CS 124/LINGUIST 180
From Languages to Information

DAN JURAFSKY
PROFESSOR OF COMPUTER SCIENCE
PROFESSOR OF LINGUISTICS
STANFORD UNIVERSITY

WINTER 2026

INTRODUCTION AND COURSE OVERVIEW

What is this class?

- **LLMs and their components**
  - Transformers, Neural Networks, Attention, Conv Nets, Sampling, Language Model Loss, RAG, Tokenization
- **LLMs and their relation to society**
  - Ethical Issues in the use of LLMs
  - LLMs/other tools for computational social science
- **Other language-related tools**
  - Social networks
  - Information retrieval
  - Recommendation engines
  - Speech recognition

What is this class?

The very broad undergrad intro to (at least) 12 grad classes!

cs224C: NLP for Computational Social Science (Yang)

cs224N: Natural Language Processing with Deep Learning (Choi/Yang)

cs224U: Natural Language Understanding (Potts)

cs224V: Conversational Virtual Assistants with Deep Learning (Lam)

cs224S: Spoken Language Processing (Maas)

cs246: Mining Massive Data Sets (Leskovec)

cs224W: Graph Neural Networks (Leskovec)

cs276: Information Retrieval (Manning)

cs329R: Race and Natural Language Processing (Jurafsky/Eberhardt)

cs329X: Human-Centered LLMs (Yang)

cs336: Language modeling from scratch (Hashimoto/Liang)

cs384: Social and Ethical Issues in NLP (Jurafsky)

What is this class? The **Commercial** World

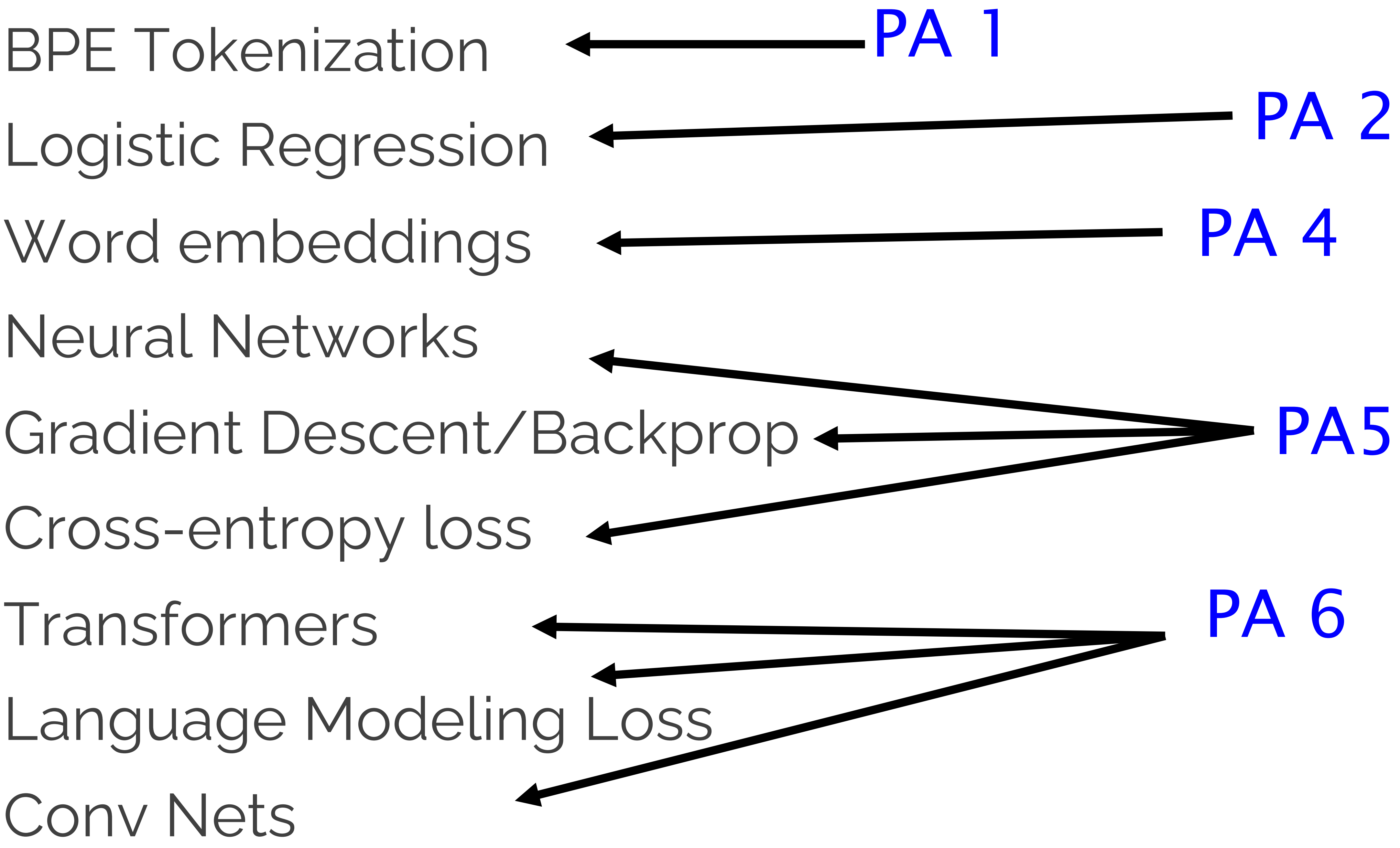
What is this class?

The rise of LLMs has completely changed everything in

- Natural Language Processing (NLP)
- AI
- Information Retrieval (IR)
- Recommendation Systems
- Speech Recognition

This class starts from scratch and builds up how LLMs work and how they are applied!

What is this class?
Intro to the algorithmic components of LLMs



BPE Tokenization

PA 1

Logistic Regression

PA 2

Word embeddings

PA 4

Neural Networks

Gradient Descent/Backprop

PA5

Cross-entropy loss

Transformers

PA 6

Language Modeling Loss

Conv Nets

Example Topic: LLMs and Transformers

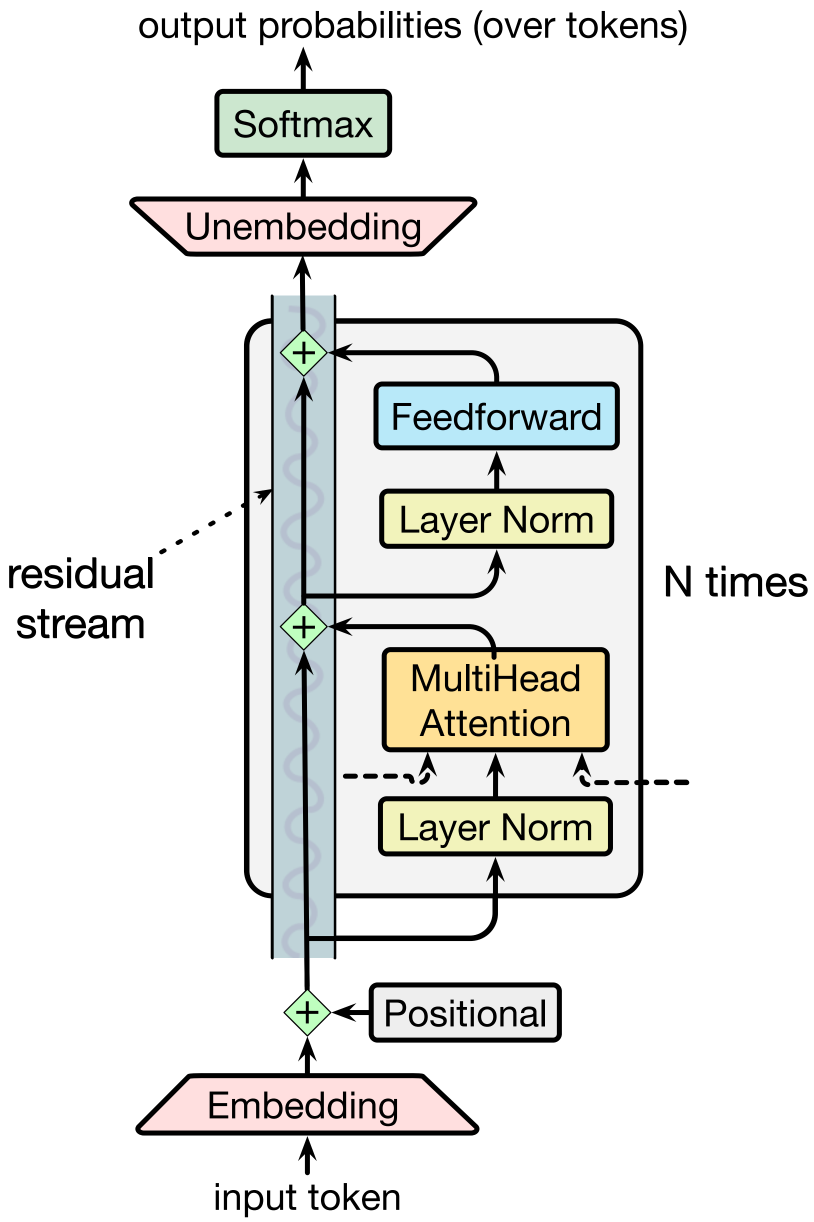
What is attention and how does the transformer work?

