



| IBM Research

Enabling the 21st Century Healthcare IT Revolution

Rakesh Agrawal, IBM Fellow
Intelligent Information Systems Research
IBM Almaden Research Center

Almaden Research Center

Based on joint work with



- Roberto Bayardo
- Alvin Cheung
- Alexandre Evfimievski
- Tyrone Grandison
- Christopher Johnson
- Jerry Kiernan
- Kristen Lefevre
- Ramakrishnan Srikant
- Yirong Xu



Thesis

- Database technology has a central role to play in addressing challenges of the 21st Century, such as healthcare and education.
- We must move our focus from managing bits to deriving value from bits.



Agenda

- Review the PITAC report on *Revolutionizing Healthcare through Information Technology*.
- Illustrate how *Hippocratic Database* technologies can help fulfill the PITAC vision.
- Outline research challenges.



Revolutionizing Healthcare Through Information Technology
President's Information Technology Advisory Committee, June 04



PITAC Framework for 21st Century Health Care Information Infrastructure

44,000-98,000 die every year from medical errors in hospitals alone

Medication errors in 1 of every 5 doses, 7% of those life threatening

17%-49% diagnostic lab tests performed because medical history and earlier test results not available



Health insurance costs risen by over 10% in each of past three years

No nation-wide monitoring to identify epidemics, patterns of adverse drug reactions, bio-terrorist incidents

**Lower Cost
Fewer Errors
Higher Quality**

PITAC Framework

Elements

Electronic Health Record

Clinical Decision Support

Computerized Provider Order Entry

Secure, Private, Interoperable Health Information Exchange

Findings and Recommendations

Economic Incentives for Investment in Healthcare IT

Health Information Exchange

Facilitating Sharing of EHR Technologies

Leveraging Federal Health IT Investments

Standardized Clinical Vocabulary

Standardized, Interoperable EHRs

The Human-Machine Interface and EHR

Coordination of Federal NHII Development

Unambiguous Patient Identification

Encrypted Internet Communications

Trust Hierarchy and Authentication

Tracing Access Requests

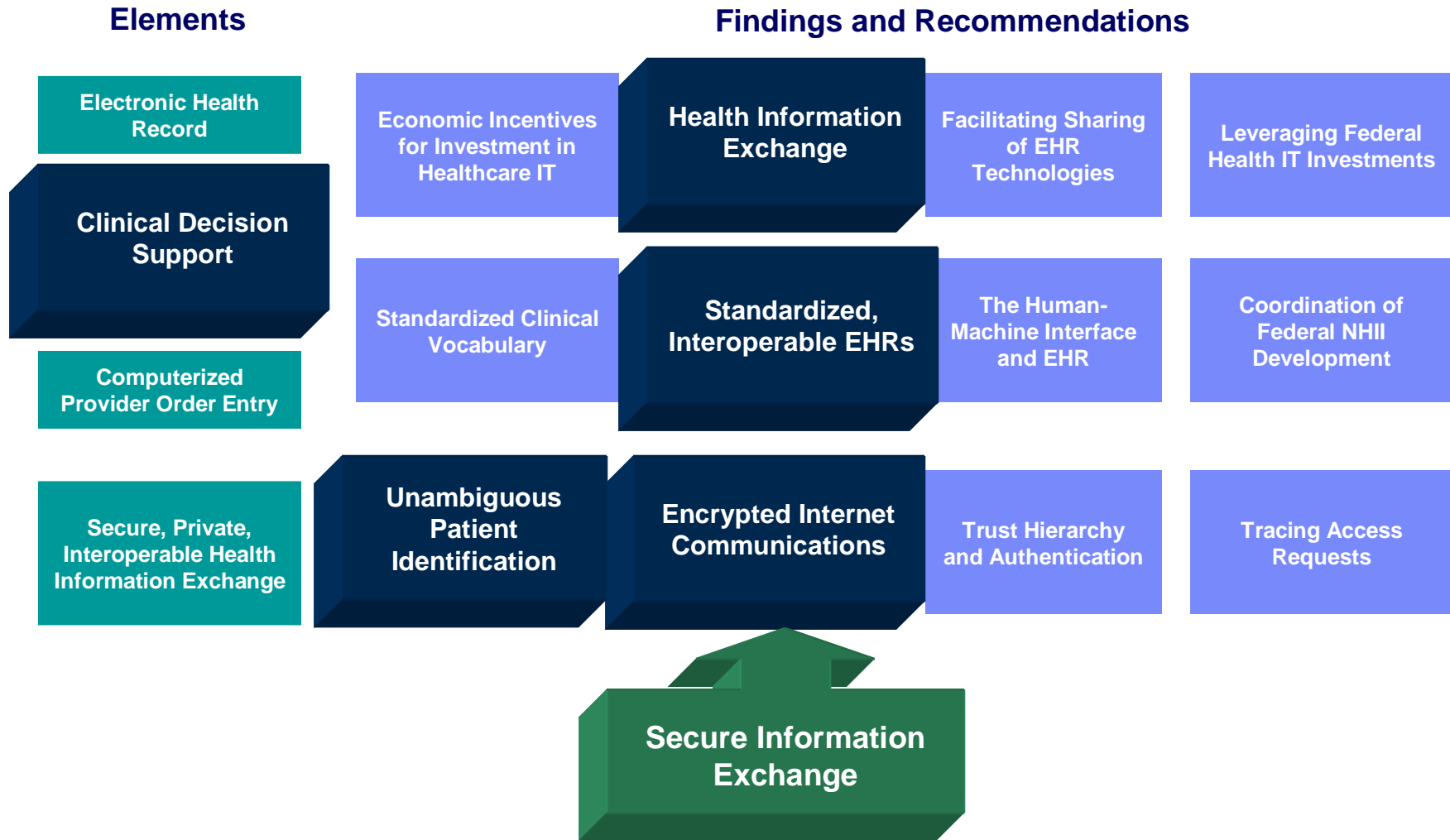
Hippocratic Database Technologies in the PITAC Framework

Elements

Findings and Recommendations



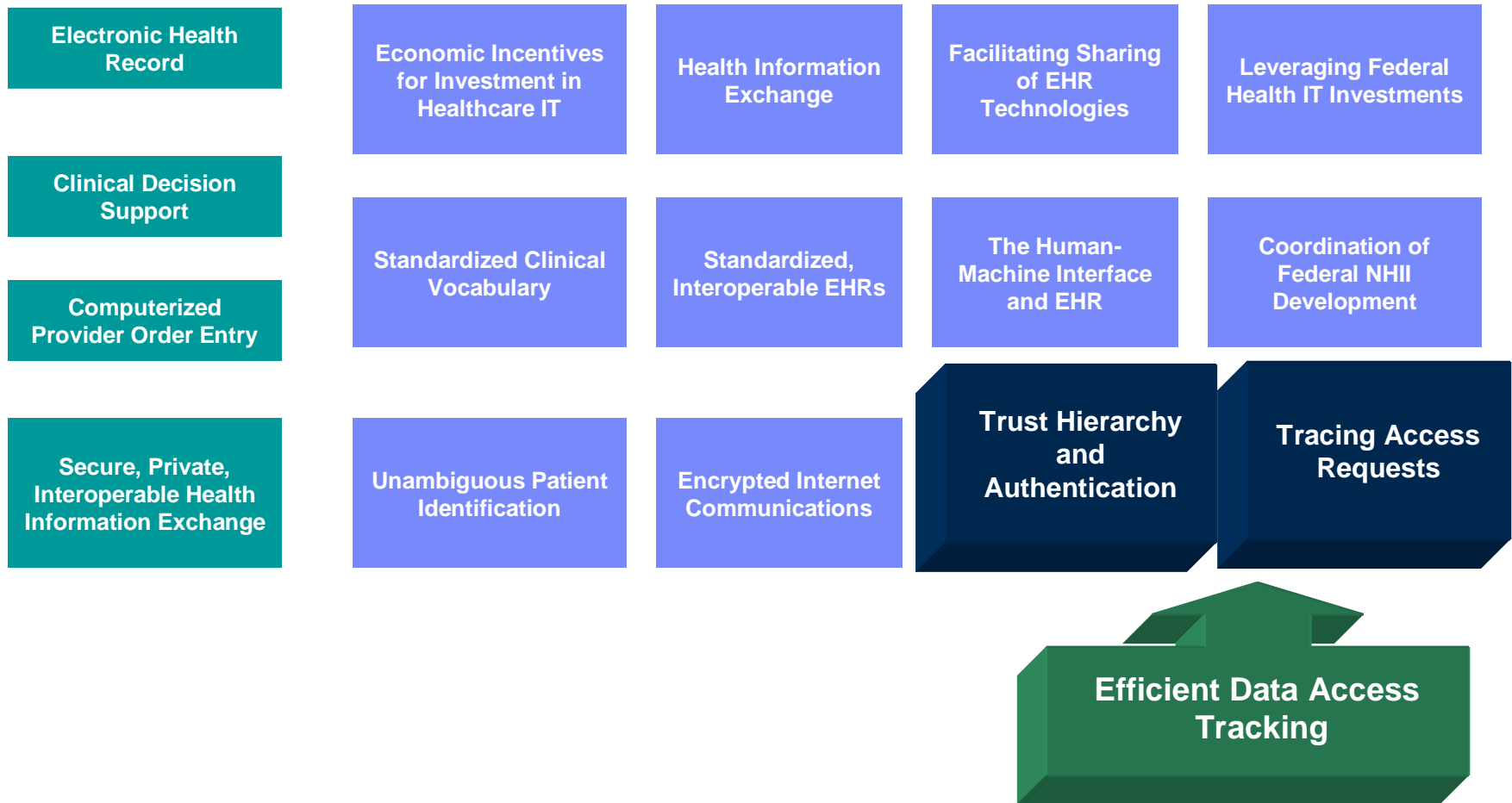
Hippocratic Database Technologies in the PITAC Framework



Hippocratic Database Technologies in the PITAC Framework

Elements

Findings and Recommendations





Hippocratic Database Technologies

Create a new generation of information systems that protect the privacy, security, and ownership of data while not impeding the flow of information.

