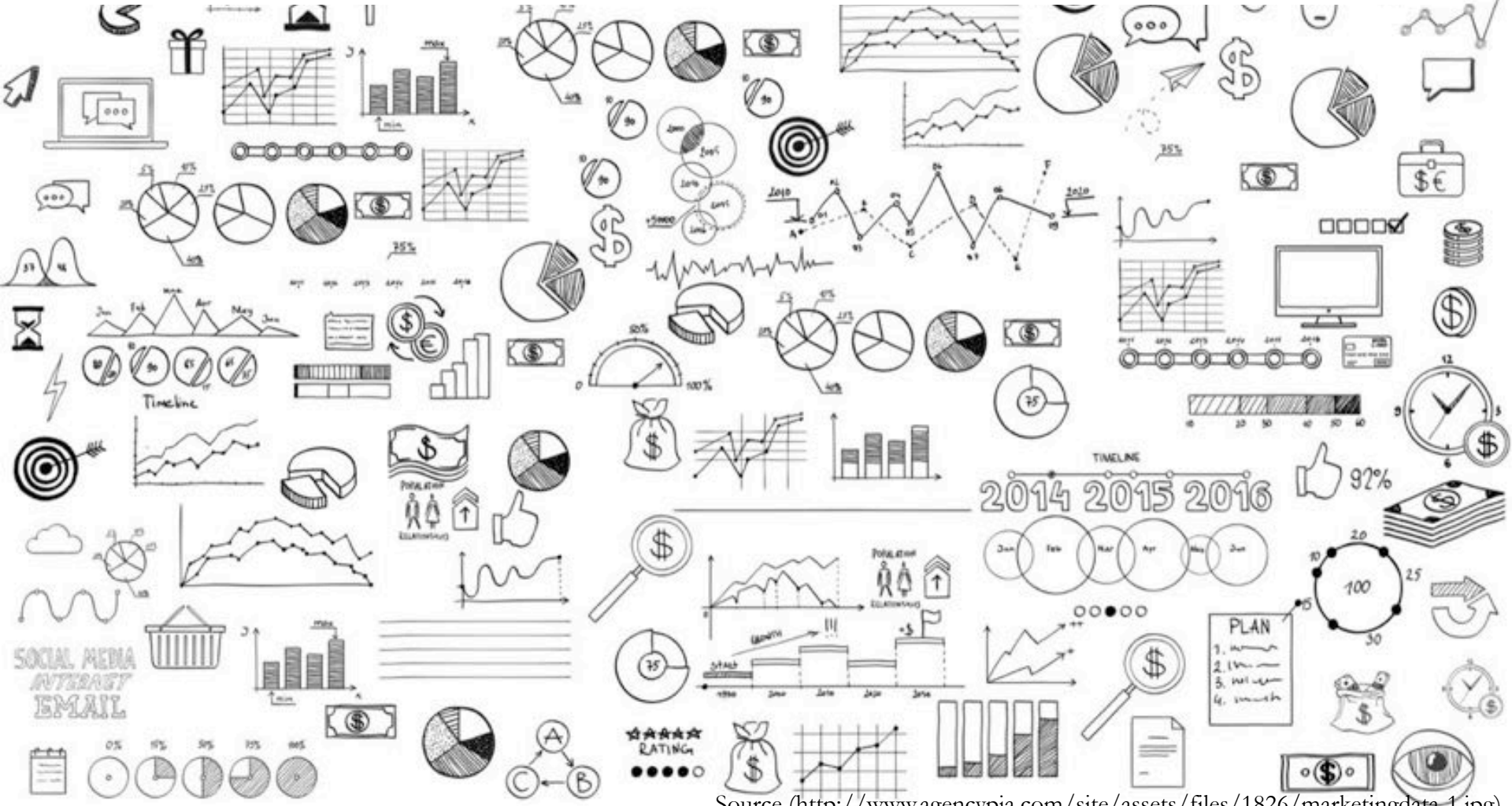
Differential Privacy

Ashwin Machanavajjhala

ashwin@cs.duke.edu

Our world is increasingly data driven



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

Personal data is collected in a variety of ways

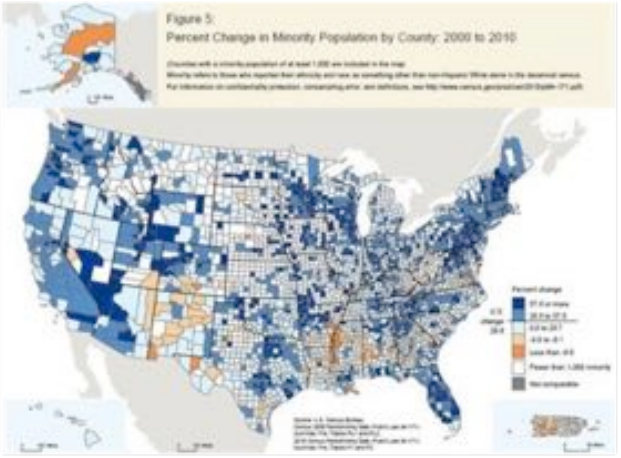

nest
amazon.com
LINKSYS
nest
HEATING
20
amazon

Aggregated Personal Data …

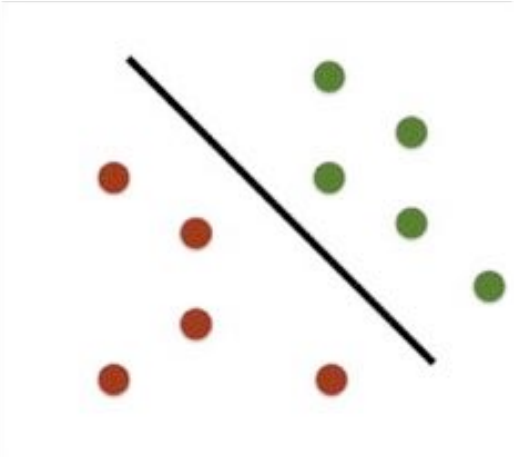
… is made publicly available in many forms.

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

Statistics
(e.g., demographic)



Predictive models
(e.g., advertising)



… but privacy breaches abound

US military reviewing security practices after fitness app reveals sensitive info

By Joshua Berlinger and Maegan Vazquez, CNN

… but privacy breaches abound

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

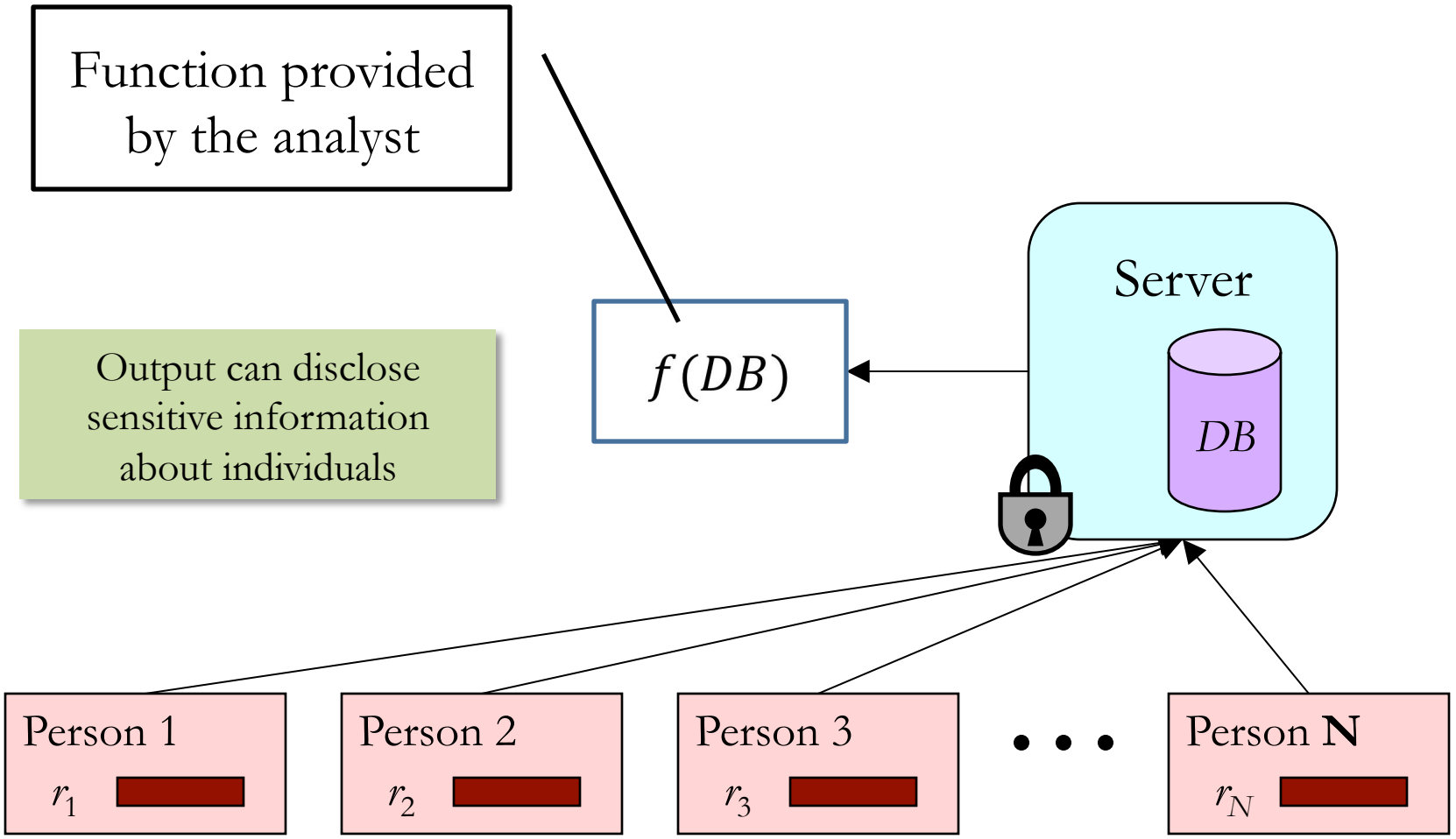
The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013

