

BIODS215

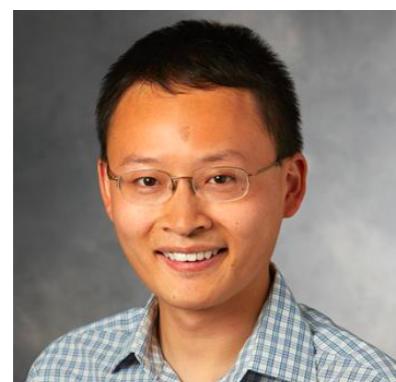
Topics in Biomedical Data Science: Large-scale inference

Winter Quarter 2018

Course Instructors



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

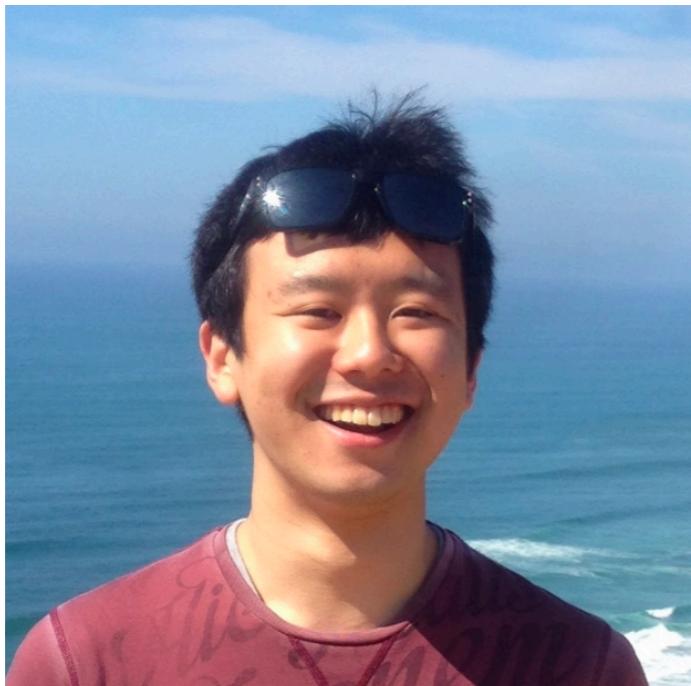


Prof. James Zou
MSOB X325
jamesz@stanford.edu
<https://sites.google.com/site/jamesyzou/>



Prof. Julia Salzman
279 Campus Drive
Beckman Center B473
julia.salzman@stanford.edu
<http://salzmanlab.stanford.edu/>

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%)

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

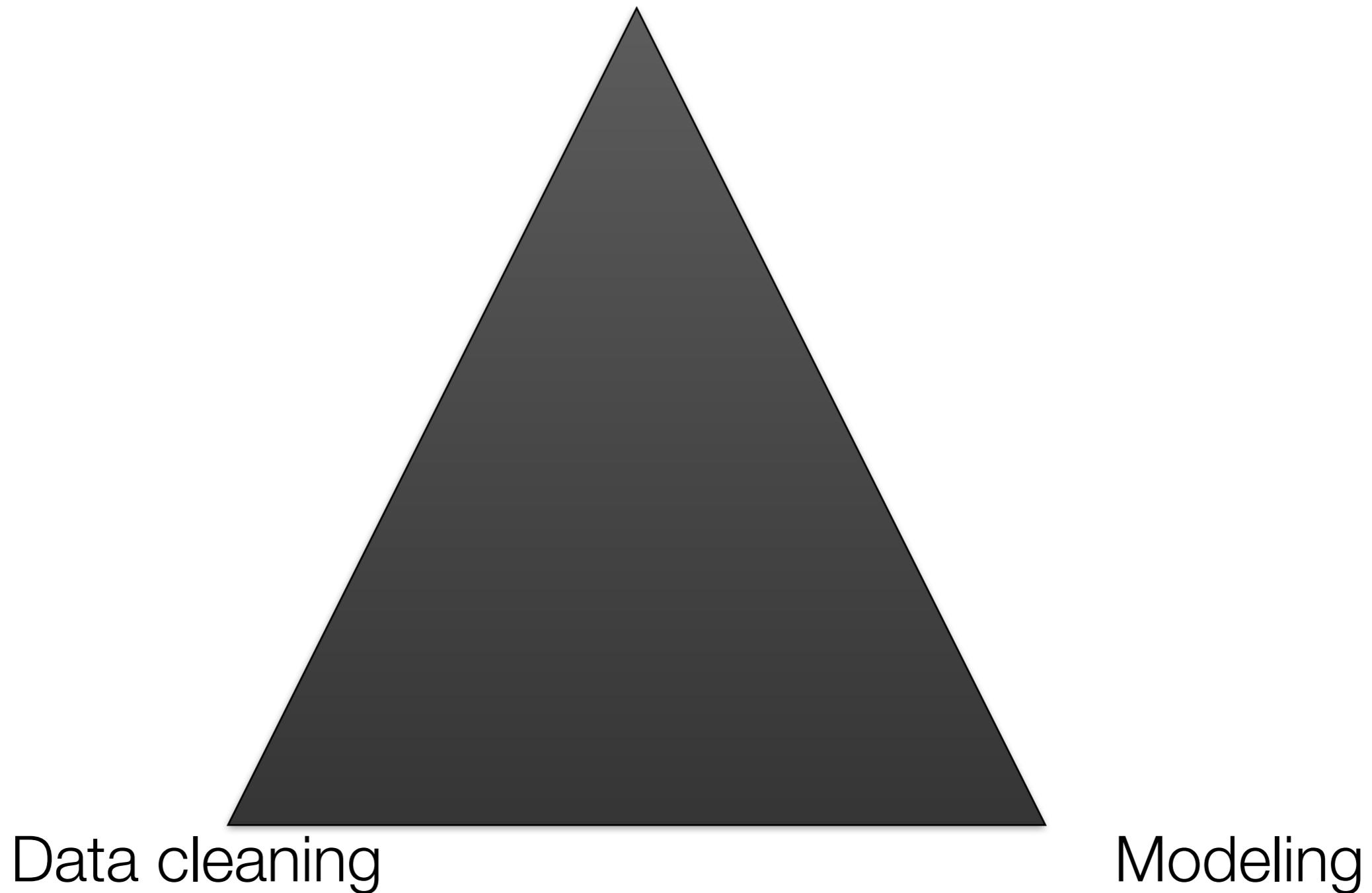


Prediction: To be able to predict what the responses are going to be to future input variables

