

BU CS591 S1
NEU CS 7880

Foundations of Private Data Analysis
Spring 2021

Lecture 01: Introduction

Jonathan Ullman

NEU

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
 - Undergrads: discuss your background with instructor.

Administrivia

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

Every lecture

- 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 with camera on
 - 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

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 reading general background, e.g., on probability, calculus, etc)
- 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

To do list for this week

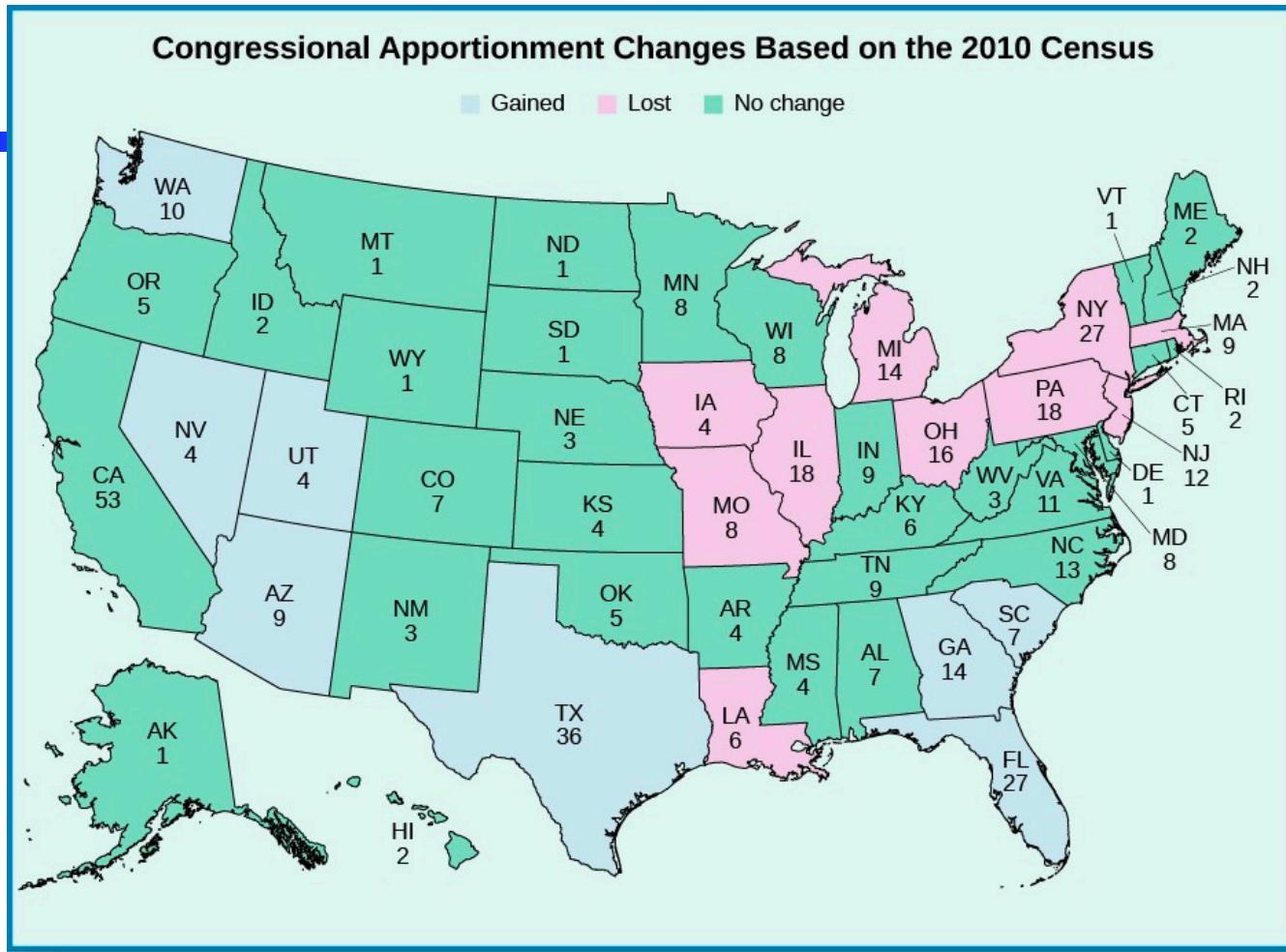
- Make sure you have access to Piazza, Gradescope
- Read the syllabus
- Fill Gradescope background survey
 - By Thursday
- Watch videos, read notes, answer questions for Lecture 2
 - By next lecture (Thu/Fri)

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, ...

