Experimental Results on Privacy-Preserving Statistics Computation

Rebecca Wright
Computer Science Department
Stevens Institute of Technology

11 December, 2003

Joint work with Canetti, Ishai, Kumar, Reiter, and Rubenfeld; and Subramaniam and Yang.

Surveillance and Data Mining

- Analyze large amounts of data from diverse sources.
- Law enforcement and homeland security:
  - detect and thwart possible incidents before they occur
  - recognize that an incident is underway
  - identify and prosecute criminals/terrorists after incidents occur
- Biomedical research
- Marketing, personalized customer service

Erosion of Privacy

"You have zero privacy. Get over it."

- Scott McNealy, 1999

- Changes in technology are making privacy harder.
  - reduced cost for data storage
  - increased ability to process large amounts of data
- Especially critical now (because decisions for surveillance systems are being made, and because of relevant legislation)

Privacy-Preserving Data Mining

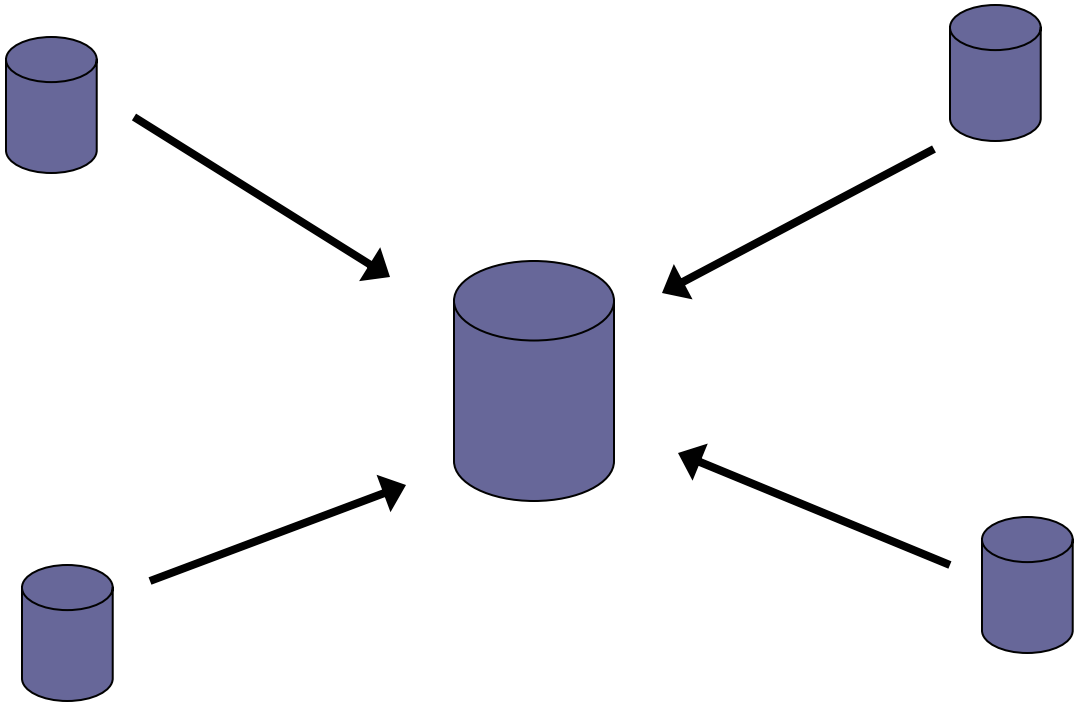
Allow multiple data holders to collaborate to compute important information while protecting the privacy of other information.

- Security-related information
- Epidemiological information
- Marketing information
- etc.

