UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Legal Text Analytics
Challenges and Pitfalls

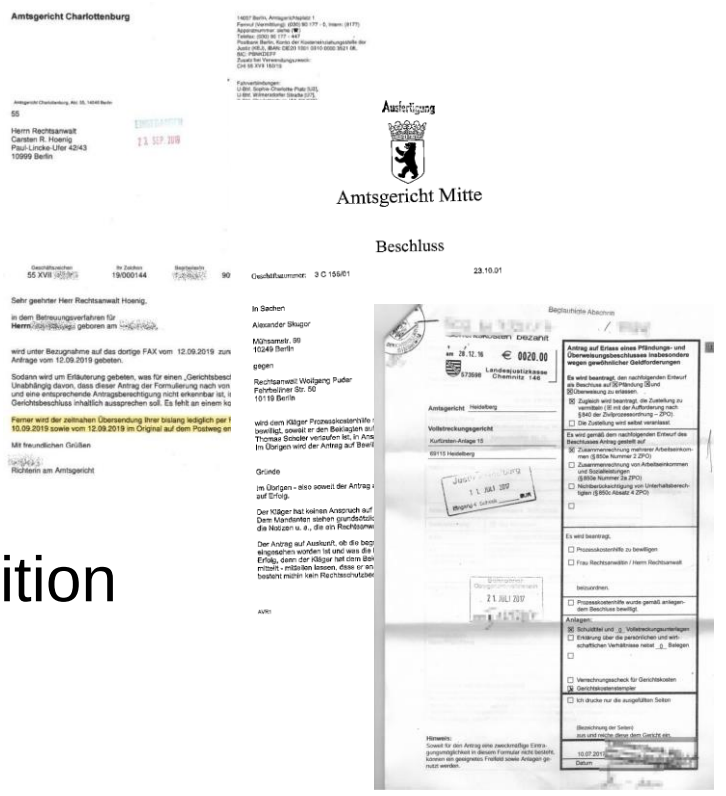
Prof. Dr. Michael Gertz
Institute of Computer Science
Heidelberg University
gertz@informatik.uni-heidelberg.de

Text Data


Schmidt
EStG
29. Auflage
2010
Kreutziger
Schaffner
Stephany
BewG
2. Auflage
Meincke
ErbStG
15. Auflage
BUNJES/GEIST
UStG
7. Auflage
STRECK
KStG
6. Auflage
KLEIN
AO
8. Auflage

The Digitized Version

The First Challenge...



Amtsgericht Charlottenburg

Amtsgericht Mitte

Beschluss

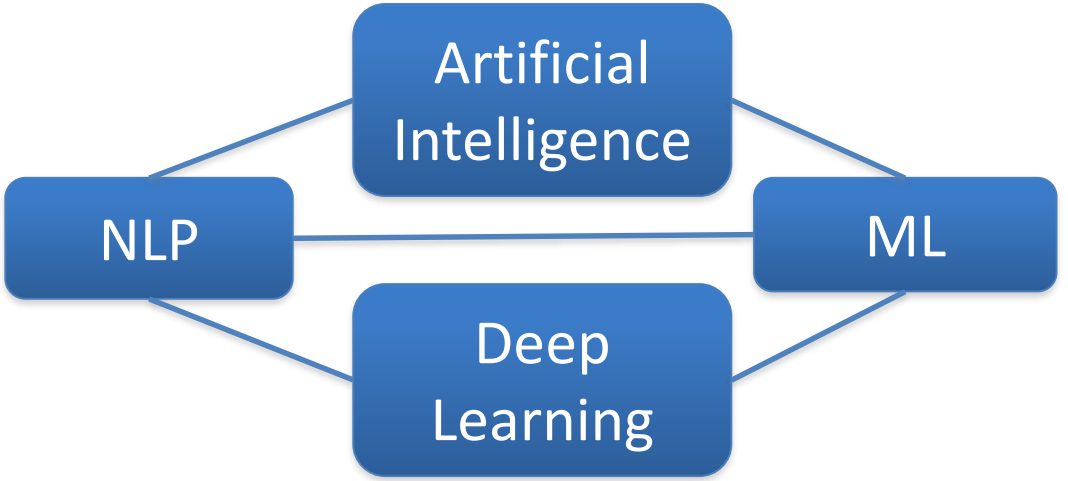
Good Optical Character Recognition (OCR) software is still a must...

