gesis Leibniz-Institut für Sozialwissenschaften

Introduction to Text Mining

Meet the Experts! – GESIS online talks

Arnim Bleier ▪ November 4, 2021

Leibniz Gemeinschaft

Speaker

Dr. Arnim Bleier

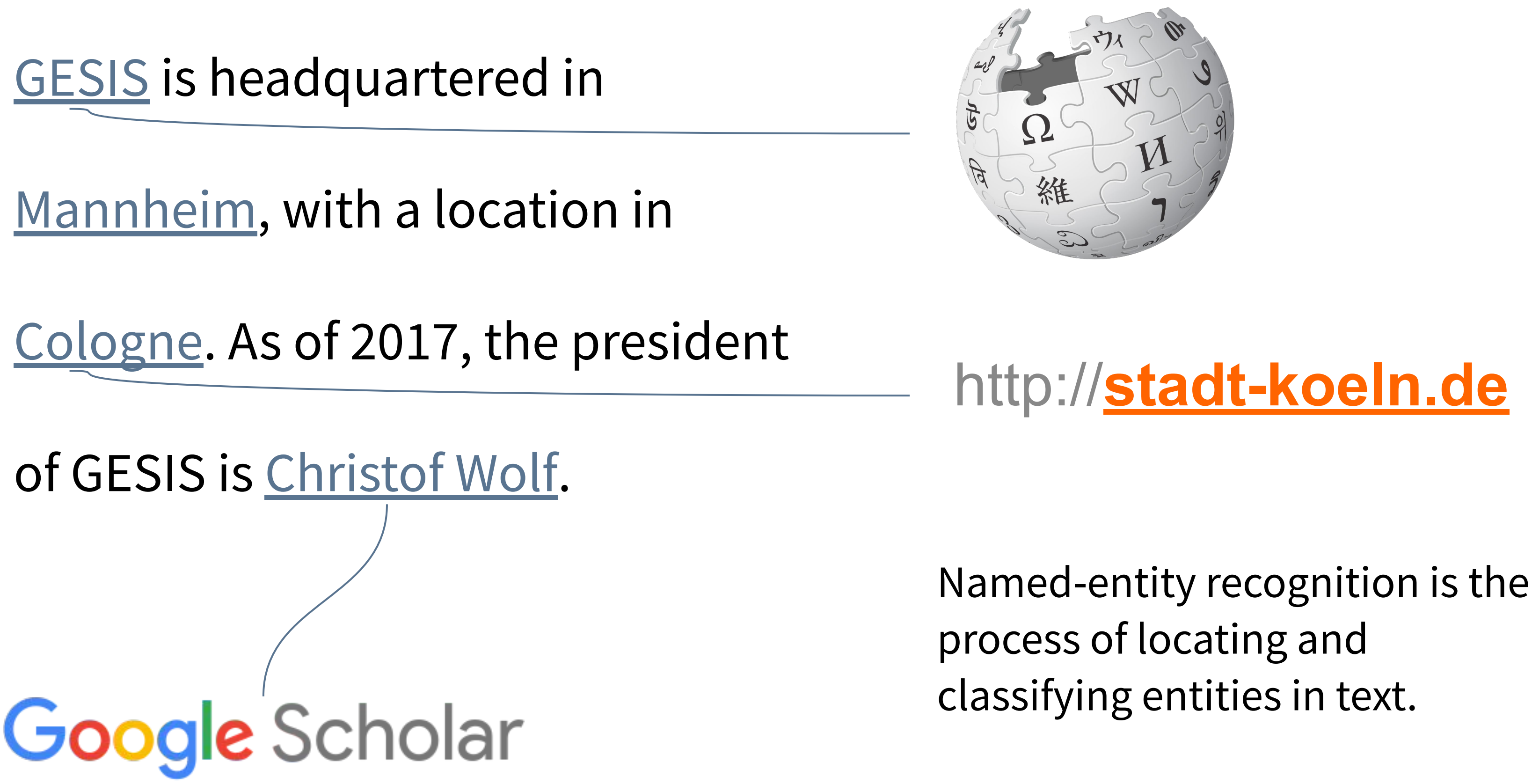
- Senior Researcher in the team Designed Digital Data, Department Computational Social Science
- PhD in statistics
- computational replicability, statistics
- Contact: arnim.bleier@gesis.org

What is Text Mining?

