CS 528 (Fall 2021)
Data Privacy & Security

Yuan Hong
Department of Computer Science
Illinois Institute of Technology

Chapter 3
Differential Privacy

RECAP

- **Anonymization or De-Identification (Input Perturbation)**
  - Linkage attacks
  - Homogeneity attacks
  - Background knowledge attacks
  - Skewness attacks
  - Similarity attacks

- **k-Anonymity, l-Diversity, t-Closeness**

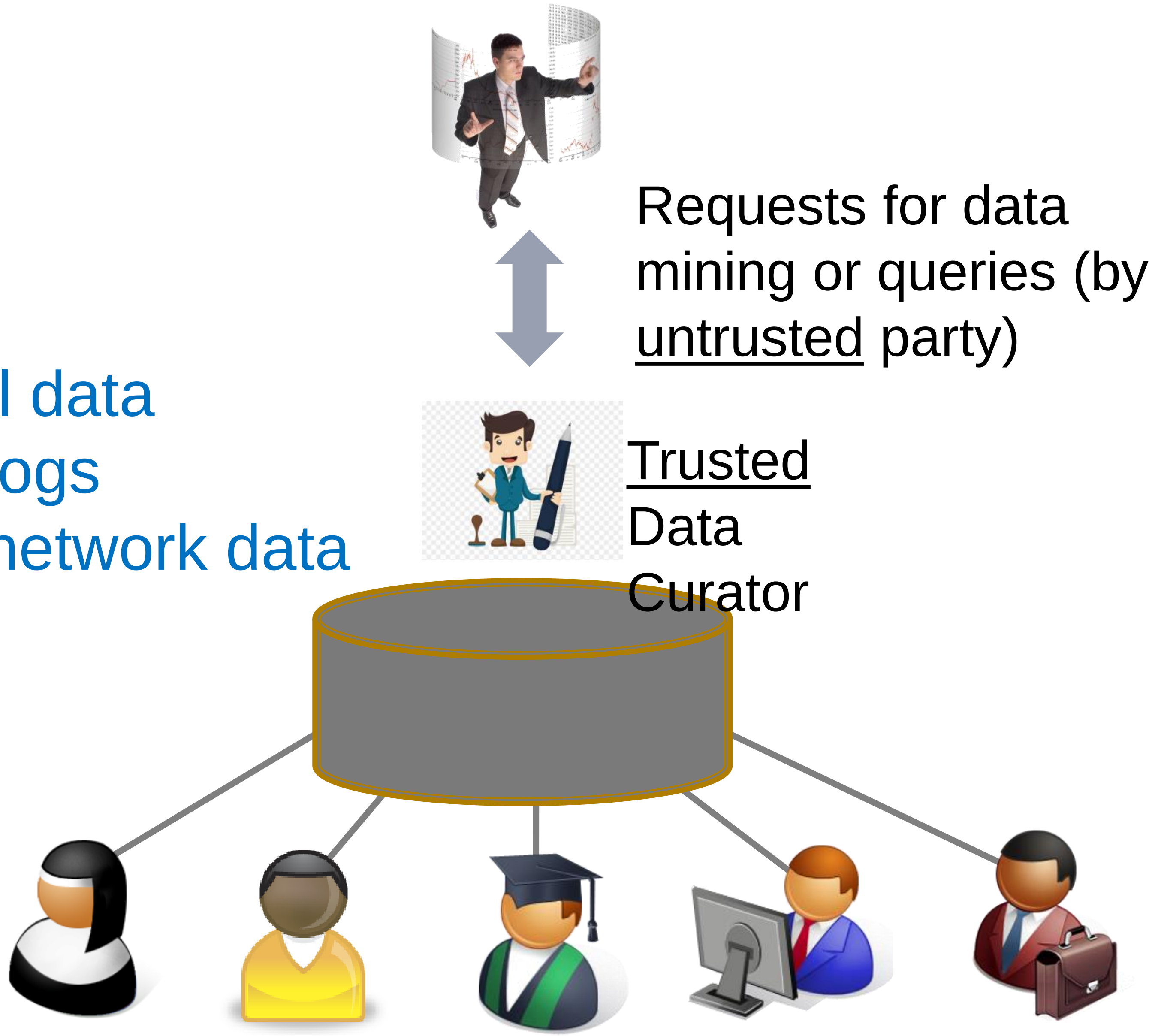
OUTLINE

Differential Privacy for Centralized Data

1. Threat Model and Architecture
2. Differential Privacy Definition
3. Basic Techniques
4. Composition Theorems
5. Other DP Variants

GENERAL SETTING (INTERACTIVE)

Medical data
Query logs
Social network data
…



STATISTICAL ANALYSES 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/Browsi ng	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

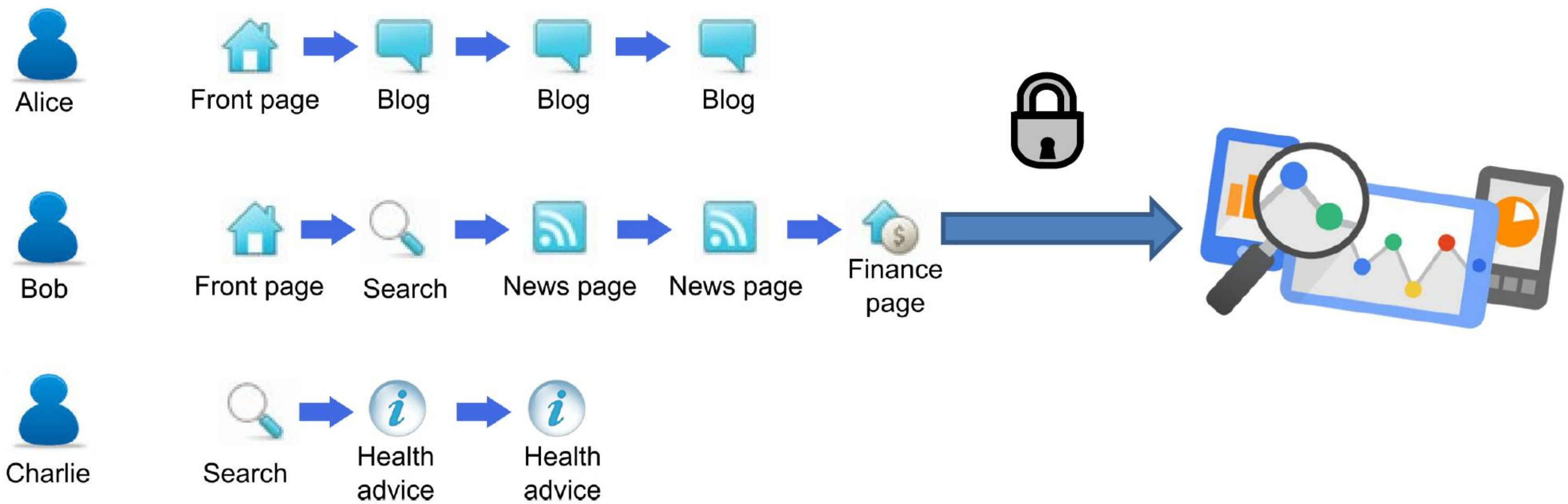
DIFFERENTIAL PRIVACY

- Promise: an individual will not be affected, adversely or otherwise, by allowing his/her data to be used in any study or analysis, no matter what other studies, datasets, or information sources, are available.
  - Protection against arbitrary background knowledge

- Paradox: learning nothing about an individual while learning useful statistical information about a population.

