Building a ***search engine*** to find ***environmental*** and ***phenotypic factors*** associated with ***disease and health***

Chirag J Patel

Northeast Big Data Innovation Hub Workshop

02/24/17

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org

$$P \equiv G + E$$

Phenotype

P = G + E

Type 2 Diabetes
Cancer
Alzheimer's
Gene expression

Phenotype *Genome*

P = G + E

Type 2 Diabetes
Cancer
Alzheimer's
Gene expression

Variants

Phenotype *Genome* *Environment*

P = G + E

Type 2 Diabetes
Cancer
Alzheimer's
Gene expression

Variants

Infectious agents
Nutrients
Pollutants
Drugs

G

We are great at ***G*** investigation!

over **2400**
Genome-wide Association Studies (GWAS)
https://www.ebi.ac.uk/gwas/

E: ???

Nothing comparable to elucidate ***E*** influence!

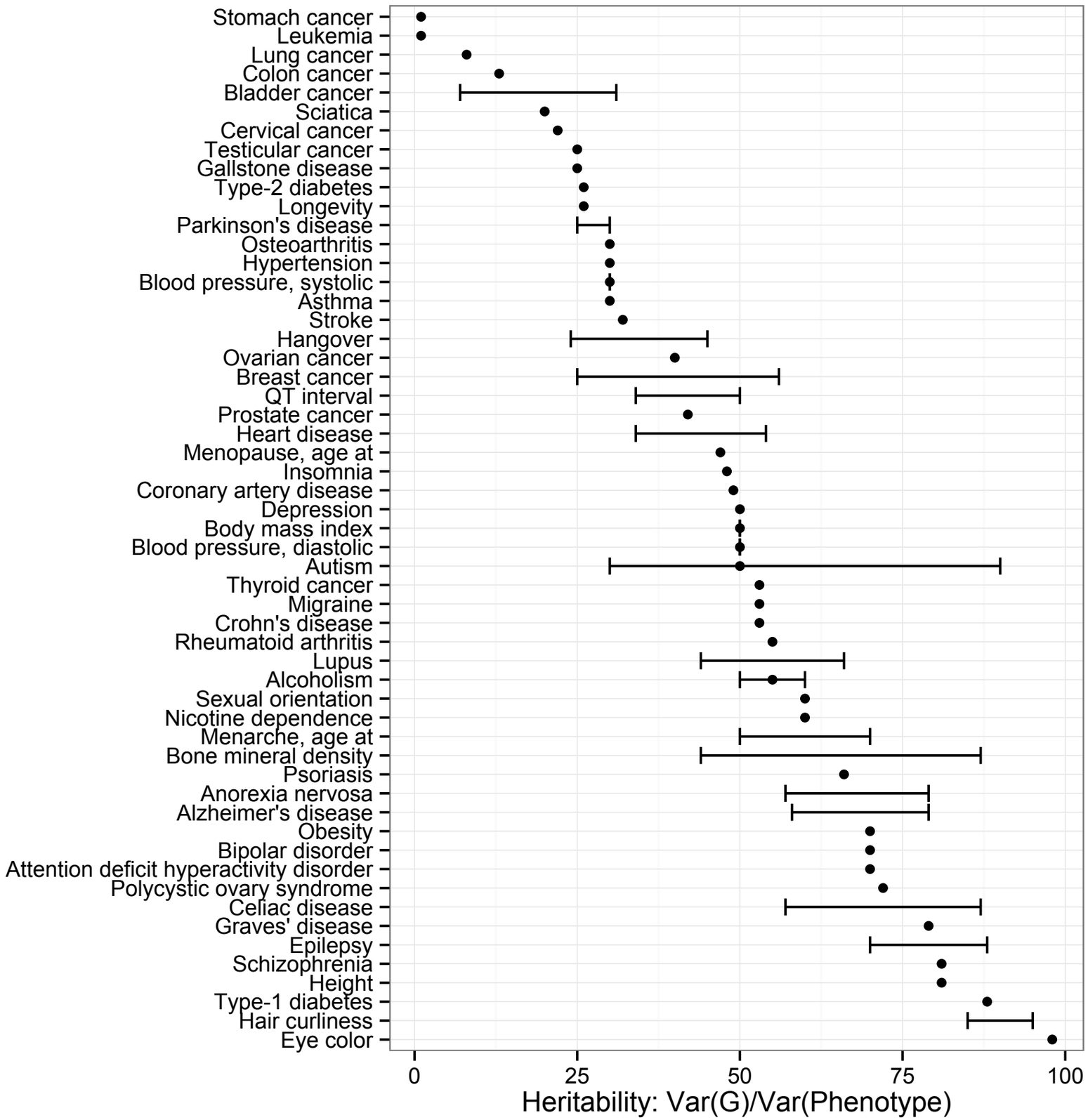
We lack high-throughput methods
and data to discover new ***E*** in ***P…***
until now!

Heritability (H^2) is the range of phenotypic variability attributed to genetic variability in a population

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

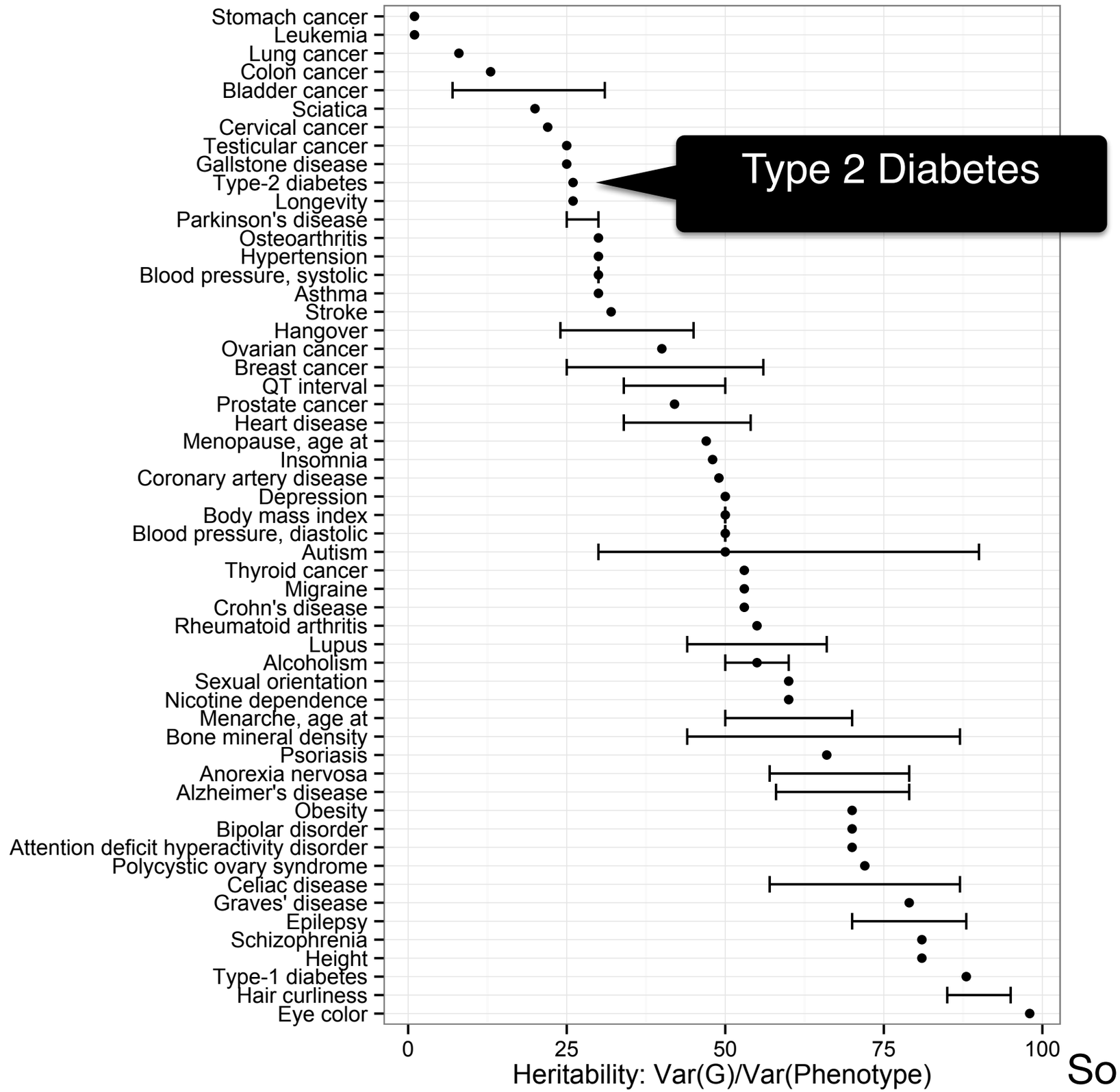
Indicator of the proportion of phenotypic differences attributed to **G**.

G estimates for burdensome diseases are **low and variable:** massive opportunity for ***E*** *discovery*



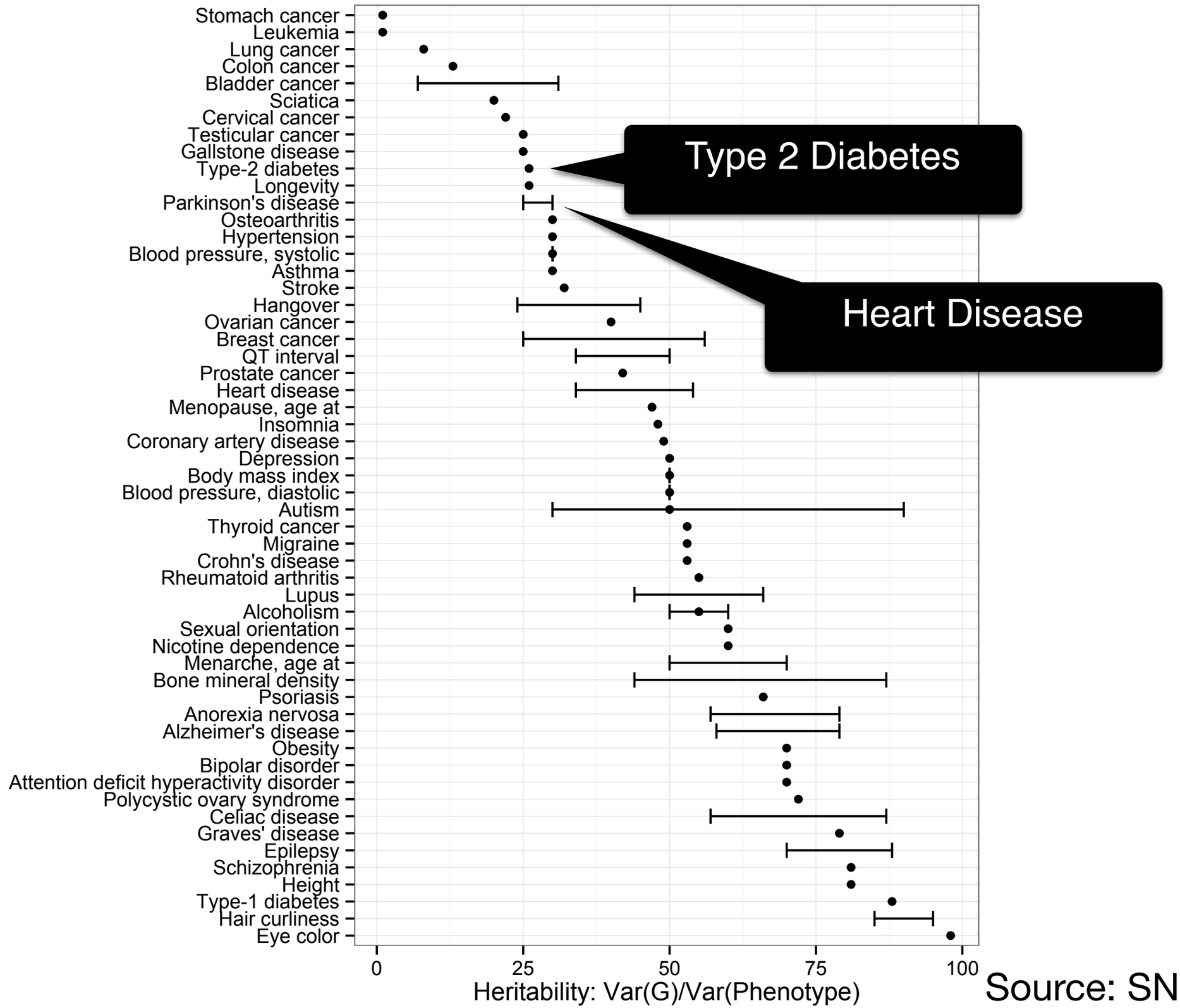
Source: SNPedia.com

G estimates for burdensome diseases are **low and variable:** massive opportunity for ***E*** *discovery*



