

The (Explicit & Implicit) Costs of Data Privacy

Bradley Malin, Ph.D.

July 11, 2019

Vanderbilt University

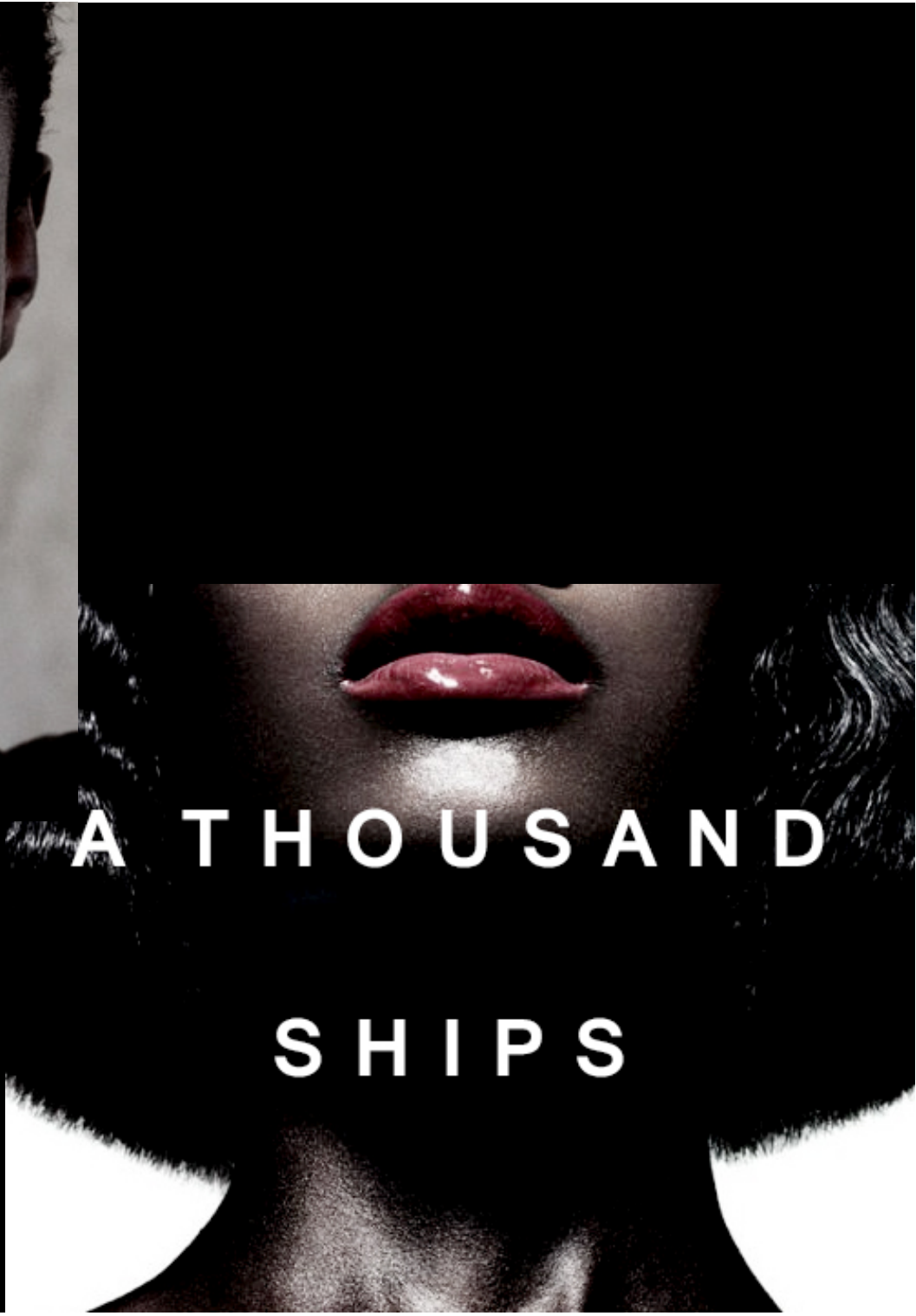
Professor of Biomedical Informatics, Biostatistics, & Computer Science

Co-Director, Center for Genetic Privacy & Identity in Community Settings (GetPreCiSe)

Co-Director, Center for Health Data Science (Heads)



