Privacy Games:
Leveraging Economics to Balance Health Data Privacy & Utility

Bradley Malin, Ph.D.

Vanderbilt University Medical Center
Vice Chair, Department of Biomedical Informatics
Co-Director, Center for Health Data Science
Co-Director, Center for Genetic Privacy & Identity in Community Settings

Vanderbilt University
Professor of Biomedical Informatics, Biostatistics, and CS
Co-Director, Big Biomedical Data Science Training Program

10 October 2019

The future of health begins with you

The *All of Us* Research Program is a historic effort to gather data from one million or more people living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, environment, and biology, researchers will uncover paths toward delivering precision medicine.

JOIN NOW

Interested in the *All of Us* Research Program?

LEARN MORE

Sign up to be notified of announcements, events, funding news and more.

SUBSCRIBE

We are building a research program of 1,000,000+ people

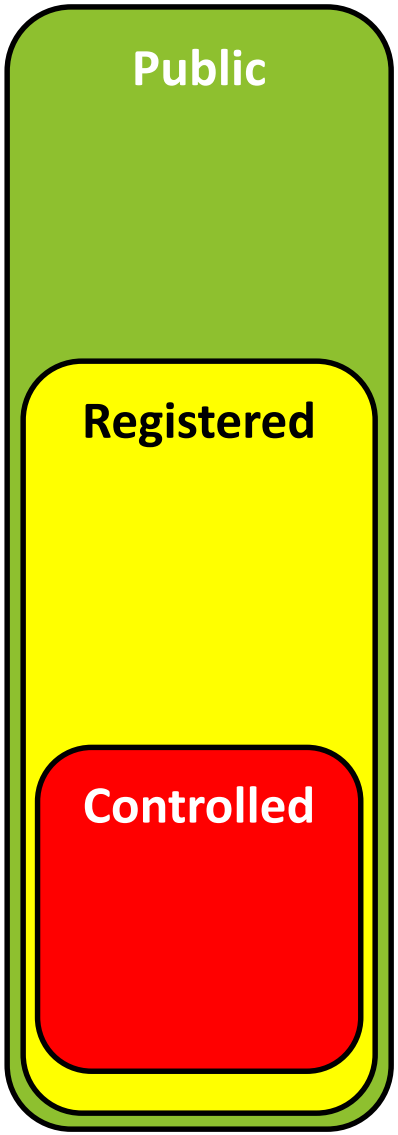
Tiered Levels of Access

- Public
  - Can be accessed without logging in
  - Summary statistics only

- Sandbox Environments

Google Cloud Platform

  - Registered
    - Available to anyone who proves their identity
    - Individual-level data with low risk of re-identification
  - Controlled
    - Available to trusted investigators (ERA Commons ID)
    - More detail, more risk



https://databrowser.researchallofus.org/

Search Across Data Types

Keyword Search

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

FAQs

Introductory Videos

User Guide

EHR Domains:

Conditions

12,735
medical concepts

36,260 participants in this domain

View Top Conditions

Drug Exposures

14,967
medical concepts

33,440 participants in this domain

View Top Drug Exposures

Labs and Measurements

7,733
medical concepts

32,480 participants in this domain

View Top Labs and Measurements

Procedures

13,202
medical concepts

35,320 participants in this domain

View Top Procedures

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

8
physical measurements

81,460 participants in this domain

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

The AOL Search Log Case of 2006

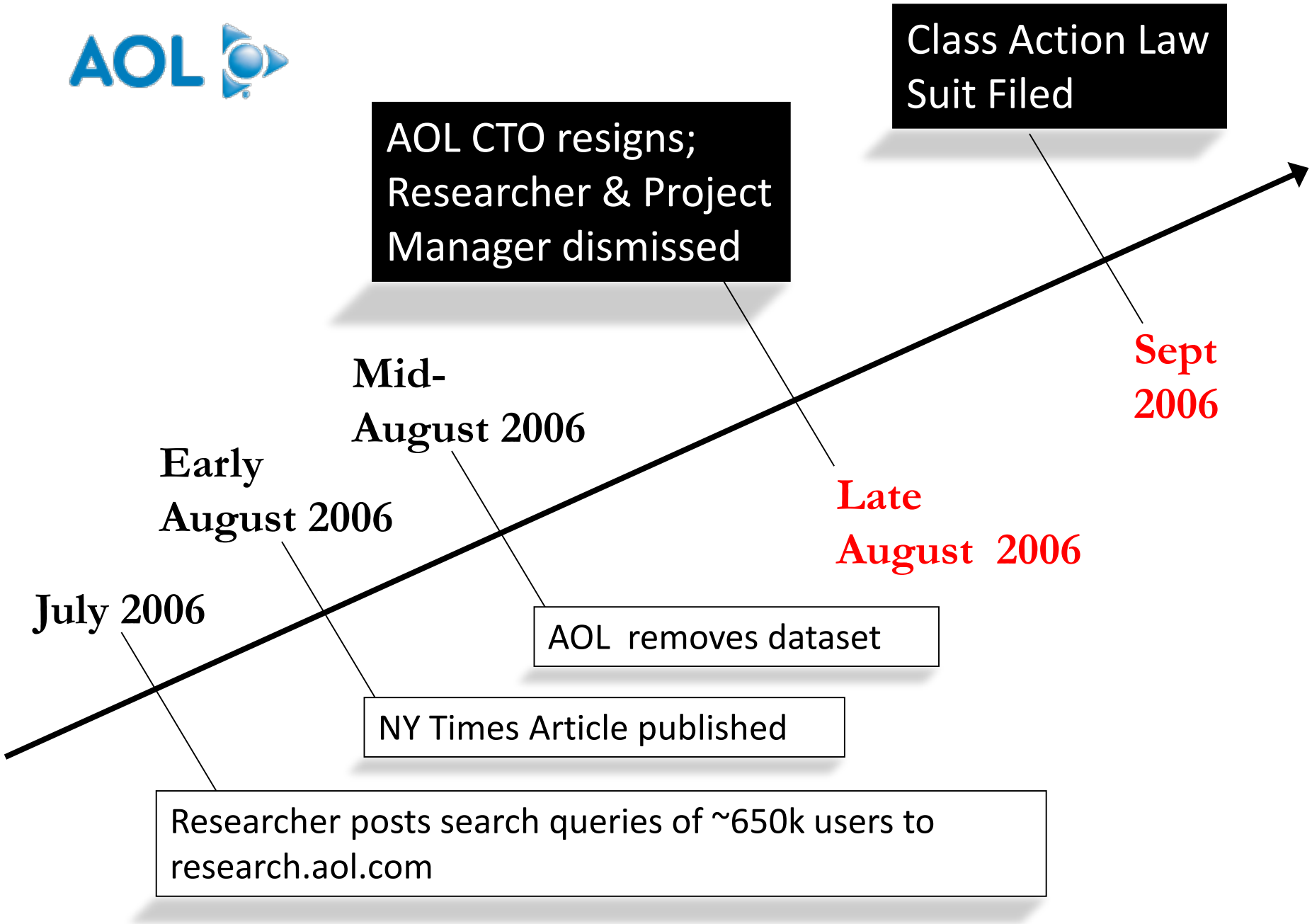
Pseudo	Name	Query	Date	Time
1	[redacted]	Books	1/2/05	16:52
2	[redacted]	Payscale	1/4/05	23:41
1	[redacted]	Porn	1/8/05	03:15

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

7


AOL
Class Action Law Suit Filed
AOL CTO resigns; Researcher & Project Manager dismissed
Sept 2006
Mid-August 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 Netfli 09)

- Published m eudonyms
- Re-identifica tions

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

Welcome Google User
Here are more stories related to your searc
• Netflix Settles Privacy Lawsuit, Canc
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.

