COMS BC1016
Introduction to Computational Thinking and Data Science

Special Topics: Data Privacy

Dec 8, 2025

Last class!

- Office hours are happening this week (Mon, Tues, Wed, Thurs)
    - Murad’s office hours will be remote today
- Final Reports are due on Friday
    - Don’t forget to do the peer review!!
      Link: https://forms.gle/JtxDWfHEAQBwuXus6
- Jupyter Hub will shutdown sometime after the semester ends
    - Download anything you want to keep from the server!

Jupyter on your own

- Anaconda is a distribution platform for Python
  - Install Anaconda https://www.anaconda.com/download
  - Open up the Anaconda Navigator
  - Launch a new notebook
  - Install `datascience` package:
    - `pip install datascience`

- https://edblogs.columbia.edu/eescx3050-001-2015-3/category/classes/class-1-intro/

Data Science

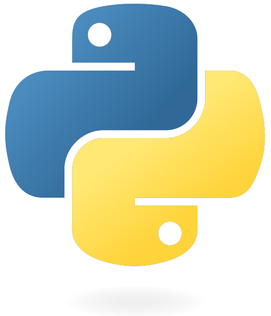
- Unprecedented access to data means we can make new discovers and more informed decisions

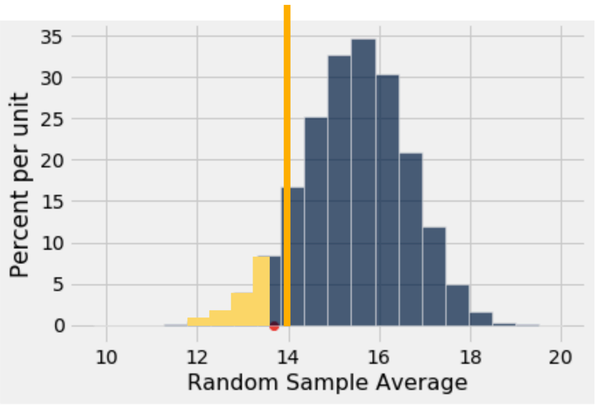
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference

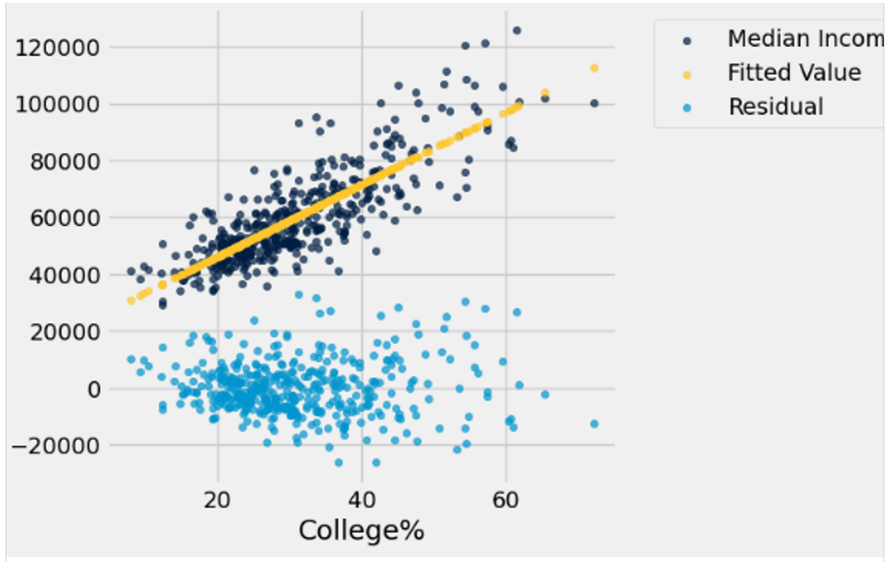
    - People can agree on evidence and measurement

We’ve learned a lot through this course

- **Computational Thinking**: Python programming!
- **Exploration**: Identifying patterns and trends using data (e.g., through visualizations)
- **Inference**: Drawing reliable conclusions using statistics
- **Prediction**: Making informed guesses about patterns using models







Limitations of Data Science

- Evidence and measurements are critical ingredients for good decision-making

  - … But they’re not enough by themselves!

- Data science is a powerful complement to quantitative analysis, but it’s not a replacement

How to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods

- Visualize, then quantify!

- Interpretation of the results in the language of the domain without statistical jargon

    - Perhaps the most important part!