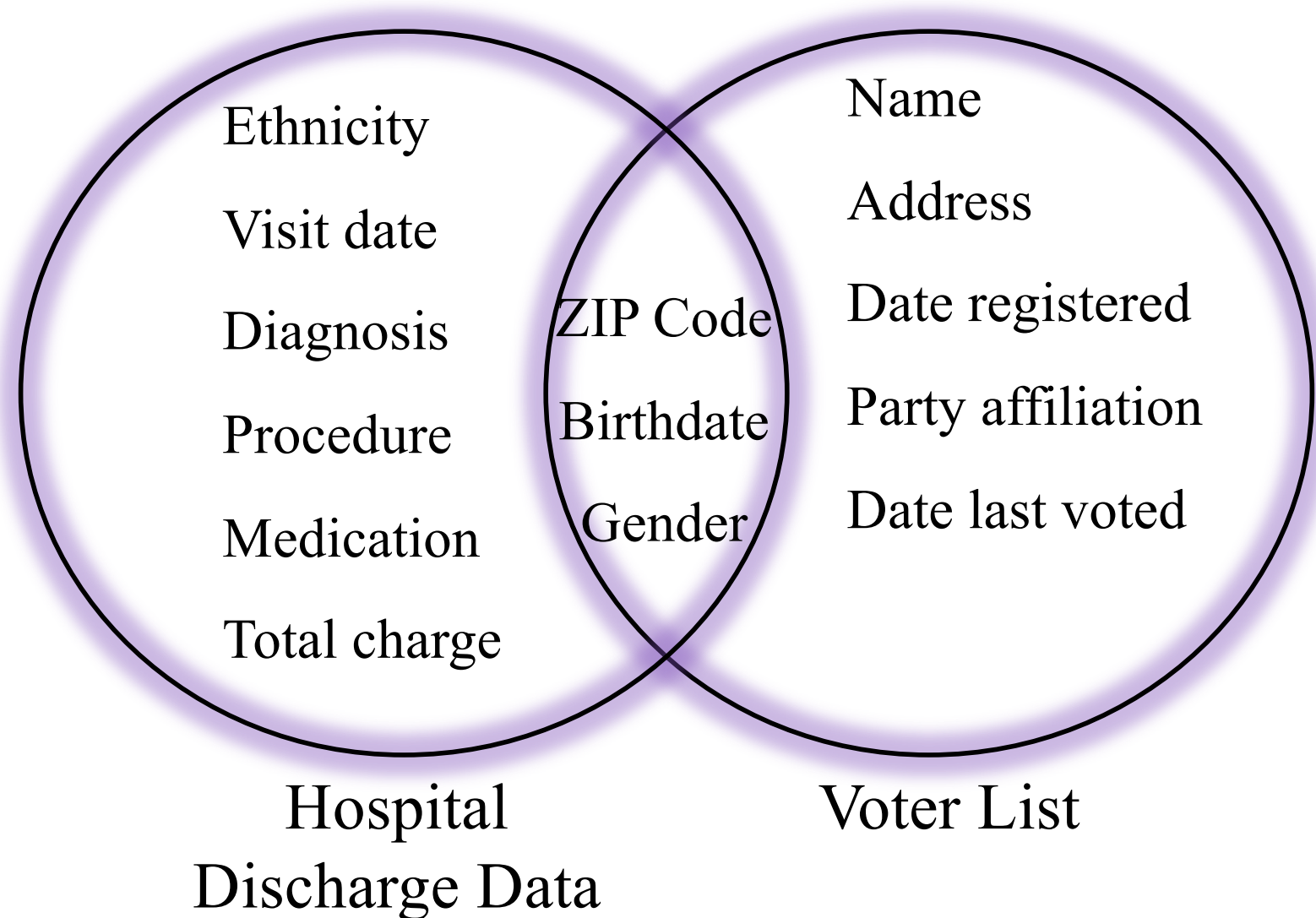
THE FACE
THAT
LAUNCHED



A THOUSAND
SHIPS



A “Quasi-identifier” Conundrum



5-Digit ZIP

+ Birthdate

+ Gender

63-87% of USA
estimated to be unique

Set the
World
on



The AOL Search Log Case (2006)

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

Goal: Support web information retrieval research

- 650 K customers, 20 M queries, 3 MONTH period
- Names replaced with persistent pseudonyms

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



Thelma Arnold
&
Dudley



AOL CTO resigns;
Researcher & Project
Manager dismissed

Class Action Law
Suit Filed

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

- Published in
- Re-identification

- Class action

Forbes
U.S. EUROPE ASIA
Home Lists Business **Tech**
Breakth

Welcome **Google** User
Here are more stories related to your search
• **Netflix Settles Privacy Lawsuit, Cancels**
See all related stories >

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

 Share 8
67 retweet



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.

2009)

pseudonyms
combinations

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study



Adam Tanner Contributor ⓘ

Apr 25, 2013, 03:47pm • 22,581 views

f

🐦

in

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

From the onset, the [Personal Genome Project](#), set up by Harvard Medical School Professor of Genetics George Church, has warned participants of the risk that someone someday could identify them, meaning anyone could look up the intimate medical histories that many have posted along with their genome data. That day arrived on Thursday.



Harvard Professor Latanya Sweeney

MUST READ [WINDOWS 10: THESE TWO NEW BUILDS DELIVER LOTS OF BUG FIXES](#)

Re-identification possible with Australian de-identified Medicare and PBS open data

Using publicly known information, a team of researchers from the University of Melbourne have claimed to re-identify seven prominent Australians in an open medical dataset.



By [Chris Duckett](#) | December 18, 2017 -- 01:01 GMT (17:01 PST) | Topic: [Security](#)

1

f 3

in



The dataset containing historic longitudinal medical billing records of one-tenth of all Australians, approximately 2.9 million people, has been found to be re-identifiable by a team from the University of Melbourne, with information such as child births and professional sportspeople undergoing surgery to fix injuries often made public.

The team, consisting of Dr Chris Culnane, Dr Benjamin Rubinstein, and Dr Vanessa Teague, warned that they expect similar results with other data held by the government, such as Census data, tax records, mental health records, penal data, and Centrelink data.

"We found that patients can be re-identified, without decryption, through a process of linking the unencrypted parts of the record with known information about the individual such as medical procedures and year of birth," Dr Culnane said.

"This shows the surprising ease with which de-identification can fail, highlighting the risky

[Your Favorite Feature] Distinguishes You!!

- Demographics (Sweeney '97; Bacher '02; Golle '06; El Emam '08; Koot '10; Li '11, Sweeney '13, Sweeney '17, Soo '18)
- Diagnosis Codes (Loukides '10; Tamersoy '10, '12; Heatherly '16)
- Laboratory Tests (Cimino '12; Atreya '13)
- DNA (Malin '00, Lin '04; Homer '08; Gymrek '13, Ayday '14, Huttenhower '15; Shringapure '15; Lippert '17, Erlich '18)
- Pedigree Structure (Malin '06, Ayday '13)
- RNA (Backes '16a; Backes '16b)
- Proteome (Li '16)
- Health Survey Responses (Solomon '12)

