CS 6431

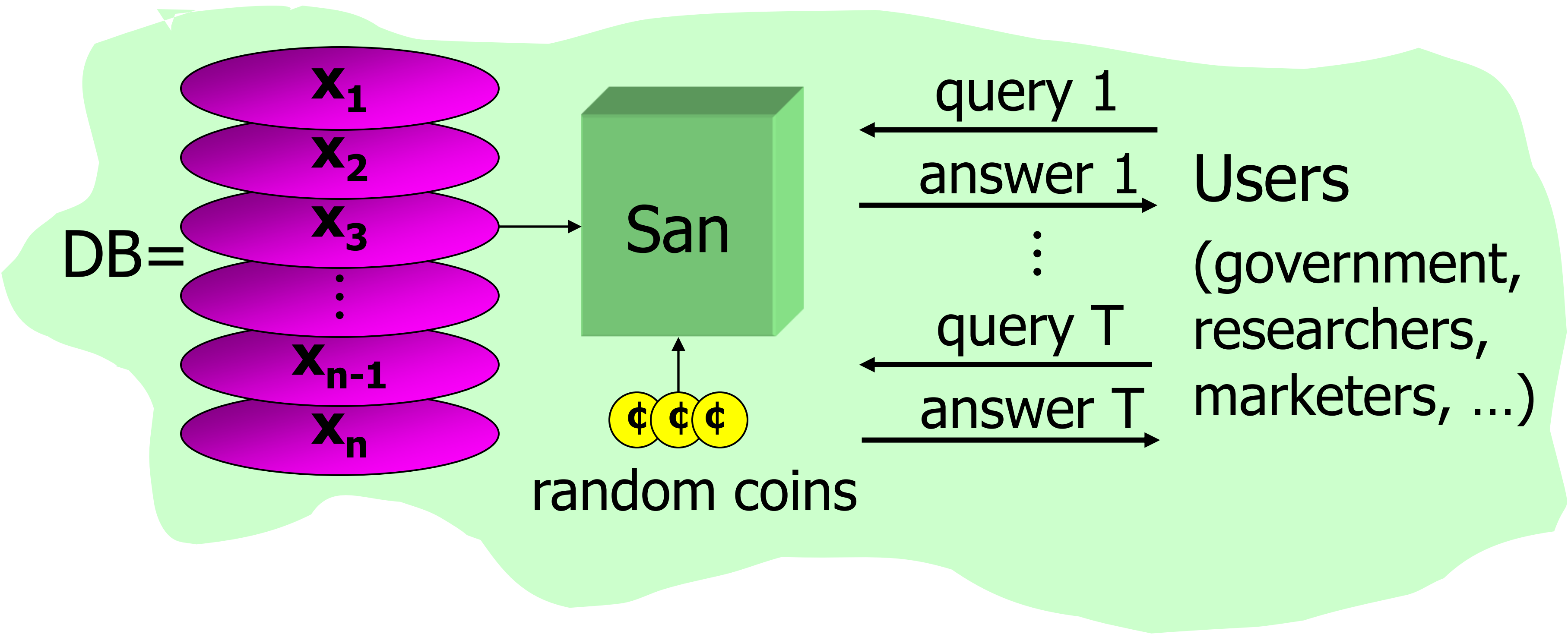
Data Privacy

Vitaly Shmatikov

Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases …
- Genetic datasets
  - $1000 genome, HapMap, DeCODE …
- Demographic datasets
  - U.S. Census Bureau, sociology studies …
- Search logs, recommender systems, social networks, blogs …
  - AOL search data, online social networks, Netflix movie ratings, Amazon …

Basic Setting



Examples of Sanitization Methods

- Input perturbation
  - Add random noise to database, release
- Summary statistics
  - Means, variances
  - Marginal totals
  - Regression coefficients
- Output perturbation
  - Summary statistics with noise
- Interactive versions of the above methods
  - Auditor decides which queries are OK, type of noise

Data “Anonymization”

