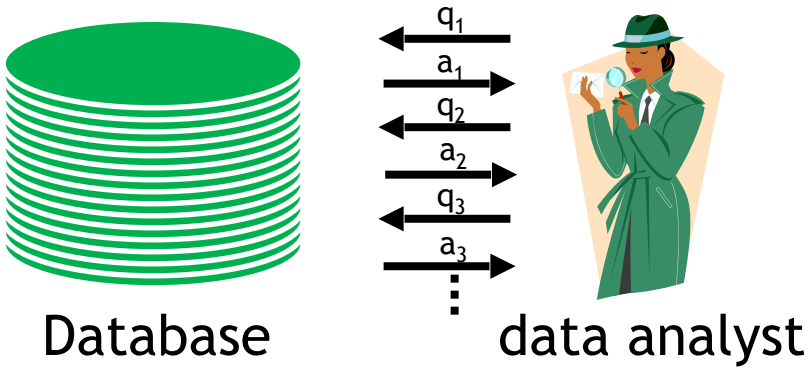
Privacy-Preserving Data Analysis

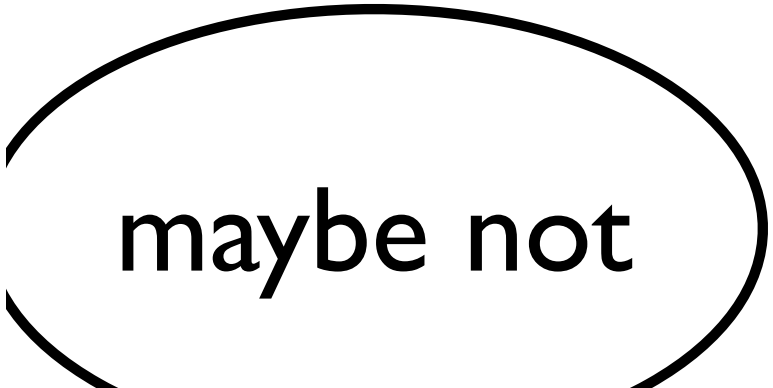


- Census, epidemic detection based on OTC drug purchases; analysis of loan application data for evidence of discrimination,…
- 50+ year old problem

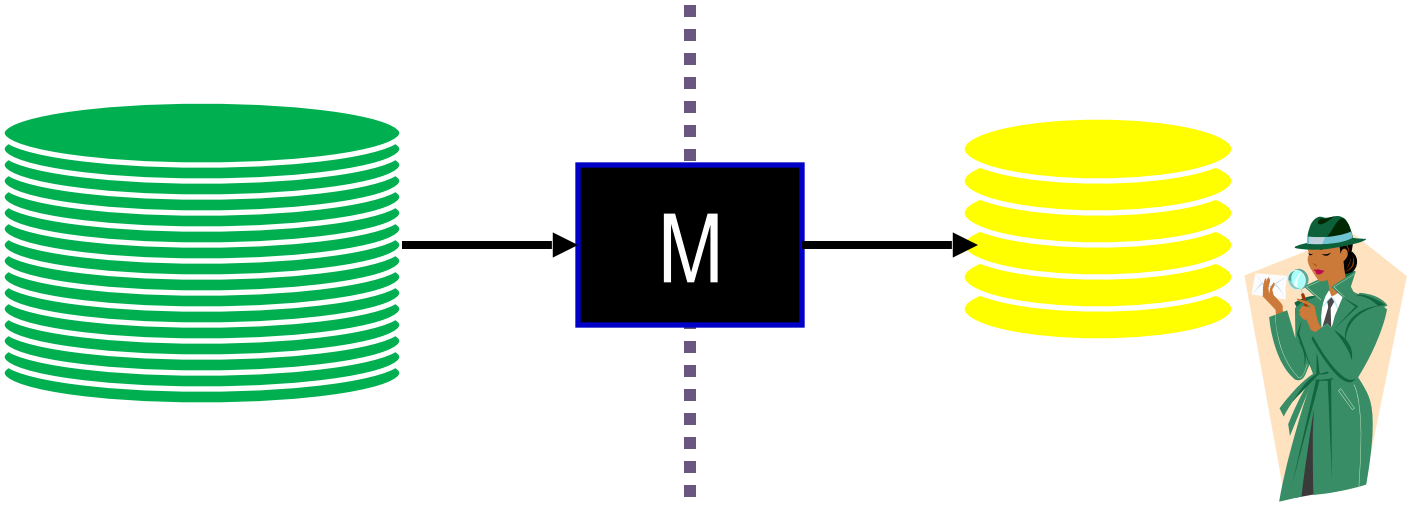
What analyses on a database might violate privacy? What analyses are privacy-preserving?

what to promise?

delete identifying information



“De-Identification”?



De-Identified data isn’t.

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	[illegible]city	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: [illegible]e-identifying anonymous data by linking to external data

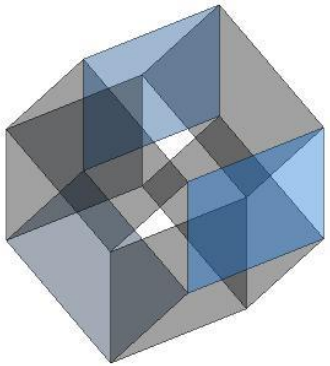
Public voter dataset

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

Curse of Dimensionality

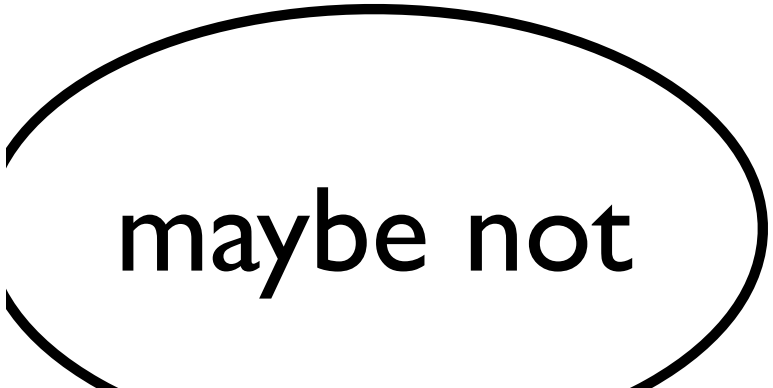
Aggarwal (VLDB 2005)



- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless

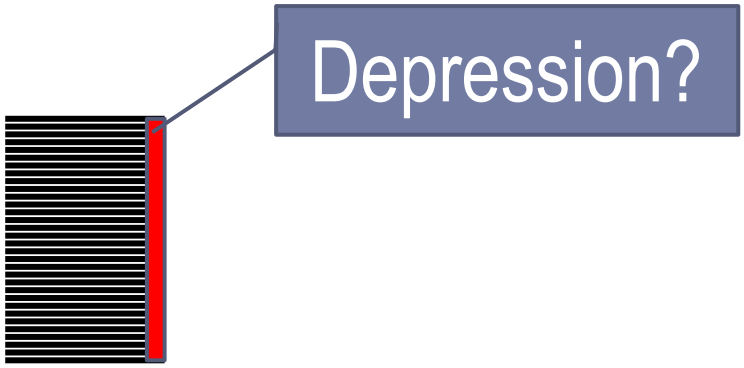
what to promise?

only ask questions that pertain to large populations



The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*
- Needle in a Haystack
  - Determine presence of an individual's genomic data in GWAS case group
- The Big Bang attack
  - Reconstruct "depression" bit column



Dinur and Nissim'03 Homer+'08

Fundamental Law of Info Recovery

- "Overly accurate" estimates of "too many" statistics is blatantly non-private.

what to promise?

access to the output should not enable one to learn anything about an individual that could not be learned without access

is this possible?

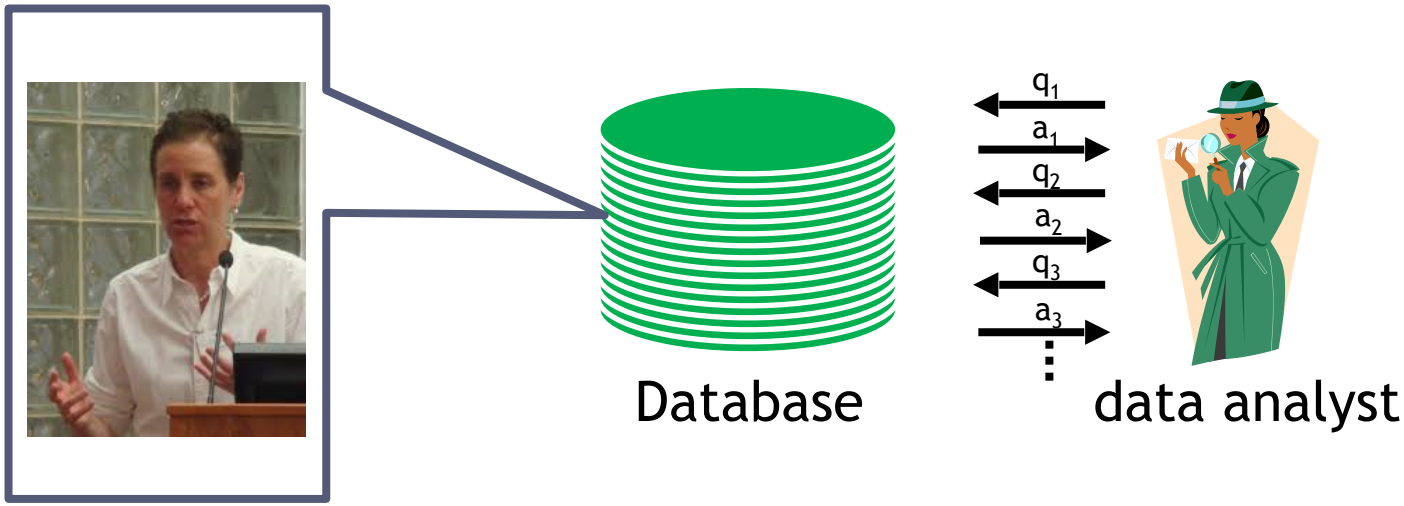
hint: *either* privacy or utility separately is easy

what to promise?

access to the output should not enable one to learn anything about an individual that could not be learned without access

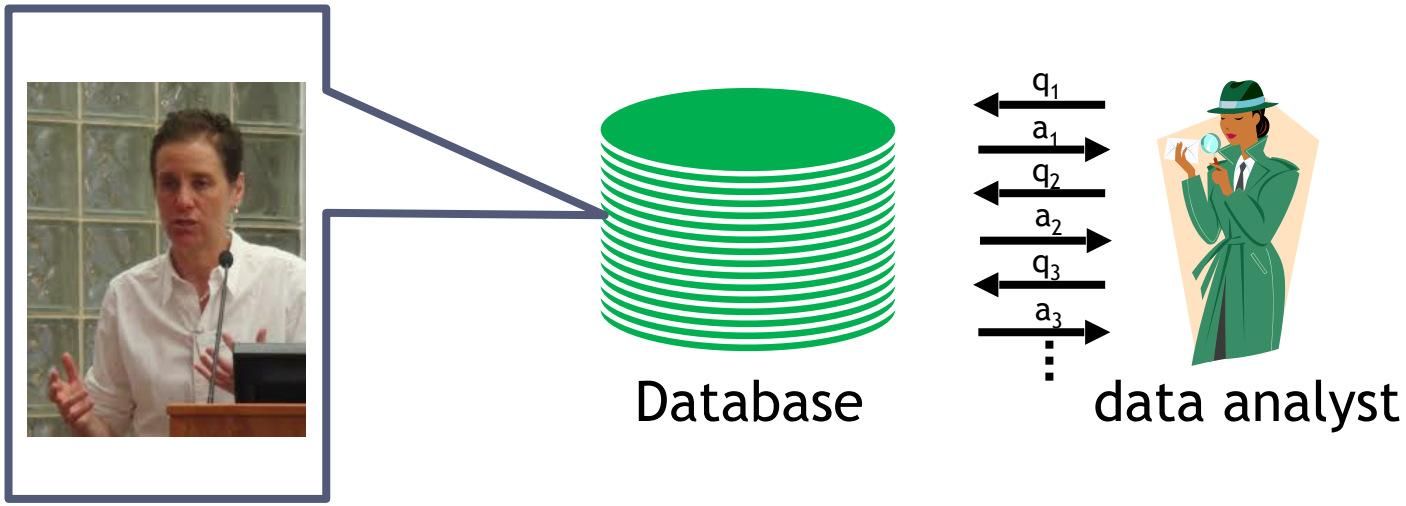
is this desirable?