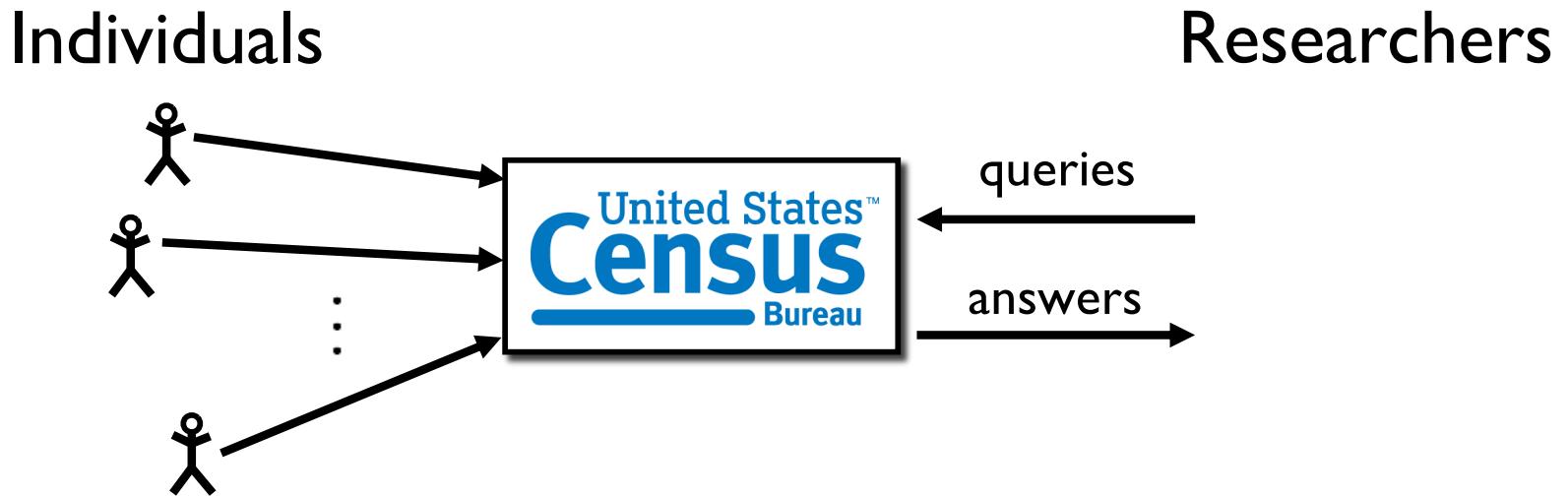
Machine learning on your devices

- Statistical models trained using data from your phones
- Statistical models trained from other personal data

Machine learning on your devices

- Statistical models trained using data from your phones
 - Offer sentence completion
 - Convert voice to speech
 - Select, for you and others to see,
 - Content (e.g. FB newsfeed)
 - 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

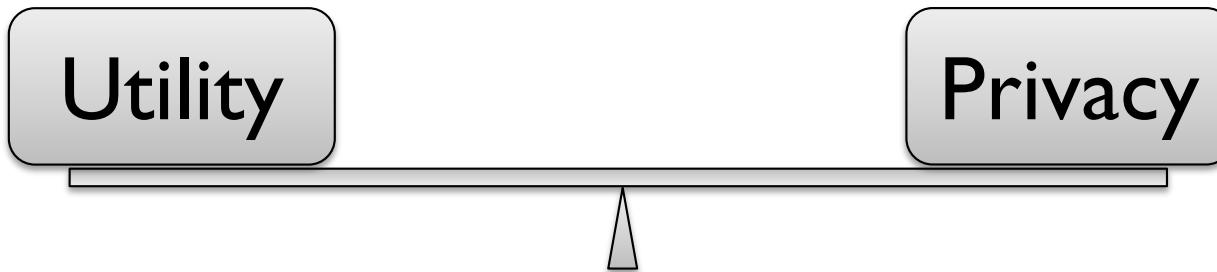
- census data
- medical/public health
- social networks
- education

