From Hospitals to Molecules:

Learning Biology through Observational Clinical Data

Rami Vanguri

Department of Biomedical Informatics
Columbia University

OSG All Hands Meeting
March 7, 2017
San Diego Supercomputer Center, La Jolla, CA

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

My Background

- Undergraduate at UCSD and worked for fkw on CDF
- PhD at Penn on ATLAS
- Currently Postdoctoral Research Scientist at Columbia University working for Nicholas Tatonetti
- The result is that I know something about computing, next to nothing about biology

What is biomedical informatics?

- “Biomedical informatics is the study of information and computation in biology and health. Healthcare research is experiencing a deluge of new data — such as a patient’s genome sequence, electronic medical records, or the complete genomic and metabolic characterization of a tumor — which necessitate the development of novel methods to interrogate, integrate, analyze, and organize this diverse information.”
- Design and implement novel quantitative and computational methods to solve wide array of problems in biology and medicine

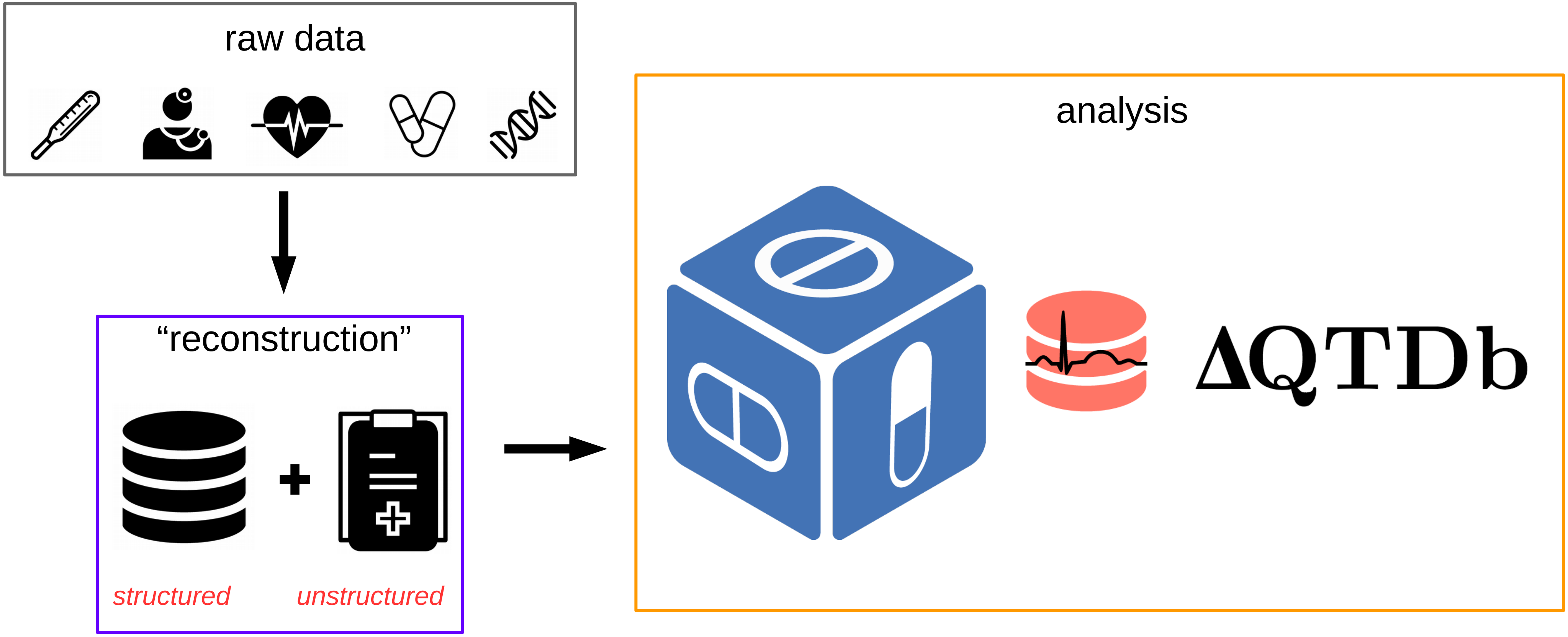
What does our lab do?

- Translational bioinformatics: integrate medical observations with systems and chemical biology models to further biological understanding
- “Bench to bedside”

Why big computing?

