Privacy-Utility Tradeoffs of Data Sources

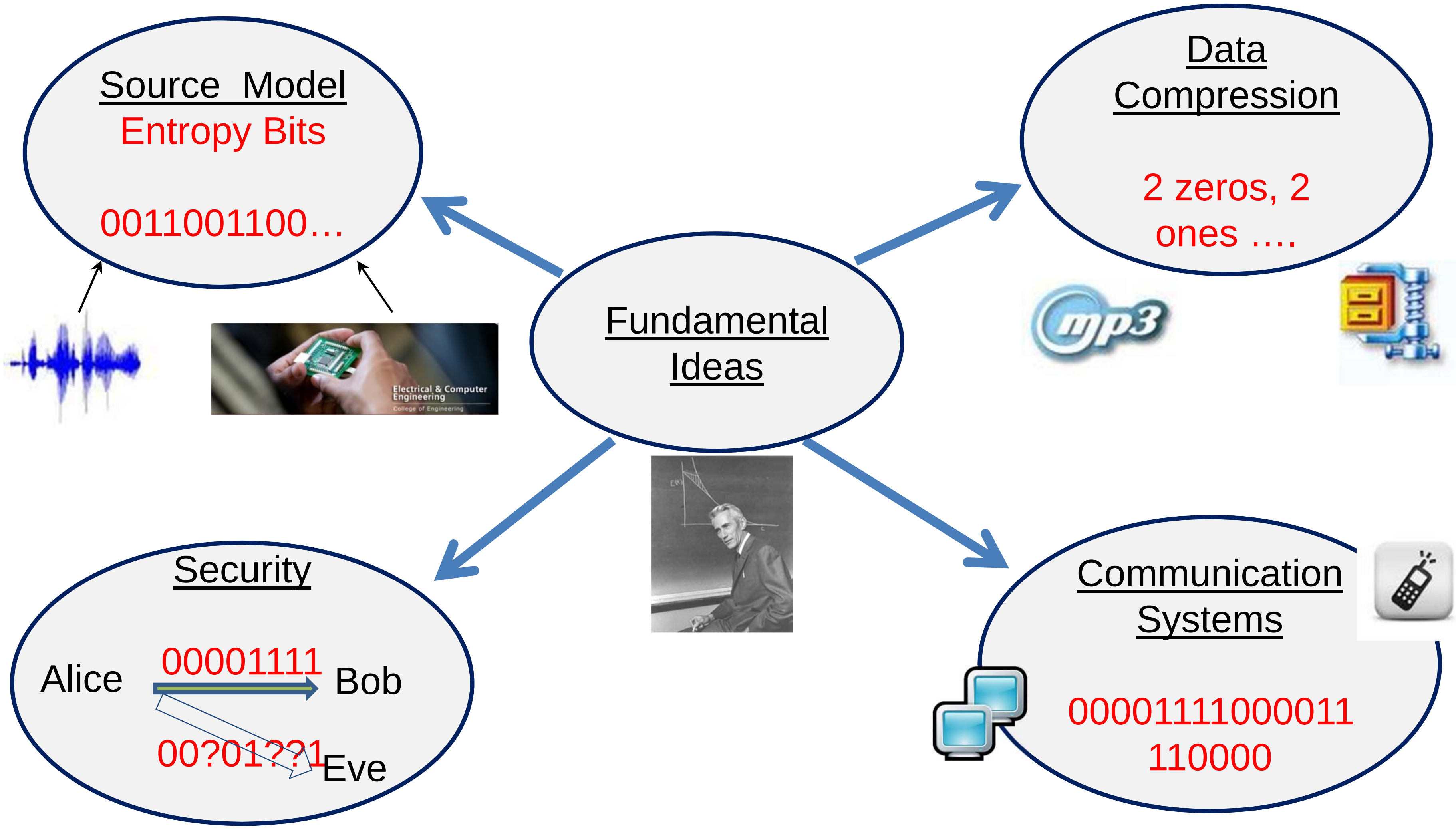
H. Vincent Poor

Princeton University

Joint work with Lalitha Sankar, et al.

Supported by NSF Grant CCF-10-16671

The Information Revolution



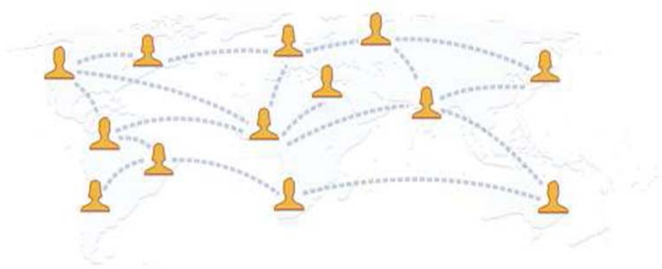
Electronic Data Repositories

- Technological leaps in information processing, storage, and communications has led to the creation of vast electronic data repositories.

By simply clicking on a **blue button** icon, users will be able to download their medical (Medicare/Medicaid) data to their personal computers. – (PubMed Central)

The Privacy Problem

- Explosive growth in electronic information sources that are publicly accessible
  - Google, Facebook, open governance, DMV records, etc.



- These electronic information sources can also leak private information!

Utility vs. Privacy

- Utility (benefit) of data repositories is in allowing legitimate users access to statistical/processed data.
  - e.g., census data
- However, individual information needs to be kept private
  - Private information (e.g., SSN, DoB, credit card) can be potentially inferred from revealed data.
- Private information is application-specific
  - DoB is private for medical but not DMV databases.
  - Census publications may not reveal name, SSN, DoB, address, tel. no. of any individual.
- Need a framework that precisely quantifies the utility-privacy tradeoffs for any application.

Talk Outline

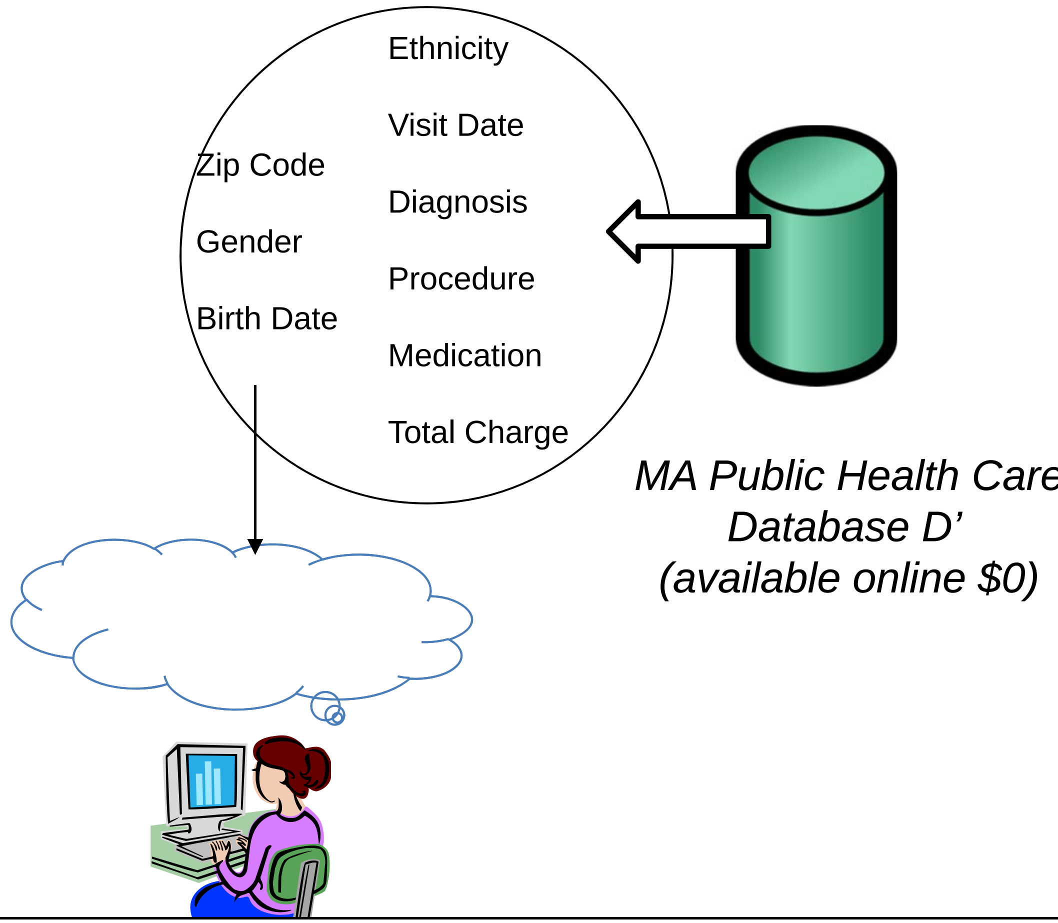
- Database privacy problem
- Smart grid privacy problems
- Summary and future work

Talk Outline

- Database Privacy Problem
  - Source and Perturbation Model
  - Utility and Privacy Metrics
  - Examples
  - Related Results

