Course Overview

Ling 250/450: Data Science for Linguistics

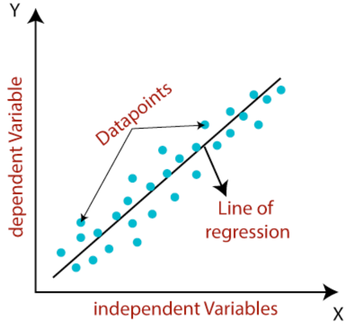
C.M. Downey

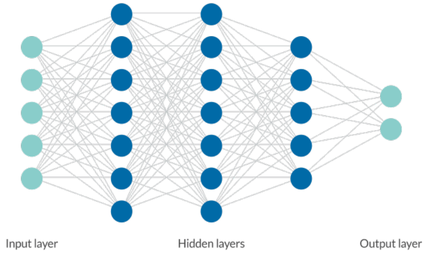
Spring 2025

What is Data Science anyway?

What is Data Science?

- Data Science is (in my opinion) somewhat **poorly defined**
- Things people will often refer to as Data Science:
  - **Collecting**, **manipulating**, and **cleaning** data
  - Running **statistical tests** and models
  - Training and using **machine learning** models
  - **Interpreting** meaning or trends from data
  - **Visualizing** data in useful and novel ways



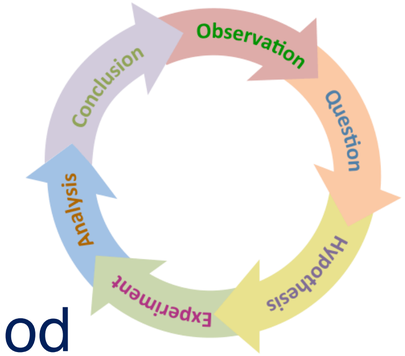


- All of these skills are used in **many other fields!** (And most pre-date “Data Science”)
- Data Scientists might only specialize in **one or a few** of these skills

Is it actually science?

- Again in my opinion, “Data Science” might be a **bit of a misnomer**
- These techniques are **often** but **not always** applied to science
  - An important distinction is that science follows the **Scientific Method**, and mostly acquires knowledge through **hypothesis testing**
  - Most of the hallmarks of “Data Science” can be used **without** following the Scientific Method
- While the name “Data Science” is here to stay, it might (sometimes) be more accurate to talk about **data {analysis | exploration | engineering}**

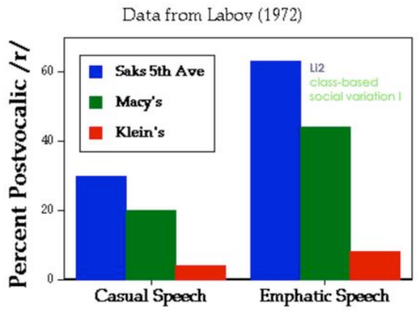
“Data Science” and Science



- Data Science techniques can be an integral part of the Scientific Method
  - Prior to experimentation: **explore**, **analyze**, and **visualize** patterns in data and previous experimental results in order to form a new **testable hypothesis**
  - During experimentation: **collect**, **clean**, and **organize** data that allows you to properly **test your hypothesis**, ideally with **statistical methods**
  - After experimentation: **analyze** results to tell if your hypothesis is supported. Create **visualizations** to share your results. **Organize** your data for dissemination and experimental **replication**
- We will read later about the importance of **separating exploratory studies** and **hypothesis testing** (but we will engage in both in this course!)

Data Science and Linguistics

- Linguistics (mostly) acts as the **science of language**
  - Thus, DS can be used to **explore and test hypotheses** about language and how it works
- Some linguistic sub-fields have **long embraced DS methods**
  - e.g. Sociolinguistics, Psycholinguistics, and “Lab”/Experimental Phonetics