The process of extracting relevant information from text

GESIS Library

Example: Named-entity recognition

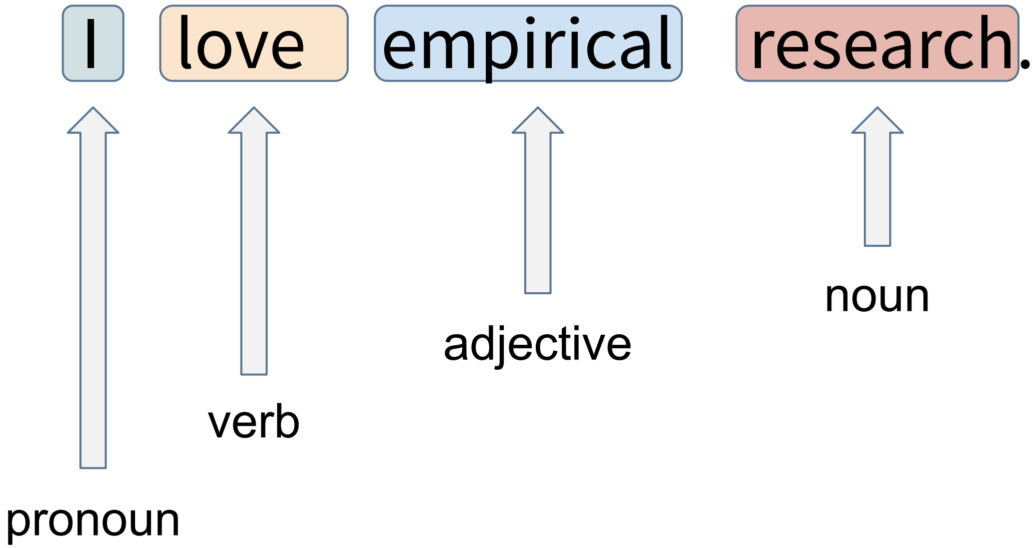


GESIS is headquartered in Mannheim, with a location in Cologne. As of 2017, the president of GESIS is Christof Wolf.

http://stadt-koeln.de

Named-entity recognition is the process of locating and classifying entities in text.

Example: Part-of-speech tagging



I love empirical research.

pronoun

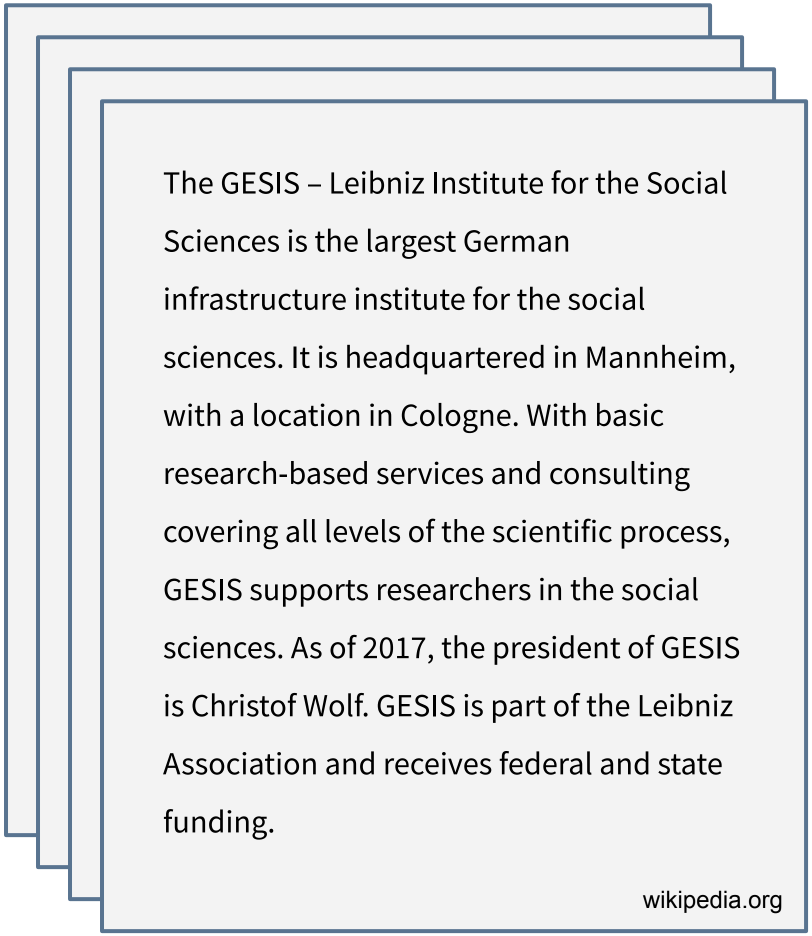
verb

adjective

noun

Part-of-speech tagging is the process of inferring the particular part of speech for a word in a text.

Example: Document classification



The GESIS – Leibniz Institute for the Social Sciences is the largest German infrastructure institute for the social sciences. It is headquartered in Mannheim, with a location in Cologne. With basic research-based services and consulting covering all levels of the scientific process, GESIS supports researchers in the social sciences. As of 2017, the president of GESIS is Christof Wolf. GESIS is part of the Leibniz Association and receives federal and state funding.

wikipedia.org

Labels:

germany

research

infrastructure

Document classification is the process of inferring for a document the membership to one or more groups.

Text Mining (typically) …

- is best with a clear goal
- reuses already existing data
- enables us to work with large datasets
- turns language into numbers
- uses machine learning models
- benefits from validation
- supports summarization and visualization
- is a diverse field of research and comprises more than one technique

What can Text Mining do for us?

Are our views on vaccination polarized?

How far are the positions of parties apart?