The Massachusetts Example

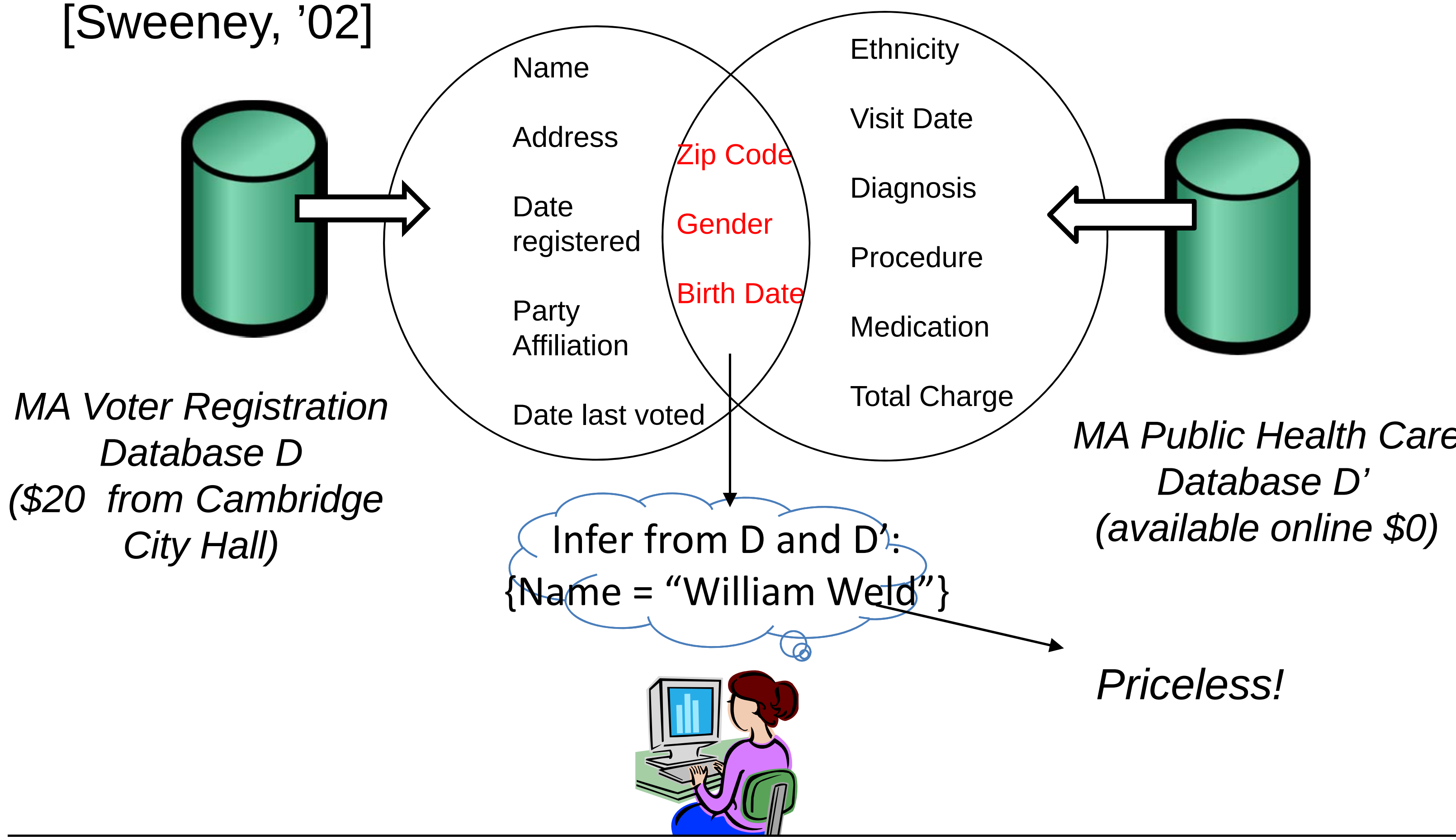
- Is it sufficient to hide personal information? [Sweeney, ’02]



L. Sweeney, “*k*-anonymity: A model for protecting privacy,” *Intl. J. Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

The Massachusetts Example

- Unique identification via correlation from two public databases [Sweeney, '02]



L. Sweeney, "*k*-anonymity: A model for protecting privacy," *Intl. J. Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

More Examples

- The Netflix competition [2006] to improve movie recommendations
  - Public training data set with movie preferences of 480,000 customers
  - Data was "de-identified" – stripped of specific personal details
- V. Shmatikov and A. Narayanan [ISSP, '08]
  - Compared film preferences of some anonymous customers with personal profiles on imdb.com,
  - *Re-identification* using distinguishing information
- Netflix claimed
  - *"Anonymity of the study data is comparable to the strictest Federal standards for anonymizing personal health information."*

A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Intl. Symp. Security and Privacy*, Oakland, CA, May 2008, pp. 111–125.

More Examples… Medical Data

- *New York Times* reports
  - Sale of clinical data is a huge and growing business.
  - *De-identified* information is “repackaged” and resold.
  - *New* regulations do NOT forbid sale of de-identified data.
- The opportunities for leakage are growing
  - Query logs, genetics, …

- De-identification is NOT sufficient for safe disclosure of medical data!

The Privacy Problem is Pervasive

- Sources leak information in unforeseeable ways
  - Intra-source leaks: hidden correlations between public and personal information, e.g., electronic health systems, census (e.g. outliers)
  - inter-source leaks: correlation between sources [Sweeney, Shmatikov]

- But the electronic sources cannot be shut down
  - Tremendous utility provided.
  - Cannot shut down Google or Facebook!

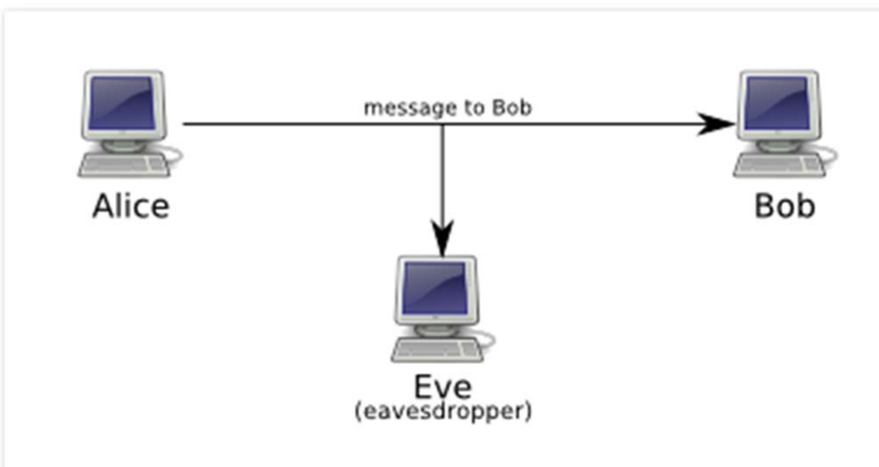
- Can we disclose (utility) while guaranteeing privacy?

Privacy vs. Secrecy!

- Privacy: the ability to prevent unwanted transfer of information (via inference or correlation) when legitimate transfers happen.

- But privacy is not secrecy!

- Secrecy Problem: Protocols and primitives clearly distinguish a malicious adversary vs intended user and secret vs non-secret data.
  - Encryption may be a solution.

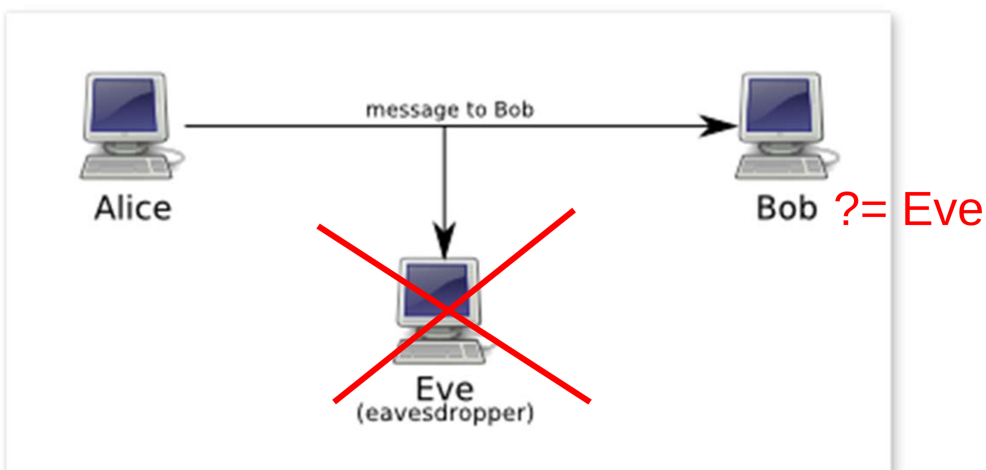


Privacy is not Secrecy!

- Privacy: the ability to prevent unwanted transfer of information (via inference or correlation) when legitimate transfers happen.

- But privacy is not secrecy!

- Privacy problem: disclosing data provides informational utility while also enabling potential loss of privacy
  - Every user is potentially an adversary
  - Encryption is not a solution!



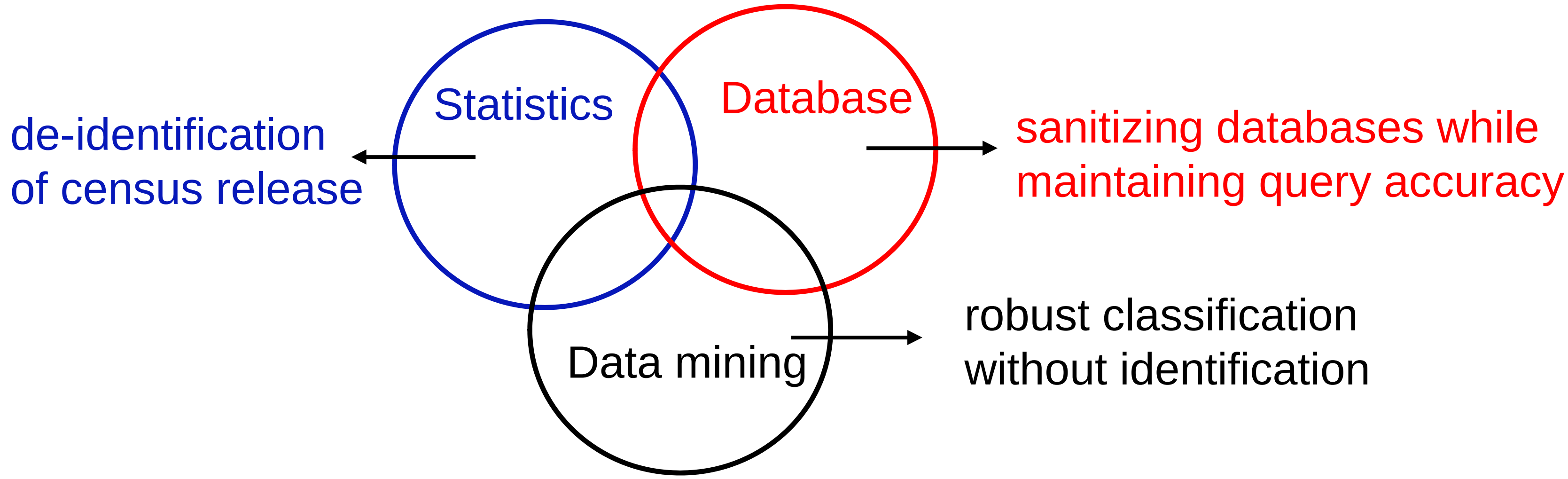
What is Utility?

- Data sources exist to be used but utility of a data source can be degraded by privacy requirements.
- "*Perfect privacy can be achieved by publishing nothing at all, but this has no utility; perfect utility can be obtained by publishing the data exactly as received, but this offers no privacy*" [Dwork '06]
- Thus, maximum utility of a data source is achieved at minimum privacy and vice versa.
- What is the utility-privacy tradeoff for a data source?

C. Dwork, "Differential privacy," in *Proc. 33rd Intl. Colloq. Automata, Language, and Programming*, Venice, Italy, July 2006.

