DIFFERENTIAL PRIVACY

THE TUTORIAL

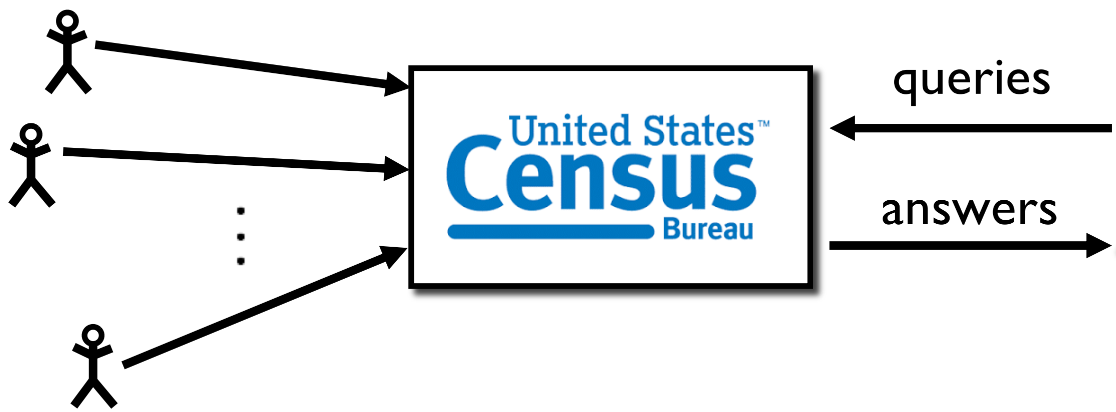
Adam Smith
Boston University
North American IT Summer School
July 2, 2019

BOSTON UNIVERSITY

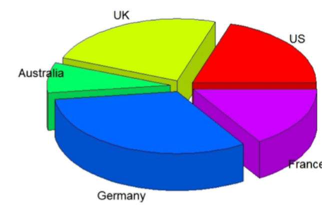
Statistical Data Privacy

Individuals

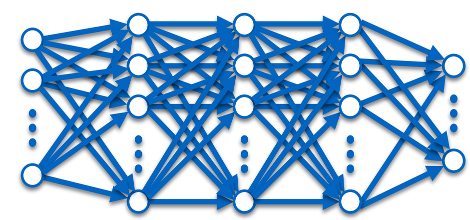
Researchers



Summaries



Complex models



Synthetic data

	Name	Birth Date	Country	State
1	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
2	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
3	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
4	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
5	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
6	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
7	Bill Smith	4-Apr-1956	United States	Texas
8	Bill Smith	4-Apr-1956	United States	Texas
9	Bill Smith	4-Apr-1956	United States	Texas
10	Bill Smith	4-Apr-1956	United States	Texas

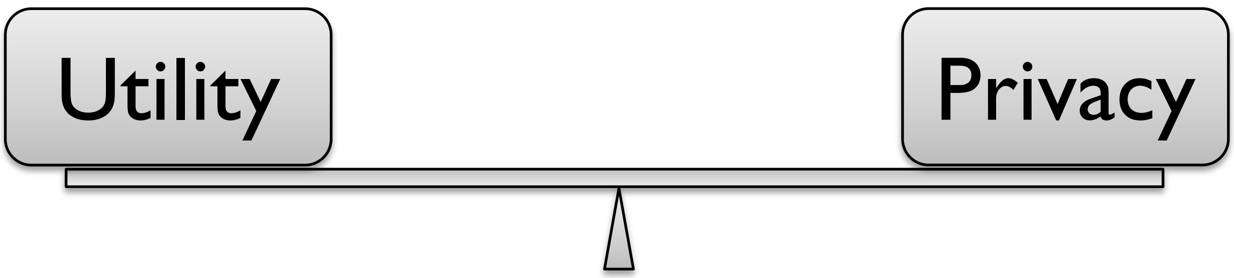
…

Large collections of personal information

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

Two conflicting goals

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

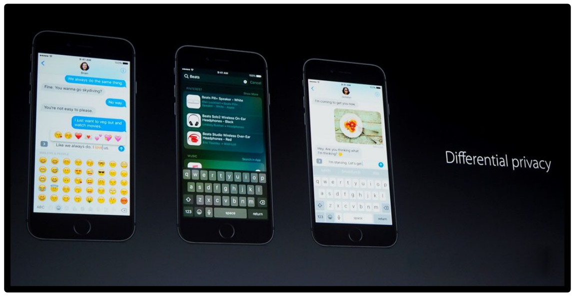


How do we define "privacy"?

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

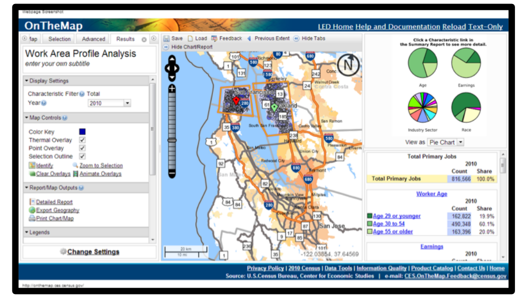
Differential Privacy [Dwork, McSherry, Nissim, *S.*, 2006]

- Several current deployments



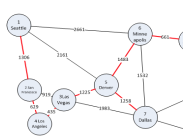
Apple

Google



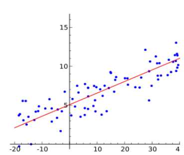
US Census

- Burgeoning field of research



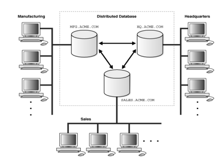
Algorithms

Crypto, security



Statistics, learning

Game theory, economics



Databases, programming languages

Law, policy

Caveats

This is a tutorial.

- Not a survey
  - ➢ Incomplete
  - ➢ If I don't cite your (or my) work, please forgive me.

- Not a broadcast
  - ➢ Ask questions!
  - ➢ Lots of material. No agenda.

Some Fantastic Resources

- Cynthia Dwork and Aaron Roth. *Algorithmic Foundations of Data Privacy, 2013* (Extended tutorial / textbook)
- Salil Vadhan. *The Complexity of Differential Privacy*, 2017.
- Aaron Roth and Adam Smith. Lecture Notes on Adaptive Data Analysis
  - http://adaptivedatanalysis.com
- *Tutorial videos:*
  - 2012 DIMACS Workshop on Differential Privacy across Computer Science.
  - 2013 Simons Workshop on Big Data and Differential Privacy
  - 2016 Newton Institute Workshop on Data Privacy and Linkage
  - 2017 Bar-Ilan Winter School on Private Data Analysis
  - 2019 Simons Institute Semester Program
    - Bootcamp + 3 workshops

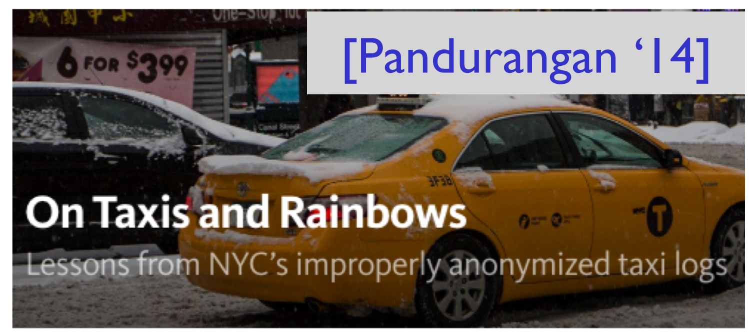
DIFFERENTIAL PRIVACY

- Episode III: Attack of the Codes
  - Reconstruction attacks
  - Membership attacks
- Episode IV: A New Hope
  - Differential privacy
- Episode VI: Return of the Algorithms
  - Algorithms for counting queries
  - Optimization and learning
- Episode VII: The Connections Awaken
  - Learning and adaptive data analysis
  - Statistics
  - Game theory
  - Law and policy

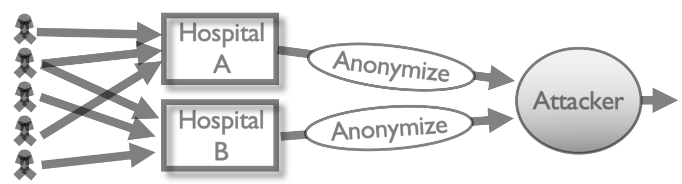
First attempt: Remove obvious identifiers

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

[Gymrek McGuire Golan Halperin Erlich '13]



Everything is an identifier

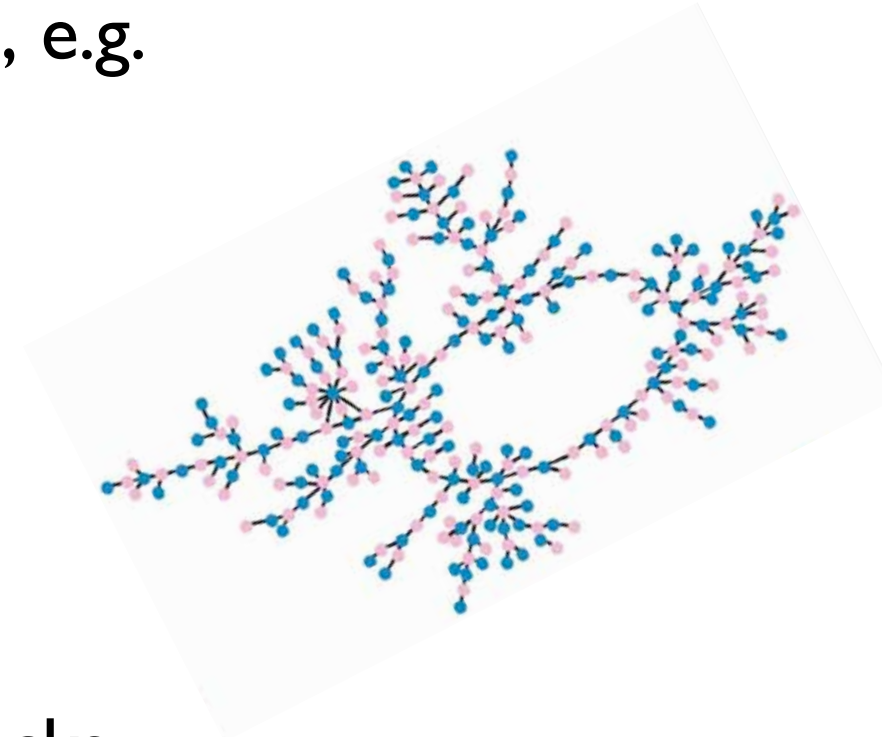


[Ganta Kasiviswanathan S '08]

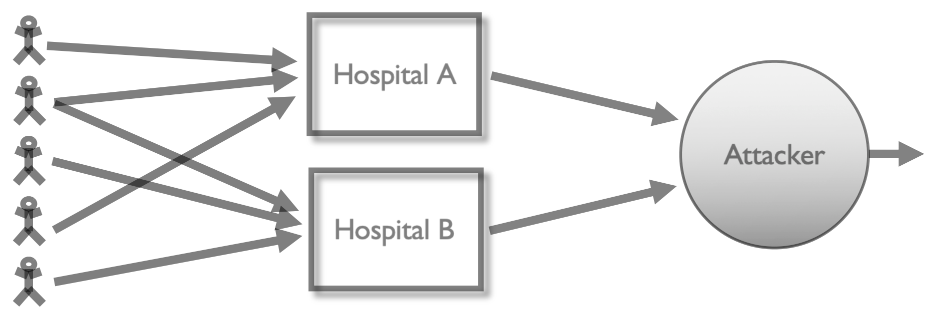
Images: whitehouse.gov, genesandhealth.org, medium.com

Other reidentification attacks

- … based on external sources, e.g.
  - ➢ Social networks
  - ➢ Computer network traces
  - ➢ Microtargeted advertising
  - ➢ Recommendation systems
  - ➢ Genetic data



- … based on composition attacks
  - ➢ Combining independent anonymized releases



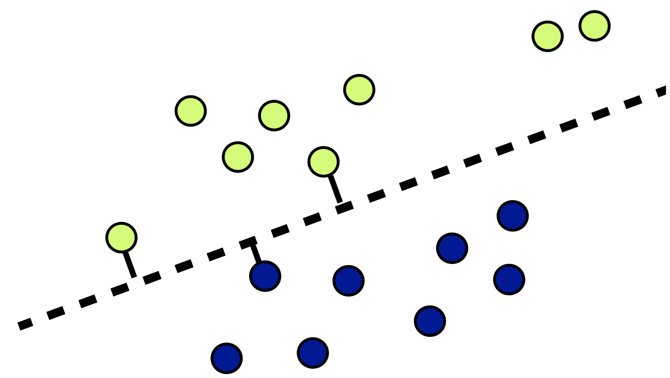
[Citations omitted]

Is the problem granularity?

What if we only release aggregate information?

Statistics together may encode data

- Average salary before/after resignation
- Support vector machine output reveals individual data points



- More generally:

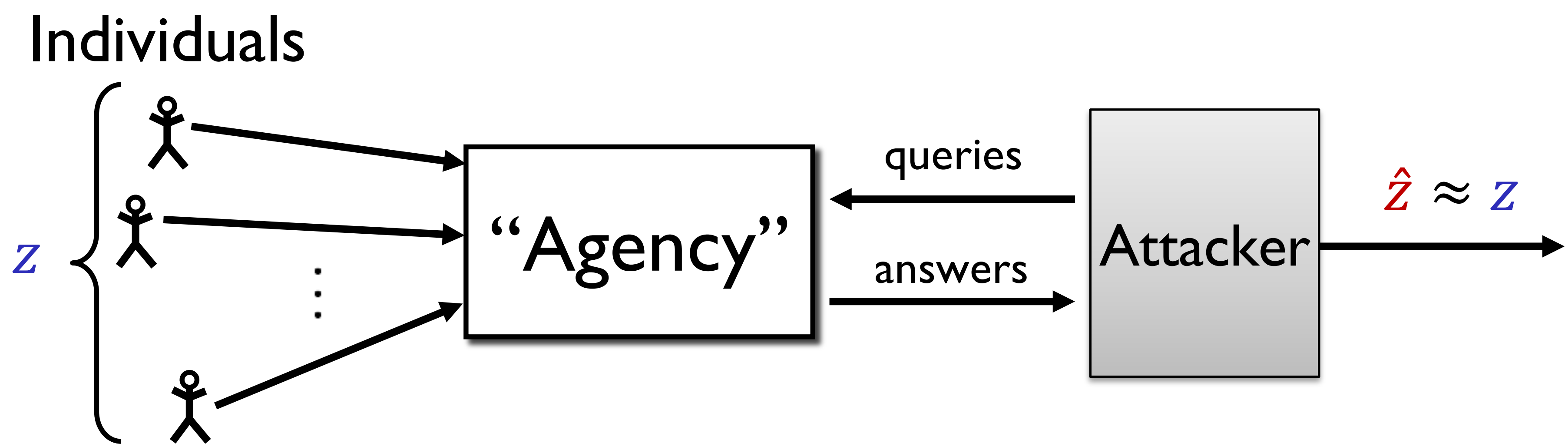
Too many, "too accurate" statistics reveal individual information

- Reconstruction attacks [Dinur Nissim 2003, …]
- Membership attacks [Homer et al. 2008, …]

Cannot release everything everyone would want to know

Reconstruction Attacks

Reconstruction Attacks [Dinur, Nissim 2003]



- If
  - ➢ Agency publishes "enough" facts
  - ➢ Facts are "sufficiently accurate"

  then attacker can reconstruct (part of) the data.
- Typically: view facts + side information as constraints
- IT angle: view agency as $\underbrace{\text{encoding}}_{\text{Computing facts}} + \underbrace{\text{noise}}_{\text{Distorting facts}}$

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Mark Hansen, New York Times, Dec. 5, 2018

In November 2016, the bureau staged something of an attack on itself. Using only the summary tables with their eight billion numbers, [they would] try to generate a record for every American that would show [their Census answers] — a “reconstruction” of the person-level data.

[…]

[The NYT was] able to perform our own reconstruction experiment on Manhattan. Roughly 1.6 million people are divided among 3,950 census blocks — which typically correspond to actual city blocks. The summary tables we needed came from the census website; we used simple tools like R and the Gurobi Optimizer; and within a week we had our first results. […]

An Abstract Setting: Linear Queries [DN'03]

The following problem arises in several settings:

- Data set has n people
- Secret vector $z \in \{0,1\}^n$
  - (1 bit per person, not the same as the data set)
- Attacker sees only

$$y = \frac{1}{n} Qz + e \quad where \begin{cases} Q \in \{0,1\}^{m \times n} \\ |e|_\infty \leq \alpha \end{cases}$$

- Under what conditions on Q, α can attacker reconstruct $\hat{z} \in \{0,1\}^n$ such that $\frac{Ham(\hat{z}, z)}{n} \to 0$?

$$y = \frac{1}{n} Qz + e$$

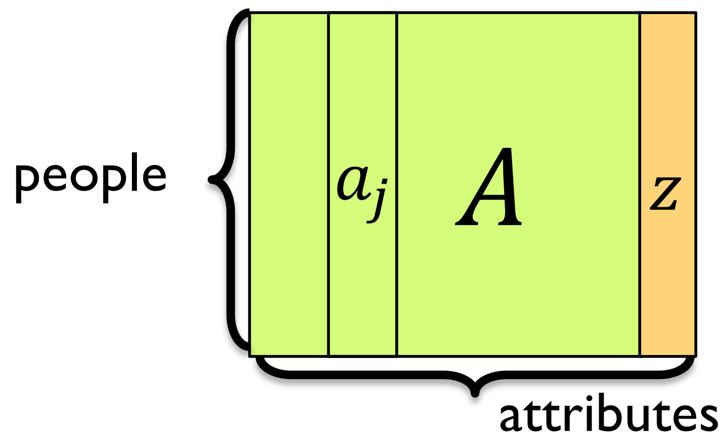
$$where \begin{cases} Q \in \{0,1\}^{m \times n} \\ |e|_\infty \leq \alpha \end{cases}$$

Example 1: Secret Attribute

- Data set is $X = (A|z)$ where $A \in \{0,1\}^{d\times n}$ is matrix of known attributes, z is secret
  - Each person's data is $d + 1$ bits

Suppose release reveals…

- Pairwise correlations
  - Attacker learns $y_j = \frac{\langle a_j, z\rangle}{n} \pm \alpha$ for each j
  - $y = A^\top z + e$ and $m = d$.
- 3-wise conjunctions
  - Attacker learns $y_{j,\ell} = \frac{\langle a_j * a_\ell\, , z\rangle}{n} \pm \alpha$
  - $y = (A * A)^\top z + e$ and $m = \binom{d}{2}$
- Convex optimization
  - Example: linear regression
  - Attacker learns $\hat{\theta} = argmin_\theta \|A\theta - z\|_2^2$
  - That is, $2(A\theta - z)^\top A \approx 0$
  - Induces (approximate) linear constraints on z



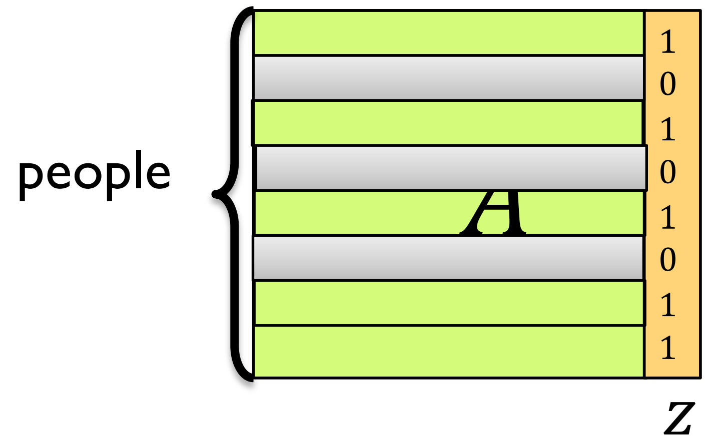
$$y = \frac{1}{n} Qz + e$$

$$where \begin{cases} Q \in \{0,1\}^{m\times n} \\ |e|_\infty \leq \alpha \end{cases}$$

Example 2: Who's In, Who's Out?

$$y = \frac{1}{n} Qz + e \quad where \begin{cases} Q \in \{0,1\}^{m\times n} \\ |e|_\infty \le \alpha \end{cases}$$

- Attacker knows superset of actual data set
  - A is matrix of superset (rows are potential individuals)
  - z is indicator vector of actual data
    - $X = diag(z_1, z_2, \dots, z_n)\ A$
  - Column sums of X are $\vec{1}^{\intercal} X = Az$
    - Approximate marginal statistics give $\Rightarrow$ approximate linear constraints on z
  - Similarly with k-way statistics and convex optimization



- Reconstructing z tells attacker who is in the data set
  - More about this type of attack later

When can we reconstruct? [DN'03]

$$y = \frac{1}{n} Qz + e$$

$$\text{where } \begin{cases} Q \in \{0,1\}^{m \times n} \\ |e|_\infty \leq \alpha \end{cases}$$