- Class actio

[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)
- Health Survey Responses (Solomon '12)
- Laboratory Tests (Cimino '12; Atreya '13; Johnson '19)
- DNA (Malin '00; Lin '02, '04; Homer '08; Gymrek '13; Ayday'14; Huttenhower '15; Shringapure '15; Lippert '17; Erlich '18; Gursoy '18; Hagestedt '19)
- Pedigree Structure (Malin '06; Ayday '13)
- RNA (Backes '16a; Backes '16b; Gursoy '19)
- Proteome (Li '16; Boonen '19)

- 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, Qian '16, Horawalavithana '18)
- Search Queries (Barbaro '06)
- Internet Use (Malin '05; Eckersley '10; Banse '11; Herrmann '12, Olejnik '12; Kirchler '16; Riederer '16; Panniello '16)
- Smart Utility Meter Usage (Buchmann '12; Faisal '15; Tudor '15)

With Enough

Time,

Effort,

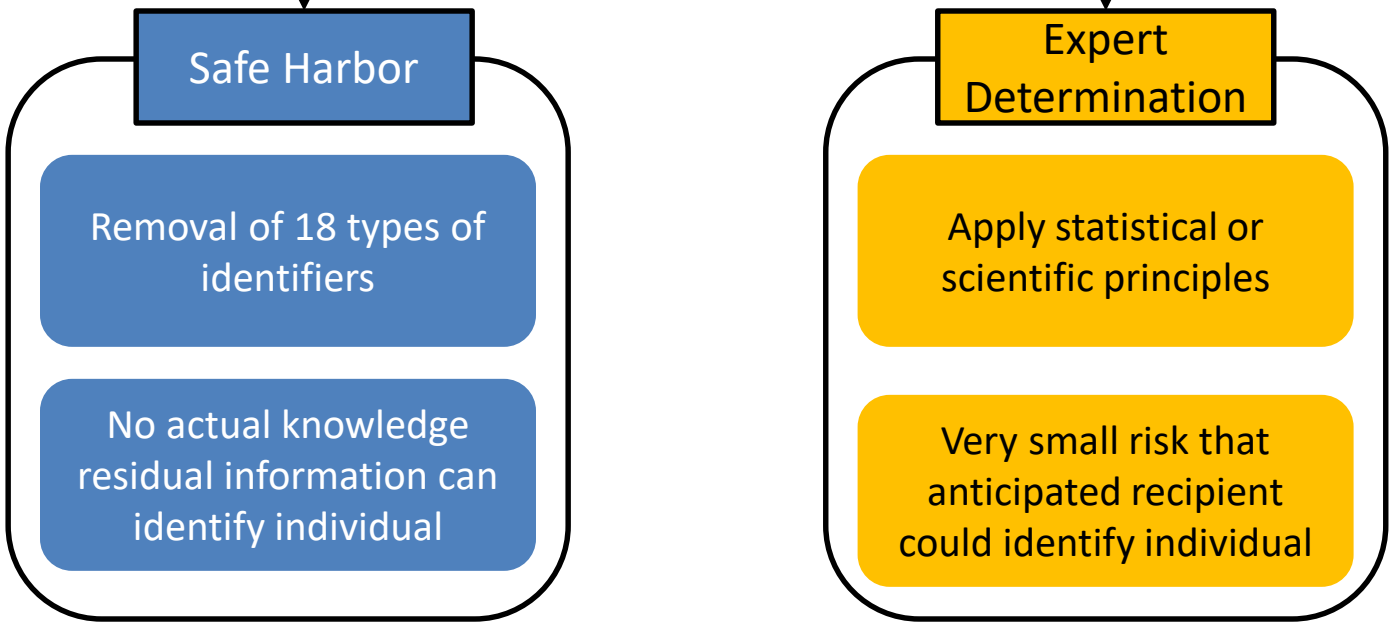
Incentive, &

MONEY...

What Does the Law Say?

According to the US Health Insurance Portability & Accountability Act (HIPAA):

"information that does not identify an individual and … no reasonable basis … information can be used to identify an individual"



According to EU (now General Data Protection Regulation [GDPR]):

"principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable"

A Simplified View on Risk

$$P(reid) \approx P(attack) * P(reid \mid attack)$$

Deterrents

- Use Agreements
- Pay for Access
- Unique Login / Pass
- Audit

Data

- Uniqueness
- Replicability
- Availability

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	●	●			

Benitez & Malin. JAMIA. 2010.

Beyond “unique”

Randomized (Golle 2006)

- Don't always have exact knowledge of what a data recipient has access to
  - Disclose sample with *{dob, gender, zip}*, but don't know the population's values.
- May know population counts, such as
  - U.S. Census aggregates for *{year of birth, gender, county}*
- Conversion: {Year of Birth, ZIP} → {Date of Birth, ZIP}

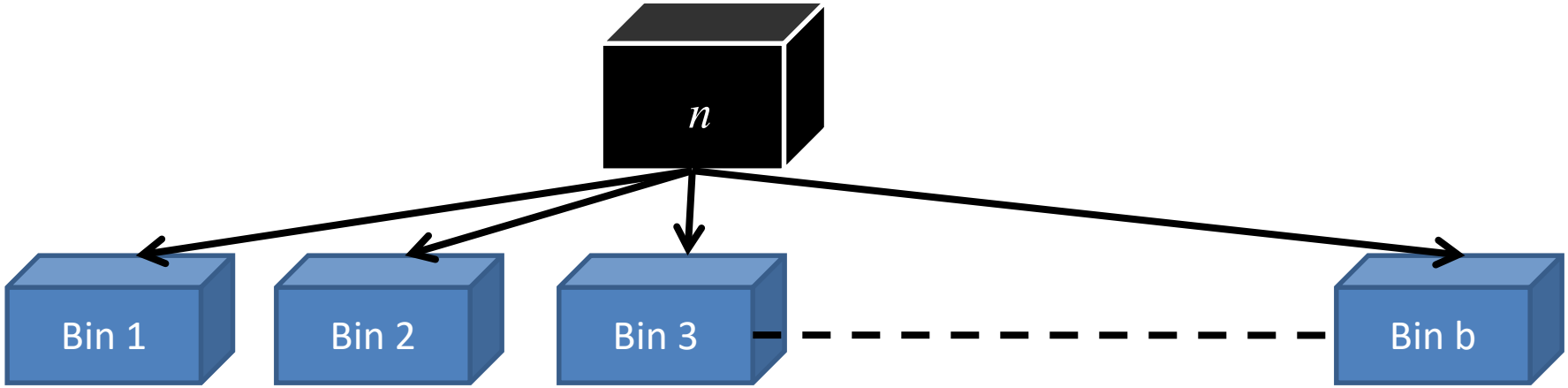
63% of the U.S. estimated as unique

		Birth Year
		1980
ZIP	**zip1**	12000
	zip2	50000
	…	
	zip m	10000

		Birthdate			SUM
		1/1/80	**…**	**12/31/80**	
ZIP	**zip1**	random		random	12000
	zip2	random		random	50000
	…				
	zip m	random		random	10000

