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?"

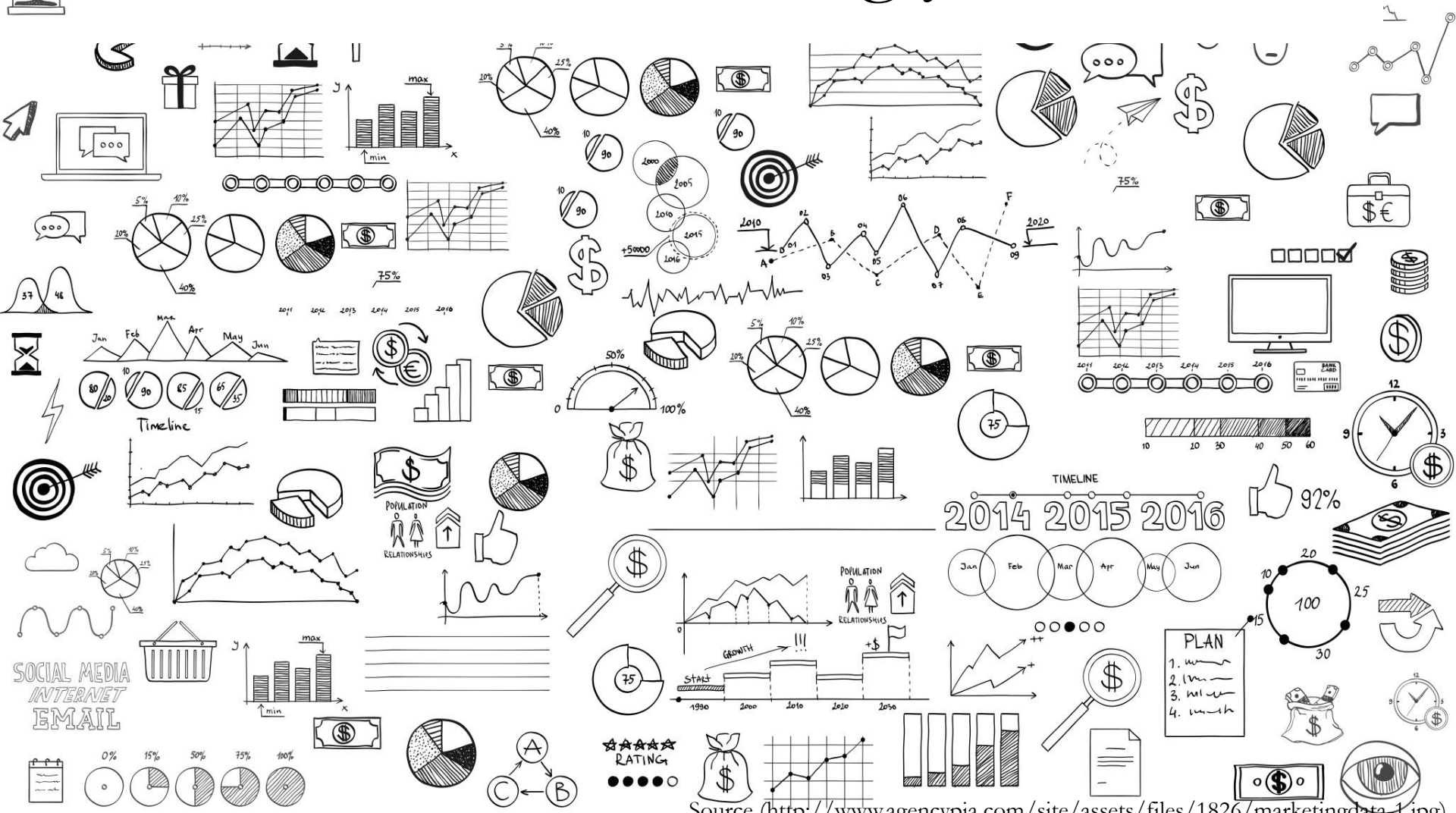
Xi He
Ph.D. Candidate, Duke University

"Can privacy algorithms work in real world systems?"

Michael Hay
Assistant Professor, Colgate University

"Can algorithms be provably private and useful?"

Our world is increasingly data driven



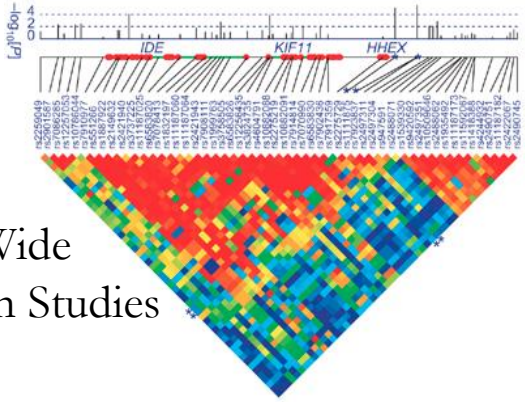
Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)

Aggregated Personal Data is **invaluable**

Advertising

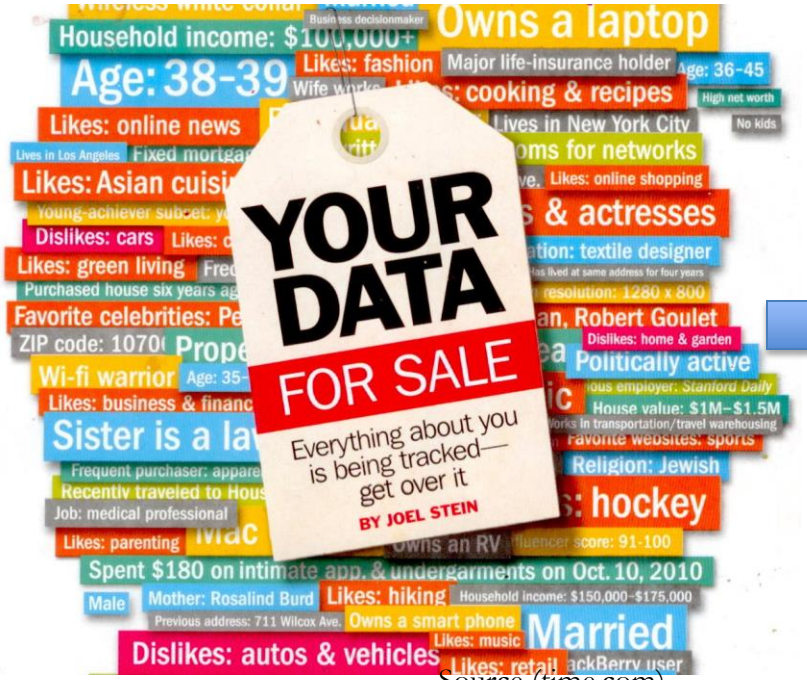
Source (esri.com)

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

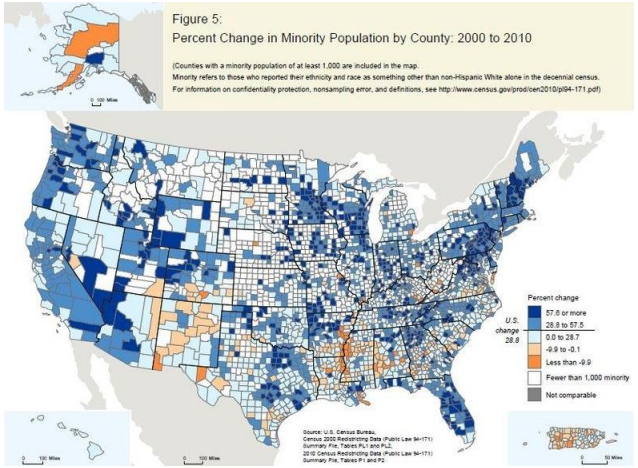
Physical/Financial Harm

Aggregated Personal Data …

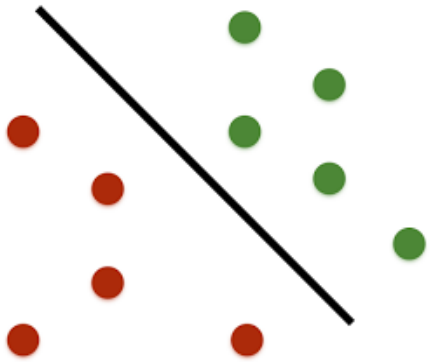
… is made publicly available in many forms.

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

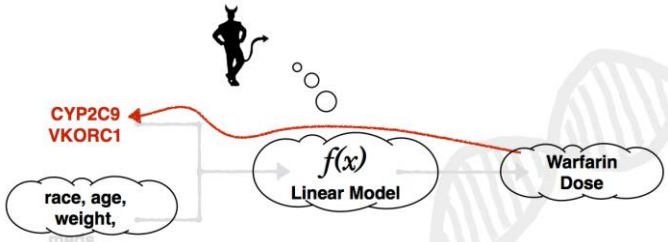
"Anonymous" Genomes Identified

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. Privacy Problem Statement
2. Differential Privacy
3. Algorithms for Tabular Data

   Break

4. Applying Differential Privacy
5. Privacy beyond tabular Data
6. Applications II

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: Applying Differential Privacy

- Real world deployments of differential privacy
  - OnTheMap 

  - RAPPOR 

- Attacks on differential privacy implementations
  - Side channel attacks
  - Floating point attacks

NEEDS TO CHANGE

Module 5: Privacy beyond tabular data

- Differential Privacy for complex data
  - Neighboring databases
  - Correlations
  - No Free Lunch Theorem

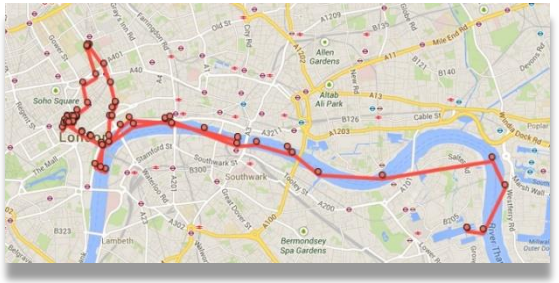
- Customizing differential privacy using Pufferfish
  - Semantic privacy definitions
  - Equivalence to DP
  - Algorithm Design

Module 6: Applications II

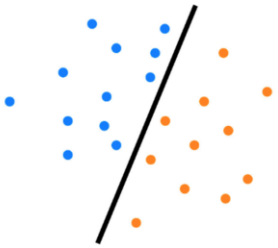
- Customized Privacy for Non-tabular Data

Social network

Location stream



- Differential Privacy for Machine Learning



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

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.

MODULE 1: PROBLEM FORMULATION

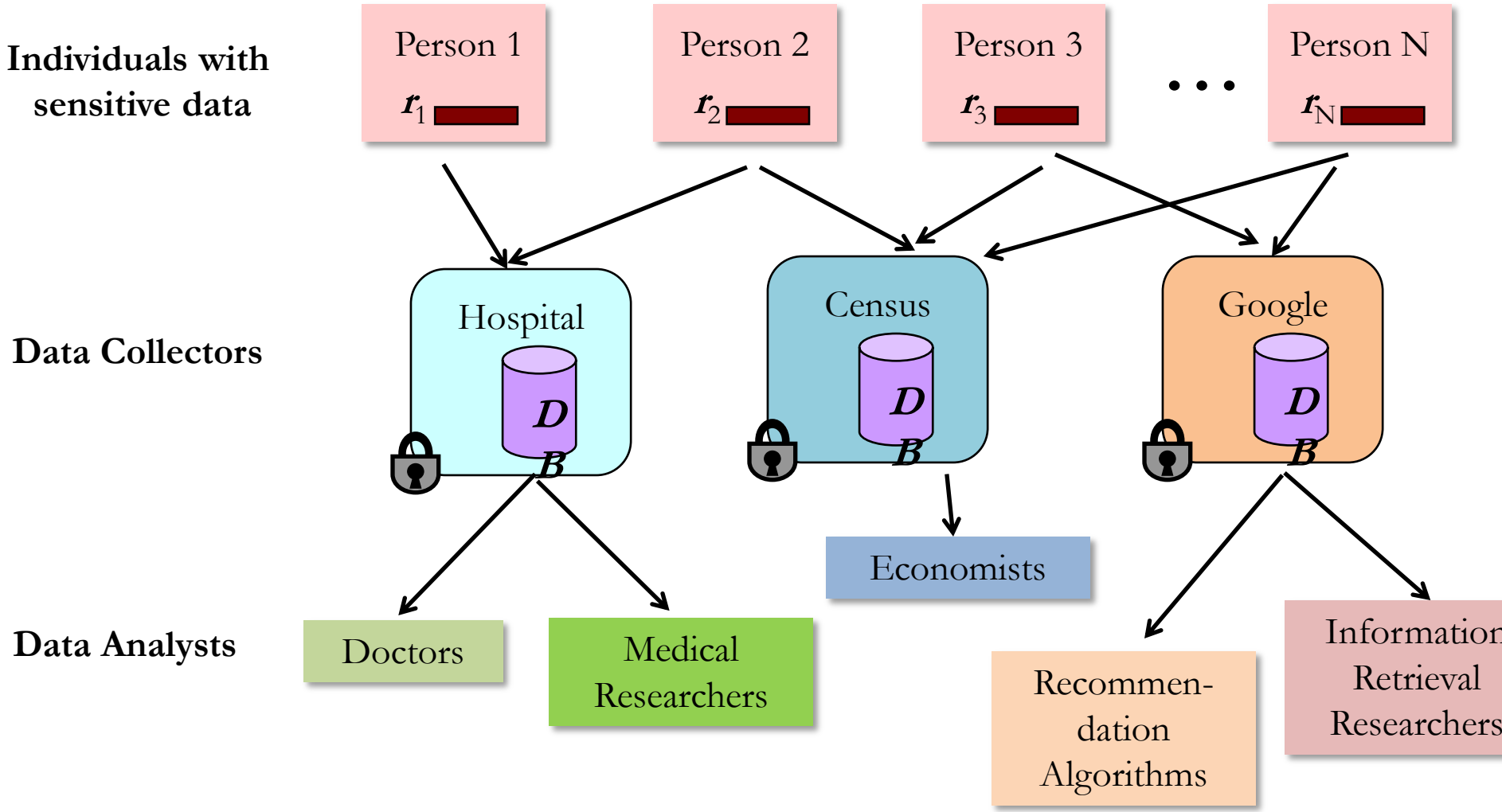
Module 1: What is Privacy?

- Privacy Problem Statement

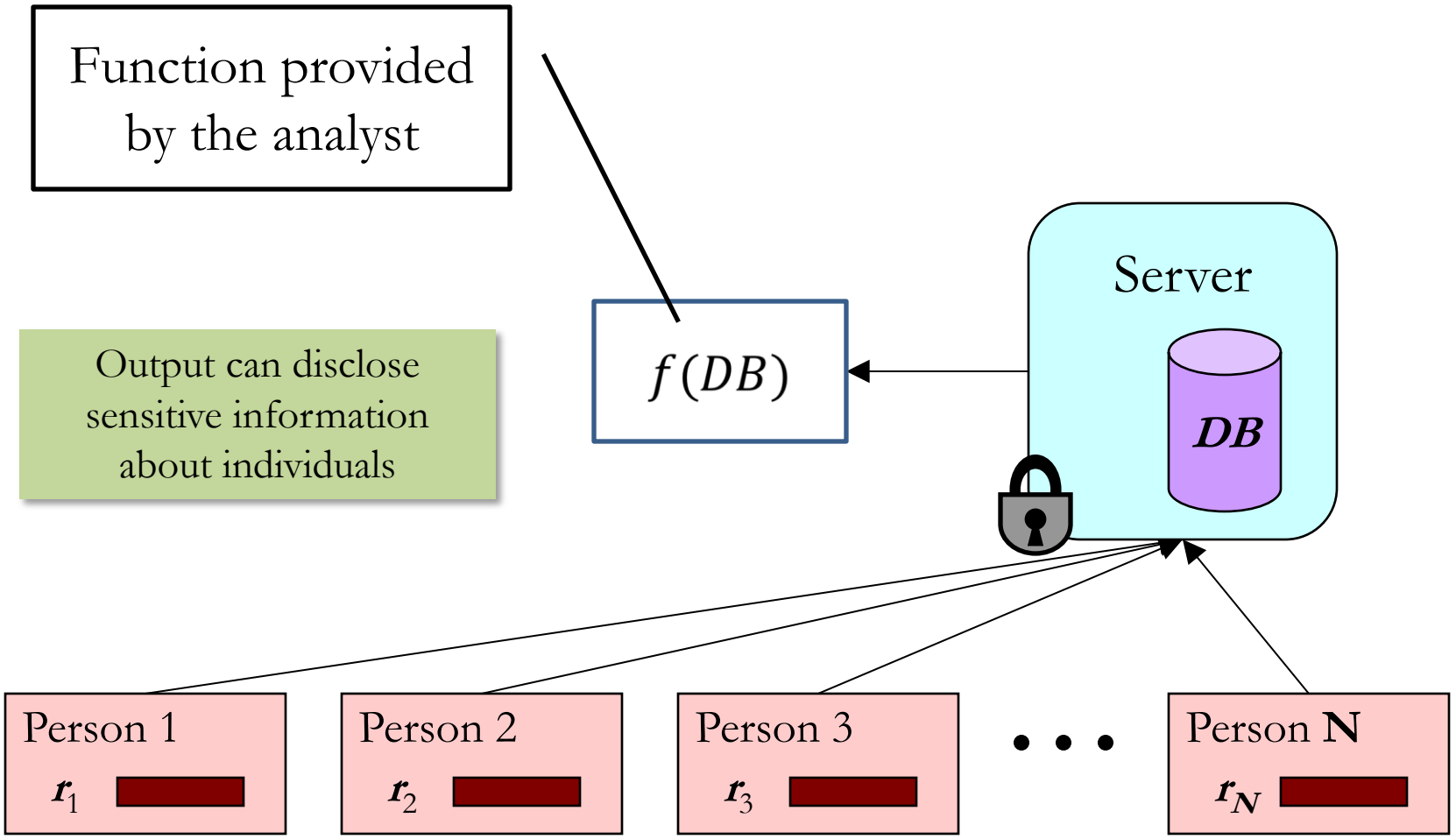
- What privacy is *not* …
  - Encryption
  - Anonymization
  - Restricted Query Answering

- What *is* privacy?

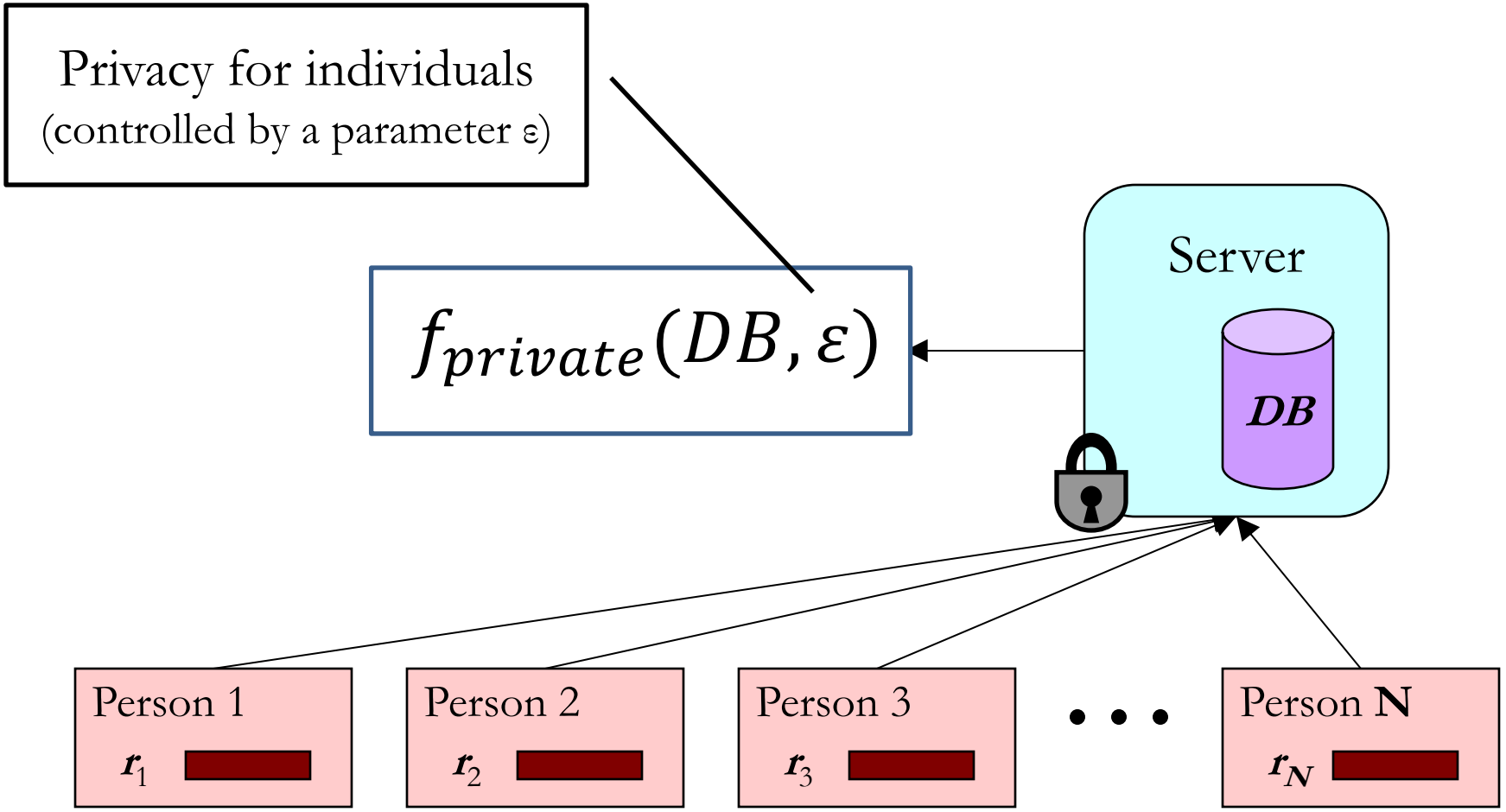
Statistical Databases



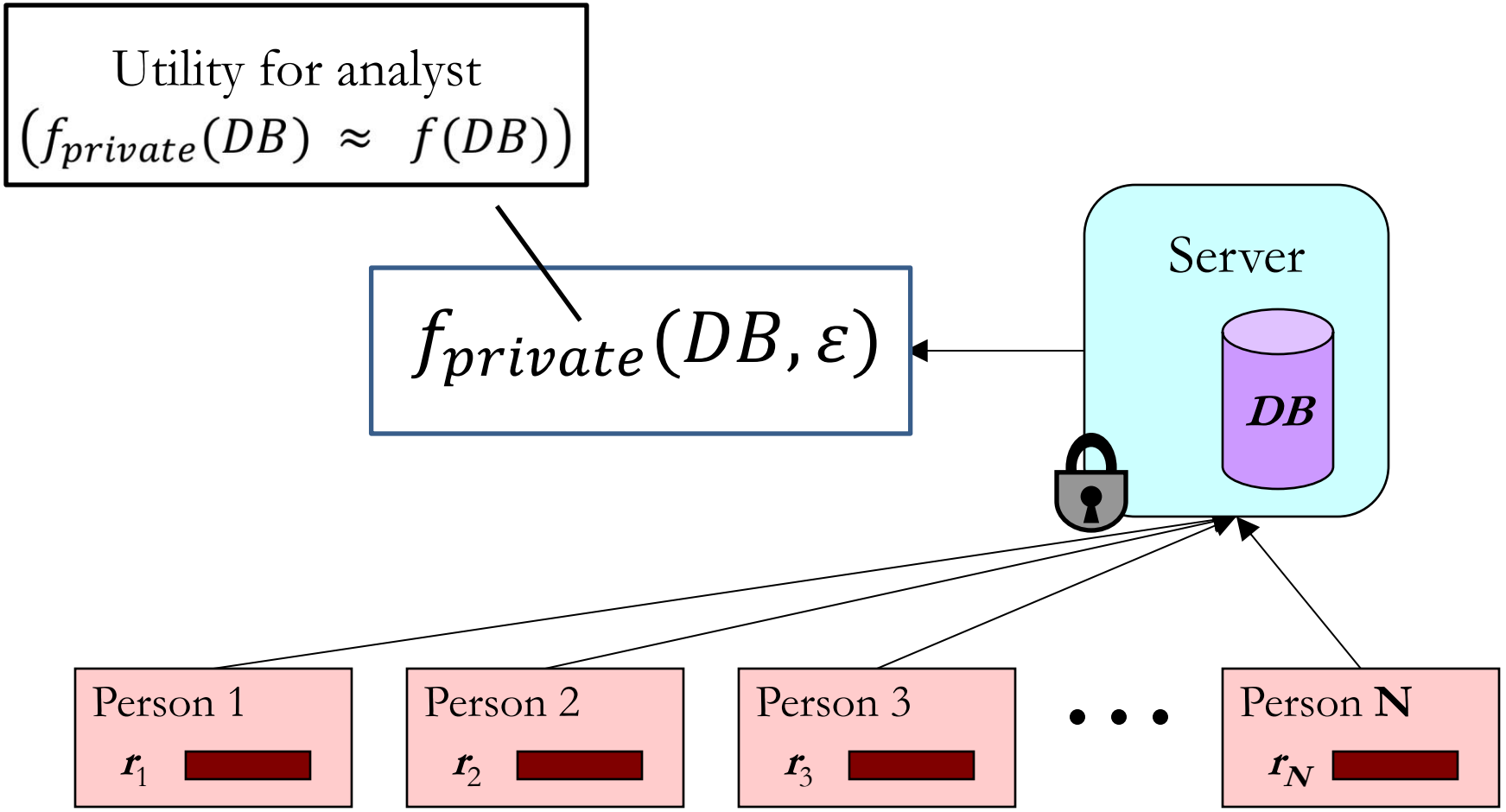
Statistical Database Privacy



Statistical Database Privacy

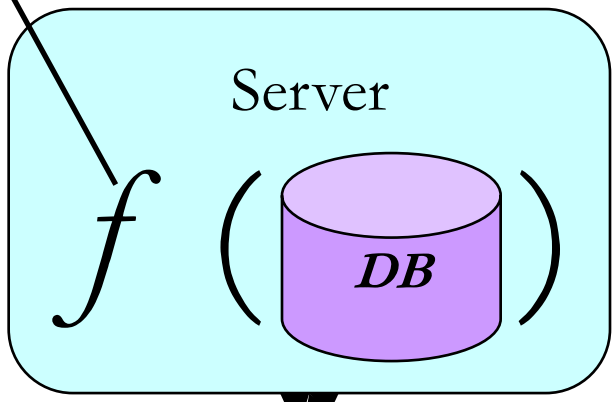


Statistical Database Privacy

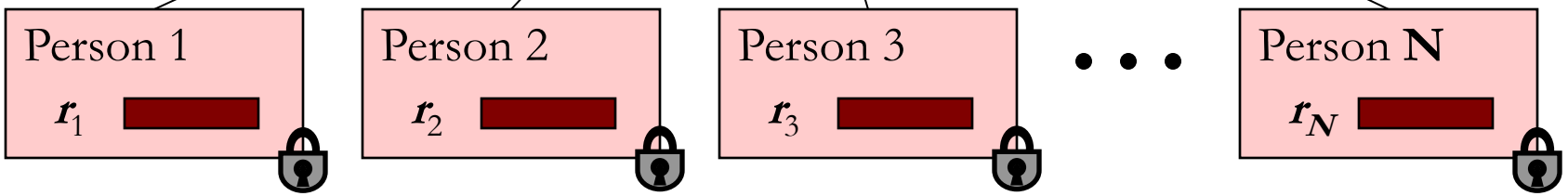


Statistical Database Privacy (untrusted collector)

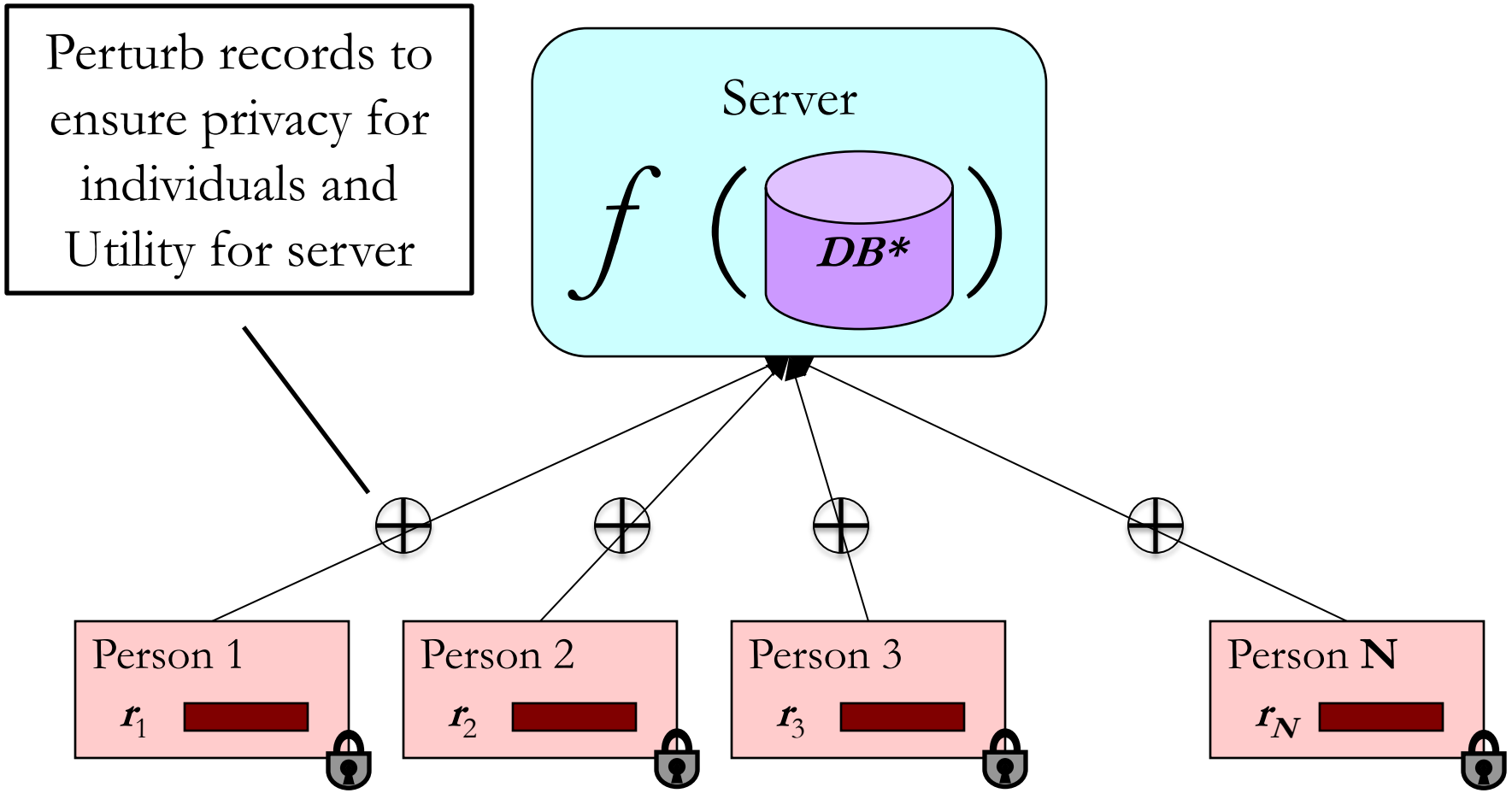
Server wants to compute f



Individuals do not want server to infer their records



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

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

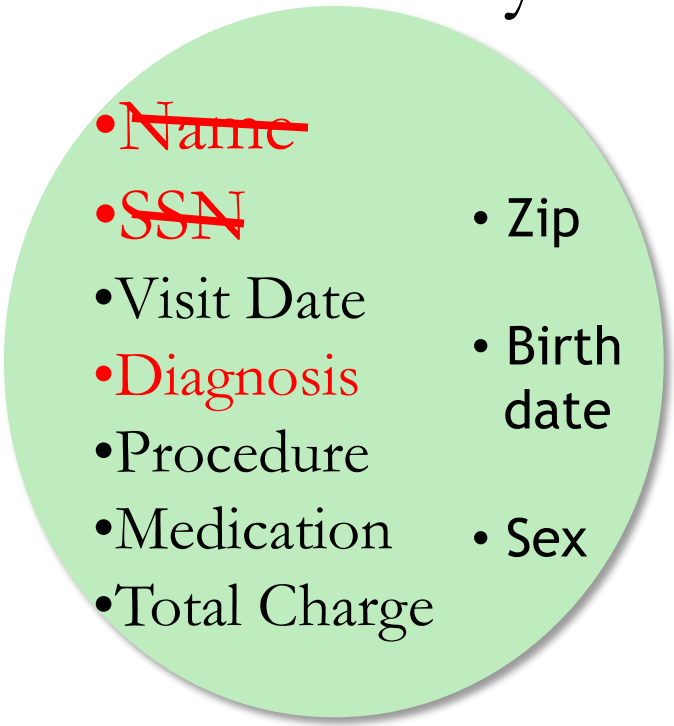
Privacy Problems

- 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.

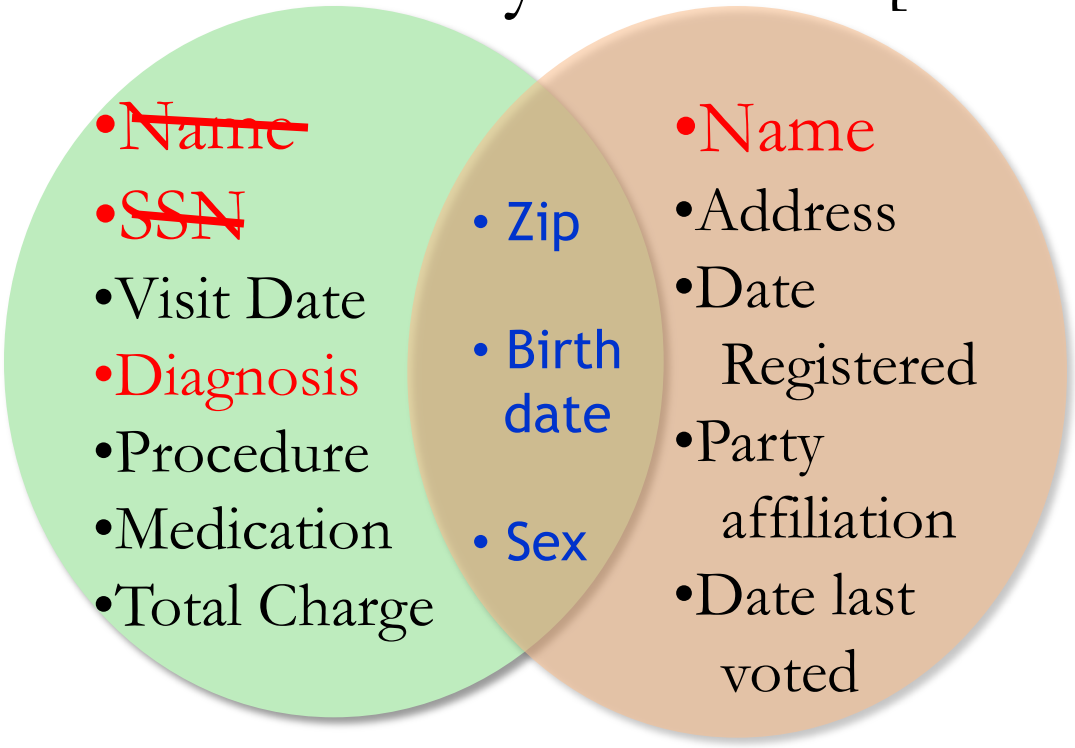
What *is* privacy?

