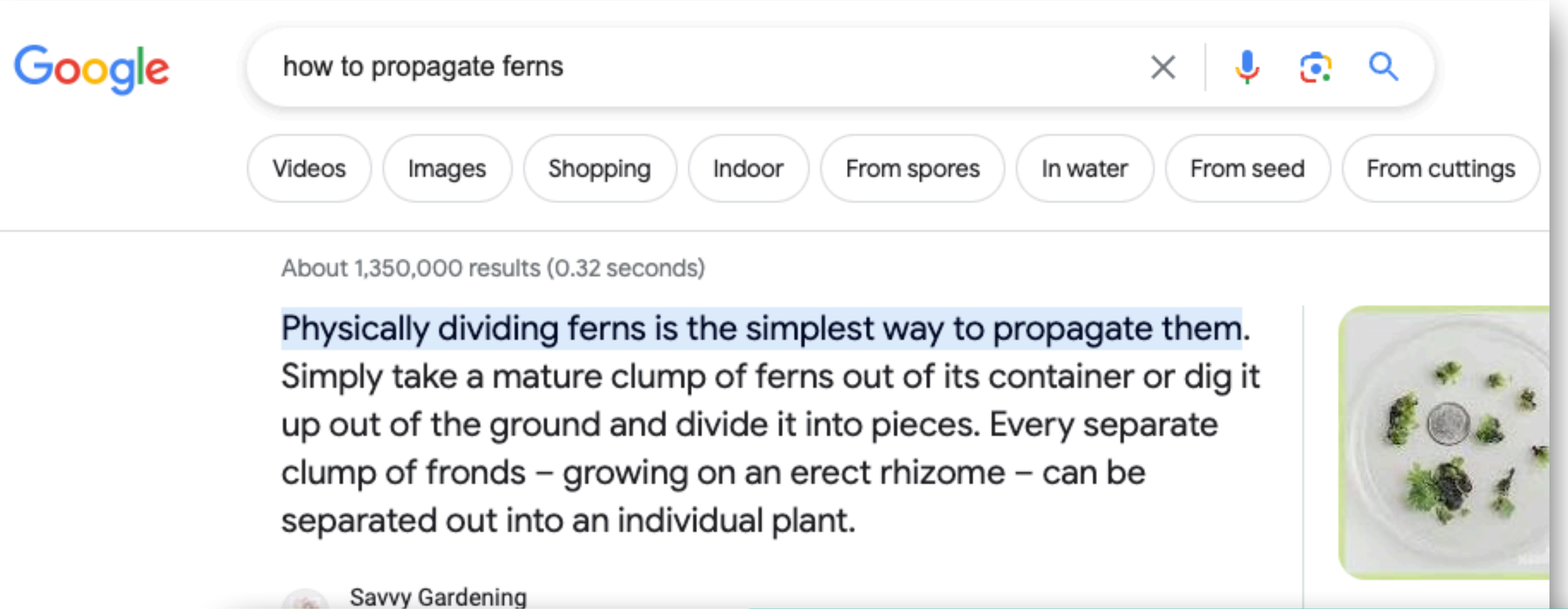
Natural Language Processing and Language Models

Swabha Swayamdipta
Assistant Professor of Computer Science

Introduction to Engineering by Prof. Darin Gray
June 23, 2023

USC Viterbi
School of Engineering

Slides adapted from Greg Durrett, UT Austin



Google

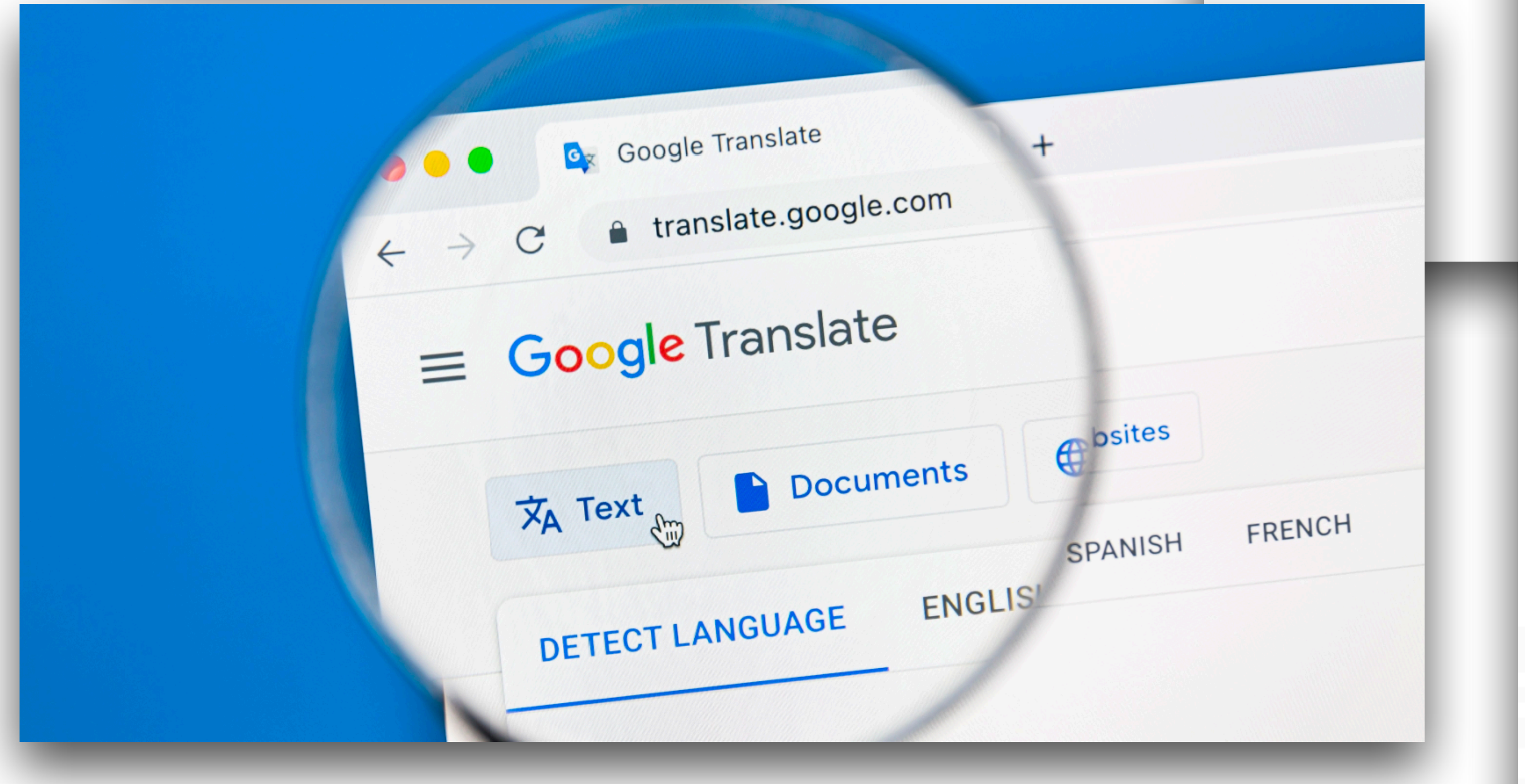
how to propagate ferns

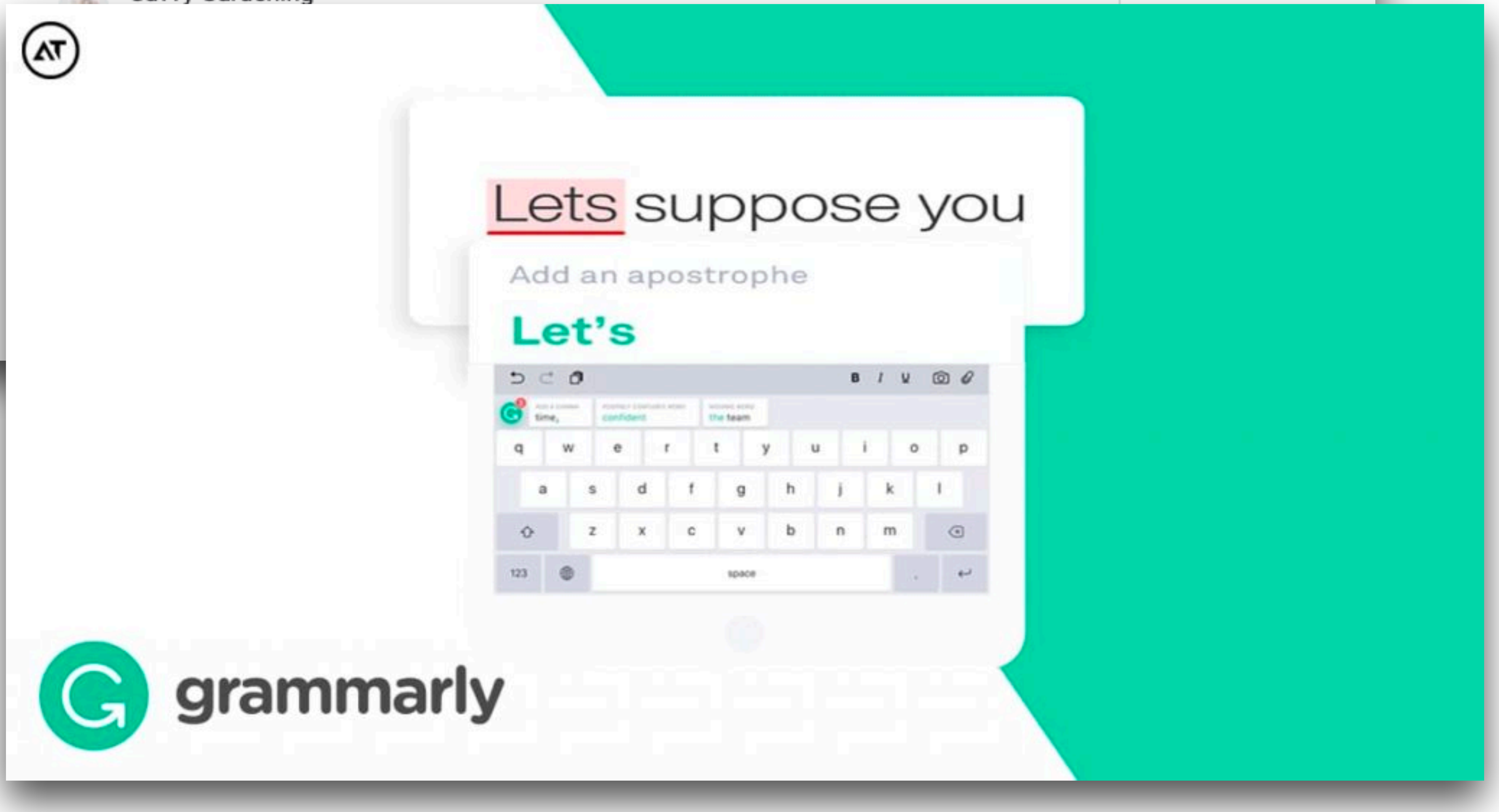
Videos Images Shopping Indoor From spores In water From seed From cuttings

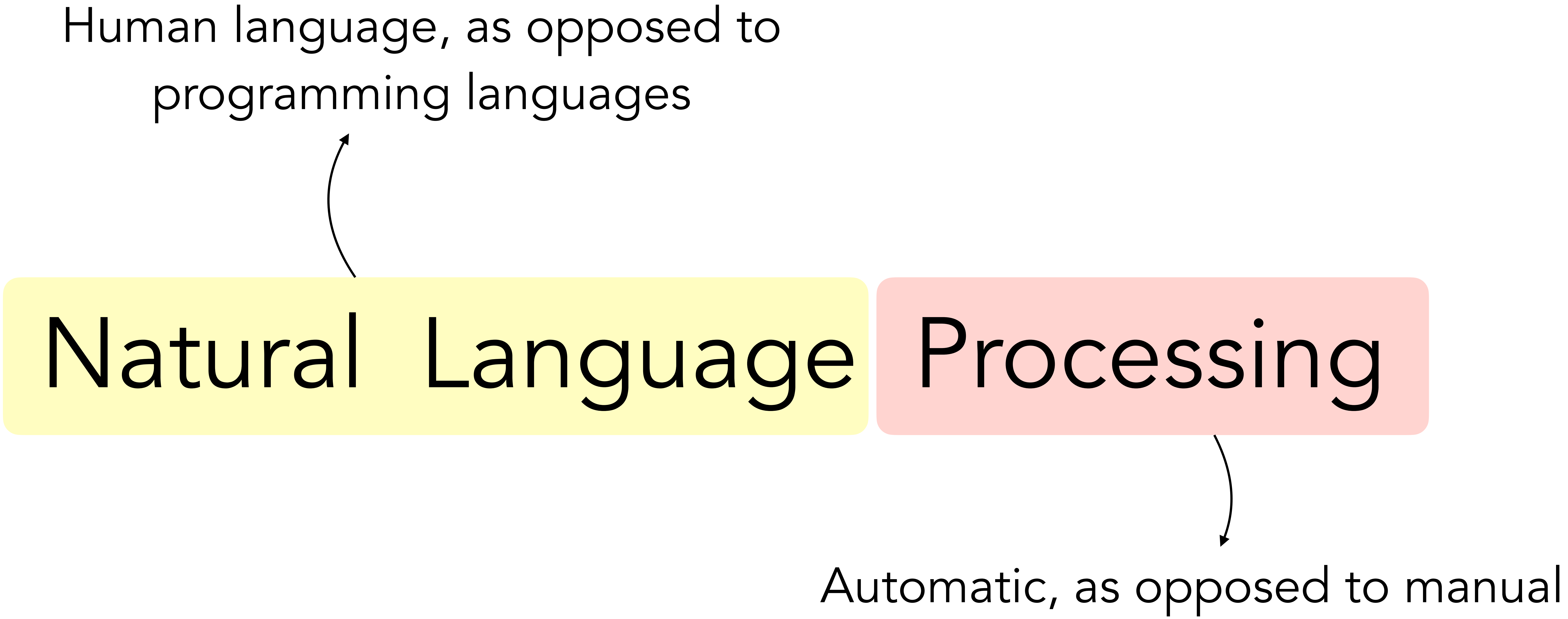
About 1,350,000 results (0.32 seconds)

Physically dividing ferns is the simplest way to propagate them. Simply take a mature clump of ferns out of its container or dig it up out of the ground and divide it into pieces. Every separate clump of fronds – growing on an erect rhizome – can be separated out into an individual plant.

