

# Beyond the Science: Infrastructure, Policy, & Cross-cutting Needs

Bradley Malin, Ph.D.

Associate Prof of Biomedical Informatics, School of Medicine

Associate Professor of Computer Science, School of Engineering

Affiliated Faculty, Center for Biomedical Ethics & Society

Vanderbilt University

Transparent

**TRUSTED**

Timely

Technologies that Mitigate  
Risk for Patients and Facilitate  
Research Workflow

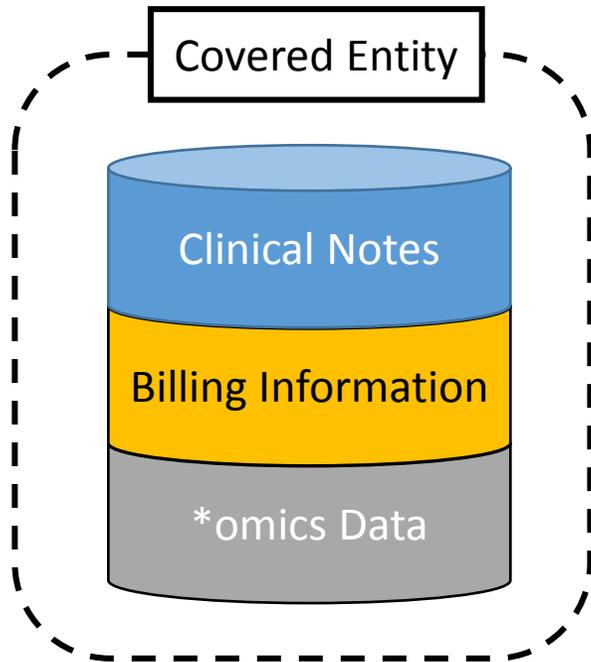
Acceptable Use Policies with  
Accountability & Legal Bite

# WHAT DO WE NEED?

Leverage 3rd Party Big Data  
Managers with Limited  
“Trust”

Engage Patients & Enhance  
Transparency in Research

# Building a Big Database for Clinical Research



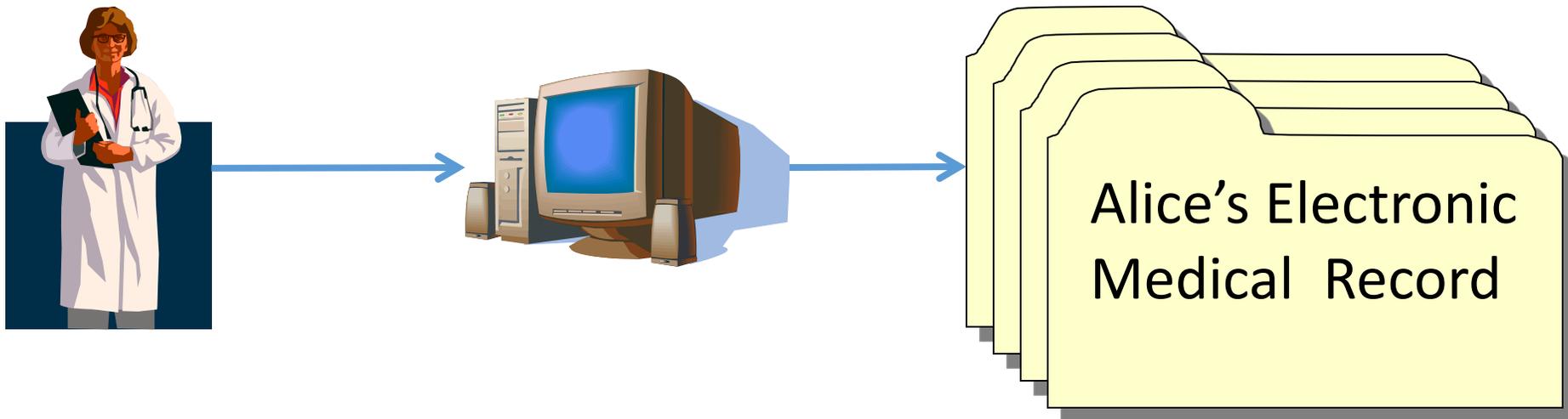
- We will learn to model health phenomena, not artifacts of healthcare operations (phenomics)
- We will tame heterogeneity of data (\*omics)
  - Transform into common language
- We will handle the magnitude of the data (data structures and trusted computing)

# More “Clinical” Data Than You Think

- Healthcare is inefficient in a variety of ways
  - Diagnosis of disease
  - Personalization of treatment
  - Management of health problems
  - *Healthcare operations → Don't forget EMR Audit Logs!*
- Growing information on which healthcare employees work with whom and how for which patients

# January 1, 2013

## Logged over 1,000,000 users' interactions



January 2, 2013

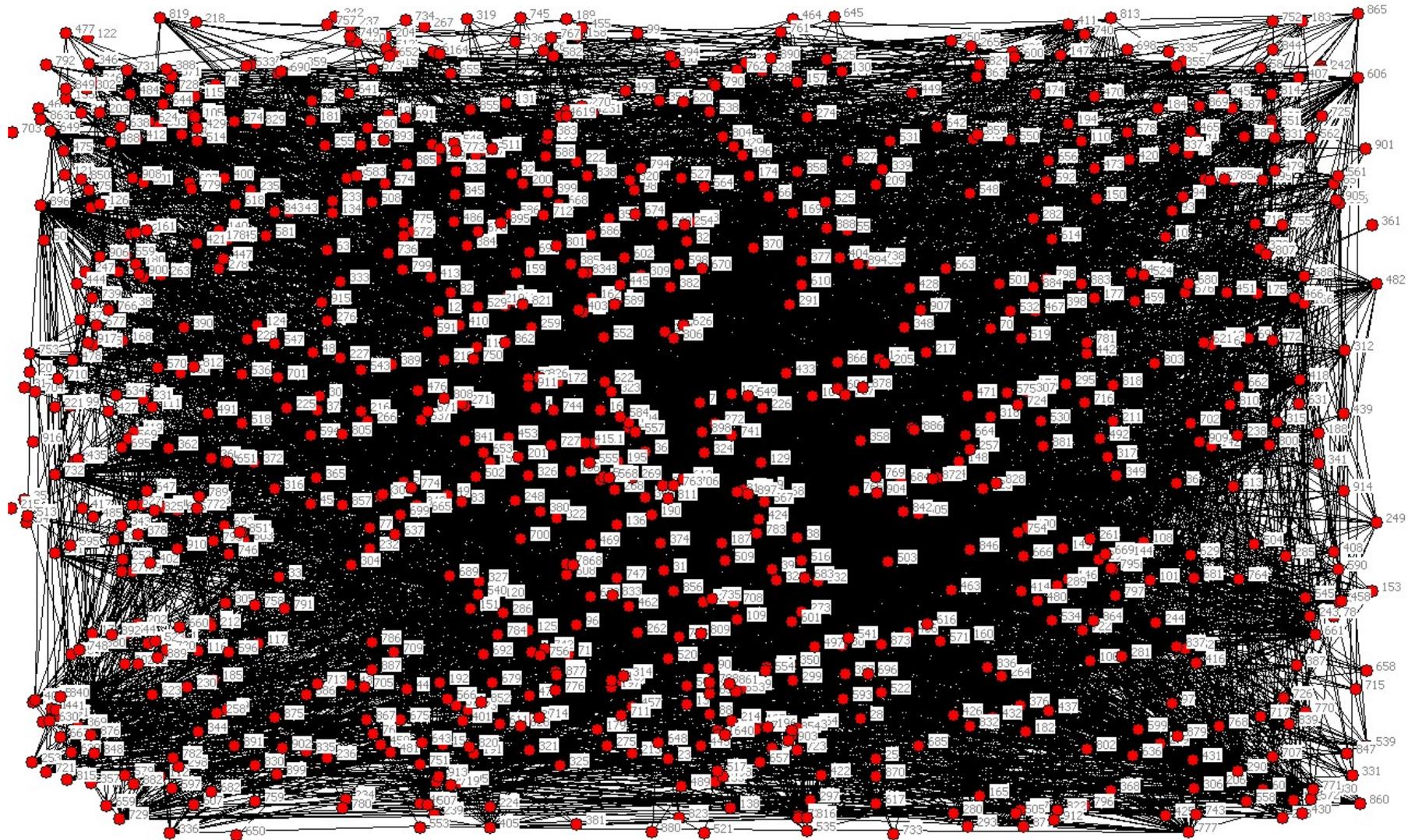
Logged over  
1,000,000 users'  
interactions