- How?
- Remove “personally identifying information” (PII)
  - Name, Social Security number, phone number, email, address… what else?
- Problem: PII has no technical meaning
  - Defined in disclosure notification laws
    - If certain information is lost, consumer must be notified
  - In privacy breaches, any information can be personally identifying
    - Examples: AOL dataset, Netflix Prize dataset

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Observation #1: Dataset Joins

- Attacker learns sensitive data by joining two datasets on common attributes
  - Anonymized dataset with sensitive attributes
    - Example: age, race, <u>symptoms</u>
  - "Harmless" dataset with individual identifiers
    - Example: name, address, age, race
- Demographic attributes (age, ZIP code, race, etc.) are very common in datasets with information about individuals

Observation #2: Quasi-Identifiers

- Sweeney's observation:

  (birthdate, ZIP code, gender) uniquely identifies 87% of US population

  - Side note: actually, only 63% [Golle, WPES '06]
- Publishing a record with a quasi-identifier is as bad as publishing it with an explicit identity
- Eliminating quasi-identifiers is not desirable
  - For example, users of the dataset may want to study distribution of diseases by age and ZIP code

k-Anonymity

- Proposed by Samarati and/or Sweeney (1998)
- Hundreds of papers since then
  - Extremely popular in the database and data mining communities (SIGMOD, ICDE, KDD, VLDB)
- NP-hard in general, but there are many practically efficient k-anonymization algorithms
- Most based on generalization and suppression

Anonymization in a Nutshell

- Dataset is a relational table
- Attributes (columns) are divided into quasi-identifiers and sensitive attributes

Race	Age	Symptoms	Blood type	Medical history
…	…	…	…	…
…	…	…	…	…

quasi-identifiers

sensitive attributes

- Generalize/suppress quasi-identifiers, don't touch sensitive attributes (keep them "truthful")

k-Anonymity: Definition

- Any (transformed) quasi-identifier must appear in at least k records in the anonymized dataset
  - k is chosen by the data owner (how?)
  - Example: any age-race combination from original DB must appear at least 10 times in anonymized DB
- Guarantees that any join on quasi-identifiers with the anonymized dataset will contain at least k records for each quasi-identifier

Two (and a Half) Interpretations

- Membership disclosure: Attacker cannot tell that a given person in the dataset
- Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute
- Identity disclosure: Attacker cannot tell which record corresponds to a given person

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers

But this does not imply any privacy!

