

Computational Methods for Linguists

Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)
04/01/21

What's your most/least favorite April 1st joke?

Class policies

(Forgot to mention last time)

- Academic integrity:
 - See the syllabus!
 - No plagiarism
 - Assignments:
 - To be completed individually
 - No copying from websites
 - We will talk about how to use e.g. StackOverflow



Reminders

- Patas cluster
 - Request access
- “Assignment 0”
- Public website:
 - Assignment descriptions
 - Schedule
- Canvas
 - Announcements
 - Introductions and other discussion
 - Homework submission
 - Blog options



Ling471 & Ling472

Compared

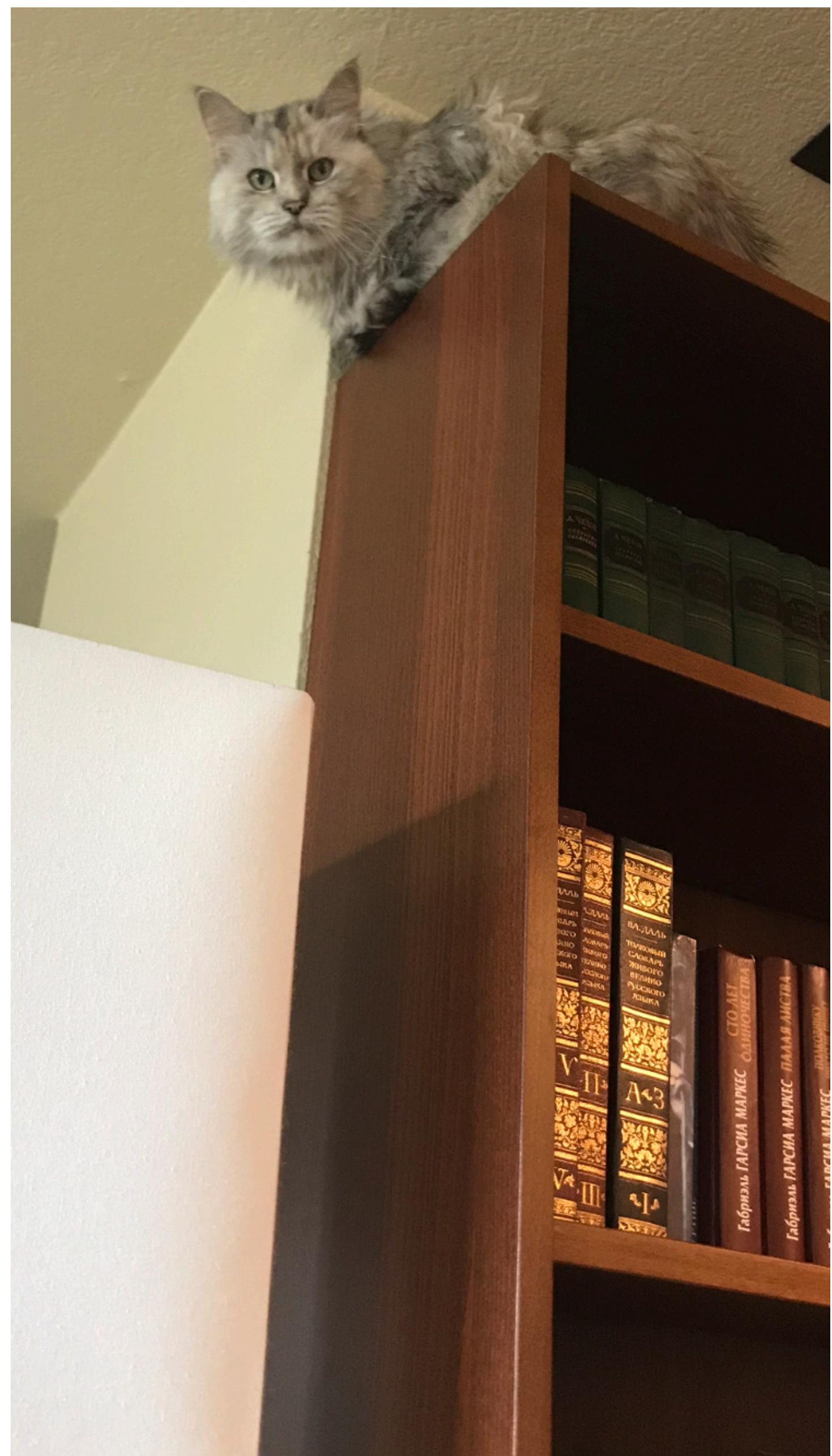
- Ling472 (Intro to CompLing)
 - Usually harder for those without programming experience
 - But this year, flexible individual goals are allowed
 - Covers statistical and formal methods in CompLing
 - Organized around levels of linguistic structure
- Ling471 (Comp. Methods for Linguists)
 - Lots of programming etc. basics
 - Focuses only on statistical methods
 - Focuses on data science and how it is related to CompLing
 - Organized as incremental skills building
- Courses may be taken concurrently
 - There will be overlap, e.g. Machine Learning lectures
- Both instructors will collaborate on connecting the courses



Plan for today

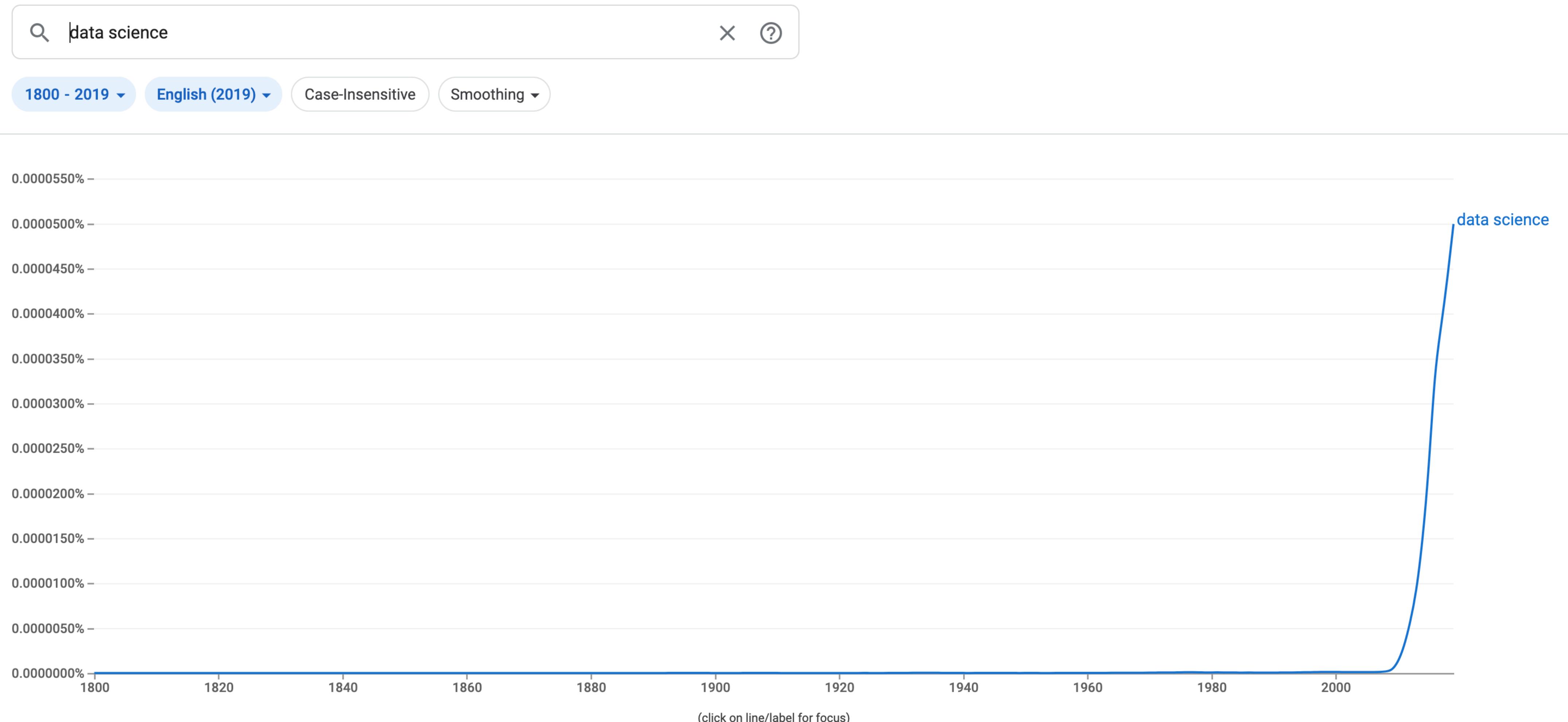
Conceptual Overview

- Data science
 - Data
 - Language data
 - Statistical methods
 - Machine learning
 - Programming
 - Systems



