existing scores to help users provide targeted document relevance feedback. More details on this mechanism are discussed in Section 5.4.4.

5.1.2 | Contributions

This chapter contributes an iterative topic refinement technique based on contrastive explanation and feedback. This visual analytics technique is tailored to domain analysts and addresses the challenges discussed in Section 3.4.2. In the following, we introduce a human-in-the-loop progressive learning technique for topic model refinement, independent of the specific topic modeling approach. Second, we present four linked task-oriented visualizations to enhance understanding of topic model parameters and provide intuitive feedback about model quality. Lastly, we validate our technique with both empirical and qualitative measures.

5.2 | Designing Visual Analytics Approaches for Domain Analysts

Domain analysts typically have a semantic understanding of their subject area and tasks. Thus, they can identify important keywords and point to relevant pieces of a document. However, as they are neither experts in the domain nor have any detailed understanding of the model's inner workings, we ought to give them an initial representative configuration for the topic modeling output and enable them to compare the results of different settings. Hence, the visual analytics technique for domain analysts focuses on facilitating *contrastive reasoning* and feedback. This chapter instantiates the **Input-Output Space** with a focus on semantics, as described in Section 3.4.2. The following subsections describe the application design space

for a topic model refinement approach tailored to domain analysts, followed by a concrete I&XAI process model depicting a contrastive topic modeling refinement workflow and the corresponding visual analytics pipeline.

5.2.1 | Application Design Space

The design space of this chapter is spanned by the aspects described in Section 3.5. The most important distinction is choosing the stakeholders to be domain analysts (D1). This decision informs some of the other explainability aspects, for example, opting for a model-agnostic (P2), contrastive explanation strategy (D8), with a high abstraction (P4). These aspects are motivated by the design rationales outlined for the **Input-Output Space** in Section 3.4.2. Our approach for designing an I&XAI technique for domain analysts is based on the contrastive interaction paradigm as depicted in Figure 5.2. Here, the analysts are always presented with two model results that compete for their satisfaction. Based on their relevance feedback, the I&XAI interface updates the worst-performing model in a new training cycle (D5) to refine it, while keeping the best-performing model as a visual anchor.

This I&XAI technique reveals more of the model complexity than the previous one (C4). However, it still does not explain the decision-making process of the model (C8), but rather calibrates the output based on the human feedback (C7). This technique allows the user to make decisions by directly comparing the relation of inputs and outputs (C2). To empower the user (C6), two models are presented, allowing for contrastive and relative decision-making (C5). In addition to displaying the changes in word relevance scores with each interaction (C3), the system keeps the model the user preferred anchored between refinement cycles (C1).

Overall, our approach avoids the *cold start problem* [Sch+02] through tailored parameter space analysis. In addition, it facilitates a step-wise understanding of the models' differences and

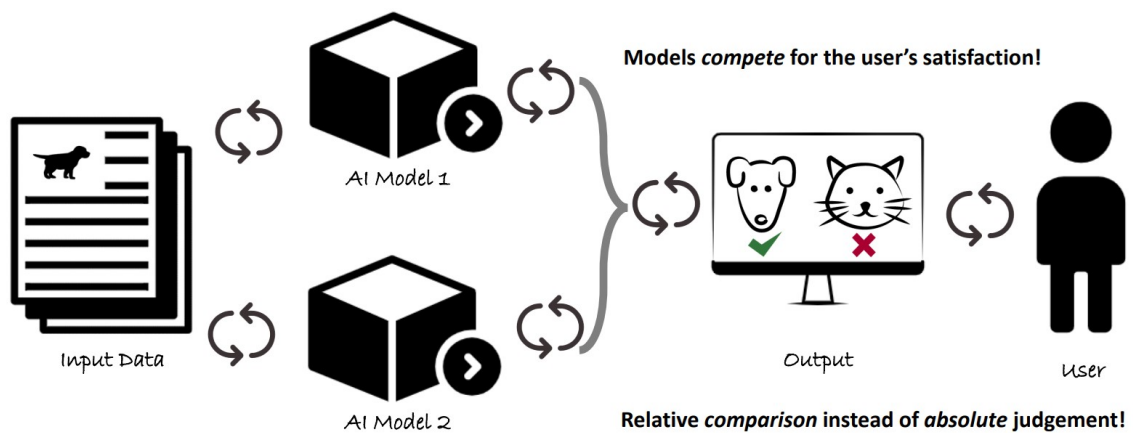


Fig. 5.2: The **contrastive** I&XAI workflow presents the users with a *comparison* between the results of two AI model configurations. This allows for *relative* instead of absolute judgments.

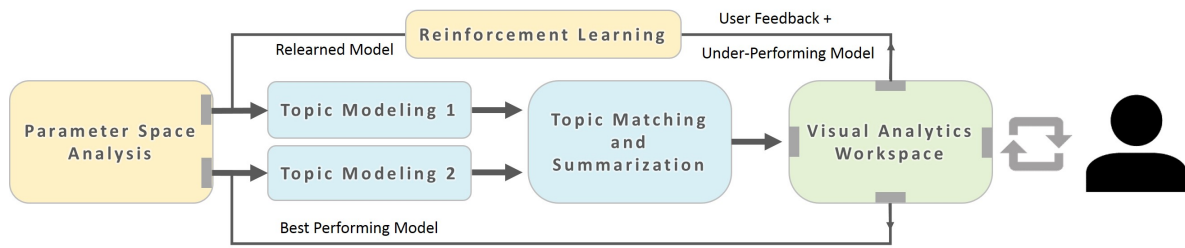


Fig. 5.3: The visual analytics pipeline shows a process model of a contrastive human-AI interaction for content model refinement in the **Input-Output Space** tailored to domain analysts. The progressive learning process includes an initial parameter space analysis and an iterative human-in-the-loop reinforcement learning process in which human annotators compare, evaluate, and optimize models using a visual analytics dashboard.

provides user guidance through quality metrics. Showing topic summary sentences contributes to avoiding harmful rationalizations (C2). And finally, it allows for relative comparisons between models and intuitive relevance feedback to refine the model’s parameters progressively.

5.2.2 | I&XAI Workflow and Visual Analytics Pipeline

We designed this progressive learning process as a mixed-initiative technique tailored to support contrastive analysis and relative feedback based on comparing topic modeling results and parameters. Figure 5.3 depicts the visual analytics pipeline of our approach. It starts with a targeted parameter space analysis for the corpus being modeled to initialize the model selection interface. After the users select two model configurations, both are trained on the given corpus and then matched using the layered topic matching algorithms. In addition, representative summary sentences for each topic are extracted. This information is used to populate the visual analytics workspace. This visual workspace is tailored to the tasks of **[T0] exploring** topic modeling output, **[T1] understanding** the topic descriptors, **[T2] diagnosing** the corpus feature distribution, and **[T3] refining** the topic models through document relevance feedback. Based on the feedback provided by the users, we can start the refinement iterations by retraining the underperforming model using reinforcement learning and keeping the best-performing model unchanged on the interface as a visual anchor for the subsequent refinement round.

5.3 | Parameter Space Analysis

Topic models typically operate in a vector space [SWY75] defined by the accumulated keyword frequency vectors of all documents in the analyzed corpus. These document descriptor vectors are constructed using a bag-of-words model, which weights every keyword in the vector by its overall frequency in the document. These weights can be adjusted by parameters, which are initialized in a preprocessing step. Topic models work best if documents can be associated strongly with one topic, and the topics generated have minimal overlap. To achieve this, we

need to find ways to make documents separable through appropriate parameterization. One common way to use parameters is to provide weights to classes of words (e.g., parts-of-speech). We call these classes of words the *features* and their parameters *feature weights*. For example, a feature weight could be used to down-weight all function words (*stop words*), effectively removing them from consideration in the modeling. Similarly, a feature weight could be used to boost the impact of all proper nouns. Discovering what the appropriate choices of feature weights are, is not well supported by topic modeling toolkits, and the values are often specific to a dataset and very sensitive to change.

In order to start with feature weighting parameters appropriate for the data, we propose a two-step strategy, in compliance with the findings of Sedlmair et al. [Sed+14]. First, we compile an automatic “educated guess” for the data-driven feature selection and weighting, which can then be adjusted by the user. Second, we generate document descriptor vectors using a scoring function selected by the user. Through this process, we configure the initial run of the topic modeling algorithms in our progressive learning process, as shown in Figure 5.3.

5.3.1 | Data-Driven Feature Selection and Weighting

In the context of our parameter space analysis, the problem of data-driven feature selection and weighting is defined over an abstract set of disjunct features $\mathcal{F} = \{\vec{f}_1, \vec{f}_2, \dots, \vec{f}_n\}$ across multiple documents $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$, with vector $\vec{f}_i = \langle w_{i,1}, w_{i,2}, \dots, w_{i,v} \rangle$ representing the complete signature vector of v words w comprising the feature. Every document is defined over the set of all features as $d_j = \{f_{1,j}, f_{2,j}, \dots, f_{n,j}\}$, with $f_{i,j}$ as the concrete feature frequency vector of the document. For example, in our work, we currently consider the following set of features $\mathcal{F} = \{\vec{f}_{nouns}, \vec{f}_{verbs}, \vec{f}_{adverbs}, \vec{f}_{adjectives}, \vec{f}_{propernouns}, \vec{f}_{namedentities}, \vec{f}_{episodes}, \vec{f}_{bigrams}, \vec{f}_{trigrams}, \vec{f}_{stopwords}\}$. Here, $\vec{f}_{nouns}$ is a vector of all *nouns* in the corpus. Hence, in analogy to the bag-of-words representation for documents, we can define every document as a set of frequency vectors given our feature set $\mathcal{F}$. The feature collection we use is driven by norms in topic modeling, but is modular and can be extended to accommodate different tasks or properties of the documents.

In order to select the appropriate features for a given corpus and weight them, we analyze their *discriminateness* which is defined by a non-uniform feature distribution across all documents. We compute this by first calculating the pairwise *feature variability* over all documents for every feature $\vec{f}_i$. We provide five different measures of feature variability in our technique. Each of our feature variability measures take two concrete feature vectors $f_{i,j}$ and output a single scalar. The default measure is a *diversity index* defined by the feature entropy [DL00]. Our experiments with different document collections confirm the finding of Oelke et al. [Oel+14] that entropy is a well-suited measure to enhance document separability. In addition, we provide alternative measures, such as feature vector distances, e.g., cosine similarity and inverse document frequency, as well as a set overlap coefficient [Sin01; MRS08]. The last

alternative measure that is available is RWPDP, a ranked and weighted penalty distance, which we introduce in Section 5.4.1.

Next, given the distribution of feature variability values across all document pairs for each feature f_i , we calculate the standard deviation σ_i of the distribution. Finally, the ratio of σ_i for every feature compared to the minimum σ_{min} across all features is proportional to the ratio of discrimination of these features for the given corpus [AZ12]. The result is that features with more diversity of values across documents (i.e., those that are more discriminative) are scaled to larger values. These ratios thus become the initial feature weights. Based on these measures, we analyze any given dataset and compute a suggestion of discriminative features and their weighting. These suggestions can be used directly or interactively refined by the users.

For some datasets, the discriminativeness of features can overemphasize different aspects of the documents, e.g., the idiosyncratic use of language by different authors or speakers. That is, topics based on these feature weights would separate utterances by speakers rather than by content. This is a common problem that also affects out-of-the-box topic modeling algorithms [Lee+17b]. In order to counteract the oversensitivity of the parameter space analysis towards linguistic nuances, such as writing styles, and to focus on a content-based separation, we introduce a globally learned parameter scale that can be weighted into the individual data-driven weights, as described in Section 5.5.3. This global score captures successful feature weight distributions for different text types, from large corpora. Depending on the analyzed text genre, such normalization can be vital for the topic modeling quality.

5.3.2 | Document Descriptor Vector Generation

Starting with the computed feature weights from the first step of the parameter space analysis, we derive *document descriptor vectors* which assign each word of the document an *importance score*, as described in Section 2.2.3. First, we multiply the concrete feature frequency vectors by the feature weights to obtain a weighted feature vector for every document. These are the default document descriptor vectors, based on the word frequency. However, as Collins et al. state, frequency is not necessarily very effective at scoring key terms of documents [CVW09]. Consequently, we allow users to select an alternative descriptor scoring function that replaces the frequency-based score in order to enhance the vectors' descriptiveness. The currently supported scoring functions include tf-idf [Jon72] and ttf-idf, an adapted version of tf-idf using the total term frequency over all documents, log-likelihood ratio [MS99], and measures based on semantic similarity such as word2vec [Mik+13]. Where the feature weighting step weights features across the corpus (e.g., up-weighting nouns), these scoring functions weight each word for each document (e.g., up-weighting “taxes” as a key term for d_1). After being calculated and normalized to integer values, the document descriptors are then used as input for the topic modeling algorithms.

5.3.3 | Initializing the Workspace

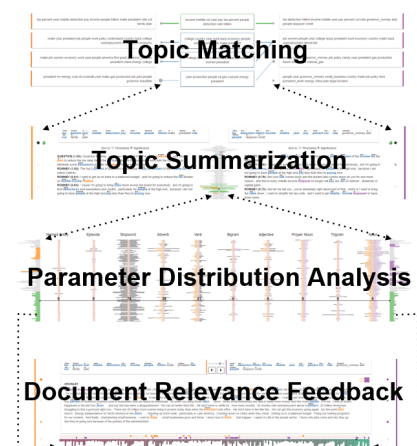
Our learning technique iteratively compares two topic models at a time. Therefore we create two initial parameter configurations through parameter space analysis to compute the two models. These configurations could be exactly the same (for example in order to examine non-deterministic topic modeling stability and robustness) or could consist of different feature distributions or weightings. Since our technique is independent of specific topic modeling algorithms, we allow the users to choose the two models (could be the same model twice) from a set of probabilistic and non-probabilistic models.

5.4 | Visual Analytics Workspace

The core component of our visual analytics technique is the visual analysis workspace. This is the interface in which users interact with the processed data and topic modeling output. We designed the visual interface as a dynamic workspace with consistent visual encoding to facilitate performing the mentally challenging exercise of comparing the different models and their document distributions. One central design consideration for both the usability and aesthetic appeal of the workspace is to use a visual linkage between the different shown components. For example, we always place the two topic models on the two

sides of the screen, referring to them as the *left* and *right* model. In addition, we use a consistent color reference for every model (**orange** for the **left** model and **purple** for the **right** one). The color is also used to indicate similarity, e.g., **blue** is used to refer to **common keywords** and **green** is used to refer to a **document overlap**. A more subtle linkage is achieved by representing all topics consistently as dots and all documents as bars. In addition to linking the visual encoding, we designed the visualization dashboard with stable visual anchors for non-changing components between views. We rely on sweeping animated transitions

between the different views and keep non-changed components anchored to preserve them as reference points for the users' analysis. Users are guided by the consistent layered interaction model, where they *peel off* layers to go deeper into the analysis. At any time users can switch to higher overview layers and go back to pick up their analysis where they left off. This is facilitated by attribute sorting, selections, and filtering which are globally effective across all levels of the visual workspace.



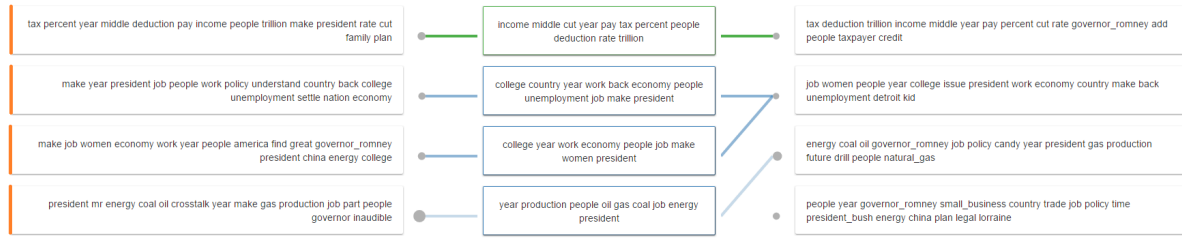


Fig. 5.4: *Topic Matching View.* Two different LDA topic models of the 2nd Obama–Romney presidential debate in 2012 are shown on the left and right sides. Their matches, ordered by decreasing match similarity (min. sim. = 0.7) while minimizing edge crossings, are shown in the middle.

5.4.1 | Topic Matching

As discussed by Alexander and Gleicher [AG16], comparing the results of two topic modeling algorithms by aligning their results is one of the most important tasks to get an overview of the results of topic modeling algorithms. Therefore, the entry point to our visual analysis workspace is the topic matching view, as shown in Figure 5.4. This visualization relies on the output of an automatic topic matching approach [El-+18c] that identifies three levels of topic matches: complete matches, similarity-only matches, and mismatches. These levels are computed based on two criteria: the topic descriptor similarity and the underlying document overlap. Hereby the descriptor similarity between two topic descriptor vectors is computed using the **Ranked and Weighted Penalty Distance** function (RWPD) that introduces a weighted penalty for every keyword that is present in only one of the two descriptor vectors. With an initial distance of 0, the final distance between the descriptor vectors of two topics ($topic_1$ and $topic_2$) is computed as follows:

$$\forall i \in topic_1, \forall j \in topic_2 : \begin{cases} i = j : \frac{|i-j|}{n} \times r \times (w(n, i) + w(n, j)) \\ i \notin topic_2 : w(n, i) \times p \\ j \notin topic_1 : w(n, j) \times p \end{cases}$$

$$w(n, pos) = \begin{cases} n > pos : \sqrt{n} - \sqrt{pos} & ; \\ else : \sqrt{n} - \sqrt{n-1} \end{cases}$$

with p representing the added penalty; n as the minimum vector length of $topic_1$ and $topic_2$; r as the maximum distance range; and i, j as descriptor keywords from $topic_1$ or $topic_2$, respectively. This distance function was developed to mimic the human perception of ranked descriptor vector similarities [El-+18c].

A complete match (represented in green) between two topics fulfills both criteria of the algorithm: it has a high descriptor similarity and a significant document overlap, whereas the document overlap in similarity-only matches (depicted in blue) is not substantial. Mismatches (shown in yellow) are defined by a high document overlap accompanied by a low descriptor similarity. In this view, the similarity threshold for topic matches can be varied interactively by the user, depending on the level of granularity of the analysis. These different levels of matching are used by analysts to identify the similarities between topics on the two considered levels.

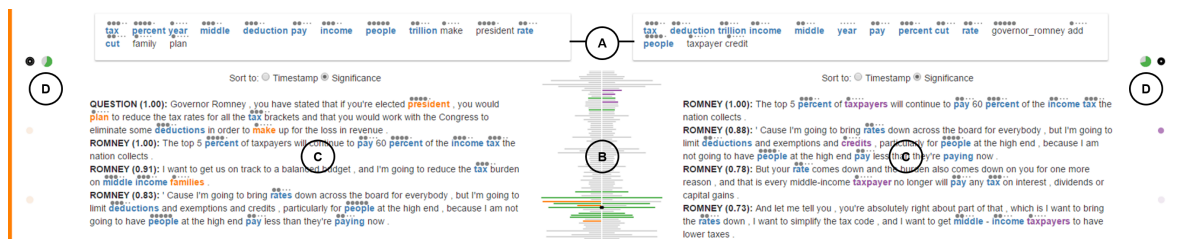


Fig. 5.5: Topic Summarization View. The descriptors of the two compared topics are shown in the top cards on the left and right (A). Descriptors appearing in both topics are colored blue, and those appearing in only one topic are orange or purple, respectively. All keywords are associated with a small glyph above them, showing their relevance score for the topic. The mirrored bar chart in the middle (B) shows all documents in the corpus, where the length of the bars is mapped to the length of the document. Documents belonging to the respective topic of both models are colored green, and documents appearing in only one of both are orange or purple. To the left and the right of the bar chart, the top 10 most representative sentences for the topics are shown (C). The pie charts (D) show the percentage of matching documents.

For example, two topics might share a large number of keywords but no document overlap, revealing a disagreement in the document-topic assignment between the two compared models. In addition to the level of matching, this view also highlights relationships between the topics in both models, such as splitting, merging, matching, and absent topics.

Figure 5.4 shows an example of the topic-matching view. In this view, the left and right topic models are aligned based on their matching scores. Every topic from the two models is represented as a box containing the ranked descriptor vector of that topic. On the inner side of the box is a circle that is scaled to the number of documents assigned to that topic. In the center of the view, the topic matches are shown with their respective color, indicating the matching level (green, blue, yellow) and the matching score which is mapped to opacity (with lower values being more opaque). Green lines indicate a **complete match**, blue lines a **similarity-only match**. The gray points encode the number of documents in a topic. Multiple edges coming to one point indicate, that the respective model combined multiple topics of the other model into one. The topic matches show the ranked set of common topic descriptors. To minimize edge crossings, the position of every topic is determined based on a priority queue that favors larger topics with higher match scores followed by topics that match an already displayed one. By hovering over a topic match, the common keywords of the two matching topics are highlighted in bold and the matching score is shown as a tooltip. In order to inspect a single topic or a pair of matching topics further, the user navigates to the next view by clicking on the object of interest.

5.4.2 | Topic Summarization

The second view of our workspace is the topic summarization view depicted in Figure 5.5. The main purpose of this view is to generate a better understanding of the topic descriptors in order to facilitate the interpretation of a topic. This is done by displaying the most significant

sentences from the documents assigned to each topic as its summary. The number of shown sentences is set to ten by default but can be adjusted by the user. These sentences are chosen to assemble a representative summary of a given topic using a tailored scoring routine. The score for topic t_i and sentence s_j is calculated as follows:

$$score(t_i, s_j) = \frac{\sum_{x \in \{w \mid w \in s_j\}} sig(x)}{\max(\bigcup_{s \in t_i} \sum_{y \in \{w \mid w \in s\}} sig(y))}$$

In other words, for every sentence, the score sums up the significance values of all unique keywords and normalizes them to the highest score of the most representative sentence of the particular topic. Using only unique keywords counteracts potential skewness toward long sentences or repetitive phrases. The keywords considered are all descriptors of the topic at hand and their significance value is given through the topic modeling algorithms and the document descriptor scoring. This scoring function determines a ranking among all sentences attributed to documents that belong to a certain topic. However, to assemble a representative collection of sentences to summarize a topic, we strive to maximize their diversity. This is achieved by introducing a penalty function for the selection of representative sentences to display in the visualization. Given the number of sentences to be shown as summaries, the function penalizes sentences that consist of exactly the same keywords as previously extracted sentences, i.e., for every set of similar sentences only the one with the highest score is displayed. This guarantees the needed diversity within the summaries in order for them to be representative and maximizes the number of topic descriptors shown across the selected sentences.

Figure 5.5 depicts the design of the topic summarization view. To remain consistent with the visual encoding of the workspace, the two topic models are assigned to the left or right side, respectively. Hence, this view shows a mirrored visualization of the two models. Since this view is a direct transition from the topic-matching view, the topics are represented using the same circles (now colored) from the previous view. To ensure linking these circles to the topics shown in the previous view, the topic summarization view is opened using an animated transition that moves the results of the two models to their corresponding sides in a sweeping motion, hiding the labels and keeping the dots. The mirrored visualization of the topic summarization view consists of a central bar chart representing all documents of the corpus, two title panels for displaying descriptors of the topic currently in focus, as well as two central panels for showing the topic summaries. In order to display a topic summary, the corresponding topic has to be pinned (📌, 📌) by selecting its representative circle. Pinning a topic has the effect of keeping them fixed and not updating the elements of the visualization through hovering. When pinned, the descriptions and summaries of the given topics are loaded in their corresponding panels. In addition, the mirrored-bar charts highlight the documents that are assigned to the pinned topic. Upon pinning two topics on either side, the corresponding document overlap is shown in green and their matching descriptors are highlighted in blue. Additionally, small pie charts indicate the model agreement: the fraction of documents shared between topics in relation to the amount of documents assigned to the pinned topics, e.g., 🍷 vs. 🍷 for the left and right topics

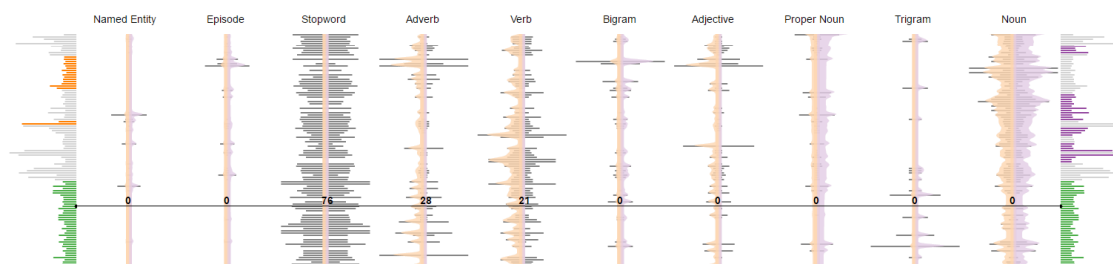


Fig. 5.6: *Parameter Distribution View* using comparative bar charts. This compact visualization technique enhances the comparison of two parameter distributions using mirrored bar charts as a baseline and two asymmetrical violin-style plots as distribution estimates. The plots are scaled using the ratio between the two compared assortments (on both sides). The larger value is scaled to the full width of the baseline and the smaller value is scaled proportionally. This figure depicts the comparison of the utterance descriptor features of the 2nd Obama–Romney presidential debate in 2012. All utterances are sorted according to their topic coherence.

of Figure 5.5, respectively. If the topic pinning is toggled off, the visualization is continuously updated when hovering over a topic circle or a document bar in the central bar chart. This feature becomes useful when exploring all relations to a given entity.

The central bar charts are an essential component of the whole workspace. Apart from the topic-matching view, these charts are used in some form in every visualization to navigate through all the documents of the corpus. Through hovering over a document bar, a position indicator (small black dot) is updated to the document's position and this document is shown in the close-reading view at the bottom of the workspace (not captured in the screenshots). By default, all documents are ordered sequentially according to their order in the corpus. However, to enhance the exploration and understanding of the topic-document relationships, the document bar charts can be reordered globally with respect to a selected measure or the overall length of the documents. For a deeper understanding of the measure and parameter distributions, the user can switch to the next view which is designed for the exploration and analysis of parameter distributions.

5.4.3 | Parameter Distribution Analysis

Figure 5.6 shows an excerpt of the parameter distribution view which uses comparative bar charts to enable the efficient comparison of multiple feature and parameter distributions across the corpus. As mentioned in the previous section, the mirrored bar charts allow navigation through the corpus while highlighting the document-topic assignments. Hence, after understanding the topic compositions, the user can further dive into the investigation of the documents by switching over to this view. An animated transition splits the topic summarization view along the central mirrored bar chart, moving the components to the left and right edge of the screen to peel off another layer. This parameter distribution view goes deeper into the structure of the corpus and allows the exploration of patterns across all document features and parameters. In keeping up this metaphor of a layered analysis, our visual analytics workspace

allows users to go up to any overview visualization at any time, then switch back to continue where they left off at a deeper layer.

To enhance the comparability of features across the corpus, we designed the comparative bar charts. This visualization technique is simple, yet has proven to be effective for the comparative analysis of two ordered distributions with an underlying baseline. As shown in the side figure, we display the baseline distribution using mirrored dark bars in the background to be a constant reference for comparison. On top, the ratio of the two compared distributions dictates the interpolation range of the violin plots that are spanned asymmetrically on either side of the symmetry axis. Hence, for two given parameter values for each of the topic models, we calculate their proportion and use the larger ratio as the maximum for the normalization of the opaque violin plots. Consequentially, the smaller value will not cover the baseline bars, leaving the size of the peaking-out bars as an indicator of the relative difference between the two parameter values.



Using this visualization technique, we can arrange all relevant parameters and features for comparison, as shown in the side figure. The document bar charts corresponding to the two topic models are situated on either side of the parameter distribution plots. Similar to the previous view, the document bar charts are used to navigate through the corpus. By hovering over a document bar, the close-reading view (not shown in the figure) is updated, as well as the navigation line which shows the concrete values of the baseline distribution for the particular document. These values could be chosen to represent the absolute frequencies of the features for every document. However, by default, they show the number of occurrences of a feature divided by the total number of words per document. This default value is chosen to emphasize the importance of that feature for the classification of the document. For example, if a ten-word document contains five adjectives, then varying the weight of adjectives will significantly impact this particular document in contrast to a longer document with the same amount of adjectives. In addition to linking and brushing, this visualization supports pinning topics and sorting all documents according to any feature or parameter. This becomes especially useful when choosing which documents to inspect further in the next view.

5.4.4 | Document Relevance Feedback

This view is the main interface for decision-making in order to adapt the topic modeling results in the next cycle of the progressive learning process, as described in Section 5.5. While the previous three visualizations were focused on understanding and comparing the two topic modeling results, the document relevance feedback view requires the users to actively vote for the most suitable model using their acquired knowledge. This is done using the interface shown in Figure 5.7. The document relevance feedback view is always present in a minimized form at the bottom of the workspace, serving as an interactive close-reading view.

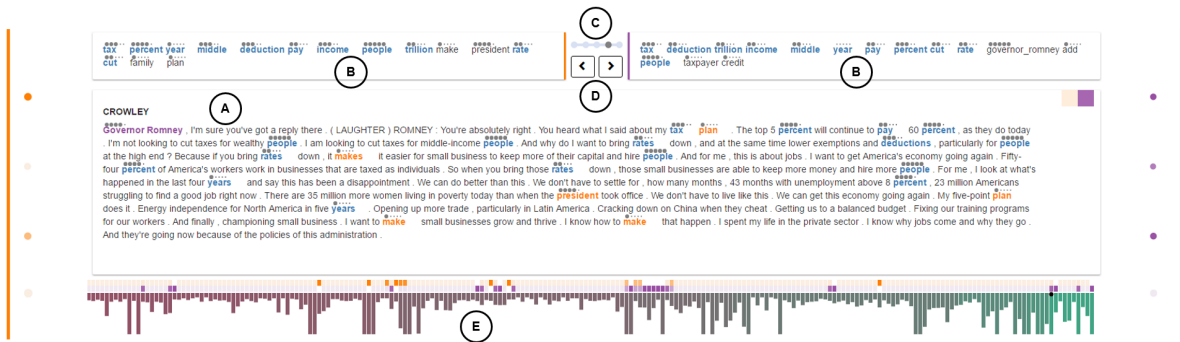


Fig. 5.7: The *Document Relevance Feedback View*. (A) The document in review; (B) the topic descriptors of the associated topics; (C) the decision slider; (D) the navigation arrows. The bar chart (E) shows the documents sorted and color-coded by their topic coherence, from bad (red) to good (green). Users rate topics for the current document by selecting a slider position, and can navigate between documents with the arrow buttons (D).

To activate the functionalities for the relevance feedback, the minimized close-reading panel is transitioned from the bottom of the workspace to the center, preserving the visual linkage of the anchored topics. When extended, this view is accompanied by a *decision-slider* between the two anchored topic-label boxes, along with a horizontal document bar chart which is sorted and colored according to the selected global sorting measure. In the center of this visualization is the currently selected document, with its keywords highlighted. In order to start the relevance feedback, the user selects a document from the horizontal bar chart (or by navigating to the previous/next document using the arrow buttons) and moves the decision-slider to the topic that yields a better description of the selected document. Our technique is designed to accommodate individual optimization strategies that depend on the user's analysis goal, time budget, expertise and familiarity with the analyzed corpus, and the noisiness of the document collection. However, in order to assist the users in choosing representative documents for the relevance feedback task, we rely on two quality metrics, similar to the ones introduced in Section 2.3.3.

Document Quality Metrics In order to minimize the manual effort of the users, we introduce two quality metrics with the intention to direct users to the documents which will be most affected by their decisions. Given a document and its top topics from each model, these metrics assess the controversy among the two topics and between the document and the topics, respectively. The first metric is topic coherence, which describes the agreement between two topics assigned to a document. The second metric is topic certainty, which measures the compatibility of the two topics with the given document. We developed both metrics to comply with the human perception of similarity between two descriptor vectors and thus utilize the ranked and weighted penalty distance function (RWPD, introduced in Section 5.4.1) for the calculation of the measurements. Hence the two quality metrics are defined for every document as follows: $TopicCoherence = RWPD(topic_1, topic_2)$; $TopicCertainty = avg(RWPD(topic_1, doc), RWPD(topic_2, doc))$.

Decision-Making The decision-making process is guided by two basic questions that the user has to keep in mind. The central question for the document relevance feedback is *which topic describes the current document better?* Beyond the scope of a single document, the user can further consider the question: *Would you like to endorse or eliminate single descriptors to guide the topic modeling?* These questions are addressed by two interactions: the decision-slider and the single-word boosts. These interactions are directly translated to actions for the refinement and learning of a new model in the next cycle and are accompanied by an immediate visual confirmation to manifest the changes caused by the users' decision.

The decision-slider simplifies the model steering to a binary decision between two models for every document. To accommodate the probabilistic nature of some topic modeling algorithms, this binary decision is extended to be based on the assigned topic probabilities. Regardless of the reviewing scheme, the user is only required to make singular decisions for one document at a time. The reinforcement learning then generalizes from the users' verdict on a sample of documents to the overall corpus. Therefore, it is essential to point the users to representative documents to consider for review. Our proposed optimization strategy is to examine outliers and documents with a high level of controversy, i.e., documents at the lower end of the quality metric scale introduced in the previous section. Accordingly, a time-efficient and successful optimization strategy we observed, is to order all documents based on their certainty or coherence scores and examine longer documents with a low score for these measures.

When analyzing a document, a decision is made by moving the slider toward the better-suited topic on a discreet five-point scale. If both topic models have an equal quality regarding the analyzed document, the slider can be kept in the middle, in order to remain neutral. Moving the slider towards a topic implies a keyword boost in favor of that topic. As described in Section 5.5.1 in more detail, the boosting differentiates between three keyword types: matching keywords from both topics; keywords contained only in the favored topic; and keywords contained only in the rejected topic. The matching keywords are positively boosted (by a factor of two) since they are clearly agreed upon by the two models. The undesirable keywords from the rejected topic get a negative boost (by a factor of one), while the favored keywords from the endorsed topic receive the highest positive boost (by a factor of three). In addition to the keyword types, the weighting of the slider amplifies the keyword boosting in the case of a polarizing decision.

Using the decision-slider affects the topic descriptors defined by the two models and leverages the users' preferred keyword compositions to sustain them. However, in order to accelerate the topic convergence, users might want to globally sanction or endorse single keywords. This is achieved through single-word boosts and penalties. Such functionality gives the users a higher degree of freedom to optimize the topic modeling. However, by considering the significance of keywords when applying the boosts, our system counteracts the potential overfitting of the descriptors. In contrast to the decision-slider, single-word boosts and penalties can be applied to any word (or n-gram) in the corpus. Therefore, besides spurring the topic convergence, this

functionality is also extensively used to remove undesirable and nonsensical words from the topic descriptors in further optimization cycles.

Visual Confirmation Since the interactions in the decision-making phase have a wide-ranging implications on the quality of the relearned model of the next processing cycle, we incorporated immediate, responsive visual feedback for the actions performed in that step. On the one hand, we track the movement of the decision-slider for every document and represent its assigned topic preference using a small icon on the top right corner of the text panel, as well as in the horizontal document bar chart on the bottom, as depicted in Figure 5.7. These icons use the two topic colors to highlight the favored model for every document, in addition to showing the certainty of the user's decision using opacity, e.g., . Additionally, we highlight documents that have been visited by the user  (without scoring). On the other hand, we show the scoring of every keyword through word-size glyphs, e.g., , , and . The glyph consists of five dots that represent the binned score of the keywords, e.g.,  for a score of 2/5. This glyph is immediately updated when changing the slider to indicate the effect of the change on the keyword scoring. The score is determined by the keyword significance for the current topic or document, combined with the overall score of that keyword (which gets affected by the boosts). Having instant visual confirmations had a positive impact on the usability and understandability of the visualization.

5.5 | Task-Driven Topic Convergence

As shown in Figure 5.3, the users' input from the Visual Analytics Workspace is used to update the inputs before starting a new topic modeling cycle. By endorsing topic descriptors or single keywords, the users are able to steer the topic modeling in order to converge toward a more intuitive and understandable topic modeling result. As described in more detail in the following sections, this goal is achieved through reinforcement learning in iterative optimization cycles.

5.5.1 | Reinforcement Learning

Using the relevance feedback described in Section 5.4.4 we constantly learn and update feature weights throughout each cycle. Whenever the users rate the topic models with respect to a given document by moving the decision-slider, we update our data structures in the background in order to prepare for the next cycle. The resulting changes are immediately presented to the users and enable them to quickly understand the impact of their actions on the topic models, despite these changes only affecting the new topic modeling cycle. Before the first cycle begins, each topic keyword in the corpus is assigned an initial score. While this value is equal for all keywords in the current implementation, it can be adjusted by the users. However, it is important that this score be larger than zero, as the scores of the document descriptors are adjusted proportionally to this value before the topic modeling algorithms are instantiated.

Each update of the decision-slider is reflected in the scores of the affected keywords: topic descriptors of the rejected topic model are penalized, while topic descriptors of the favored model are boosted by three times the value of the penalty. Keywords appearing in both sets of topic descriptors are boosted and penalized at the same time, resulting in a boost of two times the penalty. These ratios for boosts and penalties have yielded promising results in our experiments but could be easily adjusted by the users to quicken or repress the learning rate. The decision-slider has five possible positions to allow users to show a strong or subtle preference for one topic model over the other. The left- and rightmost positions of the slider correspond to a strong preference for the left or right model, respectively, and lead to boosts and penalties twice as high as the values for the more subtle preferences mentioned above. In the middle position, the scores are not updated.

While boosting and penalizing existing topic descriptors is already a very powerful tool, it is not always sufficient. An additional feature that has been often used during our user studies is the option to promote arbitrary words from any document that have not yet been recognized as a good topic descriptor by the current models. In the same way, (key)words can be penalized to ensure that they will not be part of the set of topic descriptors in a future run. These *single-word boosts* give the users a very direct way of incorporating their domain knowledge and correcting inherent biases of the topic models, which would otherwise be very hard to compensate for. When promoting single words, we first make sure to boost them to the base score associated with topic keywords before adding another boost of three times the maximum boost that can be achieved by one slider movement.

Between the runs of two cycles, we use reinforcement learning to update all parameters, as described in more detail in Section 5.5.2. While in the current implementation, the learning rate is fixed to a constant 40%, our technique can easily accommodate more complex and sophisticated learning strategies. For example, the “Win or Learn Fast” principle has previously been successfully applied in user-driven topic modelings by Tripolitakis et al. [TC16]. If the results are positively rated by their users, they reduce the learning rate and drastically increase it whenever performance worsens. Additional random changes in the parameters occurring with very low probability help to escape local maxima.

5.5.2 | Iterative Topic Evolution

To ensure the robustness of our progressive learning process and to guide the users through the optimization space, our technique is designed to relearn only the under-performing topic model while keeping the better model as an anchor for the next cycle. When the users decide to finish a cycle and restart the processing loop, we assess all slider positions chosen by the users in order to determine which topic model has—according to the user—performed better on the given corpus. For example, assuming the overall average slider position was on the left, meaning the users preferred the left topic model, we keep this model as a baseline for the next

cycle and recompute the right one. Before restarting the topic modeling, we adjust the input by applying the boosts computed in Section 5.5.1 to all words of all documents. This is done using the document descriptor generator component (Section 5.3.2) by directly adjusting the computed frequencies of every word, or repeating it in the input text before re-applying the scoring functions, which has the equivalent effect. Words that have been penalized a lot and, as a result, are associated with a zero or even negative score, will be removed by the document descriptor generator and are not taken into account for the topic models of the new cycle.

In addition to boosting keywords, we also update the feature weights (Section 5.3.1) for the under-performing model, i.e., the influence of word classes on the topic model. After each update of the decision-slider through the users, we retrieve the associated features of the topic descriptors in order to update the ratio of boosted features. This is done independently for the two active topic models. Before we start a new cycle, we collect the ratios of feature weights of the better model and use them to update the feature weights of the underperforming model. This update happens with a user-adjustable learning quota that is initially set to 40% to ensure fast convergence after a limited number of cycles. Between two cycles, all collected data, as well as key metrics, such as topic coherence and certainty scores, topic assignments on a per-document basis, and the current corpus keyword scores, are persisted in a log file. Such detailed data collection enables the evaluation of the topic model at a later stage and makes provenance tracking possible. Additionally, this data is of further interest to users, as expressed by a political scientist during our user study, who would like to further analyze the details of the topic development between cycles.

5.5.3 | Global Learning of Parameter Scales

One disadvantage of the “educated guess” for feature weights as introduced in Section 5.3 is the fact that they are initially extracted on a per-corpus base. As a result, they are relatively susceptible to changes in linguistic nuances, such as writing styles, that are specific to a given corpus. We counteract this bias towards certain features by introducing a set of globally trained feature weights. These global weights are automatically updated after a successful set of training cycles has been finished by the users. In the current implementation, the global feature weights are updated using a relatively low learning rate.

Before starting the first cycle the users are presented with the feature weights that have been automatically extracted for their given corpus. They can then either use them as is or leverage their domain knowledge to decide that they are a non-optimal fit. In this case, they can either manually adjust the weights, or mix in the global feature weights by a definable percentage. This leads the descriptor extractor to emphasize certain features and consequently increases the explainability and quality of the model. Additionally, it enables users to start the feedback cycle without having to pre-process their corpus, by only using the global feature weights.

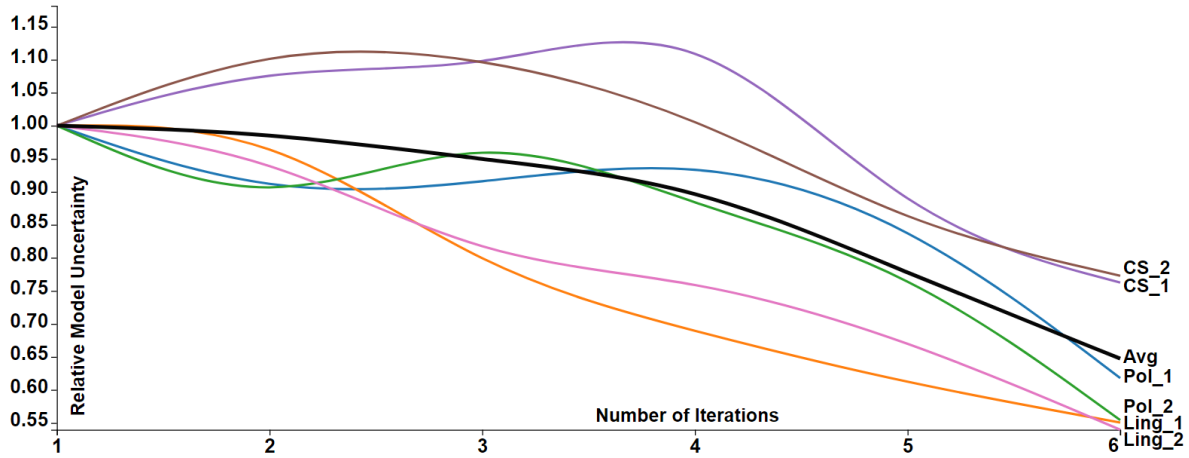


Fig. 5.8: The *Automatic Model Uncertainty* shows the positive development of the topic assignments, especially when trained by experts. Values at iteration six: $CS_1 = 0.76$; $CS_2 = 0.77$; $Pol_1 = 0.61$; $Pol_2 = 0.55$; $Ling_1 = 0.55$; $Ling_2 = 0.53$; $Avg = 0.64$.

Once the users see the current better-performing topic model as a good fit for their use case, they might decide to finish the feedback loop. They are then given the option to update the previously learned global feature weights with the ones they just trained for the better-performing model. In case they decide to do so, the global weights are adjusted for the newly learned weights with a user-definable learning quota.

This concept could be extended in order to maintain multiple “global” feature weights for different classes of documents. In this case, the users would have to assign one or multiple classes to their current corpus, before updating the respective weights. This would be particularly interesting for users working on different document classes, such as political speeches, news stories, and books. We also plan to globally learn the scores associated with keywords described in Section 5.5.1 instead of restarting that process for every corpus. Guided by updates with a user-definable learning quota we plan to move away from uniform starting scores for all keywords, instead utilizing the results from previous runs as an improved starting point. Such a feature would be helpful for users analyzing similar corpora, e.g., multiple presidential debates, where otherwise the first runs will always be spent on retraining the same keyword scores.

5.6 | Evaluation

Due to the modularity of our technique and the subjectivity of the interpretation of topic models, we chose to evaluate our technique with a mixed-methods study, as advised by Isenberg et al. [Ise+13]. We empirically measured the effectiveness of the progressive learning process with automatic metrics, as well as a manual assessment of model quality by expert annotators. We also gathered qualitative feedback on the usability of our visual analysis workspace.

Participant	Topic Precision Improvement	Automatic Model Certainty Improvement	Topic	Topic Precision Improvement
Pol ₁	33.3%	38.2%	Taxes	33.3%
Pol ₂	25.0%	44.6%	Unemployment	5.6%
CS ₁	38.3%	23.8%	Education	13.9%
CS ₂	33.3%	22.7%	Gun Law	19.4%
Ling ₁	45.0%	45.0%	Energy	58.3%
Ling ₂	48.3%	46.0%	Women's Rights	25.0%
Avg	37.2%	35.3%	Immigration	36.1%
			Attack in Libya	72.2%
			Crosstalk	8.3%
			No Topic	-100.0%
			Avg	37.2%

(a) Average improvements between the 1st and last optimization cycles. (b) Avg. improvements per topic.

Table 5.1: User study results indicating a clear improvement for all participants and topics. These are based on an empirical analysis of the logged data.

