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

Johannes Muhr, July 10th 2017, Munich

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

Outline

1. Motivation
2. Research Questions
3. Research Approach & Objectives
4. Knowledge Base
5. LexML
6. Evaluation

Motivation I

- Processes of legal experts (scientists and lawyers) are
  - time-intensive
  - knowledge-intensive
  - data-intensive

- Legal Documents
  - are growing strongly in recent years
  - are changing over time
  - are becoming more complex

→ Therefore, searching for relevant information is
  - expensive [1]
  - difficult [2]

[1] Roitblat, H. L., et al.
[2] Paul, G.L. and J.R. Baron.

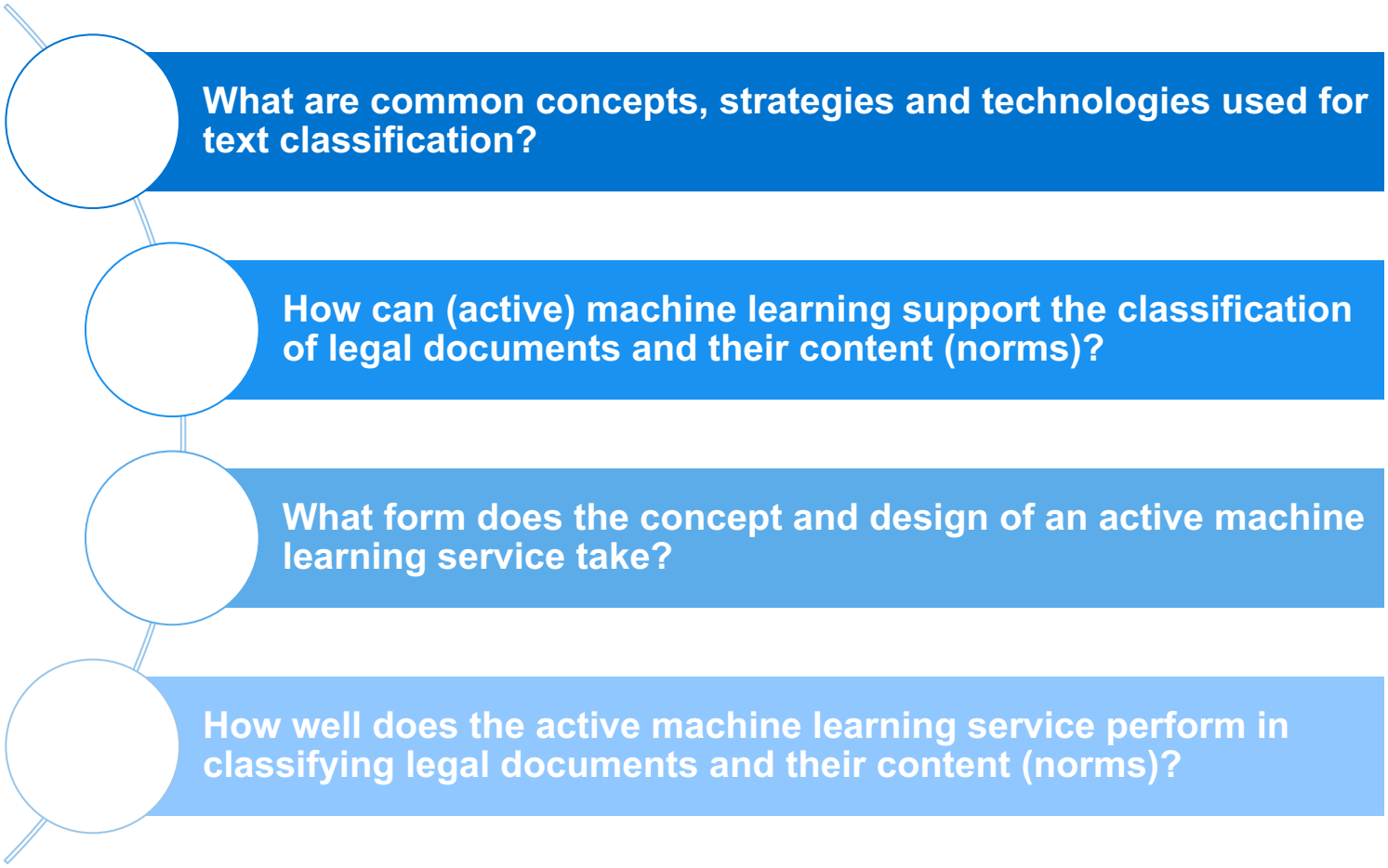
Motivation II

"There seems to be little question that machine learning will be applied to the legal sector. It is likely to fundamentally change the legal landscape." [3]

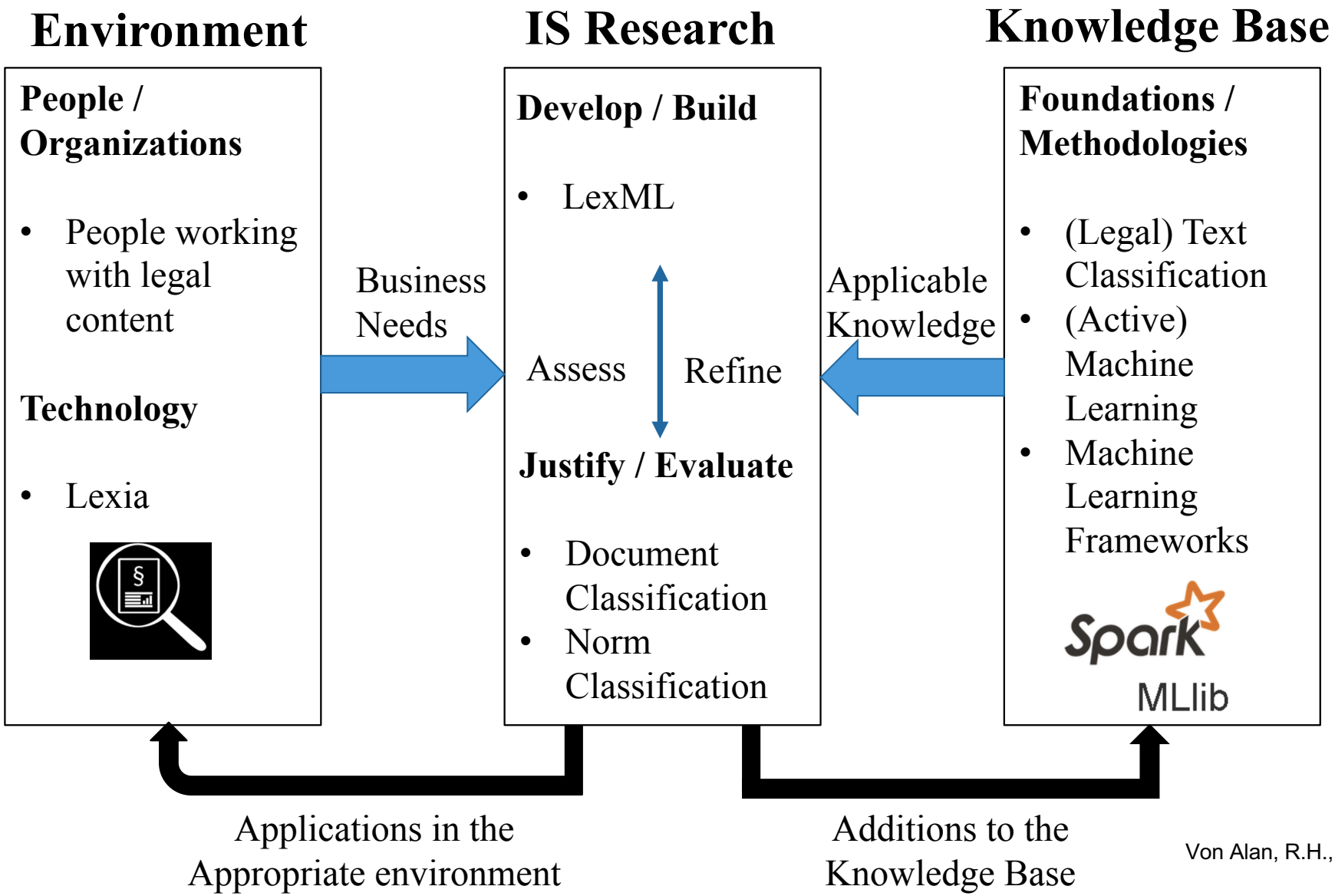
- Legal Data Science in incorporating **machine learning** is gaining more and more attention, because
  - process time and memory space are cheap
  - algorithms can process textual data fast and accurately
  - suitable frameworks are available

[3] Liu, B.

