The State of Data Privacy

Srivatsan Laxman
Microsoft Research India

Privacy – a basic human right

- Article 12, **Universal Declaration of Human Rights**, UN General Assembly, Paris, 1948:

"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks."

- Privacy awareness is increasing
- Growing adoption/debate about privacy regulations

Data is valuable for research

- Who collects the data?
  - Government agencies
  - Businesses
  - Non-profit organizations
- What can be considered as legitimate use?
  - Disclosure of aggregate statistics of sensitive attributes
  - Discovering aggregate trends in the data
  - Disclosure of well-defined (user-chosen) information from individual records
  - Impossible to generate an exhaustive list (hence the privacy problem must be tackled)

1995: The case of a Maryland banker

- Source: Woodward, B., 1995
- A Maryland banker obtained a list of cancer patients
- Matched cancer patients against a list of clients with outstanding loans
- When a match was found, banker called the loan
- Today, HIIPA regulations are in-place in the US:
  - Private medical information cannot be released to public
  - Institutions that release data must take steps to protect privacy of individuals in the data
- De-identification is a basic first-step – no names, social security numbers, home addresses, etc.

2002: Massachusetts Group Insurance privacy breach

- Source: Sweeney, L., 2002
- Massachusetts Group Insurance data was released publicly for research purposes
- It was possible to link released data to publicly available voting records
- Medical information of former Massachusetts governor William Weld was compromised
- Estimated 87% (216 million of 248 million) of US population can be uniquely identified by just 3 seemingly non-sensitive attributes: ZIP code, gender and DOB

2006: AOL search history privacy breach

- Source: Declan McCullough, CNET News.com, Aug 7, ’06
- AOL published search histories of 650,000 users
  - Names and user identities were supposedly removed
  - Random, but unique ID associated with each user
- What AOL knows about user with ID 710794:
  - Overweight golfer
  - Owns 1986 Porsche & 1998 Cadillac SLS
  - Fan of Univ. of Tennessee basketball team
  - Interested in Cherokee County School District, Canton, GA
  - Regularly searches for “lolitas” (porn)