Privacy-Preserving Data Analysis?



- "Can't learn anything new about Helen"?
- Dalenius, 1977; Goldwasser and Micali: semantic security 1982

Privacy-Preserving Data Analysis?

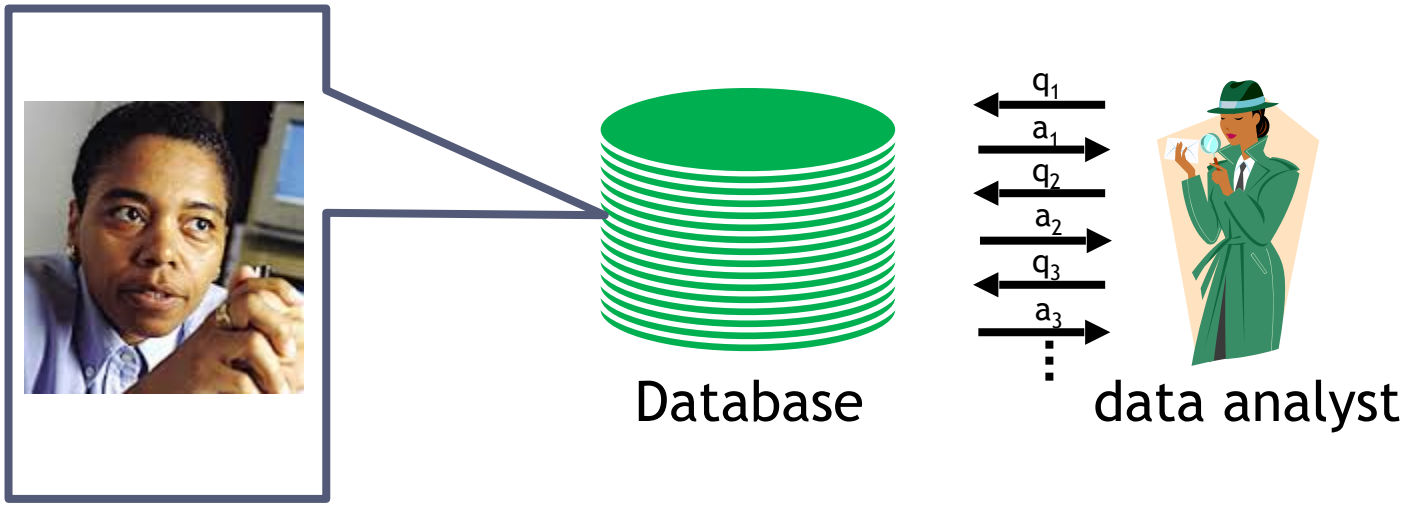


- "Can't learn anything new about Helen"?
- Then what is the point?

what to promise?

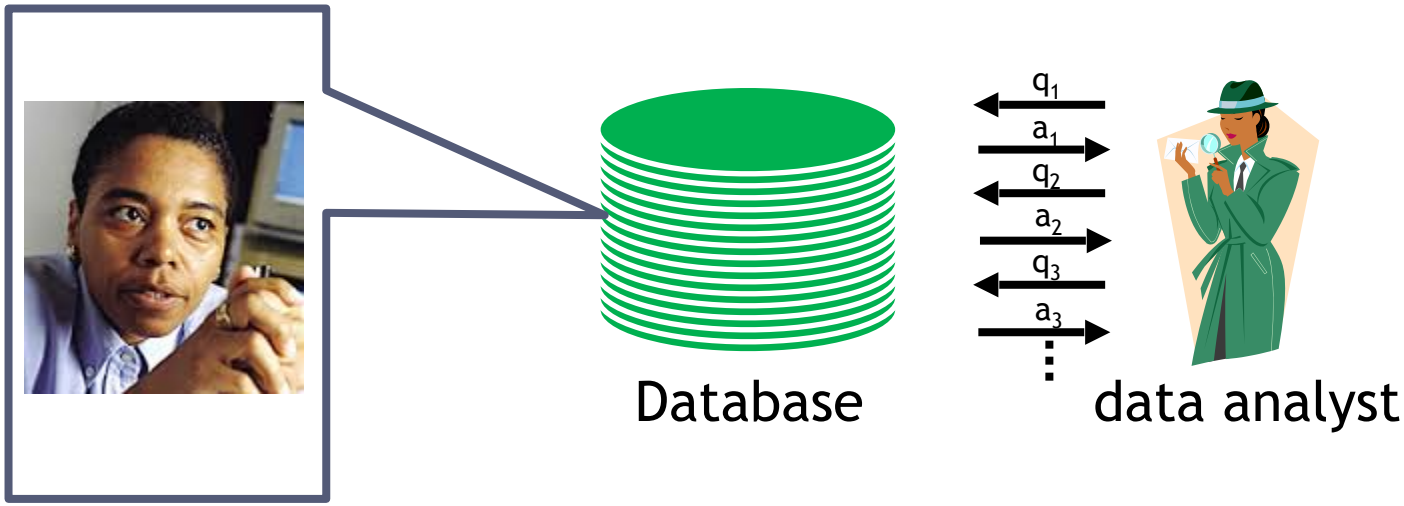
access to the output should not enable one to learn much more about an individual than could be learned via the same analysis omitting that individual from the database

Privacy-Preserving Data Analysis?