How to Analyze Data after 1016

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods

- Visualize, then quantify! **Do both using computation**

- Interpretation of the results in the language of the domain without statistical jargon

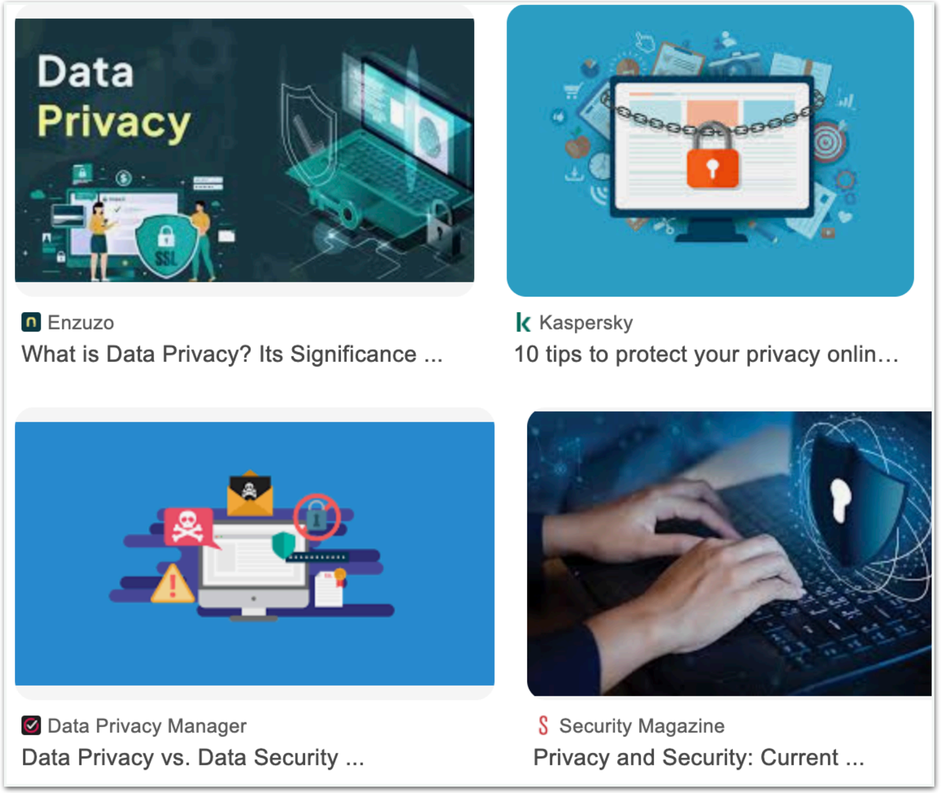
    - Perhaps the most important part!

I hope this skills are useful for you in your respective domains :)

Data Privacy

Privacy

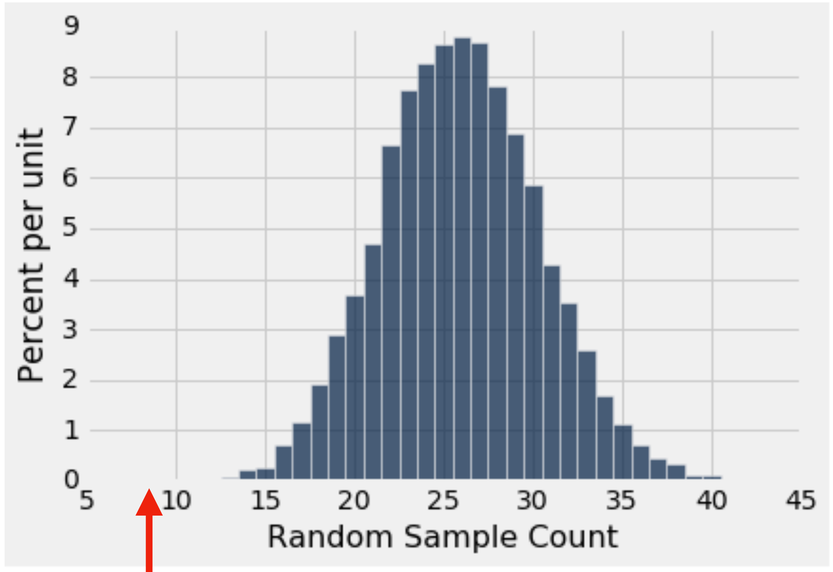
- Everyone has an intuitive notion of “privacy”
- Why is it important?
  - Freedom: I can form opinions without external influence
  - Personal dignity/autonomy/independence
  - Interpersonal relationships (decide what to share and with whom)
  - …



How to balance privacy and usefulness?

