# CS 333: Natural Language Processing

Fall 2025

Prof. Carolyn Anderson

Wellesley College

# My Work in NLP

## Text-to-Code Models

Knowledge Transfer from High-Resource to Low-Resource Programming Languages for Code LLMs
F Cassano, J Gouwar, F Lucchetti, C Schlesinger, CJ Anderson, ...
arXiv preprint arXiv:2308.09895

StarCoder: may the source be with you!
R Li, LB Allal, Y Zi, N Muennighoff, D Kocetkov, C Mou, M Marone, C Akiki, ...
arXiv preprint arXiv:2305.06161

StudentEval: A Benchmark of Student-Written Prompts for Large Language Models of Code
HML Babe, S Nguyen, Y Zi, A Guha, MQ Feldman, CJ Anderson
arXiv preprint arXiv:2306.04556

MultiPL-E: a scalable and polyglot approach to benchmarking neural code generation
F Cassano, J Gouwar, D Nguyen, S Nguyen, L Phipps-Costin, D Pinckney, ...
IEEE Transactions on Software Engineering

Non-expert programmers in the generative AI future
MQ Feldman, CJ Anderson
Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer …

## LLM Evaluation

Solving and Generating NPR Sunday Puzzles with Large Language Models
J Zhao, CJ Anderson
arXiv preprint arXiv:2306.12255

Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis) information in a Large Language Model
S Nguyen, CJ Anderson
Proceedings of the Society for Computation in Linguistics 6 (1), 400-402

Phd knowledge not required: A reasoning challenge for large language models
Z Wu, F Lucchetti, A Boruch-Gruszecki, J Zhao, CJ Anderson, J Biswas, ...
arXiv preprint arXiv:2502.01584

## Interdisciplinary Work

GlyphPattern: An Abstract Pattern Recognition Benchmark for Vision-Language Models
Z Wu, Y Kim, CJ Anderson
Association for Computational Linguistics

Components of Character: Exploring the Computational Similarity of Austen's Characters
CJ Anderson
Journal of Data Mining & Digital Humanities, 2025

Tell me everything you know: a conversation update system for the rational speech acts framework
CJ Anderson
Proceedings of the Society for Computation in Linguistics 2021, 244-253

# Course Goal

**To make you into a skilled NLP practitioner who can:**

- Understand and implement core NLP algorithms and models.
- Explore the challenges posed by different aspects of human language.
- Analyze ethical concerns about language technology.
- Complete a series of projects to implement and improve NLP models.