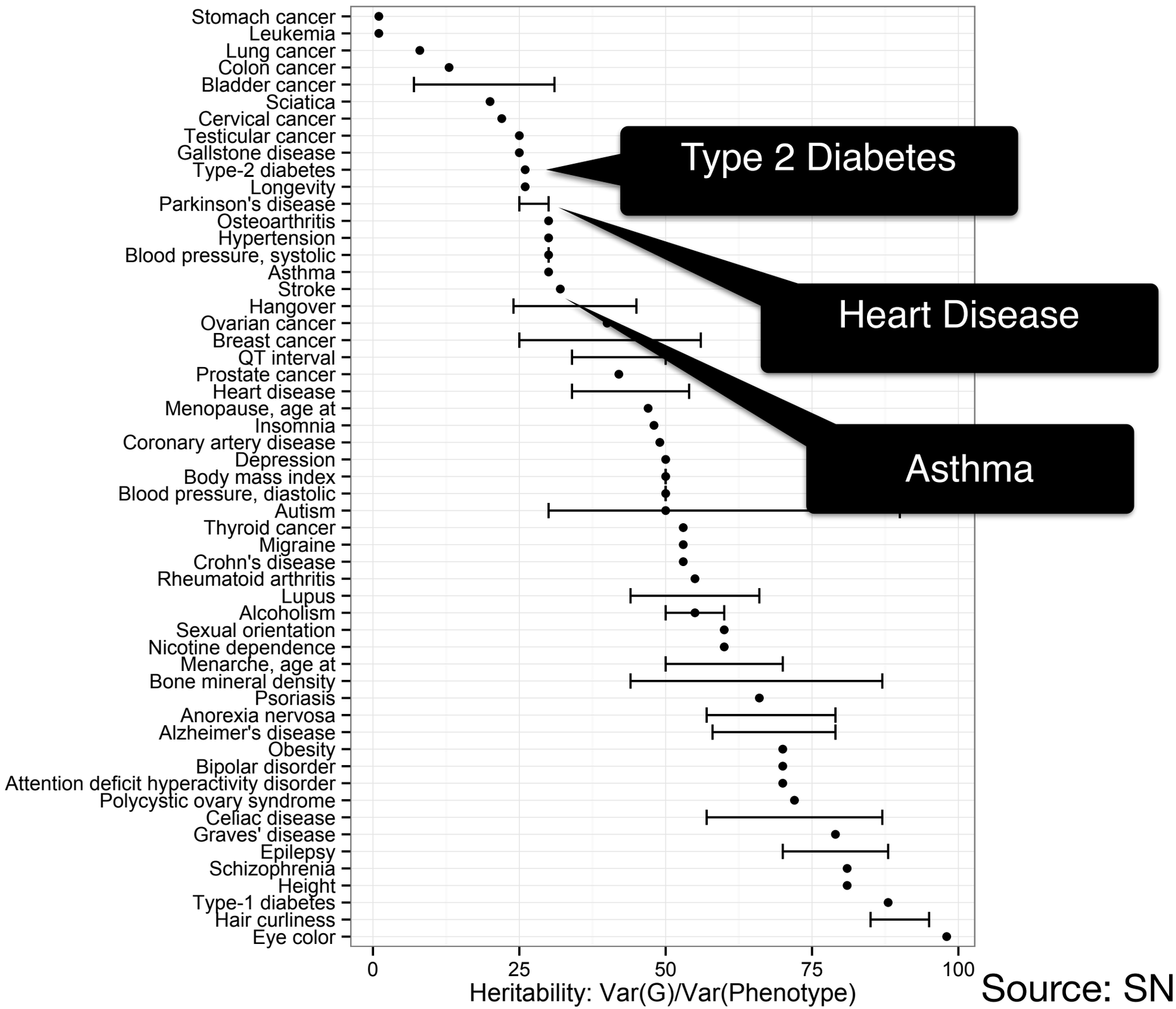
Source: SNPedia.com

G estimates for burdensome diseases are **low and variable:** massive opportunity for ***E*** *discovery*



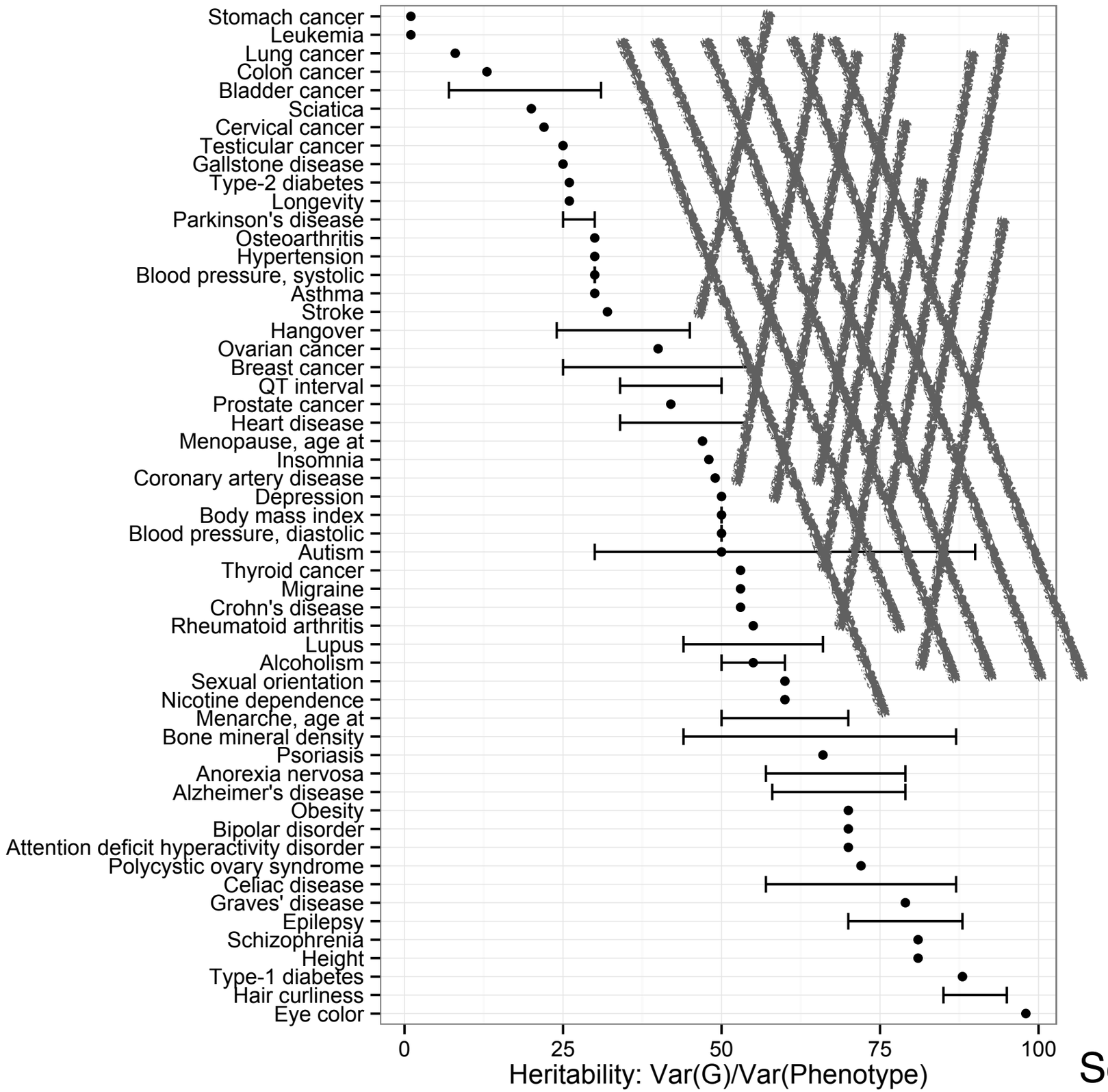
Source: SNPedia.com

G estimates for burdensome diseases are **low and variable:** massive opportunity for ***E*** *discovery*



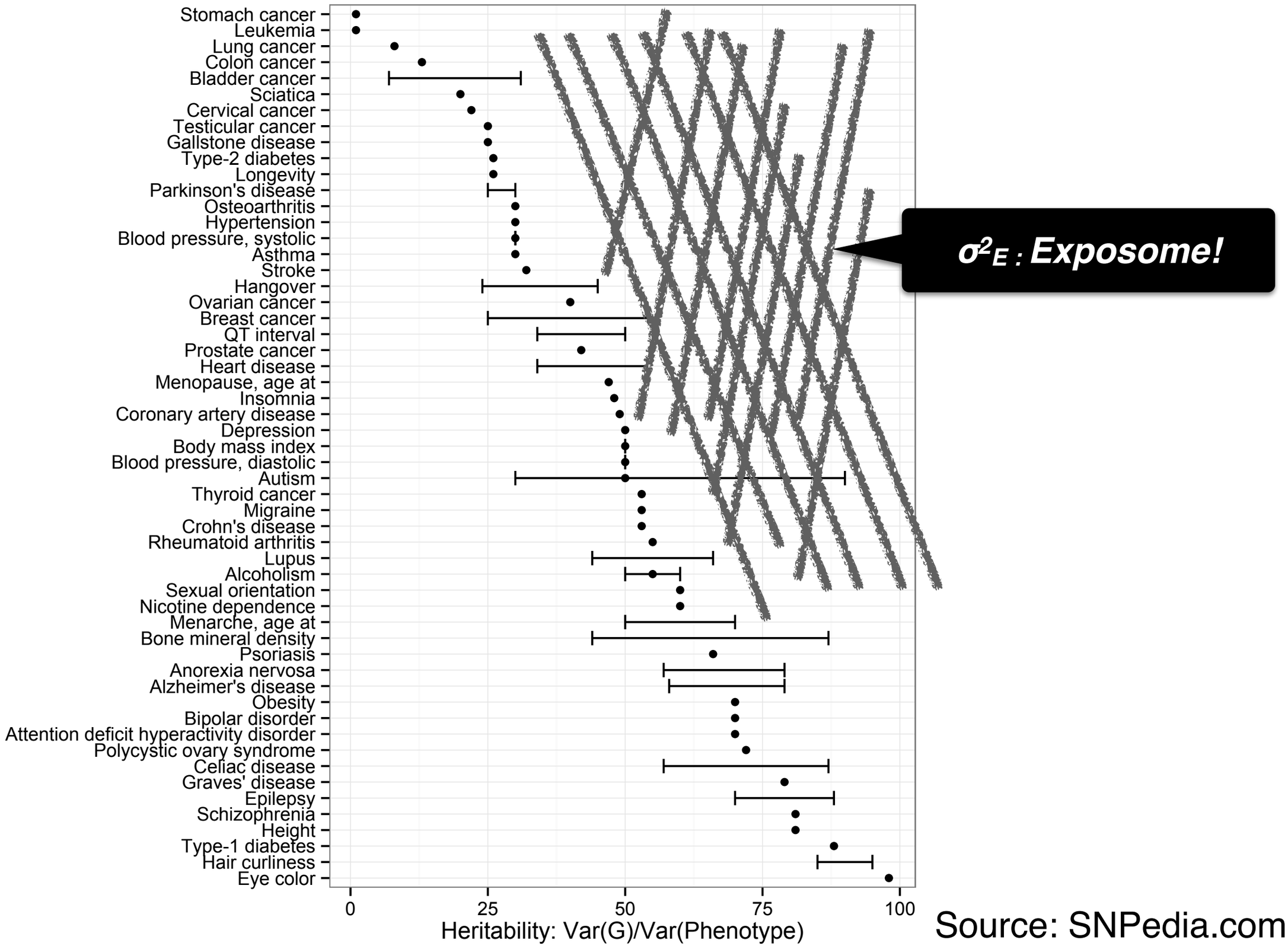
Source: SNPedia.com

G estimates for complex traits are **low and variable:** massive opportunity for *high-throughput **E** discovery*



Source: SNPedia.com

G estimates for complex traits are **low and variable:** massive opportunity for *high-throughput **E** discovery*



Source: SNPedia.com

How can we drive **discovery** of
environmental factors (**E**) in disease phenotypes (**P**)?