- **All queries:** what if Q contains all possible $q \in \{0,1\}^n$?
  - $\boxed{m = 2^n}$
  - We know $\left|y - \frac{1}{n} Qz\right|_\infty \leq \alpha$
- **Attack**: Given y, Q: Set $\hat{z} = argmin_w \left|\frac{1}{n} Qw - y\right|_\infty$
- **Theorem:** $Ham(\hat{z}, z) \leq 4\alpha n$
- Proof:
  - $Ham(\hat{z}, z) \leq 2 \max_{q \in \{0,1\}^n} |q^t(\hat{z} - z)| \leq 4\alpha n$
  - But $\max_{q \in \{0,1\}^n} |q^t(\hat{z} - z)| = n \left|\frac{1}{n} Q(\hat{z} - z)\right|_\infty$

    $\leq \left|\frac{1}{n} Q\hat{z} - y\right|_\infty + \left|\frac{1}{n} Qz - y\right|_\infty \leq 2\alpha n$
  - Get $\frac{Ham(\hat{z}, z)}{n} \to 0$ as long as $\alpha \to 0$
- To release anything that allows one to answer all counting queries, even approximately, you have to release the data

How well can we reconstruct?

$$y = \frac{1}{n} Qz + e \quad \text{where} \begin{cases} Q \in \{0,1\}^{m \times n} \\ |e|_\infty \leq \alpha \end{cases}$$

- What if m is close to n, not 2^n?
- General strategy

$$\hat{z} = round\left(argmin_{w \in [-1,1]^n} \left\| y - \frac{1}{n} Qw \right\|_p\right) \text{ for a } p \in [1, \infty]$$

- Rough rule: If $m > Cn$ and Q is "nice", then $\frac{Ham(\hat{z},z)}{n} \leq O(\alpha^2 n)$
  - When $\alpha \ll \sqrt{n}$, the error goes to 0.
  - Beautiful connections to compressed sensing and discrepancy
- What's "nice"?
  - Large min eigenvalue [DY]
  - Bounded "partial discrepancy" [MN]
  - Restricted isometry properties (beyond ℓ_∞ bounds on error) [DMT,De]
- What kinds of matrices?
  - Random [DiNi,DMT,…]
  - Random conjunctions [KRSU]
  - Hadamard [DY]

# queries m	2^n	$O(n)$
Error $\frac{Ham(\hat{z},z)}{n}$	4α	$2\alpha^2 n$
Running time	$\Omega(2^n)$	$O(n \log n)$

Hadamard Queries [DY08]

- Queries given by rows of ± 1 Hadamard matrix:

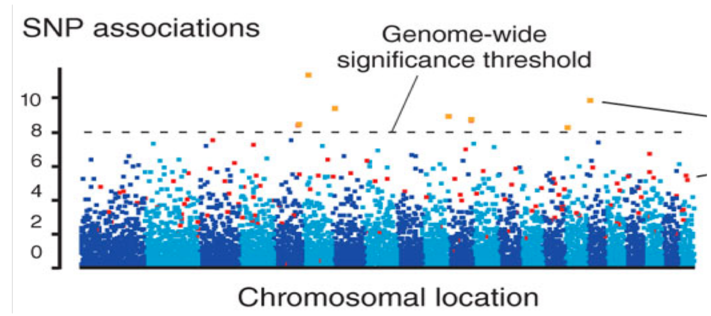
$$H_1 = (1) \qquad H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix}$$

- Attacks gets $y = \frac{1}{n} H_n z + e$ where $|e|_\infty \leq \alpha$
  - $\tilde{z} = argmin_{\mathrm{w}} \left\| \frac{1}{n} H_n w - y \right\|_2 = n H_n^{-1} y = z + n H_n^{-1} e$
  - $\hat{z} = round(\tilde{z})$
- Running time: $O(n \log n)$ by divide and conquer (FFT)
- Error
  - $\frac{Ham(round(\tilde{z}),z)}{n} \leq \frac{2}{n} \|\tilde{z} - z\|_2^2$ by Markov argument
  - Eigenvalues of H_n are $\pm\sqrt{n}$ since $(H_n)^2 = nI$
    - $\|\tilde{z} - z\|_2 \leq \sqrt{n} \|e\|_2 = \alpha n$ since $\|e\|_2 = \alpha\sqrt{n}$
  - $\frac{Ham(round(\tilde{z}),z)}{n} \leq 2\alpha^2 n$

Membership Testing Attacks

A Few Membership Attacks

- [Homer et al. 2008]
  Exact high-dimensional summaries
  allow an attacker
  to test membership in a data set

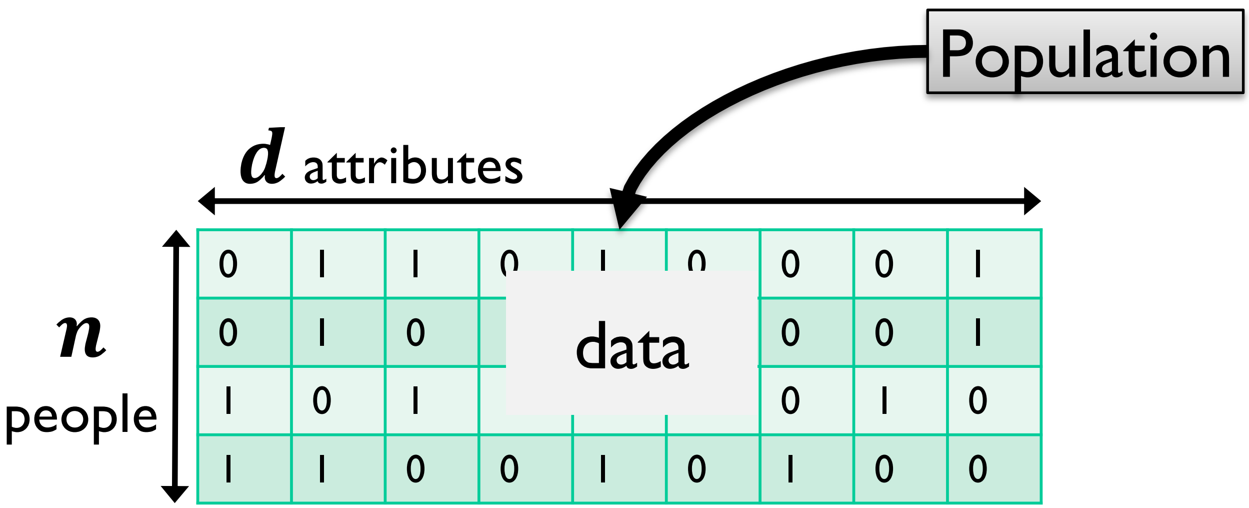


  - ➢ Caused US NIH to change data sharing practices

- [Dwork, **S**, Steinke, Ullman, Vadhan, *FOCS* '15]
  Distorted high-dimensional summaries
  allow an attacker
  to test membership in a data set

- [Shokri, Stronati, Song, Shmatikov, *Oakland* 2017]
  Membership inference using ML as a service
  (from exact answers)
  - ➢ Several follow-up papers in the security literature

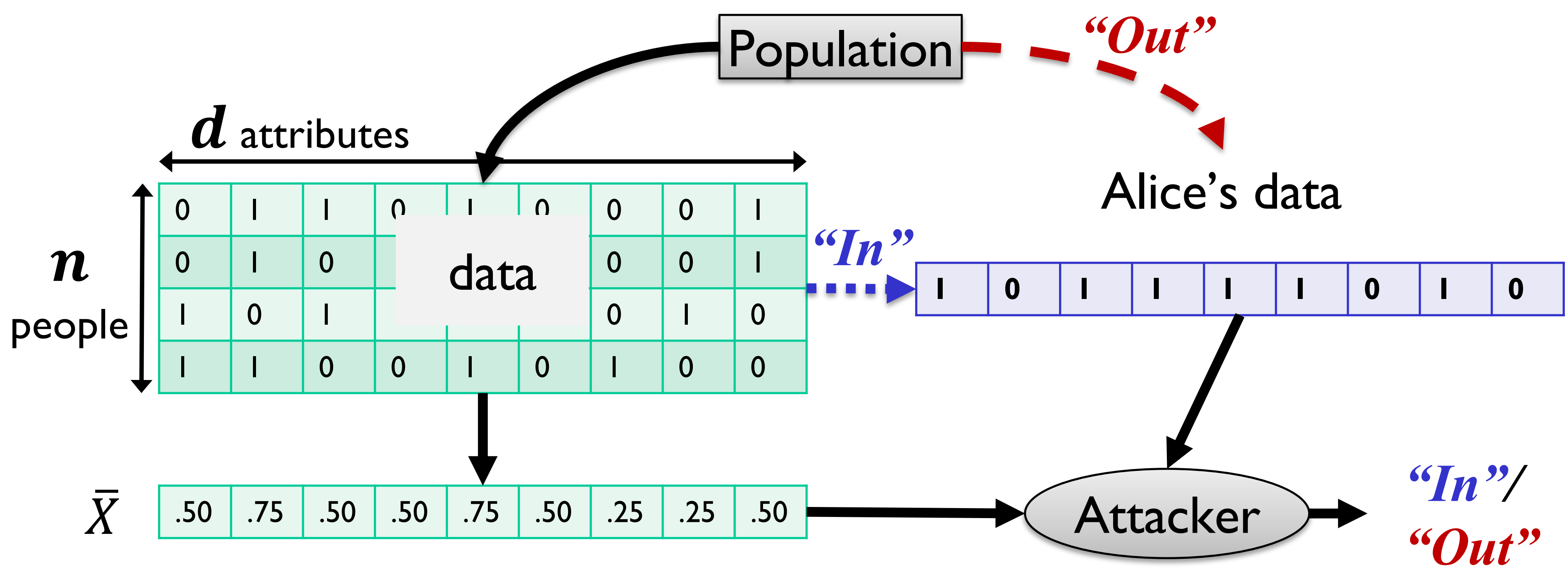
Membership Attacks



Suppose

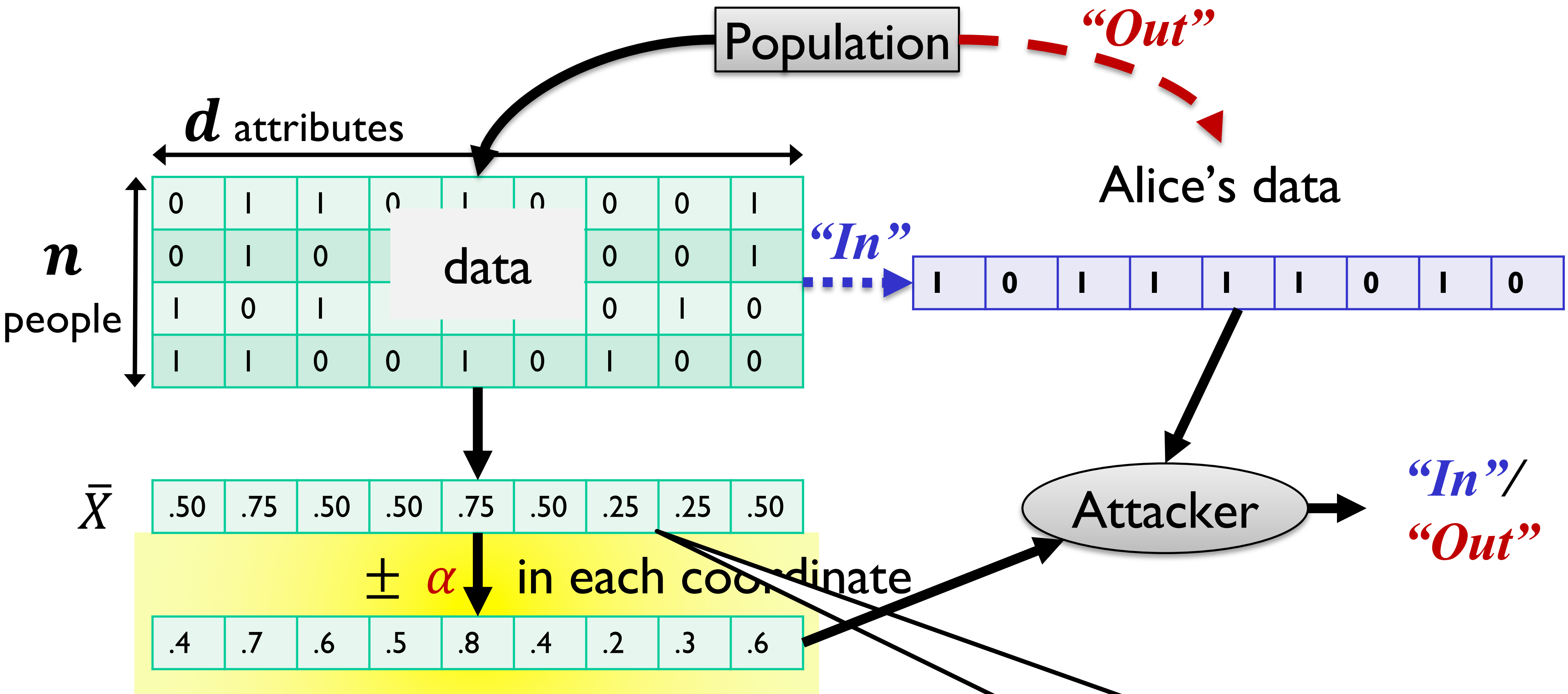
- We have a data set in which membership is sensitive
  - ➢ Participants in clinical trial
  - ➢ Targeted ad audience
- Data has many binary attributes for each person
  - ➢ Genome-wide association studies
    $d = 1\,000\,000$ (“SNPs”), $n < 2000$

Membership Attacks



- Release exact column averages
- Attacker succeeds with high probability when there are more attributes than people

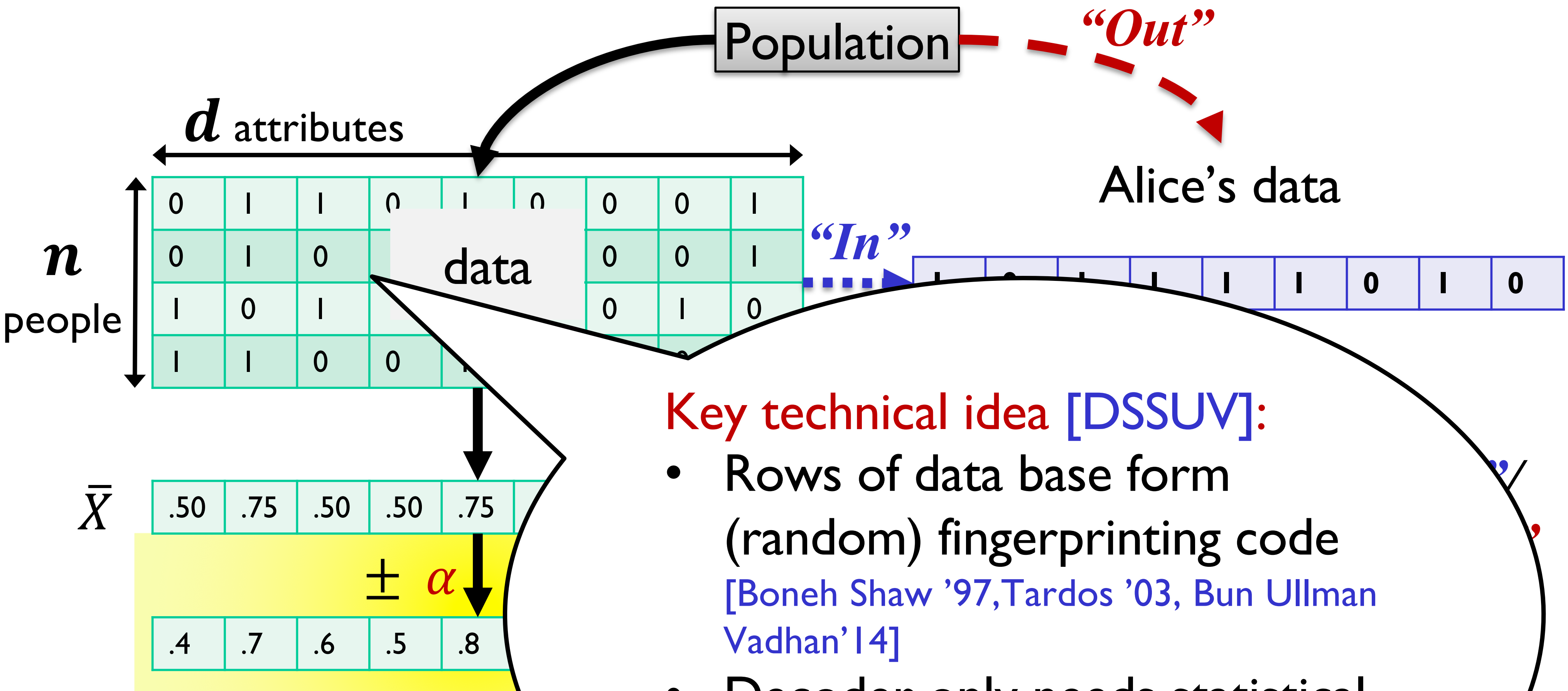
Membership Attacks



- Release ~~exact~~ distorted column avera
- Attacker succeeds with high probab there are more attributes than people and $\alpha \ll \sqrt{d}/n$

No matter how distortion performed

Membership Attacks



Key technical idea [DSSUV]:

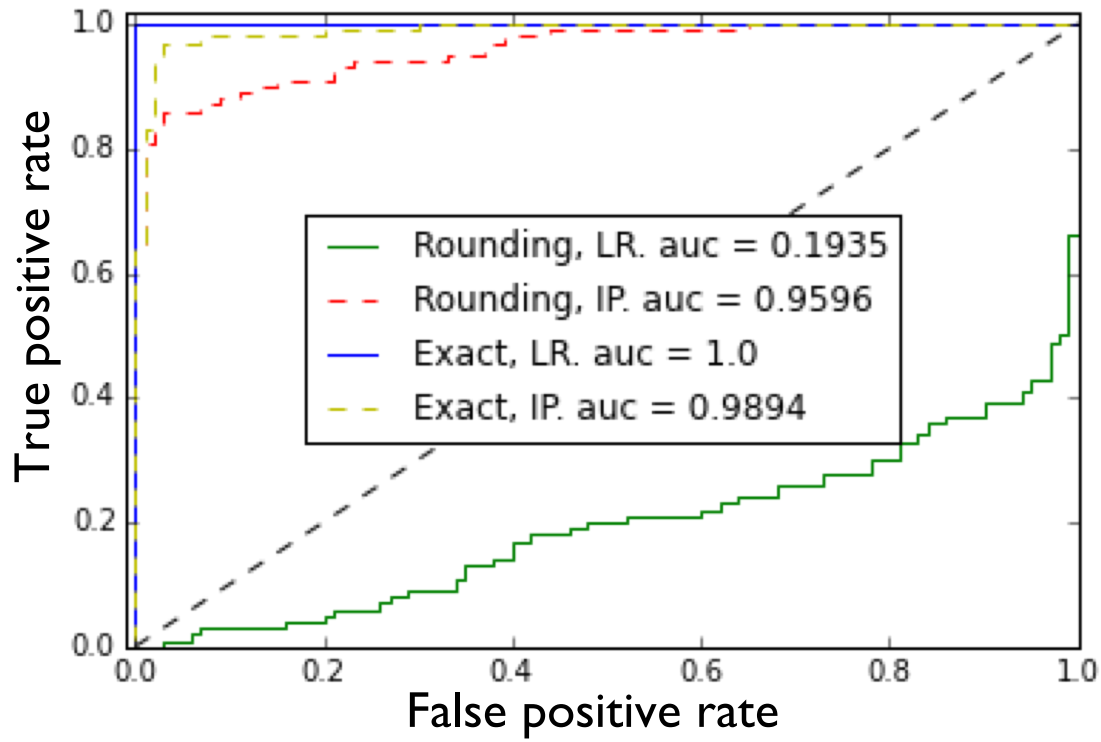
- Rows of data base form (random) fingerprinting code [Boneh Shaw '97, Tardos '03, Bun Ullman Vadhan'14]
- Decoder only needs statistical summaries
- Analysis is subtle