- Location Visits and Mobility Traces (Malin '04; Golle '09; El Emam '11; de Montjoye '15; Kondor '17; Murakami '17)
- Movie Reviews (Narayanan '08)
- Social Network Structure (Backstrom '07; Narayanan '09; Yang '12; Cecaj '14, '16)
- Search Queries (Barbaro '06)
- Internet Browsing (Malin '05; Eckersley '10; Banse '11; Herrmann '12, Olejnik '12; Kirchler '16; Riederer '16)
- Smart Utility Meter Usage (Buchmann '12; Faisal '15; Tudor '15)

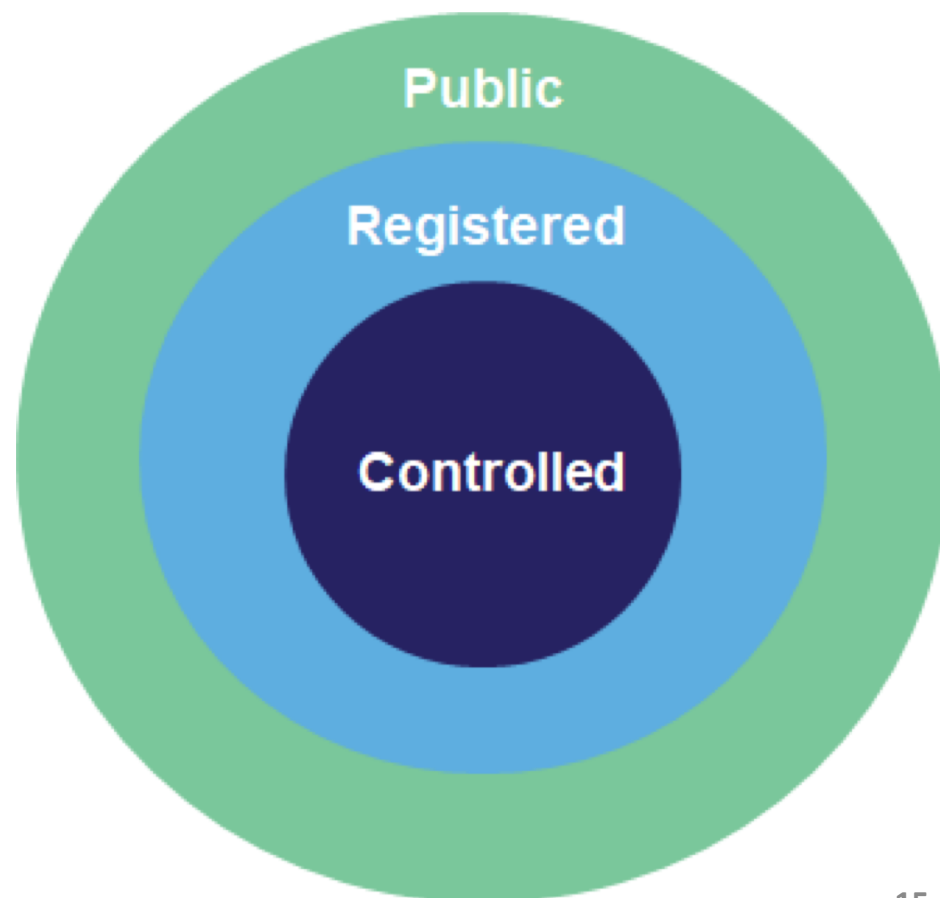
Possible

doesn't imply

Probable

Tiered Levels of Access

- Public
 - Can be accessed without logging in
 - Summary statistics only
- Sandbox Environments (on Google Cloud)
 - Registered
 - Individual level records with minimal risk to participant identification
 - Controlled
 - Individual level records with more risk to participant identification, but expected to be low



<https://databrowser.researchallofus.org/>

atabrowser.researchallofus.org

Sign On Log In to Fidelity.com ACM TOPS Eater Nashville V Medical Student Ca... Pr Vanderbilt Applicati... Vanderbilt IRB SunTrust Bank | Per... Malin, Bradley (CHI... McKesson Specialty... McKesson Specialty... Inbox (1) - brad.mal... Econometrix

Data based on Curated Data Repository (CDR) dated 11/13/2018 with 116,460 total participants.

FAQsIntroductory VideosUser Guide

EHR Domains:

Conditions

15,679

medical concepts

36,260 participants in this domain

View Top 100

Drug Exposures

14,967

medical concepts

33,440 participants in this domain

View Top 100

Measurements

7,733

medical concepts

32,480 participants in this domain

View Top 100

Procedures

13,229

medical concepts

35,320 participants in this domain

View Top 100

Survey Questions:

The Basics

14

survey questions

104,440 participants in this domain

Survey includes participant demographic information.

View Complete Survey

Overall Health

16

survey questions

101,420 participants in this domain

Survey provides information about how participants report levels of individual health.

View Complete Survey

Lifestyle

7

survey questions

100,460 participants in this domain

Survey includes information on participant smoking, alcohol and recreational drug use.

View Complete Survey

Program Physical Measurements:

Physical Measurements

Participants have the option to provide a standard set of physical measurements as part of the enrollment process ("program physical measurements").

