

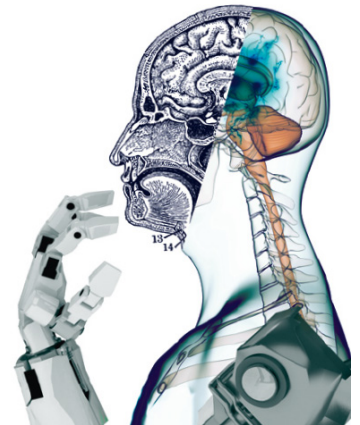
Automated extraction of semantic information from German legal documents

Bernhard Walzl et al., 09.03.2017

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

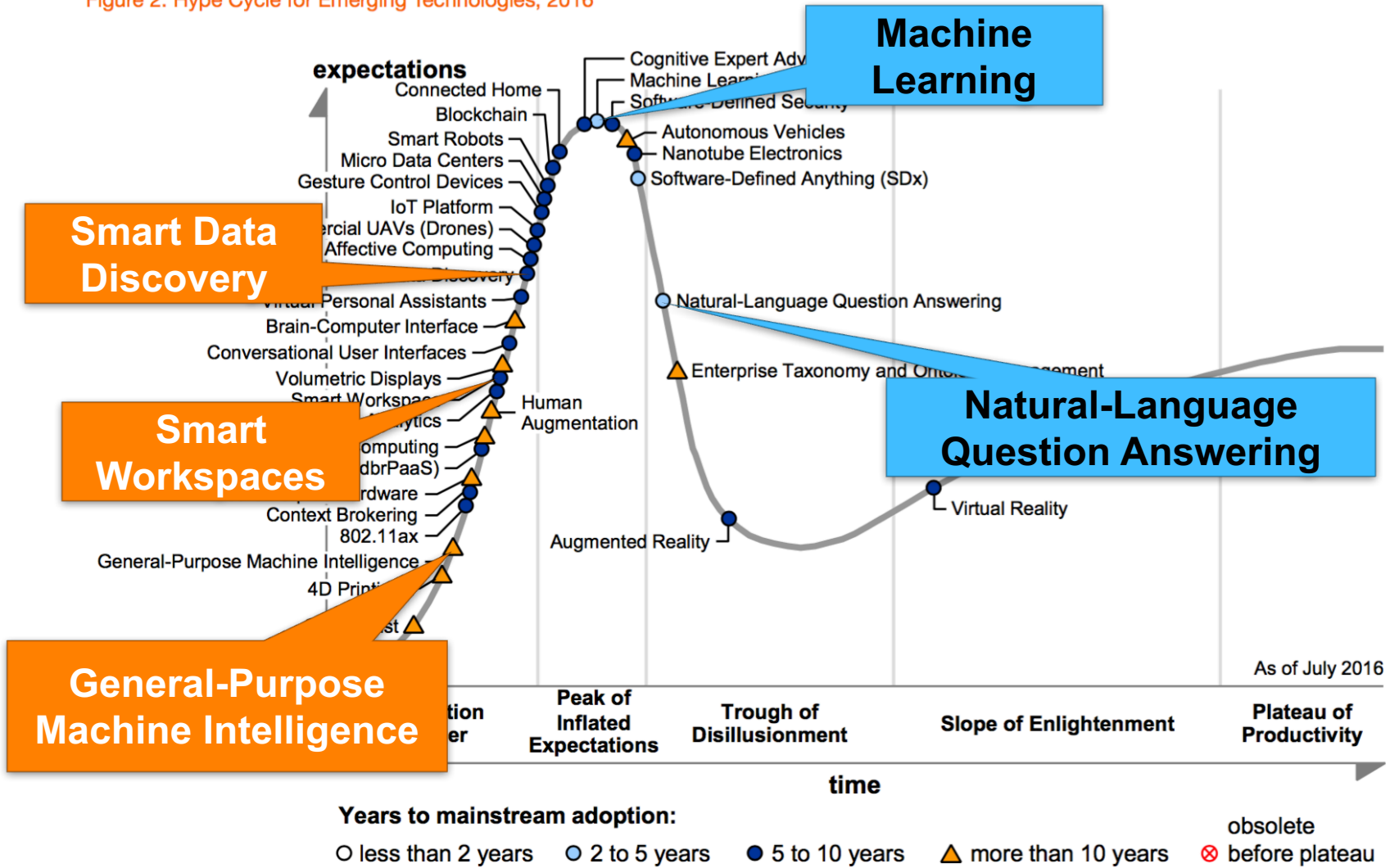


- Processes of Legal Experts (Scientists and Lawyers) are...
 - ... time-intensive
 - ... knowledge-intensive
 - ... data-intensive
- Legal Data Science is becoming more and more attractive, because
 - ... process time and memory space are cheap
 - ... algorithms can process data fast and accurate
- 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

