BU CS599 S1
Foundations of Private Data Analysis
Spring 2025

Lecture 01: Introduction

Adam Smith

BU

Today

- Course Intro
- A taste of the syllabus
  - Attacks on information computed from private data
  - A first private algorithm: randomized response

This Course

- Intro to research on privacy in ML and statistics
  - Mathematical models
    - How do we formulate nebulous concepts?
    - How do we assess and critique these formulations?
  - Algorithmic techniques
- Skill sets you will work on
  - Theoretical analysis
  - Critical reading of research literature in CS and beyond
  - Programming
- Prerequisites
  - Comfort writing proofs about probability, linear algebra, algorithms
  - Programming (in Python)
  - MS/undergrads: discuss your background with instructor.

Administrivia

- Web page: https://dpcourse.github.io/2025-spring/
  - Communication via Piazza
  - Course work on Gradescope
- Your jobs
  - Lecture preparation, attendance, participation
  - Homework
  - Project

Coursework

- Lecture prep and in-class work
- Homework
  - Due Fridays every 2 weeks
  - Limited collaboration is permitted
    - Groups of size ≤ 4
  - Academic honesty: You must
    - **Acknowledge collaborators** (or write "collaborators: none")
    - **Write your solutions yourself**, and be ready to explain them orally
      - Rule of thumb: walk away from collaboration meetings with no notes.
    - Use only course materials (except for general background, e.g., on probability, calculus, etc)
    - Any use of outside tools (e.g. GPT, Mathematica, experiments in Python) to help answer questions should be documented
- Project (details TBA)
    - Read and summarize a set of 2-3 related papers
    - Identify open questions
    - Develop new material (application of a technique to a new data set, work on open question, show some assumption is necessary, ...)
    - Presentation in last week of class

Theory v Practice

- This is a theory course on a topic of current relevance
- There will be programming assignments, as well as reading on recent developments
- You have lots of flexibility in the course project to pursue which ever direction you find most compelling

For flipped classroom lectures

- Ahead of time
  - Watch video
    - Engage actively and take notes by hand as you watch
  - Read lecture notes
  - Answer Gradescope pre-class questions
- In class
  - Be present
    - Let us know on Piazza if that is an issue in general or for specific lectures. Default is attendance at every class
  - Actively participate in problem-solving
    - Problems will be posted ahead of time
  - Take notes on your work
- After class
  - Submit your notes (photo or electronic) on Gradescope

For traditional lectures

- In class
  - Be present
    - Let us know on Piazza if that is an issue in general or for specific lectures. Default is attendance at every class
  - Bring questions
  - Actively participate in problem-solving and feedback questions
- After class
  - Work on the homework!

To do list for this week

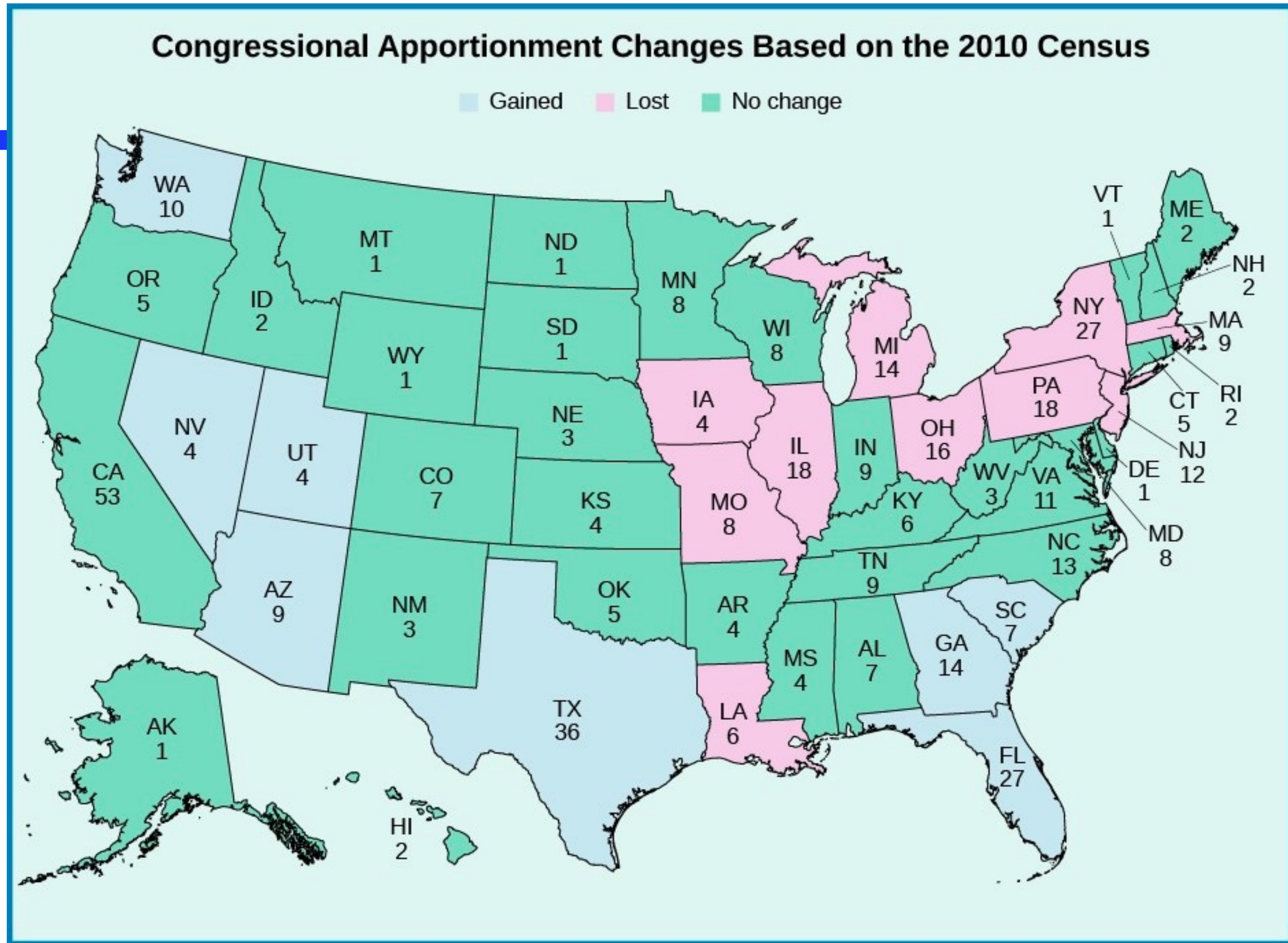
- Make sure you have access to Piazza, Gradescope
- Read the syllabus
- By Tuesday:
  - Fill background survey (see Piazza)
  - Watch videos, read notes, answer questions for Lecture 2