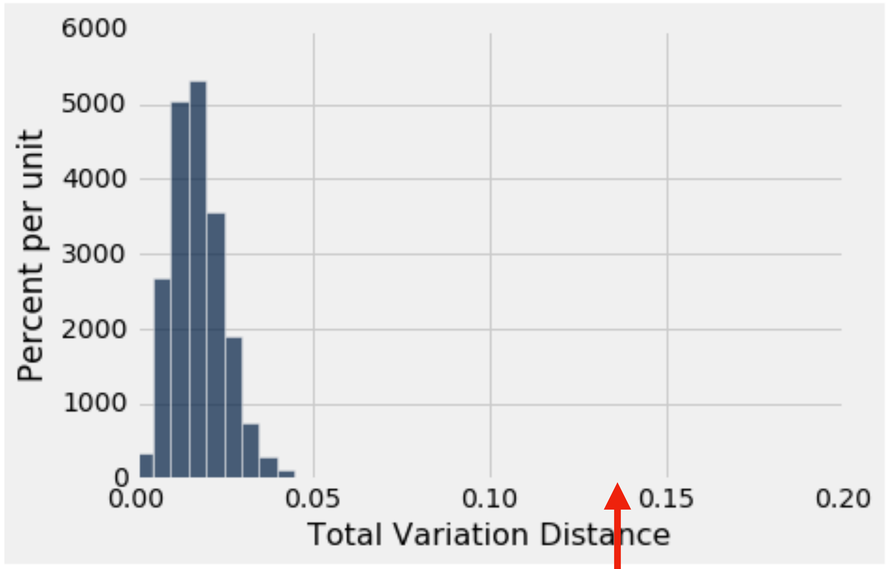
- Throughout this class we've seen how useful data has been

Swain v Alabama



Observed Number (8)

Alameda Jury



Observed TVD (0.14)

How to balance privacy and usefulness?

- Consider a hospital that collects patients' personal data and symptoms
- A data analyst wants to use this data set to answer if lung disease is linked with high blood pressure
- What are possible solutions?

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

Solution 1: Give the analyst the entire data set

- Analyst can perform the study
- However, this completely violates patient privacy

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

Solution 2: Remove PII from the data

- For many years, removing personally identifiable information (PII) was the approach taken to allow people to analyse data

- What could go wrong?

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

Name	Blood Pressure	Lung Disease
[redacted]	125	Yes
	140	Yes
	145	No

Can you guess the person?

Name	Birthday	Hometown	Occupation
[redacted]			

Can you guess the person?

Name	Birthday	Hometown	Occupation
[redacted]	Dec 13, 1989		

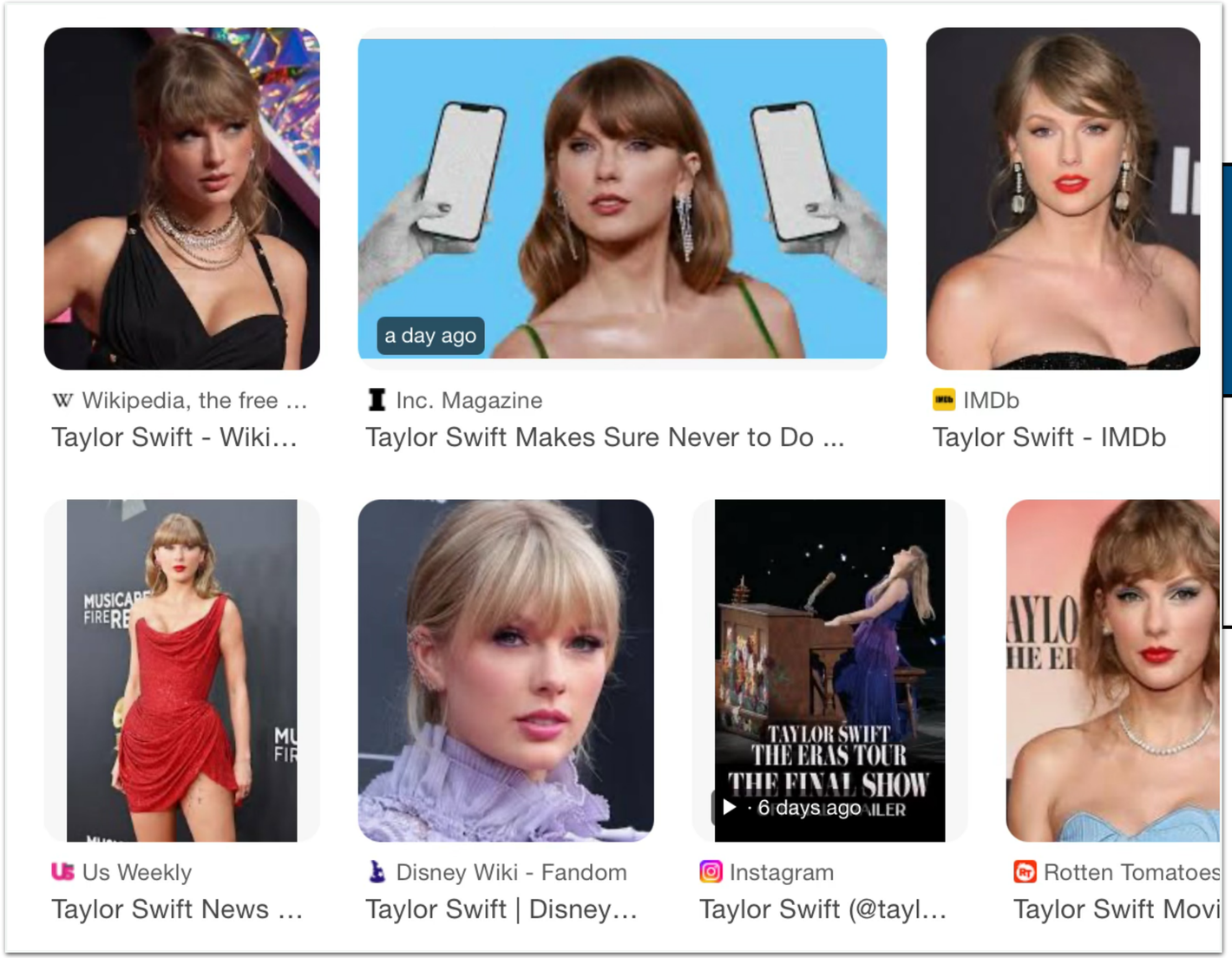
Can you guess the person?