The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]

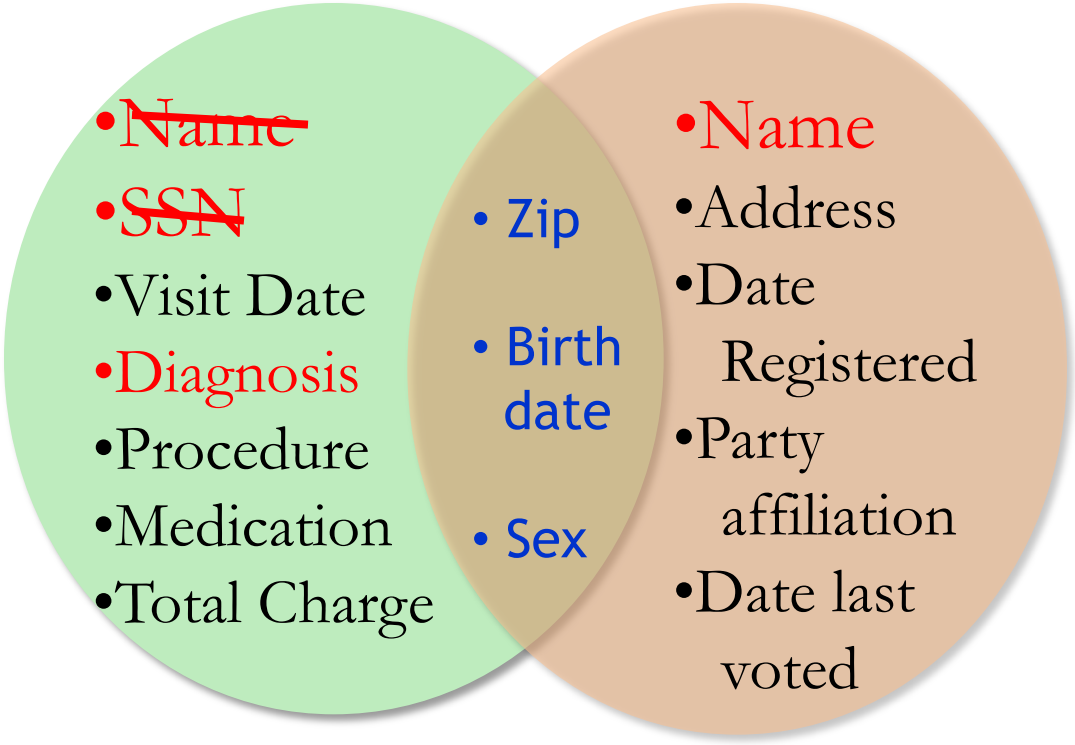


Medical Data Release

The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]



Linkage Attack



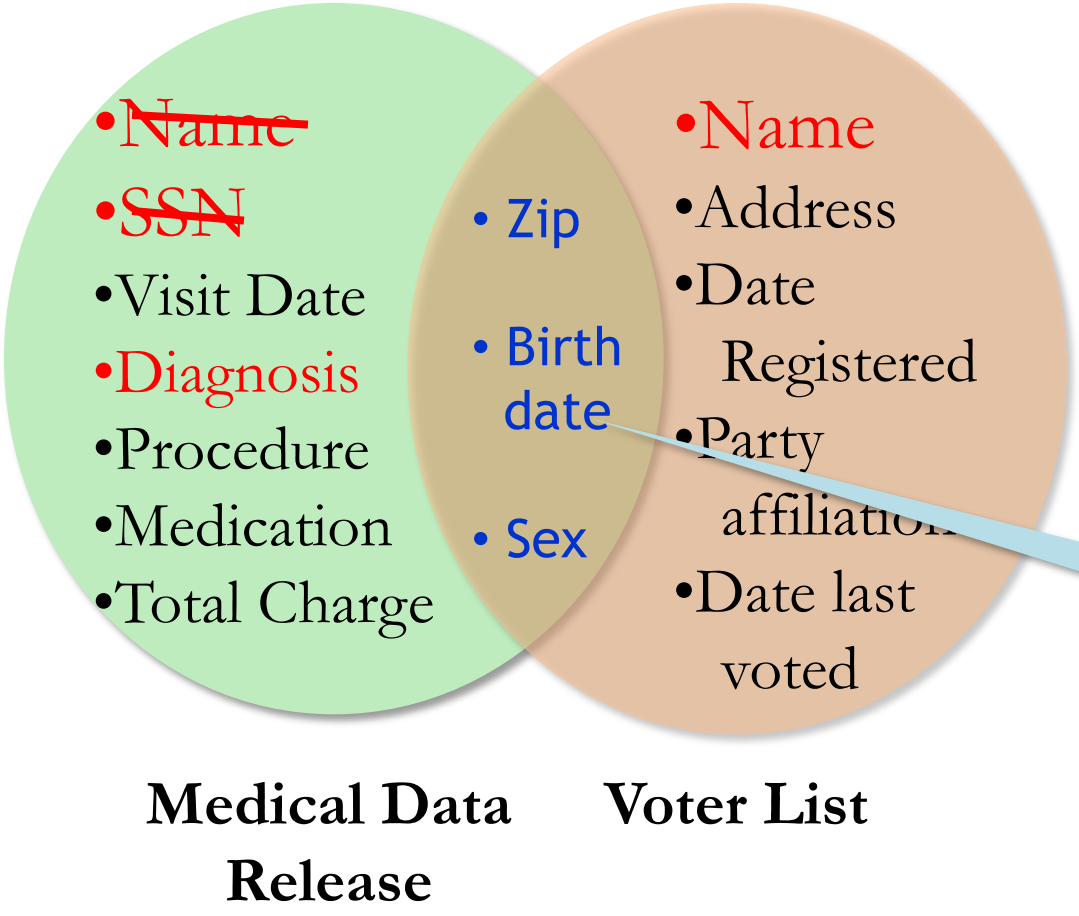
Medical Data Release

Voter List

- 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

Privacy Breach: Informal Definition

A privacy mechanism M(D)

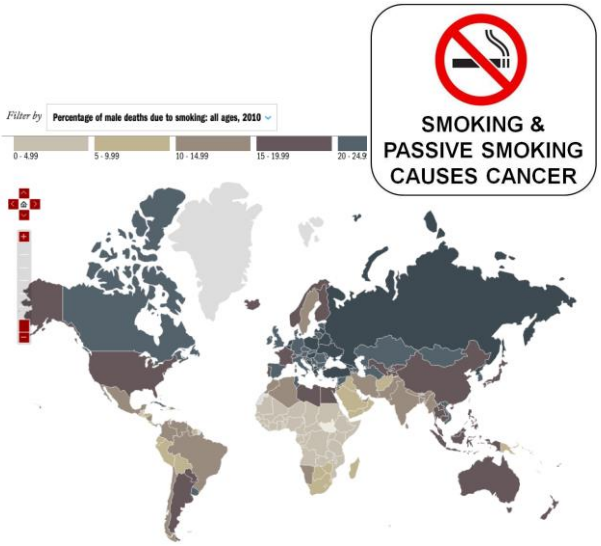
that allows

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

Problem: Background knowledge

Adversary knows prior knowledge about Umeko

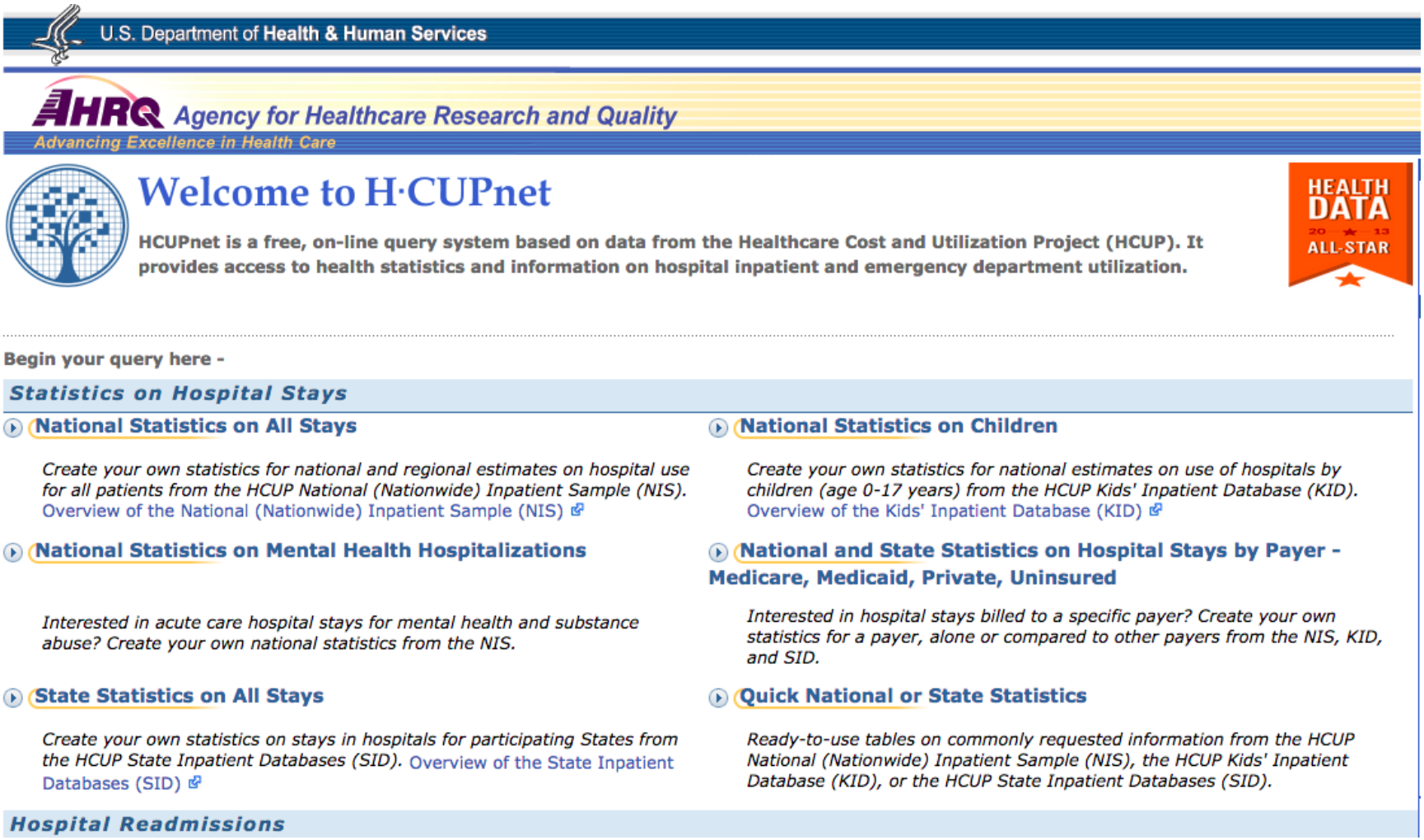
Adversary learns Umeko has Cancer

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]	*	*	*	*	*

Can reconstruct tight bounds on rest of data

[Vaidya et al AMIA 2013]

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.

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

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

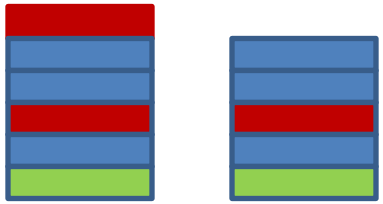
- Differential Privacy Definition
- Basic Algorithms
  - Laplace & Exponential Mechanism
  - Randomized Response
- Composition Theorems

Differential Privacy

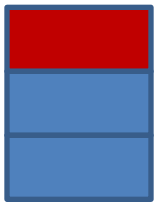
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

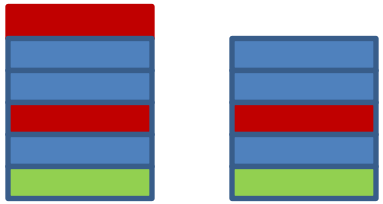
Adversary should not be able to distinguish between any D_1 and D_2 based on any O

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

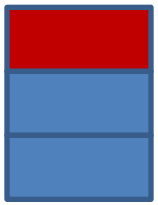
Why pairs of datasets *that differ in one row*?

For every pair of inputs that differ in one row

For every output …



D_1 D_2

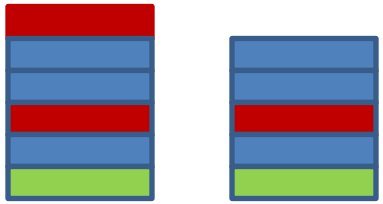


O

Simulate the presence or absence of a single record

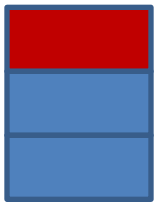
Why *all* pairs of datasets …?

For every pair of inputs that differ in one row



D_1 D_2

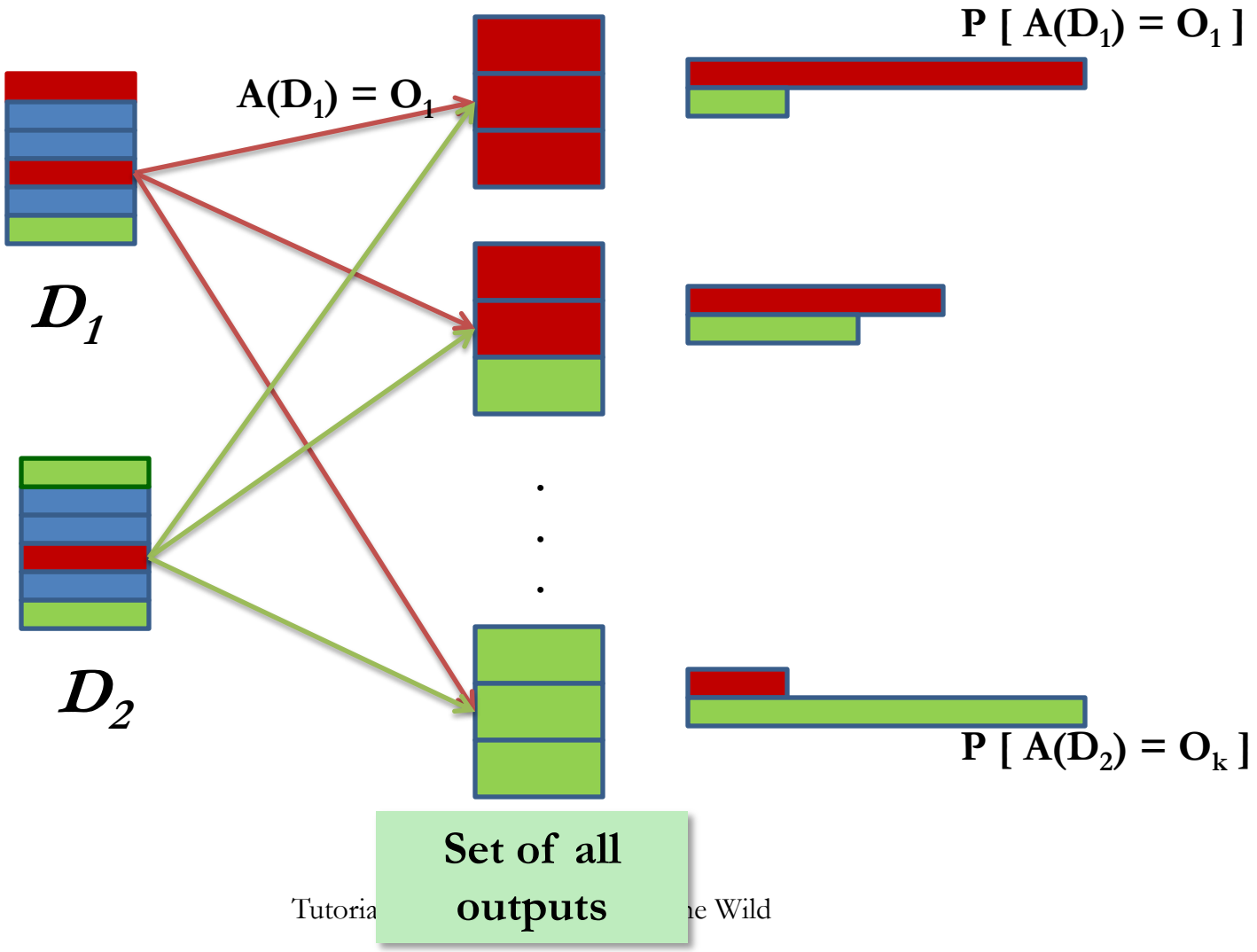
For every output …



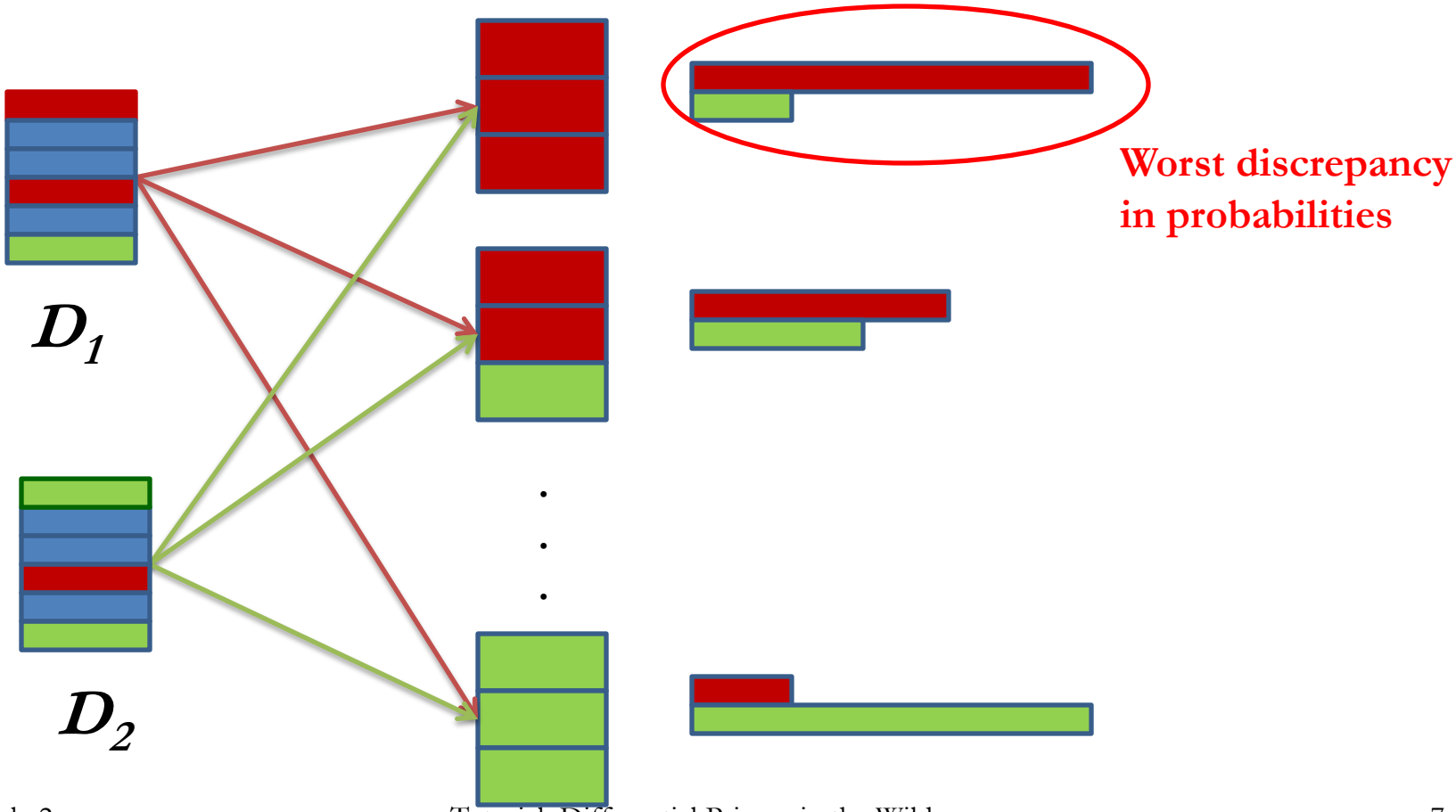
O

Guarantee holds no matter what the other records are.

Why *all* outputs?



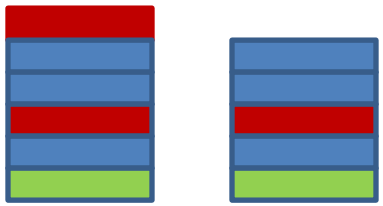
Should not be able to distinguish whether input was D_1 or D_2 no matter what the output



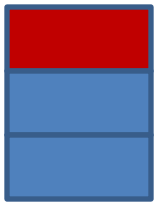
Privacy Parameterε

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 the ε more the privacy (and better the utility)

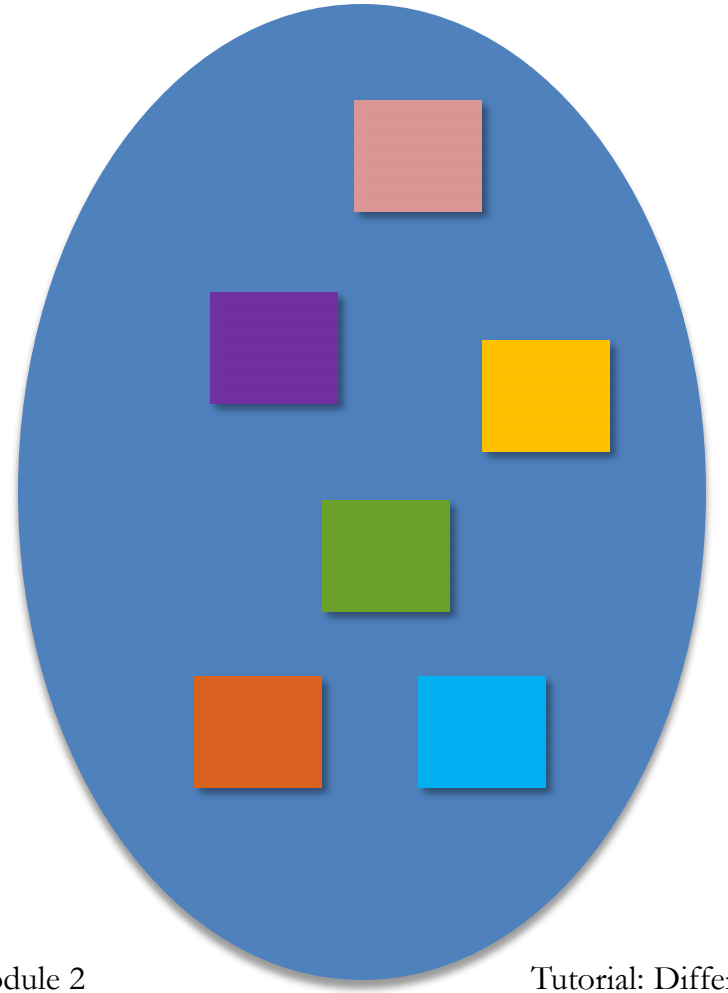
Outline of the Module 2

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

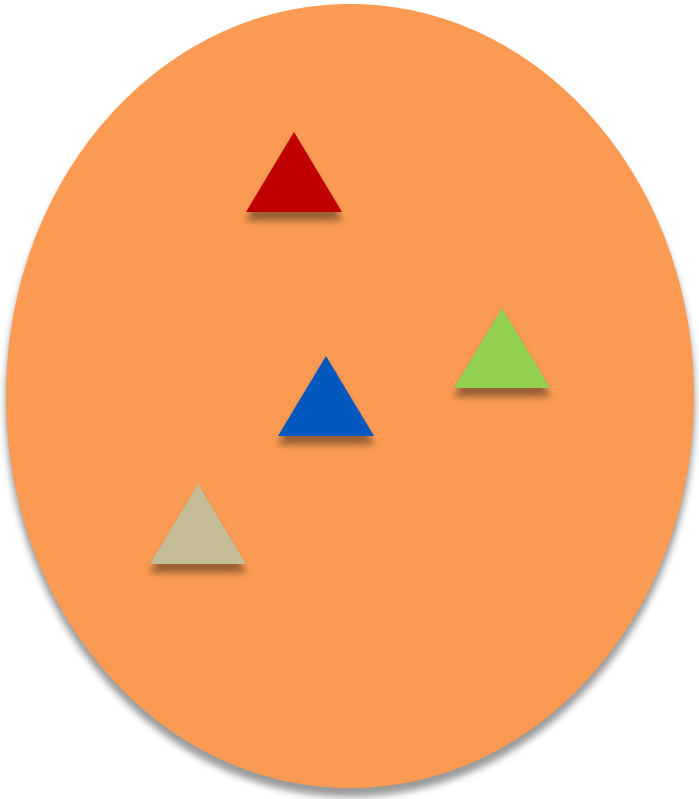
Can deterministic algorithms satisfy differential privacy?

Deterministic Algorithms do not satisfy differential privacy

Space of all inputs

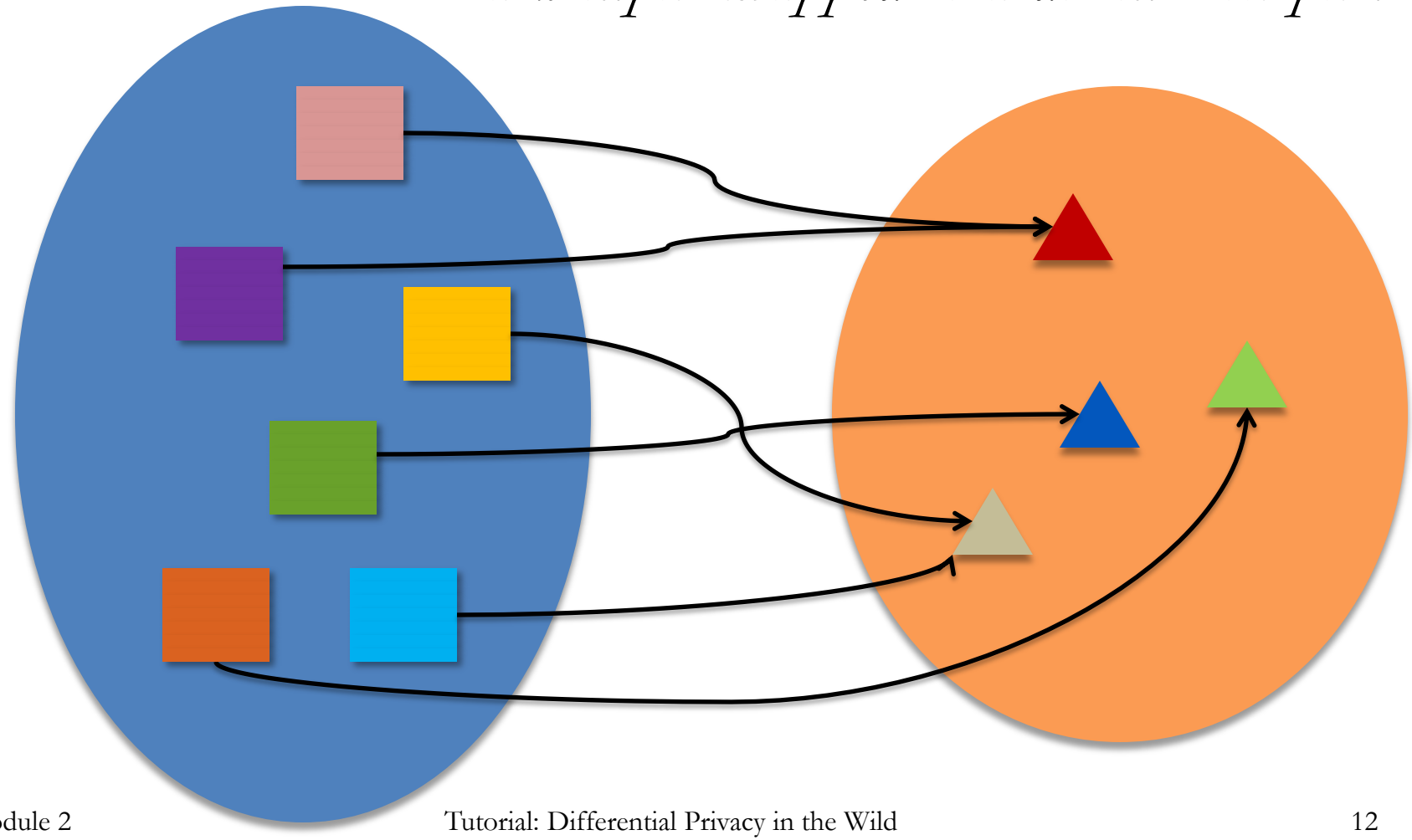


Space of all outputs (at least 2 distinct ouputs)

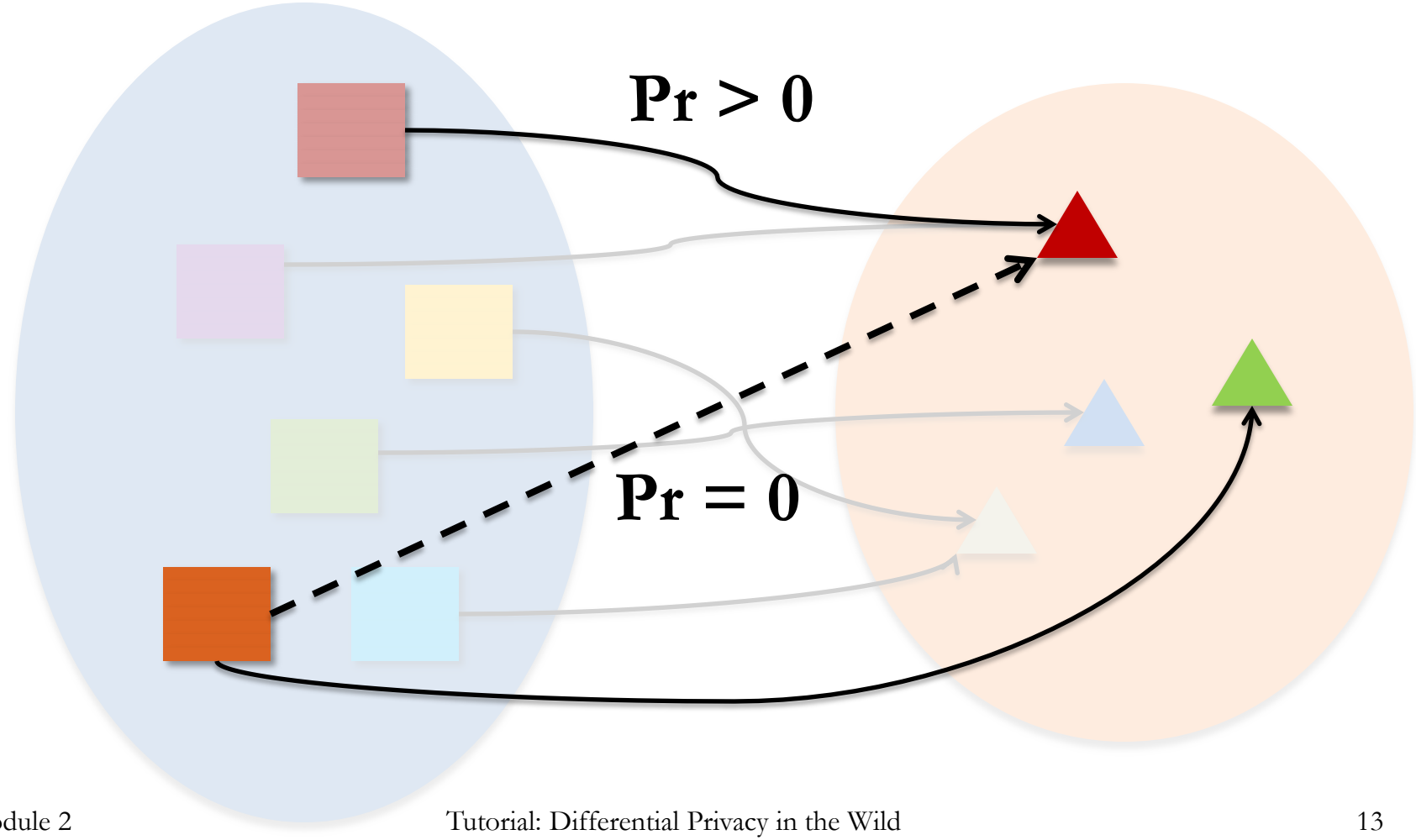


Deterministic Algorithms do not satisfy differential privacy

Each input mapped to a distinct output.



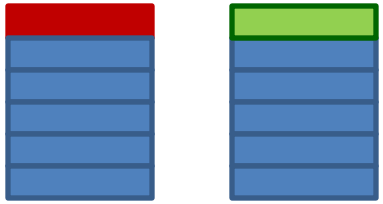
There exist two inputs that differ in one entry mapped to different outputs.



Random Sampling …

… also does not satisfy differential privacy

Input



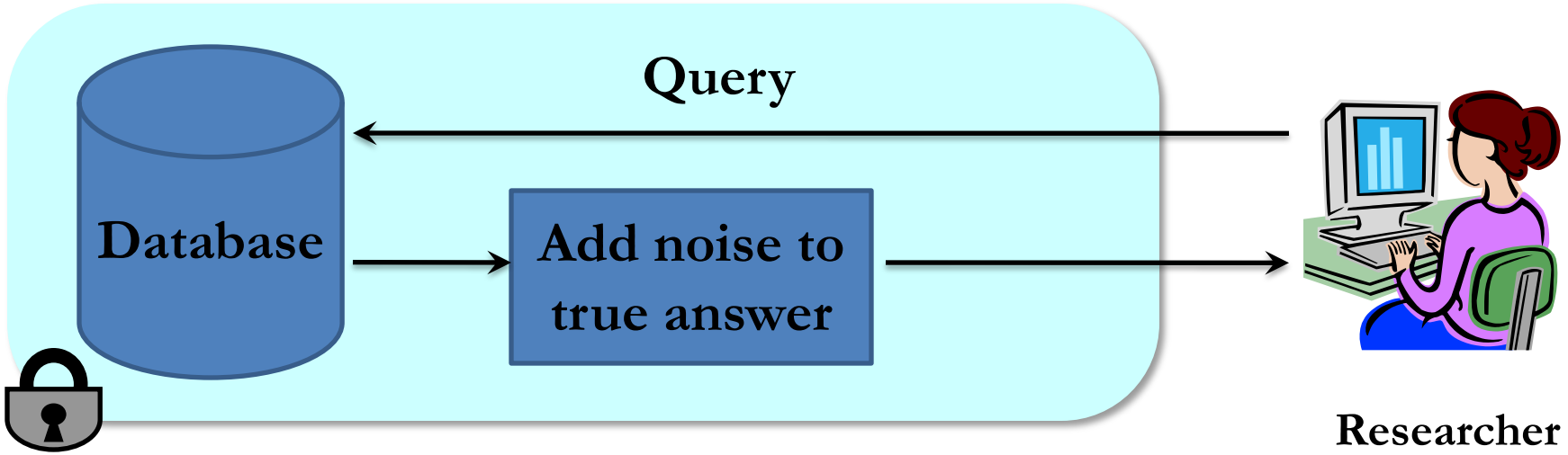
D_1 D_2

Output

O

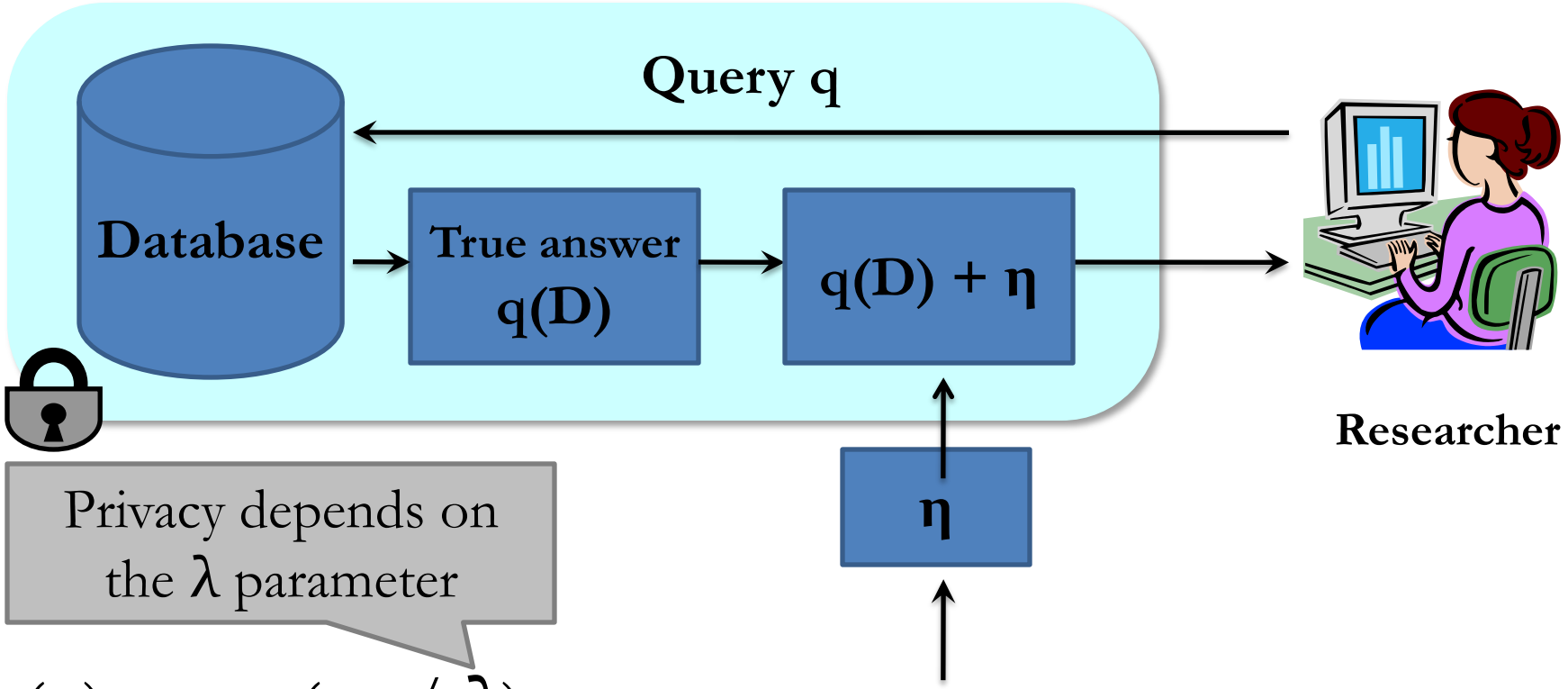
$$\Pr[D_2 \rightarrow O] = 0 \text{ implies } \log\left(\frac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]}\right) = \infty$$

Output Randomization



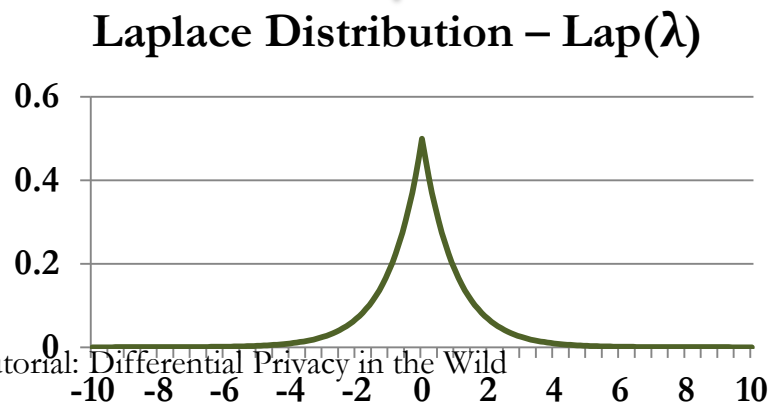
- Add noise to answers such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers.