8

physical measurements

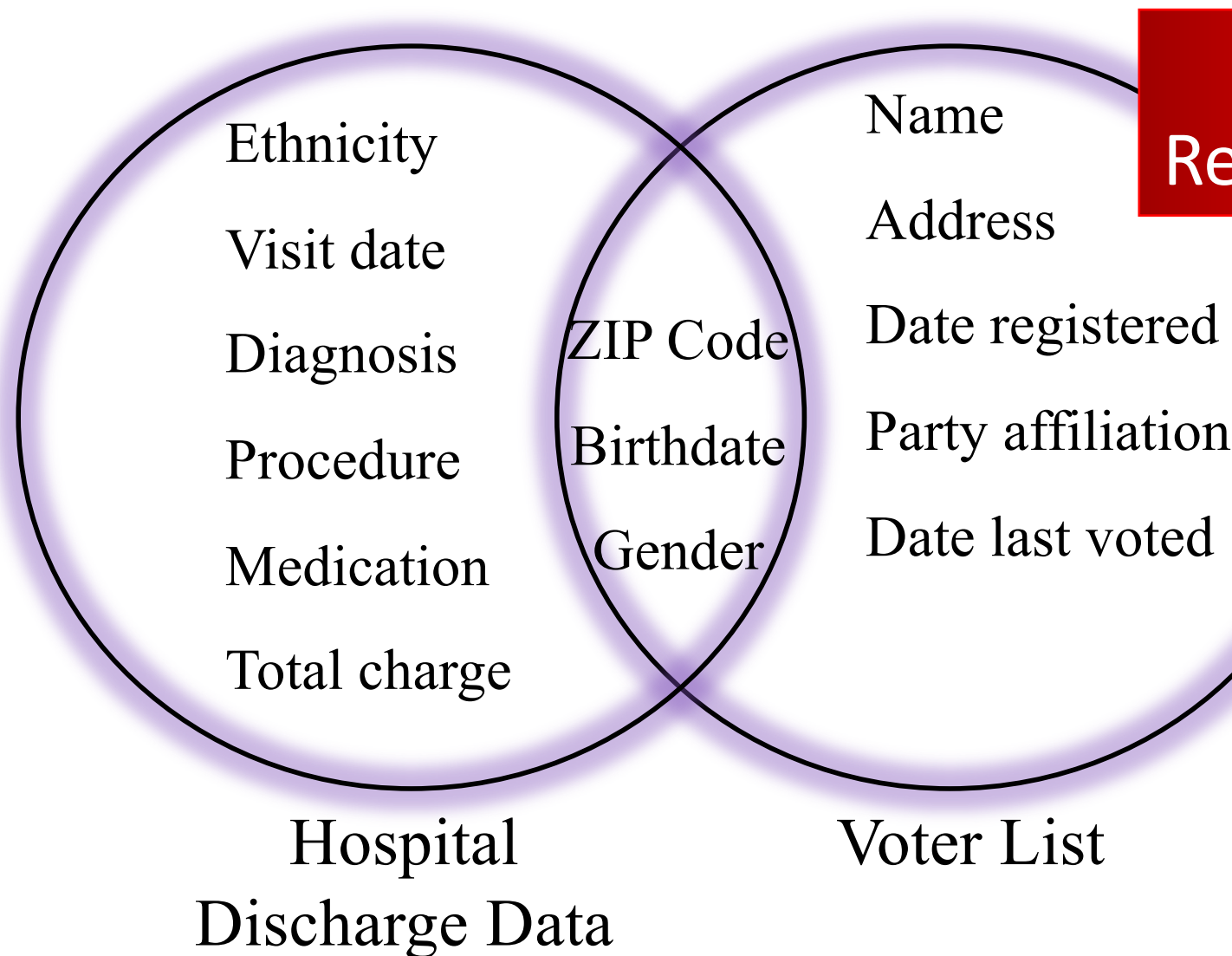
View Program Physical Measurements



We are Driven By Incentives

(under rational assumptions)

Remember Me?



High Profile
Re-identification



The Cost of Demographics Varies! (Voter Registration Lists)

	Illinois	Minnesota	Tennessee	Washington	Wisconsin
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
Name	●	●	●	●	●
Address	●	●	●	●	●
Date of Birth	●	○	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk ???</i>

Attack Strategy A
<i>Utility A</i>
<i>Risk A</i>

Attack Strategy B
<i>Utility B</i>
<i>Risk B</i>

⋮

Attack Strategy C
<i>Utility C</i>
<i>Risk C</i>

Strategies:

- Generalize Demographics
- Perturb Statistics
- Apply Data Use Agreement
- ...
- Charge for Access