Dataset To choose an appropriate dataset, we envisioned a corpus that fulfills three criteria. First, we wanted broadly-familiar content to ensure understanding by participants and annotators. Second, we sought document collections with shorter documents in order to fit multiple optimization cycles into a two-hour session (as the reinforcement learning scales in time with the corpus length). Third, in order to empirically validate the results of the study, we wanted a corpus with a known topic distribution as a gold standard. Thus we chose to use a presidential debate for the study, specifically the second US presidential debate between Romney and Obama in 2012. This debate discussed known domestic affairs, has been widely studied and the topics are accessible to a non-expert reader. And lastly, it fulfills the final criterion if we consider the document granularity on an utterance level.

Controls In order to control our study to focus on model improvement through the iterative learning process, we controlled for the topic modeling algorithm and the initial parameter settings and model across all participants. We chose to run the study based on LDA, as it is the most common baseline across the literature. We initialized both topic models to LDA with 9 topics (determined during pilot testing), as using the same model on both sides leads to more predictable behavior. We initialized one with a feature weighting based on entropy and the other with a manually selected feature weighting expected to be helpful for this dataset (to simulate manual tuning of features, which is the normal process without our technique)¹.

Method and Participants Before conducting our formal study, we ran a pilot study with three graduate students to test all conditions and refine the usability based on their feedback. Our study was broken into two tasks, each completed by different participants. In the first stage, the *Model Improvement Task*, we conducted six two-hour-long sessions with two experts each from political science, linguistics, and computer science. Participants in this phase had varied experiences with topic modeling, from novice (computer scientists who had embedded topic models in tools but not tuned them) to expert (political scientists who use topic models and manually tune them). In the absence of a standard benchmark, we created a second stage, the *Model Assessment Task*, where we evaluated the outputs of the first stage using both automatic

¹verbs, adj, adv, and stopwords = 0 – everything else = 100

quantitative metrics and manual quality coding carried out by three annotators from linguistics, knowledgeable on topic modeling and trained to recognize word relations.

Model Improvement Task This task centered around participants using the technique to perform an iterative optimization of the given dataset and models. Each study session was divided into three parts. In the first 30 minutes, we started by asking the participants about their experiences with topic modeling, then continued explaining the learning process and the visual interface design. This was followed by a brief initial feedback round to gather first impressions. In the second part (1 hour), we asked participants to use the visual interface to optimize the topic modeling results of the presidential debate in an iterative cycle. They could use all interface features and execute as many refinement cycles as they wanted. During this phase, we asked participants to ‘think aloud’ as they worked with the interface, describing their choices and any usability problems they encountered. We also collected interaction logs, screen capture videos, and the topic models generated for each optimization round. In the final part of the study, we collected their feedback on the performance of our technique and their satisfaction with the results, considering the time spent on optimization.

Model Assessment Task In this phase, we invited annotators to assess the quality of the topic models generated in the first phase. After a brief introduction to the study, annotators were given 12 worksheets, each listing 9 sets of topic keywords in rank order. Each worksheet represented the output of either an initial or final topic model from phase one (1 initial, 1 final, for each of the 6 phase one participants). The worksheets were not labeled and were presented in random order. Annotators were also given a selection of *gold standard* topics widely accepted to be represented in this dataset. The annotation task was to examine each topic and determine the best matching gold standard topic, and give a rating from 0–4 of the topic match to the standard. Next, the annotator circled all words in the topic descriptor vector which did not fit the gold standard. This process was repeated for all topics and topic models (9 topics X 12 topic models X 3 annotators = 324 ratings). Due to the demanding nature of the task, we did not annotate intermediate topic model states. The measures collected in this phase were topic model precision change, average irrelevant words change, and inter-annotator agreement.

5.6.1 | Quantitative Results: Model Improvement

We calculated the change in model certainty using the automatic certainty score described in Section 5.4.4. The results, charted across all iterations of reinforcement learning, are shown in Figure 5.8. The average uncertainty (black line) decreased monotonically through each iteration. The uncertainty improvement achieved by participants varied, but for all of them, the overall model uncertainty was lower at the end of the study than at the beginning (average 35.3%). That is, the models consistently improved through the learning iterations. While

there are too few participants to confirm a trend, we did observe a better improvement from participants who had more understanding of topic modeling before the experiment.

Expert scoring of topic precision also showed an improvement across all phase-one participants and all topics; the average before optimization was 2.31 ($\sigma = 0.46$), and after the last cycle, 3.68 ($\sigma = 0.67$) on a [0,4] scale, with an inter-annotator agreement on topic precision of 88.3%. For comparison across automatic and manual measures, Table 5.1b contains both metrics in a percent scale and reveals a strong agreement between them. The precision improvement by topic is also shown. The number of irrelevant words per topic decreased by 7.2% on average between the initial and final models, further indicating model improvement.

Due to the probabilistic nature of the topic modeling algorithm used in the study and the lack of an annotated gold-standard dataset on the granularity of the examined documents, we can only make a reliable conclusion using a relative scale to achieve a stable baseline for measuring the effect of topic improvement over time. Hence, given these starting conditions, the decrease in the model uncertainty is statistically significant, however, due to the unavoidable variance in our baseline, we do not have the experimental power to make a statistically-reliable claim. Therefore, we present our quantitative results as relative changes over time indicating a trend of substantial model improvement over all optimization cycles. This trend was confirmed in our qualitative results.

5.6.2 | Qualitative Results: Expert Feedback

Initial Feedback Regardless of their expertise in using topic models, all participants of the study saw an immediate benefit in having such a visual analytics process. Political scientists, who had more experience with topic models, reported that they often spend hours in a trial-and-error cycle to get satisfying results. When asked about their usage of topic models, computer scientists reported that they heavily relied on the automatic output of topic models to embed topics in their tools, not considering model uncertainties or fitness to the data. However, all participants uniformly agreed that one of their major concerns with topic models is the reliability of the outcome and trustworthiness of the black box. Especially the linguists were mistrustful. One of them commented that she is unaware of successful optimization strategies that would help her validate topic modeling results using her data.

Visualization Design and Usability Participants appreciated the visual anchoring we employ throughout the different stages as it gave fixed points to concentrate on and helped in guiding them through the process. Especially in combination with the layered analysis allowing them to work on a higher level, or get more detailed information on demand, it enhanced the orientation during the visual analysis. They especially liked the steadily-visible, interactive close-reading panel on the bottom of the view for keeping the analysis in context. One feature that was extensively used by all participants of the study was the option to boost or penalize

individual words as a form of direct relevance feedback to influence the topic model results toward being more intuitive. This confirms the preferred feedback mechanisms discovered by Lee et al. [Lee+17b]. It was especially useful to boost words that had not yet been recognized as good topic descriptors by any of the models.

Most users noted the steep learning curve due to the diverse functionality, the number of included visualizations and their rich set of interactions, and the density of the available information. However, all participants were able to achieve proficiency with the tool over the course of the study session. One political scientist (Pol_1) commented that “*in order to have such an expressive visualization dashboard for the analysis [he is] willing to take into account learning to use a new system.*” He added, “*if we establish such a technique as a norm for that analysis and use of topic models [he expects] that our visual workspace will be improved and extended by the demands that would arise from a community of active users*”.

Although appreciating the degree of freedom in the analysis and the serendipity of individual optimization strategies, one of the computer scientists (CS_1) suggested incorporating an option for more guided optimizations by adding explicit system recommendations. This trade-off between serendipitous and guided discovery for the analysis of topic models has been exploited by Alexander et al. [Ale+14], who argue for a more open and serendipitous analysis and exploration process.

General Assessment After using the tool for a while the users reported they were *unobtrusively* learning more about the dataset due to the effectiveness of the workspace design and the richness of on-demand information. User POL_1 positively commented that he “*could spend hours exploring a dataset with that interface*”.

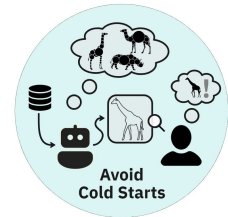
For the users, it was intuitive to keep the better-performing topic model while restarting the one that had underperformed. It gave them an easy way to track the changes introduced in the new cycle. We noticed that most of the time users preferred the relearned model over the best-performing one of the previous cycles, indicating that they immediately noticed an improvement in the topic model with every iteration. All users predicted that they would expect even better results with some additional optimization cycles. They also noted that they deemed the trade-off between the additional time needed to complete a new cycle and the resulting benefit justifiable and reported that they would use the tool regularly on their data.

5.7 | Discussion: Investigating I&XAI Design Patterns

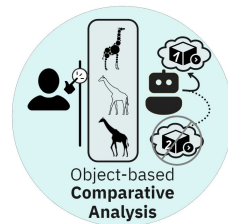
Based on the presented technique, in this section, we discuss three potentially generalizable I&XAI design patterns. These are tailored to interactive and explainable applications for the

domain analysts in the **Input-Output Space** with a focus on the domain semantics. An overview of all design patterns discussed in this dissertation can be found in Section 8.4

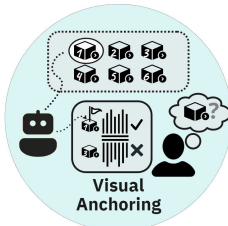
(2a) Cold Start Avoidance Analysts focusing on solving a domain problem with concrete data and tasks usually do not have the time or expertise to configure a suitable model parameterization to perform their analysis. Hence, it is desirable when designing applications for this user group to apply a *parameter space analysis* on the input data and set reasonable defaults for the analysis context.



(2b) Object-based Comparative Analysis To facilitate the effectiveness of the relevance feedback provided by analysts, we aim to elicit comparative feedback on the output suggestions they are shown. Allowing for such a comparative analysis enables the users to contribute *relative judgments instead of absolute ones*. This is an easier cognitive task and is less error-prone, especially for non-experts.



(2c) Visual Anchoring of Favored Solutions Designing responsive interfaces that show a visual change when users' feedback is a good practice to retain their focus and avoid confusion. Especially for progressive model refinement, it is imperative to keep the model configurations the users favored anchored on the interface while retraining underperforming models. This design also minimizes analysis context shifts.



In addition to these three design patterns, in this chapter, we also discussed the importance of a layered analysis with a stepwise increase in the explanation complexity and the level of detail. Lastly, we also showcased the importance of using summarization to describe topics, avoiding potentially harmful rationalizations.

Input-Output Space

Model Decisions

” *By far the greatest danger of Artificial Intelligence
is that people conclude too early that they understand it.*

— Eliezer Yudkowsky
in [Yud08]

Modeling Reply Chains to Untangle Conversational Text

” If understanding language [...] through statistical analysis does not count as true understanding, then humans have no understanding either.

— Ray Kurzweil

How to Create a Mind [Kur13] p. 157

Summary

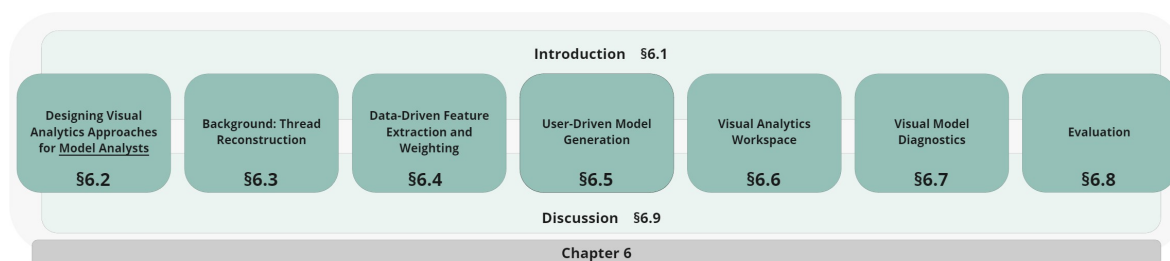
For the **model analysts**, we present *ThreadReconstructor*, a technique designed in the **Input-Output Space** (with a focus on the *model*). It supports a collaborative human-machine decision-making process during the training of content models for detecting and analyzing the implicit conversational structure of discussions. We combine supervised and unsupervised machine learning models to generate a basic structure that is enriched by user-defined queries and rule-based heuristics. Depending on the data and tasks, users can modify and create various reconstruction models that are presented and compared in the visualization interface. Our tool enables the exploration of the generated threaded structures and the analysis of the untangled reply chains, comparing different models and their agreement. To understand the inner workings of the models, we visualize their *decision spaces*, including all considered candidate relations.

Contributions

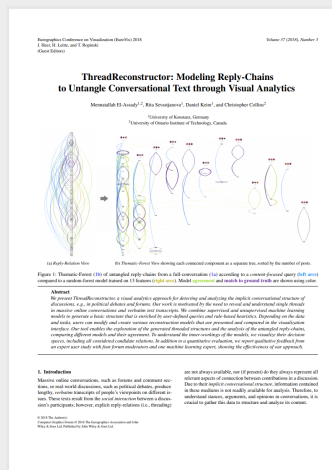
- (i) An I&XAI model refinement workflow tailored to model analysts in Section 6.2.
- (ii) An thread reconstruction technique for task-driven *feature-selection* and *interactive model-generation* in Section 6.4 and Section 6.5.
- (iii) An interactive visual workspace for thread reconstruction *model generation and comparison* in Section 6.6 and Section 6.7.
- (iv) An expert user study confirming the applicability of our approach in Section 6.8.

Contents

6.1	Introduction	142
6.1.1	Research Questions and Challenges	143
6.1.2	Contributions	145
6.2	Designing Visual Analytics Approaches for Model Analysts	145
6.2.1	Application Design Space	145
6.2.2	I&XAI Workflow and Visual Analytics Pipeline	146
6.3	Background: Thread Reconstruction	147
6.4	Data-Driven Feature Extraction and Weighting	149
6.4.1	Reply Relation Features	149
6.4.2	Message Categorization	150
6.4.3	Data-Driven Feature Weighting	150
6.5	User-Driven Model Generation	151
6.5.1	Classifier Model	151
6.5.2	User-Defined Queries	151
6.5.3	Rule-Based Heuristics	152
6.6	Visual Analytics Workspace	152
6.6.1	Text-Level View	153
6.6.2	Reply Relation View	154
6.6.3	Thematic Forest View	155
6.7	Visual Model Diagnostics	156
6.7.1	Parent-Candidate View	156
6.7.2	Model Optimization and Transfer	158
6.8	Evaluation	158
6.8.1	Scalability and Model Reliability	158
6.8.2	Expert User Studies	160
6.8.3	Case Studies	161
6.8.4	Feedback	165
6.8.5	Lessons Learned	167
6.9	Discussion: Investigating I&XAI Design Patterns	168



This chapter is based on the following publication:



ThreadReconstructor: Modeling Reply Chains to Untangle Conversational Text through Visual Analytics

Mennatallah El-Assady, Rita Sevastjanova, Daniel Keim, and Christopher Collins

Computer Graphics Forum, 37(3):351–365, 2018.

El-Assady et al. [El+ 18a]

Contribution Clarification

This chapter is based on the publication entitled “*ThreadReconstructor: Modeling Reply Chains to Untangle Conversational Text through Visual Analytics*” [El+ 18a]. The system presented in this paper is an extension of Rita Sevastjanova’s master’s project, which I closely supervised. The user study was conducted and analyzed by me while on a research internship, and the paper writing was conducted by me with only slight edits from my co-authors.

6.1 | Introduction

This chapter is motivated by the need for semi-automatic thread reconstruction through content modeling. Model analysts, e.g., forum moderators and political scientists, usually have to extract conversational structures in *verbatim* transcripts of people’s viewpoints on different issues. Such transcripts are found, for example, in massive online conversations, such as forums and comment sections, or real-world discussions, such as political debates. These texts result from the *social interaction* between a discussion’s participants; however, explicit reply relations (i.e., threading) are not always available, nor (if present) do they always represent all relevant aspects of the connection between contributions in a discussion. Due to their *implicit conversational structure*, the information contained in these mediums is not readily available for analysis. Therefore, to understand stances, arguments, and opinions in conversations, it is crucial to gather this data to structure and analyze its content.

To counteract this, some forums, e.g., Reddit [Red17a], provide their users with the option to create a nested reply, producing threaded conversations. However, not many forums maintain such a logical reply structure as a publicly accessible interface. Most commonly, only the temporally-ordered sequence of posts is provided for usage and further analysis. Hence, to observe some existing patterns in the data, analysts rely on automatic techniques to reconstruct the reply relation structure or manually go through the whole dataset. Moreover, even when the threaded structure is given, analysts often observe multiple intertwined discussions in supposedly coherent threads. This is due to drifts into side discussions, participants not making use of the provided “reply” functionality, or posts not strictly replying to previous messages but rather generally commenting on all previous text. In addition, different reply structures might be valid, depending on the semantic context of the analysis.

Model analysts often have to go through such lengthy transcripts on a daily basis. In our interviews with professional forum moderators, we learned that they sometimes have to resolve legal or policy issues arising from users’ misconduct. Following the flow of a discussion to understand all the relevant exchange of opinions and information is crucial to fulfilling such a task. They sometimes need to ensure that they have captured all related reply chains – favoring recall over precision. While for other tasks, they are usually interested in precisely reconstructing reply chains with high accuracy. Forum moderators, therefore, describe this as being one of the most time-consuming parts of their job. On the other hand, our political scientist colleagues use statistical models to test different theories of communication and argumentation in real-world debates [El+17a]. These models may be improved by automatically separating debates into thematically coherent threads.

Facing these challenges, we identified five requirements for an effective solution, namely; (1) Accurate Reconstruction (supporting different analytical tasks); (2) Untangling Conversational Threads (into distinct connected components); (3) Understanding Relations (e.g.,

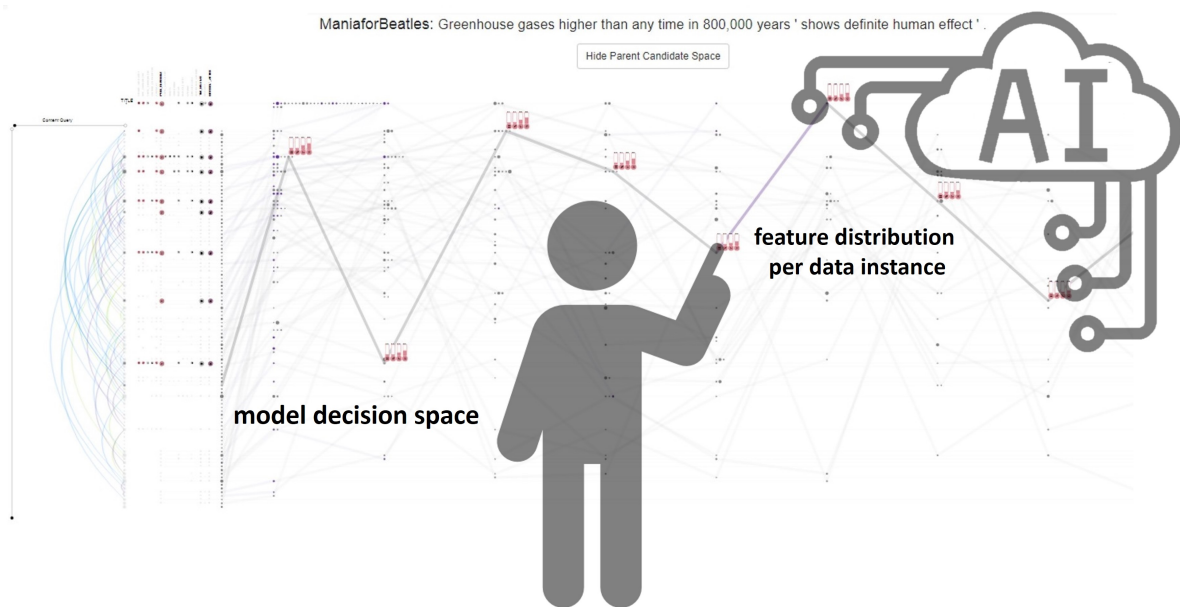


Fig. 6.1: **Model Decision Space** revealing for each instance all considered parent candidates in order of their priority for the model. This view can allow *model analysts* to inspect the decision-making of the model and its associated feature distribution.

branching-out of posts into conversations); (4) Comparing and Understanding Reconstruction-Models; (5) Optimizing Reconstruction-Models (to the given semantics of data and tasks). Addressing these requirements, we present *ThreadReconstructor*, a visual analytics approach for the semi-automatic reconstruction of discussions into threaded conversations. Our technique enables users to answer questions, such as: Which utterances are related in a transcript? What are the main discussion topics, and how are they related? Which aspects of the text did different reconstruction models favor? How do different models compare to each other, and how well did they perform? What are the main deciding factors (features) for a model?

Our proposed technique aims at tightening the human-in-the-loop process by enabling users to participate in the reconstruction of threaded conversation structures using visual analytics. In contrast to the other content chapters in this dissertation, for this technique, we widen the scope of model refinement from topic models to general content models to capture structural elements and obtain a feature distribution for explaining the model's decision space. In addition, to narrow the scope of this chapter, we are reporting the usage of our approach to forum data, more specifically, Reddit data [Red17a]. This is because it is a rich data source that provides a ground-truth structure of threaded conversations. In addition, we have worked with professional forum moderators from a company focused on forum management.

6.1.1 | Research Questions and Challenges

We present a visual analytics technique that reveals the inner workings of content modeling approaches through counterfactual explanations. They are in the form of hypothetical model decisions based on the feature distribution of all the model considerations. This approach

enables model analysts to compare and refine thread reconstruction models without revealing the full model complexity. Thus, our main research question is:

How can we enable model analysts to refine content modeling algorithms while using proxies to understand their inner workings?

To address this question, we have to tackle two challenges: First, as described in the introduction, one of the main difficulties for reconstructing threads from text transcripts is the *lack of ground truth* or the availability of multiple plausible outputs. The second challenge is facilitating the task of *model building for non-ML-experts*.

(1) Multiple Plausible Threading Structures As motivated above, untangling text transcripts to reveal their implicit conversational structure is a particularly challenging task. Besides the data quality issues, one of the main challenges is that often multiple threading structures are plausible. This can, for example, be the case when one post contributes to various threads by referring to or summarizing their main points. This challenge makes the unsupervised training of content modeling approaches notoriously difficult. In practice, model analysts need the flexibility to adjust models based on the concrete data and tasks they are addressing.

(2) Visual Model Building Model analysts are usually data scientists responsible for model building or finetuning. They often make use of existing models and do not have the expertise level of model development or debugging as model experts. Hence, one of the main challenges of this approach is to enable them to construct, diagnose, and refine different content models without requiring them to tune model-specific parameters. To address this issue, we propose a visual model-building workflow that allows for an expressive (no code) model-building through user-defined queries, as described in Section 6.5.2.

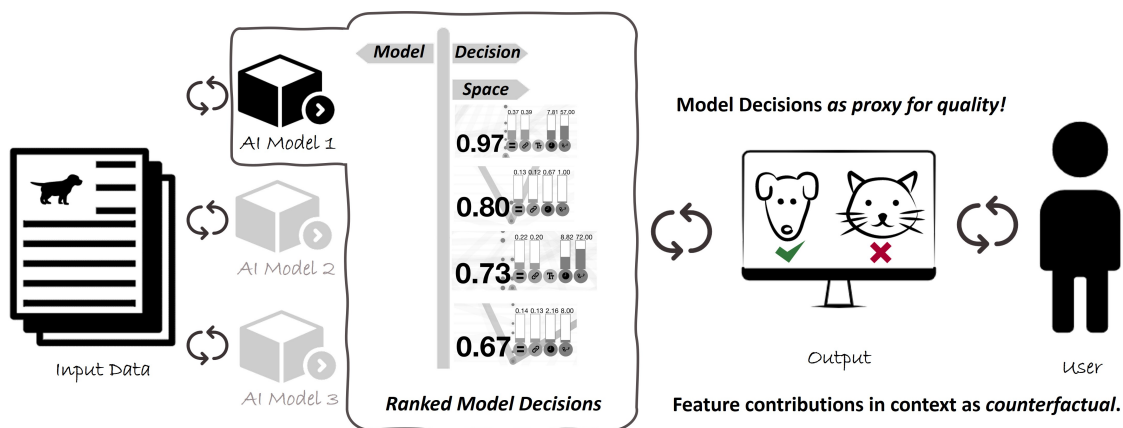


Fig. 6.2: The **contrastive** I&XAI workflow presents the users with a *ranked* list of considered model decisions and allows for the comparison between multiple model configurations. Feature contributions are used as proxies for model quality and allow for *counterfactual* explanations.

6.1.2 | Contributions

This chapter contributes an interactive visual analytics technique for thread reconstruction tailored to model analysts. Our approach relies on task-driven *feature-selection* and *interactive model-generation*. The visual interface enables the *model comparison* of thread reconstruction models, as well as a visualization of *connected components* in threaded conversations. Furthermore, we contribute an interactive visualization for analyzing *model decision spaces*, which we demonstrate on the task of thread reconstruction but could generalize to other domains. We evaluate the impact of our approach with quantitative metrics and through a qualitative study conducted with five model analysts.

6.2 | Designing Visual Analytics Approaches for **Model Analysts**

Model analysts usually adjust or finetune existing models based on specific tasks and data. Thus, they can configure models based on particular requirements. However, as they are neither experts in the domain nor have any detailed understanding of the model's inner workings, we ought to give them an initial representative configuration for the topic modeling output and enable them to compare the results of different settings. Hence, the visual analytics technique for model analysts focuses on facilitating *contrastive reasoning* and feedback. This chapter instantiates the **Input-Output Space** with a focus on the models, as described in Section 3.4.2. The following subsections describe the application design space for a content model refinement approach tailored to model analysts, followed by a concrete I&XAI process model depicting its workflow and the corresponding visual analytics pipeline.

6.2.1 | Application Design Space

The design space of this chapter is spanned by the aspects described in Section 3.5. The most important distinction is choosing the stakeholders to be model analysts (D1). This decision informs some of the other explainability aspects, for example, opting for a model-agnostic (P2), contrastive explanation strategy (D8), with a high abstraction (P4). These aspects are motivated by the design rationales outlined for the **Input-Output Space** in Section 3.4.2. Our approach for designing an I&XAI technique for model analysts is based on the contrastive interaction paradigm as depicted in Figure 6.2. Here, the analysts are presented with multiple models that they can compare. In addition, we reveal each model's decision space through a ranked list of considered model decisions along with their feature distribution. Based on these counterfactual explanations, the analysts can distill important features and generate new models accordingly, which are then displayed on the interface after a training cycle (D5).

This I&XAI technique reveals some underlying model decision-making factors (C8) by depicting the considered model decision, the so-called, *Model Decision Space*. This design makes this technique situated in the **Input-Output Space** with a focus on the model decisions, allowing the analysts to calibrate their understanding and trust of the model (C7). In addition, this design contributed to mitigating potentially harmful rationalizations about the models' decision-making (C2). Through using a visual query language, we simplify the model building task (C4), making it accessible for analysts not traversed in the concrete machine learning approaches used. In our study, we simplified the model features further by using different metaphors (C5) to empower the participants (C6) to critique the construction models and create new ones. Lastly, our approach supports multiple ML experience levels (C1), as evident in our user study.

6.2.2 | I&XAI Workflow and Visual Analytics Pipeline

We designed ThreadReconstructor to tackle the problem of *semi-automatic thread reconstruction in conversational text*. The architecture of our approach is depicted in Figure 6.3, and it is comprised of several stages. First, features are extracted from incoming conversation data, weighted, and fed to a selection of thread reconstruction algorithms, including machine learning and query-based models. The resulting thread structures are compared, and their overlaps and differences are presented in a visual interface, along with the features and details of the evidence used by the model to generate each thread relation. Based on the results of this process, the user can choose to save some or all relations from the model or adjust the parameters of the reconstruction algorithms and re-run the process. Each stage will be described in the following sections. Our approach aims at allowing the users to [T0] **explore** their data, [T2] get a better **understanding** of thread reconstruction models (applied to their data), [T2] generate and **diagnose** different models through comparing their decision spaces, and [T3] **refine** the classifiers using a set of descriptive features for reply relations.

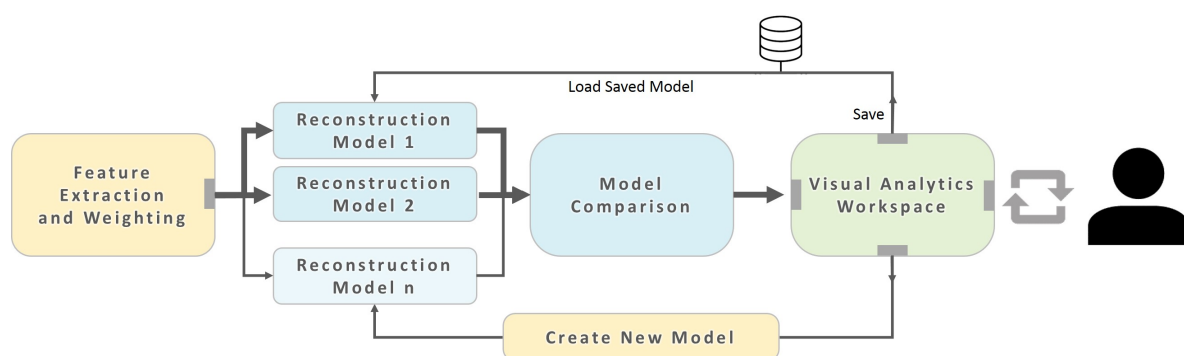


Fig. 6.3: The visual analytics pipeline shows a process model of a contrastive human-AI interaction for content model refinement in the **Input-Output Space** tailored to model analysts. The resulting visual analytics workspace enables analysts to iteratively create and compare reconstruction models, as well as analyze their outputs and considered decisions.

6.3 | Background: Thread Reconstruction

This section frames the background that specifically concerns this chapter. In contrast to the general content and topic models discussed in Section 2.3.1, the work presented in this chapter specifically addresses the task of thread reconstruction. Hence, in the following, we give an overview of reconstruction models, datasets, and evaluation methods.

Thread Reconstruction Models The reconstruction of reply relations in multi-party conversations has been an active area of research. Machine learning approaches proposed to address this research challenge mostly rely on classifier models. In general, the reconstruction of thread structure is done based on two approaches: supervised and unsupervised techniques. Both of these approaches require a set of features, which describe the likelihood that two posts in a thread are linked through a reply relation. Frequently, these features are categorized as content (also called “textual” or “intrinsic”) features (e.g., cosine similarity) or meta-data (also called “non-textual”, “extrinsic”) features (e.g., time-distance) features [ACH11; SCS09; Bal+13]. In the context of our work, we define a model as a reconstruction method that takes in a list of speaker turns (without relation information) as input and generates an acyclic reply relation graph as output. Our approach allows users to generate various models by combining pre-trained classifiers with logical queries and heuristics, as described in Section 6.5. All relations are based on a rich set of features, as described in Section 6.4.1.

Post Similarity Computation Most related works to ours use the cosine similarity function to indicate a reply relation existence. For example, Wang et al. [Wan+07] use a graph-based representation where connections between messages are determined based on inter-message similarity, which is calculated using the cosine similarity function of TF-IDF vectors. They use multiple penalization functions to remove edges between nodes that do not satisfy the selected function (e.g., *time-distance* between messages). Lin et al. [Lin+09] present a slightly different approach, which uses a sparse coding-based model named SMSS (Simultaneously Model Semantics and Structure). The model projects each message into a topic space and approximates it by a linear combination of previous messages in the same discussion. For a new unlabeled post, they compute the similarity between itself and its previous posts, rank the similarities, and then choose the top-ranked post as a candidate parent.

Supervised Reconstruction Methods Most related works on the thread structure’s reconstruction use supervised methods. Multiple machine learning algorithms have been applied for this task, such as decision trees [SMR07; ACH11], support vector machines (SVM) [ACH11], and ranking SVM [SCS09; Bal+13]. Wang et al. [Wan+11] propose a probabilistic model, *threadCRF* to predict the reply structure for threaded discussions. They use two groups of features: node and edge features. Node features depend on the observed attributes in a post. Edge features are defined as the relation between the attributes of two nodes. Liu et al. [LCC13]

use the *threadCRF* to extract a reply relation structure from patient forums. In patient forums, personal relationships are critical to understanding the discussion’s context. Therefore, features such as matching between the *address*, the *signature*, or the *role* of the person (e.g., my daughter) are important. Our work relates to these past works in providing functions such as cosine similarity as inputs to an unsupervised query-based thread reconstruction model. In addition, we have trained two classes of supervised machine learning algorithms (random forests and decision trees) using features inspired by the literature.

Reconstruction Method Evaluation Most supervised models are trained and tested on a relatively short dataset. Some authors set a limitation that the thread should be at least three messages long to be used for training the model [SCS09]. None of the mentioned works explicitly test on long threads (e.g., 100 messages per discussion and more). Some authors evaluate threads of different lengths and show that the performance of the classification model is always better for shorter threads. Wang et al. [Wan+11] emphasize that the size of a thread influences the model’s performance; if the thread length increases, the model’s performance is reduced. This problem is often reduced by restricting the reply relation distance to a fixed window. However, this has the negative effect of eliminating the possibility of modeling long-distance relations that exist in real data. In addition, most of the related work deals with the thread structure’s reconstruction task on only one specific discussion type (e.g., emails), using features tuned to that data to create a model. Different models are needed for different conversational text datasets. Our generalized approach presents a flexible query model that can be tuned to different reconstruction goals (topic, author, quotes, etc.).

Thread Reconstruction Datasets Often real-world datasets are predominately composed of “normal” examples with only a small percentage of “abnormal” or “interesting” examples [Cha+02]. That is known as the *imbalanced classes problem*. Classification of message reply relations belonging to one relatively long thread is a representative example of this issue. If a thread discussion contains n messages, then at most $n - 1$ reply relations may exist in data, assuming that one message can have at most one single parent. At the same time, $\frac{n(n-1)}{2}$ false reply relation candidates exist. If n is relatively large, then the two reply relation classes (*positive* and *negative*) are highly imbalanced. For 100 messages, $\approx 2\%$ reply relations are of the class *positive* and $\approx 98\%$ of the class *negative*. The performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced, as the model’s accuracy could reflect the underlying class distribution [WMZ07]. The model will likely predict the majority class regardless of the input features.

Usually, classification algorithms (for two-class or multi-class problems) require the data to be balanced, meaning there should be the same (or similar) number of instances representing different classes. This requirement significantly influences the classifier’s performance but reduces the classifier’s applicability in real-world scenarios. Our approach includes sampling methods used to balance the training instances for our models. We allow for arbitrary lengths

of threads and do not restrict the reply relation distance by default; however, we optionally apply heuristics found in the related work to allow for an easier comparison of the performance of our tools to past approaches. Our research demonstrates an approach that works for real-world data and provides mechanisms for human-in-the-loop analytics when working with low precision and recall, a possibility for classifiers running on imbalanced datasets.

6.4 | Data-Driven Feature Extraction and Weighting

In order to accurately reconstruct the threaded structure of a conversation, we rely on a set of tailored linguistic and statistical features extracted from the text. These features are weighted based on their distribution in the dataset being analyzed. We extract features for relations between each pair of posts and categorize the posts into 10 different categories.

6.4.1 | Reply Relation Features

Reply relation features are extracted from pairs of messages and capture patterns that can indicate parent-child thread relations. We organize them into three categories based on the **content** of the message, the **structure** of the message in relation to others, and **meta** features related to the dynamics of the conversation.

Content Features **Cosine Similarity**  measures the lexical overlap using the *cosine similarity* [Sin01] function. To improve the detection of similarity, a series of enrichment features expand messages' word vectors as follows: **WordNet Enrichment**  adds synonyms from WordNet. **URL Enrichment**  adds words from linked URLs. **Topic Enrichment**  adds words from topics represented in the document, using a topic model. **Author Enrichment**  adds words used by the same author in other posts. Additionally, **Word Embedding**  uses coreference resolution [Lee+12b] to bring messages closer together lexically by replacing all mentions from one coreference-chain with the first mention (referent). After applying these enrichment techniques, the cosine similarity of the enriched message pairs is recalculated. Finally, messages are modeled using four topic model algorithms (LDA, IHTM [El+19c], SWB [CSS06], and BTM [Yan+13]). **Topic Agreement**  is a Boolean feature that is true if at least k (a threshold set by the user) models agree that the messages belong to the same topic.

Structure Features All structural features are numeric counts of the occurrence of a relation between message pairs. **Quotes**  counts the number of direct quotes across messages, as indicated by quotation marks or ">" (greater than) characters, and the body of the quote. The **Author Name Reference**  feature similarly counts explicit references between messages, in this case, the number of times the parent message's author is explicitly mentioned in the child message. **Substring**  counts the number of common substrings of at least N (default=4) tokens. Substrings may indicate a contextual connection. If an **N-Gram**  is frequently

mentioned in the discussion, it could be an indicator that these tokens are important for the conversation's primary subject. To extract common n-grams from the dataset we count the number of co-present n-grams across message pairs. The Stanford **Named-Entity** 📍 Recognizer [FGM05] is used to extract named entities. Shared named entities are counted for each message pair. If the same word or n-gram appears densely in the discussion, it is a **Lexical Episode** 🗨️ and may indicate that messages, where the sequence appears, are part of one subtopic. Lexical episodes are extracted, then the count of occurrences of different episode words in two messages becomes the value of this feature. Finally, the Stanford CoreNLP **Coreference** 📍 resolution [Rag+10] is used to extract and count *coreferences*, which indicate shared content between messages.

Meta Features The **Distance** 📅 feature captures how many messages have been posted in the time between two messages, while the **Time Distance** 🕒 represents the time-span between replying messages in a thread. The Boolean **Different Authors** 👤 feature is true if two messages have different authors. The intuition is that participants of a discussion do not usually reply to their own previously written messages.

6.4.2 | Message Categorization

Different types of messages are more likely to be classified correctly by different models. Through categorizing messages, we are able to filter out certain types of messages and sort the corpus during analysis, in order to see how discovered relations correlate with message categories. The following message categories were designed using a sample of 10 Reddit threads of 100–200 messages each to explore the characteristics.

Messages having less than 10 tokens are classified as **short** 🗨️, and messages having more than 40 tokens are **long** 🗨️. **Quote** messages 🗨️ contain quotation marks, usually indicating a citation of a previous message. **Question** messages ❓ end with a question mark, which suggests searching for an answer in the following messages. Posts are classified as **active author** 📌 if the author has posted 3+ posts in the thread and **inactive author** 📌 if the author has only posted once. Some messages have little useful content, such as messages containing only **URLs** 🌐, messages containing mainly **special characters** 🗨️, and messages classified as **junk** [Lin+09] due to being off-topic, containing a high proportion of banned words, or being less than 3 tokens long.

6.4.3 | Data-Driven Feature Weighting

For each analyzed dataset, we pre-compute the distribution of the posts across categories and the presence of different reply relations. These are used as indicators for the users when constructing models. Additionally, these distributions are used for a data-driven weighting of the features to set tailored initial weights for every model. In alignment with the methods

introduced in Section 5.3.1, we use feature weighting, as this pre-computation step has proven to be effective for tuning models to changing data characteristics. All weights can be manually adjusted by the users in the visualization interface.

6.5 | User-Driven Model Generation

As shown in Figure 6.3, the feature extraction step builds the foundation for generating the reconstruction models. Reconstruction models are built by combining three techniques: a classifier model; a user-defined query; and a rule-based heuristic. One or more techniques must be selected, and they are executed in order as later techniques act as filters on earlier ones. All models potentially produce multiple parent candidates per child message, therefore, as a post-processing step, these candidates are ranked and the top parent is chosen to be displayed in the visualization interface.

6.5.1 | Classifier Model

Decision Tree [SK16] and Random Forest [Bre01] algorithms were trained on Reddit data to create models for classifying any pair of messages as being in a reply relation or not. Decision Trees were previously used by Aumayr and Chan [ACH11] and showed high performance. Random Forests are used as a comparison as they are more robust against overfitting. Two versions of each model were trained using WEKA [FHW16], one using 5 features defined by Aumayr and Chan [ACH11] and one with 13 features.

The training dataset contains 6926 threads, from which existing reply relations were extracted and labeled as positive. All remaining pairs of messages were linked using a created relation labeled as negative. The numerical features (e.g. *cosine similarity*, *distance*) were calculated and added to the data set. Boolean features (e.g. *different author*) were represented with 1 or 0. Following Aumayr and Chan [ACH11], we artificially create a balanced training dataset using under-sampling, to reduce the number of instances representing the majority (negative) class. The final dataset has 110,038 positive instances, and 110,038 negative ones (in total 220,076 instances). The user can apply one of the trained classifiers to extract the reply relation structure from new datasets. Current models trained on the Reddit data have a high recall, but poor precision (high false positives). As such, these models are suitable as a preprocessing step to determine parent candidates for the query-based model.

6.5.2 | User-Defined Queries

The user-defined queries are a set of criteria that are used to classify message relations as positive or negative. The query is generated in the following way: the user selects a feature subset to use for the reply relation reconstruction and creates a logical expression. For each

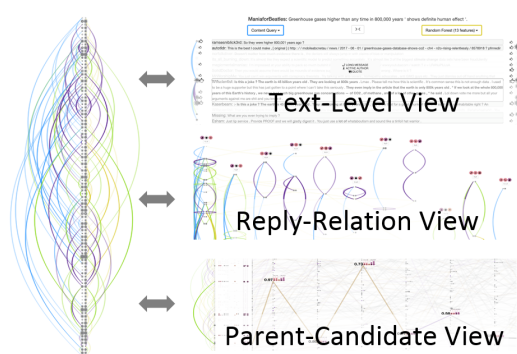
feature, they can set some parameters, such as the minimum similarity level (for content features), and maximum distance between messages (for structural and meta-features). They can weigh each feature, specifying its importance in the decision-making process. When the query is executed, for each possible reply relation the system checks if this relation satisfies the given query. All matching relations are seen as positive relation candidates. The final relation score is the weighted sum of the comparison of all features. For each message, a list of parent candidates is stored, sorted to their scores. The first message (having the highest score) is seen as the most suitable parent message. The query-based model has multiple advantages against the trained classifier. First of all, it does not overfit. The user can create different queries, and the system provides evidence of how reliable the extracted relations are, by showing the presence of features and the number of possible parent candidates for each of the child messages. The user can use this knowledge to adapt the query and improve the certainty level of the extracted reply relations. The query-based model is unsupervised. Hence, it can be applied to discussions where no ground truth information exists.

6.5.3 | Rule-Based Heuristics

Sometimes it is likely that the most frequent parent message is the first message in the thread (the message contributing to the discussion's title). For example, Balali et al. write that, "Commenters who post only one comment on a thread are more likely to reply to the root post." [Bal+13]. Thus our system will assign the title message as the parent to all messages where no appropriate parent message has been found. This heuristic is extended with the finding that an author usually does not reply to the title post in their subsequent message [Bal+13], so our rule doesn't apply to messages by the author of the title message.

6.6 | Visual Analytics Workspace

Our interactive visualization interface consists of four different views, addressing different requirements and analysis tasks. We use visual anchoring and staged animations to transition from one view to the next. We use a consistent color scheme for the **left model**, the **right model**, the **model agreement**, and the **true relations**. All transitions from one view to the next happen through the *Reply Relation View* (Figure 6.5a), which shows an overview



of the complete conversation and all reply relations. In addition to this central view, we introduce three visualizations, each tailored to address one analysis task. The *Text-Level View* is used for close reading and detailed inspection. The *Thematic Forest View* untangles a conversation into separate connected components. Lastly, the *Parent-Candidate View* is used for model

diagnostics. While in any view, users can change settings and generate models using a side-bar, and all saved models are available for selection through drop-down menus on the top **left** and **right**, as shown in Figure 6.4. In addition, all views support a rich set of interactions, e.g., selection, sorting, filtering, linking, and brushing.

Visual Design Considerations Inspired by the technique of Thread Arcs [Ker03], we represent each utterance in a conversation as a node, while the reply structure is depicted as an arc. Hence, a reply relation consists of two nodes and a link connecting them. To reduce edge crossings, the diameter of each arc is proportional to the distance between its endpoints. Nodes are by default ordered by their timestamp. Each node is shown using a circle ●, with a radius proportional to its message’s word count. As introduced by Vehlow et al. [VRW13], to indicate the certainty of a particular parent-child relation, each node encodes information about the number of parent candidates associated with it using a branching-out pattern, i.e., a node ✱ with fewer parent candidates is considered more certain than a node ✨ with more potential parents to choose from. The number of parent candidates is determined by the thread reconstruction model and is explicitly shown in the Parent-Candidate View, as described in Section 6.7. *Junk* messages are colored black so that users can see their prevalence and remove them from the view. Nodes with a reply relation connecting directly to the title message are indicated with a white circle overlaying the node ✱, and edges from these nodes are omitted in the overview to reduce edge clutter. Lastly, additional information about each node and link is shown on-demand using tooltips.

Model Generation Interface Accessible through any view on-demand, the Model Generation Interface is located in the right sidebar of the system. This component allows users to generate custom models, as described in Section 6.5. The three model generation options are enumerated and individually selectable. First, the users have the choice to select one of the trained classifiers. Next, they can choose a query. This can be one of four default queries or created using a visual-querying interface, as described in our previous work, NEREx [El+17b]. This visual-querying interface enables the users to create nested logical expressions by dragging and dropping the feature icons and logical operators. Next, the users can select to apply the rule-based heuristics in their model. Lastly, they have the choice to save a model by giving it a name. All created models immediately become accessible as selection for the **left** and/or **right** side of the visualization, as shown in Figure 6.4. In addition to the saved models, users can select to view the given ground-truth structure (if applicable), a model with all saved relations, as described in Section 6.7.2, as well as provided default models. The remainder of this section will explain each view in more detail.

