Data Privacy Decision Spaces

Why Privacy is not Dead!

Bradley Malin, Ph.D.
Assoc. Prof. & Vice Chair of Biomedical Informatics, School of Medicine
Assoc. Prof. of Computer Science, School of Engineering
Affiliated Faculty, Center for Biomedical Ethics & Society
Vanderbilt University
22/9/2015

Acknowledgements

- Edoardo Airoldi, Ph.D. (Harvard U.)
- Kathleen Benitez
- Mustafa Canim, Ph.D. (IBM)
- Ellen Wright Clayton, M.D., J.D.
- Josh Denny, M.D.
- Xiaofeng Ding, Ph.D. (U. So. Australia)
- Khaled El Emam, Ph.D. (U. Ottawa)
- Aris Gkoulalas-Divanis, Ph.D. (IBM)
- Jonathan Haines, Ph.D. (Case Western)
- Raymond Healtherly, Ph.D. (Shyft Analytics)
- Murat Kantarcioglu, Ph.D. (U. of Texas)
- Jiuyong Li, Ph.D. (U. So. Australia)
- Grigorios Loukides, Ph.D. (Cardiff U)
- Dan Masys, M.D. (U. Washington)
- Dan Roden, M.D.
- Laura Rodriguez, Ph.D. (NHGRI / NIH)
- Latanya Sweeney, Ph.D. (Harvard U.)
- Acar Tamersoy, M.S. (Georgia Tech)
- Weiyi Xia, M.S.
- Eugene Vorobeychik, Ph.D.
- Zhiyu Wan

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

Funding

- NHGRI @ NIH
  - U01 HG006385 (eMERGE)
  - U01 HG006378 (VGER)
- NLM @ NIH
  - R01 LM009989
- Trustworthy Computing @ NSF
  - CCF-0424422 (TRUST)

U.S. National Institutes of Health Data Sharing "Policy"

- 2003 Final Data Sharing Policy:
  - Receive $500k → must have data sharing plan (or say why not possible)
  - Recommended sharing data devoid of identifiers
- 2014 Genome Data Sharing Policy
  - Studies involving > $0
- Identifiable?

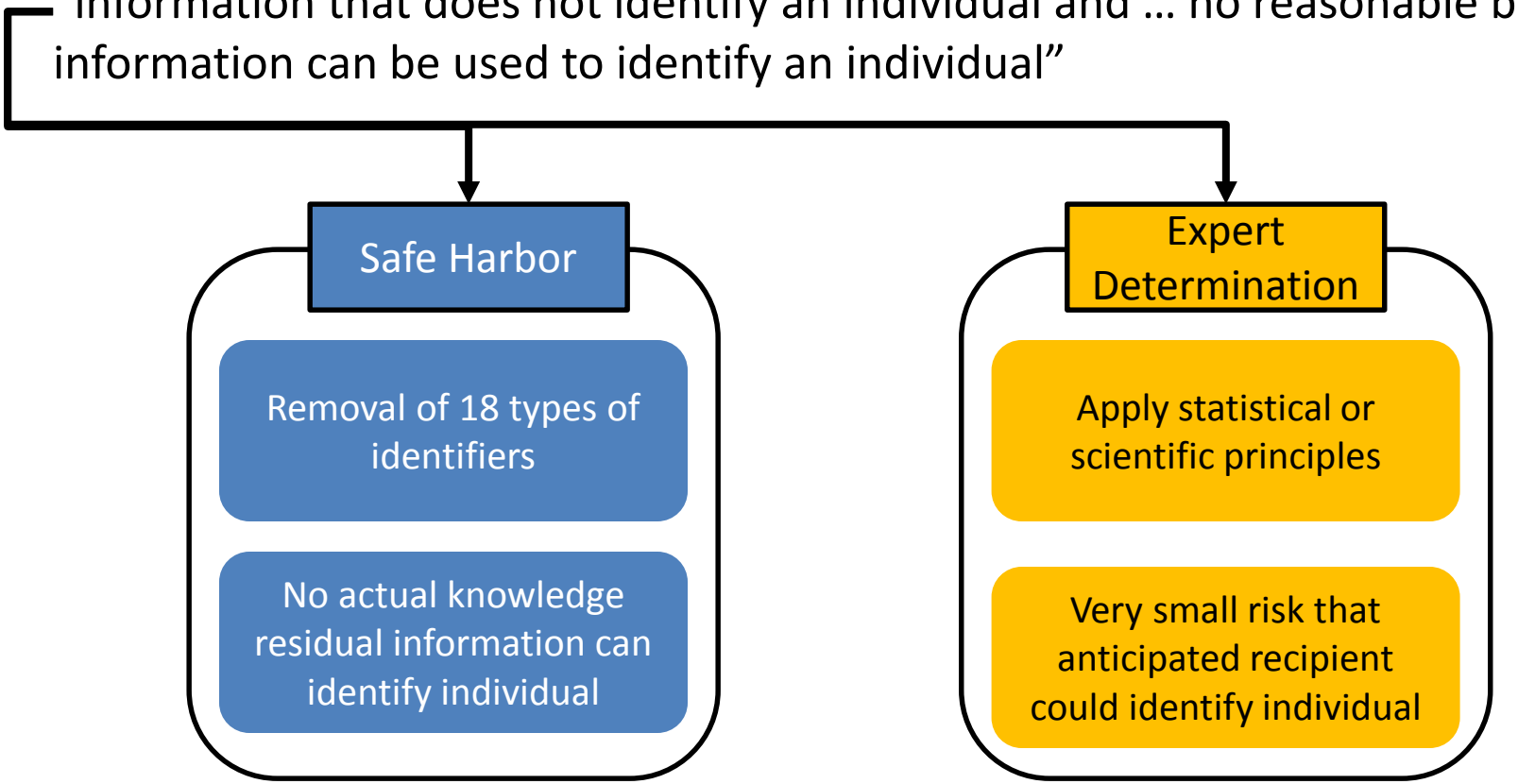
A Legal View Privacy

EU Data Protection Directive:

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

US Medical Regulation:

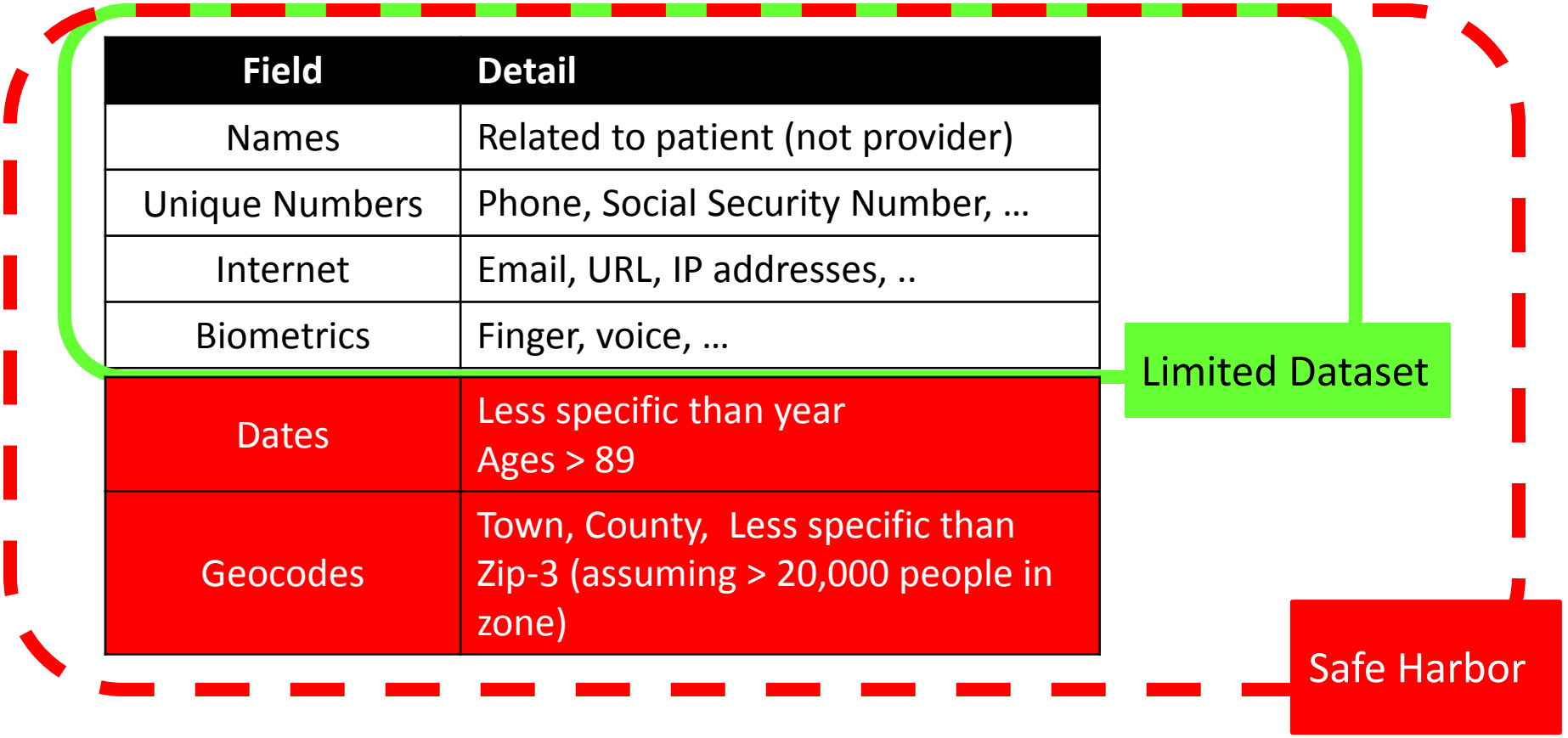
"information that does not identify an individual and … no reasonable basis … information can be used to identify an individual"



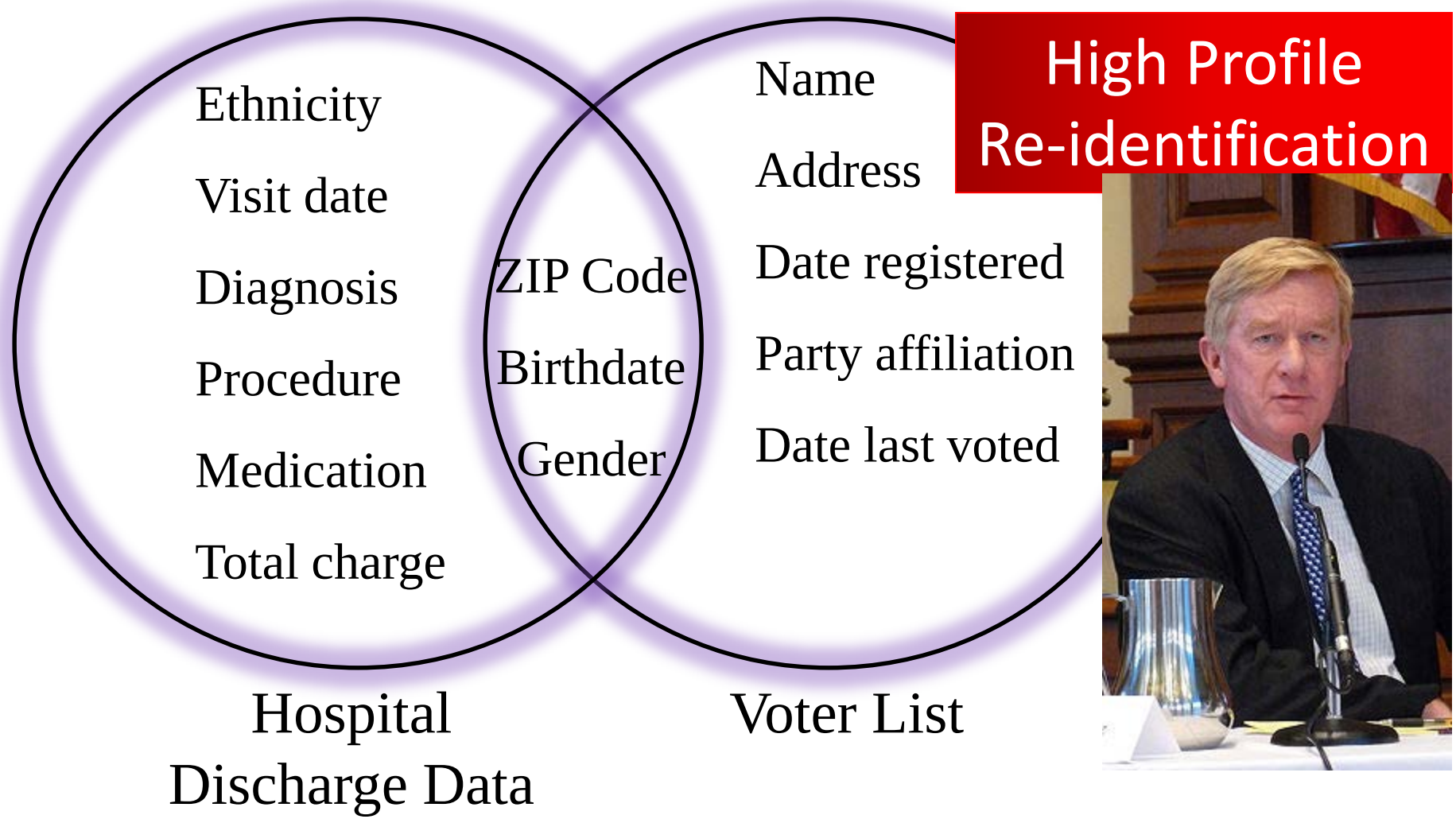
US Medical Privacy Rule

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	– See previous page

Recipes for Privacy



The Concern



High Profile Re-identification

Sweeney. Journal of Law, Medicine, & Ethics. 1997

5-Digit ZIP

+ Birthdate

+ Gender

63-87% of USA
estimated to be unique

Sweeney Tech Report 2000; Golle WPES 2006; Benitez & Malin JAMIA 2010

Set the World on

The AOL Search Log Case (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

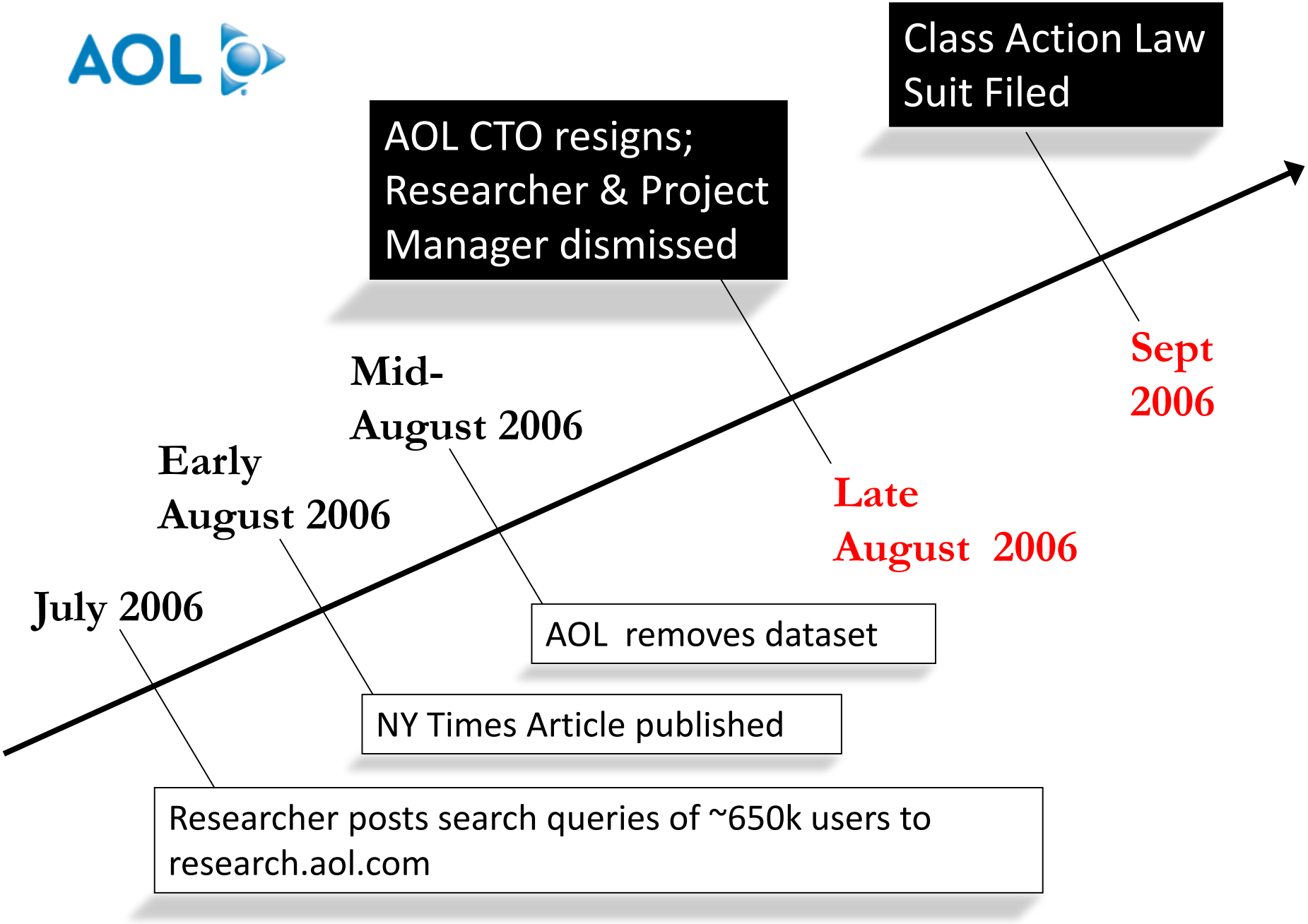
- 650 K customers, 20 M queries, 3 MONTH period
- Names replaced with persistent pseudonyms

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

Thelma Arnold

&

Dudley


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

[Your Favorite Feature] Distinguishes You!!