Today

- Course Intro
- A taste of the syllabus
  - Attacks on information computed from private data
  - A first private algorithm: randomized response

Data are everywhere

- Decisions increasingly automated using rules based on personal data



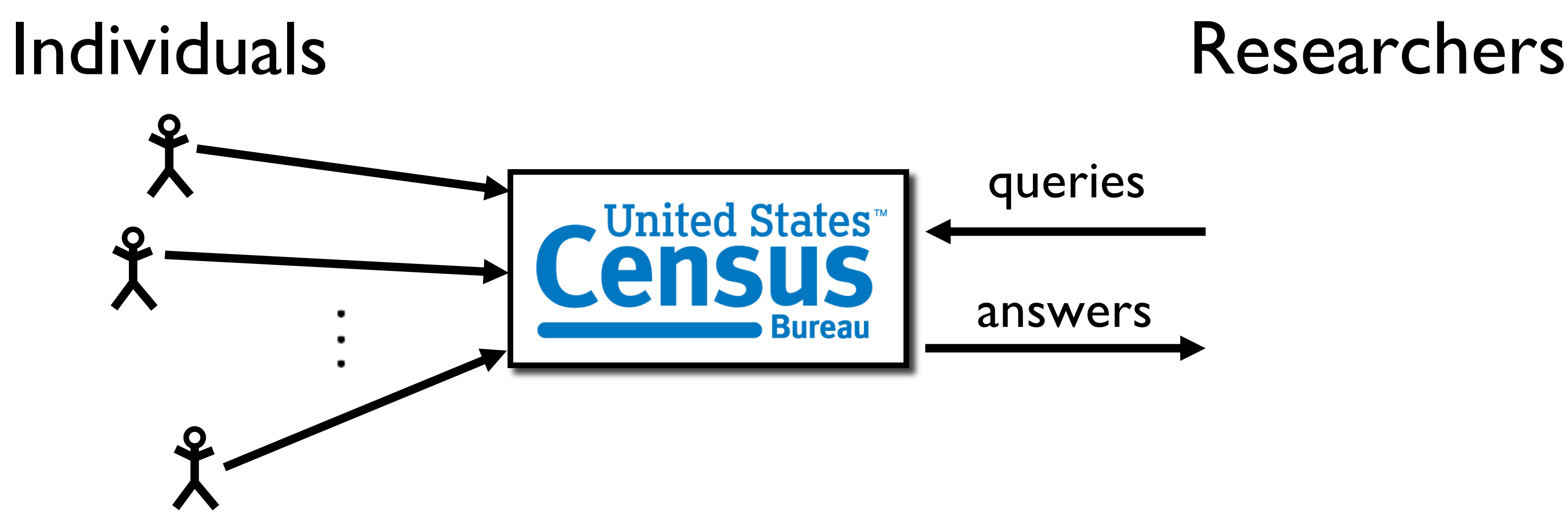
- Census data used to apportion congressional seats
  - ➢ Think about citizenship question
- Also enforce Voting Rights Act, allocate Title I funds, design state districts, …

Image © Lumen Learning

Machine learning on your devices

- Statistical models trained using data from your phones
  - ➢ Offer sentence completions
  - ➢ Convert voice to speech
  - ➢ Select, for you and others,
    - Content (e.g. newsfeed/scroll)
    - Ads
    - Recommendations for products (“You might also like…”)
- Statistical models trained from other personal data…
  - ➢ Advise judges’ bail decisions
  - ➢ Allocate police resources
  - ➢ Advise doctors on diagnosis/treatment

