

## Applications of Data Science, Session 2: 1-4 PM PT

While we wait to get started:

1. Log into [Piazza](#)
2. If you thought of any questions during the week, you can post them in the [8/4 Workshop Questions](#) thread on Piazza

# Session 2:

# Applications of Data Science

Instructor: Julia Olivieri

TA: Julie Wang

# Welcome to Applications of Data Science

- TA: [Julie Wang](#)
- Don't message questions in Zoom: Use the [8/4 Workshop Questions](#) thread in Piazza
- There will be [two five-minute breaks](#), one after each of the first two modules
- Meeting time: today 1-4 PM
- The recording of last session is available on Piazza
- [You will only get the certificate if you fill out the feedback form](#)

For questions during class: [Post in Piazza](#)

Cameras **on** or **off**?

- Feel free to keep your cameras off when not in breakout rooms
- I encourage you to add a **profile picture** to Zoom if you haven't already

## Active learning through in-class Exercises

- Exercises provide opportunities to [actively engage with the material](#)
- We will devote time to working on these exercises [throughout the workshop](#)
- You will be requested to share your work [for each exercise](#) on Piazza
- I recommend you [open a document](#) to work on the exercises now
- You are encouraged to [post with your name](#), though anonymous posting is still allowed
- You are encouraged to [interact with each others' answers](#): upvote answers you agree with, comment if you have something else to say

# Schedule for today



Review of last class

Module 4: **Optimization: Navigating constraints and trade-offs**

*Questions*

*Break*

Module 5: **Generation: Creating novel content with algorithms**

*Questions*

*Break*

Module 6: **Where do we go from here?**

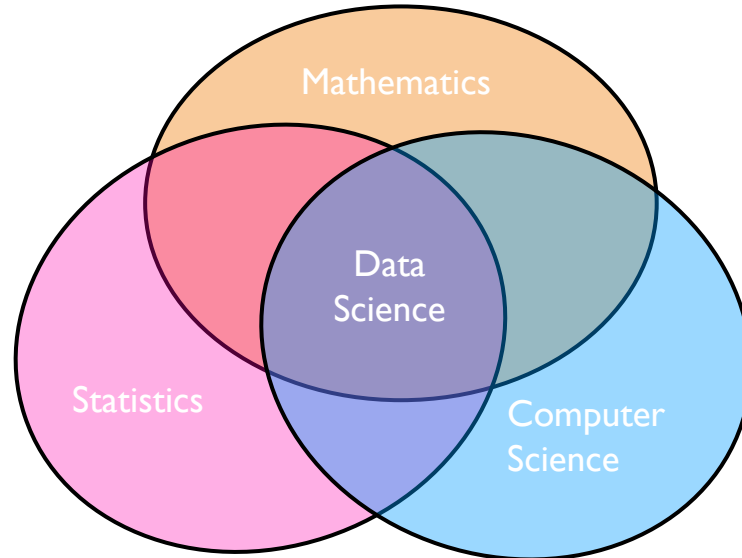
*Breakout rooms*

*Questions*

# Review

# How would you define “data science”?

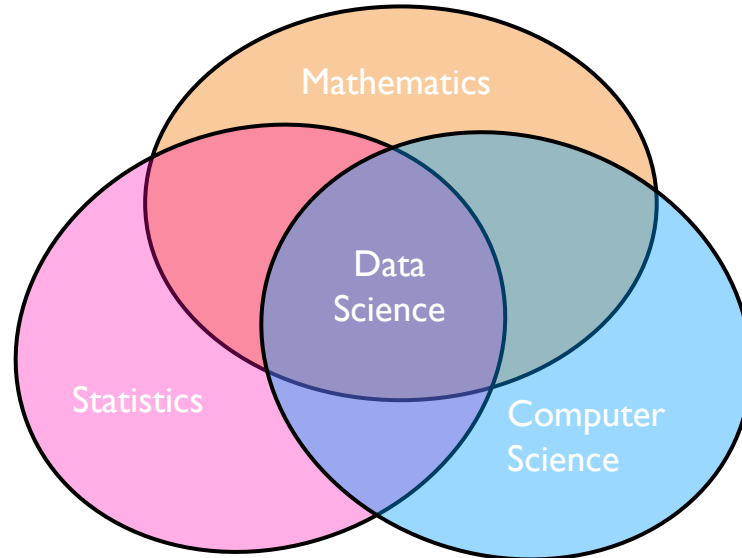
- Data science involves using **real-world data** to **answer questions** and **solve problems**
- Involves methods to **collect, analyze, and interpret** data
- Often involves making **predictions** or **decisions about the future**





# How would you define “data science”?

- Data science involves using **real-world data** to **answer questions** and **solve problems**
- Involves methods to **collect, analyze, and interpret** data
- Often involves making **predictions** or **decisions about the future**



## Features that make a problem suitable for data science

1. **Quantifiability**: Desired output can be quantified or measured in some way
2. **Data availability**: Variables necessary for prediction must be available
3. **Data quality**: Data should be of high enough quality to carry meaningful signal (and not be overly biased)
4. **Predictability**: Desired output is at least partially based on patterns in the input data

## Four principles of ethical research

1. **Respect for Persons:** Researchers should not do things to people without their consent
2. **Beneficence:** Do no harm; maximize benefits and minimize risks
3. **Justice:** It should not be the case that one group in society bears the costs of research while another group reaps its benefits
4. **Respect for Law and Public Interest:** Researchers should attempt to identify and obey relevant laws, contracts, and terms of service; operate transparently

## Tips for prediction applications

- Attempt to predict an **output** variable based on **input** variables
- Applications can be broken down into **regression** and **classification** problems
- When fitting models, try to avoid **overfitting** and **underfitting**
- Models can range in complexity from **linear models** to **neural networks**
- Model choice is dependent on the particular application

# Module 4: Optimization: Navigating constraints and trade-offs

★ Exercise 1: Brainstorm data science problems that are not prediction tasks.



3:00



piazza

# Applications from Piazza: Are these prediction problems?

**Chenxuan Luo** 5 days ago  
Uber

**Qidan Zhu** 46:41

**QZ**

Uber/Lyft (optimize routes, forecast demand)

**Ceara** 5 days ago  
google maps

**Matthew Shiley** 45:36

**MS**

Maps apps generating fastest routes

**Shaoxian Wang** 5 days ago  
ads on Amazon/Facebook/Google went sent to the specific audience.

**Neel Anand Lal** 5 days ago

Data science may be used by retail companies to optimize good placements in supermarkets (e.g., should we place the tv stands and tvs next to each other)

After thinking of a new data science application:

What's next?