Advantages of privacy protection

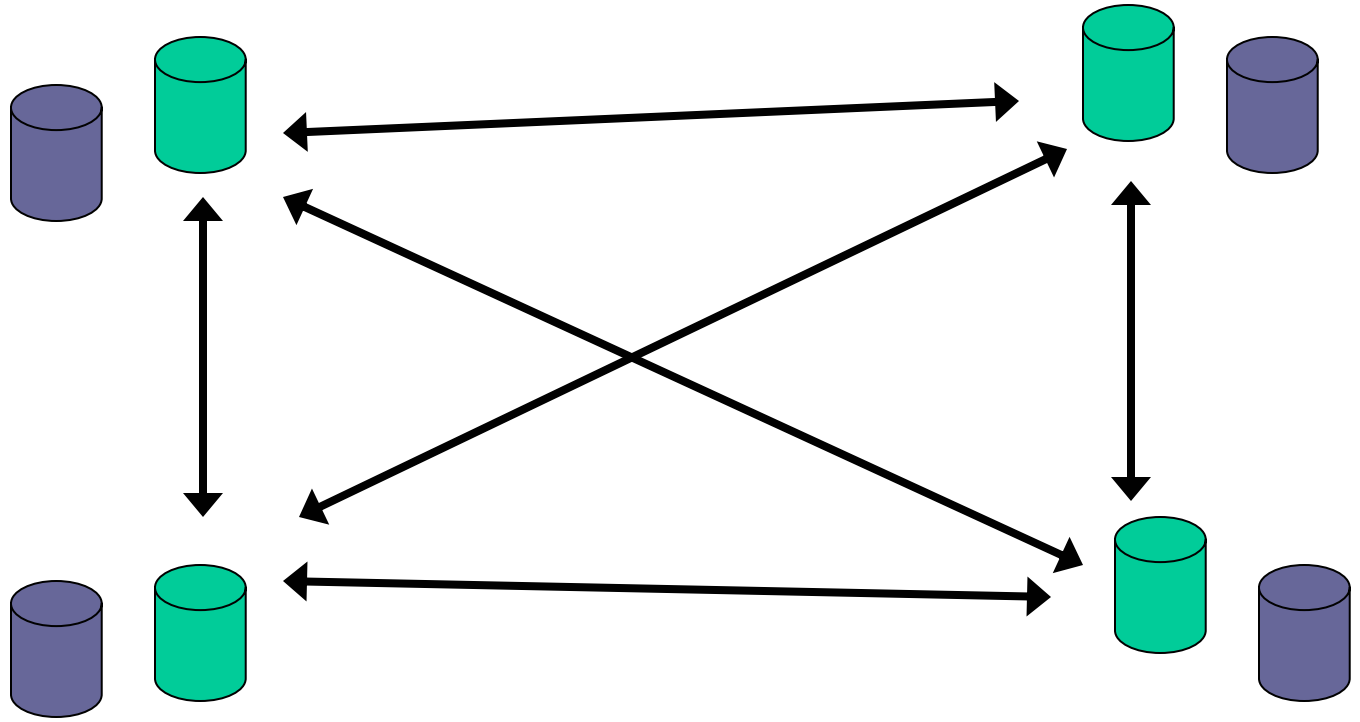
- protection of personal information
- protection of proprietary or sensitive information
- enables collaboration between different data owners (since they may be more willing or able to collaborate if they need not reveal their information)
- Compliance with legislative policies

Data Mining



- Gather and analyze data from diverse sources.

Privacy-Preserving Data Mining



- Enable analysis of data from diverse sources, without requiring original data to be gathered in a single place.

Cryptographic Approach

- Using cryptography, provably does not reveal anything except output of computation.
  - Privacy-preserving computation of decision trees [LP00]
  - Secure computation of approximate Hamming distance of two large data sets [FIMNSW01]
  - Privacy-preserving statistical analysis [CIKRRW01]
  - Privacy-preserving set intersection [FP03]

