Team for Research in Ubiquitous Secure Technology

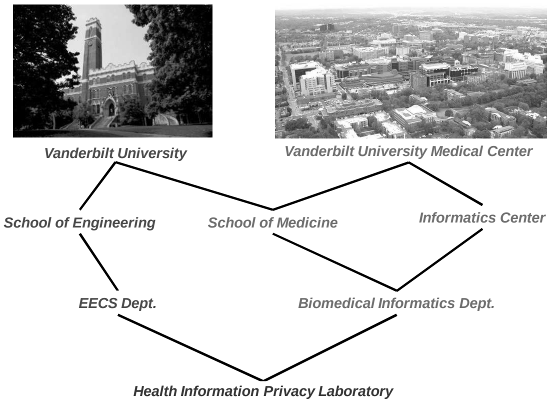
Women's Institute in Summer Enrichment

Building Systems to Support Health Information Privacy

Brad Malin, Ph.D.
Assistant Professor
Department of Biomedical Informatics, School of Medicine
Department of Computer Science, School of Engineering
Vanderbilt University

Disclaimer

- **Privacy** is an overloaded word
- Today: **privacy** in the context of a specific domain
- Healthcare
  - Health Insurance Portability & Accountability Act (HIPAA)
  - NIH Data Sharing Policy
  - NIH Genome Wide Association Study Data Sharing Policy

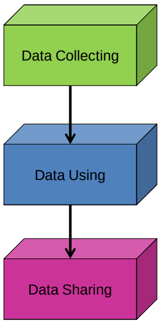

Vanderbilt University
Vanderbilt University Medical Center
School of Engineering
School of Medicine
Informatics Center
EECS Dept.
Biomedical Informatics Dept.
Health Information Privacy Laboratory

What's Going On?

- We study "privacy" in various operational realms
  - Primary Care
    - Clinical Information Systems Design
    - "Intelligent" Auditing
  - Secondary Uses
    - De-identification / Re-identification / Anonymization
    - Secure Data Integration and Analysis

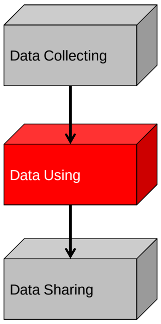
Privacy Everywhere

- We do not always control who gets, and has access to, our information
- Legally, however, data collectors may be required to maintain your privacy



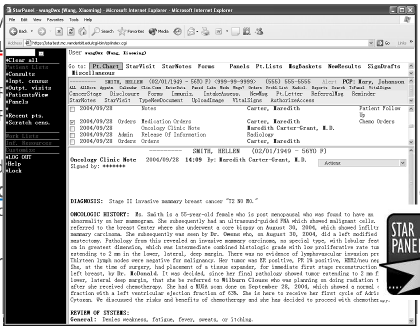
Privacy Everywhere

- Let's begin with data already in the system



Electronic Medical Records – Hooray!

- At V
  - L
  - I
  - F
  - F
- And
  - C ent
    c
- Incr
  ove


StarPanel - wang0wx (Wang, Xiaoming) - Microsoft Internet Explorer - Microsoft Internet Explorer

File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://startest.mc.vanderbilt.edu/cgi-bin/sp/index.cgi
Go Links
+Clear all
Patient Lists
+Consults
+Inpt. census
+Outpt. visits
+PatientsView
+Panels
+Recent pts.
+Scratch cens.
Work Lists
Inf. Resources
Customize
+LOG OUT
+Help
+Lock
Go to: Pt.Chart StarVisit StarNotes Forms Panels Pt.Lists MsgBaskets NewResults SignDrafts Miscellaneous
ALL AllDocs Appntm. Calendar Clin.Comm EnterData Faxed Labs Meds Msgs? Orders Probl.List Radiol. Reports Search ToPanel VitalSigns
CancerStage Disclosure Forms Immuniz. IntakeAssess. NewMsg Pt.Letter ReferralMsg Reminder
StarNotes StarVisit TypeNewDocument UploadImage VitalSigns AuthorizeAccess
Actions:

User wang0wx (Wang, Xiaoming)

--------- SMITH, HELLEN (02/01/1949 - 56YO F) <999-99-9999> (555) 555-5555 Alert PCP: Mary, Johanson

2004/09/28		Notes	Carter, Maredith	Patient Follow Up
2004/09/28	Orders	Medication Orders	Carter, Maredith	Chemo Orders
2004/09/28		Oncology Clinic Note	Maredith Carter-Grant, M.D.	
2004/09/28	Admin	Release Of Information	Radiology	
2004/09/28	Orders	Orders	Carter, Maredith	

---------- SMITH, HELLEN (02/01/1949 - 56YO F)

Oncology Clinic Note 2004/09/28 14:09 By: Maredith Carter-Grant, M.D.
Signed by: *******

DIAGNOSIS: Stage II invasive mammary breast cancer "T2 N0 M0."

