Differential Privacy in the Wild

A Tutorial on Current Practices and Open Challenges

About the Presenters

Ashwin Machanavajjhala
Assistant Professor, Duke University

"What does privacy mean … mathematically?"

Michael Hay
Assistant Professor, Colgate University

"Can algorithms be provably private and useful?"

Xi He
Ph.D. Candidate, Duke University

"Can privacy algorithms work in real world systems?"

Our world is increasingly data driven

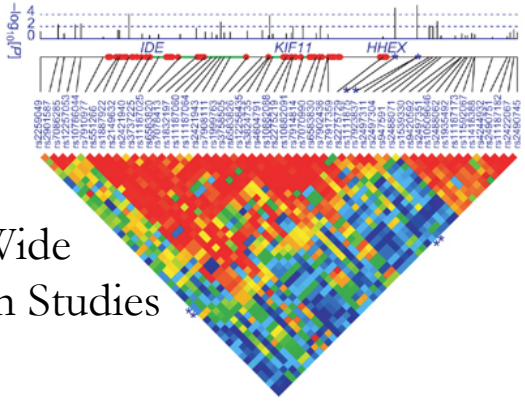
Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)

Aggregated Personal Data is **invaluable**

Advertising

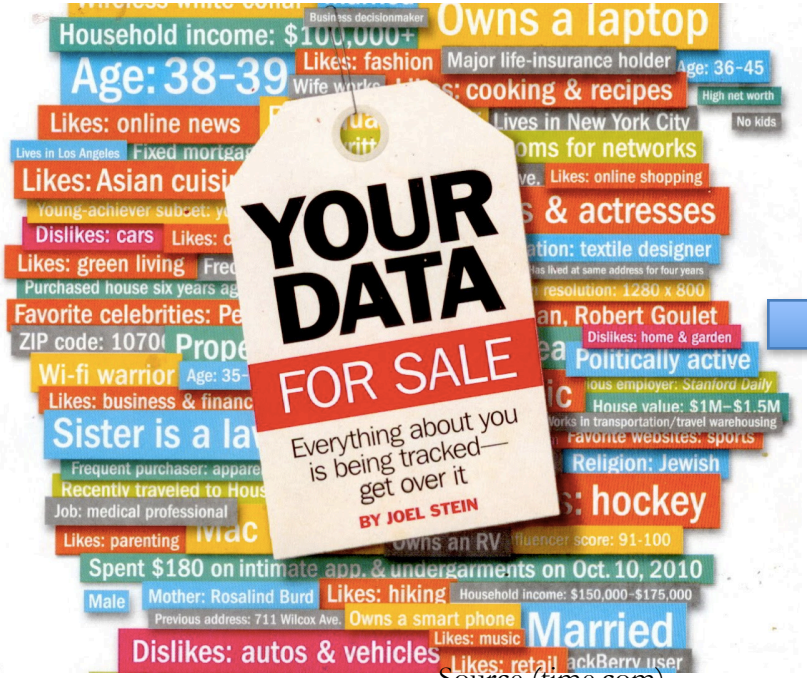
Source (esri.com)

Genome Wide Association Studies



Human Mobility analysis

Personal data is … well … **personal**!



Source (time.com)

Age
Income
Address
Likes/Dislikes
Sexual Orientation
Medical History

Redlining

Discrimination

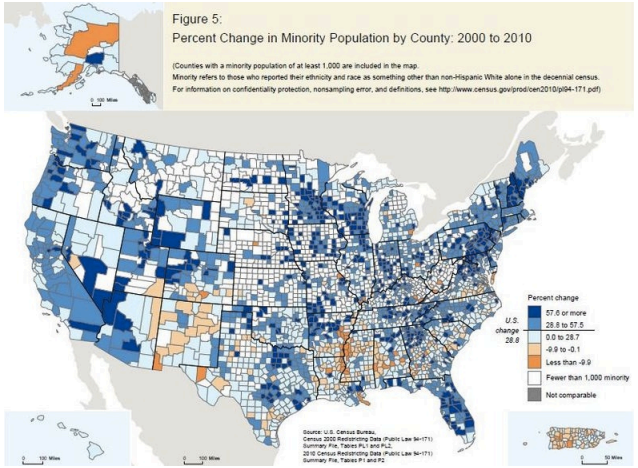
Physical/Financial Harm

Aggregated Personal Data …

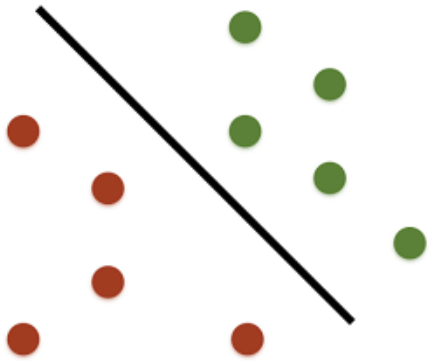
… is made publicly available in many forms.

De-identified records
(e.g., medical)

Statistics
(e.g., demographic)



Predictive models
(e.g., advertising)



That's fine … I am anonymous!

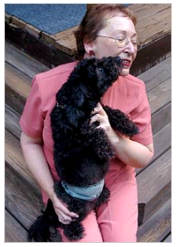
Source (http://xkcd.org/834/)

Anonymity is not enough …

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS



NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The Scientist » The Nutshell

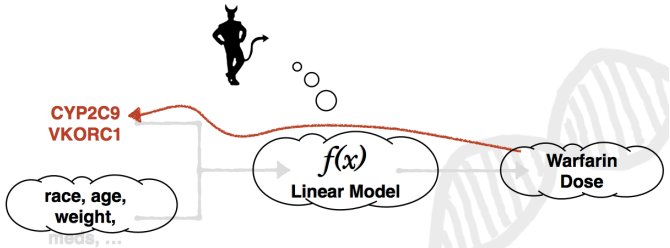
"Anonymous" Genomes Identified

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013

… and predictive models can breach privacy too

Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing



Need data analysis algorithms that can mine aggregated personal data with provable guarantees of privacy for individuals.

This is the goal of Differential Privacy.

Outline of the Tutorial

1. What is Privacy?
2. Differential Privacy
3. Answering Queries on Tabular Data

   Break
4. Applications I: Machine Learning
5. Privacy in the Real World
6. Applications II: Networks and Trajectories

Module 1: What is Privacy?

- Privacy Problem Statement

- What privacy is *not* …
  - Encryption
  - Anonymization
  - Restricted Query Answering

- What *is* privacy?

Module 2: Differential Privacy

- Differential Privacy Definition
- Basic Algorithms
  - Laplace & Exponential Mechanism
  - Randomized Response
- Composition Theorems

Module 3: Answering queries on Tabular data

- Answering query workloads on tabular databases
- Theory: two seminal results
- Survey of algorithm design ideas
  - Low dimensional range queries
  - Queries on high dimensional data
- Open Questions

Module 4: Applications I

- Private Empirical Risk Minimization
  - E.g. SVM, logistic regression
  - Make a specific learning approach private
- Private Stochastic Gradient Descent
  - E.g. Deep learning
  - Make a general purposed fitting technique private
- Other Important Problems in Private Learning

Module 5: Privacy in the Real World

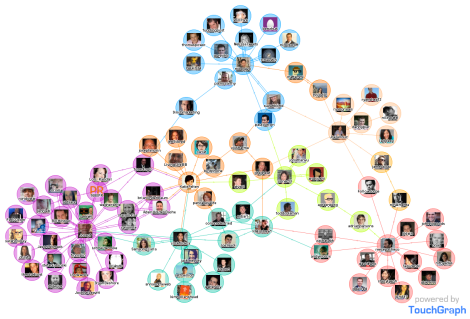
- Real world deployments of differential privacy
  - OnTheMap 

  - RAPPOR  chrome
- Privacy beyond Tabular Data
  - No Free Lunch Theorem
  - Customizing differential privacy using Pufferfish

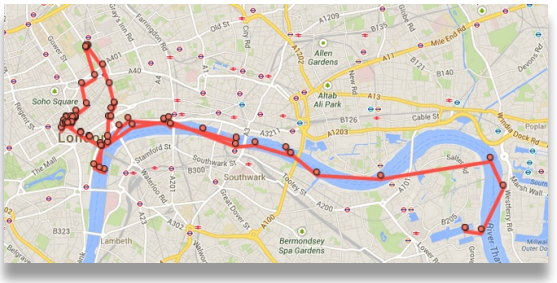
Module 6: Applications II

- Pufferfish Privacy for Non-tabular Data

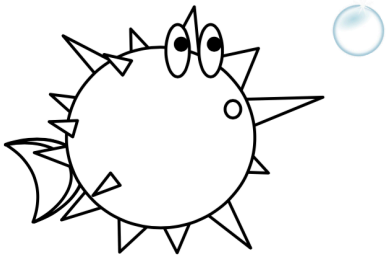
Social network



Location trajectories



- Blowfish Privacy



Scope of the Tutorial

What we do not cover:

- Securing data using encryption
- Computation on encrypted data
- Computationally bounded DP
- De-anonymization
- Anonymization schemes (k-anonymity, l-diversity, etc.)
- Access control

MODULE 1: PROBLEM FORMULATION

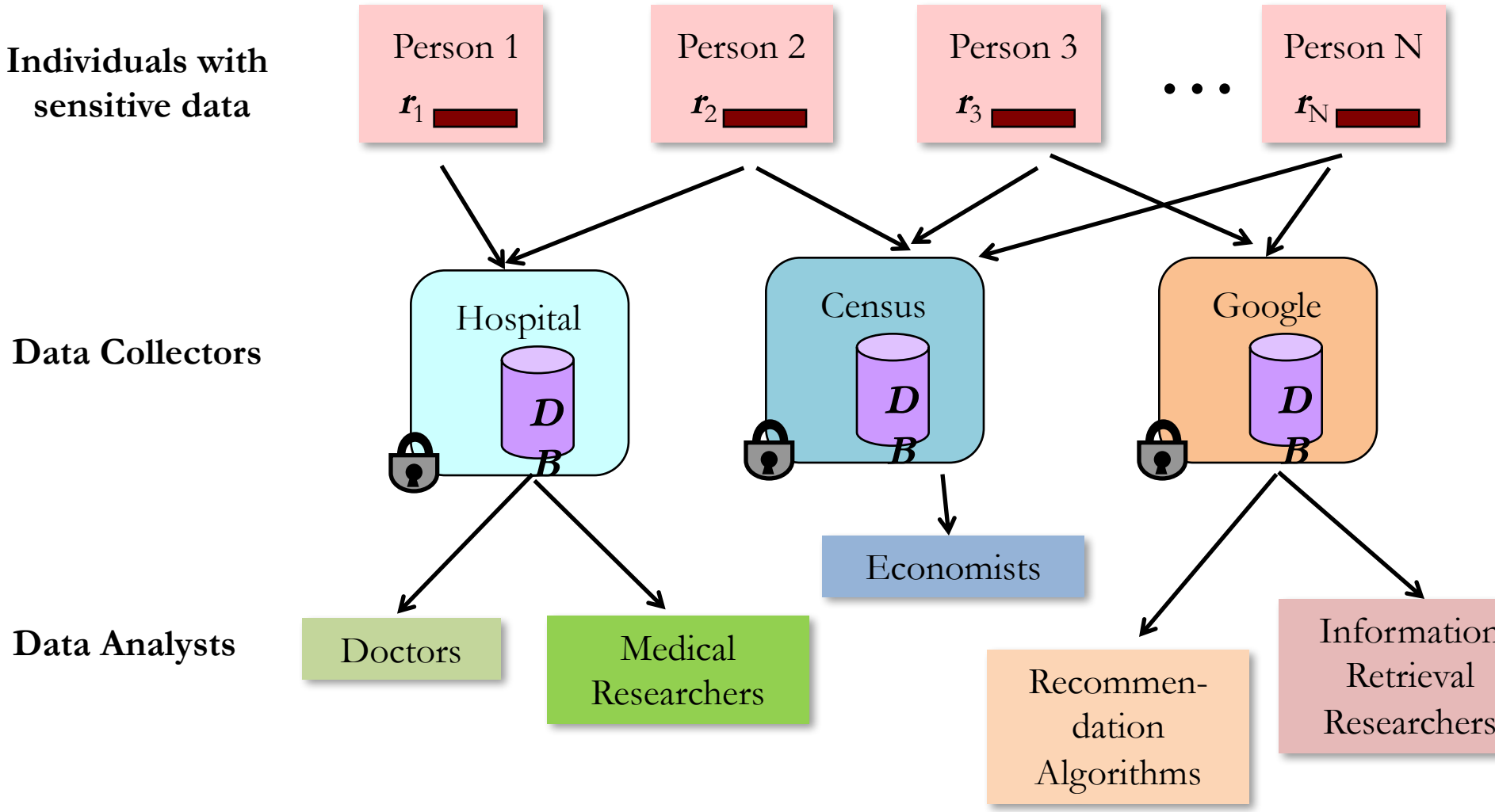
Module 1: What is Privacy?

- Privacy Problem Statement

- What privacy is *not* …
  - Encryption
  - Anonymization
  - Restricted Query Answering

- What *is* privacy?

Statistical Databases



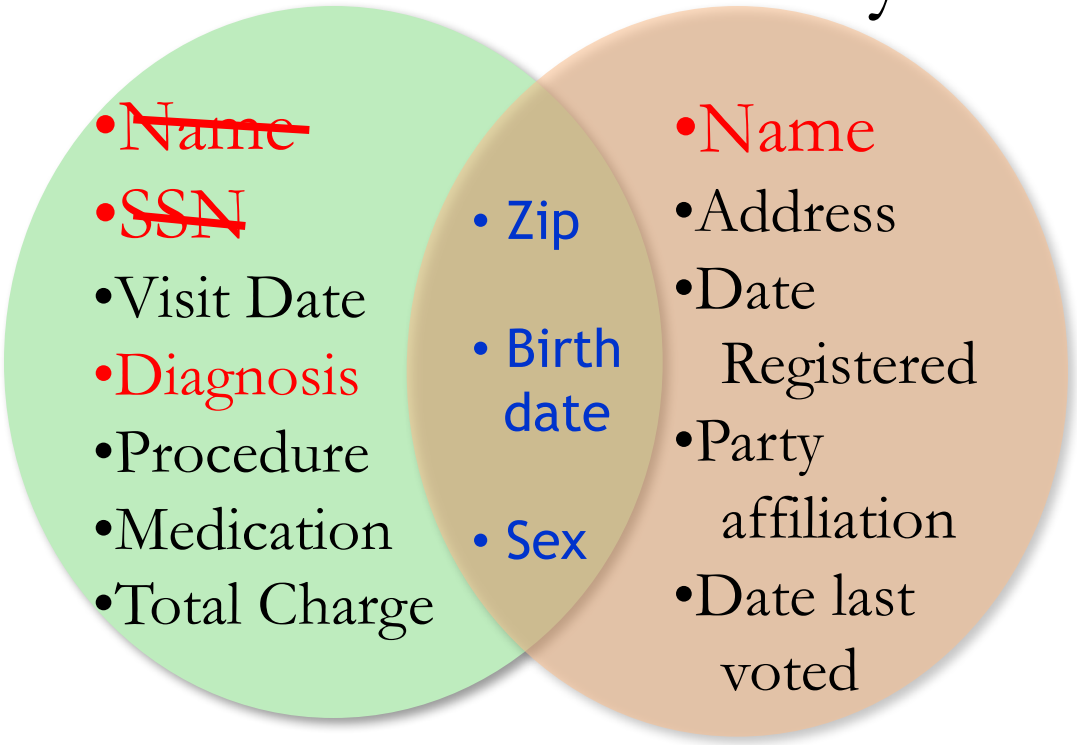
The Massachusetts Governor Privacy Breach

- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

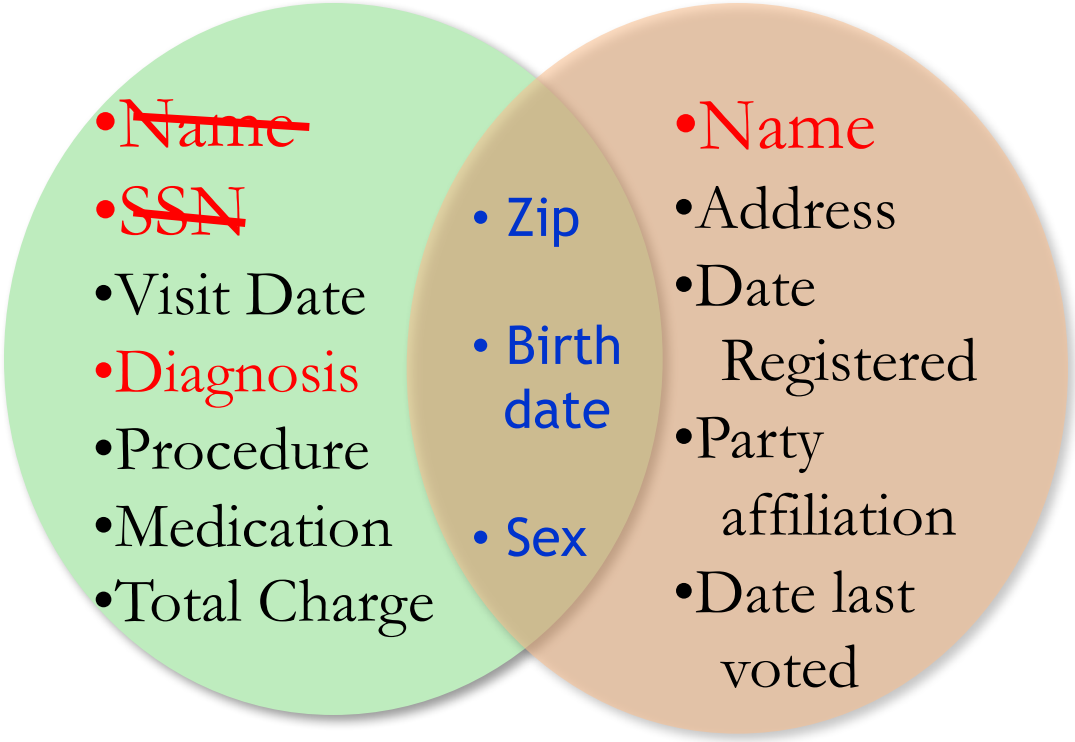
- Zip
- Birth date
- Sex

Medical Data Release

The Massachusetts Governor Privacy Breach



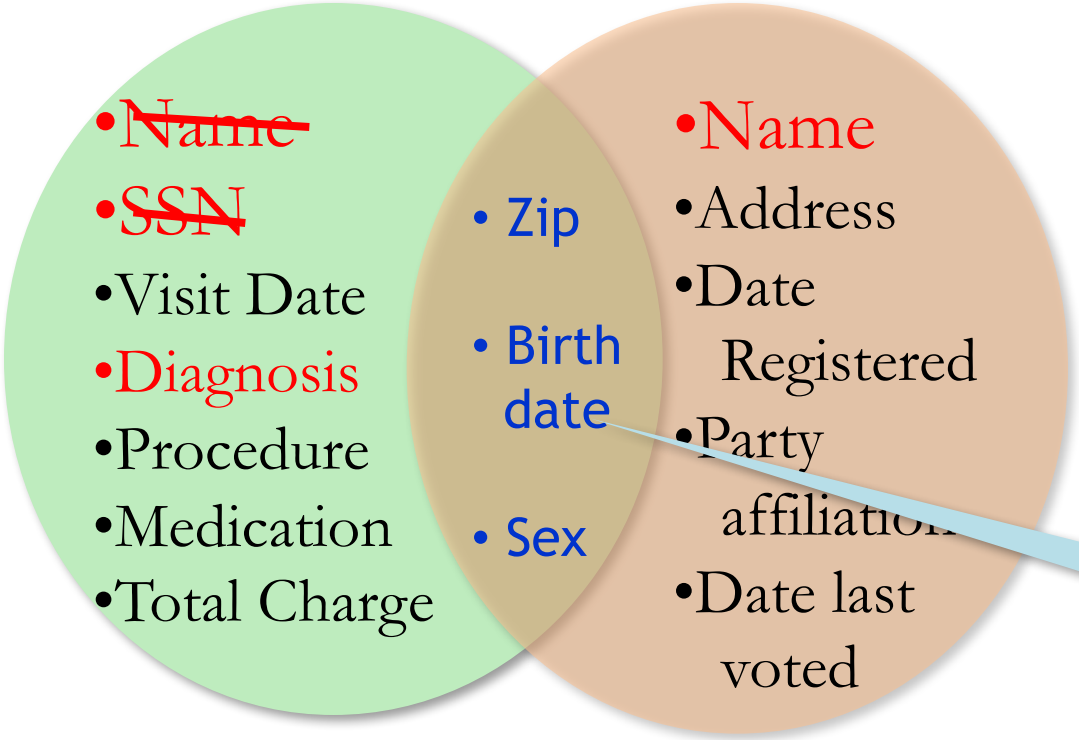
Linkage Attack



- Governor of MA **uniquely identified** using ZipCode, Birth Date, and Sex.

Name linked to Diagnosis

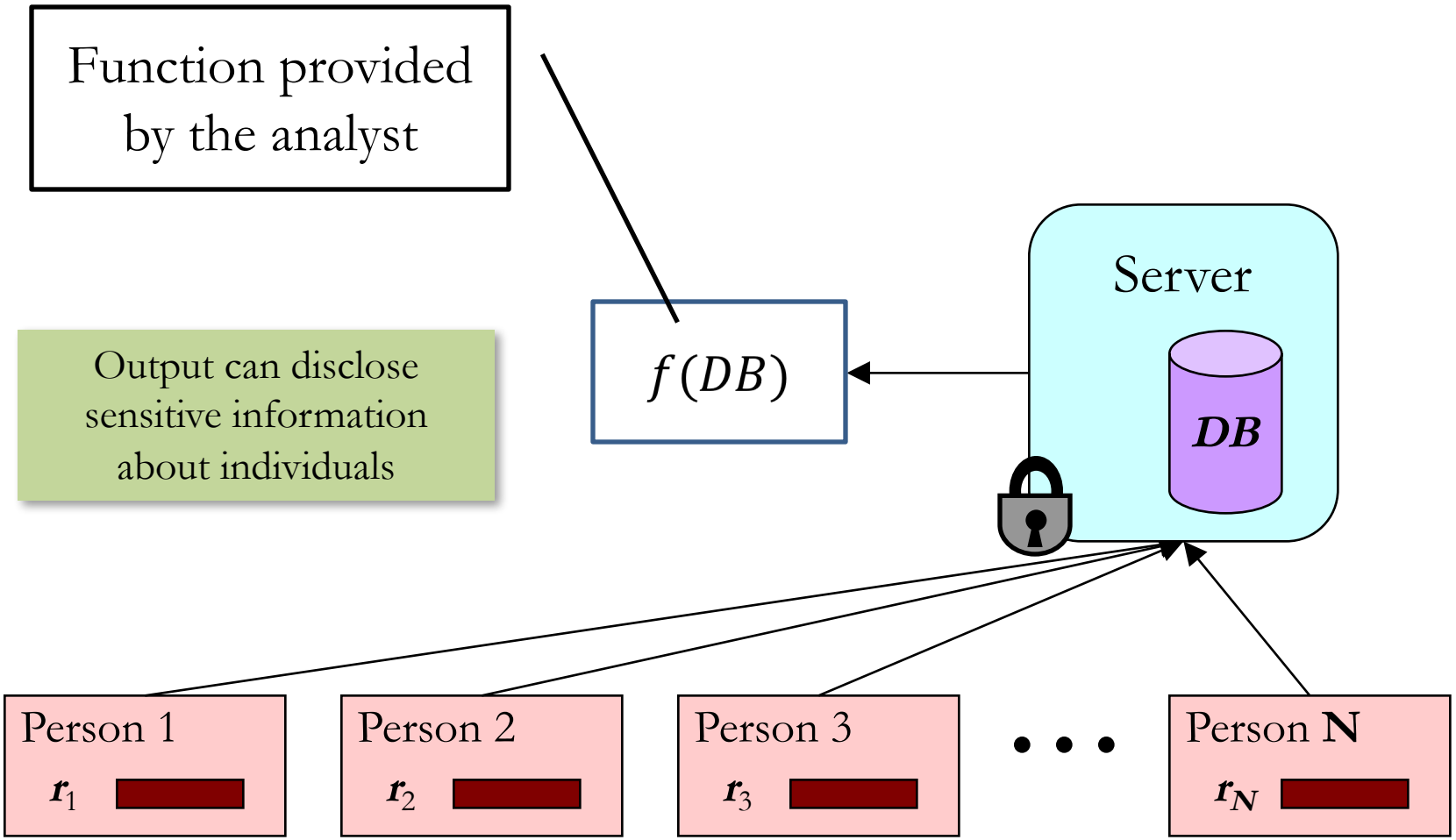
Linkage Attack



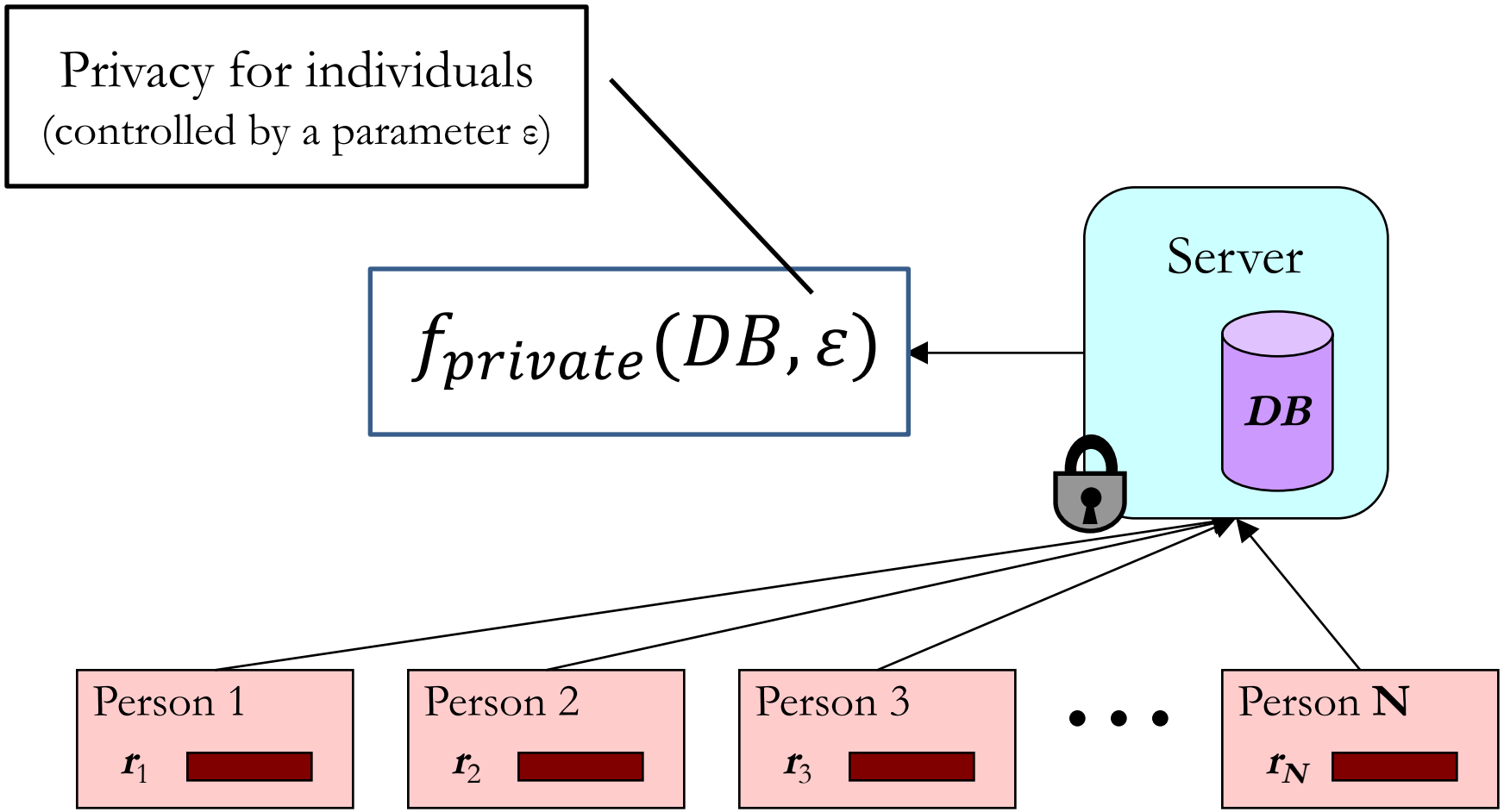
- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