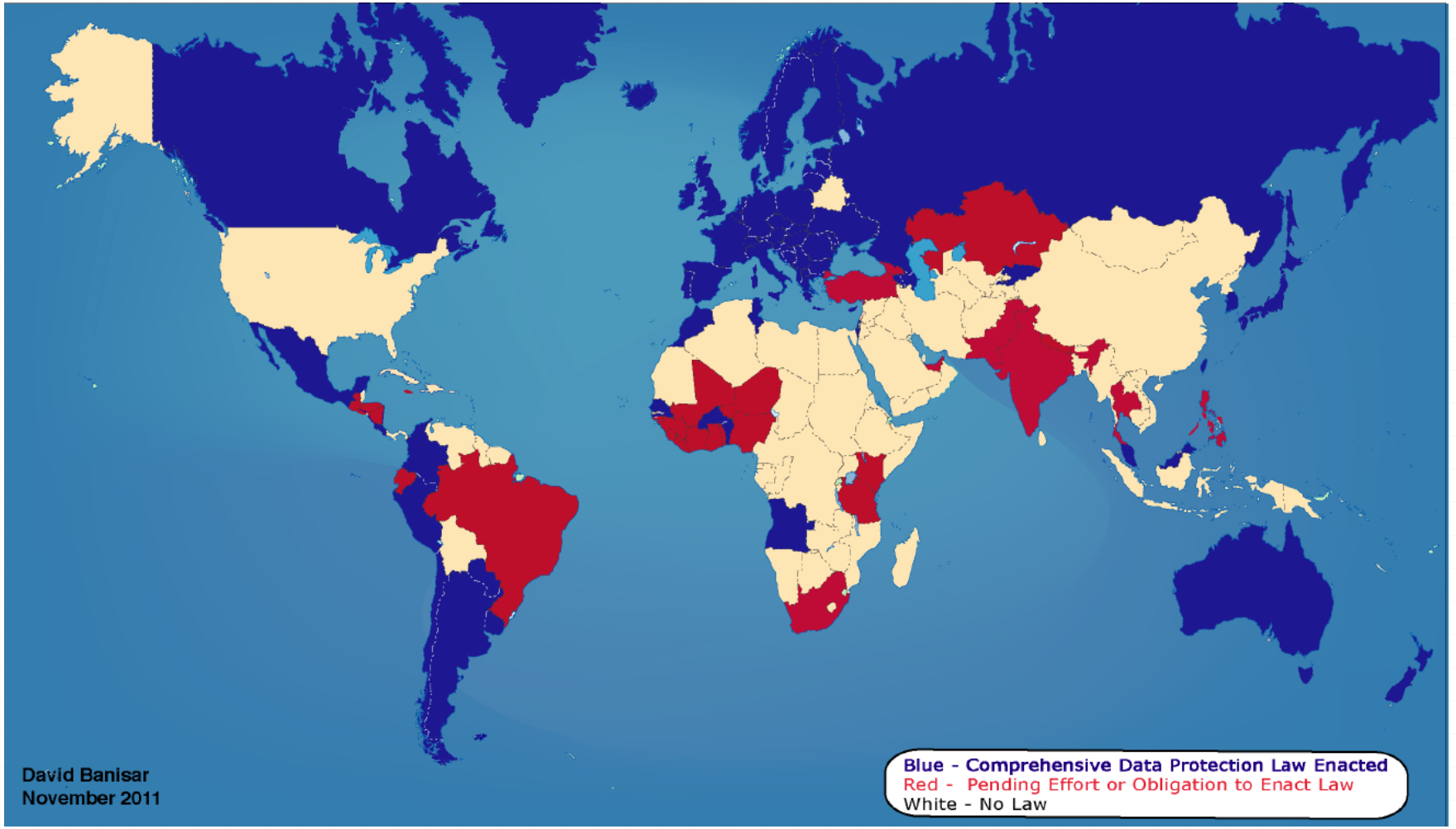
2008: Attack on Netflix data

- Source: Narayanan and Shmatikov (2008)
- Netflix data has 100 million ratings from 48000 randomly chosen anonymous users on nearly 18000 movie titles
- Date of each rating, movie title & year of release
- IMDB also has ratings, users need not be anonymous
- Linkage attacks:
  - For 89% of users, just two ratings and dates are enough to reduce the plausible set to 8 (out of the original 500,000)
  - 96% of users can be uniquely identified given just eight ratings and dates with up to 3-days of error

Debate: Data Privacy or Data Protection?

- Comprehensive data protection laws in Europe/Canada/Australia:
  - Data collection needs authorization by law or consent of individual
  - Individuals can view/update/delete their data
  - No transmission of data to locations without adequate protection
- No such all-encompassing law in the US:
  - Domain-specific laws such as HIIPA (Healthcare), COPPA (children's online privacy), FCRA (credit information), ECPA (electronic communications)
  - Organization that collects data essentially owns it
  - Laws favor information flow over individual rights to control use