# Natural Language Processing

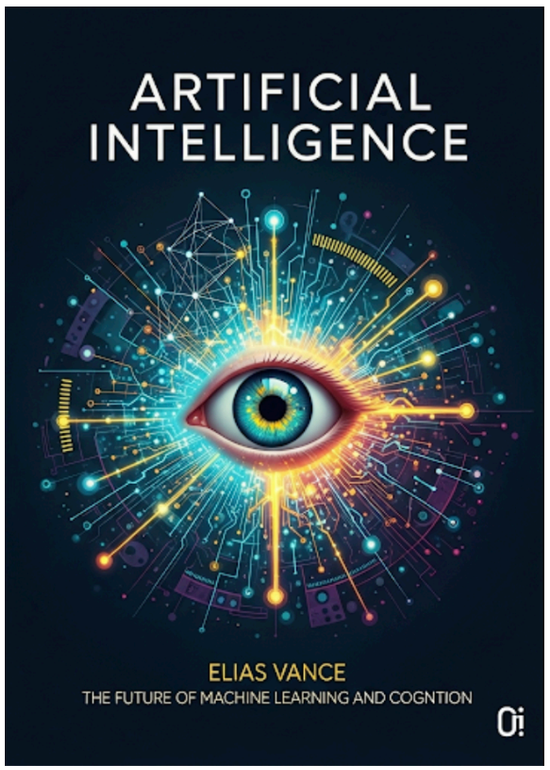
Human language

# Natural Language Processing

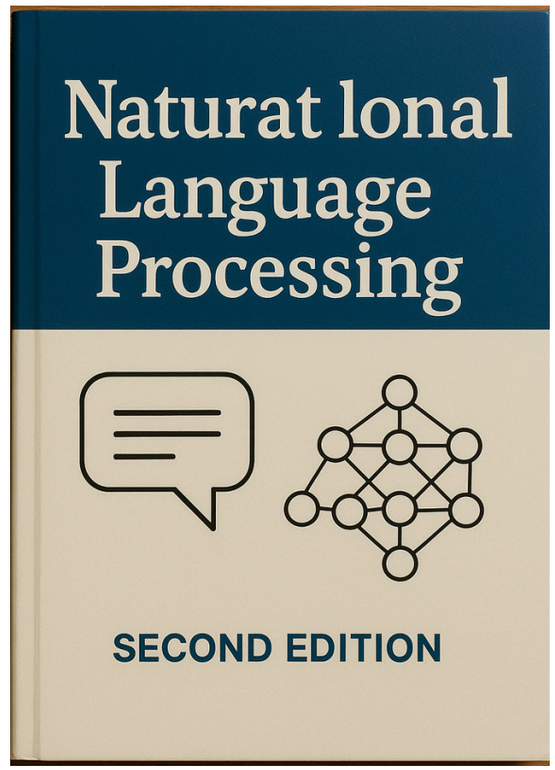
Doing stuff with language data

# Is NLP AI?

## Artificial Intelligence

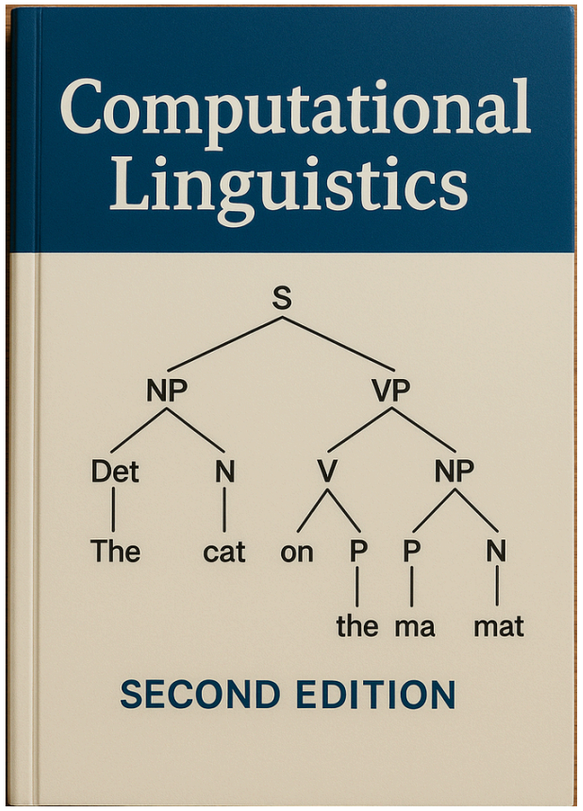



## Natural Language Processing

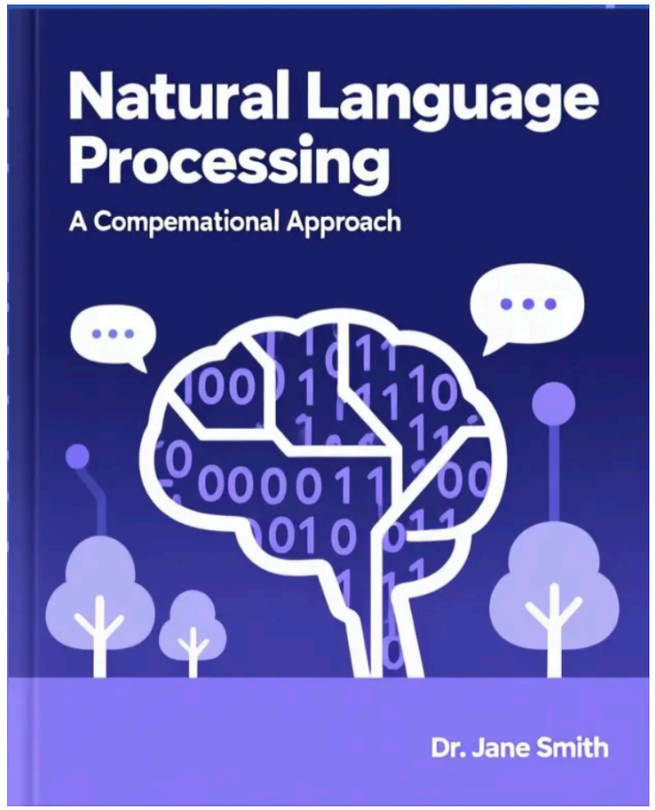


# Is NLP Computational Linguistics?

Computational Linguistics

