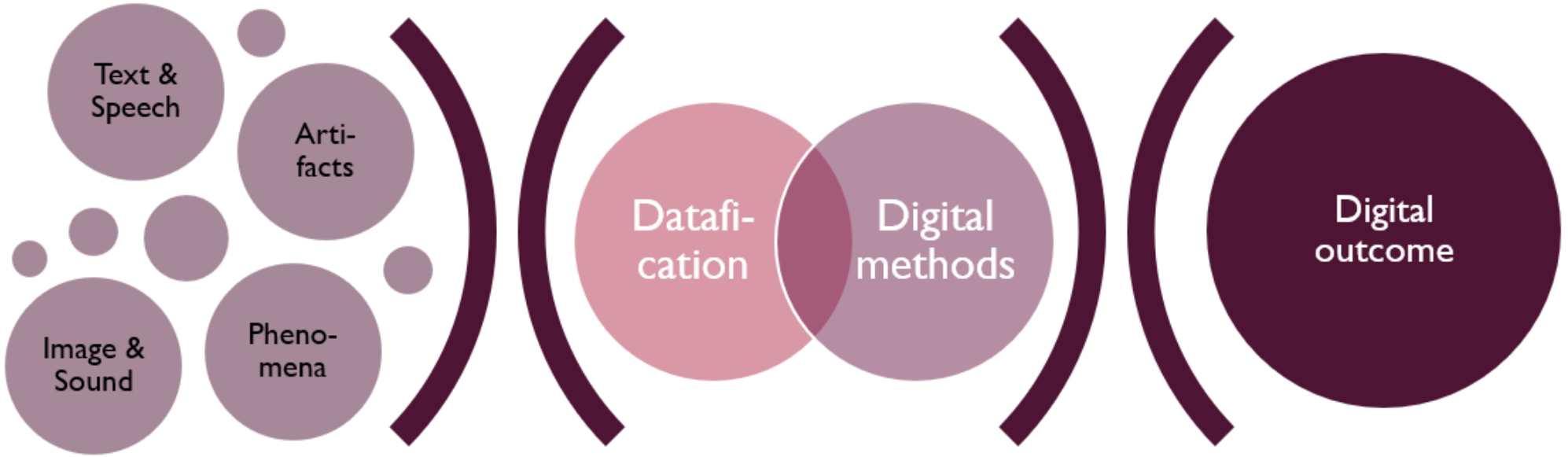
TEXT & DATA MINING
A FIRST VIEW

DOI: 10.5281/zenodo.5592495

KATHI WOITAS
DIGITAL SCHOLARSHIP SERVICES
23.09.2021
KATHI.WOITAS@UNIBE.CH

KATHI WOITAS

- Digital Scholarship Services @ University Library Bern
- Library Science + Cultural Anthropology (MA)
- Advanced training in Data Science

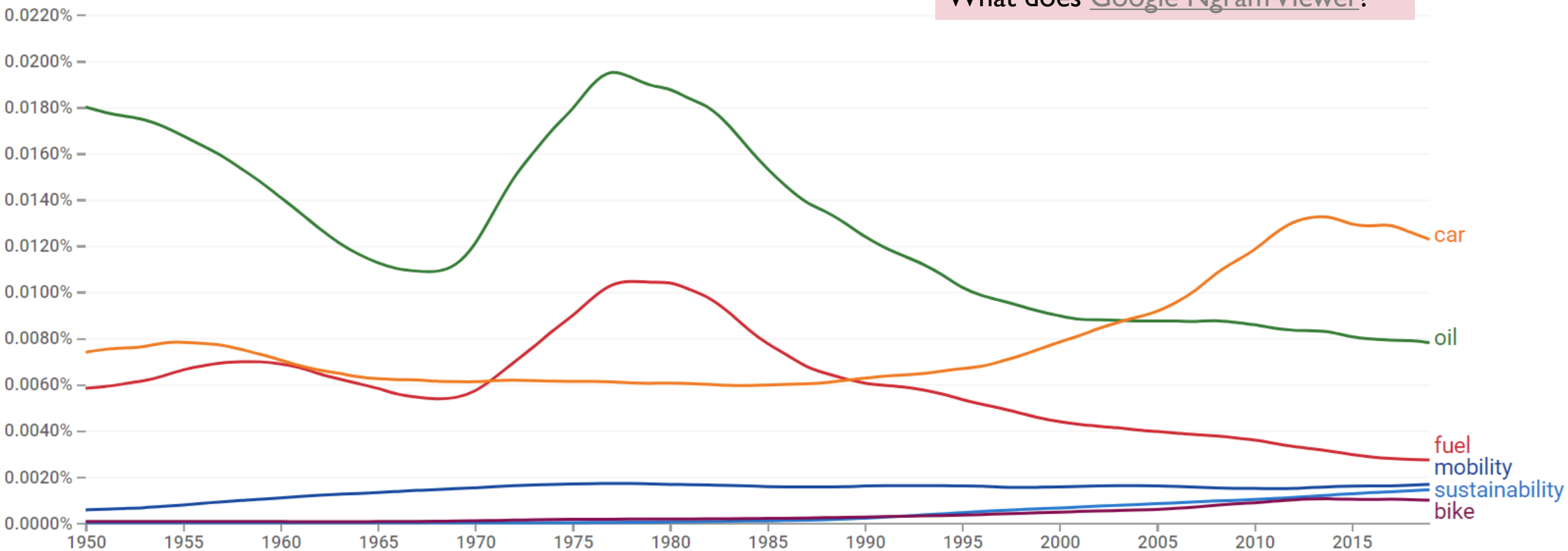


Google Books Ngram Viewer

sustainability,fuel,oil,car,mobility,bike

1950 - 2019 | English (2019) | Case-Insensitive | Smoothing

What does Google Ngram Viewer?



