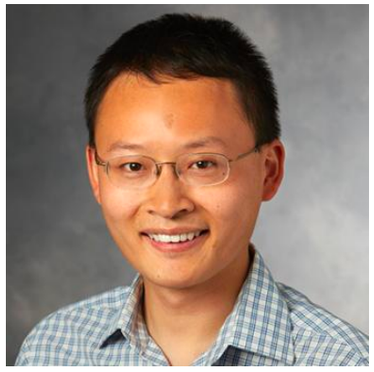
BIODS215

Topics in Biomedical Data Science: Large-scale inference

Winter Quarter 2020

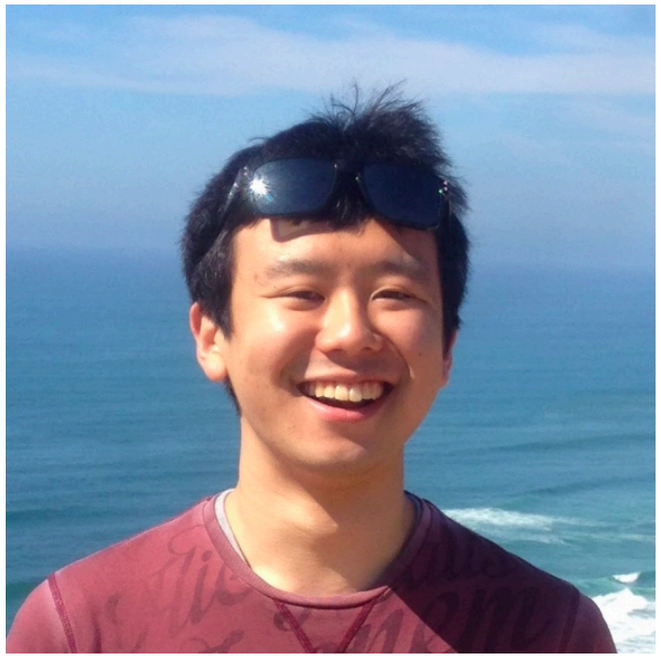
Course Instructors

Prof. Manuel A. Rivas
MSOB X321
mrivas@stanford.edu
rivaslab.stanford.edu



Prof. James Zou
MSOB X325
jamesz@stanford.edu
https://sites.google.com/site/jamesyzou/

Teaching Assistant



Yosuke Tanigawa
MSOB 3rd floor
ytanigaw@stanford.edu
rivaslab.stanford.edu

Lecture structure

~20-45 minutes motivating biomedical example

~30-55 minutes statistical inference concept lecture

5-7 minute break in the middle

Course requirements and grading

Two homework assignments (40%)

Final project (50%)

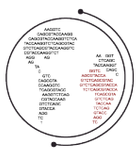
Class participation (10%)

Announcements from TA

- Canvas access
  If you don't have access to Canvas, please see Yosuke

- Gradescope for assignments
  https://gradescope.com/
  Entry code: MZGP3Y

- TA Office hour
  Tuesdays 4:30-5:20pm at MSOB x321



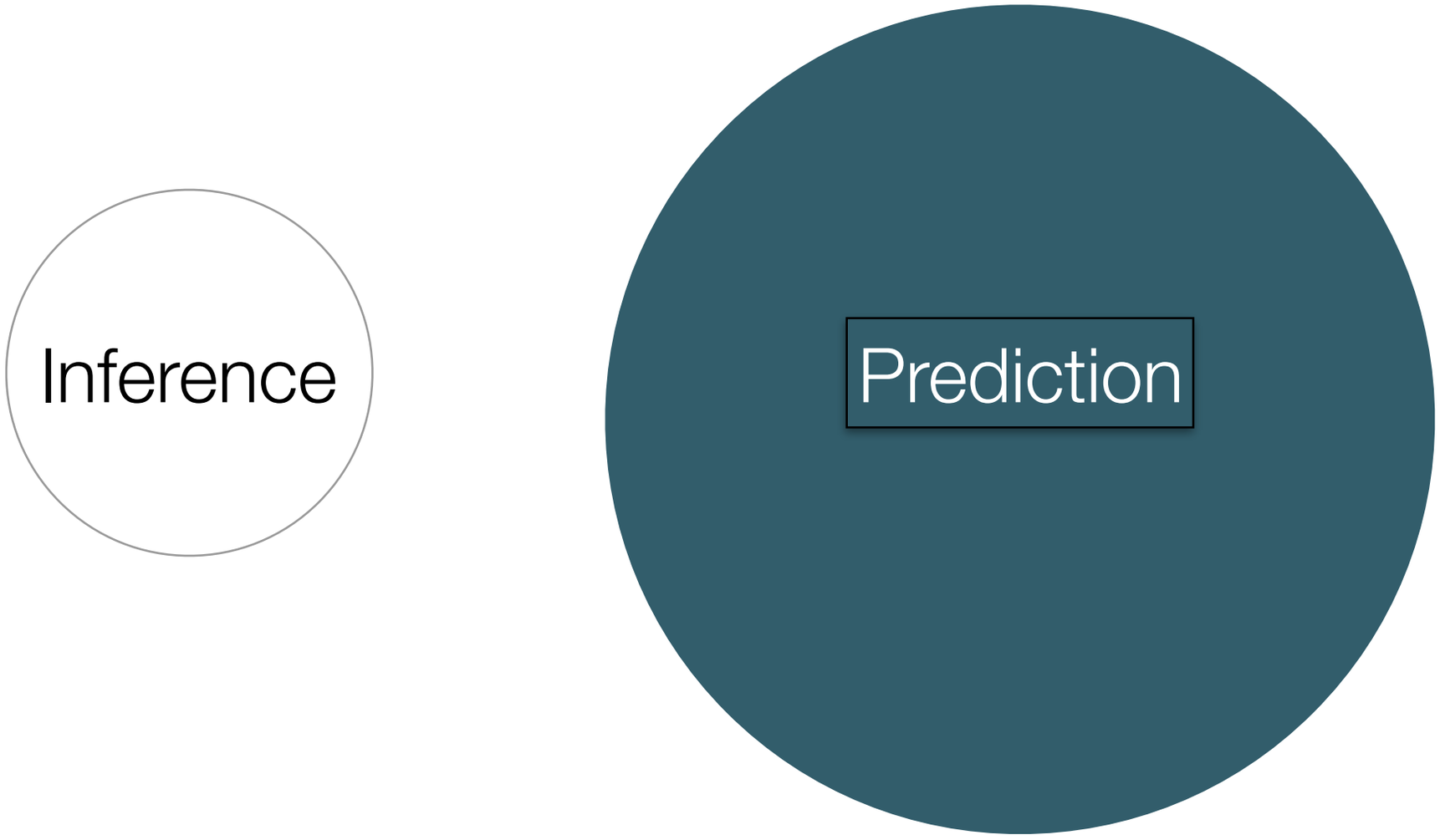
RIVASLAB

Data explosion and worldview across fields

Inference: To [infer] how nature is associating the response variables to the input variables

Statisticians, Biomedicine (therapeutics)

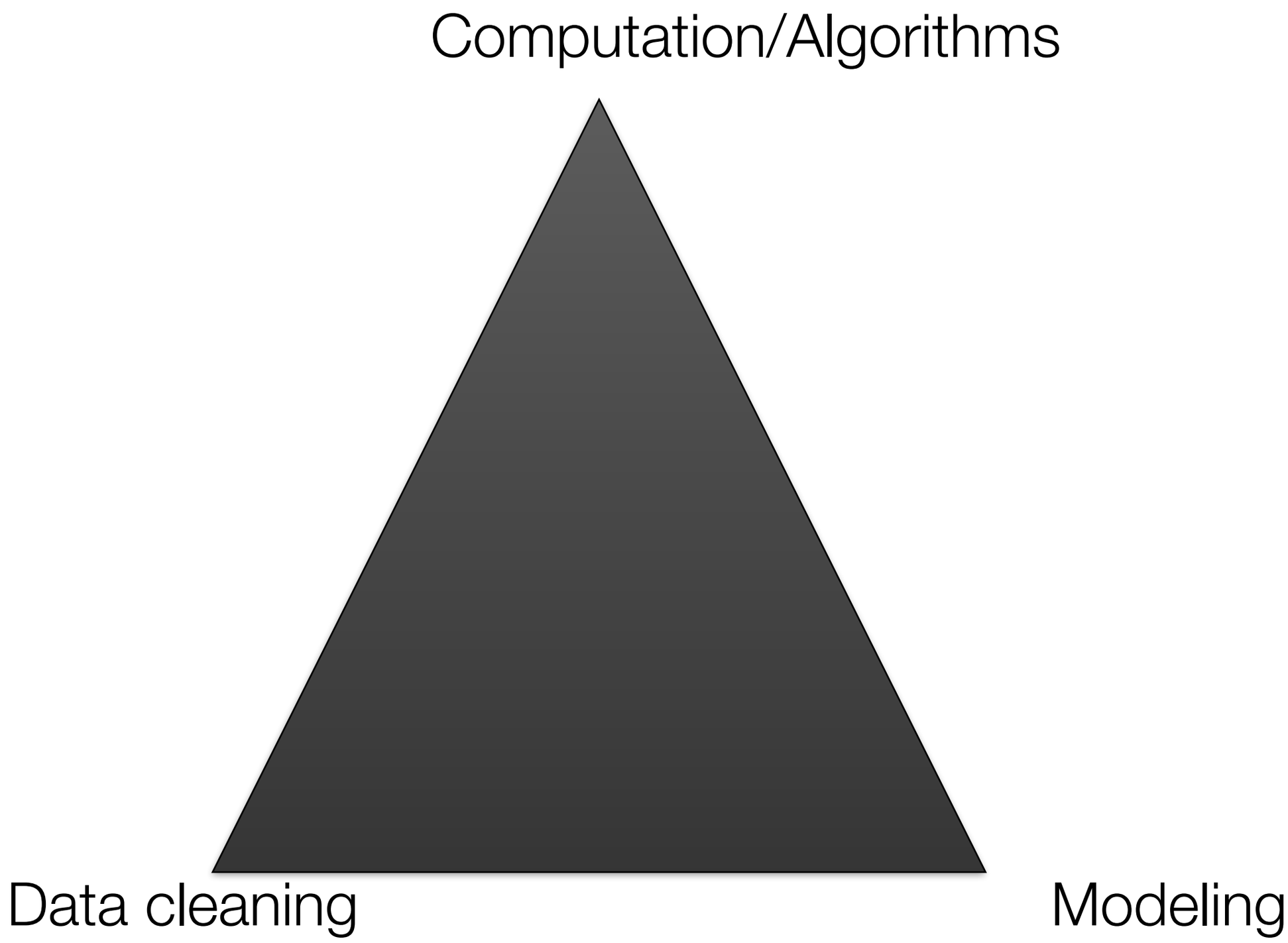
Data explosion and worldview across fields



Prediction: To be able to predict what the responses are going to be to future input variables