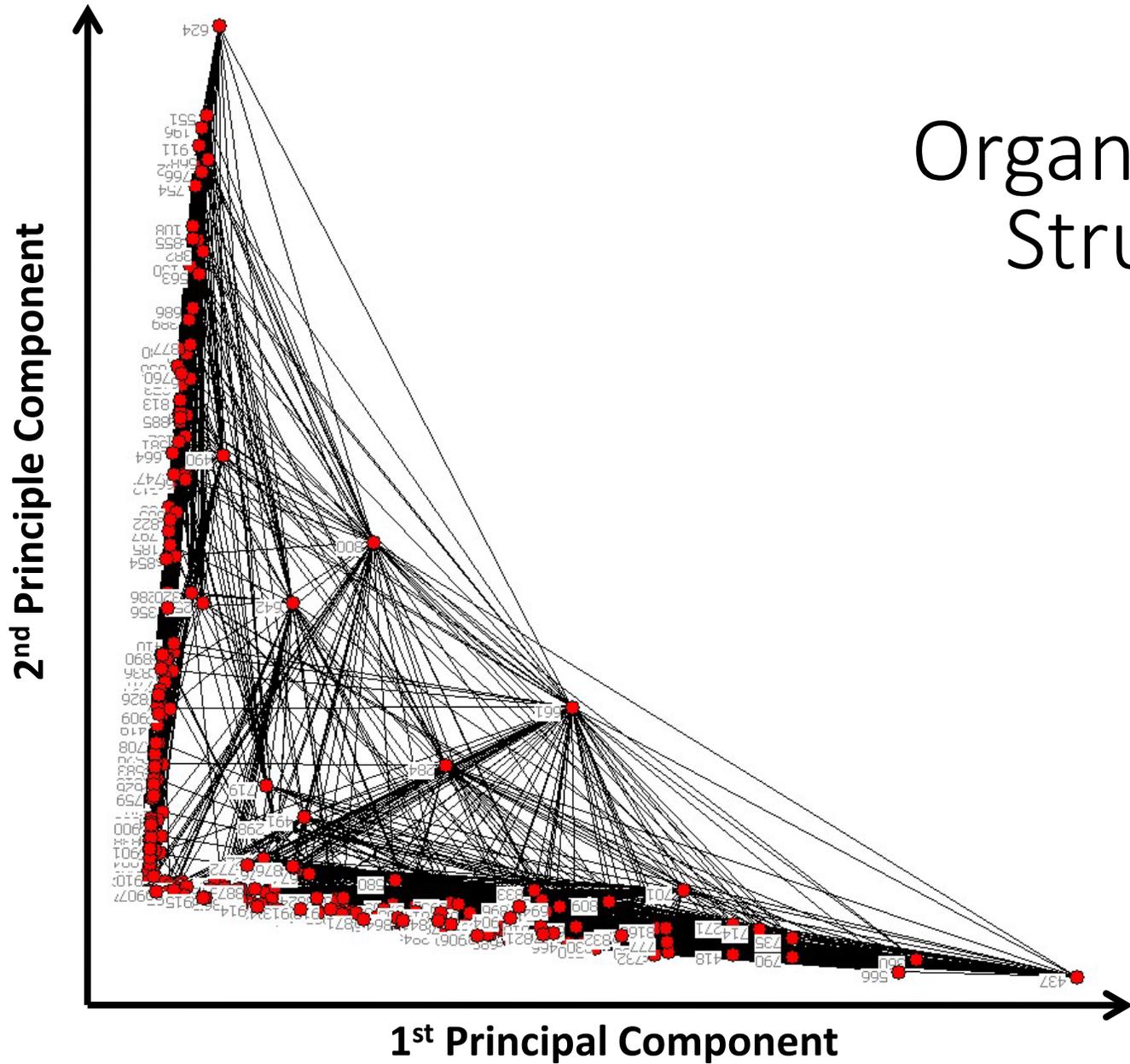
January 3, 2013

Logged over  
1,000,000 users'  
interactions

Jan 1

- EMR users linked if they accessed at least 1 common record





Organizational  
Structure

# HIPAA

## (One of) the Elephants in the Room

“Secondary use” of clinical data is possible, but it varies...

Identified  
Patient Data

- Waiver of consent: data is “on the shelf”
- Consent is impracticable 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

# HIPAA

## (One of) the Elephants in the Room

“Secondary use” of clinical data is possible, but it varies...

