

Design, Prototypical Implementation, and Evaluation of an Active Machine Learning Service in the Context of Legal Text Classification

Johannes Muhr, Feb 13th 2017, Munich

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

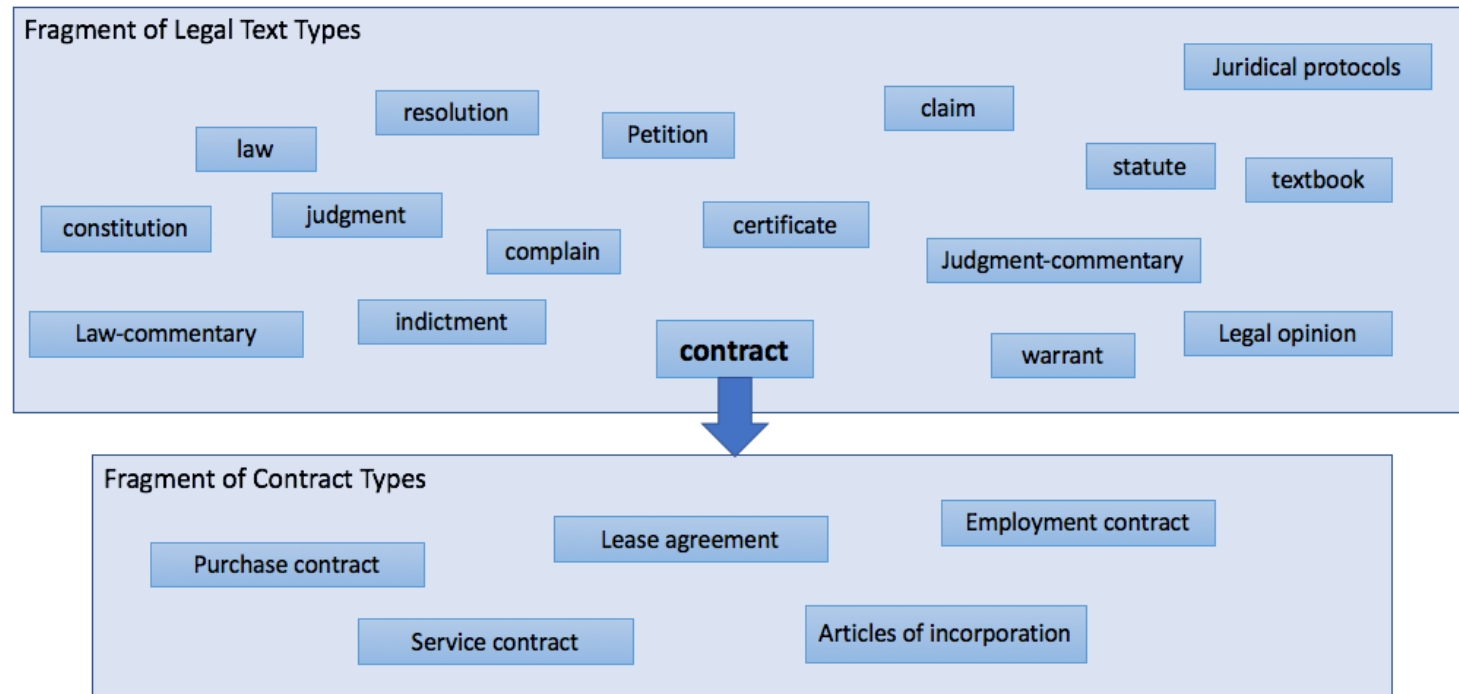
- Title (German) Design, Prototypische Implementierung, und Evaluation eines Active Machine Learning Services im Kontext von Rechtstexten

- Advisor Bernhard Walzl
- Supervisor Prof. Dr. Florian Matthes
- Project LexAlyze – Analysis of Legal Texts
- Chair Software Engineering for Business Information Systems (SEBIS)

- Student Johannes Muhr
- Start January, 15th 2017
- Submission July, 15th 2017

1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap

- Huge amount of legal documents are produced every day
- Many different kinds of legal documents