Quasi Identifier

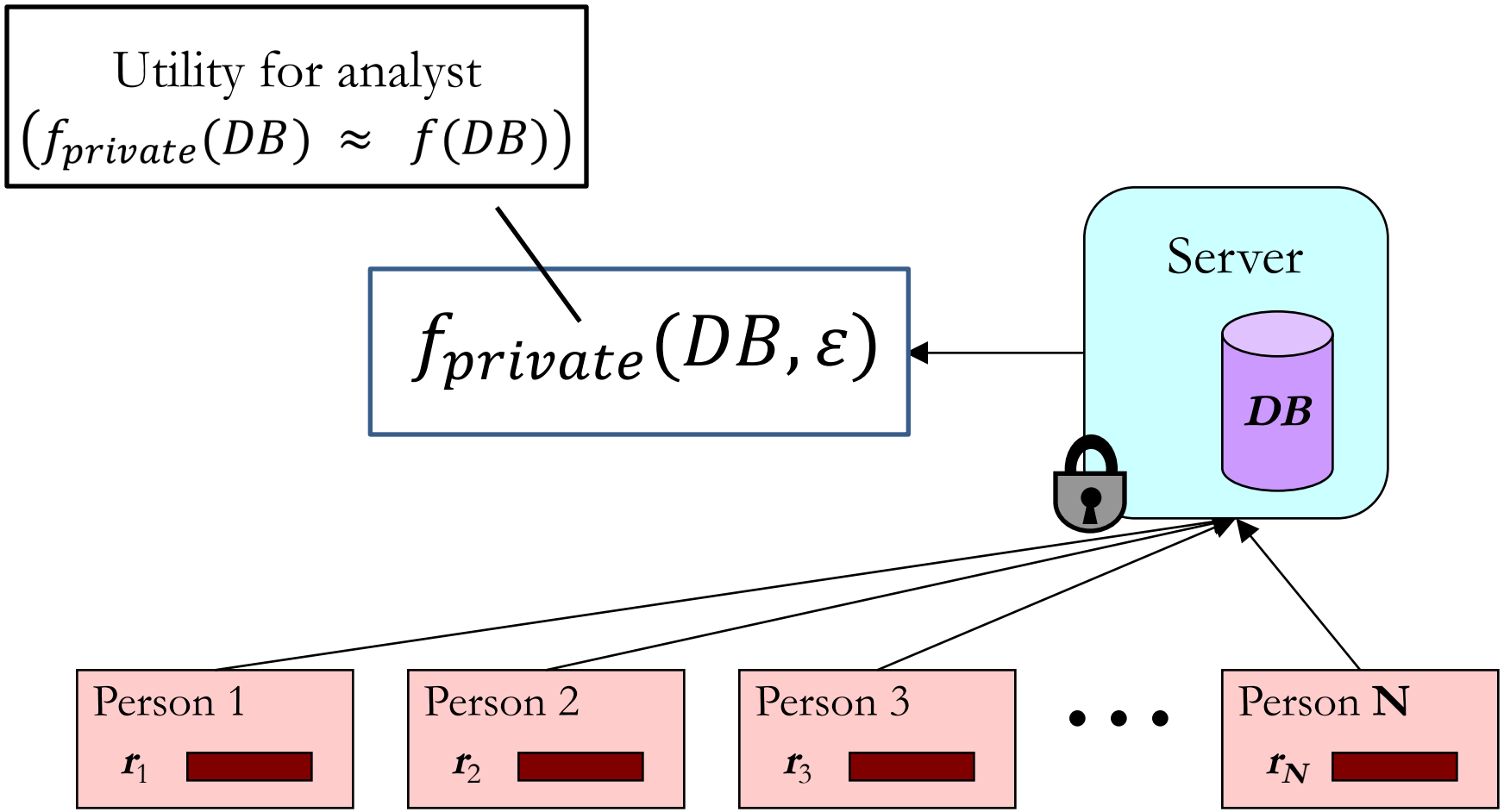
Statistical Database Privacy



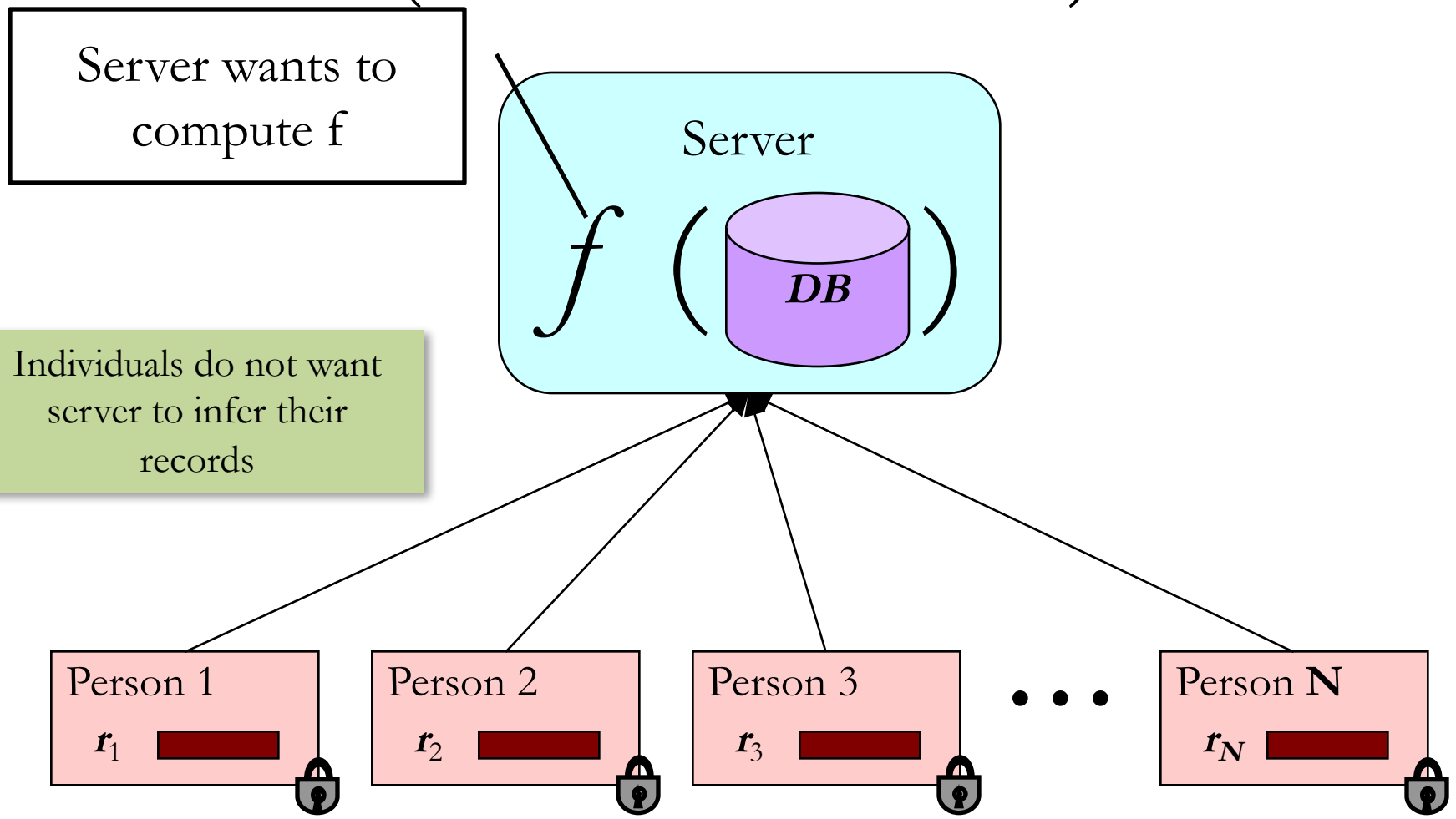
Statistical Database Privacy



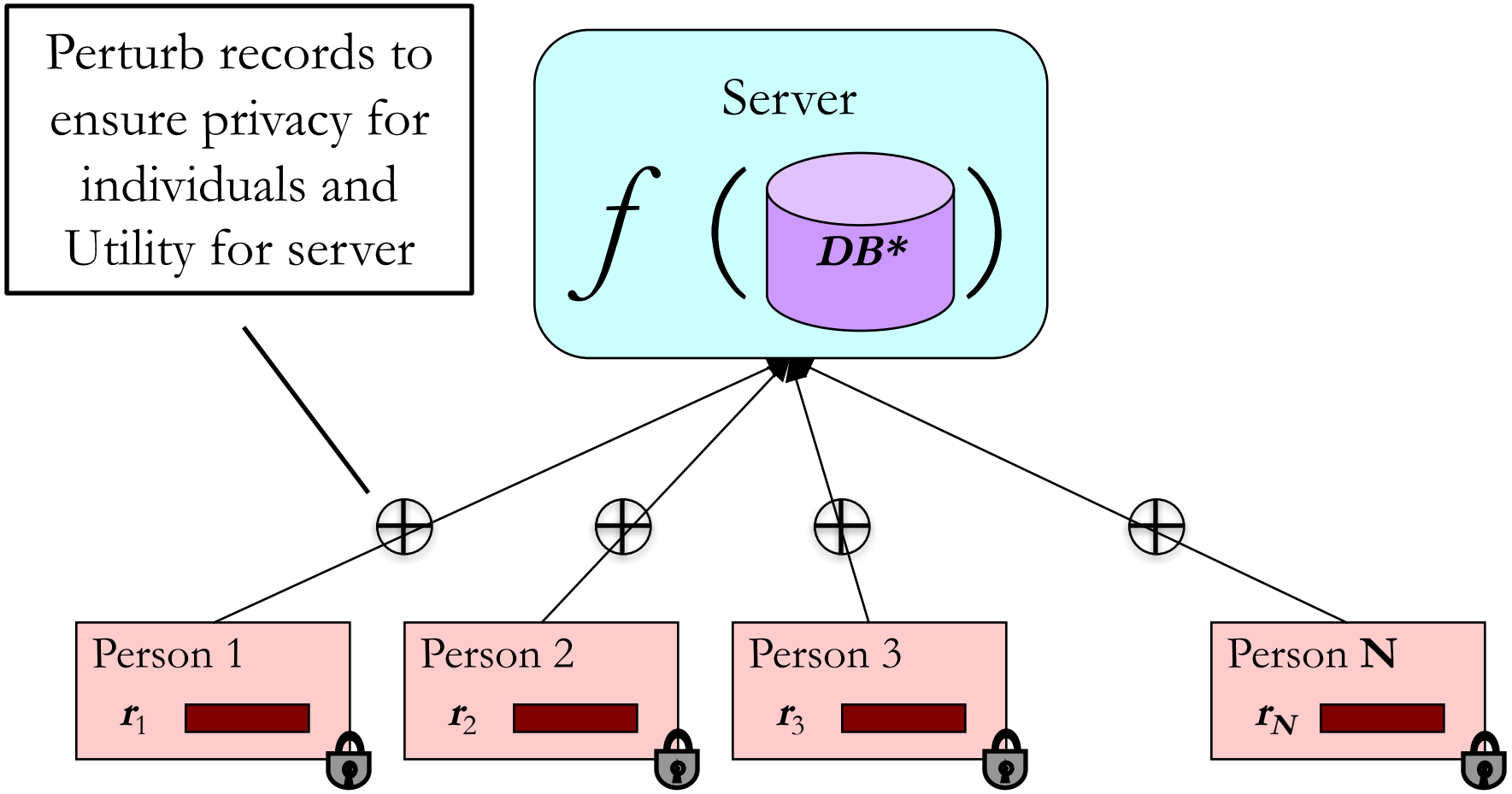
Statistical Database Privacy



Statistical Database Privacy (untrusted collector)



Statistical Database Privacy (untrusted collector)



Statistical Databases in real-world applications

Application	Data Collector	Private Information	Analyst	Function (utility)
Medical	Hospital	Disease	Epidemiologist	Correlation between disease and geography
Genome analysis	Hospital	Genome	Statistician/ Researcher	Correlation between genome and disease
Advertising	Google/FB/Y!	Clicks/Browsing	Advertiser	Number of clicks on an ad by age/region/gender …
Social Recommen-dations	Facebook	Friend links / profile	Another user	Recommend other users or ads to users based on social network

Statistical Databases in real-world applications

- Settings where data collector may not be trusted (or may not want the liability …)

Application	Data Collector	Private Information	Function (utility)
Location Services	Verizon/AT&T	Location	Traffic prediction
Recommen-dations	Amazon/Google	Purchase history	Recommendation model
Traffic Shaping	Internet Service Provider	Browsing history	Traffic pattern of groups of users

Privacy is *not* …

Statistical Database Privacy is not …

- Encryption:

Statistical Database Privacy is not …

- Encryption:
  Alice sends a message to Bob such that Trudy (attacker) does not learn the message. Bob should get the correct message …

- Statistical Database Privacy:
  Bob (attacker) can access a database
  - Bob must learn aggregate statistics, but
  - Bob must not learn new information about individuals in database.

Statistical Database Privacy is not …

- Computation on Encrypted Data:

Statistical Database Privacy is not …

- Computation on Encrypted Data:
  - Alice stores encrypted data on a server controlled by Bob (attacker).
  - Server returns correct query answers to Alice, without Bob learning *anything* about the data.

- Statistical Database Privacy:
  - Bob is allowed to learn aggregate properties of the database.

Statistical Database Privacy is not …

- The Millionaires Problem:

Statistical Database Privacy is not …

- Secure Multiparty Computation:
  - A set of agents each having a private input xi …
  - … Want to compute a function f(x1, x2, …, xk)
  - Each agent can learn the true answer, but must learn no other information than what can be inferred from their private input and the answer.

- Statistical Database Privacy:
  - Function output *must not disclose* individual inputs.

Statistical Database Privacy is not …

- Access Control:

Statistical Database Privacy is not …

- Access Control:
  - A set of agents want to access a set of resources (could be files or records in a database)
  - Access control rules specify who is allowed to access (*or not access*) certain resources.
  - 'Not access' usually means no information must be disclosed

- Statistical Database:
  - A single database and a single agent
  - Want to release aggregate statistics about a set of records without allowing access to individual records

Privacy Problems

- In todays cloud context a number of privacy problems arise:
  - Encryption when communicating data across a unsecure channel
  - Secure Multiparty Computation when different parties want to compute on a function on their private data without using a centralized third party
  - Computing on encrypted data when one wants to use an unsecure cloud for computation
  - Access control when different users own different parts of the data

- Statistical Database Privacy:
  Quantifying (and bounding) the amount of information disclosed about individual records by the output of a valid computation.

What *is* privacy?

Privacy Breach: Informal Definition

A privacy mechanism M(D)

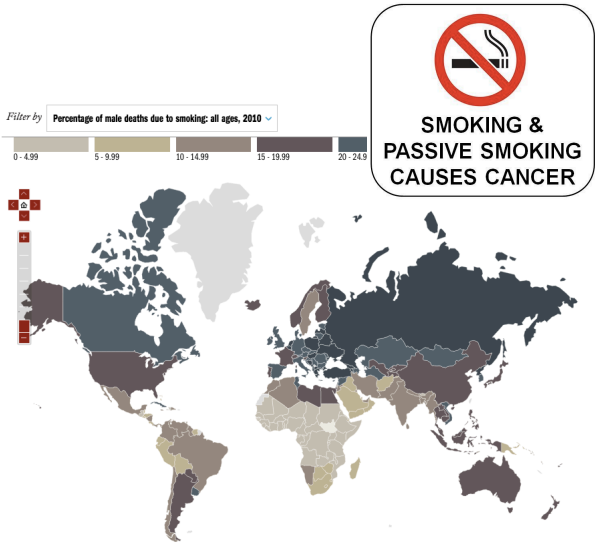
that allows

an unauthorized party 

to learn sensitive information about any individual in D,

which  could not have learnt without access to M(D).

Alice



Alice has Cancer

Is this a privacy breach? *NO*

Privacy Breach: Revised Definition

A privacy mechanism M(D) that allows
an unauthorized party 
to learn sensitive information about
any individual Alice in D,

which  could not have learnt without access to M(D)
if Alice was *not in the dataset*.

K-Anonymity: Avoiding Linkage Attacks

- If every row corresponds to one individual …

  … every row should look like k-1 other rows based on the *quasi-identifier* attributes

K-Anonymity

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

[MKGV 06]

Problem: Background knowledge

Adversary knows prior knowledge about Umeko

Adversary learns Umeko has Cancer

Name	Zip	Age	Nat.
Umeko	13053	25	Japan

Zip	Age	Nationality	Disease
130**	<30	*	~~Heart~~
130**	<30	*	~~Heart~~
130**	<30	*	Cancer
130**	<30	*	Cancer
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

A privacy mechanism must be able to protect individuals' privacy from attackers who may possess **background knowledge**

Healthcare Cost and Utilization Project

#Hospital discharges in NJ of ovarian cancer patients, 2009

Counts less than k are suppressed achieving k-anonymity

Age	#discharges	White	Black	Hispanic	Asian/Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	*	19	22
1-17	*	*	*	*	*	*	*	*
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	**1**	19	22
1-17	**3**	**1**	*	*	*	*	*	*
18-44	70	40	13	*				*
45-64	330	236	31	32			1	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

= 535 – (40+236+229+29)

#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	**1**	19	22
1-17	**3**	**1**	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

#Hospital discharges in NJ of ovarian cancer patients, 2009

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	**1**	19	22
1-17	**3**	**1**	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	[1-3]	*	*	*	*	*

[VSJO 13]

Can reconstruct tight bounds on rest of data

In fact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in new Jersey with more than 435 beds in 2009. Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.

Multiple Release problem

- Privacy preserving access to data must necessarily release some information about individual records (to ensure utility)

- However, k-anonymous algorithms can reveal individual level information even with two releases.

A privacy mechanism must satisfy **composition** …

… or allow a graceful degradation of privacy with multiple invocations on the same data.

[DN03, GKS08]

Postprocessing the output of a privacy mechanism must not change the privacy guarantee

[KL10, MK15]

Privacy must not be achieved through **obscurity.**

Attacker must be assumed to know the algorithm used as well as all parameters

Summary

- Statistical database privacy is the problem of releasing aggregates while not disclosing individual records
- The problem is distinct from encryption, secure computation and access control.
- Defining privacy is non-trivial
  - Desiderata include resilience to background knowledge and composition and closure under postprocessing.

MODULE 2: DIFFERENTIAL PRIVACY

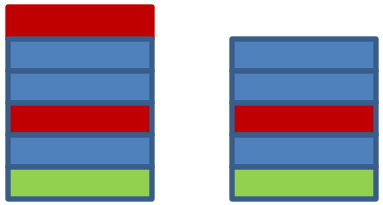
Module 2: Differential Privacy

- Differential Privacy Definition
- Basic Algorithms
  - Laplace & Exponential Mechanism
  - Randomized Response
- Composition Theorems

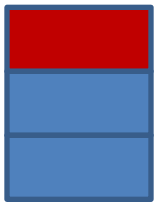
Differential Privacy

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

If algorithm A satisfies differential privacy then

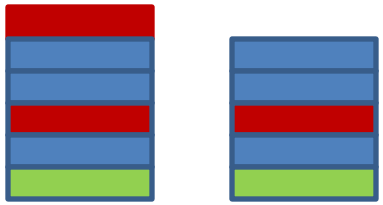
$$\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} < \exp(\varepsilon) \quad (\varepsilon > 0)$$

Intuition: adversary should not be able to use output O to distinguish between any D_1 and D_2

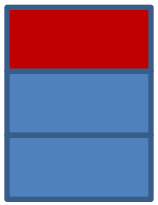
Why pairs of datasets *that differ in one row*?

For every pair of inputs that differ in one row

For every output …



D_1 D_2

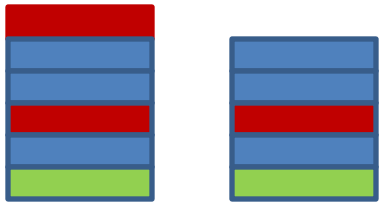


O

Simulate the presence or absence of a single record

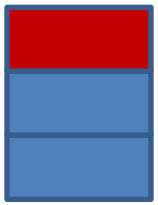
Why *all* pairs of datasets …?