Savvy Gardening







Human language, as opposed to programming languages

Natural Language Processing

Automatic, as opposed to manual

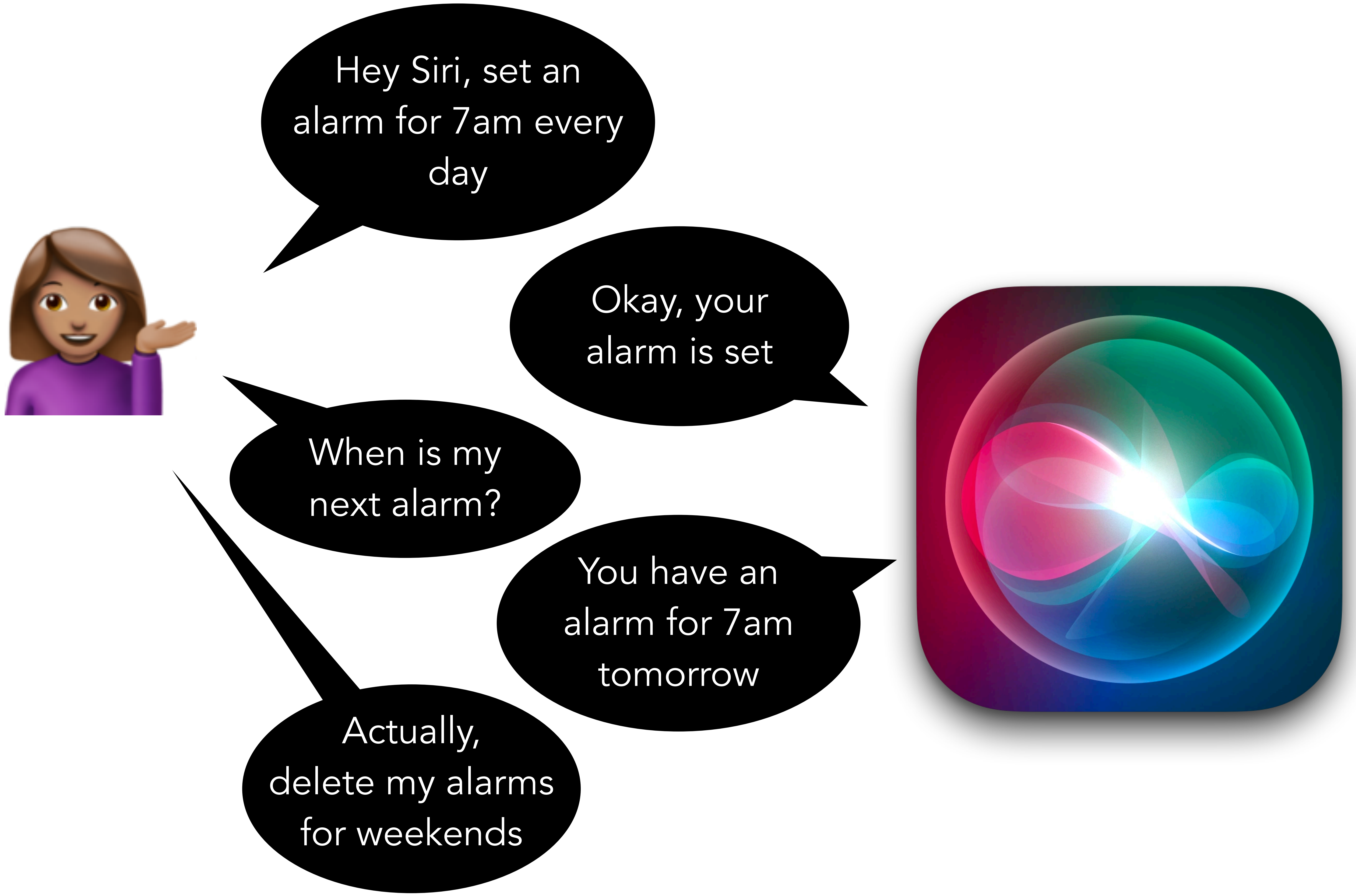
What is Natural Language Processing ?

- Field at the intersection of Computer science, AI (especially machine learning) and Linguistics
- Goal: for computers to process human language, similar to human understanding, towards performing useful tasks
- Challenge: understanding and representing the meaning of language is something even humans struggle with

Slide adapted from Chris Manning

Apple's Siri

- Understands the user
- Remembers what the user said earlier
- Can understand which alarm she is referring to



Hey Siri, set an alarm for 7am every day

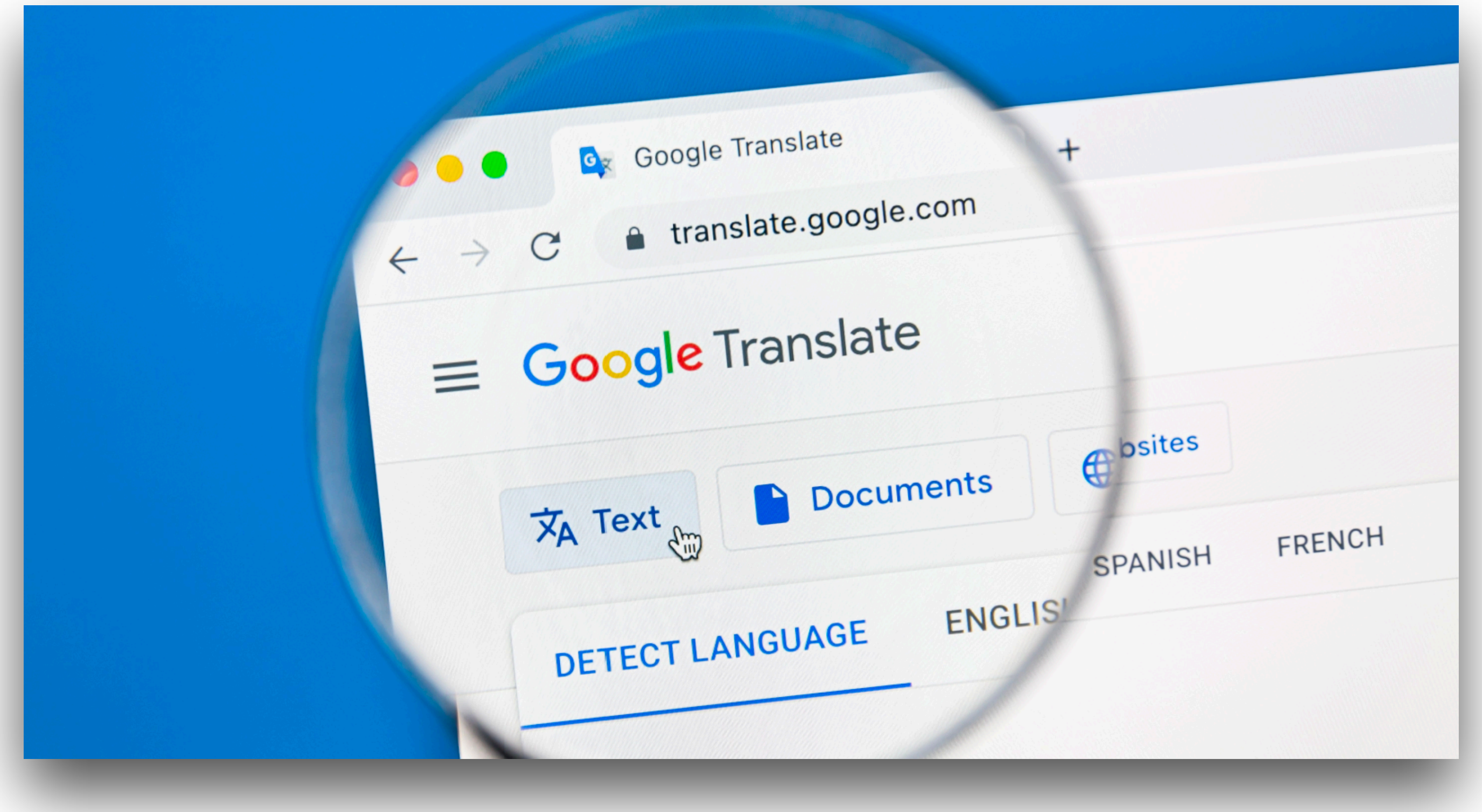
Okay, your alarm is set

When is my next alarm?

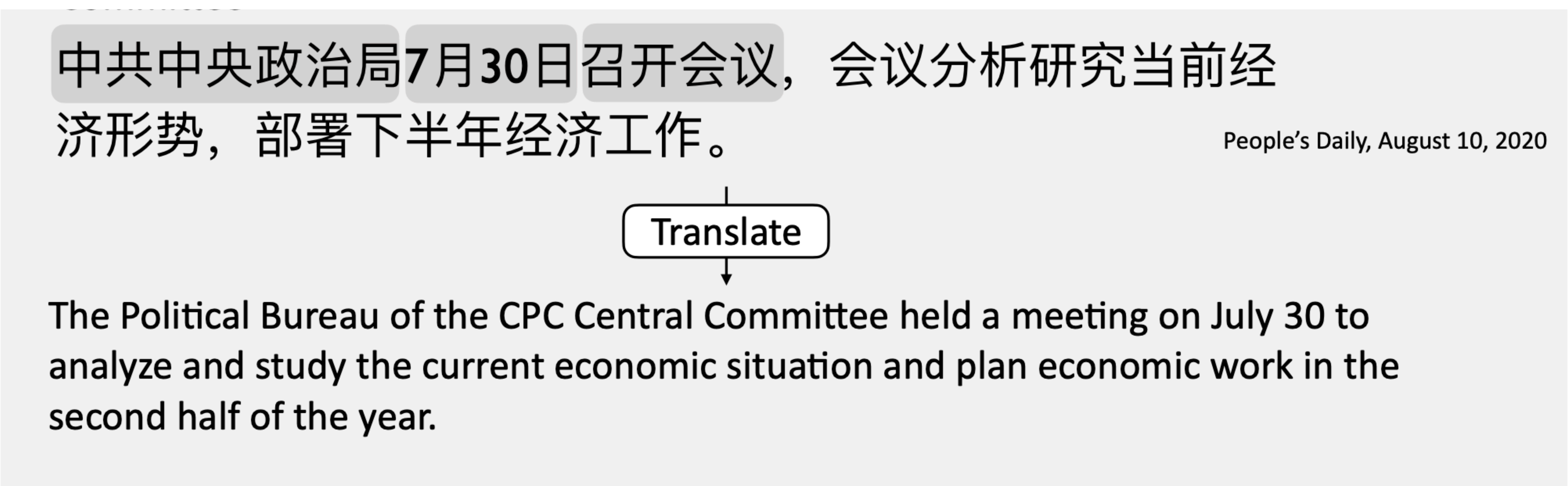
You have an alarm for 7am tomorrow

Actually, delete my alarms for weekends

Google Translate



- Detects language automatically
- Can reorder spans in text on the fly



中共中央政治局7月30日召开会议，会议分析研究当前经济形势，部署下半年经济工作。

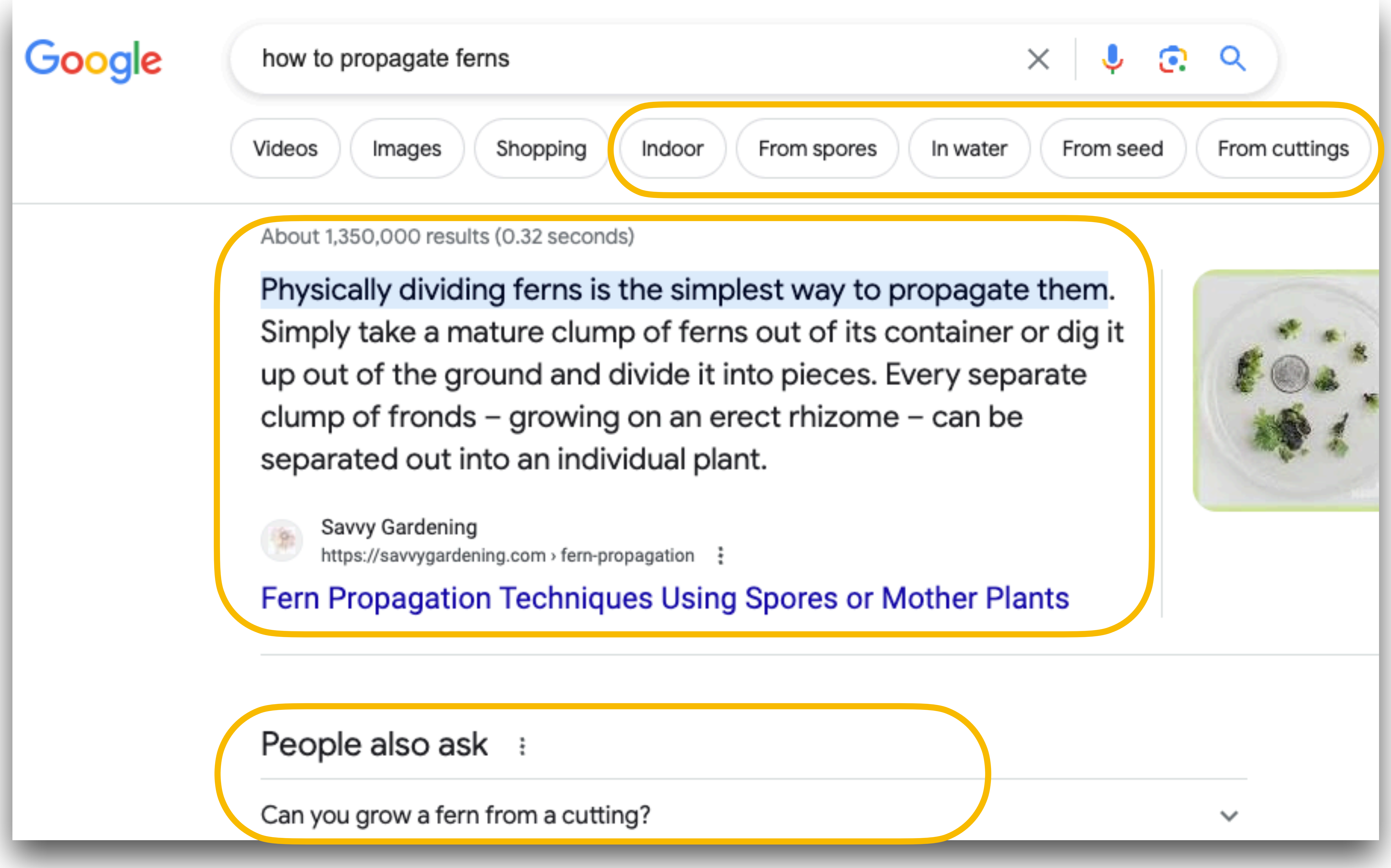
People’s Daily, August 10, 2020

Translate

The Political Bureau of the CPC Central Committee held a meeting on July 30 to analyze and study the current economic situation and plan economic work in the second half of the year.

Google Search

- Understands that a fern can be indoor, can be propagated either from seed or from cuttings
- Can find the exact passage in a webpage that answers the questions
- Can find related (in meaning) questions



how to propagate ferns

About 1,350,000 results (0.32 seconds)

Physically dividing ferns is the simplest way to propagate them. Simply take a mature clump of ferns out of its container or dig it up out of the ground and divide it into pieces. Every separate clump of fronds – growing on an erect rhizome – can be separated out into an individual plant.