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



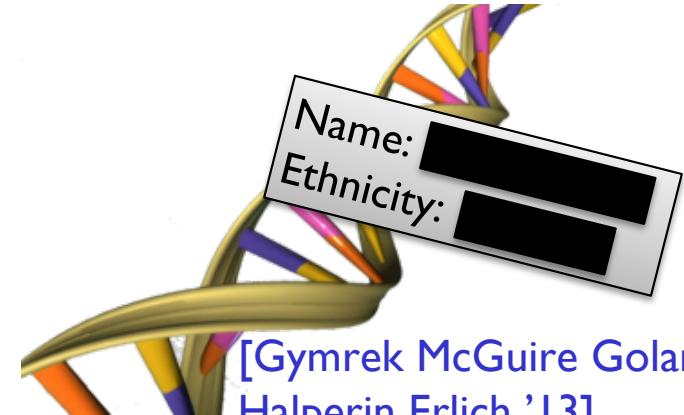
How do we define “**privacy**”?

- 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]

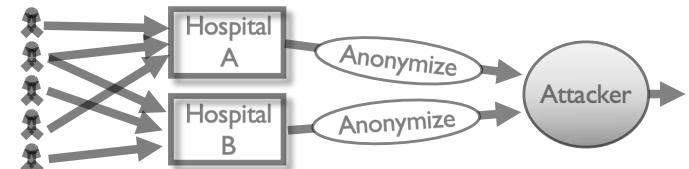


[Gymrek McGuire Golan
Halperin Erlich ’13]



[Pandurangan ’14]

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

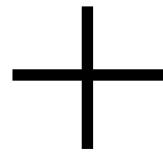


[Ganta Kasiviswanathan S ’08]

Reidentification attack example

[Narayanan, Shmatikov 2008]

👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	👎
		👎	👍		



👍			👍		
	👍				
👍			👍		👍
👍				👎	
👍				👎	
		👎			

Alice
Bob
Charlie
Danielle
Erica
Frank

Anonymized
NetFlix data

Public, incomplete
IMDB data



👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	👎
		👎	👍		

Alice
Bob
Charlie
Danielle
Erica
Frank

Identified NetFlix Data

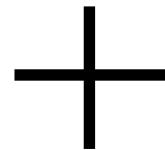
On average,
four movies
uniquely
identify user

Reidentification attack example

[Narayanan, Shmatikov 2008]

+	-	+	-	+
+	-			
+	-		+	+
+		-		
+	-	-	-	
	-	+		

Anonymized
NetFlix data



+	-		+	+
+		-		
+			+	
+			-	
			-	-

Public, incomplete
IMDB data



+	-	+	+	
+	-			
+		-	+	+
+		-		
+	-	-	-	
	-	+		

Alice
Bob
Charlie
Danielle
Erica
Frank

Identified NetFlix Data

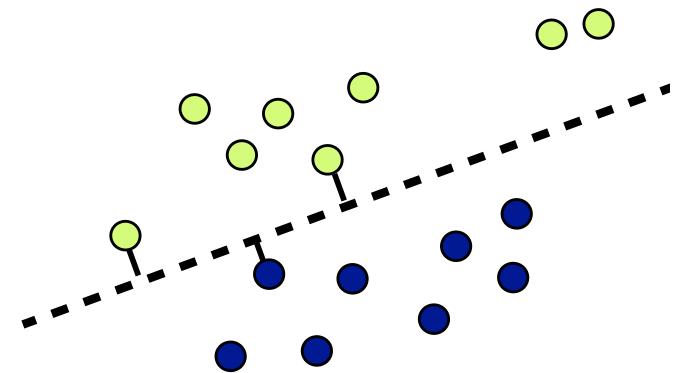
On average,
four movies
uniquely
identify user

Is the problem granularity?

What if we only release **aggregate** information?

Problem I: Models leak information

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



Models Leak Information

Somali	↔	English
ag ag ag ag ag ag ag ag ag	↔	And its length was one hundred cubits at one end
ag	↔	And they came to be at the gates of the valley by the valley by the valley

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

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**

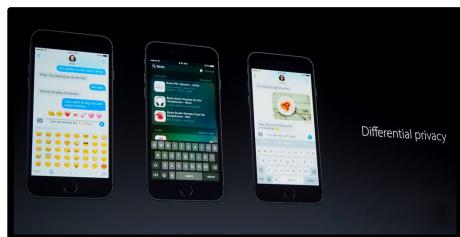
- Reconstruction attacks
 - Reconstruct all or part of data
- Membership attacks
 - Determine if a target individual is in (part of) the data set