Publisher

Recipient

Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk ???</i>

Attack Strategy A
<i>Utility A</i>
<i>Risk A</i>

Attack Strategy B
<i>Utility B</i>
<i>Risk B</i>

Recipient's
Best Strategy

⋮

Attack Strategy C
<i>Utility C</i>
<i>Risk C</i>

Publisher

Recipient

Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk B</i>

Attack Strategy A
<i>Utility A</i>
<i>Risk A</i>

Attack Strategy B
<i>Utility B</i>
<i>Risk B</i>

Recipient's
Best Strategy

⋮

Attack Strategy C
<i>Utility C</i>
<i>Risk C</i>

Publisher

Recipient

Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk B</i>

Sharing Strategy 2
<i>Utility 2</i>
<i>Risk ???</i>

Publisher

Attack Strategy A
<i>Utility A</i>
<i>Risk A</i>

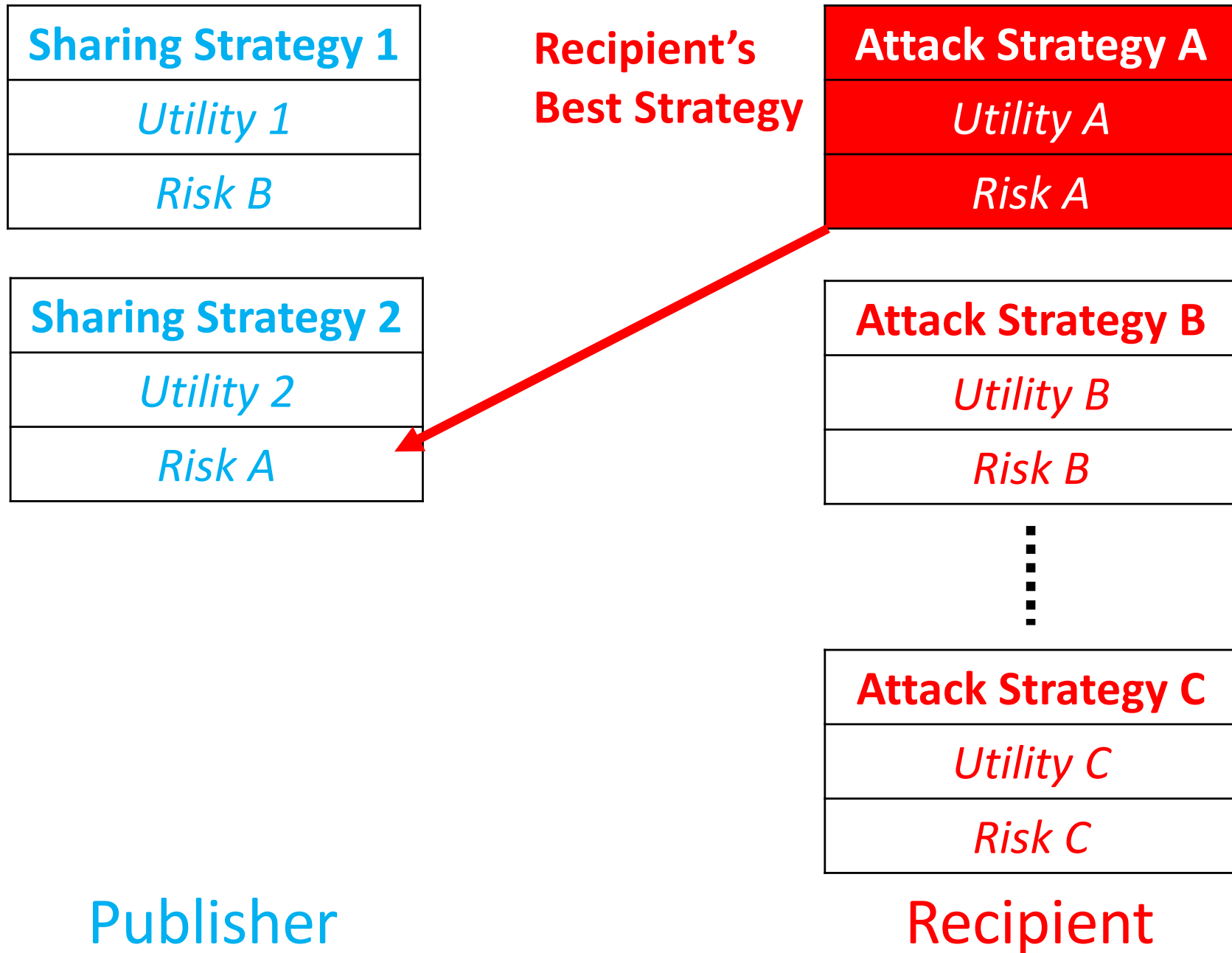
Attack Strategy B
<i>Utility B</i>
<i>Risk B</i>

⋮

Attack Strategy C
<i>Utility C</i>
<i>Risk C</i>

Recipient

Privacy “Games”



Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk B</i>

Sharing Strategy 2
<i>Utility 2</i>
<i>Risk A</i>

⋮

Sharing Strategy Z
<i>Utility Z</i>
<i>Risk Z</i>

Publisher

Privacy “Games”

Sharing Strategy 1
<i>Utility 1</i>
<i>Risk B</i>

Sharing Strategy 2
<i>Utility 2</i>
<i>Risk A</i>

⋮

Sharing Strategy Z
<i>Utility Z</i>
<i>Risk Z</i>

Publisher

Choose strategy that maximizes overall benefit

Optimizes the Risk-Utility tradeoff

Case Study

{Date of Birth, Gender, Geocode, Race}

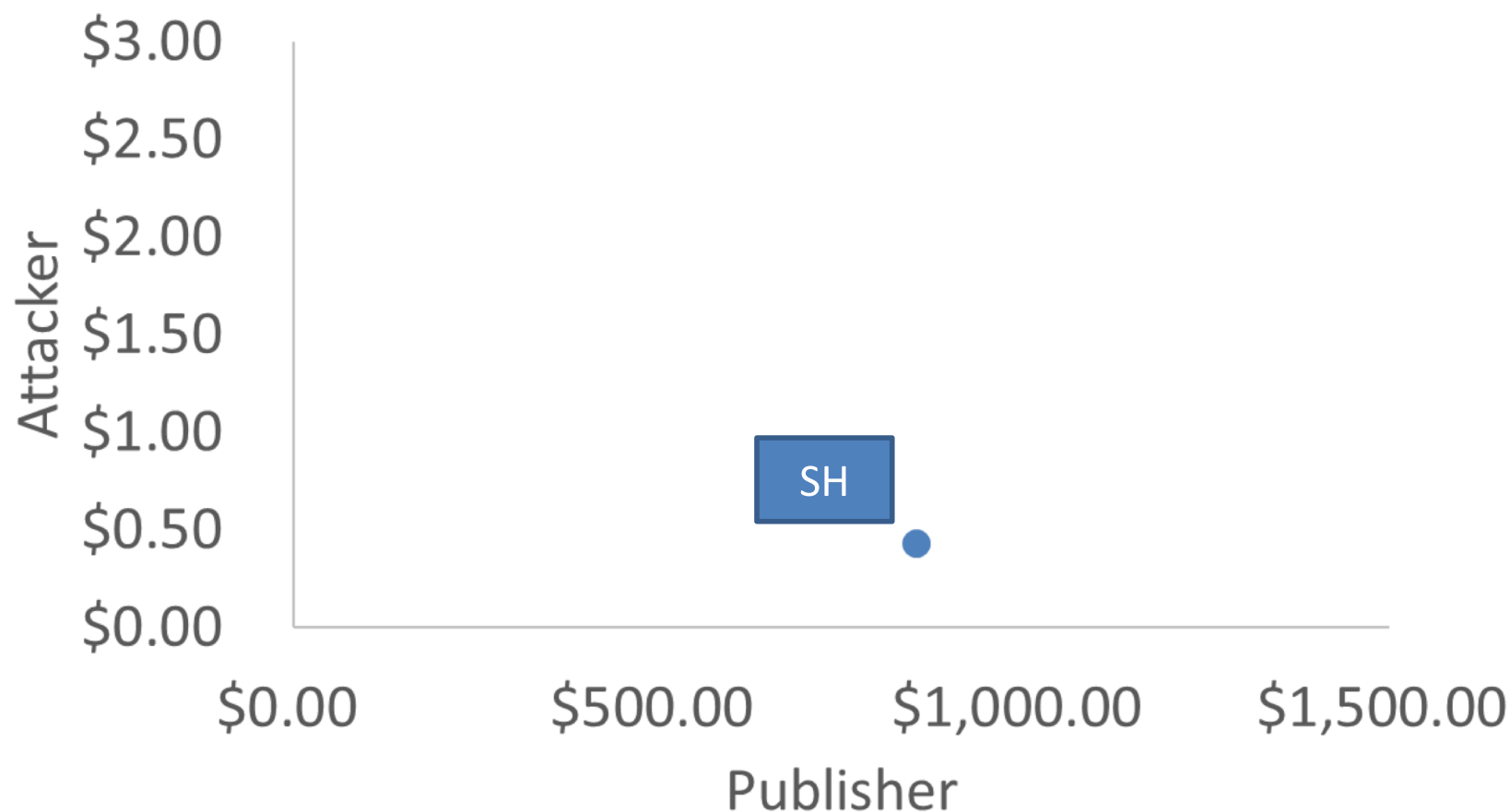
- ~30,000 Census records
- Average Payoff Per Record
- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record