Policy-Based Private Data Management

Active Enforcement
Database-level enforcement of disclosure policies and patient preferences

Privacy Preserving Data Mining
Preserves privacy at the individual level, while still building accurate data mining models at the aggregate level

Secure Information Exchange

Sovereign Information Sharing
Selective, minimal sharing across autonomous data sources, without trusted third party

Optimal k -anonymization
De-identifies records in a way that maintains truthful data but is not prone to data linkage attacks

Efficient Data Access Tracking

Compliance Auditing
Determine whether data has been disclosed in violation of specified policies

Database Watermarking
Tracks origin of leaked data by tracing hidden bit pattern embedded in the data

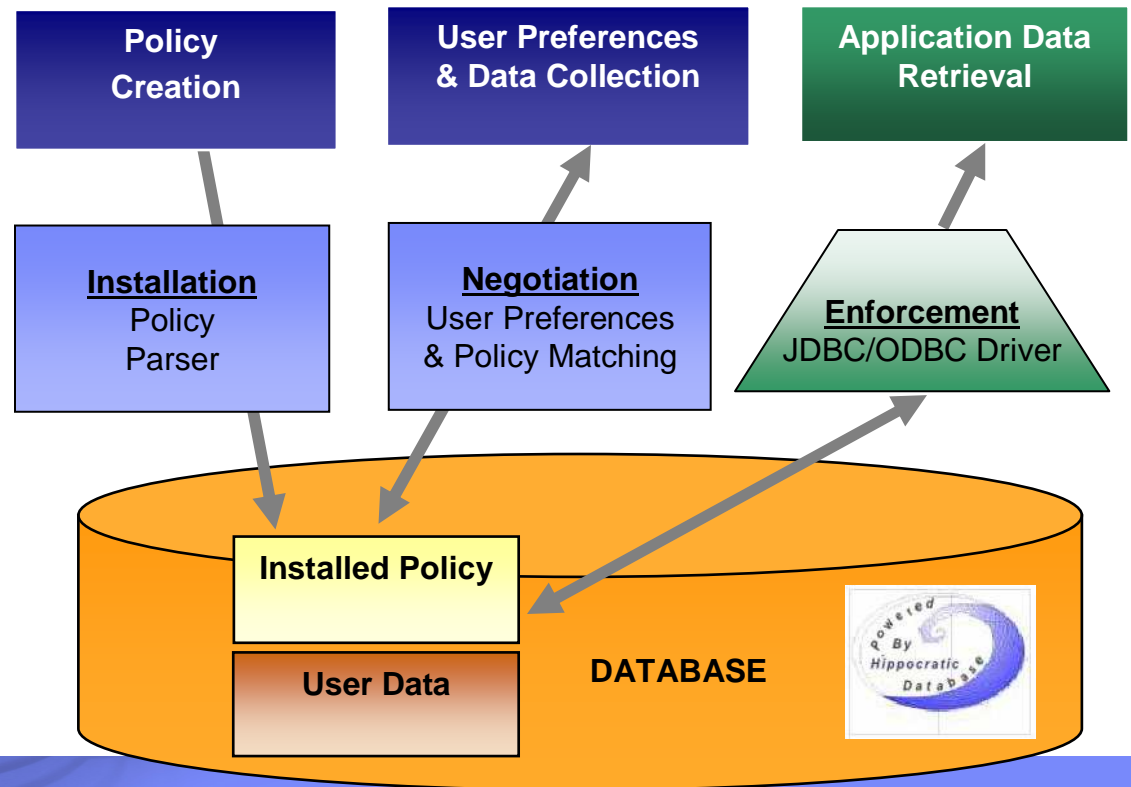
HDB Active Enforcement

- **Privacy Policy** Organizations define a set of rules describing to whom data may be disclosed (recipients) and how the data may be used (purposes)
- **Consent** Data subjects given control over who may see their personal information and under what circumstances
- **Disclosure Control** Database ensures that privacy policy and data subject consent is enforced with respect to all data access
 - Limits the outflow of information

- **Implementation intercepts and rewrites incoming queries to factor in policy, user choices, and context (e.g. purpose).**
- **Rewritten queries benefit from all the optimizations and performance enhancements provided by underlying engine (e.g. parallelism).**

- **Disclosure control at cell-level**
- **Applications do not require any modification.**
- **Database agnostic; does not require any change in the database engine.**

#	Name	Age	Phone
1	Adam	25	111-1111
3	Bob	-	333-3333
4	Daniel	40	-



VLDB 02, WWW 03, VLDB 04

Table Semantics (Informal)

Table "Patients"

Patient #	Name	Age	Address	Phone
1	Michael Bell	19	Palo Alto	111-1111
2	Natalie Lewis	22	Berkeley	222-2222
3	Robert Thorpe	23	Cambridge	333-3333
4	Jenny Thompson	31	New York	444-4444

#	Patient#	Name	Age	Address	Phone
1	√	√	√	√	√
2	X	X	X	X	X
3	√	X	X	√	√
4	√	√	X	X	X

Mask prohibited cells with null

Patient#	Name	Age	Address	Phone
1	Michael Bell	19	Palo Alto	111-1111
3			Cambridge	333-3333
4	Jenny Thompson			

Filter rows where the primary key is prohibited

Patient#	Name	Age	Address	Phone
1	Michael Bell	19	Palo Alto	111-1111
3			Cambridge	333-3333
4	Jenny Thompson			

Query Semantics Enforcement

Mask prohibited cells with null

Patient#	Name	Age	Address	Phone
1	Michael Bell	19	Palo Alto	111-1111
3			Cambridge	333-3333
4	Jenny Thompson			

Issue Query:
SELECT Name, Age
FROM Patients

Name	Age
Michael Bell	19
Jenny Thompson	

Filter rows that are entirely null from result set

Name	Age
Michael Bell	19
Jenny Thompson	

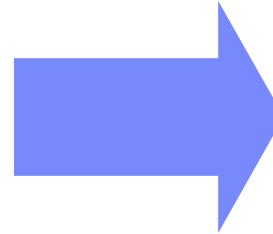
Query Semantics

Name	Age
Michael Bell	19
Jenny Thompson	

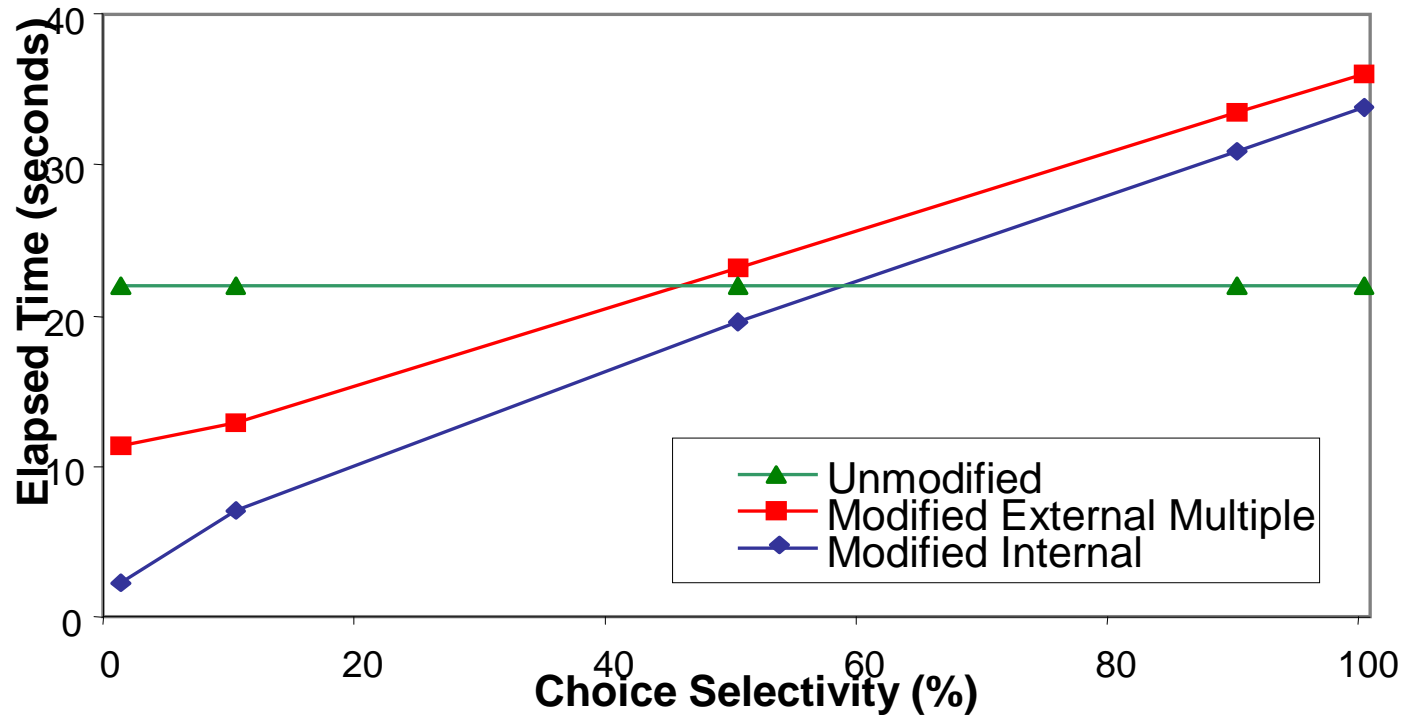
Table Semantics

Query Modification Example (Table Semantics)

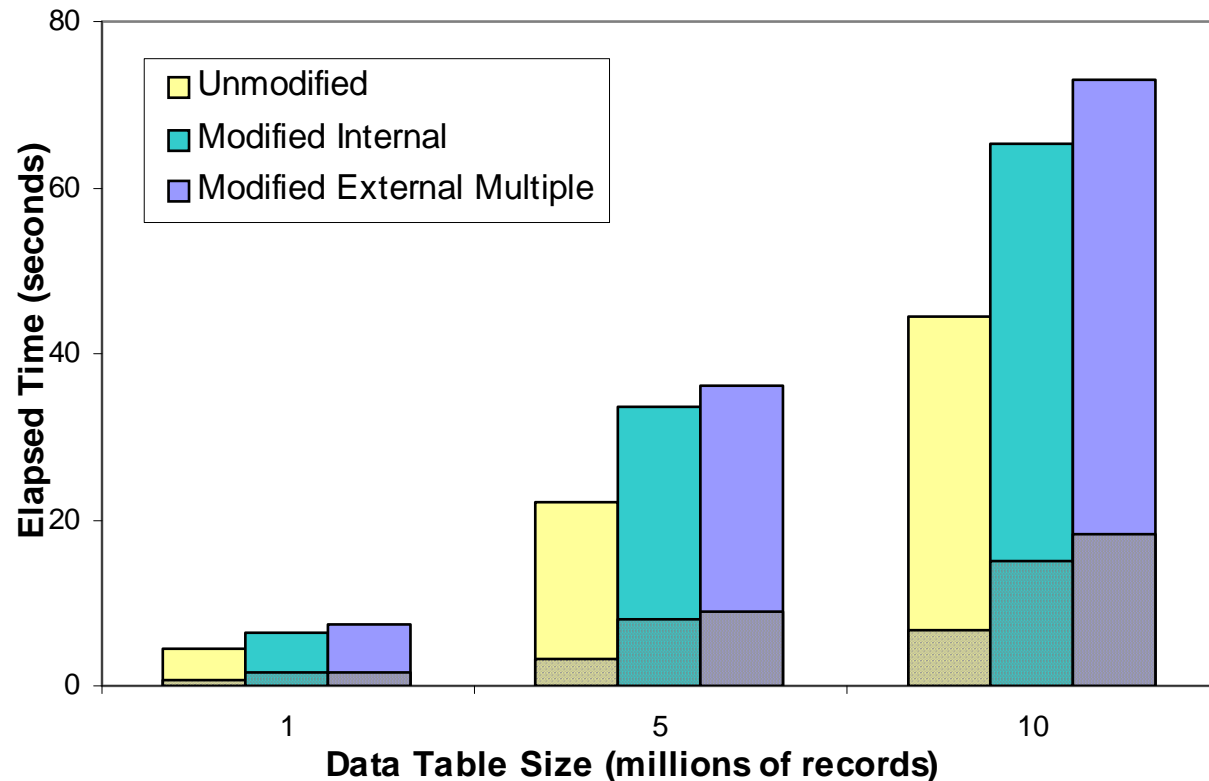
```
SELECT Name  
FROM Patients  
WHERE Age < 20
```



```
SELECT  
CASE WHEN EXISTS  
  (SELECT Name_Choice  
   FROM Patient_Choices  
   WHERE Patients.Patient# = Patient_Choices.Patient#  
   AND Patient_Choices.Name_Choice = 1)  
THEN Name ELSE null END  
FROM Patients  
WHERE Age < 20  
AND EXISTS  
  (SELECT Patient#_Choice  
   FROM Patient_Choices  
   WHERE Patients.Patient# = Patient_Choices.Patient#  
   AND Patient_Choices.Patient#_Choice = 1)
```



- Measured performance of a query selecting all records from a 5 million-record table
- Compared performance of original and modified queries for varied choice selectivity
- Not surprisingly, performance actually better for modified queries when we use privacy enforcement as an additional selection condition
 - Able to use indexes on choice values
- Shows the importance of database-level privacy enforcement for performance