Savvy Gardening
https://savvygardening.com › fern-propagation

Fern Propagation Techniques Using Spores or Mother Plants

People also ask

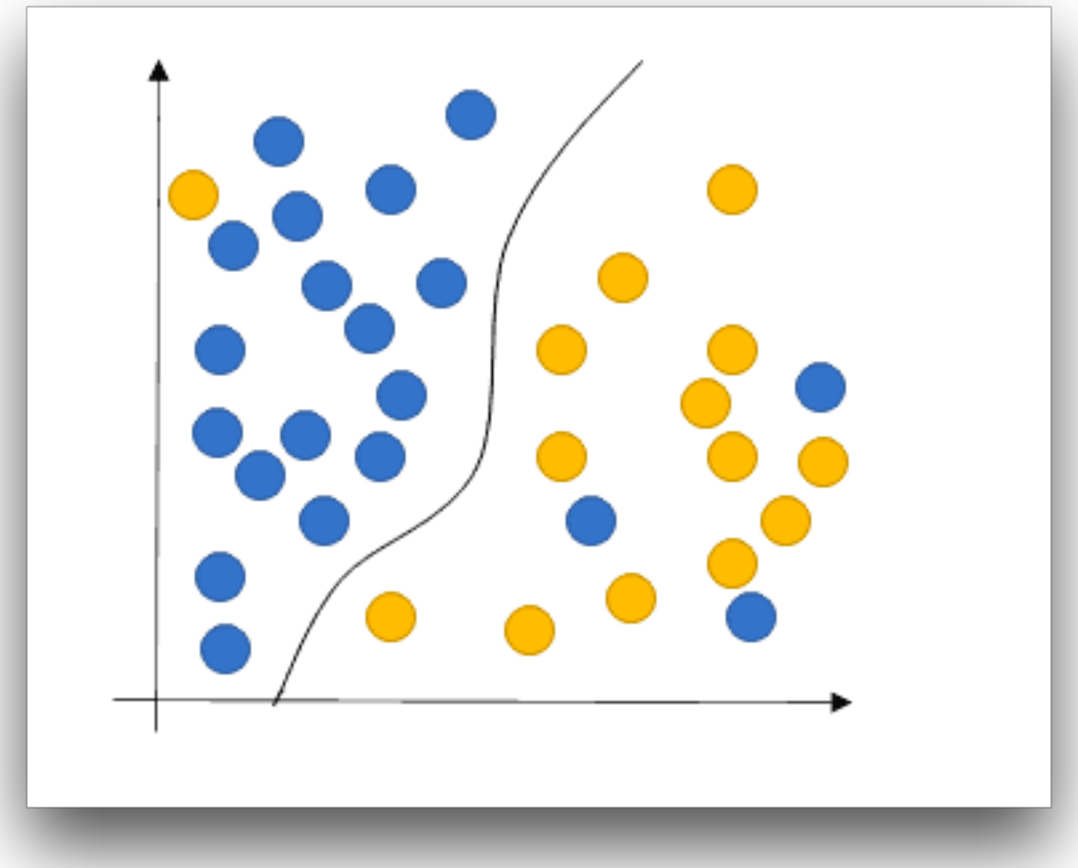
Can you grow a fern from a cutting?

Concrete Outcomes

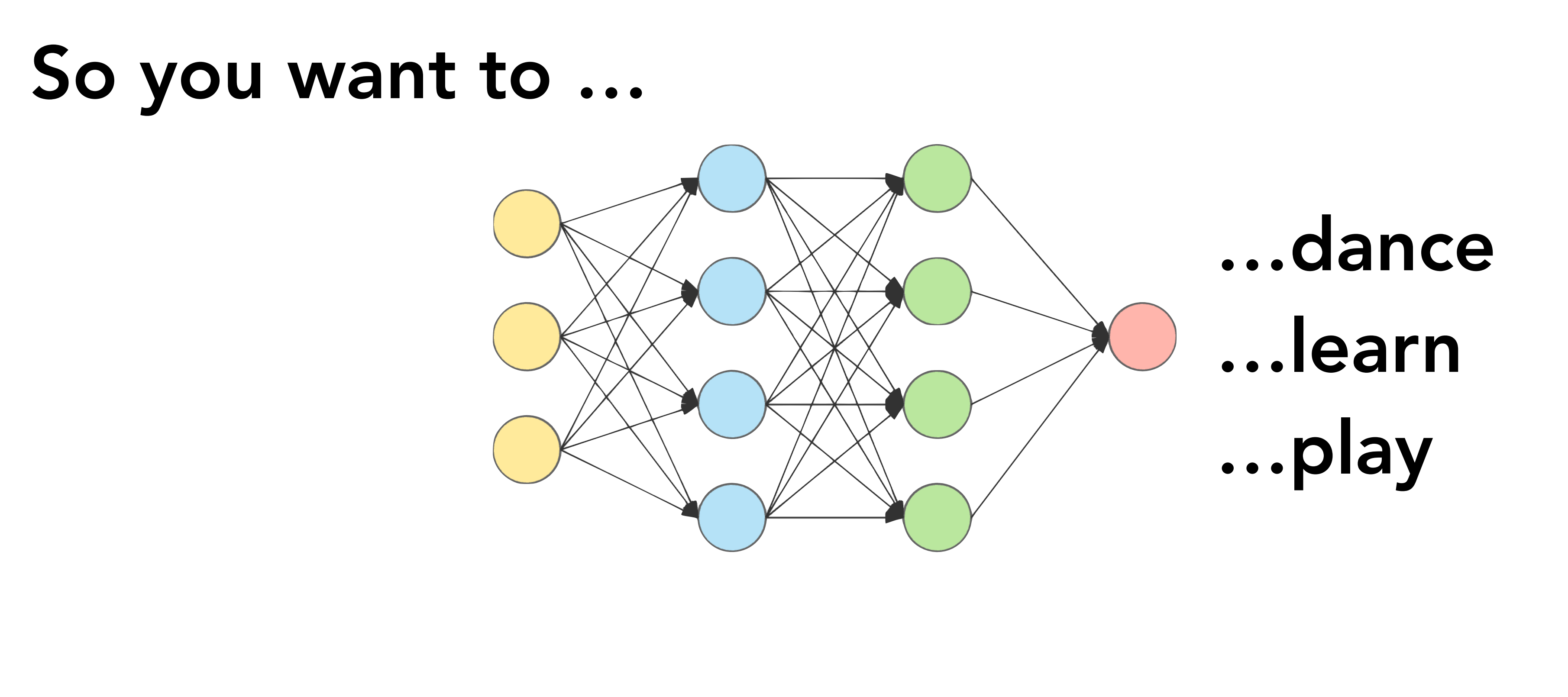
- Learn what NLP is about ✓
- Learn some basic ideas of machine learning (a statistical model) ✓
- See how a statistical model for predictive text works (what word should come next in this sentence?) ✓
- Learn the connections between this language model and models such as OpenAI's ChatGPT / GPT-4 models ✓

Outline

Natural Language Processing



Machine Learning



So you want to …

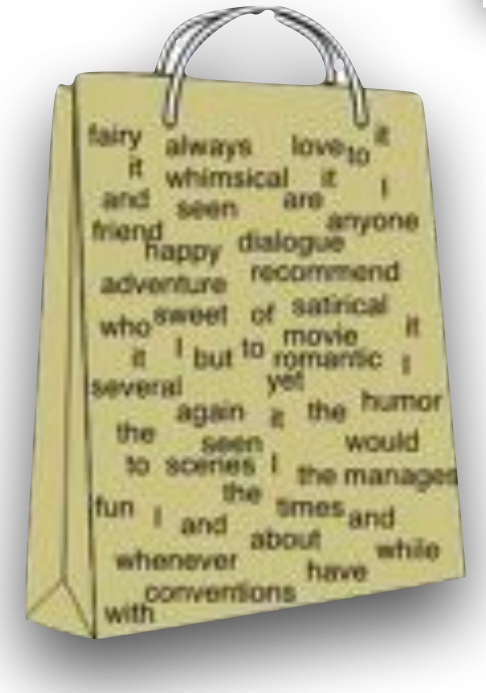
…dance

…learn

…play

Language Modeling

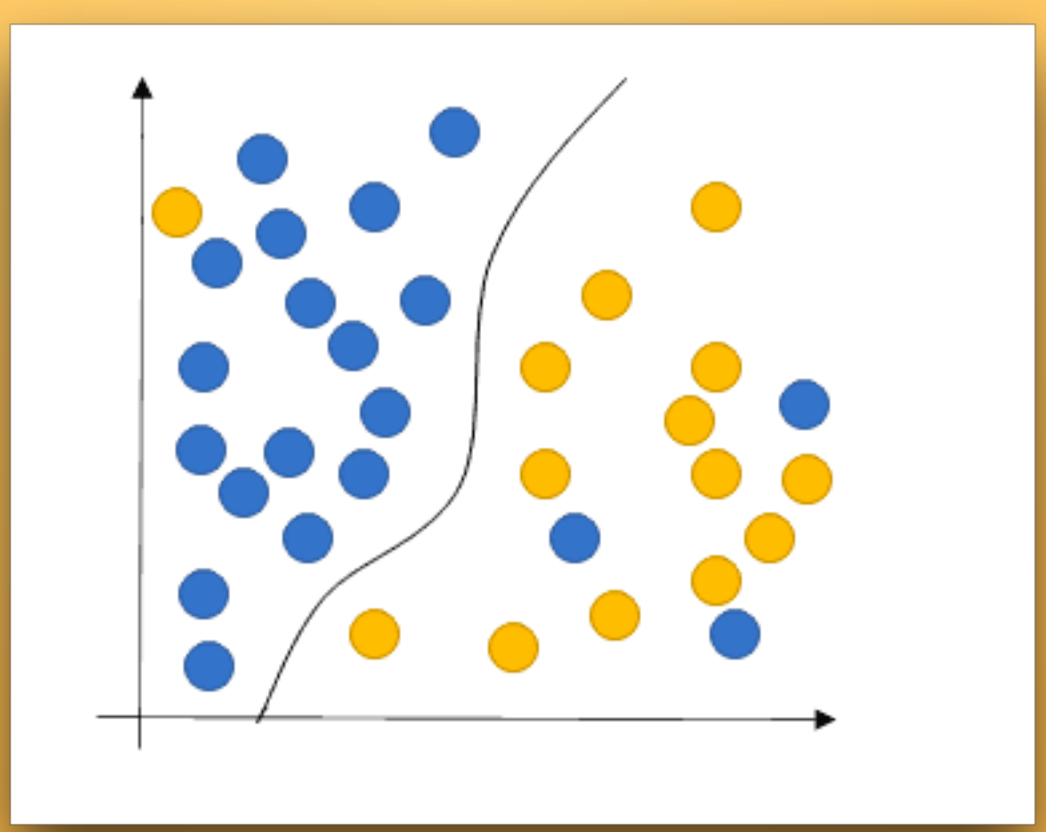
What's Next?