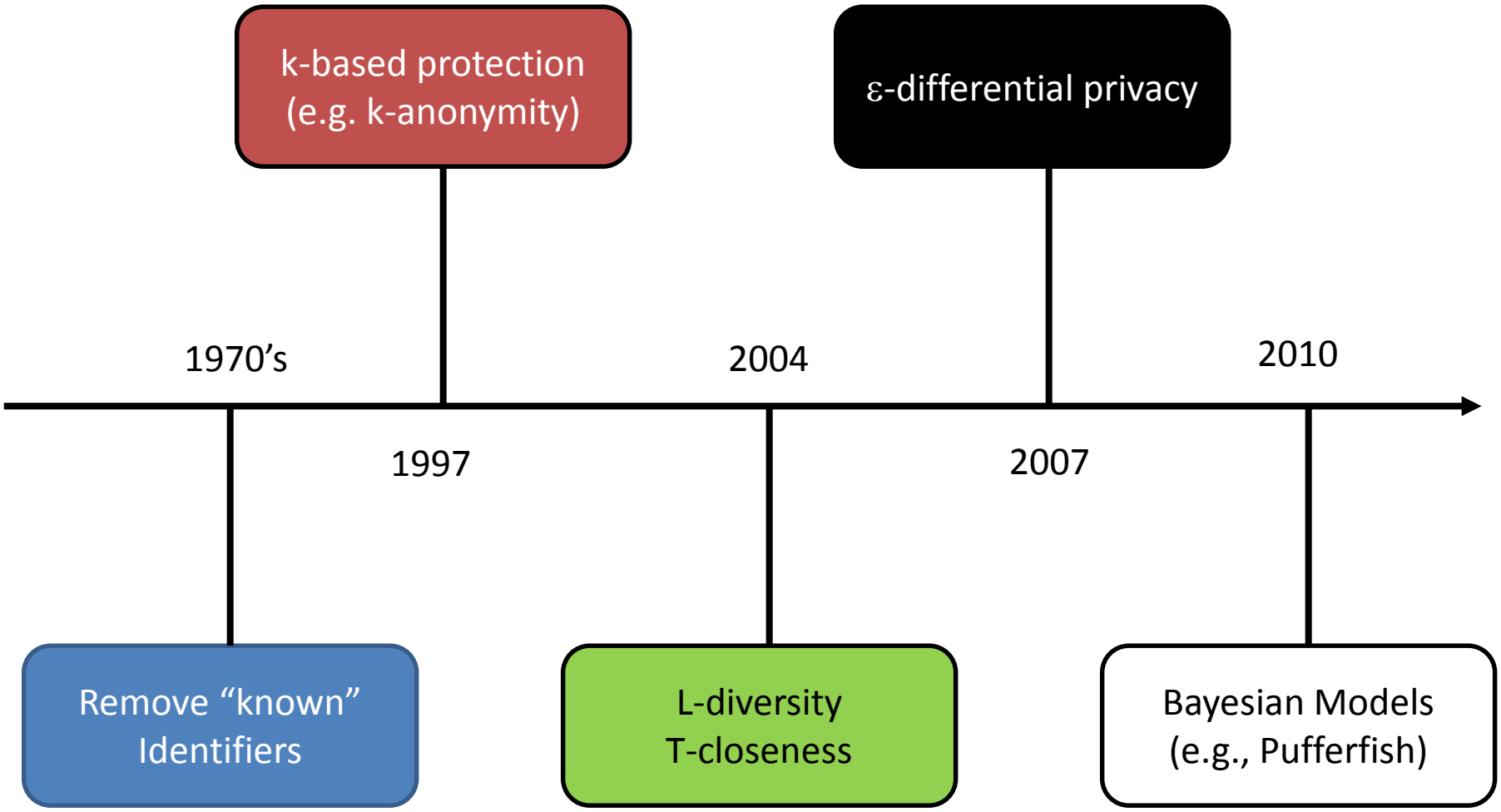
- Demographics (Sweeney ‘97; Bacher ‘02; Golle ‘06; El Emam ‘08; Koot ’10; Li ‘11)
- Diagnosis Codes (Loukides ’10; Tamersoy ‘10, ‘12)
- Laboratory Tests (Cimino ’12, Atreya ’13)
- DNA (Malin ‘00, Lin ‘04; Malin ‘05; Homer ‘08; Gymrek ’13, Ayday’14, Huttenhower ‘15)
- Health Survey Responses (Solomon ‘12)
- Location Visits (Malin ‘04; Golle ‘09; El Emam ‘11)
- Pedigree Structure (Malin ’06, Ayday ‘13)

- Movie Reviews (Narayanan ‘08)
- Social Network Structure (Backstrom ‘07; Narayanan ’09; Yang ‘12)
- Search Queries (Barbaro ‘06)
- Internet Browsing (Malin ‘05; Eckersley ’10; Banse ‘11; Herrmann ‘12, Olejnik ‘12)
- Smart Utility Meter Usage (Buchmann et al ‘12)

HIPAA Expert Determination (abridged)

Certify via "generally accepted statistical and scientific principles & 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."

A Brief History of Data Protection Models



k-Based Models

Age	Sex	Zip
30	M	15213
33	M	15217
33	F	15213
30	M	15213

Private Records

Age	Sex	Zip
30	M	15213
33	*	1521*
3*	*	15213
30	M	15213

2-Abiguous

Age	Sex	Zip
30	M	15213
33	*	1521*
33	*	1521*
30	M	15213

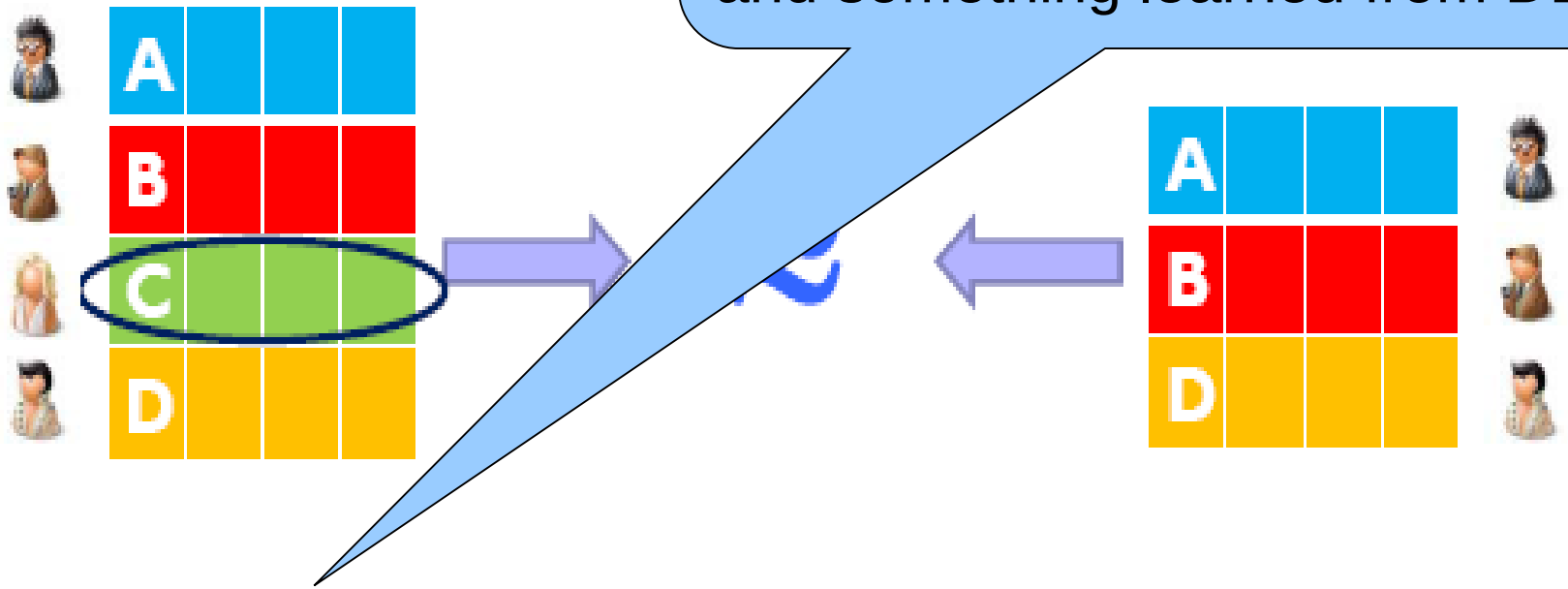
2-Anonymous

Thanks to Vitaly Shmatikov

Differential Privacy (informal)

Output is similar whether any is included in the database or

If there is already some risk of revealing a secret of C by combining auxiliary information and something learned from DB



C is no worse off because her record is included in the computation

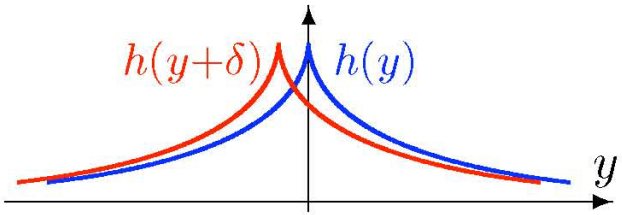
Thanks to Vitaly Shmatikov

Achieving DP with Laplace Noise

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$



Sliding property of $\mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

$\delta = f(x) - f(x') \leq \mathsf{GS}_f$

Science 7 August 2015:
Vol. 349 no. 6248 pp. 636-638
DOI: 10.1126/science.aaa9375

REPORT

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork[1,*], Vitaly Feldman[2,*], Moritz Hardt[3,*], Toniann Pitassi[4,*], Omer Reingold[5,*], Aaron Roth[6,*]

[1]Microsoft Research, Mountain View, CA 94043, USA.
[2]IBM Almaden Research Center, San Jose, CA 95120, USA.
[3]Google Research, Mountain View, CA 94043, USA.
[4]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada.
[5]Samsung Research America, Mountain View, CA 94043, USA.
[6]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA.
[*]Corresponding author. E-mail: dwork{at}microsoft.com (C.D.); vitaly{at}post.harvard.edu (V.F.); m{at}mrtz.org (M.H.); toni{at}cs.toronto.edu (T.P.); omer.reingold{at}gmail.com (O.R.); aaroth{at}cis.upenn.edu (A.R.)

Testing hypotheses privately

Large data sets offer a vast scope for testing already-formulated ideas and exploring new ones. Unfortunately, researchers who attempt to do both on the same data set run the risk of making false discoveries, even when testing and exploration are carried out on distinct subsets of data. Based on ideas drawn from differential privacy, Dwork *et al.* now provide a theoretical solution. Ideas are tested against aggregate information, whereas individual data set components remain confidential. Preserving that privacy also preserves statistical inference validity.

viewpoints

Communications of the ACM

DOI:10.1145/1743546.1743558

Arvind Narayanan and Vitaly Shmatikov

Privacy and Security

Myths and Fallacies of "Personally Identifiable Information"

Developing effective privacy protection technologies is a critical challenge for security and privacy research as the amount and variety of data collecte individuals increase exponentially.

THE DIGITAL ECONOMY relies on the collection of personal data on an ever-increasing scale. Information about our searches, browsing history, social relationships, medical history, and so forth is collected and shared with advertisers, researchers, and government agencies. This raises a number of interesting privacy issues. In today's data protection practices, both in the U.S. and internationally, "personally identifiable information" (PII)—or, as the U.S. Health Insurance Portability and Accountability Act (HIPAA) refers to it, "individually identifiable" information—has become the *lapis philosophorum* of privacy. Just as medieval alchemists were convinced a (mythical) philosopher's stone can transmute lead into gold, today's privacy practitioners believe that records containing sensitive individual data can be "de-identified" by removing or modifying PII.

Any information that distinguishes one person from another can be used for re-identifying data.

What is PII?

For a concept that is so pervasive in both legal and technological discourse on data privacy, PII is surprisingly difficult to define. One legal context is provided by breach-notification laws. California Senate Bill 1386 is a representative example: its definition of personal information includes Social Security numbers, driver's license numbers, financial accounts, but not, for example, email addresses or telephone numbers. These laws were enacted in response to security breaches involving customer data that could enable identity theft. Therefore, they focus solely on the types of data that

American Journal of Bioethics

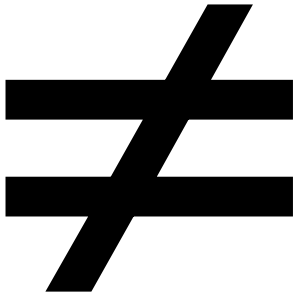
Published in final edited form as:
Am J Bioeth. 2010 September ; 10(9): 3–11. doi:10.1080/15265161.2010.494215.

Is Deidentification Sufficient to Protect Health Privacy in Research?

Mark A. Rothstein
University of Louisville School of Medicine

The revolution in health information technology has enabled the compilation and use of large data sets of health records for genomic and other research. Extensive collections of health records, especially those linked with biological specimens, are also extremely valuable for outcomes research, quality assurance, public health surveillance, and other

Possible



Probable

De-identification May Be Safe

- Reviewed all **actual** re-identification attempts
- Attacks on health data
  - 14 published re-identification attacks on any type of data
  - 11 of 14 were conducted by researchers as demo attacks
  - Only 2 of 14 attacks followed any standard
  - Only case with health data subject to “Safe Harbor” had a success rate of 0.00013

K. El Emam, et al. PLoS One. 2011.

A Case Study on Demographics

- Details at http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf
- Challenge issued by U.S. Dept. Health & Human Services
- Gave 15,000 Safe Harbor records to investigative team at University of Chicago
- Team purchased public records from commercial broker
- Correctly identified 2 people

Considerable Knowledge

- Identifiability is proportional to
  - Uniqueness (must distinguishable) x
  - Replicability (must be reproducible) x
  - Availability (must be accessible)

- *A drug dose may be unique, but may not be accessible to the public in any known resource*