6.6.1 | Text-Level View

This view, shown in Figure 6.4, is the first component users see. It is designed to facilitate the close reading of texts. By default, all messages are ordered according to their posting time,



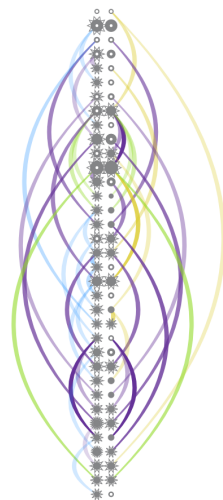
Fig. 6.4: Text-Level View showing all connections of a selected post, according to a *content-focused* query and a random forest model.

and the first line of each message is shown as one box in the center of the view. Hovering on a message reveals the remainder of the hidden text meta-information in tooltips and edge highlights. Selecting a message pin it and all its relations in the view for closer inspection. Edges are the connections created by the chosen models on the left and right sides, respectively.

This view also indicates influential posts. These are found using a score predicting up-voted content that correlates with the number of children. This score is model-dependent and is encoded in the size of a glyph . Messages categorized as junk messages are marked with a black border to indicate that they have insufficient information for accurate classification. After getting an impression of the dataset, users can switch to other views with the button in the upper center, which collapses the text by closing the two models on the side into the middle, creating the Reply Relation View.

6.6.2 | Reply Relation View

As shown in Figure 6.5a, this visualization is the central overview component of the system. It displays the connections of the left and right models on a zoomable canvas to allow for scalability. As the main overview, the reply relation View is the connecting element that enables transitioning from and to all three other visualization components. The side figure shows an example of the connections displayed in this view. The thread diameter reflects the length of a connection and its color indicates the type of the relationship. On the left side of the tree, the tool displays a details bar (not depicted in the figures) with additional, on-demand information about nodes and edges in focus. Additionally, when hovering on a relation, an informative tooltip is shown, depicting all feature values of the relation and common words and phrases shared by the parent and child of this relation. From this view, users can select to transition to other views, such as the untangled Thematic Forest View, the Parent-Candidate View, or go back to the text for close reading.



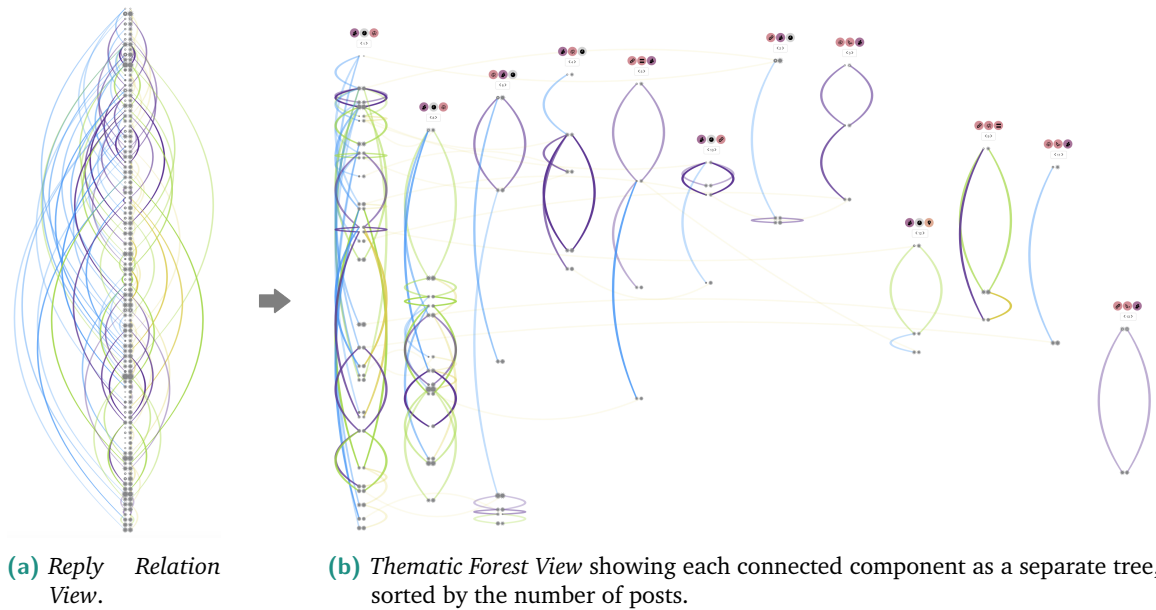


Fig. 6.5: Thematic Forest (6.5b) of untangled reply chains from a full-conversation (6.5a) according to a *content-focused* query (**left arcs**) compared to a random-forest model trained on 13 features (**right arcs**). Model **agreement** and **match to ground truth** are shown using color.

6.6.3 | Thematic Forest View

Figure 6.5b shows an example of a Thematic Forest. This visualization shows the untangled reply chains of a conversation. As the left and right models might not agree on the separation of a discussion into connected components, the layout of this view is based on one model. Using a staggered animation, this view pulls apart the different connected components from the Reply Relation View, resulting in a forest of trees. To reduce clutter, posts that reply only to the title and have no children are joined in one separate connected component. Using the *sort* operation, the users can rearrange the trees to order them according to their number of posts, highlighting the largest trees. In order to get a more compact view, users can choose to *move* all trees vertically closer together, disregarding the temporal positioning, as shown in Figure 6.8. In addition, users can *rebase* the untangled structure from the left to the right model or vice versa, to compare their differences.

Each tree in the forest is annotated with the top three features its relations share. Common words and phrases as well as the particular features common in that tree, for example, if all reply relations of a tree share a specific quote, are shown on demand. Features that are used in the current model are shown as colored icons, while the ones that are common in a tree but not used in the model are shown as gray icons (e.g., Figure 6.8a). This is an indicator that certain features might be interesting to consider adding during model refinement steps. Each tree in the forest can be analyzed in the same manner as the overview. By clicking an expansion button on top of the tree, a Text-Level view is opened in a separate window, as a new analysis sandbox, showing only the nodes of the selected component.

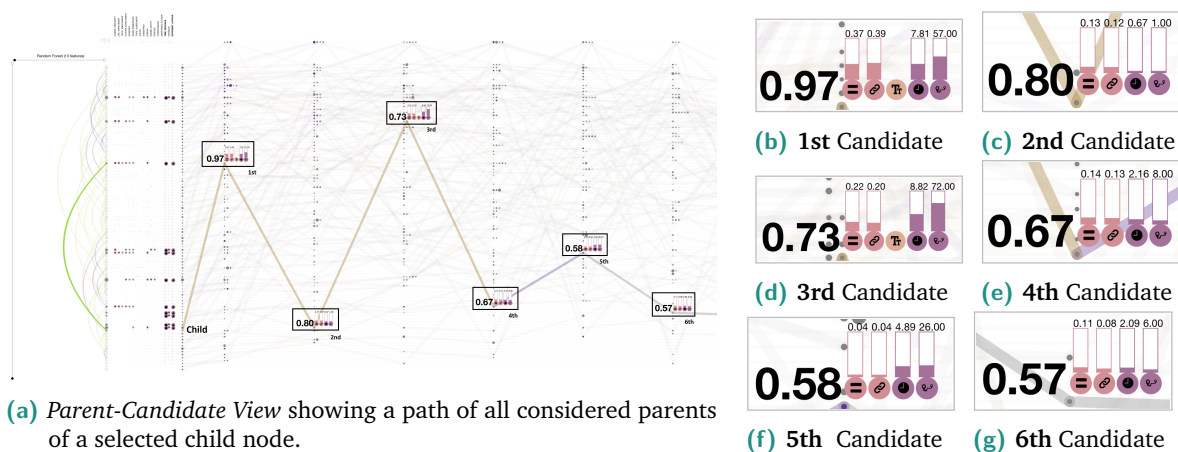


Fig. 6.6: Close-up views on the Parent-Candidate View shown in Figure 6.7b, using one selected child node as an example. The 5th candidate matches the ground truth.

6.7 | Visual Model Diagnostics

Reviewing automatic and semi-automatic machine learning approaches has become an essential task to ensure model reliability and trust [Kra+17]. However, to perform an educated critique of different models, users have to narrow down the number of considered models or embark on an extensive educational journey to understand all relevant aspects of the internal workings of many models. In this section, we present our technique that simplifies the task of model tuning by taking this action from the *Model Space* into the *Model Decision Space*. This approach maps all relevant information from each model into a comparable space, revealing the considerations that models took to reach a decision.

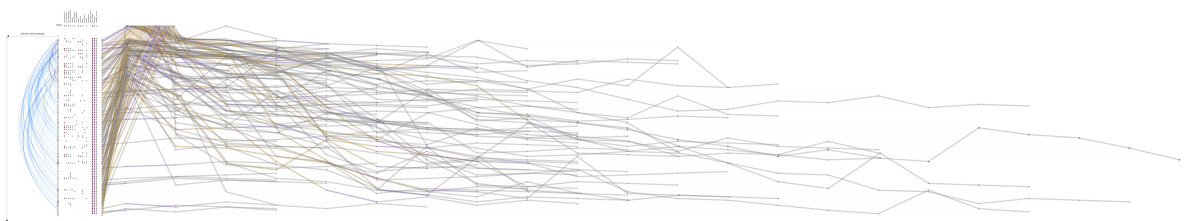
6.7.1 | Parent-Candidate View

As an instance of a *Model Decision Space*, the Parent-Candidate View visualizes the inner workings of thread reconstruction models. In particular, it shows all considered parent candidates for each child, ordered by the internal model certainty, as shown in Figure 6.6. For every node in the conversation, all parent candidates are ordered in columns according to their model-based ranking. In a representation reminiscent of a parallel-coordinate plot, each column can be treated as one dimension, with the first containing the child node followed by one dimension for each candidate ordered by their likelihood. The vertical position is determined based on the corpus order, e.g., sorted according to the time-stamp of the posts. Hence, the first two columns of the Parent-Candidate View contain information about the child and its connection to the chosen parent candidate.

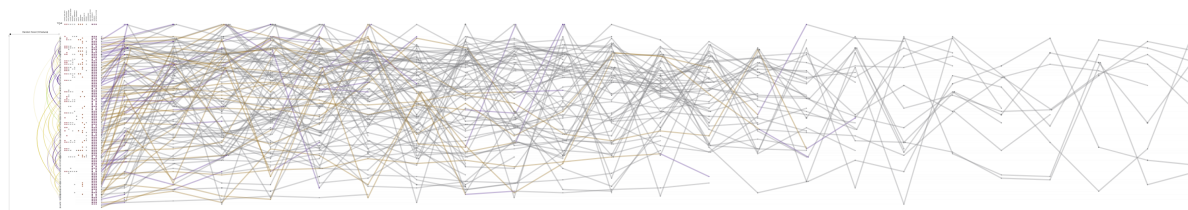
As shown in Figure 6.6a, this view is divided into three parts. On the left, the reply relation from the selected model is shown for visual linking to previous views. In the middle, a feature-distribution pane displays the arrangement of features across the conversation. Lastly, on the right, the parent-candidate space is shown. In this visualization, each node has a stable

vertical position. If multiple child messages have the same ranking for one parent, a node for each child will be packed in the vertical position of that parent within the same column. By default, all parent candidates for a selected child node are connected with a gray edge. If this **path is leading to a correct parent candidate** it is shown in color. The **last path segment before a true parent** is colored in purple, as is the **true parent node**. While a node or path is selected, the corresponding features for each parent, as well as their model certainty are shown. For example, Figure 6.6b shows that the random forest model ranked the first parent for the selected node with a certainty of 0.97. Independent of the features considered by the model, a list of relation features and their scores are available for each parent-child relation. These features highlight the sensitivity of the model to particular attributes. For this particular child, the correct parent node according to the ground truth is the fifth-in-line (Figure 6.6f). Hence, the model would pick this parent only after considering four other nodes.

In addition to exploring single utterances, the parent space view is designed to give an overview of general patterns on a model level. For example, Figure 6.7 shows two models in comparison. Displaying the decision spaces of the models over a complete discussion allows us to compare their biases and understand their particular sensitivities. General patterns, like edge-bundling, zigzag connections, and others, show us differences in the internal scoring of models and helps users assess their quality. It also shows the influence of certain posts on the discussion. Additionally, the length of the produced paths shows the breadth of the models' search space. In addition to using semi-transparent edges, as well as linking and brushing, a *slice&dice* technique is used to reduce the amount of displayed data to reduce visual clutter. Moving the sliders (on the left and top) reduces the considered data to a specific subset of messages to be explored.



(a) **Decision Tree Model** trained on 5 features. **Precision** for top-candidates is **0.17**. Considering all parent candidates **Recall** improves from **0.18** to **0.60**.



(b) **Random Forest Model** trained on 13 features. **Precision** for top-candidates is **0.25**. Considering all parent candidates **Recall** improves from **0.26** to **0.65**.

Fig. 6.7: Comparison of two classifier models in *Parent-Candidate View*. **Model 6.7a** relates most of the children to parents from the beginning of the conversation, while showing a long descending line pattern. In contrast, **Model 6.7b** shows zigzag patterns with more spread-out relations, considering more parent candidates. **Paths leading to a correct parent**, as well as the **agreement between the two models**, are color-coded.

6.7.2 | Model Optimization and Transfer

In combination with the other views, the Parent-Candidate View enables users to understand how models choose parents for the reply relations. Learning about the effect of different features on the reply relations, as well as the sensitivity of models towards particular aspects of the data, generates an opportunity for model tuning and optimization. This is either done through adjusting parameter threshold for existing models or by refining a model by editing its features. Additionally, understanding the feature distribution in a dataset facilitates the creation of new models based on aspects that became relevant through an exploratory analysis.

Our system supports these tasks through an option to save, load, and edit models for existing and new datasets. In addition, users can combine two models based on their agreement to form a new refined one. To support a deeper analysis, we allow users to save relationships during the exploration process. These can be picked from any view, including all parent candidates. The collection of saved relationships is accessible by selecting 'Saved Relationship' from the model drop-down menu. This provides an optimization strategy that is based on the data characteristics, i.e., creating many specialized models to capture different types of relations and combining their results. These can be applied to the whole conversation or to only selected types of messages, e.g., posts containing a question mark and longer posts (assuming we are searching for question-and-answer patterns).

Lastly, our system supports the creation and training of different models on a dataset with a ground-truth structure to be transferred to datasets with no known threaded structure. This is especially useful for our political science collaborators, as most debates they analyze do not typically have a known threaded structure. Using a manually annotated debate would allow them to train several models and transfer them to a broader dataset of discussions.

6.8 | Evaluation

This evaluation section reports the results of our *quantitative* experiments to verify the scalability and reliability of our thread reconstruction models in Section 6.8.1, as well as the results of our *qualitative* user study in Section 6.8.2, Section 6.8.3, and Section 6.8.4. Lastly, we sum up the lessons learned from our evaluation in Section 6.8.5.

6.8.1 | Scalability and Model Reliability

To quantify the performance of our approach in comparison to related techniques, we studied the scalability of our tool on long conversations and the overall reliability of generated reconstruction models. First, we examined the precision, recall, and f-score values of the created

classifier models on a balanced training dataset with 6926 Reddit discussions. These ranged in their length from 50 to 500 posts per file. Overall, we observed, as a result of 10-fold cross-validation, that both decision tree and random forest models achieved higher precision and recall values when trained on 13 features as opposed to five. The highest precision value (0.79) in this artificial setting was achieved by a decision tree model, in contrast to 0.74 for the random forest. However, the recall of the random forest model was slightly higher (0.70) than that of the decision tree (0.68). Both models resulted in a similar f-score (0.72 and 0.73). These results are comparable with the current state-of-the-art in machine learning.

However, while testing these models with imbalanced real-world data (40 Reddit discussions with lengths from 100 to 200), we observed a noticeable decline in the models' performance, as expected on imbalanced data. Both decision trees and random forests had comparably poor precision (0.16) and recall values (0.16). When relaxing the measures to include the top 10 parent candidates instead of the top one, the recall increased, but the precision dropped further. We observed that the quality of the models varied across different datasets, however, none of the models achieved a result on real-world data which was comparable to results with balanced relations. This problem is a known issue in machine learning [Dal+15].

Examining the results of the default query-based models on the same sample of data revealed they often outperform the trained classifiers on certain aspects. For example, the precision-based query achieves a precision value of 0.87, while recall-based queries achieve recall values of up to 0.28. As with the classifiers, these results fluctuate across different datasets. However, queries can be designed as tailored models, optimized to reconstruct a specific aspect of connection. This finding manifests the importance of a visual analytics approach, as only the combination of different model generation methods allows for accurate thread reconstructions by harnessing the strengths of each method.

In contrast to the related work [SMR07; ACH11], our approach showed promising results in reconstructing threads in conversations up to a length of 200 posts. To investigate the scalability of our system in relation to thread length, we conducted a small experiment, analyzing the effect of the length of a conversation on the reliability of the models. We chose to use the same sample of 40 Reddit discussions used in the previous investigations and extracted from each file a new data record containing only the top 10 or the top 30 messages, respectively. Using the default query models, we investigated the effect of the length of a conversation on the precision and recall values of the models. The results were apparent and confirmed findings by Wang et al. [Wan+11]. Since shorter conversations contained substantially fewer false-positive relation candidates, both the precision and recall values of models reconstructing threads in shorter discussions increased. For example, a content query applied to our data sample had an average precision of 0.36; 0.56; 0.70 and recall of 0.29; 0.48; 0.66, when applied on all posts; the top 30; the top 10, respectively.

6.8.2 | Expert User Studies

To evaluate the effectiveness and usability of our system, we conducted a qualitative expert user study. Through a collaboration with a company that owns and operates a significant amount of online community forums, we gained a unique insight into their work and the challenges that they face while dealing with massive online data. After a description of the study arrangement, we discuss the case studies, feedback gathered, and lessons learned.

Methodology Due to the number of available views and settings in addition to the limited time of qualified experts to learn all aspects of a new system, we chose to conduct a pair analytics study [KF14]. We conducted five two-hour sessions in which a member of our team (henceforth referred to as **visual analytics expert**, short **VAE**) worked with the one domain expert (henceforth referred to as **subject matter expert**, short **SME**). Each session was divided into three parts. In the first 30 minutes, after a quick introduction, the VAE explained the functionality of the tool while gathering initial feedback on the overall utility and design choices through a semi-structured interview. After making sure the SME understood the basic functionality and controls of the system, in the next hour the SME and VAE embarked on an open-ended, exploratory analysis of one dataset using the tool, guided by the interests of the SME. The SME had control over the interface with occasional input from the VAE to clarify certain interaction possibilities. During the analysis, the SMEs were encouraged to think aloud and explain the tasks they were doing. Questions from the VAE occasionally guided this study toward new analysis tasks. The last part consisted of a 30-minute feedback session, reflecting on the initial feedback and the performed analyses. This last part was guided by questions from the VAE to get a general assessment of the system, and its utility and visual design from the SME. All study sessions were audio-recorded and screen-captured for further analysis.

Datasets and Controls To ensure the validity of our study we sought datasets that fulfill certain criteria used as controls. First, we were looking for discussions with generally familiar content. Second, the lengths of these conversations ought to be between 100 to 200 utterances, to ensure long enough reply chains while remaining manageable for the users. Third, to have a reference for evaluation, we required conversations with previously known thread structures as ground-truth data. We, therefore, selected two Reddit debates regarding the topics of climate change [Red17b] and immigration [Red17c], respectively. The first dataset is a discussion of a news article entitled: “*Greenhouse gases higher than any time in 800,000 years ‘shows definite human effect’*” [Sla17], which highlights research results showing a rising level of greenhouse gases found through studying air trapped in ice cores. The second dataset is a debate on another news article entitled: “*Trump is deporting fewer immigrants than Obama, including criminals*” [Sac17]. Using these datasets across the five study sessions (3x climate change, 2x immigration) allowed us to balance the study to the effects of the content on the outcome. Generally, we observed that the usage of our system did not vary for different datasets.

Participants All the participants in our study work for a company that manages a substantial amount of forum data. However, the variation in their educational backgrounds, job responsibilities, as well as their self-declared technical proficiency varied a lot. Four out of five participants (in the following referred to as **forum moderators**, short **FM1–FM4**) worked in forum moderation (**model analysts**) with duties ranging from customer support to legal aid. Their highest educational degree also varied from secondary school across different levels of higher education. All four participants work with computers on a daily basis and are privately interested in forums. In contrast, as a control, our fifth participant (in the following referred to as **machine learning expert**, short **MLE**) works as a research scientist (**developer and model expert**) at the company and holds a Ph.D. in machine learning.

Tasks Based on our problem characterization (Section 6.3), the focus of this study was to facilitate four high-level analysis tasks, namely; data exploration [**T0**]; model generation and comparison [**T1**]; model decision understanding [**T2**]; and untangling forum discussions [**T3**]. Throughout the exploration process, the SMEs were interested in different aspects of the data and tool. Nevertheless, the VAE guided the analysis towards tasks that were not covered by the SMEs through unobtrusive questions, e.g., *“How about we take a look at the different trees this model generated?”*

6.8.3 | Case Studies

In the following, we describe a selection of case studies chosen from all five sessions. Since many scenarios were repetitive, we picked representative cases and categorized them according to the high-level tasks mentioned in the previous section.

Data Exploration

Branching out from a Topic of Interest The first view all users see in the system is the Text-Level View. Typically, users read over the title and some posts to get a feeling of the underlying data. FM3 started his analysis with the same strategy. After finding an intriguing post accusing deniers of climate change of never being convinced regardless of the amount of available evidence, he mentioned that he *“want[s] to follow this post and see what people respond to that.”* He then selected the post to examine its relations according to the default models. Becoming excited about the topic, FM3 announced that *“ [he would] create a model to look for people who agreed with each other quickly.”* The generated model was based on a query [ && ( ||  || ) &&  && ] that favored content agreement and shorter time distances. After comparing this model to existing ones, he deemed it a good one for further exploration, using the Thematic Forest View. While analyzing the untangled discussion, he commented that *“[...] the discussion about how deniers are wrong is interesting, you can see how much people are frustrated by deniers of climate change in the conversation.”* Examining more content-related reply chains, he mentioned that *“[he] learned something about the ‘earth cooling cycle’”* from

the debate, while commenting as an enthusiast for forum discussions: “*This tool is interesting for me, [...] I would like to participate in this conversation because I can see the history of how the discussion evolved.*”

Model Generation and Comparison

Refining Default Queries All SMEs relied on the default queries for their first experiments with the model generation, comparing their results to the preloaded classifier models. Through this initial analysis, they got a feel for the sensitivities of the models and their usability. After forming more concrete questions and hypotheses during this analysis phase, they would continue with generating models. FM4, for example, got interested in analyzing the relations of named entities in the discussion and how different entities propagate throughout reply chains. Starting with a default content-focused query, FM4 continued exploring the relations in the Thematic Forest. Through investigating the most prominent features in each connected component (see Figure 6.8), she concluded that she would add the time-distance as a feature of her next query, as this feature was present in the data but missed by the current query.

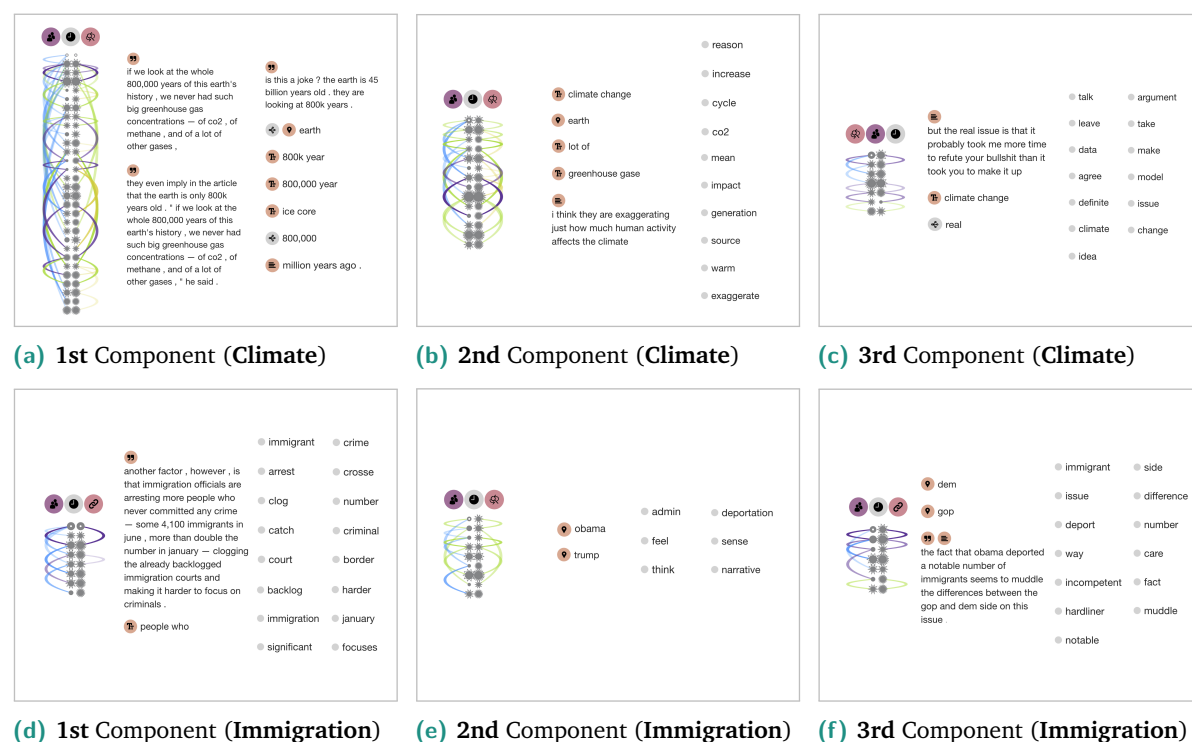


Fig. 6.8: Thematic Forest View showing the three first connected components from the untangled view of Figure 6.5b and their top features (in the top row). They are based on a discussion about the reasons and impact of climate change and indicate the altered arguments of the debate, i.e. (a) the validity of the scientific work, (b) the impact of greenhouse gases, (c) the frustration with the deniers of climate change. The second row shows the first three connected components of an untangled discussion about immigration in the US and the political strategy of Donald Trump in comparison to Barack Obama, especially concerning border safety and the number of arrests of undocumented immigrants.

Verifying Hypotheses FM1 started her analysis with a hypothesis on the behavior of forum users, saying: *“People quoting each other are either agreeing or disagreeing.”* Consequently, she embarked on a mission to examine the aspects of agreement and disagreement around the topic of climate change. She first created a broad, query-based model [(👤 || 🗨️ || 🔄) && 🧑] that is focused on textual or entity repetitions from different authors, commenting that she *“[wants] to create a reliable model based on real information.”* Afterwards, she created a second, more restricted, query-based model [(👤 || 🗨️) && 🗨️] in which she was searching for *“posts that are quoted [...] using lexical episodes to filter out aspects like science or politics.”* Both created models had a high amount of true relations. However, FM1 was interested in their agreement, which only contained true relations. After untangling the discussion, she closely read the connected posts (e.g., Figure 6.4 and Figure 6.8a) and concluded that *“[she] thought more people would argue against each other but through looking at the reply-chain I found that in this particular conversation people are supporting the claims more than I would have expected on such a controversial topic.”*

Untangling Forum Discussions

Searching for Content Patterns After reading the title of the discussion, FM2 mentioned that *“[he wants] to explore what the different side-discussions are about.”* Expecting noticeable patterns in the content of the debate around climate change, he started exploring the connected components separated by a content-focused query, as shown in Figure 6.5b. Taking a closer look at the common features of each subtree, he concluded that *“everyone is trying to prove their own theory, they are not really referring to the other theories mentioned. It’s a distribution of ‘random facts’.”* Seeing a post from a denier of climate change (Figure 6.4), he analyzed its connections (Figure 6.8a) and mentioned that *“this is a person who sparked a heated debate at the beginning of the thread. Many people are outraged in their replies that this person did not understand that the article was referring to 800k years as a reference to the age of the ice cores not the age of the earth.”* During this close-reading process, he occasionally pointed to different utterances saying: *“Here you can see a thread of very argumentative posts.”* In addition, he also found many connected components sharing common URLs, commenting that *“many people refer to the same URL to prove a theory.”* After realizing this, he declared: *“The model we created worked out very well, maybe I will use the URL feature for the next model.”*

Understanding Connections FM2 was also interested in understanding why some models relate certain posts to each other and others do not. He referred to it as *“helping [him] build more trust in the models.”* To explore this, he chose two models and used the transition between their respective Thematic Forest Views to investigate changes in the composition of the forum according to each model. He routinely checked selected reply chains and commented: *“I can see why these posts would connect, that makes sense. The considered features are working.”* or *“This model seems to be biased towards quotes.”* During this analysis, he made another finding. He found messages posted by the *Reddit Helper-Bot* and concluded that *“[...] in some cases posts were automatically removed because participants used a lot of profanity or bad language.”*

Model Decision Understanding

Following the KDD Pipeline While the analysis of all forum moderators was driven by the content of the discussions, the MLE was more interested in the model optimization task and found the Parent-Candidate View the most interesting. She described this view as showing the “*soft links between datasets.*” She commented that “*it visually highlights the sensitivities of the models towards certain aspects of the data.*” For example, Figure 6.6b indicates that the model’s decision was mostly impacted by a higher cosine similarity of the child and first parent-candidate. Based on her prior knowledge of the used algorithms, she assumed that “*the decision tree classifier will be less robust because it tends to overfit the training data.*” Exploring this aspect of model fitness, she continued her analysis through the lens of the KDD pipeline [FPS96]. She first started exploring the feature distribution across the given dataset, using the feature panel of the Parent-Candidate View, as shown in Figure 6.6a. She was interested in the overall distribution of features, investigating the features used in a particular model (highlighted in color). She commented that “*the first step in studying the performance of a model is to understand the underlying data distribution, as this is the signal the model is receiving.*”

She continued by analyzing high-level patterns across different models, as shown in Figure 6.7. She observed a different pattern for decision trees and random forest and commented that “*the behavior of these two models is as [she expected]. Both models were trained on the same features. However, the decision tree is learning a very narrow pattern, which is a sign of overfitting. In contrast, the random forest model is showing a wider range of considered candidates, indicating a broader search space.*” The last step of the KDD pipeline is the evaluation and interpretation, where the MLE gave an opinion about the two models, she explained that “*[she] suspect[s] that overall the random forest model is a more reliable model for thread reconstruction as it is more robust towards overfitting.*” This confirms our independent observation in all other studies. When asked about model overfitting, she commented that “*optimizing models to achieve higher f-scores is often not practical, as the ground-truth is only one possible connection option, other alternatives might make sense for different tasks. A good connection does not always have to be about content.*” She continued, “*people often fixate on a specific aspect and treat it as the truth. However, models might show how different concepts propagate through a corpus, which helps in understanding the data.*”

Exploring the Magic Box Both FM1 and FM4 were interested in understanding more about the classifier models. Without prior knowledge of the specific ML model, FM1 mentioned that “*[he was] intrigued to understand more about the models, seeing the parent candidates definitely allows me to gain more trust in some models.*” He reached this conclusion after having spent some time talking about his mistrust of automatic, black-box models as a reason for not using them. He also said: “*If I had a more practical demonstration of what these models can do, I would have more faith in it – this application is getting pretty high marks right now.*”

While FM1 was discussing model trust, FM4 was interested in choosing the correct model and optimizing her existing ones. She used the Parent-Candidate View to analyze all candidates picked by a model for a particular post, as shown in Figure 6.6. She commented that *“reading the text of the parent candidates and seeing their feature weighting [gave her] more ideas on how to create new models.”* Despite not understanding how the classifier models work, FM4 was still interested in using and comparing them. She built herself an abstracted mental model to deal with the classifiers, calling them *“Magic-Boxes”*. Hence, her exploration of these models was purely visual. She commented that *“the random forests look much better in the Parent-Candidate View than the decision-trees.”*, referring to their high-level patterns, as shown in Figure 6.7. After examining the patterns more closely, she concluded for the random-forest model (Figure 6.6a) that *“the zigzag-path shows two conflicting aspects that the model is considering—it is trying to connect the post to others (from the top) that are related to the same content (entity = climate change) but also to create connections to posts (from the bottom) that are using the same angry language to prove a counter-argument.”*

6.8.4 | Feedback

Throughout the study, we collected feedback from all SMEs in the form of semi-structured interview questions. All tasks performed during the analysis sessions were motivated by the requirements mentioned in Section 6.1. The SME expertise covered the stakeholder groups anticipated for the technique; all FMs were fulfilling the roles of model analysts, while the MLE took the role of a developer and model expert. In this section, we are reporting a summary of the feedback received during all user study sessions.

Initial Feedback During the first 30 minutes of every study session, we gathered initial feedback from the SMEs to capture their first impression of our system and the general research direction. All participants highlighted the importance of having such a system for their work and endorsed its relevance for the company. FM2 explained that *“this tool will make it a lot easier to pin down who is saying what to whom, which will be really handy for forum moderation, especially when dealing with legal issues.”* This was echoed by FM1, who said: *“I would use it for all the harassment claims that I get because these are often forums that are many pages long and take a lot of time to deal with.”* The same sentiment was shared by FM3, who complained that *“sorting through comments is always difficult, you regularly feel like there is more in there that you must have overlooked.”* He emphasized that *“having a tool like this would help [him] to find the topics [he is] actually looking for and see all their connection even in a larger forum.”*

Hence, all FMs needed to reconstruct reply relations in forums (as such a structure is not maintained by their company-owned forums). They all reported a very time-consuming, manual process for doing this job and did not support fully automating this task. FM2 was particularly skeptical of automatic reconstruction algorithms, saying: *“I don’t have enough confidence that these approaches completely cover all aspects of the discussions I am looking for. Especially when*

dealing with legal issues, I have to have all points covered. Using such a system seems to be a good compromise between efficiency and thoroughness.” He later on clarified that *“the visualization reveals all relevant aspects of the data and confirms [his] mental model, which makes [him] have more trust and confidence in using the system.”*

The MLE commented on the relevance of a gold-standard training dataset, saying: *“Looking at the agreement between the models highlights their certainty, even if the ground-truth is not agreeing with all the connections, they might still be relevant and definitely worth exploring.”* She mentioned that *“[she could] clearly see the benefit of such a tool to find patterns in many different ways, not just the expected ones.”*

Visualization Design and Usability After the analysis, we asked all users to comment on their experience. Except for minor issues (e.g., changing button labels), all participants praised the design and usability of the tool. With comments like; *“That’s very beautiful!”* or *“I really like how pretty the visualizations are!”*, we were pleased to see that the general aesthetics, color choices, and icons were well-perceived. In particular, FM4 commented that *“[she enjoyed] using the visual query interface because it makes it seem very easy to create a model.”* FM1 appreciated having general relation features to choose for the query rather than the tag-based system she usually uses, describing that *“using tags to filter and create query assumes that the users know something about the content of the data, which is often not the case.”* All users also approved of the concept of a layered analysis through visual anchoring and appreciated the transitions between different views highlighting that it preserves the context of the analysis while making differences visible.

General Assessment Despite the tool’s usefulness, most SMEs noticed a steep learning curve during the introduction of the system. However, all participants had no difficulty using all functionalities explained. When asked about their willingness to learn using such a system, all of them were affirmative, highlighting the benefits they gained from such an expressive tool.

Every user found a set of specific tasks they would like to perform using the tool. For example, FM3 commented that *“busier forums often get confusing to handle and when dealing with harassment issues, [he] does not want to miss a comment in a conversation”*, depicting the system as *“a way for wiretapping into conversations.”* He also mentioned that *“this tool helps in revealing the influence of certain posts on a conversation. [He would] use it, therefore, to stop fights happening on forums because of misunderstandings.”* Lastly, he highlighted, that *“often people banned from a forum come back using a different account to continue talking about the same topic. [He saw] potential for the tool to reveal such cases of fraud.”*

When asked what they would do when presented with more analysis time, all users were interested in creating new models to discover other aspects of connection in the data. Surprisingly, all SMEs agreed that while optimizing a model towards a given ground-truth structure could

be effective for machine learning, from their practical experience such a structure makes little sense in following a conversation flow. As FM1 explained, this is *“because people often hit the wrong reply button or do not bother at all.”* Therefore, they confirmed that automatizing the reconstruction based on precision and recall values did not fulfill their needs from a thread reconstruction system. FM2 commented that *“this given structure often reveals the order of replies but does not show general content patterns. That is why having user-defined queries is so helpful.”* The MLE highlighted that *“a post can be influenced by many previous messages, so it technically could have multiple parents,”* finding the Parent-Candidate view particularly useful for exploring this aspect.

6.8.5 | Lessons Learned

Addressing the requirements identified in the introduction, our tool supports the reconstruction of different reply chains based on the user’s tasks, allowing them to untangle conversations using their own mental models and semantics. The study shows that all users created different models, compared them, and tuned them based on their insights. We also observed that through this analytical process, our system could increase the model’s trustworthiness based on a better understanding of the generated structures.

From our analysis, we see the success of supporting users across a variety of levels of expertise and depth of analysis—from getting users interested in using a tool through appropriate default settings and a starting view that is familiar, to a selection of more advanced features further in the analysis. However, even with the provisions, we made for model diagnosis and understanding, model trust remains an important concern for some SMEs, especially when dealing with critical data such as legal and harassment issues.

We observed a process of learning through discovery. While we attempted to provide appropriate training in an introduction session, participants found themselves using features even when they were not fully understood. However, through the process of experimentation, the participants learned about feature extraction and machine learning. The exploratory nature of the interface, allowing for low-cost experimentation with queries and models, likely afforded this online learning.

Participants mentioned the aesthetics of the interface as an important aspect of their desire to incorporate it into their regular workflows. When building future tools to integrate visual analytics into existing workspaces, aesthetics should be carefully considered from both the ability to accurately read the visualizations, but also from the user experience point of view.

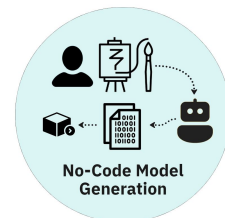
Some of the specific interests of participants in the study point to the potential for even more methods for thread reconstructions. As we entered this project, we had the traditional machine-learning view that the goal would be to recreate the ground truth. What we learned was

the importance of flexibility in creating thread “*reconstructions*” that, while not representing ground-truth data, did connect messages in meaningful ways for user tasks. An example of a potential new reconstruction method arising from our study results would be to create a model which detects rapid back-and-forth discussions between conversation participants, which may be indicative of a fight.

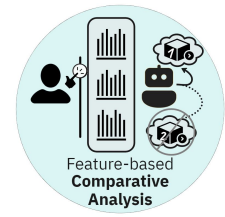
6.9 | Discussion: Investigating I&XAI Design Patterns

Based on the presented technique, in this section, we discuss three potentially generalizable I&XAI design patterns. These are tailored to interactive and explainable applications for the *model analysts* in the **Input-Output Space** with a focus on the model. An overview of all design patterns discussed in this dissertation can be found in Section 8.4

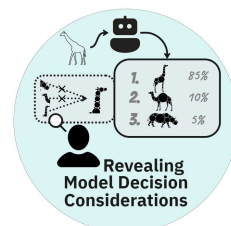
(3a) No-Code Model Generation To allow analysts to perform targeted problem-solving, we need to facilitate their access to creating and refining ML models. Hence, we propose a *visual, no-code environment* for model generation. Here, users can rely on provided options and examples to configure their own models and test them out against other models.



(3b) Feature-based Comparative Analysis To understand the decision-making of different models, we need to compare the aspects they consider when making these decisions. Hence, it is crucial to provide model analysts with a comparative analysis interface that facilitates *feature-level comparisons*, true to the ML models’ decision-making.



(3c) Revealing Model Decision Considerations Instead of revealing all of the inner workings of ML models to analysts and potentially overwhelming or confusing them, a good compromise is to reveal all of the model’s decision considerations. This allows analysts to get a deeper understanding of the model and work their way towards “*looking under the hood.*”



Model Space

” *If you can't explain it simply
you don't understand it well enough.*

— **Albert Einstein**
(attributed)

Topic Model Optimization based on User-Steerable Speculative Execution

” *It is the ability to make predictions about the future that is the crux of intelligence.*

— Jeff Hawkins

On Intelligence [HB04] p.6

Summary

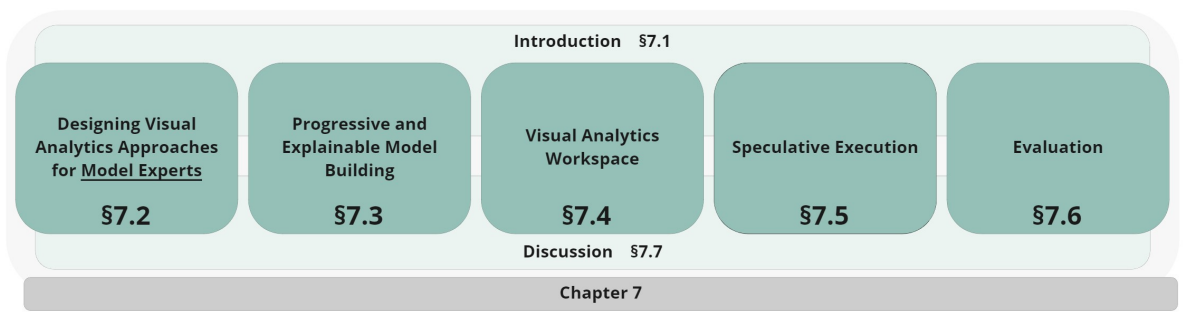
For the **model experts**, we present an intelligible topic modeling technique designed in the **Model Space**. It relies on an incremental hierarchical topic modeling algorithm to visualize the algorithmic decision-making process. Model experts can use this system to understand the model’s inner workings based on different parameters. To effectively assess the potential consequences of human interventions, we introduce *speculative execution* as a paradigm for creating user-steerable previews. Whenever the measured model quality deteriorates, the system automatically triggers a speculative execution of various optimization strategies, requesting external intervention. Experts compare the proposed optimizations to the current model state, previewing their effect on the next model iterations before applying one.

Contributions

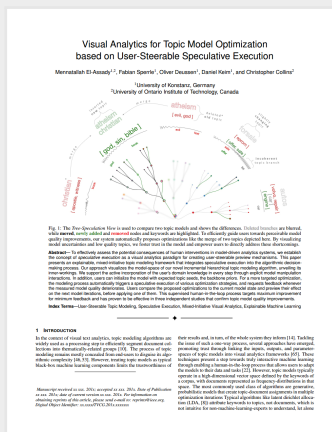
- (i) An I&XAI workflow and visual analytics pipeline for topic model refinement tailored to model experts in Section 7.2.
- (ii) An explainable algorithm for *incremental topic modeling* in Section 7.3.
- (iii) A visual workspace for *progressive* analytics and feedback in Section 7.4.
- (iv) The *Speculative Execution* paradigm for semi-supervised optimization based on model agency and allowing for informed decision-making Section 7.5.
- (v) An expert user study confirming the applicability of our approach in Section 7.6.

Contents

7.1	Introduction	176
7.1.1	Research Questions and Challenges	177
7.1.2	Contributions	178
7.2	Designing Visual Analytics Approaches for Model Experts	178
7.2.1	Application Design Space	179
7.2.2	I&XAI Workflow and Visual Analytics Pipeline	180
7.3	Progressive and Explainable Model Building	180
7.3.1	Similarity Computation	181
7.3.2	Quality Measurement	182
7.3.3	IHTM Evaluation	183
7.4	Visual Analytics Workspace	184
7.4.1	IHTM-Backbone Priors	186
7.4.2	Model-Space View	187
7.4.3	Model Quality Monitoring	190
7.5	Speculative Execution	190
7.5.1	Automatic Optimization Strategies	191
7.5.2	Comparative Speculation Mode	192
7.6	Evaluation	194
7.6.1	Qualitative Results: Expert Feedback	195
7.6.2	Quantitative Results: Model Quality Assessment	197
7.6.3	Lessons Learned	197
7.7	Discussion: Investigating I&XAI Design Patterns	198



This chapter is based on the following publication:



Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution

Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen,
Daniel Keim, and Christopher Collins

IEEE Transactions on Visualization and Computer
Graphics, 25(1):374–384, 2019.

El-Assady et al. [El-+19c]

Contribution Clarification

This chapter is based on the publication entitled “*Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution*” [El-+19c]. The system presented in this paper is an extension of Fabian Sperrle’s master’s project, which I closely supervised. The user study was conducted and analyzed by me, and the paper writing was conducted by me with only slight edits from my co-authors.

7.1 | Introduction

Model developers and machine learning experts are often tasked with refining topic models. Their main challenge is that they often cannot see into the black box of topic modeling, opting to instead use metrics (e.g., precision, recall, coherence, etc.) as a proxy for quality, as described in Section 2.3.3. These metrics do not always capture human intuition and understanding as they disguise the algorithmic complexity and nuanced decision-making [Chu+12]. Hence, in addition to abstract measures of quality, it is essential to interpret the models' decision-making plausibility in the context of the high-dimensional vector space defined by the keywords of a corpus. Without visualizing this process, it is notoriously difficult to get an intuition on the effect of changing algorithmic parameters of these models, as “seemingly small changes on the user side could have unpredictable and nonsensical cascading side effects” [Lee+17b].

Hence, to provide complete transparency for model experts, in this chapter, we present a technique that unveils the inner workings and decision-making processes of topic modeling. This is achieved by using a directly explainable and visualizable algorithm, the **Incremental Hierarchical Topic Model (IHTM)**. It is a deterministic, similarity-based algorithm developed to tightly integrate the human in every step of the algorithmic decision-making process. This user-steerable model was developed to maintain a competitive performance with state-of-the-art topic models. Furthermore, IHTM is able to request human intervention when required to leverage human decision-making efficiently. Thus, we present a mixed-initiative visual analytics technique for direct, *bidirectional* interactions between the visualization and the machine learning components.