Text Analytics

Text analytics are techniques that employ methods from

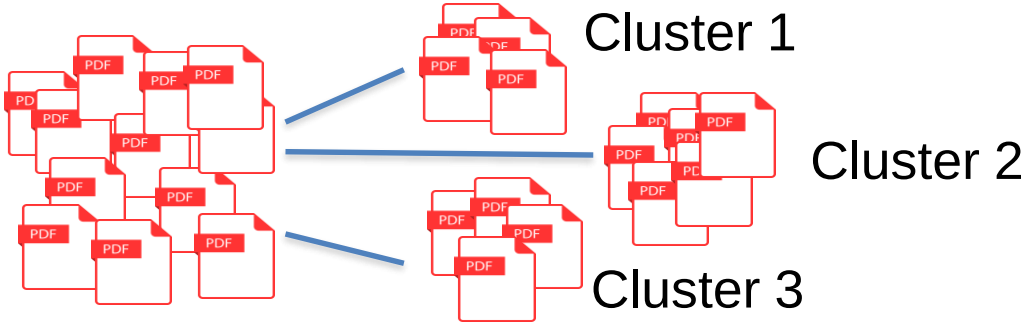
- **natural language processing** (NLP),
- **machine learning** (ML), and
- **computational linguistics** (CL)

to extract relevant information from text data.

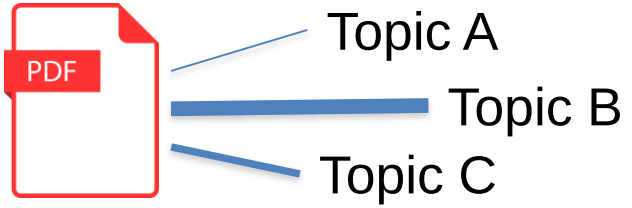


Text Analytics: Methods

- **Document clustering**: determine groups of documents such that documents in a group are similar (*unsupervised*)

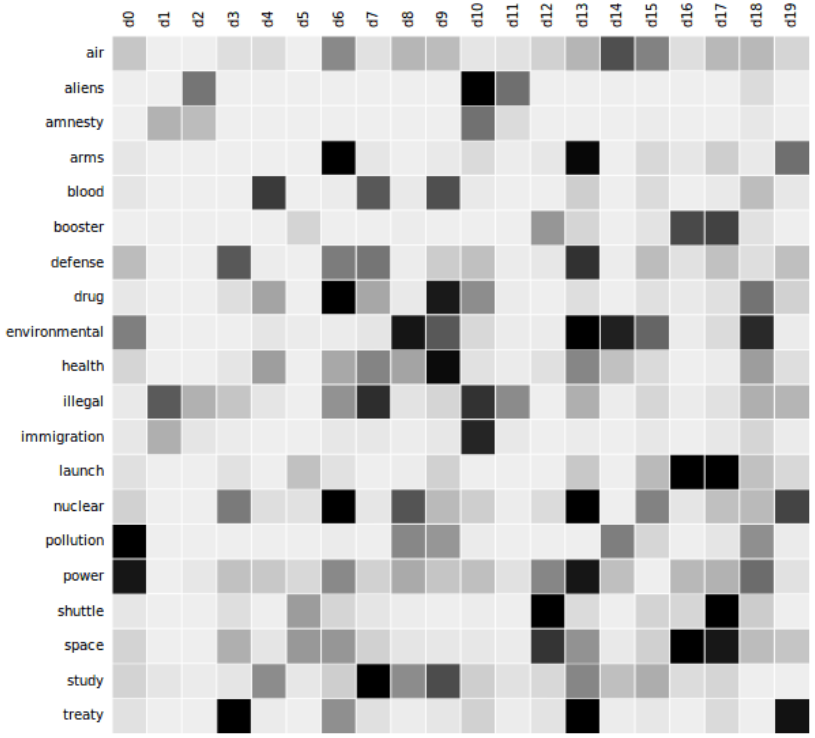


- **Document classification**: determine the topic(s) or class label(s) for a given a document (*supervised*)



Text Analytics: Methods (2)

- **Topic detection**: for a collection of documents, determine the themes or topics the documents are about.