Alternative: random distribution of values across the cells

It's an Occupancy Problem



- n people in aggregated bin
- b disaggregated bins
- the expected # of bins with exactly i people

$$f_i(n) = \binom{n}{i} b^{1-n} b^{n-i}$$

- Total number of people in a group of size less than k

$$r_k(n) = \sum_{i=1}^{k-1} f_i(n)$$

HIPAA Privacy Rule Allows Tiered Access for Secondary Uses of Data

Identified Patient Data

- Waiver of consent: data is “on the shelf”
- Consent is impractical to obtain

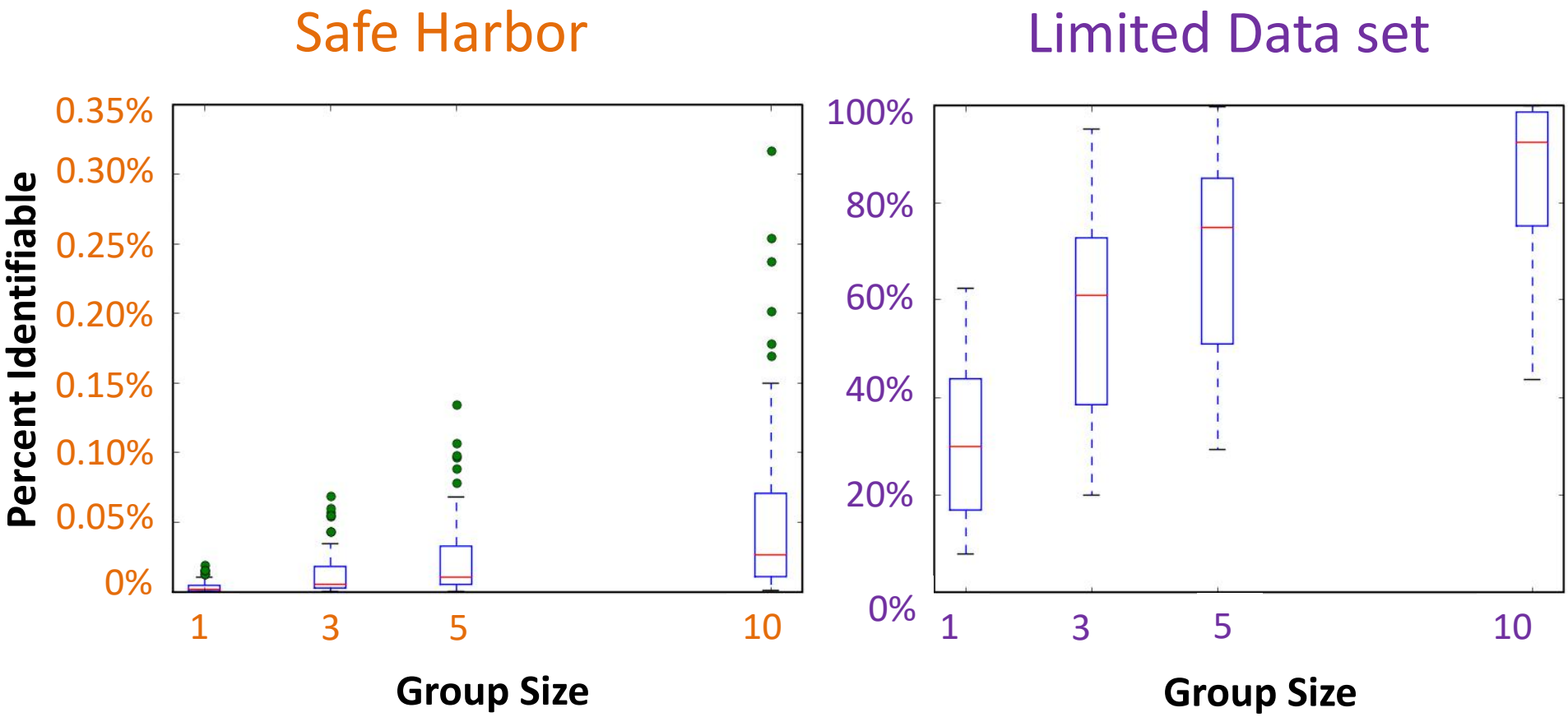
Limited Data Set

- Removal of 16 designated attributes
- Recipient signs data use contract

De-identified Data

- Option 1: Safe Harbor
- Option 2: Expert Determination

All U.S. States



Benitez & Malin, JAMIA. 2010.

Risk…

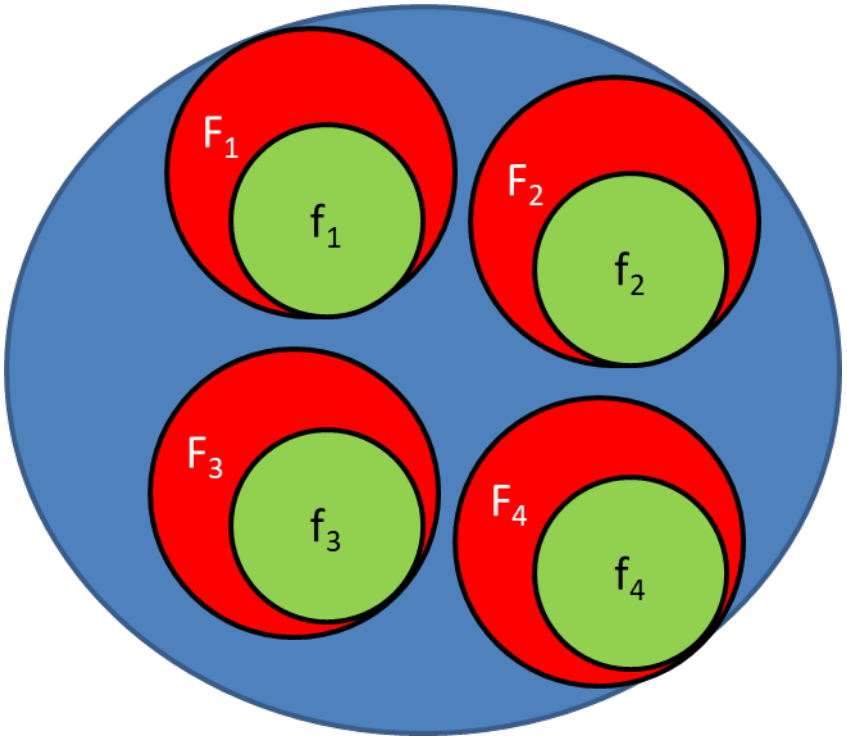
… means something different to everyone

- Can be modeled in various ways (Dankar & El Emam, 2010)
  - Prosecutor ← $1/\min_i(f_i)$
  - Journalist ← $1/\min_i(F_i)$
  - Marketer ← $n^{-1}\sum_i \frac{f_i}{F_i}$

