Enabling Oncology Research Through De-identification

Bradley Malin, Ph.D.
Assoc. Prof. & Vice Chair of Biomedical Informatics, School of Medicine
Assoc. Prof. of Computer Science, School of Engineering
Director, Health Information Privacy Laboratory
Vanderbilt University
2/24/2014

Disclosure

- Paid consultancies

  Celgene (2013 – present)

  Sanofi-Aventis (2013 – present)

  * de-identification services for oncology clinical trials data

  Office for Civil Rights, U.S. Dept. of HHS (2009 – 2013)

  ♣ development of de-identification guidance

Given Enough Effort

Given Enough

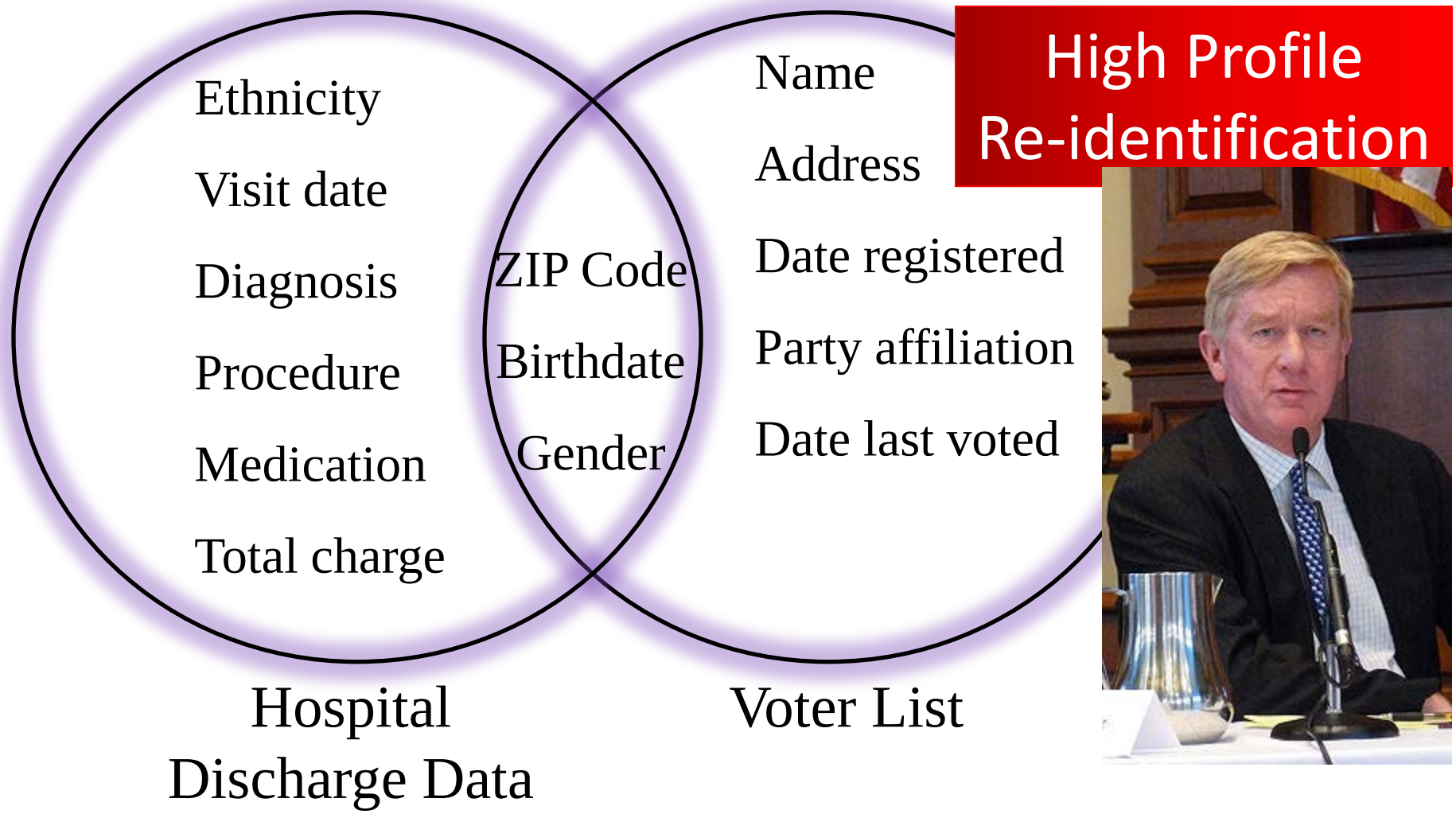
Effort,

Time,

Incentive,

Money…

Claim: De-identification Has Failed



High Profile Re-identification

Sweeney, JLME 1997

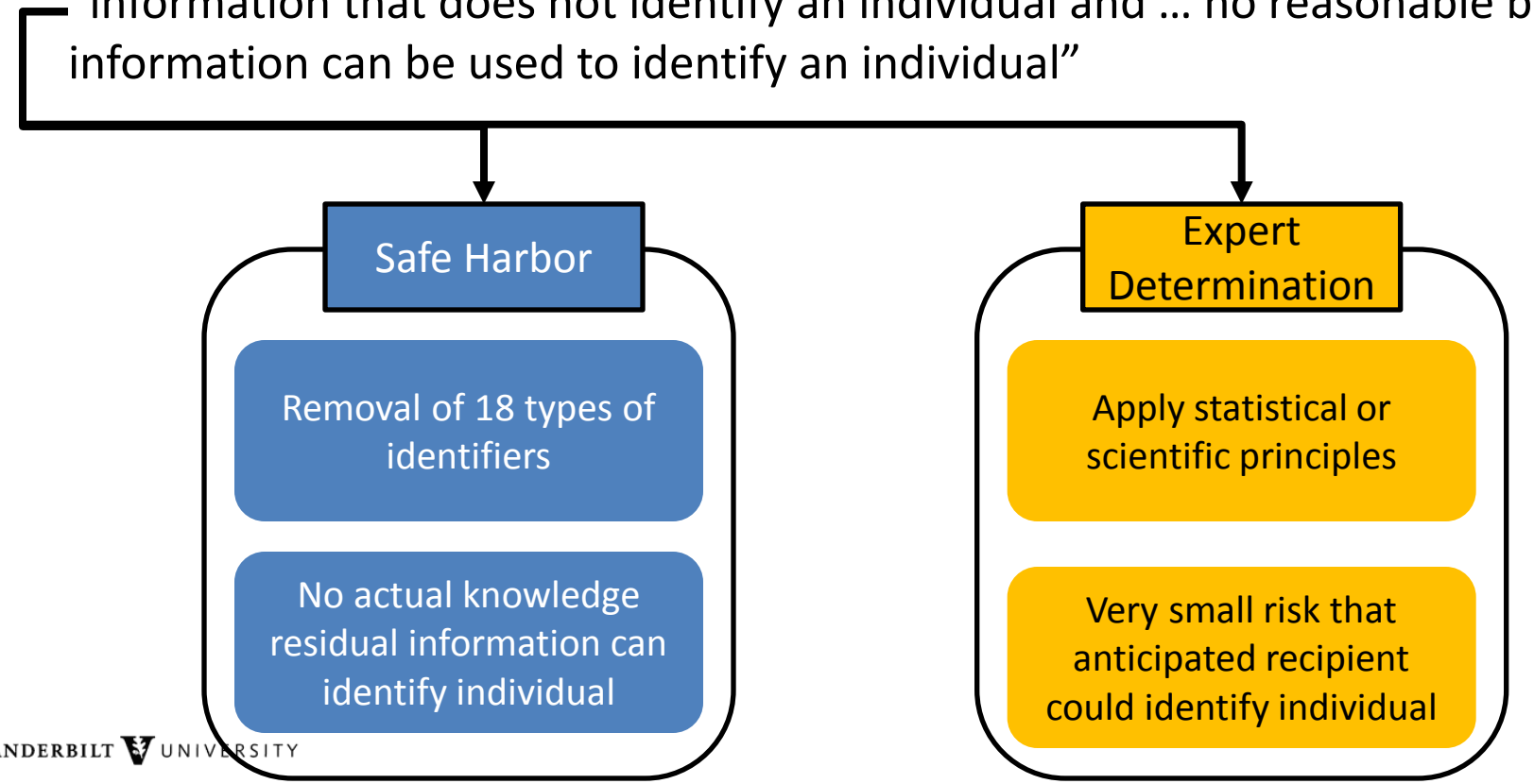
What is De-identification?

