Differential Privacy

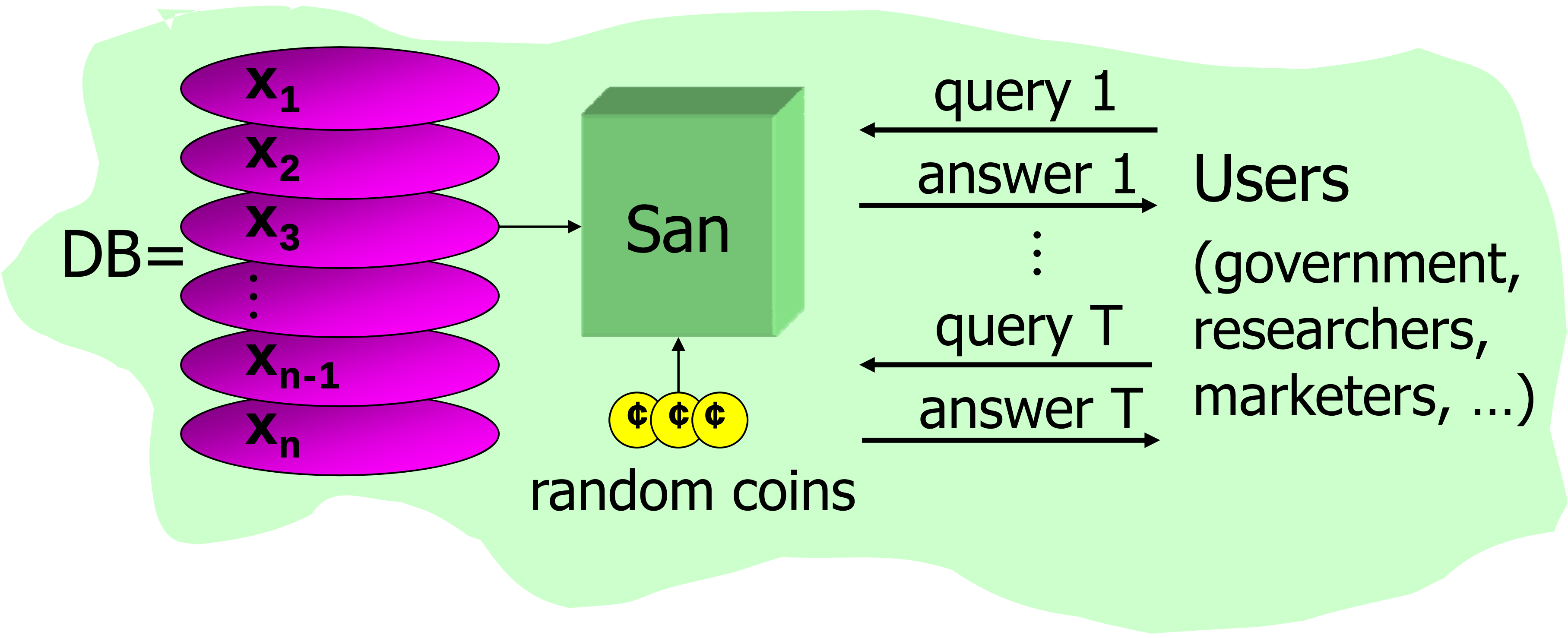
Vitaly Shmatikov

most slides from Adam Smith (Penn State)

Reading Assignment

- Dwork. "Differential Privacy" (invited talk at ICALP 2006).

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

Strawman Definition

- Assume $x_1,\ldots,x_n$ are drawn i.i.d. from unknown distribution
- Candidate definition: sanitization is safe if it only reveals the distribution
- Implied approach:
  - Learn the distribution
  - Release description of distribution or re-sample points
- This definition is tautological!
  - Estimate of distribution depends on data… why is it safe?

Blending into a Crowd

Frequency in DB or frequency in underlying population?