Identified  
Patient Data

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

Limited Data  
Set

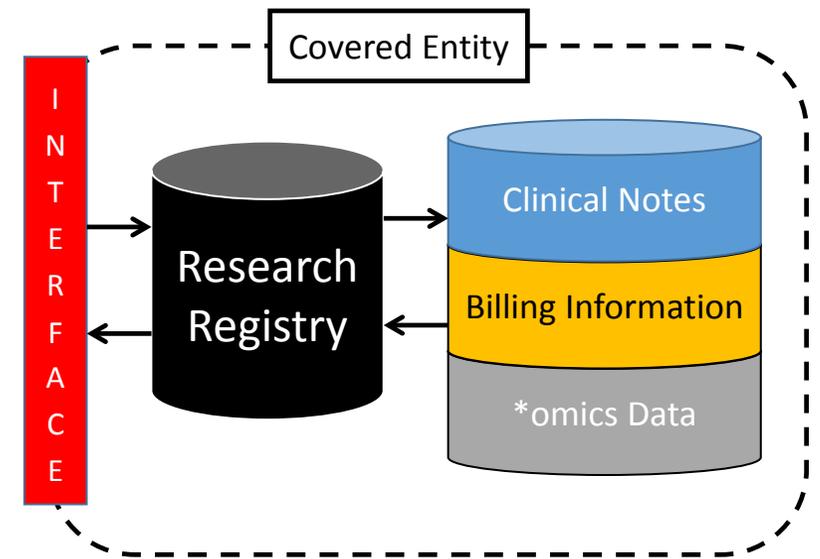
- Removal of 16 designated attributes
- Recipient signs data use contract

De-identified  
Data

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

# Transparency

- Patients may **not** want to know how data about them is used all the time... but it doesn't hurt to ask.
- They **do** want the ability to audit how data about has (or is) used
- *Should all research protocols be documented, indexed and searchable by... patient? ... even if de-identified?*
- Who would manage such a resource? The site which collected the data? HHS? OHRP?



# Consent (a couple of words)

- Is it “impracticable” to obtain consent from 1 million patients?
- Consent before a visit... we need scalable consent management information systems
  - How fine-grained (i.e., “specific”) should consent be?
  - Information altruism (thank Altman & Kohane for the suggestion!)
- Consent not received... should we go back to patients?
  - Historically difficult to keep tabs on patients, but modern networking technologies are changing the game



# HIPAA

## (One of) the Elephants in the Room

“Secondary use” of clinical data is possible, but it varies...

Identified  
Patient Data

- Waiver of consent: data is “on the shelf”
- Consent is impracticable 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

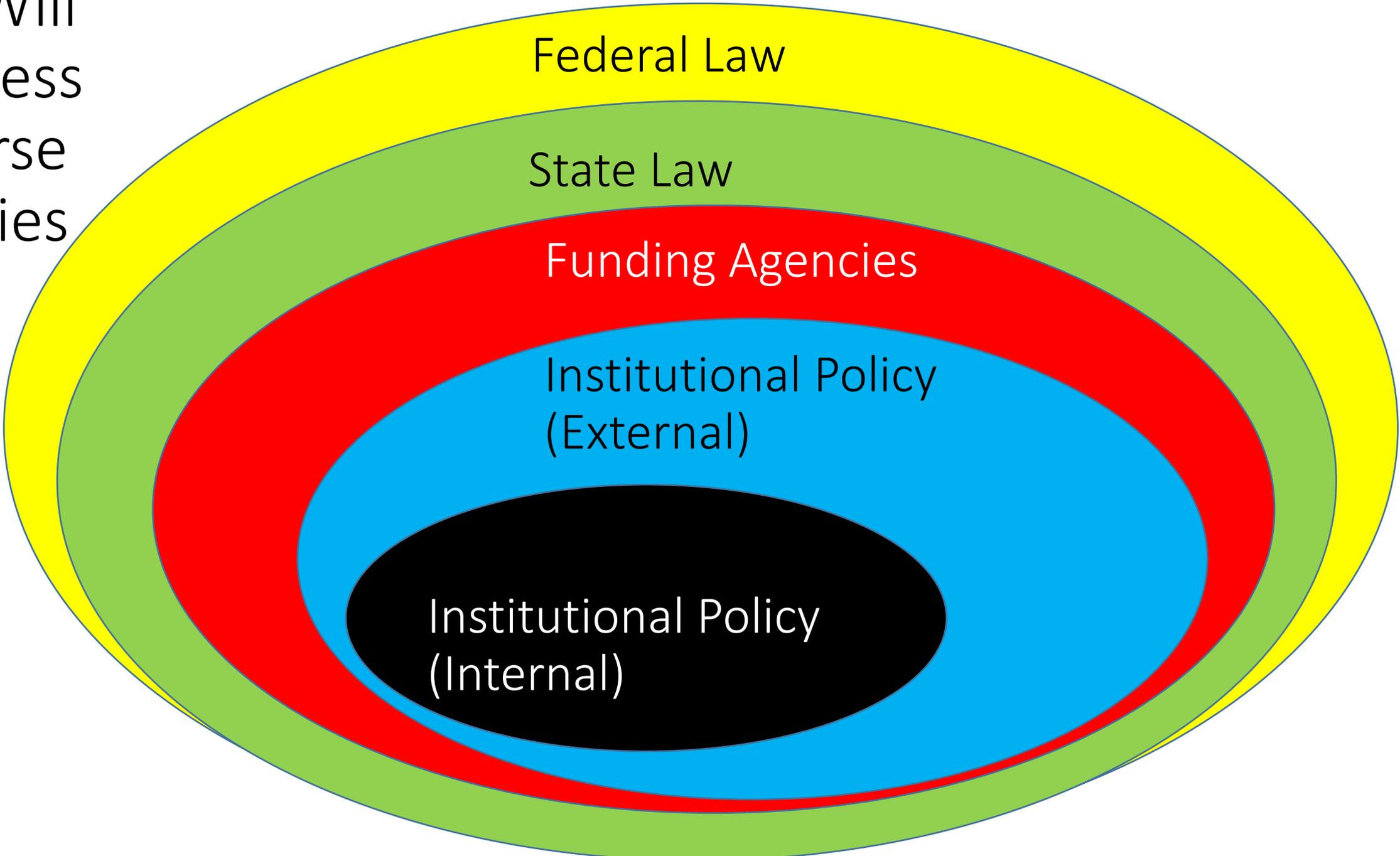
# The Limited Dataset ala HIPAA

Field	Detail
Names	Related to patient (not provider)
Unique Numbers	Phone, SSN, MRN, ...
Internet	Email, URL, IP addresses, ..
Biometrics	Finger, voice, ...