For every pair of inputs that differ in one row



D_1 D_2

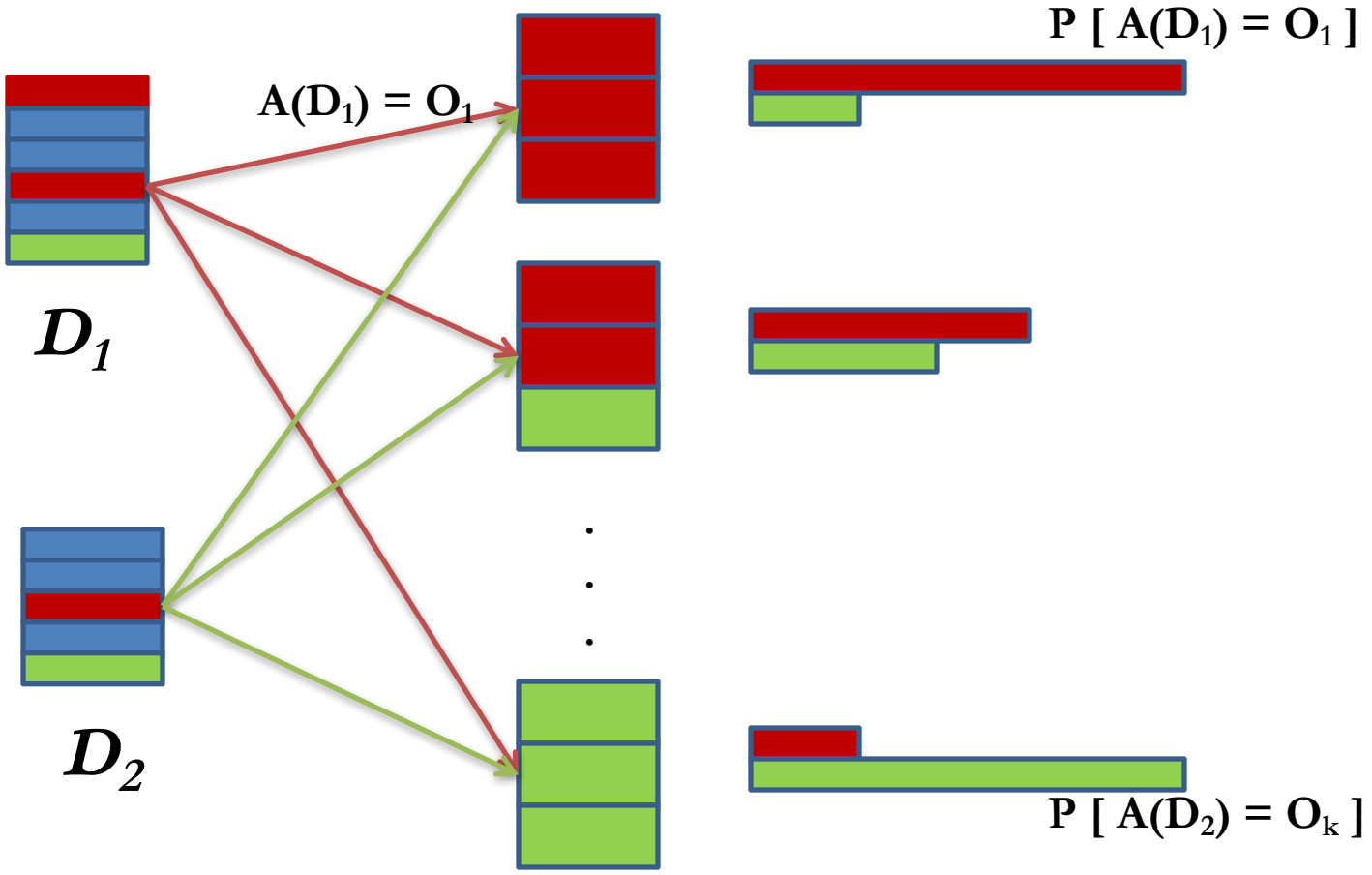
For every output …



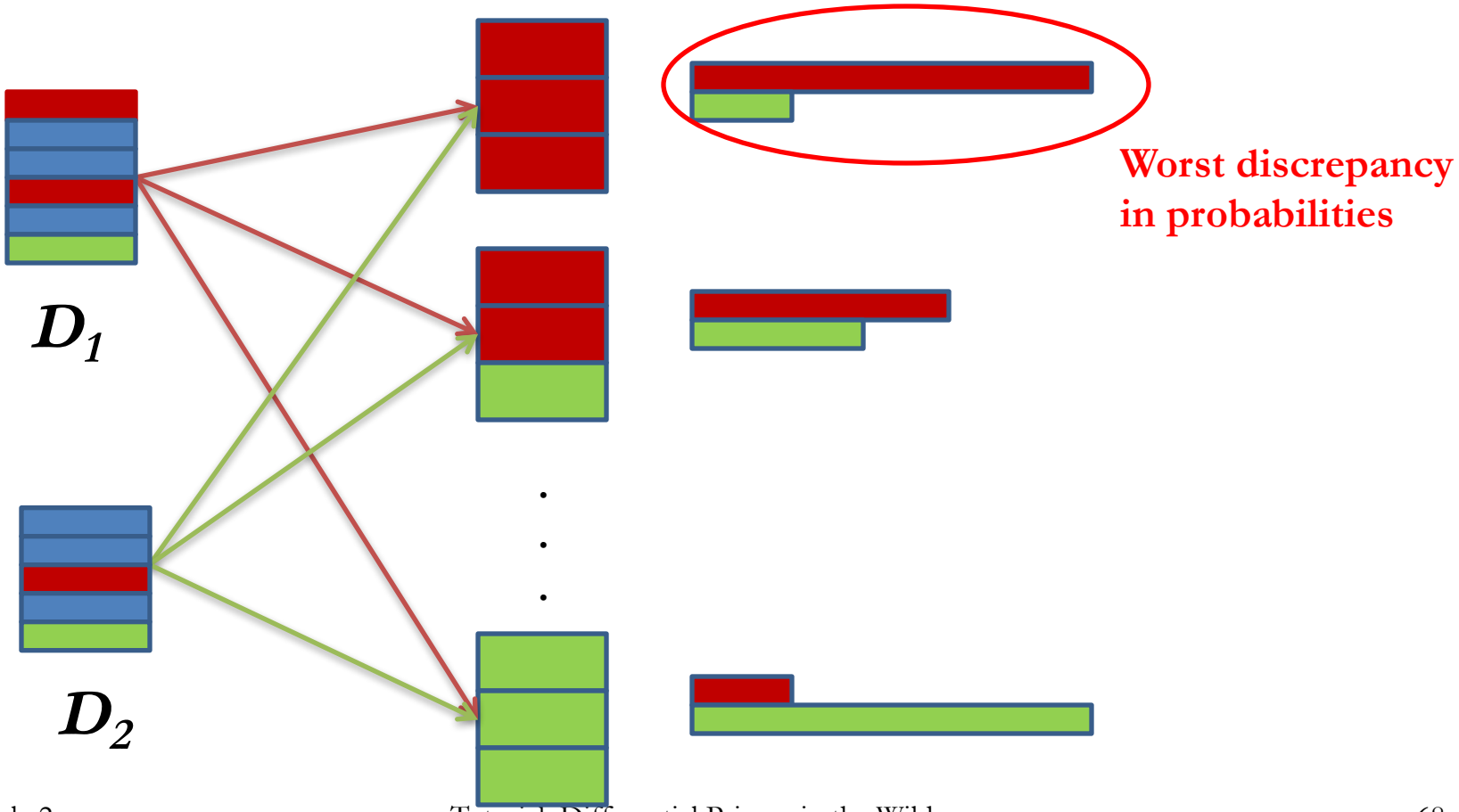
O

Guarantee holds no matter what the other records are.

Why *all* outputs?



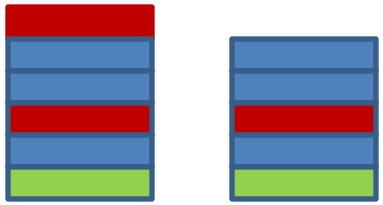
Should not be able to distinguish whether input was D_1 or D_2 no matter what the output



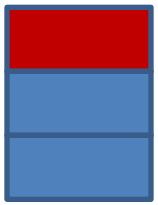
Privacy Parameter ε

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

$$\Pr[A(D_1) = O] \leq e^{\varepsilon} \Pr[A(D_2) = O]$$

Controls the degree to which D_1 and D_2 can be distinguished.
Smaller ε gives more privacy (and worse utility)

Outline of the Module 2

- Differential Privacy

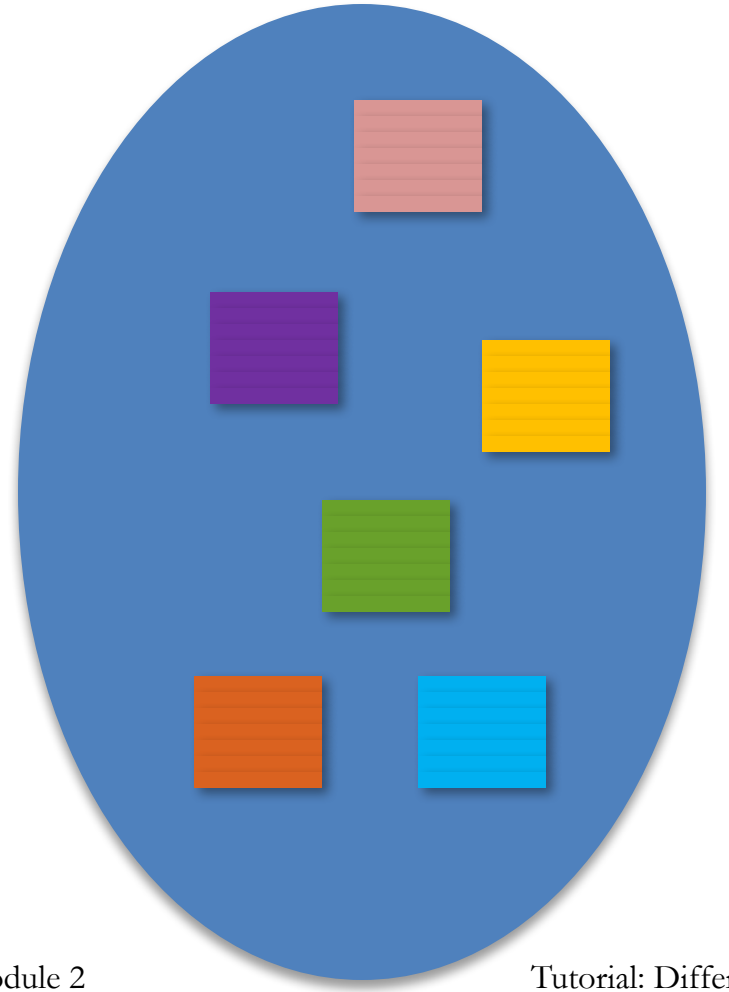
- Basic Algorithms
  - Laplace & Exponential Mechanism
  - Randomized Response

- Composition Theorems

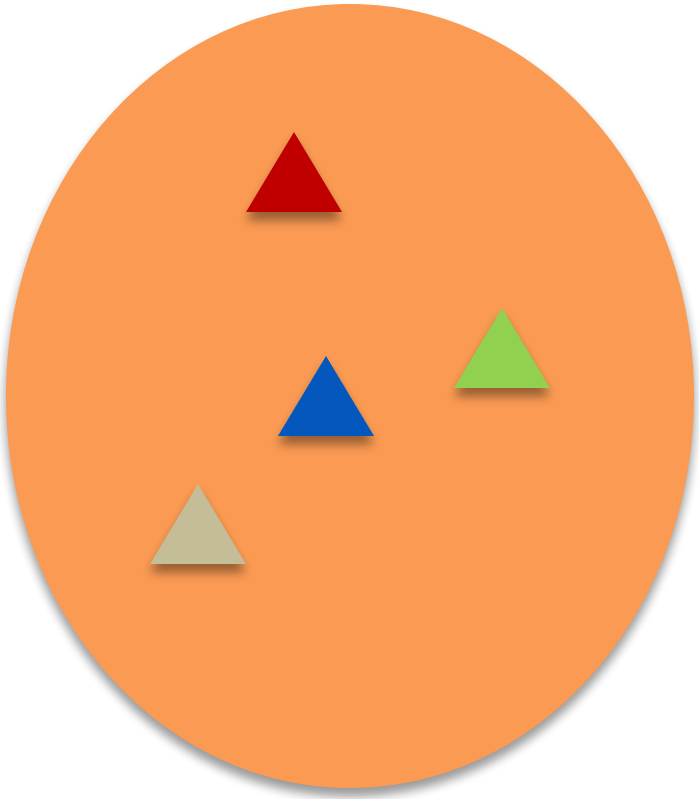
Can deterministic algorithms satisfy differential privacy?

Non trivial deterministic algorithms do not satisfy differential privacy

Space of all inputs

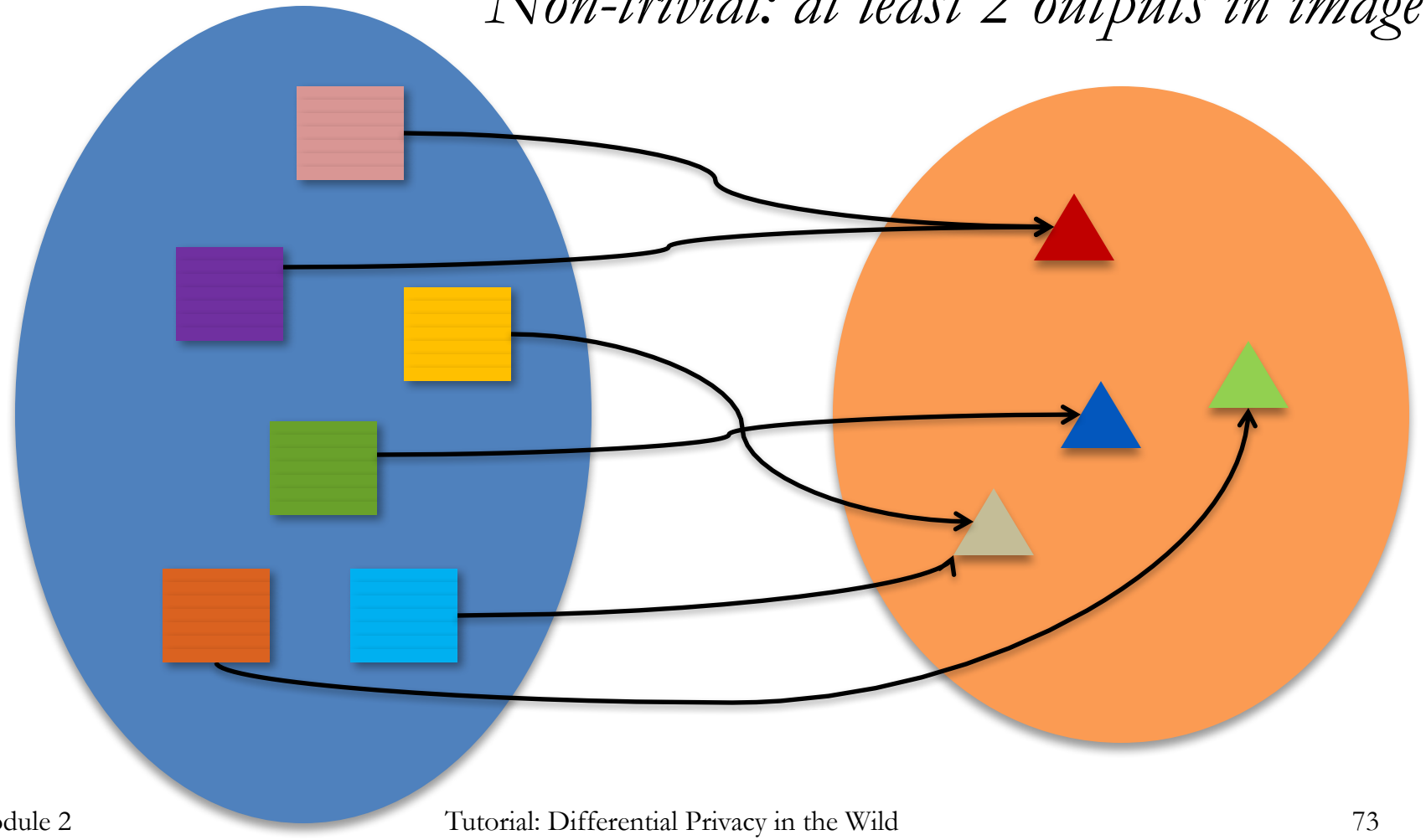


Space of all outputs

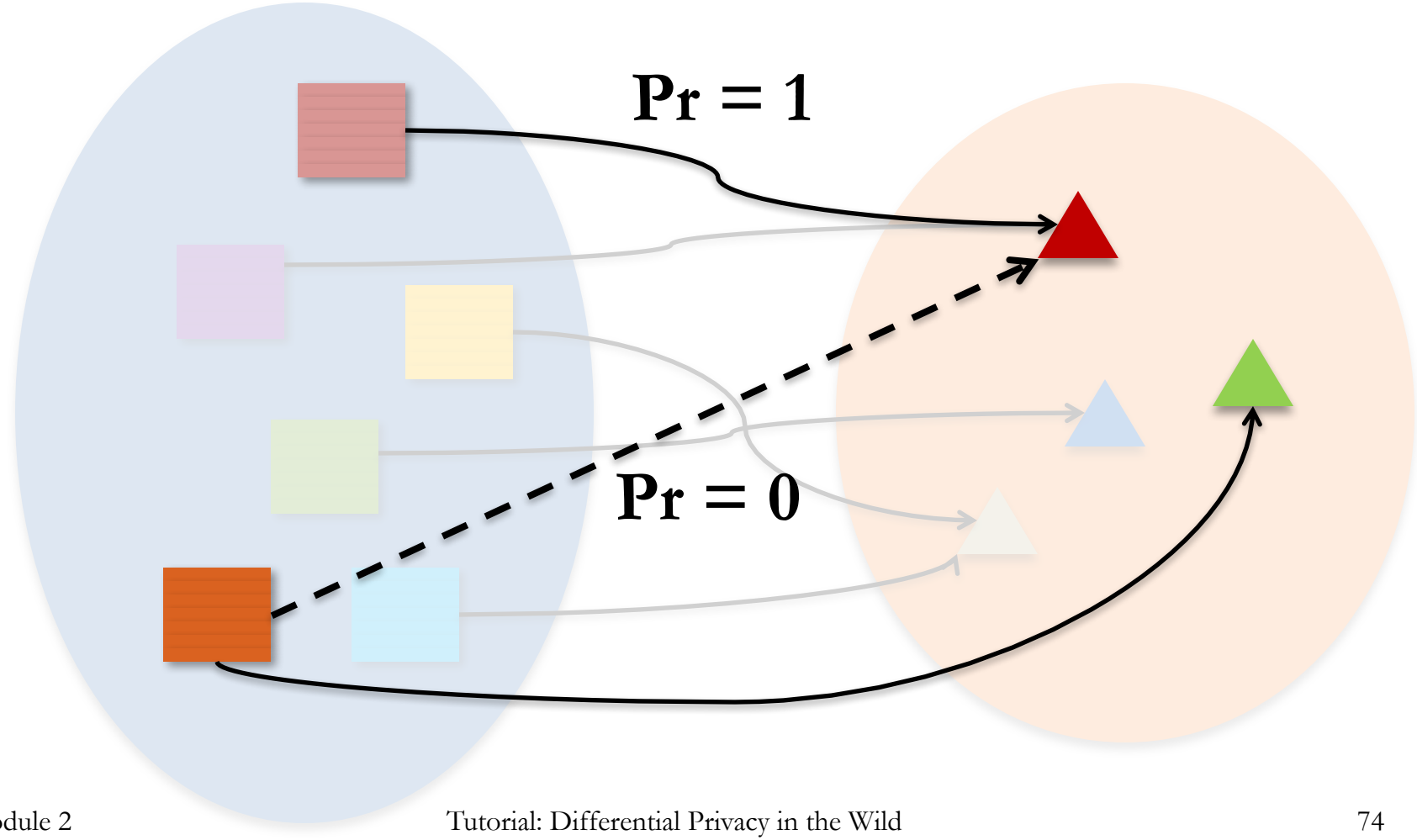


Non-trivial deterministic algorithms do not satisfy differential privacy

Non-trivial: at least 2 outputs in image

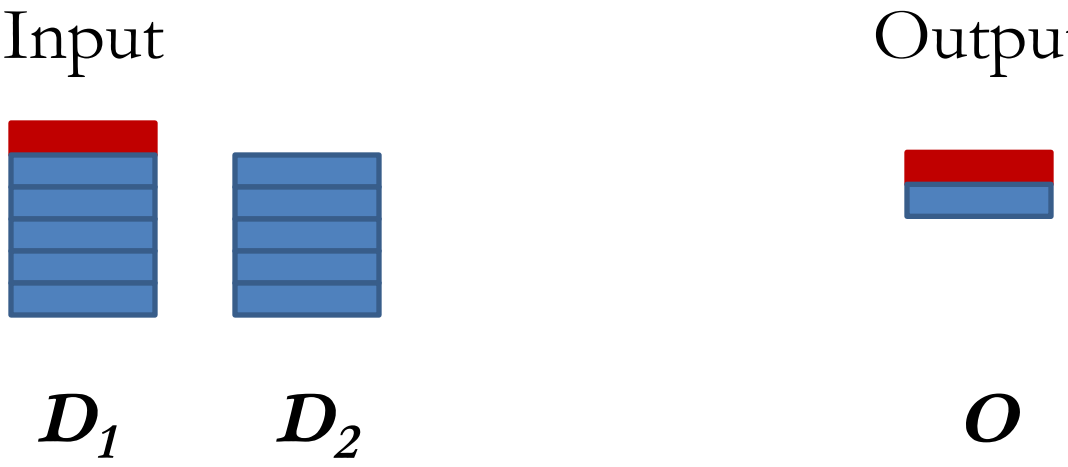


There exist two inputs that differ in one entry mapped to different outputs.