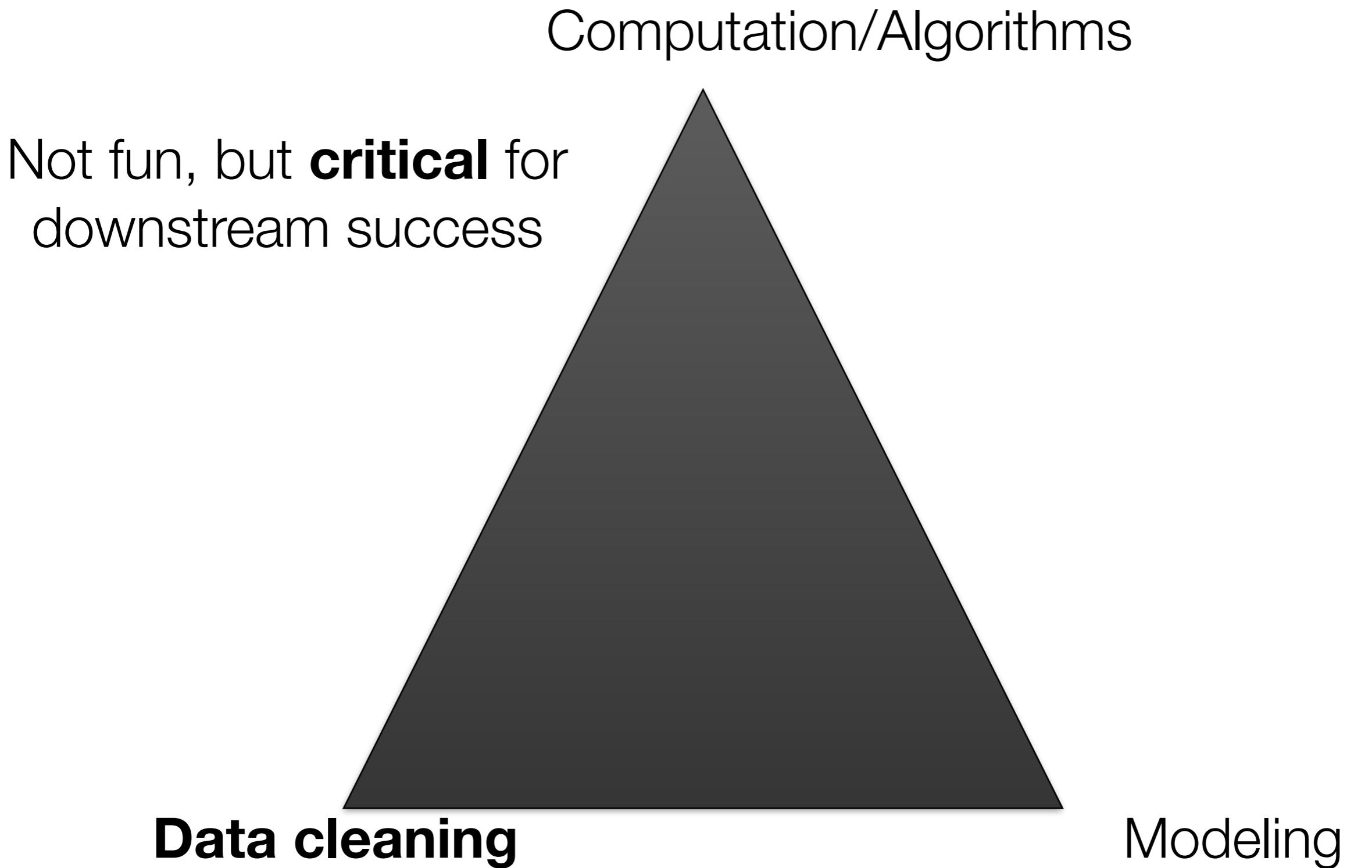
Machine Learning, Computer science

Learning objectives

Computation/Algorithms



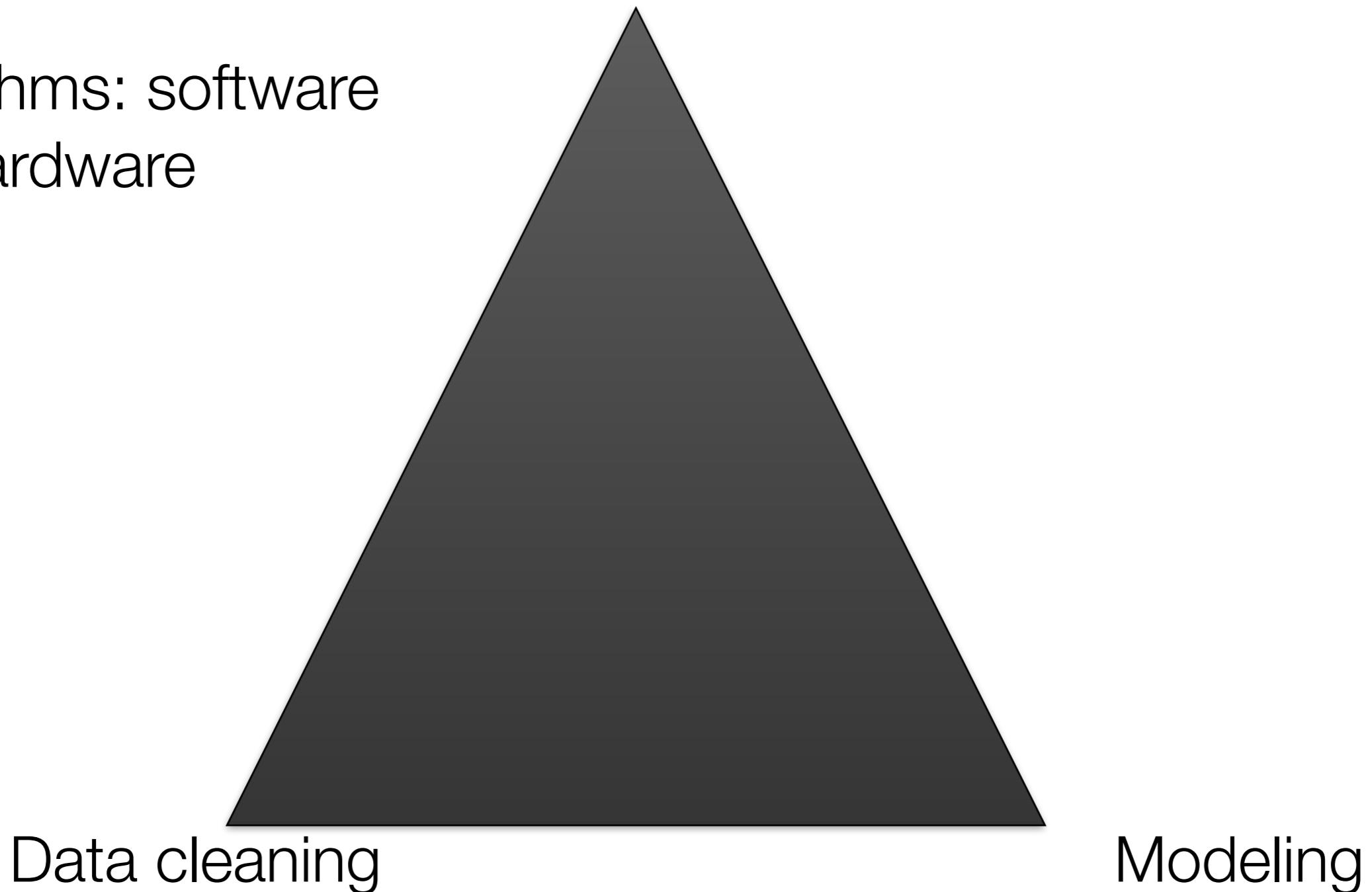
Learning objectives



Learning objectives

Computation/Algorithms