- Intuition: "I am safe in a group of k or more"
  - k varies (3… 6… 100… 10,000?)
- Many variations on theme
  - Adversary wants predicate g such that $0 < \#\{i \mid g(x_i)=\text{true}\} < k$

- Why?
  - Privacy is "protection from being brought to the attention of others" [Gavison]
  - Rare property helps re-identify someone
  - Implicit: information about a large group is public
    - E.g., liver problems more prevalent among diabetics

Clustering-Based Definitions

- Given sanitization S, look at all databases consistent with S
- Safe if no predicate is true for all consistent databases
- k-anonymity
  - Partition D into bins
  - Safe if each bin is either empty, or contains at least k elements
- Cell bound methods
  - Release marginal sums

	brown	blue	Σ
blond	2	10	12
brown	12	6	18
Σ	14	16	

⇩

	brown	blue	Σ
blond	[0,12]	[0,12]	12
brown	[0,14]	[0,16]	18
Σ	14	16	

Issues with Clustering

- Purely syntactic definition of privacy
- What adversary does this apply to?
  - Does not consider adversaries with side information
  - Does not consider probability
  - Does not consider adversarial algorithm for making decisions (inference)

"Bayesian" Adversaries

- Adversary outputs point $z \in D$
- Score = $1/f_z$ if $f_z > 0$, 0 otherwise
  - f_z is the number of matching points in D
- Sanitization is safe if $E(\text{score}) \leq \varepsilon$
- Procedure:
  - Assume you know adversary's prior distribution over databases
  - Given a candidate output, update prior conditioned on output (via Bayes' rule)
  - If $\max_z E(\text{ score } | \text{ output }) < \varepsilon$, then safe to release

Issues with “Bayesian” Privacy

- Restricts the type of predicates adversary can choose
- Must know prior distribution
  - Can one scheme work for many distributions?
  - Sanitizer works harder than adversary
- Conditional probabilities don’t consider previous iterations
  - Remember simulatable auditing?

Classical Intution for Privacy

- "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." [Dalenius 1977]
  - 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 UTCS graduate student has three arms?
- How much is "too much?"
  - Can't achieve cryptographically small levels of disclosure <u>and</u> keep the data useful
  - Adversarial user is <u>supposed</u> to learn unpredictable things about the database

Impossibility Result

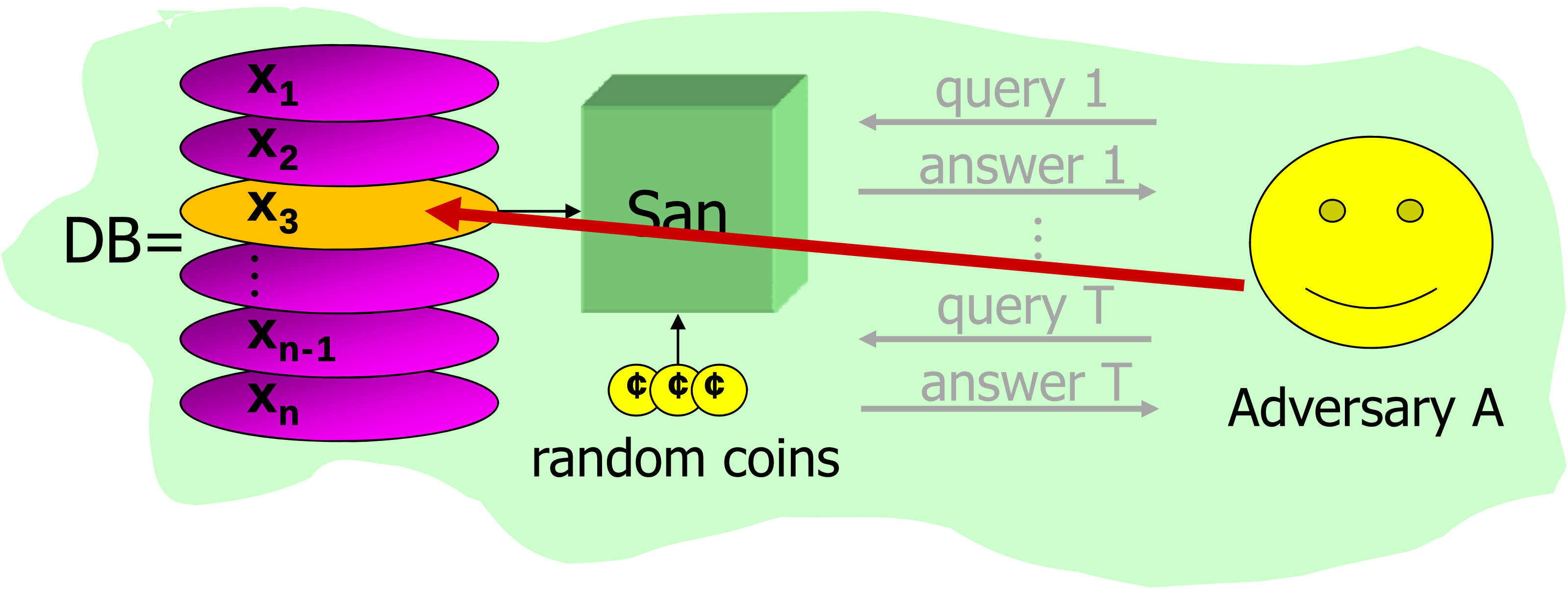
[Dwork]