Only 1 in 650,000 people identified

Thelma Arnold

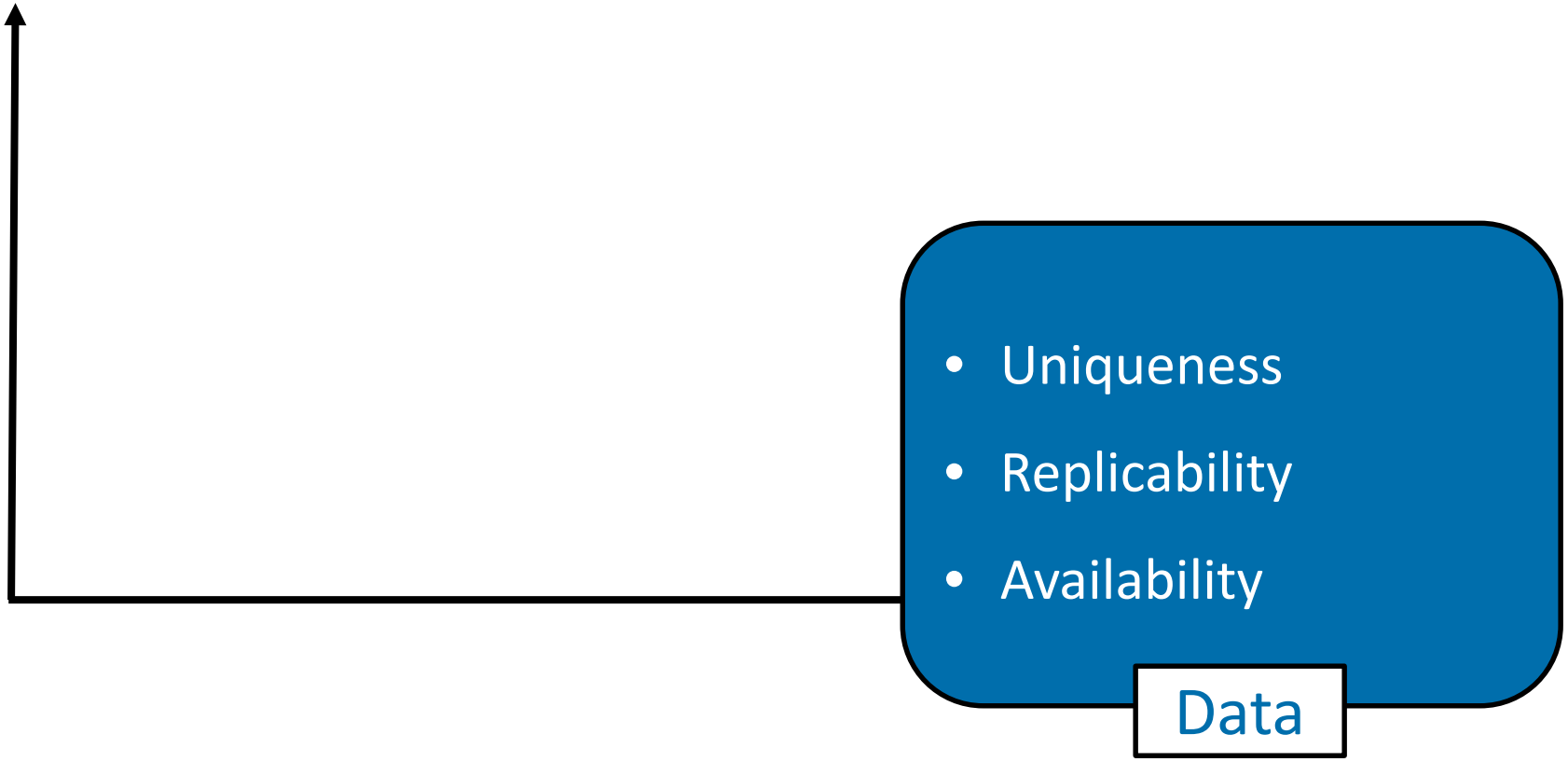
&

Dudley

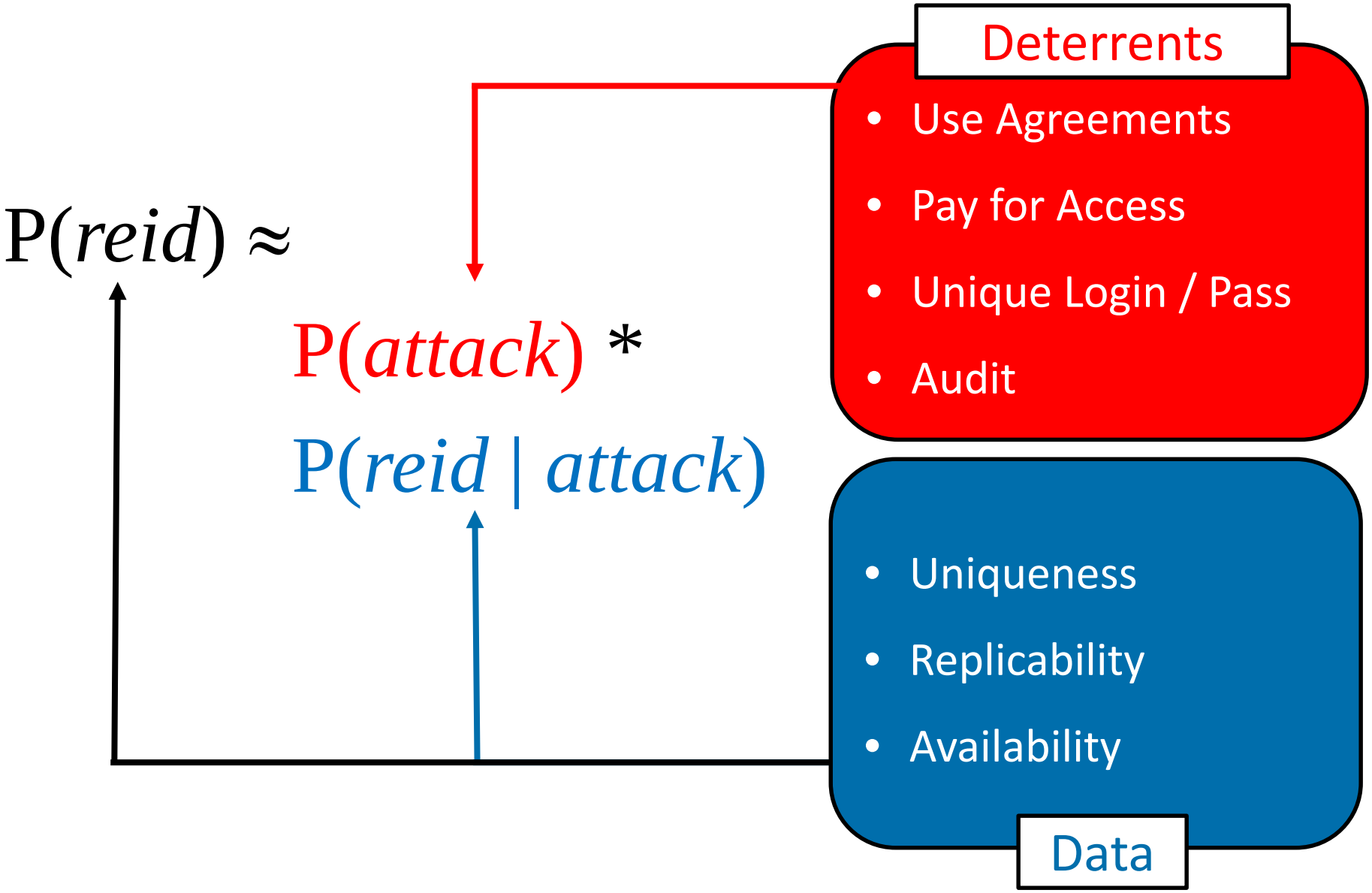
What’s Going On?

A Very Simplified View on Risk

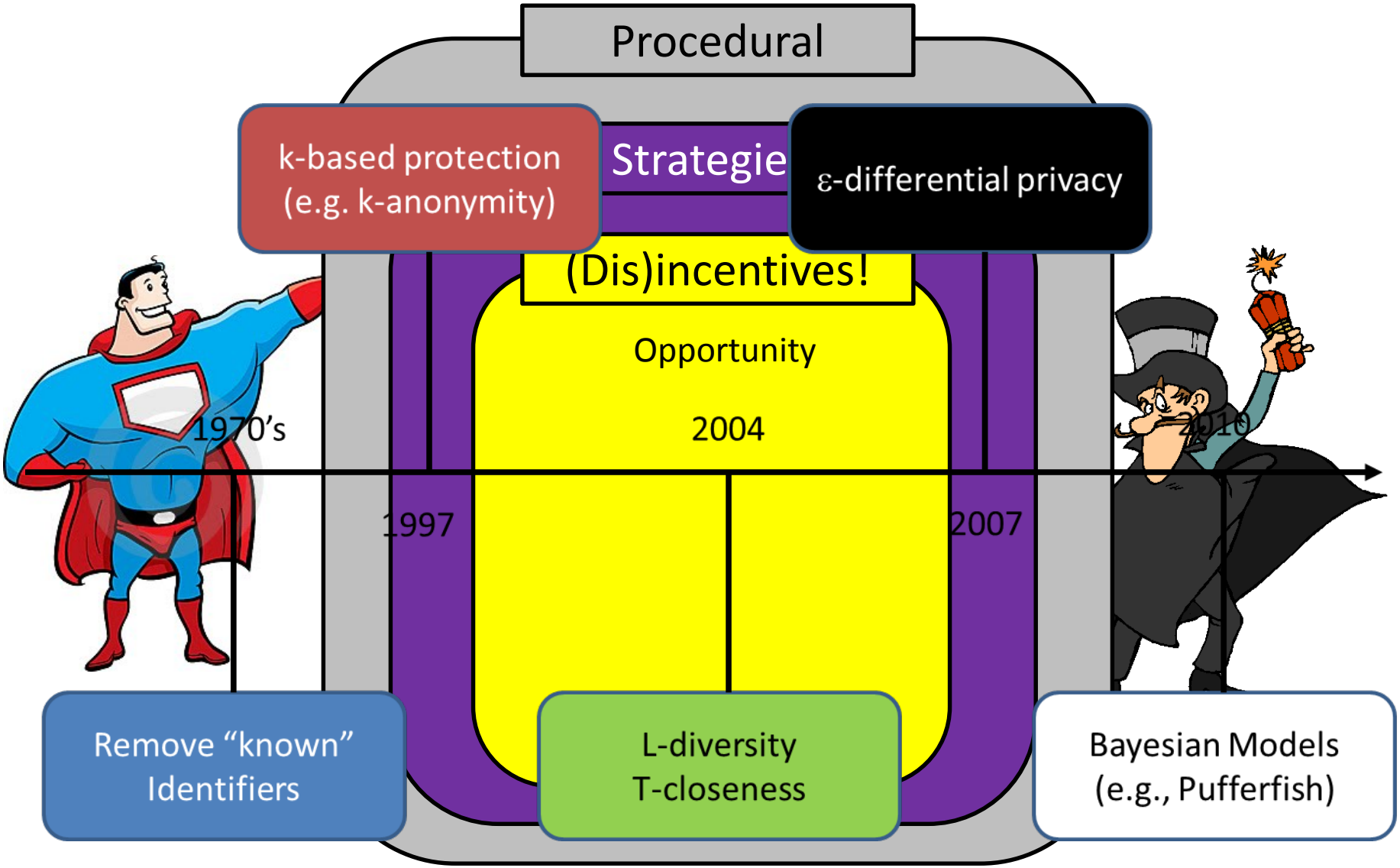
$$\mathrm{P}(reid) \approx$$



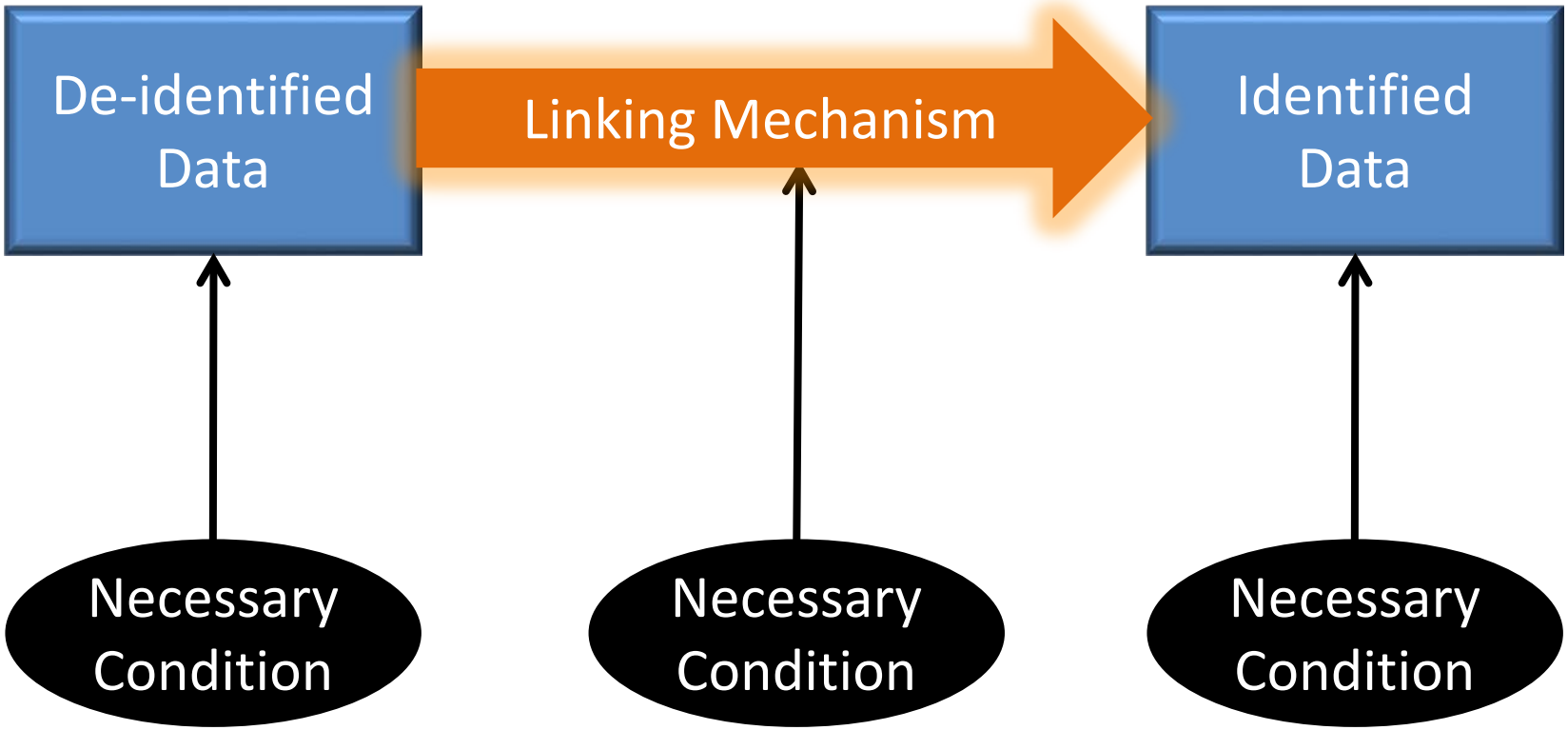
A Very Simplified View on Risk



An Augmented View of Data Privacy



Central Dogma of Re-identification



Malin, Benitez, Loukides, Clayton. Human Genetics. 2011.

…We've Been Looking at Worst Case Scenarios...

- How would you ***use*** demographics?

- Could link to registries
  - Birth
  - Death
  - Marriage
  - Professional (Physicians, Lawyers)

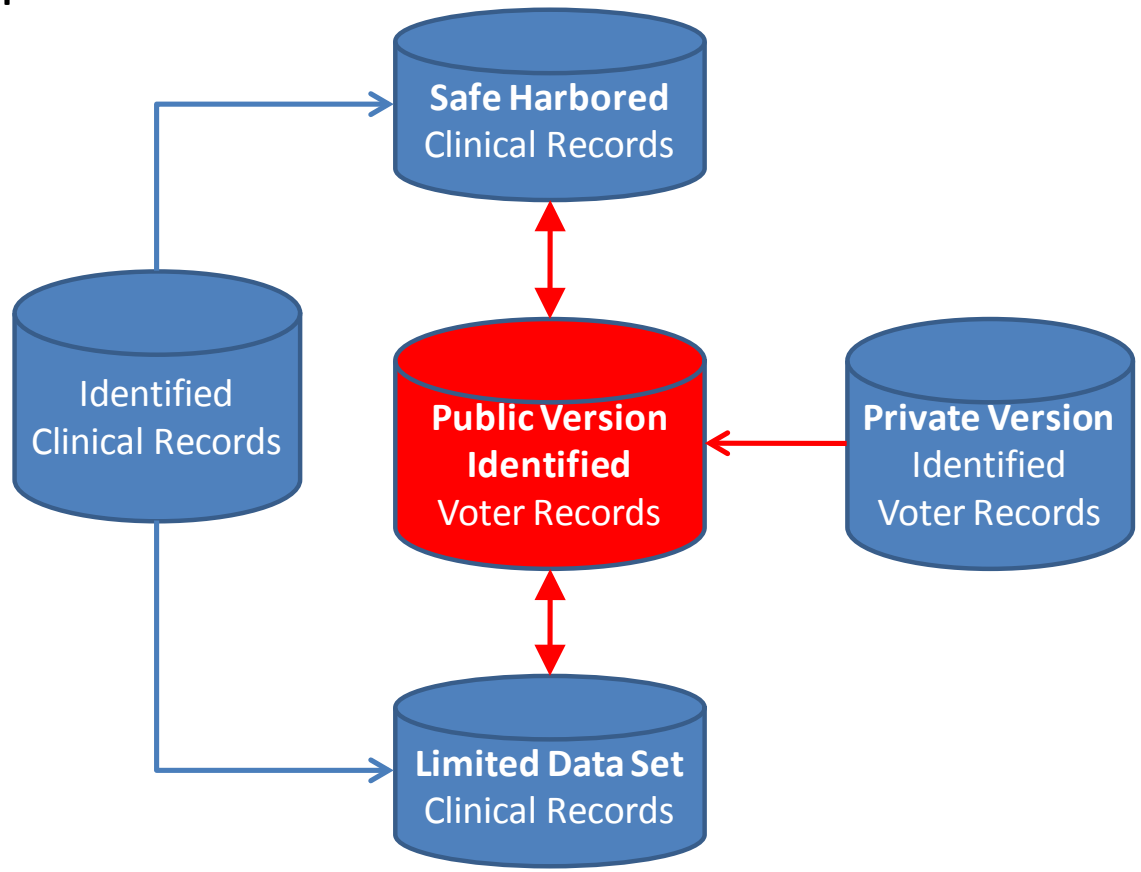
- What's in vogue?

  Back to voter registration databases

Going to the Source

- We polled all U.S. states for what voter information is collected & shared

- What fields are shared?
- Who has access?
- Who can use it?
- What's the cost?



Benitez & Malin. JAMIA. 2010.