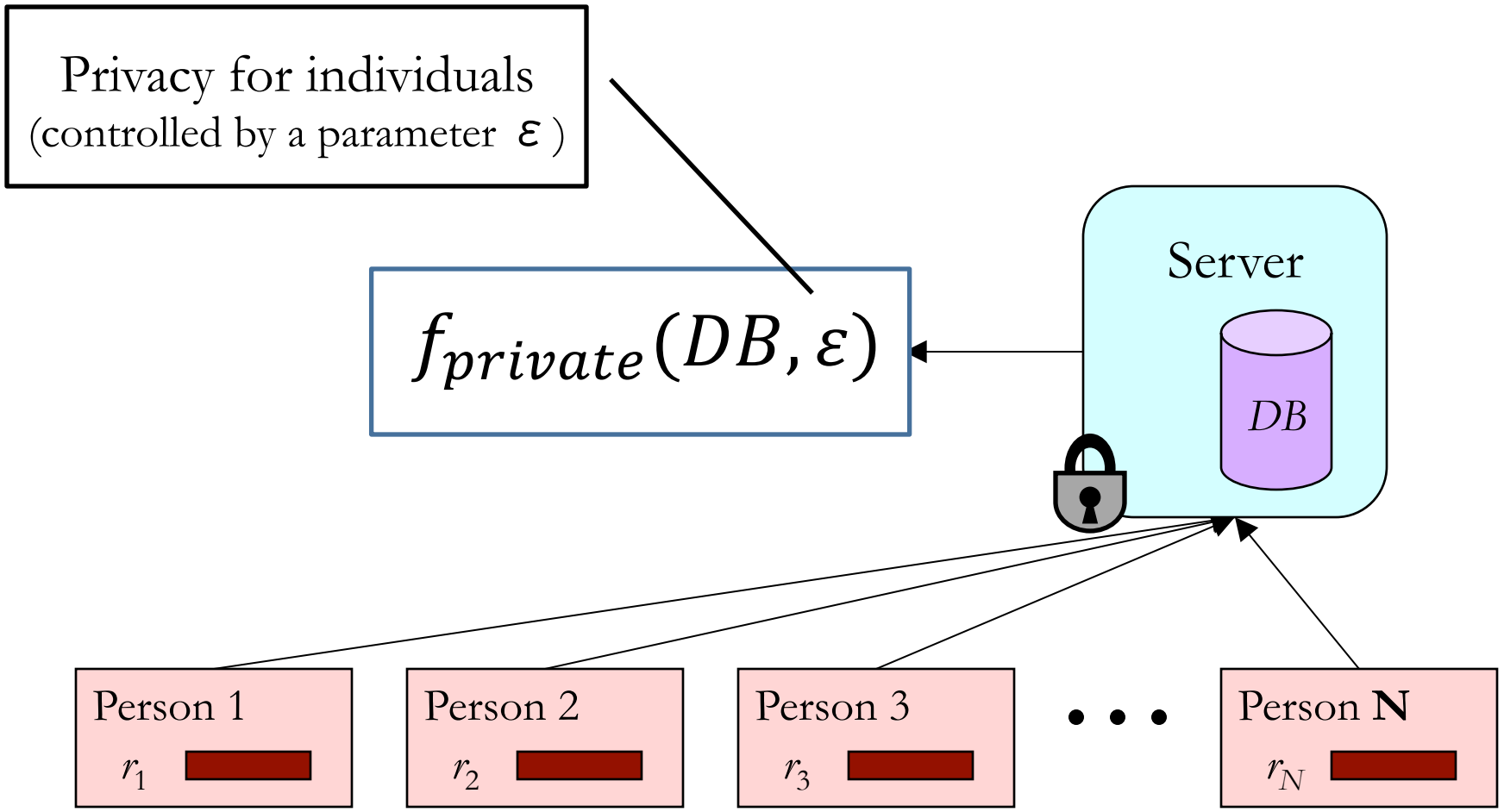
US military reviewing security practices after fitness app reveals sensitive info

By Joshua Berlinger and Maegan Vazquez, CNN

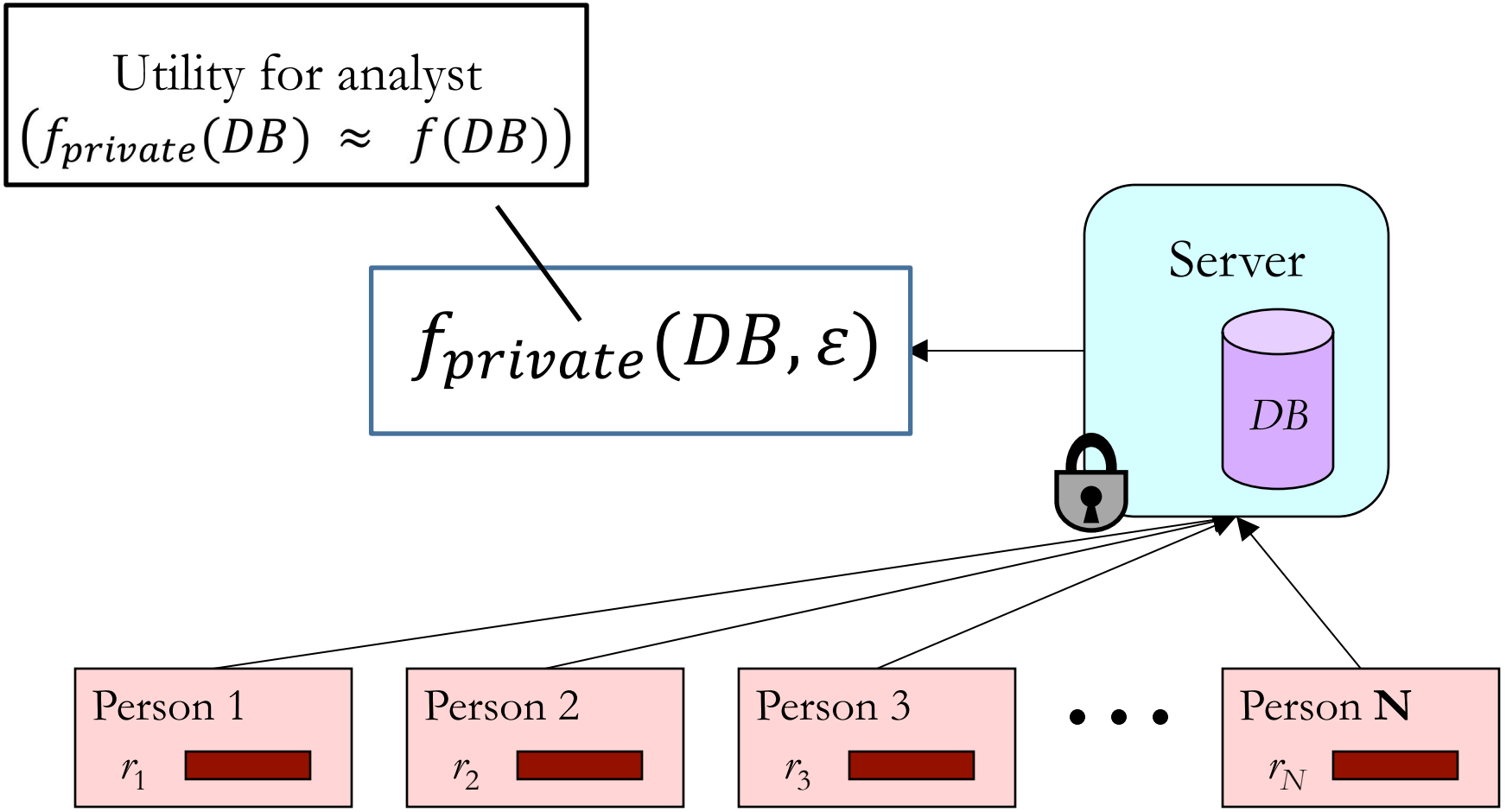
Statistical Database Privacy



Statistical Database Privacy



Statistical Database Privacy



Statistical Database Privacy

- Goal: **Learn aggregate properties** of the dataset (or the population from which dataset is drawn) with **provable privacy for individuals** in the data

- Can't be solved by:
  - Secure Computation/FHE
  - Access control / Privacy preferences
  - Naïve anonymization/Swapping