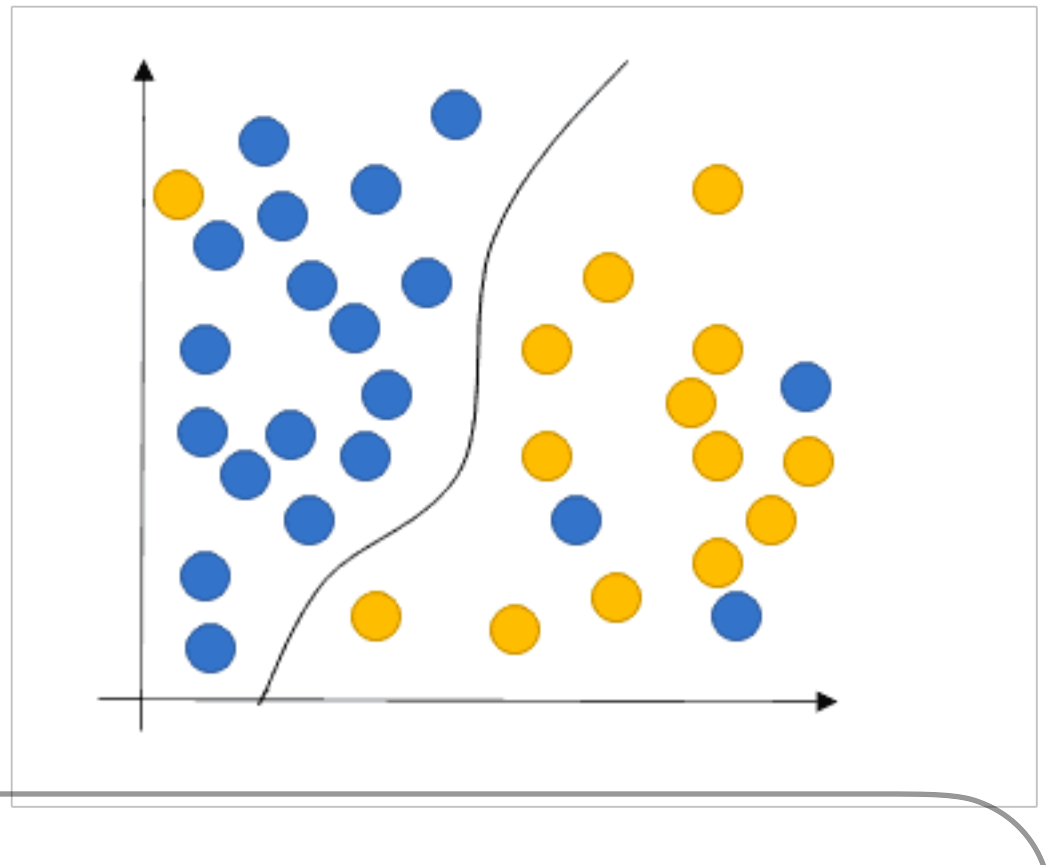
n-gram Language Models

Chat-GPT and other Large Language Models

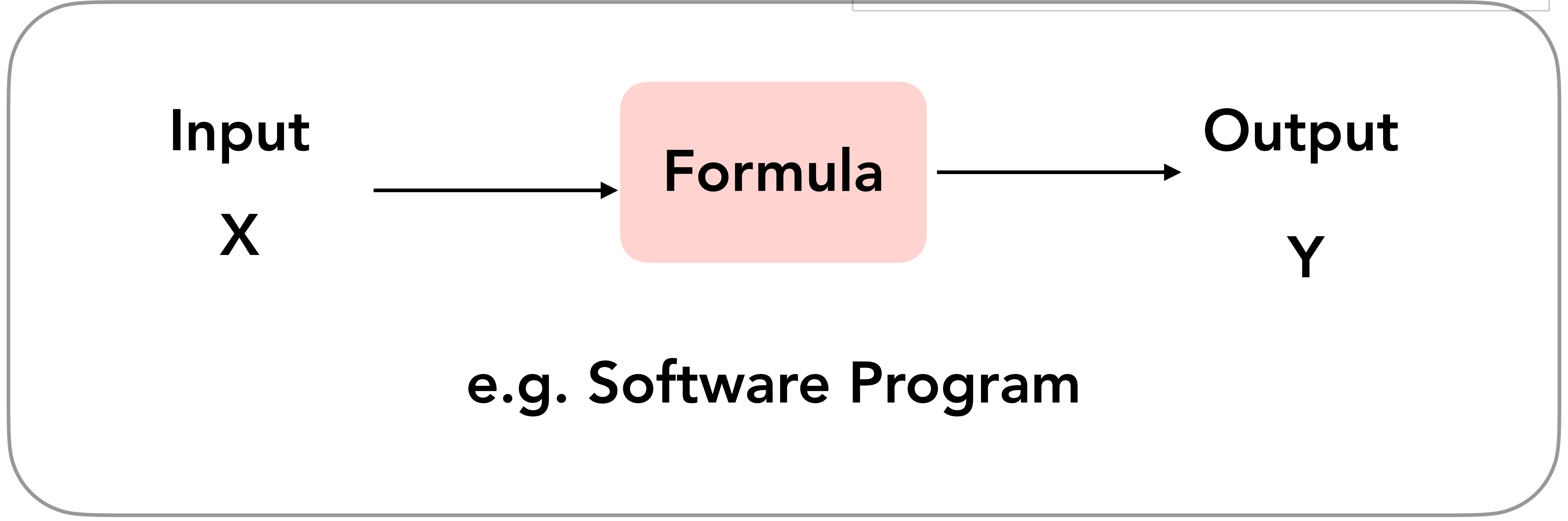
Machine Learning



Machine Learning

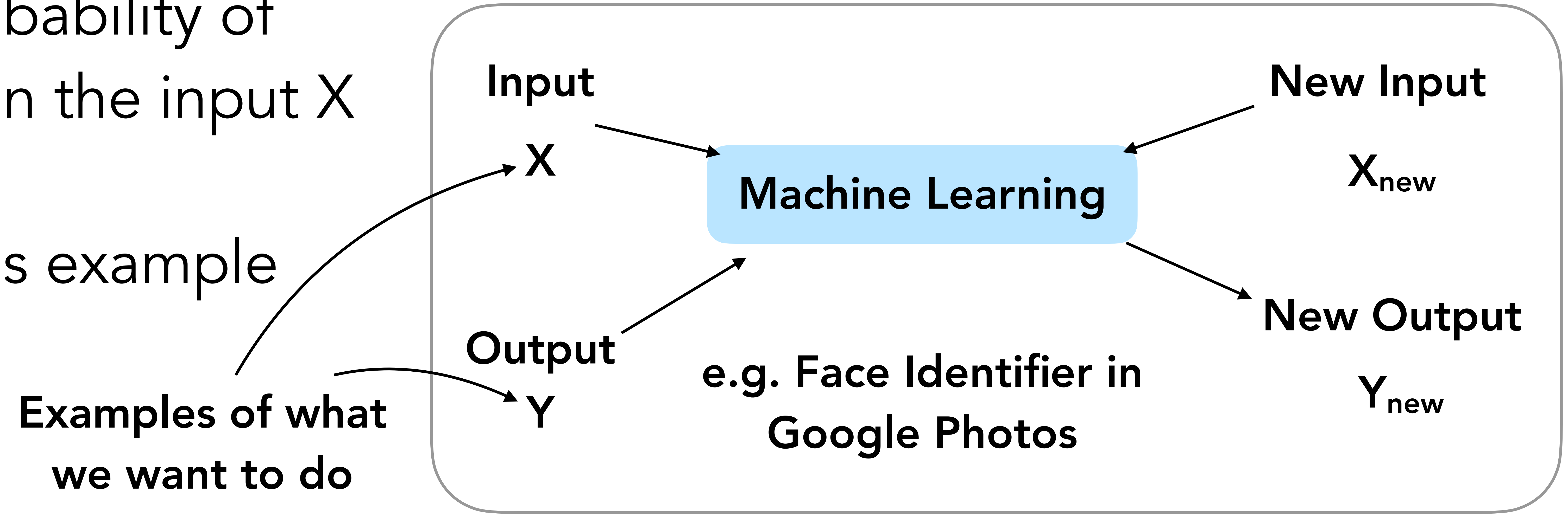


- All about predictions: Input X and Output Y
- In most real life problems, there is no simple formula to obtain Y from X
- Machine learning uses statistical analysis to figure out what would be the probability of the output Y, written as p(Y), given the input X
- Statistical Analysis: Lots of data as example pairs of input X and output Y



Input X → Formula → Output Y

e.g. Software Program



Input X → Machine Learning

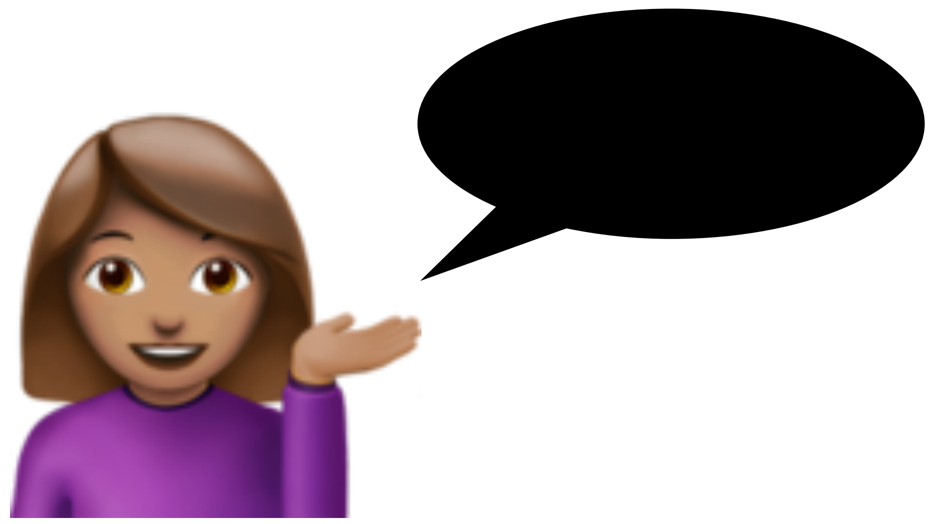
Output Y → Machine Learning

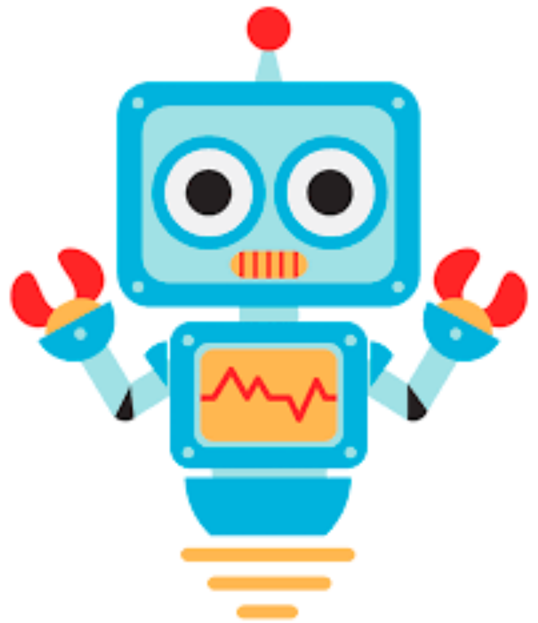
New Input X_{new} → Machine Learning → New Output Y_{new}

Examples of what we want to do

e.g. Face Identifier in Google Photos

Natural Language Processing and Machine Learning

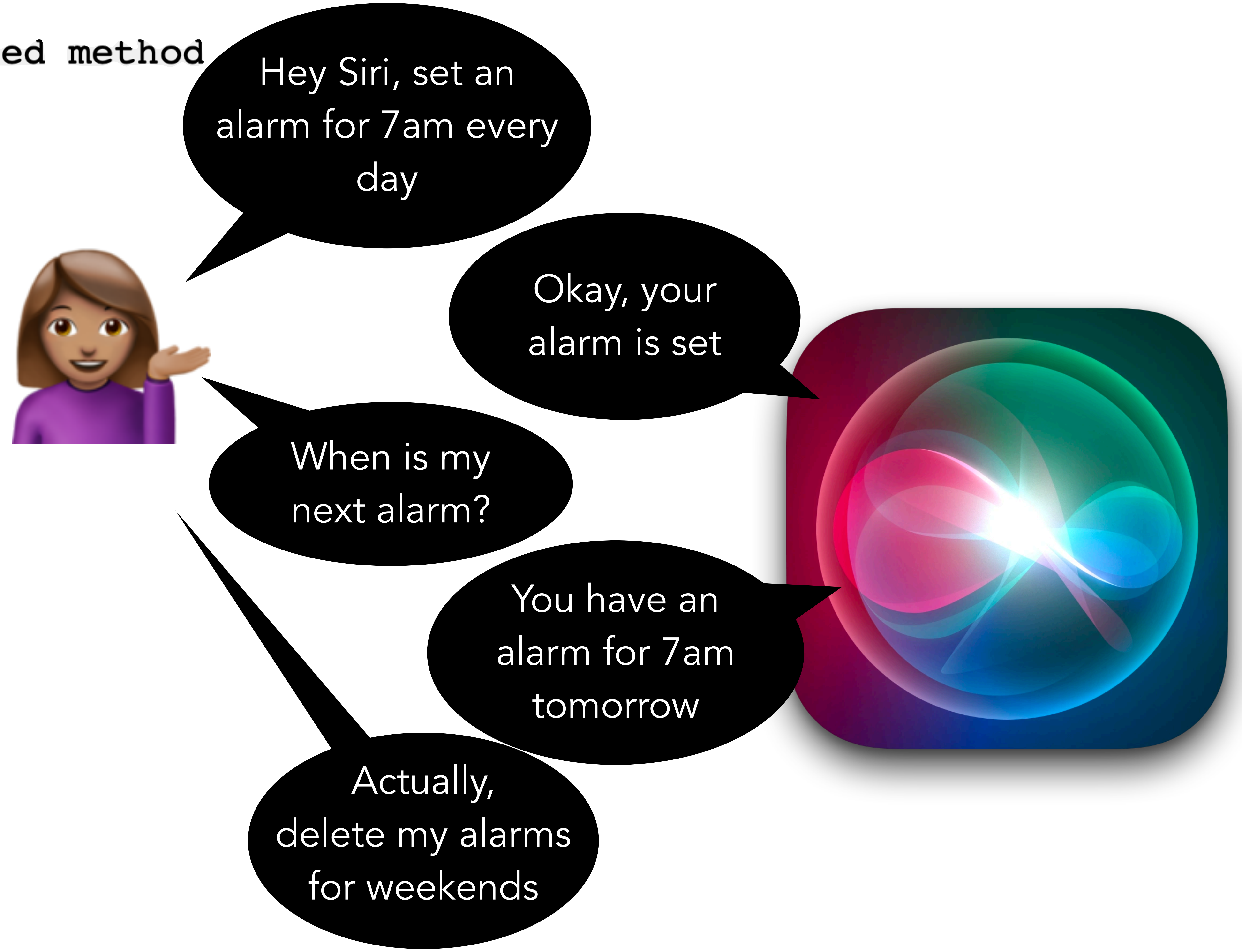




- Natural language processing uses a lot of ideas from machine learning
- Humans are good at understanding language. Computers are bad at it and it's hard to program them.
- If we see lots of examples of how humans do a task, can we teach a computer how to do it?

Building Siri

```
// Start by reading the user input with a predefined method
String userStr = readUserInput();
if (userStr.startsWith(“set a timer”))
  startTimerDialogue();
else if (userStr.startsWith(“set an alarm”) ||
         userStr.startsWith(“wake me up at”))
  startAlarmDialogue();
else [...]
```


Hey Siri, set an alarm for 7am every day
Okay, your alarm is set
When is my next alarm?
You have an alarm for 7am tomorrow
Actually, delete my alarms for weekends

- Too hard to list every case here!
- This is where machine learning comes in!

Analyze Movie Review Sentiment

Spider-Man: Across the Spider-Verse is an absolute triumph that takes everything we loved about the original film and cranks it up to a whole new level. This stunning sequel is a true testament to the power of animation, storytelling, and the enduring legacy of everyone's favorite web-slinger.

★★★★★

The Little Mermaid: To anyone who is planning on seeing this movie, I'd highly recommend to just wait until it comes out on Disney+ or something so you don't waste your money. I only went to see the movie because of my daughter and we can both say that this movie did not live up to our expectations. Furthermore, we both did not enjoy the majority of the movie at all.

★☆☆☆☆

Spider-Man: Across the Spider-Verse is an absolute triumph that takes everything we loved about the original film and cranks it up to a whole new level. This stunning sequel is a true testament to the power of animation, storytelling, and the enduring legacy of everyone's favorite web-slinger.