Privacy in Statistical Databases



Large collections of personal information

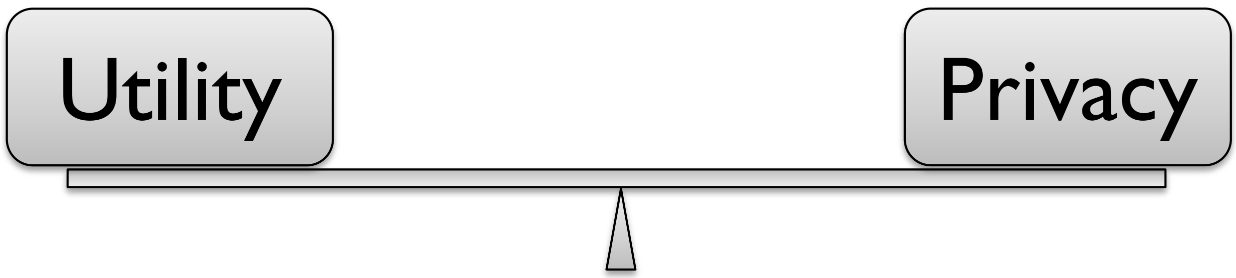
- census data
- medical/public health
- social networks
- education

Statistical analysis benefits society

Valuable because they reveal so much about our lives

Two conflicting goals

- Utility: release aggregate statistics
- Privacy: individual information stays hidden



How do we define "**privacy**"?

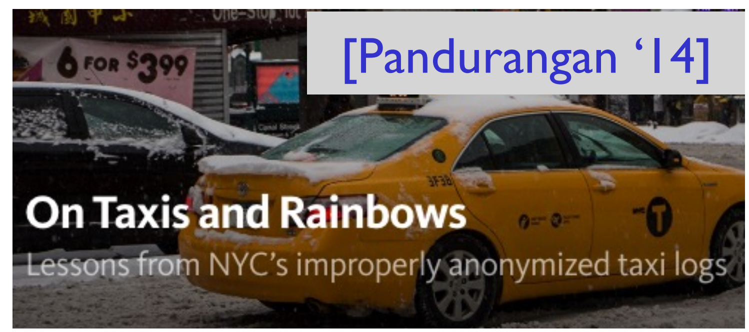
- Studied since 1960's in
  - ➢ Statistics
  - ➢ Databases & data mining
  - ➢ Cryptography
- This course section: Rigorous foundations and analysis

First attempt: Remove obvious identifiers

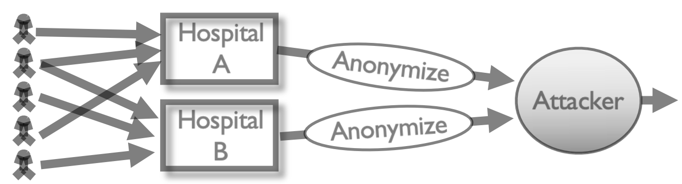
"AI recognizes blurred faces"
[McPherson Shokri Shmatikov '16]

[Gymrek McGuire Golan Halperin Erlich '13]

[Pandurangan '14]



- Everything is an identifier
- Attacker has external information
- "Anonymization" schemes are regularly broken

