EHRs and Genomic Science: eMERGE and Beyond

Justin Starren, MD, PhD, FACMI

Director, Northwestern University Biomedical Informatics Center (NUBIC)

Chief, Division of Biomedical Informatics, Department of Preventive Medicine

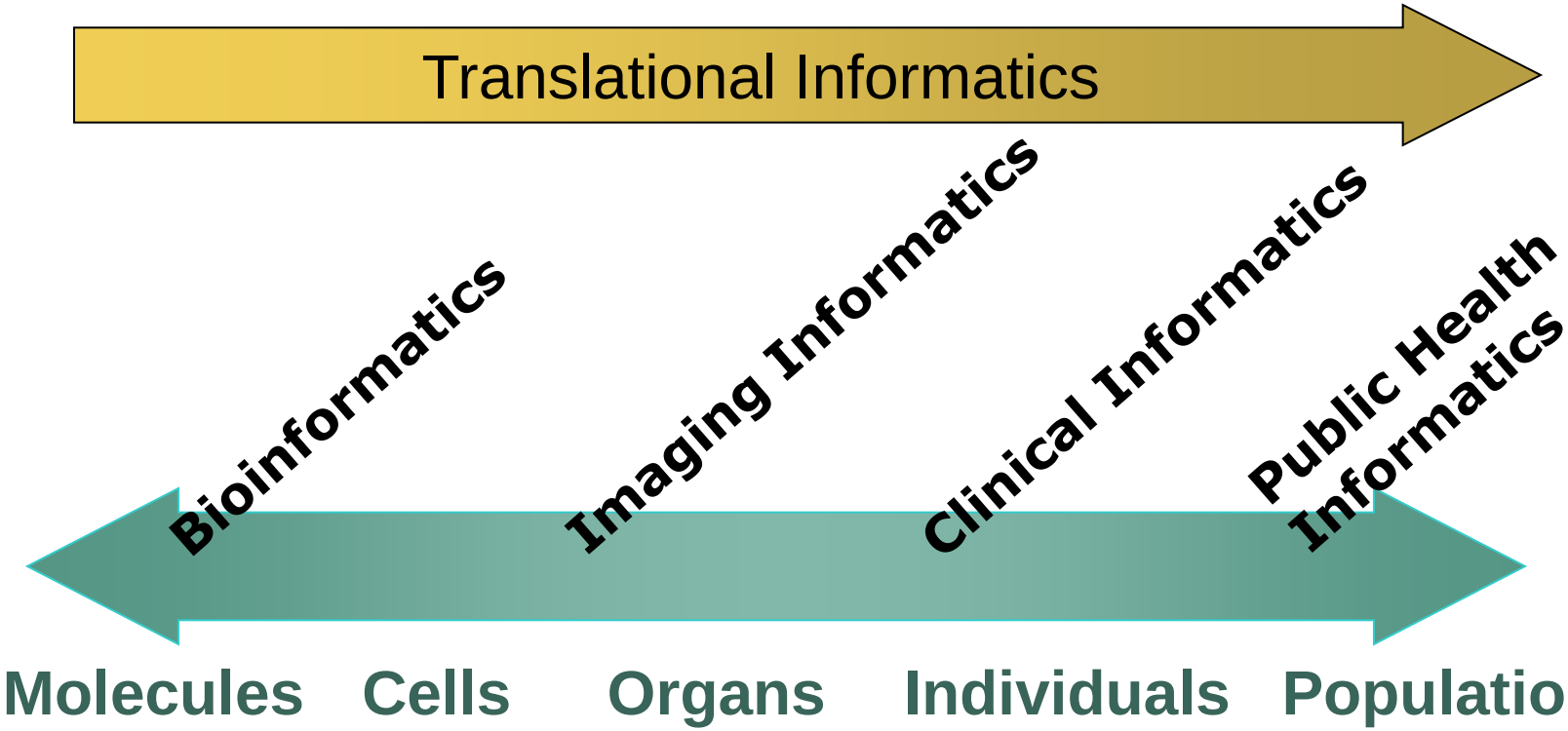
Formerly

Director, Biomedical Informatics Research Center, Marshfield Clinic

(an eMERGE site)

Terminological Plea

Biomedical Informatics



Acknowledgements

Many thanks to:
Dan Masys
Brad Malin
Terrie Kitchner
Marilyn Ritche
Cathy Schneider
Carol Waudby
Russell Wilke
Peggy Peissig
Luke Rasmussen
Richard Berg
James Linneman
Cathy McCarty
Carl Christensen
Able Kho

- This work funded by NIH grant 5U01HG004608-02 from NHGRI.
- MedLEE was supported in part by grants R01LM007659 and R01LM008635 from NLM.

The eMERGE Network

electronic Medical Records & Genomics

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies

Main Menu

- Home
- Member Sites
- About eMERGE Network
- Links
- Contact Us

User Menu

- Log in

URL:

www.gwas.net

The eMERGE Network

The eMERGE Network is a national consortium formed to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research.

The mapping of the human genome has enabled new exploration of how genetic variations contribute to health and disease. To better realize this promise, researchers must now determine ways in which genetic make-up gives some individuals a greater chance of becoming sick with chronic conditions such as diabetes, Alzheimer's, or heart disease, in order to ultimately improve patient care.

There are a number of studies conducted routinely to uncover the association between disease and a person's genetic make-up, but they are typically costly and take a long time to complete. This consortium will use data from the EMR – clinical systems that represent actual health care events, an alternative methodology, which is highly cost and time-efficient, to propel this research. Electronic medical records are one of the most exciting potential resources for research data.