⭐⭐⭐⭐⭐

- Let’s try something simple:
  - (numberOfGoodWords, numberOfBadWords)

```
int numberOfGoodWords = computeNumGoodWords(review);
int numberOfBadWords = computeNumBadWords(review);
if (numberOfGoodWords > 3 && numberOfBadWords < 2)
  return “4 stars”;
else if (numberOfGoodWords > 2 && numberOfBadWords < 3)
  return “3 stars”;
else [...]
```

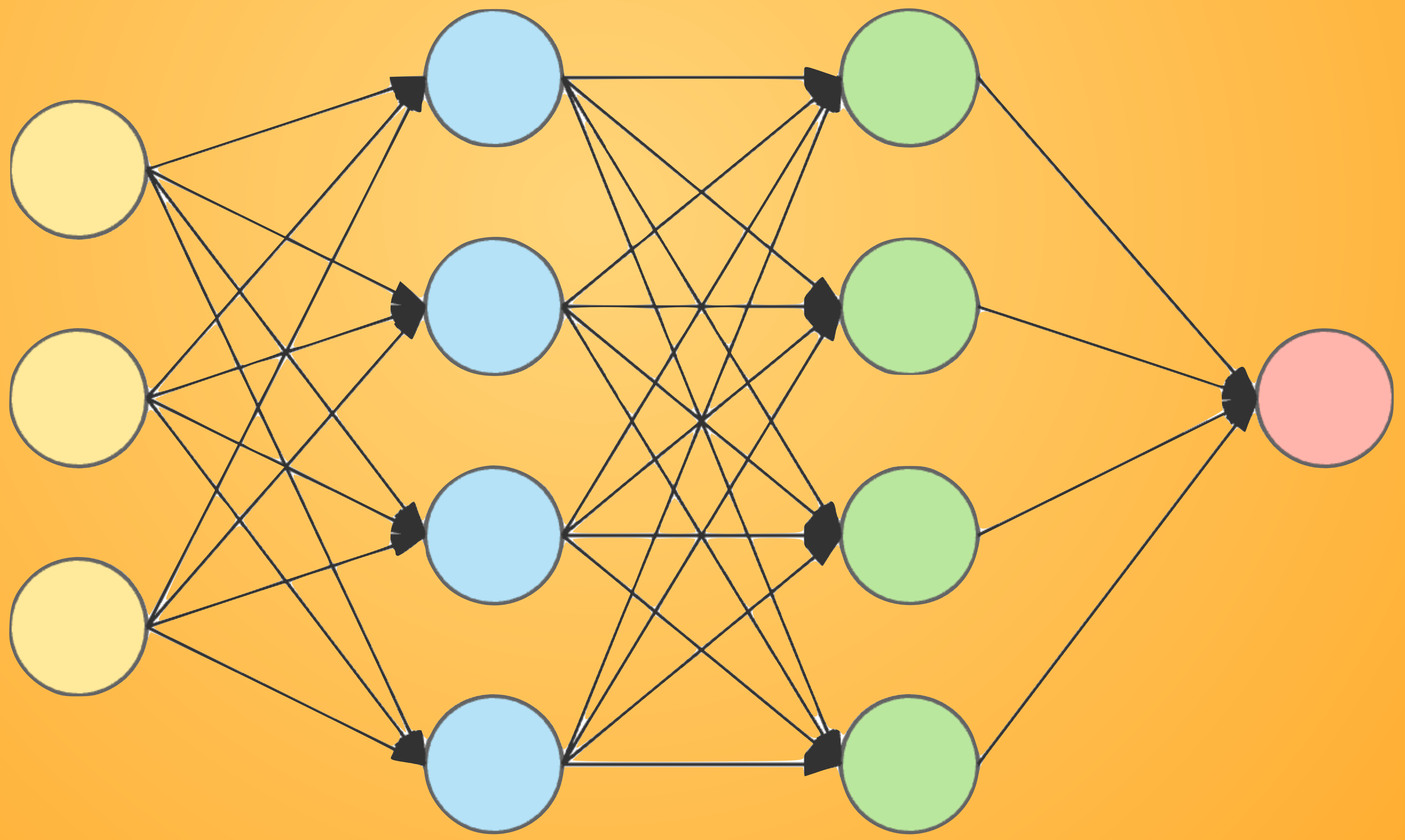
We can automatically generate this program!
(It’s called a decision tree)

Machine Learning Overview

- Lots of different models: decision trees, neural networks, Bayes Networks, …
- Machine Learning starts with a **feature representation** of this data: how do we represent it to a system?
  - We did for sentiment analysis with our variables, $(\texttt{numberOfGoodWords}, \texttt{numberOfBadWords})$
- **Neural networks** will view this as thousands of numbers (similar to how computers view programs as boolean codes) associated with each word.
- Let's use a probabilistic model for language modeling…
  - Very little math to implement…

Language Models

So you want to …



…dance
…learn
…play

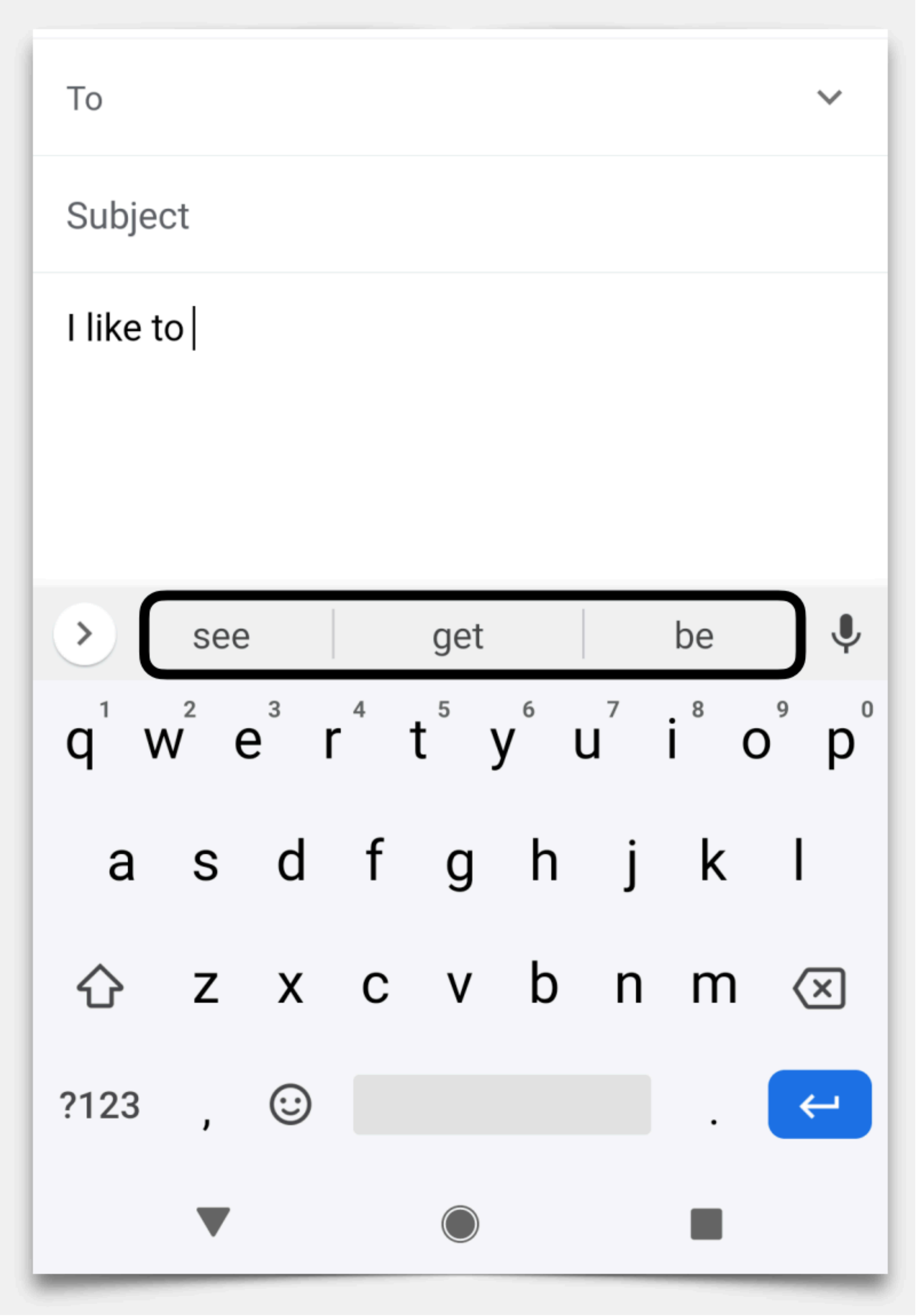
Language Models

- Task: Given a sequence of words so far (**the context**), predict what comes next.

- Like autocomplete!

- We never know for sure what comes next, but we can still make good guesses!

- Question: what is X and what is Y here?



To

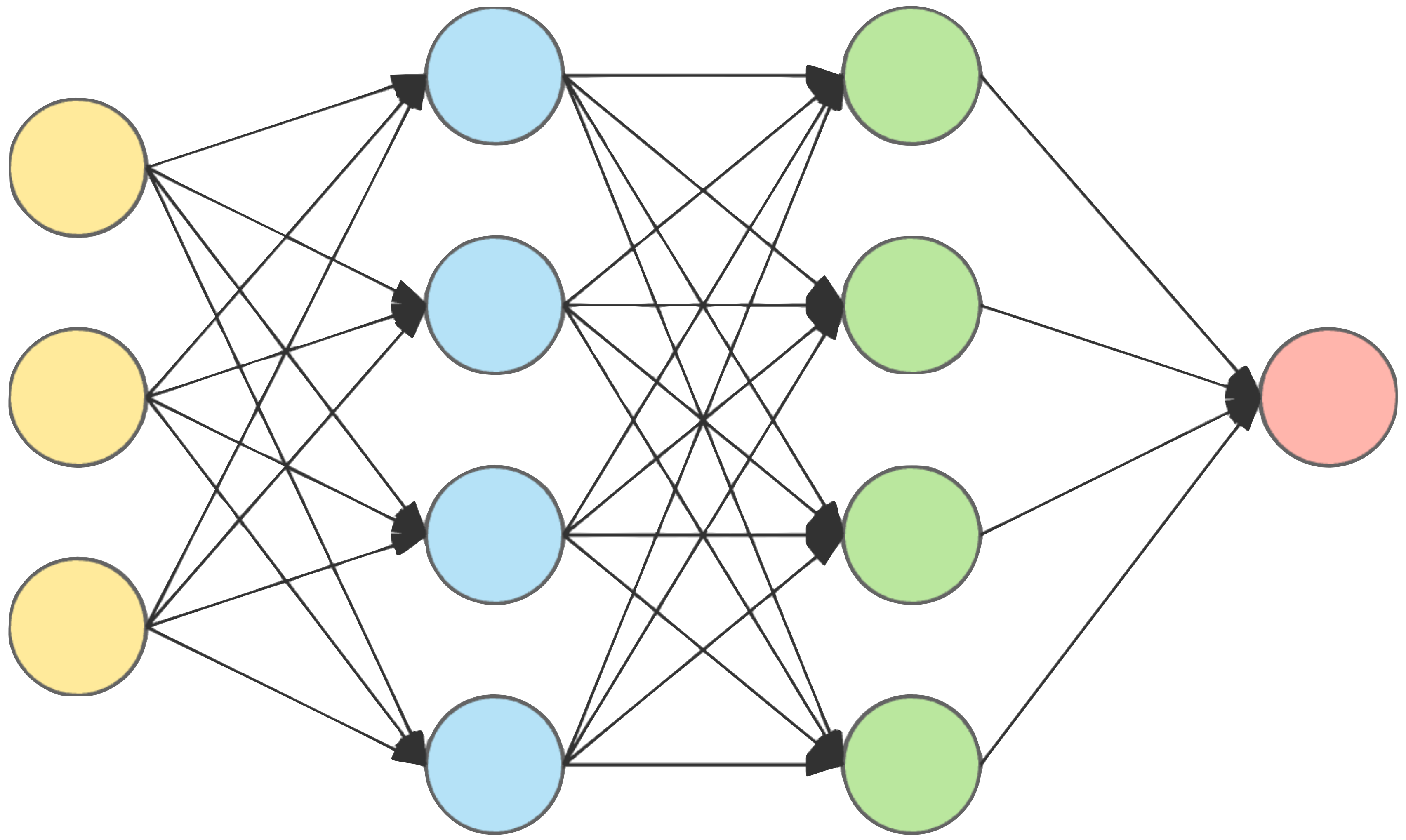
Subject

I like to

Building a Language Model

What words can follow this?

I want to …

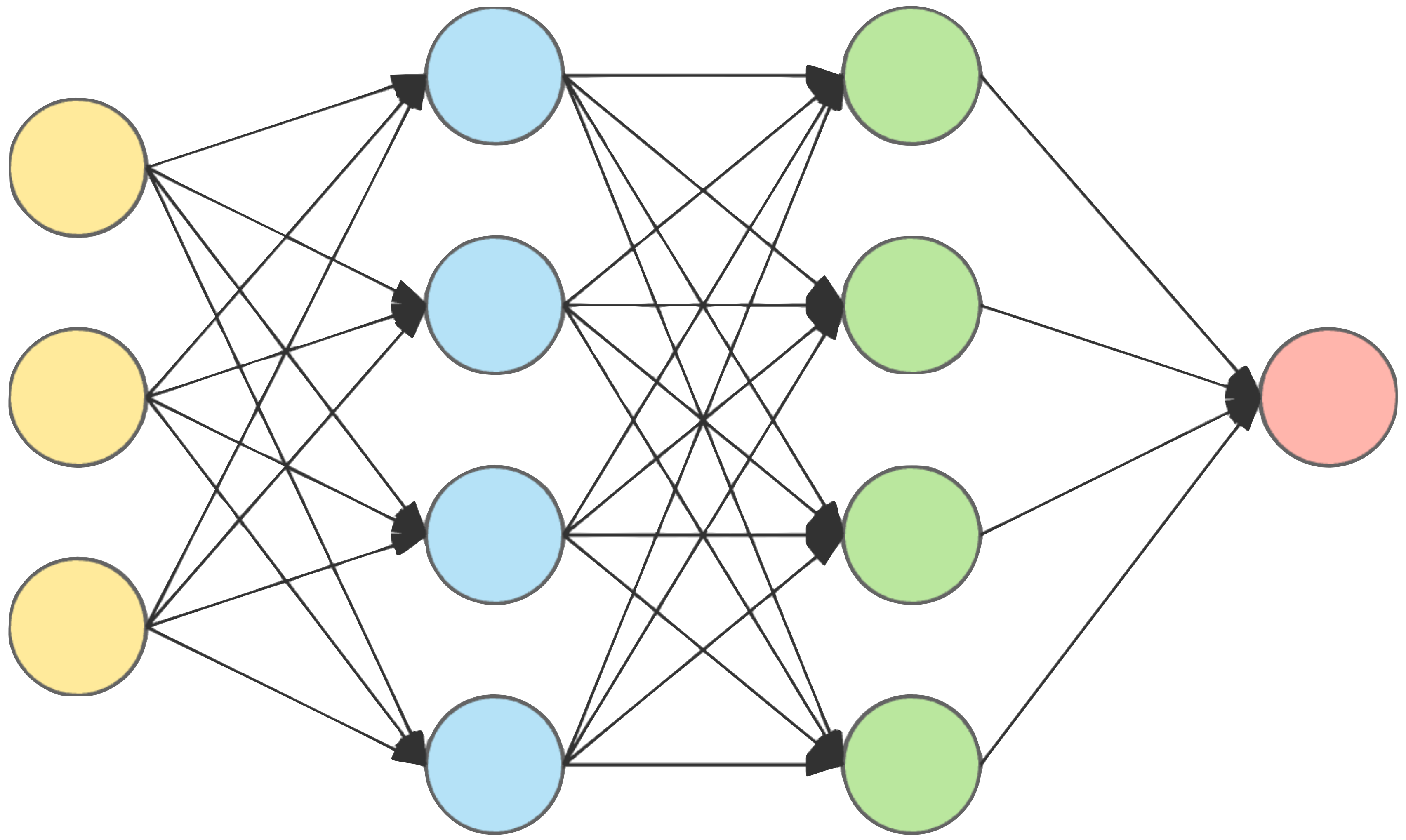


…dance
…learn
…play

What is common to these words?

Building a Language Model

What words can follow this?



The 44th President of United States was

…Barack Obama

I want to ______

Why Language Modeling?

Code computer programs!

Summarize articles, podcasts or presentations

Draft emails

Script social media posts

Create a title for an article

Play games.

