Privacy Cognizant Information Systems

Rakesh Agrawal
IBM Almaden Research Center

Jt. work with Srikant, Kiernan, Xu & Evfimievski

Thesis

- There is increasing need to build information systems that
  - protect the privacy and ownership of information
  - do not impede the flow of information
- Cross-fertilization of ideas from the security and database research communities can lead to the development of innovative solutions.

Outline

- Motivation
- Privacy Preserving Data Mining
- Privacy Aware Data Management
- Information Sharing Across Private Databases
- Conclusions

Drivers

- Policies and Legislations
  - U.S. and international regulations
  - Legal proceedings against businesses
- Consumer Concerns
  - Consumer privacy apprehensions continue to plague the Web … these fears will hold back roughly $15 billion in e-Commerce revenue." Forrester Research, 2001
  - Most consumers are "privacy pragmatists." Westin Surveys
- Moral Imperative
  - The right to privacy: the most cherished of human freedom -- Warren & Brandeis, 1890

Outline

- Motivation
- Privacy Preserving Data Mining
- Privacy Aware Data Management
- Information Sharing Across Private Databases
- Conclusions

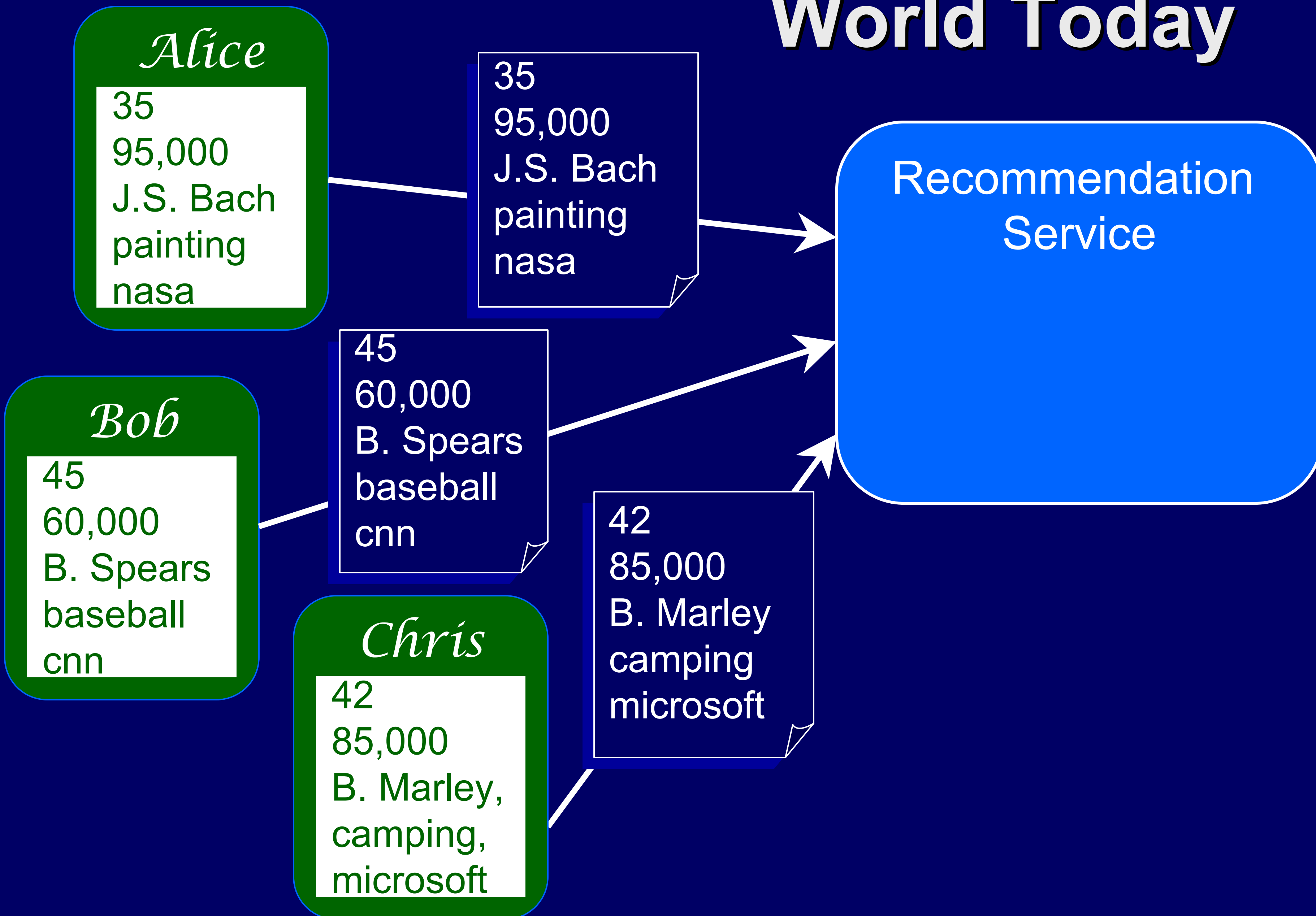
Data Mining and Privacy

- The primary task in data mining:
  - development of models about aggregated data.
- Can we develop accurate models, while protecting the privacy of individual records?

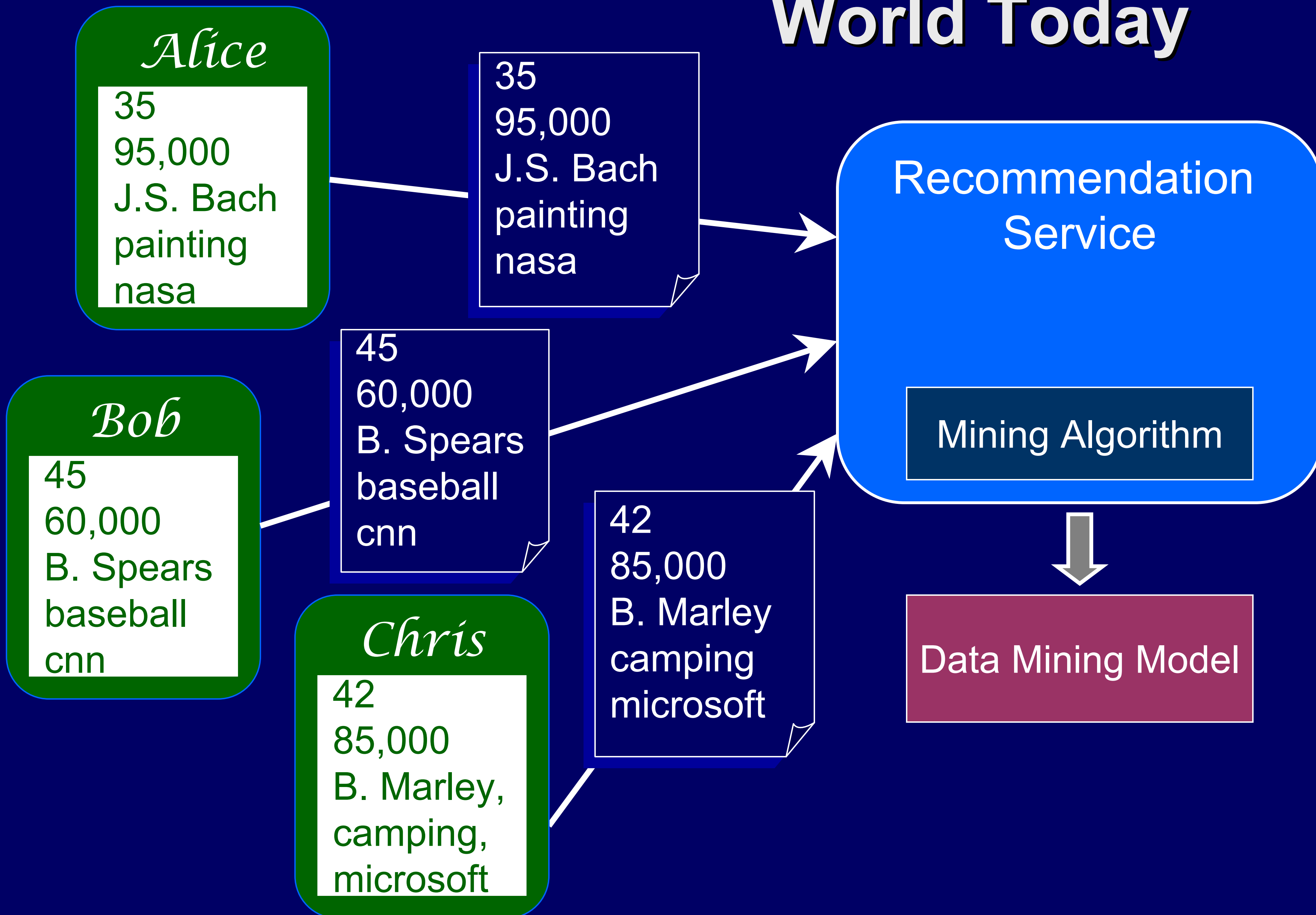
Setting

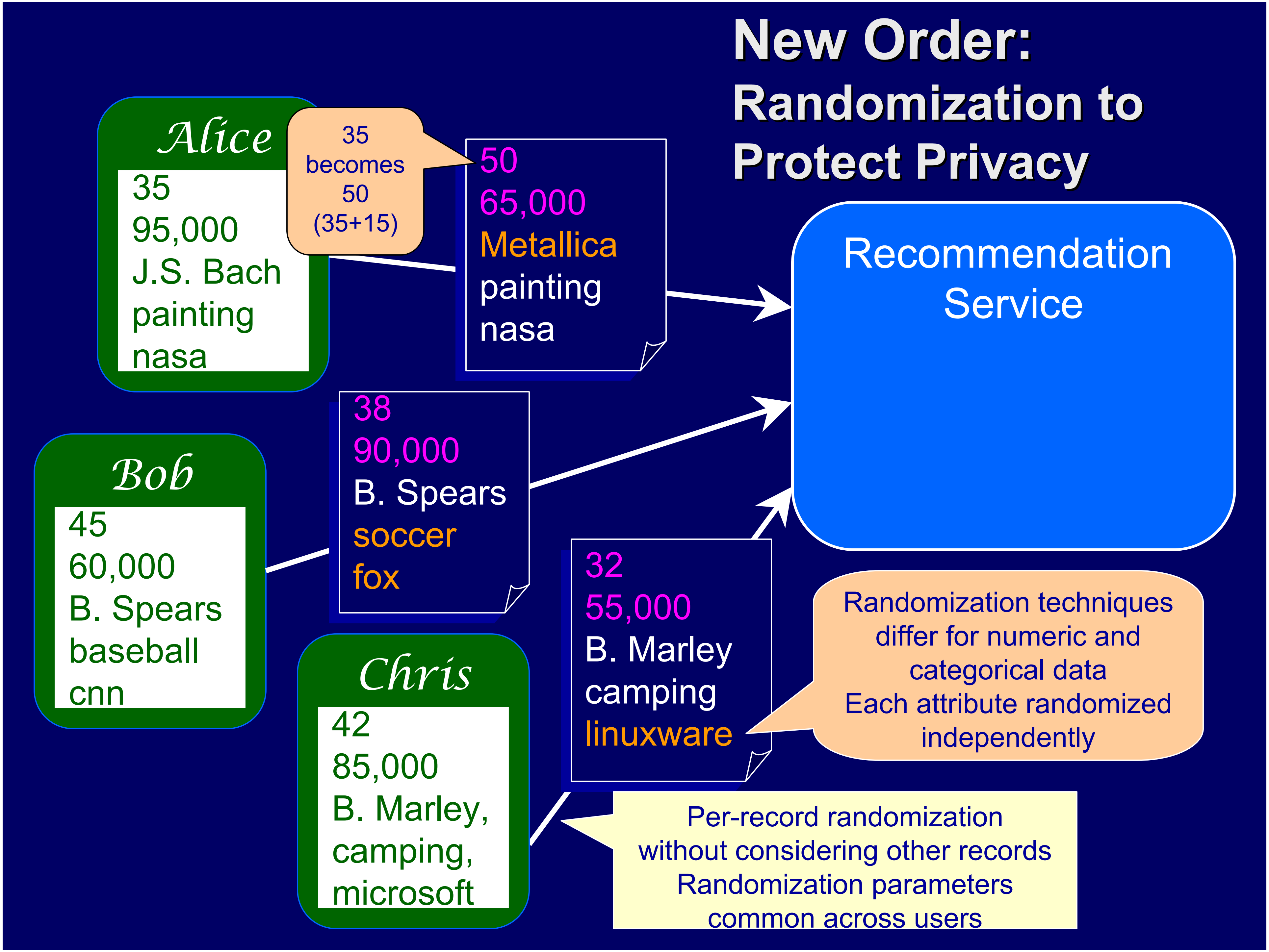
- Application scenario: A central server interested in building a data mining model using data obtained from a large number of clients, while preserving their privacy
  - Web-commerce, e.g. recommendation service
- Desiderata:
  - Must not slow-down the speed of client interaction
  - Must scale to very large number of clients
- During the application phase
  - Ship model to the clients
  - Use oblivious computations