How are language models trained?

How is text tokenized

Programming Assignment 6: Transformers!

Programming Assignment 1: Tokens!



Example Topic: Speech

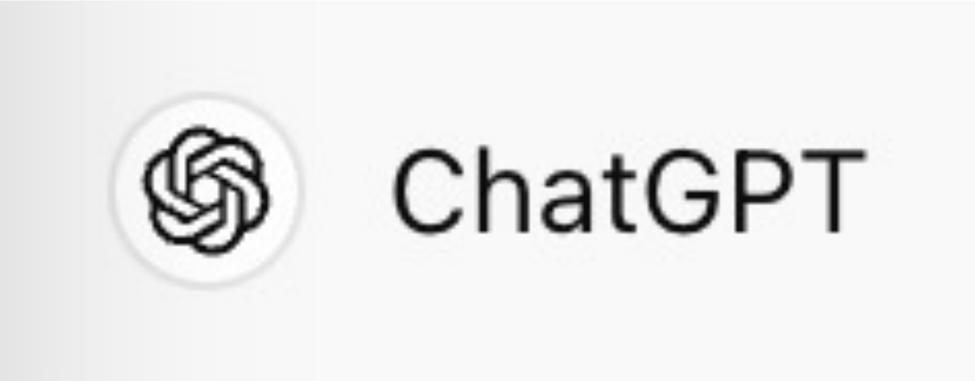
How does speech recognition work?

What different algorithms do we need to deal with speech than text?

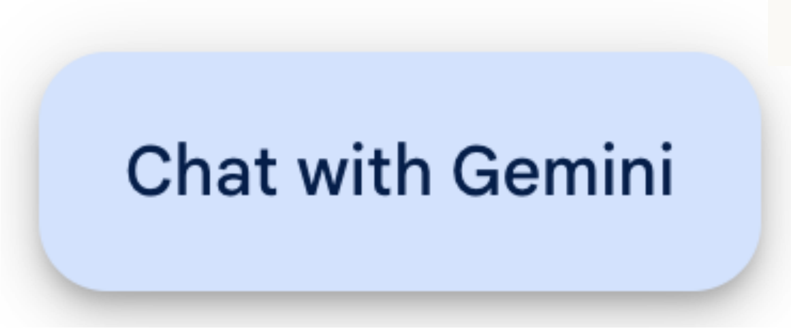
Programming Assignment 6b: Speech!

Large language models!

What can LLMs do?
What can't they do?


ChatGPT

Claude



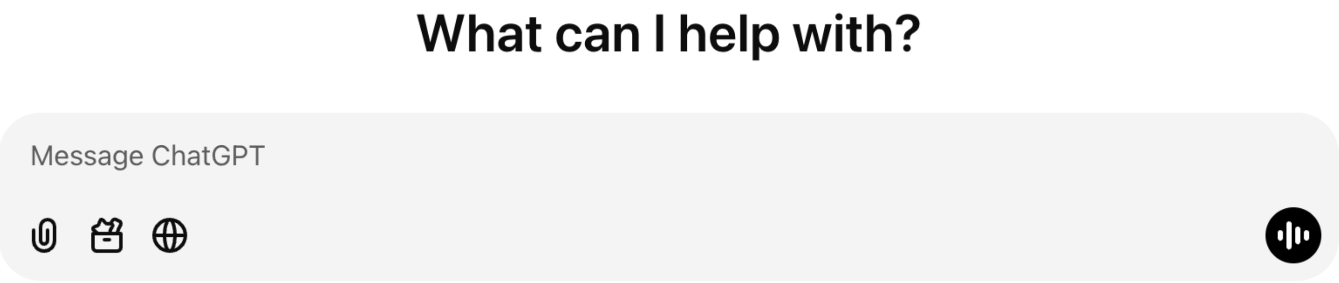
Llama 3

Lab 4
PA 7

Agentic LLMs and Personal Assistants

amazon alexa

ChatGPT



What can I help with?

Message ChatGPT

Siri

What can I help you with?

Listening..

PA 7

What is this class?
Intro to more crucial language algorithms

Regular Expressions

Minimum Edit Distance

Information Retrieval/ WebSearch

Network algorithms

- PageRank & Centrality
- Power Laws & Clustering

Recommendation engines

- Collaborative filtering

Example Topic: Information Retrieval

Text-based information retrieval (IR) for web search

Probably the most frequently used algorithm in the history of the planet

6,586,013,574 web searches every day

How does it work? We'll learn:

- classic **TF/IDF**
- modern **dense retrieval**
- LLM-based **RAG**

Programming Assignment 3: Search!