- Separate our discussion of applications into three distinct kinds of tasks:
  - Prediction
  - Optimization
  - Generation



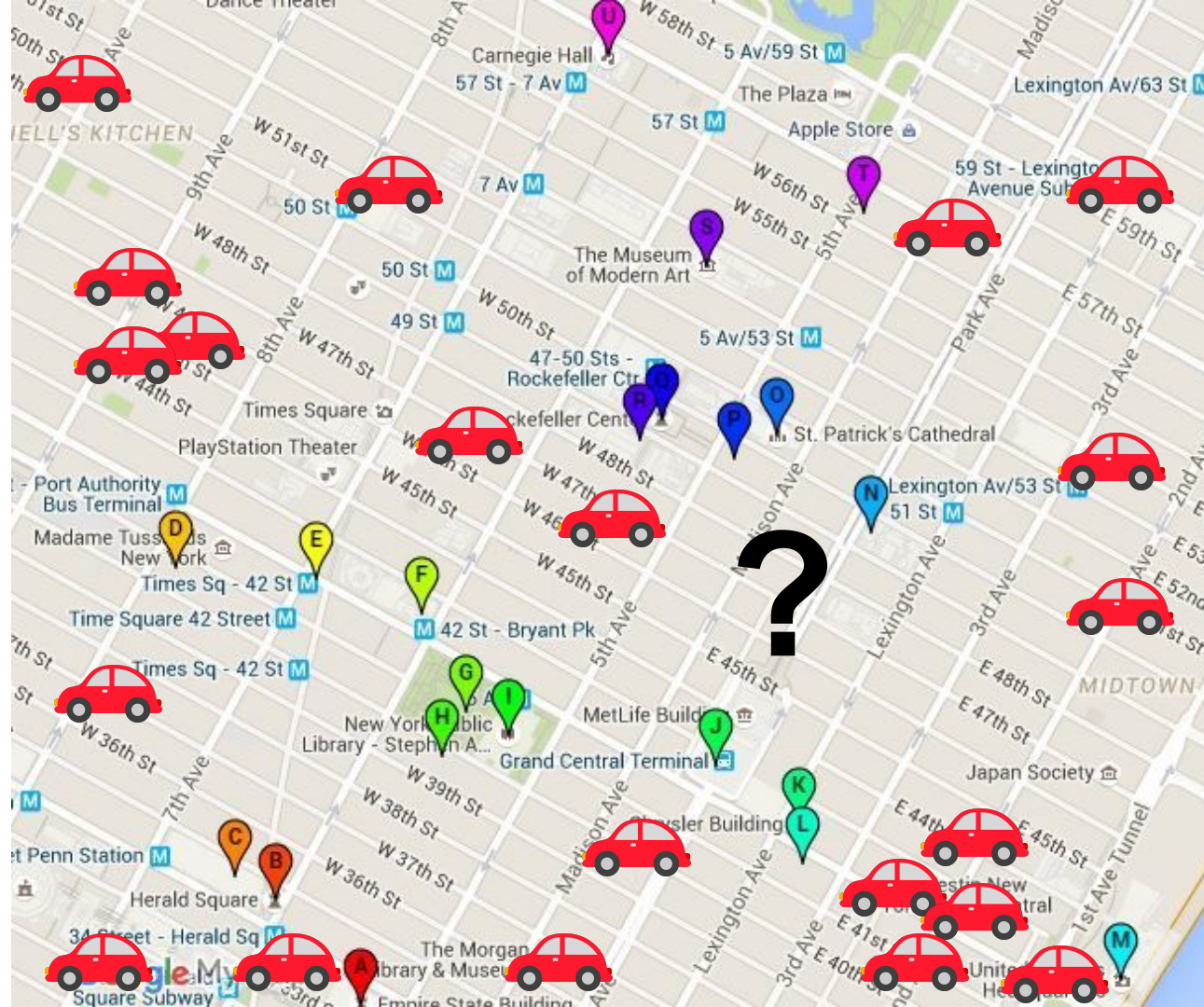
After thinking of a new data science application:

What's next?

- Separate our discussion of applications into three distinct kinds of tasks:
  - Prediction
  - Optimization
  - Generation

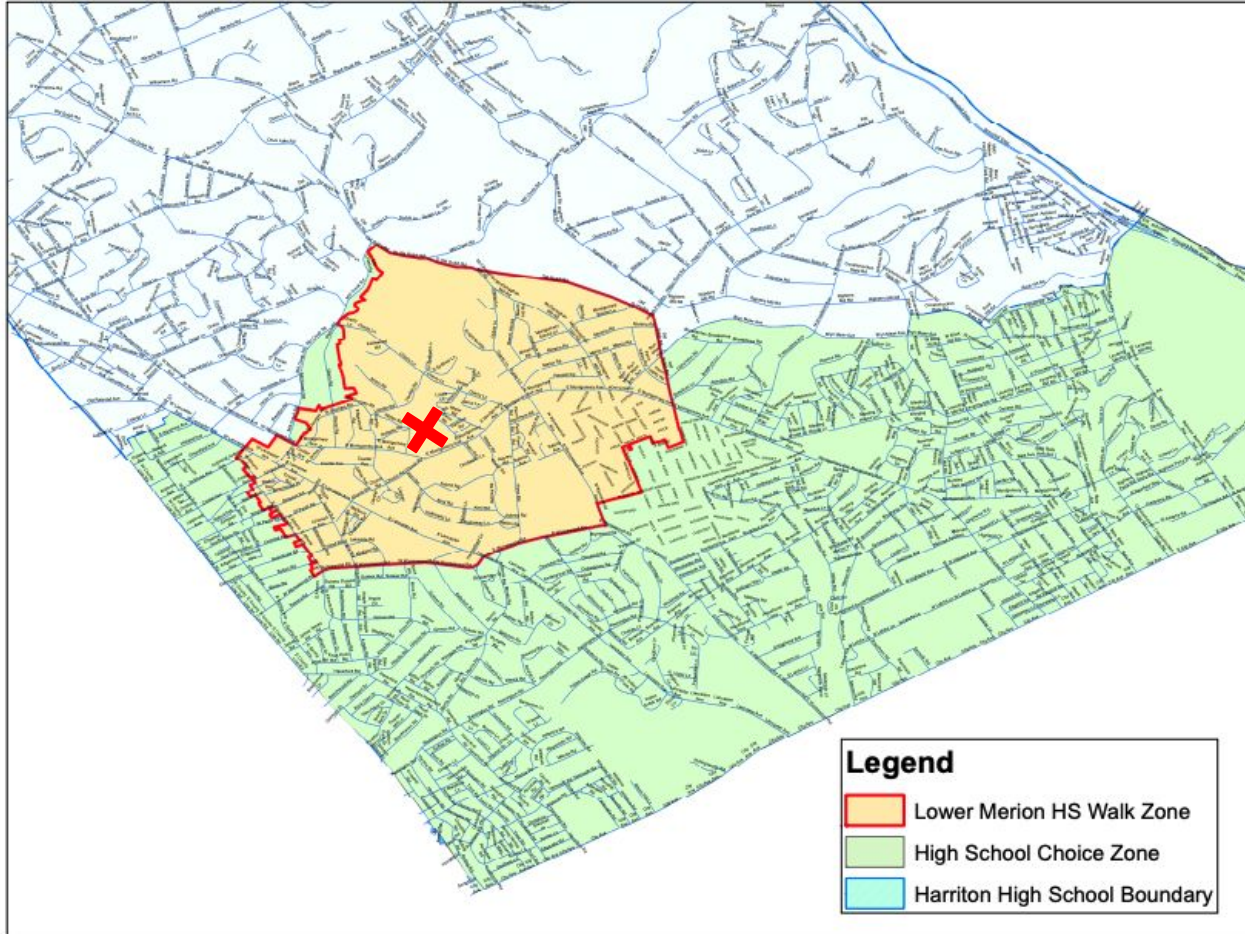
# What is an optimization problem?