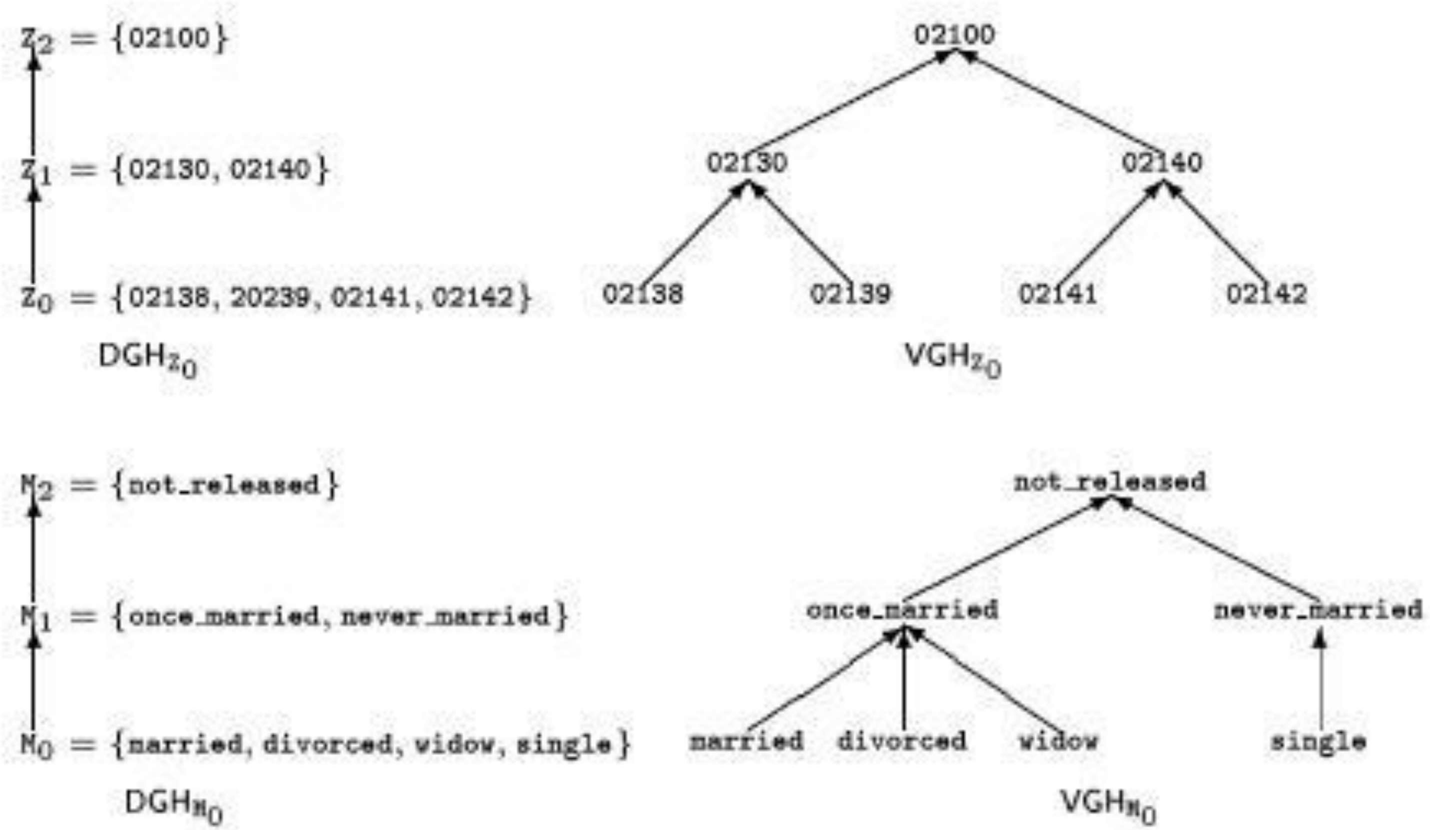
Example: k clinical records, all HIV+

Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with more general values until get k identical values
    - Example: area code instead of phone number
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with "outliers" (come back to this later)
- Lots of algorithms in the literature
  - Aim to produce "useful" anonymizations

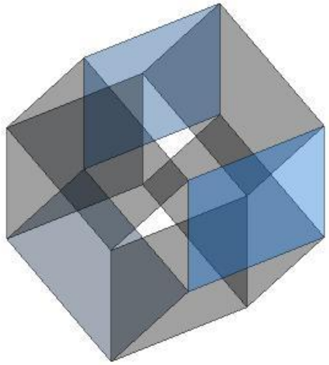
    … usually without any clear notion of utility

Generalization in Action



Curse of Dimensionality

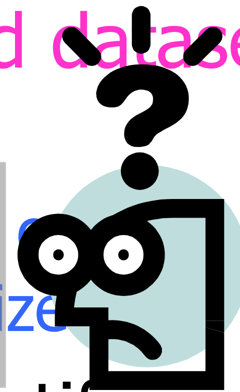
[Aggarwal VLDB '05]



- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info $\Rightarrow$ k-anonymized datasets are useless

k-Anonymity: Definition

- Any (transformed) quasi-identifier must appear in at least k records in the anonymized dataset
  - k is chos… he data owner (how?)
  - E… DB
    n… ize



This definition does not mention sensitive attributes at all!

