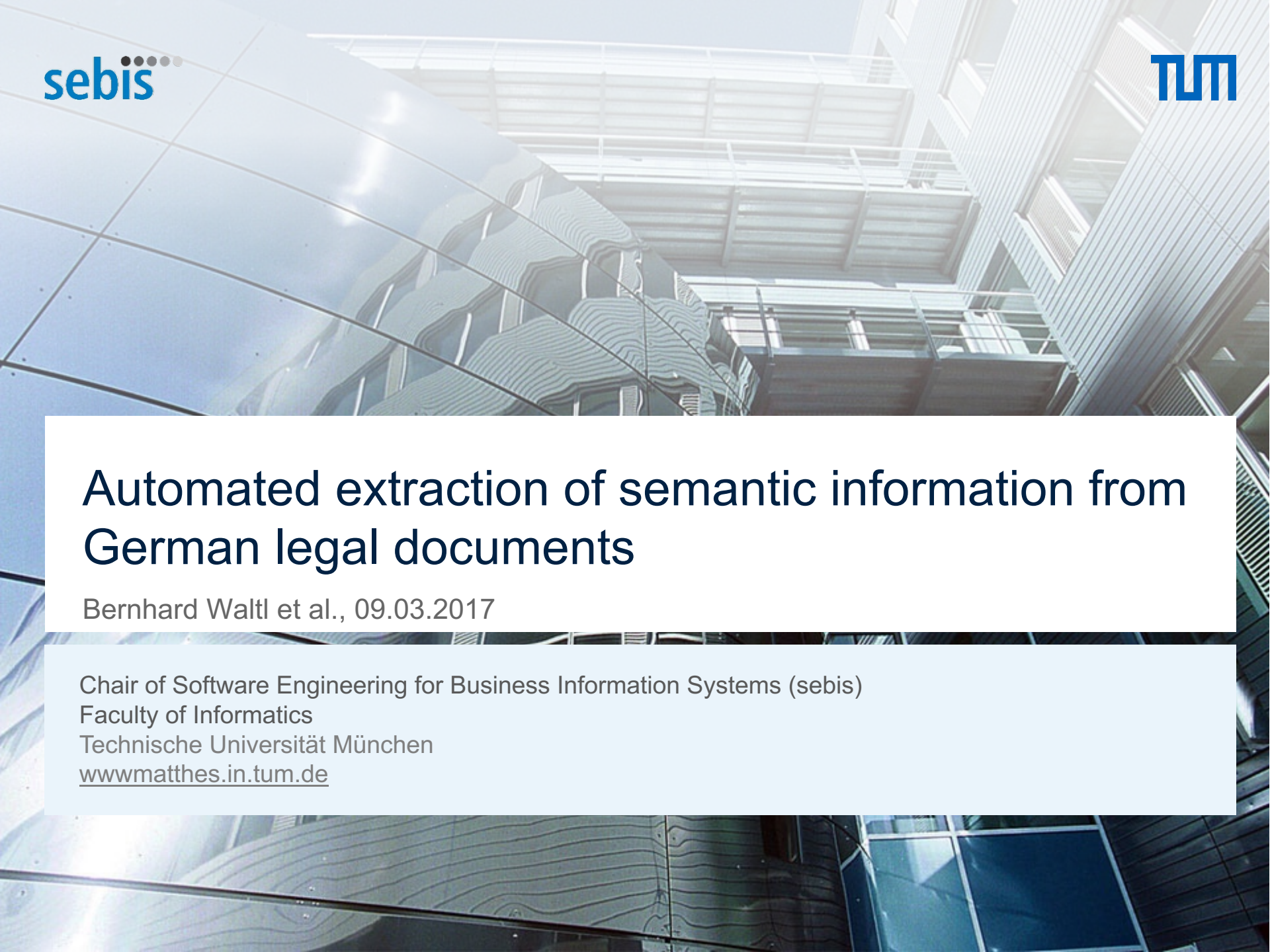


Automated extraction of semantic information from German legal documents

Bernhard Waltl et al., 09.03.2017

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

Agenda

1. Introduction

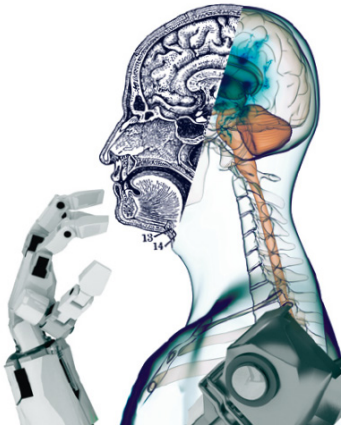
2. Data

3. Interdisciplinary research method

4. Two different use cases
   - Support of editorial staff to structure tax law judgments
   - Extract definitions and defining contexts of legal terms

