2585524

slido

Join at slido.com
#2585524

Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

CC BY NC SA

2585524

LECTURE 27

Intro to LLMs & Conclusion

What you learned in this class, and what you will learn in the future. A farewell.

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

Credit for the LLM slides:
Joseph E. Gonzalez
jegonzal@berkeley.edu

CC BY NC SA

2585524

Today's Roadmap

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models

Logistics

What did you learn in Data C100/C200?

What's Next?

Join course staff!

CC BY NC SA

2585524

Introduction to LLM

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models

Logistics

What did you learn in Data C100/C200?

What's Next?

Join course staff!

CC BY NC SA

slido

2585524

How often do you use GenAI (e.g., Gemini, ChatGPT, Claude)

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

CC BY NC SA

slido

2585524

Which GenAI technologies have you used?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

CC BY NC SA

Machine Learning with Features

2585524

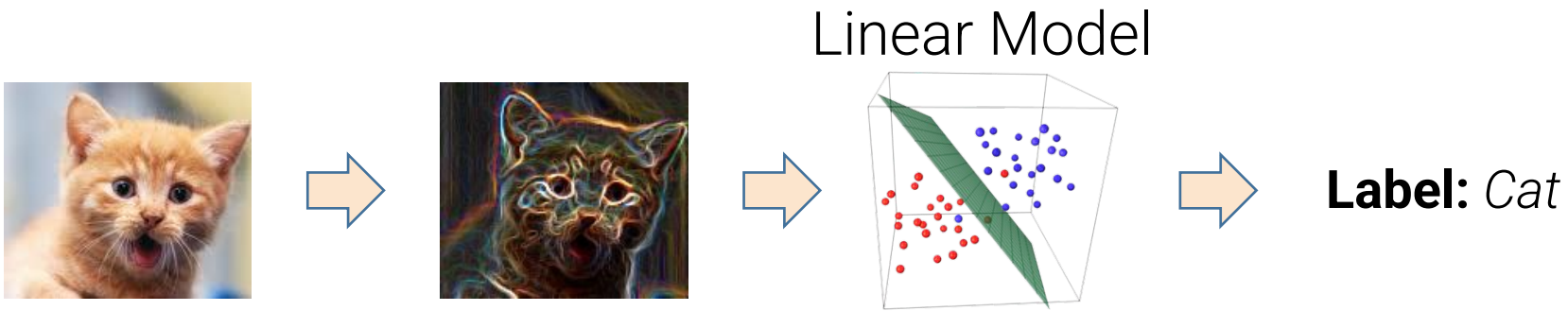
- Break function into features and learned model

$$f_\theta(x) = g_\theta\left(\phi(x)\right)$$

Generic Easy to Train Function (Linear Model)

Hand Engineered Features

- Example

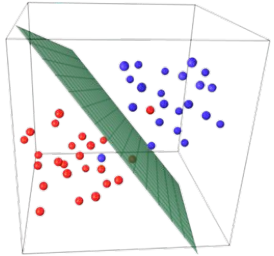


CC BY NC SA

Function Approximation Pipeline

2585524

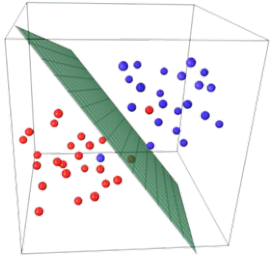
Linear Model



Label: *Cat*

Often build multiple layers of features to abstract the input.

Linear Model



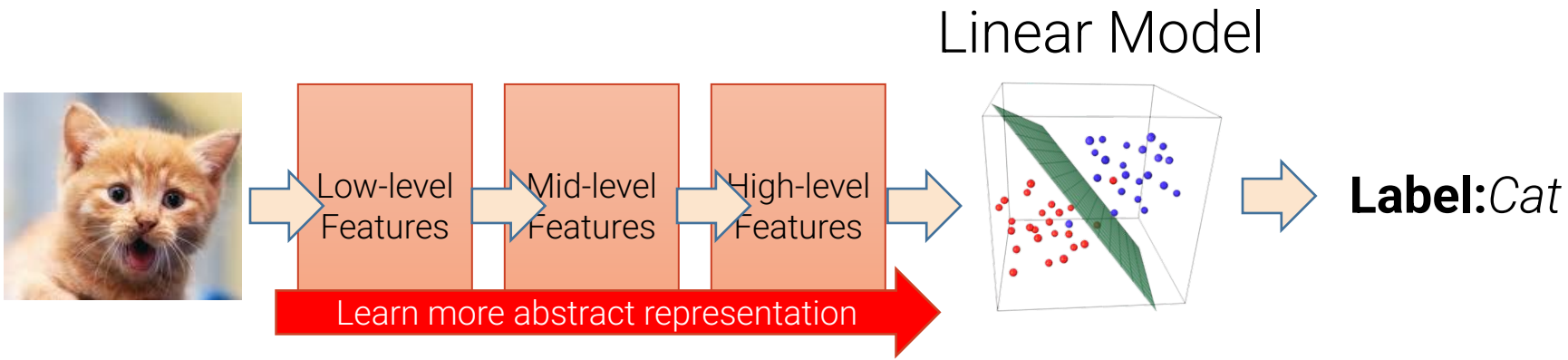
Label: *Cat*

Deep Learning tries to automate this process.

CC BY NC SA

Function Approximation Pipeline

2585524

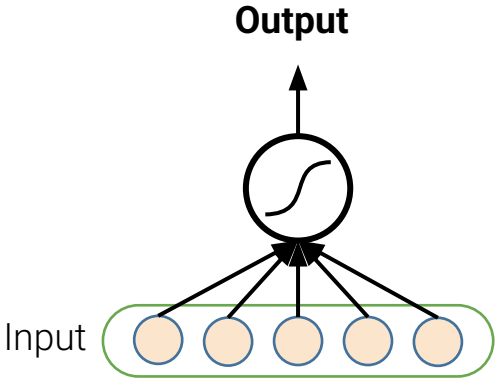


Deep Learning: automatically *learns a deep hierarchy of abstract features* as part of the model.

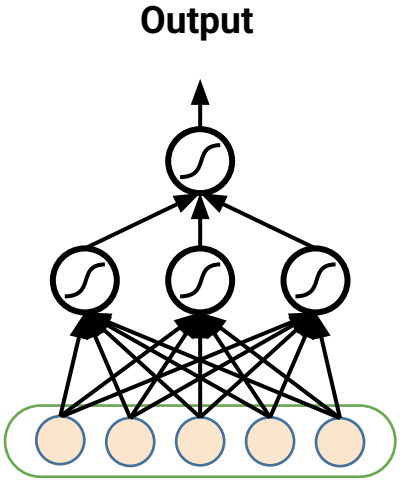
- Typically using neural networks
  - Composable general-purpose function

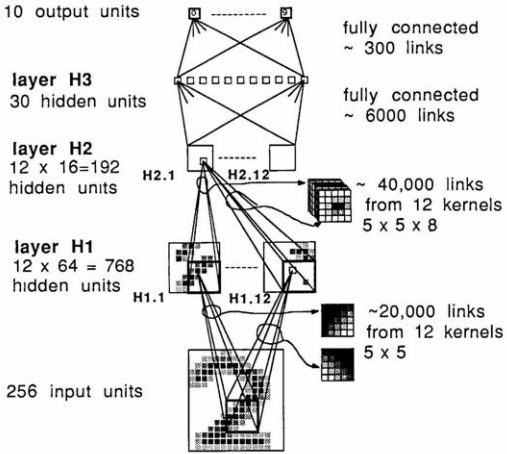
CC BY NC SA

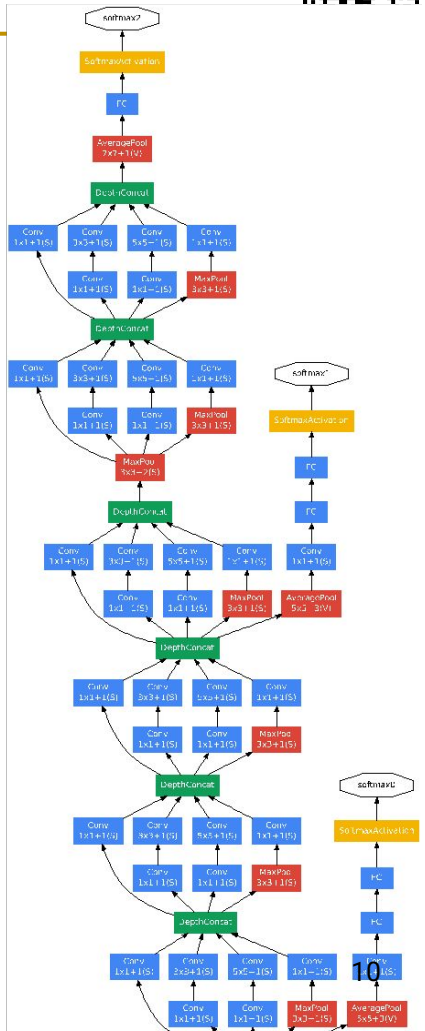
Deep Learning → Many layers …



Logistic Regression



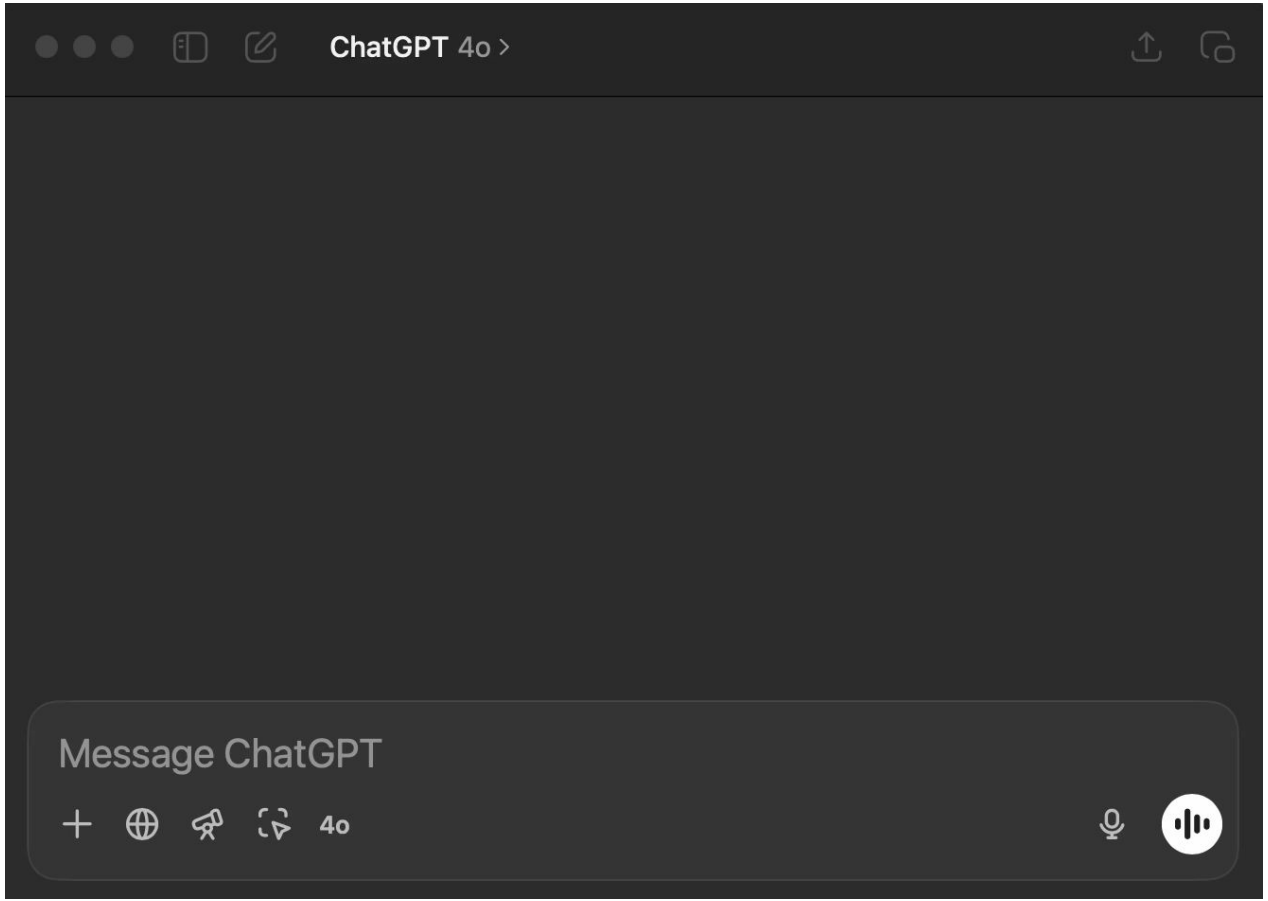




CC BY NC SA

What is ChatGPT?