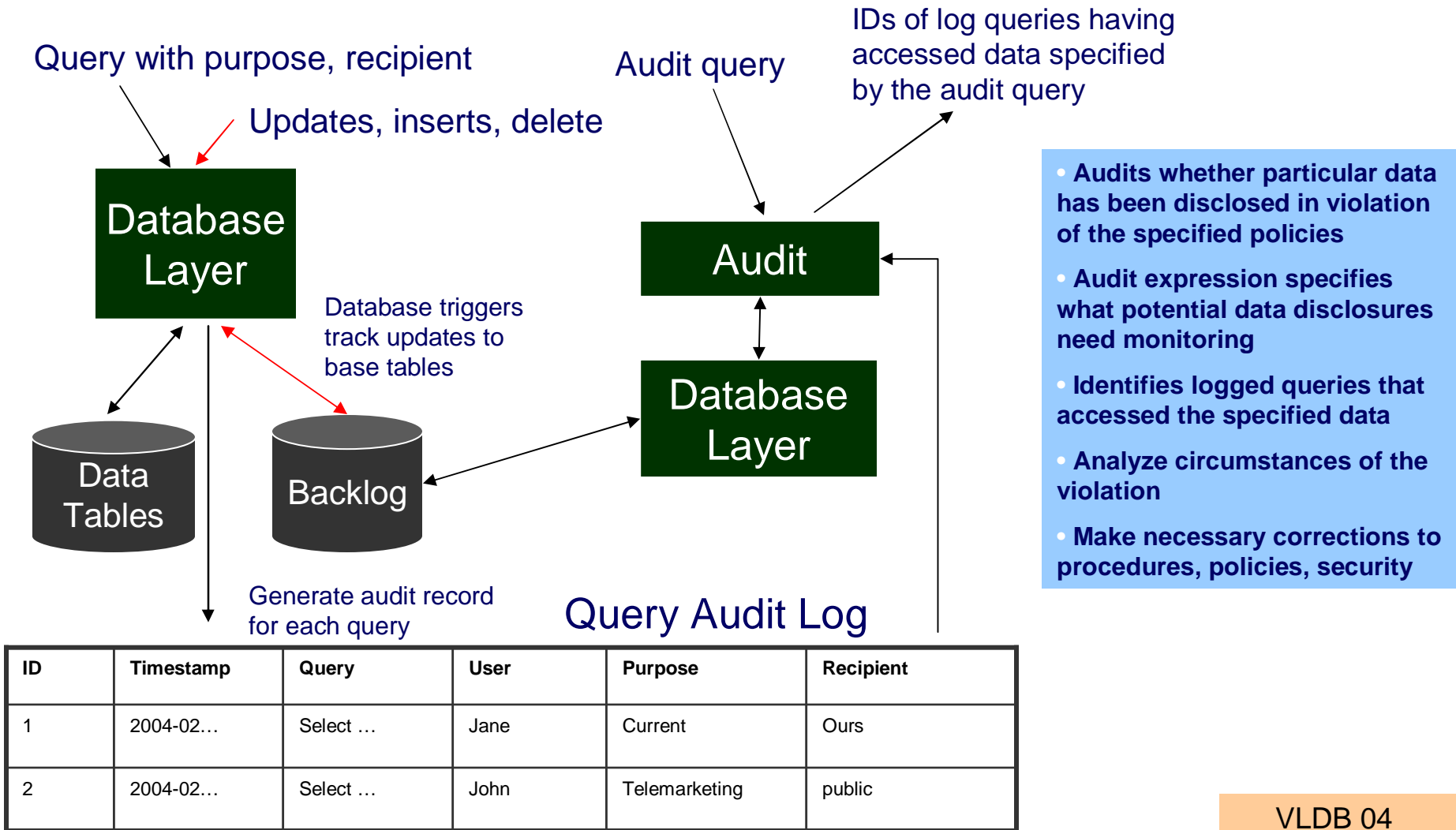
- Measured overhead cost using a query that selects all records
- Choice selectivity = 100%
 - Observed worst-case scenario where no rows are filtered due to privacy constraints, but incur all costs of cell-level checking
- Full bar represents elapsed time
- Bottom portion of bar is CPU time
- Much of the cost of privacy enforcement is CPU cost, so scales well as queries become more I/O intensive

Summary (Active Enforcement)

- Limited Disclosure is a necessary component of a comprehensive data privacy management system
- Hippocratic database technology provides a framework for automatically limiting disclosure at the database level
 - More efficient and flexible than application-level disclosure control
 - Techniques also have broader use for other applications requiring policy-driven fine-grained disclosure control
- Framework can be deployed to an existing environment with minimal modification to legacy applications
- Query modification and consent storage approaches efficient enough to be viable in practice



HDB Compliance Auditing



- Audits whether particular data has been disclosed in violation of the specified policies
- Audit expression specifies what potential data disclosures need monitoring
- Identifies logged queries that accessed the specified data
- Analyze circumstances of the violation
- Make necessary corrections to procedures, policies, security

Audit Scenario

The doctor must now review

Sometime later, Jane

The doctor uncovers that Jane's blood sugar level is high and suspects diabetes

ph... ear... take

company, proposing over

Jane com... the counter diabetes... Health and Human

Services... tests

sharing he... companies for

Jane has not been feeling well and decides to consult her doctor



Audit Expression

Who has accessed Jane's disease information?

audit T.disease
from Customer C, Treatment T
where C.cid=T.pcid **and** C.name = 'Jane'

Problem Statement

- Given
 - A log of queries executed over a database
 - An audit expression specifying sensitive data

- Precisely identify
 - Those queries that accessed the data specified by the audit expression

Definitions (Informal)

- “Candidate” query
 - Logged query that accesses all columns specified by the audit expression
- “Indispensable” tuple (for a query)
 - A tuple whose omission makes a difference to the result of a query
- “Suspicious” query
 - A candidate query that shares an indispensable tuple with the audit expression

Example:

Query Q : Addresses of people with diabetes
Audit A : Jane’s diagnosis

Jane’s tuple is indispensable for both; hence query Q is “suspicious” with respect to A

Suspicious SPJ Query

The candidate SPJ query Q and the audit expression A are of the form:

$$Q = \bar{\pi}_{CoQ}(\sigma_{P_Q}(T \times R))$$

$$A = \bar{\pi}_{CoA}(\sigma_{P_A}(T \times S))$$

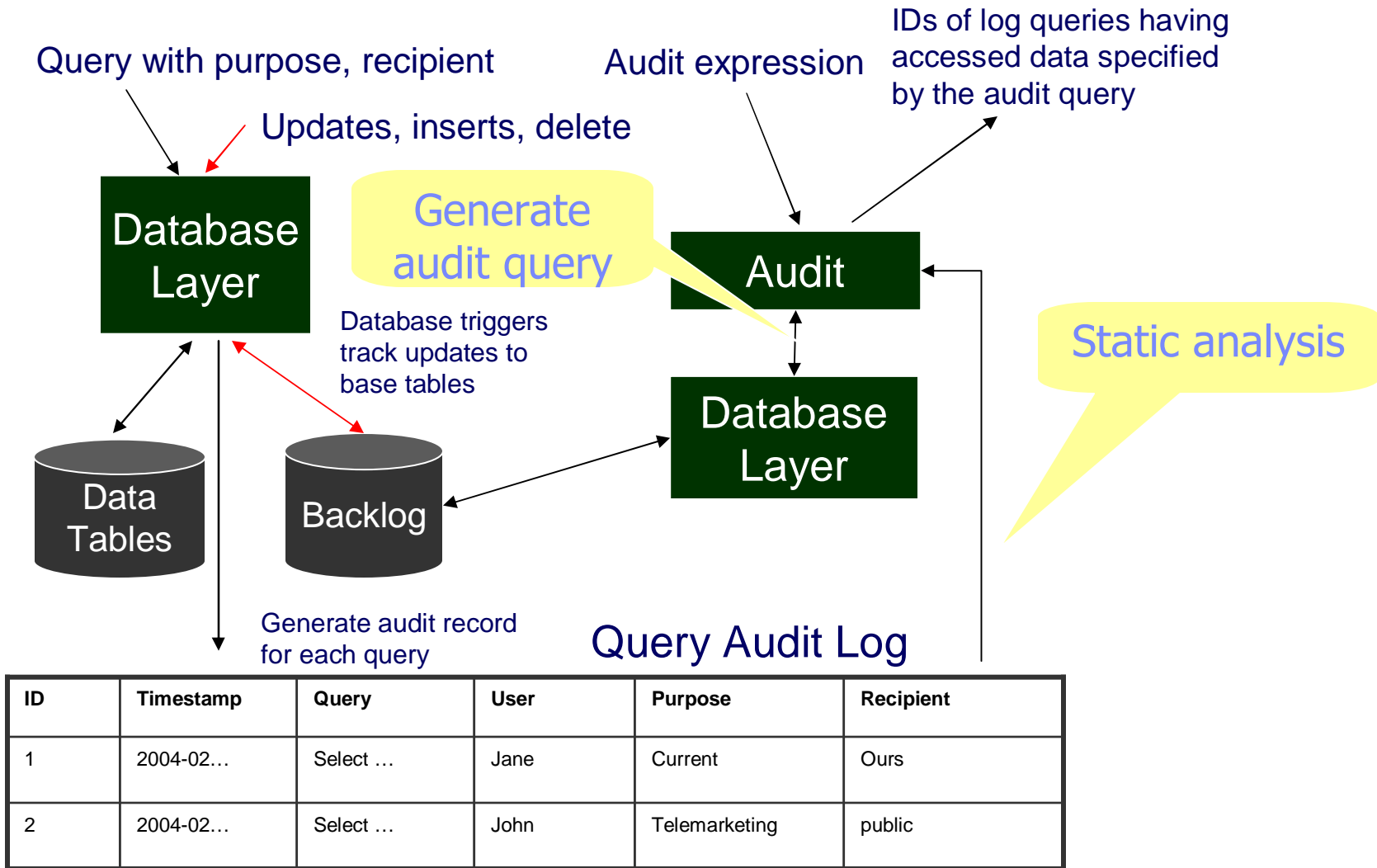
Theorem - A candidate SPJ query Q is suspicious with respect to an audit expression A iff:

$$\sigma_{P_A}(\sigma_{P_Q}(T \times R \times S)) \neq \emptyset$$

QGM rewrites Q and A into:

$$\pi^{'' Q_i''}(\sigma_{P_A}(\sigma_{P_Q}(T \times R) \times S))$$

System Overview



Static Analysis

Query Log

ID	Timestamp	Query	User	Purpose	Recipient
1	2004-02...	Select ...	James	Current	Ours
2	2004-02...	Select ...	John	Telemarketing	public

Audit expression

Accomplished by examining only the queries themselves (i.e., without running the queries)