Research Questions



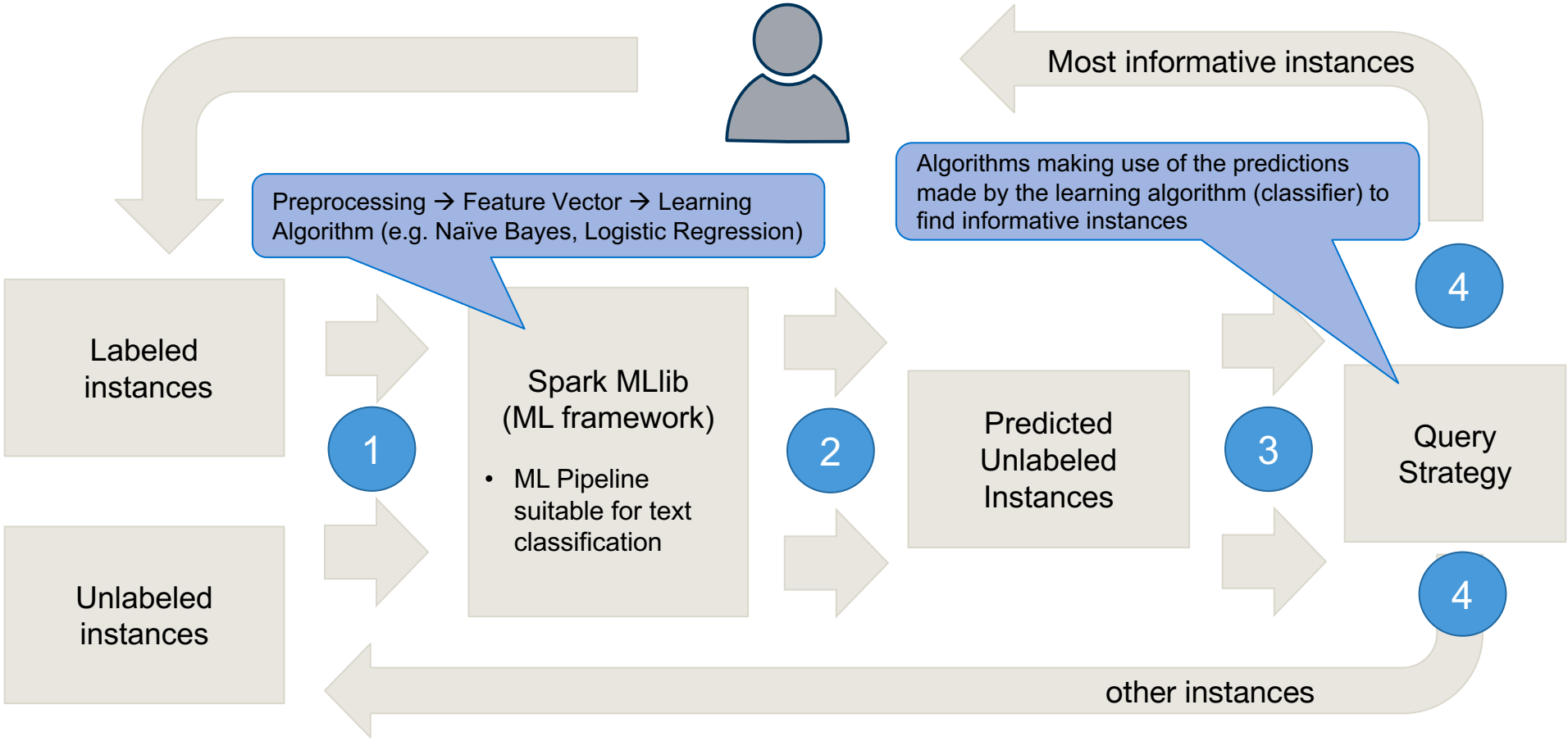
Research Approach & Objective



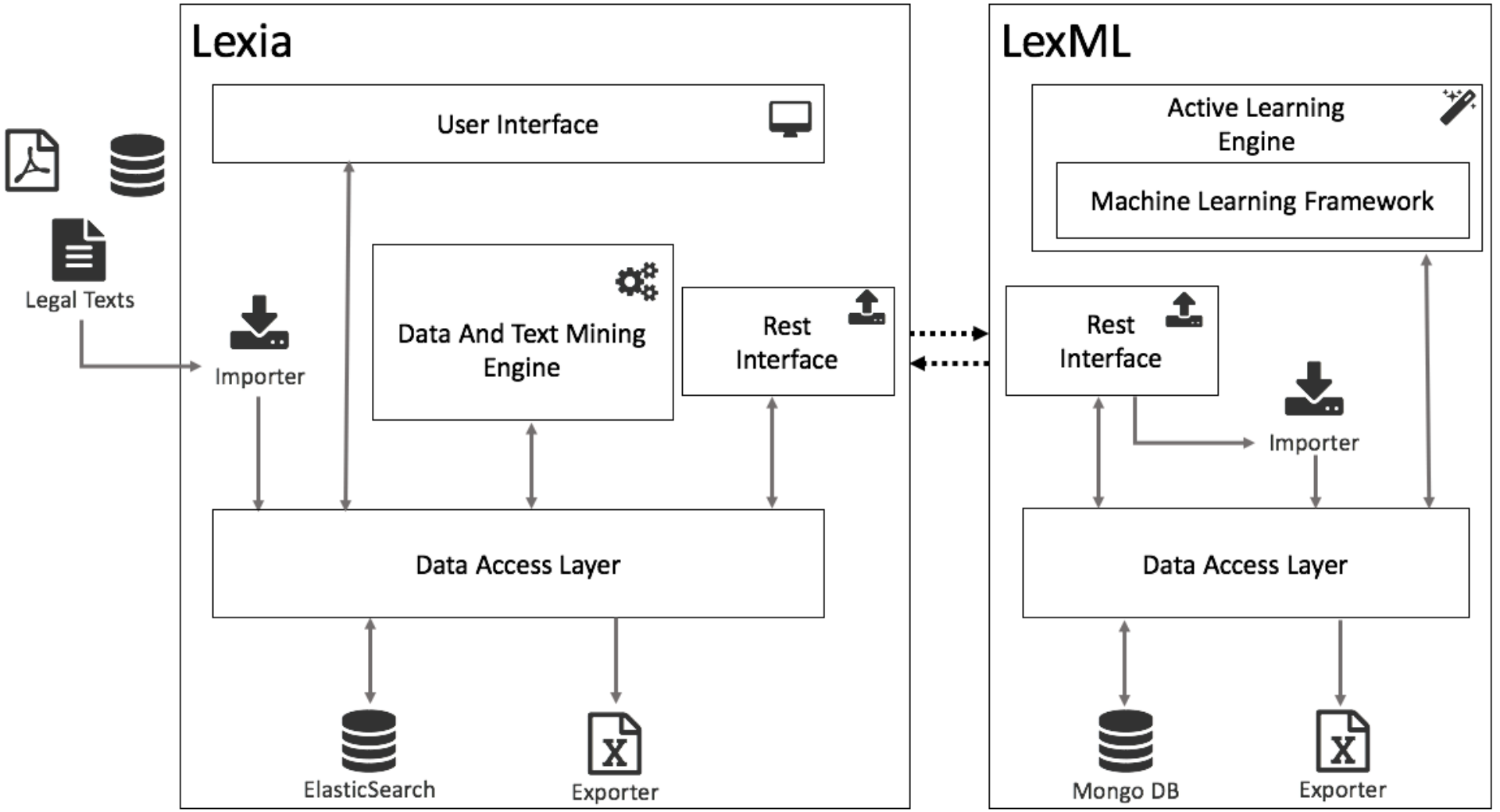
Von Alan, R.H., et al.

Knowledge Base

Iterative and **interactive** machine learning concept utilizing the strategy that the **learning algorithm** can **select the data** from which it learns.



Environment and IS Research



LexML – Basics

- Independent service using Apache Spark MLlib as machine learning framework
- Implementing the logic for active learning to perform multiclass (legal) text classification
- Accessible via Rest API

- Configurable via Lexia
  - **Label**
  - **Classifier**
    - Naïve Bayes, Logistic Regression, Multilayer Perceptron
  - **Data**
  - **Query strategy**

- Export of evaluation results in a xlsx-file

Evaluation – Experimental Setup

- **Two experiments**
  - Document classification
  - Norm classification
  - **Simulating** active learning using the available labeled dataset

- Evaluation of **twelve** different **machine learning combinations**
  - Nine active learning strategies
  - Three passive learning strategies

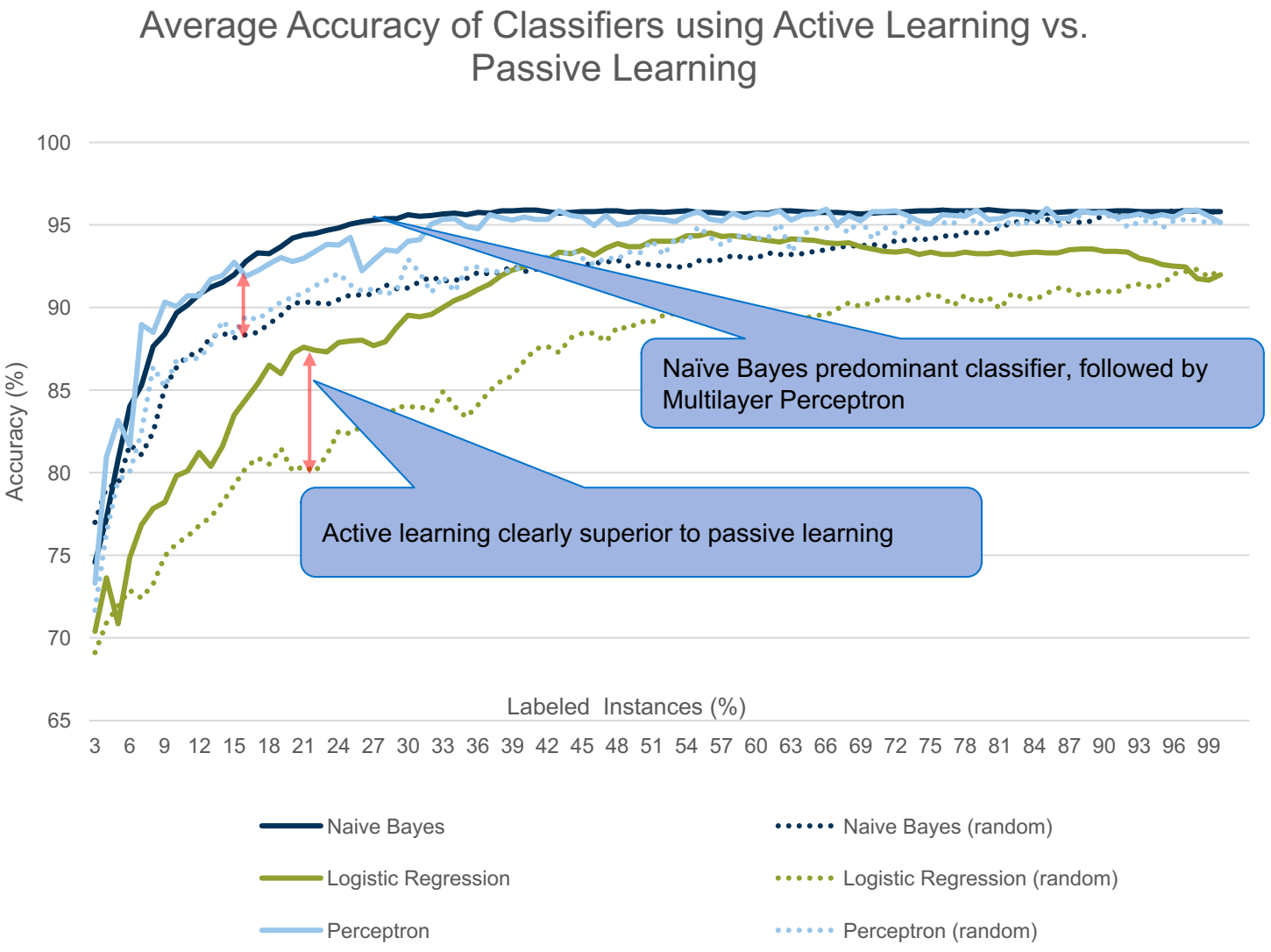
- **Evaluation process**
  - After each learning round, the resulting model was applied to the **test data**
  - **Export** of common evaluation measures (F_1, accuracy, precision & recall)
  - Use the **average result** of five rounds per combination as reference value

Evaluation – Document Classification I

- Classifying **1000 random documents** from the datev corpus into one of three classes using a subset of words of the document
- Random, but even split in training- and test dataset

Class	Sub-classes	Support
Law (“Gesetz”)	Gesetzestext (Gesamttext), Richtlinie (Gesamttext), EU-Richtlinie (Gesamttext),	12.8%
Judgment (“Urteil”)	Urteil, Beschluss, Gerichtsbescheid	65.5%
Generic legal document (“Sonstige Dokumente”)	Aufsatz, Anmerkung, Verfügung, Verfügung (koordinierter Ländererlass), Kurzbeitrag, Erlass, Erlass (koordinierter Ländererlass), Schreiben, Schreiben (koordinierter Ländererlass), Übersicht, Mitteilung	21.6%

Evaluation – Document Classification II

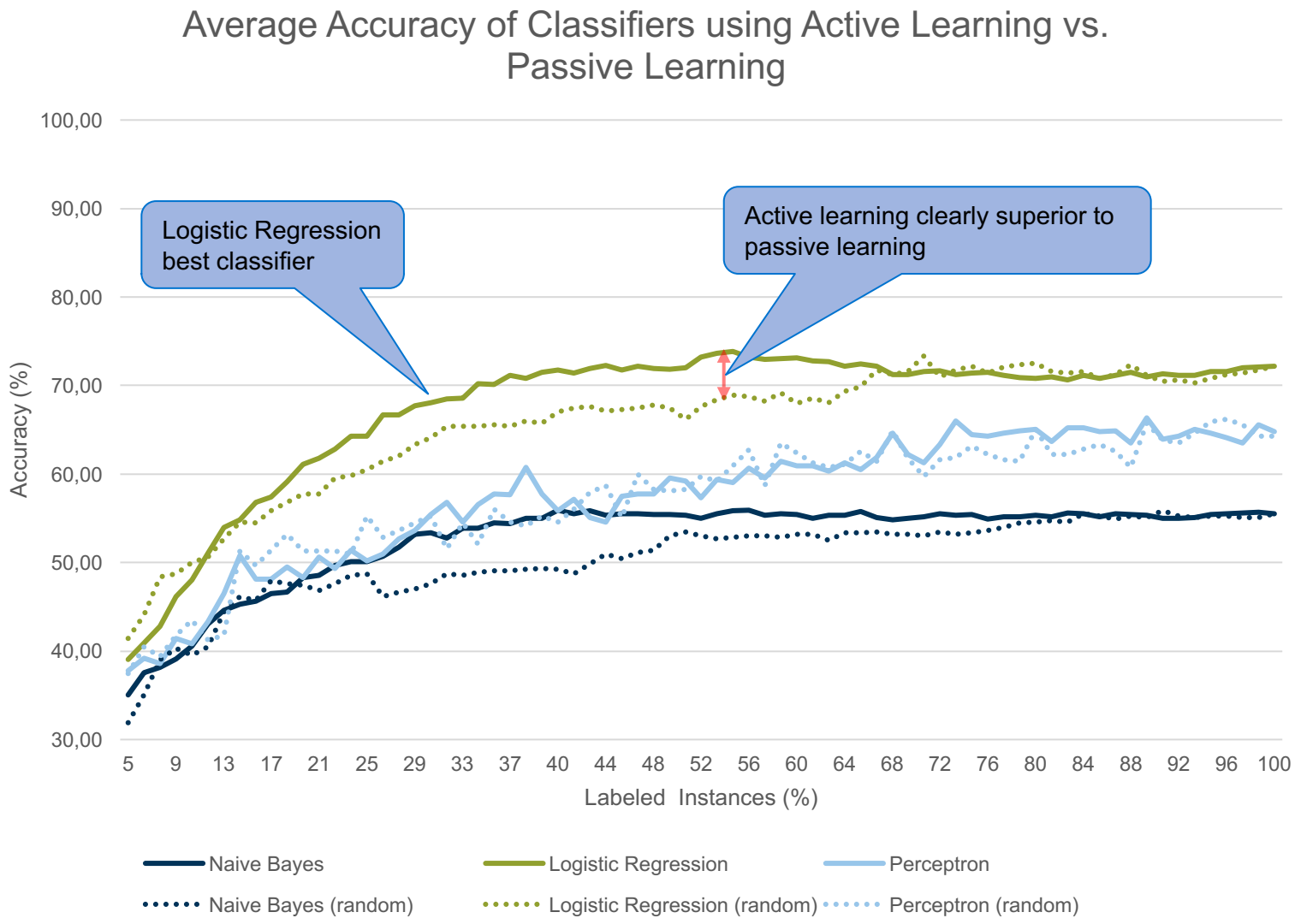


Evaluation – Norm Classification I

- Classifying **504 sentences (norms)** of the law of tenancy section of the German civil code (BGB) as one of eight possible classes
- Random, but even split in training- and test dataset