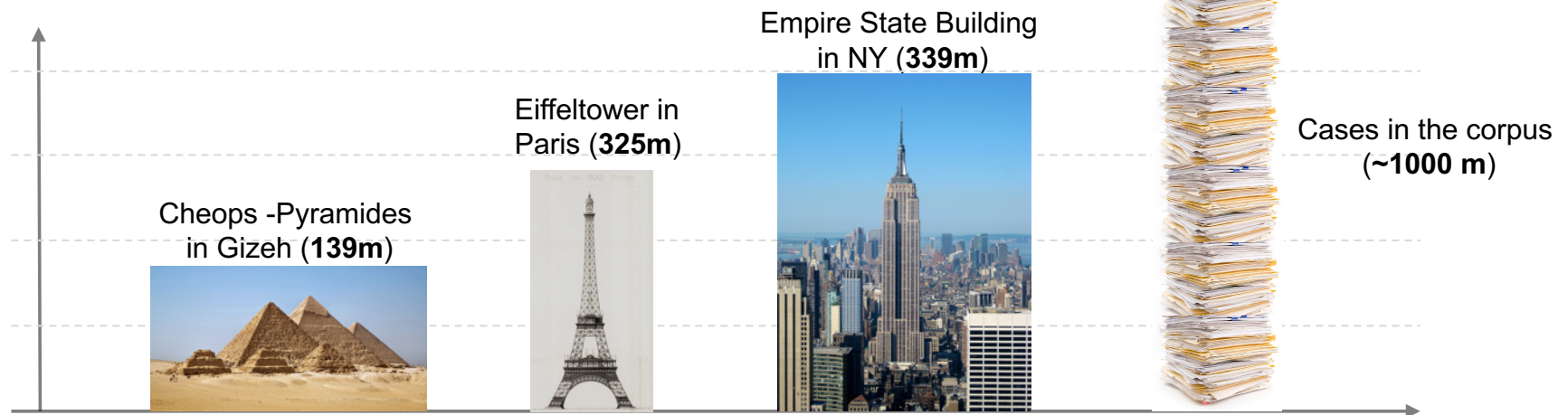
Figure 2. Hype Cycle for Emerging Technologies, 2016



Source: Gartner (July 2016)