- Release ~~exact~~ dis
- Attacker succeeds w

  there are more attributes th $\ll \sqrt{d}/n$

Robustness to perturbation

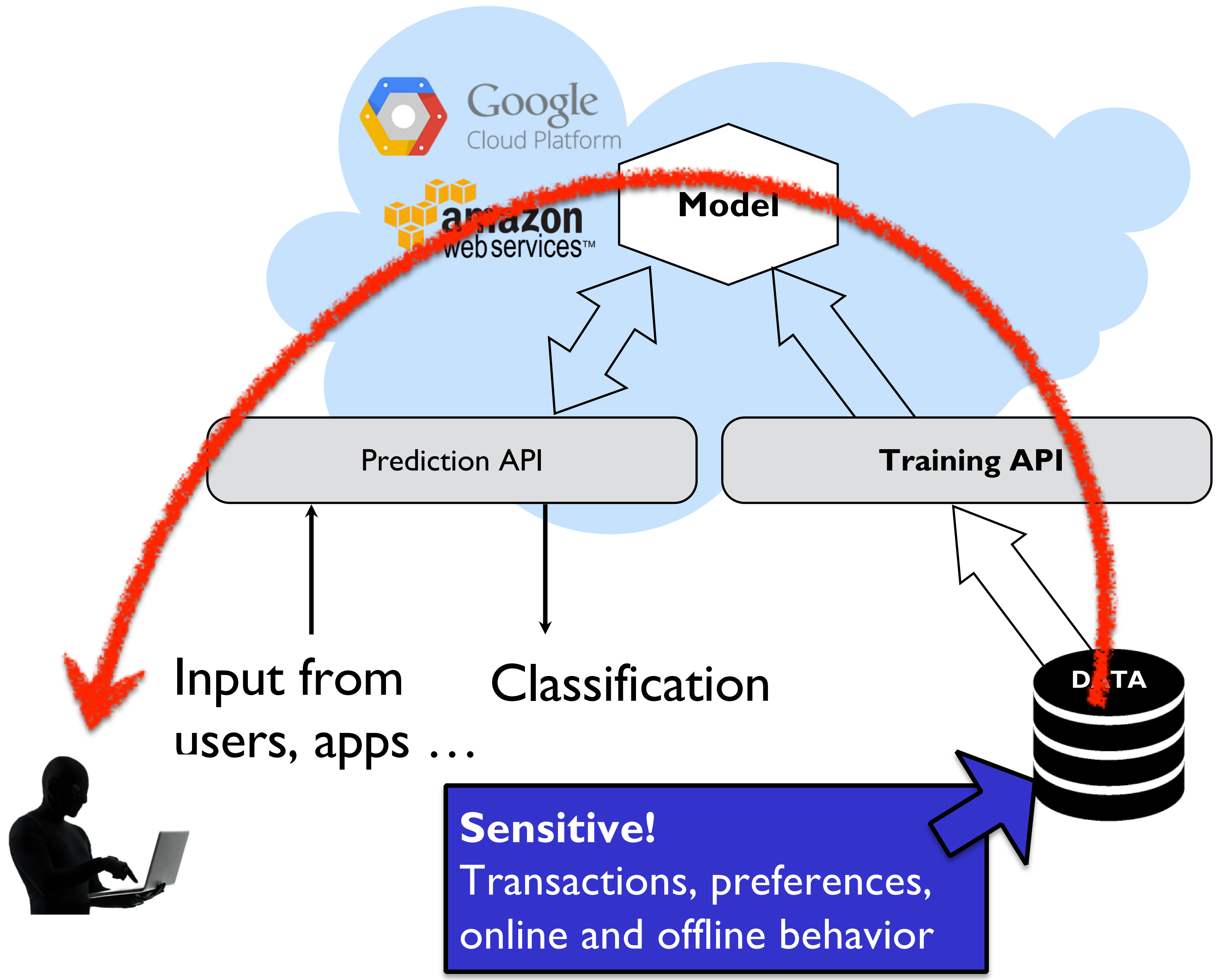
- $n = 100$
- $m = 200$
- $d = 5{,}000$
- Two tests
  - LR [Sankararam et al'09]
  - IP [DSSUV'15]



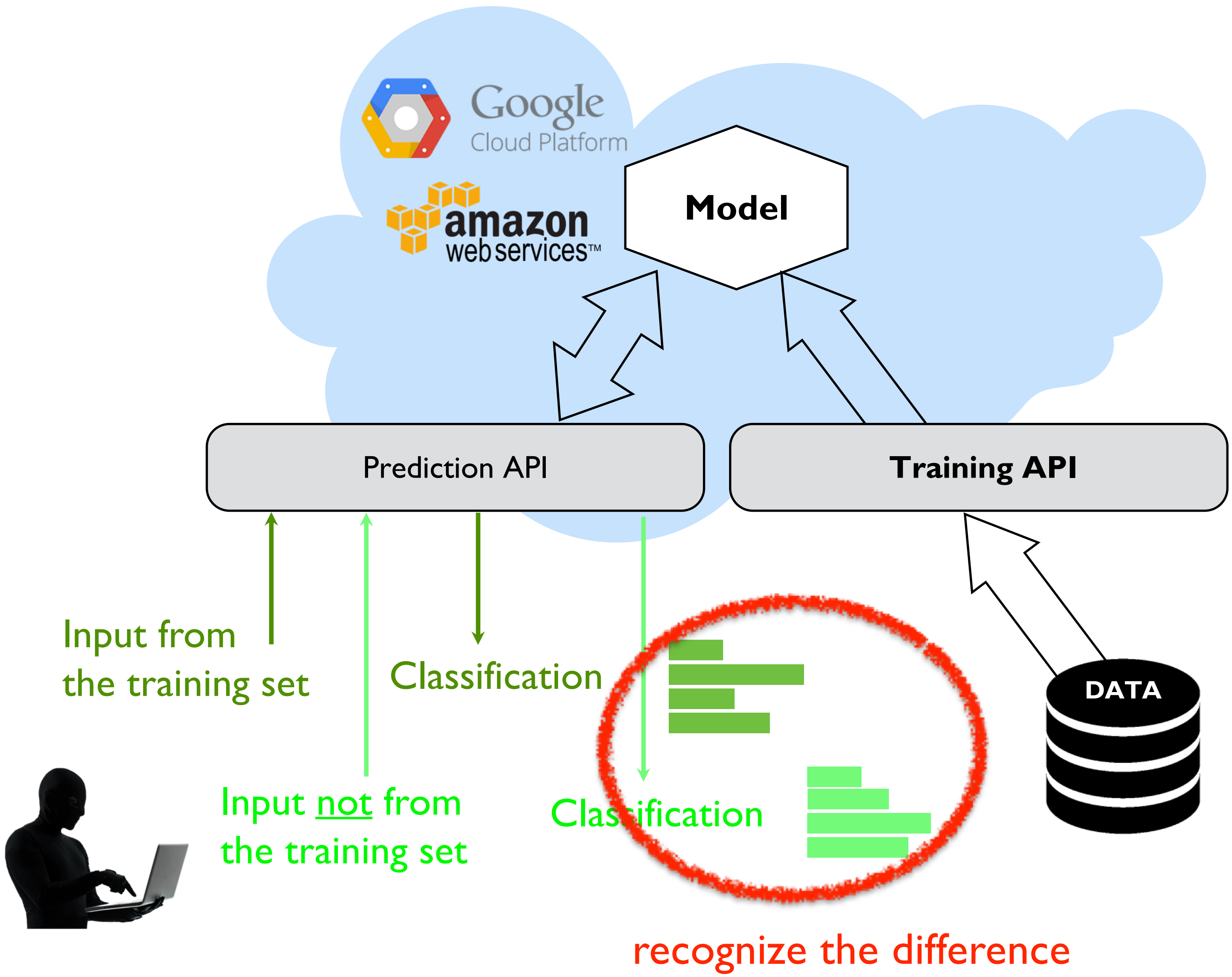
- Two publication mechanisms
  - Rounded to nearest multiple of 0.1 (red / green)
  - Exact statistics (yellow / blue)

Conclusion: IP test is robust.
Calibrating LR test seems difficult

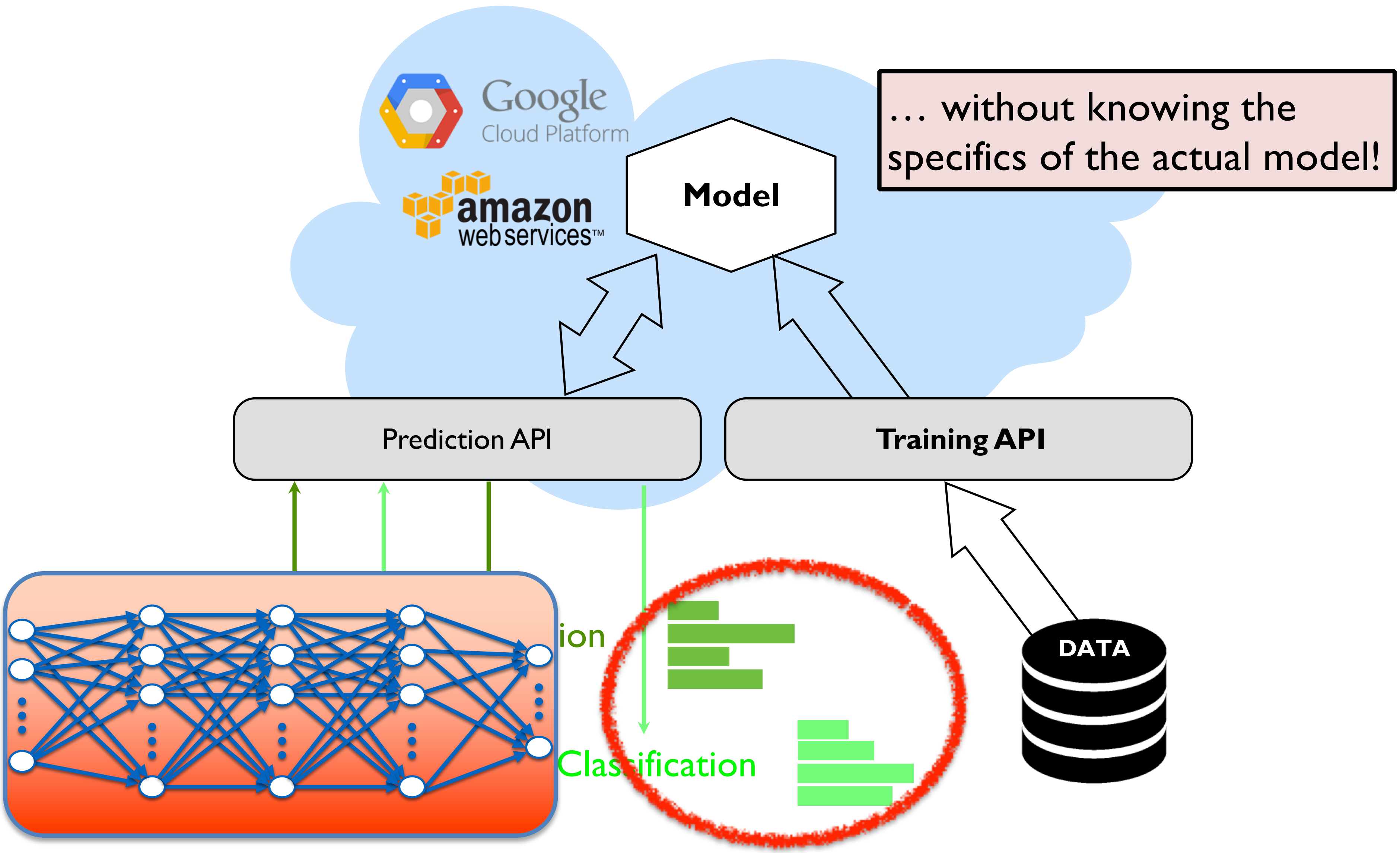
Machine Learning as a Service



Exploiting Trained Models



Exploiting Trained Models



Train a model to…

recognize the difference

Lessons

1. "Too many, too accurate" statistics allow one to reconstruct the data

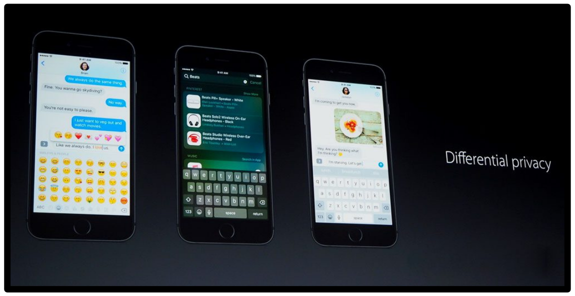
2. "Aggregate" is hard to pin down

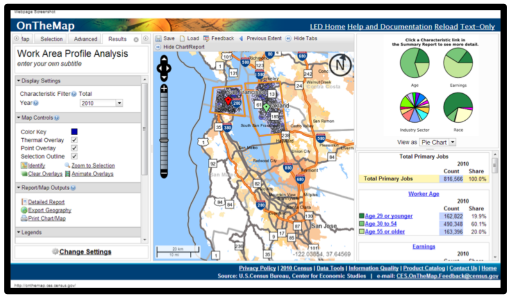
DIFFERENTIAL PRIVACY

- Episode III: Attack of the Codes
  - Reconstruction attacks
  - Membership attacks
- Episode IV: A New Hope
  - Differential privacy
- Episode VI: Return of the Algorithms
  - Algorithms for counting queries
  - Optimization and learning
- Episode VII: The Connections Awaken
  - Learning and adaptive data analysis
  - Statistics
  - Game theory
  - Law and policy

Differential Privacy

- Several current deployments

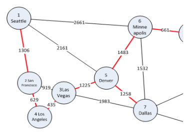




Apple Google US Census

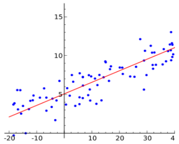
In the works: Uber, Yahoo, Microsoft, LinkedIn, …

- Burgeoning field of research



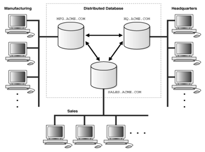
Algorithms

Crypto, security



Statistics, learning

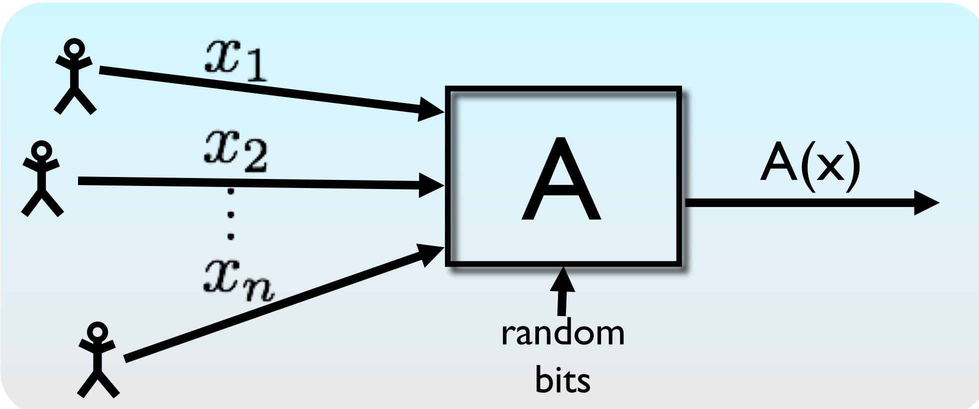
Game theory, economics



Databases, programming languages

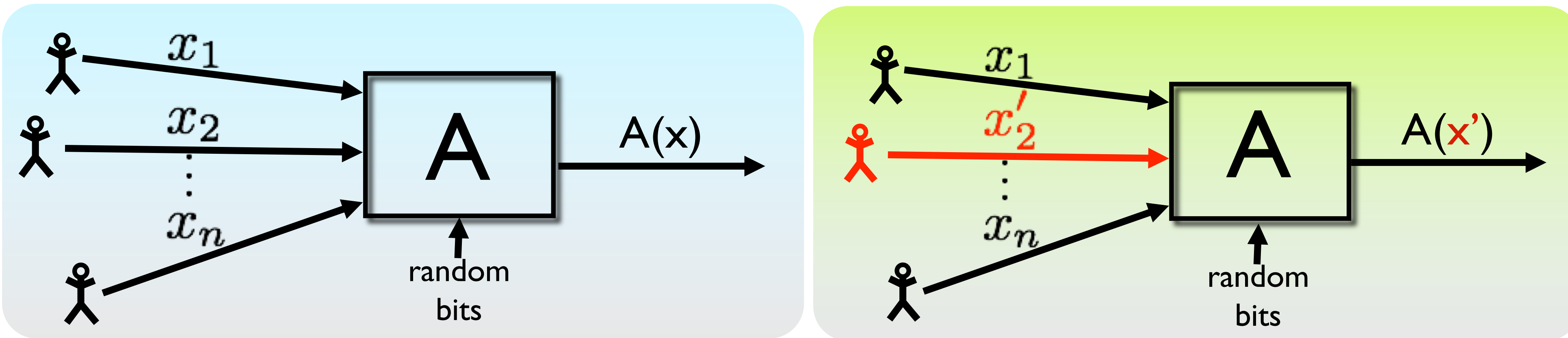
Law, policy

Differential Privacy

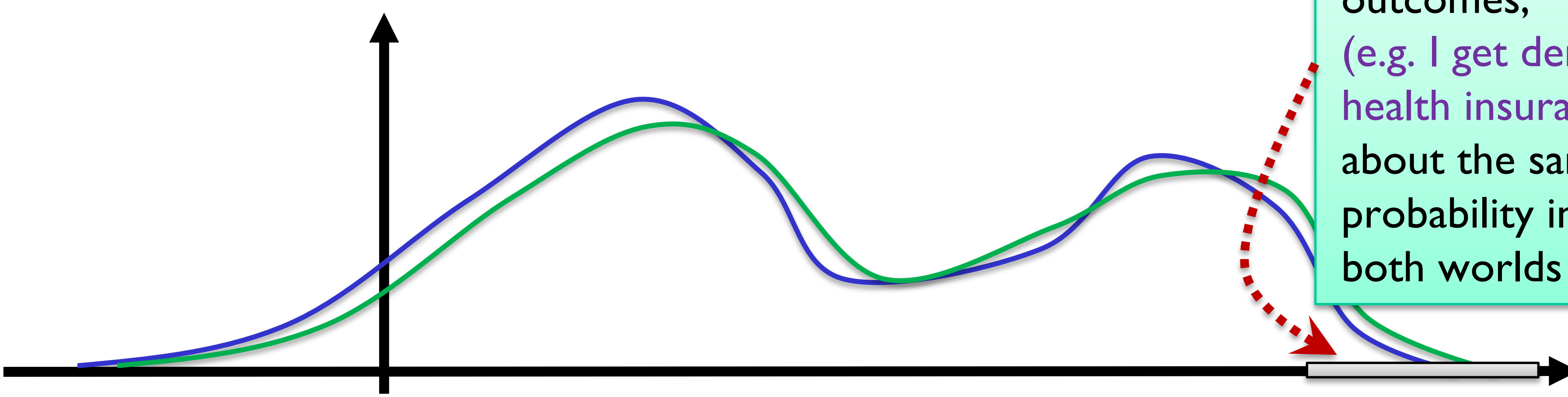


- Data set $x = (x_1, ..., x_n) \in D^n$
  - Domain D can be numbers, categories, tax forms
  - Think of x as **fixed** (not random)
- A = **probabilistic** procedure
  - A(x) is a random variable
  - Randomness might come from adding noise, resampling, etc.

Differential Privacy

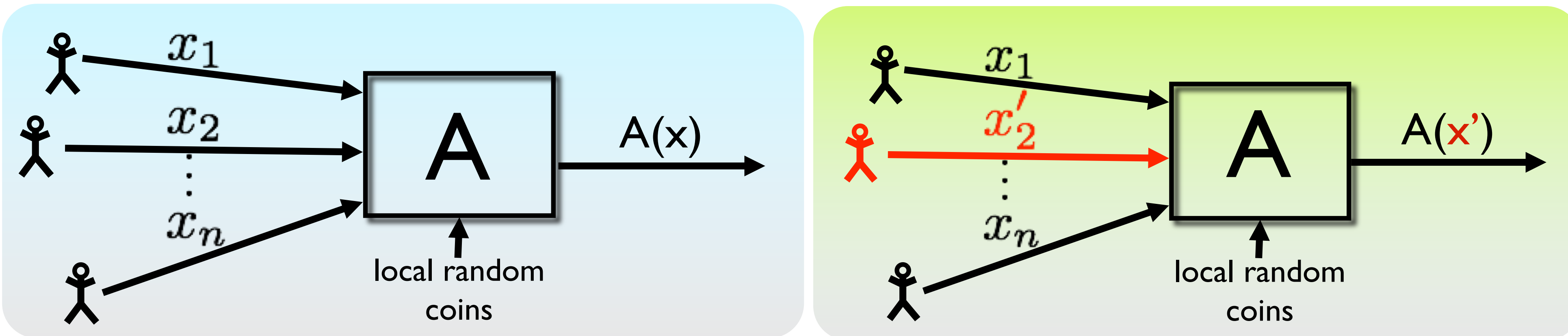


- A thought experiment
  - ➢ Change one person's data (or add or remove them)
  - ➢ Will the **probabilities of outcomes** change?