Class	Sub-classes	Support
Recht	Gebot (positiv/negativ/Soll-Pflicht (H)), Verbot (/Duldung (U))	25.0%
Pflicht	Erlaubnis (/Erlaubnis beschränkt), Ermächtigung	21.6%
Einwendung (Einw rh)	Unwirksamkeit (/Wirksamkeit beschränkt (sachlich)/(persönlich)), Unzulässigkeit (/Zulässigkeit beschränkt), Ausschluss (/Ausschluss beschränkt), Nichtigkeit	18.3%
Rechtsfolge (RF vAw)	Rechtzuweisung (/Rechtsübergang), Pflichtzuweisung (/Pflichterweiterung), Rechtseigenschaft, Freistellung	9.9%
Verfahren	Form, Frist (/Fälligkeit), Maßstab (inkl. Berücksichtigung), Entscheidungskompetenz	9.7%
Verweisung	Direkt § /direkt Vorschriftenkomplex, Analog § /analog Vorschriftenkomplex, Negativ	9.1%
Fortführungsnorm	Ausnahme, Erweiterung, Einschränkung, Gleichstellung	3.8%
Definition	Direkt, Indirekt, Negativ	2.6%

Evaluation – Norm Classification II



Summary

Conclusions

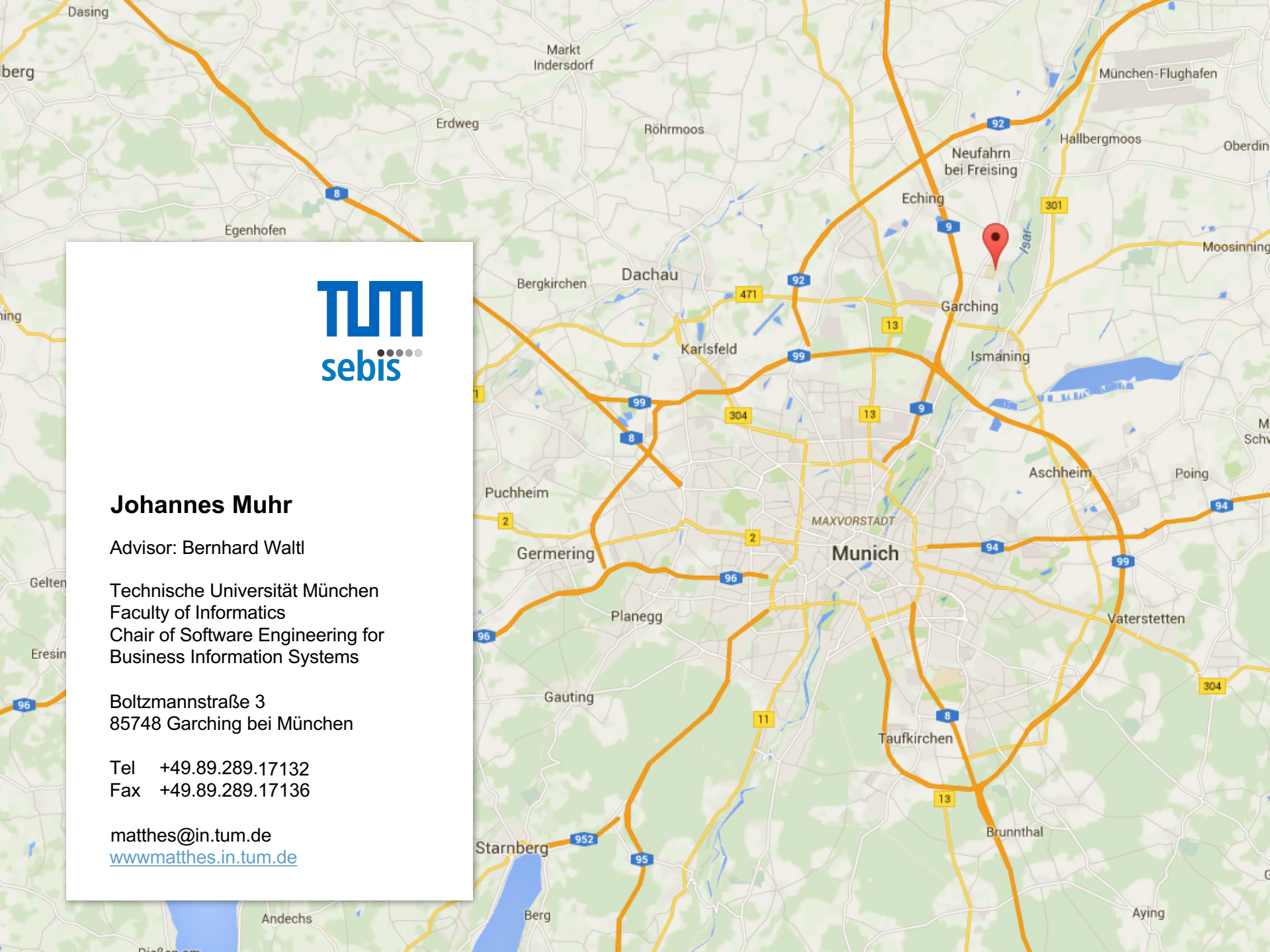
- Active Learning is a promising approach to conduct legal text classification
- Better results using less instances
- The choice of the parameters is a very complex issue

Limitations

- Evaluation with “only” one default setting
- Reusability of trained model
- Black-box classifier

Future Work

- Conduct of additional evaluations rounds with (varying learning settings)
- Implementation of additional features
  - Query strategies
  - Learning algorithms
- Combination of (active) machine learning and rule-based learning



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
wwwmatthes.in.tum.de

Bibliography I

- 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.
- de Maat, E., K. Krabben, and R. Winkels. *Machine Learning versus Knowledge Based Classification of Legal Texts*. in *JURIX*. 2010
- Francesconi, E. and A. Passerini, *Automatic classification of provisions in legislative texts.* Artificial Intelligence and Law, 2007. **15**(1): p. 1-17.
- 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
- Liu, B., *Machine Learning and the Future of Law*, L.T. Blog, Editor. 2015.
- 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.
- Paul, G.L. and J.R. Baron, *Information inflation: Can the legal system adapt.* Rich. JL & Tech., 2006. **13**: p. 1.
- Ratner, A., *Leveraging Document Structure for Better Classification of Complex Legal Documents*

Bibliography II

- Roitblat, H.L., A. Kershaw, and P. Oot, *Document categorization in legal electronic discovery: computer classification vs. manual review.* Journal of the American Society for Information Science and Technology, 2010. **61**(1): p. 70-80.

- 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

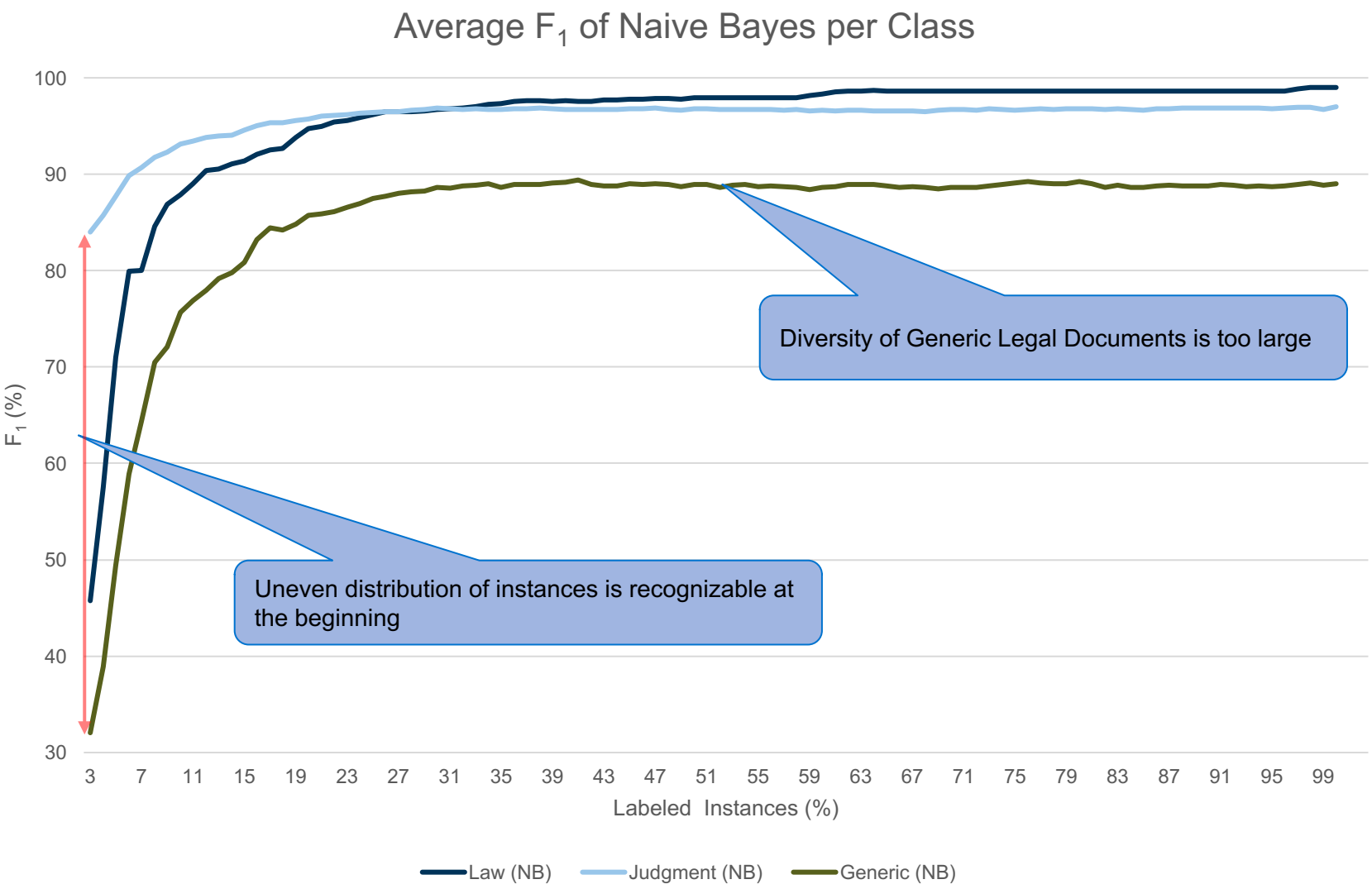
- Vern R. Walker, Ji Hae Han, Xiang Ni, Kaneyasu Yoseda (2017). Semantic Types for Computational Legal Reasoning: Propositional Connectives and Sentence Roles in the Veterans' Claims Dataset, Paper presented at the ICAIL

- Von Alan, R.H., et al., *Design science in information systems research.* MIS quarterly, 2004. **28**(1): p. 75-105.