f_i = size of group in sample

F_i = size of group in population

n = sample size

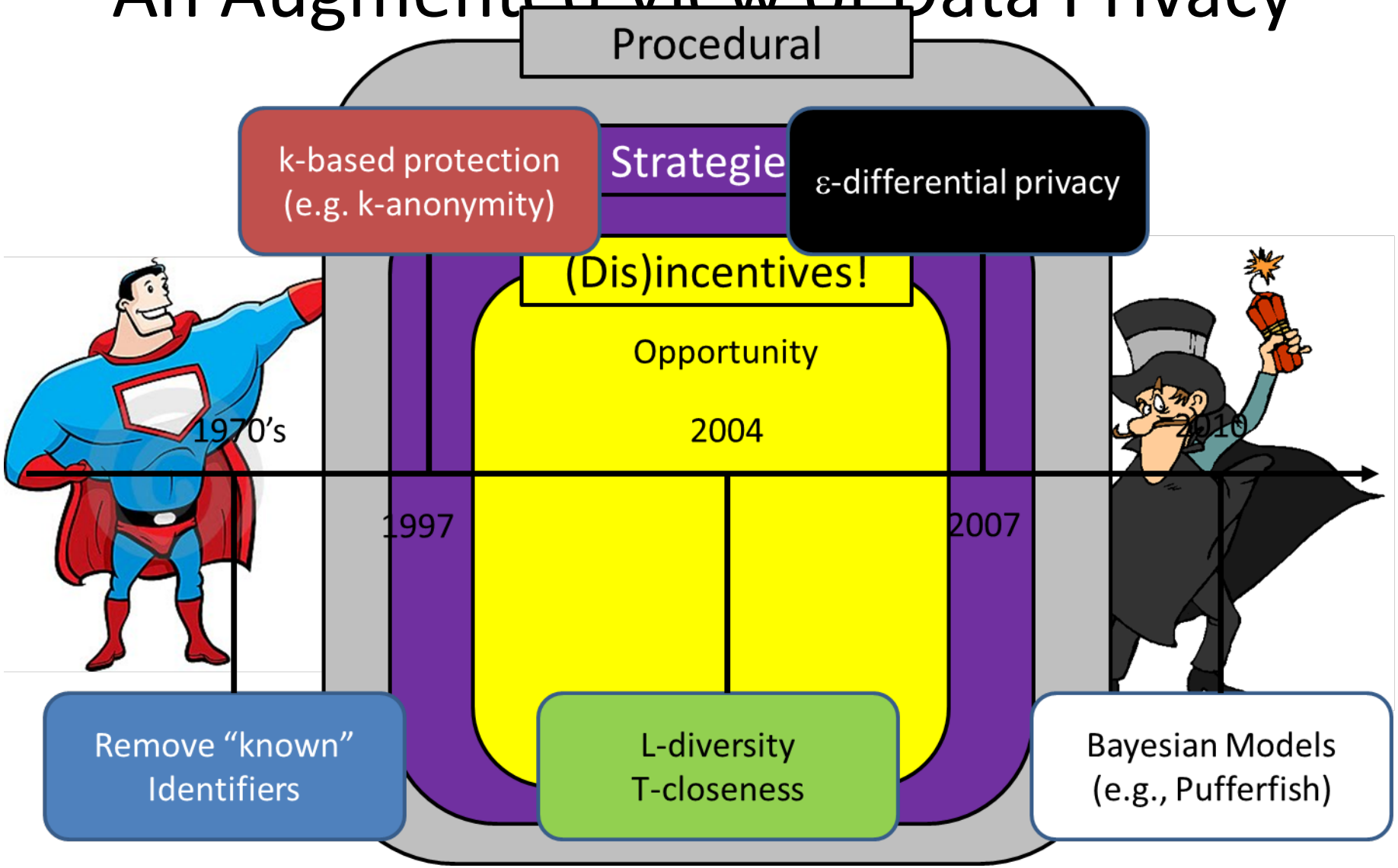


So What About Cost?

U.S. State	Limited Dataset		Safe Harbor	
	At Risk	Cost per Re-id	At Risk	Cost per Re-id
Virginia	3159764	$0	221	$0
South Carolina	2231973	$0	1386	$0
Wisconsin	72	$174	2	$6,250
West Virginia	55	$309	1	$17,000

Benitez & Malin. JAMIA. 2010.

An Augmented View of Data Privacy



Stackelberg Game

Sharing Strategy 1
Utility 1
Risk ???

Attack Strategy A
Utility A
Risk A

Strategies:

- Hold back genomic regions
- Generalize demographics
- Perturb statistics
- Apply data use agreement

…

- Charge for Access

Attack Strategy B
Utility B
Risk B

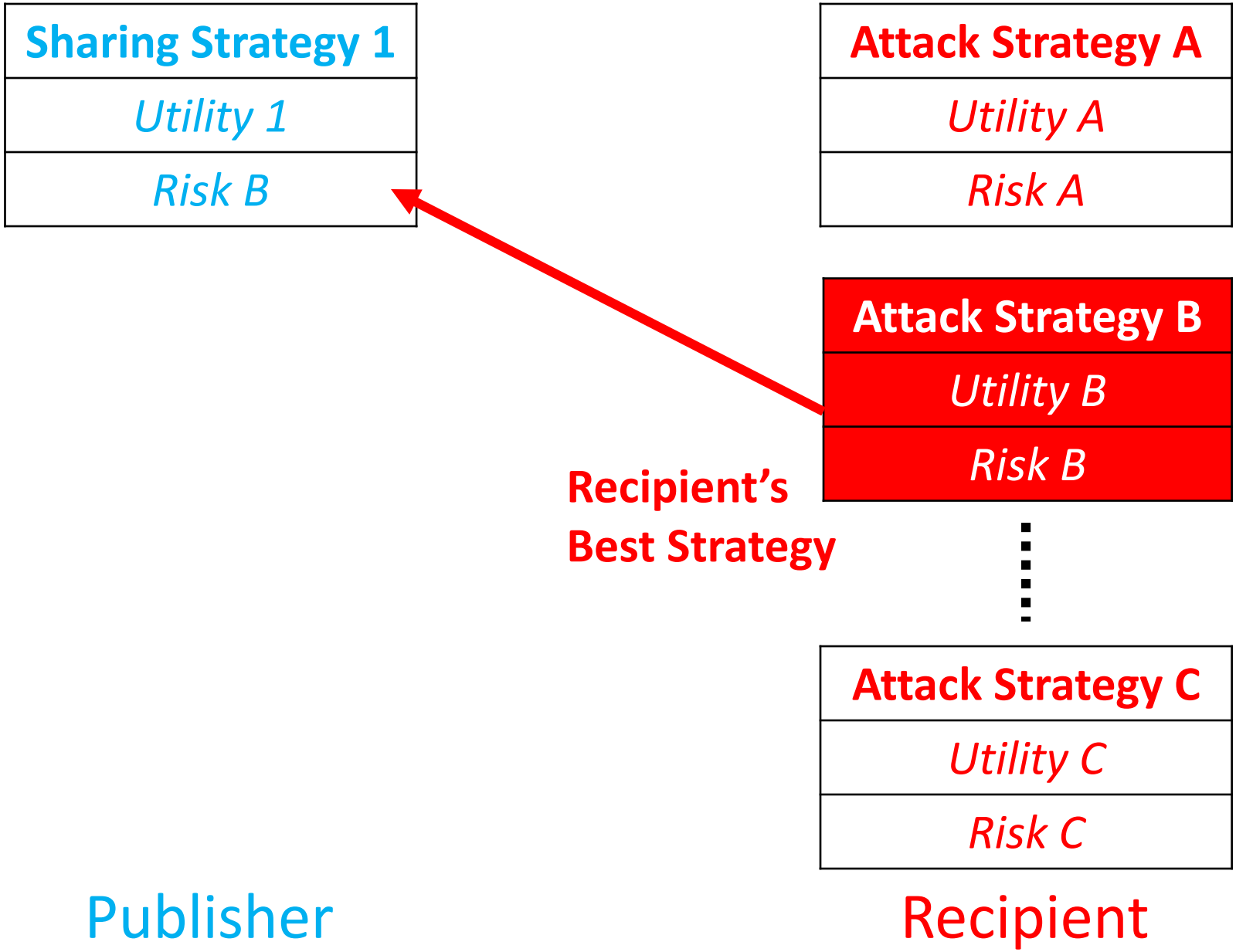
⋮

Attack Strategy C
Utility C
Risk C

Publisher

Recipient

Stackelberg Game



Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk A

Recipient's Best Strategy

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Publisher

Recipient

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk A

⋮

Sharing Strategy Z
Utility Z
Risk Z

Publisher

Choose strategy that maximizes overall benefit

Optimizes the Risk-Utility tradeoff

Adversarial Models Require *Assumptions*

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

Wan et al, PLoS One. 2015

INTELIUS
Live in the know.™

Like 28k G+1 17k

Help | Sign

★ Bookmark this S

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 required | 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!