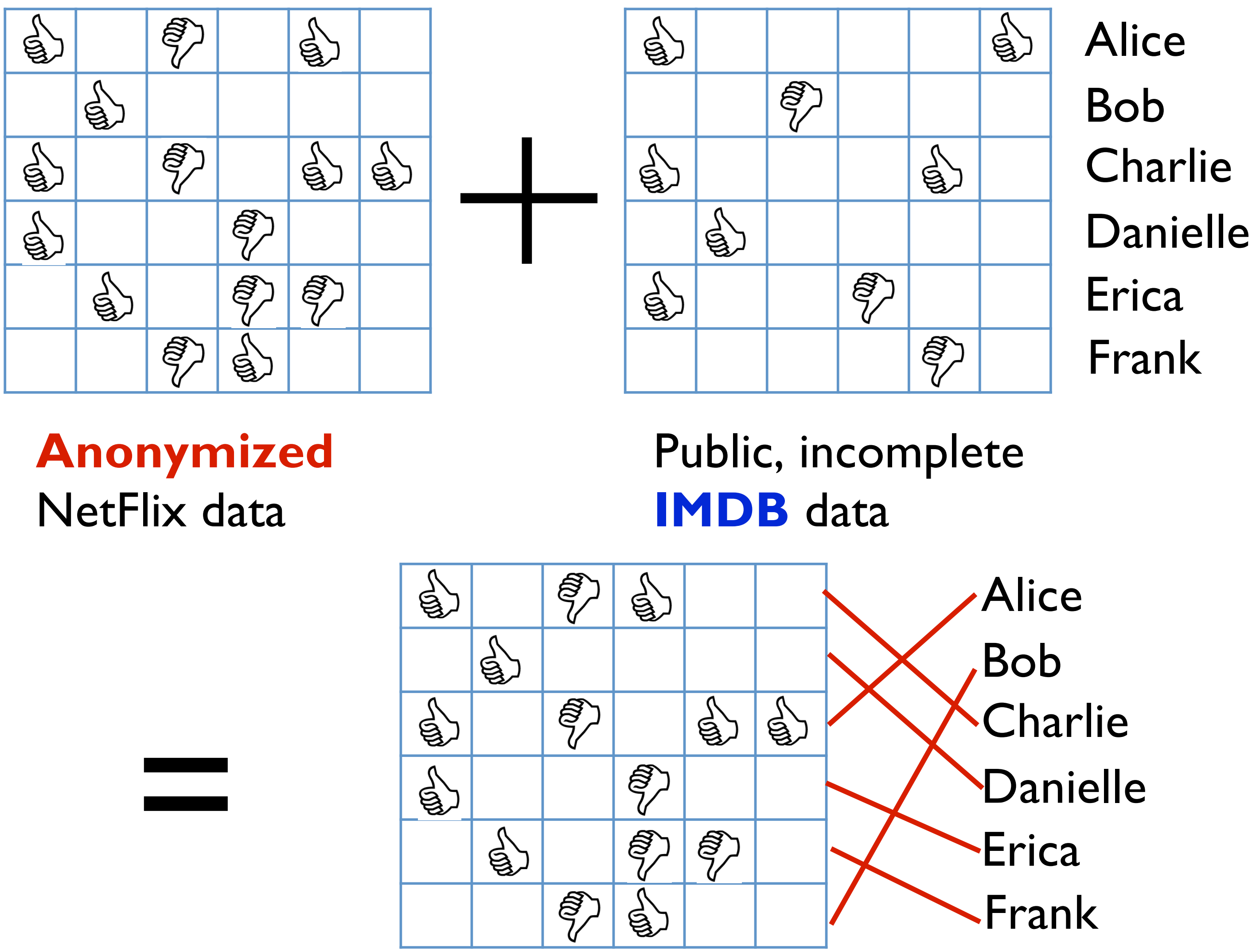


[Ganta Kasiviswanathan S '08]

Images: whitehouse.gov, genesandhealth.org, medium.com

Reidentification attack example

[Narayanan, Shmatikov 2008]



Identified NetFlix Data

On average, four movies uniquely identify user

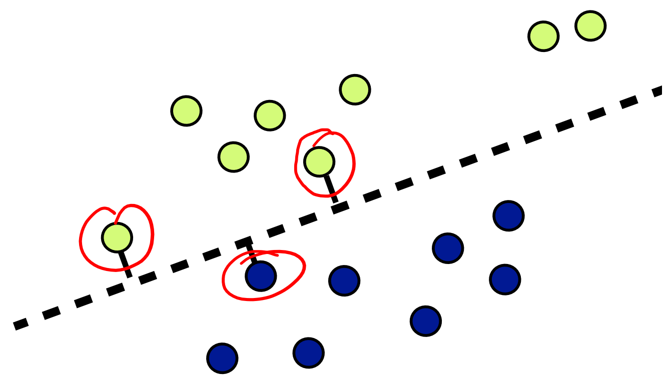
Image credit: Arvind Narayanan

Is the problem granularity?

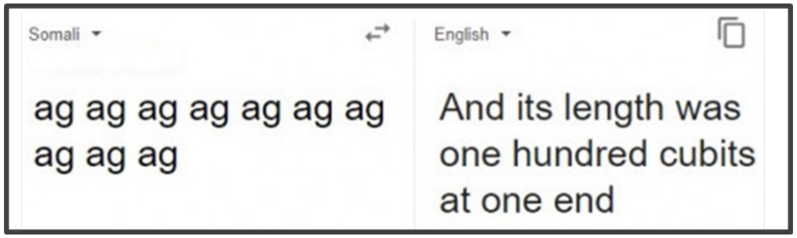
What if we only release aggregate information?

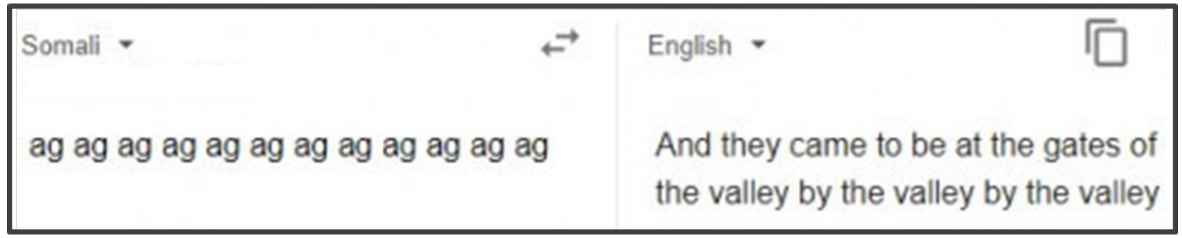
Problem 1: Models leak information

- Support vector machine output reveals individual data points
- Deep learning models reveal even more