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

Differential Privacy



x' is a neighbor of x
if they differ in one data point

Neighboring databases induce **close** distributions on outputs

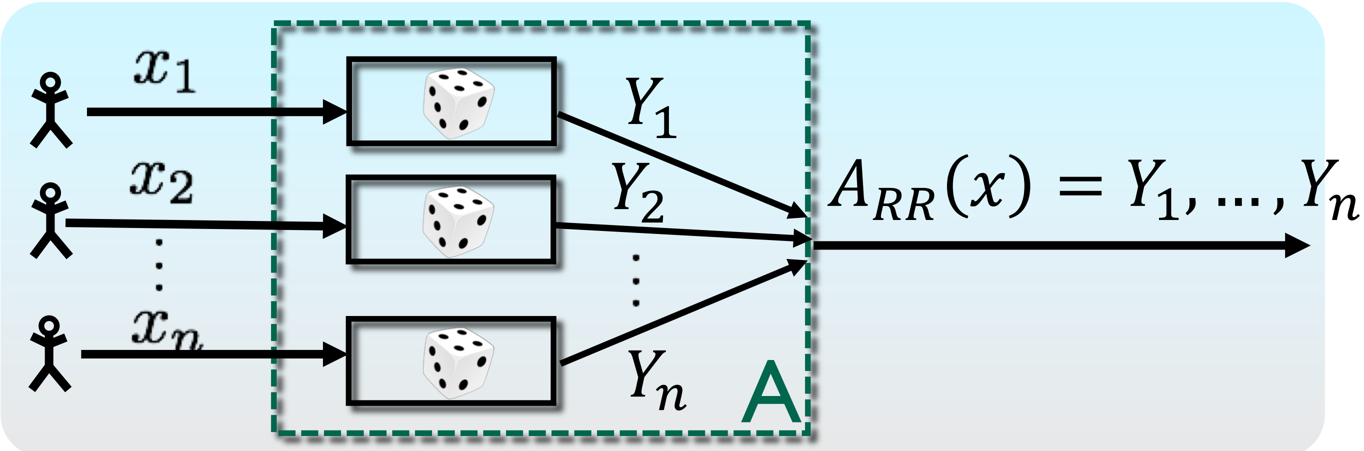
Definition: A is ϵ-differentially private if,
for all neighbors x, x',
for all subsets S of outputs

$1 + \epsilon$

$$\Pr(A(x) \in S) \le e^{\epsilon} \Pr(A(x') \in S)$$

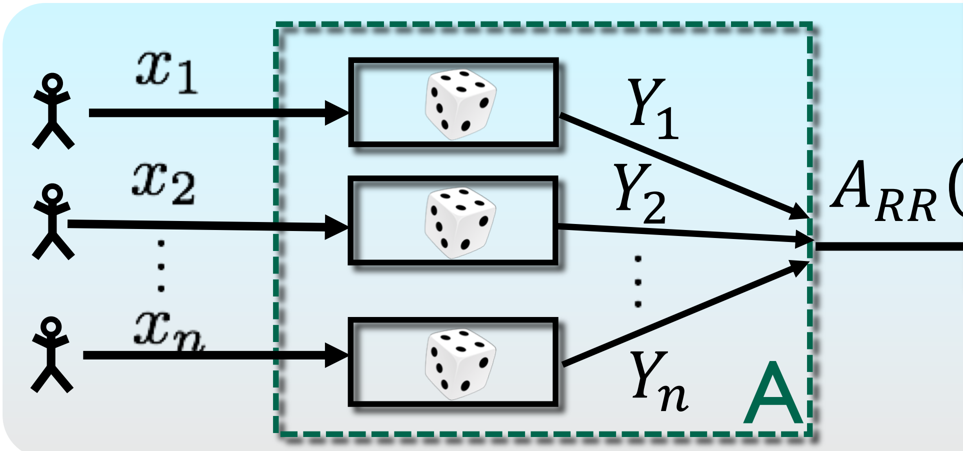
ϵ is a leakage measure

Randomized Response [Warner 1965]



- Want to release the fraction of students who've cheated on a test
  - 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 $Y_i =$ **true** value x_i
    - 5 or 6: Report $Y_i =$ opposite value $1 - x_i$
  - Output = list of reported values $Y_1, \dots, Y_n$

Randomized Response

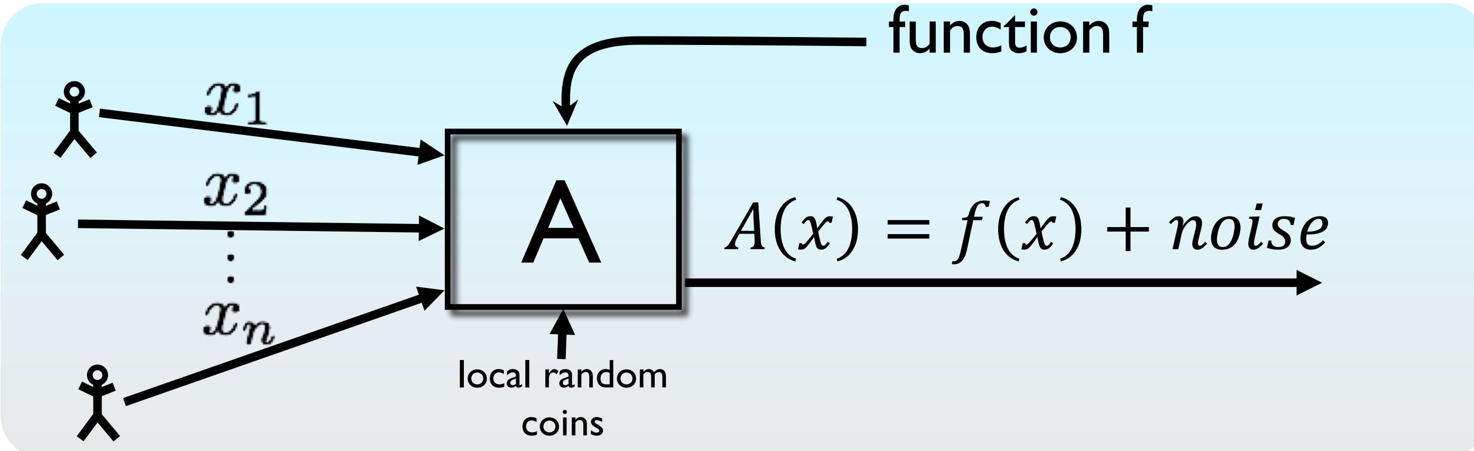


Each individual rolls a die

- $1, 2, 3$ *or* 4:
  Report **true** value x_i
- 5 *or* 6:
  Report opposite value $\overline{x_i}$

- Why is it "private"?
  - Thought experiment: Change x_{Adam} from 0 to 1
    - $Y_{Adam} = 1$ happens with probability $\frac{2}{3}$ instead of $\frac{1}{3}$
      ∴ Plausible deniability
  - Satisfies ϵ-DP with $\epsilon \approx 0.7$
- Why is it "useful"?
  - Can estimate fraction of x_i's that are 1
  - Exercise: Find f such that $\mathrm{E}\left|f(A_{RR}(\vec{x})) - \frac{1}{n}\sum_i x_i\right| = \Theta(\frac{1}{\epsilon\sqrt{n}})$

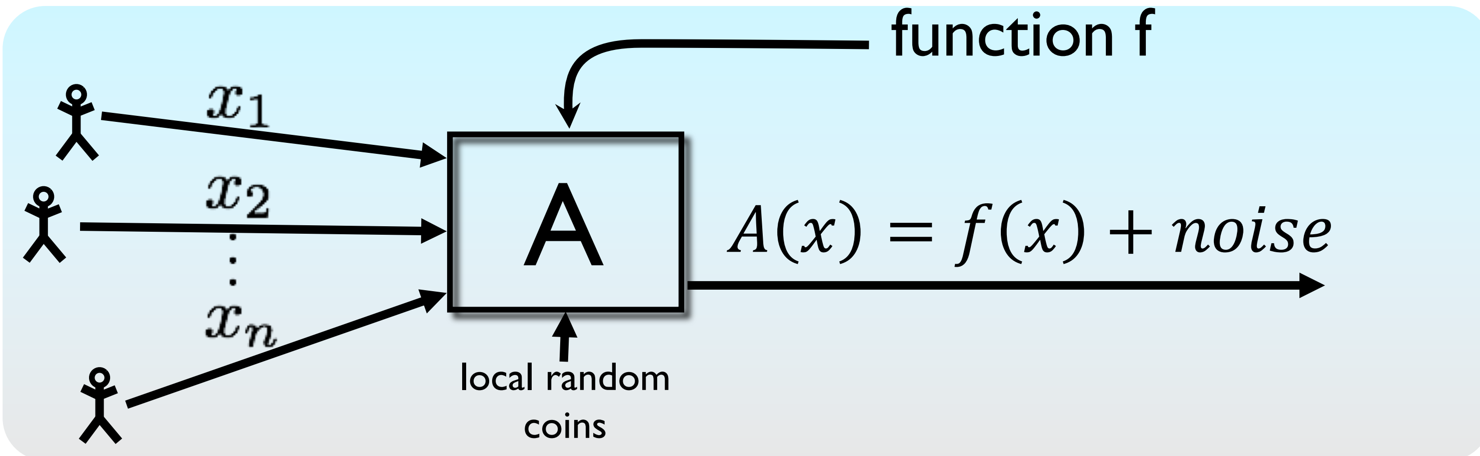
Laplace Mechanism



- Say we want to release a summary $f(x) \in \mathbb{R}^d$
  - e.g., proportion of cheaters: $x_i \in \{0,1\}$ and $f(x) = \frac{1}{n}\sum_i x_i$

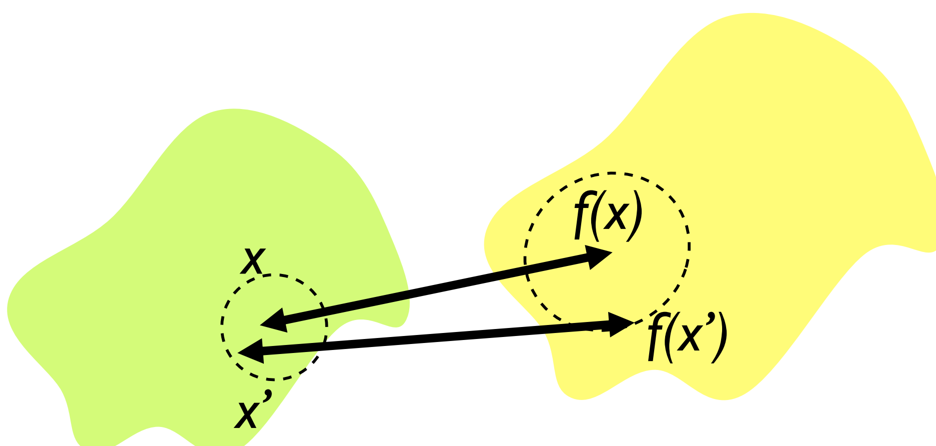
- Simple approach: add noise to $f(x)$
  - How much noise is needed?
  - Idea: Calibrate noise to some measure of f's volatility

Laplace Mechanism

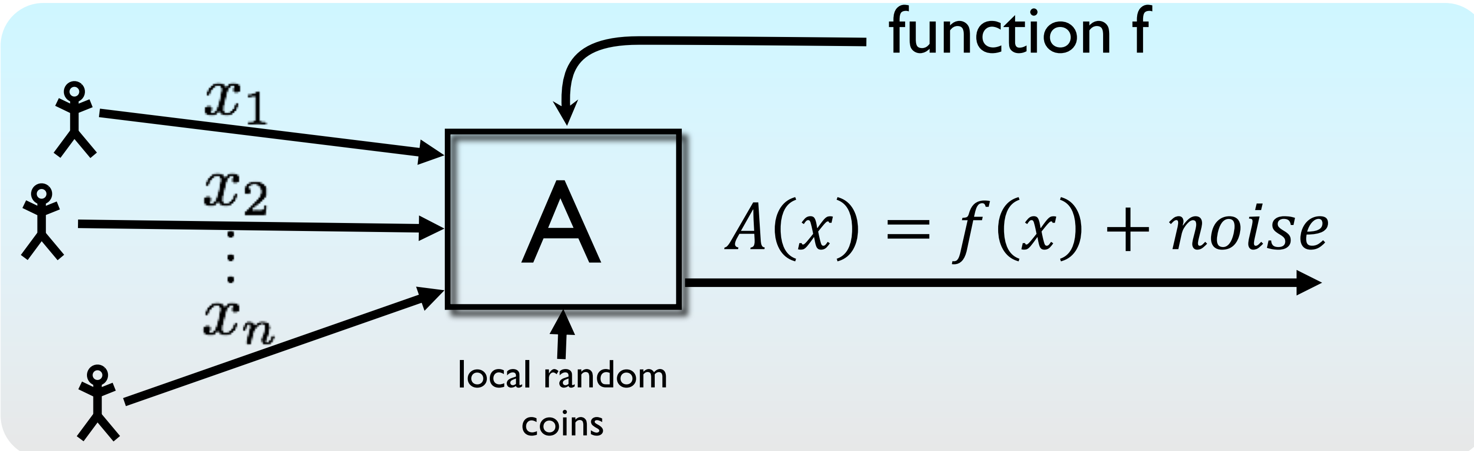


- Global Sensitivity: $\mathrm{GS}_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$

  - Example: $\mathrm{GS}_{\text{proportion}} = \frac{1}{n}$



Laplace Mechanism



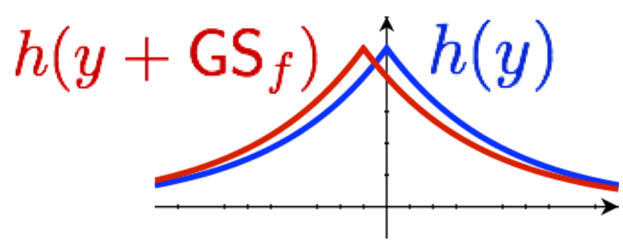
- Global Sensitivity: $\mathrm{GS}_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$

  - Example: $\mathrm{GS}_{\text{proportion}} = \frac{1}{\mathrm{n}}$

Theorem: $\mathrm{A}_{\text{Lap}}(x) = f(x) + Lap\left(\frac{GS_f}{\epsilon}\right)$ is ϵ-DP.

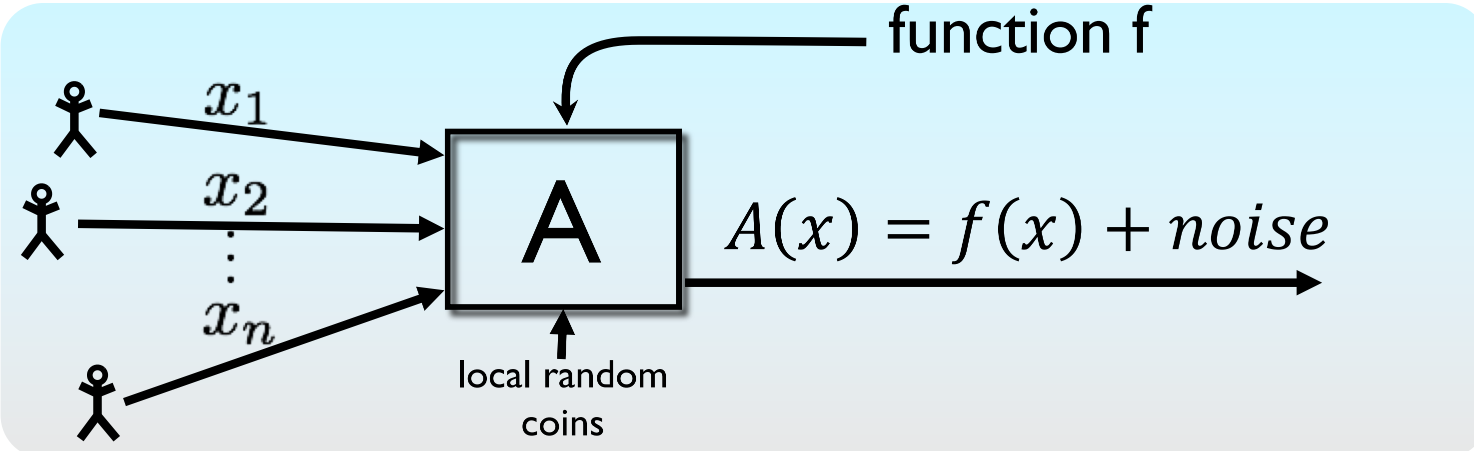
- Requires noise from Laplace distribution

$$h(y) = \frac{1}{2\lambda} e^{-|y|/\lambda}$$

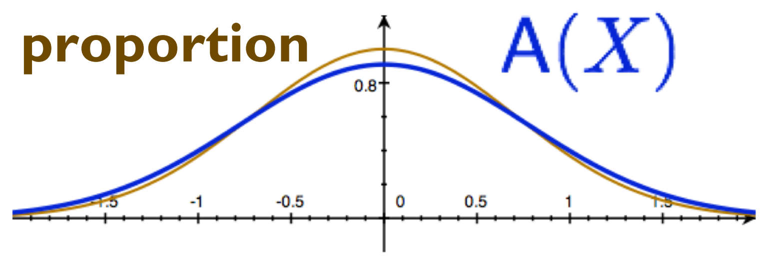


- Changing one value translates curve

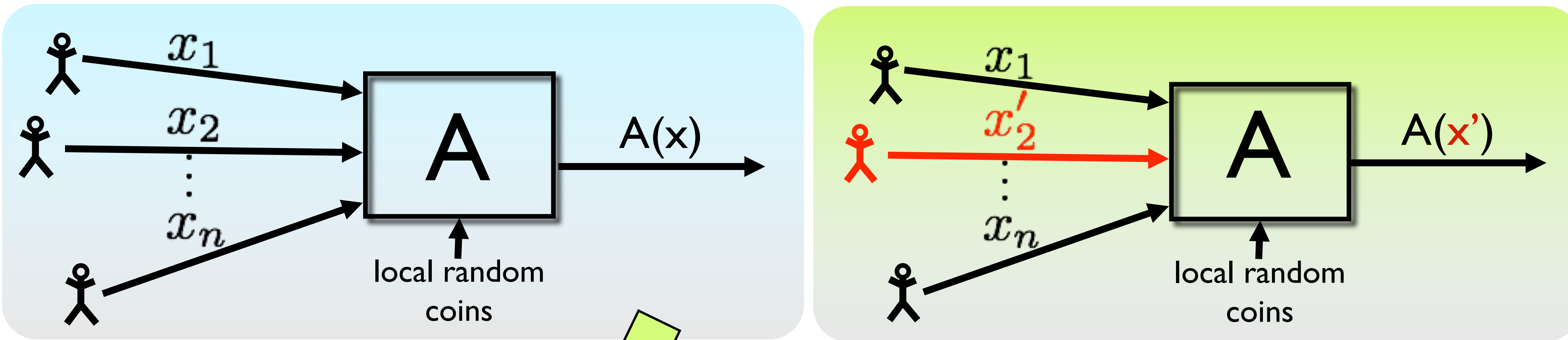
Laplace Mechanism



- Example: proportion of diabetics
  - ➢ $\text{GS}_{\text{proportion}} = \frac{1}{n}$
  - ➢ Release $\text{A(x)} = \text{proportion} \pm \frac{1}{\epsilon n}$
- Is this **a lot**?
  - ➢ If x is a random sample from a large underlying population, then **sampling noise** $\approx \frac{1}{\sqrt{n}}$
  - ➢ A(x) "as good as" real proportion



Differential Privacy



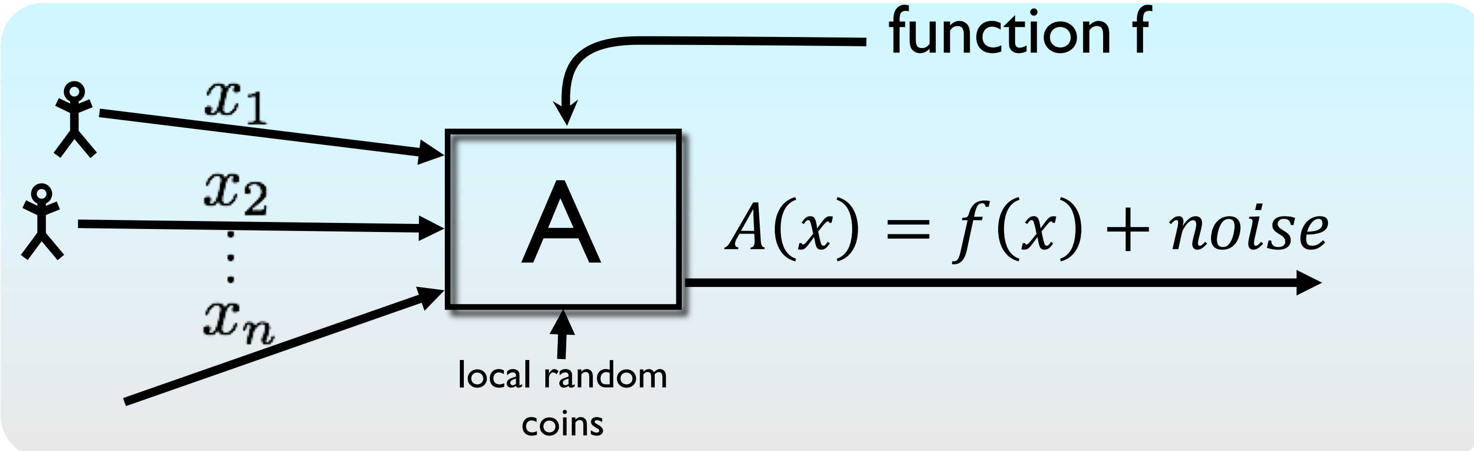
x' is a neighbor of x
if they differ in one data point

Neighboring databases induce **close** distributions on outputs

Definition: A is (ϵ, δ)-differentially private if,
for all neighbors x, x',
for all subsets S of outputs

$$\Pr(A(x) \in S) \leq e^{\epsilon} \cdot \Pr(A(x') \in S) + \delta$$

Gaussian Noise



Gaussian noise addition satisfies (ϵ, δ)-DP
with Euclidean sensitivity $GS_{f,\ell_2} = \max_{neighbors\ x,x'} \|f(x) - f(x')\|_2$

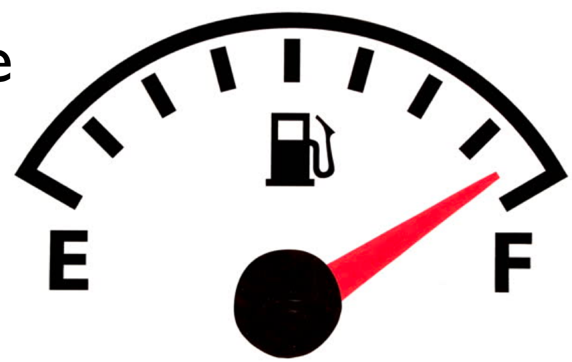
Useful Properties

- **Composition:** If $A_1, A_2, \dots A_k$ are (ϵ, δ)-differentially private, then joint output $A_1(x), A_2(x), \dots, A_k(x)$ is
  - $(k\epsilon, k\delta)$- differentially private [JL09,MM09], and
  - $\approx \left(\epsilon\sqrt{k}\sqrt{\ln 1/\delta},\ k\delta\right)$-differentially private [DRV10]
- **Post-processing:** If A is ϵ-differentially private, then so is $g(A)$ for any function g

Consequence 1: Modular design!