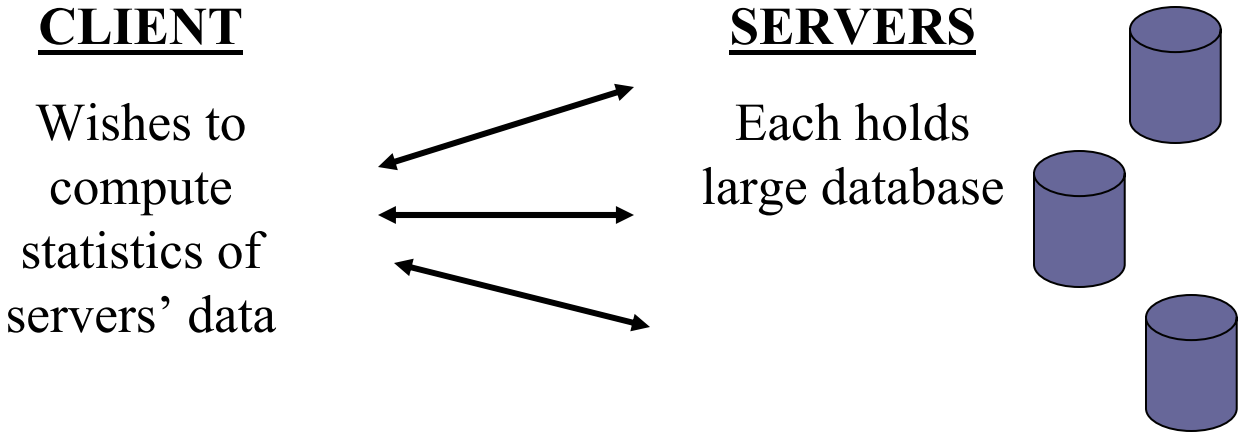
Secure Multiparty Computation

- Elegant and general cryptographic tool for performing private distributed computation.
- Allows multiple input holders to collaborate to learn the output of some function ***f*** applied to their inputs, without each party learning anything not implied by its own input and output.
- Solutions exist for any ***f*** with polynomial overhead in complexity of ***f*** and size of inputs.

Revealing Partial Information

- Many instances of partial information do not leak anything sensitive.
- But, it may not be known in advance which information will and won't be sensitive. (In the HIPAA context, this is less true.)
- Cryptographic definitions do not allow partial information leakage, unless it is explicitly defined as part of the output.

Privacy-Protecting Statistics [CIKRRW01]



- Parties communicate using cryptographic protocols designed so that:
  - Client learns desired statistics, but learns nothing else about data (including individual values or partial computations for each database)
  - Servers do not learn which fields are queried, or any information about other servers' data
  - Computation and communication are very efficient (linear computation, linear or even sublinear communication).

Privacy Concerns

- Protect clients from revealing type of sample population, type of specific data used
- Protect database owners from revealing unnecessary information or providing a higher quality of service than paid for
- Protect individuals from large-scale dispersal of their personal information

Non-Private and Inefficient Solutions

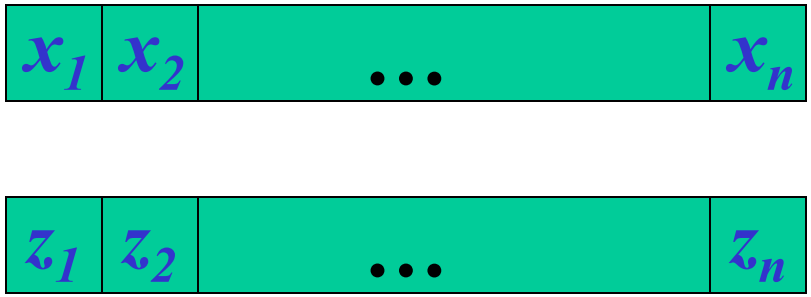
- Database sends client entire database (violates database privacy)
- For sample size *m*, use SPIR to learn *m* values (violates database privacy)
- Client sends selections to database, database does computation (violates client privacy, won't work in general for multiple databases)
- General secure multiparty computation (not efficient for large databases)

Homomorphic Encryption

- Certain computations on encrypted messages correspond to other computations on the cleartext messages.
- For Paillier encryption,
  - $E(m_1) \cdot E(m_2) = E(m_1 + m_2)$
  - also implies $E(m)^x = E(mx)$

Privacy-Preserving Statistics Protocol

- To learn mean and variance: enough to learn sum and sum of squares
- Server stores:



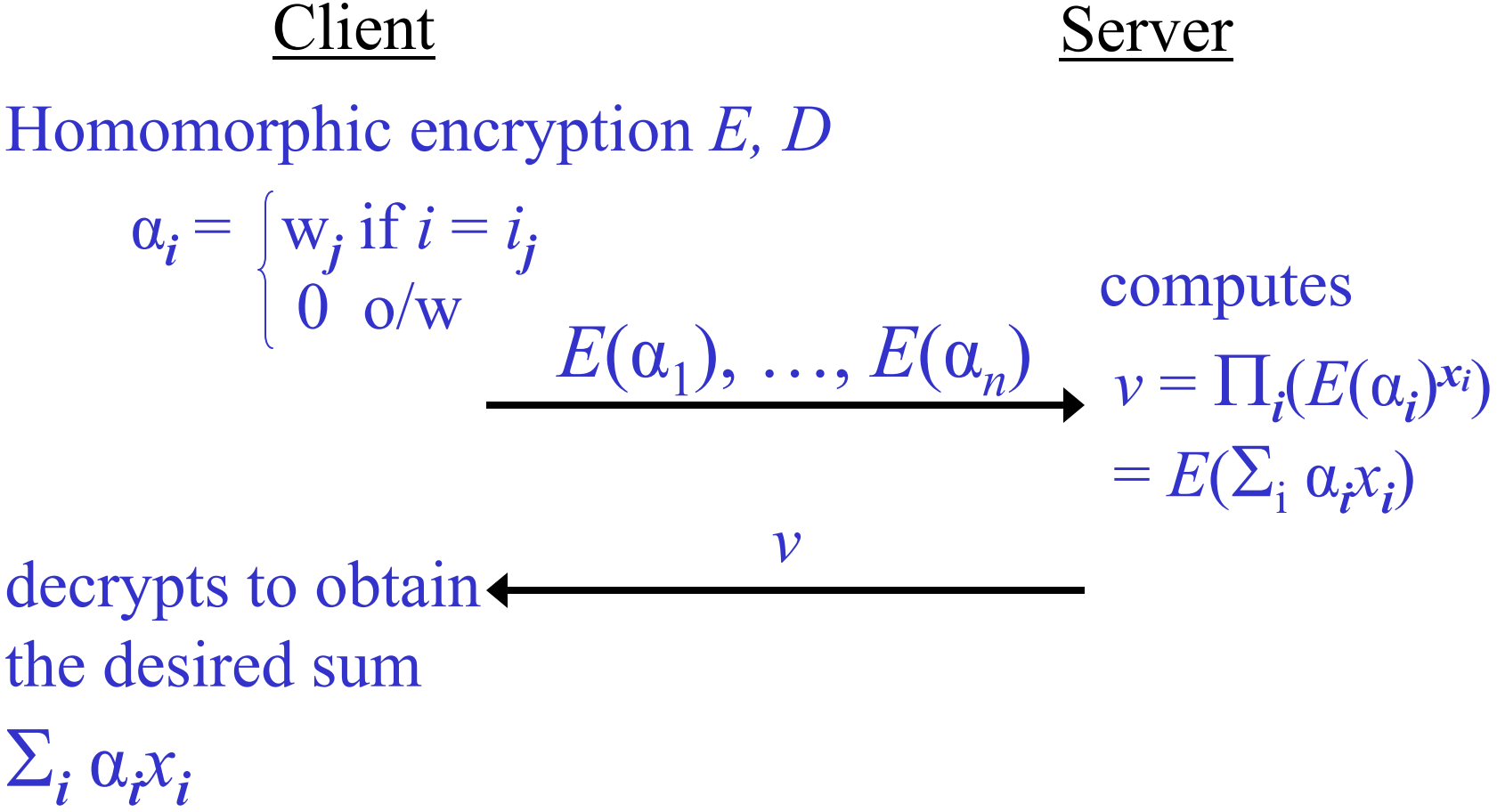
$(z_i = x_i^2)$

and responds to queries from both

- Efficient protocol for sum ⟶ efficient protocol for mean and variance

Weighted Sum Protocol

Client wants to compute selected linear combination of m items:



Efficiency

- Linear communication and computation (feasible in many cases)
- If n is large and m is small, would like solutions whose complexity depends on m rather than n. We have solutions with sublinear communication (i.e. polynomial in m, not n) and polynomial computation.
- These solutions work for any function f, not just statistics functions.