Models Leak Information

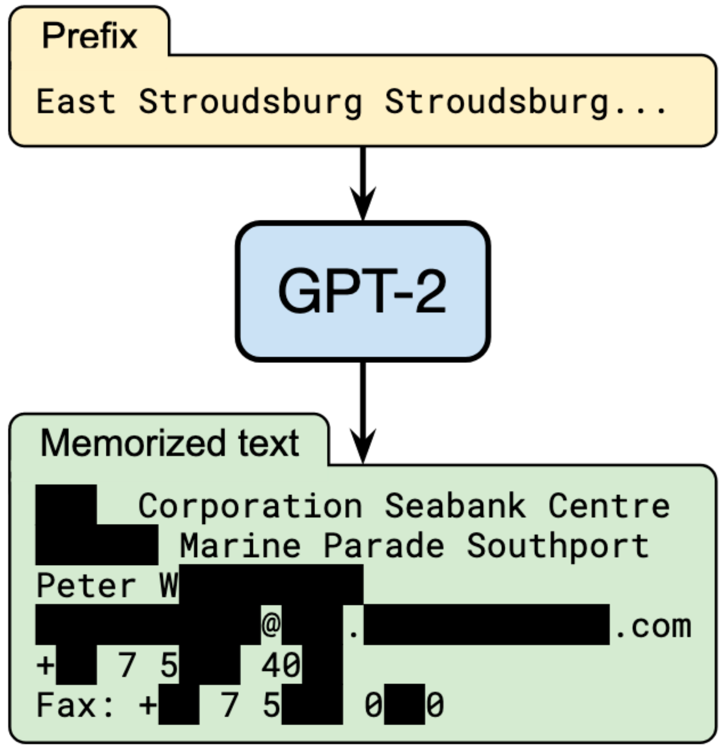




Models can leak information about training data
in unexpected ways

- Example: Smart Compose in Gmail
  - ➢ Haven't seen you in a while.
    Hope you are doing well
  - ➢ John Doe's SSN is 920-24-1930
    [Carlini et al. 2018]
- Modern deep learning algorithms often "memorize" inputs

Sources: Top 2- https://www.vice.com/en_us/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies (inspiration: Ilya Mironov)



[Carlini et al. 20]

Current language models memorize irrelevant information.

Is the problem granularity?

What if we only release aggregate information?

Problem 1: Models leak information

Problem 2: Statistics together may encode data

- Example: Average salary before/after resignation
- More generally:

Too many, "too accurate" statistics
reveal individual information

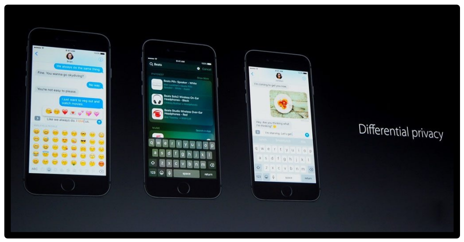
- ➢ Reconstruction attacks
  - Reconstruct all or part of data
- ➢ Membership attacks
  - Determine if a target individual is in (part of) the data set

Cannot release everything
everyone would want to know

Differential privacy

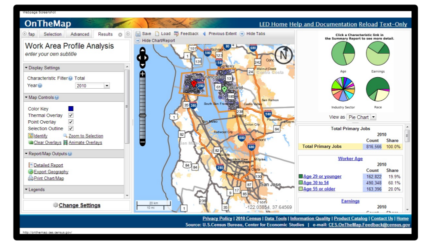
Differential Privacy

- Robust notion of "privacy" for algorithmic outputs
  - ➢ Meaningful in the presence of arbitrary side information
- Several current deployments