- **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



Topic I

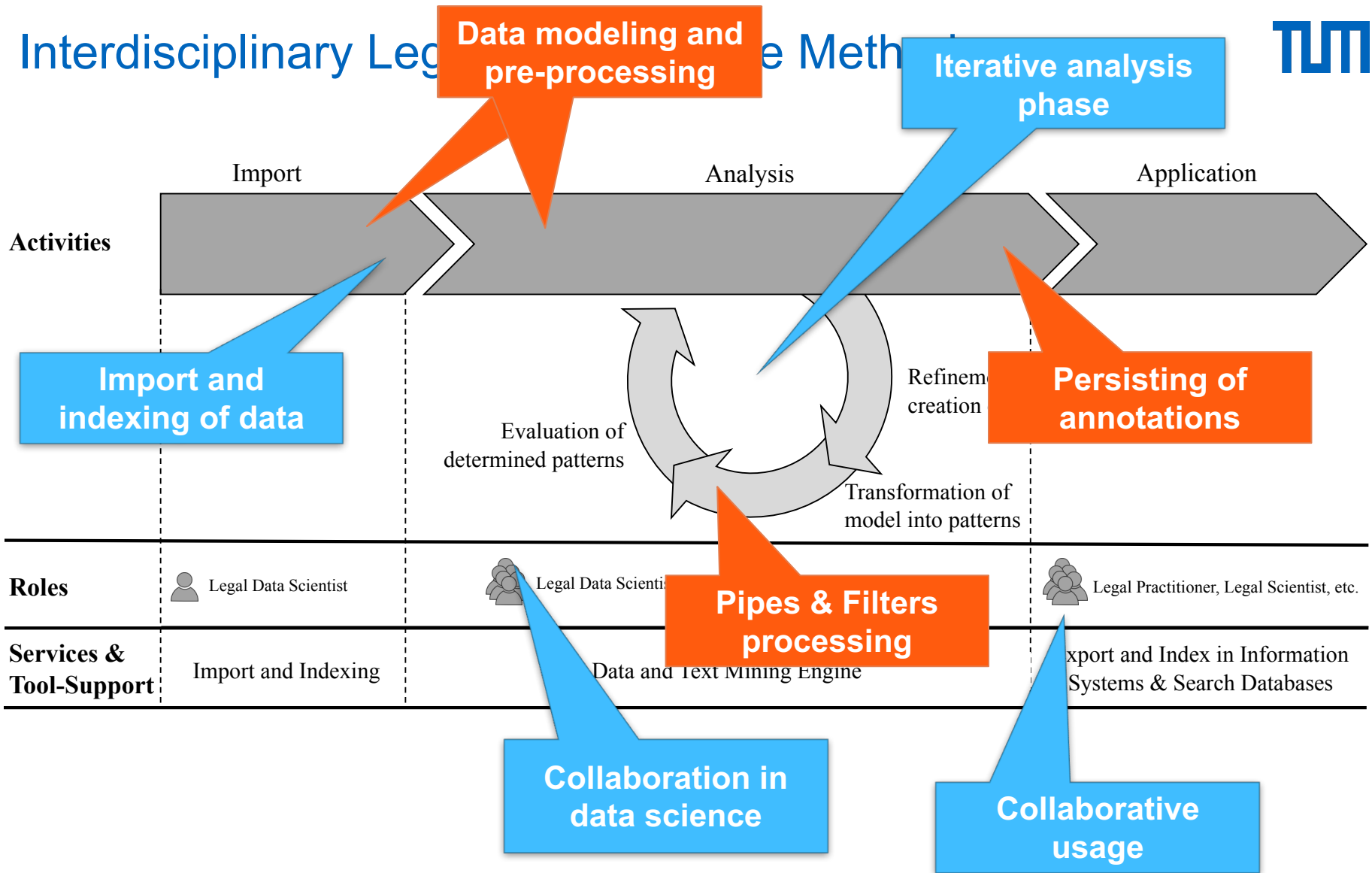
- Automated detection of relevant information
 - Year of dispute (Streitjahr, Veranlagungszeitraum)
 - Re-occurring problem: triggering based on linguistic patterns

→ 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



Semantic Analysis of Legal Data

Detection of year of dispute

SECTIONS

Open Close

SEMANTICS

- Linguistic
- Legal Information
 - AntragssatzIndicator
 - AntragssatzWithYear
 - Day
 - Jahre**
 - Konnektor
 - Month
 - PreIndicator
 - SpecificDate
 - Streitjahr
 - StrictPostIndicator
 - StrictPreIndicator
 - Timespan
 - Year
- Comments

Steuerliche Anerkennung von Verlusten...

2016/05/20

unberücksichtigt bleiben.

Statement of facts

Die Beteiligten streiten im Anschluss an eine Außenprüfung über die steuerliche Anerkennung von Verlusten aus einer Finanzanlage sowie die Aktivierung vertraglicher Zinsansprüche.

Die Klägerin ist Obergesellschaft der Y-Gruppe. Sie hat ein Stammkapital von 6.000.000 DM. Gesellschafter sind Frau A und Herr B. Alleinvertretungsberechtigter Geschäftsführer ist ihr Vater C. Die Klägerin hat ein vom Kalenderjahr abweichendes Wirtschaftsjahr (1. Februar bis 31. Januar des Folgejahres). Zum 1. Oktober 2001 erwarb die Klägerin von ihren Gesellschaftern aus deren Privatvermögen Anteile an drei Aktiengesellschaften, welche Plantagen in Übersee betreiben bzw. deren Eigentümer sind, und zwar der E-AG, der F-AG und der G-AG.

Mit Vertrag vom 22. Februar 2002 veräußerte die Klägerin die zuvor erworbenen Plantagengesellschaften an die H-AG schuldenfrei zum Preis von 30.000.000 US-\$. Die Zahlung des Kaufpreises sollte gemäß einem beigefügten Ratenzahlungsplan über die Dauer von 20 Jahren gestreckt werden. Mit der Tilgung des Kaufpreises sollte ab dem 30.06.2008 begonnen werden. Im Vertrag ist eine Verzinsung des offenen Kaufpreises für die ersten 10 Jahre in Höhe von 5 % p.a. und für die Jahre 11 bis 20 in Höhe von 6 % p.a. festgelegt. Als Sicherheit für den Verkäufer wurde die Sicherungsübereignung der veräußerten Aktien vereinbart.