Existing Approaches

- Privacy problem lies at the intersection of multiple communities.

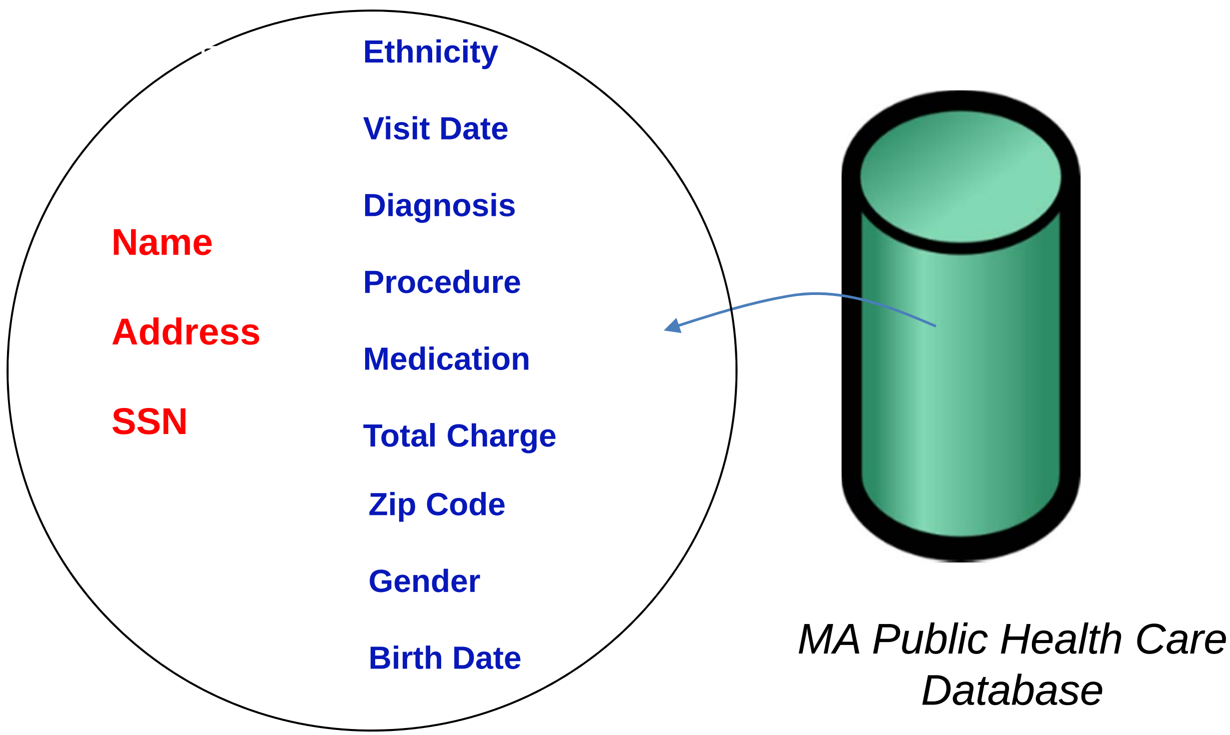


  – Application-specific approaches without universal guarantees
- CS Theory: *differential privacy* – cryptography motivated definition
  – How to guarantee non-identification
  – Privacy paramount

- Utility vs. privacy tradeoff remains unsolved.

Privacy Problem: A New Insight

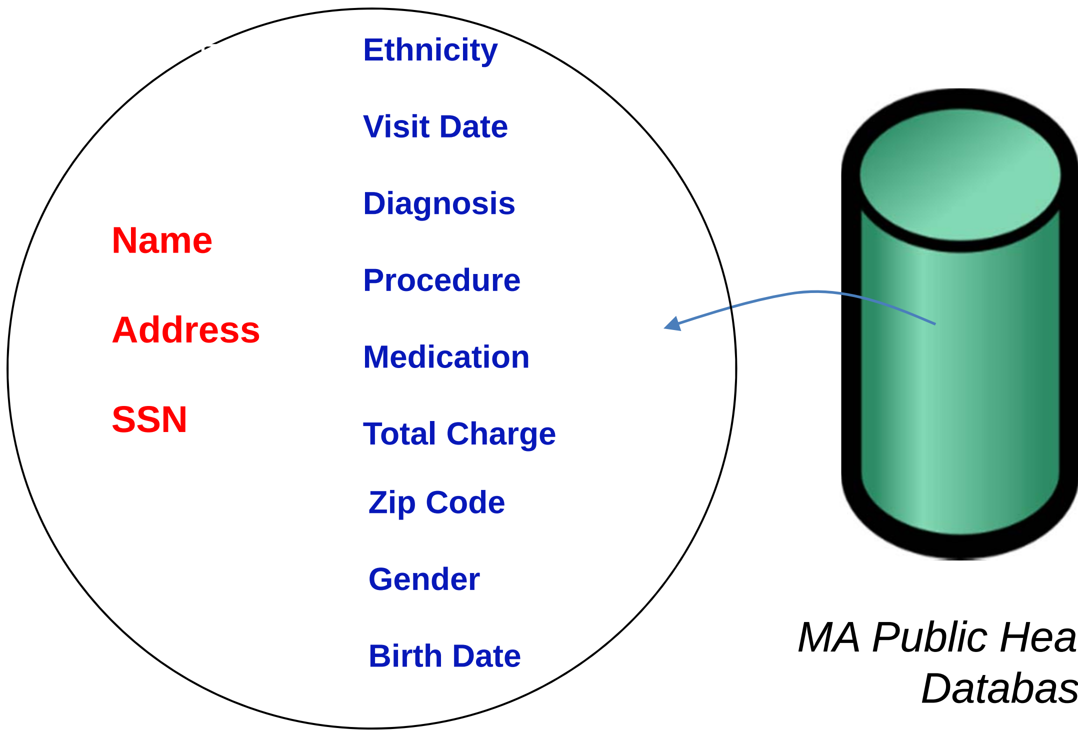
- Any data source has public and private attributes



L. Sankar, S. R. Rajagopalan, and H. V. Poor. "Utility and privacy of data sources: Can Shannon help conceal and reveal information?," *ITA Workshop*, La Jolla, CA, Feb. 2010.

Privacy Problem: A New Insight

- Any data source has public and private attributes
- Want to reveal public attributes maximally without revealing the private attributes



MA Public Health Care Database

L. Sankar, S. R. Rajagopalan, and H. V. Poor. "Utility and privacy of data sources: Can Shannon help conceal and reveal information?," *ITA Workshop*, La Jolla, CA, Feb. 2010.

Privacy Problem: A New Insight

- But… private and public attributes are correlated.
- Controlling privacy leakage amounts to controlling the correlation.
- Correlation can be controlled via perturbation of public attributes.
- Best U-P tradeoff: finding the minimal perturbation that achieves a desired correlation.
- Our contribution: a framework based on rate-distortion theory with universal metrics for utility and privacy.

L. Sankar, S. R. Rajagopalan, and H. V. Poor. “Utility and privacy of data sources: Can Shannon help conceal and reveal information?,” *ITA Workshop*, La Jolla, CA, Feb. 2010.

The Database Privacy Problem

- A database is a table – rows: individual entries (total of n); columns: attributes for each individual (total of K)

Attributes $\longrightarrow$ K total attributes

Entries $\downarrow$

Entries	Gender	Visit Date	Diagnosis	…	Medication
1					
2					
.					
.					
.					
n					

$\longleftarrow$ Numeric and non-numeric data $\longrightarrow$

Query

User

Response

Database: Source Model

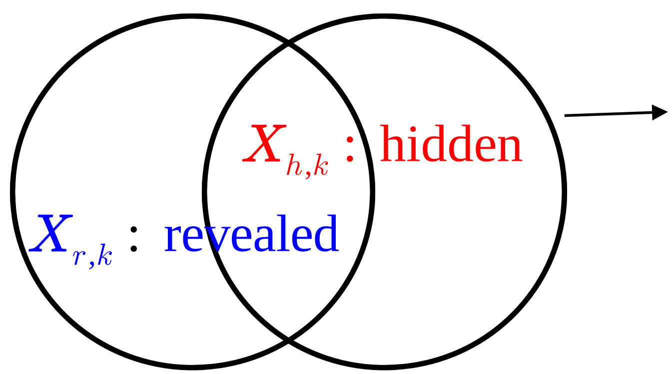
- A *real* database d is (typically) a table with $\underline{n>>1}$ rows (entries) and K columns (attributes)

Our model:

- Database d with n rows is a sequence of n i.i.d. observations of a vector random variable $X = (X_1\ X_2\ \ldots\ X_K)$ with the distribution

$$\boxed{p_X(\boldsymbol{x}) = p_{X_1 X_2 \ldots X_K}(x_1, x_2, \ldots, x_K)}$$

- Attributes divided into K_r public (revealed) and K_h private (hidden) variables, typically not disjoint



$$k^{th} \text{ entry} : X_k = \left(X_{r,k}, X_{h,k} \right)$$

L. Sankar, S. R. Rajagopalan, and H. V. Poor. "A theory of utility and privacy of data sources," *Proc. of IEEE Intl. Symp. Inform. Theory*, Austin, TX, Jun. 13-18 2010.

Database: Utility vs. Privacy

- The Utility-Privacy Problem:
  - How to reveal the public variables while hiding the private variables given that the two sets are correlated?

Database: Utility vs. Privacy

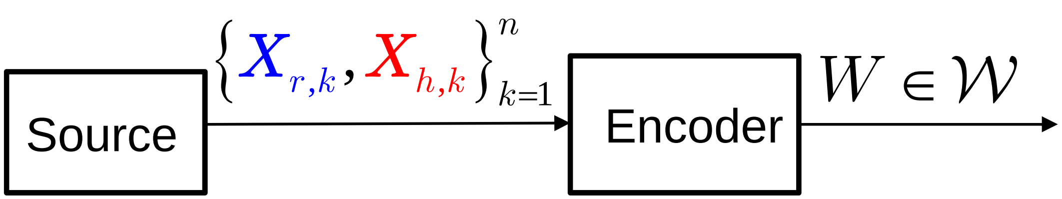
- The Utility-Privacy Problem: Rate distortion theory with privacy is a natural fit!