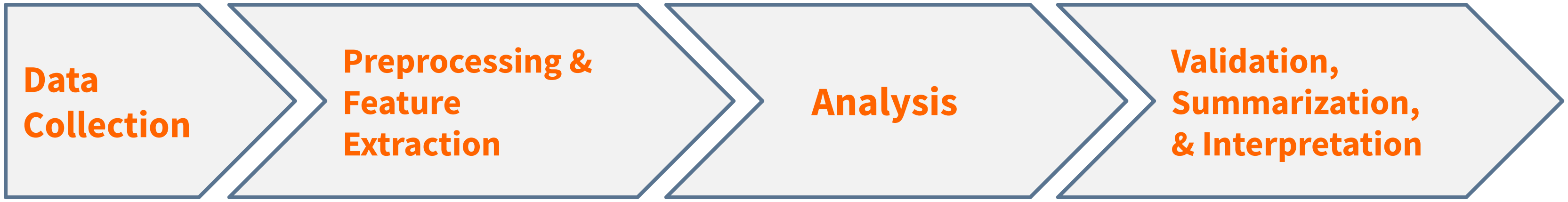
Is Wikipedia sexist and can we measure it?

Maybe …?

Computational analysis of large and suitable text corpora may enable us to answer these questions.

Text Mining Pipeline

Always start with a research question or hypothesis!



Data Collection

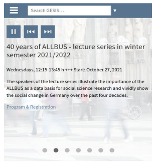
Preprocessing & Feature Extraction

Analysis

Validation, Summarization, & Interpretation

Data Sources

traditional sites



Search GESIS...

40 years of ALLBUS - lecture series in winter semester 2021/2022

Wednesdays, 12:15-13:45 h +++ Start: October 27, 2021

The speakers of the lecture series illustrate the importance of the ALLBUS as a data basis for social science research and vividly show the social change in Germany over the past four decades.

Program & Registration

new media

...

Sites such as Twitter, Reddit, or Wikipedia allow for API-based access. If this is not possible, web scraping becomes an option.

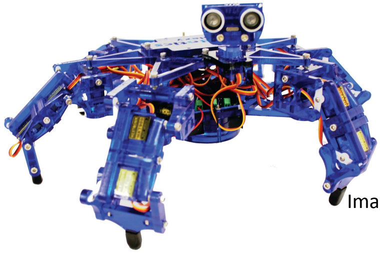
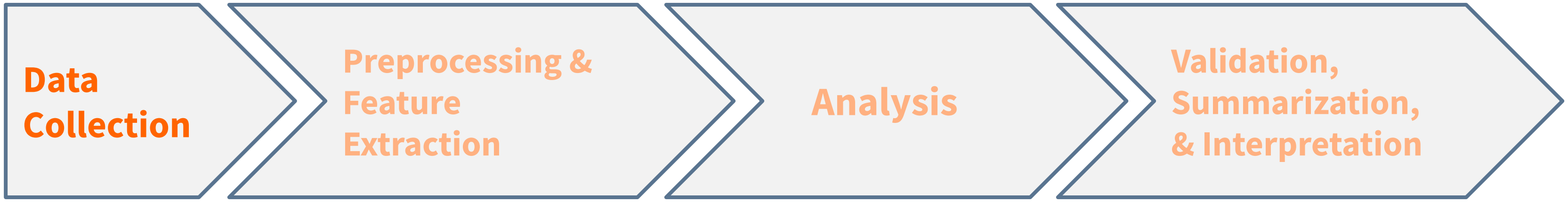
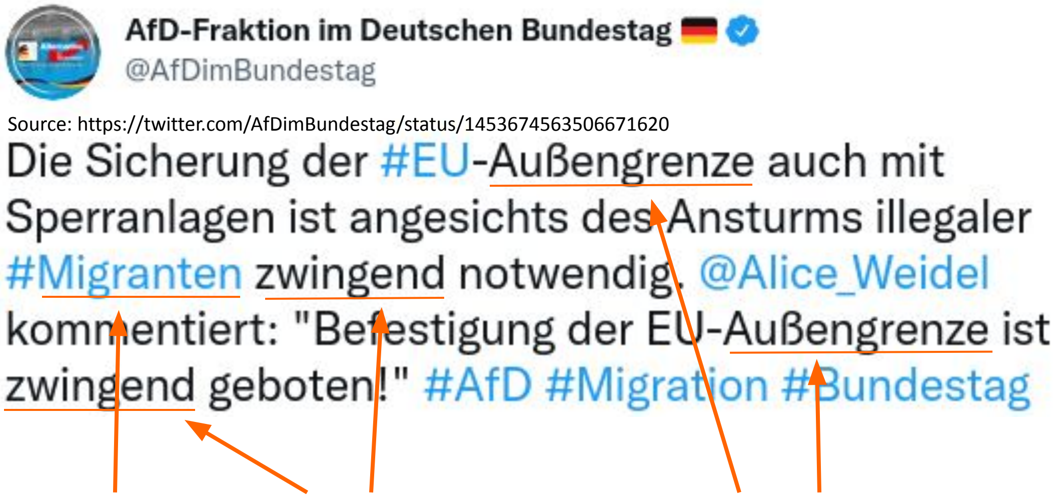