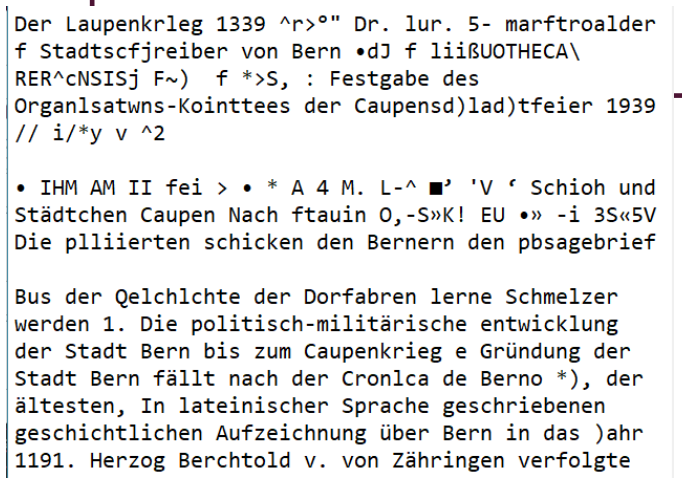
TEXT & DATA

Unstructured

Image/Sound

OCRized

Text



```
Der Laupenkrleg 1339 ^r>°" Dr. lur. 5- marftroalder
f Stadtscfjreiber von Bern •dJ f liißUOTHECA\
RER^cNSISj F~)  f *>S, : Festgabe des
Organlsatwns-Kointtees der Caupensd)lad)tfeier 1939
// i/*y v ^2

• IHM AM II fei > • * A 4 M. L-^ ■’ 'V ‘ Schioh und
Städtchen Caupen Nach ftauin O,-S»K! EU •» -i 3S«5V
Die plliierten schicken den Bernern den pbsagebrief

Bus der Qelchlchte der Dorfabren lerne Schmelzer
werden 1. Die politisch-militärische entwicklung
der Stadt Bern bis zum Caupenkrieg e Gründung der
Stadt Bern fällt nach der Cronlca de Berno *), der
ältesten, In lateinischer Sprache geschriebenen
geschichtlichen Aufzeichnung über Bern in das )ahr
1191. Herzog Berchtold v. von Zähringen verfolgte
```

Semistructured

Mark up

(e.g. xml, json)

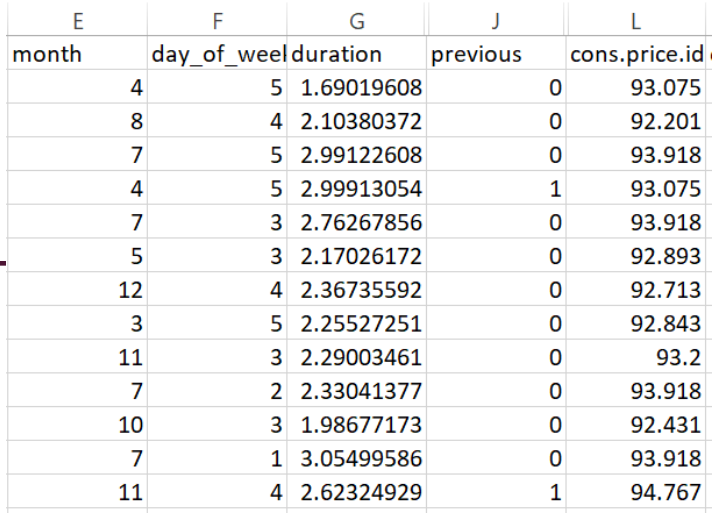
Graphs

```
diese fallend"], "publisher": "[Verlag nicht
ermittelbar]", "date": ["1670", "1720"],
"type": ["Text", "Book"], "format": "16
ungez\u00c3\u00a4hlte Seiten ; 16 cm
(8\u00c2\u00b0)", "identifier": ["doi:10.
3931/e-rara-90056", "https://www.e-rara.ch/
bes_1/doi/10.3931/e-rara-90056",
```