Name	Birthday	Hometown	Occupation
[redacted]	Dec 13, 1989	West Reading, PA	

Can you guess the person?

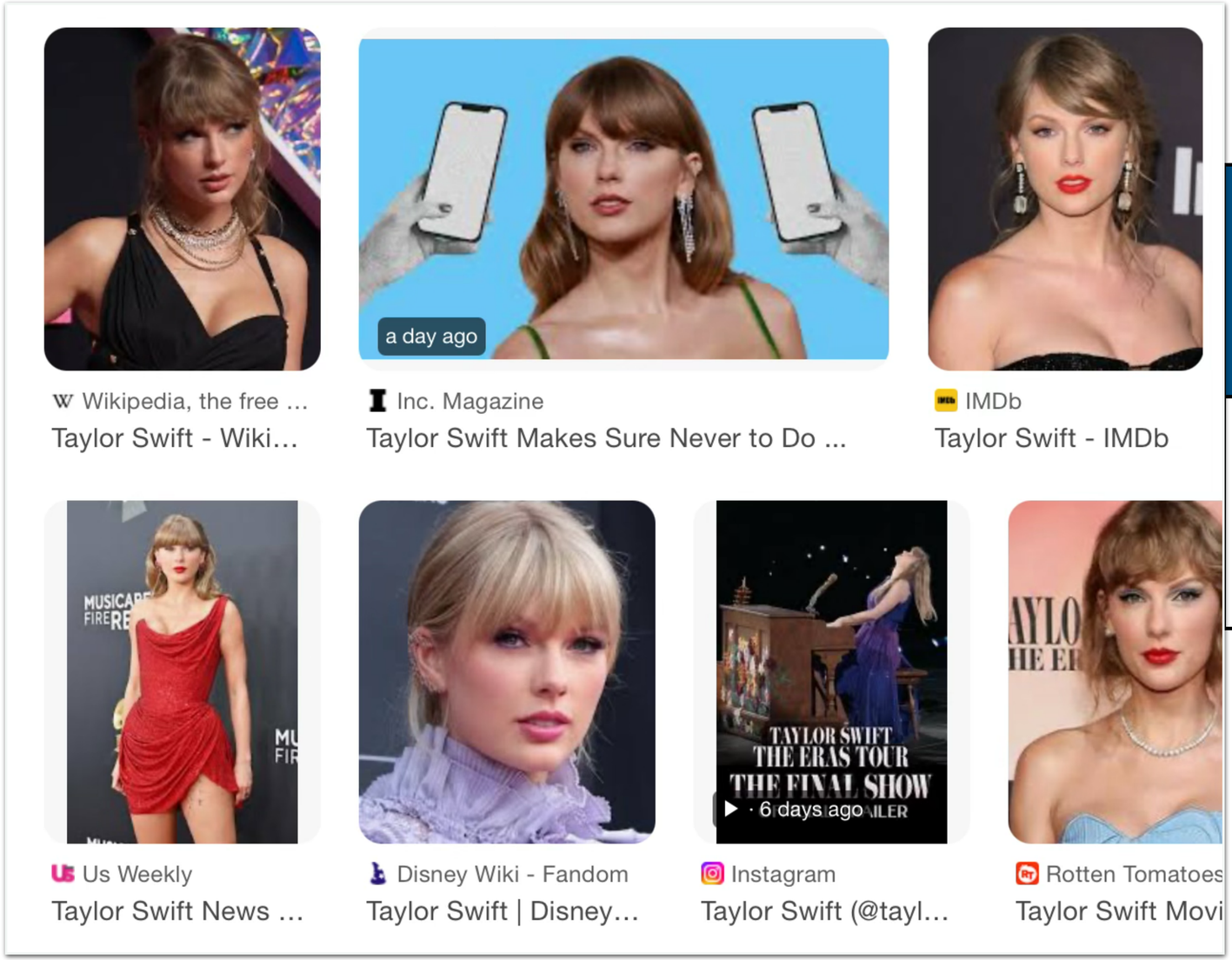
Name	Birthday	Hometown	Occupation
[redacted]	Dec 13, 1989	West Reading, PA	Singer-songwriter