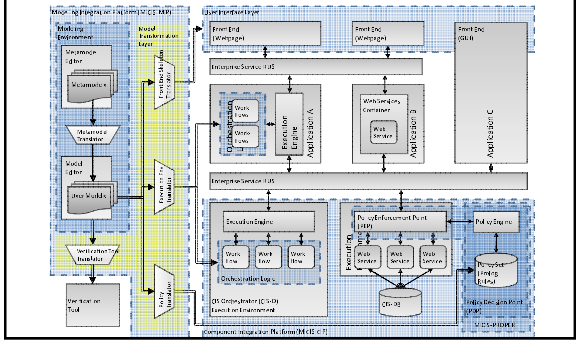
ONCOLOGIC HISTORY: Ms. Smith is a 55-year-old female who is post menopausal who was found to have an abnormality on her mammogram. She subsequently had an ultrasound-guided FNA which showed malignant cells. referred to the breast Center where she underwent a core biopsy on August 30, 2004, which showed infiltr mammary carcinoma. She subsequently was seen by Dr. Owens who, on August 30, 2004, did a left modified mastectomy. Pathology from this revealed an invasive mammary carcinoma, no special type, with lobular feat cm in greatest dimension, which was intermediate combined histologic grade with low proliferative rate tum extending to 2 mm in the lower, lateral, deep margin. There was no evidence of lymphovascular invasion pre Thirteen lymph nodes were negative for malignancy. Her tumor was ER positive, PR 1% positive, HER2/neu neg She, at the time of surgery, had placement of a tissue expander, for immediate first stage reconstruction left breast, by Dr. McDonald. It was decided, since her final pathology showed tumor extending to 2 mm f lower, lateral deep margin, that she be referred to Wilburn Clouse who was planning on doing radiation t after she received chemotherapy. She had a MUGA scan done on September 28, 2004, which showed a normal e fraction with a left ventricular ejection fraction of 63%. She is here to receive her first cycle of Adria Cytoxan. We discussed the risks and benefits of chemotherapy and she has decided to proceed with chemotherapy.

REVIEW OF SYSTEMS:
General: Denies weakness, fatigue, fever, sweats, or itching.

STAR PANEL

Clinical Information Systems Design

(Duncavage 2007; Mathe et al, 2008; Werner et al, 2007, 2008)

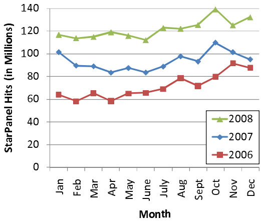


Surveillance (Paulett, Malin)

- Very little role-based access control in large academic medical centers! (why?)
- Most auditing is done manually! (why?)

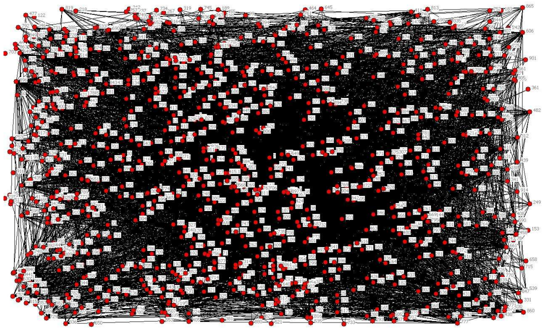
> 1.5 million patient records

> 20,000 authorized users



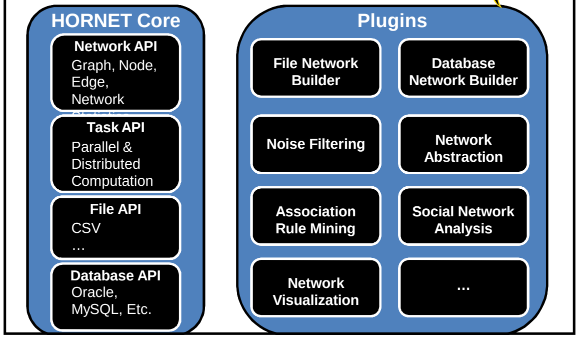
Jan 1, 2006

- Users linked if accessed 1 common patient (~ 900 users, 2000 patients)



HORNET: Healthcare Organizational Research Toolkit

(http://code.google.com/p/hornet/)


HORNET Core

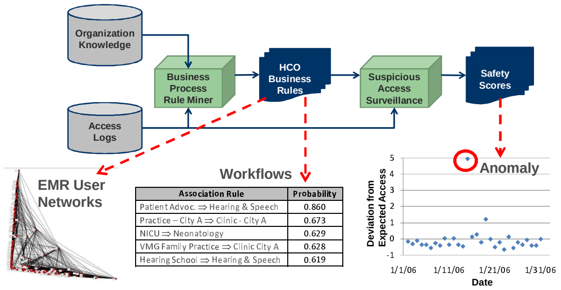
- Network API
  Graph, Node, Edge, Network
- Task API
  Parallel & Distributed Computation
- File API
  CSV
  …
- Database API
  Oracle, MySQL, Etc.