Assist with job searches, including writing resumes and

Ask trivia questions.

Compose music!!!

Extremely powerful: can in many cases replace laborious manual efforts

ate product descriptions.

Describe complex topics more simply.

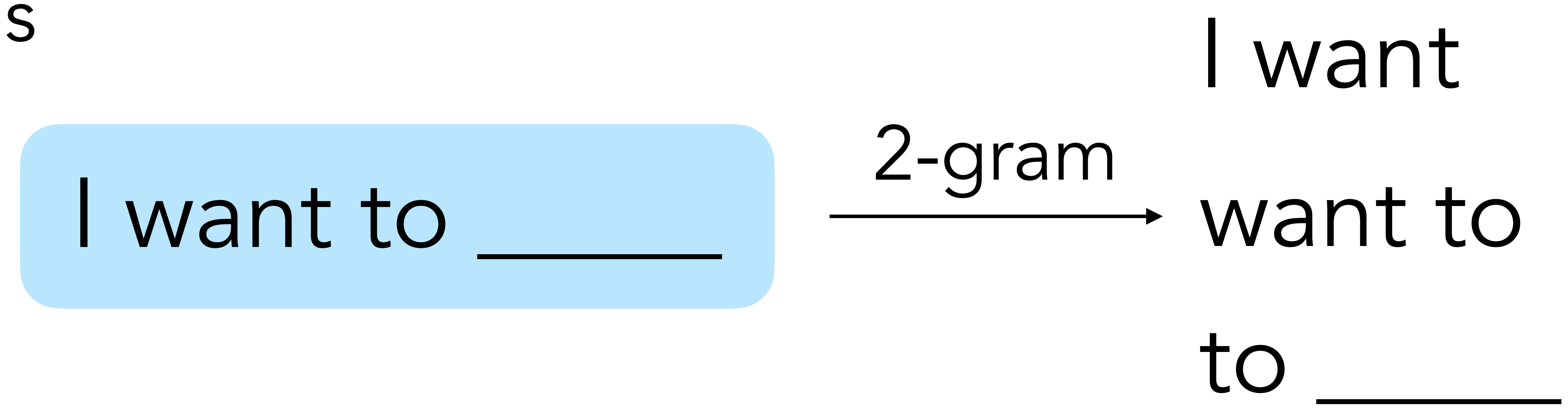
Solve math problems

Create articles, blog posts and quizzes for websites.

Reword existing content for a different medium, such as a presentation transcript for a blog post.

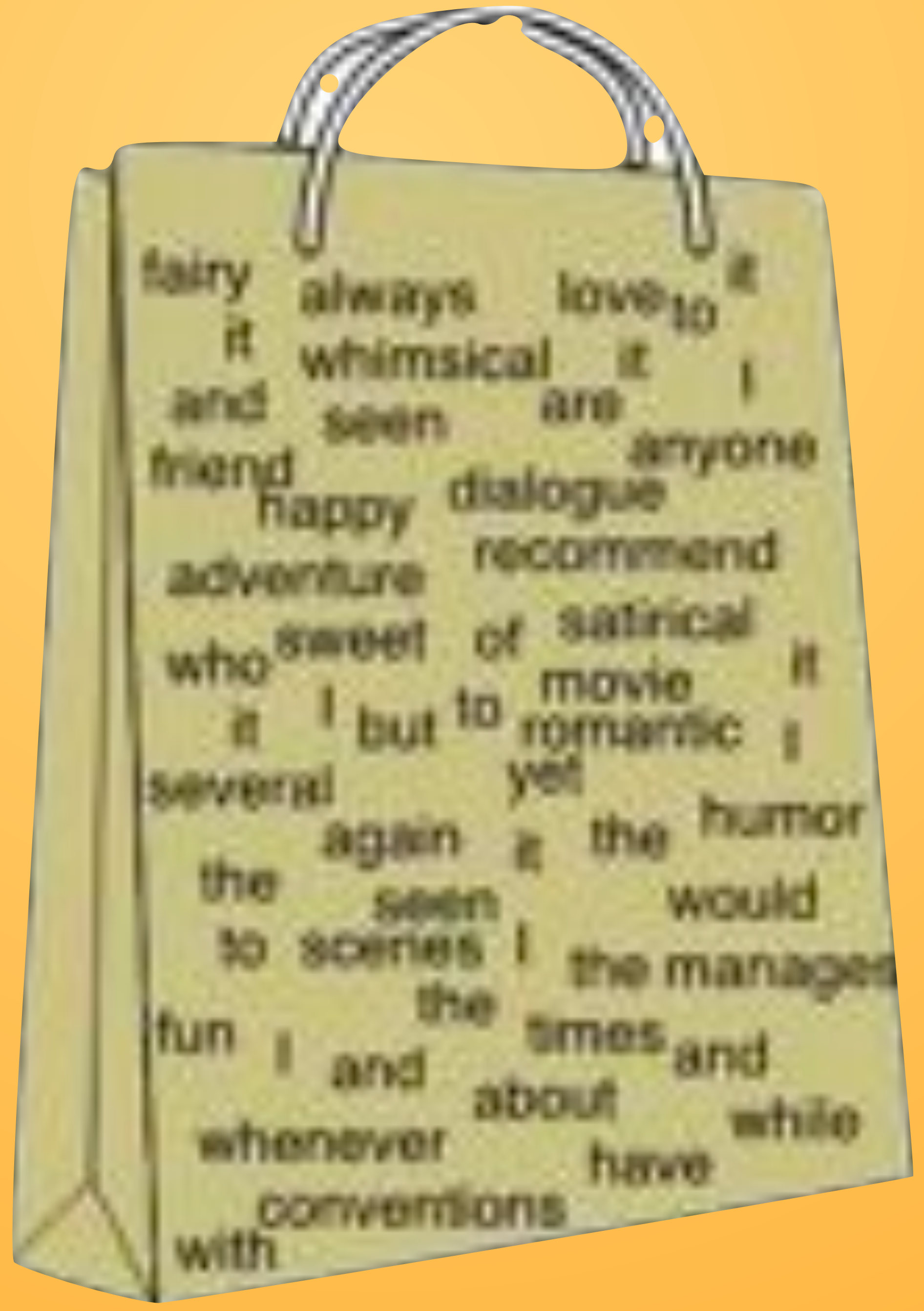
n-gram Language Modeling

- Our focus: build a model that predicts the next word based on the previous one or two words
- **n-gram**: a sequence of n words
  - *I like to* = 3-gram
  - *I really want to go* = 5-gram
- **n-gram language model**: predict the next word based on the previous n-1 words



How does a bigram context change the words that might follow?

Building n-gram Language Models



2-gram language models

This is a **conditional probability distribution**:

P(next word = y | previous word = x)

“the probability of the next word is y given that the previous word is x”

I want to _____

P(next word = was | previous word = to) = 0.0

P(next word = LA | previous word = to) = 0.2

P(next word = Europe | previous word = to) = 0.1

P(next word = Mexico | previous word = to) = 0.1

P(next word = eat | previous word = to) = 0.1 …

These have to add up to 1 over the vocabulary (every possible word y could be) “if we see to I think there’s a 20% chance the next word is LA”

Assume a **fixed vocabulary** of ~30,000 words

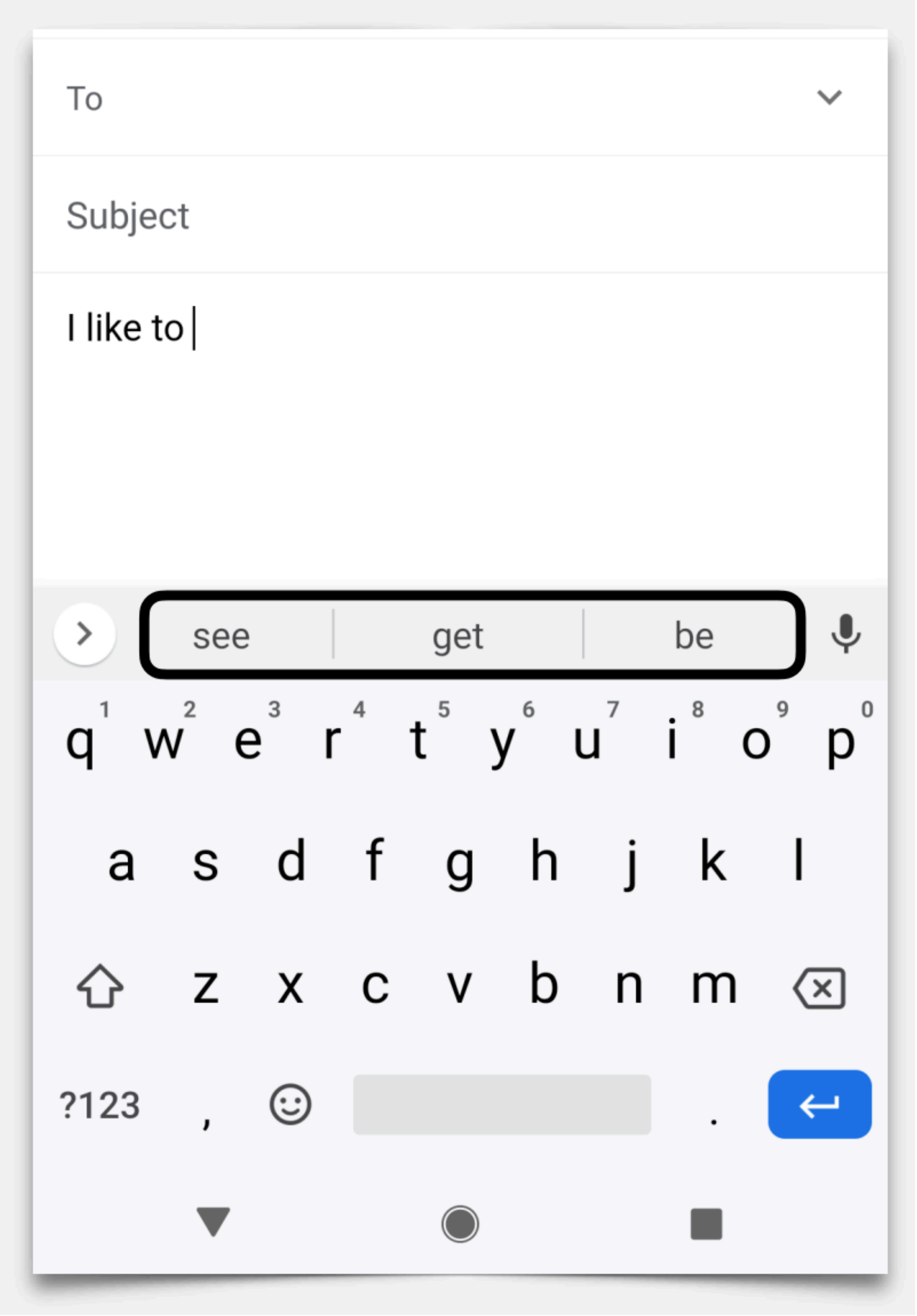
2-gram language models

- If we have these probabilities, we can build our predictive text system:

P(next word = _ | previous word = to)

Check all the possible words from that list, pick the ones with the highest probability (most likely next words)

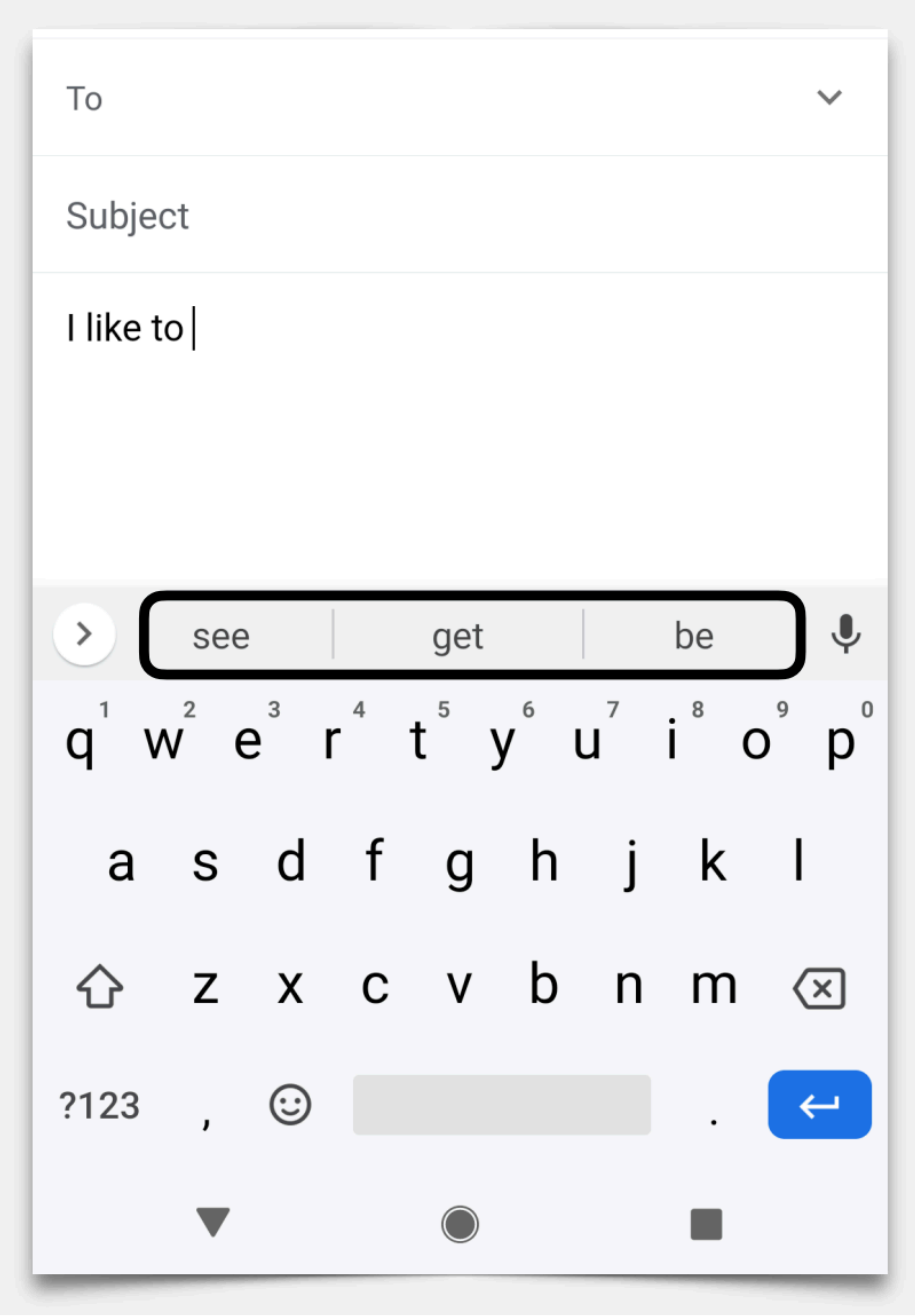
- Where do these probabilities come from? We're going to **learn them** from a bunch of text data we see



Lots and lots of text data

→ 2-gram LM probabilities →

Probability Estimation (Statistical Modeling)



To

Subject

I like to

Probability Estimation (Statistical Modeling)

Suppose we have a biased coin that's heads with probability p. p is a number between 0 and 1, and for a normal coin, p = 0.5 (equal probability of heads or tails).