- Others have been historically **more resistant** (especially Chomsky)
  - “One’s ability to produce and recognize grammatical utterances is **not based on notions of statistical approximations** and the like” (Chomsky 1957)
  - Syntax and Generative Grammar have tended towards methods more similar to **math** and **philosophy** than (experimental) science

$$
\begin{aligned}
\langle\text{SENTENCE}\rangle &\rightarrow \langle\text{NOUN-PHRASE}\rangle\langle\text{VERB-PHRASE}\rangle \\
\langle\text{NOUN-PHRASE}\rangle &\rightarrow \langle\text{CMPLX-NOUN}\rangle \mid \langle\text{CMPLX-NOUN}\rangle\langle\text{PREP-PHRASE}\rangle \\
\langle\text{VERB-PHRASE}\rangle &\rightarrow \langle\text{CMPLX-VERB}\rangle \mid \langle\text{CMPLX-VERB}\rangle\langle\text{PREP-PHRASE}\rangle \\
\langle\text{PREP-PHRASE}\rangle &\rightarrow \langle\text{PREP}\rangle\langle\text{CMPLX-NOUN}\rangle \\
\langle\text{CMPLX-NOUN}\rangle &\rightarrow \langle\text{ARTICLE}\rangle\langle\text{NOUN}\rangle \\
\langle\text{CMPLX-VERB}\rangle &\rightarrow \langle\text{VERB}\rangle \mid \langle\text{VERB}\rangle\langle\text{NOUN-PHRASE}\rangle \\
\langle\text{ARTICLE}\rangle &\rightarrow \texttt{a} \mid \texttt{the} \\
\langle\text{NOUN}\rangle &\rightarrow \texttt{boy} \mid \texttt{girl} \mid \texttt{flower} \\
\langle\text{VERB}\rangle &\rightarrow \texttt{touches} \mid \texttt{likes} \mid \texttt{sees} \\
\langle\text{PREP}\rangle &\rightarrow \texttt{with}
\end{aligned}
$$

Rationalism and Empiricism in Linguistics

- Chomsky's school of Linguistics uses **rationalist** techniques for insight on language, while experimental and data-driven methods are considered **empiricist**. (Very) roughly:
  - Rationalism: knowledge is obtained through **introspection** and **logical processes** like deduction and induction
  - Empiricism: knowledge is obtained through observation of **external experience**, especially the results of **controlled experiments**
- Note that these philosophies are **not mutually exclusive**, but researchers have historically tended to argue that one may be more reliable than the other

Why be an empiricist linguist?

- Rationalist approaches to Linguistics often rely on the **introspective judgements** from just one or a few people
  - e.g. “is this sentence grammatical?”, “do these words mean the same thing?”
- Most linguists eventually have to confront the major insight of Sociolinguistics: **language varies extensively**
  - (Between people, communities, genres, conversation partners, etc.)
  - The same person might have different judgements at different times!
- This makes it questionable what the introspection of **any one person** can tell us about **language in general**
- Empirical techniques are inherently better suited for **generalizing from variable data**

Ex: Categorical grammar rules

- A key technique of Generative syntax is to use **intuitive judgements** to figure out which sentences are **"grammatical" vs. "ungrammatical"**
- Examples from Pollard and Sag (1994)
  - We consider Kim to be an acceptable candidate
  - We consider Kim an acceptable candidate
  - We consider Kim quite acceptable
  - We consider Kim among the most acceptable candidates
  - *We consider Kim as an acceptable candidate
  - *We consider Kim as quite acceptable
  - *We consider Kim as among the most acceptable candidates
  - *?We consider Kim as being among the most acceptable candidates