Each center participating in the consortium, organized by the National Human Genome Research Institute

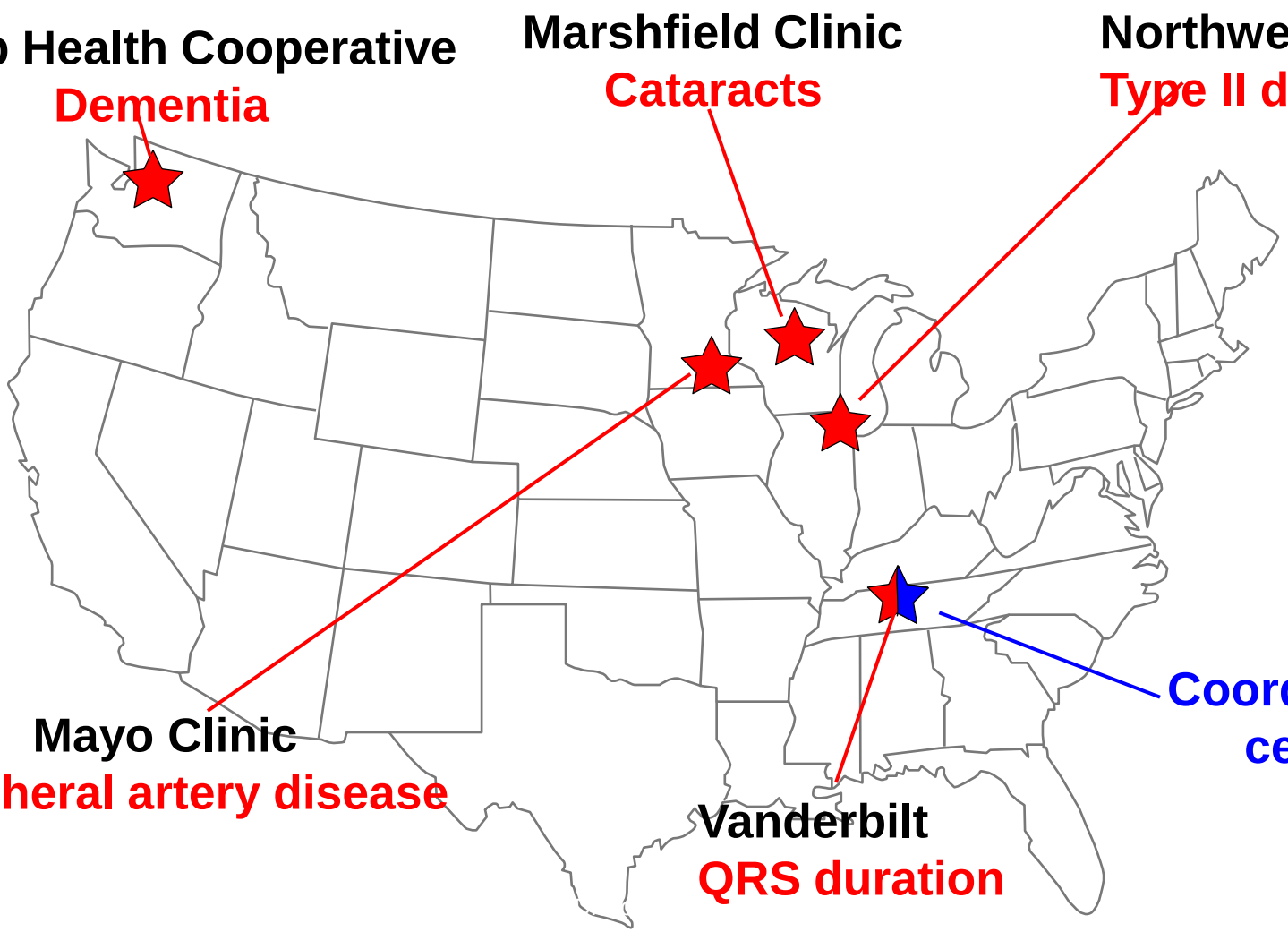
eMERGE I Questions

- Technical
  - Is the information in the EMR?
  - How to get it out?
  - Does it work across institutions?
- Ethics, Legal, Social (ELSI)
  - Recruiting (Purposeful / Opportunistic)
  - Consenting (Opt in / Opt out)
  - Privacy
  - Data Use