Privacy legislation world map



Searching for a good privacy definition

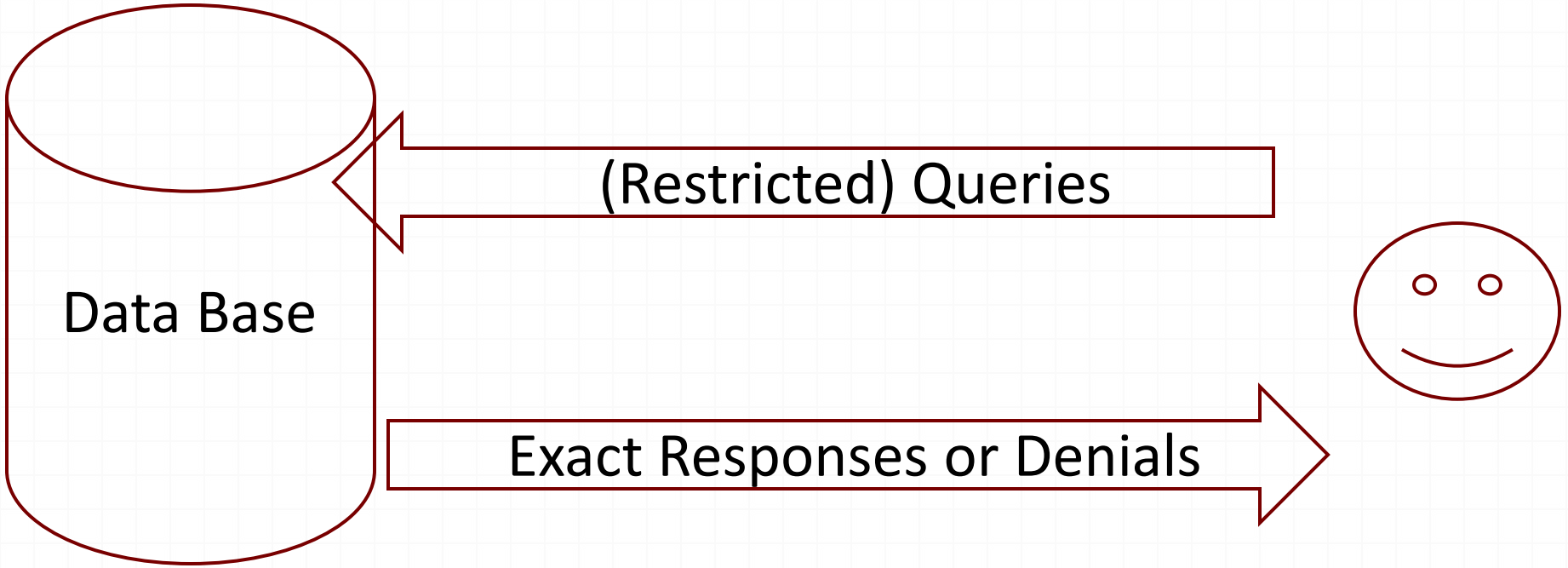
- Fellegi (1972): Right to determine what information about ourselves we will share with others
- Dalenius (1977): Anything learnt about a respondent from a database should also be learnable *without* access to it
- Goldwasser & Micali (1984): [Semantic Security] Must not be able to infer any more about message from *cypher-text & auxiliary information* than from just the auxiliary information
- Adam & Wortmann (1989): Disclosure occurs if through answers to queries a *snooper* is able to infer an exact or more accurate value of a confidential attribute
- Samarati & Sweeney (1998): Every combination of non-sensitive attributes must occur, if at all, in several records
- Dwork et al. (2006): Compare risk to an individual when included in, versus when not included in, the database
- …

This talk

- What this talk is about:
  - Introduction to formal definitions of privacy
  - Difficulties and challenges in defining privacy
  - Comparisons of different privacy definitions

- What this talk is not about:
  - Mechanisms for achieving privacy
  - Historical evolution of privacy research
  - Computational or algorithmic aspects of privacy
  - A comprehensive survey of privacy definitions (Disclaimer)

Query set restriction



Query-set-size control

- Respond *only if* response is bigger than *K* (for *L*-size database)
  - Applicable to numeric/non-numeric data
  - Applicable to multiple attributes
  - No perturbation, so responses are accurate
- Poor disclosure control (Denning et al., 1979):
  - Confidentiality breach even for K=L/2
  - Easy to construct 2 queries (both large), when put-together, reveals specific information about a record
- Query-set-overlap control: Limit overlap b/w queries
  - But attackers can collude
  - Expensive to check overlap for each new query