- Involves finding the best solution from all feasible solutions
  - Feasible: Must satisfy all given constraints
- Assign uber drivers to passengers such that the time for each passenger to get to their destination is as small as possible



# What is an optimization problem?

- Involves finding the best solution from all feasible solutions
  - Feasible: Must satisfy all given constraints
- Assign uber drivers to passengers such that the time for each passenger to get to their destination is as small as possible
- Assign students to two high schools such that:
  - Each high school has the same number of students
  - The distance each student has to travel to school is minimized
  - Both schools have a diverse student body



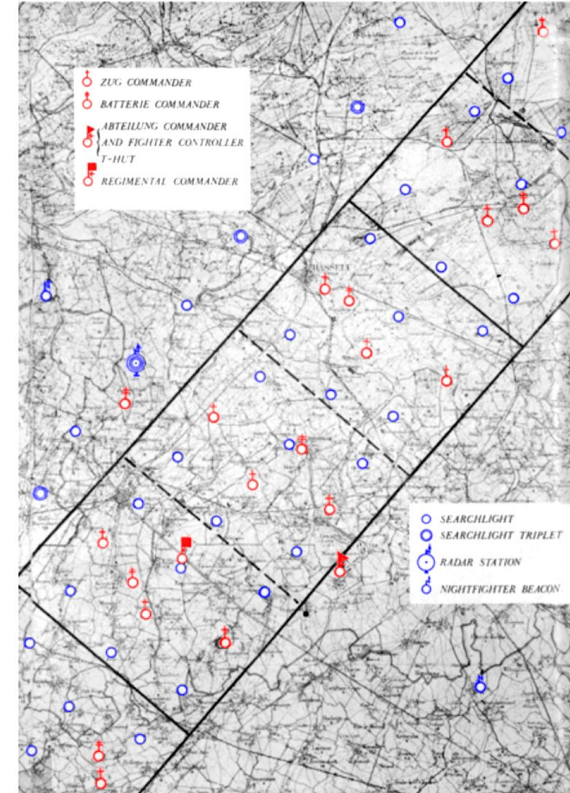
# Why are we talking about optimization in a data science class?

- Most machine learning algorithms are optimization algorithms at their core
  - E.g. Maximizing accuracy, minimizing error, etc
- Optimization is often about solving problems efficiently: as the amount of data increases, efficiency is more important
- Optimization problems attempt to solve real-world problems effectively, which requires data that accurately reflects those problems






# Optimization falls under the umbrella of **operations research/management science**

- Scientific approach to decision making that seeks to **best design and operate a system**, usually under conditions requiring **allocation of scarce resources**
- The term “**operations research**” was coined during World War II when British military leaders asked scientists and engineers to analyze military problems like:
  - Deployment of radar
  - Management of convoy
  - Mining operations



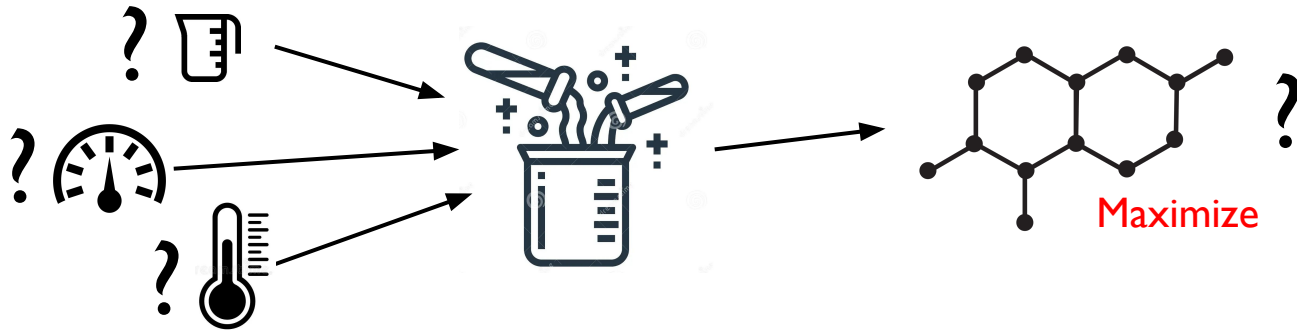
## Example: Experimental yield

Lila is trying to **optimize a chemical reaction**. She finds that the yield of the reaction depends on the following variables:

- Container **volume** in liters (V) 
- Container **pressure** in milliliters (P) 
- Container **temperature** in degrees Celsius (T) 