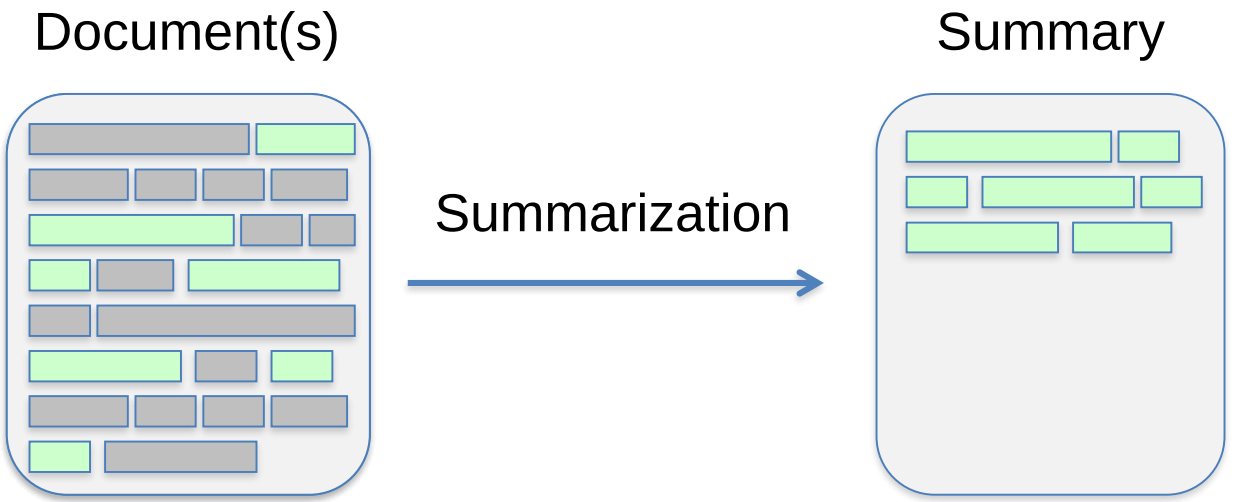


Document–Term Matrix

- Many statistical methods
- Co-occurrence of words or keywords
- Ranked list of words that best describe a topic

Text Analytics: Methods (3)

- **Summarization**: reduce length and detail of a document or collection while preserving its key points and meaning.



Text Analytics: Methods (4)

- **Question answering**: given a natural language query and a set of documents, find the best answer to the query. Result can be
  - excerpts from a document, e.g., a sentence, or
  - or summary

Goes into the area of natural language generation

Information Extraction

- Methods to extract structured information from documents
- Focus dates back to the beginning of NLP in the 70s
- Most fundamental task is
  Named Entity Recognition (NER)

- Extracted information builds backbone of many subsequent text analytics tasks and methods.

Named Entity Recognition

- Strongly depends on application domain, available ontologies and taxonomies, standard vocabularies...
- Common types of **named entities** in text:
  - Persons
  - Organizations
  - Locations
  - Times and dates
  - Monetary values
  - (Legal) concepts
  - ...

Named Entity Recognition – Example

Der Zeuge *Heiß* hat gegenüber dem Ausschuss bestätigt, dass sich die Delegation nach der Reise mit dem damaligen Kanzleramtsminister *Pofalla* getroffen und von der Reise berichtet habe. *Pofalla* habe besonders der „gesamte Ablauf des Gesprächs mit den Amerikanern“ interessiert. Dabei sei das Angebot der USA eine „wichtige Nachricht“ gewesen, die er „nicht enttäuscht“ aufgenommen habe.[1850] Auf den dabei geäußerten Informationen beruhte nach Aussage des Zeugen *Heiß* dann das Pressestatement, das *Ronald Pofalla* am 12. August 2013 abgab.[1851] Er selbst hat sich in der Vernehmung durch den Ausschuss als Zeuge folgendermaßen erinnert:

Preparatory Steps in Text Analytics

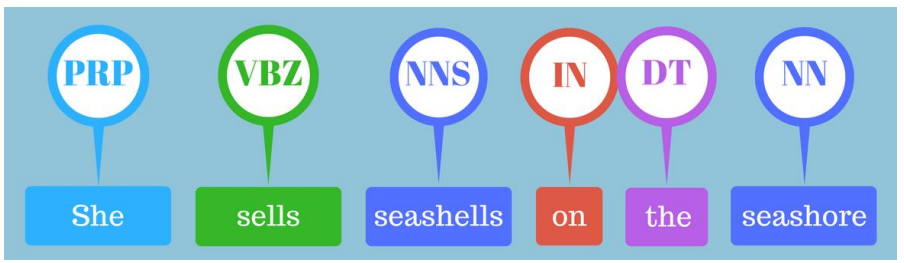
1. Sentence Segmentation

   “I don’t like Mondays.” “It is cold, i.e., freezing cold.”