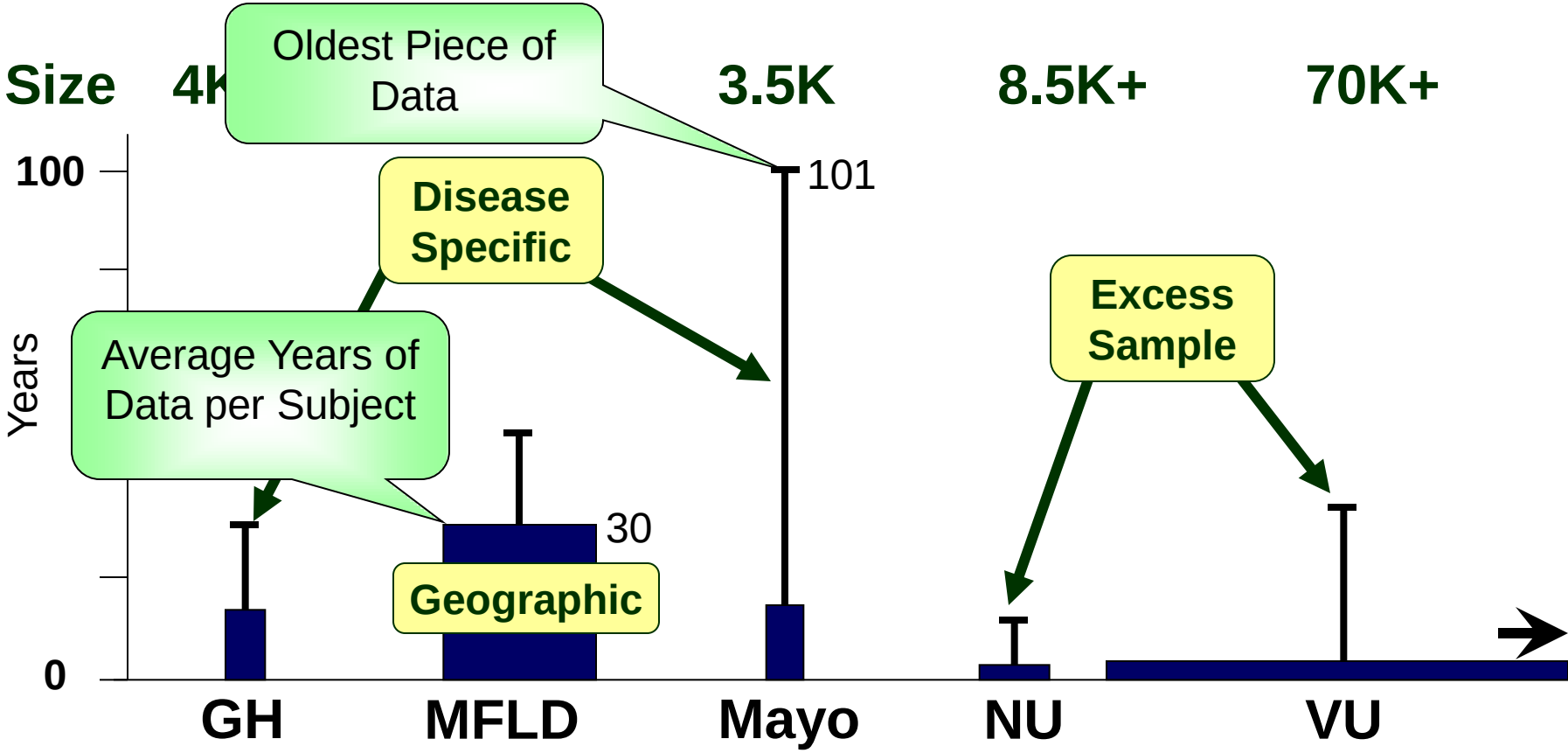
The eMERGE Network

electronic Medical Records & Genomics

A consortium of biorepositories linked to electronic medical records data for conducting genomic studies



Biorepository Overview



EMR:	EPIC	Custom	GE/ Custom	EPIC/ Cerner	Custom

Why EHR data for Genome Research?

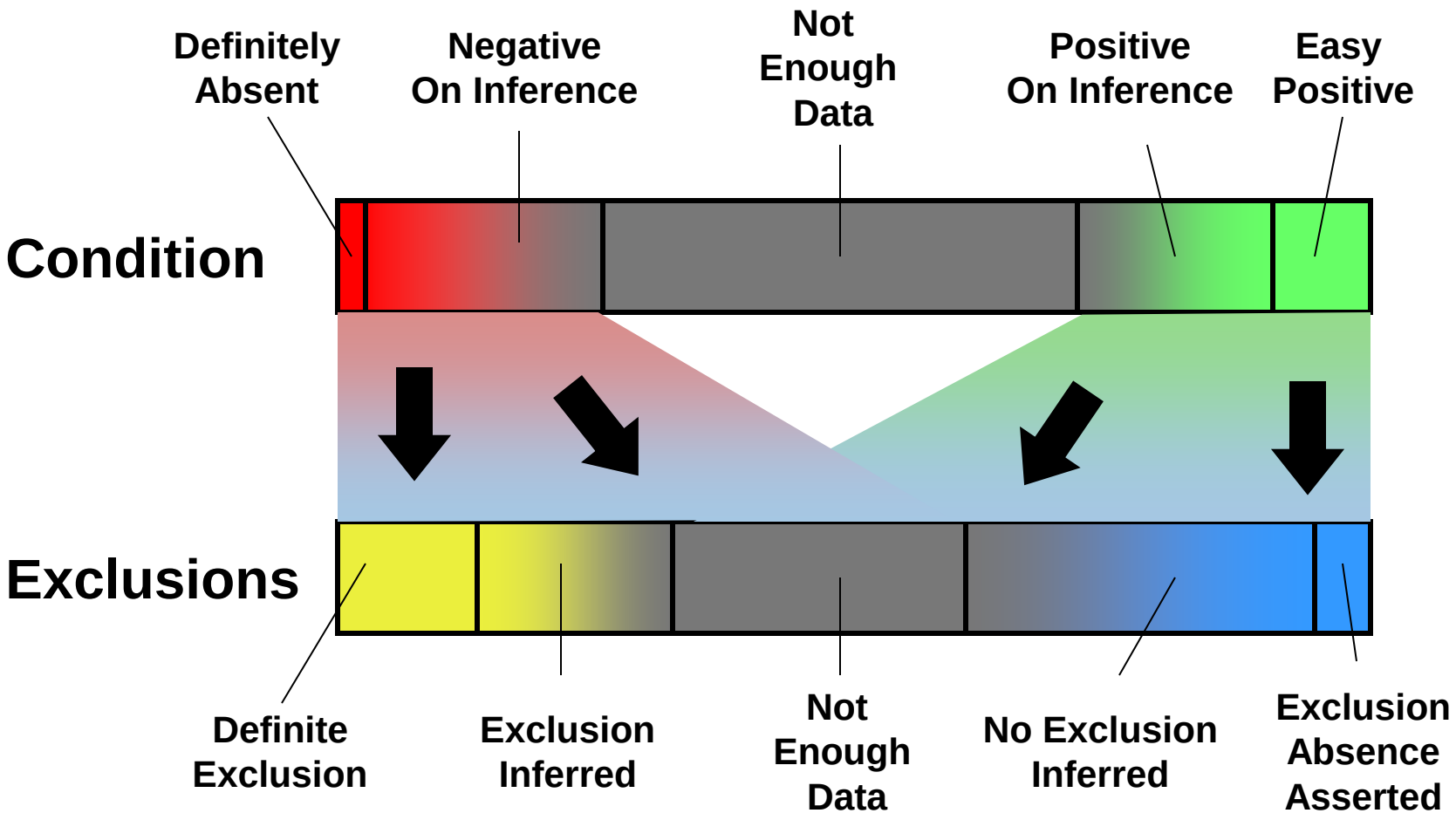
- Data is already there
  - Reduced cost
  - Increased speed
- Possibility for iterative refinement
  - Co-morbidities
  - Confounders
  - Incorporate new knowledge
- Evaluate temporal changes
- Represents “real-world” experience

Poly Phenotyping
(Associations between 2+ diseases)