How can we drive **discovery** of
environmental factors (**E**) in disease phenotypes (**P**)?

Enhance **accessibility** of **clinical and environmental data,**
and **analytic artificial intelligence tools!**

Enhance **accessibility** of **large open data** and **tools** to drive **discovery** of environmental factors (**E**) in disease phenotypes (**P**)

Noémie Elhadad, PhD
Columbia

Greg Cooper, MD, PhD
Pittsburgh

Chirag Patel, PhD
Harvard

Vasant Honavar, PhD
Penn State

Where do we get disease (***P***) data?

wearable.com

Where do we get disease ***P*** data?

Health record data from your doctor!

Where do we get disease ***P*** data?
Health record data from your doctor!

- Longitudinal data on ***millions*** of patients

Where do we get disease ***P*** data?
Health record data from your doctor!

- Longitudinal data on ***millions*** of patients
  - ***diagnoses***, ***prescriptions***, ***lab reports***, ***notes***

Where do we get disease ***P*** data?
Health record data from your doctor!

- Longitudinal data on ***millions*** of patients
  - ***diagnoses***, ***prescriptions***, ***lab reports***, ***notes***
- Sitting there in institutional IT infrastructure

Where do we get disease ***P*** data?
Health record data from your doctor!

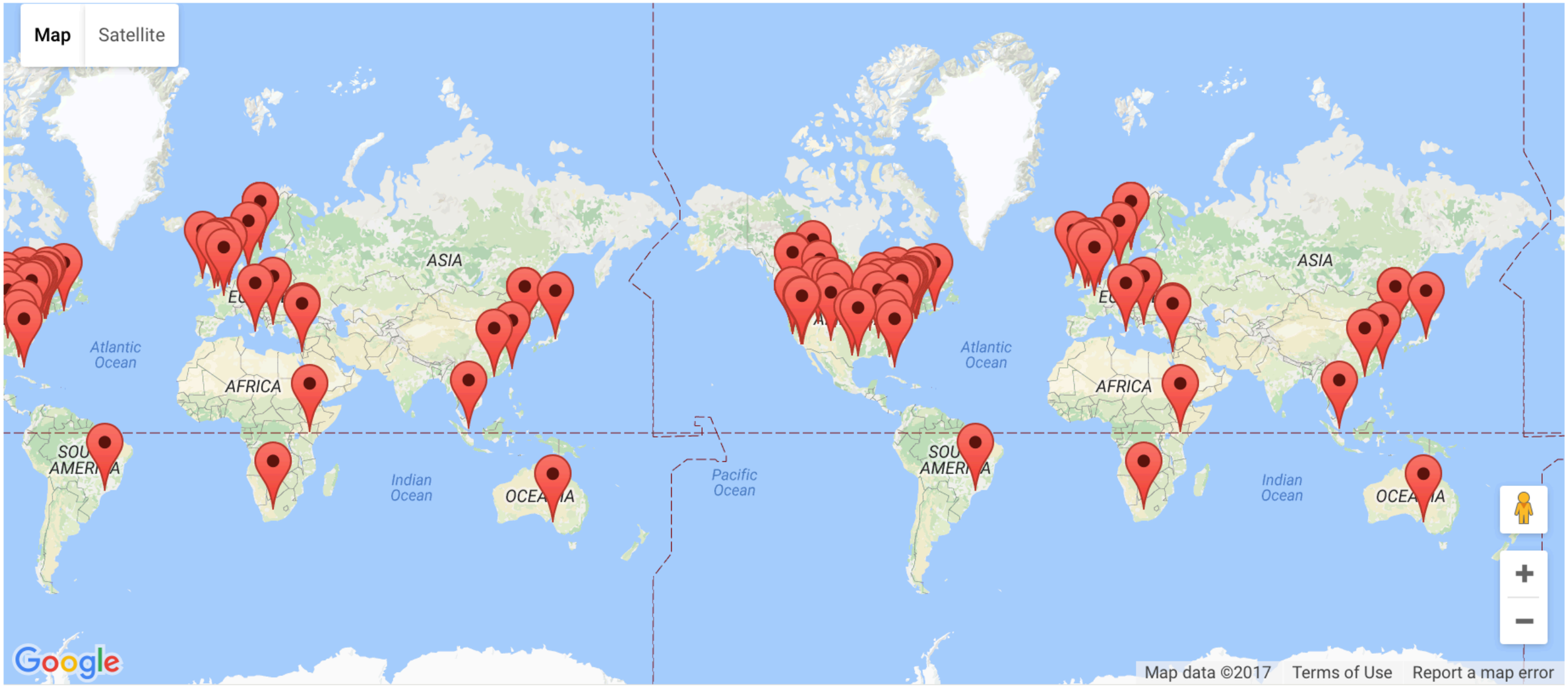
- Longitudinal data on ***millions*** of patients
  - ***diagnoses***, ***prescriptions***, ***lab reports***, ***notes***
- Sitting there in institutional IT infrastructure
- ***OHDSI*** provides a ***unified model*** to **access** data ***across institutions***, *enhancing the scientific process*!

OHDSI

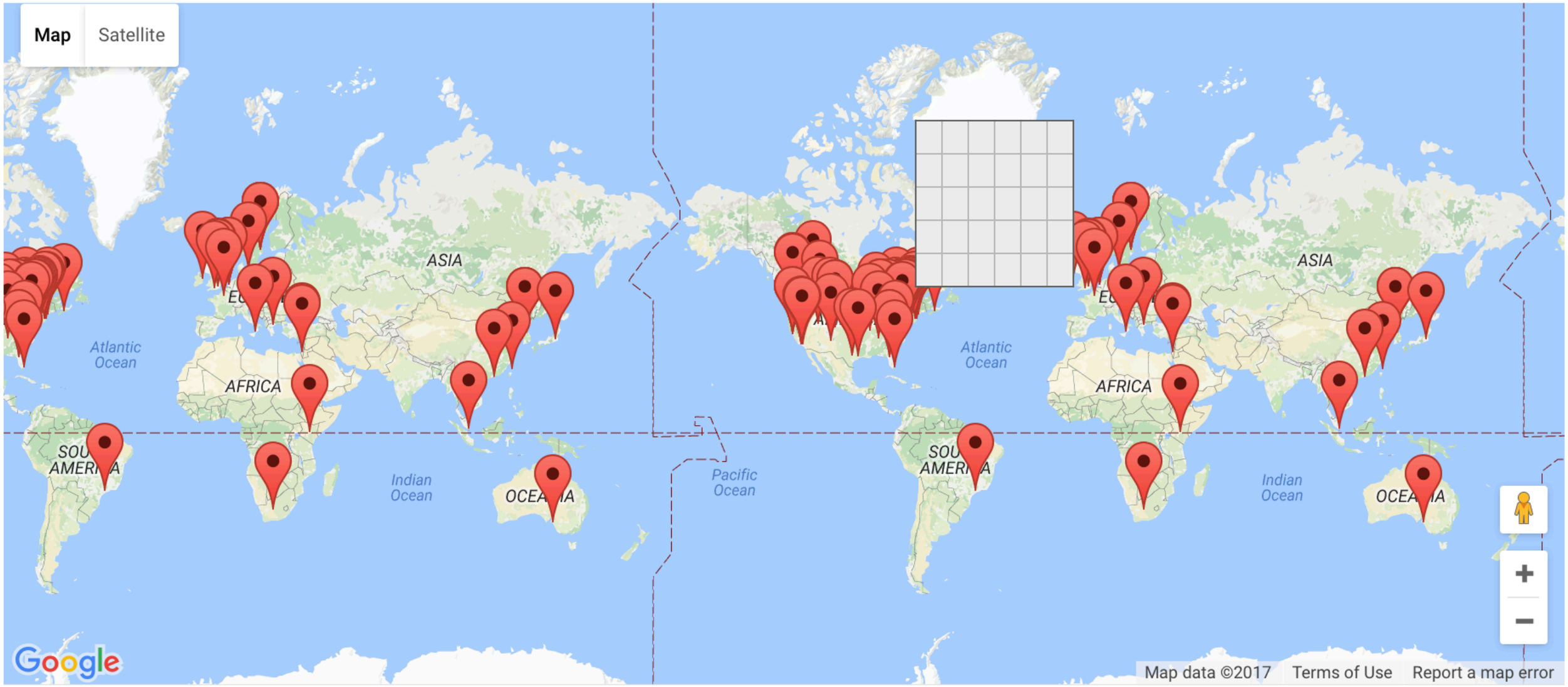
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Noémie Elhadad, PhD
Columbia

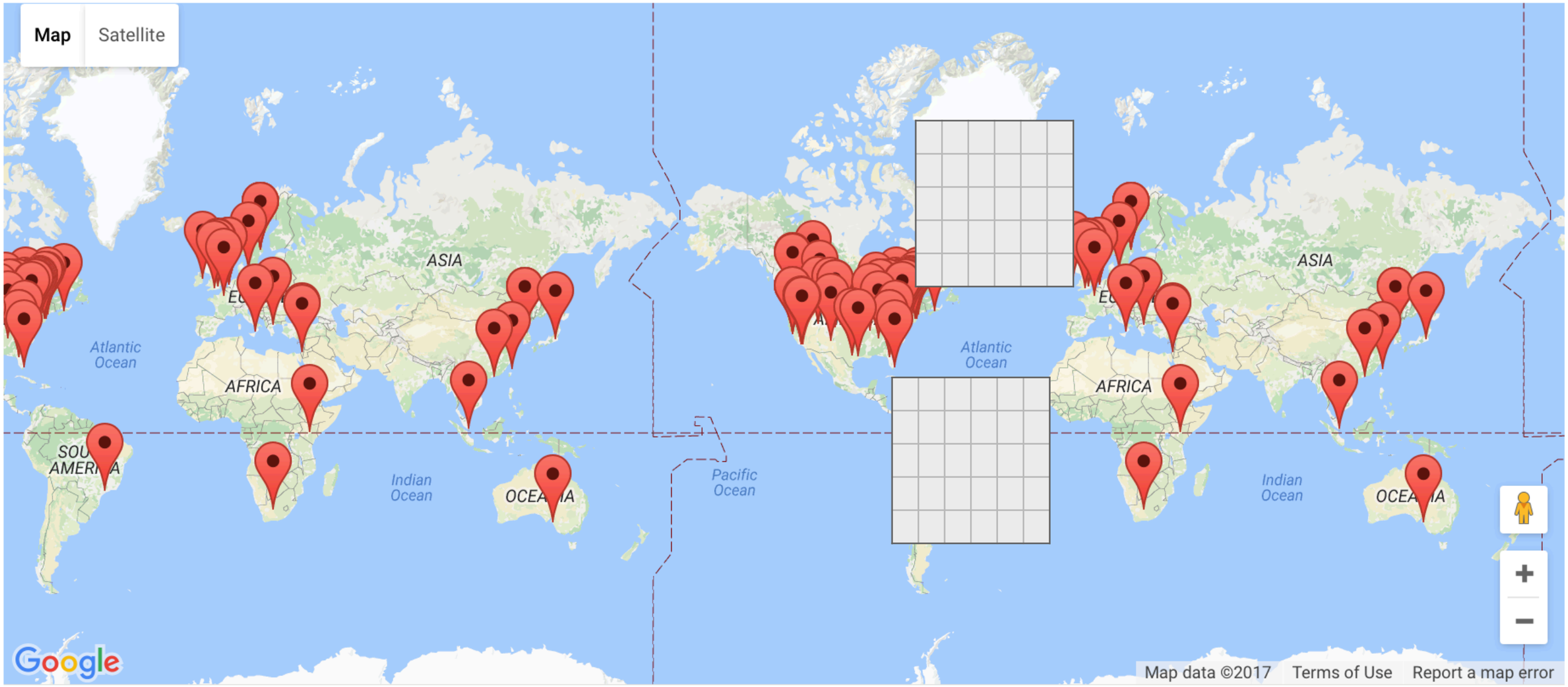
Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