- Guarantees that any join on quasi-identifiers with… in at leas…

Assumes that attacker will be able to join only on quasi-identifiers

Does not say anything about the computations that are to be done on the data

Membership Disclosure

- With large probability, quasi-identifier is unique in the population
- But generalizing/suppressing quasi-identifiers in the dataset does not affect their distribution in the population (obviously)!
  - Suppose anonymized dataset contains 10 records with a certain quasi-identifier …

    … and there are 10 people in the population who match this quasi-identifier
- k-anonymity may <u>not</u> hide whether a given person is in the dataset

Sensitive Attribute Disclosure

Intuitive reasoning:

- k-anonymity prevents attacker from telling which record corresponds to which person
- Therefore, attacker cannot tell that a certain person has a particular value of a sensitive attribute

This reasoning is fallacious!

3-Anonymization

Caucas	78712	Flu
Asian	78705	Shingles
Caucas	78754	Flu
Asian	78705	Acne
AfrAm	78705	Acne
Caucas	78705	Flu

Caucas	787XX	Flu
Asian/AfrAm	78705	Shingles
Caucas	787XX	Flu
Asian/AfrAm	78705	Acne
Asian/AfrAm	78705	Acne
Caucas	787XX	Flu

This is 3-anonymous, right?

Joining With External Database

…	…	…
Rusty Shackleford	Caucas	78705
…	…	…

+

Caucas	787XX	Flu
Asian/AfrAm	78705	Shingles
Caucas	787XX	Flu
Asian/AfrAm	78705	Acne
Asian/AfrAm	78705	Acne
Caucas	787XX	Flu

Problem: sensitive attributes are not "diverse" within each quasi-identifier group

Another Attempt: l-Diversity

[Machanavajjhala et al. ICDE '06]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Entropy of sensitive attributes within each quasi-identifier group must be at least L

Still Does Not Work

Original database

…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Cancer
Q2	Flu

99% cancer ⇒ quasi-identifier group is <u>not</u> "diverse"
…yet anonymized database does not leak anything

50% cancer ⇒ quasi-identifier group is "diverse"
This leaks a ton of information

Try Again: t-Closeness

[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Trick question: Why publish quasi-identifiers at all??

Anonymized "t-Close" Database

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous,
l-diverse and t-close…

…so secure, right?

What Does Attacker Know?

Bob is white and
I heard he was
admitted to hospital
with flu…

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
		HIV+	Shingles
Caucas	787XX	HIV-	Acne
		HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is against the rules!
“flu” is not a quasi-identifier

Yes… and this is yet another
problem with k-anonymity

Issues with Syntactic Definitions

- What adversary do they apply to?
  - Do not consider adversaries with side information
  - Do not consider probability
  - Do not consider adversarial algorithms for making decisions (inference)
- Any attribute is a potential quasi-identifier
  - External / auxiliary / background information about people is very easy to obtain

Classical Intution for Privacy

- Dalenius (1977): "If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S, a disclosure has taken place"
  - Privacy means that anything that can be learned about a respondent from the statistical database can be learned without access to the database
- Similar to semantic security of encryption
  - Anything about the plaintext that can be learned from a ciphertext can be learned without the ciphertext

Problems with Classic Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
  - What if my (incorrect) prior is that every Cornell graduate student has three arms?
- How much is "too much?"
  - Can't achieve cryptographically small levels of disclosure and keep the data useful
  - Adversarial user is supposed to learn unpredictable things about the database