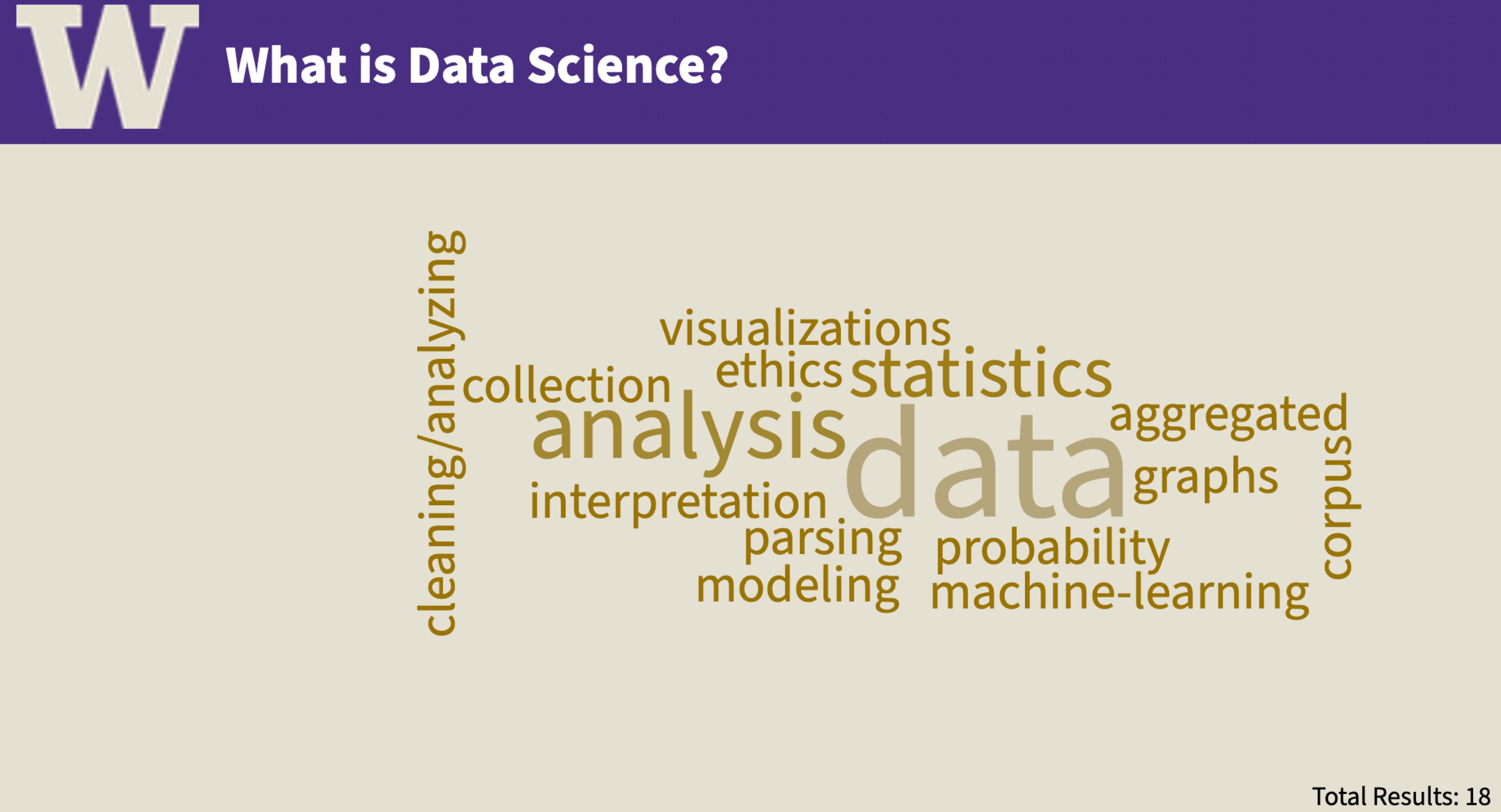
What is Data Science?

Google Books Ngram Viewer



Activity

<https://PollEv.com/olzama> (NB: it splits by word!)



Wikipedia:

“Interdisciplinary field”



The screenshot shows the Wikipedia page for "Data science". At the top, there is a circular diagram illustrating the interdisciplinary nature of data science, featuring icons for scientific thinking, innovative technologies, data science, statistical methods, and computer engineering. Below the diagram, the page title "Data science" is displayed, followed by a subtitle "Field of study". The main content area describes data science as an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. It is related to data mining, machine learning and big data. A link to the "Wikipedia" article is provided. Further down, sections include "Mathematics", "Career path", "Skills", and "Employment outlook", each with a dropdown arrow indicating more content. The "Data science books" section features five book covers: "Practical Statistics for Data Scientists", "An Introduction to Statistical Learning", "R for Data Science", "Python Data Science Handbook", and "The Elements of Statistical Learning". A "View 45+ more" link is also present.

Data science

Field of study

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

[Wikipedia](#)

Ms salary: \$92,500 [forbes.com](#)

Mathematics

Career path

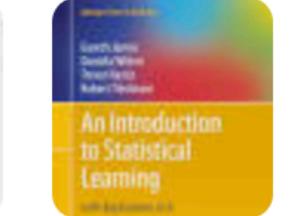
Skills

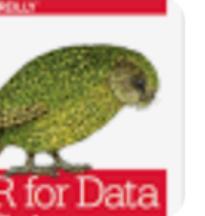
Employment outlook

Data science books

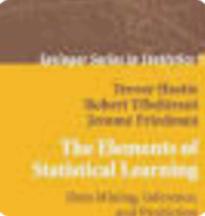
[View 45+ more](#)

 Practical Statistics for Data Scientists

 An Introduction to Statistical Learning

 R for Data Science

 Python Data Science Handbook

 The Elements of Statistical Learning

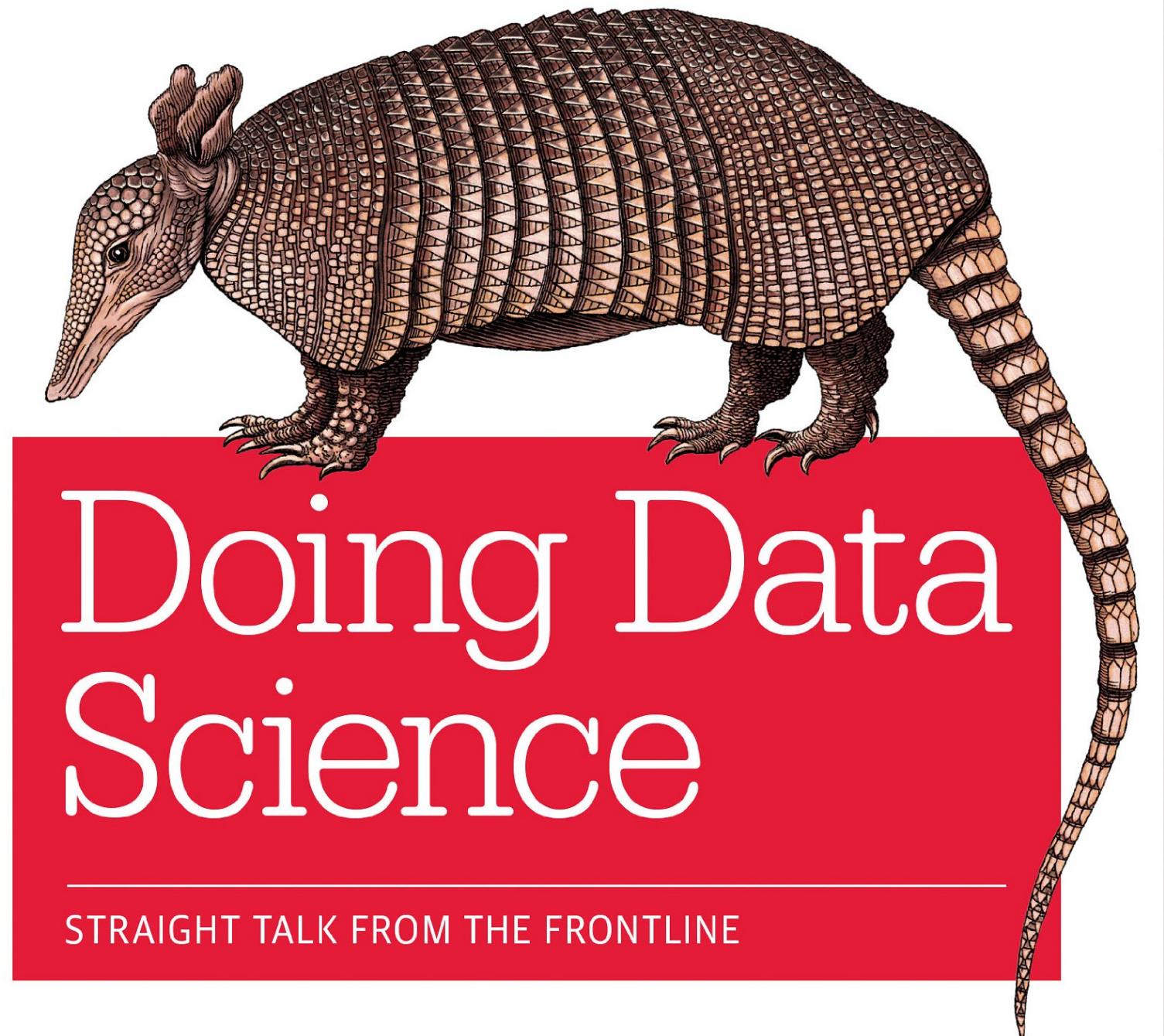
O'Neil & Schutt:

Doing Data Science

*Data science is a **practice** where people with different types of expertise come together to solve a problem through analyzing large amounts of data.*

(NB: This is not our textbook or anything! You don't need to buy it.)

O'REILLY®



Cathy O'Neil & Rachel Schutt

Compare:

- *Interdisciplinary field that uses ... algorithms...to extract knowledge*
- *A practice when different specialists solve a problem through data*
- How do these definitions differ? Which one do you like more? Why?



Another definition

(by Denis Altudov)

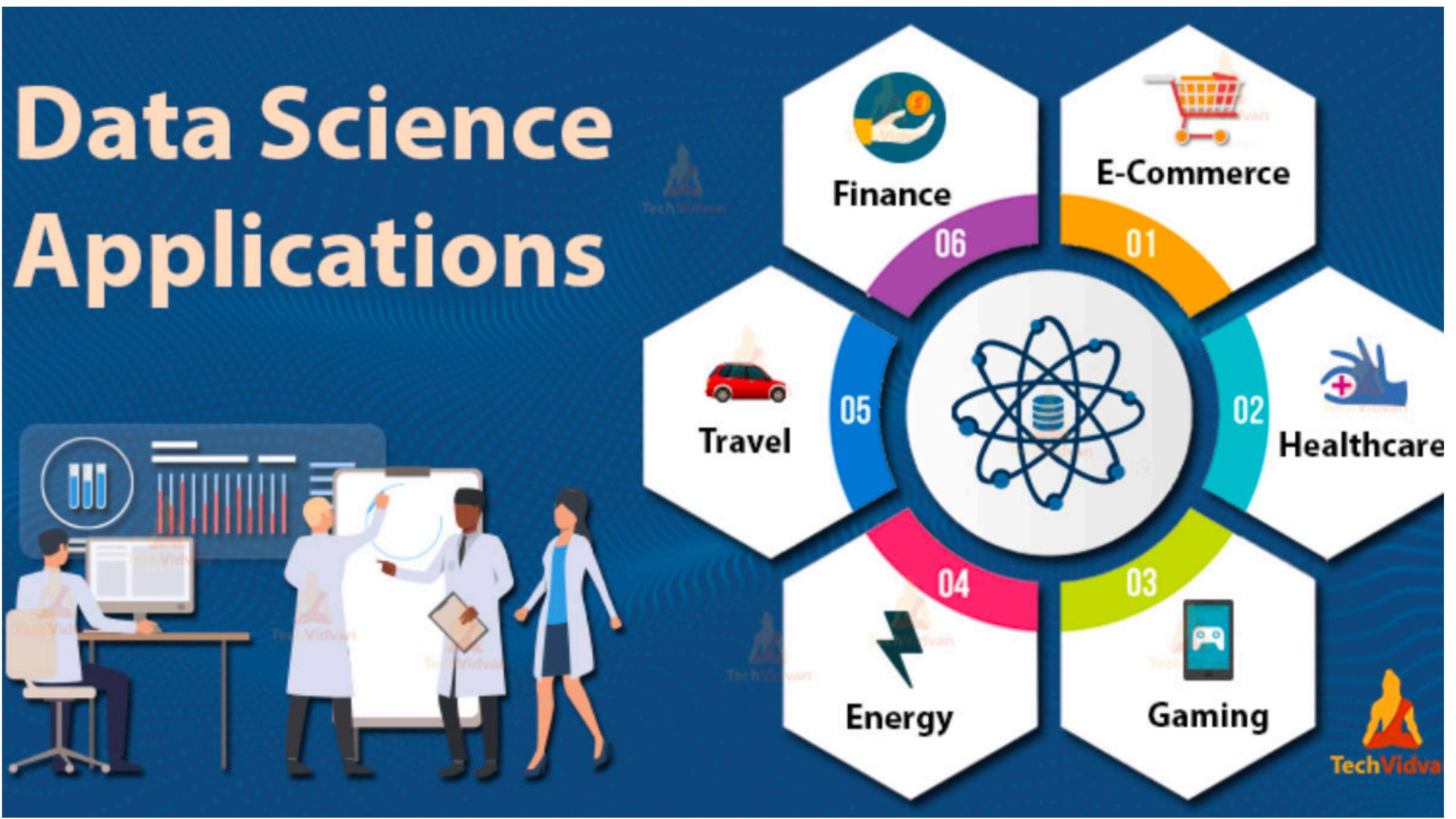
“Data science is application of rigorous statistical techniques to fuzzy objects of study”



Data Science

Examples

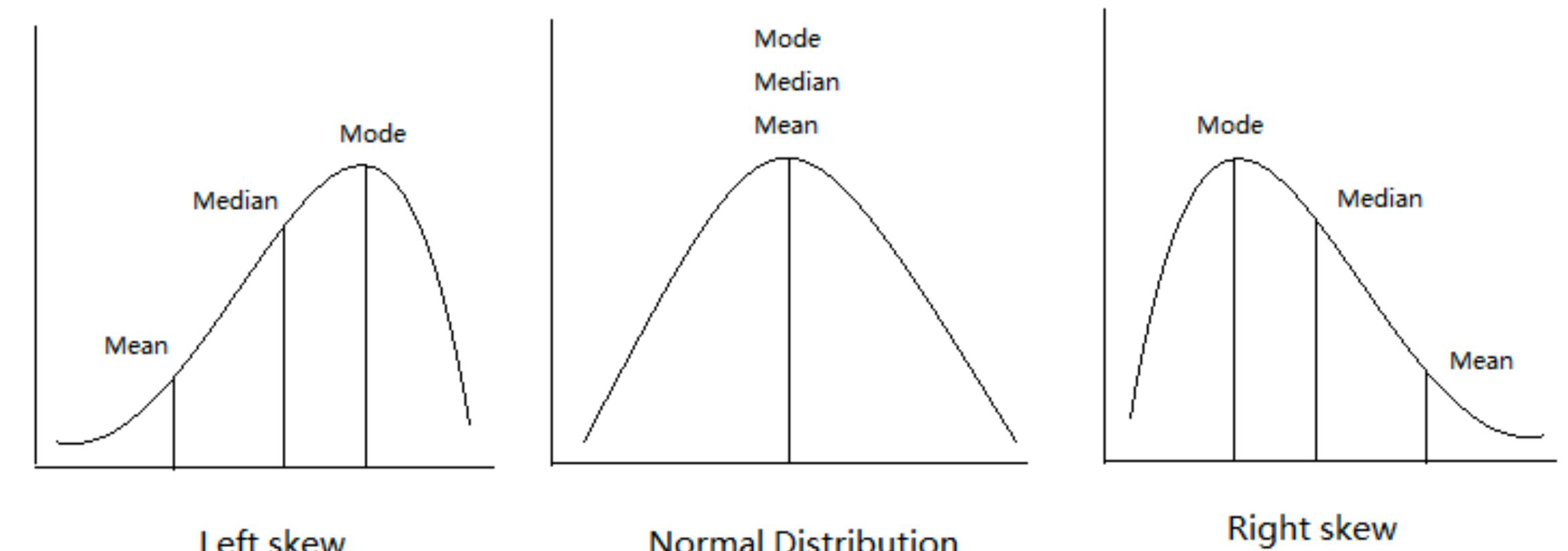
- Analyze data **statistically**
- Predict something
 - e.g. how people will behave
 - how something will impact people
- Retrospectively:
 - What happened in the past
 - What it means for the future