Payoffs!

Gain	No Attack	Attack
Publisher	v_g	$v_g - L\pi_g$
Attacker	0	$L\pi_g - c$

g : Generalization level

v_g : Value of record at g

π_g : Probability of successful attack at g

L : Loss to publisher for successful attack

c : Cost to run attack

Wan et al, PLoS One. 2015

Game Variations

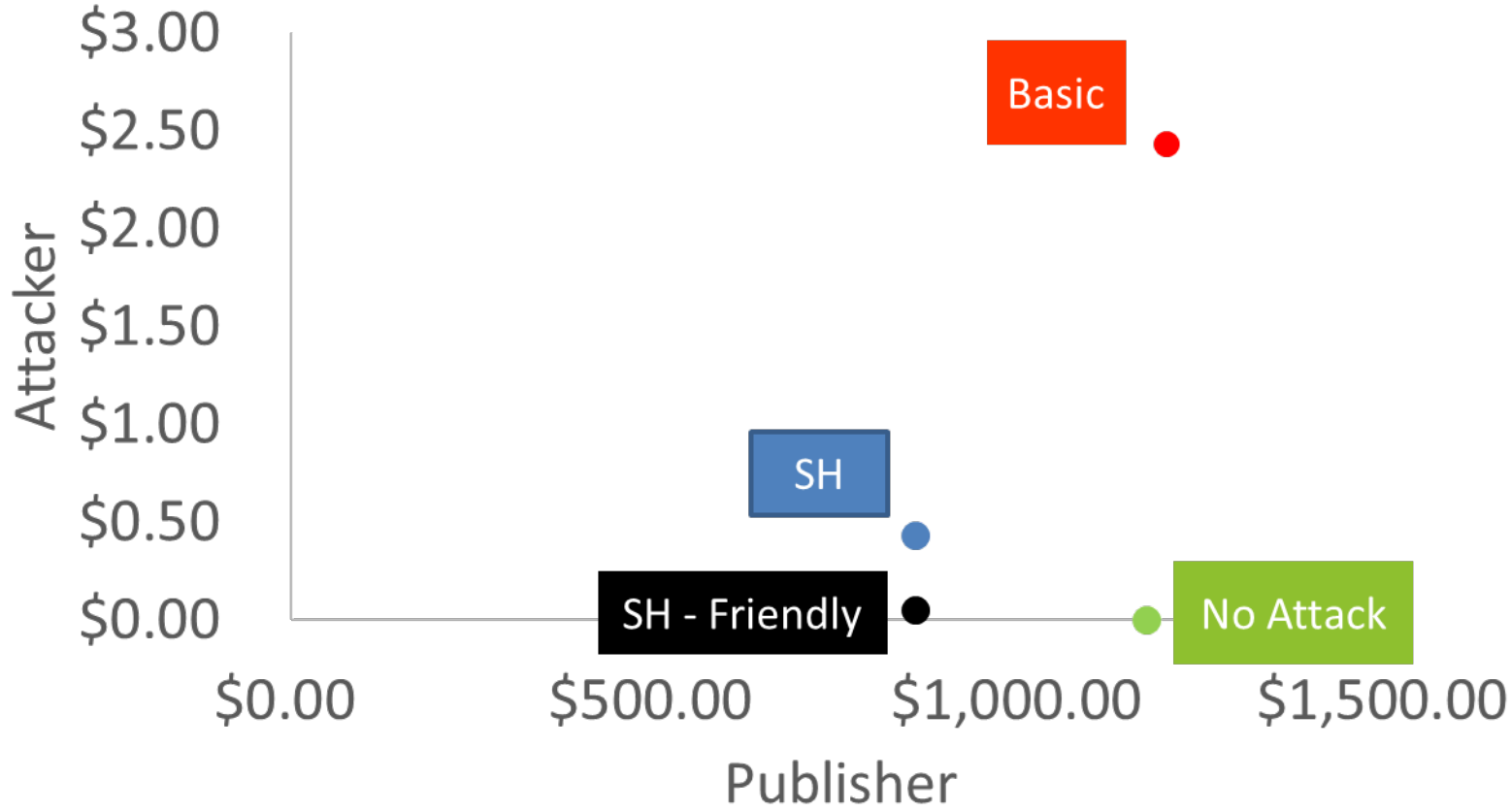
- Safe Harbor (SH) Game
  - Defender shares data according to federal policy
- Basic Game
  - Defender shares data to maximize overall payoff
- SH-Friendly
  - Defender constrains strategy space to disclose no greater detail than SH
- No Attack
  - Defender constrains strategy space to disclose no greater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

Wan et al, PLoS One. 2015

Demographic Case Study

- ~30,000 Census records
- Average Payoff Per Record

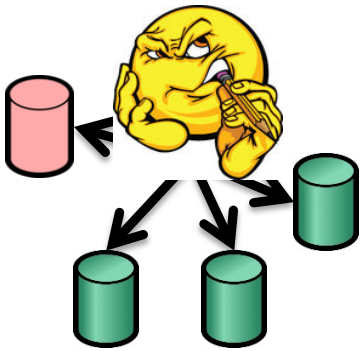
- $1200: Benefit per record
- $300: Cost per violation
- $4: Access cost per record



Adversaries Make Sequential Decisions!

(Xia et al, CIKM 2015)