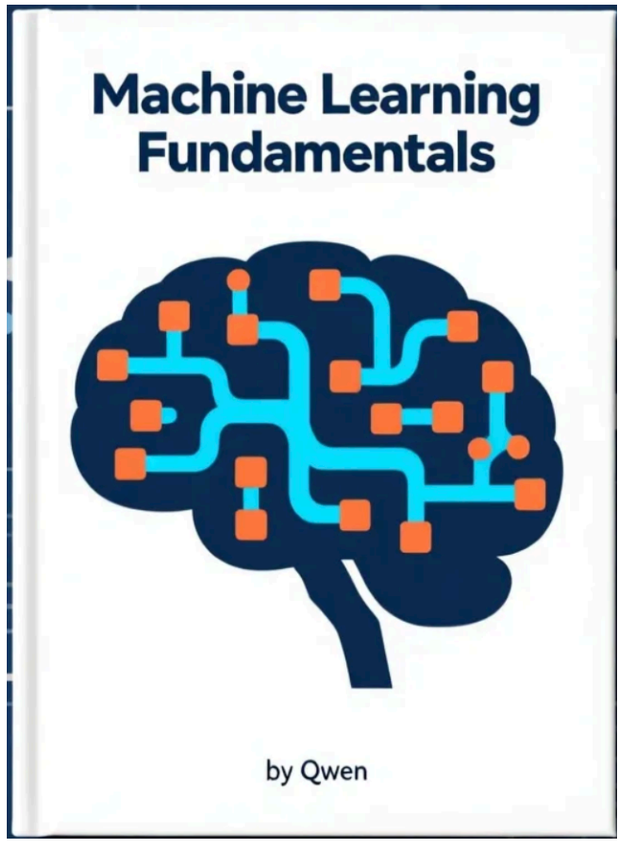



Natural Language Processing

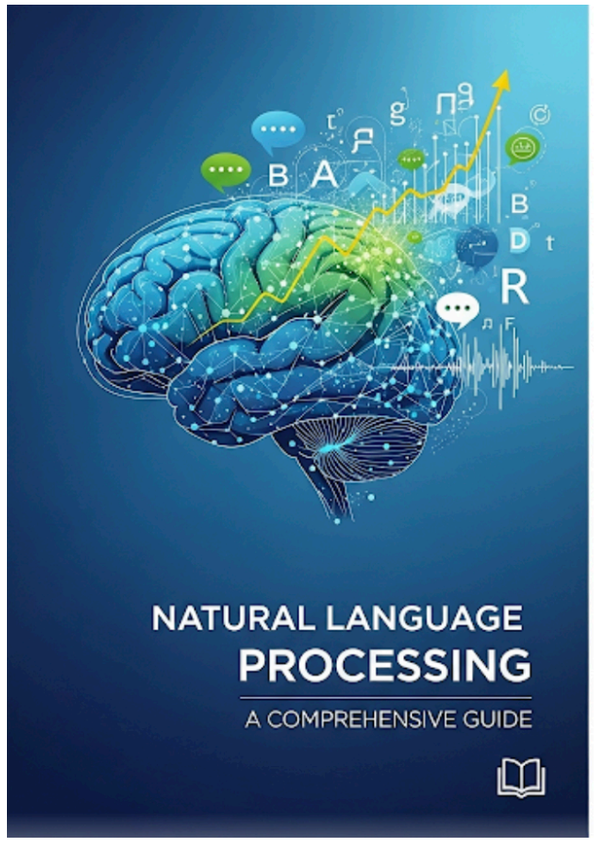


# Is NLP Machine Learning?

Machine Learning



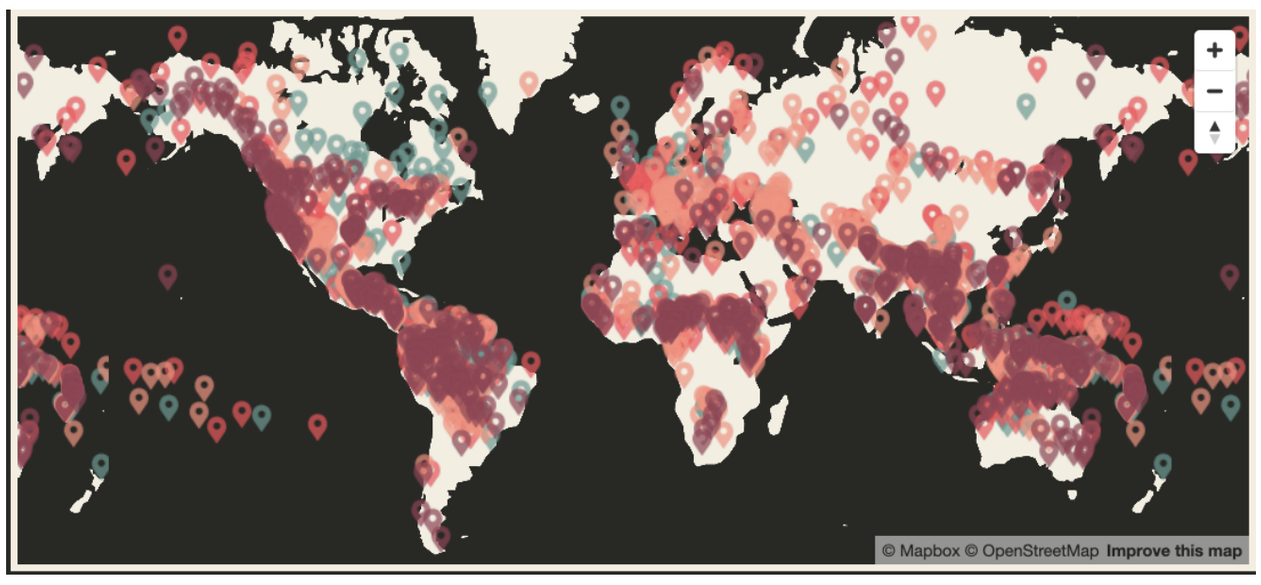

Natural Language Processing



# Natural Language

# Natural Language

- There are around 7000 human languages
- 50% of the world's languages are **endangered**
- Languages can be spoken or signed


