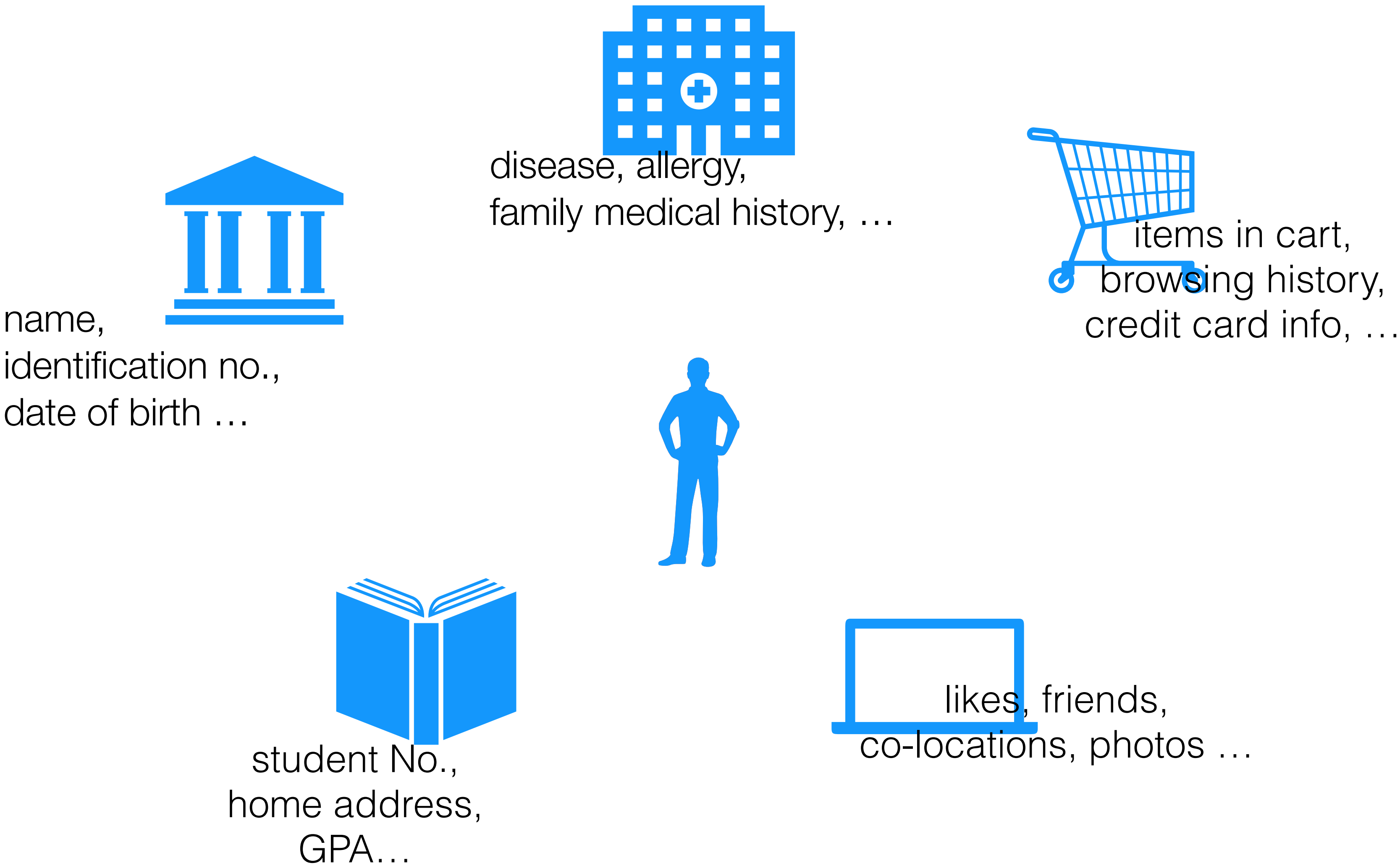
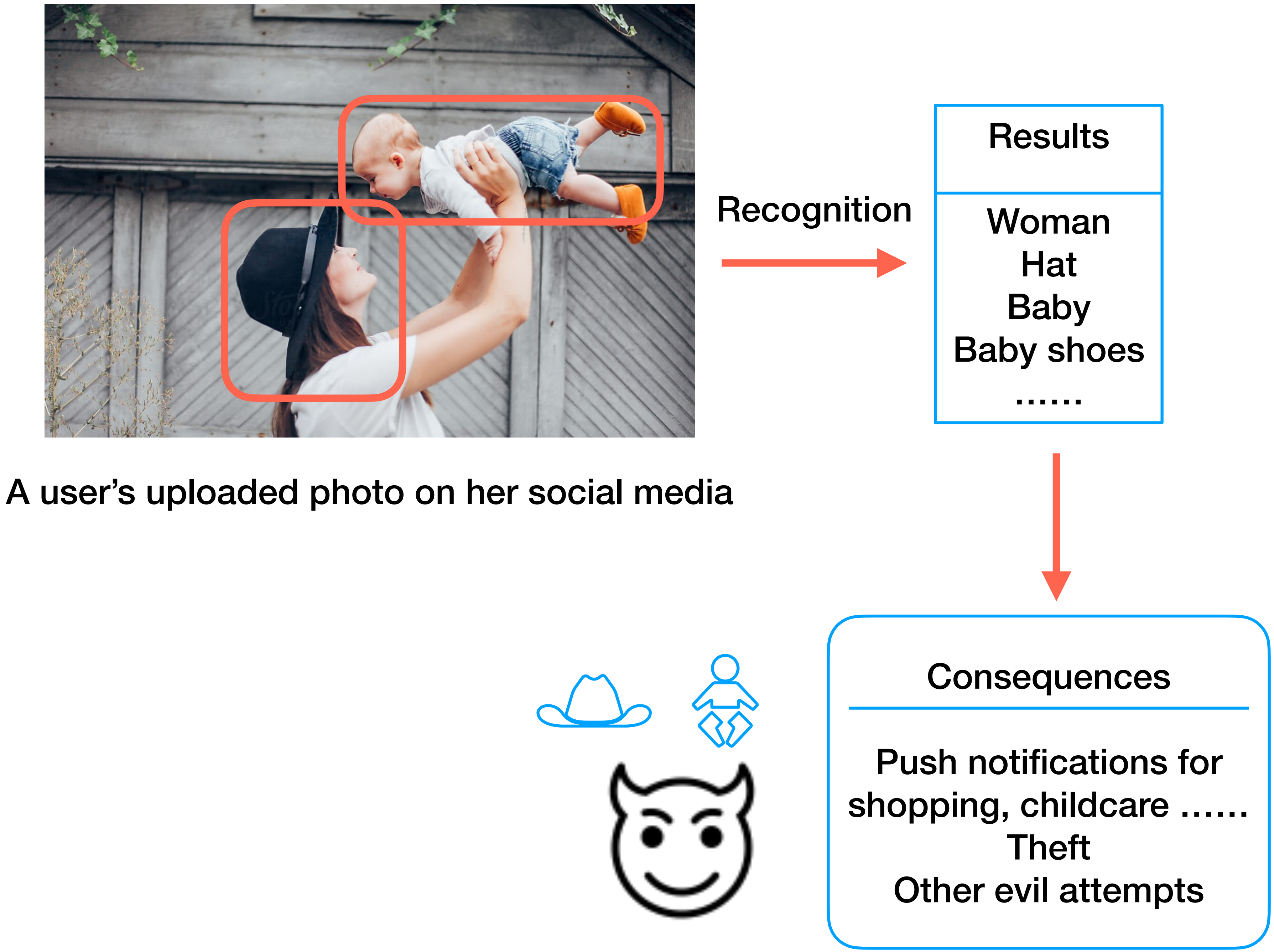
Privacy in Data Mining