Computational Biology: Comparing Sequences

```
AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
 || |||||||  ||||x|  || |||  |||||
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

Sequence comparison is key to

- Finding genes
- Determining function
- Uncovering evolutionary processes

This is also how we evaluate LLM functions like speech recognition

Minimum edit distance (Quiz 1)

Example Topic: Logistic Regression for Text Classification

Disaster Response!

Haiti Earthquake 2010

Classifying SMS messages

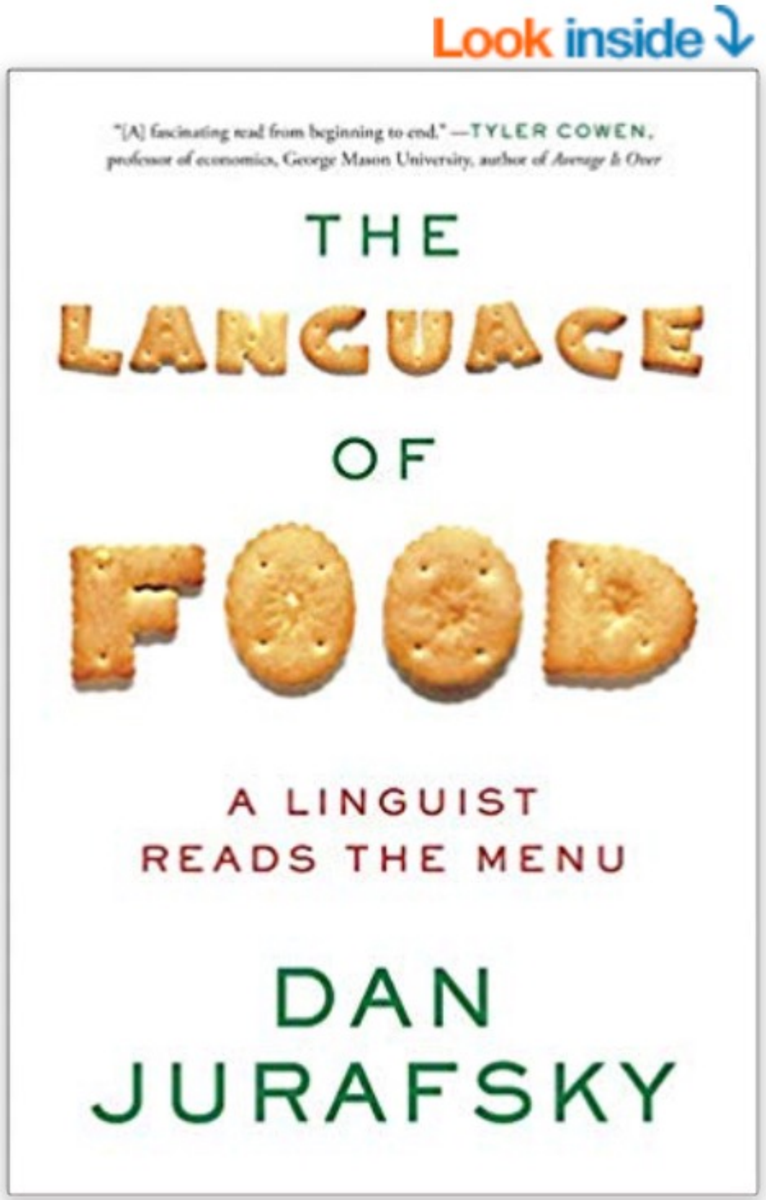
Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen

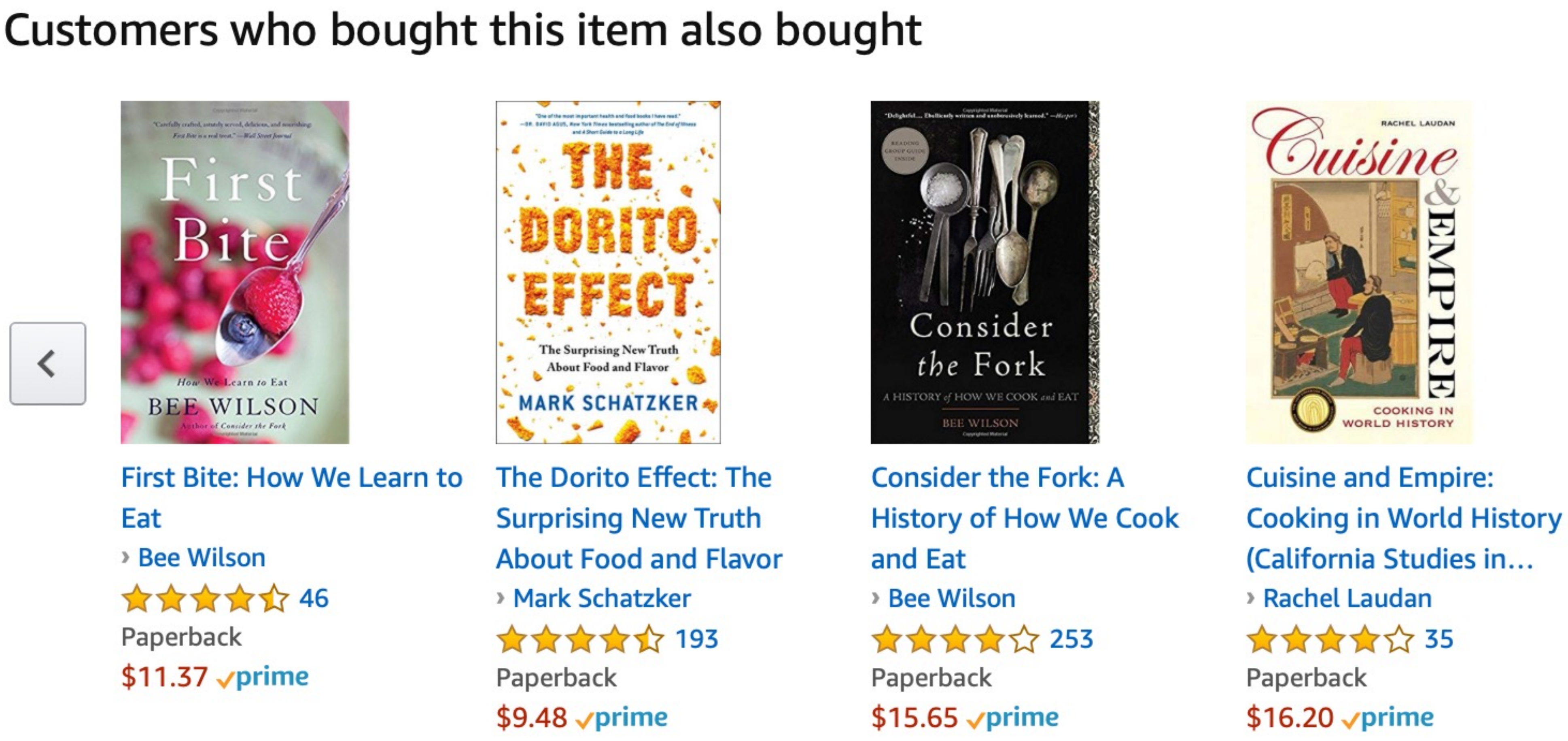
> I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water.

Programming Assignment 2: Triage!

Recommendation Engines: The Good

If you bought....


Look inside
"[A] fascinating read from beginning to end." —TYLER COWEN, professor of economics, George Mason University, author of Average Is Over
THE LANGUAGE OF FOOD
A LINGUIST READS THE MENU
DAN JURAFSKY



Customers who bought this item also bought

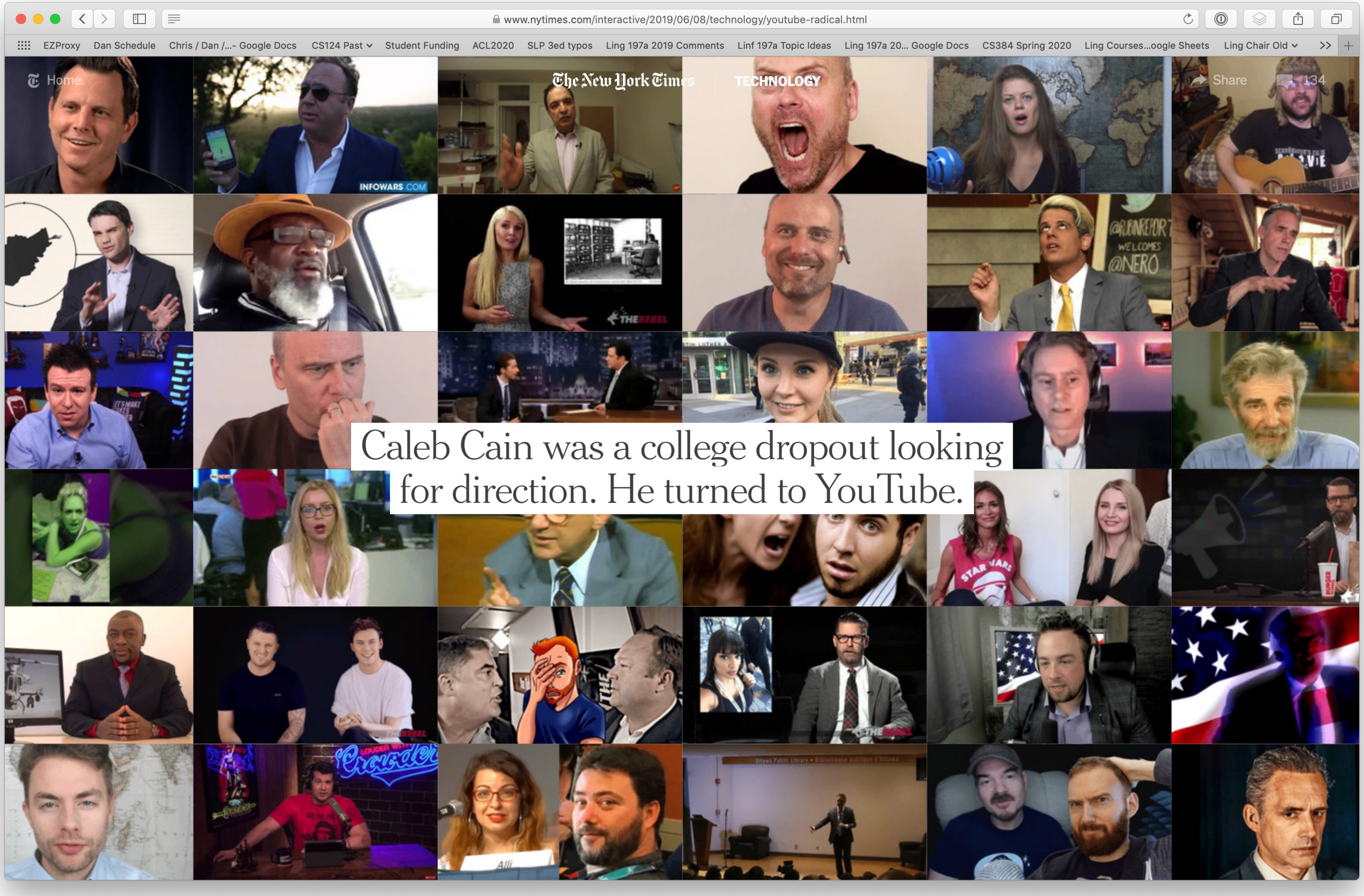
First Bite: How We Learn to Eat
› Bee Wilson
★★★★☆ 46
Paperback
$11.37 ✓prime

The Dorito Effect: The Surprising New Truth About Food and Flavor
› Mark Schatzker
★★★★☆ 193
Paperback
$9.48 ✓prime

Consider the Fork: A History of How We Cook and Eat
› Bee Wilson
★★★★☆ 253
Paperback
$15.65 ✓prime

Cuisine and Empire: Cooking in World History (California Studies in...
› Rachel Laudan
★★★★☆ 35
Paperback
$16.20 ✓prime

And the dark side: YouTube Radicalization



Caleb Cain was a college dropout looking for direction. He turned to YouTube.

You'll implement LLM agents

Using the collaborative filtering algorithm for recommendations

PA 7 and Quiz 8

What is this class?
Introduction to Social NLP and Computational Social Science

NLP and LLMs can be **socially aware** and **social actors.**

- This can lead them to be biased
- But we can also use them to analyze human biases

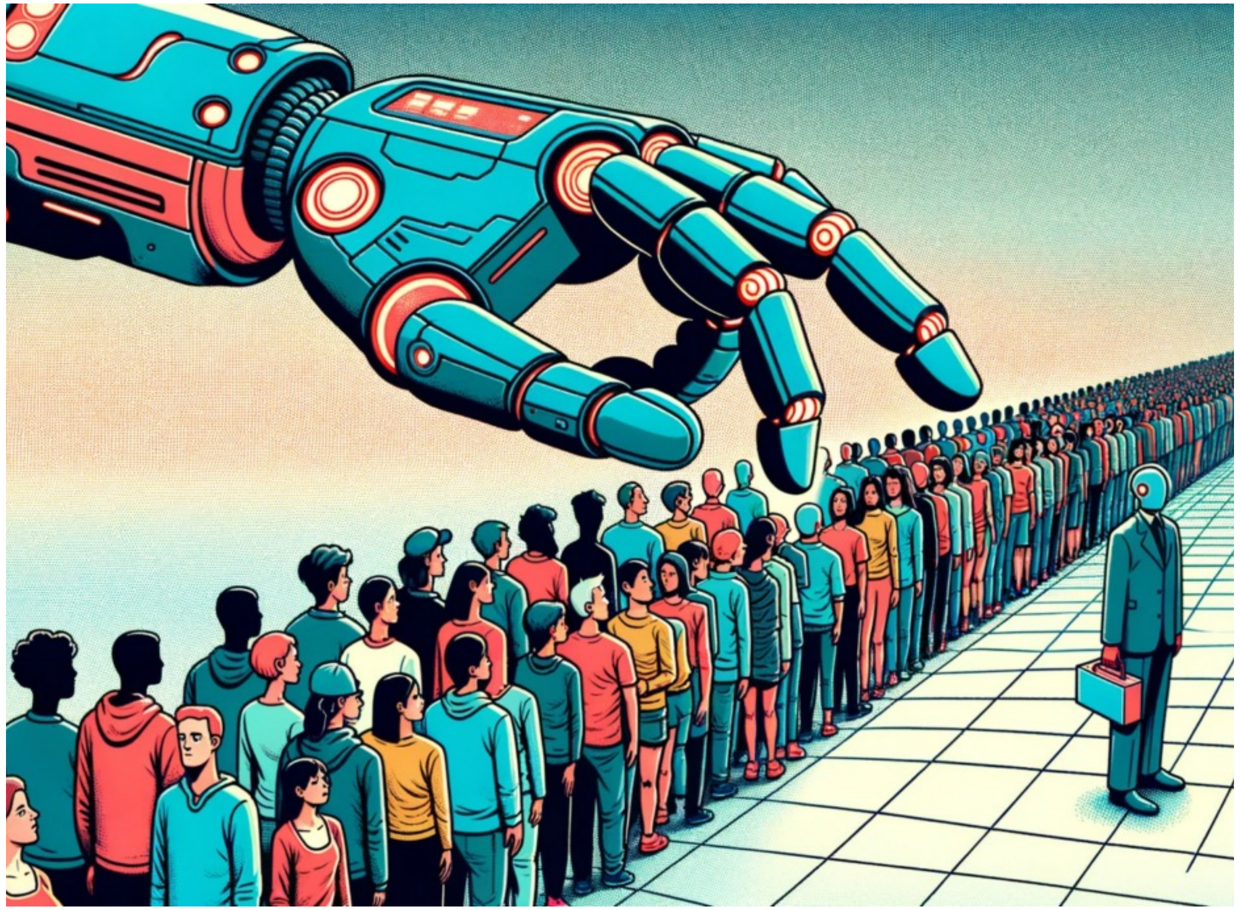
We'll study social flaws in LLM

LLMs hallucinate

LLMs are overconfident

LLMs are sycophantic

LLMs display stereotypes about every group (Asians, Muslim, Blacks, women)



The Decoder, Matthias Bastian, created by Dall-E

Example topic: Applying social NLP to humanities, social science, cultural analytics, text data science!

Detecting latent meaning

Sentiment in Restaurant Reviews

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. First Monday 19:4

900,000 Yelp reviews online

A very bad (one-star) review:

The bartender....... absolutely horrible........ we waited 10 min before we even got her attention........ and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response ....

What is the language of bad reviews?

Negative sentiment language

horrible awful terrible bad disgusting

Past narratives about people

waited, didn't, was

he, she, his, her,

manager, customer, waitress, waiter

Frequent mentions of we and us

… **we** were ignored until **we** flagged down a waiter to get **our** waitress …

Other narratives with this language

A genre using:

Past tense, we/us, negative, people narratives