Ex: Categorical grammar rules

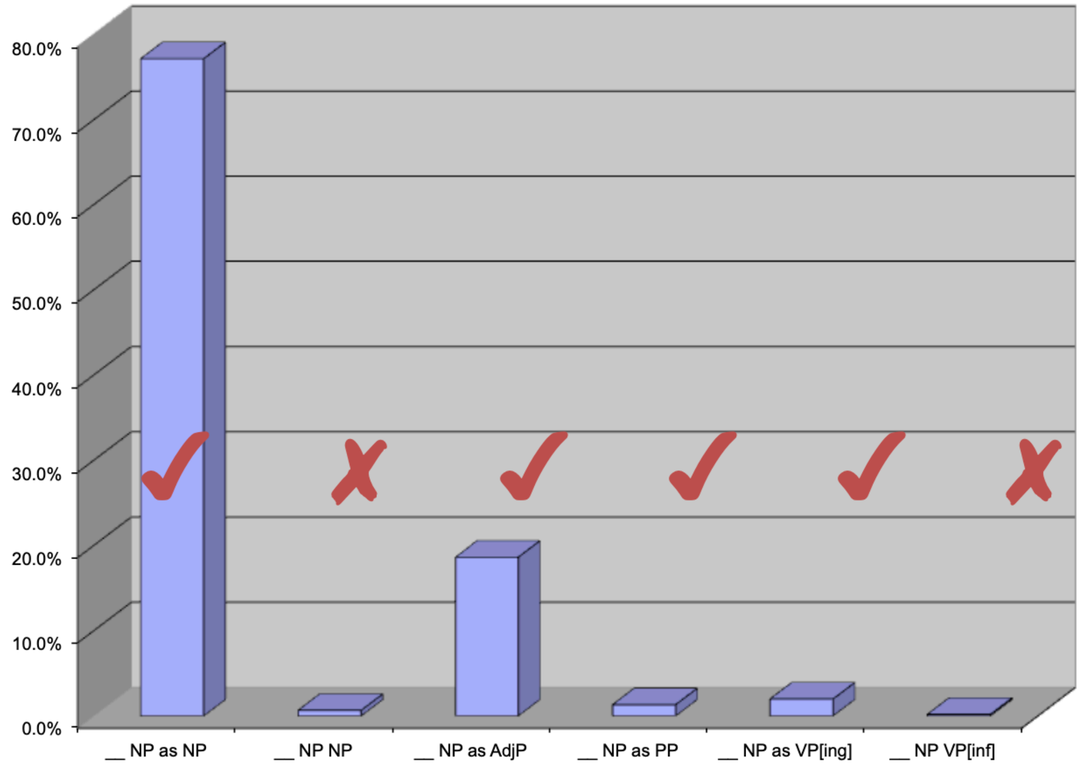
- Actual examples from the *New York Times*:
  - The boys consider her as family and she participates in everything we do
  - I don't consider it as something that gives me great concern
  - We consider that as part of the job
  - He considers them as having championship potential
  - Culturally, the Croats consider themselves as belonging to the West
- Investigating language from a **data-driven** approach reveals that it is **much more flexible** than the introspection-driven account

Problems with categorical theories

- **They tend to claim too much**
  - They place a hard categorical boundary of grammaticality, where really there is a **fuzzy edge**, determined by many **conflicting constraints**
- **They tend to explain too little**
  - They say very little about the **soft constraints** that explain how people **choose to say things** in given situations. These types of soft constraints have long been of interest to e.g. sociolinguists

Probability distribution for *regard*

- It might be more informative to construct a **probability distribution** over different variants of a sentence
- This distribution is calculated from a large **corpus** (dataset) of documented language “in the wild”
- We will discuss various standard corpora and **corpus linguistics methods** in this course