2. Word Tokenization

   “I”, “do”, “n’t”, “like”, “Mondays”, “.”

3. Part-of-Speech Tagging

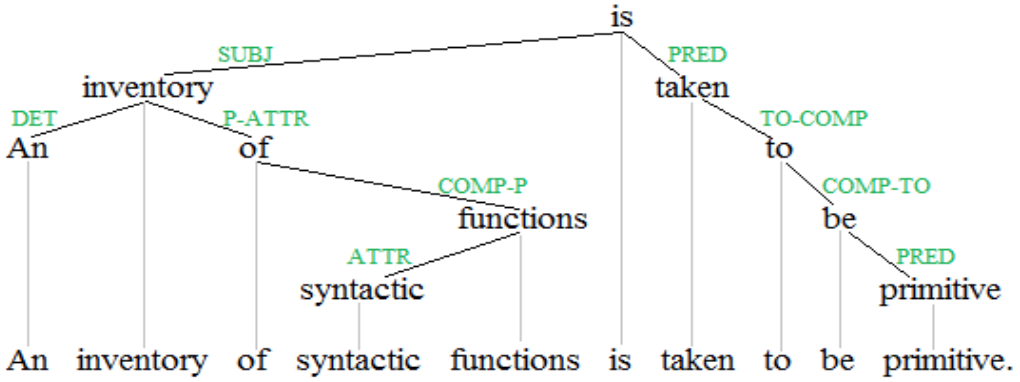


Preparatory Steps in Text Analytics (2)

4. (Optional) Lemmatization or stemming

   “He came with two lawyers” ➔ “He come with two lawyer”

5. Dependency Parsing (important for NER)



6. Chunking (detect composite phrases)

   “The trainee lawyer’s baby steps on civil law.”

Legal Text Analytics

In Legal Text Analytics, legal documents are of prime interest:

- Statutes
- Contracts
- Complaints
- Court decisions
- Directives
- Comments
- Patents
- …

Legal Text Analytics Tasks

- **Legal research**: “process of identifying and retrieving information necessary to support legal decision-making”
  - primary sources of law (statues, cases, ...)
  - secondary sources (law reviews, ...)
- **Problem**: how to guide search, formulate the “right” query, detect relevant sources, organize search results, ...?

Legal Text Analytics Tasks (2)

- **Electronic discovery**: determining electronically-stored information that is relevant for a lawsuit or investigation. “Sifting through files...”
- **Technology-Assisted Review (TAR)**: uses (supervised) machine learning to determine relevance of a document, aka “predictive coding”

Legal Text Analytics Tasks (3)

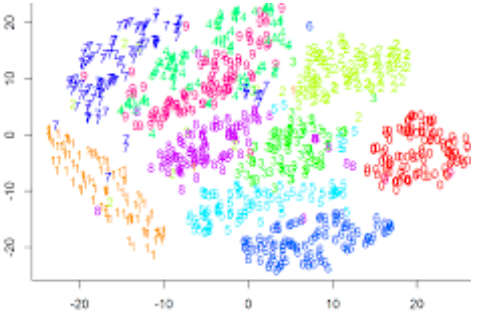
- **Contract review**: decompose contract into individual clauses and provisions to
  - compare against standard clauses
  - extract key information

- Can become quite complex, e.g., due diligence

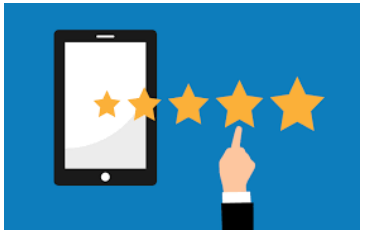
- **Document automation**: enable automatic generation of legal documents using fill-in-the-blanks template mechanisms.

It all boils down to...