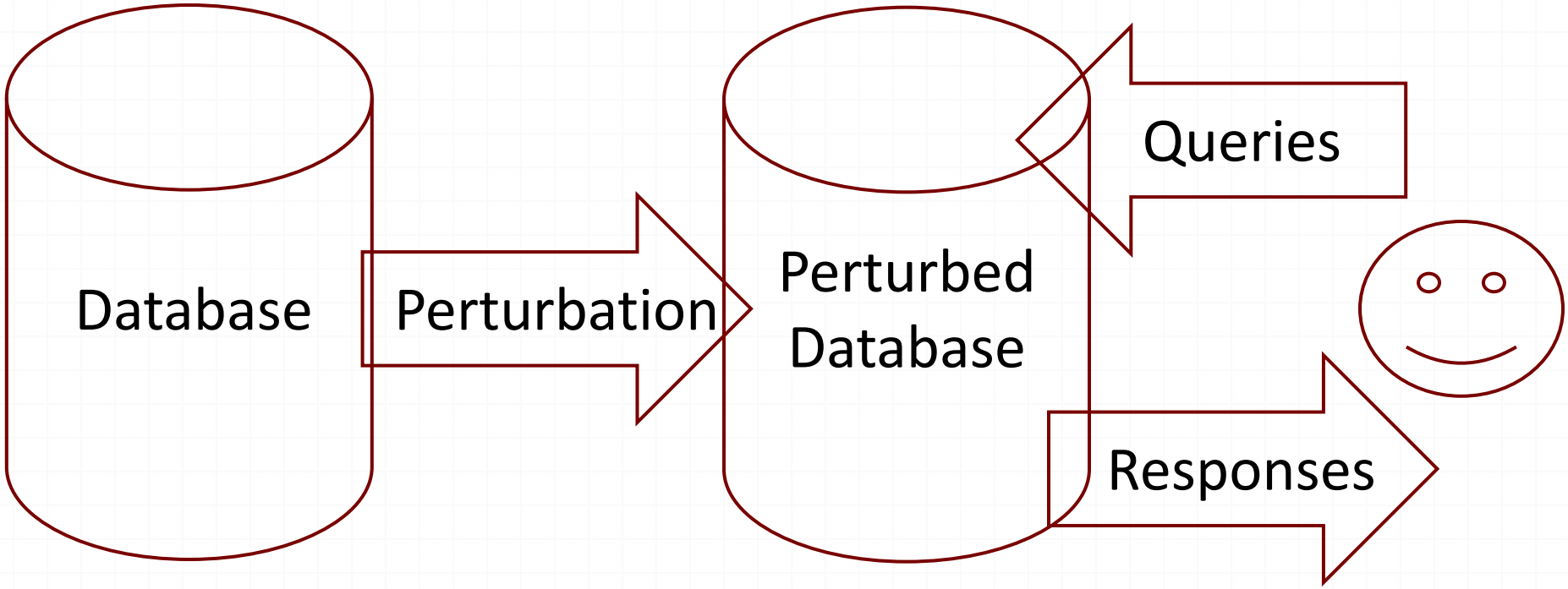
Query auditing

- Maintain logs of which queries issued by which users
- Check for possible compromise when new query arrives
- Advantages
  - Unperturbed response
  - Numerical/non-numerical attributes
- Disadvantages
  - Computation intensive
  - Storage intensive
  - Attackers can collude
- Example: Sum queries have been extensively studied
  - Chin et al. 1982, Denning 1983, McLeish 1983

Data partitioning & suppression

- Cluster records into mutually exclusive subsets called *atomic populations*
  - Drop attributes to make grouping possible
  - Can also add dummy entries to the database (noisy responses)
- Only statistical properties of atomic populations can be queried by users of the database
- Small-size atomic population can lead to partial disclosure
- Clustering-step may be computation intensive
- Examples: Yu & Chin (1977), Schlorer (1983)
- Cell suppression:
  - Cells with too few observations are suppressed
  - Examples: Cox (1980), Mugge (1983)

