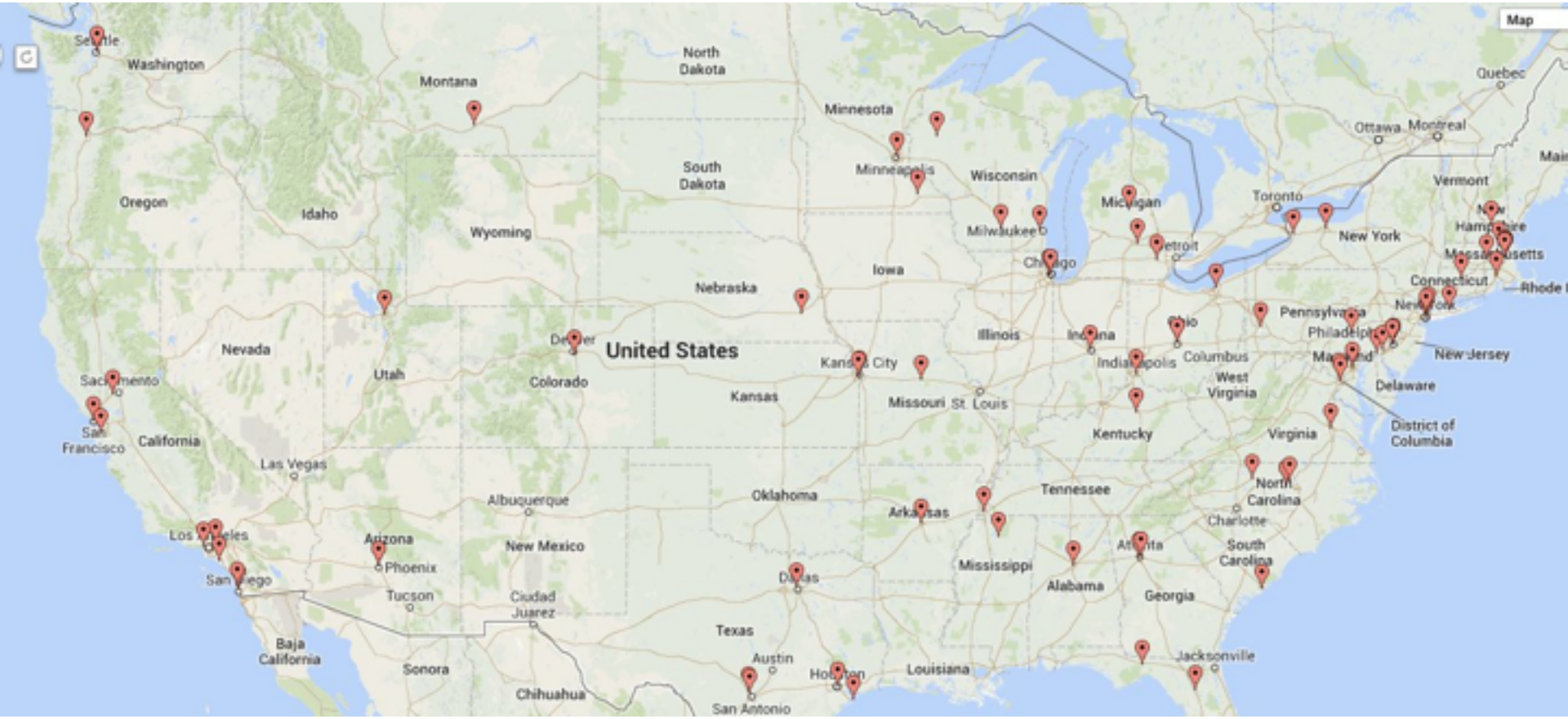


cbmi.med.harvard.edu

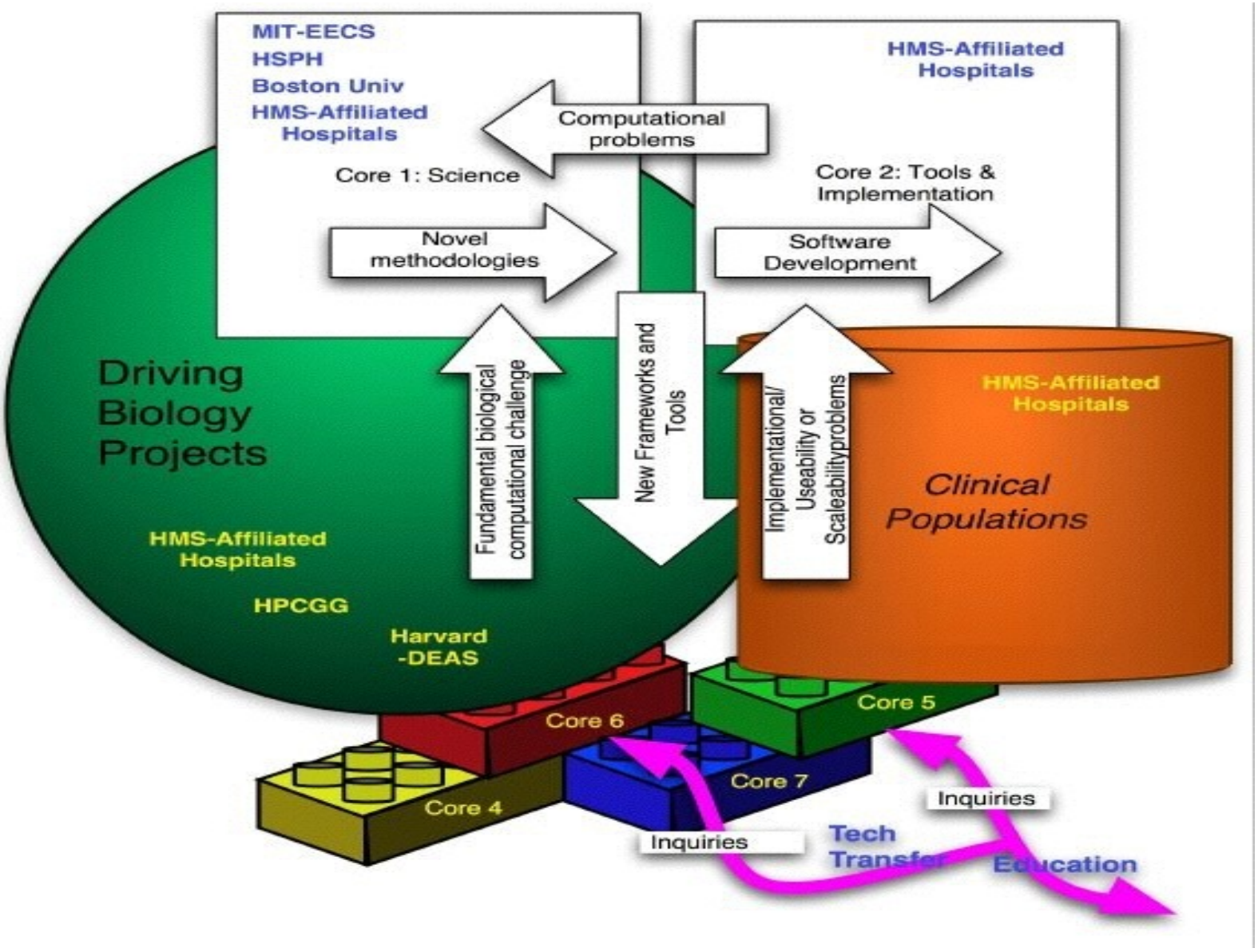
HARVARD MEDICAL SCHOOL

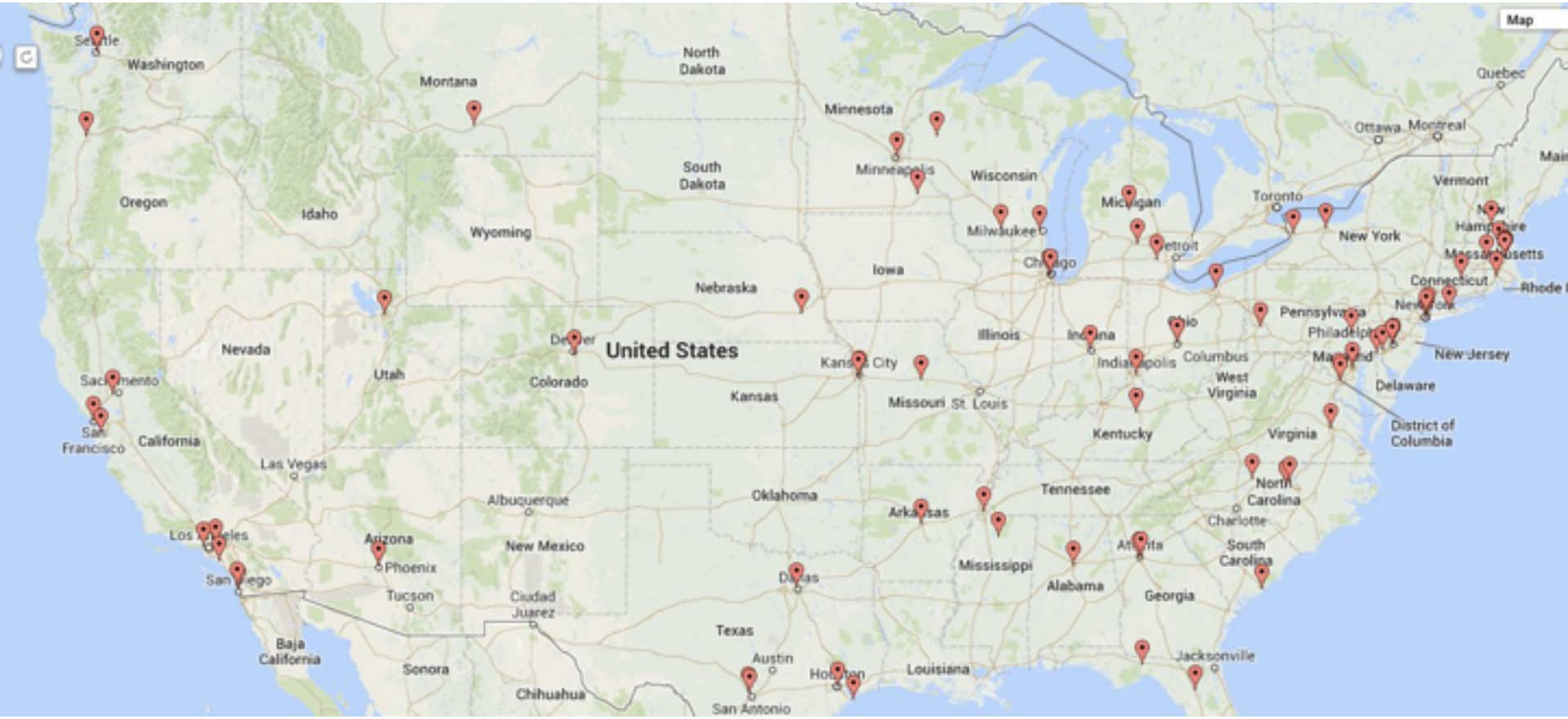
i2b2
Back to the Future

Isaac S. Kohane, MD, PhD

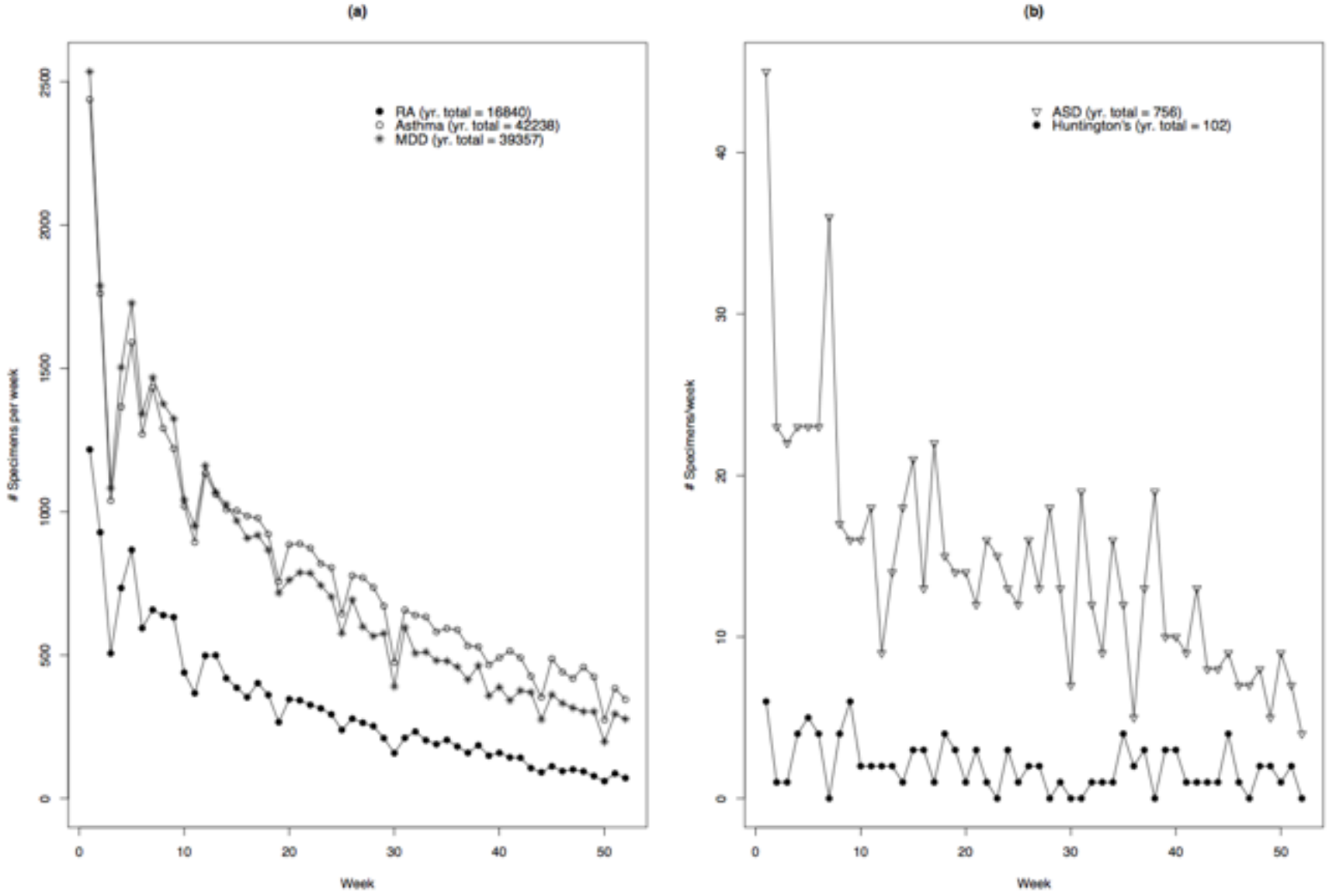
cbmi.med.harvard.edu

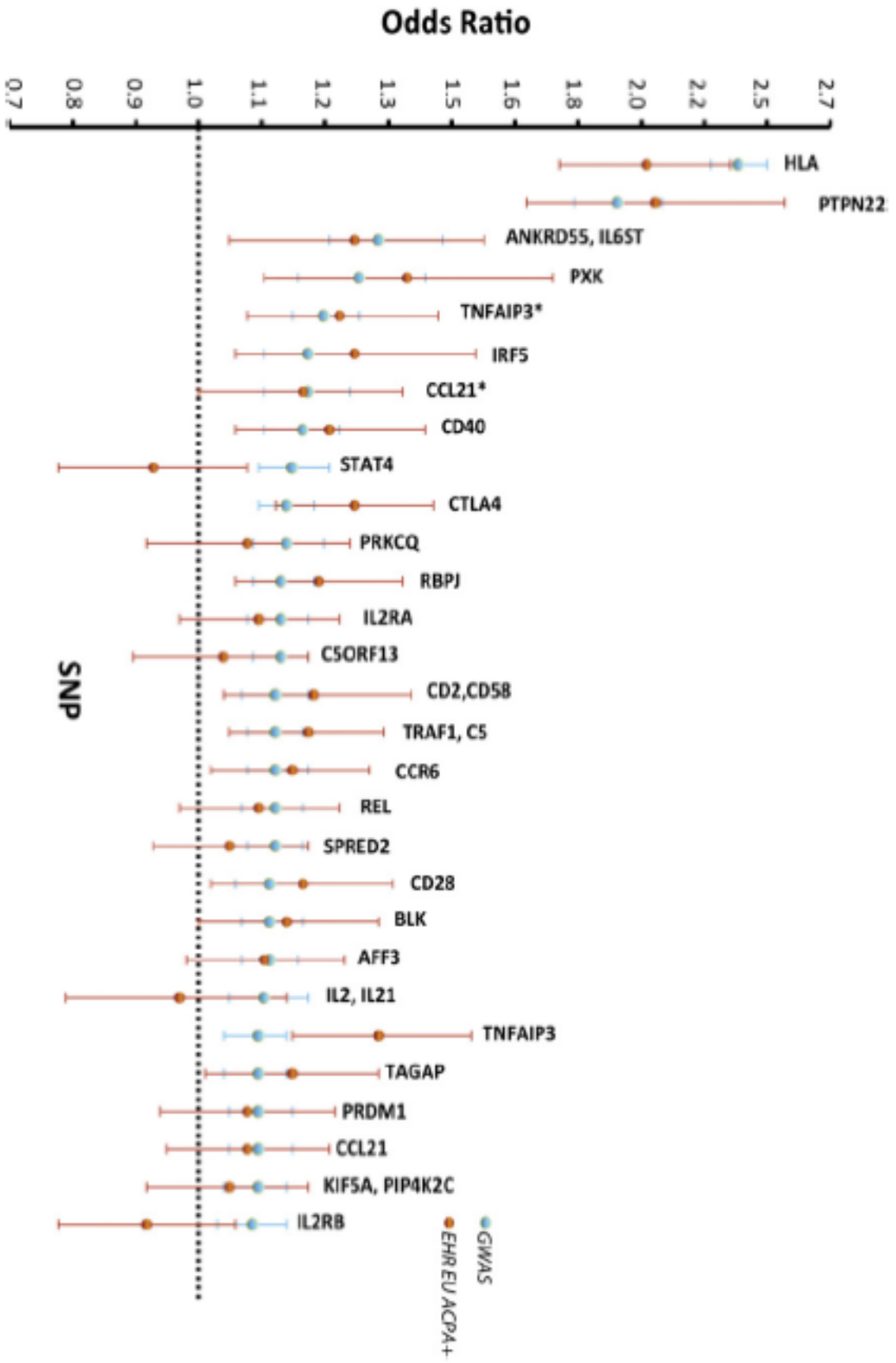
HARVARD MEDICAL SCHOOL


MIT-EECS
HSPH
Boston Univ
HMS-Affiliated Hospitals
Computational problems
Core 1: Science
HMS-Affiliated Hospitals
Core 2: Tools & Implementation
Novel methodologies
Software Development
Driving Biology Projects
Fundamental biological computational challenge
New Frameworks and Tools
Implementational/ Useability or Scaleabilityproblems
HMS-Affiliated Hospitals
Clinical Populations
HMS-Affiliated Hospitals
HPCGG
Harvard -DEAS
Core 6
Core 5
Core 7
Core 4
Inquiries
Inquiries
Tech Transfer
Education