- Extracting **features** from documents and text
  - structure, keywords, named entities, …
  - each document or parts thereof live in a high-dimensional vector space



- Employing **similarity measure** to
  - determine **relevancy** of a document with respect to a query (aka **ranking**)

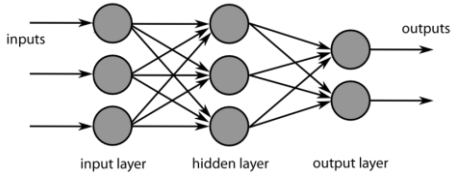


AI now solves all these problems, right?

- Natural language is **complex**.

  “Environmental regulators grill business owner over illegal coal fires.”

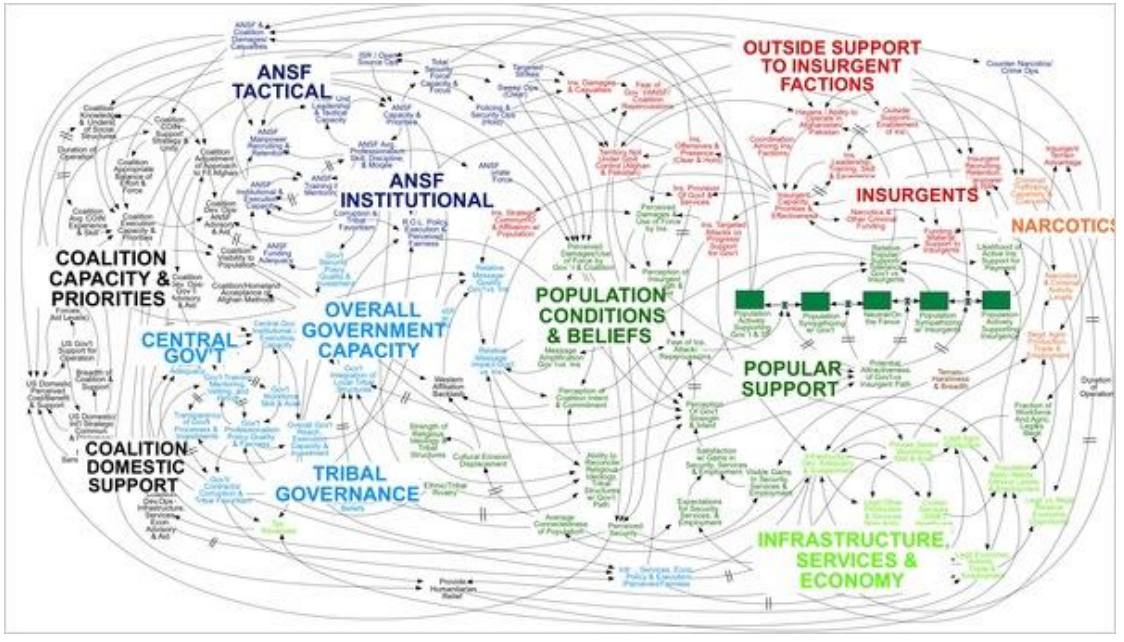
- What **text features** are relevant is quite subjective.
- Recent **Deep Learning** approaches need a lot of data for **training language models**.
  - They help to improve key analytics tasks such as sentence splitting, NER, or chunking.
  - They even can capture semantics (e.g., synonyms)



Legal Information Networks

Observations

- Law, society, and state are complex, often very dynamic, systems with many heterogeneous interacting agents.
- For investigating a legal aspect, the context matters a lot.



Context Matters

Testimonies

Deutscher Bundestag – 18. Wahlperiode – 509 – **Drucksache 18/12850**

Der Datentransfer nach § 4 Artikel 10-Gesetz stellt einen weiteren Eingriff (zusätzlich zur Erfassung und Speicherung) in das Grundrecht aus Art. 10 Abs. 1 GG dar.[2274] Die Vorschrift stellt einen abschließenden Katalog von Übermittlungssachverhalten dar.[2275]

Andere Stellen i. S. d. § 4 Abs. 4 Artikel 10-Gesetz, an die G 10-Daten aus Einzelfallerfassung durch das BfV übermittelt werden dürfen, sind Behörden, die mit präventiv-polizeilichen Aufgaben oder mit der Verfolgung von Straftaten betraut sind.[2276] Die Gesetzesbegründung führt nur Behörden mit präventiv-polizeilichen und Strafverfolgungsaufgaben als zulässige Empfänger der Daten an.[2277]