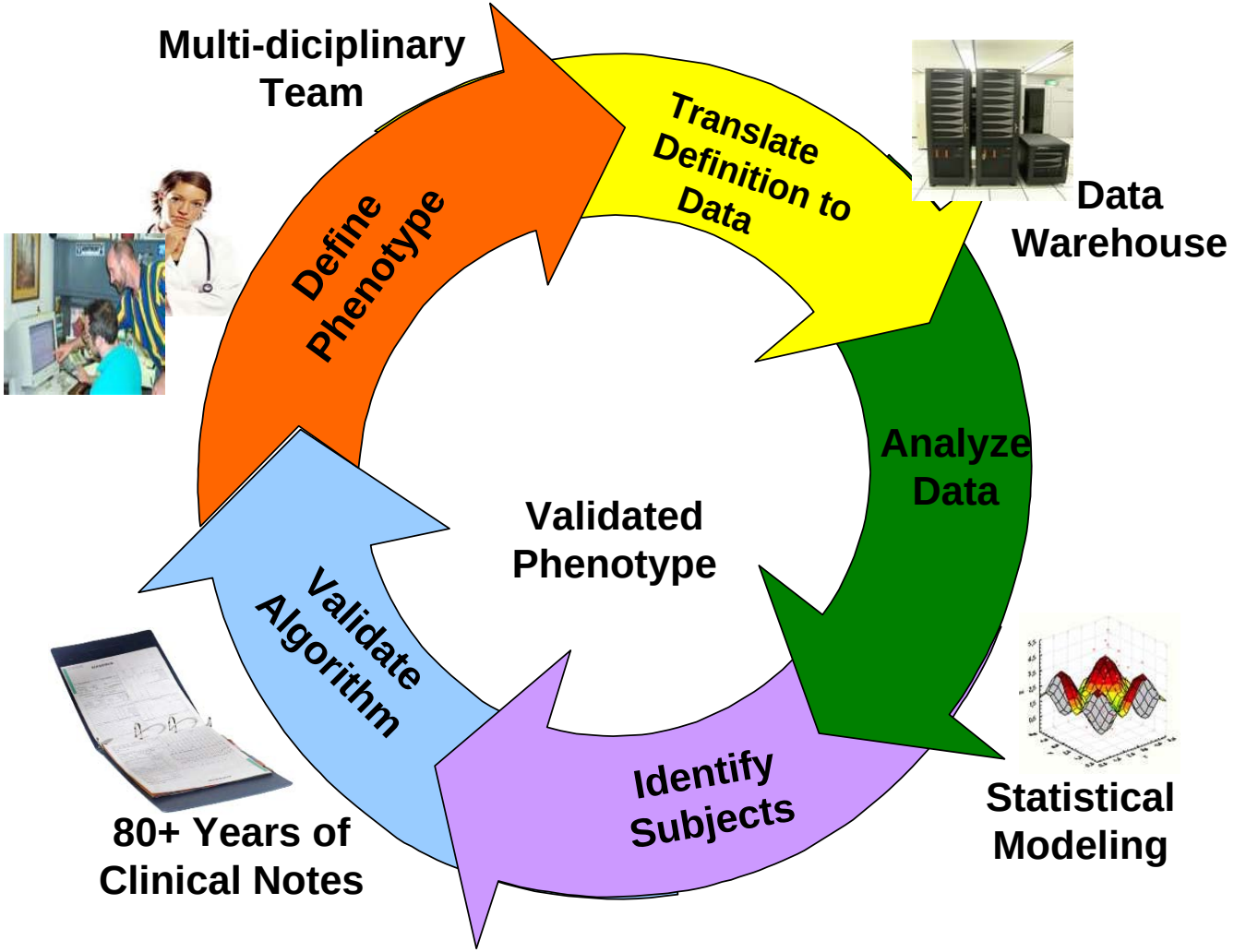
Phenotype	Total	Two Phenotypes									
		Obesity	Increased Fasting Gluc	Increased Triglyceride	Low HDL	Hypertension	Diabetes DX	MI	All CAD	Statin Exposure	Asthma
Obesity	**7353**										
Increased Fasting Gluc	**3390**	2015									
Increased Triglyceride	**5864**	3258	2263								
Low HDL	**6232**	3443	2220	4198							
Hypertension	**11671**	5510	3071	5122	5137						
Diabetes Dx	**2592**	1614	1789	1778	1800	2442					
MI	**998**	468	436	660	704	939	470				
All CAD (includes MI)	**2227**	1058	934	1383	1431	2094	962	948			
Statin Exposure	**5173**	2605	2080	3597	3340	4642	1843	860	1795		
Asthma	**2612**	1195	555	904	1000	1775	481	144	362	758	
All Cancers	**3556**	1412	1070	1635	1652	2138	905	428	953	1621	506