Game Variations

- Safe Harbor (SH) Game
 - Defender shares data according to federal policy
- Basic Game
 - Defender shares data to maximize overall payoff
- SH-Friendly
 - Defender constrains strategy space to disclose no greater detail than SH
- No Attack
 - Defender constrains strategy space to disclose no greater detail than SH

Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record
- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record



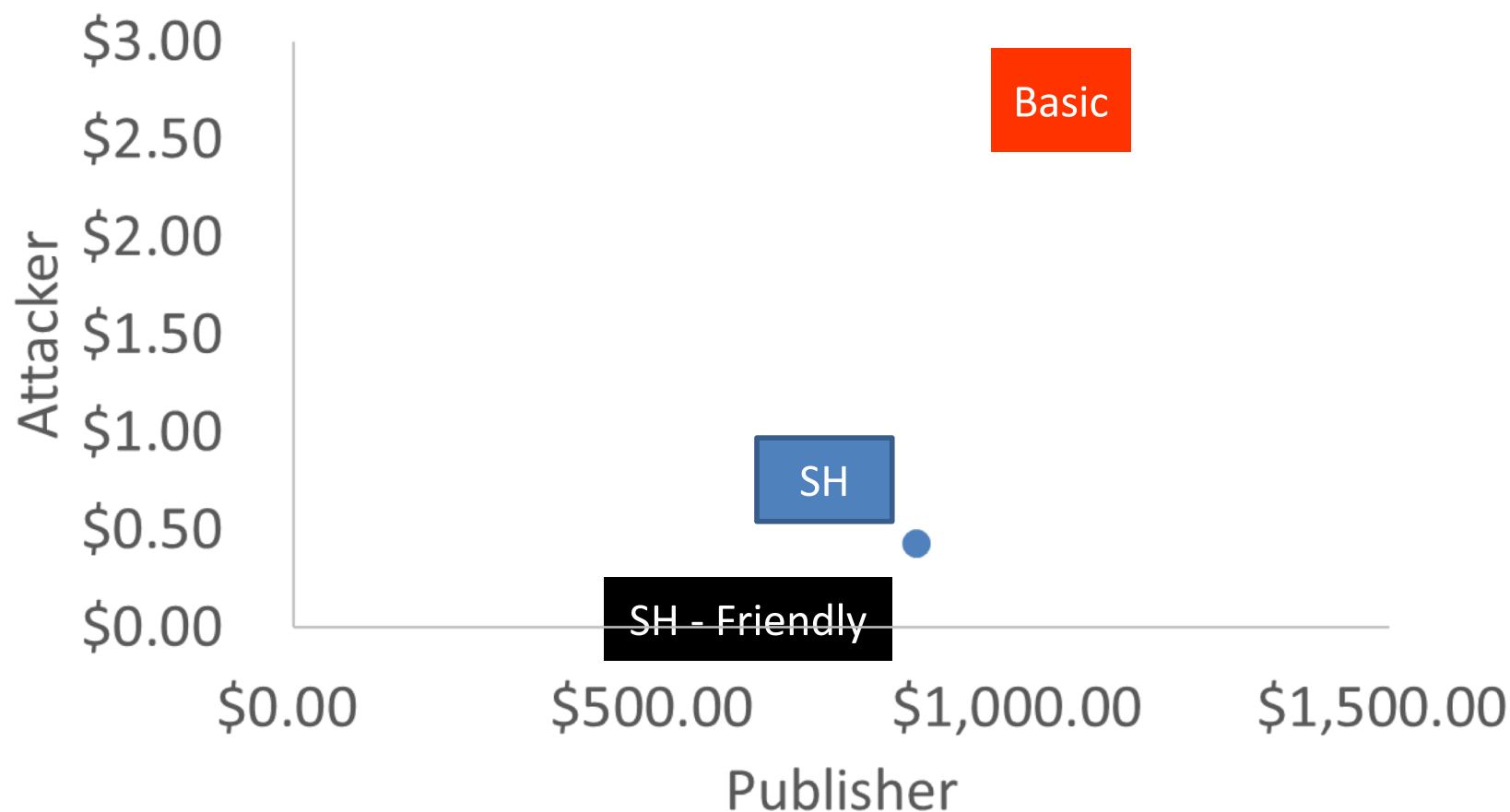
Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record
- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record