For effective human-in-the-loop decision-making, we further introduce a novel visual analytics guidance mechanism. Based on the continuous monitoring of the model stress level, the system can prompt users when it detects a problem with the algorithmic decisions, requesting their feedback on a diverse set of automatically prepared model optimizations. To support the user's decision-making, speculations on the effect of possible decisions are presented, allowing them to perform more informed refinements. This **speculative execution (SpecEx)** highlights the different outputs that may emerge from a user's feedback *before changes take effect*, altering the operational path of the model. Hence, to estimate the potential impact of interactions, users can rely on SpecEx as an efficient preview instrument. In this chapter, we present a two-fold speculative execution; one on the effect of different optimization strategies and another temporal speculation on the next steps of the IHTM algorithm.

For visual analytics solutions incorporating SpecEx, it is essential that the model is understandable, deterministic, and explainable so that users will be able to understand how to interact with it. SpecEx analytics tools require incremental processing so that users can intervene and steer the model development as the data is processed. In order to enable speculative execution, there needs to be some metrics related to the task which can be predicted. In the topic modeling

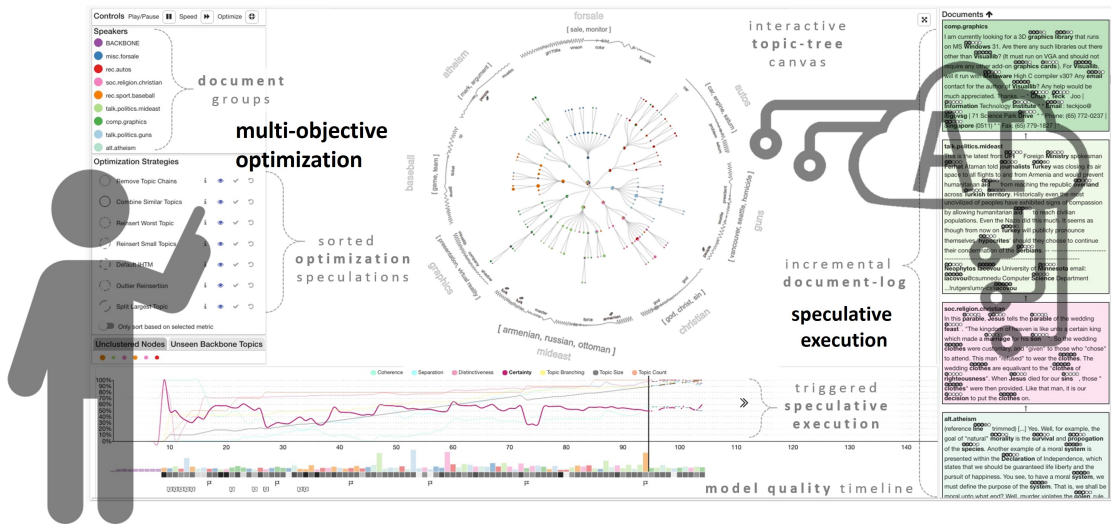


Fig. 7.1: Speculative Execution of topic model optimizations for *model experts*. Users preview the effect of optimization before applying it. This technique opens the black box of topic modeling, revealing the model's inner workings.

scenario, these are measures of quality, as introduced in Section 2.3.3. Figure 7.1 depicts the visual preview mechanism enabled by speculative execution for the topic model refinement technique presented in this chapter.

We evaluated our technique with three different approaches. First, we validated the IHTM topic model against common probabilistic models, with all annotators confirming that IHTM is competitive and achieves results of comparable quality. Second, we verified that experts could use our tool to achieve model improvements through a user study with six participants optimizing a model of a familiar dataset, followed by a semi-structured interview. Finally, we evaluated the model quality improvements by having four annotators rank quality across a variety of configurations, revealing that supervised speculative execution resulted in better models than fully automatic approaches.

7.1.1 | Research Questions and Challenges

In this chapter, we present a visual analytics technique that involves the users in each model decision while incrementally constructing a hierarchical topic model. It allows model experts to understand algorithmic optimizations in different analysis contexts and accordingly perform model refinements through steering the internal multi-objective optimization of the topic modeling. Thus, our main research question is:

How can we enable model experts to understand and refine topic modeling algorithms with minimal feedback for maximum gain?

To address this question, we have to tackle two challenges: First, enabling the model to assess its quality at every processing step to allow for *effective user feedback*. Second, providing appropriate *guidance* to the users in order to facilitate their model steering decisions. We address these challenges by employing the following two design choices:

(1) Model Quality Monitoring One of the main challenges when allowing users to interact directly with the model’s decision-making is finding the appropriate time and place for them to provide feedback. Hence, to streamline this process, we introduce the continuous **model quality monitoring** (MQM) component, as discussed in Section 7.4.3. It constantly observes the internal model quality of the IHTM algorithm to catch changes in the quality of the model and inform the overall system to alert the users, e.g., by triggering a new optimization cycle through speculative execution.

(2) Speculative Execution Speculative execution, in parallel computing, is used to reduce latency in I/O tasks by pre-fetching and executing probable instructions [OG88; CG99]. In visual analytics, we propose to adapt speculative execution to provide real-time previews of multiple paths to guide user decisions. Alexander et al. [Ale+14] argue against too much guidance and seek to foster serendipitous interactions for exploratory tasks. For the model optimization task, however, a more guided approach may be helpful. Lee et al. [Lee+17b] introduce preview mechanisms for topic modeling. Building on this, our speculative execution provides previews of alternative paths for topic model optimizations, as discussed in Section 7.5.

7.1.2 | Contributions

This chapter contributes a visual analytics technique for topic model refinement tailored to model experts. It relies on speculative execution as a semi-supervised optimization method which can reveal possible alternative outcomes for automated processes to inform model steering. Specifically, we present the explainable and visualizable IHTM algorithm, including techniques for incorporating the users’ domain knowledge throughout the model-building process. This is embedded in a tailored visual analytics workspace for the progressive understanding, diagnosis, and refinement of the modeling process. Our technique was evaluated in a multi-stage mixed-methods study confirming the effectiveness of the presented approach.

7.2 | Designing Visual Analytics Approaches for Model Experts

Model experts have a sophisticated understanding of their models and can diagnose intricate details to optimize the machine learning models in various contexts. To enable such a human-AI interaction, we need to design visual analytics workspaces that enable inductive reasoning

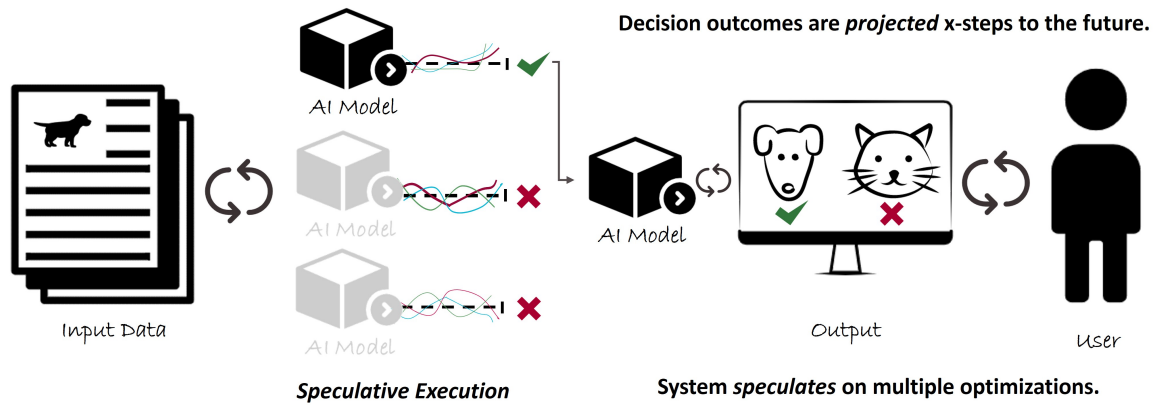


Fig. 7.2: The **inductive** I&XAI workflow incrementally processes the input data to model it continuously. Whenever the model requires human input, it *speculates* different optimization strategies, projects their results to the future, and presents them to the users, who can then make an informed decision, steering the model.

and interactions. However, when allowing users to intervene in the training process of an ML model through model-aware explanations, we need to design effective interaction techniques that balance the human and the machine agency. This chapter instantiates the **Model Space** as described in Section 3.4.3. The following subsections describe the application design space for a topic model refinement approach tailored to model experts, followed by a concrete I&XAI process model depicting a SpecEx workflow and the corresponding visual analytics pipeline.

7.2.1 | Application Design Space

The design space of this chapter is spanned by the aspects described in Section 3.5. The most important distinction is choosing the stakeholders to be model experts (D1). This decision informs some of the other explainability aspects, for example, opting for a model-aware (P2), inductive explanation strategy (D8), with a low abstraction (P4). These aspects are motivated by the design rationales outlined for the **Model Space** in Section 3.4.3. Our approach for designing an I&XAI technique for model experts is based on the inductive interaction paradigm of *speculative execution*, see Figure 7.2. Here, the experts weigh in on the multi-objective optimization process of the model and provide feedback for each analysis context based on their understanding. Hence, the I&XAI interface allows for *bidirectional* interaction between the human and the model during training (D5).

This I&XAI technique opens the black box of topic modeling (C1) and allows the model experts to inspect every decision made by the model (C8). However, instead of requiring the users to excerpt too much mental effort (C4) to find suitable intervention points for refinement, the model halts whenever it is uncertain and provides ranked optimization suggestions (C7). In addition, to allow for informed decision-making (C2), users can inspect the results of a decision before accepting it (C6). This is done through running every optimization suggestion in a speculative execution sandbox (C5) and generating a diff-view to the current model (C3).

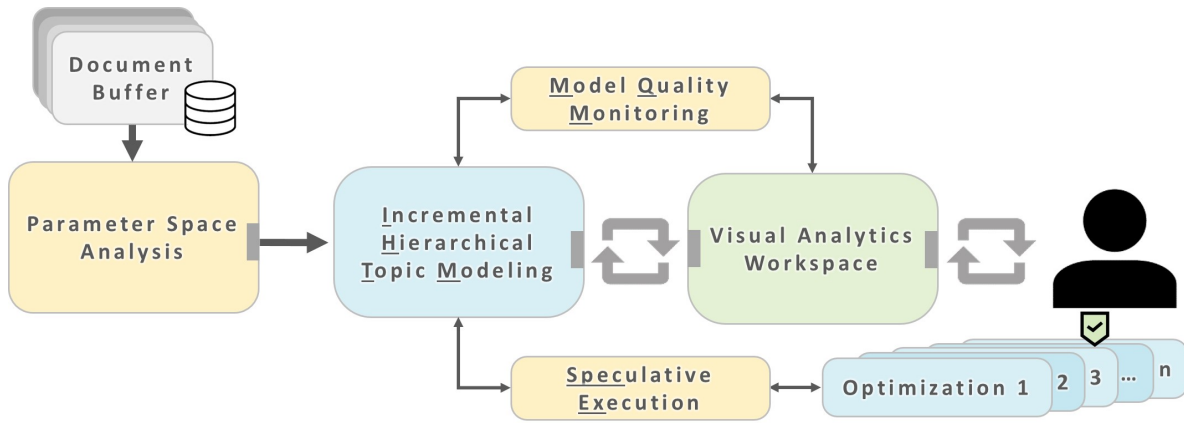


Fig. 7.3: The visual analytics pipeline shows a process model of an inductive human-AI interaction for content model refinement in the **Model Space** tailored to model experts. The user-steerable topic modeling optimization process includes constant model quality monitoring that triggers speculative optimization strategies, whenever the measured model quality deteriorates.

7.2.2 | I&XAI Workflow and Visual Analytics Pipeline

In alignment with the previous chapters, this technique addresses the following tasks: [T1] topic model **understanding**, [T2] model quality monitoring (**diagnosis**), and [T3] topic tree **refinement**. Figure 7.3 depicts our proposed technique for mixed-initiative topic modeling using visual analytics. The document buffer holds batch segments of the corpus, ready for analysis. These could be coming from a real document stream or a sorted document collection. Based on the parameter space analysis proposed in Section 5.3, every batch of documents is prepared for further processing. The IHTM then sequentially assigns each document to a topic branch, constructing a hierarchical topic structure. Every insertion is directly visualized, evaluated, and can be interactively adjusted. The model quality monitoring component constantly tracks twelve quality measures, evaluating the current *stress-level* of the algorithm. This component informs the IHTM and visual workspace of the current model quality and triggers the algorithm to halt whenever its heuristics detect a deterioration in the measured model quality. The main design rationale behind this process is to target *minimum feedback* for *maximum improvement*.

7.3 | Progressive and Explainable Model Building

In this section, we introduce the **Incremental Hierarchical Topic Model (IHTM)** (Algorithm 1) as the foundation of our mixed-initiative visual analytics technique. This explainable model progressively builds a hierarchical topic tree by iteratively inserting each document from a collection. The tree's leaf nodes represent individual documents, while all inner nodes group these documents into topics. The root node is thus an aggregation of the entire corpus. The document inserts are based on a monotonic similarity function, making the algorithm's output

deterministic, with each document belonging to exactly one topic. Documents contributing to multiple topics can be segmented into sub-documents using a topic-based document split routine [Gal+03], with each sub-document being inserted into one topic.

Algorithm IHTopicModeling(*document-buffer*)

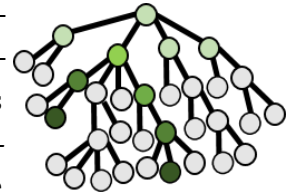
```

1  root = initialize the topic tree with an empty root;
2  while document-buffer is not empty do
3      node = create node from next doc;                                // incoming node
4      if numKeywords(node) ≥ minKeywords then
5          similarNodeSet = recursiveTreeDescend(node,root);
6          updateTree(node,similarNodeSet);                                // insert
7          extractTopicDescriptors(root);                                // update topics
8          updateMQM();                                                    // calculate all metrics
9      else unclusteredNodes.add(node)                                // doc too short
10         updateVisualisation();                                          // show insert decisions
11         if qualityDeclined() then
12             optimize();                                                // trigger optimization
13         incorporateUserFeedback();                                      // bidirectional
14     applyFinalTreePruningAndSegment();                                  // best tree cut
15     return extractTopicDescriptors(root);

```

Algo 1: Incremental Hierarchical Topic Modeling

When inserting new documents, the algorithm determines the similarity between the new document node and existing tree nodes based on the similarity of their respective keyword vectors. Starting with the root, the model finds the most similar branch within the tree for inserting the incoming node. To guarantee to find these branches without having to compare every node to the whole tree, we rely on the monotonicity property of the similarity function [BBV08]. It is defined as $\forall d, p, c \subseteq T : \min(\text{sim}(d, p), \text{sim}(d, c)) \geq \text{sim}(d, p \cup c)$, with d as *incoming document-vector* and p, c as *direct parent* and *child* nodes in the tree (T), respectively — $p \cup c$ is the normalized sum of the vectors of p and c . This guarantees that the similarity cannot increase when going from leaf nodes to the root, making each parent node in the tree more general than its child. Hence, for every new insertion, the algorithm starts with the root, calling a recursive insertion function `recursiveTreeDescend`. It evaluates the similarity of the incoming node to the current node and all its direct children. If one of the children is found to be more similar than the current node, the algorithm descends its respective tree branch; otherwise, recursion is terminated.

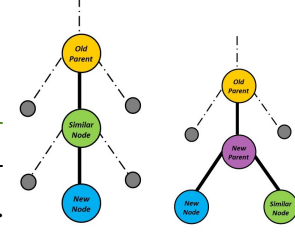


7.3.1 | Similarity Computation

Cosine similarity is our default similarity function but can be exchanged to suit the task and genre. Documents are represented as weighted keyword vectors, as proposed in Section 5.3.

The keyword-vector $V \in \mathbb{R}^w$ of each inner-node i is the sum of the vectors of all its leaf-node children $l(i)$: $V_i = \sum_{c \in l(i)} V_c$. Here, w is the number of corpus keywords.

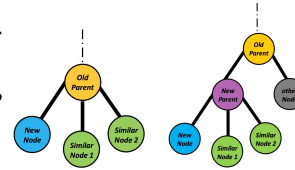
The result of this similarity computation is a set containing a single most similar node or a set of equally-similar nodes, called `similarNodeSet`. The model then distinguishes between three cases to complete the insert in the `updateTree` routine, as described in the following. (1) If the **incoming node** is similar to one other node, it differentiates whether the **similar node** is an inner node or a leaf node. In the case of an inner node, the incoming node gets appended as a direct child (1a).



(1a)

(1b)

Otherwise, a **new mutual parent** for the incoming node and the similar leaf node is created to ensure that every document remains a leaf node in the tree (1b). (2) If **multiple sibling nodes** are equally **similar**, the **incoming node** is appended depending on whether these are the only children of their **parent** or only a subset of the children. For the former, the node gets inserted as another sibling node (2a), and for the latter, the node and all similar siblings get inserted under a **new parent** (2b) (which becomes a direct child of the **previous parent** of these siblings). (3) If the tree has more than one equally-similar node and not all of them are siblings, heuristics determine the most likely insert branch, selecting one node (e.g., closest time-stamp or same author) and proceeding with the first insert strategy.



(2a)

(2b)

7.3.2 | Quality Measurement

We measure the quality of the model after an insert by calculating multiple quality metrics introduced in Section 2.3.3 (`updateMQM()`) and evaluate the last insert position using the *insert certainty*:

$$\text{insert_certainty}(d) = \frac{\max_{n \in \text{similarNodeSet}} (\text{sim}(d, n))}{\left(\frac{\sum_{n \in m} \text{sim}(d, n)}{|m|} \right)}$$

with m as the set of all nodes visited by `recursiveTreeDescend`. It measures how certain the insertion into a topic was by comparing the similarity to the chosen parent node with all other similarities encountered during the insertion process. This triggers an update of the visual analytics workspace (`updateVisualization()`), as discussed in the next section. As our system is uncertainty-aware, it automatically pauses the computation whenever necessary, requesting the users' feedback. The IHTM treats all direct children of the root as high-level topics and uses a scoring function (log-likelihood ratio [MS99]) after each document-insertion to extract the top keywords for these topics as *descriptors* (`extractTopicDescriptors()`).

Following the recommendation of Lee et al. [Lee+17b], our model is designed to be an explainable topic modeling approach and support all their suggested update and refinement operations, including adding, removing, and changing keywords; moving documents within the topic hierarchy; and lastly, splitting, merging, removing, and creating topics. Thus, the IHTM can be tightly coupled with a visual analytics system that constantly interacts with the algorithm throughout its progressive model-building process. IHTM is based on an iterative, distance-based clustering technique that yields a visualizable *model-space* — in contrast to typical probabilistic, generative topic models. As documents are incrementally inserted into the tree, the IHTM does not require the specification of an expected number of topics before the start of the modeling process. This is an important property for the use of topic models in a streaming context or when visualizing every model-decision step iteratively, as is described in more detail in the next section.

7.3.3 | IHTM Evaluation

To assess the correctness and perceived quality of the basic IHTM algorithm, we conducted a comparative study based on expert annotations from three political scientists. For this study, we chose two corpora with different characteristics, namely, (1) a collection of 200 news articles from the COHA-corpus [Dav10] (sorted by timestamp), and (2) a transcribed presidential debate [Uni12]. Where applicable, the models were restricted to 50 and 10 topics, respectively, based on previous knowledge of the data. All the experts were familiar with the content of the two datasets. We asked the participants to rate the quality of each topic (per model) on a scale from 1 (worst) to 10 (best). We gave them printed sheets, each containing the output of one topic run (on one dataset). The sheets were randomized and anonymized. The annotators reported their judgments were influenced by *whether they could find all expected themes in separate topics* and based on the *number of meaningful keywords per topic descriptor*. This study compared the result of the IHTM topics (no optimization, no interaction) with three other probabilistic topic models, HDP [Teh+06], hPAM [Min+09], and LDA [BNJ03]. These were chosen as a baseline since they have been previously considered as reference models in other comparative evaluation studies [Ble12b]. Furthermore, as our evaluation was based on expert feedback on each topic model result as a whole, we did not consider other clustering techniques that do not perform a cluster-labeling step (as these would require a time-consuming comparison of cluster memberships).

As shown in Table 7.1, the IHTM output achieved an average score of 8, outperforming the three other models. On average, annotators rated IHTM 40% better than the LDA baseline (an improvement from 5.7 to 8.0) and 3.9% better than HDP (an improvement from 7.7 to 8.0). The average inter-annotator agreement was 84%. IHTM performed best on transcribed discourse data, as this type of text has a linear structure based on the social interaction between speakers. Such data aligns well with our incremental processing approach. These results

Algo	Annotation Score [1-10]			Improvement to LDA		
	News [Dav10]	Debate [Uni12]	Avg	News [Dav10]	Debate [Uni12]	Avg
LDA [BNJ03]	5.9 (0.5)	5.6 (0.7)	5.7 (0.6)	0%	0%	0%
HDP [Teh+06]	7.9 (1.3)	7.5 (1.2)	7.7 (1.2)	35.4%	32.9%	34.3%
hPAM [Min+09]	6.7 (0.8)	8.0 (0.5)	7.4 (0.6)	15.2%	42.1%	28.6%
IHTM	7.7 (0.7)	8.4 (0.8)	8.0 (0.8)	31.1%	49.4%	40.3%

Table 7.1: Comparison of IHTM against state-of-the-art topic models. Annotators scored the topics (scale: 1–worst, 10–best; standard deviation in parenthesis). The improvements shown in the table are relative to the LDA baseline.

indicate that the basic IHTM algorithm produces topic modeling results that are competitive with state-of-the-art models.

7.4 | Visual Analytics Workspace

The visual analytics workspace is the primary window of interaction with which the users have to steer the topic modeling process. Supporting this complex task, we strive for a single-screen interface to reduce the users’ cognitive load during the analysis. As motivated in the introduction, the visual analytics workspace supports the three analysis tasks of **[T1]** understanding, **[T2]** diagnosis (through model quality monitoring), and **[T3]** refinement. Thus we designed different visualization components tailored to aid in performing these tasks.

Figure 7.4 depicts an overview of the visualization interface, including all views. It shows both the model space and the algorithm’s decision-making process, as well as topic hierarchies, keywords, and uncertainties. The *model quality timeline* tracks the metric development over time and forecasts the effects of speculative executions (see Section 7.5). A *control panel* on the left-hand side ranks optimizations based on their effectiveness for the current model and enables steering the speculative execution. Optimizations can be viewed , accepted  and rejected . The *control panel* on top allows pausing the modeling process , setting its speed , and manually triggering a speculative execution . A color map shows which *document group* the nodes in the tree belong to. *Unclustered nodes* that have not yet been added to the model are listed and can be added via drag and drop. In addition to this interface, panels on each side of the screen open up on demand, enabling users to adjust parameters (settings bar, right) and see details of interactively selected items (details bar, left).

Depending on the current topic modeling state, the workspace enters one of two modes, varying the functionality of the views and, in turn, the possible interactions. First, the **operation mode** is the primary state in which the IHTM is running, updating *one model*. Second, the **speculation mode** is a state where the IHTM is waiting for the user’s feedback, showing *candidate-models* as alternatives that the user can compare and choose from.

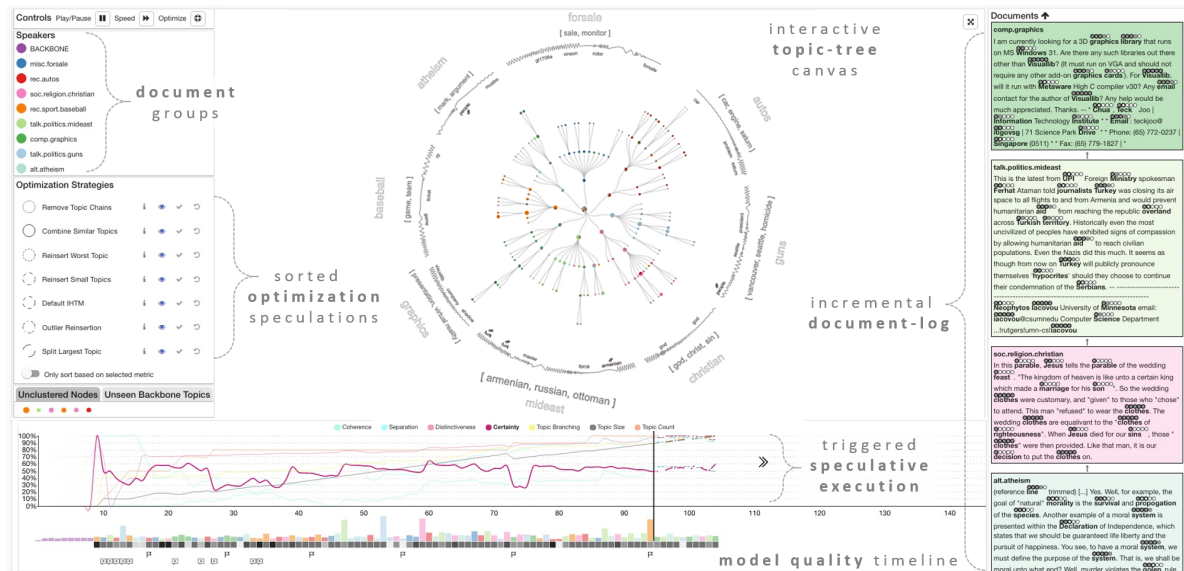


Fig. 7.4: In the IHTM workspace, new documents are added to the *document log* on the right-hand side before they are shown in the *topic tree*, the central component of the workspace.

While the users can incorporate their feedback in both modes, the expected interactions and feedback types differ. In the operation mode, only one instance of the topic tree is available for interaction. Therefore, the workspace supports more fine-grained, direct tree manipulations, as described in Section 7.4.2.

In contrast, in the speculation mode, various candidate trees are available as possible optimizations of the current tree. The central target for the interaction in this mode is the comparative analysis of two trees. This is supported by tailored difference-visualizations, e.g., Figure 7.7. This section focuses on describing the visual workspace while in the default operation mode, followed by Section 7.5 that describes the system state while in speculation mode.

Design Rationale To achieve a simple yet expressive design, we incorporated a set of visual cues throughout the interface. The first and most important cue is the *highlighting of keywords and descriptors* in all visualization components. This provides the users with an intuition of the *unit of analysis* considered by the topic modeling, i.e., users should become aware that topic models reduce documents to vectors of keyword frequencies. To highlight this, we designed *keyword-glyphs* that encode the global (topic-model relevance score) and local (frequency in the document) importance score of keywords. Both scores are discretized onto a five-point scale, encoding the global score to the outside of the glyph (black circle-border) and the local score to the inside of the glyph (gray circle), e.g., (global score 1, local score 4) and (global score 5, local score 2). These glyphs are positioned on top of keywords in several view components, as in Figure 7.4. They are used, for example, on hover, in detail-bars, and in the *document-log* — our close-reading [Jän+15] view.

In addition, as the IHTM computes uncertainty information for nodes, edges, and insert-operations, we designed *uncertainty-aware visualization* components using a concept of *visual-fuzziness*, as introduced by Vehlow et al. [VRW13]. The different visual elements encoding this uncertainty information will be introduced within the remainder of this section. Lastly, we strive for *complete transparency* when displaying the modeling results. We, therefore, include information on unclustered documents or backbone topics that users expect but that have not been detected by the algorithm in a panel for investigation.

Interaction Design All visualization components are highly interactive to support the users' exploration and analysis. For example, as shown in Figure 7.4, each document is represented in three coordinated views in the workspace; in the document-log, in the topic tree (as a leaf node), and in the timeline. In addition to *linking-and-brushing*, we provide *two levels of details-on-demand* for topics and documents. First, a detail panel on hover that contains only essential information such as keyword distributions; second, a detail sidebar with all relevant quality measures and other statistics. To further reduce complexity, the workspace is initialized with sensible *default parameters* for all components. However, we also provide an *expert-interaction mode* that shows a wide range of controls for more advanced adaptations of the visualization or of the algorithm. Lastly, to ensure data-provenance [Sti+16], tracking,sec:ch07-parameter-space-analysis and *trust-building*, we track and show all user interactions, displaying all instances of optimization and direct model-manipulations in the timeline view (see Section 7.4.3).

7.4.1 | IHTM-Backbone Priors

As motivated by Andrzejewski et al. [AZC09], introducing “must-link” and “cannot-link” constraints to topic models yields substantial improvement as it offers a user-defined notion of relevance for the semantics of the document-collection at hand. However, Lee et al. [Lee+17b] found that this method of specifying complex semantics is unintuitive to users. Jagarlamudi et al. [JDU12] suggest allowing users to generate topic keyword lists to direct models. The IHTM algorithm was, therefore, designed with this concept of domain-knowledge incorporation in mind. Based on our experience working with social science scholars on topic model tuning, we deduced that one potential place to include users' feedback on their expected topics in a corpus is *before starting* the model-building. Hence, we introduce topic backbone priors as a technique to prime the model, avoiding a cold start. Figure 7.5 shows the interface where users can construct the topic backbones. On the side, we list the most important keywords of the corpus, from where they can be dropped into the topics. Keywords are colored based on whether they originate from the **corpus** or have been added **manually** or through **word2vec** as similar keywords. Users can create a hierarchical topic structure by adding  topics. These could be derived from an external data source (e.g., a discussion agenda) or could be automatically computed by our system using heuristics and other topic models. For example, users can provide some representative documents and determine a topic distribution using

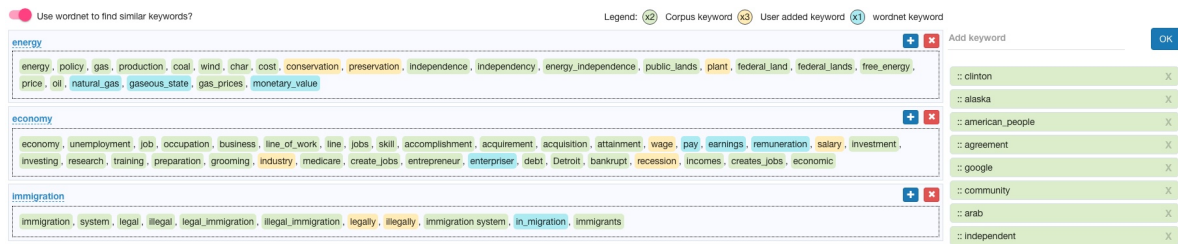


Fig. 7.5: To prime the model with their domain knowledge, users can provide a topic backbone consisting of expected topics, each characterized by representative keywords.

LDA [BNJ03]. They can then directly adjust the suggestion in place by manually editing the keywords, declaring topics irrelevant, or completely removing  them. They are additionally assisted by an automatically-ranked keyword list that extracts the most discriminative keywords based on the G^2 -metric [Dun93]. Similar keywords could be computed using *Wordnet* [Mil95] to support the users in generating a comprehensive backbone. Users can specify names for the backbone topics, which are visualized in the topic tree, as discussed in Section 7.4.2.

The backbone technique allows the users to adapt and guide the IHTM to their preferred topic granularity. For example, when analyzing news articles, users might choose to create a high-level backbone (sports, technology, crime, etc.), or they might prefer a more specific one about concrete events (a particular game, a new phone, etc.). This allows the users to adjust the topic modeling process to their data and tasks for more targeted analysis. The backbone keywords are weighted and used to initialize the IHTM topic tree before starting the document insertion loop of the algorithm. However, these backbones are considered subtle indicators for the model. Hence, if the model recognizes documents as not being similar to any user-defined backbone topic, it will construct a new topic branch. Therefore, users cannot accidentally manipulate the model into generating nonexistent topics. Unseen backbone topics are indicated in the visual interface to guide the user (see Figure 7.4). Finally, any backbone topics that users have created can be saved and loaded back in the next session, enabling users to create collections of their domain knowledge that can be modularly applied to their data.

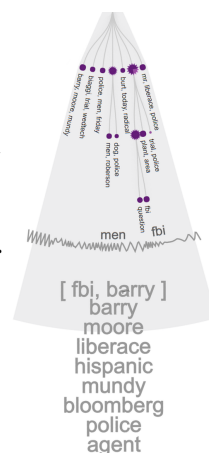
7.4.2 | Model-Space View

As described in Section 7.3, the IHTM algorithm updates the visual workspace based on the model's internal state and all performed tree-update operations after each insert step. Visualizing the model space of the algorithm, the radial topic tree is incrementally built as a central part of the workspace, as depicted in Figure 7.4. To guarantee scalability, the tree is placed on a zoomable canvas, and users can collapse branches interactively to hide topics or remove tree imbalances. The tree root represents an aggregation of the entire corpus, while the leaves are documents, and the inner nodes represent the topic hierarchy.

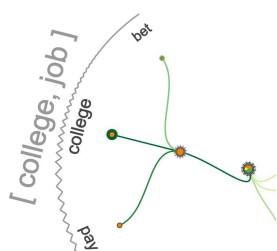
The default color scheme of the topic tree mirrors the distribution to document authors (or utterance speakers for conversation transcripts) per node. The node colors can be changed to show the inner-node variance. In addition, the size of each node represents the amount of text accumulated in all of its associated children. Applying the metaphor of node-fuzziness [Gör+18], we encode the within-cluster diversity, which we described to users as uncertainty for simplicity of explanation, of each topic (variance of pairwise-similarities of its children) to differentiate nodes with potential for improvement ✨ from certain topic-nodes ●. The same concept is applied to branches using *squiggly lines* on the outside of the tree. Line-segments showing higher frequency and amplitude indicate *noisiness* or higher variance in sub-topic branches. In addition to such confidence indicators, the outside of the tree shows three levels of keywords and descriptors, as shown in Figure 7.6. The keywords closest to the tree (on the inside of the line) correspond to the second-level sub-topics in the tree, i.e., the different branches of the main topics, which are, in turn, all direct children of the root.

Due to space constraints, we only show these keywords for larger topics and not for every leaf node. The main topic descriptors are represented as an array on the outside of the line, e.g., [coal, oil]. Users can choose to name each topic. These names are then shown on the outermost segment, e.g., *Energy*. Besides getting the topic names directly from the user, the algorithm derives names for the topics based on defined backbone priors (Section 7.4.1).

To reduce visual clutter, the topic tree only depicts the top keywords and descriptors for topics and sub-topics, respectively. Keywords appearing in multiple topics are highlighted with a glyph . However, these first few keywords of the topic vectors might not reveal the nuanced differences between their content. We, therefore, integrate a rich set of interactions. This includes hovering over keyword vectors and topic names to reveal more descriptors. In addition, all nodes in the tree show a detail box on hover to enable an understanding of why they were positioned in their respective branch in the tree. Furthermore, clicking an element on the topic tree opens a detail panel from the screen's left side.



Our visual analytics approach visualizes the process of topic modeling incrementally. Therefore, we use staged animations to show each document's insertion into the topic tree. Each document enters the workspace as a new textbox in the document log. Using an animated transition, a node is formed out of this box and traverses down the tree, from root to leaf. The visualization



shows all decisions the algorithm is taking during the recursive similarity computation as described in Algorithm 1, line 5. Starting from the root, similar candidate branches are highlighted in green (darker color showing more similarity), indicating the decision path of the IHTM algorithm. In addition, the most similar nodes along this path are highlighted. After the insert-position of the incoming node is found, we employ a clock metaphor and place it as the rightmost child of its

new parent. The document circle is animated into its new position and the tree updates all keywords and node positions. This step-by-step animation directly explains the inner workings of the model. To speed up the modeling process, users can adjust the speed of the animation or skip the branch highlighting phase, ensuring scalability to larger corpora.

While observing the model-building process, users can stop the algorithm to directly interact with the topic tree and incorporate their feedback. This can be done through dragging-and-dropping any node in the tree (or any unclustered node) to a new parent. To assist this process, the system highlights on demand the outlier children in a sub-topic or similar nodes to a selected one, enabling users to move outlier children to a new branch or bring similar nodes from other branches together. If users find a document or a sub-topic in the tree that is incorrectly classified, they can trigger an automatic delete-and-reinsert action. This positions

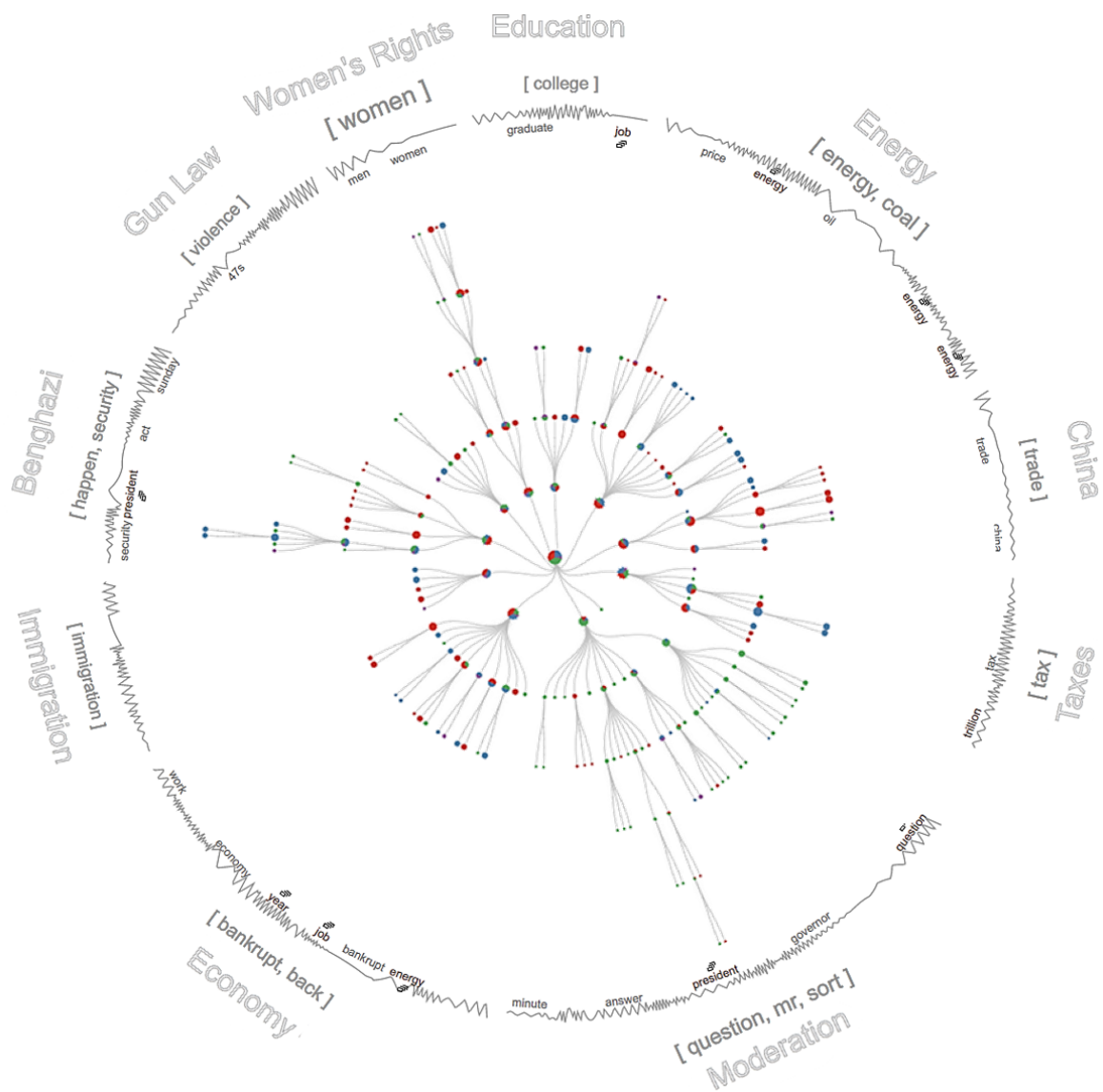
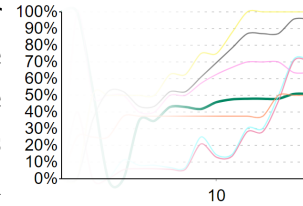


Fig. 7.6: IHTM result (with backbone; supervised SpecEx optimizations; no direct tree manipulations) for the 2nd presidential debate between Romney and Obama in 2012 [Uni12].

the deleted node and all its children at the front of the processing queue, forcing the IHTM algorithm to reinsert them into the current tree. This interaction is particularly useful for overcoming earlier errors resulting from the incremental model-building process.

7.4.3 | Model Quality Monitoring

Monitoring the quality of the topic modeling is a vital guidance element of our visual analytics approach. The IHTM algorithm issues an update for the twelve model quality measures during every insert-loop (Algorithm 1, line 8). To visualize and interact with these measures, we include a timeline view, as depicted in Figure 7.4. This timeline shows every inserted document as a bar (scaled to its text length and colored like its corresponding tree node). It indicates its insert certainty as a small square below. The darker such a certainty square, the higher the uncertainty of the insert. In addition, to provide provenance histories, we note all interaction events with icons at the bottom of the timeline. These events include the creation of a new topic , manual movement of nodes by the user , a node reinsert requested by the user , and the start of a new speculative execution cycle . Furthermore, all user-selected measures are visualized as line charts. Due to the changing scale of the measures during the incremental topic modeling process, all scales are automatically stabilized, normalizing the measures between 0% and 100%. To facilitate comparison, all measures are adjusted to be optimized towards 100%, i.e., a rising line for a measure indicates a quality improvement. One measure can be highlighted (as the tracked measure), fading out the color for the other lines. Hovering over any position on the timeline shows details of all the measured scores for the respective document associated with the selected timestamp. Lastly, as quality measurements are typically unstable for the first documents, we fade in the measures after the eighth document has been inserted to avoid misinterpretations. Through interaction with the MQM, users inherently perform a *multi-objective optimization* of all metrics to their *subjective notion of quality*.



7.5 | Speculative Execution

Speculative execution is a method that has long been used in software engineering and process optimization [OG88]. With our work, we showcase the application of SpecEx in a human-in-the-loop decision-making context using visual analytics. More specifically, for the use-case of topic model optimization, Lee et al. advise that “one possibility to allow users to manage cascading side effects is if the system can provide an estimate of the potential impacts of a refinement before it is applied.” [Lee+17b] Following their suggestion, we present a two-fold speculative execution within the model-building process of IHTM. First, the speculation shows the impact of different optimization strategies, allowing users to compare their effects before deciding on accepting one of them. Second, temporal speculation shows the further development of

different versions of the topic tree over several processing steps. Combining both preview mechanisms provides a powerful tool for efficient, mixed-initiative topic model optimization.

Such speculation can be manually triggered by the user at any point during the incremental topic modeling. However, for a more targeted optimization, we introduce an automatic triggering technique that prompts the topic modeling to halt when its quality deteriorates. The IHTM algorithm continuously assesses the change in its model quality after each insertion. As described in Algorithm 1, line 11, an optimization can be automatically triggered when a decline is detected. To measure this decline, we have defined and implemented 16 trigger strategies and evaluated their performance against a set of manually selected trigger points on two datasets. These triggers consider moving averages, as well as various groups of quality metrics. Some heuristics showed promising results, with both high precision and recall values for detecting a decline within a window of ± 2 documents from the manually-specified points. We, therefore, chose to combine the two best-performing heuristics as default triggers for SpecEx. These are “Majority of Visible Metrics Decreases” (MVM) and “Rapidly Falling Slope for Any Visible Metric” (RFa), reacting to quality declines of more than five percent each for all metrics in the timeline, and an increasing negative delta for any metric, respectively. In combination, they achieve an f-score of 0.706 (precision of 0.667 and recall of 0.750) when tested against the manually annotated corpora.

When triggered, the speculative execution stops the algorithm and starts optimizing parallel copies of the current model state. Hence, at the time of speculation, there are multiple model candidates available, as highlighted in Figure 7.3. These candidates are generated by applying different optimization strategies on copies of the latest topic tree, as introduced in Section 7.5.1. Simultaneously, the visual analytics workspace switches from operation mode to speculation mode, activating comparative instances of the different visualization components, as described in Section 7.5.2. The user is then prompted to compare the different optimized models to choose one of them. When the user accepts a certain optimized model, the current model instance is replaced with the chosen model. Our system guides users by ranking the optimizations according to their potential to improve the current model quality based on all measures (or optionally on the tracked metric). However, choosing the optimization that yields the highest measured improvement does not necessarily coincide with the users’ perceived model quality, so exploration is suggested.

7.5.1 | Automatic Optimization Strategies

Our approach implements six optimization strategies, as shown in Figure 7.4. These have been developed based on observations of manual optimization intentions. Each of these strategies gets automatically applied to a copy of the current IHTM model. The visual analytics workspace also includes the default IHTM model to allow the users to compare the temporal speculation

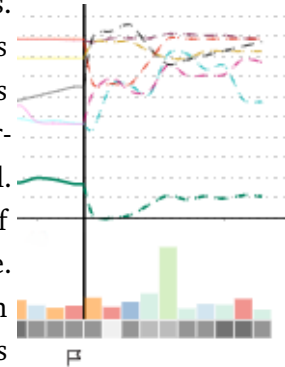
of keeping the current model running without optimization. In the following, we explain all the options users can compare and choose from during speculation mode:

- (0) **Default IHTM**: Shows the future of the current model without adjustments.
- (1) **Combine Similar Topics**: Combines the most similar topics by moving them to a common parent node.
- (2) **Reinsert Worst Topics**: Identifies the *worst* topic under the currently tracked metric and redistributes its documents.
- (3) **Reinsert Small Topics**: Reinserts all documents belonging to small topics, i.e., ones with three or fewer documents.
- (4) **Outlier Reinsertion**: Identifies and reinserts *outlier-nodes* with the algorithm used when creating the *squiggly lines* in the visualization.
- (5) **Split Largest Topic**: Splits the largest topic (based on the number of documents) into two new topics.
- (6) **Remove Topic Chains**: Identifies *topic-chains*, a clustering artifact, and moves the affected nodes to a common parent.

