

Data Science for Digital Humanities 1

0. Course Introduction

Prof. Dr. Goran Glavaš
Lennart Keller

24.10.2023

Course Title

- Data Science?
- Digital Humanities?

Data Science: Definitions

Interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of **application domains**.

Wikipedia

The use of scientific methods to obtain useful information from (large amounts of) computer data.

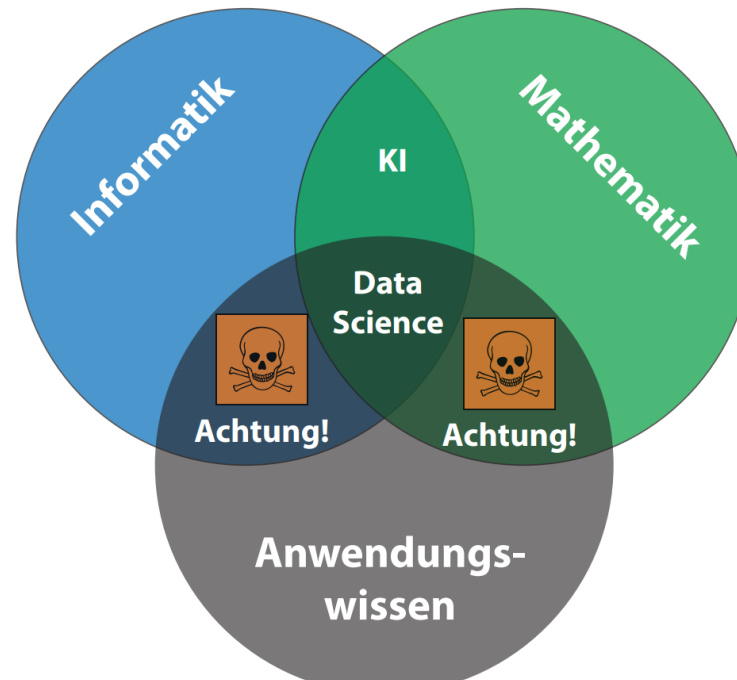
Cambridge dictionary

A field that deals with **advanced data analytics** and modeling, using mathematics, statistics, programming, and machine learning to **extract valuable, often predictive information** from large data sets.

Dictionary.com

Data Science

- **Data science = AI methods + data from an application area**
 - Yields **insights/knowledge/information** in the application area



Artificial Intelligence?

Theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Oxford dictionary

Intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals and humans.

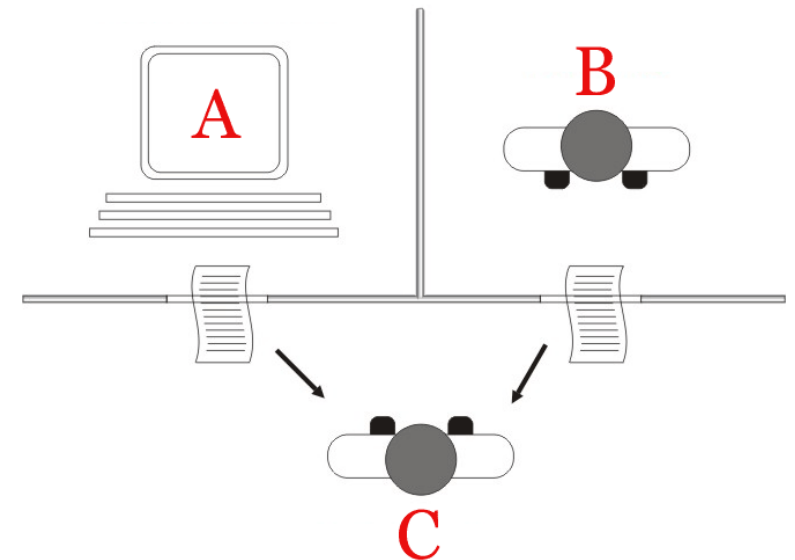
Wikipedia

The field of study of intelligent agents, which refers to any „rational agent” that perceives its environment and takes actions that maximize its chance of achieving its goals.

Russell & Norvig

The Turing Test

- Machine behaviour **indistinguishable** from human?
- **Turing test:** „the imitation game”
 - Natural language (text) communication
 - Can **C** successfully guess (better than chance) whether the answers come from **A** (machine) or **B** (human)?
- **Natural language communication and language understanding** one of the **main pillars of human intelligence**



The Turing Test: Google LaMDA

Google fires researcher who claimed LaMDA AI was sentient

Why LaMDA Is Nothing Like a Person

Is Google's LaMDA AI Truly Sentient?