Image sources: https://freesvg.org/newspaper-vector-image
https://www.facebook.com
https://www.reddit.com



Data Collection
Preprocessing & Feature Extraction
Analysis
Validation, Summarization, & Interpretation

Text as Data



AfD-Fraktion im Deutschen Bundestag 🇩🇪
@AfDimBundestag

Source: https://twitter.com/AfDimBundestag/status/1453674563506671620

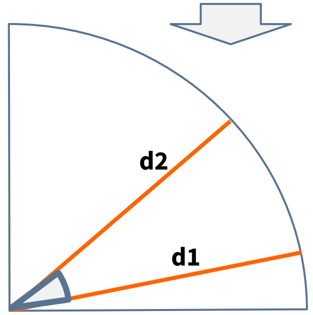
Die Sicherung der #EU-Außengrenze auch mit Sperranlagen ist angesichts des Ansturms illegaler #Migranten zwingend notwendig. @Alice_Weidel kommentiert: "Befestigung der EU-Außengrenze ist zwingend geboten!" #AfD #Migration #Bundestag

migranten	*zwingend*	*inflation*	*außengrenze*	
1	**2**	**0**	**2**	d1
2	**1**	**3**	**0**	d2
1	**0**	**0**	**1**	d3

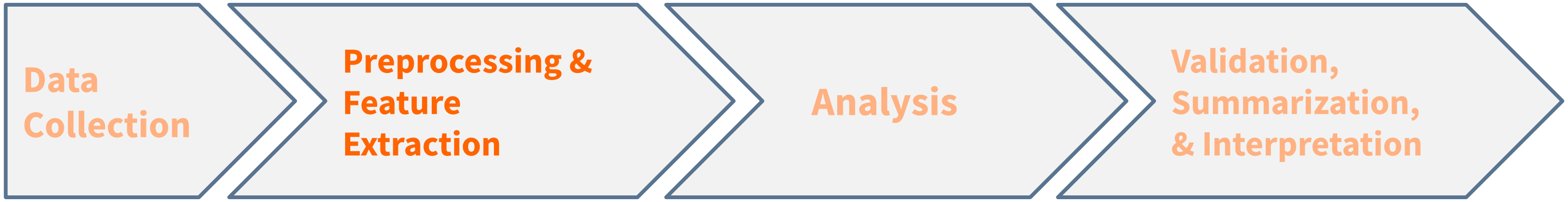
Feature extraction:

Bag-of-words

Vector-space



Document-term matrix



Data Collection

Preprocessing & Feature Extraction

Analysis

Validation, Summarization, & Interpretation

Machine Learning

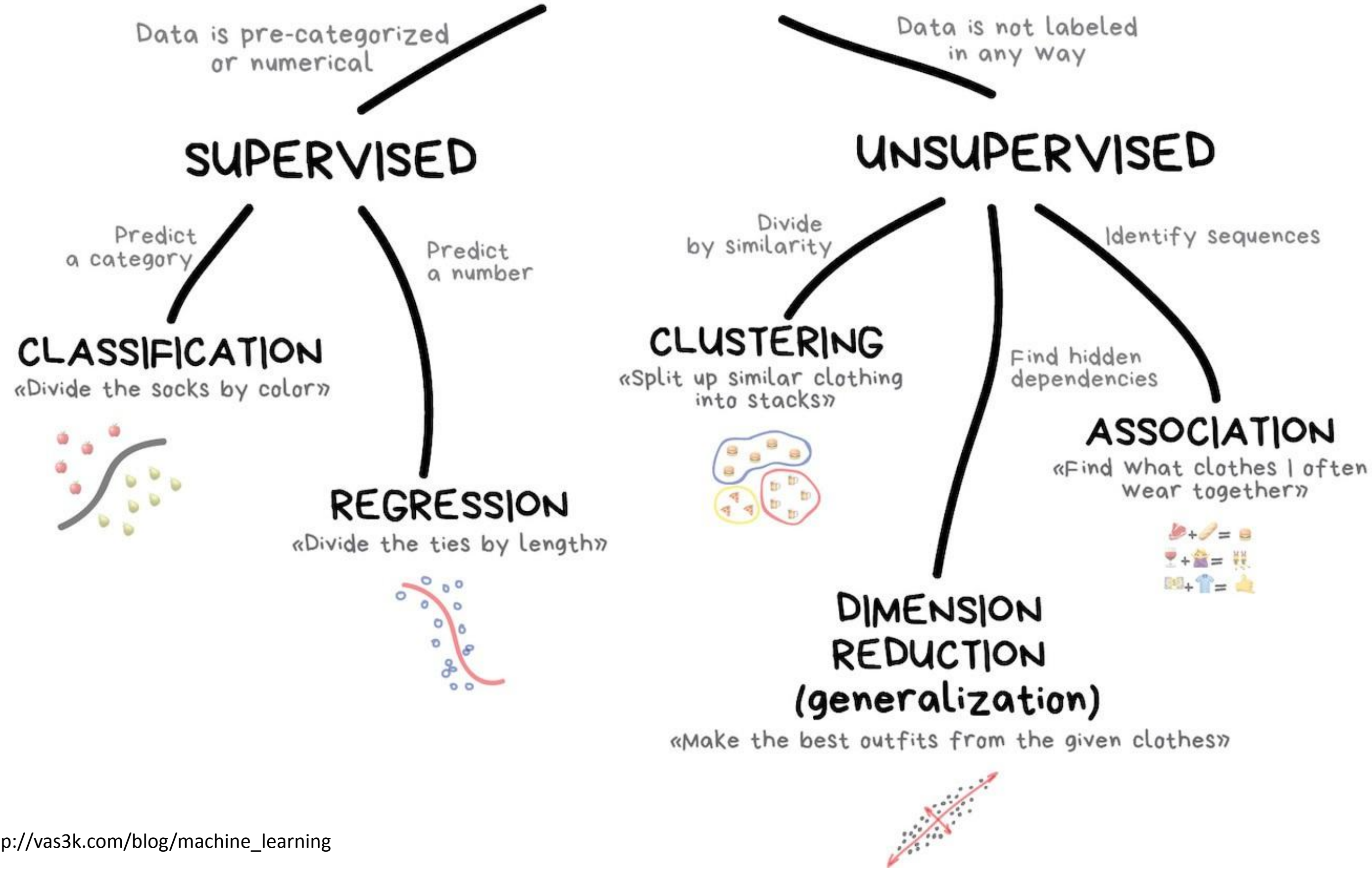
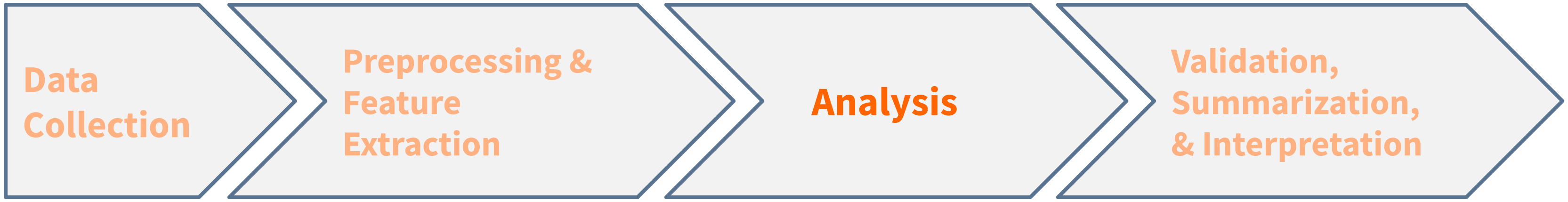
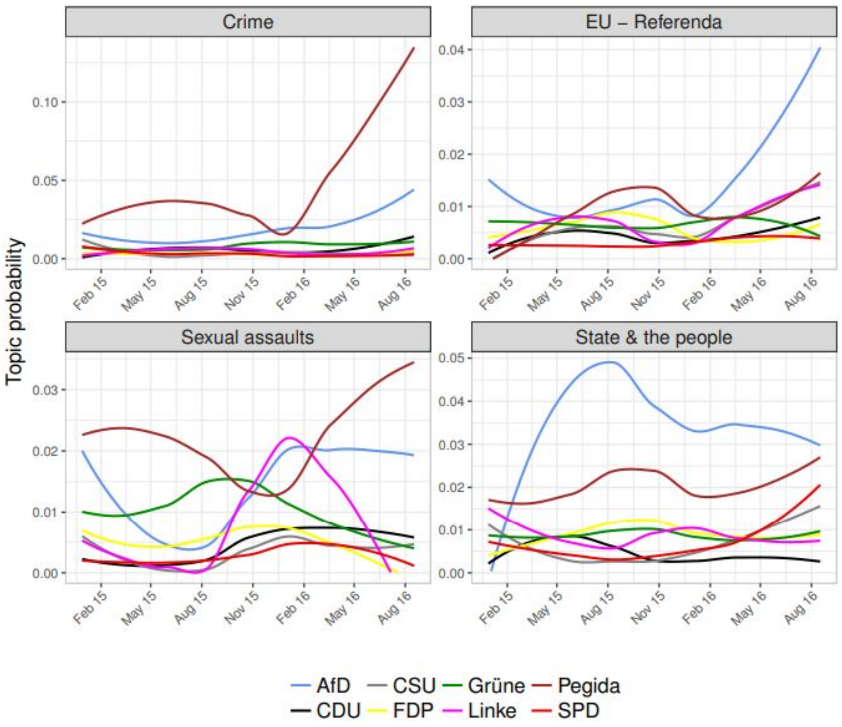
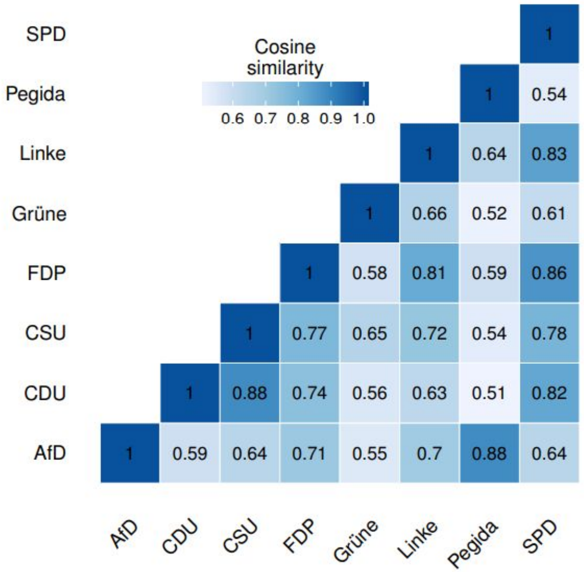