Can you guess the person?



Hometown	Occupation
West Reading, PA	Singer-songwriter

Can you guess the person?



Hometown	Occupation
West Reading, PA	Singer-songwriter

Certain combinations of attribute are uniquely identifiable!

Simple Demographics Often Identify People Uniquely

Around 2000, Latanya Sweeney used 1990 US Census summary data to estimate how uniquely identifiable certain attribute combinations were:

- 87% of the population likely uniquely identifiably by only 5-digit ZIP, gender, and date of birth

- 53% uniquely identified by only place (city, town, or municipality), gender, and date of birth

- 18% uniquely identifiable by county, gender, and date of birth

Relinking de-identified data

- Many states have legislative mandates to collect and release hospital-level data

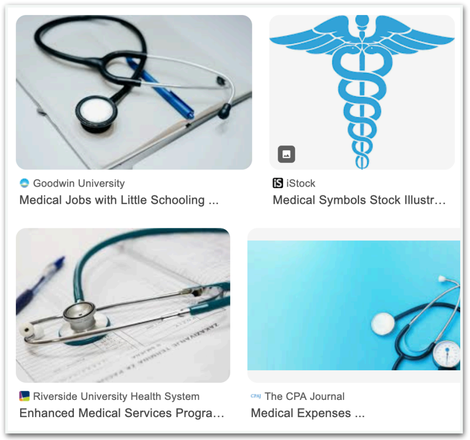
- Quasi identifiers like name, address and social security numbers were removed

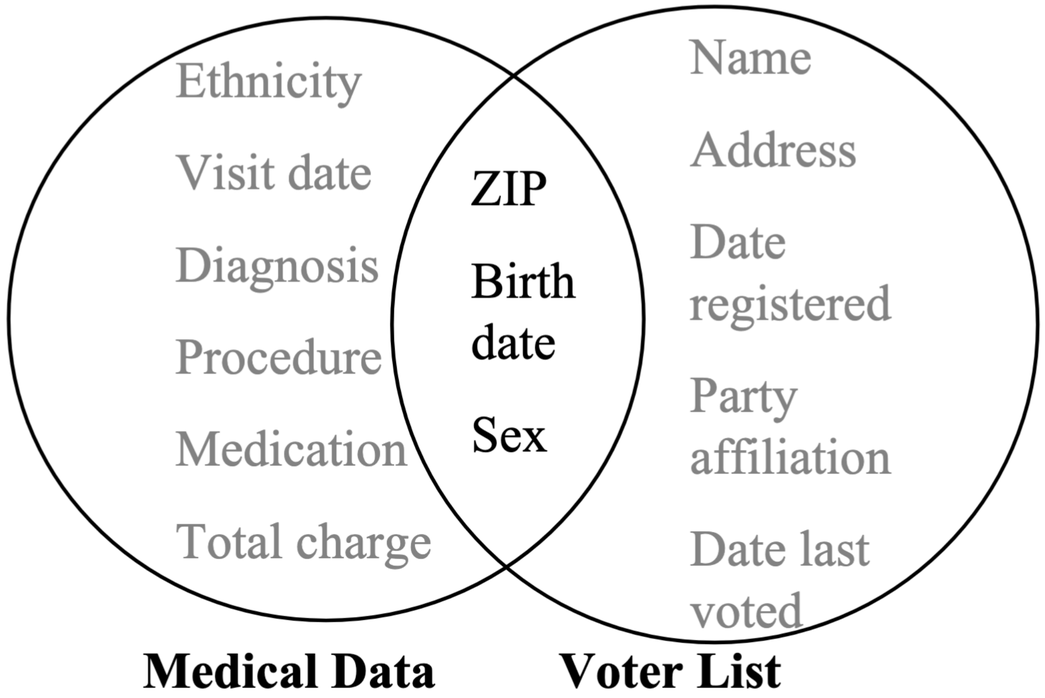
- Data still had zip code, date of birth, and gender

    - What if we combined this medical data with (semi-)publicly available data?