7.5.2 | Comparative Speculation Mode

In contrast to the operation mode (Section 7.4), the main focus of the *speculation mode* is to facilitate *comparative* analysis. Users can efficiently *flip through*, compare, and interact with multiple optimized models without any of these candidates affecting the model-building process, unless explicitly confirmed. This section describes the transformation of the topic tree and quality timeline from single-model visualizations into comparative views.

Timeline-Speculation View Whenever a speculative execution is triggered, the timeline shows a preview of the effects of choosing one of the optimization strategies. In addition, supporting the two-fold speculation, the timeline also shows how the measures would be affected over the next ten insert operations for each optimization. This preview relies on loading the next ten (buffer-size) documents from the IHTM-buffer, for each optimization, in parallel. For a more distant look into the future, users can request more batches of ten documents from the buffer (by clicking ») or increase the buffer size. As seen in Figure 7.4, all speculative measure developments are shown after the trigger line (which indicates the time point the algorithm was stopped). These measure indicators get updated whenever the user selects another optimization strategy to compare. Hence, this view allows users to estimate the effect an optimization will have on every measured quality metric before accepting such an optimization.



Tree-Speculation View Comparing two topic trees in the same view is a challenging task, as attested by previous approaches, e.g., TreeVersity2 [Gue+13] or CandidTree [Lee+07].

In contrast to methods proposing juxtaposed views [Mun+03], we aimed at designing a comparative tree visualization that comprehensively highlights the differences between two models within the same topic tree to reduce the cognitive load, as advised by [Gle18].

As shown in Figure 7.7, our proposed *tree-speculation view* highlights the *changes* an optimization strategy introduces to the current model-state. Such differences include: split, merged, deleted, or newly added topic branches; changes on the document level; or keyword changes. **Deleted branches** are blurred, while **moved** (from a deleted branch), **newly added** (including ‘future documents’ from the temporal speculation) and **removed** branches and keywords are highlighted, using different colors, respectively. By visualizing model uncertainties and low-quality topics, we enable the users to foster trust in the model and empower them to directly address these shortcomings. This design focuses the attention of the users on the newly created (optimized) topic tree while promoting the tracking and spotting of changes. To further aid the users’ understanding of the ongoing tree transformation, we designed a set of tailored interactions. Notably, hovering a deleted or a moved node highlights its new position in the tree and connects it with its original position. An example of such an interaction is shown in Figure 7.7, where the (old) topics of ‘Christian’ and ‘Atheism’ were merged into a new topic by moving their high-level topic nodes under a new, joint parent node. In order to focus the analysis task of this view on the comparison, most interactions from the topic tree component are disabled. Hence, users cannot perform any direct tree manipulations while in speculation mode. Once an optimization strategy is accepted, the tree-speculation view transitions into a topic tree view, and all interactions of the operation mode get enabled again.

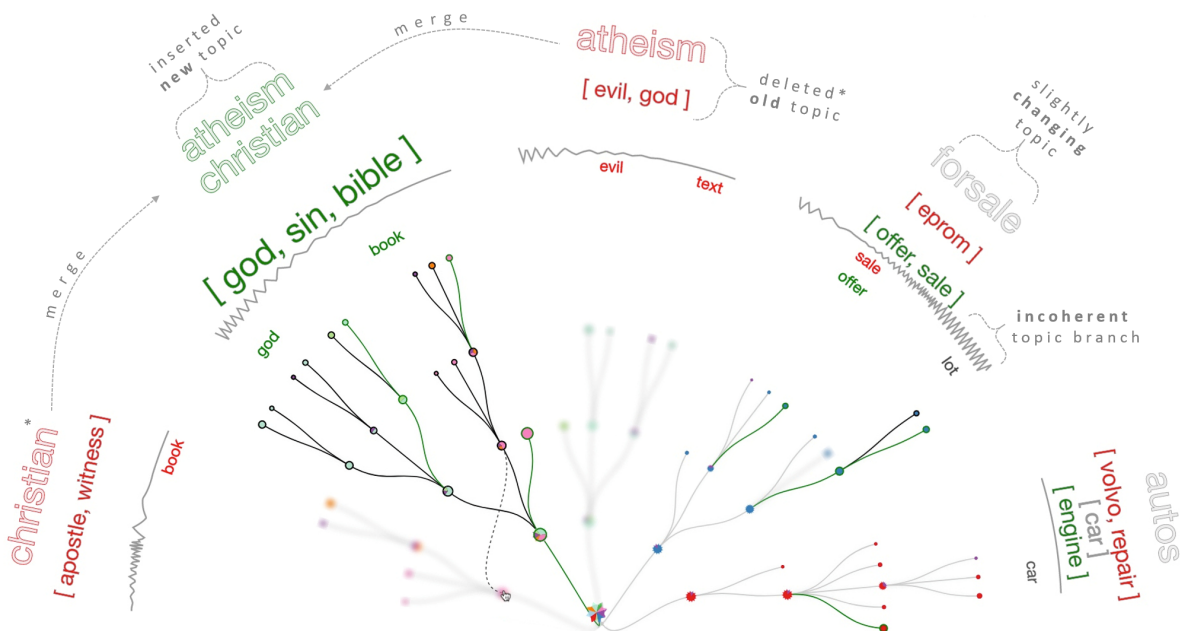


Fig. 7.7: The *Tree-Speculation View* is used to compare two topic models and shows the differences. Deleted branches are blurred, while moved, newly added and removed nodes and keywords are highlighted. To efficiently guide users towards perceivable model quality improvements, our system automatically proposes optimizations like the merge of two topics depicted here.

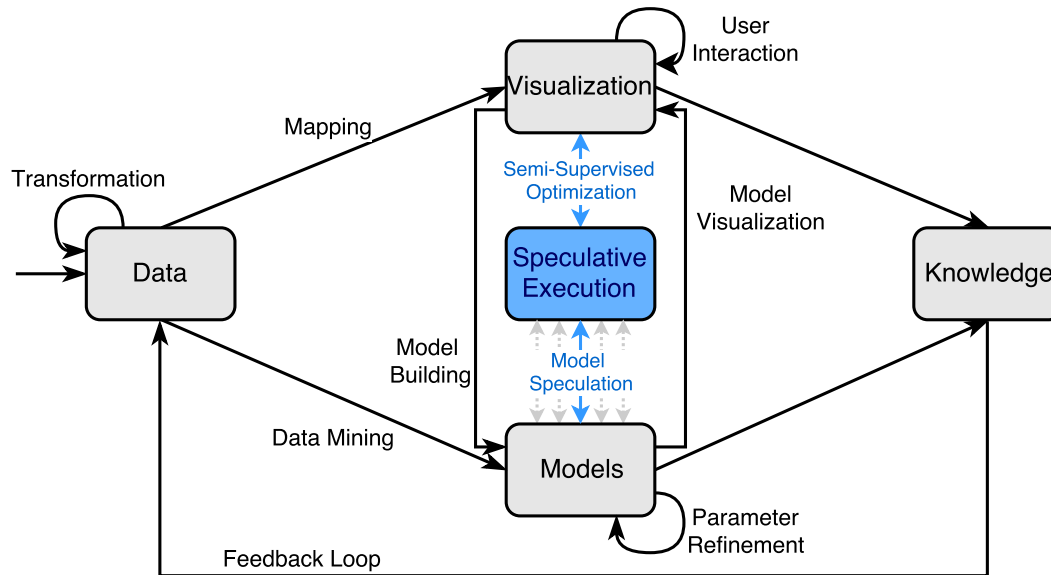


Fig. 7.8: Extension of VA-Pipeline by Keim et al.[Kei+08] to **SpexEx-VA**, with speculative execution as semi-supervised optimization technique.

SpecEx as a Visual Analytics Technique Our instantiation of SpecEx can be seen as an extension of the Visual Analytics Pipeline [Kei+08] into a mixed-initiative guidance approach. When there are modeling decisions to make or multiple models to consider, SpecEx provides an additional bridge between the models and the visualization, allowing for model speculation and leading to a semi-supervised optimization, as shown in Figure 7.8. Speculation is possible across several different analytic options, such as different algorithms, temporal predictions on streaming data, alternative parameter settings, or multiple runs of non-deterministic algorithms. Speculation may be useful in scenarios where deeper model understanding is required to build trust, such as in explainable machine learning scenarios. Thus, SpecEx can reveal the relevant, alternative pathways considered by a model to enhance user insight into its inner workings.

7.6 | Evaluation

Due to the modularity of our approach and the subjectivity of defining good-quality topic models, we chose to perform three independent evaluations, each focusing on one relevant aspect of model quality. The first study targets the evaluation of the IHTM algorithm, allowing us to establish a baseline for the other studies. The second study is a qualitative expert study with political scientists, linguists, and computer scientists. Lastly, the third study is a controlled quantitative study to assess the relative model quality improvement of using our system against the IHTM algorithm baseline.

7.6.1 | Qualitative Results: Expert Feedback

This study focused on participants using our approach to optimize a topic model. We recruited six different participants from different disciplines: two experts each from linguistics, political science, and computer science. Each participant completed a two-hour session in three parts: (1) interface explanation and first-impression feedback [30 mins], (2) using the visual workspace to optimize a topic model [60 mins], and (3) a semi-structured interview on the design and usability of the system [30 mins]. All participants performed their analysis on the transcript of the second US presidential debate between Romney and Obama in 2012. This dataset was chosen for the familiarity of its content and due to the corpus length (approx. 280 utterances). To set the same baseline for the model-building process, all participants began with predefined backbone priors that included all topics on the agenda of this debate [Uni12]. During all study sessions, the default automatic triggers for speculative execution were used. Finally, all sessions were audio-recorded, screen-captured, and logged for later analysis. In the remainder of this section, we report aggregated feedback gathered from all sessions combined. Figure 7.6 shows one example of a topic tree regenerated based on the settings of the study.

Initial Feedback After explaining the task of the study and the goals of our system, all participants immediately recognized the added value of a visually-explainable topic modeling process. In particular, the trust-building aspect was consistently mentioned in the feedback from most participants. Some of them rely heavily on topic modeling for their research, building complex statistical models based on the results of topic models. For example, *Pol2* reported that “[he] usually invest[s] an extensive amount of time to manually refine the results of the topic modeling, because [he] has to ensure that the foundation of [his] statistical computation is solid to avoid wrong inferences.” He continued that “[he] usually know[s] exactly which topic distribution to expect for [his] datasets, as [he is] the one collecting the [experimental] data.” In particular, this participant was enthusiastic about the idea of setting backbone priors for guiding the algorithm. The same general sentiment was shared by other participants who agreed that setting a “semantic frame” (*Ling2*) to the analysis is intuitive and is expected to yield an improvement in the modeling results.

Observations During the Optimization Process We witnessed different patterns across the groups of participants. For example, the linguists were tempted to analyze the keyword distributions in the topic tree more closely than other participants. This resulted in them using direct tree manipulations to group different keyword clusters at the beginning of the optimization process, extending the idea of priming the algorithm with semantic knowledge. They also reported relying on the document log and the topology of the topic tree to judge the quality of the model. In contrast, the two computer scientists were much more focused on choosing optimizations that did not negatively affect the measured model quality. One of them admitted at the beginning of the session that he would trust the system recommendations more than his own intuition. However, through interacting with the system and understanding

User Rank	Backbone	SpecExec	Measured Rank	Improvement
1.00 (0.00)	yes	supervised	2.00 (1.00)	67.72% (35.98)
3.00 (1.00)	yes	no	2.75 (0.97)	67.01% (65.84)
3.83 (1.64)	yes	automatic	1.67 (1.15)	91.75% (75.36)
4.17 (1.11)	no	supervised	4.33 (0.58)	9.65% (2.04)
4.42 (1.56)	no	no	6.00 (0.00)	0.00% (0.00)
4.83 (0.94)	no	automatic	4.00 (1.73)	14.6% (11.05)

Table 7.2: Result quality for six different models, averaged over three corpora. Users ranked the models (scale: 1–best, 6–worst) according to their perception of quality. The measured rank is calculated based on the improvements in quality metric scores. For each value, the standard deviation across the three corpora is shown in parentheses.

the model decision-making process, this model expert became more skeptical and started comparing different optimizations based on his understanding of the data. He later reported that “[he] would have more confidence in the results of such a transparent process that [he] can understand than in a hidden model.” Lastly, a general pattern observed in the analysis of the political scientists is that they both were very keen on understanding the specifics of the quality measurement to select “some important ones” to focus on. During the analysis session with one of them, we observed that he, too, relied on the automatic triggers to stop and judge the model quality. However, in contrast to the computer scientists, his choice of an optimization strategy relied mostly on a concrete action plan which he derived from analyzing and comparing the topic trees. While these usage patterns are not indicative or generalizable to research fields or broad user groups, observing the differences showed the broad applicability of our system. Regardless of the analysis strategies, they all reported perceiving an increase in the model quality, even when some of the quality metrics decreased. Furthermore, they praised the rich set of interactions provided in the technique and the amount of details provided.

General Assessment After going through a topic model optimization process, all users confirmed that they perceived an improvement in the model quality. When asked whether they performed the three tasks targeted by our technique, all of them confirmed that they did. In addition, some of them said that they also did an exploratory analysis and gained insight into both the dataset and the modeling process. They additionally stated that they had built trust in the system: not only did they learn to rely on its guidance (requesting feedback when the model quality declines), but most of them also learned to have more confidence in judging when the algorithm is making wrong decisions (e.g., because of missing semantics). They all approved of the SpecEx process, indicating that it is intuitive to use and that it helped them in making informed decisions. Furthermore, some model experts also suggested additional requested features to be added to the technique. Notably, a score measuring the impact of the topic backbone on the model result, a score indicating the differences between two topic trees in speculation mode, and an option to jump back in time and change previous decisions, allowing to undo operations. All of these suggestions are useful extensions to our approach.

7.6.2 | Quantitative Results: Model Quality Assessment

In preparation for our third study, a topic modeling expert ran six different configurations of the modeling process for each input corpus. Three of the configurations used a topic backbone, while the other three did not. Within the two groups, speculative execution was once disabled and once performed automatically. “Automatically”, here, means that the top-rated strategy offered by the system was always accepted. The last two configurations used supervised SpecEx, meaning that the expert inspected the proposed optimizations and selected the most fitting one, performing a semi-supervised optimization.

We chose three different datasets for this study: (1) the presidential debate from the first study, (2) a collection of 125 newsgroup articles from the 20news dataset [Lan], and (3) 100 news articles from the Associated Press [Ble]. The insert order of the documents in both news corpora was determined by randomly shuffling their documents. For each corpus, we let four annotators (two topic modeling experts and two novice users) rank the model outputs created by the expert on a scale from one (best) to six (worst), leading to 72 annotations. The results, presented in Table 7.2, are averaged over the three corpora, sorted to the annotators’ ranking, and show that annotators rate models with a backbone higher than those without. The quality metrics confirm this, reporting higher increases in quality for models with a backbone when compared to an IHTM baseline with no backbone and no SpecEx. Independently of whether a backbone was provided or not, users preferred models with supervised SpecEx over no SpecEx, ranking them 1.0 and 3.0, and 4.17 and 4.42, respectively. They even found that automatic SpecEx deteriorated the model quality and consistently ranked it worse than what the baseline IHTM returned, indicating that optimizing according to numerical values without semantics does not work with current quality measures. Contrary to the users’ perception of quality, the metrics improve most under automatic SpecEx (91.75% and 14.6%). Nonetheless, the metrics also indicate most improvement of quality through supervised SpecEx when compared to models with no SpecEx, finding a 9.65% improvement if no backbone is specified.

7.6.3 | Lessons Learned

Through evaluating our work with three independent measures, we confirmed the competitiveness of IHTM and the effectiveness of our visual analytics workspace. Using these different evaluations, we obtained diverse perspectives on relevant quality criteria. Our quantitative evaluation has established that the quality of topic modeling outputs based on a backbone structure was better received. Furthermore, our evaluation shows that annotators favor optimized topic structures that are computed using a supervised speculative execution over fully-automated optimizations. This, however, stands in contrast to the automatically computed quality measures that indicate that automatic SpecEx yields better results. More details on the differences in the measured quality over the three configurations of SpecEx can be found in Table 7.3. The results indicate that no single metric perfectly aligns with the human ranking,

Corpus	Presidential Debate [Uni12]			AP Corpus [Ble]			20news Corpus [Lan]		
SpecExec	no	auto.	superv.	no	auto.	superv.	no	auto.	superv.
User Rank	3 (0.8)	2.75 (0.96)	1 (0)	3 (1.41)	4.5 (1.91)	1 (0)	2.25 (0.5)	4.25 (1.71)	1 (0)
Avg. Change	16.2	22.74	32.50	141.38	172.16	104.41	43.50	80.35	66.24
Coherence	+5.95	+10.34	+26.40	-33.07	-54.65	-24.58	+11.37	+22.91	+13.80
Separation	+0.86	-12.73	-16.02	+21.79	+39.18	+1.45	+2.44	+4.21	-0.29
Distinct.	+7.24	-2.10	+14.10	-2.63	-10.00	-18.56	+15.59	+35.18	+15.68
PMI	+8.44	-4.20	+1.89	-31.45	-45.10	-28.23	-0.53	+6.68	+2.05
Certainty	+5.19	+0.85	-4.65	-38.97	-41.33	-37.65	-7.87	-9.04	-9.85
Branching	+29.82	+54.09	+68.10	+492.52	+457.14	+297.96	+135.93	+231.02	+206.94
Compact.	+19.76	+1.79	+18.20	+206.52	+150.09	+148.77	+14.59	+89.57	+59.04
Topic Size	+6.16	+51.37	+65.13	+333.33	+550.00	+286.10	+157.74	+236.30	+210.87
Speakers	+63.64	+80.30	+87.33	+368.00	+582.50	+317.30	+67.14	+114.73	+97.32
Topics	0	-29.41	-35.29	-76.92	-84.62	-74.36	-61.54	-69.23	-69.23

Table 7.3: Relative quality metric change when applying a backbone and using SpecEx, compared to the original IHTM result as a baseline. The “user rank” (scale: 1–best, 6–worst) is not predicted by any individual quality metric and does not align with the measured improvements.

validating our claim that users should perform a *multi-objective optimization* instead of focusing on a single metric. Often, metrics track opposing goals: Coherence, for example, is typically high for very small topics and can achieve its best possible score when attributing all documents to their own topic. Topic size, in contrast, aims for large, general topics that provide a good overview. Both metrics can never be optimal at the same time, and it is the user’s task to select a middle-ground appropriate for their use case.

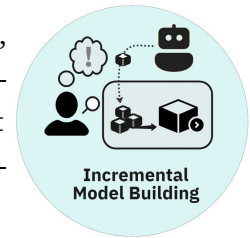
In addition, our qualitative study highlights the effectiveness of speculative execution in a visual analytics system as a mechanism for user guidance and trust-building. Interviewing users with different backgrounds displayed many different usage patterns of our visual analytics technique. However, regardless of their expertise discipline, all users approved of our incremental, animated visualization of the model-building process. Almost all users praised the different possibilities for domain-knowledge incorporation, making use of all three methods (backbone priors, direct topic tree manipulations, and automatic optimizations) during the model optimization sessions of the studies. Such an explainable machine learning technique not only supports users in adapting the models to their tasks and data but also opens up a venue for using explorative visual analytics for education.

7.7 | Discussion: Investigating I&XAI Design Patterns

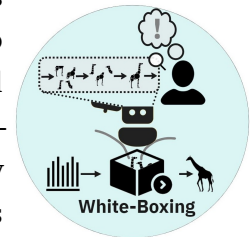
Based on the presented technique, in this section, we discuss three potentially generalizable I&XAI design patterns. These are tailored to interactive and explainable applications for

the *model experts* in the **Model Space**. An overview of all design patterns discussed in this dissertation can be found in Section 8.4

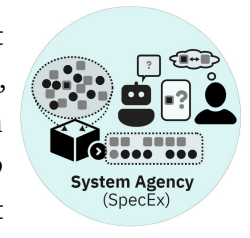
(4a) Incremental Model Building To allow users to understand, diagnose, and refine all aspects of the model’s decision-making, we ought to enable them to inspect its learning process. Hence, one crucial aspect that facilitates model-aware explanations is to incrementally build models, visualizing all intermediate decisions and changes during training.



(4b) White-Boxing In addition to an inherently explainable model, it is essential to show the model’s internal metrics and decision certainties to facilitate transparency. This will enable model experts to keep up and follow along with the modeling process while inspecting the model’s objectives during training. Note that usually, the modeling objectives are only approximations of the human’s intuition and understanding. Hence, it is vital to allow users to examine when their expectations differ from the model’s internal metrics.



(4c) System Agency To enable an efficient modeling process, we cannot rely on humans to control the modeling and feedback process. Therefore, it is advisable to aim for a system agency, i.e., the system requests human intervention and feedback when it is uncertain. Such a *human-in-the-loop* process can further be improved by designing appropriate decision-support mechanisms for the users, e.g., through *speculative execution*.



Closing

” *Truth is much too complicated
to allow anything but approximations.*

— **John von Neumann**
(as paraphrased by Farah [Far14])

Summary

” *The idea that the future is unpredictable is undermined every day by the ease with which the past is explained.*

— **Daniel Kahneman**
Thinking, Fast and Slow [Kah11]

This dissertation explored the research question of *how to empower different stakeholders to interact with machine learning models for collaborative problem-solving and decision-making*. It addressed it by mapping out different *levels of explainability for human-AI interaction in visual text analytics*, as its title suggests. In particular, the presented framework was framed in a theoretical chapter. This included a general design space spanning eight dimensions, a formalization of the explainability aspects from the design space into the *explAIner* pipeline with five properties, and a conceptual framework encompassing three *explainability spaces*. The main content chapters of the dissertation instantiated the conceptual framework for topic and content model refinement. Each of these four chapters presented a novel visual analytics technique tailored to a specific stakeholder. All techniques were designed to tackle the tasks of *understanding, diagnosis, and refinement of content and topic models*.

Contents

8.1	The Design Space of I&XAI Applications	204
8.1.1	I&XAI Design Dimensions	204
8.1.2	<i>explAIner</i> Properties	204
8.1.3	Explainability Considerations	205
8.2	The Explainability Spaces Framework	205
8.2.1	The Semantic Space	205
8.2.2	The Input-Output Space	205
8.2.3	The Model Space	206
8.3	I&XAI Techniques for Content & Topic Modeling	206
8.3.1	Designing for Domain Experts: <i>Semantic Concept Spaces</i>	206
8.3.2	Designing for Domain Analysts: <i>Progressive Learning</i>	207
8.3.3	Designing for Model Analysts: <i>ThreadReconstructor</i>	208
8.3.4	Designing for Model Experts: <i>Speculative Execution</i>	208
8.4	I&XAI Design Patterns	208

8.1 | The Design Space of I&XAI Applications

In this dissertation, we discussed different aspects to consider for interactive and explainable machine learning. These form a comprehensive design space and inform different levels of explainability that build the foundation of our conceptual framework. This section recaps the three aspects that build the design space of I&XAI applications, as depicted in Figure 8.1.

8.1.1 | I&XAI Design Dimensions

The first set of aspects is the I&XAI design dimensions introduced in Section 3.2. These comprise the following eight dimensions: (D1) Users and Stakeholders; (D2) Analytical Tasks and Application Domains; (D3) Data Characteristics; (D4) Types and Classes of ML Models; (D5) Stages in the KDD Process; (D6) Explainability Goals and Values; (D7) Explainability Methods and Processes; and (D8) Explainability Strategies and Mediums.

These dimensions describe aspects related to the domain and application context (D1-3), the used machine learning process (D4,5), and the explainability that mediates between them (D6-8). The primary contribution associated with introducing and describing these I&XAI dimensions is to enable researchers and developers to situate their approaches within the design space. This will facilitate the comparison of works across the field and accelerate the distillation of best practices. Doing so can, over time, enable the I&XAI field to foster structured methodologies, as discussed further in Section 10.2.2.

8.1.2 | *explAIner* Properties

Next, to operationalize the explainability dimensions described above, in Section 3.3, we introduce the *explAIner* pipeline. This describes how I&XAI systems can modularize their explainability and interaction components into, so-called, *explAIners*. In addition, we derive five properties for *explAIners*, namely: (P1) Number of Model States; (P2) Parts of the Model

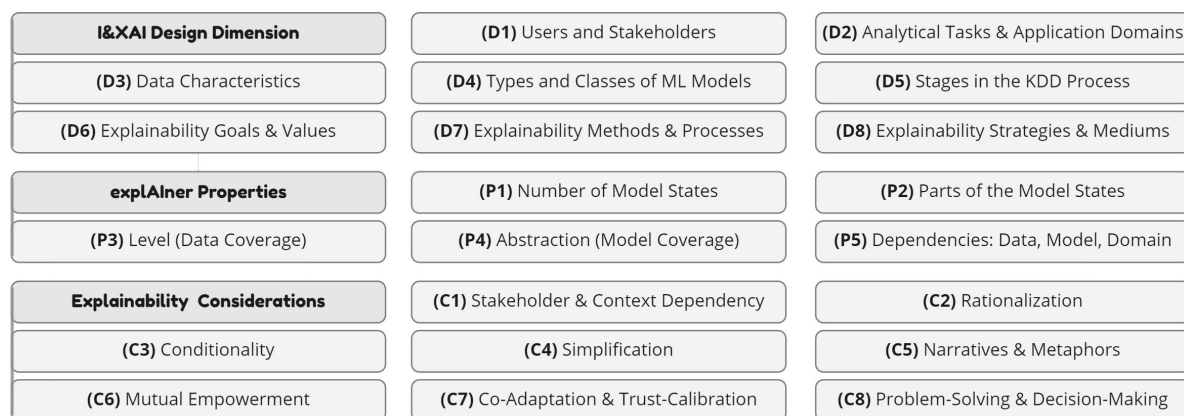


Fig. 8.1: The design space for I&XAI applications as proposed in this dissertation.

States; (P3) Level (Data Coverage); (P4) Abstraction (Model Coverage); and (P5) Dependencies (Data, Model, Domain). Using these pipelines, researchers and developers can discuss their design decisions and define modular interfaces for I&XAI components.

8.1.3 | Explainability Considerations

Lastly, we introduced eight explainability considerations derived from surveying the literature in Section 2.1.3. They describe aspects that impact the effectiveness of human-AI interaction, namely: (C1) Stakeholder and Context Dependency; (C2) Rationalization; (C3) Conditionality; (C4) Simplification; (C5) Narratives and Metaphors; (C6) Mutual Empowerment; (C7) Co-Adaptation and Trust-Calibration; and (C8) Problem-Solving and Decision-Making. We utilized the trade-offs described in these considerations to inform the design of the presented I&XAI workflows and techniques in the main content chapters of this dissertation.

8.2 | The Explainability Spaces Framework

One of the main contributions of this dissertation is the *Explainability Spaces* conceptual framework, described in Section 3.4. It argues for the deliberate design of tailored I&XAI workflows depending on the target user groups. This framework describes three distinct design spaces for I&XAI applications. However, these spaces form a spectrum of possible designs and techniques that can fall at any place on this spectrum, utilizing a mixture of aspects from the different spaces. In the following, we summarize the three spaces of the framework.

8.2.1 | The Semantic Space

As depicted in Figure 8.2, the first space is the **Semantic Space**. It is tailored to *domain experts* and enables them to externalize their domain knowledge and understanding. The main mental model we ascribe to this stakeholder group is that they aspire to *see the big picture* when solving their abstract domain problems. They typically work in a *theory-driven, top-down manner* and thus prefer *model-agnostic explanations* to support their *deductive reasoning*. This is usually enabled through a *concept-level interaction*, for example, based on *ontologies*. These allow for an *explicit knowledge representation*. More details on this space are discussed in Section 3.4.1.

8.2.2 | The Input-Output Space

The second space, see Figure 8.2, is the **Input-Output Space**. It is the most common in practice and is tailored to both *domain and model analysts*. This space is designed for concrete problem-solving and decision-making. Hence, the main mental model attributed to its stakeholder groups is that they would like to *focus on the task and data issues* at hand. Therefore, they typically work in a *task-driven manner*, preferring *counterfactual explanations* to support their

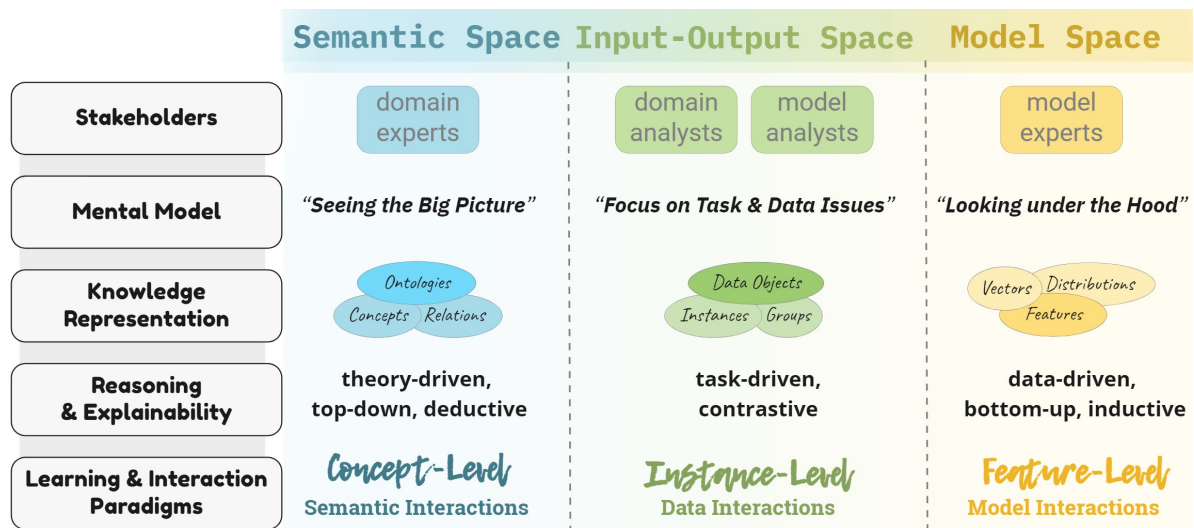


Fig. 8.2: Summary of the *Explainability Spaces* conceptual framework, describing five aspects for every one of the three spaces, as discussed in Section 3.4.

contrastive reasoning. This is usually enabled through a *instance-level interaction*, for example, based on *data objects and groups*. More details on this space are discussed in Section 3.4.2.

8.2.3 | The Model Space

As depicted in Figure 8.2, the third space is the **Model Space**. It is tailored to *model experts* and enables them to examine the inner workings of ML models. Thus, their primary mental model is to *look under the hood* of the model to observe its problem-solving and decision-making processes. Model experts typically work in a *data-driven, bottom-up manner* and thus prefer *model-aware explanations* to support their *inductive reasoning*. This is usually enabled through a *feature-level interaction*, for example, based on *vectorizations and distributions*. These allow for an *implicit knowledge representation*. More details on this space are discussed in Section 3.4.3.

8.3 | I&XAI Techniques for Content & Topic Modeling

Based on the described design space and conceptual framework, in the main content chapters of this dissertation, we introduced four I&XAI techniques for content and topic model refinement. In the following, we briefly summarize these approaches and their contributions. For each of the four techniques, Figure 8.3 depicts a process model describing the suggested I&XAI workflow, the visual analytics pipeline of the application, and the proposed interface design.

8.3.1 | Designing for Domain Experts: *Semantic Concept Spaces*

Domain experts usually detect implicit relations between objects and attributes that they strive to incorporate into the ML models. However, typically, providing feedback is constrained by

algorithmic parameters and is challenging for non-machine-learning experts. As described in Chapter 4, *Semantic Concept Spaces* contribute an approach to bring the expression of user semantics *closer* to the actual ML model. Users can see the *big picture* of the concept space and generate new ground truth data through their interactions and knowledge externalization, a form of *machine teaching*. The modifications applied to the concept space become transferable knowledge that can be used to initiate models used on other data. Our proposed design is based on two parallel hierarchies, one for user-defined concepts and another for model-derived topics. The interactions take place in the *Semantic Space* and are thus *model-agnostic*.

8.3.2 | Designing for Domain Analysts: *Progressive Learning*

Domain analysts typically focus on a concrete domain task and data, aiming to use them to perform a targeted adaptation of ML models. In Chapter 5, we propose a *progressive learning* technique to refine topic modeling parameters. Our technique supports a layered analysis based on the data and tasks of the users. This stepwise analysis is provided through a visual analysis workspace, consisting of four visualizations that are tailored to the analysis tasks of topic matching, topic summarization, parameter distribution analysis, and document relevance feedback. The workspace is backed up by a *reinforcement learning feedback loop* enabling users to optimize topic models and obtain more easily understandable results. This technique is designed in the *Input-Output Space* with a focus on the domain semantics.

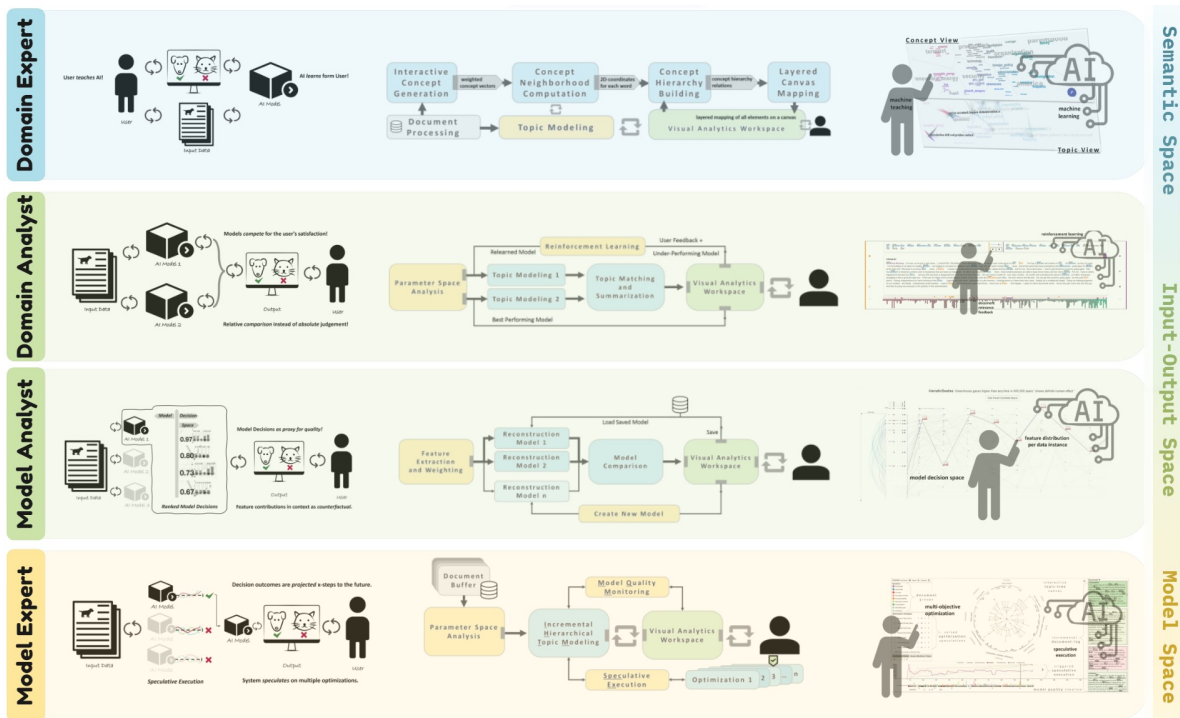


Fig. 8.3: Overview of the four I&XAI techniques presented in this dissertation with different I&XAI workflows, visual analytics pipelines, as well as interface and interaction designs.

8.3.3 | Designing for Model Analysts: *ThreadReconstructor*

Model analysts also focus on targeted problem-solving and decision-making but with a stronger emphasis on the ML model. In Chapter 6, we presented *ThreadReconstructor*, an I&XAI technique for reconstructing threads, untangling conversations, and understanding model decisions. This approach is based on a set of tailored features that lay the foundation for *user-driven model generation*. Models iteratively created with our tool are customized to the data and tasks of the users. In the visualization interface, all created models can be compared and optimized using four different views. Through revealing the *decision space* of a model, we allow the comparison of its internal workings, while simplifying its complexity. This technique is designed in the **Input-Output Space** with a focus on the ML model.

8.3.4 | Designing for Model Experts: *Speculative Execution*

Lastly, for *model experts* we proposed a tailored I&XAI technique for topic model optimization that reveals its inner workings. The *speculative execution* approach proposed in Chapter 7 explains and visualizes every step of the model decision-making. This allows for a tight integration of users' feedback and understanding into the ML process. We introduced a custom visual analytics workspace that interactively displays all intermediate results of the topic modeling, allowing users to understand, diagnose, and refine them. In addition to direct manipulations of the built topic tree, our system enables users to *prime the topic modeling algorithm with expected outputs*, integrating their expected data semantics into the modeling process. For targeted optimization, we further introduce *speculative execution* as a novel interaction paradigm that acts as a *preview mechanism for efficient user-steerable optimization*. The interactions and explanations are *model-aware* and take place in the **Model Space**.

8.4 | I&XAI Design Patterns

Based on the proposed techniques, we observed several potentially generalizable I&XAI design patterns. These are discussed at the end of each of the four content chapters. An overview of all design patterns is depicted in Figure 8.4.

For the *domain experts*, in Section 4.7, we propose to rely on a *machine teaching* paradigm for interaction; to keep the *user's and system's knowledge represented separately* to allow for effective interaction mapping, and to allow for an AI-in-the-loop analysis through giving users the agency to *request guidance* when they see fit.

For *domain analysts*, in Section 5.7, we propose meaningfully initializing the workspace to *avoid cold start problems*; to enable *comparative analysis* on an object-level; and to use *visual anchoring* as a strategy to avoid distraction by context-switching.

For *model analysts*, in Section 6.9, we propose model generation through a *visual, no-code interface*; to enable *comparative analysis* on a feature-level; and to reveal *model decision considerations* for a more fine-grained inspection.

Lastly, for the *model experts*, in Section 7.7, we propose *incremental model building* to enable *white-boxing* the model and reveal its inner workings. In addition, we propose facilitating *system agency* to request user feedback only when needed in a human-in-the-loop manner.

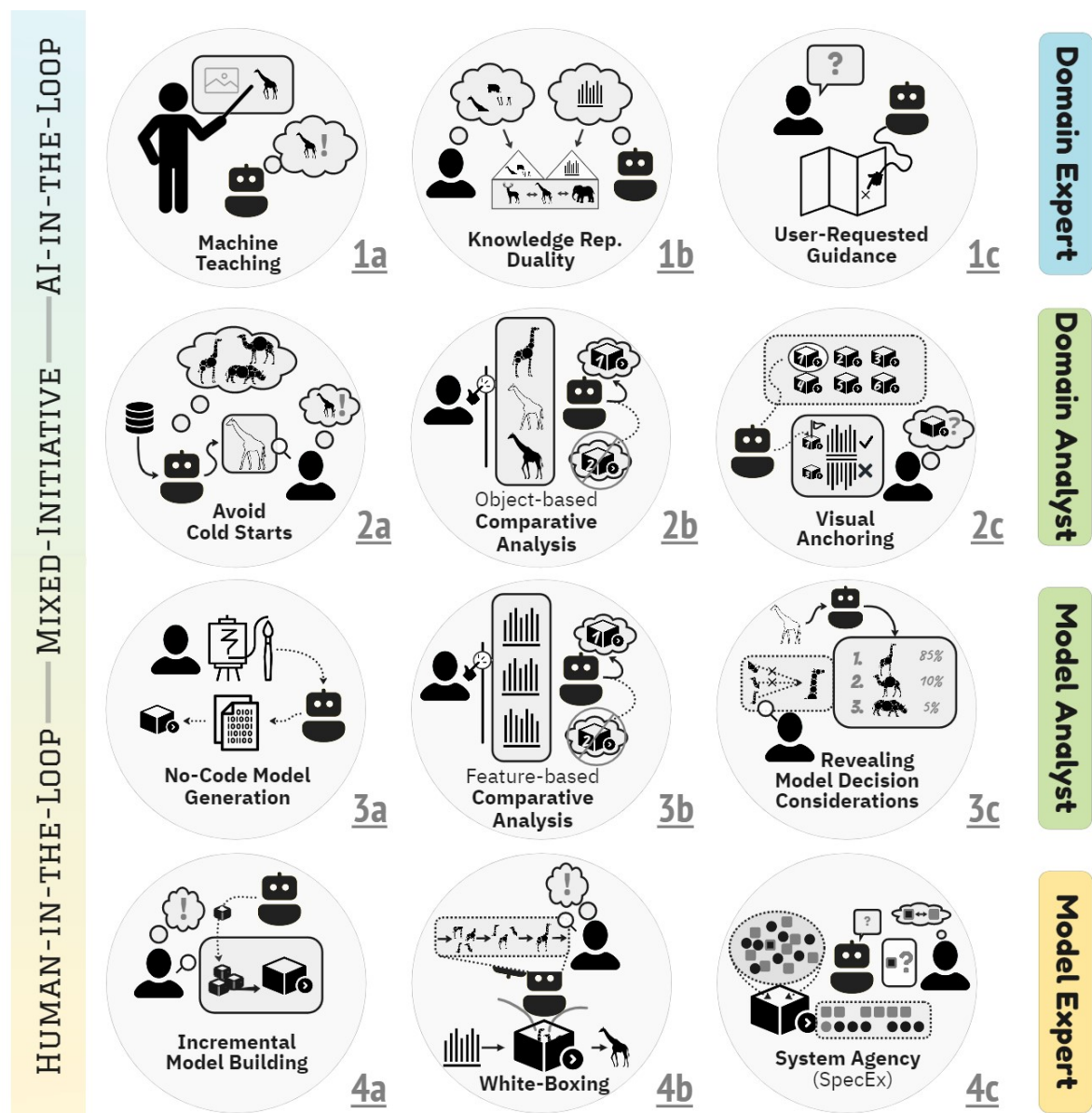


Fig. 8.4: Overview of I&XAI design patterns grouped by the four content chapters.

Discussion

” *The only true wisdom is in knowing you know nothing.*

— **Socrates**
(attributed)

In this dissertation, we have presented four general contributions (1) a comprehensive description of the design space of I&XAI; (2) the conceptual explainability spaces framework; (3) four distinct visual analytics techniques for content and topic model refinement; and (4) twelve generalizable design patterns for I&XAI derived from the best practices of our proposed applications. With these contributions, we aim to make a step towards enabling *seamless human-AI interaction for collaborative problem-solving and decision-making*. In this discussion chapter, we examine the limitations of our approach, highlight lessons learned, and point out open research questions and challenges.

Contents

9.1	Limitations	212
9.1.1	I&XAI Framework and Design Space Methodology	212
9.1.2	Evaluation and Generalizability	212
9.1.3	Targeted Human Modeling for Personalization and Guidance	213
9.2	Lessons Learned	213
9.2.1	I&XAI Workflows for Intelligence Augmentation	213
9.2.2	Co-Adaptive Analytics: I&XAI as Dynamic Systems	214
9.2.3	Reasoning and Knowledge Representation in Human-AI Communication	216
9.3	Open Research Questions and Challenges	216
9.3.1	Detecting and Mitigating Human-AI Miscommunication Issues	216
9.3.2	Interaction Intent Identification to Bridge Knowledge Gaps	217
9.3.3	Trustworthy Decision-Making Support	217

9.1 | Limitations

This dissertation proposes a broad set of concepts and techniques. We had to navigate several tradeoffs to find a reasonable scope for the work presented. Thus, this section discusses three main limitations of our work, paving the way for future research opportunities.

9.1.1 | I&XAI Framework and Design Space Methodology

Due to the rapid evolvement of the field of I&XAI, there is still no consensus in the literature about the design space of this young research area. Thus, by consolidating proposed theories and concepts, we derived the I&XAI framework and design space, presented in Chapter 3. These enable us to structure and position I&XAI techniques to compare them and derive design patterns and best practices. However, some of the proposed aspects and dimensions could be refined and adjusted after being scrutinized and applied to a broader set of tools and applications. In addition, we carefully considered the names of all the components presented. However, some name changes might lead to refining our understanding of their underlying concepts. For example, the explainability spaces are spanned from domain semantics to the AI-model semantics. We named the outer spaces **Semantic Space** and **Model Space** but could, for instance, have chosen the name *Domain Space* instead of **Semantic Space**. We ended up with the current name as an homage to the concept of *Semantic Interactions* [End+15].

Overall, in our future work, we aim to further audit, refine, and verify the claims, concepts, and names for our I&XAI framework and design space. This will be achieved by surveying I&XAI techniques to position them in our framework and iteratively refine it, as well as conducting expert workshops and interviews to solidify our findings. In Section 10.2.2, we discuss our future research direction of developing a robust and structured methodology for I&XAI.

9.1.2 | Evaluation and Generalizability

Evaluating I&XAI approaches is a notoriously challenging issue [CS19] as these techniques are usually complex constructs that combine innovations in multiple pipeline stages [Spe+21a]. As depicted in Figure 9.1, there are various objectives and desired outcomes from an I&XAI workflow. We would like to achieve *accurate modeling* of the data, *high-fidelity explanations* of the model [WB19], *precise mapping* to the interface, and *effective interactions* to collect human feedback. Hence, in our presented techniques, we conducted multi-staged studies, each covering a specific aspect of the I&XAI workflow. These evaluations confirm our proposed techniques' effectiveness and applicability for refining content and topic models. However, they do not guarantee generalizability beyond this application domain. In addition, we evaluated the techniques without specifically recruiting representatives of the four stakeholder groups.

