Introduction to Differential Privacy

Sayak Ray Chowdhury
Assistant Professor
Department of Computer Science and Engineering
IIT Kanpur

Theory CS Winter School 2024 @ IISc Bangalore

Content

- Overview of Data Privacy
  - Why do we need privacy?
  - How can privacy be breached?
  - Adversarial Attacks

- Differential Privacy as an answer
  - Definition of Differential Privacy
  - Properties of Differential Privacy
  - Basic mechanisms, and their privacy and utility guarantees

- More DP mechanisms
  - A variant of DP definition

Acknowledgement

Materials are based on

- The Algorithmic Foundations of Differential Privacy, by Cynthia Dwork and Aaron Roth
- Privacy in Statistics and Machine Learning, taught by Adam Smith and Jonathan Ullman
- Privacy Preserving Machine Learning, taught by Aurélien Bellet
- Algorithms for Private Data Analysis, taught by Gautam Kamath
- Applied Privacy for Data Science, taught by James Honaker and Salil Vadhan

Suggestions are welcome

Data Privacy

The ability of an individual to seclude themselves or to withhold information about themselves

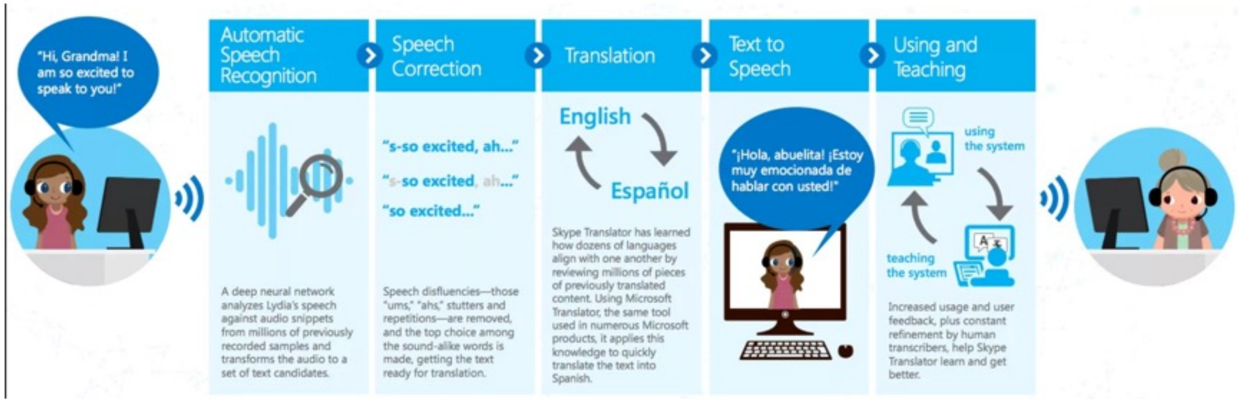
Data are everywhere

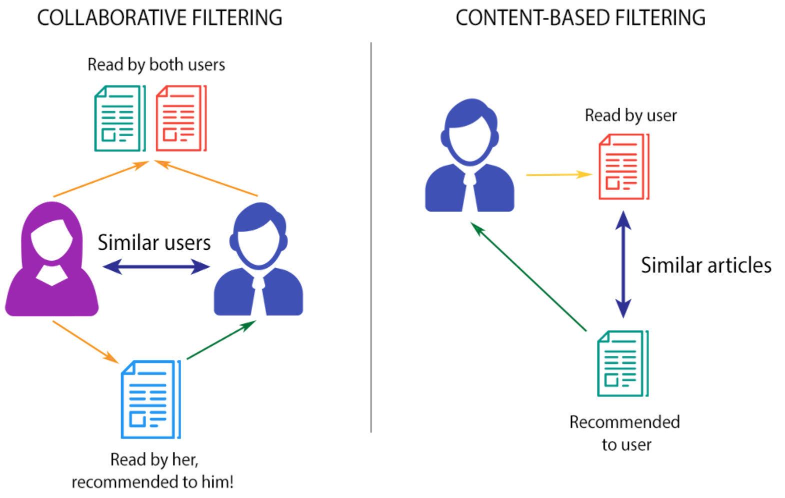
Massive collection of personal data by companies and public organizations, driven by the progress of data science and AI

Data is increasingly sensitive and detailed: browsing history, purchase history, social network posts, speech, geolocation, health...

Machine Learning on our Data

Real-time Speech Translation

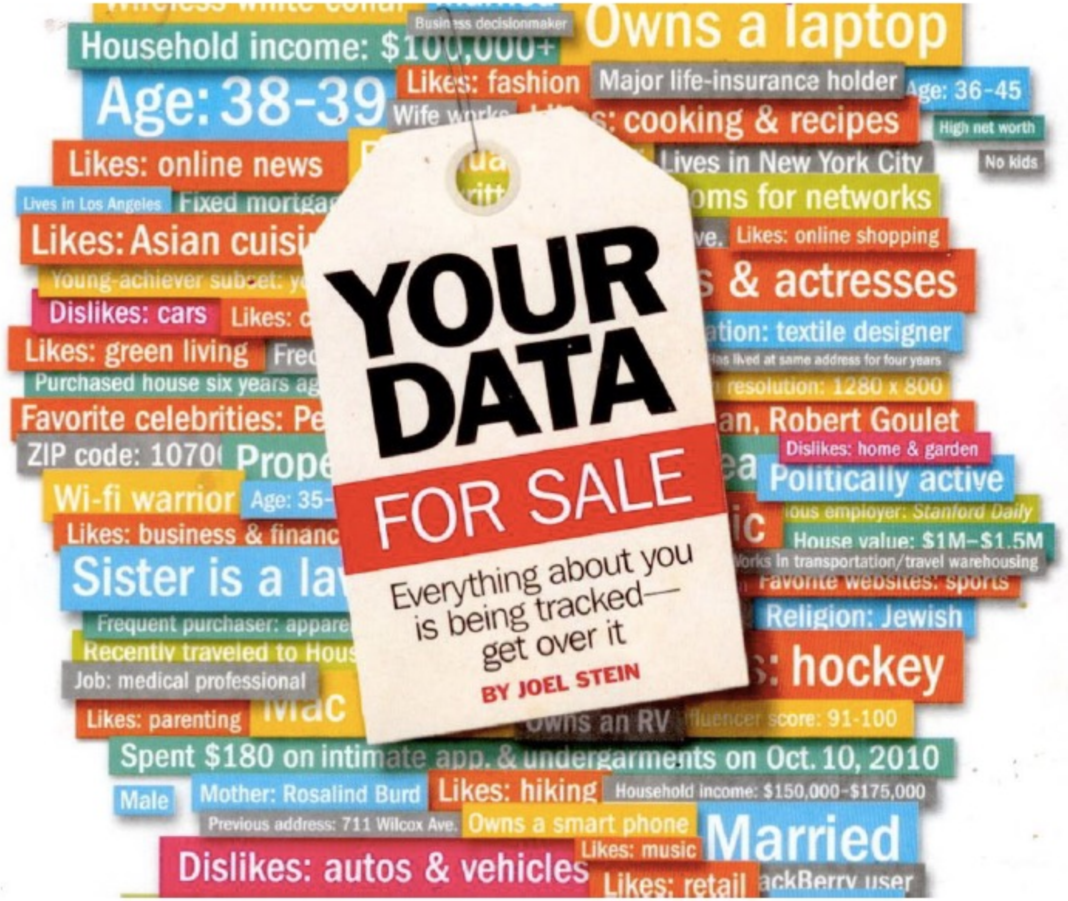




Autonomous Driving

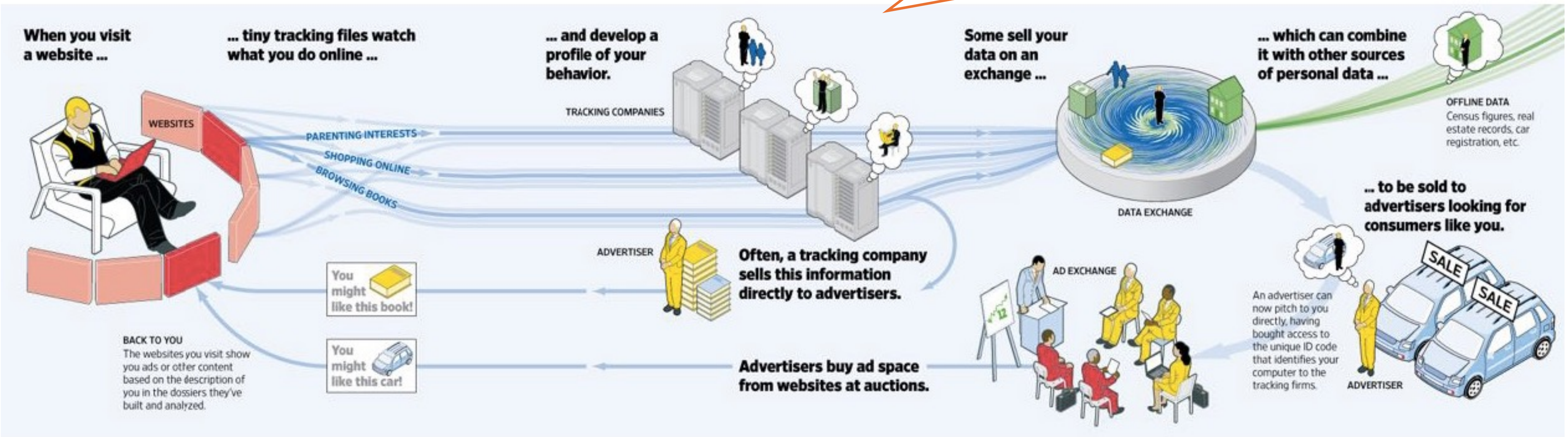
Conversational Systems

Data Privacy: The Problem



Data Privacy: The Problem

How our data are collected?

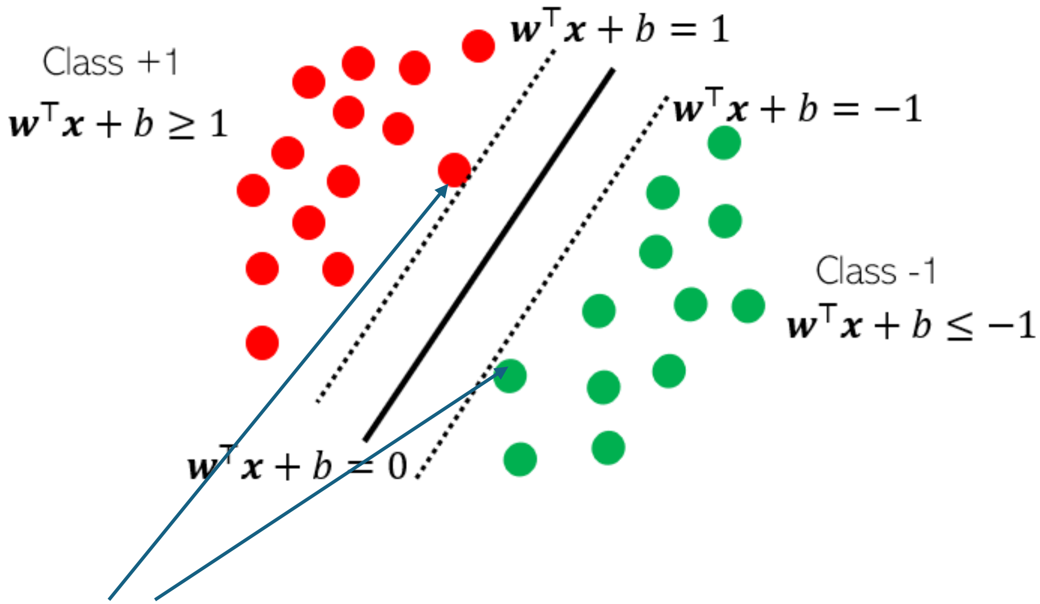


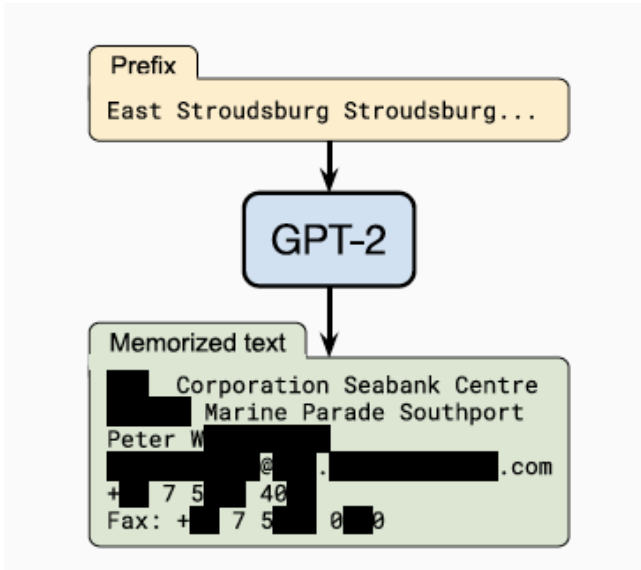
Data Privacy: The Problem

Websites that track our data

Site	Exposure Index	Trackers
dictionary.com	Very High	234
merriam-webster.com	High	131
comcast.net	High	151
careerbuilder.com	High	118
photobucket.com	High	127
msn.com	High	207
answers.com	Medium	120
yp.com	Medium	89
msnbc.com	Medium	117
yahoo.com	Medium	106
aol.com	Medium	133
wiki.answers.com	Medium	72
cnn.com	Medium	72
about.com	Medium	83
cnet.com	Medium	81
verizonwireless.com	Medium	90
imdb.com	Medium	55
live.com	Medium	115
att.com	Medium	58
walmart.com	Medium	66
bbc.co.uk	Medium	45
ebay.com	Medium	42
ehow.com	Medium	55

Data Privacy: The Problem





Support vectors reveal training data

LLMs reveal Sensitive information
(by adversarial prompting)

Modern ML models almost memorize inputs (e.g. Autocomplete feature in Gmail)

Data Privacy: The Problem

Given a database with sensitive information such as

Credentials

- credit card number, passwords, ...