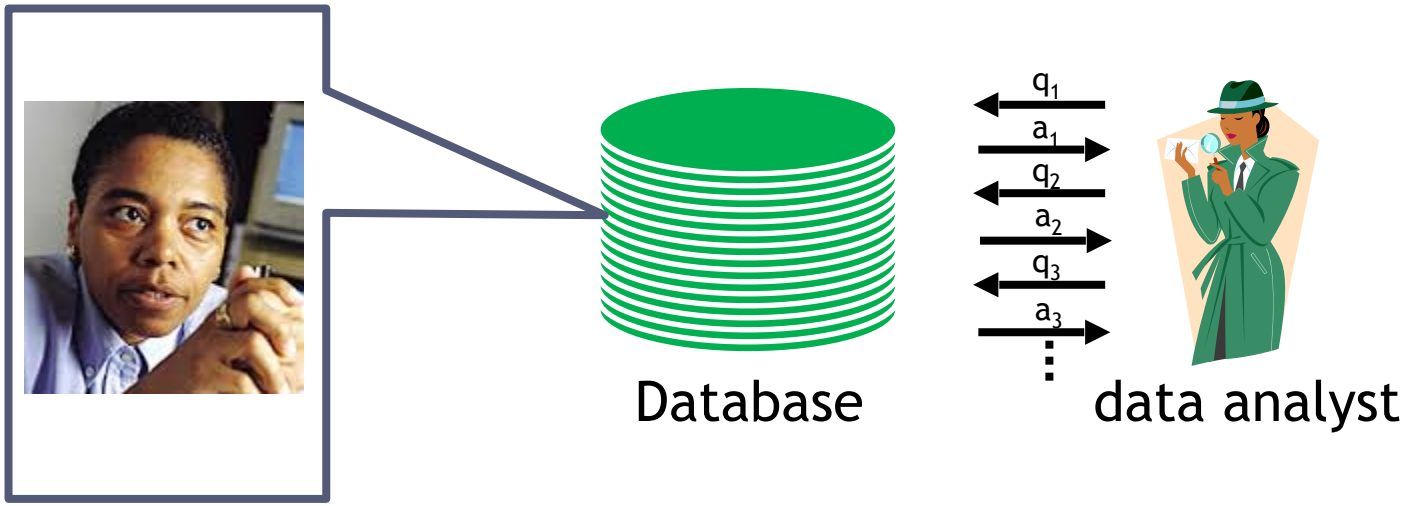
- Ideally: learn same things if Helen is replaced by another random member of the population

Privacy-Preserving Data Analysis?



- Ideally: learn same things if Helen is replaced by another random member of the population ("stability")

Privacy-Preserving Data Analysis?



- Stability preserves Helen's privacy AND prevents over-fitting
- Privacy and Generalization are aligned!

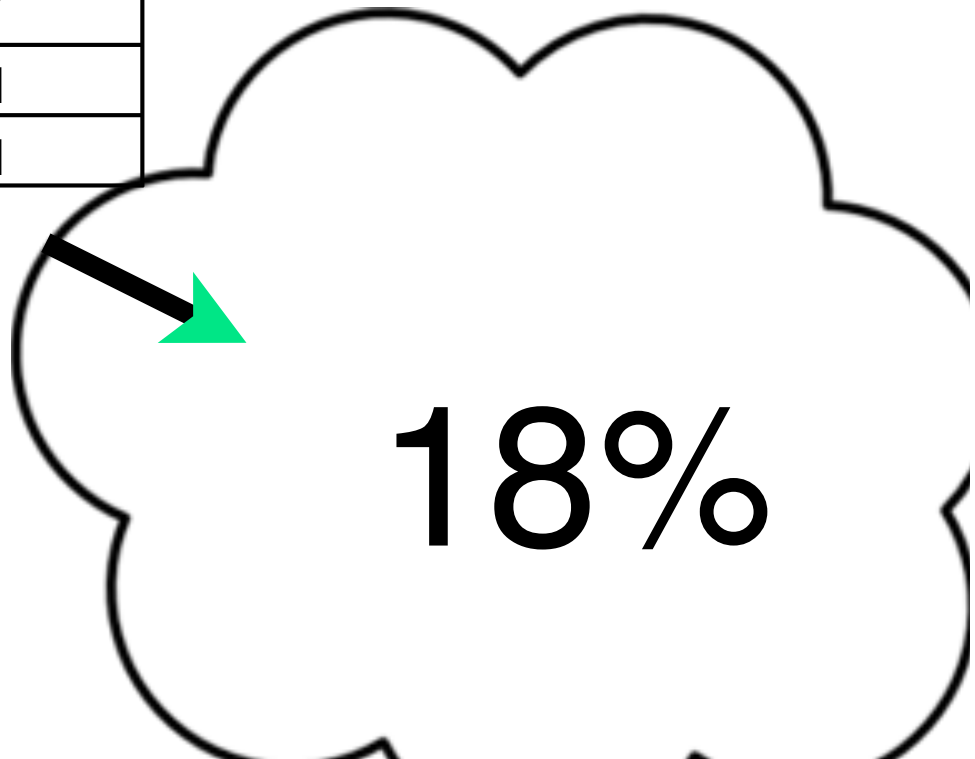
Differential Privacy

- The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset.
  - Helen goes away, Latanya joins, Helen is replaced by Latanya

what to promise?

