VANDERBILT
UNIVERSITY

Brookings Institution Expert Workshop on Statistical & Epidemiological Issues in Active Medical Product Surveillance

A Brief Introduction to Privacy Enhancing Technologies for Surveillance Purposes

Bradley Malin, Ph.D.

Assistant Prof. of Biomedical Informatics, School of Medicine

Assistant Prof. of Computer Science, School of Engineering

Vanderbilt University

February 16, 2011

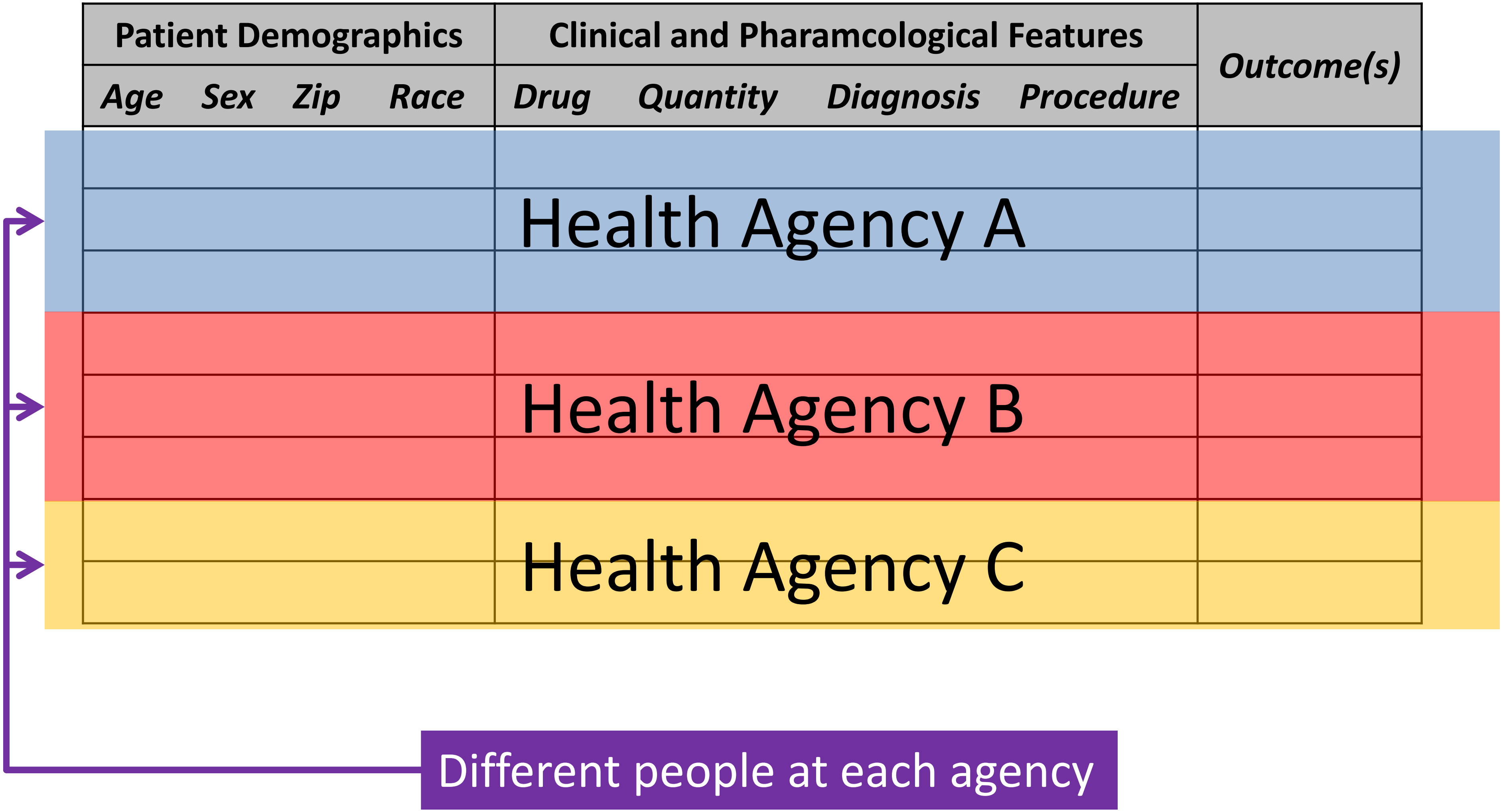
Privacy Preserving Data Mining in Application

- There are generic “solutions” that provide provable privacy and utility
- They often need to be tailored to specific applications
- Simply because there may be no published solution for Sentinel needs specifically … does not mean that adaptation cannot be achieved (or is difficult)

A Generic Data View

Patient Demographics				**Clinical and Pharamcological Features**				***Outcome(s)***
Age	***Sex***	***Zip***	***Race***	***Drug***	***Quantity***	***Diagnosis***	***Procedure***	