Consequence 2: Privacy is a consumable resource

- ϵ measures leakage
- can be treated as a "privacy budget"
- Each analysis consumes some

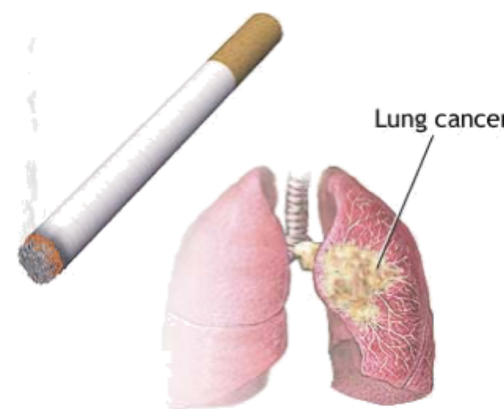


Interpreting Differential Privacy

- A naïve hope:

  ~~Your beliefs about me are the same after you see the output as they were before~~

- Impossible
  - Suppose you know that I smoke
  - Clinical study: "smoking and cancer correlated"
  - You learn something about me
    - Whether or not my data were used



- Differential privacy implies:
  No matter what you know ahead of time,

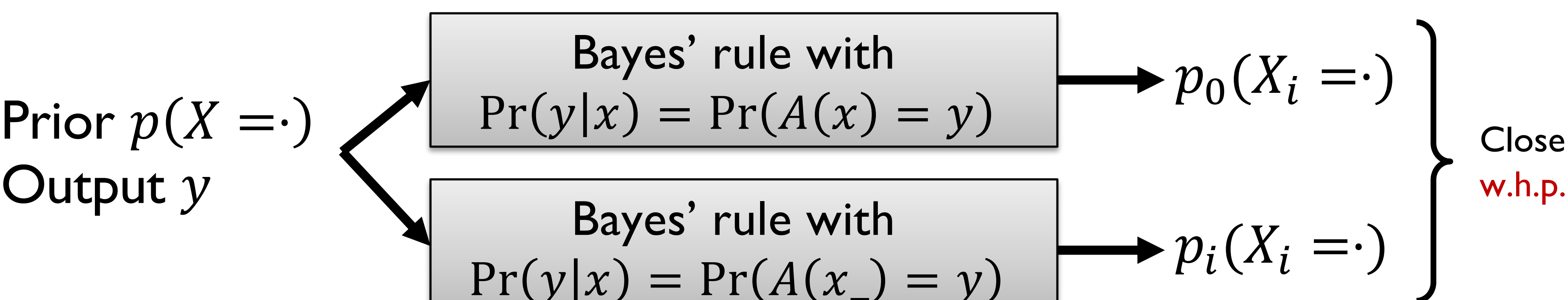
  You learn (almost) the same things about me whether or not my data are used

  - Provably resists attacks mentioned earlier

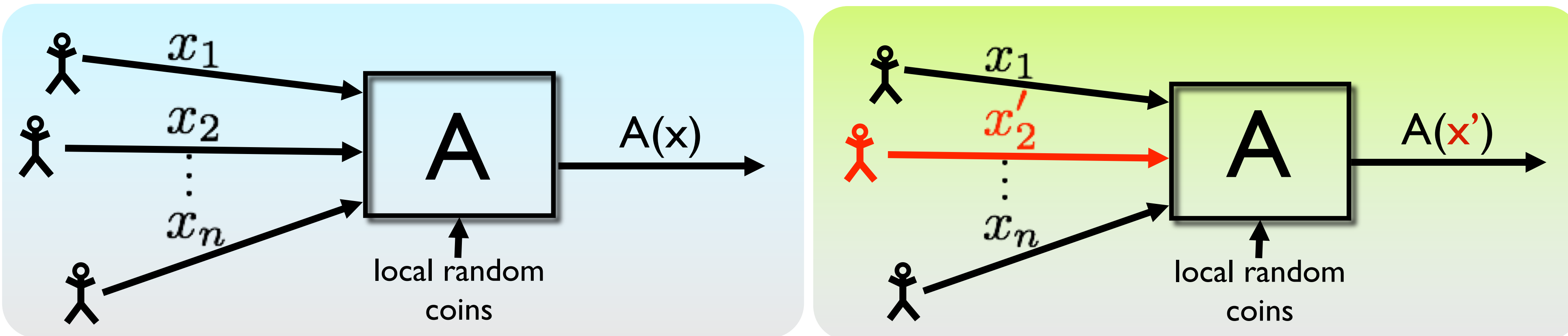
Bayesian Interpretation [KS08]

- Suppose you are an attacker
  - "Background knowledge" = prior distribution $p(X = \cdot)$
  - "Conclusions about i on output a" = $p(X_i = \cdot \mid A(X) = a)$
  - Experiment 0: Run $A(X)$
  - Experiment i: Run $A(X_{-i})$ with $x_{-i} = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$
- **Theorem**: If A is (ϵ, δ)-DP with $\delta \ll \frac{1}{n}$, then for all i,

$$X_i \Big|_{A(X)=a} \approx_{\epsilon', \delta'} X_i \Big|_{A(X_{-i})=a} \quad \textit{with prob.} \geq 1 - \sqrt{\delta n}$$



What can we compute privately?

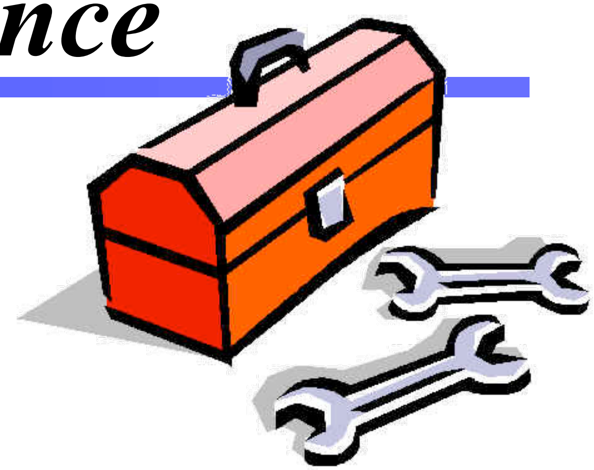


- "Privacy" = change in one input leads to small change in output distribution

  What computational tasks can we achieve privately?

- Lots of recent work, interesting questions
  - ➢ Across different fields: statistics, data mining, machine learning, cryptography, algorithmic game theory, networking, information theory

A Broad, Active Field of Science



- Algorithmic tools and techniques
- Theoretical foundations
  - 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

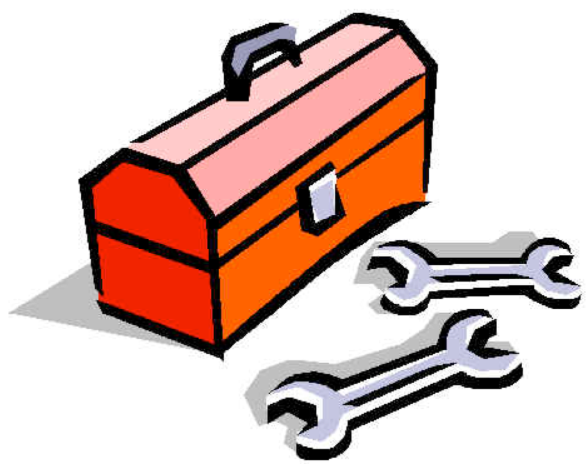
Google Scholar:
1,000+ articles with
"differential privacy"
in the title

13,000+ articles with
"differential privacy"
in text

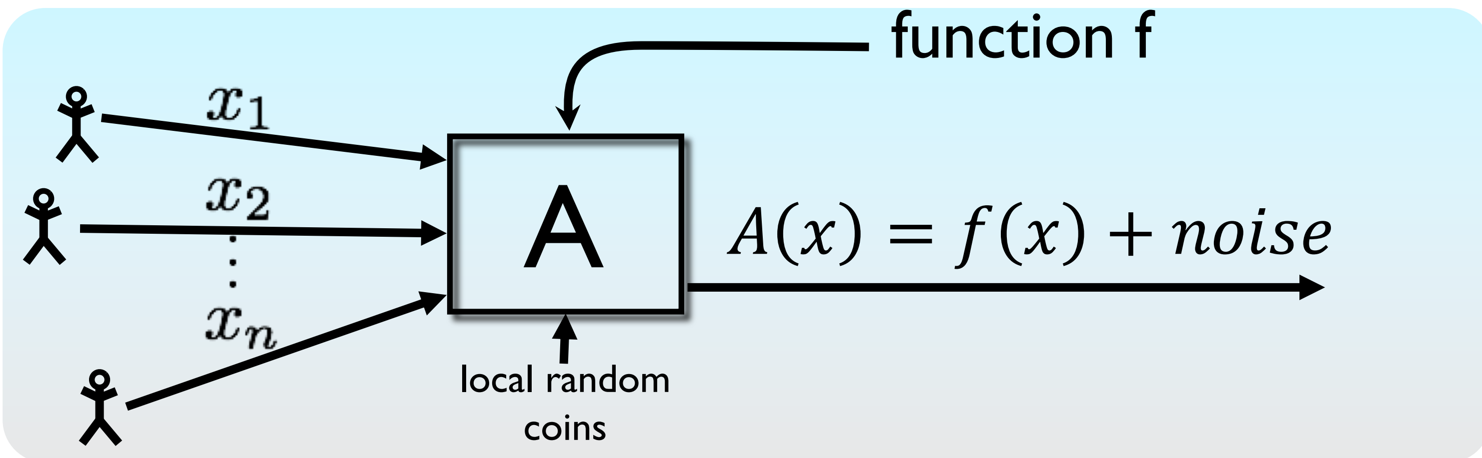
DIFFERENTIAL PRIVACY

- Episode III: Attack of the Codes
  - Reconstruction attacks
  - Membership attacks
- Episode IV: A New Hope
  - Differential privacy
- Episode VI: Return of the Algorithms
  - Algorithms for counting queries
  - Optimization and learning
- Episode VII: The Connections Awaken
  - Learning and adaptive data analysis
  - Statistics
  - Game theory
  - Law and policy

Basic Technique 1:
Noise Addition



Laplace + Gaussian Mechanisms



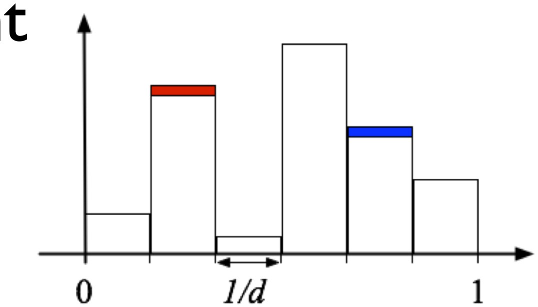
- Global Sensitivity: $GS_{f,p} = \max_{x,x' \text{neighbors}} \|f(x) - f(x')\|_p$
  - Example: $\mathrm{GS}_{\text{proportion}} = \frac{1}{n}$

Theorem:

- $A_{\text{Lap}}(x) = f(x) + Lap\left(\frac{GS_{f,1}}{\epsilon}\right)$ is ϵ-DP.
- $A_{\text{Gauss}}(x) = f(x) + N\left(0, \left(2GS_{f,2}\sqrt{\log 1/\delta}\,/\epsilon\right)^2\right)$ is (ϵ, δ)-DP.

Example: Histograms

- Say $x_1, \dots, x_n$ in domain D
  - Partition D into d disjoint bins
  - $f(x) = (n_1, \dots, n_d)$ where $n_j = \#\{i\colon\, x_i \text{ in } j\text{–th bin}\}$
  - $GS_{f,1} = GS_{f,2} = 1$
  - Sufficient to add noise $Lap\left(\frac{1}{\epsilon}\right)$ to each count
- Examples
  - Histogram on the line
  - Populations of 50 states
  - Marginal tables
    - bins = possible combinations of attributes

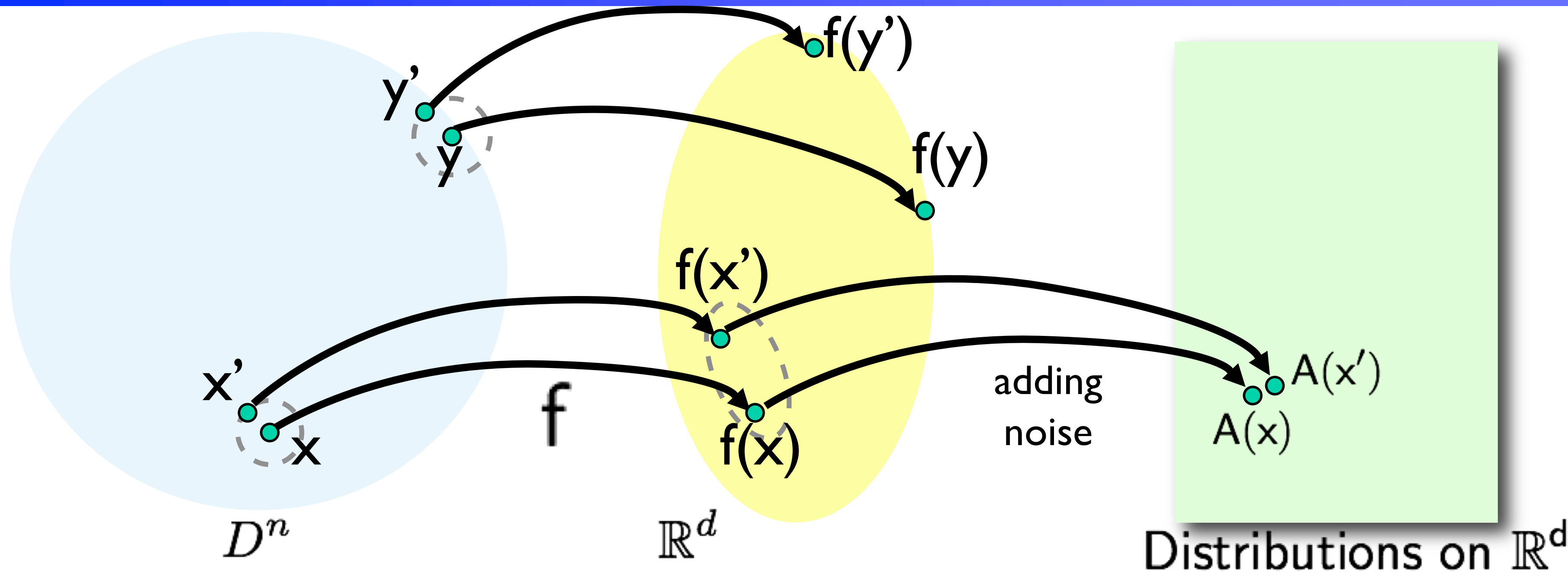


ABO and Rh Blood Type Frequencies in the United States

ABO Type	Rh Type	How Many Have It	
O	positive	38%	45%
O	negative	7%	
A	positive	34%	40%
A	negative	6%	
B	positive	9%	11%
B	negative	2%	
AB	positive	3%	4%
AB	negative	1%	

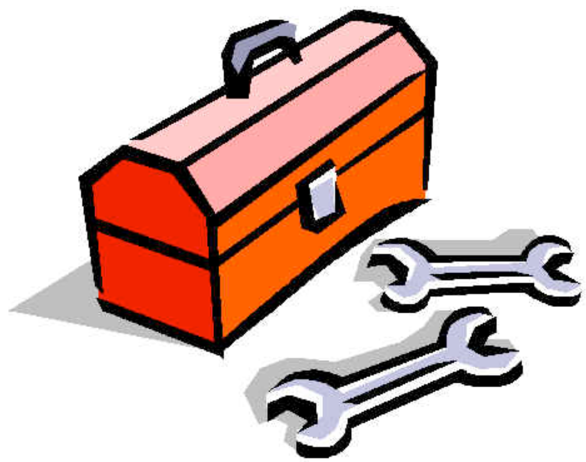
(Source: American Association of Blood Banks)

Global versus local [NRS07]



- Global sensitivity is worst case over inputs
- Local sensitivity: $$\mathsf{LS}_f(x) = \max_{x' \text{ neighbor of } x} \|f(x) - f(x')\|_1$$
- Reminder: $\mathsf{GS}_f(x) = \max_x \mathsf{LS}_f(x)$
- [NRS'07,DL'09, ...] Techniques with error ≈ local sensitivity
  - ➢ Basis of best algorithms for graph data

Basic Technique 2:
Exponential Sampling



Exponential Sampling *[McSherry, Talwar '07]*

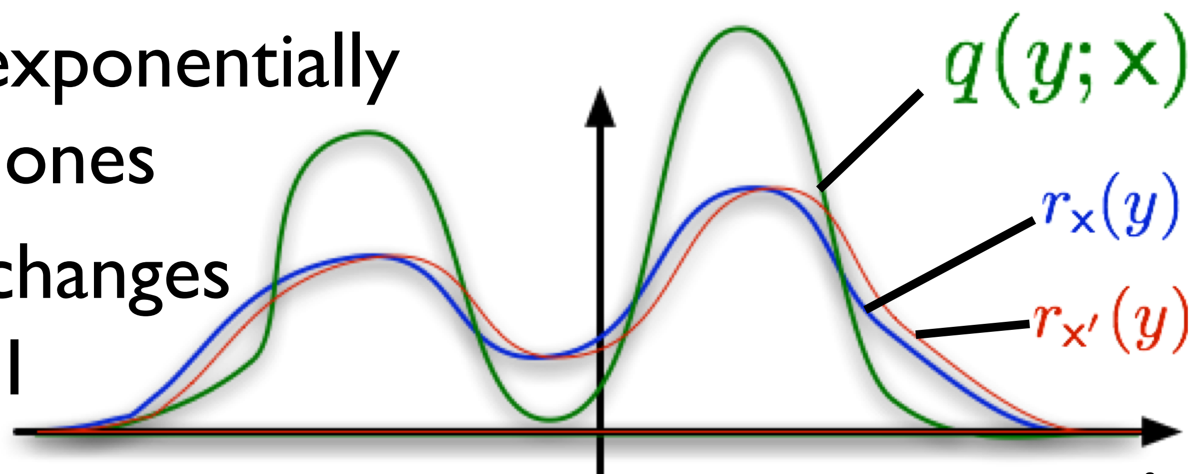
- Sometimes noise addition makes no sense
  - ➢ mode of a discrete distribution
  - ➢ minimum cut in a graph
  - ➢ classification rule
- [MT07] Motivation: auction design
- Subsequently applied very broadly

Exponential Sampling

- Data: x_i = {websites visited by student i today}
- Range: Y = {website names}
- "Score" of y: $q(y;\ x)\ =\ |\{i\ :\ y\ \subseteq\ x_i\}|$
- Goal: output a site with $q(y;\ x) \approx \max_y q(y; x)$

- **ExpMech**: Given x,
  Output website y with probability $r_x(y) \propto e^{\epsilon q(y;x)}$

- Utility: Popular sites exponentially more likely than rare ones
- Privacy: One person changes websites' scores by ≤1



Analysis

- **Lemma**: ExpMech is $(2\epsilon, 0)$-differentially private.
- Proof:
  - Look at ratio $\frac{r_x(y)}{r_{x'}(y)} = \frac{\exp(\epsilon q(y;x))}{\exp(\epsilon q(y;x'))} \cdot \frac{C_{x'}}{C_x}$ where $C_x = \sum_y \exp(\epsilon q(y;x))$
  - Each term contributes at most e^ϵ to ratio.
- **Prop**: Let $OPT_x = \max_y q(y;x)$. For all $\beta > 0$, $\hat{y} = ExpMech(x)$ satisfies $q(\hat{y};x) \geq OPT_x - \ln\left(\frac{|Y|}{\beta}\right)/\epsilon$ with probability $\geq 1 - \beta$.
- Proof: Let $G_t = \{y \in Y : q(y;x) \geq OPT_x - t\}$
  - Consider the ratio $\frac{\Pr(\overline{G_t})}{\Pr(G_t)} \leq |Y|e^{-\epsilon t}$.

