

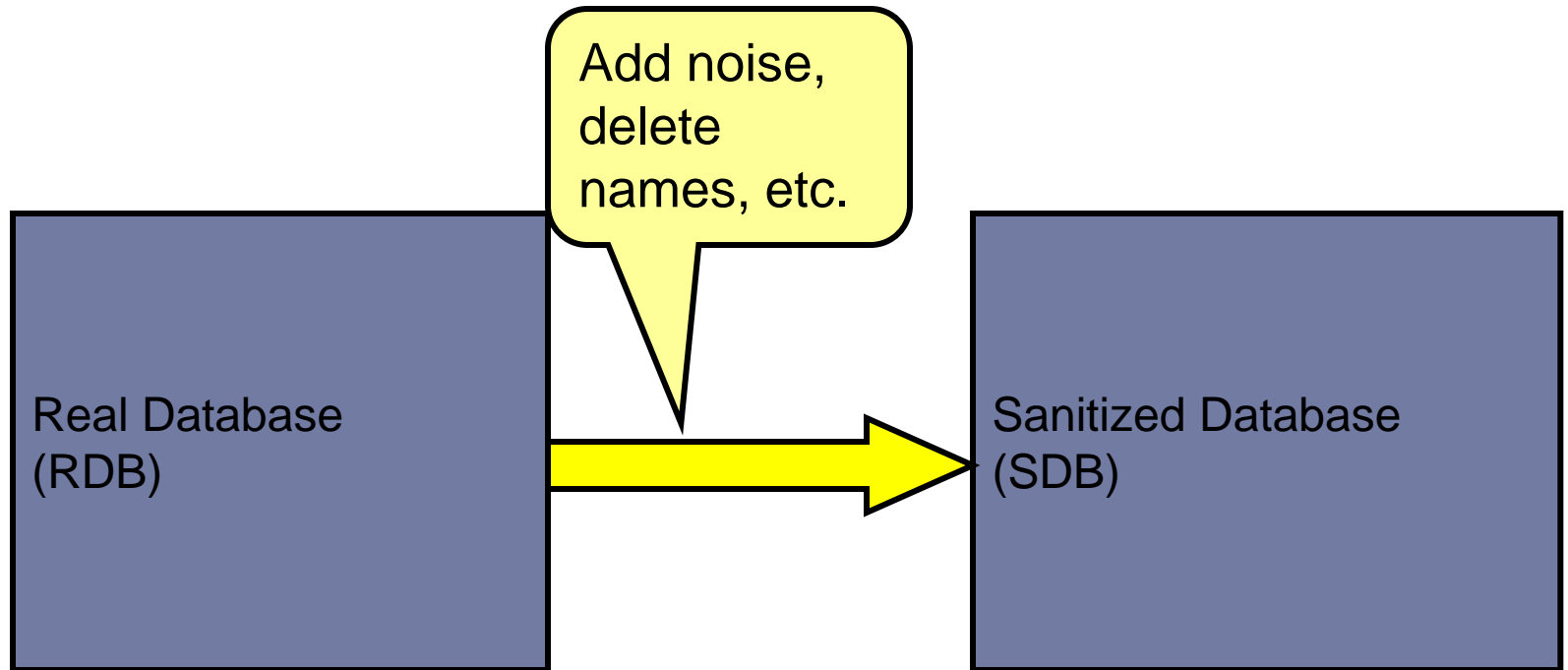
18739A: Foundations of Security and Privacy

Differential Privacy

Anupam Datta

Fall 2011

Sanitization of Databases



Health records

Census data

Protect privacy

Provide useful
information
(utility)



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, \dots, 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

- ▶ Intuition: “I am safe in a group of k or more”
 - ▶ k varies (3... 6... 100... 10,000?)