<u>According to EU</u> (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"

<u>According to HIPAA (Privacy Rule)</u>:

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



HIPAA "Cookbook" Standards

Field	Detail
Names	Related to patient (not provider)
Unique Numbers	Phone, SSN, MRN, …
Internet	Email, URL, IP addresses, ..
Biometrics	Finger, voice, …
Dates	Less specific than year Ages > 89
Geocodes	Town, County, Less specific than Zip-3 (assuming > 20,000 people in zone)
"Catch all"	"Any other unique identifying number, characteristic, or code"

Limited Dataset

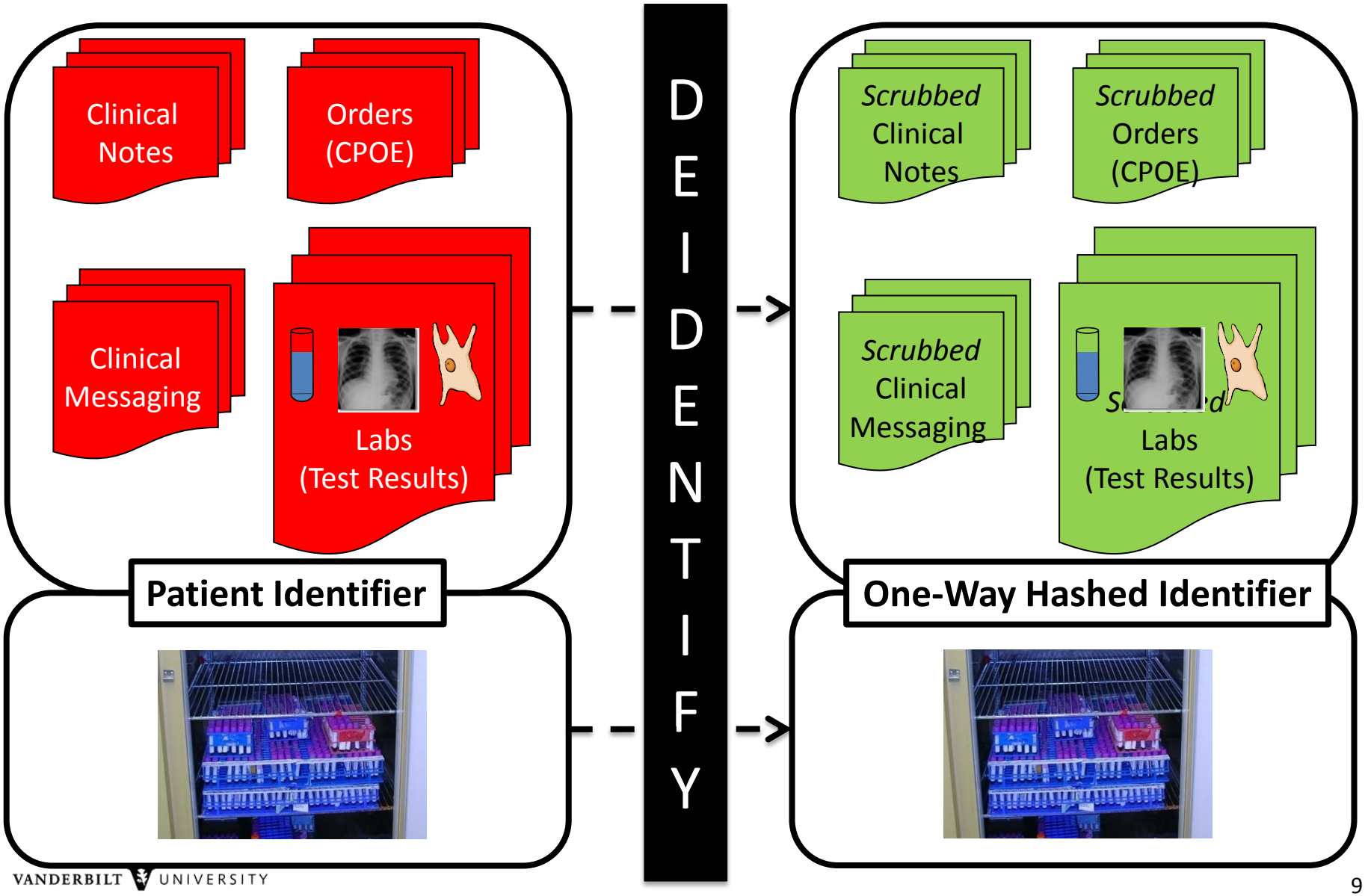
Safe Harbor

*** Must have no *actual knowledge* the remaining data can be used to identify

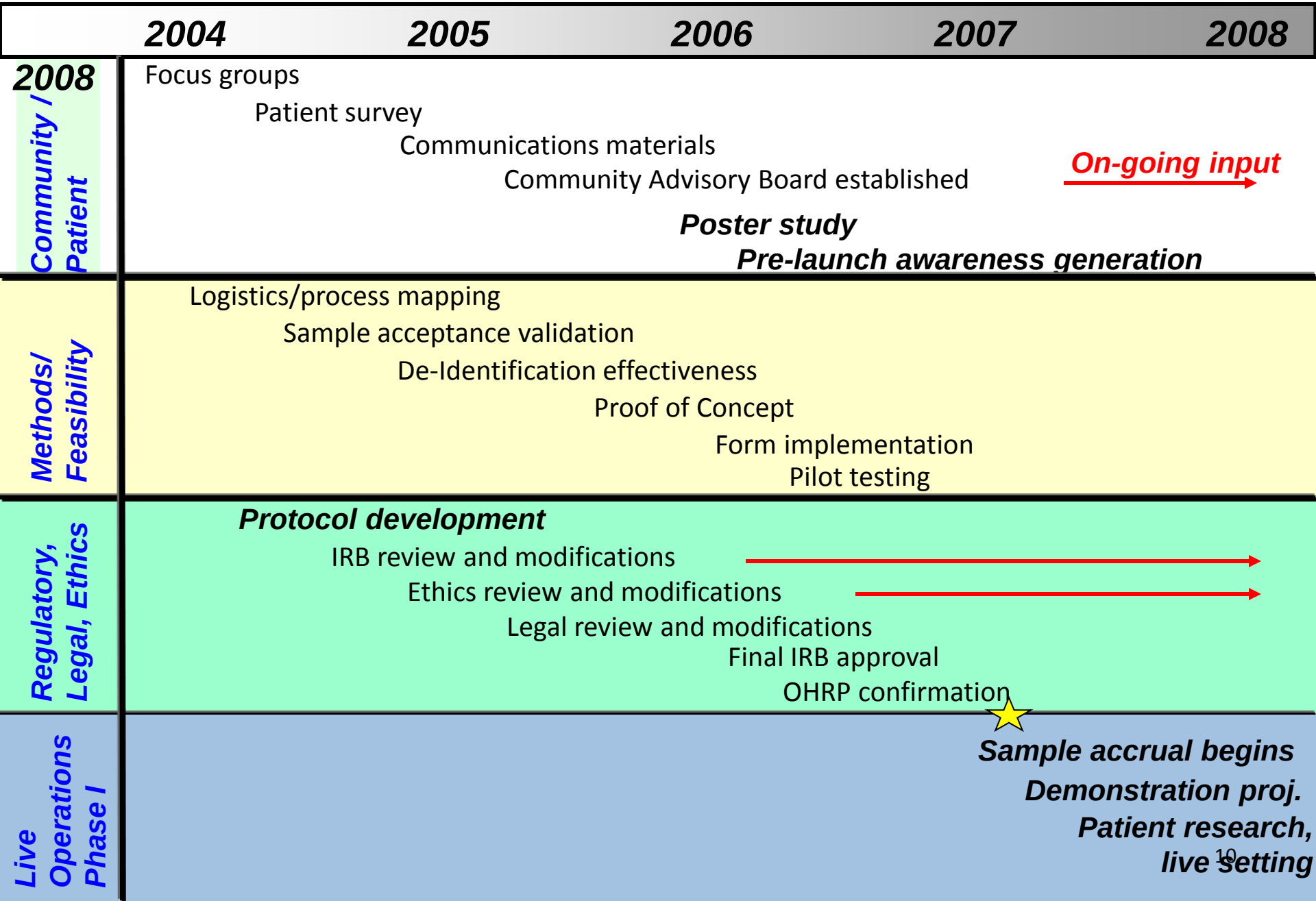
Practice What You Preach

Vanderbilt's BioVU

(~2 million patients records → over 100 TB of data)



Vanderbilt De-identified EMR + DNA

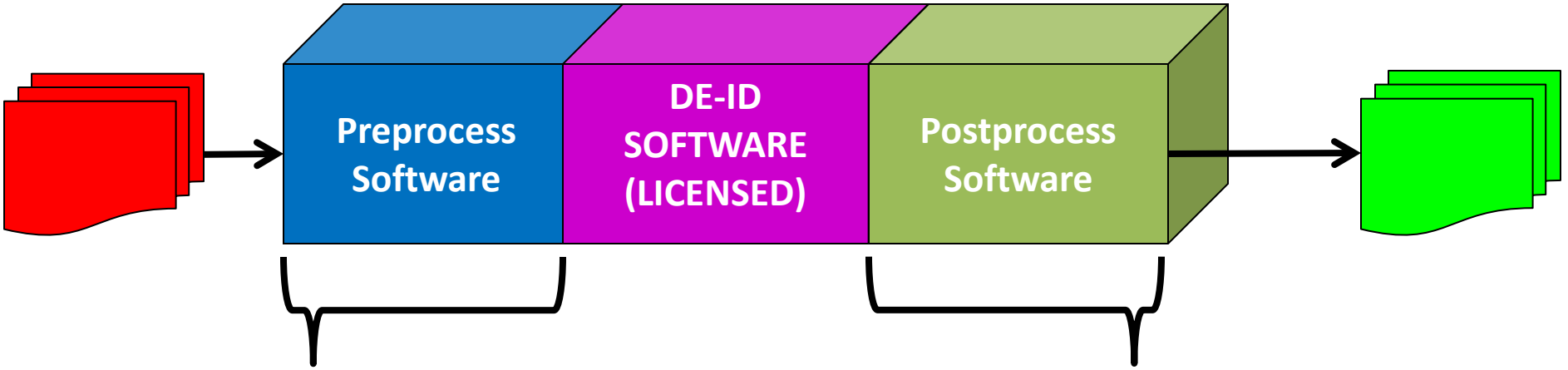


Redaction in Natural Language

Original PHI

Smith, **61** yo ...
daughter, **Lynn**, to ...
oncologist **Dr. White** ...
5/13/10 to consider ...
SWOG protocol **1811**, ...
was randomized **5/10** ...
to call **Mr. Smith** on ...
PLAN:**Dr White** and I ...