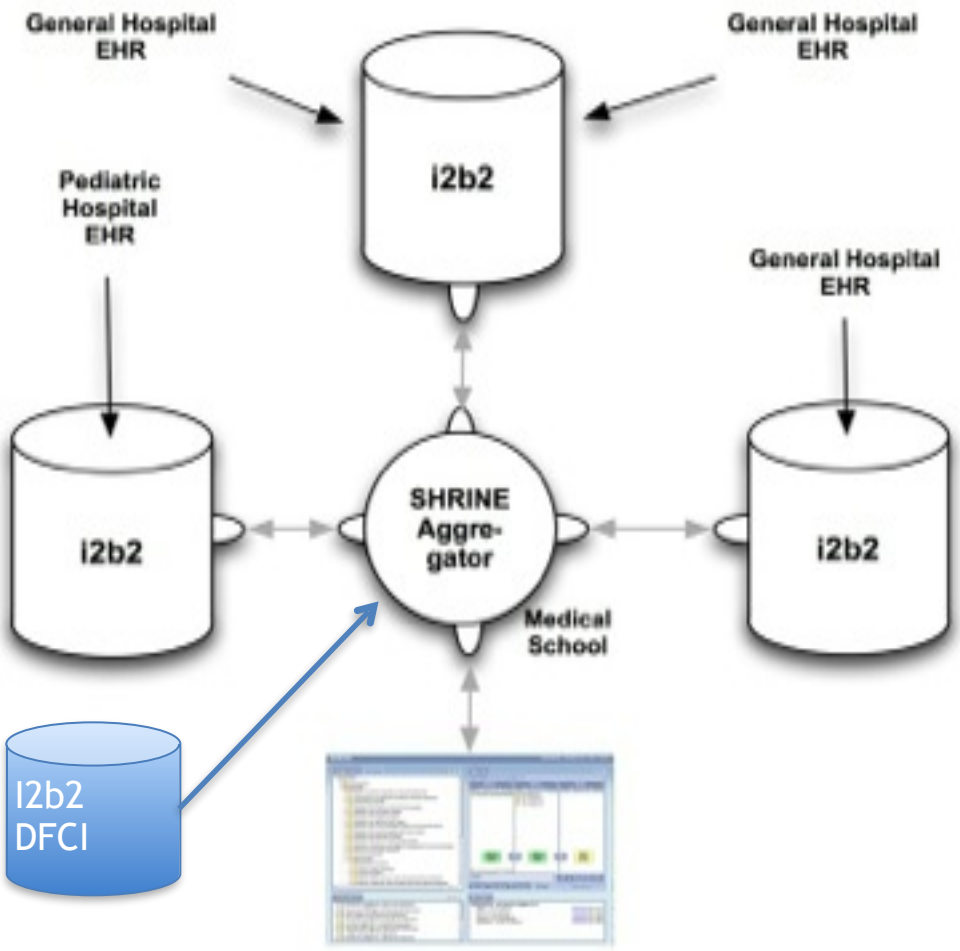
cbmi.med.harvard.edu


(a)
RA (yr. total = 16840)
Asthma (yr. total = 42238)
MDD (yr. total = 39357)
Specimens per week
Week
(b)
ASD (yr. total = 756)
Huntington's (yr. total = 102)
Specimens/week
Week


Odds Ratio
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.5
1.6
1.8
2.0
2.2
2.5
2.7
SNP
HLA
PTPN22
ANKRD55, IL6ST
PXK
TNFAIP3*
IRF5
CCL21*
CD40
STAT4
CTLA4
PRKCQ
RBPJ
IL2RA
C5ORF13
CD2,CD58
TRAF1, C5
CCR6
REL
SPRED2
CD28
BLK
AFF3
IL2, IL21
TNFAIP3
TAGAP
PRDM1
CCL21
KIF5A, PIP4K2C
IL2RB
EHR EU ACPA+
GWAS

High throughput phenotyping to find rare patients

- Search routine clinical records from 5 major hospitals for:
  - Demographics
  - Diagnosis
  - Medications
  - Lab Results
- Reach N
  - Rare Dx
  - Small Effects
- 10 billion FACTS
- 6 million patients



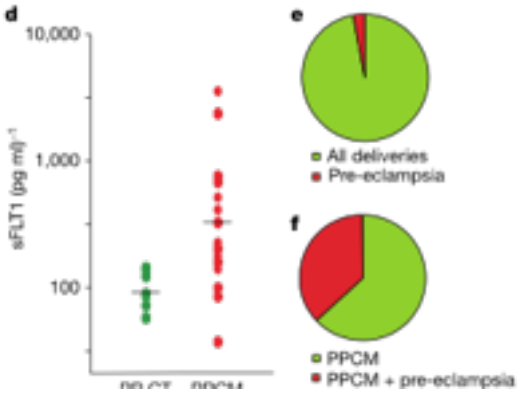
Finding rare events of interest.

Importance of real-time exploration

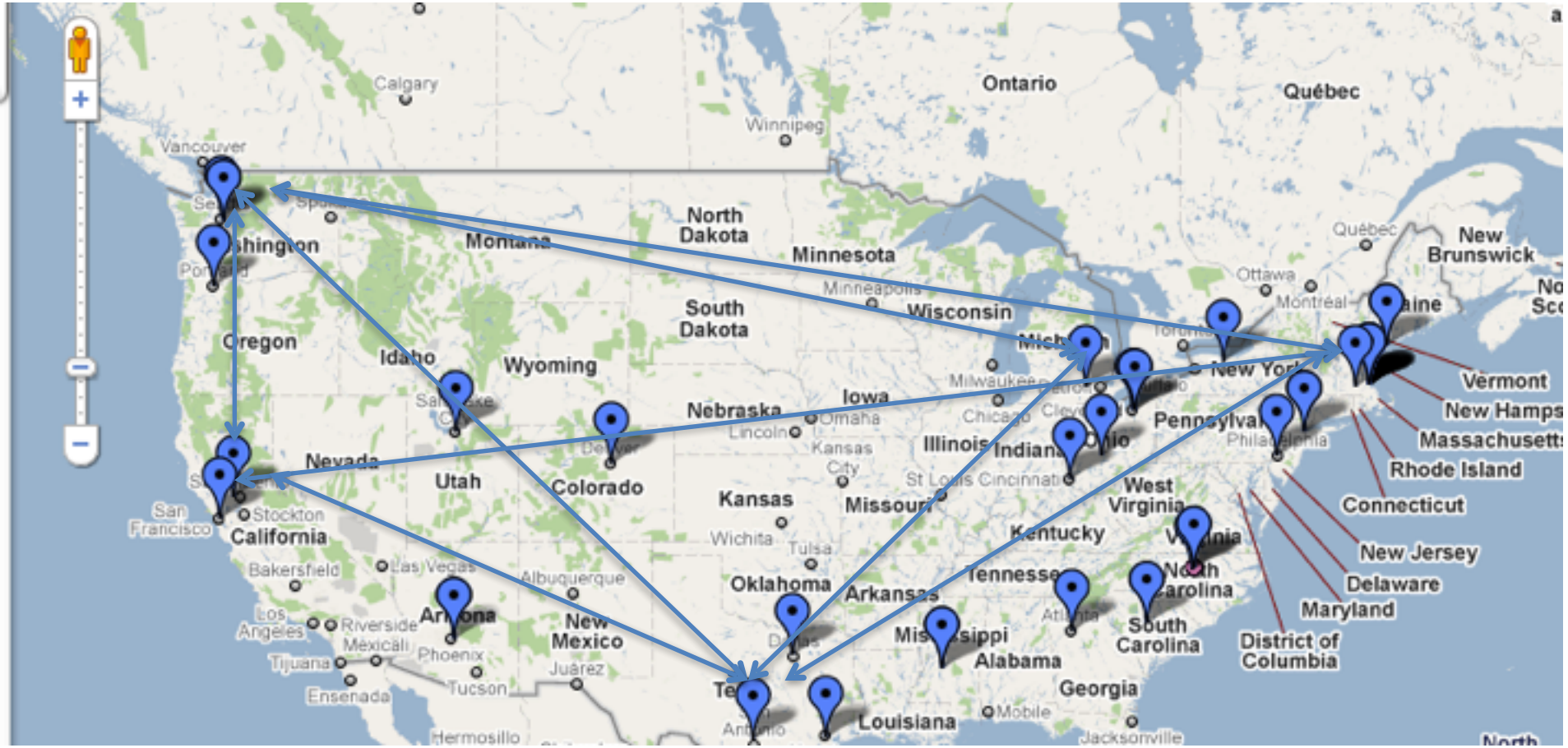
Cardiac angiogenic imbalance leads to peripartum cardiomyopathy

Ian S. Patten[1,7*], Sarosh Rana[1*], Sajid Shahul[1], Glenn C. Rowe[2], Chiboon Jang[3], Laura Liu[3], Michele R. Hacker[4], Julie S. Rhee[5], John Mitchell[4], Feroze Mahmood[6], Philip Hess[6], Caitlin Farrell[1], Nicole Koulbis[1], Eliyahu V. Khankin[1], Suzanne D. Burke[1,8], Igor Tudorache[6], Johann Bauersachs[7], Federica del Monte[2], Denise Hilfiker-Kleiner[7], S. Ananth Karumanchi[1,8] & Zoltan Arany

samples from subjects with PPCM have been previously described[5]. Patients in both studies were predominantly Caucasian. Retrospective analyses of PPCM and pre-eclampsia in the Harvard teaching hospitals were performed using the Harvard Shared Health Research Information Network (SHRINE)[53], a de-identified repository of aggregate patient information.

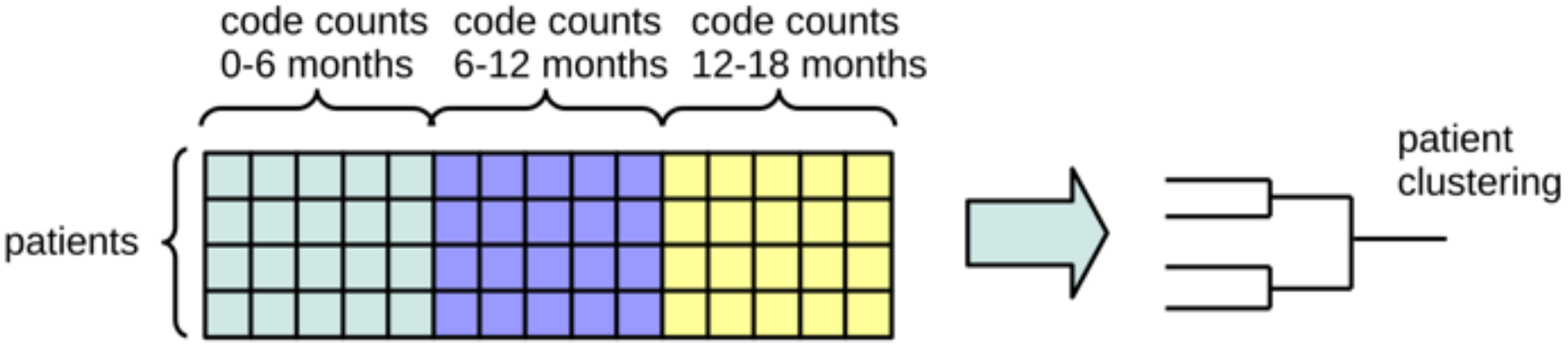


2012

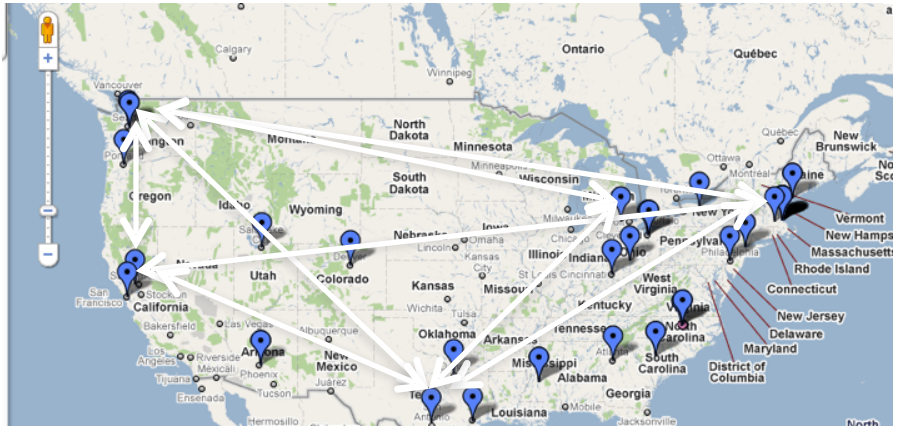


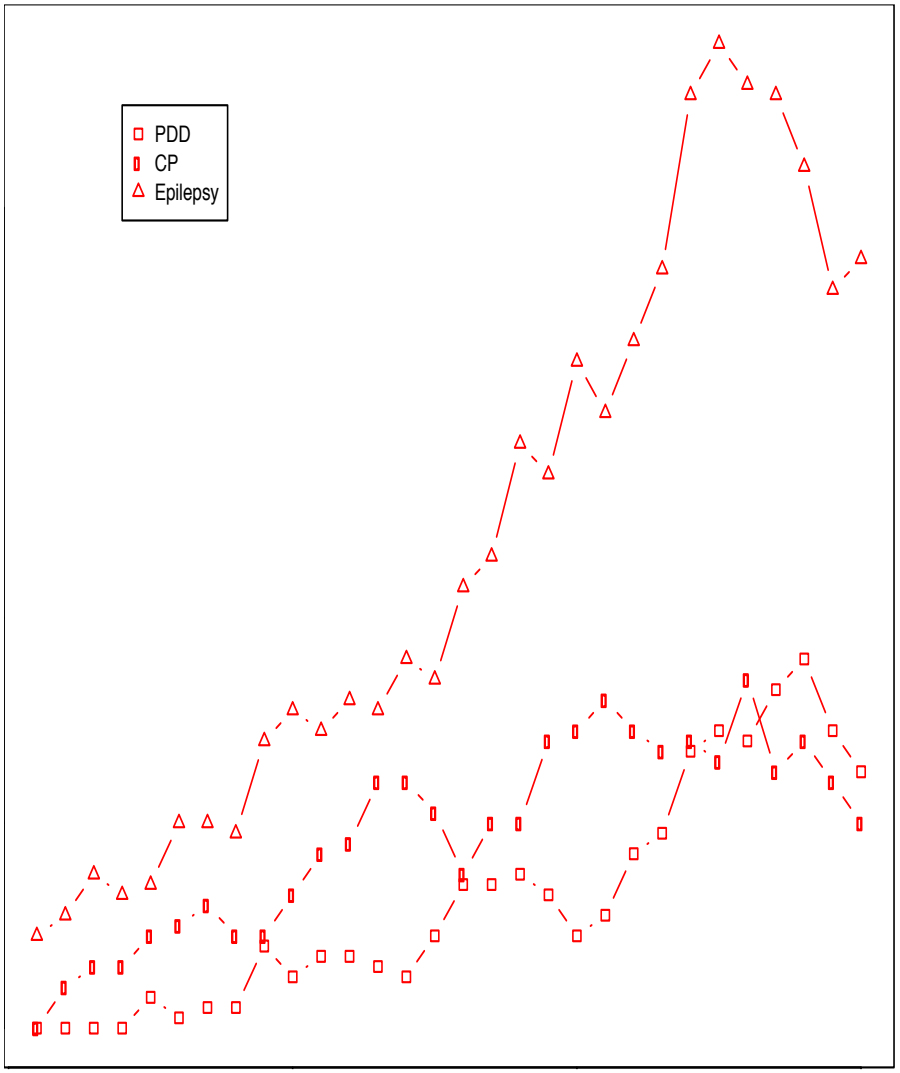
Calgary
Ontario
Québec
Winnipeg
Vancouver
North Dakota
Minnesota
Montana
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Nebraska
Iowa
Nevada
Utah
Colorado
Kansas
Missouri
Illinois
Indiana
Ohio
Pennsylvania
New York
Vermont
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
West Virginia
Kentucky
Tennessee
North Carolina
South Carolina
Georgia
Alabama
Mississippi
Arkansas
Oklahoma
Louisiana
New Mexico
Arizona
California
San Francisco
Stockton
Bakersfield
Las Vegas
Los Angeles
Riverside
Mexicali
Tijuana
Ensenada
Phoenix
Tucson
Albuquerque
Juárez
Hermosillo
Wichita
Tulsa
Omaha
Lincoln
Kansas City
St Louis
Cincinnati
Chicago
Milwaukee
Minneapolis
Ottawa
Montréal
Québec
Philadelphia
Mobile
Jacksonville