[1] Gruner, 2008, A Client's Analysis and Discussion of a Multi-Million Dollar Federal Lawsuit

- **Manual document classification is very expensive and time consuming**
 - 13,5 Million \$ were spent for classifying 1,6 Million items needing 4 month (= 8,50\$ per document) [1]
- **A lot of time is wasted with (document) discovery [2]**

	Senior Partner	Associate	Senior Associate	Junior Partner	Junior Partner	Senior Partner	All Others	Total	
	A	A	A	B	A	B	A & B	Hours	Dollars
Discovery									
Documents	60	311	162	95	57	39	197	921	294,455
Depositions	105	125	125	59	83	2	26	525	189,513
Total	165	436	287	154	141	41	223	1,446	483,968
Communications									
Internal	155	150	101	53	54	35	80	628	237,314
Opposition	66	149	86	24	44	0	41	411	137,216
Client	82	45	46	33	30	13	7	256	104,334
Total	303	344	232	110	128	48	128	1,295	478,864
Pleadings & Research									
Pleadings	35	282	162	44	37	18	65	643	197,190
Legal Research	8	126	114	1	2	1	126	377	99,909
Total	43	408	276	44	38	19	191	1,019	297,099
Expert Witness Support	35	21	25	253	22	70	31	457	203,303
Administration	-	6	4	-	-	-	973	983	130,663
All Other									
Settlement	32	69	19	21	5	14	-	159	58,244
Other	12	24	29	12	4	7	20	109	38,205
Hearings	7	3	5	-	2	-	1	18	6,977
Total	51	96	53	33	11	21	21	286	103,425
Total Hours	597	1,310	878	595	341	199	1,566	5,486	1,697,322
Total Dollars	305,117	301,202	258,724	250,005	146,439	138,305	297,531		

Hours:

$$\frac{1\,446}{5\,486} = 26,4\%$$

Dollars:

$$\frac{483\,986}{1\,697\,322} = 28,5\%$$

[1] Roitblat, H. L., et al. (2010). Document categorization in legal electronic discovery: computer classification vs. manual review.

[2] Gruner, (2008). A Client's Analysis and Discussion of a Multi-Million Dollar Federal Lawsuit

➤ Result

- **Document and Sentence classification** is a hot topic
- Manual classification is very expensive and time-consuming
- Machine learning approach is supposed to help here