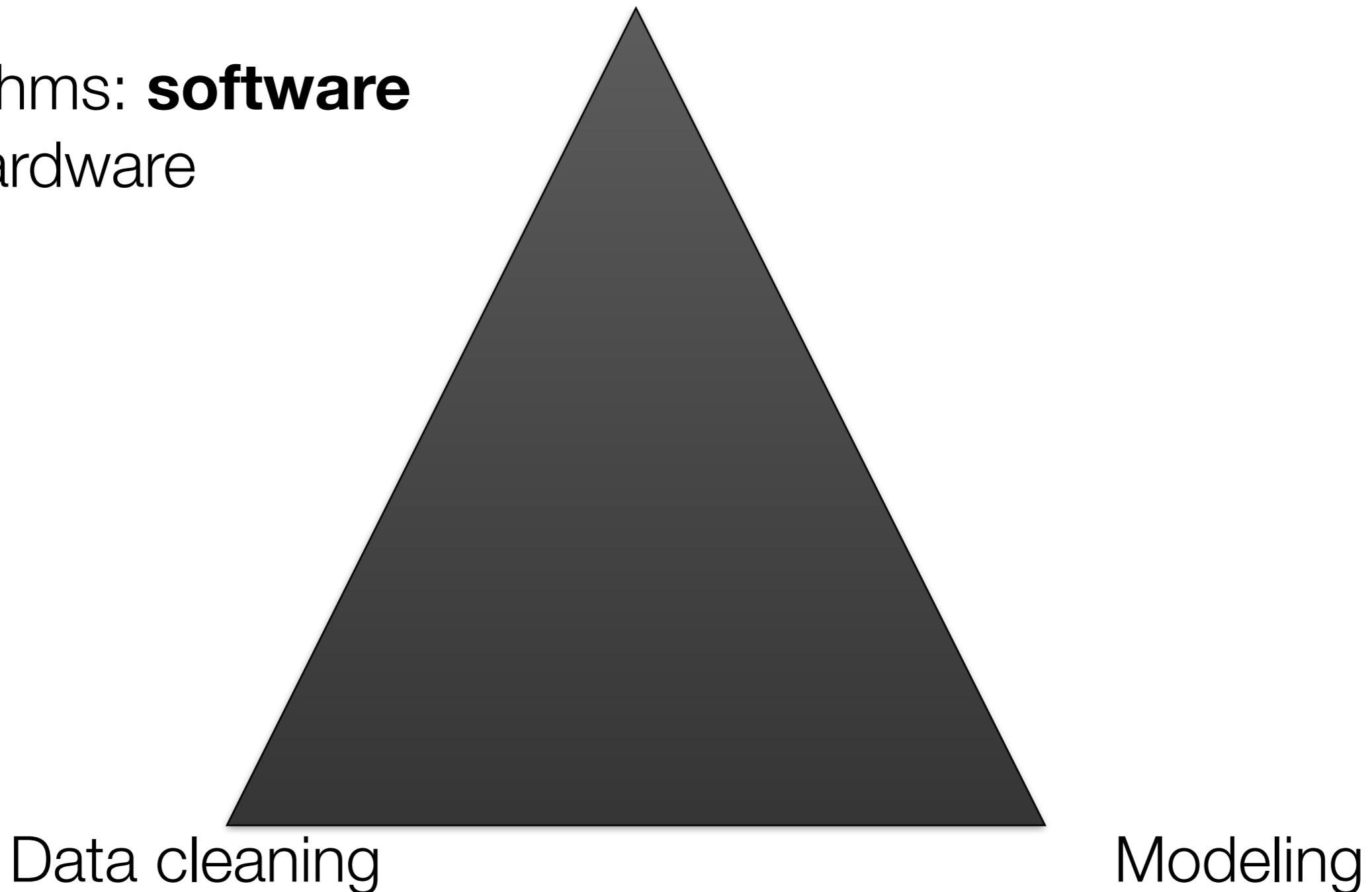
Algorithms: software
and hardware



Learning objectives

Computation/Algorithms

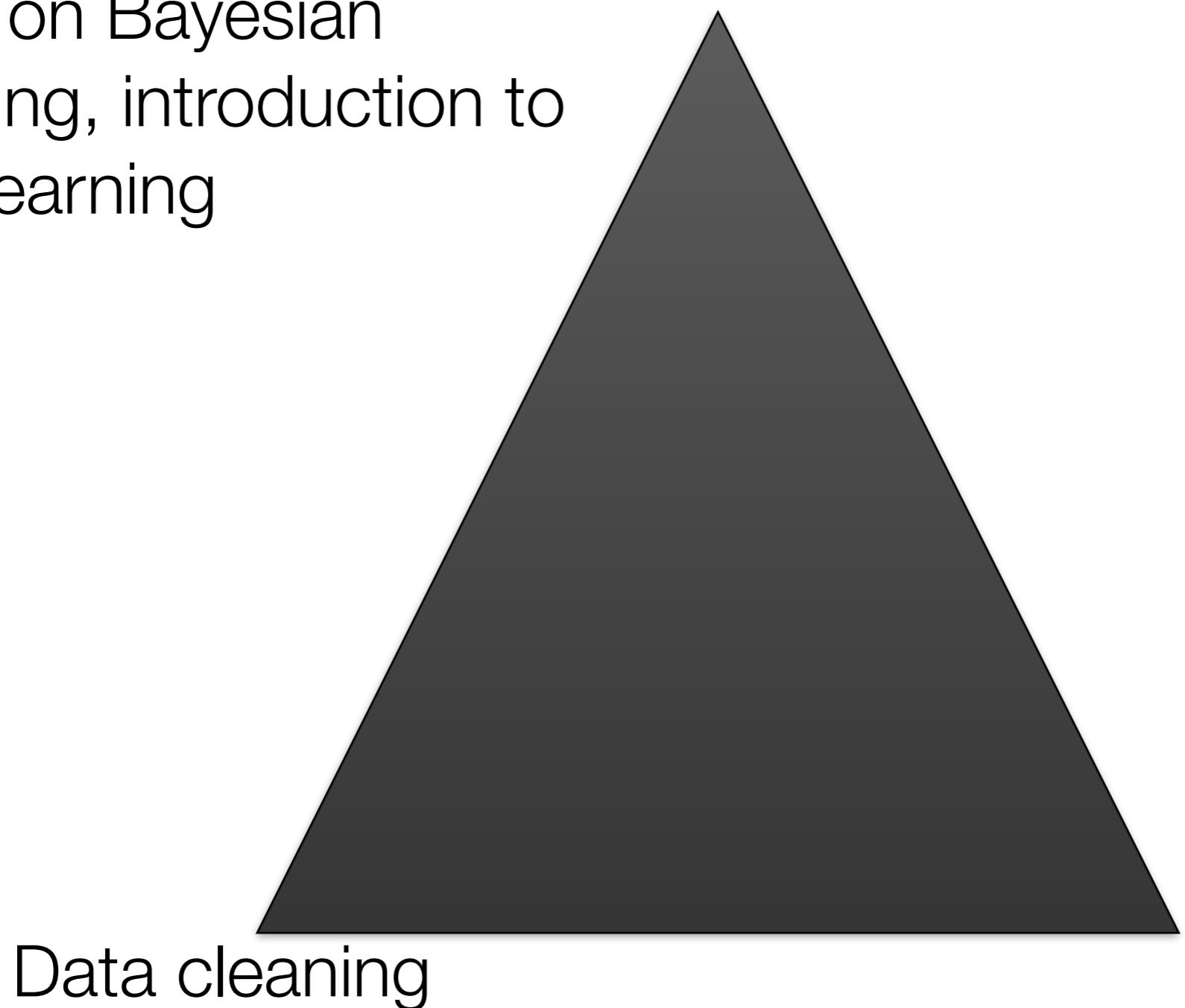
Algorithms: **software**
and hardware



Learning objectives

Focus on Bayesian modeling, introduction to deep learning