think of output as randomized

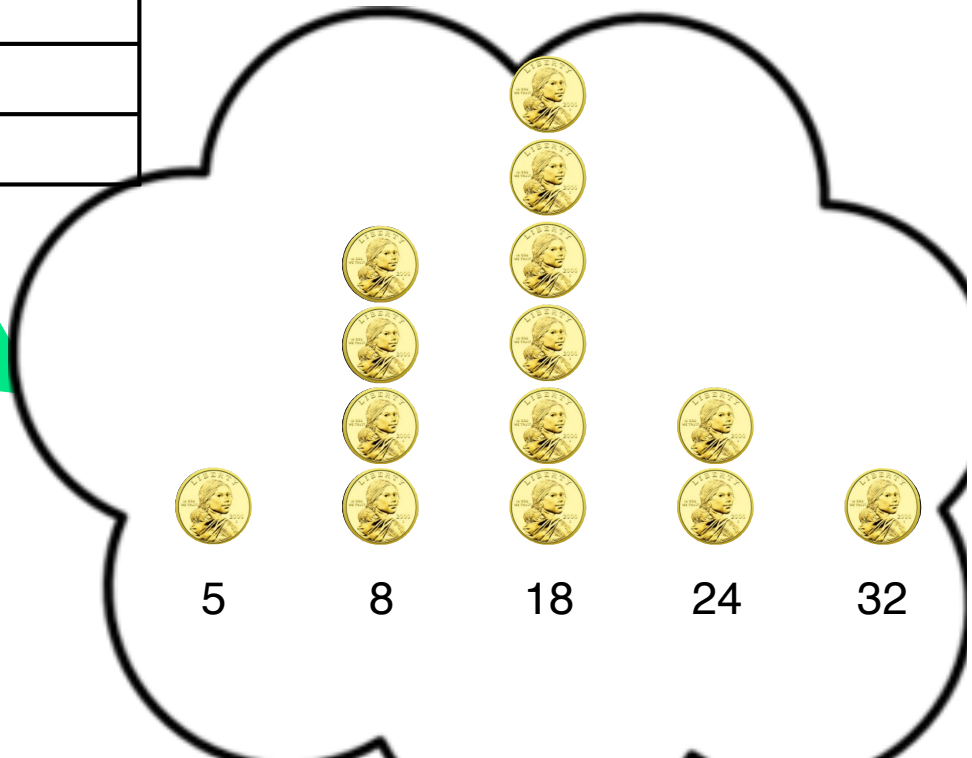
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



what to promise?

think of output as randomized

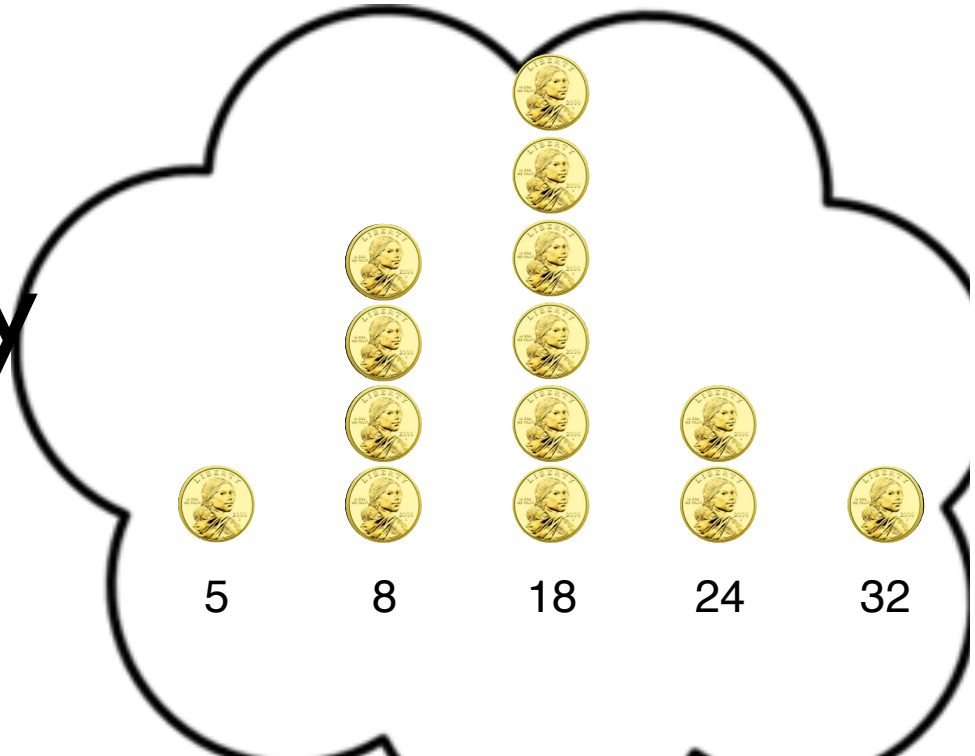
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



what to promise?

think of output as randomized

promise: if you leave the database, no outcome will change probability by very much



statistical database model

X set of possible entries/rows

one row per person

database x a set of rows; $x \in \mathbb{N}^{|X|}$

(histogram)