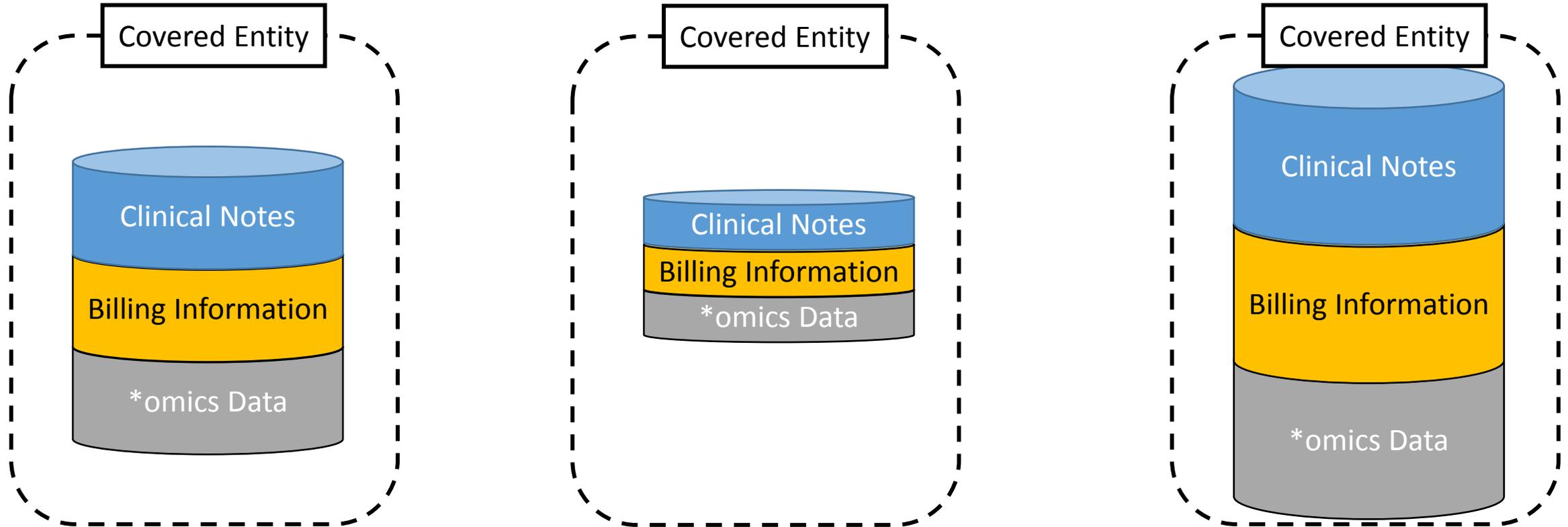
Limited Dataset

Researcher also agrees not to identify patient or misuse data

We Will  
Address  
Diverse  
Policies



# We Will Move From One to Many



What is the incentive to play when you pay disproportionately?

# Use Agreements → Policies

- Policies and laws are long and getting longer and difficult to interpret
- If we are going to integrate and distribute big data over many systems we must develop...
  - ... policies that are codified in computer-manageable languages
- Usability: must be modular and configurable
- Flexibility: must leave room for interpretation, but such room must be clearly demarcated

# NIH Data Sharing Policies (a quick refresher)

- 2003 Final Data Sharing Policy:
  - Studies with > \$500k/yr → Investigators must have data sharing plan or explain why it's not possible
  - Recommends sharing data devoid of identifiers
- 2007 GWAS Policy
  - Studies involving > \$0
  - Recent considerations for extending this to all sequencing data
- Identifiable?

NIH



HIPAA

# HIPAA

## (One of) the Elephants in the Room

“Secondary use” of clinical data is possible, but it varies...

Identified  
Patient Data

- Waiver of consent: data is “on the shelf”
- Consent is impracticable 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

# HIPAA “Cookbook” Standards

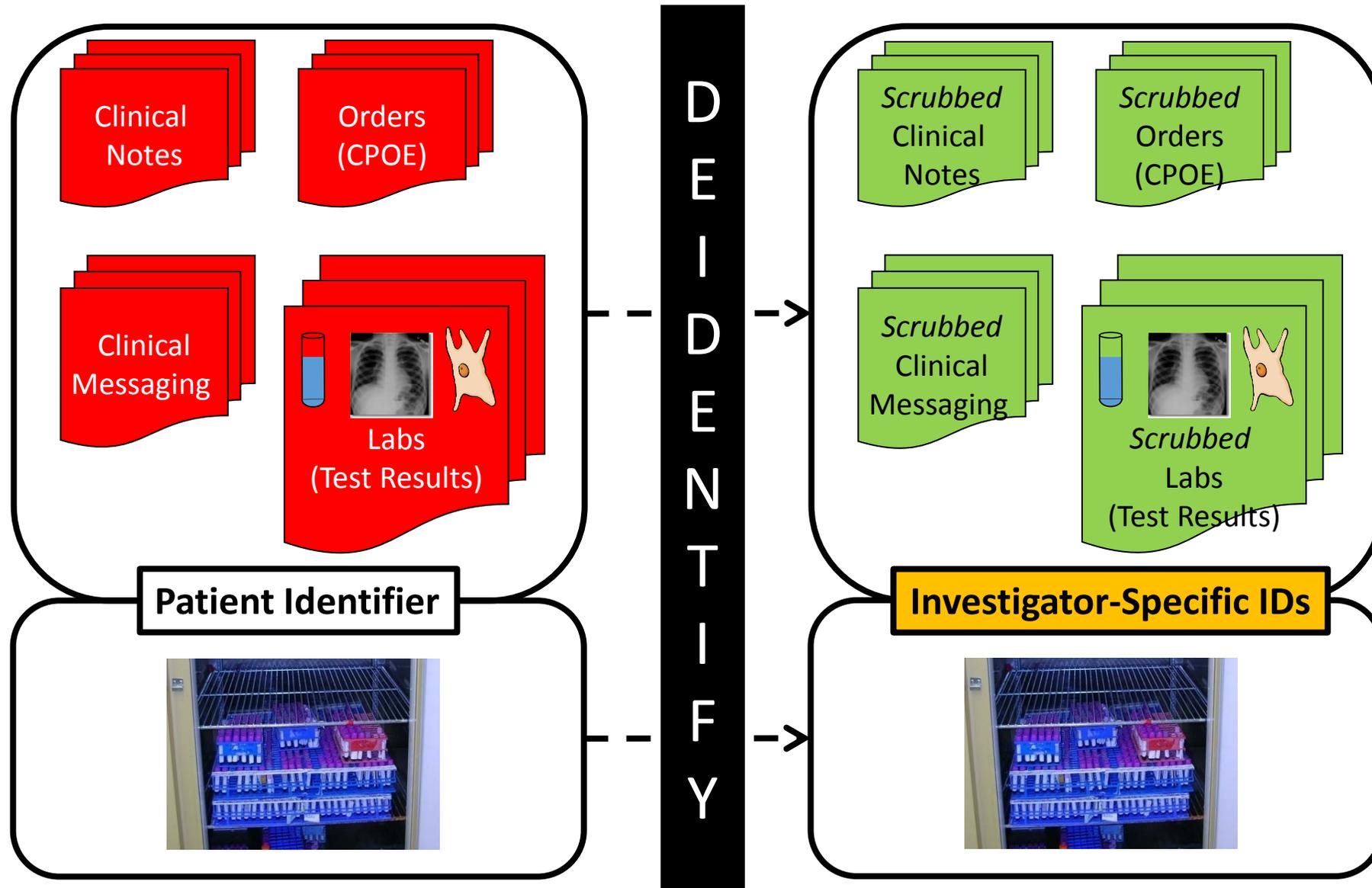
Field	Detail
Names	Related to patient (not provider)
Unique Numbers	Phone, SSN, MRN, ...
Internet	Email, URL, IP addresses, ..
Biometrics	Finger, voice, ...
Dates	Less specific than year Ages > 89
Geocodes	Town, County, Less specific than Zip-3 (assuming > 20,000 people in zone)
“Catch all”	“Any other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