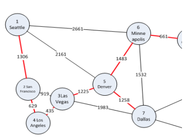
Apple

Google



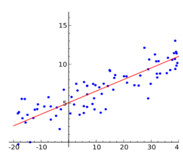
US Census

- Burgeoning field of research



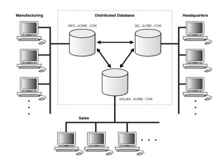
Algorithms

Crypto, security



Statistics, learning

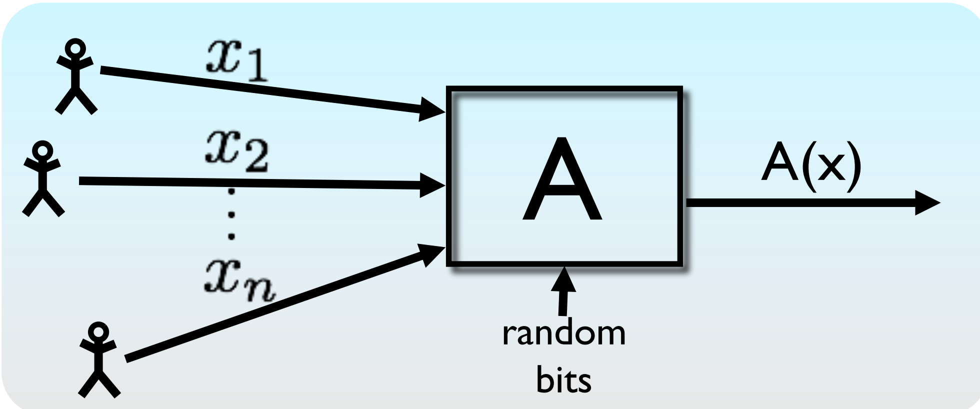
Game theory, economics



Databases, programming languages

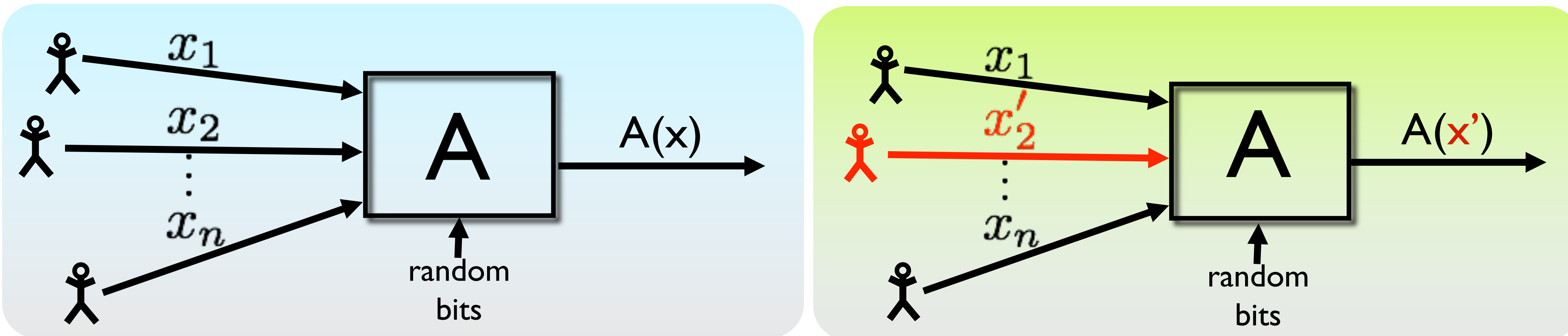
Law, policy

Differential Privacy

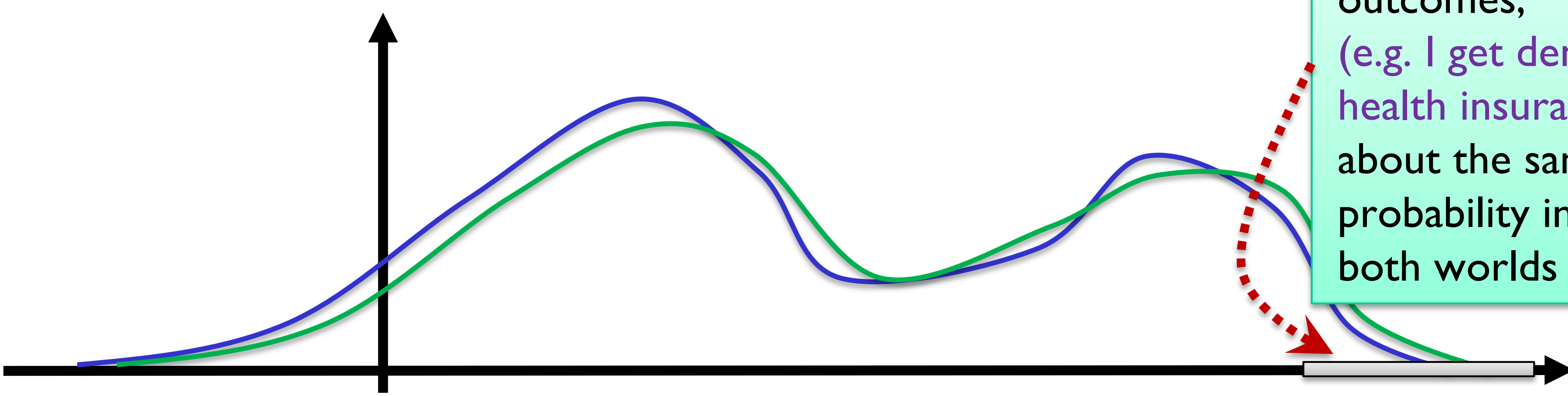


- Data set $x = (x_1, \ldots, x_n) \in \mathcal{X}$
  - Domain $\mathcal{X}$ can be numbers, categories, tax forms
  - Think of x as **fixed** (not random)
- A = **probabilistic** procedure
  - $A(x)$ is a random variable
  - Randomness might come from adding noise, resampling, etc.