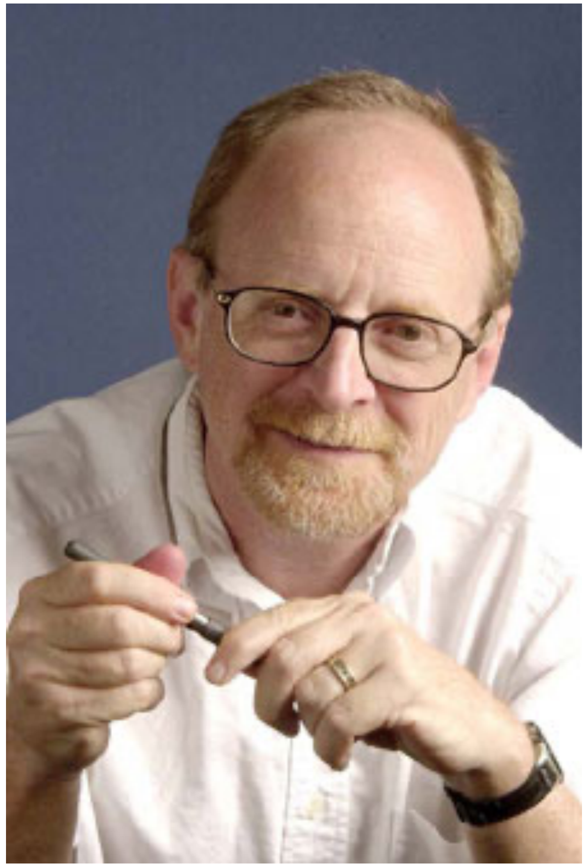
Texts written by **people suffering trauma**

- James Pennebaker lab at UT Austin
- Past tense is used for "distancing"
- Use of “we”: seeking solace in community

1-star reviews are trauma narratives!

The lesson of reviews:

It’s all about personal interaction



What about positive reviews? Sex, Drugs, and Dessert

addicted to pepper shooters

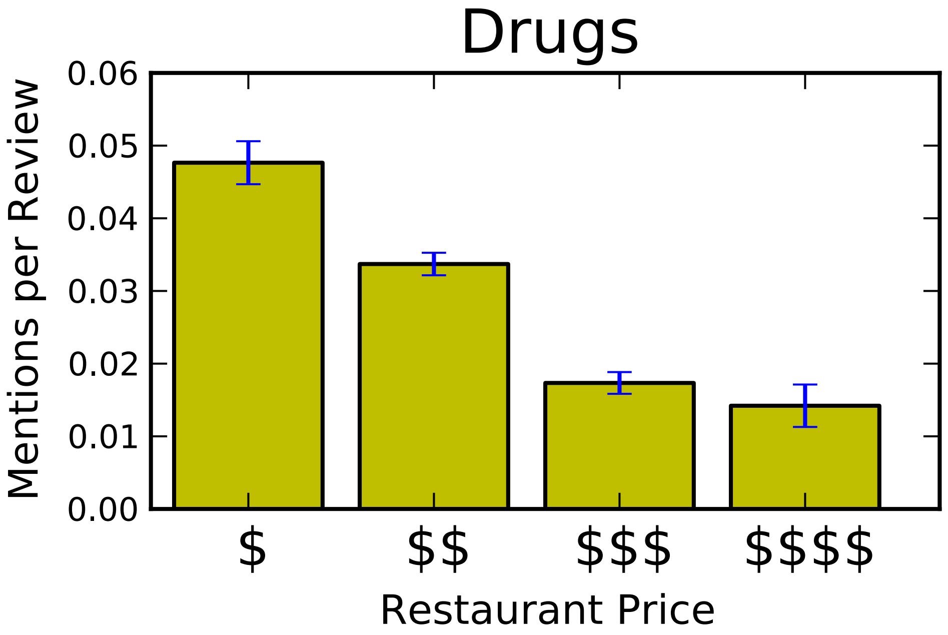
garlic noodles… my *drug of* choice

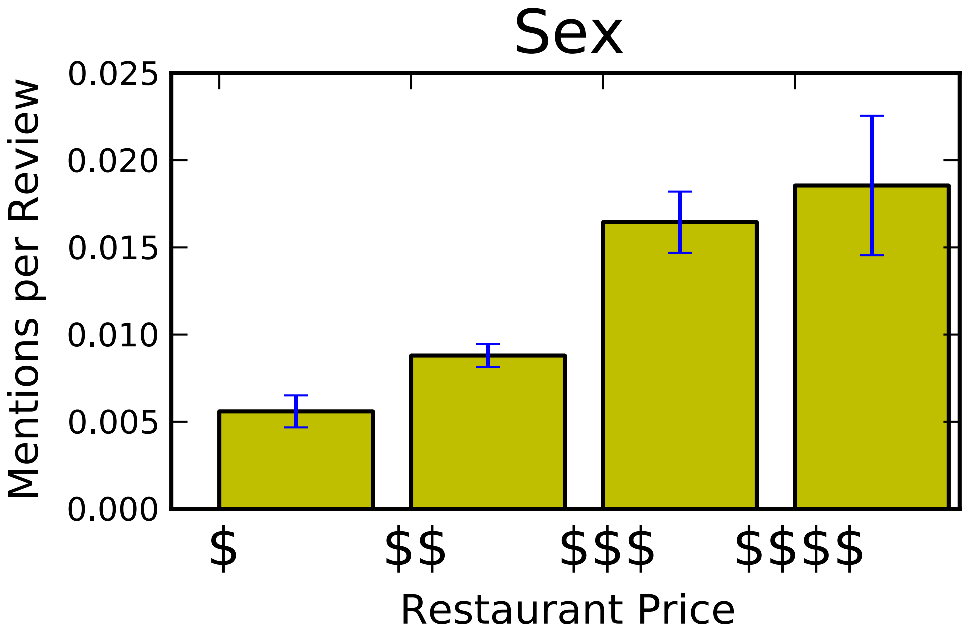
the fries are *like crack*

orgasmic pastry

sexy food

seductively seared fois gras





Computational Social Science
Help improve Police-Community Interaction

Problem:

Inappropriate use of force by police, especially to Black Americans

NLP can help!

- Prof. Jennifer Eberhardt (Psych and GSB) has shown to measure and improve police-community relations.
- Together we apply NLP to do this at scale!
  - Analyze speech from body-worn cameras to quantify police-community interactions
  - Develop officer training
  - Reduce the chances of violence

Yet another topic: Social Networks

The network formed by your friends or other relations offline or online

- Can we compute properties of these networks?
- Extract information from them?

  - *Network algorithms (Quiz 9)*

How does language modeling work?

Let's start by thinking about the language modeling task.

Why is it so remarkable?

What makes language interpretation hard?

Ambiguity

Language is ambiguous

Often as language users we don't even notice this

Resolving ambiguity is hard

Yet language models do this efficiently

Some very simple kinds of ambiguity

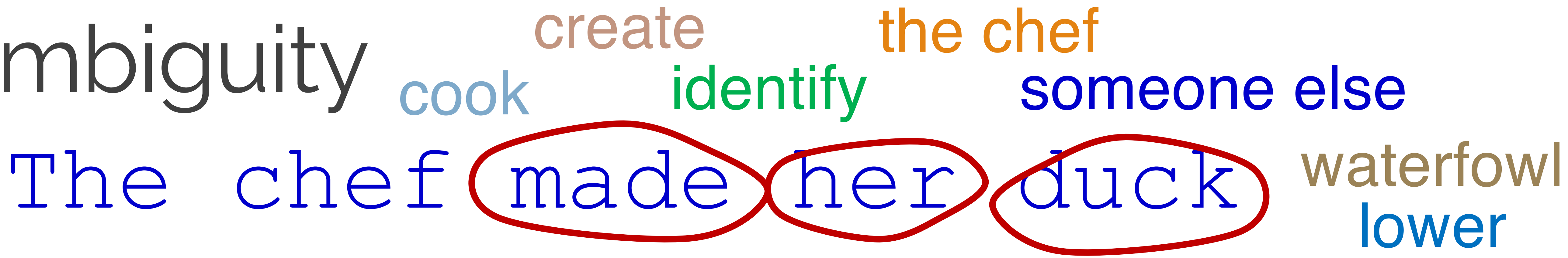
There are at least half a dozen meanings of this sentence:

`The chef made her duck`

Go here and type (and vote for) some definitions

https://pollev.com/danjurafsky451

Ambiguity



cook

create

identify

the chef

someone else

The chef made her duck

waterfowl

lower

The cook cooked waterfowl for a different woman X (person using "she/her" pronouns) to eat

The cook cooked waterfowl belonging to X

The cook cooked waterfowl belonging to the cook

The cook created the (plaster?) waterfowl that X owns

The cook caused X to quickly lower X's head or body

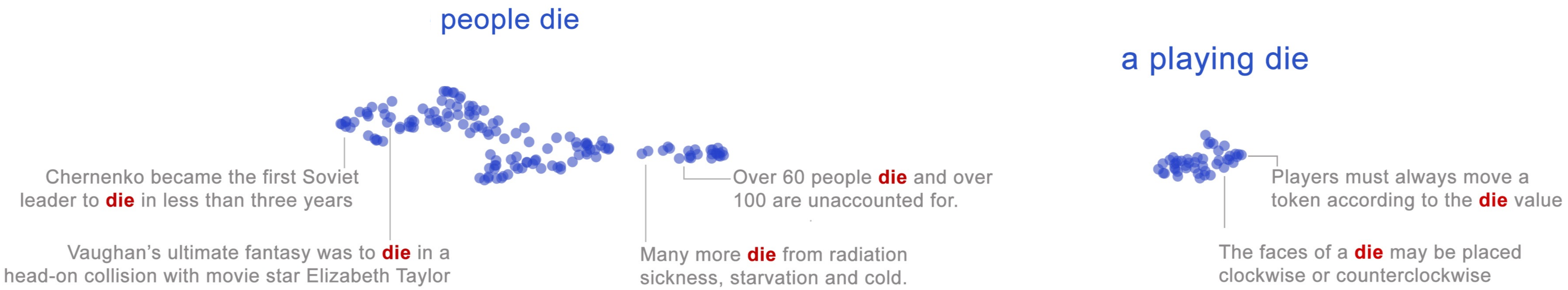
The cook uncovered the true identity of the cook's spy waterfowl

The cook waved their magic wand and turned X into undifferentiated waterfowl

How do LLMs deal with the complexity of meaning?

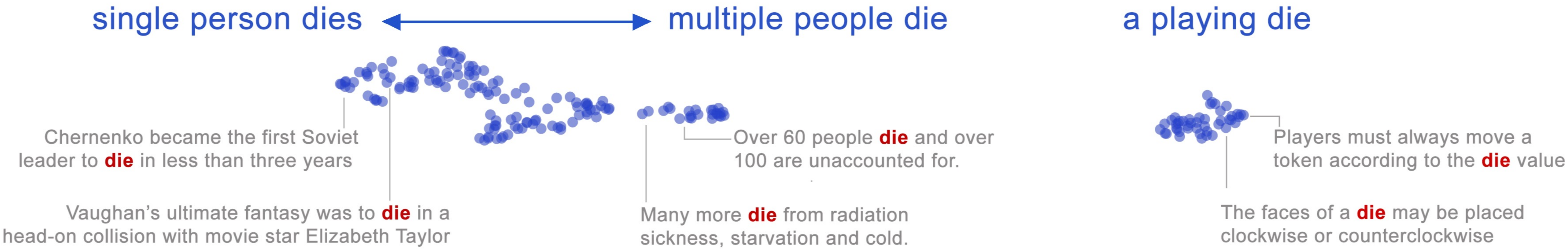
Neural **word embeddings**

A word's meaning represented as a region in 1000-dim space!
Here's the word "die" in 2D:



Word embeddings

But not just two discrete senses



Chernenko became the first Soviet leader to **die** in less than three years

Vaughan's ultimate fantasy was to **die** in a head-on collision with movie star Elizabeth Taylor

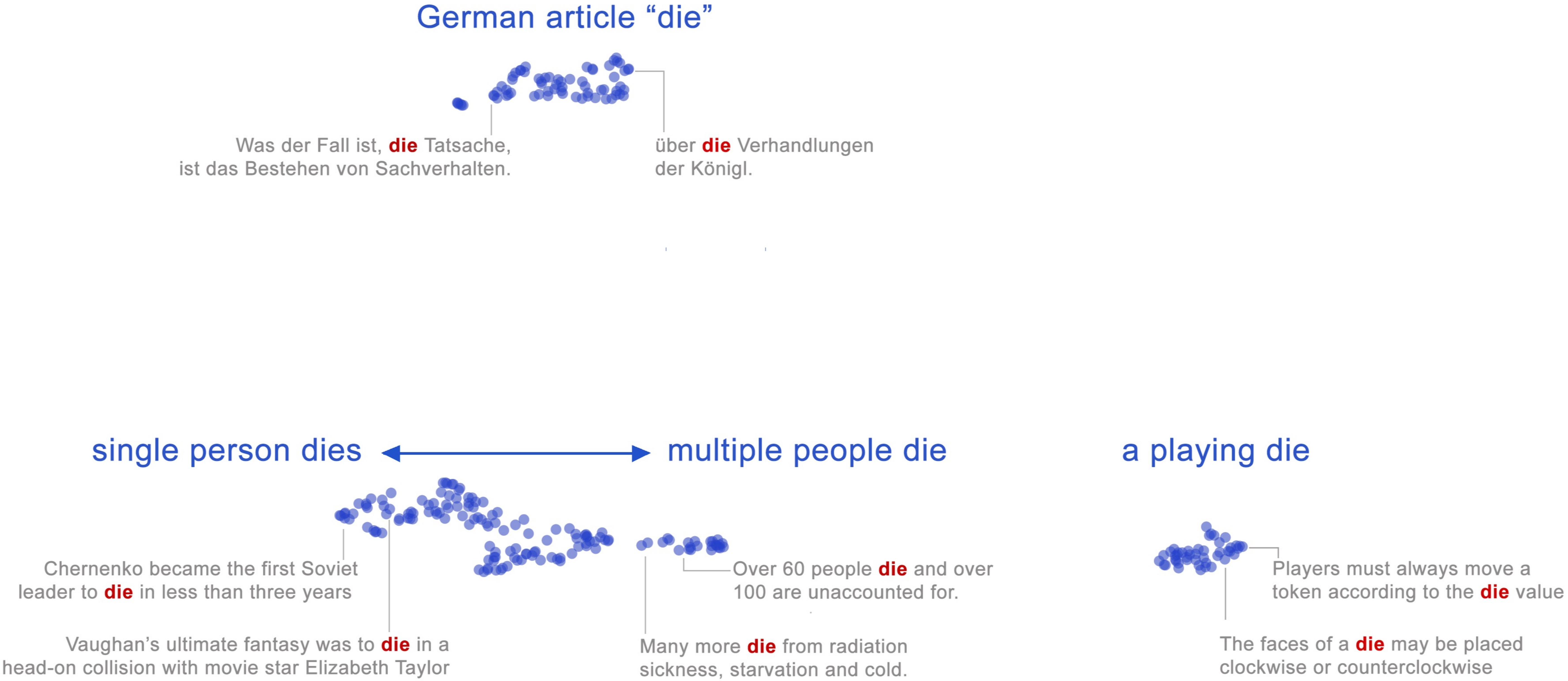
Over 60 people **die** and over 100 are unaccounted for.

Many more **die** from radiation sickness, starvation and cold.

Players must always move a token according to the **die** value

The faces of a **die** may be placed clockwise or counterclockwise