Cannot release everything
everyone would want to know

Differential privacy

Differential Privacy

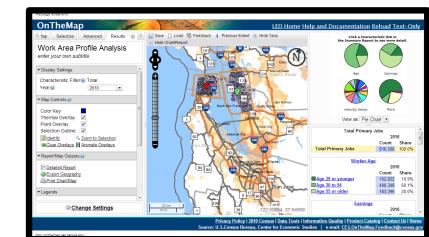
- Robust notion of “privacy” for algorithmic outputs
 - Meaningful in the presence of arbitrary side information
 - Several current deployments



Apple

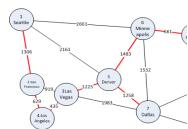


Google



US Census

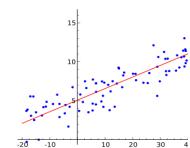
- Burgeoning field of research



Algorithms



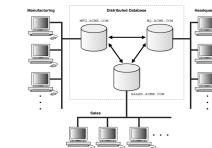
Crypto, security



Statistics, learning



Game theory, economics

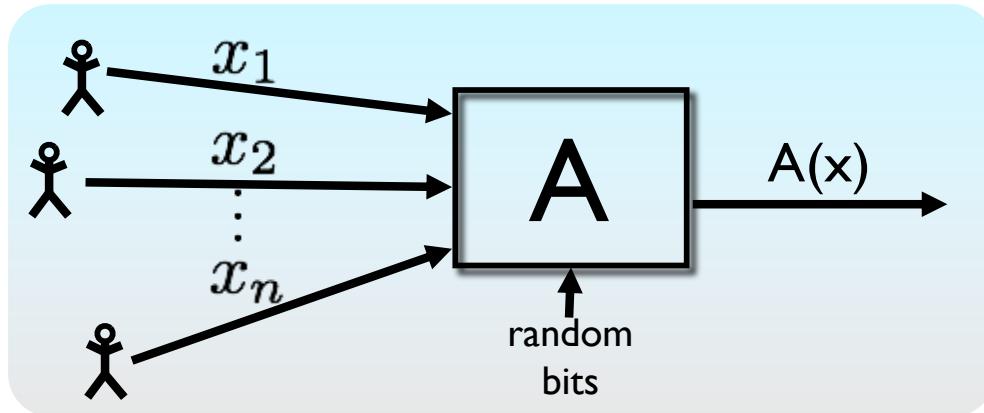