Identification Information

- name, age, gender, bank details, biometrics, ...

Sensitive Information

- medical records, political opinions, religious beliefs, ...

How can we

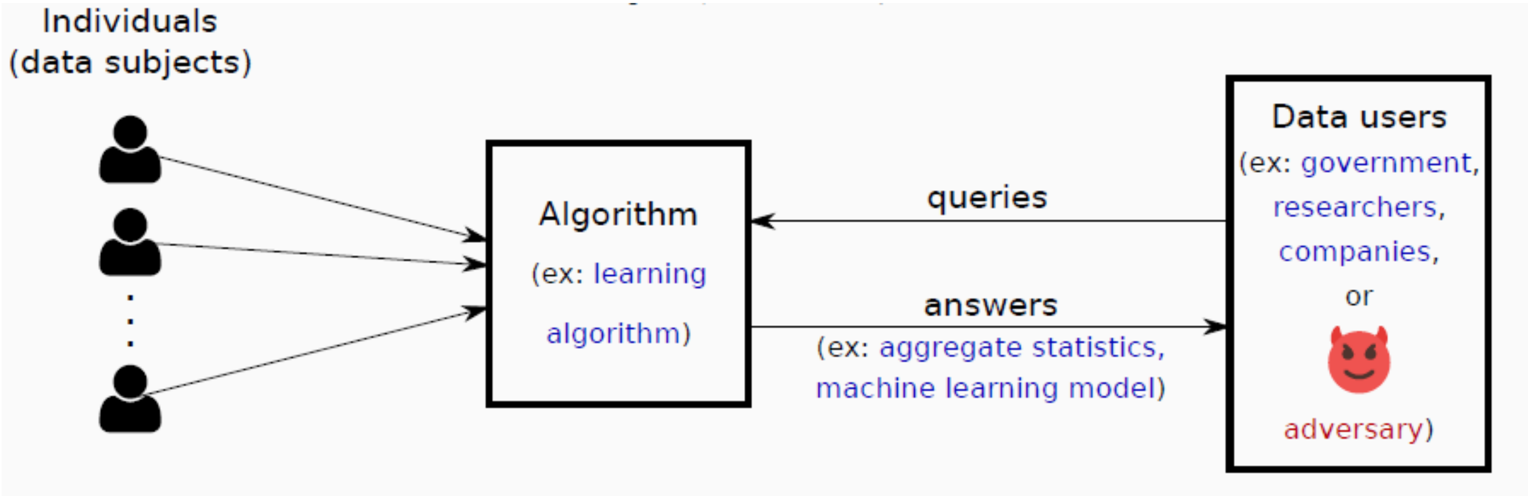
Policy formation, Clinical trials, Sentiment analysis, Searching for fraud, Academic research, ...

- ensure desirable uses of the data

Hiding individual information

- while protecting the privacy of the data subjects?

Privacy in Statistical Databases

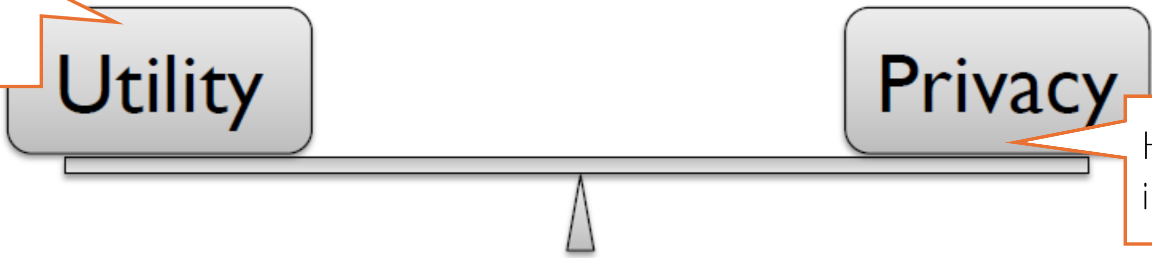


Statistical analysis benefits society

Large collection of personal information

Two Conflicting Objectives

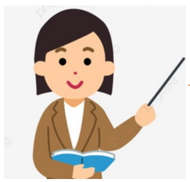
Releasing Aggregate Statistics



Hiding Individual information

Goal: How to achieve utility while maintaining privacy?

But, before that: How do we define **privacy**?



This lecture series; foundation and analysis

1st Attempt: Data Anonymization

Remove obvious identifiers (name, social security number) that uniquely identify an individual before publishing the data

Convince ourselves that data cannot be fully anonymized AND remain useful

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Name	Postal Code	Age	Sex	Has Disease?
	02445	36	F	1
	02446	18	M	0
	02118	66	M	1
	⋮	⋮	⋮	⋮
	02120	40	F	1

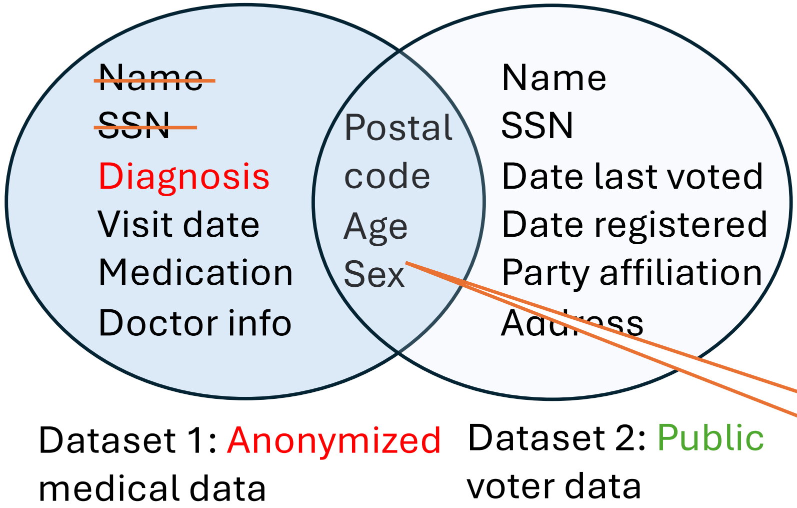
Zora has the disease

Now, we can't know that Zora has the disease

or, can we?

Is Data anonymization Safe?

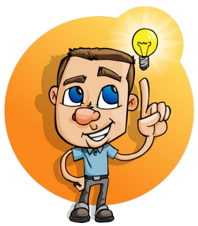
Linkage Attack



Reidentification via Linkage: uniquely linking a record in the anonymized dataset to a record in a public dataset

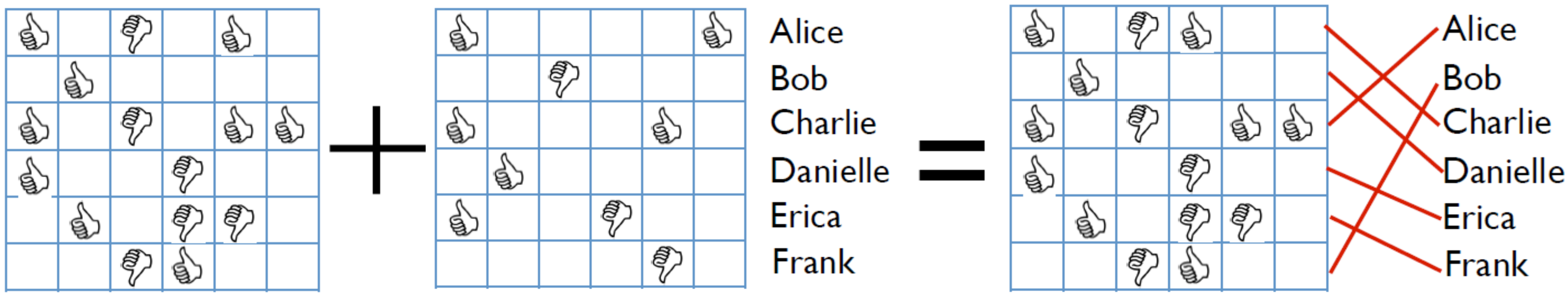
An estimated 87% of the US population is uniquely identified by the combination of their age, sex, and postal code

Quasi Identifiers



The Massachusetts Governor's privacy breach [Sweeney 2002]

Linkage in Practice: The Netflix Challenge



Anonymized NetFlix data

Public, incomplete **IMDB** data

Identified NetFlix Data

Challenge: Improve the Netflix Recommender system
Prize: US$1,000,000

- On average 4 movies uniquely identify a user [Narayanan Shmatikov 2008]
- Reveal information on users' movie-watching history, which they chose not to reveal publicly

2nd Attempt: K-Anonymization

Identifier | Quasi-Identifier | Sensitive attribute

Name	Postal Code	Age	Sex	Has Disease?
	024			1
	024			0
	021			1
	X			X
	021			1

Sweeny 2002:

Suppress/Generalize attributes to make every record in the dataset indistinguishable from at least $k-1$ other records with respect to the Quasi Identifiers

Now, we can't know that Zora has the disease, or, can we?

No! Can still infer that Zoya has the disease (everyone in the group has it)

Pitfalls of K-Anonymization: Composition

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

2 hospital release K anonymous tables for patients' medical history

A 28 year old person visited both hospitals

The person has AIDS

Ganta, Kashivishwanathan, Smith 2008

3rd Attempt: Release Aggregate Statistics

Is granularity the problem?
What if we only release aggregate statistics about many individuals?

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Data with the health insurance provider of a company

The company can ask for information like:

- How many females have the disease?
- How many females living in [postal code] have the disease?
- How many females living in [postal code] and aged [year] have the disease?

Now, can we know that Zora has the disease?

Are releasing aggregate statistics safe?

- Differencing Attack
- Reconstruction Attack
- Membership Inference Attack

Differencing Attack

Company asks: How many females living in 02120 and aged 40 have the disease?

Known to the company | Sensitive

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Data with the health insurance provider of a company

Say the answer is 1. Then it is very likely that the company learned about Zora's disease

Counter-argument: If the answer is 1, are we aggregating anything? What if the answer is 5?

The company now asks:

- How many females living in 02120 and aged $\geq \mathbf{40}$ have the disease? → Answer: 5
- How many females living in 02120 and aged $\geq \mathbf{41}$ have the disease? → Answer: 4

Zora's privacy is breached if she is the only 40 years old female employee living in 02120

Reconstruction from Statistical Table

Are specific questions the problem?
What if we ask for some "benign" information?

"Attack" on statistical disclosure methods used by US Census [Garfinkel et al 2019]

Group	Age Count	Age Median	Age Mean
Total Population	7	30	38
Female	4	30	33.5
Male	3	30	44
Black or African American	4	51	48.5
White	3	24	24
Single Adults	(D)	(D)	(D)
Married Adults	4	51	[illegible]
Black or African American Female	3	36	36.7

Count=1

1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	[illegible]	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

Already reveals a lot of information

Census releases tabulation of Statistics
What can we learn from this table?

Prior knowledge: $1 \leq M1, M2, M3 \leq 125$ → 341376 possible choices for M1, M2, M3

From table: $M2 = 30$, $M1+M2+M3 = 132$, $M1+M3 = 102$ → 30 choices only

Reconstruction Attack

Identifiers (z) — Secrets (s)

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

e.g. $z_n = \{Zora, 02120, 40, F\}, s_n = 1$

Recap: queries are of the form

How many individuals older than 40 have disease?

We want to release Count Statistics of the form

Query condition $\in \{0,1\}$

$$\sum_{j=1}^{n} \phi(z_j)\, s_j$$

What is $\phi(z_i)$ for the above query?

$$= [\phi(z_1), \ldots, \phi(z_n)] \cdot [s_1, \ldots s_n]$$

$F \in \{0,1\}^n$ $\quad$ $s \in \{0,1\}^n$

A General Reconstruction Attack:

Input: k query vectors $F_1, \ldots, F_k \in \{0,1\}^n$ and k answers $a_1, \ldots, a_k \in$ R

Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_i|$

Reconstruction Accuracy

Reconstruction Attack:

Input: k query vectors $F_1, \ldots, F_k \in \{0,1\}^n$ and k answers $a_1, \ldots, a_k \in \mathrm{R}$

Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_i|$

Hypothesis: each query is answered within error αn, that is, $\max_{i \in [k]} |F_i \cdot s - a_i| \leq \alpha n$

Then the reconstruction error is at most $4\alpha n$ if the attacker makes $k = 2^n$ queries

number of entries where the vectors s & $\tilde{s}$ differ

Powerful attack: Recovers 96% of secret bits even from answers with 1% error (think $\alpha = \frac{1}{100}$)

Reconstruction using all possible queries

But is this attack realistic?