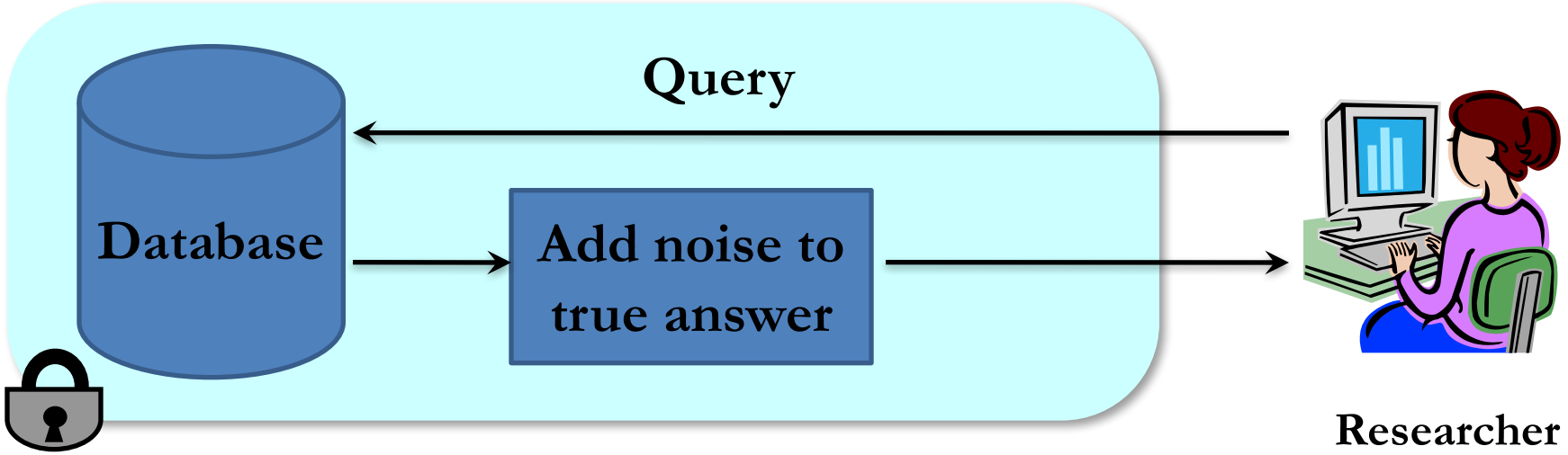
Random Sampling …

… also does not satisfy differential privacy



$$\Pr[D_2 \rightarrow O] = 0 \text{ implies } \frac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]} = \infty$$

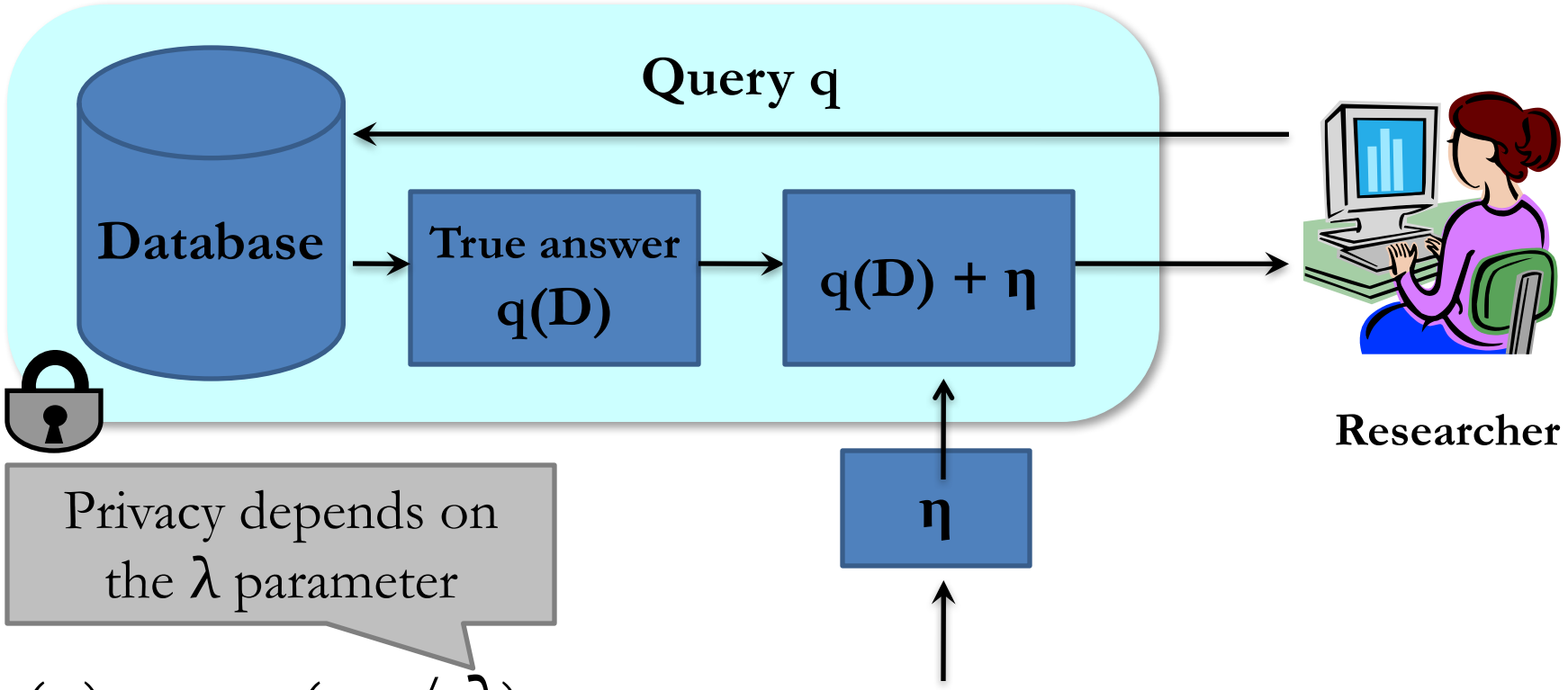
Output Randomization



- Add noise to answers such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers.

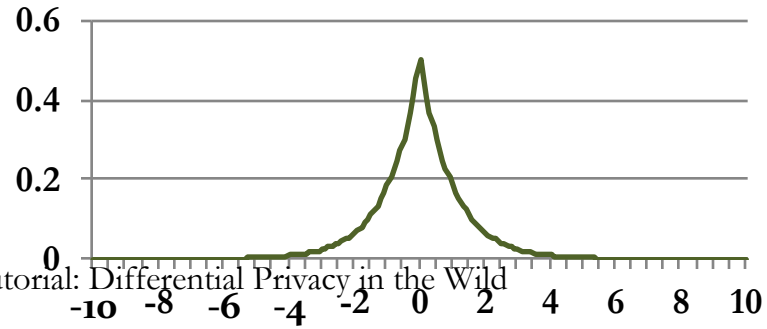
[DMNS 06]

Laplace Mechanism



$h(\eta) \propto \exp(-\eta / \lambda)$

Mean: 0,
Variance: $2\lambda^2$



How much noise for privacy?

[Dwork et al., TCC 2006]

Sensitivity: Consider a query q: $I \rightarrow R$. S(q) is the smallest number s.t. for any neighboring tables D, D',

$$| q(D) - q(D') | \leq S(q)$$

Thm: If **sensitivity** of the query is **S**, then the following guarantees **ε**-differential privacy.

$$\lambda = S/\varepsilon$$

Sensitivity: COUNT query

- Number of people having disease
- Sensitivity = 1

- Solution: $3 + \eta$,
  where η is drawn from Lap(1/ε)
  - Mean = 0
  - Variance = $2/\varepsilon^2$

D

Disease (Y/N)
Y
Y
N
Y
N
N

Sensitivity: SUM query

- Suppose all values x are in [a,b]

- Sensitivity = b

Privacy of Laplace Mechanism

- Consider neighboring databases D and D'
- Consider some output O

$$\frac{\Pr\left[A(D) = O\right]}{\Pr\left[A(D') = O\right]} = \frac{\Pr\left[q(D) + \eta = O\right]}{\Pr\left[q(D') + \eta = O\right]}$$

$$= \frac{e^{-|O-q(D)|/\lambda}}{e^{-|O-q(D')|/\lambda}}$$

$$\leq e^{|q(D)-q(D')|/\lambda} \leq e^{S(q)/\lambda} = e^{\varepsilon}$$

Utility of Laplace Mechanism

- Laplace mechanism works for **any function** that returns a real number

- Error: $E(\text{true answer} - \text{noisy answer})^2$

$$= \text{Var}(\text{Lap}(S(q)/\varepsilon))$$

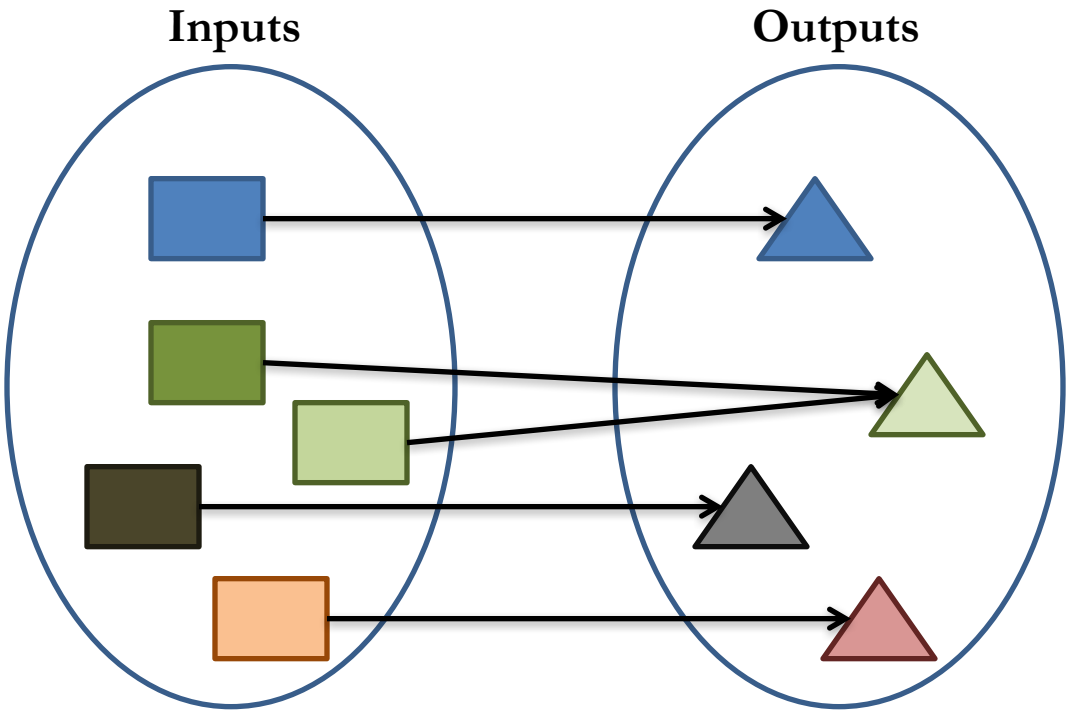
$$= 2*S(q)^2 / \varepsilon^2$$

Exponential Mechanism

- For functions that do not return a real number …
  - "what is the most common nationality in this room": Chinese/Indian/American…
- When perturbation leads to invalid outputs …
  - To ensure integrality/non-negativity of output

Exponential Mechanism

Consider some function f (can be deterministic or probabilistic):



How to construct a differentially private version of f?

Exponential Mechanism

- Scoring function *w: Inputs* $\times$ *Outputs* $\rightarrow$ *R*

- D: nationalities of a set of people
- #(D, O): # people with nationality O
- f(D): most frequent nationality in D
- w(D, O) = |#(D, O) - #(D, f(D))|

Exponential Mechanism

- Scoring function *w: Inputs* $\times$ *Outputs* → *R*

- Sensitivity of w

$$\Delta_w = \max_{O\&D,D'} |w(D,O) - w(D,O')|$$

where D, D' differ in one tuple

Exponential Mechanism

Given an input D, and a scoring function w,

Randomly sample an output O from *Outputs* with probability

$$\frac{e^{\frac{\varepsilon}{2\Delta} \cdot w(D,O)}}{\sum_{Q \epsilon Outputs} e^{\frac{\varepsilon}{2\Delta} \cdot w(D,Q)}}$$

- Note that for every output O, probability O is output > 0.

Randomized Response (a.k.a. local randomization)

D

Disease (Y/N)
Y
Y
N
Y
N
N

With probability p,
Report true value

With probability 1-p,
Report flipped value

O

Disease (Y/N)
Y
N
N
N
Y
N

Differential Privacy Analysis

- Consider 2 databases D, D' (of size M) that differ in the j^{th} value
  - D[j] ≠ D'[j]. But, D[i] = D'[i], for all i ≠ j

- Consider some output O

$$\frac{P(D \to O)}{P(D' \to O)} \leq e^{\varepsilon} \Leftrightarrow \frac{1}{1+e^{\varepsilon}} < p < \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$$

Utility Analysis

- Suppose n1 out of N people replied "yes", and rest said "no"
- What is the best estimate for π = fraction of people with disease = Y?
- Extract an estimate through *post-processing*

$$\pi_{hat} = \{n1/n - (1-p)\}/(2p-1)$$

- $E(\pi_{hat}) = \pi$
- $Var(\pi_{hat}) = \frac{\pi(1-\pi)}{n} + \frac{1}{n(16(p-0.5)^2 - 0.25)}$

Sampling **Variance due to coin flips**

Laplace Mechanism vs Randomized Response

Privacy

- Provide the same ε-differential privacy guarantee

- Laplace mechanism assumes data collected is trusted
- Randomized Response does not require data collected to be trusted
  - Also called a *Local* Algorithm, since each record is perturbed

Laplace Mechanism vs Randomized Response

Utility

- Suppose a database with N records where μN records have disease = Y.
- Query: # rows with Disease=Y

- Std dev of Laplace mechanism answer: $O(1/\varepsilon)$
- Std dev of Randomized Response answer: $O(\sqrt{N})$

Outline of the Module 2

- Differential Privacy
- Basic Algorithms
  - Laplace & Exponential Mechanism
  - Randomized Response
- Composition Theorems

Why Composition?

- Reasoning about privacy of a complex algorithm is hard.

- Helps software design
  - If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.

A bound on the number of queries

- In order to ensure utility, a statistical database must leak some information about each individual
- We can only hope to bound the amount of disclosure

- Hence, there is a limit on number of queries that can be answered

[DN03]

Dinur Nissim Result

- A vast majority of records in a database of size n can be reconstructed when $n \log(n)^2$ queries are answered by a statistical database …

  … even if each answer has been arbitrarily altered to have up to $o(\sqrt{n})$ error

Sequential Composition

- If M_1, M_2, ..., M_k are algorithms that access a private database D such that each M_i satisfies ε_i -differential privacy,

  then running all k algorithms sequentially satisfies ε-differential privacy with $\varepsilon=\varepsilon_1+...+\varepsilon_k$

Privacy as Constrained Optimization

- Three axes
  - Privacy
  - Error
  - Queries that can be answered

- E.g.: Given a fixed set of queries and **privacy budget** ε, what is the minimum error that can be achieved?