Zum 1. Mai 2002 verbuchte die Klägerin aus dem Geschäft mit der H-AG gemäß § 8 b KStG steuerfreie Veräußerungsgewinne ("Erträge aus Beteiligungen") in Höhe von 8.512.195,78 € und 8.226.000 € = 16.738.195,78 €. Im Rahmen einer früheren Außenprüfung für die Veranlagungszeiträume **1999 bis 2003** wurde der erklärte steuerfreie Veräußerungsgewinn im Umfang zusätzlicher Anschaffungskosten infolge einer verdeckten Einlage des Gesellschafters B um 14.914.000 € auf 1.824.195,78 € gekürzt.

In ihrer Bilanz zum 31. Januar 2003 wies die Klägerin wegen der Kaufpreisforderung aus dem Veräußerungsgeschäft mit der H-AG eine Fremdwährungsforderung in Höhe von 33.215.235 € (= 30 Mio. US-\$) aus. Mit Kauf- und Abtretungsvertrag vom 16. Juni 2003 übertrug die Klägerin 1/3 ihrer Forderung gegenüber der H-AG auf die C-Holding GmbH. Hierdurch wurden im Wirtschaftsjahr 2003/2004 Wechselkursverluste in Höhe von 2.676.849 € realisiert. Mit Vertrag vom 21. November 2003 wurde die C-Holding GmbH auf die Klägerin verschmolzen. Die Klägerin wurde hierdurch wieder vollständige Inhaberin der Forderung gegenüber der H-AG.

Voraussetzungen möglich, die hier nicht vorlagen. Eine Teilwertabschreibung i.H.v. 25.189.235 € auf die

HIGHLIGHT

Keyword

Next Prev Clear

INFORMATION

QUANTIFICATION

- Linguistic KPIs
- Semantic KPIs

SEMANTIC LABELS

1999 bis 2003

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	-

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

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

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

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



Trade-off

SECTIONS

Show Close

SEMANTICS

- Linguistic
- Legal Information
 - InternalReference
 - LegalDefinition
 - LegalEntity
- All None
- Other annotations
- Comments

§ 89 Haftung für Organe; Insolvenz

(1) Die Vorschrift des § 31 findet auf den Fiskus sowie auf die Körperschaften, Stiftungen und Anstalten des öffentlichen Rechts entsprechende Anwendung.(2) Das Gleiche gilt, soweit bei Körperschaften, Stiftungen und Anstalten des öffentlichen Rechts das Insolvenzverfahren zulässig ist, von der Vorschrift des § 42 Abs. 2.

§ 90 Begriff der Sache

Sachen im Sinne des Gesetzes sind nur körperliche Gegenstände.

§ 90a Tiere

Tiere sind keine Sachen. Sie werden durch besondere Gesetze geschützt. Auf sie sind die für Sachen geltenden Vorschriften entsprechend anzuwenden, soweit nicht etwas anderes bestimmt ist.

§ 91 Vertretbare Sachen

Vertretbare Sachen im Sinne des Gesetzes sind bewegliche Sachen, die im Verkehr nach Zahl, Maß oder Gewicht bestimmt zu werden pflegen.

§ 92 Verbrauchbare Sachen

(1) Verbrauchbare Sachen im Sinne des Gesetzes sind bewegliche Sachen, deren bestimmungsmäßiger Gebrauch in dem Verbrauch oder in der Veräußerung besteht.(2) Als verbrauchbar gelten auch bewegliche Sachen, die zu einem Warenlager oder zu einem sonstigen Sachinbegriff gehören, dessen bestimmungsmäßiger Gebrauch in der Veräußerung der einzelnen Sachen besteht.

§ 93 Wesentliche Bestandteile einer Sache

Bestandteile einer Sache, die voneinander nicht getrennt werden können, ohne dass der eine oder der andere zerstört oder in seinem Wesen verändert wird (wesentliche Bestandteile), können nicht Gegenstand besonderer Rechte sein.

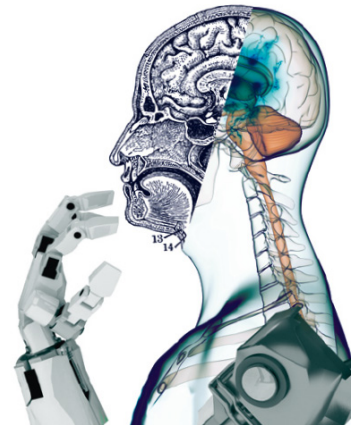
QUANTIFICATION

- Law KPIs
- Annotation Count

ANNOTATION LIST

- Information
 - Sachen im Sinne des Gesetze...
 - Tiere sind keine Sachen.
 - Vertretbare Sachen im Sinne...
 - (1) Verbrauchbare Sachen im...
 - Eine Sache ist nicht Zubehö...
 - Sachen
 - Tiere
 - Vertretbare Sachen
 - Verbrauchbare Sachen
 - Sache

- 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



- 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!