Filter Queries

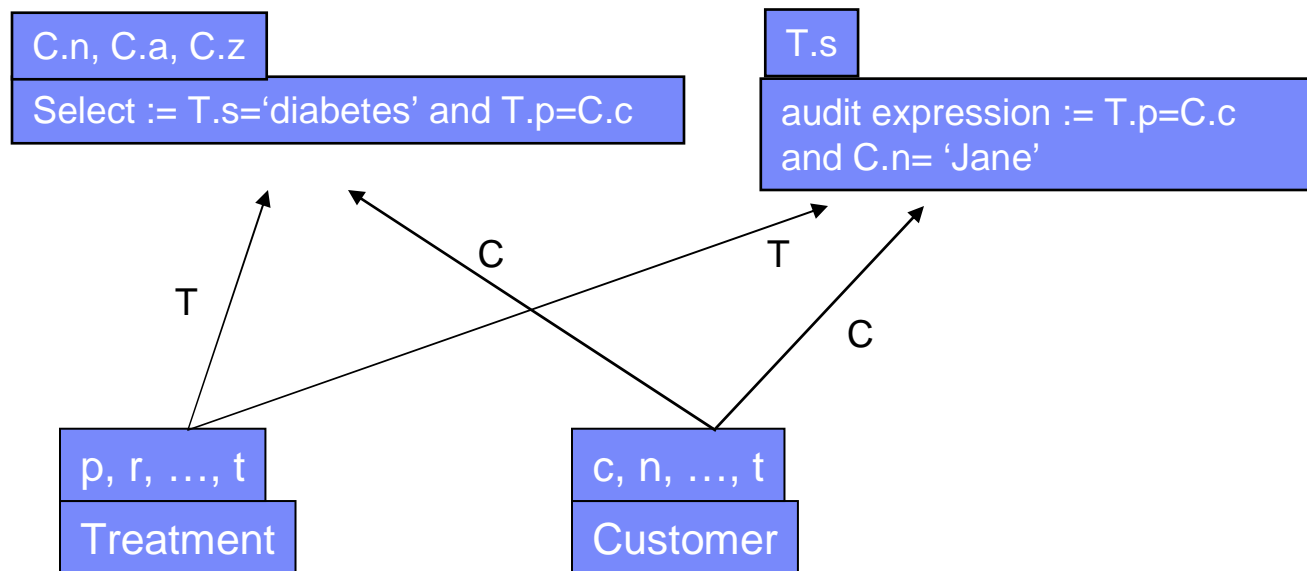
Eliminate queries that could not possibly have violated the audit expression

$$C_Q \supseteq C_{oA}$$

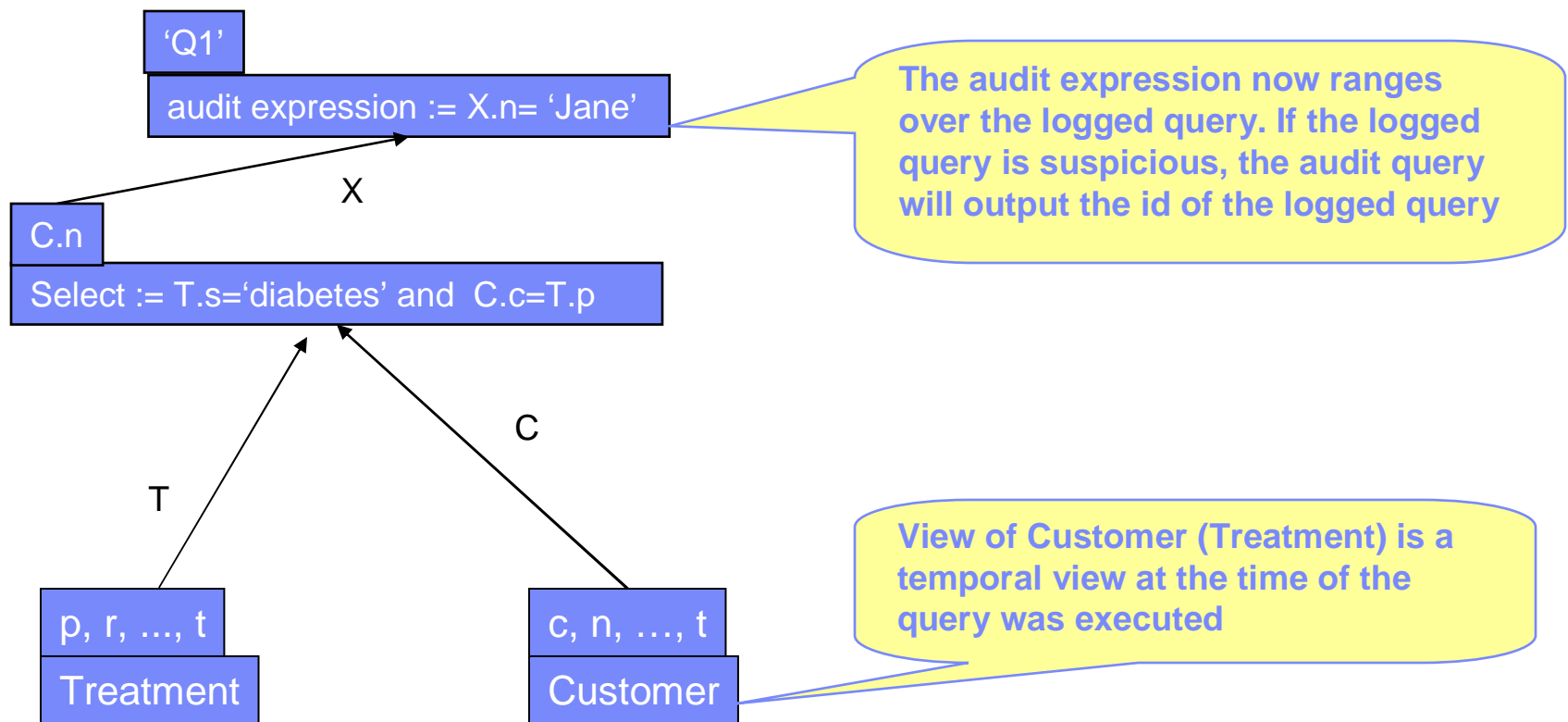
Candidate queries

Merge Logged Queries and Audit Expression

Merge logged queries and audit expression into a single query graph



Transform Query Graph into an Audit Query



Suspicious SPJ Query

The candidate SPJ query Q and the audit expression A are of the form:

$$Q = \bar{\pi}_{CoQ}(\sigma_{P_Q}(T \times R))$$

$$A = \bar{\pi}_{CoA}(\sigma_{P_A}(T \times S))$$

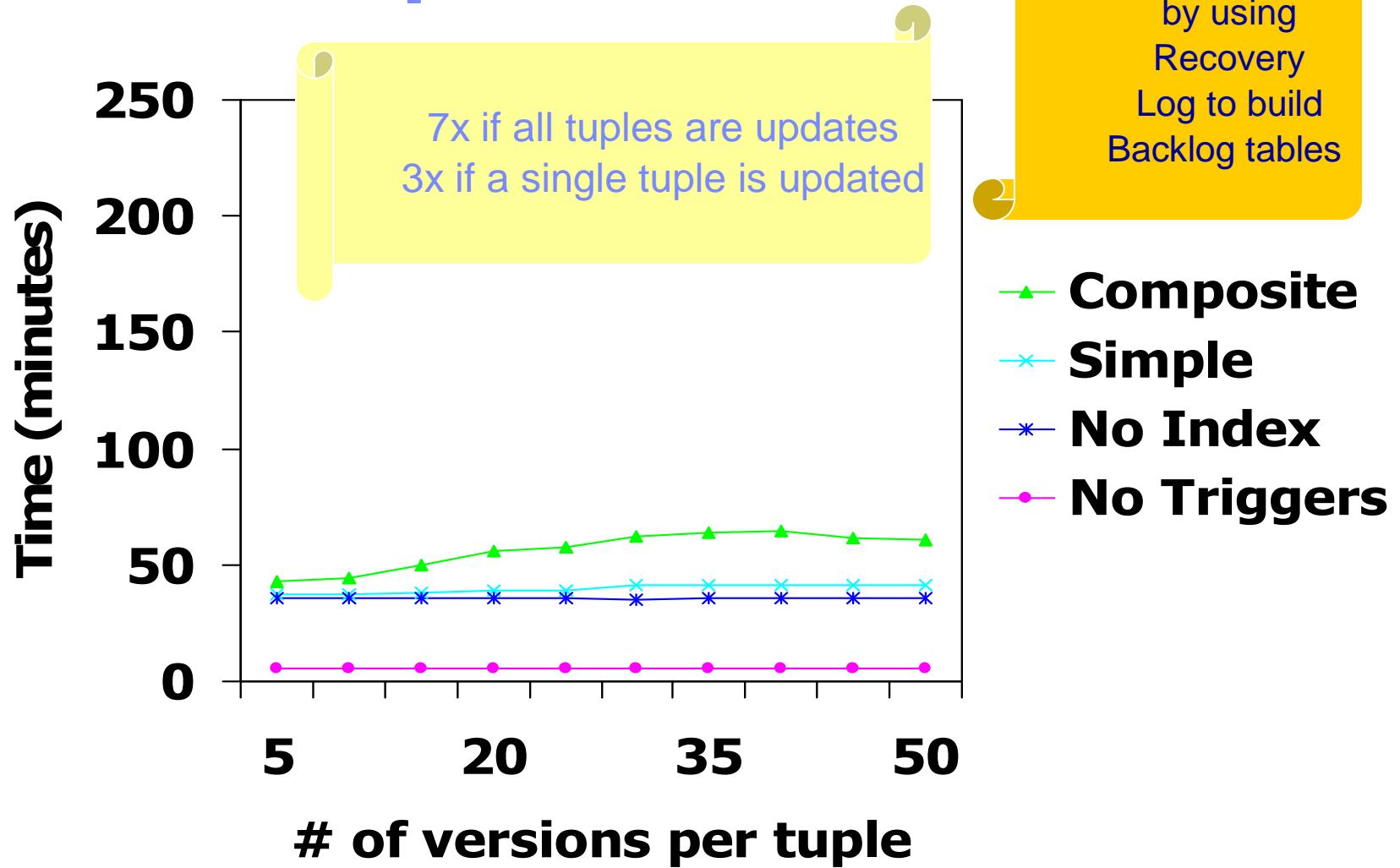
Theorem - A candidate SPJ query Q is suspicious with respect to an audit expression A iff:

$$\sigma_{P_A}(\sigma_{P_Q}(T \times R \times S)) \neq \emptyset$$

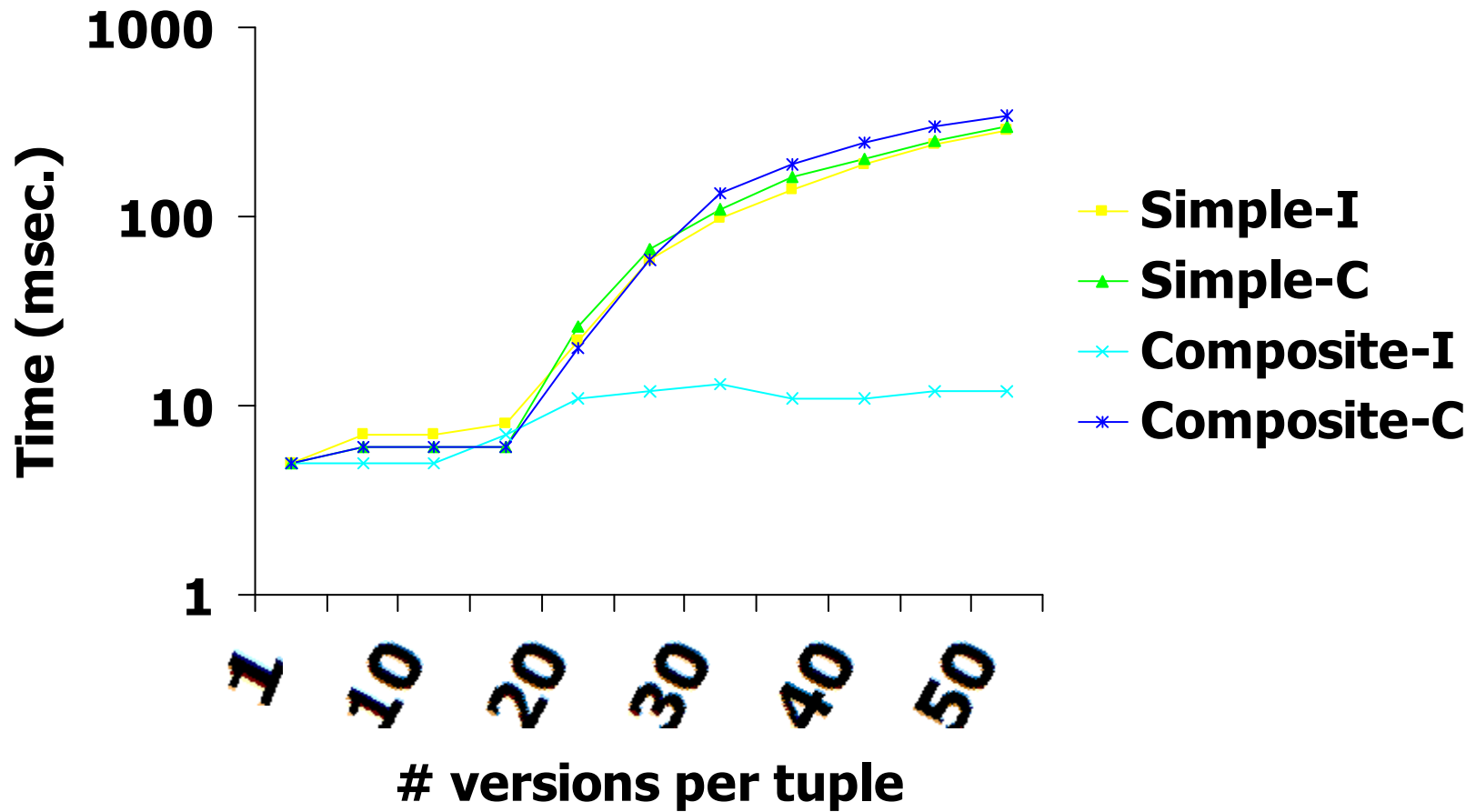
QGM rewrites Q and A into:

$$\pi^{Q_i}(\sigma_{P_A}(\sigma_{P_Q}(T \times R) \times S))$$

Overhead on Updates



Audit Query Execution Time



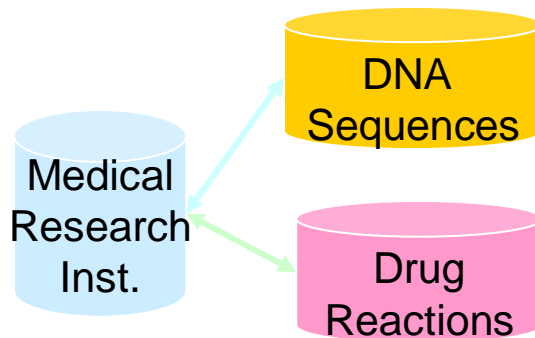
Summary (Compliance Auditing)

- Fast and precise audits (including *reads*)
- Non disruptive
 - Minimal performance impact on normal operations
- Fine grained



HDB Sovereign Information Sharing

- **Separate databases due to statutory, competitive, or security reasons.**
 - Selective, minimal sharing on need-to-know basis.
- **Example: Among those who took a particular drug, how many had adverse reaction and their DNA contains a specific sequence?**
 - Researchers must not learn anything beyond counts.
- **Algorithms for computing joins and join counts while revealing minimal additional information.**



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.

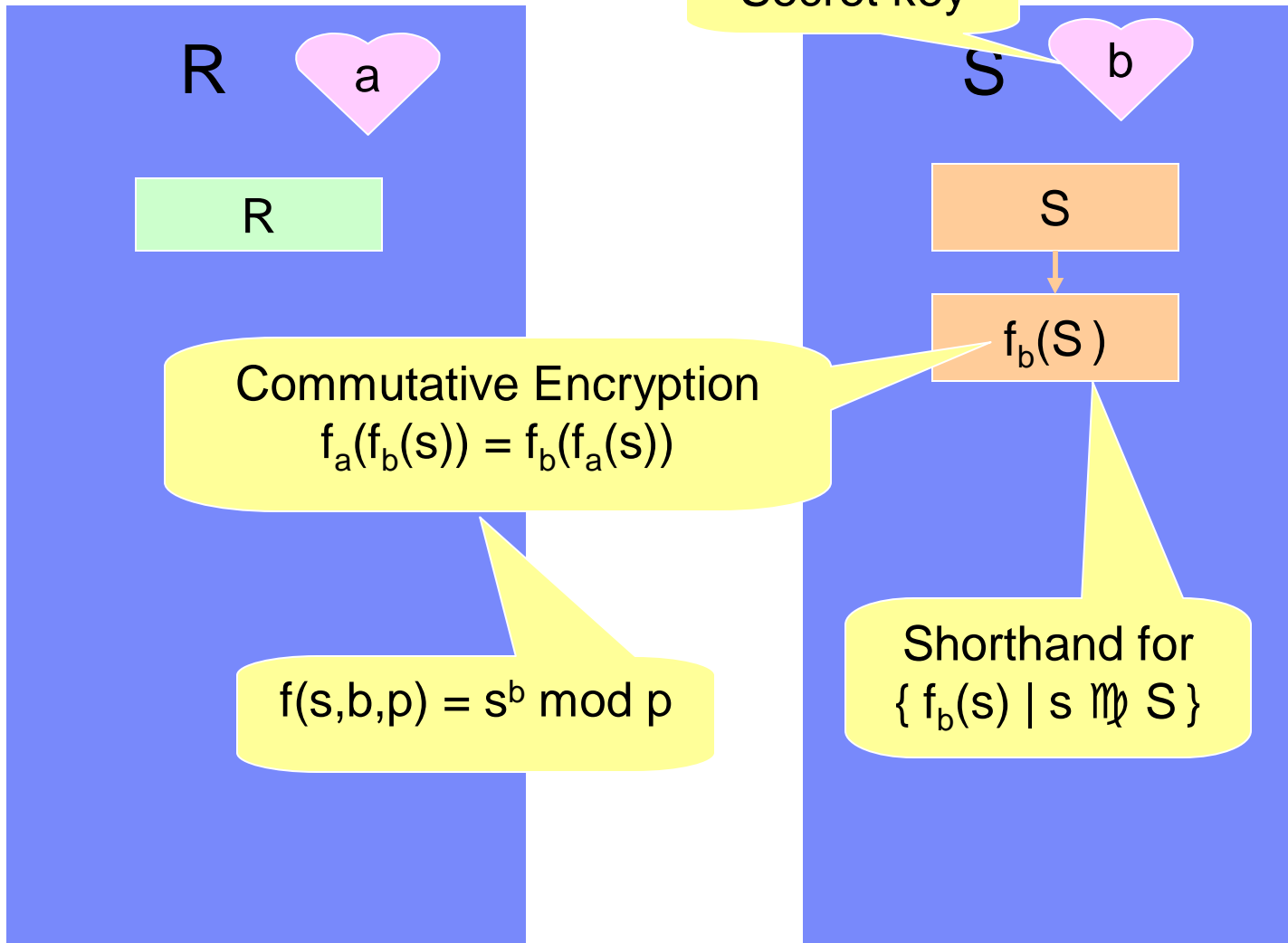
Sigmod 03, DIVO 04

Problem Statement: Minimal Sharing

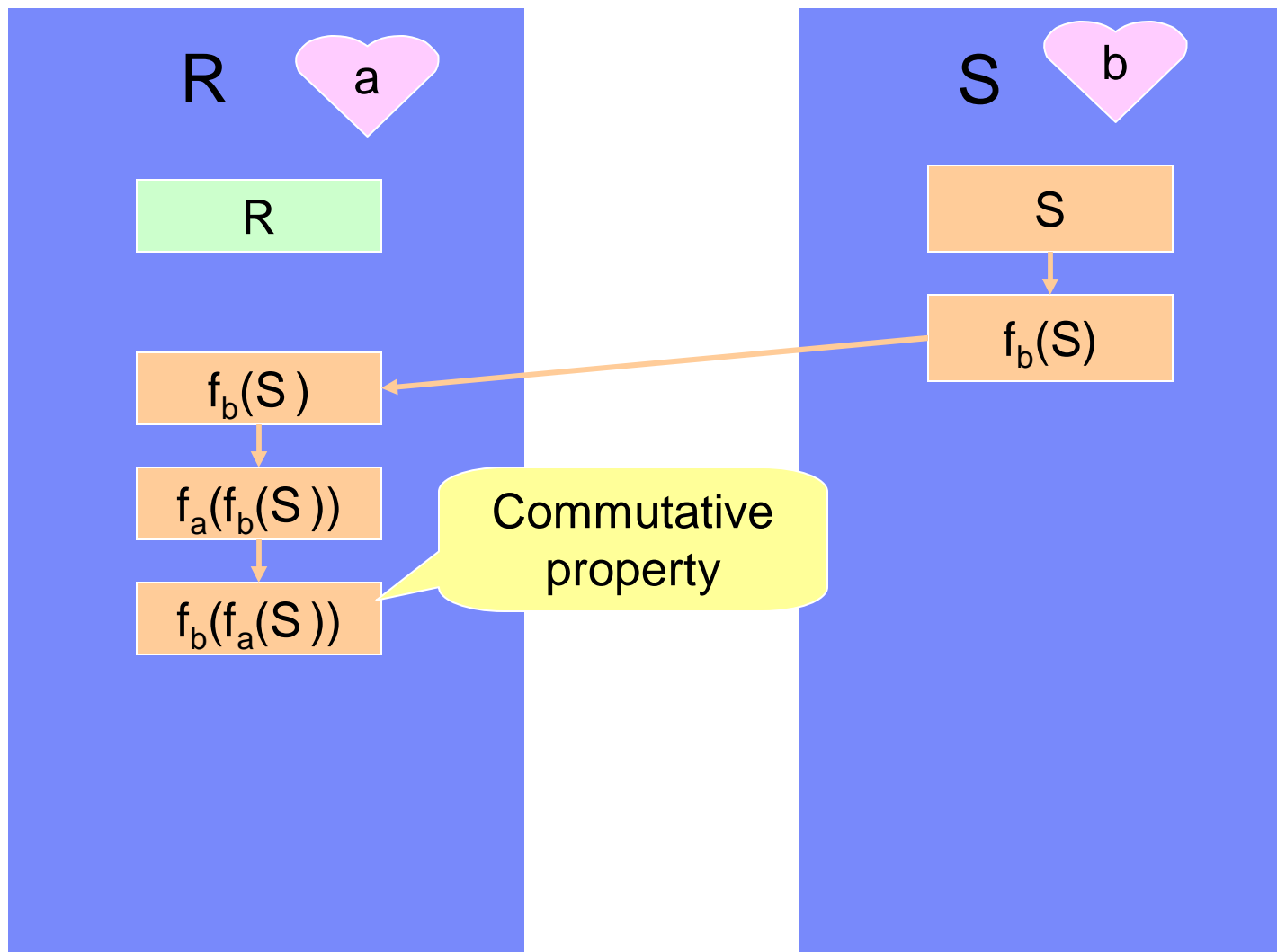
- Given:
 - Two parties (honest-but-curious): R (receiver) and S (sender)
 - Query Q spanning the tables R and S
 - *Additional (pre-specified) categories of information I*

- 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 example, in the upcoming intersection protocols
 $I = \{ |R| , |S| \}$