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

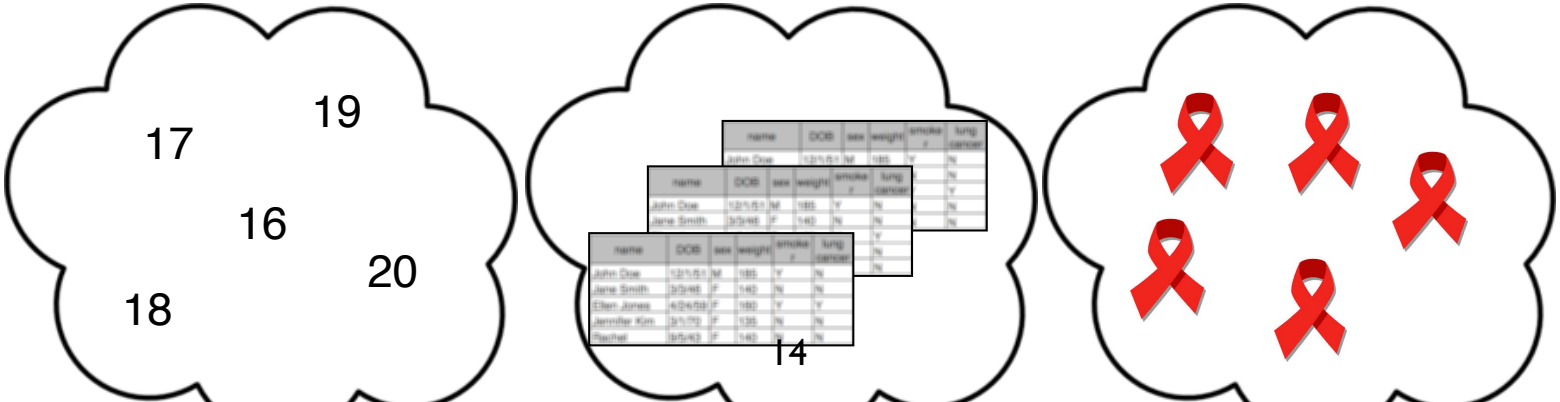
analyst objective

wishes to compute on $x \in \mathbb{N}^{|X|}$

fit a model, compute a statistic, share "sanitized" data

preserve privacy of individuals

design randomized algorithm M mapping x to into outcome space, that masks small changes in x



neighboring databases

what's a small change?

require nearly identical behavior on neighboring databases differing by the addition or removal of a single row:

$$||x - y||_1 \leq 1$$

for $x, y \in \mathbb{N}^{|X|}$

differential privacy

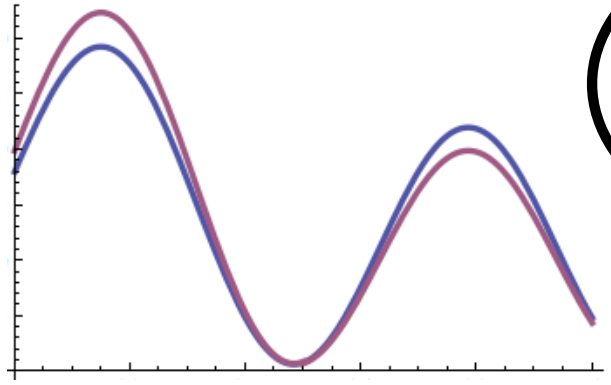
[DinurNissim03, DworkNissimMcSherrySmith06, Dwork06]

ε-Differential Privacy for algorithm M:

for any two neighboring data sets x_1, x_2, differing by the addition or removal of a single row