Image source: http://vas3k.com/blog/machine_learning


Data Collection
Preprocessing & Feature Extraction
Analysis
Validation, Summarization, & Interpretation

Summarization and Visualization

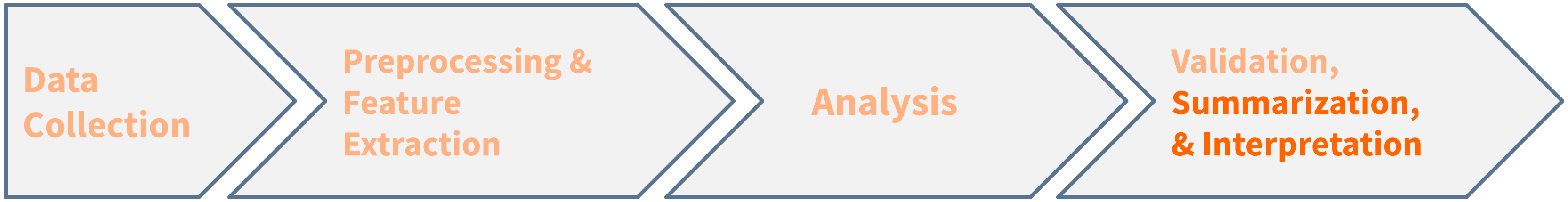




	AfD	CDU	CSU	FDP	Grüne	Linke	Pegida	SPD
SPD								1
Pegida							1	0.54
Linke						1	0.64	0.83
Grüne					1	0.66	0.52	0.61
FDP				1	0.58	0.81	0.59	0.86
CSU			1	0.77	0.65	0.72	0.54	0.78
CDU		1	0.88	0.74	0.56	0.63	0.51	0.82
AfD	1	0.59	0.64	0.71	0.55	0.7	0.88	0.64

Cosine similarities between topic distributions of Pegida and political parties.

Sebastian Stier, Lisa Posch, Arnim Bleier, Markus Strohmaier 2017. When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties



Data Collection

Preprocessing & Feature Extraction

Analysis

Validation, **Summarization, & Interpretation**

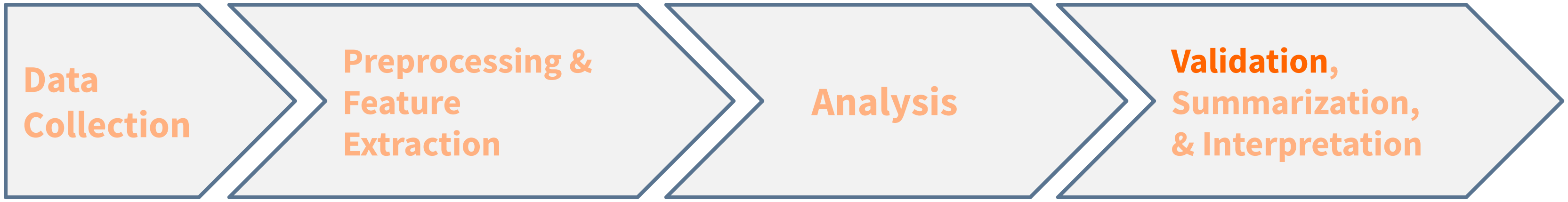
A word of caution

Text data from social media has been used to infer expression of political support, the onset of depression , or signs immanent stock market movements. “If true, this would make the microblogging service the most universally applicable concoction since the discovery of snake oil.”[1]

We have to ask ourselves:

- Is the data suitable to answer our research question?
- Have the right features been extracted?
- Are we measuring what we intend to measure?
- Are our conclusions the only one that are supported by the data?