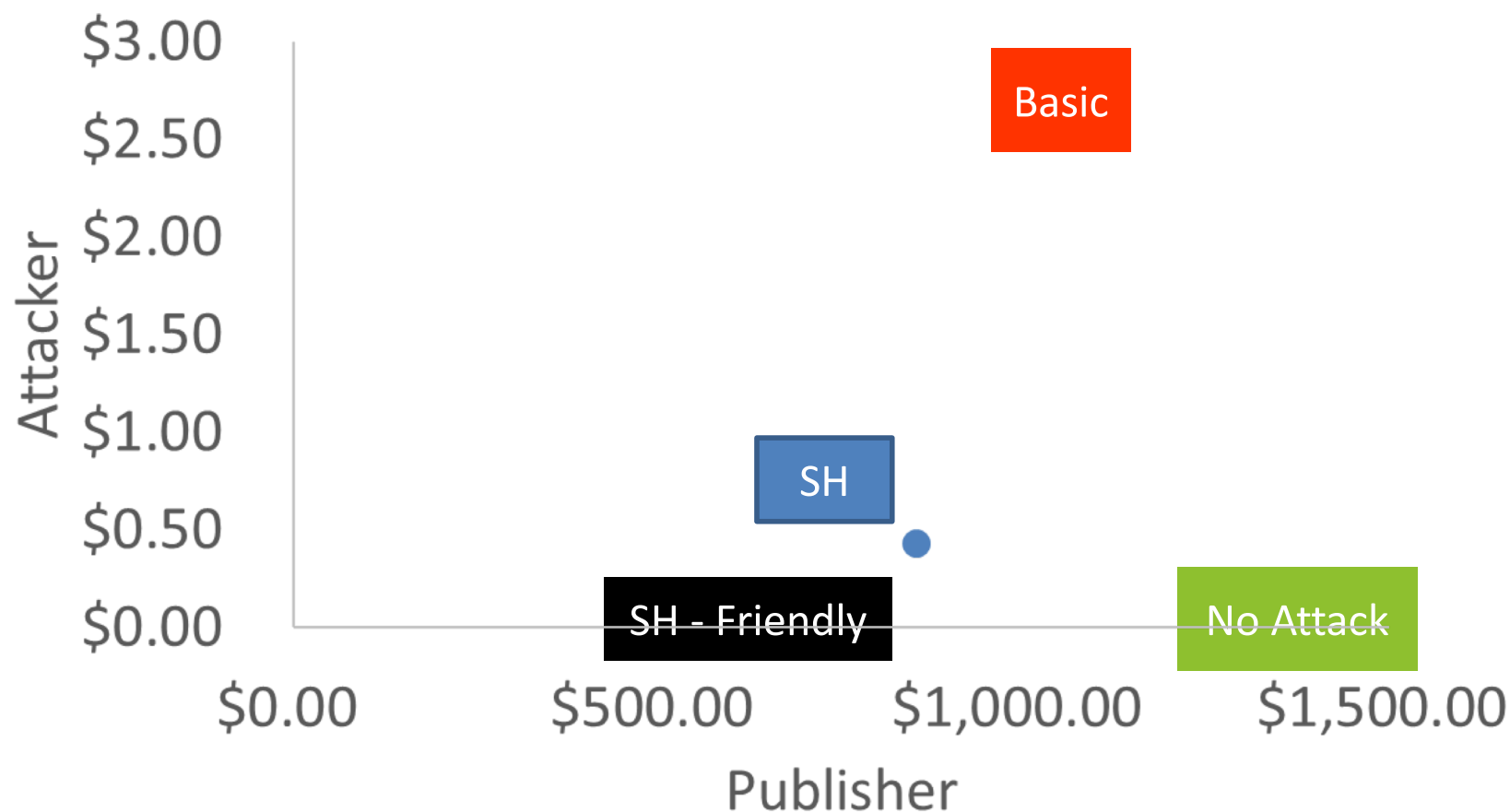
Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record
- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record



Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record
- \$1200: Benefit per record
- \$300: Cost per violation
- \$4: Access cost per record



De-identification is NOT a Panacea

- There is *always* a risk of re-identification
- But risk exists in any security setting
- The challenges are
 - Determine an appropriate level of risk
 - Ensure accountability
- Combine with data use agreements
- Risk is proportional to anticipated recipient trustworthiness (public vs. vetted investigator)

many ways to

Manipu

late

Alternative Data Protection Frameworks

- Encrypted Computation
 - Homomorphic encryption
 - Secure multiparty computation (SMC)
 - In a nutshell: learn aggregate answers without seeing individual records
- Secure Hardware
 - Push data into a tamper resistant environment
 - Intel SGX (software guard extension)
- Verifiable Provenance
 - Blockchain
 - It doesn't protect privacy... but it does provide lineage... sorta

Costs...

- De-identification: loss of data utility
- Encryption: loss in functionality
- Secure Environments: loss in efficiency
- Do (almost) nothing:
 - loss of privacy
 - loss of money due to litigation and remuneration
 - loss of societal trust
 - loss of scientific opportunity

Technology Lock-in

Questions?

b.malin@vanderbilt.edu

Center for Genetic Privacy and Identity in Community Settings

<http://www.vumc.org/getprecise/>

Vanderbilt Health Data Science Center

<http://www.vumc.org/heads/>

Vanderbilt Health Information Privacy Laboratory

<http://www.hiplab.org/>

[People Search](#)
[Background Check](#)
[Criminal Records](#)
[Reverse Lookup](#)
[Intelius Premier](#)
[Identity Protection](#)
[Employee Screening](#)
[People Search](#)
[Email Lookup](#)
[Social Network Search](#)
[Property Records](#)
[24-Hour People Search Pass](#)

People Search - Updated Daily, Accurate and Fast!