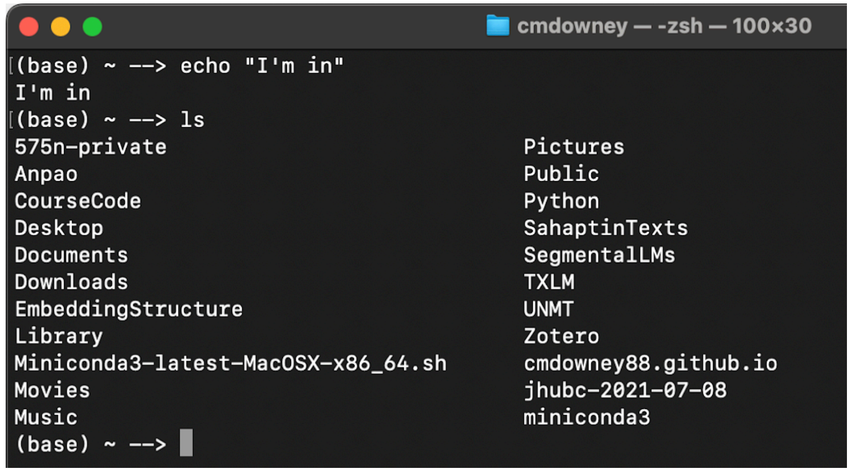
About this course

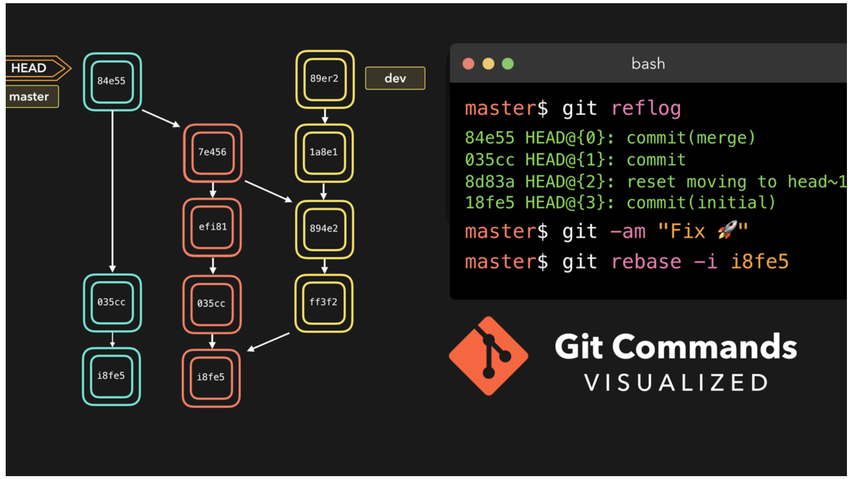
Course goals

- This course is meant to serve as a **tools-based** introduction to **data-driven methods** in Linguistics
- The tools we introduce will help you:
  - **Collect, clean, organize, and understand** linguistic data
  - **Gain access** to existing corpora and understand **annotation conventions**
  - **Visualize** patterns in language, and **effectively communicate** your findings
  - **Test linguistic hypotheses** with statistical robustness
- The course will also serve as a gentle introduction to the **Python** and **R** programming languages

Larger theme: Basic programming tools

- Covered in the first part of the course
- Using the **command line**
  - Will learn how to interface with computers in a new and **more powerful** way
  - Is sometimes the **only way** to interface with powerful servers for data science!
- **Version control** with **Git and Github**
  - The standard way to manage **collaboration, edits,** and **differing versions** of software





Larger theme: Data manipulation in Python

- Very basic introduction to **Python**
  - The dominant language for **Science** and **Data Engineering**
  - Easy language to **get started**
- Introduction to relevant **Python libraries**
  - Regular Expressions
  - Natural Language Tool Kit
  - Pandas

```
>>> nltk.corpus.sinica_treebank.tagged_words()
[('ä', 'Neu'), ('åæ', 'Nad'), ('åç', 'Nba'), ...]
>>> nltk.corpus.indian.tagged_words()
[('মহিষের', 'NN'), ('সন্তান', 'NN'), (':', 'SYM'), ...]
>>> nltk.corpus.mac_morpho.tagged_words()
[('Jersei', 'N'), ('atinge', 'V'), ('m\xe9dia', 'N'), ...]
>>> nltk.corpus.conll2002.tagged_words()
[('Sao', 'NC'), ('Paulo', 'VMI'), ('(', 'Fpa'), ...]
>>> nltk.corpus.cess_cat.tagged_words()
[('El', 'da0ms0'), ('Tribunal_Suprem', 'np0000o'), ...]
```

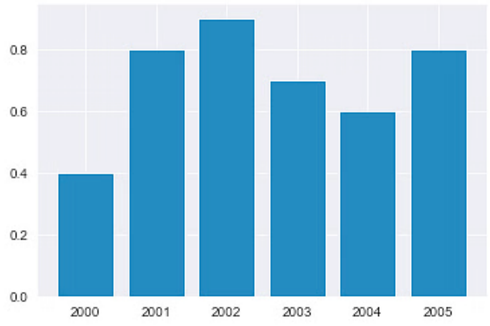
```
years = range(2000, 2006)
apples = [0.35, 0.6, 0.9, 0.8, 0.65, 0.8]
oranges = [0.4, 0.8, 0.9, 0.7, 0.6, 0.8]
```

```
plt.bar(years, oranges)

plt.xlabel('Year')
plt.ylabel('Yield (tons per hectare)')

plt.title("Crop Yields in Kanto")
```