* Based on Dec. 2007 dataset,

Ins and Outs of EHR Phenotyping



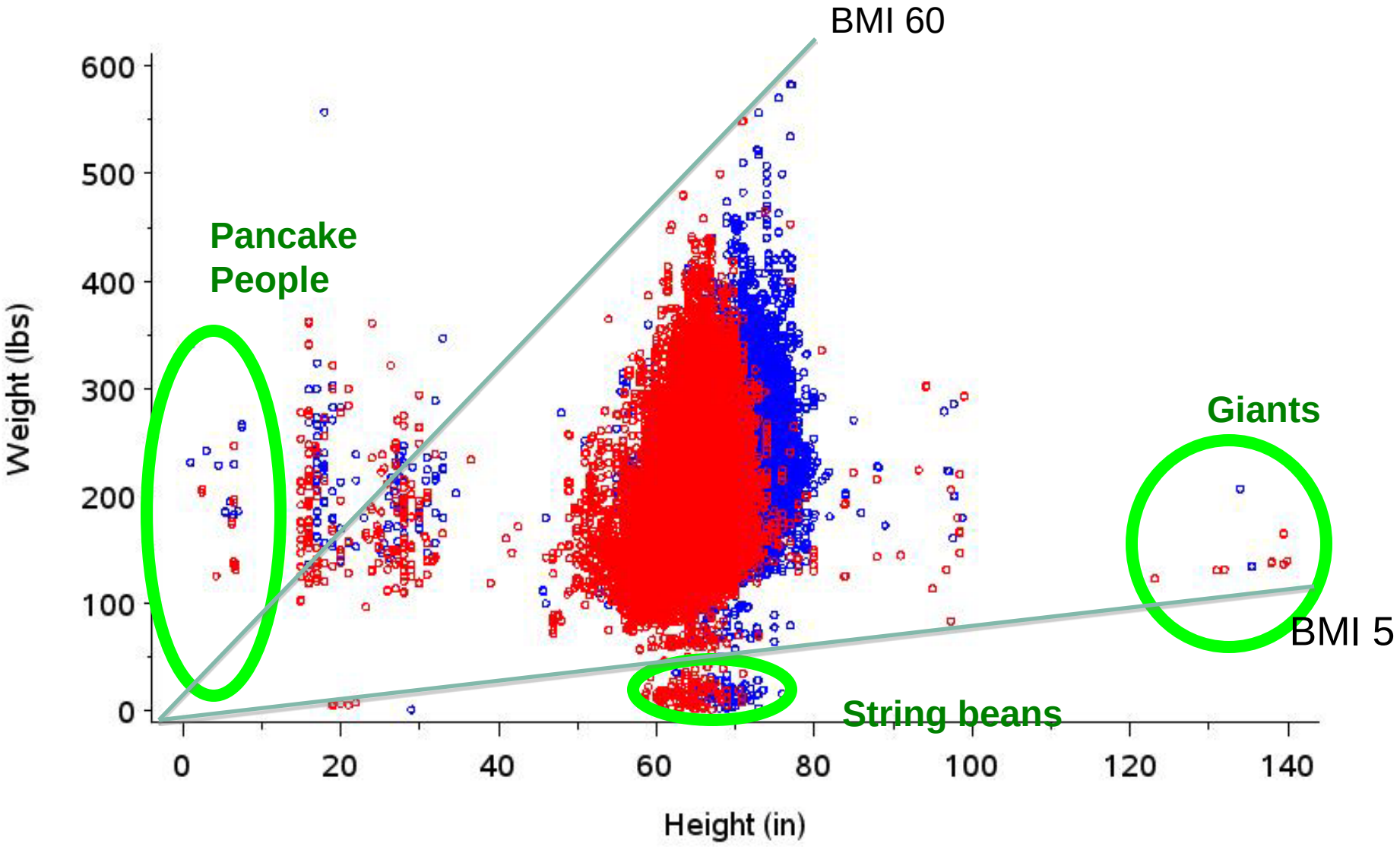
Iterative Phenotyping



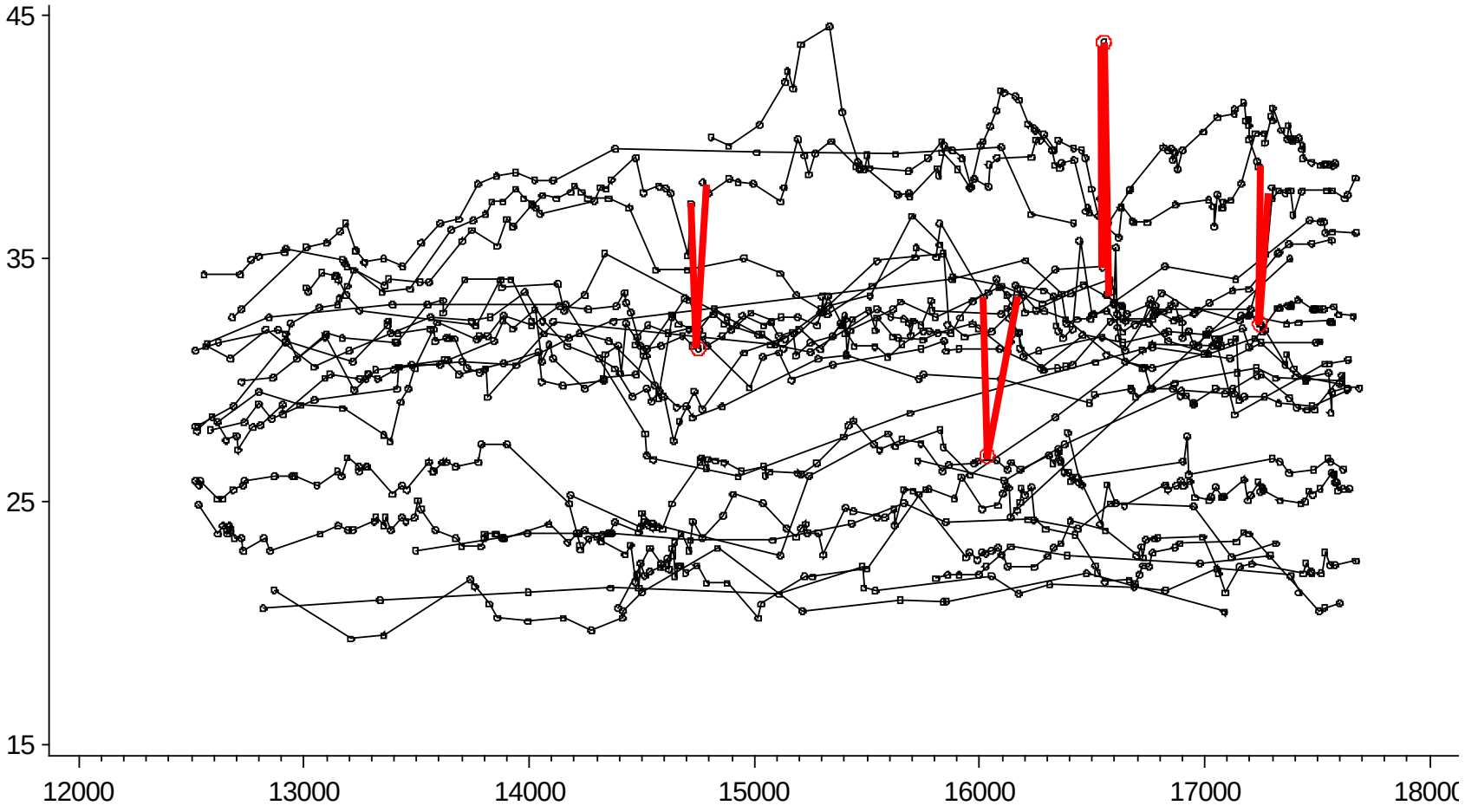
EHR Data Challenges:

- Data are collected for **clinical use**
  - Clinical treatment may not contribute to genetic or research study aims
  - **"Absence of evidence is not evidence of absence."** *Carl Sagan*
- Lack of standardized responses
- Multiple data points collected
  - Single data entry with minimal error checking
  - Duplication or disagreement of facts