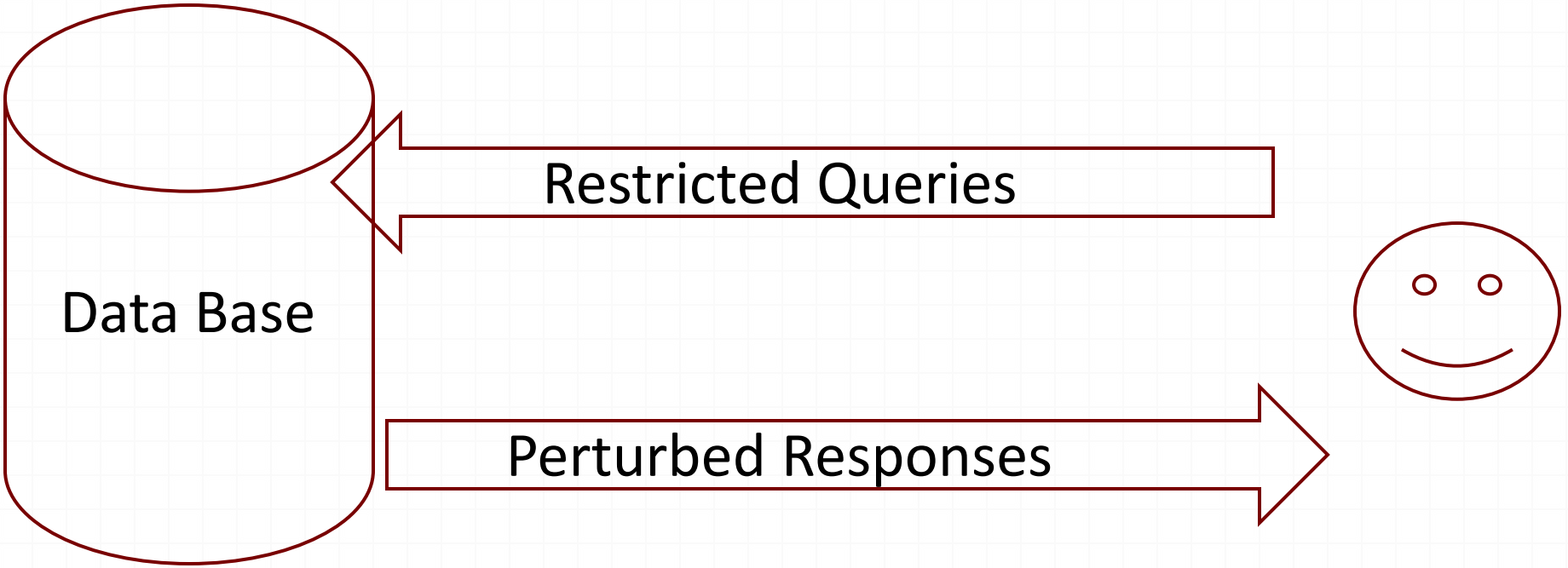
Data perturbation



Generating noisy databases

- Data swapping/shuffling
  - Reiss (1984), Fienberg et al. (2004), Sarathy et al. (2002)
  - Random swapping of data entries
  - Preserve some order statistics (such as row sums)
  - Can release contingency tables of swapped data
  - Sub-population data and/or microdata may look like non-sense
- Sampling: Liew et al. (1985), Matloff (1986), Rubin (1993), Reiter (2005), Matthews et al. (2010)
  - Estimate a distribution for the data
  - Generate synthetic data sample from learnt distribution
  - Release synthetic sample
- Explicit noise addition
  - Randomized response for categorical data (Warner, 1965)
  - Additive/multiplicative perturbation for numerical data (Taub et al., 1984)
  - Matrix masking for microdata: Y = AXB + C (Cox 1980, Fuller 1993)

Output perturbation



Noisy responses (interactive setting)

- Random-sample queries
  - Sub-sample query response before release (Denning 1980)
  - Insert some random spurious responses (Leiss 1982)
- Additive perturbation for numeric queries (Beck 1980)
- Rounding (Haq 1975, Achugbue and Chine 1979)
  - Query response is rounded-up or down to nearest multiple of some chosen base
- May require addition of less noise than data perturbation due to restricted class of queries
- May be possible to calibrate noise w.r.t number of queries that need to be answered

Types of disclosure risks

- Risk of re-identification
  - Individual can be accurately identified in the released data (or query responses)
- Risk of predictive disclosure
  - Value of some sensitive attribute can be estimated with reasonable accuracy
  - Can occur with or without links to external sources
  - Concerns information of both an individual and of a group (or population)
- Can we define privacy measures to assess such risks?

Confidentiality measures

- Spruill (1982), Paass (1988)
  - Proportion of entries that can be linked to the original data or to data from auxiliary source
  - Based on "distance" of each record in the released data to records in the original/auxiliary data
- Duncan & Lambert (1989), Reiter (2005)
  - Decision-theoretic approach
  - Prior distribution over sensitive attribute $p(s)$
  - Loss function for $L(t, s)$
  - Minimum Expected Loss is used as a measure of confidentiality (greater min loss implies greater confidentiality)

Confidentiality measures (contd.)