Liyao Xiang
xiangliyao.cn
Shanghai Jiao Tong University

Big Data Era



Individual information is everywhere


Recognition
Results
Woman
Hat
Baby
Baby shoes
......
A user's uploaded photo on her social media
Consequences
Push notifications for
shopping, childcare ......
Theft
Other evil attempts

- Let’s take a look at some privacy violation cases

AOL Search Debacle

- AOL Research released a compressed text file
- Containing 20 million search keywords for over 650,000 users over a 3-month period intended for research purposes
- Personally identifiable information was present
- The New York Times was able to locate an individual from the released and anonymized search records by **cross referencing** them with phonebook listings

user-ct-test-collection-06.txt

```
1998497 anthony burger  2006-03-05 13:01:36     2       http://www.anthonyburger.com
1998497 gaither 2006-03-05 13:02:22     4       http://www.bill.gaither.com-music.homepages.org
1998497 allegiant air   2006-03-05 15:27:59     1       http://www.allegiantair.com
1998497 gaithe  2006-03-05 17:07:32
1998497 gaither 2006-03-05 17:07:44     7       http://www.gaither.com
1998497 gaithe  2006-03-05 17:09:53
1998497 gaither 2006-03-05 17:10:03     7       http://www.gaither.com
1998497 allegiant air   2006-03-05 18:22:26     1       http://www.allegiantair.com
1998497 disney coronado springs resort orlando fl       2006-03-07 14:09:08     5       http://hotels.about.com
1998497 www.hli.com     2006-03-10 09:05:39
1998497 heritage lottery international  2006-03-10 09:06:56     1       http://blog.supersurge.com
1998497 googlemaps.com  2006-03-11 00:12:28     1       http://www.googlemaps.com
1998497 amy grant       2006-03-11 19:29:34     7       http://www.mindspring.com
1998497 amy grant       2006-03-11 19:29:34     2       http://www.amygrant.com
1998497 amy grant       2006-03-11 19:29:34     5       http://en.wikipedia.org
1998497 david phelps    2006-03-11 19:33:55     1       http://www.davidphelps.com
1998497 imercer.com socil security      2006-03-12 13:58:18
```

aol search database

Ads by Google | Web Database | ASP Database | Access Query Example | ColdFusion Random | Police Records

Search Results

Your search returned **6454** hits. Displaying **1-30**.

User ID	Search Keywords	Date	Website
9461954	as of 2003 the fda had approved more than fifty drugs for the treatment of hiv aids or aids-related conditions. when taken the right way these drugs can drive the hiv virus below detectable levels. the bad news is that these drugs are very expensive and	2006-05-03 20:56:35	
2856400	antipsychotic drugs safe for patient with bradycardia	2006-03-26 09:46:41	http://www.drugs.com
20837908	getting high on otc drugs	2006-05-20 19:01:04	https://www.totse.com/en/drugs/otc/index.html
15737462	nude photos free boobs or breasts or tits -drugs boobs or breasts or tits -drugs boobs or breasts or tits -drugs boobs or breasts or tits -drugs boobs or breasts or tits -drugs	2006-04-16 22:57:15	http://www.aviationespace.net
21309272	prescription drugs and side effects	2006-05-06 23:25:29	http://www.drugs.com

NetFlix Privacy Lawsuit

- $1 million Netflix prize for movie recommendation challenge
- Netflix published 10 million movie rankings by 500,000 customers
- Anonymized by removing personal details and replacing names with random numbers
- Cancelled for customer privacy invasion
    - A woman sued Netflix, for Netflix made it possible for her to be identified
    - Researchers de-anonymized some of the Netflix data by comparing rankings and timestamps with public info in IMDb

https://arxiv.org/pdf/cs/0610105.pdf

Privacy Violation

- AOL search
- Netflix competition
- High-dimensional data is unique

Anonymity is NOT enough!
Linkage Attack

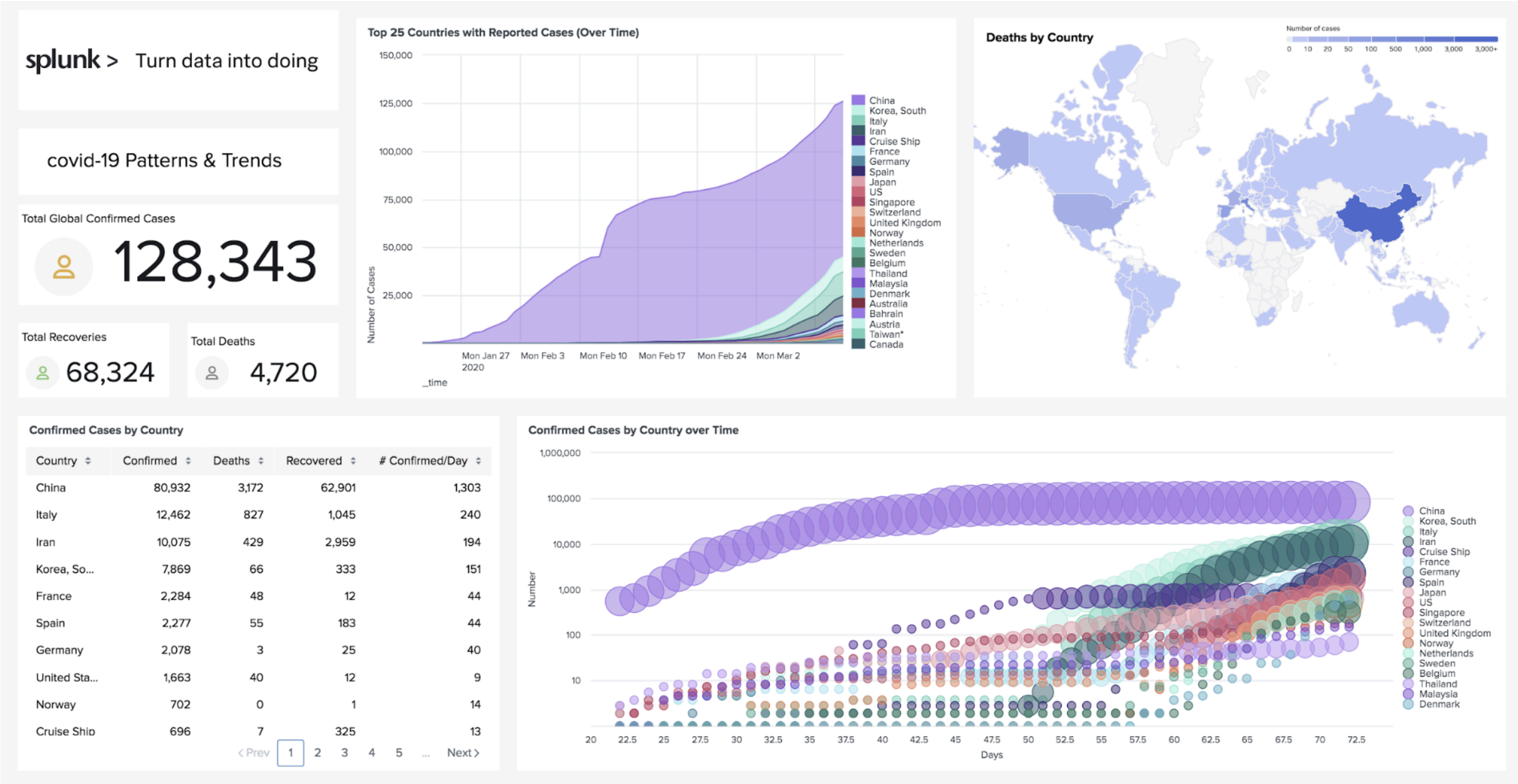
Example: John Center Employee Salary Table