Relinking de-identified data

In 1997, Latanya Sweeny famously de-anonymization a MA hospital discharge database by joining it with a public voter database





AOL Search Log Release

- In 2006, AOL Research released 20 million search queries of over 650,000 users over a three-month period
    - The file was removed by not before being reposted many places
- Users were labeled with numeric IDs
    - However, many search queries contained PII

August 10, 2006

A Face Is Exposed for AOL Searcher No. 4417749

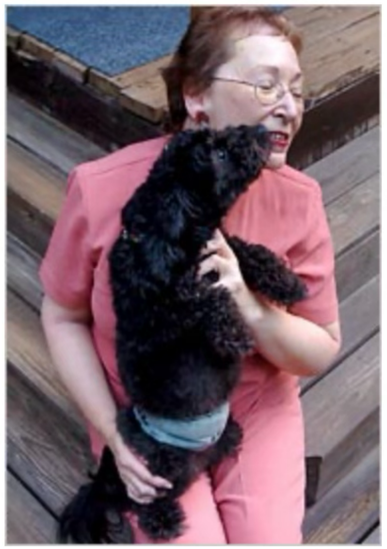
By MICHAEL BARBARO and TOM ZELLER Jr.

GO TO LESSON PLAN

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Netflix Prize

- In 2006, Netflix published 100 million movie rankings by ~500,000 users as part of a million dollar competition for a better recommendation system
    - Data was anonymized by removing personal details and replacing names with random numbers
- Narayanan and Shmatikov gave an de-anonymization attack using IMDB as auxiliary information
    - The attack is robust to perturbations in the data and partial mistakes in background knowledge
    - Can identify 99% of records in the dataset with only 8 movie ratings

But does privacy of movie ratings matter?

- Different information may be deduced from public information vs private information

- Privacy isn't always necessarily about the *average* person, but if there is *anyone* who may care

    - A user may be okay voluntarily revealing *some* of their movie preferences publicly but not everything

- For example, a person who was successfully de-anonymized had privately rated movies with predominate gay themes

    - There may be settings the user does not want this information to be public

Implications to anonymized data

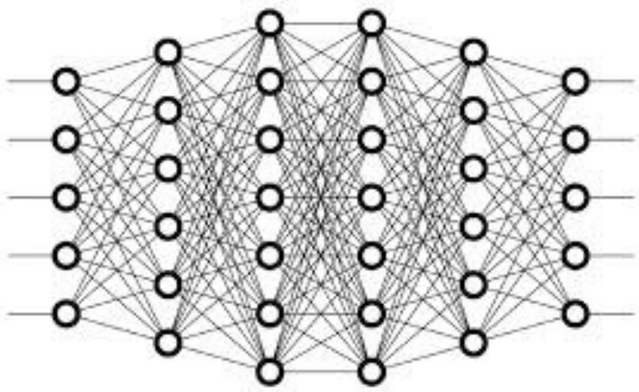
- It doesn’t take much to de-anonymize data!
  - Naive anonymization mechanisms don’t work
- Once a user has been de-anonymized, the user can’t disclose non-trivial information even via a pseudonym without being linked back to their real identity
  - Once any piece of data has been linked to the *real* identity, any association between that data and a pseudonym breaks anonymity
  - Notion known as **forward secrecy**

Solution 3: Machine Learning?

- What if we use the data to train a neural network and only give people the neural network?

- Unfortunately, there's an entire line of research attacking these models!