- 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





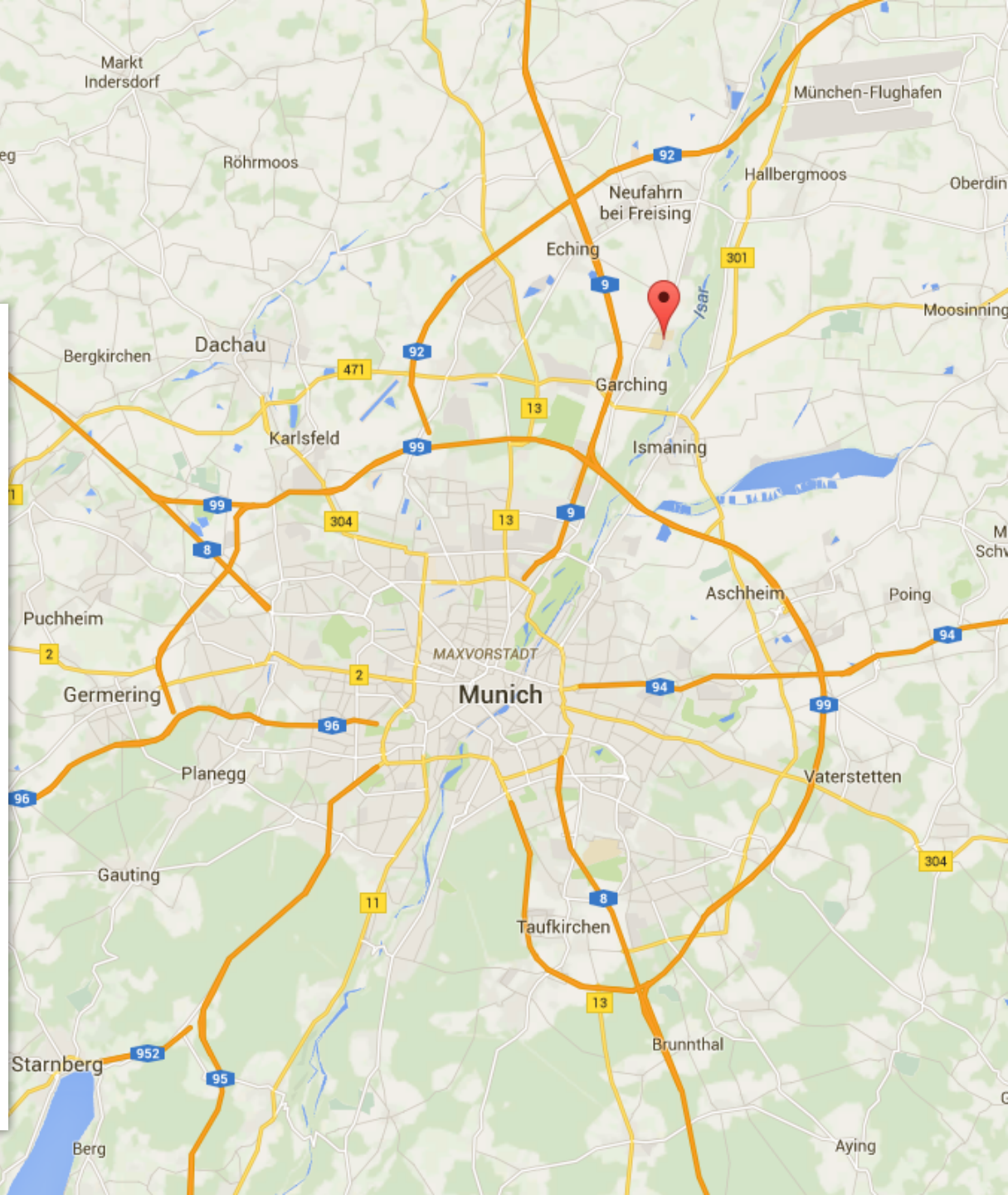
Bernhard Waltl
Research Associate

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
www.matthes.in.tum.de



Reference Architecture

Overall



Implementation Details

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