Position	Gender	Depart.	Year of Entry	Teaching	Salary
Faculty	Female	John Center	2018	CS	—

One employee (Me) fits description!

Release Statistics

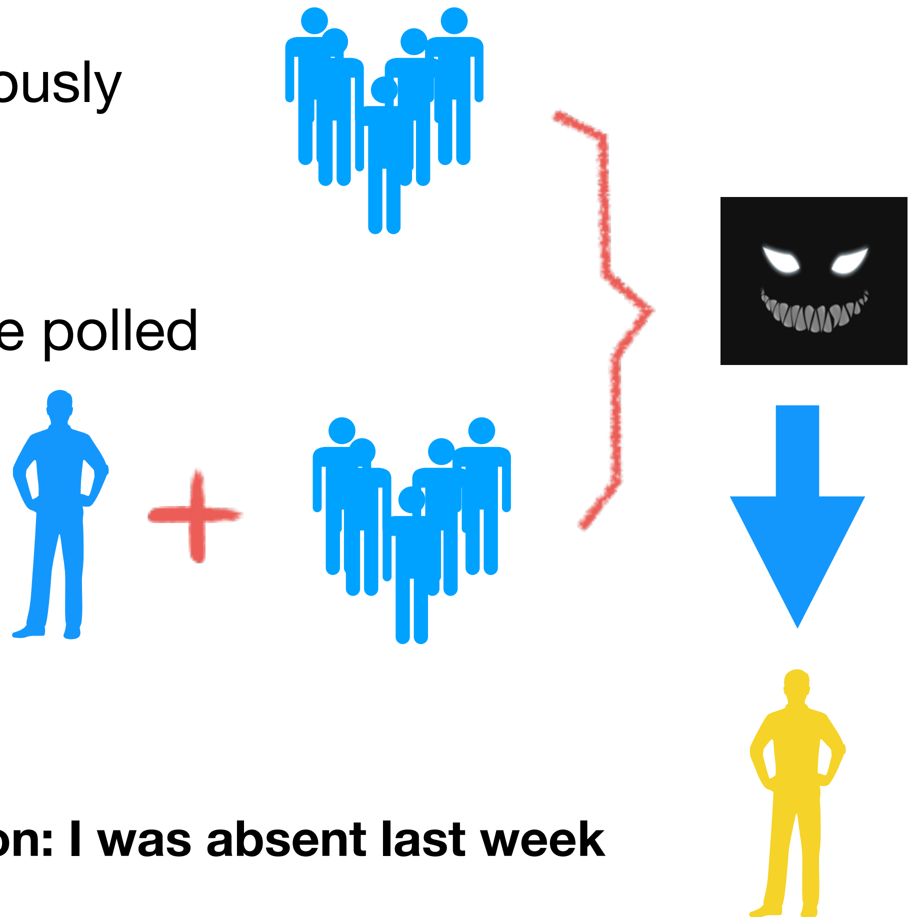
- Not release dataset. How about releasing statistics?
- Can the statistics be used to track an individual?



Country	Confirmed	Deaths	Recovered	# Confirmed/Day
China	80,932	3,172	62,901	1,303
Italy	12,462	827	1,045	240
Iran	10,075	429	2,959	194
Korea, So...	7,869	66	333	151
France	2,284	48	12	44
Spain	2,277	55	183	44
Germany	2,078	3	25	40
United Sta...	1,663	40	12	9
Norway	702	0	1	14
Cruise Ship	696	7	325	13

Side Information May Leak Privacy

- Eve polled our class last week: are you using a Macbook?
- 70 of us answered anonymously
- Eve got 20 Yes
- I came in this week, and Eve polled again
- Now he got 21 Yes
- My secret is leaked!



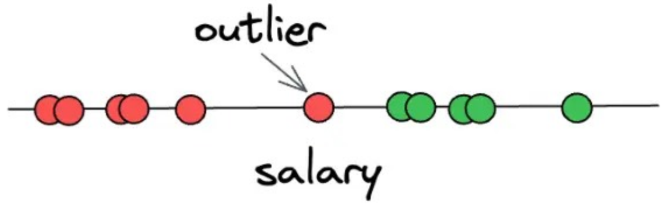
Side information: I was absent last week

Privacy Leakage is Everywhere

- Anonymization may not work
  - identify an individual by collection of fields, attributes, zip code, date of birth, gender …
  - A linkage attack to match “anonymized” records with non-anonymized records
- Re-Identification may not be the only risk
  - A collection of medical records on a given date list a small number of diagnoses. Additional information of visiting the facility on the date narrows range of possible diagnoses