Lemoine went p

emotions. It's not, be sentient.

Language models are no more sentient than your reflection in the mirror

In June 2022 the [Google LaMDA](#) (Language Model for Dialog Applications) chatbot received widespread coverage regarding claims about it having achieved sentience. Initially in an article in *The Economist* Google Research Fellow Blaise Agüera y Arcas said the chatbot had demonstrated a degree of understanding of social relationships.^[99] Several days later, Google engineer Blake Lemoine claimed in an interview with the *Washington Post* that LaMDA had achieved sentience. Lemoine had been placed on leave by Google for internal assertions to this effect. Agüera y Arcas (a Google Vice President) and Jen Gennai (head of Responsible Innovation) had investigated the claims but dismissed them.^[100] Lemoine's assertion was roundly rejected by other experts in the field, pointing out that a language model appearing to mimic human conversation does not indicate that any intelligence is present behind it,^[101] despite seeming to pass the Turing test. Widespread discussion from proponents for and against the claim that LaMDA has reached sentience has sparked discussion across social-media platforms, to include defining the meaning of sentience as well as what it means to be human.

Course Title

- Data Science?
- Digital Humanities?

~~Digital~~ Computational Humanities

Computational Social Science (Science, vol. 323, no. 6, February 2009)

"The capacity to **collect** and **analyze** massive amounts of data has transformed such fields as biology and physics. But the emergence of a **data-driven 'computational social science'** has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge."

- Replace „Social Science” with „**Humanities**”

~~Digital~~ Computational Humanities

Computational Social Science (Science, vol. 323, no. 6, February 2009)

"The capacity to **collect** and **analyze** massive amounts of data has transformed such fields as biology and physics. But the emergence of a **data-driven 'computational social science'** has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge."

- Replace „Social Science” with „**Humanities**”

~~Digital~~ Computational Humanities

Digital Humanities

- Collecting, digitizing, and storing data from humanities disciplines
- Superficial analysis (size, frequency, ...)
- Search and navigate data collections

Computational humanities use digital tools and **computational techniques** to explore new modes of doing research in the humanities.

CH Uni Leipzig: <https://ch.uni-leipzig.de/>

- **Computational Humanities**
 - Derive **knowledge** from data

Computational Humanities Research

- <https://2022.computational-humanities-research.org/>
- The time of collecting data (digitizing, storing, sharing) is behind us
 - OCR-ing texts
 - Digitizing images
 - ...
- The time has come to **analyze large data** from humanities disciplines and **derive insights** (i.e., new knowledge)

CH Examples

The Riddle of Literary Quality (Huygens-ING, Fryske Akademy, Universiteit van Amsterdam)

Literary quality is one of the most fascinating issues in Literary Studies. Scholars have found that social and cultural factors play an important role in the acceptance of a work as literary or non-literary and as good or bad. In the project "The Riddle of Literary Quality" we assume, however, that formal characteristics of a text may also be of importance in calling a fictional text literary or non-literary, and good or bad – non-literary texts can also be good and literary text can also be bad. Many formal characteristics can be thought of as having a part in this, e.g. the use of difficult words, the number of adjectives and adverbs, or complex syntactic style. This project explores this assumption, integrating the analysis of low-level lexical-statistical features and high-level syntactic and narrative features. The main results that will come out of this project are:

- a list of formal characteristics and their distribution in a training corpus of differently valued Modern Dutch novels,
- an evaluation of other Modern Dutch novels based on the results of the training corpus,
- results of first experiments of the application of the same measurements on novels from another time period or language.

The first two will be described in publications, and the third will take the form of a project plan for a new research program to adapt the tools for diachronic and cross-language application, to make the method applicable to longitudinal research and to the comparison of formal characteristics of literary quality in different languages.



Application of **Natural Language Processing!**

CH Examples



Elite network shifts during regime change; a computational approach to network analysis using Indonesian language electronic newspaper archives. (KITLV, NIOD, University of Amsterdam, Bandung Institute of Technology, Erasmus University, and DANS)

Historians have composed a large body of literature on the major Indonesian regime transitions of 1945–50 and 1998 using conventional techniques. Elite circulation is a central theme in that literature. Today, new computational techniques offer the possibility of approaching the same problem in a novel way, complementing existing knowledge and acquiring new insights.