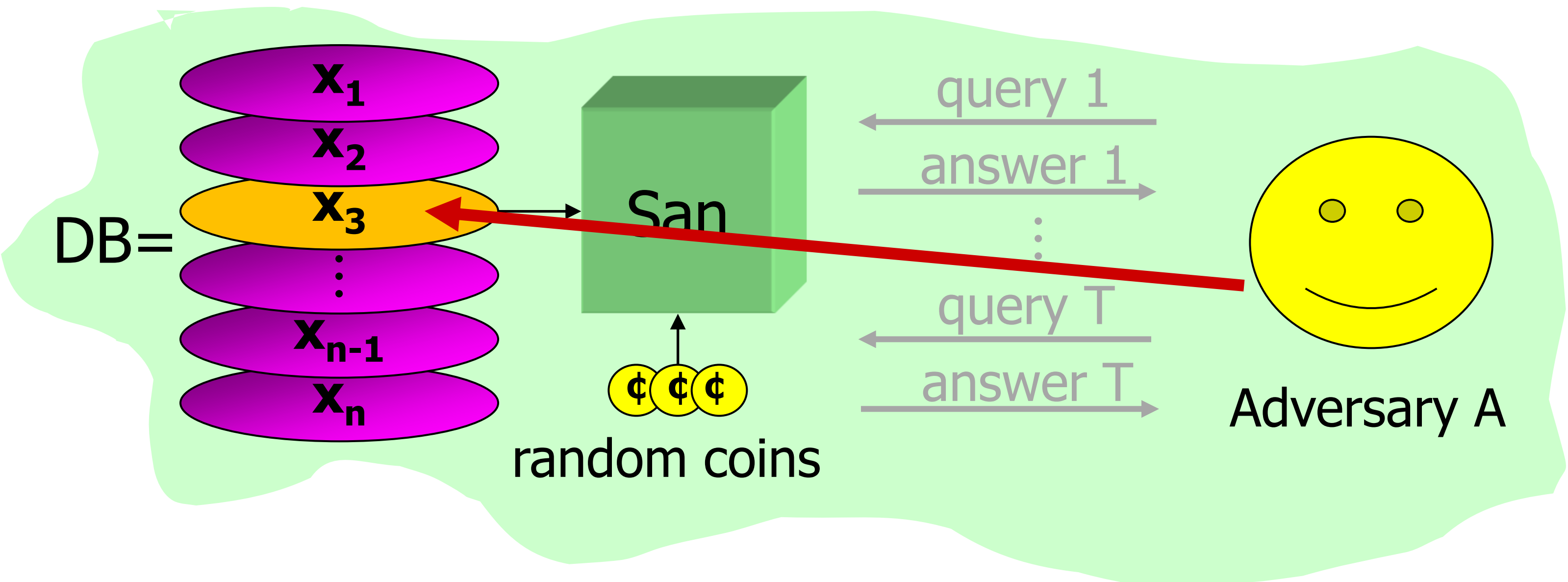
Absolute Guarantee Unachievable

[Dwork]

- Privacy: for some definition of "privacy breach," $\forall$ distribution on databases, $\forall$ adversaries A, $\exists$ A′ such that $\Pr(A(San)=breach) - \Pr(A'()=breach) \le \varepsilon$
  - For reasonable "breach", if San(DB) contains information about DB, then some adversary breaks this definition
- Example
  - I know that you are 2 inches taller than the average Russian
  - DB allows computing average height of a Russian
  - This DB breaks your privacy according to this definition… even if your record is not in the database!