Nach der ständigen Rechtsauffassung des BfV stellte § 4 Abs. 4 Artikel 10-Gesetz indes eine „völlig hinreichende Rechtsgrundlage" zur Übermittlung auch an ausländische Nachrichtendienste dar.[2278] Im BfV war man der Ansicht, die in § 4 Abs. 4 Artikel 10-Gesetz genannten Übermittlungszwecke könnten auch durch eine Übermittlung an ausländische Stellen erzielt werden.[2279] Die Hinzufügung einer eigenständigen adressatenbezogenen – im Bezug auf die Zusammenarbeit mit ausländischen Partnerdiensten – Übermittlungsvorschrift sei „systemfremd".[2280] Die Individualüberwachung und strategische Überwachung seien „wesensmäßig verschiedene Regelungsmaterien", die im Artikel 10-Gesetz unterschiedlich behandelt würden, sodass sich aus dem Vorliegen einer konkreten Regelung beim BND für die Übermittlung von Daten aus der strategischen Fernmeldeüberwachung an ausländische Stellen (§ 7a Artikel 10-Gesetz) kein Umkehrschluss dahingehend ziehen ließe, dass eine solche für das BfV ebenfalls erforderlich sei.[2281]

Nach einem Schreiben des BfV an das BMI vom 22. Januar 2014 sei die G 10-Kommission über diese Rechtsauffassung und die damit verbundene Übermittlungspraxis des BfV im Rahmen mehrerer Kontrollbesuche zwischen 2009 und Herbst 2013 informiert worden und habe im Jahr 2011 in einem Einzelfall einen Kennzeichnungsverzicht gemäß § 4 Abs. 3 Artikel 10-Gesetz erklärt.[2282]

ccc) Dienstvorschrift zur Datenweitergabe an ausländische Nachrichtendienste

Die Datenweitergabe an ausländische Nachrichtendienste wird im BfV in der sog. Dienstvorschrift über die Beziehungen des Bundesamtes für Verfassungsschutz zu ausländischen Nachrichtendiensten (sog. DV-Ausland) geregelt.[2283] Die einschlägigen Bestimmungen dieser Dienstvorschrift, Stand 1. November 1998, lauteten:

„Personenbezogene Daten dürfen gemäß § 19 Abs. 3 BVerfSchG im übrigen an ausländische Nachrichtendienste übermittelt werden, wenn:

Vorabfassung - wird durch die endgültige Fass

2281) Schreiben BfV (L3 / 3B) an BMI vom 22. Januar 2014, MAT A BfV-9/2 (Tgh.-Nr. 39/14 – GEHEIM), Ordner 4, Bl. 393 (394), (VS-NfD – insoweit offen).
2282) Schreiben BfV (L3 / 3B) an BMI vom 22. Januar 2014, MAT A BfV-9/2 (Tgh.-Nr. 39/14 – GEHEIM), Ordner 4, Bl. 393 (394), (VS-NfD – insoweit offen).
2283) sog. DV-Ausland, MAT A BfV-5 (VS-NfD – insoweit offen).

Statutes

§ 4 Prüf-, Kennzeichnungs- und Löschungspflichten, Übermittlungen, Zweckbindung

(1) Die erhebende Stelle prüft unverzüglich und sodann in Abständen von höchstens sechs Monaten, ob die erhobenen personenbezogenen Daten im Rahmen ihrer Aufgaben allein oder zusammen mit bereits vorliegenden Daten für die in § 1 Abs. 1 Nr. 1 bestimmten Zwecke erforderlich sind. Soweit die Daten für diese Zwecke nicht erforderlich sind und nicht für eine Übermittlung an andere Stellen benötigt werden, sind sie unverzüglich unter Aufsicht eines Bediensteten, der die Befähigung zum Richteramt hat, zu löschen. Die Löschung ist zu protokollieren. Die Protokolldaten dürfen ausschließlich zur Durchführung der Datenschutzkontrolle verwendet werden. Die Protokolldaten sind am Ende des Kalenderjahres, das dem Jahr der Protokollierung folgt, zu löschen. Die Löschung der Daten unterbleibt, soweit die Daten für eine Mitteilung nach § 12 Abs. 1 oder für eine gerichtliche Nachprüfung der Rechtmäßigkeit der Beschränkungsmaßnahme von Bedeutung sein können. In diesem Fall ist die Verarbeitung der Daten einzuschränken; sie dürfen nur zu diesen Zwecken verwendet werden.