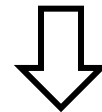
- ▶ Why?
 - ▶ Privacy is “protection from being brought to the attention of others” [Gavison]
 - ▶ Rare property helps re-identify someone
 - ▶ Implicit: information about a large group is public
 - ▶ E.g., liver problems more prevalent among diabetics



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

	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)



Classical Intuition 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



Impossibility Result

[Dwork, Naor 2006]

- ▶ Privacy: For some definition of “privacy breach,”
 \forall distribution on databases, \forall adversaries A , $\exists A'$
such that $\Pr(A(\text{San})=\text{breach}) - \Pr(A'(\text{DB})=\text{breach}) \leq \epsilon$
 - ▶ **Result**: For reasonable “breach”, if $\text{San}(\text{DB})$ contains information about DB, then some adversary breaks this definition
- ▶ **Example**
 - ▶ Terry Gross is two inches shorter than the average Lithuanian woman
 - ▶ DB allows computing average height of a Lithuanian woman
 - ▶ This DB breaks Terry Gross’s privacy according to this definition... **even if her record is not in the database!**



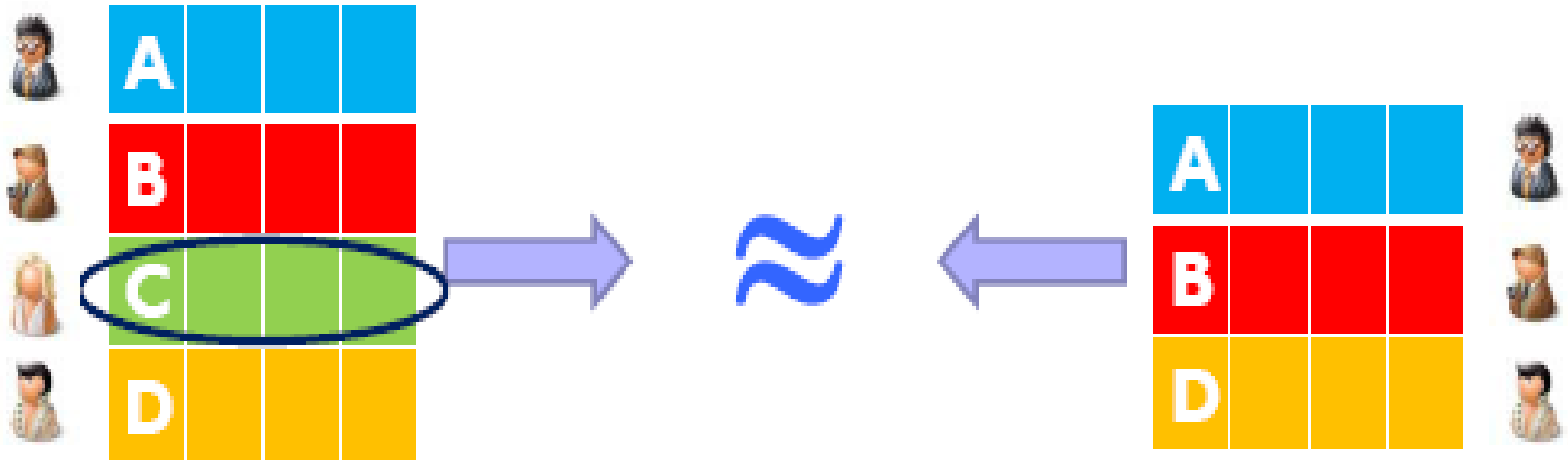
(Very Informal) Proof Sketch

- ▶ Suppose DB is uniformly random
- ▶ “Breach” is predicting a predicate $g(\text{DB})$
- ▶ Adversary knows $r, H(r ; \text{San}(\text{DB})) \oplus g(\text{DB})$
 - ▶ H is a suitable hash function, $r=H(\text{DB})$
- ▶ By itself, does not leak anything about DB
- ▶ Together with $\text{San}(\text{DB})$, reveals $g(\text{DB})$



Differential Privacy (informal)

Output is similar whether any single individual's record is included in the database or not



C is **no worse off** because her record is included in the computation



Differential Privacy [Dwork et al 2006]