Scrubbing Process



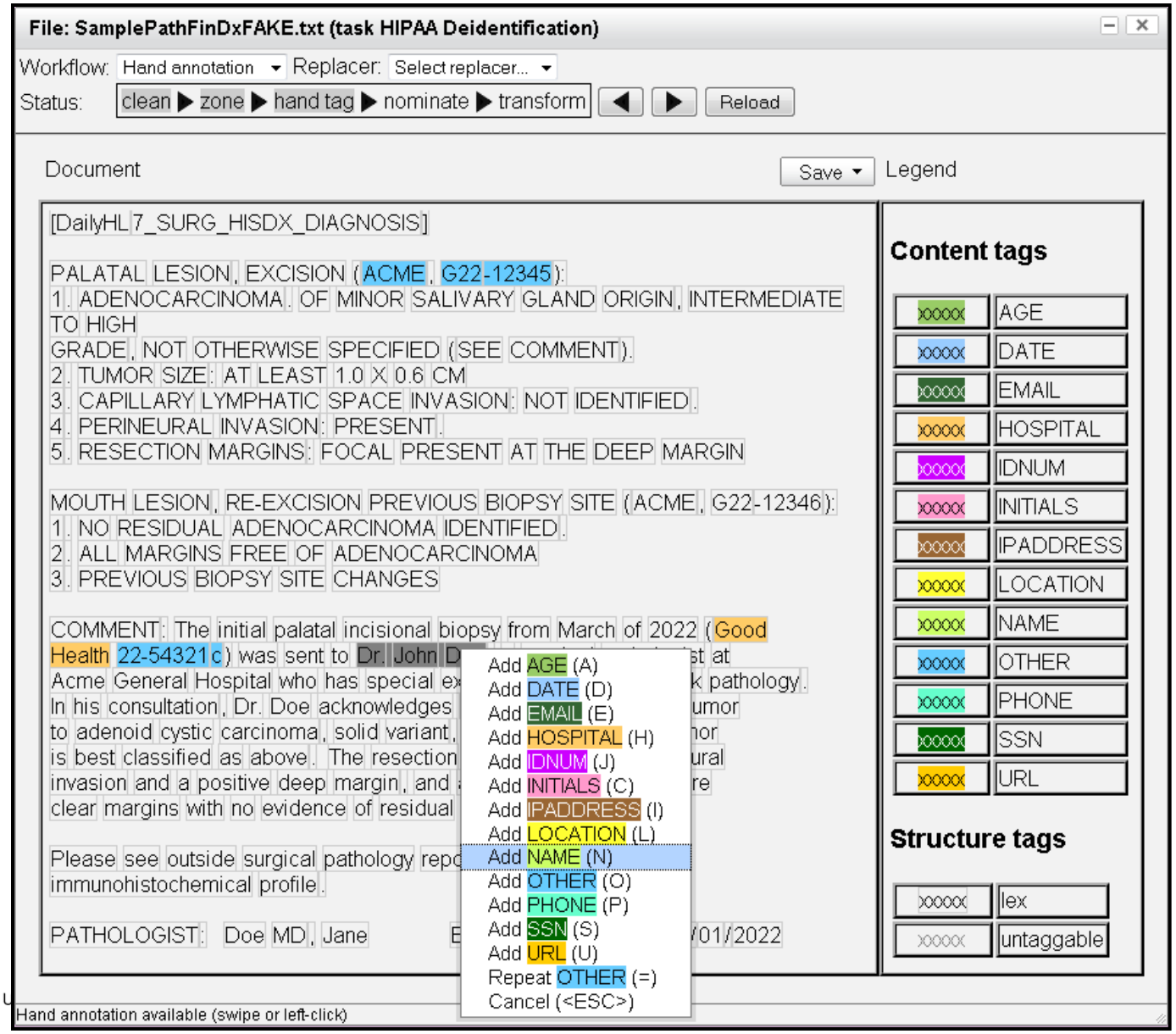
- Convert records to standard format
- Remove uninformative terms (e.g., “cc:”, “sincerely”)
- Add **PROTECTED[begin] & **PROTECTED[end] tags to retain necessary information

- Random Offset of **DATE
- Addition of hashed pseudonym

Recall = 0.999

Software - MIST (MITRE Identification Scrubbing Toolkit)

Aberdeen et al, IJMI 2010



Does Machine Learning Work?
(Vanderbilt EMR – *No Dictionaries*)

	Discharge	Laboratory	Letter	Order	All
Train	200	400	200	400	1200
Test	50	100	50	100	300
Precision	0.946	0.905	0.931	0.993	0.943
Recall	0.986	0.966	0.956	0.999	0.978

Precision: 0.91 – 0.99

Recall: 0.95 – 0.99

Aberdeen et al. IJMI, 2010

Negligible Impact on Medication Extraction

- Conditional Random Field

  (@ Cincinnati Children's Hospital)
- ~3500 clinical notes over 22 note types

	Original Notes	Scrubbed Notes
Precision	96.3	96.3 – 96.5
Recall	89.3	88.9 – 89.5
F-measure	92.6	92.5 – 92.7

Deleger et al. JAMIA., 2013

Redaction Has its Limits

Original PHI	**Redacted PHI & Leaked PHI
Smith, 61 yo ...	**pt_name<A>, **age<60s> yo ...
daughter, Lynn, to ...	daughter, Lynn, to ...
oncologist Dr. White ...	oncologist Dr. **MD_name<C> ...
5/13/10 to consider ...	**date<5/28/10> to consider ...
SWOG protocol 1811, ...	SWOG protocol **other_id, ...
was randomized 5/10 ...	was randomized 5/10 ...
to call Mr. Smith on ...	to call Mr. **pt_name<A> on ...
PLAN:Dr White and I ...	PLAN:Dr White and I ...

Redaction Has its Limits…
but it Isn't the Only Option

Original PHI

Smith, 61 yo ...
daughter, Lynn, to ...
oncologist Dr. White ...
5/13/10 to consider ...
SWOG protocol 1811, ...
was randomized 5/10 ...
to call Mr. Smith on ...
PLAN:Dr White and I ...

**Redacted PHI & Leaked PHI

**pt_name<A>, **age<60s> yo ...
daughter, Lynn, to ...
oncologist Dr. **MD_name<C> ...
**date<5/28/10> to consider ...
SWOG protocol **other_id, ...
was randomized 5/10 ...
to call Mr. **pt_name<A> on ...
PLAN:Dr White and I ...

Surrogate PHI & Hidden PHI

Jones, a 64 yo ...
daughter, Lynn, for ...
oncologist Dr. Howe ...
5/28/10 to consider ...
SWOG protocol 1798, ...
was randomized 5/10 ...
to call Mr. Jones on ...
PLAN:Dr White and I ...

Idea: Inject surrogated information to hide the leaks!

Carrell et al., JAMIA 2013

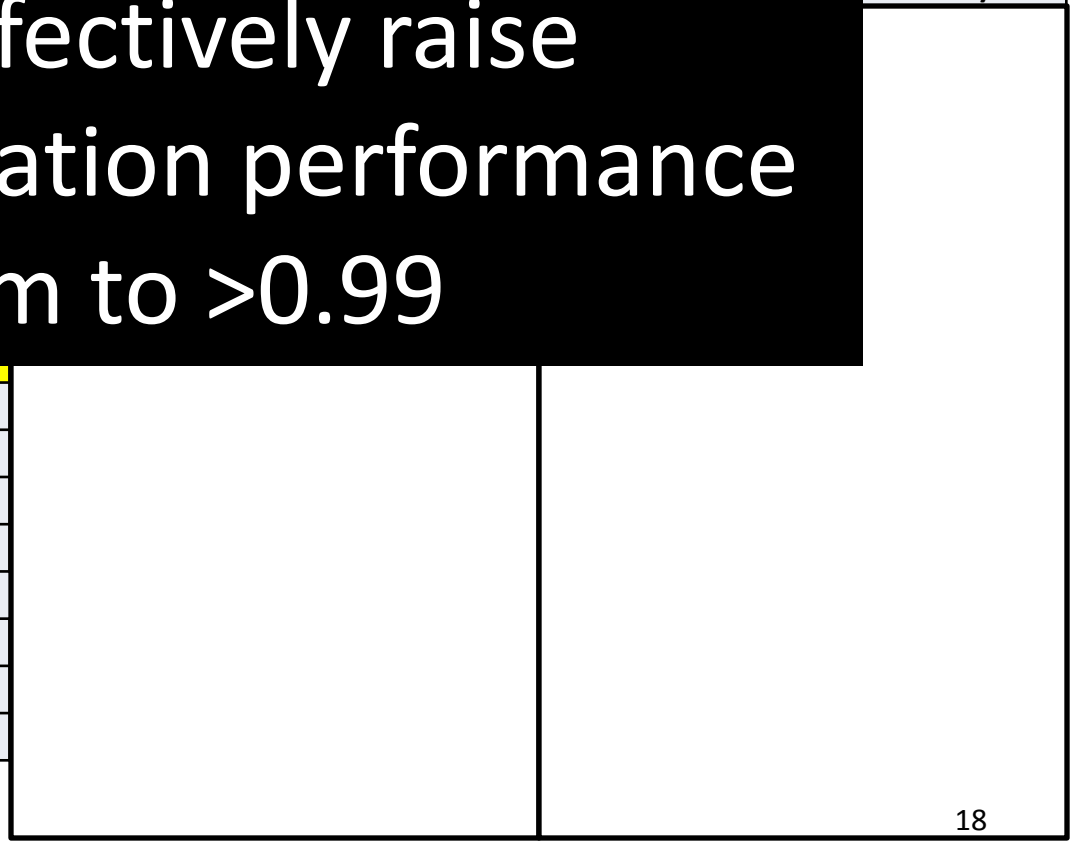
Hiding in Plain Sight [HIPS]