2585524



CC BY NC SA

What Is ChatGPT?

2585524

Chat: natural language system

Explained Today

G: Generatively – Designed to model the creation of text

P: Pretrained – Trained on lots of naturally occurring data

T: Transformer – A kind of neural network architecture

Goals for Today

What is it?

How does it work?

Why is it changing the world?

Chat GPT is just one example of a

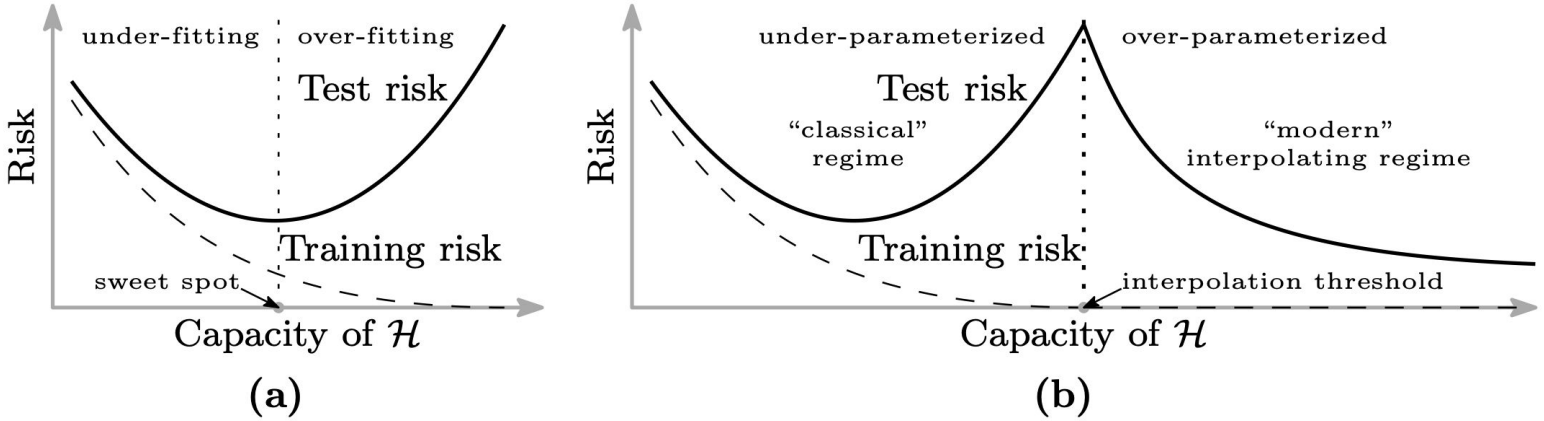
Large Language Model (LLM)

CC BY NC SA

What Is a Large Language Model?

2585524

- **Large**: The model parameters (θ) are *BIG!*
  - *BILLIONS of PARAMETERS!!!!*
  - *Parameters: numbers encoding the neurons in the model*
- **Language Model:** Predicting language (e.g., words)



Reconciling modern machine learning practice and the bias-variance trade-off

Deep Double Descent: Where Bigger Models and More Data Hurt

CC BY NC SA

Predicting the Next Word is *Knowledge*

2585524

The capital of California is ___Sacramento___

San Francisco (1862)
Benicia (1853)
Vallejo (1852)

Predicting the next word allows you to:

- Answer questions
- Tell stories
- Accomplish tasks

Generative AI

CC BY NC SA

2585524

Tell a short story about a fairy princess named Alice.

Once

BY NC SA

2585524

Generative AI Thesis

If you can generate the world you must understand the world.

Starting
Small:

How do we predict the next ~~word~~ token?

CC BY NC SA

Modeling Tokens not Words

2585524

Tokens represent words, word parts, and special characters.

"The smallest tokenizer!" →

Tokens: ["The", "small", "est", " token", "izer", "!"]

- Constructed based on frequency of words on the web
- Allows for new words, misspelings, ☺️, and numbers
- Vocabulary Size of GPT: ~100K unique tokens

CC BY NC SA

Generative Language Modeling

2585524

The best class at UC Berkeley is

Data-100!
Data 8.
any class taught by…
…

- More than one possible next token.
- Model the probability of all next tokens

CC BY NC SA

What Is a Large Language Model?

2585524

- **Large**: The model parameters (θ) are *BIG!*
  - *BILLIONS of PARAMETERS!!!!*
  - *Parameters: numbers encoding the neurons in the model*
- **Language Model:** Predicting language (e.g., words)

The capital of California is ___Sacramento___

San Francisco (1862)
Benicia (1853)
Vallejo (1852)

CC BY NC SA

2585524

The best class at UC Berkeley is _____

Context (ordered tokens)

<u>Model:</u> Pr(“Data-100” | “The best class at UC Berkeley is ”)

Next Token

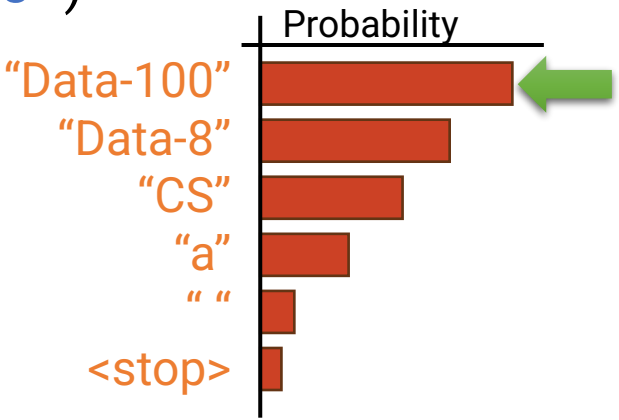
Context (ordered tokens)

NeuralNetwork(“The best class at UC Berkeley is ”)

How do we pick the **next token**?

Randomly according to the distribution.

Output:



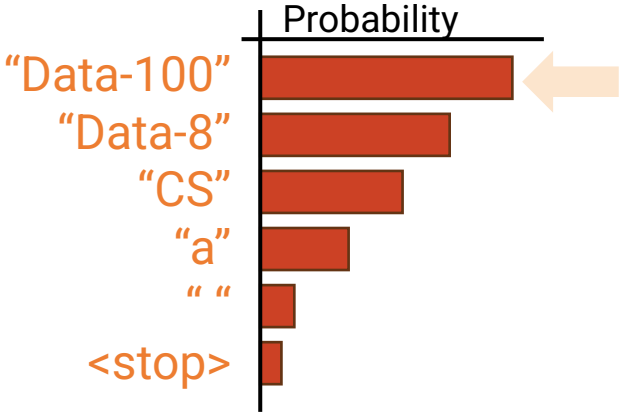
CC BY NC SA

Recap: Generative Language Modeling

2585524

- Model **the probability of the next token** (word) in a sentence
  - Using a **neural network** (magic…)
- We **randomly pick** the next word using the distribution.

NN(“The best class at UC Berkeley is”)



How do we go from predicting words to writing essays?

CC BY NC SA

2585524

Predict one word (token) at a time!

Generative AI writes in stream of “consciousness”

CC BY NC SA

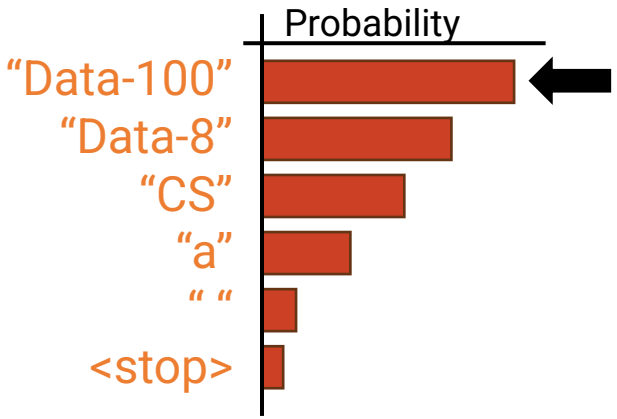
Auto-regressive Decoding

2585524

Sample **one token at a time** and add to context

Model: NN(“The best class at UC Berkeley is”)

Context



Decode one word:

The best class at UC Berkeley is Data-100

CC BY NC SA

Auto-regressive Decoding

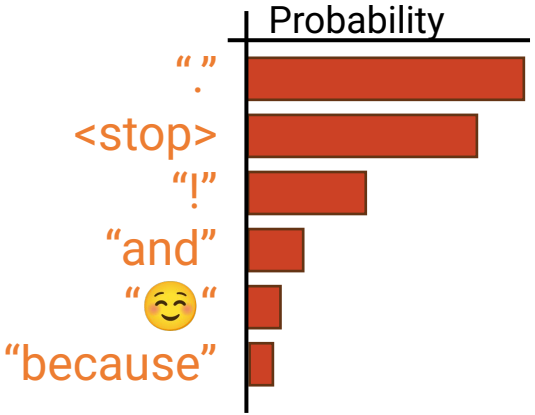
2585524

Sample **one token at a time** and **add to context.**

<u>Model:</u> NN(“The best class at UC Berkeley is Data-100”)

Context

Output:



CC BY NC SA

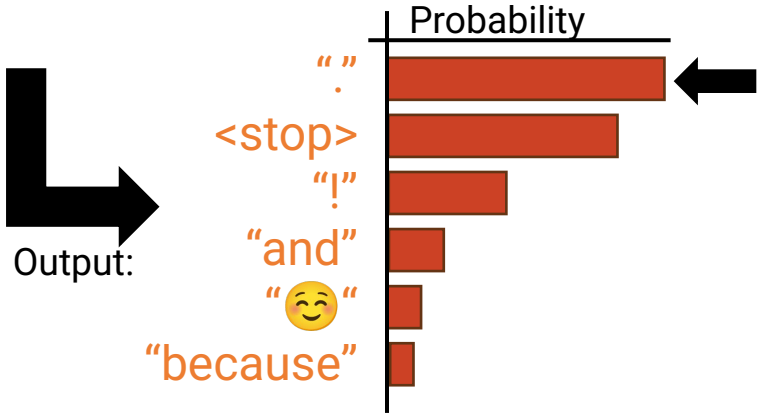
Auto-regressive Decoding

2585524

Sample **one token at a time** and **add to context.**

<u>Model:</u> NN(“The best class at UC Berkeley is Data-100”)

Context



<u>Decode one word:</u>

The best class at UC Berkeley is Data-100.

CC BY NC SA

Auto-regressive Decoding

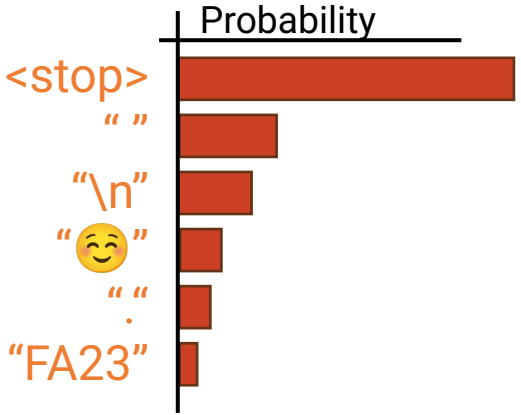
2585524

Sample **one token at a time** and **add to context.**

<u>Model:</u> NN(“The best class at UC Berkeley is Data-100.”)