Motivation: Real world use cases

- Releasing Census Data

- Healthcare Analytics

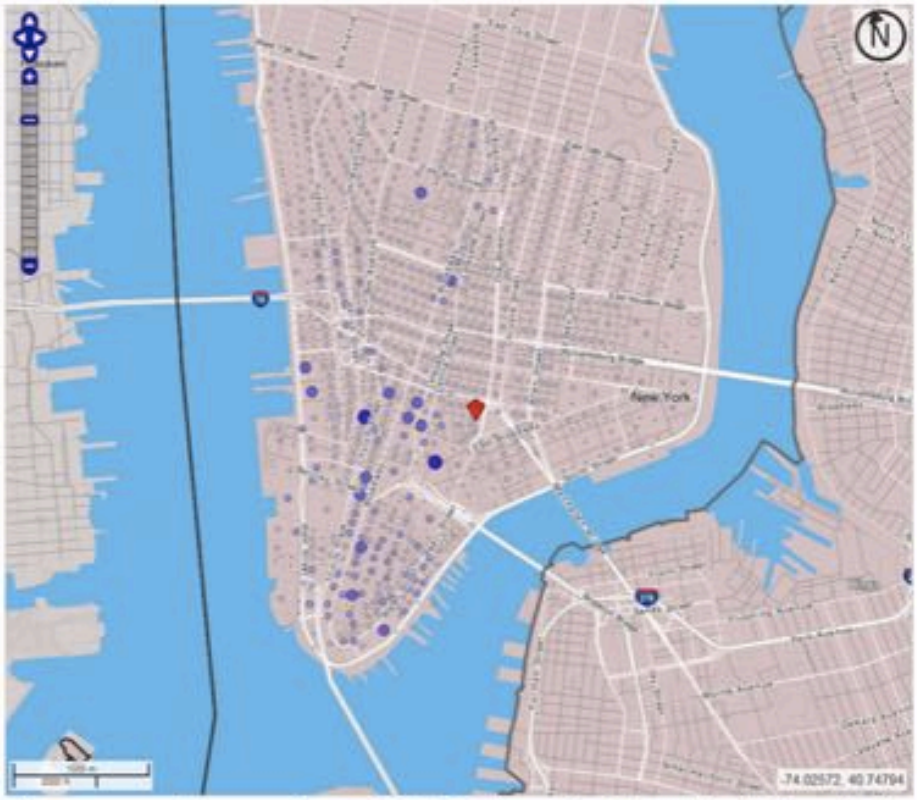
- Web data collection

United States Census Bureau

OnTheMap

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

Employment in Lower Manhattan



Residential pattern of workers employed in Lower Manhattan

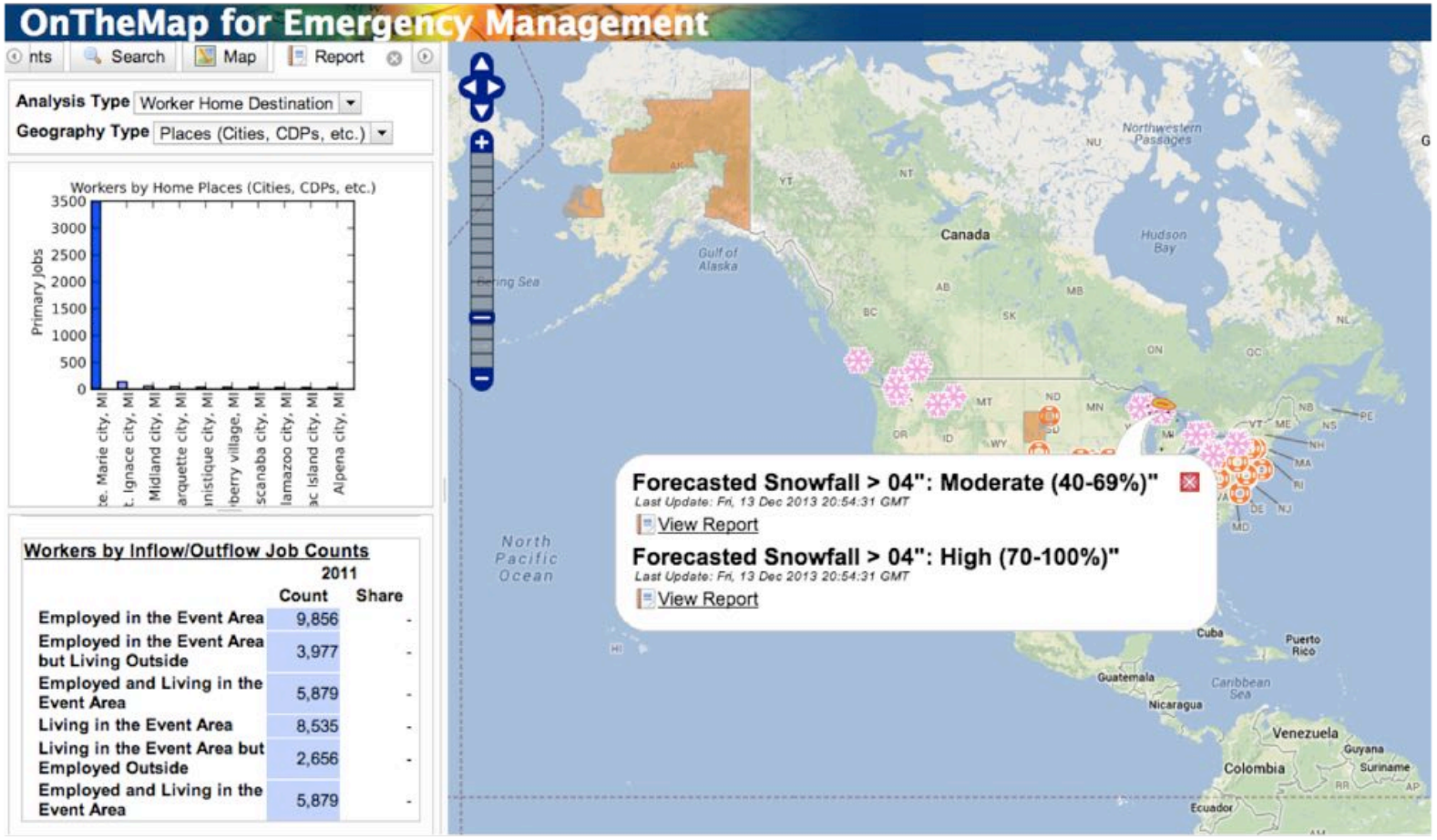
The maps above show LODES data in New York City in the OnTheMap application. The map on the left shows employment by census block in Lower Manhattan (in dense urban areas one census block is often equivalent to one city block). Large, dark dots have more employment than small, light dots. The map on 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.

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 release such data?



Why privacy is needed?

US Code: Title 13 CENSUS

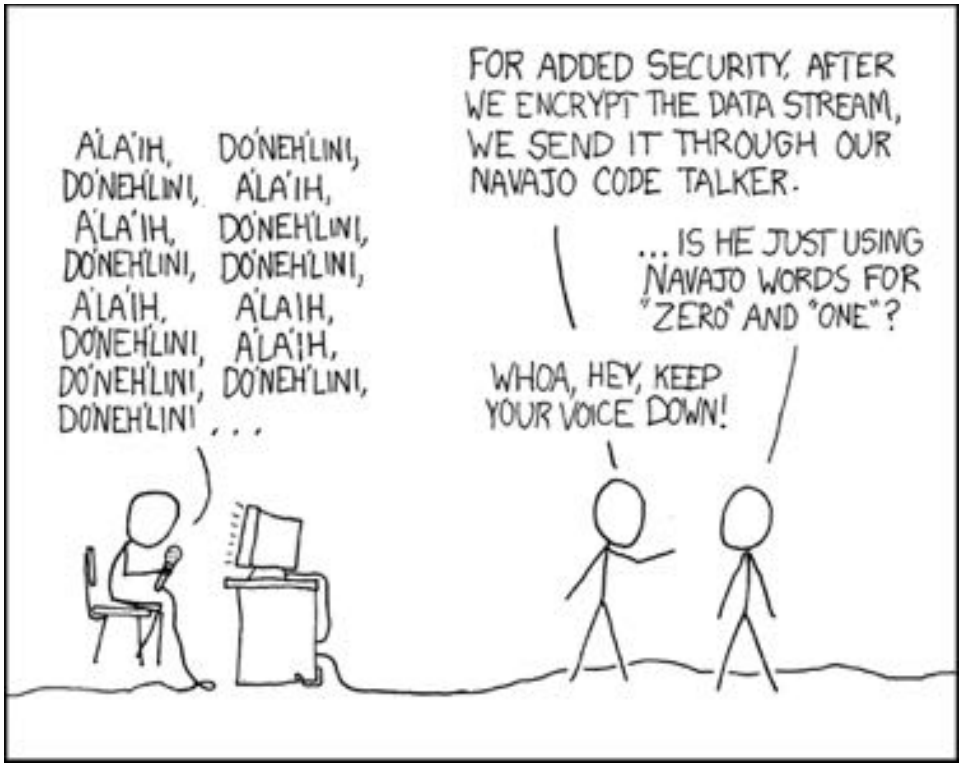
It is against the law to make any publication whereby the data furnished by any particular establishment or individual under this title can be identified.

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

Machanavajjhala, Kifer, Abowd, Gehrke, Vilhuber, "Privacy: Theory to Practice on the map", ICDE 2008

Challenge

- Current algorithm for data release with *no provable guarantees* and parameters used have to be kept *secret*



Motivation: Real world use cases

- Releasing Census Data

- **Healthcare Analytics**

- Web data collection

Healthcare Cost and Utilization Project

Why Privacy?


HIPAA
Health Insurance Portability
and Accountability Act

#Hospital discharges in NJ of ovarian cancer patients, 2009

Counts less than k are suppressed

Age	#discharges	White	Black	Hispanic	Asian/Pcf HInder	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	*	*	*	*	*	*

Vaidya, Shafiq, Jiang, Ohno-Machado, "Identifying inference attacks against healthcare data repositories", AMIA 2013

#Hospital discharges in NJ of ovarian cancer patients, 2009

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

= 535 – (40+236+229+29)

#Hospital discharges in NJ of ovarian cancer patients, 2009

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

#Hospital discharges in NJ of ovarian cancer patients, 2009

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

[VSJO 13]

Can reconstruct tight bounds on rest of data

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

Multiple Release problem

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

- Multiple releases can lead to database reconstruction
  - *Death by a thousand cuts*