The Availability of Demographics Varies…

	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
Name	●	●	●	●	●
Address	●	●	●	●	●
Date of Birth	●	○	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

Benitez & Malin. JAMIA. 2010.

Who is Like You?

Time for Estimation (Golle 2006)

- Disclose sample with {*dob*, *gender*, *zip*}, but don't know the population's values
- Don't always have exact knowledge of what a data recipient has access to

		Birthdate		
		1/1/80	…	12/31/80
ZIP	zip1			
	zip2			
	…			
	zip m			

Time for Estimation (Golle 2006)

- May know population counts, such as
  - Census aggregates for {*year of birth, gender, county*}

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

		Birthdate		
		1/1/80	…	12/31/80
ZIP	zip1			
	zip2			
	…			
	zip m			

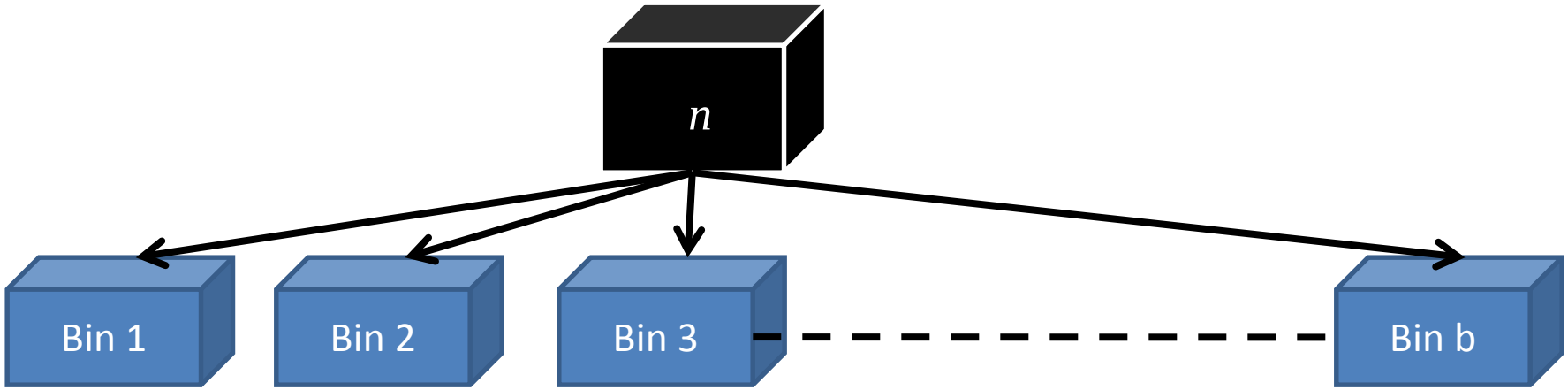
Time for Estimation (Golle 2006)

- May know population counts, such as
  - U.S. Census aggregates for {*year of birth, gender, county*}
- Use a disaggregation model to estimate

		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

It's an Occupancy Problem (Golle 2006)



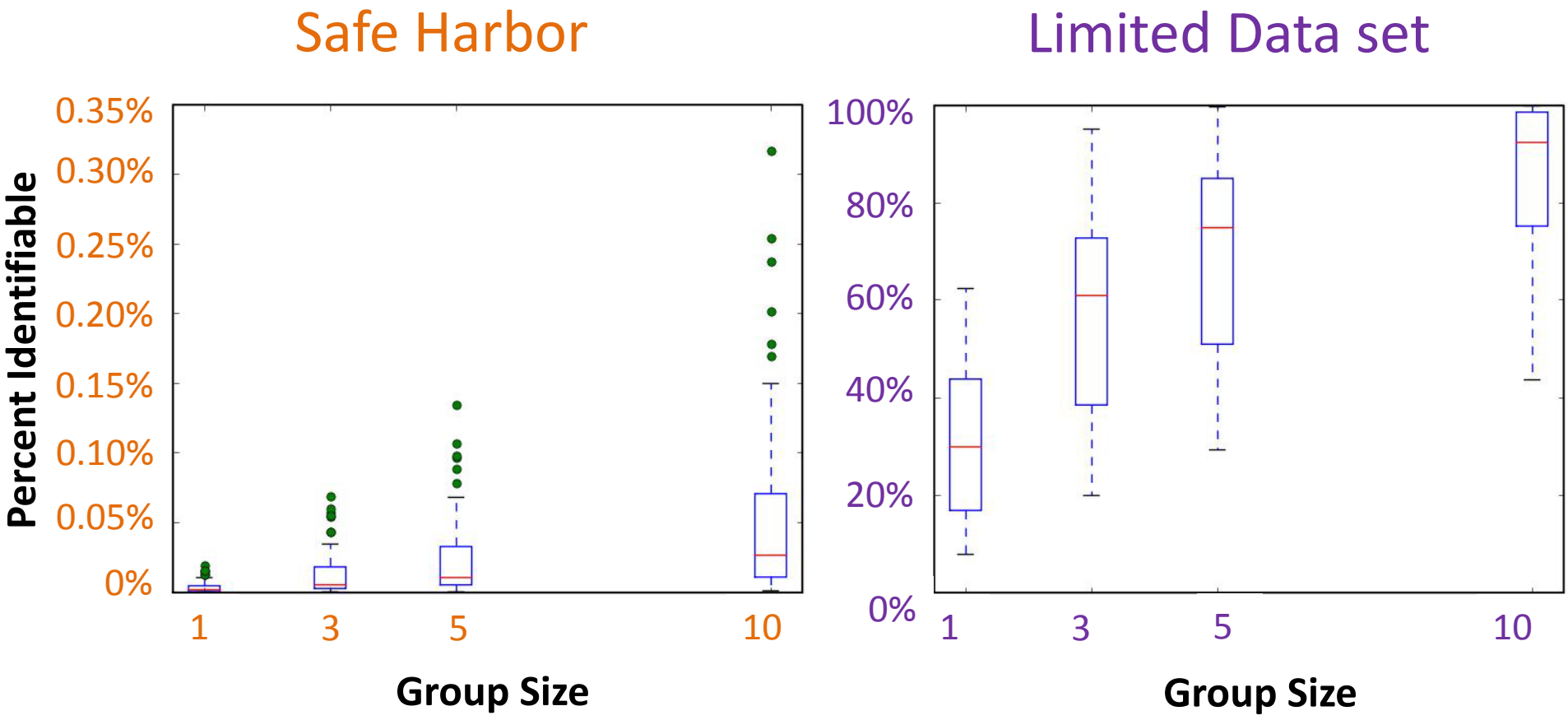
- n people in aggregated bin
- b disaggregated bins
- the expected # of bins with exactly i people

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

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

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

All U.S. States



Benitez & Malin, JAMIA. 2010.

Risk…

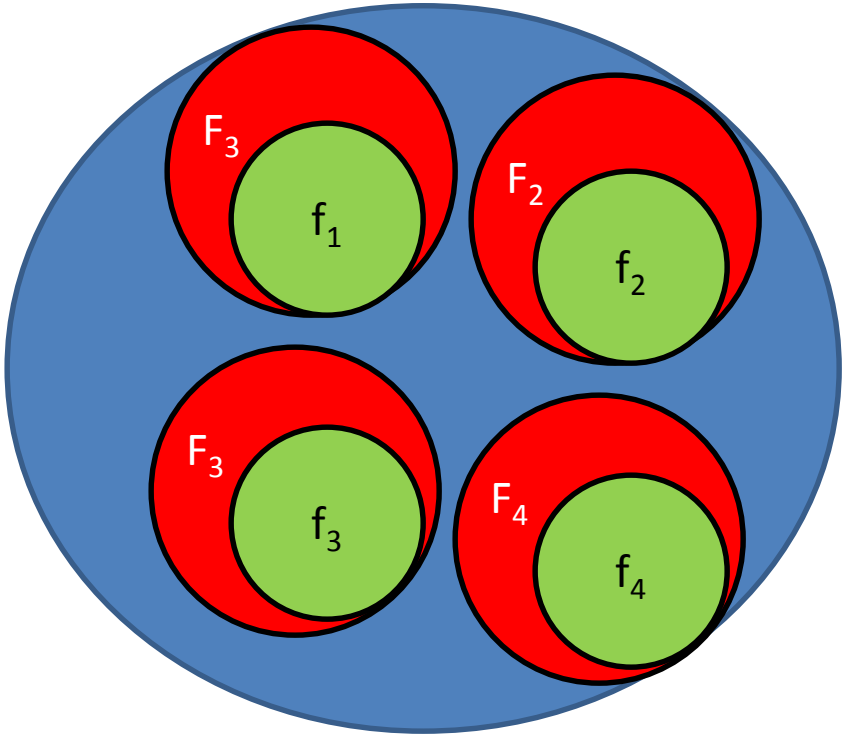
… means something different to everyone

- 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



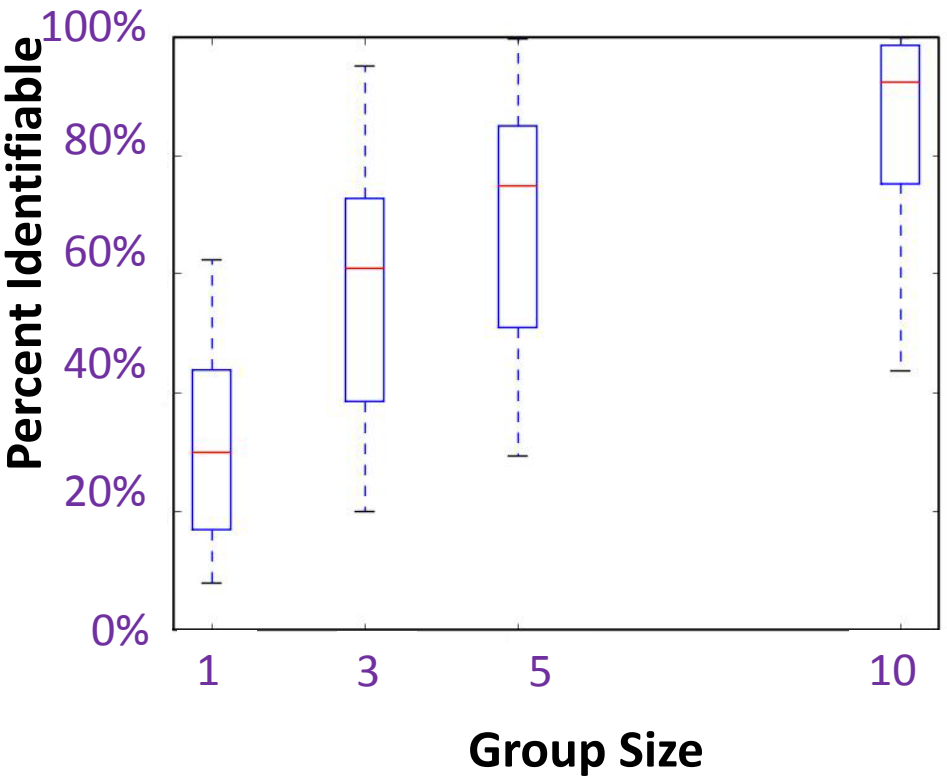
The Availability of Demographics Varies…

	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
Name	●	●	●	●	●
Address	●	●	●	●	●
Date of Birth	●	○	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

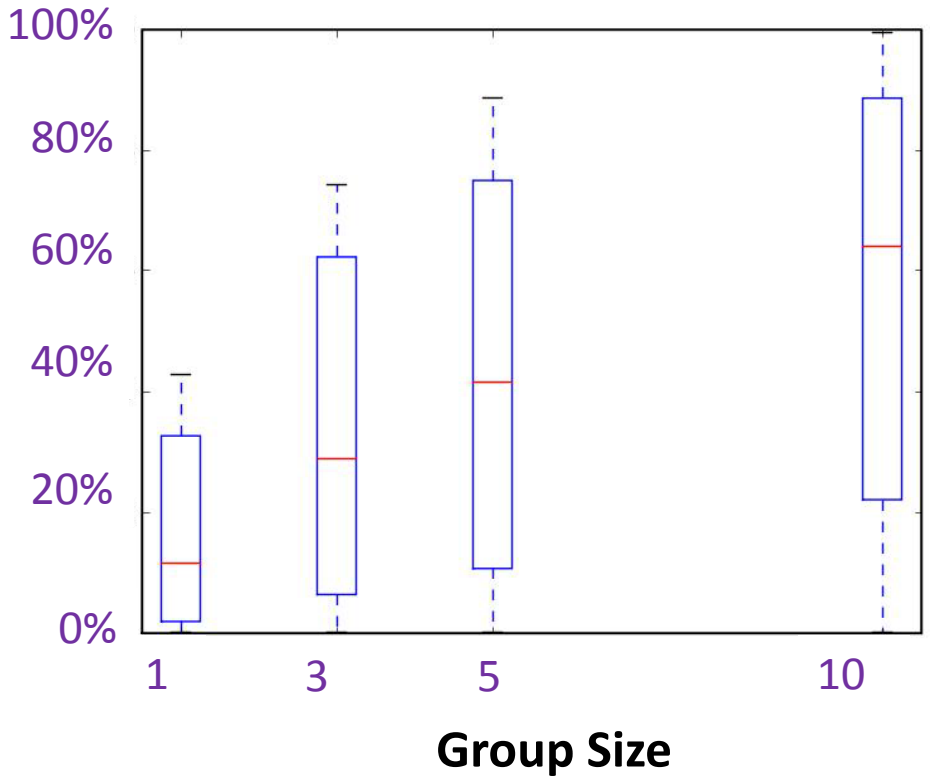
Benitez & Malin. JAMIA. 2010.

Identifiability Changes!

Limited Data Set



Limited Data Set ←→ Voter Reg.



Benitez & Malin. JAMIA. 2010.

What About Cost?

(Consider Marketer Risk)

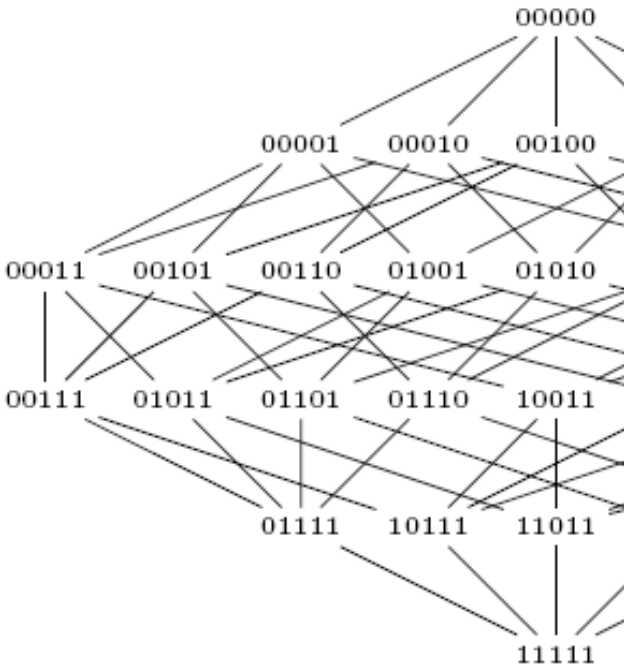
U.S. State	Limited Dataset		Safe Harbor	
	At Risk	Cost per Re-id	At Risk	Cost per Re-id
Virginia	3159764	$0	221	$0
South Carolina	2231973	$0	1386	$0

Benitez & Malin. JAMIA. 2010.

We Need Policy Alternatives

How Can we Find Policies?