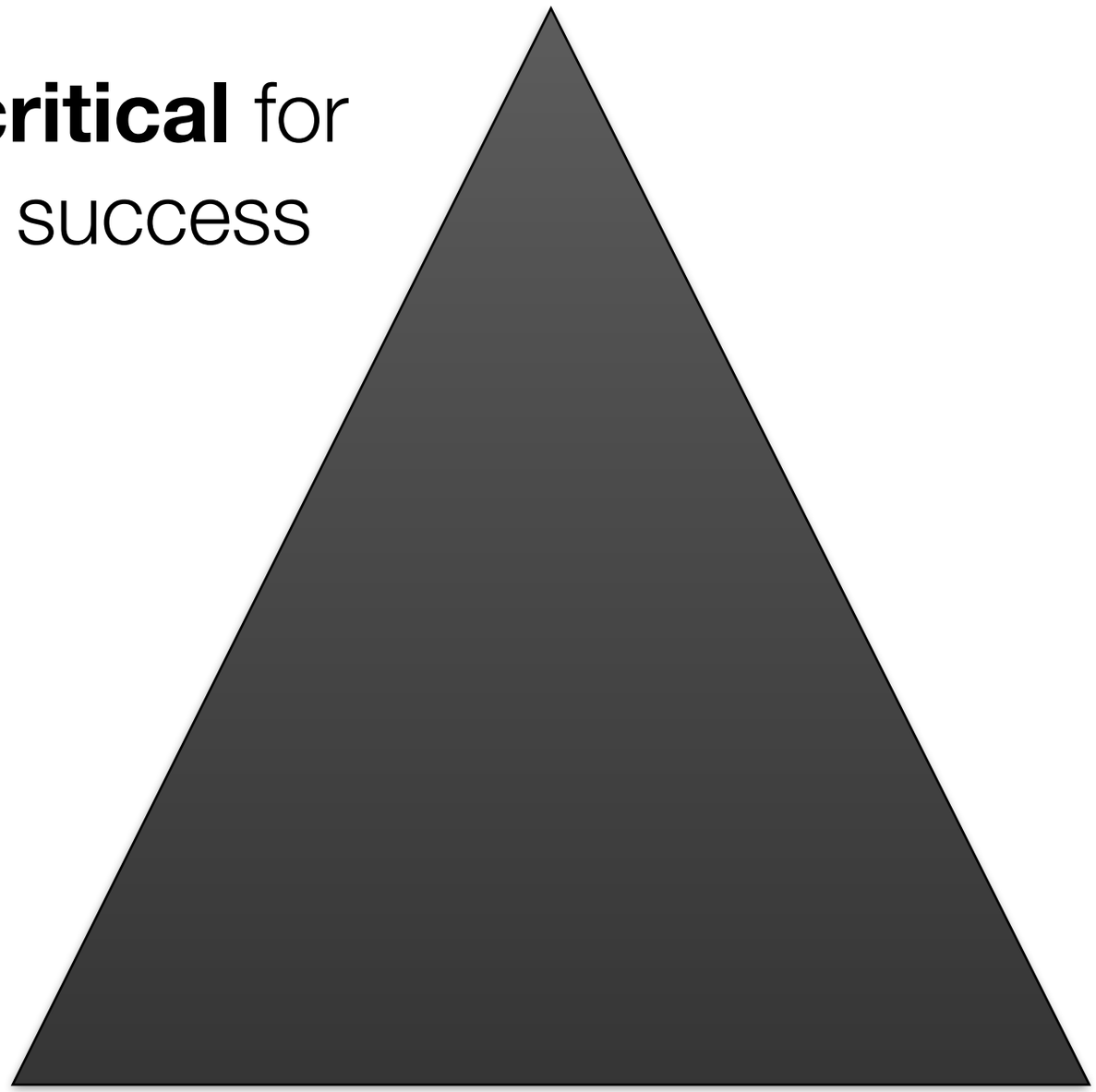
Machine Learning, Computer science

Learning objectives



Learning objectives

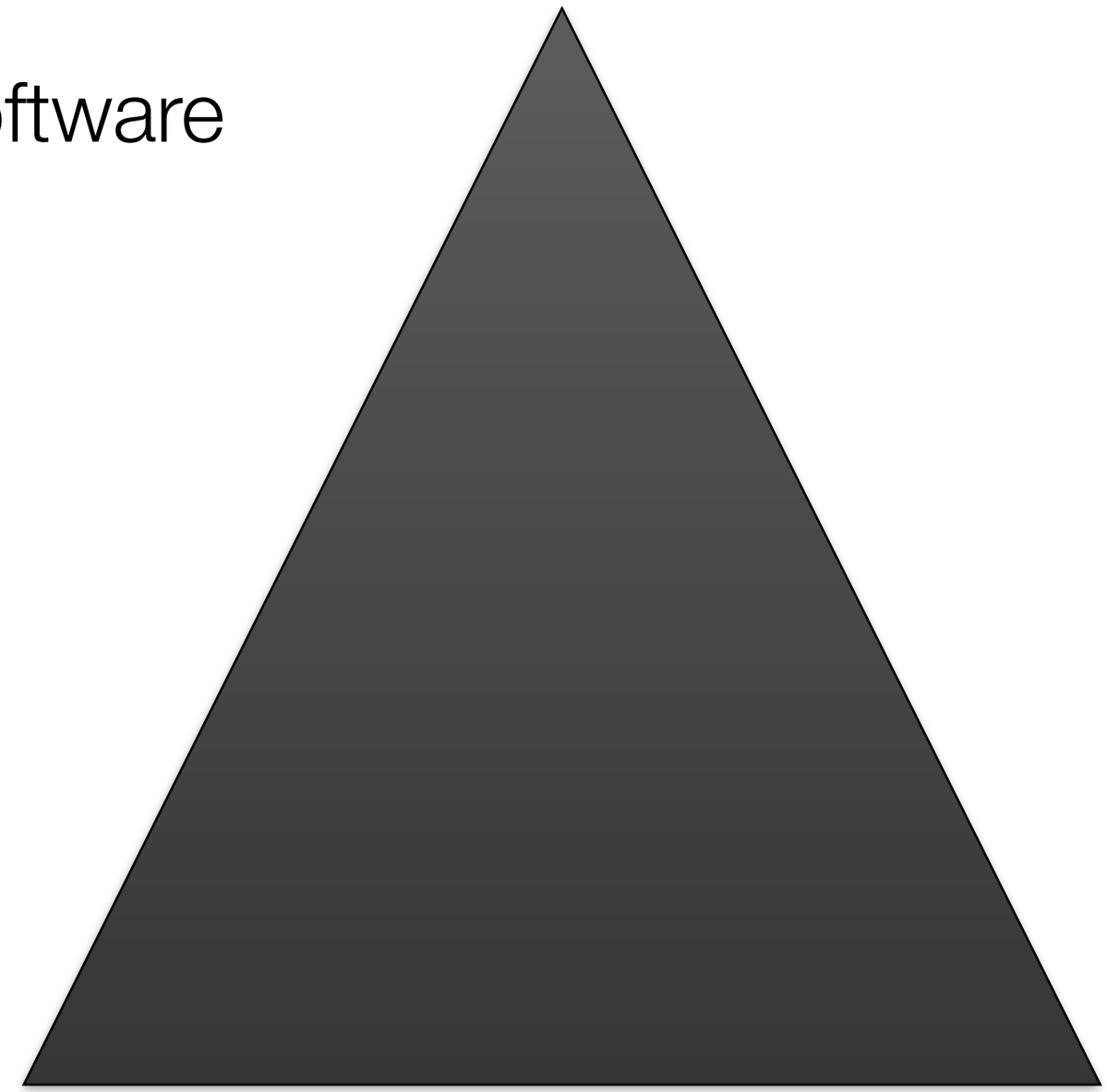
Not fun, but **critical** for downstream success