Motivation: Real world use cases

- Releasing Census Data

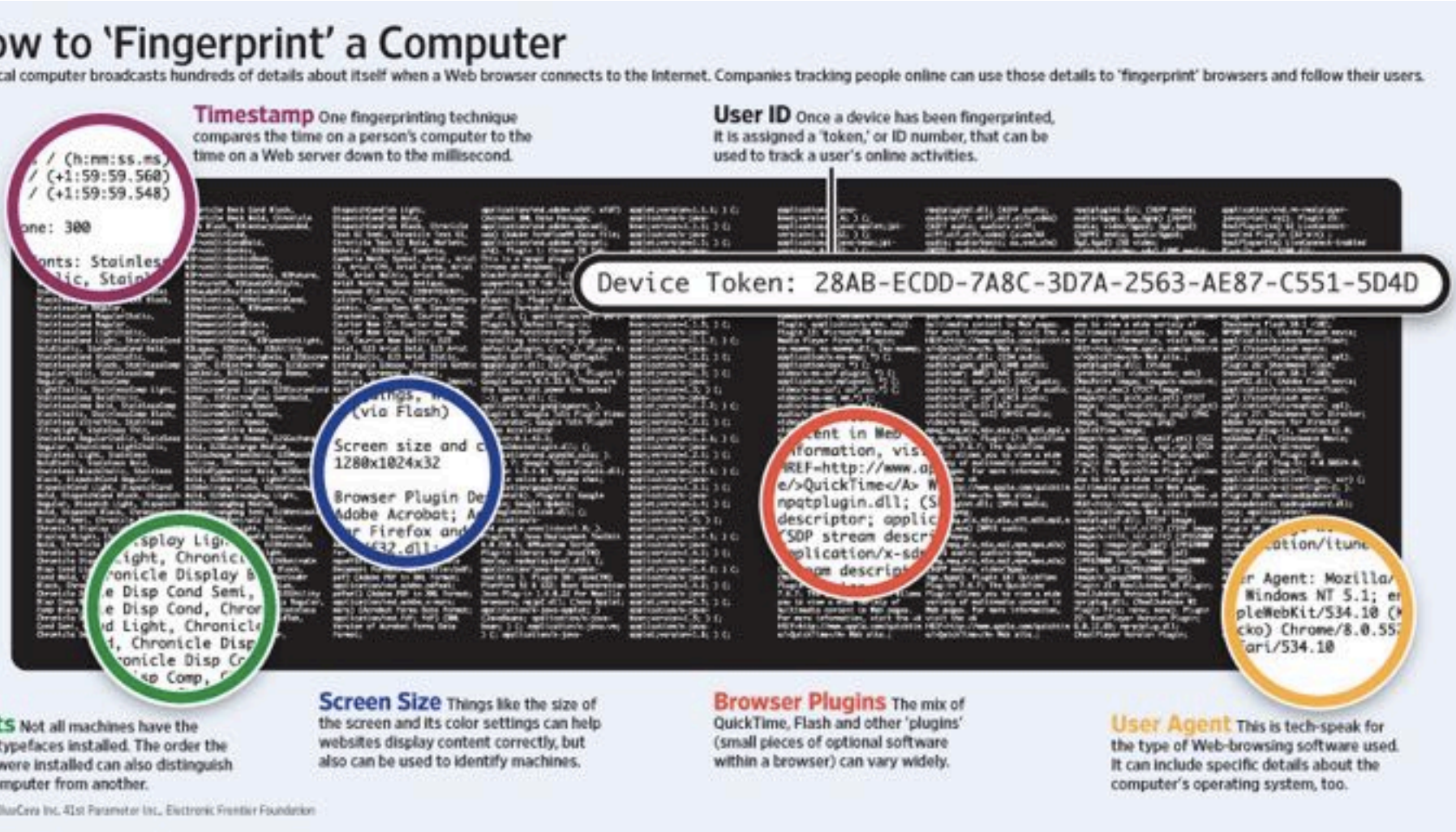
- Healthcare Analytics

- **Web data collection**

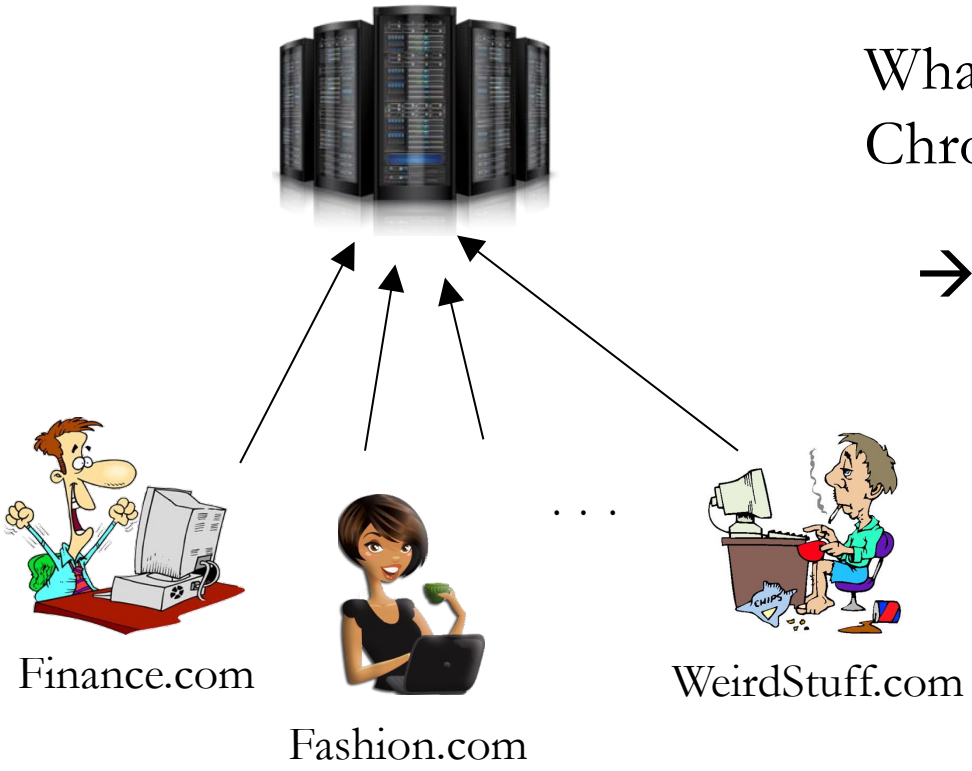
A dilemma

- 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

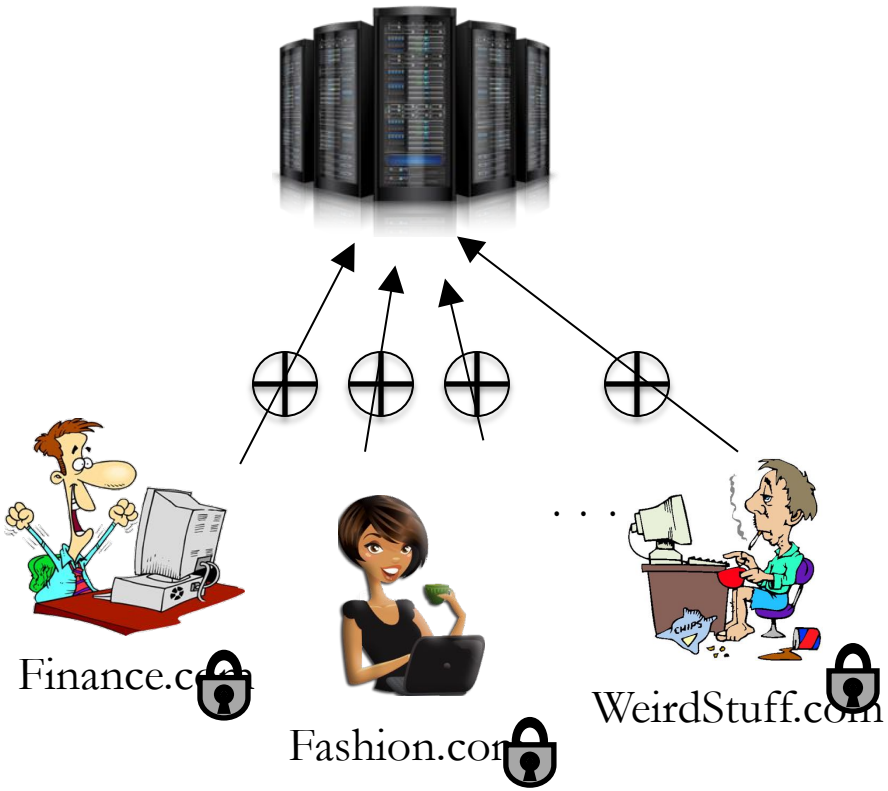


What are the *frequent* unexpected Chrome homepage domains?

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

Erlingsson, Pihur, Korolova, "RAPPOR: Randomized Aggregatable Privacy Preserving Ordinal Response", CCS 2014

Why privacy is needed?



Liability (for server)

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

Challenge

- Need to constantly collect/analyze/release data
  - Aggregate data collected may not reveal something sensitive about you … but over time, your presence may be revealed
  - *Death by a thousand cuts*

- Data collected about individuals can be linked with other external data
  - Door opening + your face on camera = you entered a room

Need a way to reason about privacy

- Data analytics … with *provable* privacy guarantees.
  - Define privacy … mathematically.
  - Explore and quantify privacy *vs* utility tradeoffs.
  - Build systems with provable privacy guarantees.

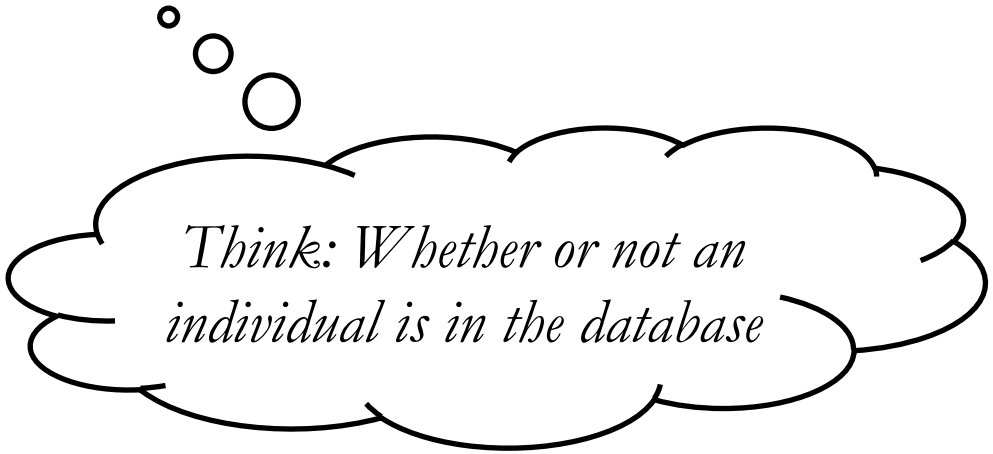
DIFFERENTIAL PRIVACY

Outline

- Differential Privacy Definition
- Basic Algorithms
  - Laplace Mechanism and Sensitivity
  - Randomized Response
- Composition Theorems

Defining Privacy: Differential Privacy

The output of an algorithm should be *insensitive* to adding or removing a record from the database.