World Today

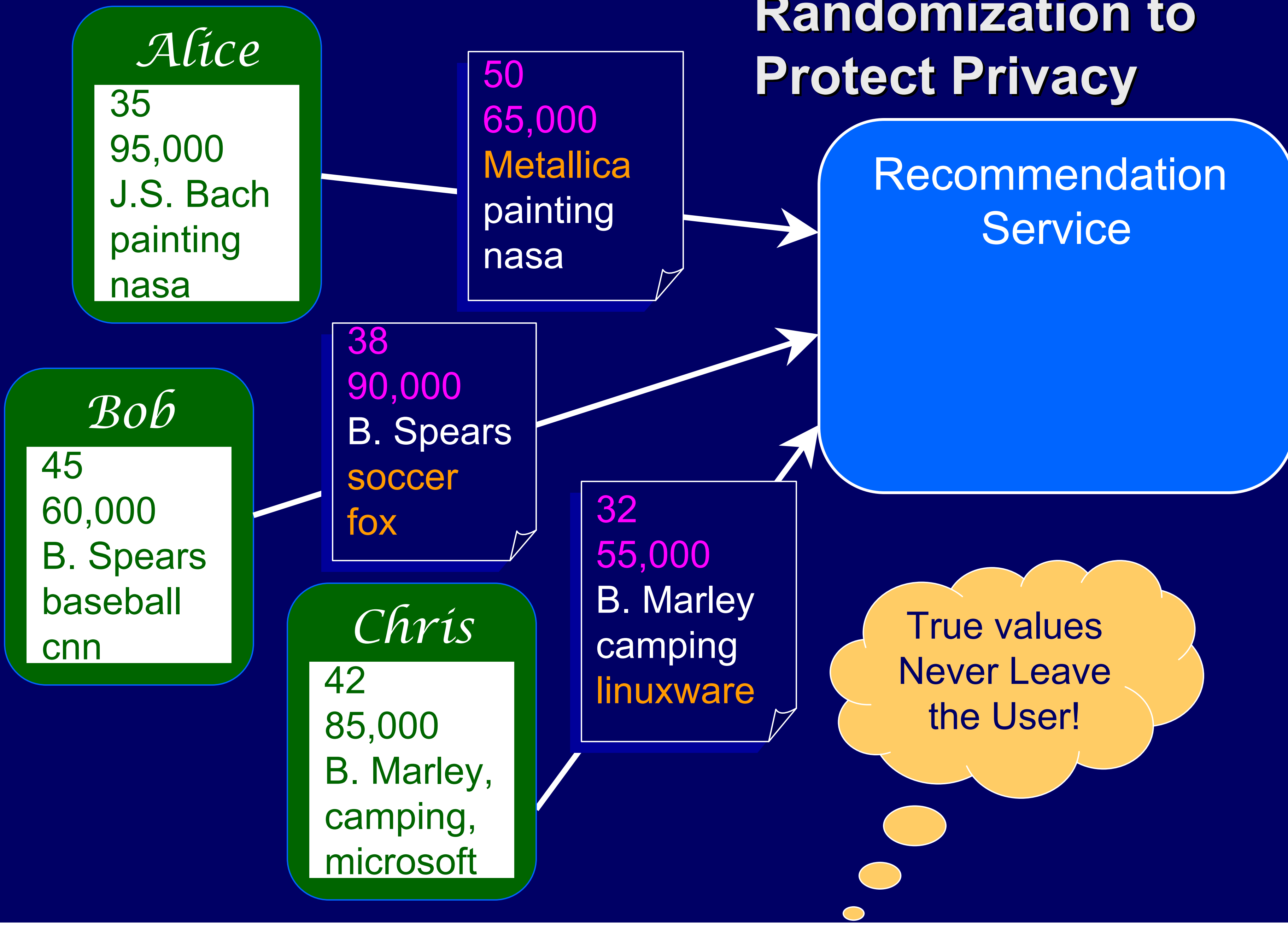

Alice
35
95,000
J.S. Bach
painting
nasa
35
95,000
J.S. Bach
painting
nasa
Recommendation Service
Bob
45
60,000
B. Spears
baseball
cnn
45
60,000
B. Spears
baseball
cnn
Chris
42
85,000
B. Marley,
camping,
microsoft
42
85,000
B. Marley
camping
microsoft

World Today

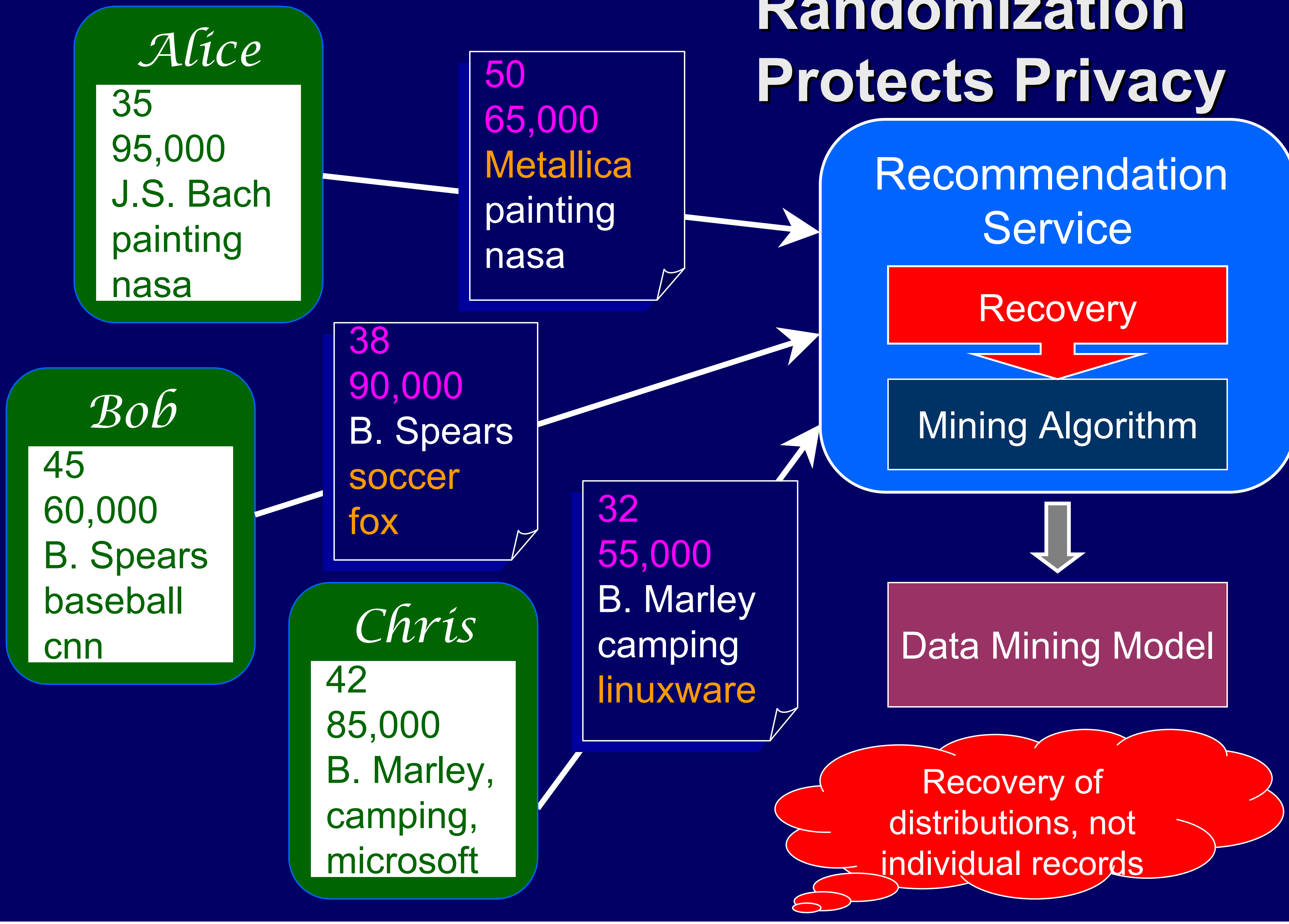

Alice
35
95,000
J.S. Bach
painting
nasa
Bob
45
60,000
B. Spears
baseball
cnn
Chris
42
85,000
B. Marley,
camping,
microsoft
35
95,000
J.S. Bach
painting
nasa
45
60,000
B. Spears
baseball
cnn
42
85,000
B. Marley
camping
microsoft
Recommendation Service
Mining Algorithm
Data Mining Model


New Order: Randomization to Protect Privacy
Alice
35
95,000
J.S. Bach
painting
nasa
35 becomes 50 (35+15)
50
65,000
Metallica
painting
nasa
Recommendation Service
Bob
45
60,000
B. Spears
baseball
cnn
38
90,000
B. Spears
soccer
fox
Chris
42
85,000
B. Marley,
camping,
microsoft
32
55,000
B. Marley
camping
linuxware
Randomization techniques differ for numeric and categorical data
Each attribute randomized independently
Per-record randomization without considering other records
Randomization parameters common across users

New Order: Randomization to Protect Privacy



New Order: Randomization Protects Privacy



Reconstruction Problem (Numeric Data)

- Original values $x_1, x_2, ..., x_n$
  - from probability distribution X (unknown)
- To hide these values, we use $y_1, y_2, ..., y_n$
  - from probability distribution Y
- Given
  - $x_1+y_1, x_2+y_2, ..., x_n+y_n$
  - the probability distribution of Y

  Estimate the probability distribution of X.