- <u>Privacy</u>: for some definition of "privacy breach," $\forall$ distribution on databases, $\forall$ adversaries A, $\exists$ A′ such that $Pr(A(San)=breach) - Pr(A'()=breach) \leq \varepsilon$
  - For reasonable "breach", if San(DB) contains information about DB, then some adversary breaks this definition
- Example
  - Vitaly knows that Alex Benn is 2 inches taller than the average Russian
  - DB allows computing average height of a Russian
  - This DB breaks Alex's privacy according to this definition… even if his record is <u>not</u> in the database!

(Very Informal) Proof Sketch

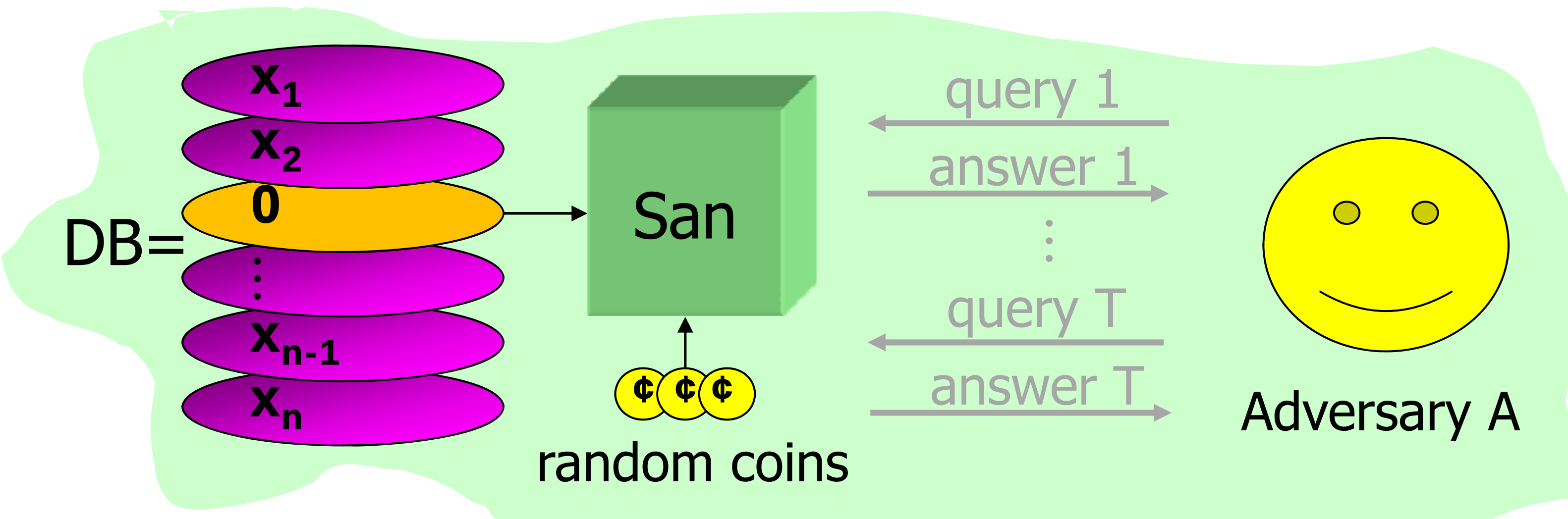
- Suppose DB is uniformly random
  - Entropy I(DB ; San(DB)) > 0
- "Breach" is predicting a predicate g(DB)
- Adversary knows r, H(r ; San(DB)) $\oplus$ g(DB)
  - H is a suitable hash function, r=H(DB)
- By itself, does not leak anything about DB (why?)
- Together with San(DB), reveals g(DB) (why?)