“Horizontally” Partitioned Data



Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

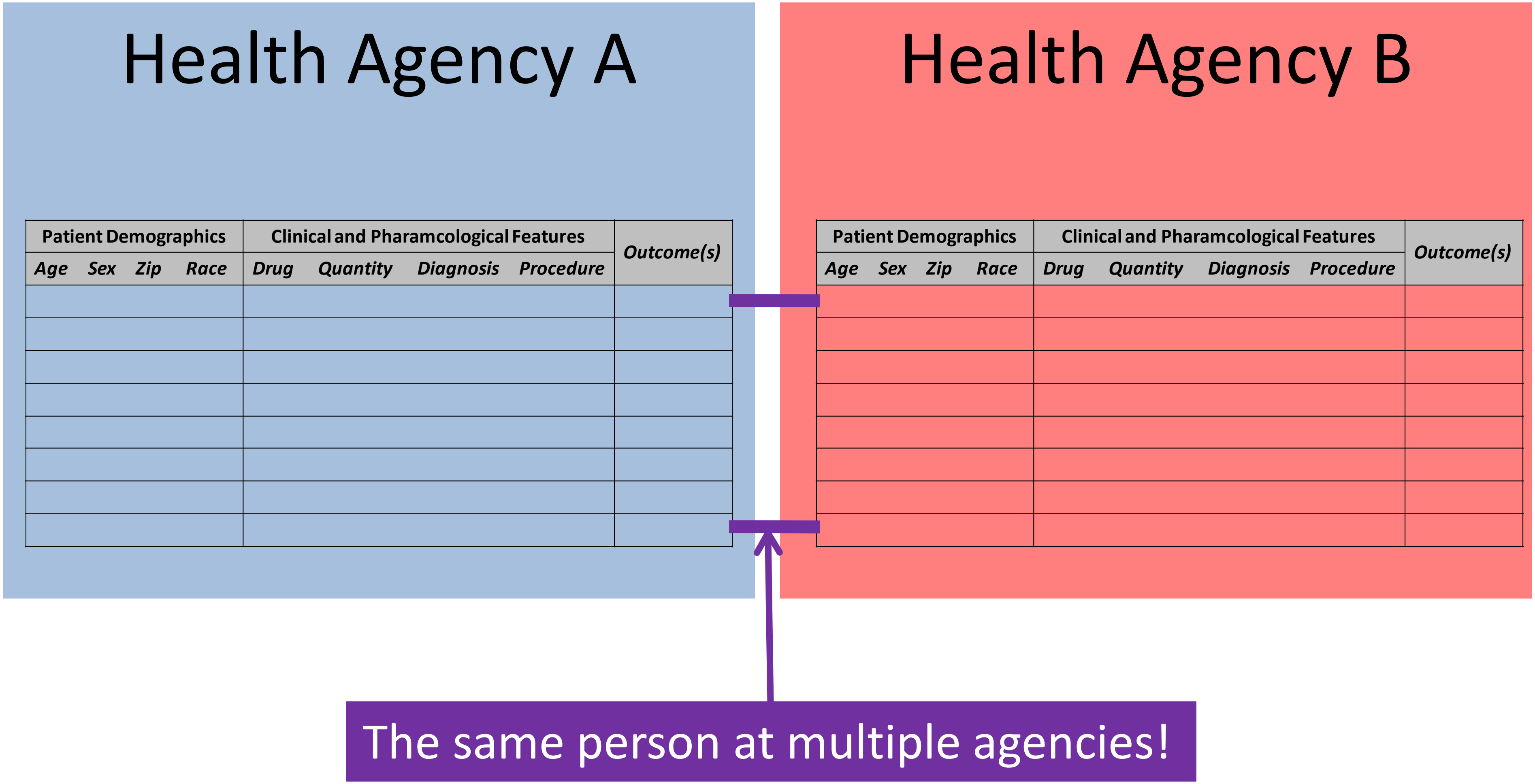
Health Agency A

Health Agency B

Health Agency C

Different people at each agency

“Vertically” Partitioned Data



Health Agency A

Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

Health Agency B

Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

The same person at multiple agencies!

Aspects of Solutions for Horizontal Partitioning

Manipulation

- Transformation
- Generalization
- Randomization
- Cryptographic

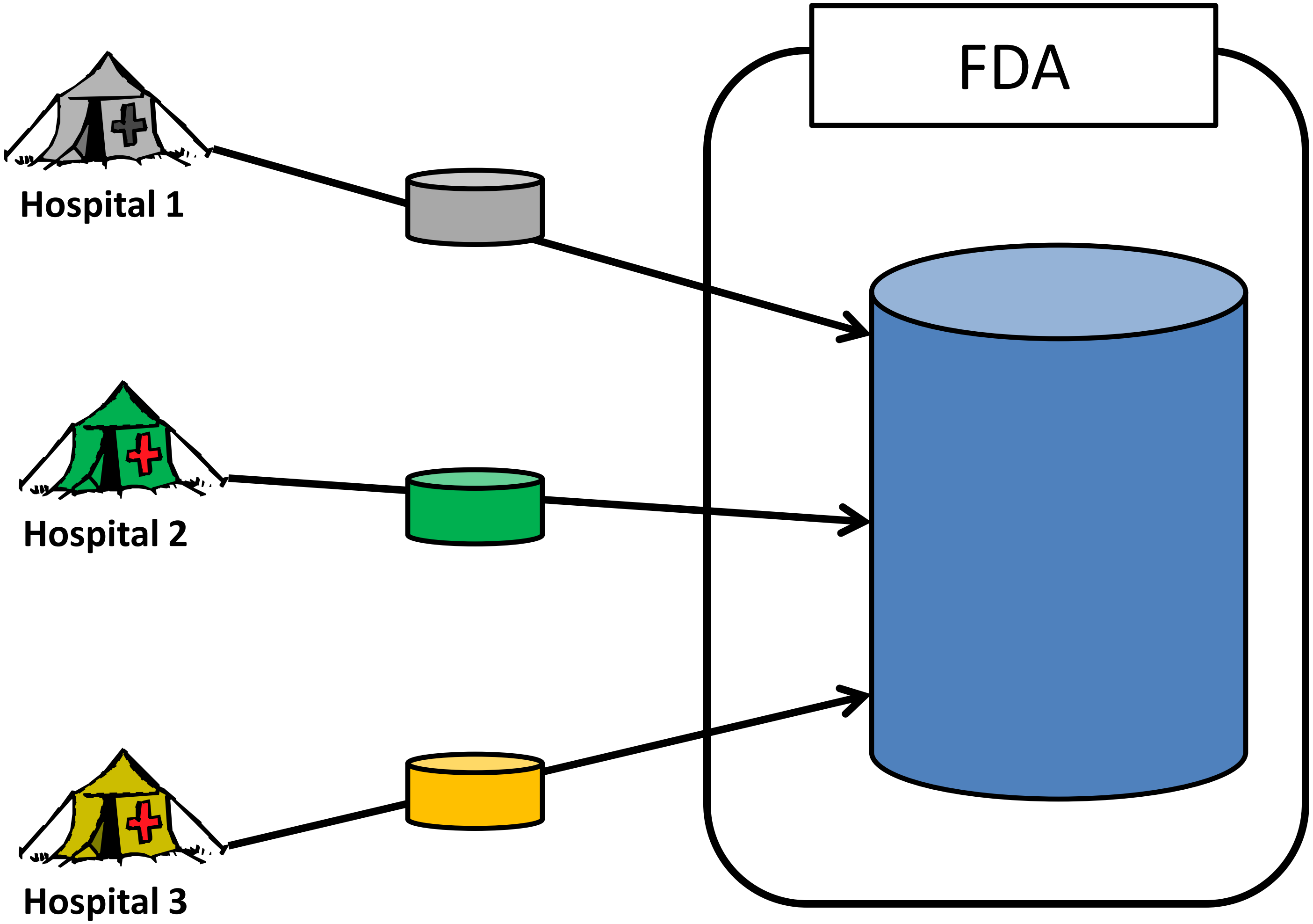
Information Shared

- Data
- Models

Interaction

- Interactive Agencies
- Intermediaries
- Third Parties

Non-Interactive



Generalization of Data

- Reveal abstractions of actual values

e.g., 5-digit zip code → 3-digit zip code

e.g., 1-year age range → 5-year age range

- Can be formalized to guarantee protection for each record shared

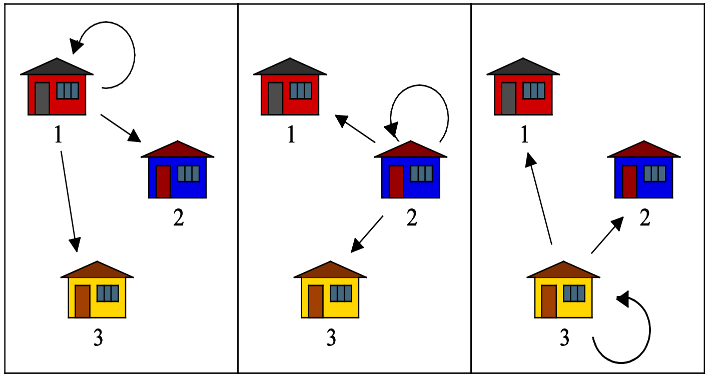
e.g., every record equivalent to k-1 other records [*k*-anonymity principle (Sweeney 2002)]

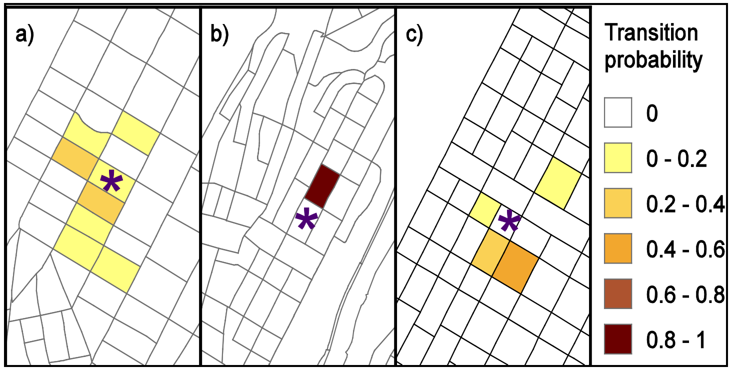
- Concept was used to support the Essence-II biosurveillance system (Lombardo 2003)

Randomization of Demographics

(Wieland et al., PNAS 2008)

- Can “move” patients to formally mitigate identification risks in sharing biosurveillance data.
- Frame the process as a linear programming problem





- Can control the probability that any location from the randomized data set originated from any specific individual in the underlying population
- Experimental evidence indicates the data is still useful for cluster detection

Randomized Response

(Warner 1965; Du & Zhan 2003)

- Used in the survey community for decades, but recently updated for data mining algorithms
- Randomly “change” an agency’s answer according to a known distribution
- Supply results and randomization distribution to recipient.
- Can use distribution to infer the aggregate answer, but not any particular answer
- Note: Based on central limit theorem, so it requires a decent amount of data

A Cryptographic Solution

(Paillier 1999)

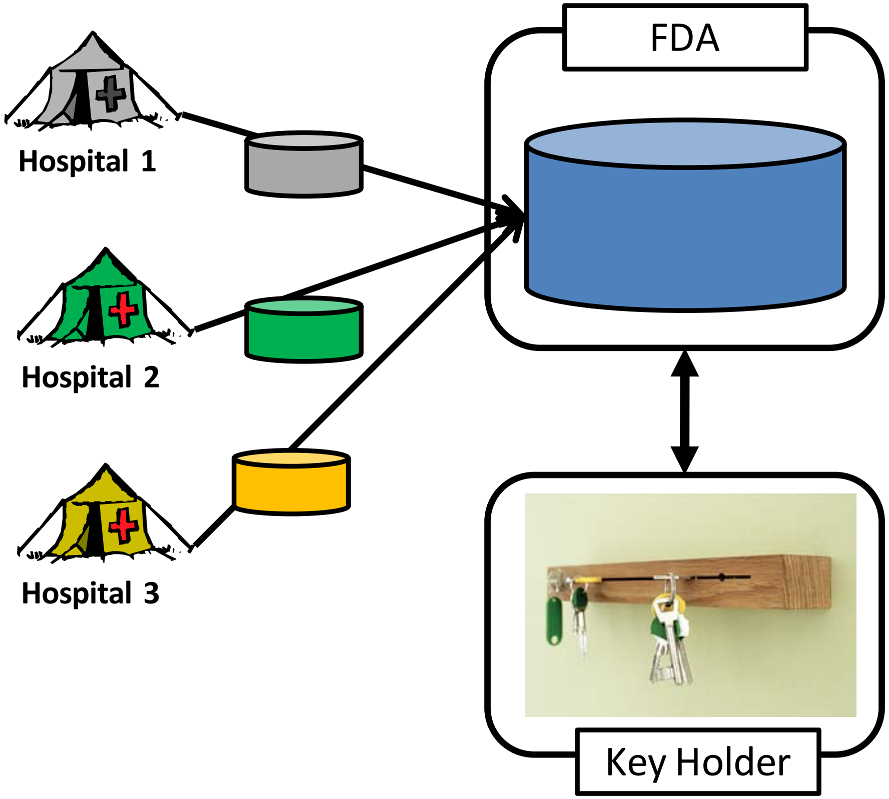
(Genomics Application: Kantarcioglu, Jiang, Liu, & Malin, 2008)

- Agencies send encrypted versions of cases and controls
- Useful variant of crypto in this case is "homomorphic" cryptosystem:

$$E(a+b) = E(a) + E(b)$$

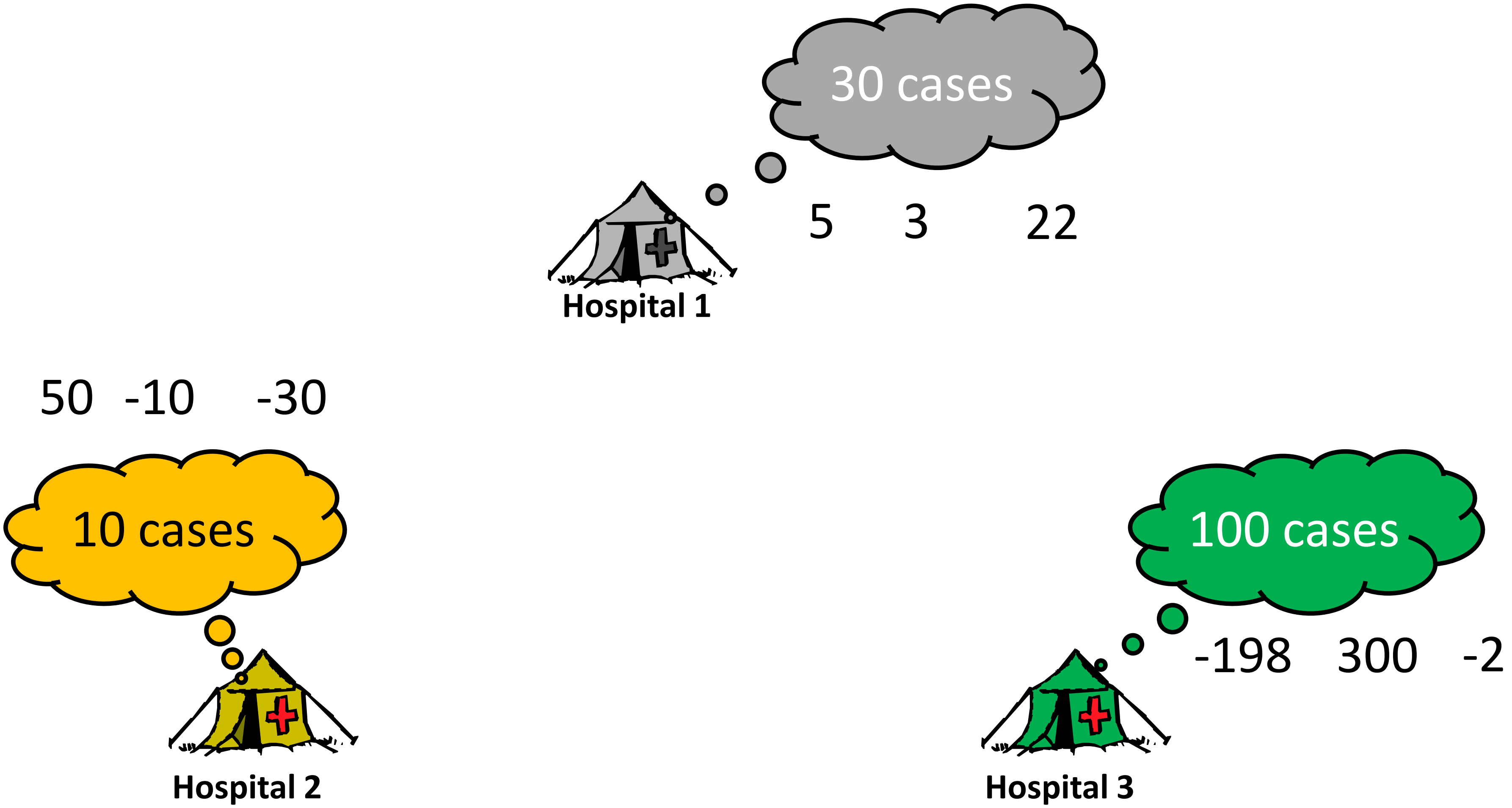
$$D(E(a+b)) = a + b$$

- FDA can "sum" results without learning what any record contributes
- A "key holder" party can report on the decrypted results.
- Known application of such approach in e-voting systems



An Interactive Solution: Secret Sharing

(Shamir 1979)



30 cases

5 3 22

Hospital 1

50 -10 -30

10 cases

Hospital 2

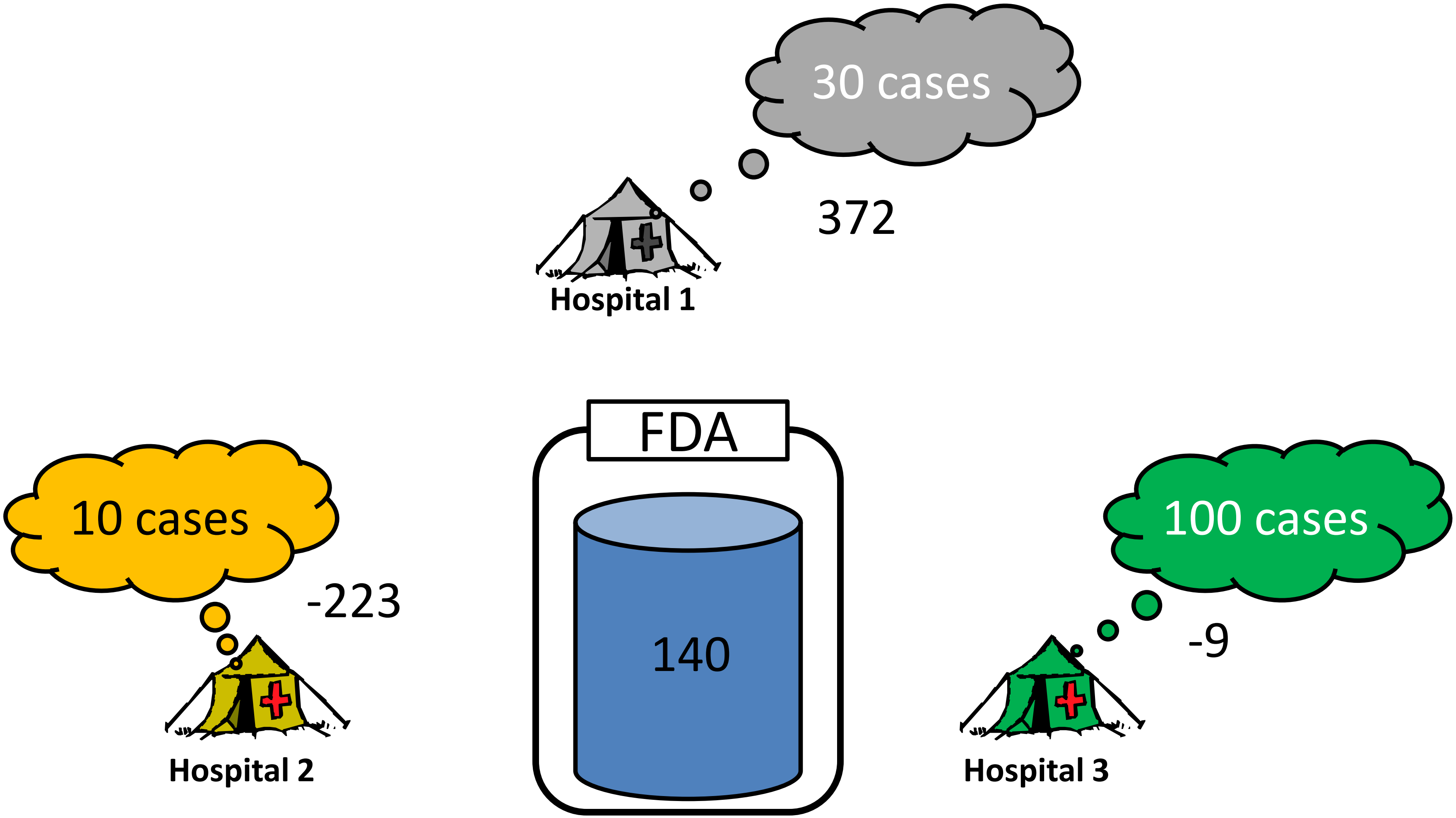
100 cases

-198 300 -2

Hospital 3

An Interactive Solution: Secret Sharing

(Shamir 1979)



30 cases

372

Hospital 1

FDA

140

10 cases

-223

Hospital 2

100 cases

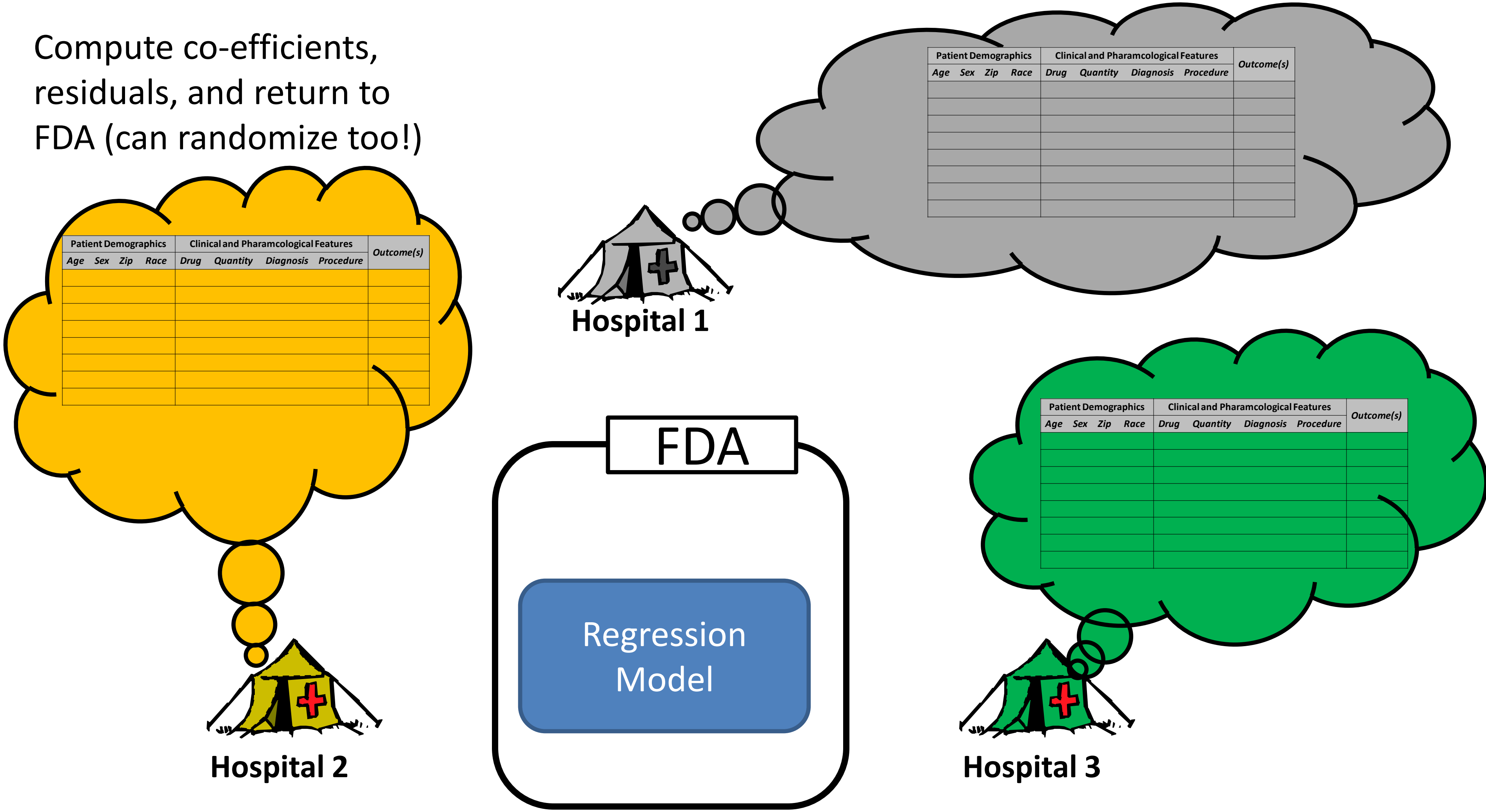
-9

Hospital 3

Model-Based Interaction

(Karr, Lin, Reiter, Sanil 2005)

Compute co-efficients, residuals, and return to FDA (can randomize too!)



Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

Patient Demographics				Clinical and Pharamcological Features				*Outcome(s)*
Age	*Sex*	*Zip*	*Race*	*Drug*	*Quantity*	*Diagnosis*	*Procedure*	

A Couple of Notes on Vertical Partitioning

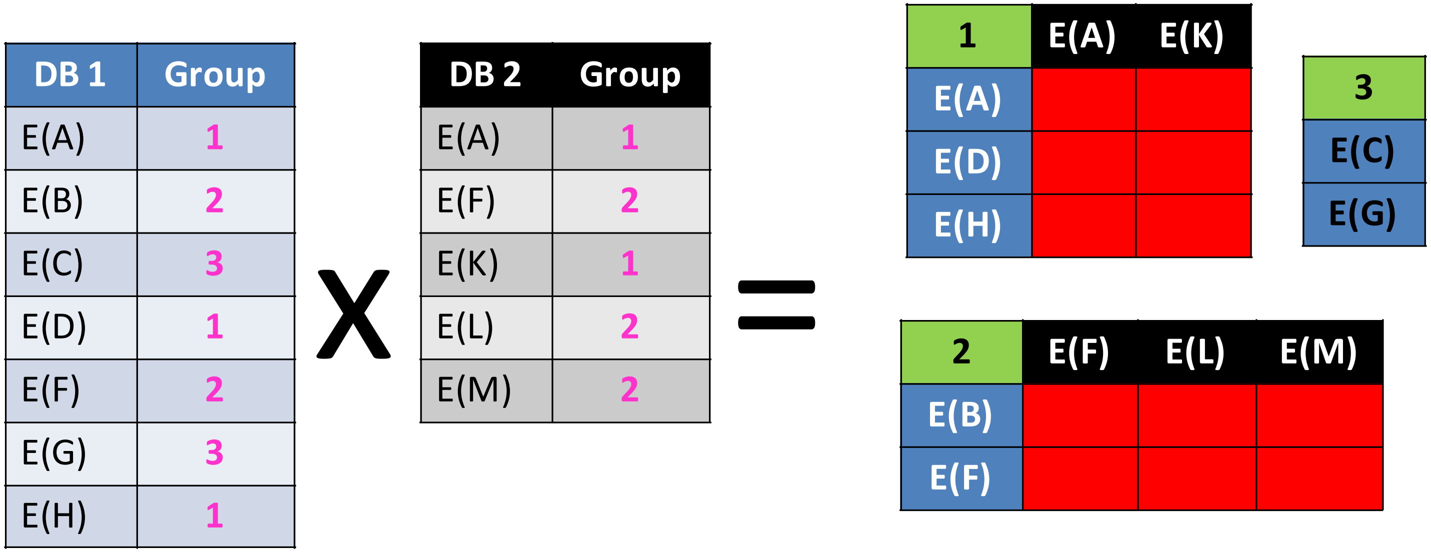
Extension to “Join”

(Kantarcioglu, Inan, Jiang, & Malin 2009)

- Can extend framework to evaluate:

$$E(John) = E(John)$$

- Use de-identified patient information to partition the space (e.g., reveal “all 30 year-old males”)



DB 1	Group
E(A)	1
E(B)	2
E(C)	3
E(D)	1
E(F)	2
E(G)	3
E(H)	1

X

DB 2	Group
E(A)	1
E(F)	2
E(K)	1
E(L)	2
E(M)	2

=

1	E(A)	E(K)
E(A)	🟥	🟥
E(D)	🟥	🟥
E(H)	🟥	🟥

3
E(C)
E(G)

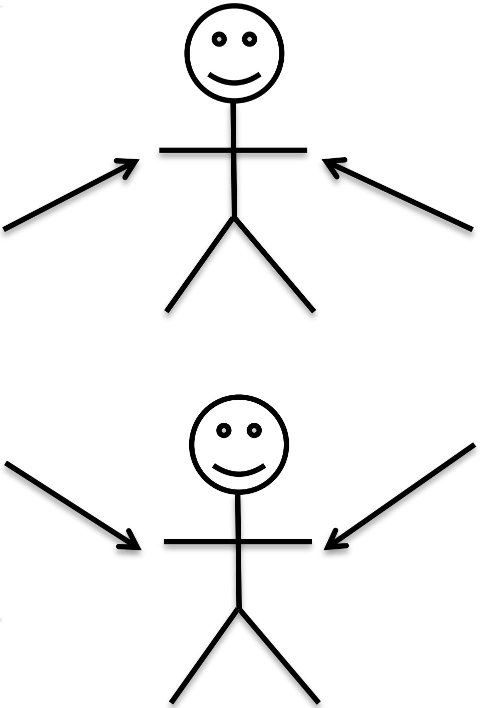
2	E(F)	E(L)	E(M)
E(B)	🟥	🟥	🟥
E(F)	🟥	🟥	🟥

- Experiments with data from the U.S. Census indicate over 1500 times faster than non-partitioned (~ 3 hours for 15000 records)

But Real Patient Information is Messy!

Set of records from Vanderbilt

First Name	Last Name
john	**smith**
lucille	ball
bill	**clinton**
hillary	clinton