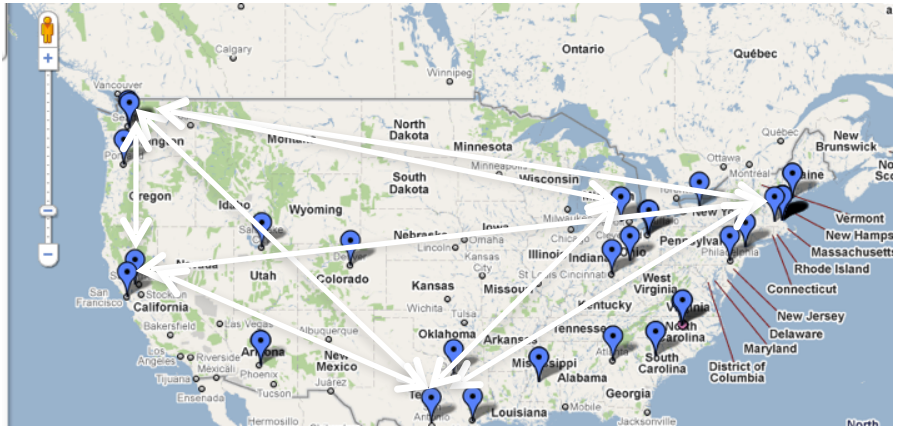
Unbiased clustering

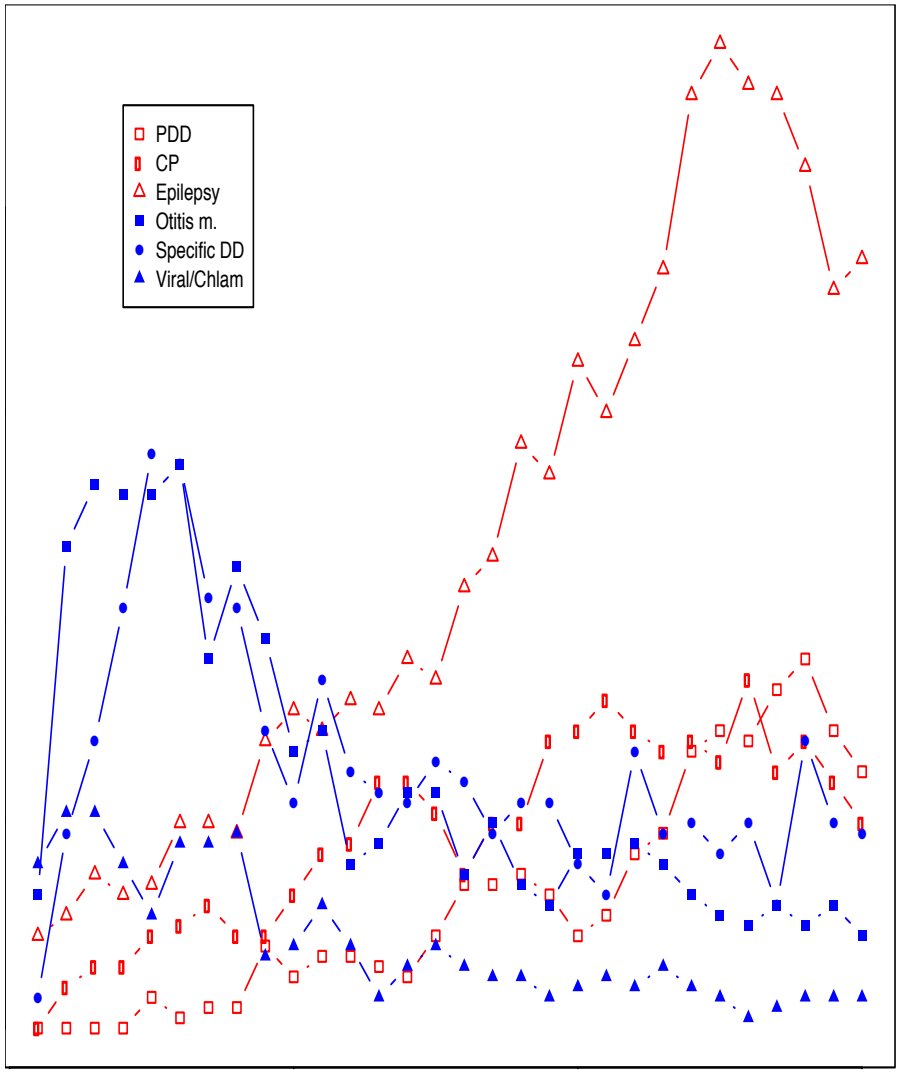


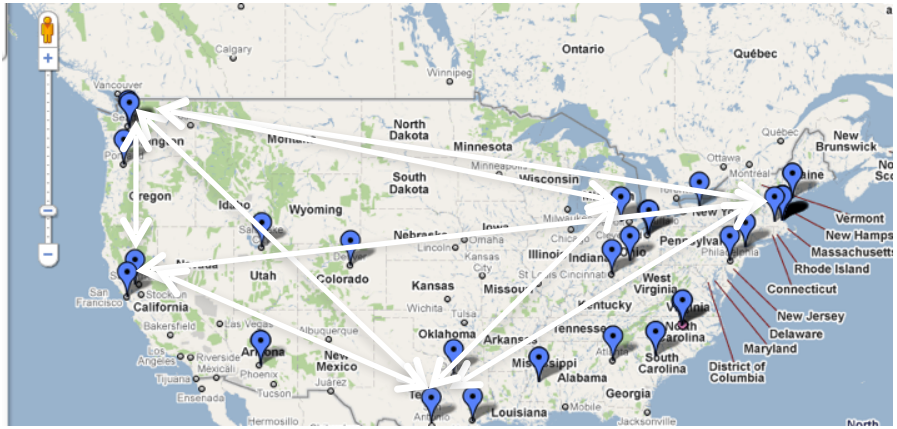
Autism or Autisms?

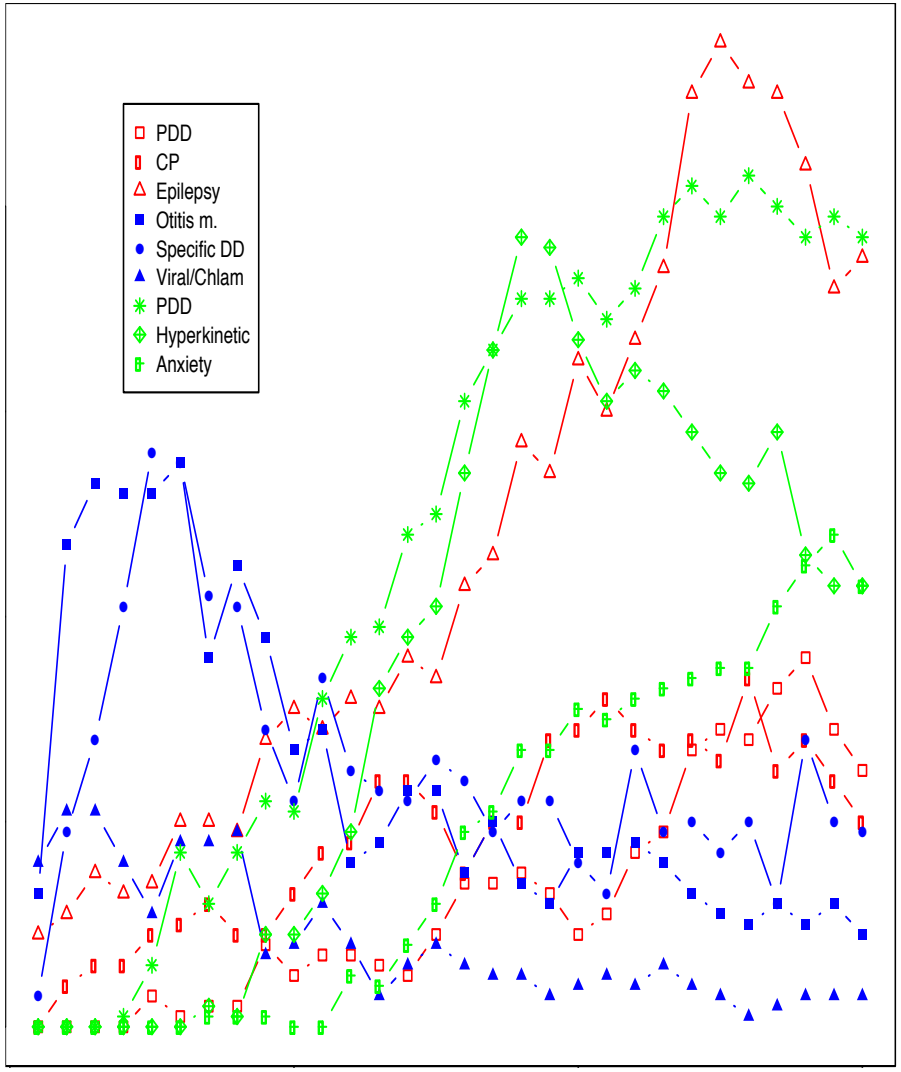




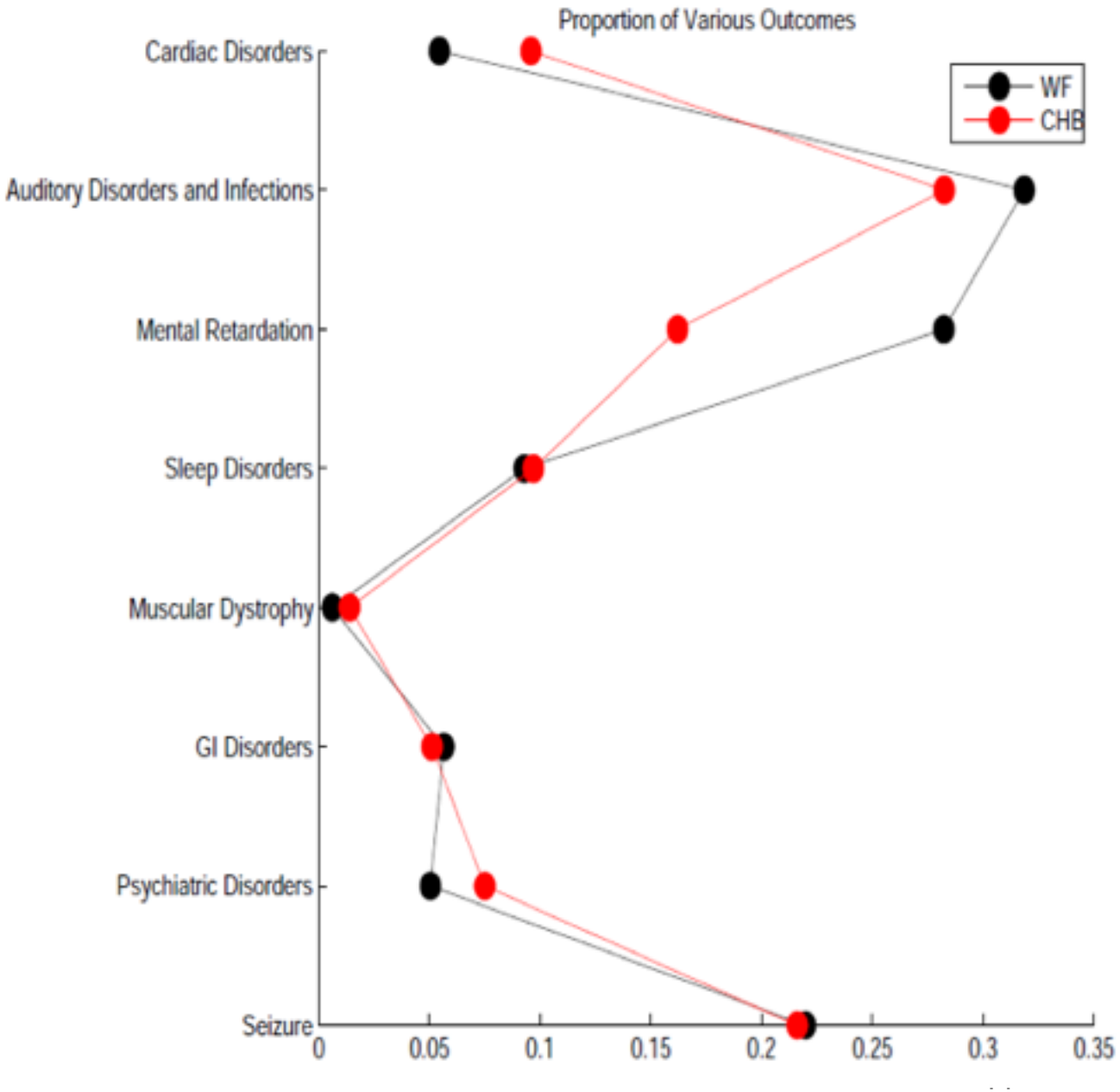

Ontario
Québec
Calgary
Winnipeg
Vancouver
North Dakota
Minnesota
South Dakota
Wisconsin
Wyoming
Utah
Colorado
Kansas
Oklahoma
Arkansas
Louisiana
Georgia
Alabama
South Carolina
West Virginia
Vermont
New Hamps
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
New Brunswick
California
Arizona
New Mexico
Phoenix
Tucson
Ottawa
Montréal
Québec
Omaha
Lincoln
Kansas City
Wichita
Tulsa
Albuquerque
Las Vegas
Bakersfield
Los Angeles
Riverside
Mexicali
Tijuana
Ensenada
Juárez
Hermosillo
Mobile
Jacksonville
Philadelphia
Cincinnati
St Louis
Chicago
Milwaukee
Minneapolis
San Francisco
Stockton


PDD
CP
Epilepsy
Otitis m.
Specific DD
Viral/Chlam

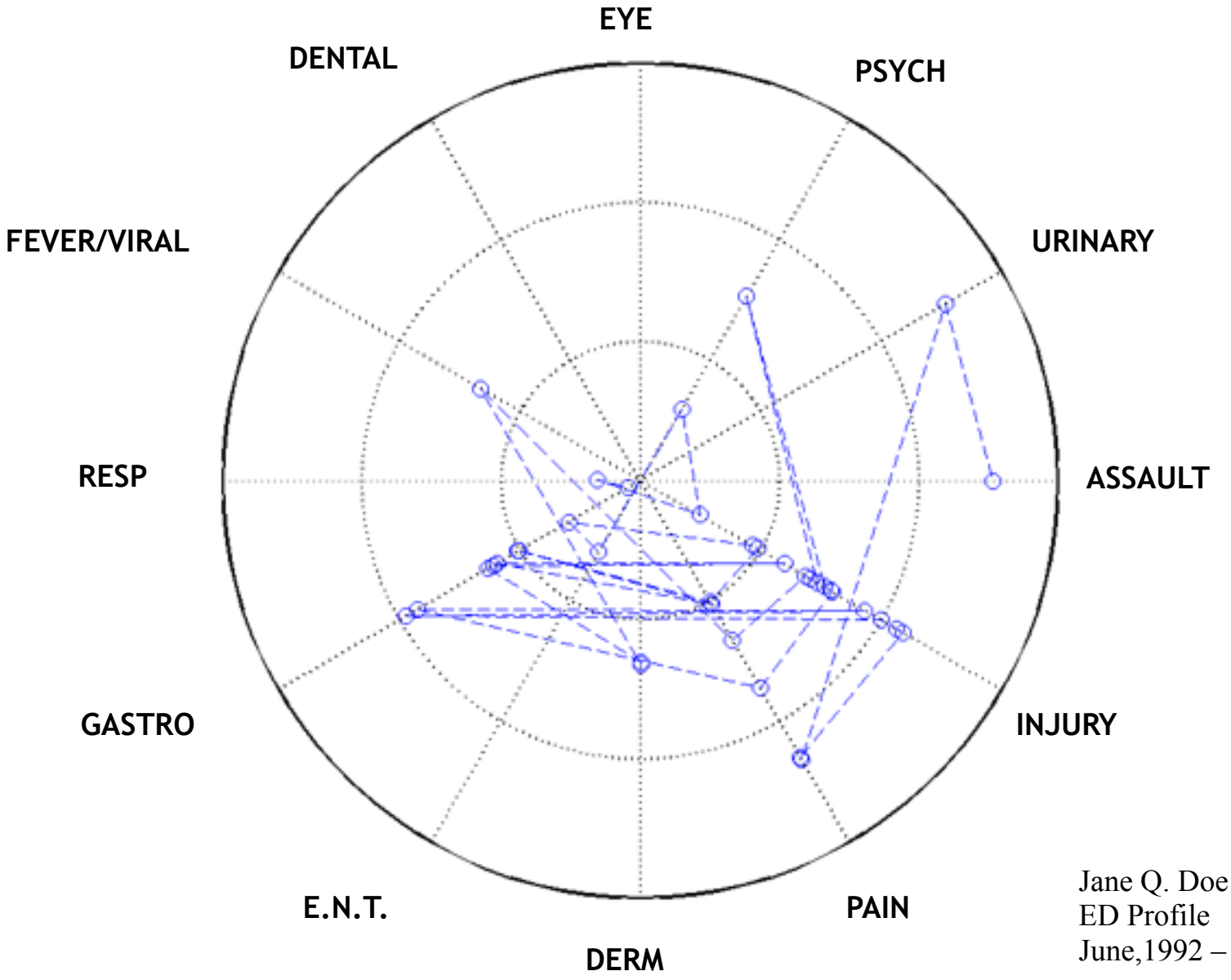

Ontario
Québec
Calgary
Winnipeg
Vancouver
North Dakota
Minnesota
South Dakota
Wisconsin
Wyoming
Utah
Colorado
Kansas
Oklahoma
Arkansas
Louisiana
Georgia
Alabama
South Carolina
Maryland
Delaware
New Jersey
Connecticut
Rhode Island
Massachusetts
Vermont
New Brunswick
Ottawa
Montréal
Québec
Kentucky
West Virginia
Illinois
Indiana
Ohio
Pennsylvania
Wichita
Tulsa
Omaha
Lincoln
Kansas City
Albuquerque
Las Vegas
Los Angeles
Riverside
Phoenix
Tucson
Juárez
Tijuana
Ensenada
Hermosillo
Mobile
Jacksonville
Bakersfield
San Francisco
Stockton
California
Arizona
New Mexico
Mississippi
Tennessee
North Carolina
District of Columbia