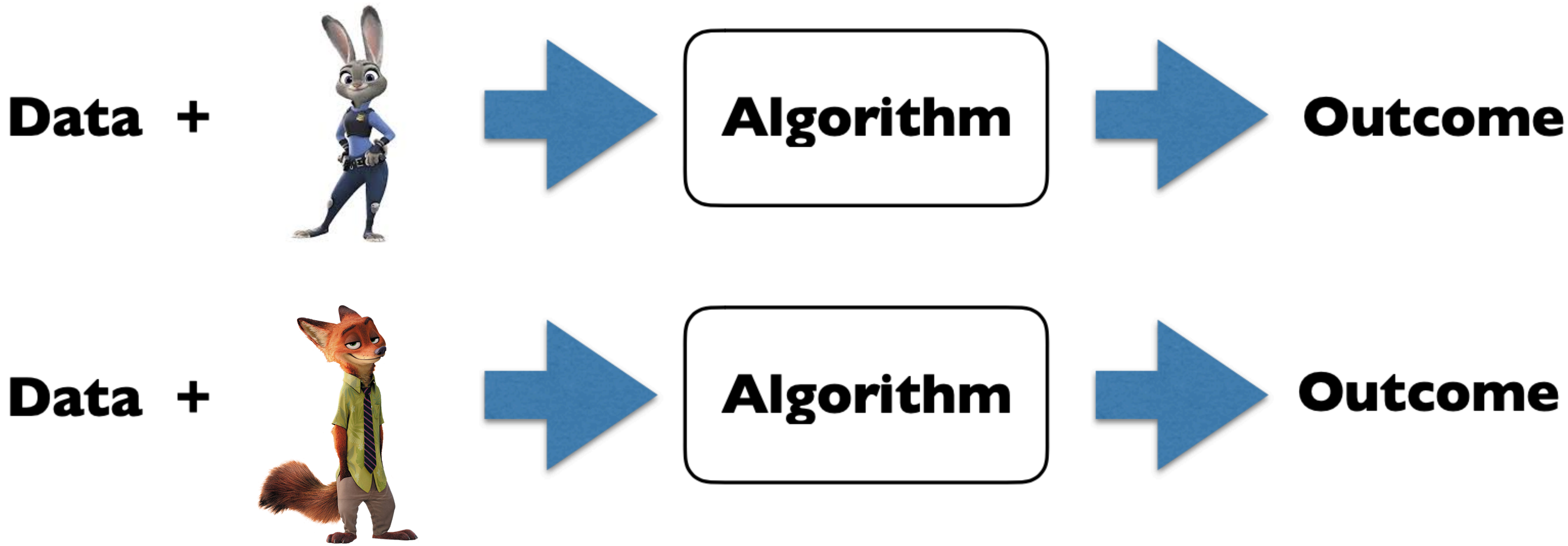
Privacy Leakage is Everywhere

- "Ordinary" facts are not OK
  - Bob regularly buys candies over years until suddenly switching to rarely buying candies — most likely be diagnosed with diabetes
- "Just a few" is not OK
  - Outliers may be more important!
- Queries over large sets may be risky
  - differencing attack to two large sets, one w/ X, one w/o X



- What are the privacy-preserving techniques in data mining?

Differential Privacy



Participation of a person does not change outcome

An adversary cannot decide if the person is in the dataset

Differential Privacy

- Randomness

A(**Data** + )

Random variables

have close distributions

A(**Data** + )

Randomness: Added by randomized algorithm **A**

Closeness: Bound likelihood ratio at each observed point

Basic Terms

- A **trustworthy** curator holds data of individuals in database D
- Each row corresponds to an individual
- **Goal**: Protect every individual row while permitting statistical analysis of D
- Non-interactive model: Curator releases summary statistics, or “sanitized database” once and for all
- Interactive model: permit asking queries adaptively, decide which query to ask next based on observed responses

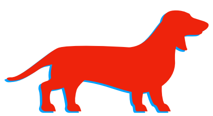
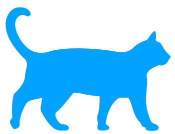
Name	Occupation	Date of Birth	Gender
Alice	Student	2001.1.1	Female
Bob	Faculty	1990.2.3	Male
Eve	Staff	1995.6.7	Male

A privacy mechanism is an algorithm that takes as input a database, the set of all possible database rows, random bits, a set of queries, and produces an output string.

Defining Privacy

- Privacy: data analysis knows no more about an individual **after** analysis is completed than **before** the analysis was begun
- Formally, adversary's prior and posterior views about an individual should **not** be "too different"
- Reminiscent of semantic security for a cryptosystem:
    - semantic security says nothing is learned about the plaintext from the ciphertext
        - e.g., if side information says the ciphertext is an encryption of "dog" or "cat," the ciphertext leaks nothing about which of "dog" or "cat" has been encrypted
        - Adversary simulator has the same odds of guessing as does the eavesdropper

Ciphertext: 911376011023607

Difference

- Semantic security
    - 3 parties: message sender, receiver, eavesdropper
- Privacy
    - 2 parties: curator & data analyst
    - data analyst can be adversary
    - given as **auxiliary information** the encryption of a secret using **random pad**, the analyst can decrypt the secret, but the adversary simulator learns nothing
    - careful in deciding “reasonable” auxiliary knowledge

Plausible Deniability

- “Privacy” comes from plausible deniability of any outcome. Report if one has property P by:

  1. Flip a coin
  2. If **tails**, then report truthfully ("Yes" if having P, "No" if not having P)
  3. If **heads**, then flip a second coin and report “Yes” if heads and “No” if tails

- What is the expected number of “Yes”?

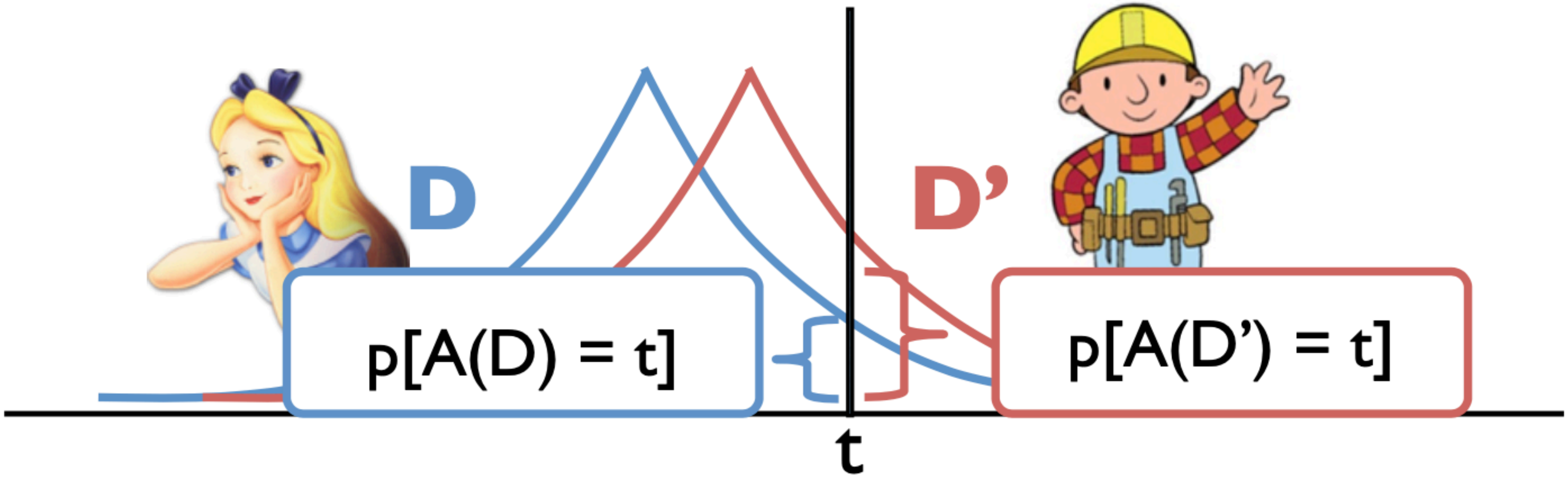
  ➡The expected number of “Yes” is 1/4 × total no. of participants “who do not has P” + 3/4 × total no. of participants “who has P”

  ➡if p is the true fraction of having P, the expected number of “Yes” is (1/4) + p/2

Randomized Alg.