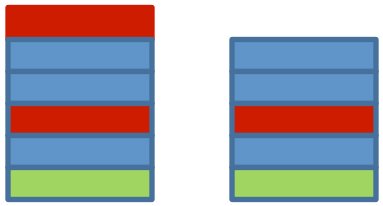


Dwork, McSherry, Nissim, Smith. "Calibrating Noise to Sensitivity for Private Data", TCC 2006,
Gödel Prize 2017

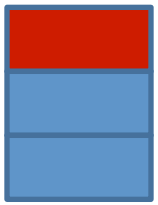
Differential Privacy

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

If algorithm A satisfies differential privacy then

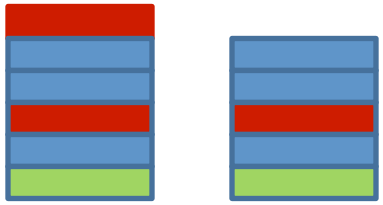
$$\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} < \exp(\varepsilon) \quad (\varepsilon > 0)$$

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

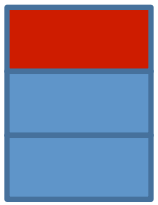
Why pairs of datasets *that differ in one row*?

For every pair of inputs that differ in one row

For every output …



D_1 D_2

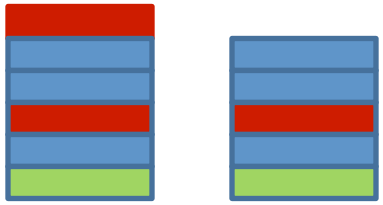


O

Simulate the presence or absence of a single record

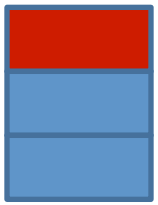
Why *all* pairs of datasets …?

For every pair of inputs that differ in one row



D_1 D_2

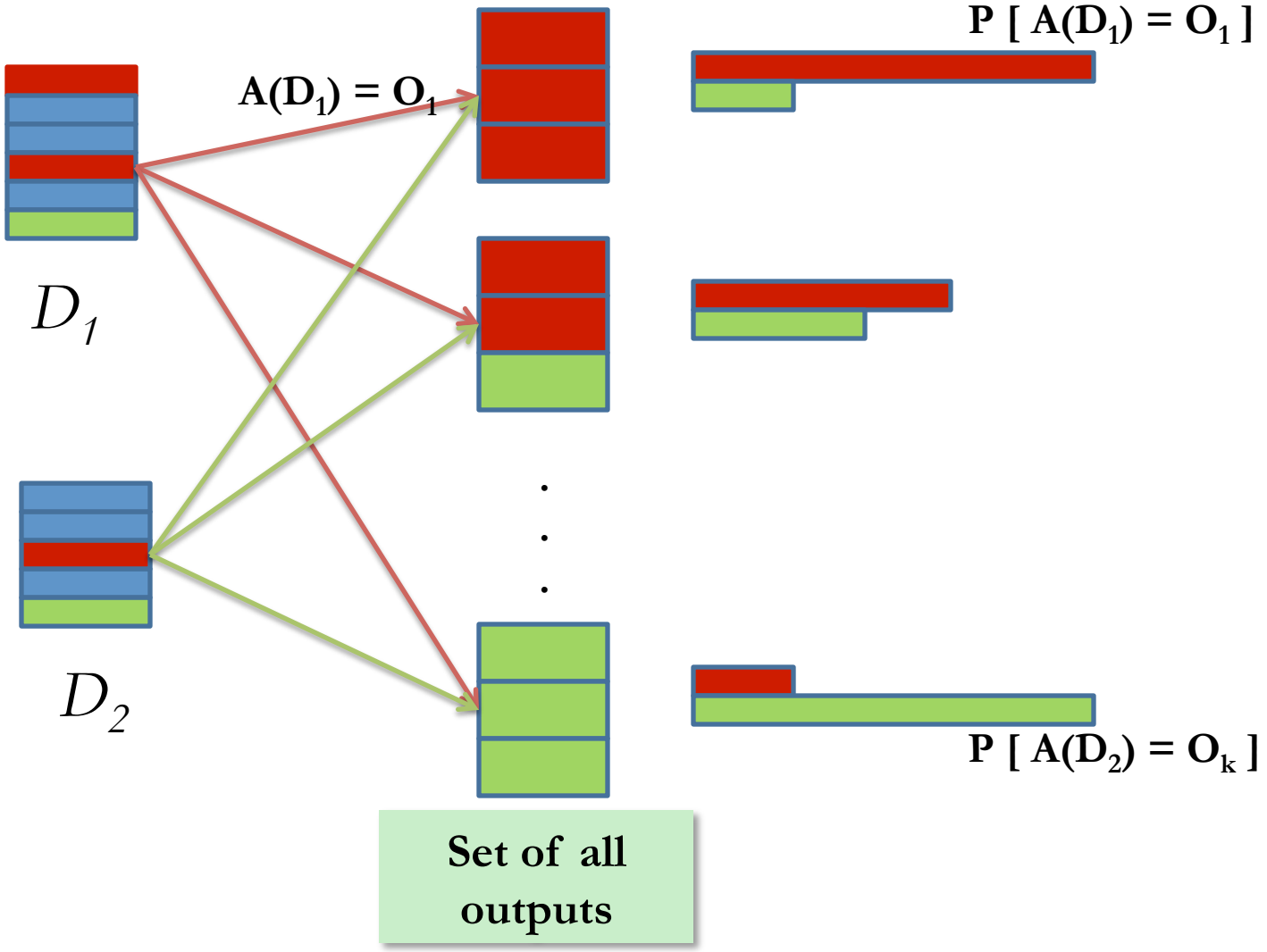
For every output …



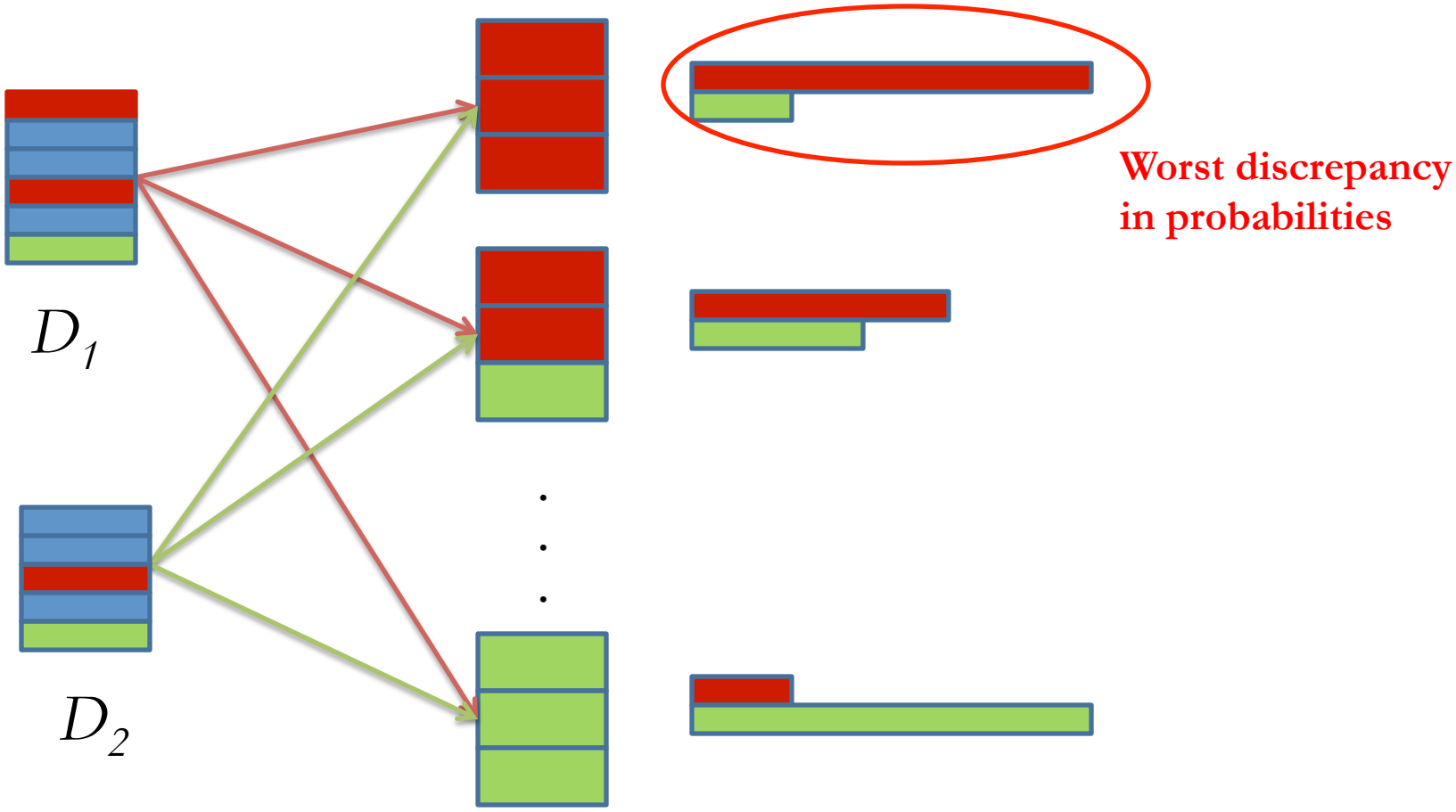
O

Guarantee holds no matter what the other records are.