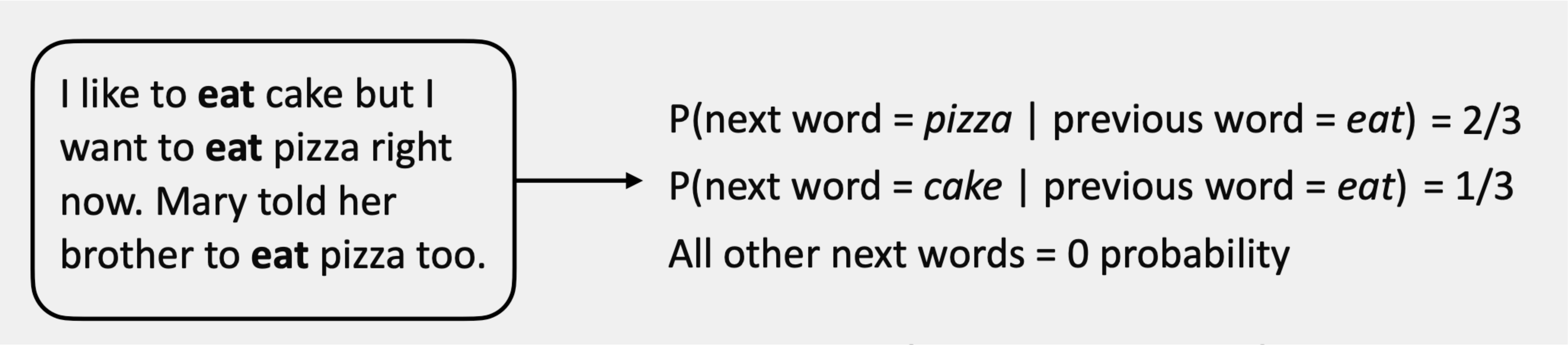
Suppose we flip the coin four times and see (H, H, H, T)

1. What do you think the probability p of heads is with this coin? Take a guess!

- We don't know what p is — p could be 0.5! But p = 3/4 = 0.75 maximizes the probability of the data. We'll say "this is the most likely value of p"
- The probability of the data is p*p*p*(1-p) — if you've taken calculus, you can take the derivative and set it equal to zero and find p = 0.75

n-gram Language Model

The decision for what words occur after a word w is exactly the same as the biased coin, but with 33,000 possible outcomes (different words) instead of 2.



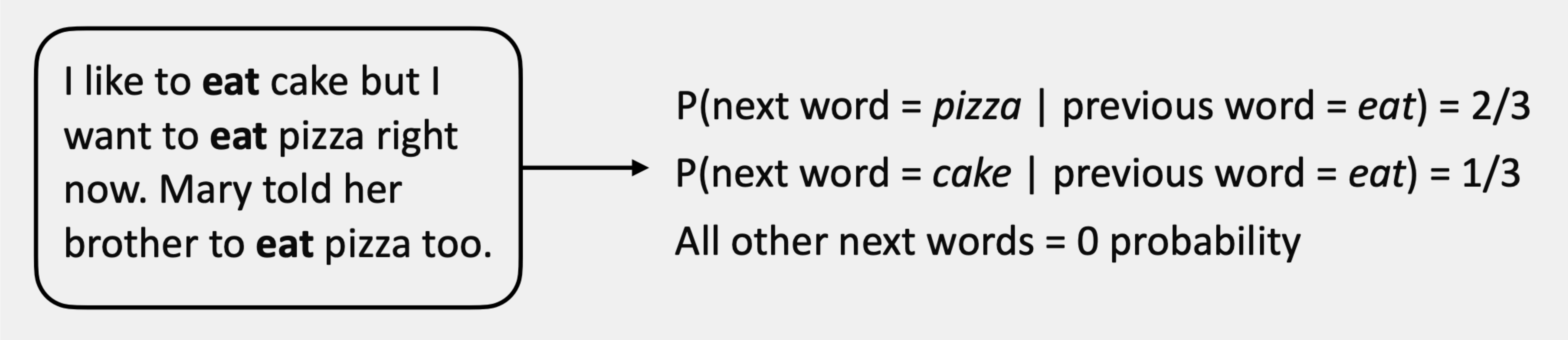
I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.

P(next word = *pizza* | previous word = *eat*) = 2/3

P(next word = *cake* | previous word = *eat*) = 1/3

All other next words = 0 probability

Smoothing



I like to **eat** cake but I want to **eat** pizza right now. Mary told her brother to **eat** pizza too.

P(next word = *pizza* | previous word = *eat*) = 2/3

P(next word = *cake* | previous word = *eat*) = 1/3

All other next words = 0 probability

- All other 29,998 words getting 0 probability just doesn't seem right. We want to assign some probability to other words
- We want to smooth the distribution from our counts

$$P(w \mid w_{\text{prev}}) = \lambda \frac{\text{count}(w_{\text{prev}}, w)}{\text{count}(w_{\text{prev}})} + (1-\lambda)\frac{\text{count}(w)}{\text{total word count}}$$

a number between 0 and 1 (like 0.9)

what we had before

a *unigram* LM

Neural Network Language Models

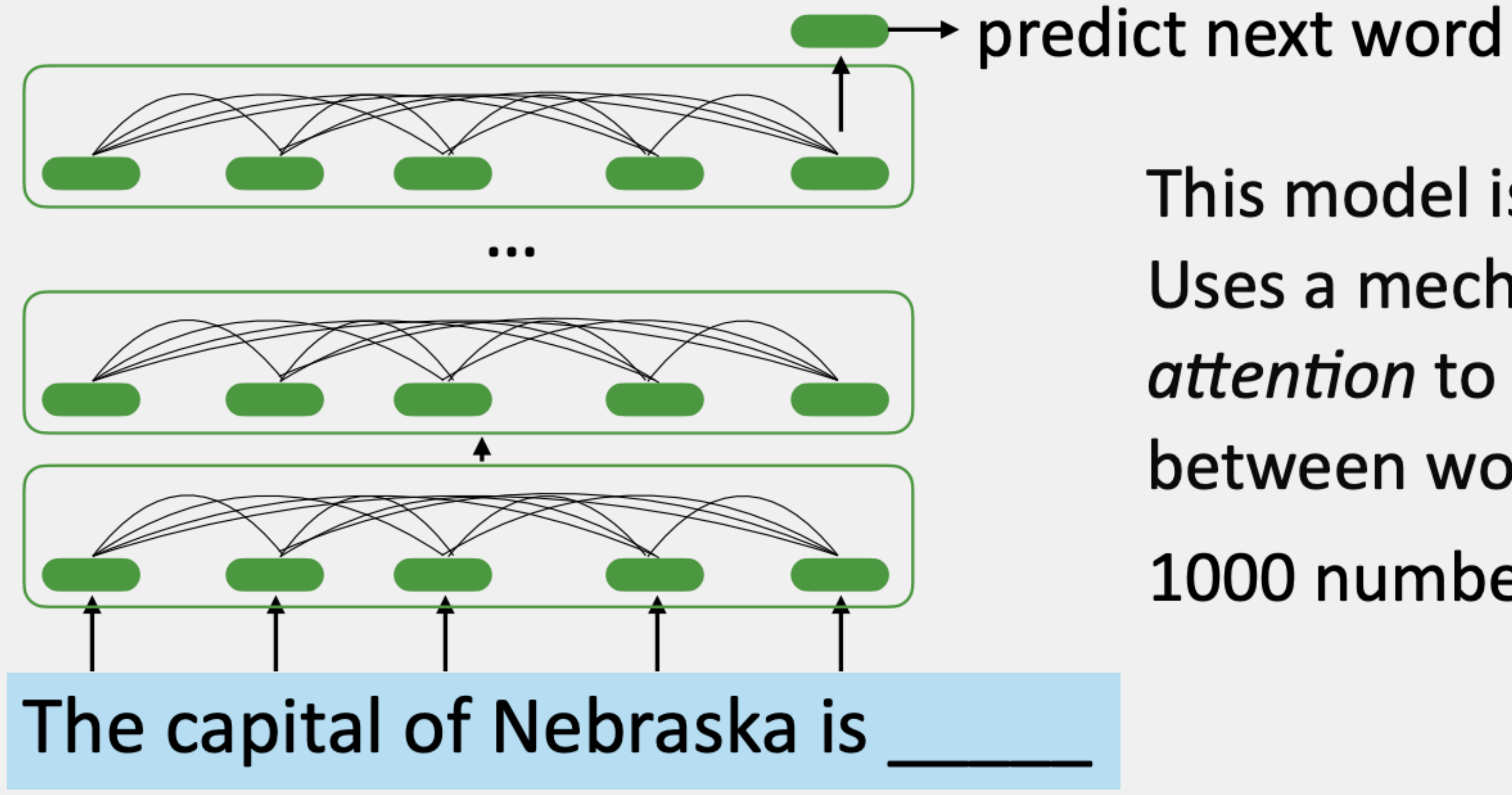
Neural networks are function that map f(context) → prediction

f is very, very complicated!

f(x) = 2x+3 has one input (x) and 2 parameters (2 and 3)

The f we use here has >1000 inputs and >1 million parameters!

These can be learned from data using derivatives from calculus



The capital of Nebraska is _____

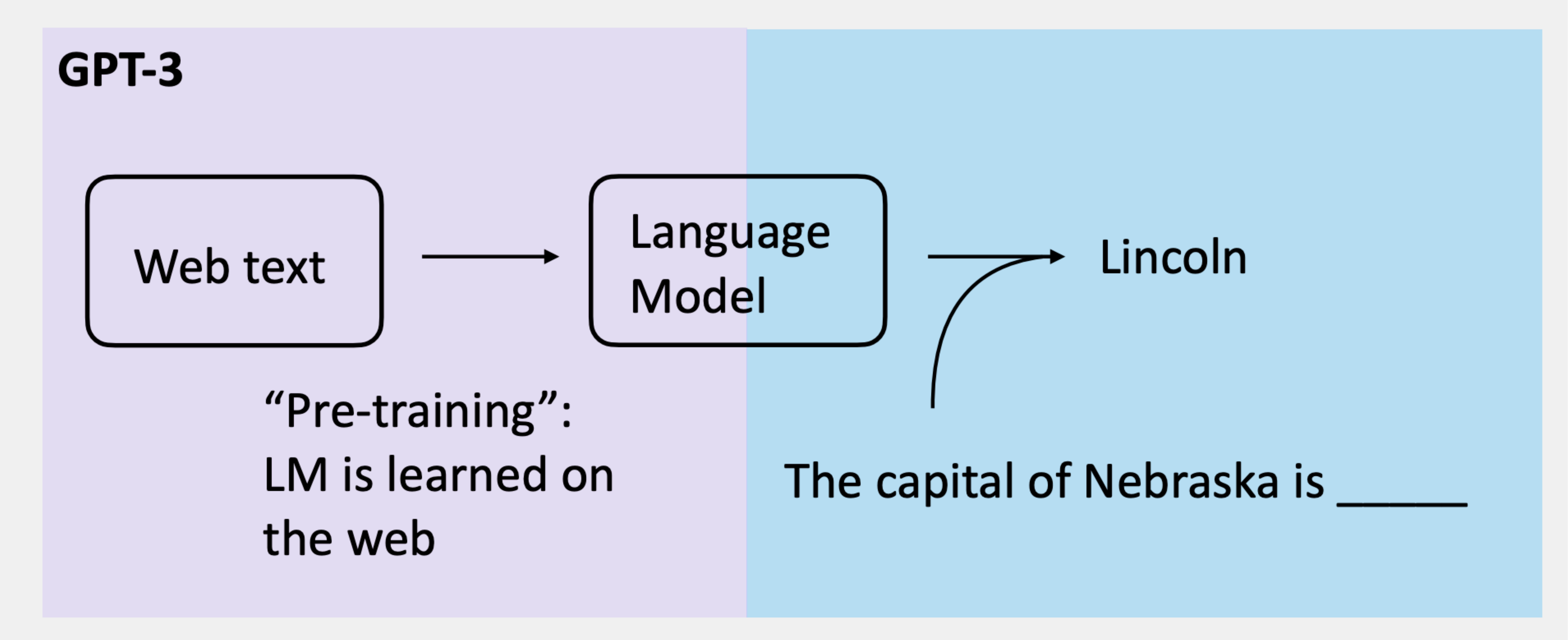
This model is called a Transformer. Uses a mechanism called self *attention* to capture interactions between words

1000 numbers x 5 words = 5000 inputs

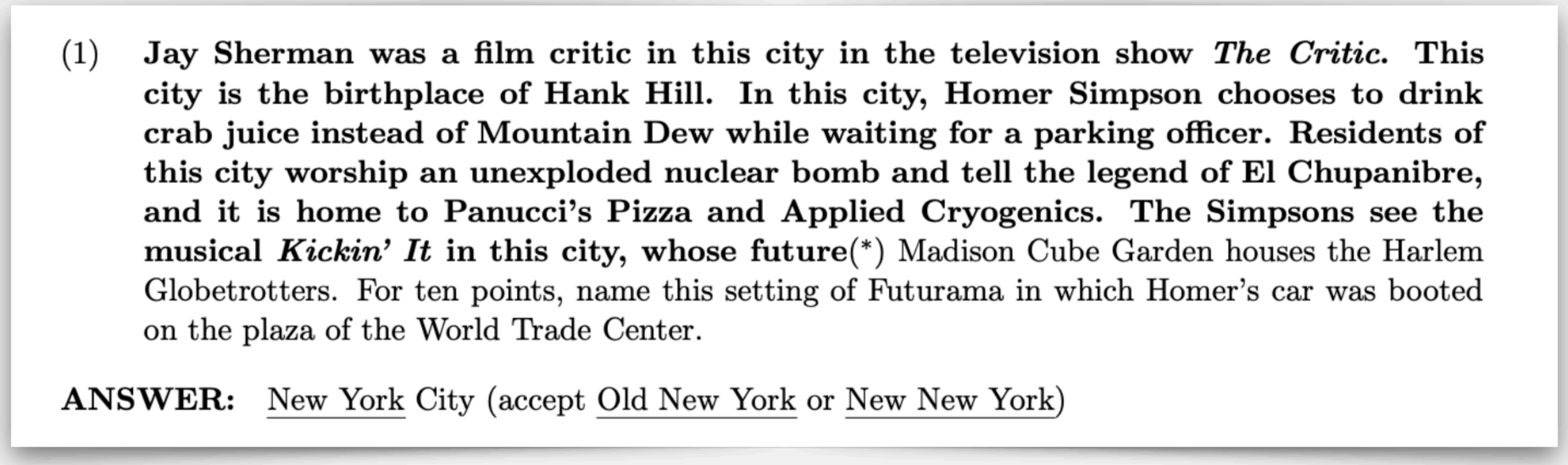
Chat-GPT and other Large Language Models

Using Large Language Models

- These models are trained over a ton of data (a curated scrape of the web). So they will have seen information about Nebraska and Lincoln.
- A big enough model can answer questions even without being trained to do so. What else can we get these models to do?