➤ Solution Approaches

1. Use of (Ruta) Rules
2. Active Machine Learning (AL)
3. Combination of Ruta Rules and AL



1. Motivation

2. Active Learning

3. Research Questions

4. Solution Approach

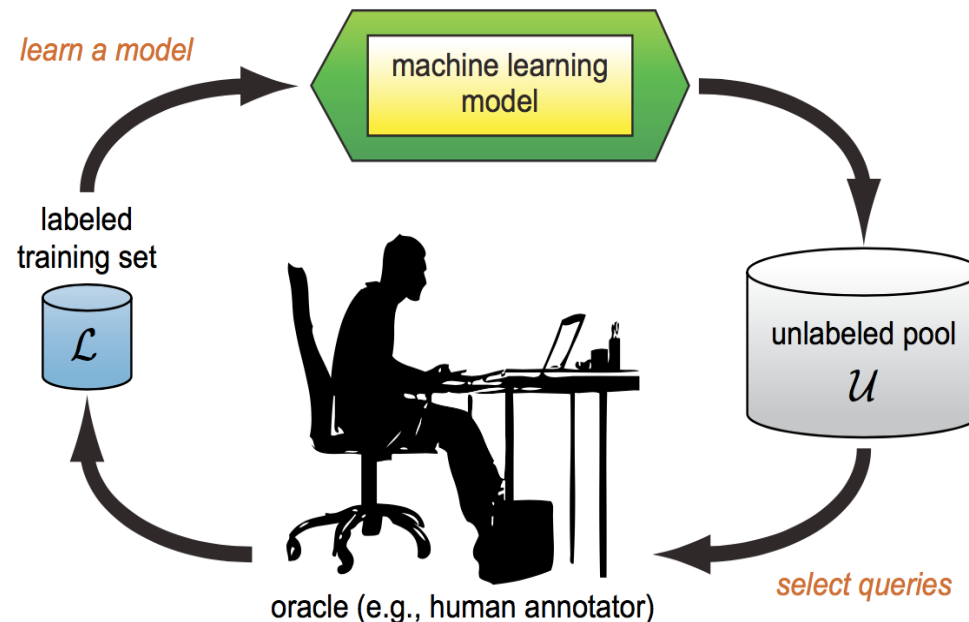
5. Roadmap

- **Why using Active Machine Learning for Document- & Sentence Classification?**
 - Detection of rules is limited
 - Minor linguistic variations are enough that sentences are not classified accordingly
 - “Im Sinne des Gesetzes” != “Im Sinne der Gesetze”
 - Active learning has already been successfully applied in
 - ✓ text classification [3]
 - ✓ and also within the legal environment [4]

[3] Novak, Mladenič, & Grobelnik, 2006; S. Tong & Koller, 2002; Segal, Markowitz, & Arnold, 2006

[4] Cardellino, Villata, Alemany, & Cabrio, 2015; Šavelka, Trivedi, & Ashley, 2015; Sunkle et al., 2016

- Subfield of machine learning with people in the loop (iterative & interactive form)
- **Goal:** Reduce size of needed trainings data by labelling those instances that are especially helpful
- Many influencing factors need to be considered (e.g. classifier, query strategy)



- **Document classification**
 - >100 000 documents
 - Manually labelled set of documents received from Datev
- **Sentence Classification**
 - Available from laws (Lexia)
 - Manual classification with the help of Elena Scepankova