- Added a surrogation component to MIST*
- ~130 oncology notes from Group Health Coop of Puget Sound

**MIST forced into a dumbed-down state for assessment*

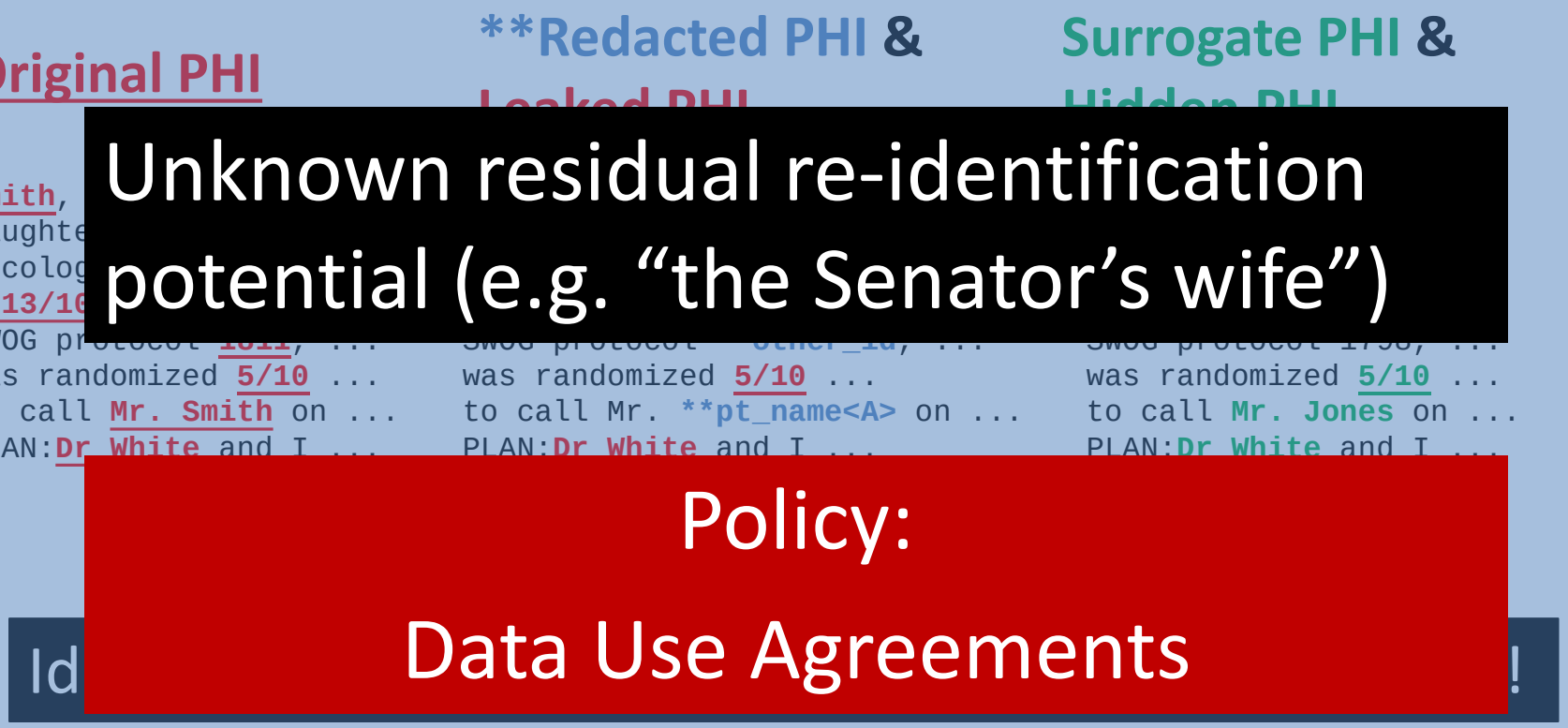
Identifier type			
HIPAA			
Pat. name			
Age			
Phone #			
Address			
Date			
MRN	3	3	1.00
Acct. #	1	1	1.00
Other ID #s	10	9	0.90
ALL	323	47	0.15
OTHER			
Prac name	82	9	0.11
Org. name	27	20	0.74
ALL	109	29	0.27

Can effectively raise de-identification performance from to >0.99



Carrell et al., JAMIA 2013

Even HIPS has Limits

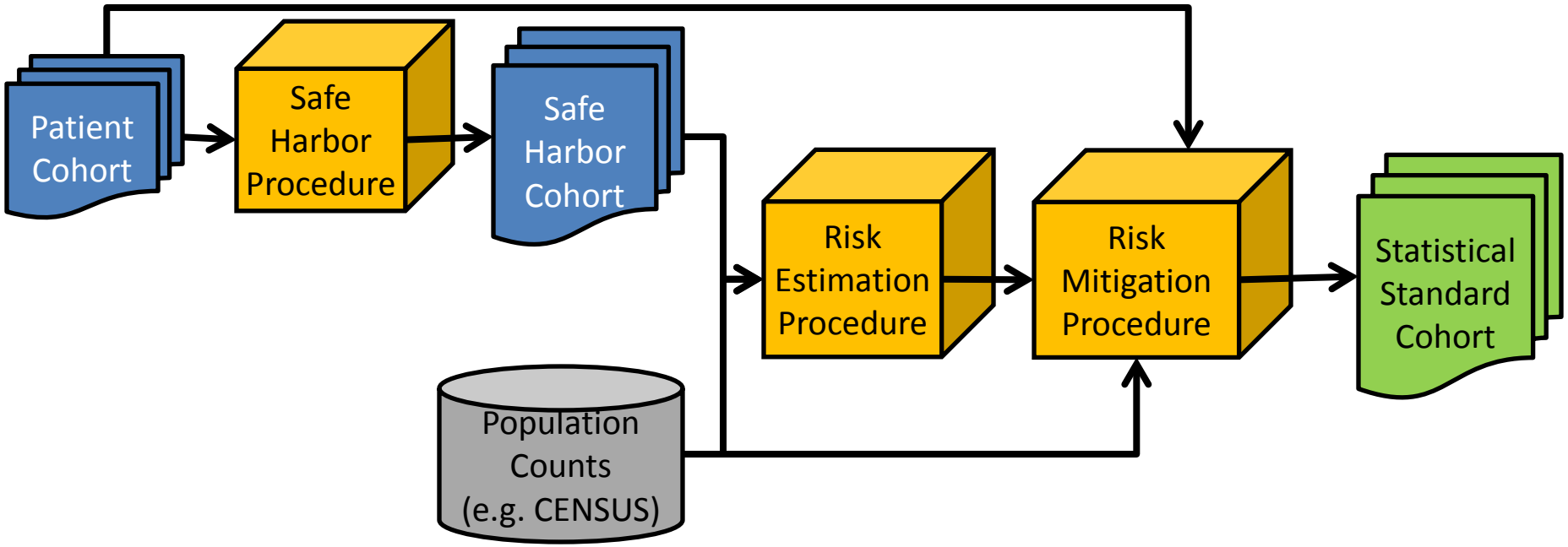


Carrell et al., JAMIA 2013

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."