People Search



First Name	M.I.	Last Name <i>required</i>	City and/or State	Search

Reverse Phone Lookup



Phone Number	Search
() -	

More ways to get info you need:

- ✦ [Perform a Background Check](#)
- ✦ [Run a Background Check by SSN](#)
- ✦ [Perform an Address Lookup](#)
- ✦ [Do a Reverse Phone Lookup](#)

What is People Search?

It's a confidential way to find people so you can reconnect or just get more info on a person. People Search reports can include phone numbers, address history, age & date of birth, relatives, and more. Find a person you're curious about – search today!



What is Reverse Phone Lookup?

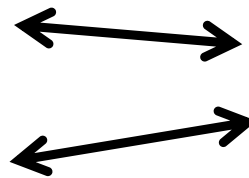
It's a confidential way to find out who a phone number belongs to. Reverse phone search works for landline, unlisted & non-published numbers, and cell phone lookups. Reports can include phone type, owner name, address & more. Curious? Do a phone number lookup!



**Big(ger) Data
Can Enable
Privacy**

An Attack on Diagnoses

Identified EMR data		
i	ID	ICD9
1	Jim	333.4
2	Jack	333.4
3	Mary	401.0,401.1
4	Anne	401.1,401.2,401.3
5	Tom	571.40,571.42
6	Greg	571.40,571.43

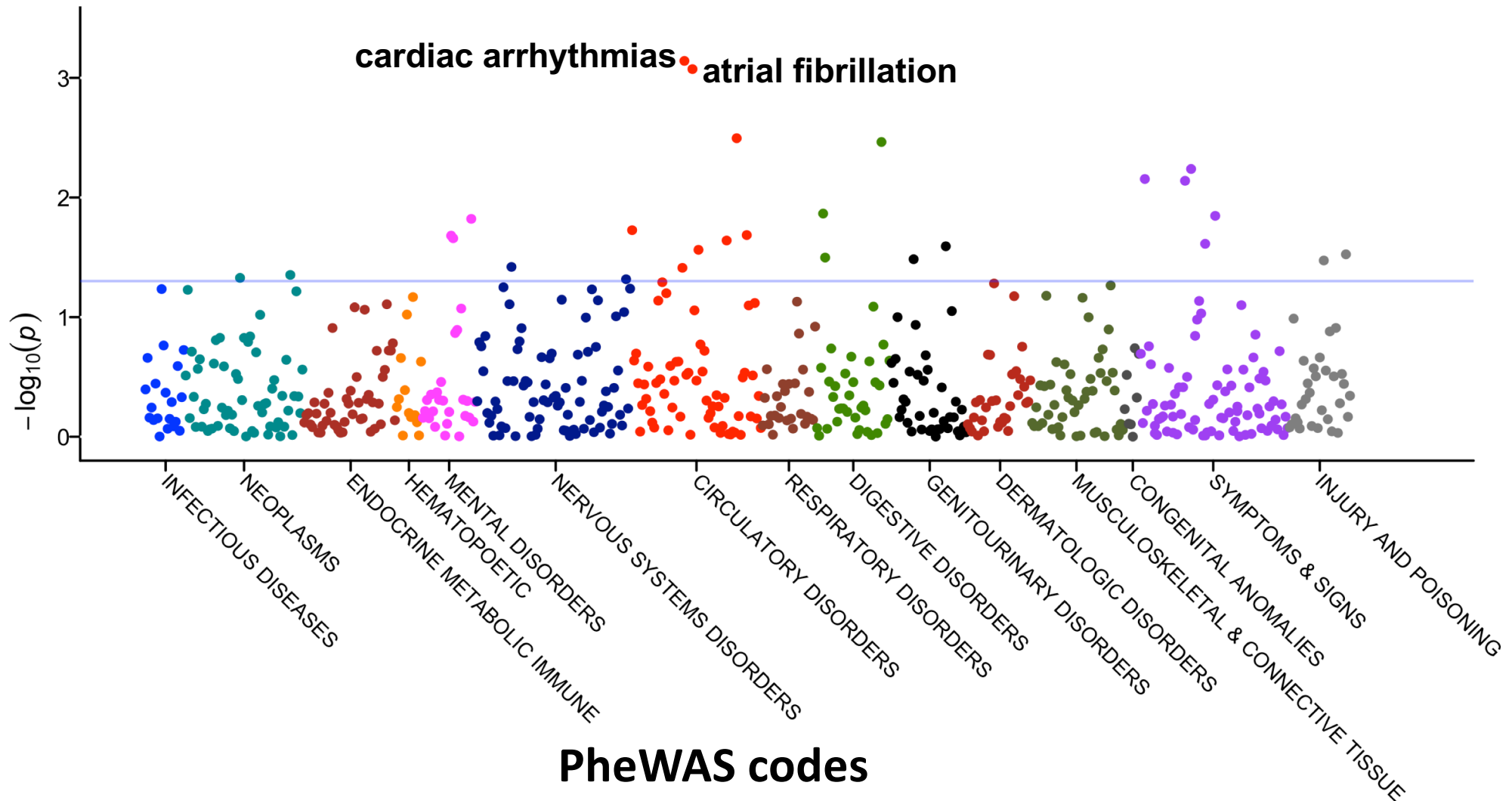


De-identified Research data		
j	ICD9	DNA
1	333.4	CT...A
2	401.0,401.1	AC...T
3	571.40,571.42	GC...A

More than ½ of Vanderbilt patients are unique!

Phenome Wide Association Studies

(associated with longer QRS duration in normal hearts)



Thanks to Josh Denny

A Little Realism Goes a Long Way

- Suppress diagnoses if < 5 patients
- Validation of 192 genome – phenotype associations
- X-axis: original p-values
- Y-axis: protected p-values