Plugins

- File Network Builder
- Database Network Builder
- Noise Filtering
- Network Abstraction
- Association Rule Mining
- Social Network Analysis
- Network Visualization
- …

Learning Policies and Anomalies from EMR Access Logs

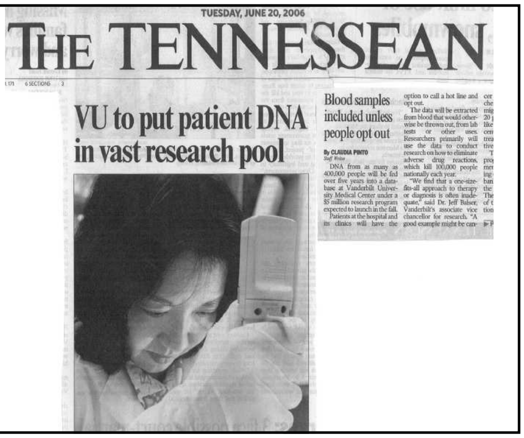
- Extracts patterns from medical record access logs to model policies & detect "privacy" violations.



Association Rule	Probability
Patient Advoc. ⇒ Hearing & Speech	0.860
Practice – City A ⇒ Clinic - City A	0.673
NICU ⇒ Neonatology	0.629
VMG Family Practice ⇒ Clinic City A	0.628
Hearing School ⇒ Hearing & Speech	0.619

What's Going On?

- "Privacy" in various operational settings
  - Primary Care
    - ~~Clinical Information Systems Design~~
    - ~~"Intelligent" Auditing~~
  - Secondary Uses
    - De-identification / Re-identification / Anonymization
    - Secure Data Integration and Analysis



TUESDAY, JUNE 20, 2006

THE TENNESSEAN

VU to put patient DNA in vast research pool

Blood samples included unless people opt out

By CLAUDIA PINTO
Staff Writer

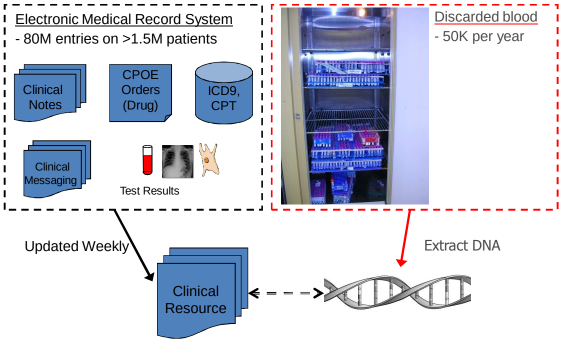
DNA from as many as 400,000 people will be fed over five years into a database at Vanderbilt University Medical Center under a $5 million research program expected to launch in the fall.

Patients at the hospital and its clinics will have the option to call a hot line and opt out.

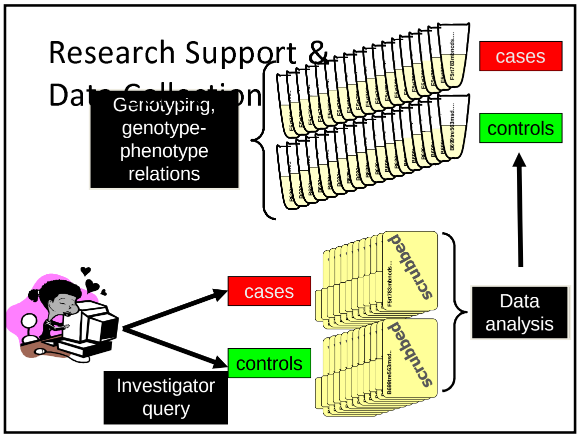
The data will be extracted from blood that would otherwise be thrown out, from lab tests or other uses. Researchers primarily will use the data to conduct research on how to eliminate adverse drug reactions, which kill 100,000 people nationally each year.

"We find that a one-size-fits-all approach to therapy or diagnosis is often inadequate," said Dr. Jeff Balser, Vanderbilt's associate vice chancellor for research. "A good example might be can-

Information Integration



Research Support & Data Collection



Holy Hand Grenades! How Did You…

- Initially an institutionally funded project
- Office for Human Research Protections designation as Non-Human Subjects Research under 45 CFR 46 ("HIPAA Common Rule")*
  - ***Samples & data not linked to identity***
  - Conducted with IRB & ethics oversight

*Roden D et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008; 84(3): 362-369.

HIPAA

(the **ELEPHANT** in the room)

- Health Insurance Portability & Accountability Act
- Resolve state laws hampering standardization, transfer, & sharing of health information
- "covered entity" cannot use or disclose protected health information (PHI)
  - data "explicitly" linked to a particular individual, or
  - could reasonably be expected to allow individual identification

HIPAA - Secondary Data Sharing

- Safe Harbor
- Limited Release
- Statistical or Scientific Standard

HIPAA Safe Harbor

- Data that can be given away without oversight
- Requires removal of eighteen attributes
  - Names / Initials
  - Street address, city, county, precinct code and equivalent geocodes
  - All elements of dates, except year, and all ages over 89
  - #'s: Phone, Fax, Social Security, Medical Record, Health Plan ID, Account, License, Serial, Device
  - Web: Email, URL, IP addresses
  - Biometric identifiers: finger, voice prints
  - Full face photo images and comparable images
  - Any other unique identifying number, characteristic, or code
    - A code is an identifier if the person holding the coded data can re-identify the individual