Context

Output:



CC BY NC SA

Auto-regressive Decoding

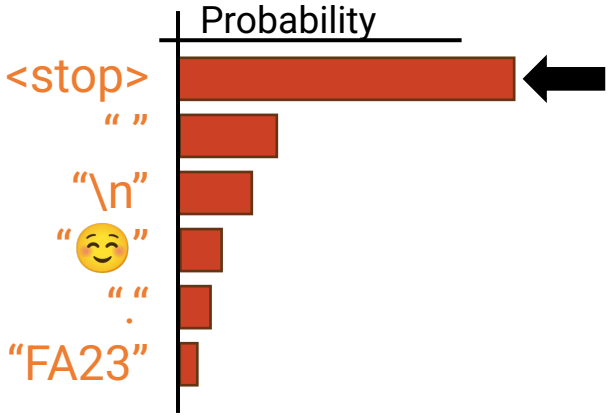
2585524

Sample **one token at a time** and **add to context.**

Model: NN("The best class at UC Berkeley is Data-100.")

Context

Output:



Decode one word:

The best class at UC Berkeley is Data-100.<stop>

Stop decoding!

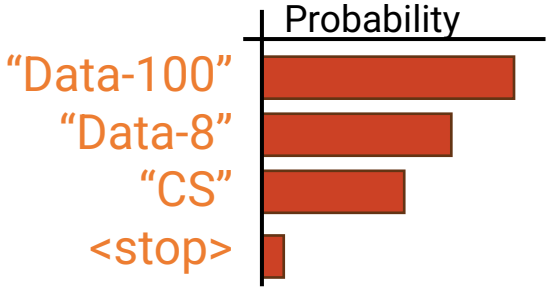
CC BY NC SA

Quick Recap:

2585524

- **Generative Language Models** encode the distribution over the next word (token) given the text before it (context).

NN("The best class at UC Berkeley is")



- By **repeatedly sampling** the next word **we can write essays**!

The best class at UC Berkeley is

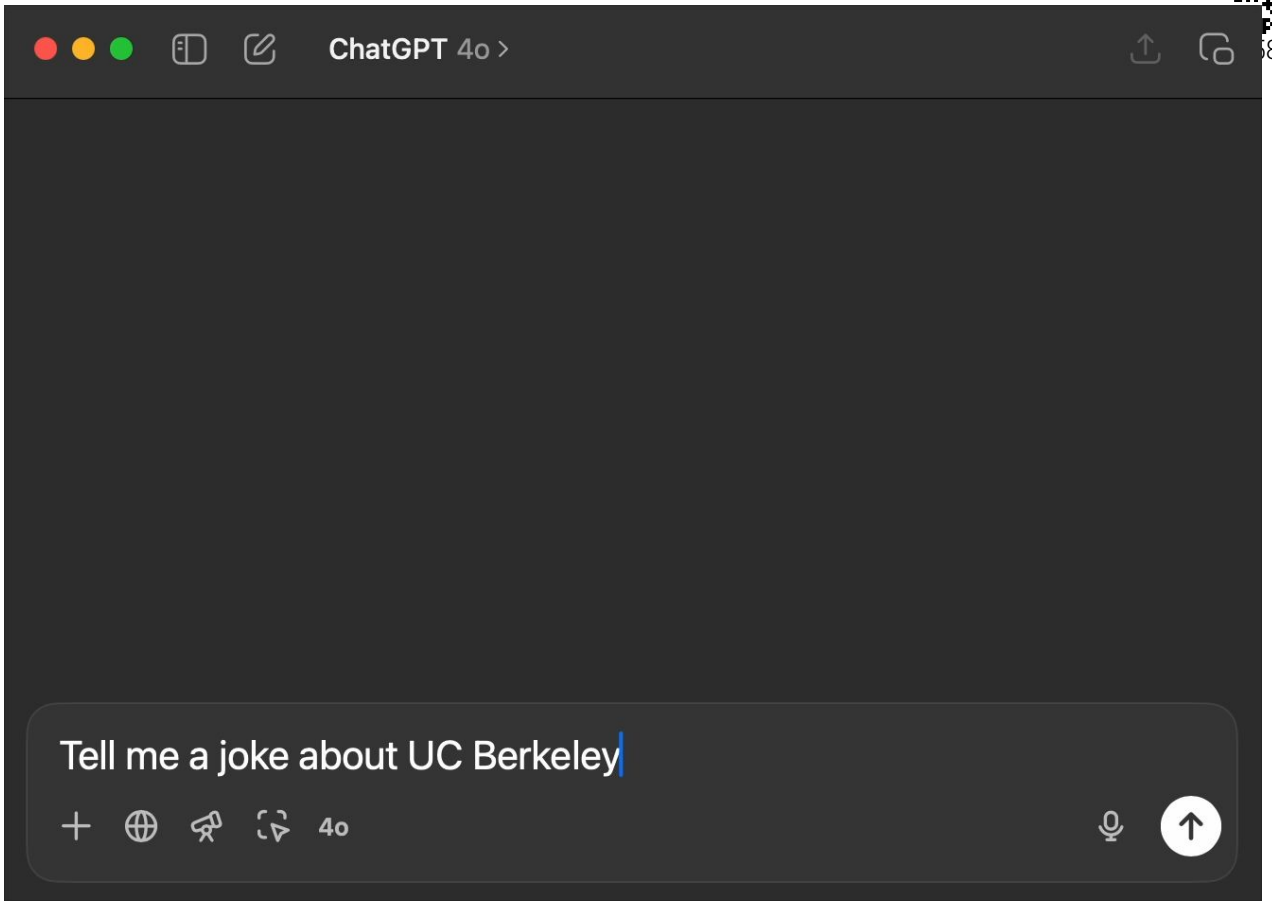
The best class at UC Berkeley is Data-100

The best class at UC Berkeley is Data-100.

The best class at UC Berkeley is Data-100!<stop>

CC BY NC SA

One Token at a Time



85524

CC BY NC SA

2585524

Language Modeling

is basically auto-complete where we always accept the next word.

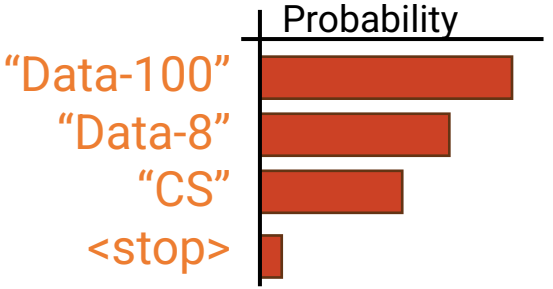
Predictive text; tap a suggestion to apply.

CC BY

Quick Recap:

- **Generative Language Models** encode the distribution over the next word (token) given the text before it (context).

NN(“The best class at UC Berkeley is”)



- By repe[illegible]pling the next word **we can write essays**!

How does the Neural Network Work?

The best class at UC Berkeley is Data-100.

The best class at UC Berkeley is Data-8!<stop>

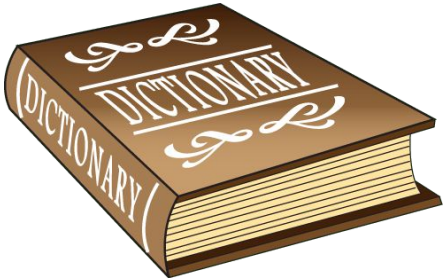
2585524

How Does the Neural Network Work?

Convert the input (context) into **lists of numbers** (vectors)

2585524

NN("The best class at UC Berkeley is")



1.2	2.0	-0.8	-3.4	2.5	-0.7	-5.2
3.7	-1.4	3.1	4.1	-5.1	-9.9	-0.7
-2.1	-0.9	0.6	-0.1	-4.2	3.1	-1.5
...	...	...	...	...	...	...
-0.6	7.1	-2.3	9.4	-2.0	4.7	3.0

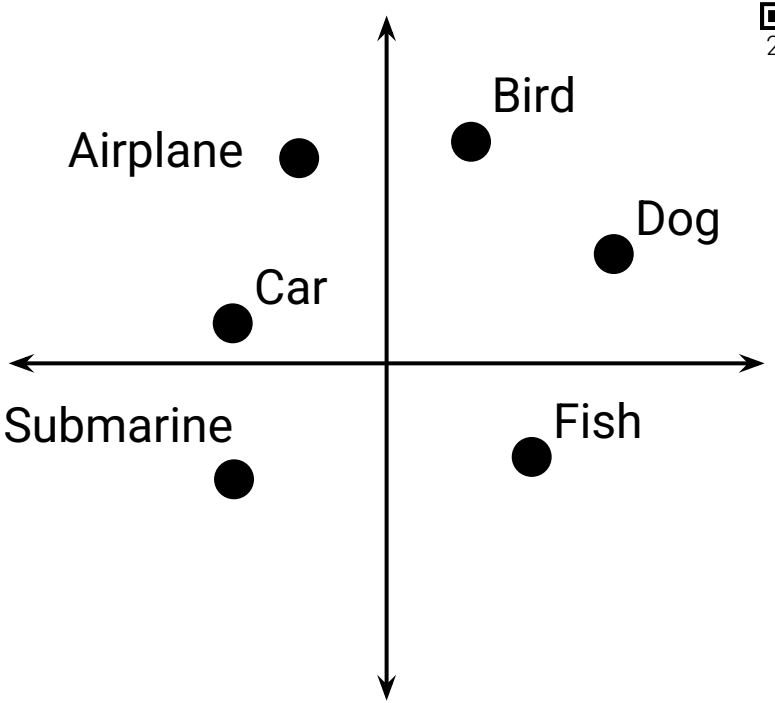
Initially random numbers → but we will learn to encode meaning.

CC BY NC SA

Words in a Vector Space

2585524

Simplified 2d example.
(In practice we use
thousands of dimensions.)