1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap

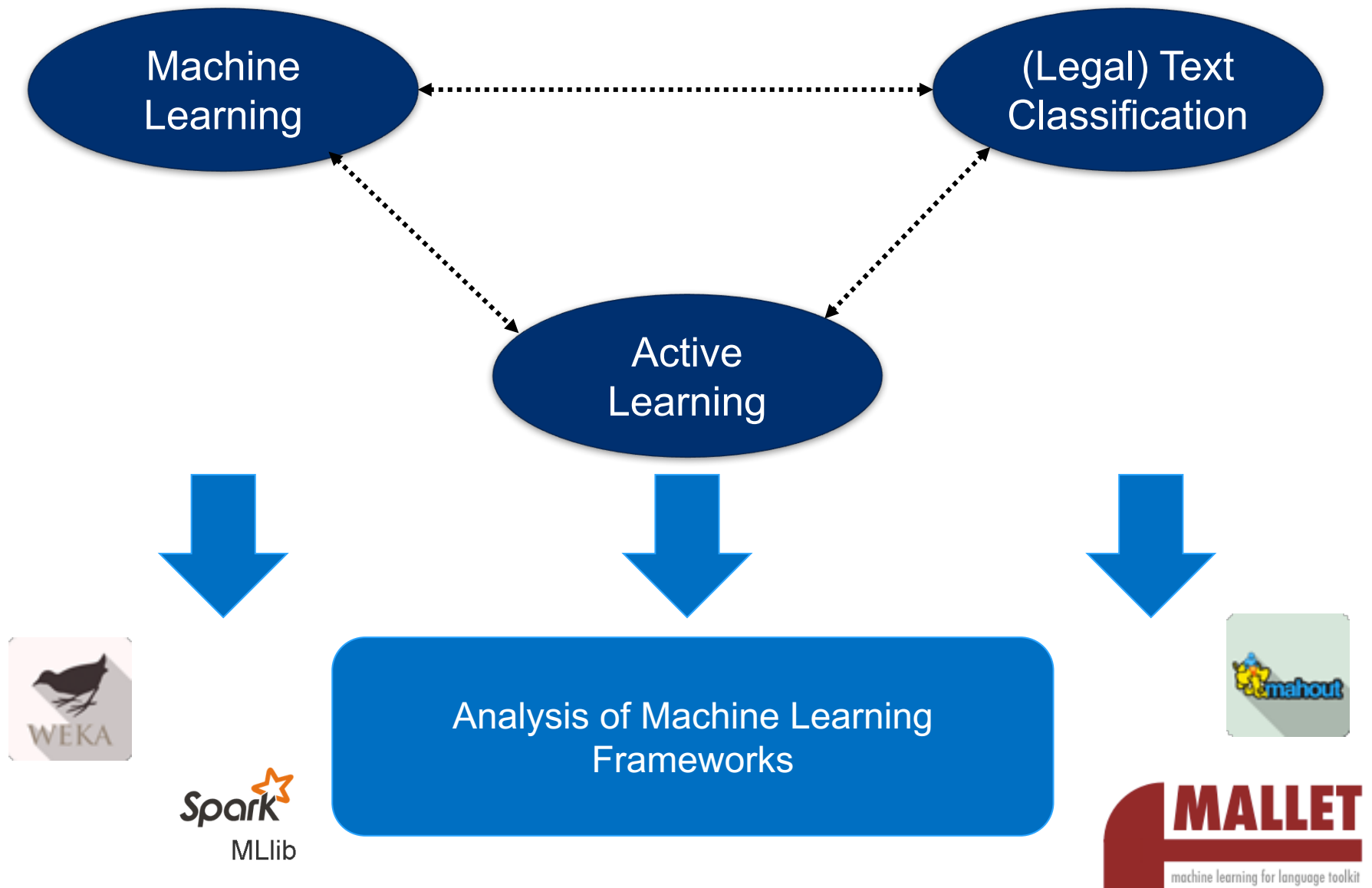


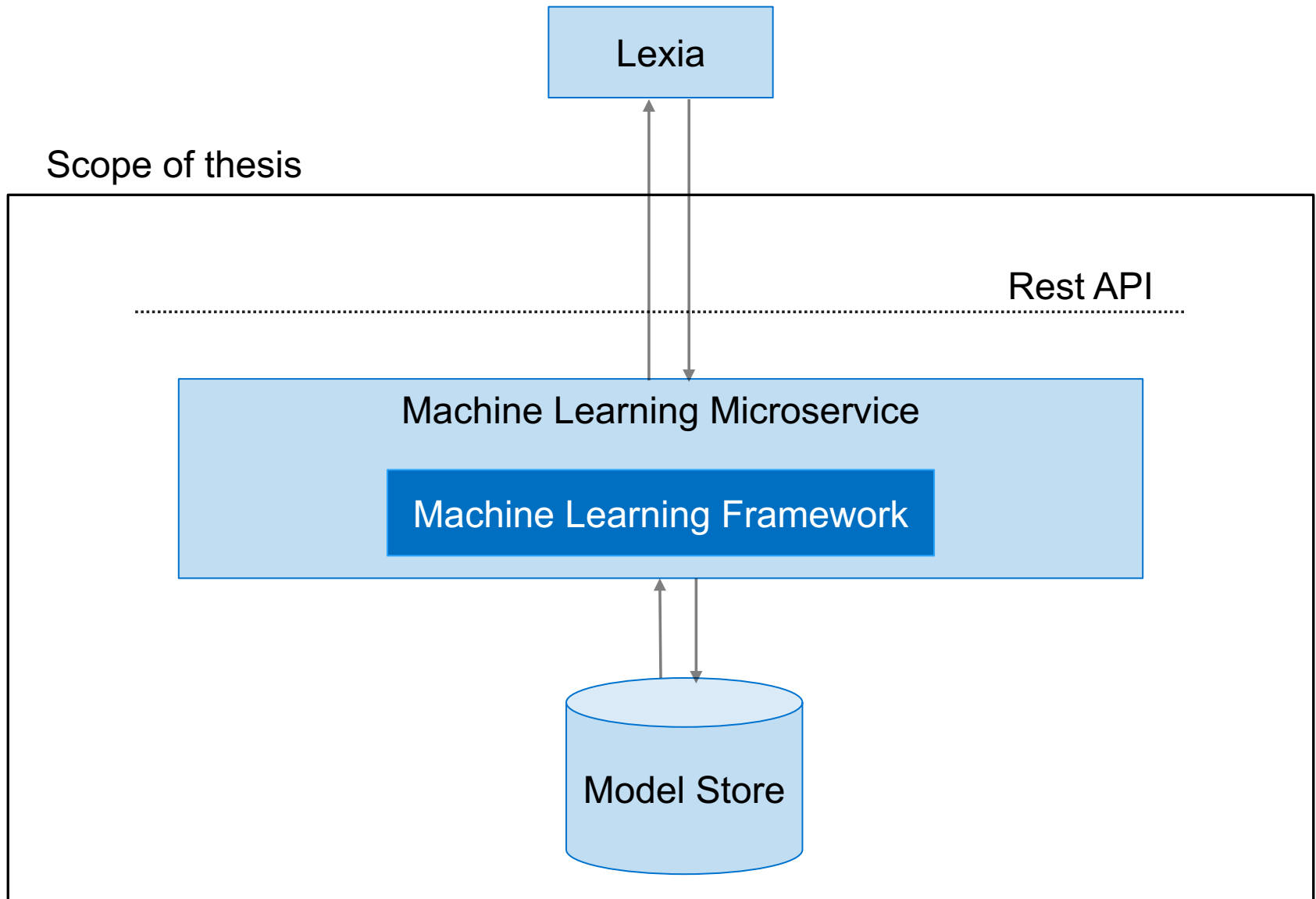
- ? **What are common concepts, strategies and technologies used in the context of text classification?**
- ? **How can (active) machine learning support the classification of legal documents and their content (sentences)?**
- ? **What does the concept and design of an active machine learning service look like?**
- ? **How well does the active machine learning service in the classification of legal documents and their content (sentences) perform?**