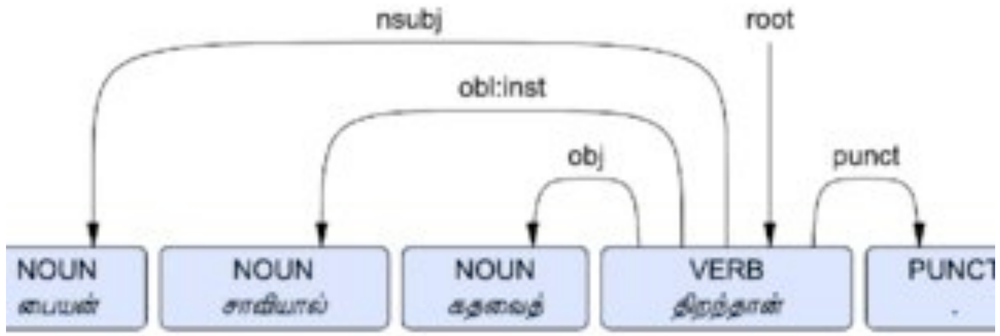
```
<BarContainer object of 6 artists>
```



Larger theme: Linguistic annotation

- Introduction to **common annotation schemes** for Linguistic data
- Discussion of why it is important to have **machine-readable** data
- Discussions of the **challenges** and **reliability** of annotation

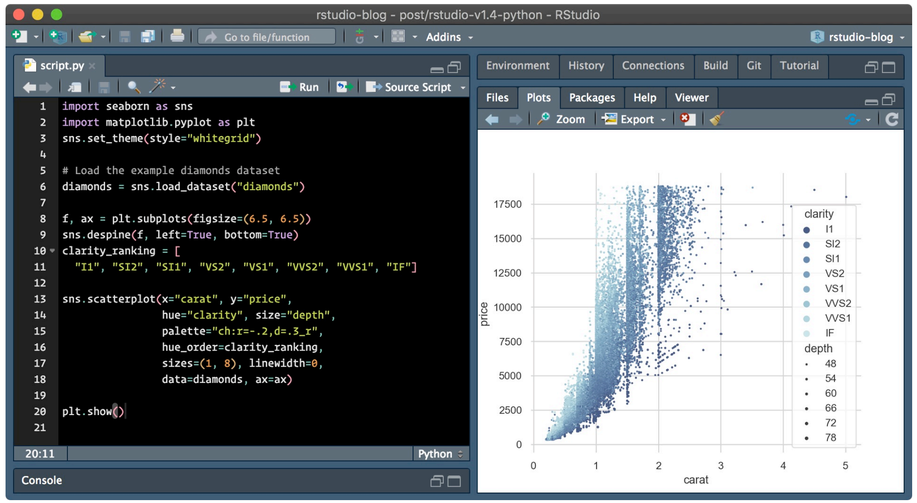
POS	XPOS	Morph-features
NOUN	_	Case=Nom|Gender=Masc|Number=Sing|Person=3
NOUN	_	Case=Ins|Gender=Neut|Number=Sing|Person=3
NOUN	_	Case=Acc|Gender=Neut|Number=Sing|Person=3
VERB	_	Gender=Masc|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin|Voice=Act
PUNCT	_	PunctType=Peri