Represent **words as vectors:**

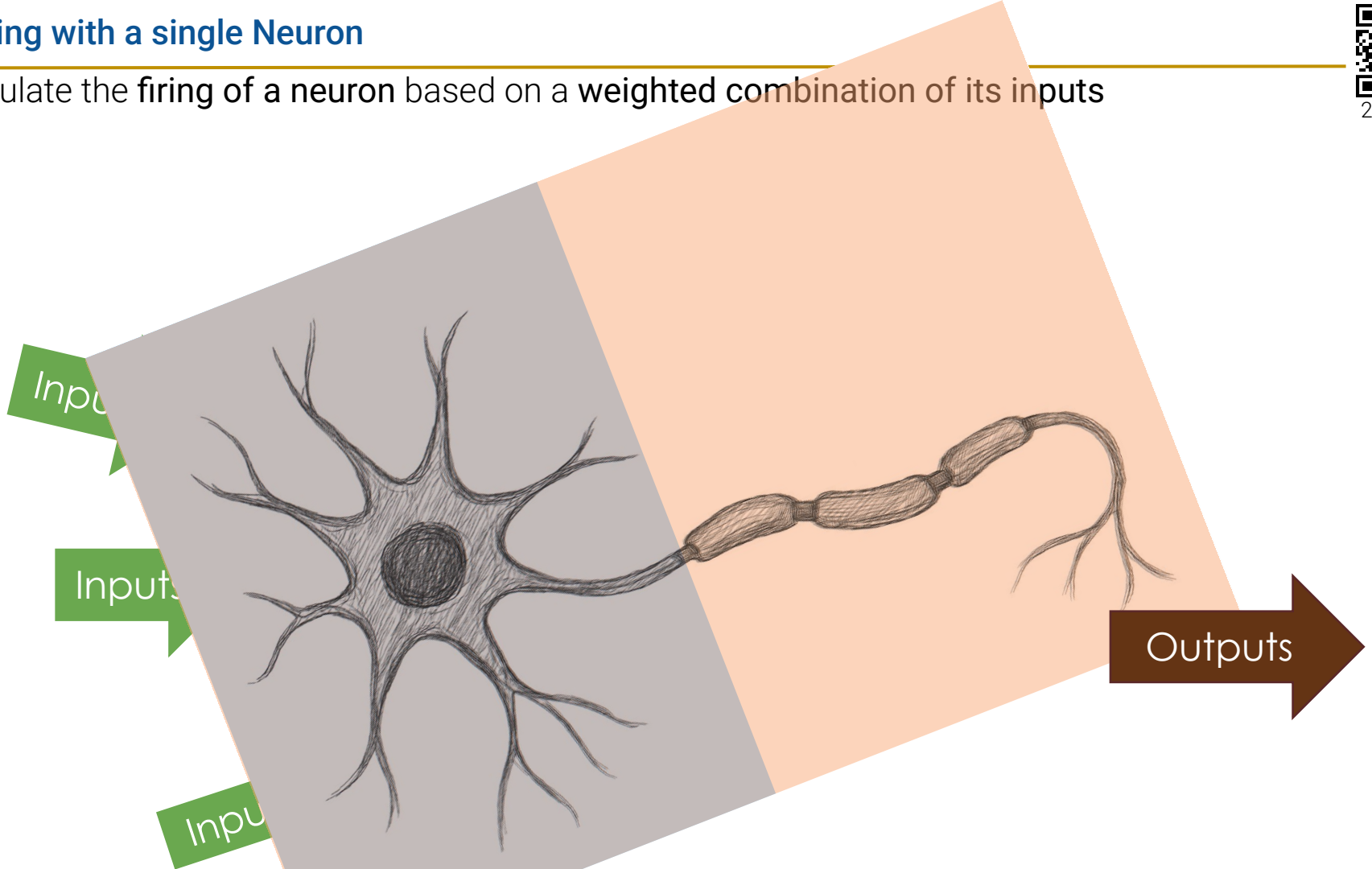
- Each word in the **dictionary has a vector**
- Nearby words are **semantically related**
- Idea dates back to Latent Semantic Analysis (1988)
  - Earlier in 1950s Linguistics

CC BY NC SA

Starting with a single Neuron

2585524

Simulate the **firing of a neuron** based on a **weighted combination of its inputs**



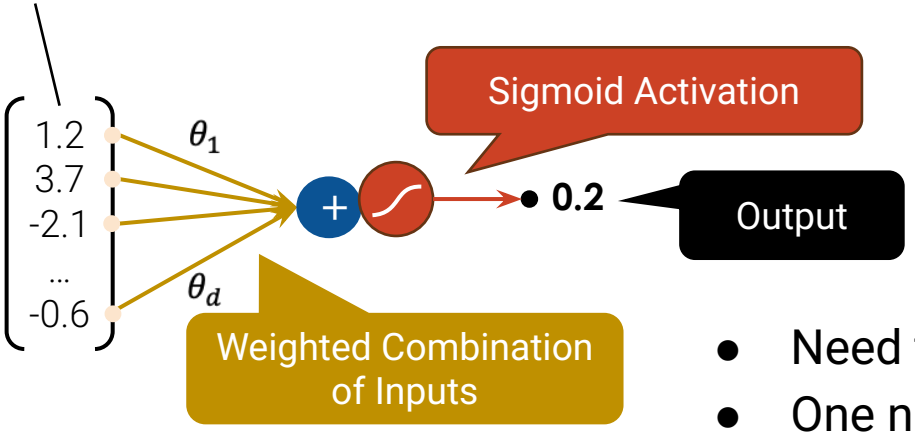
BY NC SA

Starting with a single Neuron

2585524

Simulate the **firing of a neuron** based on a **weighted combination of its inputs**

NN("The best class at UC Berkeley is"**)**



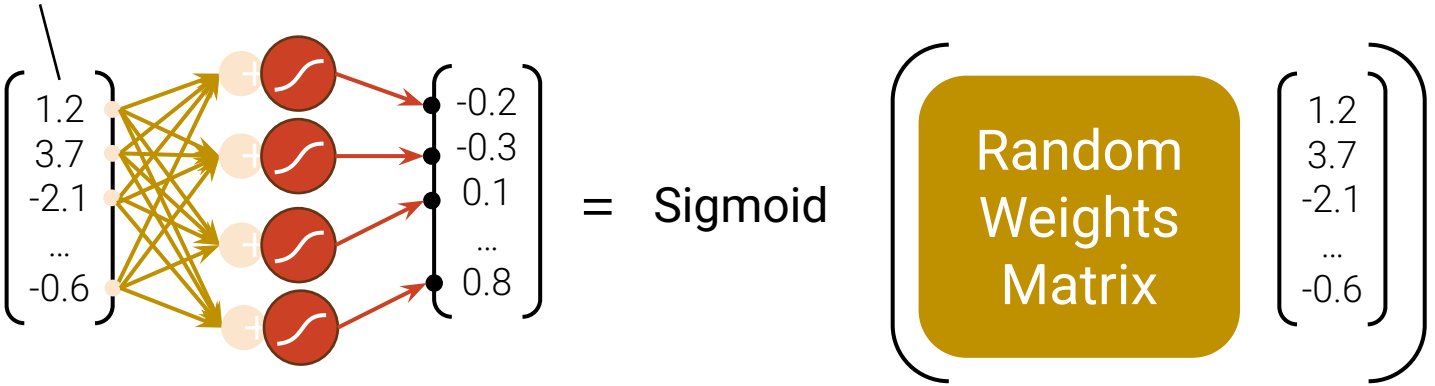
- Need to learn these weights
- One neuron is not enough
  - Connect many different neurons

CC BY NC SA

Adding many Neurons to A Word

2585524

NN("The best class at UC Berkeley is"**)**



Need to repeat this process for every word!

CC BY NC SA

NN("The best class at UC Berkeley is")

2585524

Sigmoid

Random Weights Matrix

1.2	2.0	-0.8	-3.4	2.5	-0.7	-5.2
3.7	-1.4	3.1	4.1	-5.1	-9.9	-0.7
-2.1	-0.9	0.6	-0.1	-4.2	3.1	-1.5
...	...	...	...	...	...	...
-0.6	7.1	-2.3	9.4	-2.0	4.7	3.0

=

Artificial Neural Network Outputs

0.2	0.1	0.2	-0.4	-0.2	0.3	0.3
-0.3	-0.7	0.4	-0.1	-0.5	-0.9	0.4
0.1	0.3	0.5	0.6	0.2	0.7	0.5
...	...	...	...	...	...	...
0.8	-0.7	-0.3	0.9	0.4	-0.1	-0.4

How do we go **from all these vector outputs** to **predicting the next word?**

CC BY NC SA

What is Chat GPT?

2585524

Chat: natural language system

✅ G: Generatively – Designed to model the creation of text

P: Pretrained – Trained on lots of naturally occurring data

✅ T: Transformer – A kind of neural network architecture

Chat GPT is just one example of a

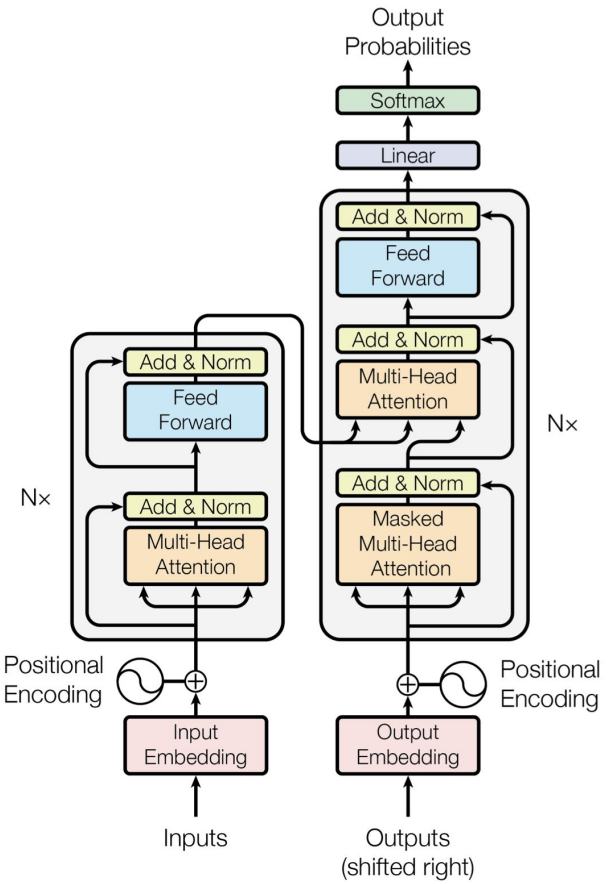
✅ Large Language Model (LLM)

CC BY NC SA

2585524

The Transformer Model

Explained using intuition only…



*famously difficult to explain.

CC BY NC SA

2585524

What words in the context most predict the next word? (Where should we attend?)

Tell a short story about a fairy princess named Alice.

Once upon a time there was a fairy princess named ____

CC BY NC SA

2585524

What words in the context most predict the next word? (Where should we attend?)

Tell a short story about a fairy princess named Alice.

Once upon a time there was a fairy princess named _____

Big Idea:
Use the context to focus on the important words.

CC BY NC SA

Transformer Combines Neuron Outputs

Use **learned attention** (more neurons) to **combine the outputs** into a single vector.

2585524

NN("The best class at UC Berkeley is")

Transformer

$$\left(\begin{bmatrix}0.2\\-0.3\\0.1\\\dots\\0.8\end{bmatrix}\begin{bmatrix}0.1\\-0.7\\0.3\\\dots\\-0.7\end{bmatrix}\begin{bmatrix}0.2\\0.4\\0.5\\\dots\\-0.3\end{bmatrix}\begin{bmatrix}-0.4\\-0.1\\0.6\\\dots\\0.9\end{bmatrix}\begin{bmatrix}-0.2\\-0.5\\0.2\\\dots\\0.4\end{bmatrix}\begin{bmatrix}0.3\\-0.9\\0.7\\\dots\\-0.1\end{bmatrix}\begin{bmatrix}0.3\\0.4\\0.5\\\dots\\-0.4\end{bmatrix}\right) = \begin{bmatrix}0.1\\-0.4\\0.3\\\dots\\0.7\end{bmatrix}$$