Structured

Tabular

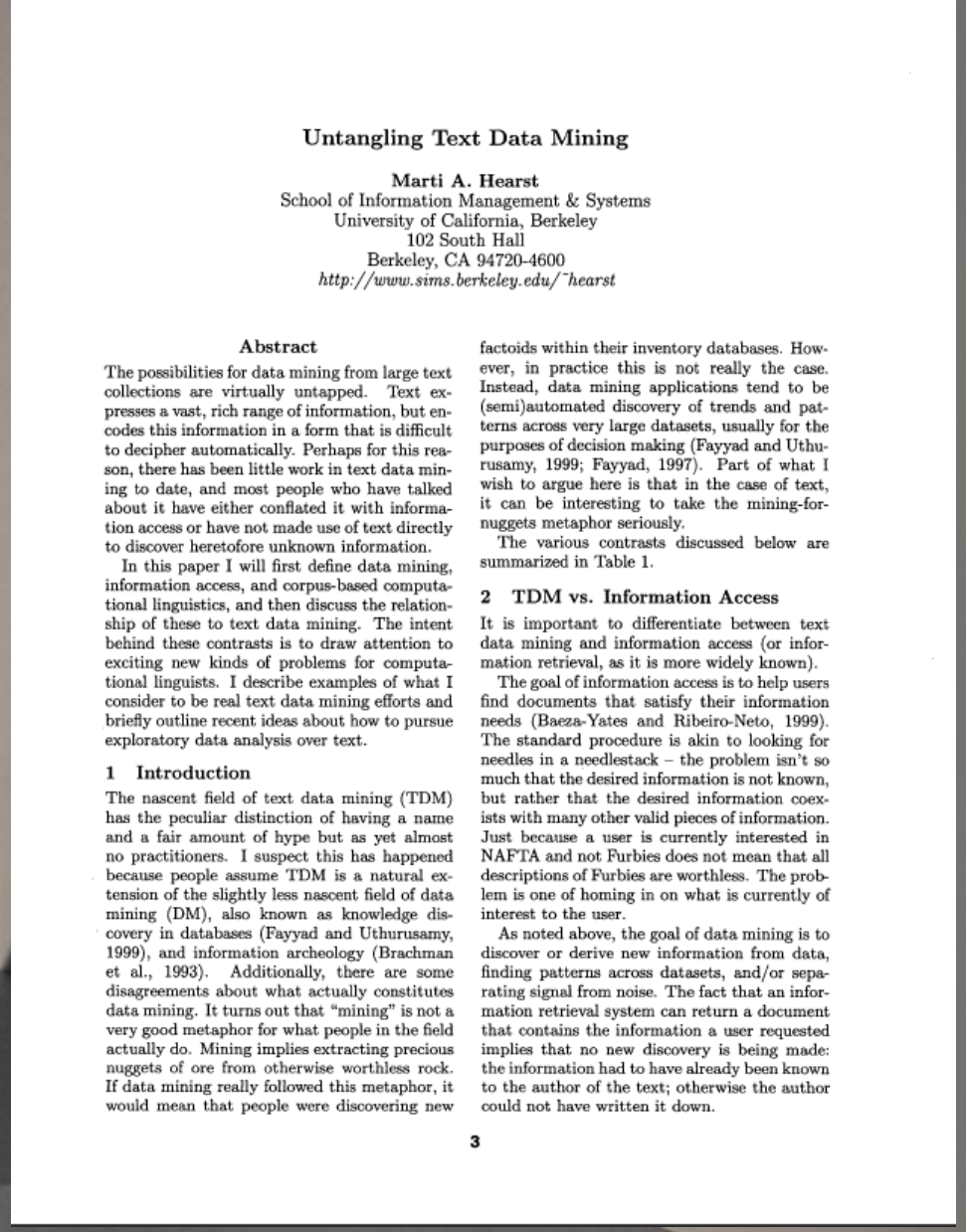
Numeric



E	F	G	J	L
month	day_of_weel	duration	previous	cons.price.id
4	5	1.69019608	0	93.075
8	4	2.10380372	0	92.201
7	5	2.99122608	0	93.918
4	5	2.99913054	1	93.075
7	3	2.76267856	0	93.918
5	3	2.17026172	0	92.893
12	4	2.36735592	0	92.713
3	5	2.25527251	0	92.843
11	3	2.29003461	0	93.2
7	2	2.33041377	0	93.918
10	3	1.98677173	0	92.431
7	1	3.05499586	0	93.918
11	4	2.62324929	1	94.767

DATA MINING

"THE GOAL OF DATA MINING IS TO DISCOVER OR DERIVE NEW INFORMATION FROM DATA, FINDING PATTERNS ACROSS DATASETS, AND/OR SEPARATING SIGNAL FROM NOISE."



Untangling Text Data Mining

Marti A. Hearst
School of Information Management & Systems
University of California, Berkeley
102 South Hall
Berkeley, CA 94720-4600
http://www.sims.berkeley.edu/~hearst

Abstract

The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically. Perhaps for this reason, there has been little work in text data mining to date, and most people who have talked about it have either conflated it with information access or have not made use of text directly to discover heretofore unknown information.

In this paper I will first define data mining, information access, and corpus-based computational linguistics, and then discuss the relationship of these to text data mining. The intent behind these contrasts is to draw attention to exciting new kinds of problems for computational linguists. I describe examples of what I consider to be real text data mining efforts and briefly outline recent ideas about how to pursue exploratory data analysis over text.

1 Introduction

The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM), also known as knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and information archeology (Brachman et al., 1993). Additionally, there are some disagreements about what actually constitutes data mining. It turns out that "mining" is not a very good metaphor for what people in the field actually do. Mining implies extracting precious nuggets of ore from otherwise worthless rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets, usually for the purposes of decision making (Fayyad and Uthurusamy, 1999; Fayyad, 1997). Part of what I wish to argue here is that in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously.

The various contrasts discussed below are summarized in Table 1.

2 TDM vs. Information Access

It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known).

The goal of information access is to help users find documents that satisfy their information needs (Baeza-Yates and Ribeiro-Neto, 1999). The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. Just because a user is currently interested in NAFTA and not Furbies does not mean that all descriptions of Furbies are worthless. The problem is one of homing in on what is currently of interest to the user.