Towards a Risk-Based De-identification Model



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

Vandy ECG Case Study

Who	State	State Population Size (2010 Census)	Cohort Size	Patients >89 years old
Vanderbilt	TN	~6 million	~3,000	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 in Multiple Populations

- 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.

Risk Model: Uniques

Is 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

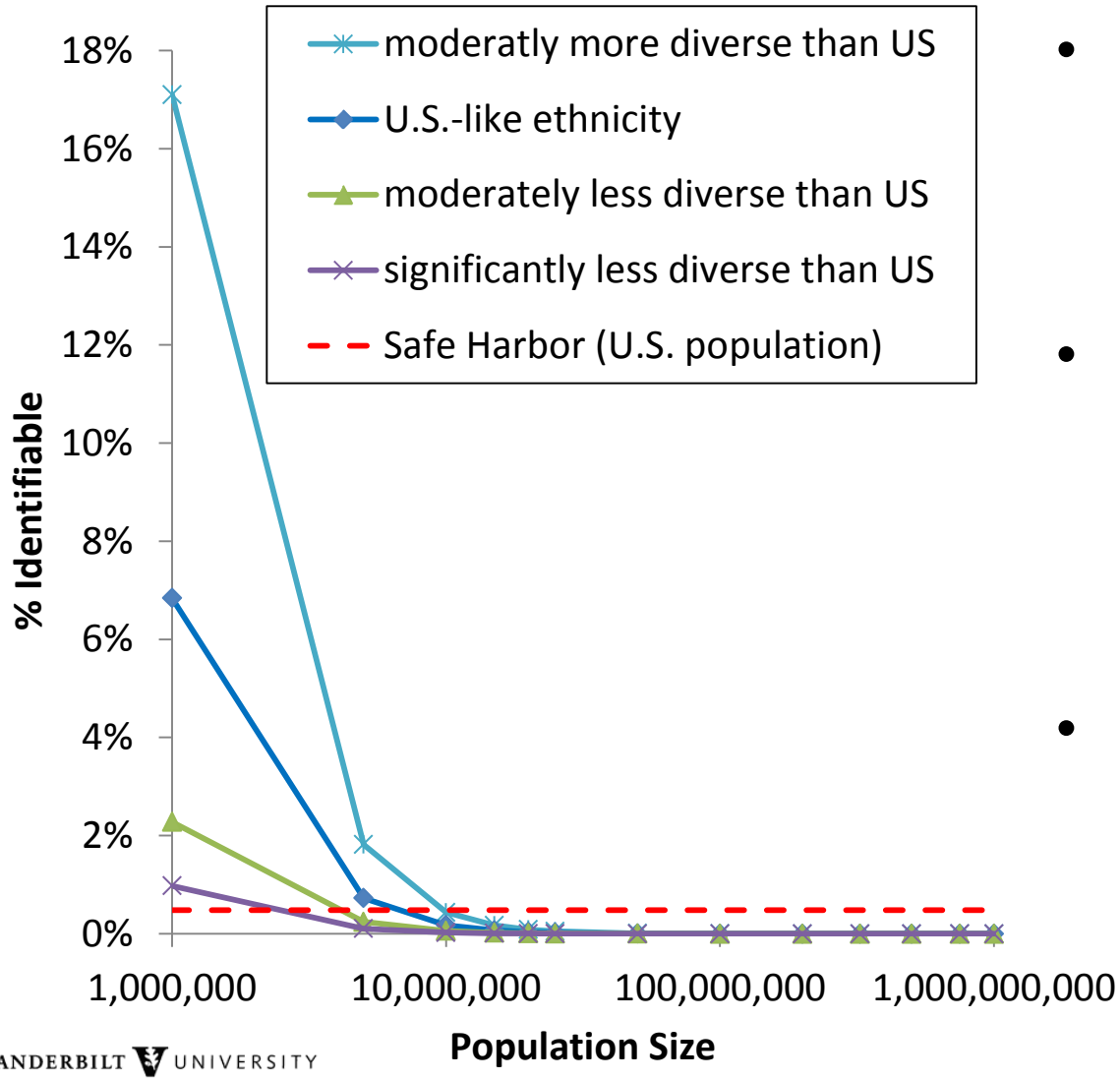
Green = risk no worse than Safe Harbor

Malin, Benitez, & Masys. JAMIA. 2011.

Forthcoming Data from Sanofi

- Oncology clinical trial data for Project Data Sphere
- De-identification Decisions
  - Only field-structured data (no free text)
  - Suppression of contact information (e.g., phone #, medical record #)
  - Coarsen geographic area:
    - North America, South America, Western Europe, Eastern Europe, & Other
  - Age reported at year, but top-coded as 85+
  - Dates of trial-related events permitted, but
  - Death events limited to one-week interval
- Proof of Protection
  - Use population and dataset-specific distributions to show re-identification risk is no worse than Safe Harbor
  - Safe Harbor: 0.00029% of U.S. population estimated to be unique
  - Sanofi: ~0.000001% " " " " "

Risk in a Multinational Setting



- Risk analysis initially performed using US population statistics
- Extrapolated analysis by simulating the diversity of various demographic distributions (e.g., age, race)
- **Decision: no region less than 10M people**

Prepping for Expert Determination

- 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*

- "Adversaries" have incomplete knowledge

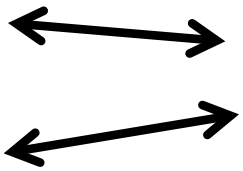
[Your Favorite Feature] Distinguishes You!!

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

- 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)

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*

- ~50% of Vanderbilt patients with at least 1 diagnosis code are unique!
- ~75% " " " " " " " 2

Loukides, Denny, and Malin. JAMIA. 2010.