Differential Privacy (1)



- Example with Russians and Alex Benn
  - Adversary learns Alex's height even if he is not in the database
- Intuition: "Whatever is learned would be learned regardless of whether or not Alex participates"
  - Dual: Whatever is already known, situation won't get worse

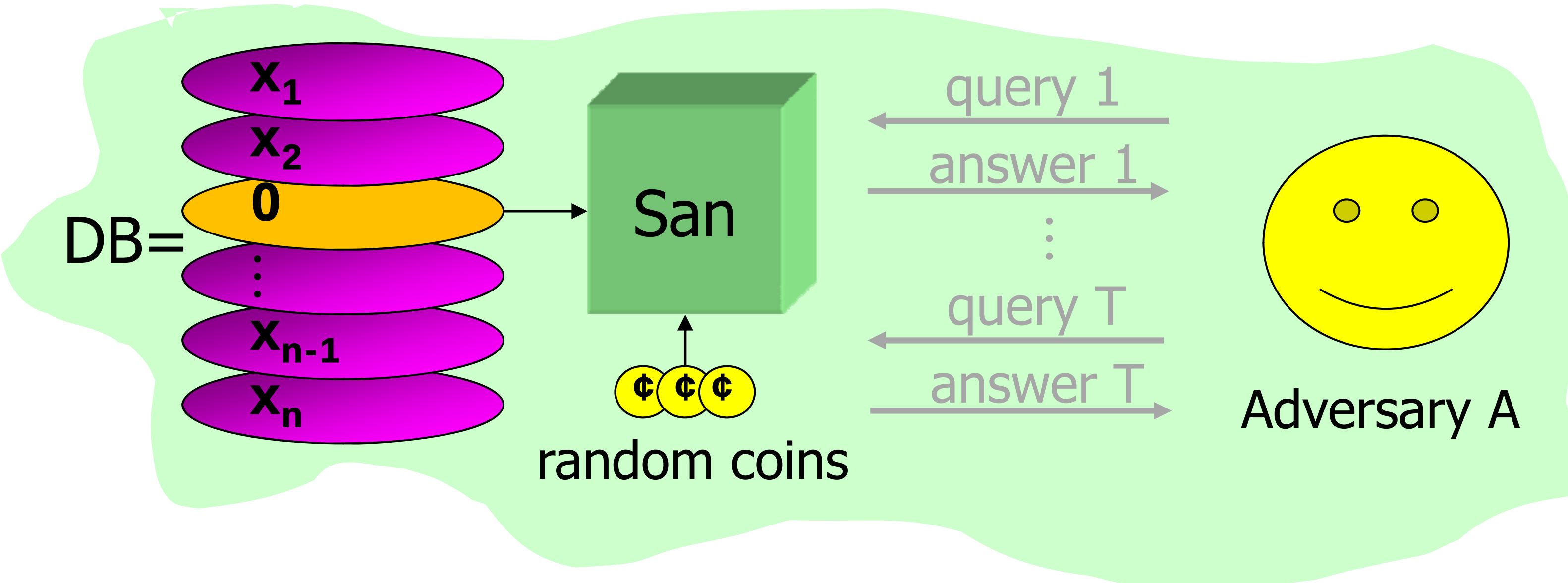
Differential Privacy (2)