Reconstruction Algorithm

f_X^0 := Uniform distribution
j := 0
repeat

$$f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y((x_i+y_i)-a)\,f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i+y_i)-a)\,f_X^j(a)}$$

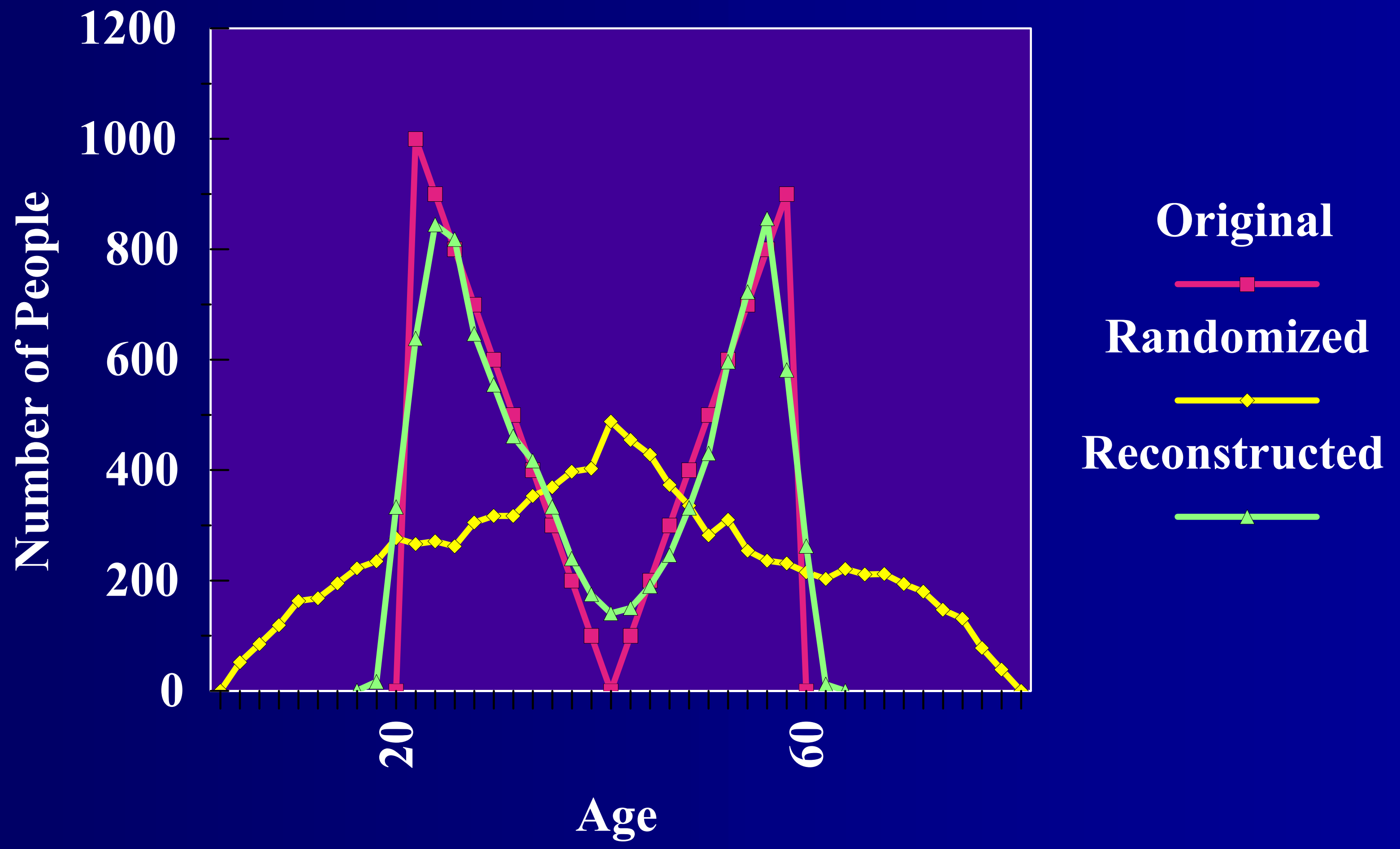
Bayes' Rule

j := j+1
until (stopping criterion met)

(R. Agrawal & R. Srikant, SIGMOD 2000)

- Converges to maximum likelihood estimate.
  - D. Agrawal & C.C. Aggarwal, PODS 2001.

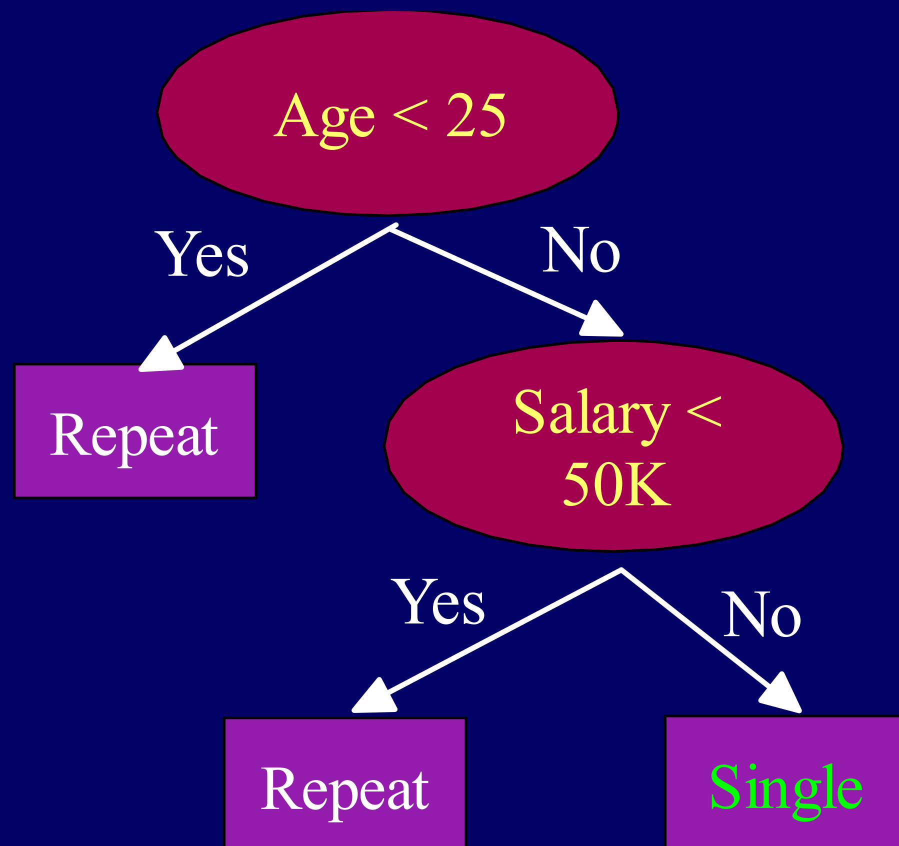
Works Well



1200
1000
800
600
400
200
0
Number of People
20
60
Age
Original
Randomized
Reconstructed

Decision Tree Example

Age	Salary	Repeat Visitor?
23	50K	Repeat
17	30K	Repeat
43	40K	Repeat
68	50K	Single
32	70K	Single
20	20K	Repeat



Algorithms

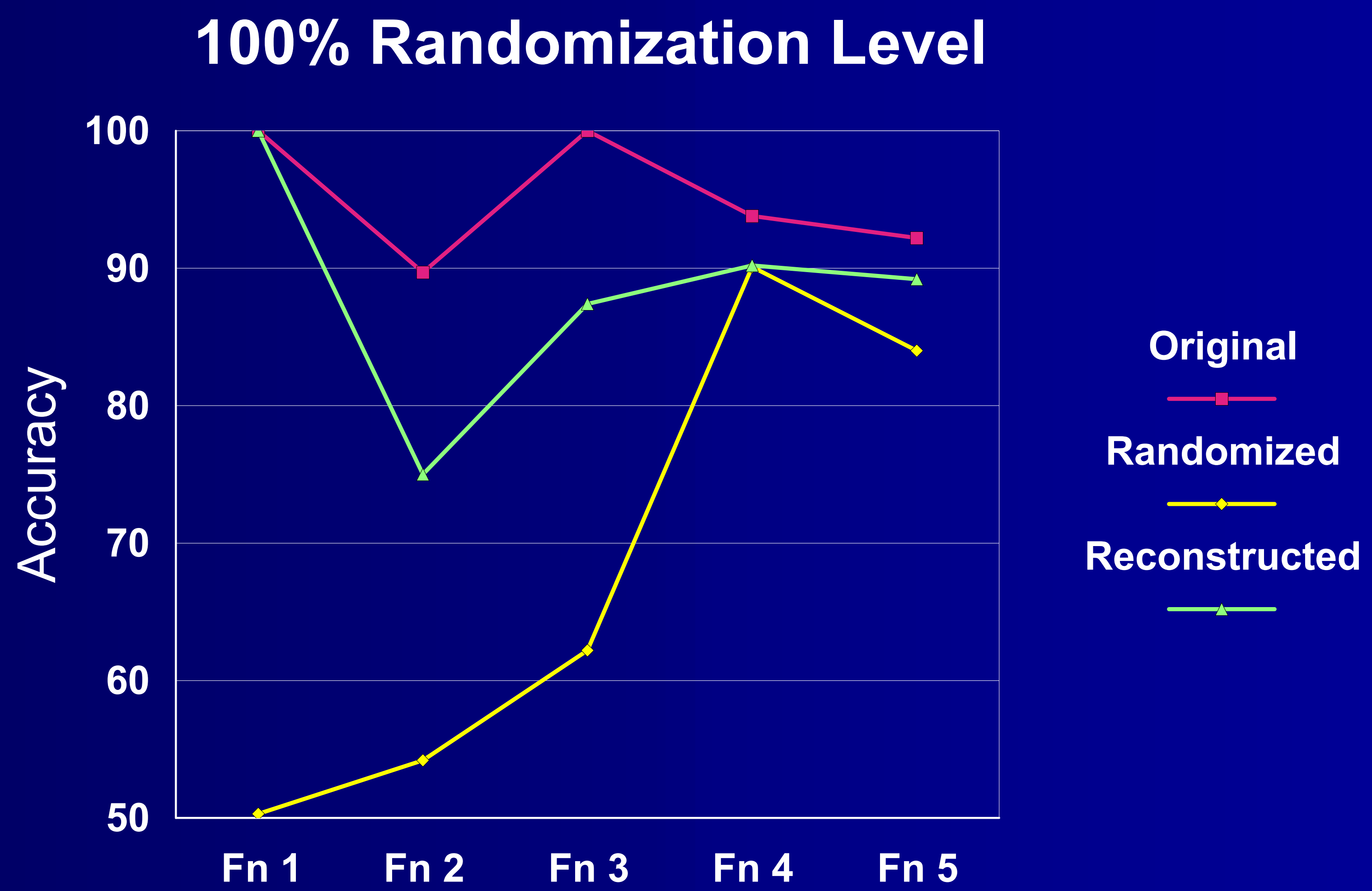
- Global
  - Reconstruct for each attribute once at the beginning
- By Class
  - For each attribute, first split by class, then reconstruct separately for each class.
- Local
  - Reconstruct at each node

See SIGMOD 2000 paper for details.