- Computational jobs are becoming larger
  - Used to be able to use 2 servers with ~100 CPUs
  - Reached limitations, went to AWS and OSG
- Deep learning extremely powerful tool, efficient via GPU

datasets are
heterogenous!



raw data

“reconstruction”

structured

unstructured

analysis

ΔQTDb

Clinical Data Challenges

- Missingness, incomplete, messy
- Heterogeneous data types (genetics, EHR, protein networks)
- Protected Health Information – HIPAA concerns
- Electronic health records stored in SQL tables

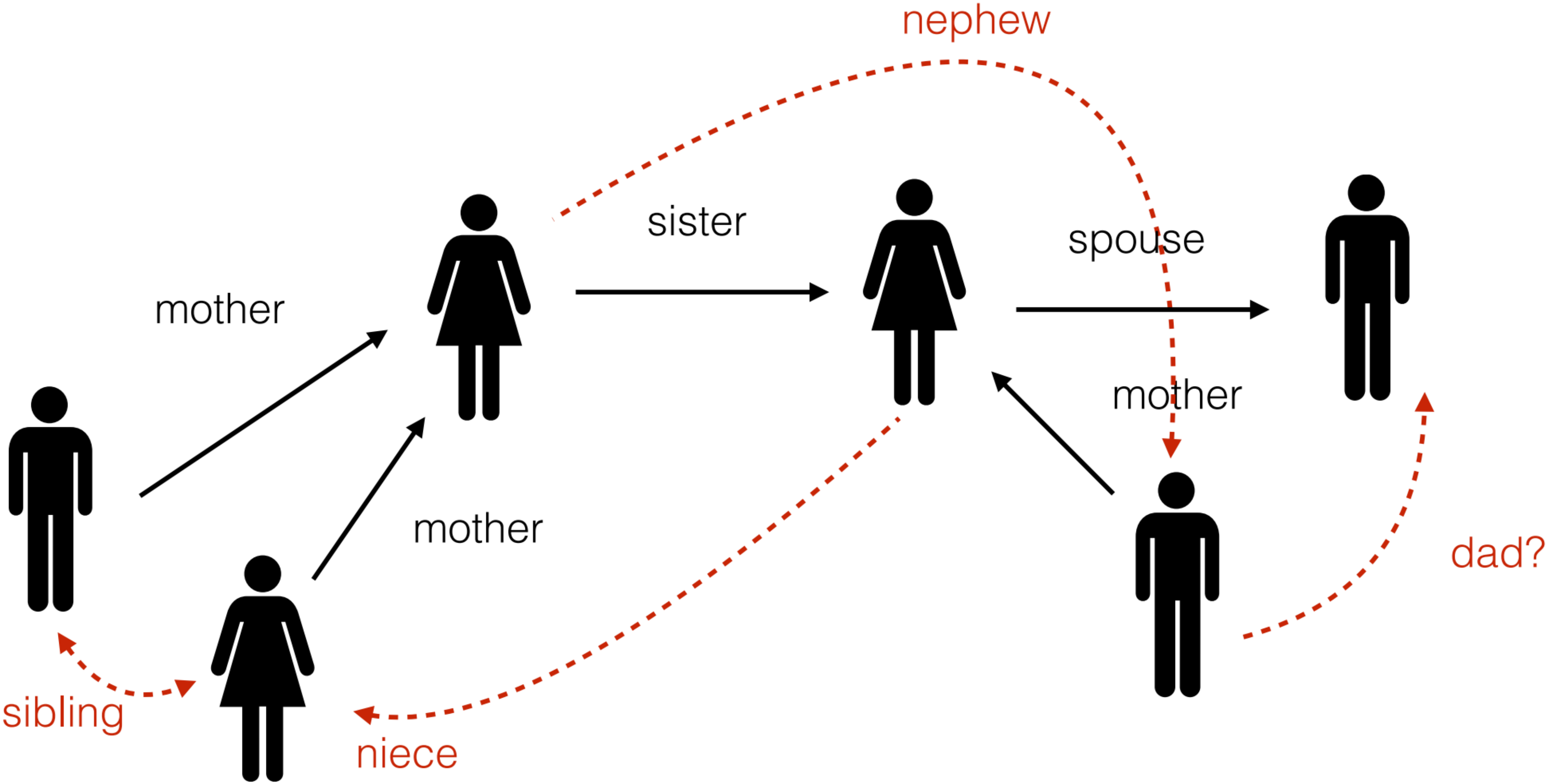
Clinical Data Analysis Example: h_2

- Heritability estimates the amount of variation in a trait is due to genetics (vs environment), known as h_2
  - Estimating heritability usually involves in-depth dedicated studies (twins, mice, etc)
  - Limited sample size

By using emergency contact information in Columbia University Medical Center electronic health records, we can infer 4.7M familial relationships and use them to estimate various disease heritabilities.