https://interactive.howwegettonext.com/endangeredlanguages/

# Layers of Linguistic Representation

# Up Top: Pragmatics


B asks a yes-no question, but A does not respond with yes or no.

# Up Top: Pragmatics


A literally says: it is 7 in the morning.

# Up Top: Pragmatics


A implies: it's way too early for ice cream.

# Up Top: Pragmatics


Pragmatics: the meaning of sequences of sentences.

# Next Layer: Semantics


How do we know what B's question means?

# Next Layer: Semantics


What does the sequence of words *wanna go get ice cream* mean?

# Next Layer: Semantics


Too complicated to explain here-- go take semantics!

# Next Layer: Semantics


What does the sequence of words *it's 7 am* mean?

# Next Layer: Semantics


[[it's 7am]] = NOW(7am)

# Next Layer: Semantics


$[[\text{it's 7am}]]^{c,w} = \text{True if } w_t == \text{7am else False}$

# Next Layer: Semantics


Basically: *it's 7am* is a **function** that takes a **world** and returns true for some worlds and false for others.

# Next Layer: Semantics


Semantics: the **meaning** of a sentence is its **truth conditions** (the conditions under which it is true).

# Middle Layer: Syntax

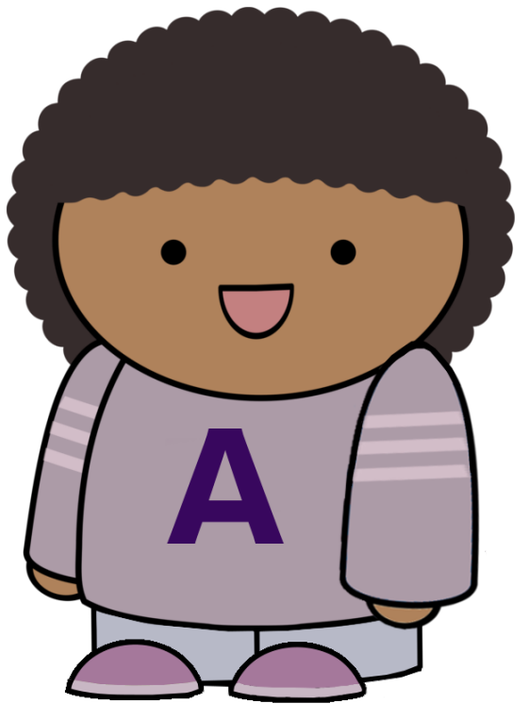

How do we determine the order of the words?

# Middle Layer: Syntax


How do we determine the order of the words?

# Middle Layer: Syntax

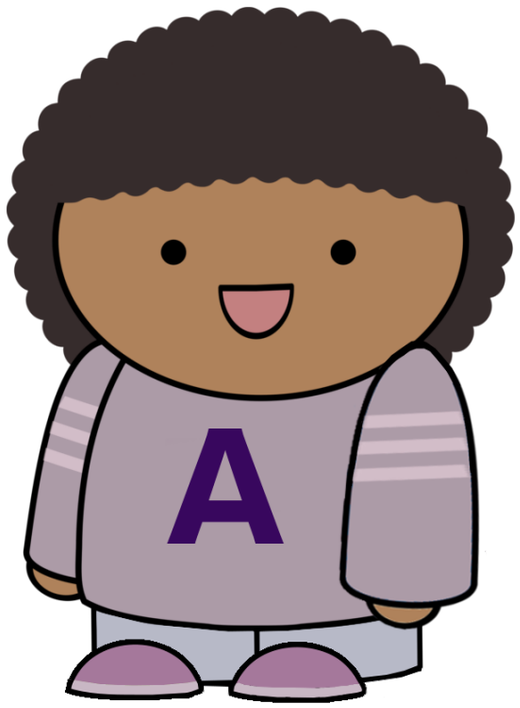

*Must be German...*




How do we determine the order of the words?