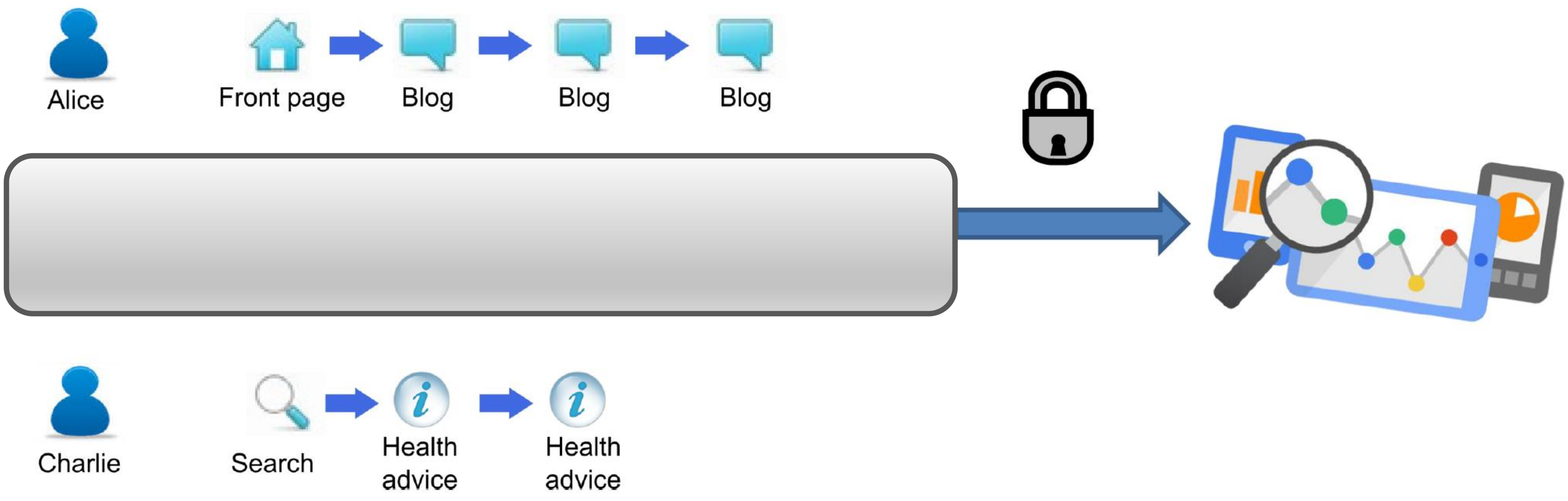
DIFFERENTIAL PRIVACY

- Statistical outcome is <u>indistinguishable</u> regardless whether a particular user (record) is included in the data or not.



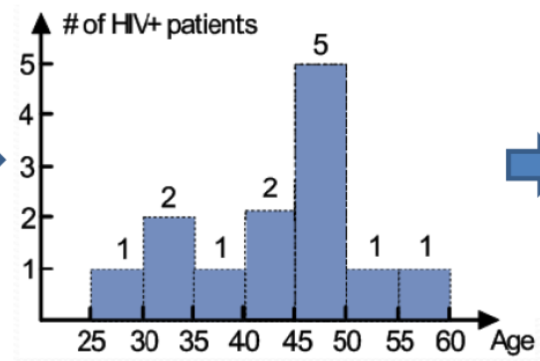
DIFFERENTIAL PRIVACY

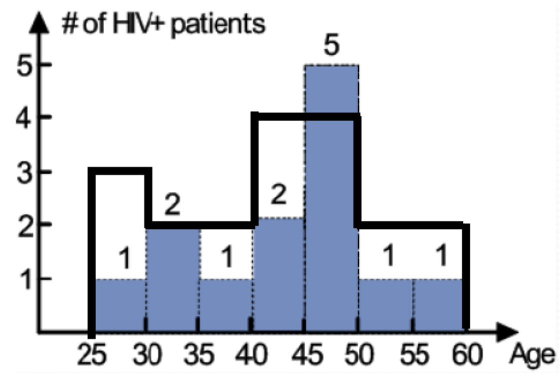
- Statistical outcome is indistinguishable regardless whether a particular user (record) is included in the data or not.



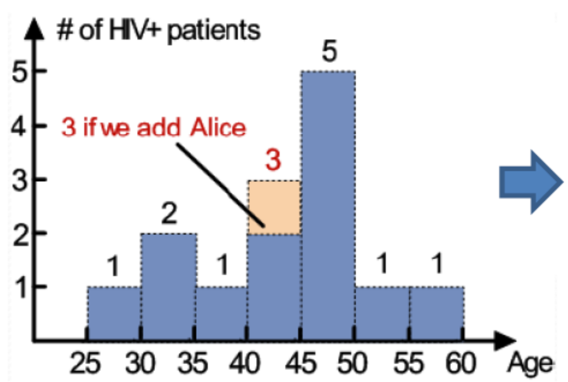
DIFFERENTIAL PRIVACY: AN EXAMPLE

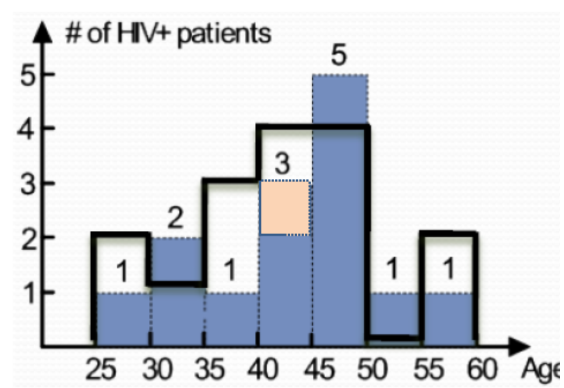
Name	Age	HIV+
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…





Name	Age	HIV+
Alice	43	Y
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…





Original records

Original histogram

Perturbed histogram with differential privacy

OUTLINE

Differential Privacy for Centralized Data

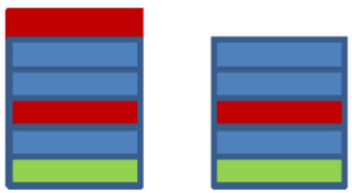
1. Threat Model and Architecture
2. Differential Privacy Definition
3. Basic Techniques
4. Composition Theorems
5. Non-Interactive DP Mechanisms

DIFFERENTIAL PRIVACY (PROBABILISTIC)

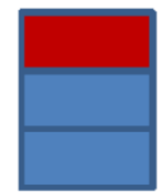
[Dwork ICALP 2006]

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

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

WHY PAIRS OF DATASETS THAT DIFFER IN ONE ROW/RECORD

For every pair of inputs that differ in one row

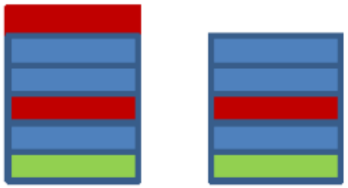
For every output …

Guarantee holds no matter what the other records are.