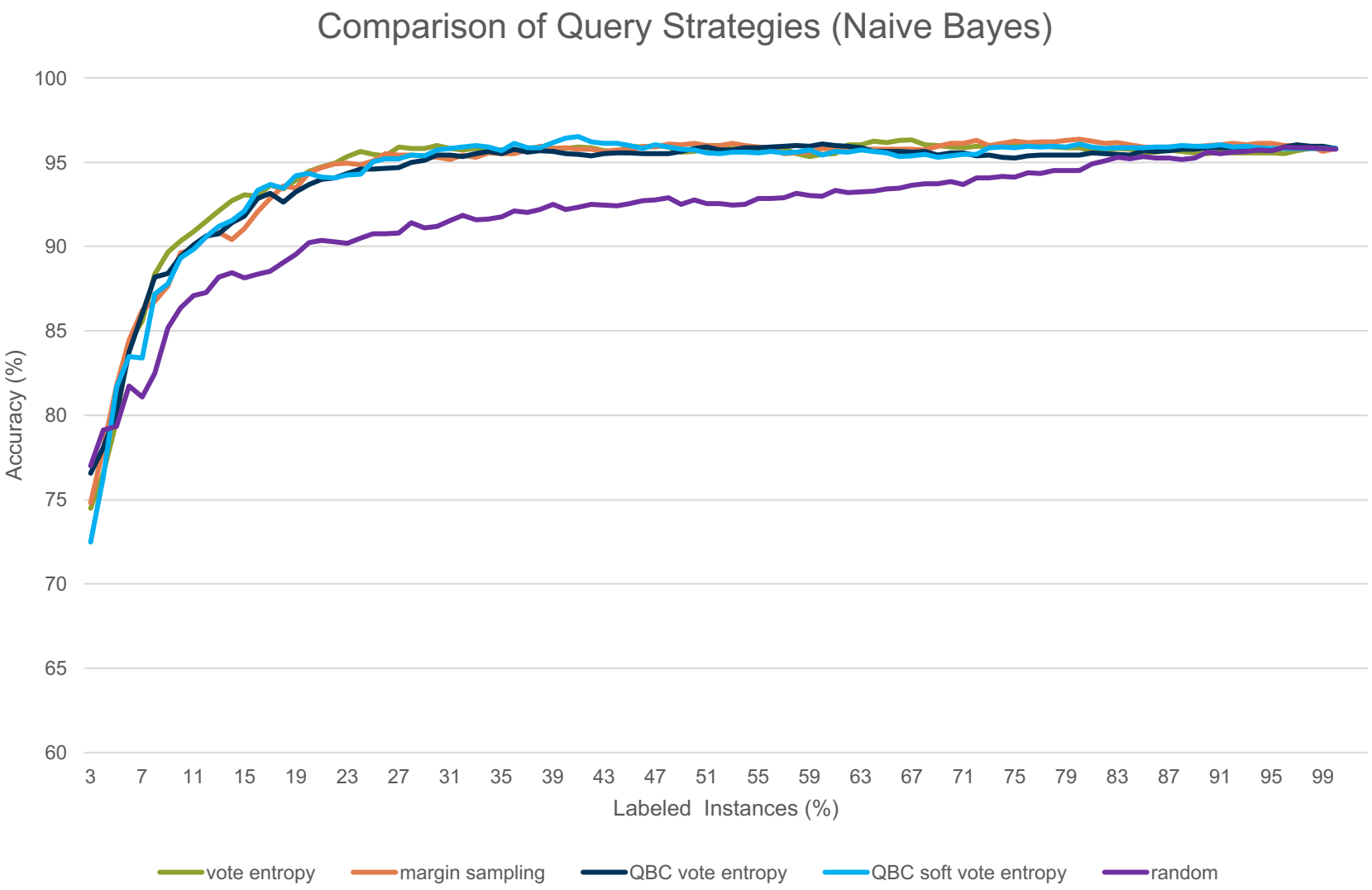
Backup – Related Work

- Ratner (2014) performed an experiment in which he classified **legal contract documents into categories**, such as “Arbitration Agreements” or “Manufacturing Contract”
- Šavelka, Trivedi and Ashley (2015) examined the relevance and applicability of the **individual statutory provisions**
- Roitblat, Kershaw and Oot (2010) compared **the classification accuracy** of computers relative to traditional human manual review
- De Maat, Krabben and Winkels (2010) conducted a **norm classification** experiment classifying sentences of the **dutch law**
- Walker, Han, Ni and Yoseda (2017) examined **semantic types** for **argument mining in legal texts** to provide a common basis for machine learning