No: it requires 2^n queries (exponential in the size of the dataset)

What if the number of queries are $\ll 2^n$?

Reconstruction Accuracy

Reconstruction Attack:

Input: k query vectors $F_1, \dots, F_k \in \{0,1\}^n$ and k answers $a_1, \dots, a_k \in \mathrm{R}$

Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_i|$

Hypothesis: each query is answered within error αn, that is, $\max_{i \in [k]} |F_i \cdot s - a_i| \leq \alpha n$

Then the reconstruction error is at most $\mathrm{O}(\alpha^2 n^2)$ with probability $1 - 2^{-n}$ if the attacker makes $k = O(n)$ queries chosen uniformly at random from the set 2^n possible queries

Powerful attack: Recovers nearly all secret bits (reconstruction error $\ll n$) from answers with error $\ll \sqrt{n}$ (think $\alpha \ll \frac{1}{\sqrt{n}}$)

But is this attack Computationally feasible?

No: it requires to search over 2^n possible vectors

How can we make the attack run in time polynomial in n?

Reconstruction Attack (Compute Friendly)

Reconstruction Attack:

Input: k query vectors $F_1, \dots, F_k \in \{0,1\}^n$ and k answers $a_1, \dots, a_k \in R$

~~Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_i|$~~

Output: a vector of secrets $\hat{s} \in \mathbb{R}^n$ that minimizes $\max_{i \in [k]} |F_i \cdot \hat{s} - a_i|$ & round-off to $\tilde{s} \in \{0,1\}^n$

Hypothesis: each query is answered within error $\ll \sqrt{n}$, that is, $\max_{i \in [k]} |F_i \cdot s - a_i| \ll \sqrt{n}$

Then nearly all secret bits are recovered with a very high probability if the attacker makes $k = O(n)$ queries chosen uniformly at random from the set 2^n possible queries

Runs in time polynomial in dataset size

Linear programming in n variables & $k = O(n)$ constraints

Rounding-off is Linear in n

But why does reconstruction attack work when error $\ll \sqrt{n}$?

What happens if we allow error $\geq \sqrt{n}$?

Membership Inference attacks

Reconstruction in Practice: The Diffix Challenge

Diffix: System for computing statistic

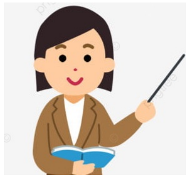
Make SQL queries on a database while
Preventing disclosures about individuals

Secrets

```
SELECT count(*) FROM loans
WHERE loanStatus = 'C'
AND clientId BETWEEN 2000 and 3000
```

Identifiers

Diffix add noise to the counts



Can the system provide exact counts?

No. Think about differencing attacks

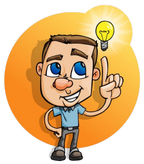
Diffix knew about this

Recall: Efficient reconstruction requires random queries like

```
SELECT count(*) FROM loans
WHERE loanStatus = 'C'
AND (clientId = 2007
OR clientId = 2018
...
OR clientId = 2991)
```

Queries of this form will be answered with $\geq \sqrt{k}$ error (# of terms in query $= k$)

If terms are randomly selected, then $k = O(n)$ and hence error $\geq \sqrt{n}$



No Reconstruction!

Cohen and Nissim 2018

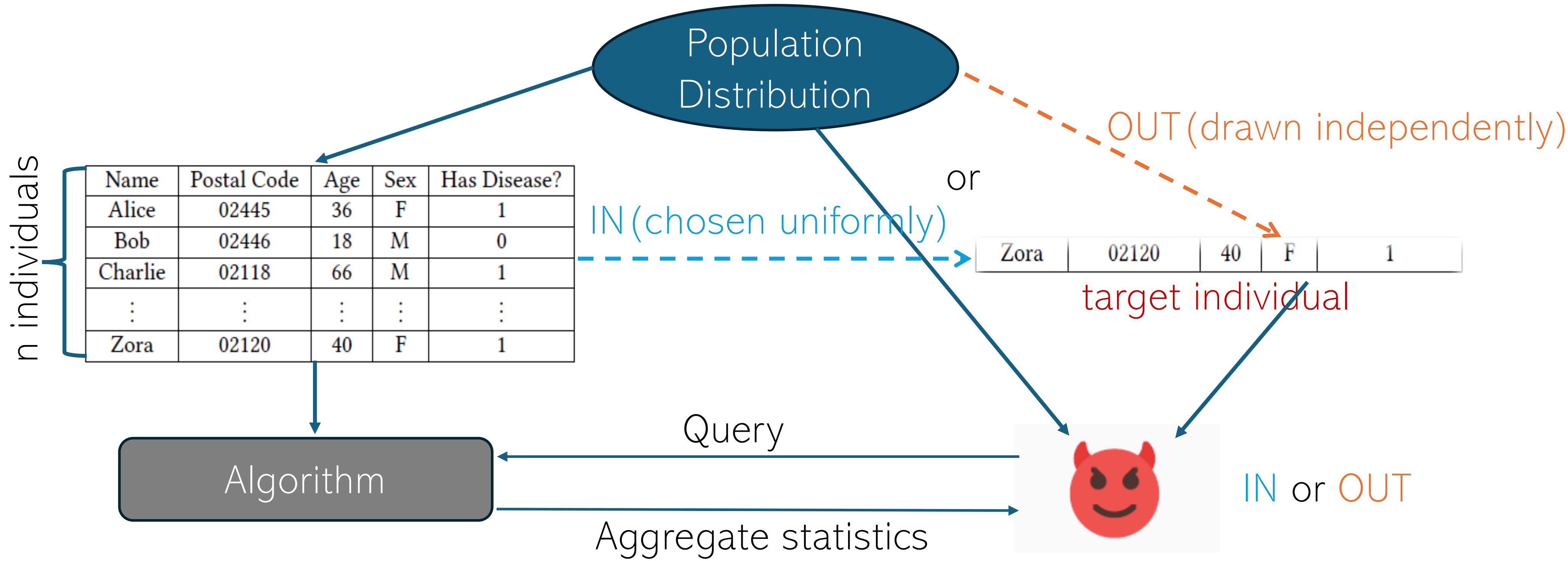
Random enough query

```
SELECT COUNT(clientId) FROM loans
WHERE FLOOR(100 * ((clientId * 2)^0.7))
    = FLOOR(100 * ((clientId * 2)^0.7) + 0.5)
AND clientId BETWEEN 2000 and 3000
AND loanStatus = 'C'
```

Constant # of terms in query
Answered with error $O(1)$

Full victim to Reconstruction!

Membership Inference Attack

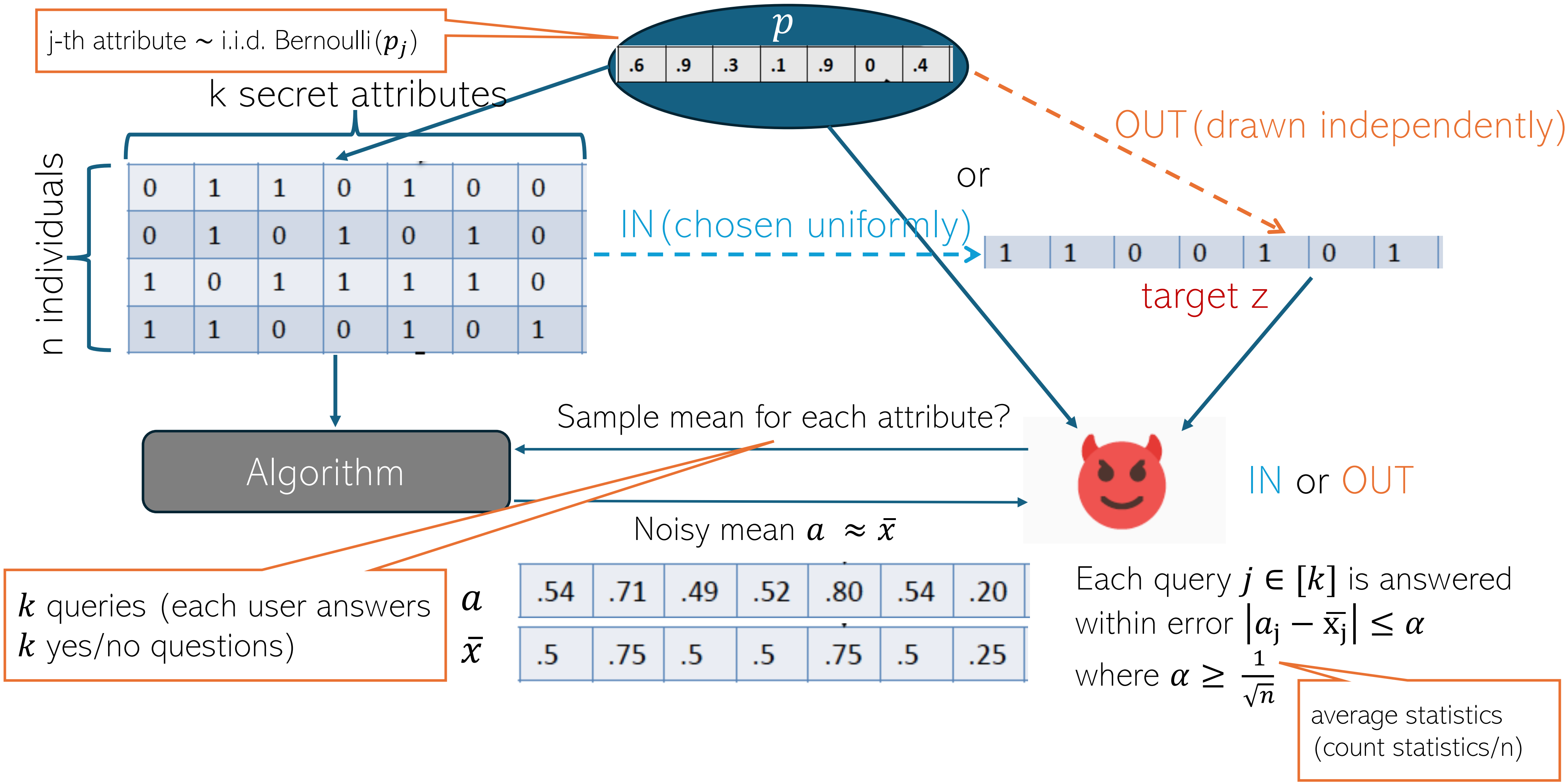


Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

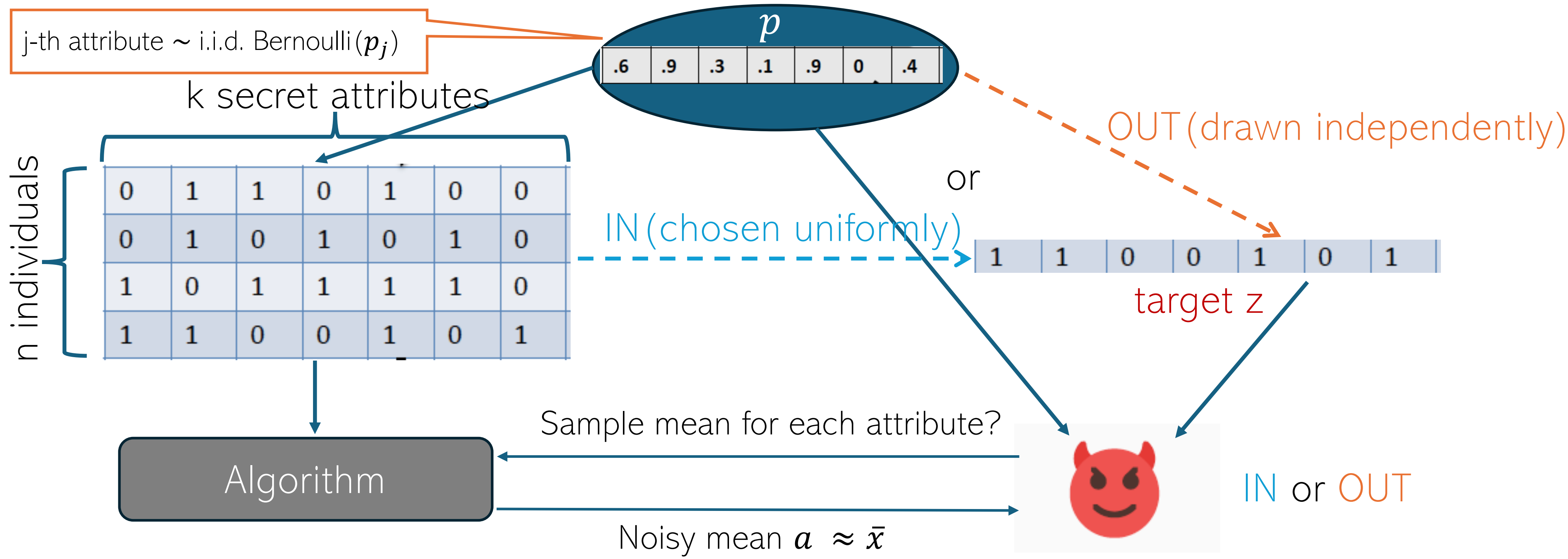
Attacker gets
- Access to Algorithms output
- Zora's data
- Auxiliary information about population