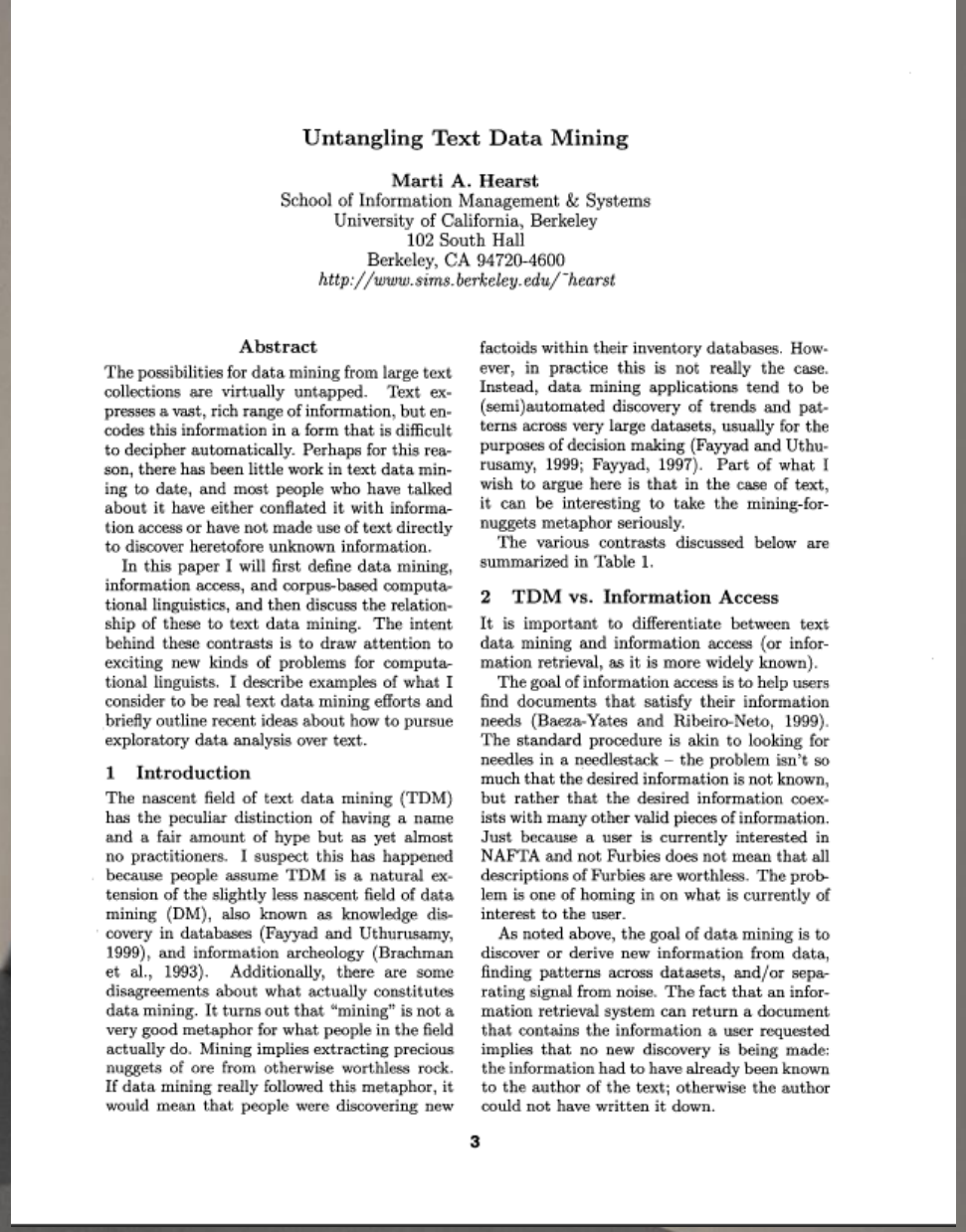
As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested implies that no new discovery is being made: the information had to have already been known to the author of the text; otherwise the author could not have written it down.

3

Hearst, M.A. Untangling text data mining. in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* 3–10 (Association for Computational Linguistics, 1999). doi:10.3115/1034678.1034679.

TEXT MINING

"IF WE EXTRAPOLATE FROM DATA MINING … ON NUMERICAL DATA TO DATA MINING FROM TEXT COLLECTIONS, WE DISCOVER THAT THERE ALREADY EXISTS A FIELD ENGAGED IN TEXT DATA MINING: COMPUTATIONAL LINGUISTICS!"



Untangling Text Data Mining

Marti A. Hearst
School of Information Management & Systems
University of California, Berkeley
102 South Hall
Berkeley, CA 94720-4600
http://www.sims.berkeley.edu/~hearst

Abstract

The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically. Perhaps for this reason, there has been little work in text data mining to date, and most people who have talked about it have either conflated it with information access or have not made use of text directly to discover heretofore unknown information.

In this paper I will first define data mining, information access, and corpus-based computational linguistics, and then discuss the relationship of these to text data mining. The intent behind these contrasts is to draw attention to exciting new kinds of problems for computational linguists. I describe examples of what I consider to be real text data mining efforts and briefly outline recent ideas about how to pursue exploratory data analysis over text.

1 Introduction

The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM), also known as knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and information archeology (Brachman et al., 1993). Additionally, there are some disagreements about what actually constitutes data mining. It turns out that "mining" is not a very good metaphor for what people in the field actually do. Mining implies extracting precious nuggets of ore from otherwise worthless rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets, usually for the purposes of decision making (Fayyad and Uthurusamy, 1999; Fayyad, 1997). Part of what I wish to argue here is that in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously.

The various contrasts discussed below are summarized in Table 1.

2 TDM vs. Information Access

It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known).