Embeddings: not just one language



German article “die”

Was der Fall ist, **die** Tatsache,
ist das Bestehen von Sachverhalten.

über **die** Verhandlungen
der Königl.

single person dies ⟷ multiple people die

a playing die

Chernenko became the first Soviet
leader to **die** in less than three years

Vaughan’s ultimate fantasy was to **die** in a
head-on collision with movie star Elizabeth Taylor

Over 60 people **die** and over
100 are unaccounted for.

Many more **die** from radiation
sickness, starvation and cold.

Players must always move a
token according to the **die** value

The faces of a **die** may be placed
clockwise or counterclockwise

LLMs and Embeddings

LLMs learn to develop vector models of meaning through training

These models let them represent sophisticated and subtle differences in meaning

We can also use these for computational social science!

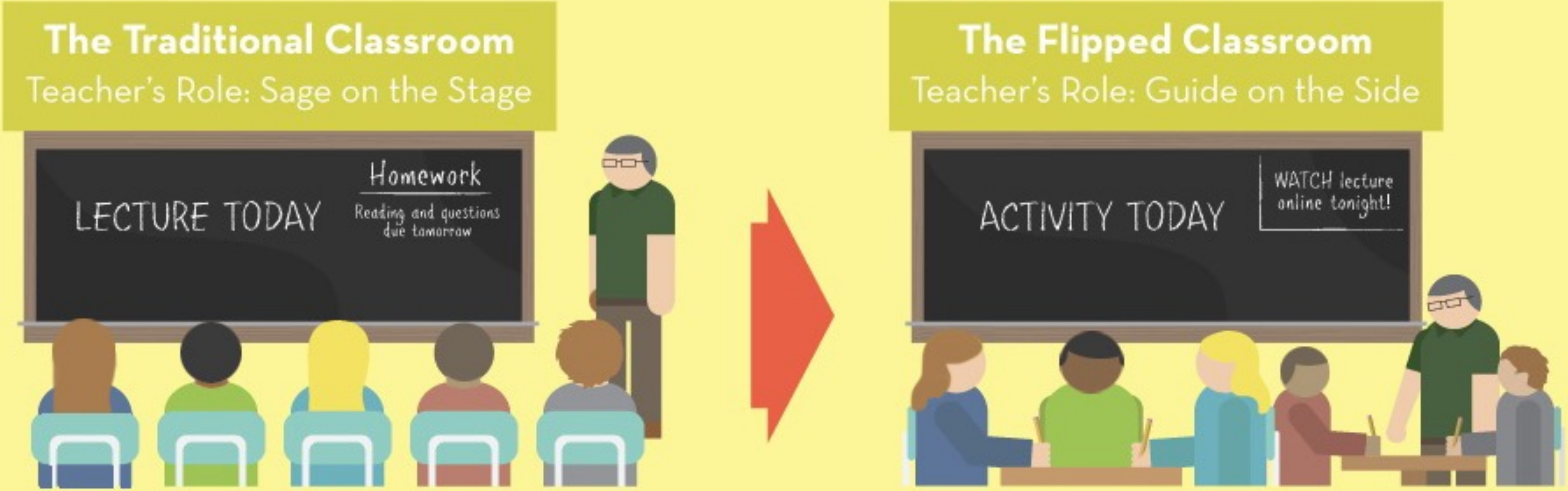
PA 4 (Embeddings!)

What is this class?
Evidence Based Pedagogy!

WHAT IS THE FLIPPED CLASSROOM?

The flipped classroom inverts traditional teaching methods, delivering instruction online outside of class and moving "homework" into the classroom.

THE INVERSION



The Traditional Classroom
Teacher's Role: Sage on the Stage

LECTURE TODAY
Homework
Reading and questions due tomorrow

The Flipped Classroom
Teacher's Role: Guide on the Side

ACTIVITY TODAY
WATCH lecture online tonight!

From (defunct) www.knewton.com/flipped-classroom/

Why the flipped classroom (1)

Mastery learning: Learn until you master

Benjamin Bloom, 1968

Bloom's mastery learning

Personalized, **goal-driven practice**, driven by **feedback**

1. Watch (and re-watch) lectures at your own pace and learn when it's best for you
2. Videos have embedded miniquizzes. If you get it wrong, it gives you feedback about why you misunderstood.
3. You have **infinite** chances at each weekly Tuesday Quiz, so you can go back to the lecture and retake them.
4. With programming assignments you can see your performance on the training and dev set to see what you might be doing wrong on the test set!

Master learning: Grading

I don't grade on a curve.

Your grade describes whether you mastered the material.

I expect you all to master the material.

It is very easy to get an A in this class, most people do

Why the videos have embedded quizzes: “summative” vs “formative” assessment

Summative assessment

- Final exams/midterms: goal is grading

Formative assessment

- Along the way: goal is for **you** to find out what you don’t know so you can learn

Why I don't have a midterm or final

Multiple-choice timed tests don't reflect real life tasks

Scores don't correlate well with ability to do the task

They are stressful and annoying

They invite cheating

They waste an entire week that we instead use for content

Why the flipped classroom (2)

Attention span: everyone spaces out during long lectures

- Middendorf and Kalish, 1995, Johnstone and Percival 1976, Burns 1985

“the class started 1:00. The student sitting in front of me took copious notes until 1:20. Then he just nodded off… motionless, with eyes shut for about a minute and a half, pen still poised. Then he awoke and continued his rapid note-taking as if he hadn’t missed a beat.”