Computation/Algorithms



Transformation of many industries

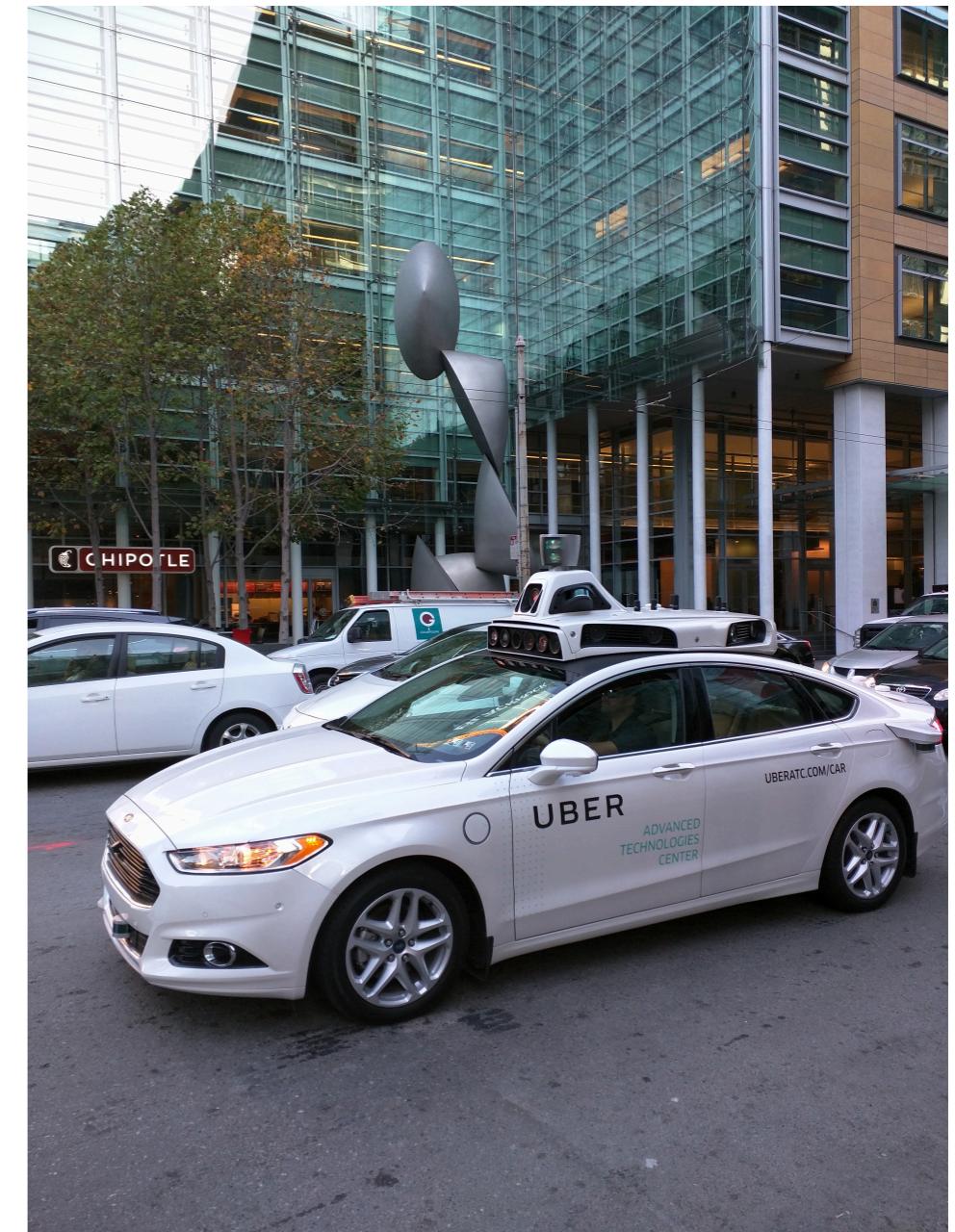


XING[®]
POWERING RELATIONSHIPS

Google+



Google Cloud Platform Live

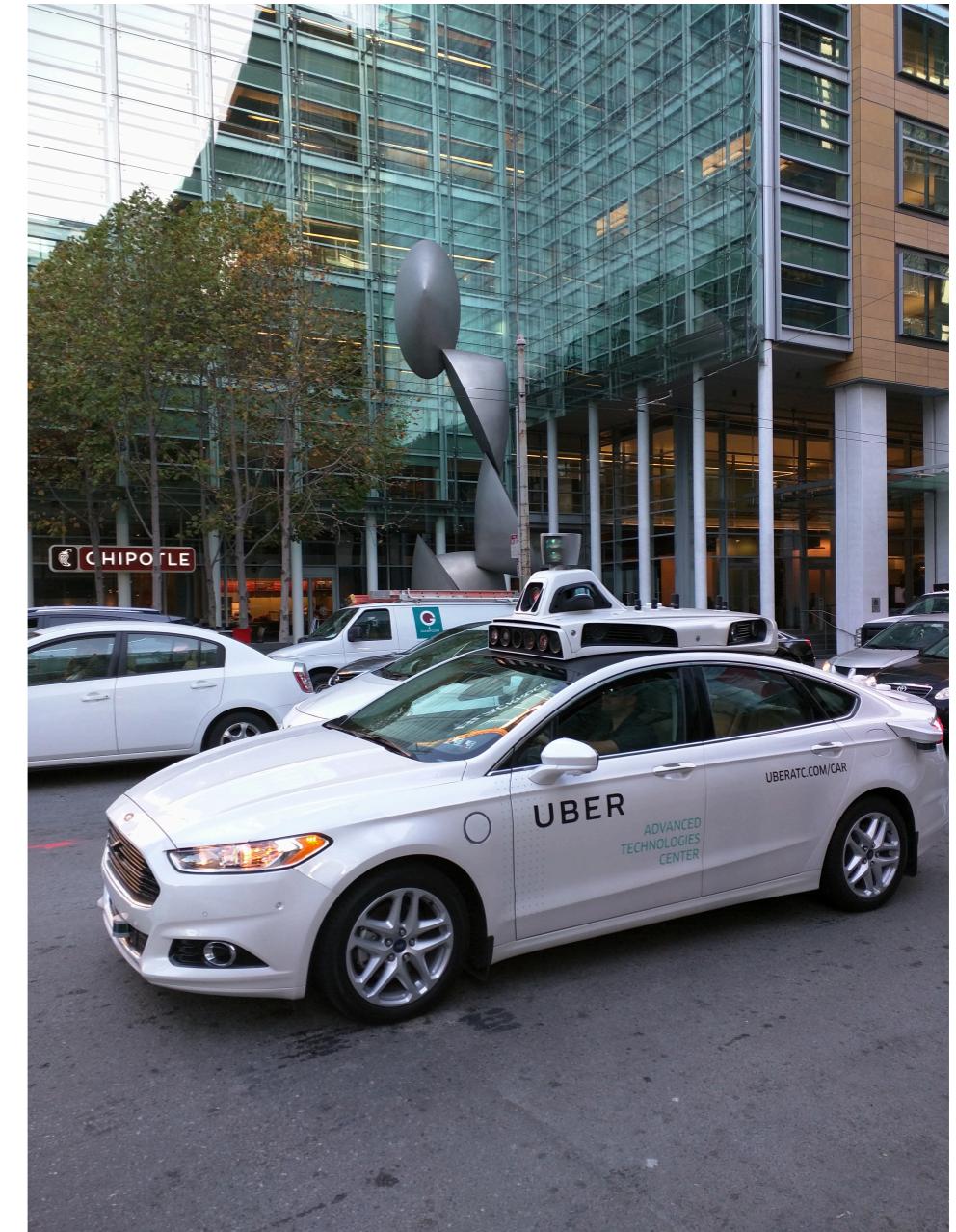


Transformation of many industries