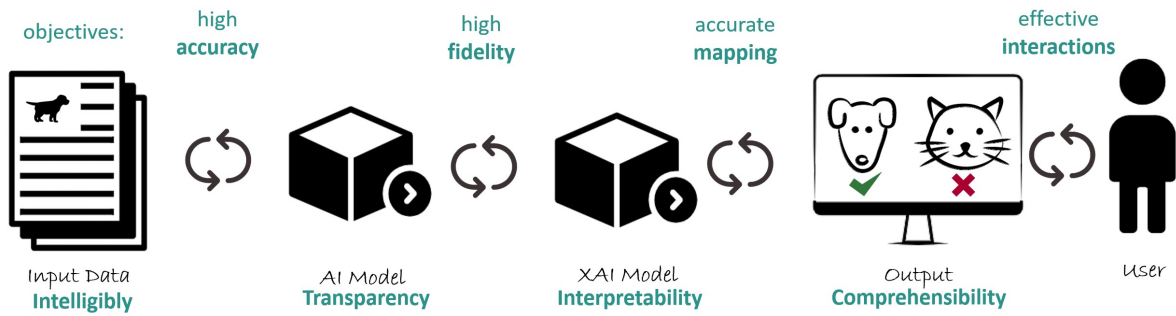


Fig. 9.1: I&XAI workflow objectives and desired outcomes.

Thus, in the future, we aim to develop a structured evaluation approach to compare across stakeholders, design patterns, and application domains. This will confirm the generalizability of our findings and contribute to establishing best practices and guidelines for this research field, as discussed further in Section 10.2.1.

9.1.3 | Targeted Human Modeling for Personalization & Guidance

This dissertation categorizes humans into four broad stakeholder groups, reflecting their knowledge, backgrounds, and expertise. However, for true intelligence augmentation, we must incorporate better *human modeling approaches* beyond our four stakeholder groups. For example, we need to consider other human factors, such as personality traits, learning effects, subjective preferences, etc. This opens up future research opportunities for contextualized and personalized analysis, as discussed in Section 9.2.2, as well as establishing structured human-AI communication methods to enable *seamless* collaboration, as illustrated in Section 10.2.3.

9.2 | Lessons Learned

Based on the work conducted for this dissertation, in this section, we summarize the main lessons learned to inspire future research directions.

9.2.1 | I&XAI Workflows for Intelligence Augmentation

In this dissertation, we proposed four adaptations to the traditional linear I&XAI workflow, e.g., as depicted in Figure 9.1. Each of the techniques we introduced in the content chapters proposes a novel I&XAI pipeline to allow for deductive, contrastive, or inductive analysis. These workflow innovations contributed to enabling our goal of intelligence augmentation for the different stakeholders. Hence, one of our main lessons learned is that ***tailoring I&XAI workflows is crucial for effective intelligence augmentation.***

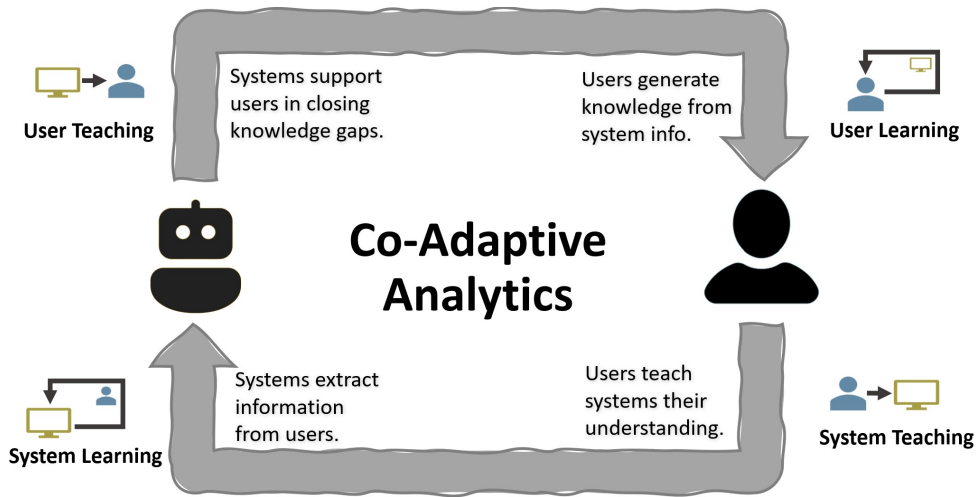


Fig. 9.2: Co-adaptive analytics enables dynamic analysis and knowledge development over time. It includes the dynamics of user teaching, user learning, system teaching, and system learning.

9.2.2 | Co-Adaptive Analytics: I&XAI as Dynamic Systems

The systems presented in this dissertation all progressively evolve their models throughout every analysis session. This temporal dynamic enables human-AI collaboration for joint problem-solving and decision-making. Hence, another lesson we learned is that ***I&XAI approaches must be developed as dynamic systems***. As such, we model the evolving knowledge of both the users and the AI models throughout an analysis session.

Generalizing this concept, we propose that I&XAI systems should conduct *co-adaptive analytics* [Spe+21b]. This entails modeling four interaction dynamics to reflect the evolving knowledge of the users and models. The dynamics reflect that both the user and the system are *learning from* and *teaching* each other. Hence, as depicted in Figure 9.2, we have to model systems to perform user teaching, user learning, system teaching, and system learning.

These interaction dynamics can be modeled implicitly through system design, as in our four techniques, or explicitly, i.e., modeling the learning and teaching evolution. However, one pressing challenge for such systems is to address the question of *how much agency each of the two actors should have*. In our proposed techniques, we presented approaches with varying human and system agency ratios, as illustrated in Figure 8.4. The span ranged from a higher human agency (AI-in-the-loop), through mixed-initiative techniques, to a higher system agency (human-in-the-loop). However, none of our tools had an evolving or adapting agency, for instance, starting with a higher human and lower system agency and evolving to progressively increase the system agency while decreasing the human one.

We presented such a *co-adaptive analytics* system for topic model refinement as a follow-up approach to this dissertation’s techniques. This I&XAI technique is designed in the **Input-Output Space** and relies on *co-adaptive guidance agents* [Spe+21c]. It reveals the model-internal multi-objective optimization process to the user through six single-objective agents. Each agent

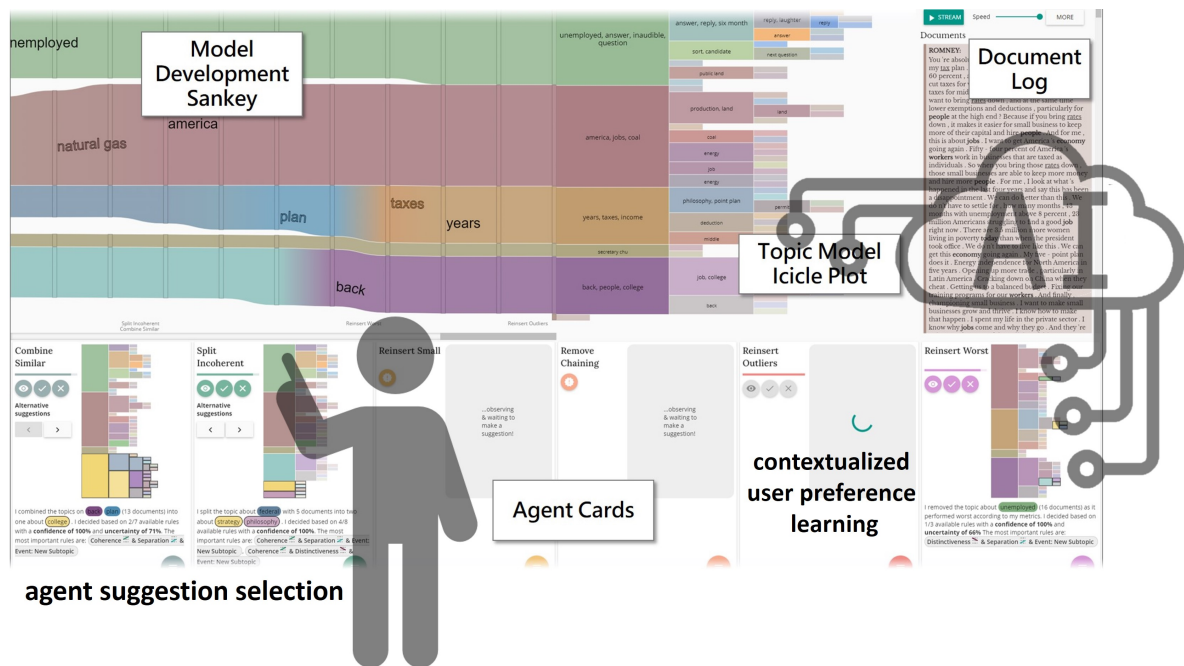


Fig. 9.3: Adaptive Guidance Agents [Spe+21c] of topic model optimizations for analysts. Each agent explains their decision-making using a verbalization of the main decisive factors and metrics that lead to the suggestion. The technique's I&XAI workflow is depicted in Figure 9.4.

(shown as an agent card on the interface, see Figure 9.2) is trained to optimize the topic tree based on a different objective. Agents use vectorizations of the analysis context to learn when and which of their suggestions leads to an optimized model. When an agent detects a situation where it can make a suggestion, its card gets populated with an explanation about the suggestion, the expected output, and the underlying metrics. This is presented both visually and verbally. The users can compare agent suggestions throughout the refinement process, accepting and rejecting them to influence the topic tree. Based on the user feedback, the system learns a contextualized user preference model to calibrate future agents' timing and suggestions. The *contrastive* I&XAI workflow, depicted in Figure 9.4, allows for agency evolution, where at the beginning of the analysis, humans are more in control until the agents learn their preferences and can start automating objectives with high model confidence.

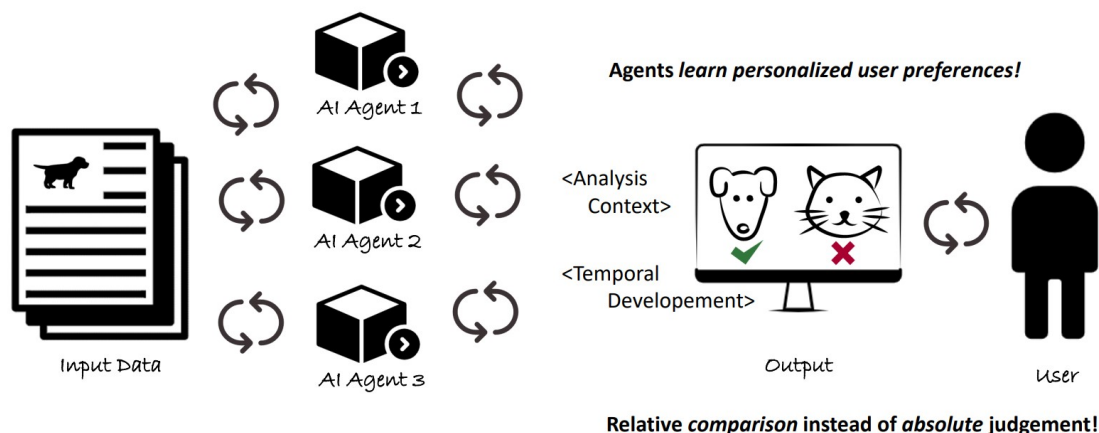


Fig. 9.4: This **contrastive** I&XAI workflow presents the users with *contextualized optimization suggestions* and learns user preference profiles based on their interactions.

9.2.3 | Reasoning and Knowledge Representation in Human-AI Communication

Communication is inherently about translation from one semantic to another; if both align, communication can become smooth and effective, but if they do not, it can be subject to miscommunication pitfalls. As opposed to human-human communication, human-AI communication translates between usually non-aligned semantics.

Therefore, it is crucial to consider the knowledge representation of the stakeholders interacting with an AI when modeling interactions and designing explainability. Hence, our last major lesson learned in this dissertation is to ***customize explainability and interaction mapping to the stakeholders' mental models***. These mental models, reasoning paradigms, as well as the resulting knowledge representations are described in detail in Section 3.4.

9.3 | Open Research Questions and Challenges

Based on the conducted research for this dissertation, we derived several findings as summarized in the previous sections. However, we also encountered several open research questions and challenges beyond the scope of this work. Hence, to promote future research opportunities and open up new frontiers, in this section, we discuss these open challenges.

9.3.1 | Detecting & Mitigating Human-AI Miscommunication Issues

As highlighted in the previous sections, human-AI communication is notoriously challenging. For instance, as described in Section 3.4.1, to model the *soft knowledge* of humans, we require semantically-accurate depictions, e.g., using *object-centric representations*. Hence, to avoid miscommunication issues, we ought to detect situations that require adjusting the explanation strategy or interaction modality used in communication. Some open questions regarding these challenges are; *how do we detect potential miscommunication*, *how do we dynamically adapt the communication to fit the analysis context*, and *how do we study best practices while taking into account subjective preferences*.

We can overcome some of these issues through *co-adaptive analytics*, i.e., by modeling the communication through learning and teaching dynamics. Nevertheless, it remains unclear *how effective visual versus verbal communication is for I&XAI* [Sev+18b] and *whether we can rely on rhetorical and argumentation strategies for compelling explanations* [El+19a]. To address these open questions, we need to study specific design patterns and their impact on I&XAI.

9.3.2 | Interaction Intent Identification to Bridge Knowledge Gaps

A notable open challenge for this research area is the accurate detection and modeling of user analysis intents. To design effective intelligence augmentation systems, we need to know the *rationale behind the users' interactions* and their *knowledge gaps* to be able to bridge them adequately [CGM19]. Hence it is still an open question *how to design interaction intent identification modules* to capture such information. Some prior works have relied on analyzing interaction logs or using eye-tracking techniques [Bus+12]. However, such approaches often provide error-prone or uncertain estimates of the users' behavior. Hence an open challenge for I&XAI is to figure out *how to design uncertainty-aware systems* to combat such issues.

9.3.3 | Trustworthy Decision-Making Support

According to the *General Data Protection Regulation (GDPR)*, when using machine learning models for high-stake decisions, users are entitled to be provided with “*meaningful explanations of the logic involved*” [GF17]. However, it remains unclear *how much we should show the users*. In particular, for I&XAI systems that aim to provide adequate decision-making support, it is imperative to study *how much control we should give which users and in which situation*. Especially if we rely on human input for learning, we need to investigate *how to avoid biases and reasoning pitfalls* [San+14] and *how to detect mental model changes or opinion shifts* [WB19].

We, therefore, rely on *co-adaptive guidance* to increase the users' confidence and trust in problem-solving and decision-making [Spe+21b]. Nonetheless, an open challenge to investigate is *whether more confident user decisions lead to better ML models*. We need to figure out *whether users reporting that they trust the models, actually means that they understand its inner workings* or maybe that they are rationalizing their understanding. This leads to the question of *whether we can trust the users to interact directly with the ML models*, i.e., should we only care about the users trusting the models or also about the *models trusting the users* [Spe+19]. Lastly, it remains unclear *how explainability affects trust* and *how we can measure these effects* [Spe+20a].

Conclusion

” *In literature and in life we ultimately pursue,
not conclusions, but beginnings.*

— **Sam Tanenhaus**
Literature Unbound [Tan84]

In conclusion, this dissertation presented a comprehensive *design space* for interactive and explainable machine learning approaches. This design space facilitated the introduction of our *conceptual framework*, spanning three *explainability spaces* and four *stakeholder groups*. Instantiating this framework, we presented four I&XAI techniques for *content and topic model refinement*. All techniques were designed to address the three tasks of *understanding, diagnosis, and refinement* of the ML models. Our *multi-staged evaluations* based on quantitative metrics and qualitative expert user studies verified the applicability of the presented techniques. Lastly, we derived twelve generalizable *design patterns* for I&XAI applications. In the following, we conclude the dissertation by pointing out the main take-home messages and by highlighting future research directions.

Contents

10.1 Take-Home Messages	220
10.1.1 Challenging Problems Require Shared Human-AI Agency	220
10.1.2 Stakeholders and Application Contexts Determine I&XAI Workflows	220
10.1.3 Tailored Visual Analytics System Design Empowers Users	221
10.2 Future Research Directions	221
10.2.1 Distilling Best Practices, Design Patterns, and Guidelines	221
10.2.2 Developing a Structured I&XAI Methodology	221
10.2.3 Supporting Seamless Human-AI Collaborations	221

10.1 | Take-Home Messages

After discussing the summary and main lessons learned in the previous chapters, in this section, we summarize the three primary take-home messages of our work.

10.1.1 | Challenging Problems Require Shared Human-AI Agency

Our I&XAI techniques have prominently illustrated that some problems, such as content and topic modeling, require both human and AI problem-solving capabilities. Such challenging and *ill-structured* [Sim73] problems usually have to satisfy multiple optimization objectives, and often these objectives do not sufficiently capture human intuition, experience, and expertise. To adequately address these challenges, we require a *shared human-AI agency*, for example, through *intelligence augmentation*.

10.1.2 | Stakeholders and Application Contexts Determine I&XAI Workflows

This dissertation has highlighted the potential for configuring custom I&XAI workflows to suit the targeted stakeholder group and application context. Designing workflows that go beyond the traditional linear one, opens up novel interaction and explanation designs that can appropriately configure and contextualize our techniques. In this dissertation, we have illustrated various workflows using deductive, contrastive, and inductive strategies. Our tailored workflows highlight the necessity to *adapt our designs to the intended user domains, data, and application contexts*.

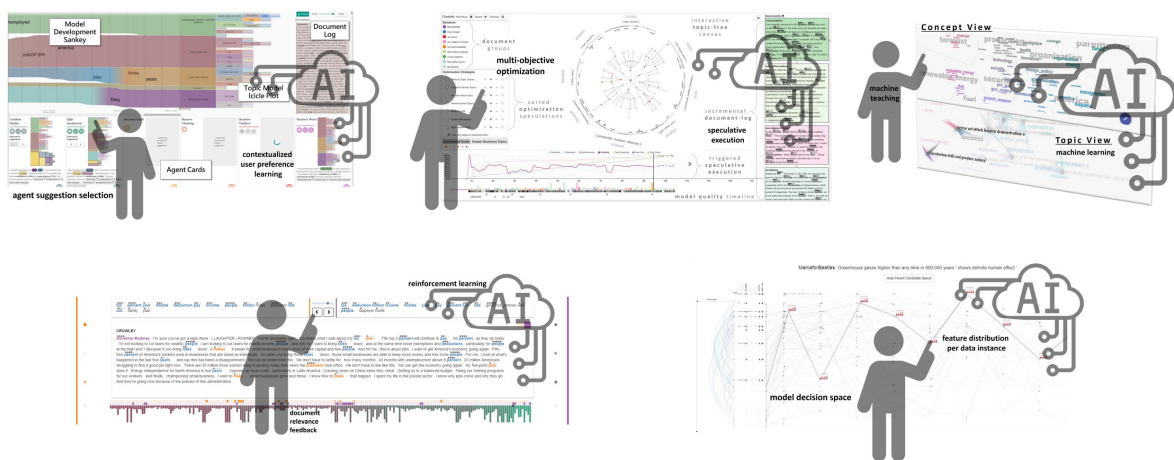


Fig. 10.1: I&XAI interfaces for content and topic model refinement. The interfaces are designed to tackle the same tasks but vary in their I&XAI workflows and strategies.

10.1.3 | Tailored Visual Analytics System Design Empowers Users

Through evaluating our proposed I&XAI techniques in user studies, we learned that *tailored designs could empower users to perform tasks that they would usually not be able to do* due to a lack of technical knowledge and literacy. Hence, as depicted in Figure 10.1, we need to study customizing our explanation strategies and interaction designs to different paradigms, empowering users to understand, diagnose, and refine machine learning models at their respective levels of experience, knowledge, and expertise.

10.2 | Future Research Directions

Lastly, based on the work conducted in this dissertation, in Section 9.3, we identified multiple research gaps. In this section, we highlight three opportunities for future research.

10.2.1 | Distilling Best Practices, Design Patterns, and Guidelines

The first research opportunity is to collect and survey I&XAI best practices and design patterns to formulate guidelines for the design of I&XAI applications. These guidelines will allow researchers and practitioners to discuss their designs, introduce novel contributions, and abstract away from one application domain to another, assessing the guidelines' generalizability.

10.2.2 | Developing a Structured I&XAI Methodology

Second, as discussed in Section 9.1.1, one of the limitations of our proposed I&XAI design space, framework, and design patterns, is the lack of a structured methodology. Hence, our next research opportunity is the development of a robust I&XAI methodology based on the groundwork provided in this dissertation. As part of our future work, we intend to further audit, refine, and verify the claims, concepts, and names for our I&XAI framework and design space. In addition, we will evaluate the methodology by surveying existing techniques to position them in the proposed design space and by conducting expert studies and workshops.

10.2.3 | Supporting Seamless Human-AI Collaborations

Our last future direction is to engage in interdisciplinary research, bringing computer scientists with scholars from psychology, philosophy, social science, the humanities, and others, to address the challenges of intelligence augmentation. This will enable us to work towards co-adaptive and multi-modal interaction and explanations that can be contextualized and personalized. Achieving such *seamless human-AI communication* will expedite effective collaboration because *together, human-AI teams can solve more challenging problems than each by themselves*.

Bibliography

- [AB18] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160 (cit. on pp. 21, 22, 46).
- [Abb+13] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, and Sepideh Mahabadi. “Real-Time Recommendation of Diverse Related Articles”. In: *Proceedings of the International Conference on World Wide Web*. 2013, pp. 1–12 (cit. on p. 28).
- [Abd+13] Alfie Abdul-Rahman, Julie Lein, Katherine Coles, et al. “Rule-Based Visual Mappings — With a Case Study on Poetry Visualization”. In: *Computer Graphics Forum* 32.3 (2013), pp. 381–390 (cit. on p. 29).
- [Abd+18] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. “Trends and Trajectories for Explainable, Accountable and Intelligible Systems : An HCI Research Agenda”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2018, 582:1–582:18 (cit. on p. 54).
- [ACH11] Erik Aumayr, Jeffrey Chan, and Conor Hayes. “Reconstruction of Threaded Conversations in Online Discussion Forums”. In: *Proceedings of the International AAAICAI Conference on Weblogs and Social Media* (2011), pp. 26–33 (cit. on pp. 147, 151, 159).
- [AG16] Eric Alexander and Michael Gleicher. “Task-Driven Comparison of Topic Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 320–329 (cit. on pp. 37, 112, 120).
- [AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. “Learning Topic Models – Going beyond SVD”. In: *IEEE Annual Symposium on Foundations of Computer Science*. IEEE, 2012, pp. 1–21 (cit. on p. 34).
- [Agr+09] Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, et al. “Exploiting Web Search Engines to Search Structured Databases”. In: *Proceedings of the International Conference on World Wide Web*. 2009, pp. 501–510 (cit. on p. 28).
- [Aiv+19] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, et al. “FairWashing: the Risk of Rationalization”. In: *Proceedings of the International Conference on Machine Learning (ICML)* 97 (2019), pp. 161–170 (cit. on p. 25).
- [AL19] Yongsu Ahn and Yu-Ru Lin. “FairSight: Visual Analytics for Fairness in Decision Making”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.01 (2019), pp. 1086–1095 (cit. on p. 24).

- [ALA17] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “Learning Bilingual Word Embeddings with (Almost) No Bilingual Data”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. 2017, pp. 451–462 (cit. on p. 29).
- [Ale+14] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. “Serendip: Topic Model-Driven Visual Exploration of Text Corpora”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2014, pp. 173–182 (cit. on pp. 36, 96, 134, 178).
- [Ale+17] Eric Alexander, Chih-Ching Chang, Mariana Shimabukuro, et al. “Perceptual Biases in Font Size as a Data Encoding”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.8 (2017), pp. 2397–2410 (cit. on p. 93).
- [Als+11] Jamal Alsakran, Yang Chen, Ye Zhao, Jing Yang, and Dongning Luo. “STREAMIT: Dynamic Visualization and Interactive Exploration of Text Streams”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2011, pp. 131–138 (cit. on p. 28).
- [Als+18] Bilal Alsallakh, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. “Do Convolutional Neural Networks Learn Class Hierarchy?” In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 152–162 (cit. on p. 24).
- [AM01] Javed A Aslam and Mark Montague. “Models for Metasearch”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2001, pp. 276–284 (cit. on p. 28).
- [Ame+14] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. “Power to the People: The Role of Humans in Interactive Machine Learning”. In: *AI Magazine* 35.4 (2014), pp. 105–120 (cit. on pp. 23, 46).
- [Ame+19] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, et al. “Guidelines for Human-AI Interaction”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019, 3:1–3:13 (cit. on pp. 25, 26, 46, 50, 55, 62).
- [Ami+04] Einat Amitay, Nadav Har’El, Ron Sivan, and Aya Soffer. “Web-A-Where: Geotagging Web Content”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2004, pp. 273–280 (cit. on p. 28).
- [And+02] Keith Andrews, Wolfgang Kienreich, Vedran Sabol, et al. “The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities”. In: *Proceedings of IEEE Conference on Information Visualization* 1.3–4 (2002), pp. 166–181 (cit. on p. 29).
- [And+11] David Andrzejewski, Wisconsin Madison, Mark Craven, and Benjamin Recht. “A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic”. In: *Proceedings International joint Conference on Artificial Intelligence*. 2011 (cit. on p. 33).
- [Aod+14] Oisín Mac Aodha, Vassilios Stathopoulos, Gabriel J. Brostow, et al. “Putting the Scientist in the Loop – Accelerating Scientific Progress with Interactive Machine Learning”. In: *International Conference on Pattern Recognition*. 2014, pp. 9–17 (cit. on p. 21).
- [AOP12] Aretha B. Alencar, Maria Cristina F. de Oliveira, and Fernando V. Paulovich. “Seeing Beyond Reading: A Survey on Visual Text Analytics”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012), pp. 476–492 (cit. on p. 28).

- [Arr+20] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115 (cit. on pp. 22, 46, 52–54, 72).
- [Ary+19] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, et al. “One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques”. In: *arXiv preprint arXiv:1909.03012* (2019) (cit. on pp. 21, 24).
- [ASW12] Daniel Angus, Andrew Smith, and Janet Wiles. “Conceptual Recurrence Plots: Revealing Patterns in Human Discourse”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.6 (2012), pp. 988–997 (cit. on pp. 28, 29).
- [Aue20] Jan Auernhammer. “Human-Centered AI: The Role of Human-Centered Design Research in the Development of AI”. In: *Synergy - DRS International Conference*. 2020 (cit. on p. 23).
- [AWS06] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. “Knowing the User’s Every Move: User Activity Tracking for Website Usability Evaluation and Implicit Interaction”. In: *Proceedings of the International Conference on World Wide Web*. 2006, pp. 203–212 (cit. on p. 28).
- [AZ12] Charu C Aggarwal and ChengXiang Zhai. “A Survey of Text Clustering Algorithms”. In: *Mining Text Data*. Springer, 2012, pp. 77–128 (cit. on pp. 33, 118).
- [AZC09] David Andrzejewski, Xiaojin Zhu, and Mark Craven. “Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors”. In: *Proceedings International Conference on Machine Learning*. 2009, pp. 25–32 (cit. on pp. 33, 100, 186).
- [Bac+15] Sebastian Bach, Alexander Binder, Grégoire Montavon, et al. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PloS one* 10.7 (2015), e0130140 (cit. on p. 60).
- [Bal+13] Ali Balali, Hesham Faili, Masoud Asadpour, and Mostafa Dehghani. “A Supervised Approach for Reconstructing Thread Structure in Comments on Blogs and Online News Agencies”. In: *Computacion y Sistemas* 17.2 (2013), pp. 207–217 (cit. on pp. 147, 152).
- [Ban+19] Gagan Bansal, Besmira Nushi, Ece Kamar, et al. “Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (2019), pp. 2–11 (cit. on p. 26).
- [Bas+07] Holger Bast, Alexandru Chitea, Fabian Suchanek, and Ingmar Weber. “ESTER: Efficient Search on Text, Entities, And Relations”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2007, pp. 671–678 (cit. on p. 28).
- [BB05] Mary Bazire and Patrick Brézillon. “Understanding Context Before using It”. In: *International and Interdisciplinary Conference on Modeling and using Context*. Vol. 3554. Springer, 2005, pp. 29–40 (cit. on p. 24).
- [BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. *Clustering via Similarity Functions: Theoretical Foundations and Algorithms*. 2008 (cit. on p. 181).
- [BCB10] E. A. Bier, S. K. Card, and J. W. Bodnar. “Principles and Tools for Collaborative Entity-Based Intelligence Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.2 (2010), pp. 178–191 (cit. on p. 28).

- [BE01] Victoria Bellotti and Keith Edwards. “Intelligibility and Accountability: Human Considerations in Context-Aware Systems”. In: *Human–Computer Interaction* 16.2–4 (2001), pp. 193–212 (cit. on p. 54).
- [Bec+20] Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. “Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. ACL, 2020, pp. 60–73 (cit. on pp. 50, 53, 54, 73).
- [Beh+18] Michael Behrisch, Michael Blumenschein, Nam Wook Kim, et al. “Quality Metrics for Information Visualization”. In: *Computer Graphics Forum*. Vol. 37. 3. 2018, pp. 625–662 (cit. on p. 16).
- [Ben+03] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. “A Neural Probabilistic Language Model”. In: *Journal of Machine Learning Research* 3.Feb (2003), pp. 1137–1155 (cit. on p. 29).
- [Ben20] Roy T. Bennett. *The Light in the Heart: Inspirational Thoughts for Living your Best Life*. 2020 (cit. on p. vii).
- [BEX02] Florian Beil, Martin Ester, and Xiaowei Xu. “Frequent Term-Based Text Clustering”. In: *Proceedings ACM Conference on Knowledge Discovery and Data Mining*. ACM. 2002, pp. 436–442 (cit. on p. 33).
- [BF11] David M Blei and Peter I Frazier. “Distance Dependent Chinese Restaurant Processes”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2461–2488 (cit. on p. 29).
- [BGJ10] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. “The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies”. In: *Journal of the ACM (JACM)* 57.2 (2010), pp. 1–30 (cit. on p. 33).
- [BHM17] Jordan Boyd-Graber, Yuening Hu, and David Mimno. “Applications of Topic Models”. In: *Foundations and Trends in Information Retrieval* 11.2–3 (2017), pp. 143–296 (cit. on pp. 7, 36).
- [Bia+05] Bruno Biais, Denis J. Hilton, Karine Mazurier, and Sébastien Pouget. “Judgmental Overconfidence, Self-Monitoring and Trading Performance in an Experimental Financial Market”. In: *Experimental Studies eJournal* (2005) (cit. on p. 25).
- [Bic+17] Paulo Bicalho, Marcelo Pita, Gabriel Pedrosa, Anisio Lacerda, and Gisele L Pappa. “A General Framework to Expand Short Text for Topic Modeling”. In: *Information Sciences* 393 (2017), pp. 66–81 (cit. on p. 34).
- [BK09] Tony Bergstrom and Karrie Karahalios. “Conversation Clusters: Grouping Conversation Topics Through Human-Computer Dialog”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009, pp. 2349–2352 (cit. on p. 34).
- [BKW16] Fabian Beck, Sebastian Koch, and Daniel Weiskopf. “Visual Analysis and Dissemination of Scientific Literature Collections with SurVis”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 180–189 (cit. on p. 29).
- [BL06a] David M. Blei and John D. Lafferty. “Correlated Topic Models”. In: *Proceedings of the International Conference on Neural Information Processing Systems* (2006), pp. 147–154 (cit. on p. 34).

- [BL06b] David M. Blei and John D. Lafferty. “Dynamic Topic Models”. In: *Proceedings of the International Conference on Machine learning*. ACM. 2006, pp. 113–120 (cit. on pp. 29, 33, 34).
- [Ble12a] David M Blei. “Topic Modeling and Digital Humanities”. In: *Journal of Digital Humanities* 2.1 (2012) (cit. on pp. 36, 73).
- [Ble12b] David M. Blei. “Probabilistic Topic Models”. In: *Communications of the ACM* 55.4 (2012), pp. 77–84 (cit. on pp. 6, 7, 33, 34, 100, 183).
- [BLL12] Andrea Bunt, Matthew Lount, and Catherine Lauzon. “Are Explanations Always Important?: A Study of Deployed, Low-Cost Intelligent Interactive Systems”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 2012, pp. 169–178 (cit. on p. 25).
- [BMS17] M Berger, K Mcdonough, and L. M Seversky. “cite2vec: Citation-Driven Document Exploration via Word Embeddings”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 691–700 (cit. on pp. 28, 29).
- [BNJ03] David M. Blei, Andrew Y Ng, and Michael I Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3.3 (2003), pp. 993–1022 (cit. on pp. 7, 33, 34, 36, 87, 183, 184, 187).
- [Bos+13] Harald Bosch, Dennis Thom, Florian Heimerl, et al. “ScatterBlogs2: Real-Time Monitoring of Microblog Messages Through User-Guided Filtering”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2022–2031 (cit. on p. 28).
- [Bos17] Nick Bostrom. *Superintelligence*. Dunod, 2017 (cit. on p. 3).
- [Bou09] Gerlof Bouma. “Normalized (Pointwise) Mutual Information in Collocation Extraction”. In: *Proceedings International Conference of German Society for Computational Linguistics and Language Technology* (2009), pp. 31–40 (cit. on p. 38).
- [Box79] George EP Box. “Robustness in the Strategy of Scientific Model Building”. In: *Robustness in Statistics*. Academic Press, 1979, pp. 201–236 (cit. on p. 75).
- [Bra+18] Adam James Bradley, Mennatallah El-Assady, Katharine Coles, et al. “Visualization and the Digital Humanities”. In: *IEEE Computer Graphics and Applications* 38.6 (2018), pp. 26–38 (cit. on p. 16).
- [Bre01] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 151).
- [Bré07] Patrick Brézillon. “Context Modeling: Task Model and Practice Model”. In: *International and Interdisciplinary Conference on Modeling and using Context*. Springer. 2007, pp. 122–135 (cit. on p. 24).
- [Bré99] Patrick Brézillon. “Context in Artificial Intelligence: I. A Survey of the Literature”. In: *Computers and artificial intelligence* 18 (1999), pp. 321–340 (cit. on p. 24).
- [Bro+07] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. “A Semantic Approach to Contextual Advertising”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2007, pp. 559–566 (cit. on p. 28).
- [Bro+17] Austin J Brockmeier, Tingting Mu, Sophia Ananiadou, and John Y Goulermas. “Quantifying the Informativeness of Similarity Measurements”. In: *Journal of Machine Learning Research* 18.76 (2017), pp. 1–61 (cit. on p. 29).

- [Bus+12] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V Elst. “Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1.2 (2012), p. 9 (cit. on p. 217).
- [BYC13] James Bergstra, Daniel L K Yamins, and D Cox. “Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures”. In: *Journal of Machine Learning Research* (2013), pp. 115–123 (cit. on p. 62).
- [Cab+19] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, et al. “FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2019), pp. 46–56 (cit. on p. 24).
- [Cai+19] Carrie J. Cai, Emily Reif, Narayan Hegde, et al. “Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2019, 4:1–4:14 (cit. on pp. 20, 21, 24).
- [Cam+17] Raphael Campos, Sérgio Canuto, Thiago Salles, Clebson CA de Sá, and Marcos André Gonçalves. “Stacking Bagged and Boosted Forests for Effective Automated Classification”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2017, pp. 105–114 (cit. on p. 29).
- [Cañ+05] Alberto J Cañas, Roger Carff, Greg Hill, et al. “Concept Maps: Integrating Knowledge and Information Visualization”. In: (2005), pp. 205–219 (cit. on p. 35).
- [Cao+11] Nan Cao, D. Gotz, Jimeng Sun, Yu-Ru Lin, and Huamin Qu. “SolarMap: Multifaceted Visual Analytics for Topic Exploration”. In: *Proceedings of IEEE International Conference on Data Mining*. 2011, pp. 101–110 (cit. on p. 36).
- [Cao+15] Nan Cao, Lu Lu, Yu-Ru Lin, Fei Wang, and Zhen Wen. “Socialhelix: Visual Analysis of Sentiment Divergence in Social Media”. In: *Journal of visualization* 18.2 (2015), pp. 221–235 (cit. on p. 29).
- [CC13] Meng-Wei Chang and Christopher Collins. “Exploring Entities in Text with Descriptive Non-Photorealistic Rendering”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2013, pp. 9–16 (cit. on p. 29).
- [CCL16] Francine Chen, Patrick Chiu, and Seongtaek Lim. “Topic Modeling of Document Metadata for Visualizing Collaborations over Time”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2016, pp. 108–117 (cit. on p. 29).
- [CCO20] Roberta Calegari, Giovanni Ciatto, and Andrea Omicini. “On the Integration of Symbolic and Sub-Symbolic Techniques for XAI: A Survey”. In: *Intelligenza Artificiale* 14.1 (2020), pp. 7–32 (cit. on p. 52).
- [CCP09] Christopher Collins, Sheelagh Carpendale, and Gerald Penn. “DocuBurst: Visualizing Document Content using Language Structure”. In: *Computer Graphics Forum* 28.3 (2009), pp. 1039–1046 (cit. on p. 29).
- [CD18] Marco Cavallo and Çagatay Demiralp. “A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2018), pp. 1–4 (cit. on p. 82).
- [Cen+18] Davide Ceneda, Theresia Gschwandtner, Thorsten May, et al. “Guidance or No Guidance? A Decision Tree can Help”. In: *Proceedings of Eurographics Conference on Visualization, Workshop on Visual Analytics (EuroVA)*. Eurographics Association, 2018, pp. 19–23 (cit. on p. 46).

- [CG99] Fay Chang and Garth A. Gibson. “Automatic I/O Hint Generation through Speculative Execution”. In: *Proceedings Symp. on Operating Systems Design and Implementation*. 1999, pp. 1–14 (cit. on p. 178).
- [CGM19] Davide Ceneda, Theresia Gschwandtner, and Silvia Miksch. “A Review of Guidance Approaches in Visual Data Analysis: A Multifocal Perspective”. In: *Computer Graphics Forum* 38.3 (2019), pp. 861–879 (cit. on p. 217).
- [CH19] Miruna-Adriana Clinciu and Helen Hastie. “A Survey of Explainable AI Terminology”. In: *Proceedings of the Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI)*. 2019, pp. 8–13 (cit. on p. 54).
- [Cha+02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. “SMOTE: Synthetic Minority Over-Sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357 (cit. on p. 148).
- [Cha+09] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. “Reading Tea Leaves: How Humans Interpret Topic Models”. In: *Proceedings of the International Conference on Neural Information Processing Systems* (2009), pp. 288–296 (cit. on pp. 7, 37, 38, 73).
- [Cha+12] Junghoon Chae, Dennis Thom, Harald Bosch, et al. “Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2012, pp. 143–152 (cit. on p. 28).
- [Che+06] David Cheng, Ravi Kannan, Santosh Vempala, and Grant Wang. “A Divide-and-Merge Methodology for Clustering”. In: *ACM Trans. on Database Systems (TODS)* 31.4 (2006), pp. 1499–1525 (cit. on p. 33).
- [Che+08] Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. “Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning”. In: (2008) (cit. on p. 35).
- [Che+13] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, et al. “Leveraging Multi-Domain Prior Knowledge in Topic Models.” In: vol. 13. 2013, pp. 2071–77 (cit. on p. 33).
- [Che+14] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. “BTM: Topic Modeling Over Short Texts”. In: *IEEE Transactions on Knowledge and Data Engineering* 26.12 (2014) (cit. on p. 35).
- [Che+17] Siming Chen, Shuai Chen, Zhenhuang Wang, et al. “D-Map: Visual Analysis of Ego-Centric Information Diffusion Patterns in Social Media”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2017, pp. 41–50 (cit. on p. 29).
- [Che+18] S. Chen, J. Li, G. Andrienko, et al. “Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling”. In: *IEEE Transactions on Visualization and Computer Graphics* (2018), pp. 2499–2516 (cit. on pp. 26, 62).
- [Che15] Y. Chen. “Visual Opinion Analysis of Threaded Discussions”. In: *IEEE International Conference on Data Mining Workshop*. 2015, pp. 646–651 (cit. on p. 30).
- [CHF10] Fanny Chevalier, Stephane Huot, and Jean-Daniel Fekete. “WikipediaViz: Conveying Article Quality for Casual Wikipedia readers”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2010, pp. 49–56 (cit. on p. 29).

- [Chi+15] Ming-Te Chi, Shih-Syun Lin, Shiang-Yi Chen, Chao-Hung Lin, and Tong-Yee Lee. “Morphable Word Clouds for Time-Varying Text Data Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 21.12 (2015), pp. 1415–1426 (cit. on pp. 28, 29).
- [Cho+13] Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. “UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 1992–2001 (cit. on pp. 36, 37, 80).
- [Cho+16] Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda, and William Ribarsky. “VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 210–219 (cit. on p. 29).
- [Cho00] Freddy YY Choi. “Advances in Domain Independent Linear Text Segmentation”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2000, pp. 26–33 (cit. on p. 34).
- [Chr11] Brian Christian. *The Most Human Human: What Talking with Computers Teaches us about what it Means to be Alive*. 2011 (cit. on p. 77).
- [Chu+06] Jennifer Chu-Carroll, John Prager, Krzysztof Czuba, David Ferrucci, and Pablo Duboue. “Semantic Search via XML Fragments: A High-Precision Approach to IR”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2006, pp. 445–452 (cit. on p. 28).
- [Chu+12] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. “Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2012, pp. 443–452 (cit. on pp. 6, 7, 29, 37, 176).
- [Chu+13] Jason Chuang, Sonal Gupta, Christopher D Manning, and Jeffrey Heer. “Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2013, pp. 612–620 (cit. on p. 37).
- [CJH19] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. “The Effects of Example-Based Explanations in a Machine Learning Interface”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 2019, pp. 258–262 (cit. on p. 21).
- [CMH12] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. “Termite: Visualization Techniques for Assessing Textual Topic Models”. In: *Proceedings of International Conference on Advanced Visual Interfaces*. ACM. 2012, pp. 74–77 (cit. on pp. 29, 36).
- [CMM83] Jaime G Carbonell, Ryszard S Michalski, and Tom M Mitchell. “An Overview of Machine Learning”. In: *Machine learning* (1983), pp. 3–23 (cit. on p. 52).
- [CMQ20] Furui Cheng, Yao Ming, and Huamin Qu. “DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2020), pp. 1438–1447 (cit. on p. 70).
- [CNP06] Giuseppe Carenini, Raymond T Ng, and Adam Pauls. “Interactive Multimedia Summaries of Evaluative Text”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2006, pp. 124–131 (cit. on p. 28).
- [Col+18] Christopher Collins, Natalia Andrienko, Tobias Schreck, et al. “Guidance In the Human–Machine Analytics Process”. In: *Visual Informatics* 2.3 (2018), pp. 166–180 (cit. on p. 82).

- [CPC09] Christopher Collins, Gerald Penn, and Sheelagh Carpendale. “Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1009–1016 (cit. on p. 29).
- [Cra+12] Steven P Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. “Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond”. In: *Mining Text Data*. Springer, 2012, pp. 129–161 (cit. on p. 34).
- [Cra96] Mark W. Craven. “Extracting Comprehensible Models from Trained Neural Networks”. PhD thesis. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 1996 (cit. on p. 53).
- [Cro+11] Patricia J Crossno, Andrew T Wilson, Timothy M Shead, and Daniel M Dunlavy. “Topicview: Visually Comparing Topic Models of Text Collections”. In: *Tools with Artificial Intelligence (ICTAI), IEEE International Conference on*. IEEE. 2011, pp. 936–943 (cit. on p. 37).
- [CS19] Michael Chromik and Martin Schuessler. “A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI”. In: (2019) (cit. on p. 212).
- [CSS06] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. “Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. MIT Press, 2006, pp. 241–248 (cit. on pp. 34, 149).
- [CSS08] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. “Combining Concept Hierarchies and Statistical Topic Models”. In: (2008), pp. 1469–1470 (cit. on p. 34).
- [Cua+07] Ana M Cuadros, Fernando V Paulovich, Rosane Minghim, and Guilherme P Telles. “Point Placement by Phylogenetic Trees and Its Application to Visual Analysis of Document Collections”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2007, pp. 99–106 (cit. on p. 29).
- [Cui+11] Weiwei Cui, Shixia Liu, Li Tan, et al. “TextFlow: Towards Better Understanding of Evolving Topics in Text”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2412–2421 (cit. on p. 36).
- [Cui+14] Weiwei Cui, Shixia Liu, Zhuofeng Wu, and Hao Wei. “How Hierarchical Topics Evolve in Large Text Corpora”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2281–2290 (cit. on pp. 34, 36).
- [Cus19] Fiery Andrews Cushman. “Rationalization Is Rational”. In: *Behavioral and Brain Sciences* 43 (2019), e28 (cit. on p. 25).
- [CVW09] Christopher Collins, Fernanda B Viégas, and Martin Wattenberg. “Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE. 2009, pp. 91–98 (cit. on pp. 28, 29, 118).
- [CW20] Dennis Collaris and Jarke J van Wijk. “Explainexplore: Visual Exploration of Machine Learning Explanations”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. IEEE. 2020, pp. 26–35 (cit. on pp. 25, 48).
- [Dal+15] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. “Calibrating Probability with Undersampling for Unbalanced Classification”. In: *IEEE Symposium Series on Computational Intelligence (IEEE SSCI)*. IEEE. 2015, pp. 159–166 (cit. on p. 159).