Parallel Composition

- If $M_1, M_2, ..., M_k$ are algorithms that access disjoint databases $D_1, D_2, \ldots, D_k$ such that each M_i satisfies ε_i -differential privacy,

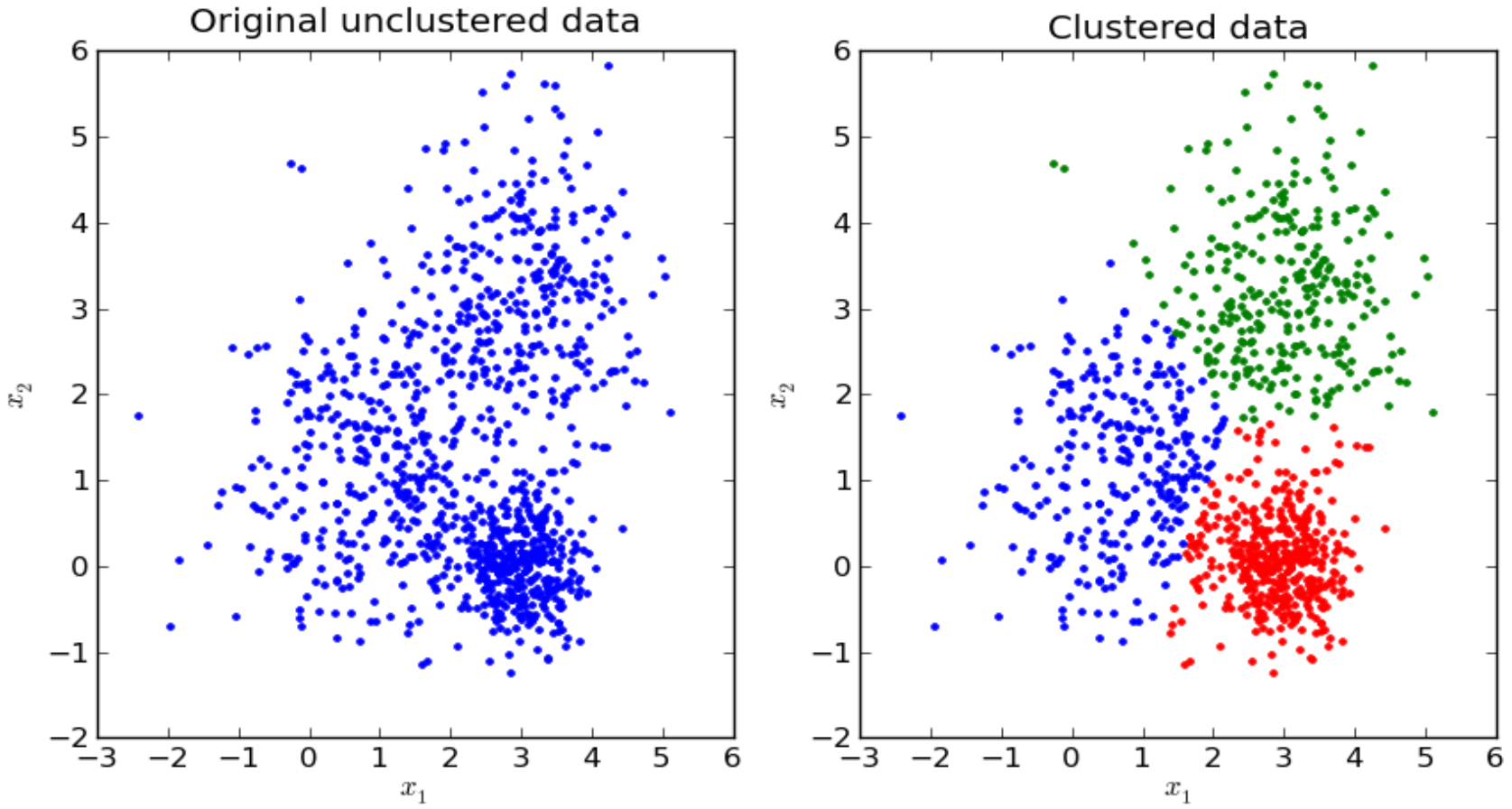
  then running all *k* algorithms in "parallel" satisfies ε-differential privacy with $\varepsilon = \max\{\varepsilon_1,...,\varepsilon_k\}$

Postprocessing

- If M_1 is an ε-differentially private algorithm that accesses a private database D,

  then outputting $M_2(M_1(D))$ also satisfies ε-differential privacy.

Case Study: K-means Clustering



Kmeans

- Partition a set of points $x_1, x_2, \ldots, x_n$ into k clusters $S_1, S_2, \ldots, S_k$ such that the following is minimized:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|_2^2$$

Mean of the cluster S_i

Kmeans

Algorithm:

- Initialize a set of k centers
- Repeat
  Assign each point to its nearest center
  Recompute the set of centers
  Until convergence …

- Output final set of k centers

[BDMN 05]

Differentially Private Kmeans

- Suppose we fix the number of iterations to T
- In each iteration (given a set of centers):
  1. Assign the points to the new center to form clusters
  2. Noisily compute the size of each cluster
  3. Compute noisy sums of points in each cluster

Differentially Private Kmeans

- Suppose we fix the number of iterations to T

Each iteration uses ε/T privacy budget, total privacy loss is ε

- In each iteration (given a set of centers):

1. Assign the points to the new center to form clusters
2. Noisily compute the size of each cluster
3. Compute noisy sums of points in each cluster

Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

- In each iteration (given a set of centers):

  1. Assign the points to the new center to form clusters

  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster

Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

- In each iteration (given a set of centers):

1. Assign the points to the new center to form clusters — NO
2. Noisily compute the size of each cluster — YES
3. Compute noisy sums of points in each cluster — YES

Differentially Private Kmeans

What is the sensitivity?

- In each iteration (given a set of centers):

  1. Assign the points to the new center to form clusters

  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster

1

Domain size

Differentially Private Kmeans

- Suppose we fix the number of iterations to T

Each iteration uses ε/T privacy budget, total privacy loss is ε

- In each iteration (given a set of centers):

1. Assign the points to the new center to form clusters
2. Noisily compute the size of each cluster

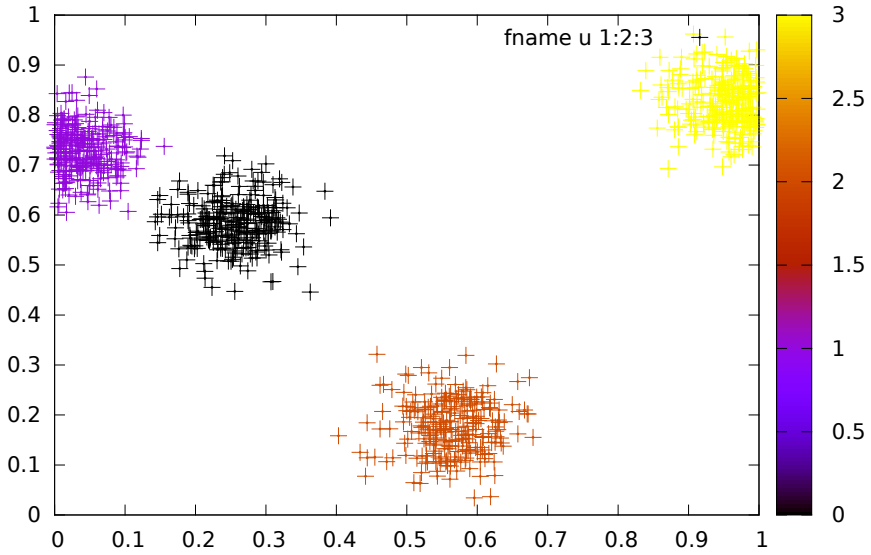
Laplace(2T/ε)

3. Compute noisy sums of points in each cluster

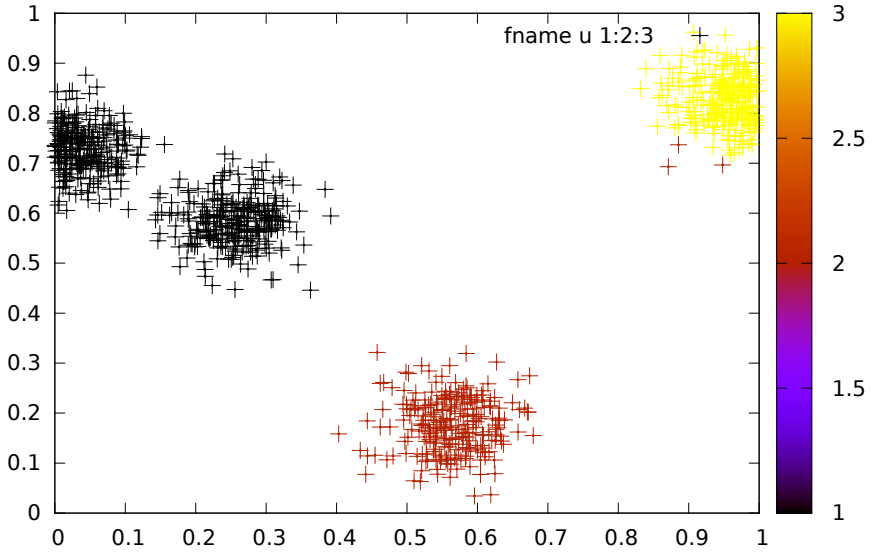
Laplace(2T |dom|/ε)

Results (T = 10 iterations, random initialization)

Original Kmeans algorithm



Laplace Kmeans algorithm



- Even though we noisily compute centers, Laplace kmeans can distinguish clusters that are far apart.

- Since we add noise to the sums with sensitivity proportional to |dom|, Laplace k-means can't distinguish small clusters that are close by.

Summary

- Differentially private algorithms ensure an attacker can't infer the presence or absence of a single record in the input based on any output.

- Building blocks
  - Laplace, exponential mechanism and randomized response
- Composition rules help build complex algorithms using building blocks

MODULE 3: ANSWERING QUERIES ON TABULAR DATA

Module 3: Answering queries on Tabular data

- Answering query workloads on tabular databases
- Theory: two seminal results
- Survey of algorithm design ideas
  - Low dimensional range queries
  - Queries on high dimensional data
- Open Questions

Problem Formulation

- **Input**:
  - Private database *D* consisting of a single table (each tuple represents data of single individual)
  - Workload *W* of counting queries* with arbitrary predicates

```
SELECT COUNT(*) FROM D WHERE <P>;
```

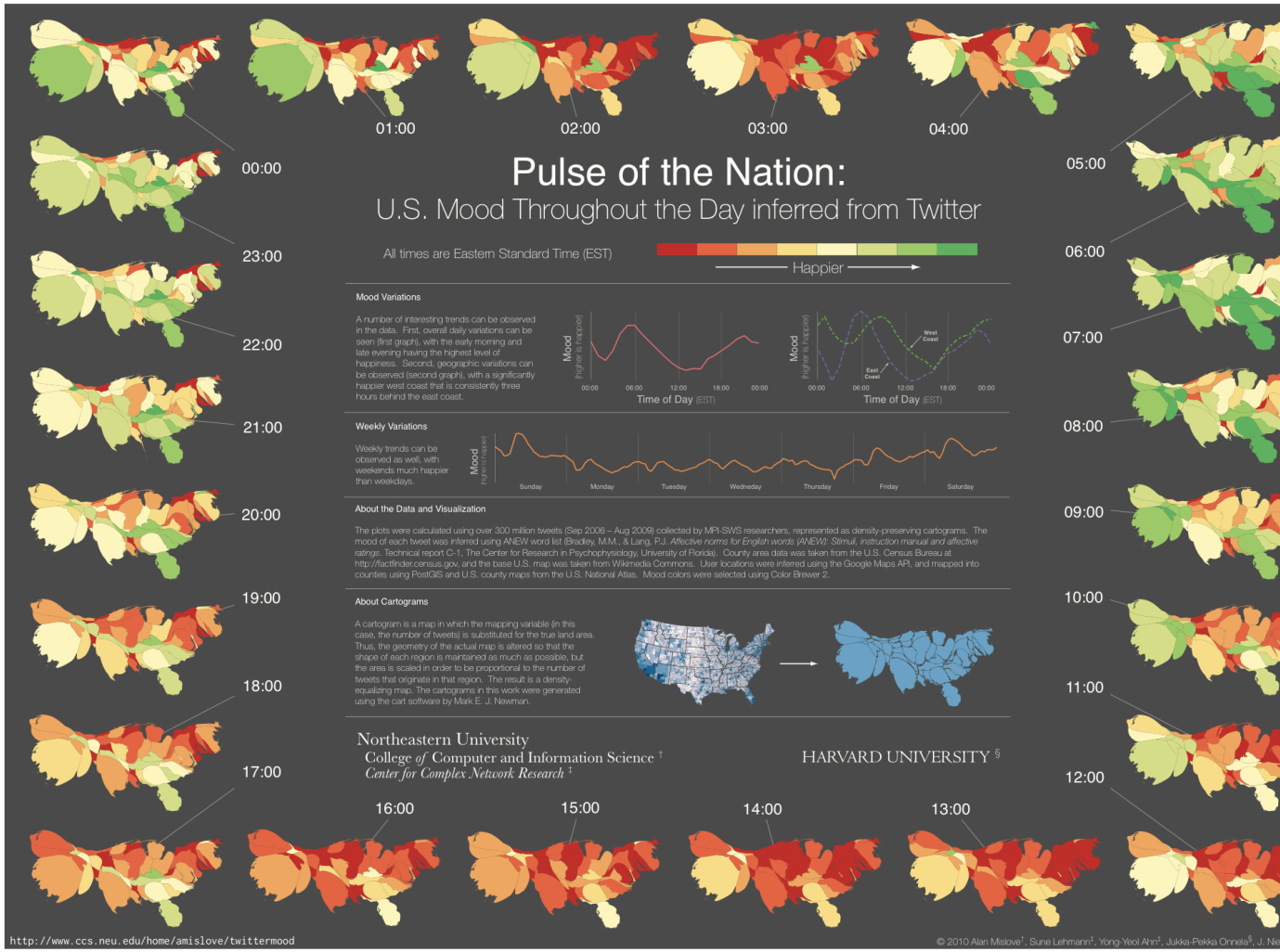
- **Output**: (noisy) answers to *W*
- Requirement: query answering algorithm satisfies differential privacy

* Many techniques can also support *linear queries*: `SELECT SUM(f(t)) FROM D` where user-defined f maps tuple to [0,1]

Analysis of temporal & spatial patterns

Counting query:

```
SELECT COUNT(*)
FROM Tweets
WHERE
moodScale=k
AND t <= time
AND time < t+1
AND UScounty = C
```



http://www.ccs.neu.edu/home/amislove/twittermood/

Statistical agencies: data publishing

- U.S. Census Bureau publishes statistics that can typically be derived from marginals
- A **marginal** over attributes $A_1, \ldots, A_k$ reports count for each combination of attribute values.
  - aka cube, contingency table
  - E.g. 2-way marginal on *EmploymentStatus* and *Gender*
- ***Thousands*** of marginals released

U.S. Census Bureau

AMERICAN FactFinder

DP03 | SELECTED ECONOMIC CHARACTERISTICS

2010-2014 American Community Survey 5-Year Estimates

Subject	ZCTA5 13346			
	Estimate	Margin of Error	Percent	Percent Margin of Error
EMPLOYMENT STATUS				
Population 16 years and over	5,676	+/-301	5,676	(X)
In labor force	2,715	+/-223	47.8%	+/-3.7
Civilian labor force	2,715	+/-223	47.8%	+/-3.7
Employed	2,529	+/-228	44.6%	+/-3.6
Unemployed	186	+/-93	3.3%	+/-1.7
Armed Forces	0	+/-16	0.0%	+/-0.5
Not in labor force	2,961	+/-288	52.2%	+/-3.7
Civilian labor force	2,715	+/-223	2,715	(X)
Percent Unemployed	(X)	(X)	6.9%	+/-3.4
Females 16 years and over	2,921	+/-216	2,921	(X)
In labor force	1,312	+/-140	44.9%	+/-4.5
Civilian labor force	1,312	+/-140	44.9%	+/-4.5
Employed	1,245	+/-135	42.6%	+/-4.3
Own children under 6 years	325	+/-117	325	(X)
All parents in family in labor force	241	+/-99	74.2%	+/-17.3
Own children 6 to 17 years	476	+/-102	476	(X)
All parents in family in labor force	389	+/-95	81.7%	+/-8.5
COMMUTING TO WORK				
Workers 16 years and over	2,449	+/-217	2,449	(X)
Car, truck, or van -- drove alone	1,518	+/-176	62.0%	+/-5.2
Car, truck, or van -- carpooled	116	+/-57	4.7%	+/-2.3
Public transportation (excluding taxicab)	17	+/-19	0.7%	+/-0.8
Walked	531	+/-116	21.7%	+/-4.3
Other means	132	+/-58	5.4%	+/-2.4
Worked at home	135	+/-64	5.5%	+/-2.5
Mean travel time to			X)	(X)
OCCUPATION				
Civilian employed population 16 years and over	2,529	+/-228	2,529	(X)

https://factfinder.census.gov/

[WWLTRD09]

Genome Wide Association Studies

- **Goal**: to study genetic factors associated with a given disease
- Collect subsets of the population with diseases (*case*) and without (*control*)
- Extract SNPs (specific DNA subsequences)
  - For each SNP, usually find 2 *alleles* (alternative forms of the gene)

- **Counting queries**: Compute allele frequencies in both the case and control groups (marginal over SNPxDisease)

	Disease	
SNP	Control	Case
0 (Other allele)	C_{00}	C_{01}
1 (Risk allele)	C_{10}	C_{11}

- Perform association test using these frequencies (e.g., Chi Square Test) to identify SNPs highly associated with disease

Problem variant: offline vs. online

- **Offline** (batch):
  - Entire W given as input, answers computed in **batch**
- **Online** (adaptive):
  - W is sequence $q_1, q_2, \ldots$ that arrives online
  - **Adaptive**: analyst's choice for q_i can depend on answers $a_1, \ldots, a_{i-1}$
- Answering linear queries online is strictly harder than answering them offline [BSU16].