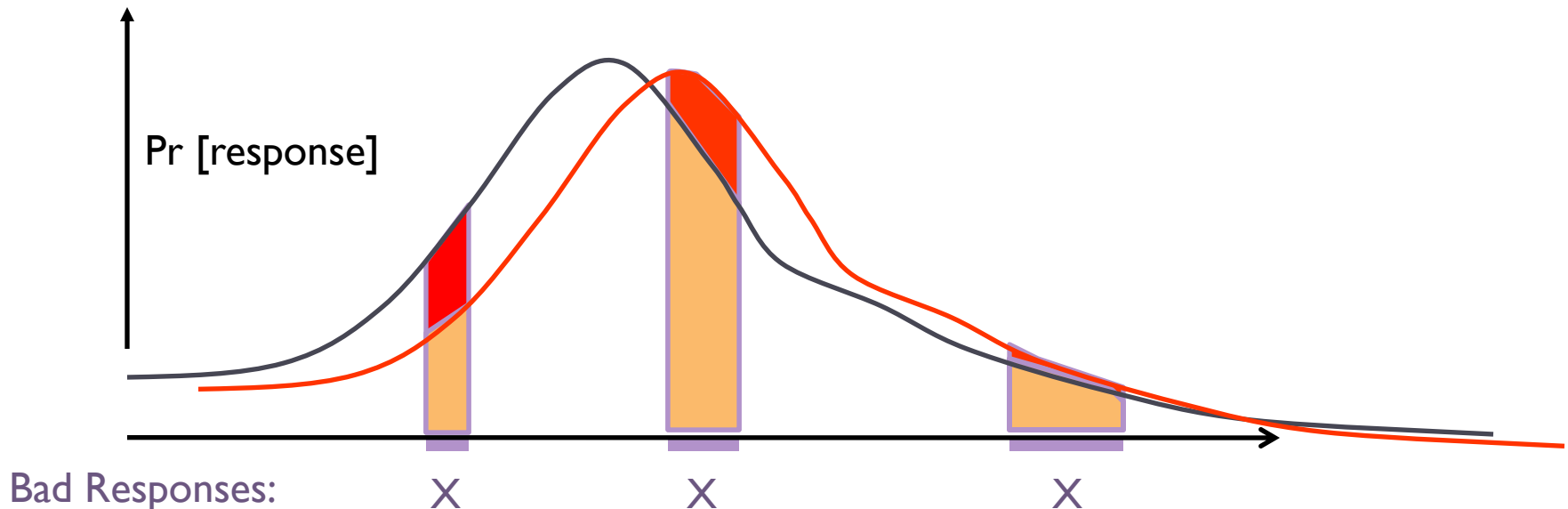
Randomized sanitization function κ has ϵ -differential privacy if for all data sets $D1$ and $D2$ differing by at most one element and all subsets S of the range of κ ,