PDD
CP
Epilepsy
Otitis m.
Specific DD
Viral/Chlam
PDD
Hyperkinetic
Anxiety

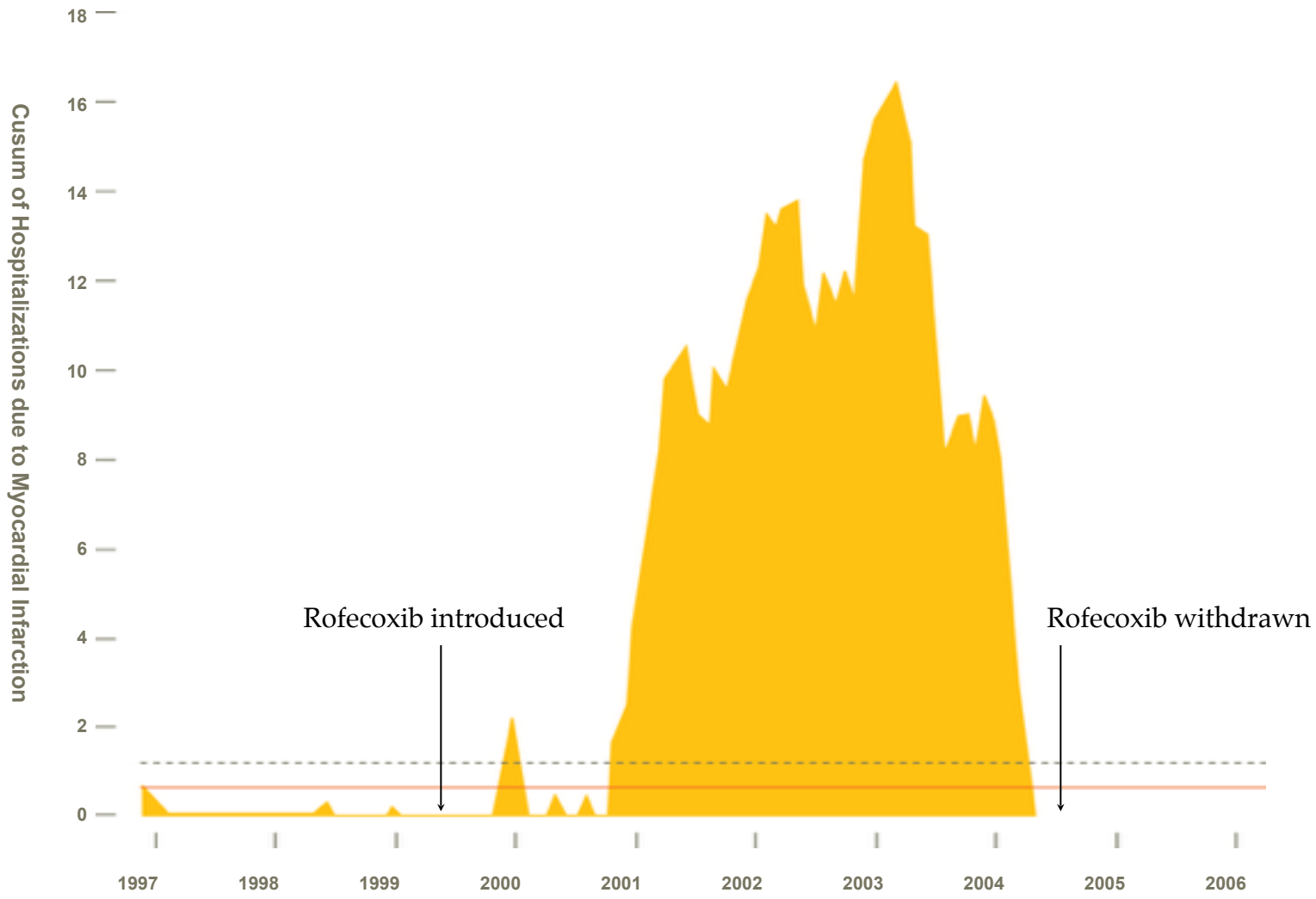
Reproducibility



Reis, et al., BMJ 2009

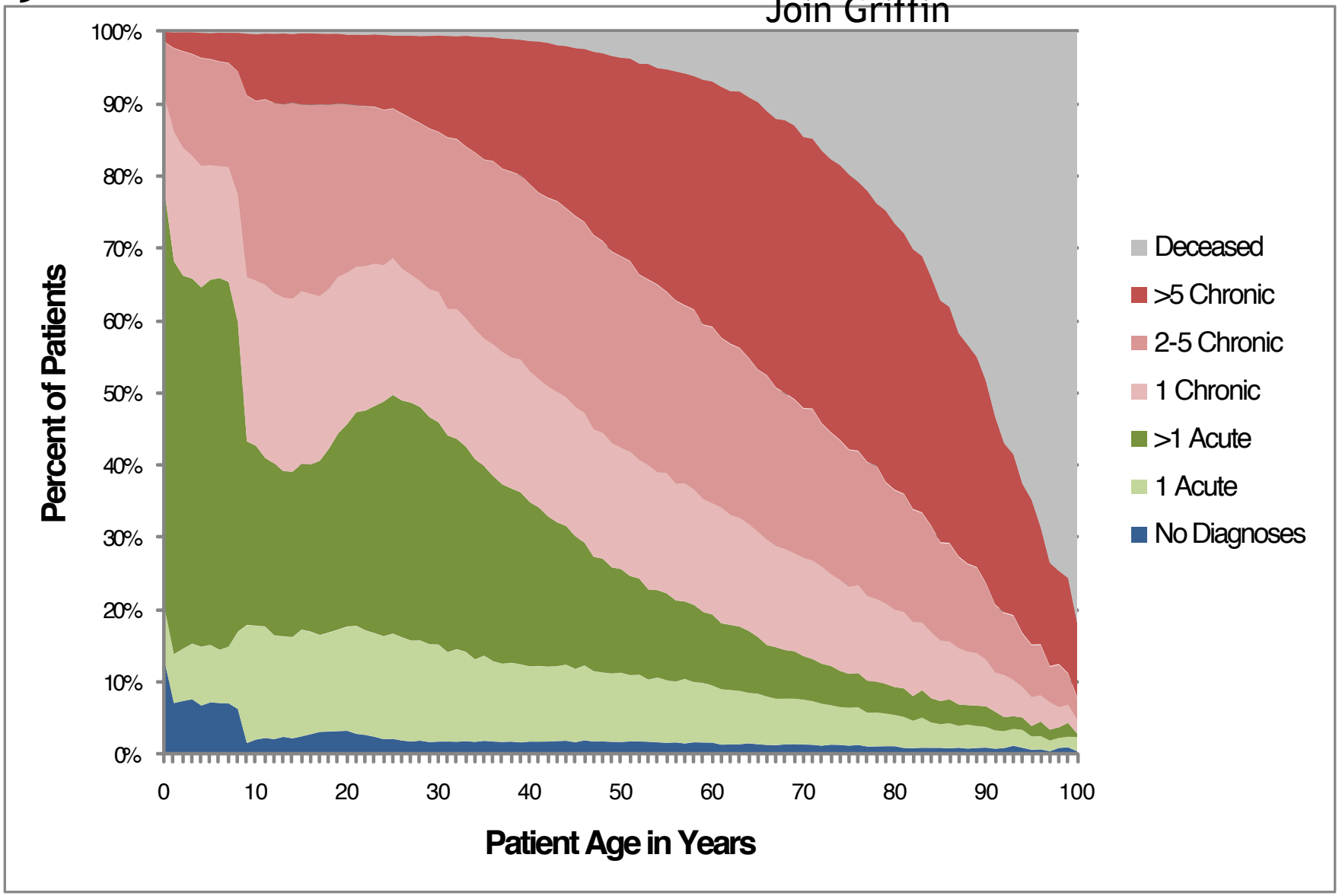


Jane Q. Doe
ED Profile
June,1992 – Feb,1999

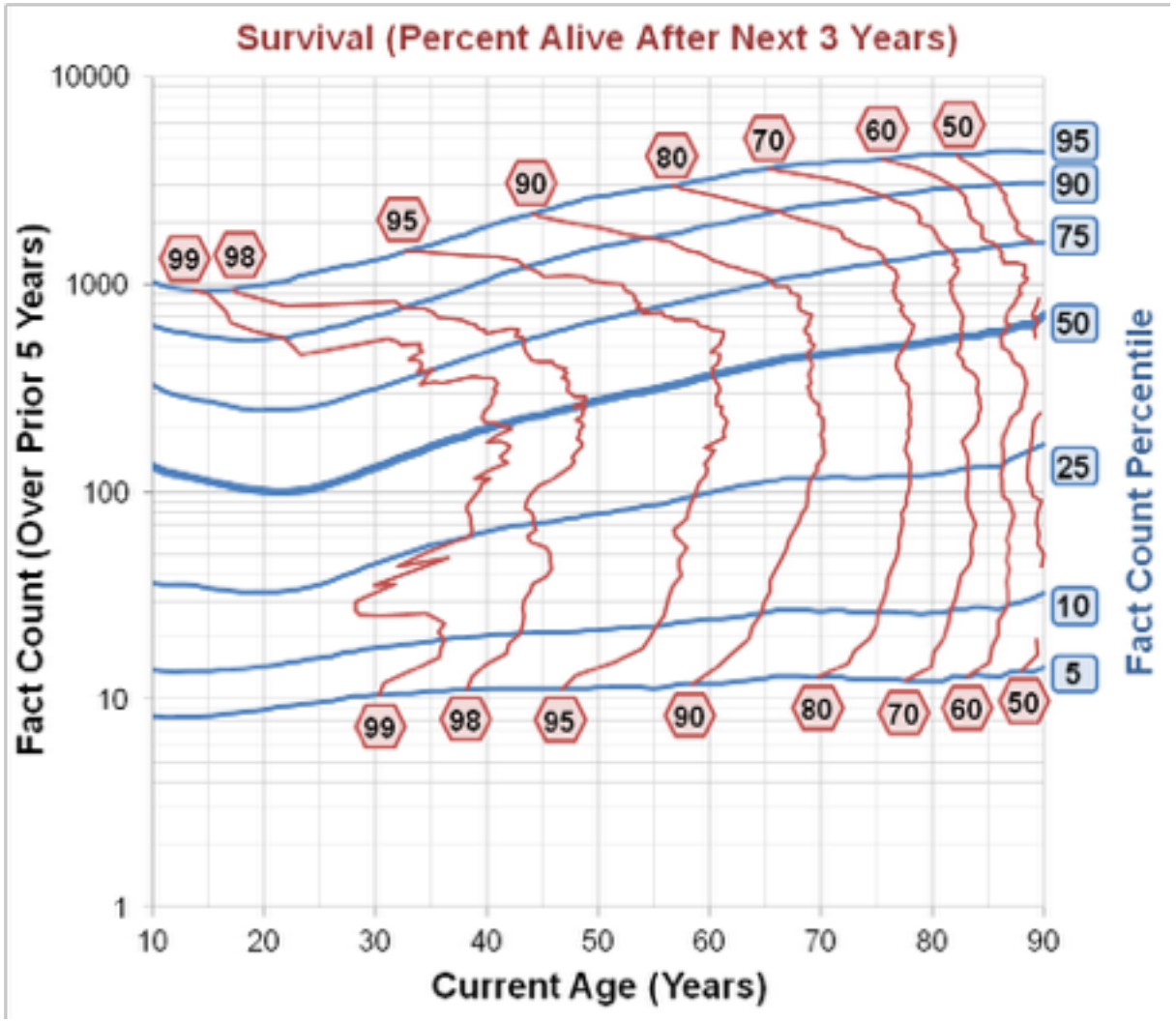


Brownstein, PLoS One, 2007

Care Behavior + Health => Healthcare Dynamics

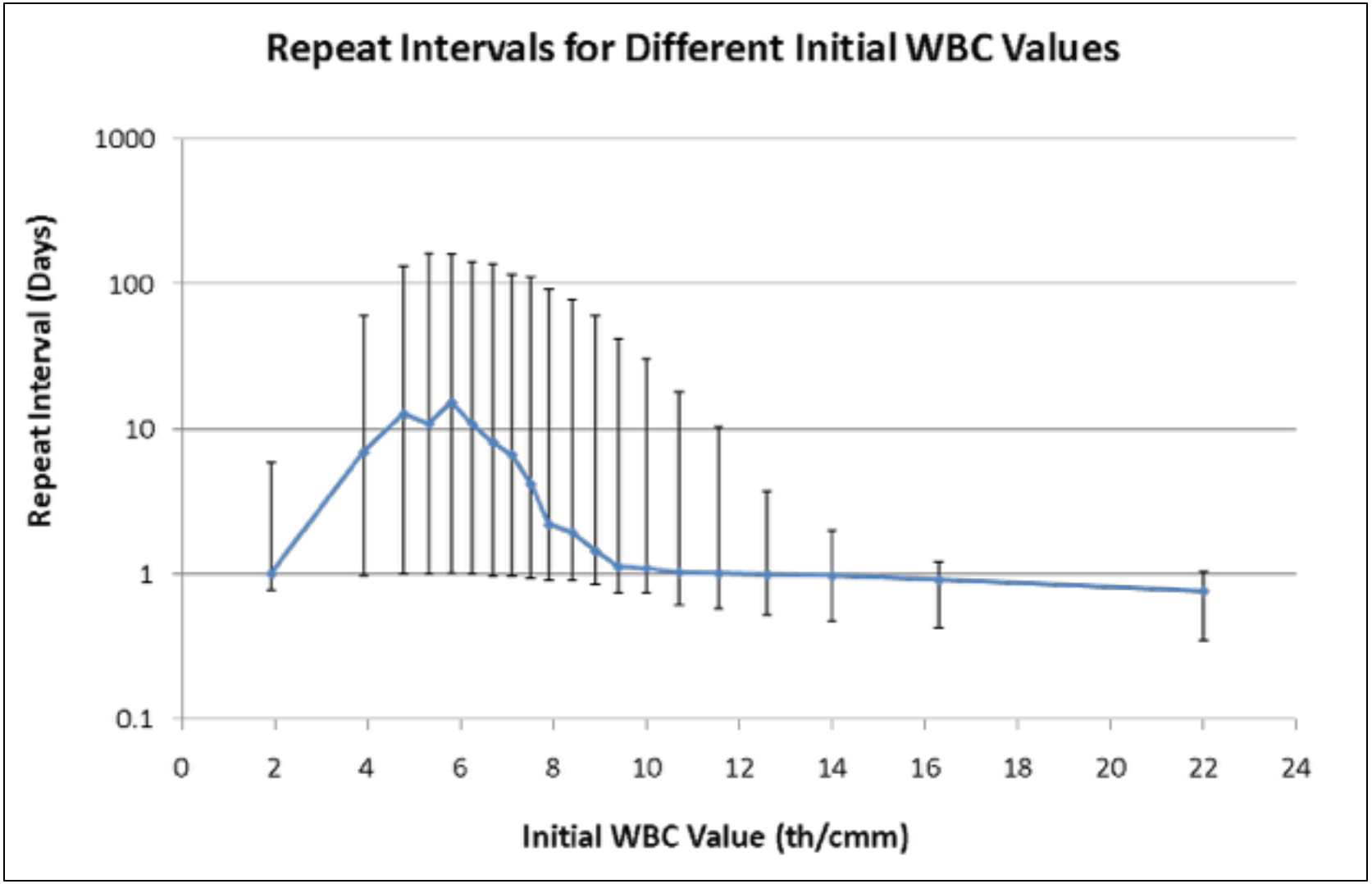


Fact Count Growth & Survival Chart (Male)

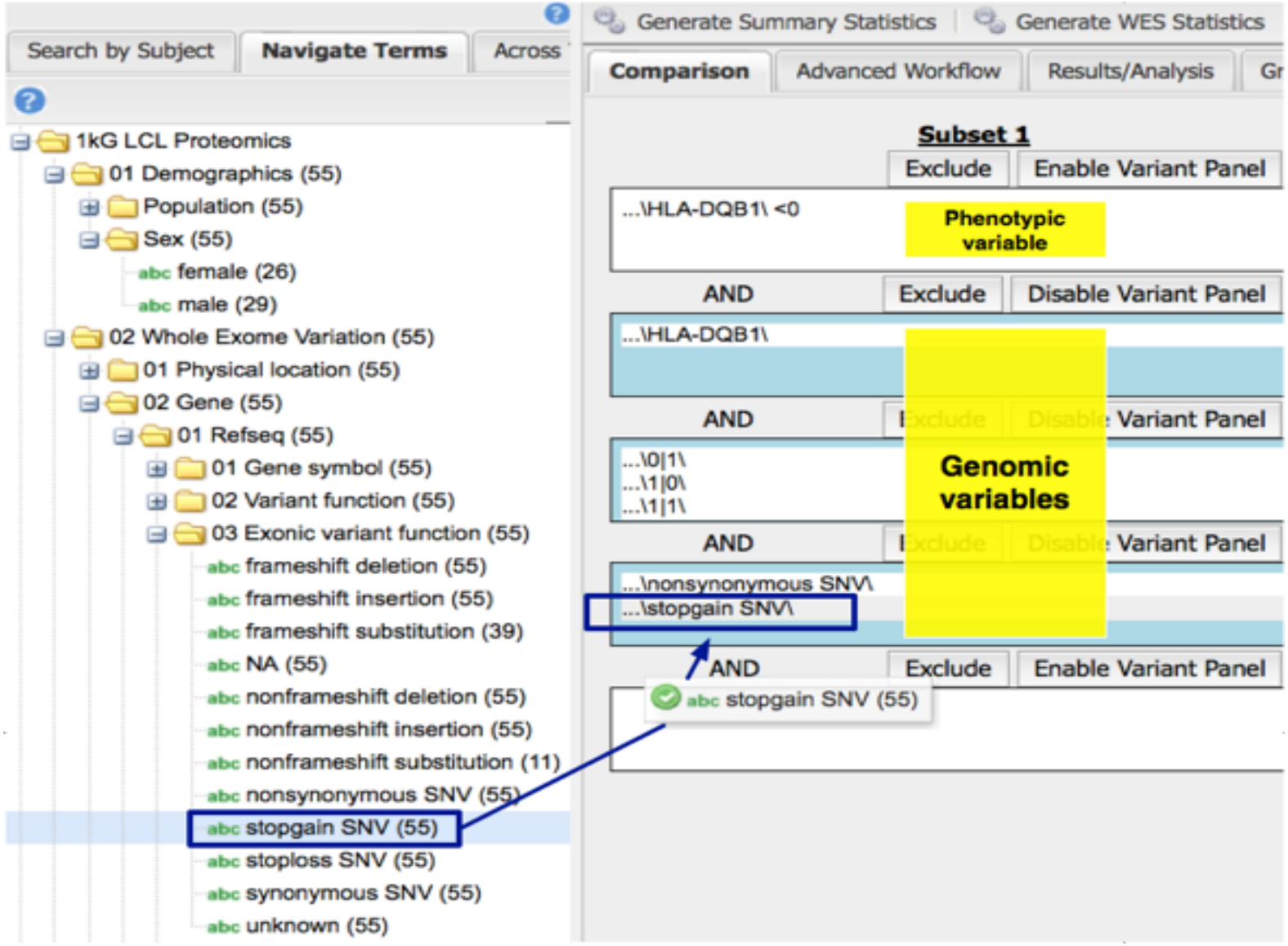


Horizontal blue curves with rectangular labels are five year fact count percentiles. Vertical red curves with hexagonal labels are three year survival curves. Data are from BWH and MGH from 7/28/2001 to 7/27/2009. All patients had at least one fact between 7/28/2005 and 7/27/2006. The “current” age is the patient age on 7/27/2006. This chart represents only male patients.

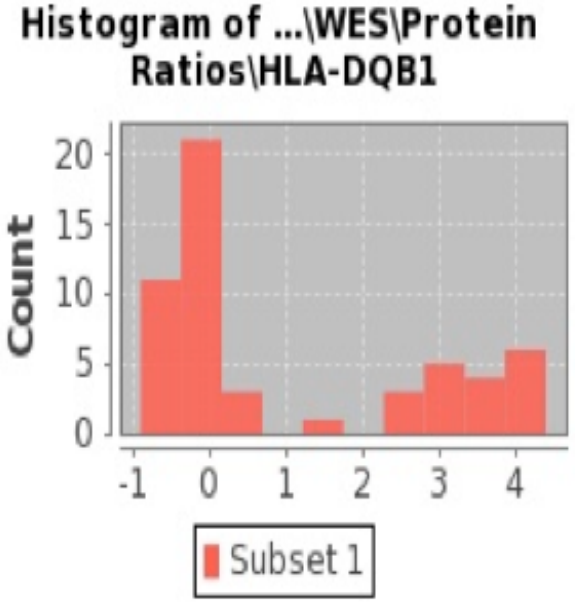
Normal Physician Behavior

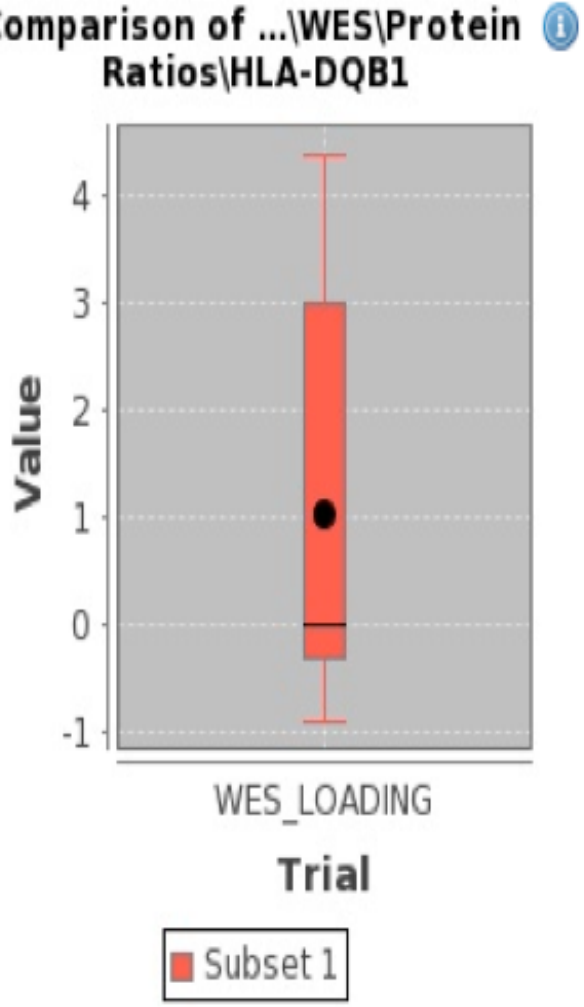


Weber GM, Kohane IS. Extracting physician group intelligence from electronic health records to support evidence based medicine. PLoS One. 2013 May 29;8(5):e64933.