Why *all* outputs?



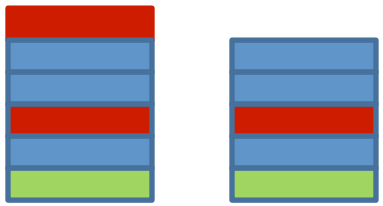
Should not be able to distinguish whether input was D_1 or D_2 no matter what the output



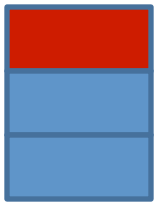
Privacy Parameter ε

For every pair of inputs that differ in one row

For every output …



D_1 D_2



O

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

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

Outline

- Differential Privacy

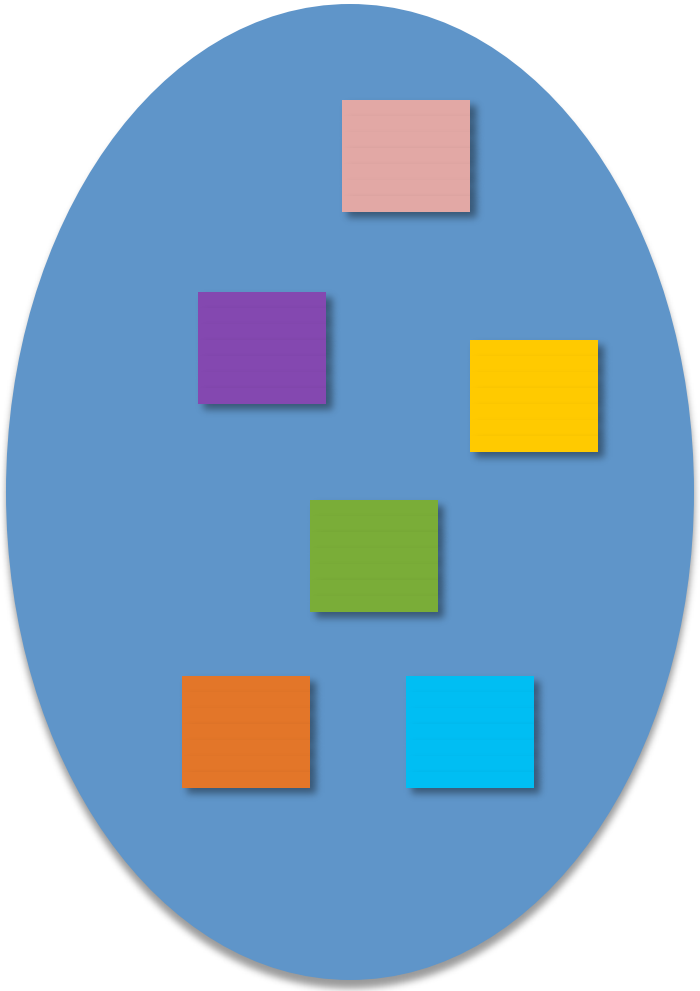
- Basic Algorithms
  - Laplace Mechanism & sensitivity
  - Randomized Response

- Composition Theorems

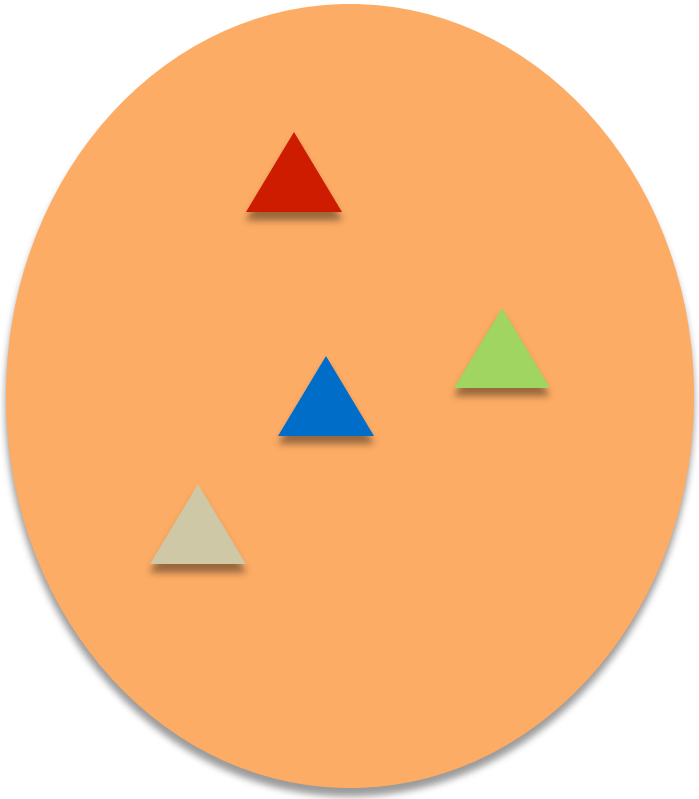
Can deterministic algorithms satisfy differential privacy?