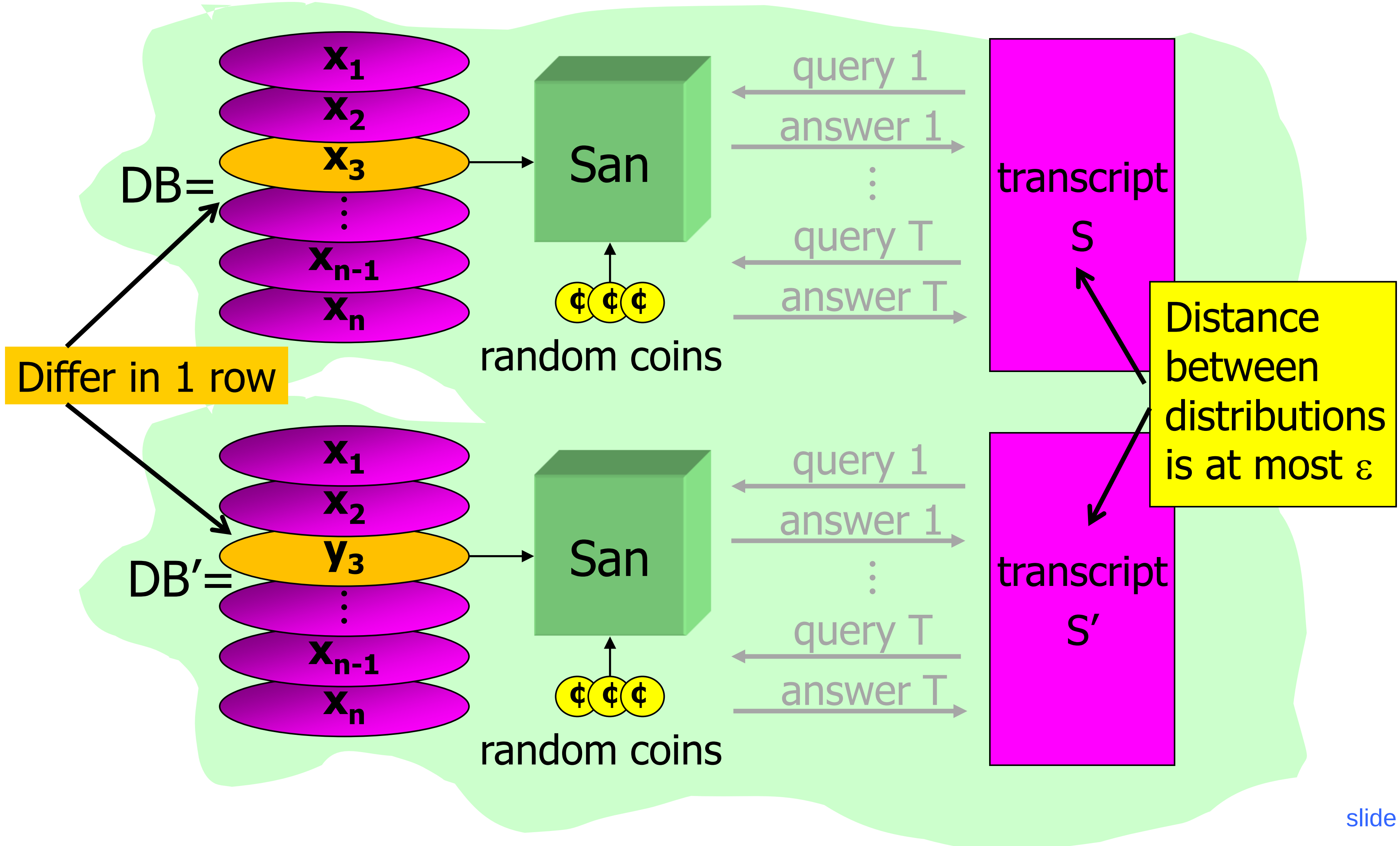
Differential Privacy

[Dwork]



- Absolute guarantees are problematic
  - Your privacy can be "breached" (per absolute definition of privacy) even if your data is not in the database
- Relative guarantee: "Whatever is learned would be learned regardless of whether or not you participate"
  - Dual: Whatever is already known, situation won't get worse

Indistinguishability



Which Distance to Use?

- Problem: ε must be large
  - Any two databases induce transcripts at distance $\leq n\varepsilon$
  - To get utility, need $\varepsilon > 1/n$
- Statistical difference $1/n$ is not meaningful!
  - Example: release a random point from the database
    - $San(x_1,\ldots,x_n) = (j, x_j)$ for random j
  - For every i, changing x_i induces statistical difference $1/n$
  - But some x_i is revealed with probability 1
    - Definition is satisfied, but privacy is broken!