Big Data

≠

End of Privacy

Simple Expert Model

k-Anonymity (Sweeney, 2002)

Ensure *k* record for every set of identifiers

"Guaranteed" Privacy

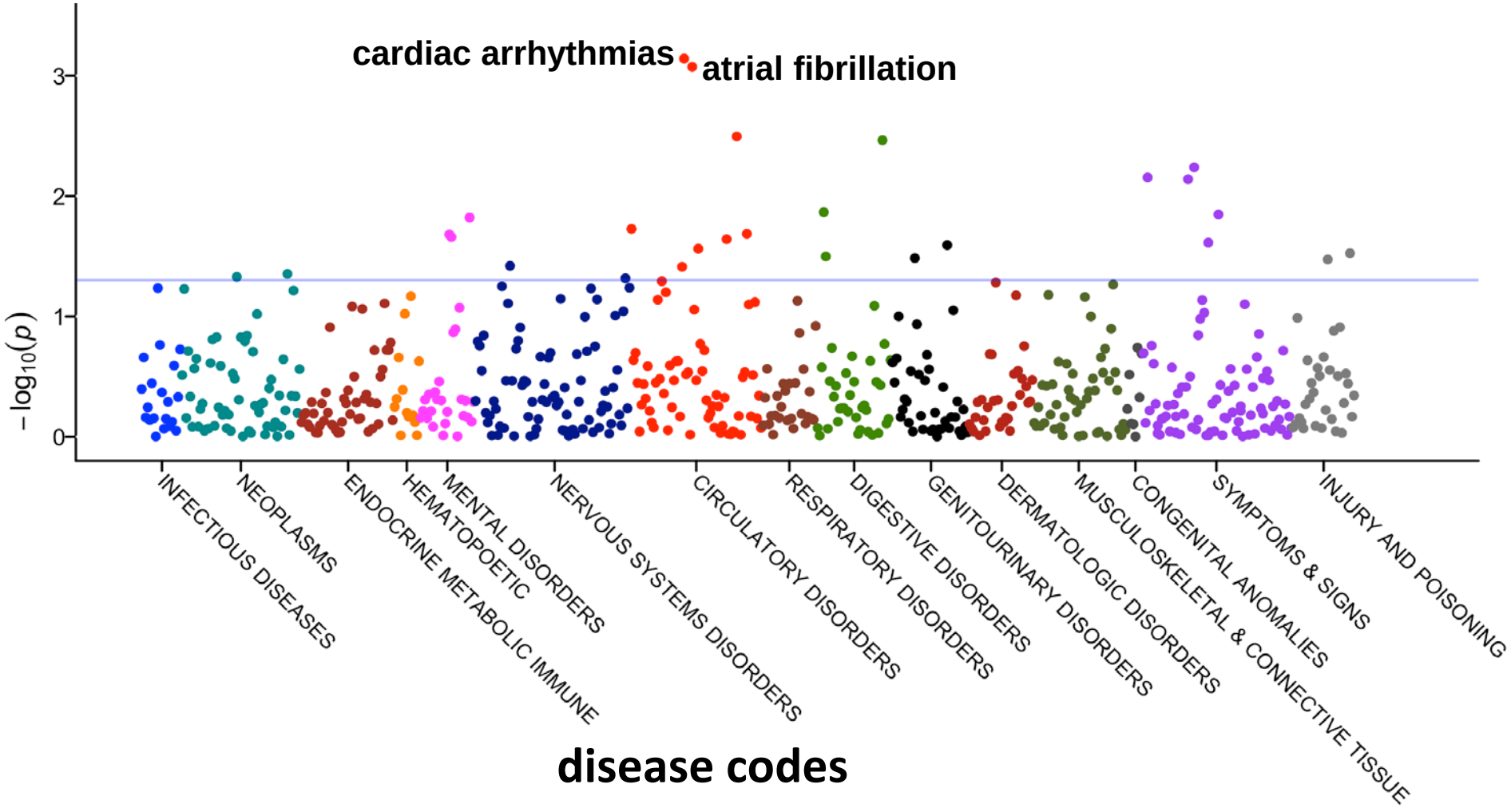
- Privacy: No record links to < k people using diagnoses
- Utility: Retain diagnoses codes for genome-phenome "validation"
- Cohort: 3000 Vanderbilt patients in a QRS study
- Results shown for k = 5

Phenotype	Intelligent
Asthma	✓
ADHD	
Bipolar	✓
Bladder cancer	
Breast cancer	✓
Coronary Disease	✓
Diabetes 1	✓
Diabetes 2	✓
Lung Cancer	✓
Pancreatic Cancer	✓
Platelet Related Phenotype	
Preterm Birth	✓
Prostate Cancer	✓
Psoriasis	✓
Renal Cancer	✓
Schizophenia	✓
Sickle-Cell Disease	✓

Loukides, Gkoulalas-Divanis, & Malin. PNAS. 2010.

Phenome Wide Association Studies

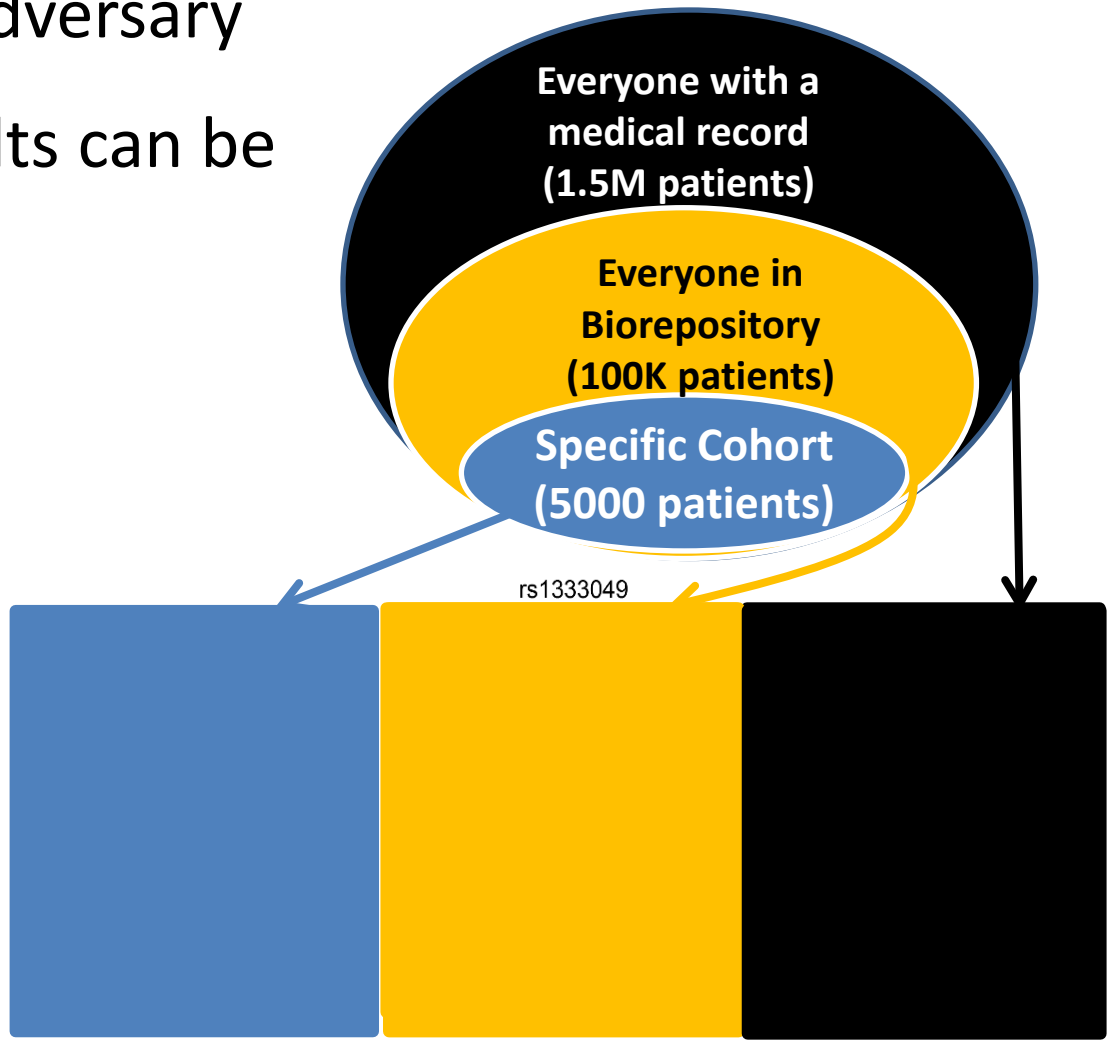
(associated with longer QRS duration in normal hearts)