Non trivial deterministic algorithms do not satisfy differential privacy

Space of all inputs

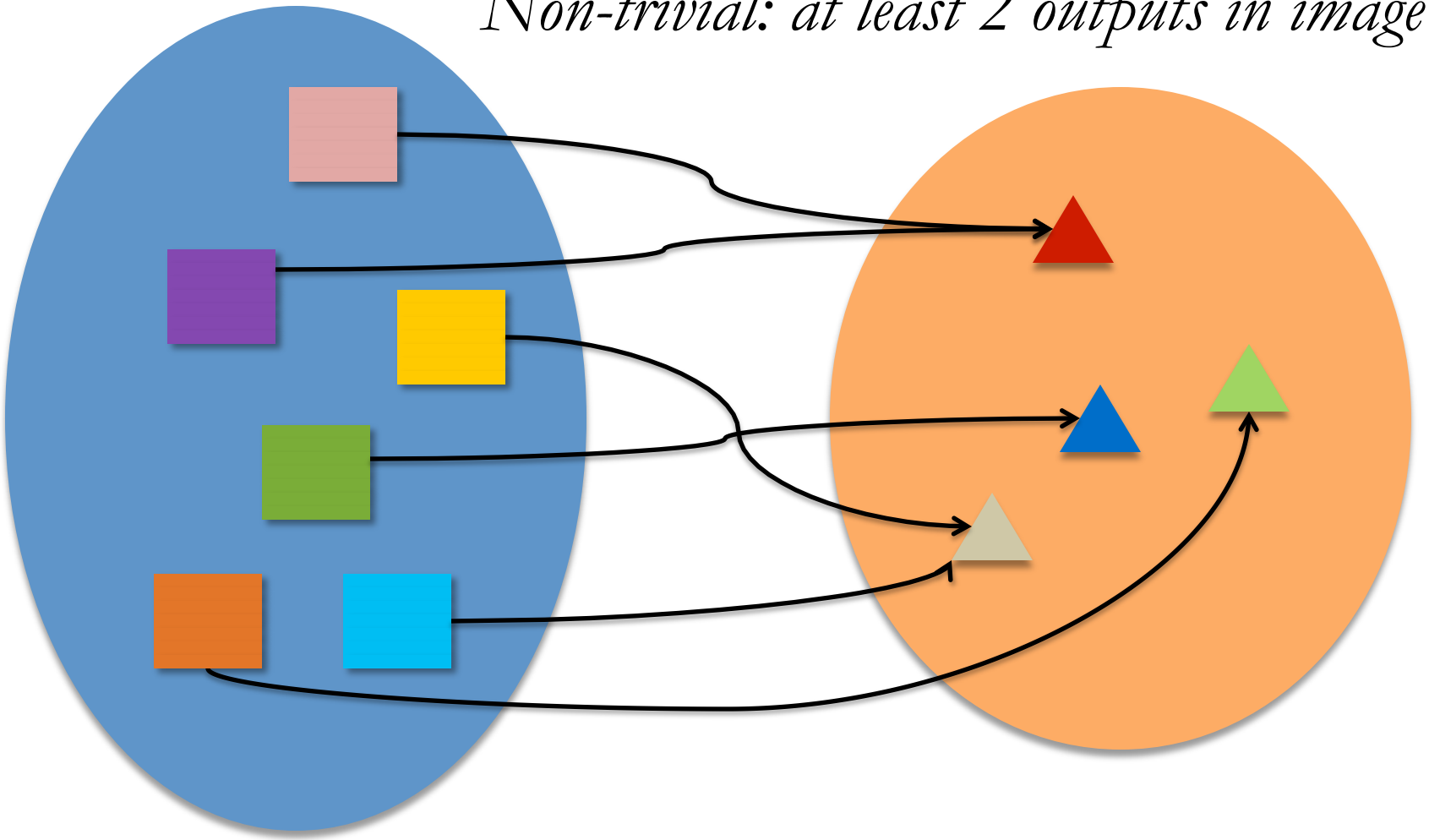


Space of all outputs

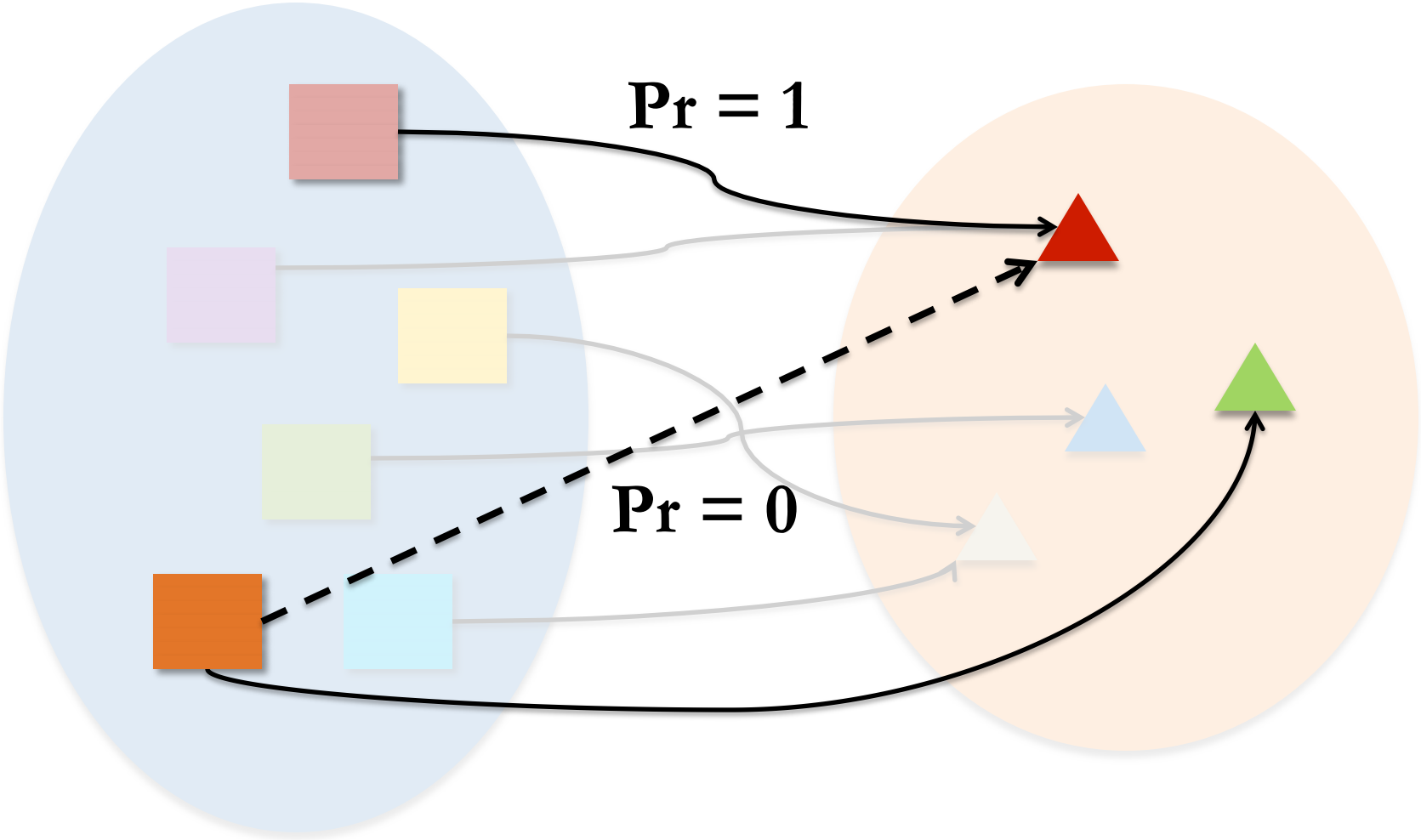


Non-trivial deterministic algorithms do not satisfy differential privacy

Non-trivial: at least 2 outputs in image



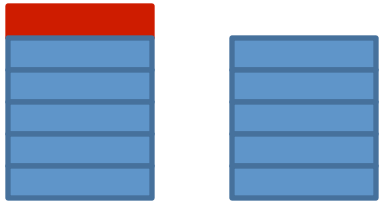
There exist two inputs that differ in one entry mapped to different outputs.



Random Sampling …

… also does not satisfy differential privacy

Input



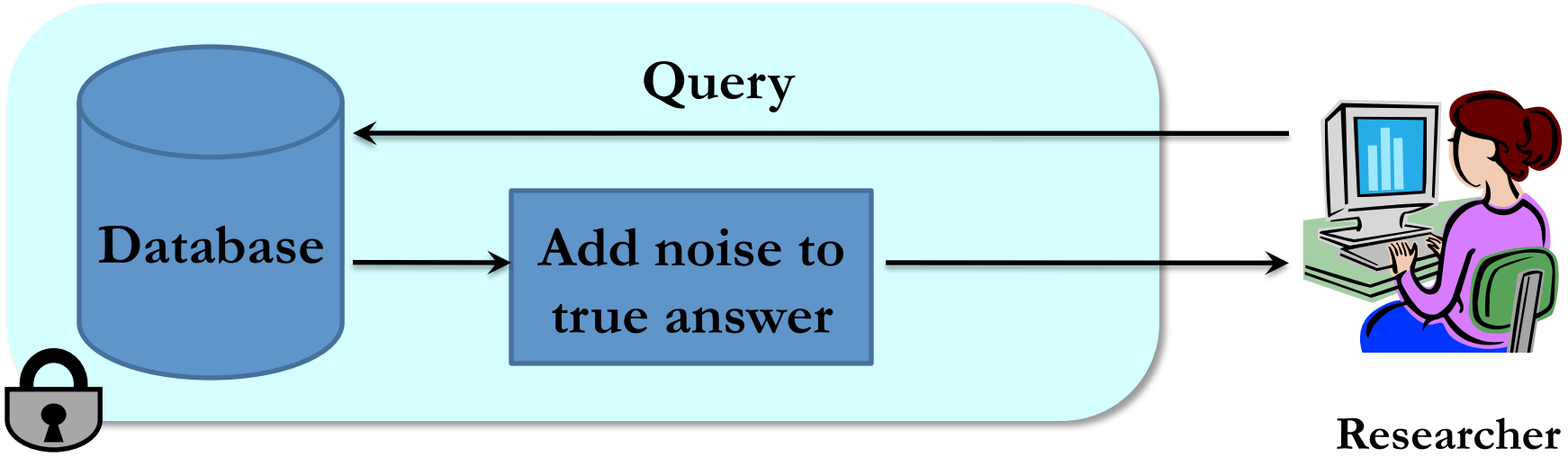
D_1 D_2

Output

O

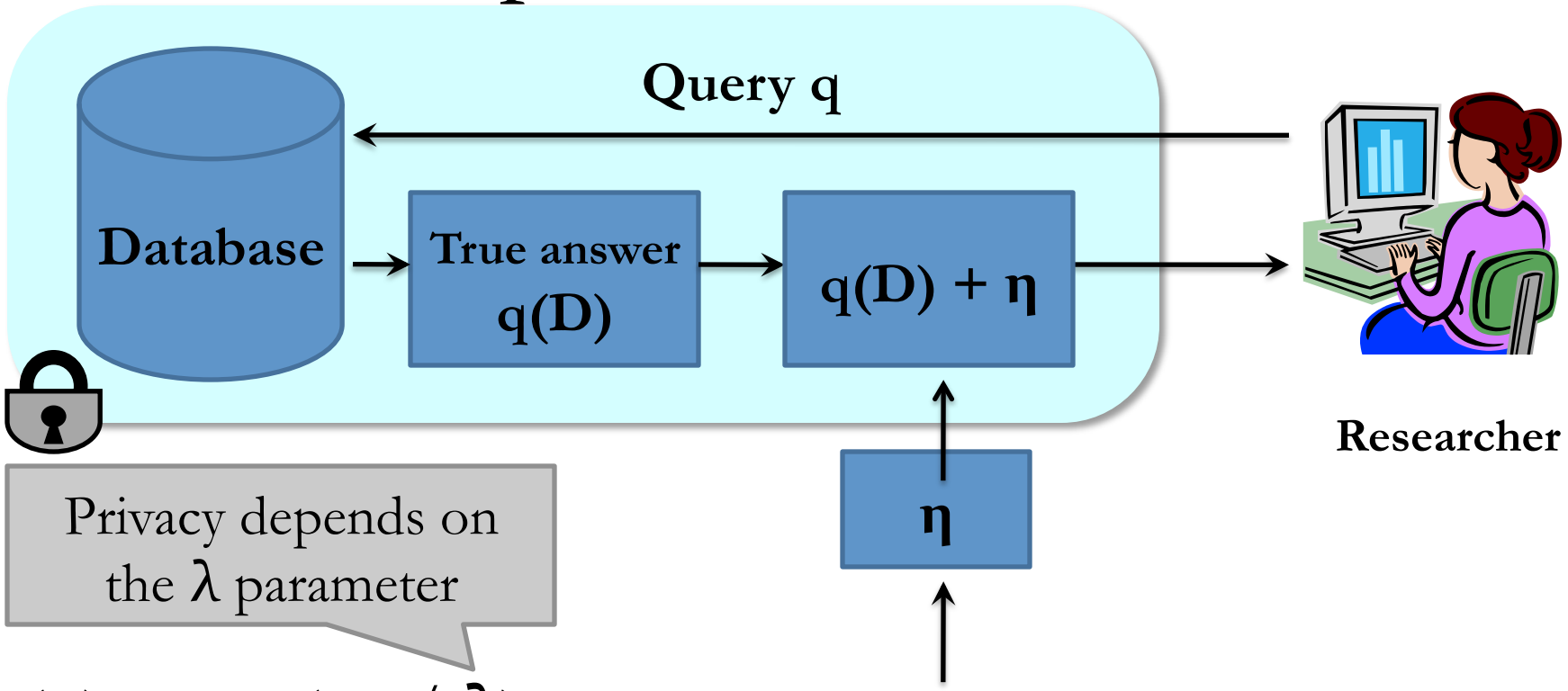
$\Pr[D_2 \rightarrow O] = 0$ implies $\dfrac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]} = \infty$

Output Randomization



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

Laplace Mechanism

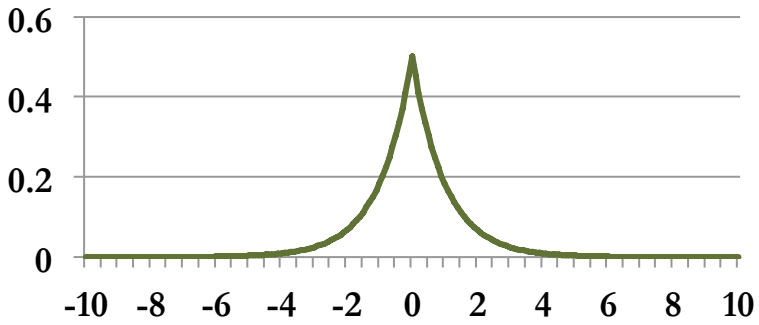


Privacy depends on the λ parameter

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

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

Laplace Distribution – Lap(λ)



How much noise for privacy?