access the external dataset



link the record to the external dataset

exploit an individual that matches the record

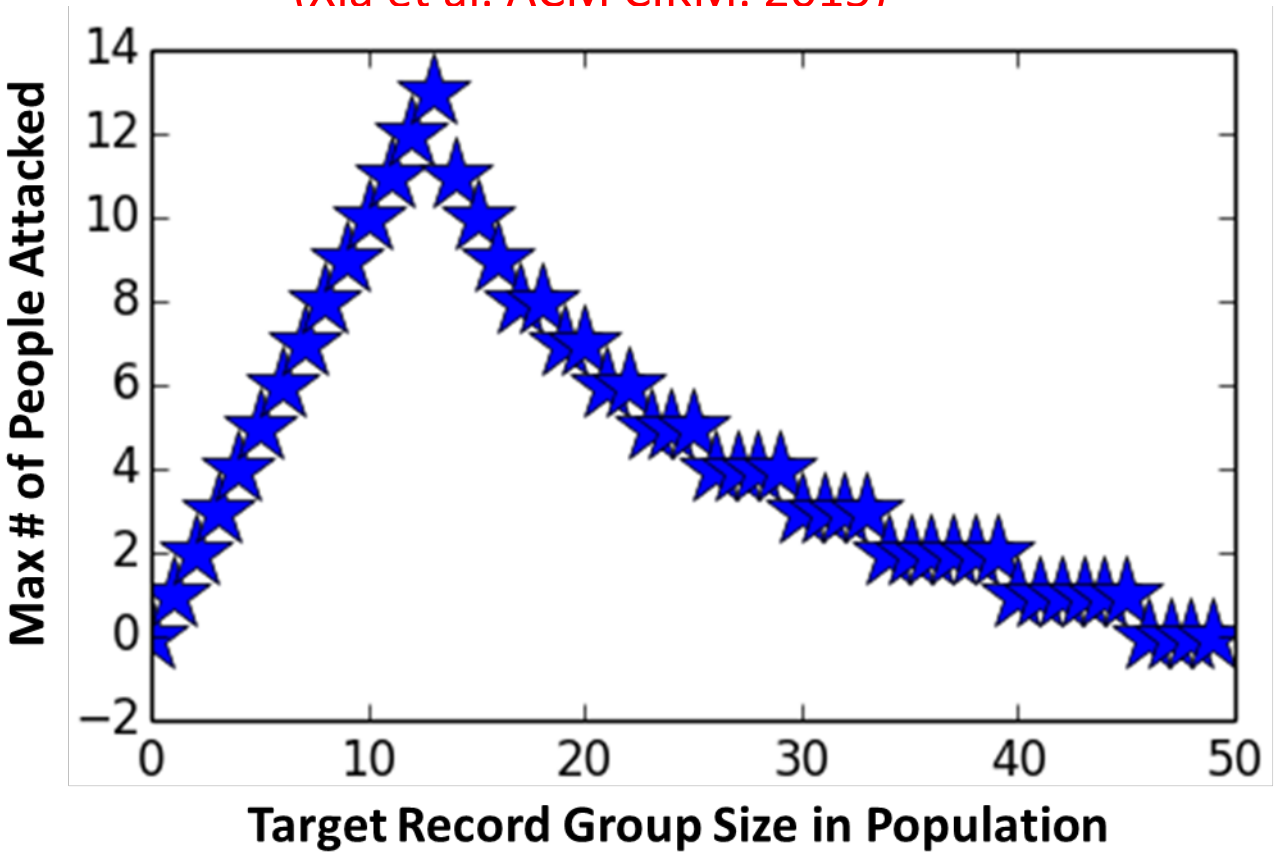
Intelius	$3.95/record
USSearch	$1.45/record
NY voter registration	$0
broker	?

t2.micro	$0.013/hour
t2.medium	$0.052/hour
m3.large	$0.133/hour

phone marketing/confirmation	$0.013/mi
email marketing	
Publish the information	
Penalty	$?

Comparison to Baseline Stackelberg

(Xia et al. ACM CIKM. 2015)



	Sequential	Baseline
Exploit	≥ 1 individual	1 individual
Decisions	Sequence	Single attack

Wan, et al. 2015.

Prior Probabilities Matter

ALOT

*omics Data is High-Dimensional!

AG AT CA AA CA TG GC AA TA AT CA GT AG GG AT AT CC AC AG TT AA TA CT
CC AG CA GC AG AT AA CA GG TA GG GC GA AA GA GA GC CA TA CT CC CC CC
CA TT AG TT GT AA GG GG AT AC CA GC GG AC AC AC AG TT TC AA CC AC AT
CG TG AC AT GG TC AT CC GC GA TC AC CT CC CG TG TT TT GC TC TC AC AC
AT AT GA CA TG CT TA AC CT AG TC CC AC TT CA AA CG TG TA AA TT GA TA
TC CT AC AT CA CT CA CC TC AG TC TA CA CC CG TG TA TT AG GT CC TT TA
TG TA [illegible] CC GG GG AG TT GG [illegible] CA CG TT CC CC AA CG CC [illegible] AT TG
AT A[illegible]T CC CC[illegible]C CA CA TC AA GT [illegible] GG CT CC AC CT TA GT[illegible] GC CA
GT [illegible] AA CT TG [illegible] GC [illegible] CA [illegible] GG [illegible]G CC A[illegible]AT AG GG GT [illegible] TA GG
GA [illegible] TT TT TA [illegible] AC [illegible] GT [illegible] TT [illegible] AT [illegible] GC [illegible] CC AT CT AC [illegible] AT TT
TT [illegible] CT GA CG [illegible] TT [illegible]C TG [illegible] AG [illegible] GG [illegible] TC[illegible] AC TA TG TC [illegible] GT GT
CG [illegible] AA AT GG [illegible] CC [illegible]T TT [illegible] CA [illegible] AT [illegible] GC[illegible] A CA TA AC AC [illegible] AG TC
TG A[illegible]G TA G[illegible]C GT [illegible]G TG A[illegible] GG [illegible] CG TC[illegible] AG TG CG GG CC [illegible] CC GG
AG GA [illegible] TC TG[illegible]C GT A[illegible] TT [illegible] AG CT[illegible] AT TG TT AG[illegible] AG
AA GT CA AC GC CC TT TC GA GC GT GG AT GT[illegible] TC AC AT GA AC AT CT GA
TC CA GA CA GA CA AT AT TG AC GG CC CA A[illegible]T CT TT GT GT CT AG CT CG
CA GT AC CC CT GA CG CC GT GC TA CC CC AG TG GG GG TA GA TT CG AC CG
CG GG TC AG TA GC TC CT GA AG GG AC TA AT TA CT TC TG TT GT GT GC GT
CC CT TC GC TA GG AC GA AT CT TG GA TG GG AA AA TA AT CT CA AA AC CC
TC AT CG AG GG TG CC GC GT AG AC CT TC CA AG GC GA TA TG AC TA AA CG
AA GA CA GA TA TG AT AT CA CC AC AC GG TT TA GG GC CG CC TG TA TG TT
GT CC CT AC TT CC CT TA CA AC TG CG GG TC AT GA CC

Only 1

Lin, Owen, & Altman. Science 2004.

Populations

Individual

It's not just unique sequences… summary stats (Homer et al. 2008)

PRIVATE "adversary"

Name	SNP 1	SNP 2		SNP n
Alice	AA	CG	…	TT
Bob	AT	GG		GT

Association Studies (e.g., dbGaP)

CONTROLS (e.g., No diabetes)	SNP 1			SNP 2				SNP n		
Allele	AA	AT	TT	CC	CG	GG	…	GG	GT	TT
Probability	0.5	0.3	0.2	0.7	0.2	0.1		0.6	0.3	0.1

CASES (e.g., Diabetes)	SNP 1			SNP 2				SNP n		
Allele	AA	AT	TT	CC	CG	GG	…	GG	GT	TT
Probability	0.3	0.35	0.35	0.6	0.25	0.15		0.55	0.35	0.1

General Population (e.g., 1000 Genomes)

	SNP 1			SNP 2				SNP n		
Allele	AA	AT	TT	CC	CG	GG	…	GG	GT	TT
Probability	0.3	0.35	0.35	0.6	0.25	0.15		0.55	0.35	0.1

SPHINX Genes Pathways Drugs Methods Terms of Use eMERGE Network Login

https://www.emergesphinx.org/

SPHINX

A resource of the *e*MERGE Network

Search by Gene, Pathway, or Drug | Search

List all: genes, pathways, drugs

***S**equence and **Ph**enotype **In**tegration E**x**change (**SPHINX**) is a web-based tool for exploring data for hypothesis generation, especially around drug response implications of genetic variation across the eMERGE PGx cohort.*

9 sites **82** genes **38,112** variants **60** pathways **515** drugs

9,015 unique samples - cohort complete

Last update: July 2015

The eMERGE-PGx project was a multi-center pilot of implementing pharmacogenetic sequencing in clinical practice to improve health care. SPHINX is a searchable catalog of observed inherited variants in a 9015 subject population, large enough to reflect even rare variation. The participants' constitutional DNA was sequenced using the PGRNseq assay, a targeted megabase of sequence in 82 PGx genes, genes identified as important for pharmcogenomics.