- Define n+1 games
  - Game 0: Adv. interacts with San(DB)
  - Game i: Adv. interacts with San(DB_{-i}); $DB_{-i} = (x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n)$
- Given S and prior p() on DB, define n+1 posterior distrib's

$$p_i(DB|S) = p(DB|S \text{ in Game } i) = \frac{p(San(DB_{-i}) = S) \times p(DB)}{p(S \text{ in Game } i)}$$

Differential Privacy (3)



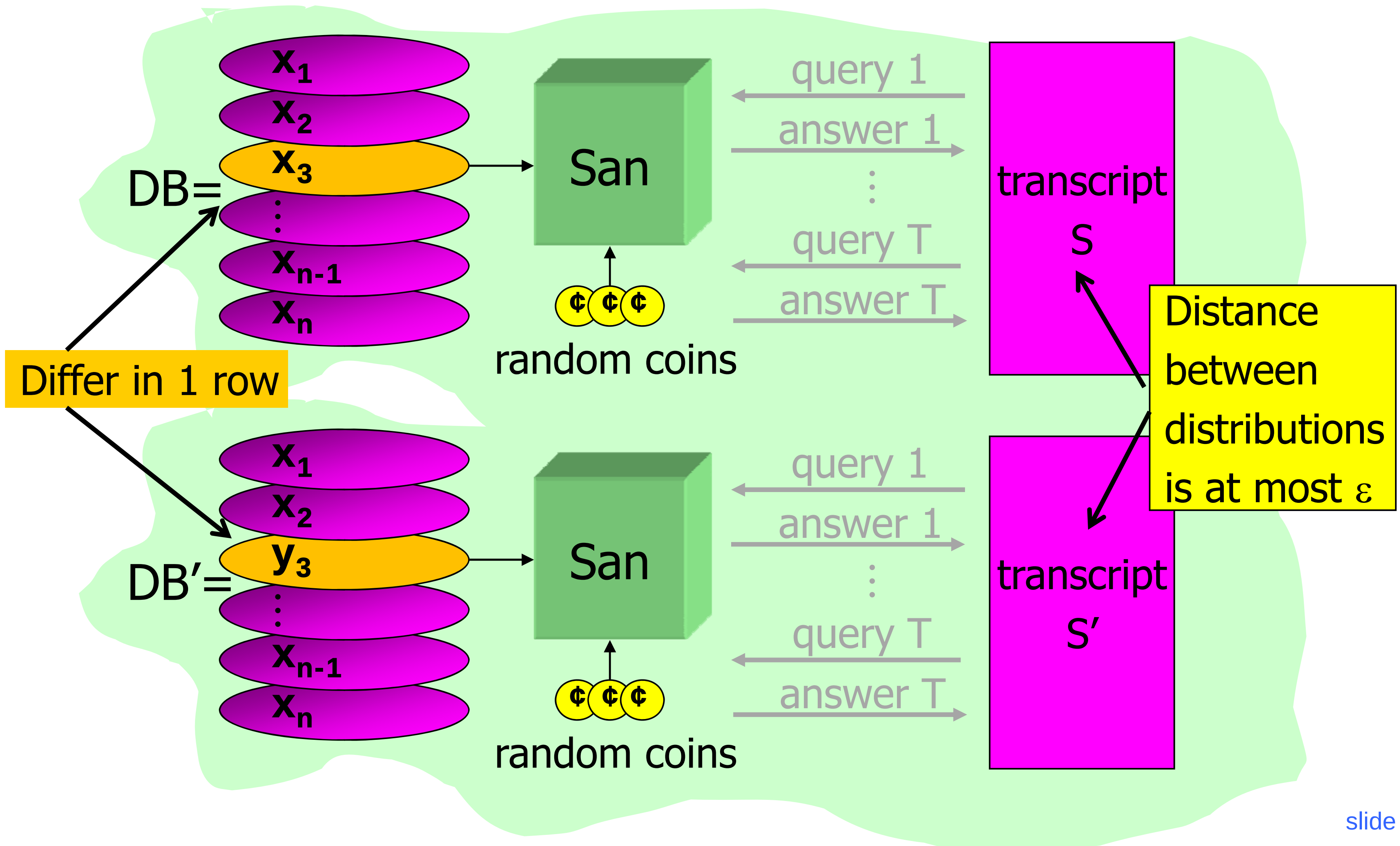
Definition: San is safe if

∀ prior distributions p(¢) on DB,

∀ transcripts S, ∀ i =1,…,n

$$\text{StatDiff}(\ p_0(¢|S)\ ,\ p_i(¢|S)\ ) \le \varepsilon$$

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 random point in 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