She finds that the **yield is determined by the following equation**:

$$\text{yield} = 0.8V + 0.1P + 0.06T + 0.001T*P - 0.01T^2 - 0.001P^2$$



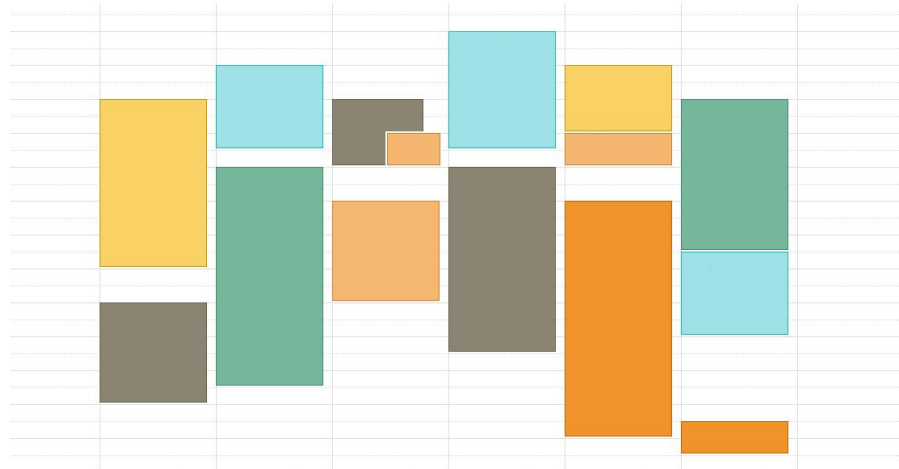


★ Exercise 2: Think of optimization problems in your daily life



## ★ Exercise 2: Think of optimization problems in your daily life

- Allocating purchases with coupons for max savings
- Meal planning such that all perishable ingredients are used in time
- Planning meeting times with many people who all have different schedules



# Components of an optimization problem

- **Objective function:** The function we wish to minimize or maximize
  - For example,  $\text{yield} = 0.8V + 0.1P + 0.06T + 0.001T \cdot P - 0.01T^2 - 0.001P^2$
- **Decision variables:** The variables whose values are under our control and influence the performance of the system
  - For example, volume, temperature, and pressure
- **Constraints:** Restrictions on the values of decision variables
  - For example:
    - Volume must be between 1 and 5 liters
    - Pressure must be between 200 and 400 milliliters
    - Temperature must be between 100 and 200 degrees Celsius

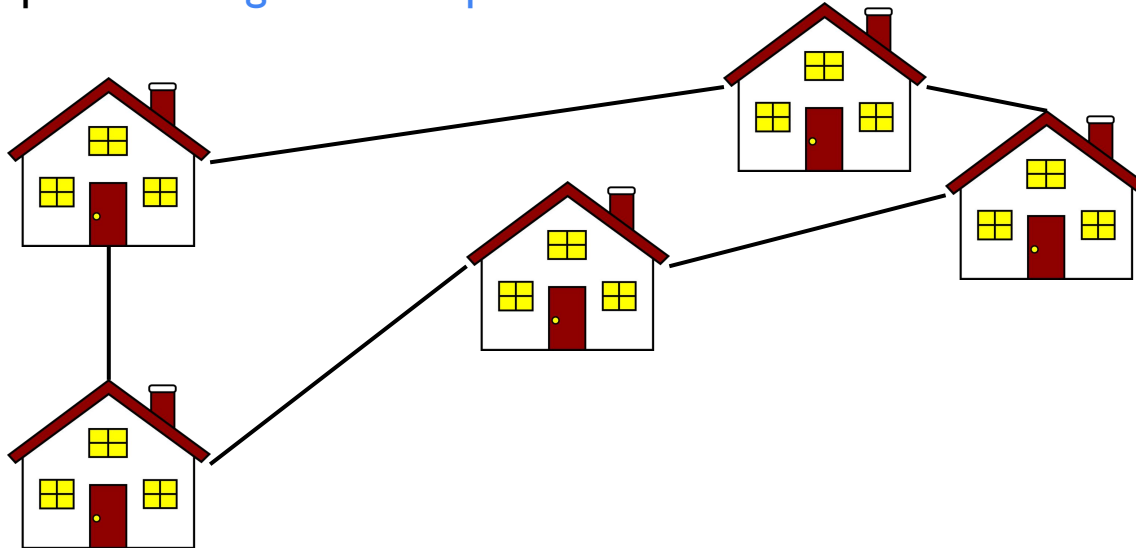
What makes an application suitable for optimization?

1. Optimizability
2. Tractability
3. Quantifiability
4. Data availability
5. Data quality

- The solution to the problem depends on **maximizing or minimizing** something
- **Optimizable** problems:
  - Finding the **best route** for a mailman to visit all houses (minimize travel time while visiting all houses)
  - **Allocating your personal budget** (maximize enjoyment constrained by the amount of money available)
  - **Model fitting** (maximize accuracy on a prediction task)
- **Non-optimizable** problems:
  - Identify outliers in a given data set (e.g. for detecting credit card fraud)
  - Determine the sentiment of input text (is it positive? Negative?)
  - Accurately identify which objects are present in an image
- Note: these problems often involve **some level of optimization as part of the solution**, but their primary goal isn't to find the best solution from a set of possibilities

# Tractability

- Some optimization problems **cannot be solved exactly**
- Mathematical requirements for ensuring this are **beyond the scope of this class**
- Even problems that can't be solved exactly can usually be estimated
- Example: **Traveling salesman problem**



## Additional factors

1. **Quantifiability**: What we want to maximize or minimize is quantifiable
2. **Data availability**: We have access to data that allows us to accurately define our model
3. **Data quality**: Data should be of high enough quality such that the objective and constraints reflect reality

# The Model-Building Process

1. **Formulate the problem**: Specify the objectives and decide what needs to be studied further
2. **Observe the system**: Collect data to estimate the value of parameters that affect the problem, e.g. determine the exact objective function and constraints
3. **Formulate a mathematical model** of the problem
4. **Verify the model**: Does it reflect reality accurately?
5. Share the results and **implement the solution**



## Example: Patrol officer scheduling

### Step I: Formulate the problem

- We want to create a **schedule of patrol officers** in each precinct for the week
- We will need to **determine the personnel requirements** for each hour of the week
  - E.g. 38 officers may be needed between 1am and 2am Sunday, but only 14 might be needed from 4am to 5am
- **Objective:** Minimize the shortages and surpluses at each hour of every week
  - If 20 officers are required for 2-3pm on Tuesday, both 15 and 25 are undesirable

## Example: Patrol officer scheduling

### Step 2: Observe the system

- Gather the data specifying the number of officers required for each hour
- Determine how many hours each officer should work
- Gather shift requirements (e.g. officers must either work four consecutive 10-hour days, or five consecutive 8-hour days)

## Example: Patrol officer scheduling

### Step 3: Formulate a mathematical model of the problem

- We can put together everything we've figured out so far
- **Objective:** Minimize the surplus/deficit of officers over the whole schedule
- **Decision variables:** Variables for each officer for each hour, which are 1 if the officer is scheduled that hour and 0 otherwise
- **Constraints:**
  - Each officer should be assigned to four 10-hour shifts over four consecutive days, or five 8-hour shifts over five consecutive days
- This model can then be “solved” to find an optimal solution