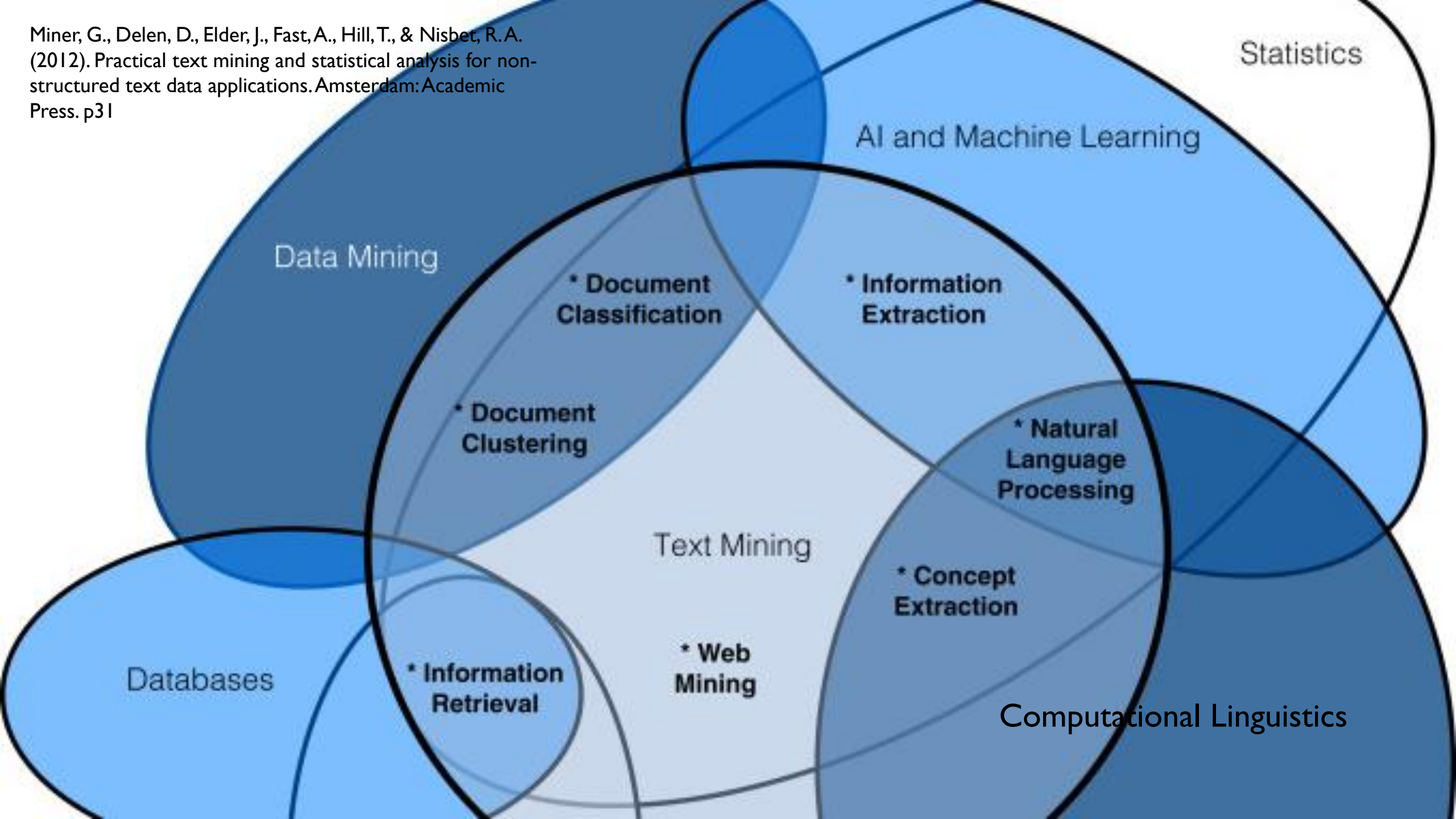
The goal of information access is to help users find documents that satisfy their information needs (Baeza-Yates and Ribeiro-Neto, 1999). The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. Just because a user is currently interested in NAFTA and not Furbies does not mean that all descriptions of Furbies are worthless. The problem is one of homing in on what is currently of interest to the user.

As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested implies that no new discovery is being made: the information had to have already been known to the author of the text; otherwise the author could not have written it down.

3

Hearst, M.A. Untangling text data mining. in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* 3–10 (Association for Computational Linguistics, 1999). doi:10.3115/1034678.1034679.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. A. (2012). Practical text mining and statistical analysis for non-structured text data applications. Amsterdam: Academic Press. p31



Classical linguistic approach	Statistical approach
Analyze via syntactic, semantic, … interpretation	Analyze via a large but limited set of recurring patterns
Create a rule-based language model • Mimic human language processing • Model contains words, grammar (syntax) and meaning (semantics) as layers with increasing complexity	**Learn language patterns from a vast number of documents** • Language = repetition of patterns • Occurence of patterns trigger the *most likely* rule/action

MACHINE LEARNING IN A NUTSHELL

"The goal of data mining is to **discover or derive new information** from data,

finding patterns across datasets, and/or separating signal from noise."