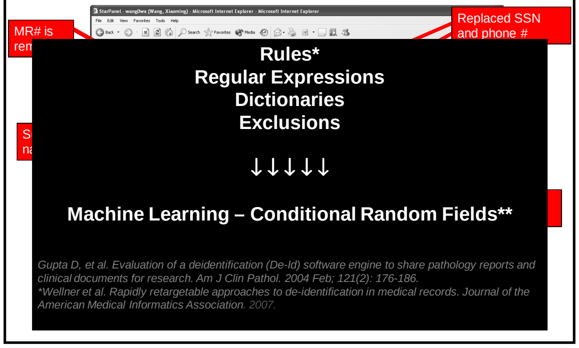
HIPAA Limited Dataset

- Includes more specific information than Safe Harbor
- Can include
  - Dates of birth, death, service
  - Geographic Info: Town, Zip code, County
- **Requires Contract:** Research entity provides assurances that it will not use or disclose the information for purposes other than research and will not identify or contact the individuals who are the subjects

HIPAA Statistical / Scientific Standard

- Certify via "generally accepted statistical and scientific principles and methods, that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by the anticipated recipient to identify the subject of the information."
- "Must document the methods and results of the analysis that justify such a determination"
- "Must not disclose the key or other mechanism that would have enabled the information to be re-identified"
  - includes pseudo-random number algorithms and seed values

"Scrubbing" Medical Records


StarPanel - wang0wx (Wang, Xiaoming) - Microsoft Internet Explorer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
MR# is rem
Replaced SSN and phone #
S na

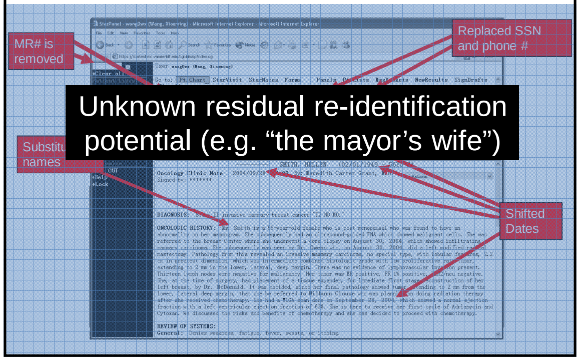
Rules*
Regular Expressions
Dictionaries
Exclusions

↓↓↓↓↓

Machine Learning – Conditional Random Fields**

Gupta D, et al. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004 Feb; 121(2): 176-186.
**Wellner et al. Rapidly retargetable approaches to de-identification in medical records. Journal of the American Medical Informatics Association. 2007.*

"Scrubbed" Medical Record


StarPanel - wang0wx (Wang, Xiaoming) - Microsoft Internet Explorer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
https://startest.mc.vanderbilt.edu/cgi-bin/sp/index.cgi
MR# is removed
Replaced SSN and phone #
Substitu names
Shifted Dates
Clear all
OUT
Help
Lock
Actions

Unknown residual re-identification potential (e.g. "the mayor's wife")

User wang0wx (Wang, Xiaoming)

Go to: Pt.Chart StarVisit StarNotes Forms Panels PtLists MsgBaskets NewResults SignDrafts

SMITH, HELLEN (02/01/1949 - 56YO)

Oncology Clinic Note 2004/09/28 By: Maredith Carter-Grant, M.D.
Signed by: *******

DIAGNOSIS: Stage II invasive mammary breast cancer "T2 N0 M0."

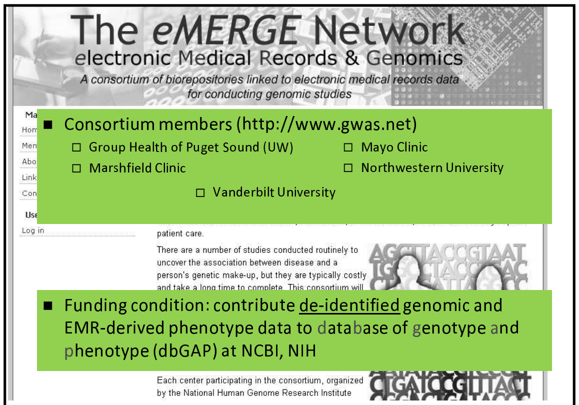
ONCOLOGIC HISTORY: Ms. Smith is a 55-year-old female who is post menopausal who was found to have an abnormality on her mammogram. She subsequently had an ultrasound-guided FNA which showed malignant cells. She was referred to the breast Center where she underwent a core biopsy on August 30, 2004, which showed infiltrating mammary carcinoma. She subsequently was seen by Dr. Owens who, on August 30, 2004, did a left modified radical mastectomy. Pathology from this revealed an invasive mammary carcinoma, no special type, with lobular features, 2.2 cm in greatest dimension, which was intermediate combined histologic grade with low proliferative rate tumor, extending to 2 mm in the lower, lateral, deep margin. There was no evidence of lymphovascular invasion present. Thirteen lymph nodes were negative for malignancy. Her tumor was ER positive, PR 1% positive, HER2/neu negative. She, at the time of surgery, had placement of a tissue expander, for immediate first stage reconstruction of her left breast, by Dr. McDonald. It was decided, since her final pathology showed tumor extending to 2 mm from the lower, lateral deep margin, that she be referred to Wilburn Clouse who was planning on doing radiation therapy after she received chemotherapy. She had a MUGA scan done on September 28, 2004, which showed a normal ejection fraction with a left ventricular ejection fraction of 63%. She is here to receive her first cycle of Adriamycin and Cytoxan. We discussed the risks and benefits of chemotherapy and she has decided to proceed with chemotherapy.