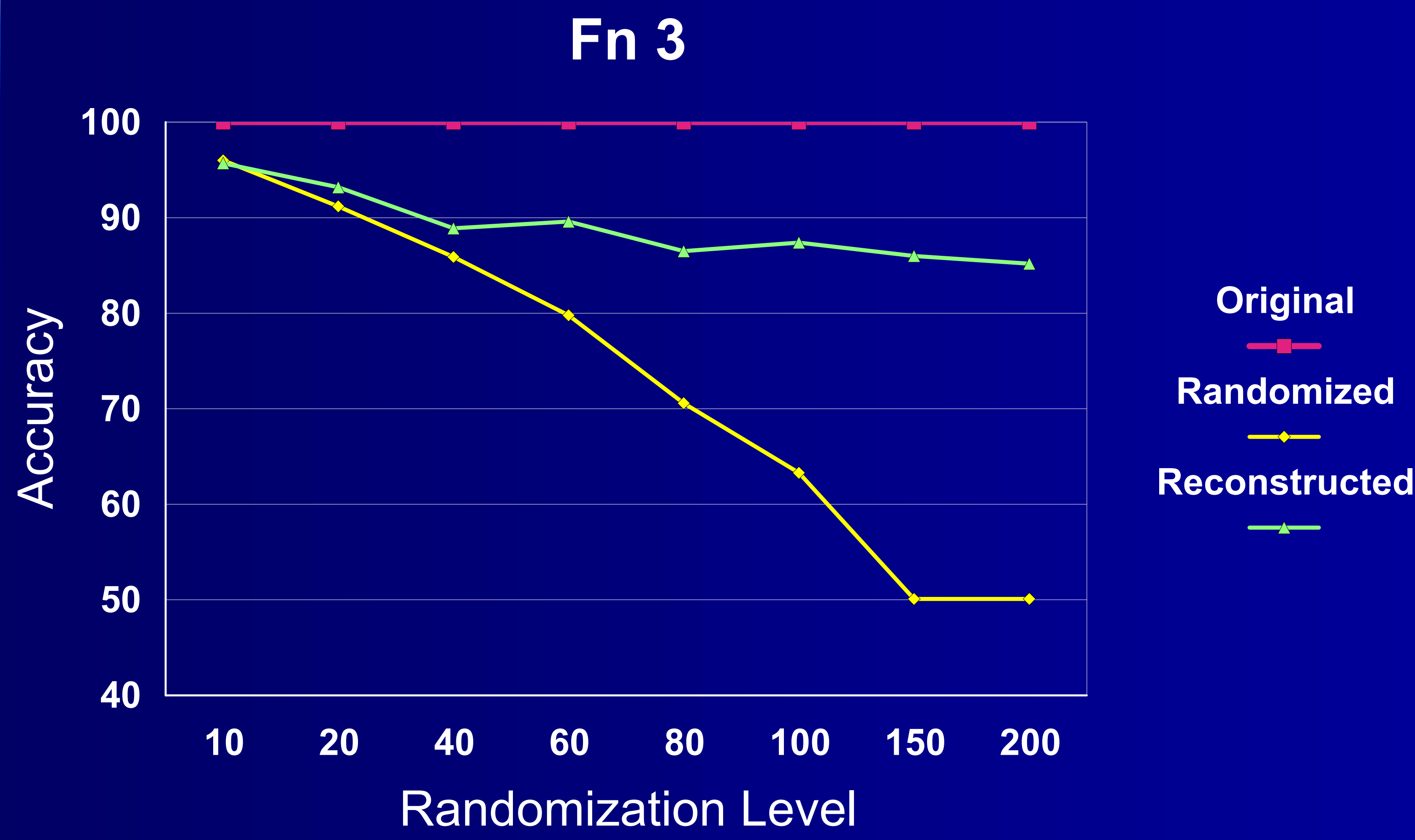
Experimental Methodology

- Compare accuracy against
  - Original: unperturbed data without randomization.
  - Randomized: perturbed data but without making any corrections for randomization.
- Test data not randomized.
- Synthetic benchmark from [AGI+92].
- Training set of 100,000 records, split equally between the two classes.

Decision Tree Experiments



Accuracy vs. Randomization



More on Randomization

- Privacy-Preserving Association Rule Mining Over Categorical Data
  - Rizvi & Haritsa [VLDB 02]
  - Evfimievski, Srikant, Agrawal, & Gehrke [KDD-02]
- Privacy Breach Control: Probabilistic limits on what one can infer with access to the randomized data as well as mining results
  - Evfimievski, Srikant, Agrawal, & Gehrke [KDD-02]
  - Evfimievski, Gehrke & Srikant [PODS-03]

Related Work:
Private Distributed ID3

- How to build a decision-tree classifier on the union of two private databases (Lindell & Pinkas [Crypto 2000])
- Basic Idea:
  - Find attribute with highest information gain privately
  - Independently split on this attribute and recurse
- Selecting the Split Attribute
  - Given v1 known to DB1 and v2 known to DB2, compute (v1 + v2) log (v1 + v2) and output random shares of the answer
  - Given random shares, use Yao's protocol [FOCS 84] to compute information gain.
- Trade-off
  - \+ Accuracy
  - – Performance & scaling

Related Work: Purdue Toolkit

- Partitioned databases (horizontally + vertically)
- Secure Building Blocks
- Algorithms (using building blocks):
  - Association rules
  - EM Clustering
- C. Clifton et al. Tools for Privacy Preserving Data Mining. SIGKDD Explorations 2003.

Related Work: Statistical Databases

- Provide statistical information without compromising sensitive information about individuals (AW89, Sho82)
- Techniques
  - Query Restriction
  - Data Perturbation
- Negative Results: cannot give high quality statistics and simultaneously prevent partial disclosure of individual information [AW89]

Summary

- Promising technical direction & results
- Much more needs to be done, e.g.
  - Trade off between the amount of privacy breach and performance
  - Examination of other approaches (e.g. randomization based on swapping)

Outline

- Motivation
- Privacy Preserving Data Mining
- Privacy Aware Data Management
- Information Sharing Across Private Databases
- Conclusions

Hippocratic Databases

- Hippocratic Oath, 8 (circa 400 BC)
  – What I may see or hear in the course of treatment … I will keep to myself.
- What if the database systems were to embrace the Hippocratic Oath?

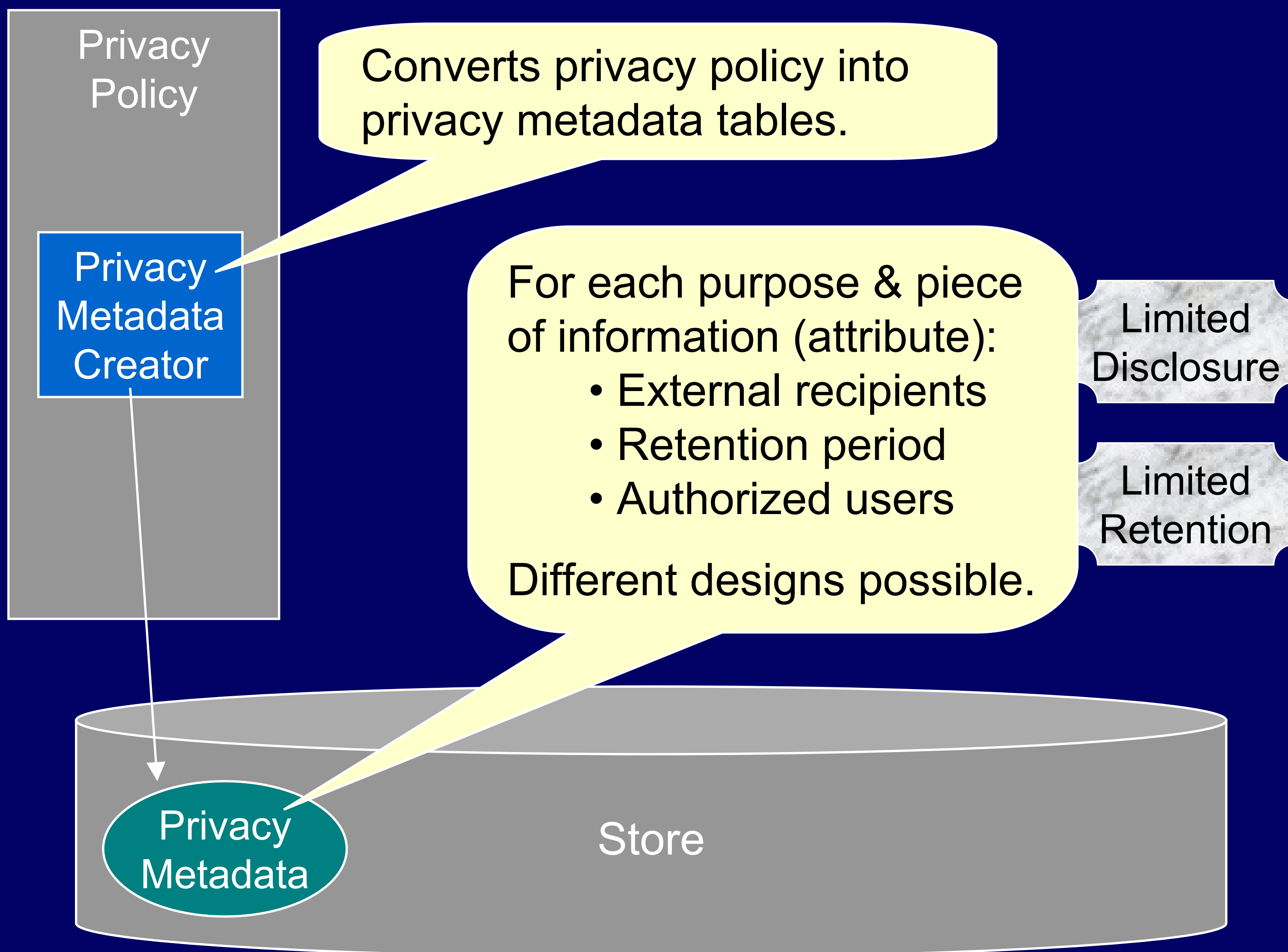
- Architecture derived from privacy legislations.
  – US (FIPA, 1974), Europe (OECD , 1980), Canada (1995), Australia (2000), Japan (2003)

- Agrawal, Kiernan, Srikant & Xu: VLDB 2002..

Architectural Principles

- **Purpose Specification**
  Associate with data the purposes for collection
- **Consent**
  Obtain donor's consent on the purposes
- **Limited Collection**
  Collect minimum necessary data
- **Limited Use**
  Run only queries that are consistent with the purposes
- **Limited Disclosure**
  Do not release data without donor's consent
- **Limited Retention**
  Do not retain data beyond necessary
- **Accuracy**
  Keep data accurate and up-to-date
- **Safety**
  Protect against theft and other misappropriations
- **Openness**
  Allow donor access to data about the donor
- **Compliance**
  Verifiable compliance with the above principles