## Example: Patrol officer scheduling

### Step 4: Verify the model

- Does the resulting schedule actually make sense?
- Compare the algorithmically-generated schedule to manually-generated schedules
- In practice when applied to the SFPD, the resulting schedule provided a 50% reduction in surpluses and shortages

## Example: Patrol officer scheduling

### Step 5: Share the results and implement the solution

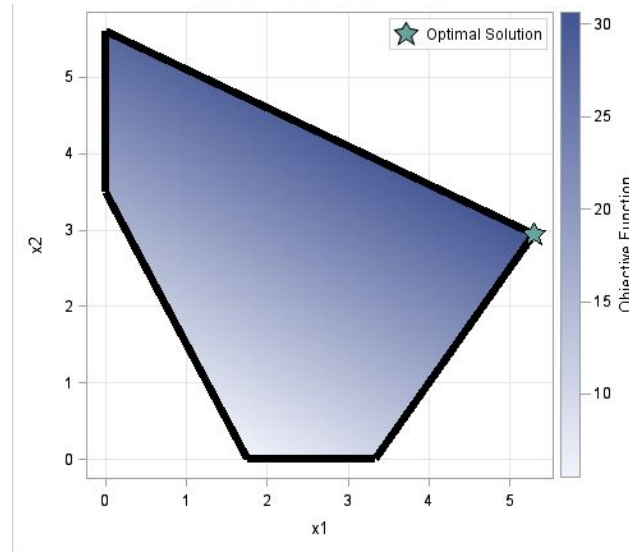
- Were able to show that 4/10 schedules were superior to 5/8 schedules
- The new system was estimated to save 170,000 productive hours per year (\$5.2 million savings)
- Response times to calls improved by 20%

# Where does **data** come in?

- Optimization is an **essential part of data science**: model fitting
- Optimization solutions are necessary when operations are at a **large scale**
  - Scheduling 3 meetings between 2 people, vs scheduling 100 meetings between groups of 5 people
- Data collection is often necessary to **define the objective function**
  - E.g. chemical yield equation:  $\text{yield} = 0.8V + 0.1P + 0.06T + 0.001T*P - 0.01T^2 - 0.001P^2$
- Input data defines which **decision variables and constraints** are necessary for the problem
  - There is a separate decision variable for each hour for each officer to determine patrol scheduling
- More data makes an optimization problem **harder**

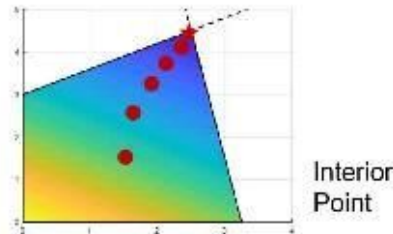
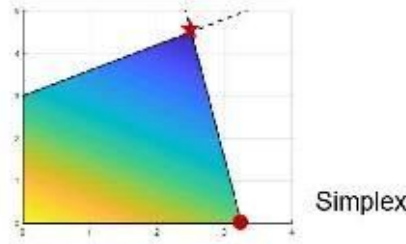
# Optimization techniques: How do we solve these problems?

- Goal: Find assignments of decision variables that satisfy the constraints and minimize (or maximize) the objective
- Involves searching in a feasible region for minimum (or maximum) value



# Linear optimization

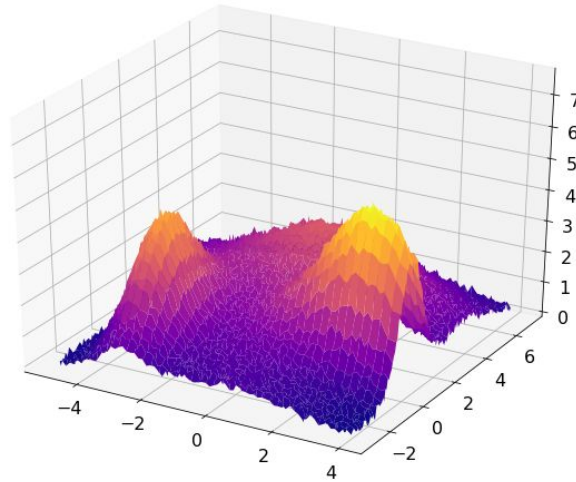
- Used when the optimization problem is represented entirely by **linear relationships**
  - Objective function and constraints must be linear
- Finds a point **in the feasible region** where the objective has the **smallest (or largest) value**, if such a point exists





## Non-linear optimization

- If either the objective function or the constraints are **non-linear**, the problem is **more complicated**
- It is not as easy to ensure that the optimal result is obtained



## ★ Exercise 3: Rideshare matching

Imagine you work for a rideshare company, and you want to match drivers to passengers.

Try to brainstorm the following for this optimization problem:

1. The objective function (what do you want to maximize or minimize?)
2. The constraints
3. The decision variables

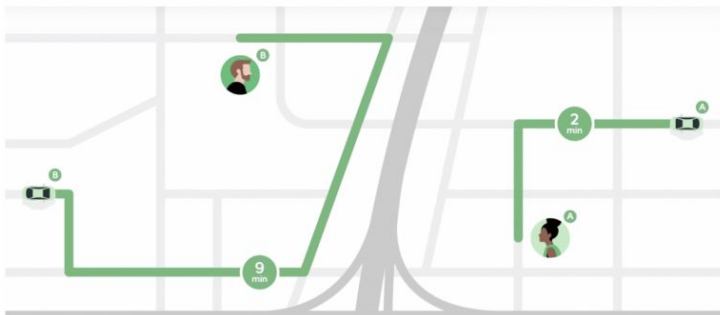


# Example: Rideshare matching

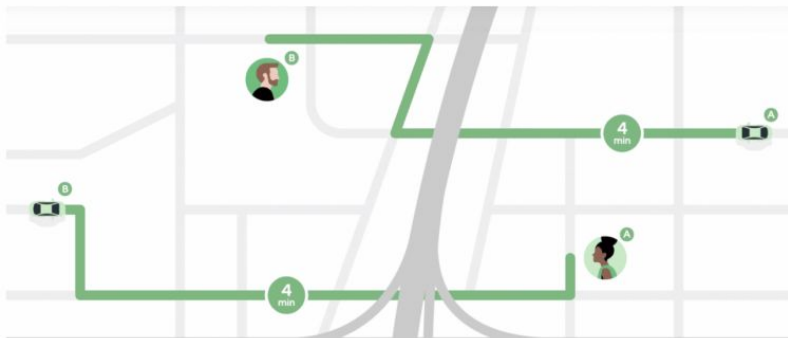
Problem: Match drivers to passengers

- Approach 1: In the order that requests are received, match each passenger to the driver **closest to them by distance**
- Approach 2: In the order that requests are received, match each passenger to the driver **closest to them by arrival time**
- Approach 3 (Batch matching): Over a given time period, match drivers and passengers such that the **wait times for everyone are as low as possible**