REVIEW OF SYSTEMS:
General: Denies weakness, fatigue, fever, sweats, or itching.

Technology + Policy

- Databank access restricted to Vanderbilt employees
- Must sign use agreement that prohibits “re-identification”
- Operations Advisory Board and Institutional Review Board approval needed for each project
- All data access logged and audited per project

To Vanderbilt …. and Beyond



The *eMERGE* Network

electronic Medical Records & Genomics

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies

patient care.

There are a number of studies conducted routinely to uncover the association between disease and a person's genetic make-up, but they are typically costly and take a long time to complete. This consortium will

Each center participating in the consortium, organized by the National Human Genome Research Institute

- Consortium members (http://www.gwas.net)
  - □ Group Health of Puget Sound (UW)
  - □ Mayo Clinic
  - □ Marshfield Clinic
  - □ Northwestern University
  - □ Vanderbilt University
- Funding condition: contribute <u>de-identified</u> genomic and EMR-derived phenotype data to database of genotype and phenotype (dbGAP) at NCBI, NIH

Data Sharing Policies

- Feb ‘03: National Institutes of Health Data Sharing Policy
  - ***“data should be made as widely & freely available as possible”***
  - ***researchers who receive >= $500,000*** *must develop a data sharing plan* ***or describe why data sharing is not possible***
  - Derived data must be shared in a manner that is devoid of “identifiable information”

- Aug ‘06: NIH Supported Genome-Wide Association Studies Policy
  - □ Researchers who received >= $0 for GWAS

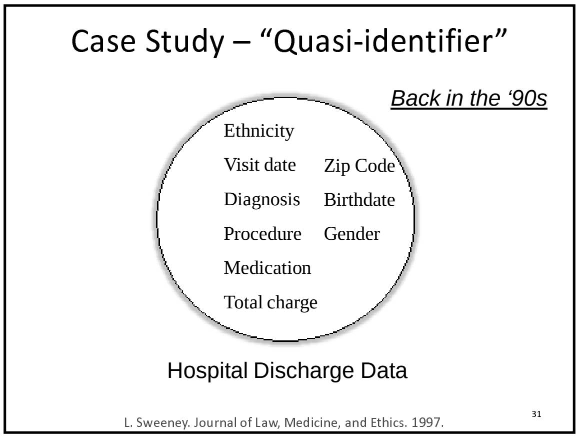
The Face that Launched a Thousand Ships

29

Healthcare Reform At Work

- In 1997, approx. 44 of 50 states collected and disseminated hospital discharge data
- In 2005, approx. 47 of 50 states “ “
- Attributes recommended by *National Association of Health Data Organizations* for disclosure
  - Patient Zip Code
  - Patient Birth Date
  - Patient Gender
  - Patient Racial Background
  - Patient Number
  - Visit Date
  - □ Principle Diagnosis Codes (ICD-9)
  - □ Procedure Codes
  - □ Physician ID Number
  - □ Physician Zip Code
  - □ Total Charges

J. Schoenman et al. The value of hospital discharge databases. NORC & NAHDO. 2005. http://www.hcup-us.ahrq.gov/reports/final_report.pdf



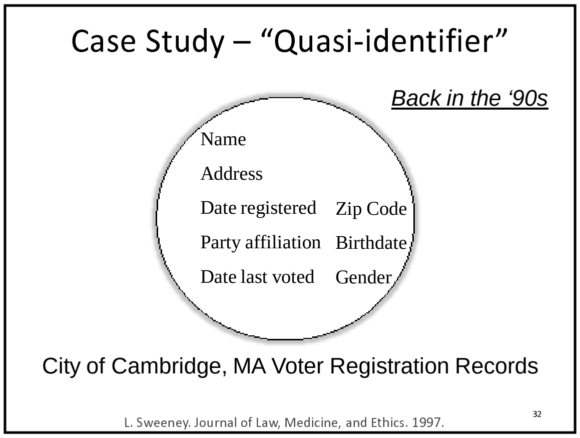
Case Study – "Quasi-identifier"