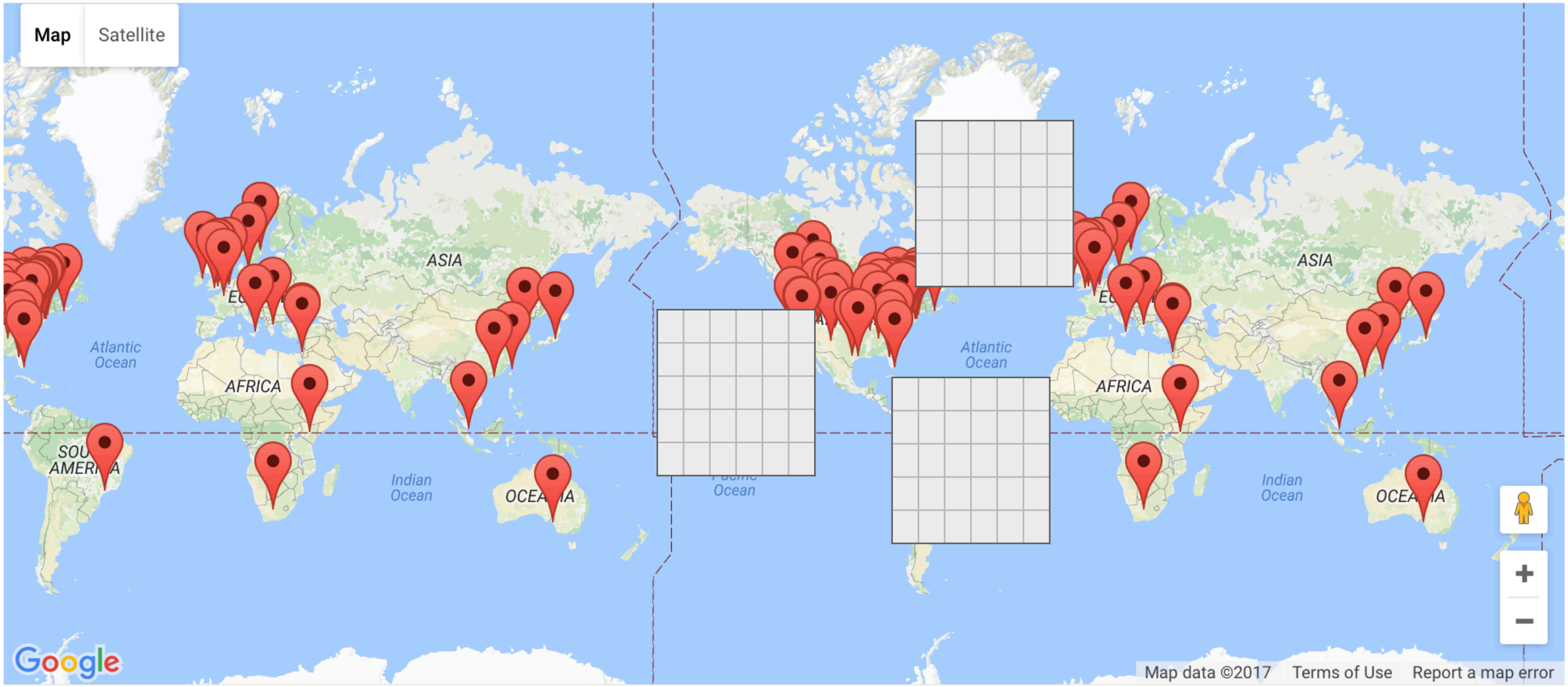
Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



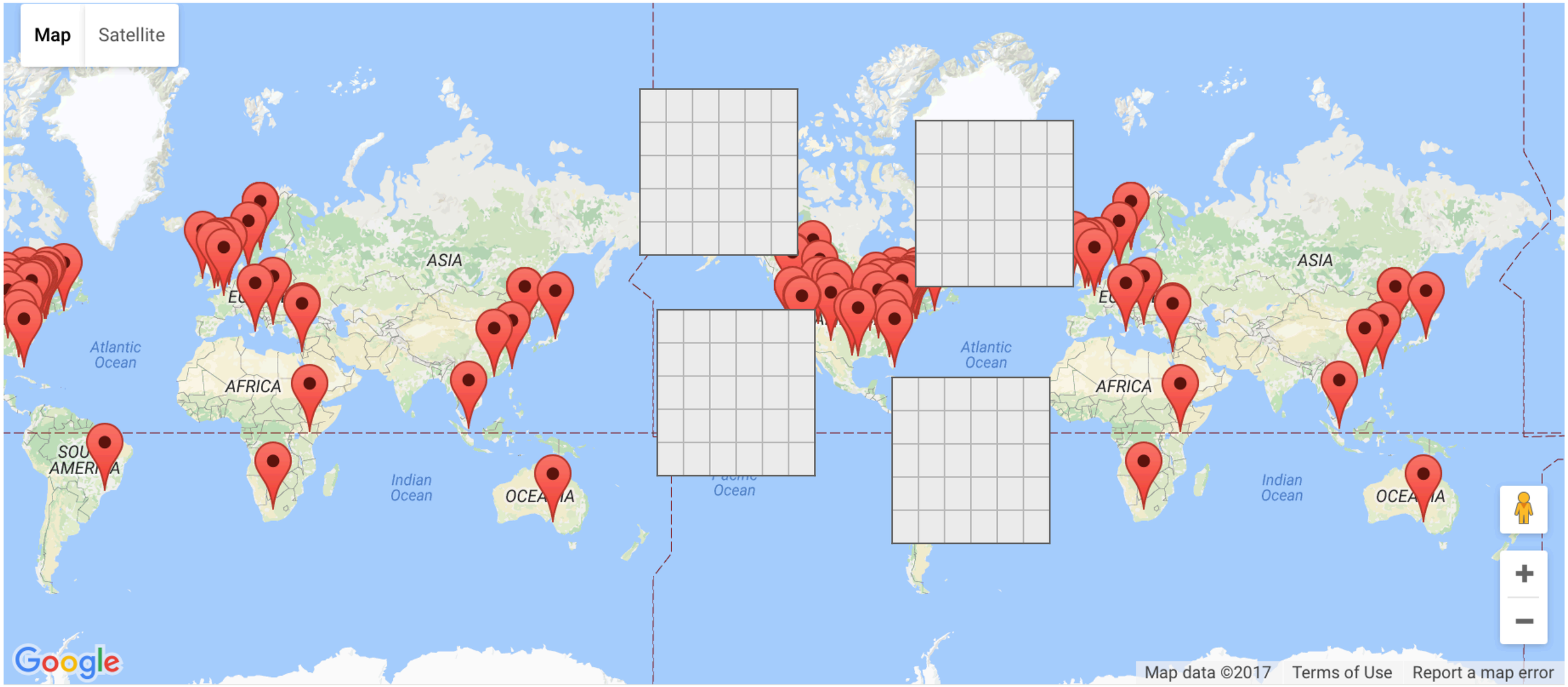
Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



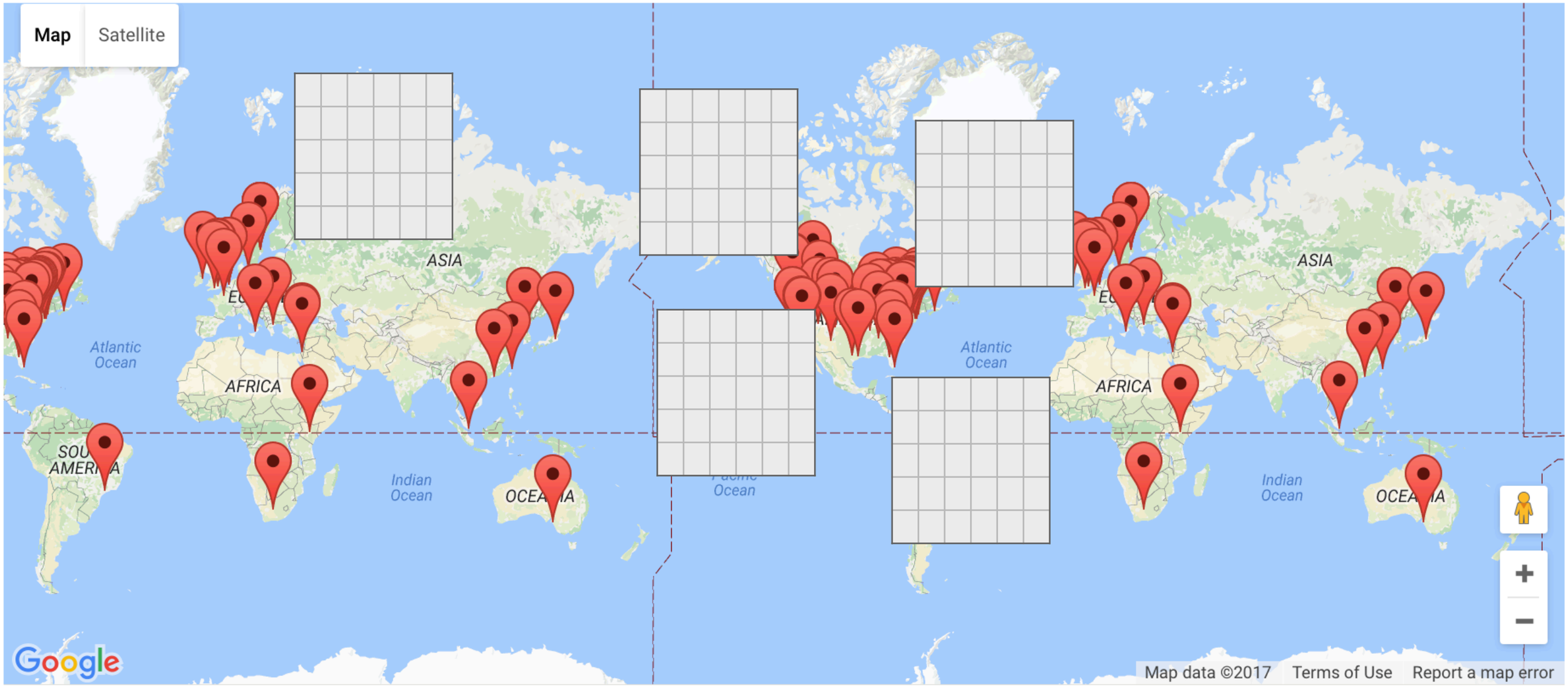
Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



Capitalize on ***digitalized health record data***
(from around the world)!
High-powered dataset(s) for discovery.



Where do we get ***environmental (E)*** data?