Differential Privacy



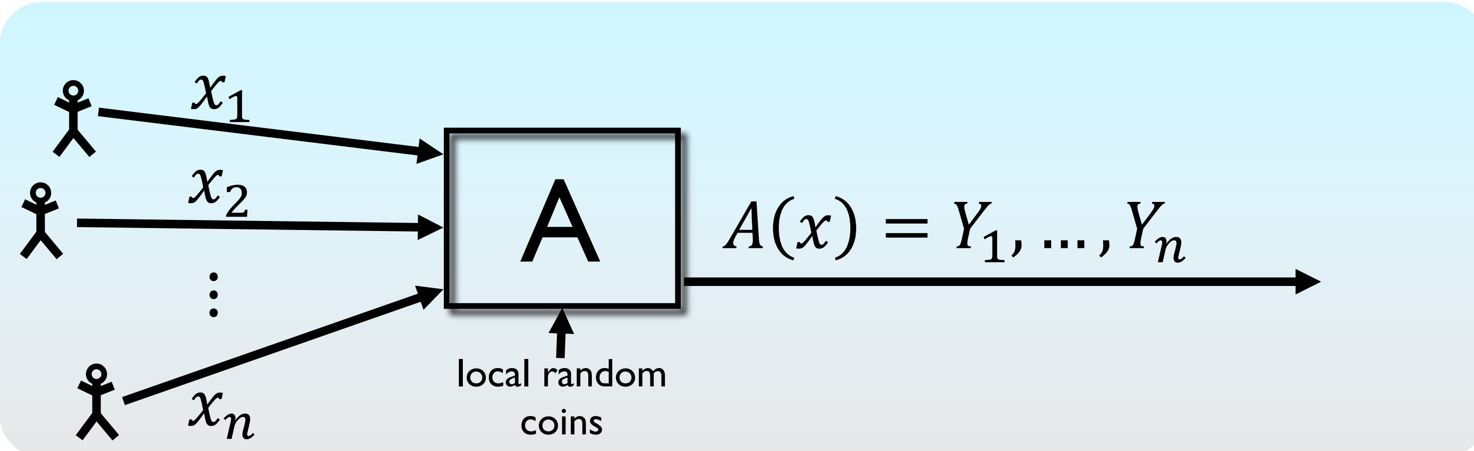
- A thought experiment
  - Change one person's data (or add or remove them)
  - Will the **probabilities of outcomes** change?



For any set of outcomes, (e.g. I get denied health insurance) about the same probability in both worlds

A First Algorithm: Randomized Response

Randomized Response (Warner 1965)



- Say we want to release the proportion of diabetics in a data set
  - Each person's data is a bit: $x_i = 0$ or $x_i = 1$
- Randomized response: each individual rolls a die
  - 1, 2, 3 or 4: Report true value x_i
  - 5 or 6: Report opposite value $1 - x_i$
- Output is list of reported values $Y_1, \dots, Y_n$
  - It turns out that we can estimate fraction of x_i's that are 1 when n is large

Randomized Response

i	x_i	Die roll	Y_i
1	**0**	5	**yes**
2	**1**	1	**yes**
3	**1**	3	**yes**
4	**1**	2	**yes**
5	**0**	6	**yes**
6	**0**	4	**no**
7	**1**	2	**yes**
8	**0**	3	**no**
9	**1**	2	**yes**
10	**1**	5	**no**

10	**0**	3	**no**

What sort of privacy does this provide?