5. Evaluation

6. Next Steps: What to expect?



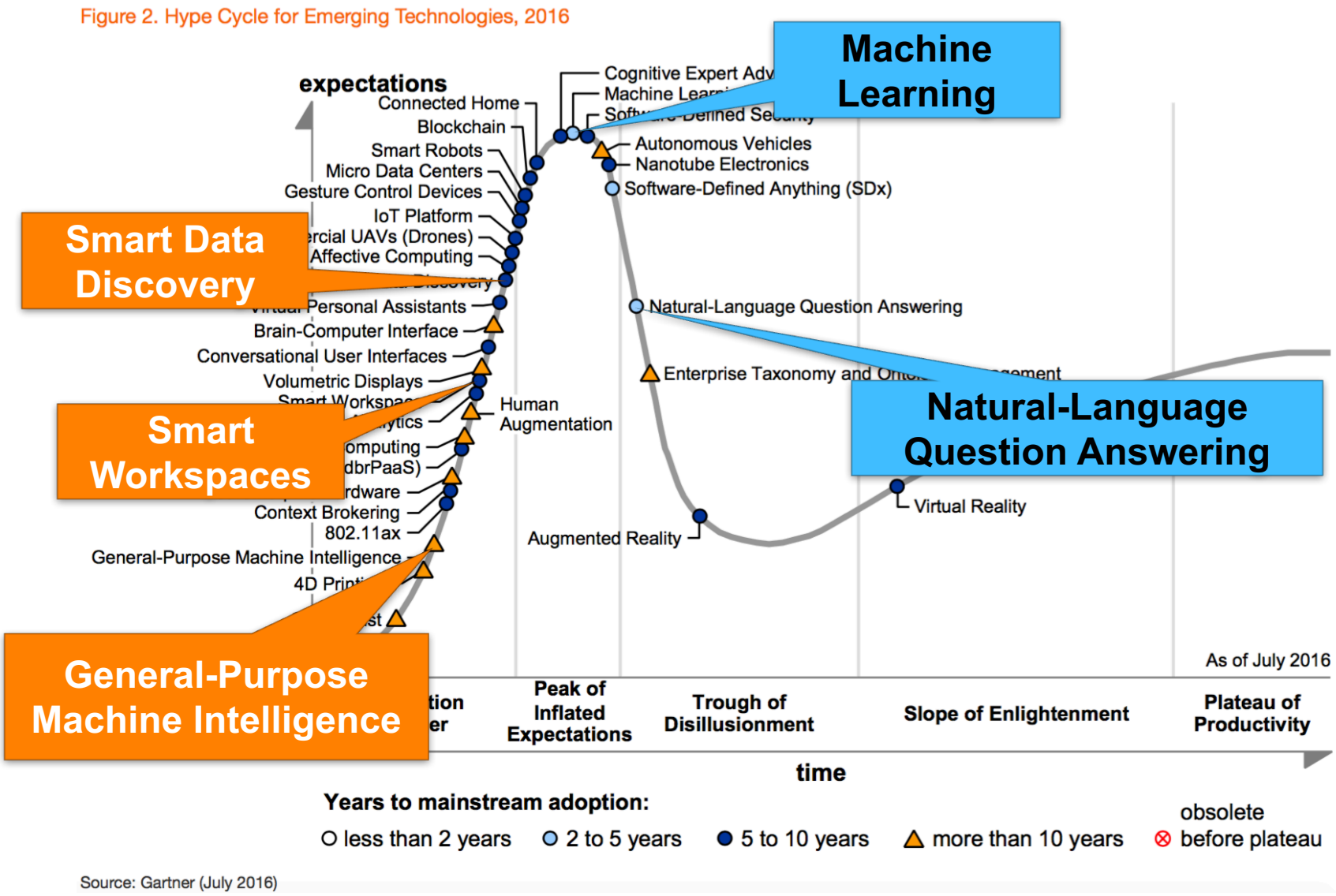
Introduction

- Processes of Legal Experts (Scientists and Lawyers) are…
  - ... time-intensive
  - ... knowledge-intensive
  - ... data-intensive

- Legal Data Science is becoming more and more attractive, because
  - ... process time and memory space are cheap
  - ... algorithms can process data fast and accurate

- In order to achieve highest accuracy,
  - algorithms (e.g., splitter & segmenter),
  - models and patterns (e.g., machine learning, pattern recognition),
- have to be adapted.