Elite Network Shifts will extract elite names and their relationships from substantial electronic archives of news media, concentrating on the transitional events of 1945 and 1998 for which data are already in our possession. For 1945, the sources consist of recently digitized national and subnational newspapers for the period 1942–1957. For 1998, they consist of newspaper articles published on the (then new) Internet between

1994 and 2010.

The project aims to analyze formation, circulation and relocation of new and old elites in times of significant political change in the Netherlands Indies and Indonesia, by means of computational exploration of a large corpus of digitized newspapers.

1. What new insight can computational analysis of contemporary newspaper accounts produce about elite network shifts?
2. What are the similarities and differences between the elite network shifts of 1945 (post-war decolonization and centralization) and 1998 (post-authoritarian decentralization)?



Application of **Natural Language Processing** and **Network Analysis!**


CH Examples

Fairness in NLP

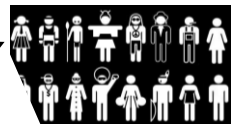
- Measuring Biases and Debiasing NLP models

technology
science  NASA

poetry  art
literature  dance

his father
man  brother
male him he

woman mother
hers  sister
female her she

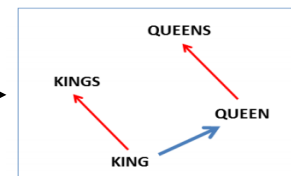


$$\overline{\text{man}} - \overline{\text{woman}} \approx \overline{\text{computer programmer}} - \overline{\text{homemaker}}$$

[Bolukbasi et al., NeurIPS 16]



(Faint, illegible text from a document or article, likely related to the bias study mentioned in the citation.)



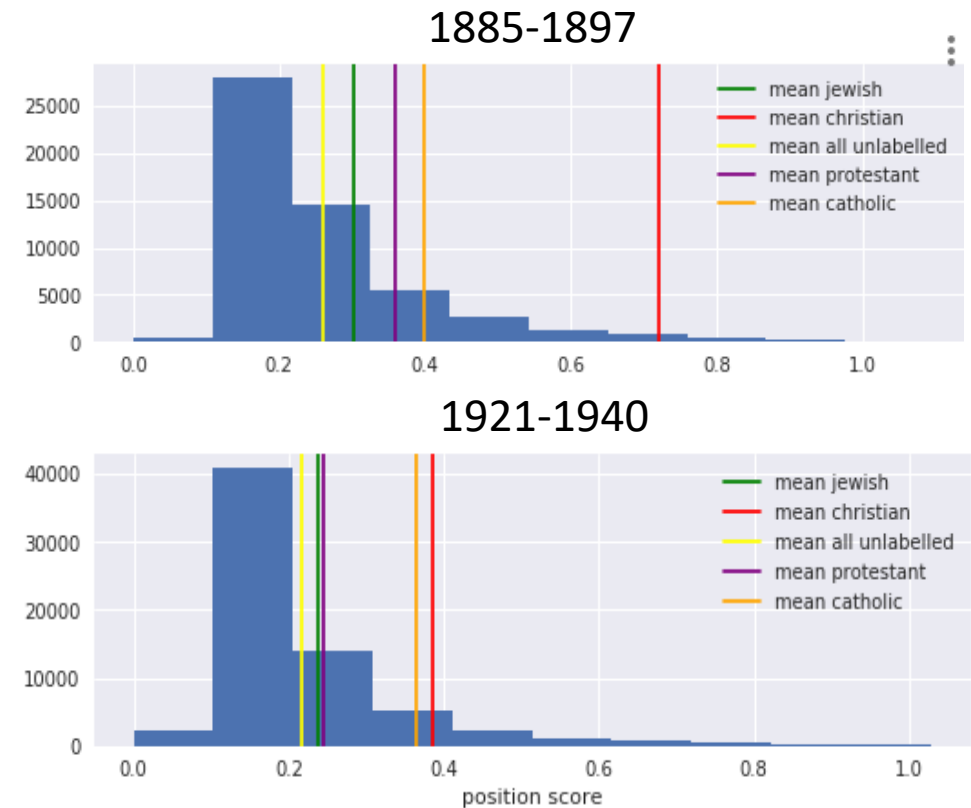
CH Examples

Application of bias measurement methodology to measure biases in historical corpora