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

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

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

$$\lambda = S/\varepsilon$$

Sensitivity: COUNT query

- Number of people having disease
- Sensitivity = 1

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

D

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

Sensitivity: SUM query

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

- Sensitivity = b

Privacy of Laplace Mechanism

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

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

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

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

Utility of Laplace Mechanism

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

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

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

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

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

Warner, "Randomized Response" JASA 1965

Differential Privacy Analysis

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

- Consider some output O

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

Utility Analysis

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

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

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

Sampling **Variance due to coin flips**

Laplace Mechanism vs Randomized Response

Privacy

- Provide the same ε -differential privacy guarantee

- Laplace mechanism assumes data collector is trusted
  - Like in the case of US Census Bureau
- Randomized Response does not require data collected to be trusted
  - Like in the Google Chrome case
  - Also called a *Local* Algorithm, since each record is perturbed

Laplace Mechanism vs Randomized Response

Utility

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

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

Outline

- Differential Privacy
- Basic Algorithms
  - Laplace Mechanism & sensitivity
  - 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

- 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

Dinur, Nissim, "Revealing information while preserving privacy", PODS 2003

Sequential Composition

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

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

Privacy as Constrained Optimization

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

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

Parallel Composition

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

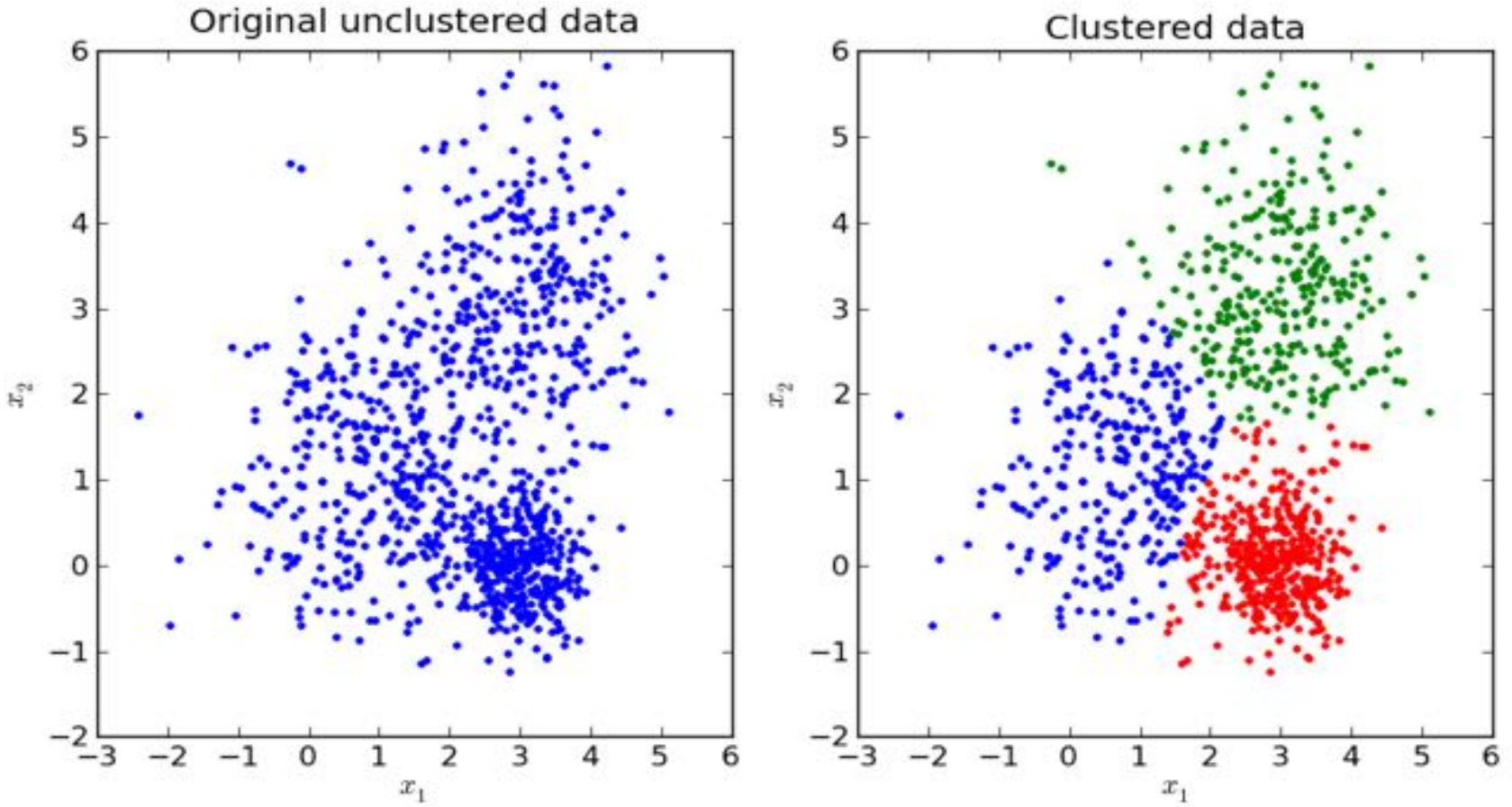
  then running all k algorithms in "parallel" satisfies ε-differential privacy with $\varepsilon = \max\{\varepsilon_1, ..., \varepsilon_k\}$

Postprocessing

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

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

Case Study: K-means Clustering



Blum, Dwork, McSherry, Nissim, "Practical privacy: the SuLQ framework", PODS 2005

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

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

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

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

Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

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

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

  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster

Differentially Private Kmeans

Exercise: Which of these steps expends privacy budget?

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

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

Differentially Private Kmeans

What is the sensitivity?

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

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

  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster

1

Domain size

Differentially Private Kmeans

- Suppose we fix the number of iterations to T

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

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

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

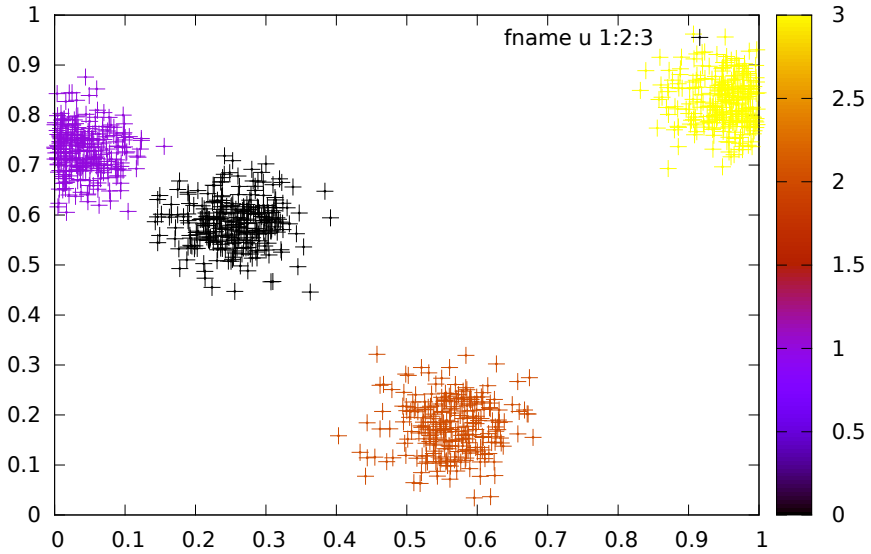
Laplace(2T/ ε)

3. Compute noisy sums of points in each cluster

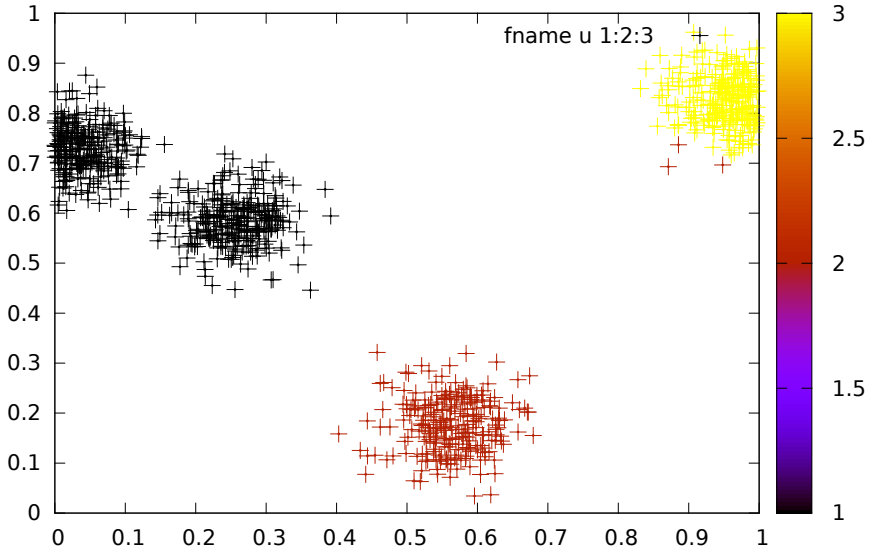
Laplace(2T |dom|/ ε)

Results (T = 10 iterations, random initialization)

Original Kmeans algorithm



Laplace Kmeans algorithm



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

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

Summary

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

- Building blocks
  - Laplace mechanism and randomized response

- Composition rules help build complex algorithms using building blocks