- Skinner and Elliot (2002)
  - $\Theta = \frac{\sum_j I[f_j=1]}{\sum_j F_j I[f_j=1]}$
  - f_j is the j^{th} identifying feature in the released sample
  - F_j its frequency in the original population
- Skinner and Shlomo (2008)
  - In-practice, f_j is observable, while F_j is not
  - Fit log-linear models to estimate F_j

k-anonymity (Sweeney 2002)

- Shows an attack on Massachusetts Group Insurance data (by linking to unique attributes from voter records)
- Sweeney proposed *k-anonymity* as a solution
- Dalenius (1986) introduced "quasi-identifiers"
  - Attributes that can be used to match with an external DB
- Definition: *A table is k-anonymous if every combination of quasi-identifiers appears at least k times*

Algorithms for k-anonymity

- Sweeney (2002), Hundepool et al. (2005)
- Based on two approaches:
  - Generalization: Group identifying attributes to broader categories (e.g., use states instead of city for location)
  - Suppression: Abstain from releasing a specific sensitive values
- Must not distort data too much
- Hardness result (Meyerson & Williams, 2003):
  - k-anonymity under minimum distortions is NP Hard

k-anonymity to ℓ-diversity

- Two attacks on Machanavajjhala et al. (2007) describe two simple attacks on k-anonymity
- Homogeneity attack: Sensitive attribute lacks diversity
- Background knowledge attack: Under some background information, uncertainty of a sensitive attribute may dramatically reduce

Example attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

ℓ-diversity (Machanavajjhala et al. 2007)

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Figure 3. 3-Diverse Inpatient Microdata

ℓ-diversity (contd.)

- Basic idea: Each equivalence class E must exhibit diversity with respect to sensitive attributes
- Three specific instantiations:
  - Distinct ℓ-diversity: At least ℓ distinct sensitive values in E
  - Entropy ℓ-diversity: Entropy of sensitive attribute in E is at least $\log(\ell)$
  - Recursive (c, ℓ)-diversity: E satisfies $r_1 < c(r_\ell, r_{\ell+1}, \cdots, r_m)$
    - r_i is frequency of i^{th} most frequent value in the class
- Algorithms based on adaptation of k-anonymity methods (& exploit a non-monotonicity property)

Criticism & attacks (Li et al. 2007)

- ℓ-diversity may be difficult to achieve, even unnecessary
  - Skewed distribution in sensitive attribute may imply different values have different levels of sensitivity
  - No need to protect majority value
- Skewness attack:
  - Consider a sensitive (binary) attribute with 99-01 bias
  - If data has an equivalence class with a 50-50 bias
    - Distinct and entropy 2-diversities, recursive (c,2)-diversity
    - But leads to obvious disclosure (more info
- Similarity attack:
  - Sensitive values within a class may be distinct but semantically related (e.g., different stomach-related diseases)

t-closeness (Li et al., 2007)

- Definition: An equivalence class E satisfies t-closeness if distance between the distribution of sensitive values in E and that in the whole data is within threshold t. When all equivalence classes satisfy this, data is said to satisfy t-closeness property
- Instantiation based on Earth Mover's Distance (EMD)
  - Minimum total distance points must move to transform one distribution into the other
- Drawbacks:
  - Complications arise under multiple sensitive attributes
  - What if adversary knows more than overall distribution of sensitive attribute?
  - What about multiple releases after insertions/deletions in data?
    - Led Xiao et al. (2007) to m-invariance

What is going wrong here?

- Definitions are syntactic conditions on released data
  - No formalization of any *semantic* notion of privacy
- Semantic Security (Goldwasser & Micali 1984)
  - Defined in a secure communication setting
  - Must not be able to learn anything more about a message given its cypher-text & auxiliary information, compared to what can be learnt from just auxiliary information
  - Fundamental concept in modern-day cryptography
- Can we also formalize privacy along these lines?