Database: Utility vs. Privacy

- The Utility-Privacy Problem: Rate distortion theory with privacy is a natural fit!
- Encoder maps d (X^n) to a "sanitized" database (SDB) d'

$$\boxed{\text{Encoder}: X^n \rightarrow \mathcal{W} = \{SDB_1, SDB_2, \ldots, SDB_M\}}$$

  - M: number of revealed ("quantized") databases



L. Sankar, S. R. Rajagopalan, and H. V. Poor. "A theory of utility and privacy of data sources," *Proc. of IEEE Intl. Symp. Inform. Theory*, Austin, TX, Jun. 13-18 2010.

Database: Utility vs. Privacy

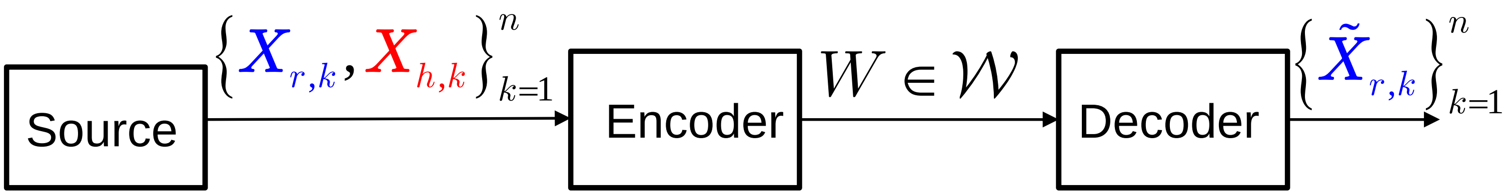
- The Utility-Privacy Problem: Rate distortion theory with privacy is a natural fit!
- Encoder maps d (X^n) to a "sanitized" database (SDB) d'

$$\text{Encoder}: X^n \to \mathcal{W} = \{SDB_1, SDB_2, \ldots, SDB_M\}$$

  - M: number of revealed ("quantized") databases

- Decoder: Uses d' to obtain a "reconstructed" database (for query processing)

$$\text{Decoder}: \mathcal{W} \to \tilde{X}_h^n$$



L. Sankar, S. R. Rajagopalan, and H. V. Poor. "A theory of utility and privacy of data sources," *Proc. of IEEE Intl. Symp. Inform. Theory*, Austin, TX, Jun. 13-18 2010.

Utility Metric

- Map utility to fidelity
  - Utility is a measure of closeness of d and d'.
  - Fidelity is affected by added noise, limited precision, suppression.

Encoding Constraint:

- Utility constraints $\Delta_d \rightarrow$ avg. distortion per entry (row)

$$\boxed{\Delta_d \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(X_{r,i},\tilde{X}_{r,i}\right)\right] \leq D+\varepsilon}$$

  - ρ: distance-based function (e.g.: Hamming, Euclidean, K-L)
  - D: upper bound on the avg. distortion per entry

- More generally, can bound distortion on all subsets of X_r

Privacy Metric

- Map privacy to equivocation
  - Privacy is a measure of 'uncertainty' about hidden data given revealed data.

Encoding Constraint:

- Privacy constraints $\Delta_p \rightarrow$ equivocation on average per entry (row)

$$\boxed{\Delta_p \equiv \frac{1}{n} H\left(\boldsymbol{X}_h^n \,|\, W\right) > E - \varepsilon}$$

  - E: lower bound on the avg. privacy per entry

- More generally, can bound equivocation on all subsets of X_h

The Utility-Privacy Tradeoff

- Utility-privacy tradeoff region ($\mathcal{T}$) is

$$\boxed{\mathcal{T} \equiv \{(D,E) : (D,E) \text{ is feasible}\}}$$

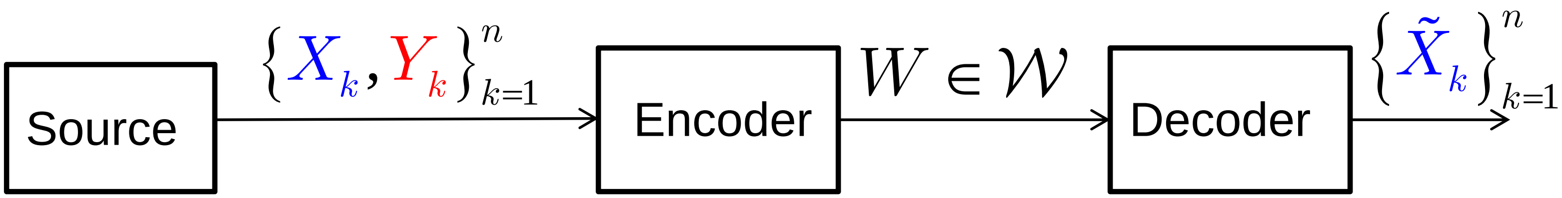
- How do we compute $\mathcal{T}$?

- Consider the following source coding problem with privacy constraints .….

A Source Coding Problem with Privacy

- A source (X,Y) wishes to reveal X subject to a fidelity constraint while keeping Y as private as possible.
  - Revealing X will result in information leakage about Y

- Problem first studied by Yamamoto [IT, ’83]



H. Yamamoto, “A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers,” *IEEE Trans. Inform. Theory*, 29(6), Nov. 1983.

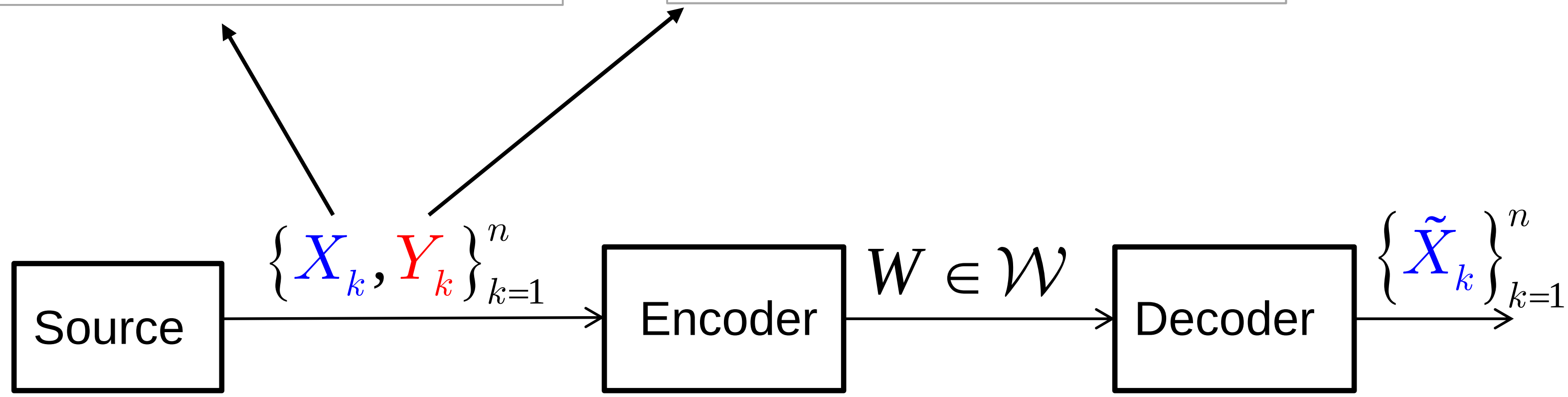
A Source Coding Problem with Privacy

Distortion

$$\Delta_d \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(X_i,\tilde{X}_i\right)\right] \le D+\varepsilon$$

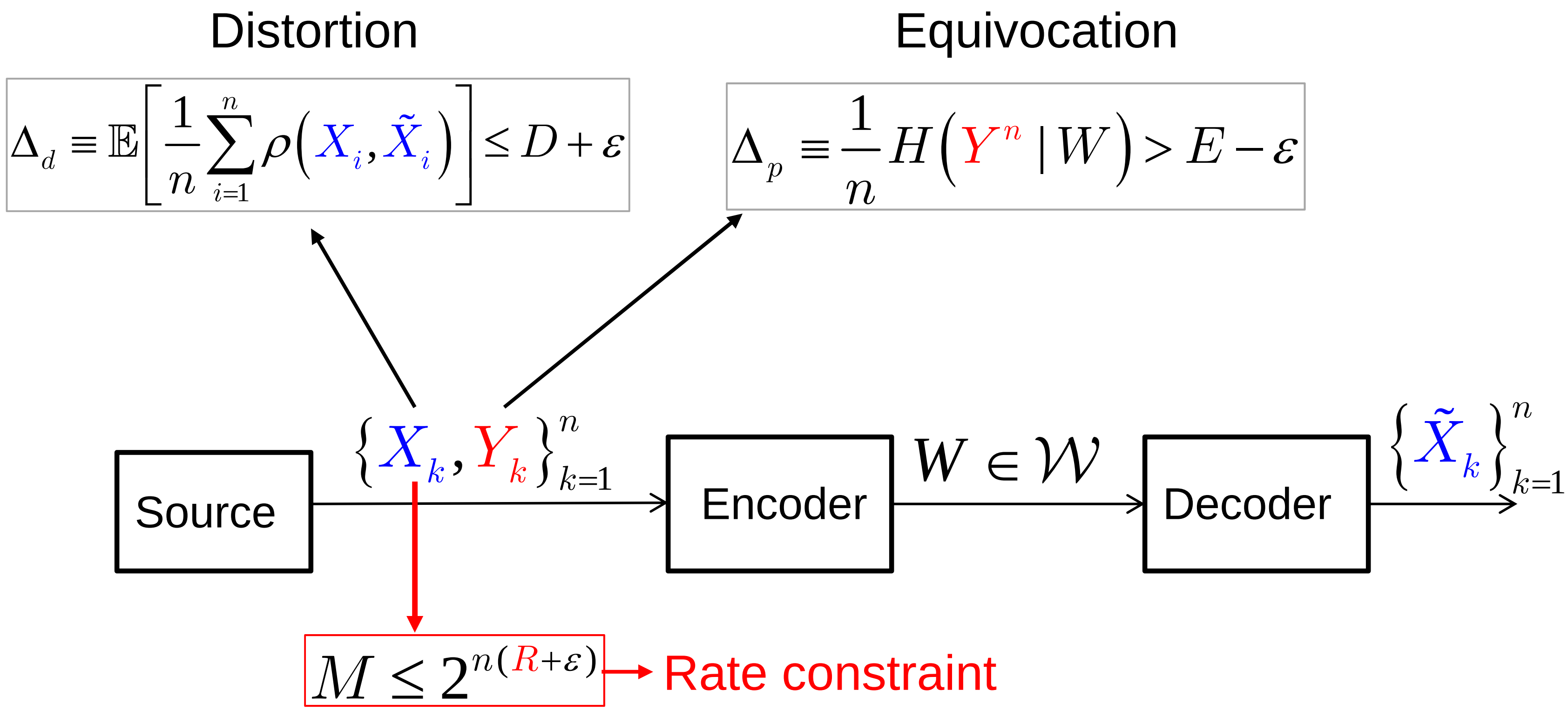
Equivocation

$$\Delta_p \equiv \frac{1}{n}H\left(Y^n \mid W\right) > E-\varepsilon$$



- Simplified version of the database privacy problem with…..
  - one private and one public attribute