Outline



1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap

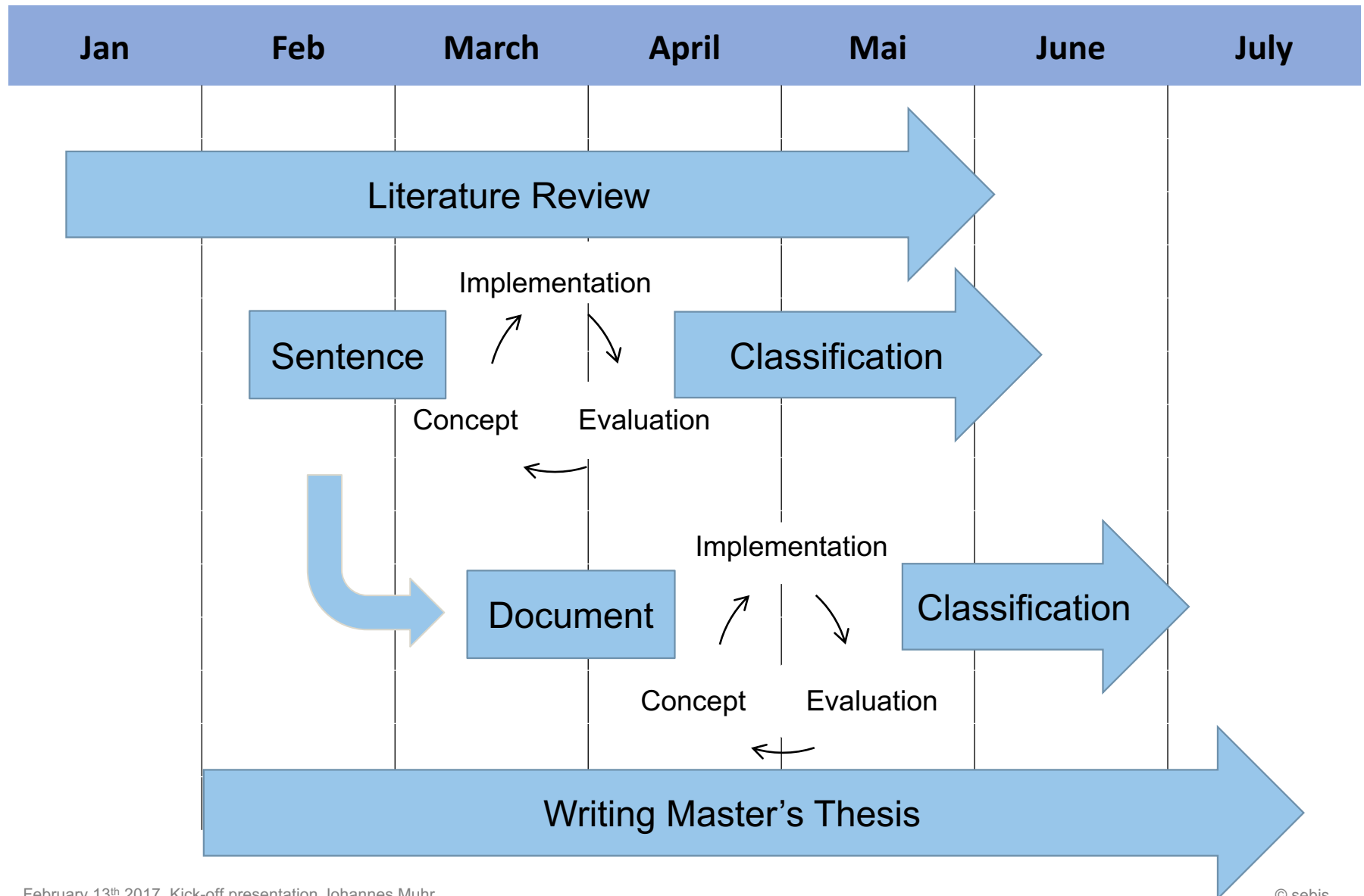




Outline



1. Motivation
2. Active Learning
3. Research Questions
4. Solution Approach
5. Roadmap





Johannes Muhr

Advisor: Bernhard Waltl

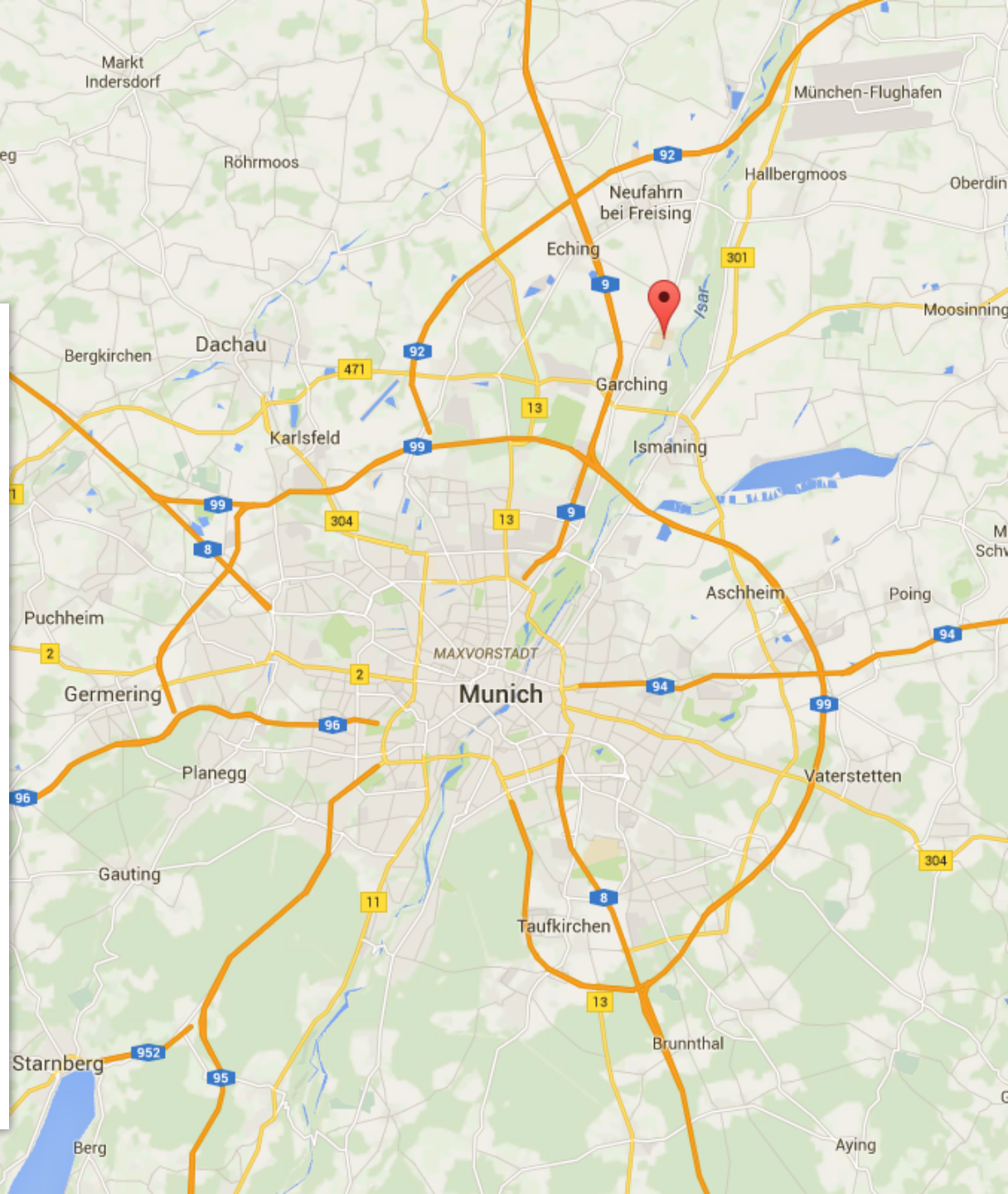
Technische Universität München
Faculty of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17132