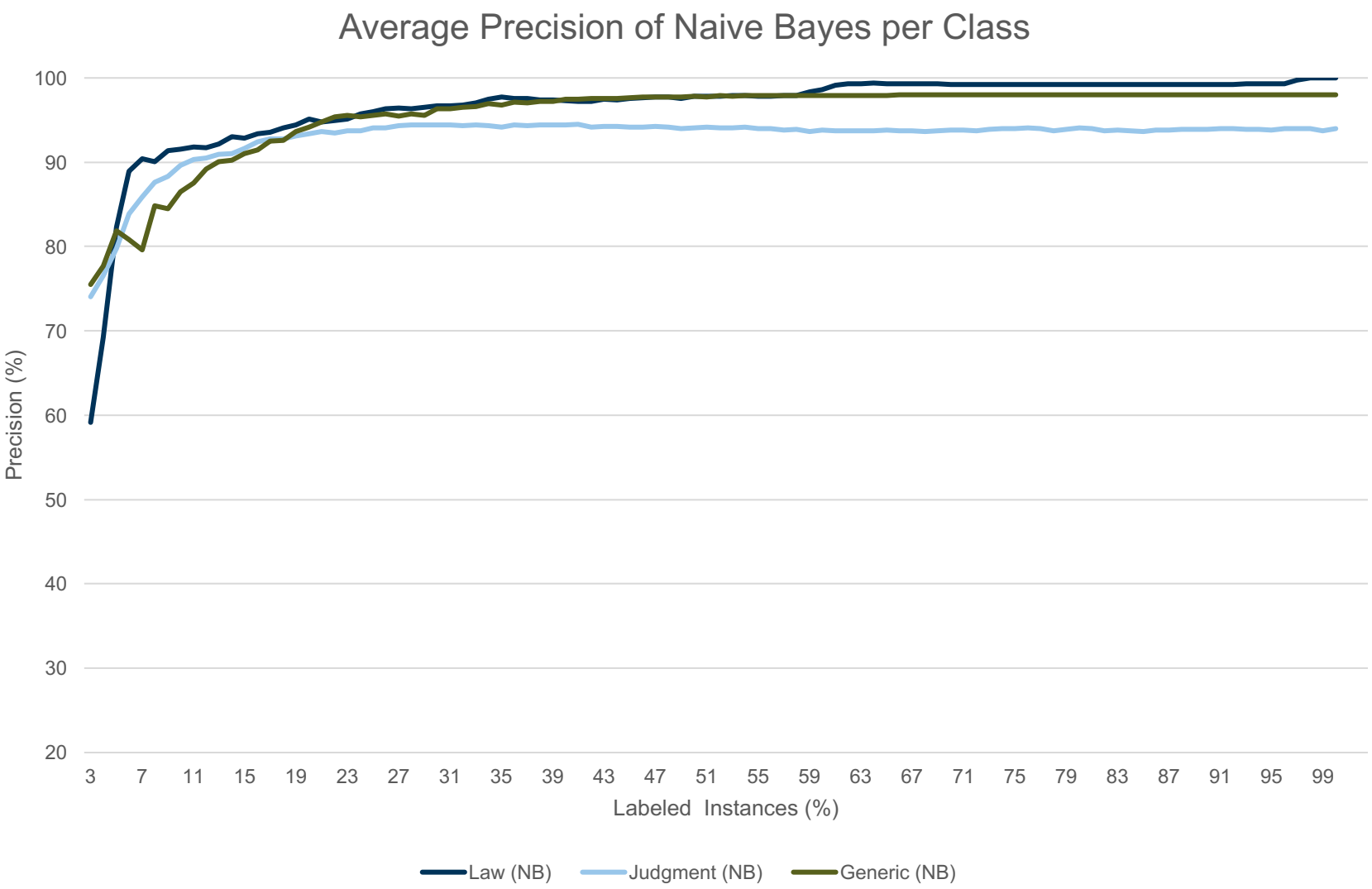
Backup– Document Classification I



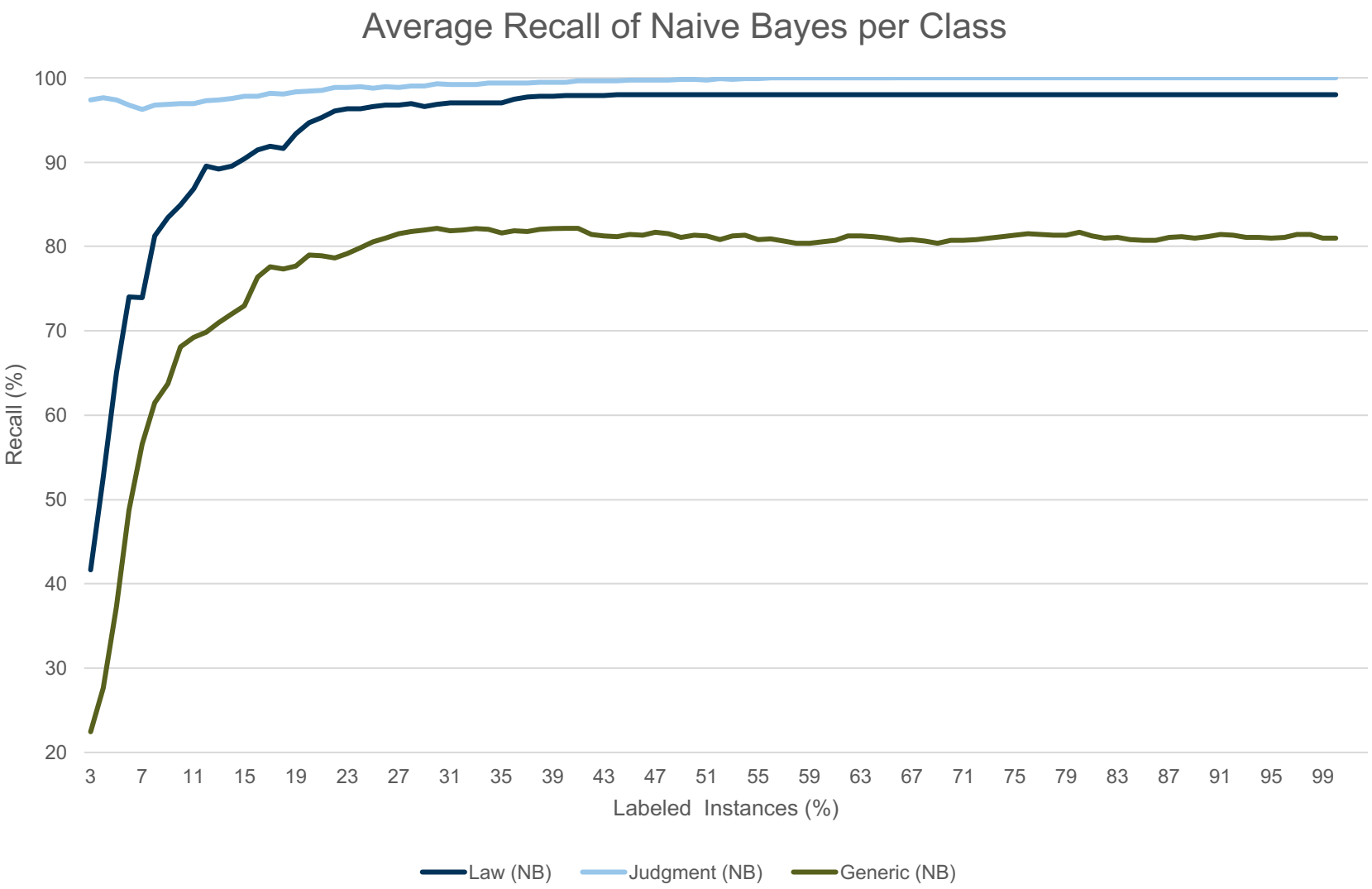
Backup – Document Classification II



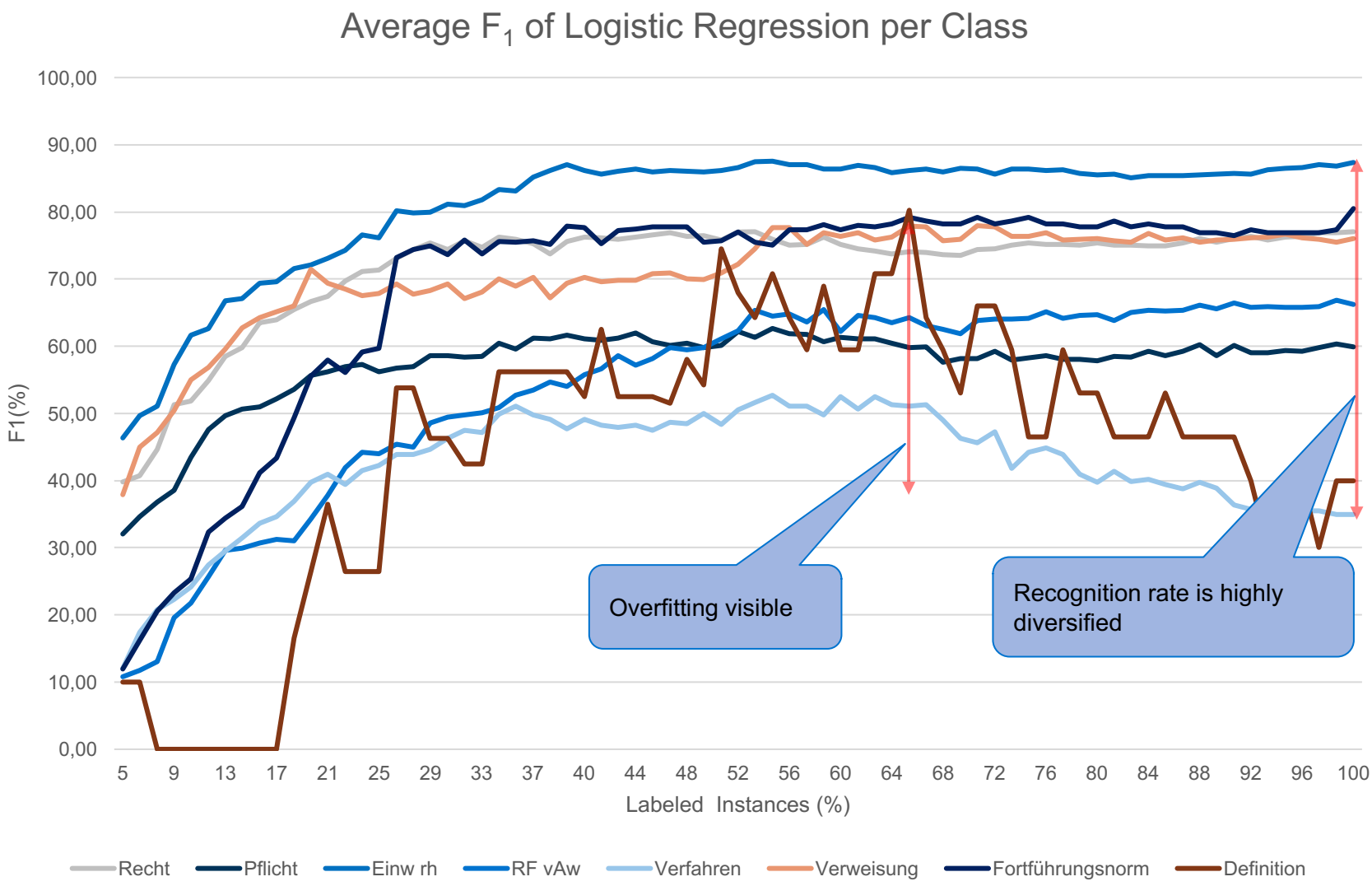
Backup – Document Classification III



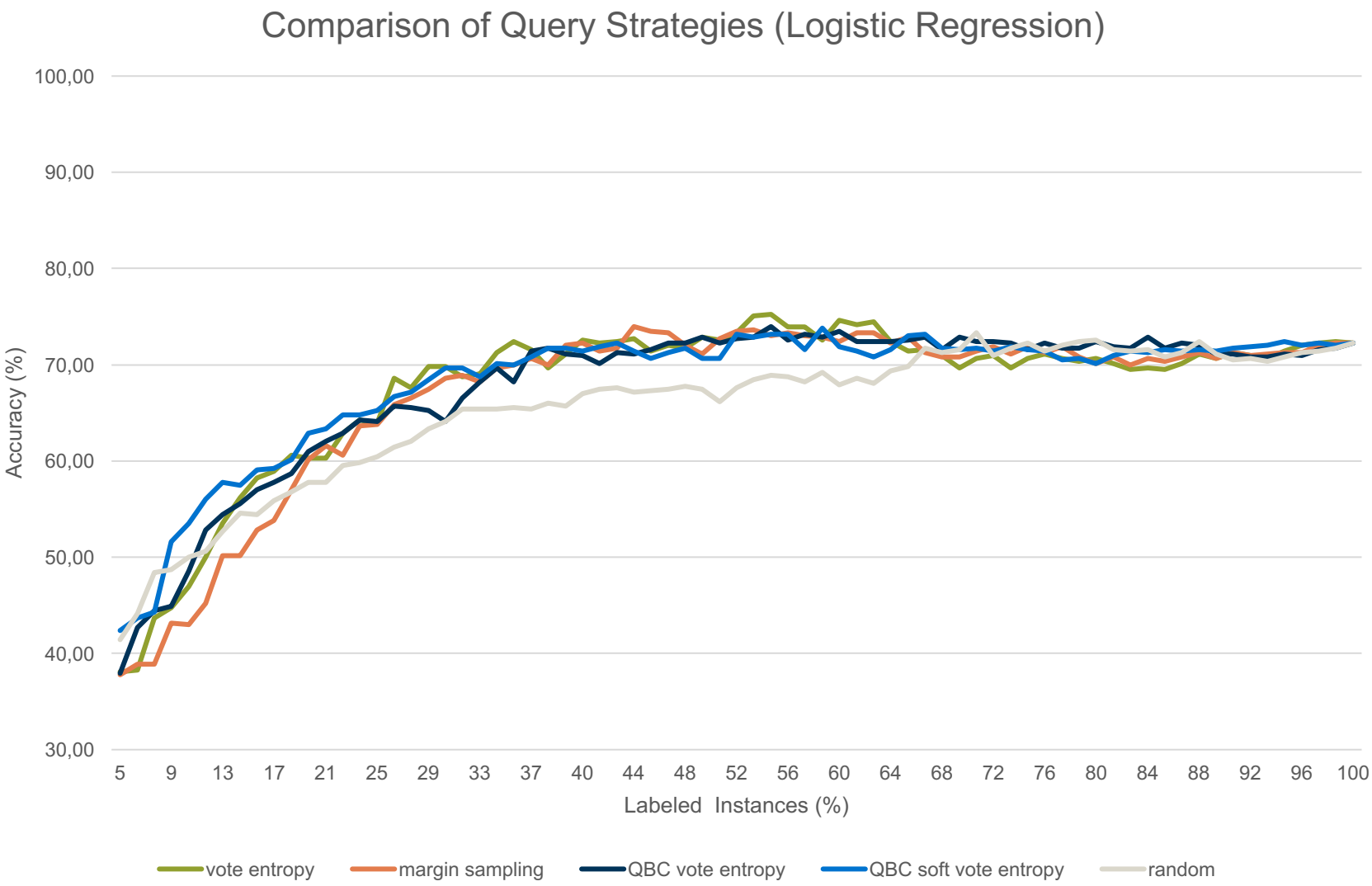
Backup – Document Classification IV



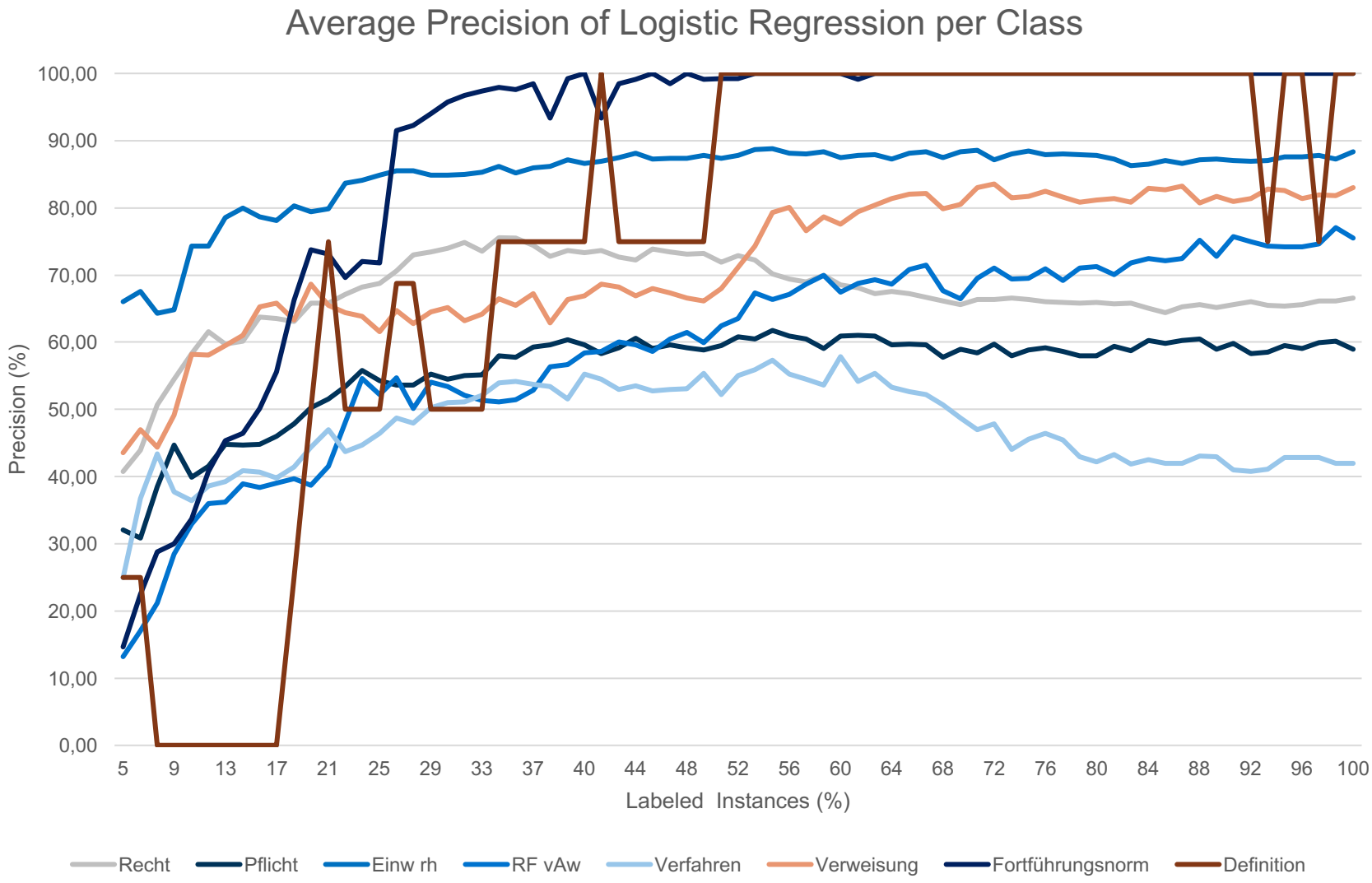
Backup – Norm Classification I



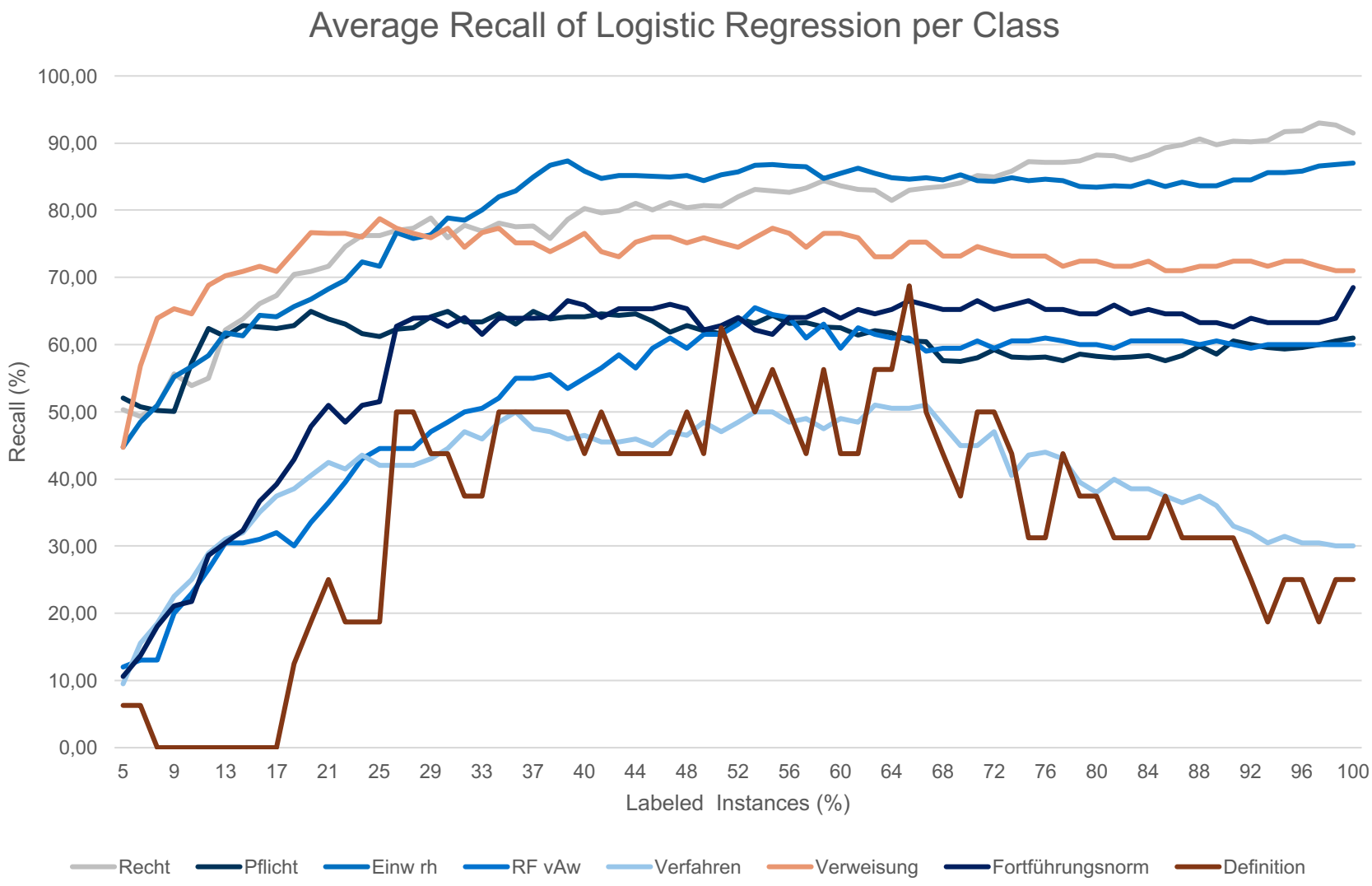
Backup – Norm Classification II



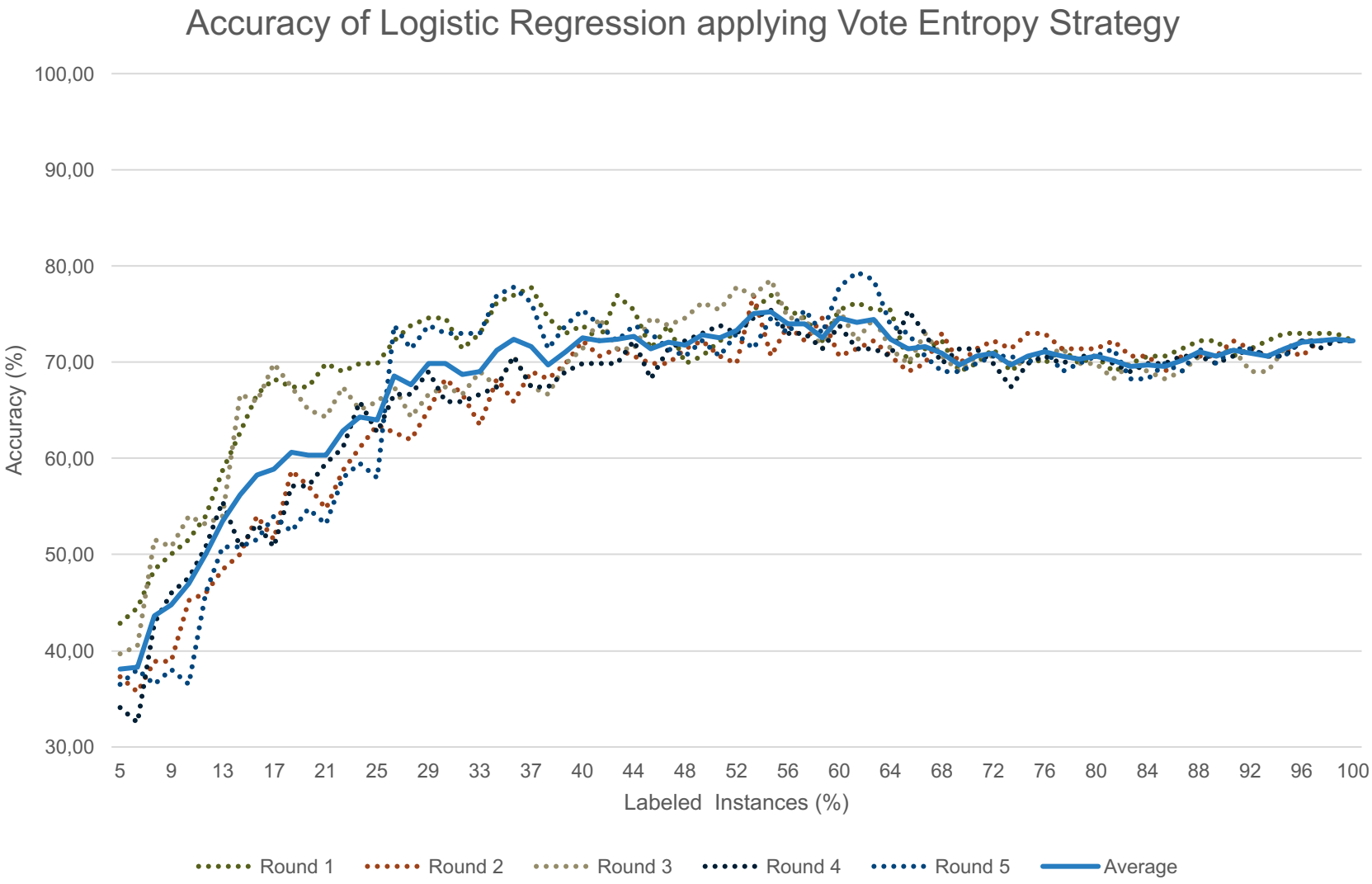
Backup – Norm Classification III



Backup – Norm Classification IV



Backup – Norm Classification V



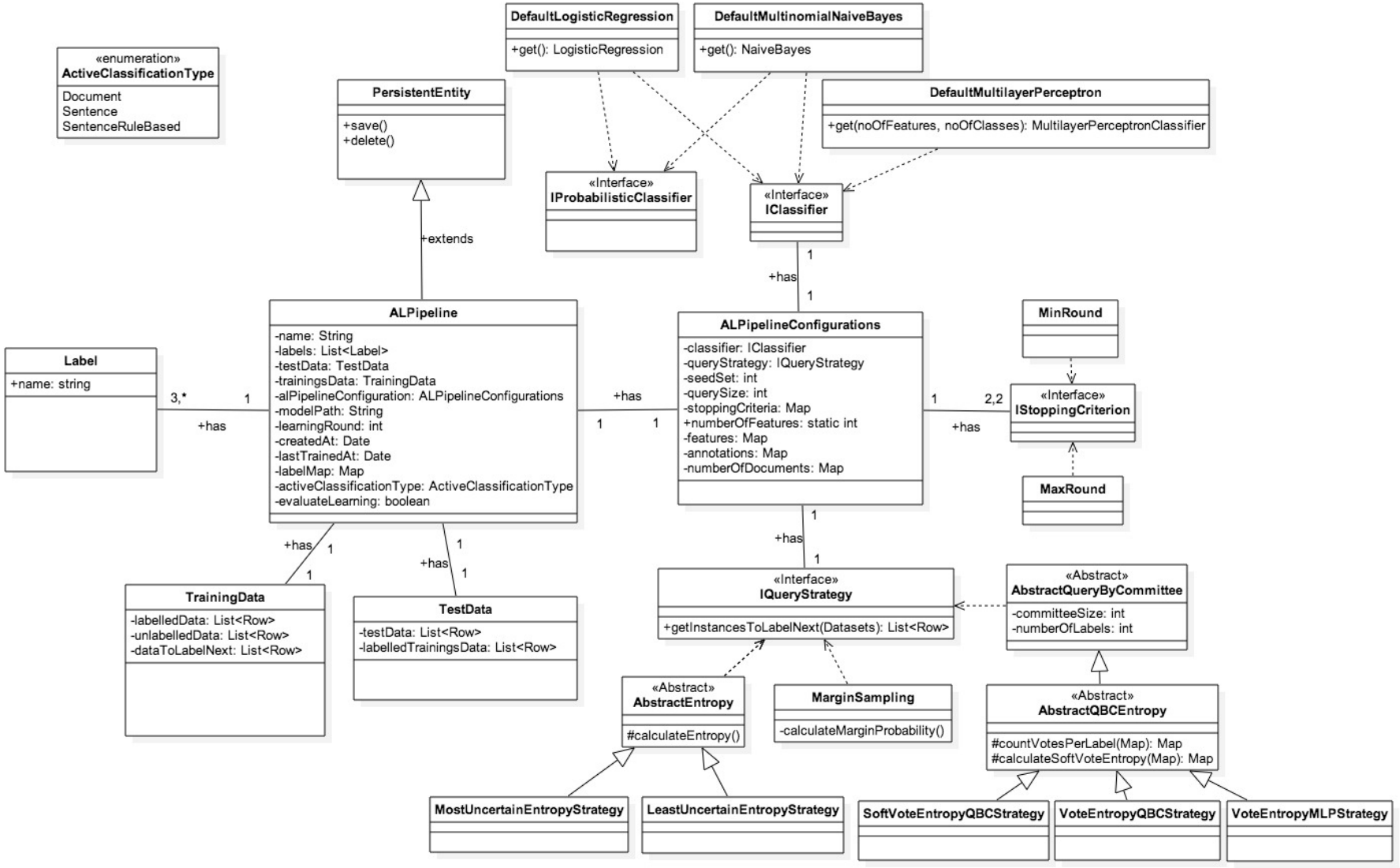
Backup – Evaluation Measures

- Confusion Matrix

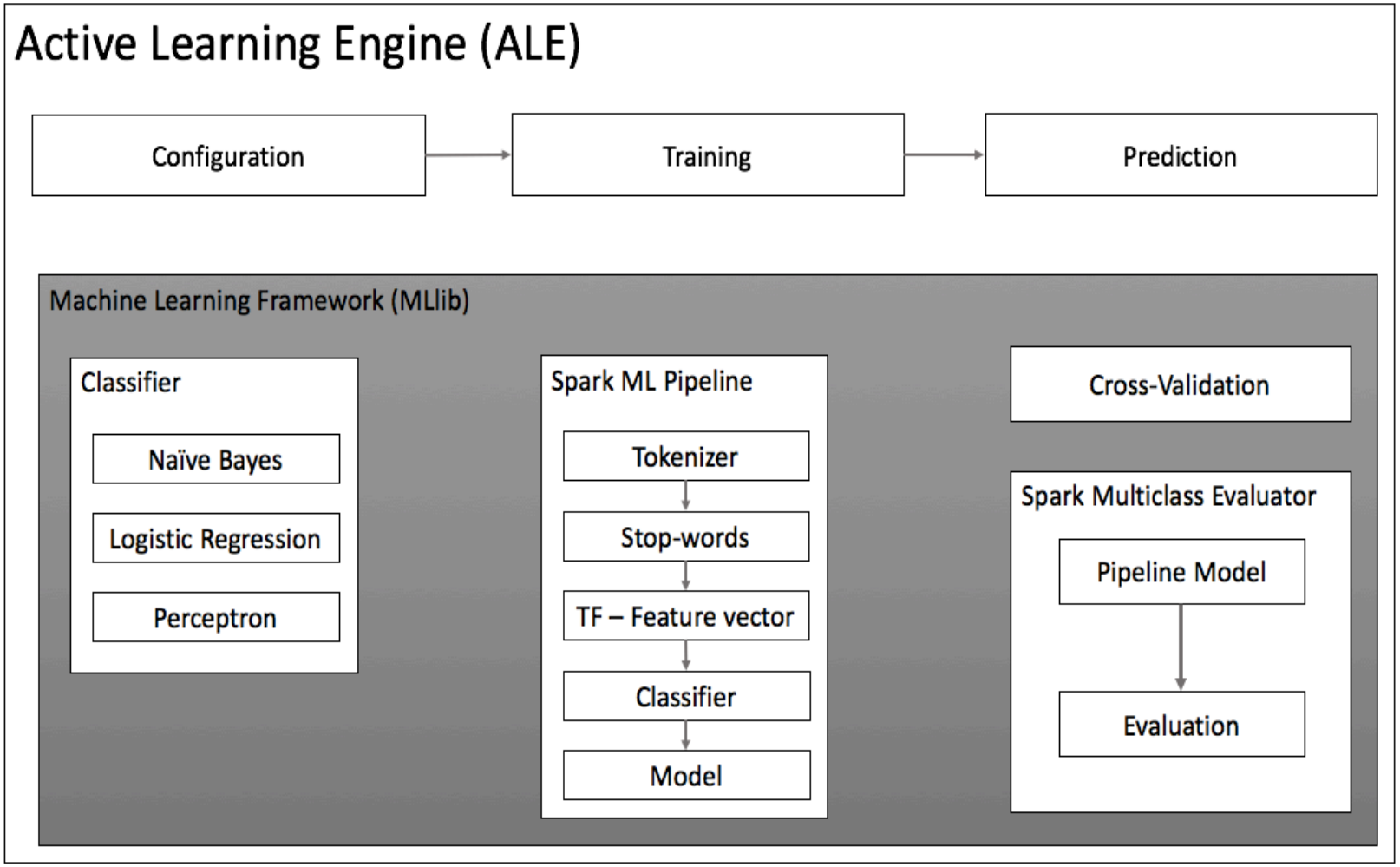
		True Class	
		A	B
Predicted class	A	TP	FP
	B	FN	TN

- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- F_1 (F-score) $= \frac{2*Precision*Recall}{Precision+Recall}$