Student remembered only the first 15-20 minutes

Why the flipped classroom (3)

Active learning: Be in charge of your learning

- Most important: programming assignments
- Active learning (“constructivism”), learning by doing

Collaborative learning: Learn from each other

- Use class **lab** time for group problem-solving
- “Small group active learning”
- You **must** do **PA7** in groups of 3-4
- We encourage pair programming on PA1-6 and quizzes

Why the flipped classroom (4)

Constructivism and Labs

Labs tend to involve working through algorithms on mini-problems by hand so you understand them deeply

cs124: Flipped classroom

1. **Prerecorded video lectures on Canvas:**
   - About 80 ~10-minute lectures by me
   - About ~90 minutes/week of video lectures
   - Another 10 lectures by the TAs
2. **Live sessions: (none are recorded)**
   - 5 required in-person lectures
   - 5 required in-class labs("active learning")
     - Lab #1 (Unix text tools) next Tuesday is **required in person**
     - Labs #2, #3, #4 are required but attendance is extra credit (you can do at home).
     - Lab #4 February 24(Git and PA7) is **required in person**

Logistics More Specifically

Online Video **Lectures** w/embedded non-graded questions (watch **before** relevant class/lab/quiz)

20 pages of **reading** a week (read before quiz is due)

Weekly online **quizzes** (due Tue of following week)

7 Python programming assignments (PAs) (due Fri of following week)

- Except PA 7 you get extra time, 2+ weeks

Why you should read the textbook and watch the videos and come to class

Students who do everything report learning more

Also: because we put so much effort into making all the materials!!!

- **The textbook**: gets updated every summer and every fall, taking 100s of hours
- **The videos**: a large percentage get updated every year, each suite of 8 10-minute lectures takes about 40 hours to develop every summer and fall.
- **The live lectures**: change from year to year

(Why so much updating? Our field is in massive flux!!!!!!)

Learning Goals

At the end of this course, you will be able to:

Learning goals

Understand training and inference of large language models and their social implications

Learning goals

Be able to prompt large language models, reason about what they can and can't do, and about their social implications

Learning goals

Write efficient regular expressions to solve any kind of text-based extraction task

Learning goals

Build a supervised classifier to do classification

Learning goals

Build a neural network and train it using stochastic gradient descent

Learning goals

Build a search engine

Learning goals

Build a recommendation engine

Learning goals

Build a computational model of word meaning using neural word embeddings

Learning goals

Understanding agent-based LLM modeling

Learning goals

Understand and implement PageRank and other social network functions

Learning goals

Become expert at working together on computational projects and use group tools like github

Work in our field is rarely done alone!