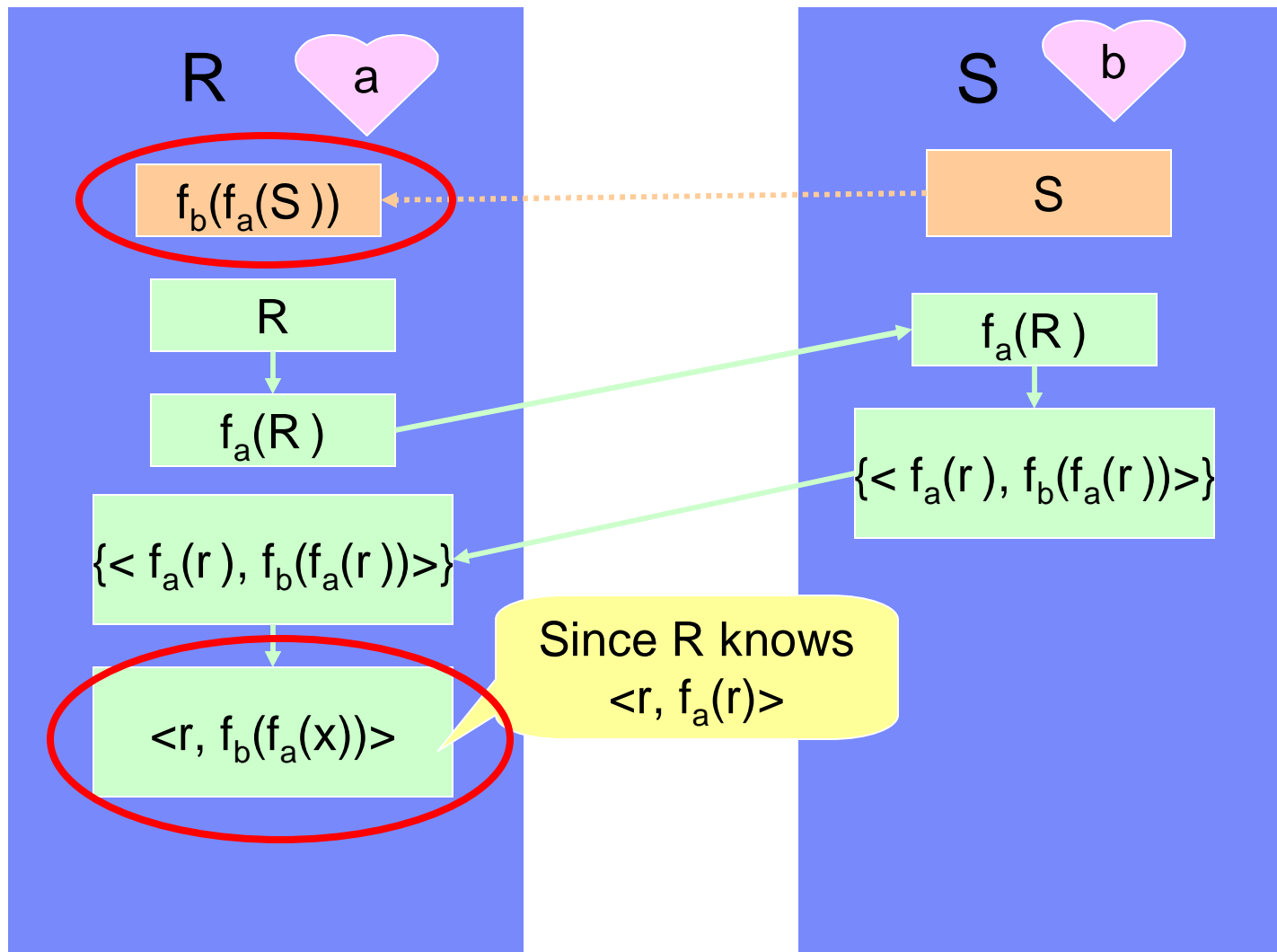
Intersection Protocol



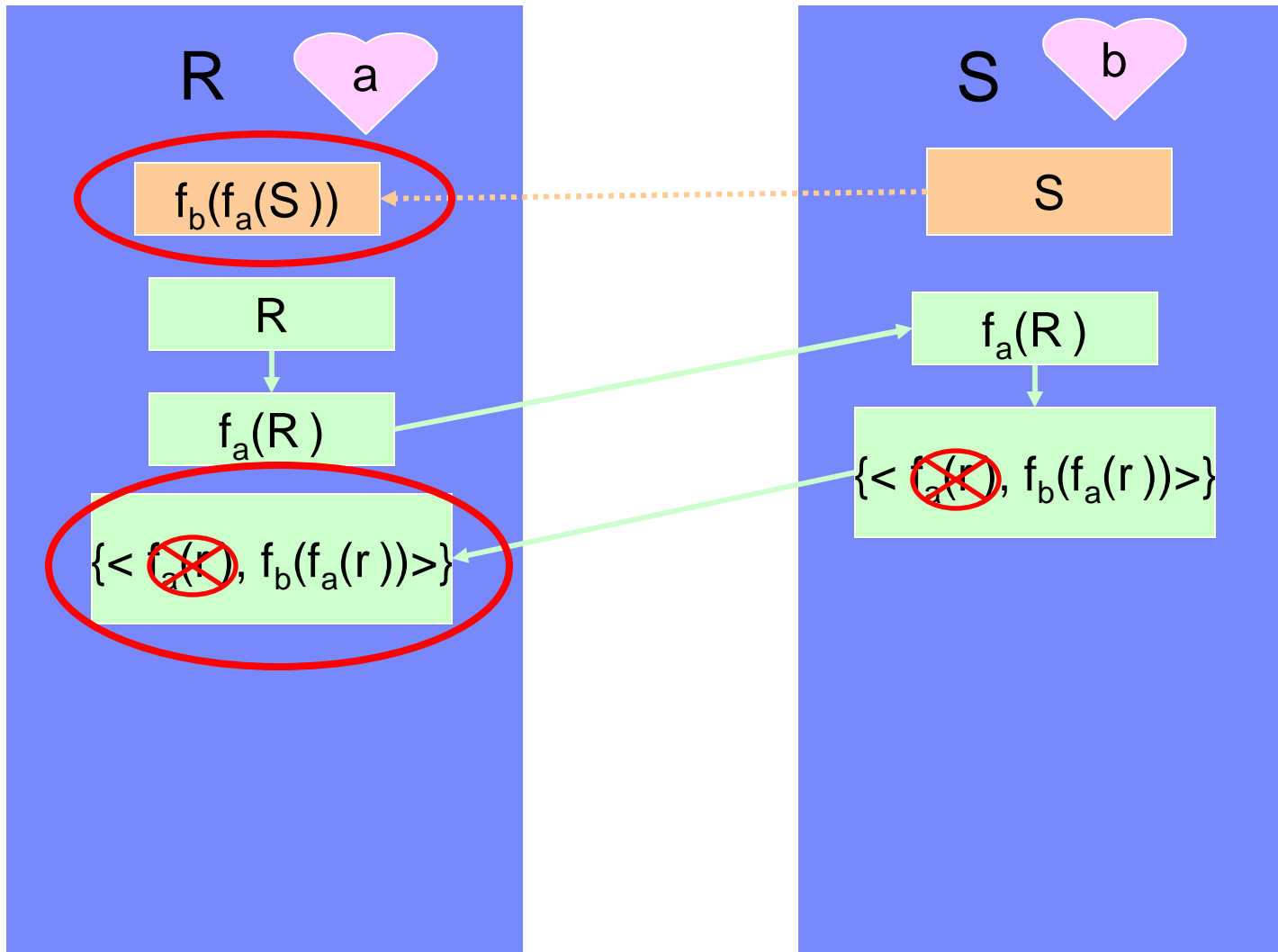
Intersection Protocol



Intersection Protocol

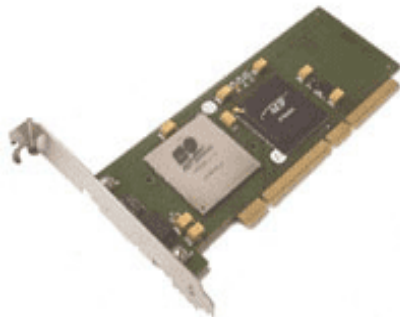


Intersection Size



Performance

- **Airline application: 150,000 (daily) passengers and 1 million people in the watch list:**
 - 120 minutes with one accelerator card
 - 12 minutes with ten accelerator cards
- **Epidemiological research: 1 million patient records in the hospital and 10 million records in the Genbank:**
 - 37 hours with one accelerator cards
 - 3.7 hours with ten accelerator cards



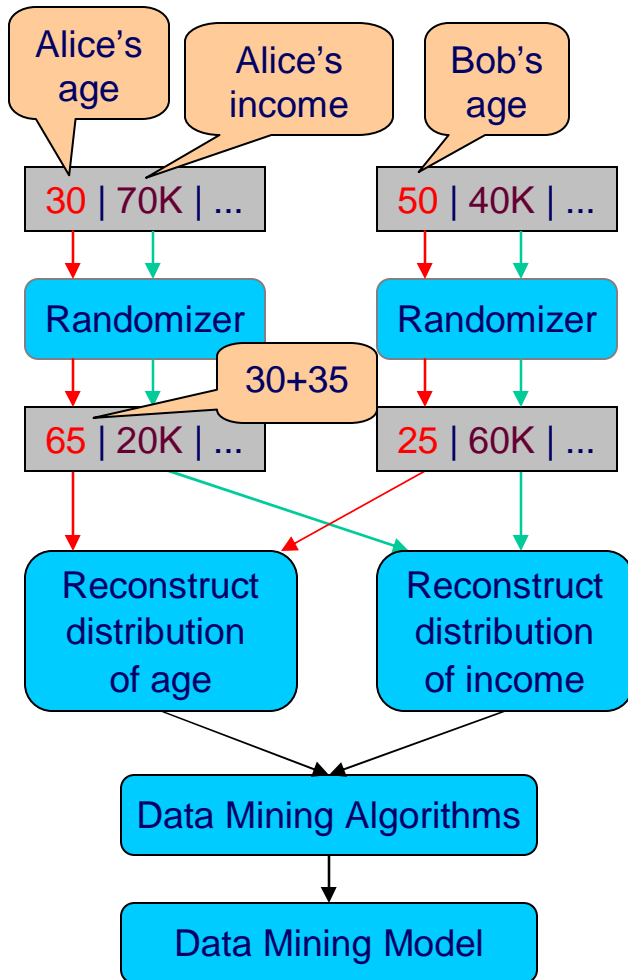
AEP SSL CARD Runner 2000 ≈ \$2K
20K encryptions per minute
10x improvement over software implementation

Summary (Sovereign Information Integration)

- New applications require us to go beyond traditional Centralized and Federated information integration:
Sovereign Information Integration
- Need further study of tradeoff between efficiency and
 - information disclosed
 - approximation

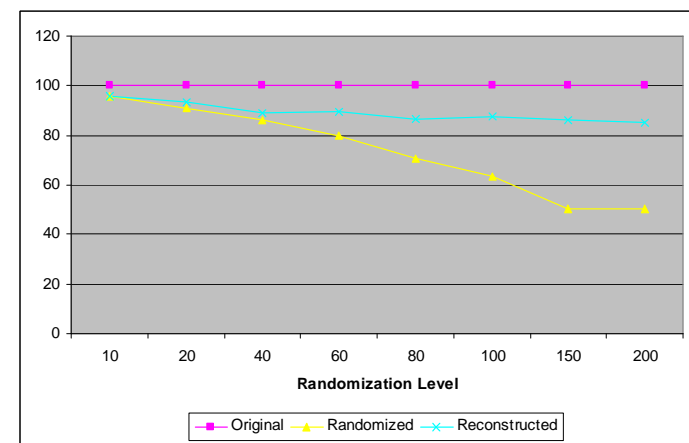
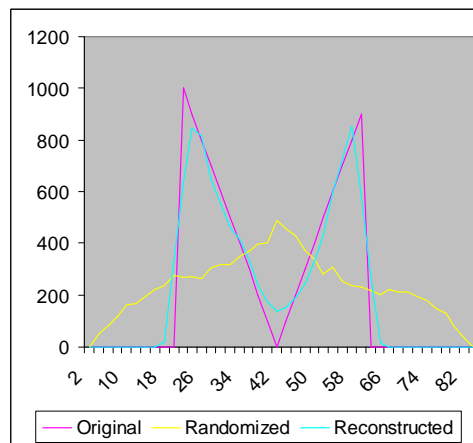


HDB Privacy Preserving Data Mining



Sigmod00, KDD02, Sigmod05

- Insight: Preserve privacy at the individual level, while still building accurate data mining models at the aggregate level.
- Add random noise to individual values to protect privacy.
- EM algorithm to estimate original distribution of values given randomized values + randomization function.
- Algorithms for building classification models and discovering association rules on top of privacy-preserved data with only small loss of accuracy.