More Neuron Weights to Learn

Predicted Word Vector

CC BY NC SA

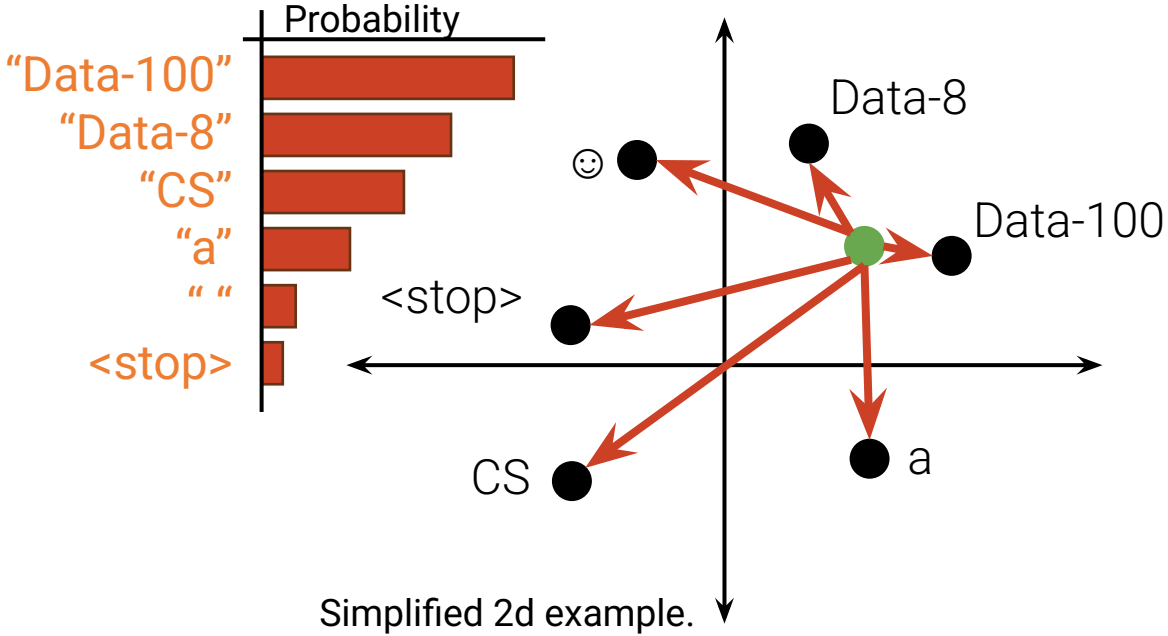
Transformer Combines Neuron Outputs

Nearest word vector **in the dictionary** is the **most likely word**

2585524

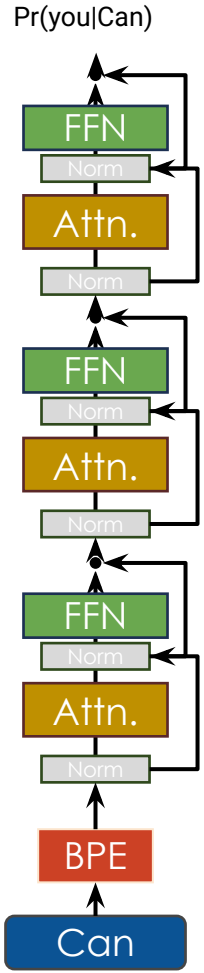
$$= \begin{pmatrix} 0.1 \\ -0.4 \\ 0.3 \\ \dots \\ 0.7 \end{pmatrix}$$

Predicted Word Vector



CC BY NC SA

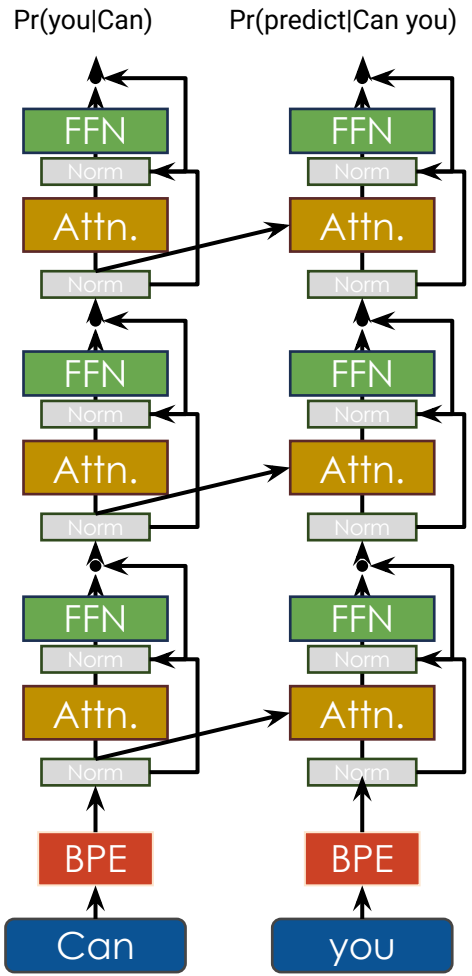
Unrolling the Model



2585524

CC BY NC SA

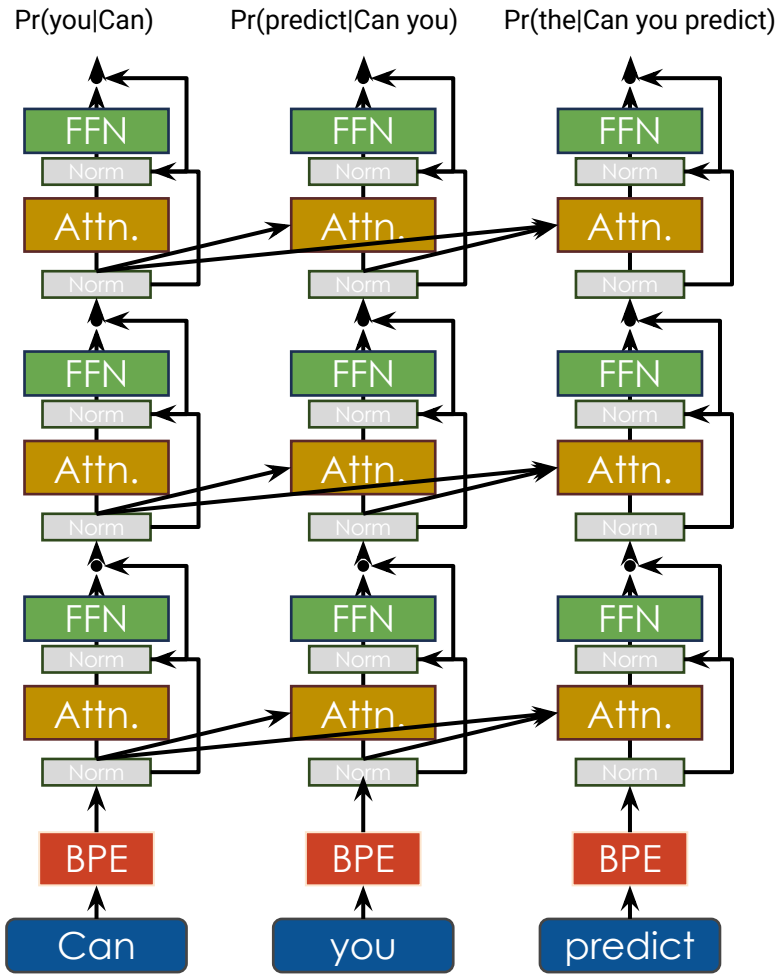
Unrolling the Model



2585524

CC BY NC SA

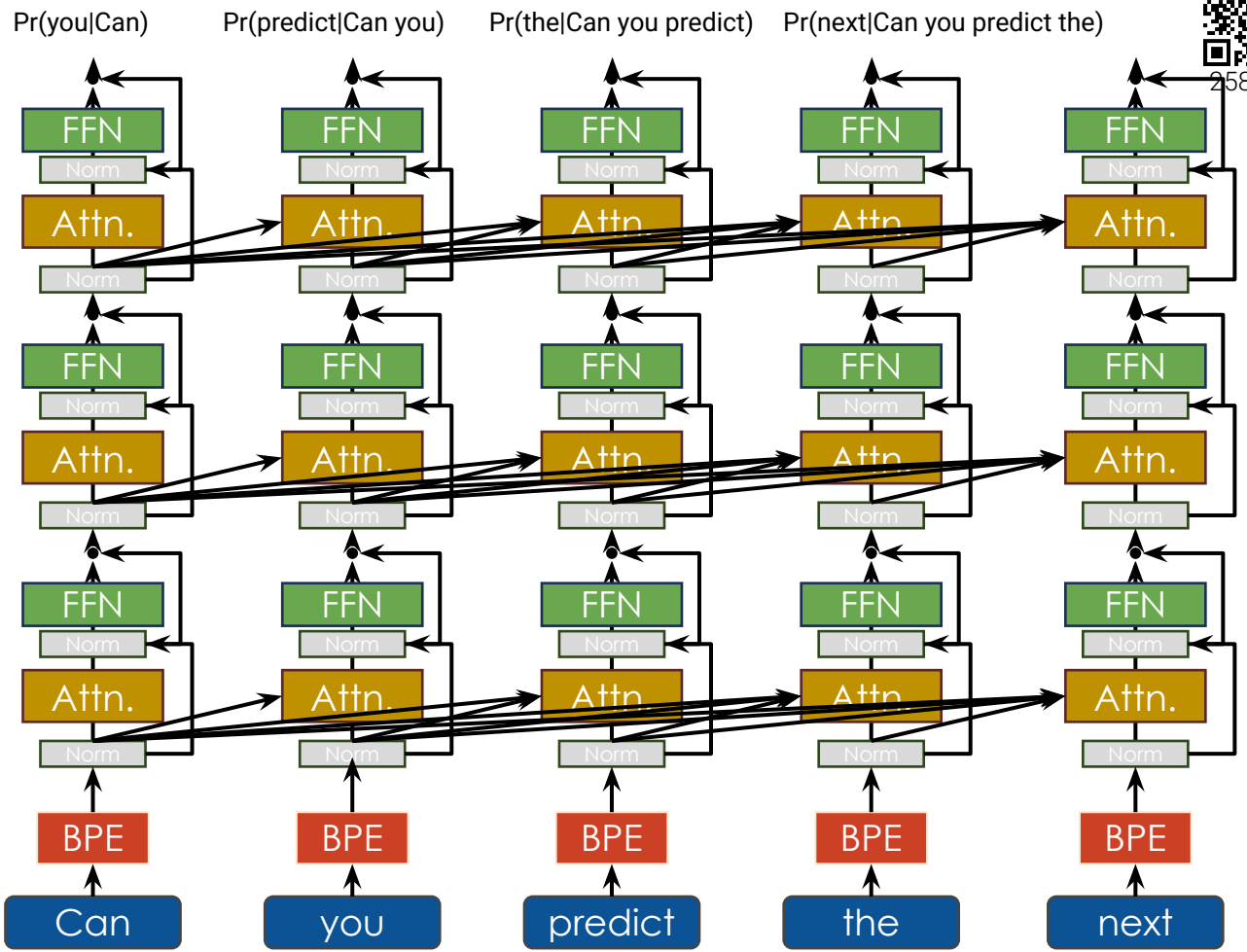
Unrolling the Model



2585524

CC BY NC SA

Unrolling the Model



2585524

CC BY NC SA

What is Chat GPT?

2585524

Chat: natural language system

✅ G: Generatively – Designed to model the creation of text

P: Pretrained – Trained on lots of naturally occurring data

✅ T: Transformer – A kind of neural network architecture

Chat GPT is just one example of a

✅ Large Language Model (LLM)

CC BY NC SA

How to train your neural network

2585524

- Collect a very **large dataset**
- Initialize all weights to **random numbers**
  - Initial model **makes terrible predictions**
- **Define an objective** that we want to improve
  - **Maximize the probability** of the **next word** over the dataset
- **Iteratively adjust the weights** to **improve the objective**
  - Small **tweaks to hundreds of billions of weights** using **thousands of expensive GPUs over many weeks**

CC BY NC SA

Pre-training on *Everything**

2585524

- Train the model on a **large collection** of data to learn **generalizable patterns**

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Reddit Links

OpenAI GPT3 Data Mix

- Llama-3 "open-source" models trained on **15.6T tokens from an unknown data mix.**
  - 405-billion parameter model trained on **16K H100s ($25K each)**
  - **39.3M GPU hours**

*Everything that is legal to use for training hopefully…

CC BY NC SA

Generative Pre-training

2585524

Single Passage of Text

We the People of the United States, in Order to form a more perfect Union, establish Justice, …

Tune the model parameters to **maximize the likelihood** of the next token

Each token
is a training example

Pr("We" | "")

Pr("the" | "We")

Pr("People" | "We the")

Pr("of" | "We the People")

Pr("the" | "We the People of")

…

CC BY NC SA

2585524

Now you know

What is Chat GPT

There is still one more thing

Chat: natural language system

- ✅: Generatively – Designed to model the creation of text
- ✅: Pretrained – Trained on lots of naturally occurring data
- ✅: Transformer – A kind of neural network architecture

GPT alone can’t chat!

