2. What is data science? LING 471

Today's goals

- Define data science (and/or explain why it is hard to define)
- Define a dataset and identify linguistic data
- Define and recognize a data statement in a dataset
- Describe what linguistic annotations are used for in NLP
- Run a program in VSCode

UW Data Science minor

About the Minor

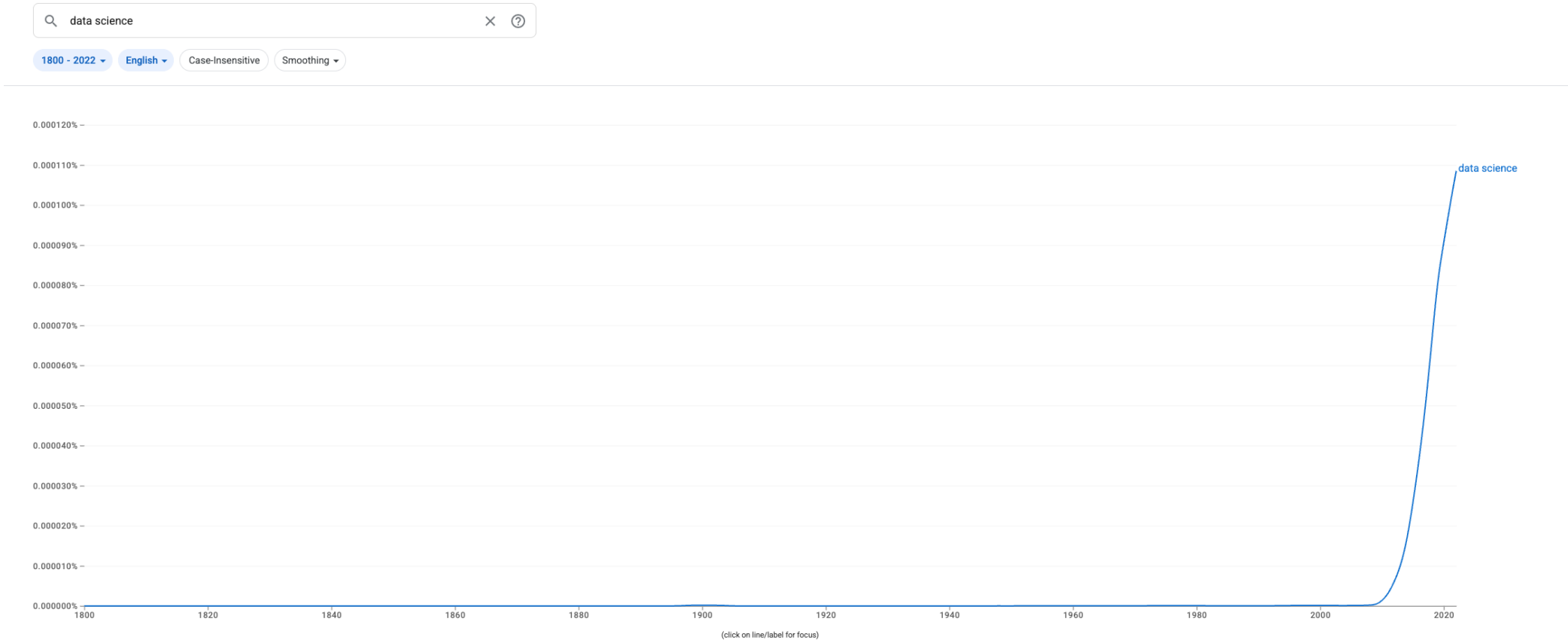
Students in the Data Science minor will gain literacy and fluency in data science methods and understand their implications for society and the world. This minor helps students leverage familiarity with data science in fields outside of data science, and gain skills and fluency to work with data in their major domain of study.

Outcomes

- Help students to leverage familiarity with data science in fields outside of data science.
- Help students gain skills and fluency to work with data in their major domain of study.
- Set students up for success in emergent "translator" roles on heterogeneous teams comprised of data scientists and domain experts. For example, they will be poised to provide expertise in policy, design, social theory, ethics, etc. on a data science team.
- Serve as a stepping stone toward acquiring more advanced skills and degrees in data science.