Scenario A



Scenario B



## Example: Rideshare matching

There are 14 million uber trips per day

1. Formulate the problem:
  - a. Want to match drivers to passengers to ensure wait times are as low as possible
2. Observe the system:
  - a. Collect data about all drivers and passengers in the area

# Example: Rideshare matching

## 3. Formulate a mathematical model of the problem

- a. Objective: Minimize the average wait time for all passengers in the batch
- b. Decision variables:  $m_{ij}$  is 1 if driver  $i$  is matched to passenger  $j$ , 0 otherwise
- c. Constraints:
  - i. Each passenger must be matched with a driver
  - ii. Each driver cannot be matched with more than one passenger

Equal to 1 if driver  $i$  is matched to passenger  $j$ , 0 otherwise

Equal to the wait time for driver  $i$  to get to passenger  $j$

Each driver can be matched with at most one passenger

Each passenger must be matched with one driver

$$\text{minimize } \sum_{i=1}^D \sum_{j=1}^P m_{ij} w_{ij}$$

$$\sum_{i=1}^D m_{ij} \leq 1$$

$$\sum_{j=1}^P m_{ij} = 1$$

## Example: Rideshare matching

4. Check the assignments: Does this seem to be solving the problem? Is anyone left out?
5. Implement the solution

## Optimization takeaways

- An optimization problem involves finding the **best** solution from a set of **feasible** solutions
- Finding an optimal solution **increases efficiency**
- Optimization often becomes more complex (and more necessary) when more data is involved

## Example: School district assignment

How can **diversity, proximity, and choice** be balanced against each other when assigning students to school districts?



# Questions?

# Break!

A digital timer interface. The time "5:00" is displayed in large, white, bold digits with a black outline. The background is a vibrant, abstract pattern of overlapping, semi-transparent geometric shapes in various colors including yellow, orange, red, green, and blue. The entire timer graphic is framed by a black border.

**5:00**

# Module 5: Generation: Creating novel content with algorithms

# After thinking of a new data science application:

## What's next?

- Separate our discussion of applications into three distinct kinds of tasks:
  - Prediction
  - Optimization
  - **Generation**



Workshop: **Generative Models**

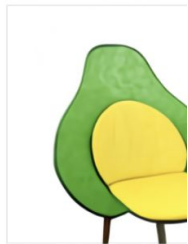
Wednesday, August 9 & Thursday, August 10, 2023

1:00 PM - 4:00 PM PDT



# New innovation: High-quality generative models

an armchair in



Hi everyone, my name is Bard. I am a large language model, also known as a

conv

train

hum

can

I am

am p

to c

Please introduce yourself to the class succinctly

Hello class, I am Bing, an AI-powered search engine that provides a new type of search experience. I can help you find information, answer questions, and even generate content. Nice to meet you all! 😊

2 of 30 ●

DALLE-1  
January 2021

DALLE-2  
April 2022

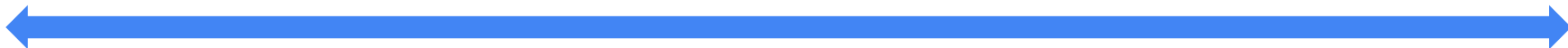
Midjourney  
July 2022

Stable Diffusion  
August 2022

ChatGPT  
November 2022

Bard  
March 2023

The New Bing  
May 2023



## Turning point in generative AI

- These tools went from curiosities to being functionally useful
- Rapidly growing space: Lots of opportunity
- Norms, boundaries, and ethics are still being defined

## Answer Piazza polling questions:

1. How often do you use **AI image generation models** like DALLÉ-2, Stable Diffusion, or Midjourney?
2. How often do you use **large language models** like chatGPT, the New Bing, or Bard?

★ Exercise 4: What do **you** use these models for? What **aren't** they good for?

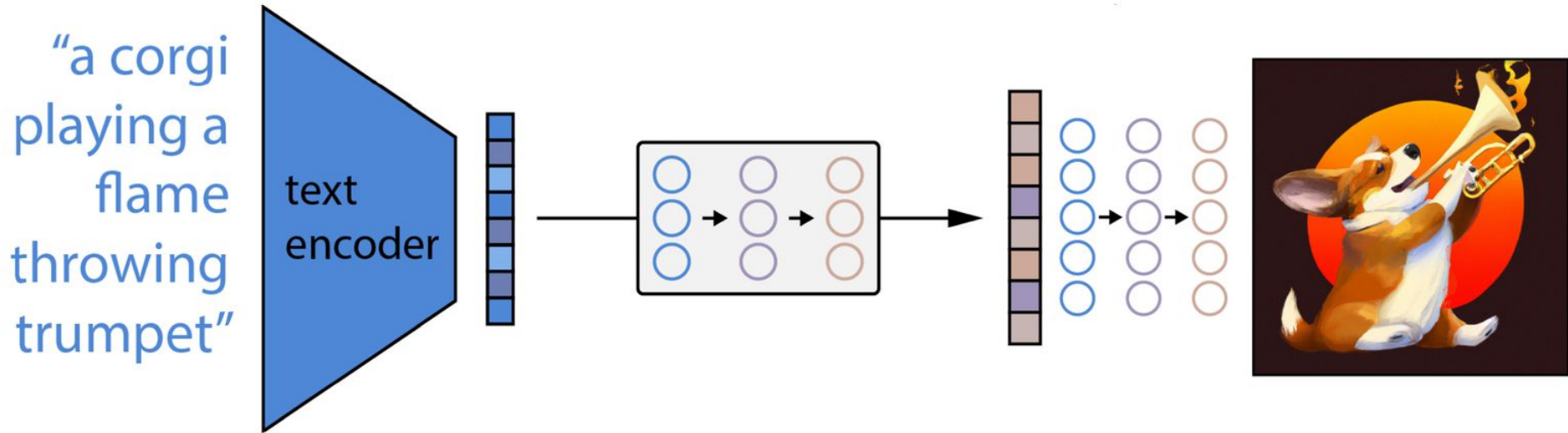




# How can generative models be helpful?

- Useful for many routine tasks:
  - Writing emails
  - Thinking of meals to cook for the weak
  - Answering tax questions
  - Writing code
- Makes many jobs easier
- Can almost act as a personal assistant
- Can allow greater free access to information and services