Back in the '90s

Hospital Discharge Data

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

31



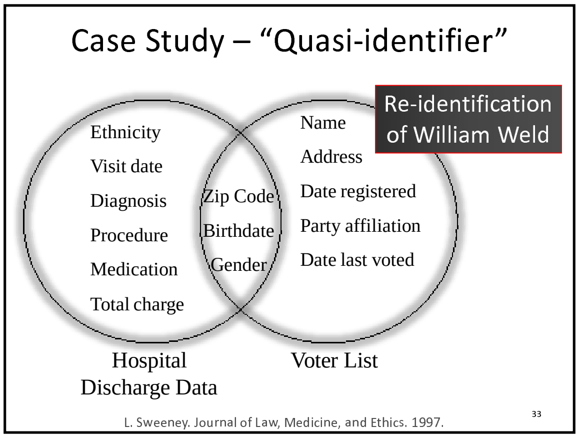
Case Study – "Quasi-identifier"

Back in the '90s

City of Cambridge, MA Voter Registration Records

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

32

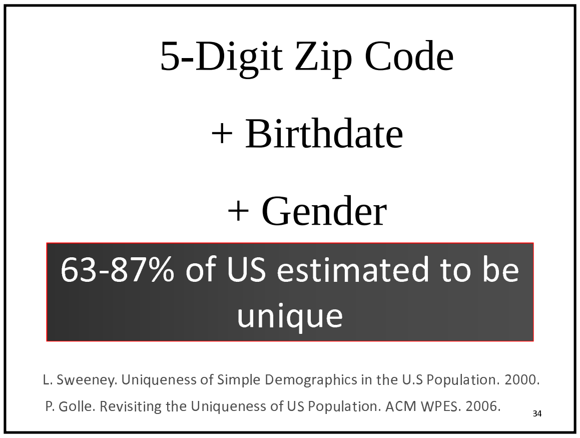


Case Study – "Quasi-identifier"

Re-identification of William Weld

L. Sweeney. Journal of Law, Medicine, and Ethics. 1997.

33



5-Digit Zip Code

+ Birthdate

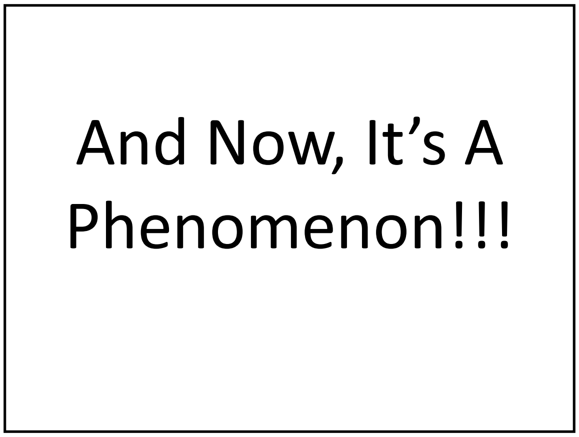
+ Gender

63-87% of US estimated to be unique

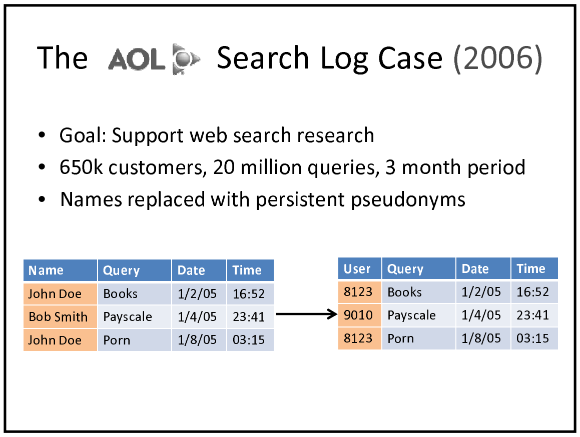
L. Sweeney. Uniqueness of Simple Demographics in the U.S Population. 2000.

P. Golle. Revisiting the Uniqueness of US Population. ACM WPES. 2006.

34



And Now, It's A Phenomenon!!!



The AOL Search Log Case (2006)

- Goal: Support web search research
- 650k customers, 20 million queries, 3 month period
- Names replaced with persistent pseudonyms

Name	Query	Date	Time
John Doe	Books	1/2/05	16:52
Bob Smith	Payscale	1/4/05	23:41
John Doe	Porn	1/8/05	03:15

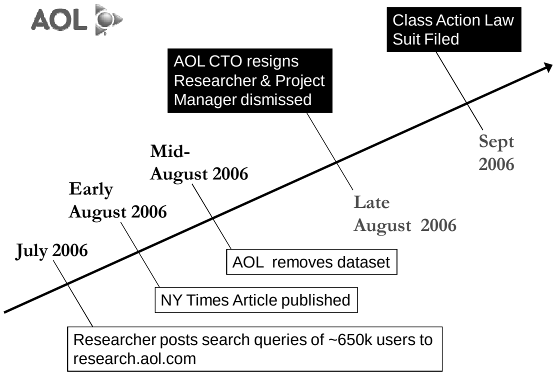
→

User	Query	Date	Time
8123	Books	1/2/05	16:52
9010	Payscale	1/4/05	23:41
8123	Porn	1/8/05	03:15

Barbaro & Zeller. A face exposed for AOL searcher no. 4417749. **New York Times**. Aug 9, 2006.