- PA1-6: Pair programming is encouraged
- PA7: Must be done in groups of 3-4

Can I use LLMs to do my homeworks?

The class policy is:

- You should use LLMs like you use the TA: to give you help, answer your questions, improve your understanding.
- Don't directly paste LLM code.

Why:

- **Learning Goals**! If the LLM does your homeworks for you, you won't achieve your learning goals
- And then you will have a very bad time when you have to whiteboard during job interviews.

What is this class?

The very broad undergrad intro to (at least) 12 grad classes!

cs224C: NLP for Computational Social Science (Yang)

cs224N: Natural Language Processing with Deep Learning (Choi/Yang)

cs224U: Natural Language Understanding (Potts)

cs224V: Conversational Virtual Assistants with Deep Learning (Lam)

cs224S: Spoken Language Processing (Maas)

cs246: Mining Massive Data Sets (Leskovec)

cs224W: Graph Neural Networks (Leskovec)

cs276: Information Retrieval (Manning)

cs329R: Race and Natural Language Processing (Jurafsky/Eberhardt)

cs329X: Human-Centered LLMs (Yang)

cs336: Language modeling from scratch (Hashimoto/Liang)

cs384: Social and Ethical Issues in NLP (Jurafsky)

Should I take 124 or 224N or something else?

CS124 is designed for sophomores or juniors

- It's gentle (I explain everything) and broad (covering many topics, not just NLP/LLMs but also recommendation engines, IR, social networks, social computing)
- Mastery learning, quizzes, programming assignments with starter code and scaffolding.
- No research project, but a fun chatbot final homework

CS224N is a deeper, laser focused, grad course

- They assume you are very familiar with ML; 1st homework jumps right into optimization
- More focus on systems/implementation/scaling, you code more advanced things

CS224N/U/V/S/W, 246, 336, 329R

- Learning via research: novel research projects as a large component

CS324X (Human Centered NLP), CS346 (Social and Ethical Issues in NLP) require 224N or 224U

CS224C: more applied focus, applying NLP to social science: (NLP for Computational Social Science)

(You should of course take all of them!!)

Logistics: Instructor

Instructor: Dan Jurafsky (he/him)

Professor in CS and Linguistics

My office hours:

- This week and next week
  - **Tuesday** after class 4:30-5:30
- Then every **Thursday** classtime 3-4:20
- Margaret Jacks Hall 117
- Book times at calendly.com/jurafsky

How to pronounce my name:

Picture by Ross Petukhov

Course Staff

ıot

Dan Jurafsky
Professor

Linda Liu
Head TA

Belinda Yeung
TA / Student Liaison

Amelie Byun
Course Manager

Adi Badlani
TA

Sri Jaladi
TA

Riya Karumanchi
TA

Ishan Khare
TA

Isabel Sieh
TA

Isha Sinha
TA

Sunny Yu
TA

Esidore Eneinyang
Ethics TA

More logistics: Prereqs

- CS106B,
- Python (at the level of CS106A),
- CS109 (or equivalent background in probability),
- Programming maturity and knowledge of UNIX equivalent to CS107 (this can be waived for PhD students)

Should I come to class if I am sick?

No!

Grading:

A: 93% and above of the total points

A-: 90% and above of the total points

B+: 87% and above of the total points

B: 83% and above of the total points