Important aspects of problem: Data and query complexity

- Data complexity
  - Dimensionality: number of attributes
  - Domain size: number of distinct attribute combinations
  - Many techniques specialized for *low dimensional data*
- Query complexity
  - Many techniques designed to work well for a specific *class* of queries
  - Classes (in rough order of difficulty): histograms, range queries, marginals, counting queries, linear queries

Solution variants: query answers vs. synthetic data

Two high-level approaches to solving problem

1. **Direct**:
   - Output of the algorithm is list of query answers
2. **Synthetic data**:
   - Algorithm constructs a *synthetic dataset D'*, which can be queried directly by analyst
   - Analyst can pose additional queries on *D'* (though answers may not be accurate)

Theory

- Given negative result of Dinur-Nissim, is there any hope?
- Yes!
  - The key to Dinur-Nissim is that query answers have ***independent*** noise (which can cancel out)
  - To answer more queries, query error must be ***correlated***
- Examples of correlation
  - Use some query answers to approximate others
  - Construct a synthetic database that is approximately accurate for queries of interest

[BLR08]

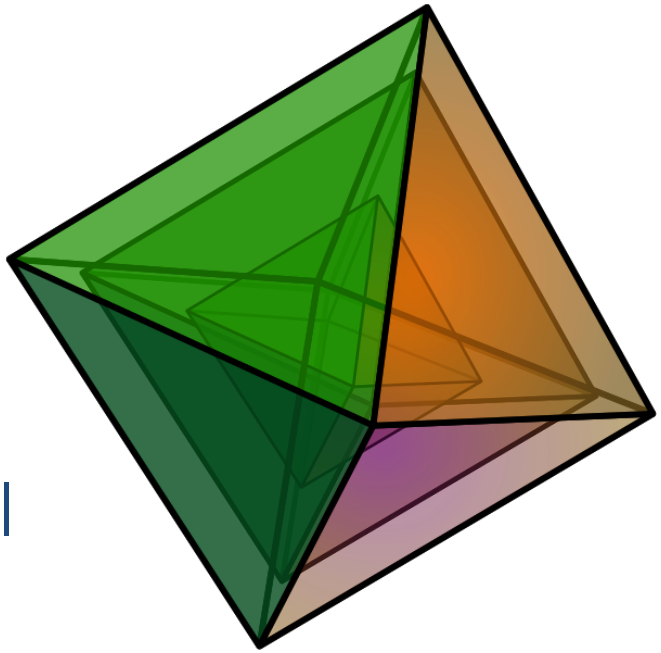
Answering Exponentially Many Queries Offline

- Key technical insight: For any set of count queries W, there exists a *small* database D' *consistent* with D on every query in W.
  - Small: $O(\log(|W|)/\alpha^2)$
  - Consistent: error for any q in W is at most α
- Result follows from learning theory:
  - Estimates on small random sample will generalize to population

[BLR08]

Answering Exponentially Many Queries Offline

- Input: W, ε
- Output: D'
- The Mechanism:
  - T = {all small databases D'}
  - $f(D, D') = -max_{q \in W} |q(D) - q(D')|$
  - Output $D' \in T$ using Exponential Mechanism applied to f



- Theorem: Is ε-private and w.h.p. error α is at most

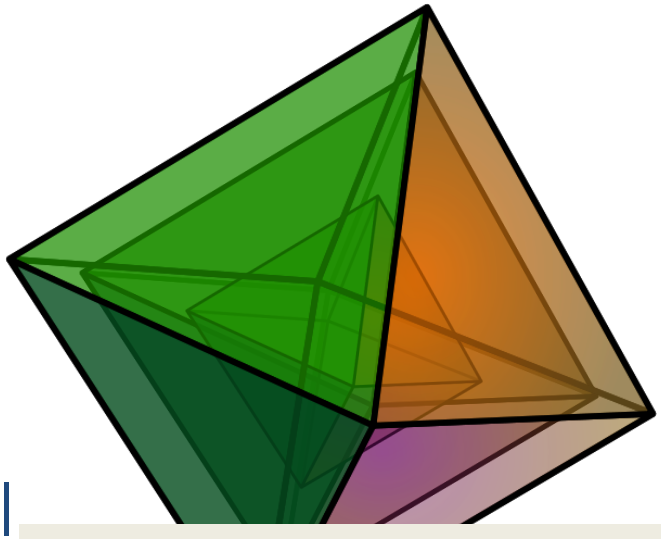
$$O\left(\frac{\log|\text{domain}|\log|W|}{\varepsilon|D|}\right)^{1/3}$$

[BLR08]

Answering Exponentially Many Queries Offline

- Input: W, ε
- Output: D'
- The Mechanism:
  - T = {all small databases D'}
  - $f(D, D') = -max_{q \in W} |q(D) - q(D')|$
  - Output $D' \in T$ using Exponential Mechanism applied to f
- Theorem: Is ε-private and w.h.p.

$$O\left(\frac{\log|\text{domain}|\log|}{\varepsilon|D|}\right.$$



Limitations

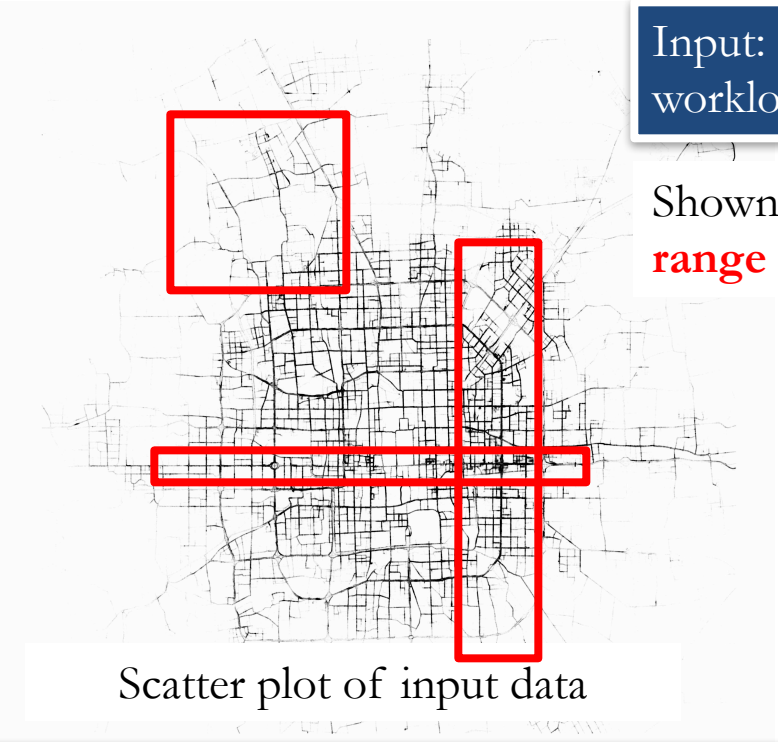
- Impractical: runtime exponential
- Offline (for online see [HR10])

Case study: range queries over spatial data

Input: sensitive data *D*

	Latitude	Longitude
1		
2	39.98105	116.30142
3	39.9424	116.30587
4	39.93691	116.33438
5	39.94354	116.3532
6	...	...

BeijingTaxi dataset[1]: 4,268,780 records of (lat,lon) pairs of taxi pickup locations in Beijing, China in 1 month.



Scatter plot of input data

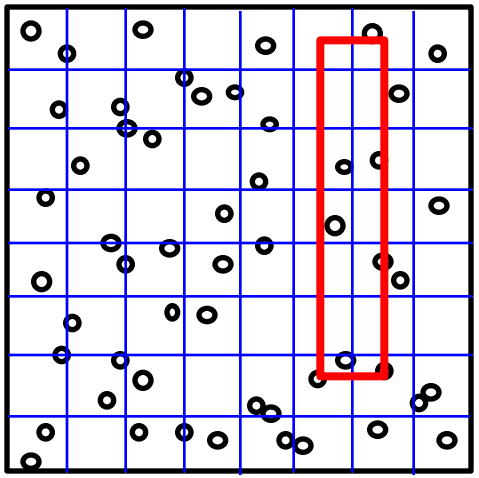
Input: range query workload *W*

Shown is workload of **3 range queries**

Task: compute answers to workload *W* over private input *D*

[1]Raw data from: Taxi trajectory open dataset, Tsinghua university, China. http://sensor.ee.tsinghua.edu.cn, 2009.

Baseline algorithm



Scatter plot of input data

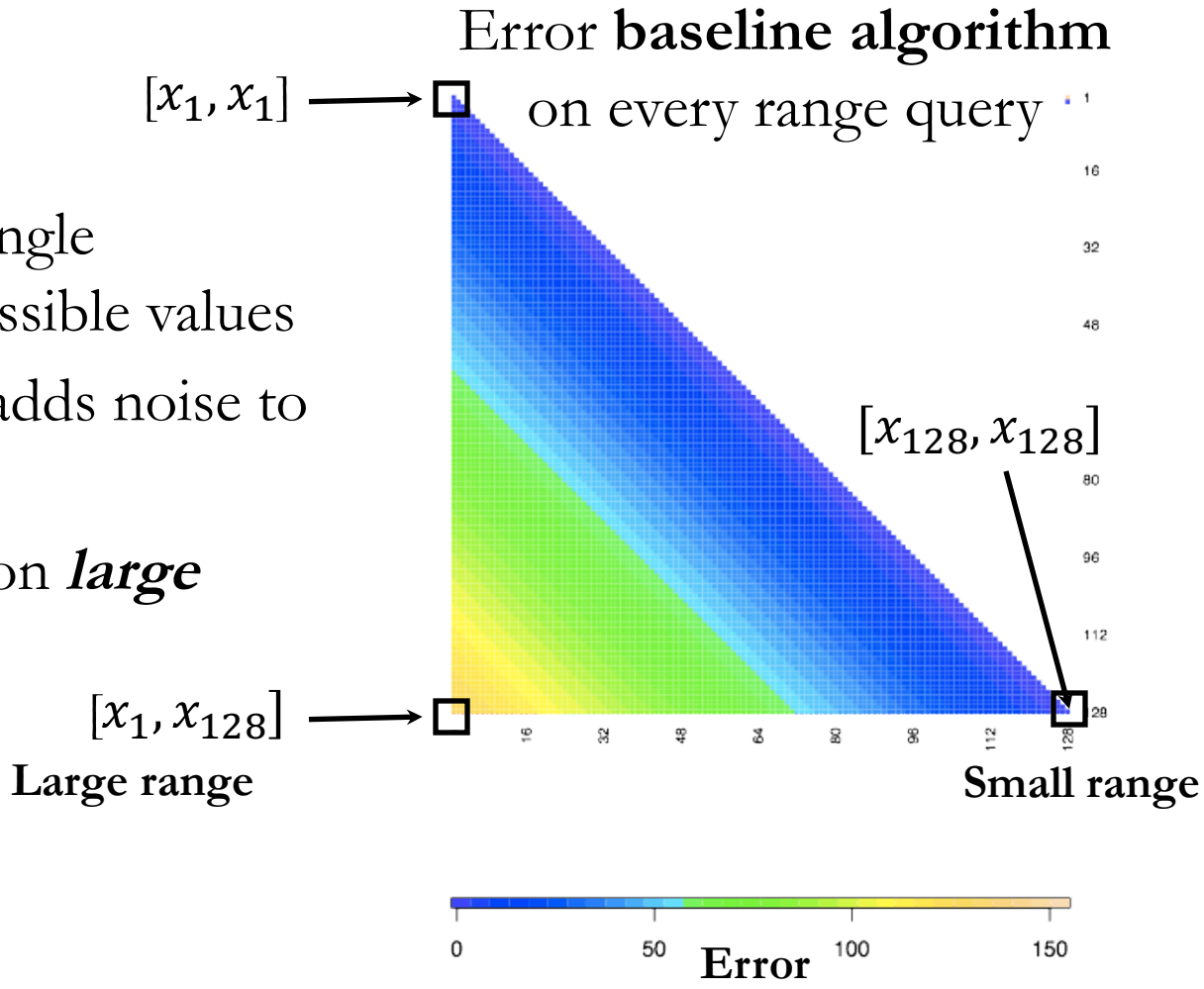
Limitations

- Granularity of discretization
  - Coarse: detail lost
  - Fine: noise overwhelms signal
- Noise accumulates: squared error grows *linearly* with range

1. Discretize attribute domain into cells
2. Add noise to cell counts (Laplace mechanism)
3. Use noisy counts to either…
   1. Answer queries directly (assume distribution is uniform within cell)
   2. Generate synthetic data (derive distribution from counts and sample)

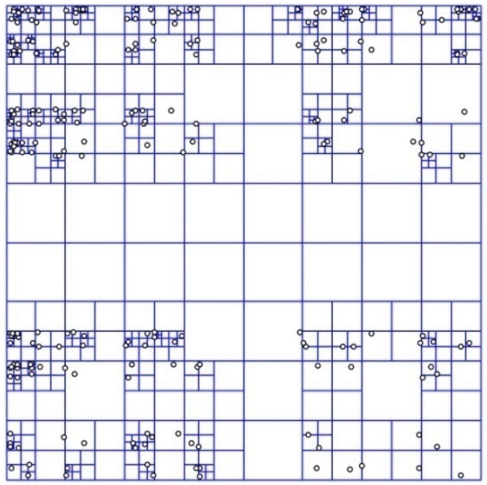
Error analysis

- Dataset is 1D: single attribute, 128 possible values
- Baseline simply adds noise to every cell count
- Incurs *high error* on ***large ranges***



[CPSSY12]

Improved algorithm: private quad-tree

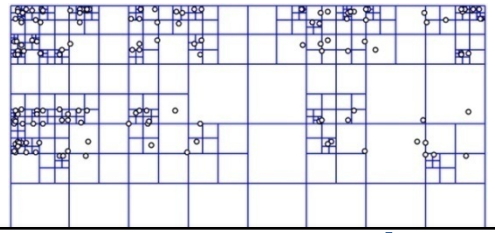


quadtree

- **Input**: maximum height h, minimum leaf size L, data set D
- Recurse on nodes of tree:
  - Add $Lap(1/\varepsilon)$ noise to node count
  - Split node domain into quadrants
  - Create child nodes
- Stop when:
  - Noisy count of node $\leq L$
  - Max height h is reached
- **Intuition**:
  - Early stopping controls granularity of discretization
  - To answer long range queries, leverage hierarchy of noisy counts

[CPSSY12]

Improved algorithm: private quad-tree



Exercise: Let $h' \leq h$ be height of resulting tree. Algorithm satisfies ε'-differential privacy for ε' equal to

1. $\varepsilon h'$
2. εh
3. $4\varepsilon h$
4. $\varepsilon 4^h$

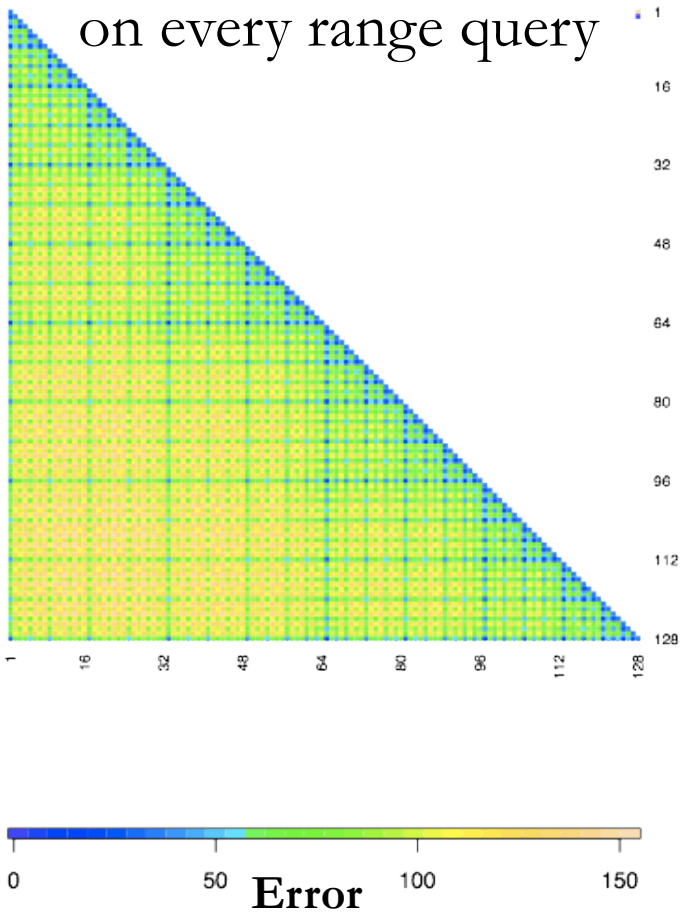
- **Input**: maximum height $\boldsymbol{h}$, minimum leaf size $\boldsymbol{L}$, data set D
- Recurse on nodes of tree:
  - Add $Lap(1/\varepsilon)$ noise to node count
  - Split node domain into quadrants
  - Create child nodes
- Stop when:
  - Noisy count of node $\leq \boldsymbol{L}$
  - Max height $\boldsymbol{h}$ is reached
- **Intuition**:
  - Early stopping controls granularity of discretization
  - To answer long range queries, leverage hierarchy of noisy counts

Using tree structure for range queries

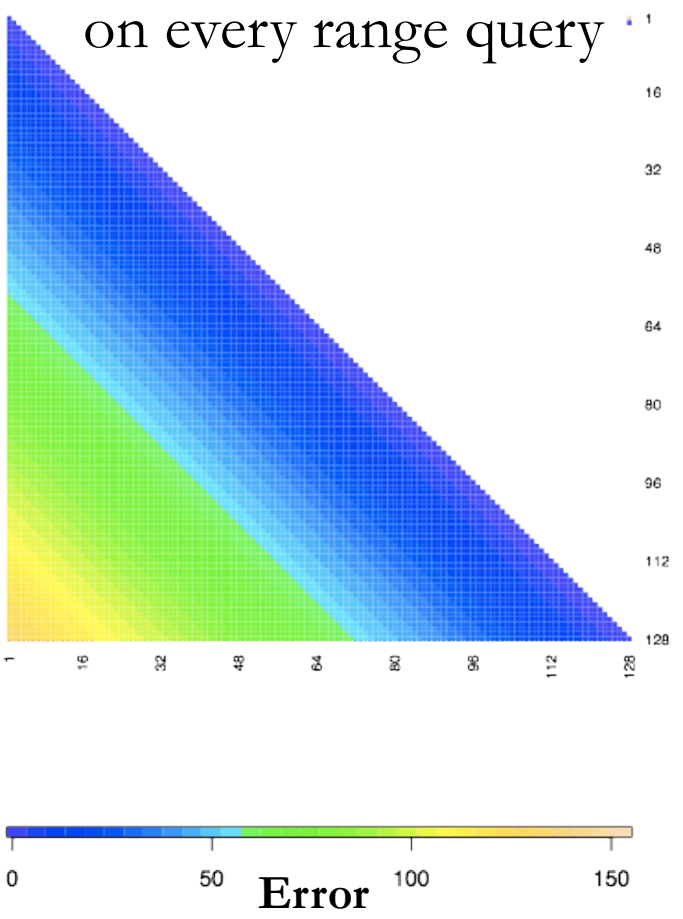
- Quad-tree, each node has noisy count. How use to answer range queries?
- Idea 1: given range query q, find smallest set of noisy counts from tree that "cover" q
- Idea 2 (better):
  - Observation: node's noisy count and *sum* of children's noisy counts are two estimates of same quantity
  - Combine estimates using **statistical inference**