Databases, programming languages



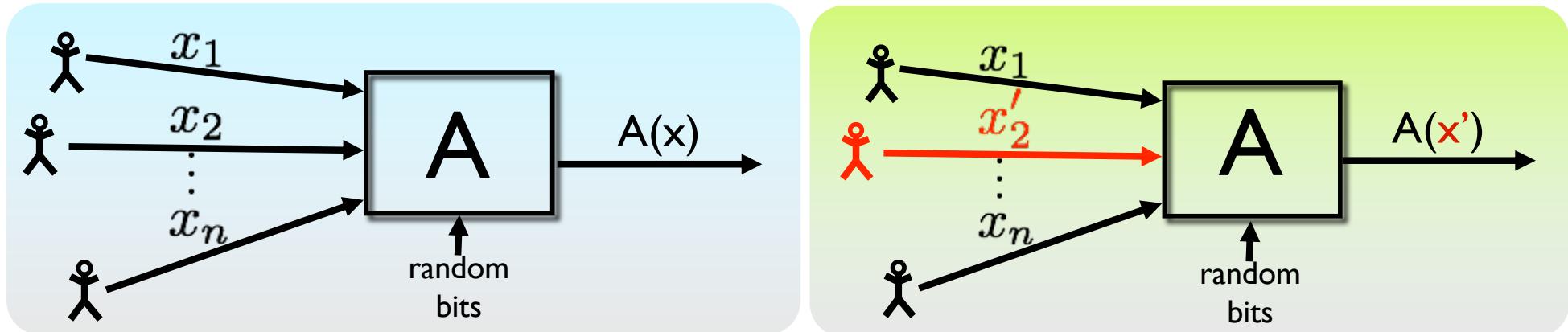
Law, policy

Differential Privacy

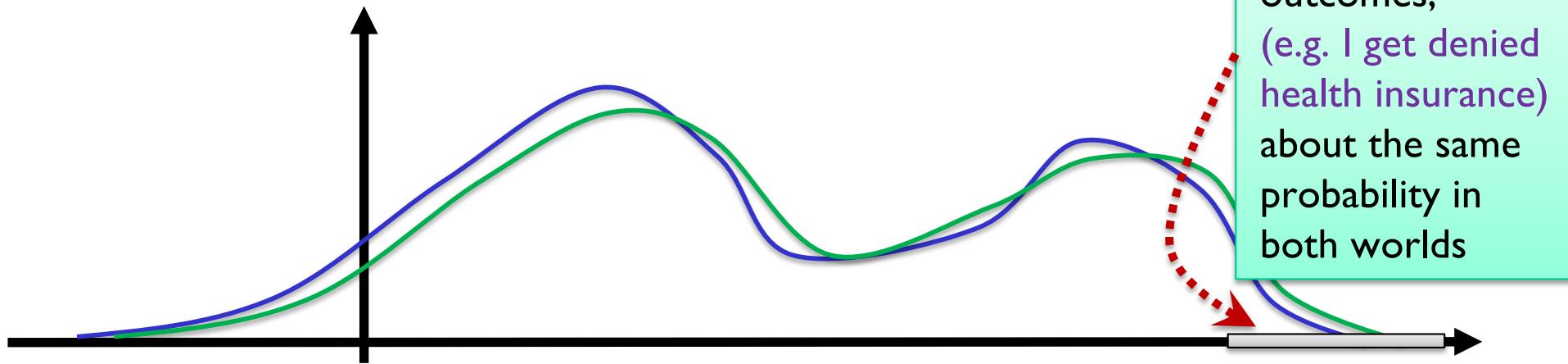


- Data set $x = (x_1, \dots, x_n) \in D^n$
 - Domain D can be numbers, categories, tax forms
 - Think of x as **fixed** (not random)
- **A = probabilistic procedure**
 - $A(x)$ is a random variable
 - Randomness might come from adding noise, resampling, etc.

Differential Privacy

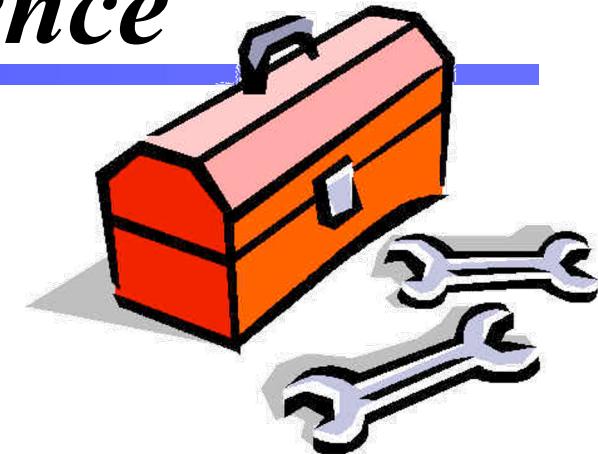


- A thought experiment
 - Change one person's data (or add or remove them)
 - Will the **probabilities of outcomes** change?