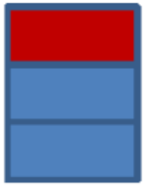
WHY ALL PAIRS OF DATASETS

For every pair of inputs that differ in one row

For every output …



D_1 D_2

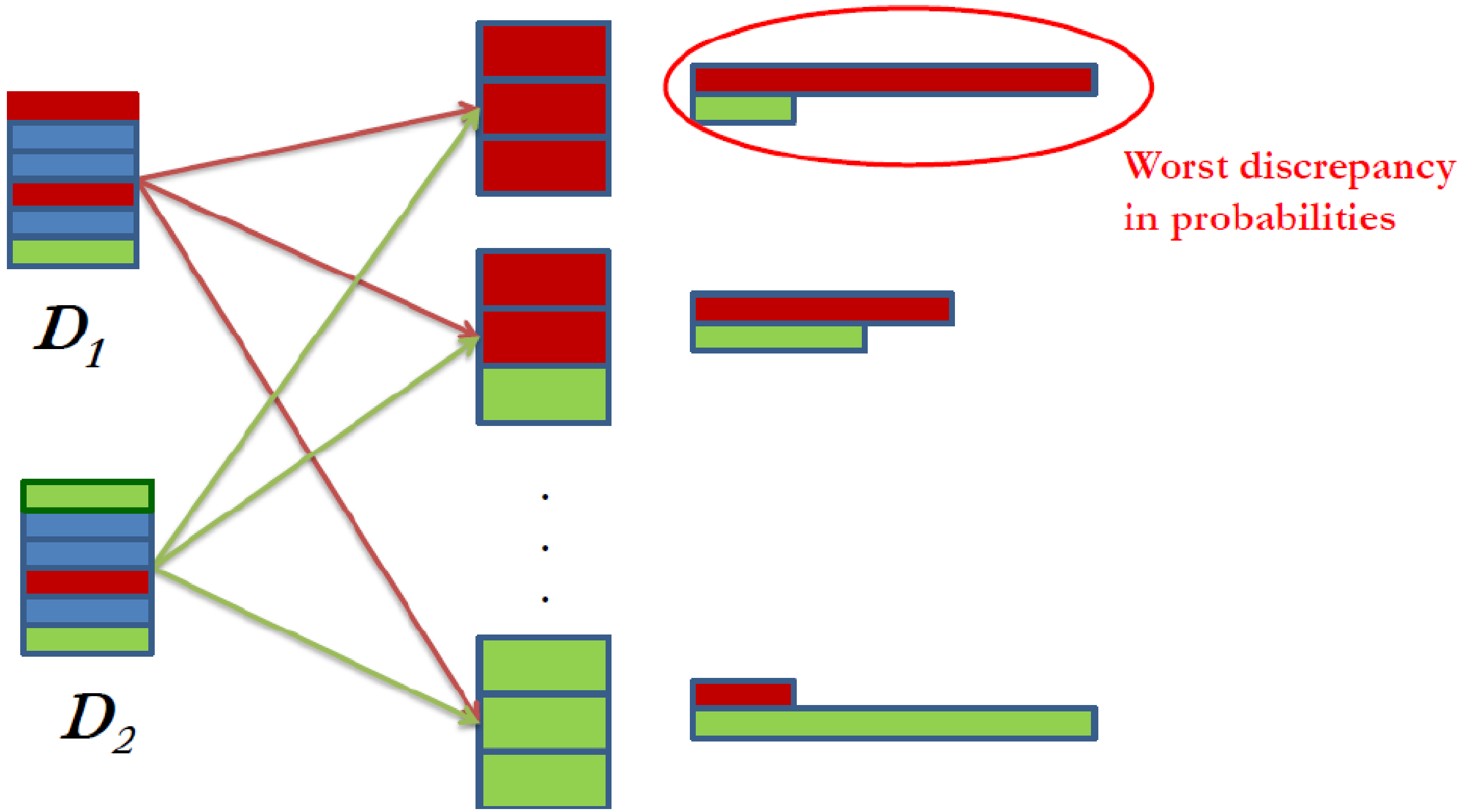


O

Simulate the presence or absence of a single record

WHY ALL OUTPUTS

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



PRIVACY PARAMETER ε

- Controls the degree to which D1 and D2 can be distinguished. Smaller the ε, more privacy (less utility)

For every pair of inputs that differ in one row

For every output …

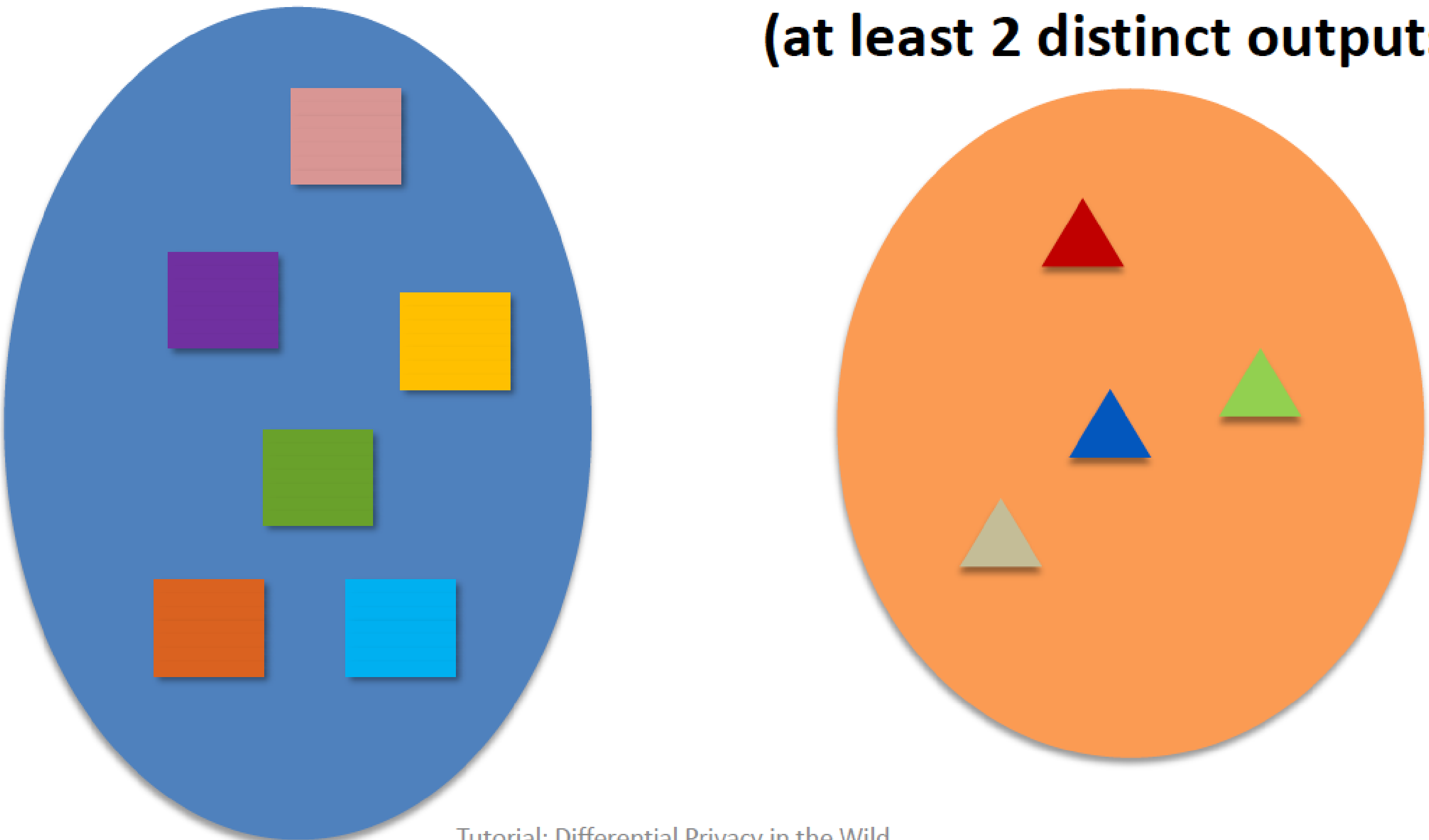
D_1 D_2 O

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

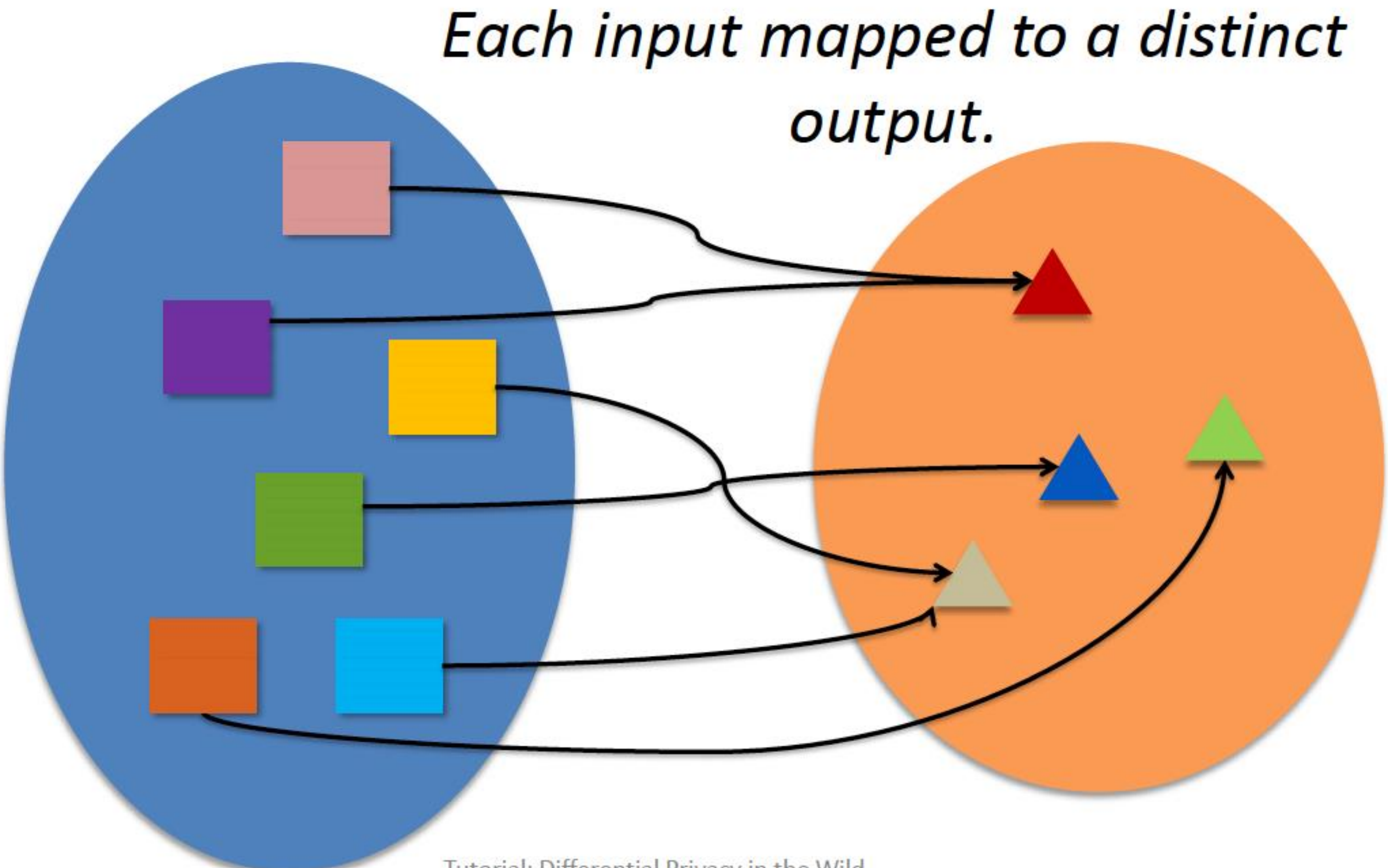
NON TRIVIAL DETERMINISTIC ALGORITHM DOES NOT SATISFY DIFFERENTIAL PRIVACY

Space of all inputs

Space of all outputs (at least 2 distinct outputs)