Learning objectives

Algorithms: software and hardware

Computation/Algorithms



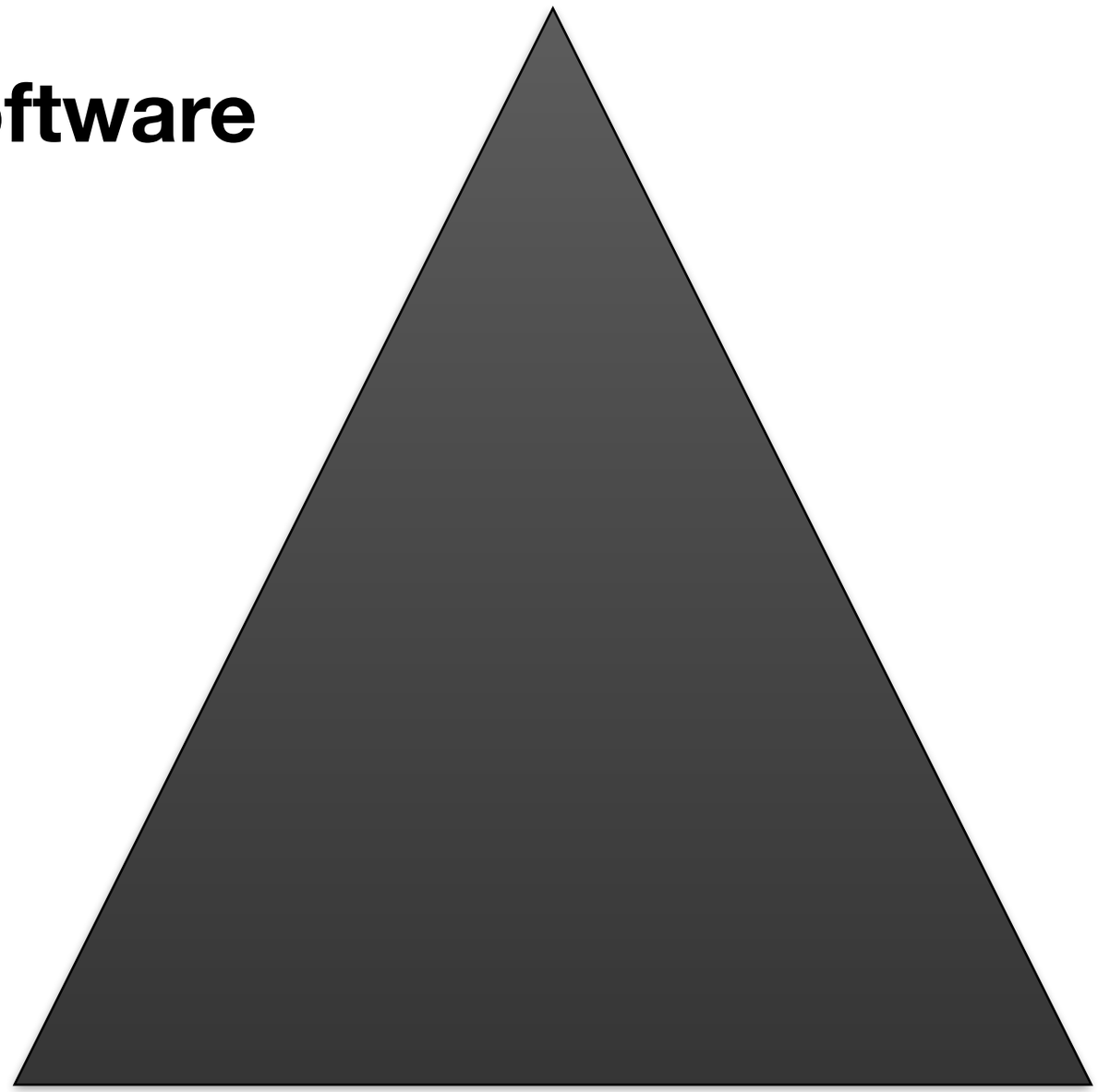
Data cleaning

Modeling

Learning objectives

Algorithms: **software** and hardware

Computation/Algorithms

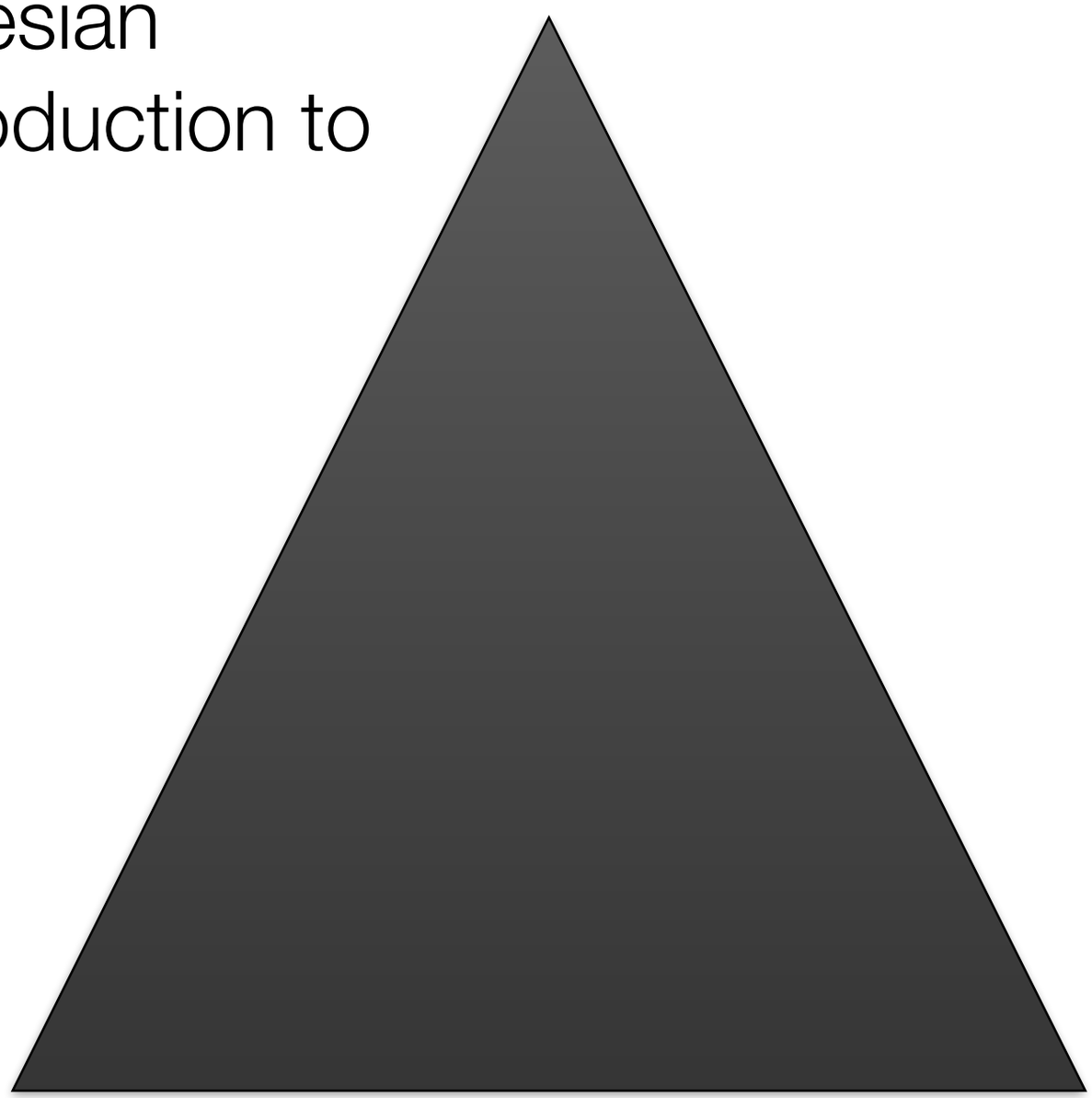


Data cleaning

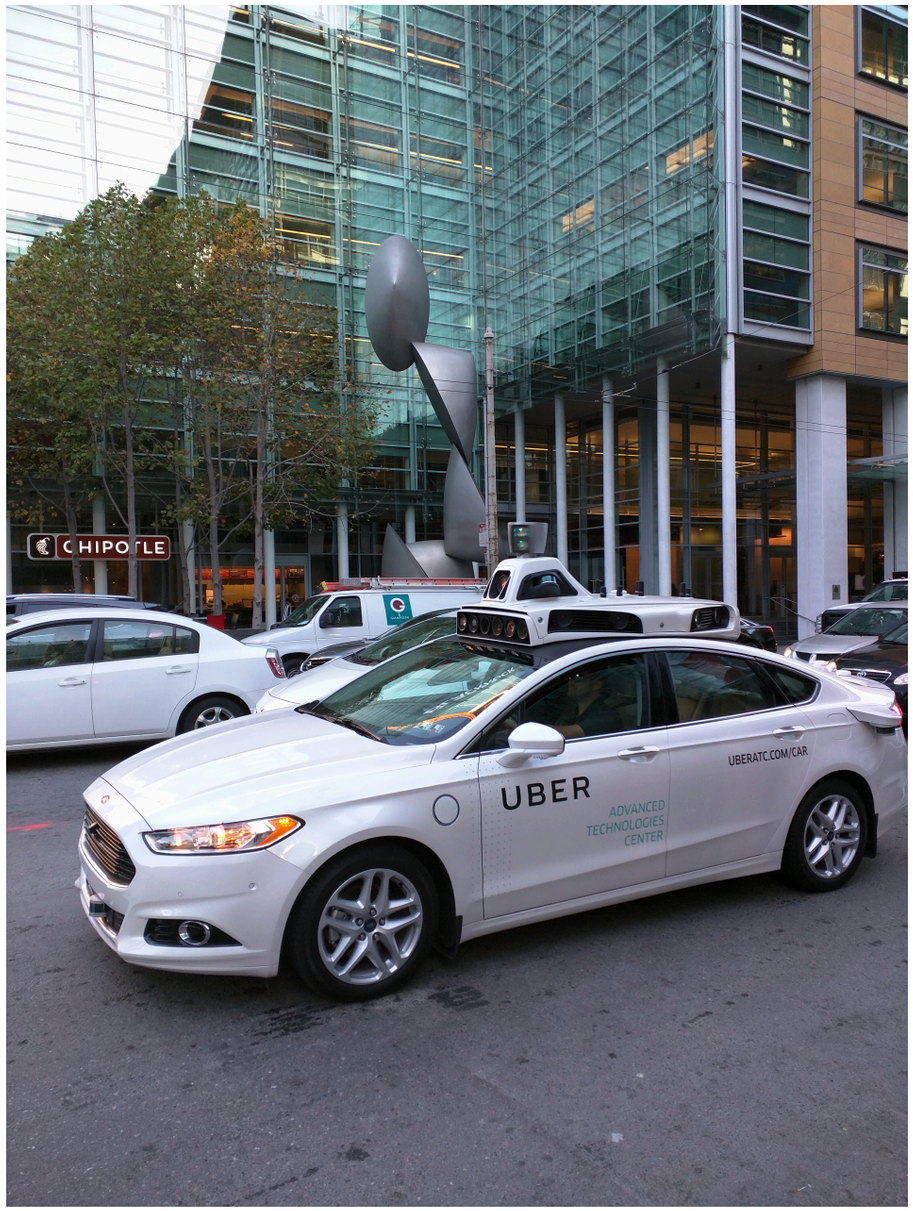
Modeling

Learning objectives

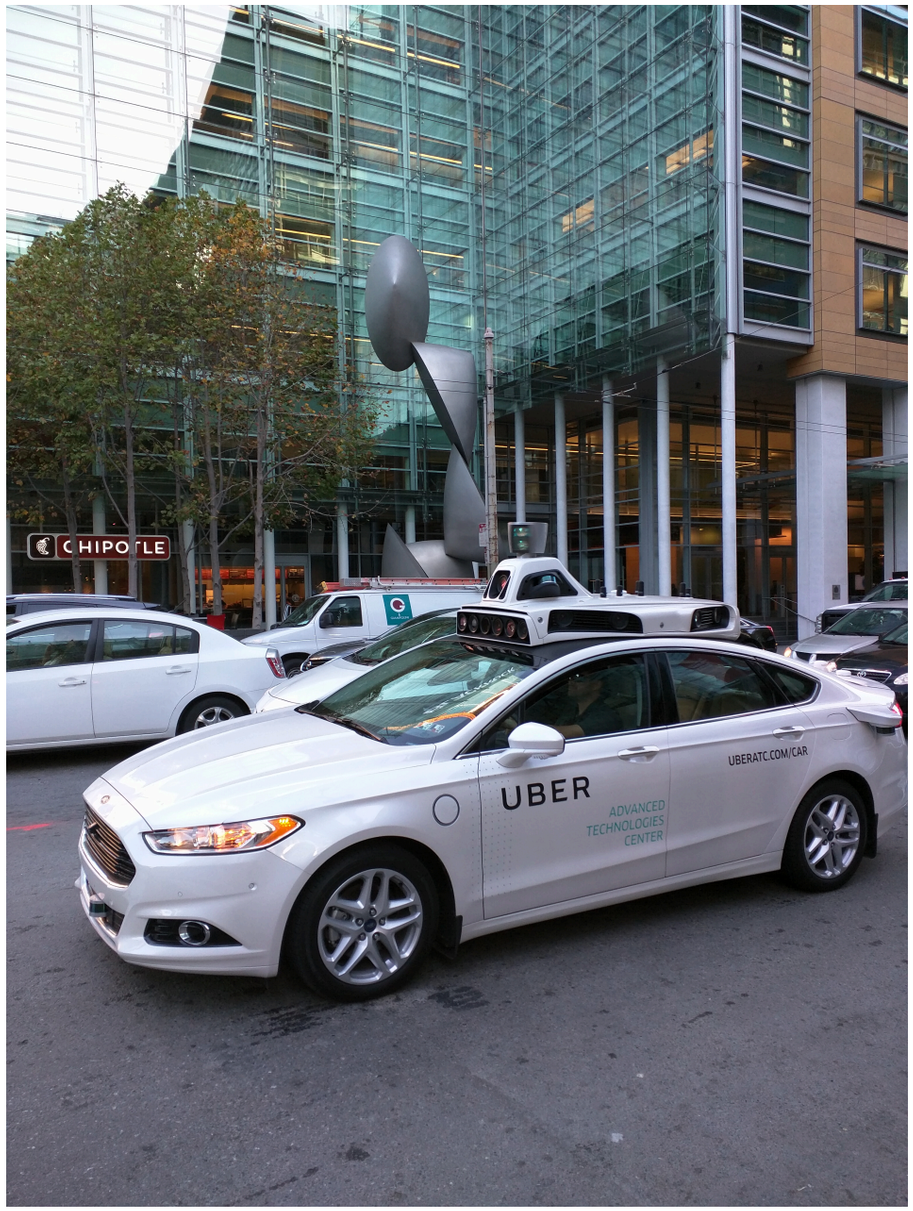
Focus on Bayesian modeling, introduction to deep learning



Transformation of many industries



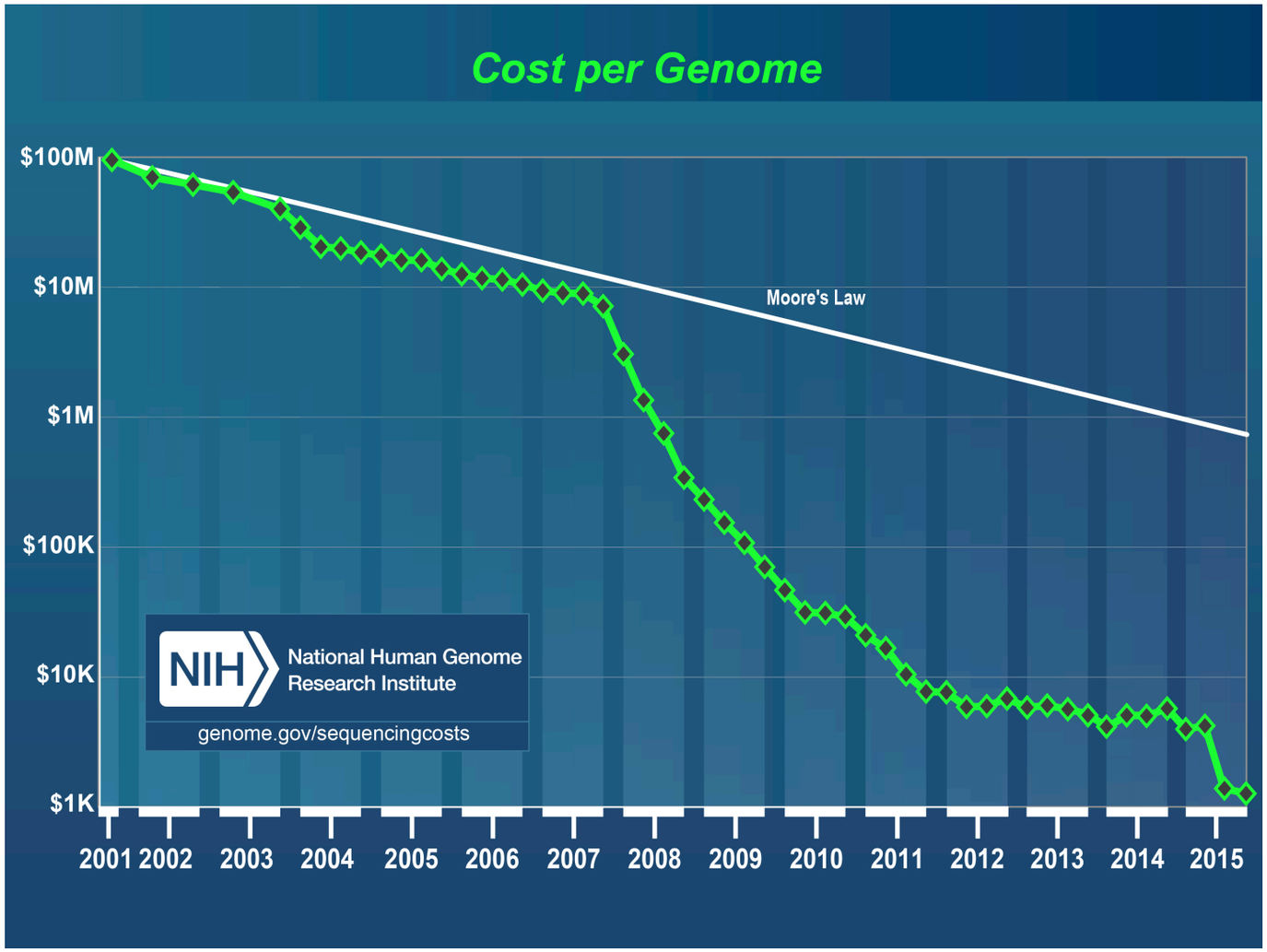
Transformation of many industries



What is missing?

Technologies transforming biomedicine

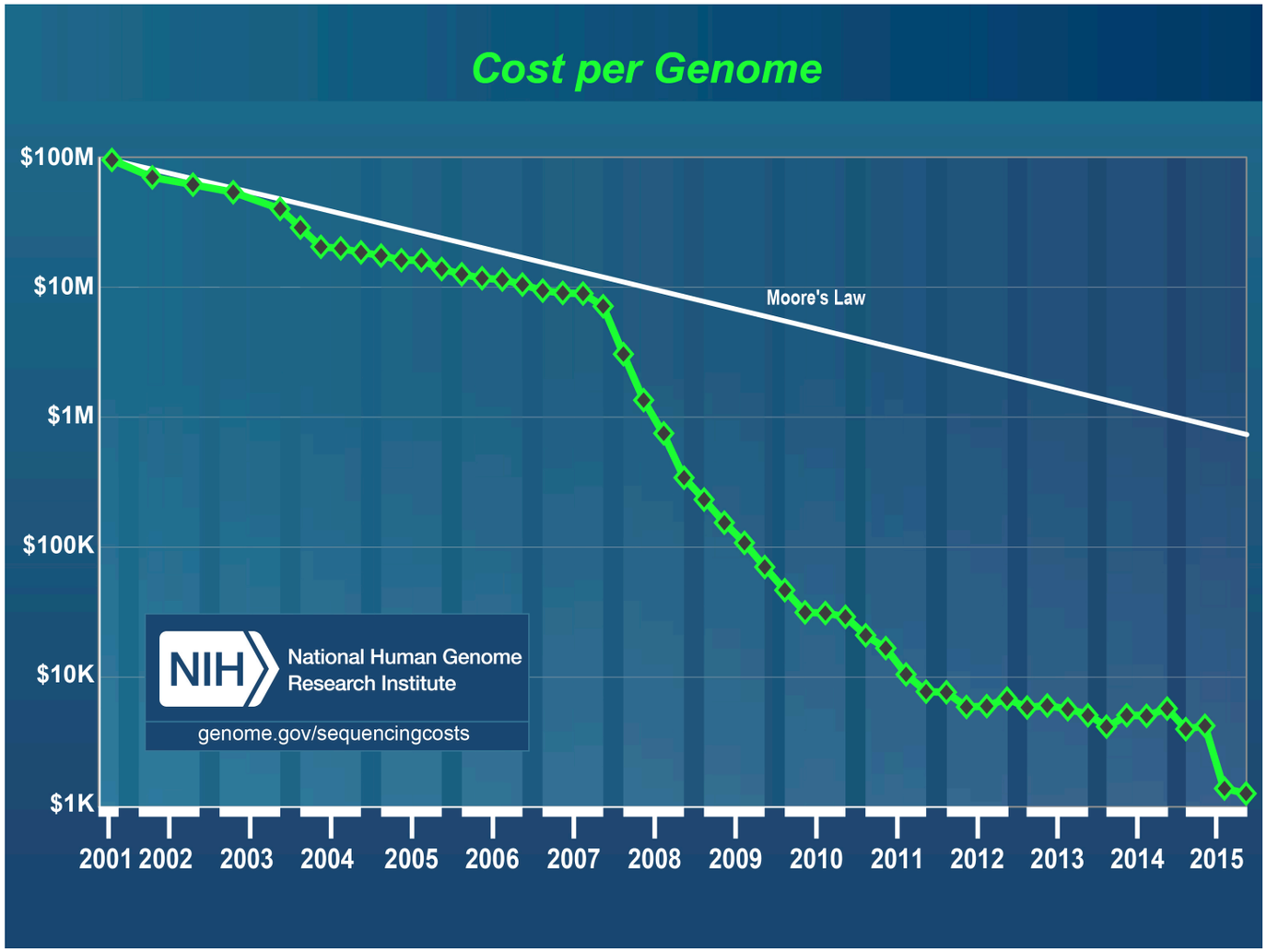
Cost of sequencing has plummeted over the past 15 years



