



A Privacy Preserving Framework for Integrating, Storing and Querying Biological Data

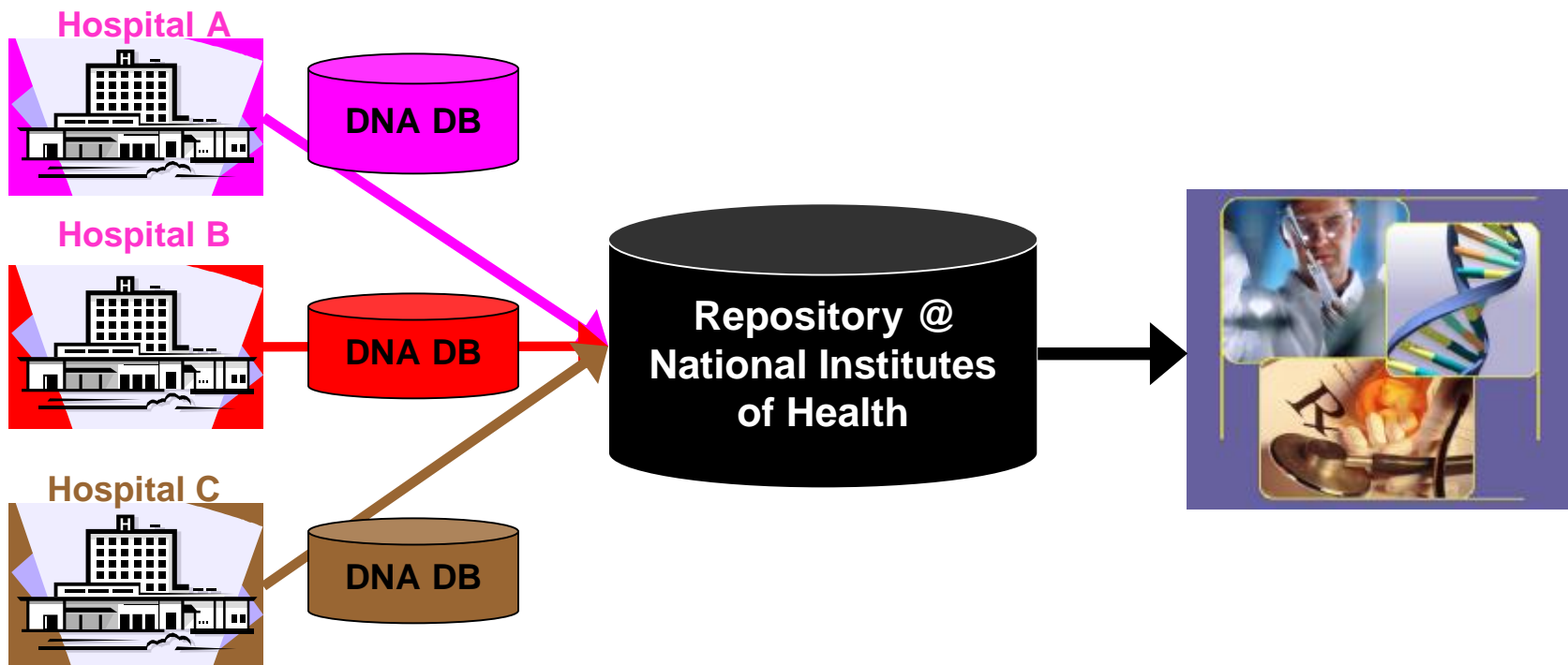
Murat Kantarcioglu, Ph.D.
University of Texas at Dallas

National Academy of Sciences' workshop on collecting, storing, protecting
and accessing biological data collection

November 17-18, 2008

Example

- n **Goal:** Construct repositories of person-specific DNA for pharmacogenetic and biomedical research



- n **Challenge:** Need to merge, store, query records securely without violating privacy

Privacy Enhancing Technologies

n Cryptographic techniques

- Symmetric key systems
- Public key systems
- Homomorphic Encryption
 - n Certain operations on the encrypted data sets are possible using Homomorphic encryption
- Id-based encryption
 - n Any string (bob@company.com) could be public key in Id-based encryption
- Cryptographic Hardware
 - n Encryption keys can be stored securely.

Privacy Enhancing Technologies

n **Anonymization techniques**

- Non-individually identifiable data sets with formal privacy guarantees could be released using techniques such as k-anonymity.

n **Data analysis & Digital Forensics techniques**

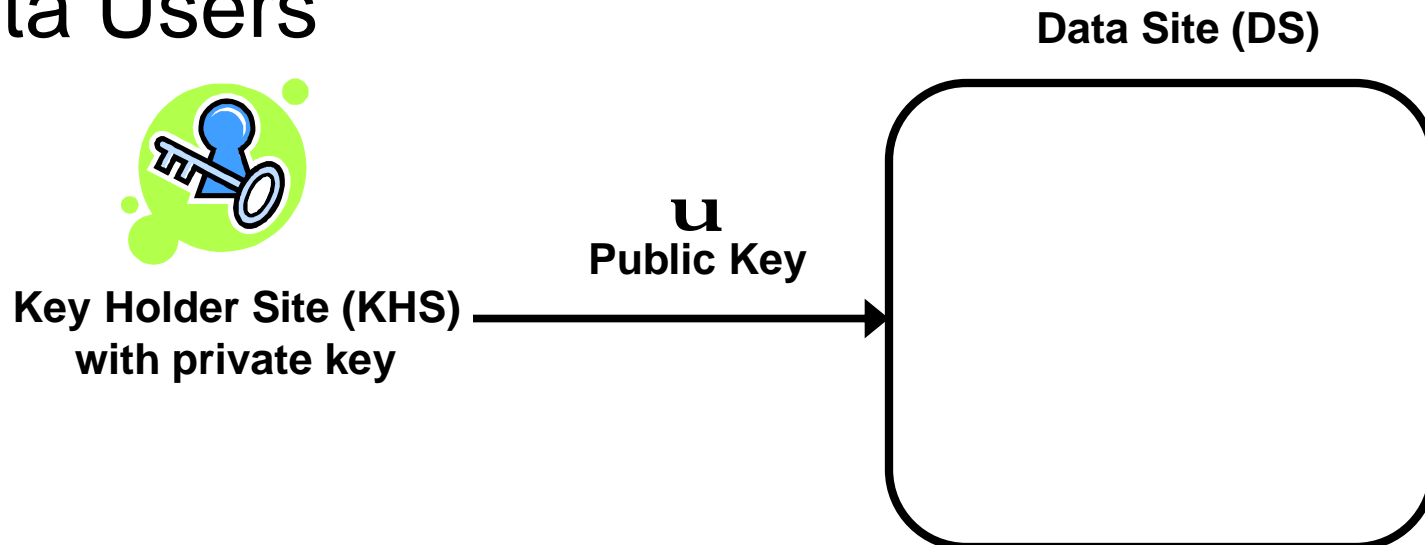
- Audit logs
- Online auditing tools
- Digital Forensics techniques

Example:

Secure Record Management

(Kantarcioglu, Jiang, Liu, Malin, IEEE TITB 2008)

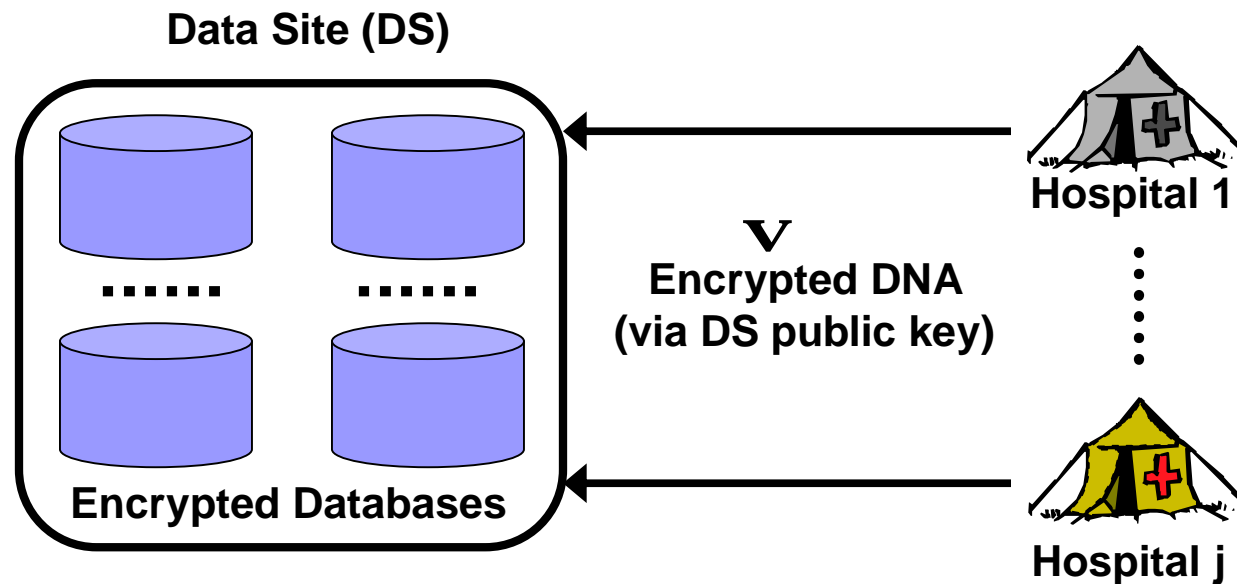
- n Data Providers
- n Third Party Data Managers β Required
- n Data Users



Key Generation: <public, private>

Architecture

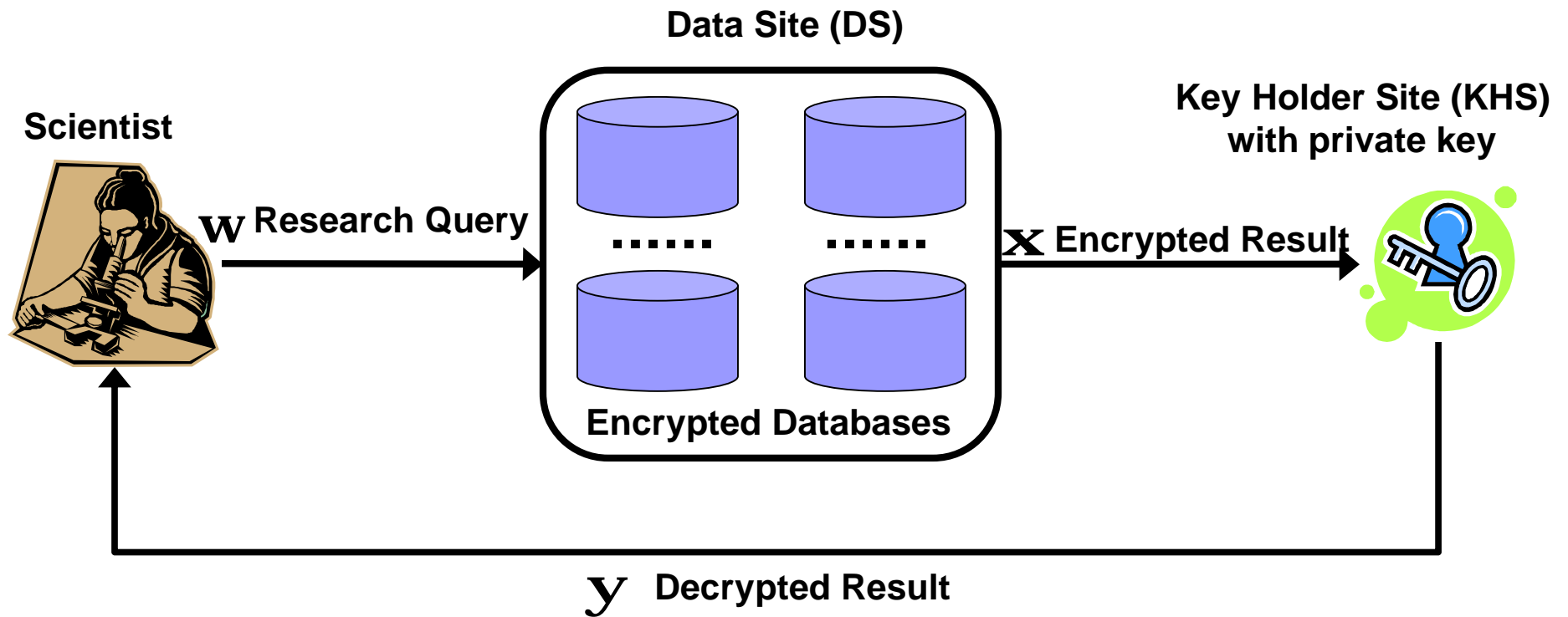
(Kantarcioglu, Jiang, Liu, Malin, IEEE TITB 2008)



Data Encryption

Architecture

(Kantarcioglu, Jiang, Liu, Malin, IEEE TITB 2008)



Query Issuance Result Decryption Query Processing



Architecture

(Kantarcioglu, Jiang, Liu, Malin, IEEE TITB 2008)

n **Other capabilities of our system**

- .. Privacy-preserving audits
- .. Efficient privacy-preserving data integration using anonymized data.

n **Current limitations**

- .. Slower compared to non-secure versions.
- .. More storage is needed.

Conclusions

- n Existing technologies **enable** privacy-preserving biological data integration, storage and querying.
- n The trade-off is between cost versus privacy
 - Almost any task could be achieved **without violating privacy**.
- n With more research, all the potential benefits of “biological data” could be unlocked at a reasonable cost without violating individual privacy.

References

- n For the details of our proposed system , please see the following references. (**Joint work with Wei Jiang** (Purdue University), and **Bradley Malin**, (Vanderbilt University))
 - Murat Kantarcioglu, Wei Jiang, Ying Liu, and Bradley Malin, "**A Cryptographic Approach to Securely Share and Query Genomic Sequences**", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 5, pp 606-617 (2008)
 - Murat Kantarcioglu, Wei Jiang, and Bradley Malin, "**A Privacy-Preserving Framework for Integrating Person-Specific Databases** ", Privacy in Statistical Databases, 2008, LNCS 5262, pp. 298–314, 2008.