- Accuracy $= \frac{TP+TN}{P+TN+FP+FN}$

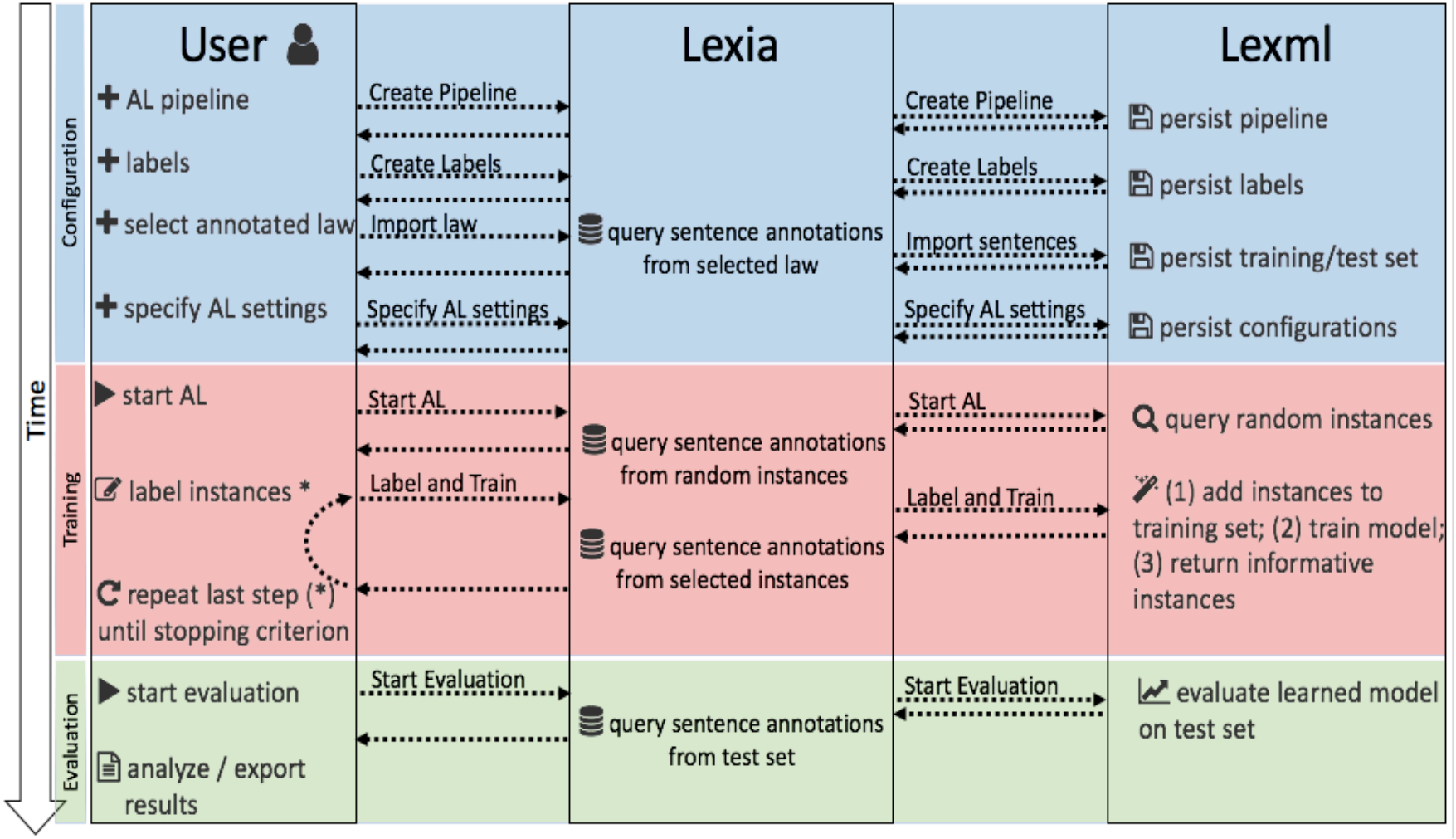
Backup – Data Model



Backup – Active Learning Engine



Backup – Workflow Norm Classification



Backup – Query Strategies I

Uncertainty Sampling

- Margin Sampling
  - $x_M^* = \underset{x}{argmax} [\, P\text{ø}\,(\hat{y}_2|x) - P\text{ø}\,(\hat{y}_1|x)]$
  - Ambiguous instances with small margins should help the model to discriminate between them.

- Vote Entropy
  - $x_H^* = \underset{x}{argmax} - \sum_y P\text{ø}\,(y|x) \log P\text{ø}\,(y|x)$
  - Approach based on the Shannon entropy, measuring the variable's average information content:

Backup – Query Strategies II

Query By Committee

- Vote Entropy
  - $x^*_{VE} = \underset{x}{argmax} - \sum_y \frac{V(y,x)}{|C|} \, log \frac{V(y,x)}{|C|}$