- Probability Simplex: given a discrete set O, the probability simplex over O is denoted as $\Delta(O)$
- A randomize alg. A with domain $\mathcal{D}$ and discrete range O is associated with a mapping: $\mathcal{D} \rightarrow \Delta(O)$. On input $x \in \mathcal{D}$, alg. $\mathcal{A}$ outputs $A(x) = t$ with probability $(A(x))_t$ for each $t \in O$
- Distance between databases: the l1-norm of a database D is $\| D \|_1$. The l1 distance between D and D' is $\| D - D' \|_1$.
  - a measure of how many records differ between D & D'

Differential Privacy



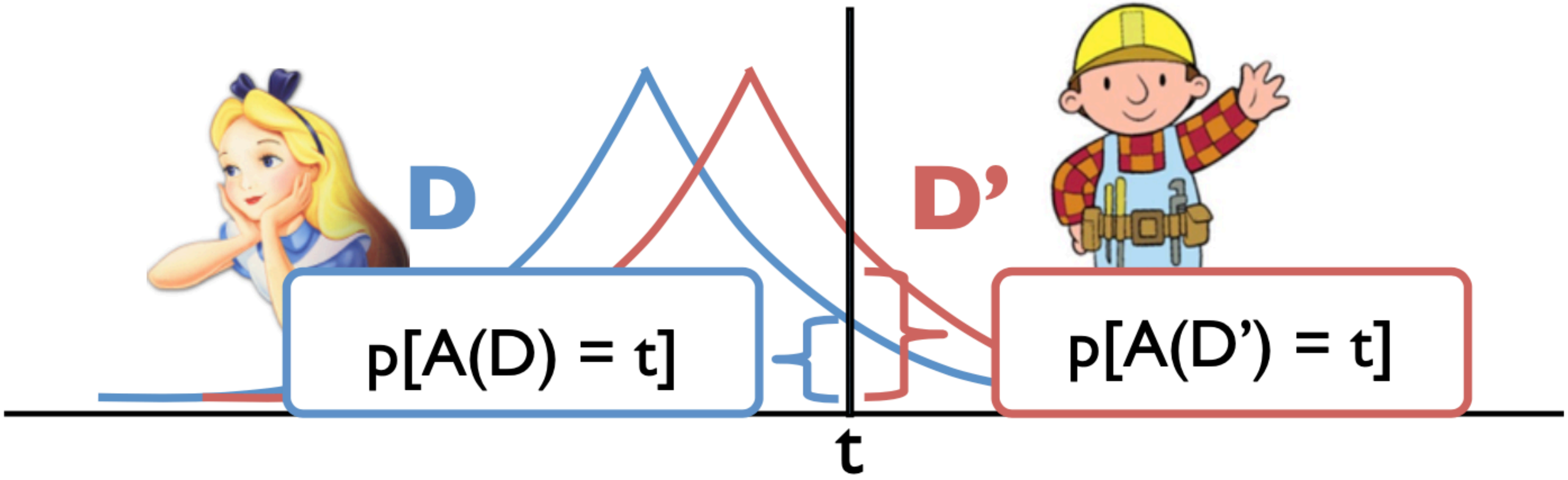
For all D, D' that differ in one person's value,

If A = ϵ-differentially private randomized algorithm, then:

Max-divergence of p(A(D)) and p(A(D'))

$$\boxed{\sup_t \left| \log \frac{p(A(D)=t)}{p(A(D')=t)} \right|} \leq \epsilon$$

Approx. Differential Privacy



For all D, D' that differ in one person's value,

If A = (ϵ, δ)-differentially private randomized algorithm, then:

$$\max_{S, \Pr(A(D) \in S) > \delta} \left[\log \frac{\Pr(A(D) \in S) - \delta)}{\Pr(A(D') \in S} \right] \leq \epsilon$$

Formal Definition

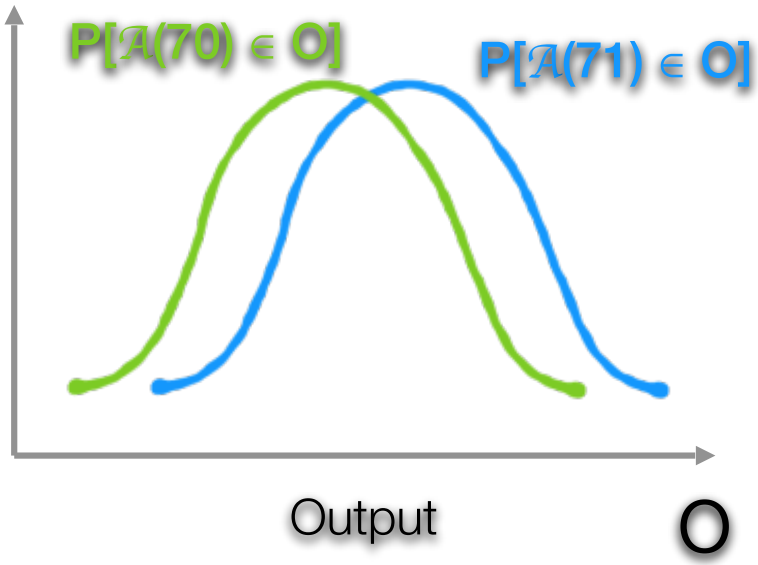
- A randomized alg. A with domain $\mathcal{D}$ is (ε, δ)-differentially private if

  for all $O \subseteq \text{Range}(A)$ and for all $D, D' \in \mathcal{D}$ such that $\| D - D' \|_1 \leq 1$:

$$P[A(D) \in O] \leq e^{\varepsilon} P[A(D') \in O] + \delta$$

for every pair of **adjacent databases** D, D', the posterior distributions should be close

δ: residual probability, should be small



Adjacent Databases

- Consider differential privacy at a level of individuals
  - insensitive to the addition or removal of any individual
    - e.g., a differentially-private movie recommendation system could protect data at:
      - event level: hiding the rating of a single movie, but not one's preference for the romantic movies. Hide a person's rating for a movie
      - user level: hiding an individual's entire ratings. Hide a person's likes and dislikes
- Protection against **arbitrary threats** including re-identification
- Automatic mitigation of linkage attacks
- **Quantification of privacy loss**, allows comparisons among different privacy-preserving techniques

Properties

- Post-Processing:
  - Let $\mathcal{A}$ be a randomized alg. that is (ε, δ)-differentially private. Let f be an arbitrary randomized mapping. Then $f \circ \mathcal{A}$ is (ε, δ)-differentially private
- Group privacy for (ε, 0)-differentially private mechanisms:
  - Any (ε, 0)-differentially private mechanism $\mathcal{A}$ is ($k\varepsilon$, 0)-differentially private for groups of size k
- Composition: combination of two differentially private alg. is differentially-private
  - Let $\mathcal{M}_1$, $\mathcal{M}_2$ be an $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$-differentially-private alg. respectively. Their combination $\mathcal{M}_{1,2}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is ($\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2$)-differentially private

Composition

- What do we mean by composition?