Limited Dataset

Safe Harbor

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

# Vanderbilt's: BioVU Model



# Software: From Theory to Practice

HIDE (Gardner & Xiong, DKE 2009)

MIST (Aberdeen et al, IJMI 2010)

The image displays two side-by-side browser windows. The left window is titled "HIDE™: Health Information DE-identification - Mozilla Firefox" and shows the website at <http://www.mathcs.emory.edu/hide/>. It features a navigation menu with links for Home, Publications, Software, and People, and a "Project Overview" section. The right window is titled "MIST: The MITRE Identification Scrubber Toolkit - Mozilla Firefox" and shows the website at <http://mist-deid.sourceforge.net/>. It has a large "MIST" logo and the text "The MITRE Identification Scrubber Toolkit". Below this is a list of questions: "What is it?", "How does it work?", "Why did we build it?", "Where can I get it?", "What license does it have?", and "How mature is it?". To the right of these questions is a paragraph explaining that MIST is a suite of tools for identifying and redacting personally identifiable information (PII) in free-text medical records, and an example of a document snippet with PII redacted.

# NLP for De-id (MIST)

**File: SamplePathFinDxFAKE.txt (task HIPAA Deidentification)**

Workflow: Hand annotation ▾ Replacer: Select replacer... ▾

Status: clean ► zone ► hand tag ► nominate ► transform ◀ ▶ Reload

Document Save ▾ Legend

[DailyHL7\_SURG\_HISDX\_DIAGNOSIS]

PALATAL LESION, EXCISION (ACME, G22-12345):

- ADENOCARCINOMA, OF MINOR SALIVARY GLAND ORIGIN, INTERMEDIATE TO HIGH GRADE, NOT OTHERWISE SPECIFIED (SEE COMMENT).
- TUMOR SIZE: AT LEAST 1.0 X 0.6 CM
- CAPILLARY LYMPHATIC SPACE INVASION: NOT IDENTIFIED.
- PERINEURAL INVASION: PRESENT.
- RESECTION MARGINS: FOCAL PRESENT AT THE DEEP MARGIN

MOUTH LESION, RE-EXCISION PREVIOUS BIOPSY SITE (ACME, G22-12346):

- NO RESIDUAL ADENOCARCINOMA IDENTIFIED.
- ALL MARGINS FREE OF ADENOCARCINOMA
- PREVIOUS BIOPSY SITE CHANGES

COMMENT: The initial palatal incisional biopsy from March of 2022 (Good Health 22-54321c) was sent to Dr. John Doe at Acme General Hospital who has special expertise in oral pathology. In his consultation, Dr. Doe acknowledges the tumor to adenoid cystic carcinoma, solid variant, and is best classified as above. The resection is clear margins with no evidence of residual tumor.

Please see outside surgical pathology report for immunohistochemical profile.

PATHOLOGIST: Doe MD, Jane E. Doe MD 01/2022

Hand annotation available (swipe or left-click)

**Content tags**

xxxxx	AGE
xxxxx	DATE
xxxxx	EMAIL
xxxxx	HOSPITAL
xxxxx	IDNUM
xxxxx	INITIALS
xxxxx	IPADDRESS
xxxxx	LOCATION
xxxxx	NAME
xxxxx	OTHER
xxxxx	PHONE
xxxxx	SSN
xxxxx	URL

**Structure tags**

xxxxx	lex
xxxxx	untaggable

Add AGE (A)  
Add DATE (D)  
Add EMAIL (E)  
Add HOSPITAL (H)  
Add IDNUM (J)  
Add INITIALS (C)  
Add IPADDRESS (I)  
Add LOCATION (L)  
Add NAME (N)  
Add OTHER (O)  
Add PHONE (P)  
Add SSN (S)  
Add URL (U)  
Repeat OTHER (=)  
Cancel (<ESC>)

# Is NLP for De-id Feasible?

(Aberdeen et al. 2010)

- EMR Records (No Name or Place Dictionaries invoked)
- Machine learning based on “conditional random fields”
- Four document classes: Discharge Summaries (DS), Letters, Labs, Orders

	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***

# 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!

# 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									Extractor)		
HIPAA									Recall	Precis.	
Pat. name									67	.33	
Age									00	.00	
Phone #									50	1.00	
Address									00	--	
Date									06	.03	
MRN									00	--	
Acct. #									00	--	
Other ID #s	10	9	0.90	0	0	.00	--	2	0	.00	.00
ALL	323	47	0.15	7	0	.00	.00	62	6	.13	.10
OTHER											
Prac name	82	9	0.11	5	4	.44	.80	8	4	.44	.50
Org. name	27	20	0.74	8	6	0	.75	3	1	0	.33
ALL	109	29	0.27	13	10	0	.77	11	5	.17	.45