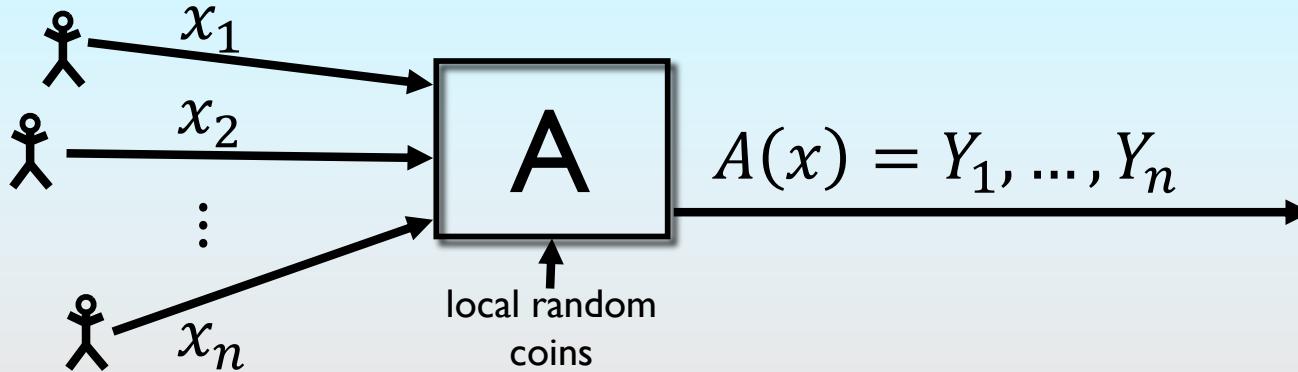
A Broad, Active Field of Science

- Algorithmic tools and techniques
 - Feasibility results: Learning, optimization, synthetic data, statistics
 - Variations on the definition
- Design tools
 - Programming/query languages, logics, evaluation platforms
- Domain-specific algorithms
 - Networking, clinical data, social networks, geographic data, mobile traces ...
- Connections to other areas
 - Law and policy
 - “Adaptive” generalization bounds
 - Game theory



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 1 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

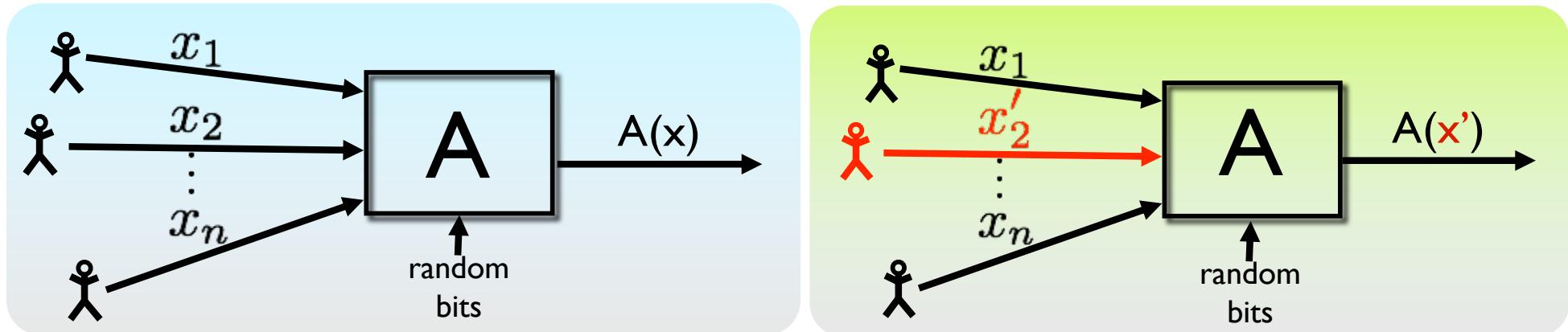
One approach:
Plausible deniability

- x_{10} could have been 0
- x_8 could have been 1
- Suppose we fix everyone else's data $x_1, \dots, x_9 \dots$

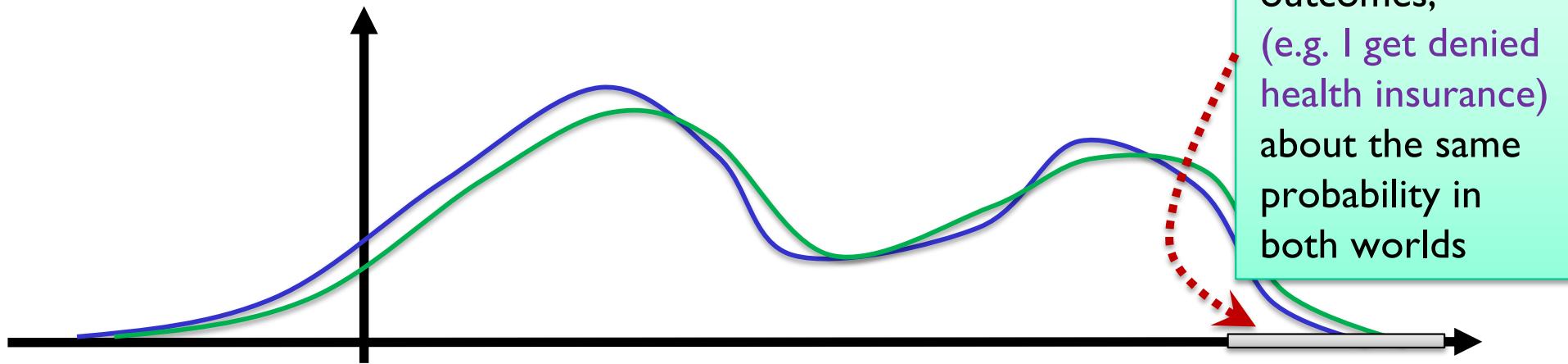
- What is

$$\frac{\Pr(Y_{10} = \text{no} | x_{10} = 1)}{\Pr(Y_{10} = \text{no} | x_{10} = 0)} ?$$

Differential Privacy



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



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 $\frac{\Pr(A(\vec{x})=\vec{a})}{\Pr(A(\vec{x}')=\vec{a})} \in \left\{\frac{1}{3}, \frac{2}{3}\right\}$.