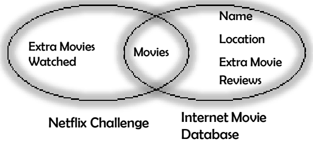
User 4417749 issued hundreds of

Thelma Arnold & Dudley


AOL
Class Action Law Suit Filed
AOL CTO resigns
Researcher & Project Manager dismissed
Mid-August 2006
Sept 2006
Early August 2006
Late August 2006
July 2006
AOL removes dataset
NY Times Article published
Researcher posts search queries of ~650k users to research.aol.com

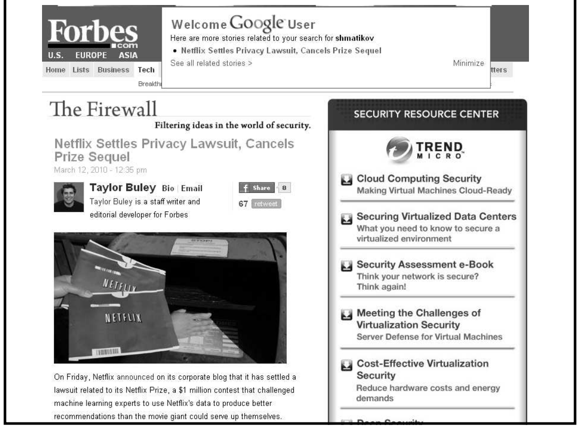
The Netflix Challenge (2008-2009)

- Netflix published movie selections of ~450,000 pseudonymized subscribers
- Re-identification via uniqueness of movie combinations



- Class action filed December 2009

A. Narayanan & V. Shmatikov. IEEE Security and Privacy Conference. 2008.



Welcome Google User

Here are more stories related to your search for **shmatikov**

- Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

See all related stories >

Minimize

The Firewall

Filtering ideas in the world of security.

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

March 12, 2010 - 12:35 pm

Taylor Buley Bio | Email

Taylor Buley is a staff writer and editorial developer for Forbes

67

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

SECURITY RESOURCE CENTER

TREND MICRO

Cloud Computing Security
Making Virtual Machines Cloud-Ready

Securing Virtualized Data Centers
What you need to know to secure a virtualized environment

Security Assessment e-Book
Think your network is secure? Think again!

Meeting the Challenges of Virtualization Security
Server Defense for Virtual Machines

Cost-Effective Virtualization Security
Reduce hardware costs and energy demands

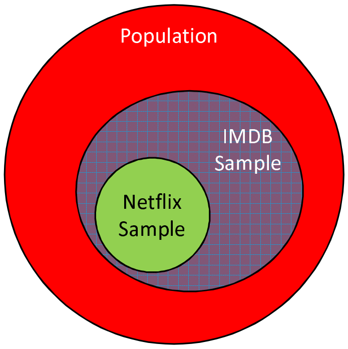
Deep Security

Samples
Samples
Samples
Samples
Samples

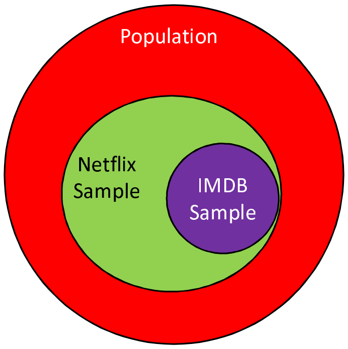
and

Populations

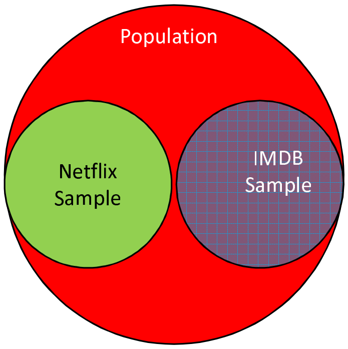
Re-identification ?



Re-identification ?

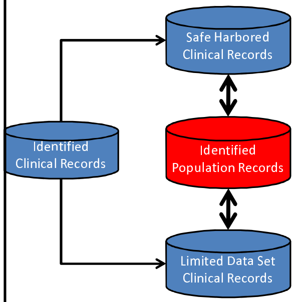


Re-identification ?



Now Back to Your Regularly Scheduled Programming

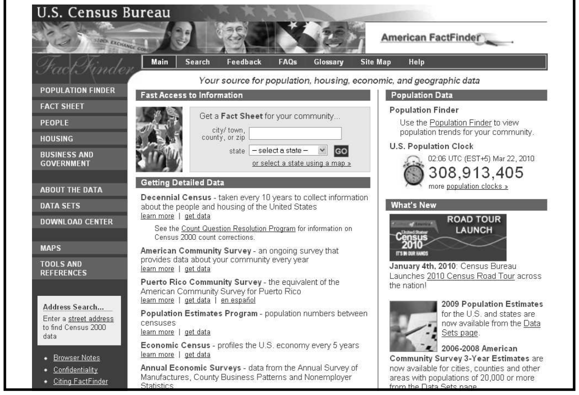
Attacks on Demographics


Safe Harbored Clinical Records
Identified Clinical Records
Identified Population Records
Limited Data Set Clinical Records

- We use population estimates according the US Census
- It's not perfect, but it's a start.