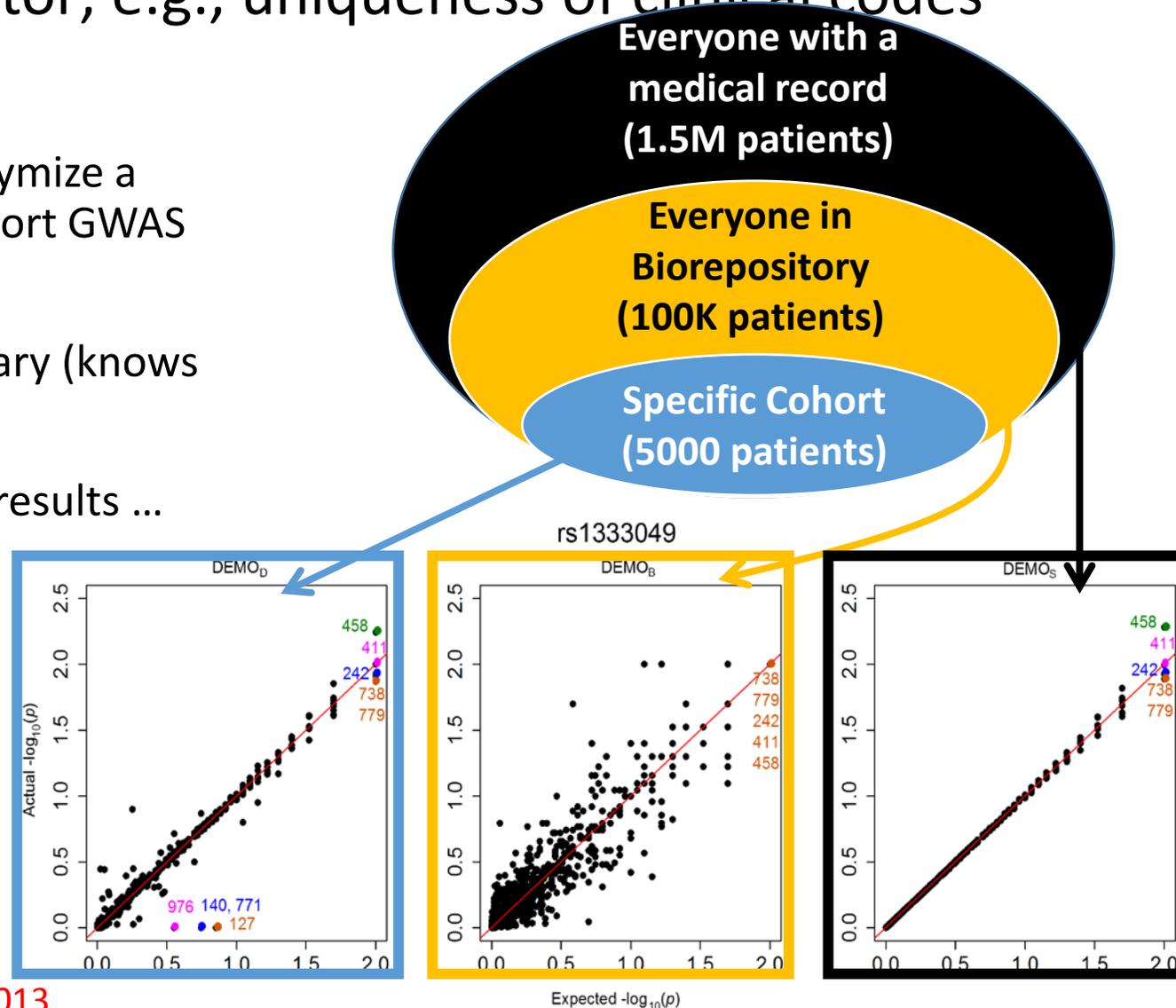
Can effectively raise  
de-identification performance  
from to 0.99 ... but will IRBs  
accept it?

# HIPAA Expert Determination (abridged)

Certify via “generally accepted statistical and scientific principles and methods, that the **risk is very small** that the information could be used, alone or in combination with other **reasonably available information**, by the **anticipated recipient** to identify the subject of the information.”

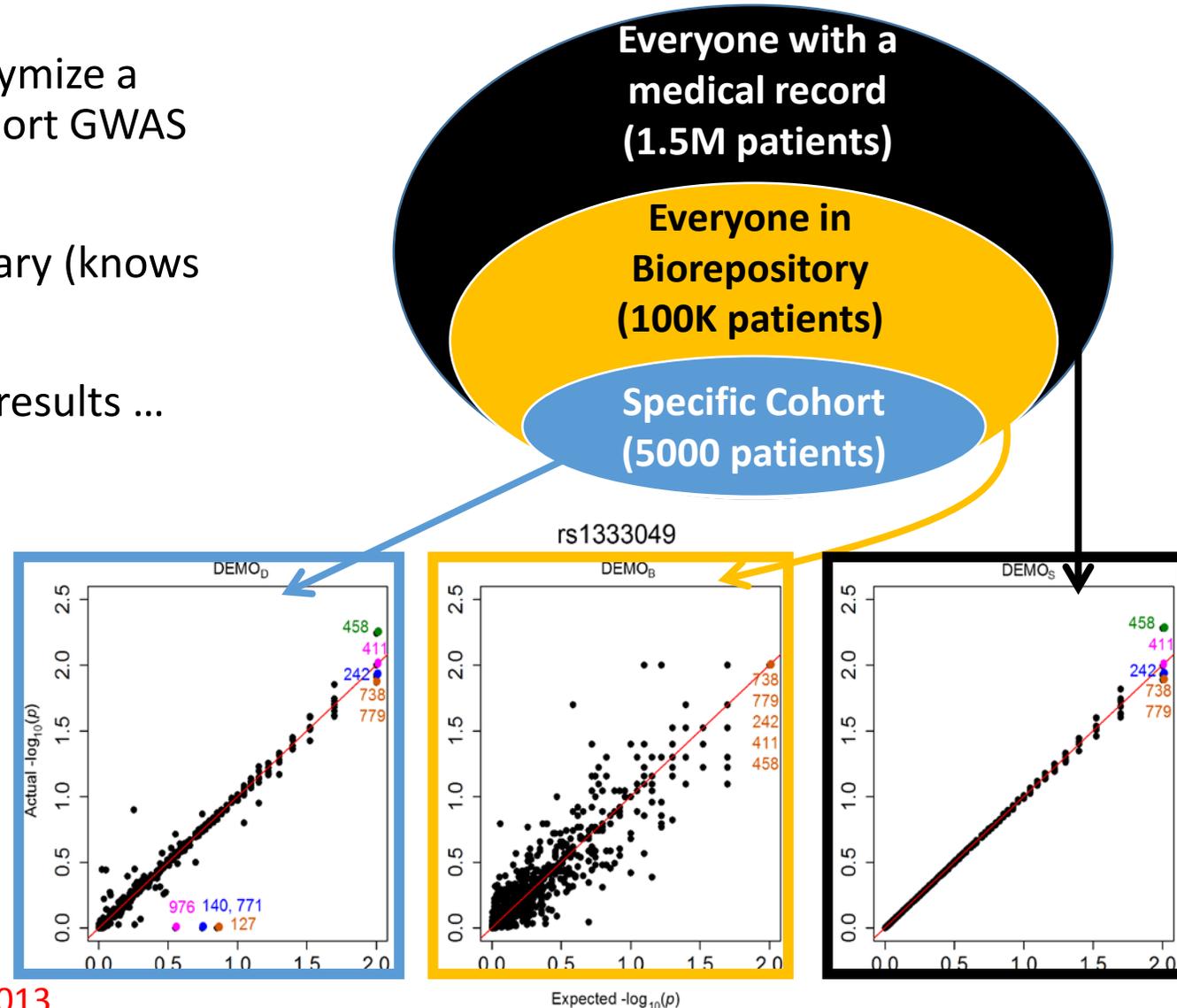
# Many Ways to Formally “Anonymize”

- Must know the attack vector; e.g., uniqueness of clinical codes
- Previous work showed how to anonymize a patients’ set of billing codes to support GWAS validation
- but it assumed a very strong adversary (knows who’s in cohort)
- But you can achieve almost perfect results ...
- ... when adversary is more realistic
- Validation of 192 SNP – phenotype associations



# Anonymized Clinical Data in Big Groups

- Previous work showed how to anonymize a patients' set of billing codes to support GWAS validation
- but it assumed a very strong adversary (knows who's in cohort)
- But you can achieve almost perfect results ...
- ... when adversary is more realistic
- Validation of 192 SNP – phenotype associations



But are we Witnessing the  
Death of Privacy?