Code	Description	Count
MOT	MOTHER	386180
SPO	SPOUSE	275870
FAT	FATHER	153900
CHI	CHILD	80797
SIB	SIBLING	59773
UNK	UNKNOWN	57175
18	SELF	40662
LIF	LIFE PARTNER	29232
PAR	PARENT	21885
GRP	GRANDPARENT	21345
AUN	AUNT / UNCLE	10428
NNE	NIECE/NEPHEW	10415
Total		**~1.1 million**

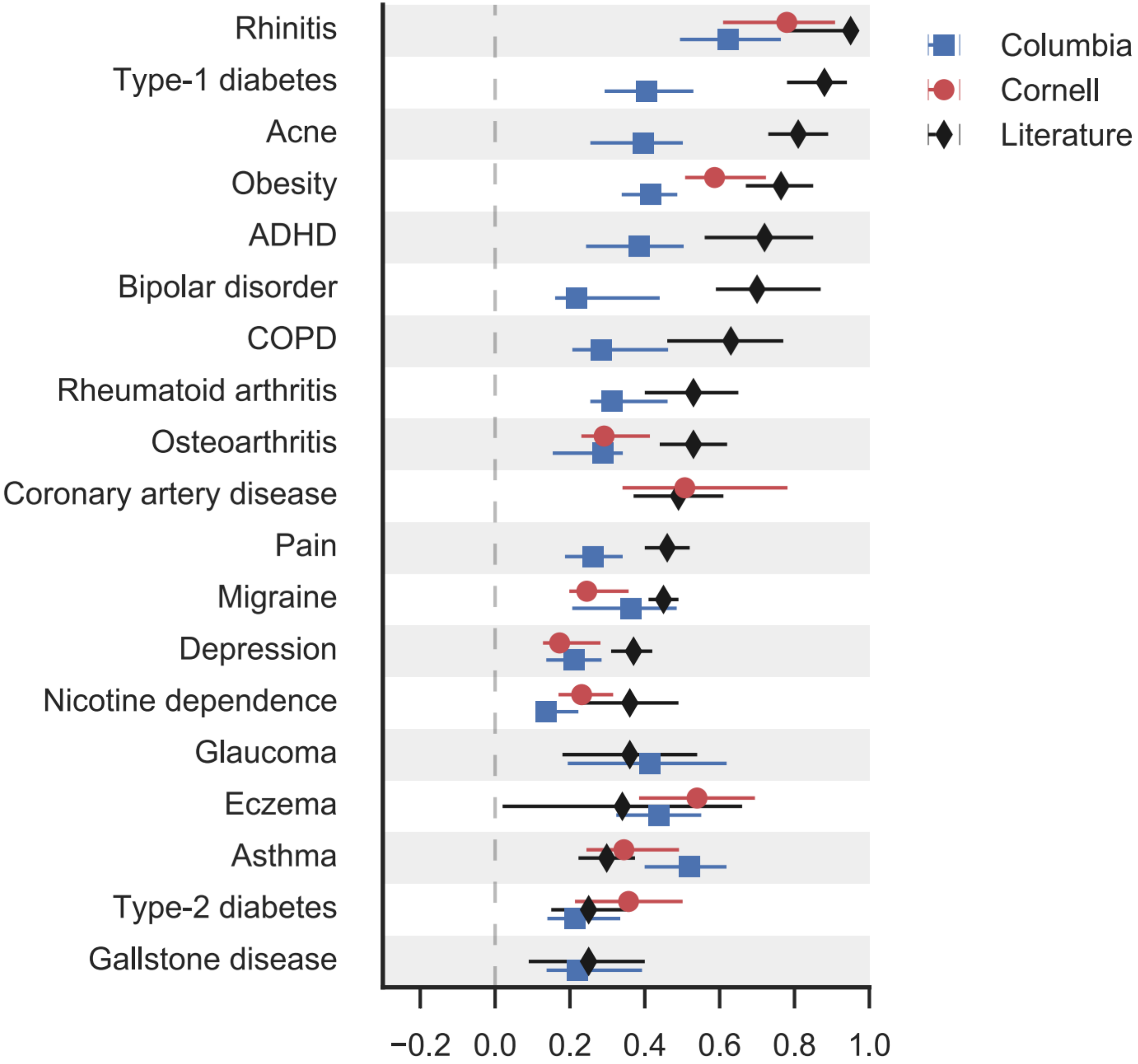
Inferred Relationships



Description	**Columbia**
Child	482,308
Parent	482,308
Sibling	424,242
Aunt/Uncle	185,822
Nephew/Niece	185,822
Spouse	169,017
Cousin	142,435
Sibling/Sibling-in-law	132,538
Grandparent	117,139
Grandchild	117,139
Grandaunt/Granduncle	96,675
Grandnephew/Grandniece	96,675
First cousin once removed	85,679
Parent/Parent-in-law	52,174
Great-grandparent	45,053
Total	**~3.2 million**

Calculating Heritability

- Traits are assigned in electronic health records via insurance billing codes (ICD-9)
- Observational heritability: estimate of h_2 where the phenotypes are from observational data
    - Access to traits not able to evaluate with traditional studies (such as neurological)