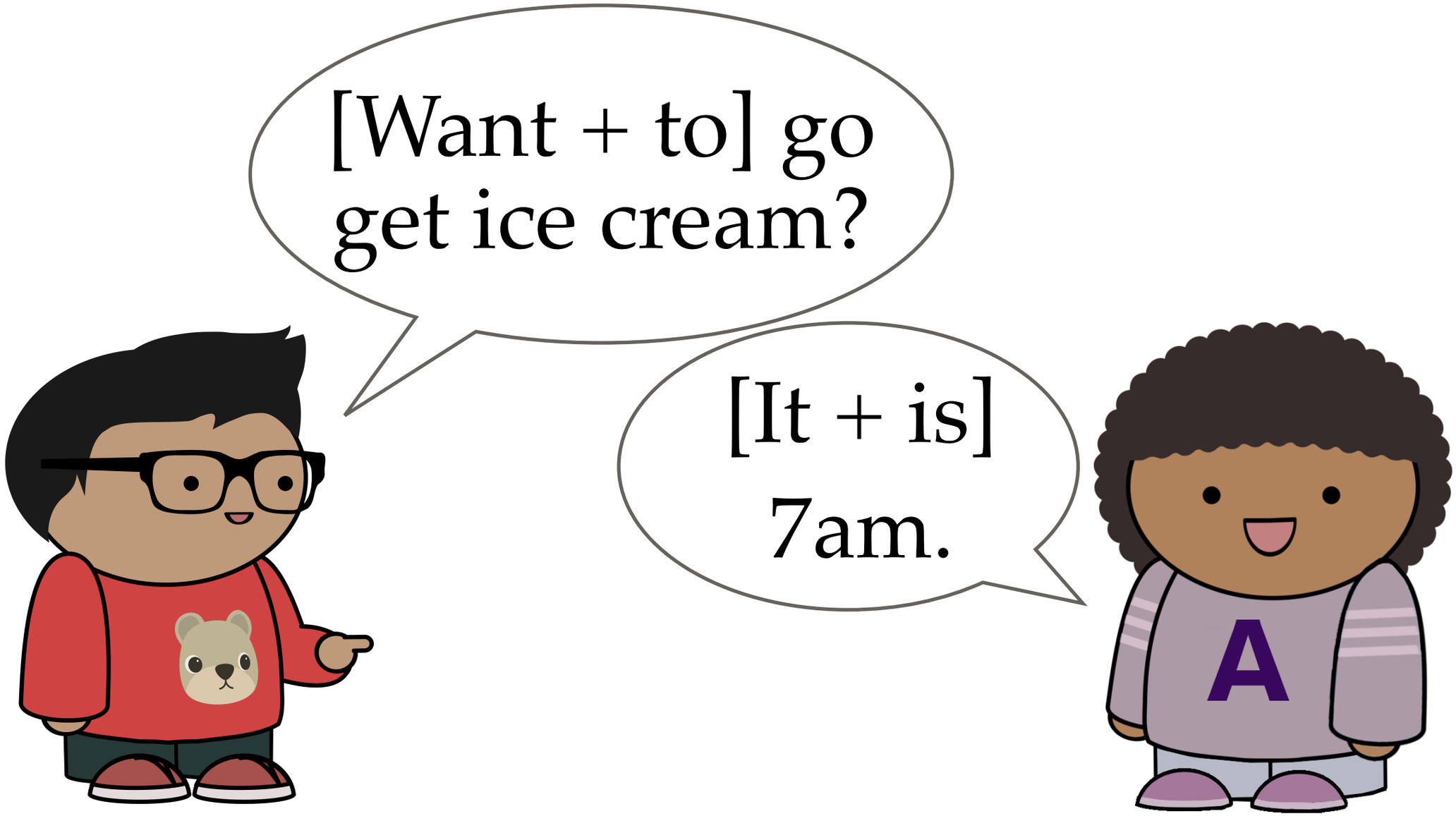
# Middle Layer: Syntax


*Must be German...*




Syntax: the structure of a sentence is determined by a set of language-specific syntactic rules.

# Glue Layer: Morphology




Morphology: the rules that determine how words are formed.

# Lower Layer: Phonology


Phonology: the rules that determine how the sounds/signs of a language are organized

# Lower Layer: Phonology


Phonology: the rules that determine how the sounds/signs of a language are organized

# Foundation: Phonology


Phonetics: how do language users produce the building blocks of language?

# Layers of Linguistic Abstraction

PRAGMATICS — It's 7am ⟹ it's a weird time for ice cream so I don't know how to respond.

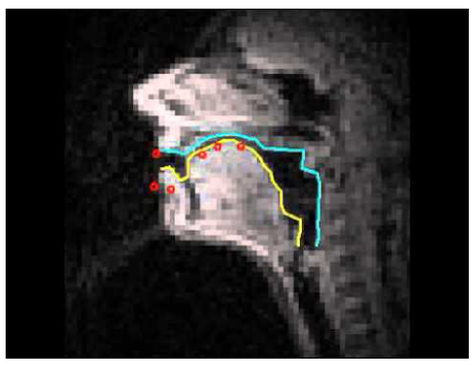
SEMANTICS — [[It's 7am]] → True if now(w) == 7am else False

SYNTAX — {7, it's, am} → It's 7am.