Ritchie et al., Circulation 2013

Big Data Can Mean Big Privacy

- Often use very strong adversary
- But almost perfect results can be achieved…
- *… in real world*
- Validation of 192 SNP – phenotype associations



*Heatherly, et al. PLoS One. 2013

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)

De-identification Can Be Safe

- Reviewed all **actual** re-identification attempts and rates of success
- All attacks through 2010
  - 14 published re-identification attacks on any type of data
  - 11 were conducted by researchers as demo attacks
  - Only 2 datasets followed any standard
  - Only case with health data subject to Safe Harbor had a success likelihood of 0.00013

K. El Emam, E. Jonker, K. Arbuckle, and B. Malin. PLoS One. 2011.

Challenges for De-identification

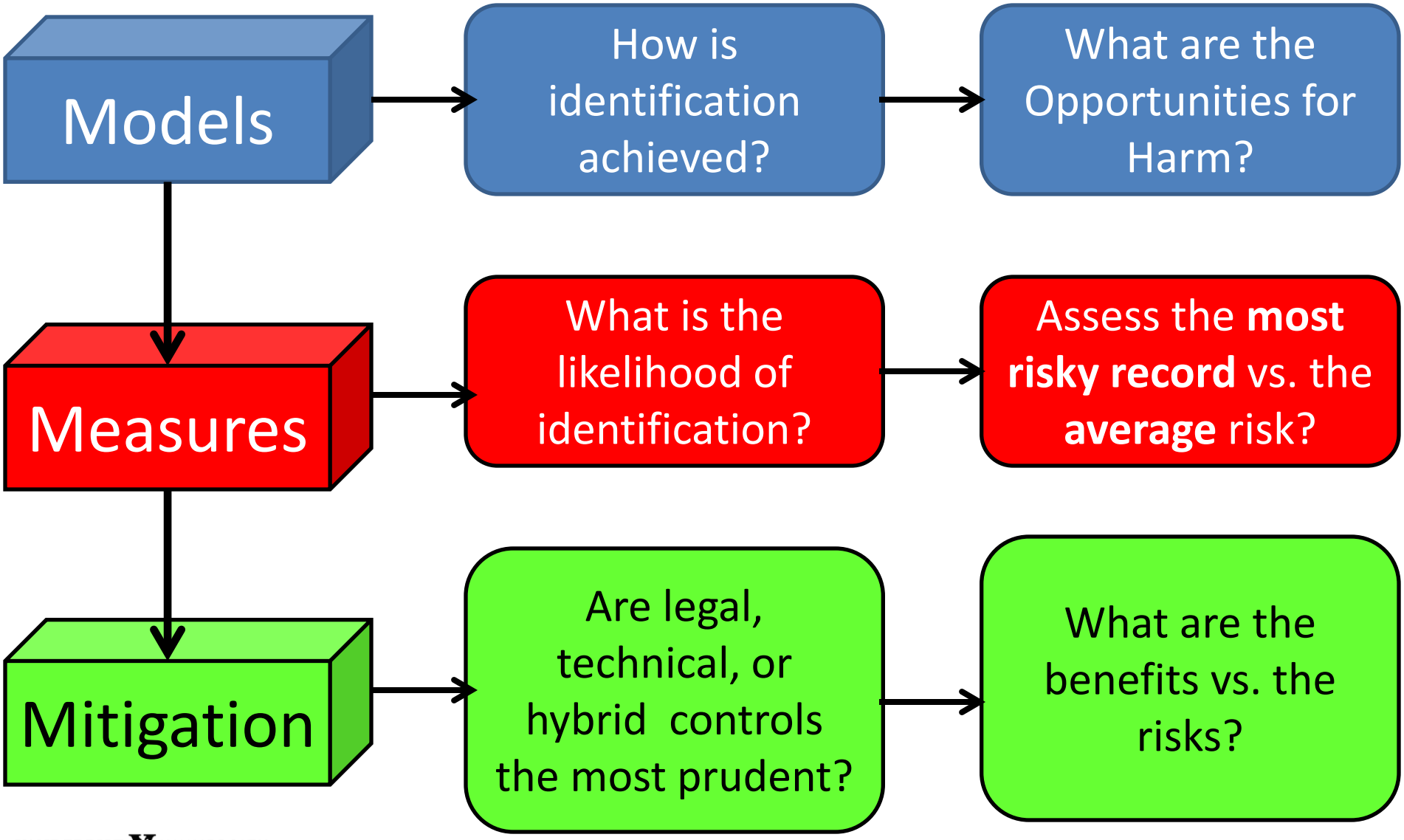
- 2014 recent report from NRC Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences
- HIPAA calls for protection from identity disclosure… but does not address utility of the data
- No definitive standard for
  - Risk Assessment
  - De-identification Methodology (but the Office for Civil Rights issued HIPAA guidance in November 2012)
- Need for national clearinghouse of models, methods, and evaluations
- Protections should be proportional to harm, recipients, and generally the **contex**t
- Case studies are needed!

NRC Recommendations

- Data Protection Plans
  - Degree of identifiability
  - Computing environment where data is shared
  - Location & method of data storage
  - Controls to the data
  - Secure transmission of data
  - Methods of output (paper vs. electronic)
  - Mechanisms for audit and oversight

- Researchers should honor confidentiality agreements, but no further consent should be necessary for secondary use (including linkage to other resources, unless specified from the outset)

We Must be Reasonable & Practical



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.
- Raymond Healtherly, Ph.D.
- Murat Kantarcioglu, Ph.D. (U. of Texas)
- Jiuyong Li, Ph.D. (U. So. Australia)
- Grigorios Loukides, Ph.D. (Cardiff U)
- Steve Nyemba
- Dan Masys, M.D. (U. Washington)
- Muqun Li
- Dan Roden, M.D.
- Laura Rodriguez, Ph.D. (NHGRI / NIH)
- Latanya Sweeney, Ph.D. (Harvard U.)
- Acar Tamersoy (Georgia Tech)
- Wei Xie
- Weiyi Xia
- Wen Zhang
- Zhiyu Wan

- NRC Committee on revisions to the Common Rule

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)

Questions?

b.malin@vanderbilt.edu

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