- Model acceptable data abstractions as a lattice and "search" for low risk



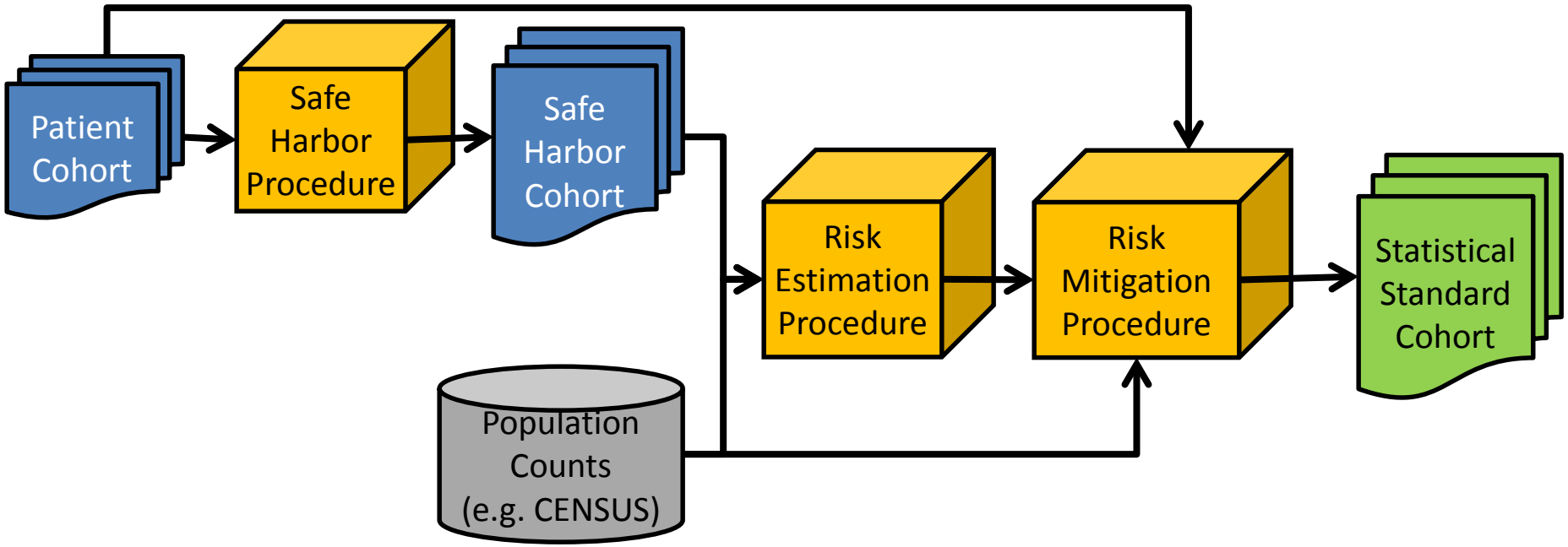
Original Values			
Name	Gender	Age	Diagnosis
Sister Susie	F	24	(L)eukemia
Jack Sprat	M	20	(H)ypertension
Mary Contrary	F	20	(M)yocardial Infarction
Boy Blue	M	21	(M)yocardial Infarction
King Cole	M	23	(D)iabetes
Jill Hill	F	22	(D)iabetes
Jack Hill	M	22	(H)ypertension

Policy 00000		
Gender	Age	Diag.
[M,F]	[20-24]	L
[M,F]	[20-24]	H
[M,F]	[20-24]	M
[M,F]	[20-24]	M
[M,F]	[20-24]	D
[M,F]	[20-24]	D
[M,F]	[20-24]	H

Policy 10010		
Gender	Age	Diag.
F	[23-24]	L
M	[20-22]	H
F	[20-22]	M
M	[20-22]	M
M	[23-24]	D
F	[20-22]	D
M	[20-22]	H

Policy 11111		
Gender	Age	Diag.
F	24	L
M	20	H
F	20	M
M	21	M
M	23	D
F	22	D
M	22	H

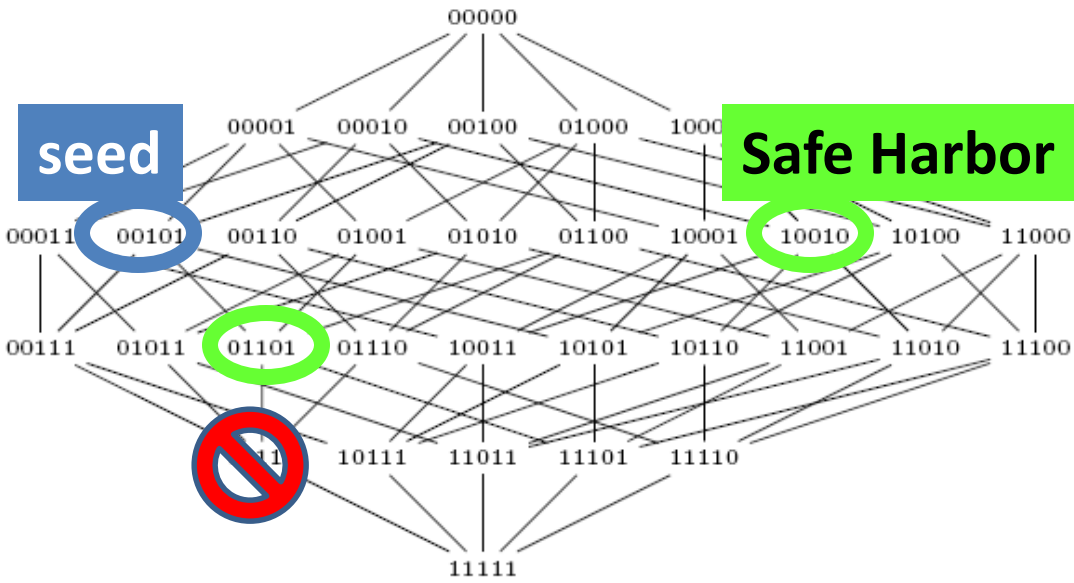
Simple Discovery Decision



Benitez, Loukides, and Malin. ACM IHI. 2010.
Malin, Benitez, and Masys. JAMIA. 2011.
Xia, et. Al. ACM CODASPY. 2013

Searching the Lattice

- Risk is monotonic on the graph
- Search space is huge and there are multiple "optimal" solutions
- We can search for "good" solutions using a ILP
- Faster - bisecting strategy



Benitez, Loukides, and Malin. ACM IHI. 2010.

Vandy ECG Case Study

Who	State	State Population Size (2000 Census)	Cohort Size	Patients >89 years old
Vanderbilt	TN	5,689,283	2,983	12

Policy	Generalizations			Risk
	Gender	Race	Age	
Safe Harbor	∅	∅	[90 - 120]	0.909
Alternative 1	[M or F]	∅	∅	0.476
Alternative 2	∅	[Asian or Other]	∅	0.857
Alternative 3	∅	∅	[52 - 53]	0.875

Benitez, Loukides, and Malin. ACM IHI. 2010.

Evaluation of De-id Model

- Cohorts from the Electronic Medical Records and Genomics Consortia (http://www.gwas.net)

Pheno.	Cohort	Who	State	State Population Size (2000 Census)	Clinical Finding of Interest	Cohort Size	Patients >89 years old
Primary	G_{Dem}	GHC	WA	5,894,121	Dementia	3,616	1,483
	R_{Cat}	Marshfield	WI	5,363,675	Cataracts	2,646	269
	Y_{PAD}	Mayo	MN	4,919,479	Peripheral Arterial Disease	3,412	29
	N_{T2D}	Northwestern	IL	1,2519,293	Type-II Diabetes	3,383	6
	V_{QRS}	Vanderbilt	TN	5,689,283	QRS Duration	2,983	12
Quality Control	N_{QRS}	Northwestern	IL	1,2519,293	QRS Duration	149	0
	V_{T2D}	Vanderbilt	TN	5,689,283	Type-II Diabetes	2,015	18

Malin, Benitez, & Masys. JAMIA. 2011.

Analysis for eMERGE

Risk Model: Uniques

Are the number of uniques expected to be greater than Safe Harbor?

Disclosure Policy	Acceptable?						
	G_{DEM}	R_{CAT}	Y_{PAD}	N_{T2D}	V_{QRS}	N_{QRS}	V_{T2D}
Generalized Ethnicity (Black, White, Other)							
Age at 5 Year Bins							
Generalized Ethnicity AND Age at 5 year bins							
Age at 10 Year Bins							

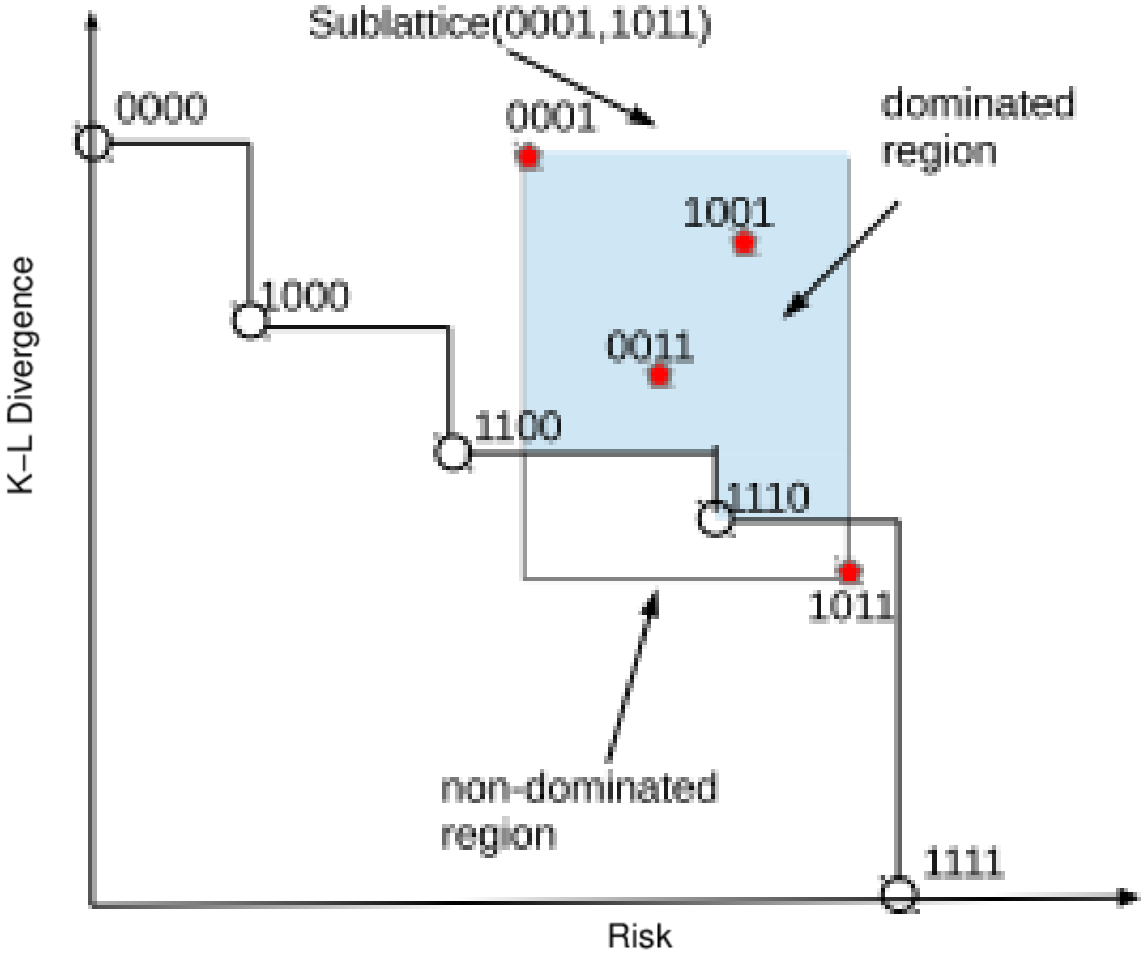
Red = more risk than Safe Harbor

Green = risk no worse than Safe Harbor

Malin, Benitez, & Masys. JAMIA. 2011.

Dual Optimization Extension (Risk-Utility Frontier)

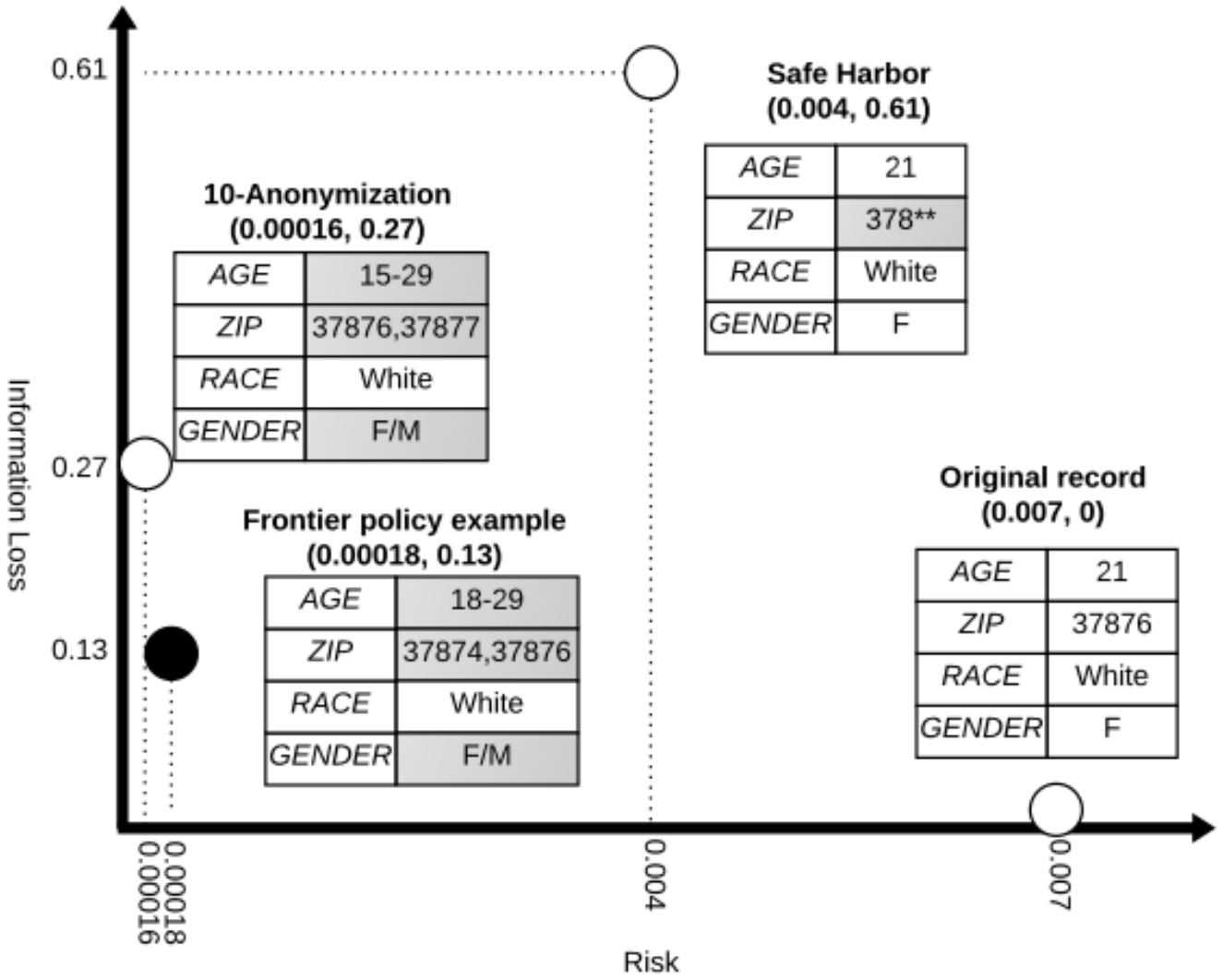
- Privacy = Marketer Risk
- Utility = KL-divergence of quasi-ID distribution
- Both are monotonic over paths in the lattice
- Apply sublattice heuristic to search for regions of interest (proportion of region with potential to dominate)



Xia, Heatherly, Ding, Li, and Malin; 2013; 2015

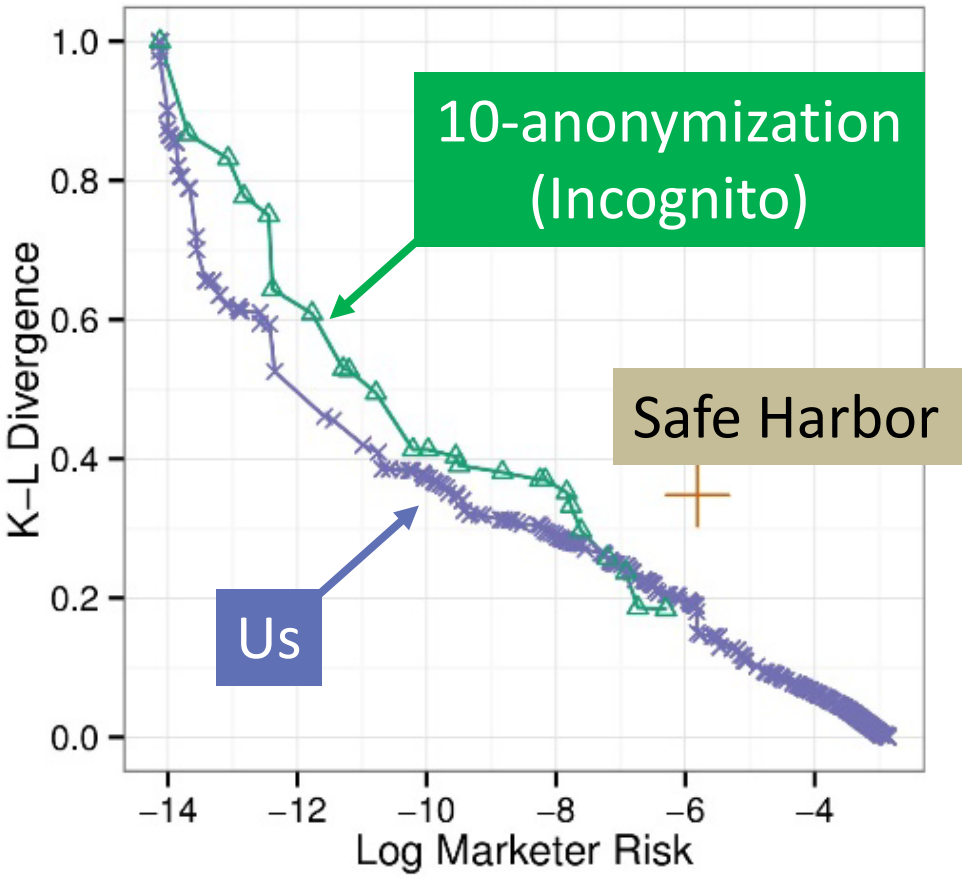
Enhancement: Risk-Utility Frontiers

(Xia, et al. ACM CODASPY. 2013; JAMIA. 2015)



Example: State of Hawaii

- Simulation of 30,000 records from the Adult Census Database
  - {Age, Race, Gender}
- Appended 5-digit ZIP proportional to public use statistics
- Space of 2^{700} policies
- Frontier of ~400 policies discover by evaluating ~20,000



We are Driven By Incentives

(under rational assumptions)

Stackelberg Game

Sharing Strategy 1
Utility 1
Risk ???