What Can I Do with SPHINX?

See the lists of genes, drugs, and pathways in the catalog

Search the catalog of variants by:

- Gene
- Drug interactions (from PharmGKB data)
- Pathway (from PharmGKB data)

See for each gene:

- Observed single nucleotide variants
- Pathways involved
  - Other genes in each pathway
  - Other variants in each pathway
- Drug interations
  - Other genes with that drug interaction
  - Other variants with that drug interaction

The sites participating in eMERGE and the eMERGE-PGx project include:

- **Children's Hospital of Pennsylvania**
- **Cincinnati Children's Medical Center with Boston Children's Hospital**
- **Essentia Rural Health with Marshfield Clinic and The Pennsylvania State University**
- **Geisinger Health System**
- **Group Health Cooperative with University of Washington**
- **Mayo Clinic**
- **Mount Sinai School of Medicine**
- **Northwestern University**
- **Vanderbilt University**

Summary Scenario

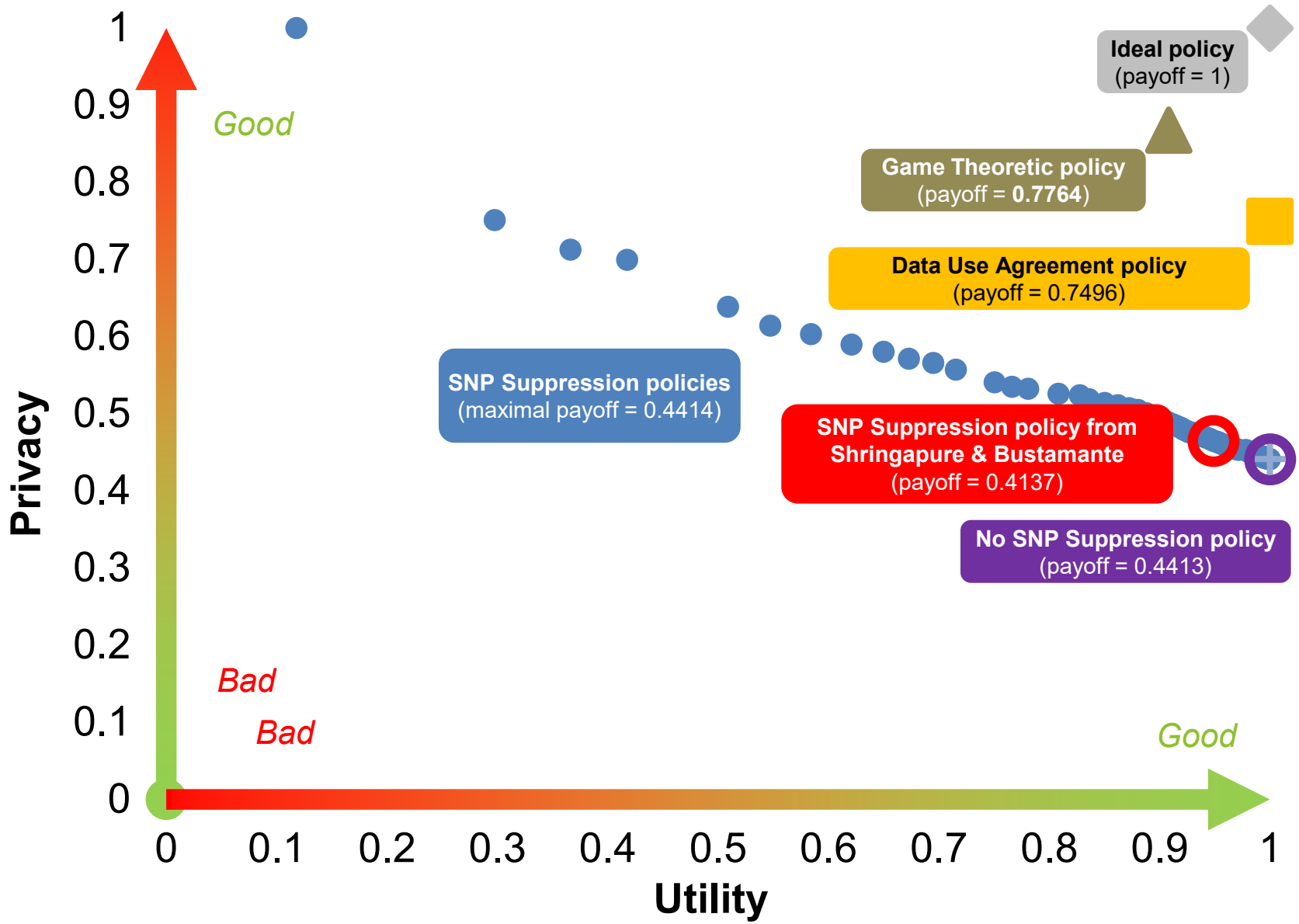
(Wan et al, AJHG 2017)

- 8194 individuals in SPHINX
- 2,500 SNPs to publish
  - Total of 38,112 genomic variants regions
    - 51,826 SNPs

- $45,000 max for publisher
- $360 for successful detection
- $180 for access to data
- Max allowable FPR of 10% and TPR of 50%
- Prior of 0.05 (estimated from data)