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($1/\varepsilon$)
  - 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: $E(\text{true answer} - \text{noisy answer})^2$

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

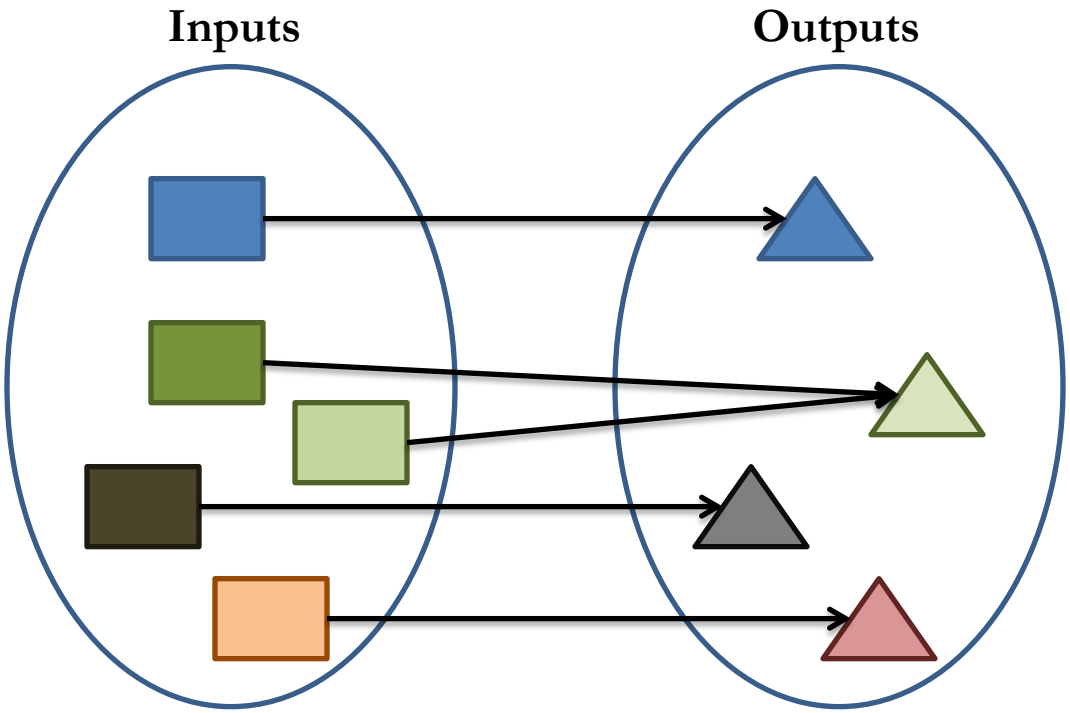
$$= 2*S(q)^2 / \varepsilon^2$$

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* $\rightarrow$ *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?

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

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

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

Dinur Nissim Result

[Dinur-Nissim PODS 2003]

- 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 the combination of their outputs 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, …, D_k such that each M_i satisfies ε_i -differential privacy,

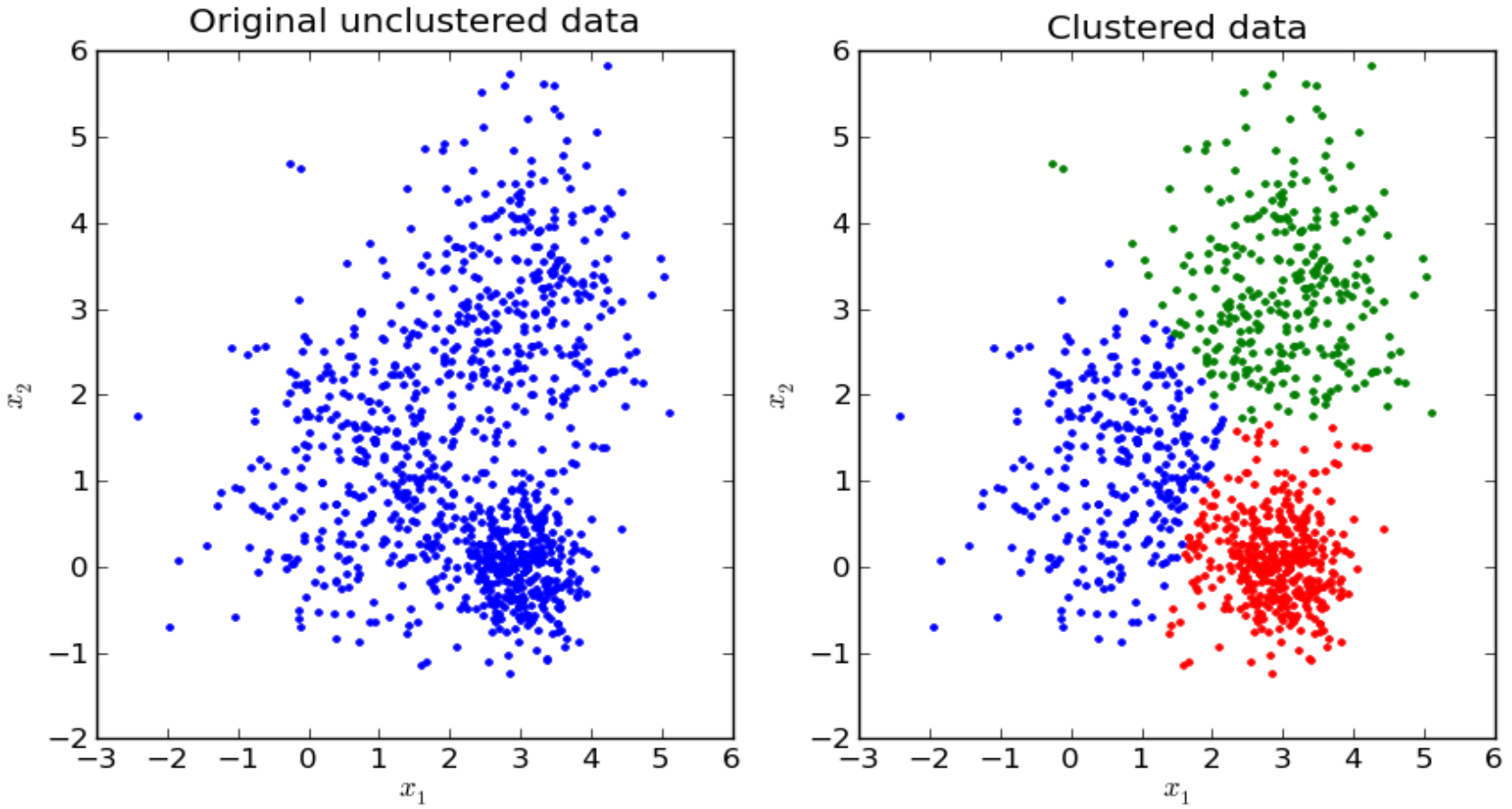
  then the combination of their outputs satisfies ε-differential privacy with$\varepsilon = \max\{\varepsilon_1,...,\varepsilon_k\}$

Postprocessing

- If M_1 is an εdifferentially 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

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

If 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

- Suppose we fix the number of iterations to T
  (Each iteration uses a privacy budget of ε / T)

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

  1. Noisily compute the size of each cluster
  2. Compute noisy sums of points in each cluster
  3. Recompute new clusters based on 1. and 2.

Sensitivity = 1

Sensitivity = size of domain = |dom|

Postprocessing
(no impact on privacy)

Differentially Private Kmeans

- Suppose we fix the number of iterations to T
  (Each iteration uses a privacy budget of ε / T)

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

  1. Noisily compute the size of each cluster
  2. Compute noisy sums of points in each clust
  3. Recompute new clusters based on 1. and 2.

S(q) = 1	ε / 2T
S(q) = |dom|	ε / 2T
S(q) = 0	0

Differentially Private Kmeans

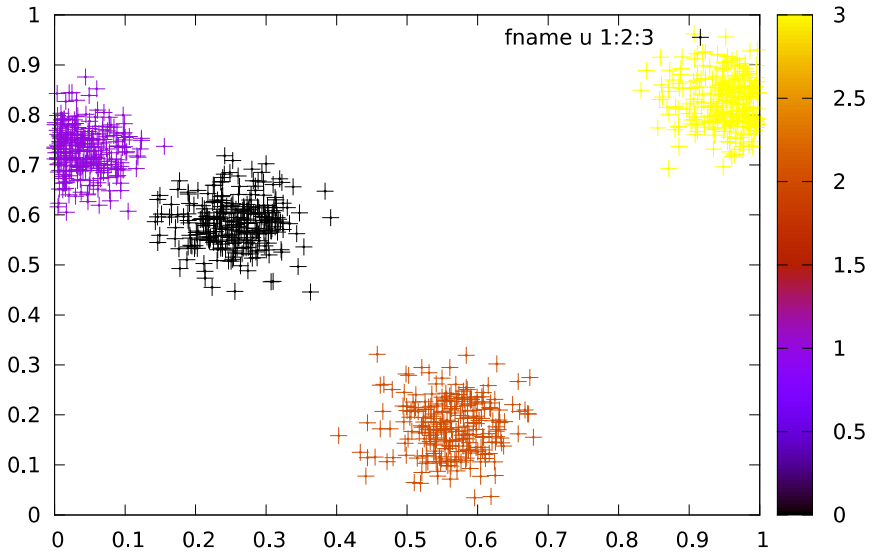
- Suppose we fix the number of iterations to T
  (Each iteration uses a privacy budget of ε / T)

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

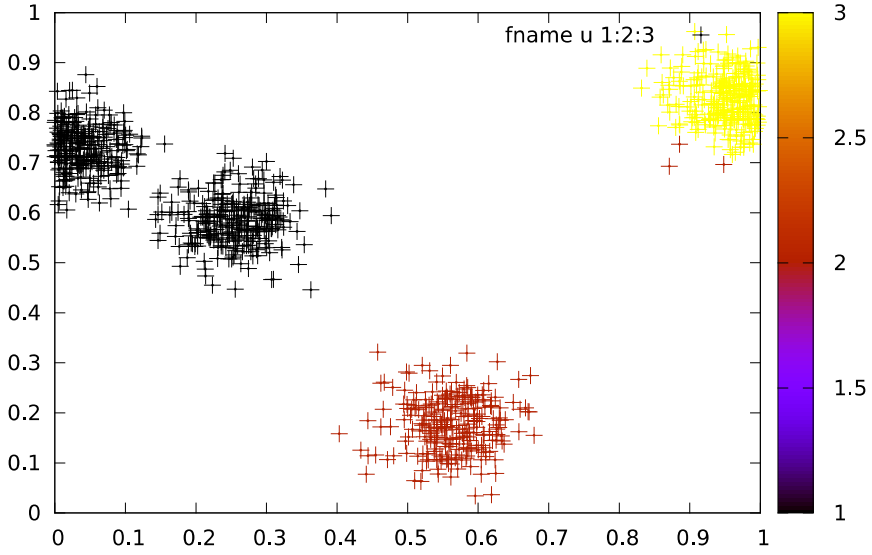
  1. Noisily compute the size of each cluster — Laplace(2T/ε)
  2. Compute noisy sums of points in each cluster — Laplace(2T |dom| /ε)
  3. Recompute new clusters based on 1. and 2. — Compute exactly

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.