    1. Repeated use of differentially-private alg. on the same database

    2. Repeated use of differentially-private alg. on different databases that may contain information relating to the same individual

- Model composition where the adversary can adaptively affect the databases being input to future mechanisms

- A probabilistic adversary $\mathcal{A}$ for i = 1, …, k:

    1. $\mathcal{A}$ outputs two adjacent databases x_i^0 and x_i^1, a mechanism $\mathcal{M}_i$ and parameters w_i

    2. $\mathcal{A}$ receives $y_i \in \mathcal{M}_i (w_i, x_i^b)$

Composition

- $\mathcal{A}$'s view of the experiment: coin tosses b & all outputs (y_1, …, y_k)

- Consider $\mathcal{A}$ chooses x_i^0 to hold Bob's data and x_i^1 to differ only in that Bob's data are deleted. Differential privacy requires the two experiments to be "close" to each other, i.e., $\mathcal{A}$ cannot tell, given the output of all k mechanisms, whether Bob's data was ever used

For a fixed view v = ($r, y_1, \ldots, y_k$)

b = 0, the view of A is $V^0 = (R^0, Y_1^0, \ldots, Y_k^0)$

b = 1, the view of A is $V^1 = (R^1, Y_1^1, \ldots, Y_k^1)$

$$
\begin{aligned}
&\ln\left(\frac{\Pr[V^0 = v]}{\Pr[V^1 = v]}\right) \\
&= \ln\left(\frac{\Pr[R^0 = r]}{\Pr[R^1 = r]} \cdot \prod_{i=1}^{k} \frac{\Pr[Y_i^0 = y_i | R^0 = r, Y_1^0 = y_1, \ldots, Y_{i-1}^0 = y_{i-1}]}{\Pr[Y_i^1 = y_i | R^1 = r, Y_1^1 = y_1, \ldots, Y_{i-1}^1 = y_{i-1}]}\right) \\
&= \sum_{i=1}^{k} \ln\left(\frac{\Pr[Y_i^0 = y_i | R^0 = r, Y_1^0 = y_1, \ldots, Y_{i-1}^0 = y_{i-1}]}{\Pr[Y_i^1 = y_i | R^1 = r, Y_1^1 = y_1, \ldots, Y_{i-1}^1 = y_{i-1}]}\right) \\
&\stackrel{\text{def}}{=} \sum_{i=1}^{k} c_i(r, y_1, \ldots, y_i).
\end{aligned}
$$

$$
c_i(r, y_1, \ldots, y_{i-1}, y_i) = \ln\left(\frac{\Pr[\mathcal{M}_i(w_i, x_i^0) = y_i]}{\Pr[\mathcal{M}_i(w_i, x_i^1) = y_i]}\right)
$$

- What are the basic mechanisms?

Randomized Response

- Report if one has property P by:
  1. Flip a coin
  2. If **tails**, then report truthfully
  3. If **heads**, then flip a second coin and report "Yes" if heads and "No" if tails
- The above mechanism is (ln3, 0)-differentially private
- Proof:

$$\frac{\Pr[\text{Response} = \text{Yes} | \text{Truth} = \text{Yes}]}{\Pr[\text{Response} = \text{Yes} | \text{Truth} = \text{No}]}$$

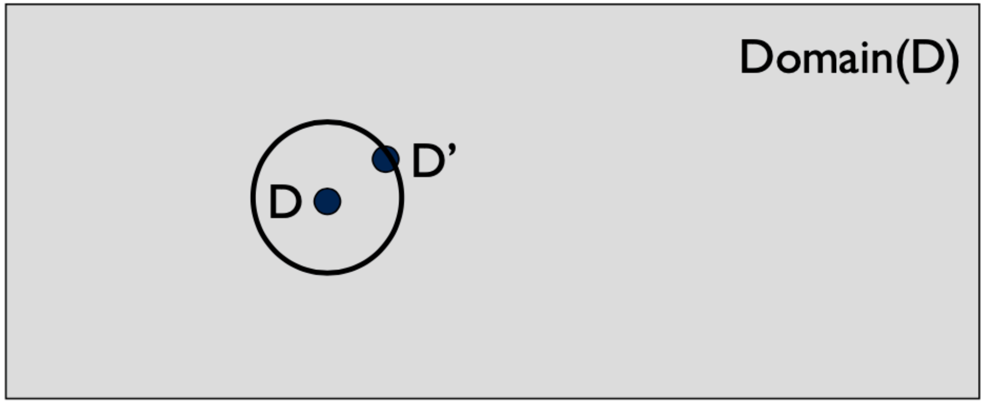
When the truth is "Yes" the outcome will be "Yes" if the 1st coin comes up tails (prob. 1/2) or the 1st & 2nd coin comes up heads (prob. 1/4)

$$= \frac{3/4}{1/4} = \frac{\Pr[\text{Response} = \text{No} | \text{Truth} = \text{No}]}{\Pr[\text{Response} = \text{No} | \text{Truth} = \text{Yes}]} = 3.$$

Global Sensitivity Method

- Given function f, sensitive dataset D
- Find a differentially-private approximation to f(D)
  - E.g., f(D) = mean of data points in D
  - Define dist(D, D') = #records that D, D' differ by **Global Sensitivity of f**:

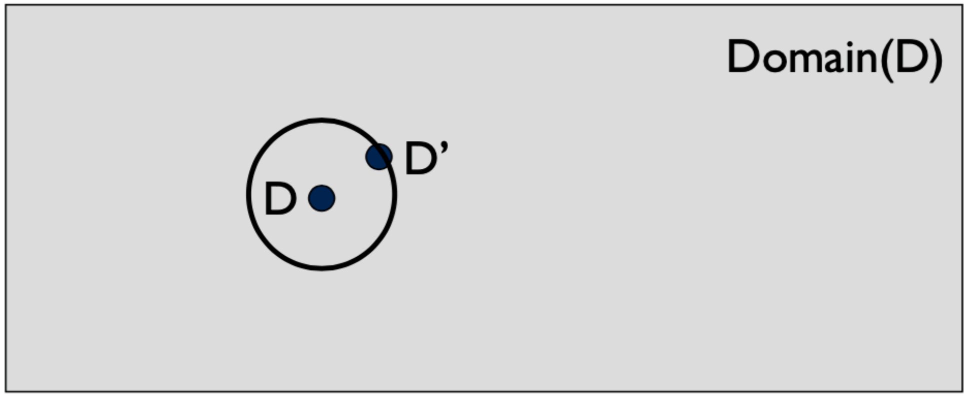
S(f) = | f(D) - f(D')|



Global Sensitivity Method

- Given function f, sensitive dataset D
- Find a differentially-private approximation to f(D)
  - E.g., f(D) = mean of data points in D
  - Define dist(D, D') = #records that D, D' differ by **Global Sensitivity of f**:

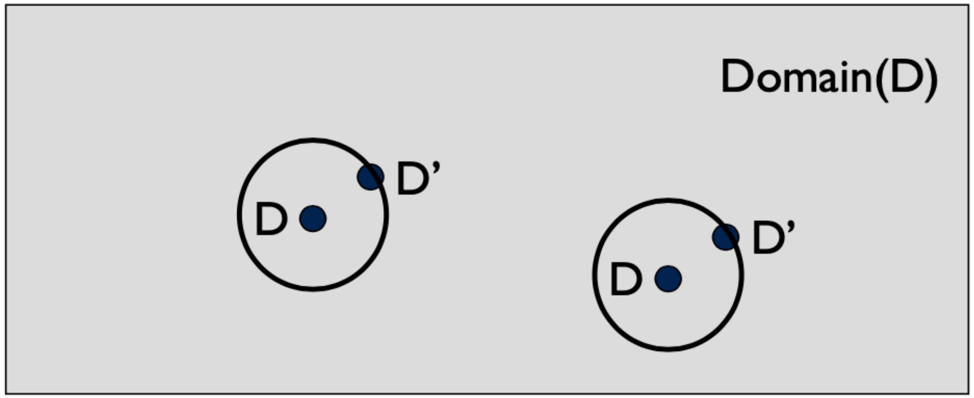
$$S(f) \quad = \quad \underset{dist(D, D') = 1}{} \; | f(D) - f(D')|$$



Global Sensitivity Method

- Given function f, sensitive dataset D
- Find a differentially-private approximation to f(D)
  - E.g., f(D) = mean of data points in D
  - Define dist(D, D') = #records that D, D' differ by **Global Sensitivity of f**:

$$S(f) = \max_{dist(D, D') = 1} |f(D) - f(D')|$$



Laplace Mechanism

- Counting queries "How many elements in the database satisfy Property P?"