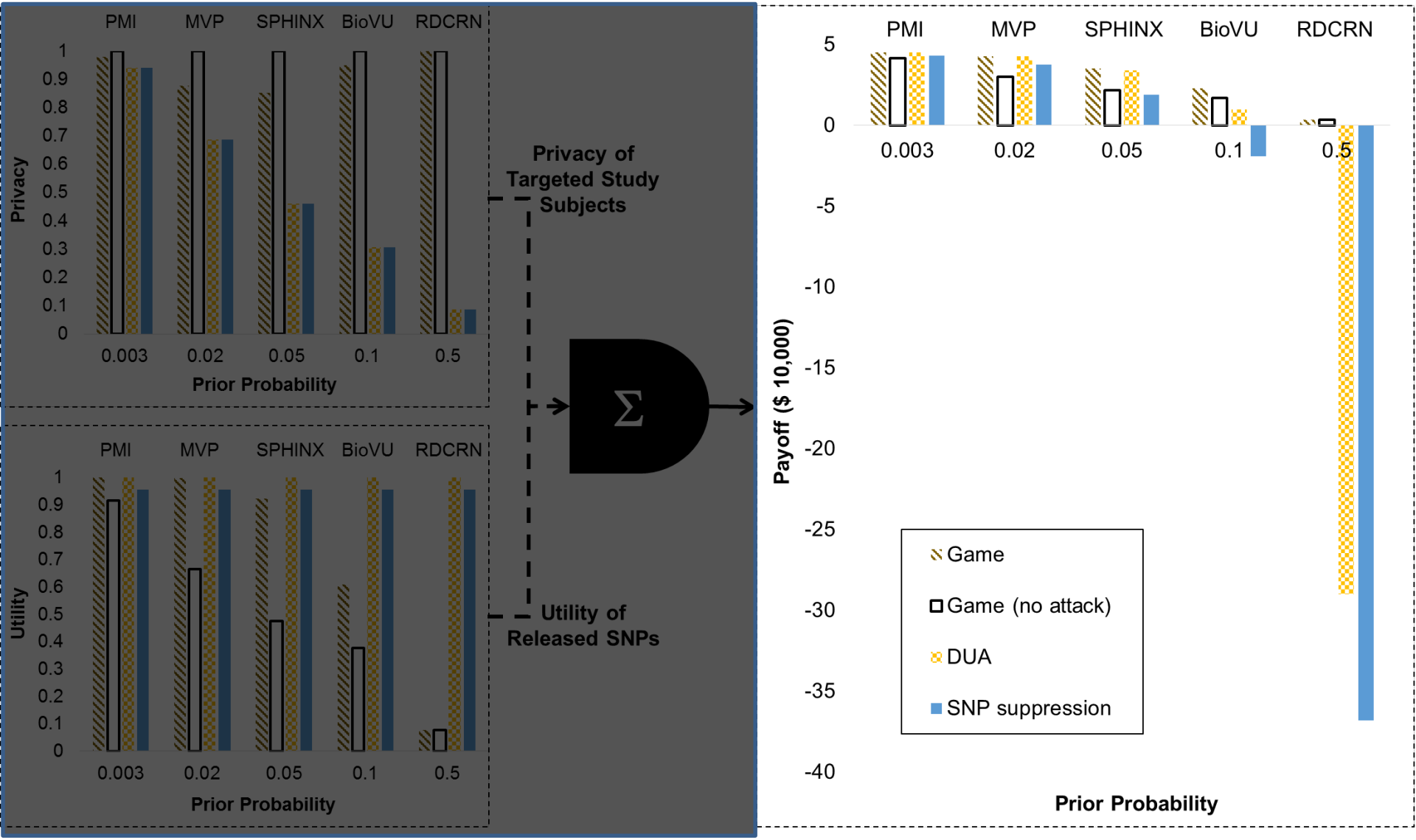
SPHINX Policy Analysis

(Wan et al, AJHG 2017)



Sensitivity Analysis on Prior Probability

(Wan et al. American Journal of Human Genetics. 2017)

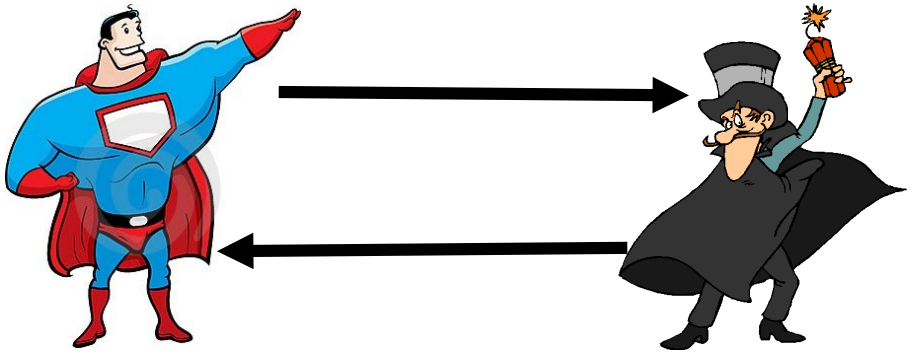


I’m yummy!

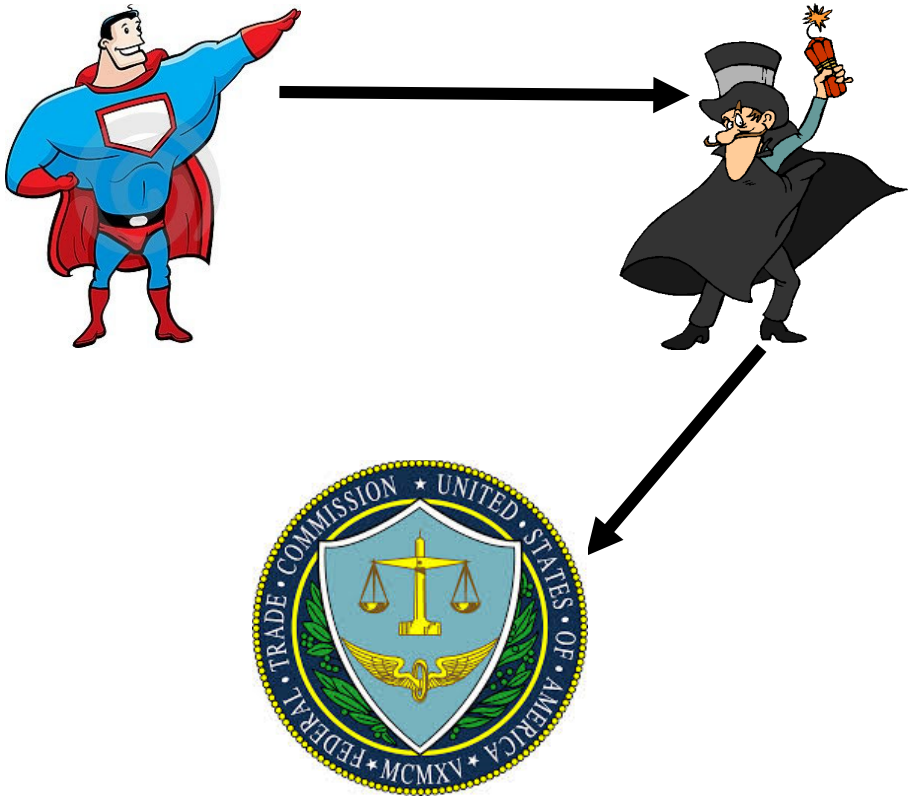
Market Design Matters

(Li et al, JBI 2016)

I'll Sue You for Breach!



Regulator Oversight



What Else to Worry About

- The number and diversity of “investigators” will grow
- Data will be more detailed and complex
- Risk analysis helps … but we need agreement on societally acceptable levels
- Alternative models of computation:
  - Secure multiparty computation
  - Synthetic records generation
  - … but there are always tradeoffs
- Breaches will happen … and responses need to be prepared

Acknowledgements

- John Aberdeen (MITRE)
- Kathleen Benitez
- David Carrell, Ph.D. (U. Washington)
- Ellen Wright Clayton, M.D., J.D.
- Josh Denny, M.D., M.S.
- Khaled El Emam, Ph.D. (U. Ottawa)
- Aris Gkoulalas-Divanis, Ph.D. (IBM)
- Jonathan Haines, Ph.D. (Case Western)
- Paul Harris, Ph.D.
- Raymond Heatherly, Ph.D. (Shyft Analytics)
- Lynette Hirschman, Ph.D. (MITRE)
- Murat Kantarcioglu, Ph.D. (U. of Texas)
- Jacqueline Kirby, M.S.
- Rachel Li, Ph.D.
- Jiuyong Li, Ph.D. (U. So. Australia)
- Yongtai Liu, M.S.
- Grigorios Loukides, Ph.D. (Kings College London)
- Dan Masys, M.D. (U. Washington)
- Dan Roden, M.D.
- Laura Rodriguez, Ph.D. (Geissinger)
- Weiyi Xia, Ph.D.
- Eugene Vorobeychik, Ph.D. (Washington U. of St. Louis)
- Zhiyu Wan, M.S.
- Zhijun Yin, Ph.D.

eMERGE Teams

- Boston Children's Hospital
- Children's Hospital of Philadelphia
- Cincinnati Children's Hospital
- Geisinger Health System
- Group Health Research Institute / U. Washington
- Marshfield Clinic
- Mayo Clinic
- Mt. Sinai Medical Center / Columbia University
- Northwestern University
- Vanderbilt University Medical Center

Funding

- NHGRI @ NIH
  - RM1 HG009034 (GetPreCiSe)
  - R01 HG00684 (RISK)
  - U01 HG006385 (eMERGE)
  - U01 HG006378 (VGER)
- NLM @ NIH
  - R01 LM009989

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/