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>;
```

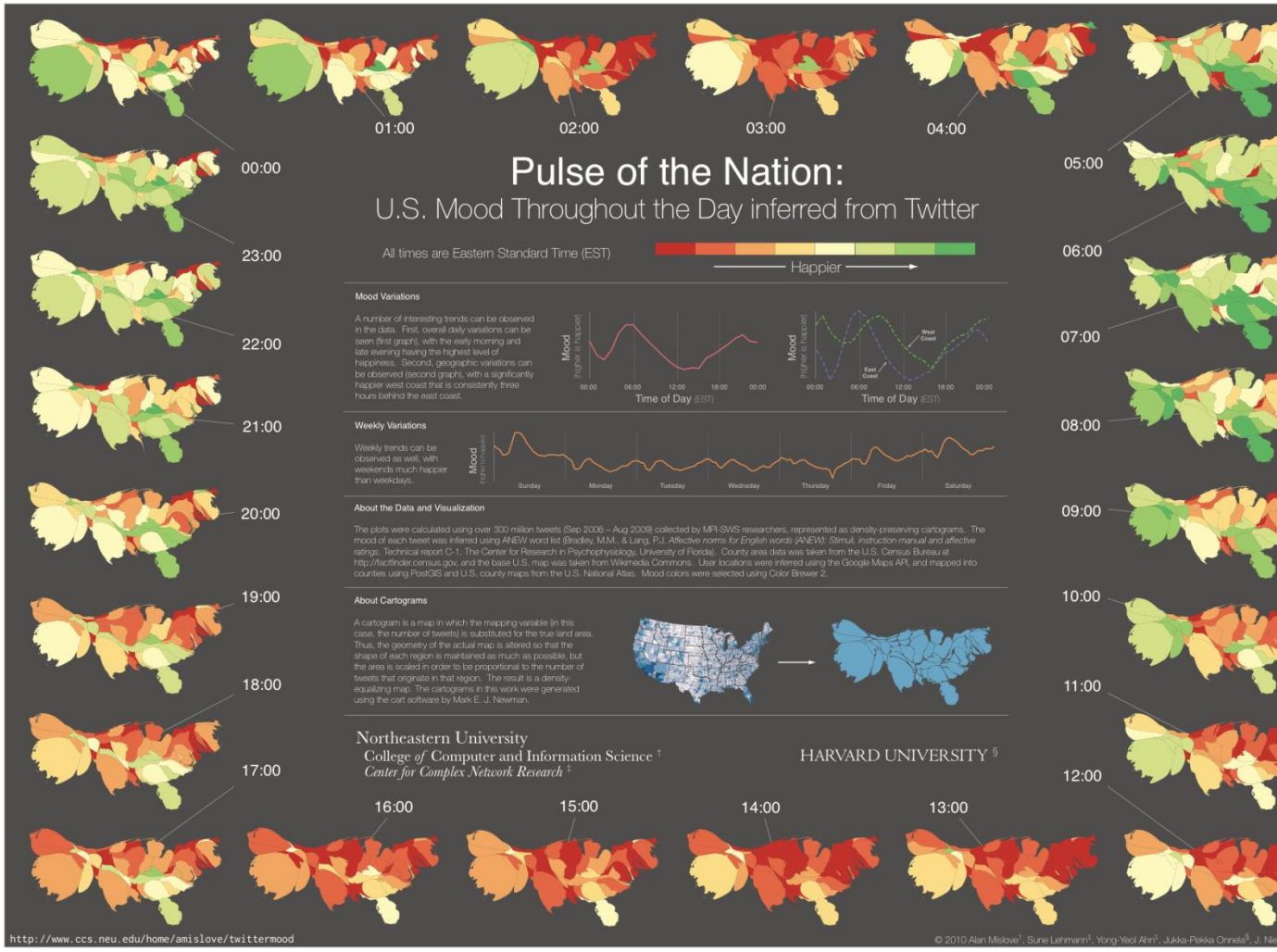
- **Output**: (noisy) answers to *W*
- Requirement: query answering algorithm satisfies differential privacy

* Many techniques can also support *linear queries*: compute SUM of user-defined function that maps each 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

- 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*
- U.S. Census Bureau statistics can typically be derived from *k*-way marginal over different combinations of available attributes
- ***Hundreds*** 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

- Pick a disease
- Pick a subsets of the population with (case) and without (control) the disease
- Extract SNPs (specific DNA subsequences)
  - For each SNP, usually 2 alleles are found (major – 1 and minor – 0)
- **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
  - Logistic Regression, Chi Square Test, …
- Goal: identify SNP 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

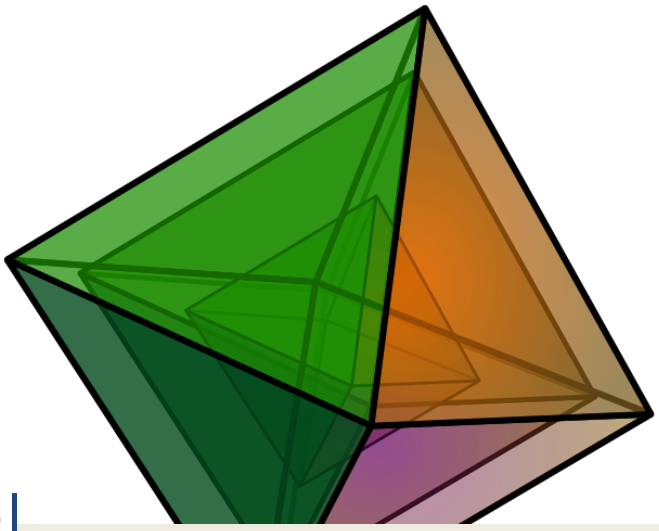
- Given negative result of Dinur-Nissim, is there any hope?
- Yes! Realistic workloads are not *adversarial*
- Key technical insight of BLR08: For any set of count queries W, there exists a small database D' consistent with D on every query in W.
  - Small: $O(VCDIM(W)/\varepsilon^2)$
    - VC-Dimension of W is independent of $|D|$!
  - $VCDIM(W) \leq \log(|W|)$

[BLR08]

Answering Exponentially Many Queries Offline

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

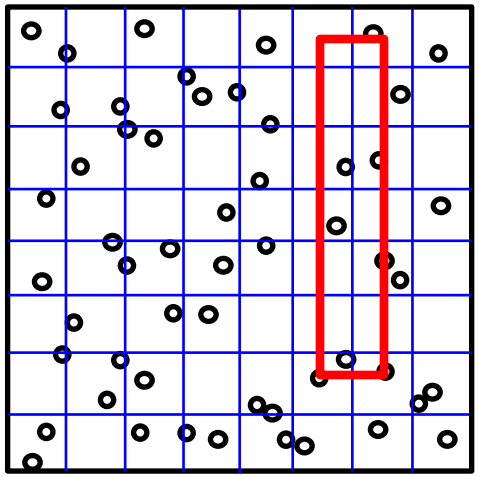
Input: range query workload *W*

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

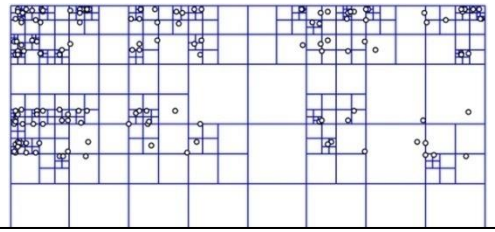
Limitations

- 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)

[CPSSY12]

Improved algorithm: private quad-tree



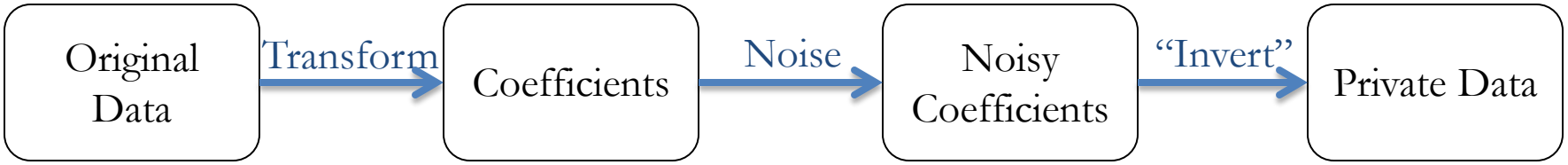
Exercise: Let $h' \leq h$ be height of resulting tree. Algorithm satisfies ε'-differential privacy for ε' equal to

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

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

Credit: Cormode

Data transformations



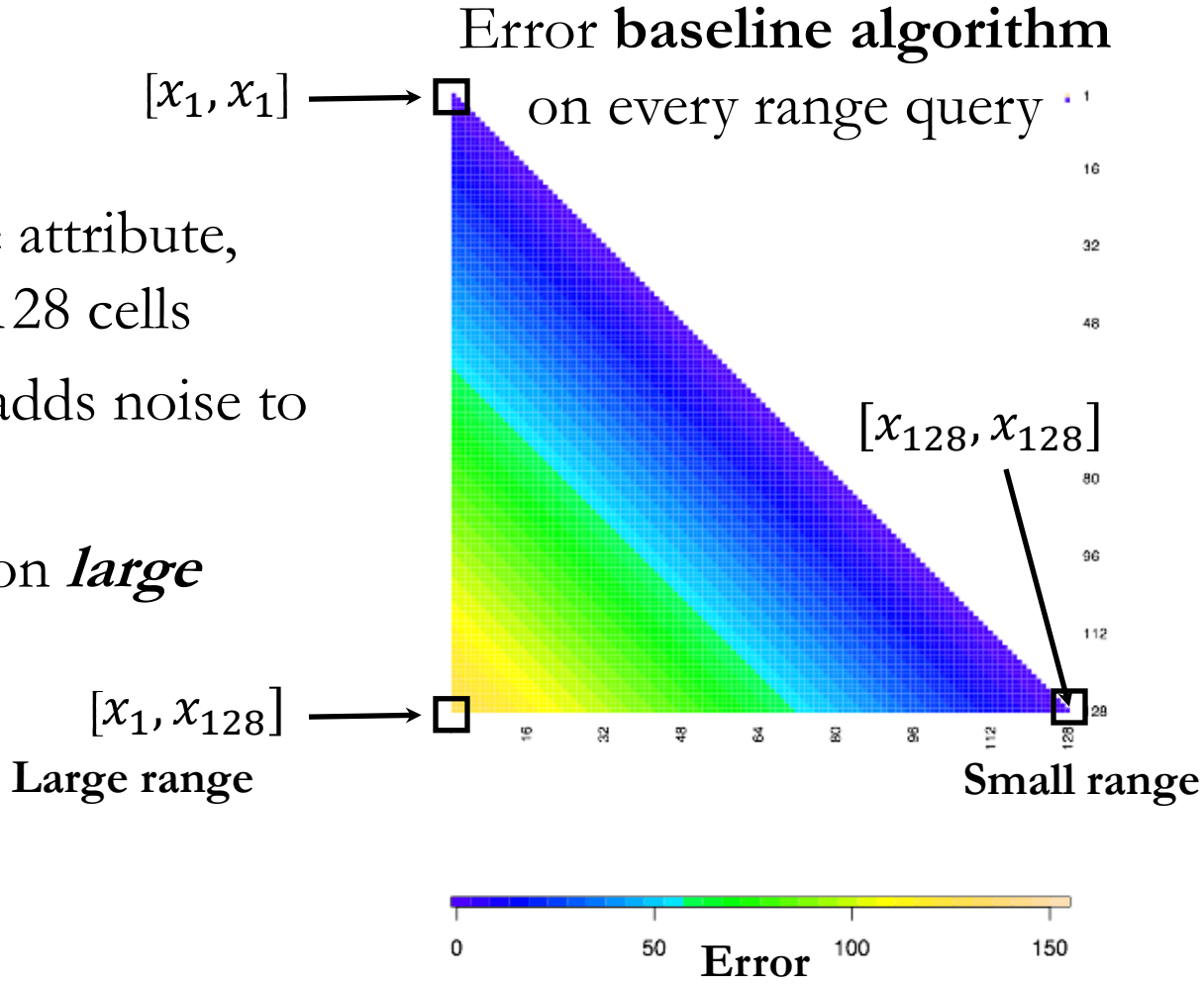
- Can think of trees as transform of input
- Can apply other data transformations
- **General idea:**
  - Apply transform of data
  - Add noise in the transformed space (based on sensitivity)
  - Publish noisy coefficients, or "invert" transform
- **Goal**: pick a low sensitivity transform that preserves good properties of data

Linear transformations

- Approach
  - Discretize domain to finest granularity cells
  - Use Laplace mechanism to answer batch of queries, each of which is linear combination of cell counts
- 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")

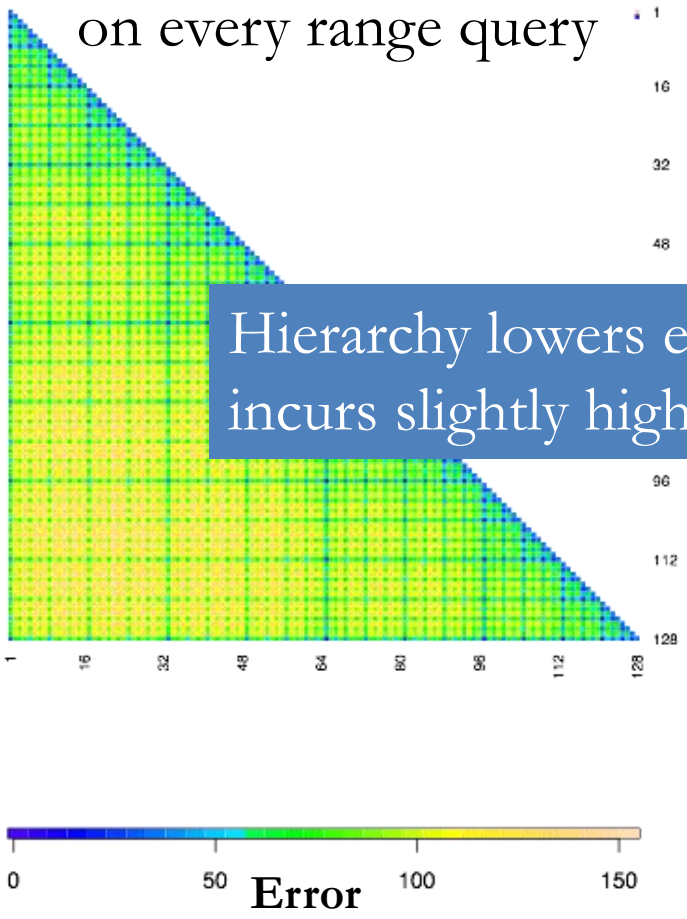
Error analysis

- Domain is single attribute, discretized into 128 cells
- Baseline simply adds noise to every cell count
- Incurs *high error* on ***large ranges***

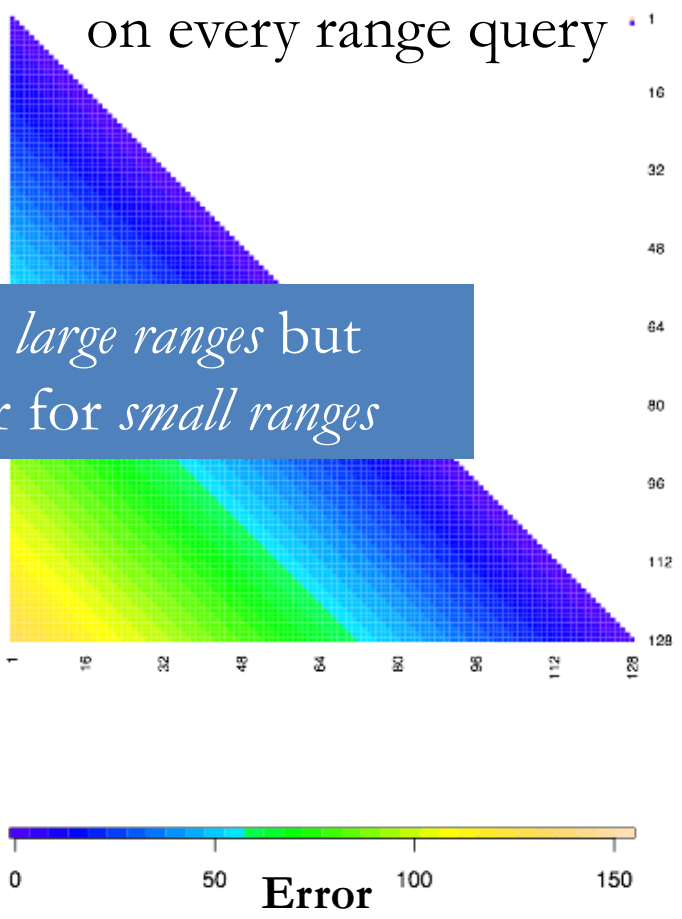


Error comparison

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*

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

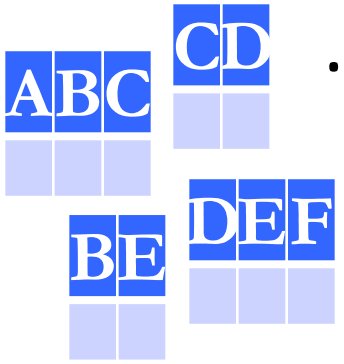
- Generally an under-studied area
- Two algorithms, both synthetic data generators
  - PrivBayes [ZCPSX14]
  - DualQuery [GAHRW14]
- Common properties
  - Limited to binary attributes
  - Designed to support low-order marginals (and other workloads well approximated by marginals, such as classification)

[ZCPSX14]

PrivBayes

High-dimensional table T

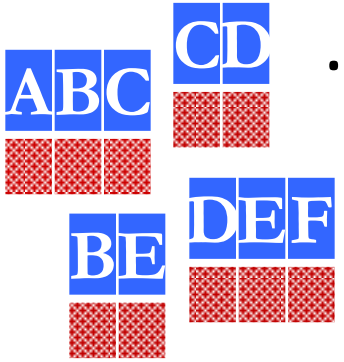
decompose



Low-dimensional tables

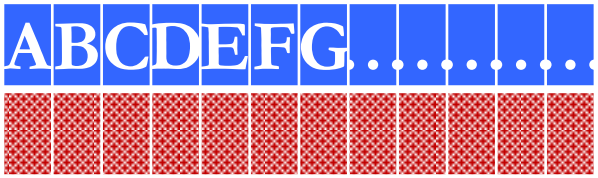
Add noise

- 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

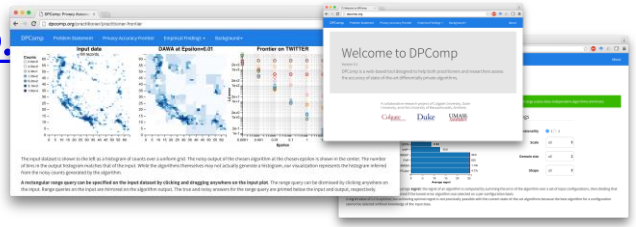
1. Let Q^t be distribution over queries (Q^1 is uniform)
2. For t = 1… T
   a) S ← Query player samples s queries from Q^t — Privacy
   b) Data player finds record x_t that maximizes total answer on S — NP-Hard
   c) Q^{t+1} ← Query player updates(x_t, D, Q^t)
3. Output synthetic database x_1, …, x_t

- 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

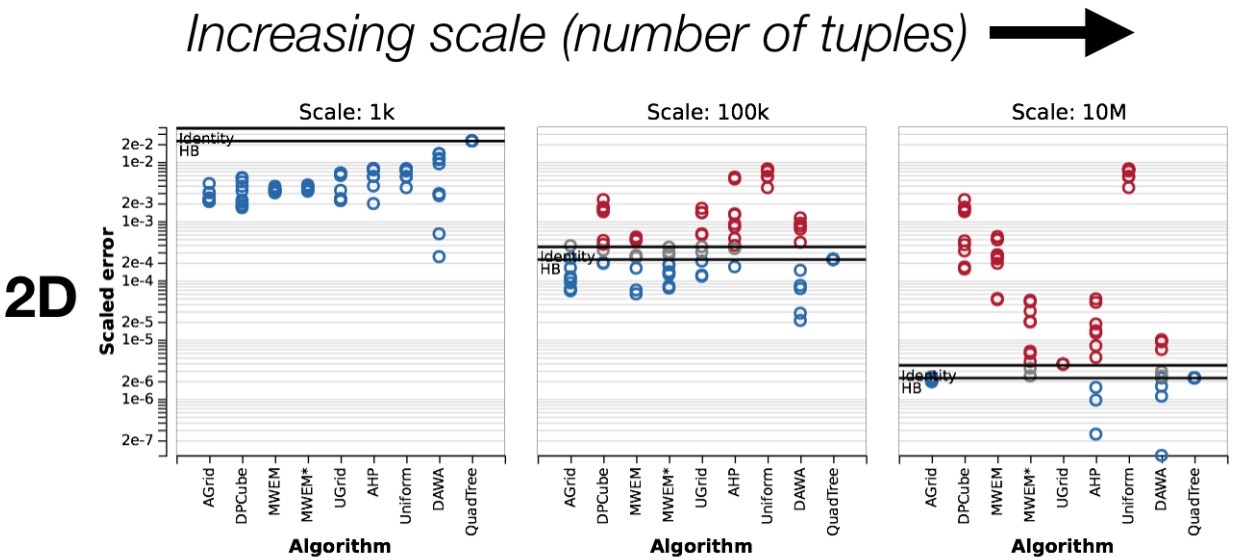
[HMMCZ16]

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 https://www.dpcomp.



One finding from [HMMCZ16]: Some data-dependent algorithms fail to offer benefits at larger scales (no. of tuples)



Open questions

- Robust and private algorithm selection
- Error bounds for data-dependent algorithms
- Empirical evaluation of algorithms for high dimensional data

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.

Differential Privacy in the Wild (Part 2)

A Tutorial on Current Practices and Open Challenges

MODULE 4:
APPLYING DIFFERENTIAL PRIVACY

Outline of the Tutorial

1. Privacy Problem Statement
2. Differential Privacy
3. Algorithms for Tabular Data

   Break

4. Applying Differential Privacy
5. Privacy beyond tabular Data
6. Applications II

Module 4: Applying Differential Privacy

- Real world deployments of differential privacy
  - OnTheMap 

  - RAPPOR 

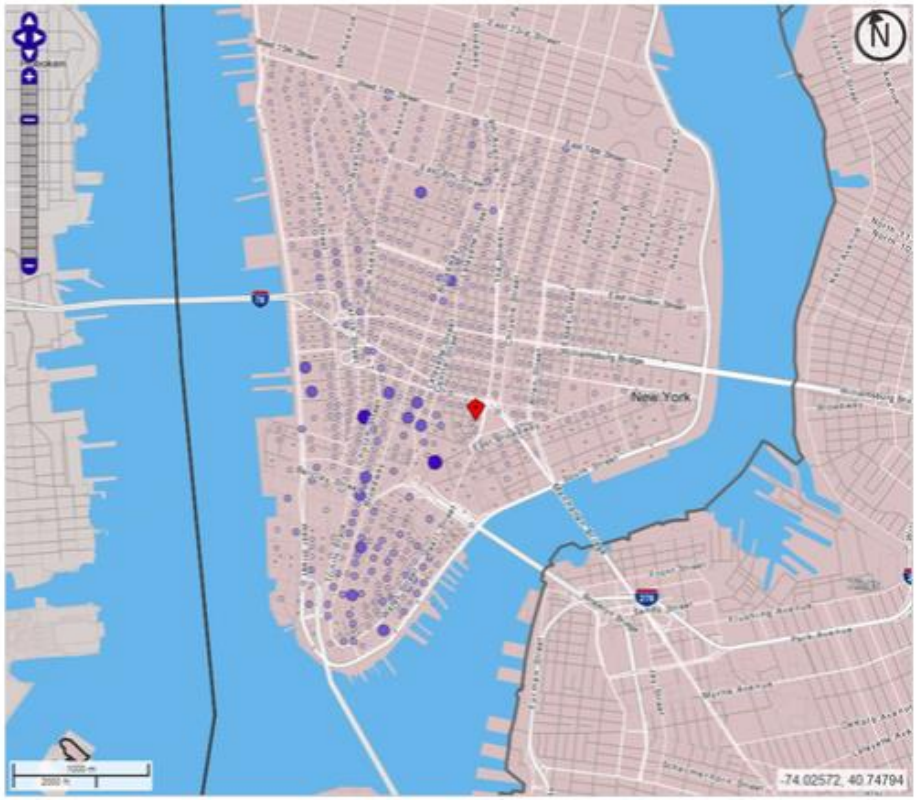
- Attacks on differential privacy implementations
  - Side channel attacks
  - Floating point attacks

United States Census Bureau

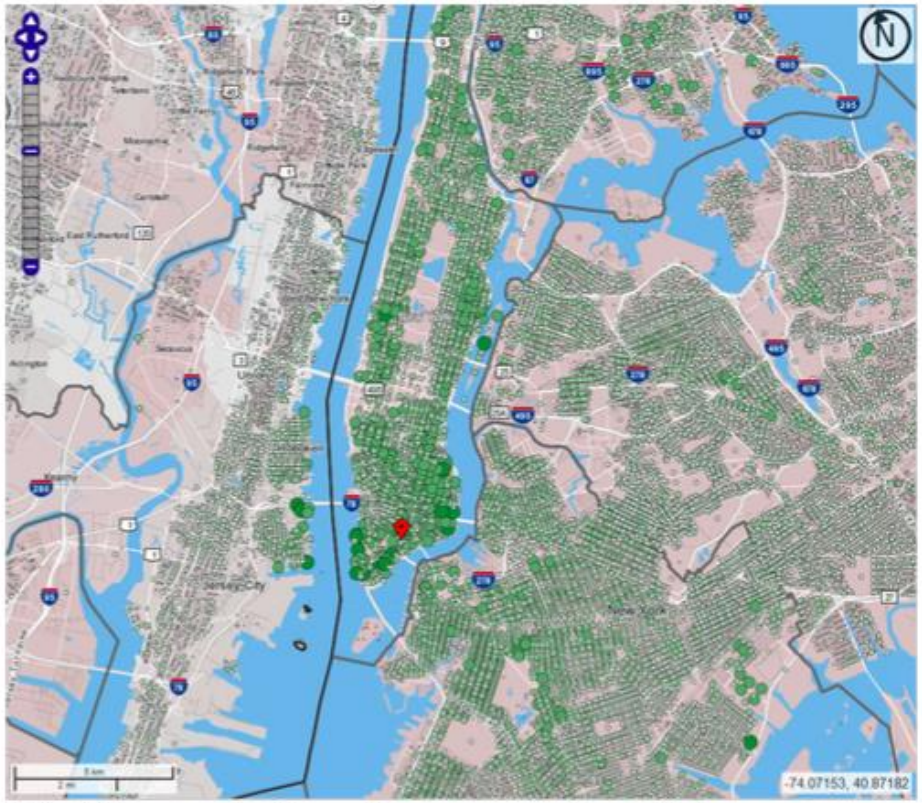
OnTheMap

http://onthemap.ces.census.gov/

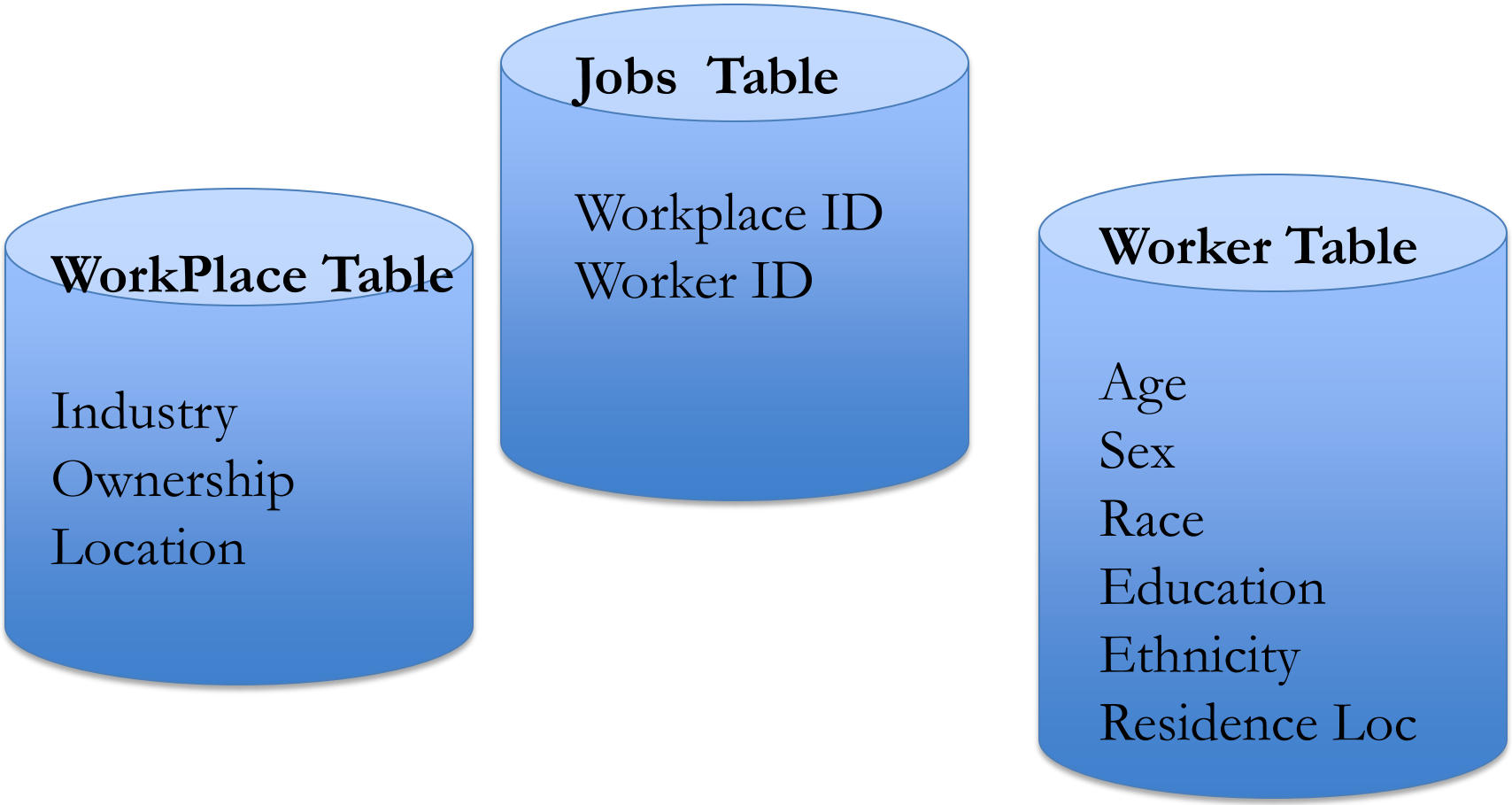
Employment in Lower Manhattan



Residential pattern of workers employed in Lower Manhattan



The maps above show LODES data in New York City in the OnTheMap application. The map on the left shows employment by census block in Lower Manhattan (in dense urban areas one census block is often equivalent to one city block). Large, dark dots have more employment than small, light dots. The map on the right shows the residential patterns of the same workers (those employed in Lower Manhattan). Workers employed in Lower Manhattan live throughout New York City as well as in New Jersey and other areas of New York state.


Jobs Table
Workplace ID
Worker ID
WorkPlace Table
Industry
Ownership
Location
Worker Table
Age
Sex
Race
Education
Ethnicity
Residence Loc

Why release such data?

- Quarterly Workforce Indicators
  - Total employment
  - Average Earnings
  - New Hires & Separations
  - Unemployment Statistics

E.g., Missouri state used this data to formulate a method allowing **QWI to suggest industrial sectors where transitional training might be most effective** … to proactively reduce time spent on unemployment insurance …

Why privacy is needed?

US Code: Title 13 CENSUS

It is against the law to make any publication whereby the data furnished by any particular establishment or individual under this title can be identified.

Violating the statutory confidentiality pledge can result in fines of up to $250,000 and potential imprisonment for up to five years.

Publish a Synthetic database

- Rather than have a system for performing query answering, release a synthetic dataset.
  - Analyst can now perform arbitrary analysis on this synthetic dataset.
- Very popular amongst statisticians
  - For ensuring privacy
  - For imputing (filling in) missing values
- Approach used by U.S. Census Bureau for several data products

Publish a Synthetic Database

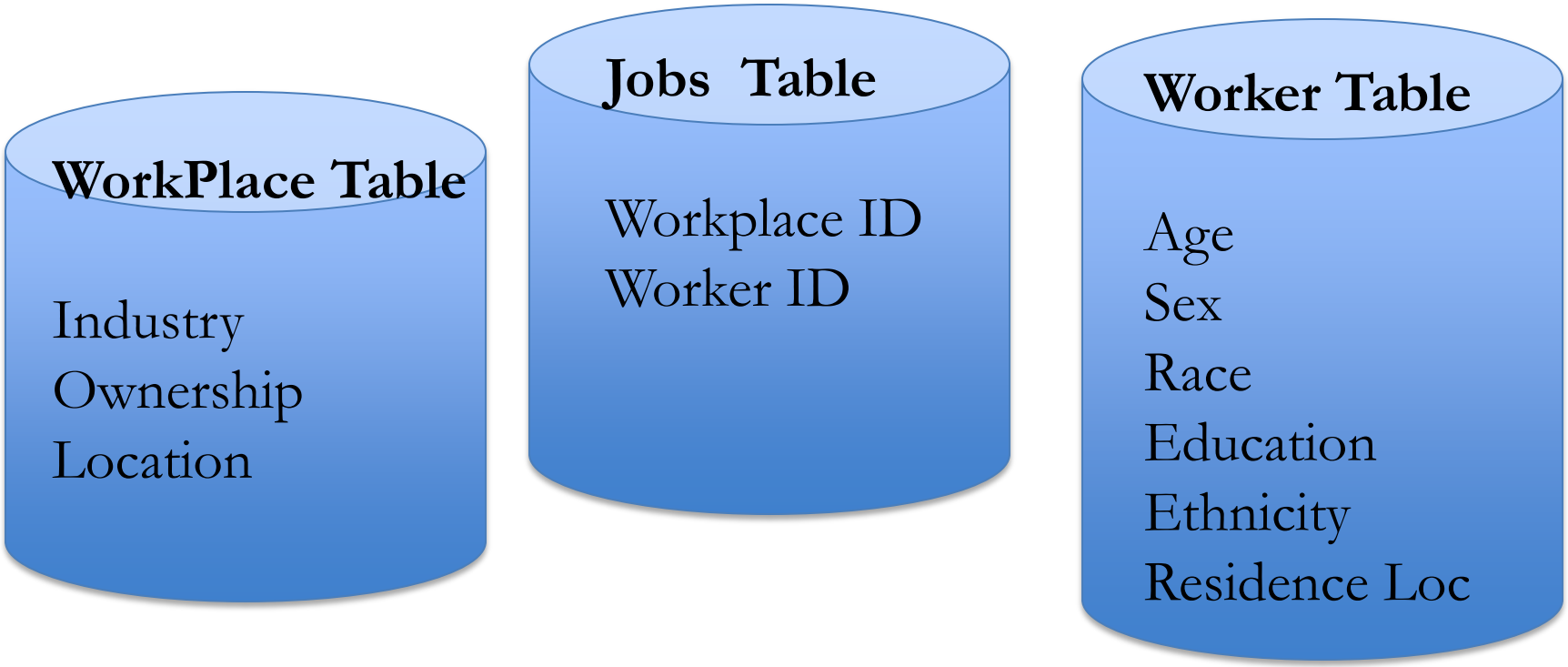
- Sanitize the dataset one time
- Analyst can perform arbitrary computations on the synthetic datasets

- Unlike in query answering systems
  - No need to maintain state (of queries asked)
  - No need to track privacy loss across queries or across analysts

Synthetic Data and US Census

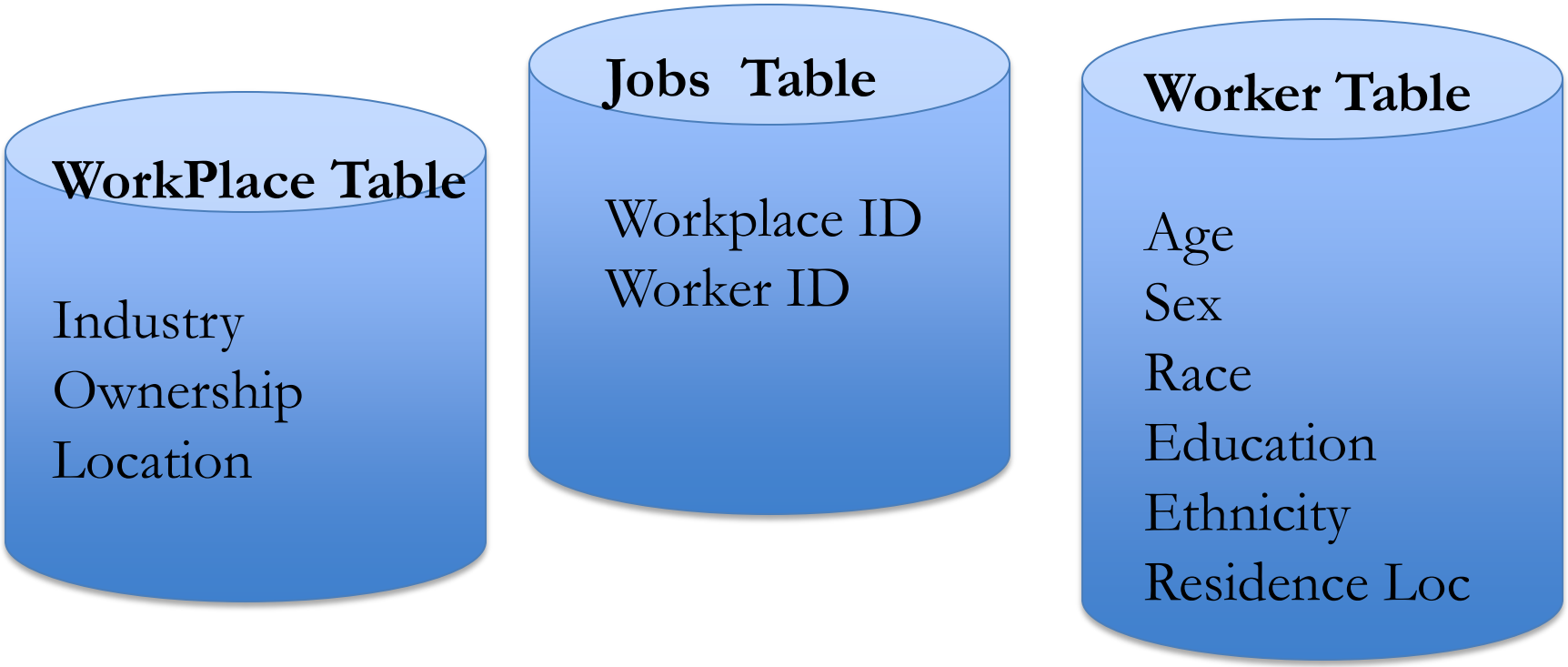
- U.S. Census Bureau uses synthetic data to share data from Survey of Income and Program Participation, American Community Survey, Longitudinal Business Database and OnTheMap

- Only OnTheMap has formal guarantee of privacy.



Worker ID	Residence	Workplace
1223	MD11511	DC22122
1332	MD2123	DC22122
1432	VA11211	DC22122
2345	PA12121	DC24132
1432	PA11122	DC24132
1665	MD1121	DC24132
1244	DC22122	DC22122

[MKAGV08] proposed differentially private algortihms to release residences in 2008.



[MKAGV08] proposed differentially private algortihms to release residences in 2008.

[HMKGAV15] proposed differentially private algorithms to release the rest of the attributes.

OnTheMap

Residence (Sensitive)

Workplace (Quasi-identifier)

Worker ID	Origin	Destination
1223	MD11511	DC22122
1332	MD2123	DC22122
1432	VA11211	DC22122
2345	PA12121	DC24132
1432	PA11122	DC24132
1665	MD1121	DC24132
1244	DC22122	DC22122

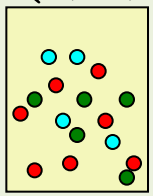
Census Blocks

A Synthetic Data Generator

(Dirichlet resampling)

Step 1: **Noise Addition** *(for each destination)*

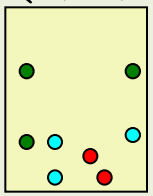
D (7, 5, 4)



Multi-set of Origins for workers in Washington DC.

+

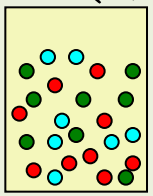
A (2, 3, 3)



Noise (fake workers)

=

D+A (9, 8, 7)



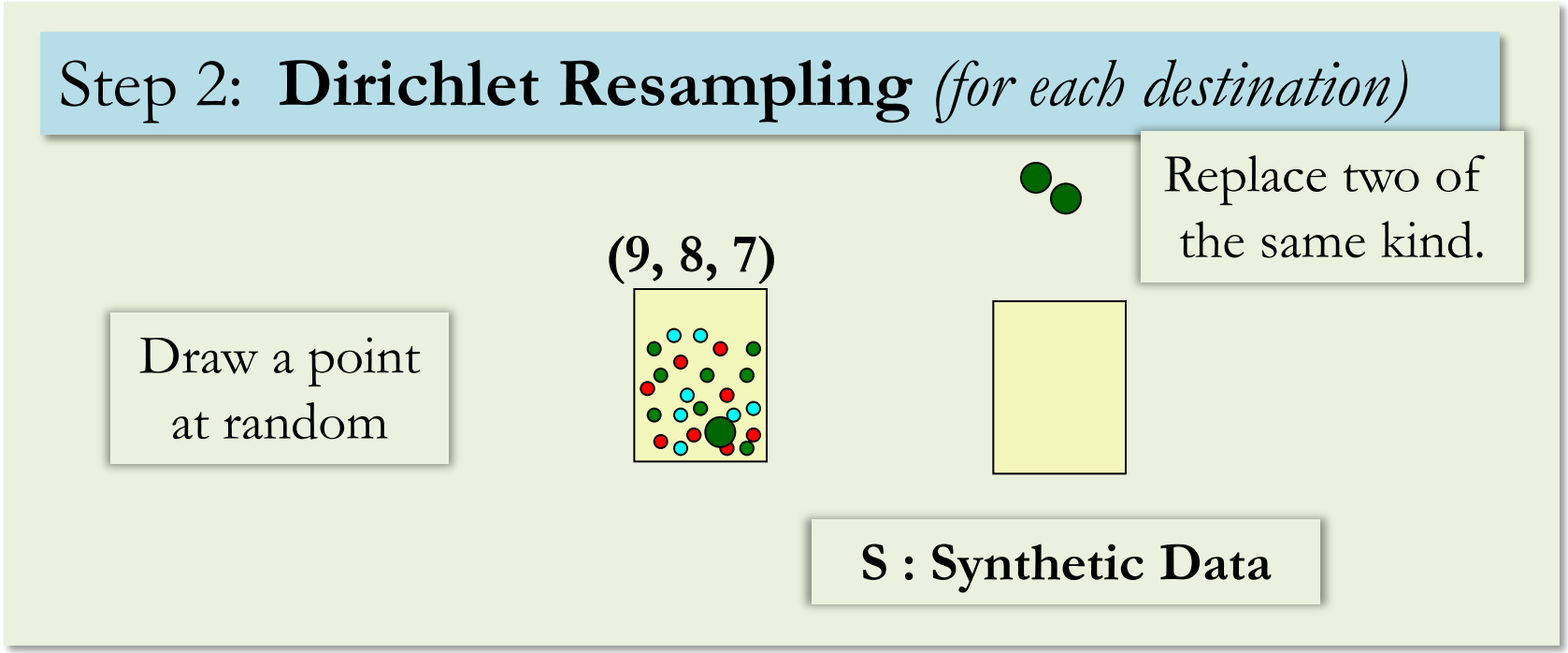
Noise infused data

● **Washington DC** ● **Somerset** ○ **Fuller**

If some residence block has count 0, then no noise is added.

A Synthetic Data Generator

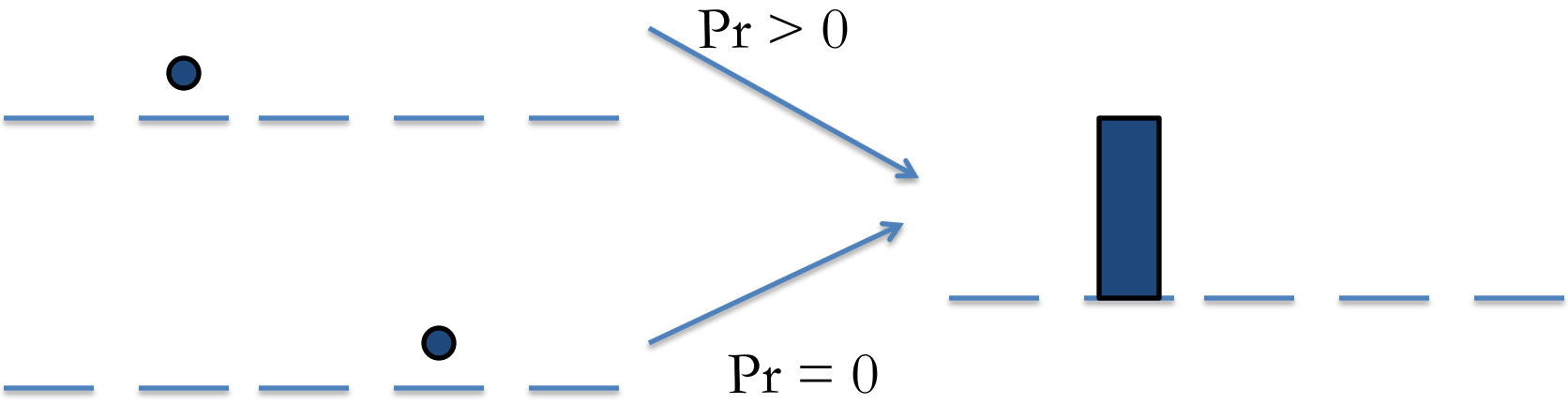
(Dirichlet resampling)



If some residence block has count 0, then no synthetic point is sampled from it.

Attack

- A synthesizer that does not add noise to blocks that have 0 count results in re-identification attacks.



Dirichlet Resampling has high sensitivity

- Noise required per block: (differential privacy)

1 million original and synthetic workers.

lesser privacy →

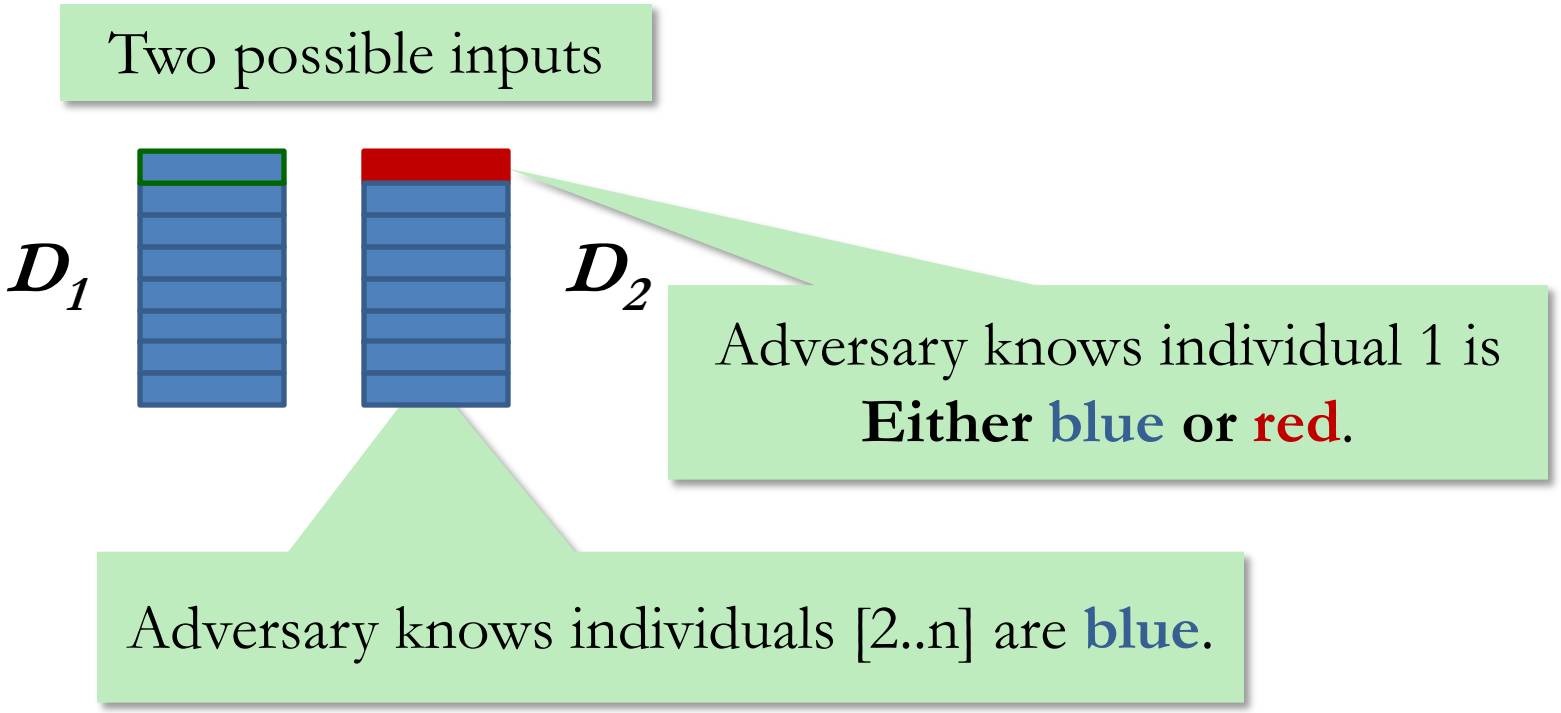
Privacy (e^{ε} =)	5	10	20	50
Noise per block	$25x10^4$	$11x10^4$	$5x10^4$	$2x10^4$

- Add noise to every block on the map.

  There are 8 million Census blocks on the map!

  1 million original workers and 16 billion fake workers!!!

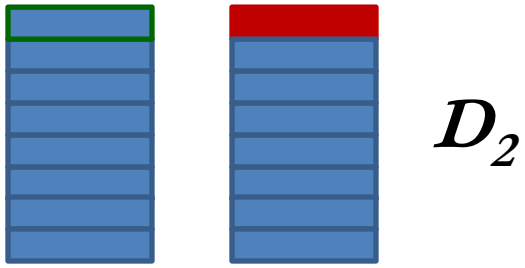
Intuition



blue and red are two different origin blocks.

Intuition

Two possible inputs



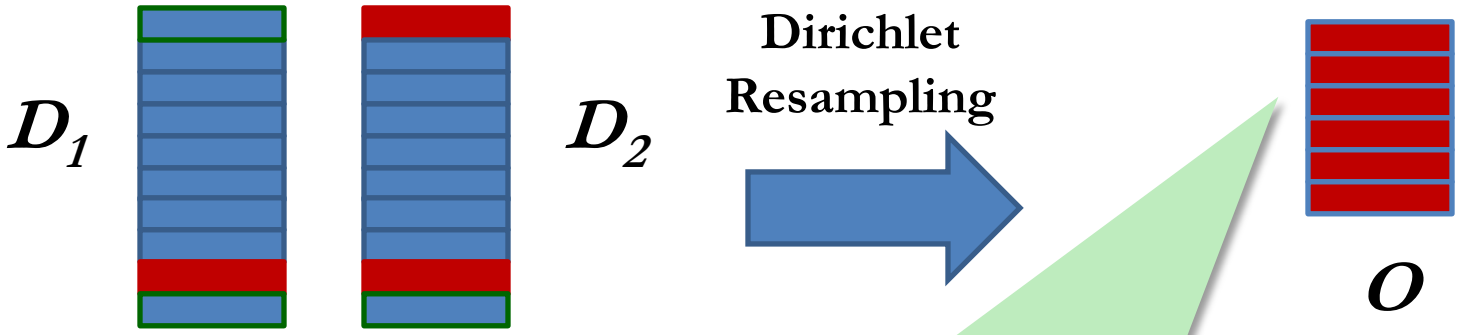
Noise Addition

blue and red are two different origin blocks.

Intuition

Noise infused inputs

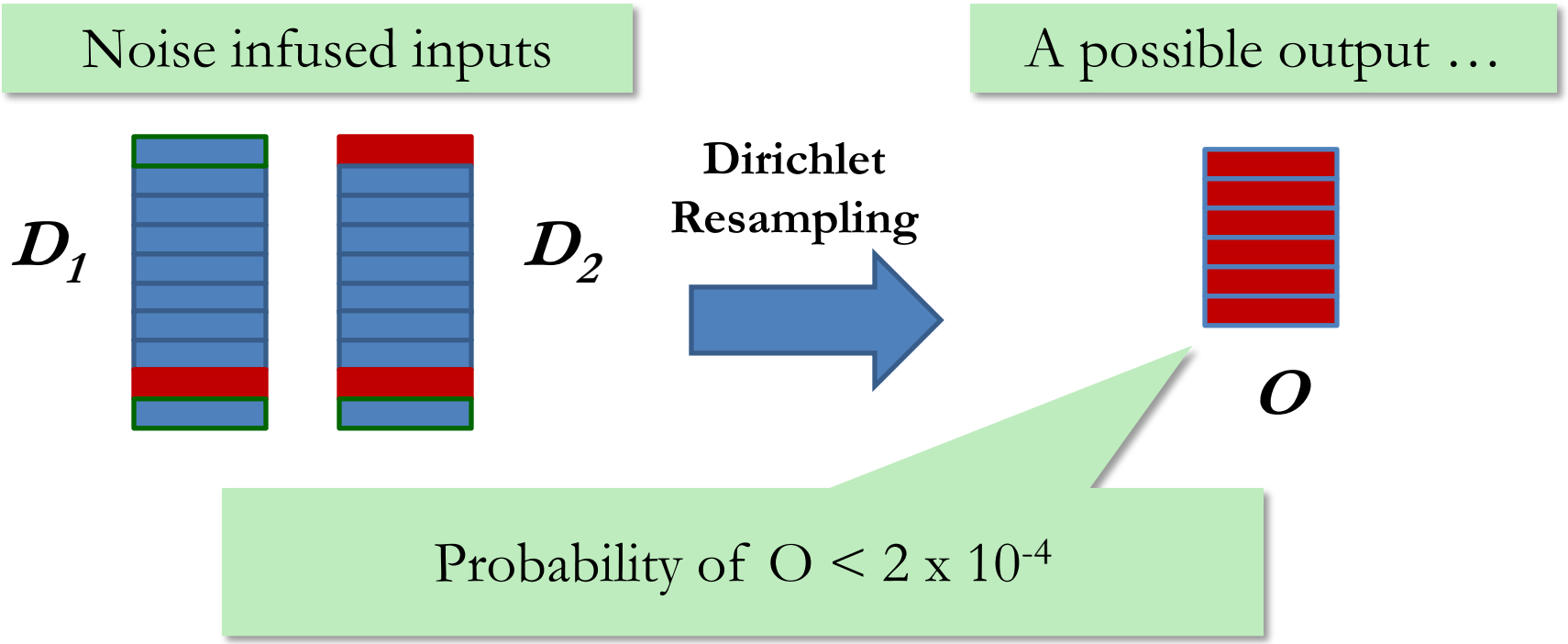
A possible output …



Adversary infers that it is very likely individual 1 is red …
… unless noise added is very large.

blue and red are two different origin blocks.

But, breach occurs with very low probability.

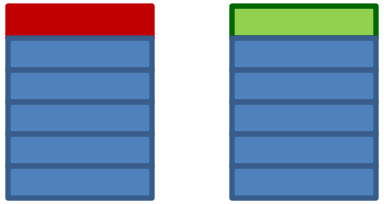


blue and red are two different origin blocks.

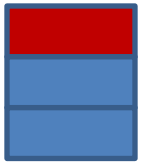
Probabilistic Differential Privacy

For every pair of inputs that differ in one value

For every ***probable*** output



D_1 D_2



O

Let Safe(D_1, ϵ) be (informally) the set of outputs where the differential privacy guarantee holds for D_1 and its neighbors. Probabilistic Differential Privacy requires that for all D_1,

$$Pr[\mathcal{A}(D_1) \in Safe(D_1, \epsilon)] > 1 - \delta$$

How much noise is added per block?

- Noise required per block

lesser privacy →

Privacy (e^{ε} =)	5	10	20	50
Noise	$25x10^4$	$11x10^4$	$5x10^4$	$2x10^4$ ← **Differential Privacy**
Noise	17.5	5.5	2.16	0.74 ← **Probabilistic Differential Privacy ($\delta = 10^{-5}$)**

1 million original and synthetic workers.

Where to add noise?

- To every block?
  - There are about **8 million** blocks on the map!
  - Total noise added is about **6 million** (even with ln
- Only to blocks that appear in the data?
  - Breach privacy!
- Ignoring outliers?
  - Degrades utility
  - Outliers contribute to about half the workers.

Final Solution

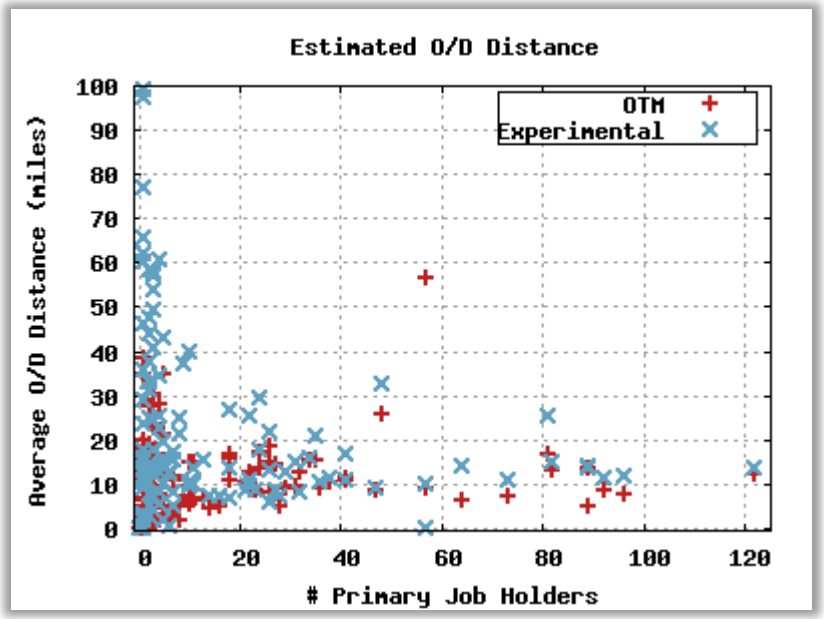
- Step 1 : Coarsen the domain
  - Based on an existing public dataset (Census Transportation Planning Package, CTPP).

- Step 2: Probabilistically drop blocks with 0 support
  - Pick a function *f* (based on external data)
  - For every block with support 0, *ignore with probability f(b)*

Step 2 incurs some additional privacy loss that can be formally quantified (details omitted).

Evaluation

- Utility measured by average commute distance for each destination block.



Experimental Setup:

- **OTM**: Currently published OnTheMap data used as original data.
- All destinations in Minnesota.
- 120, 690 origins per destination.
  - chosen by pruning out blocks that are > 100 miles from the destination.
- $\varepsilon = 4.3$, $\delta = 10^{-5}$
- Additional leakage due to probabilistic pruning = 4 *(min f(b) = 0.0378)*

Module 4: Applying Differential Privacy

- Real world deployments of differential privacy
  - OnTheMap 

  - RAPPOR  chrome
- Attacks on differential privacy implementations
  - Side channel attacks
  - Floating point attacks

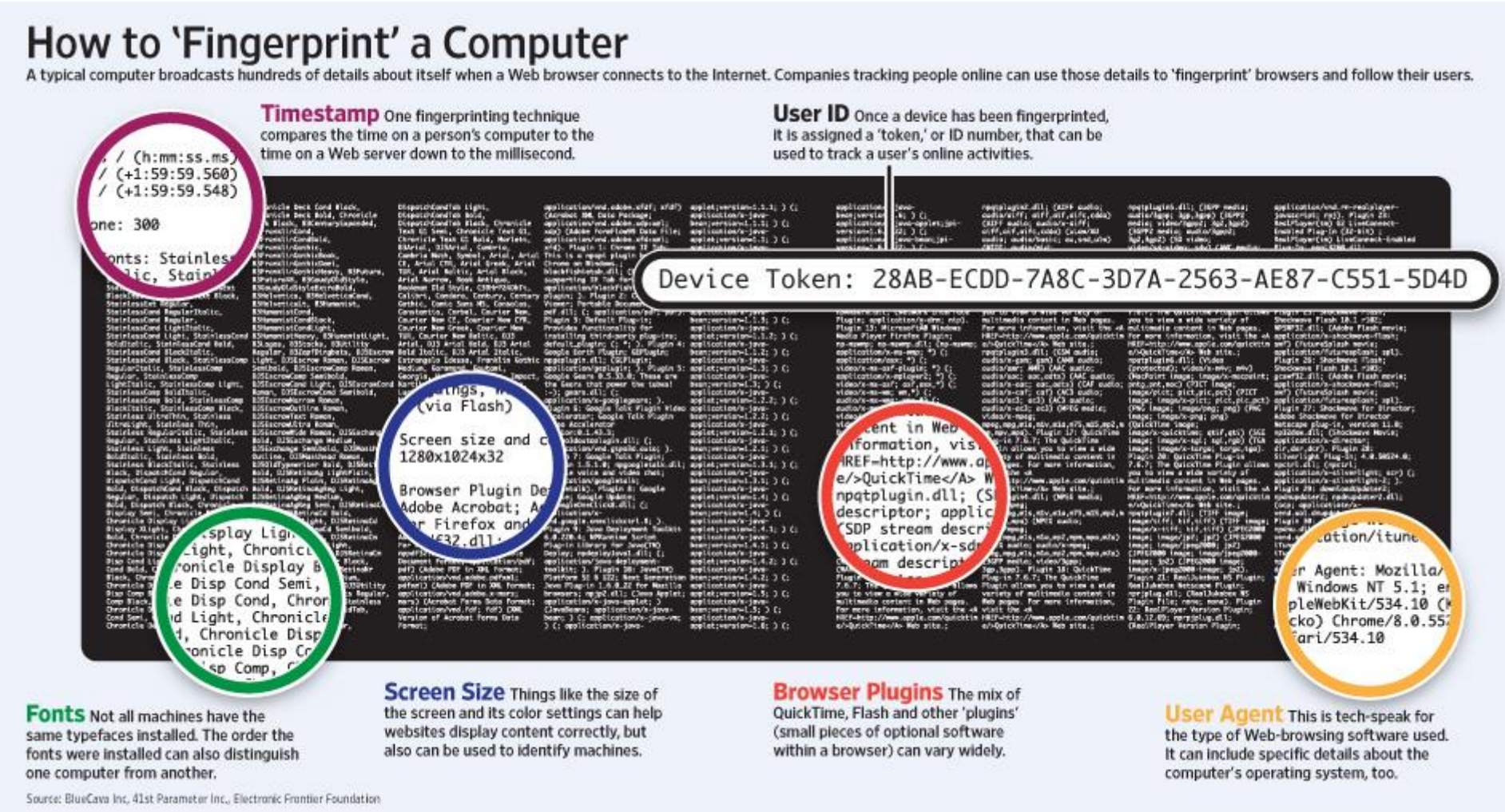
A dilemma 

- Cloud services want to protect their users, clients and the service itself from abuse.

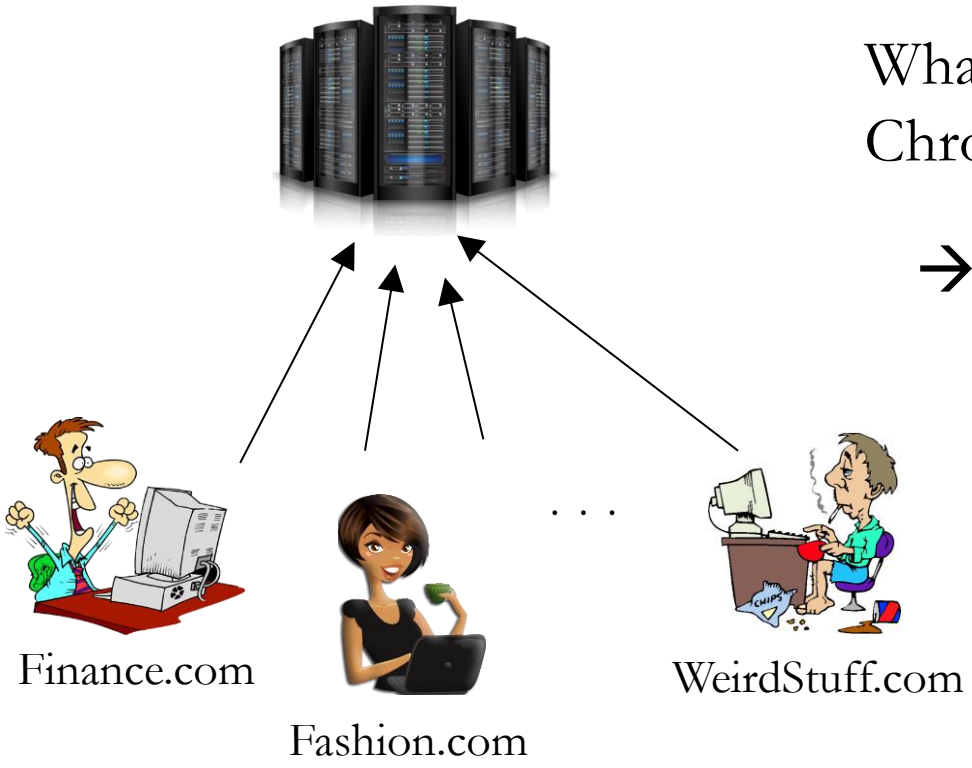
- Need to monitor statistics of, for instance, browser configurations.
  - Did a large number of users have their home page redirected to a malicious page in the last few hours?

- But users do not want to give up their data

Browser configurations can identify users



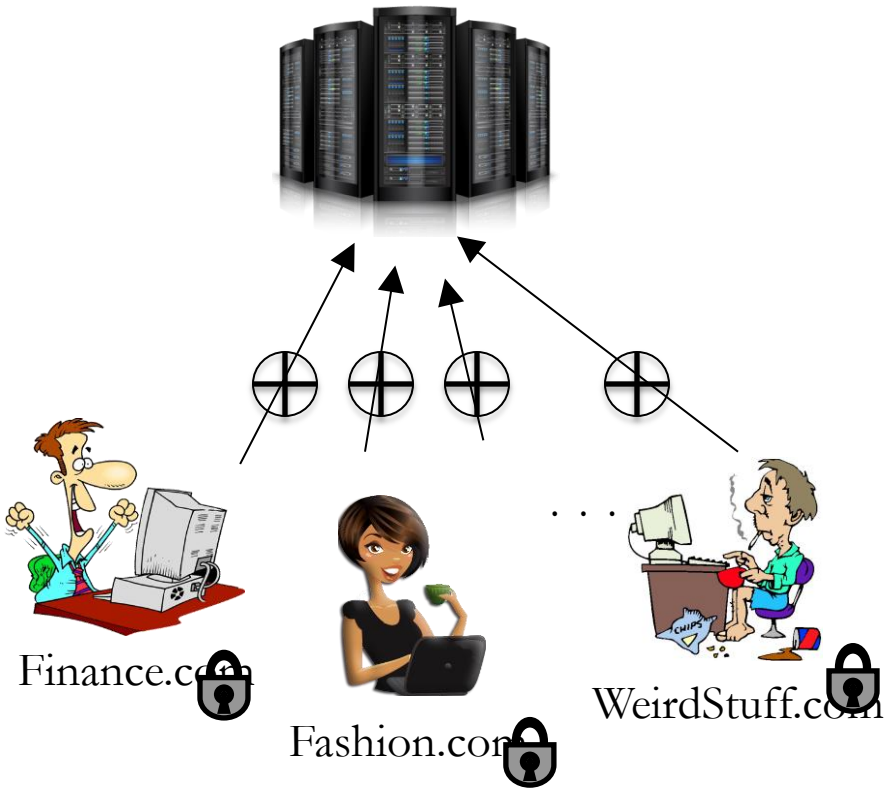
Problem [Erlingsson et al CCS'14]



What are the frequent unexpected Chrome homepage domains?

→ To learn malicious software that change Chrome setting without users' consent

Why privacy is needed?



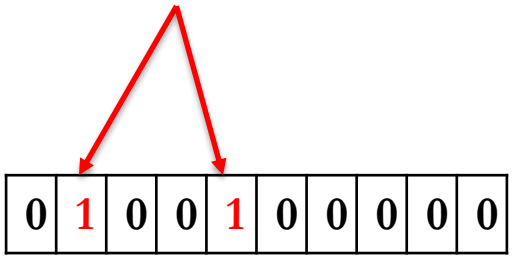
Liability (for server)

Storing unperturbed sensitive data makes server accountable (breaches, subpoenas, privacy policy violations)

Client Input Perturbation

- Step 1: Compression: use ***h*** hash functions to hash input string to ***k***-bit vector (Bloom Filter)

Finance.com



Bloom Filter *B*

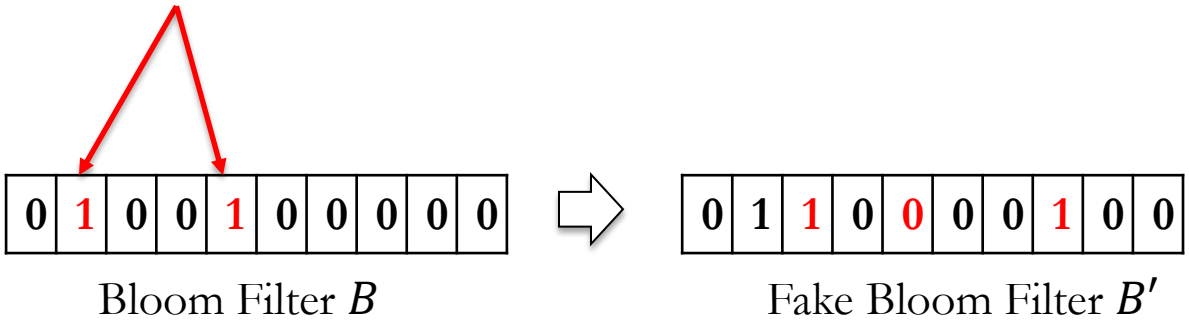
Why Bloom filter step?

Simple randomized response does not scale to large domains (such as the set of all home page URLs)

Permanent RR

- Step 2: Permanent randomized response $B \rightarrow B'$
  - Flip each bit with probability f/2
  - ***B'*** is memorized and will be used for all future reports

Finance.com



Instantaneous RR

- Step 4: Instantaneous randomized response $B' \rightarrow S$
  - Flip bit value 1 with probability 1-q
  - Flip bit value 0 with probability 1-p

Finance.com

0	1	0	0	1	0	0	0	0	0

Bloom Filter B

0	1	1	0	0	0	0	1	0	0

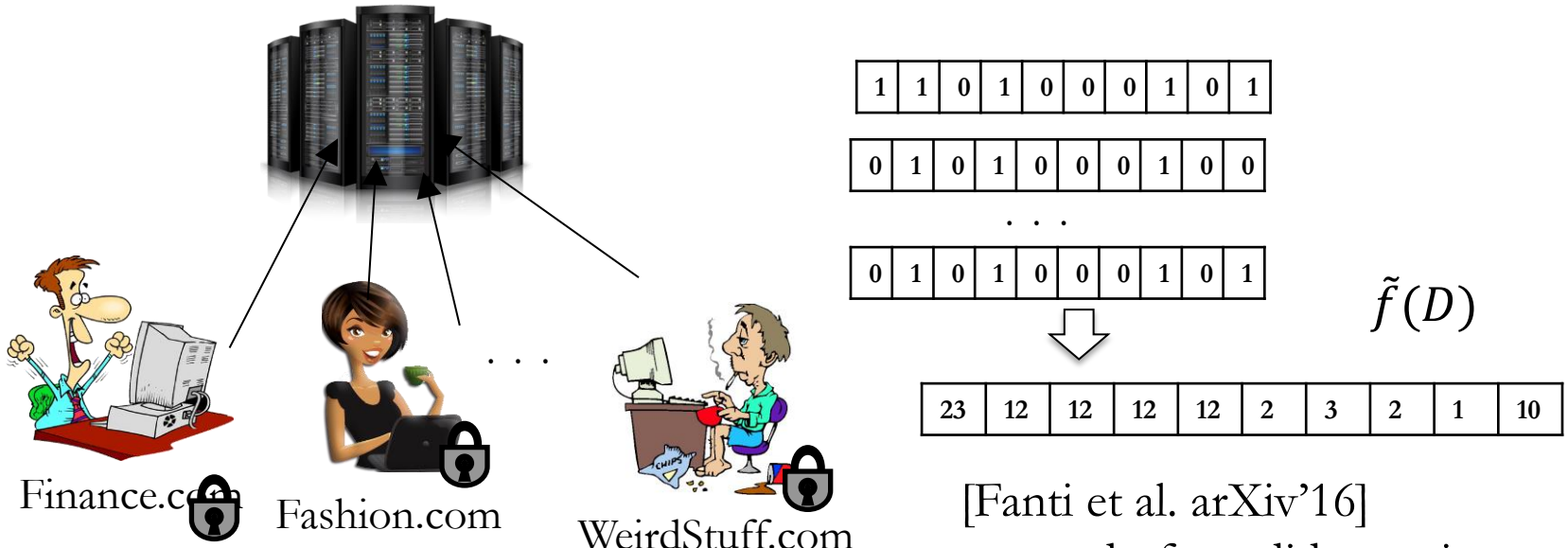
Fake Bloom Filter B'

1	1	0	1	0	0	0	1	0	1

Report sent to server S

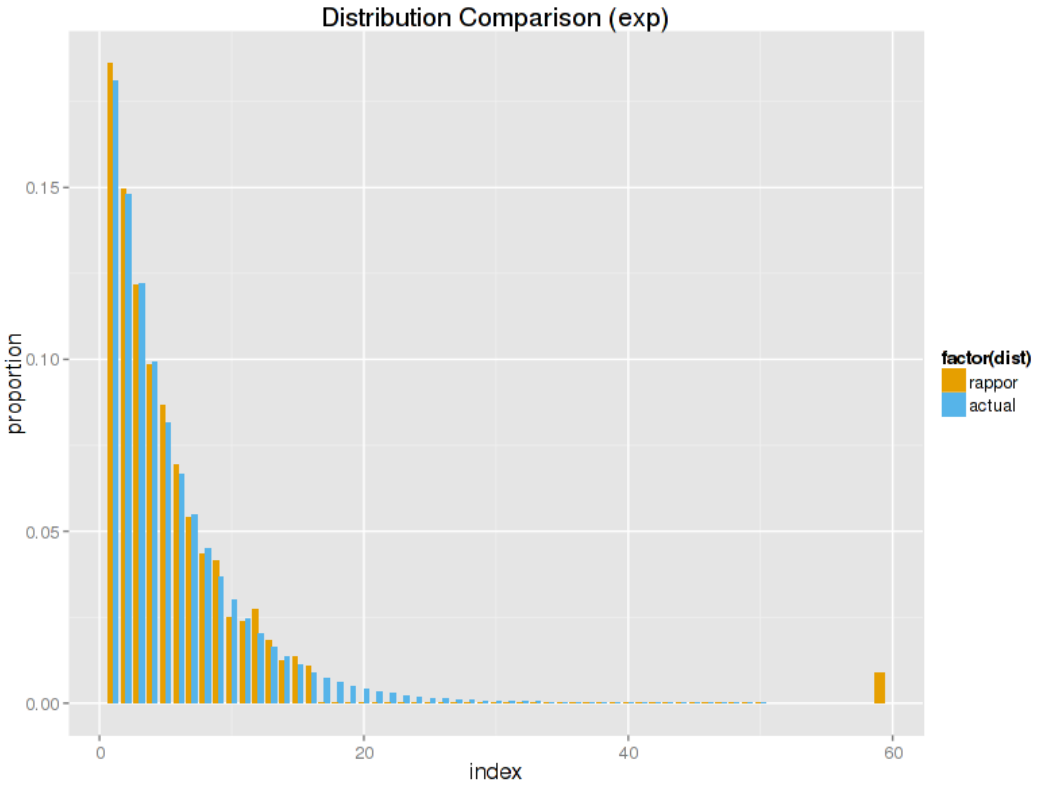
Server Report Decoding

- Step 5: estimates bit frequency from reports $\tilde{f}(D)$
- Step 6: estimate frequency of candidate strings with regression from $\tilde{f}(D)$



RAPPOR Demo

http://google.github.io/rappor/examples/report.html



Simulation Input

Number of clients	100,000
Total values reported / obfuscated	700,000
Unique values reported / obfuscated	50

RAPPOR Parameters

k	Size of Bloom filter in bits	16
h	Hash functions in Bloom filter	2
m	Number of Cohorts	64
p	Probability p	0.5
q	Probability q	0.75
f	Probability f	0.5

Parameter Selection (Exercise)

- Recall RR for *a single bit* (Module 2)
  - RR satisfies ε-DP if reporting flipped value with probability $1-p$, where $\frac{1}{1+e^{\varepsilon}} \leq p \leq \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$

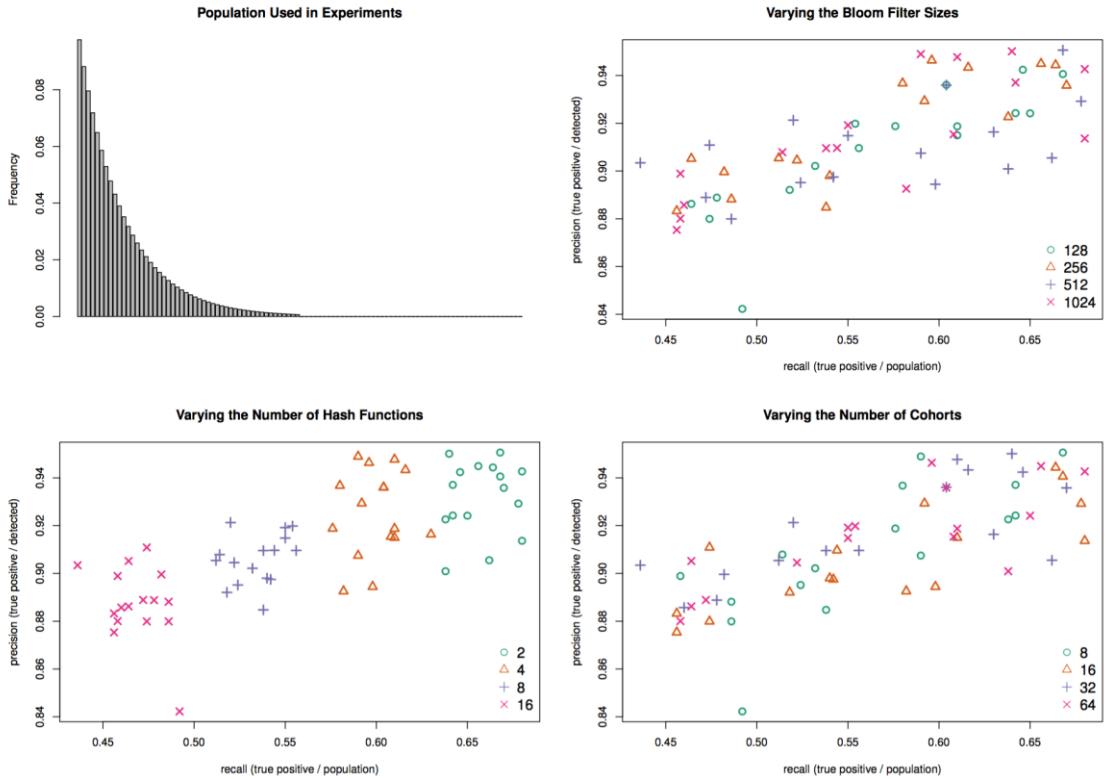
- Question 1: if Permanent RR flips each bit in the k-bit bloom filter with probability 1-p, which parameter affects the final privacy budget?
  1. # of hash functions: h
  2. bit vector size: k
  3. Both 1 and 2
  4. None of the above

Parameter Selection (Exercise)

- Answer: # of hash functions: h
  - Remove a client's input, the maximum changes to the true bit frequency is h.
  - Permanent RR satisfies $(h\varepsilon)$-DP.

Parameter Selection (Exercise)

- In addition, h affects the utility most compared to other parameters



Other Real World Employments

- Differentially private password Frequency lists [Blocki et al. NDSS '16]
  - release a corpus of 50 password frequency lists representing approximately 70 million Yahoo! users
  - varies from 8 to 0.002
- Human Mobility [Mir et al. Big Data '13]
  - synthetic data to estimate commute patterns from call detail records collected by AT&T
  - 1 billion records ~ 250,000 phones
- Apple will use DP [Greenberg. Wired Magazine '16]
  - in iOS 10 to collect data to improve QuickType and emoji suggestions, Spotlight deep link suggestions, and Lookup Hints in Notes
  - in macOS Sierra to improve autocorrect suggestions and Lookup Hints

Outline of Module 4

- Real world employments of differential privacy
  - OnTheMap 
  - RAPPOR 

- Attacks on differential privacy implementations
  - Side channel attacks
  - Floating point attacks

Covert Channel [Haeberlin et al SEC '11]

- Key assumption in differential privacy implementations:
  - **The querier can *only* observe the result of the query, and nothing else.**
  - This answer is guaranteed to be differentially private.

- In practice: The querier can observe other effects.
  - E.g, Time taken by the query to complete, power consumption, etc.
  - Suppose a system takes 1 minute to answer a query if Bob has cancer and 1 micro second otherwise, then based on query time the adversary may know that Bob has cancer.

Threat Model

- Assume the adversary (querier) does not have physical access to the machine.
  - Poses queries over a network connection.

- Given a query, the adversary can observe:
  - Time that the response arrives at their end of the connection
  - Answer to their question
  - The system's decision to execute the query or deny (since the new query would exceed the privacy budget)

Timing Attack