DATA – A CONTINUUM

Unstructured

Image/Sound

OCRized

Text

```
Der Laupenkrleg 1339 ^r>°" Dr. lur. 5- marftroalder
f Stadtscfjreiber von Bern •dJ f liißUOTHECA\
RER^cNSISj F~)  f *>S, : Festgabe des
Organlsatwns-Kointtees der Caupensd)lad)tfeier 1939
// i/*y v ^2

• IHM AM II fei > • * A 4 M. L-^ ■’ 'V ‘ Schioh und
Städtchen Caupen Nach ftauin O,-S»K! EU •» -i 3S«5V
Die plliierten schicken den Bernern den pbsagebrief

Bus der Qelchlchte der Dorfabren lerne Schmelzer
werden 1. Die politisch-militärische entwicklung
der Stadt Bern bis zum Caupenkrieg e Gründung der
Stadt Bern fällt nach der Cronlca de Berno *), der
ältesten, In lateinischer Sprache geschriebenen
geschichtlichen Aufzeichnung über Bern in das )ahr
1191. Herzog Berchtold v. von Zähringen verfolgte
```

Semistructured

Mark up

(e.g. xml, json)

Graphs


```
diese fallend"], "publisher": "[Verlag nicht
ermittelbar]", "date": ["1670", "1720"],
"type": ["Text", "Book"], "format": "16
ungez\u00c3\u00a4hlte Seiten ; 16 cm
(8\u00c2\u00b0)", "identifier": ["doi:10.
3931/e-rara-90056", "https://www.e-rara.ch/
bes_1/doi/10.3931/e-rara-90056",
```


Structured

Tabular

Numeric

E	F	G	J	L
month	day_of_weel	duration	previous	cons.price.id
4	5	1.69019608	0	93.075
8	4	2.10380372	0	92.201
7	5	2.99122608	0	93.918
4	5	2.99913054	1	93.075
7	3	2.76267856	0	93.918
5	3	2.17026172	0	92.893
12	4	2.36735592	0	92.713
3	5	2.25527251	0	92.843
11	3	2.29003461	0	93.2
7	2	2.33041377	0	93.918
10	3	1.98677173	0	92.431
7	1	3.05499586	0	93.918
11	4	2.62324929	1	94.767

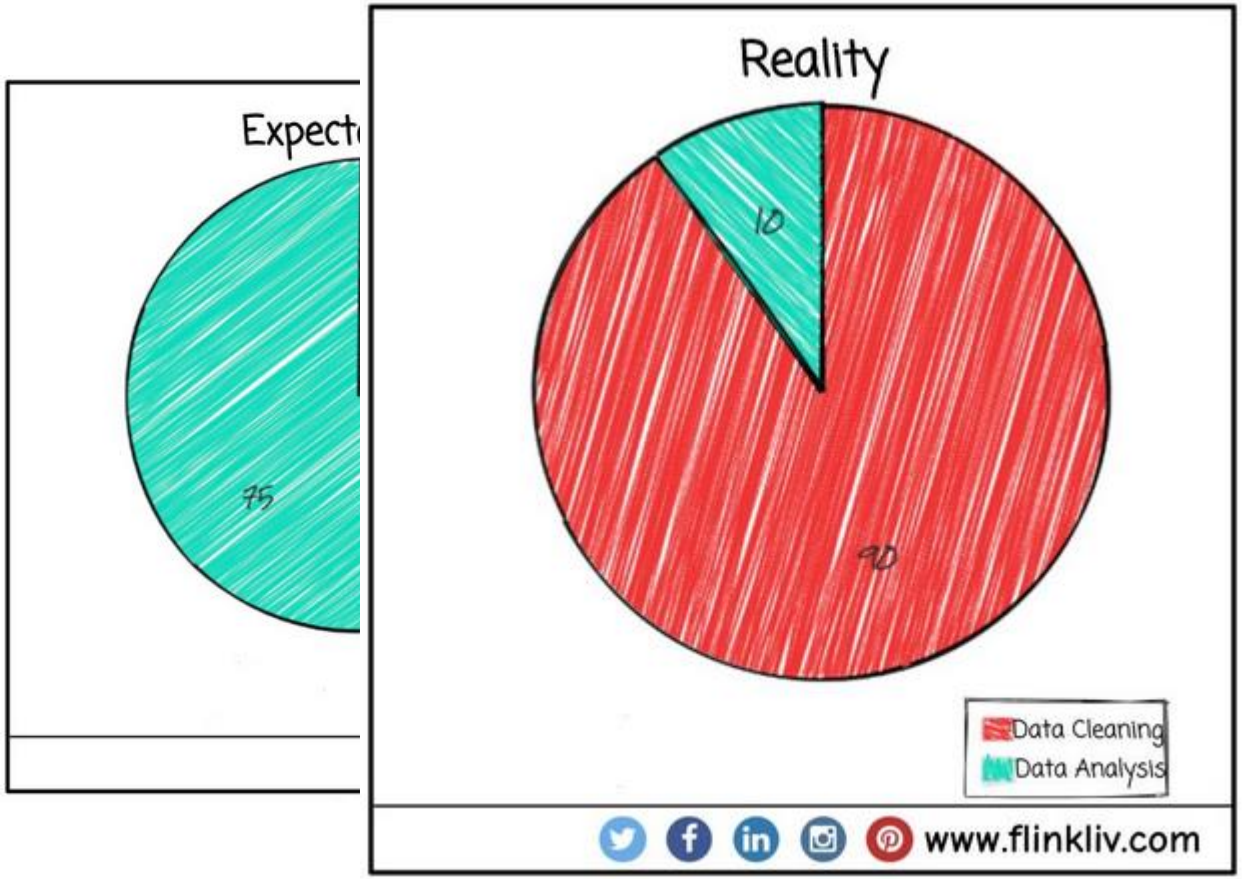
TDM WORKFLOW IN A NUTSHELL

1 Retrieval & Access

2 (Pre)Processing

3 Extraction

4 Analysis



TDM WORKFLOW: 1 RETRIEVAL & ACCESS

Find suitable resources

- Often primary resources (digital collections/archives, free or licensed)
- Look for TDM rights and download/scraping regulations!
- If in doubt or facing problems: Get in touch with UB Bern.