# Identifying Personal Genomes by Surname Inference

Melissa Gymrek,<sup>1,2,3,4</sup> Amy L. McGuire,<sup>5</sup> David Golan,<sup>6</sup> Eran Halperin,<sup>7,8,9</sup> Yaniv Erlich<sup>1\*</sup>

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases.

We show that this technique can be used to identify individuals who rely on family identifiers with high accuracy.

## RESEARCH ETHICS

Surnames are a common way to identify individuals in recreational genealogy databases. Basic patrilineal

## The Complexities of Genomic Identifiability

Laura L. Rodriguez,<sup>1</sup> Lisa D. Brooks,<sup>1</sup> Judith H. Greenberg,<sup>2</sup> Eric D. Green<sup>1\*</sup>

Sharing research data has long been fundamental to the advancement of science. In today's scientific culture, making research data available broadly and efficiently via the internet has become the standard for many data types, including genomic and some other "omic"-type data produced by high-throughput methods. The acceleration of research progress and the resulting public benefit achieved through such broad data-sharing have been transformative for the scientific enterprise (1–3). However, sharing data generated from human research participants must be done in a manner that appropriately protects participant interests.

Several recent studies have suggested that some analyses of high-dimensional molecular data can raise



## POLICYFORUM

Recent work reveals the need to re-examine the current paradigms for managing the potential identifiability of genomic and other "omic"-type data.

data and genealogic information derived from these individuals and their relatives (10, 11). The CEPH participants whose samples were included in the HapMap Project (and then in the 1000 Genomes Project) underwent a process of re-consent to inform them about the plans for providing very broad and open access to the genomic data derived from their samples and for the in-depth genomic analyses that would be performed on those data. The inability to guarantee privacy and the possibility—then seen as remote—that individual identification might eventually become feasible were described explicitly. Despite this hypothetical and assumed low risk of identification, Gymrek *et al.* have now shown that it is possible to identify some participants of a genomics research

By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (10) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (11) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigen-

and western European and originally had their Centre d'Etude du were later reconstructed from the HapMap project.] did not pursue full resolution.

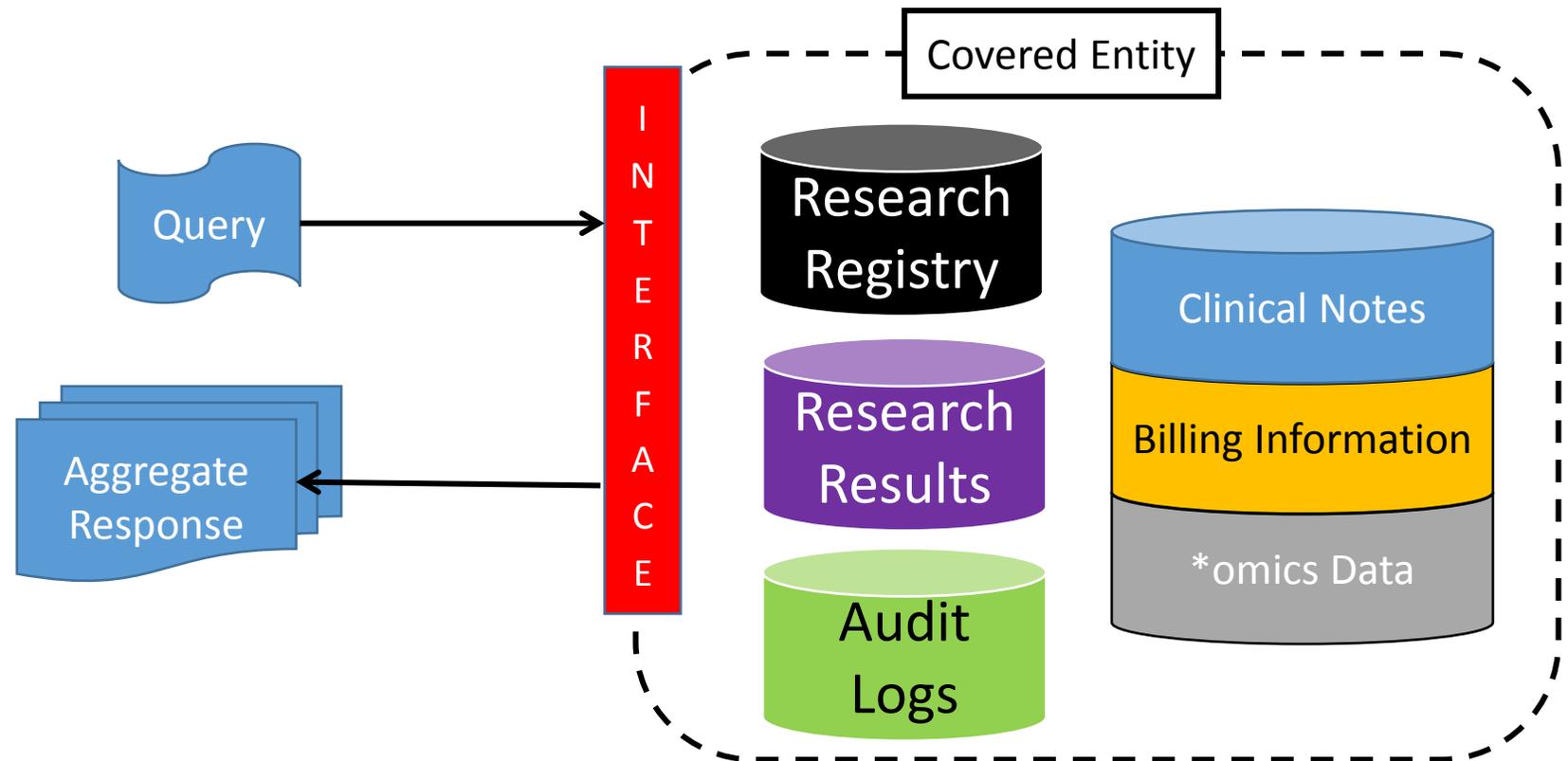
vely approach the population, apply some data sets, and identification of individuals. We show that genomes can be derived from recreational genealogy databases allowed by Internet 1 individuals were A samples, the individuals had signed stated and the data usage ication. Representatives that funded the and confirmed the their guidelines (12). surname inference, (mysearch.org) and