MORPHOLOGY — {7 → 7, [it is] → it's, am → am}

PHONOLOGY — ɪts ˈsɛvən ə ɛm.

PHONETICS



# Natural and Artificial Language Learning

# How do people learn language?

Humans learn language instinctively:

- Language has a **critical acquisition period**
- Language acquisition begins before birth and follows predictable developmental stages
- Humans can't decide **not** to learn language
- Language acquisition does not seem to correlate with intelligence
- All human cultures have language; no other species do
- All human languages are equally expressive

# Example: Child Language Acquisition

## Example 1

cj and ember manning liked

**Gareth Roberts** @garicgymro · 45m

Just overheard from two of my kids:
Osian (5;1): Look how I catched Mickey!
Eirwen (8;2): Do you mean caught?
Osian: … yeah.
Eirwen: But you can keep saying catched!
Osian: Look how I catched him!

## Example 2

# Learning Language

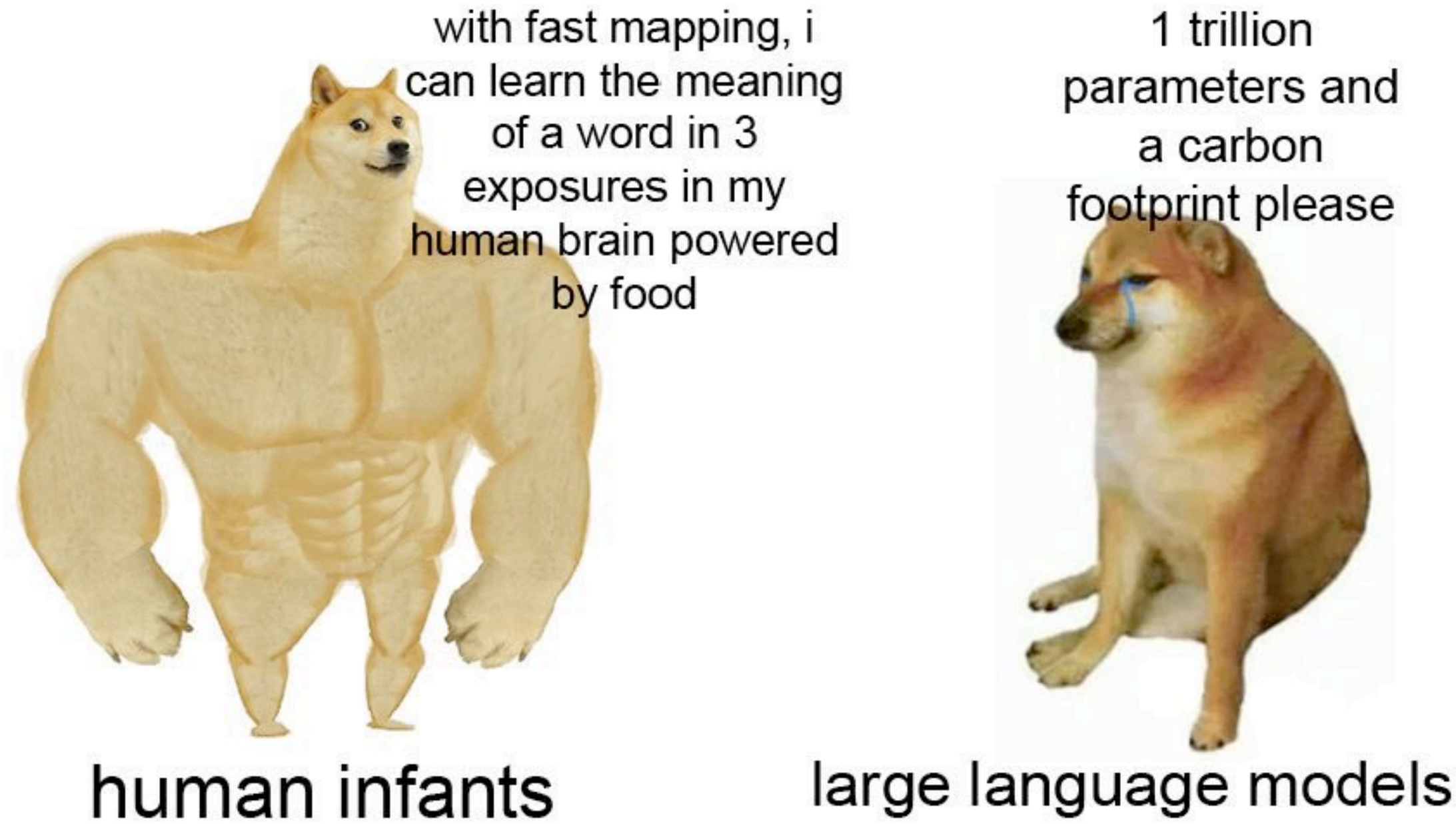



photo credit: Josef Fruehwald

# Practicalities

# Schedule

- Room: SCI L039
- Lecture: 9:55-11:10 on Tuesdays and Fridays
- Assignments are due on **Thursdays at 10 PM**