Difficulties in formalizing a semantic notion

- Unlike secure communication, we need to release (publicly) useful information about the data
- The bugbear of auxiliary information:
  - Consider a database with information about heights of individuals in a population
  - Goal: Make database available for “average height” queries
  - Auxiliary information: Terry Gross is two inches shorter than the average Lithuanian woman
  - Release of average height of Lithuanian woman breaches privacy of Terry Gross (who may not even be in the data)!
- Rigorous impossibility result exists

Differential Privacy (Dwork et al. 2006)

- Definition: A randomized function $\mathcal{A}$ is ϵ-differentially private if for all data sets D_1 and D_2 that differ in at most one element, and for all $S \subseteq Range(\mathcal{A})$, we have

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D_2) \in S]$$

- Key features:
  - Interactive setting
  - Agnostic to auxiliary information
  - Can handle multiple queries (composition)
  - Any DP mechanism must add noise to response (output perturbation)

Example

- Let $f(D)$ denote the mean of a database D
- D' is a database differing from D in just one record
- $\Delta f = \max_{D,D'} \|f(D)\|_1$ (L_1-sensitivity)
- $\hat{f}(D) = f(D) + Laplace(0, \sigma)$
- $\hat{f}(D)$ satisfies $(\frac{\Delta f}{\sigma})$-differential privacy

DP versions of many useful algorithms

- PCA, k-means, decision trees (Blum et al. 2005)
- Contingency tables (Barak et al. 2007)
- PAC learning algorithms (Kasiviswanathan et al. 2007)
- Learning half-spaces (Blum et al. 2008)
- Recommender systems (McSherry and Mironov 2009)
- Frequent pattern mining (Bhaskar et al. 2010)
- …

Drawbacks

- Side-steps issue of how much more adversary learns after receiving query responses
- Noise addition
  - Found to be unacceptable high in many settings
  - Increases substantially for multiple queries
  - Cannot release entire database (non-interactive setting)
  - Noise calibrated assuming adversary knows all-but-one record
- Auxiliary information
  - No way around the Terry Gross example!
  - Agnostic to what adversary may know
  - No meaningful guarantee under arbitrary auxiliary information

Evolution of DP-style privacy definitions

- (ϵ, δ)-DP (Nissim et al. 2007):
  - $\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D_2) \in S] + \delta$
- (ϵ, δ)-probabilistic DP (Machanavajjhala et al. 2008):
  - Probability over possible outputs that fail the ϵ-DP criterion is at most δ
- Abstract DP and Generic DP (Kifer and Lin 2010):
  - Axiomatization of privacy and utility guarantees
  - Leads to acceptable relaxations of DP

ϵ-Generalized DP (Bhaskar et al. 2011)

- Data often exhibits statistical properties
  - Exploited successfully in Machine Learning for years
  - Can this represent a source of uncertainty for privacy?
- Definition: A mechanism $\mathcal{A}: \mathcal{D} \rightarrow \mathcal{Y}$ satisfies ϵ-Generalized DP under a distribution D over $\mathcal{D}$ and auxiliary information Aux, if for all $S \subseteq \mathcal{Y}$ and $a, a' \in \mathcal{D}$,

$$\Pr_{\mathrm{T}\sim\mathrm{D},\mathrm{c_f}}[\mathcal{A}(T) \in S | t_\ell = a] \leq \Pr_{\mathrm{T}\sim\mathrm{D},\mathrm{c_f}}[\mathcal{A}(T) \in S | t_\ell = a']$$

  where t_ℓ is ℓ^{th} entry of T & c_f denotes random coins of f
- Reduces to DP when Aux is all-but-one entry and D is a point distribution
- Noiseless Privacy possible in many non-trivial settings

Are we there yet?

- Clearly not. Why?
  - Many privacy definitions (including DP) have been around for a while, but none adopted in-practice
  - Debate concerning what auxiliary information to support
  - Debate concerning addition of noise
  - Data-release mechanisms may be regulated, but cannot legislate directly against privacy breaches
- Privacy monitoring systems?
  - May be an important line to pursue
  - No practical privacy definition may suffice
  - Must be able to detect privacy breaches at the earliest (like in computer/network security)

Thank you!!