Technologies transforming biomedicine

Cost of sequencing has plummeted over the past 15 years

~$1000 cost point projected for 18/19



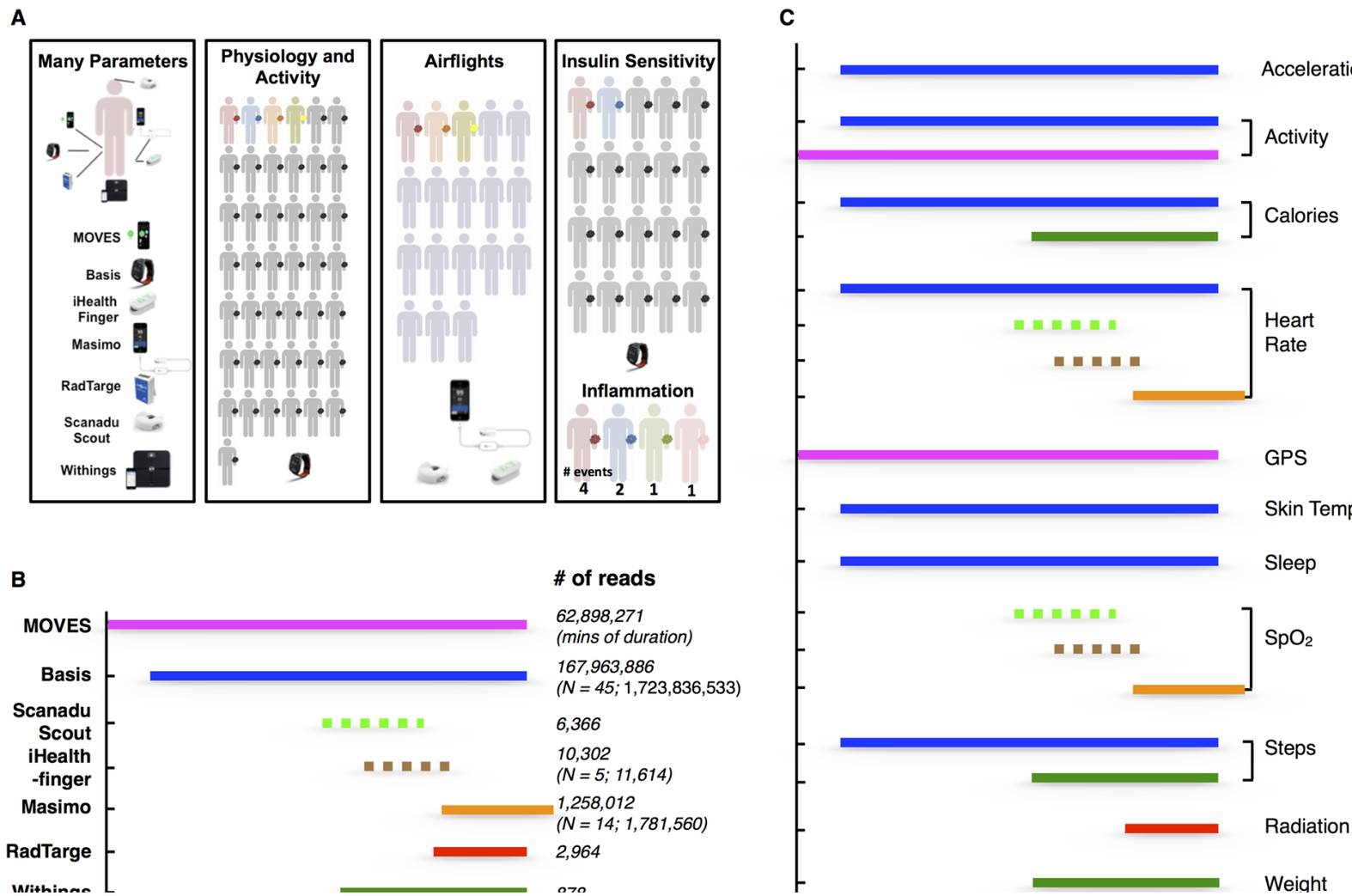
Technologies transforming biomedicine

Wearables and sensors

Ability to continuously monitor health measurements

Technologies transforming biomedicine

Li et al. 2017, PLoS Biology



Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information

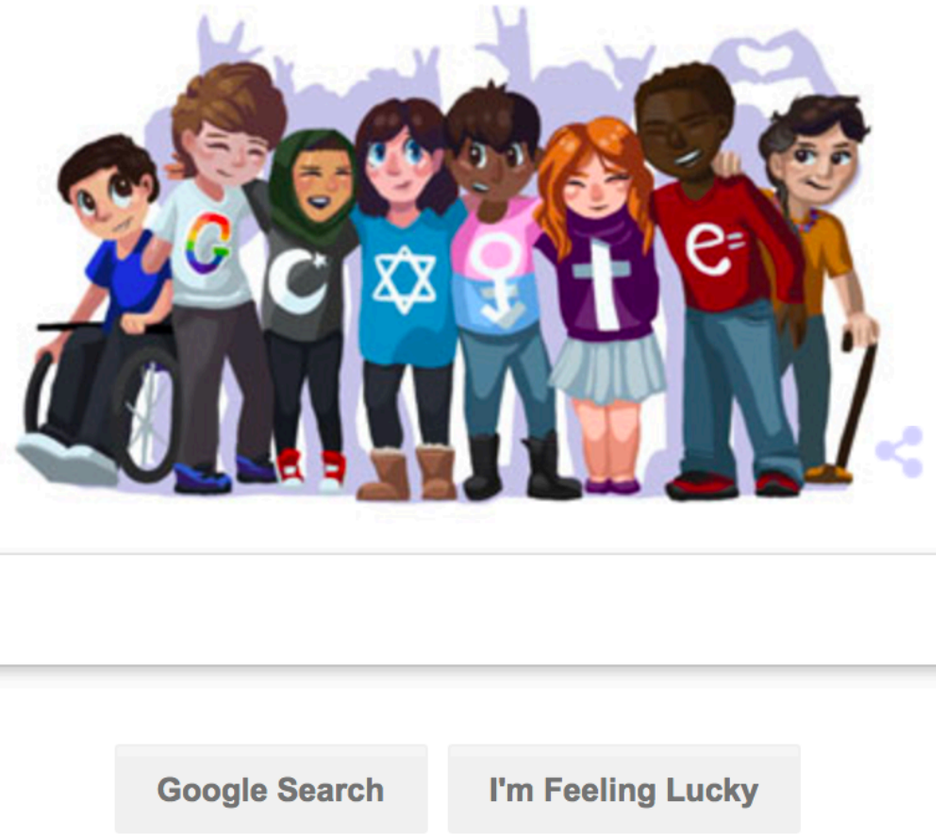
Technologies transforming biomedicine

Data streams from individuals participating in social networks

Social network data

Technologies transforming biomedicine

Data streams
from individual's
search activity



Search engine data

Technologies transforming biomedicine

Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention

Search engine data

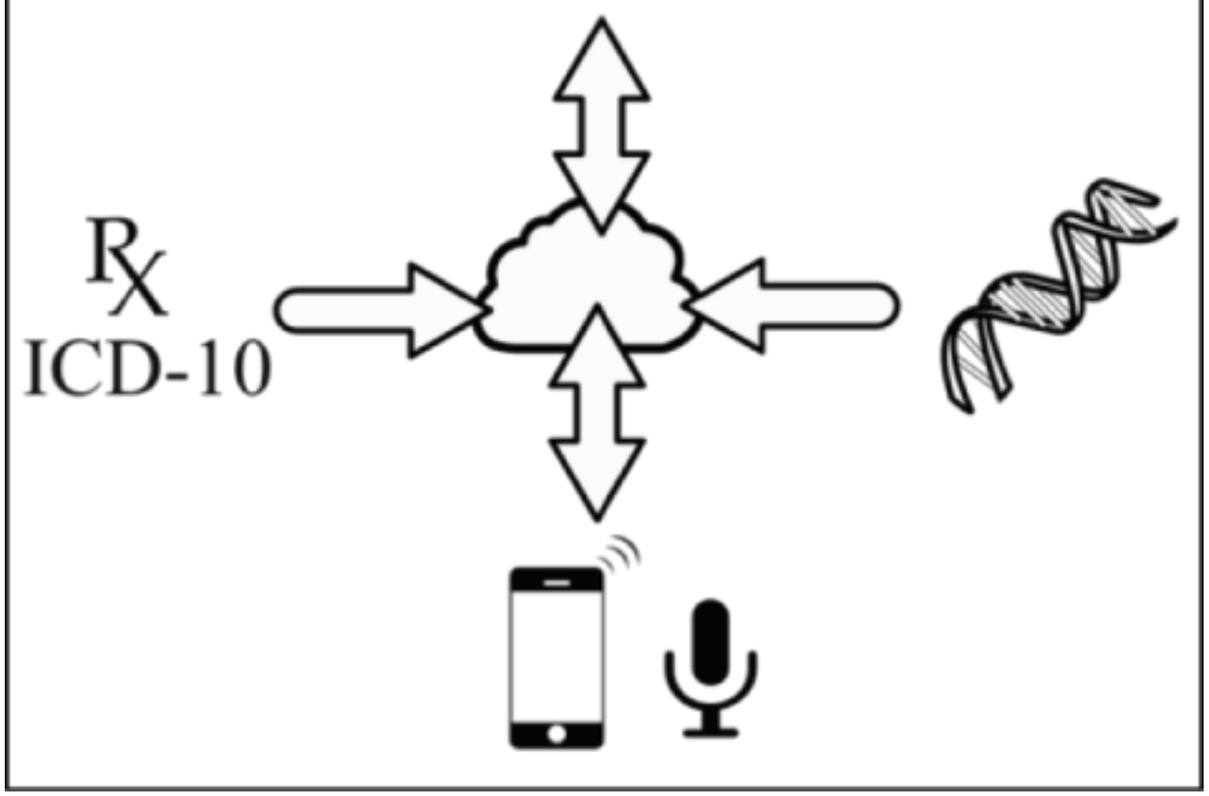
State of health records in many places

Old technologies - health records across many regions of the world are annotated in pencil and paper

How to digitize and put data into action?

$$\int_{\beta_1} P\left(\mathbf{Data}|\beta_1\right) P\left(\beta_1\right) d\beta_1 = cf_{\mathcal{N}}\left(\hat{\beta}; \mu, \widehat{\mathbf{V}}_{\beta} + \mathbf{U}\right)$$

$$y_{ijm}|G_j \sim \sum_{c=1}^{\infty} \pi_{jc} \mathcal{N}\left(\sigma_a b_c, \Sigma_c\right)$$



Challenge for this generation

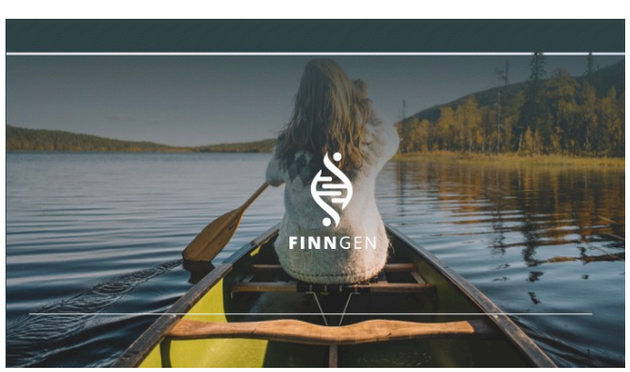
Precision health and biobank initiatives

Precision health and biobank initiatives

UK Biobank

China Kadoorie Biobank

Precision Medicine Initiative



FinnGen

Introduction to the UK Biobank project

Major source of data for this course

About the UK Biobank

National and international health resource

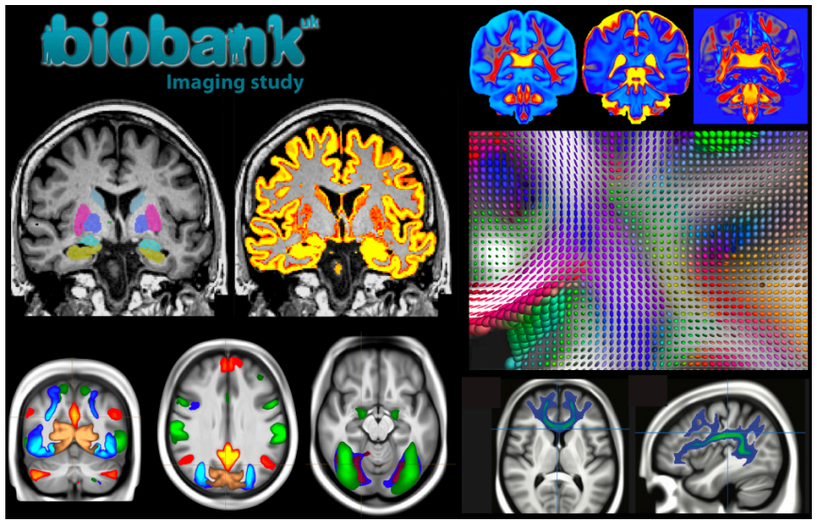
About the UK Biobank

National and international health resource

Hospital in-patient record

Primary care clinical notes

About the UK Biobank



National and international health resource

Hospital in-patient record

Primary care clinical notes

Imaging

~10,000 individuals -> 100,000

About the UK Biobank

National and international health resource

Hospital in-patient record

Primary care clinical notes

Imaging

Physical activity

About the UK Biobank

National and international health resource

Hospital in-patient record

Primary care clinical notes

Imaging

Physical activity

Biomarkers, etc

UK Biobank data showcase webpage

http://biobank.ctsu.ox.ac.uk/crystal/

Please visit

UK Biobank data showcase webpage

biobank uk

Index | Browse | Search | Catalogues | Downloads | Help

Welcome to the online showcase of UK Biobank resources. If you are new to using the showcase we recommend you begin by reading the short introductory **User Guide**. Please note that the showcase contains only anonymous summary information.

Essential Information

Information regarding timelines, updates, release schedules etc.

Browse

Find data items by navigating according to their category of origin.

Search

Find data items by searching on keywords and other characteristics.

Catalogues

Simple listings of database contents and additional resources.

Downloads

Download supporting utilities.

Login

Request data access and view cross-tabulations.

Improving the health of future generations

UK Biobank data showcase webpage

biobank uk

Index | Browse | Search | Catalogues | Downloads | Help

Browse by Primary Category of Origin

Category	Items
Population characteristics	8
UK Biobank Assessment Centre	2023
Biological samples	184
Genomics	12
Genotyping process	6
Genotyping intensities	27
Genotype confidences	25
Genotype calls & imputation	26
Online follow-up	466
Additional exposures	221
Health-related outcomes	149
Returned datasets	1

Top Level

Level 1

Level 2

Level 3

Summary generated 4 February 2017

Improving the health of future generations

UK Biobank data showcase webpage

biobank uk

Index | Browse | Search | Catalogues | Downloads | Help

Browse by Primary Category of Origin

Category	Items
Population characteristics	8
UK Biobank Assessment Centre	0
Recruitment	13
Touchscreen	385
Verbal interview	31
Physical measures	396
Cognitive function	69
Imaging	1108
Biological sampling	10
Procedural metrics	11
Biological samples	184
Genomics	96
Online follow-up	466
Additional exposures	221
Health-related outcomes	149
Returned datasets	1

Top Level

Level 1

Level 2

Level 3

Summary generated 4 February 2017

Improving the health of future generations

UK Biobank data showcase webpage

biobank uk

Index | Browse | Search | Catalogues | Downloads | Help

Browse by Primary Category of Origin

Category	Items
Population characteristics	8
UK Biobank Assessment Centre	2023
Biological samples	184
Genomics	96
Online follow-up	466
Additional exposures	221
Health-related outcomes	0
Hospital in-patient	121
Death register	6
Cancer register	8
Algorithmically-defined outcomes	14
Returned datasets	1

Top Level

Level 1

Level 2

Level 3

Summary generated 4 February 2017

Improving the health of future generations

UK Biobank data showcase webpage

biobank uk

Index | Browse | Search | Catalogues | Downloads | Help

Browse by Primary Category of Origin

Category	Items
Population characteristics	8
UK Biobank Assessment Centre	2023
Biological samples	184
Genomics	96
Online follow-up	0
Diet by 24-hour recall	317
Cognitive function follow-up	48
Work environment	101
Mental health	0
Additional exposures	221
Health-related outcomes	149
Returned datasets	1

Summary generated 4 February 2017

Top Level

Level 1

Level 2

Level 3

Improving the health of future generations

UK Biobank announcement

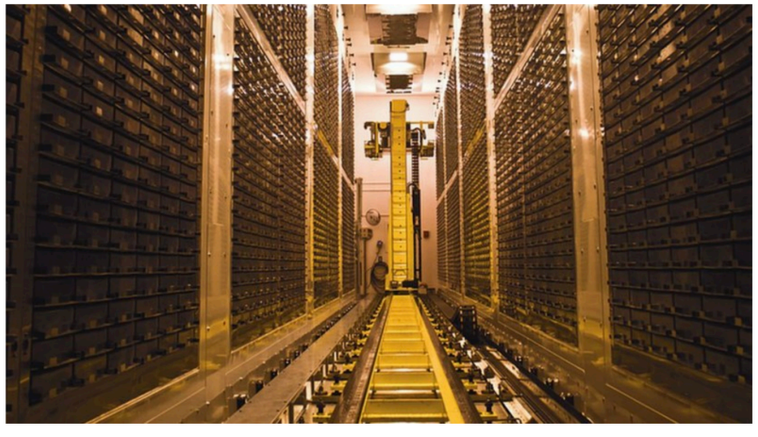
Drug Company Consortium To Sequence The Genes Of 500,000 Britons Over Next Two Years

Matthew Herper, FORBES STAFF
I cover science and medicine, and believe this is biology's century. FULL BIO

Adi Gaskell, Contributor
A London based innovation scout

Medical Consortium Aim To Find Treasure In UK Biobank Data

01/09/2018 02:55 am ET



The power of genetics is something that I've touched on a number of times. Technology

Rewriting Life

500,000 Britons' Genomes Will Be Public by 2020, Transforming Drug Research

Yancopoulos calls the slow start by the U.S. "a national embarrassment." The U.K. data trove is set to dominate "for the foreseeable future, the next five to 10 years," he says. "It's going to be the best resource. It's the first place people will go."