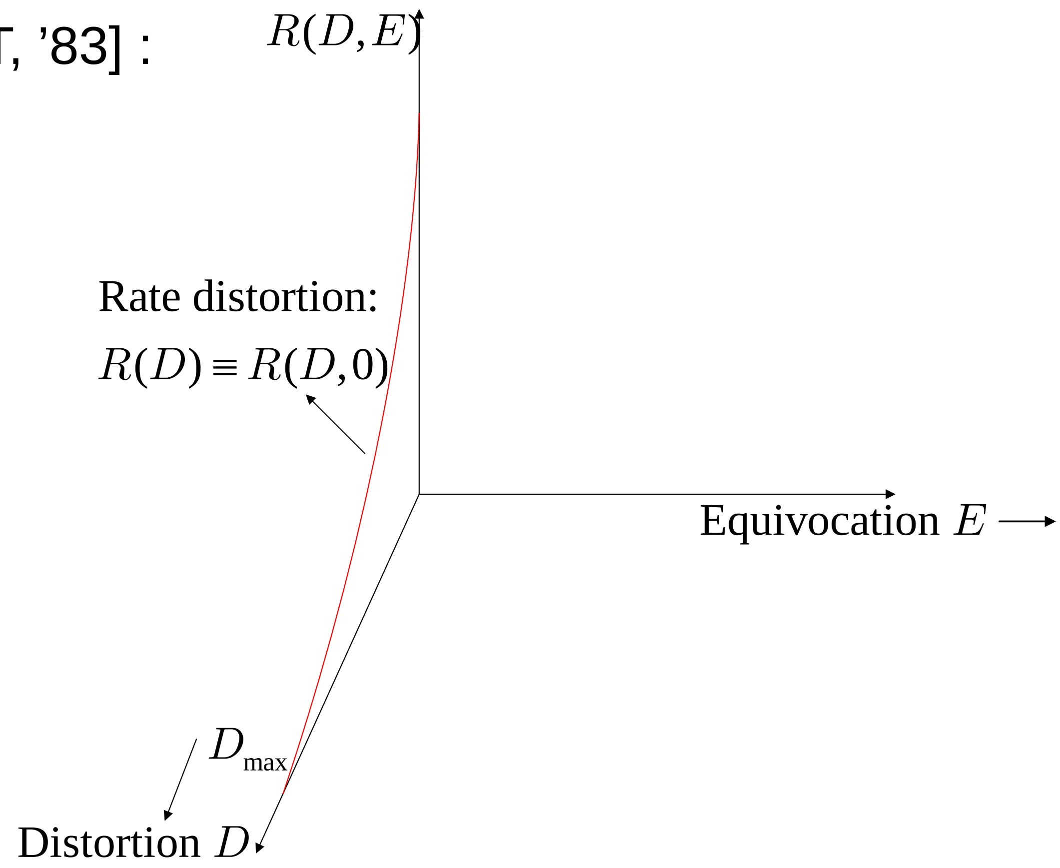
A Source Coding Problem with Privacy



- Simplified version of the database privacy problem with additional rate constraint
  - Rate constraint bounds the number of “quantized” sequences
  - For U-P tradeoff this seems superfluous

Rate-Distortion-Equivocation (RDE)

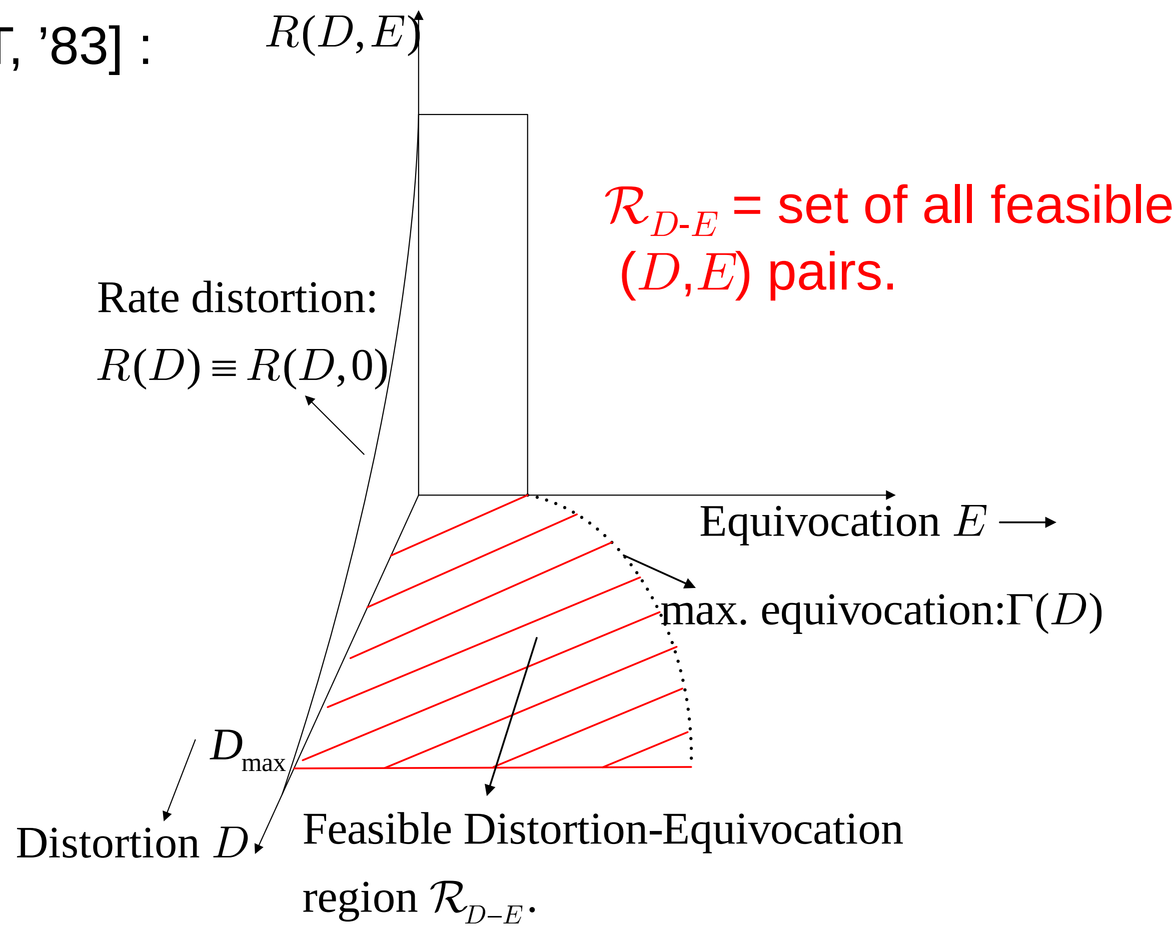
- Yamamoto [IT, '83] :



- $R(D)$ is the minimal compression rate for a distortion D

Rate-Distortion-Equivocation (RDE)

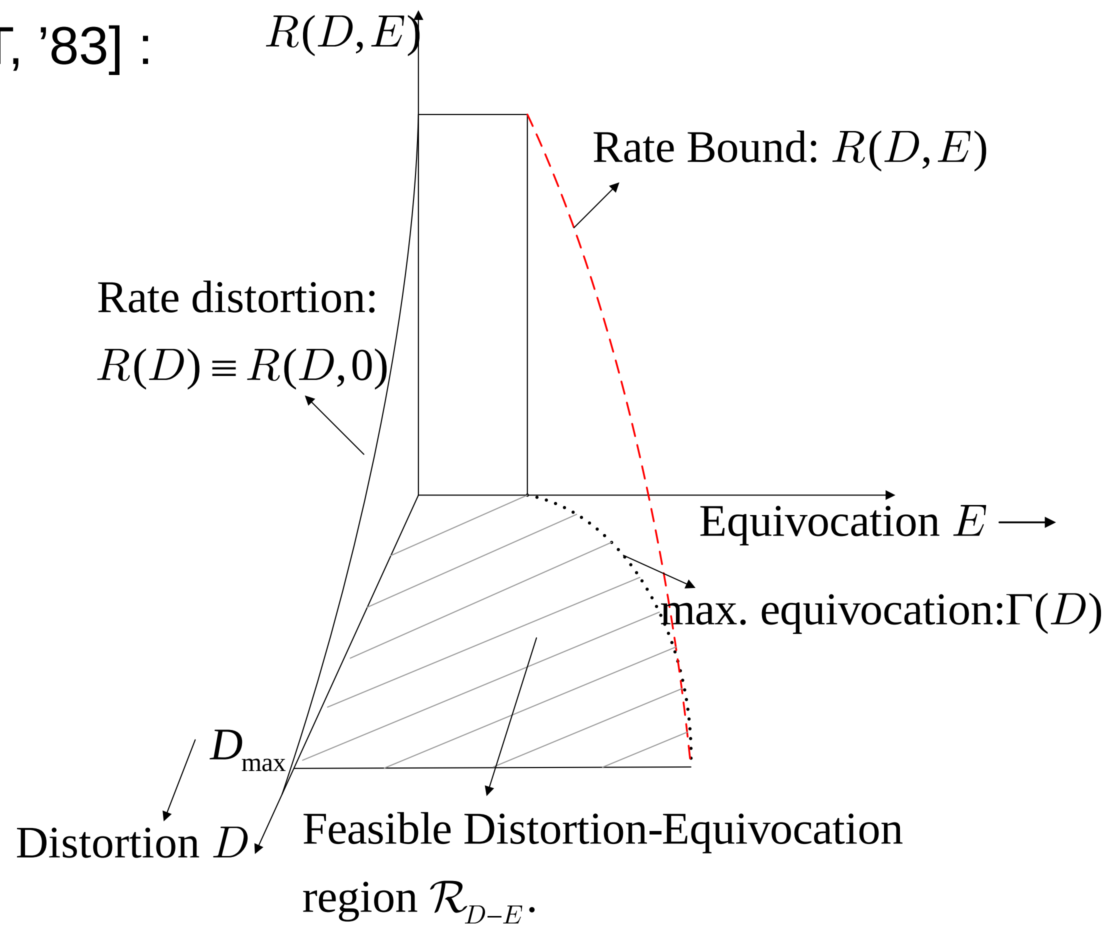
- Yamamoto [IT, '83] :