Attacker decides if Zora's data is in the dataset or not

Membership Inference Attack



Membership Inference Attack



Each query $j \in [k]$ is answered within error $|a_j - \bar{x}_j| \leq \alpha$ where $\alpha \geq \frac{1}{\sqrt{n}}$

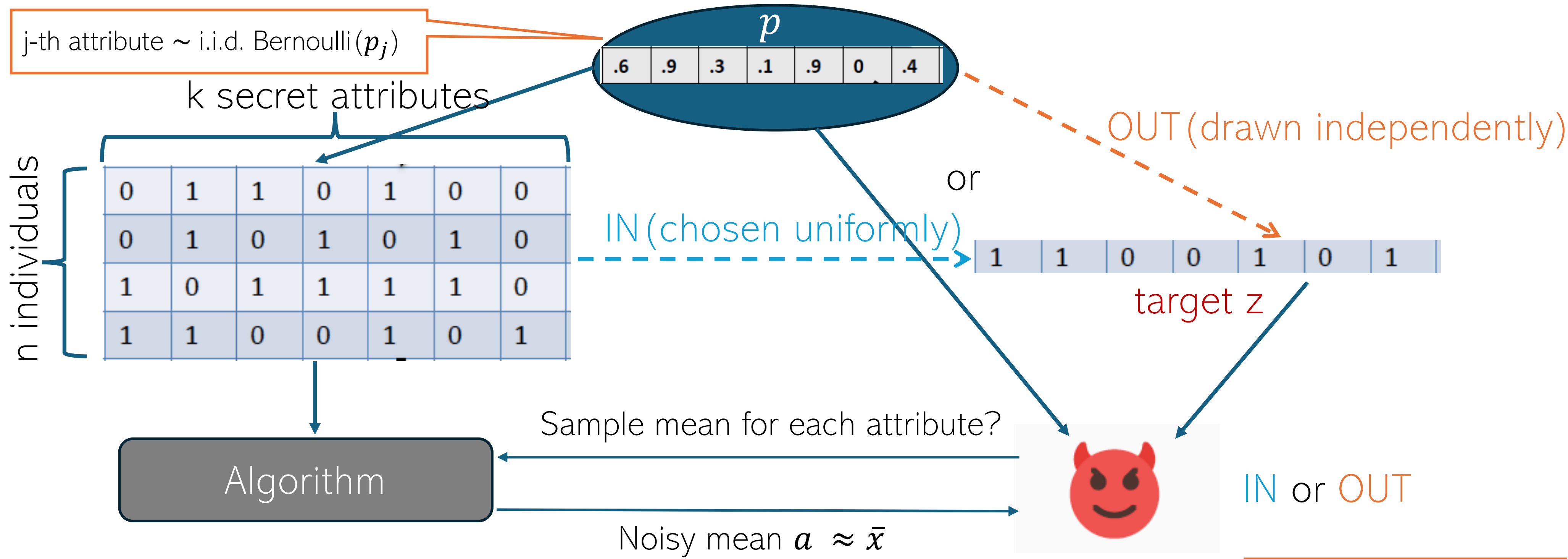
Dwork et al 2015:

There exists an attack such that when $k \geq n$ and $\alpha < \frac{\sqrt{k}}{n\sqrt{\log(1/\delta)}}$:

- If target = IN, then $P[\text{IN}] \geq \frac{1}{\alpha^2 n}$ (True Positive)
- If target = OUT, the $P[\text{IN}] \leq \delta$ (False Positive)

succeeds with 90% probability when $k = 1.1n$

Membership Inference Attack



Dwork et al 2015:

There exists an attack such that when $k \geq n$ and $\alpha < \frac{\sqrt{k}}{n\sqrt{\log(1/\delta)}}$:

- If target = IN, then $P[\text{IN}] \geq \frac{1}{\alpha^2 n}$ (True Positive)
- If target = OUT, the $P[\text{IN}] \leq \delta$ (False Positive)

Choose each $p_j \sim$ U[0,1]

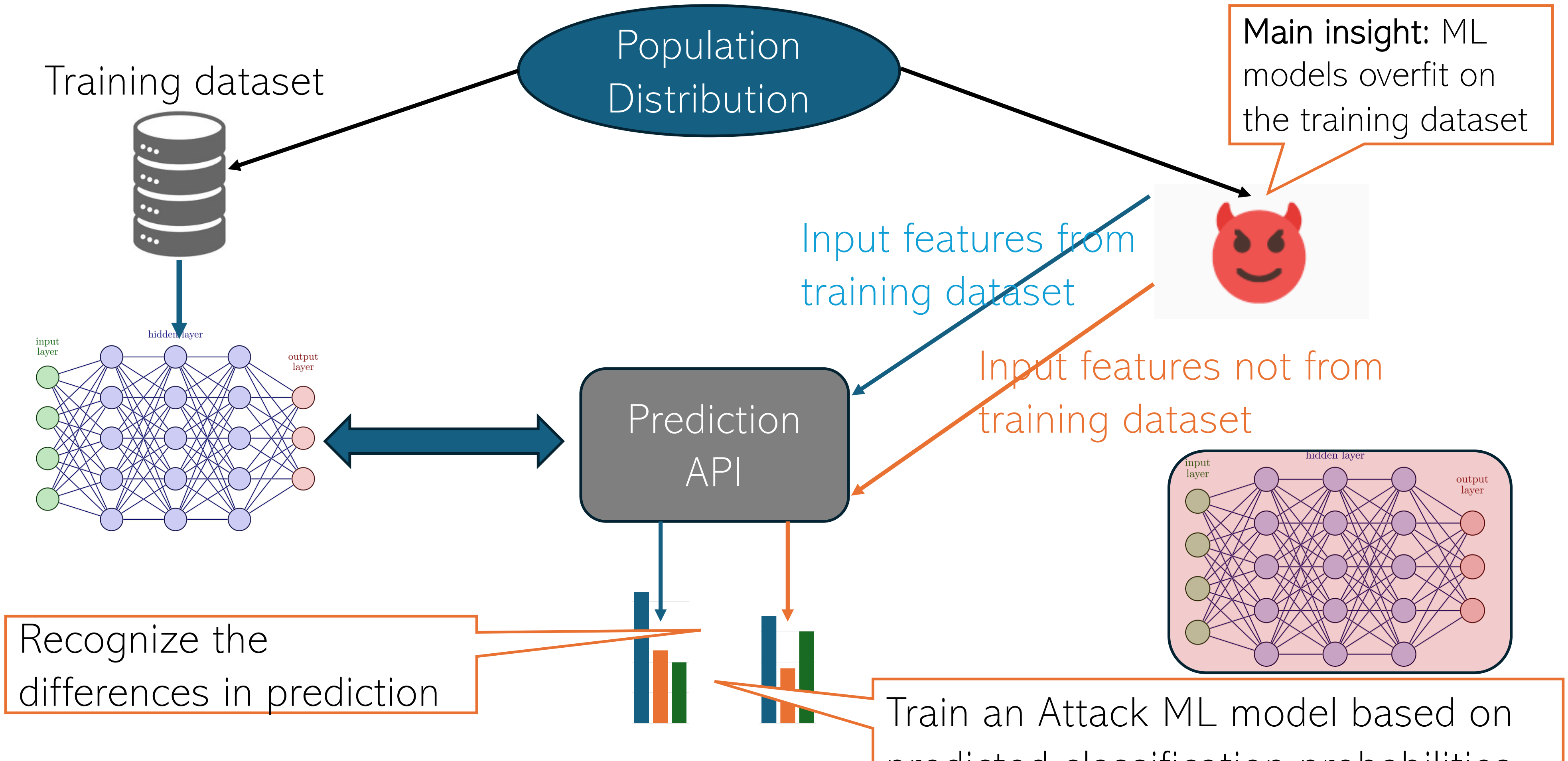
The Attack:

If $(a - p) \cdot (z - p) \geq \tau$ return IN

else return OUT

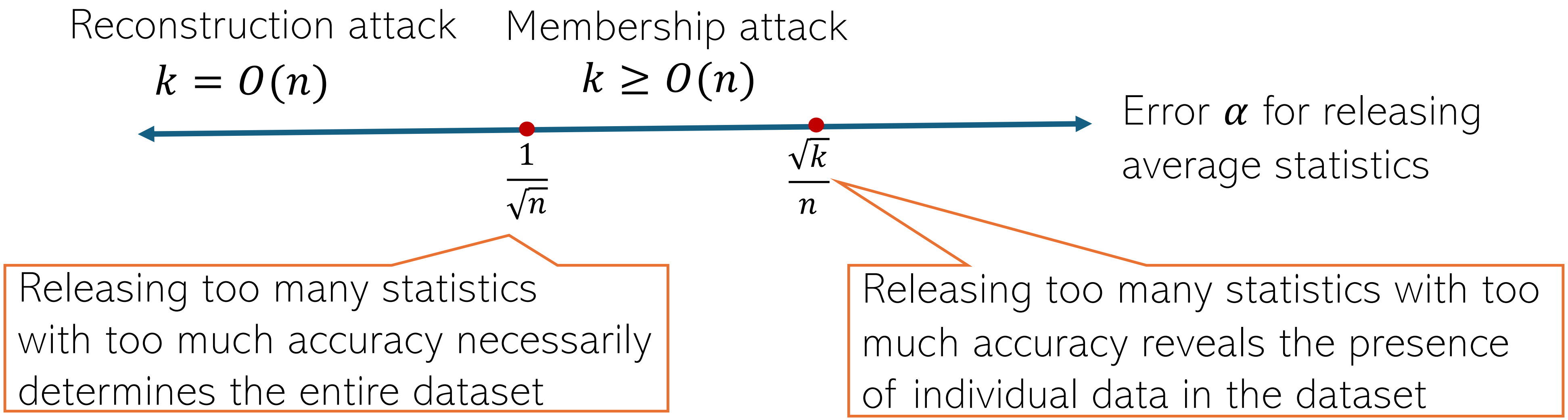
Set $\tau \approx \sqrt{k \log(1/\delta)}$ to make false positive probability δ

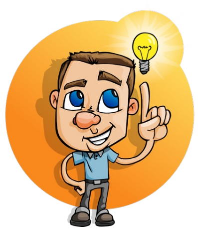
Membership Inference in Practice: ML vs. ML



Shokri et al. 2017

The Attack Landscape

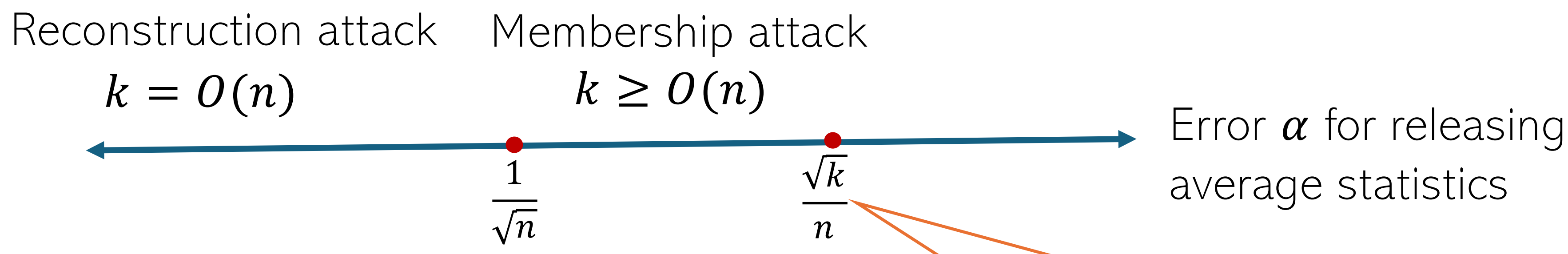




- Every statistic released yields a (hard or soft) constraint on the dataset
- We need a quantitative theory that tells us "how much is too much" and "how many is too many"

End of Lecture 1

Recap: The Attack Landscape



Releasing too many statistics with too much accuracy necessarily determines the entire dataset

Releasing too many statistics with too much accuracy reveals the presence of individual data in the dataset

We need a quantitative theory for "how much is too much" and "how many is too many"

Recap: Reconstruction Attack

Identifiers (z) — Secrets (s)

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Recap: we want to answer **k** queries of the form

$$F \cdot s = \sum_{j=1}^{n} \phi(z_j)\, s_j \quad \text{(count statistics)}$$

Reconstruction Attack:
Input: queries $F_1, \dots, F_k$ and answers $a_1, \dots, a_k$
Output: secrets $\tilde{s}$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_i|$

Hypothesis: each query is answered within error $\ll \sqrt{n}$, that is, $\max_{i \in [k]} |F_i \cdot s - a_i| \ll \sqrt{n}$

Then nearly all secret bits are recovered with a very high probability if the attacker makes $k = O(n)$ queries chosen uniformly at random from the set 2^n possible queries

Reconstruction attack works when error $\ll \sqrt{n}$

Preventing Reconstruction Attack

Identifiers (z) — Secrets (s)

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

m random subsamples → For each $j \in [m]$, pick an $i \in [n]$ uniformly at random & set $(z_j', s_j') = (z_i, s_i)$

Recap: we want to answer **k** queries of the form

$$F \cdot s = \sum_{j=1}^{n} \phi(z_j)\, s_j \quad \text{(count statistics)}$$

Reconstruction Attack:

Input: queries $F_1, \ldots, F_k$ and answers $a'_1, \ldots, a'_k$

Output: secrets $\tilde{s}$ that minimizes $\max_{i \in [k]} |F_i \cdot \tilde{s} - a'_i|$

Release answer:

$$a' = \frac{n}{m} \sum_{j=1}^{m} \phi(z_j') s_j'$$

Note: With high probability $\max_{i \in [k]} |F_i \cdot s - a_i'| \leq \sqrt{n \log k}$

error $\approx \sqrt{n \log n}$ for $k = O(n)$ queries

But, does subsampling give privacy guarantee?