K. Benitez and B. Malin. Evaluating re-identification risk with respect to the HIPAA privacy policies. Journal of the American Medical Informatics Association. 2010; 17: 169-177.

The Census? http://factfinder.census.gov/



Your source for population, housing, economic, and geographic data

Fast Access to Information

Get a **Fact Sheet** for your community...

city/ town, county, or zip

state – select a state –

or select a state using a map »

Getting Detailed Data

Decennial Census - taken every 10 years to collect information about the people and housing of the United States
learn more | get data

See the Count Question Resolution Program for information on Census 2000 count corrections.

American Community Survey - an ongoing survey that provides data about your community every year
learn more | get data

Puerto Rico Community Survey - the equivalent of the American Community Survey for Puerto Rico
learn more | get data | en español

Population Estimates Program - population numbers between censuses
learn more | get data

Economic Census - profiles the U.S. economy every 5 years
learn more | get data

Annual Economic Surveys - data from the Annual Survey of Manufactures, County Business Patterns and Nonemployer Statistics

Address Search...

Enter a street address to find Census 2000 data

- Browser Notes
- Confidentiality
- Citing FactFinder

Population Data

Population Finder

Use the Population Finder to view population trends for your community.

U.S. Population Clock

02:06 UTC (EST+5) Mar 22, 2010

308,913,405

more population clocks »

What's New

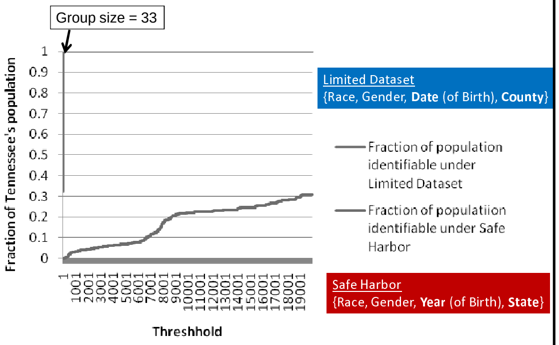
January 4th, 2010: Census Bureau Launches 2010 Census Road Tour across the nation!

2009 Population Estimates for the U.S. and states are now available from the Data Sets page.

2006-2008 American Community Survey 3-Year Estimates are now available for cities, counties and other areas with populations of 20,000 or more from the Data Sets page

Beyond "unique"

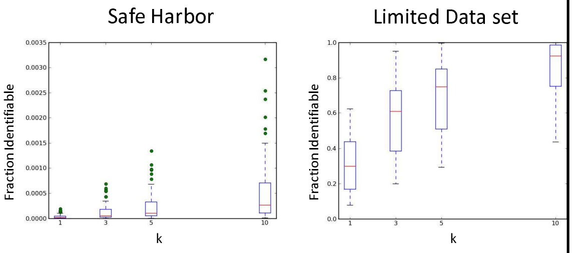
Case Study: Tennessee



Limited Dataset
{Race, Gender, **Date** (of Birth), **County**}

Safe Harbor
{Race, Gender, **Year** (of Birth), **State**}

All U.S. States


Safe Harbor

Limited Data set

Policy Analysis via “Trust Differential”…

$$\frac{\underline{\text{Risk(Limited Dataset)}}}{\text{Risk(Safe Harbor)}}$$

- Uniques
  - *Delaware’s risk increases by a factor* *~1,000*
  - *Tennessee’s* *“ “ “ “* *~2,300*
  - *Illinois’s “* *“ “ “ “* *~65,000*
- ≤20,000
  - *Delaware’s risk does not increase*
  - *Tennessee’s risk increases by a factor of ~8*
  - *Illinois’s risk increases by a factor of ~37*

State Policy

	IL	MN	TN	WA	WI
WHO???	Registered Political Committees (ANYONE – In Person)	MN Voters	Anyone	Anyone	Anyone
Format	Disk	Disk	Disk	Disk	Disk
Cost	$500	$46; “use ONLY for elections, political activities, or law enforcement”	$2500	$30	$12,500
Voter ID	●	●	●	●	●
Name	●	●	●	●	●
Address	●	●	●	●	●
Voter Status	●	●	●	●	●
District Information	●	●	●	●	●
Election History	●	●	●	●	●
Date of Birth	●	○	●	●	
Date of Registration	●	●	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

Cost?

State	Limited Dataset		Safe Harbor	
	Marker Risk	Cost per Re-id	Total Risk	Cost per Re-id
VA	3159764	$0	221	$0
NY	2905697	$0	221	$0
SC	2231973	$0	1386	$0
WI	72	$174	2	$6,250
WV	55	$309	1	$17,000
NH	10	$827	1	$8,267

A Couple of Parting Thoughts

- The application of technology must be considered within the systems and operational processes they will be applied
- One person’s vulnerability is another person’s armor (variation in risks)
- It is possible to inject privacy into health information systems – but it must be done early (see “privacy by design)!
- Sometimes theory needs to be balanced with practicality