Rate-Distortion-Equivocation Region

Rate-Distortion-Equivocation (RDE)

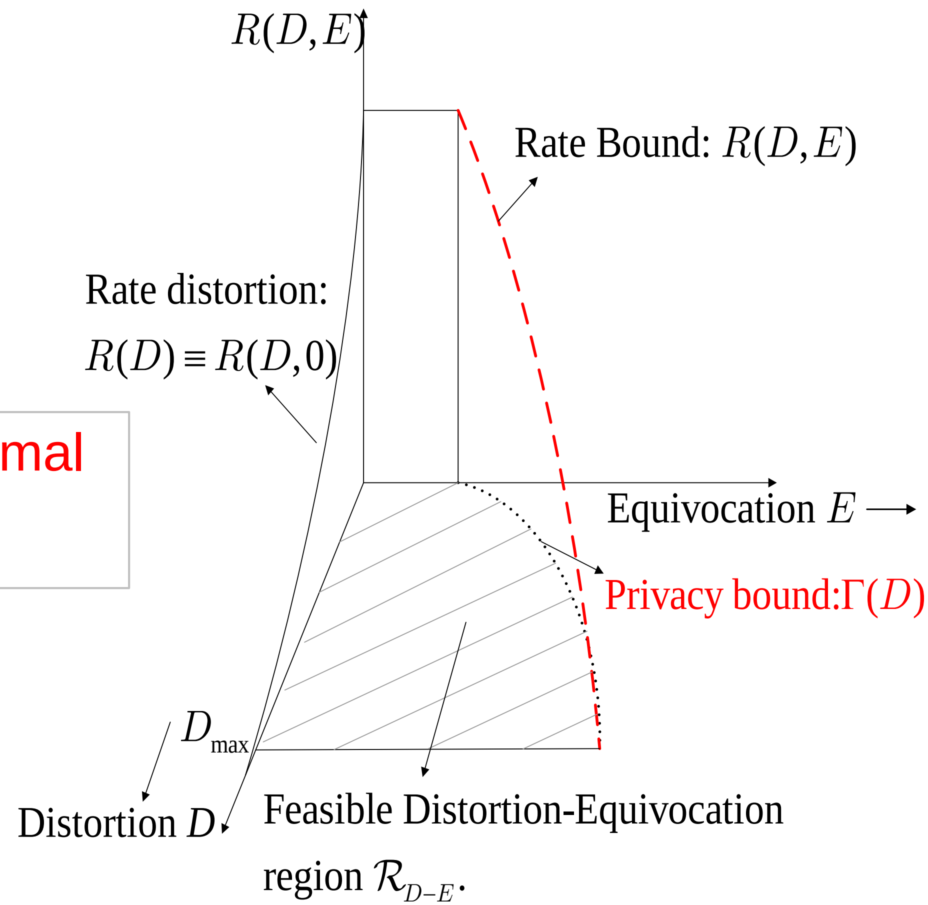
- Yamamoto [IT, '83] :



Rate-Distortion-Equivocation Region

Rate-Distortion-Equivocation (RDE)

- SRP [ISIT, '10] :



Distortion D determines the maximal achievable privacy $\Gamma(D)$

L. Sankar, S. R. Rajagopalan, and H. V. Poor. "A theory of utility and privacy of data sources," *Proc. of IEEE Intl. Symp. Inform. Theory*, Austin, TX, Jun. 13-18 2010.

The Utility-Privacy Tradeoff

- Recall: utility-privacy tradeoff region $\mathcal{T}$ is

$$\boxed{\mathcal{T} \equiv \{(D,E) : (D,E) \text{ is feasible}\}}$$

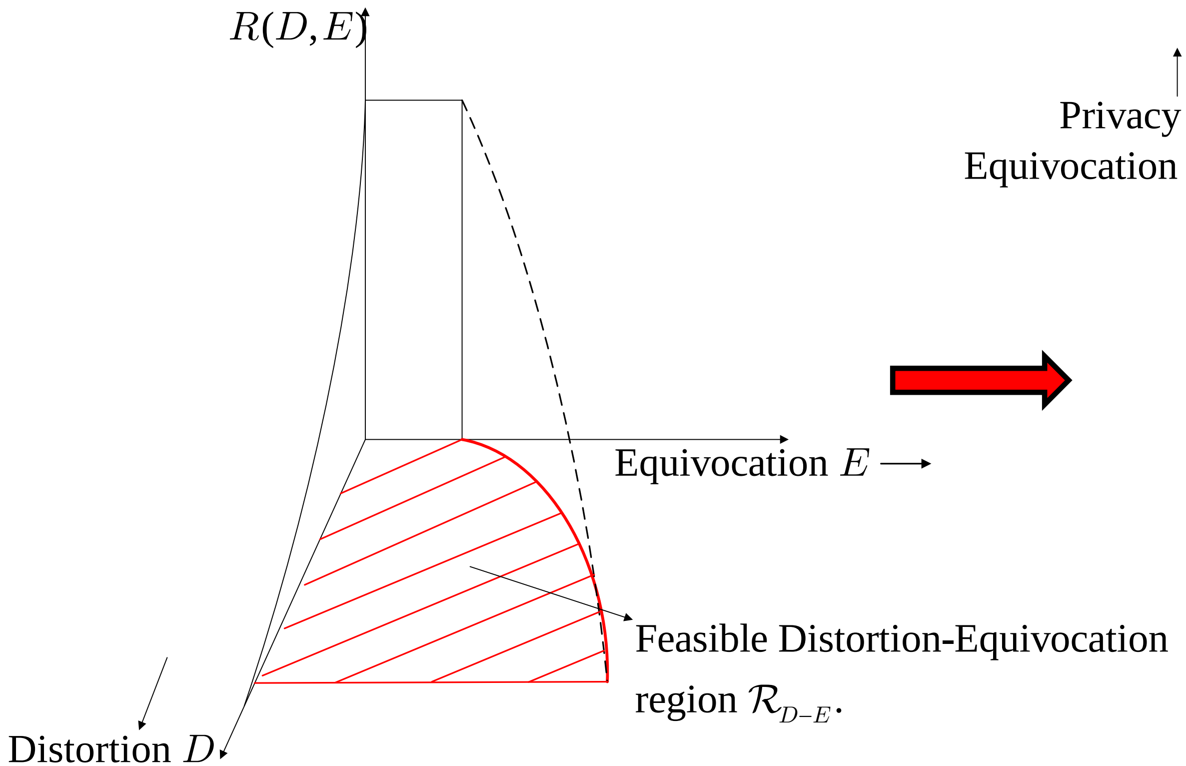
- Recall: $\mathcal{R}_{D\text{-}E}$: feasible distortion-equivocation pairs

- Theorem [SRP, ISIT '10] :

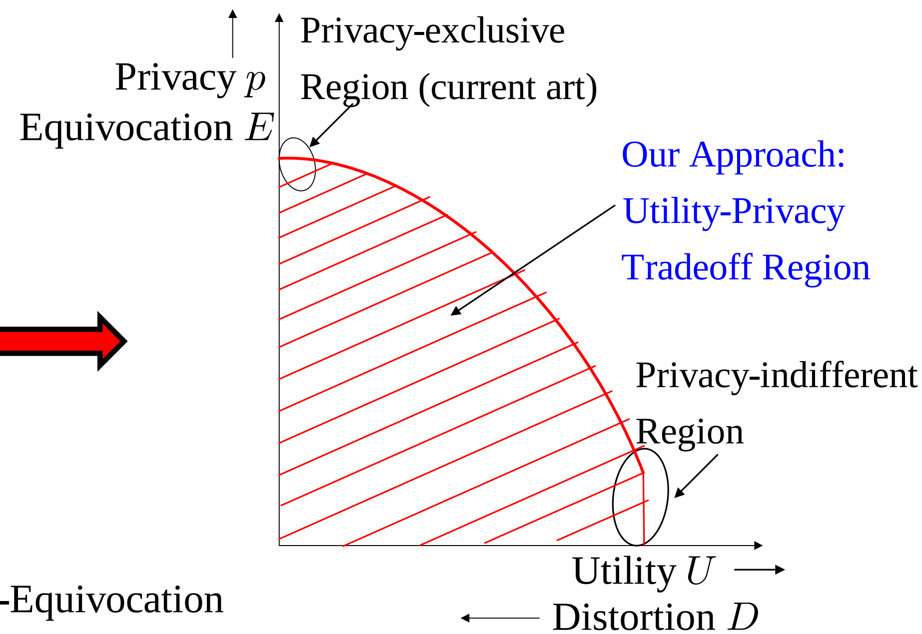
For a database with utility and privacy constraints, $\mathcal{T} = \mathcal{R}_{D-E}$.

L. Sankar, S. Raj Rajagopalan, H. V. Poor, "A theory of privacy and utility in databases," submitted to the *IEEE Trans. Inform. Theory*, Feb. 2011.

Utility-Privacy/RDE Regions



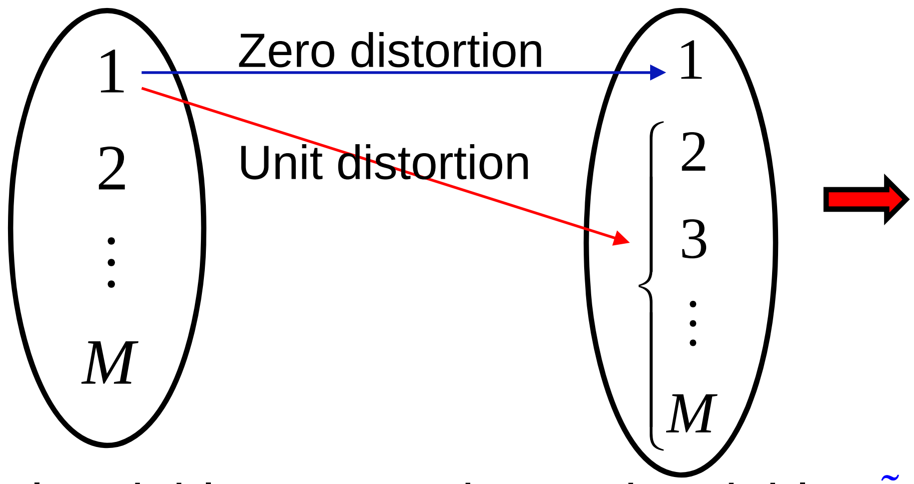
(a): Rate-Distortion-Equivocation Region



(b): Utility-Privacy Tradeoff Region

Example 1: Categorical Database

- Categorical data: finite alphabet data with discrete distribution
  - e.g.: SSN, zipcode, etc.