```
Function is_f(Record r){
    if(r.name = Bob && r. disease = Cancer)
            sleep(10 sec);
    return f(r);
}

Function countf(){
    var  fs = from record in data
            where (is_f(record))
    print fs.NoisyCount(0.1);
}
```

Timing Attack

```
Function is_f(Record r){
    if(r.name = Bob && r. disease = Cancer)
        sleep(10 sec);
    return f(r);
}
```

If Bob has Cancer, then the query takes **> 10 seconds**;

If Bob does not have Cancer, then query takes **less than a second**.

Global Variable Attack

```
Boolean found = false;
Function f(Record r){
    if(found)  return 1;
    if(r.name = Bob && r.disease = Cancer){
            found = true; return 1;
    } else return 0;
}

Function countf(){
    var  fs = from record in data
            where (f(record))
    print fs.NoisyCount(0.1);
}
```

Global Variable Attack

```
Boolean found = false;
Function f(Record r){
    if(found)  return 1;
    if(r.name = Bob && r.disease = Cancer){
        found = true; return 1;
    } else return 0;
```

Typically, the Where transformation does not change the sensitivity of the aggregate (each record transformed into another value).

But, this transformation changes the **sensitivity** – if Bob has Cancer, then all subsequent records return 1.

Privacy Budget Attack

```
Function is_f(Record r){
    if(r.name = Bob && r.disease = Cancer){
        run a sub-query that uses a lot of the privacy budget;
    }
    return f(r);
}

Function countf(){
    var  fs = from record in data
        where (f(record))
    print fs.NoisyCount(0.1);
}
```

Privacy Budget Attack