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,
- **Socio-Economic**

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,
- **Socio-Economic**
  - *US Census* American Community Survey (ACS)

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,
- **Socio-Economic**
  - *US Census* American Community Survey (ACS)
- **Epidemiological**

Examples of sources of disparate ***external*** exposome datasets available in the ***Exposome Data Warehouse***

- **Geological**
  - *NASA* - Cloud and Atmosphere Profiles
  - *NOAA* Climate Data
- **Pollution**
  - *EPA* Air Quality Surveillance Data Mart, or *AirData*,
- **Socio-Economic**
  - *US Census* American Community Survey (ACS)
- **Epidemiological**
  - *CDC* Wonder, *USDA* Food Atlas

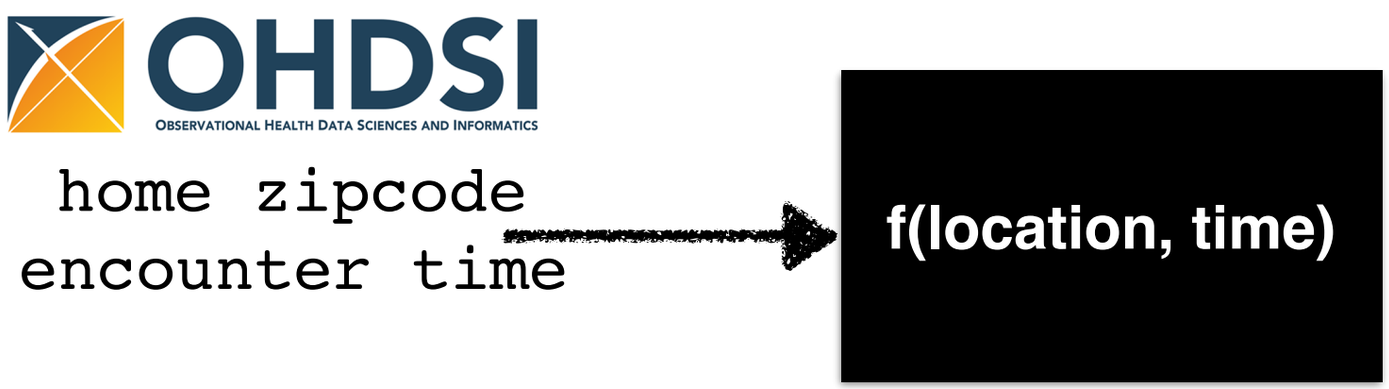
A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***

A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

```
home zipcode
encounter time
```

A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***



A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***

OHDSI
Observational Health Data Sciences and Informatics

home zipcode
encounter time

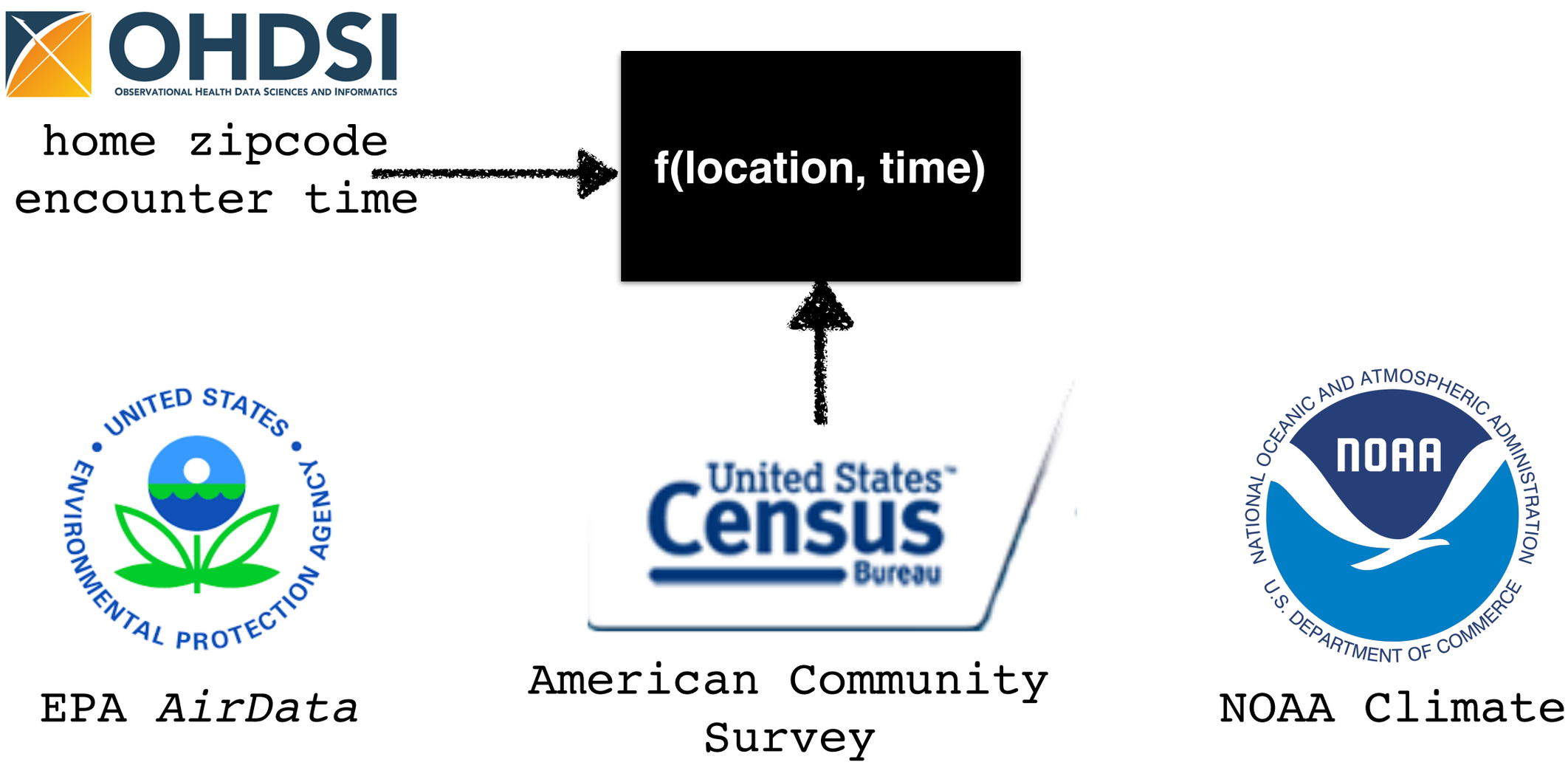
→ **f(location, time)**

EPA *AirData*

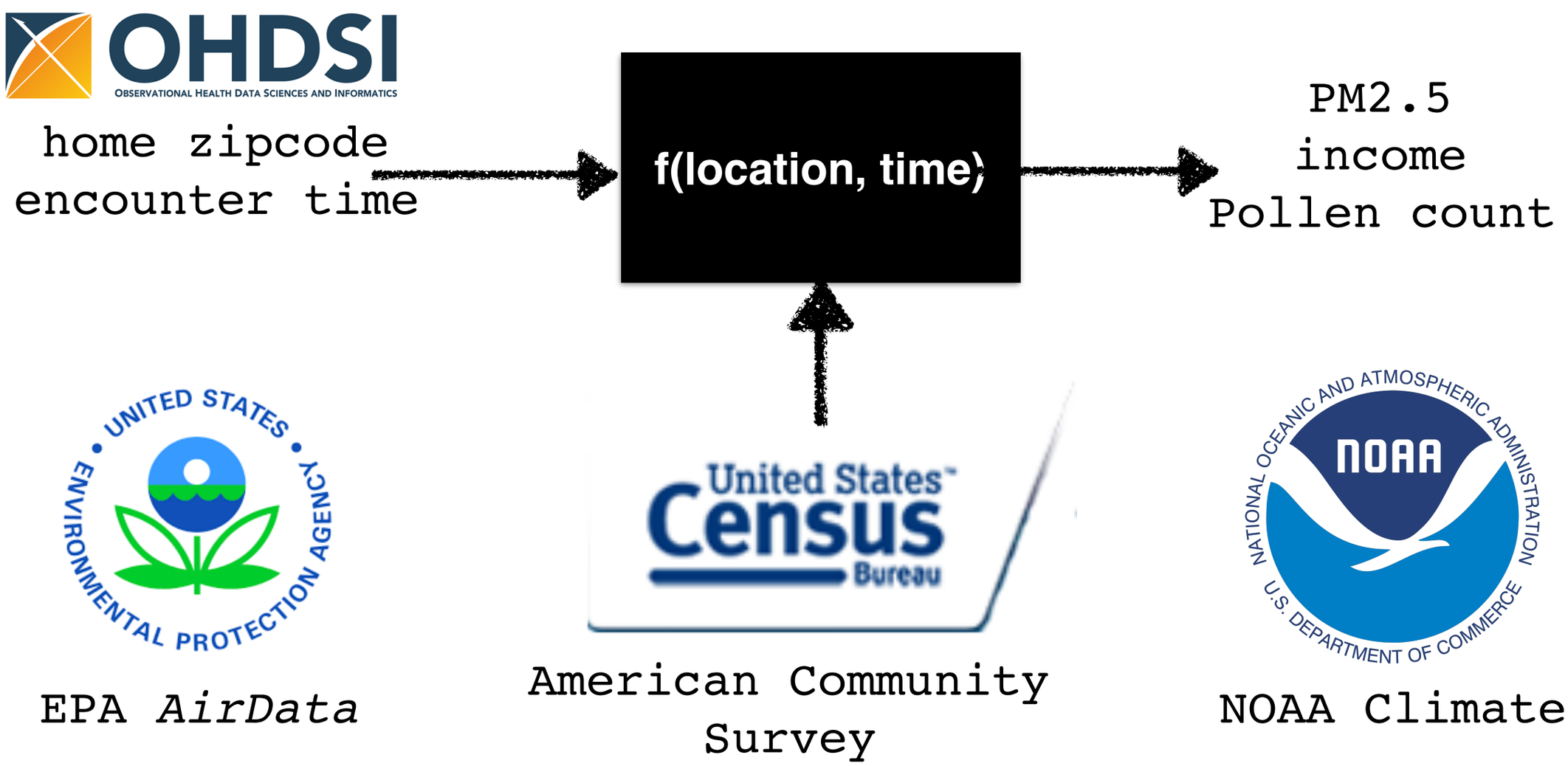
American Community Survey

NOAA Climate

A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***



A key challenge:
mashing up ***Exposome Data Warehouse*** with patient data from ***OHDSI***



OHDSI

millions of patients

	age	**sex**	**E?**	***Time(E)***	***zip***
individual$_1$	21	F	no	12/11/2015	02215
individual$_2$	35	M	yes	1/1/2016	95376
individual$_3$	75	M	yes	3/5/1998	02124
..					
..					
..					
..					
...					
individual$_n$					

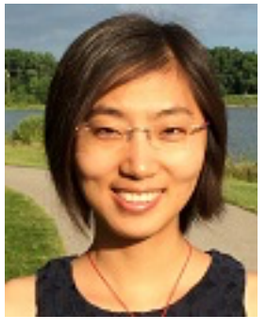
millions of patients

	OHDSI					EPA		NOAA		Census	
	age	**sex**	**E?**	***Time(E)***	***zip***	**Pb**	**PM2.5**	**Temp**	**Wind**	**hh income (K)**	**gini index**
$individual_1$	21	F	no	12/11/2015	02215	23	50	20	0	50	0.20
$individual_2$	35	M	yes	1/1/2016	95376	10	23	70	15	100	0.1
$individual_3$	75	M	yes	3/5/1998	02124	0	3	55	30	30	0.50
..											
..											
..											
..											
...											
$individual_n$											

Will it work? yep!

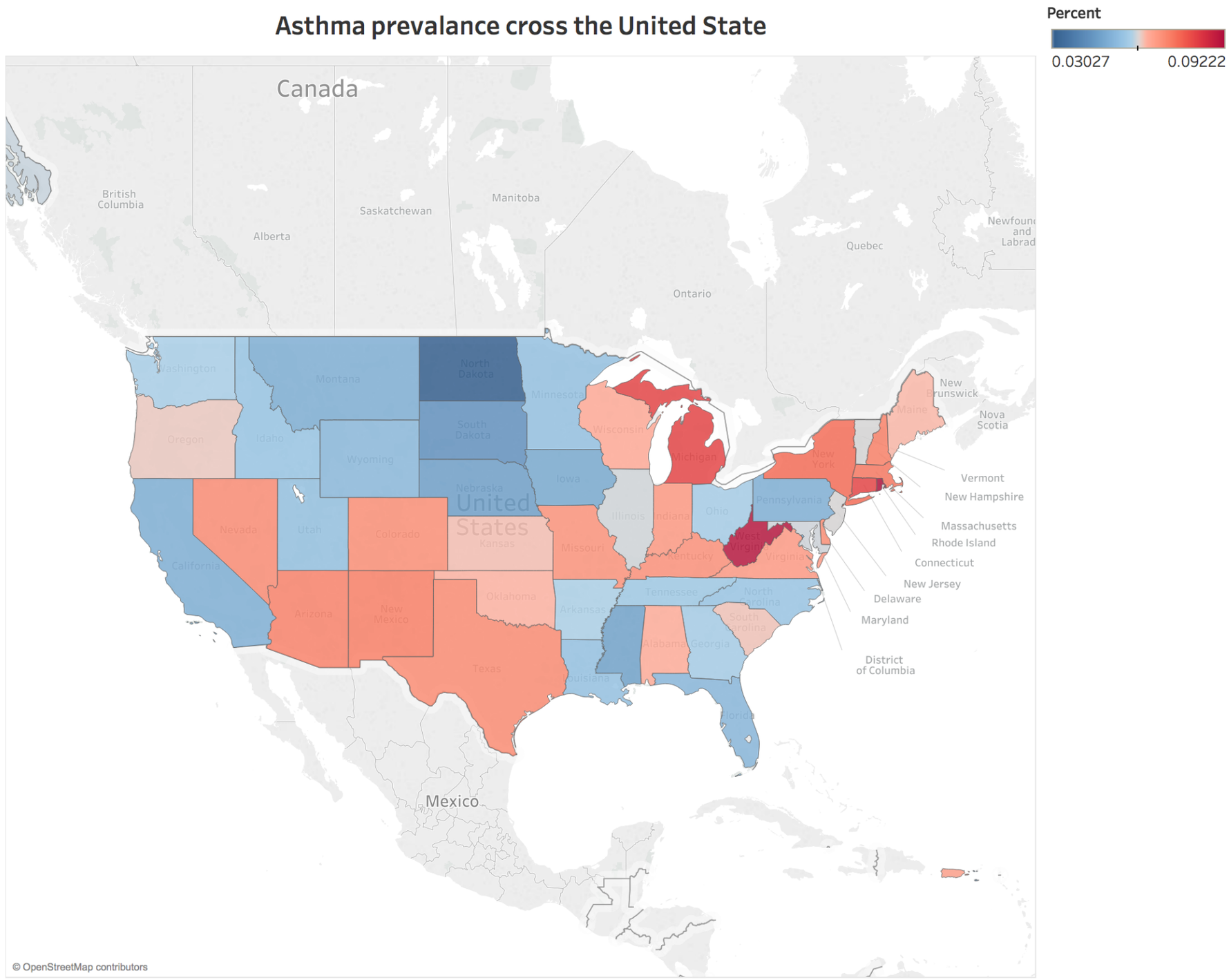
Does temperature (and weather) influence asthma-related pediatric ER visits?