Motivation
Gartner Hype Cycle July 2016



Dataset
Cooperation between Datev eG & TUM

- **Dataset**
  - > 130 000 documents from German tax law
  - > 50 different document categories
    - Cases, judgments, decisions, laws, EU-regulations, administrative guidelines, scientific articles, etc.
  - Timespan: 1919 – 2016
  - Data format: XML & JSON
    - fully digitized
    - no scans and no OCR

Semantic Analysis of Legal Data
Cooperation between Datev eG & TUM

Topic I

- Automated detection of relevant information
  - Year of dispute (Streitjahr, Veranlagungszeitraum)
  - Re-occurring problem: triggering based on linguistic patterns

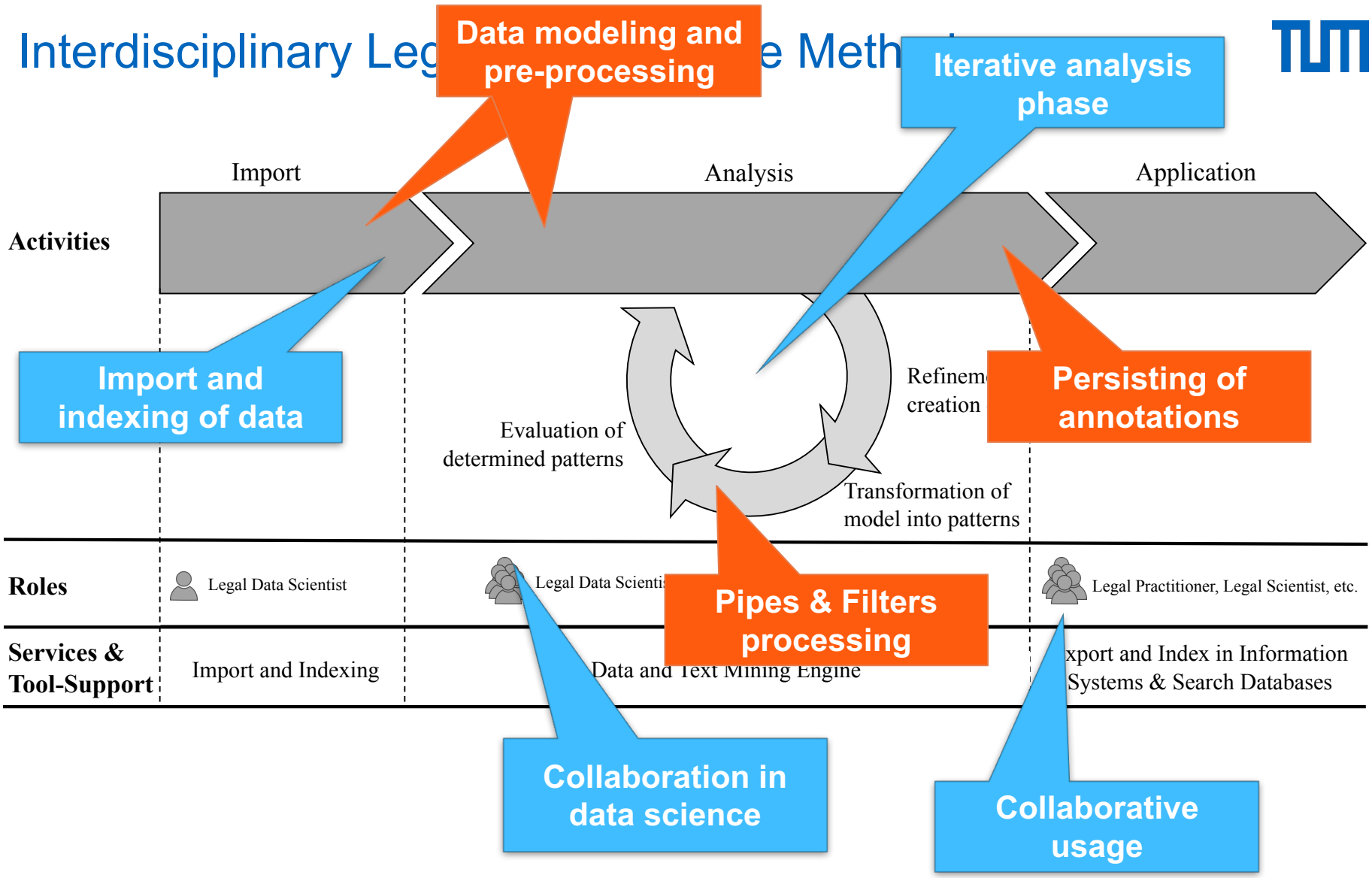
➔ Support of editorial staff

Topic II

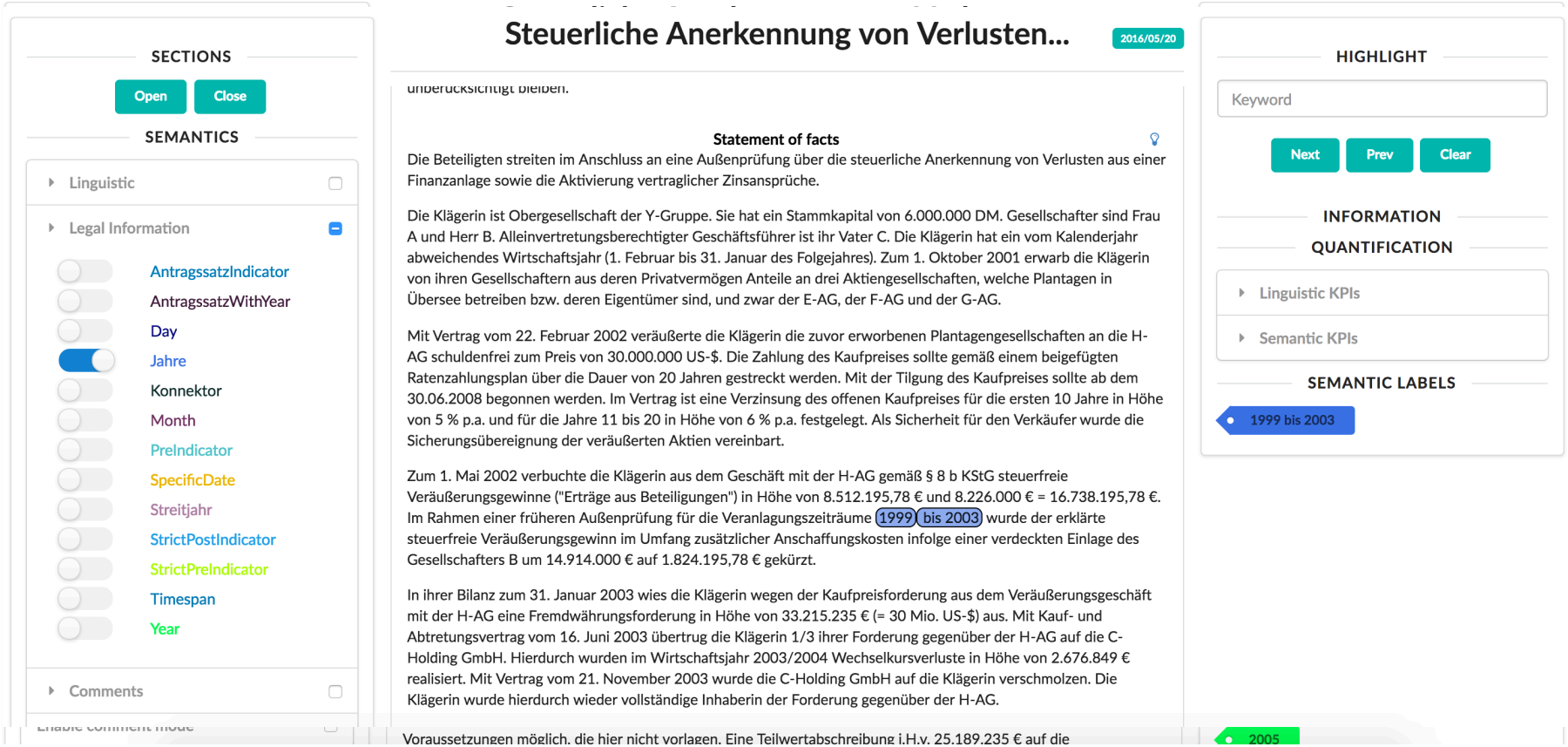
- Detection of legal definitions and terms in context of a definition
  - Interpretation support
  - Highlighting of text (context is decisive)