Rare diseases run in families

If you have **cystic fibrosis**, what is your risk for:

Your neighbor (unrelated)?

Your sibling?

Your twin?

Variation in your DNA influences your risk

Common diseases also run in families

If you have **cystic fibrosis**, what is your risk for:

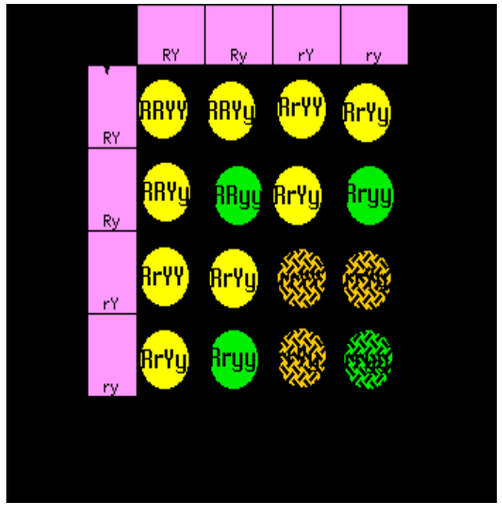
Your neighbor (unrelated)?

Your sibling?

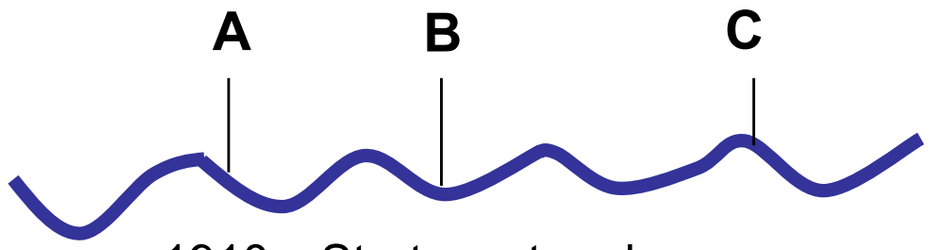
Your twin?

Variation in your DNA influences your risk

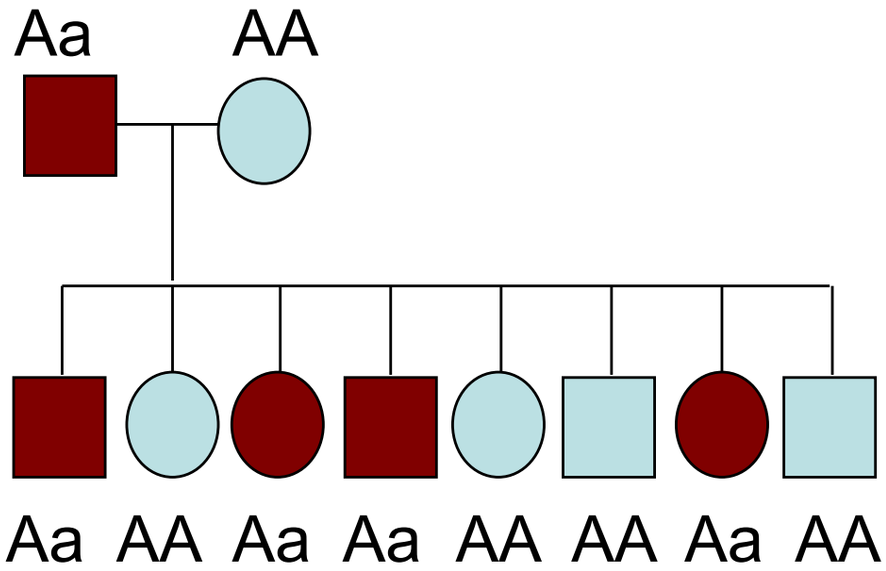
20th century genetics



1860s: Mendel's laws of inheritance – discrete, transmissible units of inherited variation resulting in phenotypic differences



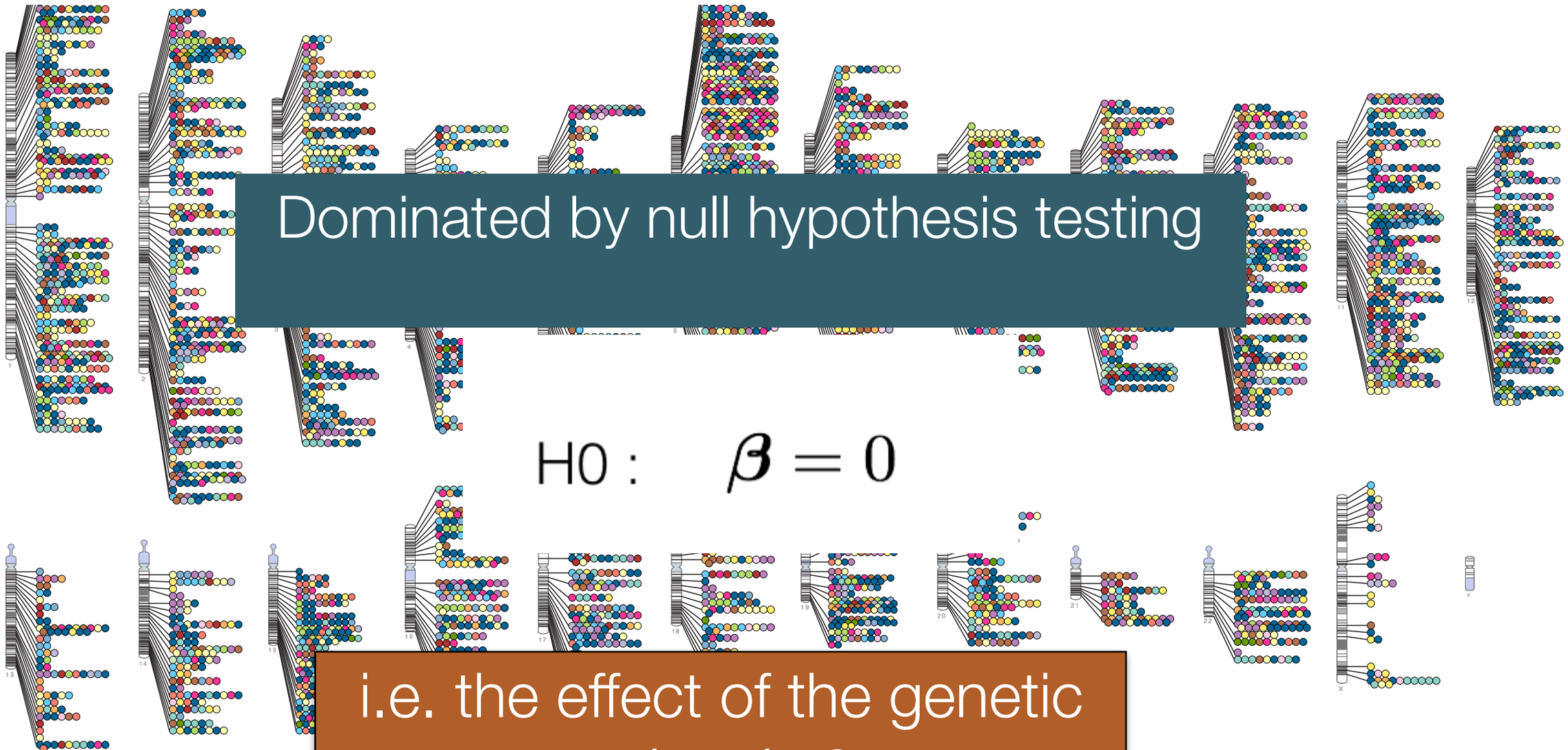
1910s: Sturtevant and Morgan create the first genetic map