# Risk-Based Privacy

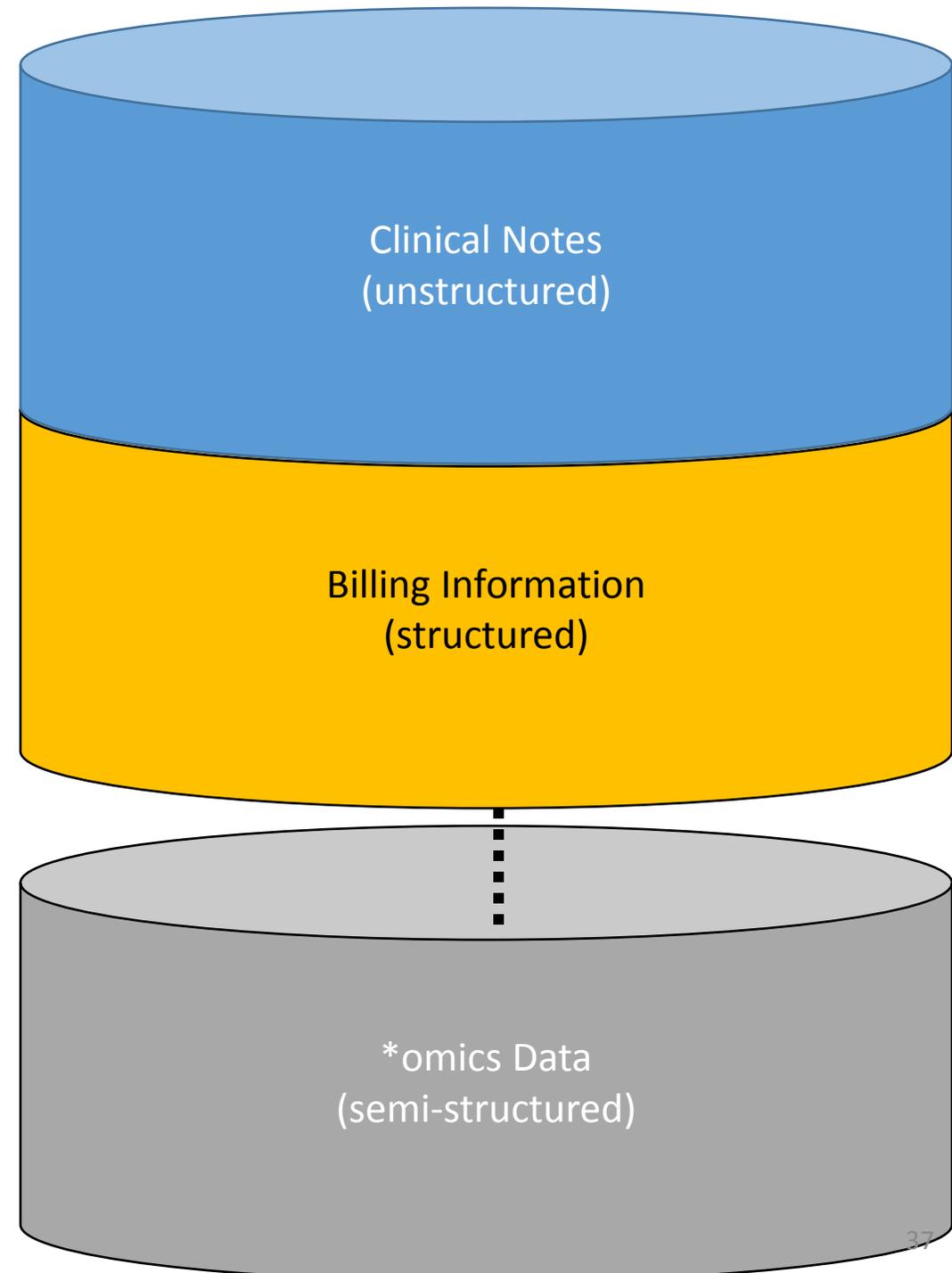
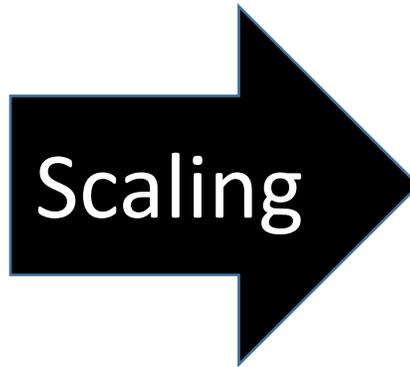
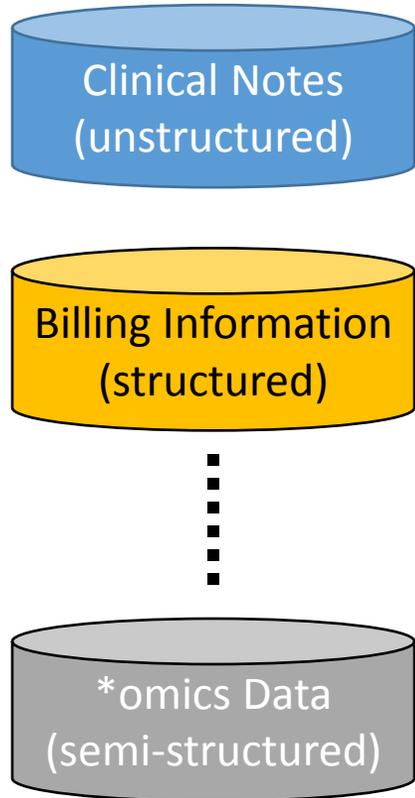
- It's not always “easy” to identify patients. We can quantify the risk.
- IRBs need training in “informational risk” and data-based studies...
- But if the IRBs will not take responsibility, then who?
- National center(s) of excellence for de-identification method vetting and assistance?
- De-identification too strict – let's say “data protection”

# Query-Response DB's

- Hold all of the data local and let people query it...
- Provide only aggregate responses
- What are the “right” methods?
- What if the user overuses the system?



# Big Can Get Really Big!



# Quick, Robin! To the CloudMobile!

- Computer and network security risks have always existed
- So why should we be concerned about the cloud?
  - Lack of institutional oversight for physical security
  - Collocation of data from disparate organizations
  - Concerns over liability for breaches



Batmobile, circa 1989

# Can Encryption Be Our Friend?

- Keep the data encrypted at all times
- We can analyze / mine encrypted data (e.g., “homomorphic” cryptosystems)
- Opportunities
  - Breakthroughs in crypto are making it faster everyday
  - Reduce trust required in 3<sup>rd</sup> party manager
- Challenges
  - Will users trust clinical data they can't see?
  - What functions / mining methods are need to support clinical research workflows?

Privacy v Utility  
Optimization

Policy Management

Trustworthy  
Frameworks

Engagement  
Environments

# Acknowledgements

## Vanderbilt

- Ellen Wright Clayton
- Josh Denny
- Jonathan Haines
- Raymond Heatherly
- Grigorios Loukides\*
- Dan Roden
- Reyyan Yeniterzi\*

## Beyond

- John Aberdeen
- Sam Bayer
- David Carrell
- Cheryl Clark
- Lynette Hirschman
- Li Xiong
- Ben Wellner