Unfiltered Adult Height/Weight

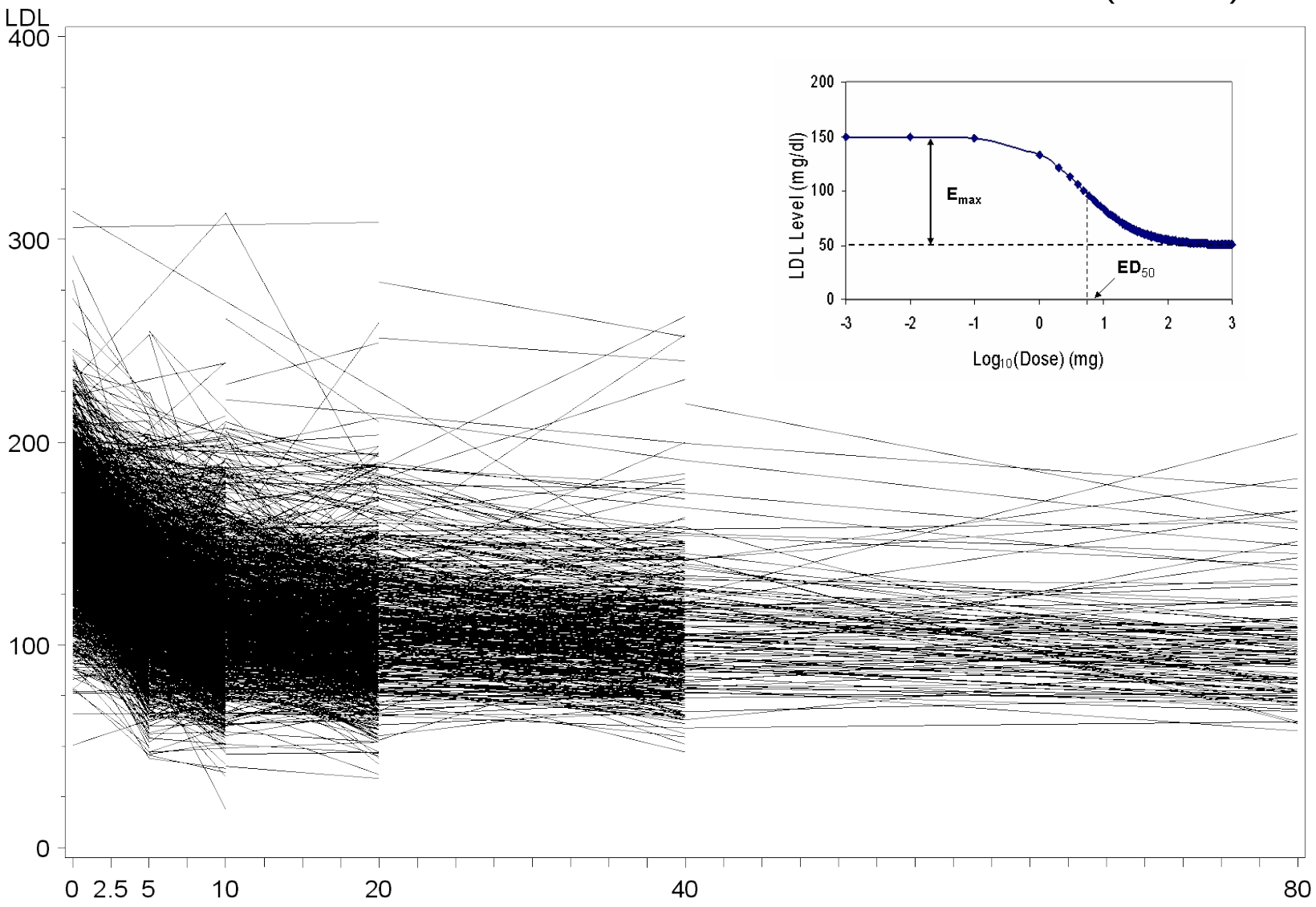


Temporal BMI Patterns



Population Dose-Response

Median LDLs for all Patients with Multiple Valid Dose Results (n=2822)



Phenotyping

	GH	MC	Mayo	NWU	VU
Primary phenotype:	Dementia	Cataracts	PAD	T2 Diabetes	QRS Duration
Phenotyping Approach:					
Billing/Claims:	x	x	x		x
Diagnoses:	x	x	x	x	x
Procedures:		x	x	x	x
Medication:	x	x	x	x	x
Labs:		x	x	x	x
NLP:	X	x	x	X	x
Regular Expression:		x		x	
ICR:		x			

Cross-network phenotype: Hypothyroidism

	GHC	Mfld	Mayo	NU	VU	Total
Cases	310	592	293	97	185	**1,477**
Controls	1,223	649	2905	516	1,476	**7,669**

9,146

- Already Genotyped Samples
- Additional Samples Available
  - 2834 Additional Cases
  - 15,062 Additional Controls

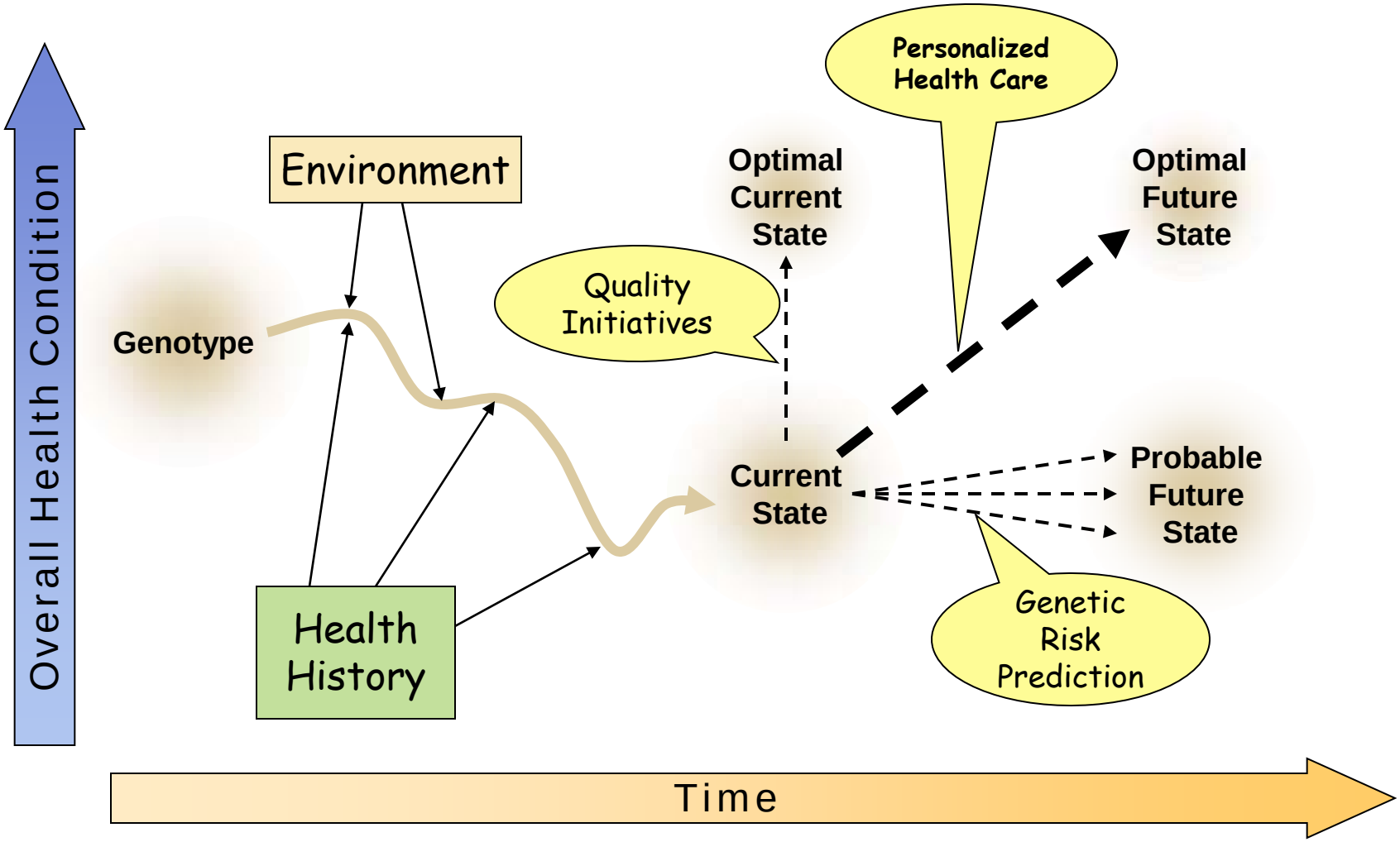
What did I learn from eMERGE?