1) Jungherr, Andreas 2018. Normalizing digital trace data

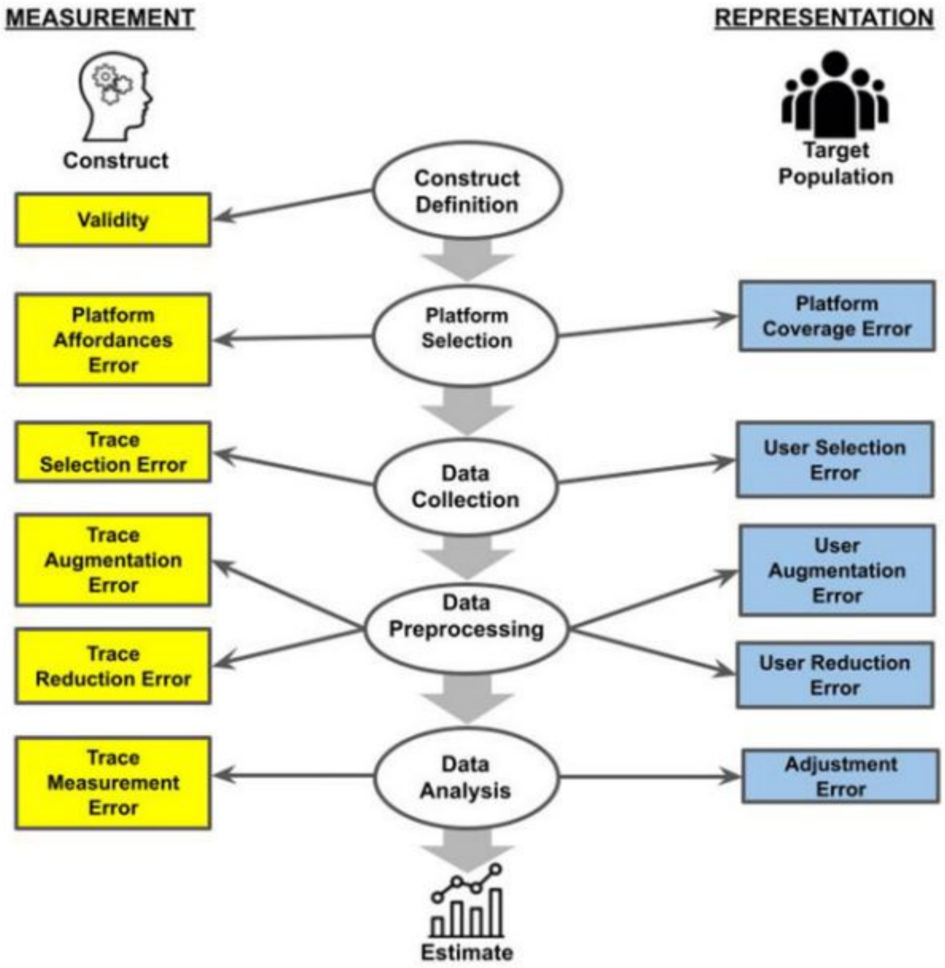

Data Collection

Preprocessing & Feature Extraction

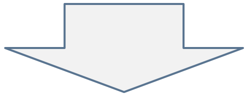
Analysis

Validation, Summarization, & Interpretation

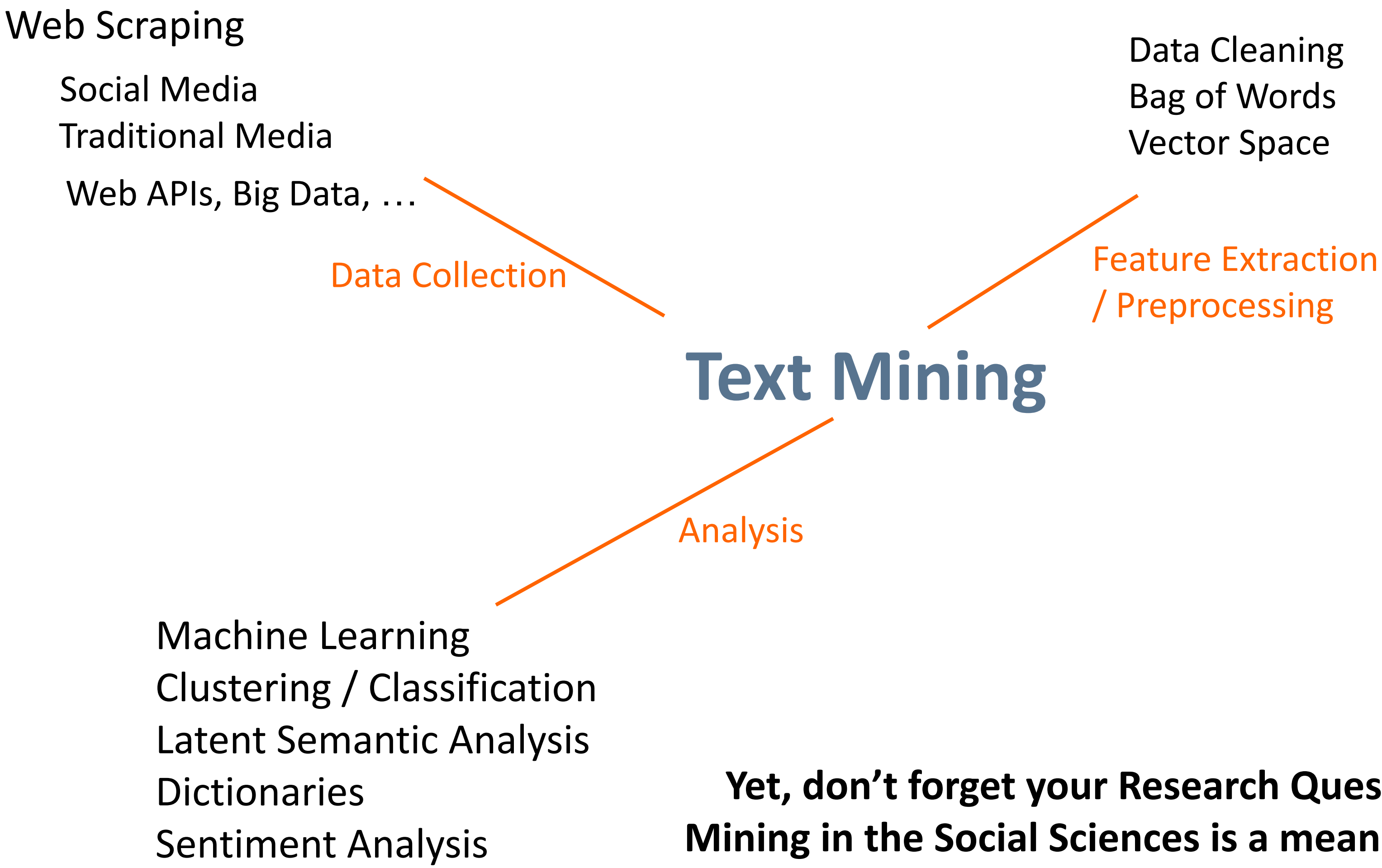
Accessing the Total Error



Groves and Lyberg "Total Survey Error: Past, Present, and Future".



Sen, et al. "A total error framework for digital traces of human behavior on online platforms".



Web Scraping

Social Media
Traditional Media
Web APIs, Big Data, …

Data Collection

Data Cleaning
Bag of Words
Vector Space

Feature Extraction / Preprocessing

Text Mining

Analysis

Machine Learning
Clustering / Classification
Latent Semantic Analysis
Dictionaries
Sentiment Analysis

Yet, don’t forget your Research Question. Text Mining in the Social Sciences is a means to an end.

Conclusion

- Clearly formulate your research question.
- Ensure you have an understanding of all stages of the process.
    - Have you selected the right data?
    - Do you have enough data?
    - Was the data cleaning step carried out the way you think?
    - Have you selected the right features?
    - Are there equivalent analysis models that may have resulted in different results?
- Ensure that all stages of your analysis are documented.
- Think about how your work could be replicated.
    - Is the data you have used available to others?
    - Can you publish your analysis, is it even possible to publish the used analysis code?

Thank you !

gesis Leibniz-Institut für Sozialwissenschaften

Leibniz Gemeinschaft

Expert Contact & GESIS Consulting

Contact: you can reach the speaker/s via e-mail: arnim.bleier@gesis.org

GESIS Consulting: GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle.

Please visit our website www.gesis.org for more detailed information on available services and terms.