Error comparison

Error **tree algorithm** on every range query

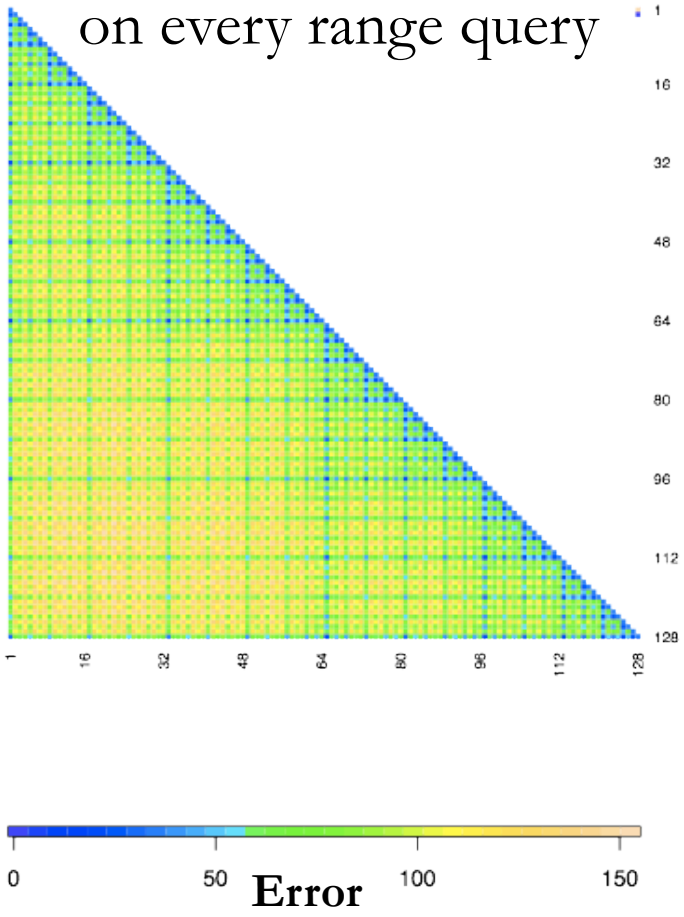


Error **baseline algorithm** on every range query

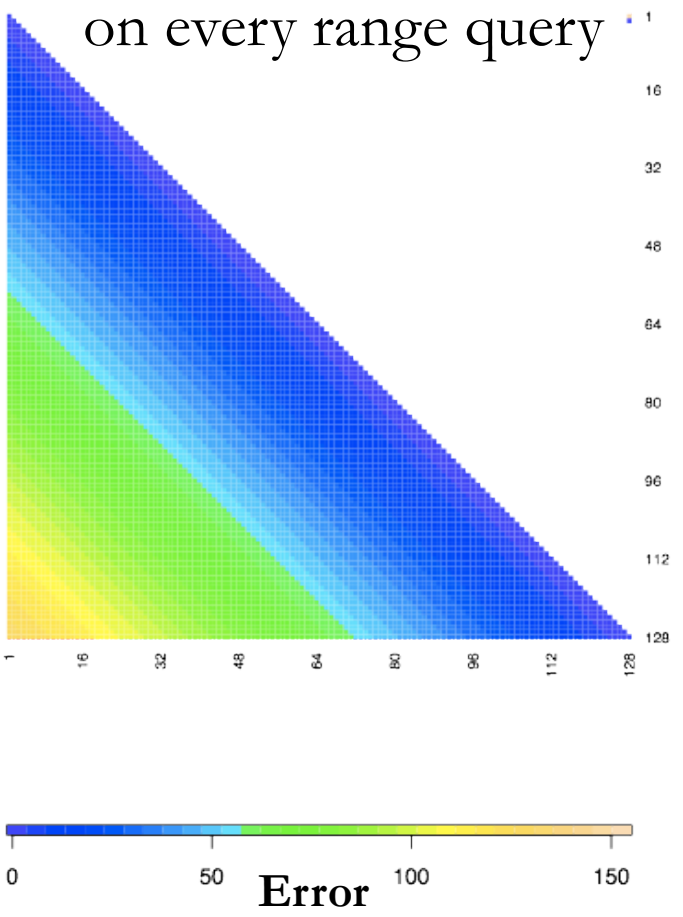


Hierarchy lowers error on *large ranges* but incurs slightly higher error for *small ranges*

Error **tree algorithm** on every range query



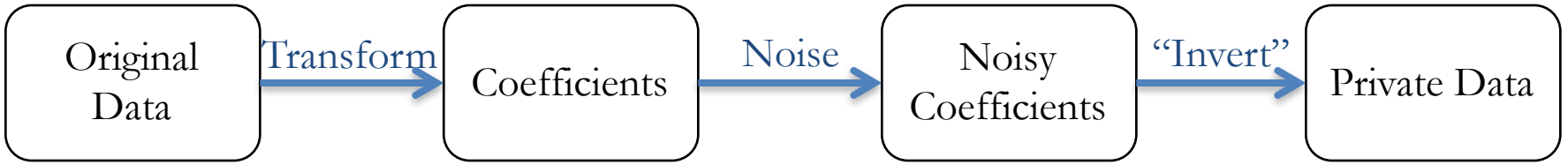
Error **baseline algorithm** on every range query



Credit: Cormode

Data transformations

- Can think of trees as transform of input



- Can apply other data transformations
- **Goal**: pick a low sensitivity transform that preserves good properties of data

Linear transformations

- Examples
  - Hierarchical: Trees [HRMS10,QYL13], full height quadtree [CPSSY12]
  - Haar Wavelet [XWG10]
  - Discrete Fourier transform [BCDKMT07]
- Inverting transformation
  - Some transformations (e.g. tree) have redundancy (over-constrained), so require pseudo-inverse
- Matrix Mechanism [LHRMM10,LM12,LM13]
  - Formalizes problem of designing a linear transform that is tailored to the queries
- Error rates are *independent* of input (assumes linear transform is "full rank")

Lossy transformations

- Variants
  - Drop "small" coefficients:
    - Quad-tree with early stopping (noisy count threshold)
    - Fourier coefficients: EFPA [ACC12], [RN10]
  - Data-adaptive discretization:
    - PrivTree [ZXX16], KD-Tree [CPSSY12], DAWA [LHMY14], [DNRR15], [QYL13], [BLR08]
  - Data-adaptive measurement:
    - MWEM [HLM12], DualQuery [GAHRW14]
  - Randomized transforms: sketches and compressed sensing
    - JL Transform [BBDS12], Compressive mechanism [LZWY11]
- "Inverting" transformation
  - Because lossy, they are under-constrained, requires estimation
- Error rates *depend on input*
  - Can be much lower (trades off small bias for lower variance)
  - Warrants careful empirical evaluation; algorithms are "**data dependent**"

High dimensional data

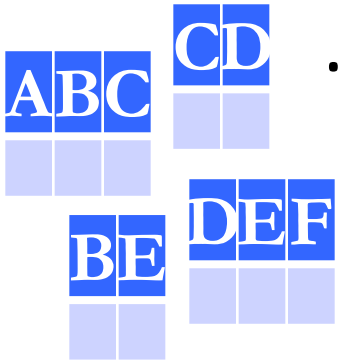
- Generally an under-studied area
- Two algorithms, both synthetic data generators
  - PrivBayes [ZCPSX14]
  - DualQuery [GAHRW14]
- Common properties
  - Limited to binary attributes
  - Designed to support low-order marginals (and other workloads well approximated by marginals, such as classification)

[ZCPSX14]

PrivBayes

High-dimensional table T

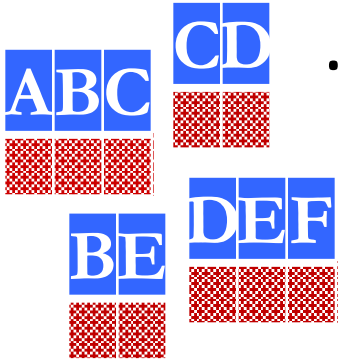
decompose



Low-dimensional tables

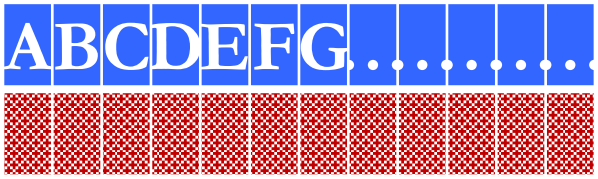
Add noise

- Method:
  - Use **Bayesian network** to learn data distribution
  - After BN learned, generate *synthetic data* by sampling from BN
- **Challenge**: *privately* choosing *good* decomposition



Noisy tables

reconstruct



Noisy table T^*

[GAHRW14]

Dual Query

- Problem of generating synthetic data formulated as a zero sum game between
  - Data player: generates synthetic records to reduce query error
  - Query player: chooses queries with high error (on current synthetic dataset)
  - Theoretical analysis of utility comes from studying equilibrium of game

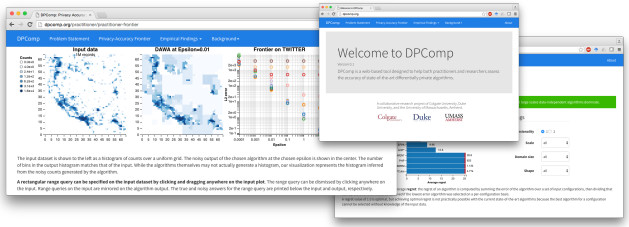
1. Let Q^t be distribution over queries (Q^1 is uniform)
2. For t = 1… T
   a) S ← Query player samples s queries from Q^t — Privacy
   b) Data player finds record x_t that maximizes total answer on S — NP-Hard
   c) Q^{t+1} ← Query player updates(x_t, D, Q^t)
3. Output synthetic database x_1, …, x_t

- What makes it practical?
  - Unlike some prior work[HLM12], avoids storing distribution over domain (exponential)
  - Approximate solution may be good enough!
  - Optimization problem can be solved with off-the-shelf solvers
- Case study on 500K 3-way marginals over 17K binary attributes, using CPLEX solver

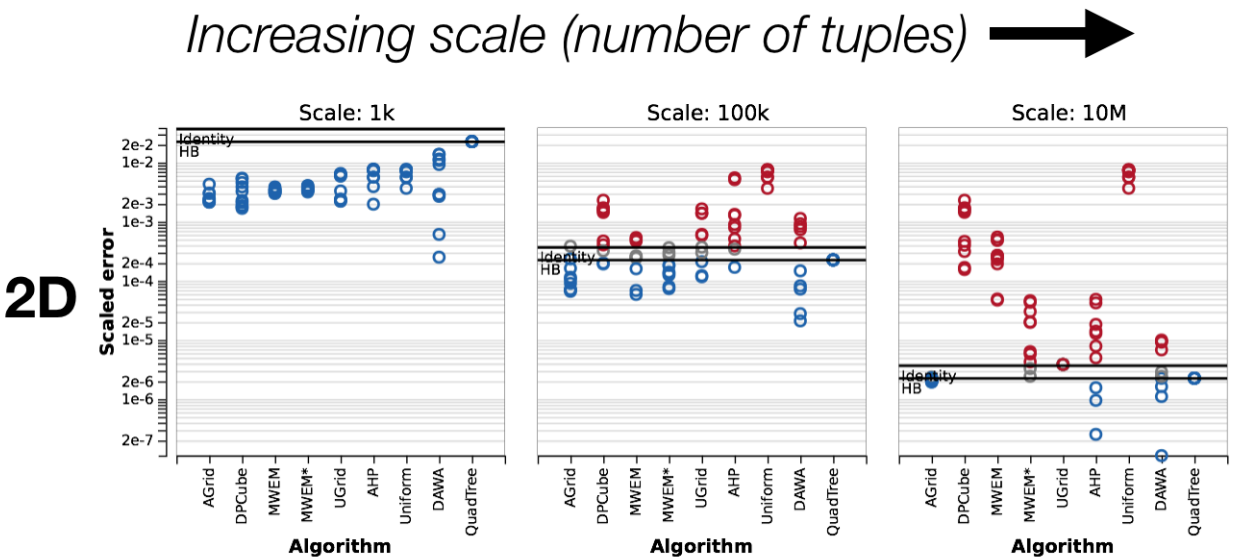
[HMMCZ16]

Empirical benchmarks

- [HMMCZ16] propose a novel evaluation framework for standardized evaluation of privacy algorithms.
- Study of algorithms for range query answering over 1 and 2D
- Benchmark website www.dpcomp.org



Some data-dependent algorithms fail to offer benefits at larger scales (no. of tuples)



Open questions

- Robust and private algorithm selection
  - Pythia (Thursday 2 PM DP Session)
- Error bounds for data-dependent algorithms
- Empirical evaluation of algorithms for high dimensional data

Differential Privacy References

CACM Articles

- [D11] Dwork, A firm foundation for private data analysis. In CACM, 2011.
- [MK15] Machanavajjhala & Kifer, Designing statistical privacy for your data. In CACM, 2015.

Book

- [DR14] Dwork & Roth, The Algorithmic Foundations of Differential Privacy. In Foundations and Trends, 2014.

Tutorials

- [C13] Graham Cormode. Building blocks of privacy: Differentially private mechanisms. In PrivDB, 2013.
- [YZMWX12] Yang et al., Differential privacy in data publication and analysis. In SIGMOD, 2012.
- [HLMPT11] Hay et al., Privacy-aware Data Management in Information Networks. In SIGMOD, 2011.
- [CS14]K. Chaudhuri and A. D. Sarwate. Differential privacy for signal processing and machine learning. In WIFS, 2014.
- [KNRS13] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In TCC, 2013.

Module 1 References

- [S02] Sweeney, "K-anonymity", IJFUKS 2010
- [DN03] Dinur, Nissim, "Revealing information while preserving privacy", PODS 2003
- [D06] Dwork, "Differential Privacy", ICALP 2006
- [MKGV06] Machanavajjhala, Kifer, Gehrke, Venkitasubramaniam, "L-Diversity" ICDE 2006
- [GKS08] Ganta, Kasiviswanathan, Smith, "Composition attacks and auxiliary information in data privacy", KDD 2008
- [KL10] Kifer, Lin, "Towards an Axiomatization of Statistical Privacy and Utility.", PODS 2010
- [VSJO13] Vaidya, Shafiq, Jiang, Ohno-Machado, "Identifying inference attacks against healthcare data repositories", AMIA 2013
- [MK15] Machanavajjhala, Kifer, "Designing statistical privacy for your data", CACM 2015

Module 2 References

- [W65] Warner, "Randomized Response" JASA 1965
- [DN03] Dinur, Nissim, "Revealing information while preserving privacy", PODS 2003
- [BDMN05] Blum, Dwork, McSherry, Nissim, "Practical privacy: the SuLQ framework", PODS 2005
- [D06] Dwork, "Differential Privacy", ICALP 2006
- [DMNS06] Dwork, McSherry, Nissim, Smith, "Calibrating noise to sensitivity in private data analysis", TCC 2006
- [MT07] McSherry, Talwar, "Mechanism Design via Differential Privacy", FOCS 2007

Module 3 References

- [ACC12] Ács et al. Differentially private histogram publishing through lossy compression. In *ICDM*, 2012.
- [BBDS12] Blocki et al. The johnson-lindenstrauss transform itself preserves differential privacy. In FOCS, 2012.
- [BCDKMT07] Barak et al. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In PODS, 2007.
- [BLR08] Blum et al. A learning theory approach to noninteractive database privacy. In STOC, 2008.
- [DNRR15] Dwork et al. Pure Differential Privacy for Rectangle Queries via Private Partitions. In ASIACRYPT, 2015.
- [CPSSY12] Cormode et al. Differentially Private Spatial Decompositions. In ICDE, 2012.
- [GAHRW14] Gaboardi et al. Dual Query: Practical Private Query Release for High Dimensional Data. In ICML, 2014.
- [HLM12] Hardt et al. A simple and practical algorithm for differentially private data release. In NIPS, 2012.
- [HMMCZ16] Hay et al. Principled Evaluation of Differentially Private Algorithms using DPBench. In SIGMOD, 2016.
- [HRMS10] Hay et al. Boosting the accuracy of differentially private histograms through consistency. In PVLDB, 2010.
- [LHMY14] Li et al. A data- and workload-aware algorithm for range queries under differential privacy. In PVLDB, 2014.
- [LHRMM10] Li et al. Optimizing linear counting queries under differential privacy. In PODS, 2010.
- [LM12] Li et al. An adaptive mechanism for accurate query answering under differential privacy. In PVLDB, 2012.
- [LM13] Li et al. Optimal error of query sets under the differentially-private matrix mechanism. In ICDT, 2013.
- [LZWY11] Li et al. Compressive mechanism: utilizing sparse representation in differential privacy. In WPES, 2011.
- [QYL13] Qardaji et al. Understanding hierarchical methods for differentially private histograms. In PVLDB, 2013.
- [QYL13] Qardaji et al. Differentially private grids for geospatial data. In ICDE, 2013.
- [RN10] Rastogi et al. Differentially private aggregation of distributed time-series with transformation and encryption. In *SIGMOD*, 2010.
- [WWLTRD09] Wang et al. Privacy-preserving genomic computation through program specialization. In CCS, 2009.
- [XWG10] Xiao et al. Differential privacy via wavelet transforms. In ICDE, 2011.
- [ZCPSX14] Zhang et al. PrivBayes: private data release via bayesian networks. In SIGMOD, 2014.
- [ZXX16] Zhang et al. PrivTree: A Differentially Private Algorithm for Hierarchical Decompositions. In SIGMOD, 2016.