Formalizing Indistinguishability

Definition: San is ε-indistinguishable if

$\forall$ A, $\forall$ DB, DB′ which differ in 1 row, $\forall$ sets of transcripts S

$$p(\, San(DB) \in S \,) \in (1 \pm \varepsilon)\; p(\, San(DB') \in S \,)$$

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

Indistinguishability ⇒ Diff. Privacy

Definition: San is safe if

∀ prior distributions p(¢) on DB,

∀ transcripts S, ∀ i =1,…,n

StatDiff(p_0(¢|S) , p_i(¢|S)) ≤ ε

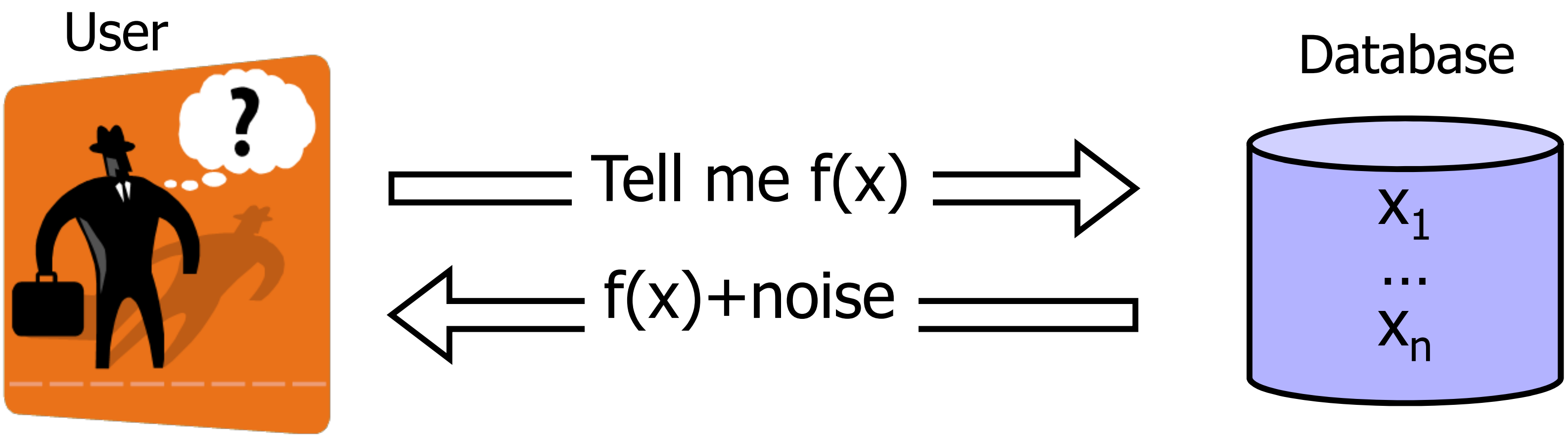
$$p_i(DB|S) = p(DB|S \text{ in Game } i) = \frac{p(San(DB_{-i}) = S) \times p(DB)}{p(S \text{ in Game } i)}$$