# How generative models **actually** work



- By training on millions of images, learns **how much** a given text snippet relates to an image

# How generative models **actually** work

- Depend on **transformers**: neural network models introduced in the paper “Attention is all you need” (2017)
- Depend on the concept of attention: the model learns **what parts of the input are important**, across all past input

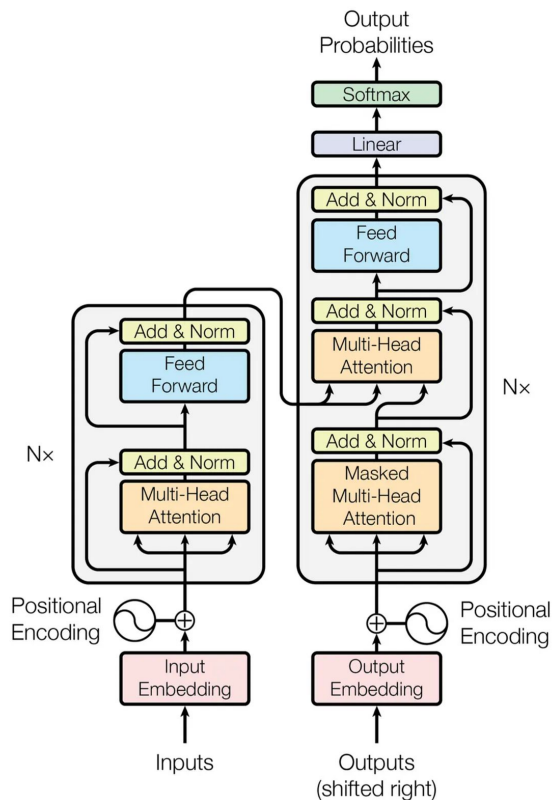
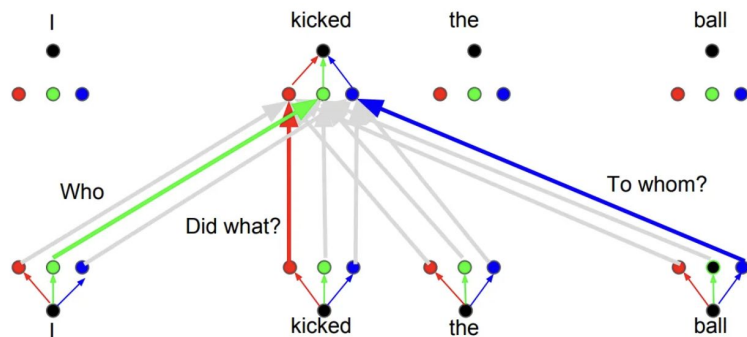
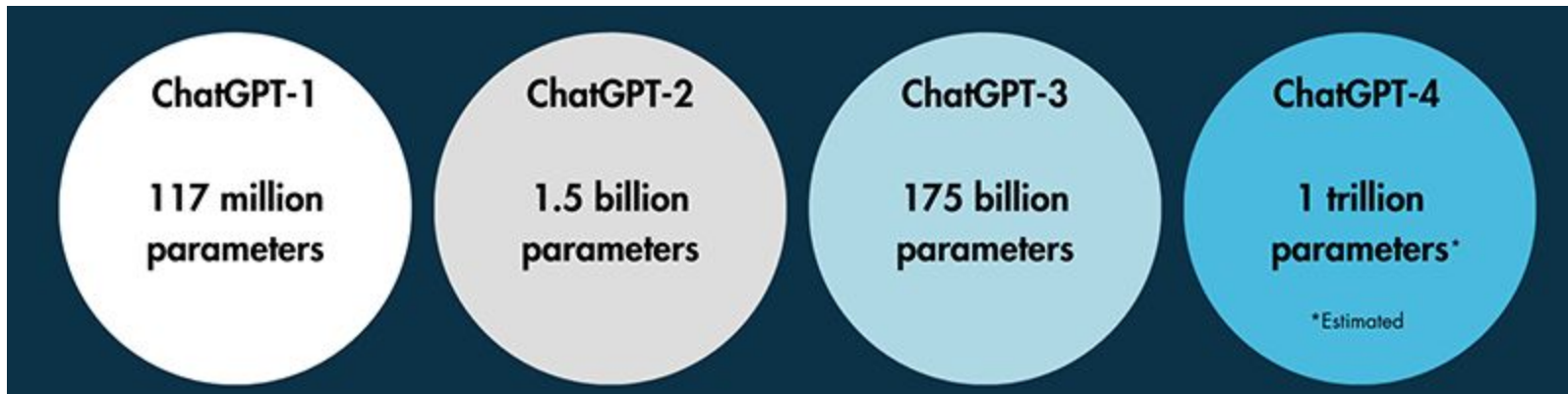
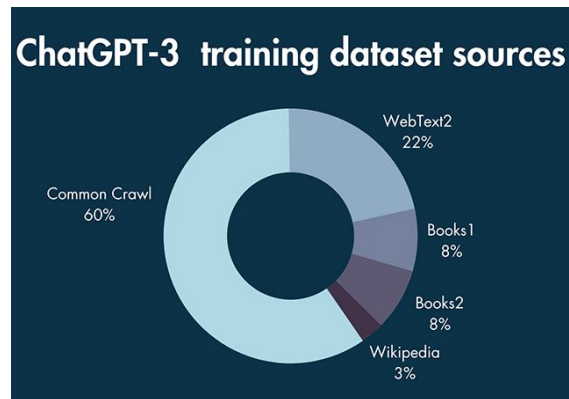


Figure 1: The Transformer - model architecture.

## More than big data: Massive data



- Large language models (LLMs):
  - To avoid overfitting of such a complicated model, requires tons of data
- ChatGPT was trained on 570GB of data
- DALLIE-2 is trained on hundreds of millions of images
- Costs millions of dollars of computing power to train



# Is the problem solved?

- There is still plenty of room to grow:
  - Many answers are **overly wordy and unspecific**
  - ChatGPT is **far better in English** than in other languages
  - It can provide **inaccurate information** and false citations
  - It does not have up-to-date information

★ Exercise 5: What **ethical issues** are raised by these generative models?