Problem Statement (Numeric Data)

- To hide original values x_1, x_2, \dots, x_n
 - from probability distribution X (unknown)

we use y_1, y_2, \dots, y_n

- from probability distribution Y
- Problem: Given
 - $x_1+y_1, x_2+y_2, \dots, x_n+y_n$
 - the probability distribution of Y

Estimate the probability distribution of X .

Reconstruction Algorithm

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

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

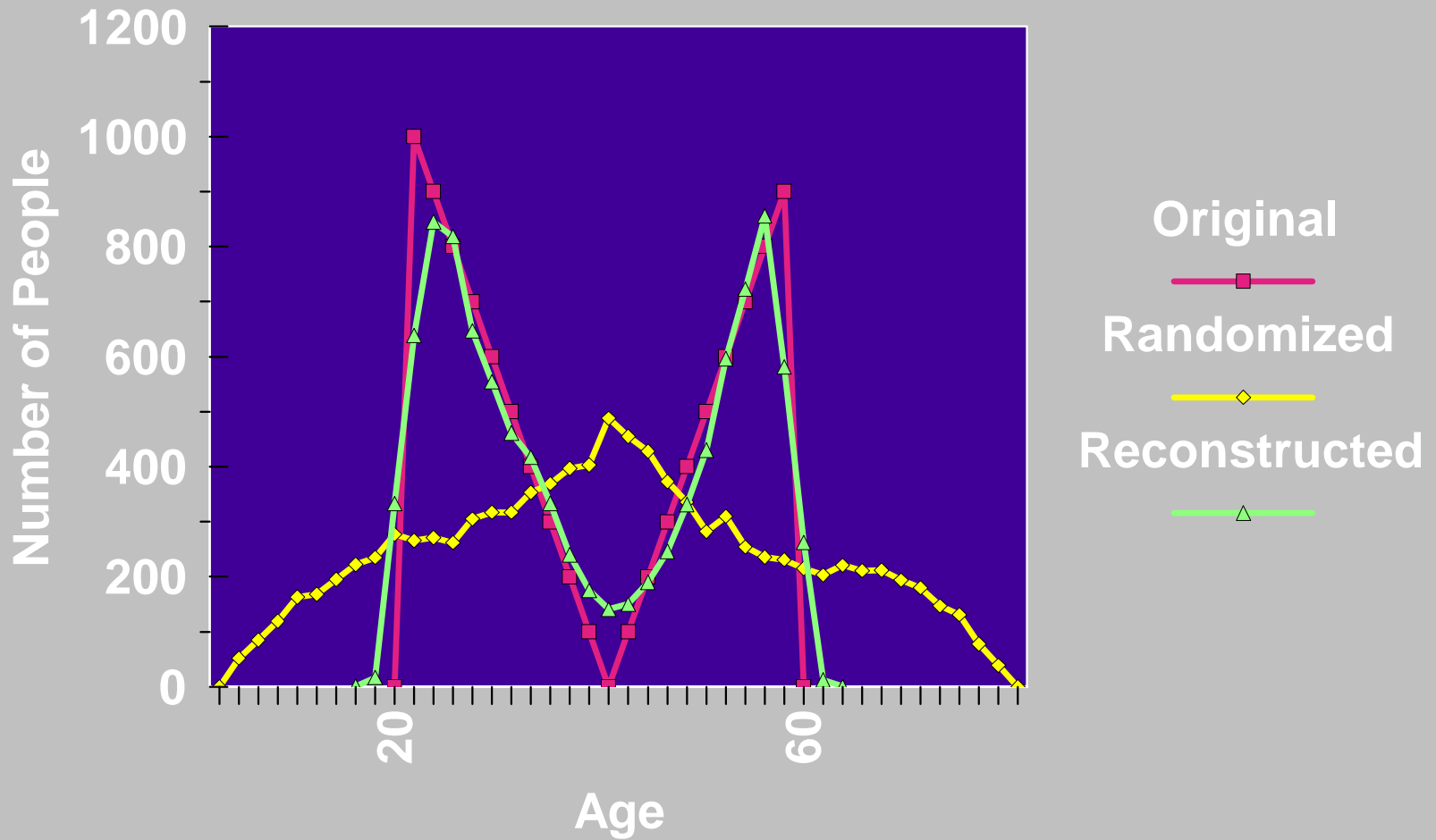
$j := j+1$

until (stopping criterion met)

(R. Agrawal, R. Srikant. *Privacy Preserving Data Mining*. SIGMOD 2000)

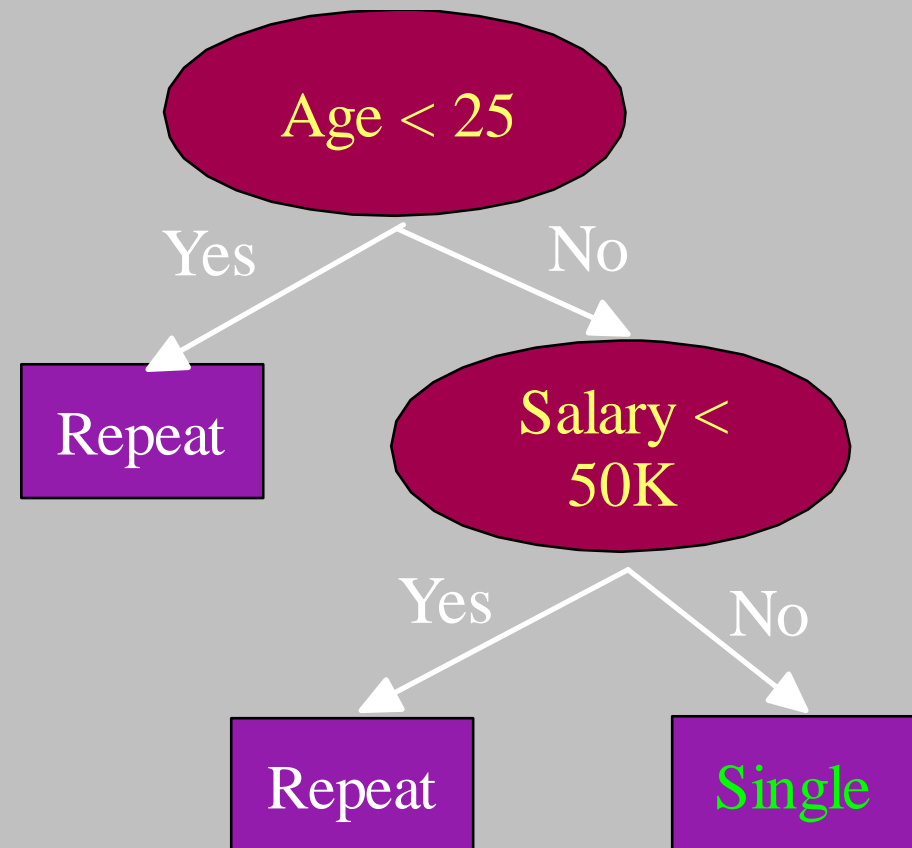
- Converges to maximum likelihood estimate.
(D. Agrawal & C.C. Aggarwal, PODS 2001)

Works Well



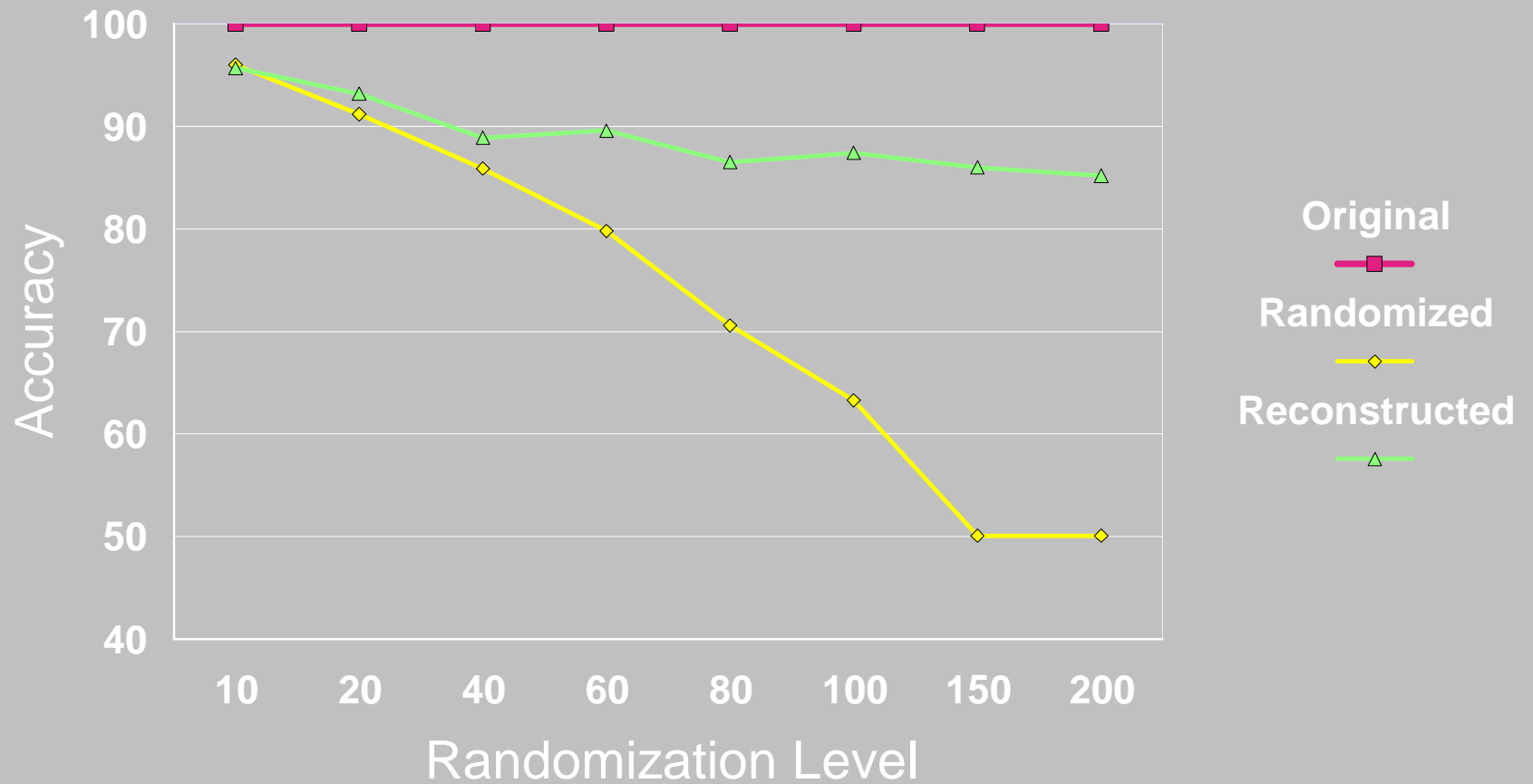
Application to Building Decision Trees

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



Accuracy vs. Randomization

Fn 3



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]*

- Privacy-Preserving OLAP
 - *Agrawal, Srikant, Thomas [Sigmod 05]*

HDB Optimal k -Anonymization

- **Goal:** De-identify data such that it retains integrity, but is resistant to data linkage attacks.
- **Motivation:** Naïve methods are resistant to data linkage attacks, in which combine subject data with publicly available information to re-identify represented individuals.
- **Samarati and Sweeney k -anonymity* method**
 - A k -anonymized data set has the property that each record is indistinguishable from at least $k-1$ other records within the data set.
- **Optimal k -anonymization**
 - We have developed a k -anonymization algorithm that finds optimal k -anonymizations under two representative cost measures and variations of k .