Zora's data lie in the subsample with probability $\frac{m}{n}$

So, the same privacy concern remains

But, reconstruction attack requires error $\ll \sqrt{n}$

Hence, subsampling prevents reconstruction

We need a theory to give accurate answers with rigorous privacy guarantees

Requirements of Privacy

Protection against auxiliary knowledge: we need to be robust to whatever knowledge an attacker may have since we cannot predict what she knows or might know in the future

Protection against multiple analyses: we need to be able to track how much information is leaked when asking several questions about the same data

Achieving utility: we need to be able to do “meaningful statistical analysis” of datasets

Privacy Definition: Attempt 1

An analysis of a dataset is private if the attacker's belief about an individual stays the same after they see the result as it were before (no matter what they know before time)

Impossible to reveal nothing if the result is to depend on the data (else we don't get any utility)

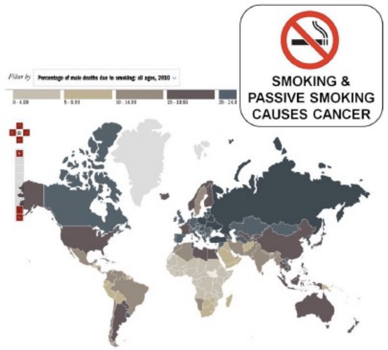
Before and after requirement unachievable after auxiliary knowledge

Not quite there!

Health insurance company knows Alice is a smoker

Alice

+



=

company raises Alice's insurance premium

Does this breach Alice's privacy?

No: The company would have raised the premium regardless of Alice's participation

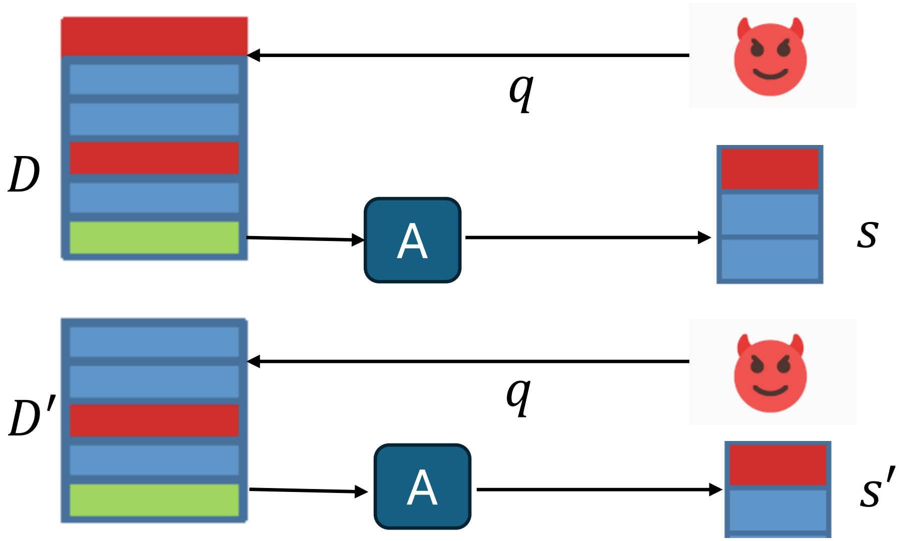
Such correlations are the kind of things we want to be able to learn

Privacy Definition: Attempt 2

An analysis of a dataset is private if the attacker would draw almost same conclusions about an individual whether or not her data were used in the analysis (no matter what they know before time)

can't infer membership of an individual in the dataset or can't reconstruct any attribute about her

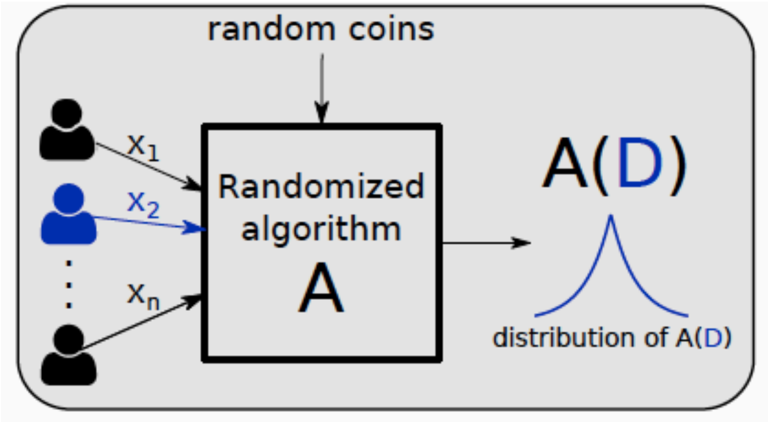
Randomization is necessary to be robust to auxiliary knowledge

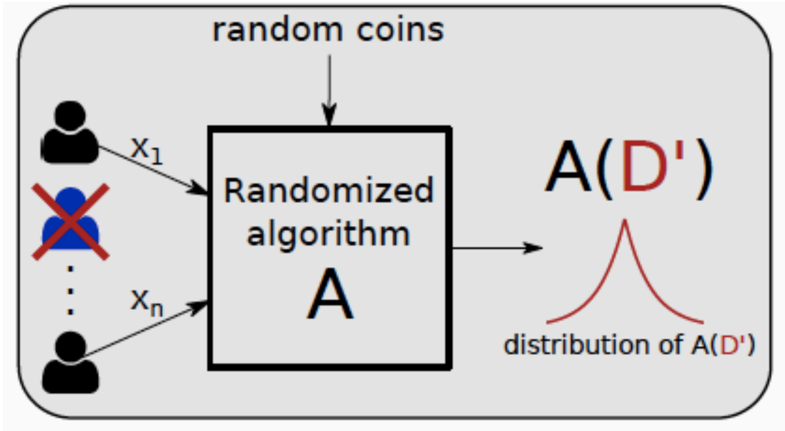


- Say, A is a non-trivial deterministic algorithm
- For datasets D, D' differing only in a single record, the same query q yields different outputs s, s'
- An adversary knowing that the dataset is one of D, D' can learn the differing record

Differential Privacy (DP)

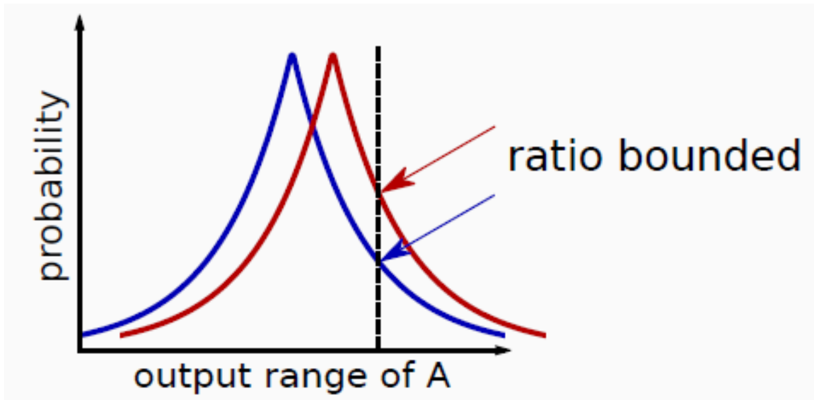
Dwork, McSherry, Nissim and Smith [2006]





A thought experiment:

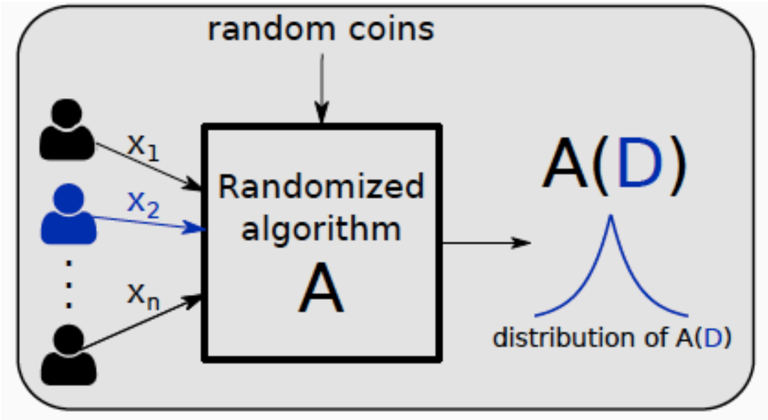
- Change, add or remove one person's data
- Will the probabilities of the outcomes change?

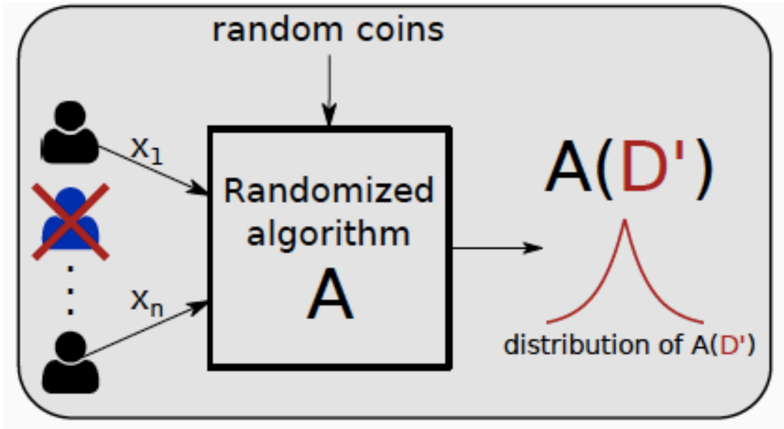


Requirement of DP: Both distributions should be close

Differential Privacy (DP)

Dwork, McSherry, Nissim and Smith [2006]





The randomized algorithm A is ϵ-differentially private
if for all neighboring datasets D, D' and for all outputs S:

A thought experiment:

- Change, add or remove one person's data
- Will the probabilities of the outcomes change?