Genome wide association studies have been very successful

Over 1000 genetic associations

Genome wide association studies have been very successful

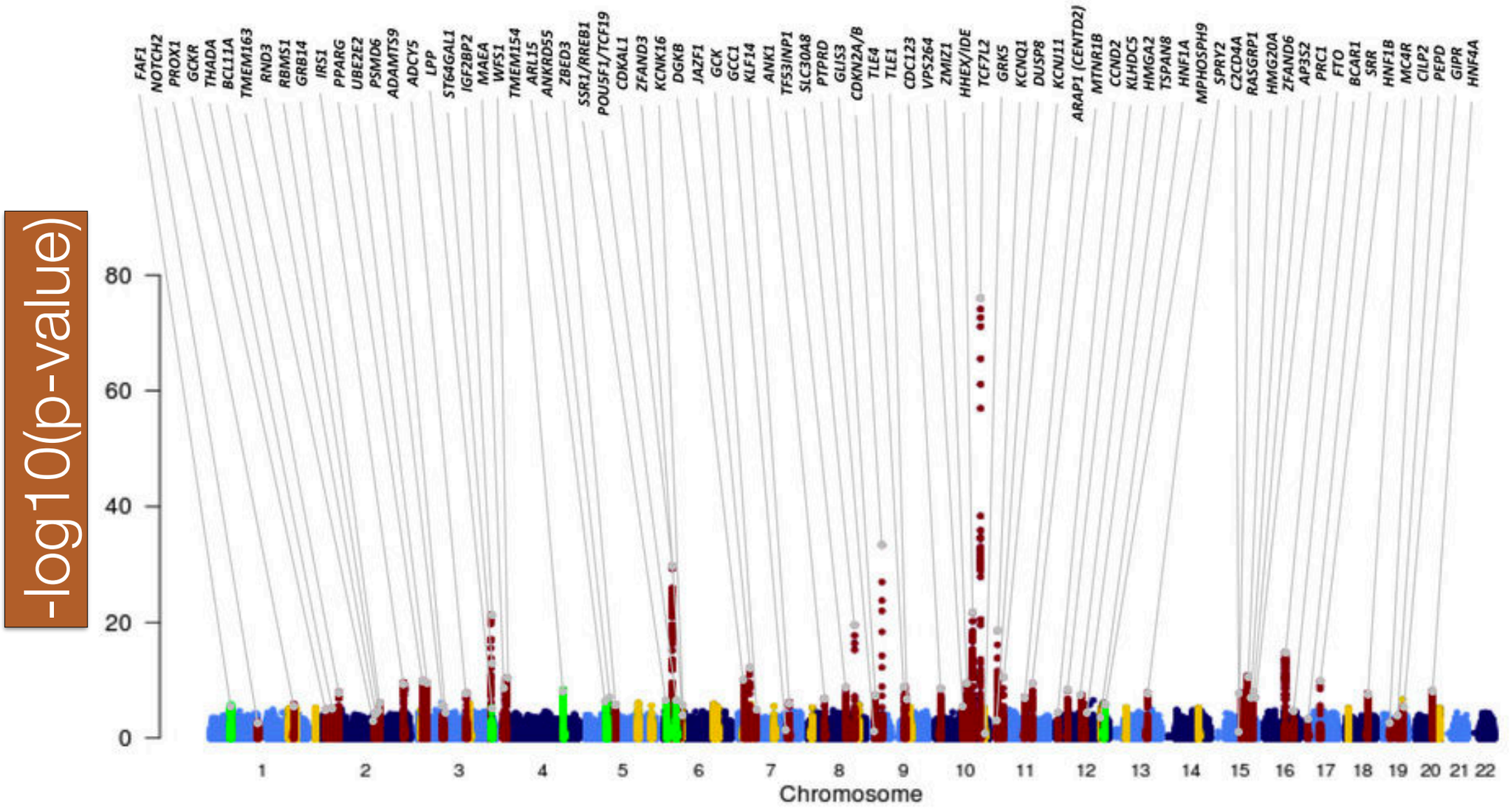


Dominated by null hypothesis testing

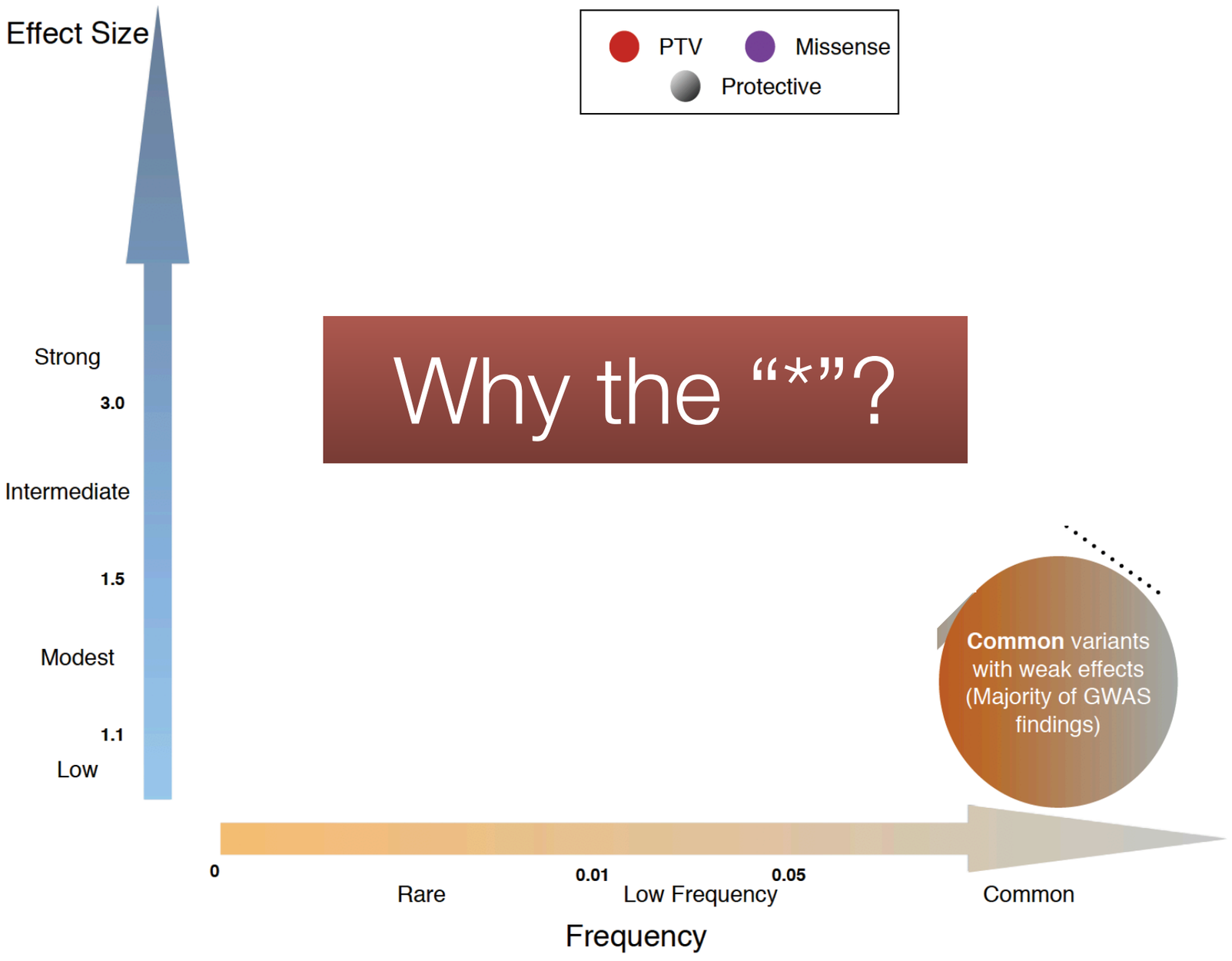
$$H0: \quad \beta = 0$$

i.e. the effect of the genetic variant is 0.

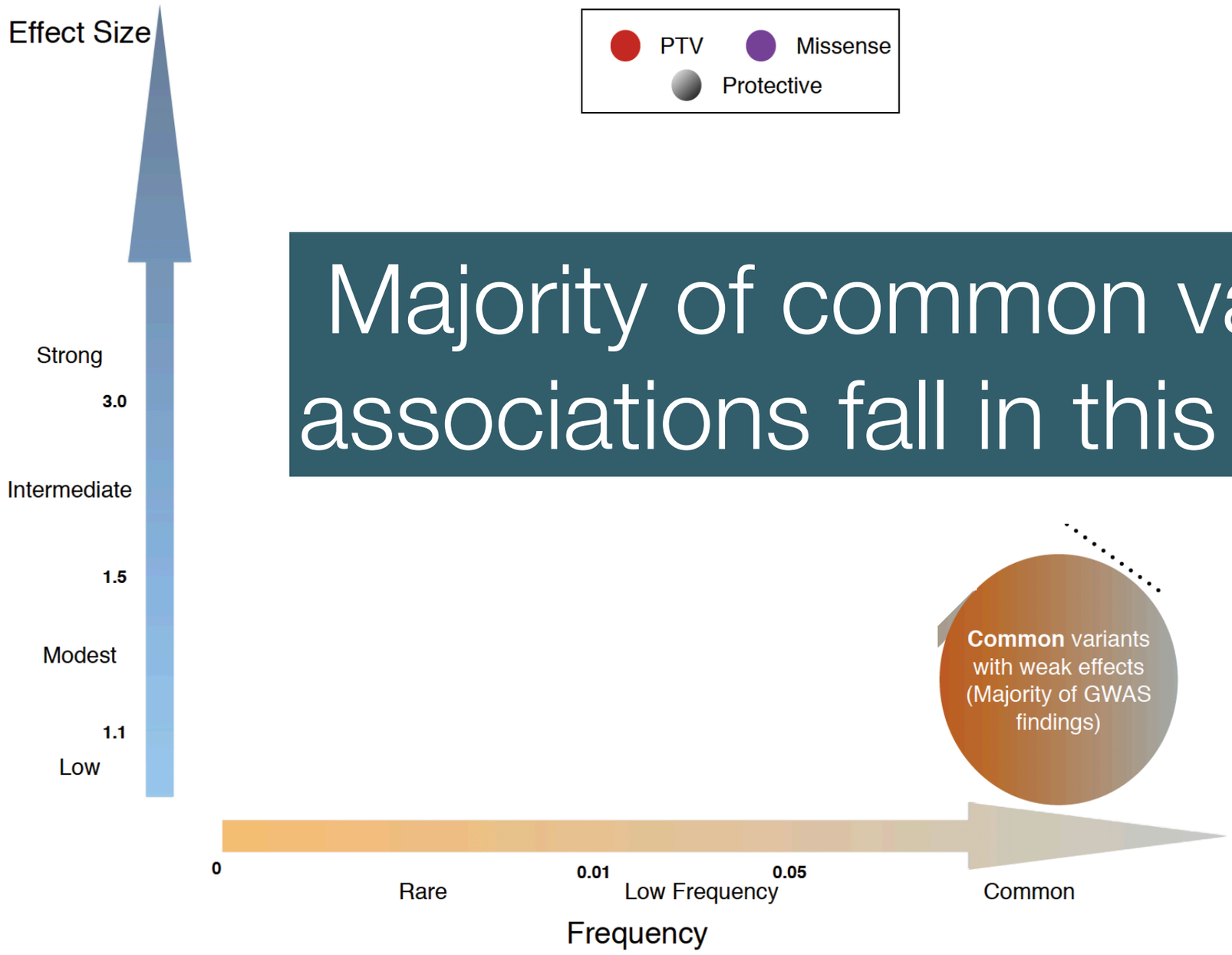
Manhattan plot



Genome wide association studies have been very successful*



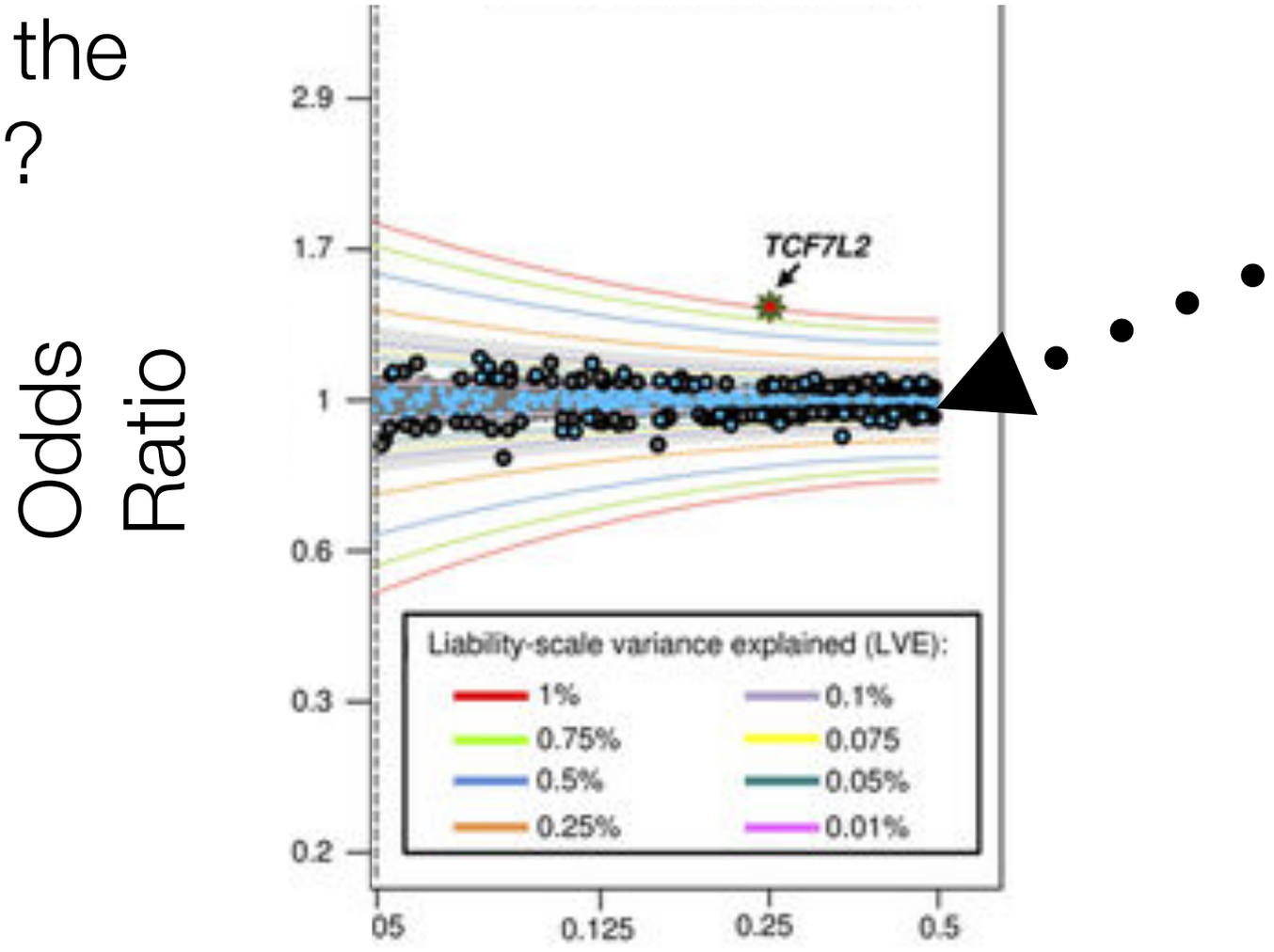
Genome wide association studies have been very successful*



Why the “*”?

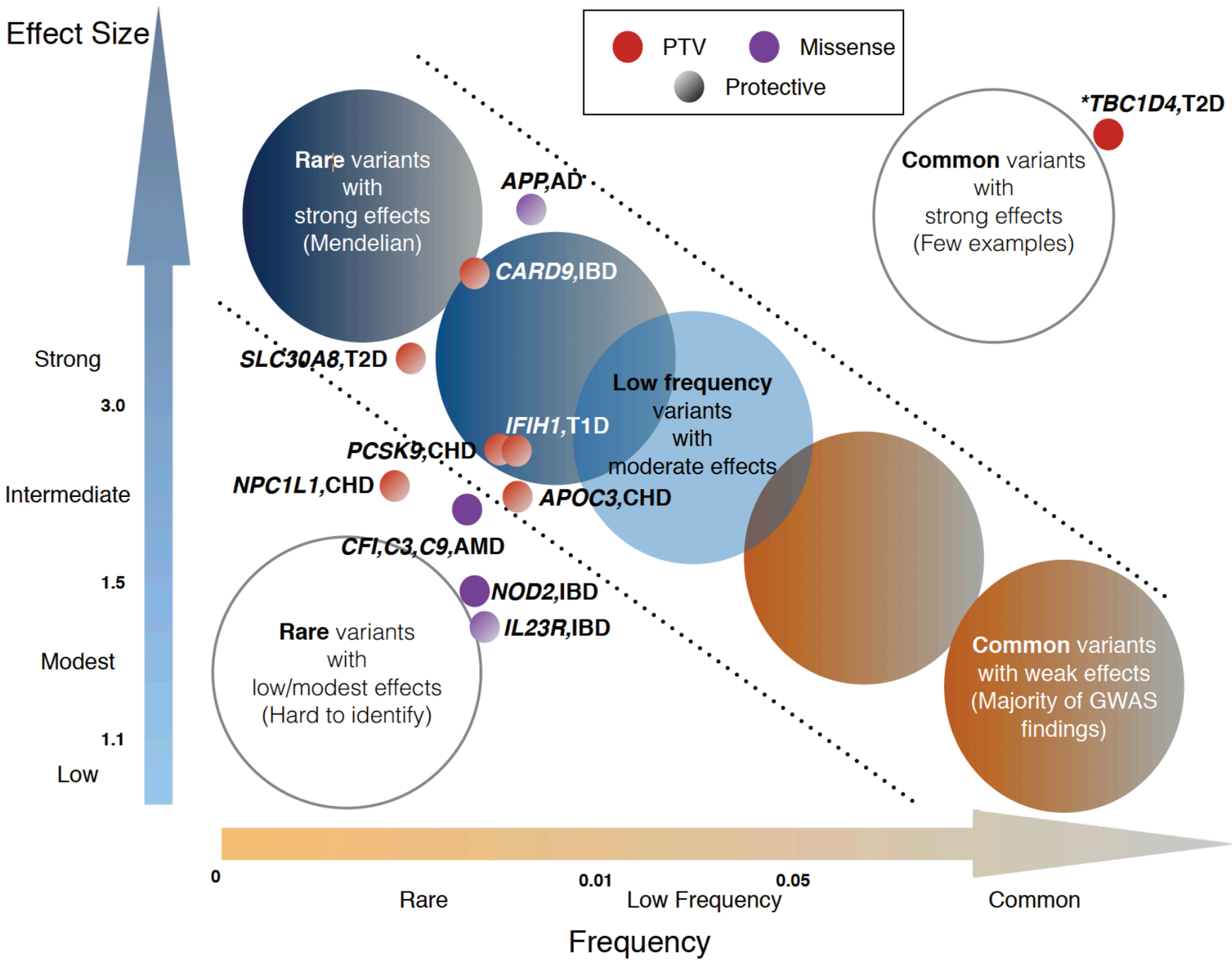
In the context of **type 2 diabetes** all common variant associations were tiny

Why the "*"?