B-: 80% and above of the total points

C+: 77% and above of the total points

C: 73% and above of the total points

C- (= Credit): 70% and above of the total points

Grading: A+

A+: It is very very easy to get an A in this class but **hard to get an A+.** For an A+ you must do **all** of the following:

- Have perfect scores on all the PAs and quizzes
- Have perfect attendance (or absences excused) at 12 classes (that means all lectures, all labs, and all tutorials, i.e., even the non-required labs and tutorials)
- Have given at least 5 **substantive** and helpful answers to students on the class Ed forum
- Have turned in and gotten credit on extra credit problems on at least 3 of the labs, quizzes, or PAs

Syllabus

cs124.stanford.edu

Where do I find all the programming assignments and quizzes and readings?

Everything is on the webpage cs124.stanford.edu

Except the videos which are on Canvas Modules!

In other words:

- Lectures slides: **webpage**
- Lab instructions: **webpage** (points to git where they live)
- Tutorial information: **webpage**
- Readings: **webpage**
- Programming assignments: **webpage** (points to git where they live)
- Weekly quizzes: **webpage** (points to gradescope where they live)
- Videos: **Canvas**

Coming up this week: Thursday

Optional tutorial on jupyter notebooks and PA0, getting ready for PA1

Come to class **with your laptops** and we'll go through PA0 together!

This tutorial will be led by amazing TA Sri Jaladi!!! But I and many other CAs will be there!

Action Items Before Thursdays class!

1) Read the syllabus webpage at cs124.stanford.edu

2) Look at PA0 (you can find it from the webpage)

3) Watch Canvas Videos on "PA0 Mac Setup" (or "PA0 Windows Setup"), also pointed to by webpage

Coming up next week (Tuesday)

"Unix for poets":

- grep
- sort

Key UNIX tools for dealing with text files and regular expressions.

Plus a brief exercise on n-gram LMs

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

Action Items before next Tuesday's class!

1) Watch the "week 1" videos on Canvas by Sunday (since the quiz is also due Tuesday)

3) Download this file to your laptop

http://cs124.stanford.edu/nyt_200811.txt

4) If you don't know UNIX yet (haven't had cs107):

- For people using a Windows 10 machine, if you don't have Ubuntu on your machine:
  - Watch the pa0 Windows video about how to download and install Ubuntu (it's pointed to from the website)
- Watch Chris Gregg's excellent UNIX videos here: Logging in, first 7 File System, and first 8 useful commands

  https://web.stanford.edu/class/archive/cs/cs107/cs107.1186/unixref/

PA1: Spam Lord and Tokenize!

Write regular expressions to spread evil* throughout the galaxy!

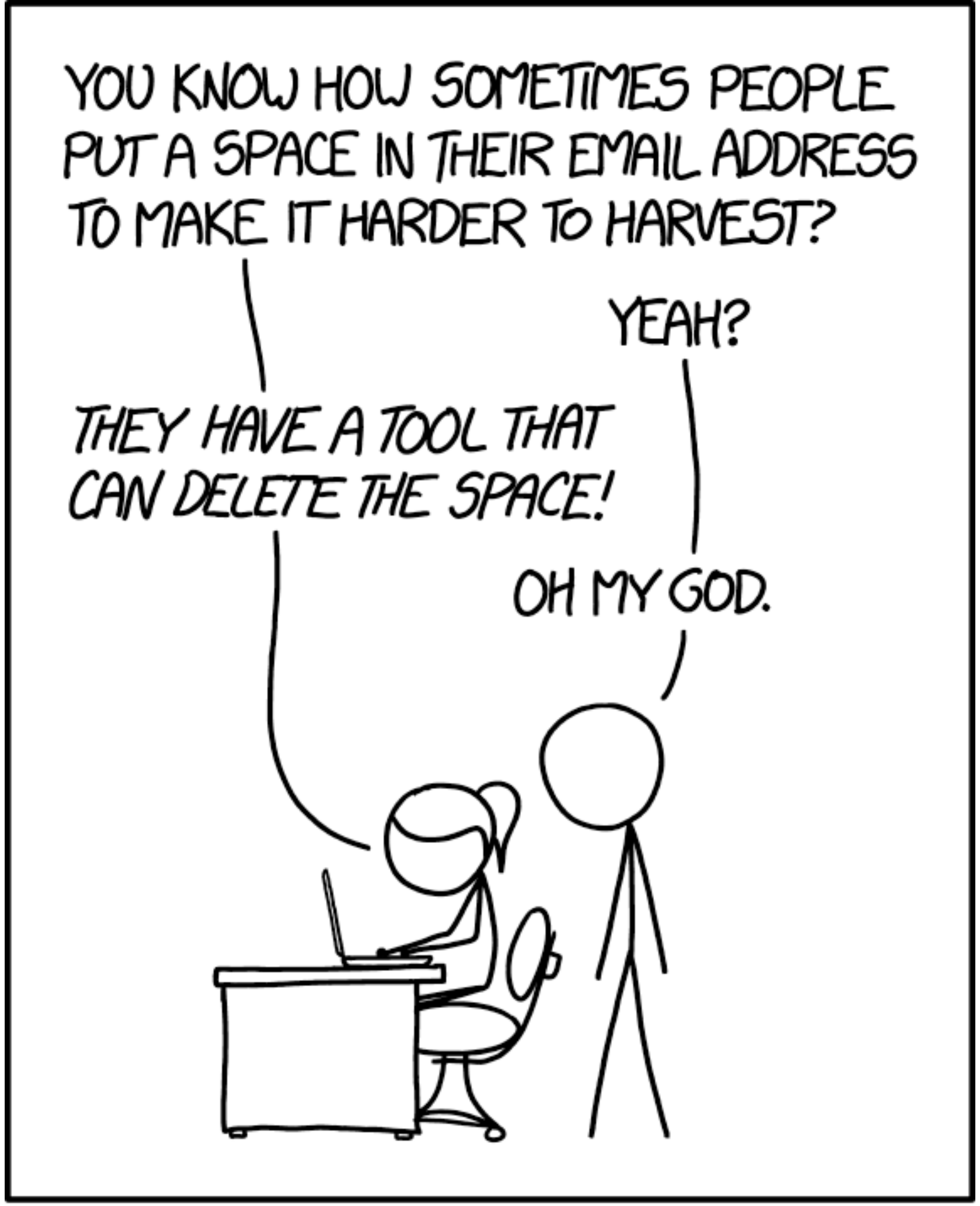
By extracting email addresses from the web!

```
jur a fs ky at st anford dot e d u
```

Also learn how to tokenize like an LLM

Goes live Friday 5pm!

*Just kidding; don't be evil


YOU KNOW HOW SOMETIMES PEOPLE PUT A SPACE IN THEIR EMAIL ADDRESS TO MAKE IT HARDER TO HARVEST?
YEAH?
THEY HAVE A TOOL THAT CAN DELETE THE SPACE!
OH MY GOD.

LESS-DRAMATIC REVELATIONS FROM THE CIA HACKING DUMP