CC BY NC SA

Need to Teach GPT to Follow Instructions (and be fun!)

2585524

- Generative pre-training **captures knowledge**

- The pre-trained model finishes statement but **doesn't follow instructions.**

  You ask: "What is attorney client privilege?"

  GPT Replies: "Why is it important? Provide a concise example for full credit on this exam."

- We can improve model behavior and capabilities with **in-context learning** and **post-training**

CC BY NC SA

Fine-tuning

2585524

Running **additional training iterations** with a specific task

- The task differs from the original pre-training task
  - **New objective**: translate sentence, follow instruction
  - **New training data** from new source domain
- **Smaller learning rate** (as you get older you learn slower?)
  - Avoid “catastrophic forgetting” (new information causes forgetting pre-training information)

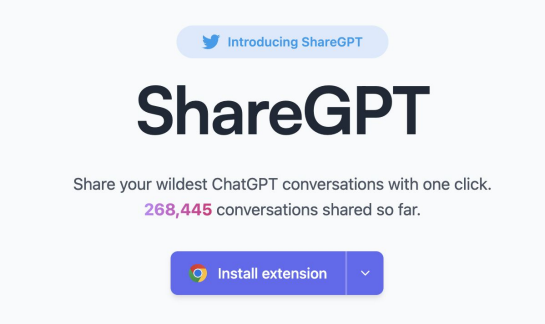
CC BY NC SA

Vicuña

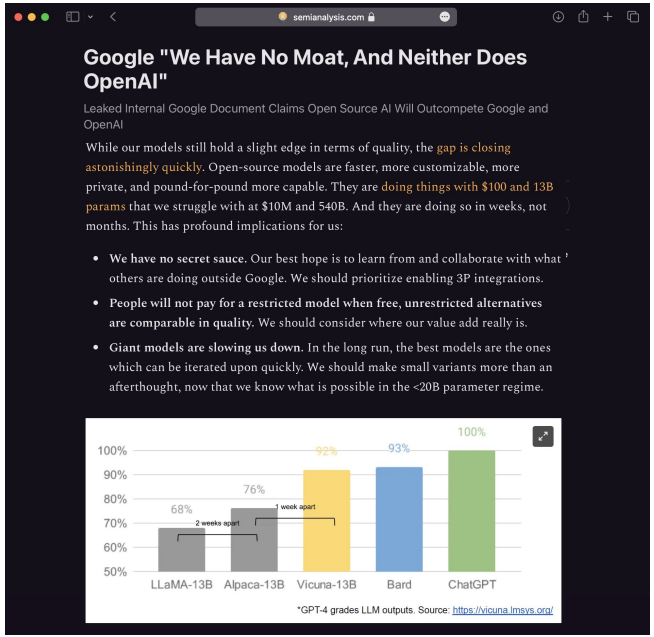
Open-source Instruction Fine-Tuned LLMs and LLM-as-a-judge

2585524

- First open-source model that was "comparable" to ChatGPT
  - Fine-tuned LlaMA-13B on ShareGPT Data



Tiny
High Quality Data
70K conversations (~800MB)



- Helped launch academic open-source GenAI research

CC BY NC SA

2585524

Now you know

What is Chat GPT

Chat: natural language system

- ✅: Generatively – Designed to model the creation of text
- ✅: Pretrained – Trained on lots of naturally occurring data
- ✅: Transformer – A kind of neural network architecture

How are people using it?

CC BY NC SA

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:        ← task descriptio
cheese =>                           ← prompt
```

Zero-shot relies on model already "knowing" how to complete the task.

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← example
cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           ← prompt
```

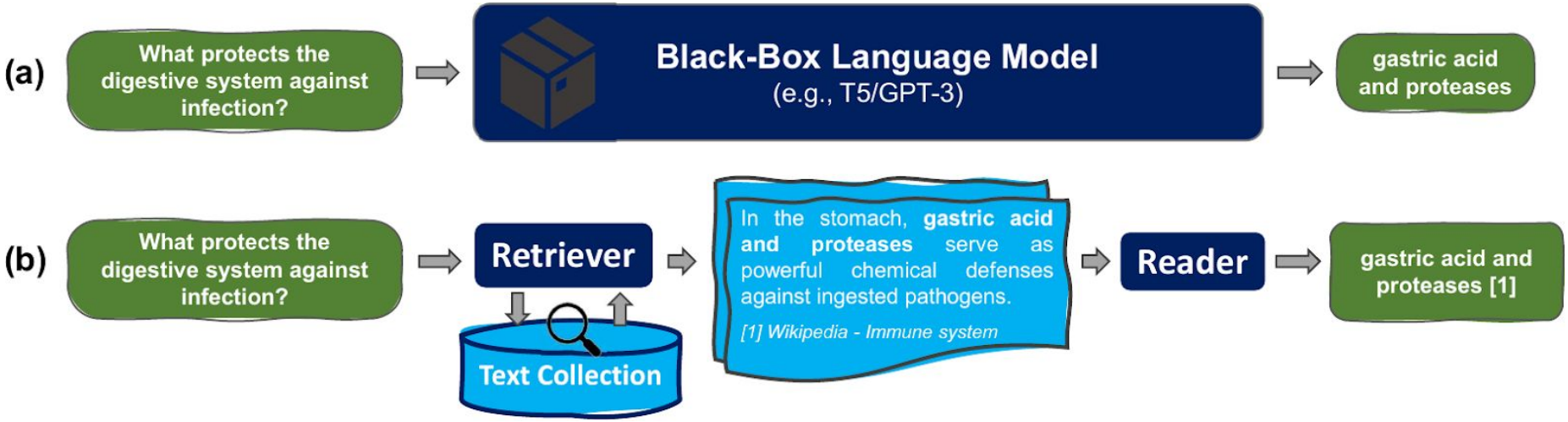
In-context Learning

CC BY NC SA

Retrieval Augmented Generation (RAG)

2585524

- Use external data to augment LLMs



https://acl2023-retrieval-lm.github.io

CC BY NC SA

2585524

Logistics

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models

Logistics

What did you learn in Data C100/C200?

What's Next?

Join course staff!

CC BY NC SA

Announcements

2585524

Final Exam Logistics

Friday, May 16th: 11:30am - 2:30pm

In-person only

Final Exam **Review Sessions**:

- Tuesday and Thursday of RRR same time as lecture in Wheeler 150
  - Tuesday: Post-Midterm Content Review
  - Thursday: Problem-Solving Practice

Course evaluations extra credit

- +0.5 point of EC for completing the internal final survey AND official course evaluations
- +0.5 points of EC if 90% of the class fills out both.

CC BY NC SA

2585524

What did you learn in Data C100/C200?

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models

Logistics

What did you learn in Data C100/C200?

What's Next?

Join course staff!

CC BY NC SA

What Were We Supposed to Teach You?

2585524

Prepare

Prepare students for advanced Berkeley courses in **data management**, **machine learning**, and **statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**.

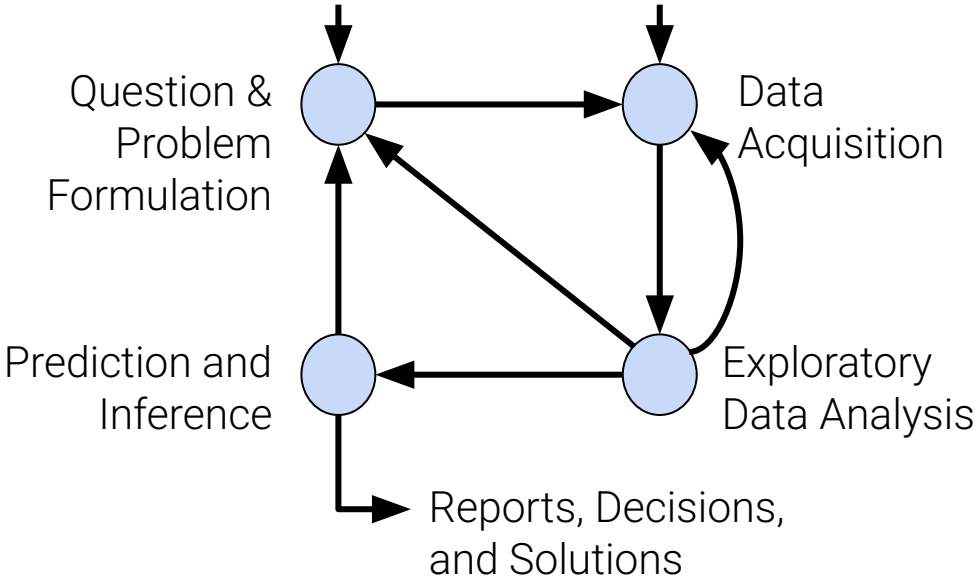
Empower

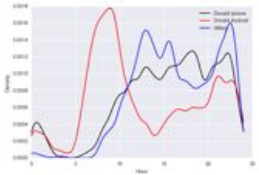
Empower students to apply computational and inferential thinking to address **real-world problems**.