- [Dau+17] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. “Language Modeling with Gated Convolutional Networks”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2017, pp. 933–941 (cit. on p. 29).
- [DC00] Susan Dumais and Hao Chen. “Hierarchical Classification of Web Content”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2000, pp. 256–263 (cit. on p. 28).
- [DCP08] Mark Dredze, Koby Crammer, and Fernando Pereira. “Confidence-Weighted Linear Classification”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2008, pp. 264–271 (cit. on p. 29).
- [DD06] Darina Dicheva and Christo Dichev. “TM4L: Creating and Browsing Educational Topic Maps”. In: *British Journal of Educational Technology* 37 (2006), pp. 391–404 (cit. on p. 35).
- [Del+08] Matthias Deller, Stefan Agne, Achim Ebert, et al. “Managing a Document-Based Information Space”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2008, pp. 119–128 (cit. on p. 28).
- [Der+14] Michael Derntl, Nikou Günnemann, Alexander Tillmann, Ralf Klamma, and Matthias Jarke. “Building and Exploring Dynamic Topic Models on the Web”. In: *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 2014, pp. 2012–2014 (cit. on p. 35).
- [Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. ACL, 2019, pp. 4171–4186 (cit. on p. 35).
- [Dha+21] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, et al. “Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle”. In: *Proceedings Conference on Designing Interactive Systems*. ACM, 2021, pp. 1591–1602 (cit. on p. 22).
- [DK17] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv preprint arXiv:1702.08608* abs/1702.08608 (2017) (cit. on p. 23).
- [DK18] John J. Dudley and Per Ola Kristensson. “A Review of User Interface Design for Interactive Machine Learning”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.2 (2018), 8:1–8:37 (cit. on pp. 21–23).
- [DK95] Thomas G Dietterich and Eun Bae Kong. *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*. Tech. rep. Technical report, Department of Computer Science, Oregon State University, 1995 (cit. on p. 62).
- [DL00] Manoranjan Dash and Huan Liu. “Feature Selection for Clustering”. In: *Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2000, pp. 110–121 (cit. on p. 117).
- [DLP08] Chris Ding, Tao Li, and Wei Peng. “On the Equivalence Between Non-Negative Matrix Factorization and Probabilistic Latent Semantic Indexing”. In: *Computational Statistics & Data Analysis* 52.8 (2008), pp. 3913–3927 (cit. on p. 33).
- [Doe+10] Marian Doerk, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. “A Visual Backchannel for Large-Scale Events”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1129–1138 (cit. on p. 36).

- [Dom+19] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, et al. “Explanations can be Manipulated and Geometry is to Blame”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 22).
- [Dos+17] Finale Doshi-Velez, Mason Kortz, Ryan Budish, et al. “Accountability of AI Under the Law: The Role of Explanation”. In: *Berkman Center Research Publication* 18.07 (2017) (cit. on p. 54).
- [Dou+11] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. “ParallelTopics: A Probabilistic Approach to Exploring Document Collections”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2011, pp. 231–240 (cit. on p. 36).
- [Dou+13] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. “HierarchicalTopics: Visually Exploring Large Text Collections using Topic Hierarchies”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2002–2011 (cit. on pp. 29, 36).
- [Dou+15] Wenwen Dou, Isaac Cho, Omar Eltayeb, et al. “DemographicVis: Analyzing Demographic Information Based on User Generated Content”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2015, pp. 57–64 (cit. on pp. 28, 29).
- [Dow+18] Michelle Dowling, John Wenskovich, Peter Hauck, et al. “A Bidirectional Pipeline for Semantic Interaction”. In: *Proceedings of the IEEE Visualization Conference, Workshop on Machine Learning from User Interaction for Visualization and Analytics*. Vol. 11. 2018 (cit. on p. 80).
- [Du+17] Nan Du, Yingyu Liang, Maria-Florina Balcan, et al. “Scalable Influence Maximization for Multiple Products in Continuous-Time Diffusion Networks”. In: *Journal of Machine Learning Research* 18.2 (2017), pp. 1–45 (cit. on p. 29).
- [Dun93] Ted Dunning. “Accurate Methods for The Statistics of Surprise and Coincidence”. In: *Computational Linguistics* 19.1 (1993), pp. 61–74 (cit. on p. 187).
- [DWM04] Kushal Dave, Martin Wattenberg, and Michael Muller. “IBM Research Report: Flash Forums and ForumReader : Navigating a New Kind of Large-Scale Online Discussion”. In: *IBM Watson Research Center* 23305 (2004), pp. 1–11 (cit. on p. 30).
- [DXT17] Yali Du, Chang Xu, and Dacheng Tao. “Privileged Matrix Factorization for Collaborative Filtering”. In: *Proceedings International joint Conference on Artificial Intelligence*. 2017, pp. 1610–1616 (cit. on p. 28).
- [DYY16] Zhi-Hong Deng, Hongliang Yu, and Yunlun Yang. “Identifying Sentiment Words using an Optimization Model with L1 Regularization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2016, pp. 115–121 (cit. on p. 29).
- [EBH21] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. “How to Support Users in Understanding Intelligent Systems? Structuring the Discussion”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2021 (cit. on p. 23).
- [EBN13] Alex Endert, Lauren Bradel, and Chris North. “Beyond Control Panels: Direct Manipulation for Visual Analytics”. In: *IEEE Computer Graphics and Applications* 33.4 (2013), pp. 6–13 (cit. on p. 82).
- [EC08] Gonenc Ercan and Ilyas Cicekli. “Lexical Cohesion Based Topic Modeling for Summarization”. In: *International Conference on Intelligent Text Processing and Computational Linguistics* (2008), pp. 582–592 (cit. on p. 34).

- [Edw54] Ward Edwards. “The Theory of Decision Making”. In: *Psychological bulletin* 51.4 (1954), p. 380 (cit. on p. 26).
- [EFN12] Alex Endert, Patrick Fiaux, and Chris North. “Semantic Interaction for Visual Text Analytics”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 473–482 (cit. on pp. 29, 82).
- [EH19] Mennatallah El-Assady and Annette Hautli-Janisz. *Discourse Maps—Feature Encoding for the Analysis of Verbatim Conversation Transcripts*. 2019 (cit. on p. 31).
- [Ehr17] Patrik Ehrencrona Kjellin. “Interactive Topic Modeling for Source Code Analysis”. MA thesis. Uppsala University, 2017 (cit. on p. 35).
- [Ehs+18] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. “Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations”. In: *Proceedings of the Conference on AI, Ethics, and Society*. ACM, 2018, pp. 81–87 (cit. on p. 25).
- [Ehs+19] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. “Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 2019, pp. 263–274 (cit. on p. 53).
- [Ehs+21] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. “Expanding Explainability: Towards Social Transparency in AI Systems”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2021 (cit. on pp. 23, 25, 49, 68).
- [Eis+12] Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. “TopicViz: Interactive Topic Exploration in Document Collections”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Extended Abstracts)*. 2012, pp. 2177–2182 (cit. on p. 36).
- [El+16] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. “ConToVi: Multi-Party Conversation Exploration using Topic-Space Views”. In: *Computer Graphics Forum* 35.3 (2016), pp. 431–440 (cit. on pp. 16, 30, 31).
- [El+17a] Mennatallah El-Assady, Annette Hautli-Janisz, Valentin Gold, et al. “Interactive Visual Analysis of Transcribed Multi-Party Discourse”. In: *Proceedings of the Association for Computational Linguistics, System Demonstrations*. ACL, 2017, pp. 49–54 (cit. on pp. 16, 31, 142).
- [El+17b] Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. “NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations”. In: *Computer Graphics Forum*. Vol. 36. 3. 2017, pp. 213–225 (cit. on pp. 16, 29, 31, 153).
- [El+18a] Mennatallah El-Assady, Rita Sevastjanova, Daniel Keim, and Christopher Collins. “Thread-Reconstructor: Modeling Reply-Chains to Untangle Conversational Text through Visual Analytics”. In: *Computer Graphics Forum* 37.3 (2018), pp. 351–365 (cit. on pp. 15, 141).
- [El+18b] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. “Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 382–391 (cit. on pp. 15, 87, 111).

- [El-+18c] Mennatallah El-Assady, Fabian Sperrle, Rita Sevastjanova, Michael Sedlmair, and Daniel Keim. “LTMA: Layered Topic Matching for the Comparative Exploration, Evaluation, and Refinement of Topic Modeling Results”. In: *International Symposium on Big Data Visual and Immersive Analytics*. 2018, pp. 1–10 (cit. on pp. 112, 114, 120).
- [El-+19a] Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, et al. “Towards Explainable Artificial Intelligence: Structuring the Processes of Explanations”. In: *ACM CHI Workshop on Human-Centered Machine Learning Perspectives*. 2019, pp. 1–9 (cit. on pp. 24, 45, 55, 216).
- [El-+19b] Mennatallah El-Assady, Wolfgang Jentner, Fabian Sperrle, et al. “lingvis.io - A Linguistic Visual Analytics Framework”. In: *Proceedings of the Association for Computational Linguistics, System Demonstrations*. ACL, 2019, pp. 13–18 (cit. on pp. 16, 27, 31, 38, 73, 87).
- [El-+19c] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. “Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 374–384 (cit. on pp. 15, 87, 88, 94, 100, 149, 175).
- [El-+20] Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. “Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 1001–1011 (cit. on pp. 15, 79).
- [EMH19] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. “Neural Architecture Search: A Survey”. In: *Journal of Machine Learning Research* 20.55 (2019), pp. 1–21 (cit. on p. 62).
- [End+15] Alex Endert, Remco Chang, Chris North, and Michelle Zhou. “Semantic Interaction: Coupling Cognition and Computation through Usable Interactive Analytics”. In: *IEEE Computer Graphics and Applications* 35.4 (2015), pp. 94–99 (cit. on pp. 80, 82, 212).
- [End+17] Alex Endert, W. Ribarsky, Cagatay Turkay, et al. “The State of the Art in Integrating Machine Learning into Visual Analytics”. In: *Computer Graphics Forum* 36.8 (2017), pp. 458–486 (cit. on pp. 22, 25, 46).
- [End16] Alex Endert. “Semantic Interaction for Visual Analytics: Inferring Analytical Reasoning for Model Steering”. In: *Synthesis Lectures on Visualization* 4.2 (2016), pp. 1–99 (cit. on p. 68).
- [EOS13] Sandra Garcia Esparza, Michael P. O’Mahony, and Barry Smyth. “CatStream: Categorising Tweets for User Profiling and Stream Filtering”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2013, pp. 25–36 (cit. on p. 28).
- [ER20] Upol Ehsan and Mark O. Riedl. “Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach”. In: *Lecture Notes in Computer Science*. Springer, 2020, pp. 449–466 (cit. on pp. 23, 25, 46).
- [Esl+15] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, et al. “I Always Assumed that I Wasn’t Really that Close to [Her]: Reasoning About Invisible Algorithms in News Feeds”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2015, pp. 153–162 (cit. on p. 25).
- [Fan+16] Hanyin Fang, Fei Wu, Zhou Zhao, et al. “Community-Based Question Answering via Heterogeneous Social Network Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2016, pp. 122–128 (cit. on p. 28).

- [Far14] Ilijas Farah. “Absoluteness, Truth, and Quotients”. In: vol. 25. 2014, pp. 1–24 (cit. on p. 201).
- [FB74] Raphael Finkel and Jon Bentley. “Quad Trees a Data Structure for Retrieval on Composite Keys”. In: *Acta Informatica* 4.1 (1974), pp. 1–9 (cit. on p. 90).
- [FBM16] Johanna Fulda, Matthew Brehmer, and Tamara Munzner. “TimeLineCurator: Interactive authoring of visual timelines from unstructured text”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 300–309 (cit. on p. 29).
- [Fed+17a] Paolo Federico, Florian Heimerl, Steffen Koch, and Silvia Miksch. “A Survey on Visual Approaches for Analyzing Scientific Literature and Patents”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (2017), pp. 2179–2198 (cit. on p. 28).
- [Fed+17b] Paolo Federico, Markus Wagner, Alexander Rind, et al. “The Role of Explicit Knowledge: A Conceptual Model of Knowledge-Assisted Visual Analytics”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, pp. 92–103 (cit. on p. 80).
- [Fer+19] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. “Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?” In: *IEEE Computational Intelligence Magazine* 14.1 (2019), pp. 69–81 (cit. on p. 53).
- [FFB18] Cristian Felix, Steven Franconeri, and Enrico Bertini. “Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 657–666 (cit. on p. 92).
- [FG01] Norbert Fuhr and Kai Großjohann. “XIRQL: A Query Language for Information Retrieval in XML Documents”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2001, pp. 172–180 (cit. on p. 28).
- [FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. “Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. 2005, pp. 363–370 (cit. on p. 150).
- [Fis+08] Danyel Fisher, Aaron Hoff, George Robertson, and Matthew Hurst. “Narratives: A Visualization to Track Narrative Events as they Develop”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2008, pp. 115–122 (cit. on pp. 28, 29).
- [Fis87] Douglas H. Fisher. “Knowledge Acquisition Via Incremental Conceptual Clustering”. In: *Machine Learning* 2.2 (1987), pp. 139–172 (cit. on p. 33).
- [FJ03] Jerry Alan Fails and Dan R. Olsen Jr. “Interactive Machine Learning”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. ACM. 2003, pp. 39–45 (cit. on pp. 21, 46, 62).
- [FK12] Gerlinde Fellner and Sebastian Krügel. “Judgmental Overconfidence: Three Measures, One Bias?” In: *Journal of Economic Psychology* 33 (2012), pp. 142–154 (cit. on p. 25).
- [For87] Steven Fortune. “A Sweepline Algorithm for Voronoi Diagrams”. In: *Algorithmica* 2 (1987), pp. 153–174 (cit. on p. 92).
- [FPS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. In: *Proceedings of International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996, pp. 82–88 (cit. on pp. 46, 52, 164).

- [Fu+17] Siwei Fu, Jian Zhao, Weiwei Cui, and Huamin Qu. “Visual Analysis of MOOC Forums with iForum”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 201–210 (cit. on p. 30).
- [Fuj+08] Ko Fujimura, Shigeru Fujimura, Tatsushi Matsubayashi, Takeshi Yamada, and Hidenori Okuda. “Topigraphy: Visualization for Large-Scale Tag Clouds”. In: *Proceedings of the International Conference on World Wide Web*. 2008, pp. 1087–1088 (cit. on p. 29).
- [FY99] Radu Florian and David Yarowsky. “Dynamic Nonlocal Language Modeling via Hierarchical Topic-Based Adaptation”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. ACL. 1999, pp. 167–174 (cit. on p. 34).
- [FZB11] Xiaoliang Fan, Ruisheng Zhang, and Patrick Brézillon. “Contextualizing Scientific Workflows in Cooperative Design”. In: *Proceedings of the International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE. 2011, pp. 17–22 (cit. on p. 25).
- [GA19] David Gunning and David Aha. “DARPA’s Explainable Artificial Intelligence (XAI) Program”. In: *AI Magazine* 40.2 (2019), pp. 44–58 (cit. on pp. 8, 20).
- [Gal+03] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. “Discourse Segmentation of Multi-party Conversation”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. 2003, pp. 562–569 (cit. on p. 181).
- [Gan+14] Qihong Gan, Min Zhu, Mingzhao Li, et al. “Document Visualization: An Overview of Current Research”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 6.1 (2014), pp. 19–36 (cit. on p. 28).
- [Gar+18] Rafael Garcia, Alexandru Cristian Telea, B. B. D. Silva, Jim Tørresen, and João Luiz Dihl Comba. “A Task-and-Technique Centered Survey on Visual Analytics for Deep Learning Model Engineering”. In: *Computers & Graphics* (2018), pp. 30–49 (cit. on p. 50).
- [GAZ19] Amirata Ghorbani, Abubakar Abid, and James Zou. “Interpretation of Neural Networks is Fragile”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (2019), pp. 3681–3688 (cit. on p. 22).
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 52).
- [GBD92] Stuart Geman, Elie Bienenstock, and René Doursat. *Neural Networks and the Bias Variance Dilemma*. 1992 (cit. on p. 62).
- [GC17] Derek Greene and James P Cross. “Exploring the Political Agenda of the European Parliament using a Dynamic Topic Modeling Approach”. In: *Political Analysis* 25.01 (2017), pp. 77–94 (cit. on p. 35).
- [GF17] Bryce Goodman and Seth Flaxman. “European Union Regulations on Algorithmic Decision-Making and a ”Right to Explanation””. In: *AI magazine* 38.3 (2017), pp. 50–57 (cit. on pp. 20, 217).
- [GKH15] Dominic Girardi, Josef Kueng, and Andreas Holzinger. “A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop”. In: *Brain Informatics and Health*. Springer, 2015, pp. 389–398 (cit. on p. 21).
- [GL13] Andrew Gelman and Eric Loken. “The Garden of Forking Paths: Why Multiple Comparisons can be a Problem, Even when there is No ”Fishing Expedition” or ”P-Hacking” and the Research Hypothesis was Posited ahead of Time”. In: *Department of Statistics, Columbia University* (2013) (cit. on p. 62).

- [Gla06] Malcolm Gladwell. “Blink: The Power of Thinking without Thinking”. In: (2006) (cit. on p. 43).
- [Gle16] Michael Gleicher. “A Framework for Considering Comprehensibility in Modeling”. In: *Big data* 4.2 (2016), pp. 75–88 (cit. on p. 53).
- [Gle18] Michael Gleicher. “Considerations for Visualizing Comparison”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 413–423 (cit. on pp. 82, 193).
- [GLF89] John H. Gennari, Pat Langley, and Doug Fisher. “Models of Incremental Concept Formation”. In: *Artificial intelligence* 40.1 (1989), pp. 11–61 (cit. on p. 33).
- [Glu+18] Michael Glueck, Mahdi Pakdaman Naeini, Finale Doshi-Velez, et al. “PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 371–381 (cit. on p. 35).
- [Gom+14] E Gomez-Nieto, Roman F San, P Pagliosa, et al. “Similarity Preserving Snippet-Based Visualization of Web Search Results”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.3 (2014), pp. 457–470 (cit. on p. 28).
- [Gör+18] Jochen Görtler, Christoph Schulz, Daniel Weiskopf, and Oliver Deussen. “Bubble Treemaps for Uncertainty Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 719–728 (cit. on p. 188).
- [GPC15] Ashwinkumar Ganesan, Shimei Pan, and Jian Chen. “LDAExplore : Visualizing Topic Models Generated using Latent Dirichlet Allocation”. In: *Proceedings of the International Conference on Intelligent User Interfaces, Workshop on Visual Text Analytics* (2015) (cit. on p. 36).
- [Gre+12] Brynjar Gretarsson, John O’donovan, Svetlin Bostandjiev, et al. “TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling”. In: *Transactions on Intelligent Systems and Technology* 3.2 (2012), pp. 23–49 (cit. on p. 35).
- [GRE15] Valentin Gold, Christian Rohrdantz, and Mennatallah El-Assady. “Exploratory Text Analysis using Lexical Episode Plots”. In: The Eurographics Association, 2015, pp. 85–89 (cit. on pp. 31, 32).
- [Gru09] Jonathan T. Grudin. “AI and HCI: Two Fields Divided by a Common Focus”. In: *AI Magazine* 30 (2009), pp. 48–57 (cit. on p. 23).
- [GS06] Tom Griffiths and Mark Steyvers. “Probabilistic Topic Models”. In: *Latent Semantic Analysis: A Road to Meaning* (2006) (cit. on p. 33).
- [GST07] Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. “Topics in Semantic Representation.” In: *Psychological review* 114.2 (2007), p. 211 (cit. on p. 35).
- [Gue+13] John Alexis Guerra-Gomez, Michael L. Pack, Catherine Plaisant, and Ben Shneiderman. “Visualizing Change over Time using Dynamic Hierarchies: TreeVersity2 and the StemView”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2566–2575 (cit. on p. 192).
- [Gui+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, et al. “A Survey Of Methods For Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5 (2018) (cit. on pp. 21, 22, 53, 54).
- [Gun16] David Gunning. *Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53*. Tech. rep. Defense Advanced Research Projects Agency (DARPA), 2016 (cit. on p. 20).

- [GX12] Yuhong Guo and Min Xiao. “Cross Language Text Classification via Subspace Co-Regularized multi-view learning”. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2012), pp. 1615–1622 (cit. on p. 28).
- [Har+02] Knut Hartmann, Stefan Schlechtweg, Ralf Helbing, and Th Strothotte. “Knowledge-Supported Graphical Illustration of Texts”. In: *Proceedings of International Conference on Advanced Visual Interfaces*. 2002, pp. 300–307 (cit. on p. 29).
- [Har15] Adam W. Harley. “An Interactive Node-Link Visualization of Convolutional Neural Networks”. In: *International Symposium on Visual Computing*. 2015, pp. 867–877 (cit. on p. 61).
- [Hav+01] Susan Havre, Elizabeth Hetzler, Ken Perrine, Elizabeth Jurrus, and Nancy Miller. “Interactive Visualization of Multiple Query Results”. In: *Proceedings of IEEE Conference on Information Visualization*. 2001, pp. 105–112 (cit. on p. 28).
- [Hav+02] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. “ThemeRiver: Visualizing Thematic Changes in Large Document Collections”. In: *IEEE Transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 9–20 (cit. on pp. 29, 36).
- [Hav02] Taher H Haveliwala. “Topic-Sensitive Pagerank”. In: *Proceedings of the International Conference on World Wide Web*. 2002, pp. 517–526 (cit. on p. 28).
- [HB04] Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. Macmillan, 2004 (cit. on p. 173).
- [HBV02] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. “Clustering Validity Checking Methods: part II”. In: *ACM Sigmod Record* 31.3 (2002), pp. 19–27 (cit. on p. 98).
- [HC14] Enamul Hoque and Giuseppe Carenini. “ConVis: A Visual Text Analytic System for Exploring Blog Conversations”. In: *Computer Graphics Forum*. Vol. 33. 3. 2014, pp. 221–230 (cit. on p. 30).
- [HC15] Enamul Hoque and Giuseppe Carenini. “ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations”. In: *Proceedings of the International Conference on Intelligent User Interfaces*. ACM. 2015, pp. 169–180 (cit. on p. 36).
- [HDA13] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. “Contextifier: Automatic Generation of Annotated Stock Visualizations”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2013, pp. 2707–2716 (cit. on p. 29).
- [He+17] Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P Xing. “Efficient Correlated Topic Modeling with Topic Embedding”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2017, pp. 225–233 (cit. on p. 29).
- [Hea06] Marti A. Hearst. “Clustering Versus Faceted Categories for Information Exploration”. In: *Communications of the ACM* 49 (2006), pp. 59–61 (cit. on p. 68).
- [Hei+12] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. “Visual Classifier Training for Text Document Retrieval”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2839–2848 (cit. on pp. 28, 35).
- [Hei+16] Florian Heimerl, Markus John, Qi Han, Steffen Koch, and Thomas Ertl. “DocuCompass: Effective Exploration of Document Landscapes”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2016, pp. 11–20 (cit. on p. 29).

- [HF96] Morten Hertzum and Erik Frøkjær. “Browsing and Querying in Online Documentation: A Study of User Interfaces and the Interaction Process”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 3 (1996), pp. 136–161 (cit. on p. 68).
- [HG16] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *arXiv preprint arXiv:1610.02136* (2016) (cit. on p. 51).
- [HHL17] Quang Minh Hoang, Trong Nghia Hoang, and Kian Hsiang Low. “A Generalized Stochastic Variational Bayesian Hyperparameter Learning Framework for Sparse Spectrum Gaussian Process Regression”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 2007–2014 (cit. on p. 29).
- [Hil04] Denis J. Hilton. “Knowledge-Based Causal Attribution : The Abnormal Conditions Focus Model”. In: 2004 (cit. on p. 25).
- [Hil90] Denis J. Hilton. “Conversational Processes and Causal Explanation”. In: *Psychological Bulletin* 107.1 (1990), pp. 65–81 (cit. on p. 25).
- [HJ07] Denis J Hilton and L McCLURE John. “The Course of Events: Counterfactuals, Causal Sequences, and Explanation”. In: *The psychology of counterfactual thinking*. Routledge, 2007, pp. 56–72 (cit. on p. 25).
- [HK04] Khaled M Hammouda and Mohamed S Kamel. “Efficient Phrase-Based Document Indexing for Web Document Clustering”. In: *IEEE Transactions on Knowledge and Data Engineering* 16.10 (2004), pp. 1279–1296 (cit. on p. 33).
- [Hof+13] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. “Trust in Automation”. In: *IEEE Intelligent Systems* 28.1 (2013), pp. 84–88 (cit. on p. 54).
- [Hof+18] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. “Metrics for Explainable AI: Challenges and Prospects”. In: *arXiv:1812.04608 [cs]* (2018) (cit. on p. 54).
- [Hoh+18] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. “Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2018), pp. 2674–2693 (cit. on pp. 24, 48).
- [Hoh+19a] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. “Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019, 579:1–579:13 (cit. on pp. 21, 22).
- [Hoh+19b] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. “Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2019), pp. 1096–1106 (cit. on pp. 22, 24).
- [Hol16] Andreas Holzinger. “Interactive Machine Learning for Health Informatics: When Do we Need the Human-in-the-Loop?” In: *Brain Informatics* 3.2 (2016), pp. 119–131 (cit. on p. 21).
- [Hor99] Eric Horvitz. “Principles of Mixed-Initiative User Interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999, pp. 159–166 (cit. on p. 20).

- [HR03] Geoffrey E Hinton and Sam T Roweis. “Stochastic Neighbor Embedding”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2003, pp. 857–864 (cit. on pp. 85, 88).
- [Hsi+17] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, et al. “Collaborative Metric Learning”. In: *Proceedings of the International Conference on World Wide Web*. 2017, pp. 193–201 (cit. on p. 29).
- [HSK95] Denis J. Hilton, Richard H. Smith, and Sung Hee Kin. “Processes of Causal Explanation and Dispositional Attribution.” In: *Journal of Personality and Social Psychology* 68 (1995), pp. 377–387 (cit. on p. 25).
- [HSS10] Derek L. Hansen, Ben Shneiderman, and Marc Smith. “Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights”. In: *Active Media Technology*. Springer, 2010, pp. 47–62 (cit. on p. 30).
- [Hu+14] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. “Interactive Topic Modeling”. In: *Machine learning* 95.3 (2014), pp. 423–469 (cit. on p. 36).
- [Hua+10] Jian Huang, Jianfeng Gao, Jiangbo Miao, et al. “Exploring Web Scale Language Models for Search Query Processing”. In: *Proceedings of the International Conference on World Wide Web*. 2010, pp. 451–460 (cit. on p. 29).
- [Hub08] Alexander C. Hubmann-Haidvogel. “ThreadVis for Thunderbird: A Thread Visualization Extension for the Mozilla Thunderbird Email client”. MA thesis. Graz University of Technology, 2008 (cit. on p. 30).
- [Hus+17] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. “Imitation Learning: A Survey of Learning Methods”. In: *ACM Computing Surveys (CSUR)* 50.2 (2017), pp. 1–35 (cit. on p. 46).
- [IBL05] Renáta Iváncsy, Attila Babos, and Csaba Legány. “Analysis and Extensions of Popular Clustering Algorithms”. In: *International Symposium of Hungarian Researchers on Computational Intelligence*. 2005 (cit. on p. 98).
- [IPT16] Francesca Ieva, Anna Maria Paganoni, and Nicholas Tarabelloni. “Covariance-Based Clustering in Multivariate and Functional Data Analysis”. In: *Journal of Machine Learning Research* 17.1 (2016), pp. 4985–5005 (cit. on p. 34).
- [Ise+13] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. “A Systematic Review on the Practice of Evaluating Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827 (cit. on pp. 100, 113, 130).
- [Ise+17] Petra Isenberg, Florian Heimerl, Steffen Koch, et al. “vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (2017), pp. 2199–2206 (cit. on p. 29).
- [ISU17] Katsuhiko Ishiguro, Issei Sato, and Naonori Ueda. “Averaged Collapsed Variational Bayes Inference”. In: *Journal of Machine Learning Research* 18.1 (2017), pp. 1–29 (cit. on p. 29).
- [Jän+15] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges”. In: (2015), pp. 83–103 (cit. on pp. 6, 185).

- [Jän+17] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. “Visual Text Analysis in Digital Humanities”. In: *Computer Graphics Forum* 36.6 (2017), pp. 226–250 (cit. on p. 28).
- [JBS17] Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. “MEmbER: Max-Margin Based Embeddings for Entity Retrieval”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2017, pp. 783–792 (cit. on p. 28).
- [JDU12] Jagadeesh Jagarlamudi, Hal Daumé Iii, and Raghavendra Udupa. “Incorporating Lexical Priors into Topic Models”. In: *Proceedings Conference of the European Chapter of the Association for Computational Linguistics* (2012), pp. 204–213 (cit. on pp. 34, 186).
- [Jen+17] Wolfgang Jentner, Mennatallah El-Assady, Bela Gipp, and Daniel A Keim. “Feature Alignment for the Analysis of Verbatim Text Transcripts”. In: *Proceedings of Eurographics Conference on Visualization, Workshop on Visual Analytics (EuroVA)*. 2017, pp. 13–17 (cit. on p. 31).
- [Jen+18] Wolfgang Jentner, Rita Sevastjanova, Florian Stoffel, et al. “Minions, Sheep, and Fruits: Metaphorical Narratives to Explain Artificial Intelligence and Build Trust”. In: *Proceedings of the IEEE Visualization Conference, Workshop on Machine Learning from User Interaction for Visualization and Analytics*. 2018 (cit. on pp. 24, 26, 62).
- [JKM12] Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios. “Relative N-Gram Signatures: Document Visualization at the Level of Character N-Grams”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2012, pp. 103–112 (cit. on p. 28).
- [JLC19] Liu Jiang, Shixia Liu, and Changjian Chen. “Recent Research Advances on Interactive Machine Learning”. In: *Journal of Visualization* 22.2 (2019), pp. 401–417 (cit. on p. 22).
- [Jon72] Karen Spärck Jones. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”. In: *Journal of Documentation* 28.1 (1972), pp. 11–21 (cit. on pp. 32, 54, 87, 99, 118).
- [JSG06] Alejandro Jaimes, Nicu Sebe, and Daniel Gatica-Perez. “Human-Centered Computing: A Multimedia Perspective”. In: *Proceedings of the ACM International Conference on Multimedia*. 2006, pp. 855–864 (cit. on p. 23).
- [JW17] Stefan Jänicke and David Joseph Wrisley. “Interactive Visual Alignment of Medieval Text Versions”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2017, pp. 127–138 (cit. on p. 29).
- [JZ03] Xiang Ji and Hongyuan Zha. “Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2003, pp. 322–329 (cit. on p. 34).
- [JZ15] Rie Johnson and Tong Zhang. “Semi-Supervised Convolutional Neural Networks for Text Categorization via Region Embedding”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2015, pp. 919–927 (cit. on p. 29).
- [KAC17] Natchaya Kijmongkolchai, Alfie Abdul-Rahman, and Min Chen. “Empirically Measuring Soft Knowledge in Visualization”. In: *Computer Graphics Forum* 36.3 (2017), pp. 73–85 (cit. on p. 66).
- [Kah+18] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Polo Chau. “ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 88–97 (cit. on p. 24).

- [Kah+19] Minsuk Kahng, Nikhil Thorat, Duen Horng Polo Chau, Fernanda B. Viegas, and Martin Wattenberg. “GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 310–320 (cit. on p. 24).
- [Kah11] Daniel Kahneman. *Thinking, Fast and Slow*. Vol. 1. Farrar, Straus and Giroux New York, 2011 (cit. on pp. 67, 72, 203).
- [Kat+11] Saurabh S Kataria, Krishnan S Kumar, Rajeev R Rastogi, Prithviraj Sen, and Srinivasan H Sengamedu. “Entity Disambiguation with Hierarchical Topic Models”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2011, pp. 1037–1045 (cit. on p. 29).
- [Kau+20] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, et al. “Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–14 (cit. on p. 23).
- [KBK11] Milos Krstajic, Enrico Bertini, and Daniel Keim. “CloudLines: Compact Display of Event Episodes in Multiple Time-Series”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2432–2439 (cit. on p. 29).
- [KCP15] Da Kuang, Jaegul Choo, and Haesun Park. “Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering”. In: *Partitional Clustering Algorithms*. Springer, 2015, pp. 215–243 (cit. on p. 33).
- [KDK03] Krishna Kummamuru, Ajay Dhawale, and Raghu Krishnapuram. “Fuzzy Co-Clustering of Documents and Keywords”. In: *Proceedings of IEEE Conference on Fuzzy Systems*. Vol. 2. IEEE. 2003, pp. 772–777 (cit. on p. 33).
- [Kei+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, et al. “Visual Analytics: Definition, Process, and Challenges”. In: *Proceedings of IEEE Conference on Information Visualization*. Vol. 4950. Springer, 2008, pp. 154–175 (cit. on pp. 5, 194).
- [Ker03] Bernard Kerr. “Thread Arcs: An Email Thread Visualization”. In: *Proceedings of IEEE Conference on Information Visualization*. 2003, pp. 211–218 (cit. on pp. 30, 153).
- [KF14] Linda T. Kaastra and Brian D. Fisher. “Field Experiment Methodology for Pair Analytics”. In: *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM, 2014, pp. 152–159 (cit. on pp. 100, 160).
- [KHY15] Younghoon Kim, Jiawei Han, and Cangzhou Yuan. “TOPTRAC: Topical Trajectory Pattern Mining”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2015, pp. 587–596 (cit. on p. 28).
- [Kim+17] Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas Elmqvist. “TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 151–160 (cit. on pp. 29, 36).
- [KK15] Kostiantyn Kucher and Andreas Kerren. “Text Visualization Techniques: Taxonomy, Visual Survey, and Community Insights”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2015, pp. 117–121 (cit. on pp. xxii, 6, 27, 28).
- [KM16] Santosh KC and Arjun Mukherjee. “On the Temporal Dynamics of Opinion Spamming: Case Studies on Yelp”. In: *Proceedings of the International Conference on World Wide Web*. 2016, pp. 369–379 (cit. on p. 29).

- [Koc+11] S Koch, H Bosch, M Giereth, and T Ertl. “Iterative Integration of Visual Insights During Scalable Patent Search and Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.5 (2011), pp. 557–569 (cit. on p. 29).
- [Koc+14] Steffen Koch, Markus John, Michael Wörner, Andreas Müller, and Thomas Ertl. “VarifocalReader – In-Depth Visual Analysis of Large Text Documents”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1723–1732 (cit. on p. 28).
- [KP02] April Kontostathis and William M. Pottenger. *A Mathematical View of Latent Semantic Indexing: Tracing Term Co-Occurrences*. Tech. rep. 2002 (cit. on p. 34).
- [KP13] Da Kuang and Haesun Park. “Fast Rank-2 Nonnegative Matrix Factorization for Hierarchical Document Clustering”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2013, pp. 739–747 (cit. on p. 34).
- [Kra+17] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. “A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, pp. 162–172 (cit. on pp. 20, 24, 156).
- [KRS04] Jaap Kamps, Maarten de Rijke, and Börkur Sigurbjörnsson. “Length Normalization in XML Retrieval”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2004, pp. 80–87 (cit. on p. 28).
- [KSS21] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A Flaw in Human Judgment*. Little, Brown, 2021 (cit. on p. 109).
- [KUG14] M. Ali-ud-din Khan, Muhammad Fahim Uddin, and Navarun Gupta. “Seven V’s of Big Data Understanding Big Data to Extract Value”. In: *Proceedings of Conference of the American Society for Engineering Education*. 2014, pp. 1–5 (cit. on p. 51).
- [Kul+13] Todd Kulesza, Simone Stumpf, Margaret M. Burnett, et al. “Too Much, Too Little, or Just Right? Ways Explanations Impact End Users’ Mental Models”. In: *IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 2013, pp. 3–10 (cit. on p. 23).
- [Kun+19] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. “Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019, 487:1–487:12 (cit. on p. 70).
- [Kur13] Ray Kurzweil. *How to Create a Mind: The Secret of Human Thought Revealed*. Penguin Books London, 2013 (cit. on p. 139).
- [Kus+15] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. “From Word Embeddings to Document Distances”. In: *International Conference on Machine Learning*. 2015, pp. 957–966 (cit. on p. 38).
- [Kwo+16] B. C. Kwon, S-H Kim, S Lee, et al. “VisOHC: Designing Visual Analytics for Online Health Communities”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 71–80 (cit. on pp. 28, 29).
- [Kwo+19] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, et al. “RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 299–309 (cit. on p. 24).

- [LA99] Bjornar Larsen and Chinatsu Aone. “Fast and Effective Text Mining using Linear-Time Document Clustering”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 1999, pp. 16–22 (cit. on p. 34).
- [Lai+15] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. “Recurrent Convolutional Neural Networks for Text Classification”. In: 2015, pp. 2267–2273 (cit. on p. 29).
- [Lam+07] Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, and Frédéric Saubion. “SegGen: A Genetic Algorithm for Linear Text Segmentation”. In: *Proceedings International joint Conference on Artificial Intelligence*. 2007, pp. 1647–1652 (cit. on p. 34).
- [Lau+10] Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. “Best Topic Word Selection for Topic Labelling”. In: *Proceedings of the International Conference on Computational Linguistics: Posters*. ACL. 2010, pp. 605–613 (cit. on p. 33).
- [Lau+11] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. “Automatic Labelling of Topic Models”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics* (2011), pp. 1536–1545 (cit. on p. 33).
- [LBL16] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. “Interpretable Decision Sets: A Joint Framework for Description and Prediction”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM. 2016, pp. 1675–1684 (cit. on p. 21).
- [LCC13] Yunzhong Liu, Feng Chen, and Yi Chen. “Learning Thread Reply Structure on Patient Forums”. In: *Proceedings of International Workshop on Data management & Analytics for Healthcare* (2013), pp. 1–4 (cit. on p. 147).
- [LD97] Thomas K Landauer and Susan T. Dutnais. “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge”. In: *Psychological Review* (1997), pp. 211–240 (cit. on p. 34).
- [Lee+07] Bongshin Lee, George G. Robertson, Mary Czerwinski, and Cynthia Sims Parr. “CandidTree: Visualizing Structural Uncertainty in Similar Hierarchies”. In: *Proceedings of IEEE Conference on Information Visualization 6.3* (2007), pp. 233–246 (cit. on p. 192).
- [Lee+10] Bongshin Lee, Nathalie Henry Riche, Amy K Karlson, and Sheelagh Carpendale. “Spark-Clouds: Visualizing Trends in Tag Clouds”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1182–1189 (cit. on p. 29).
- [Lee+12a] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. “iVisClustering: An Interactive Visual Document Clustering via Topic Modeling”. In: *Computer Graphics Forum* 31.3 (2012), pp. 1155–1164 (cit. on p. 29).
- [Lee+12b] Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. “Joint Entity and Event Coreference Resolution across Documents”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2012), pp. 489–500 (cit. on p. 149).
- [Lee+15] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. “Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1603–1612 (cit. on p. 20).

- [Lee+17a] Juho Lee, Creighton Heaukulani, Zoubin Ghahramani, Lancelot F. James, and Seungjin Choi. “Bayesian Inference on Random Simple Graphs with Power Law Degree Distributions”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2017, pp. 2004–2013 (cit. on p. 29).
- [Lee+17b] Tak Yeon Lee, Alison Smith, Kevin Seppi, et al. “The Human Touch: How Non-Expert Users Perceive, Interpret, and Fix Topic Models”. In: *International Journal of Human-Computer Studies* 105 (2017), pp. 28–42 (cit. on pp. 6, 7, 37, 73, 112, 118, 134, 176, 178, 183, 186, 190).
- [Lep+18] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. “Fair, Transparent, and Accountable Algorithmic Decision-making Processes”. In: *Philosophy & Technology* 31 (2018), pp. 611–627 (cit. on p. 54).
- [LG14] Omer Levy and Yoav Goldberg. “Dependency-based word embeddings”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. Vol. 2. 2014, pp. 302–308 (cit. on p. 84).
- [LGM20] Q. Vera Liao, Daniel Gruen, and Sarah Miller. “Questioning the AI: Informing Design Practices for Explainable AI User Experiences”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020 (cit. on pp. 23, 70).
- [Li+13] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He. “Exploiting Hybrid Contexts for Tweet Segmentation”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2013, pp. 523–532 (cit. on p. 34).
- [Li+15] Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. “Sentence-level Emotion Classification with Label and Context Dependence”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2015, pp. 1045–1053 (cit. on p. 29).
- [Li+17] Chenliang Li, Yu Duan, Haoran Wang, et al. “Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings”. In: *ACM Transactions on Information Systems* 36.2 (2017), pp. 1–30 (cit. on pp. 33, 35).
- [Lim+19] Brian Y. Lim, Qian Yang, Ashraf M. Abdul, and Danding Wang. “Why these Explanations? Selecting Intelligibility Types for Explanation Goals”. In: *Proceedings of the International Conference on Intelligent User Interfaces, Workshop on Visual Text Analytics*. 2019 (cit. on p. 21).
- [Lim11] Brian Lim. “Improving Understanding, Trust, and Control with Intelligibility in Context-Aware Applications”. In: *Human-Computer Interaction* (2011) (cit. on p. 54).
- [Lin+08] Xu Ling, Qiaozhu Mei, ChengXiang Zhai, and Bruce Schatz. “Mining Multi-Faceted Overviews of Arbitrary Topics in a Text Collection”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2008, pp. 497–505 (cit. on p. 35).
- [Lin+09] Chen Lin, Jiang-Ming Yang, Rui Cai, Xin-Jing Wang, and Wei Wang. “Simultaneously Modeling Semantics and Structure of Threaded Discussions: A Sparse Coding Approach and Its Applications”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)* (2009), pp. 131–138 (cit. on pp. 147, 150).
- [Lip18] Zachary C. Lipton. “The Mythos of Model Interpretability”. In: *Journal of acmqueue* 61.10 (2018), pp. 36–43 (cit. on pp. 22, 54).

- [Liu+15] Xiaosheng Liu, Jia Zeng, Xi Yang, Jianfeng Yan, and Qiang Yang. “Scalable Parallel EM Algorithms for Latent Dirichlet Allocation in Multi-Core Systems”. In: *Proceedings of the International Conference on World Wide Web*. 2015, pp. 669–679 (cit. on p. 29).
- [Liu+16] Shixia Liu, Jialun Yin, Xiting Wang, et al. “Online Visual Analytics of Text Streams”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.11 (2016), pp. 2451–2466 (cit. on pp. 29, 30).
- [Liu+17a] Mengchen Liu, Jiabin Shi, Zhen Li, et al. “Towards Better Analysis of Deep Convolutional Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 91–100 (cit. on p. 24).
- [Liu+17b] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. “Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective”. In: *Visual Informatics* 1.1 (2017), pp. 48–56 (cit. on pp. 7, 22, 24, 37, 46, 53, 55, 57).
- [Liu+18a] Mengchen Liu, Shixia Liu, Hang Su, Kelei Cao, and Jun Zhu. “Analyzing the Noise Robustness of Deep Neural Networks”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2018, pp. 60–71 (cit. on pp. 24, 53).
- [Liu+18b] Mengchen Liu, Jiabin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. “Analyzing the Training Processes of Deep Generative Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 77–87 (cit. on p. 24).
- [Liu+18c] Shixia Liu, Xiting Wang, Christopher Collins, et al. “Bridging Text Visualization and Mining: A Task-driven Survey”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.7 (2018), pp. 2482–2504 (cit. on pp. 6, 16, 27–29, 34, 50).
- [LL17] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017, pp. 4768–4777 (cit. on p. 22).
- [LM06] Wei Li and Andrew McCallum. “Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. ACM, 2006, pp. 577–584 (cit. on p. 34).
- [LNB14] Jey Han Lau, David Newman, and Timothy Baldwin. “Machine Reading Tea Leaves : Automatically Evaluating Topic Coherence and Topic Model Quality”. In: *Proceedings Conference of the European Chapter of the Association for Computational Linguistics* (2014) (cit. on pp. 35, 37).
- [LNG09] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. “Topic-Link LDA: Joint Models of Topic and Author Community”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2009, pp. 665–672 (cit. on p. 34).
- [Loc+20] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, et al. “Object-Centric Learning with Slot Attention”. In: *Proceedings of the International Conference on Neural Information Processing Systems* 33 (2020), pp. 11525–11538 (cit. on p. 68).
- [Lod+02] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. “Text Classification using String Kernels”. In: *Journal of Machine Learning Research* 2 (2002), pp. 419–444 (cit. on p. 29).
- [Lu+15] Zhongqi Lu, Zhicheng Dou, Jianxun Lian, Xing Xie, and Qiang Yang. “Content-Based Collaborative Filtering for News Topic Recommendation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2015, pp. 217–223 (cit. on p. 28).