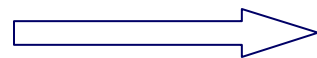
Process of k -anonymization

- **Data suppression** - involves deleting cell values or entire tuples.
- **Value generalization** - entails replacing specific values such as a phone number with a more general one, such as the area code alone.

Advantages of Optimal k -anonymization

- **Truthful** - Unlike other disclosure protection techniques that use data scrambling, swapping, or adding noise, all information within a k -anonymized dataset is truthful.
- **Secure** - More secure than other de-identification methods, which may inadvertently reveal confidential information.

Name	Phone	Diagnosis
Rob	408-402-3456	HIV
Ed	408-888-2367	Rubella
Sam	408-767-1231	Asthma



**k -anonymization
($k=3$, on name+phone)**

Name	Phone	Diagnosis
-	408-***-****	HIV
-	408-***-****	Rubella
-	408-***-****	Asthma

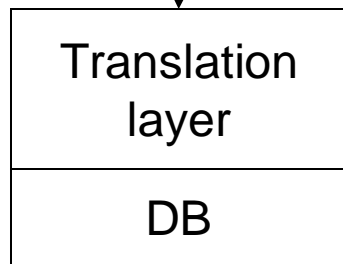
* P. Samarati and L. Sweeney. "Generalizing Data to Provide Anonymity when Disclosing Information." In Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 188, 1998.

HDB Order Preserving Encryption

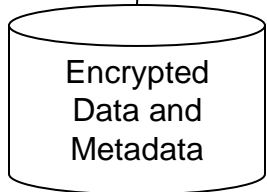


Plaintext Queries

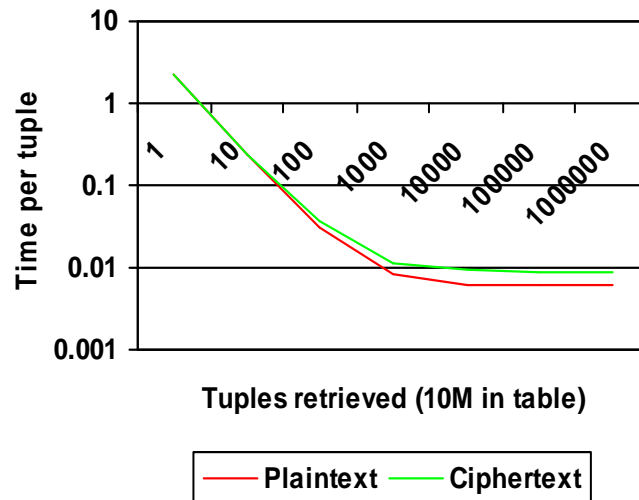
Select name from Emp where sal > 100000



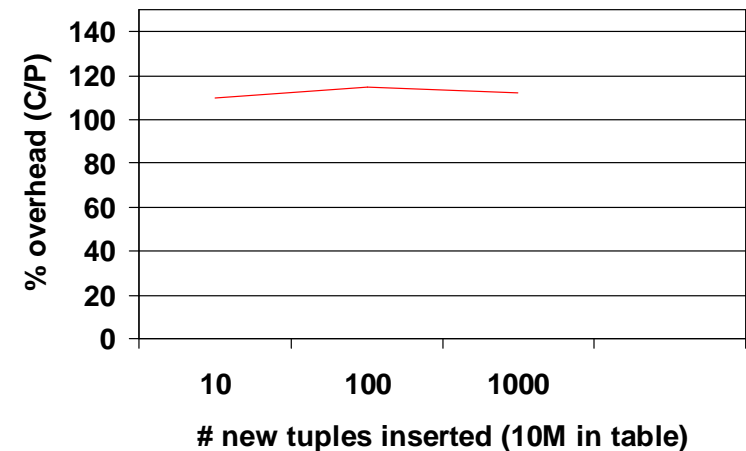
Select decrypt ("xsxx", key1)
from "cwlxss"
Where
"xescs" >
OPESencr(100000, key2)



Sigmod04

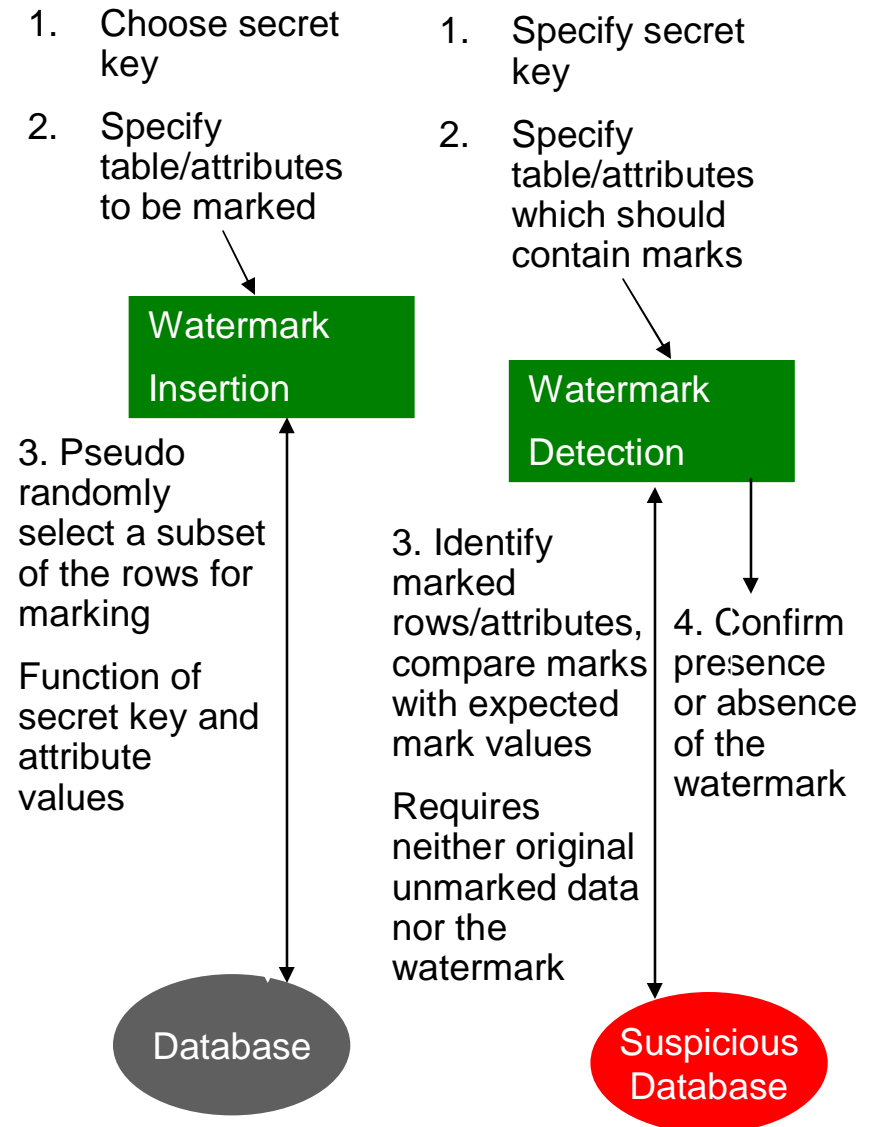


- Translation of plaintext queries into equivalent queries over encrypted data and metadata
- Use of regular as well as order preserving encryption for efficient evaluation of range queries over encrypted columns
- OPES encryption effectively hides the distribution of original plaintext values by encrypting input plaintext values into any chosen target distribution



HDB Watermarking

- **Goal: Deter data theft and assert ownership of pirated copies.**
- **Watermark – Intentionally introduced pattern in the data.**
 - Very unlikely to occur by chance.
 - Hard to find => hard to destroy (robust against malicious attacks).
- **Existing watermarking techniques developed for multimedia are not applicable to database tables.**
 - Rows in a table are unordered.
 - Rows can be inserted, updated, deleted.
 - Attributes can be added, dropped.
- **New algorithm for watermarking database tables.**
 - Watermark can be detected using only a subset of the rows and attributes of a table.
 - Robust against updates, incrementally updatable.



Challenges

*Asking questions is easy:
it's answering them that's hard.*

Policy Specification & Inference Control

- How to determine if the policy specification correctly captures the intent? (The person specifying the policy is usually not a Computer Scientist!).
- How to help the consumer understand what he is consenting to?
- For what classes of queries and policies and under what practical assumptions, can we guarantee safety from inference?
- How to use auditing for inference control?





Data Pointillism

Name	Phone
Bob	394-1015
Alice	396-1012
Alice	396-1112

Phone	Address	City
396-1012	Maple St	Chatham
394-1015	-	Madison
396-1112	Maple St	Madison

Patient	Policy#
Alice	AAA1035
Bob	AAA1035
Alice	UHG1035



- > 14B records with Choicepoint
- Data from > 22,000 sources in RDC's GRID
- >550 companies compiling databases of pvt information

Bob	394-1015	Maple St	Madison	AAA1035
Alice	396-1012	Maple St	Chatham	UHG1035
Alice	396-1112	Maple St	Madison	AAA1035

- Accuracy? Limits?
- How to allow someone to verify data?
- Identifying and correcting errors?
- Usage control?

Kafkaesque Nightmare or Solomonic Talisman?

Massively Distributed Data Management

- What if personal data lives on a personal device?
- On demand data sharing
- Safety of data on the device
- Distributed backup in the network



512MB SanDisk Cruzer
\$47.99



Transcend 40GB Portable Hard Disk USB
95mm x 71.5mm x 15mm, \$189

Privacy & Game Theory

- Assume that parties are rational and want to achieve the best result for themselves.
- What mechanisms can be designed so that the best strategy for any party (Nash equilibrium) is not to cheat?



Concluding Remarks

- Database technology has opportunity to play crucial role in addressing major challenges of the 21st Century, such as improving Healthcare and Education.
- We need to focus on:
 - Deriving value from bits we know how to manage so well.
 - Demonstrating what could not be done earlier.
- Will we live up to the challenge?



References

- R. Agrawal, R. Srikant. "Privacy Preserving OLAP." *ACM Int'l Conf. On Management of Data (SIGMOD)*, June 2005.
- R. Bayardo, R. Agrawal. "Data Privacy Through Optimal k-Anonymization." *Proc. of the 21st Int'l Conf. on Data Engineering*, Tokyo, Japan, April 2005.
- R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzaou, R. Srikant. "Auditing Compliance with a Hippocratic Database." *30th Int'l Conf. on Very Large Databases (VLDB)*, Toronto, Canada, August 2004.
- K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, D. DeWitt. "Limiting Disclosure in Hippocratic Databases." *30th Int'l Conf. on Very Large Databases (VLDB)*, Toronto, Canada, August 2004.
- R. Agrawal, J. Kiernan, R. Srikant, Y. Xu. "Order Preserving Encryption of Numeric Data." *ACM Int'l Conf. On Management of Data (SIGMOD)*, Paris, France, June 2004.
- 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.
- 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. "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.
- 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.

Thank you!