XING[®]
POWERING RELATIONSHIPS

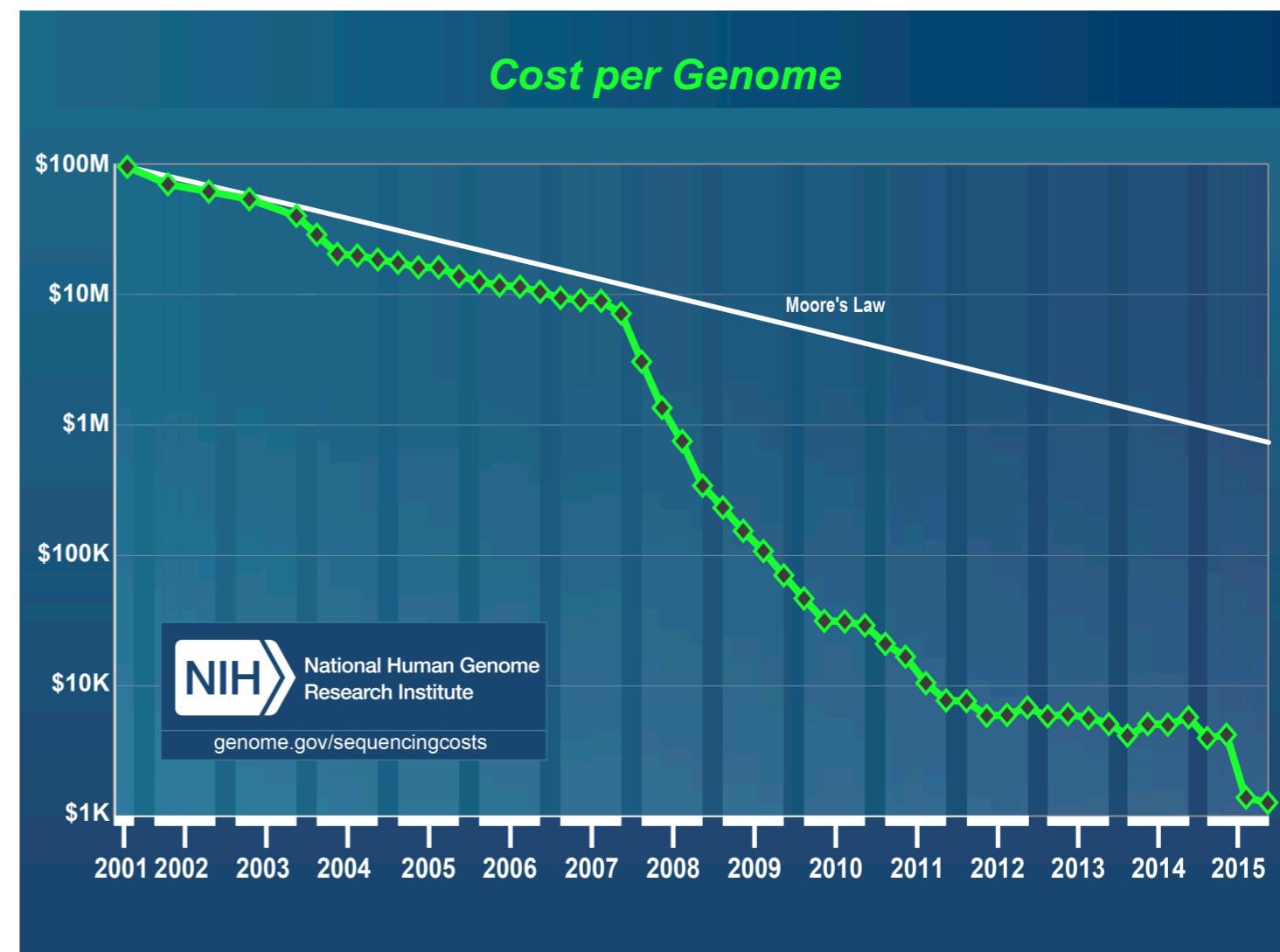
Google+



What is missing?

Technologies transforming biomedicine

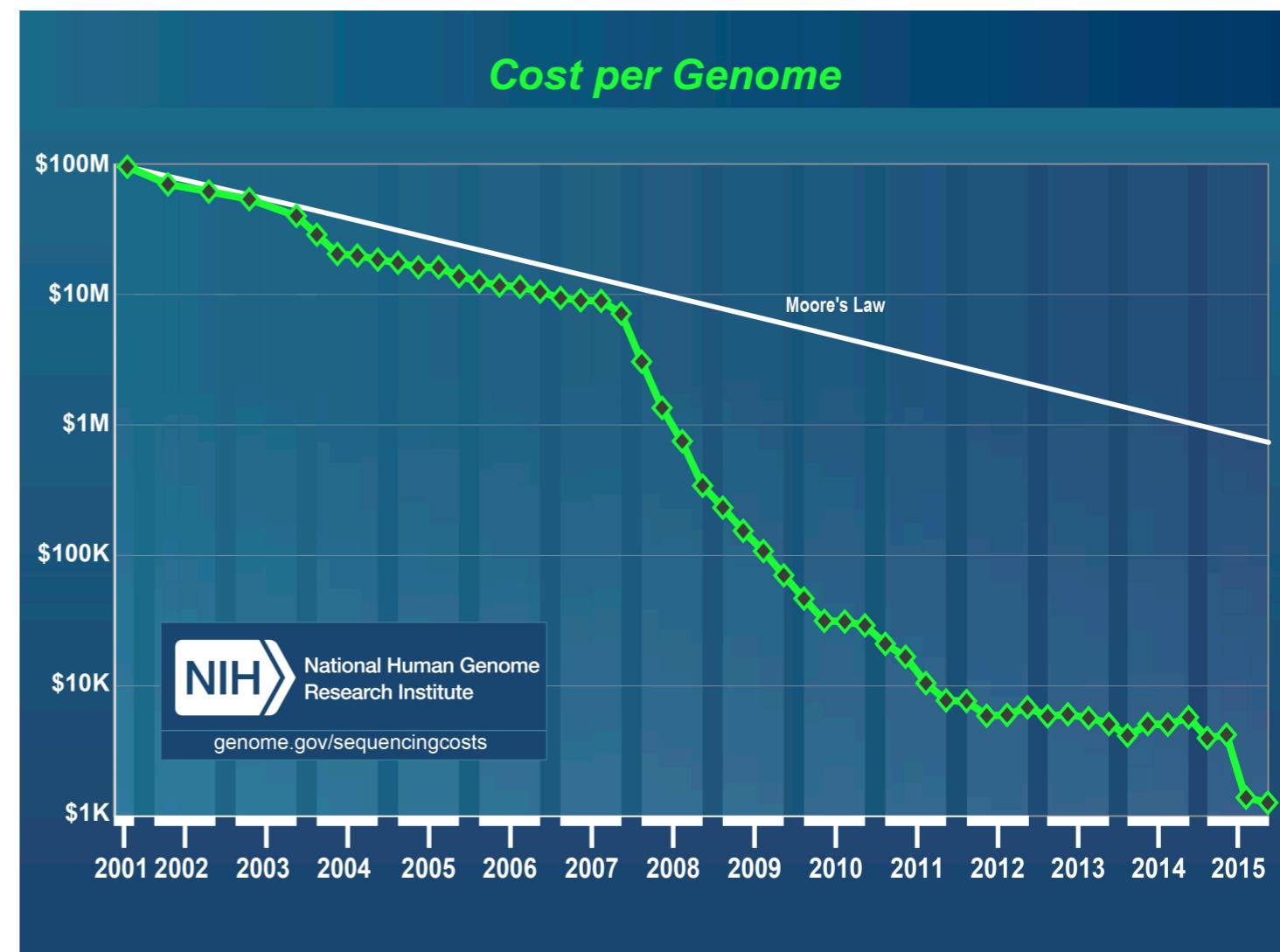
Cost of sequencing has plummeted over the past 15 years