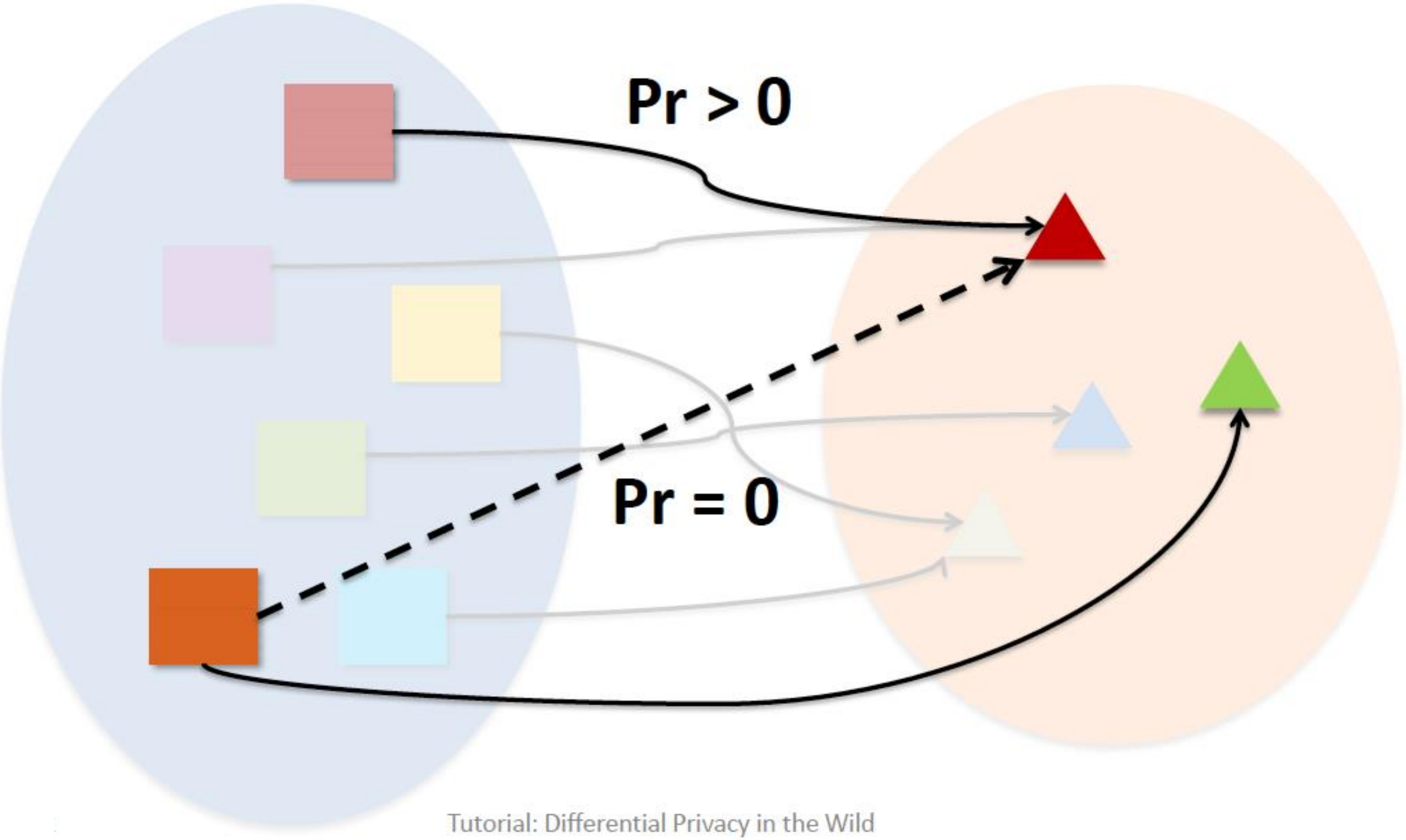
NON TRIVIAL DETERMINISTIC ALGORITHM DOES NOT SATISFY DIFFERENTIAL PRIVACY



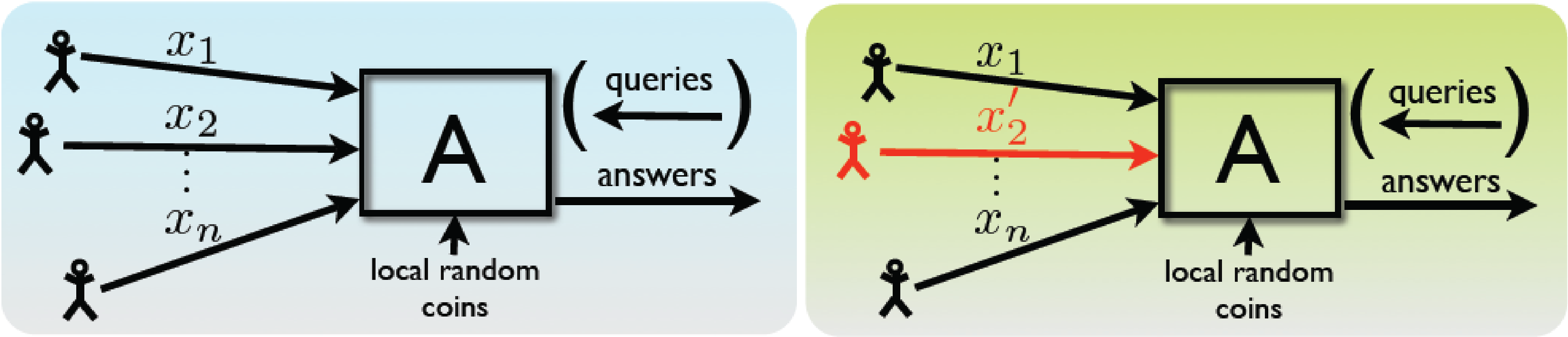
Tutorial: Differential Privacy in the Wild

NEIGHBORING INPUTS

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



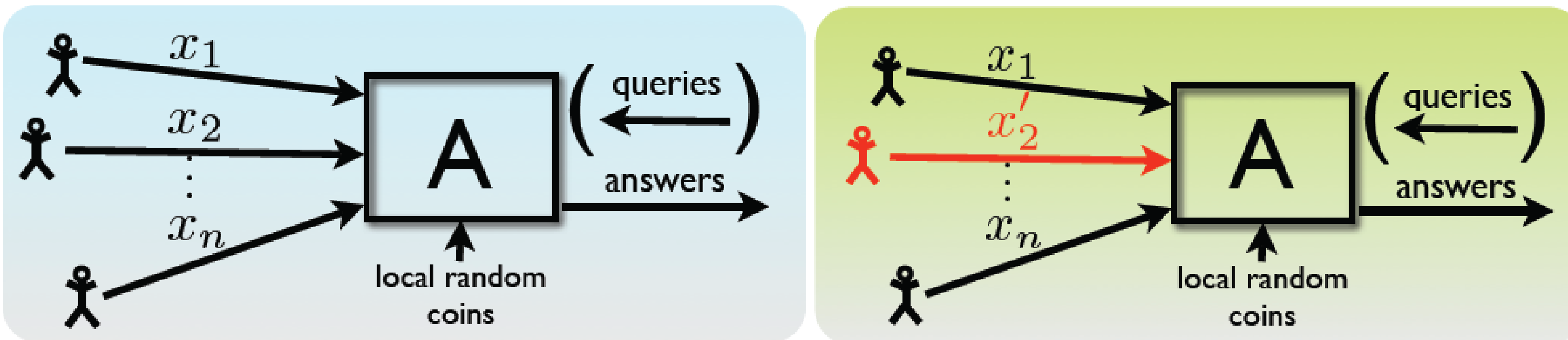
DIFFERENTIAL PRIVACY (INDISTINGUISHABILITY)



x' is a neighbor of x
if they differ in one row

From the released statistics, it is hard to tell which case it is.