any $S \subseteq$ range(M),

$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$

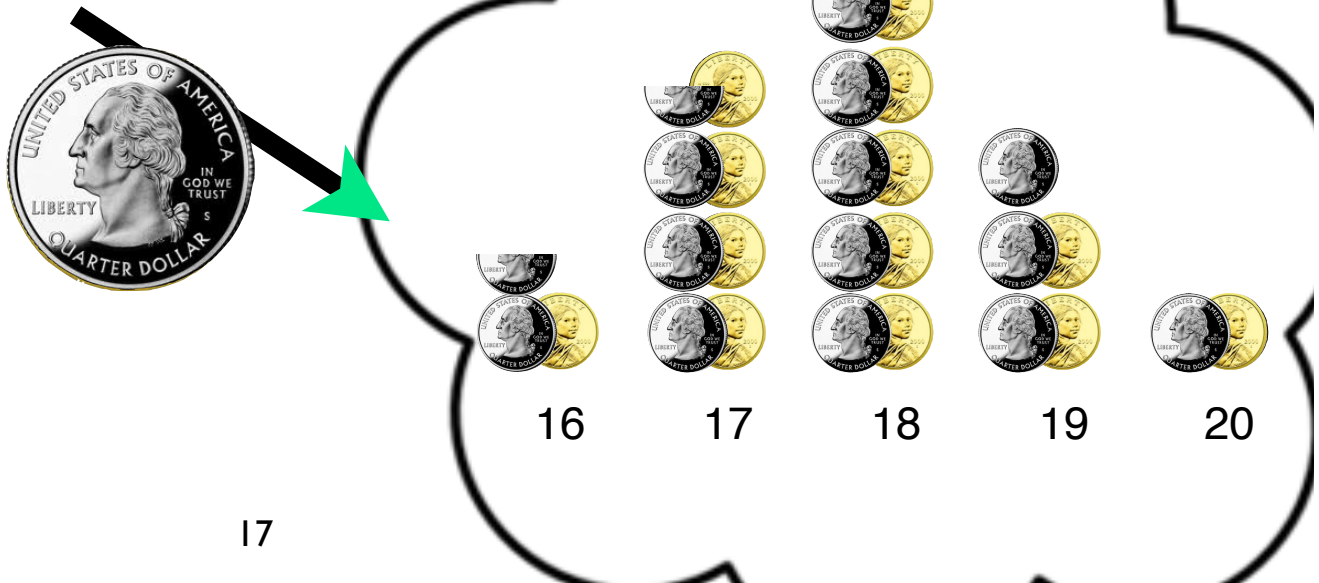


$e^{\varepsilon} \sim (1 + \varepsilon)$

differential privacy

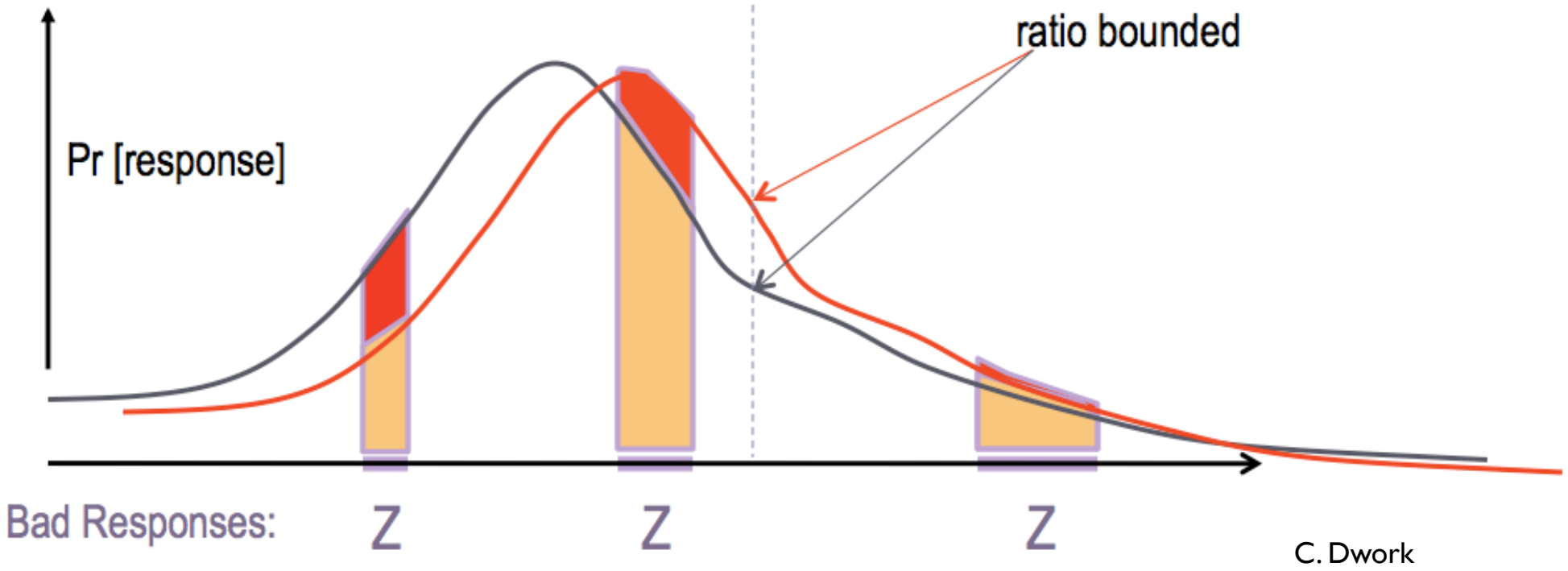
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
~~Ellen Jones~~	~~4/24/59~~	~~F~~	~~160~~	~~Y~~	~~Y~~
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



differential privacy

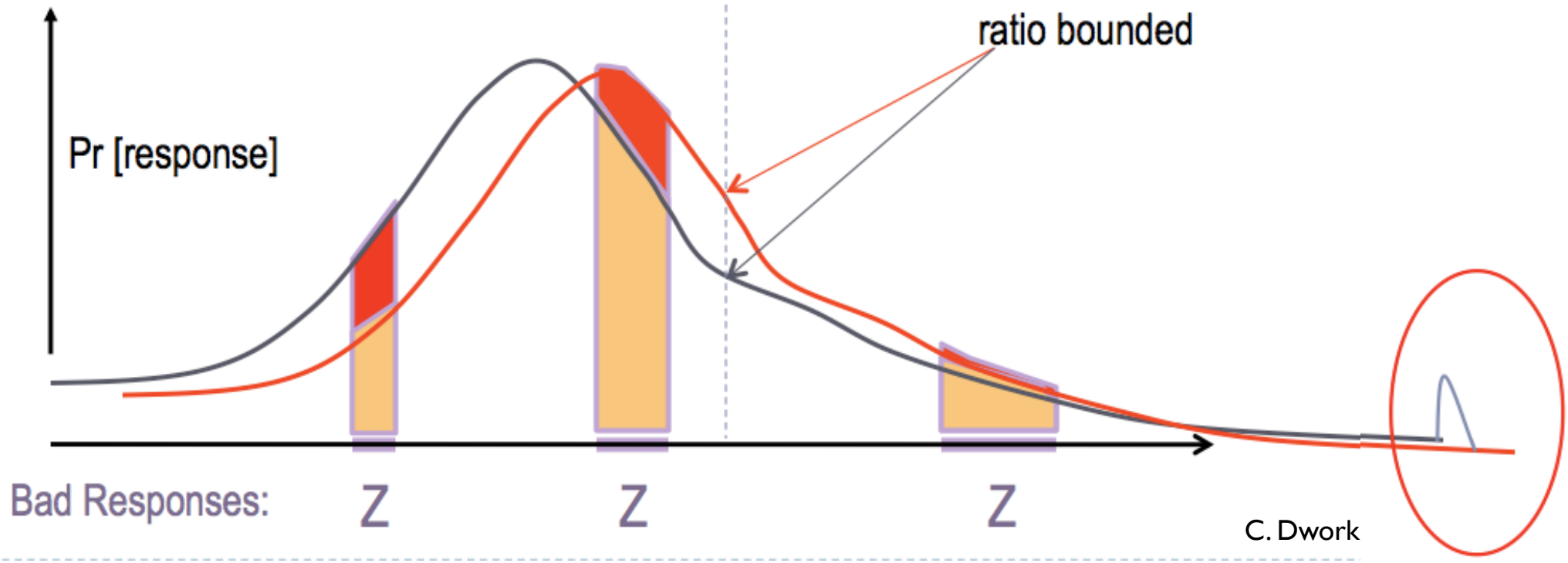
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$



C. Dwork

(ε, δ)-differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S] + \delta$$



differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

Is a statistical property of mechanism behavior

- unaffected by auxiliary information
- independent of adversary's computational power

differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

promise: if you leave the database, no outcome will change probability by very much

is this achievable?