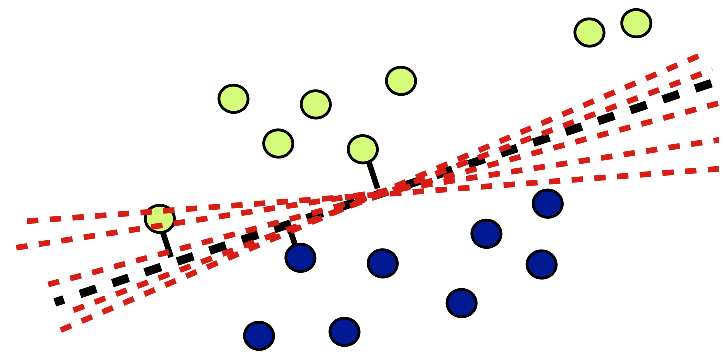
Exponential Sampling, in General

Ingredients:

- Set of outputs Y with prior distribution p(y)
- Score function $q(y;x)$ such that
  for all y, neighbors x, x': $|q(y;x) - q(y;x')| \le \Delta$

ExpMech: Given x,

- Output y from Y with probability
$$r_x(y) \propto p(y) e^{\frac{\epsilon q(y;x)}{\Delta}}$$



- **Prop**: Let $OPT_x = \max_y q(y;x)$. For all $\beta > 0$, $\hat{y} = ExpMech(x)$ satisfies $q(\hat{y};x) \ge OPT_x - \Delta \frac{\ln\left(\frac{|Y|}{\beta}\right)}{\epsilon}$ with probability $1 - \beta$.

Using Exponential Sampling

- Mechanism above very general
  - Every differentially private mechanism is an instance!
  - Still a useful design perspective

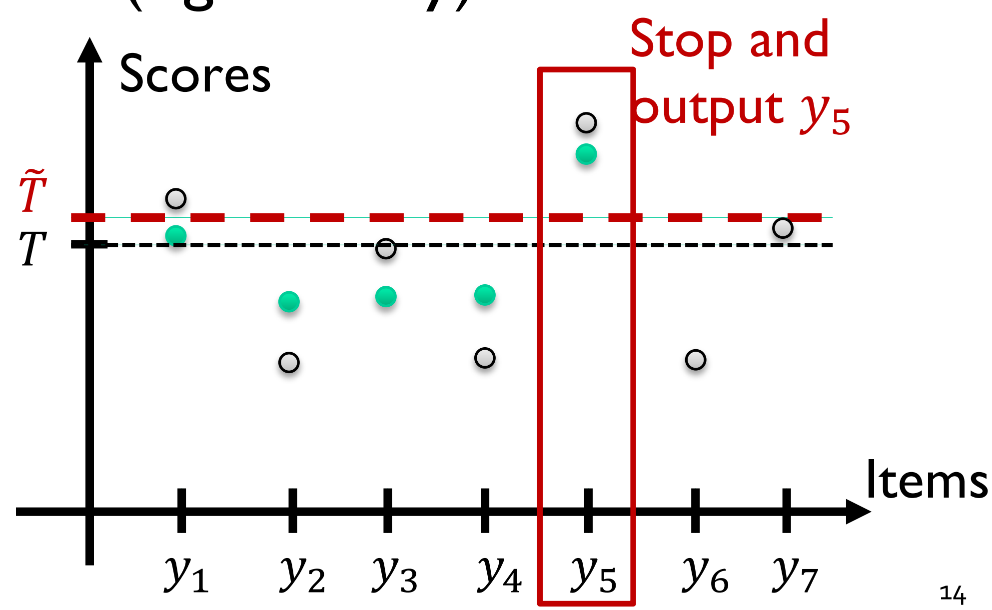
- Perspective used explicitly for
  - Learning discrete classifiers [KLNRS'08]
  - Synthetic data generation [BLR'08,...,HLM'10]
  - Convex Optimization [CM'08,CMS'10]
  - Frequent Pattern Mining [BLST'10]
  - Genome-wide association studies [FUS'11]
  - High-dimensional sparse regression [KST'12]
  - ...

About the Exponential Mechanism

- ExpMech is "Gibbs sampling"
  - Maximizes expected score subject to entropy constraint
- Alternative Implementation: "Report Noisy Max"
  - Add noise $Lap\left(\frac{\Delta}{\epsilon}\right)$ to each score
  - Report argmax of noisy scores
    - Basically the same distribution as Gibbs!
- Lower bound
  - Every (ϵ, δ)-DP algorithm, in worst case, outputs $\hat{y}$ with $q(\hat{y}; x) \leq OPT_x - \Omega\left(\frac{\Delta \ln(|Y|)}{\epsilon}\right)$.
- Generalizations
  - "Online" version ("sparse vector technique")
  - Variants do much better on specific classes of inputs
  - Can handle scores with different sensitivities smoothly

Sparse Vector Technique [RR'10, HR'10]

- "Online" variant of exponential mechanism
- Input:
  - Data set x
  - Public score $q(\cdot;\cdot)$, threshold T
  - Set Y arrives as a public sequence $y_1, y_2, \ldots$ with private scores $q(y_j; x)$
- Goal:
  - Output the first item with score (significantly) above T
- Algorithm
  - $\tilde{T} \leftarrow T + Lap\left(\frac{4\Delta}{\epsilon}\right)$
  - For $i = 1,2, \ldots, |Y|$:
    - $\tilde{q}_i \leftarrow q(y_i; x) + Lap\left(\frac{2\Delta}{\epsilon}\right)$
    - If $\tilde{q}_i > \tilde{T}$, **stop** and output y_i



Linear Queries

Case Study

Collections of linear queries

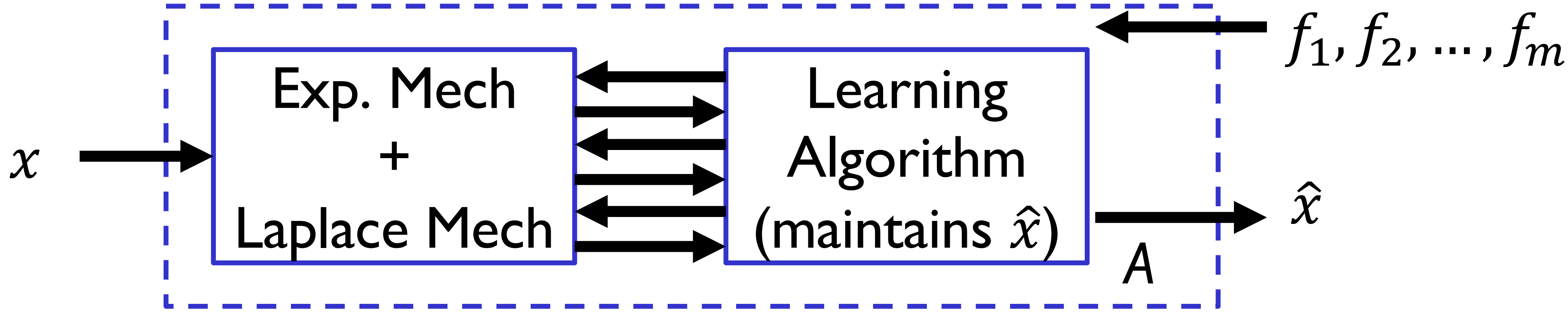
Could also look at ℓ_1 or ℓ_2 errors

- Data is a multi-set in domain D
- Represented as a histogram $\vec{x} \in \mathbb{N}^{|D|}$ where $x(i) = (\#\text{ occurrences of } i \text{ in } x)$
- Linear query is given by a function $f: D \to [0,1]$,
  - Answer to f on x is $\sum_{i \text{ in } x} f(i) = \langle f, x \rangle$
- Goal: Given a workload of queries $f_1, \dots, f_m$, release $\hat{f}_1, \dots, \hat{f}_m$ to minimize $\alpha = \frac{1}{n} \max_j \left| \hat{f}_j - \langle f_j, x \rangle \right|$
  - Captures releasing collections of contingency tables, means, covariance matrices, etc
- How low can the error be
  - in terms of $n, m, |D|$?
  - for a particular collection of queries?

Error bounds for linear queries

- Goal: Given $f_1, \ldots, f_m$, minimize $\alpha n = \max_j \left|\hat{f}_j - \langle f_j, x\rangle\right|$
  - Alternately, find n necessary for given error α
- Laplace mechanism + composition results
  - Require $n \geq O\left(\frac{m \log m}{\alpha\epsilon}\right)$ or $n \geq O_\delta\left(\frac{\sqrt{m \log m}}{\alpha\ \epsilon}\right)$
  - Best possible when $n \gg m$
  - Time $O(mn)$
- "Learn the data" paradigm [BLR'08, DNNRV'09, RR'10,HR'10,HLM'11]
  - $n \geq O\left(\frac{\log^3(m\cdot|D|)}{\epsilon\alpha^3}\right)$ or $n \geq O_\delta\left(\frac{\sqrt{\log m}\cdot\log|D|}{\alpha^2\ \epsilon}\right)$
    - Allows exponentially many queries ☺
  - Time $O(mn|D|)$
    - Can be exponential ☹

Idea: "Learn the data" [DNNRV'09, HR'10]



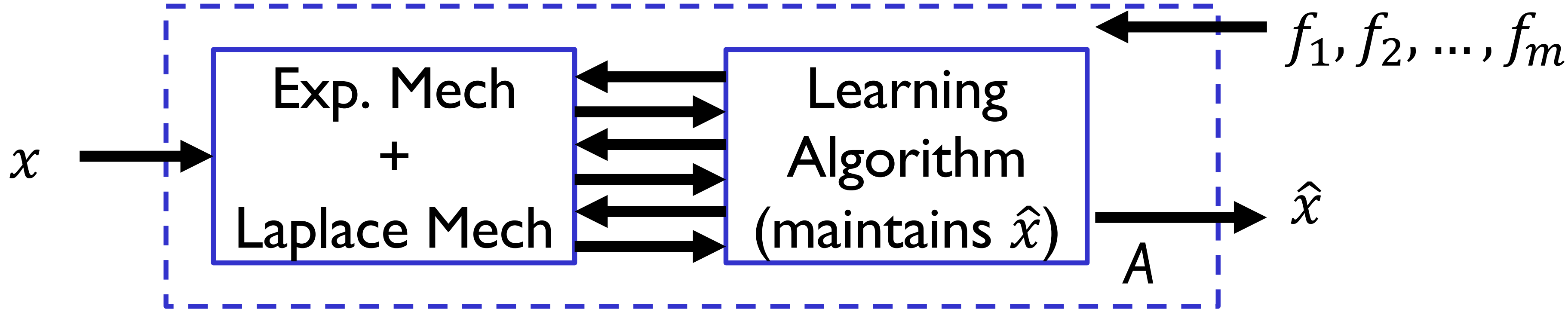
Release mechanism learns a "model" of x through DP interface

- Search for $\hat{x}$ to minimize $error(\hat{x}) = \max_j \left| \langle f_j, \hat{x} \rangle - \langle f_j, x \rangle \right|$
  (Generally do not get $\hat{x} \approx x$)

Traditional learning	"Learn for privacy"
Parameters of classifier	Data model $\hat{x}$
Training data	User's queries f_j
Gradient computations	Actual data access

- Learner computes a sequence of estimates $\hat{x}_0, \hat{x}_1, \hat{x}_2, \ldots$
- Gradient: $\nabla error(\hat{x}_t) = \pm f_{j^*}$
  where $j^* = argmax_j \left| \langle f_j, \hat{x}_t \rangle - \langle f_j, x \rangle \right|$

"Learn the data" as a game



- Can think of this as a two-player game
  - Learner plays generative model $\hat{x}$
  - Data holder uses DP algorithm to find query that distinguishes $\hat{x}$ from real data x
- Similar to generative adversarial networks (GANs)
- Game perspective leads to current best algorithms for creating synthetic data, e.g.
  - E.g. [Gaboardi, Arias, Su, Roth, Wu 2014, Beaulieu-Jones, Wu, Williams, Greene, 2017, Boob, Cummings, Kimpara, Tantipongpipat, Waites, Zimmerman, 2019, McKenna, Sheldon, Miklau 2019, Jordon, Yoon, van der Schaar, 2019]

"Geometric" approaches I

- Consider matrix W with columns $f_1, \dots, f_m$
  - Goal: find $\hat{w}$ such that $\|\hat{w} - Wx\|$ is small
- Define sensitivity polytope [Hardt Talwar 10]
$$K = conv(\pm f_1, \dots, \pm f_m)$$
- Observe:
  - Sensitivity: for x, x' neighbors, $Wx - Wx' \in K$
  - Range: if x has n records, then $Wx \in n \cdot K$
- This suggests two general approaches
  - [HT'10] Release noise scaled to K-norm: $A(x) = Wx + Z$ where $p_Z(y) \propto \exp(-\epsilon \|y\|_K)$ and $\|y\|_K = \min\{r \geq 0 : y \in r \cdot K\}$
  - Projection [Nikolov Talwar Zhang '13]:
$$A(x) = Proj_{nK}(Wx + noise)$$
    where $Proj_{nK}(z) = argmin\{\|y - z\|_2 : y \in n \cdot K\}$
- Variations on these are known to be (close to) optimal in several settings [HT'10,BDKT'12,NTZ'13]

"Geometric" approaches II

- Given W, "matrix mechanism" [Li, Miklau, Hay, McGregor, Rastogi, 10] and follow-ups have 3 stages:
  - Select "good" $k \times |D|$ matrices A and B such that $W = BA$
  - Measure $y = Ax + Z$ where Z is Laplace/Gaussian
  - "Reconstruct" $\tilde{w} = By$
  - Output projected value $\hat{w} = Proj_{nK}(\tilde{w})$

- Selection of A, B depends on kind of noise and error goal
  - For Gaussian noise and ℓ_2 error, objective is $||B||_{Fr} \cdot ||A||_{1\to 2}$
    - $||B||_{Fr}$ is sum of squared entries
    - $||A||_{1\to 2}$ is maximum norm of columns in A

- Basis of current Census implementations

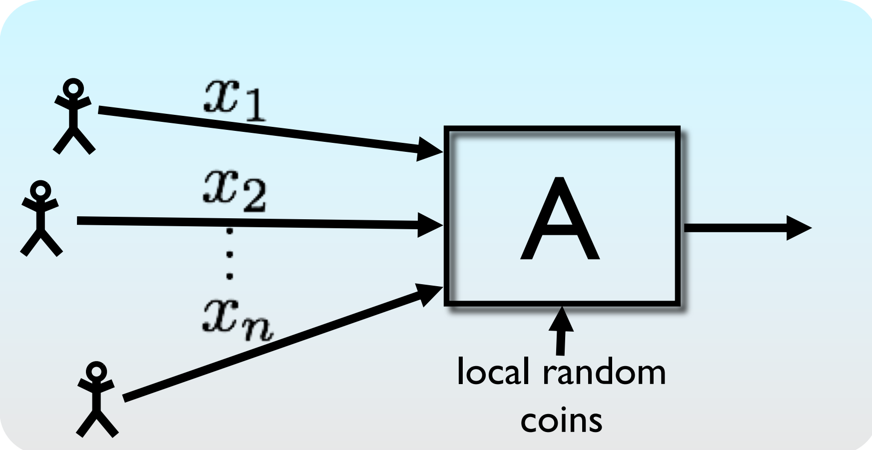
DIFFERENTIAL PRIVACY

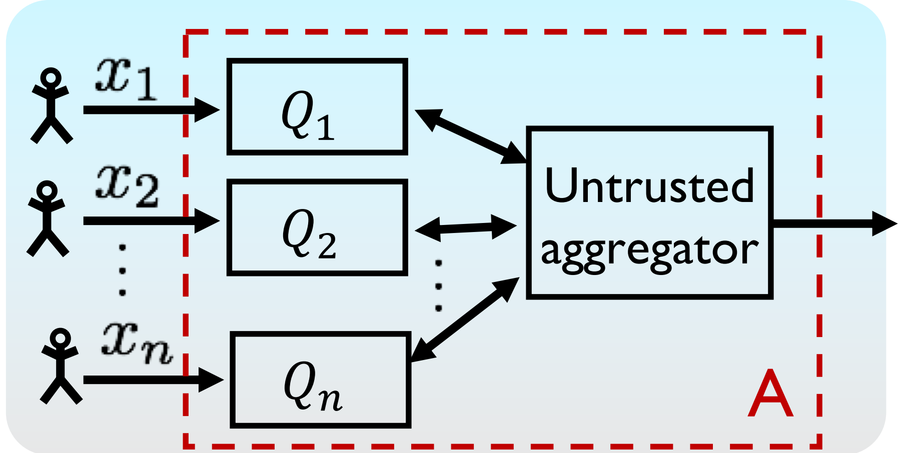
- Episode III: Attack of the Codes
  - Reconstruction attacks
  - Membership attacks
- Episode IV: A New Hope
  - Differential privacy
- Episode VI: Return of the Algorithms
  - Algorithms for counting queries
  - Optimization and learning
- Episode VII: The Connections Awaken
  - Learning and adaptive data analysis
  - Statistics
  - Game theory
  - Law and policy

The Local Model for Differential Privacy

Local Model for Privacy

Equivalent to [Efvimievski, Gehrke, Srikant '03]





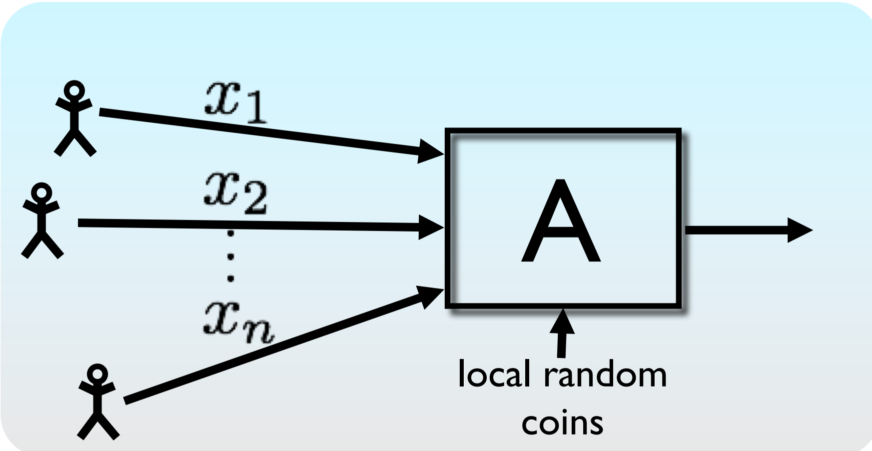
- "Local" model
  - Person i randomizes their own data
  - Attacker sees everything except player i's local state

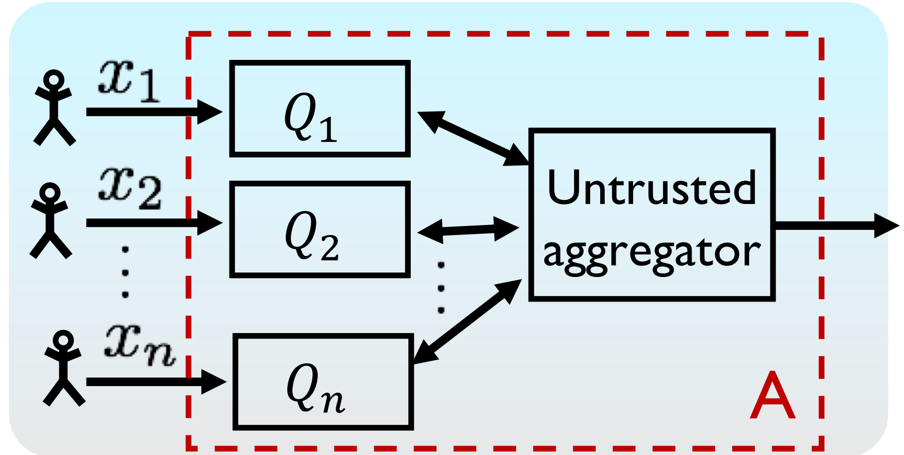
- Definition: A is ϵ-locally differentially private if for all i:
  - for all neighbors $\boldsymbol{x}$, $\boldsymbol{x}'$,
  - for all local coins $\boldsymbol{r}_{-i}$ of all other parties,
  - for all transcripts t:

$$\Pr_{\text{coins } r_i}(A(\boldsymbol{x}, r_{-i}) = t) \leq e^{\epsilon} \cdot \Pr_{\text{coins } r_i}(A(\boldsymbol{x}', r_{-i}) = t)$$