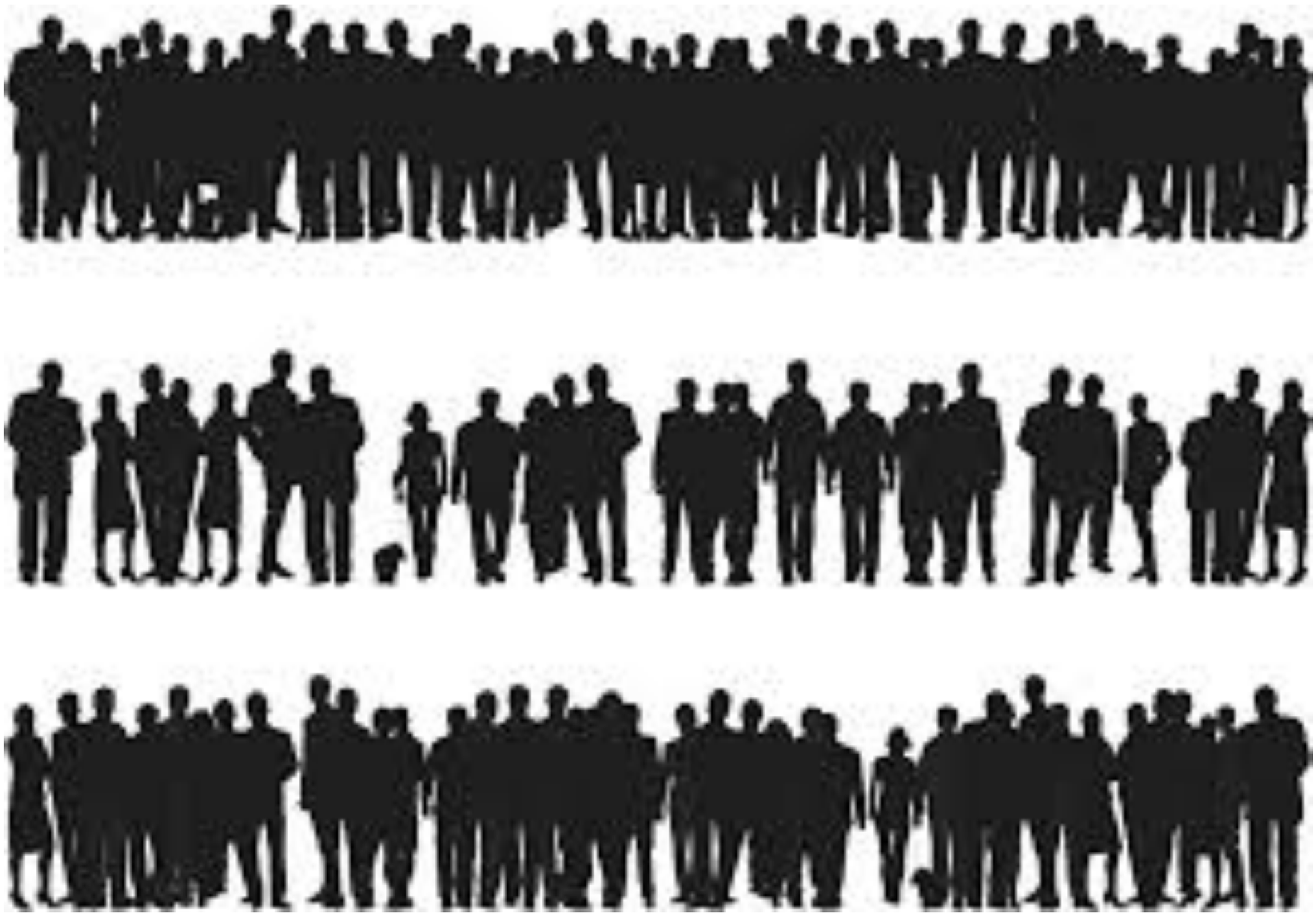


Tiny effect sizes for all associated

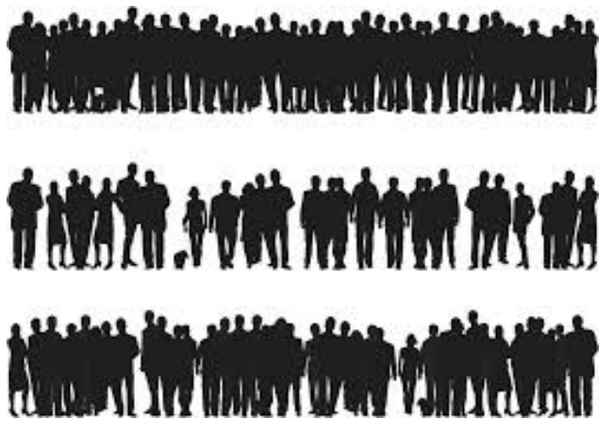
Additional signals started emerging from rare variants



Precision Medicine



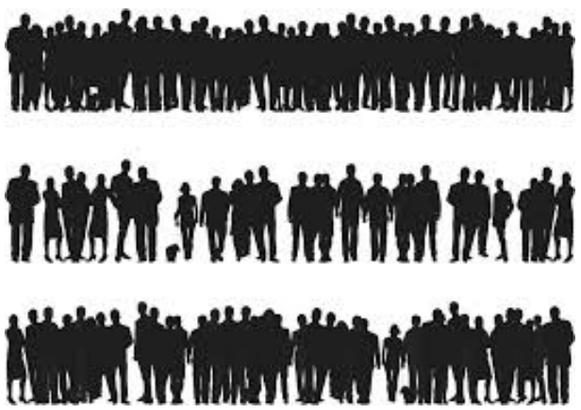
“Experiments of nature” that protect can guide selection of drug targets



Lower risk for disease

Examples of protective mutations

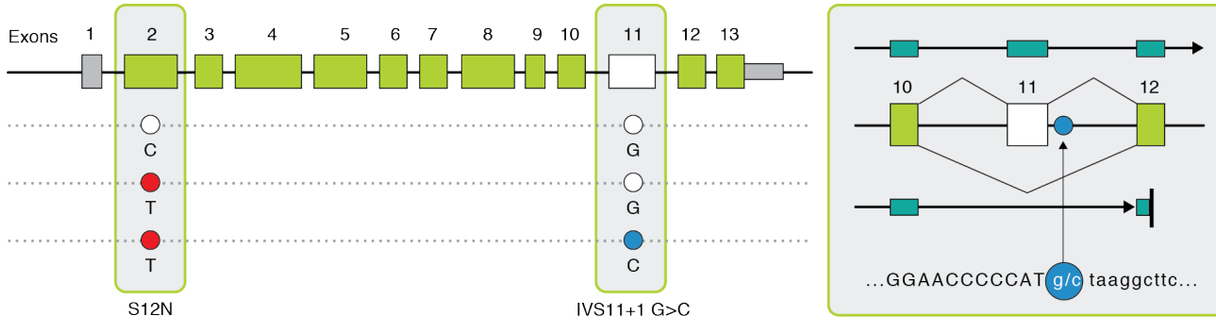
PCSK9 for LDL and MI
Nav 1.7 for pain
CARD9 for Crohn's disease and ulcerative colitis
RNF186 for ulcerative colitis
CCR5 for HIV



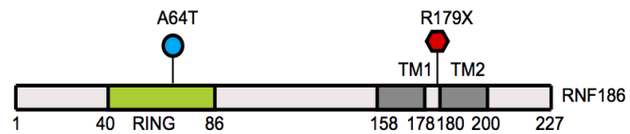
Lower risk for disease

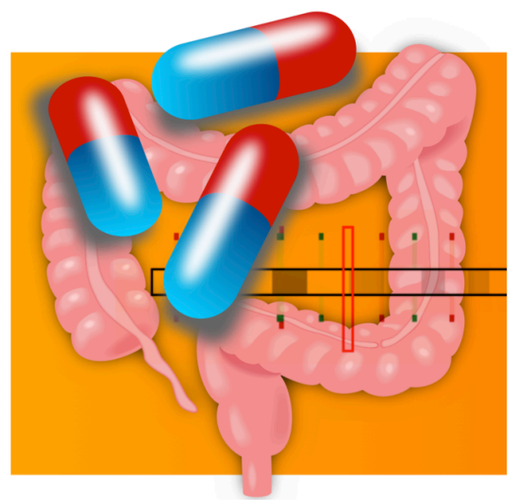
Rare, strong acting alleles provide interpretation of the GWAS results

- Splice variant in *CARD9* cause premature truncating of protein and ***strongly protects*** against the development of Crohn's disease and ulcerative colitis ($p < 10^{-16}$).



- Protective genetic variants reveal process that is:
  - **safe** (naturally occurs in healthy adults)
  - **effective** (proven to reduce risk of disease).





Population medical data combined with **human genetic data** empowers novel **data science technologies**

Global Biobank Engine

UK Biobank Array European

Select an association set | Submit

Search for a gene or variant or region or phenotype

Examples for UK Biobank array - Gene: F5, Variant: 1:169519049-T-C, RS ID: rs6025, Region: 10:114686614-114786614, Phenotype: Asthma

Genetic Association Results

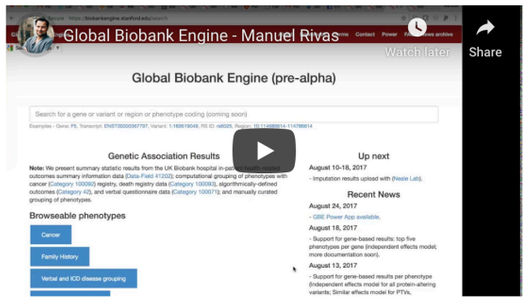
Note: We have aggregated summary statistics from over 750,000 individuals across three population cohorts: UK Biobank, Million Veterans Program and Biobank Japan. We are continuously adding data from other population cohorts in Global Biobank Engine. Please contact us if you want it to be featured.

For UK Biobank we present summary statistic results from the UK Biobank hospital in-patient health-related outcomes summary information data (Data-Field 41202); computational grouping of phenotypes with cancer (Category 100092) registry, death registry data (Category 100093), algorithmically-defined outcomes (Category 42), and verbal questionnaire data (Category 100071); and manually curated grouping of phenotypes.

Terms of use

We encourage use of the initial browser but note that case-control results are provided as general guides and specific results may not have yet been subjected to the data quality, statistical and population genetics review that would normally be required for publication or clinical inference.

Tutorial



Recent News | Lab manuscripts

bioRxiv preprints

9 November 2019

- Sex-specific genetic effects across biomarkers

17 October 2019

- Polygenic risk modeling with latent trait-related genetic components

14 October 2019

- Reported CCR5-Δ32 deviation from Hardy-Weinberg equilibrium is explained by poor genotyping of rs6262503

21 August 2019

- Assessing digital phenotyping to enhance genetic studies of human diseases

20 August 2019

- WhichTF is dominant in your open chromatin data?

26 June 2019

- Rare protein-altering variants in ANGPTL7 lower intraocular pressure and protect against glaucoma

5 June 2019

- Genetics of 38 blood and urine biomarkers in the UK Biobank

7 May 2019

- A Fast and Flexible Algorithm for Solving the Lasso in Large-scale and Ultrahigh-dimensional Problems

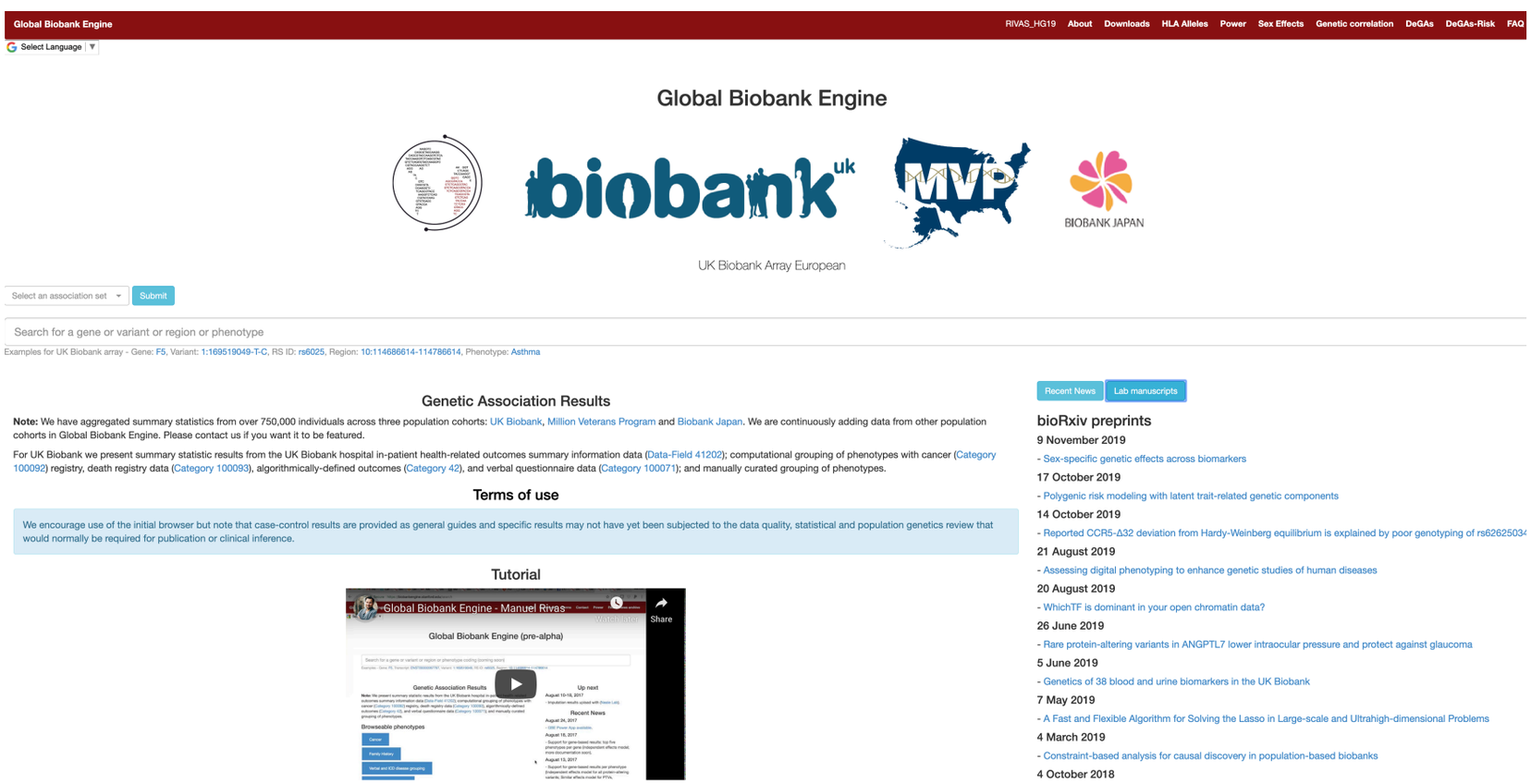
4 March 2019

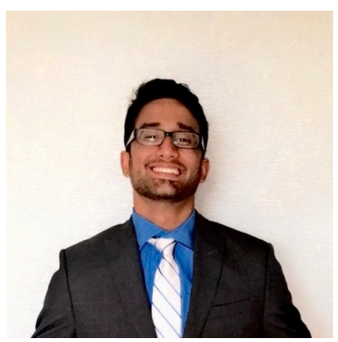
- Constraint-based analysis for causal discovery in population-based biobanks