<https://techvidvan.com/tutorials/data-science-applications/>

Statistical analysis

- Data science is **inherently** statistical
- Statistics interprets quantitative data
 - Basic examples:
 - Mean and Median
 - “What’s the average salary in Seattle?”
 - “Are women more talkative than men?”
 - What about **qualitative** data?
 - Qualitative **methods** play a role
 - Data tends to be **digitized**
 - But the **nature** of data may remain **qualitative**



<https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b>

What we need to do data science:

- Datasets
 - BIG ones!
 - Why?
- Domain expertise
 - E.g. linguistics
 - This is **the whole point**
- Statistics (and probability theory!)
 - Statistics **interprets quantitative data**
- Programming
 - **Large** quantities of data cannot be processed with pen and paper
- Visualization and communication
 - Because **business**
- But wait! What about machine learning and deep learning?!
 - Yes, but they apply as **statistical methodologies**



Data science profiles

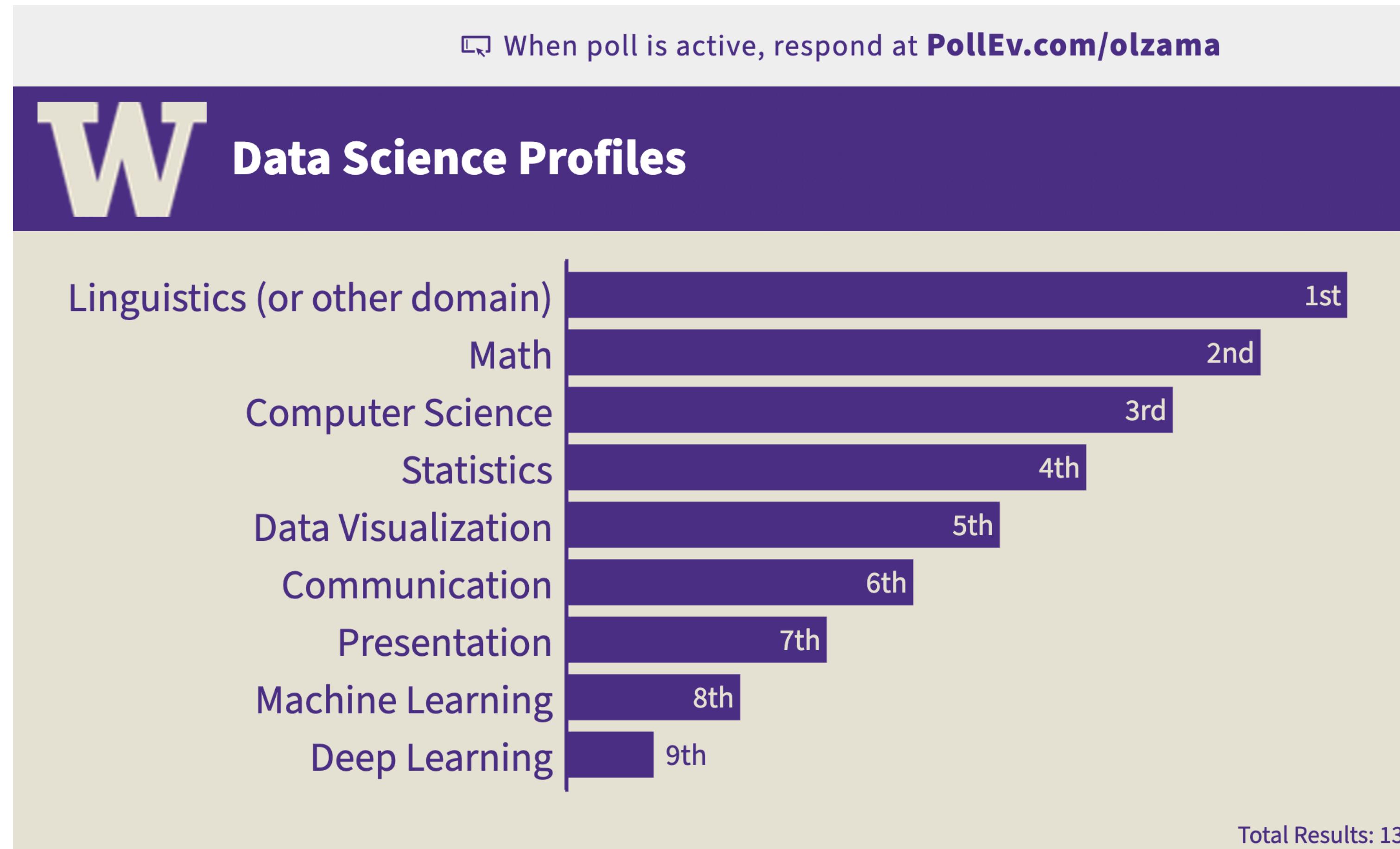
Activity

- Data science requires **different** skills
- Rank yours in relative terms of strength, to **visualize** your data science “profile”
- NB: This is your profile **the way it actually is**; not what you think it “should” be!



Our collective Data Science profile:

<https://PollEv.com/olzama>



Data and Datasets

- What counts as “data” in Linguistics?



Activity

<https://PollEv.com/olzama> (NB: it splits by word!)

W

What counts as "data" in linguistics?

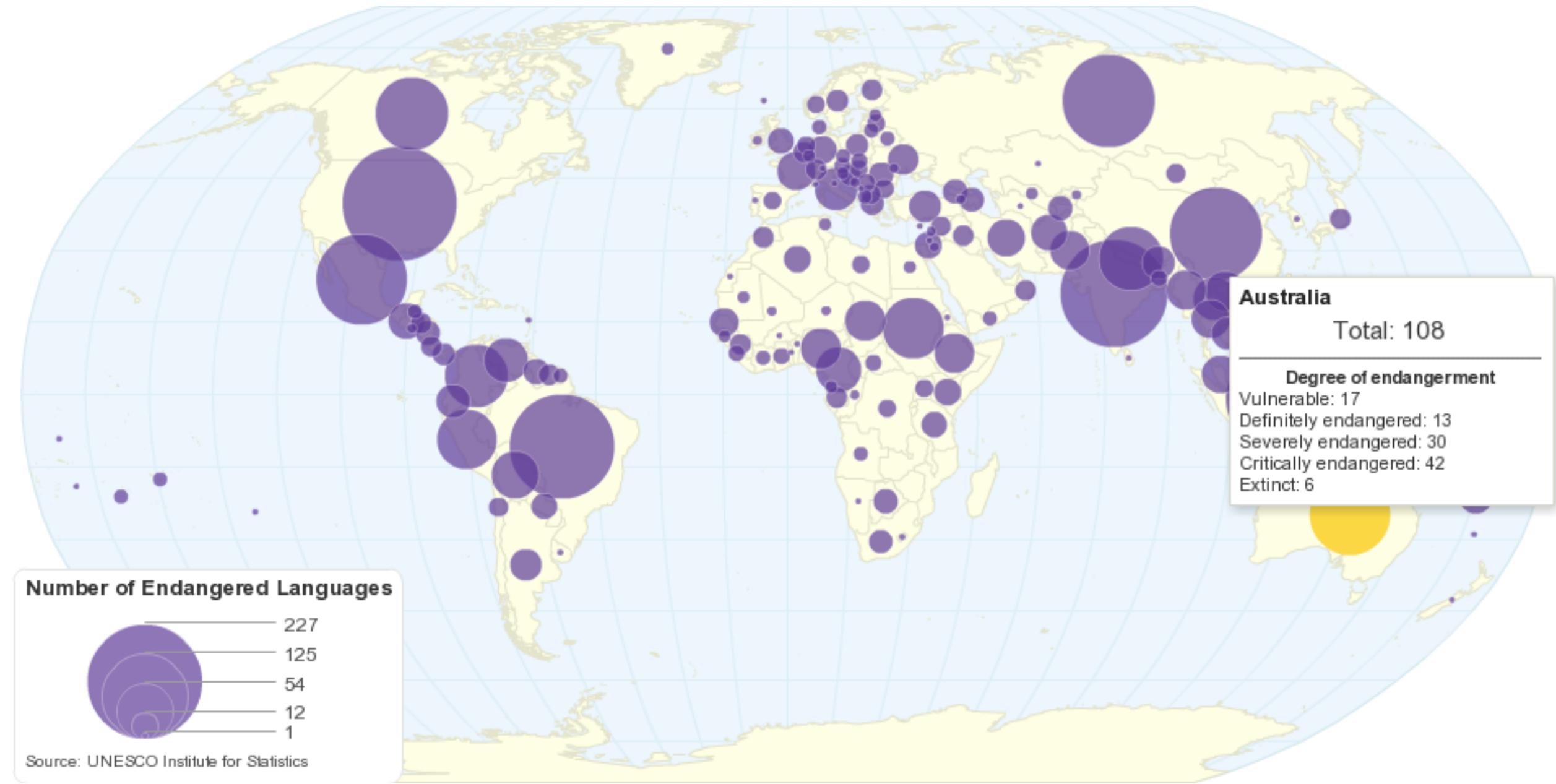
A word cloud visualization where the size and position of words represent their frequency or importance. The most prominent words are 'speech', 'interviews', 'corpus', and 'text'. Other visible words include 'eye-tracking', 'brain waves', 'people', 'sentences', 'eye-movement', and 'comprehension'.

eye-tracking
interviews brain waves
speech text
corpus people sentences
eye-movement
comprehension

Total Results: 16

Data In Linguistics

- Recorded speech and text
 - Why in this order?
 - Most languages of the world aren't written!
 - Is text more structured than speech?
 - Is it easier to work with text or with speech?
 - What does it mean to focus on text vs speech?



<http://chartsbin.com/view/1339>

Data in linguistics and data science

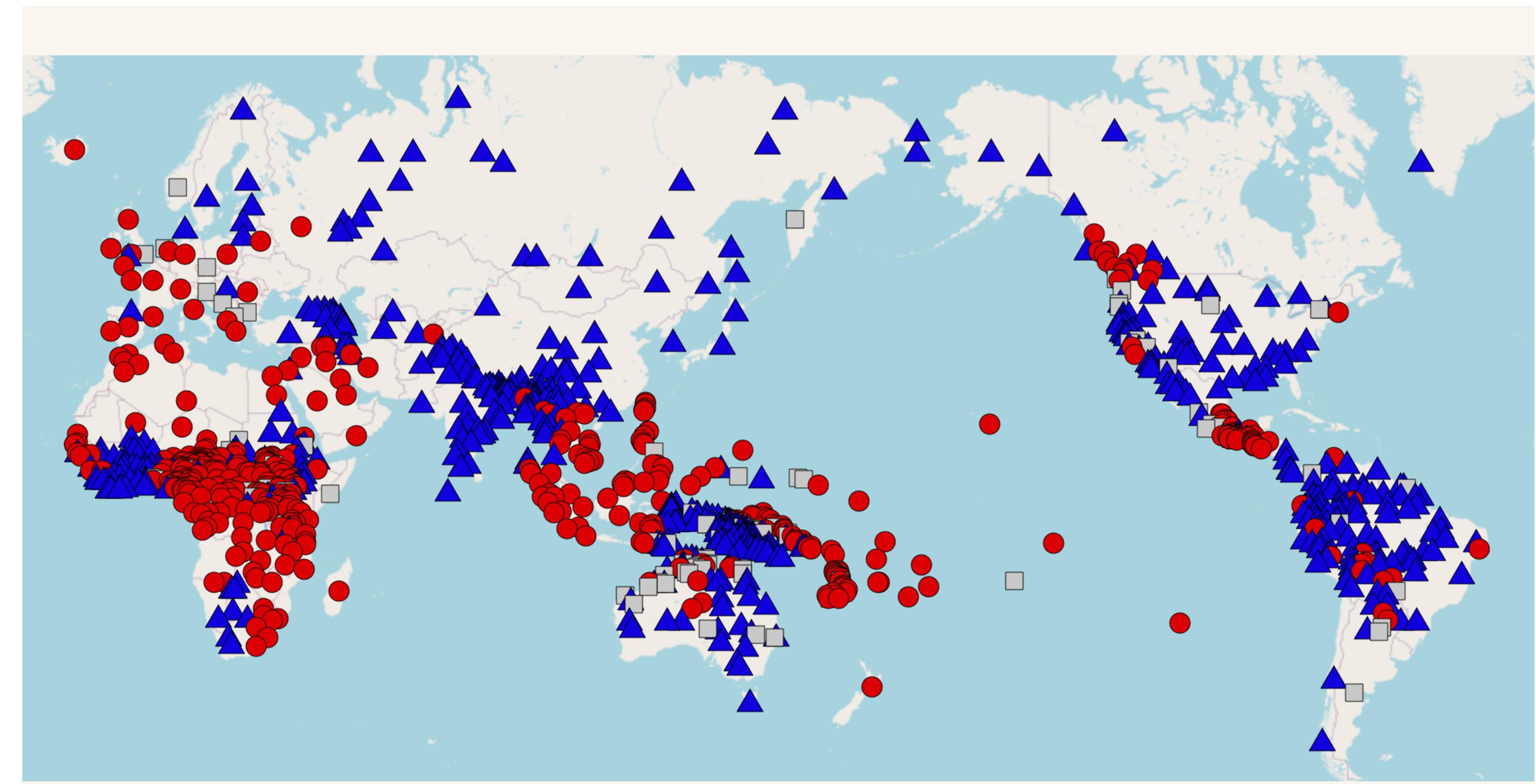
- Data science usually means LOTS of data
 - Why?
- Which areas of linguistics have LOTS of data?
 - **Most of them**
 - ...potentially



<https://depts.washington.edu/ldplab/>

Linguistics and data science

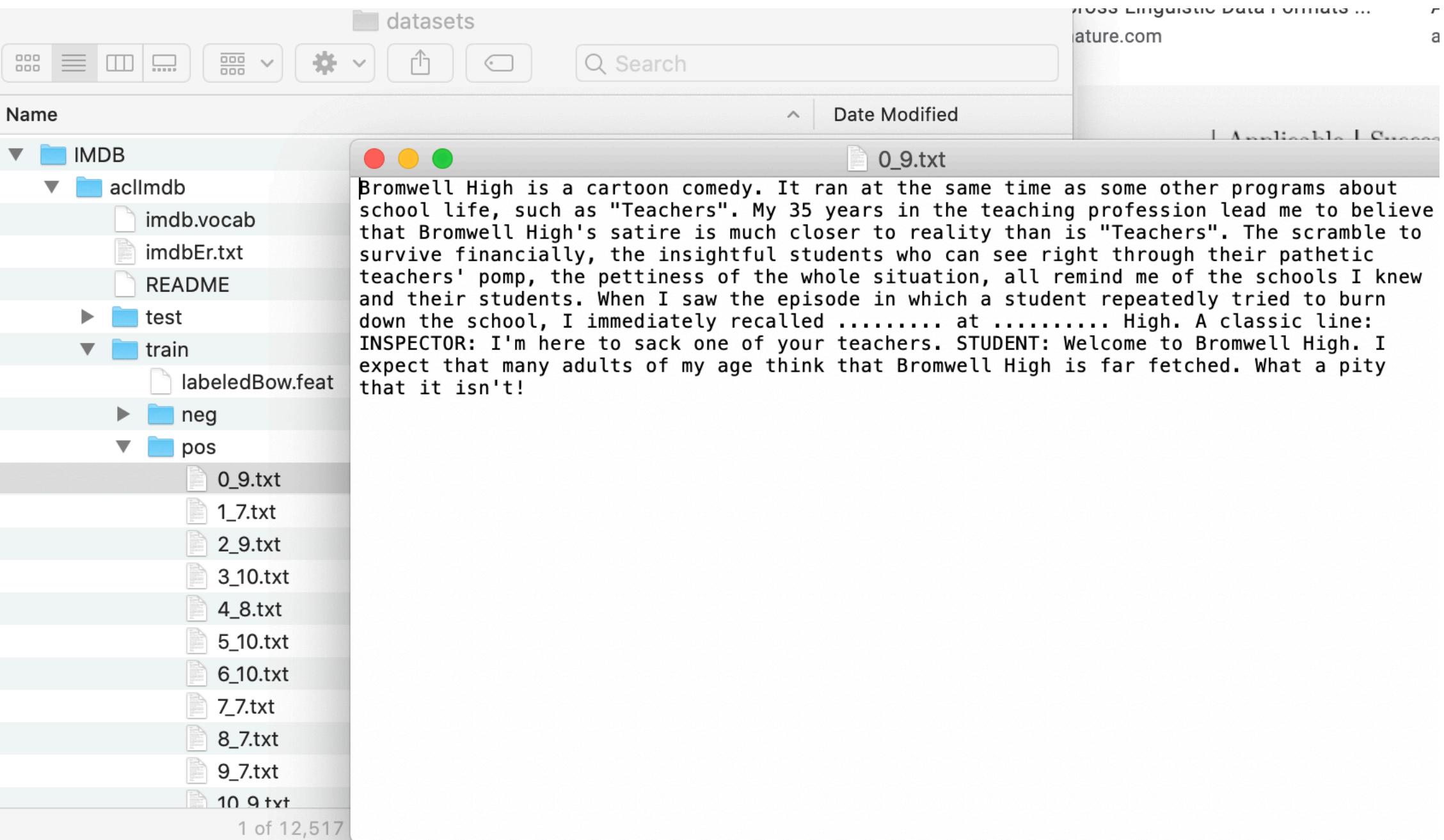
- “Corpus linguistics”
 - Various subfields; statistical analysis over large texts
- Sociolinguistics
 - Statistically significant correlations between sociolinguistic variables
- Historical (“dyachronic”) linguistics
- Linguistic typology
- What else?
 - Almost **everything, potentially**
 - So long as the data can be **managed**



<https://wals.info/feature/86A?v1=t00d&v3=sccc#2/21.0/152.9> (Dryer, 2005. WALS. Order of Genitive and Noun)

Datasets

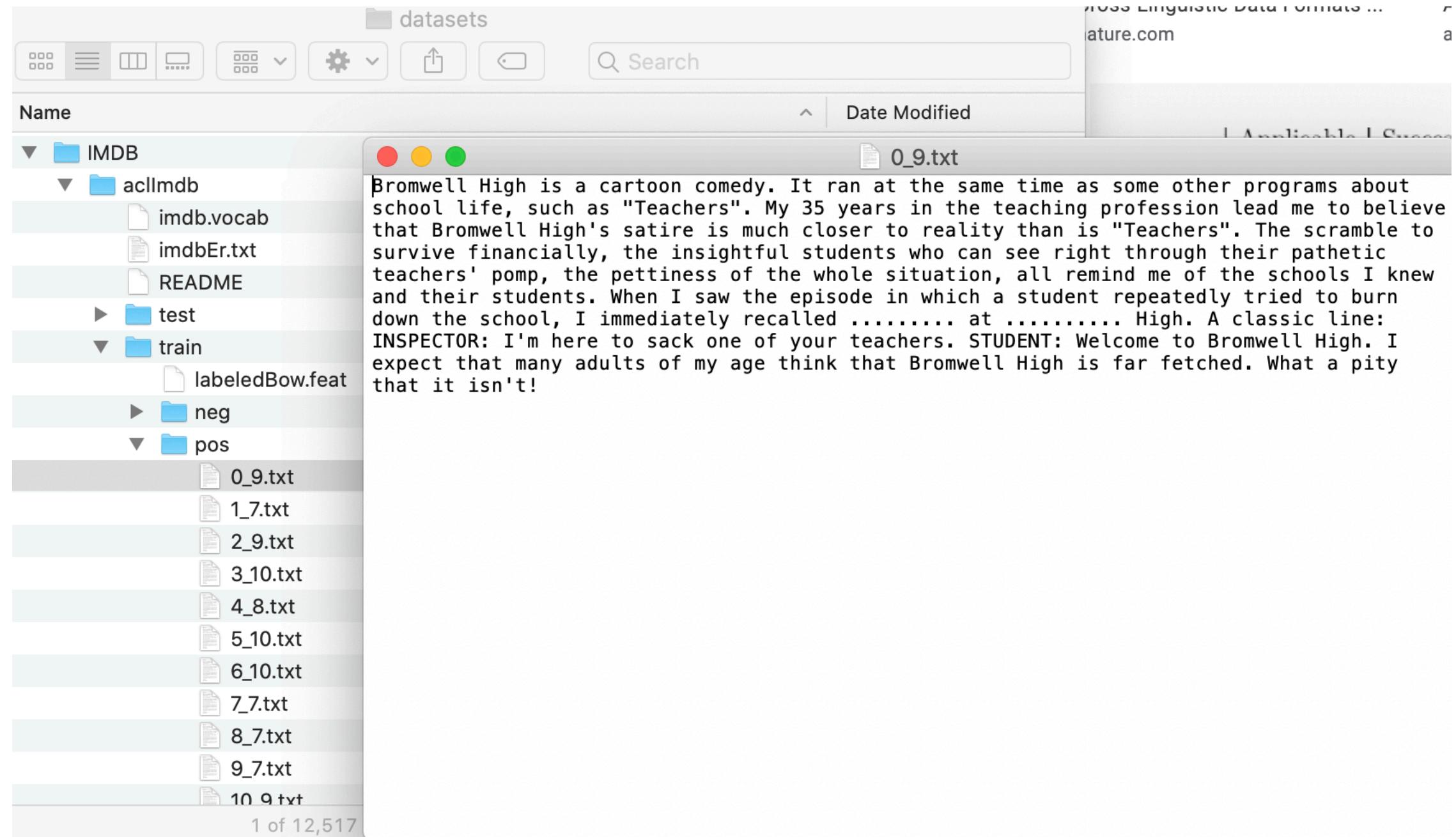
- Datasets are sets of data :)
 - No really! Anything that's data is a dataset
 - But, in practice, datasets are **workable** data
 - Size, annotation, structure, consistency, noise
 - Some datasets are used again and again in research
 - The IMDB reviews dataset!
 - That's why, datasets should be and are recognized as a **contribution to science**



A screenshot of an IMDB movie review page for the film "You've Got Mail" (1998). The page header shows the movie's title and year. Below the title is a thumbnail image of the two lead actors. The main content area displays user reviews. One review is highlighted, showing a 10/10 rating and the text: "Unbelievable film! Such a warmth and happiness!". The review is attributed to "SRVforever" and dated "30 December 2005". Below the review, there is a detailed summary of the film's qualities, mentioning its beauty in words and moments, excellent acting, and feel-good atmosphere. At the bottom of the page, there is a note indicating 114 out of 148 users found the review helpful, and links for "Sign in to vote" and "Permalink".

Datasets

- Datasets take many forms
- But they all have **provenance**
- **Language** datasets are produced by people
 - People have age and a variety of identities
 - ...influencing **how** they talk and write
 - The provenance of the dataset biases the system that's based on it
 - Can introduce **additional** biases to "balance it out"
 - But **no such thing** as an unbiased system



A screenshot of an IMDb movie review page for the film 'You've Got Mail' (1998). The page shows a summary of the movie, a user rating of 10/10, and a single review by 'SRVforever' dated 30 December 2005. The review is highly positive, praising the film's warmth and happiness. Below the review, there is a note indicating 114 out of 148 users found it helpful.

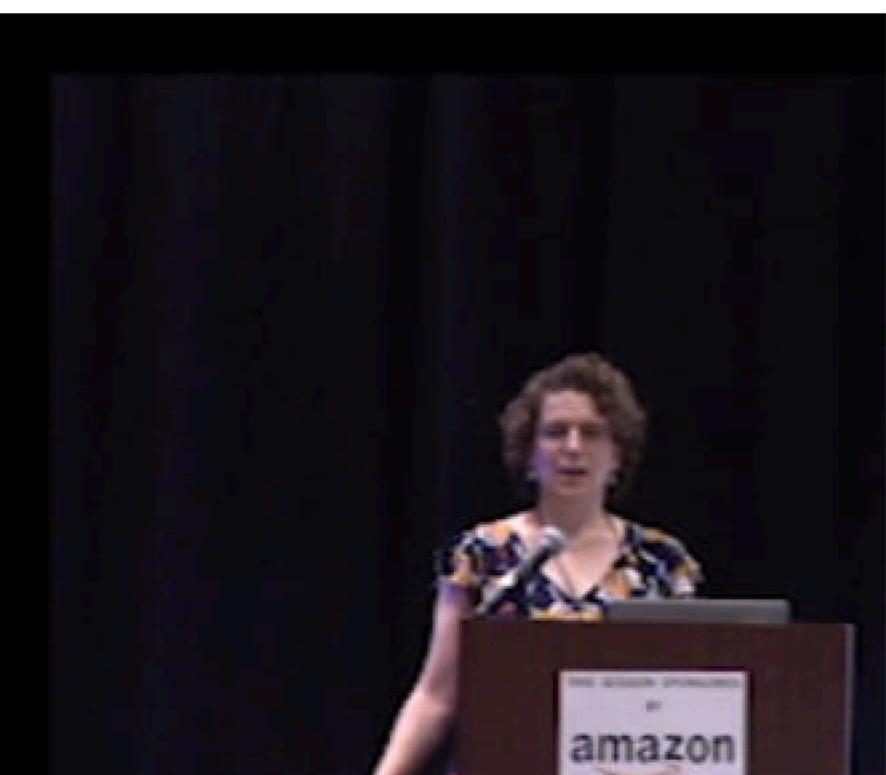
Data statements

Bender & Friedman 2019

- Reading assigned “for April 8”
 - But will discuss now (not on April 8)
 - Read it soon (to do Assignment 1)
- A **practice** of stating facts about the dataset used to “train” the system
 - “Training” is a machine learning concept
 - The system “trains” by getting feedback on how well it did something **with respect to the data**

Data statements alone won't 'solve' bias, but if we do not make a commitment to data statements or a similar practice for making explicit the characteristics of datasets, then we will single-handedly undermine the field's ability to address bias.

Emily M. Bender giving a talk on Data Statements for NLP in 2019 at NAACL



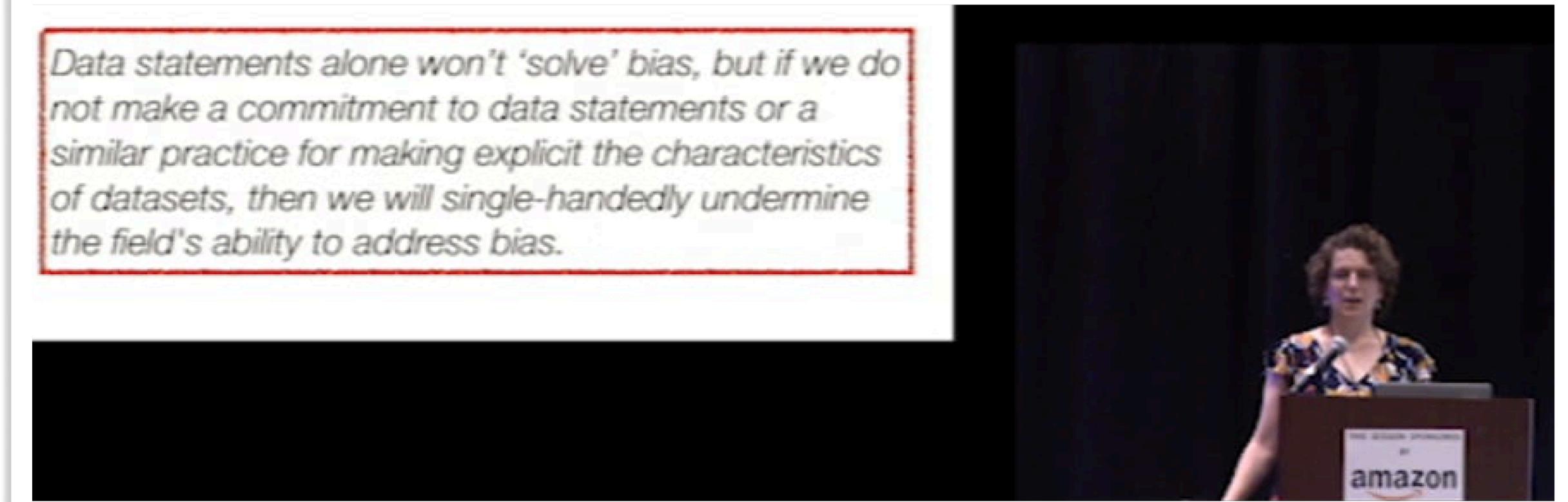
Data statements

Bender & Friedman 2019

- Schema:
 - Curation rational
 - How were data selected and why
 - Speaker demographics
 - Annotator demographics
 - Many datasets are **annotated** in some way
 - e.g. positive or negative review?
 - Could the judgment here differ?!
 - Speech situation
 - Text genre etc.
 - Recording quality

Data statements alone won't 'solve' bias, but if we do not make a commitment to data statements or a similar practice for making explicit the characteristics of datasets, then we will single-handedly undermine the field's ability to address bias.

Emily M. Bender giving a talk on Data Statements for NLP in 2019 at NAACL



Annotated data

- Speaker characteristics (e.g. region)
- L1/L2 speech
- Part of speech
- Named entity
- Sentence structure
- Discourse structure
- Some of the above are what we want the system to predict; other things are useful **downstream**

Manchester United produced a stunning comeback from 2-0 down to beat Newcastle United - and potentially save manager José Mourinho from the sack. Second half strikes from Juan Mata and Anthony Martial levelled proceedings, before an injury-time winner from Alexis Sánchez sent the Old Trafford crowd into raptures. After the international break, the Red Devils travel to Stamford Bridge where they face a stern test against unbeaten Chelsea.

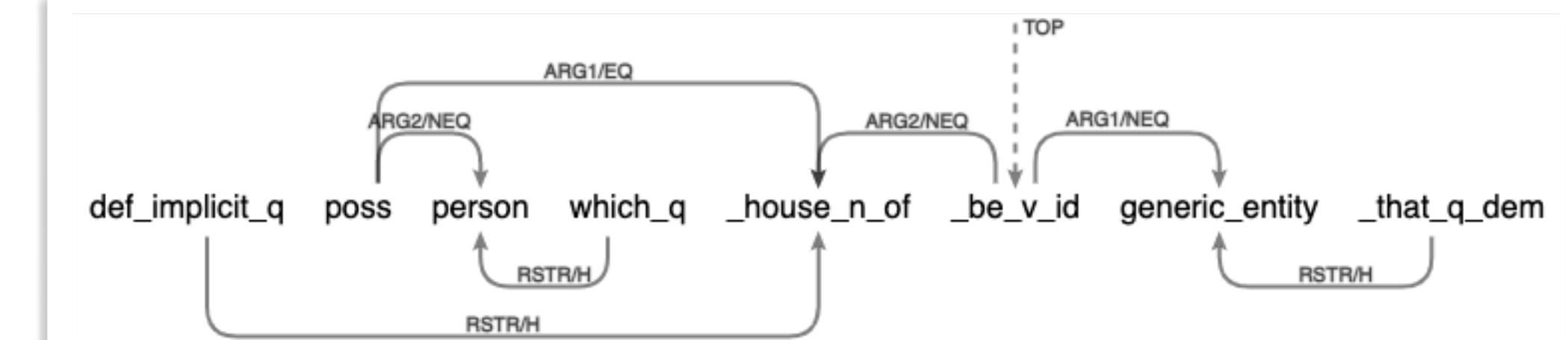
<https://lionbridge.ai/articles/training-data-is-key-for-natural-language-processing-algorithms/>

Annotated data In linguistics

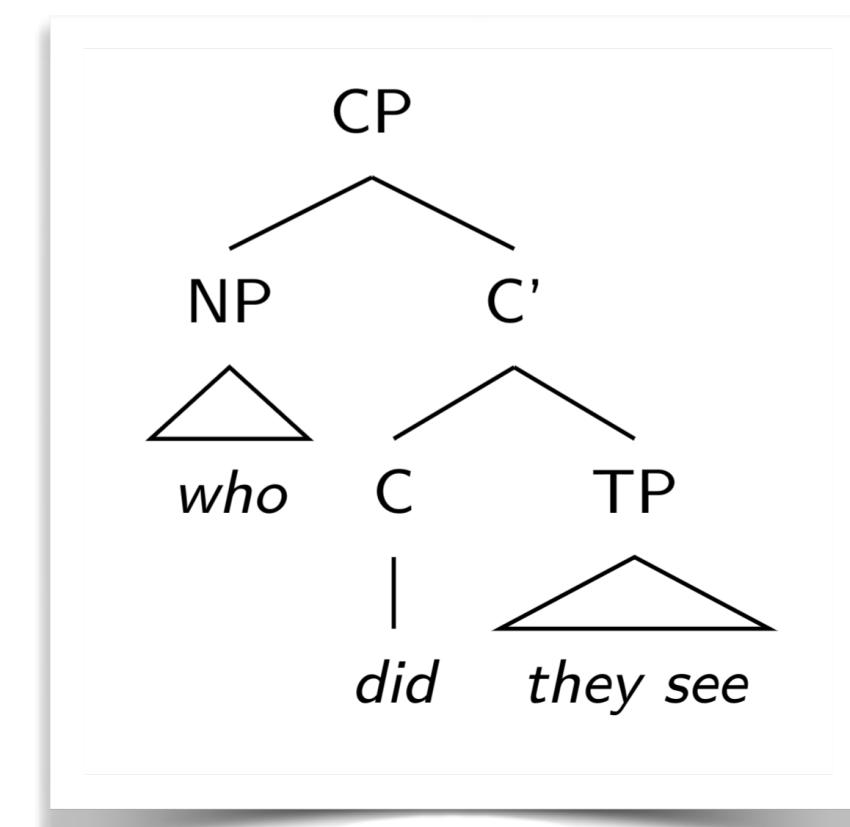
- Recorded speech and text associated with sociolinguistic variables:
 - Gender, age, geographic region...
- Interlinearized Glossed Text
 - Linguistic **analysis** and **annotation**
- **Structural** annotations
- What about syntax trees?
- In **computational** linguistics?
 - In **NLP**, emphasis on **raw** data
 - Why?
 - NLP is a computer science discipline
 - Deep learning

mā?ā-nī sàá =∅ nè =V
who-INDEP house =be there =Q
'Whose house is that?' [bxl]