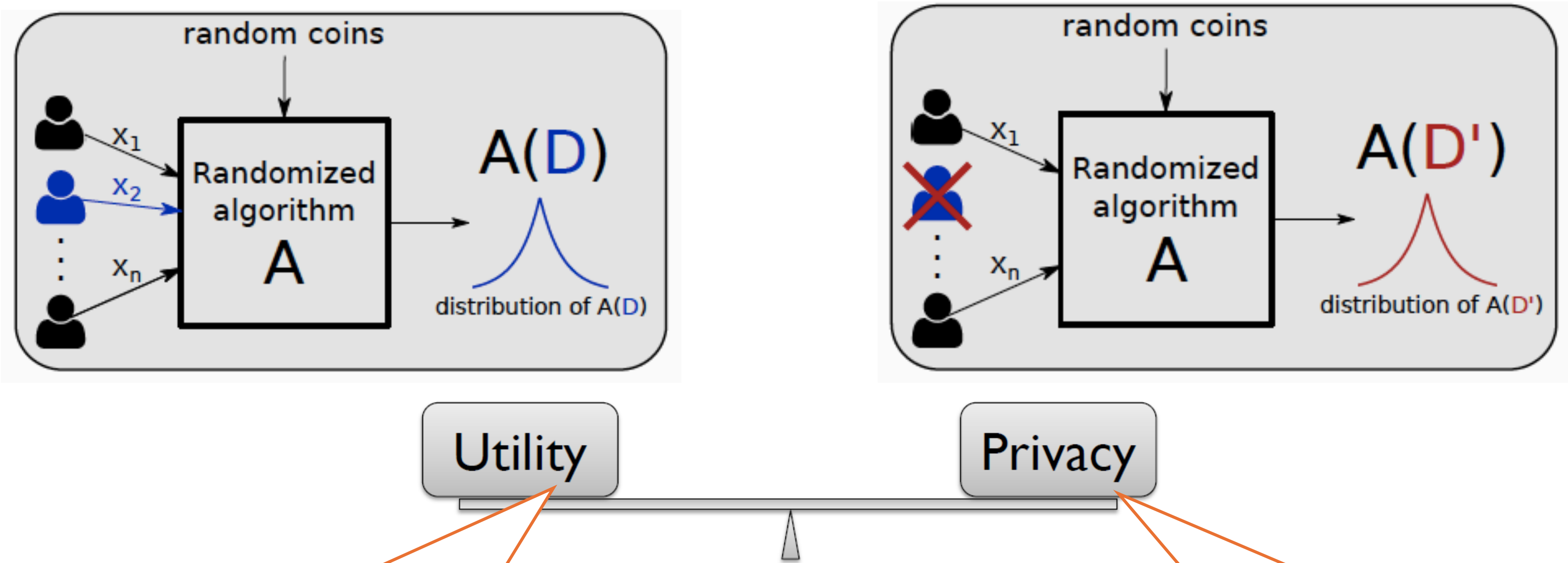
Neighboring datasets

(a) $P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$

(b) $P[A(D') \in S] \leq e^{\epsilon} \cdot P[A(D) \in S]$

Requirement of DP: Both distributions should be close $(\epsilon \approx 0)$

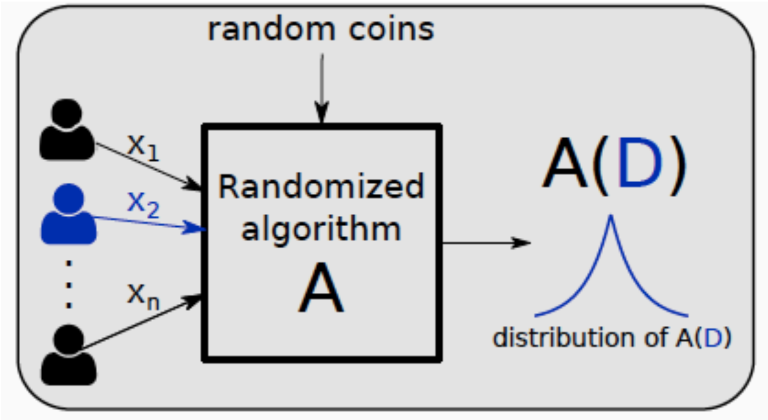
Two Conflicting Objectives

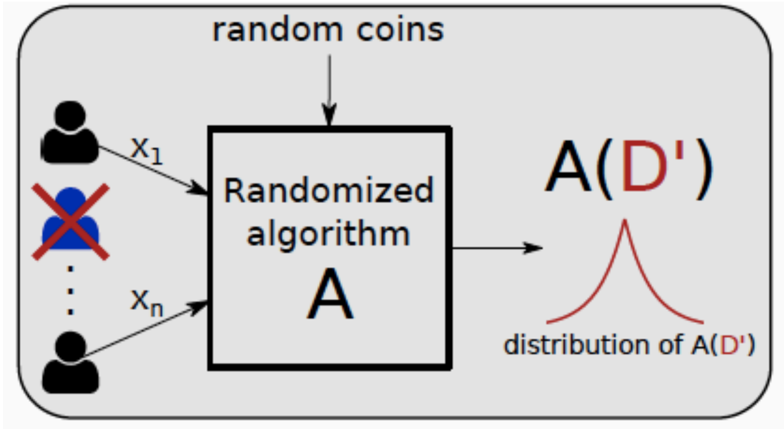


Enable statistical analysis of datasets e.g. inference about population, training ML models

Protect individual level data against all attack strategies and auxiliary information

Promises (and not) of DP





$$P[A(D) \in S] \overset{e^{\epsilon}}{\approx} P[A(D') \in S]$$

What DP promises …

- Whatever an attacker learns about me, it could have learned from everyone else's data
- Protection from the attacker's auxiliary knowledge
- Graceful composition for multiple queries (k repetitions)

What DP doesn't promise…

- Protection for information that is not localized to a few records
- Giving privacy where none previously exists
- Guarantee that individuals won't be "harmed"

DP Research and Deployments

Algorithms

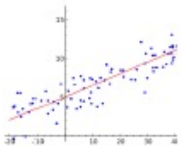
- Approximation algorithms
- Singular value decomposition
- Streaming Algorithms

.....

Crypto, security

- Multi-party computation
- Floating point arithmetic
- Computational primitives

.....



Statistics, learning

- Histogram
- Contingency tables
- Regression
- Estimation
- Clustering

.....

Game theory, economics

- Social network analysis
- Mechanism design
- Multi-agent systems

.....

iOS 10 and Safari (2016)

US census (2020)

RAPPOR for Chrome Statistics (2014)

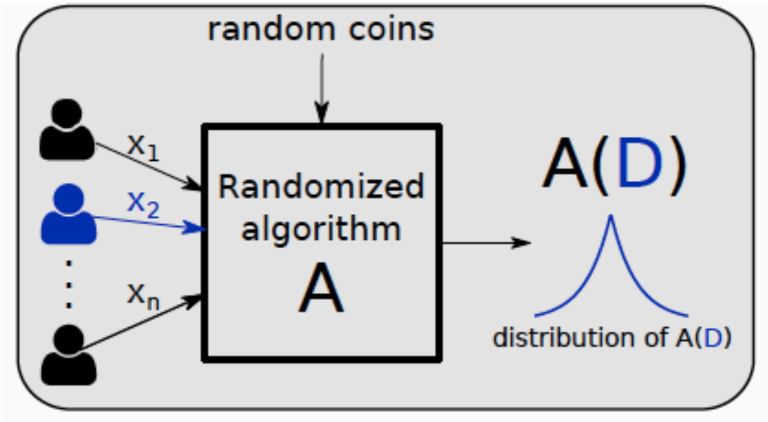
Growing interest from many communities in seeing whether DP can be brought into practice (databases, programming languages, medical informatics, law, social science, …)

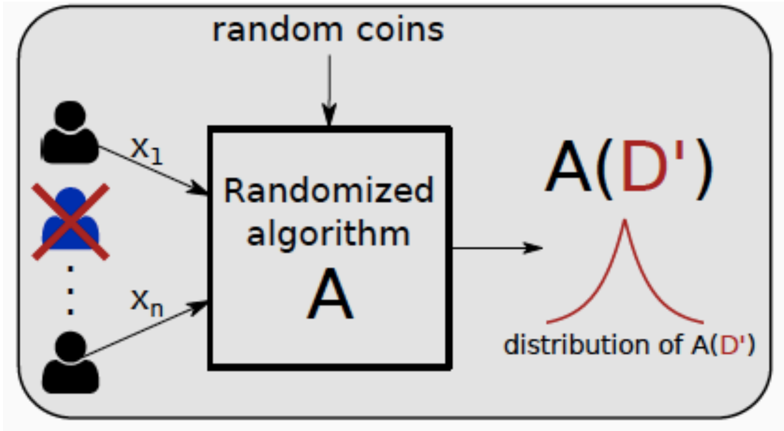
Comparison with other Privacy Models

Model	Utility	Privacy	Data holder
Differential Privacy	Statistical analysis of dataset	Individual information	Trusted server
Secure Function Evaluation	Any given query	Everything other than result of the query	Users
Homomorphic Encryption	Any given query	Everything	Untrusted server

Key principle: DP is a property of analysis and not of a particular output

Recap: Differential Privacy (DP)





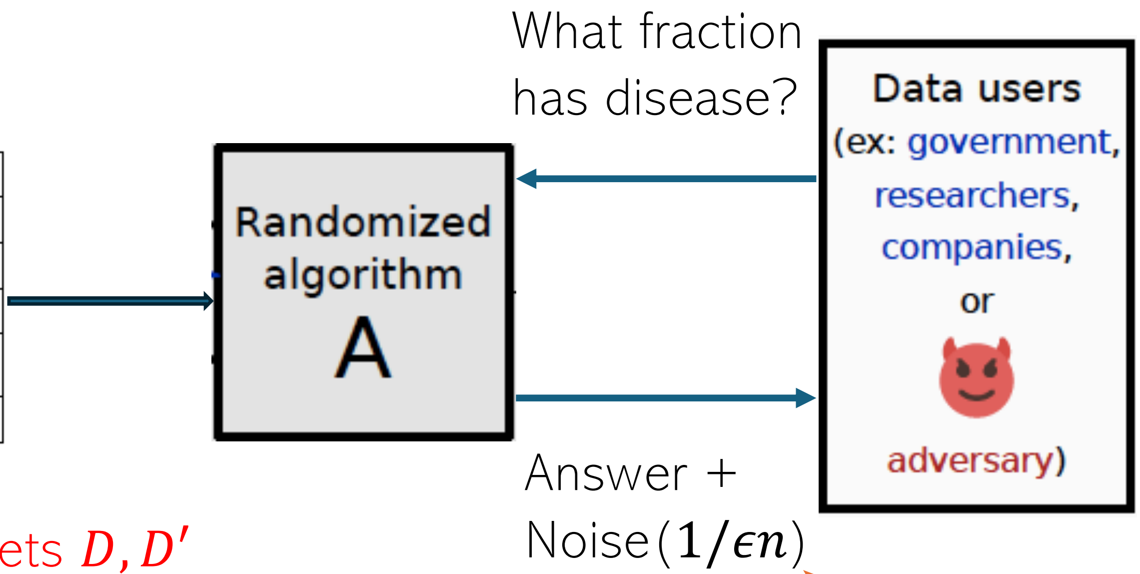
The randomized algorithm A is ϵ-differentially private
if for all neighboring datasets D, D' and for all outputs S:

$$P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$$

How to achieve DP?

n individuals; non-sensitive: Name, Postal Code, Age, Sex; sensitive: Has Disease?

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1



Requirement: for all neighboring datasets D, D' and for all outputs S:

$$P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$$

For meaningful privacy guarantee: $0 < \epsilon \leq 1$

Very little noise needed to hide one individual as $n \to \infty$

Laplace Mechanism

non-sensitive | sensitive

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

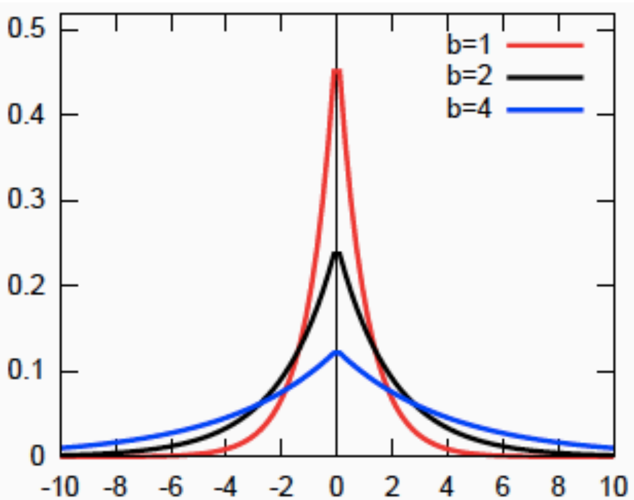
Randomized algorithm
A

What fraction has disease?

Data users
(ex: government, researchers, companies, or

adversary)

Answer + Laplace$(1/\epsilon n)$



PDF (with scale b):

$$p(y; b) = \frac{1}{2b} \exp\left(-\frac{|y|}{b}\right)$$

Density at $y \propto \exp(-\epsilon n|y|)$

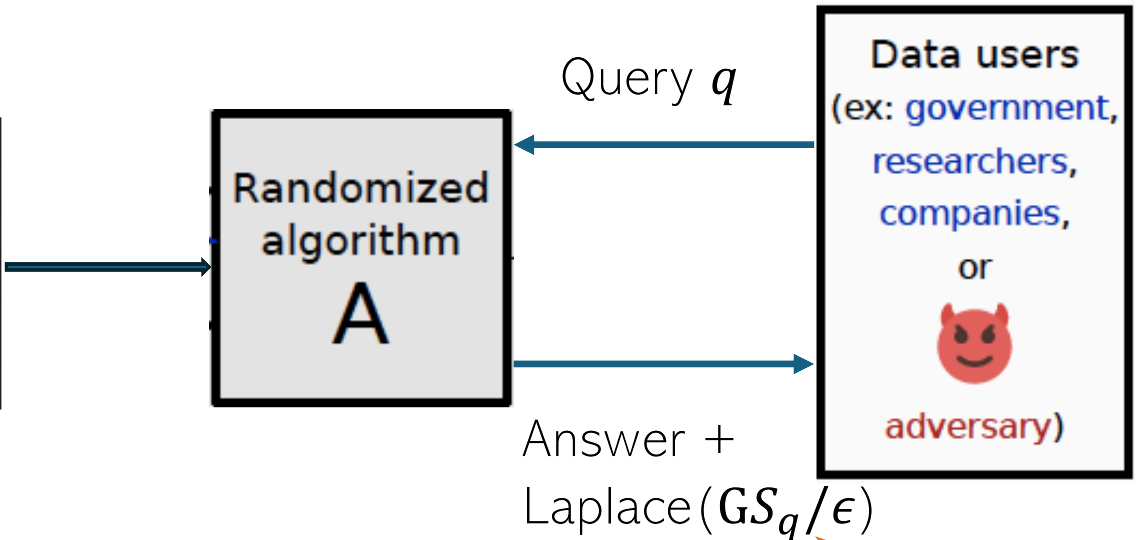
How much noise should we add for a given query q?

Laplace Mechanism

non-sensitive | sensitive

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1



Global sensitivity of a query q:

$$GS_q = \max_{D\sim D'} |q(D) - q(D')|$$

Neighboring datasets

How sensitive a query is to change in one record in the dataset?