Formalizing Indistinguishability

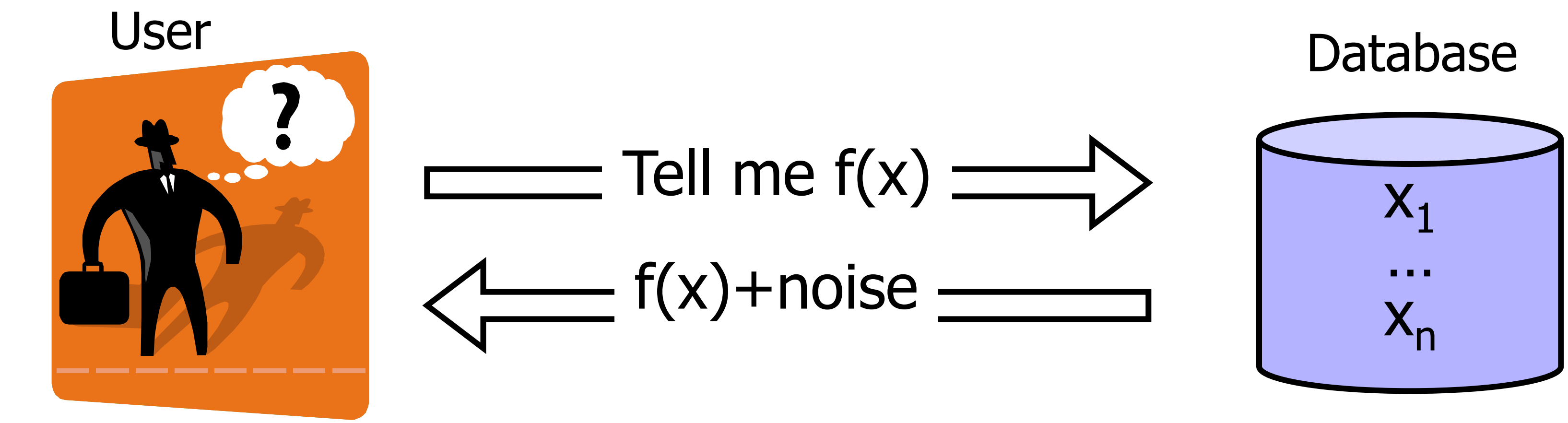
Definition: San is ε-indistinguishable if

$\forall$ A, $\forall$ $\underline{\text{DB}}$, $\underline{\text{DB}}'$ which differ in 1 row, $\forall$ sets of transcripts S

$$p(\ \text{San}(\text{DB}) \in S\ ) \in (1 \pm \varepsilon)\ p(\ \text{San}(\text{DB}') \in S\ )$$

Equivalently, $\forall$ S: $\dfrac{p(\ \text{San}(\text{DB}) = S\ )}{p(\ \text{San}(\text{DB}') = S\ )} \in 1 \pm \varepsilon$

Laplacian Mechanism



- Intuition: f(x) can be released accurately when f is insensitive to individual entries $x_1, \ldots x_n$
- Global sensitivity $GS_f = \max_{\text{neighbors } x,x'} ||f(x) - f(x')||_1$
  - Example: $GS_{\text{average}} = 1/n$ for sets of bits

Lipschitz constant of f

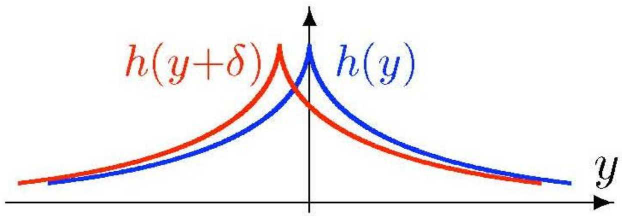
- Theorem: $f(x) + Lap(GS_f/\varepsilon)$ is ε-indistinguishable
  - Noise generated from Laplace distribution

Sensitivity with Laplace Noise

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

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



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

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

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

Differential Privacy: Summary

- San gives ε-differential privacy if for all values of DB and Me and all transcripts t:

$$\frac{\Pr[\, \mathcal{S}an(\mathrm{DB} - \mathrm{Me}) = t\,]}{\Pr[\, \mathcal{S}an(\mathrm{DB} + \mathrm{Me}) = t\,]} \leq e^{\varepsilon} \approx 1 \pm \varepsilon$$