Heath 2017. *A grammar of Jalkunan (Mande)*

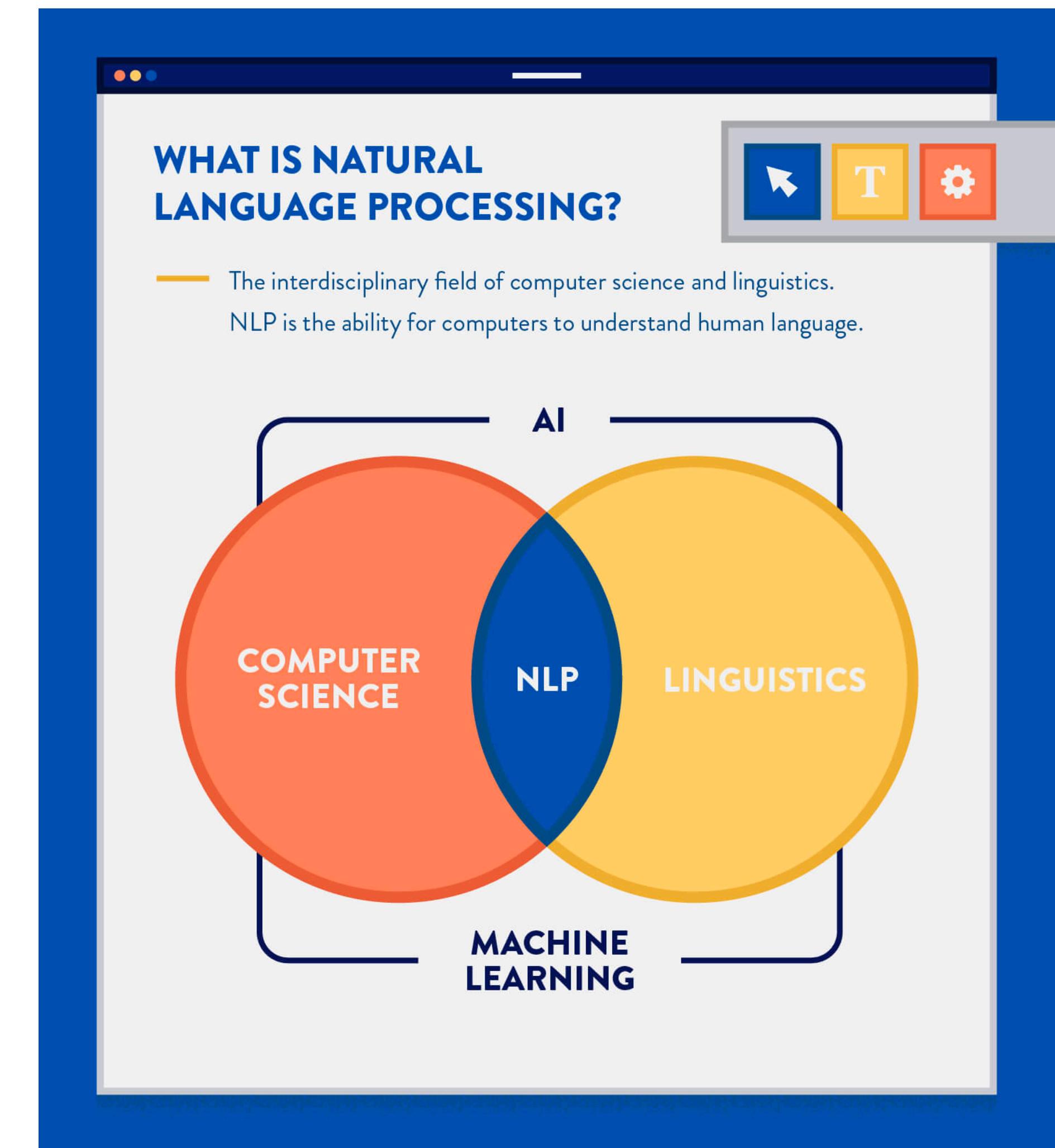


A dependency graph for the above sentence
One of the **most** useful data formats in NLP!



Computational linguistics and natural language processing

- Is it the same thing?
 - For many people, yes! Interchangeable use...
 - For some:
 - **CL:** What can we learn about human language with computer aid?
 - E.g. How did pronunciation in Wales changed over the last 50 years?
 - Is this data science?
 - **NLP:** What can we learn about the world through the lens of language data? (Mostly text...)
 - E.g. How likely is the person to enjoy one movie based on **how they talk** about another movie?
 - Is this data science?



<https://clevertap.com/blog/natural-language-processing/>

NB: Olga **doesn't** think computers can “understand” human language!
(But that's a philosophical debate.)

Natural language processing and data science

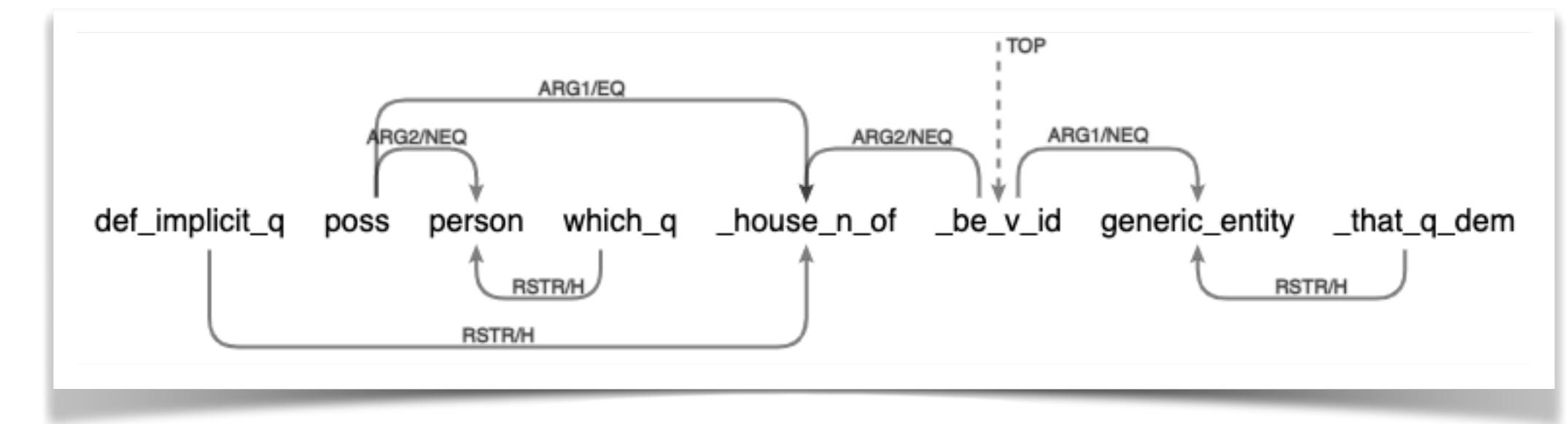
- Predicting people's behavior based on how they talk and/or write:
 - Advertisement, news, **movie** suggestions, etc.
 - IMDB reviews dataset
 - Each review is **labeled** as “positive” or “negative”
- “NLP is a branch of data science...”
 - maybe... **if** data science is a discipline!
 - Atuomatic speech recognition?
 - Machine translation?



<https://www.translatemedia.com/translation-blog/machine-translation-multilingual-sentiment-analysis-projects/>

Raw data in NLP

- Goal: **Automation**
 - Reduce/eliminate the need in annotation
 - Pros and Cons?
 - Cheaper products and services
 - Fewer jobs
- Despite the goal, NLP **relies** on annotation:
 - For training some machine learning algorithms
 - For **evaluation**



A dependency graph

Questions?