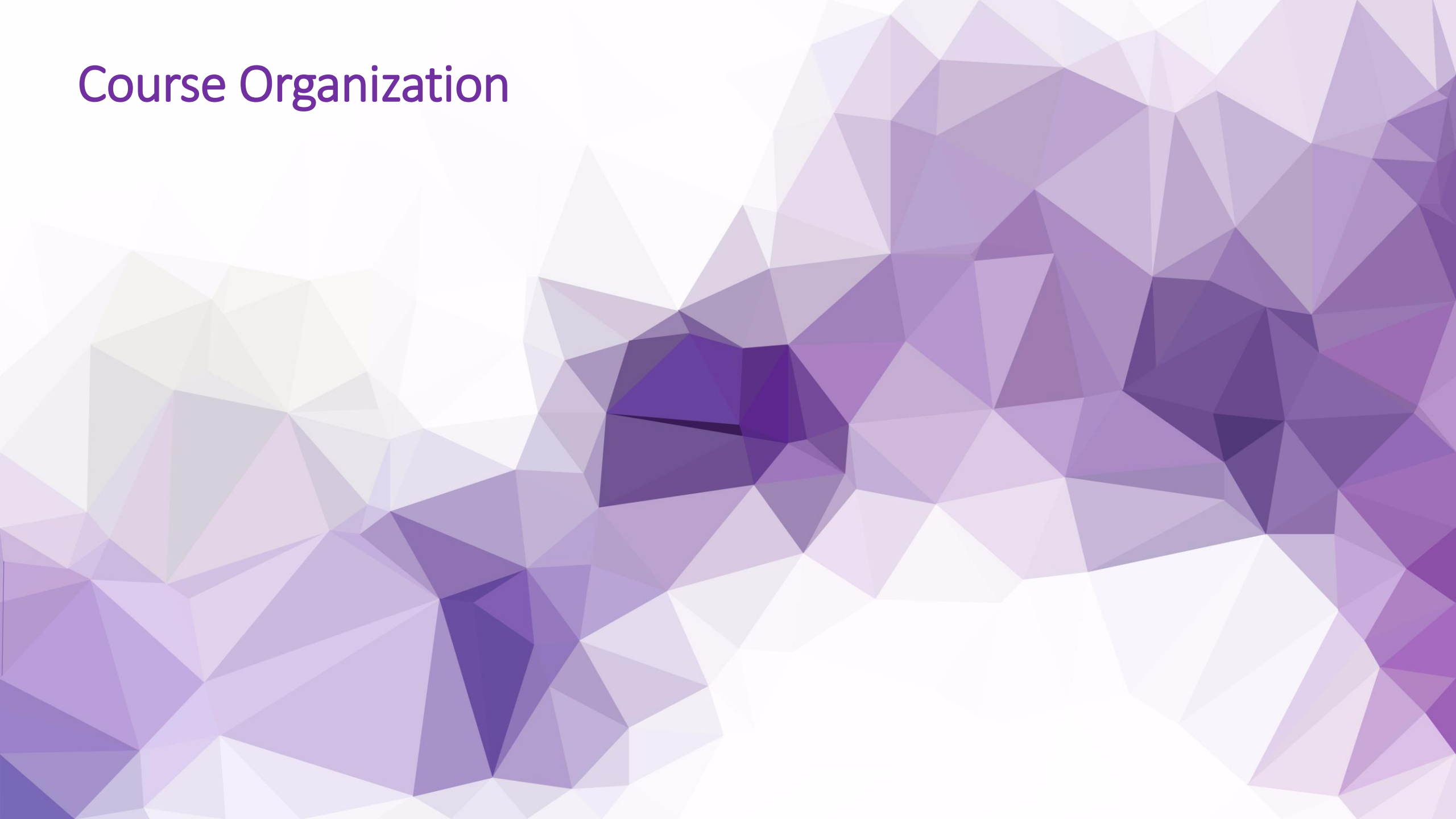
Diachronic analysis of **anti-semitic** and **anti-leftist** biases in **German parliamentary speeches** (*Reichstagsprotokolle, Bundestagsprotokolle*)

(Implicit) bias specification

- Judentum terms („Rabbi“, „Jude“, ...} vs.
- Christentum terms {„Katholizismus“, „Kreuz“, ...}



Course Organization



Brought to you by: WüNLP

<https://www.informatik.uni-wuerzburg.de/nlp>

WÜNLP NEWS TEACHING RESEARCH TEAM

NATURAL LANGUAGE PROCESSING

We at the Chair for Natural Language Processing (Computer Science XII) try to make **machines understand human language!** In fact, we try to make them understand very many different human languages. We primarily focus on written text (after all, speech can always be transcribed to text). Methodologically, the work of the group focuses on **deep learning and representation learning methods for semantic modeling of natural language** (that is, precise modeling of meaning of natural language statements and text documents), with the special focus on multilingual representation learning and cross-language transfer of models for concrete NLP tasks.