Rhinitis
Type-1 diabetes
Acne
Obesity
ADHD
Bipolar disorder
COPD
Rheumatoid arthritis
Osteoarthritis
Coronary artery disease
Pain
Migraine
Depression
Nicotine dependence
Glaucoma
Eczema
Asthma
Type-2 diabetes
Gallstone disease
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Columbia
Cornell
Literature

Specifics on Computing Needs

- Small data input (list of individuals with/without trait), small data output (h_2), long processing time
- Thousands of jobs – time for each job (trait) depends on number of affected individuals
- Difficult to know runtime a priori

Next project (nSIDES)

- Mine public FDA dataset for statistically significant drug effects
- Deep learning is used to to calculate bias space in FDA reports
  - We have a GPU test bed for this (Tesla K40)
  - Not sustainable for the number of models we need to generate

Specifics on Computing Needs

- GPU jobs, take hours each
  - ~4500 initial jobs to calculate single drug effects
  - Many more to calculate drug interactions
- AWS mechanism to connect instances will be used to supplement OSG resources

Biomedical Translator

NIH funded program to accelerate biomedical translation for the research community. Existing biomedical data spanning clinical, genetic and fundamental biology will be integrated to form disease classification that can be targeted by various preventative and therapeutic interventions.

Biomedical Translator

- Spans 11 universities including Columbia and UCSD (Trey Ideker)
- We will use nSIDES to form prototype for translator – DeepLink

(Marticain)

www.twosides.io/?drug=␣,␣&AE=CUI␣

all/any/multi-sides

[(Drug1, Drug2), AE]

- Timeline of AE signal over time
  - PRR
    - PSM
    - naive (behind a layer)
  - # reports
- Similar drug(s):
  - Chem. similarity
  - Drug Class

- drug1 (+drug2)
- PRR
- pvalue
- a,b,c,d

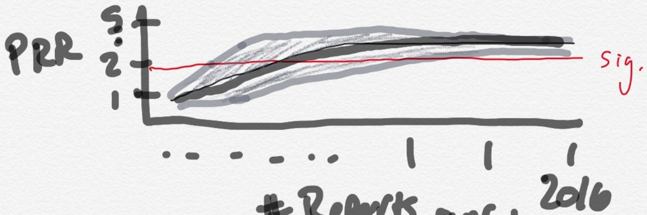
per year

[[Drug 1, Drug2 ⟷ Long QT Syndrome

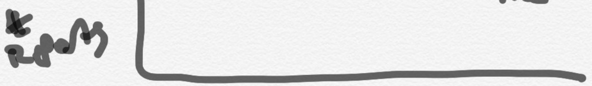
(CUI on mouseover)

Summary: There is / is not a - - - - - (p-val)

PRR over Time



Reports over time



Similarity / Severity

Δ □

images of drugs

Pivots: $\Delta_1 \Delta_2$ $\square_1$

→ $\Delta_1 \Delta_3$ $\square_1$: sim to Δ_2

→ $\Delta_4 \Delta_5$ $\square_1$: strongest/weakest drugs for $\square_1$

→ $\Delta_1 \Delta_2$ $\square_2$: strongest/weakest AE for $\Delta_1 \Delta_2$

- Timeline of AE signal over time
  - PRR
    - PSM
    - naive (behind a layer)
  - # reports
- Similar drug(s):
  - Chem. similarity
  - Drug Class (ATC)
- image of drug (chem structure, Pill)
- outlink to PubMed + ΔQTDb

- drug1 (+drug 2)
- PRR
- pvalue
- a, b, c, d

per year

- CI

Future Projects (Clinical Notes)

- Use deep learning techniques to analyze clinical notes
  - Classify undiagnosed patients
  - Discover distinct disease subtypes
  - Predict patient disease course
- We predict that GPUs will be the primary computing need

Future Prospects: Genomics Medicine

- Leverage clinical note analysis to recruit patients for sequencing
- Discover causal genetic variants
- Uncover mechanism

Genetic analysis and deep learning require extensive computing resources

Summary

- As machine learning has advanced, grid computing has become necessary to efficiently analyze large amounts of clinical data
- Direct implications for generating biological hypotheses, leading to better understanding of drug interactions and disease

Acknowledgements

tatonettilab.org
r.vanguri@columbia.edu

Lab Members

Nicholas Tatonetti
Kayla Quinnies
Theresa Kolek
Alexandra Jacunski
Tal Lorberbaum
Mary Boland

Yun Hao
Joseph Romano
Phyllis Thangaraj
Alexandre Yahi
Fernanda Polubriaginof
Victor Nwankwo

Funding

NIH NIGMS R01GM107145
NIH NCATS OT3TR002027
Herbert Irving Fellowship

Tatonetti Lab

at Columbia University

Columbia University
Medical Center