- **Children <= 17 y/o** with >=1 ICD9 code corresponding to 493.*
- **N=56K, >84K ER visits**
- Weather station data
  - (daily **temperature**, **wind**, **humidity**)
- **Case-crossover design** (only investigated cases)



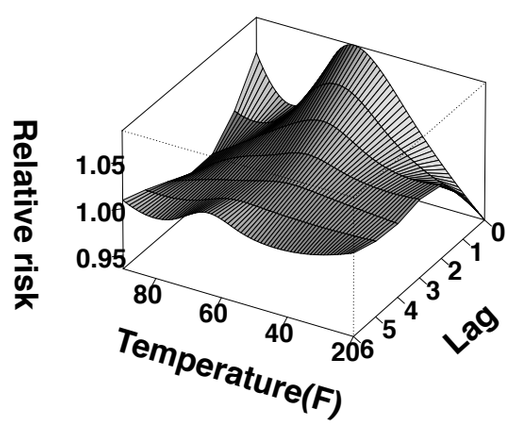
Yeran Li, PhD (MS, HSPH)
Chirag Lakhani, PhD (HMS)
Yun Wang, PhD (Post-doc, HSPH)

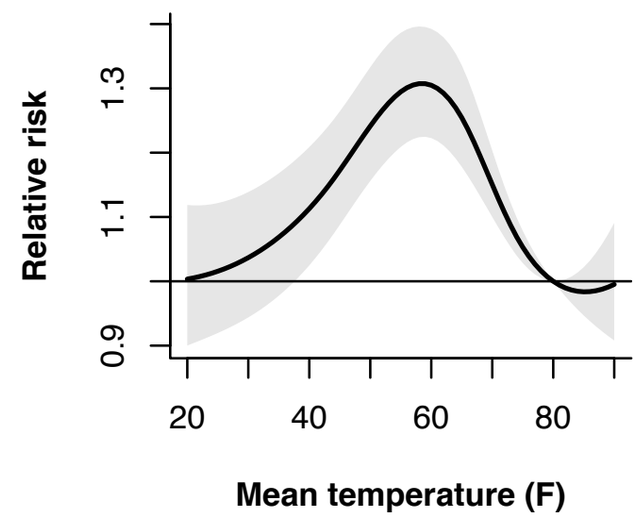
Prevalence of asthma attack varies across the US

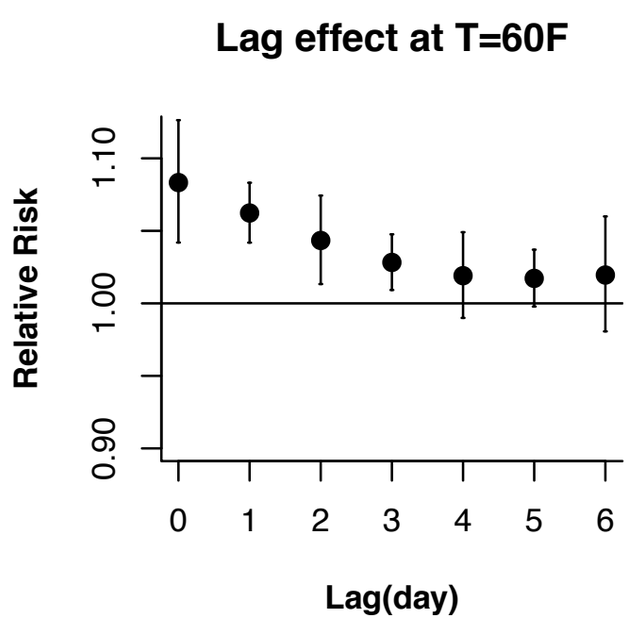


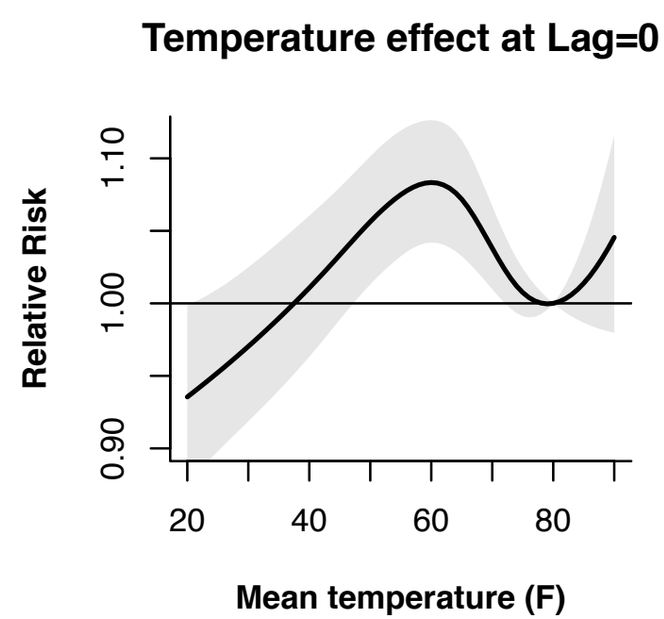
Does temperature influence asthma ER visits?: *yes*!