Original variable X Distorted variable $\tilde{X}$

Hamming Distortion D :

$$D = Pr(X = \tilde{X})$$

Equivocation E :

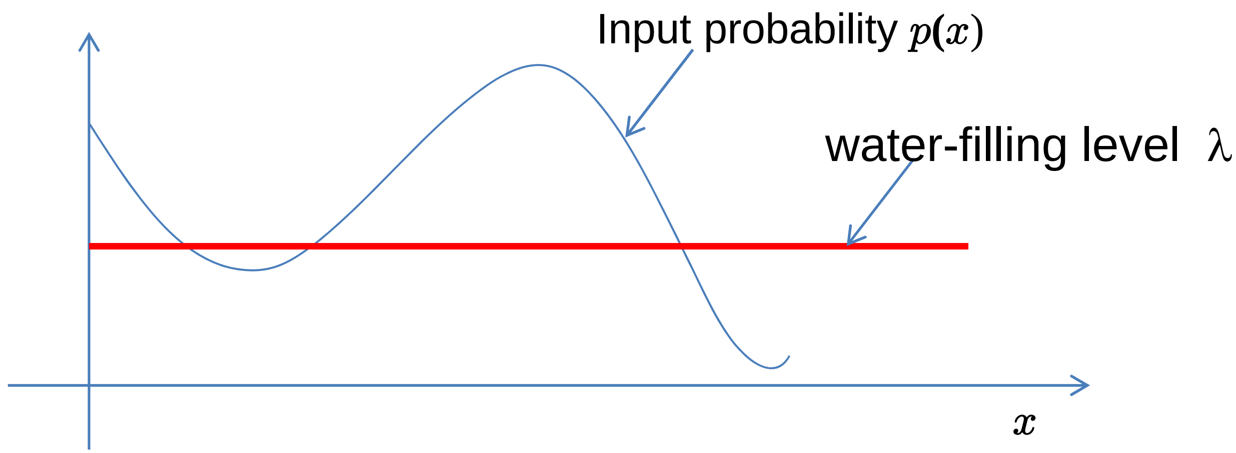
$$E = H(X \mid \tilde{X})$$

- The categorical database case has remained largely unaddressed in privacy research until now.

L. Sankar, S. R. Rajagopalan, and H. V. Poor. "An information-theoretic approach to privacy," *Proc. 48th Allerton Conf. Comm., Cntl., and Comp.*, Monticello, IL, Sep, 2010.

Example 1: Categorical Database

- Optimal input to output mapping: reverse 'water-filling'
  - Only x with $p(x) > \lambda$ revealed (λ : water-level).
- Eliminates samples with low probabilities (relative to water-level λ)
  - Equivalent to outlier aggregation/suppression (dominant statistical approaches)
  - Such samples reveal the most information
- As $D\uparrow$, $\lambda\uparrow$ (relative to distribution) to reveal fewer samples

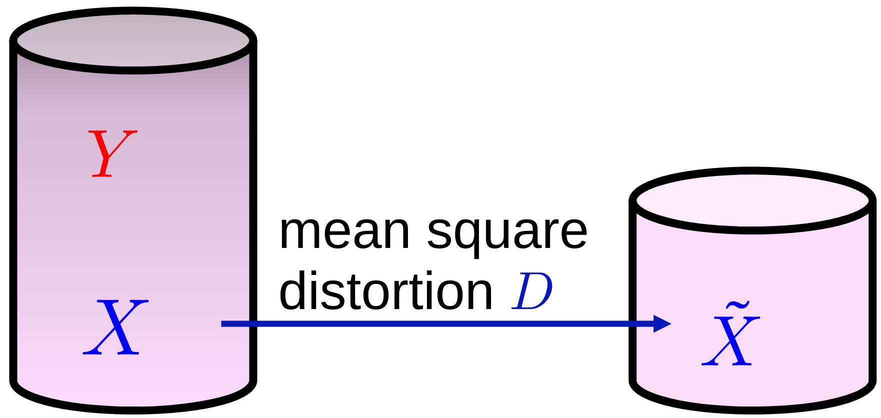


Example 2: Numerical Database

- Numerical data: finite/infinite alphabet real data
  - e.g.: results of medical tests, etc.
  - Medical research often assumes Gaussian distributed data

$$(X,Y) \sim \mathcal{N}(0,\Sigma)$$

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

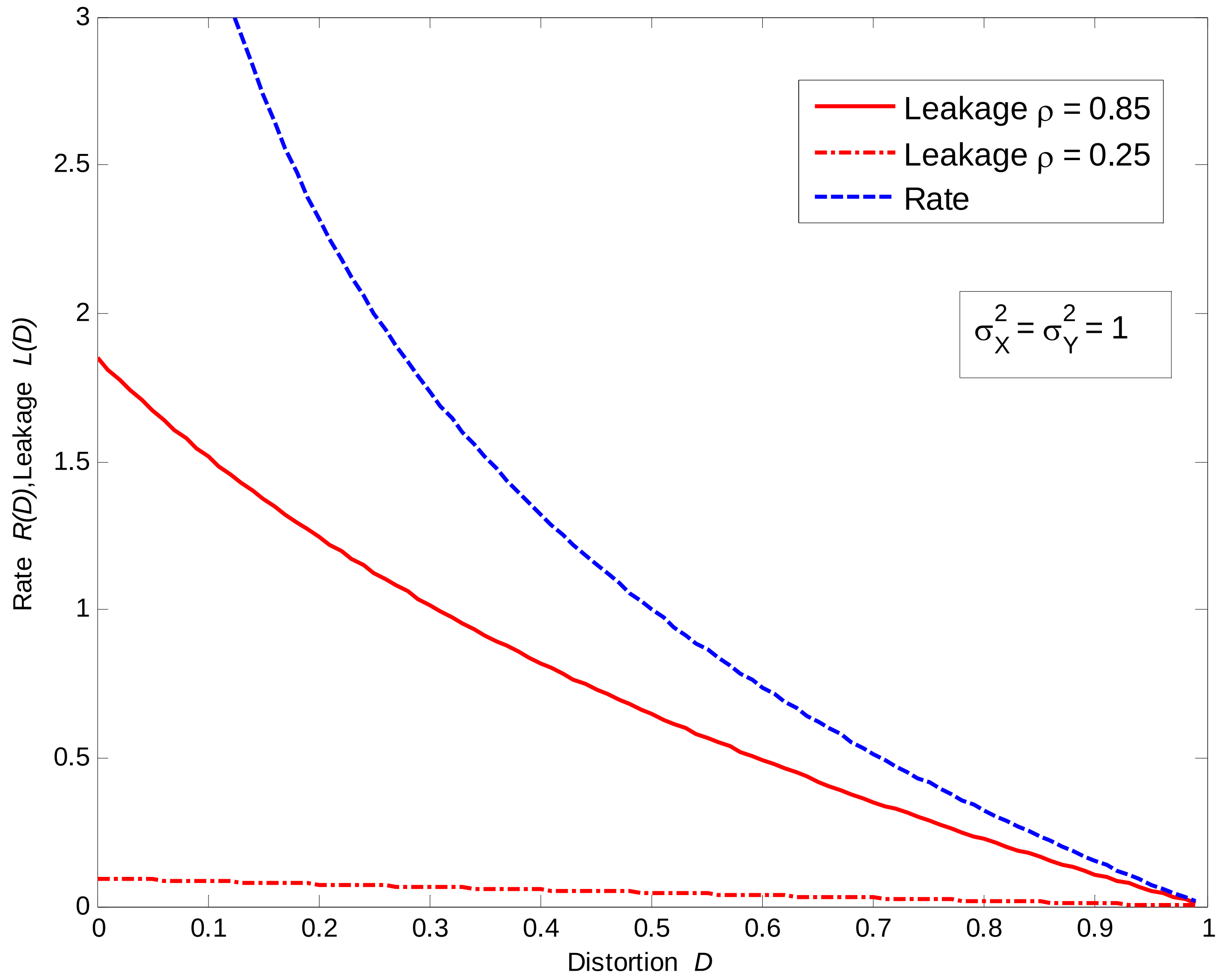


$$\boxed{\text{Privacy Leakage: } L = I(Y;\tilde{X})}$$

- Sanitized DB remains Gaussian distributed.
  - Gaussian $\tilde{X}$ achieves minimal $R(D,E)$ and maximal privacy $\Gamma(D)$

L. Sankar, S. R. Rajagopalan, and H. V. Poor. "An information-theoretic approach to privacy," *Proc. 48th Allerton Conf. Comm., Cntl., and Comp.*, Monticello, IL, Sep, 2010.