- [Lun+17] Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. “Tandem Anchoring: a Multiword Anchor Approach for Interactive Topic Modeling”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics* (2017), pp. 896–905 (cit. on p. 35).
- [Luo+12] D. Luo, J. Yang, M Krstajic, W Ribarsky, and D. A Keim. “EventRiver: Visually Exploring Text Collections with Temporal References”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.1 (2012), pp. 93–105 (cit. on p. 28).
- [Luo+17] Minnan Luo, Feiping Nie, Xiaojun Chang, et al. “Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 2308–2314 (cit. on p. 29).
- [Lv+11] Yuanhua Lv, Taesup Moon, Pranam Kolari, et al. “Learning to Model Relatedness for News Recommendation”. In: *Proceedings of the International Conference on World Wide Web*. 2011, pp. 57–66 (cit. on p. 28).
- [Ma+16] Jing Ma, Wei Gao, Prasenjit Mitra, et al. “Detecting Rumors from Microblogs with Recurrent Neural Networks”. In: *Proceedings International joint Conference on Artificial Intelligence*. 2016, pp. 3818–3824 (cit. on p. 29).
- [Ma+20] Yuxin Ma, Tiankai Xie, Jundong Li, and Ross Maciejewski. “Explaining Vulnerabilities to Adversarial Machine Learning through Visual Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 1075–1085 (cit. on p. 24).
- [MAA21] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. “A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications”. In: *International journal of environmental research and public health* 18.4 (2021), p. 2121 (cit. on pp. 21, 53).
- [Mac+11] A. M. MacEachren, A. Jaiswal, A. C. Robinson, et al. “SensePlace2: GeoTwitter Analytics Support for Situational Awareness”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2011, pp. 181–190 (cit. on p. 29).
- [Mad+14] K Madhavan, N Elmqvist, M Vorvoreanu, et al. “DIA2: Web-Based Cyberinfrastructure for Visual Analysis of Funding Portfolios”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1823–1832 (cit. on p. 28).
- [McC+16] N. McCurdy, J. Lein, K. Coles, and M. Meyer. “Poemage: Visualizing the Sonic Topology of a Poem”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 439–448 (cit. on p. 29).
- [McF+13] Daniel A McFarland, Daniel Ramage, Jason Chuang, et al. “Differentiating Language Usage Through Topic Models”. In: *Poetics* 41.6 (2013), pp. 607–625 (cit. on p. 34).
- [MCM83] Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell. “A Theory and Methodology of Inductive Learning”. In: *Machine Learning: An Artificial Intelligence Approach*. Springer, 1983, pp. 83–134 (cit. on p. 53).
- [MDB17] Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. “Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification using Convolutional Neural Network”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2017, pp. 377–387 (cit. on p. 28).
- [Meh+17] Hrim Mehta, Amira Chalbi, Fanny Chevalier, and Christopher Collins. “DataTours: A Data Narratives Framework”. In: *Proceedings of IEEE Conference on Information Visualization*. 2017 (cit. on p. 96).

- [Mei+08] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. “Topic Modeling with Network Regularization”. In: *Proceedings of the International Conference on World Wide Web*. 2008, pp. 101–110 (cit. on p. 34).
- [MH08] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605 (cit. on p. 85).
- [Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2013, pp. 3111–3119 (cit. on pp. 82, 118).
- [Mil19] Tim Miller. “Explanation in Artificial Intelligence: Insights from the Social Sciences”. In: *Artificial Intelligence* 267 (2019), pp. 1–38 (cit. on p. 25).
- [Mil95] George A Miller. “WordNet: A Lexical Database for English”. In: *Communications of the ACM* (1995), pp. 39–41 (cit. on p. 187).
- [Mim+11] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2011), pp. 262–272 (cit. on p. 38).
- [Min+09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. “Distant Supervision for Relation Extraction without Labeled Data”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. ACL. 2009, pp. 1003–1011 (cit. on pp. 183, 184).
- [Min+17] Yao Ming, Shaozu Cao, Ruixiang Zhang, et al. “Understanding Hidden Memories of Recurrent Neural Networks”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2017, pp. 13–24 (cit. on p. 24).
- [Mit+19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. “Model Cards for Model Reporting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 220–229 (cit. on p. 24).
- [MLM07] David Mimno, Wei Li, and Andrew McCallum. “Mixtures of Hierarchical Topics with Pachinko Allocation”. In: *Proceedings of International Conference on Machine Learning*. ACM. 2007, pp. 633–640 (cit. on pp. 33, 34).
- [Mol18] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. Lulu.com, 2018 (cit. on pp. 21, 54).
- [Moo16] Christopher E Moody. “Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec”. In: *arXiv preprint arXiv:1605.02019* (2016) (cit. on p. 33).
- [MQB19] Yao Ming, Huamin Qu, and Enrico Bertini. “Rulematrix: Visualizing and Understanding Classifiers with Rules”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 342–352 (cit. on pp. 22, 24).
- [MRS08] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Vol. 1. Cambridge University Press, 2008 (cit. on p. 117).
- [MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999 (cit. on pp. 32, 87, 118, 182).
- [MSG16] Gustavo Malkomes, Chip Schaff, and Roman Garnett. “Bayesian Optimization for Automated Model Selection”. In: *Proceedings of the International Conference on Neural Information Processing Systems* 29 (2016) (cit. on p. 62).

- [MSM18] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. “Methods for Interpreting and Understanding Deep Neural Networks”. In: *Digital Signal Processing* (2018) (cit. on pp. 53, 54).
- [Müh+14] Thomas Mühlbacher, Harald Piringer, Samuel Gratzl, Michael Sedlmair, and Marc Streit. “Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1643–1652 (cit. on pp. 7, 37).
- [Mun+03] Tamara Munzner, François Guimbretière, Serdar Tasiran, et al. “TreeJuxtaposer: Scalable Tree Comparison using Focus+ Context with Guaranteed Visibility”. In: *ACM Transactions on Graphics* 22.3 (2003), pp. 453–462 (cit. on p. 193).
- [Mun09] Tamara Munzner. “A Nested Model for Visualization Design and Validation”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 921–928 (cit. on p. 22).
- [Mur+18] Sugeerth Murugesan, Sana Malik, Fan Du, Eunye Koh, and Tuan Manh Lai. “DeepCompare : Visual and Interactive Comparison of Deep Learning Model Performance”. In: *Symposium on Visualization in Data Science (VDS)*. 2018 (cit. on pp. 24, 61).
- [Mur+19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. “Definitions, Methods, and Applications in Interpretable Machine Learning”. In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080 (cit. on pp. 25, 46, 53, 54).
- [MZ06] Qiaozhu Mei and ChengXiang Zhai. “A Mixture Model for Contextual Text Mining”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2006, pp. 649–655 (cit. on p. 34).
- [MZR20] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems*. 2020 (cit. on pp. 21, 22, 46, 49).
- [NBB11] David Newman, Edwin V Bonilla, and Wray Buntine. “Improving Topic Coherence with Regularized Topic Models”. In: *Proceedings of the International Conference on Neural Information Processing Systems* 24 (2011) (cit. on p. 35).
- [NBR12] Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. “SITS: A Hierarchical Nonparametric Model using Speaker Identity for Topic Segmentation in Multiparty Conversations”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2012, pp. 78–87 (cit. on p. 35).
- [Ngu+14] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, et al. “Modeling Topic Control to Detect Influence in Conversations using Nonparametric Topic Models”. In: *Machine Learning* 95.3 (2014), pp. 381–421 (cit. on p. 34).
- [Ngu+15] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. “Improving Topic Models with Latent Feature Word Representations”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 299–313 (cit. on p. 33).
- [NKK17] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. “Topic Modelling for Qualitative Studies”. In: *Journal of Information Science* 43.1 (2017), pp. 88–102 (cit. on p. 38).
- [Nor83] Donald A Norman. “Some Observations on Mental Models”. In: *Mental models*. Psychology Press, 1983, pp. 15–22 (cit. on p. 23).

- [NQ17] Andrew P Norton and Yanjun Qi. “Adversarial-Playground: A Visualization Suite Showing How Adversarial Examples Fool Deep Learning”. In: *IEEE Symposium on Visualization for Cyber Security (VizSec)*. IEEE. 2017, pp. 1–4 (cit. on p. 22).
- [OC03] Paul Ogilvie and Jamie Callan. “Combining Document Representations for Known-Item Search”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2003, pp. 143–150 (cit. on p. 29).
- [Oel+09] Daniela Oelke, Ming Hao, Christian Rohrdantz, et al. “Visual Opinion Analysis of Customer Feedback Data”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2009, pp. 187–194 (cit. on p. 28).
- [Oel+10] D Oelke, D Spretke, A Stoffel, and D. A Keim. “Visual Readability Analysis: How to Make Your Writings Easier to Read”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2010, pp. 123–130 (cit. on p. 28).
- [Oel+14] Daniela Oelke, Hendrik Strobel, Christian Rohrdantz, Iryna Gurevych, and Oliver Deussen. “Comparative Exploration of Document Collections: A Visual Analytics Approach”. In: *Computer Graphics Forum*. Vol. 33. 3. 2014, pp. 201–210 (cit. on p. 117).
- [OG88] Hiroshi G Okuno and Anoop Gupta. “Parallel Execution of OPS5 in QLISP”. In: *Proceedings Conference on Artificial Intelligence Applications*. IEEE. 1988, pp. 268–273 (cit. on pp. 178, 190).
- [OKK13] D. Oelke, D. Kokkinakis, and D. A Keim. “Fingerprint Matrices: Uncovering the Dynamics of Social Networks in Prose Literature”. In: *Computer Graphics Forum* 32.3 (2013), pp. 371–380 (cit. on p. 29).
- [OMS17] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. “Feature Visualization”. In: *Distill* (2017) (cit. on p. 22).
- [Par+18] Deokgun Park, Seungyeon Kim, Jurim Lee, et al. “ConceptVector: Text Visual Analytics via Interactive Lexicon Building using Word Embedding”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 361–370 (cit. on pp. 36, 80).
- [Per+11] Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. “Hierarchically Supervised Latent Dirichlet Allocation”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2011, pp. 2609–2617 (cit. on p. 28).
- [Pez+18] Nicola Pezzotti, Thomas Höllt, Jan C. van Gemert, et al. “DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 98–108 (cit. on p. 24).
- [Pin+03] David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. “TableExtraction using Conditional Random Fields”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2003, pp. 235–242 (cit. on p. 29).
- [Pow08] David M.W. Powers. “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation”. In: *Journal of Machine Learning Technologies* (2008), pp. 2229–3981 (cit. on p. 62).
- [Pro+17] Julia Proskurnia, Marc-Allen Cartright, Lluís Garcia-Pueyo, et al. “Template Induction over Unstructured Email Corpora”. In: *Proceedings of the International Conference on World Wide Web*. 2017, pp. 1521–1530 (cit. on p. 28).

- [Pur+06] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. “Unsupervised Topic Modelling for Multi-Party Spoken Discourse”. In: *Proceedings Annual Meeting of the Association for Computational Linguistics*. ACL. 2006, pp. 17–24 (cit. on p. 33).
- [Qui+09] Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009 (cit. on p. 62).
- [Rag+10] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, et al. “A Multi-Pass Sieve for Coreference Resolution”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2010, pp. 492–501 (cit. on p. 150).
- [Ram+04] Lakshmith Ramaswamy, Arun Iyengar, Ling Liu, and Fred Douglass. “Automatic Detection of Fragments in Dynamically Generated Web Pages”. In: *Proceedings of the International Conference on World Wide Web*. 2004, pp. 443–454 (cit. on p. 28).
- [Ram+09a] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. “Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL. 2009, pp. 248–256 (cit. on p. 33).
- [Ram+09b] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. “Topic Modeling for the Social Sciences”. In: *Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Workshop on Applications for Topic Models: Text and Beyond*. 2009 (cit. on p. 37).
- [Ram+20] Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. “Model Agnostic Multilevel Explanations”. In: *Proceedings of the International Conference on Neural Information Processing Systems* 33 (2020), pp. 5968–5979 (cit. on pp. 21, 24).
- [Rau17] Rauber, Paulo E. and Fadel, Samuel G. and Falcão, Alexandre X. and Telea, Alexandru C. “Visualizing the Hidden Activity of Artificial Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* (2017) (cit. on p. 24).
- [RC09] Mary Beth Rosson and John M Carroll. “Scenario Based Design”. In: *Human-Computer Interaction* (2009), pp. 145–162 (cit. on p. 68).
- [RD18] Andrew Slavin Ross and Finale Doshi-Velez. “Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018 (cit. on pp. 22, 54).
- [Rei+10] Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. “Spherical Topic Models”. In: *Proceedings International Conference on Machine Learning* (2010), pp. 903–910 (cit. on p. 34).
- [Ren+14] Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn van Dolen, and Maarten de Rijke. “Hierarchical Multi-Label Classification of Social Text Streams”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2014, pp. 213–222 (cit. on p. 28).
- [RG00] Paul Rayson and Roger Garside. “Comparing Corpora using Frequency Profiling”. In: *Proceedings Workshop on Comparing Corpora-Volume 9*. ACL. 2000, pp. 1–6 (cit. on p. 87).
- [Rie19] Mark O. Riedl. “Human-Centered Artificial Intelligence and Machine Learning”. In: *Human Behavior and Emerging Technologies* 1.1 (2019), pp. 33–36 (cit. on p. 23).

- [Ron14] Odora Ronald James. “Using Explanation as a Teaching Method: How Prepared are High School Technology Teachers in Free State Province, South Africa”. In: *Journal of Social Sciences*. 2014, 38:71–81 (cit. on p. 56).
- [Ros+04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. “The Author-Topic Model for Authors and Documents”. In: *Proceedings Conference on Uncertainty in Artificial Intelligence*. AUA Press. 2004, pp. 487–494 (cit. on pp. 33, 34).
- [RS08] Marc’Aurelio Ranzato and Martin Szummer. “Semi-Supervised Learning of Compact Document Representations with Deep Networks”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2008, pp. 792–799 (cit. on p. 29).
- [RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2016, pp. 1135–1144 (cit. on pp. 21, 22, 54, 60).
- [RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Anchors: High-Precision Model-Agnostic Explanations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018) (cit. on pp. 54, 60).
- [Rud19] Cynthia Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215 (cit. on p. 21).
- [RW11] Mark D Reid and Robert C Williamson. “Information, Divergence and Risk for Binary Experiments”. In: *Journal of Machine Learning Research* 12 (2011), pp. 731–817 (cit. on p. 28).
- [RW16] Md Mustafizur Rahman and Hongning Wang. “Hidden Topic Sentiment Model”. In: *Proceedings of the International Conference on World Wide Web*. 2016, pp. 155–165 (cit. on p. 29).
- [RWJ16] Yafeng Ren, Ruimin Wang, and Donghong Ji. “A Topic-Enhanced Word Embedding for Twitter Sentiment Classification”. In: *Information Sciences* 369 (2016), pp. 188–198 (cit. on p. 33).
- [Sac+14] Dominik Sacha, Andreas Stoffel, Florian Stoffel, et al. “Knowledge Generation Model for Visual Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1604–1613 (cit. on pp. 26, 46, 51, 62).
- [Sac+16] Dominik Sacha, Michael Sedlmair, Leishi Zhang, et al. “Human-Centered Machine Learning through Interactive Visualization”. In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 2016, pp. 641–646 (cit. on pp. 25, 26).
- [Sac+19] Dominik Sacha, Matthias Kraus, Daniel A. Keim, and Min Chen. “VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 385–395 (cit. on p. 22).
- [Sad+17] Afshin Sadeghi, Christoph Lange, Maria-Esther Vidal, and Sören Auer. “Integration of Scholarly Communication Metadata using Knowledge Graphs”. In: *Proceedings International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 328–341 (cit. on p. 80).

- [San+14] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms”. In: *Data and Discrimination: Converting Critical Concerns Into Productive Inquiry* (2014), pp. 1–23 (cit. on pp. 54, 217).
- [Sap+16] Upendra Sapkota, Thamar Solorio, Manuel Montes-y-Gómez, and Steven Bethard. “Domain Adaptation for Authorship Attribution: Improved Structural Correspondence Learning”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2016, pp. 2226–2235 (cit. on p. 28).
- [SBN14] Alessandro Sordoni, Yoshua Bengio, and Jian-Yun Nie. “Learning Concept Embeddings for Query Expansion by Quantum Entropy Minimization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2014, pp. 1586–1592 (cit. on p. 28).
- [SC02] Luo Si and Jamie Callan. “Using Sampled Data and Regression to Merge Search Engine Results”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2002, pp. 19–26 (cit. on p. 29).
- [SC04] Chirag Shah and W Bruce Croft. “Evaluating High Accuracy Retrieval Techniques”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2004, pp. 2–9 (cit. on p. 29).
- [SC16] Suwon Suh and Seungjin Choi. “Two-Dimensional Correlated Topic Models”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016 (cit. on p. 34).
- [Sch+02] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. “Methods and Metrics for Cold-Start Recommendations”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM. 2002, pp. 253–260 (cit. on pp. 84, 115).
- [Sch12] Benjamin M Schmidt. “Words Alone: Dismantling Topic Models in the Humanities”. In: *Journal of Digital Humanities* 2.1 (2012), pp. 49–65 (cit. on pp. 7, 35).
- [SCH17] Robert Speer, Joshua Chin, and Catherine Havasi. “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017 (cit. on p. 87).
- [Sch93] Hinrich Schütze. “Word Space”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 1993, pp. 895–902 (cit. on p. 29).
- [SCK18] Dirk Streeb, Min Chen, and Daniel A Keim. “The Biases of Thinking Fast and Thinking Slow”. In: *Cognitive Biases in Visualizations*. Springer, 2018, pp. 97–107 (cit. on p. 67).
- [SCS09] Jangwon Seo, W. Bruce Croft, and David a. Smith. “Online Community Search using Thread Structure”. In: *Proceedings of the International Conference on Information and Knowledge Management* (2009), pp. 1907–1910 (cit. on pp. 147, 148).
- [Sed+14] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. “Visual Parameter Space Analysis: A Conceptual Framework”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2161–2170 (cit. on p. 117).
- [Sel+17] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626 (cit. on p. 22).

- [Sen12] Prithviraj Sen. “Collective Context-Aware Topic Models for Entity Disambiguation”. In: *Proceedings of the International Conference on World Wide Web*. 2012, pp. 729–738 (cit. on p. 29).
- [Sev+18a] Rita Sevastjanova, Mennatallah El-Assady, Annette Hautli-Janisz, et al. “Mixed-Initiative Active Learning for Generating Linguistic Insights in Question Classification”. In: *Proceedings of the IEEE Visualization Conference, Workshop on Data Systems for Interactive Analysis (DSIA)*. 2018 (cit. on p. 31).
- [Sev+18b] Rita Sevastjanova, Fabian Beck, Basil Ell, et al. “Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models”. In: *Proceedings Workshop on Visualization for AI Explainability*. 2018 (cit. on pp. 31, 56, 216).
- [SGA14] Damiano Spina, Julio Gonzalo, and Enrique Amigó. “Learning Similarity Functions for Topic Detection in Online Reputation Monitoring”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2014, pp. 527–536 (cit. on p. 28).
- [SH10] Edward Segel and Jeffrey Heer. “Narrative Visualization: Telling Stories with Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1139–1148 (cit. on p. 24).
- [She+17a] Qiaomu Shen, Tongshuang Wu, Haiyan Yang, et al. “NameClarifier: A Visual Analytics System for Author Name Disambiguation”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 141–150 (cit. on p. 29).
- [She+17b] Yikang Shen, Wenge Rong, Nan Jiang, et al. “Word Embedding Based Correlation Model for Question/Answer Matching.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 3511–3517 (cit. on p. 28).
- [SHM14] Alison Smith, Timothy Hawes, and Meredith Myers. “Hiérarchie: Interactive Visualization for Hierarchical Topic Models”. In: *Proceedings Workshop on Interactive Language Learning, Visualization, and Interfaces*. 2014, p. 71 (cit. on p. 36).
- [Shn20] Ben Shneiderman. “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy”. In: *International Journal of Human–Computer Interaction* 36.6 (2020), pp. 495–504 (cit. on p. 23).
- [Sim+17] Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, et al. “Machine Teaching: A New Paradigm for Building Machine Learning Systems”. In: *arXiv preprint arXiv:1707.06742* (2017) (cit. on pp. 26, 46, 62, 66, 68, 82, 83).
- [Sim73] Herbert A Simon. “The Structure of Ill Structured Problems”. In: *Artificial intelligence* 4.3-4 (1973), pp. 181–201 (cit. on p. 220).
- [Sin01] Amit Singhal. “Modern Information Retrieval: A Brief Overview”. In: *Bulletin of the IEEE CS Technical Committee on Data Engineering* 24.4 (2001), pp. 35–43 (cit. on pp. 117, 149).
- [SK16] Himani Sharma and Sunil Kumar. “A Survey on Decision Tree Algorithms of Classification in Data Mining”. In: *International Journal Science and Research* 5.4 (2016), pp. 2094–2097 (cit. on p. 151).
- [SKE20] Bruno Schneider, Daniel A Keim, and Mennatallah El-Assady. “Datashiftexplorer: Visualizing and Comparing Change in Multidimensional Data for Supervised Learning”. In: *International Journal Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP*. 2020, pp. 141–148 (cit. on pp. 53, 72).

- [Smi+17] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, et al. “Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 1–15 (cit. on p. 35).
- [SMM12] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. “Design Study Methodology: Reflections from the Trenches and the Stacks”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2431–2440 (cit. on p. 50).
- [Smo87] Paul Smolensky. “Connectionist AI, Symbolic AI, and the Brain”. In: *Artificial Intelligence Review* 1.2 (1987), pp. 95–109 (cit. on pp. 67, 72).
- [SMR07] Anne Schuth, Maarten Marx, and Maarten de Rijke. “Extracting the Discussion Structure in Comments on News-Articles”. In: *Proceedings of ACM International Workshop on Web Information and Data Management* (2007), pp. 97–104 (cit. on pp. 147, 159).
- [Smu19] Nathalie A Smuha. “The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence”. In: *Computer Law Review International* 20.4 (2019), pp. 97–106 (cit. on pp. 8, 20).
- [Son+10] Yangqiu Song, Shimei Pan, Shixia Liu, et al. “Constrained Coclustering for Textual Documents”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2010, pp. 581–586 (cit. on p. 29).
- [Son+17] Jeong-Woo Son, Junkey Jeon, Alex Lee, and Sun-Joong Kim. “Spectral Clustering with Brainstorming Process for Multi-View Data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 2548–2554 (cit. on p. 29).
- [Spe+18] Fabian Sperrle, Jürgen Bernard, Michael Sedlmair, Daniel A. Keim, and Mennatallah El-Assady. “Speculative Execution for Guided Visual Analytics”. In: *Proceedings of the IEEE Visualization Conference, Workshop on Machine Learning from User Interaction for Visualization and Analytics*. 2018 (cit. on p. 62).
- [Spe+19] Fabian Sperrle, Udo Schlegel, Mennatallah El-Assady, and Daniel A. Keim. “Human Trust Modeling for Bias Mitigation in Artificial Intelligence”. In: *ACM CHI Workshop: Where is the Human? Bridging the Gap Between AI and HCI*. 2019 (cit. on pp. 26, 72, 217).
- [Spe+20a] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, et al. *Should We Trust (X)AI? Design Dimensions for Structured Experimental Evaluations*. 2020 (cit. on pp. 22, 26, 46, 217).
- [Spe+20b] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel A. Keim, and Mennatallah El-Assady. “Learning and Teaching in Co-Adaptive Guidance for Mixed-Initiative Visual Analytics”. In: *Proceedings of Eurographics Conference on Visualization, Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2020, pp. 61–65 (cit. on p. 46).
- [Spe+21a] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, et al. “A Survey of Human-Centered Evaluations in Human-Centered Machine Learning”. In: *Computer Graphics Forum* (2021), pp. 543–568 (cit. on pp. 6, 23, 50, 53, 212).
- [Spe+21b] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel Keim, and Mennatallah El-Assady. “Co-adaptive Visual Data Analysis and Guidance Processes”. In: *Computers & Graphics* 100 (2021), pp. 93–105 (cit. on pp. 26, 46, 65, 214, 217).
- [Spe+21c] Fabian Sperrle, Hanna Schäfer, Daniel A. Keim, and Mennatallah El-Assady. “Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement”. In: *Computer Graphics Forum* 40.3 (2021), pp. 215–226 (cit. on pp. 214, 215).

- [Spi+20] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. “explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 1064–1074 (cit. on pp. 14, 24, 45, 46, 48, 55).
- [SS09] Dimitri Solomatine and Durga Shrestha. “A Novel Method to Estimate Model Uncertainty using Machine Learning Techniques”. In: *Water Resources Research* (2009) (cit. on p. 62).
- [SS16] Robert J. Sternberg and Karin Sternberg. *Cognitive Psychology*. Nelson Education, 2016 (cit. on p. 56).
- [SS97] Hinrich Schütze and Craig Silverstein. “Projections for Efficient Document Clustering”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 1997, pp. 74–81 (cit. on p. 29).
- [SSS04] Stefan Schlechtweg, Petra Schulze-Wollgast, and Heidrun Schumann. “Interactive Treemaps with Detail on Demand to Support Information Search in Documents”. In: *Proceedings of Eurographics Conference on Visualization (EuroVis)*. 2004, pp. 121–128 (cit. on p. 28).
- [Sta+07] John Stasko, Carsten Görg, Zhicheng Liu, and Kanupriya Singhal. “Jigsaw: Supporting- Investigative Analysis through Interactive Visualization”. In: *IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 2007, pp. 131–138 (cit. on p. 29).
- [Ste+12] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. “Exploring Topic Coherence over Many Models and Many Topics”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012, pp. 952–961 (cit. on p. 35).
- [Sti+16] Holger Stitz, Stefan Luger, Marc Streit, and Nils Gehlenborg. “AVOCADO: Visualization of Workflow-Derived Data Provenance for Reproducible Biomedical Research”. In: *Computer Graphics Forum*. Vol. 35. 3. 2016, pp. 481–490 (cit. on pp. 62, 186).
- [Str+12] H. Strobel, M. Spicker, A. Stoffel, D. Keim, and O. Deussen. “Rolled-Out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives”. In: *Computer Graphics Forum* 31.3 (2012), pp. 1135–1144 (cit. on p. 29).
- [Str+18a] Dirk Streeb, Rebecca Kehlbeck, Dominik Jäckle, and Mennatallah El-Assady. “Distances, Neighborhoods, or Dimensions? Projection Literacy for the Analysis of Multivariate Data”. In: *Proceedings Workshop on Visualization for AI Explainability*. 2018 (cit. on p. 88).
- [Str+18b] Hendrik Strobel, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. “LST-MVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 667–676 (cit. on pp. 24, 60).
- [Str+19] Hendrik Strobel, Sebastian Gehrmann, Michael Behrisch, et al. “Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 353–363 (cit. on p. 24).
- [Sun+17] Guodao Sun, Tan Tang, Tai-Quan Peng, Ronghua Liang, and Yingcai Wu. “SocialWave: Visual Analysis of Spatio-temporal Diffusion of Information on Social Media”. In: *Transactions on Intelligent Systems and Technology* 9.2 (2017), p. 15 (cit. on p. 29).
- [Sun20] S Shyam Sundar. “Rise of Machine Agency: A Framework for Studying the Psychology of Human-AI Interaction (HAI)”. In: *Journal of Computer-Mediated Communication* 25.1 (2020), pp. 74–88 (cit. on p. 23).

- [SVZ14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *International Conference on Learning Representations (ICLR)*. 2014 (cit. on p. 22).
- [SWY75] Gerard Salton, Anita Wong, and Chung-Shu Yang. “A Vector Space Model for Automatic Indexing”. In: *Communications of the ACM* 18.11 (1975), pp. 613–620 (cit. on p. 116).
- [Sza+16] Danielle Albers Szafr, Deidre Stuffer, Yusef Sohail, and Michael Gleicher. “TextDNA: Visualizing Word Usage with Configurable Colorfields”. In: *Computer Graphics Forum* 35.3 (2016), pp. 421–430 (cit. on p. 29).
- [Tan+09] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. “Social Influence Analysis in Large-Scale Networks”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2009, pp. 807–816 (cit. on p. 34).
- [Tan+12] Jiliang Tang, Huiji Gao, Huan Liu, and Atish Das Sarma. “eTrust: Understanding Trust Evolution in an Online World”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM. ACM, 2012, pp. 253–261 (cit. on p. 20).
- [Tan84] Sam Tanenhaus. *Literature Unbound: A Guide for the Common Reader*. Ballantine Books, 1984 (cit. on p. 219).
- [TC16] Evangelos Tripolitakis and Georgios Chalkiadakis. “Probabilistic Topic Modeling, Reinforcement Learning, and Crowdsourcing for Personalized Recommendations”. In: *Proceedings European Conference on Multi-Agent Systems*. 2016 (cit. on pp. 35, 128).
- [Teh+06] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. “Hierarchical Dirichlet Processes”. In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581 (cit. on pp. 29, 33, 34, 183, 184).
- [TG10] M Trampuš and M Grobelnik. “Visualization of Online Discussion Forums”. In: *Workshop on Pattern Analysis Applications* 11 (2010), pp. 134–141 (cit. on p. 30).
- [Tho+12] Dennis Thom, Harald Bosch, Steffen Koch, Michael Wörner, and Thomas Ertl. “Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2012, pp. 41–48 (cit. on p. 29).
- [TKC17] Gary K. L. Tam, Vivek Kothari, and Min Chen. “An Analysis of Machine-and Human-Analytics in Classification”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 71–80 (cit. on p. 22).
- [TM11] Nava Tintarev and Judith Masthoff. “Designing and Evaluating Explanations for Recommender Systems”. In: *Recommender Systems Handbook*. Springer, 2011, pp. 479–510 (cit. on p. 20).
- [TSS15] Philipp Thomann, Ingo Steinwart, and Nico Schmid. “Towards an Axiomatic Approach to Hierarchical Clustering of Measures”. In: *Journal of Machine Learning Research* 16 (2015), pp. 1949–2002 (cit. on p. 34).
- [Tur50] Alan M. Turing. “Computing Machinery and Intelligence”. In: *Mind* LIX.236 (1950), pp. 433–460 (cit. on p. xxv).
- [Ver+10] Jo Vermeulen, Geert Vanderhulst, Kris Luyten, and Karin Coninx. “PervasiveCrystal: Asking and Answering Why and Why Not Questions about Pervasive Computing Applications”. In: *International Conference on Intelligent Environments (IE)*. IEEE. 2010, pp. 271–276 (cit. on p. 21).
- [Vos14] Catalin Voss. “Building Topic Models Based on Anchor Words”. 2014 (cit. on p. 35).

- [VRW13] Corinna Vehlow, Thomas Reinhardt, and Daniel Weiskopf. “Visualizing Fuzzy Overlapping Communities in Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2486–2495 (cit. on pp. 153, 186).
- [VW11] Stef Van Den Elzen and Jarke J van Wijk. “Baobabview: Interactive Construction and Analysis of Decision Trees”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2011, pp. 151–160 (cit. on p. 22).
- [Wal+09] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. “Evaluation Methods for Topic Models”. In: *Proceedings of International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112 (cit. on p. 35).
- [Wal06] Hanna M Wallach. “Topic Modeling: Beyond Bag-of-Words”. In: *Proceedings of International Conference on Machine learning*. ACM. 2006, pp. 977–984 (cit. on p. 35).
- [Wan+07] Yi-Chia Wang, Mahesh Joshi, William W Cohen, and Carolyn Rosé. “Recovering Implicit Thread Structure in Newsgroup Style Conversations”. In: *Artificial Intelligence* (2007), pp. 152–160 (cit. on p. 147).
- [Wan+08] Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. “Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2008, pp. 307–314 (cit. on p. 29).
- [Wan+09] Xin-Jing Wang, Xudong Tu, Dan Feng, and Lei Zhang. “Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2009, pp. 179–186 (cit. on p. 28).
- [Wan+11] Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. “Predicting Thread Discourse Structure Over Technical Web Forums”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2011), pp. 13–25 (cit. on pp. 147, 148, 159).
- [Wan+13] Xiting Wang, Shixia Liu, Yangqiu Song, and Baining Guo. “Mining Evolutionary Multi-Branch Trees from Text Streams”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2013, pp. 722–730 (cit. on p. 29).
- [Wan+14] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. “Knowledge Graph and Text Jointly Embedding”. In: *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2014, pp. 1591–1601 (cit. on p. 80).
- [Wan+15] Florence Ying Wang, Arnaud Sallaberry, Karsten Klein, Masahiro Takatsuka, and Mathieu Roche. “SentiCompass: Interactive Visualization for Exploring and Comparing the Sentiments of Time-Varying Twitter Data”. In: *IEEE Pacific Visualization Symposium (PacificVis)*. 2015, pp. 129–133 (cit. on p. 29).
- [Wan+16a] Xiting Wang, Shixia Liu, Yang Chen, et al. “How Ideas Flow Across Multiple Social Groups”. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2016, pp. 51–60 (cit. on p. 28).
- [Wan+16b] Xiting Wang, Shixia Liu, Junlin Liu, et al. “TopicPanorama: A Full Picture of Relevant Topics”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.12 (2016), pp. 2508–2521 (cit. on p. 36).

- [Wan+16c] Yashen Wang, Heyan Huang, Chong Feng, et al. “Cse: Conceptual Sentence Embeddings Based on Attention Model”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*. 2016, pp. 505–515 (cit. on p. 29).
- [Wan+17] Chong Wang, Yining Wang, Po-Sen Huang, et al. “Sequence Modeling via Segmentations”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2017, pp. 3674–3683 (cit. on p. 34).
- [Wan+19a] Danding Wang, Qian Yang, Ashraf M. Abdul, and Brian Y. Lim. “Designing Theory-Driven User-Centric Explainable AI”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019, 601:1–601:15 (cit. on p. 54).
- [Wan+19b] Junpeng Wang, Liang Gou, Han-wei Shen, and Hao Yang. “DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 288–298 (cit. on p. 24).
- [War+01] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. “Interactive Machine Learning: Letting Users Build Classifiers”. In: *International Journal of Human-Computer Studies* 55.3 (2001), pp. 281–292 (cit. on p. 21).
- [WB19] Daniel S. Weld and Gagan Bansal. “The Challenge of Crafting Intelligible Intelligence”. In: *Communications of the ACM* 62.6 (2019), pp. 70–79 (cit. on pp. 21, 22, 54, 212, 217).
- [WC06] Xing Wei and W Bruce Croft. “LDA-Based Document Models for Ad-Hoc Retrieval”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2006, pp. 178–185 (cit. on p. 29).
- [Wei+10] Furu Wei, Shixia Liu, Yangqiu Song, et al. “TIARA: A Visual Exploratory Text Analytic System”. In: *Proceedings ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 153–162 (cit. on pp. 7, 36).
- [Wex+20] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, et al. “The What-If Tool: Interactive Probing of Machine Learning Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 56–65 (cit. on pp. 22, 24, 70).
- [WG12] K. Wongsuphasawat and D Gotz. “Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2659–2668 (cit. on p. 29).
- [Wil+21] Wesley Willett, Bon Adriel Aseniero, Sheelagh Carpendale, et al. “Perception! Immersion! Empowerment! Superpowers as Inspiration for Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), pp. 22–32 (cit. on p. 26).
- [WMR18] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: *Harvard Journal of Law & Technology* 31 (2018), pp. 841–887 (cit. on pp. 22, 70).
- [WMZ07] Gm Weiss, K McCarthy, and Bibi Zabbar. “Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?” In: *International Conference on Data Mining* (2007), pp. 1–7 (cit. on p. 148).
- [Won+18] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, et al. “Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 1–12 (cit. on pp. 24, 61).

- [Woo+07] J. Wood, J. Dykes, A. Slingsby, and K. Clarke. “Interactive Visual Exploration of a Large Spatio-Temporal Dataset: Reflections on a Geovisualization Mashup”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1176–1183 (cit. on p. 29).
- [Woo+18] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. “A Qualitative Exploration of Perceptions of Algorithmic Fairness”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–14 (cit. on p. 54).
- [WRK11] Franz Wanner, Thomas Ramm, and Daniel A. Keim. “ForAVis: Explorative User Forum Analysis”. In: *Proceedings of International Conference on Web Intelligence, Mining and Semantics*. ACM, 2011, 14:1–14:10 (cit. on p. 30).
- [WS68] Günter Wyszecki and WS Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulas*. Wiley, 1968 (cit. on p. 91).
- [Wu+10] Yingcai Wu, Furu Wei, Shixia Liu, et al. “OpinionSeer: Interactive Visualization of Hotel Customer Feedback”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1109–1118 (cit. on p. 28).
- [Wu+14] Yingcai Wu, Shixia Liu, Kai Yan, and Mengchen Liu. “OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1763–1772 (cit. on p. 29).
- [Wu+17] Zongda Wu, Li Lei, Guiling Li, et al. “A Topic Modeling Based Approach to Novel Document Automatic Summarization”. In: *Expert Systems with Applications* 84 (2017), pp. 12–23 (cit. on p. 34).
- [WV08] Martin Wattenberg and Fernanda B Viégas. “The Word Tree, an Interactive Visual Concordance”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1221–1228 (cit. on p. 29).
- [WWL19] Jiangtao Wang, Yasha Wang, and Qin Lv. “Crowd-Assisted Machine Learning: Current Issues and Future Directions”. In: *Computer* 52.1 (2019), pp. 46–53 (cit. on p. 22).
- [WY04] Matthew O. Ward and Jing Yang. “Interaction Spaces in Data and Information Visualization”. In: *Proceedings of the Joint Eurographics - IEEE TCVG Conference on Visualization*. 2004, pp. 137–146 (cit. on p. 47).
- [XG13] Min Xiao and Yuhong Guo. “A Novel Two-Step Method for Cross Language Representation Learning”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2013, pp. 1259–1267 (cit. on p. 28).
- [Xia+17] Cao Xiao, Ping Zhang, W Art Chaovalitwongse, Jianying Hu, and Fei Wang. “Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 1590–1596 (cit. on p. 29).
- [XLG03] Wei Xu, Xin Liu, and Yihong Gong. “Document Clustering Based on Non-Negative Matrix Factorization”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM. 2003, pp. 267–273 (cit. on pp. 33, 34).
- [Xu+21] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. “From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI”. In: *ArXiv abs/2105.05424* (2021) (cit. on p. 23).
- [XX13] Pengtao Xie and Eric P. Xing. “Integrating Document Clustering and Topic Modeling”. In: *arXiv preprint arXiv:1309.6874* (2013) (cit. on p. 33).

- [XYX15] Pengtao Xie, Diyi Yang, and Eric P Xing. “Incorporating Word Correlation Knowledge into Topic Modeling”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference, Human Language Technologies* (2015), pp. 725–734 (cit. on p. 35).
- [Yan+13] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. “A Biterm Topic Model for Short Texts”. In: *Proceedings of the International Conference on World Wide Web*. 2013, pp. 1445–1456 (cit. on pp. 35, 149).
- [Yan+14] Yi Yang, Shimei Pan, Doug Downey, and Kunpeng Zhang. “Active Learning with Constrained Topic Model”. In: *Proceedings Workshop on Interactive Language Learning, Visualization, and Interfaces* (2014) (cit. on p. 35).
- [Yan+20] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. “Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–13 (cit. on p. 23).
- [Yin+12] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. “Using Social Media to Enhance Emergency Situation Awareness”. In: *IEEE Intelligent Systems* 27.6 (2012), pp. 52–59 (cit. on p. 28).
- [YN16] Mikhail Yurochkin and XuanLong Nguyen. “Geometric Dirichlet Means Algorithm for Topic Inference”. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2016, pp. 2505–2513 (cit. on p. 29).
- [Yu+03] Shipeng Yu, Deng Cai, Ji-Rong Wen, and Wei-Ying Ma. “Improving Pseudo-Relevance Feedback in Web Information Retrieval using Web Page Segmentation”. In: *Proceedings of the International Conference on World Wide Web*. 2003, pp. 11–18 (cit. on p. 28).
- [Yua+10] Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. “A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification”. In: *Journal of Machine Learning Research* 11 (2010), pp. 3183–3234 (cit. on p. 29).
- [Yud08] Eliezer Yudkowsky. *Artificial Intelligence as a Positive and Negative Factor in Global Risk*. 2008 (cit. on p. 137).
- [YYQ17] Yi Yang, Quanming Yao, and Huamin Qu. “VISTopic: A Visual Analytics System for Making Sense of Large Document Collections using Hierarchical Topic Modeling”. In: *Visual Informatics* (2017), pp. 40–47 (cit. on p. 36).
- [ZCM02] Yi Zhang, Jamie Callan, and Thomas Minka. “Novelty and Redundancy Detection in Adaptive Filtering”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM. 2002, pp. 81–88 (cit. on p. 28).
- [ZDB17] He Zhao, Lan Du, and Wray Buntine. “A Word Embeddings Informed Focused Topic Model”. In: *Proceedings of The Asian Conference on Machine Learning (ACML)*. 2017, pp. 423–438 (cit. on p. 33).
- [Zha+07] Xian Zhang, Yu Hao, Xiaoyan Zhu, Ming Li, and David R Cheriton. “Information Distance from a Question to an Answer”. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2007, pp. 874–883 (cit. on p. 28).
- [Zha+11] Yuejie Zhang, Lei Cen, Cheng Jin, Xiangyang Xue, and Jianping Fan. “Learning Inter-Related Statistical Query Translation Models for ENGLISH-CHINESE Bi-Directional CLIR”. In: *Proceedings International joint Conference on Artificial Intelligence*. 2011, pp. 1915–1920 (cit. on p. 28).

- [Zha+12] Jian Zhao, Fanny Chevalier, Christopher Collins, and Ravin Balakrishnan. “Facilitating Discourse Analysis with Interactive Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2639–2648 (cit. on p. 28).
- [Zha+14] Jian Zhao, Nan Cao, Zhen Wen, et al. “#FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1773–1782 (cit. on p. 29).
- [Zha+15] Xianchao Zhang, Linlin Zong, Xinyue Liu, and Hong Yu. “Constrained NMF-Based Multi-View Clustering on Unmapped Data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2015, pp. 3174–3180 (cit. on p. 29).
- [Zha+19] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S. Ebert. “Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 364–373 (cit. on pp. 24, 61).
- [Zhe+11] Guoqing Zheng, Jinwen Guo, Lichun Yang, et al. “Mining Topics on Participations for Community Discovery”. In: *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR)*. 2011, pp. 445–454 (cit. on p. 34).
- [Zho+21] Lina Zhou, Souren Paul, Haluk Demirkan, et al. “Intelligence Augmentation: Towards Building Human-Machine Symbiotic Relationship”. In: *AIS Transactions on Human-Computer Interaction* 13.2 (2021), pp. 243–264 (cit. on p. 5).
- [ZKF05] Ying Zhao, George Karypis, and Usama Fayyad. “Hierarchical Clustering Algorithms for Document Datasets”. In: *Data mining and knowledge discovery* 10.2 (2005), pp. 141–168 (cit. on p. 90).
- [ZP17] Nevin L Zhang and Leonard KM Poon. “Latent Tree Analysis.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 4891–4898 (cit. on p. 29).

Websites and Online Resources

- [Ble] David M. Blei. *Latent Dirichlet Allocation in C*. URL: <http://www.cs.columbia.edu/~blei/lda-c/> (visited on Mar. 31, 2018) (cit. on pp. 197, 198).
- [Dav10] Mark Davies. *Corpus of Historical American English*. 2010. URL: <http://corpus.byu.edu/coha/> (visited on Mar. 31, 2018) (cit. on pp. 183, 184).
- [FHW16] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”, Fourth Edition*. 2016. URL: https://www.cs.waikato.ac.nz/ml/weka/Witten%5C_et%5C_al%5C_2016%5C_appendix.pdf (cit. on p. 151).
- [HLE19] AI HLEG. *Ethics Guidelines for Trustworthy AI*. 2019. URL: <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1> (visited on Dec. 16, 2020) (cit. on p. 20).
- [Lan] Ken Lang. *Home Page for 20 Newsgroups Data Set*. URL: <http://qwone.com/~jason/20Newsgroups/> (visited on Mar. 31, 2018) (cit. on pp. 197, 198).
- [MM02] Andrew Kachites McCallum and David Mimno. *MALLET: A Machine Learning for Language Toolkit*. 2002. URL: <http://mallet.cs.umass.edu> (cit. on pp. 7, 38).

- [Red17a] Reddit. *Reddit*. 2017. URL: <http://reddit.com> (visited on Dec. 3, 2017) (cit. on pp. 142, 143).
- [Red17b] Reddit. *Reddit – Greenhouse Gases Higher than any Time in 800,000 Years 'Shows Definite Human Effect'*. 2017. URL: https://www.reddit.com/r/worldnews/comments/6eljby/greenhouse%5C_gases%5C_higher%5C_than%5C_any%5C_time%5C_in%5C_800000/ (visited on Dec. 3, 2017) (cit. on p. 160).
- [Red17c] Reddit. *Reddit – Trump is Deporting Fewer Immigrants Than Obama, Including Criminals*. 2017. URL: https://www.reddit.com/r/politics/comments/6sy51s/trump%5C_is%5C_deporting%5C_fewer%5C_immigrants%5C_than%5C_obama/ (visited on Dec. 3, 2017) (cit. on p. 160).
- [Sac17] Maria Sacchetti. *Washington Post – Trump is Deporting Fewer Immigrants than Obama, Including Criminals*. 2017. URL: https://www.washingtonpost.com/local/immigration/trump-is-deporting-fewer-immigrants-than-obama-including-criminals/2017/08/10/d8fa72e4-7e1d-11e7-9d08-b79f191668ed%5C_story.html (visited on Dec. 3, 2017) (cit. on p. 160).
- [Sla17] Claire Slattery. *ABC News – Greenhouse Gases Higher than any Time in 800,000 Years 'Shows Definite Human Effect'*. 2017. URL: <http://mobile.abc.net.au/news/2017-06-01/greenhouse-gases-database-shows-co2-ch4-n2o-rising-relentlessly/8578918> (visited on Dec. 3, 2017) (cit. on p. 160).
- [Uni12] CNN Political Unit. *Transcript: Second Presidential Debate*. Oct. 2012. URL: <http://politicalticker.blogs.cnn.com/2012/10/16/transcript-second-presidential-debate/> (visited on Mar. 31, 2018) (cit. on pp. 183, 184, 189, 195, 198).

DISSERTATION

Levels of Explainability for Human-AI Interaction in Visual Text Analytics

Mennatallah El-Assady