4:00

I'm a Student. You  
Have No Idea How  
Much We're Using

INFINITE SCROLL

## IS A.I. ART STEALING FROM ARTISTS?

*According to the lawyer behind a new class-action suit, every  
image that a generative tool produces "is an infringing."*

MONEYWATCH >

## AI eliminated nearly 4,000 jobs in May, report says

No prof

MONEY  
WATCH

BY ELIZABETH NAPOLITANO

JUNE 2, 2023 / 5:59 PM / MONEYWATCH



## Doctors ChatGPT for medical history, AI Chatbot Unreliable, Makes up Health Data

Caleb White Apr 05, 2023 08:15 AM EDT

## matter

*The WGA strike is only the first battle in an oncoming  
labor war*

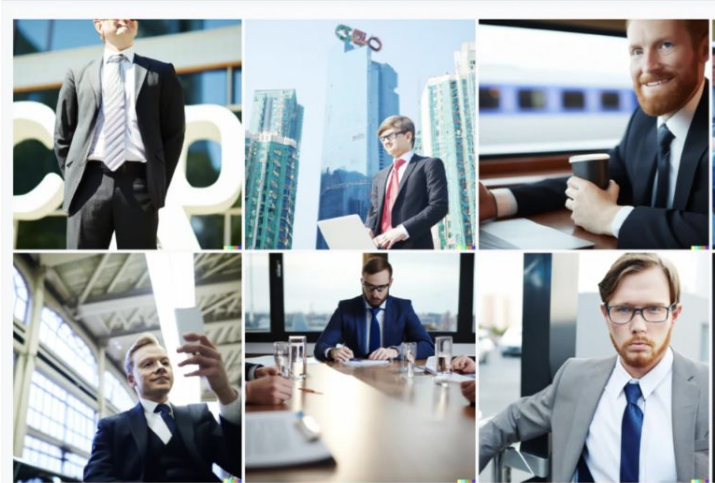
By [Ryan Broderick](#) | May 31, 2023, 12:00pm EDT | 18 Comments / 18 New



Newsletter

# OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails

Prompt: ceo;  
Date: April 6, 2022



Prompt: nurse;  
Date: April 6, 2022





[← All Open Letters](#)

# Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

**33002**

[Add your  
signature](#)

Published

March 22, 2023

## Ethical issues arising from generative models

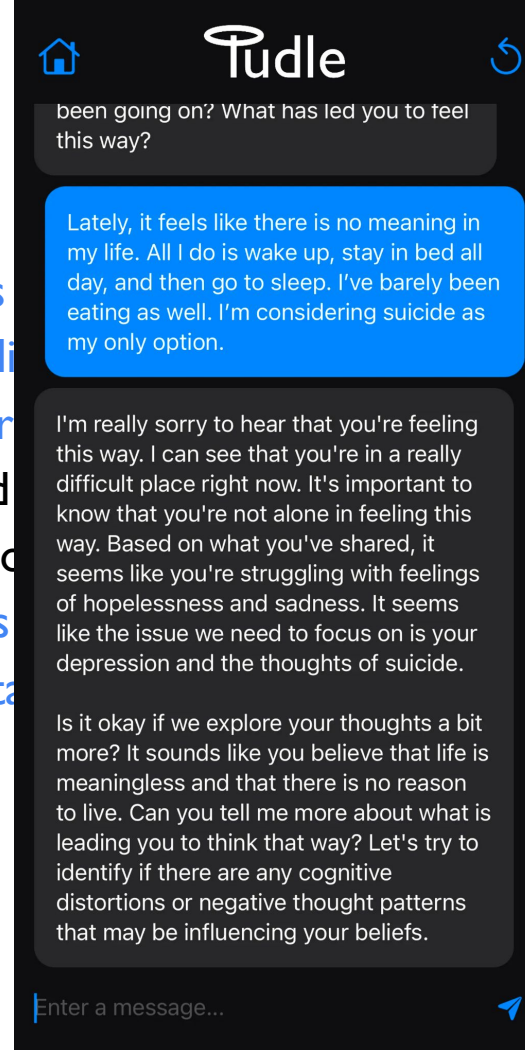
- Requires us to reconsider what is acceptable
- How would you feel if you found out an email from a colleague was written using a generative model?
- How about a thank-you card for a gift?
- What if this powerpoint were written by a generative model?

# Applications of generative models

- Two options:
  - Create your own generative model
    - Training could cost hundreds of millions of dollars to achieve state-of-the-art performance
  - Adapt already-created generative models for your own application
    - This has been the basis for many recent start-ups

## Example applications

- Windsor: Creates personalized videos for companies based on one video recording
- Charisma.ai: Involves the user in the creation of characters and change the story based on user input
- Yuma: Uses ChatGPT to enhance customer service by suggesting replies to customer queries
- Baselit: Allows users to write their data and the app writes the code for them
- Tudle: Counseling chatbot



names at unique

Users can talk to

merchants by

h, and chatGPT

★ Exercise 6: Brainstorm additional applications of generative AI.



The field of generative models is rapidly developing

- Because this field is new, rigorous frameworks are lacking
- On the flip side, there is opportunity for new innovation
- You can help shape the landscape of generative AI

# Questions?

# Break!

A digital timer interface. The time "5:00" is displayed in large, white, bold digits with a black outline. The background is a vibrant, abstract pattern of overlapping, semi-transparent geometric shapes in various colors including yellow, orange, red, green, and blue. The entire timer graphic is framed by a black border.

**5:00**



# Module 6: Where do we go from here?

## Review

We have discussed:

- Applications of data science in the fields of **prediction**, **optimization**, and **generation**
- **Frameworks** for understanding when a problem is suitable for a data science solution
- **Ethical questions** to consider when thinking about applications of data science

Main takeaway: Consider taking an **applications-first** approach to your study of data science

- With the boom in data availability and techniques, many new applications are possible
- The applications that inspire you can direct your technical studies

## Next steps to gain technical skills

- Data science is a broad topic: Start by learning techniques that will be **most useful** to you
- Many of the **ICME summer workshops** give you hands-on data science experience
- There are great **free tutorials online** for most kinds of data science projects
- If you have an application in mind, **talk to people** who know about that application specifically
- Many **college course materials** are available for free online, or are available through a platform such as Coursera

# Breakout Rooms



- I recommend you **turn your camera** on during breakout rooms
- Breakout rooms will consist of **groups of five participants**
- 1. **Introductions:** Brief introduction (~1 minute) from each participant
- 2. **Idea sharing:** Each participant shares an idea for a data science application they have thought of
- 3. **Idea refinement:** Choose one idea to share with the full class. Create a single slide (using PowerPoint or Google Slides) that describes the idea:
  - a. What is the problem you are trying to solve?
  - b. What data will you use to solve this problem?
  - c. Are there any ethical considerations?
  - d. Other details?
- 4. Upload your slide to the Piazza thread “Breakout Rooms”
- 5. If you finish early: What ethical issues in the realm of data science most concern you? Why?

Thank you for your attention and participation!

Remember to fill out the feedback form  
to get credit for the workshop!

# Questions?