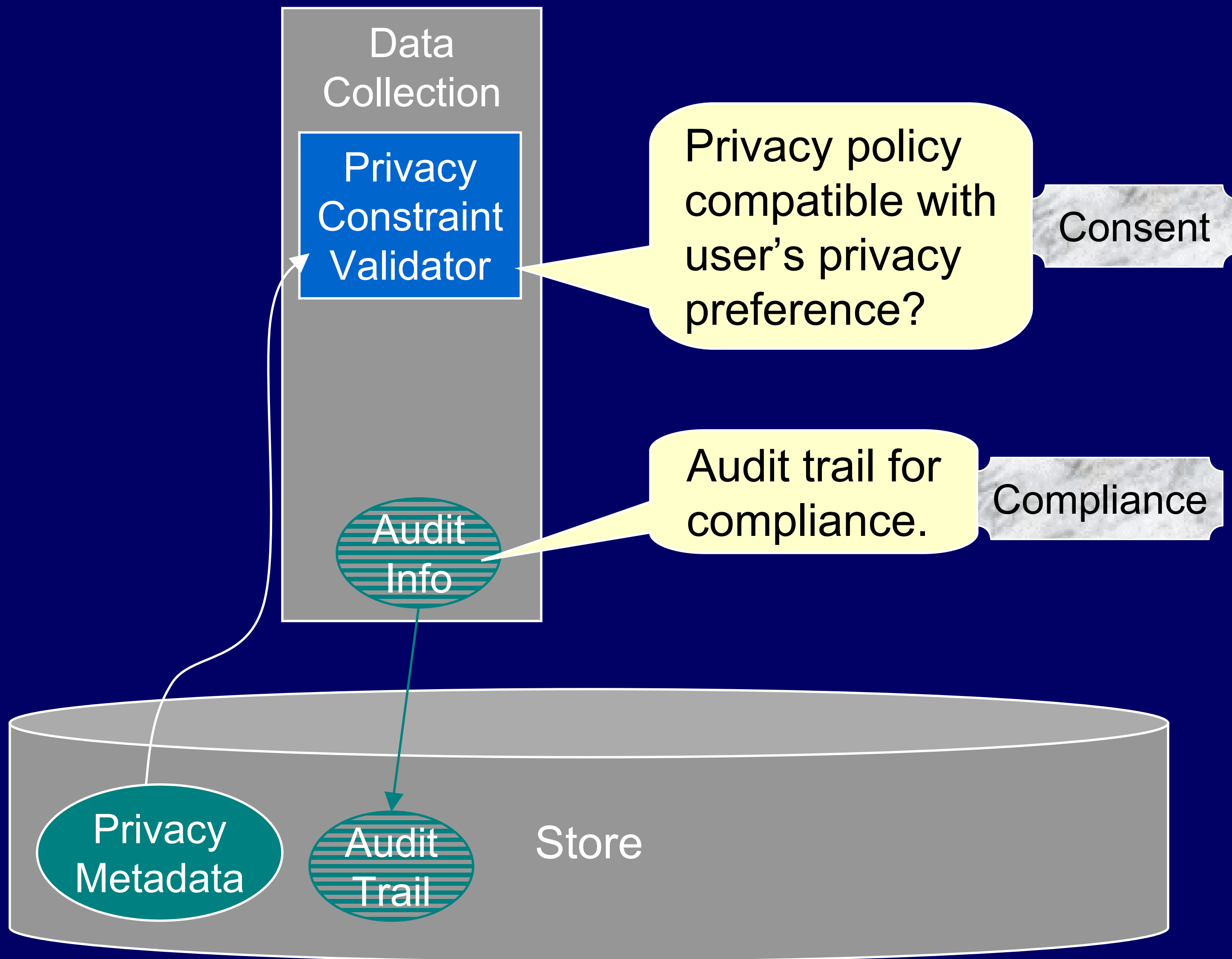
Architecture: Policy



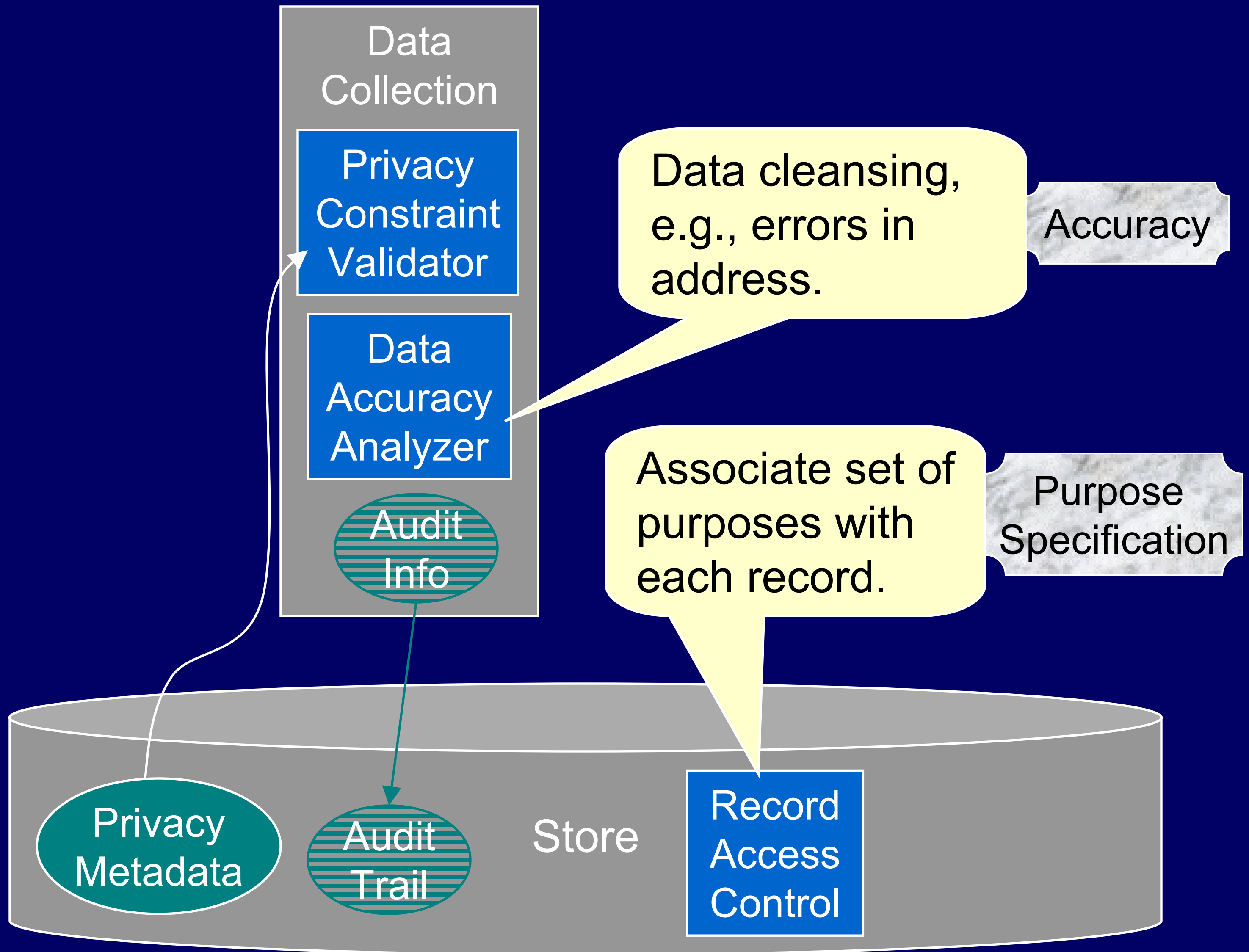
Privacy Policies Table

Purpose	Table	Attribute	External-recipients	Authorized-users	Retention
purchase	customer	name	{delivery, credit-card}	{shipping, charge}	1 month
purchase	customer	email	*empty*	{shipping}	1 month
register	customer	name	empty	{registration}	3 years
register	customer	email	empty	{registration}	3 years
recommendations	order	book	empty	{mining}	10 years