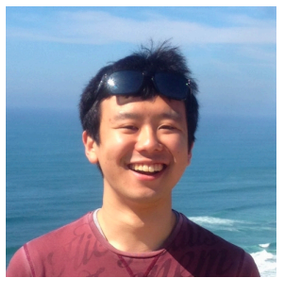
4 October 2018

https://biobankengine.stanford.edu

Global Biobank Engine







Course projects

Topics will be proposed during next week's lecture

Data are organized by Yosuke

Goal : To implement and apply techniques learned in the class to big biomedical datasets

Data available in the course

1. **UK Biobank**

   Summary statistic data for over 100 diseases

2. ENCODE

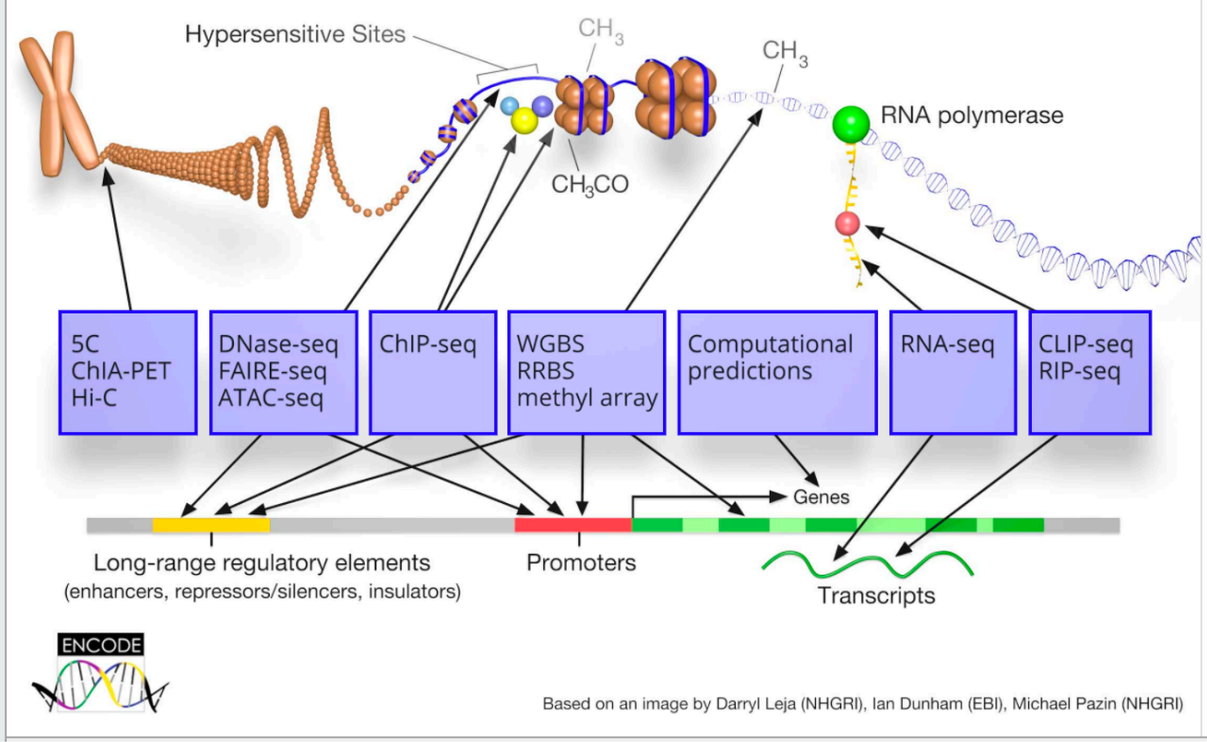
Data available in the course

1. UK Biobank

   Summary statistic data for over 100 diseases

2. ENCODE

ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Get Started

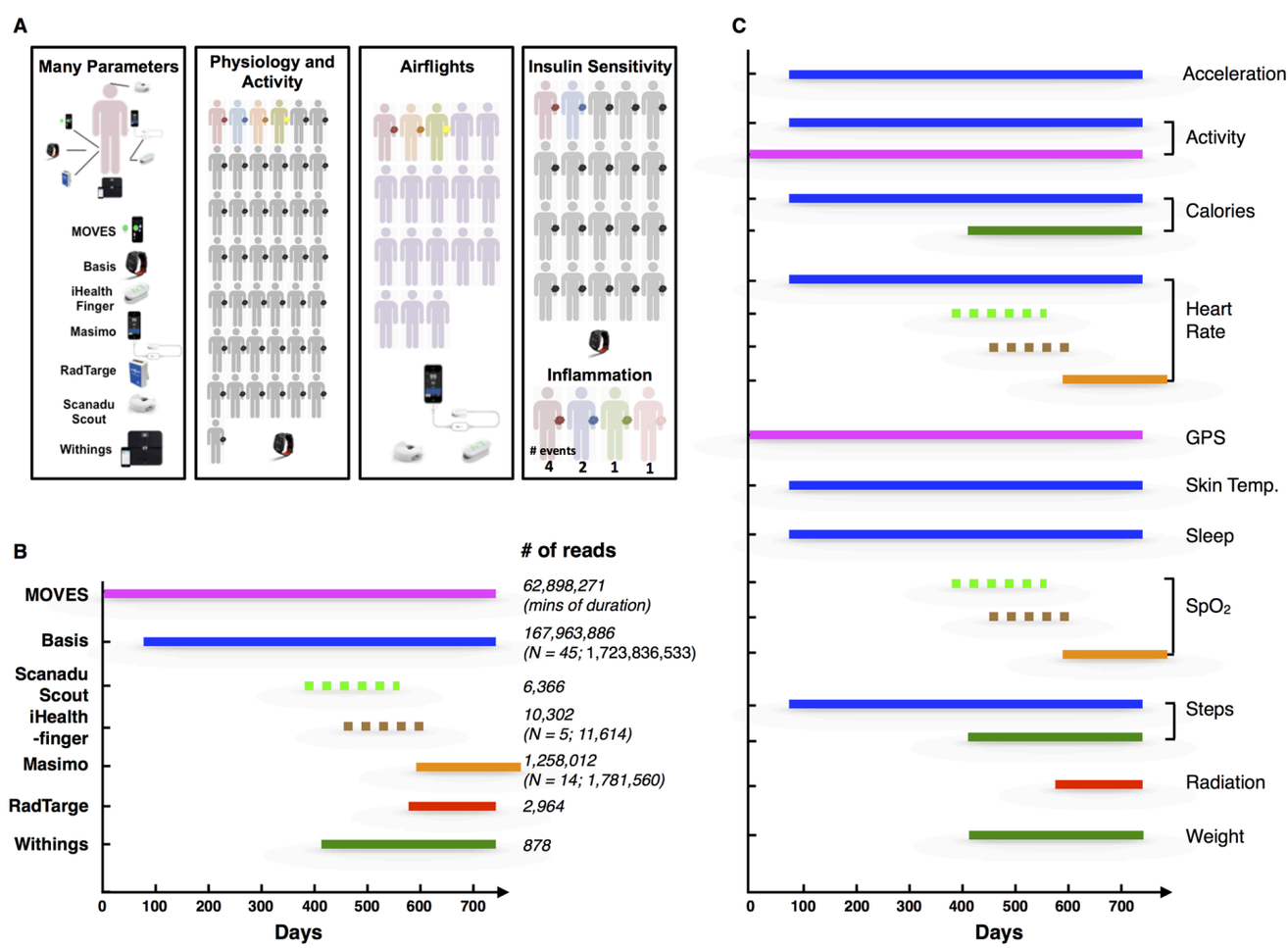
Data available in the course

1. UK Biobank

   Summary statistic data for over 100 diseases

2. ENCODE

3. Wearable Biosensor



Data available in the course

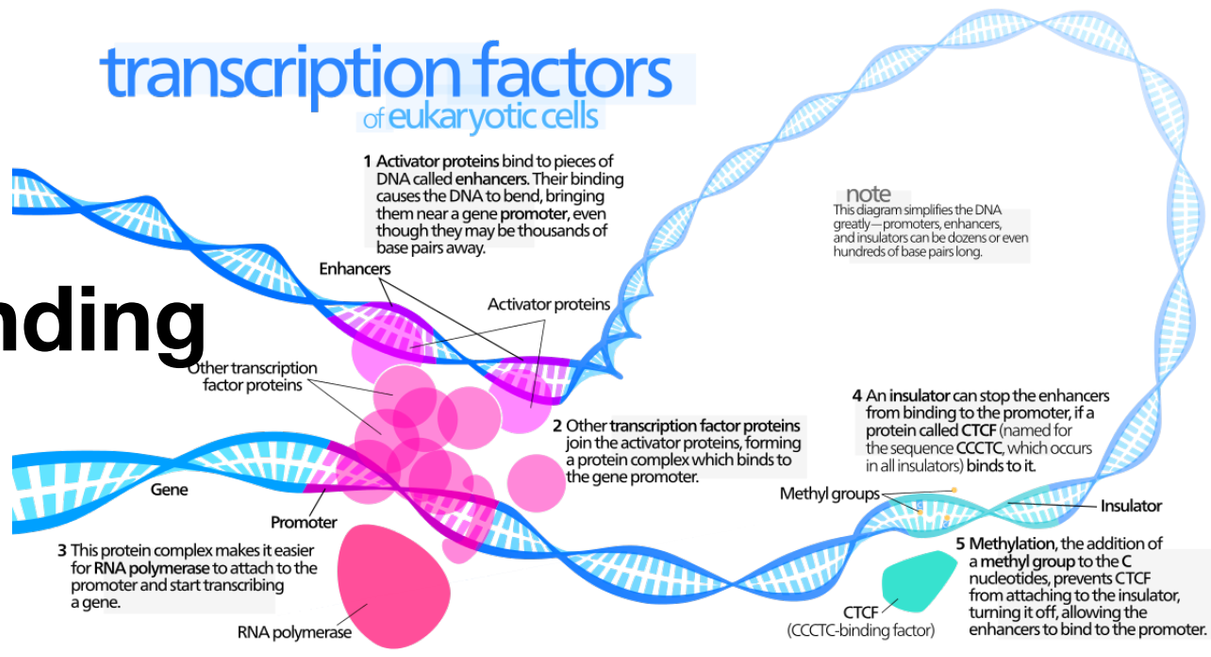
1. UK Biobank

   Summary statistic data for over 100 diseases

2. ENCODE

3. Wearable Biosensor

4. Transcription factor binding



Data available in the course

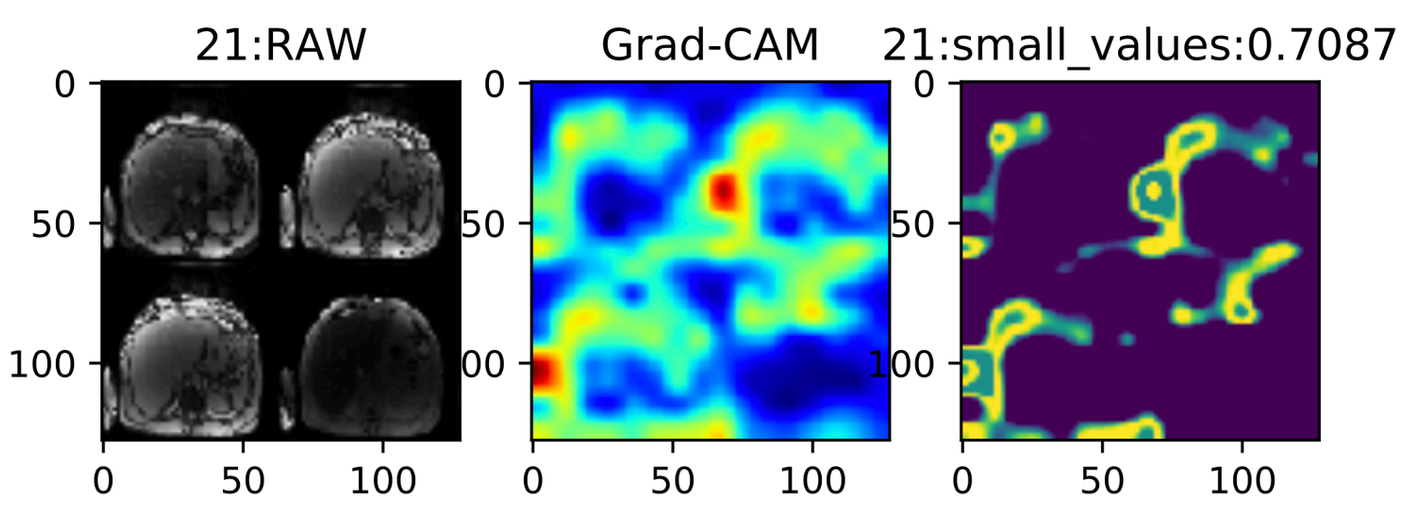
1. UK Biobank

   Summary statistic data for over 100 diseases

2. ENCODE

3. Wearable Biosens

4. Transcription fac



5. Imaging combined with genetics

50 years of Data Science

Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics

50 years of Data Science, David Donoho 2015

50 years of Data Science

Chambers: more emphasis on data preparation and presentation

50 years of Data Science, David Donoho 2015

50 years of Data Science

Breiman: more emphasis on prediction

50 years of Data Science, David Donoho 2015

50 years of Data Science

"For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt… All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical statistics) which apply to analyzing the data."
— *The Future of Data* Analysis, John Turkey 1962

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

2.Data representation and transformation

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

2.Data representation and transformation

3. Computing with data

STAN in this course

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

2.Data representation and transformation

3. Computing with data

4.Data modeling

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

2.Data representation and transformation

3. Computing with data

4.Data modeling

5. Data visualization and presentation

50 years of Data Science, David Donoho 2015

The Six Divisions of Greater Data Science

1. Data exploration and preparation

2.Data representation and transformation

3. Computing with data

4.Data modeling

5. Data visualization and presentation

6.Science about Data Science

50 years of Data Science, David Donoho 2015

The Next 50 years of Data Science

Open Science takes over

Reproducibility

Documented workflows

Science as data

50 years of Data Science, David Donoho 2015

Topics you will learn in this course

Causal inference - Mendelian randomization

Survival analysis

Bayesian multilevel modeling

Bayesian mixture models

Risk modeling

Lasso Net

Convolution Neural Networks, LSTM

Gaussian Process regression

Course website

Syllabus
Reading materials
Lecture notes https://canvas.stanford.edu/courses/113896
Problem Sets
Data links

Sample STAN programs
Application to data
Github repository

https://biods215.github.io/class_website/2020.html

Github repository for the course

https://biods215.github.io/