Density at $y \propto \exp(-\epsilon|y|/GS_q)$

Theorem: The mechanism $A(D,q) = q(D) +$ Laplace(GS_q/ϵ) is ϵ-DP

Privacy Guarantee: Proof

In Board

Utility Guarantee

In Board

Properties of DP: Robust to Auxiliary Knowledge

A is ϵ-DP if for all neighboring datasets D, D' and for all outputs S:

$$P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$$

Robust to arbitrary auxiliary knowledge

Bounds the relative advantage that an attacker gets by observing output of an algorithm

Attacker may know the dataset except one record

Attacker may have all external sources of knowledge

Algorithm A can be public (a key requirement for modern security)

Properties of DP: Postprocessing

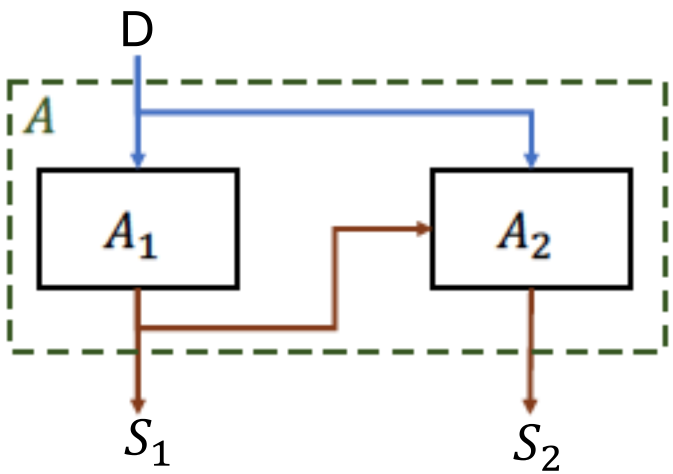
Theorem: Let an algorithm $A: D \rightarrow S$ be ϵ-DP and $f: S \rightarrow O$ be any (randomized) function. Then, the composed algorithm $f(A): D \rightarrow O$ is also ϵ-DP

Impossible to compute a function of the output of a private algorithm and make it less private

Allows data users to do whatever they want with output of a private algorithm

Proof: In Board

Properties of DP: Basic Composition



Theorem: Let $A: D \rightarrow S_1 \times S_2$ be an composed algorithm that outputs (s_1, s_2) where $s_1 = A_1(D)$ and $s_2 = A_2(s_1, D)$. Then A is $(\epsilon_1 + \epsilon_2)$-DP

Allows to control cumulative privacy for multiple queries on the same dataset

$A_1: D \rightarrow S_1$ is ϵ_1-DP

$A_2: S_1 \times D \rightarrow S_2$ is ϵ_2-DP → $A_2(s_1, \cdot)$ is ϵ_2-DP for all $s_1 \in S_1$

Extends to k such DP algorithms (one for each query): cumulative privacy scales linearly with number of queries

Can be improved using **Advanced Composition:** cumulative privacy scales sub-linearly with number of queries

Proof: Basic Composition

In Board

Privacy Accounting

Composition: If A is ϵ-DP for one query, then it is $k\epsilon$-DP for k queries

What if total allowed privacy loss is ϵ_0? Need to set $\epsilon = \epsilon_0/k$

Trade-off needed b/w accuracy and number of queries (for given privacy loss)

More queries → Smaller ϵ → Less accuracy for answering each query

Composition (+ post-processing) allow designing DP algorithms which

1. Can ask multiple low-sensitivity queries
2. Can tolerate noisy answer to the queries

Classic ML example: Stochastic Gradient Descend (SGD)

Setting ϵ: Group Privacy

Theorem: Let D_1, D_2 be two datasets of n records that differ in $1 \le k \le n$ positions. If an algorithm A is ϵ-DP, then for all outputs S, we have

$$P[A(D_1) \in S] \le e^{k\epsilon} \cdot P[A(D_2) \in S]$$

Different than composition

Need to set $\epsilon \ge \frac{1}{n}$ for reasonable utility

Hide participation of

1. An individual who contribute several records
2. Groups of people whose data are strongly correlated

Why?

DP algorithms can't give useful output for small datasets

If $\epsilon \ll \frac{1}{n}$ then regardless of number of differing positions k, the distributions of $A(D_1)$ and $A(D_2)$ are almost same

→ To ensure high privacy, the algorithm ignores its input

Proof: Group Privacy

In Board

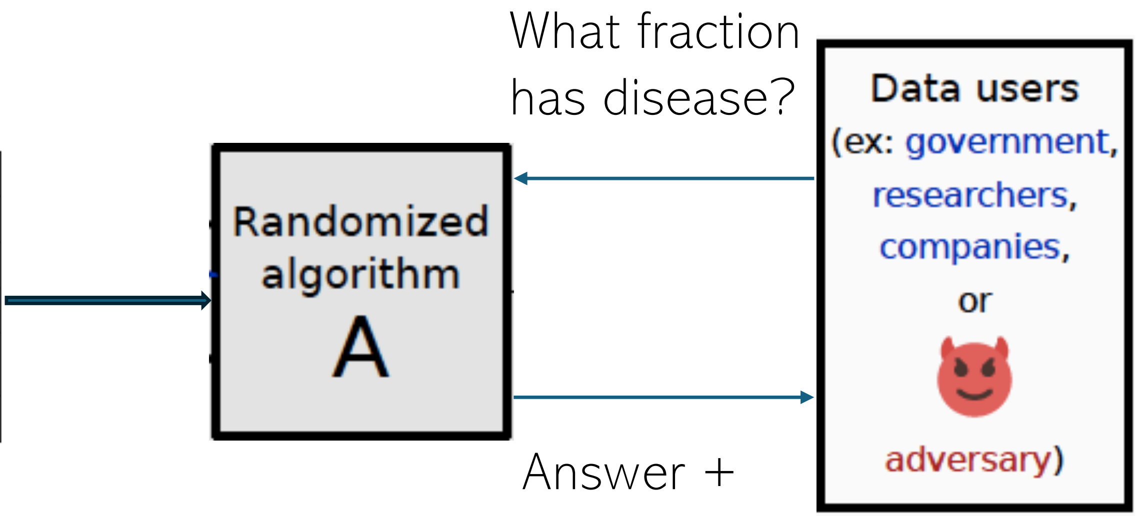
End of Lecture 2

Till Now: Numeric Queries

non-sensitive sensitive

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1



Answer + Noise$(1/\epsilon n)$

Useful only if similar answers have similar utility

Output perturbation

Not satisfied always

Buyer	Offer
Alice	3€
Bob	4€

Find profits for the following prices: {3, 3.01 , 4, 4.01 , …}

Privacy for Non-numeric Queries

Queries of the form:

1. Which CS theory lecture is popular among students?

2. What is the most popular AI model?

3. Which price would make the most profit from buyers?

Answers of the form:

Y = {Matching, Zero-knowledge protocol, Differential privacy, …}

Y = {GPT4, Llama, Phi2, Gemini, …}

Y = {3, 3.01, 4, 4.01, …}

Global Sensitivity of a utility function u:

$$GS_u = \max_{y \in Y} \max_{D \sim D'} |u(D, y) - u(D', y)|$$

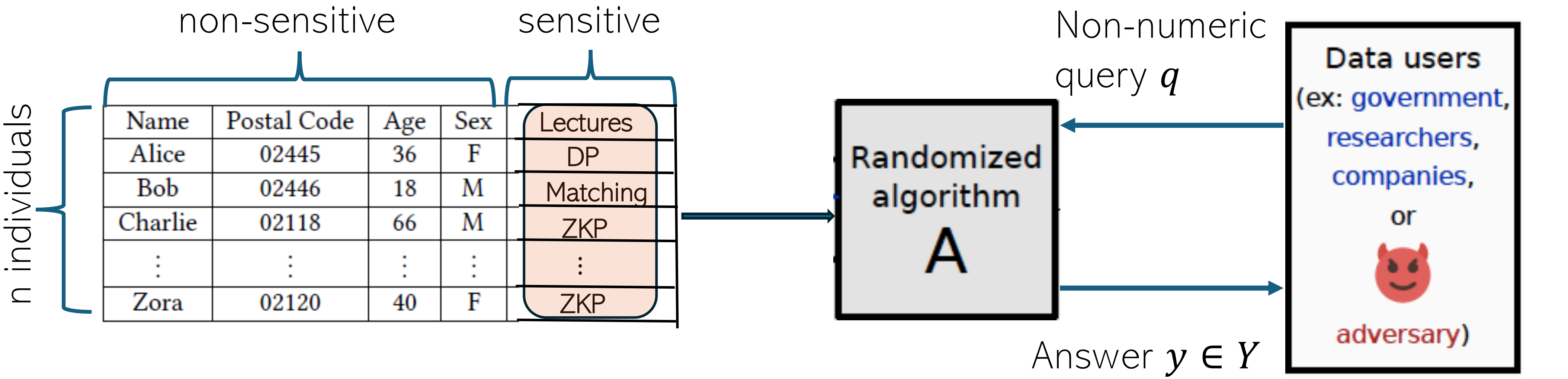
Neighboring datasets

Query $q: D \to Y$

Utility function $u: D \times Y \to \mathbb{R}$

How good is to return y when query is q?

Exponential Mechanism



Name	Postal Code	Age	Sex	Lectures
Alice	02445	36	F	DP
Bob	02446	18	M	Matching
Charlie	02118	66	M	ZKP
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	ZKP

with probability $\propto \exp(\epsilon \cdot \frac{u(D,y)}{2GS_u})$

High utility answers exponentially more likely

Global Sensitivity of a utility function u:

$$GS_u = \max_{y \in Y} \max_{D \sim D'} |u(D,y) - u(D',y)|$$

Theorem: The mechanism that answers $y \in Y$ with probability $P[A(D) = y] \propto \exp(\epsilon \cdot \frac{u(D,y)}{2GS_u})$ is ϵ-DP

Privacy Guarantee: Proof

For all $\mathbf{y} \in \mathrm{Y}$, we need to bound the ratio

$$\frac{\mathrm{P}[A(D)=y]}{\mathrm{P}[A(D')=y]}$$

We have $\mathrm{P}[A(D) = y] \propto \exp(\epsilon \cdot \frac{u(D,y)}{2\Delta})$

What is the proportionality constant here?

It is $\mathrm{C(D)} = \frac{1}{\sum_{y' \in Y} \exp(\epsilon \cdot \frac{u(D,y')}{2\Delta})}$

Upper bound on Global Sensitivity:

$$GS_u = \max_{y \in Y} \max_{D \sim D'} |u(D,y) - u(D',y)| \leq \Delta$$

See that $\frac{\mathrm{P}[A(D)=y]}{\mathrm{P}[A(D')=y]} = \frac{C(D)}{C(D')} \cdot \frac{\exp(\epsilon \cdot \frac{u(D,y)}{2\Delta})}{\exp(\epsilon \cdot \frac{u(D',y)}{2\Delta})}$

The ratio of proportionality constants is also upper bounded by $\exp(\epsilon/2)$

Re-write this as

$$\exp(\epsilon \cdot \frac{u(D,y) - u(D',y)}{2\Delta})$$

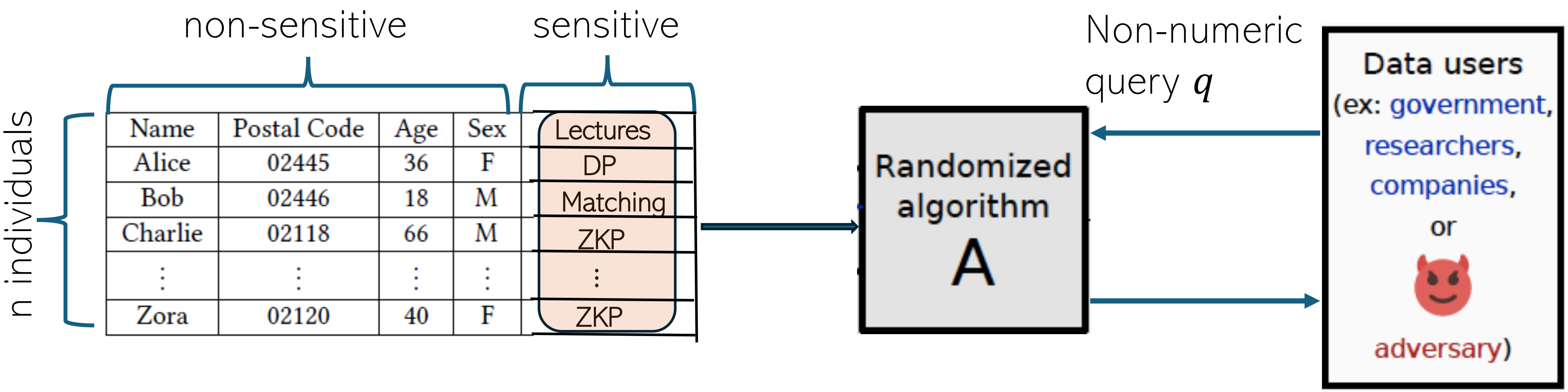
This is upper- bounded by

$$\exp(\epsilon \cdot \frac{\Delta}{2\Delta}) = \exp(\epsilon/2)$$

Then, for all $\mathbf{y} \in \mathrm{Y}$, we bound the ratio as

$$\frac{\mathrm{P}[A(D)=y]}{\mathrm{P}[A(D')=y]} \leq \exp(\epsilon/2) \cdot \exp(\epsilon/2) = \exp(\epsilon)$$