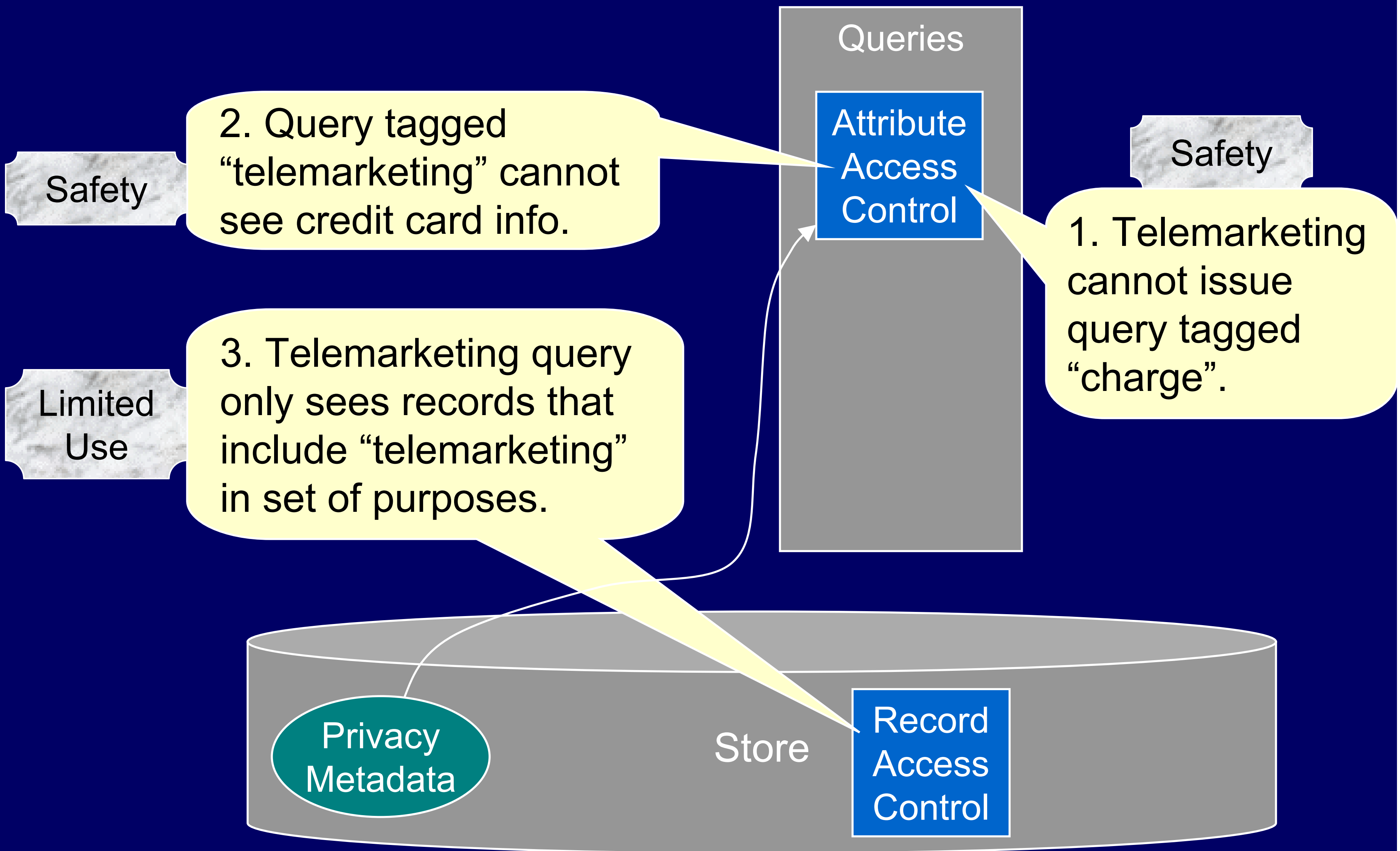
Architecture: Data Collection



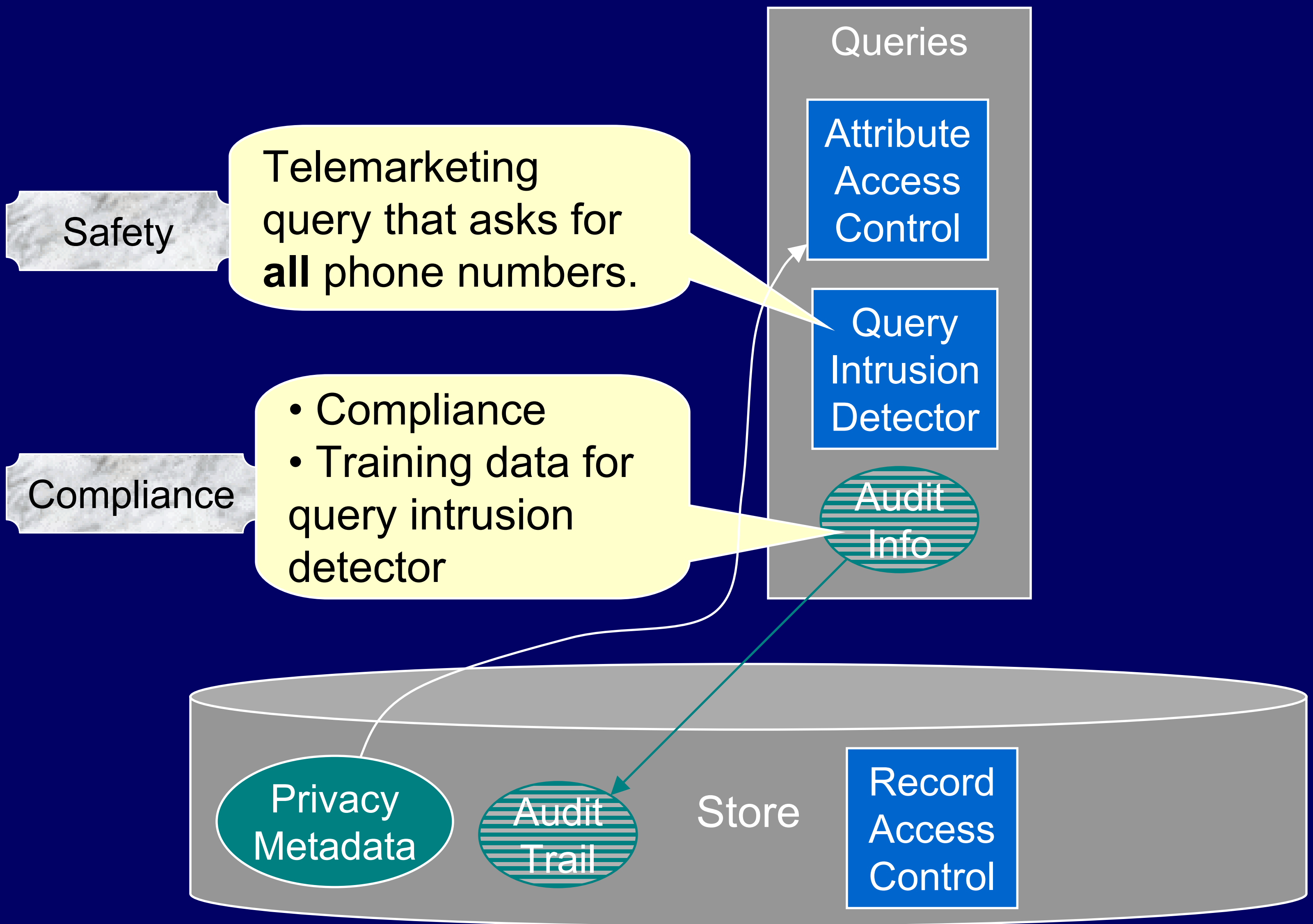
Architecture: Data Collection



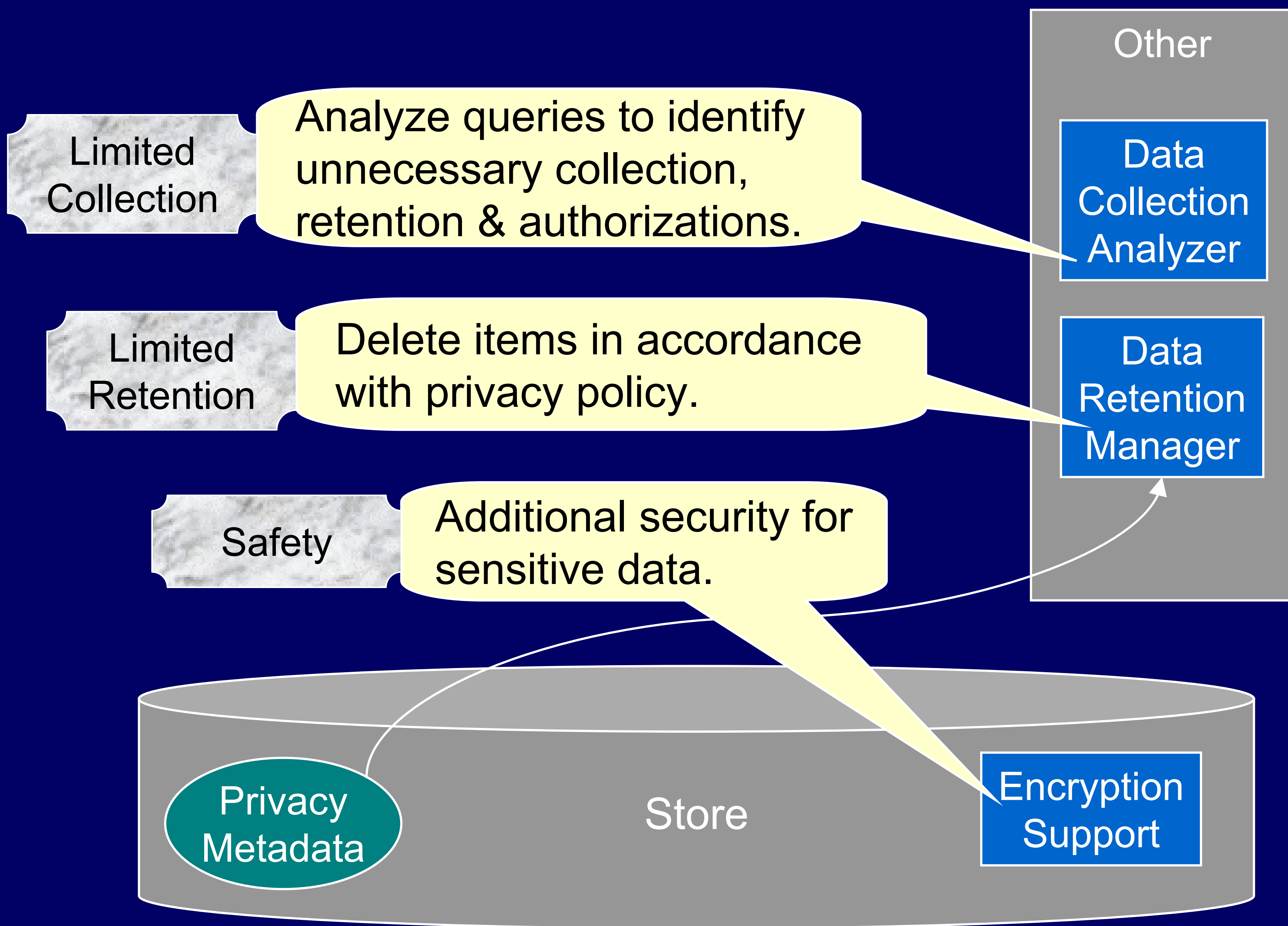
Architecture: Queries



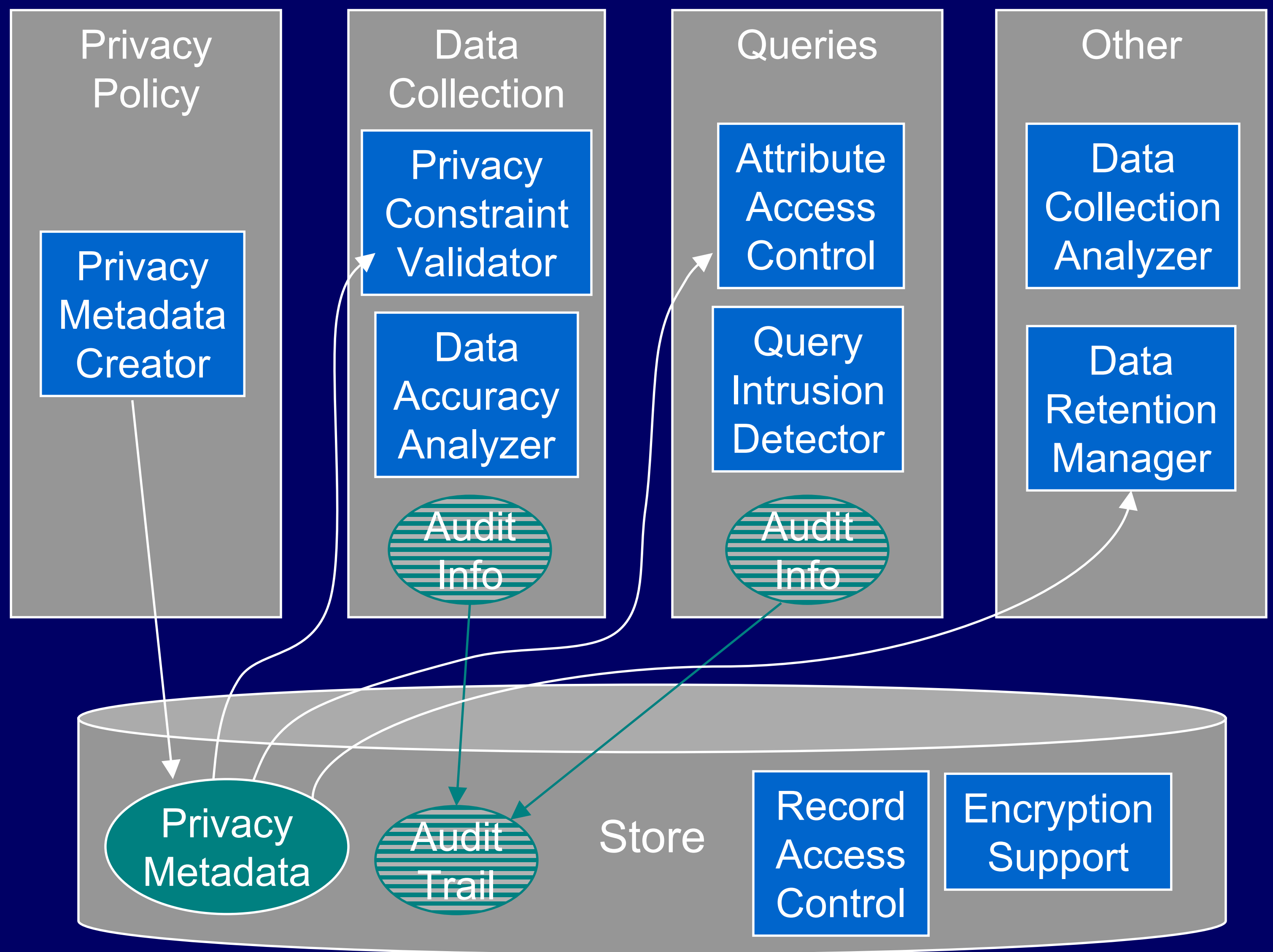
Architecture: Queries



Architecture: Other



Architecture



Related Work: Statistical & Secure Databases

- Statistical Databases
  - Provide statistical information (sum, count, etc.) without compromising sensitive information about individuals, [AW89]
- Multilevel Secure Databases
  - Multilevel relations, e.g., records tagged "secret", "confidential", or "unclassified", e.g. [JS91]
- Need to protect privacy in transactional databases that support daily operations.
  - Cannot restrict queries to statistical queries.
  - Cannot tag all the records "top secret".

Some Interesting Problems

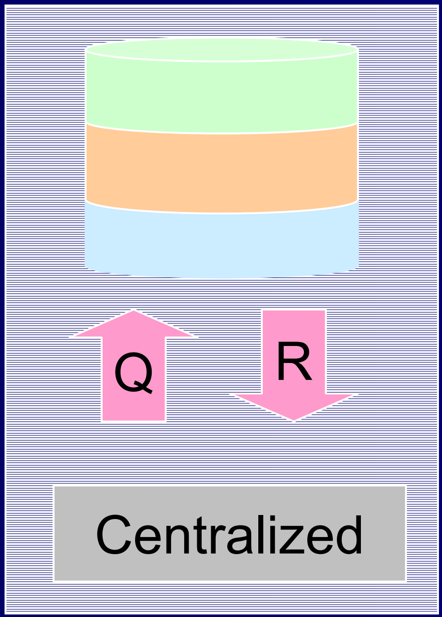
- Privacy enforcement requires cell-level decisions (which may be different for different queries)
  - How to minimize the cost of privacy checking?
- Encryption to avoid data theft
  - How to index encrypted data for range queries?
- Intrusive queries from authorized users
  - Query intrusion detection?
- Identifying unnecessary data collection
  - Assets info needed only if salary is below a threshold
  - Queries only ask “Salary > threshold” for rent application
- Forgetting data after the purpose is fulfilled
  - Databases designed not to lose data
  - Interaction with compliance

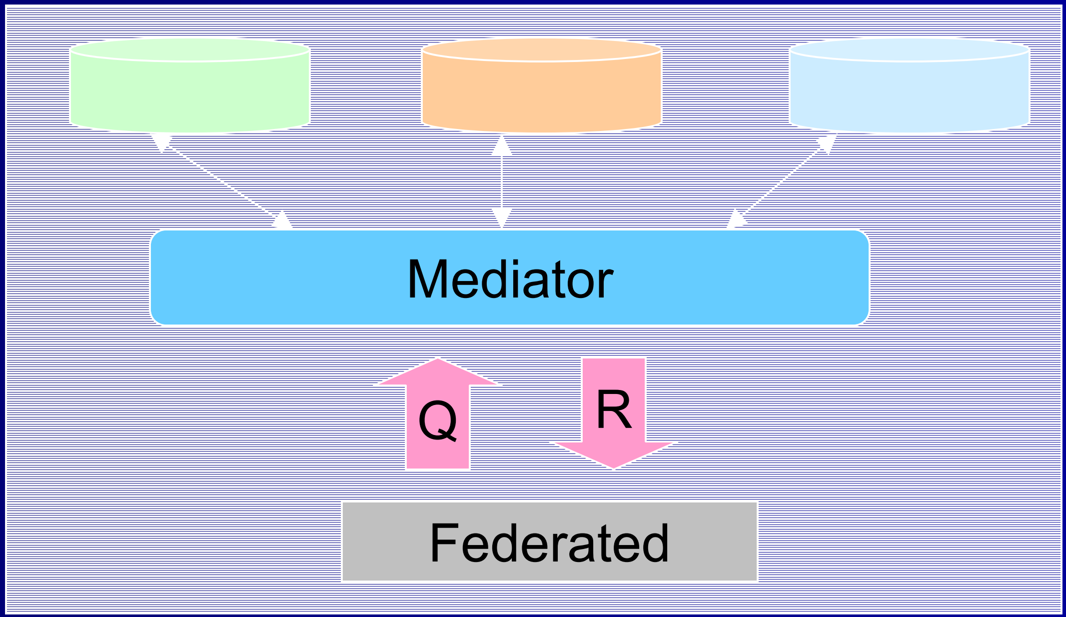
Solutions must scale to database-size problems!