➔ Next generation search

Interdisciplinary Leg… e Meth…



Semantic Analysis of Legal Data
Detection of year of dispute



Link

Ruta Script
Legal Definition (excerpt)

```
1  // Basic linguistic vocabulary
2  DECLARE ISDG;
3  "im Sinne dieses Gesetzes" -> LDSache.ISDG;
4  "im Sinne des Gesetzes" -> LDSache.ISDG;
5
6  DECLARE IST;
7  "ist|sind" -> LD.IST;
8
9  DECLARE NEG;
10 "keine|kein|nicht" -> LD.NEG;
11
12 DECLARE LDIdentifier; // Declare the indicator for legal definitions
13 DECLARE LegalEntity; // Declare the legally defined entity
14 DECLARE LegalDefinition; // Declare the legal definition
15
16 // Definition of linguistic patterns and rules
17 // {{ADJ}} {{NOUN}} im Sinne dieses|des Gesetzes ist {{Phrase}}
18 ((pos.N? pos.N) {-> LD.LegalEntity} LD.ISDG) {-> LD.LDIdentifier};
19 ((pos.ADJ+ pos.N) {-> LD.LegalEntity} LD.ISDG) {-> LD.LDIdentifier};
20
21 // {{NOUN}} ist kein {{NOUN}}
22 (pos.N {-> LD.LegalEntity} LD.IST LD.NEG pos.N) {-> LD.LDIdentifier};
23 (pos.N{-PARTOF(LD.LegalEntity) -> LD.LegalEntity} LD.ISDG){->LD.LDIdentifier};
24
25 // Annotate the sentence being a legal definition as LegalDefinition
26 Sentence{CONTAINS(LD.LDIdentifier) -> LD.LegalDefinition};
27
28 // Remove temporary annotations
29 LD.IST {-> UNMARK(LD.IST)};
30 LD.NEG {-> UNMARK(LD.NEG)};
31 LD.ISDG {-> UNMARK(LD.ISDG)};
32 LD.LDIdentifier{-> UNMARK(LD.LDIdentifier)};
```

Listing 1: Linguistic pattern descriptions (LD.ruta) for the semantic entity Legal Definition using Apache Ruta

Semantic Analysis of Legal Data
Annealing approach

1. Constrain search space (e.g., Tatbestand) using structural information
2. Determine specific dates, years and timespans
3. Determining indicating sentences and phrases
4. Decide on particular patterns and prioritized Apache Ruta rules