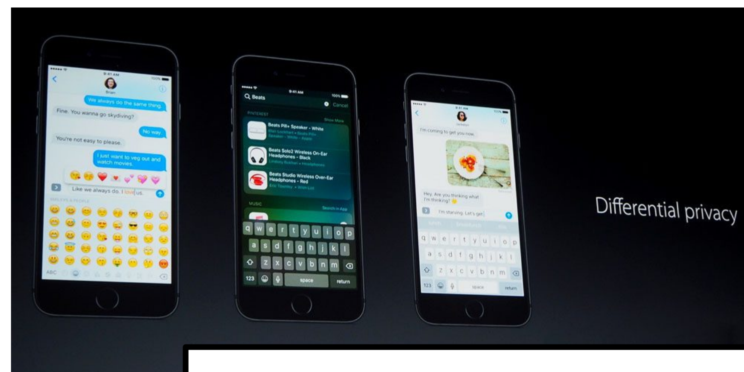
$\delta = 0$
w.l.o.g.
[BNS17]

Local Model for Privacy



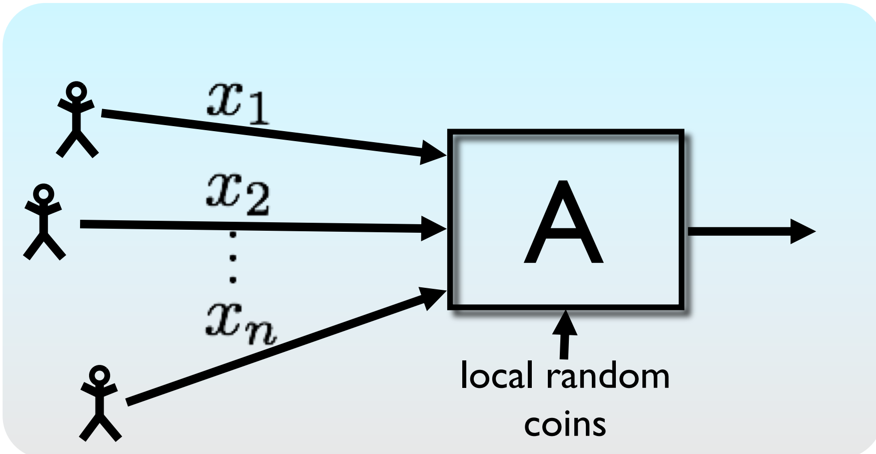


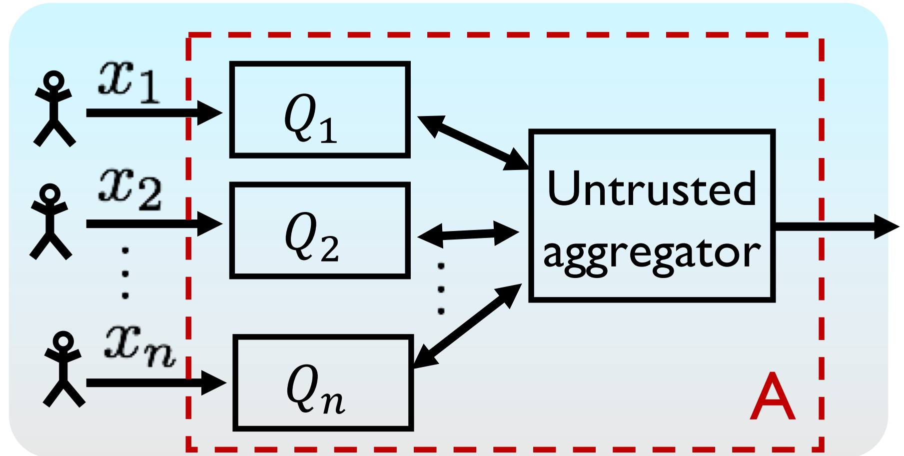
- Pros
  - No trusted curator
  - No single point of failure
  - Highly distributed
- Cons
  - Lower accuracy



Chrome

Local Model for Privacy





What can and can't we do in the local model?

Example: Randomized response

- Each person has data $x_i \in \mathcal{X}$
  - Analyst wants to know average of $f: \mathcal{X} \to \{-1,1\}$ over $\boldsymbol{x}$
  - E.g. "what is the fraction of diabetics"?
- Randomization operator takes $y \in \{-1,1\}$:

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

ratio is e^{ϵ}

- Observe:
  - If $c_{\epsilon} = \frac{e^{\epsilon}+1}{e^{\epsilon}-1}$, then $\boldsymbol{E}\left(c_{\epsilon} \cdot Q(y)\right) = y$
- How can we estimate a proportion?
  - $A(x_1, \ldots, x_n) = \frac{1}{n} \sum_i c_{\epsilon} \cdot Q\left(f(x_i)\right)$

Contrast with $\frac{1}{n\epsilon}$ in central model (via Laplace noise)

- **Proposition:** $\boldsymbol{E}\left|A(\boldsymbol{x}) - \frac{1}{n}\sum_i f(x_i)\right| \leq \frac{c_{\epsilon}}{2\sqrt{n}} \approx \frac{1}{\epsilon\sqrt{n}}$.

What can we do using noisy averages?

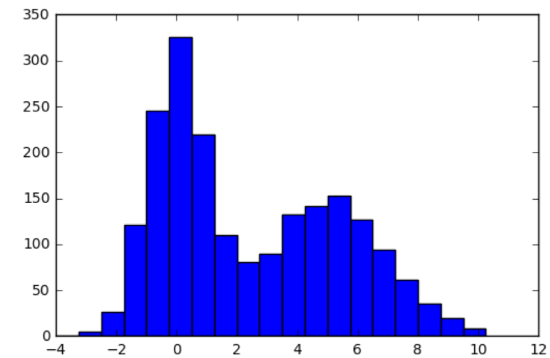
- An SQ algorithm interacts with a data set by asking a series of "statistical queries"
  - Query: $f: \mathcal{X} \to [-1,1]$
  - Response: $\hat{a} \in \frac{1}{n}\sum_i f(x_i) \pm \alpha$ where α is the **tolerance**
- Huge fraction of basic learning/optimization algorithms can be expressed in SQ form [Kearns 93]

- **Theorem** (follows [Blum Dwork McSherry Nissim '05]): Every q-query SQ algorithm with tolerance α can be simulated by ϵ-LDP protocol when $n \geq \frac{q \ln q}{\alpha^2 \epsilon^2}$.

Central model: $n \approx \frac{\sqrt{q \ln q}}{\alpha \epsilon}$

Histograms

[Mishra Sandler 2006, Hsu Khanna Roth 2012, Erlingsson, Pihur, Korolova 2014, Bassily Smith 2015, …]



- Every participant has $x_i \in \{1,2,\ldots,d\}$.
- Histogram is $h(x) = (n_1, n_2, \ldots, n_d)$ where $n_j = \#\{i: x_i = j\}$
- Straightforward protocol: Map each x_i to indicator vector e_{x_i}
  - So $h(x) = \sum_i e_{x_i}$
  - $\boldsymbol{Q'(x_i)}$: Apply $Q(\cdot)$ to each entry of e_{x_i}.

$$\downarrow x_i$$

$$e_{x_i} = (0,0,\ldots,0,1,0,\ldots,0)$$

$$Q'(e_{x_i}) = (Q(0),\ldots,Q(1),\ldots,Q(0))$$

- **Proposition:** $Q'(\cdot)$ is 2ϵ-LDP and

$$\boldsymbol{E}\left\|\sum_i Q'(x_i) - h(x)\right\|_\infty \le \frac{\sqrt{n\log d}}{\epsilon}$$

optimal

Central: $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$

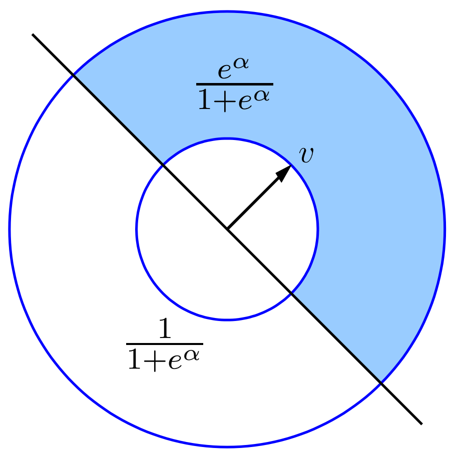
Succinctness

- Randomized response has optimal error $\frac{\sqrt{n \log d}}{\epsilon}$
  - Problem: Communication, time, and server memory $\Omega(d)$
  - How much is really needed?

- **Theorem** [Bassily, Nissim, Stemmer, Thakurta '17, Bun, Nelson, Stemmer '18]: Protocol with
  - optimal error,
  - $\tilde{O}\left(\epsilon\sqrt{n \log d}\right)$ space,
  - $\tilde{O}(n)$ total time
- Upper bound idea:
  - Connection to "heavy hitters" algorithms from streaming [Hsu, Khanna, Roth '12]
  - Two data structures:
    - estimate individual frequencies
    - Identify heavy hitters
- Experimental evaluation [cites above + Wang, Li, Jha '18]

Vector averages *[Duchi Jordan Wainright '13]*

- Suppose each input is a vector $x_i \in \mathbb{R}^d$ with $\|x_i\|_2 \leq 1$
  - ➢ How can we estimate $\frac{1}{\mathrm{n}} \sum_i x_i$?
- Use rand. response for each of the d coordinates?
  - ➢ Use $\frac{n}{d}$ players to estimate each coordinate.
  - ➢ Error $\sqrt{d/n\epsilon^2}$ per coordinate.
  - ➢ Total ℓ_2 error $E\left\|A(x) - \frac{1}{n}\sum_i x_i\right\|_2 \leq d/\sqrt{n\epsilon^2}$
- **Theorem** [DJW'13]: Can estimate to error $\sqrt{d/n\epsilon^2}$.
- Idea: Let B_d = unit ball in $\mathbb{R}^d$
  - ➢ $R(v)$ samples uniformly from either
    - $\{u \in B_d: \langle u, v\rangle \geq 0\}$ w.p. $e^\epsilon/(1+e^\epsilon)$, or
    - $\{u \in B_d: \langle u, v\rangle < 0\}$ w.p. $1/(1+e^\epsilon)$.
  - ➢ If $\|x\|_2 = 1$, then $E\big(R(x)\big) = c_{\epsilon,d} \cdot x$ where $c_{\epsilon,d} = \Theta\big(\epsilon/\sqrt{d}\big)$.



Limitations of Local Algorithms

SQ algorithms and Local Privacy [KLNRS'08]

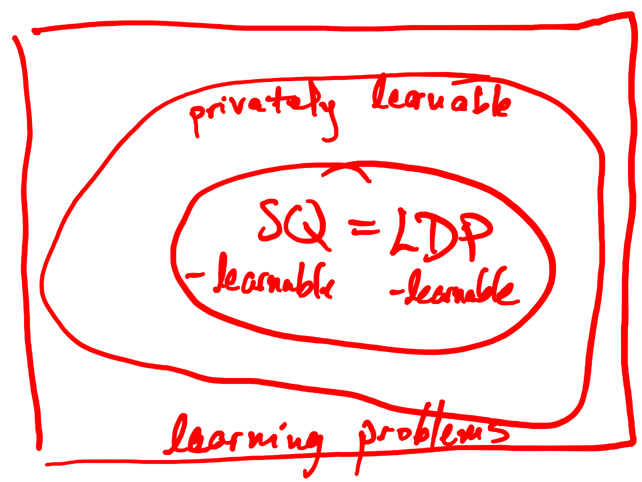
- Every SQ algorithm can be simulated by a LDP protocol.

- **Theorem:** Every LDP algorithm that assumes i.i.d. data can be simulated by SQ with $q \approx n$ and $\alpha \approx 1/n$

- **Corollary** (via [Kearns'93])**:** No LDP algorithm can learn parity with polynomially many samples ($n = 2^{\Omega(d)}$).

- "Learn parity" $\approx$ distinguish between n samples from either
  - Uniform on $\{0,1\}^d$, or
  - Uniform on $\{x \in \{0,1\}^d : x \odot z = 0 \; mod \; 2\}$ where z is a secret in $\{0,1\}^d$.

- **Theorem:** Centralized DP can learn parity with $n = O\left(\frac{d}{\epsilon}\right)$ samples.
  - "Simpler" exponential separation now known [Duchi, Jordan, Wainright'13, Ullman'17]



SQ Algorithms simulate LDP protocols

- Roughly:
  Every LDP algorithm with n data points can be simulated by an $O(n)$-query SQ algorithm with
  - Actually a distributional statement: assume that data drawn i.i.d from some distribution P

- Key piece: Transform the randomizer so only 1 bit is sent to aggregator by each participant
  - Use rejection sampling to get right distribution

- Corollary [Bun, Nelson, Stemmer'18]: In the local model,

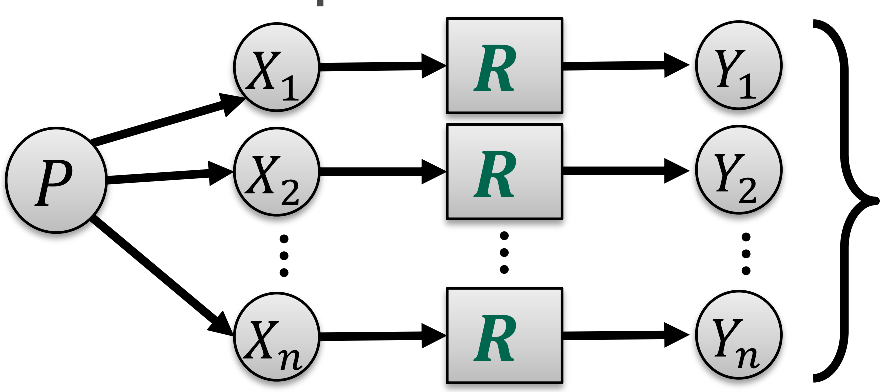
$$(\epsilon, 0)\text{-DP} \approx (\epsilon, \delta)\text{-DP}$$

Information-theoretic lower bounds

- For local DP algorithms, easiest arguments use information-theoretic framework
  [Beimel,Nissim,Omri'10, Chan,Shi,Song'12, Duchi,Jordan,Wainwright'13]
  - Tight lower bounds for many basic estimation tasks
- **Theorem:** If A is (ϵ, δ)-locally DP, then

$$\boldsymbol{E}\left|A(\boldsymbol{x}) - \frac{1}{n}\sum_i f(x_i)\right| = \boldsymbol{\Omega}\left(\mathbf{1}/\boldsymbol{\epsilon}\sqrt{\boldsymbol{n}}\right)$$

- Idea:
  - Suppose $X_1, \ldots, X_n \sim P$ i.i.d., where P is randomly chosen
  - Show that protocol leaks little information about P



Lemma: For every distribution on P,
$I(P; Y_1, \ldots, Y_n) \leq n\epsilon^2$

Main Lemmas

- **Lemma:** If R is ϵ-DP, then $I\big(X; R(X)\big) \le O(\epsilon^2)$

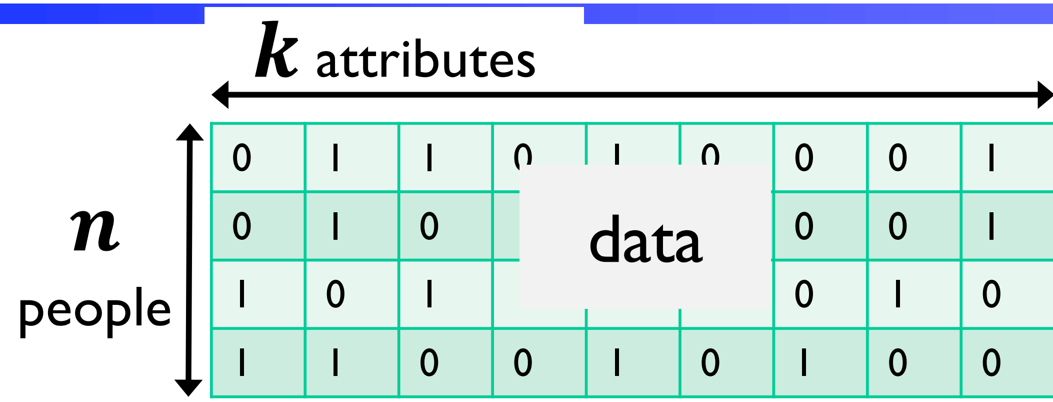
- **Stronger Lemma:** If R is ϵ-DP, and

$$W(x) = \begin{cases} x & w.p. \quad \alpha \\ 0. & w.p. 1 - \alpha \end{cases},$$

then $I\big(X; R(W(X))\big) \le O(\alpha^2 \epsilon^2)$.

- To prove $1/\epsilon\sqrt{n}$ lower bound for counting query:
  - Show that algorithm with error α leaks $\le n\alpha^2\epsilon^2$ bits
  - To estimate P, need to learn at least one bit
  - So error $\alpha \ge 1/\epsilon\sqrt{n}$

Selection Lower Bounds [DJW'13, Ullman '17]



- Suppose each person has k binary attributes
- **Goal**: Find index j with highest count $(\pm\alpha)$

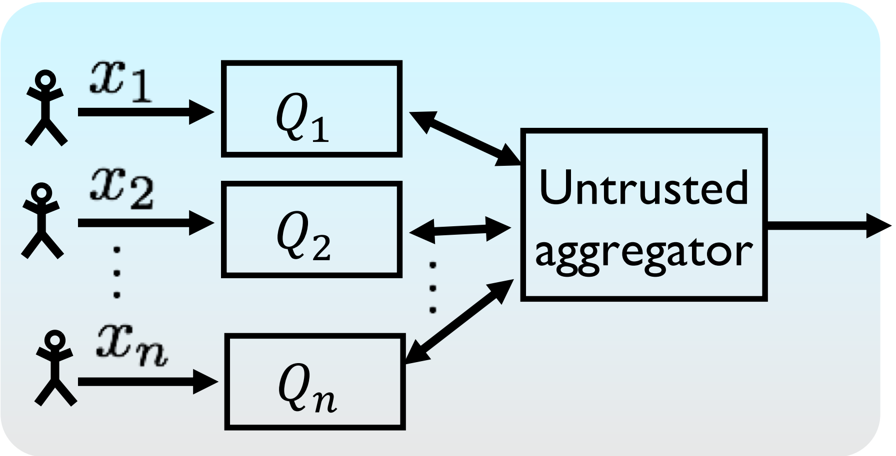
- **Central model**: $n = O(\log(k)/\epsilon\alpha)$ suffices
  [McSherry Talwar '07]
- **Local model:** Any **noninteractive** local DP protocol with nontrivial error requires

$$n = \Omega(k\log(k)\,/\epsilon^2)$$

  - [DJW'13, Ullman '17]
  - (No lower bound known for interactive protocols)

What about interaction?