CONT



For all neighboring databases D and D’, which differ in any arbitrary record,

For all possible subsets of output space: $S \subseteq \text{range}(A)$

$$\Pr[A(D) \in S] \leq e^{\varepsilon} \Pr[A(D') \in S]$$

RELAXED DIFFERENTIAL PRIVACY

Definition 2 ((ϵ, δ)-differential privacy). *A randomization algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy if for all neighboring inputs D and D' and any set of possible outputs S, we have $Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D') \in S] + \delta$ and $Pr[\mathcal{A}(D') \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D) \in S] + \delta$.*

Pr[A(D) є S]=0 or Pr[A(D') є S]=0

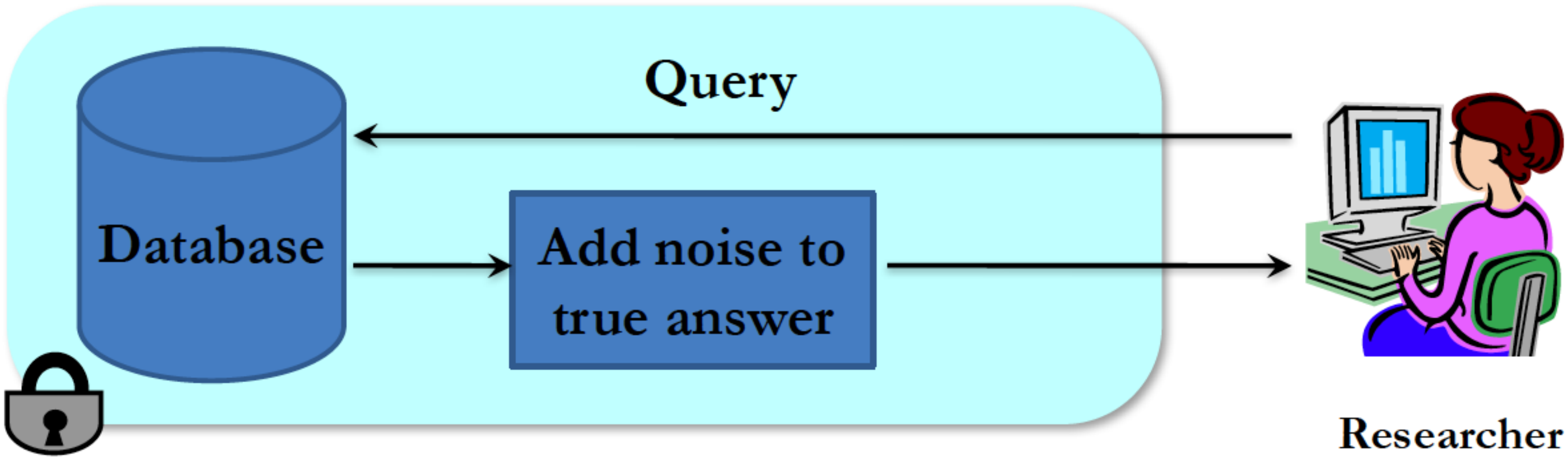
OUTLINE

Differential Privacy for Centralized Data

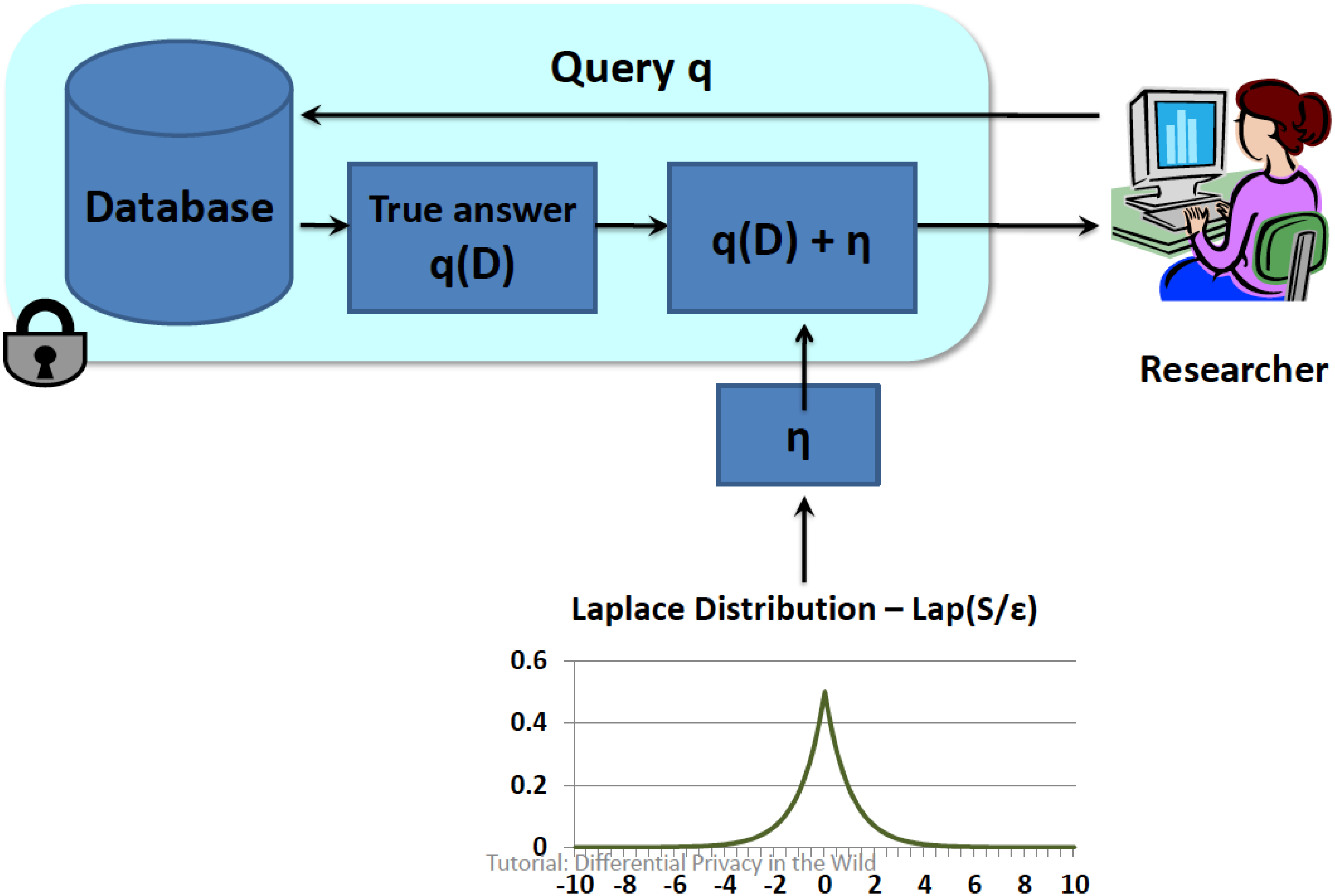
1. Threat Model and Architecture
2. Differential Privacy Definition
3. Basic Techniques
4. Composition Theorems
5. Other DP Variants

OUTPUT RANDOMIZATION

- Adding noise to answers (of queries) such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers

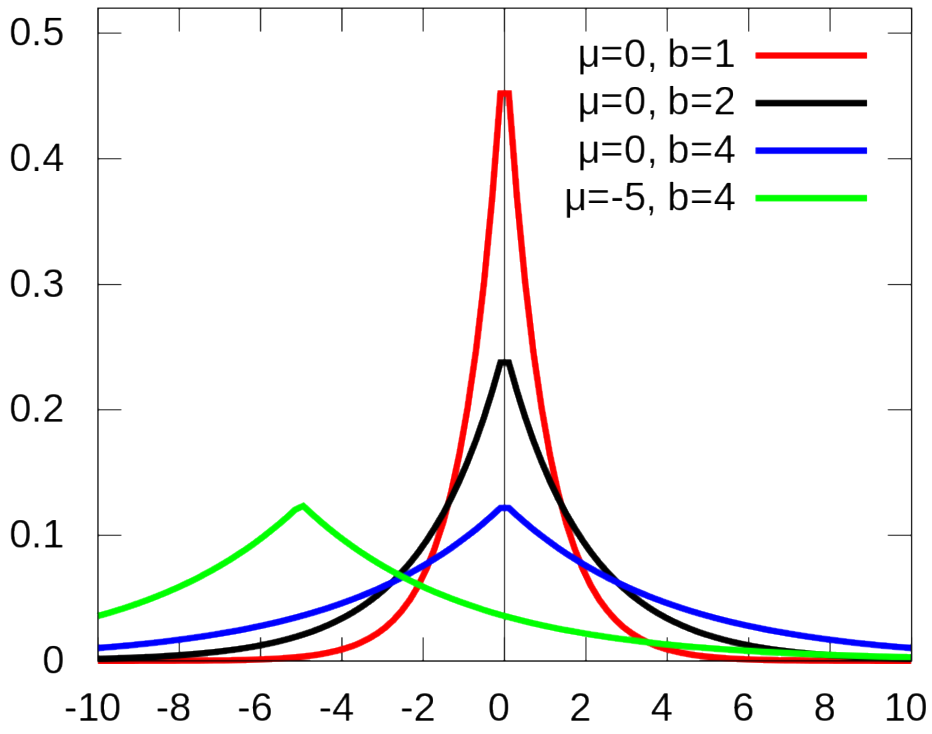


LAPLACE MECHANISM

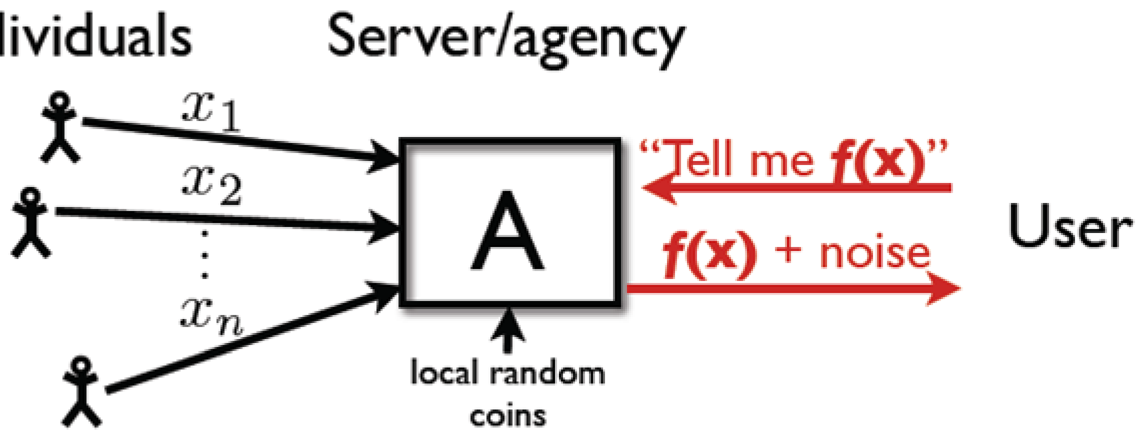


LAPLACE DISTRIBUTION

- PDF $= \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu - x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x - \mu}{b}\right) & \text{if } x \geq \mu \end{cases}$
- Denoted as Lap(b) when u=0
- Mean u
- Variance $2b^2$



GLOBAL SENSITIVITY

- 

- Global Sensitivity:

$$GS_f = max\|f(x) - f(x')\|_1$$

x, x' neighbors

Example: $GS_{avg} = \frac{1}{n}$

DIFFERENTIAL PRIVACY GUARANTEE

Theorem:

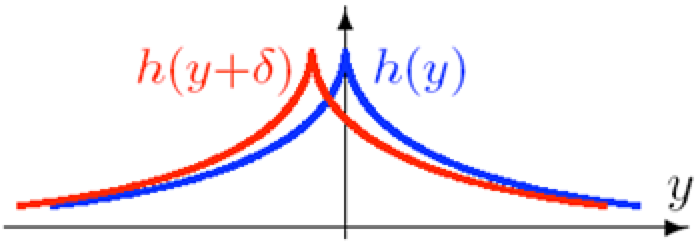
$$A(x) = f(x) + \mathrm{Lap}\left(\frac{GS_f}{\epsilon}\right) \text{ is } \epsilon\text{-DP}$$

- Intuition: add more noise when function is sensitive
- Smaller and/or larger sensitivity results in larger noise

PROOF

$$A(x) = f(x) + \mathrm{Lap}\left(\frac{GS_f}{\epsilon}\right) \text{ is } \epsilon\text{-DP}$$

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$



Sliding property of $\mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

$$\delta = f(x) - f(x') \leq \mathsf{GS}_f$$

EXAMPLE: COUNT QUERY

- Number of people having disease
- Sensitivity = 1
- Solution: 3+noise
  where noise is drawn from Lap(1/ε)
  Mean = 0
  Variance = 2/ε^2

D

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

UTILITY OF LAPLACE MECHANISM

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

- Error: E(true answer – noisy answer)^2

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

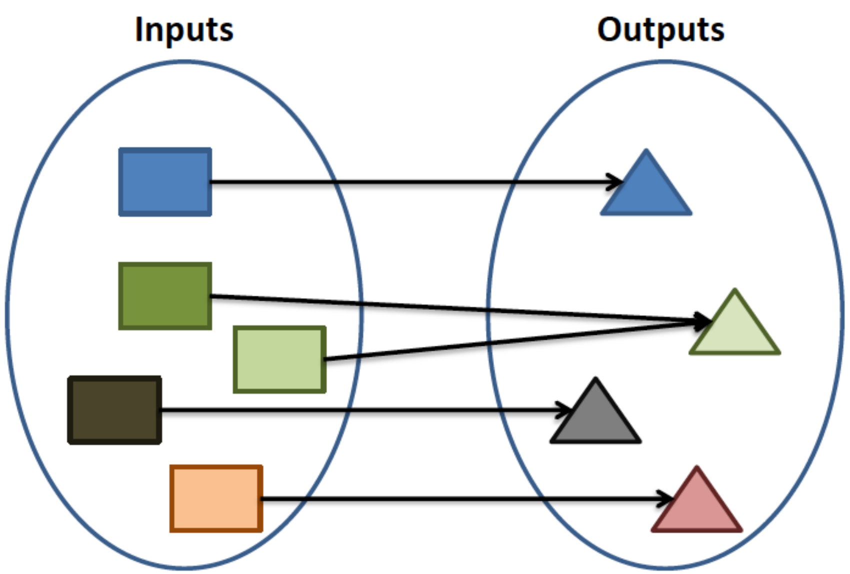
  $= 2*S(q)^2\ /\ \varepsilon^2$

Error bound: very unlikely the result has an error greater than a factor

EXPONENTIAL MECHANISM

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

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



How to construct a differentially private version of f?

EXPONENTIAL MECHANISM

- Scoring/utility function
  - ✓ w: Inputs x Outputs → R
- D: nationalities of a set of people
- f(D): most frequent nationality in D
- u(D, O) = #(D, O), the number of people with nationality O

EXPONENTIAL MECHANISM

Theorem *For a database D, output space R and a utility score function u : D×R → R, the algorithm A*

$$\Pr[A(D) = r] \propto \exp(\varepsilon \times u(D, r) / 2\Delta u)$$

satisfies ε-differential privacy, where Δu is the sensitivity of the utility score function

$$\Delta u = \max_{r \,\&\, D,D'} |u(D, r) - u(D', r)|$$

Approximately $\Pr[A(D)=r]/\Pr[A(D')=r] <= \exp(\epsilon)$

PRIVACY OF EXPONENTIAL MECHANISM

$$\frac{\Pr[\mathcal{M}_E(x,u,\mathcal{R})=r]}{\Pr[\mathcal{M}_E(y,u,\mathcal{R})=r]} = \frac{\left(\frac{\exp(\frac{\varepsilon u(x,r)}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}\right)}{\left(\frac{\exp(\frac{\varepsilon u(y,r)}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(y,r')}{2\Delta u})}\right)}$$

$$= \left(\frac{\exp(\frac{\varepsilon u(x,r)}{2\Delta u})}{\exp(\frac{\varepsilon u(y,r)}{2\Delta u})}\right)\cdot\left(\frac{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(y,r')}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}\right)$$