```
Function is_f(Record r){
    if(r.name = Bob && r.disease = Cancer){
        run a sub-query that uses a lot of the privacy budget;
    }
    return f(r);
}
```

If Bob does not has Cancer, then privacy budget decreases by **0.1**.
If Bob has Cancer, then privacy budget decreases by **0.1 + Δ**.

Even if adversary can't query for the budget, he can detect the change in budget by counting how many more queries are allowed.

Avoiding Side Channel Attacks

- Existing frameworks: Fuzz & GUPT

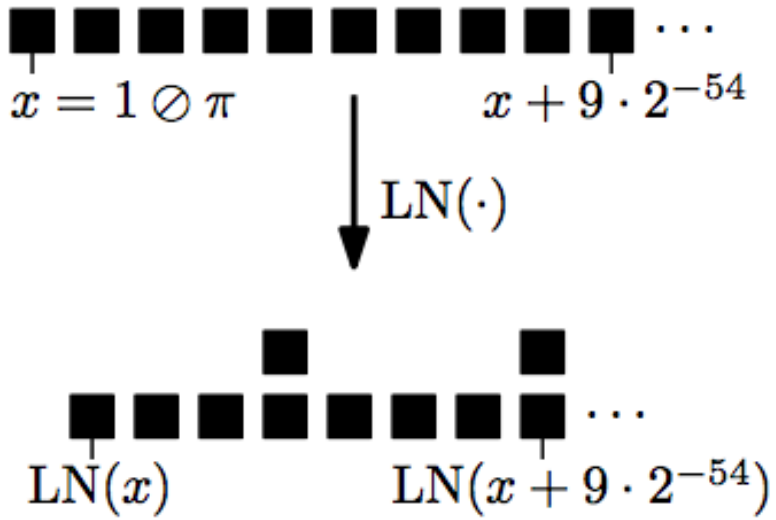
  [Haeberlin et al SEC '11] [Mohan et al SIGMOD '12]

- Timing attack
  - Ensuring each query takes a bounded time on all records or blocks
- Global variable attack
  - Global static variables are either inaccessible to adversary or disallowed
- Privacy budget attack
  - Sensitivity is statically estimated (rather than dynamically)

Least significant bits and Laplace Mechanism

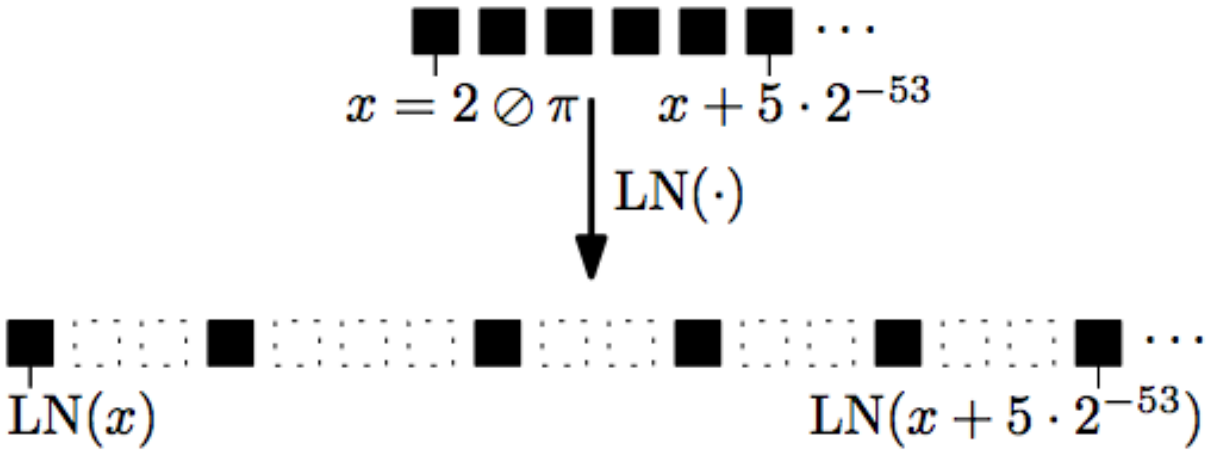
[Mironov CCS '12]

- Suppose Laplace mechanism is implemented using standard floating point,
- Certain outputs are more likely than others



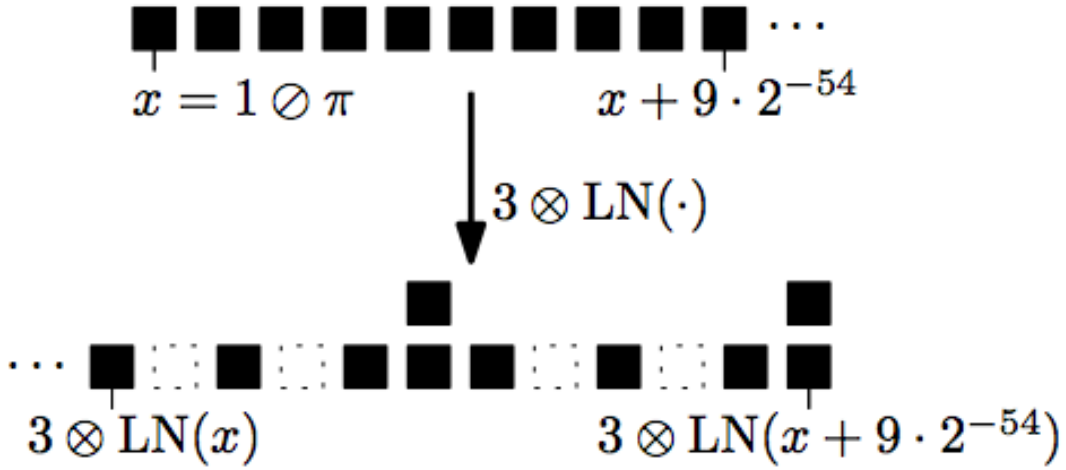
Least significant bits and Laplace Mechanism

- Suppose Laplace mechanism is implemented using standard floating point,
- Certain outputs may not appear



Least significant bits and Laplace Mechanism

- Suppose Laplace mechanism is implemented using standard floating point,
- Both can happen simultaneously



Least significant bits and Laplace Mechanism

- Sensitivity computation under floating point is also tricky (assume left to right summation in the following example):

$$n = 2^{30} + 1$$

$$x_1 = 2^{30}, x = -2^{-23}, \ldots, x_n = -2^{-23}$$

$$f(x_1, \ldots, xn) = \sum_{i=1}^{n} x_i = 2^{30} - 2^{30}\, 2^{-23} = 2^{30} - 128$$

$$f(x_1 + 1, \ldots, xn) = x_1 + 1 = 2^{30} + 1$$

References

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber, "*Privacy: From Theory to Practice on the Map",* ICDE 2008

Ú. Erlingsson, V. Pihur, A. Korolova, "*RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*", CCS 2014

G. Fanti, V. Pihur, Ú. Erlingsson, "*Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries*", arXiv:1503.01214

J. Blocki, A. Datta, J. Bonneau, "*Differentially Private Password Frequency Lists Or, How to release statistics from 70 million passwords (on purpose)*", NDSS 2016

D. J. Mir ; S. Isaacman ; R. Caceres ; M. Martonosi ; R. N. Wright, "*DP-WHERE: Differentially private modeling of human mobility*", Big Data 2013

F. McSherry, "*PINQ: Privacy Integrated Queries*", SIGMOD 2009

I. Roy, S. Setty, A. Kilzer, V. Shmatikov, E. Witchel, "*Airavat: Security and Privacy for MapReduce*", NDSS 2010

A. Haeberlin, B. Pierce, A. Narayan, "*Differential Privacy Under Fire*", SEC 2011

J. Reed, B. Pierce, M. Gaboardi, "*Distance makes types grow stronger: A calculus for differential privacy*", ICFP 2010

P. Mohan, A. Thakurta, E. Shi, D. Song, D. Culler, "*Gupt: Privacy Preserving Data Analysis Made Easy*", SIGMOD 2012

A. Smith, "*Privacy-preserving statistical estimation with optimal convergence rates*", STOC 2011

I. Mironov, "*On significance of the least significant bits for differential privacy ppt*", CCS 2012

MODULE 5: BEYOND TABULAR DATA

Module 5: Privacy beyond tabular data

- Differential Privacy for complex data
  - Neighboring databases
  - Correlations
  - No Free Lunch Theorem

- Customizing differential privacy using Pufferfish
  - Semantic privacy definitions
  - Equivalence to DP
  - Algorithm Design

Differential Privacy & Complex Datatypes

- Defining neighboring databases
  - What is a record?
- Records can be correlated
  - Unravels privacy guarantee

Neighboring Databases …

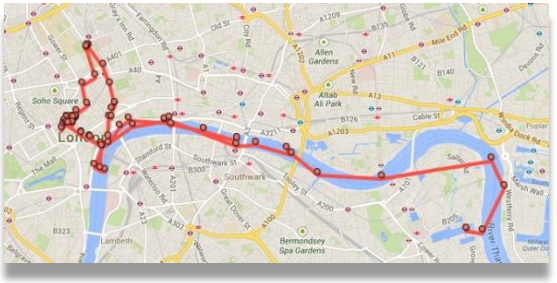
… differ in one record.

- In graphs, a records can be:
  - An edge (u,v)
  - The adjacency list of node u

Neighboring Databases …

… differ in one record.

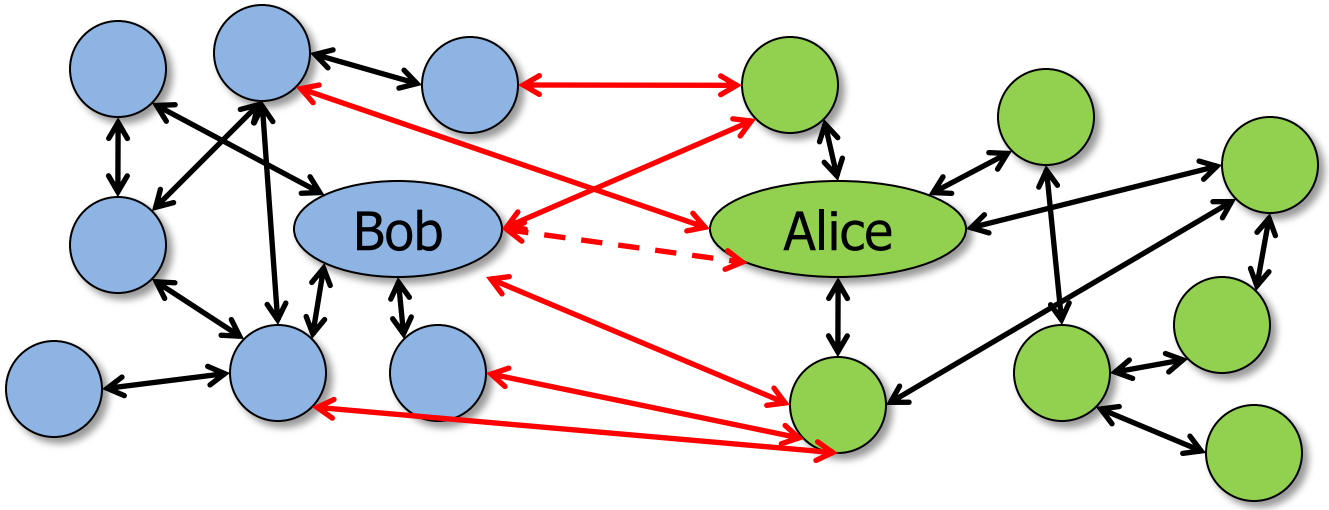
- In location trajectories, a record can be:
  - Each location in the trajectory
  - A sequence of locations spanning a window of time
  - The entire trajectory



Neighboring databases …

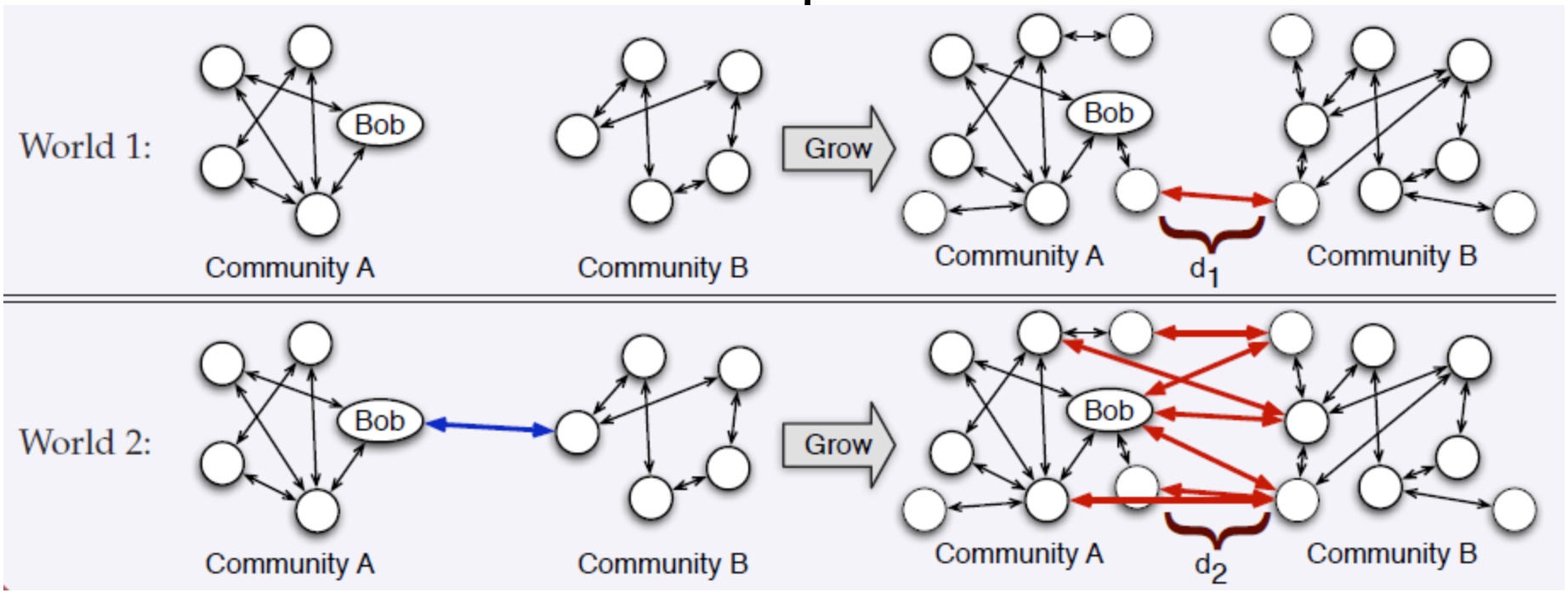
… how do they affect the privacy that is guaranteed?

Correlations and DP



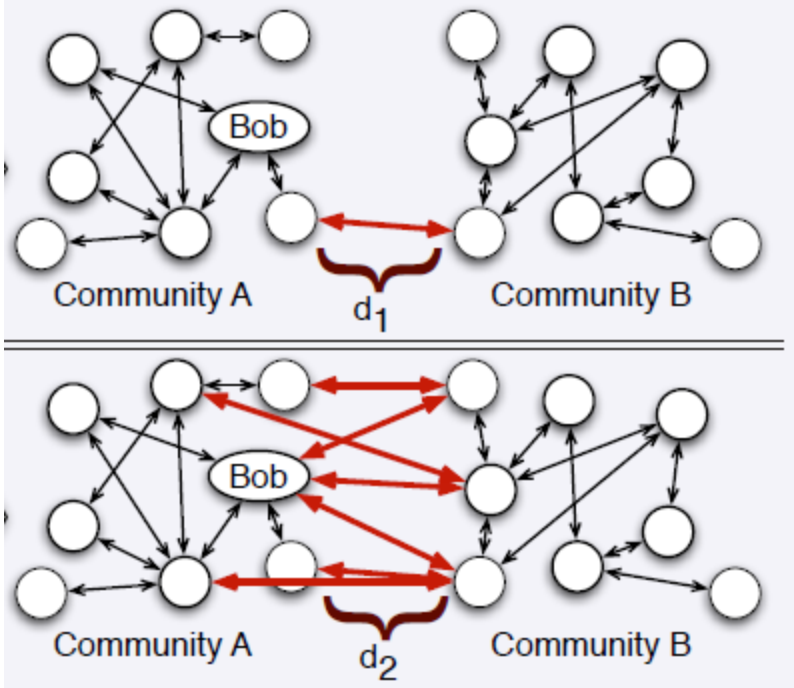
- Want to release the number of edges between **blue** and **green** communities.
- Should not disclose the presence/absence of Bob-Alice edge.

Adversary knows how social networks



- Depending on the social network evolution model, $(d_2\text{-}d_1)$ is *linear* or even *super-linear* in the size of the network.

Differential privacy fails to avoid breach



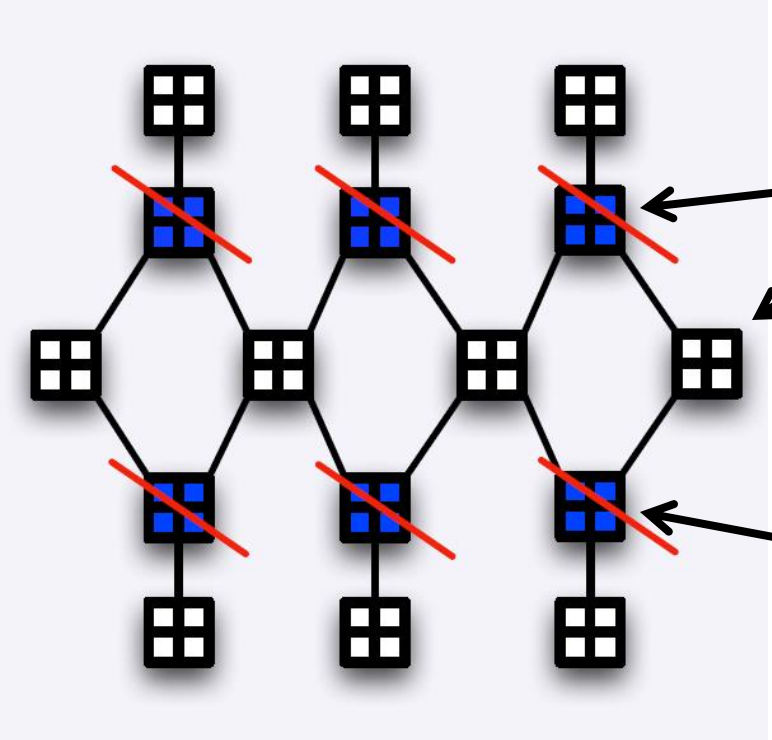
Output $(d_1 + \delta)$

δ ~ Laplace($1/\varepsilon$)

Output $(d_2 + \delta)$

Adversary can distinguish between the two worlds if $d_2 - d_1$ is large.

Reason for Privacy Breach

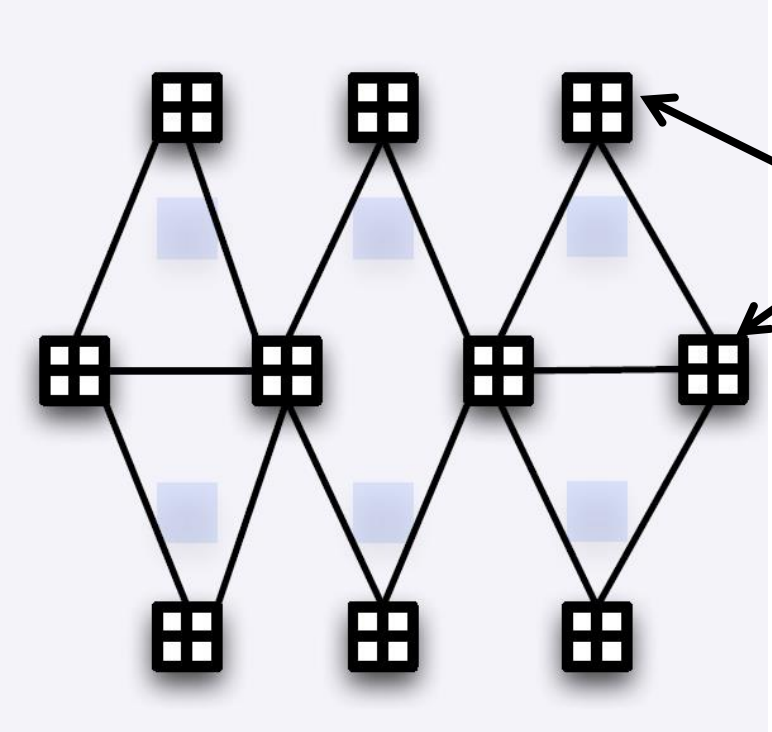


- Pairs of tables that differ in one tuple
-  cannot distinguish them

Tables that do not satisfy background knowledge

Space of all possible tables

Reason for Privacy Breach




can distinguish between every pair of these tables based on the output

Space of all possible tables

No Free Lunch Theorem

It is not possible to guarantee *any* utility in addition to privacy, *without making assumptions about*

- the data generating distribution

[Kifer-Machanavajjhala SIGMOD '11]

- the background knowledge available to an adversary

[Dwork-Naor JPC '10]

Need a formal theory to understand the privacy ensured by DP

Pufferfish

- Pufferfish (data):
  - contains tetrodotoxin (sensitive information).
- Toxin is everywhere:
  - Liver
  - Intestines
  - Skin / Muscles
- Removing all toxin = removing fish

- Chef (algorithm):
  - Processes the fish.
- Certification and license (privacy definition):
  - Rules chef must follow / restrictions on algorithm
  - Guarantees output is (relatively) safe.

- Fugu (sanitized data):
  - Tasty (high utility)
  - Minimal toxins
  - Minimal leakage of sensitive information

Pufferfish Semantics

- What is being kept secret?
- Who are the adversaries?
- How is information disclosure bounded?
  - (similar to epsilon in differential privacy)

Sensitive Information

- **Secrets**: S be a set of potentially sensitive statements
  - "individual j's record is in the data, and j has Cancer"
  - "individual j's record is not in the data"

- **Discriminative Pairs**: $S_{pairs} \subseteq S \times S$
  Mutually exclusive pairs of secrets.
  - ("Bob is in the table", "Bob is not in the table")
  - ("Bob has cancer", "Bob has diabetes")

  - Denotes an adversary's possible beliefs about a target individual.

Adversaries

- We assume a Bayesian adversary who is can be completely characterized by his/her prior information about the data
  - We do not assume computational limits

- **Data Evolution Scenarios**: set of all probability distributions that could have generated the data (… think adversary's prior).

  - *No assumptions*: All probability distributions over data instances are possible.

  - *I.I.D.*: Set of all *f* such that: $P(data = \{r_1, r_2, \ldots, r_k\}) = f(r_1) \times f(r_2) \times \ldots \times f(r_k)$

Information Disclosure

- Mechanism M satisfies ε-Pufferfish(S, Spairs, D), if

$$\forall\, w \in range(M)$$
$$\forall\, (s, s') \in S_{pairs}$$
$$\forall \theta \in D, s.t. \quad P(s|D), P(s'|D) \neq 0$$

$$P(M(\mathfrak{D}) = w|s, \theta) \leq e^{\varepsilon} P(M(\mathfrak{D}) = w|s', \theta)$$

Pufferfish Semantic Guarantee

$$e^{-\varepsilon} \leq \frac{P(s|M(\mathfrak{D}) = w, \theta)}{P(s'|M(\mathfrak{D}) = w, \theta)} \Big/ \frac{P(s|\theta)}{P(s'|\theta)} \leq e^{\varepsilon}$$

Posterior odds of s vs s'

Prior odds of s vs s'

Pufferfish & Differential Privacy

- Discriminative Pairs:
  - s_x^i : record i takes the value x
  - $s_\perp^i$: record i is not in the database
  - $S_{pairs} = \{(s_x^i, s_\perp^i) | \forall x \in dom, \forall\ record\ i\}$

- Attackers should not be able to tell whether a record is in or out of the database

Pufferfish & Differential Privacy

- Data evolution:
  - For all $\theta = [f_1, f_2, f_3, \ldots, f_k]$

$$P[Data = D|\Theta] = \prod_{r_i \in D} f_i(r_i)$$

- Adversary's prior may be any distribution that makes records **independent**

Pufferfish & Differential Privacy

- Discriminative Pairs:
  - $S_{pairs} = \{(s_x^i, s_\perp^i) | \forall x \in dom, \forall\ record\ i\}$

- Data evolution:
  - For all $\theta = [f_1, f_2, f_3, \ldots, f_k]$

$$P[Data = D|\Theta] = \prod_{r_i \in D} f_i(r_i)$$

A mechanism M satisfies differential privacy
if and only if
it satisfies Pufferfish instantiated using Spairs and {θ}

Examples: Graphs

- Neighboring graphs differ in presence/absence of one edge

- Pufferfish meaning:
  - Data: matrix of bits
  - Secrets: whether or not an edge (u,v) is in the graph (bit at (u,v) is 0 or 1)
  - Data generating distributions: All graphs where **each edge e is independently present** with probability p_e.

- But …
  - Edges are not independent in real graphs

Examples: Location Traces

- Neighboring tables differ in one location (at one point of time) of an individual

- Pufferfish meaning
  - Data: a matrix of locations
  - Secrets: Whether or not individual was at some location at some point of time
  - Data Generating Distributions: All trajectories where an individual's **location at some time is independent of all other locations** …

- But …
  - Current location depends on previous locations…

Customizing Privacy

- Setup secrets and discriminative pairs based on the requirements of what must be kept secret
- Set up data generating distributions to capture correlations known to the adversary
- Pufferfish results in privacy definition that bounds the adversary's posterior and prior odds for every discriminative pair.

Challenges

- Setting up data generating distributions are tricky
  - Adversary's knowledge is unknown
- No known general algorithms known for Pufferfish definitions
  - Need to do algorithm design from scratch for each definition
  - Do not satisfy composition

Blowfish Privacy [He et al SIGMOD 14]

- Special case of Pufferfish that satisfies sequential composition

- A framework for redefining neighboring databases for complex datatypes using a *policy graph*
  - Captures many neighboring definitions
  - Handles correlations induced by constraints on database
    - Prior data releases
    - Location constraints

Algorithm Design Simplified

[Haney et al VLDB 16]

- Transformational equivalence between Blowfish and differential privacy