yes!


chrome

source: http://www.recode.net/2016/6/15/11940908/mossberg-apple-is-still-a-world-of-its-own

source: http://edge.alluremedia.com.au/m/l/2015/03/ApplePrivacy.jpg

today

Formalizing privacy

Privacy properties and basic tools

Randomized Response

differential privacy

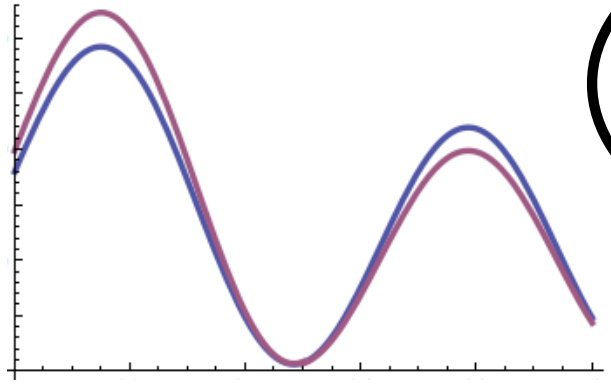
[DinurNissim03, DworkNissimMcSherrySmith06, Dwork06]

ε-Differential Privacy for algorithm M:

for any two neighboring data sets x_1, x_2, differing by the addition or removal of a single row

any $S \subseteq$ range(M),

$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$



$e^{\varepsilon} \sim (1 + \varepsilon)$

group privacy

Thm. Any $(\varepsilon, 0)$-DP mechanism M is $(k\,\varepsilon, 0)$-DP for groups of size k. i.e., for all

$$||x - y||_1 \leq k$$

and any $S \subseteq$ range(M),

$$\Pr[M(x) \in S] \leq e^{\varepsilon k} \Pr[M(y) \in S]$$

post-processing

Thm. Let $M : \mathbb{N}^{|X|} \to R$ be (ε, δ)-DP.

Let f: $R \to R'$ be an arbitrary randomized mapping.

Then $f \circ M : \mathbb{N}^{|X|} \to R'$ is (ε, δ)-DP.

composition

[DworkKenthapadiMcSherryMironovNaor06,DworkLei09]

Thm. For $i \in [k]$, let $M_i : \mathbb{N}^{|X|} \to R_i$ be $(\varepsilon_i, \delta_i)$-DP. Then the mechanism $(M_1(x), \ldots, M_k(x))$ is $(\sum_i \varepsilon_i, \sum_i \delta_i)$-DP.

actually, holds even if subsequent computations chosen as function of previous results

"advanced" version

Is bigger delta better for privacy, or worse?

What about epsilon?

today

Formalizing privacy

Privacy properties and basic tools

Randomized Response

Randomized Response [Warner65]

flip a coin

if tails, respond truthfully

if heads, flip a second coin and respond "yes" if heads; respond "no" if tails

Claim. Randomized Response is (ln 3, 0)-DP.

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

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

Randomized Response

Given database $x = x_1, \ldots, x_n$ where $x_i \in \{0,1\}$

(say $x_i = 1$ if person committed crime
$= 0$ if person did not commit crime)

Query: $\sum_i x_i / n$ (= fraction of people that committed crime)

Mechanism:

Step 1. For $i = 1 \ldots n$

Let $y_i = x_i$ with probability $3/4$

$y_i = 1 - x_i$ with probability $1/4$

Step 2. Let $f(y_1 \ldots y_n) = \sum_i y_i / n$

output $f(y_1 \ldots y_n)$

Lemma Mechanism is $(\ln 3, 0)$-dp

pf First we show that the output of step 1 $Y_1 \dots Y_n$ is $(\ln 3, 0)$-dp. Then by post processing, $f(Y_1 \dots Y_n)$ is also $(\ln 3, 0)$-dp.

Consider 2 neighboring databases
$X = x_1 x_2 \dots x_n$
$X' = x_1 \dots \bar{x}_i \dots x_n$
} differ only on coord. i

Show $\forall y_1 \dots y_n \quad \dfrac{Pr(y_1 \dots y_n | x_1 \dots x_n)}{Pr(y_1 \dots y_n | x_1 \dots \bar{x}_i \dots x_n)} \leq 3$

$$\frac{Pr(y_1 \dots y_n | x_1 \dots x_n)}{Pr(y_1 \dots y_n | x_1 \dots \bar{x}_i \dots x_n)} = \frac{Pr(y_1|x_1) \cdot Pr(y_i|x_i) \dots Pr(y_n|x_n)}{Pr(y_1|x_1) \dots Pr(y_i|\bar{x}_i) \dots Pr(y_n|x_n)} = \frac{Pr(y_i|x_i)}{Pr(y_i|\bar{x}_i)}$$

$$\frac{Pr(Y_i|X_i)}{Pr(Y_i|\bar{X}_i)} \leq \frac{3/4}{1/4} = 3 = e^{\varepsilon}$$

$$\Rightarrow \quad \varepsilon = \ln 3$$

---

accuracy of Randomized Response:

Let n' = # of respondents who say 1 $\left(= \sum_i Y_i\right)$

Let $p = \sum_i X_i / n$

$E(n') = (pn)\frac{3}{4} + (1-p)n\frac{1}{4} = p\frac{n}{2} + \frac{n}{4}$

So max likelihood estimator of p, $\hat{p}$ is $\left(n' - \frac{n}{4}\right)\frac{2}{n} = \frac{2n'}{n} - \frac{1}{2}$

and variance of $\hat{p} = \frac{p(1-p)}{n} + \frac{\frac{3}{4}\left(\frac{1}{4}\right)}{n\left(2\cdot\frac{3}{4} - 1\right)^2} = \frac{p(1-p)}{n} + \frac{3}{4n}$