For every S and DB, indistinguishability implies

$$\frac{p_i(DB|S)}{p_0(DB|S)} = \frac{p(San(DB_{-i}) = S)}{p(San(DB) = S)} \times \frac{p(S \text{ in Game } 0)}{p(S \text{ in Game } i)} \approx 1 \pm 2\epsilon$$

This implies StatDiff(p_0(¢|S) , p_i(¢| S)) ≤ ε

Diff. Privacy in Output Perturbation



- 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: $GS_{average} = 1/n$ for sets of bits
- Theorem: $f(x) + Lap(GS_f / \varepsilon)$ is ε-indistinguishable
  - Noise generated from Laplace distribution

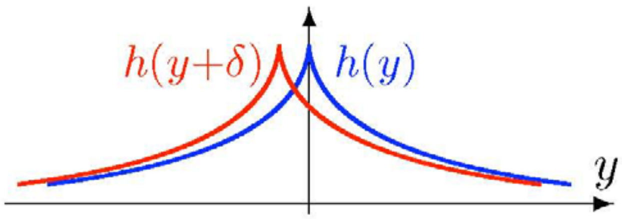
Lipschitz constant of f

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

Proof idea:

$$A(x): \text{ blue curve}$$

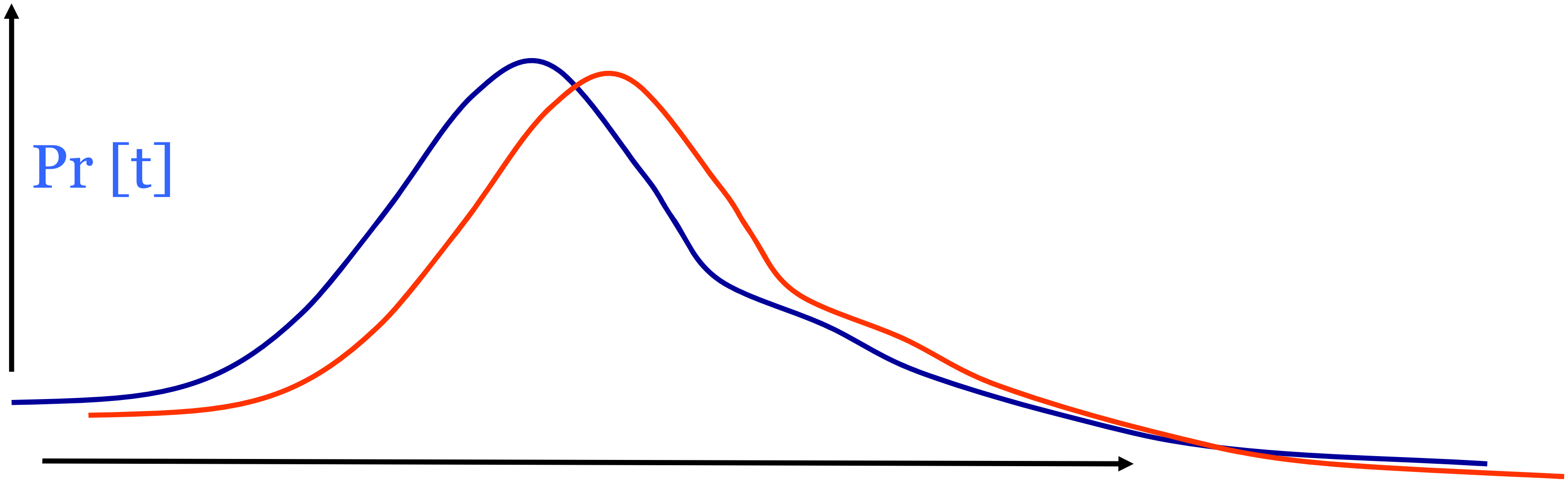
$$A(x'): \text{ red curve}$$

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

Differential Privacy: Summary

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

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



Intuition

- No perceptible risk is incurred by joining DB
- Anything adversary can do to me, it could do without me (my data)