CC BY NC SA

The Data Science Lifecycle

2585524

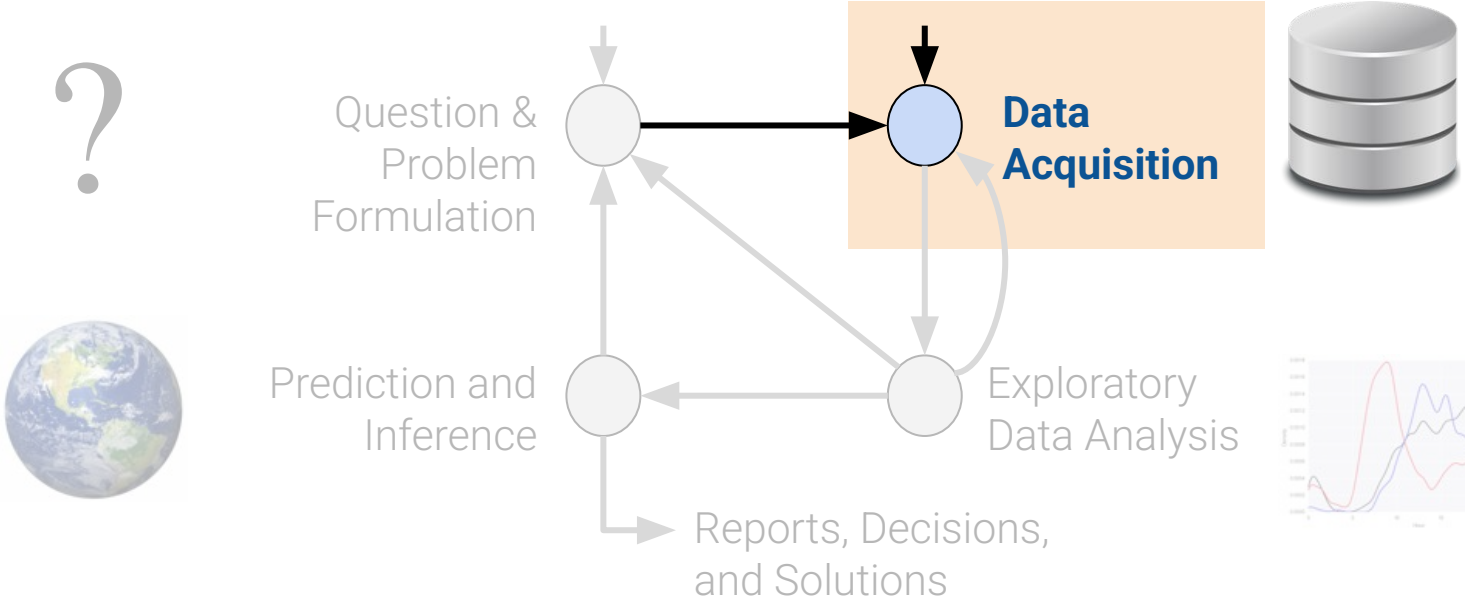




CC BY NC SA

The Data Science Lifecycle

2585524



jupyter

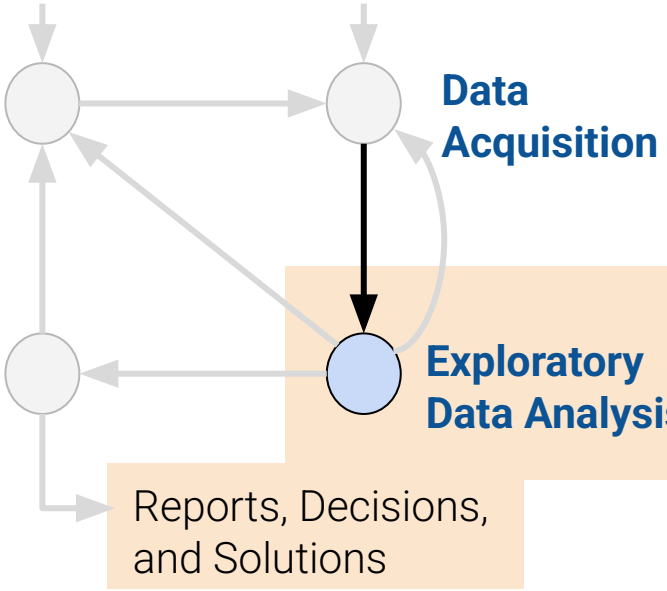
pandas

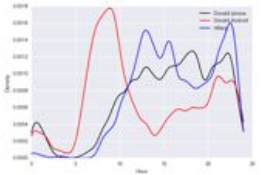
SQLite

CC BY NC SA

The Data Science Lifecycle

2585524

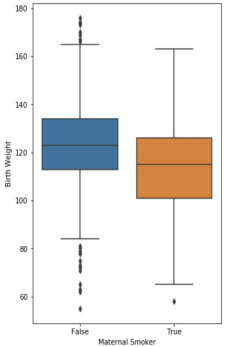


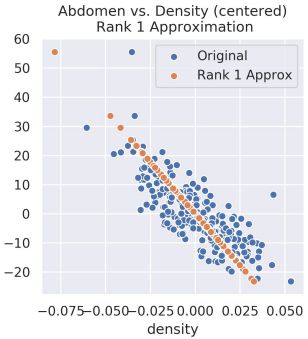


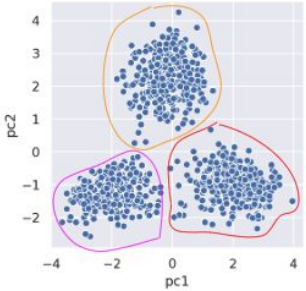
Seaborn

plotly

matplotlib



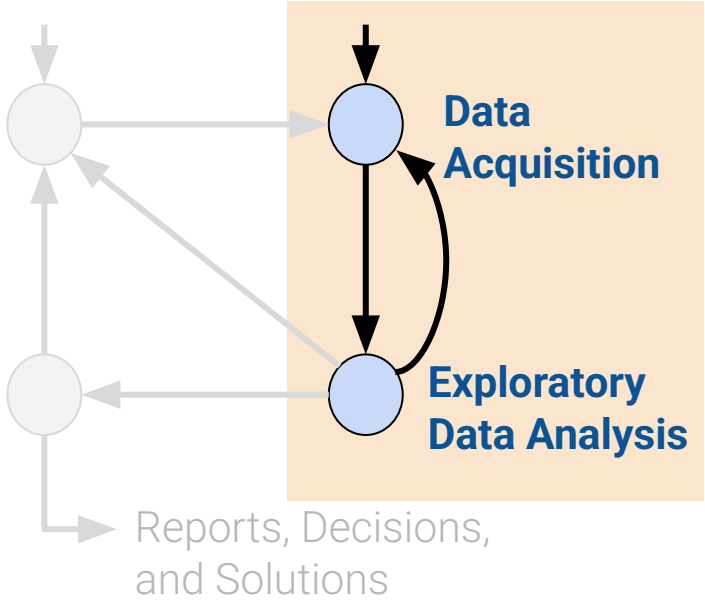


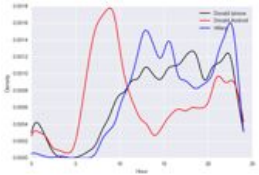


CC BY NC SA

The Data Science Lifecycle

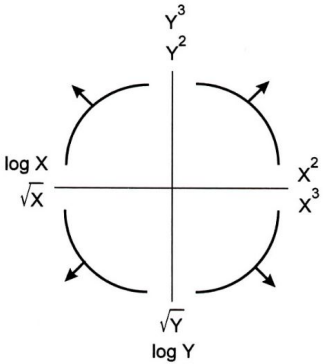
2585524





```
(?<=\.) {2,}(?=[A-Z])
```

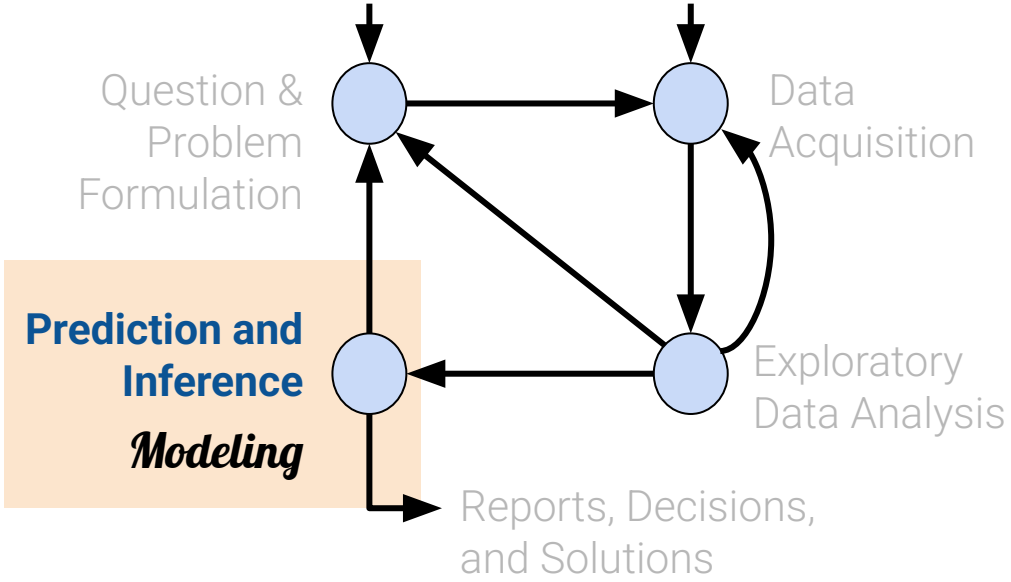
Feature Engineering

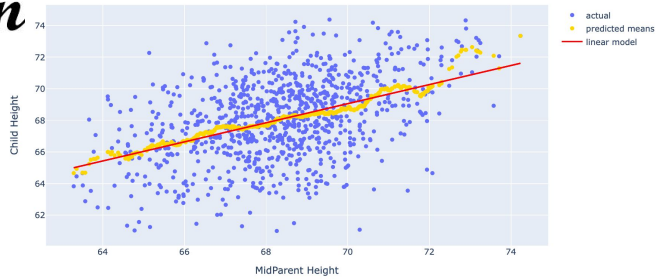


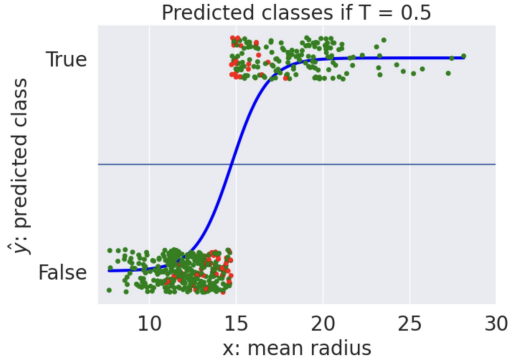
CC BY NC SA

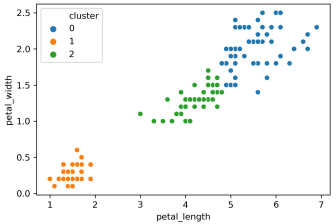
The Data Science Lifecycle

2585524





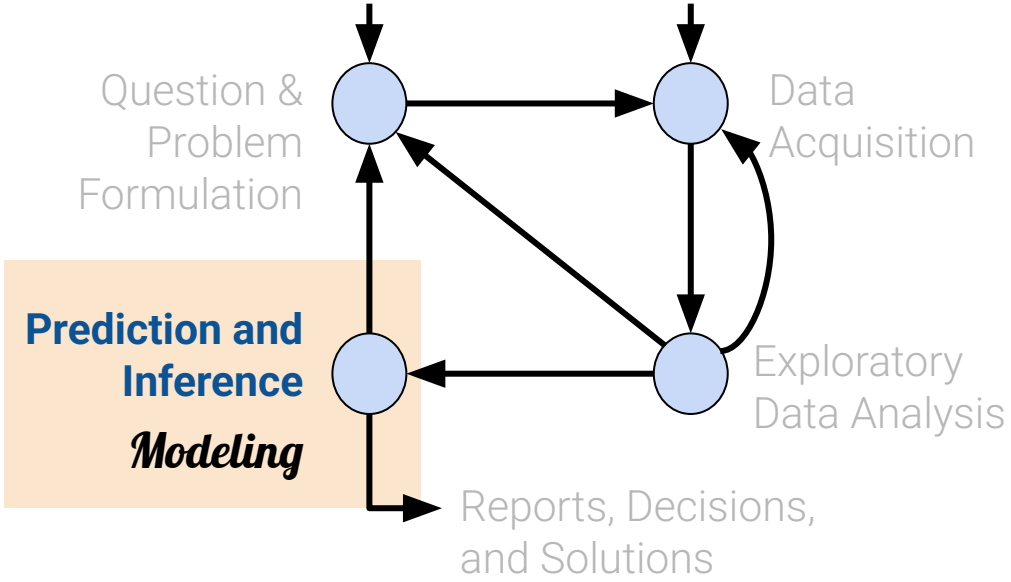




CC BY NC SA

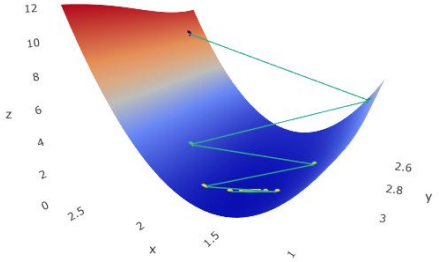
The Data Science Lifecycle

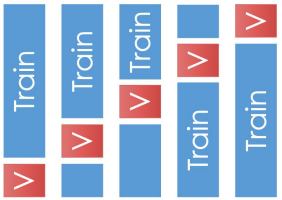
2585524



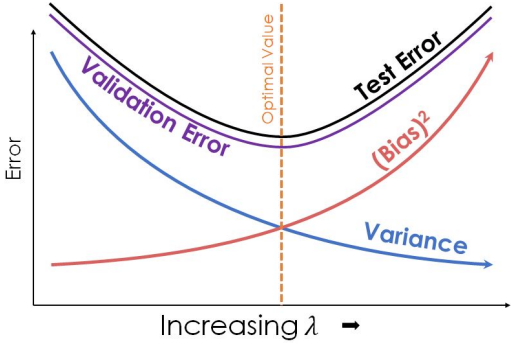
SciPy

NumPy





5-Fold Cross Validation



CC BY NC SA

What Didn't We Focus on in Data 100?