From: https://dataminor.uw.edu/about-us/

Historical usage of “data science” construction

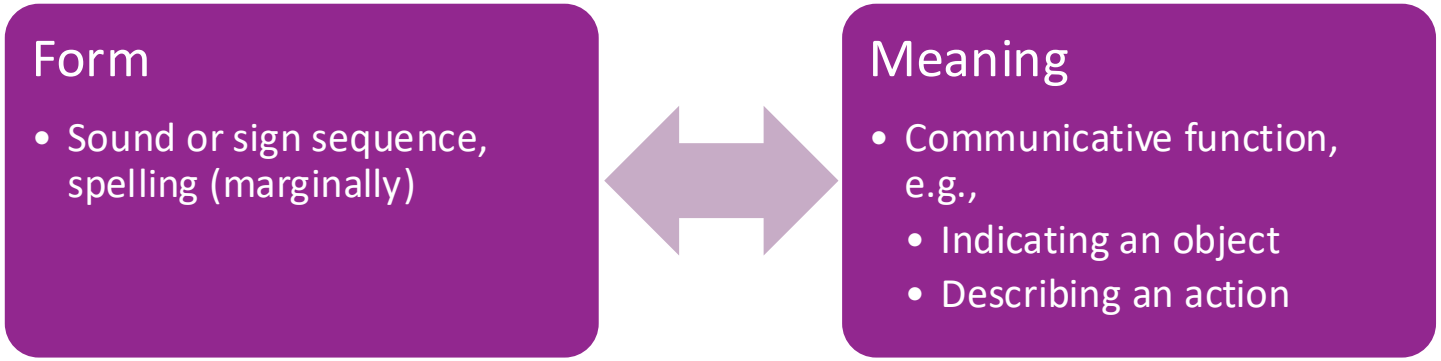


Try it out!

- Break into groups of 3-5
  - Play with Google Books Ngram Viewer
  - Pick 2-3 words and observe the patterns
  - Discuss with class

A bit of linguistics: The semantics of buzzwords

- What exactly is a word/phrase?
  - We have an intuitive sense, but actually hard to define
- Let's go with a form-meaning pair
- Frequently, meaning is not directly related to constituent parts, e.g., *understand*
  - Noncompositionality



Buzzwords (cont.)

- Languages changes! Meaning changes over time across usage communities
- Buzzwords are a particular example
  - Many have or used to have technical meanings (in the right communicative context)
    - *artificial intelligence*, *synergy*
  - Meaning changed from something specific to invoking a general and popular concept, often to make speaker appear fashionable
- **So, *data science*… buzzword?**

Examples of data science

- Statistical analysis of data
- Predicting something, e.g.,
  - Human behavior
  - Impact of technology or surgery on people
- Retrospective analyses
  - What happened in the past? Can it predict the future?
- *Predict* here means ‘best possible guess’, not ‘clairvoyant vision’

Statistical analysis

- Data science relies on **statistics**
- Statistics handles **quantitative** data
- What about **qualitative** data?
  - Can be analyzed if it can be made quantitative
    - Keep in mind when we start working with text…

What do we need to do data science?

- Datasets (big ones!)
- Domain expertise
  - E.g., linguistics, speech science
- Statistics
- Programming
- Visualization and communication
- Machine learning, deep learning?
  - Yes! These are forms of statistical learning

Data in linguistics

- Recorded speech/sign and text
  - Why is text last? Most languages are not written!
- Is text more structured than speech?
- Is text easier to work with than speech?
- What does it mean to focus on text over speech?

Linguistic data and data science

- When someone refers to data science, they often mean lots of data
  - Maybe even “big data” (another buzzword…)
- Which areas of linguistics have lots of data?
- What is “a lot” or “big”?
  - GPT-4 was trained on +10 trillion tokens
  - Ultimately, it depends on your analysis method…

Where is data science used in linguistics?

- Corpus linguistics (analyzing large corpora of newspapers?)
- Sociolinguistics (analyzing Twitter data?)
- Phonetics (classifying depression from speech?)
- Psycholinguistics (analyzing 500K+ responses to visual or auditory stimuli?)
- Syntax (parsing sentences?)
- Morphology (tagging parts of speech in large bodies of text?)
- Basically… **Everywhere!** (Depending on effort level)

Datasets

- To be tautological: a set of data
  - But, this is actually an important general definition
  - Any data point is a dataset
- In practice, a dataset needs to be workable; consider size, annotations, structure, consistency, noise
- Some datasets are used perpetually
  - IMDB reviews dataset (get used to this phrase…)
- Data sets are an important contribution to science

Datasets (cont.)

- Datasets have **provenance**
- Language datasets are created by people (and people vary!)
- Provenance of a dataset biases it (e.g., YouTube comments and flags)
  - Perhaps you can balance it out by adding additional data
  - But, there is no such thing as an unbiased system

LGBT community anger over YouTube restrictions which make their videos invisible

#YouTubeIsOverParty trends on Twitter after users say videos referencing same-sex relationships are being filtered out

Data statements (Bender and Friedman, 2019)

- Reading assigned for April 9
  - We're discussing it now because it is relevant for Assignment 1
- A **practice** of stating facts about the data set used to **train** a system
  - Training is an ML concept
  - System trained or learns by getting feedback on how well it did a task, w.r.t. the data

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

Data statements (cont.)