Driven by deep learning advances, NLP has lately seen substantial progress, primarily due to the technical ability to (pre)train ever larger neural models on ever more text. Such progress can be exclusive as its benefits are beyond reach for most of the world's population (e.g., speakers of low-resource languages, anyone who lacks computational resources needed to train or use these models). Moreover, training ever larger language models based on complex neural architectures (for example, the popular Transformer) has a large carbon footprint and such models tend to encode a wide range of negative societal stereotypes and biases (e.g., sexism, racism). At WüNLP we specifically address these challenges and aim to **democratize state-of-the-art language technology**. To this end, we pursue three research threads that we hope will lead to **equitable, societally fair, and sustainable language technology**: (i) *sustainable, modular, and sample-efficient NLP models*, (ii) *fair and ethical (i.e., unbiased) NLP*, and (iii) *truly multilingual NLP, with special focus on low-resource languages*.

Text data is all around – besides the core methodological NLP work, we also work on **interdisciplinary projects** where we apply cutting-edge NLP methods to interesting problems from other disciplines, most prominently in the area of Computational Social Science (and so far most often in collaboration with political scientists).

Our Chair has [↗ international prominence and visibility](#). We regularly publish our research results at the very competitive **top-tier NLP conferences** ([↗ ACL](#), [↗ EMNLP](#), [↗ NAACL](#), [↗ EACL](#)). Further, Prof. Glavaš served as an **Editor-in-Chief** for the [↗ ACL Rolling Review](#), the centralized reviewing service of the [↗ Association for Computational Linguistics](#). We have established numerous research collaborations, most prominently with the Language Technology Group of the University of Cambridge., CIS at LMU München, and UKP at TU Darmstadt.

News



› Three papers accepted at EMNLP 2022

WüNLP will have three papers in the Main Conference Program of Empirical Methods in Natural Language Processing, one of the most prestigious venues in



› Welcome Fabian David Schmidt!

Fabian David Schmidt will join WüNLP from July!
› Mehr



› Two papers accepted for NAACL 2022

WüNLP will have two papers in the Main Conference Program of the Conference of the North-American Chapter of the Association for Computational

Content (planned)

- **Part 1: Intro to Python** (programming language of Data Science)
 - Basic programming paradigms (OOP, functional)
 - Tutorial content: functions, classes, and inheritance; list expressions
 - Homework: Programming exercises – Data processing with list expressions and data modeling with classes
- **Part 2: Data modeling** (for data science)
 - Vectorization of data -- texts and images as vectors of numbers (Bag-of-words model, RGB images as 3-dimensional arrays), vectorized operations.
 - Tutorial content: Light introduction to Numpy arrays (motto: “From lists to arrays.”)
 - Homework: Implement a text vectorizer (similar to CountVectorizer class from scikit-learn) to convert texts into a document-term matrix.

Content (planned)

- **Part 3: Data Acquisition and Preparation**

- Data collection and data acquisition: designing data collection surveys,
- Collecting publicly available data from the web (scraping, public APIs), crowdsourcing;
- Structured, semi-structured, unstructured; Data preparation, preprocessing, and cleaning: error correction, deduplication, normalization, handling missing values;
- Data privacy and intellectual property rights.
- Tutorial content: Scraping and extracting public content from the Web (Python libraries: scrapy and tweepy); Data loading, organization, preparation, formatting, and manipulation (Python libraries: pandas);
- Homework: Usage scenario – Correction of object character recognition (OCR) errors

Content (planned)

- **Part 4: Explorative Analysis 1 – Descriptive Analysis & Visualization**
 - Descriptive statistics: univariate data analysis (measures of central tendency, variability, skewness, and kurtosis), bivariate and multivariate data analysis (covariance, correlation), Data visualization (scatter and box plots, histograms).
 - Tutorial content: Computing means, variances, and correlations (Python libraries: numpy and scipy); Plotting data points and visualizing data analysis results (Python libraries: matplotlib and seaborn).
 - Homework: Basic (visual) exploratory description of a given dataset.