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

⋮

Attack Strategy C
Utility C
Risk C

Strategies:

- Generalize Demographics
- Perturb Statistics
- Apply Data Use Agreement

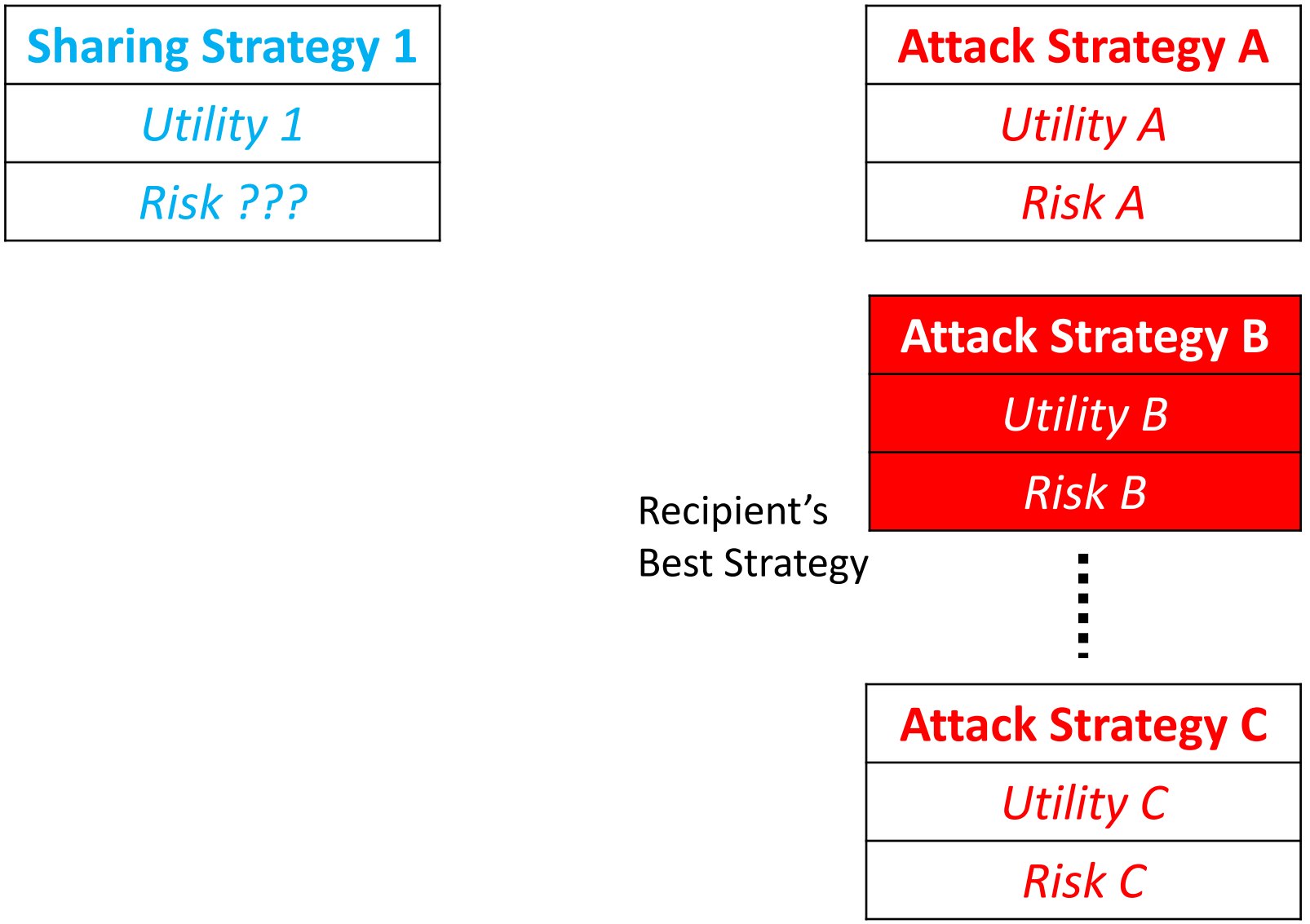
…

- Charge for Access

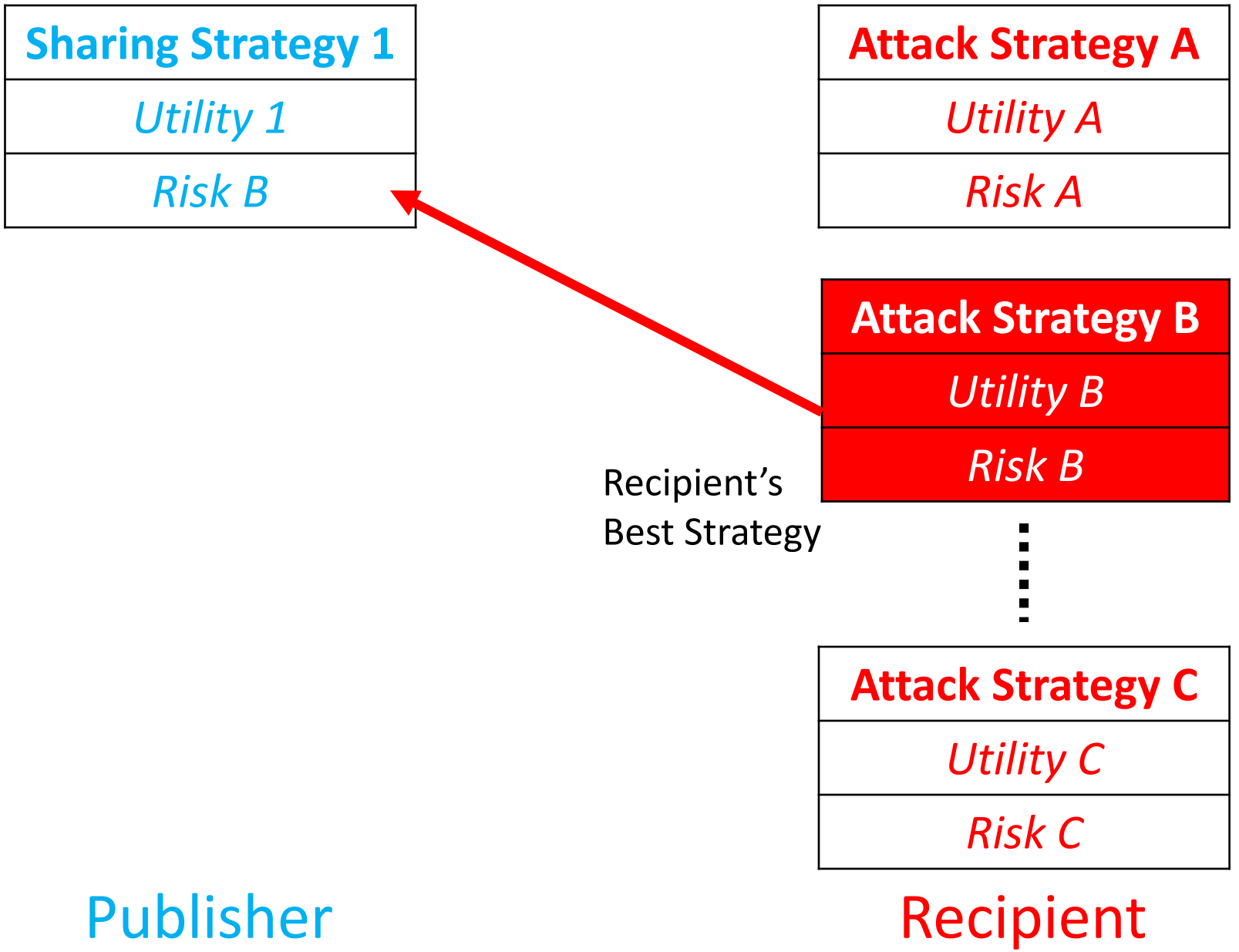
Publisher

Recipient

Stackelberg Game



Stackelberg Game



Stackelberg Game

Sharing Strategy 1
Utility 1
Risk B

Sharing Strategy 2
Utility 2
Risk ???

Attack Strategy A
Utility A
Risk A

Attack Strategy B
Utility B
Risk B

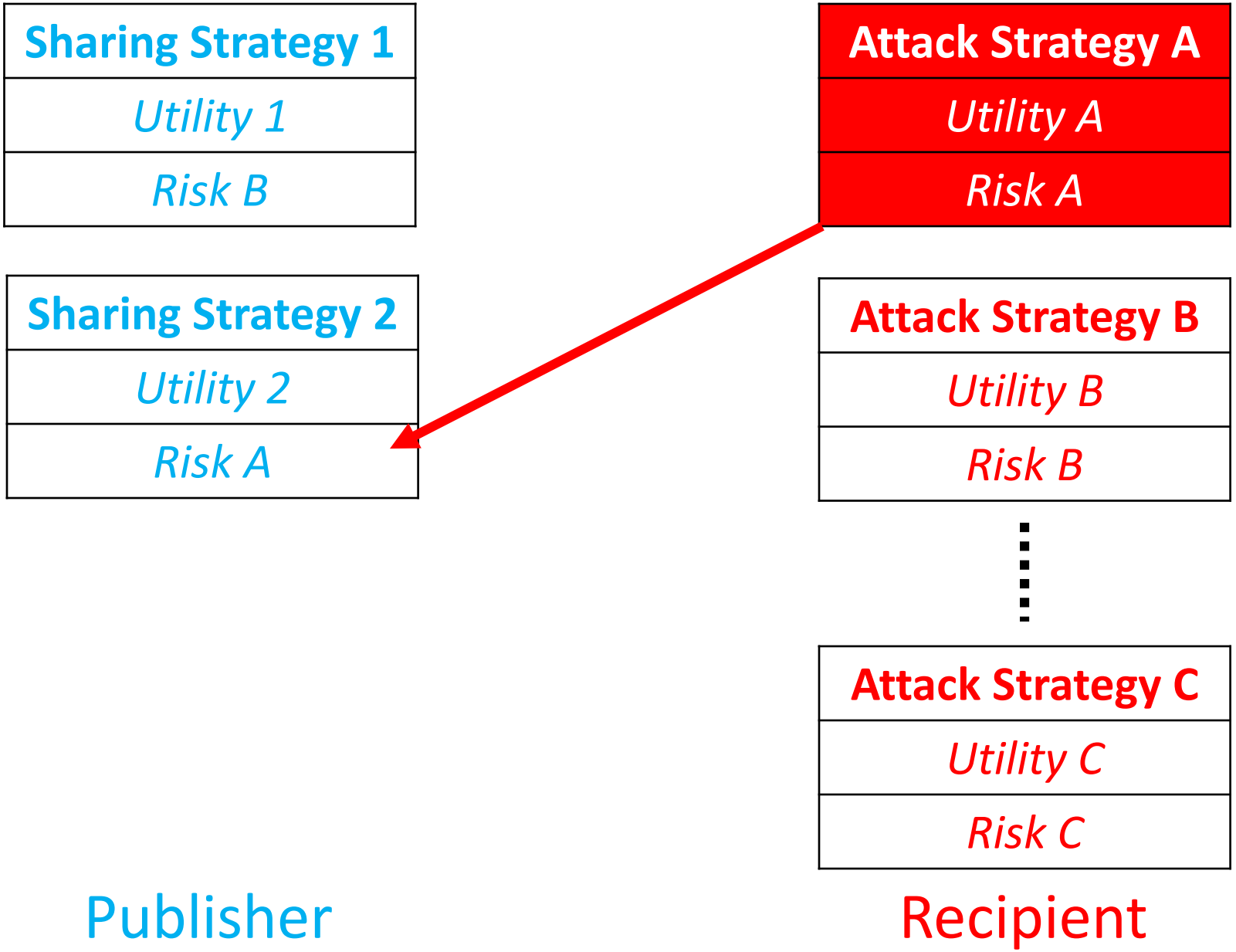
⋮

Attack Strategy C
Utility C
Risk C

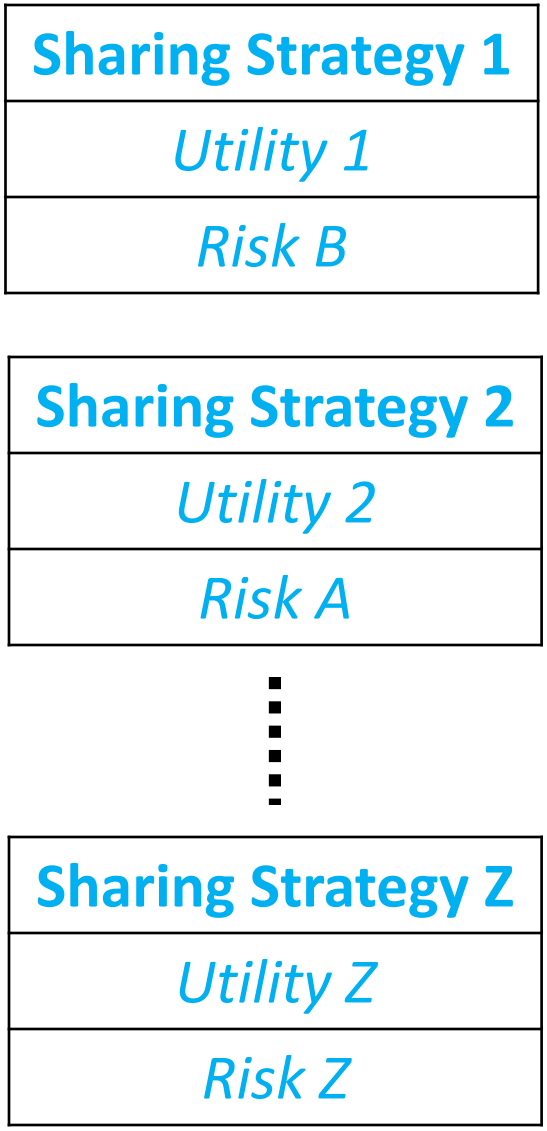
Publisher

Recipient

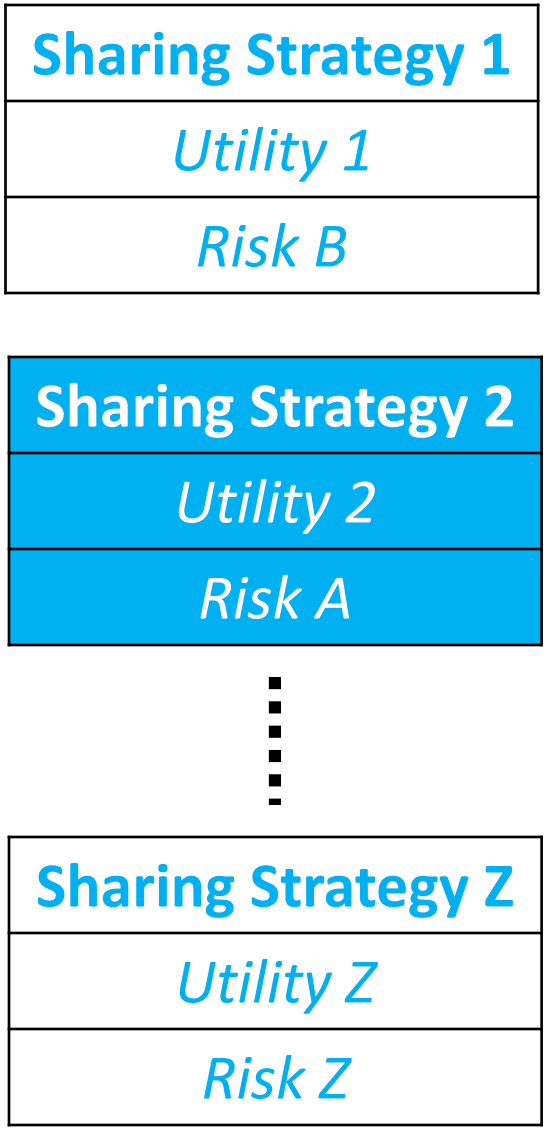
Stackelberg Game



Stackelberg Game



Stackelberg Game



Choose Strategy that maximizes overall benefit

- Optimal Utility / Risk Tradeoff

Payoffs!

Gain	No Attack	Attack
Publisher	v_g	
Attacker	0	

g : Generalization level

v_g : Value of record at g

Wan et al, PLoS One. 2015

Payoffs!

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

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

Wan et al, PLoS One. 2015

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

Solving the Game?