Technologies transforming biomedicine

Cost of sequencing has plummeted over the past 15 years

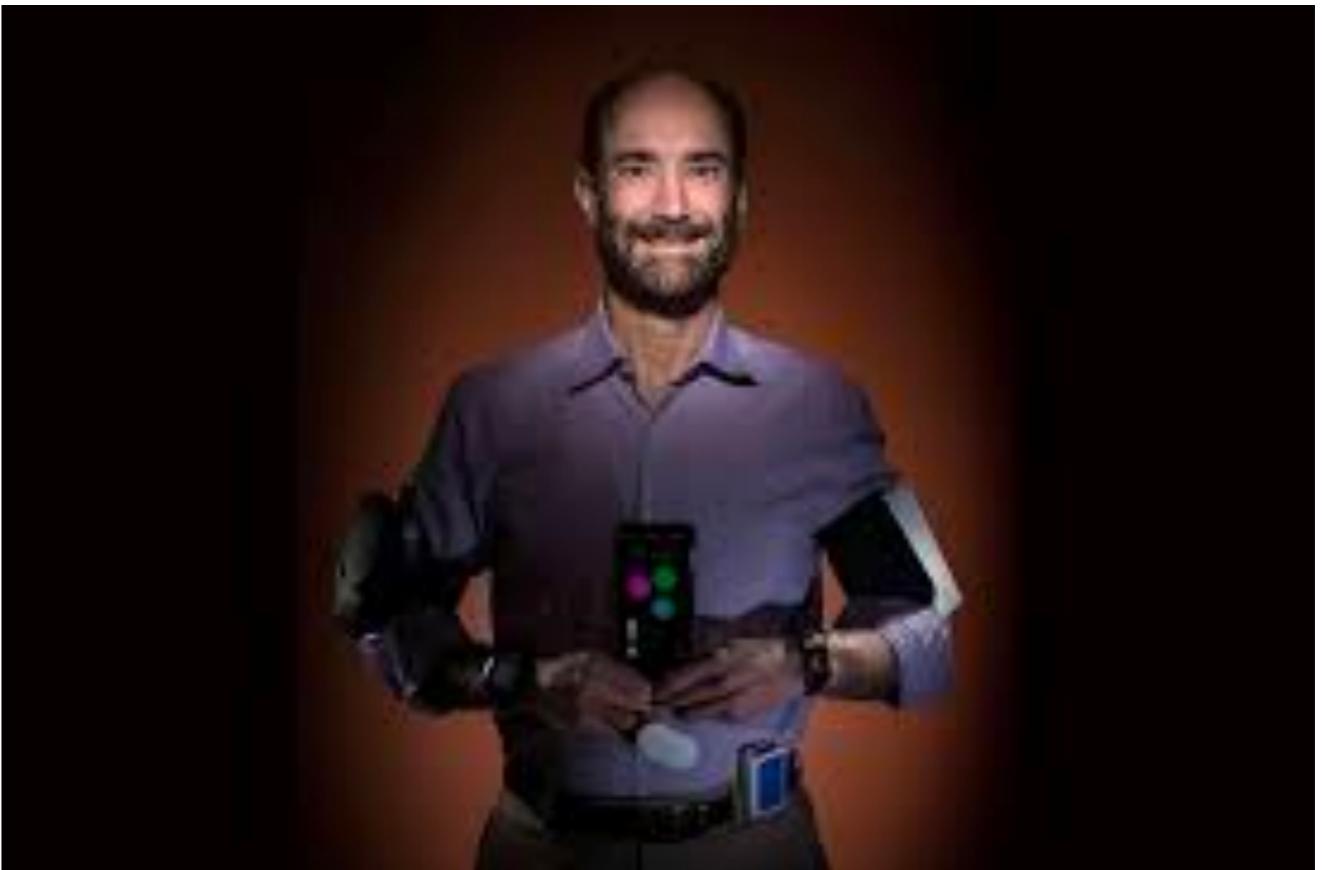
~\$1000 cost point projected for 18/19



Technologies transforming biomedicine

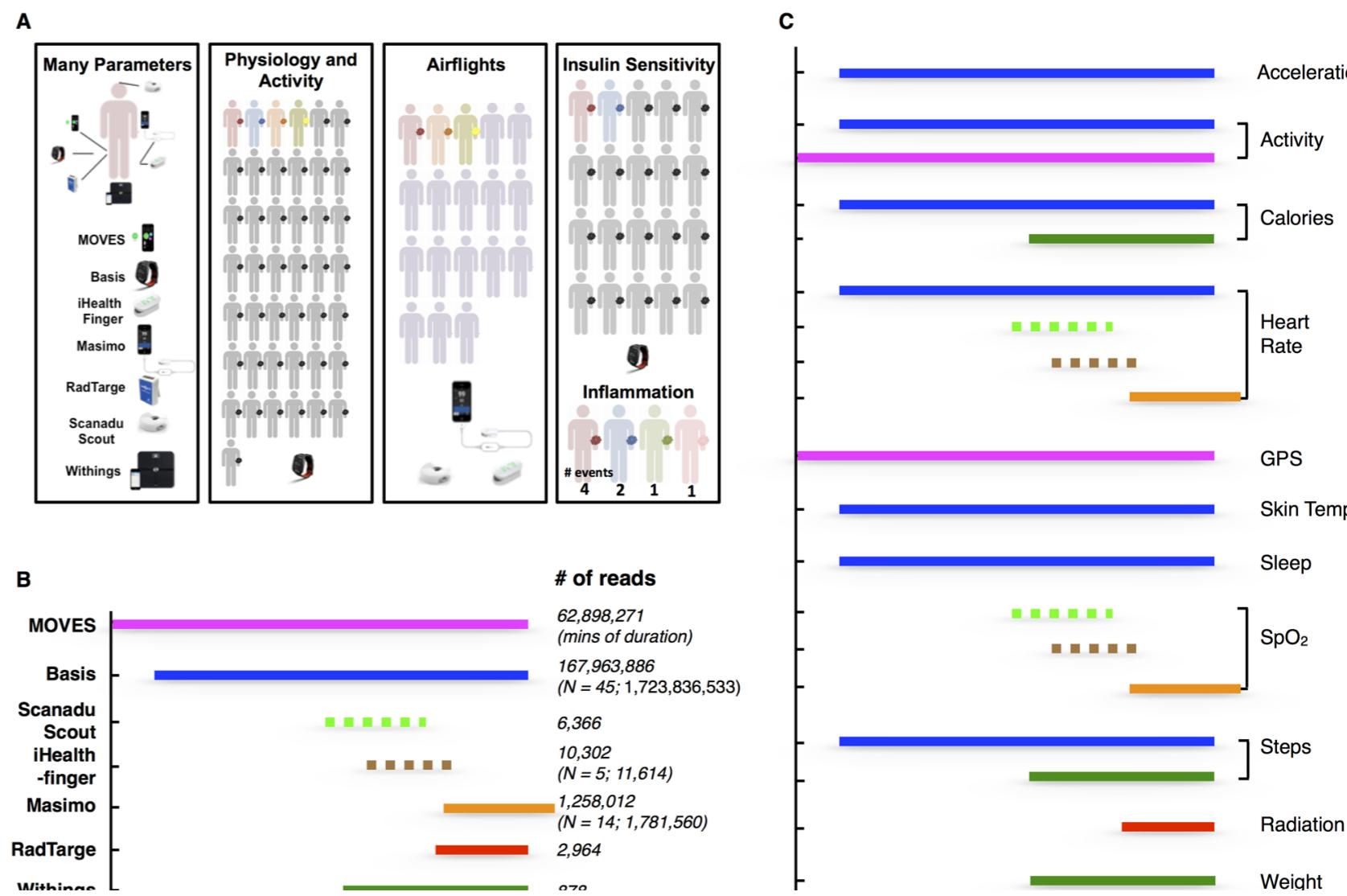
Wearables and sensors

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

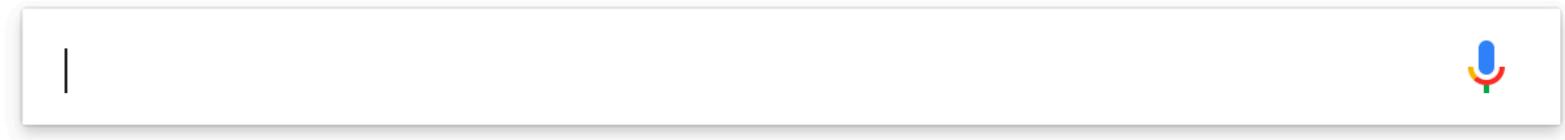
Data streams from
individuals participating
in social networks



Social network data

Technologies transforming biomedicine

Data streams
from individual's
search activity



Google Search

I'm Feeling Lucky

Search engine data

Technologies transforming biomedicine



Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer²,
Mark S. Smolinski¹ & Larry Brilliant¹