Get the raw data

- Access bulk downloads or data APIs, avoid website scraping. May take a while!
- Adapting file structure, merge files etc.

TDM WORKFLOW: 2 (PRE)PROCESSING – BUILDING A CORPUS

OCR and cleaning

- Read out or recognize text, e.g. from PDF
- Remove noise, e.g. text file header, mark up, metadata

Basic text processing: Normalization

- Tokenization: Split text into sentences and words
- Lowercase all words (language specific)
- Remove punctuation and/or stopwords (= frequent words with low semantics like 'a' or 'on')
- Normalize spelling variants, abbreviations
- Stemming (= cut all words to their stems)

The chef cooks the meat and the sous-chefs cook the soups.

The, chef, cooks, the, meat, and, the, sous-chefs, cook, the, soups

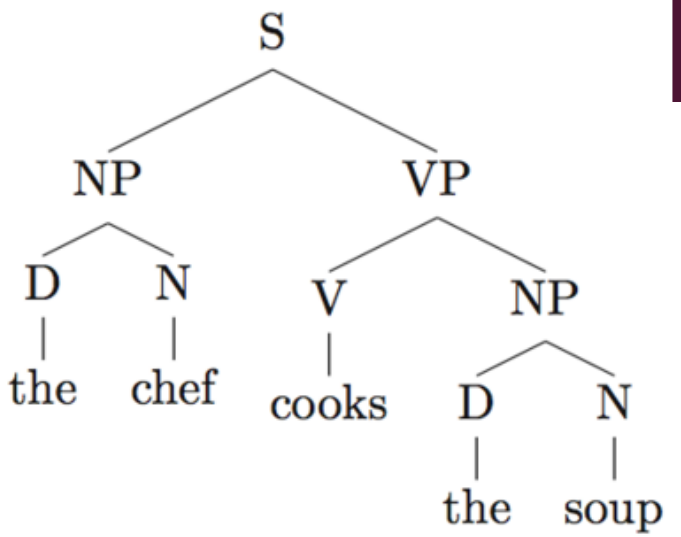
chef	1
cooks	1
cook	1
meat	1
soups	1
sous-chefs	1

chef	1
cook	2
meat	1
soup	1
sous-chef	1

TDM WORKFLOW: 2 (PRE)PROCESSING

Deeper text processing:

Morphological + syntactical analysis



- Get parts of speech (POS) of tokens
- Get syntactic features of tokens (parsing tree)
- Lemmatization (= cut to their grammatical base)

chef	1
cooks	1
cook	1
meat	1
soups	1
sous-chefs	1

chef	1
cook	2
meat	1
soup	1
sous-chef	1

TDM WORKFLOW: 3 EXTRACTION

Statistical, exploratory analysis

- e.g. word frequencies over time, co-occurrences of words
- e.g. Term frequency – Inverse document frequency (TF-IDF)

Feature engineering

- Select certain statistical or linguistic features
- Transform tokens to numerical features according to different methods

TDM WORKFLOW: 4 ANALYSIS

Unsupervised learning methods

- Find unknown patterns in documents

Use cases:

- *Document clustering*: Find groups of similar documents
- *Topic modeling*: Find groups of similar documents and get their common 'topics'