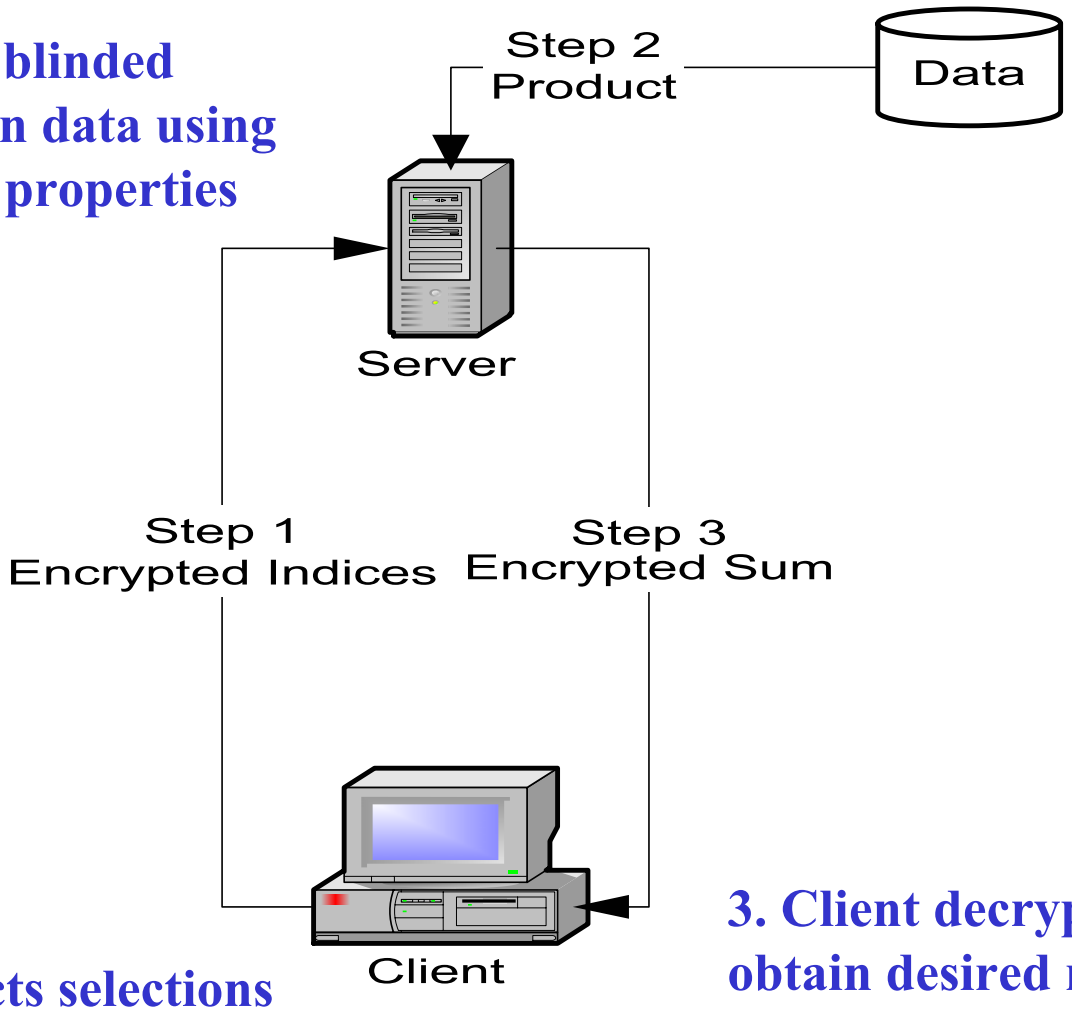
Experimental Results

- Selected sum protocol using linear communication and computation.

- Implementation
  - implementation in Java and C++
  - uses Paillier encryption
  - experiments use synthetic data

Basic Architecture

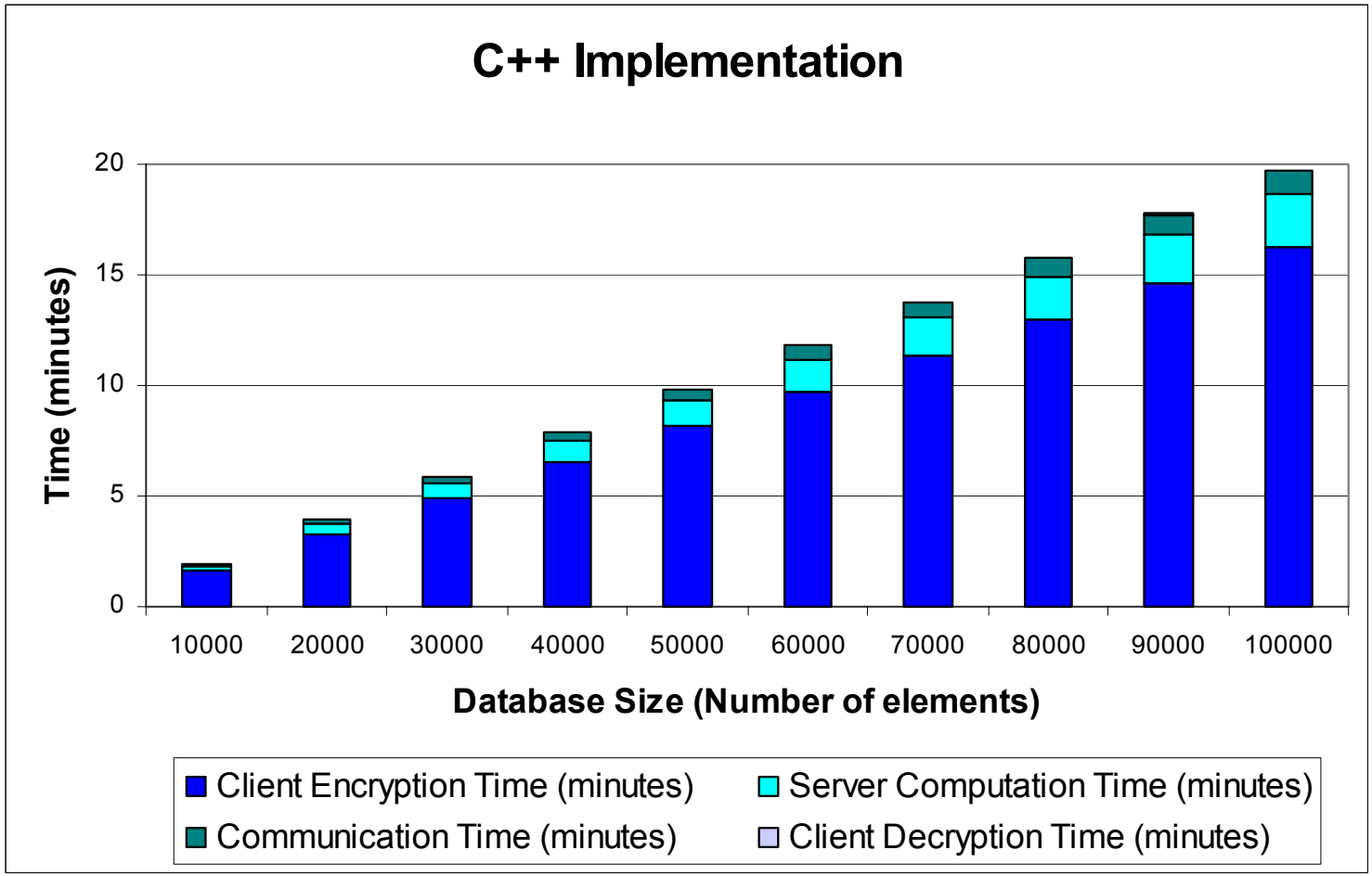
2. Server does blinded computation on data using homomorphic properties



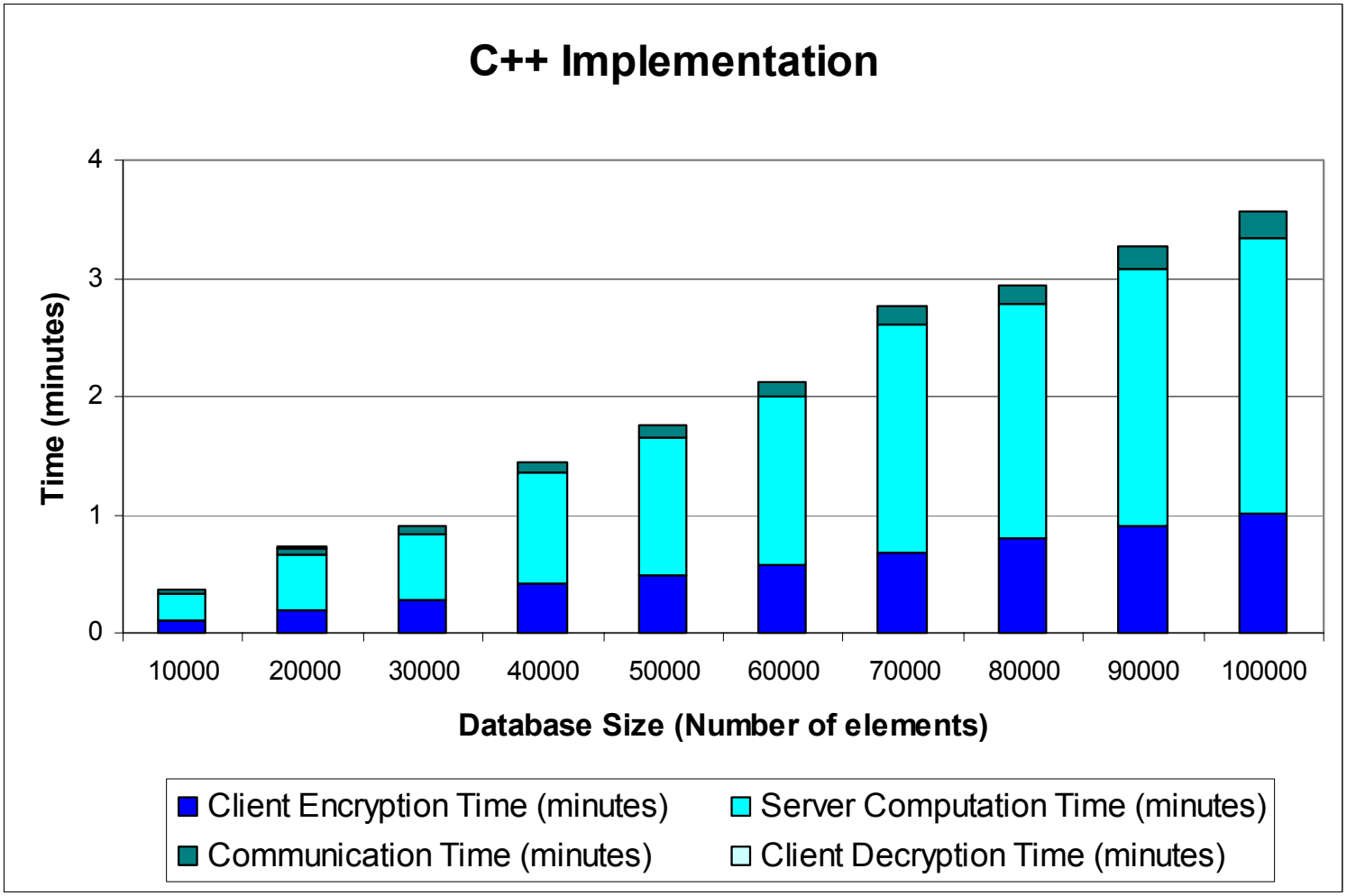
1. Client protects selections using homomorphic encryption