¹Google Inc. ²Centers for Disease Control and Prevention

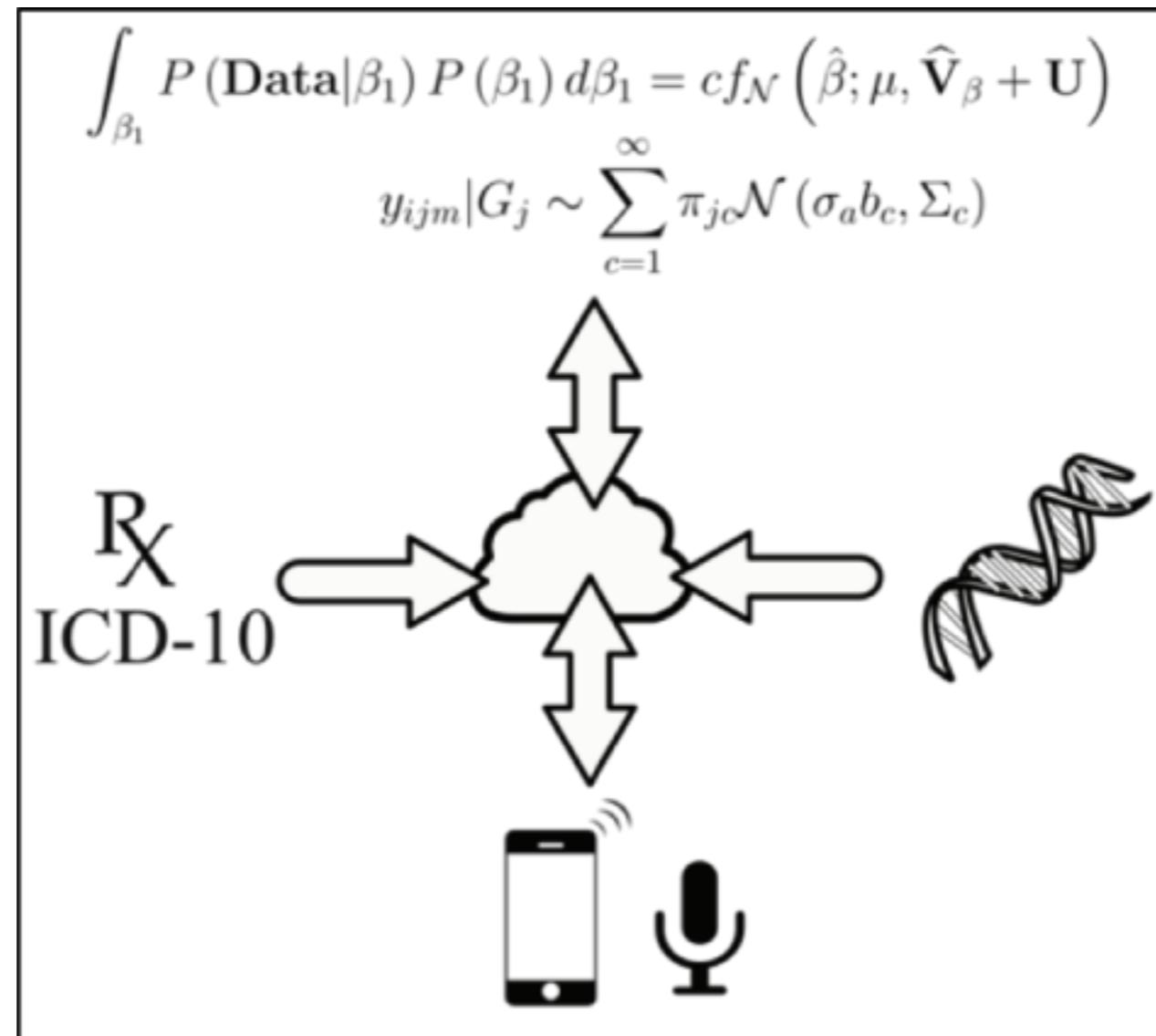
Search engine data

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?



Challenge for this generation

Precision health and biobank initiatives



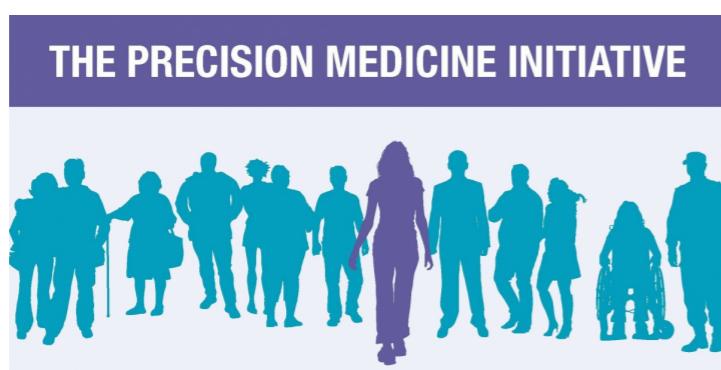
Precision health and biobank initiatives



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

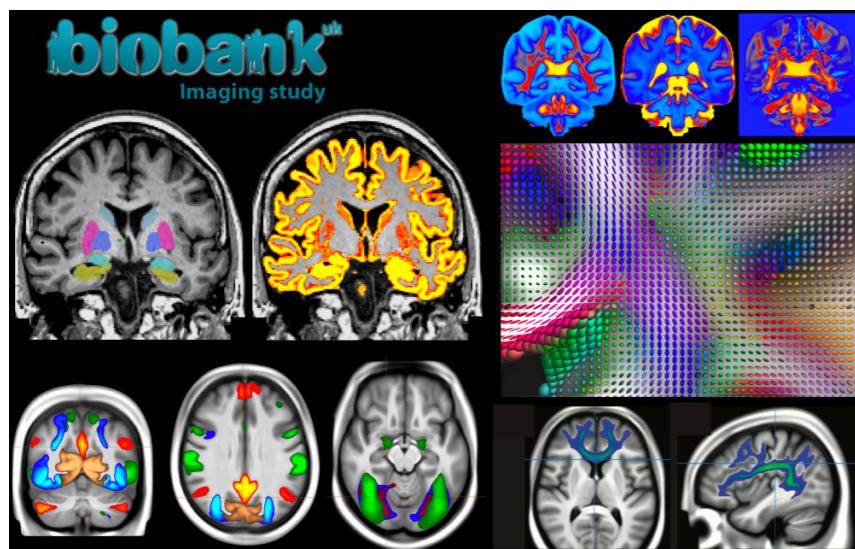


Hospital in-patient record

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

[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.

Legal notice: Without a written licence from UK Biobank, you may not copy, reproduce, republish, download, distribute, make available to the public or otherwise use any of the content displayed on this website in whole or in part or permit or assist any third party to do the same, except to the extent permitted at law.

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

Summary generated 4 February 2017

Top Level

Level 1

Level 2

Level 3

Improving the health of future generations

UK Biobank data showcase webpage

[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

[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



[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

Top Level

Level 1

Level 2

Level 3

Summary generated 4 February 2017

Improving the health of future generations

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:

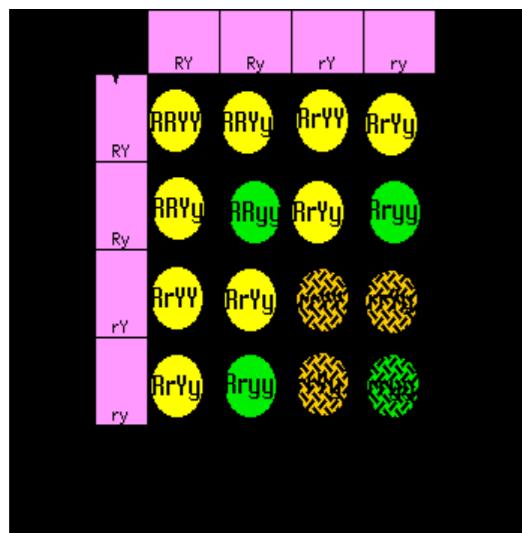
Your neighbor (unrelated)?

Your sibling?