Report Noisy Max Mechanism



Name	Postal Code	Age	Sex	Lectures
Alice	02445	36	F	DP
Bob	02446	18	M	Matching
Charlie	02118	66	M	ZKP
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	ZKP

Answer $\mathbf{argmax}_{y \in Y}\{\, u(D,y) + \mathrm{Z_y}\}$

where each $Z_y \sim \mathrm{Exp}(\frac{2GS_u}{\epsilon})$ is Independent and identically distributed

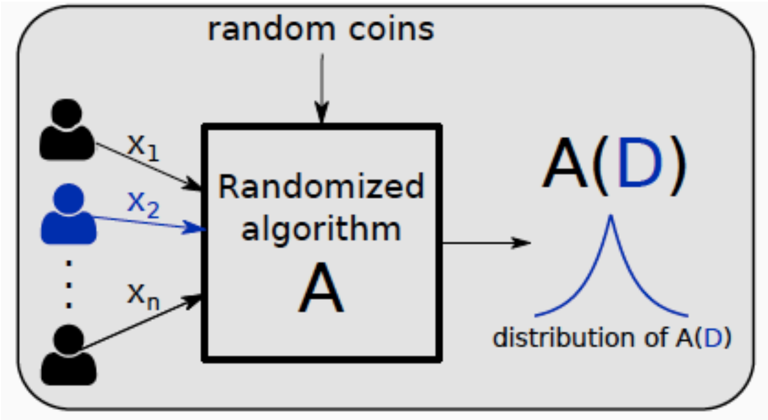
Exponential distribution has PDF: $p(y;\lambda) = \frac{1}{\lambda}\exp\left(-\frac{y}{\lambda}\right), y > 0$

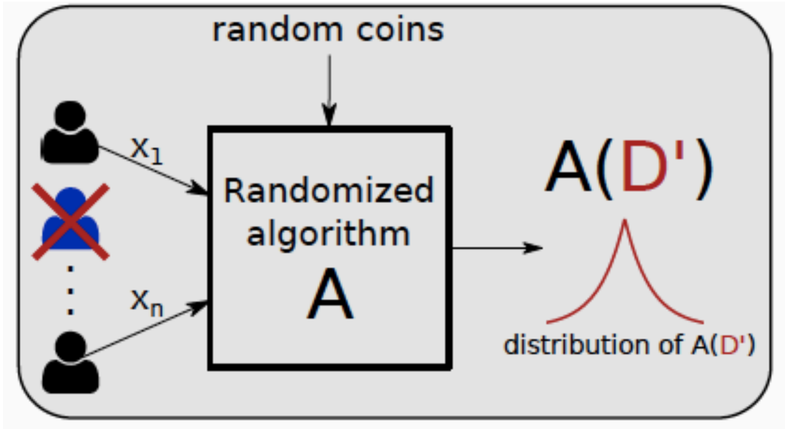
Global Sensitivity of a utility function u:

$$GS_u = \max_{y \in Y} \max_{D \sim D'} |u(D,y) - u(D',y)|$$

Theorem: The mechanism Report Noisy Max is ϵ-DP

Recap: Differential Privacy

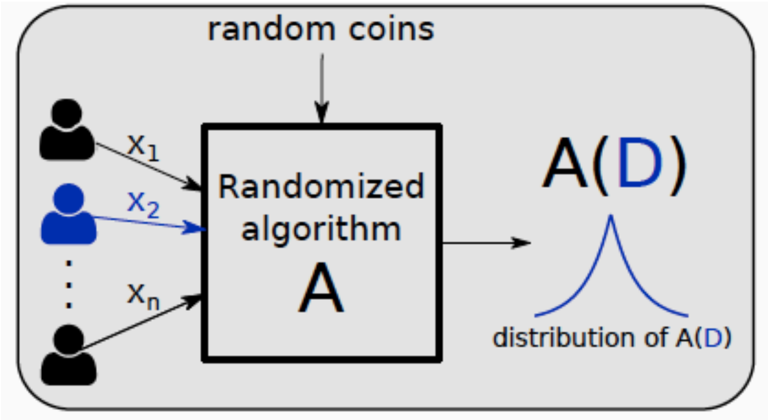


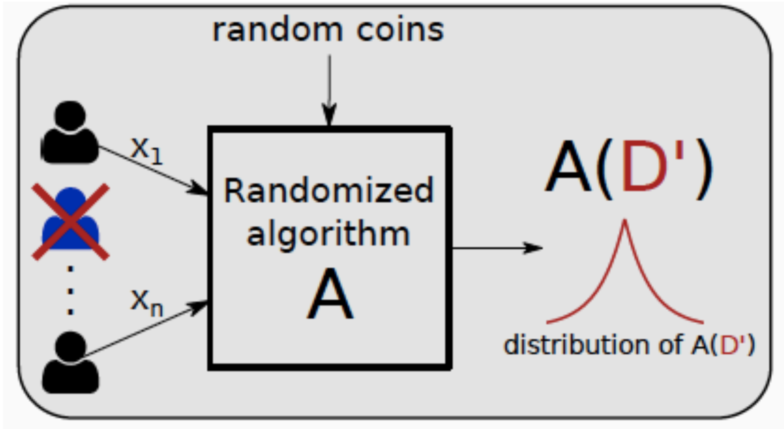


The randomized algorithm A is ϵ-differentially private
if for all neighboring datasets D, D' and for all outputs S:

$$P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S]$$

Variant: Approximate Differential Privacy





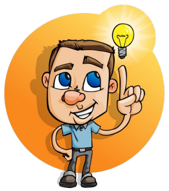
A is ϵ-DP with probability at least $1-\delta$

The randomized algorithm A is (ϵ, δ)-differentially private if for all neighboring datasets D, D' and for all outputs S:

$$P[A(D) \in S] \leq e^{\epsilon} \cdot P[A(D') \in S] + \delta$$

Makes sense only when $\delta \ll \frac{1}{n}$

Why?



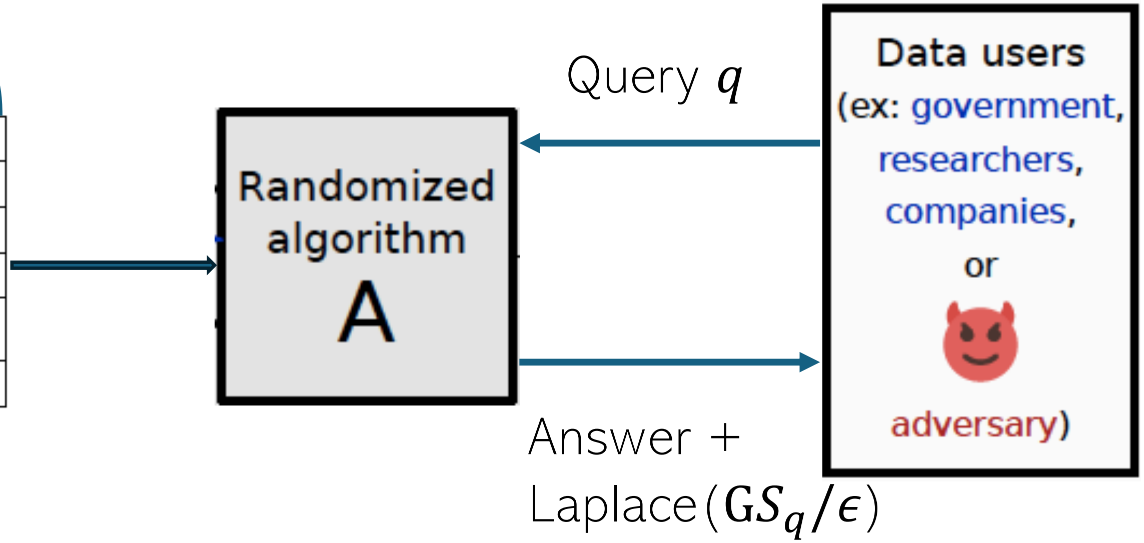
Pick a random person from the dataset and Publish her data → $\left(0, \frac{1}{n}\right)$ − DP

Recap: Laplace Mechanism for Pure DP

non-sensitive | sensitive

n individuals

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1



Global sensitivity of a query q:

$$GS_q = \max_{D \sim D'} |q(D) - q(D')|$$

Gaussian Mechanism for Approximate DP

non-sensitive | sensitive

n individuals

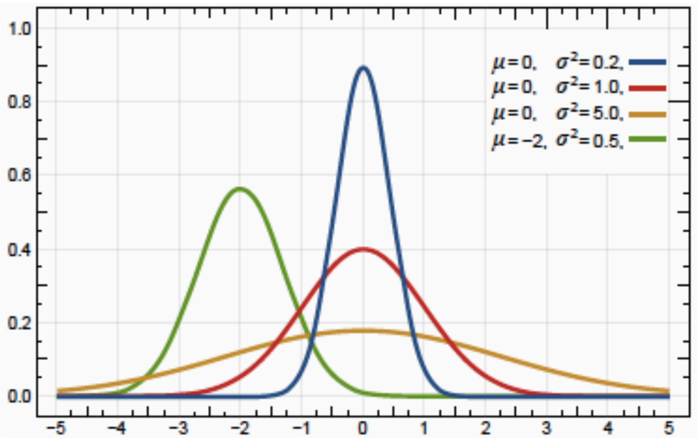
Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Randomized algorithm A

Query q

Data users (ex: government, researchers, companies, or adversary)

Answer + Gaussian$(0, 2\log(1/\delta) GS_q^2/\epsilon^2)$



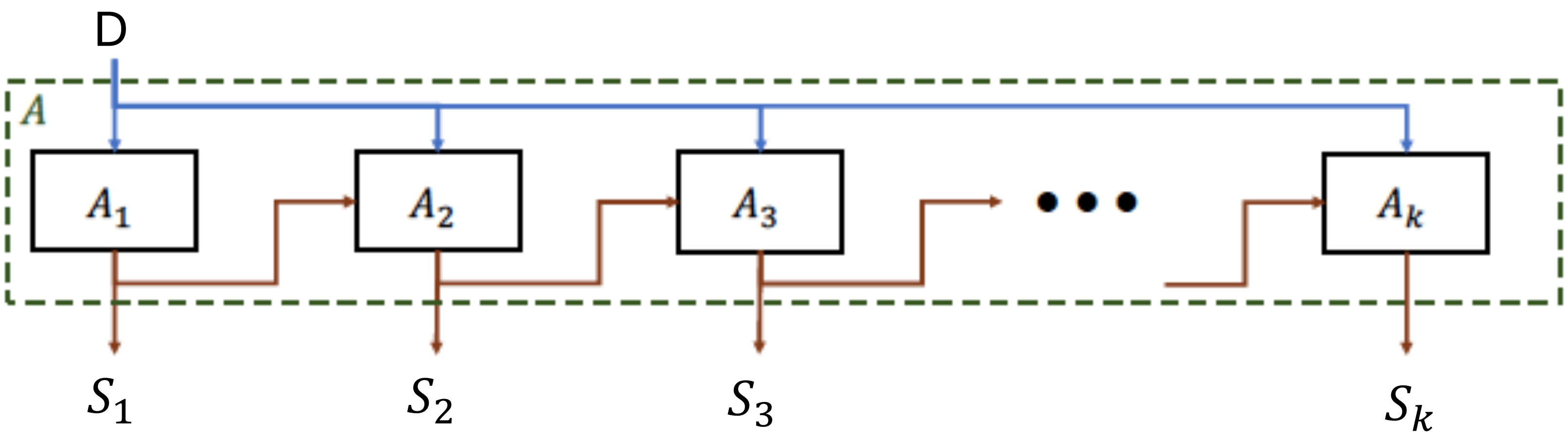
$$p(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Theorem:

$A(D,q) = q(D) + \text{Gaussian}(0, 2\log(1/\delta) GS_q^2/\epsilon^2)$

is (ϵ,δ)-DP (approximate DP)

Advanced Composition for Approximate DP



$A_1: D \to S_1$ is (ϵ, δ)-DP

$A_2: S_1 \times D \to S_2$ is (ϵ, δ)-DP

$A_3: S_1 \times S_2 \times D \to S_3$ is (ϵ, δ)-DP

$\vdots$

$A_k: S_1 \times S_2 \times \cdots S_{k-1} \times D \to S_k$ is(ϵ, δ)-DP

Theorem: Let $A: D \to S_1 \times S_2 \times \cdots \times S_k$ be an composed algorithm that outputs $(s_1, s_2, \cdots, s_k)$ where $s_1 = A_1(D)$, $s_2 = A_2(s_1, D), \ldots, s_k = A_k(s_1, \cdots, s_{k-1}, D)$.Then A is (ϵ', δ')-DP, where

$\epsilon' = \epsilon\sqrt{2k\log(1/\delta_0)} + k\epsilon\frac{e^{\epsilon}-1}{e^{\epsilon}+1}$ and $\delta' = k\delta + \delta_0$

Some constant > 0

Lower order term ($e^{\epsilon} \approx 1 + \epsilon$ for small ϵ)

Selected References

- Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy.
- Simson Gar nkel, John M Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data.
- Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets.
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models.
- Sweeney, L. (2002). k-anonymity: A model for protecting privacy.
- Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis.
- Aloni Cohen and Kobbi Nissim. Linear program reconstruction in practice.
- Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts.
- Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy.
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models.