Evaluation
Confusion Matrix

		Prediction Outcome	
		YOD	No YOD
Actual Outcome	YOD	186	21
	No YOD	11	-

$$\text{F1} \quad \frac{2 * 186}{2 * 186 + 21 + 11} \approx \mathbf{92\ \%}$$

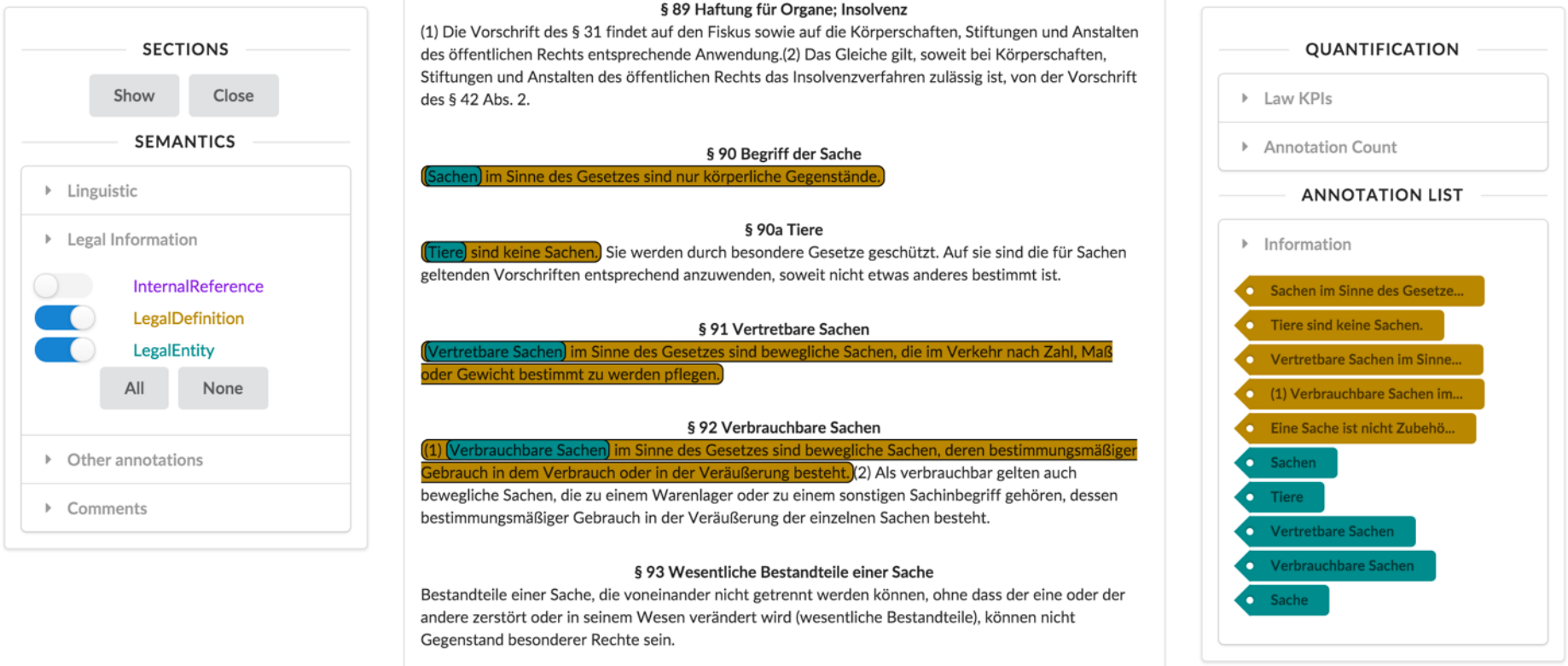
$$\text{Precision} \quad \frac{186}{186 + 11} \approx \mathbf{94\ \%}$$

$$\text{Recall} \quad \frac{186}{186 + 21} \approx \mathbf{90\ \%}$$

Table 1: Quality assessment of extraction of year of dispute (YOD) in 100 rand ly selected cases.

Trade-off

Case Study
Detection of complex patterns: The Civil Code



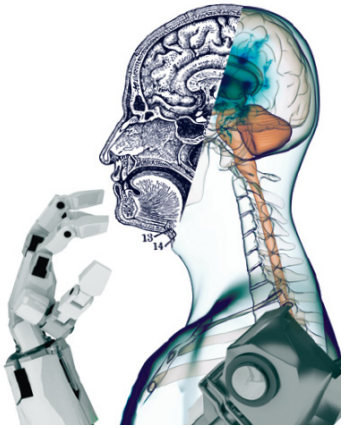
Challenges
Detection legal definitions

- Lack of an **adapted legal theory** for data science
  - UK and US legal systems are more elaborate
  - CBR – Case based reasoning
  - IBR – Issue based reasoning
  - see publications by Herbert L. Hart, Kevin Ashley, Trevor Bench-Capon, etc.

- **Inter-annotator agreement** is problematic
  - How to define a gold standard?

- **Context of interpretation** is decisive
  - What is the reason/rationale for an argument?

- High **linguistic variety**
  - Even in a restricted legal domain, e.g., tax law



Semantic Analysis
Next Steps

- Rule-based systems are well suited to determine **regularly occurring patterns** for extraction, classification, and categorization problems.

- Humans can...
  - ... understand
  - ... create
  - ... maintain and adapt the rules.

- **Next step:** Active machine learning approaches for content classification
  - E.g., Naive Bayes, Support Vector Machine, Neural Networks, etc.
  - Statistics & Probabilities (“looks like”, “is similar to”, etc.)
  - Training is different

- **Claim:** Difficulty for users should not increase!