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

$$\cdot\left(\frac{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(y,r')}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}\right)$$

$$\le \exp\left(\frac{\varepsilon}{2}\right)\cdot\exp\left(\frac{\varepsilon}{2}\right)\cdot\left(\frac{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}\right)$$

$$= \exp(\varepsilon).$$

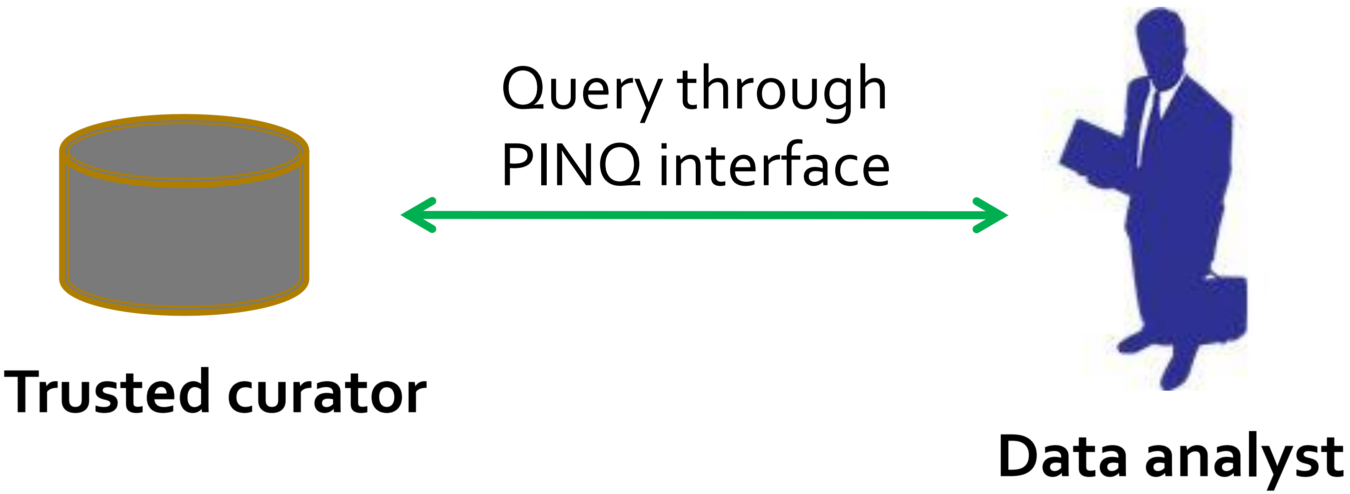
UTILITY OF EXPONENTIAL MECHANISM

- Can give strong utility guarantees, as it discounts outcomes exponentially based on utility score

PRIVACY INTEGRATED QUERIES (PINQ)

- Language for writing differentially-private data analyses
- Language extension to .NET framework
- Provides a SQL-like interface for querying data
- Goal: Hopefully, non-privacy experts can perform privacy-preserving data analytics

SCENARIO


Query through
PINQ interface
Trusted curator
Data analyst

EXAMPLE

```
static void Main(string[] args)
{
    var source = Enumerable.Range(1, 1000).AsQueryable();
    var agent = new PINQAgentBudget(1.0);

    var data = new PINQueryable<int>(source, agent);

    Console.WriteLine("count: " + data.NoisyCount(0.01));
    Console.WriteLine("count: " + data.NoisyCount(0.10));
    Console.WriteLine("count: " + data.NoisyCount(1.00));
}
```

OUTLINE

Differential Privacy for Centralized Data

1. Threat Model and Architecture
2. Differential Privacy Definition
3. Basic Techniques
4. Composition Theorems
5. Other DP Variants

WHY COMPOSITION?

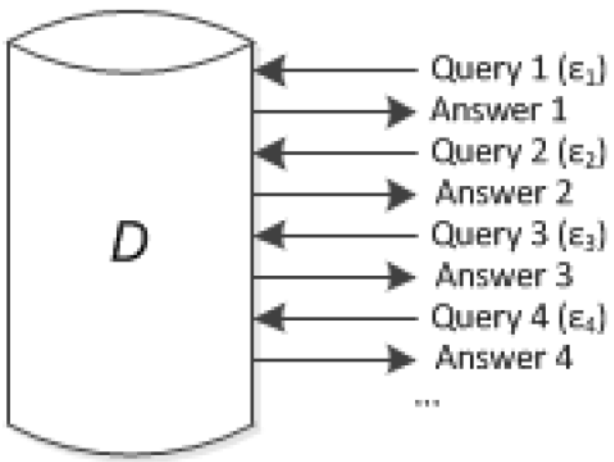
- Reasoning about privacy of a complex algorithm is hard

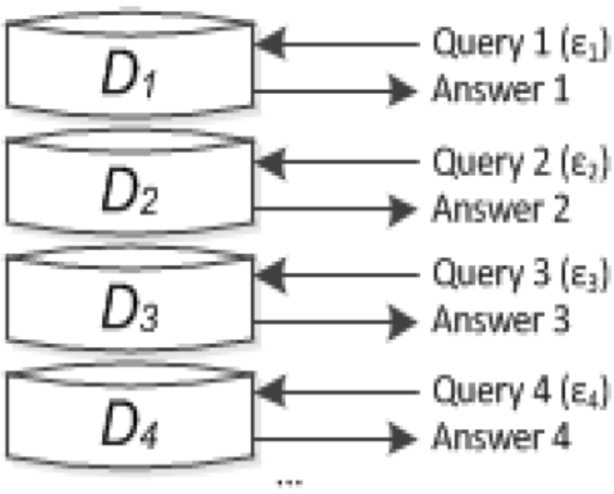
- Helps software design
  - If building blocks are proven to be private, it would be easy to reason about privacy of the complex algorithm built entirely using these building blocks.

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

COMPOSITION AND PRIVACY BUDGET





Sequential composition

$\sum_i \varepsilon_i$ –differential privacy

Parallel composition

$\max(\varepsilon_i)$–differential privacy

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 $\varepsilon=\varepsilon_1+\ldots+\varepsilon_k$

PARALLEL COMPOSITION

- If M_1, M_2, …, M_k are algorithms that access a private disjoint database D_1, D_2, …, D_k such that each M_i satisfies ε_i -differential privacy,

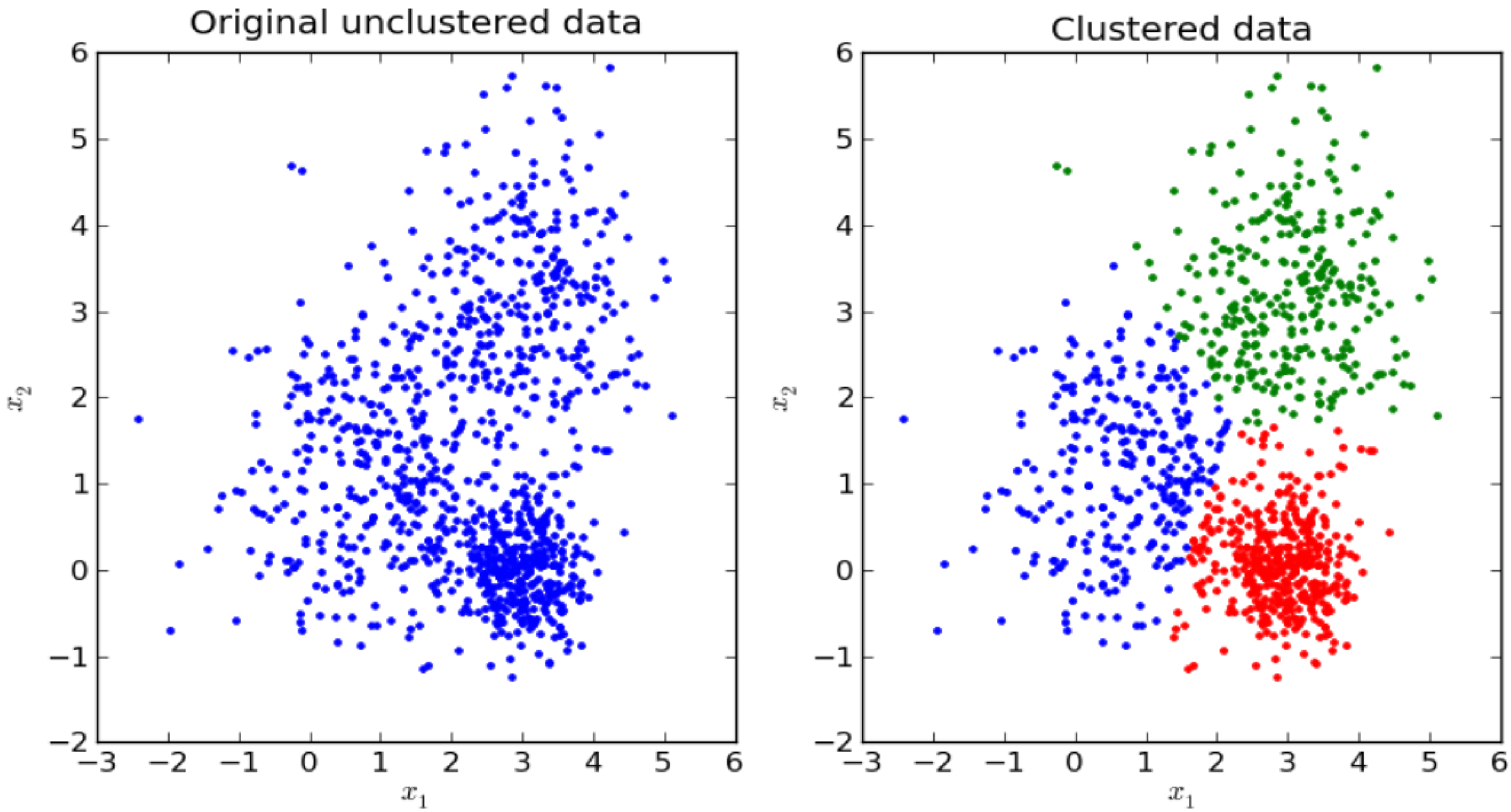
  then the combination of their outputs satisfies ε -differential privacy with $\varepsilon = \max\{\varepsilon_1,\ldots,\varepsilon_k\}$