(2) Die verbleibenden Daten sind zu kennzeichnen. Nach einer Übermittlung ist die Kennzeichnung durch den Empfänger aufrechtzuerhalten. Die Daten dürfen nur zu den in § 1 Abs. 1 Nr. 1 und den in Absatz 4 genannten Zwecken verwendet werden.

§ 19 Übermittlung personenbezogener Daten durch das Bundesamt für Verfassungsschutz

(1) Das Bundesamt für Verfassungsschutz darf personenbezogene Daten, die mit den Mitteln nach § 8 Absatz 2 erhoben worden sind, an die Staatsanwaltschaften, die Finanzbehörden nach § 386 Absatz 1 der Abgabenordnung, die Polizeien, die mit der Steuerfahndung betrauten Dienststellen der Landesfinanzbehörden, die Behörden des Zollfahndungsdienstes sowie andere Zolldienststellen, soweit diese Aufgaben nach dem Bundespolizeigesetz wahrnehmen, übermitteln, soweit dies erforderlich ist zur

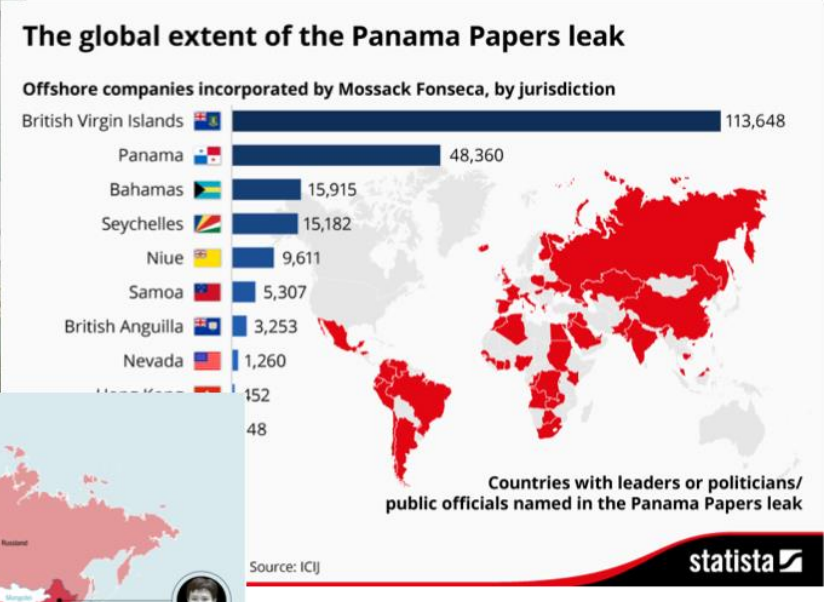
1. Erfüllung eigener Aufgaben der Informationsgewinnung (§ 8 Absatz 1 Satz 2 und 3),
2. Abwehr einer im Einzelfall bestehenden Gefahr für den Bestand oder die Sicherheit des Bundes oder eines Landes oder für Leib, Leben, Gesundheit oder Freiheit einer Person oder für Sachen von erheblichem Wert, deren Erhaltung im öffentlichen Interesse geboten ist,
3. Verhinderung oder sonstigen Verhütung von Straftaten von erheblicher Bedeutung oder
4. Verfolgung von Straftaten von erheblicher Bedeutung;

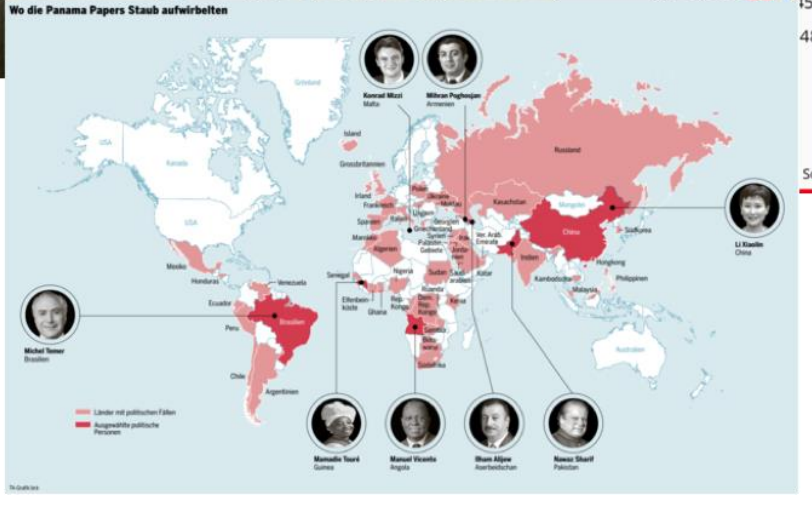
Network of related pieces of information in documents