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No



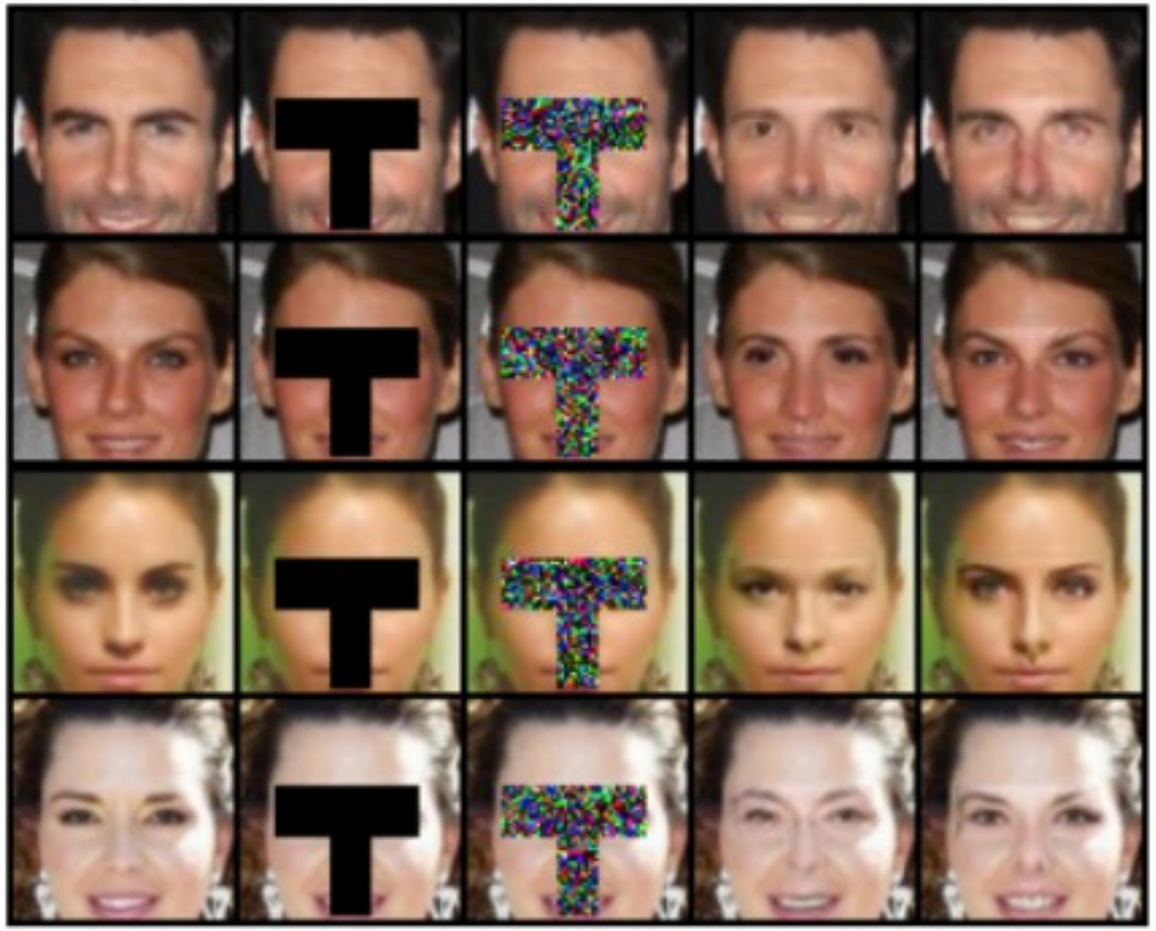
Solution 3: Machine Learning?

- Even if you don’t release the raw data, access to neural networks can reveal information about the training data!

- **Model inversion** attacks “aim to reconstruct sensitive features of training data by taking advantage of their correlation with the model output”

- **Training data extraction** attacks recover individual training data

Model Inversion Attack



https://arxiv.org/pdf/1911.07135

Training Data Extraction Attack

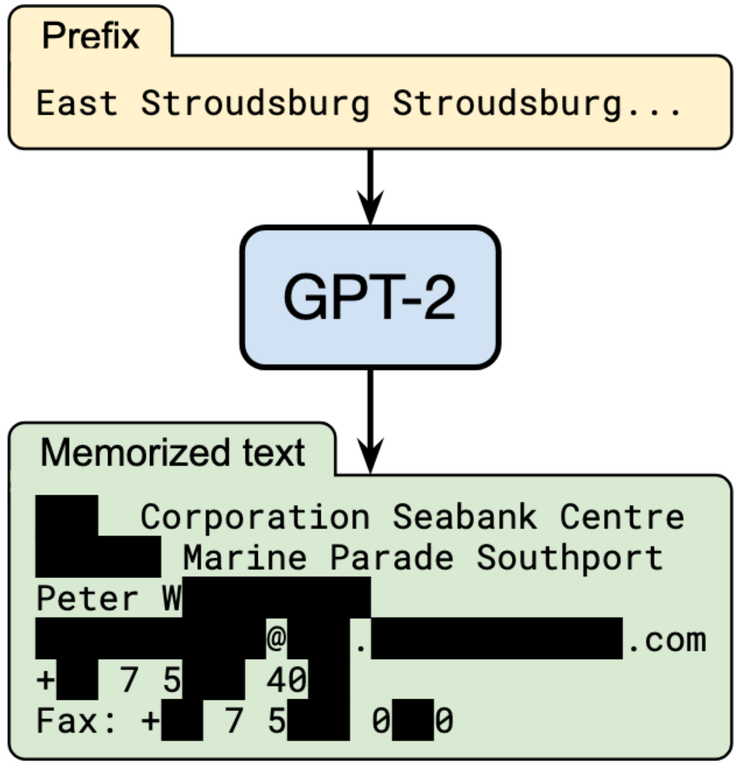


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

https://arxiv.org/pdf/2012.07805

Solution 4: Analyst queries the hospital

- Seems reasonable for broad queries, right?