Search by Subject
Navigate Terms
Across
1kG LCL Proteomics
01 Demographics (55)
Population (55)
Sex (55)
female (26)
male (29)
02 Whole Exome Variation (55)
01 Physical location (55)
02 Gene (55)
01 Refseq (55)
01 Gene symbol (55)
02 Variant function (55)
03 Exonic variant function (55)
frameshift deletion (55)
frameshift insertion (55)
frameshift substitution (39)
NA (55)
nonframeshift deletion (55)
nonframeshift insertion (55)
nonframeshift substitution (11)
nonsynonymous SNV (55)
stopgain SNV (55)
stoploss SNV (55)
synonymous SNV (55)
unknown (55)
Generate Summary Statistics
Generate WES Statistics
Comparison
Advanced Workflow
Results/Analysis
Subset 1
Exclude
Enable Variant Panel
...\HLA-DQB1\ <0
Phenotypic variable
AND
Exclude
Disable Variant Panel
...\HLA-DQB1\
...\0|1\
...\1|0\
...\1|1\
Genomic variables
...\nonsynonymous SNV\
...\stopgain SNV\
stopgain SNV (55)

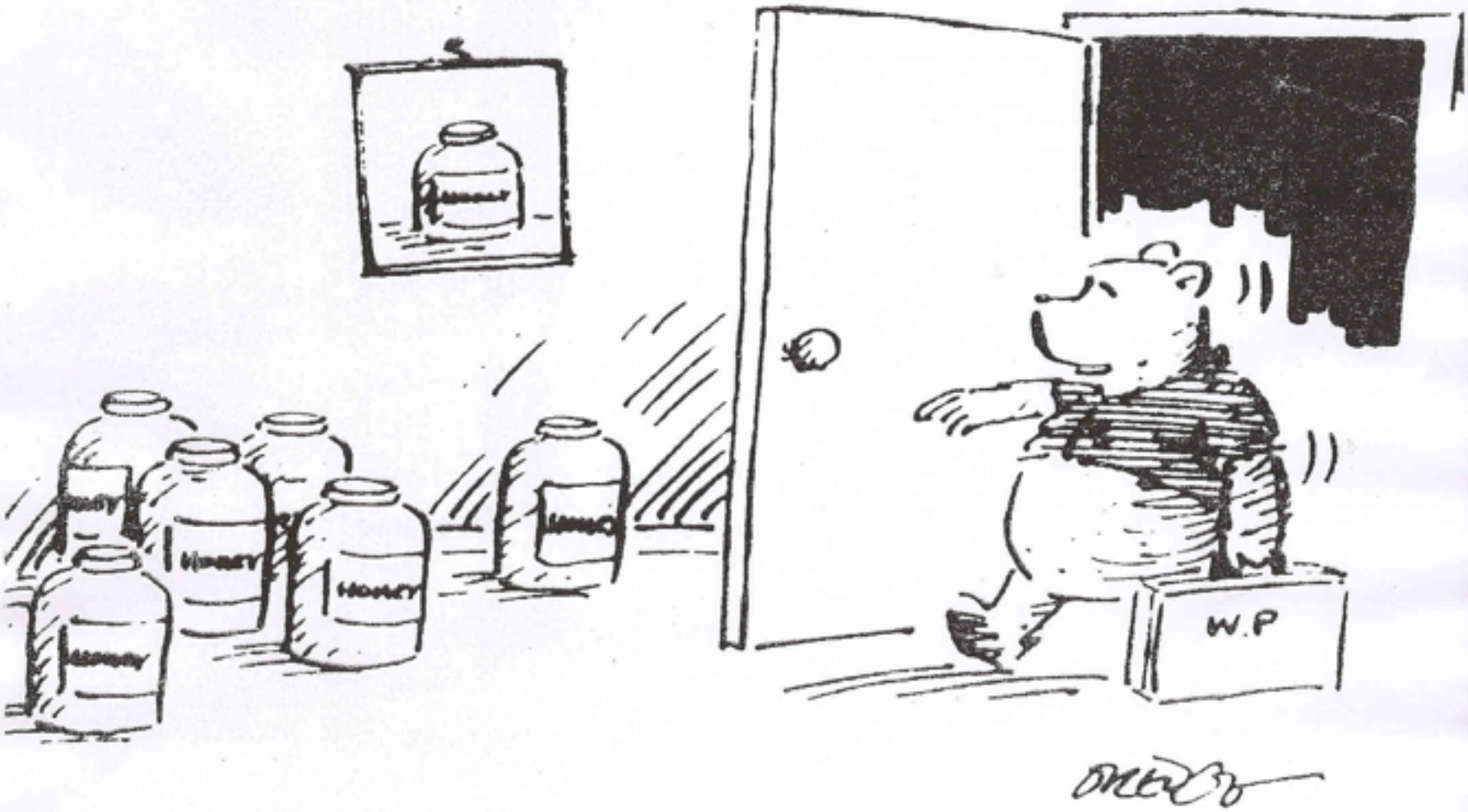
Analysis of ...\WES\Protein Ratios\HLA-DQB1 for subsets:





Subset 1

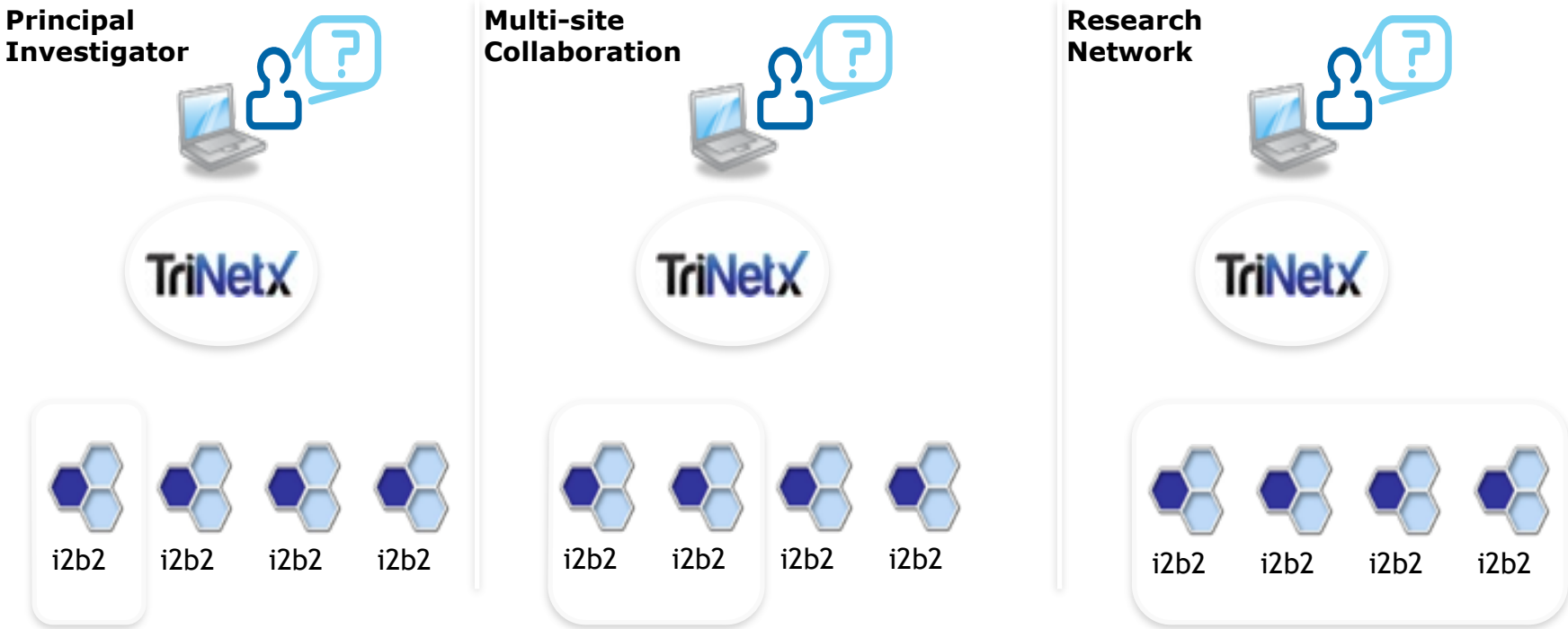
WES_LOADING
Mean: 1.05
Median: -0.01
IQR: 3.33
SD: 1.81
Data Points: 54



"Hi, honey – I'm home!"

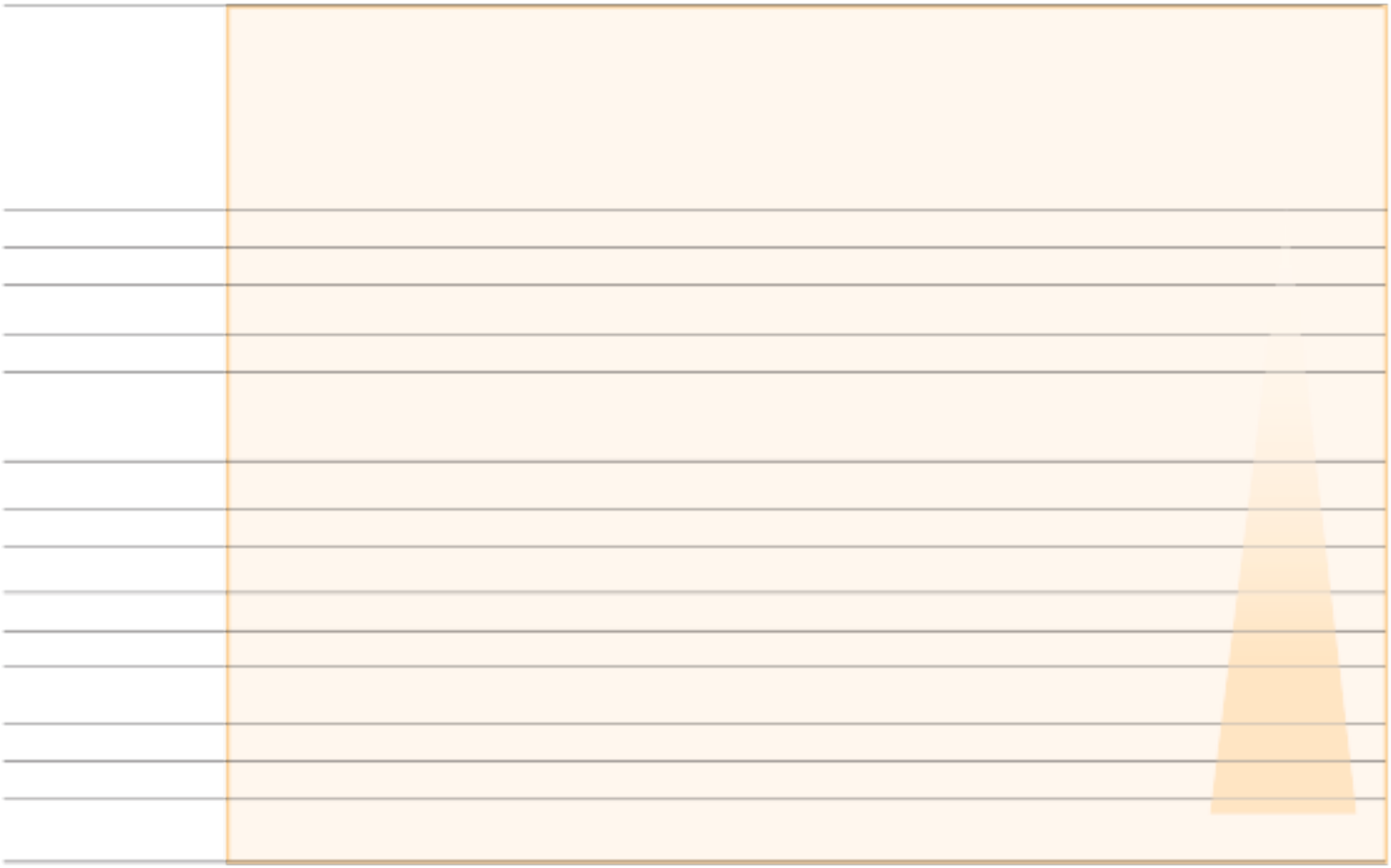
TriNetX Buids off of i2b2

- Supports local data access and formation of ad hoc research networks
- Facilitates translational research for internal and/or sponsored studies
- Contributes to sustainability of your research enterprise



Where To?

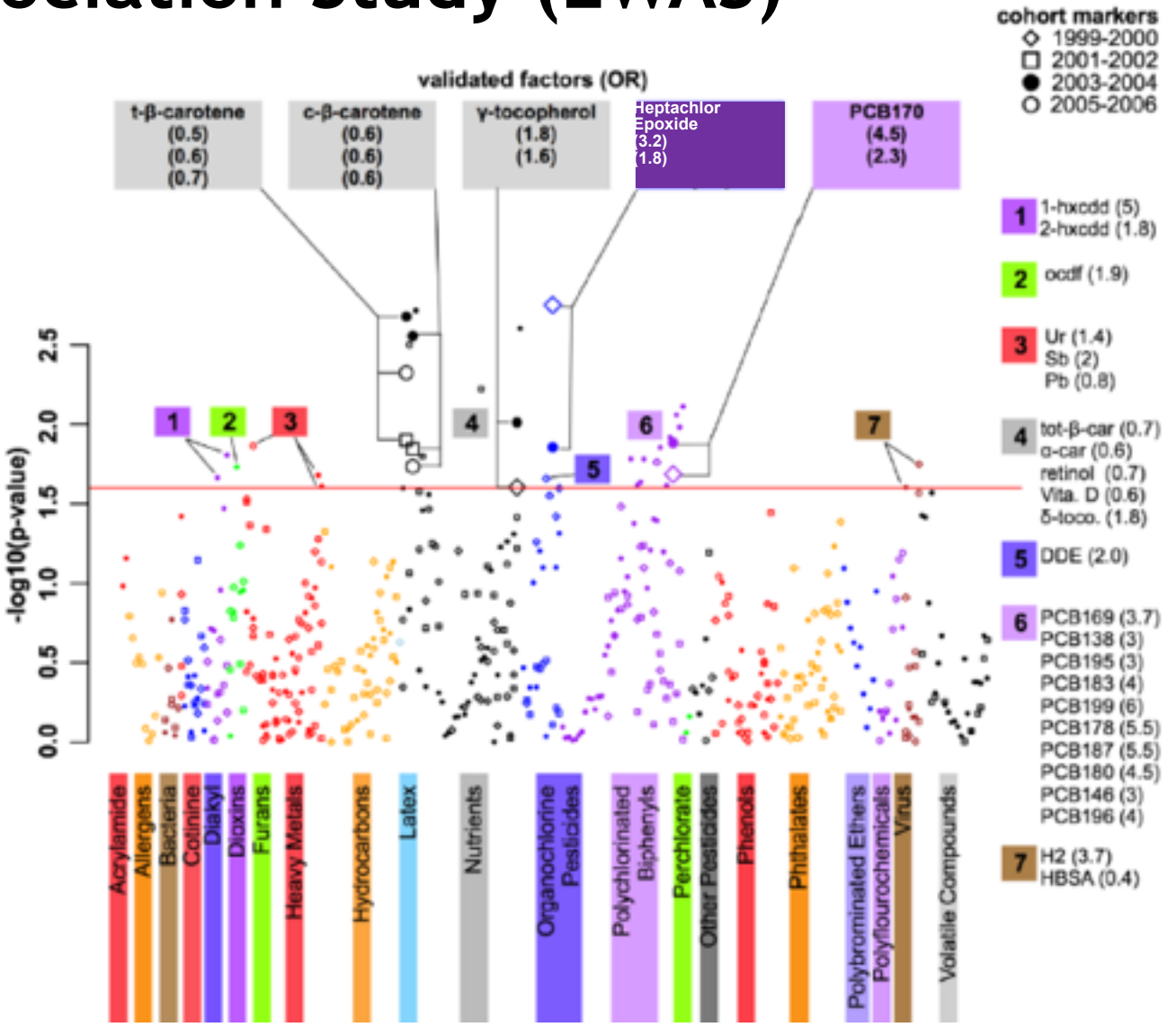
Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



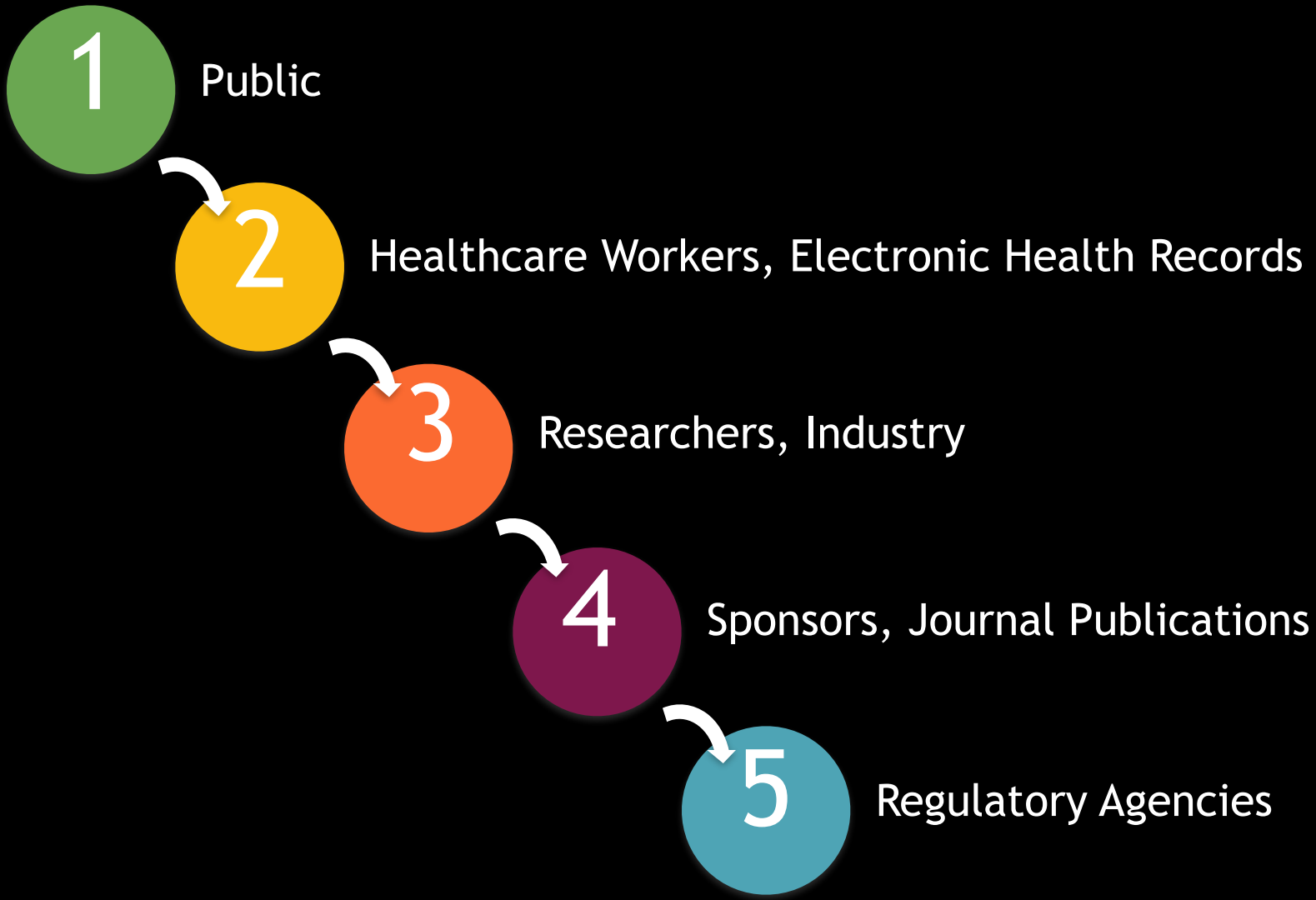
Environment-Wide Association Study (EWAS)