- The sublattice search will work… but it's not optimal

- Alternatives
  - Backward Induction Search
    - For each generalization level
      - choose the one that maximizes publisher's utility
    - Exhaustive search over combinatorial space of data representation

  - ILP or something else.

Wan et al, PLoS One. 2015

Intelius - Find People with

www.intelius.com

Apps | Barclays | ScholarOne Manusc... | HRPP/IRB | Vanderb... | VBA Express : Excel -... | Home - PubMed - ... | Other bookmarks

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!

Entity	Fine	#Records	Fine/record	Date
New York and Presbyterian Hospital	$4,800,000	6,800	$705.9	May 7, 2014
QCA Health Plan, Inc.	$250,000	148	$1689.2	Apr 22, 2014
Skagit County, Washington	$215,000	118,000	$1.8	Mar 7, 2014
Adult and Pediatric Dermatology	$150,000	2,200	$68.2	Dec 26, 2013
Affinity Health Plan, Inc.	$1,215,780	344,579	$3.5	Aug 14, 2013
WellPoint Inc.	$1,700,000	612,402	$2.8	Jul 11, 2013
Idaho State University	$400,000	17,500	$22.9	May 21, 2013
The Hospice of North Idaho	$50,000	441	$113.4	Jan 2, 2013

individuals. These breaches are now posted in a new, more accessible format that allows users to search and sort the posted breaches. Additionally, this new format includes brief summaries of the breach cases that OCR has investigated and closed, as well as the names of private practice providers who have reported breaches of unsecured protected health information to the Secretary. The following breaches have been reported to the Secretary:

Show Advanced Options

Breach Report Results

Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
Brooke Army Medical Center	TX	Healthcare Provider	1000	10/21/2009	Theft	Paper/Films
Mid America Kidney Stone Association, LLC	MO	Healthcare Provider	1000	10/28/2009	Theft	Network Server
Alaska Department of Health and Social Services	AK	Healthcare Provider	501	10/30/2009	Theft	Other, Other Portable Electronic Device
Health Services for Children with Special Needs, Inc.	DC	Health Plan	3800	11/17/2009	Loss	Laptop
L. Douglas Carlson, M.D.	CA	Healthcare Provider	5257	11/20/2009	Theft	Desktop Computer
David I. Cohen, MD	CA	Healthcare Provider	857	11/20/2009	Theft	Desktop Computer
Michele Del Vicario, MD	CA	Healthcare Provider	6145	11/20/2009	Theft	Desktop Computer
Joseph F. Lopez, MD	CA	Healthcare Provider	952	11/20/2009	Theft	Desktop Computer
Mark D. Lurie, MD	CA	Healthcare Provider	5166	11/20/2009	Theft	Desktop Computer

Case Study

- $1200: Benefit per record
- $300: Cost per violation
- Average Payoff Per Record

- $4: Access cost per record
- ~30,000 Census records

Case Study

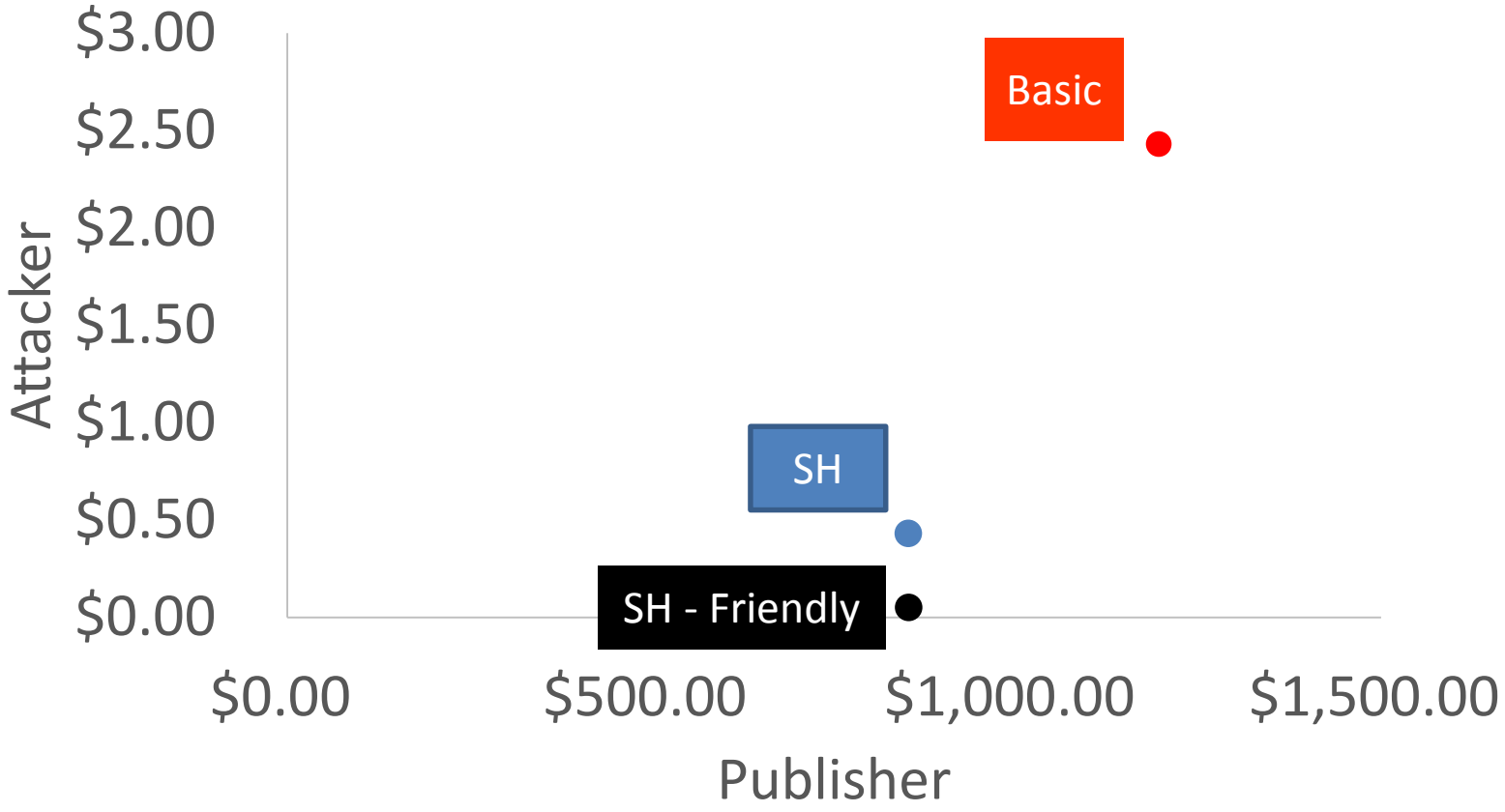
- $1200: Benefit per record
- $300: Cost per violation
- Average Payoff Per Record

- $4: Access cost per record
- ~30,000 Census records

Case Study

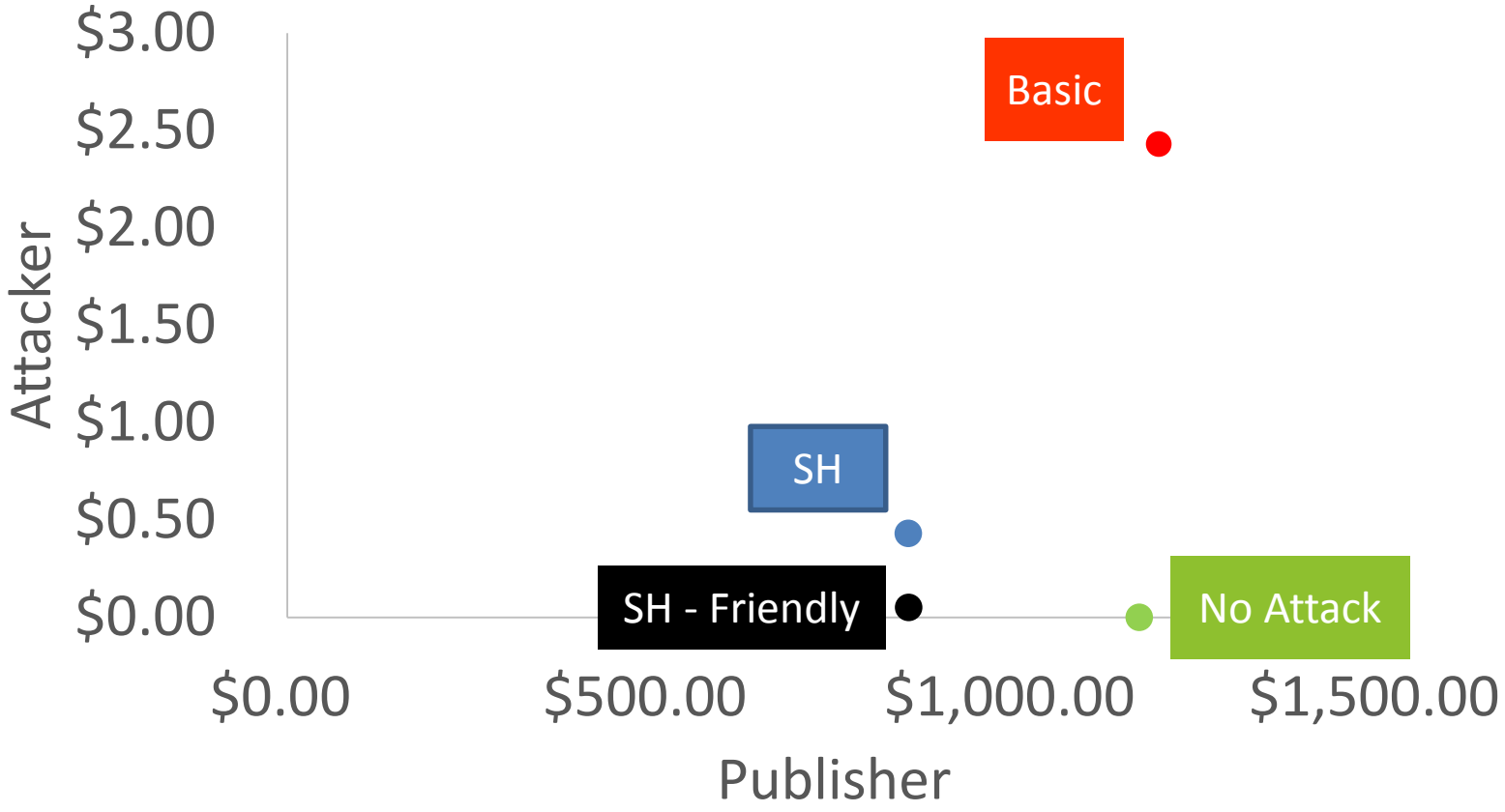
- $1200: Benefit per record
- $300: Cost per violation
- Average Payoff Per Record

- $4: Access cost per record
- ~30,000 Census records

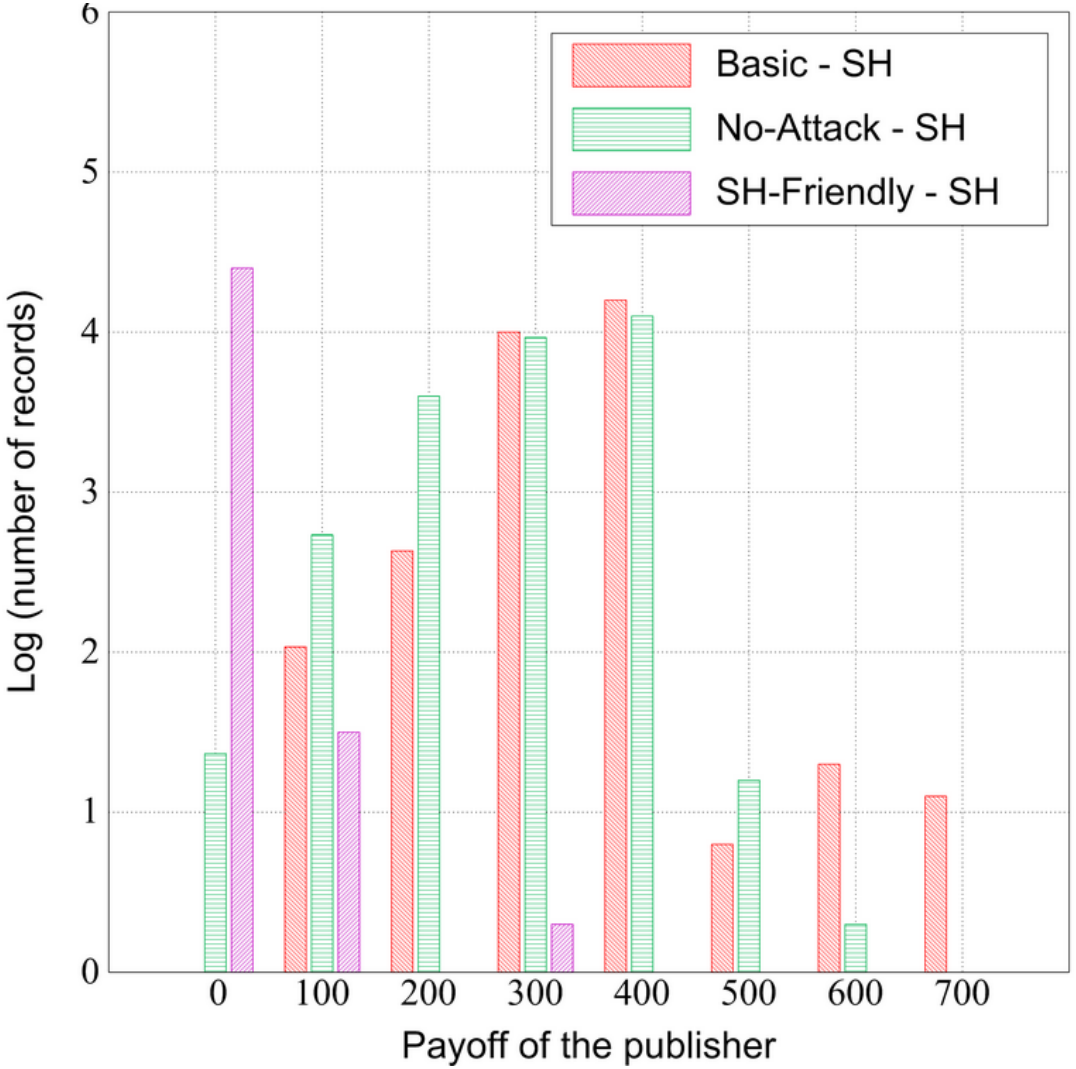


Case Study

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



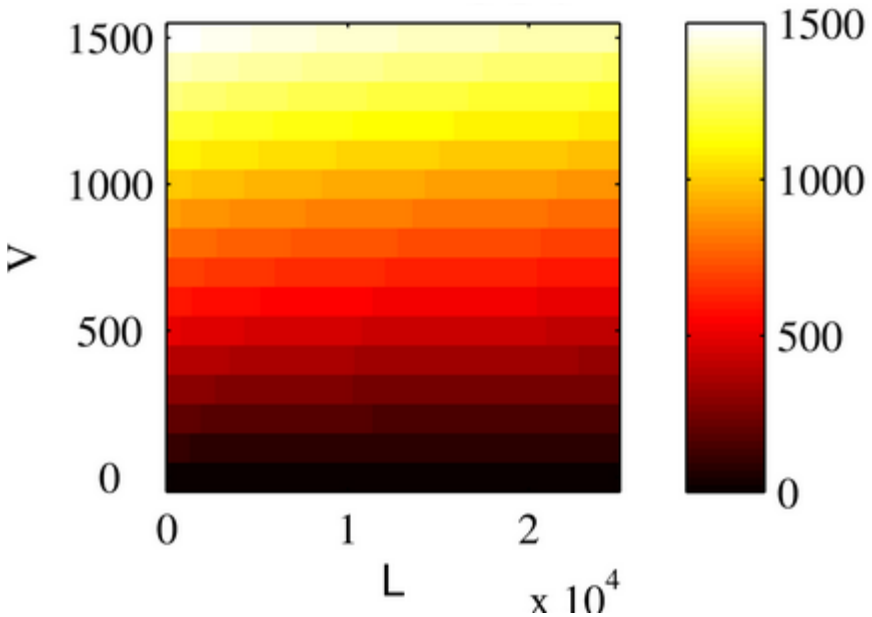
Case Study



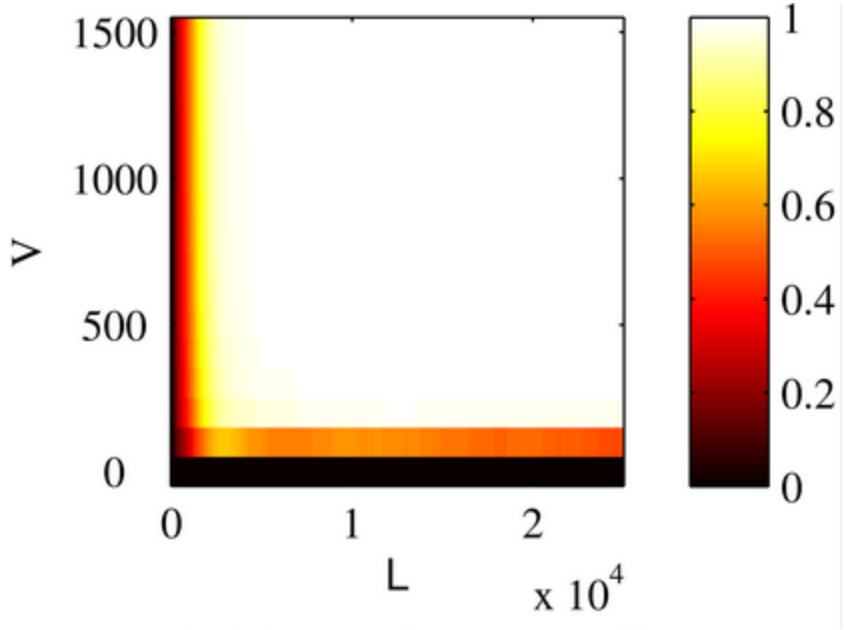
Wan et al, PLoS One. 2015

Sensitivity Analysis

Publisher Payout



Probability of Attack



V = Value of Record

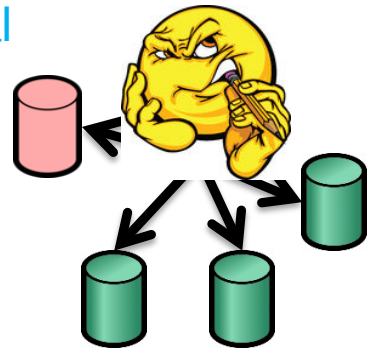
L = Loss due to re-identification

We Must Account for Process

Adversaries Make Sequential Decisions!