- Answering queries under a Blowfish privacy policy is equivalent in error to answering transformed queries under differential privacy
  - **COME TO THE TALK: Tuesday Research Session 8**

Wasserstein Mechanism

[Chaudhuri et al Corr Abs 16]

- A general algorithm for handling probabilistic constraints under Pufferfish

- Can leverage sequential composition for constraints that take a special form.

Summary

- Complex datatypes require custom privacy definitions
  - No Free Lunch theorem
  - Varied notions of neighboring databases
  - Correlations can unravel privacy ensured by DP algorithms

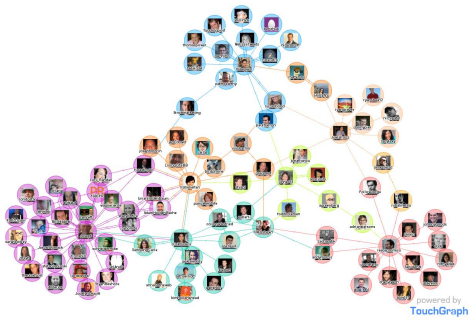
- Pufferfish is a mathematical framework for rigorous and customizable privacy definitions
  - Helps understand semantics of privacy definitions
  - Algorithms known for special classes of the framework.

MODULE 6: APPLICATIONS II

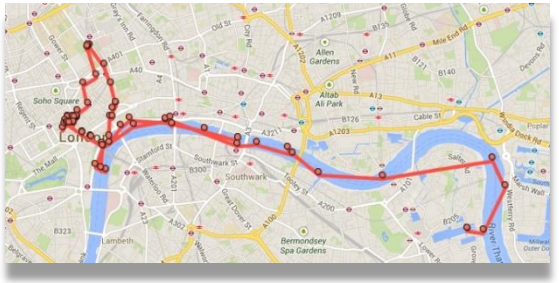
Module 6: Applications II

- Differential Privacy for Non-tabular Data

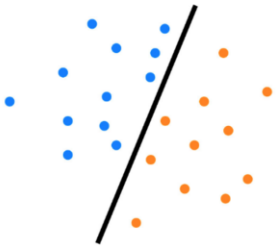
Social network



Location trajectories

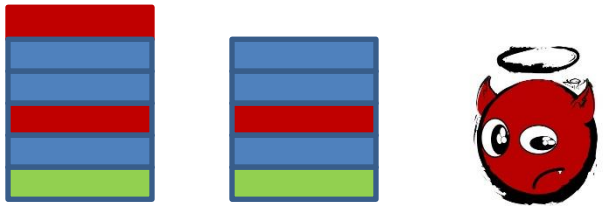


- Differential Privacy for Machine Learning



Common Themes

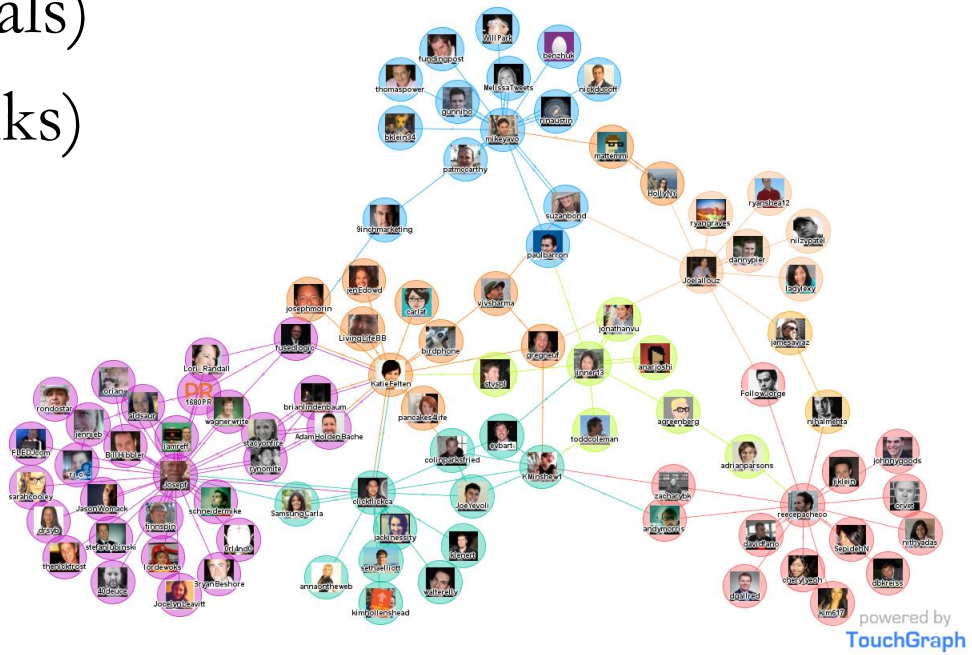
- What are secrets and neighboring datasets for different applications?



- Correlations between protected objects requires further redefinition of privacy
- New privacy definitions requires new algorithm design
- Many open questions

Social Network

- Represented using a graph $G(V, E)$
  - V: node set (individuals)
  - E: edge set (social links)



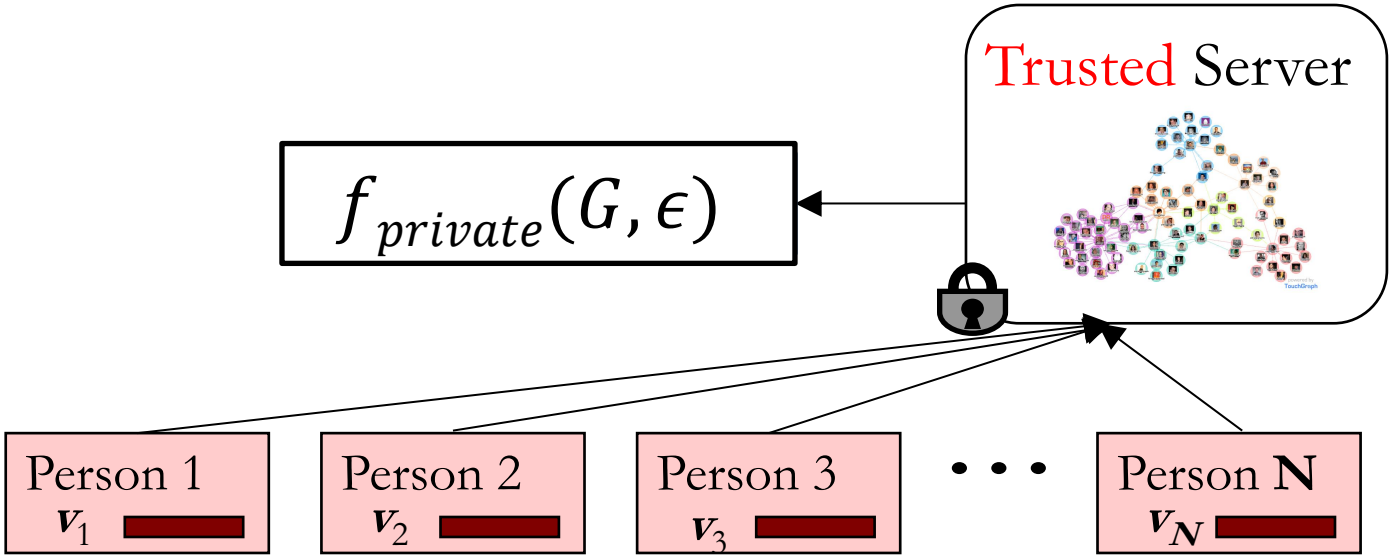
Social Network

- Attacks on graph anonymization

"it is possible for an adversary to learn whether **edges exist** or not between specific targeted pairs of nodes." [Backstrom et al WWW'07]

"a third of the **users** on both **Twitter** and **Flickr**, can be **re-identified** in the anonymous Twitter graph with only a 12% error rate." [Narayanan & Shmatikov SP'09]

Private Analysis of Social Network



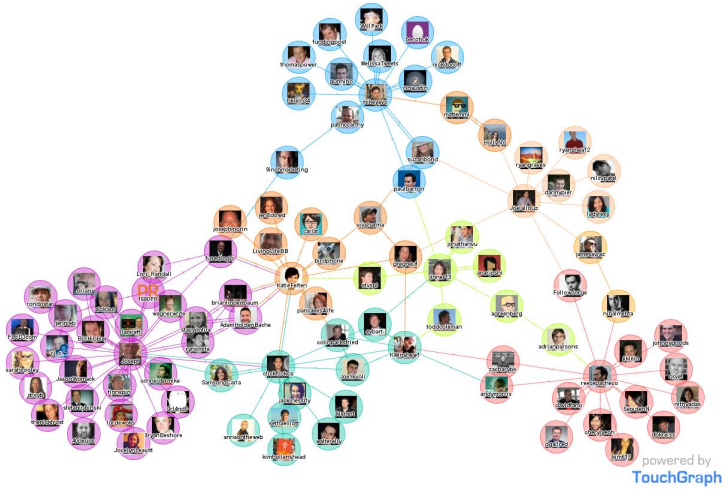
Differential Privacy: $f_{private}$ is ϵ-differentially private if for all **neighbors** G, G' and output S:

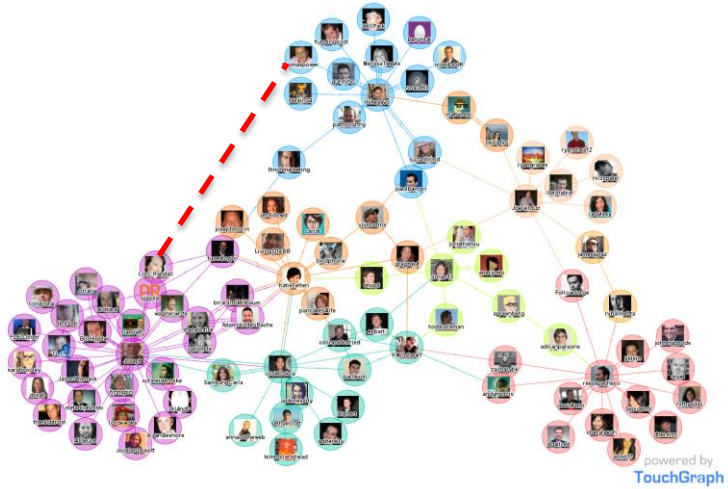
$$Pr[f_{private}(G, \epsilon) \in S] \leq e^{\epsilon} \Pr[f_{private}(G', \epsilon) \in S]$$

Variants of DP for Social Network

- Edge Differential Privacy

Secret: social links between individuals



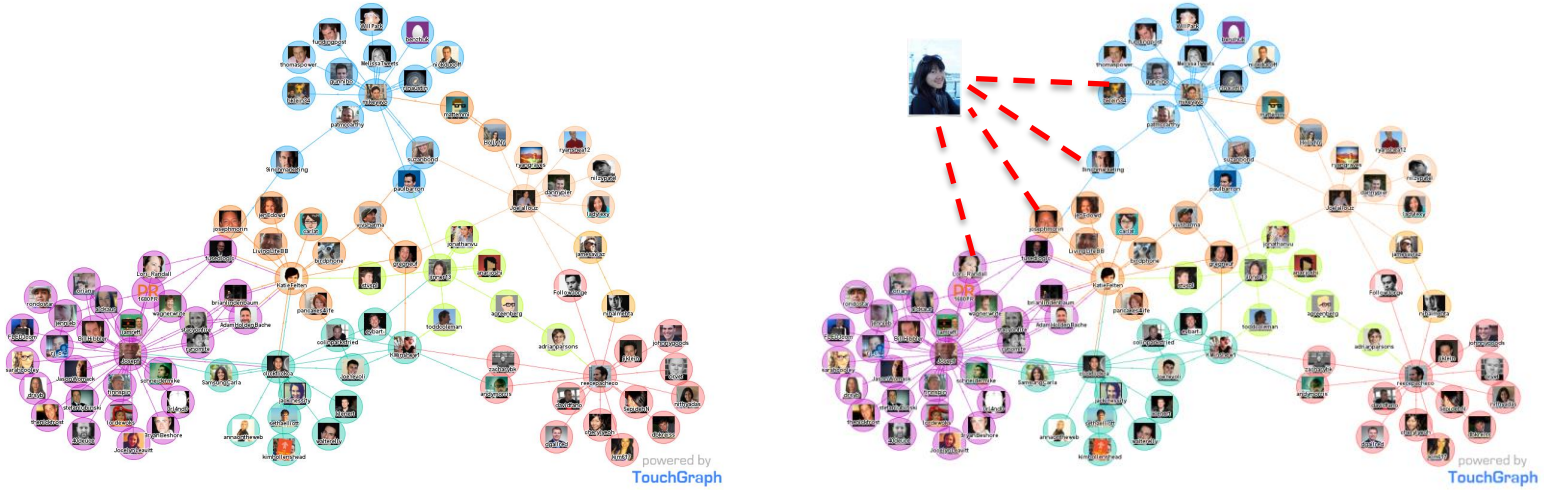


Two graphs are **neighbors** if they differ in the presence of one edge

Variants of DP for Social Network

- Node Differential Privacy

Secret: presence of an individual



Two graphs are **neighbors** if one can be obtained by another by adding or removing a node and all its edges

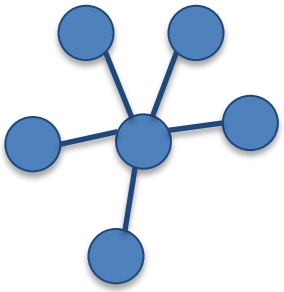
Examples for Social Network Statistics

- Degree distribution $D(G)$
- Number of edges
- Counts of small subgraphs

  e.g triangles, k-triangles, k-stars, etc.
- Cut
- Distance to nearest graph with a certain property
- Joint degree distribution

Examples for Social Network Statistics

- Degree distribution $D(G)$

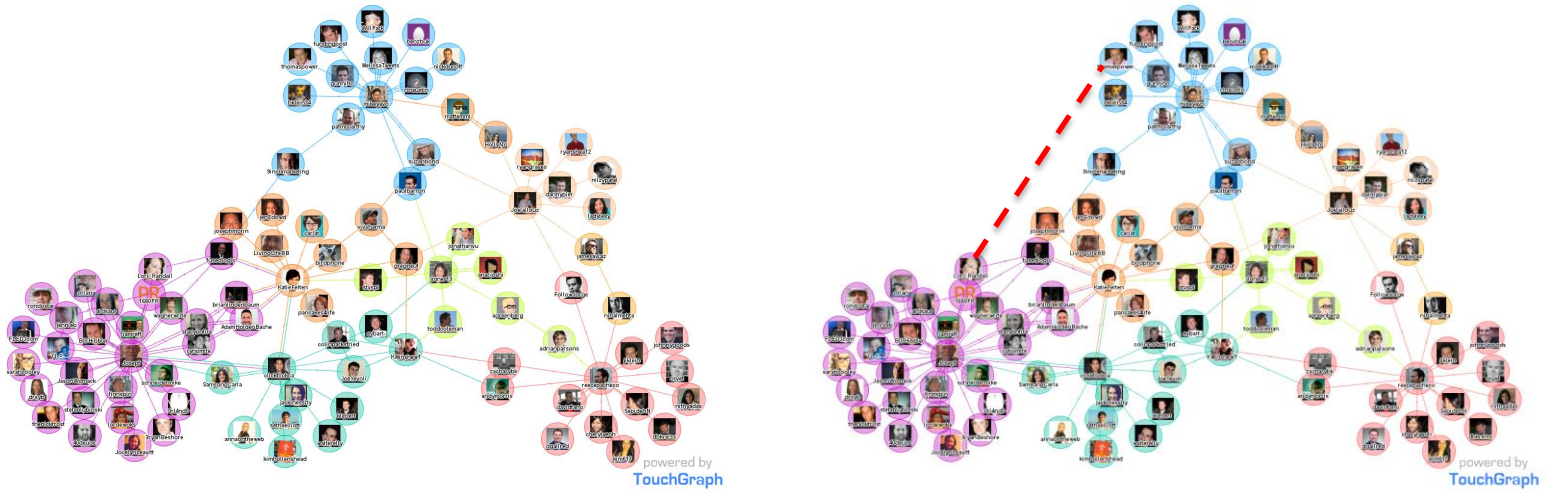


Degree	0	1	2	3	4	5
Frequency	0	5	0	0	0	1

$$D(G) = [0,5,0,0,0,1]$$

Global Sensitivity of Degree Distribution

- What is the global sensitivity of the degree distribution of $G(V, E)$, where $|V| = n$ under Edge differential privacy?



Global Sensitivity of Degree Distribution

- What is the global sensitivity of the degree distribution of $G(V, E)$, where $|V| = n$ under Edge differential privacy?

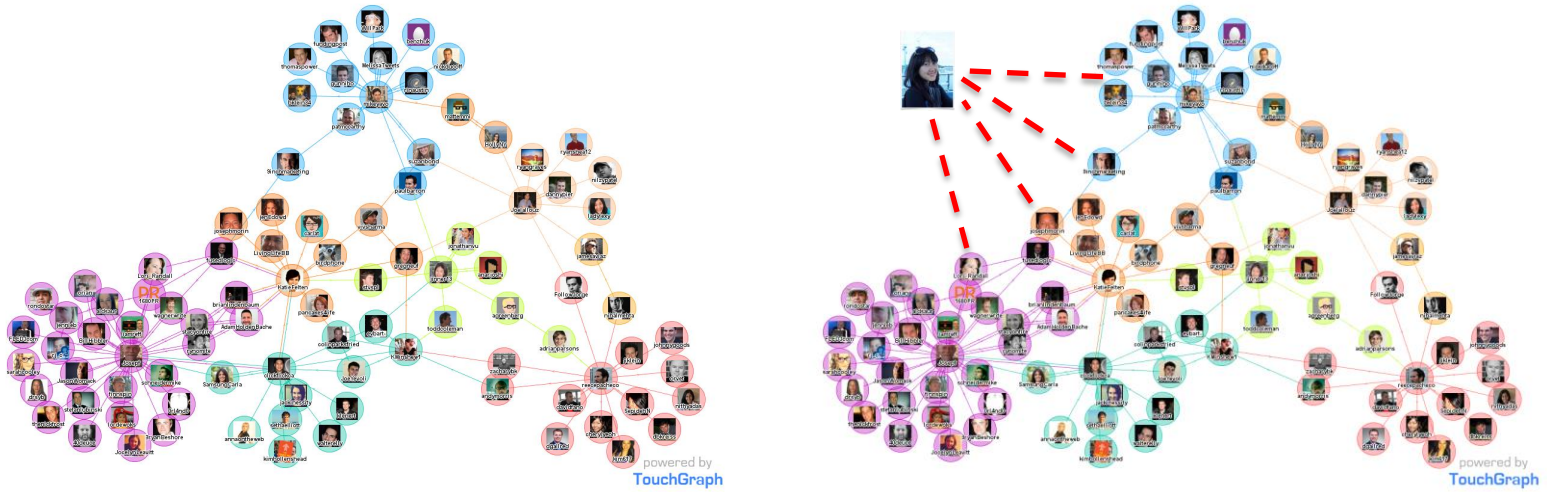
Answer: 4

Remove edge (i,j), the changes in degree frequency

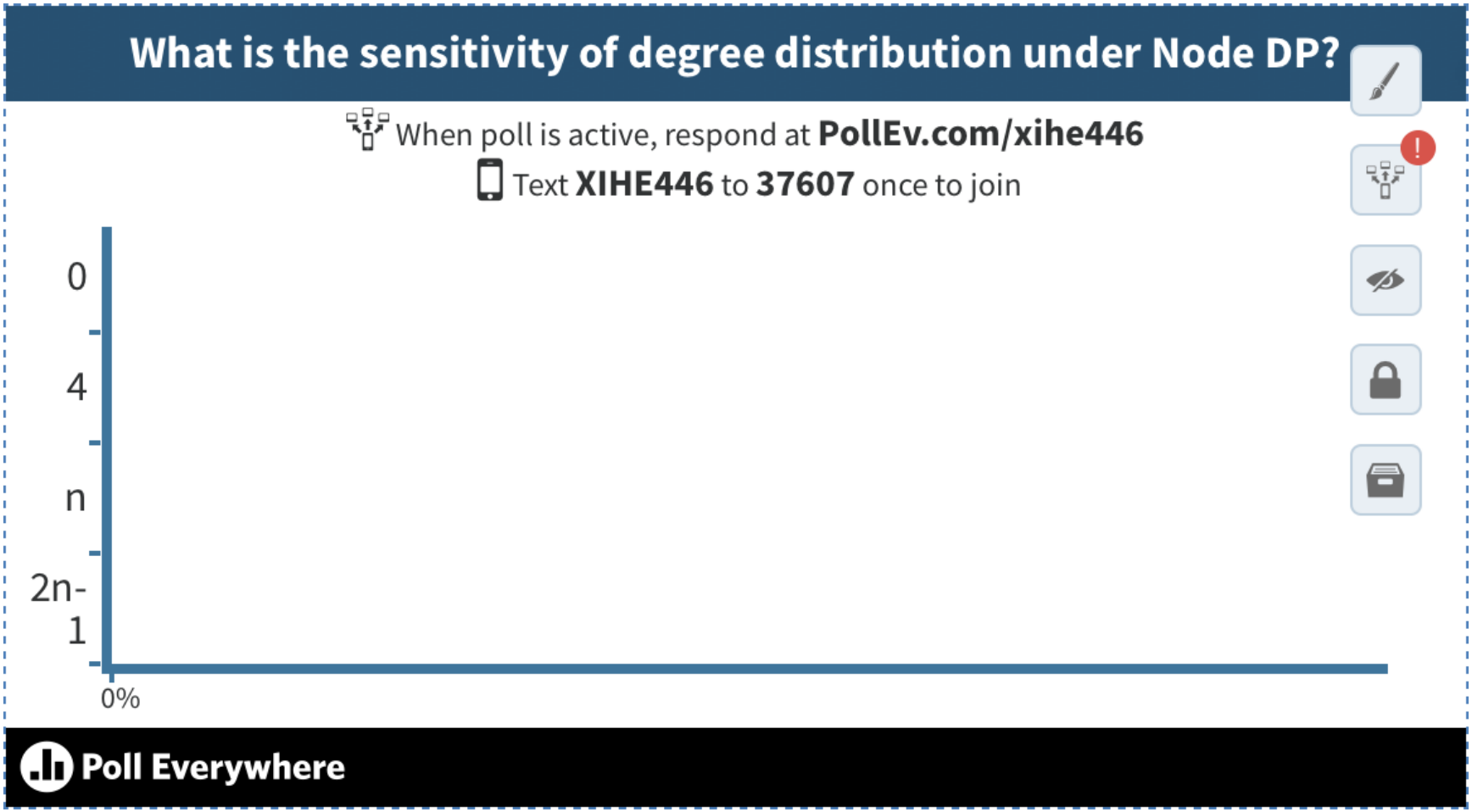
Degree	…	d_i-1	d_i	…	d_j-1	d_j	…
Frequency	…	+1	-1	…	+1	-1	…

Global Sensitivity of Degree Distribution

- What is the global sensitivity of the degree distribution of $G(V, E)$, where $|V| = n$ under Node differential privacy?



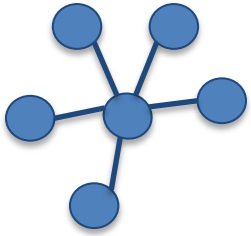
Exercise


What is the sensitivity of degree distribution under Node DP?
When poll is active, respond at PollEv.com/xihe446
Text XIHE446 to 37607 once to join
0
4
n
2n-1
0%
Poll Everywhere

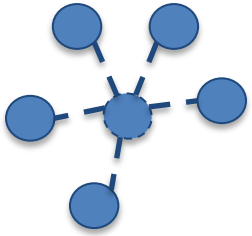
Global Sensitivity of Degree Distribution

- What is the global sensitivity of the degree distribution of $G(V, E)$, where $|V| = n$ under Node differential privacy? Answer: 2n-1

Highly Sensitive!!→ Too much noise



$D(G) = [0,5,0,0,0,1]$



$D(G') = [5,0,0,0,0,0]$

Approach to Highly Sensitive Queries

Key idea:

- Projection G on $\boldsymbol{\theta}$**-bounded graphs** G_θ
- Answer queries on G_θ instead of G

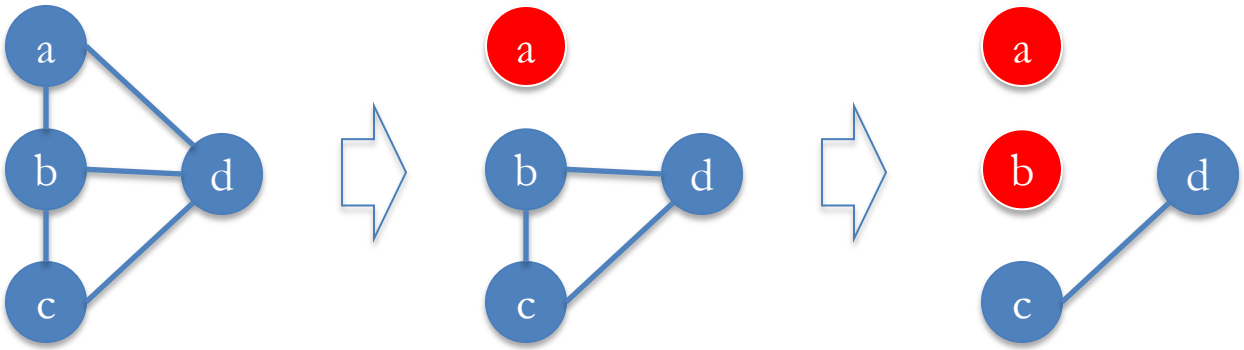
$$\widetilde{D(G)} = D(G_\theta) + noise$$

- Existing approaches for degree distribution
  - Node-based truncation [Kasiviswanathan et al TCC'13]
  - Lipschitz extensions [Raskhodnikova & Smith arXiv'15]
  - Edge-based projection [Day et al SIGMOD'16]

Node-based Truncation

[Kasiviswanathan et al TCC'13]

- $T_\theta(G)$: Remove any node with degree over θ
  - For example, $\theta = 1$



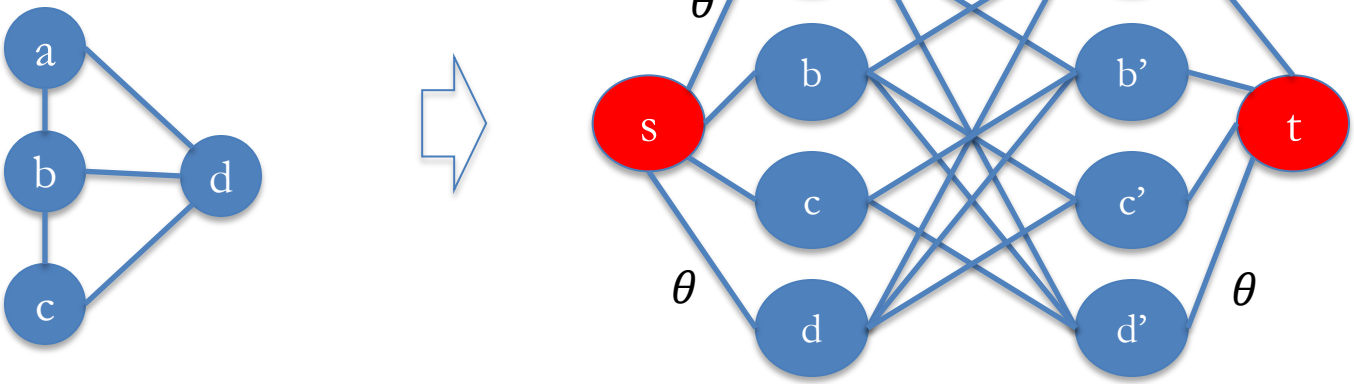
Original Graph Remove node a Remove node b

Lipschitz Extensions

[Raskhodnikova & Smith arXiv'15]

- Construct a new bipartite graph $F(G)$
  - Duplicate V, V' with source s and sink t
  - $w(s,v) = w(v,t) = \theta, w(v,u) = 1$, for $v \in V, u \in V'$

For example, $\theta = 1$



Original

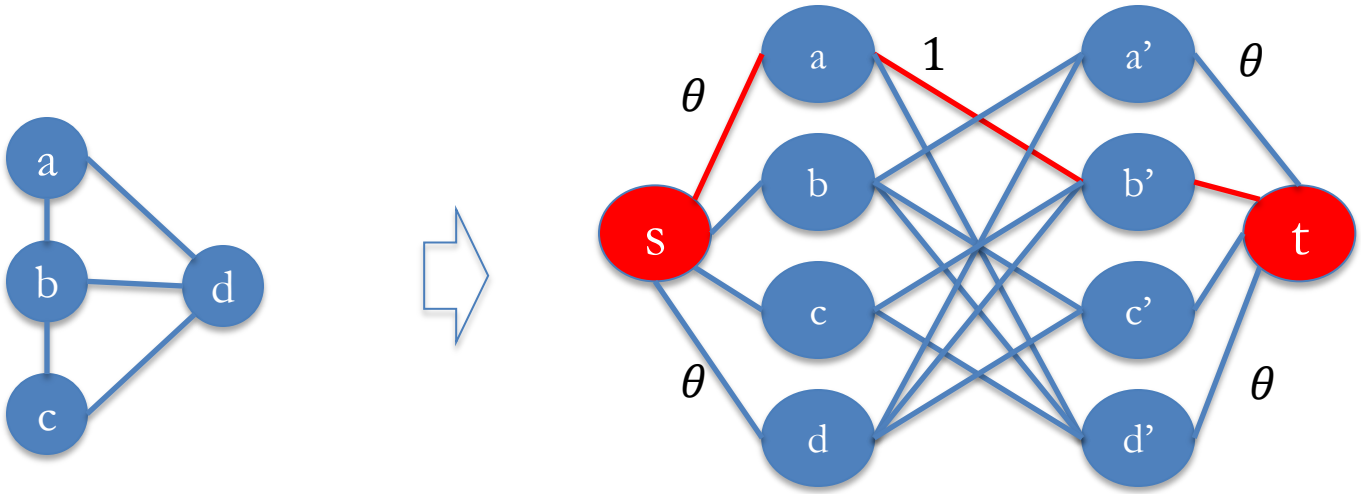
Duplicated graph with source and sink

Lipschitz Extensions

[Raskhodnikova & Smith arxiv'15]

- Construct a new bipartite graph $F(G)$
  - Run max-flow algorithm to get $F_{\theta(G)}$, while minimize

$$\sum_{v \in V} \left(\theta - f(s,v)\right)^2 + \sum_{v' \in V'} \left(\theta - f(v',t)\right)^2$$



Original

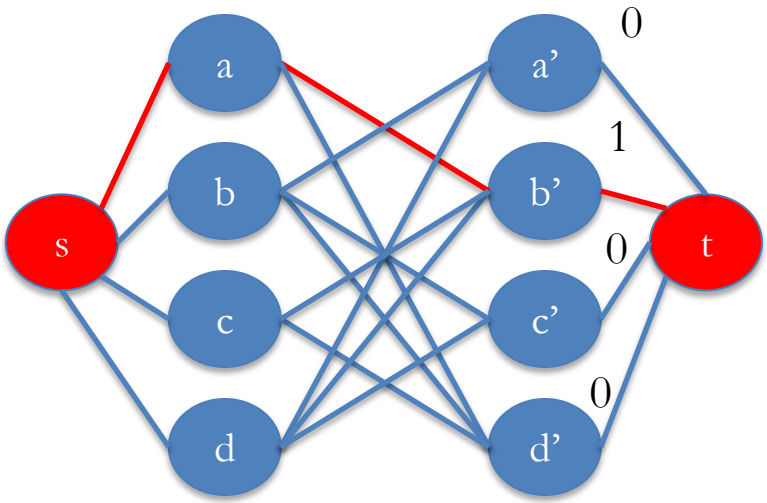
Duplicated graph with source and sink

Lipschitz Extensions

- Estimate degree list with $\hat{F}_{\theta(G)} = sort\big(f(v,t)\big)$

  E.g sorted degree list: [1,0,0,0]

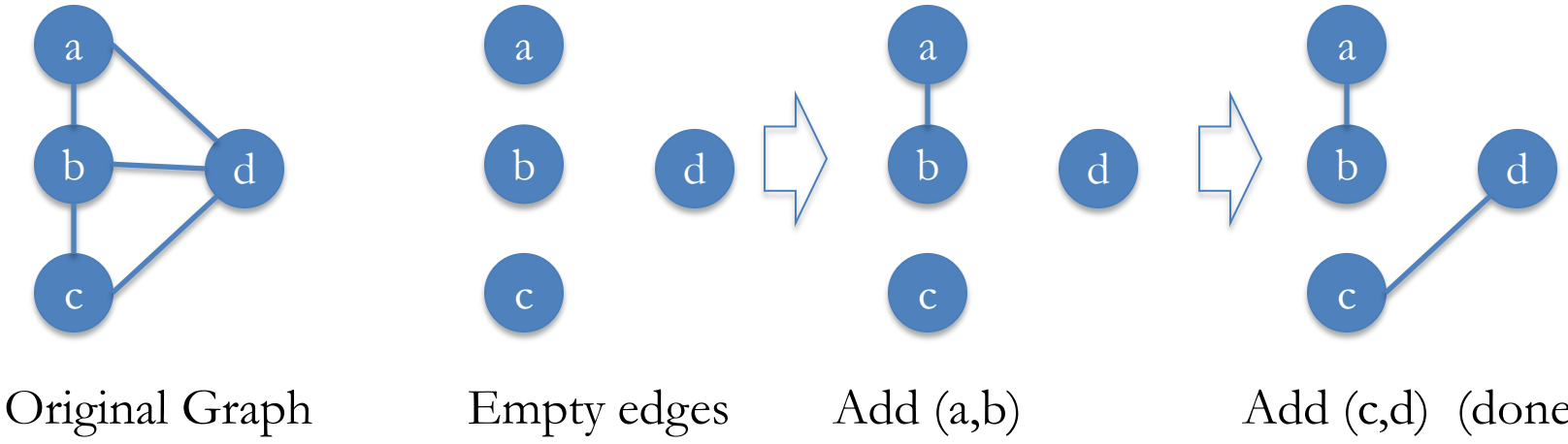
  → degree distribution: [3,1,00]



Edge-based Projection

[Day et al SIGMOD 16]

- $\pi_\theta(G)$: Add an edge (u, v) if u or v's degree dose not reach θ



How much noise?

- Answer queries on G_θ instead of G

$$\widetilde{D(G)} = D(G_\theta) + noise$$

- Sensitivity
  - Node-based truncation: $2\theta \cdot \delta$
    - Smooth sensitivity approach [Nissim et al STOC'07]
  - Lipschitz extensions: 6θ
  - Edge-based projection: $2\theta + 1$

Applicable to count
- edges
- small subgraph

[Kasiviswanathan et al TCC'13]

Work on Edge DP

Degree distribution

- Global sensitivity + Post-processing [Hay et al ICDM'09, Hay et al VLDB'10, Karwa et al Privacy in Statistical Databases'12, Kifer Lin SIGMOD'13]

Small subgraph counting

- Smooth sensitivity [Nissim et al STOC'07]
- Ladder function [Zhang et al SIGMOD'15]
- Noisy sensitivity [Behoora et al PVLDB'11, Dwork et al STOC'09]

Cut

- Random projections, global sensitivity [Blocki et al FOCS'12]
- Iterative updates [Hardt et al FOCS'10, Gupta et al TCC'12]

Releasing differentially private graph

- Exponential random graphs [Lu et al KDD'14, Karwa et al arXiv'15]

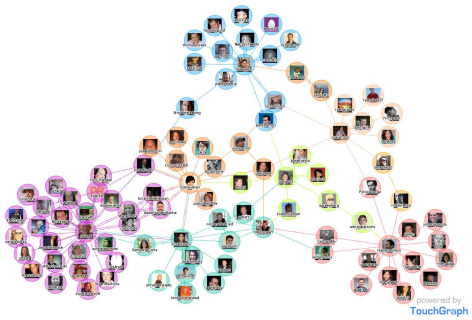
Open questions

- What other queries can be solved with Node-DP?
  - Cuts, Pairwise distances between nodes
  - Release synthetic graph?
- What are the right privacy notions for social network? Is it socially acceptable to offer weaker privacy protection to high-degree nodes?
- Social networks have node and edge attributes. What queries are useful?
- What new privacy definition is required for the correlation in the graph?

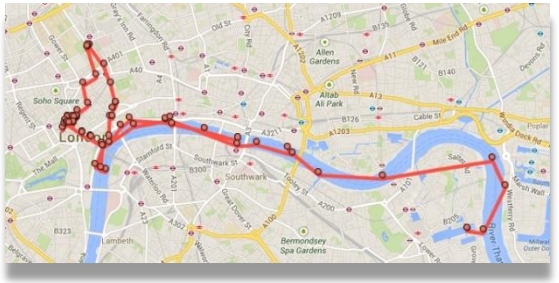
Outline of Module 6

- Differential Privacy for Non-tabular Data

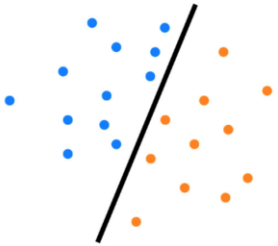
Social network



Location stream



- Differential Privacy for Machine Learning



Location Stream

High uniqueness & High predictability

Montjoye et al. Sci Rep.'13 Song et al. Science'10

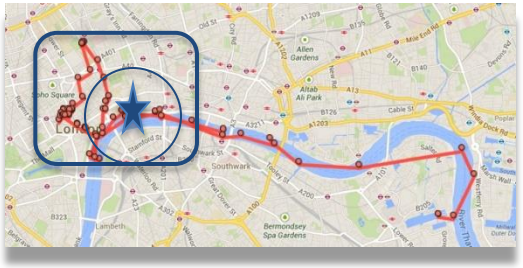
'show me how you move and  will tell you who you are'

Gambs et al. SPRINGL'11

'geosocial service "check in" dropped from 18% to 12%'

in the Pew Research Center's Internet Project, 2013

Rich Domain for Privacy Policies



What to hide?
All properties of the individual are secret e.g. Where is home location?
Properties within a small window e.g. Did user visit home in the last hour?
Properties at a specific time e.g. Did user visit home at time t?
Some properties (not all) at a specific time e.g. Did user visit near home at time ?

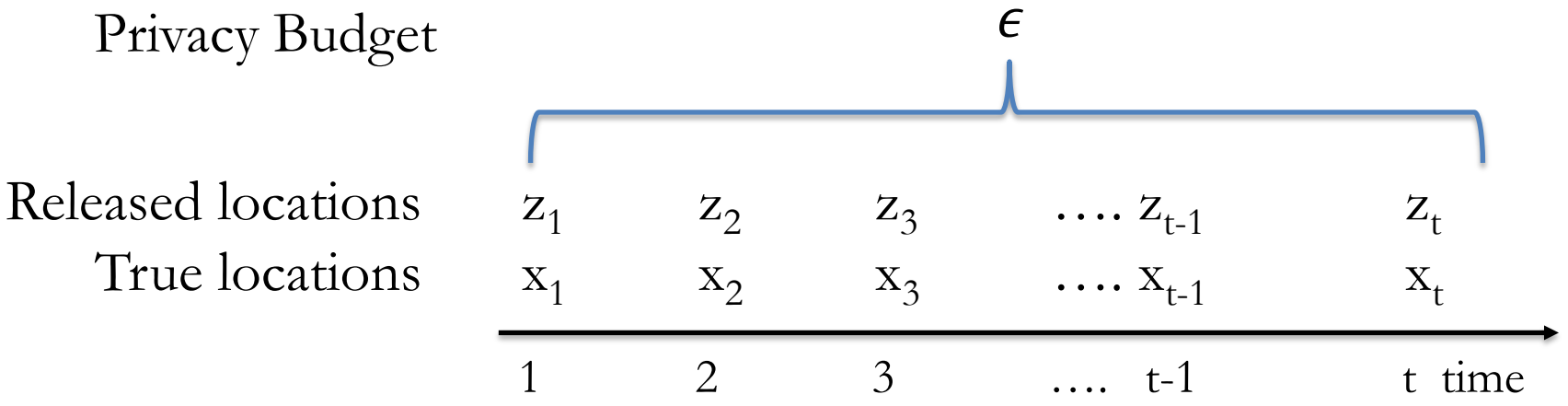
Overview of DP Definitions

Neighbors differ in	What to hide?
Trajectory	**All properties of the individual are secret** e.g. Where is home location?
Window	**Properties within a small window** e.g. Did user visit home in the last hour?
Event	**Properties at a specific time** e.g. Did user visit home at time t?
Geo-indistinguishability	**Some properties (not all) at a specific time** e.g. Did user visit near home at time ?

Different Levels of Protection

- User-level DP for entire trajectory
  - Neighboring databases D_1, D_2
    - Differ in one user's entire stream
  - Release aggregate statistics for multiple users' data

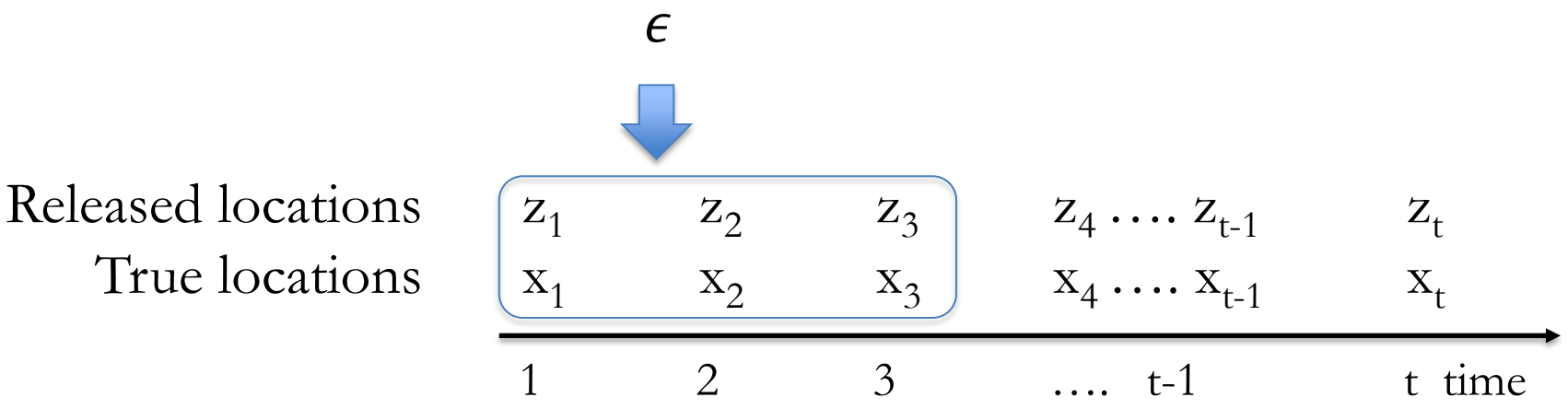
[Chen et al CCS'12, He et al VLDB'15]



Different Levels of Protection

[Kellaris et al VLDB'14]

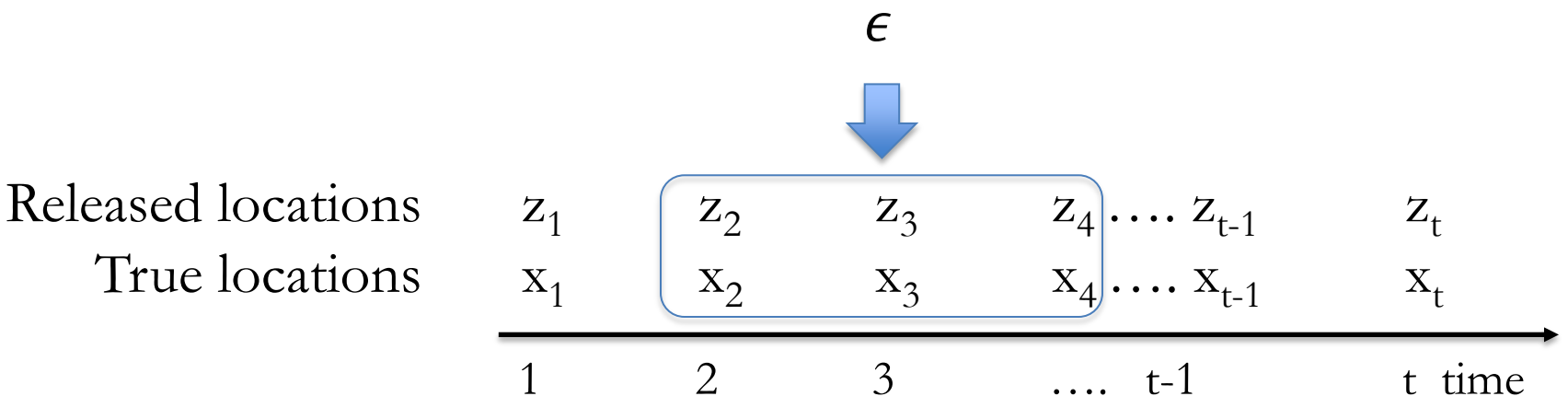
- w-event DP
  - E.g. w=3



Different Levels of Protection

[Kellaris et al VLDB'14]

- w-event DP
  - E.g. w=3

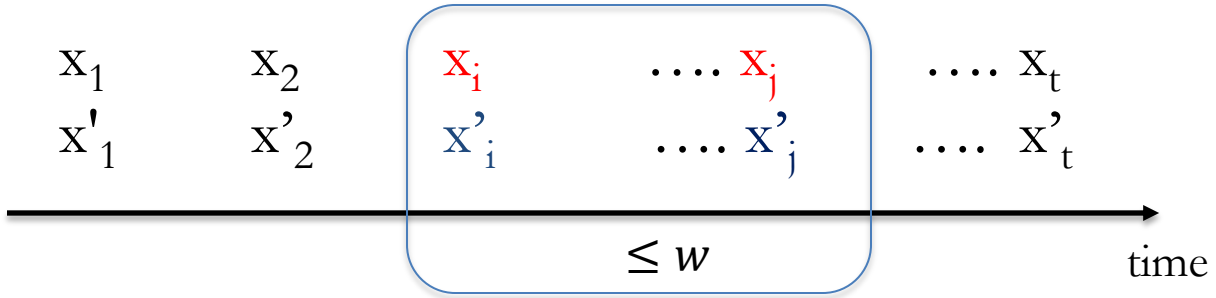


Different Levels of Protection

[Kellaris et al VLDB'14]

- w-event DP
  - **Neighboring stream prefix** $(x_1, x_2.., xt), (x'1, x'2.., x'_t)$
    - For any $i < j$, if $x_i \neq x'_i$, and $x_j \neq x'_j$ then $j - i + 1 \leq w$
    - x_i and x'_i are the same or neighboring

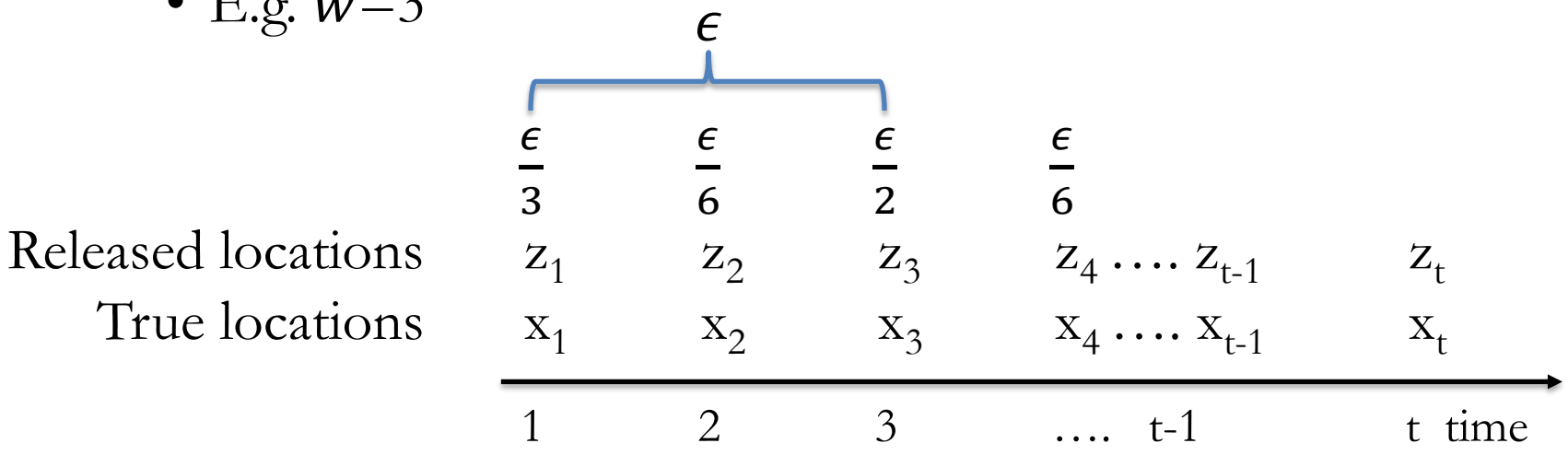
→Protect updates happening within w-event with privacy budget ϵ



Different Levels of Protection

[Kellaris et al VLDB'14]

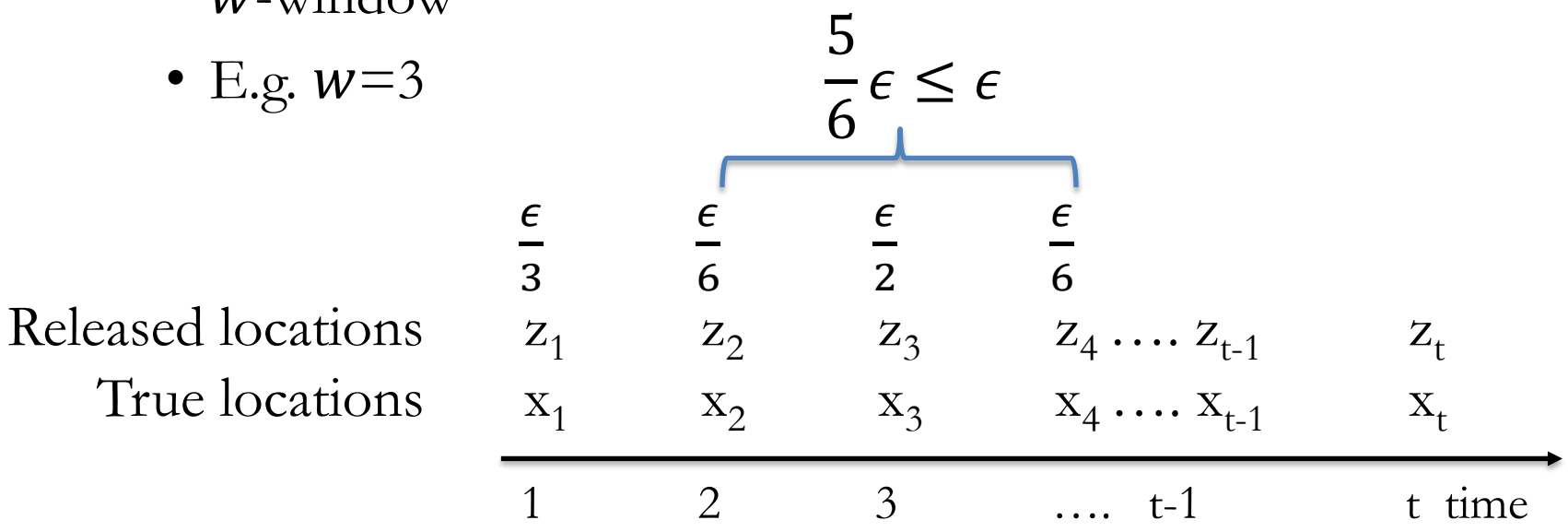
- w-event DP
  - Allow budget allocation strategy:
    - Adaptive assign privacy budgets to events within the same w-window
    - E.g. w=3



Different Levels of Protection

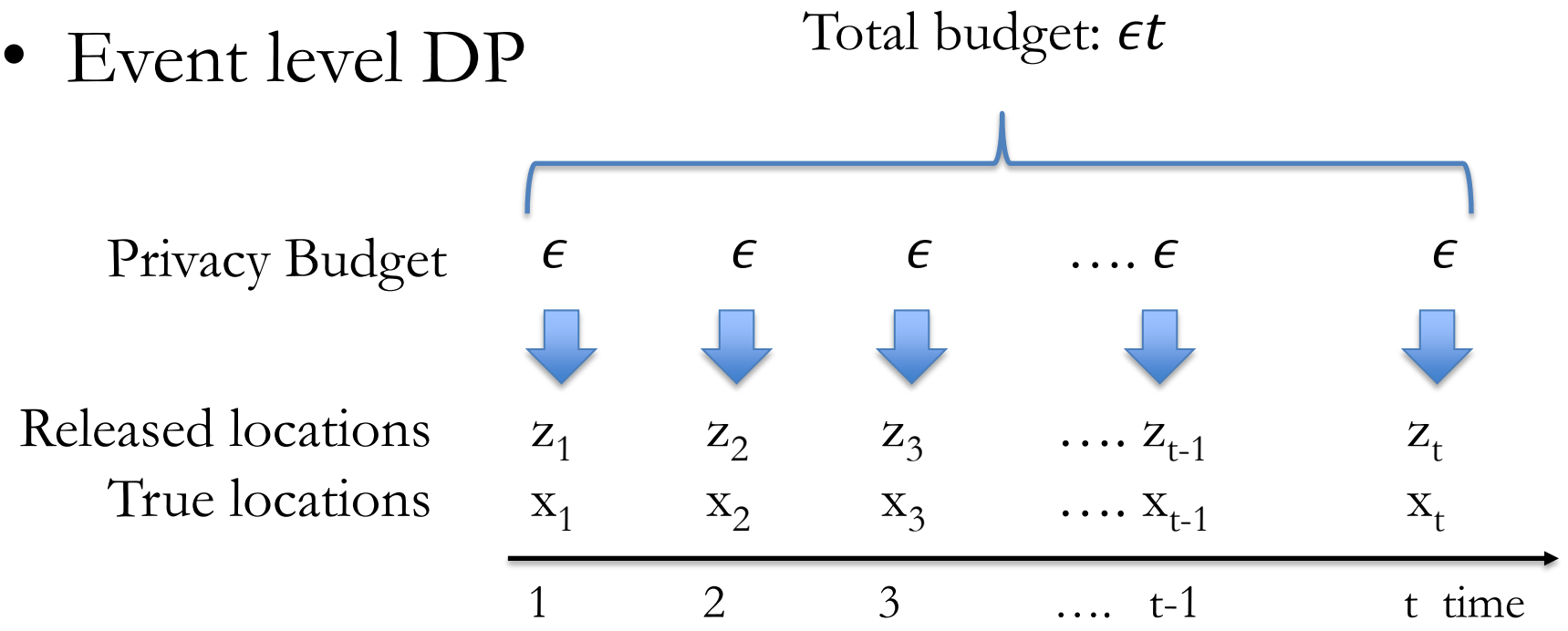
[Kellaris et al VLDB'14]

- w-event DP
  - Allow budget allocation strategy:
    - Adaptive assign privacy budgets to events within the same w-window
    - E.g. w=3



Different Levels of Protection

- Event level DP

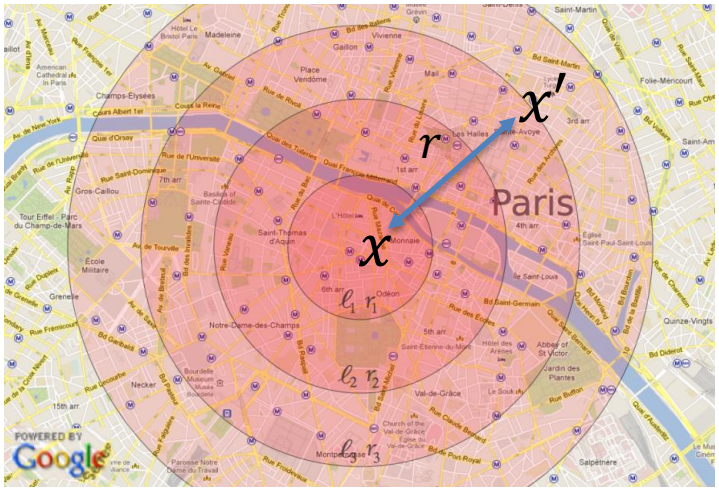


If a person staying at a location for a long time $x_1 = x_2 = \cdots = xw$, averaging $(z_1, .., z_w)$ leaks the true location.

Protect a Single Location

- Protect a single location