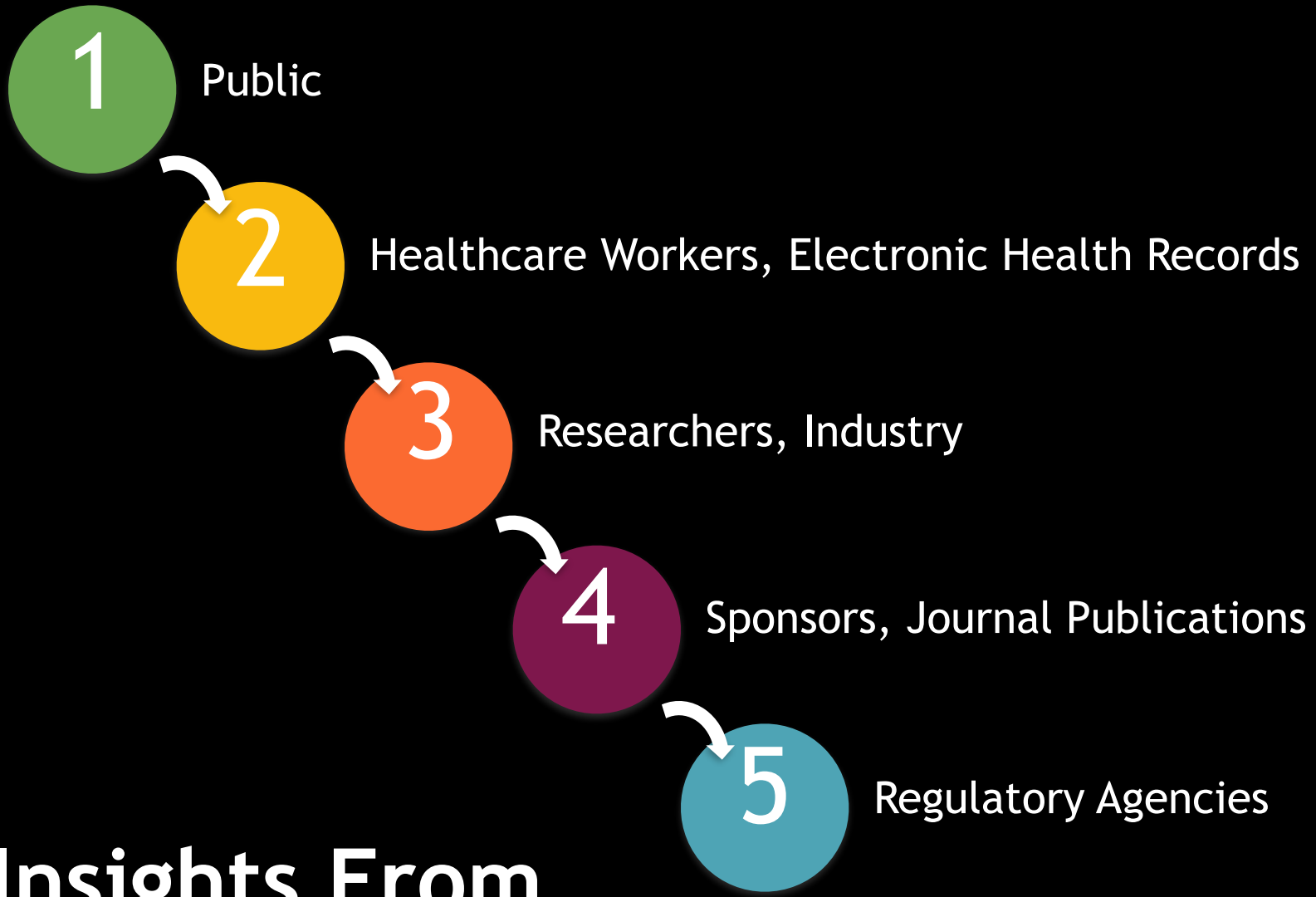
Chirag Patel

- New associations in 2+ cohorts: **gamma tocopherol, hepatochlor epoxide, PCB**
- Known associations: **vitamin D, beta carotene**
- Interesting: **hepatitis B**

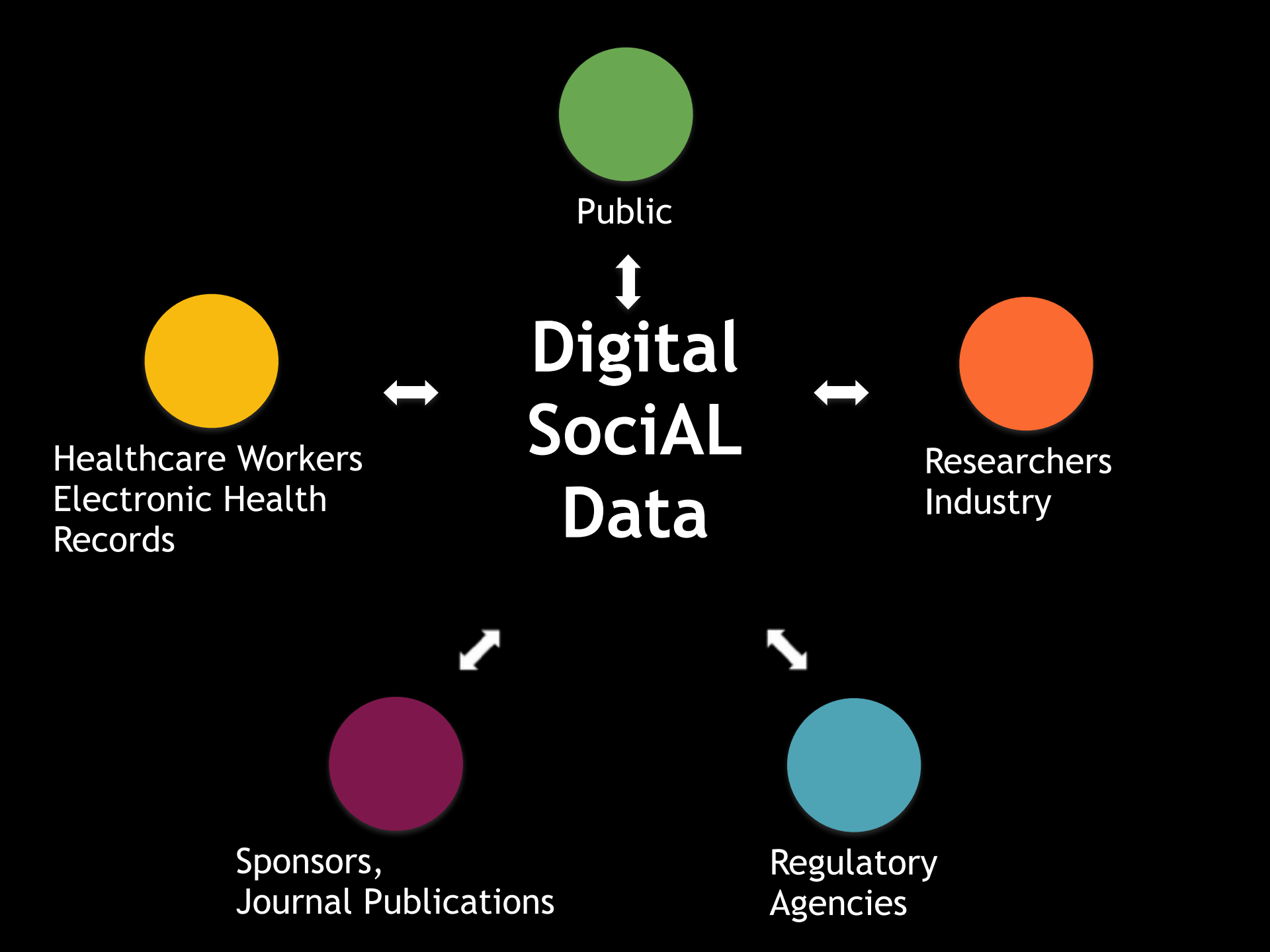


Patel C, Bhattacharya J, Butte AJ. *PLoS One*, 2010.

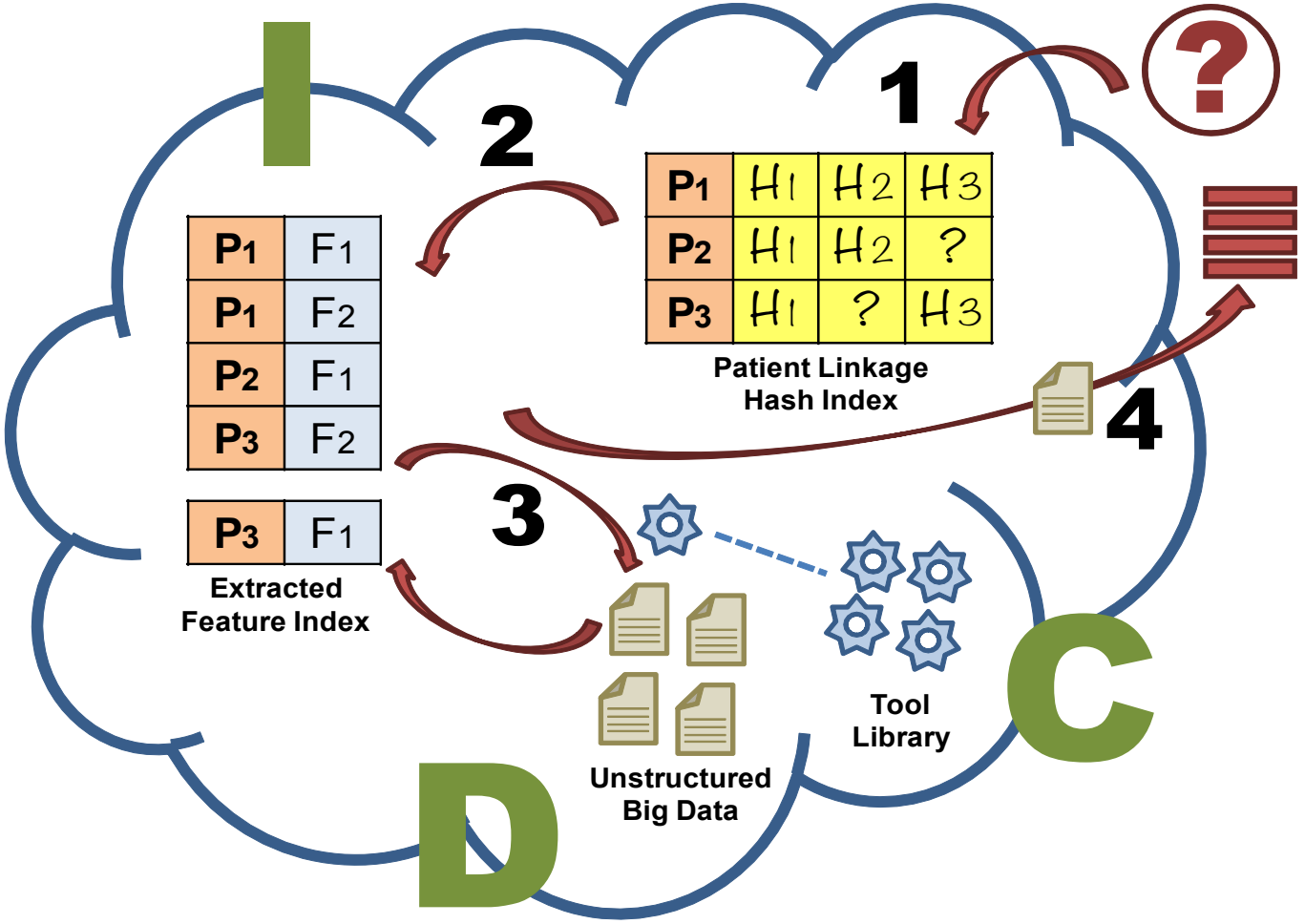

1
Public
2
Healthcare Workers, Electronic Health Records
3
Researchers, Industry
4
Sponsors, Journal Publications
5
Regulatory Agencies



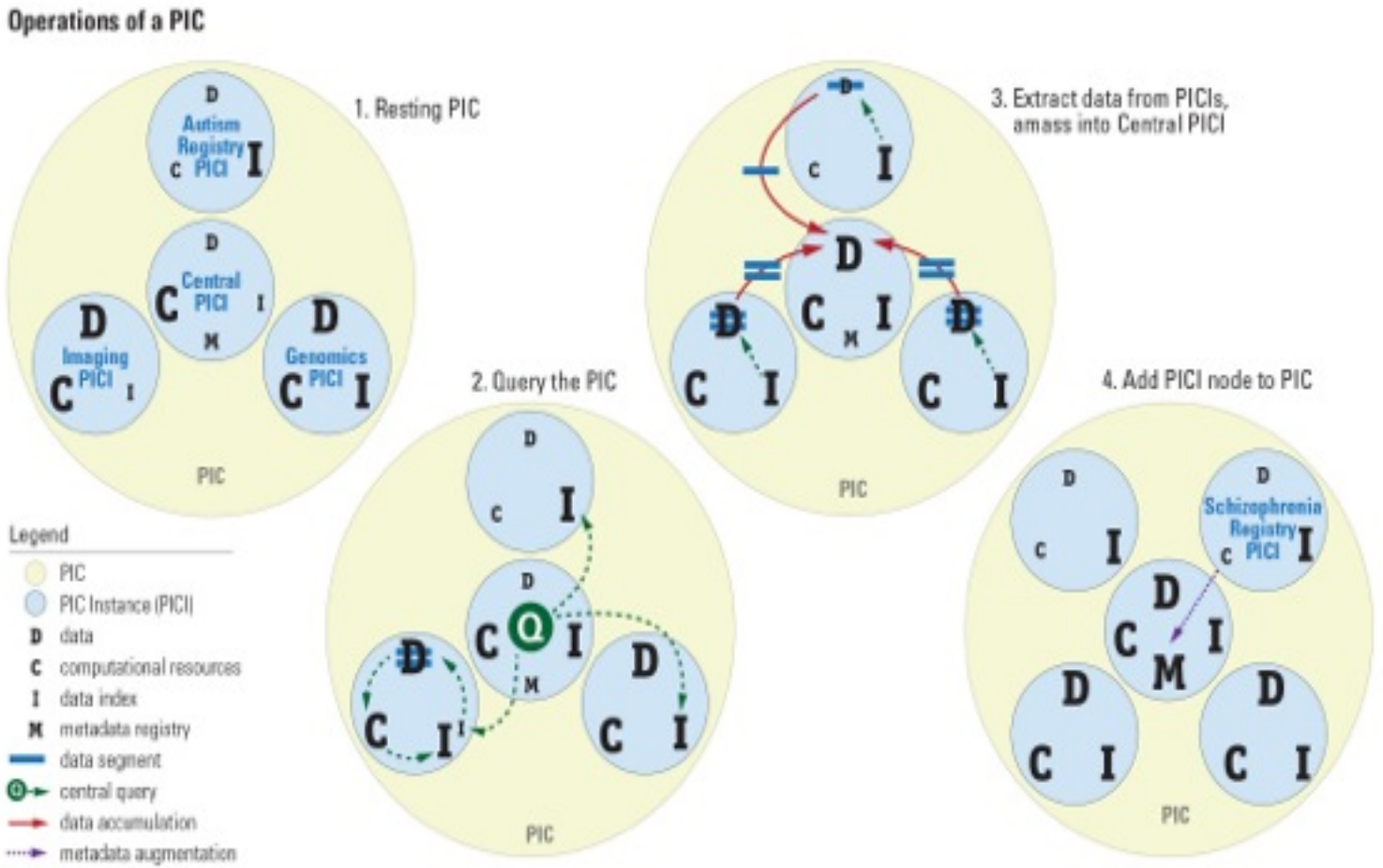
Insights From Population Health Data


Public
Digital SociAL Data
Healthcare Workers Electronic Health Records
Researchers Industry
Sponsors, Journal Publications
Regulatory Agencies

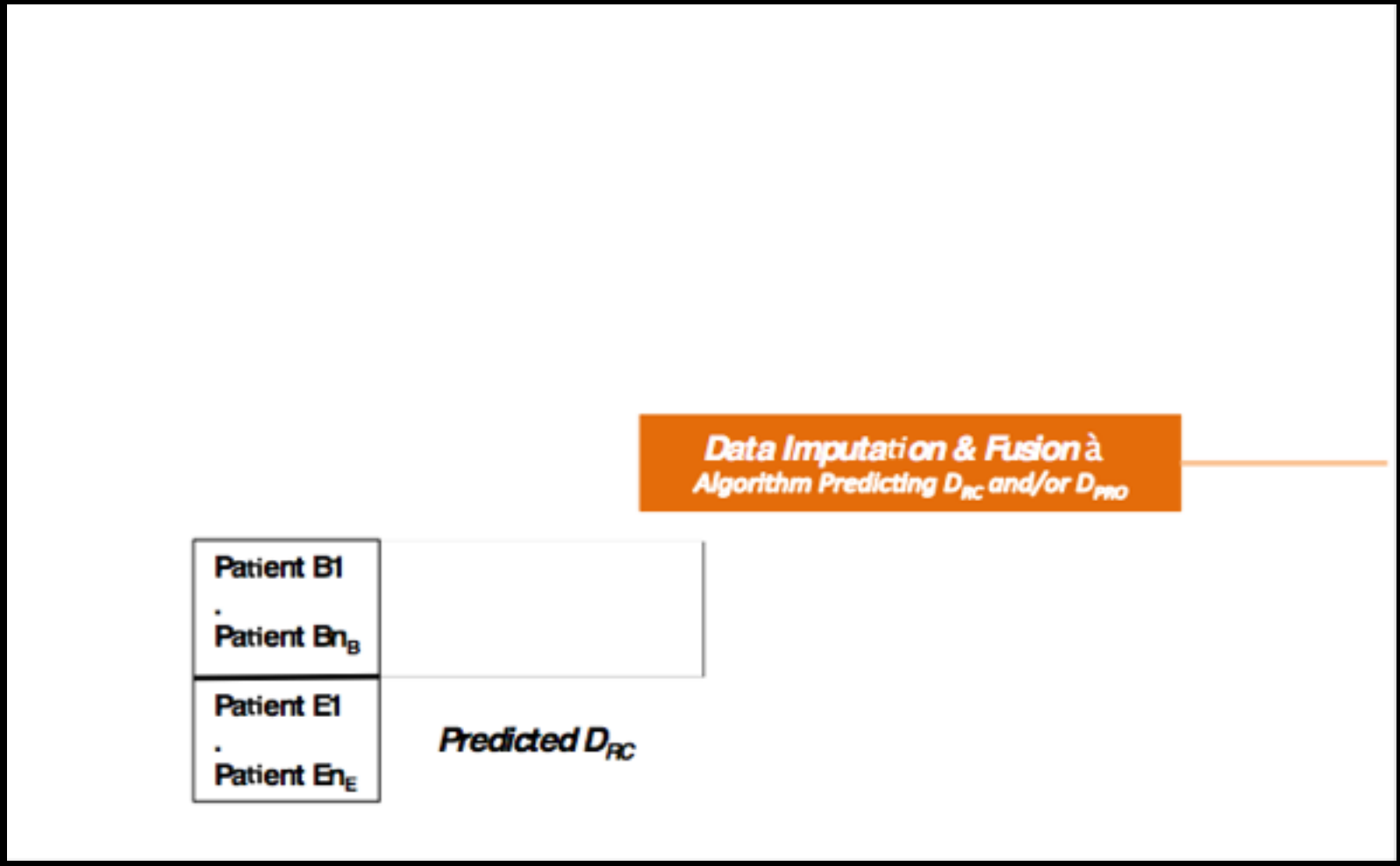
A Community of Patient-centered Information Commons