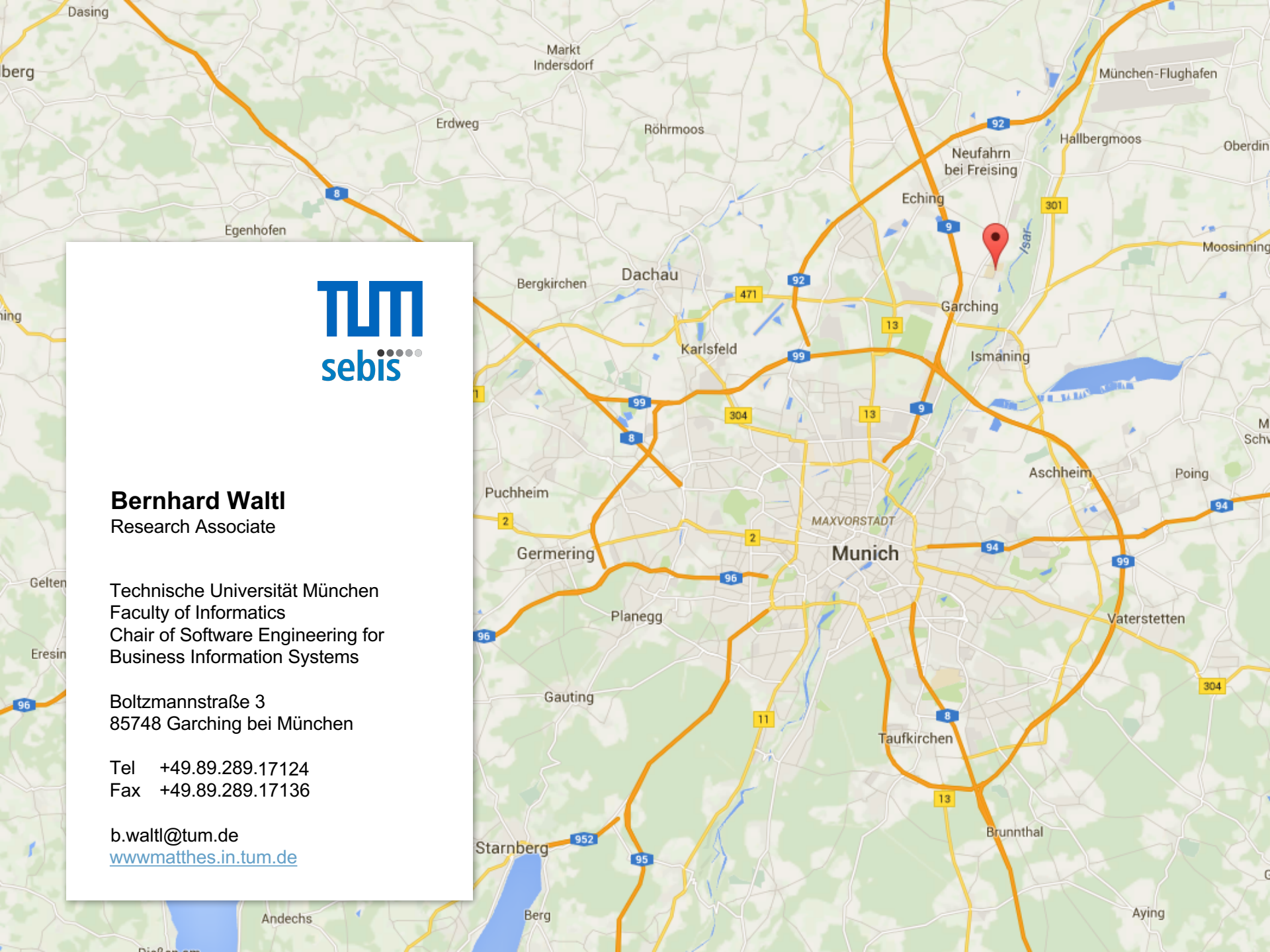
Conclusion and Outlook

- Legal data analysis is well established in legal informatics
  - Related work since decades
  - Text mining is becoming in particular relevant
  - Several approaches exist but lack of reuse in the domain

- Data Science Environment for German Legal Texts
  - Collaborative and interactive web application tailored to German legislation
  - **NOT only software but also methodology**

- Different usage scenarios
  - Extraction of information
  - Support of search in unstructured information

- Lexalyze: An interdisciplinary research program
  - www.lexalyze.de



Bernhard Waltl
Research Associate

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

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17124
Fax +49.89.289.17136

b.waltl@tum.de
wwwmatthes.in.tum.de

Reference Architecture
Overall

Implementation Details

- Java Web Application (Play Framework)
- ElasticSearch
- Apache UIMA
  - DKPro UIMAFit
  - Apache Ruta