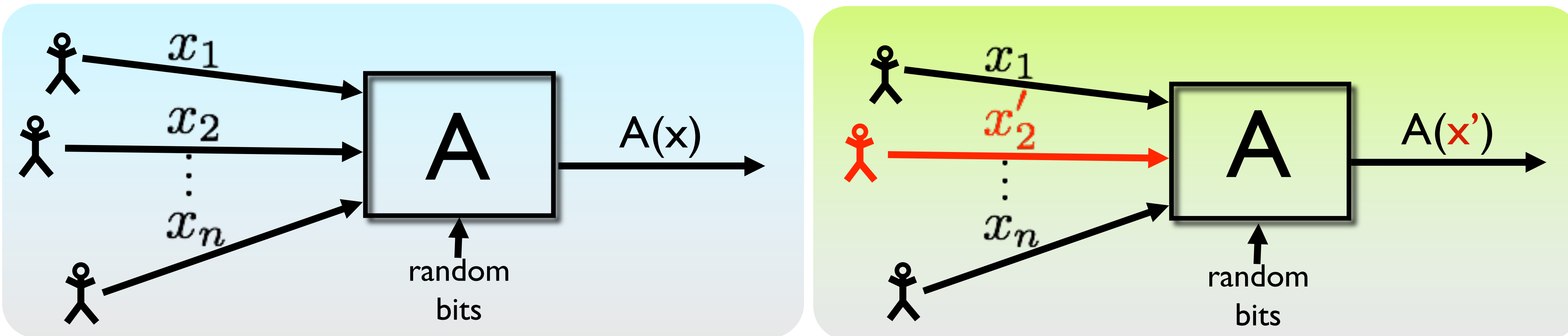
- Many possible answers

One approach:
Plausible deniability

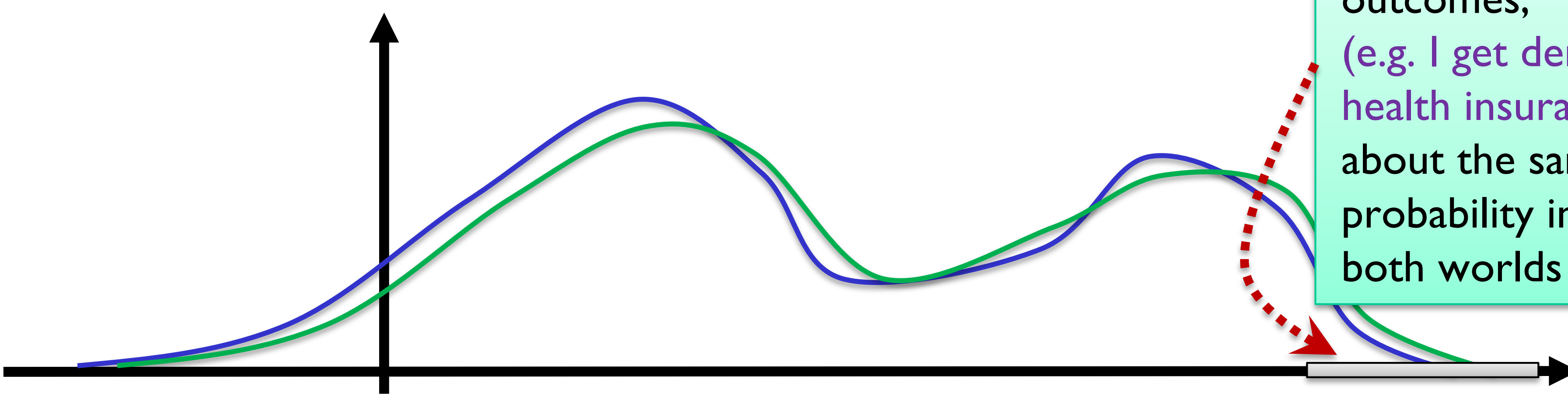
- x_{10} could have been 0
- x_8 could have been 1

- Suppose we fix everyone else's data $x_1, \ldots, x_9 \ldots$
- What is $$\frac{\Pr(Y_{10} = no | x_{10} = 1)}{\Pr(Y_{10} = no | x_{10} = 0)}?$$

Differential Privacy



- A thought experiment
  - Change one person's data (or add or remove them)
  - Will the **probabilities of outcomes** change?



For any set of outcomes, (e.g. I get denied health insurance) about the same probability in both worlds

Plausible deniability and RR

A bit more generally…

- Fix any data set $\vec{x} \in \{0,1\}^n$, and any neighboring data set $\vec{x}'$
  - Let i be the position where $x_i \neq x_i'$
  - (Recall $x_j = x_j'$ for all $j \neq i$)
- Fix an output $\vec{a} \in \{0,1\}^n$

$$\Pr(A(\vec{x}) = \vec{a}) = \left(\frac{2}{3}\right)^{\#\{j: x_j = a_j\}} \left(\frac{1}{3}\right)^{\#\{j: x_j \neq a_j\}}$$

(because decisions made independently)

- When we change one output, one term in the product changes (from $\frac{2}{3}$ to $\frac{1}{3}$ or vice versa)
- So $\frac{\Pr(A(\vec{x})=\vec{a})}{\Pr(A(\vec{x}')=\vec{a})} \in \left\{\frac{1}{2}, 2\right\}$.

For specific outputs to sets of outcomes

- Now consider a set of outputs $S \subseteq \{0,1\}^n$
  - Examples
    - $S = \{\vec{a}\colon \text{ fraction of } 1's \text{ in } \vec{a} \text{ is in } [0.8,0.9]\}$, or
    - $S = \{\vec{a}\colon \quad a_{37} = 1\}$
- For every data set $\vec{x}$, $\Pr\left(\vec{Y} \in S \middle| \vec{x}\right) = \sum_{\vec{a} \in S} \Pr(\vec{Y} = \vec{a} | \vec{x})$
  Therefore, for every pair of neighbors $\vec{x}, \vec{x}'$:

$$\frac{\Pr\left(\vec{Y} \in S \middle| \vec{x}\right)}{\Pr\left(\vec{Y} \in S \middle| \vec{x}'\right)} = \frac{\sum_{\vec{a} \in S} \Pr\left(\vec{Y} = \vec{a} \middle| \vec{x}\right)}{\sum_{\vec{a} \in S} \Pr\left(\vec{Y} = \vec{a} \middle| \vec{x}'\right)} \le \frac{\sum_{\vec{a} \in S} 2 \Pr\left(\vec{Y} = \vec{a} \middle| \vec{x}'\right)}{\sum_{\vec{a} \in S} \Pr\left(\vec{Y} = \vec{a} \middle| \vec{x}'\right)} = 2$$

- Similarly, $\frac{\Pr\left(\vec{Y} \in S \middle| \vec{x}\right)}{\Pr\left(\vec{Y} \in S \middle| \vec{x}'\right)} \ge \frac{1}{2}$

Great! We have proved that

- for every set of possible outcomes,
  the probability of that set can go up or down by a factor of at most 2
  when any one person's data is changed.
- This means the randomized response algorithm is $\ln(2)$-differentially private. We'll learn more in a few lectures.