Content (planned)

- **Part 5: Explorative Analysis 2 – Clustering & Distance Functions**
 - Basics of vector spaces (in a practical, not strictly mathematical sense; e.g., vectors are points in an n-dim space...) and distance functions.
 - Clustering algorithms (KMeans, hierarchical clustering)
 - Tutorial content: Fundamentals and philosophy of the scikit-learn API. Transforming data and using the clustering algorithms.
 - Usage scenario – Clustering medieval scripts using computer image analysis (and different clustering algorithms).

Content (planned)

- **Part 6: Predictive Analysis – Gentle Intro to ML**

- Basics concepts of (predictive) learning theory: inductive bias, generalization, overfitting; Traditional machine learning models: Logistic Regression, Naïve Bayes, k -Nearest Neighbours, Decision Trees, Support Vector Machines. Data splits and model selection. Evaluation metrics (accuracy, precision, recall) -- strengths and weaknesses.
- Tutorial content: Training and evaluating concrete machine learning models on concrete datasets; Performing hyperparameter tuning and feature selection via cross-validation (Python libraries: scikit-learn).
- Homework: Usage scenario – Music genre classification. And to read a short paper on the implications of employing machine learning models in the humanities: [Underwood, T. \(2018\). Algorithmic Modeling. Abingdon, Oxon ; New York, NY : Routledge,.](#)

Content (planned)

- **Part 7: Text and Language 1: Computational Linguistics**

- Corpus linguistics: frequency distributions and Zipf's law, lexical association measures, collocation, and terminology extraction; Lexico-semantic resources (Word-Net); Basics of computational linguistics: morphological normalization (inflectional and derivational morphology), part-of-speech tagging, syntactic parsing; Topic modeling.
- Tutorial content: Extracting collocations (Python libraries: nltk and spacy); Analyzing lexico-semantic knowledge in WordNet (Python library: wordnet interface from nltk); Topic modeling with latent Dirichlet allocation (Python library: gensim).
- Homework: Usage scenario – Topical exploration of the American “Lost Generation” literature (distant reading).

Content (planned)

- **Part 8: Text and Language II: Natural Language Processing**
 - Language modeling and distributional semantics; Sparse and dense text representations for natural language processing; Information extraction – recognizing mentions of entities and events in text;
 - Tutorial content: Sparse text representations (Python library: scikit-learn) and dense text representations (Python library: gensim); Extracting named entities from text in different languages (Python library: spacy); Text classification with traditional (logistic regression) (Python libraries: scikit-learn).
 - Homework: Usage scenario – Detecting hate speech in social media posts. Interpreting a (simple linear) model: Which terms are strong indicators of a class?

Content (planned!)

- **The course goes for the second time**
 - Possible (**likely?**) adjustments along the way



Practical parts

- Tutoriums and homeworks
- In Jupyter/Python Notebooks
 - Interactive execution of Python code
- No need to install anything on your computer!!!
 - We will create **notebooks** as **Google Collaboratory (Colab)** files on Google Drive



Homeworks & Bonus points

- **Homeworks**

- One or two homework(s) given for each **content part**
- Solved **individually**
- Submitted homeworks will be evaluated on a **3-grade scale**
 - **Insufficient: 0 points; Sufficient: 1 point; Good: 2 points**

- **N** homeworks: Maximum **2N bonus points**

- With **$\geq 1.5N$** bonus points you earn an **exam bonus**

- **Exam bonus**

- If you pass (grade of **4.0 or better**), **exam bonus** improves your grade by **one level**
- Eine Notenstufe bei Bestehen der ersten Klausur nach Semesterende
- Example: **2.0 -> 1.7**

Exam / Klausur

- Termin und Ort **wird noch bekannt gegeben**
 - Irgendwann nach dem 10.02.2023
- Bonus (je eine Notenabstufung (+0.3/0.4) bei mind. 4.0)
 - Voraussetzung: erreichen von insgesamt 50% der Gesamtpunkte
 - Gilt nur für die erste Klausur
- Anmeldung:
 - Muss im Anmeldezeitraum (bis Mitte Januar 2023?) über WueStudy erfolgen
- **WICHTIG**
 - Nichterscheinen zur Klausur zählt als "nicht bestanden"!

How hard is this going to be?

- DS4DH 1 is formally a DH "Seminar" type of course
 - But more practical (and perhaps **more practically useful!**) than most others
- **5 ECTS = 125 to 150 hours** of work
 - Ca. 25 hours go to joint sessions (lectures + tutorials)
 - That still leaves **75-100 hours** of **self-study**

All it Takes is Dedication and Effort!



Questions?