Relative risk of asthma attack by mean temperature

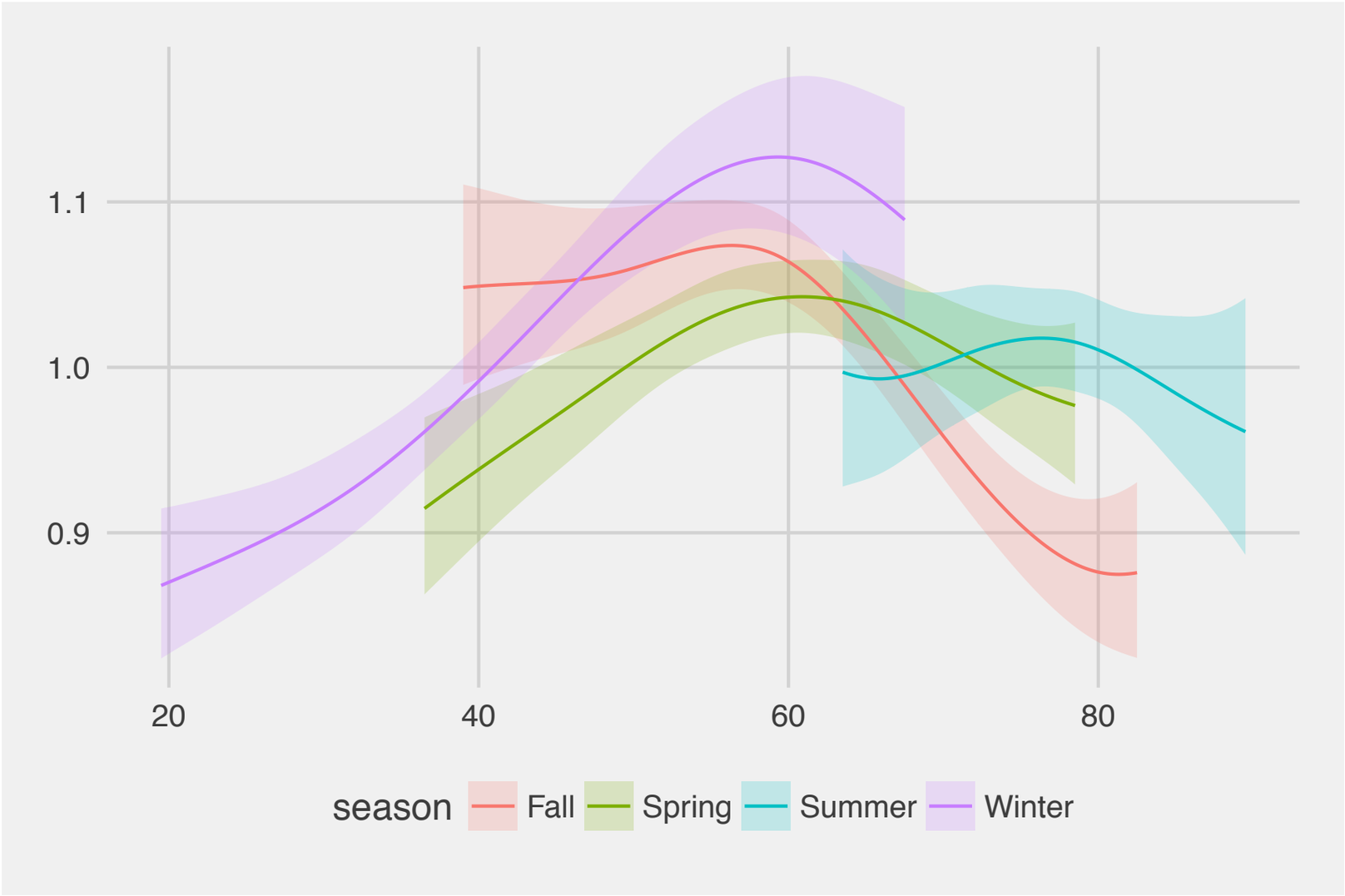




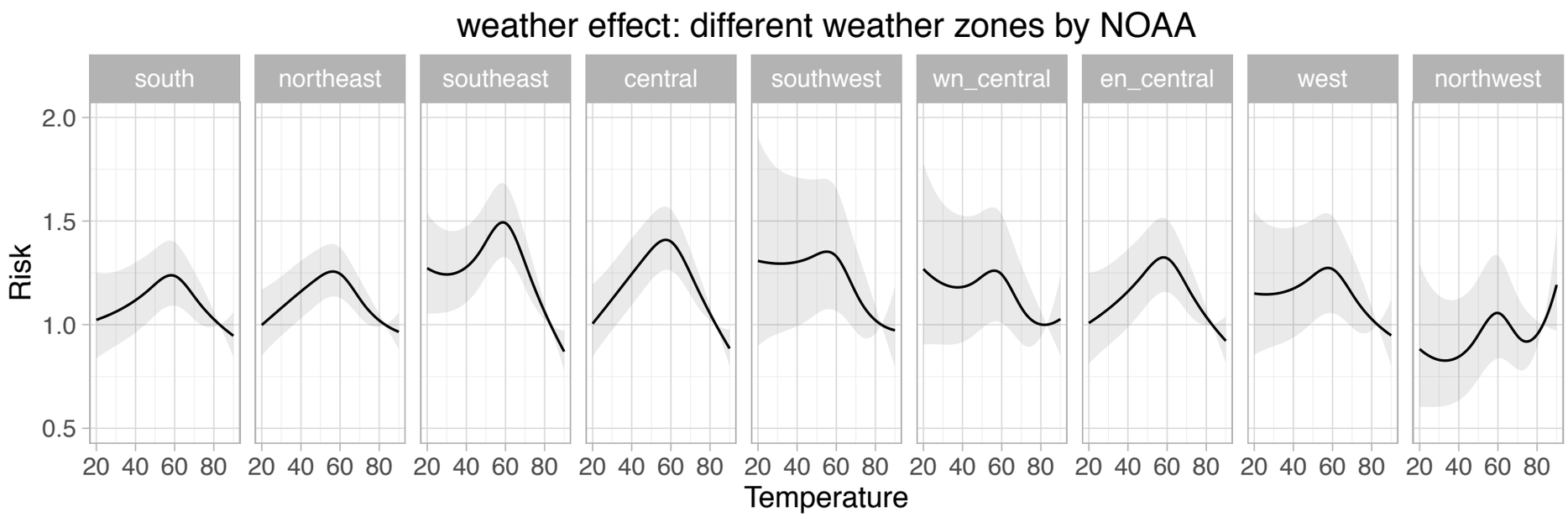


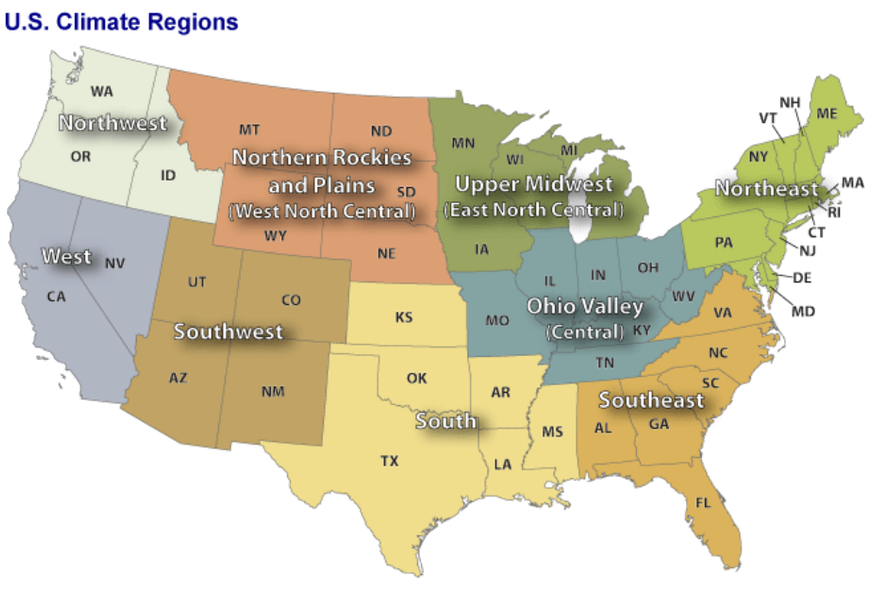


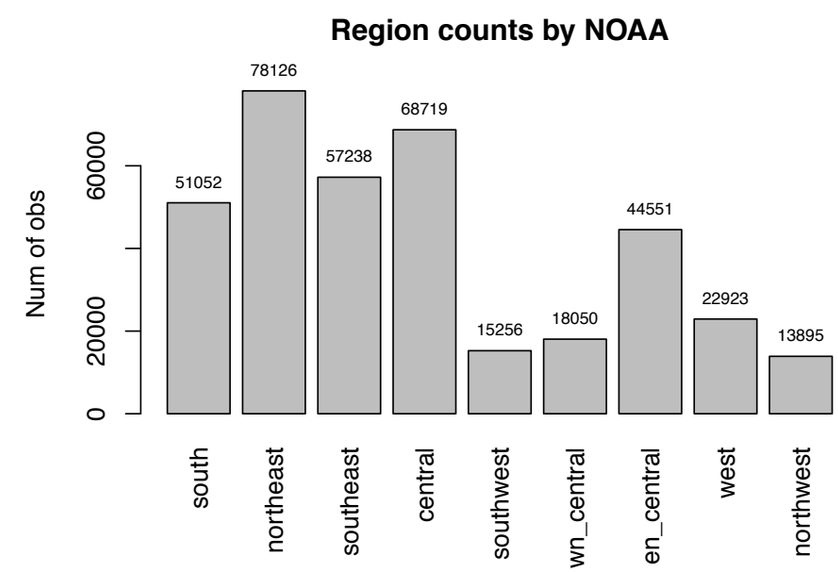
Rates of asthma attacks depend on season?: ***yes!***



Rates of asthma attacks dependent on region?: ***yes!***







https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php

weather.com

The Weather Channel | 47 ° New York, NY | 53 ° Brookline Village,... | US | °F | Search city, zip or place

WEATHER | MAPS | SEVERE | VIDEO & PHOTOS | ACTIVITIES | HEALTH | TRAVEL | SIGN UP FOR NEWSLETTER | Login / Sign Up

Today | Hourly | 5 Day | 10 Day | Weekend | Maps | Monthly | More Forecasts

SPECIAL WEATHER STATEMENT Until 10:00am , Thu Feb 23 — 2

NEW YORK, NY
as of 9:28 am EST

47°
CLOUDY
feels like **46°**
H **65°** / L **53°**
UV Index **1 of 10**

Snowstorm Underway: Blizzard Conditions Expected

RIGHT NOW

Wind	SW 4 mph
Humidity	90%
Dew Point	45°
Pressure	29.90 in ↓
Visibility	4.0 mi

RADAR | CLOUDS | RADAR & CLOUDS | WEATHER IN MOTION →

NEXT 36 HOURS

HOURLY → | 10 DAYS →

TODAY	TONIGHT	FRI	FRI NIGHT	SAT
AM CLOUDS/PM SUN				
HIGH 65°	LOW 53°	HIGH 64°	LOW 53°	HIGH 59°
20%	40%	10%	10%	90%

Cloudy skies early, then partly cloudy this afternoon. High around 65F. Winds light and variable.

WIND	HUMIDITY	UV INDEX	SUN
SSW 5 mph	72%	3 of 10	↑ 6:38 am ↓ 5:40 pm

PLAN YOUR DAY

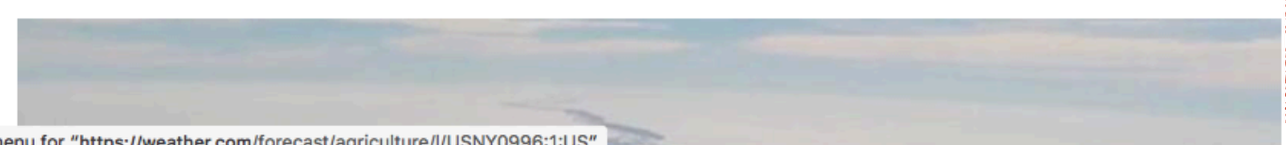
VERY GOOD — **Pollen/Breathing/Mold**
VIEW ALL ALLERGENS →
Presented by Flonase

Out of Growing Season
FARMING FORECAST →
Powered by Ram Trucks

Good Road Conditions

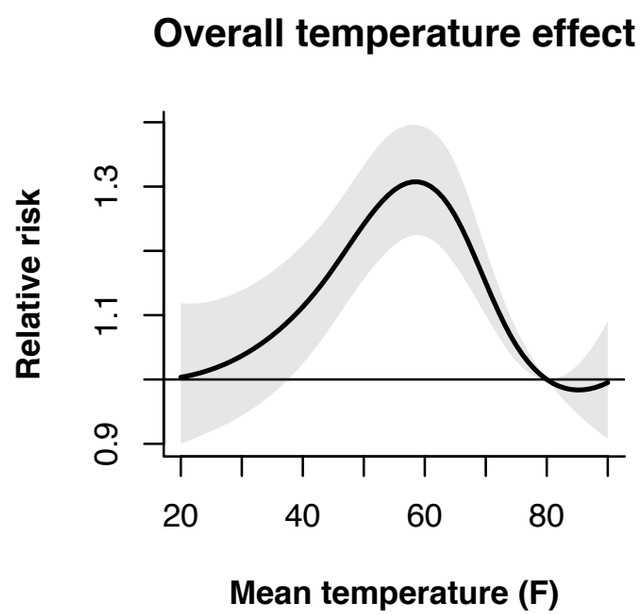
WEATHER HEADLINES

VIEW ALL →



Display a menu for "https://weather.com/forecast/agriculture/l/USNY0996:1:US"

Does temperature (and weather) influence asthma-related ER visits in kids?: the tip of the iceberg!



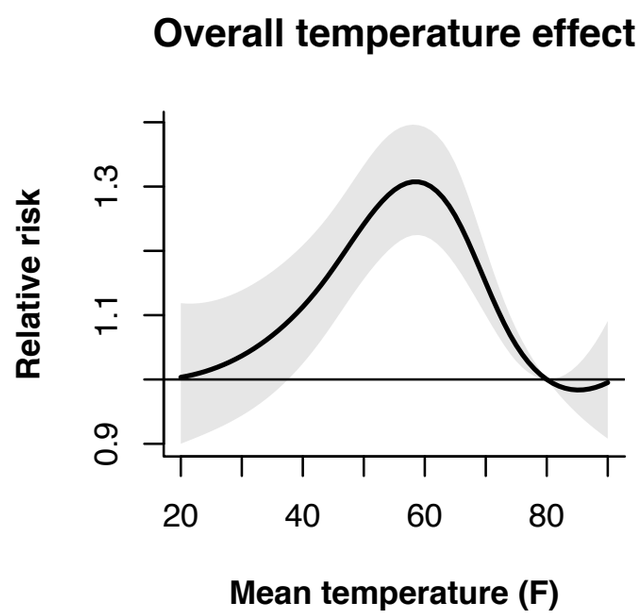
What other scientific questions?

- what is influence of **pollen**?
- what is the influence of **air pollution**?
- what about **adults**?

Can we replicate the analysis?

- different **populations**
- using different **data**
- with different **analysts**

Does temperature (and weather) influence asthma-related ER visits in kids?: the tip of the iceberg!



What other scientific questions?

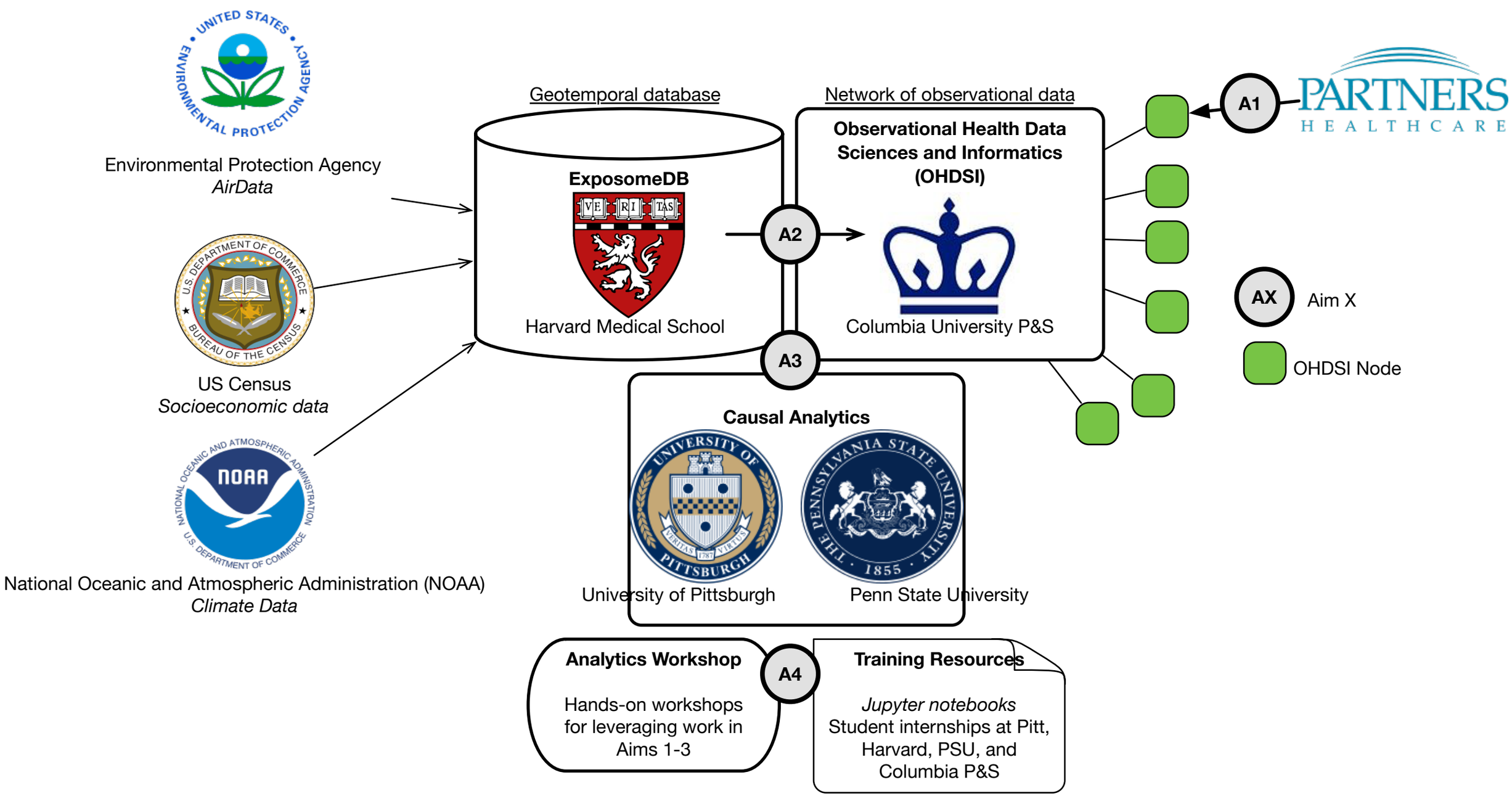
- what is influence of **pollen**?
- what is the influence of **air pollution**?
- what about **adults**?