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



K-MEANS

Partition a set of points x_1, x_2, …, x_k into k clusters S_1, …, S_k such that the following the minimized:

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

Mean of the cluster S_i

K-MEANS

Algorithm:

- Initialize a set of k centers
- Repeat
    - Assign each point to its nearest center
- Until convergence
- Output final set of k clusters

DIFFERENTIALLY PRIVATE K-MEANS

- Suppose we fix the number of iterations to T
- In each iteration (given the 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 K-MEANS

- 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 the 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 K-MEANS

Exercise: Which of these steps expends privacy budget?

- In each iteration (given the 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 K-MEANS

Exercise: Which of these steps expends privacy budget?

- In each iteration (given the 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 cluste — YES

DIFFERENTIALLY PRIVATE K-MEANS

What is the sensitivity?

- In each iteration (given the set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluste
  - 3. Compute noisy sums of points in each cluster

1

Domain size

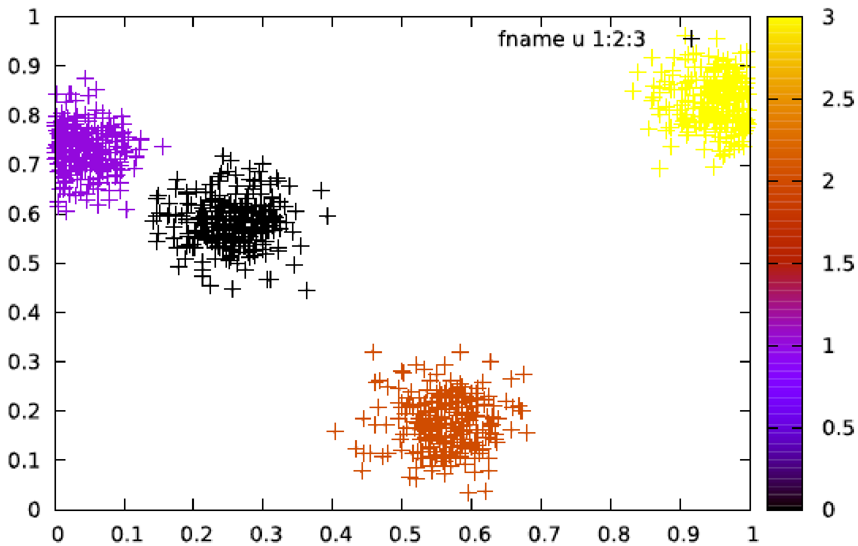
DIFFERENTIALLY PRIVATE K-MEANS

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

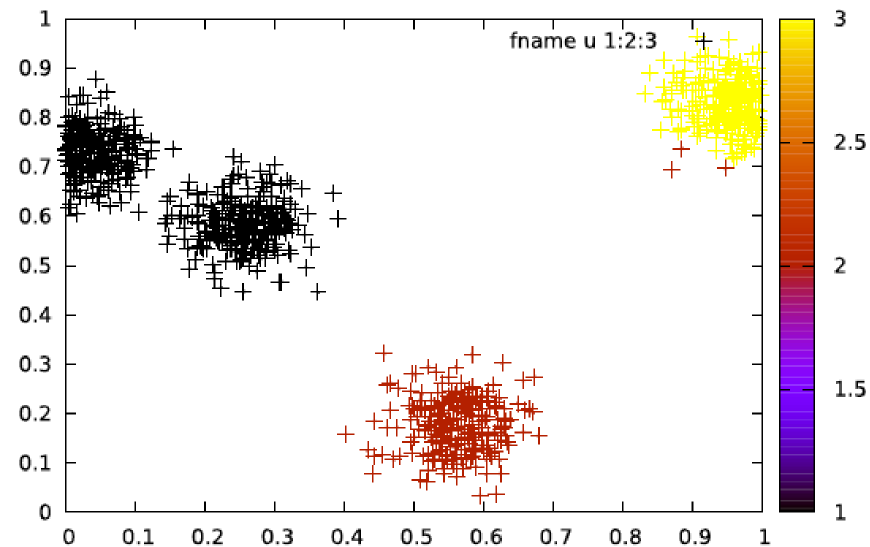
- In each iteration (given the set of centers):
  - 1. Assign the points to the new center to form clusters
  - 2. Noisily compute the size of each cluste Laplace(2T/ε)
  - 3. Compute noisy sums of points in each cluster Laplace(2T |dom|/ε)

RESULTS (T=10 ITERATIONS, RANDOM INITIALIZATION)

Original Kmeans algorithm



Laplace Kmeans algorithm



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

- Since we add noise to the sums with sensitivity proportional to |dom|, Laplace K-Means cannot distinguish small clusters that are close by.

PRIVACY AS CONSTRAINED OPTIMIZATION

- Three axes: **privacy**, **error** (**utility**), **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?
- E.g., Given a task and privacy budget ε, how to design a set of queries (functions) and allocate the budget such that the error is minimized?

OUTLINE

Differential Privacy for Centralized Data

1. Threat Model and Architecture
2. Differential Privacy Definition
3. Basic Techniques
4. Composition Theorems
5. Other DP Variants

RELAXED INDISTINGUISHABILITY DP

Definition 2 ((ϵ, δ)-differential privacy). *A randomization algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy if for all neighboring inputs D and D' and any set of possible outputs S, we have $Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D') \in S] + \delta$ and $Pr[\mathcal{A}(D') \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D) \in S] + \delta$.*

Pr[A(D) є S]=0 or Pr[A(D’) є S]=0

How to bound the probabilities?

RELAXED PROBABILISTIC DP

Given any neighboring inputs D and D’, and any output O

- If Pr[A(D) = O]>0 or Pr[A(D’) = O]>0, then Pr[A(D) = O]/Pr[A(D’) = O]<= exp (ϵ)
- If Pr[A(D) = O]=0, then for all D’, the overall probability Pr[A(D’) = O]<= δ

AN EXAMPLE OF NON-INTERACTIVE DP

Input Search Log

User-ID u_j	query-url pair Φ_i	Count c_{ij}
081	pregancy test nyc, medicinenet.com	2
	book, amazon.com	3
	google, google.com	15
082	car price, kbb.com	2
	google, google.com	7
083	google, google.com	17
	diabetes medecine, walmart.com	1
	book, amazon.com	1
	car price, kbb.com	5

Compute the ***optimal output counts*** of all the query-url pairs $x=(x_1,.., x_n)$

(Sampling is differentially private if x satisfies the conditions in Theorem 2)

Multinomial Sampling

query-url pair Φ_i	Optimal Output Count x_i	Sampled User IDs (sampled times)
~~pregancy test nyc, medicinenet.com~~	~~0~~	
book, amazon.com	3	2→081 (2), 083 (1)
google, google.com	20	20→081 (8), 082 (3), 083 (9)
~~diabetes medecine, walmart.com~~	~~0~~	
car price, kbb.com	4	3→082 (1), 083 (3)

(Maximize the output utility with a defined utility measure for the ***output counts*** of all the query-url pairs)

(a) Sanitization with Multinomial Sampling

Output Search Log

User ID u_j	query-url pair Φ_i	Count x_{ij}
081	~~pregancy test nyc, medicinenet.com~~	~~0~~
	book, amazon.com	2
	google, google.com	8
082	car price, kbb.com	1
	google, google.com	3
083	google, google.com	9
	~~diabetes medecine, walmart.com~~	~~0~~
	book, amazon.com	1
	car price, kbb.com	3

(b) A Sample Output

- Using Probabilistic DP Definition

ACKNOWLEDGMENTS

Note: Some of the slides in this lecture are adapted based on materials created by

- Dr. Michael Hay at Colgate University
- Dr. Ashwin Machanavajjhala at Duke University
- Dr. Li Xiong at Emory University