More Services from GESIS

- Get materials for capacity building in computational social science and take advantage of our expanding expertise and resources in digital behavioral data.
- Use GESIS data services for finding data for secondary analysis and sharing your own data.
- Check out the GESIS blog "Growing Knowledge in the Social Sciences" for topics, methods and discussions from the GESIS cosmos – and beyond.
- Keep up with GESIS activities and subscribe to the monthly newsletter.
-

for publications, tools & services.

gesis Leibniz-Institut für Sozialwissenschaften

Meet the Experts

More from CSS Experts in the Series

June 24	Katrin Weller: **A Short Introduction to Computational Social Science and Digital Behavioral Data**
July 01	Fabian Flöck, Indira Sen: **Digital Traces of Human Behavior from Online Platforms – Research Designs and Error Sources**
July 08	Sebastian Stier, Johannes Breuer: **Combining Survey Data and Digital Behavioral Data**
Sept 16	Katrin Weller, Oliver Watteler: **Ethics and Data Protection in Social Media Research**
Sept 30	Roberto Ulloa: **Introduction to Online Data Acquisition**
Oct 07	Roberto Ulloa: **Auditing Algorithms: How Platform Technologies Shape our Digital Environment**
Oct 14	Marius Sältzer, Sebastian Stier: **The German Federal Election: Social Media Data for Scientific (Re-)Use**
Nov 04	Arnim Bleier: **Introduction to Text Mining**
Nov 11	Haiko Lietz: **Social Network Analysis with Digital Behavioral Data**
Dec 2	Olga Zagovora, Katrin Weller: **Altmetrics: Analyzing Academic Communications from Social Media Data**
Dec 16	Andreas Schmitz: **Online Dating: Data Types and Analytical Approaches**
Jan 13	Gizem Bacaksizlar: **Political Behavior and Influence in Online Networks**
Jan 27	David Brodesser: **SocioHub – A Collaboration Platform for the Social Sciences**