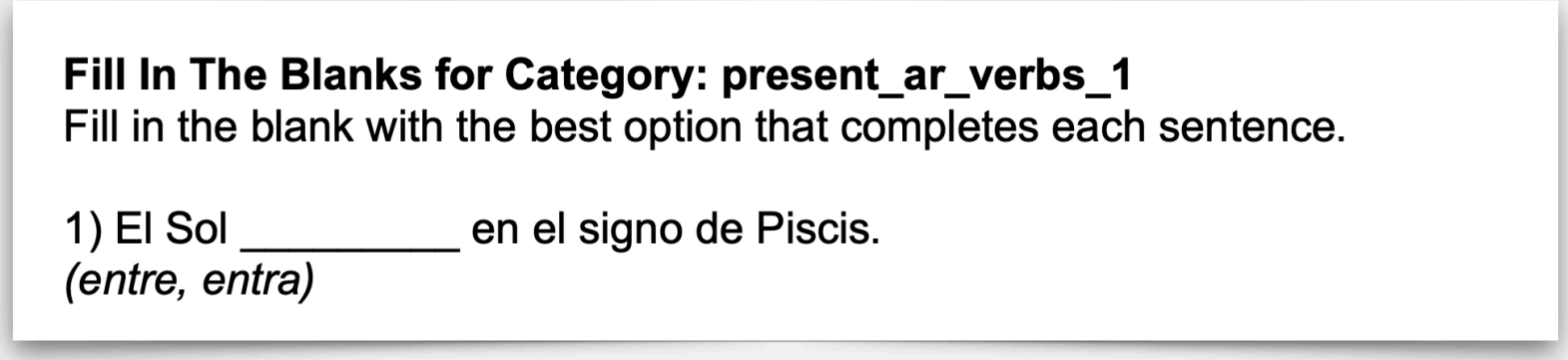
Why does this work?



(1) **Jay Sherman was a film critic in this city in the television show *The Critic*. This city is the birthplace of Hank Hill. In this city, Homer Simpson chooses to drink crab juice instead of Mountain Dew while waiting for a parking officer. Residents of this city worship an unexploded nuclear bomb and tell the legend of El Chupanibre, and it is home to Panucci's Pizza and Applied Cryogenics. The Simpsons see the musical *Kickin' It* in this city, whose future**(*) Madison Cube Garden houses the Harlem Globetrotters. For ten points, name this setting of Futurama in which Homer's car was booted on the plaza of the World Trade Center.

ANSWER: New York City (accept Old New York or New New York)

https://quizbowlpackets.com/



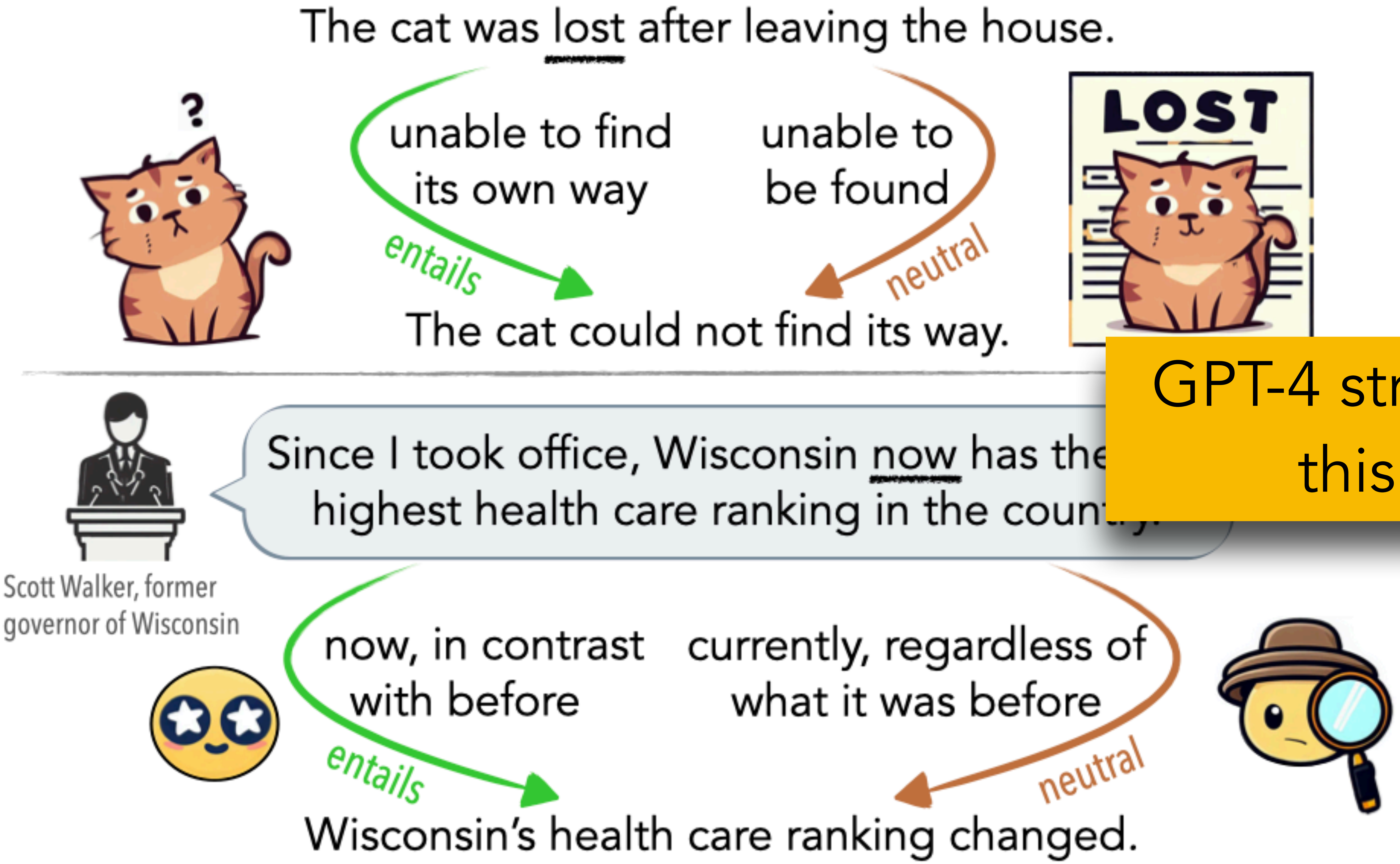
Fill In The Blanks for Category: present_ar_verbs_1
Fill in the blank with the best option that completes each sentence.

1) El Sol __________ en el signo de Piscis.
(entre, entra)

https://www.123teachme.com/spanish_worksheets/list_all

The model has really seen how to do a lot of tasks already when it was being built!

But, LLMs are not perfect…



The cat was lost after leaving the house.

unable to find its own way

unable to be found

The cat could not find its way.

Since I took office, Wisconsin now has the highest health care ranking in the country.

Scott Walker, former governor of Wisconsin

now, in contrast with before

currently, regardless of what it was before

Wisconsin's health care ranking changed.

GPT-4 struggles on this task!

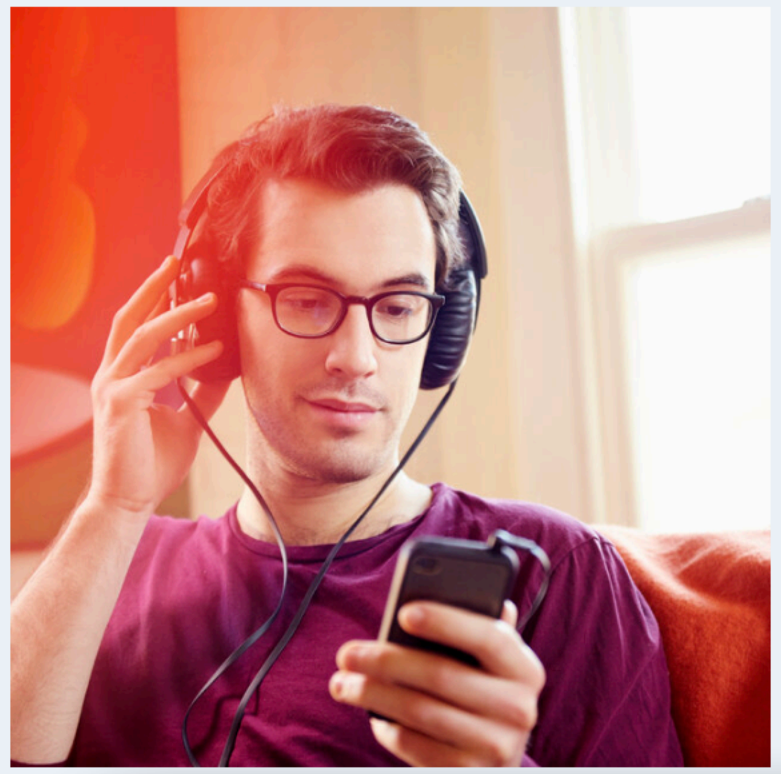
Ethical Concerns

Can We No Longer Believe Anything We See?

By Tiffany Hsu and Steven Lee Myers

April 8, 2023

Which image was created by artificial intelligence? Click on your guess



An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World

A track like "Heart on My Sleeve," which went viral before being taken down by streaming services this week, may be a novelty for now. But the legal and creative questions it raises are here to stay.

Give this article 215

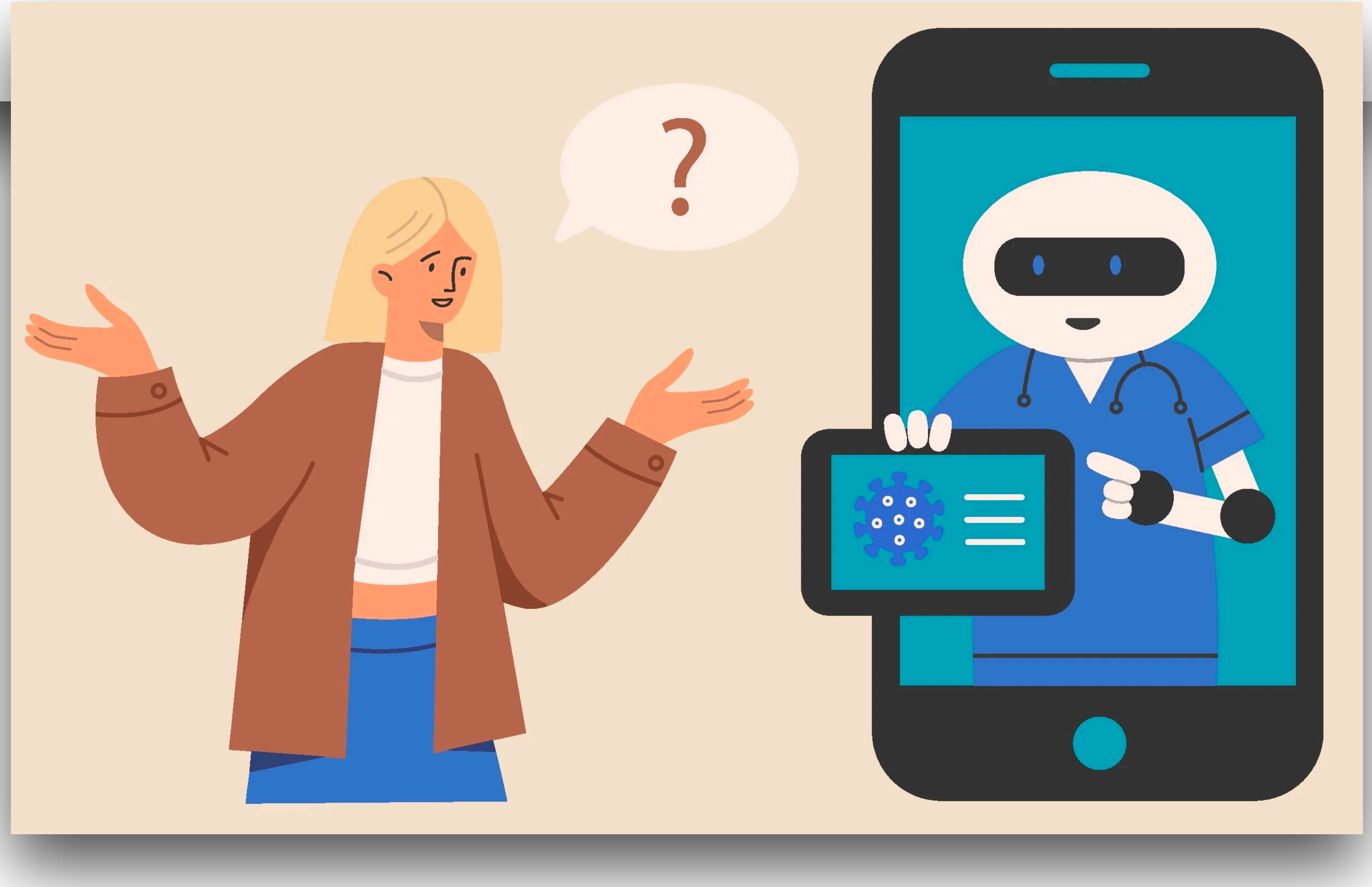
This Tool Could Protect Artists From A.I.-Generated Art That Steals Their Style

Artists want to be able to post their work online without the fear "of feeding this monster" that could replace them.

Mar 8, 2023 - Technology

Chatbot therapy, despite cautions, finds enthusiasts

Peter Allen Clark



A.I.-Generated Content Discovered on News Sites, Content Farms and Product Reviews

The findings in two new reports raise fresh concerns over how artificial intelligence may transform the misinformation landscape online.

What's Next?

Courses to take

- How to study this more?
- Math to learn: probability, linear algebra
- Machine learning or data science online courses
    - Andrew Ng's Coursera course: https://www.coursera.org/learn/machine-learning
    - Sentiment Analysis tutorial: https://realpython.com/sentiment-analysis-python/
- More programming or software engineering can help
    - Python

Further Reading

- Understanding more about neural networks: Chris Olah, Jay Alammar
    - https://colah.github.io/
    - https://jalammar.github.io/
- Latest big language models:
    - https://openai.com/blog/better-language-models/
    - https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Thank you!

My lab

· DILL ·

You can find these slides here: