PURDUE
UNIVERSITY

Health Data Privacy: Data Anonymization

Chris Clifton

18 March 2008

CERIAS
Center for Education and Research in Information Assurance and Security

The Tradeoff

- Healthcare Data is Sensitive
  - Embarassment
  - Economic
  - Legal
- Healthcare Data is Valuable
  - Research
  - Process optimization/improvement
  - Marketing

Anonymized Data

- HIPAA protects Individually Identifiable Health Information

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
Alice	47901	3/4/56	F	…
Bob	47904	4/5/67	M	…
Chris	47906	5/6/78	M	Schizophrenic

Anonymized Data

- HIPAA protects Individually Identifiable Health Information
  - Is this identifiable?
  - *Probably…*

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	47901	3/4/56	F	…
	47904	4/5/67	M	…
	47906	5/6/78	M	Schizophrenic

HIPAA: De-Identifying Data

- A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable
  - Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
  - Documents the methods and results of the analysis that justify such determination
- The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
  - Names, Location < 1st three digits of zip, dates < year, Tel/Fax/email/SSN/MRN/InsuranceID/Account/licence/VIN/License Plate Numbers, DeviceID, URL/IP, Biometric IDs, full-face photographs, any other unique identifiers; and
  - The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

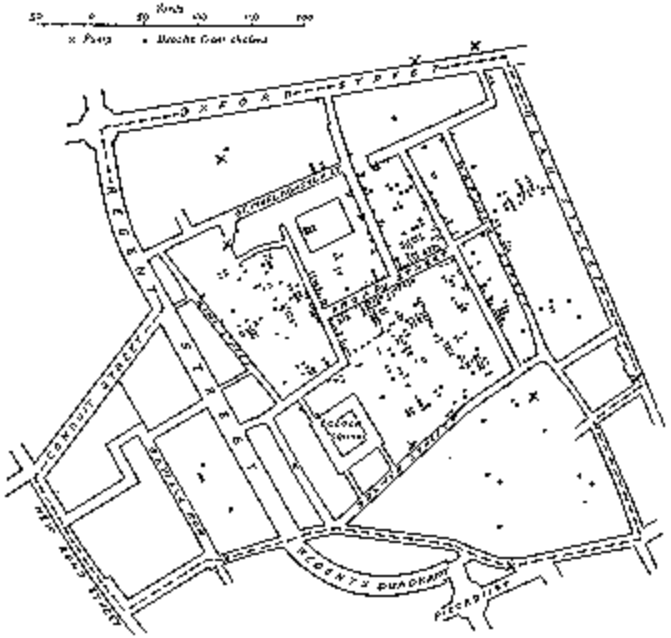
Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	479xx	56	F	…
	479xx	67	M	…
	479xx	78	M	Schizophrenic

Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?
- Dot chart by Dr. James Snow showing deaths from cholera in relation to the locations of public water pumps.
  - Observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump.



Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?
  - Is it enough?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	479xx	56	F	…
	479xx	67	M	…
	479xx	78	M	Schizophrenic

Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?
  - Is it enough?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	479xx	56	F	…
	479xx	67	M	Uses Marijuana for Pain
	479xx	78	M	Schizophrenic

Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?
  - Is it enough?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	479xx	56	F	Uses Marijuana for Phantom Pain
	479xx	67	M	Uses Marijuana for Pain
	479xx	78	M	Schizophrenic

Work in Privacy/Anonymity

- Definitions and Metrics
  - Risk-based
- Text Anonymization
  - Prevent Identifiability
  - Preserve meaning