Recall basic probability facts

- Random variables have expectations and variances

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x)$$

$$Var(X) = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)^2\right)$$

- Expectations are linear: For any rv's X_1, \dots, X_n and constants a_1, \dots, a_n :

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i)$$

- Variances add over **independent** random variables. If X_1, \dots, X_n are independent, then

$$Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 Var(X_i)$$

- The **standard deviation** is $\sqrt{Var(X_i)}$

Exercise 1: sums of random variables

- Say X_1, X_2, \dots, X_n are independent with, for all i ,

$$\begin{aligned}\mathbb{E}(X_i) &= \mu \\ \sqrt{\text{Var}(X_i)} &= \sigma\end{aligned}$$

- Then what are the expectation and variance of the average $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$?

- a) $\mathbb{E}(\bar{X}) = \mu n$ and $\sqrt{\text{Var}(\bar{X})} = n\sigma$
- b) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \sigma$
- c) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \sigma/\sqrt{n}$
- d) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{n}$
- e) $\mathbb{E}(\bar{X}) = \mu/n$ and $\sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{n}$

Exercise 2: Estimating $\sum_i x_i$ from RR

- Show there is a procedure which, given Y_1, \dots, Y_n , produces an estimate A such that

$$\sqrt{\mathbb{E} \left(A - \sum_{i=1}^n x_i \right)^2} = O(\sqrt{n}).$$

Standard deviation
of estimate

Equivalently, $\sqrt{\mathbb{E} \left(\frac{A}{n} - \bar{X} \right)^2} = O\left(\frac{1}{\sqrt{n}}\right)$

➤ Hint: What are the mean and variance of $3Y_i - 1$?

Randomized response for other ratios

- Each person has data $x_i \in \mathcal{X}$
 - Normally data is more complicated than bits
 - Tax records, medical records, Instagram profiles, etc
 - Use \mathcal{X} to denote the set of possible records
- Analyst wants to know sum of φ : $\mathcal{X} \rightarrow \{0,1\}$ over x
 - Here φ captures the property we want to sum
 - E.g. “what is the number of diabetics?”
 - $\varphi((Adam, 168 \text{ lbs.}, 17, \text{not diabetic})) = 0$
 - $\varphi((Ada, 142 \text{ lbs.}, 47, \text{diabetic})) = 1$
 - We want to learn $\sum_{i=1}^n \varphi(x_i)$



- Randomization operator takes $z \in \{0,1\}$:

$$R(z) = \begin{cases} z & w.p. \frac{e^\epsilon}{e^\epsilon + 1} \\ 1 - z & w.p. \frac{1}{e^\epsilon + 1} \end{cases}$$

Ratio is e^ϵ (think $1 + \epsilon$ for small ϵ)

For each person i ,
 $Y_i = R(\varphi(x_i))$

Randomized response for other ratios

- Each person has data $x_i \in \mathcal{X}$
 - Analyst wants to know sum of $\varphi: \mathcal{X} \rightarrow \{0,1\}$ over x
- Randomization operator takes $z \in \{0,1\}$:

$$R(z) = \begin{cases} z & w.p. \frac{e^\epsilon}{e^\epsilon + 1} \\ 1 - z & w.p. \frac{1}{e^\epsilon + 1} \end{cases}$$

- How can we estimate a proportion?

➤ $A(x_1, \dots, x_n)$:

- For each i , let $Y_i = R(\varphi(x_i))$
- Return $A = \sum_i (aY_i - b)$

➤ What values for a, b make $\mathbb{E}(A) = \sum_i \varphi(x_i)$?



We can do much better than this!
Coming up ...

- **Proposition:** $\sqrt{\mathbb{E}(A - \sum_i \varphi(x_i))^2} = \frac{e^\epsilon + 1}{e^\epsilon - 1} \sqrt{n}.$ ≈ $\frac{2\sqrt{n}}{\epsilon}$ when ϵ small