Outline

- Motivation
- Privacy Preserving Data Mining
- Privacy Aware Data Management
- Information Sharing Across Private Databases
- Conclusions

Today's Information Sharing Systems





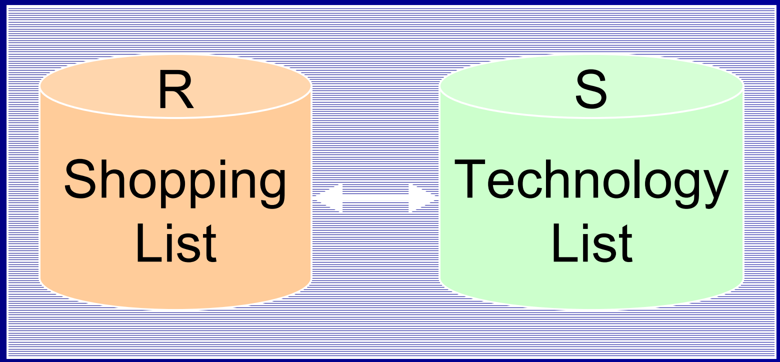
Assumption: Information in each database can be freely shared.

Minimal Necessary Information Sharing

- Compute queries across databases so that no more information than necessary is revealed (without using a trusted third party).
- Need is driven by several trends:
  - End-to-end integration of information systems across companies.
  - Simultaneously compete and cooperate.
  - Security: need-to-know information sharing
- Agrawal, Evfimievski & Srikant: SIGMOD 2003.

Selective Document Sharing

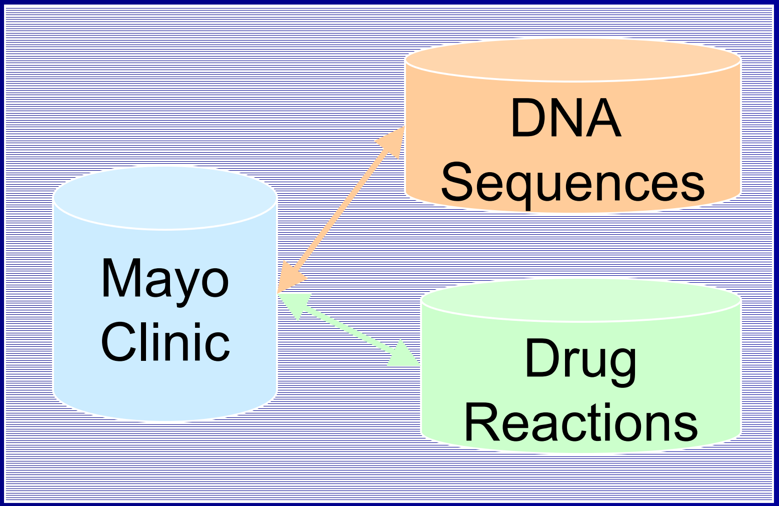
- R is shopping for technology.
- S has intellectual property it may want to license.
- First find the specific technologies where there is a match, and then reveal further information about those.



Example 2: Govt. agencies sharing information on a need-to-know basis.

Medical Research

- Validate hypothesis between adverse reaction to a drug and a specific DNA sequence.
- Researchers should not learn anything beyond 4 counts:



	Adverse Reaction	No Adv. Reaction
Sequence Present	?	?
Sequence Absent	?	?

Minimal Necessary Sharing

R	
a	
u	
v	
x	

S	
b	
u	
v	
y	

R ❁ S

- R must not know that S has b & y
- S must not know that R has a & x

R ❁ S
u
v

Count (R ❁ S)

- R & S do not learn anything except that the result is 2.

Problem Statement: Minimal Sharing

- Given:
  - Two parties (honest-but-curious): R (receiver) and S (sender)
  - Query Q spanning the tables R and S
  - Additional (pre-specified) categories of information I
- Compute the answer to Q and return it to R without revealing any additional information to either party, except for the information contained in I
  - For intersection, intersection size & equijoin,

    I = { |R| , |S| }
  - For equijoin size, I also includes the distribution of duplicates & some subset of information in $R \bowtie S$

A Possible Approach

- Secure Multi-Party Computation
  - Given two parties with inputs x and y, compute f(x,y) such that the parties learn only f(x,y) and nothing else.
  - Can be solved by building a combinatorial circuit, and simulating that circuit [Yao86].
- Prohibitive cost for database-size problems.
  - Intersection of two relations of a million records each would require 144 days

Intersection Protocol: Intuition

- Want to encrypt the value in R and S and compare the encrypted values.
- However, want an encryption function such that it can only be jointly computed by R and S, not separately.

Commutative Encryption

Commutative encryption F is a computable function

f : Key F X Dom F -> Dom F, satisfying:

- For all e, e' $\in$ Key F, $f_e \circ f_{e'} = f_{e'} \circ f_e$

  (The result of encryption with two different keys is the same, irrespective of the order of encryption)
- Each f_e is a bijection.

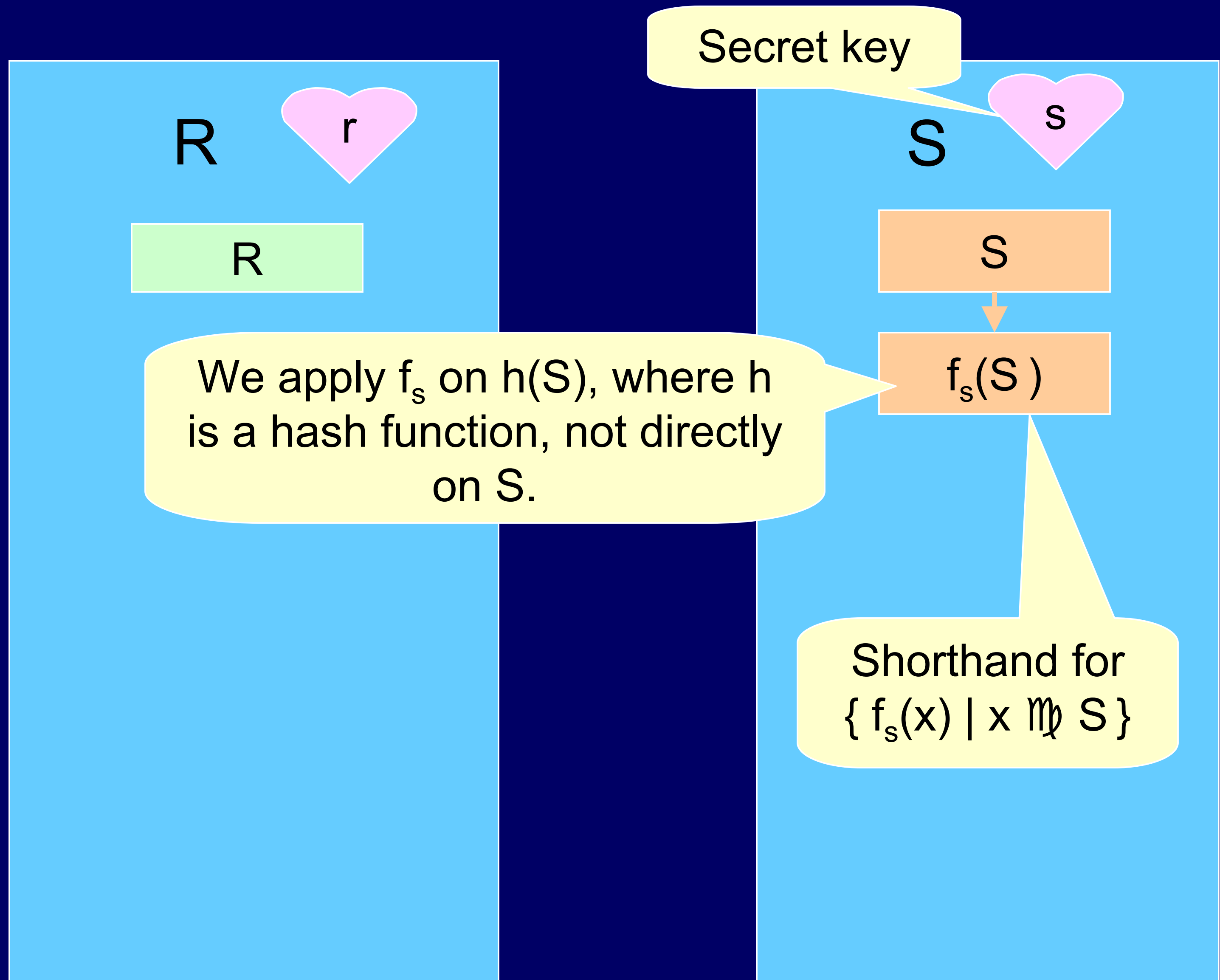
  (Two different values will have different encrypted values)
- The distribution of $<x, f_e(x), y, f_e(y)>$ is indistinguishable from the distribution of $<x, f_e(x), y, z>$; x, y, z $\in_r$ Dom F and e $\in_r$ Key F.

  (Given a value x and its encryption $f_e(x)$, for a new value y, we cannot distinguish between $f_e(y)$ and a random value z. Thus we cannot encrypt y nor decrypt $f_e(y)$.)