# Help Hours

- Mondays 4-5:30pm
- Thursdays 4-5
- By appointment (schedule using online calendar)

Come to my help hours to …

- Get help with CS333
- Talk about NLP

# Tutors

Ashley Sheng

Tutor

Caroline Witty

Tutor

# Readings

This course has required weekly readings. Most are from the course textbook: *Speech and Language Processing* by Jurfasky & Martin. The third edition is free online.

All readings are listed on the schedule.

Please finish each week's required reading before coming to class on Tuesday.

# Quizzes

There will be a **quiz** in class **every Tuesday** to test your understanding of the assigned reading.

+ open note, closed computer
+ timed but very brief (1-2 questions)

# Homework will be in Python

Please set up a Python 3.12 virtual environment.

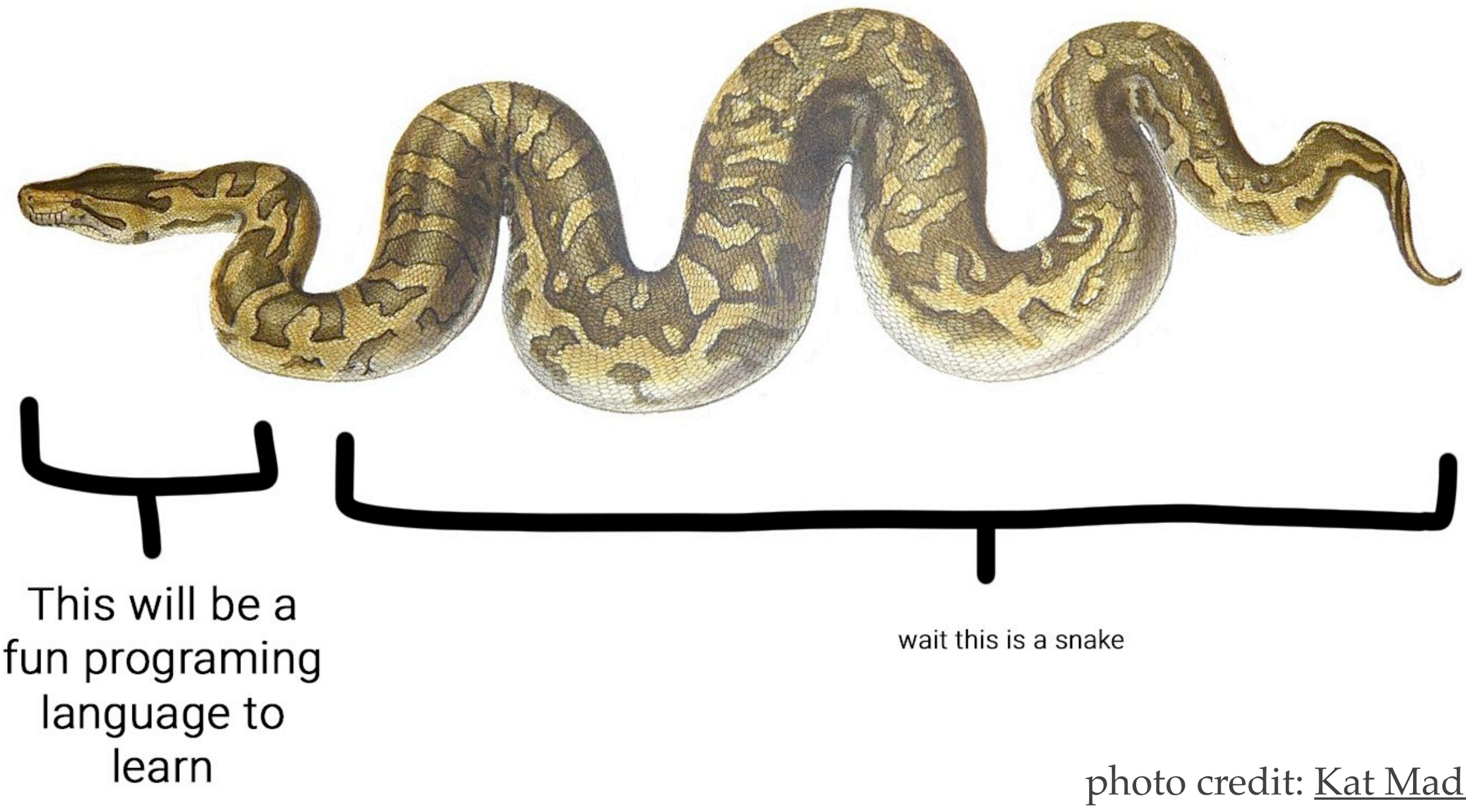



photo credit: Kat Maddox

# Assignments

- Assignments are due on **Thursdays at 10 PM**
- Homework submission will be through Gradescope.
- There are **9** weekly assignments.
- **HW 0** is due this Thursday.