- EHRs rich source of phenotypic data
- Cost effective
- Successful phenotyping is dependent on
  - Understanding your EMR environment
  - Utilizing a multi-disciplinary team
  - Utilizing creative phenotyping approaches
- Terminology
  - Coded data is always wrong
  - “It’s a dessert topping and a floor wax”
- Workflow
  - “Do it in half the time”

Things NOT to take away

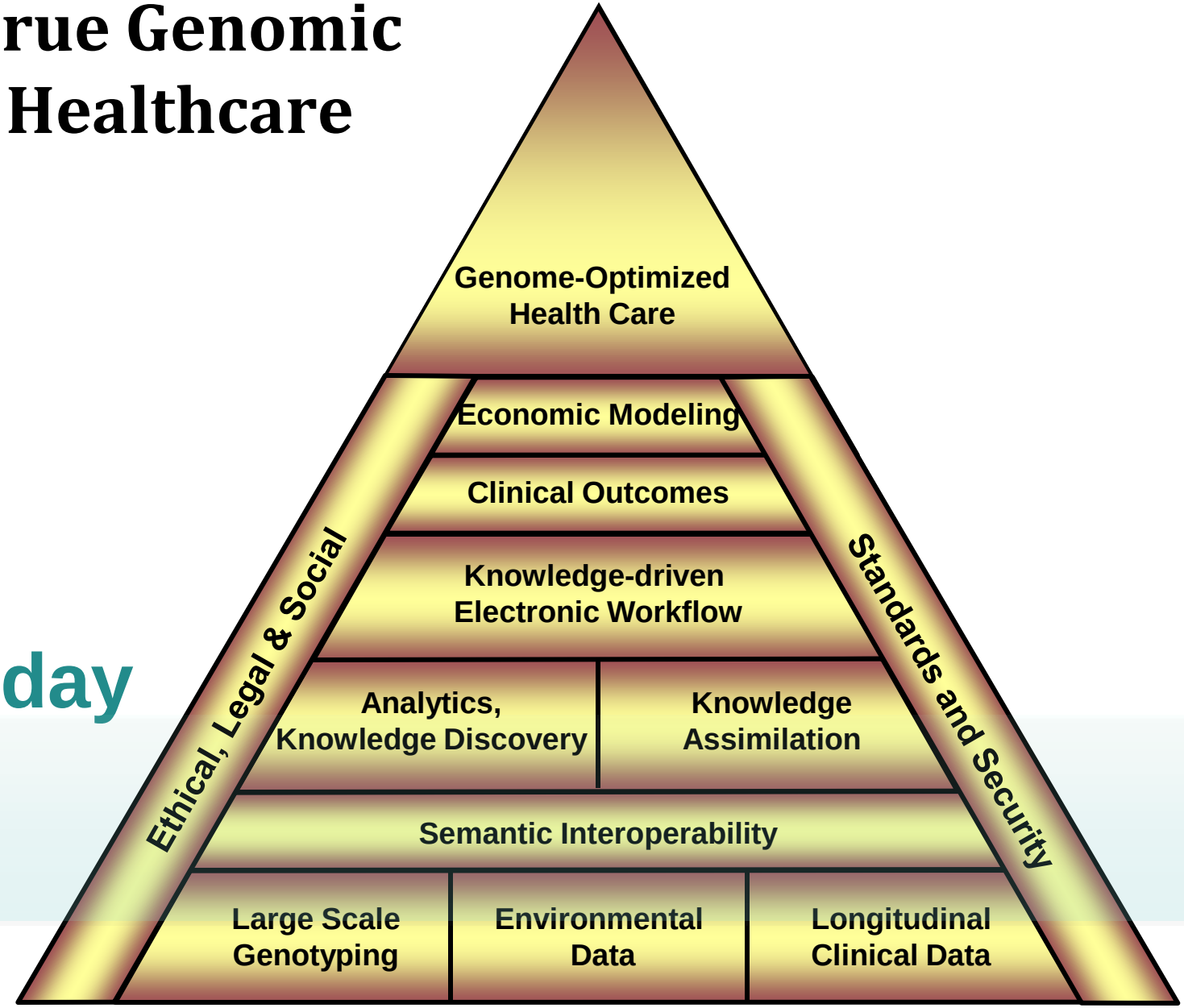
- "If everyone used the same system, we could combine the data."
- "We need to define new standards for…."
- "We need consensus."
- "We could do great research if ~~doctors~~ EHRs would only…"