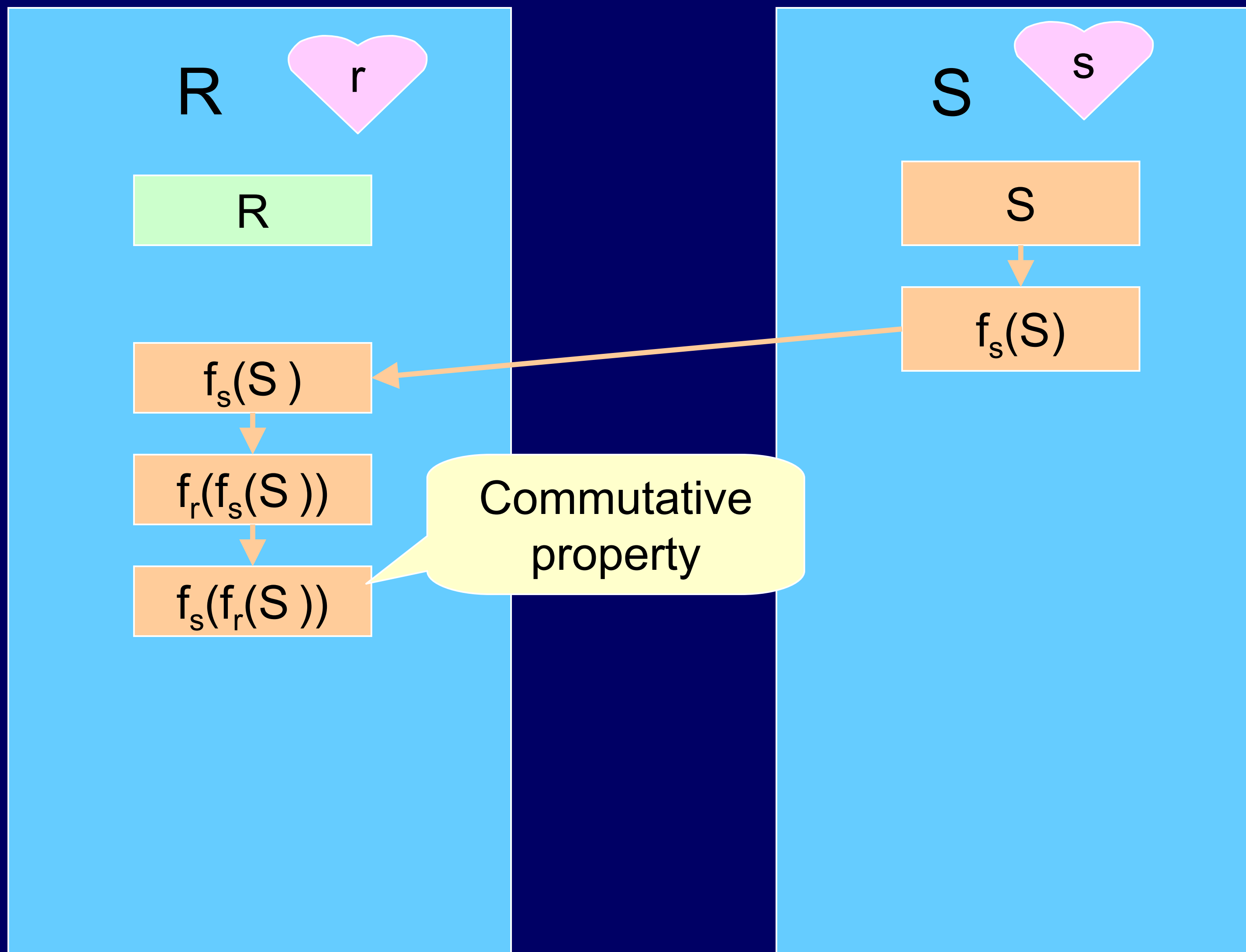
Example Commutative Encryption

- $f_e(x) = x^e \bmod p$
  where
  - p: safe prime number, i.e., both p and q=(p-1)/2 are primes
  - encryption key e ♍ 1, 2, …, q-1
  - Dom F: all quadratic residues modulo p
- Commutativity: powers commute
  $(x^d \bmod p)^e \bmod p = x^{de} \bmod p = (x^e \bmod p)^d \bmod p$
- Indistinguishability follows from Decisional Diffie-Hellman Hypothesis (DDH)

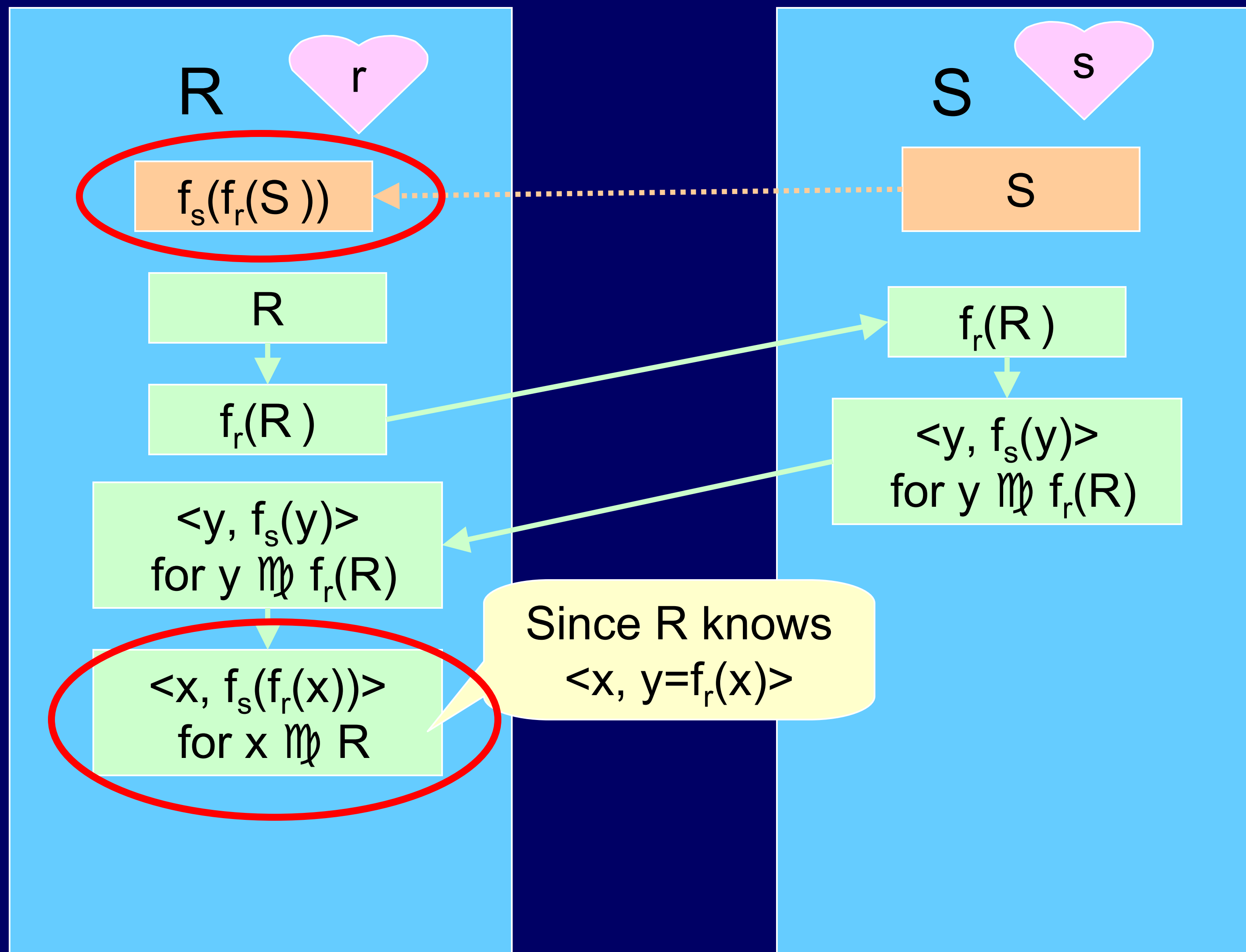
Intersection Protocol



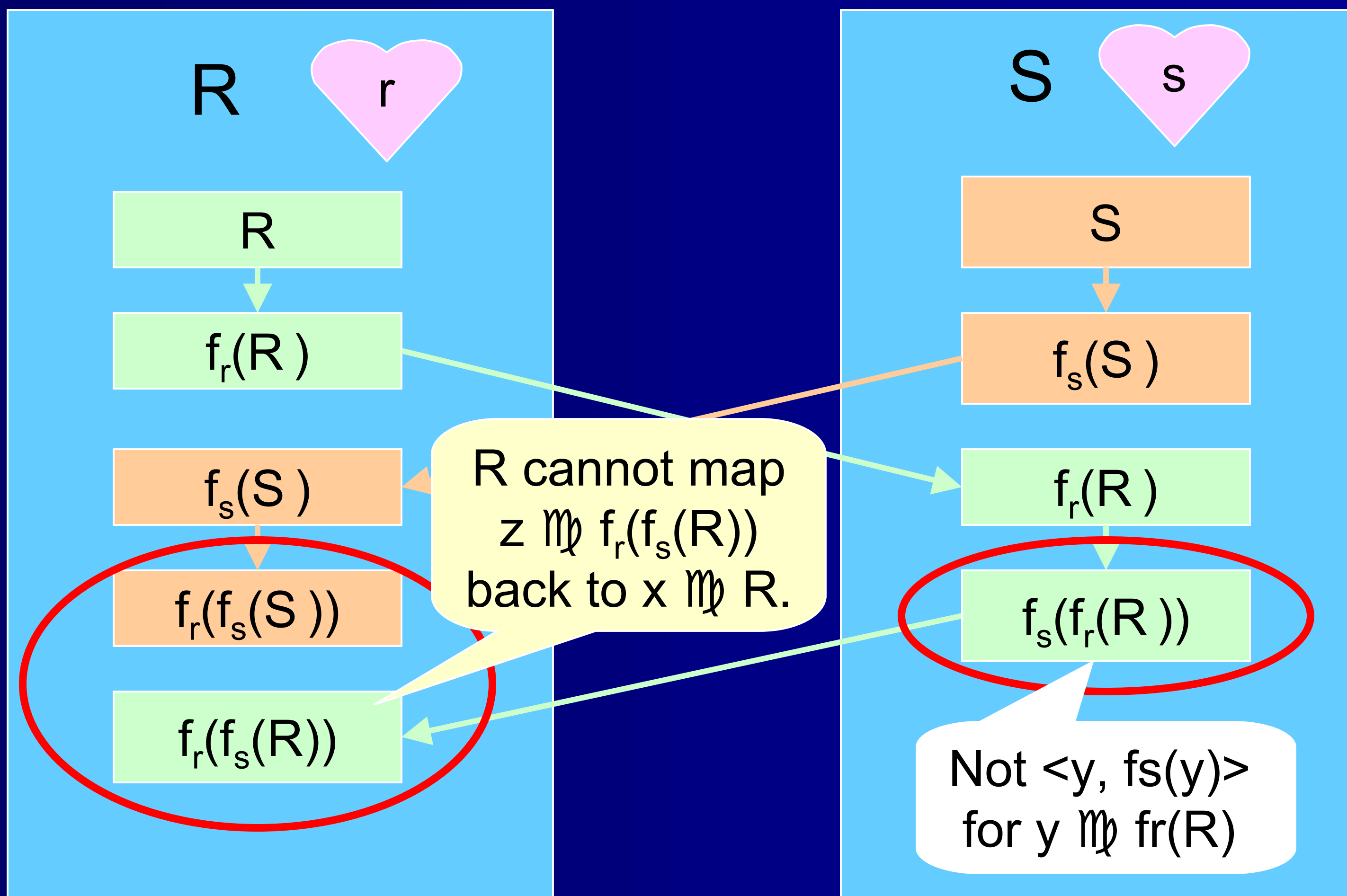
Intersection Protocol



Intersection Protocol



Intersection Size Protocol



Equi Join and Join Size

- See Sigmod03 paper
- Also gives the cost analysis of protocols

Related Work

- [NP99]: Protocols for list intersection problem
  - Oblivious evaluation of n polynomials of degree n each.
  - Oblivious evaluation of n^2 polynomials.
- [HFH99]: find people with common preferences, without revealing the preferences.
  - Intersection protocols are similar to ours, but do not provide proofs of security.

Challenges

- Models of minimal disclosure and corresponding protocols for
  - other database operations
  - combination of operations
- Faster protocols
- Tradeoff between efficiency and
  - the additional information disclosed
  - approximation

Closing Thoughts

- Solutions to complex problems such as privacy require a mix of legislations, societal norms, market forces & technology
- By advancing technology, we can change the mix and improve the overall quality of the solution
- Gold mine of challenging research problems (besides being useful)!

References

http://www.almaden.ibm.com/software/quest/

- M. Bawa, R. Bayardo, R. Agrawal. Privacy-preserving indexing of Documents on the Network. 29th Int'l Conf. on Very Large Databases (VLDB), Berlin, Sept. 2003.
- R. Agrawal, A. Evfimievski, R. Srikant. Information Sharing Across Private Databases. ACM Int'l Conf. On Management of Data (SIGMOD), San Diego, California, June 2003.
- A. Evfimievski, J. Gehrke, R. Srikant. Liming Privacy Breaches in Privacy Preserving Data Mining. PODS, San Diego, California, June 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. An Xpath Based Preference Language for P3P. 12th Int'l World Wide Web Conf. (WWW), Budapest, Hungary, May 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Implementing P3P Using Database Technology. 19th Int'l Conf.on Data Engineering(ICDE), Bangalore, India, March 2003.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Server Centric P3P. W3C Workshop on the Future of P3P, Dulles, Virginia, Nov. 2002.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. Hippocratic Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002.
- R. Agrawal, J. Kiernan. Watermarking Relational Databases. 28th Int'l Conf. on Very Large Databases (VLDB), Hong Kong, August 2002. Expanded version in VLDB Journal 2003.
- A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke. Mining Association Rules Over Privacy Preserving Data. 8th Int'l Conf. on Knowledge Discovery in Databases and Data Mining (KDD), Edmonton, Canada, July 2002.
- R. Agrawal, R. Srikant. Privacy Preserving Data Mining. ACM Int'l Conf. On Management of Data (SIGMOD), Dallas, Texas, May 2000.