Pro

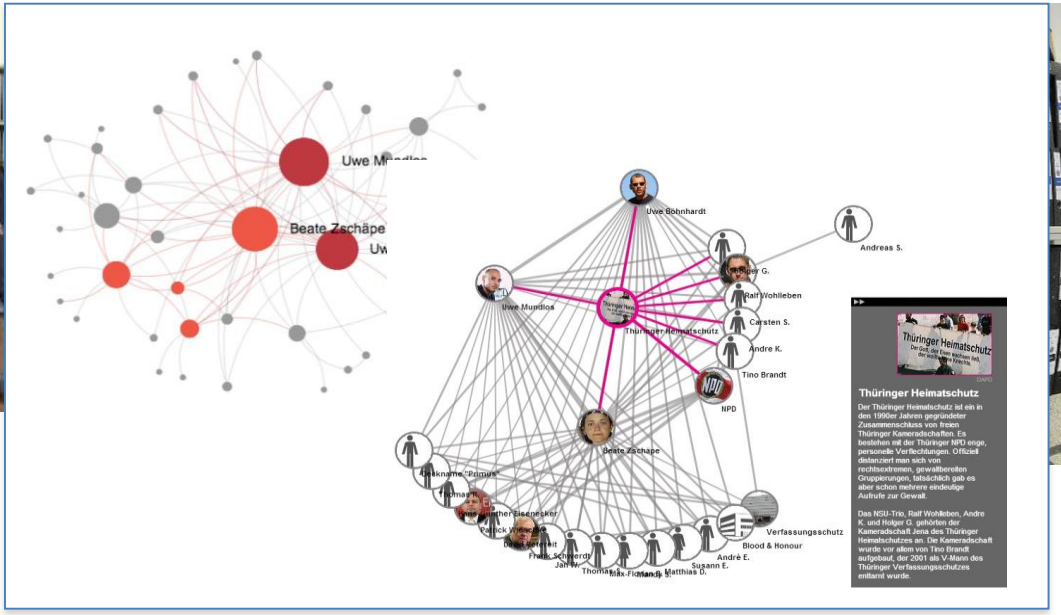
There is little sequential to getting the full picture.

Panama Papers





National Socialist Underground Trial



- over 650 folders
- close to half a million pages (2014), many additions
- 540 witnesses
- 248 admissions of evidence
- ...

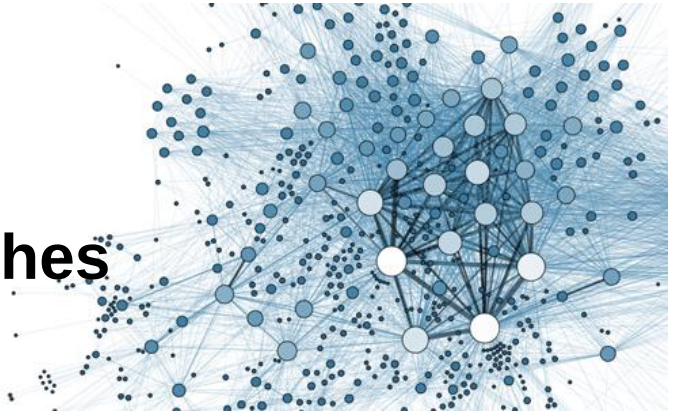
Information Networks

Hypothesis: named entities and concepts that (frequently) occur together in documents have some relationship.

Approach:

- Extract named entities and concepts (nodes)
- Frequent co-occurrence indicates relationship

Allows for several information detection and exploration approaches



Summary and outlook

- Text Analytics methods and techniques are key to almost all legal tech applications.
- Amount of text data will significantly increase (document automation!).
- Information discovery and exploration will dramatically increase in complexity.
- Many law firms are “sitting on” very valuable text data that could be exploited to improve legal businesses.

Thank you for your attention!

Questions?