- Simplest protocols have just 1 message
  - Q_i is known to player i at start of protocol
  - Called "noninteractive"

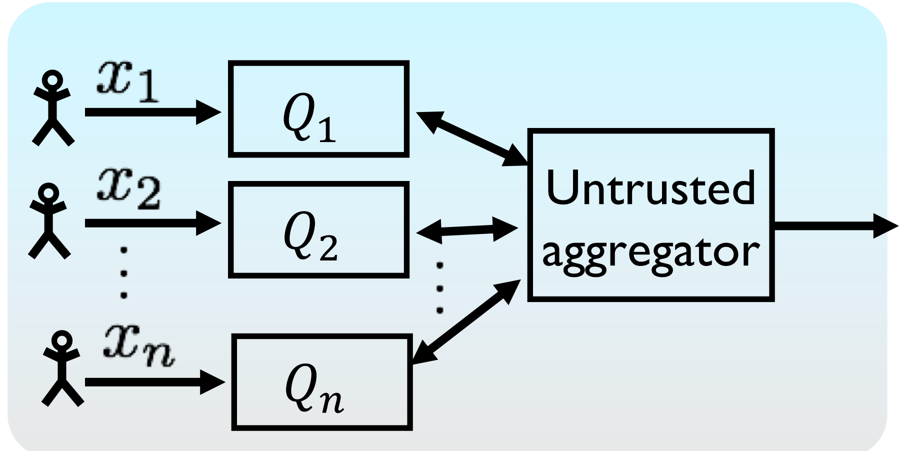


- But some protocols are interactive
  - Server might talk to each player several times
  - Server may choose Q_2 based on $Q_1(X_1)$

- Interaction is expensive
  - Latency
  - Aggregator must be online

Interaction is necessary for LDP

- [KLNRS08]
  For "hidden parity" problem,
  noninteractive LDP requires exponentially more data
  than 2-round LPD



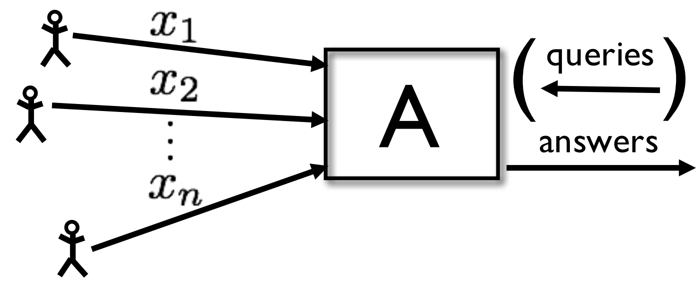
  - Proof by separating adaptive SQ from nonadaptive SQ
  - Stronger separations now known [Feldman 2019, Joseph-Mao-Roth 2019]

- Is interaction useful in practice?
  - Known protocols for convex optimization use lots of interaction [DJW'13,STU '17]
  - Lower bounds known for a subclass of protocols [**S**TU'17, McMahan, Srebro, **S.** ,Wang, Woodward'18]

Differential Privacy and Game Theory

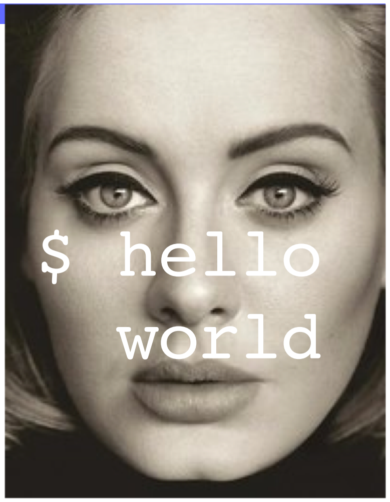
Game-theoretic interpretation of DP

- Suppose that person i is deciding whether to contribute data to a data set
- Different outputs of A have different **utility** to i
  - $U_i = u_i\big(A(X)\big)$



- If A is (ϵ, δ)-differentially private, then
  $\mathbf{E}(U_i | \text{person } i\text{'s data is used})$
  $\leq e^{-\epsilon}\mathbf{E}(U_i | \text{person } i\text{'s data is } \textbf{not} \text{ used}) - \delta$
- Implications
  - Participating in a study costs me little
  - [McSherry-Talwar '07] Every differentially private algorithm is approximately truthful
    - Little incentive to misreport values

Digital Good Auction [MT '07]



- 1 seller with a digital good
- n potential buyers
  - Each has a secret value v_i in [0,1] for song
  - Setting price p will get revenue $rev(p) = p|\{i: vi \geq p\}|$
  - How can seller set p to get revenue ≈ $OPT = \max rev(p)$?
- Straightforward bidding mechanism
  - Each player reports $v_i' = 0$
  - Lying can drastically change best price
- Instead, sample p^* from $r(p) \propto \exp(\epsilon \cdot rev(p))$
  - Approximately truthful
  - Expected revenue $\geq OPT - O\left(\frac{\ln(\epsilon n)}{\epsilon}\right)$

Economic Theory & Differential Privacy

- Mechanism Design
  - ➢ Twin goals
    - Incentive-compatibility
    - Incentive-compatibility
  - ➢ Exactly truthful mecha
    Orlandi, Smorodinsky '12, Ch
- "Pricing" privacy [Ghosh
  - ➢ Can we reward survey
  - ➢ Can we use prices to elicit values for privacy?
  - ➢ "Endogenize" ε?
- How can private information change games?
  - ➢ Equilibrium selection [Kearns, Pai, Roth, Ullman '14, Rogers Roth '14]
- Sensitive information as a public good
  - ➢ How can we decide how to use the privacy budget?

See videos of tutorials by
- Katrina Ligett at Simons Institute January 2019 "bootcamp"
- Aaron Roth at 2012 DIMACS Workshop on Differential Privacy

Law, Policy and Differential Privacy

From Law to Technical Definitions

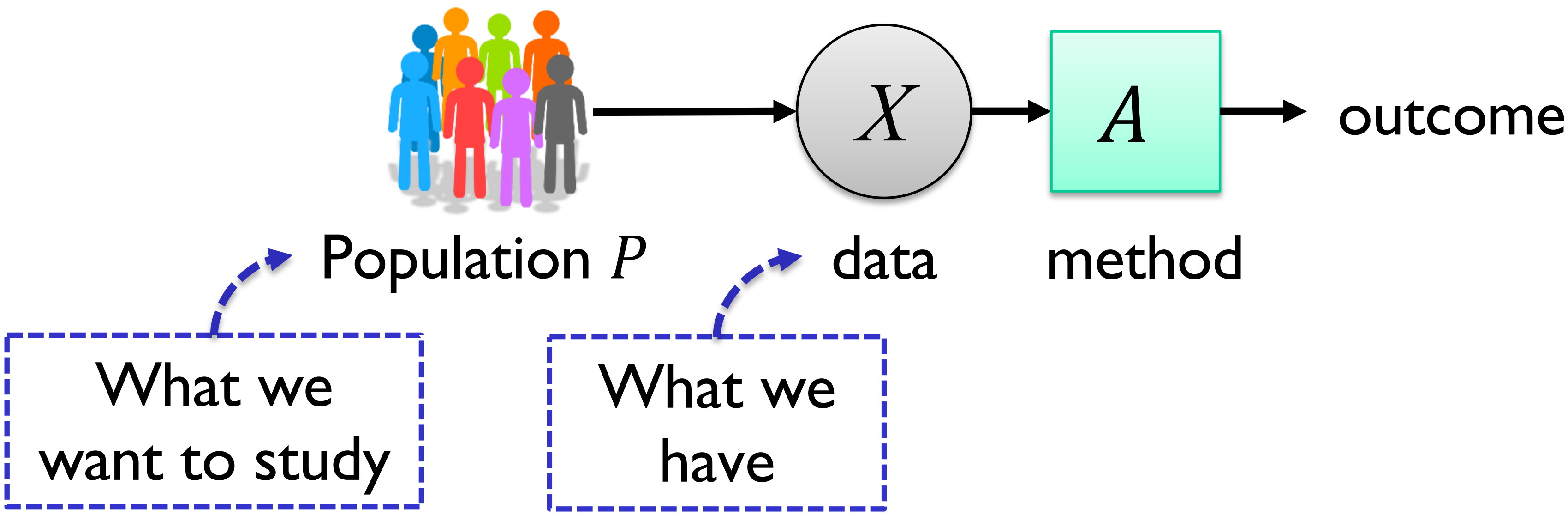
Two central challenges

1. Given a body of law and regulation, what technical definitions comply with that law?
   - E.g., GDPR
2. How should we write laws and regulations so they make sense given evolving technology?
   - E.g., Surveillance ≠ physical wiretaps

- Technical research must inform these questions
  - E.g. "personally identifiable information" is meaningless
- [Nissim et al. 2016] Mathematical formulations play an important role
  - E.g. formal interpretation of FERPA (a US law) mirrors DP
  - "Singling out" in GDPR is challenging to make sense of

Adaptive Data Analysis

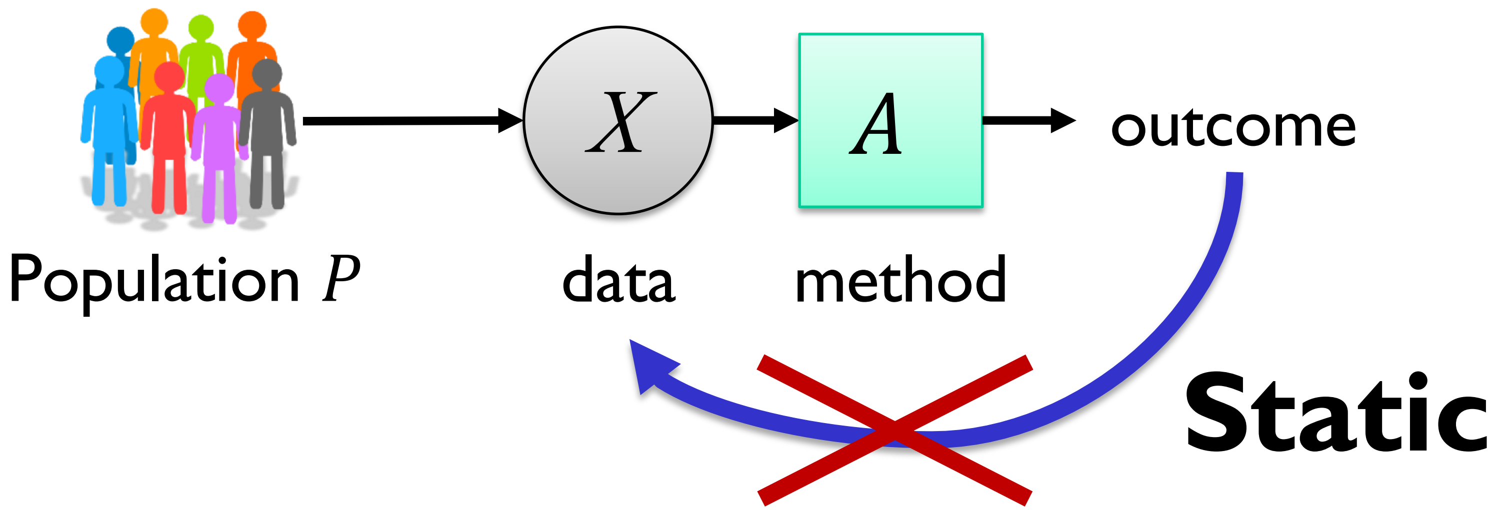
Overfitting



- **Inference:** Draw conclusions about P based on X

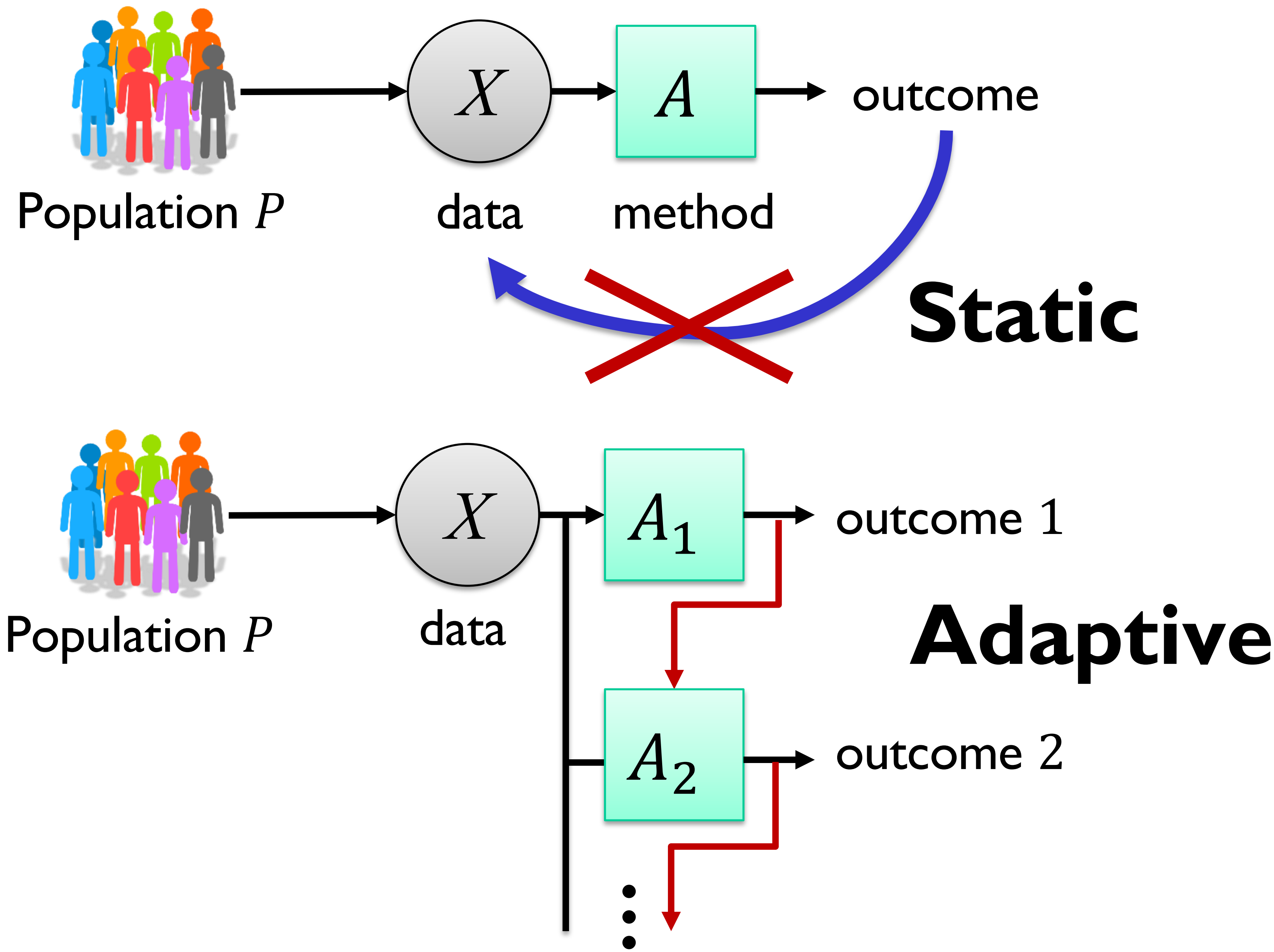
- **Overfitting / false discovery:**
  Conclusions that hold for X but not for P

Overfitting



- Decades of work on preventing overfitting
  - Cross-validation, bootstrap, multiple hypothesis testing, FDR control, …
- Designed for static data analysis
  - Assumes method selected independently of data

Overfitting



Adaptivity is common

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

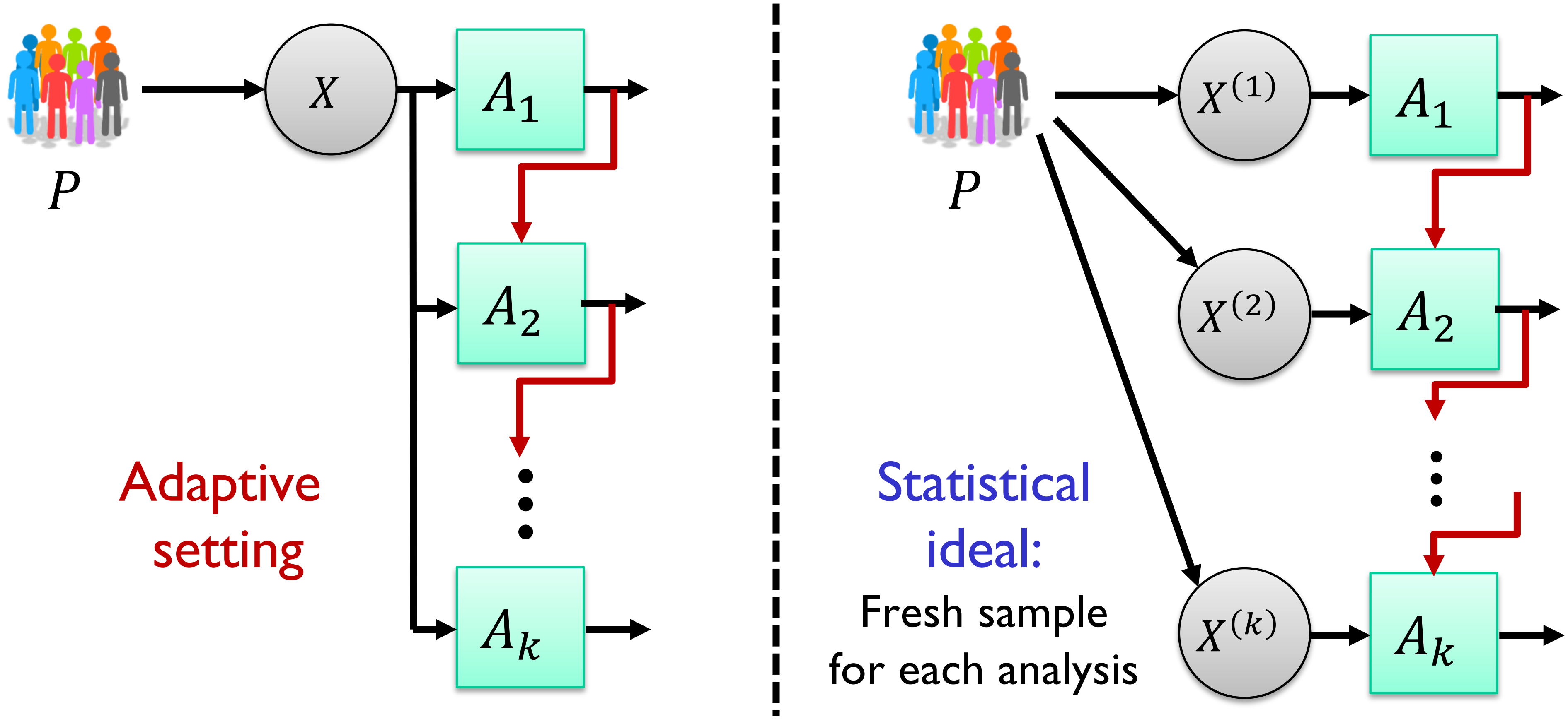
There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mis- a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed

This *multiple comparisons* issue is well known in statistics and has been called "*p*-hacking" in an influential 2011 paper by the psychology re-

American Scientist, 2014

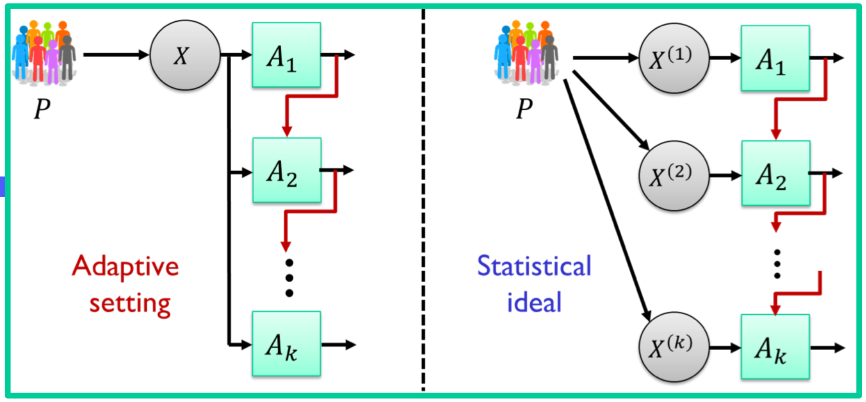
How can we provide statistically valid answers to adaptively selected analyses?

Getting a Baseline



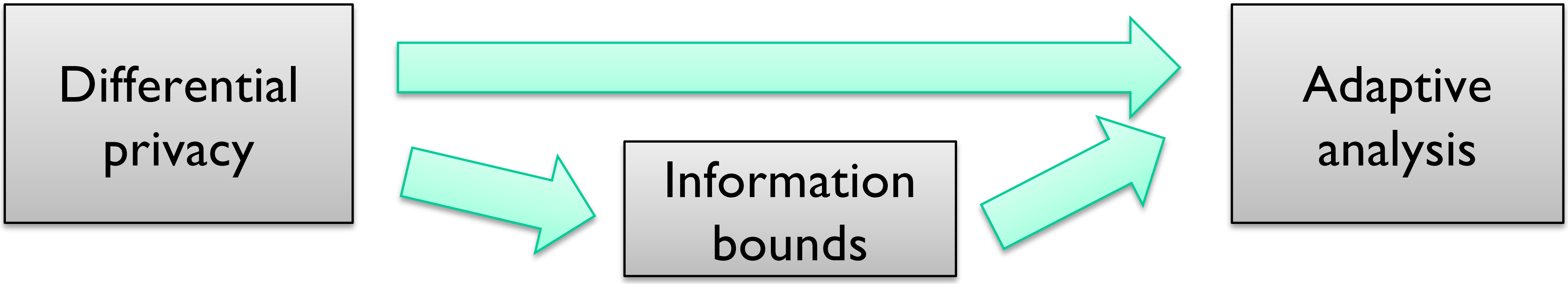
- Goal: Relate adaptive setting to statistical ideal worlds
- Understand how properties of algorithms $A_1, A_2, \ldots$ affect that relationship

Privacy and overfitting



Folklore: Differential privacy don't overfit

- Recent discovery: DP prevents adaptive overfitting [Dwork Feldman Hardt Pitassi Reingold Roth '15]
- Recent developments (my work and others…)
  - Tight connection between DP and overfitting
    - Best known bounds on accuracy
  - General information-theoretic framework
    - Unifies & generalizes known results



A few things I didn't tell you about

- Other algorithmic techniques
  - Local sensitivity, smoothed sensitivity, and Lipschitz extensions
  - Subsample and aggregate
- PAC learning
- Different access models
  - Continual release
  - Local privacy
  - Pan-privacy
- Computational notions
- Lower bounds
  - Accuracy (sometimes via information theory)
  - Computation time
- Programming tools
  - New developments in type theory
- DP in practice

Conclusions

- Define privacy in terms of my effect on output
  - Meaningful despite arbitrary external information
  - I should participate if I get benefit
- Rigorous framework for private data analysis
  - Rich algorithmic literature (theoretical and applied)
  - There is no competing theory
- What computations can we secure?
  - Differential privacy provided a surprising formalization for a previously ad hoc area
  - What other areas need formalization?