# Intellectual Curiosity Points

If you follow the homework assignment instructions, you can earn up to 90 points.

You can earn up to 10 additional points by demonstrating intellectual curiosity: doing something **beyond** what is described.

# Late Policy

You have **5 late days** for the semester, which you can use all at once, or spread across assignments. **Subsequent late days will come at a cost of 5 HW points.**

Important: I will not answer questions on late work during help hours.

If you have a prolonged illness or unexpected circumstance, let me know and **we'll work together** to make a **custom plan**.

# Sickness

**Please don’t come to class sick.**

If you are sick, let me know and **we'll work together** to make a **plan**.

# Collaboration policy

In this class, you can talk at a high-level with other students about assignments, but **you cannot show them your code**.

If you discuss a homework problem with another student, **please note this** on your assignment when you submit it.

**You may not use ChatGPT, Bard, Codex or any other AI system** unless explicitly stated in the homework assignment.

# AI policy

**You may not use ChatGPT, Bard, Codex or any other AI system** on assignments unless explicitly stated in the homework assignment.

If you would like to use generative AI for any other purpose during the semester, **you must receive permission first**. Send me an email describing how you'd like to use it, and I'll let you know what I think.

# AI Policy

- The impact of GenAI on programming education is a key research area of mine. I'm very happy to talk more about why my policy is what it is.
- My goal is to give you the necessary foundation to become an expert in NLP (if you choose). This requires significant programming expertise.
- ***Automation is about trading learning for efficiency***. This can be worthwhile! But not when your goal is to learn.

# AI Policy

- The impact of GenAI on programming education is a key research area of mine. I'm very happy to talk more about why my policy is what it is.
- My goal is to give you the necessary foundation to become an expert in NLP (if you choose). This requires significant programming expertise.
- ***Automation is about trading learning for efficiency***. This can be worthwhile! But not when your goal is to learn.
- For the **final project**, there are some tasks where the trade-off may be appropriate (for instance, webscraping to build a dataset— not a learning goal for this class).

# Midterms and Final Paper

- **Midterm 1**: in-class programming exam on Oct. 10th
- **Midterm 2**: in-class paper exam on Nov. 7th
- **Final paper** on a research topic of your choice due at the end of term

# Course Goal

**To make you into a skilled NLP practitioner who can:**

- Understand and implement core NLP algorithms and models.
- Explore the challenges posed by different aspects of human language.
- Analyze ethical concerns about language technology.
- Complete a series of projects to implement and improve NLP models.

# Next class: text processing

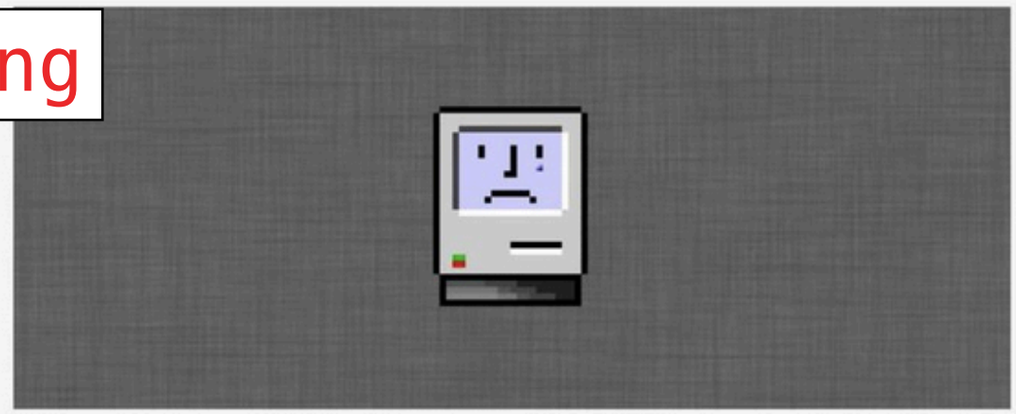
Natural Language Processing

is mostly

DATA WRANGLING

8

## Wrangling data is frustrating



## Can we make it fun?

No.

9

## OK but can we eliminate frustration?

Also no.

Sorry.