  e.g. Location-based Services (LBS) to find a restaurant
  - Not reveal the exact location
  - Revealing an approximate location is ok



- A mechanism satisfies **ϵ-geo-indistinguishability** iff for all observations $S \subseteq Z$, for all $r > 0$, for all **neighbors** $x, x' : d(x, x') \leq r$, [Andrés et al CCS'13]

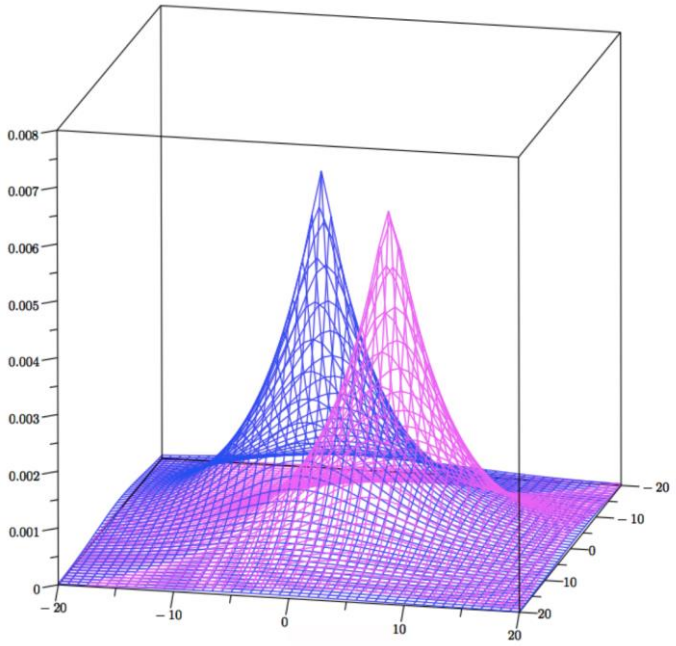
$$\Pr[S|x] \leq e^{\epsilon r} Pr\,[S|x']$$

Bivariate Laplacian

- Bivariate Laplacian

$$p_x(z) = \frac{\epsilon^2}{2\pi} e^{\epsilon d(x,z)}$$

- Efficient method to draw points based on polar coordinates



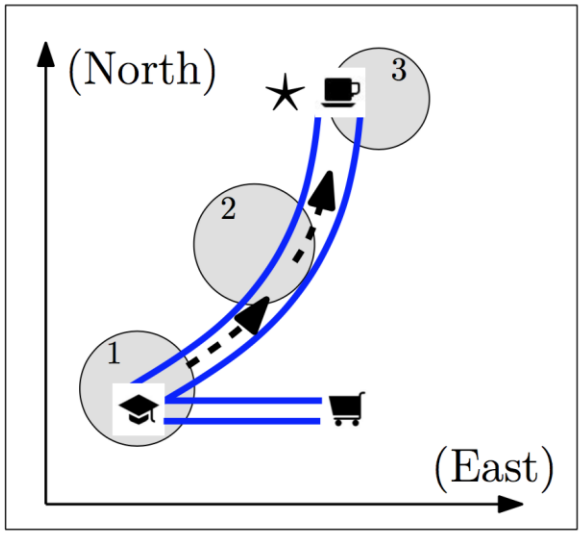
Handling Correlations within Trajectories

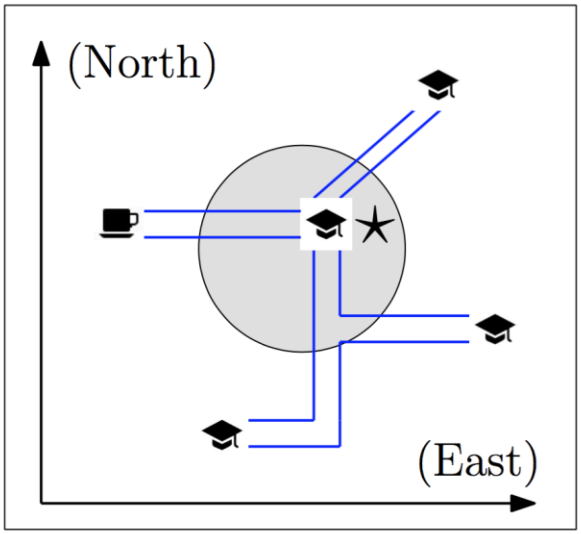
- Reasonable to assume entire trajectories are uncorrelated
- For other neighboring definitions, windows or events in a trajectory are correlated in all weaker notions
- E.g. physical constraints (alas: no wormholes ... yet) impose correlations between locations at adjacent time points

Temporal Correlations

[Xiao & Xiong CCS'15]

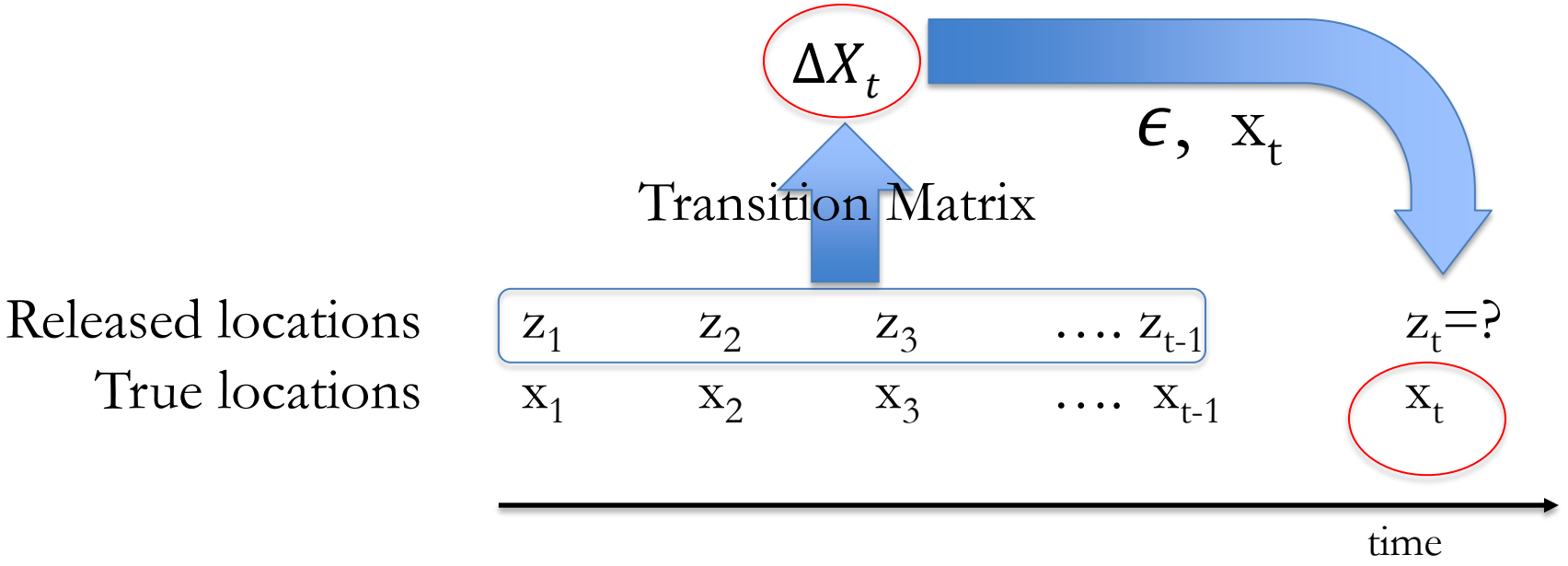
- Privacy breach on geo-indistinguishability based trajectory release, given the
  - **road network constraint**
  - **user's moving pattern**





Temporal Correlation

- **δ-Location Set**: ΔX_t a smallest set of locations a user can appear with probability sum greater than $1-\delta$, based on previously released noisy locations (z_1 ,.., z_{t-1})



Temporal Correlation

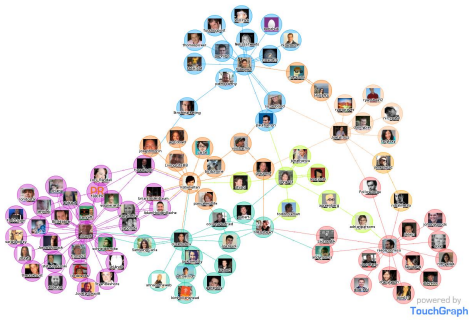
At any timestamp t, a randomized mechanism A satisfies **ϵ-differential privacy on δ-location set ΔX_t** if, for any output z_t and any two **neighboring** locations x_1 and x_2 in ΔX_t, the following holds:

$$\Pr[A(x_1) = zt] \leq e^{\epsilon} \Pr[A(x_2) = zt]$$

Different Level of Privacy Protection

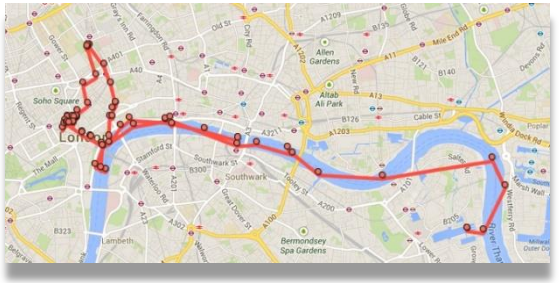
- Differential Privacy for Non-tabular Data

Social network



- Edge DP
- Node DP

Location stream

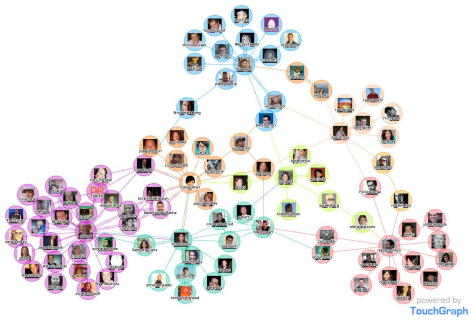


- ϵ-indistinguishability
- Event DP
- w-event DP
- User level DP

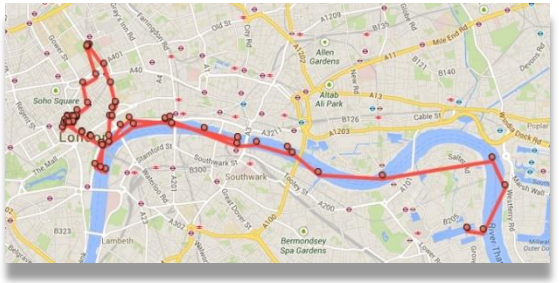
Outline of Module 6

- Differential Privacy for Non-tabular Data

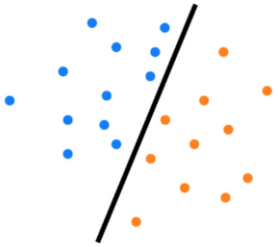
Social network



Location stream

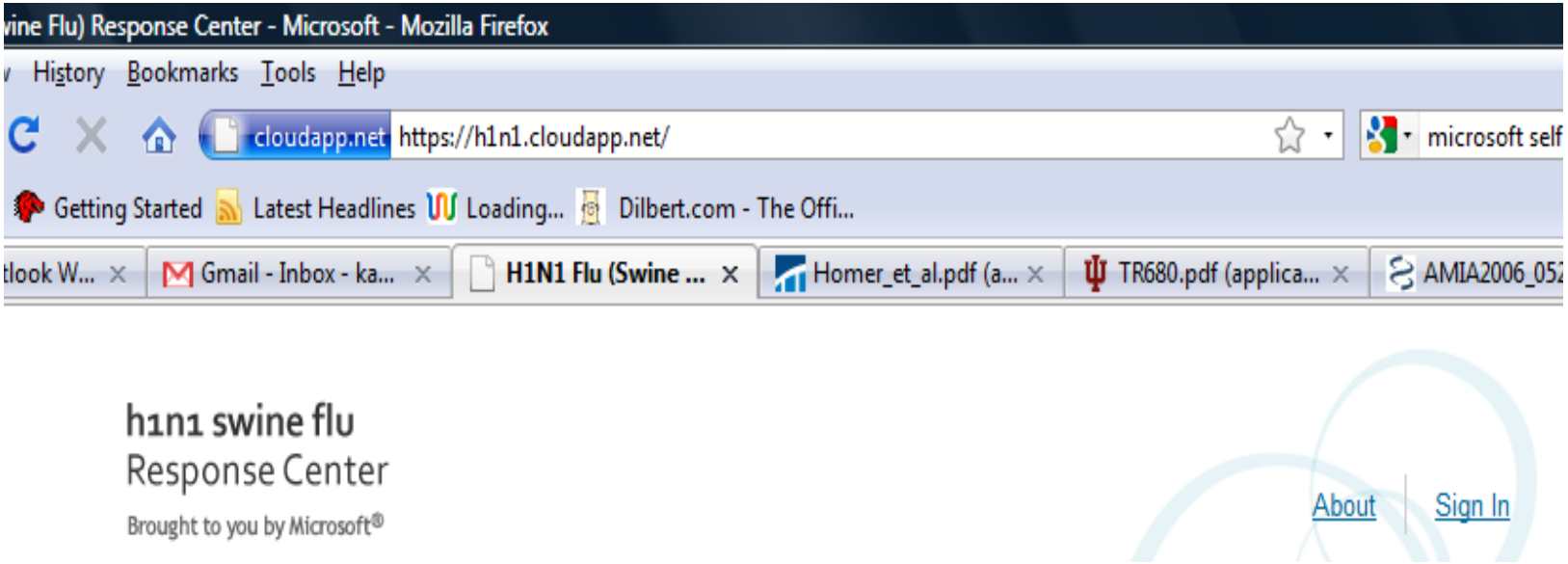


- Differential Privacy for Machine Learning



Credit: Chaudhuri

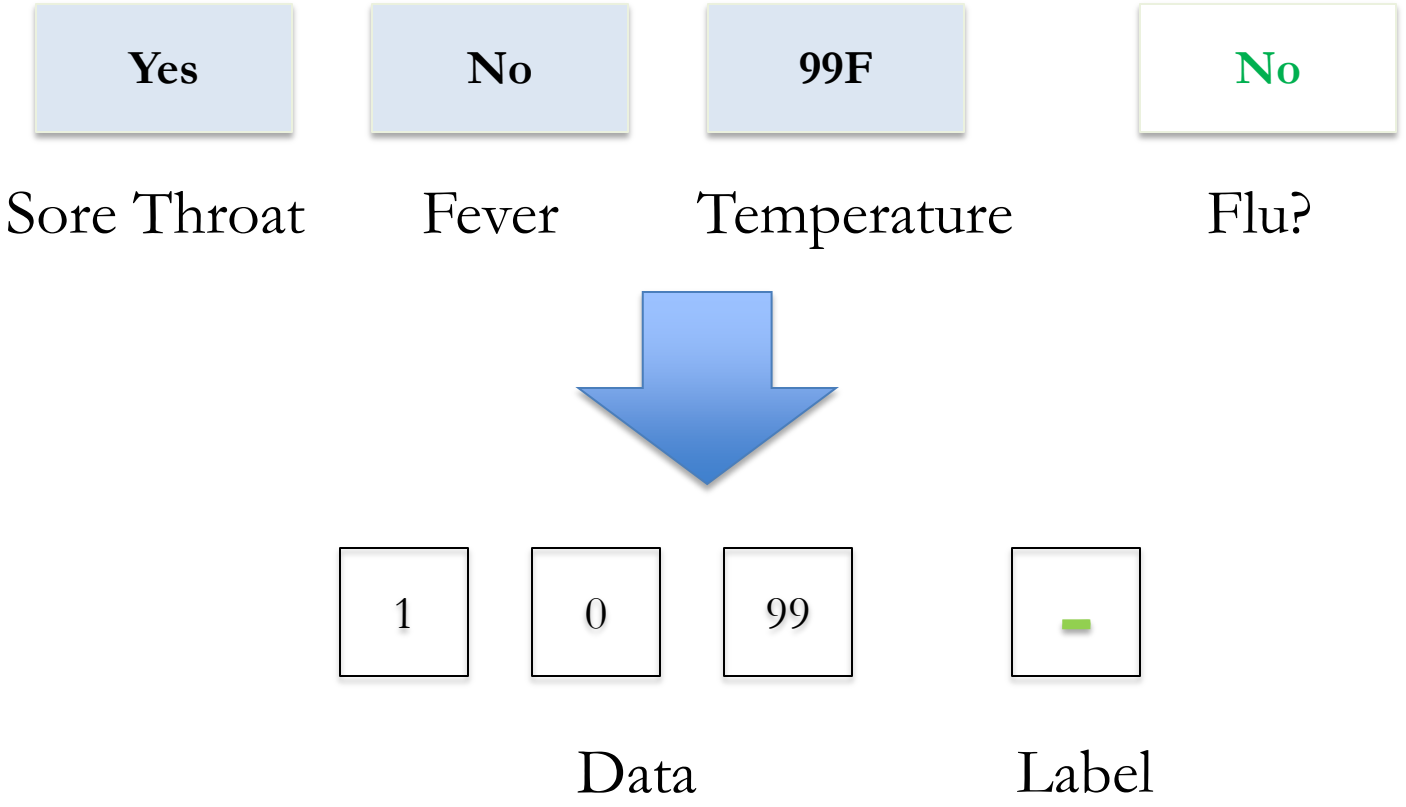
Differentially Private Machine Learning



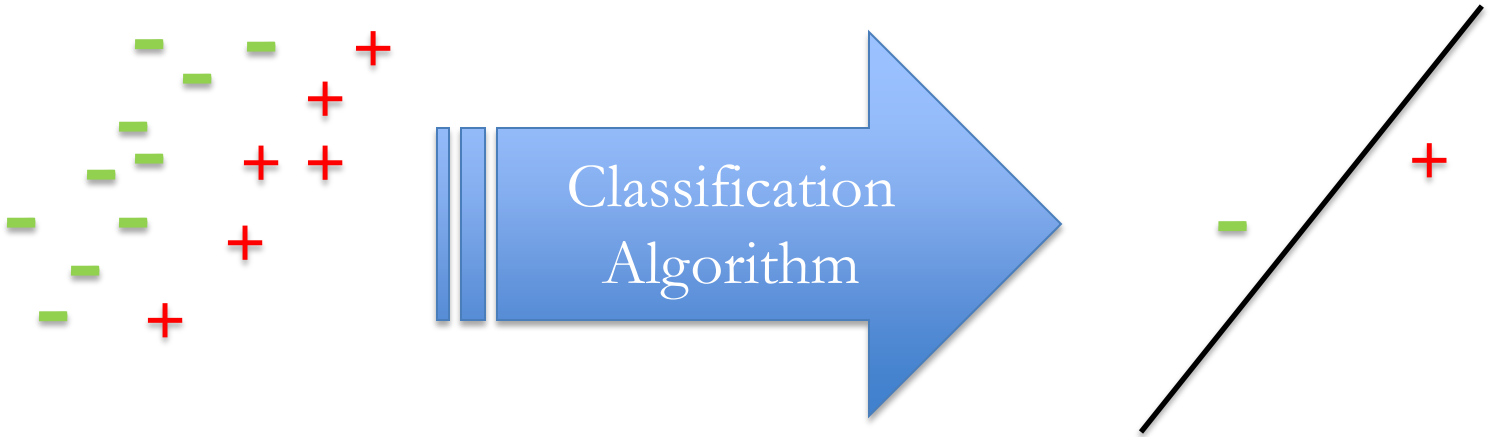
Predicts flu or not, based on patient symptoms

Trained on sensitive patient data

From Attributes to Labeled Data



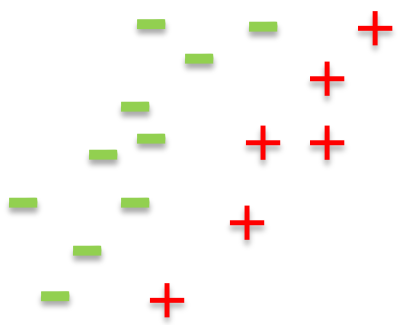
Classifying Sensitive Data



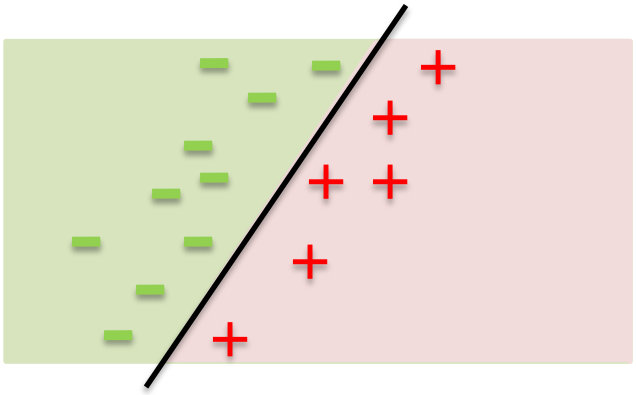
Private Data

Public Classifier

Classifying Sensitive Data



Classifying Sensitive Data



Distribution P over labelled examples

Goal: Find a vector w that separates + from - for most points from P

Key: Find a simple model to fit the samples

Empirical Risk Minimization

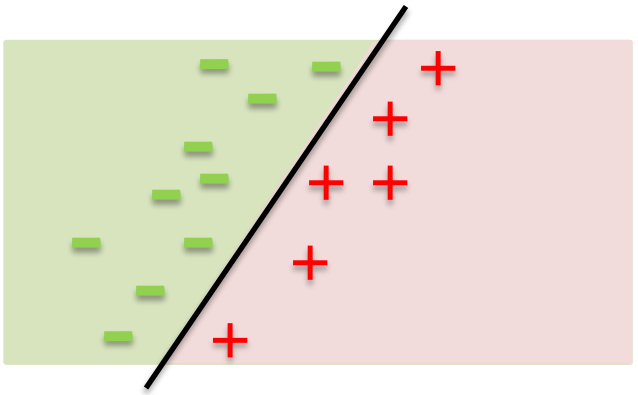
- **Goal:** Labeled data $D = \{(x_i, y_i)\}$, find

$$f(D) = argmin_{\omega} \frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} L(y_i \omega^T x_i)$$

Regularizer (Model Complexity)

Risk (Training Error)

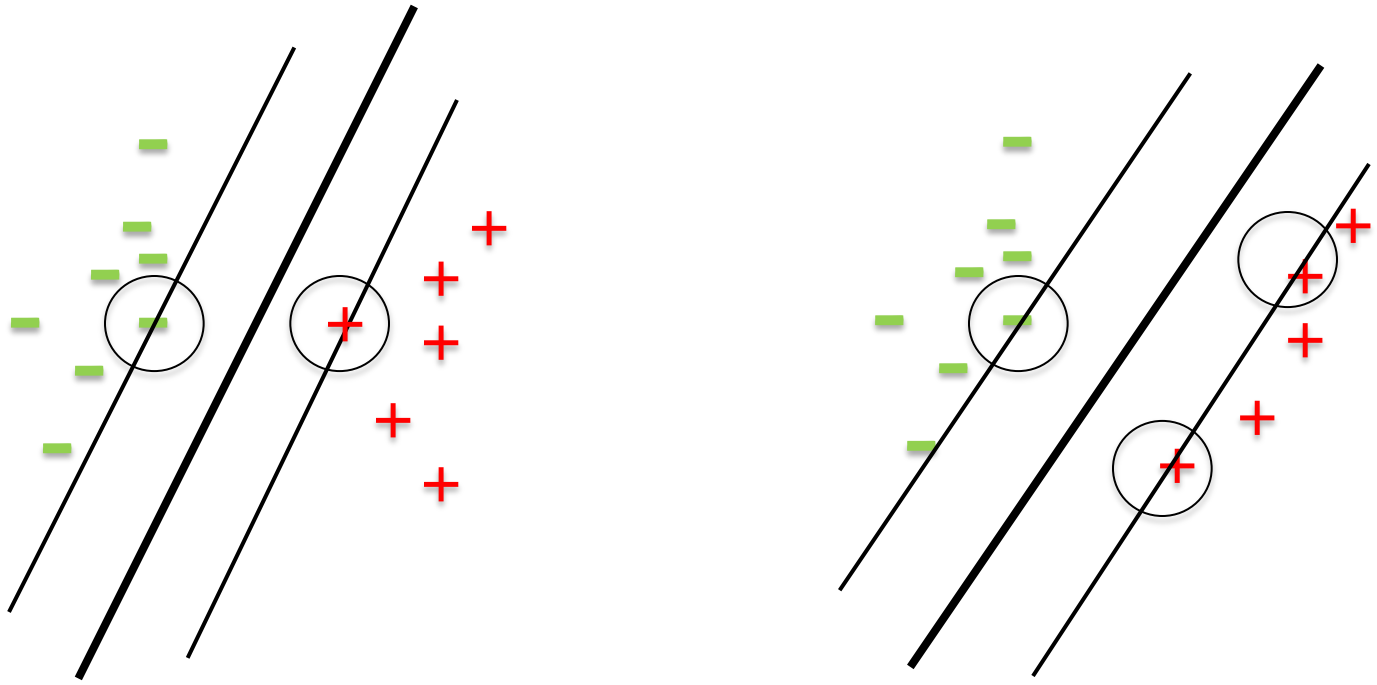
Examples of Loss Function



Risk: Hinge loss
Optimizer: Support vector machines (SVM)

Risk: Logistic loss
Optimizer: Logistic regression

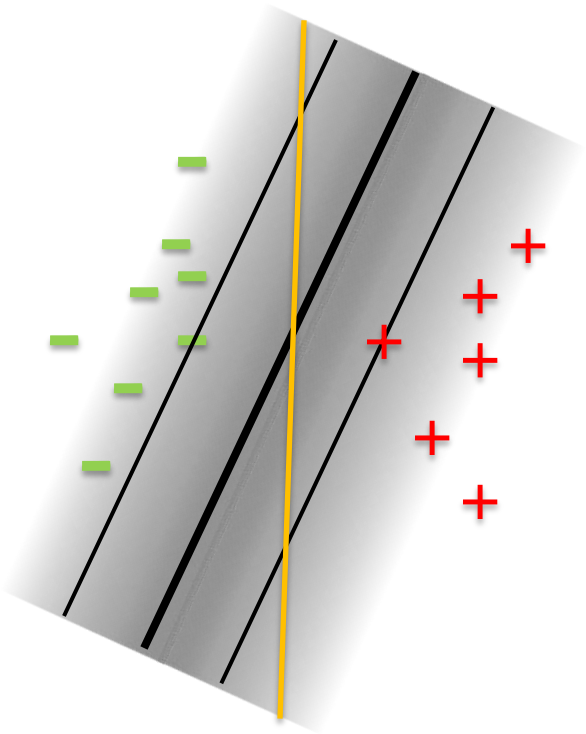
Global Sensitivity of SVM Solution



SVM solution is a combination of support vectors

If one support vector moves, solution changes

How to make ERM private?



Pick ω from distribution near the optimal solution

Output Perturbation

- What the sensitivity of this function?

$$f(D) = argmin_{\omega} \frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} L(y_i \omega^T x_i)$$

Theorem: [Chaudhuri et al MLR'11]

If $\parallel x_i \parallel \leq 1$ and L is l-Lipschitz, then for any D, D' with $\mathrm{dist}(D, D') = l$,

$$\parallel f(D) - f(D') \parallel_2 \leq \frac{2}{\lambda n}$$

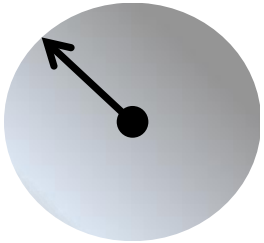
Output Perturbation

- **Goal:**

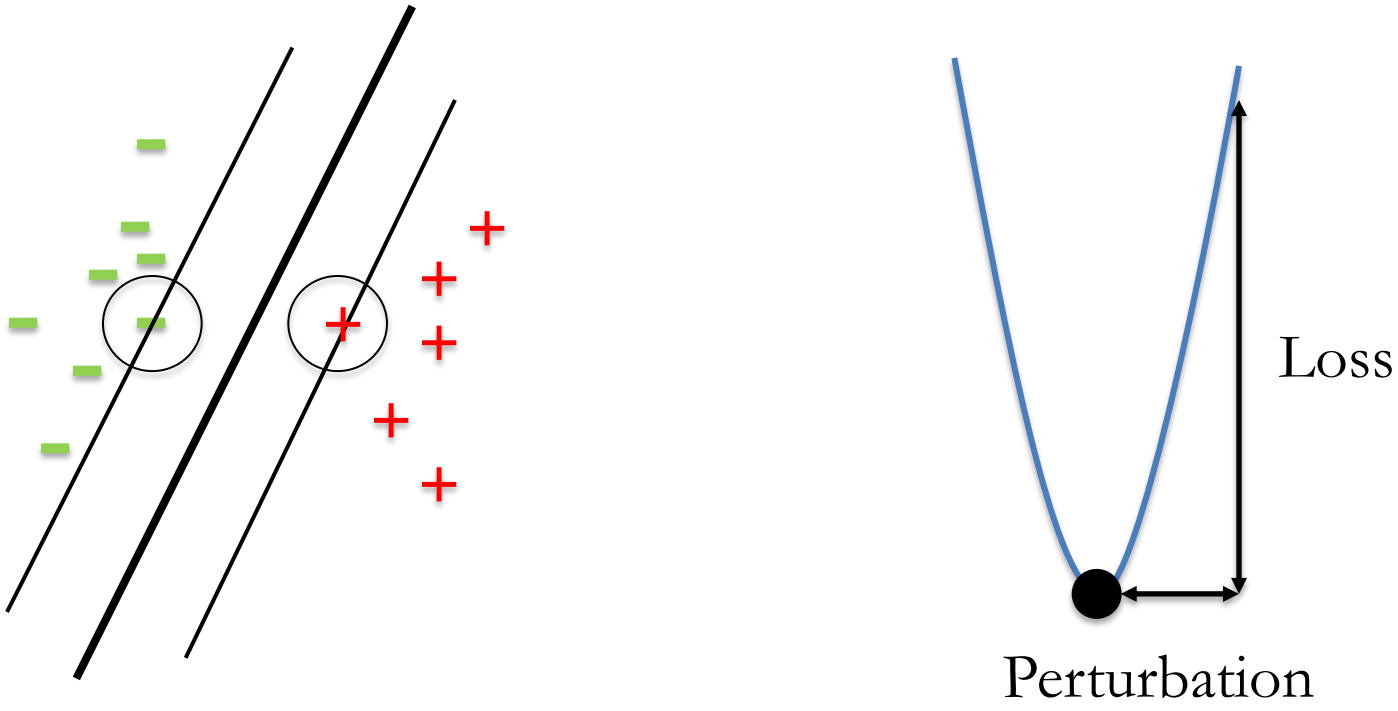
$$\tilde{f}(D) = f(D) + noise =$$

$$\left[argmin_{\omega} \frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} L(y_i \omega^T x_i) \right] + noise$$

- $noise$ drawn from
  - Magnitude: drawn from $\Gamma(\mathrm{d}, \frac{2}{\lambda n \epsilon})$
  - Direction: uniform at random



Objective Perturbation



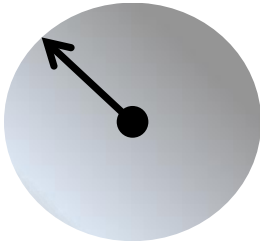
Optimization surface is very steep in some direction
High loss if perturbed in those directions

Objective Perturbation

- **Insight:** Perturb optimization surface and then **optimize**

$$\tilde{f}(D) = argmin_{\omega} \left[\frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} L(y_i \omega^T x_i) + noise \right]$$

- $noise$ drawn from
  - Magnitude: drawn from $\Gamma(\mathrm{d}, \frac{1}{\alpha})$
  - Direction: uniform at random



Objective Perturbation

- **Insight:** Perturb optimization surface and then **optimize**

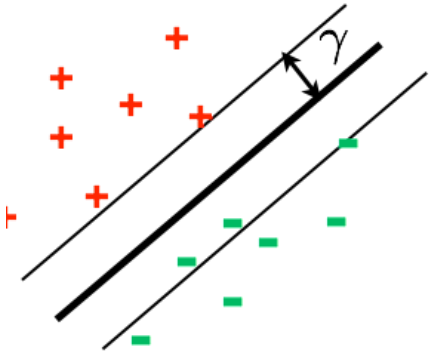
$$\tilde{f}(D) = argmin_{\omega} \left[\frac{1}{2}\lambda \parallel \omega \parallel^2 + \frac{1}{n}\sum_{i=1}^{n} L(y_i \omega^T x_i) + noise \right]$$

Theorem: If L is convex and double-differentiable with $|\mathrm{L}'(\mathrm{z})| \leq l$, $|\mathrm{L}''(\mathrm{z})| \leq c$ then Algorithm is $\alpha + 2\log\left(1 + \frac{c}{n\lambda}\right)$-differentially private. [Chaudhuri et al MLR'11]

Accuracy

- Number of samples for error ϵ
  - Fewer samples implies higher accuracy

d: #dimensions
γ: marginal
α: privacy
ϵ: error
$\gamma, \alpha, \epsilon < 1$



- **Normal SVM:** $\frac{1}{(\epsilon\gamma)^2}$
- **Objective perturbation:** $\frac{1}{(\epsilon\gamma)^2} + \frac{d}{\epsilon\alpha\gamma}$
- **Output perturbation:** $\frac{1}{(\epsilon\gamma)^2} + \frac{d}{\epsilon^{1.5}\alpha\gamma}$

Open Questions

- Solve non-convex optimization
  - Deep learning
    - Stochastic gradient descent [Abadi et al arXiv'16]
- How well does DP machine learning algorithms work for real datasets?
  - [Zhang et al SIGMOD'12, Fredrikson et al USENIX'14]
- Understand the relationship between *Stability* of learning algorithms and DP

In Summary

1. Privacy Problem Statement
2. Differential Privacy
3. Algorithms for Tabular Data
4. Applications I
5. Privacy beyond Tabular Data
6. Applications II

References

Lars Backstrom, Cynthia Dwork, Jon M. Kleinberg. "*Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*". WWW 2007

Arvind Narayanan and Vitaly Shmatikov. "De-anonymizing Social Networks." IEEE Security & Privacy 2009

S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, A. Smith. "*Analyzing Graphs with Node Differential Privacy*". TCC 2013

K. Nissim, S. Raskhodnikova, A. Smith. "*Smooth Sensitivity and Sampling in Private Data Analysis*".STOC 2007

S. Raskhodnikova, A. Smith, "*Efficient Lipschitz Extensions for High-Dimensional Graph Statistics and Node Private Degree Distributions*", arXiv:1504.07912v1, 2015

W. Day, N. Li, M. Lyu. "*Publishing Graph Degree Distribution with Node Differential Privacy*". SIGMOD 2016

K. Nissim, S. Raskhodnikova, and A. Smith. "*Smooth sensitivity and sampling in private data analysis*".STOC 2007.

V. Karwa, S. Raskhodnikova, A. Smith, G. Yaroslavtsev. "*Private Analysis of Graph Structure*". VLDB 2011

C. Dwork, J. Lei. "*Differential privacy and robust statistics*". STOC 2009

M. Hay, C. Li, G. Miklau, and D. Jensen. "*Accurate estimation of the degree distribution of private networks*" ICDM 2009.

M. Hay, V. Rastogi, G. Miklau, and D. Suciu. "*Boosting the accuracy of differentially-private queries through consistency*". PVLDB 2010.

References

V. Karwa, A. B. Slavkovic. "*Differentially Private Graphical Degree Sequences and Synthetic Graphs*". Privacy in Statistical Databases 2012: 273-285

B. Lin, D. Kifer. "*Information preservation in statistical privacy and bayesian estimation of unattributed histograms*". SIGMOD 2013

A. Gupta, A. Roth, J. Ullman. "*Iterative Constructions and Private Data Release*". TCC 2012

J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava and X. Xiao. "*Private Release of Graph Statistics using Ladder Functions*." SIGMOD 2015

V. Karwa and A. B. Slavkovi´c and P. Krivitsky. "*Differentially Private Exponential Random Graphs*". arXiv:1409.4696v2 2015

W. Lu. G. Miklau. "*Exponential random graph estimation under differential privacy*". KDD 2014

V. Karwa, P. N. Krivitsky, A. B. Slavković . "*Sharing Social Network Data: Differentially Private Estimation of Exponential-Family Random Graph Models*". arXiv:1511.02930v1. 2015

J. Blocki, A. Blum, A. Datta, O. Sheffet. "*The Johnson-Lindenstrauss Transform Itself Preserves Differential Privacy*". FOCS 2012

M. Hardt, G. N. Rothblum. " *A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis*". FOCS 2010

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. " *Unique in the crowd: The privacy bounds of human mobility*". Sci. Rep., 3(1376), 2013.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi. " *Limits of predictability in human mobility*". Science, 327(5968):1018–1021, 2010

S. Gambs, M.O. Killijian, M. Cortez. "*Show Me How You Move and I Will Tell You Who You Are*". SPRINGL 2010

References

M. Andrés, N. Bordenabe, K. Chatzikokolakis, Catuscia Palamidessi. "*Geo-Indistinguishability: Differential Privacy for Location-Based Systems*". CCS 2013

Y. Xiao, L. Xiong. " *Protecting Locations with Differential Privacy under Temporal Correlations*". CCS 2015

G. Kellaris, S. Papadopoulos, X. Xiao, D. Papadias. "*Differentially Private Event Sequences over Infinite Streams*". VLDB 2014

X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, D. Srivastava, "*DPT: Differentially private trajectory synthesis using hierarchical reference systems*", VLDB 2015

R. Chen, G. Acs, and C. Castelluccia. "*Differentially private sequential data publication via variable-length n-grams*". In CCS, 2012.

K. Chaudhuri, C. Monteleoni, A. D. Sarwate. "*Differentially Private Empirical Risk Minimization*". In Journal of Machine Learning Research 2011

R. Bassily, A. Smith, A. Thakurta. "*Private Empirical Risk Minimization, Revisited*". FOCS 2014

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D.Page, T. Ristenpart. "*Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing.*" USENIX 2014

J. Zhang, Z. Zhang, X. Xiao, Y. Yang, M. Winslett. "*Functional Mechanism: Regression Analysis under Differen6al Privacy*". PVLDB 2012

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang. "*Deep Learning with Differential Privacy*". arXiv:1607.00133v1 2006