- L1-sensitivity of counting query f:

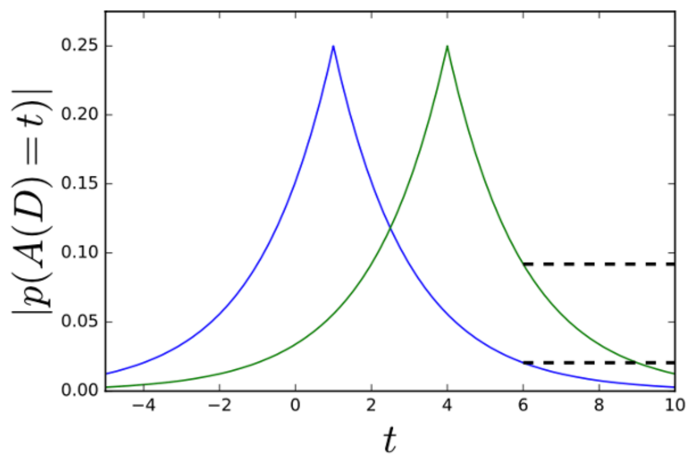
$$\Delta f = \max_{\|D-D'\|_1=1} \|f(D) - f(D')\|_1 \quad \mathbf{= 1}$$

The sensitivity of f gives an upper bound on how much we must perturb output to preserve privacy

captures the magnitude by which a single individual's data can change the function f in the **worst** case

- Laplace Distribution with scale b is the distribution with PDF:

$$\mathrm{Lap}(t|b) = \frac{1}{2b}\exp(-\frac{t}{b})$$



Laplace Mechanism

- Given query f, Laplace mechanism is defined as:

$$\mathcal{A}(\mathbf{x}, \mathbf{f}(\cdot), \boldsymbol{\varepsilon}) = \mathbf{f}(\mathbf{x}) + \mathbf{t}$$

where t is a random variable drawn from Lap(Δf/**ε**)

- The above mechanism is (**ε**, 0)-differentially private

- Proof: Let p_x denote the PDF of $\mathcal{A}$(x) and p_y denote the PDF of $\mathcal{A}$(y).

at some arbitrary point z:

$$\frac{p_x(z)}{p_y(z)} = \frac{\exp(-\frac{\epsilon|f(x)-z|}{\Delta f})}{\exp(-\frac{\epsilon|f(y)-z|}{\Delta f})} = \exp(\frac{\epsilon(|f(x)-z|-|f(y)-z|)}{\Delta f})$$

$$\leq \exp(\frac{\epsilon|f(x)-f(y)|}{\Delta f})$$

$$\leq \exp(\epsilon)$$

Fact: $\mathrm{Lap}(t|b) = \frac{1}{2b}\exp(-\frac{t}{b})$

$$b = \frac{\Delta f}{\epsilon}, t = z - f(x)$$

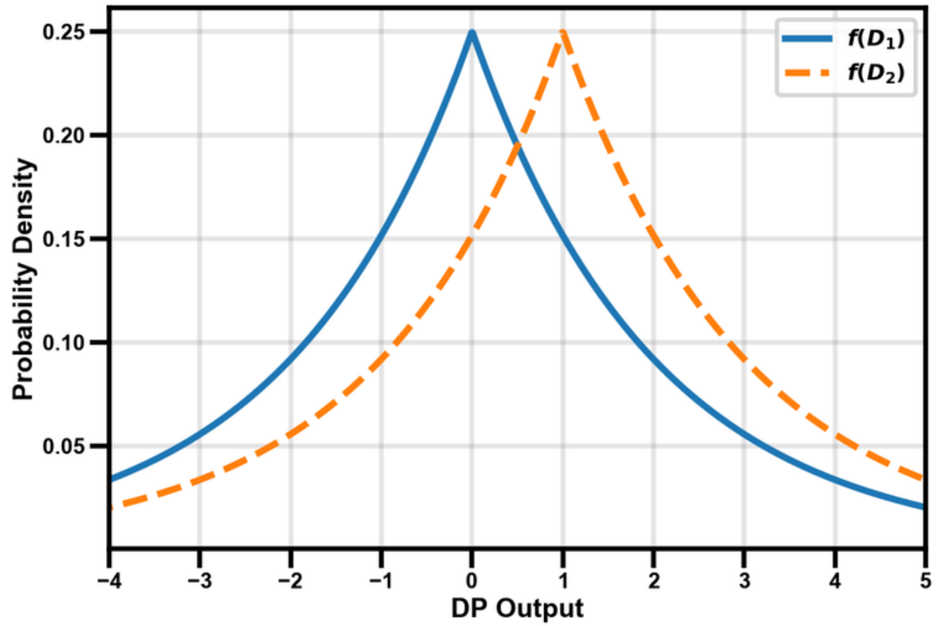
Example: Mean

M(D) = Mean(D), where each record is a scalar in [0, 1]

Global sensitivity of f = 1/n

Laplace mechanism:

Output M(D) + z, where $z \sim \frac{1}{n\epsilon} Lap(0,1)$



Accuracy Loss

- How much noise do we introduce in Laplace mechanism?
- Let query f map databases to k numbers: z = $\mathcal{M}$(x, f(·), **ε**) = f(x) + t.

  For δ ∈ (0, 1]: **output of Laplace Mechanism**

$$
\begin{aligned}
\Pr\left[\|f(x) - z\|_\infty \geq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] &= \Pr\left[\max_{i \in [k]} |Y_i| \geq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] \\
&\leq k \cdot \Pr\left[|Y_i| \geq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] \\
&= k \cdot \left(\frac{\delta}{k}\right) \\
&= \delta
\end{aligned}
$$

how much are we away from the true response?

very small since we restrict the amount of noise to be added

Fact: If t~Lap(b), then Pr[|t| >= ln(k/δ)·b] = δ/k

Example

$$\Pr\left[\|f(x) - z\|_\infty \geq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] \leq \delta$$

- We wish to calculate the frequency of the first names, from a list of 10,000 potential names
- Query $f : N^{|X|} \to R^{10000}$
- Sensitivity $\Delta f = 1$, since every person can only have at most one first name
- Calculate the frequency of all 10, 000 names with (1, 0)-differential privacy
- With probability 95%, no estimate will be off by more than an additive error of $\ln(10000/.05) \approx 12.2$