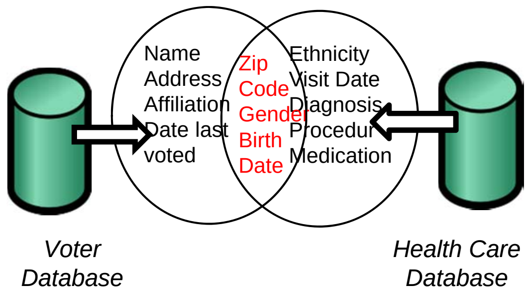
Example 2: Numerical Database



Related Results

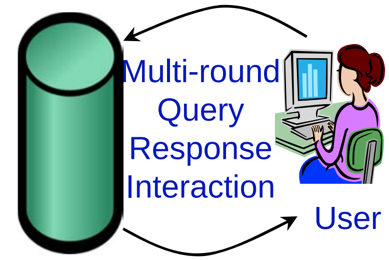
The Side Information Problem

Model and
U-P tradeoff
for decoder
side information



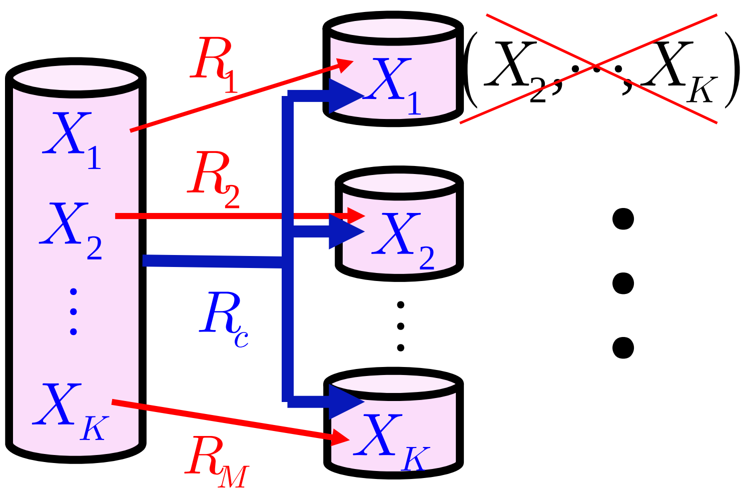
The Successive Disclosure Problem

Conditions for
no privacy leaks over
successive queries
relative to one-shot



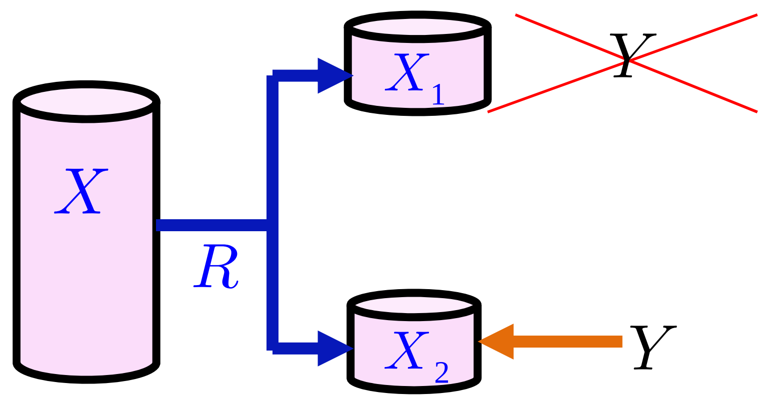
L. Sankar, S. Raj Rajagopalan, H. V. Poor, "A theory of privacy and utility in databases," submitted to the *IEEE Trans. Inform. Theory*, Feb. 2011.

Multi-user Privacy



R. Tandon, L. Sankar, H. V. Poor, "Multiuser Privacy and Common Information," submitted to *ISIT 2011.*

Discriminatory Coding and Privacy



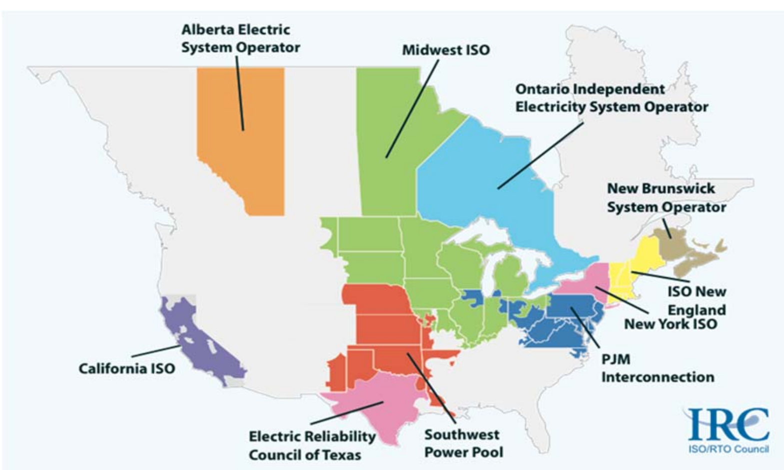
R. Tandon, L. Sankar, H. V. Poor, "Discriminatory Lossy Source Coding," submitted to *Globecom 2011.*

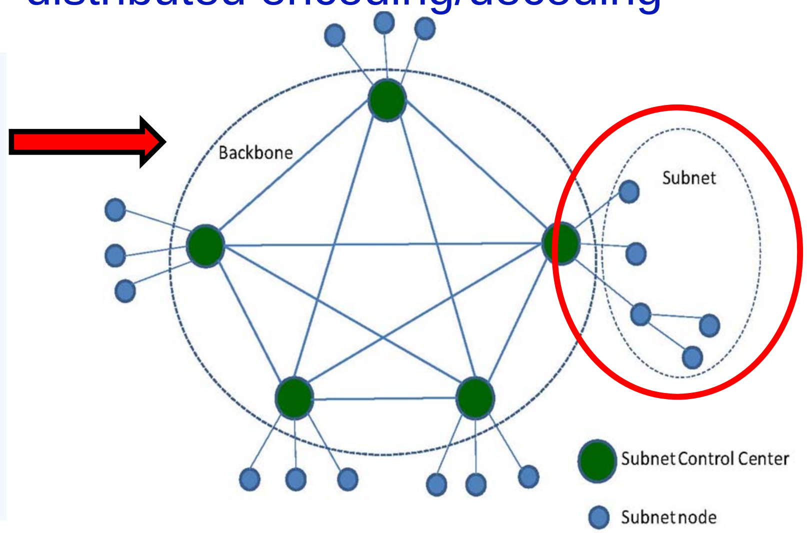
Talk Outline

- Database privacy problems
- Smart grid privacy problems
- Summary and future work

Smart Grid – Competitive Privacy

- N.A. Grid: interconnected regional transmission organizations which:
  - need to share measurements on state estimation for reliability (utility)
  - wish to withhold information for economic competitive reasons (privacy)
- Leads to a new problem of *competitive privacy*
  - Our results: precise quantification of state leakage (privacy) vs. estimation error (utility) and optimal communication scheme
  - New problem in source coding – distributed encoding/decoding

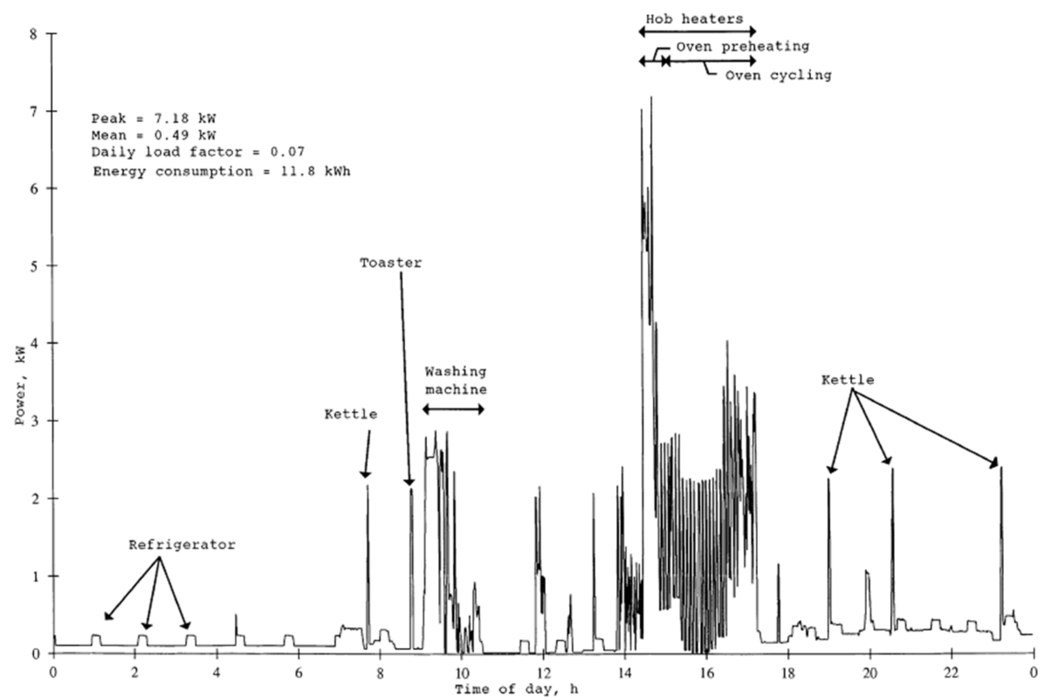




L. Sankar, S. Kar, R. Tandon, and H. V. Poor, “Competitive privacy in the smart grid: An information-theoretic approach,” submitted to *IEEE SmartGridComm*, Apr. 2011.

Smart Grid – Smart Meter Privacy

- Smart meter is a critical enabler of the Smart Grid
- For consumers: Tariff- and load-aware appliance usage
- For electricity suppliers: Load balancing; data mining (analytics)
  - Data mining: tremendous utility to supplier; huge consumer privacy risk
- Time-series data: utility-privacy tradeoff via rate-distortion for sources with memory



S. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor, "Smart meter privacy: Utility-privacy tradeoff," submitted to *IEEE SmartGridComm*, Apr. 2011.

Talk Outline

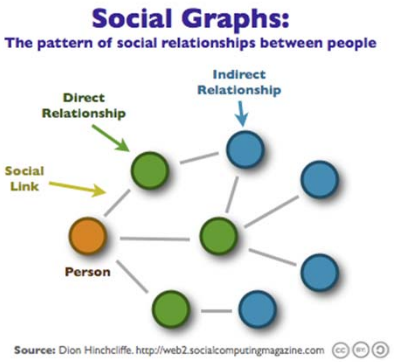
- Database privacy problem
- Smart grid privacy problems
- Summary and Future Work

Summary

- The privacy problem is immediate and here to stay … and multiply…
- One solution will not fit all applications…
- But a framework provides the much needed abstraction
- More needs to be done…

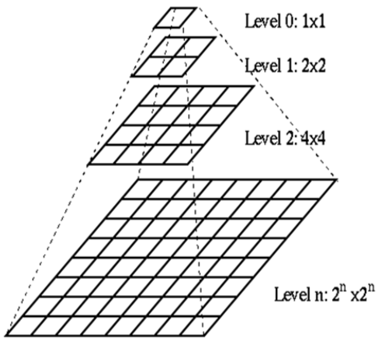
Trying to ward off regulators, the advertising industry has agreed on a standard icon — a little “i” — that it will add to most online ads that use demographics and behavioral data to tell consumers what is happening. – NY Times, Jan. 26, 2010.

Future Work



Privacy in Social Networks:

- Quantifying privacy and utility in social networks
  - Information leakage due to social graph
  - How to quantify utility?



Practical Privacy via Signal Processing:

- Compressive sensing, quantization, clustering, …
- Universal lossy coding schemes

Medical Database Privacy:

- De-identification and privacy
- Does synthetic data suffice?
- Need for re-identification?

Multi-Disciplinary Research

Communication Theory

Information and Coding Theory

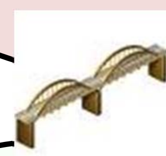
Universal Utility-Privacy Tradeoff Framework

Probability Theory and Statistics

Game Theory

Estimation Theory

Signal Processing



Computer Science Statistics

For more: … http://www.arxiv.org

Thank you!