(Xia et al, CIKM 2015)

access the external dataset



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

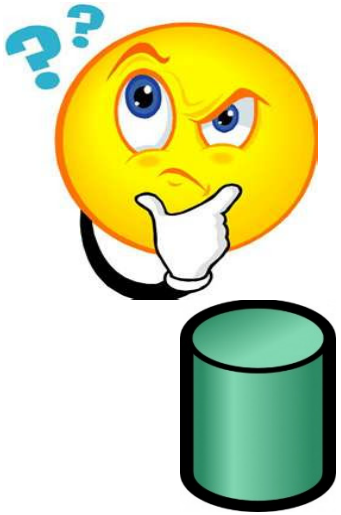
link the record to the external dataset

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

exploit an individual that matches the record

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

Outcome of Adversary's Actions Are Stochastic



- Before accessing the external dataset, the adversary may not know how many individuals can be linked to the record (unknown equivalence group size)

External dataset

	A_1	A_2	identity
p_1	?	?	?
p_2	?	?	?
p_3	?	?	?

De-identified dataset

	A_1	A_2	Sensitive information
r_1	22	F	hepatitis
r_2	56	M	tuberculosis

No match found

	r_1	r_2	No match
p_1	?	?	?
p_2	?	?	?
p_3	?	?	?

Adversary Makes a Series of Decisions to Complete an Attack

- The state of the attack can be represented by a set of variables
- Given state and action, attacker is granted a reward or pays a penalty
- Next state of attack depends on the current state & decision
  - Before accessing the external dataset, the attacker is uncertain about the content of the dataset
  - Before exploiting an individual, the attacker is not sure if the outcome will be success or fail

Factored Markov Decision Process (FMDP) to Represent Adversarial Behavior

- *Assumption:* Adversary always wants to maximize expected payoff
- MDP: models decision making in situations where outcomes are partly random and partly under the control of decision maker
- An optimal policy policy(x): action that maximizes the expected payoff across states
- Methods
  - Linear Programming, Value iteration, Policy iteration
- Challenges
  - State explosion of the FMDP
  - Dependency explosion of the Dynamic Bayesian Network

Experimental Setting

- Linking Variables: [*Age*, *ZIP-5*, *Gender*, *Race*]
- De-identified data: Adult Census + North Carolina ZIP-5
  - ~30K people
- Identified dataset: NC Voter's registration
  - ~6M people
- US census 2010 data from NC used to estimate the equivalence group size distribution

Costs & Gains

- $100 Cost to access identified dataset
- $10 Cost to conduct exploit
- $8000 Gain for successful adversary
- $10000 Penalty for adversary when exploit detected

(Sensitivity analysis in CIKM paper)

Comparison to Baseline Stackelberg

	Sequential	**Baseline***
Exploit (per de-id record)		1 individual
Decision Making		

*Wan, et al. PLoS One 2015.

Comparison to Baseline Stackelberg

	Sequential	**Baseline***
Exploit (per de-id record)	≥ 1 individual	1 individual
Decision Making		

*Wan, et al. PLoS One 2015.

Comparison to Baseline Stackelberg

	Sequential	**Baseline***
Exploit (per de-id record)	≥ 1 individual	1 individual
Decision Making		Single attack decision

$$payoff = G \times (\frac{prior_{r,D_e}}{G_{r,D_e}}) - p_{det} \times C_p - C_d - C_l - C_e$$

*Wan, et al. PLoS One 2015.

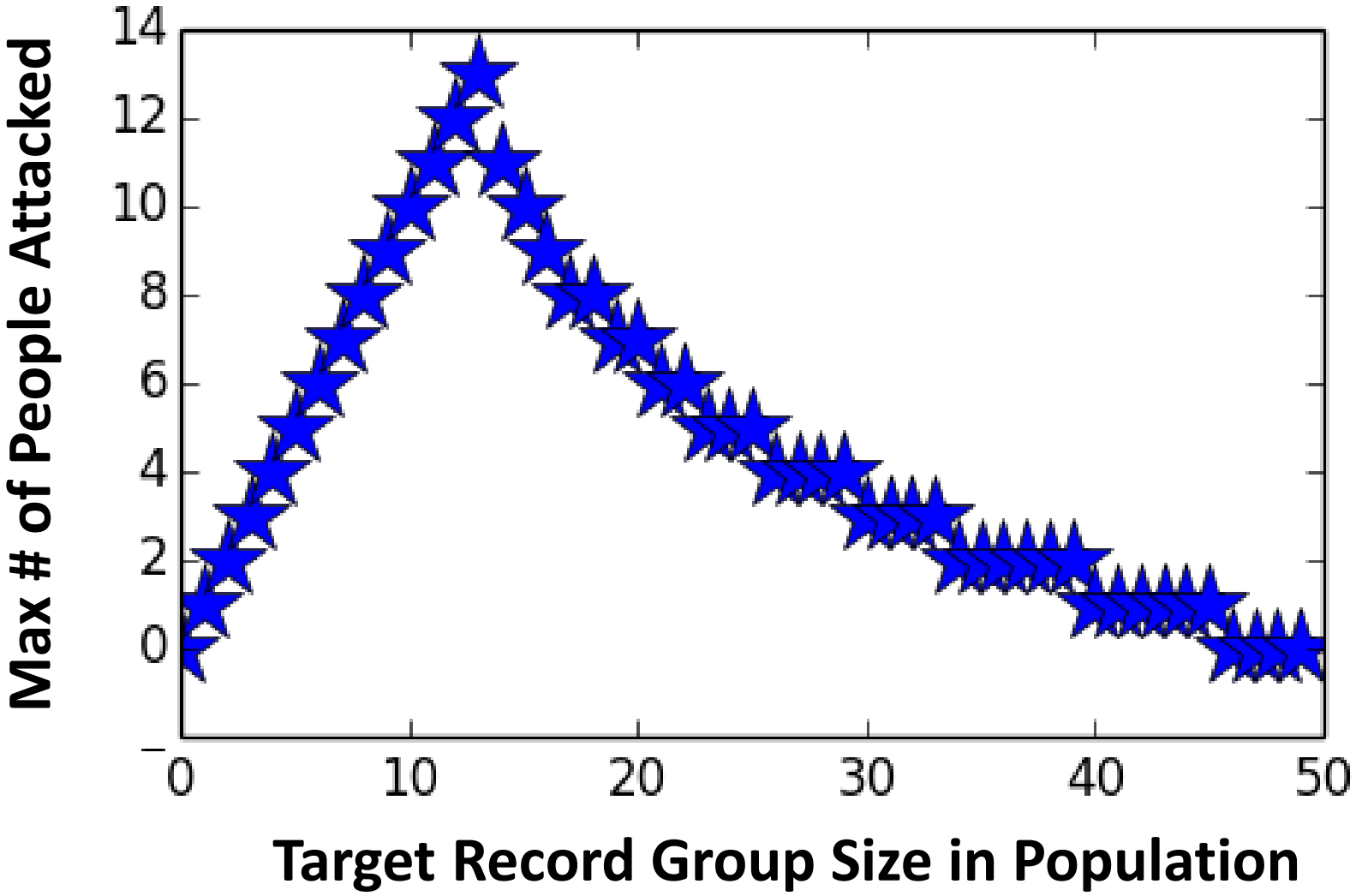
Comparison to Baseline Stackelberg

	Sequential	**Baseline***
Exploit (per de-id record)	≥ 1 individual	1 individual
Decision Making	Sequence of decisions	Single attack decision

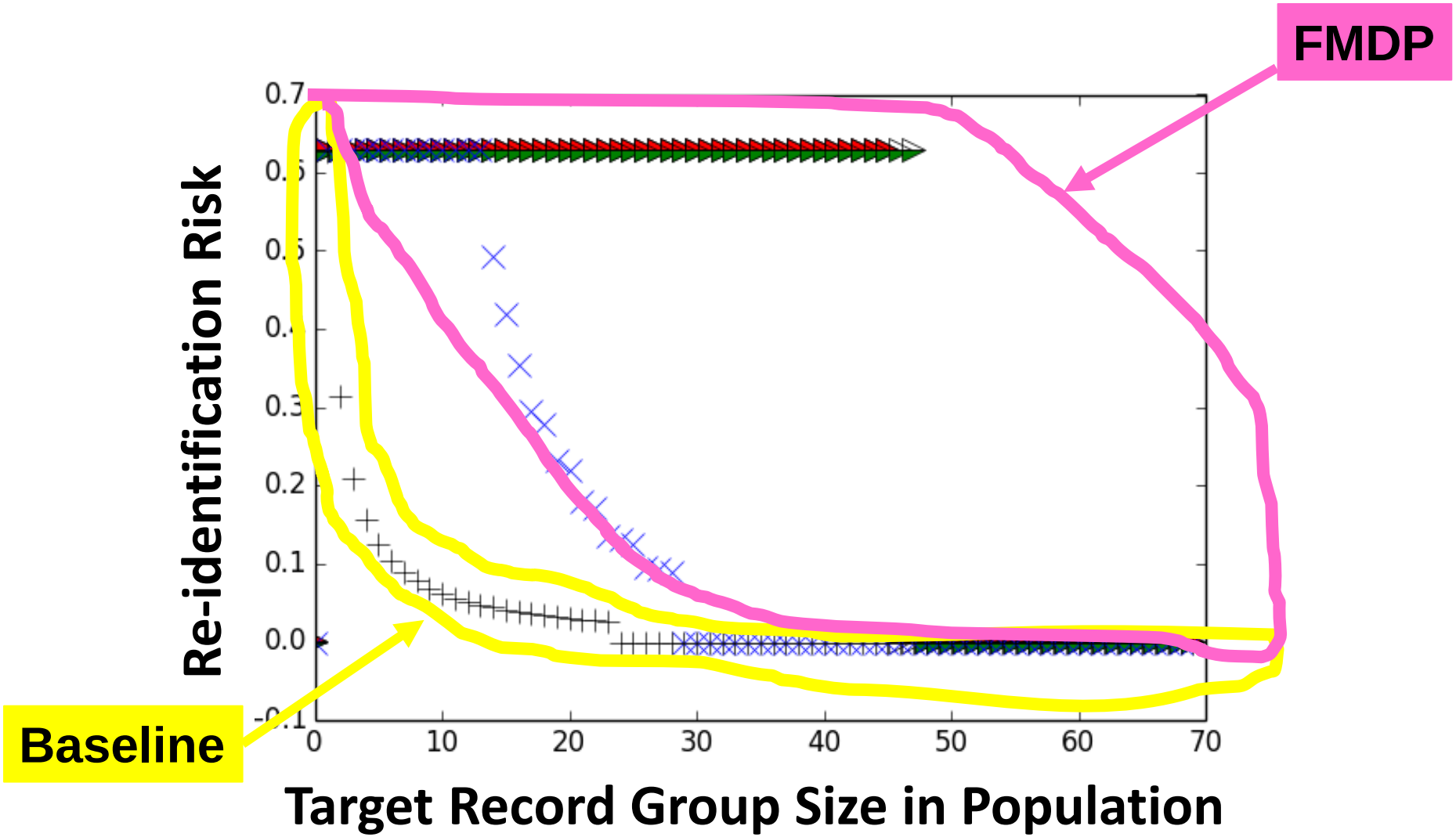
$$payoff = G \times (\frac{prior_{r,D_e}}{G_{r,D_e}}) - p_{det} \times C_p - C_d - C_l - C_e$$

*Wan, et al. PLoS One 2015.

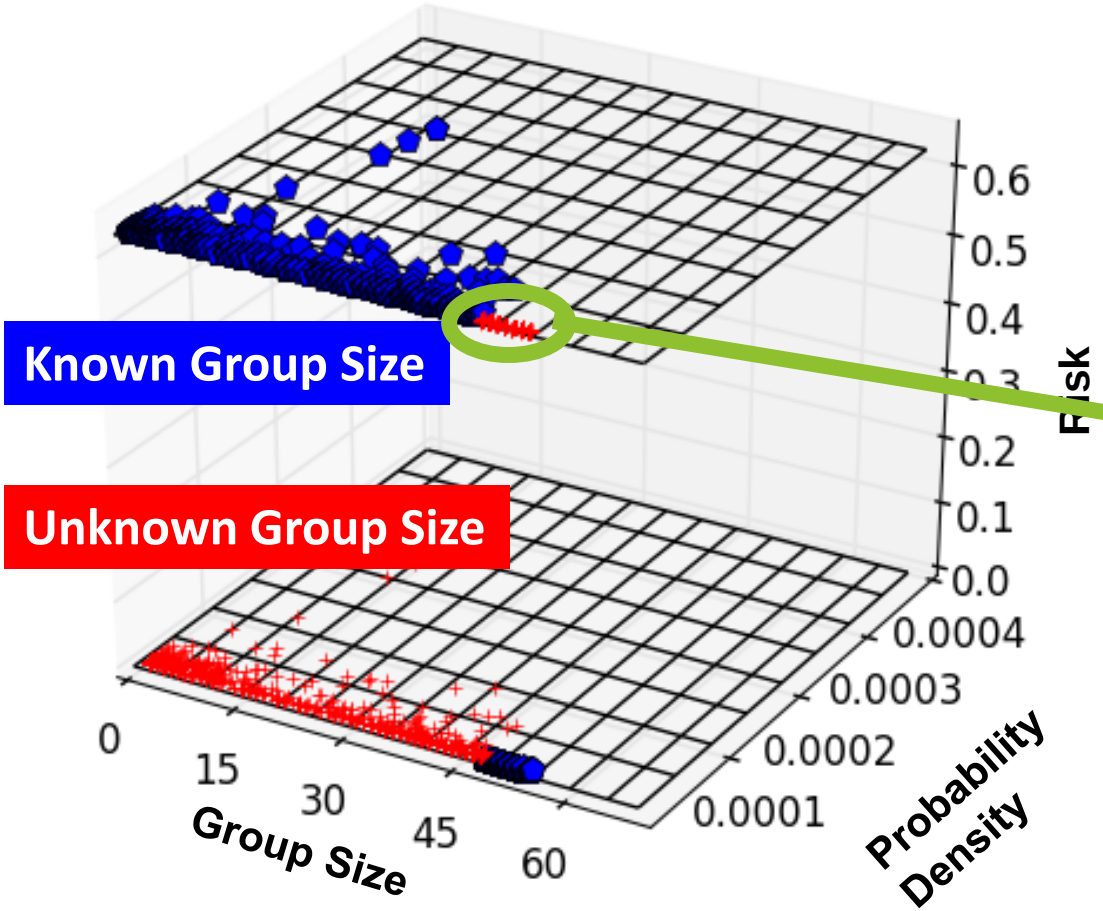
Adversary May Exploit > 1 Person!



Baseline Can Underestimate Risk!



Uncertainty in Equivalence Group Size can Lead to Higher Risk!



SUNK COST DILEMMA

So Where Are We Now?

- Privacy is NOT dead
- It is a much more complex landscape than has been suggested
- The space of options for data and social manipulation can be modeled… but it's huge!

- Challenges
  - Reliable cost estimates
  - Must Beware of "Baiting"!
  - Non-monotonic privacy and utility functions
  - State explosion in process models
  - Multiple publisher / adversary scenarios

Questions?

b.malin@vanderbilt.edu

Health Information Privacy Laboratory
http://www.hiplab.org/