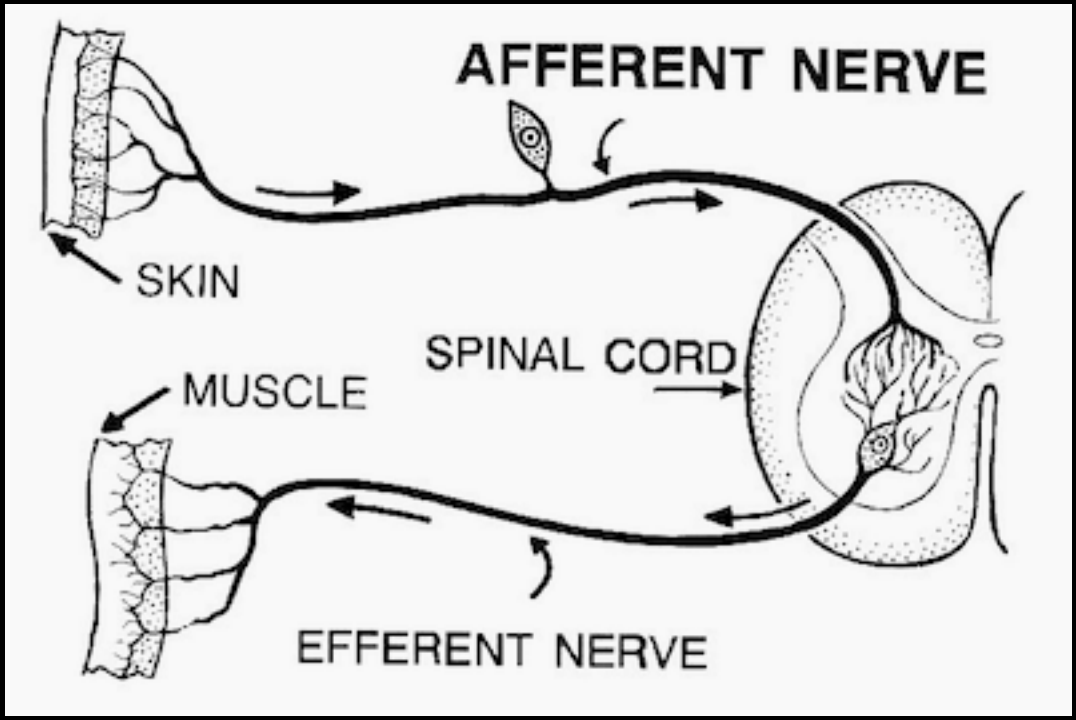
Managing the Dance Between Computational Resources and Data Location



Challenge: Data Imputation and Fusion



The second ½ of our mission:
Afferent and **Efferent** Arms of Healthcare Information



Challenge:

The NEW ENGLAND JOURNAL of MEDICINE

NEJM Volume 360:1278-1281 March 26, 2009 Number 13

No Small Change for the Health Information Economy

Kenneth D. Mandl, M.D., M.P.H., and Isaac S. Kohane, M.D., Ph.D.

Four Propositions

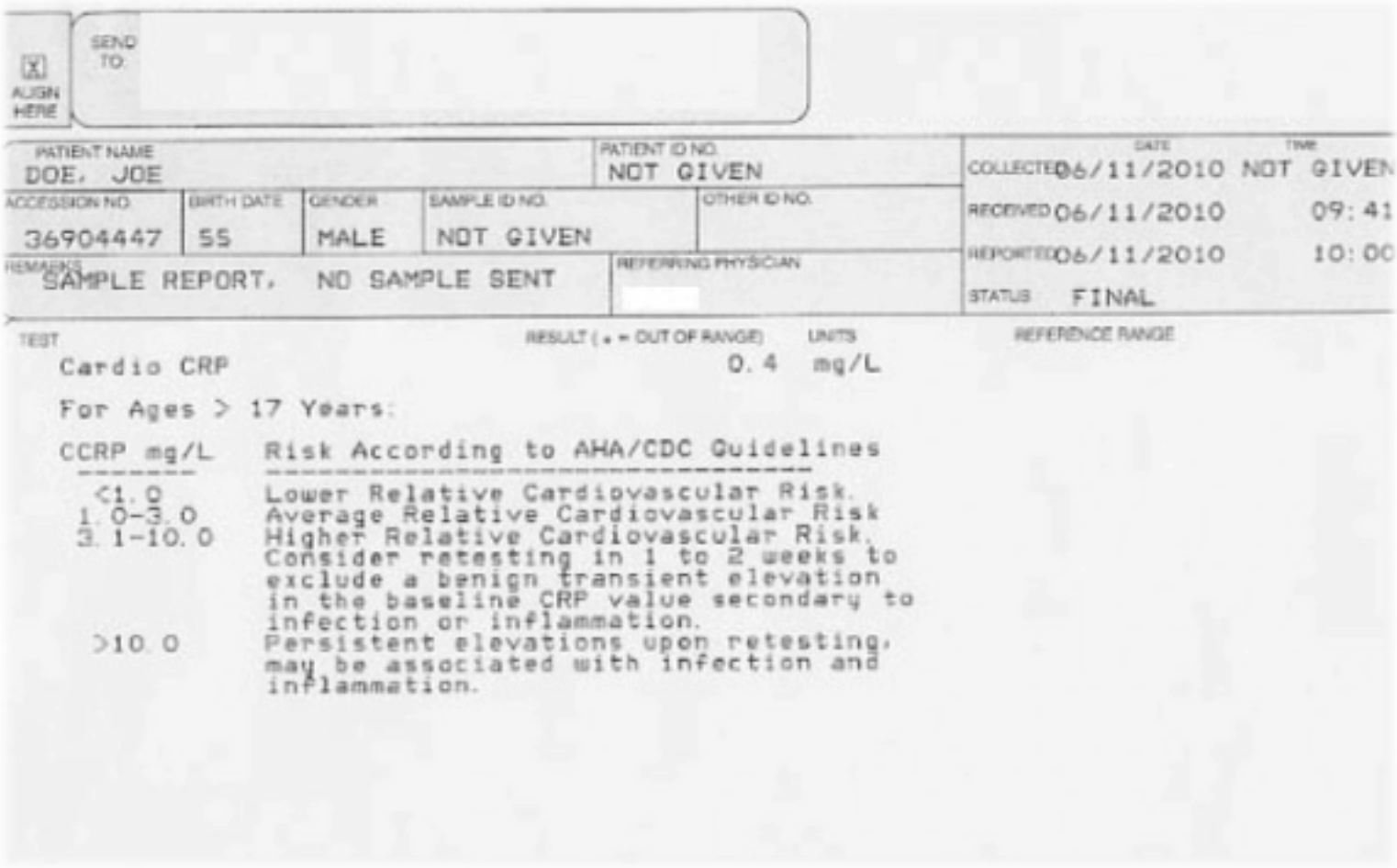
(1) **Liquidity of data**. The platform and its applications should reduce impediments to the transfer of data, in an agreed-upon form, from one system to another.

(2) **Substitutability of applications**. The system should be sufficiently modular and interoperable so that ... just as consumers may swap out applications on their iPhones, physicians should be able to readily replace one … system with another.

(3) **Platform should be built to open standards** accommodating both open-source and closed-source software.

(4) **Platform that supports diverse applications** will lead to a robust health information economy … by allowing competition and "natural selection" for high-value, low-cost products.

State-of-the-Art … in 2011???



SEND TO:

ALIGN HERE

PATIENT NAME				PATIENT ID NO.	DATE	TIME
DOE, JOE				NOT GIVEN	COLLECTED 06/11/2010	NOT GIVEN
ACCESSION NO.	BIRTH DATE	GENDER	SAMPLE ID NO.	OTHER ID NO.	RECEIVED 06/11/2010	09:41
36904447	55	MALE	NOT GIVEN		REPORTED 06/11/2010	10:00
REMARKS: SAMPLE REPORT, NO SAMPLE SENT			REFERRING PHYSICIAN		STATUS FINAL	

TEST	RESULT (* = OUT OF RANGE)	UNITS	REFERENCE RANGE
Cardio CRP	0.4	mg/L	

For Ages > 17 Years:

CCRP mg/L	Risk According to AHA/CDC Guidelines
<1.0	Lower Relative Cardiovascular Risk.
1.0-3.0	Average Relative Cardiovascular Risk
3.1-10.0	Higher Relative Cardiovascular Risk. Consider retesting in 1 to 2 weeks to exclude a benign transient elevation in the baseline CRP value secondary to infection or inflammation.
>10.0	Persistent elevations upon retesting, may be associated with infection and inflammation.

SMART Platform*

- **Data Model**
  Medical Context, Medical Record Elements, Data Coding
- **Application Programming Interface (API)**
  Consistent developer data access mechanisms
- **User Interface**
  Standards-based integration, flexibility
- **Authentication**
  In-browser, server-to-server

Our vocabulary

- **Data Sources**

  Managed by containers

- **Containers**

  Present data from data sources to apps in a uniform fashion

- **Apps**

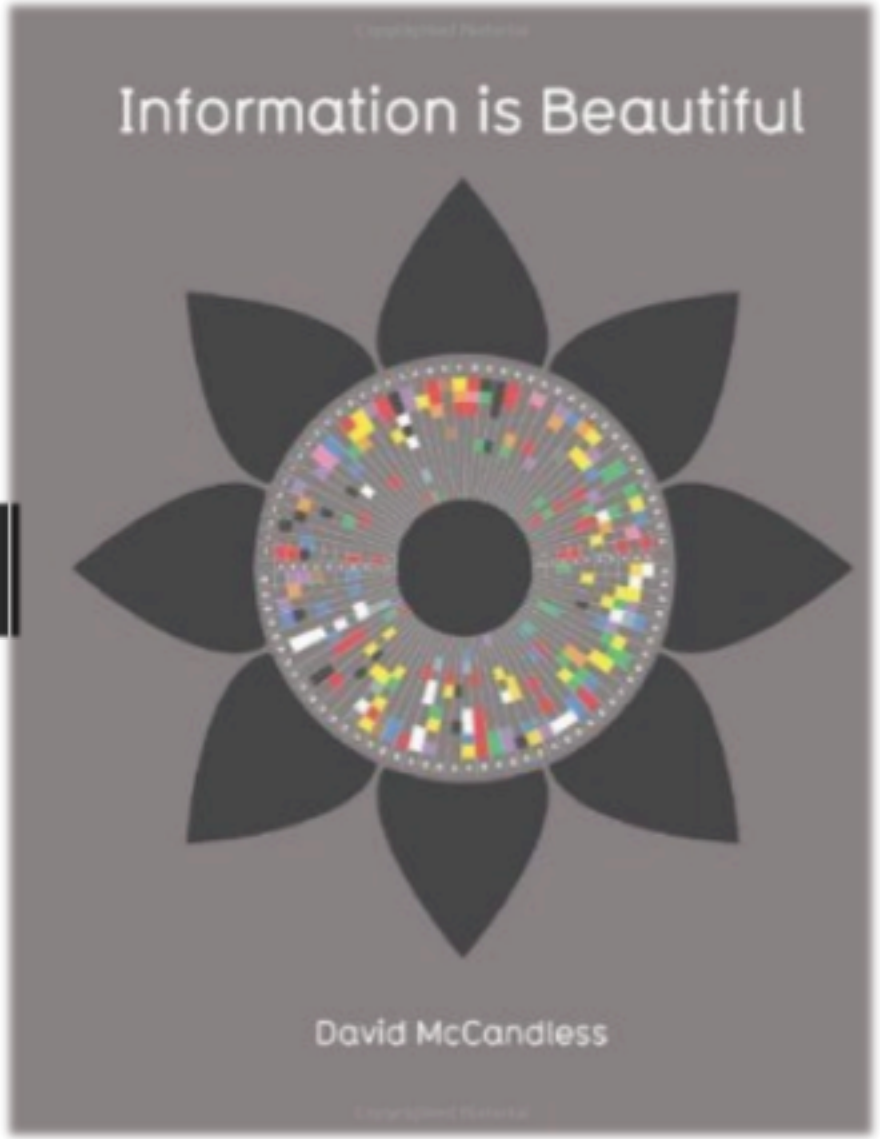
  Completely substitutable

Inspired by a

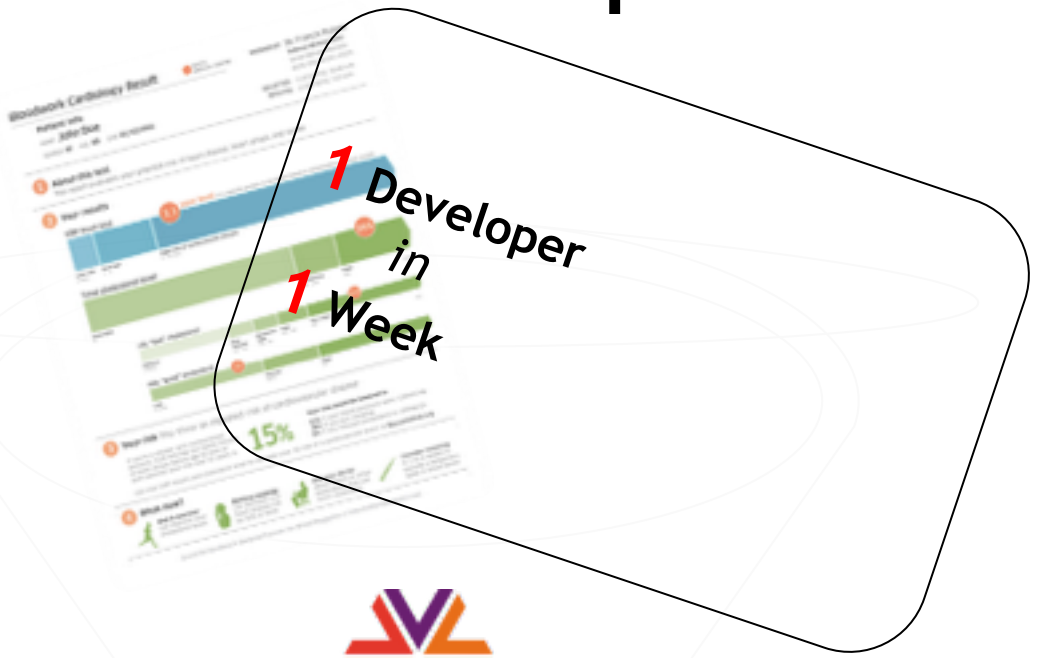
WIRED

18.12 Issue

Design Challenge



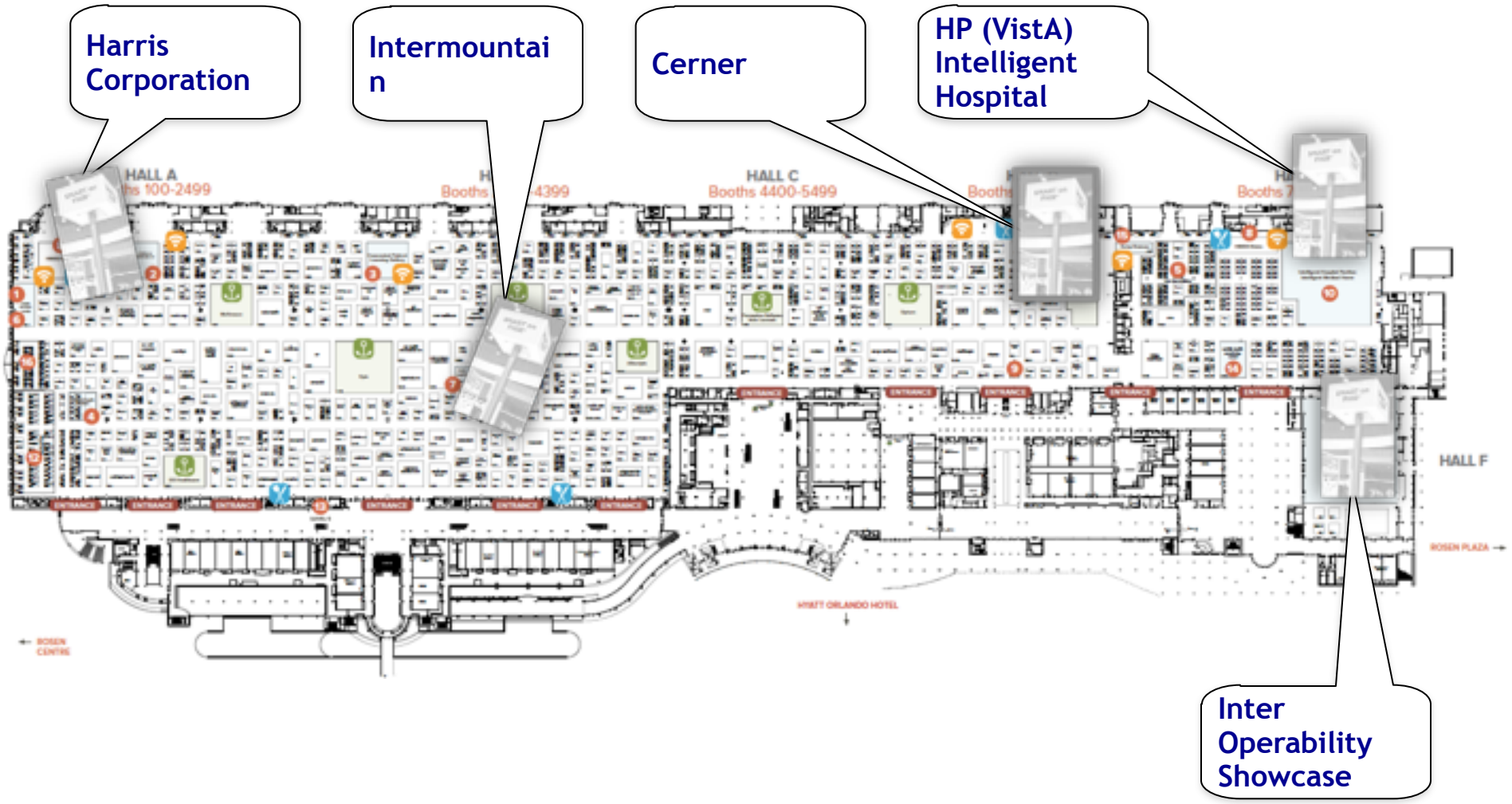
1 Design + 1 Developer + 1 Week



SMART

SMART
Cardiac
Risk
App

SMART on FHIR® at HIMSS 14



HIMSS 14 (Feb 24-27, 2014)