- Unfortunately, also not good! Analyst can make “broad” queries that when taken together can reveal information about a specific individual

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

Solution: ???

Name	Blood Pressure	Lung Disease
Abby	125	Yes
Brian	140	Yes
Cary	145	No

How do you define data privacy?

- This is a hard question!

- The people who worked on this definition ultimately won the Godel prize and influenced how everyone thinks about data privacy

What's our goal?

- Intuitively:
  - The analyst shouldn't learn too much about any *individual* person
  - The analyst should still be able to do meaningful analysis about *global* trends
- Putting these together, *whether or not you're in the dataset should produce approximately the same output*
- This is what **differential privacy** aims to do

Differential Privacy

- **Differential privacy** gives mathematical definitions for what makes a mechanism private

- Informally: Whether or not you're in the dataset produces *approximately* the same output
  - "Approximately" is now a parameter
  - Better privacy = less accurate result
  - Can quantify privacy loss

- Mechanisms usually compute the answer to the analyst's query and add noise to hide individual data but preserve functionality

Example: Randomized Response

- Let’s say I want to know what fraction of the population has done drugs

- Maybe you have, but you don’t necessarily want to reveal it in case it gets out and affects your professional life

- Luckily, I don’t actually care whether or not you specifically have done drugs! I’m just curious what fraction of the overall population has done drugs!

Example: Randomized Response

- Steps for participating in the survey:
  - Flip a coin
  - If the coin came up heads, answer truthfully to the yes/no question.
  - If the coin came up tails, flip another coin. If the second coin is heads, answer “YES”. Otherwise, answer “NO”.

Example: Random

- Steps for participating in the survey:
  - Flip a coin
  - If the coin came up heads, answer truthfully to the yes/no question.
  - If the coin came up tails, flip another coin. If the second coin is heads, answer “YES”. Otherwise, answer “NO”.

Why does this protect privacy?

Suppose you answered “YES”.
Does that mean you’ve done drugs?

Example: Random

- Steps for participating in the survey:
  - Flip a coin
  - If the coin came up heads, answer truthfully to the yes/no question.
  - If the coin came up tails, flip another coin. If the second coin is heads, answer “YES”. Otherwise, answer “NO”.

Why does this protect privacy?

Suppose you answered “YES”.
Does that mean you’ve done drugs?

How does the randomized response help?

- Two cases:
  - If you've done drugs, you answer YES 75% of the time
  - If you haven't done drugs, you answer YES 25% of the time
- Suppose $x\,\%$ respondents have done drugs and $(100-x)\,\%$ haven't
- If they follow the survey correctly, then the number of YES answers should be $0.75x + 0.25(100-x) = 25\% + 0.5x\,\%$
- If I see $30\,\%$ of people say YES, then I can calculate that roughly $10\,\%$ of people have done drugs

Why settle for noisy data?

- It’s a philosophical question, but actually all data is noisy
- All learning algorithms make mistakes
- The best we can ever hope for is approximation!
- The amount of noise you need to add to get differential privacy is smaller than the sampling error!
    - Ultimately you don’t lose anything by being private

Examples of Differential Privacy Used Today

- **Apple**: QuickType suggestions, emoji suggestions https://images.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
- **Google**: Most popular websites when launching Chrome https://venturebeat.com/ai/following-apple-google-tests-differential-privacy-in-gboard-for-android/
- **US Census**: Redistricting data https://www.census.gov/newsroom/blogs/director/2021/07/redistricting-data.html
- **Wikimedia**: Geolocation data of contributions/contributors https://foundation.wikimedia.org/wiki/Legal:Country_and_Territory_Protection_List

Does differential privacy solve everything?

- Unfortunately not!

- **Sensitivity**: Some queries can be significantly affected by whether or not the dataset is changed

    - Example: “What’s the max age of an undergraduate at Barnard?” The presence of someone who is significantly older than everyone else affects the data!

    - Some targeted questions cannot be answered in a privacy-preserving way

- **Composition**: Multiple queries or multiple mechanisms compounds the noise or privacy lose

What does this mean for us?

- Privacy is your right to limit access to your self
  - Preserves autonomy, allows for self evaluation
- If you are someone who collects data, be careful when publishing the raw data
  - Just removing PII is not sufficient for maintaining privacy
  - Consider using differential privacy
- Only takes a small named database to deanonymize a larger “anonymous” database

This is the end

- I hope you all had a good semester!
- Good luck with your final projects :)
- Office hours all this week
  - Murad’s office hours today are remote