- Schema
  - Curation rationale
  - Speaker demographics
  - Annotator demographics
  - Speech situation
  - Text genre
  - Recording quality
- Based on reporting for psych studies (and similar)

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

Annotated data

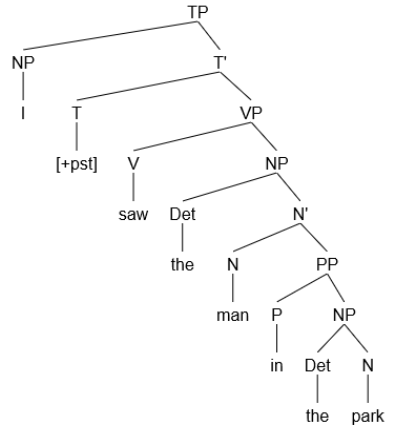
- Annotation types
  - Speaker characteristics (e.g., region)
  - L1/L2 speech
  - Part of speech
  - Named entity
  - Sentence structure
  - Discourse structure
- Some of these are what we want a system to predict

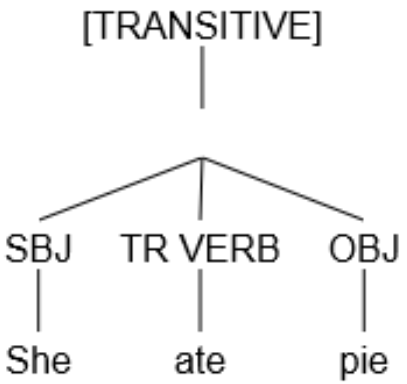
This/DET is/V an/DET annotated/ADJ sentence/N ./PUNC

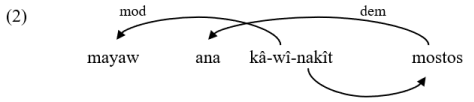
Annotated data in linguistics

- Recorded speech and text associated with (socio)linguistic variables
  - Gender, age, region, part of speech
- Interlinearized glossed text
  - Linguistic **analysis** and **annotation**
- Syntax trees?
- Comp ling emphasizes **raw** data
  - NLP is a comp sci discipline
  - Deep learning uses more raw data

(17) zono ja vova-ŋga
yesterday 1.SG.PAST speak-PROG
'Yesterday, I was speaking'







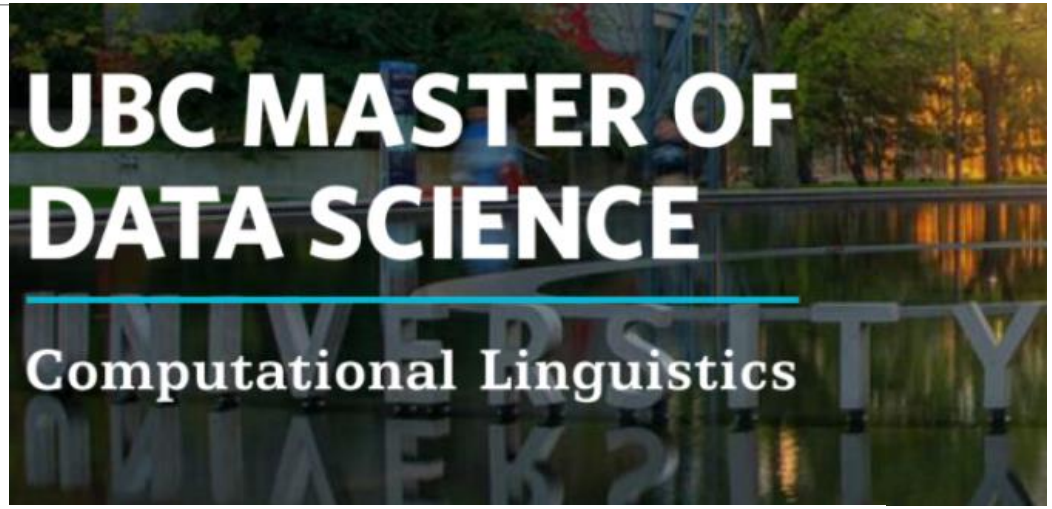
Dependency analysis of Plains Cree, credit Katie Schmirler

Computational linguistics vs. NLP

- Are they the same? For some, yes
- For others
  - CL: What can we learn about language using computational methods?
    - E.g., How can we simulate language acquisition in children?
  - NLP: What can we learn about the world through language data?
    - E.g., Do someone's tweets predict their mental health?
- Where does data science fit in?