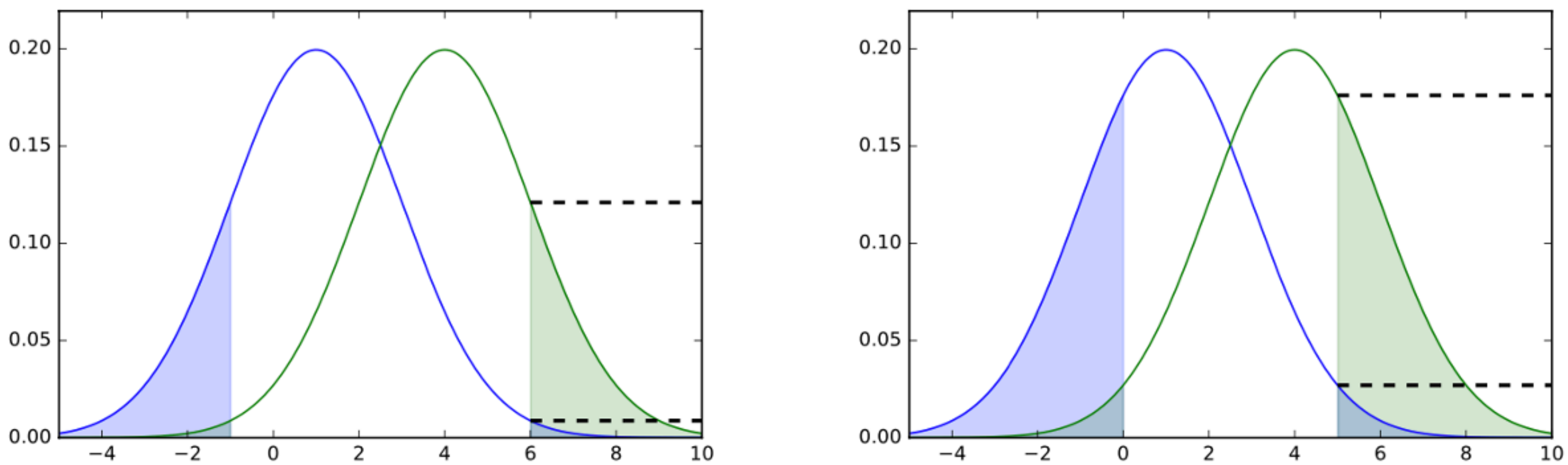
Gaussian Mechanism

Global Sensitivity of f is $\Delta f = \max_{\text{dist}(D, D') = 1} \| f(D) - f(D') \|_2$

Output $\mathcal{M}$(D) + Z where

$$Z \sim \frac{\Delta f}{\epsilon} \mathcal{N}(0, 2\ln(1.25/\delta))$$

(ϵ, δ)-differentially private



Exponential Mechanism

- We wish to choose the "best" response but adding noise directly to the computed quantity can destroy its value
    - Suppose we have an abundant supply of goods and 4 bidders: A,B,C,D, where A,B,C each bid $1.00 and D bids $3.01. What is the optimal price? At $3.01 the revenue is $3.01, at $3.00 the revenue is $3.00, but at $3.02 the revenue is 0!
- Exponential mechanism is defined w.r.t. utility function, mapping outputs to utility scores
- We only care about the sensitivity of u:

possible output r

$$\Delta u \equiv \max_{r \in \mathcal{R}} \max_{x,y:\|x-y\|_1 \leq 1} |u(x,r) - u(y,r)|$$

- Exponential mechanism: outputs r ∈ R with prob. proportional to

$$\exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)$$

Exponential Mechanism

- Exponential mechanism preserves (**ε**, 0)-differential privacy
- Proof: The privacy loss is

$$\ln \frac{\Pr[\mathcal{M}_E(x, u, \mathcal{R}) = r]}{\Pr[\mathcal{M}_E(y, u, \mathcal{R}) = r]} =$$

$$\ln\left(\frac{\exp(\varepsilon u(x, r)/\Delta u)}{\exp(\varepsilon u(y, r)/\Delta u)}\right) = \varepsilon[u(x, r) - u(y, r)]/\Delta u) \leq \varepsilon$$

Exponential Mechanism

- Problem:
- Given function f(w, D), sensitive Data D
- Find differentially-private approximation to

$$w^* = \underset{w}{\operatorname{argmax}}\, f(w, D)$$

Example: f(w, D) = accuracy of classifier w on dataset D

Exponential Mechanism

Outputs r ∈ R with prob. $\exp(\frac{\varepsilon u(x,r)}{2\Delta u})$

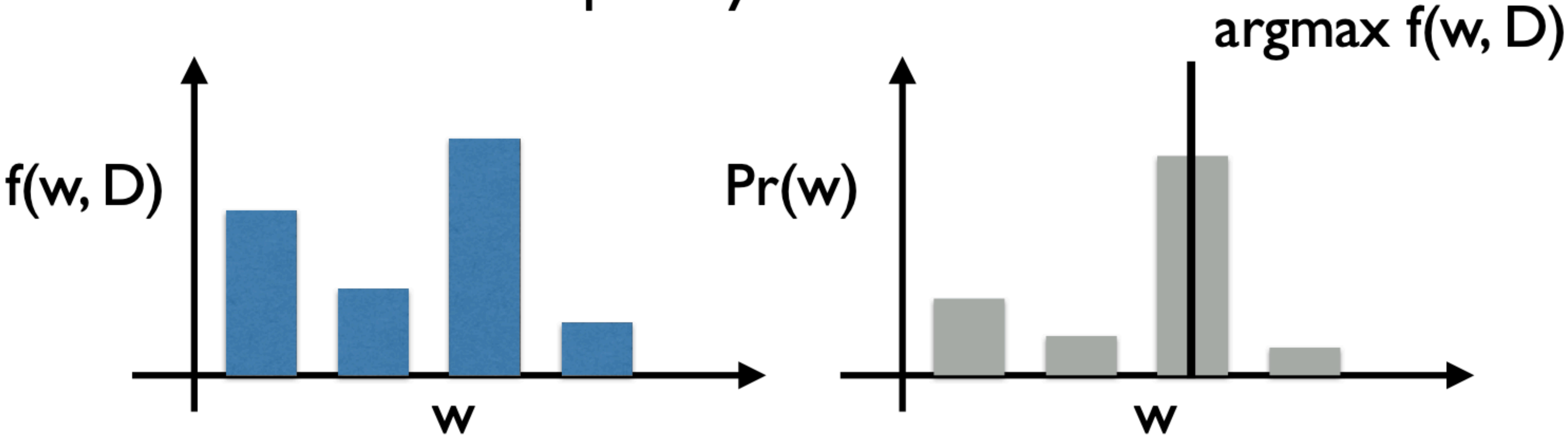
Suppose for any w,

$$|f(w, D) - f(w, D')| \leq S$$

when D and D' differ in 1 record. Sample w from:

$$p(w) \propto e^{\epsilon f(w,D)/2S}$$

for ϵ-differential privacy.



Example: Parameter Tuning

- Given validation data D, k classifiers w_1, .., w_k, privately find the classifier with highest accuracy on D
- Here, f(w, D) = classification accuracy of w on D. For any w, any D and D' that differ by one record

$$|f(w, D) - f(w, D')| \leq \frac{1}{|D|}$$

So, the exponential mechanism outputs w_i with prob:

$$\Pr(w_i) \propto e^{\epsilon |D| f(w_i, D)/2}$$

The larger D is, the higher likelihood that the optimal w is sampled

Reading

- C. Dwork and A. Roth, “The Algorithmic Foundations of Differential Privacy,” 2014, Chapter 1, 2, 3