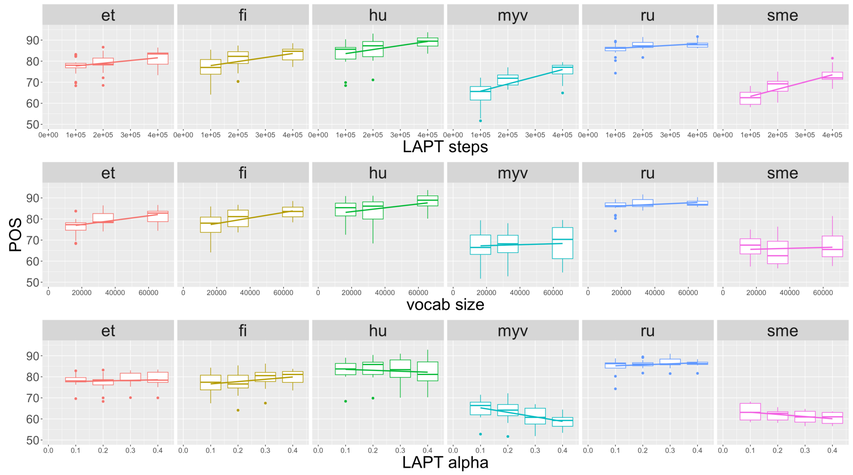


ure 03: A dependency graph for the annotation given in Tabl

Larger theme: R for testing and visualization

- Introduction to **R**
  - Most commonly used language for **statistical testing**
  - Also has the best libraries for **data visualization**
- **ggplot**
  - Library for developing **professional data visualizations**
  - Tricky at first, but gives the best results





What this course is **not**

- This is not a **machine learning** course
- This is not a **Natural Language Processing** (NLP) course
- This is not an **algorithms/data-structures** course
- This is not a **software engineering** course
- This course is not about **Large Language Models**
  - However, you can study the **output** of LLMs if you want

Policies and Organization

Basic Information

- Instructor
  - C.M. Downey
  - Assistant Professor in Linguistics & Data Science
- Meeting time
  - Monday/Wednesday 9:00-10:15am
  - **This week only** we meet **Wednesday/Friday**
- Prerequisites
  - Intro to Linguistics (LING 110)

Online Resources

- Blackboard
  - Announcements
  - Discussion boards
  - Homework submission
- Course website (cmdowney88.github.io/teaching/ling250/spring2025)
  - Detailed course schedule
  - Syllabus
  - Slides (posted after class)
  - Important links and documents

Attendance

- Because of the hands-on nature of this course, **attendance is mandatory**
- I will **keep track of attendance**, however:
  - You can be absent from **up to four** sessions without penalty
  - These absences can be for **any reason** (e.g. illness, travel, catching up on other courses), and you **do not have to approve it with me**
  - Absences beyond these four **count against your attendance grade** (5%)
- You may **request** additional excused absences for important exceptions (e.g. religious obligations, interviews, jury duty)

Participation

- Participation is also **part of your final grade** (5%)
- A large part of class time will be devoted to **hands-on demonstrations** of various software tools
  - You are expected to **follow along** with the demo on your **own device**
  - This means you will need to bring a **laptop** or **other appropriate device** to class (a tablet might work if it is running Windows; an iPad probably won't work)
- **Before Friday:** follow the **setup guide** posted on Blackboard and the course website to get your device ready

Course Work

- Homework (40%): between 6 and 8 assignments involving **written questions, practice problems, and small coding projects**
  - Students may work on these **collaboratively**, but each must **submit their own work**
- Midterm Project (20%): a **larger coding project**, in which everyone will work on the same data/problem
- Final Project (30%): a term research project which will be conducted with **data of each student's choosing** (within reason)

Deadlines and Late Work

- Unless specified otherwise, all assignments are due **11pm Eastern**
- Work submitted after the deadline will incur a penalty
  - Up to **1 hour** late: -5%
  - Up to **24 hours** late: -10%
  - Up to **48 hours** late: -20%
  - **> 48 hours** late: not graded (0 for the assignment)
- Please **feel free to request extensions**, but you must request it **before the deadline**
  - I will be more willing to grant extensions the longer before the deadline it is requested

Academic Honesty

- Homeworks can be completed collaboratively, but the on the **Midterm Project**, only **minimal collaboration** is allowed
- The use of **Large Language Models / Chatbots** for this class is **allowed**, however:
  - It is **only allowed for code** (**not** essays or short answers)
  - You are **fully responsible** for the success our failure of your code
  - We will talk more about **responsible practices** for using LLMs later

“Required” Textbooks

- You will sometimes be assigned **required readings** before class, but these will always be drawn from **free online sources**
  - Some are available online as PDFs, others through the **UR Library**
- **Links** to these sources can be found here and on the **course website:**
  - *Natural Language Processing with Python*
  - *Data Science from Scratch*
  - *Speech and Language Processing*

Questions/Discussion?