$$\Pr[\kappa(D1) \in S] \leq e^\epsilon \Pr[\kappa(D2) \in S]$$

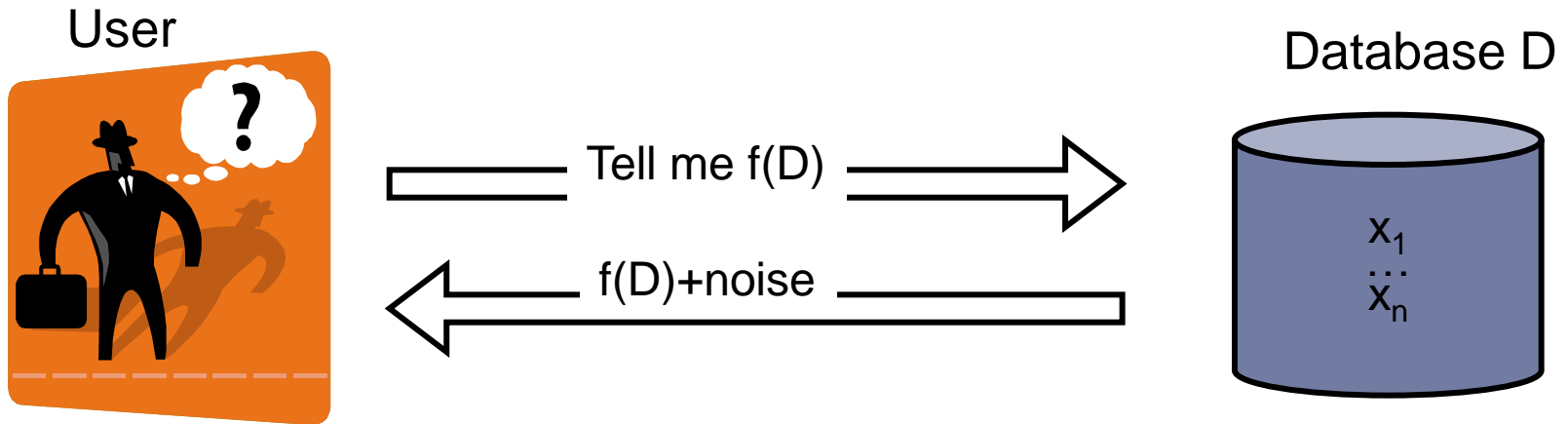


Intuition

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



Achieving Differential Privacy



- ▶ How much noise should be added?
- ▶ Intuition: $f(D)$ can be released accurately when f is insensitive to individual entries x_1, \dots, x_n (the more sensitive f is, higher the noise added)



Sensitivity of a function

We will achieve ϵ -differential privacy by the addition of random noise whose magnitude is chosen as a function of the largest change a single participant could have on the output to the query function; we refer to this quantity as the *sensitivity* of the function⁹.

Definition 3. For $f : \mathcal{D} \rightarrow \mathbb{R}^d$, the L1-sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (2)$$

for all D_1, D_2 differing in at most one element.

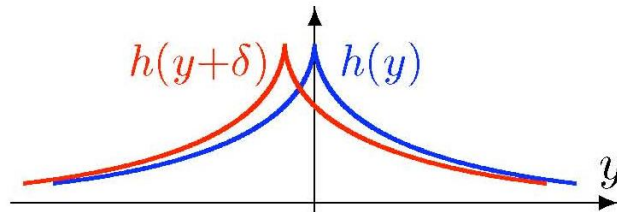
- ▶ **Examples:** $\Delta \text{count} \leq 1$, $\Delta \text{histogram} \leq 1$
- ▶ **Note:** Δf does *not* depend on the database

Sensitivity with Laplace Noise

Theorem

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

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



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

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

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

Acknowledgements

- ▶ Some slides are from Vitaly Shmatikov and Adam Smith

Achieving Differential Privacy

The privacy mechanism, denoted \mathcal{K}_f for a query function f , computes $f(X)$ and adds noise with a scaled symmetric exponential distribution with variance σ^2 (to be determined in Theorem 4) in each component, described by the density function

$$\Pr[\mathcal{K}_f(X) = a] \propto \exp(-\|f(X) - a\|_1/\sigma) \quad (3)$$

This distribution has independent coordinates, each of which is an exponentially distributed random variable. The implementation of this mechanism thus simply adds symmetric exponential noise to each coordinate of $f(X)$.

Theorem 4. *For $f : \mathcal{D} \rightarrow \mathbb{R}^d$, the mechanism \mathcal{K}_f gives $(\Delta f/\sigma)$ -differential privacy.*

Proof of Theorem 4

Proof. Starting from (3), we apply the triangle inequality within the exponent, yielding for all possible responses r

$$\Pr[\mathcal{K}_f(D_1) = r] \leq \Pr[\mathcal{K}_f(D_2) = r] \times \exp(\|f(D_1) - f(D_2)\|_1/\sigma). \quad (4)$$

The second term in this product is bounded by $\exp(\Delta f/\sigma)$, by the definition of Δf . Thus (1) holds for singleton sets $S = \{a\}$, and the theorem follows by a union bound.

“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 $\max_z E(\text{score} \mid \text{output}) < \varepsilon$, 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