Fax +49.89.289.17136

matthes@in.tum.de
www.matthes.in.tum.de



- Busse, D. (2000). Textsorten des Bereichs Rechtswesen und Justiz. In G. Antos, K. Brinker, W. Heineman, & S. F. Sager (Eds.), *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. (Handbücher zur Sprach- und Kommunikationswissenschaft)* (pp. 658-675). Berlin/New York: de Gruyter
- Cardellino, C., Villata, S., Alemany, L. A., & Cabrio, E. (2015). *Information Extraction with Active Learning: A Case Study in Legal Text*. Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics.
- Gruner, R. H. (2008). *Anatomy of a Lawsuit - A Client's Analysis and Discussion of a Multi-Million Dollar Federal Lawsuit*. Retrieved from <http://www.gruner.com/writings/AnatomyLawsuit.pdf>
- Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1), 24. doi:10.1186/s40537-015-0032-1
- Novak, B., Mladenič, D., & Grobelnik, M. (2006). Text Classification with Active Learning. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, & W. Gaul (Eds.), *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9–11, 2005* (pp. 398-405). Berlin, Heidelberg: Springer Berlin Heidelberg.

- Šavelka, J., Trivedi, G., & Ashley, K. D. (2015). Applying an Interactive Machine Learning Approach to Statutory Analysis.
- Segal, R., Markowitz, T., & Arnold, W. (2006). *Fast Uncertainty Sampling for Labeling Large E-mail Corpora*. Paper presented at the CEAS.
- Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66), 11.
- Sunkle, S., Kholkar, D., & Kulkarni, V. (2016, 5-9 Sept. 2016). *Informed Active Learning to Aid Domain Experts in Modeling Compliance*. Paper presented at the 2016 IEEE 20th International Enterprise Distributed Object Computing Conference (EDOC).
- Tong, S. (2001). *Active learning: theory and applications*. Citeseer.
- Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(1), 45-66

- **Use of online Platforms like**
 - Google Scholar,
 - Web of Science,
 - Institute of Electrical and Electronics Engineers (IEEE),
 - or Online Public Access Catalogue (OPAC) and Google Books
- **Backwards Search**