TDM WORKFLOW: 4 ANALYSIS

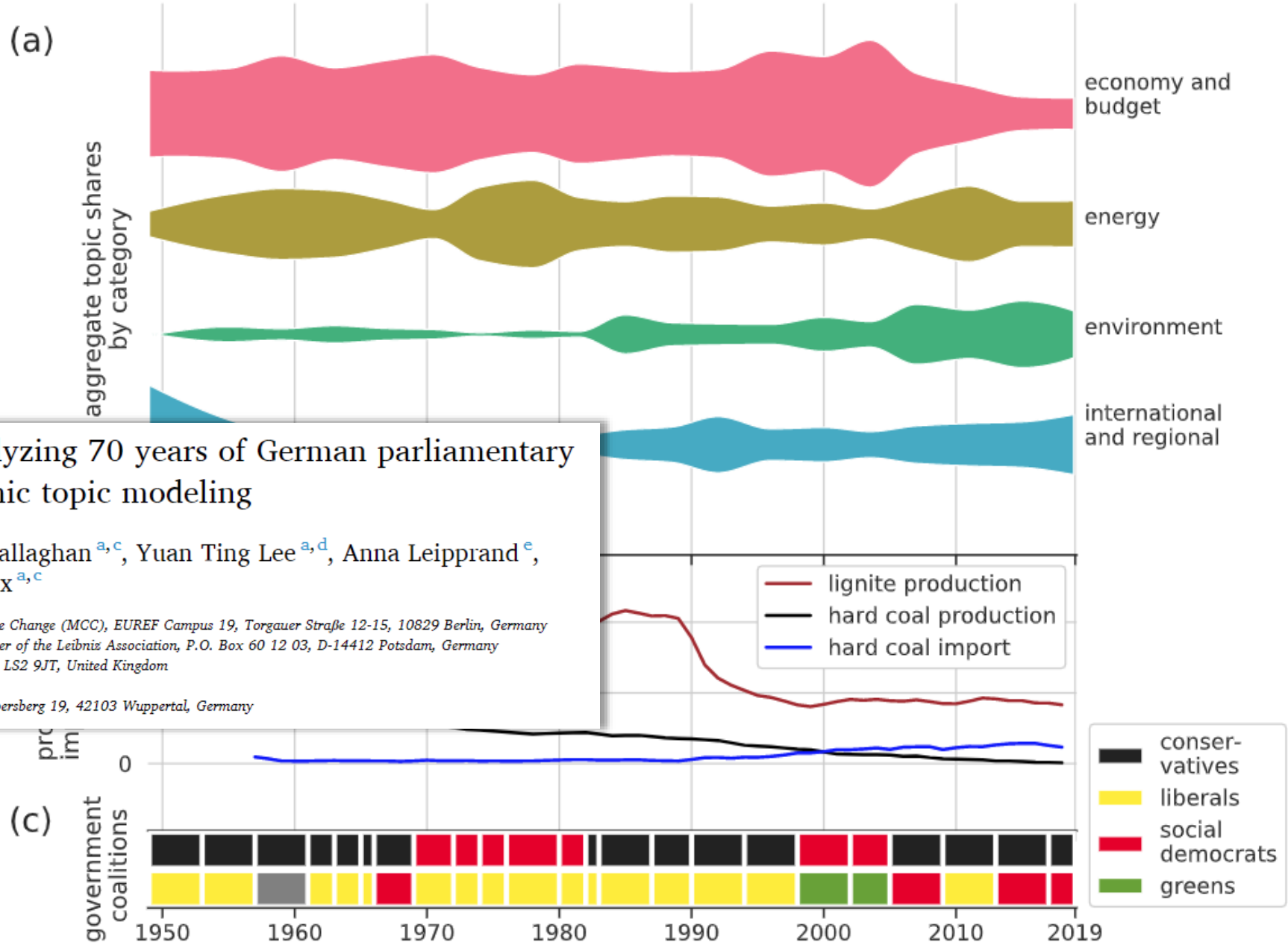
Supervised learning methods

- Find the mathematical connection between certain data values to predict outcomes of similar new data

Use Cases:

- *Sentiment analysis*: Recognize the sentiment/mood of a text (think of reviews…)
- *Text classification*: Classify new texts into known groups
- *Named Entity Recognition*: Recognize named entities like places, persons etc.

Start with reading?



Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling

Finn Müller-Hansen [a,b,*], Max W. Callaghan [a,c], Yuan Ting Lee [a,d], Anna Leipprand [e], Christian Flachsland [a,d], Jan C. Minx [a,c]

[a] Mercator Research Institute on Global Commons and Climate Change (MCC), EUREF Campus 19, Torgauer Straße 12-15, 10829 Berlin, Germany
[b] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany
[c] School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom
[d] Hertie School, Friedrichstraße 180, 10117 Berlin, Germany
[e] Wuppertal Institut für Klima, Umwelt, Energie gGmbH, Döppersberg 19, 42103 Wuppertal, Germany

Energy Research & Social Science **2021**, *72*, 101869; https://doi.org/10.1016/j.erss.2020.101869

TDM HINTS & SUPPORT

Data

- Turn to UB Bern before start scraping a database – a university-wide block may occur!
- Ask UB Bern for specific data (and digitization) demand
- Resources for TDM website (english version: work in progress)
- Vendor's TDM platforms on the rise (e.g. Nexis Data Lab with worldwide news data)

Infrastructure

- No special infrastructure needed to start with TDM (see Methods & Tools next slide)
- (for bigger data: UNIBE's High Performance Cluster UBELIX)

TDM HINTS & SUPPORT

Methods & Tools

- Wide landscape, open and proprietary tools…
- Rely on software (e.g. Voyant, Weblicht, Text analysis in SPSS)
- Start with coding (R, Python/Jupyter) – greater possibilities, better reproducibility
- Use Jupyter with Python or R without installing local: noto.epfl.ch
- Python introductory courses at Science IT Support
- UB Bern's Digital Toolbox: Jupyter Notebook Tutorials for data work (e.g. OCR, NLP)

GO FURTHER…

Tutorials & Materials

- Online-Tutorial: OpenMinTeD Introduction to TDM
- GESIS materials (videos, tutorials) Capacity Building in Computational Social Science
- GESIS current series Computational Social Science and Digital Behavioral Data
- ePol Text Mining Verfahren (eTMV) (short example studies, in German)
- Constellate Tutorials & Jupyter Notebooks for Python and TDM
- …and a myriad of other material on the web…

GO FURTHER…

Methodology

- Ignatow, G. & Mihalcea, R. *Text Mining: A Guidebook for the Social Sciences*. (Sage, 2017). doi:10.4135/9781483399782.
- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F. & Lane, J. *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. (Chapman and Hall/CRC, 2020). doi:10.1201/9780429324383.
- Lemke, M. & Wiedemann, G. Einleitung: Text Mining in den Sozialwissenschaften. in *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (eds. Lemke, M. & Wiedemann, G.) 1–13 (Springer, 2016). doi:10.1007/978-3-658-07224-7_1.
- Wiedemann, G. & Lemke, M. Text Mining für die Analyse qualitativer Daten. in *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (eds. Lemke, M. & Wiedemann, G.) 397–419 (Springer Fachmedien, 2016). doi:10.1007/978-3-658-07224-7_15.
- Manderscheid, K. Text Mining. in *Handbuch Methoden der empirischen Sozialforschung* (eds. Baur, N. & Blasius, J.) 1103–1116 (Springer, 2019). doi:10.1007/978-3-658-21308-4_79.
- Mayerl, Jochen. Bedeutet 'Big Data' das Ende der sozialwissenschaftlichen Methodenforschung? [Essay] in *Soziopolis* (…2015). link

THANK YOU!

KATHI.WOITAS@UNIBE.CH