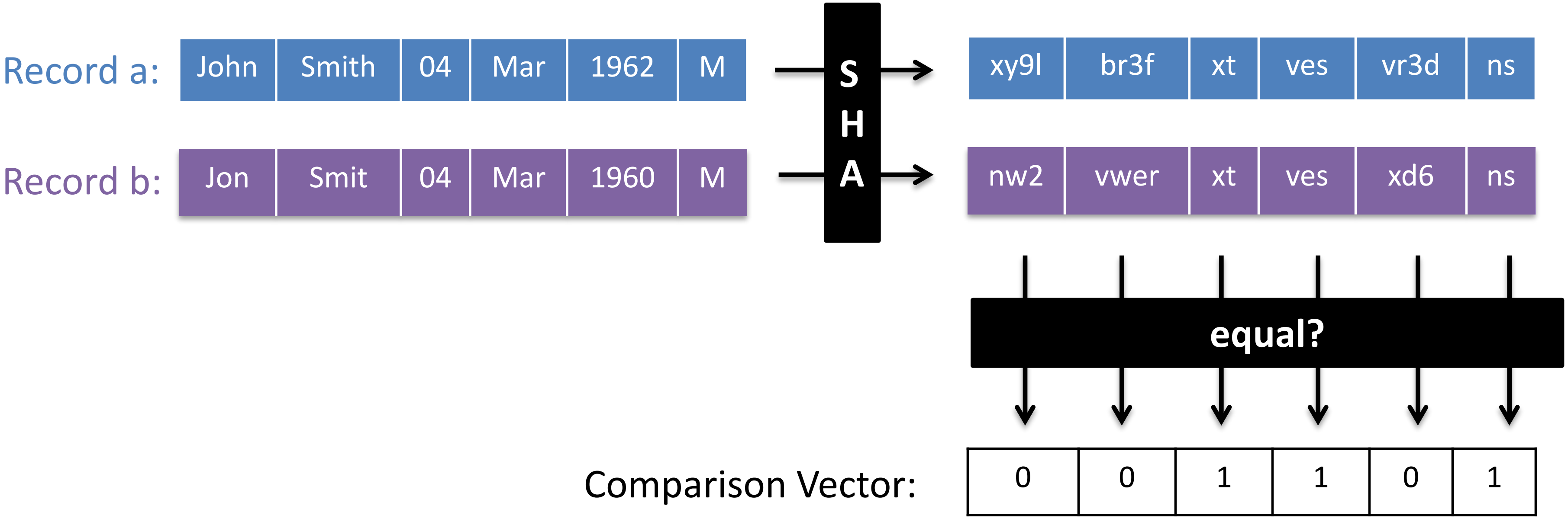


Set of records from Emory

First Name	Last Name
jon	**smyth**
taylor	swift
william	**clinton**
jon	bon jovi

Practical Computations

(Grannis et al 2003)



Record a: John | Smith | 04 | Mar | 1962 | M

Record b: Jon | Smit | 04 | Mar | 1960 | M

xy9l | br3f | xt | ves | vr3d | ns

nw2 | vwer | xt | ves | xd6 | ns

Comparison Vector: 0 | 0 | 1 | 1 | 0 | 1

where SHA is the Secure Hash Algorithm

Approximate Field Comparison with Bloom Filters

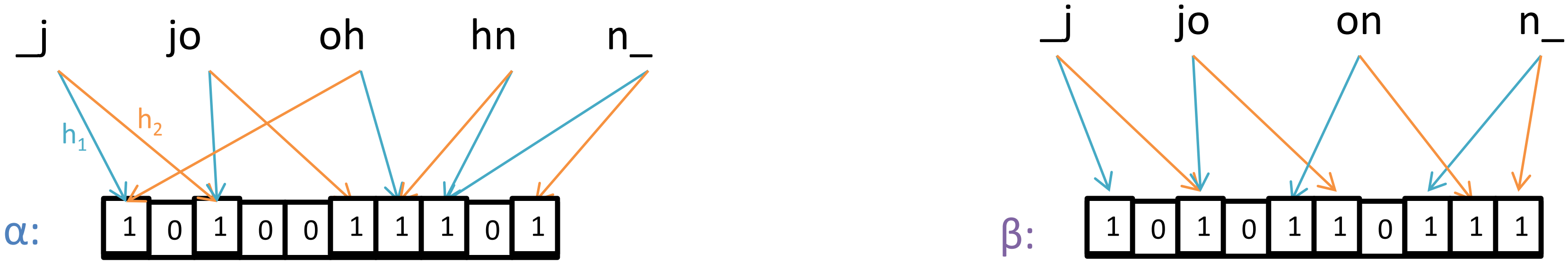
(Schnell et al 2009; Durham et al 2010)

Record a

john

Record b

jon



α: 1 0 1 0 0 1 1 1 0 1

β: 1 0 1 0 1 1 0 1 1 1

$$Dice\ coefficient = 2\left(\frac{|\alpha \cap \beta|}{|\alpha| + |\beta|}\right) = \frac{2 \times 5}{13} = 0.77$$

$where\ |*|\ is\ the\ number\ of\ bits\ set\ to\ 1\ in\ Bloom\ filter\ *$

Some Useful References

- W. Du and Z. Zhan. Using randomized response techniques for privacy preserving data mining. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2003: 505-510.
- E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin. Private medical record linkage with approximate matching. *Proceedings of the 2010 American Medical Informatics Association Annual Symposium*. 2010: 182-186.
- S. Grannis, J. Overhage, S. Hui, C. McDonald. Analysis of a probabilistic record linkage technique without human review. Proceedings of the 2003 American Medical Informatics Association Annual Symposium. 2003: 259-263.
- M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. IEEE Transactions on Information Technology in Biomedicine. 2008; 12(5): 606-617.
- M. Kantarcioglu, A. Inan, W. Jiang, and B. Malin. Formal anonymity models for efficient privacy-preserving joins. *Data and Knowledge Engineering*. 2009; 68(11): 1206-1223.
- A. Karr, X. Lin, J. Reiter, and A. Sanil. Secure regression in distributed databases. *J. Computational and Graphical Statistics.* 2005; 14: 263-279.
- J. Lombardo. The ESSENCE II disease surveillance test bed for the national capital area. Johns Hopkins APL Technical Digest. 2003; 24(4): 327-334.
- P. Paillier. Public-key cryptosystems based on composite degree-residuosity classes. *Lecture Notes in Computer Science: Advances in Cryptology – Eurocrypt '99.* 1999; 1592: 223-238.
- A. Shamir. How to share a secret. *Communications of the ACM*. 1979; 11: 612-613.
- L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness, & Knowledge-Based Systems.* 2002; 557-570.
- S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*. 1965; 60(309): 63-69.
- S. Wieland, C. Cassa, K. Mandl, and B. Berger. Revealing the spatial distribution of a disease while preserving privacy. Proceedings of the National Academy of Sciences. 2008; 105(46): 17608-17613.

Questions?

b.malin@vanderbilt.edu

Health Information Privacy Laboratory

http://www.hiplab.org/