2585524

- Causal inference.
  - How do we establish causality when we identify a correlation observed during EDA?
- Deep learning.
  - How can the machine do the hard work of picking the right features instead of requiring humans to pick them in advance?
- Decision making.
  - After we build our regression/classification/clustering algorithms, what do we do next?
  - How do we avoid false discoveries?
- Time series analysis.
- Other flavors of machine learning (e.g. reinforcement learning).
- Open-ended exploration of problems and datasets picked by you.
- Non-tabular data (e.g. images/video, sensor-generated/spatial data, natural language...).

CC BY NC SA

2585524

What’s Next?

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models

Logistics

What did you learn in Data C100/C200?

What’s Next?

Join course staff!

CC BY NC SA

More Course Advice and Recommendations

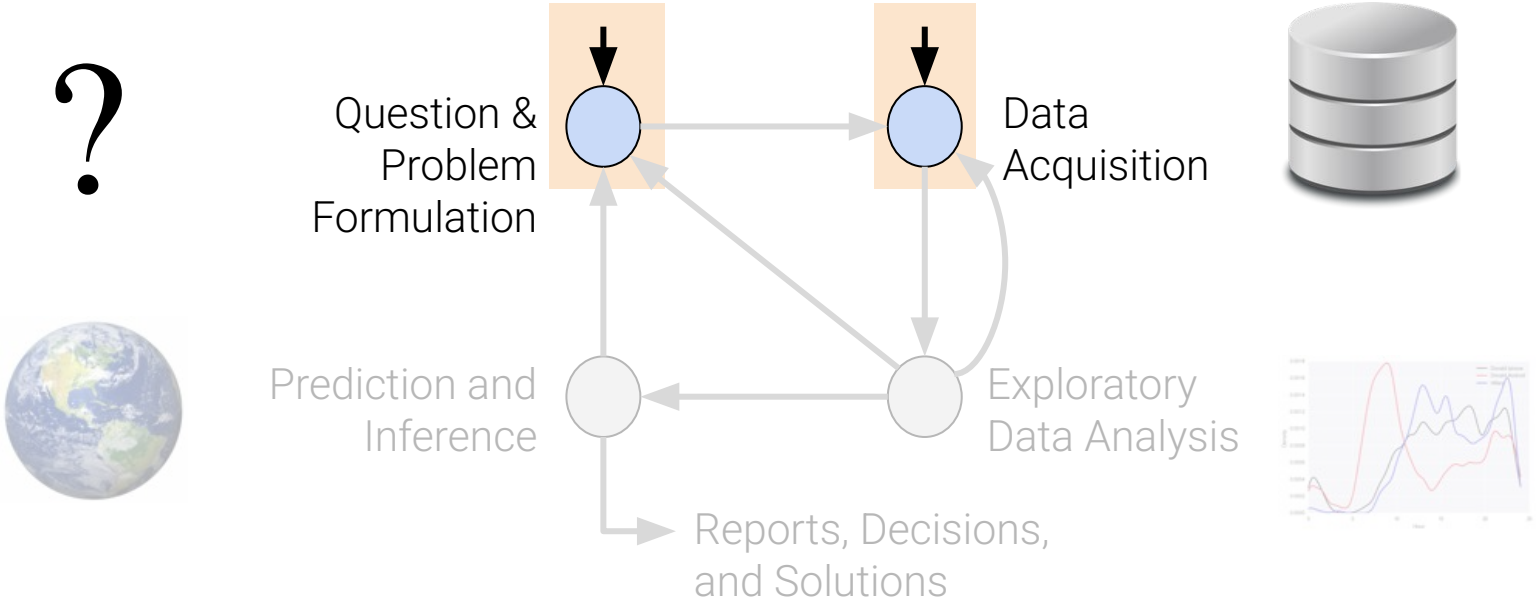
2585524

Join the Data Science Community (Data 001) on Ed:
https://edstem.org/us/join/gk5MZQ**!**

CC BY NC SA

Real-world Applications

2585524



CC BY NC SA

Real-world Applications

2585524

A great way to strengthen your knowledge of the ideas you learned in this class (and to build a portfolio to help you find jobs!) is to use your new skills to analyze real-world data.

- There are countless sources of data available on the internet.
- Find one, load it into a notebook, and get to work!

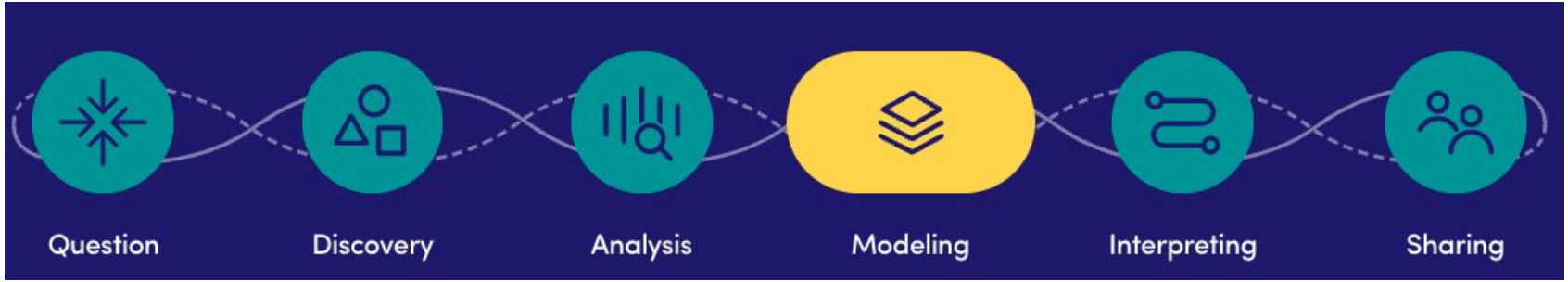
Places to look for data and applications of data science:

- Awesome Public Datasets GitHub repo.
- Kaggle.
- Google's dataset search.
- Towards Data Science.
- reddit.com/r/dataisbeautiful.
- Five Thirty Eight.
- Pudding
- Tidy Tuesday
- and so on!

CC BY NC SA

Real-World Toolkit: The Data Science *Ethos* Lifecycle

2585524



Take **DATA C104: Human Contexts and Ethics of Data** for more contexts and strategies!

Through the Lenses:

Positionality	(of the researcher)
Sociotechnical Systems	(of the academic community)
Power	(transparent/reproducible workflows and what they entail)
Narratives	(of the expected results)

https://ethos.academicdatascience.org/

CC BY NC SA

Real-world Applications

2585524

Even though you're new to the discipline, you are also among the most skilled people in the world at data science. **Use your power wisely!**

Let us know what you do next!

Spider-Man (2002), or 1st century BC
[Wikipedia]

CC BY NC SA

Do not edit
How to change the design
524

What was your favorite Data 100 topic?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension
slido

2585524

Join course staff!

Lecture 27, Data 100 Spring 2025

Introduction to Large Language Models
Logistics
What did you learn in Data C100/C200?
What's Next?
Join course staff!

CC BY NC SA

Leads

2585524

Head TA

Sam Bobick

Shreya Gupta

Meenakshi Mittal

Xiaorui Liu

Rachel Xin

Dan Nguyen

Rose Niousha

Sarah Kiefer

Sammie Smith

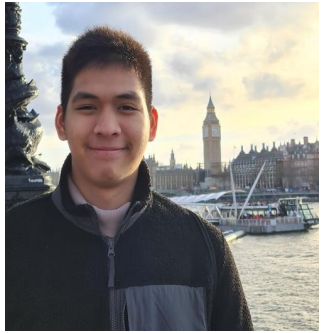
Malavikha Sudarshan

This class would not have been possible without our **Lead Course Staffs**, who led various teams to ensure everything runs smoothly

CC BY NC SA

UCS2s

2585524



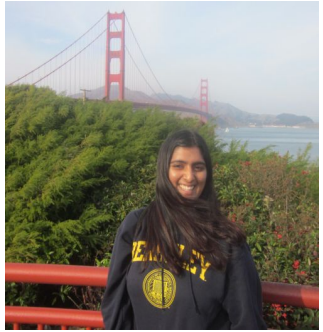
Emrie Loh

Dun-Ming Huang

Conan Smallwood

Azalea Bailey

James Geronimo



Sarika Pasumarthy

Jake Pastoria

This class would not have been possible without our **UCS2s**, who led weekly discussions and assisted the leads

CC BY NC SA

UCS1s

2585524

Cristina Prieto

Kelly Hu

Skyla Ma

Steven Luo

William Lee

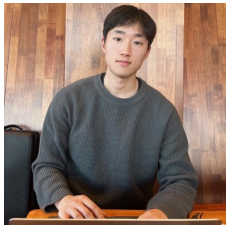
Willy Guan

Wolfgang Edholm

Kathryn Sun

Ella Hammond

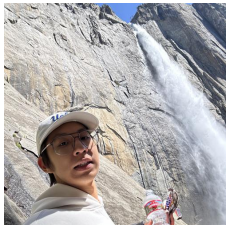
Anshul Jambula



Terry Kim

Sardaana Eginova

Noah Abji



Neil Chen

Dhruv Pendharkar

Vicky Huang

This class would not have been possible without our **UCS1s**, who reviewed and graded the assignments, answered lecture and Ed questions, and hosted office hours!

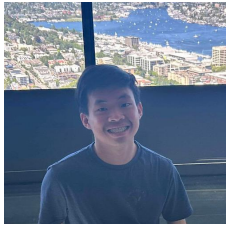
CC BY NC SA

UCS1s

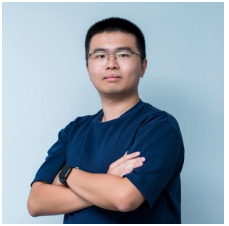
2585524

Collin Duong

Rohan Bijukumar

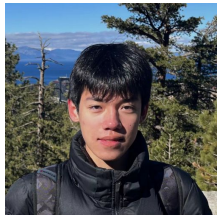


Alex Huang



Kevin Hu

Henry Hong



Haichao Li

Gisella Chan

Tyler Pham

Aiden Wang

Simon Kang

Surya Ramkumar

Satleen Gill

Alex Proshkin

This class would not have been possible without our **UCS1s**, who reviewed and graded the assignments, answered lecture and Ed questions, and hosted office hours!

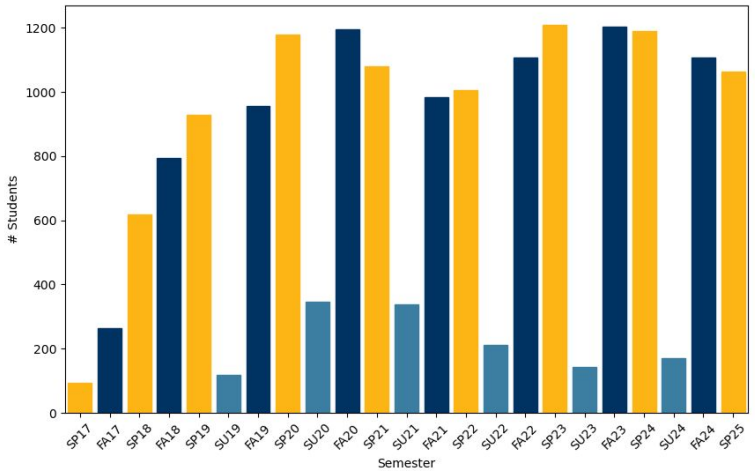
CC BY NC SA

Helping out with Data 100

2585524

Teaching is a great way to deeply understand course material.

- Data 100 has grown rapidly.
- This growth is only possible due to the help of students.
- The path to course staff starts with **you**.



We need you!

https://data.berkeley.edu/joining-data-course-staff

Fall ASE applications for DATA courses are usually released by May; Spring ASE applications are released by late October.

CC BY NC SA

2585524

From all of us: you have been an absolute delight to work with this Spring. We're so proud of how far you've come in these past few months, and are so excited to see where this class takes you.

Thank you!

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

CC BY NC SA

slido

2585524

Thank you and goodbye :)

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

CC BY NC SA