  - y involves all possible labeling, V(y,x) is the number of committees that "voted" for label y for instance x, and |C| is the committee size.

- Soft Vote Entropy
  - $x^*_{SVE} = \underset{x}{argmax} - P_C(y|x) \, log P_C(y|x)$
  - Taking the confidence of the decision into account: $P_C(y|x) = \frac{1}{|C|} \sum_{c \in C} P_c \, (y|x)$ is the average ("consensus") probability that y is correctly labeled according to the committee.

Backup – Classifier I

Naïve Bayes

- Feature d
- Class c
- P(d) constant (each feature has the same probability of being in the dataset)
- Assumes strong independence between the single features

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

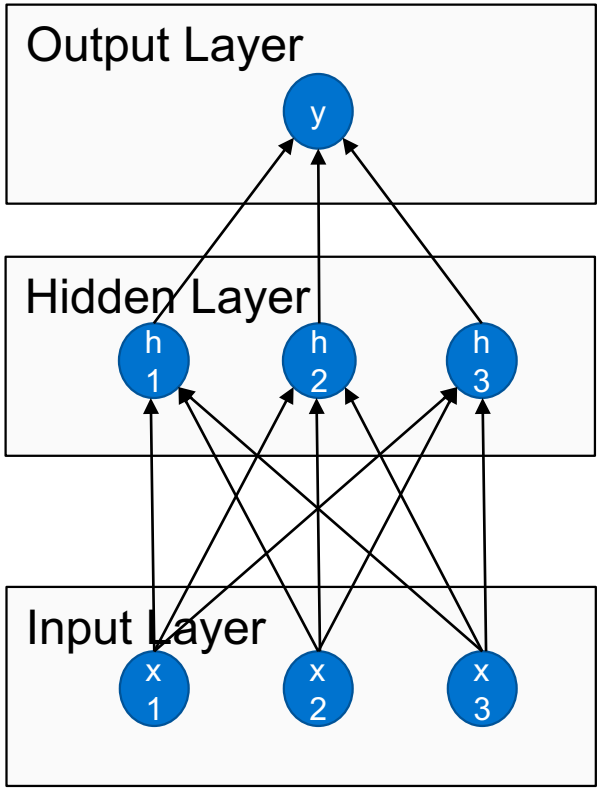
Logistic Regression

- Uses the characteristic that the probability of observing the true label (logit transformation) can be modelled as a linear combination of the attributes

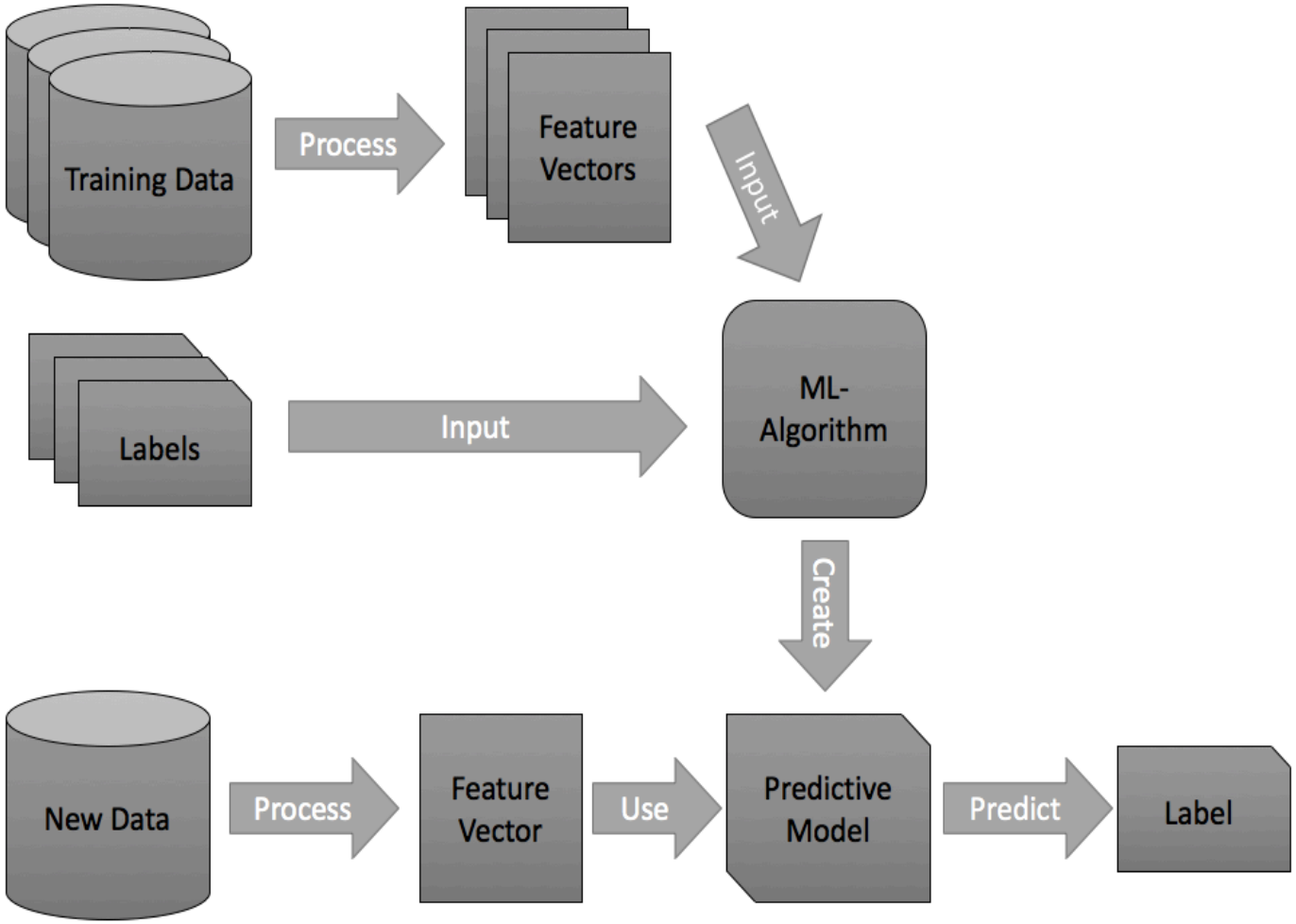
Backup – Classifier II

Multilayer Perceptron

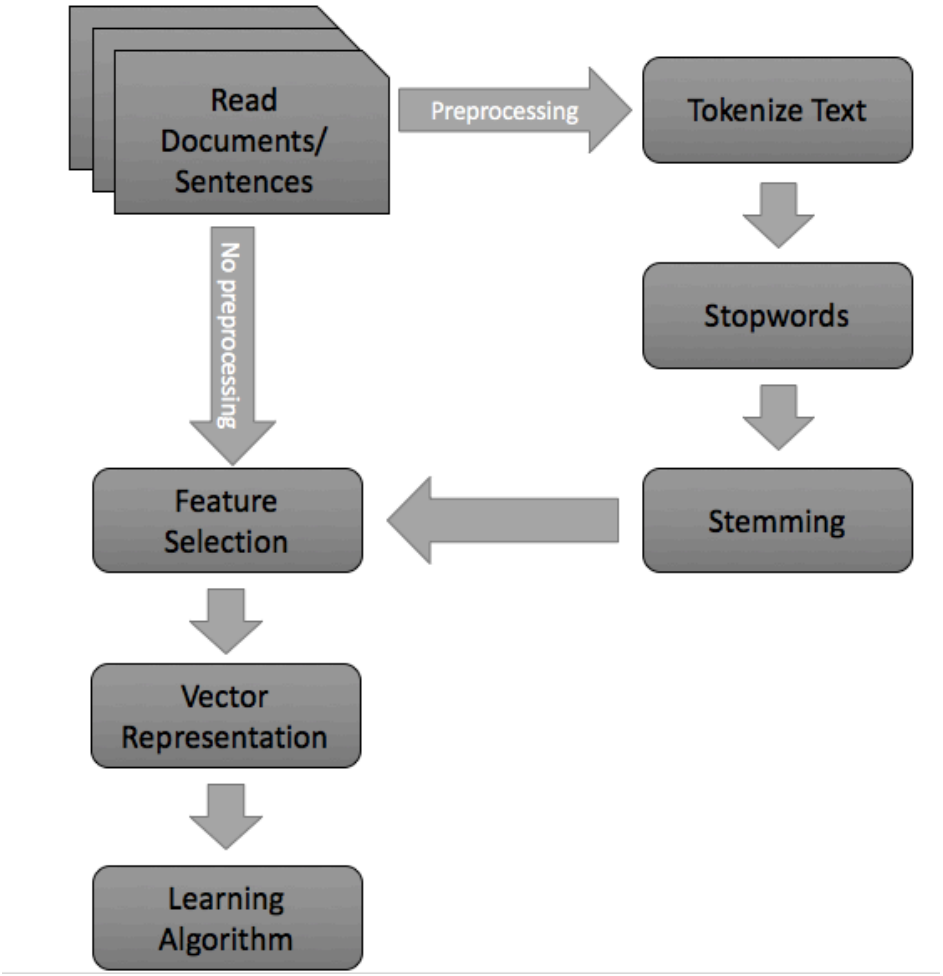
- Several layers and units (nodes) in this layer
- Weights are assigned to each edge
- **Back propagation** is used as learning algorithm
- Forward phase, weights are fixed and training instances are propagated through the network
- Backward phase, the desired output is compared to the actual output at each output neuron to calculate the error rate
- For each neuron in the network, all weights are adjusted so that the actual output better approximates the desired output



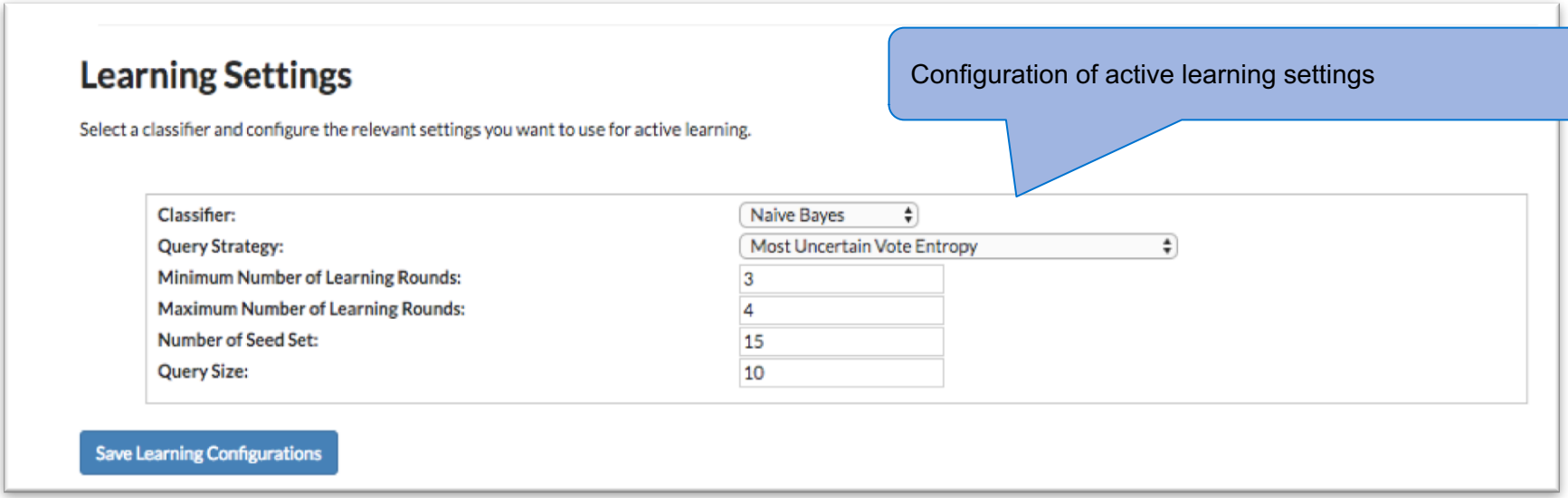
Backup – Supervised Machine Learning



Backup – Text Classification Process



Backup – Front-end (Configuration)



Backup– Front-end (Training)

Process your imported legal texts

Pipeline: DCTest

Labeling of legal documents

Please continue learning until a stopping criterion is met

Current Learning Round: 1 **Minimum Rounds To Go: 1** **Maximum Rounds To Go: 2**

Title	Promulgation Date	Prediction	Confidence (%)	Label
Keine Aufgabe des verpachteten landwirtschaftlichen Betriebs bei Veräußerung der Hofstelle; Praxis-Hinweise zur BFH-Entscheidung IV R 61/01 vom 26.06.2003	Thu Nov 13 23:00:00 CET 2003	GenericLegalDocument	0.9613485178221252	✓ Law Judgment GenericLegalDocument

Praxis-Hinweise1. Dass ein land- und forstwirtschaftlicher Betrieb auch ohne Hofstelle existieren kann, ist keine ganz neue Erkenntnis, sondern in der Praxis bereits seit vielen Jahren erprobt. Auch der BFH hat diese Veränderung zur Kenntnis genommen und deshalb bereits im Urteil vom 18.3.1999, IV R 65/98 (BStBl II 1999, 398) entschieden, dass keine Betriebsaufgabe vorliegt, wenn ein Erbe die Hofstelle erhält, ein anderer Erbe aber Stückländereien. Ebenso wenig sieht der BFH in der Veräußerung wesentlicher Betriebsflächen eine Betriebsaufgabe, sofern nicht eine absolute Minimalgrenze unterschritten wird. Diese ist bisher nicht genau definiert worden; diskutiert wird aber eine Mindestfläche von ca. 3.000 m. Seit Jahrzehnten wird darüber hinaus auch die parzellenweise Verpachtung nicht als Betriebsaufgabe behandelt, selbst wenn sie den ganzen Betrieb betrifft und die Pachtverträge nicht zu einem einheitlichen Zeitpunkt auslaufen.2. Ein Landwirt kann deshalb seinen Betrieb im Grundsatz nur aufgeben, wenn er ausdrücklich die Betriebsaufgabe erklärt. Beachten Sie dabei: Die Betriebsaufgabeerklärung wirkt nur für die Zukunft. Im Besprechungsfall hatte der Kläger erst im Einspruchsverfahren beantragt, die Versicherungsentschädigung als Aufgabegewinn zu behandeln. Darin sah der BFH keine rückwirkende Aufgabeerklärung auf den Zeitpunkt des Schadensfalls. Stattdessen liegt aber eine Aufgabe im Zeitpunkt der Erklärung gegenüber dem FA vor, die nicht mehr rückgängig gemacht werden kann. Selbst wenn der Kläger mangels Rückwirkung auf das Schadensjahr lieber auf eine Betriebsaufgabe verzichten würde, ist dies jetzt nicht mehr möglich.

Backup – Comparison of Machine Learning Frameworks I

	MLlib	Mahout	Weka	Scikit-learn	Mallet
Current Version (2nd Feb. 2016)	2.1	0.12.2	3.8	0.18	2.0.8
License	Apache Software Foundation (AFS)	Apache Software Foundation (AFS)	General Public License (GPL)	Berkeley Software Distribution (BSD)	Common Public License (CPL)
Open Source	yes	yes	yes	yes	yes
Popular Users	OpenTable, Verizon	?	Pentaho	Evernote, Spotify	?
Processing Platform	MapReduce (deprecated), Spark	Spark	wrapper for Spark available	none	none
Interface Language	Java, Scala, Python, R	Mainly Scala, Java (for older versions)	Java, R	Python	Java
Suitable for large datasets	yes	yes	partially	yes	partially
Community Support	good	moderate	good	good	moderate
Documentation	very good	moderate	good	very good	good
Support for NLP/Textual Data /AL	good	moderate (via Lucene)	good	good	very good
Configuration	simple [c]	difficult [c]	simple [c]	simple	very simple

Backup – Comparison of Machine Learning Frameworks II

	MLlib	Mahout	Weka	Scikit-learn	Mallet
Classification Algorithms					
NB	✓ [a]	✓ (Spark optimized)	✓ (batch, incremental)	✓	✓
SVM	✓ [b]	only via MLlib	✓ (batch)	✓	/
MLP	✓ [a]	only via MLlib	✓ (batch)	✓	/
Multiclass Classification	one vs. all	only via MLlib	one vs. all, one vs. one	one vs. all, one vs. one	/
Multilabel Classification	one vs. all	/	by extensions (e.g. Mulan, Meka)	one vs. all	/
Pipeline	✓ [a]	/	/	✓	✓
Total Algorithm Coverage	very high	high	very high	very high	moderate
Preprocessing					
Stemming	(by extensions (Snowball))	/	✓	✓	✓
Stop Words	✓	/	✓	✓	✓
Coverage of Feature Selection Methods	very high	low	very high	very high	moderate
Text Representation					
Word Vector	✓	✓	✓	✓	✓
TF-IDF	✓		✓	✓	
Evaluation					
Confusion Matrix	✓	✓	✓	✓	✓
ROC/AUC Curve	✓	/	✓	✓	✓

Backup – Literature study

- **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**

Backup – Hourly Billings by Task and Individual

- **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