NLP and data science

- Predicting people's behavior based on how they talk or write
  - Webads, news, movie suggestions
- "NLP is a branch of data science…"
  - Maybe… if data science is a coherent discipline
  - Automatic speech recognition?
  - Machine translation?

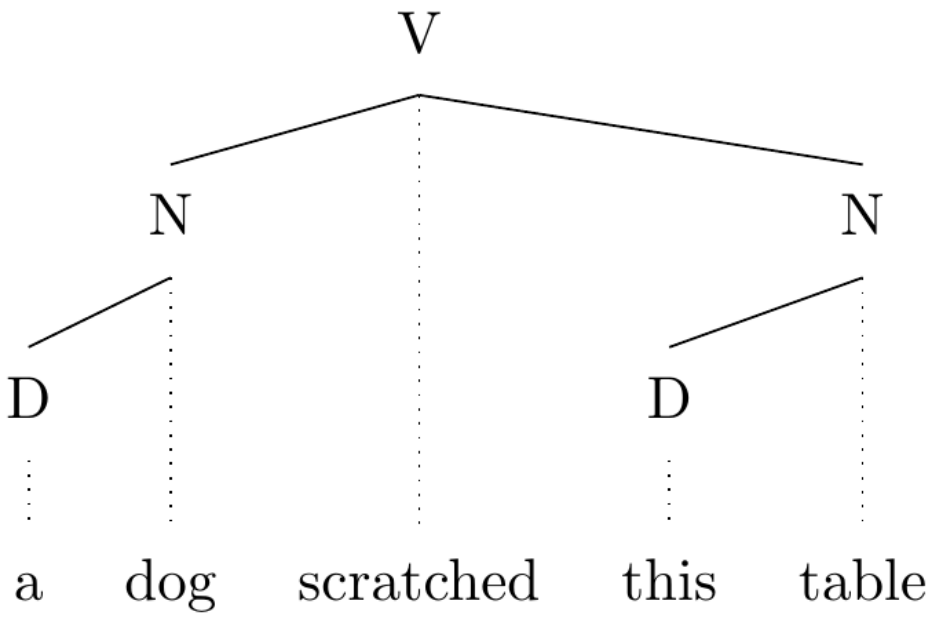


Block 5 (4 weeks + 1 week break, 4 credits)

- **Advanced Corpus Linguistics** | COLX 523
- **Computational Morphology** | COLX 525
- **Machine Translation** | COLX 531
- **Unsupervised Learning** | DSCI 563

Raw data in NLP

- Goal: **automation**
  - Reduce or eliminate need for annotation
- Pro: cheaper products and services
- Con: fewer jobs, requires ENORMOUS amounts of data
- Despite goal, NLP relies on annotations for **training** and **evaluation**

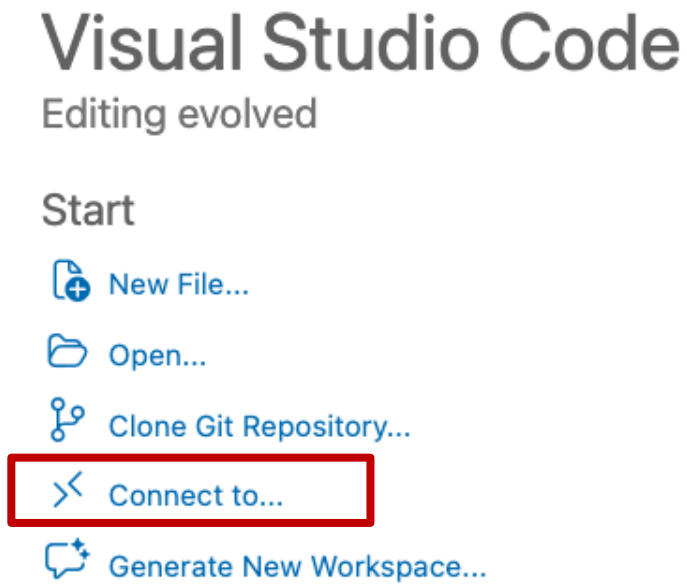


Some VSCode basics

A FIRST PROGRAM IN VSCODE

Writing your first program

- If you created a patas account during last class
  - Try accessing it



Writing your first program

- Type
  - print("Hello LING471")
- Hit the triangle on the top right

Reflection

- For some of you, that code may have been very basic
- For others of you, that might be the first time you have ever seen any lines of code
- Keep today in mind as we move through the quarter
  - We will gradually incorporate more complex programming and more complex language questions as we progress