Can we replicate the analysis?

- different **populations**
- using different **data**
- with different **analysts**

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Integrating the ***ExposomeDB*** with ***OHDSI*** and causal modeling tools to drive and demonstrate discovery.



Many hypotheses are possible to address: useful for **public health** & **planning**!

Many hypotheses are possible to address: useful for **public health** & **planning**!

How does ***socioeconomic*** context influence **hospital use**, **disease rates**, and **recovery**?

Many hypotheses are possible to address: useful for **public health** & **planning**!

How does ***socioeconomic*** context influence **hospital use**, **disease rates**, and **recovery**?

What is the effect of ***air pollution*** levels in ***disease*****?**

Many hypotheses are possible to address: useful for **public health** & **planning**!

How does ***socioeconomic*** context influence **hospital use**, **disease rates**, and **recovery**?

What is the effect of ***air pollution*** levels in ***disease*****?**

Do adverse **weather** conditions influence **hospital use**?

Many hypotheses are possible to address: useful for **public health** & **planning**!

How does ***socioeconomic*** context influence **hospital use**, **disease rates**, and **recovery**?

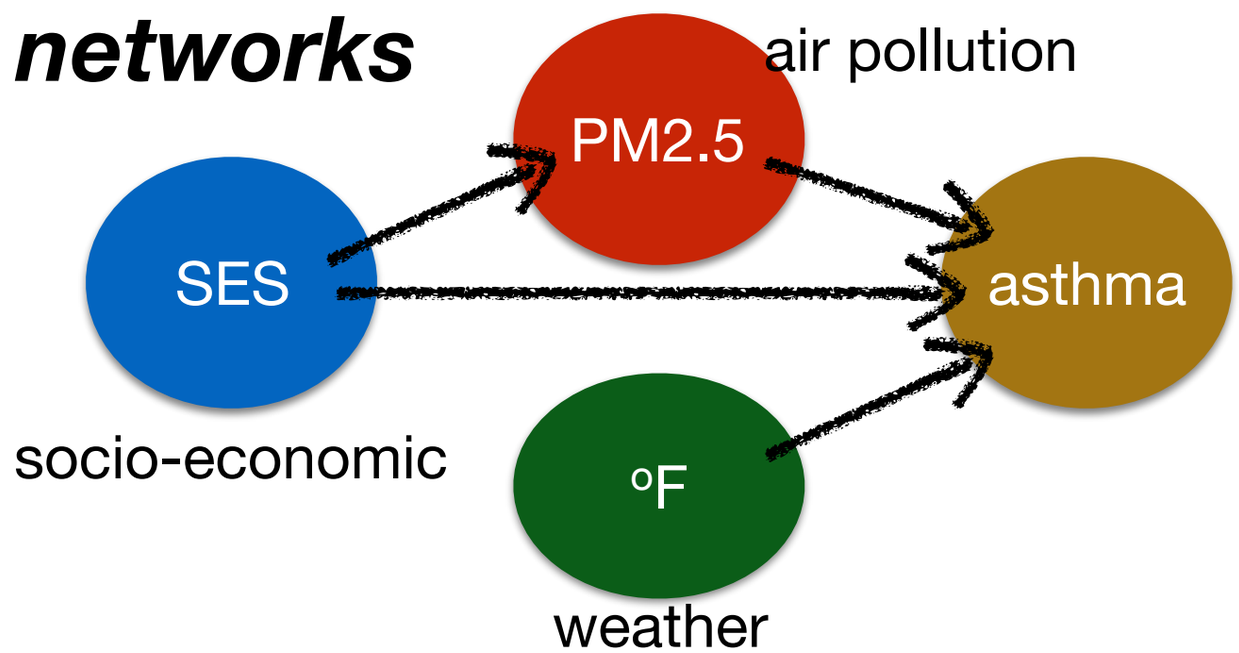
What is the effect of ***air pollution*** levels in ***disease*****?**

Do adverse **weather** conditions influence **hospital use**?

What pharmaceutical **drugs** lead to **adverse health outcomes**?

We will harness tools in machine learning extract ***signal from noise***!

bayesian networks



Greg Cooper, MD, PhD
Pittsburgh

case-crossover

Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study

Chirag J. Patel[1], Jianguang Ji[2], Jan Sundquist[2], John P. A. Ioannidis[3] & Kristina Sundquist[2]

Sci Rep 2016

Vasant Honavar, PhD
Penn State

ccd.pitt.edu

Center for Causal Discovery

About Research Education Tools Wiki People Calendar News

Causal Web

our user-friendly web application for performing causal discovery analysis on big data using large memory servers at the Pittsburgh Supercomputing Center. Use this software if you want to quickly try out a causal discovery algorithm or if you have big data that cannot be analyzed on your local hardware.

The software currently includes:

- Fast Greedy Equivalence Search (FGES) for continuous variables
- Fast Greedy Equivalence Search (FGES) for discrete variables
- Greedy Fast Causal Inference (GFCI) algorithm for continuous variables

User guide Web app

Causal Command

a Java library and command line implementation of algorithms for performing causal discovery on big data. Use this software if you are interested incorporating analysis via a shell script or in a Java-based program. The 'Software' button below leads to a comprehensive repository. Choose the 'causal-cmd-x.x.x -jar-with-dependencies.jar' from the downloads list when using this as an executable via the command line or as an API in a Java program.

The software currently includes:

- Fast Greedy Equivalence Search (FGES) for continuous variables
- Fast Greedy Equivalence Search (FGES) for discrete variables
- Greedy Fast Causal Inference (GFCI) algorithm for continuous variables

User guide Software

Display a menu

http://www.ccd.pitt.edu/tools/

Integrating the ***ExposomeDB*** with ***OHDSI*** and **analytics** to drive and demonstrate discovery **<u>by the community</u>!**

(trainees especially welcome)

Integrating the ***ExposomeDB*** with ***OHDSI*** and **analytics** to drive and demonstrate discovery **<u>by the community</u>!**
(trainees especially welcome)

- 2-day *hands-on* workshop in New York or Boston

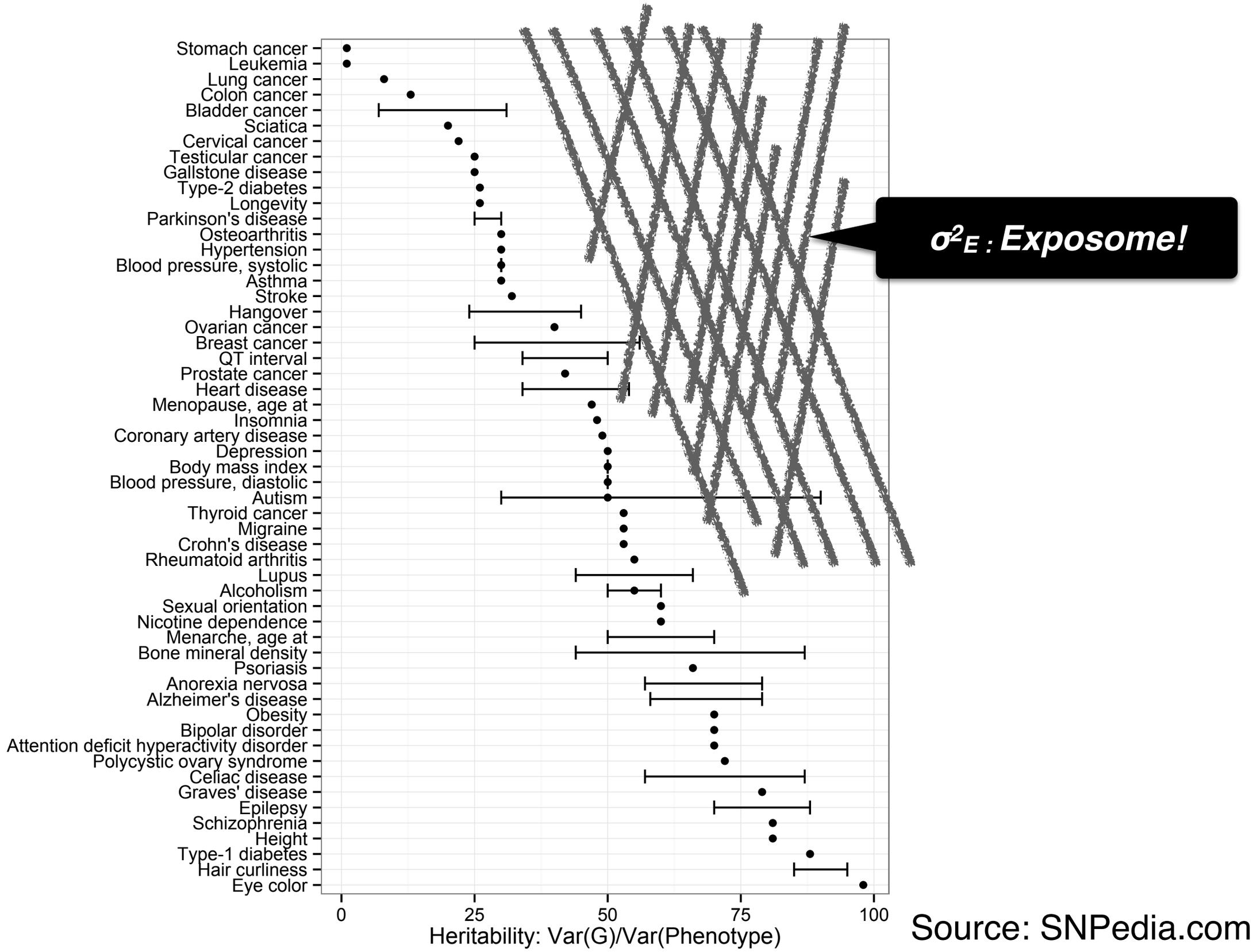
Integrating the ***ExposomeDB*** with ***OHDSI*** and **analytics** to drive and demonstrate discovery **<u>by the community</u>!**
(trainees especially welcome)

- 2-day *hands-on* workshop in New York or Boston
- remote “exchange” internship program and 2-week immersion

Integrating the ***ExposomeDB*** with ***OHDSI*** and **analytics** to drive and demonstrate discovery **<u>by the community</u>!**
(trainees especially welcome)

- 2-day *hands-on* workshop in New York or Boston
- remote “exchange” internship program and 2-week immersion
- dissemination of electronic training resources

Many hypotheses are possible to address: useful for **can we build a machine learning predictor to estimate *E*?**



Source: SNPedia.com

Thanks

RagGroup
Chirag Lakhani
Yeran Li
Shreyas Bhave
Rolando Acosta

Harvard DBMI
Isaac Kohane
Susanne Churchill
Nathan Palmer
Sunny Alvear
Michal Preminger

Noémie Elhadad (Columbia)
Vasant Honavar (PSU)
Greg Cooper (Pitt)
George Hripcsak (Columbia)

René Baston
Katie Naum
Kathleen McKeown

NIEHS National Institute of Environmental Health Sciences

NIH Common Fund
Big Data to Knowledge

Agilent Technologies

PhRMA RESEARCH • PROGRESS • HOPE

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

Chirag J Patel
chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org