3. Client decrypts to obtain desired result

Experimental Results



Run Time with Preprocessing



Shortcomings of Cryptographic Approach / Further Research

- Efficiency
- Need for efficient solutions to be constructed specifically for each desired functionality
- SMP definitions both too strong and too weak
  - require every data element to be accessed, hence linear lower bounds
  - say nothing about what output itself leaks

Advantages of Cryptographic Approach

- Strong and rigorous privacy guarantees of any information beyond what computed output itself reveals.
- Since no partial information is leaked, can be used as general building blocks without buildup of privacy loss.
- Results generalize to malicious participants (at an additional loss of efficiency).

The PORTIA Project

Privacy, Obligations, and Rights in Technologies of Information Assessment

A five-year multidisciplinary project focusing on the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners, and data users.

The PORTIA Project

- Major technical themes:
  - Privacy-preserving data mining
  - identity theft and identity privacy
  - database policy enforcement tools
  - using trusted platforms to provide trusted privacy-preserving services

PORTIA Personnel

- Academic investigators:
  - Dan Boneh, Hector Garcia-Molina, John Mitchell, Rajeev Motwani, *Stanford*
  - Joan Feigenbaum, Ravi Kannan, Avi Silberschatz, *Yale*
  - Stephanie Forrest, *University of New Mexico*
  - Helen Nissenbaum, *NYU*
  - Rebecca Wright, *Stevens Institute of Technology*

PORTIA Personnel

- Research partners
  - Jack Balkin, *Yale Law School*
  - Greg Crabb, *Secret Service*
  - Cynthia Dwork, Brian LaMacchia, *Microsoft*
  - Sam Hawala, *US Census Bureau*
  - Kevin McCurley, *IBM Research*
  - Perry Miller, *Yale Center for Medical Informatics*
  - John Morris, *Center for Democracy and Technology*
  - Benny Pinkas, *HP Labs*
  - Marc Rotenberg, *Electronic Privacy Information Center*
  - Alejandro Schaffer, *DHHS/National Institutes of Health*
  - Dan Schutzer, *Citigroup*

Conclusions and Future Plans

- Cryptographic solutions are possible, but expensive. Optimizations can help in some situations.
- Cryptography is not the whole solution, but can be a useful component: Investigate integration of cryptographic approach and other approaches.
- Enable privacy-protection for a broad range of data base computations and data mining algorithms
- Technology and policy must work together.

More information

- www.cs.stevens-tech.edu/~rwright
- The PORTIA project: crypto.stanford.edu/portia.