Relationship between Quality Initiatives and Personalized Health Care



True Genomic Healthcare

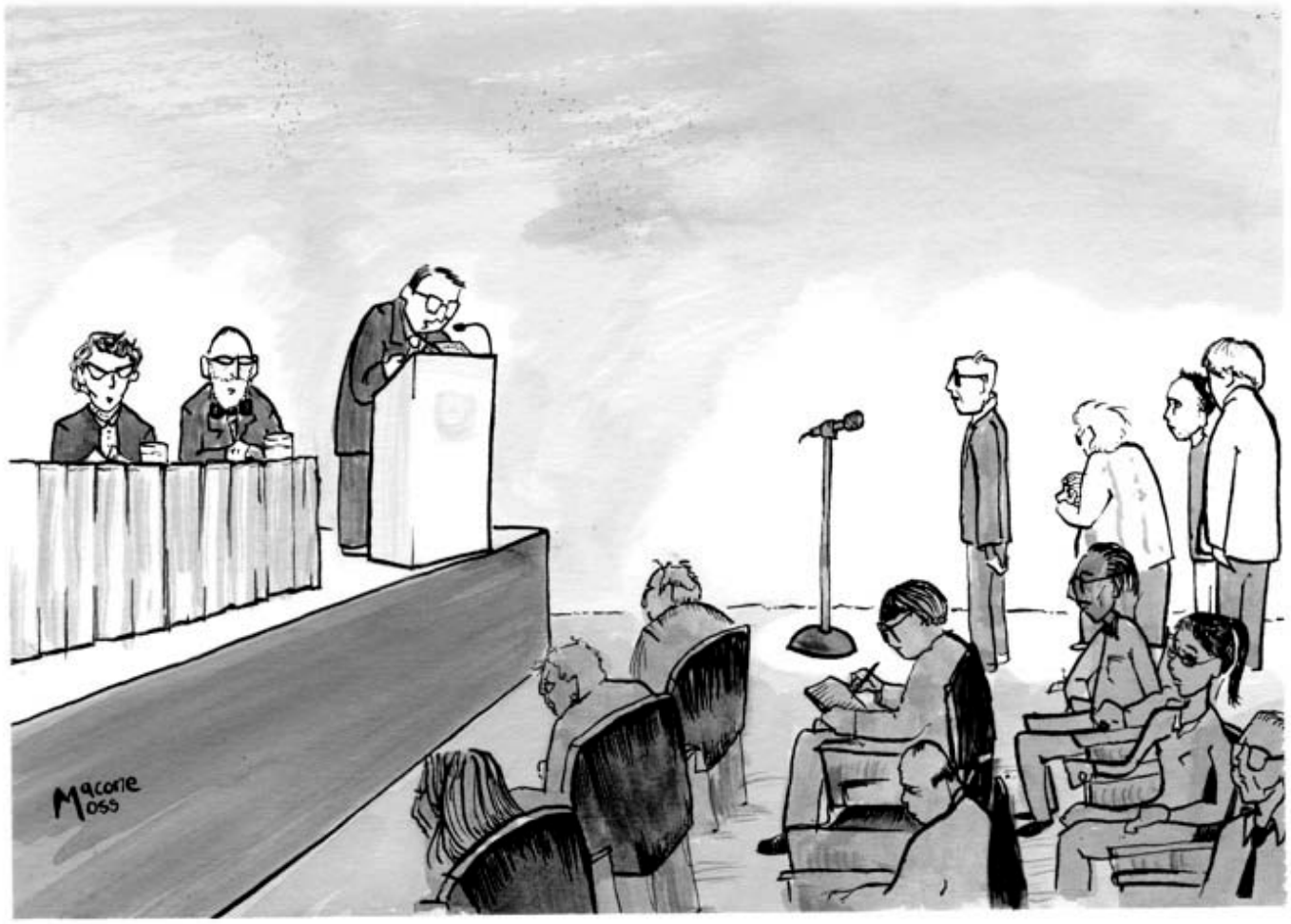
Today



“We like learning from large, noisy data sets”

--Larry Page, Google

The Economist special report on managing information
Feb. 27, 2010



We'd now like to open the floor to shorter speeches disguised as questions.

Biorepository Characteristics

	Group Health	Marshfield	Mayo	Northwestern	Vanderbilt
Name	GHC Biobank	Personalized Medicine Res Project	Bio-repository	NUgene	BioVU
Overview	Disease Specific Registry	Geographic cohort recruitment	Disease Specific Registry	Excess sample collection	Excess sample collection
Size	4,000+	20,000	3,500	8,500 +	70,000 +
Population	96% Caucasian	98% Caucasian	96% Caucasian	12% AA 8% Hispanic	11% AA
Recruit Method	Consent	Consent w/ recontact	Consent	Consent w/ recontact	Opt-out

Supplement phenotypes using genotyped samples from primary phenotypes*

*The benefit of data from routine clinical testing results recorded in EHR

	RBC/WBC	Diabetic Retinopathy	Lipid Levels & Height	GFR
GHC	3,579	230	3,114	1,713
Marshfield	3,865	213	3,693	3,929
Mayo	3,346	806	3,175	3,340
NU	2,484	139	2,816	1,485
VU	2,650	1,449	1,631	2,679

Open Issues

- Supporting personal preferences
- Motivation for data sharing
- Incorporating research into efficient clinical workflow
- Resolve the Quality / Business / Research privacy disconnect
- Standardized IRB / Data Sharing / IP processes
- Technology / Policy balance

The problem with ICD9

- **Limited Granularity**
- **False negatives:**
  - Outpatient billing limited to 4 diagnoses/visit
  - Outpatient billing done by physicians (e.g., takes too long to find the unknown ICD9)
  - Inpatient billing done by professional coders:
    - omit codes that don't pay well
    - can only code problems actually explicitly mentioned in documentation
- **False positives**
  - Diagnoses evolve over time -- physicians may initially bill for suspected diagnoses that later are determined to be incorrect
  - Billing the wrong code (perhaps it is easier to find for a busier clinician)
  - Physicians and Patients driven to find billable ICD code.
    - Example: Anti-TNF biologics (e.g., infliximab) originally not covered for psoriatic arthritis, so rheumatologists would code the patient as having rheumatoid arthritis
    - Wisconsin just mandated insurance coverage for Autism