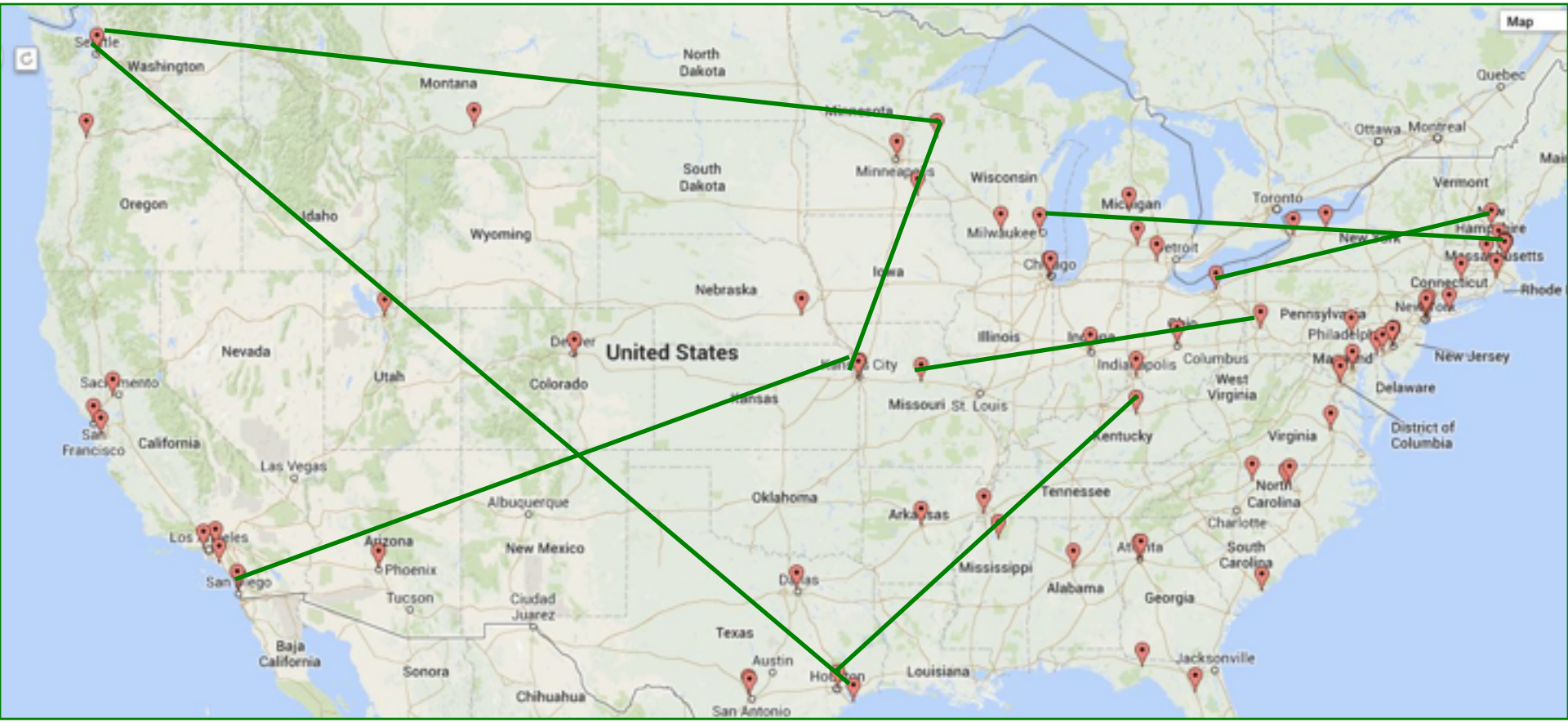
The Effect Arm Opportunity

- Clinical Trials (PCORI & CTSA)
- Direct to/from patient
- Integration with home-monitoring

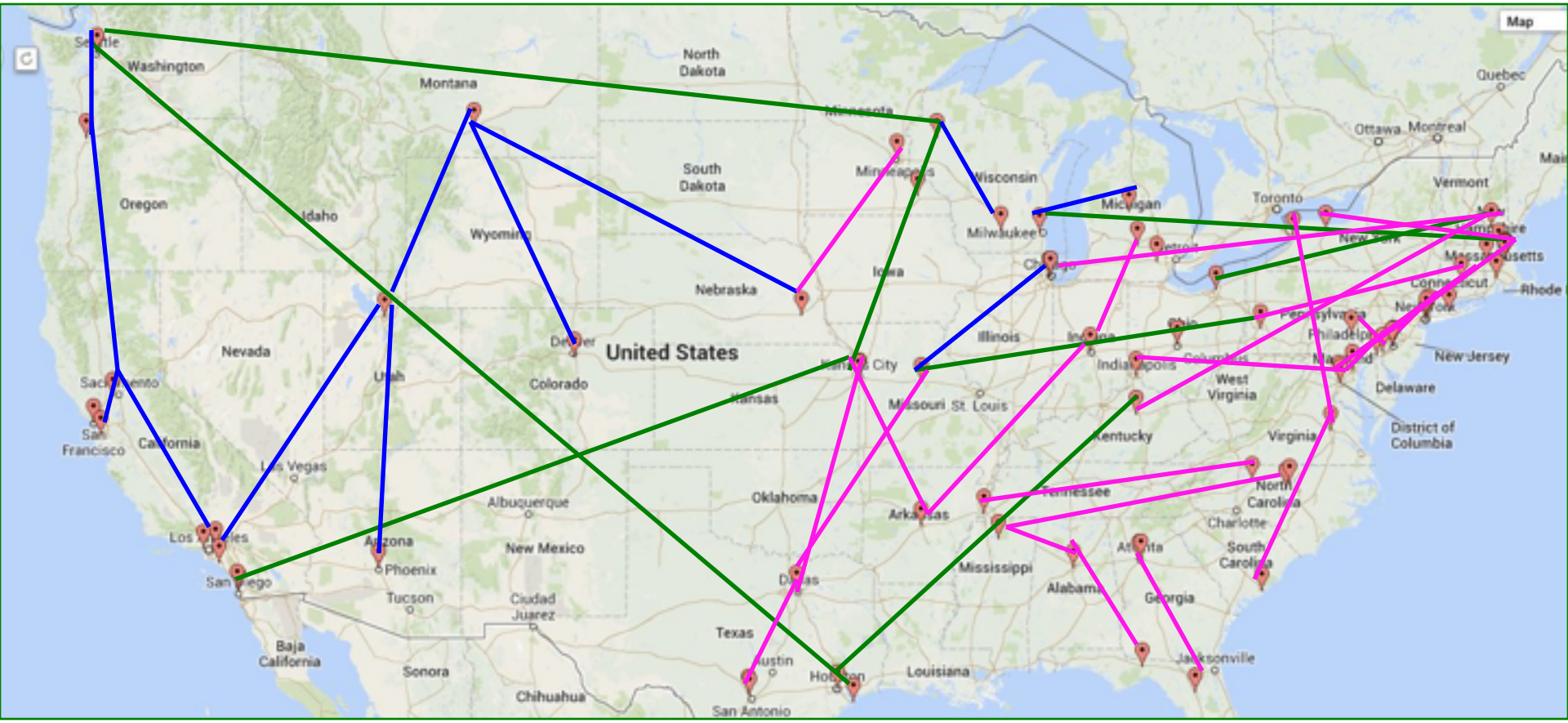
Methodological Agenda

- Biomedical relevance of data science → expansive view of what constitutes biomedical data.
- The investment in healthcare can be leveraged for data science with incremental $.
  - Challenge: Randomization in retrospective-prospective and other *in silico* trial designs.
  - Linkage across institution and data modalities
  - Generalizing for methods of data-type dependent imputation, de-noising

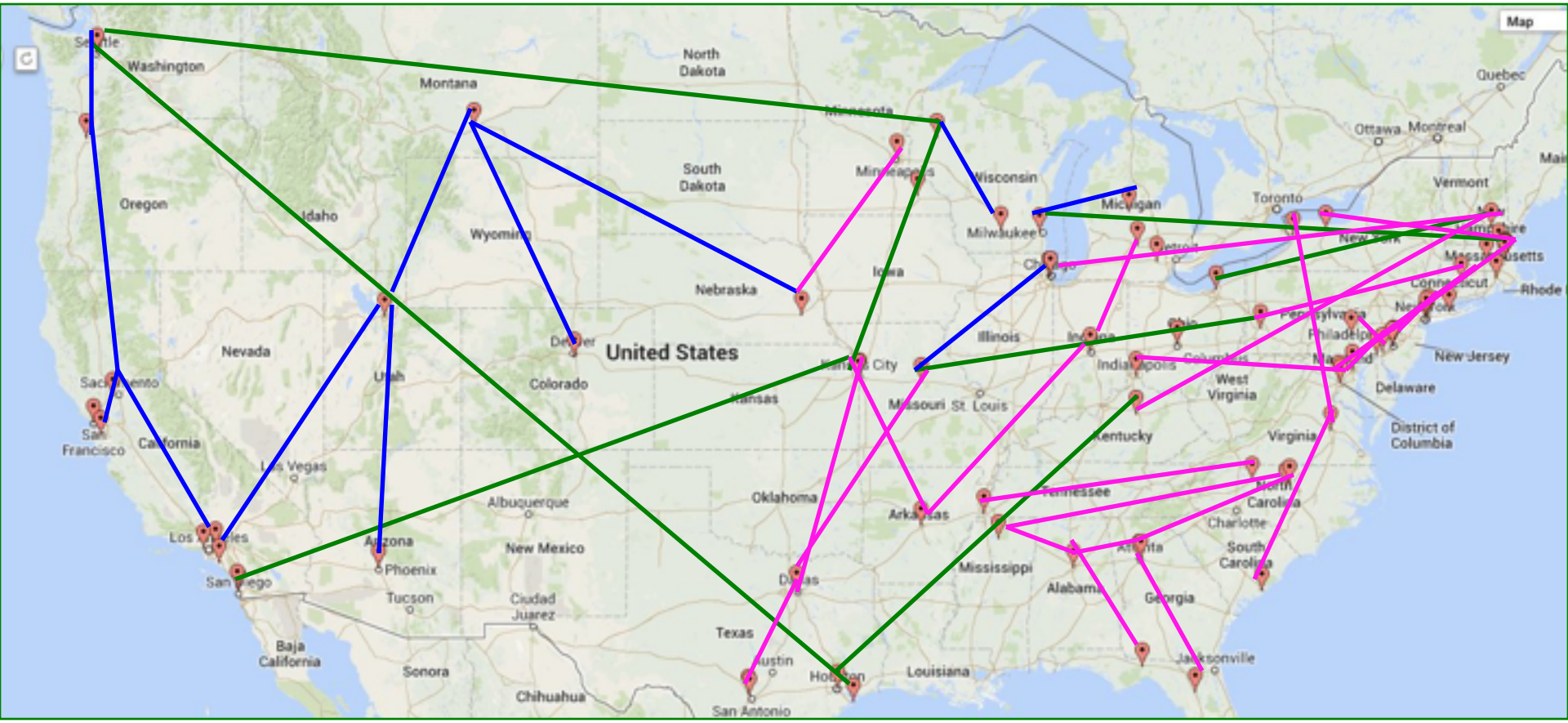
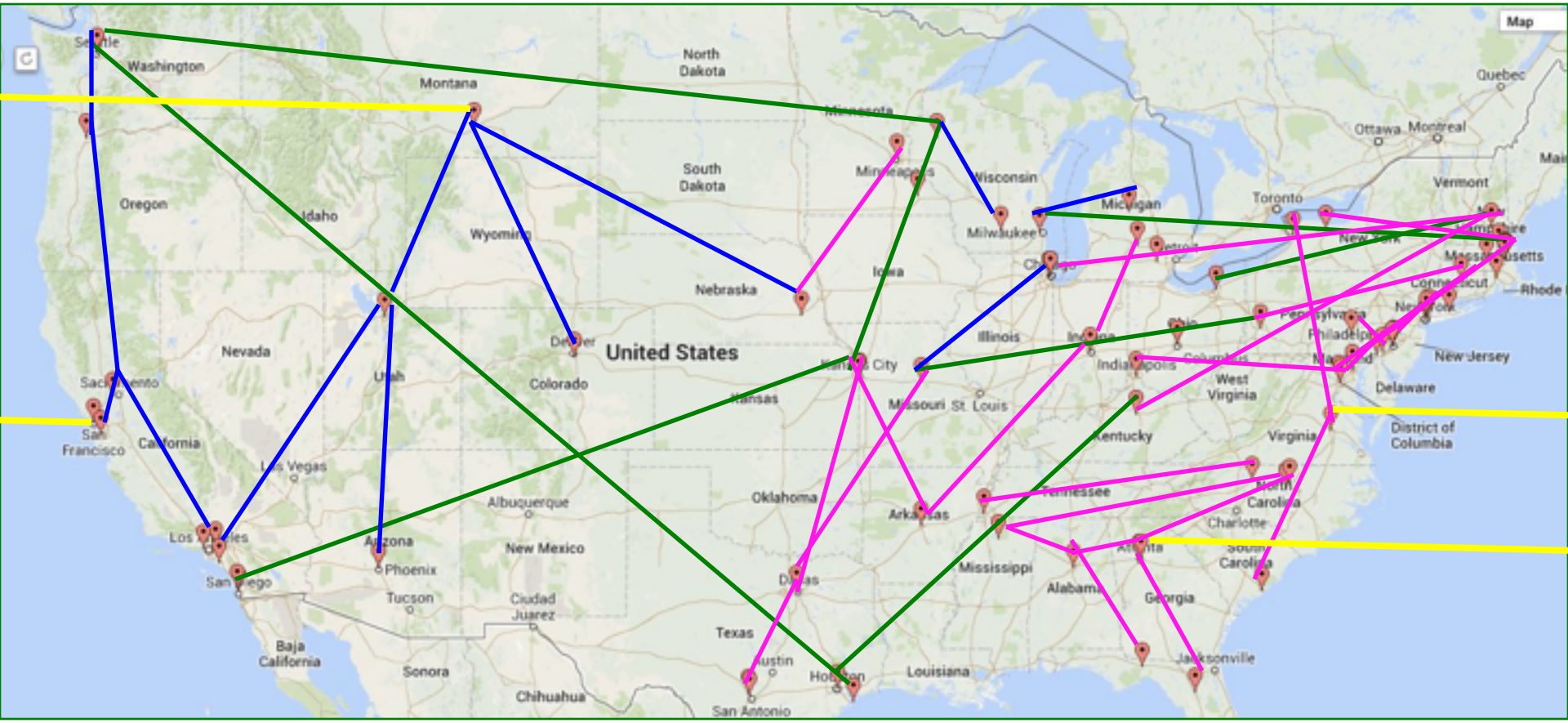
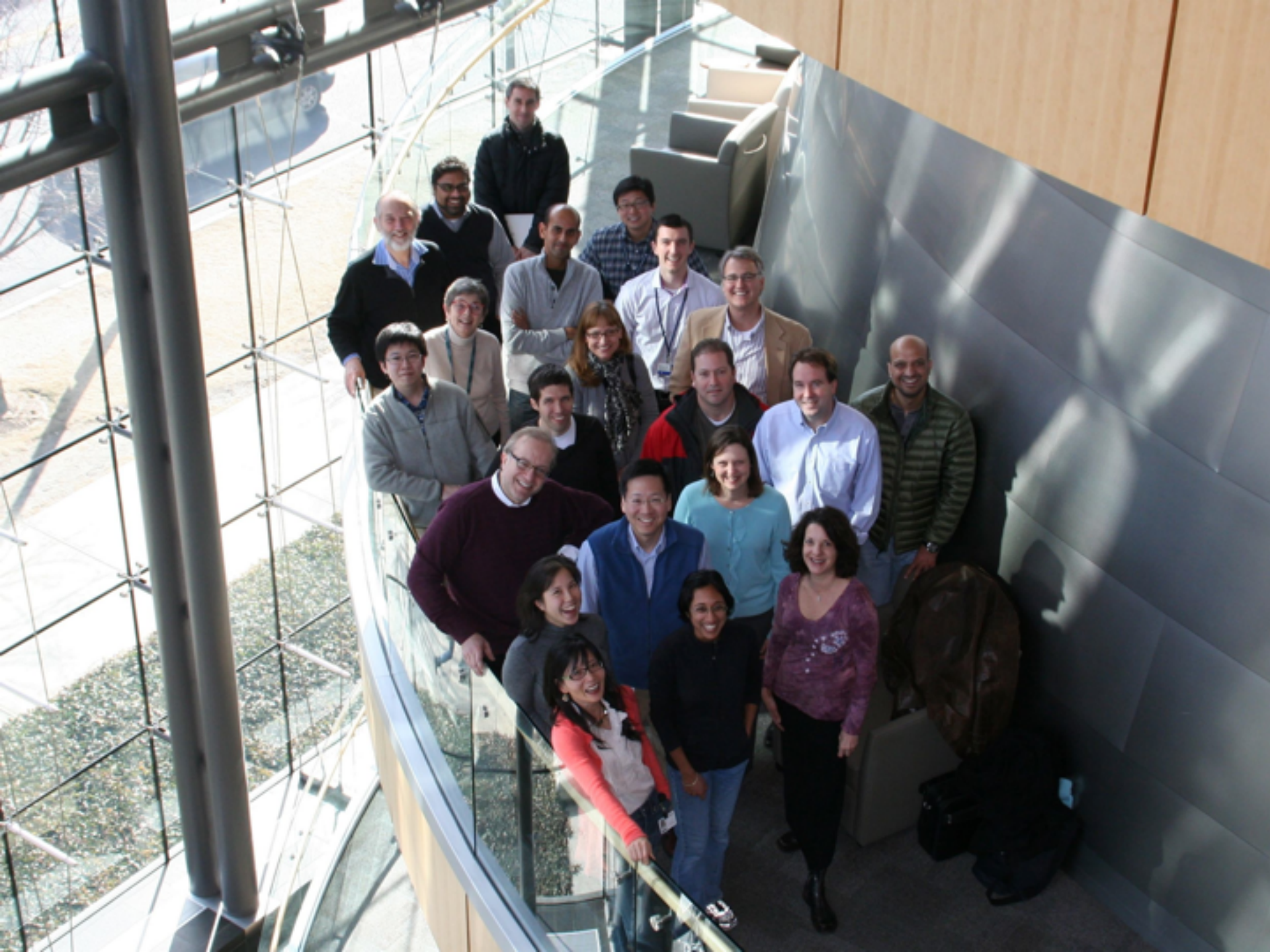
Federated and Federalist



Federated and Federalist



Federated and Federalist



Federated and Federalist