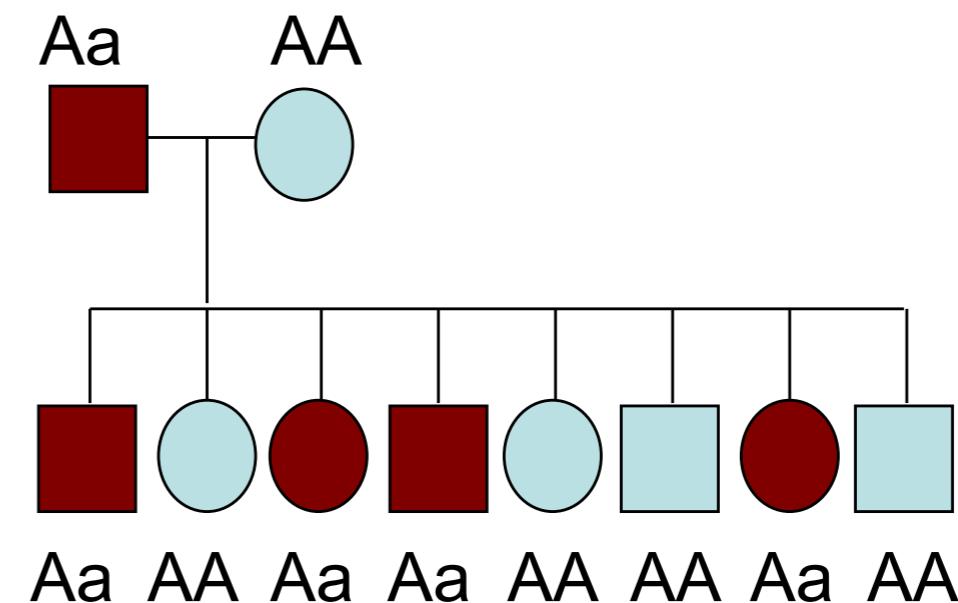
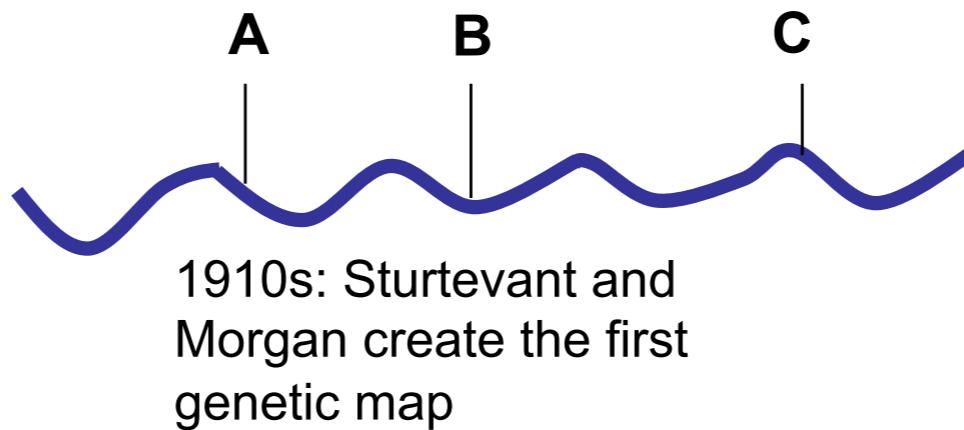
Your twin?

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



Precision Medicine



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



Lower
risk for
disease



Examples of protective mutations

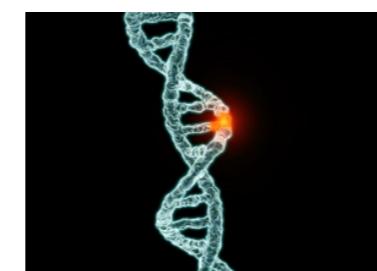
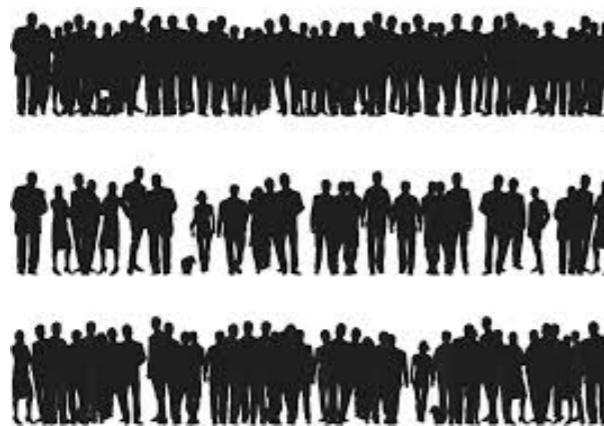
PCSK9 for LDL and MI

Nav 1.7 for pain

CARD9 for Crohn's disease and ulcerative colitis

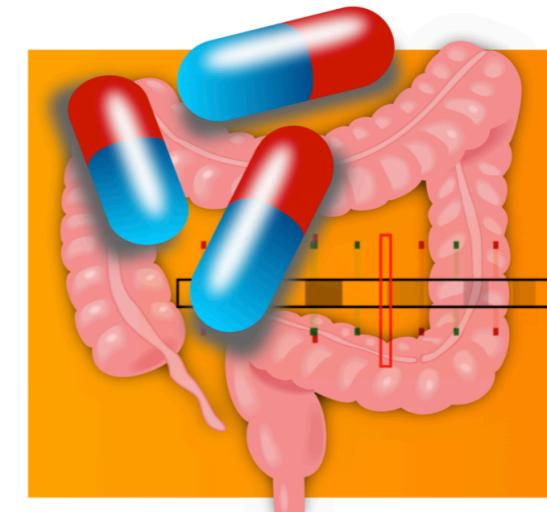
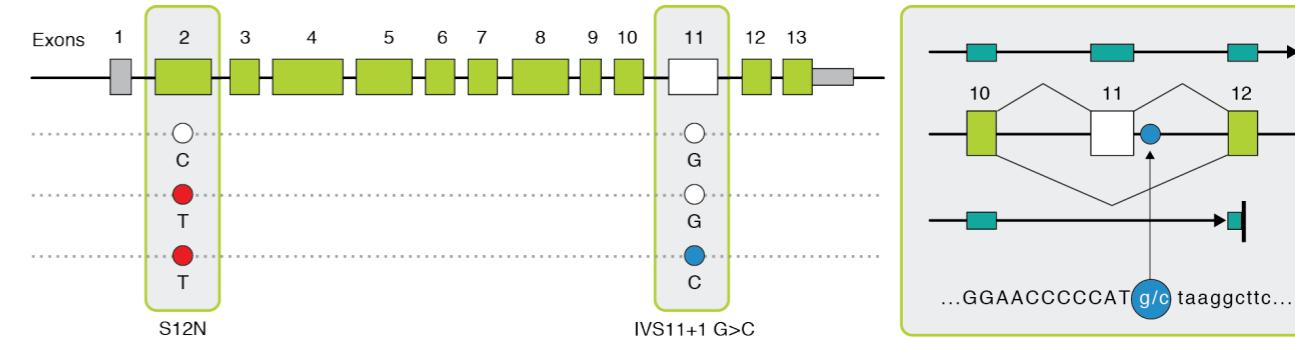
RNF186 for ulcerative colitis

CCR5 for HIV



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

[!\[\]\(85b0f06a00119c268054533155f06449_img.jpg\) Select Language ▾](#)

Global Biobank Engine (pre-alpha)

Search for a gene or variant or region or phenotype coding (coming soon)

Examples - Gene: [F5](#), Transcript: [ENST00000367797](#), Variant: [1:169519049](#), RS ID: [rs6025](#), Region: [10:114686614-114786614](#)

Genetic Association Results

Note: 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.

Browseable phenotypes



Up next

October 10-18, 2017

- Imputation results upload with ([Neale Lab](#)).

Recent News

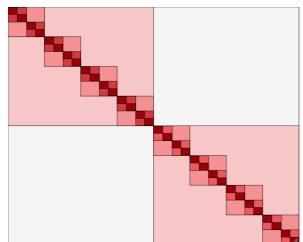
October 18, 2017

- Release of genetic parameters app I: [genetic correlation](#).

October 18, 2017

- Release of [decomposition app](#).

<https://biobankengine.stanford.edu>



Global Biobank Engine

Global Biobank Engine

About Downloads Terms Contact Power FAQ & News archive

Select Language ▾

Global Biobank Engine (pre-alpha)

Search for a gene or variant or region or phenotype coding (coming soon)

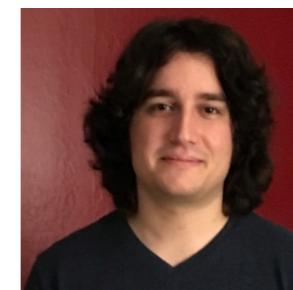
Examples - Gene: [F5](#), Transcript: [ENST00000367797](#), Variant: [1:169519049](#), RS ID: [rs6025](#), Region: [10:114686614-114786614](#)

Genetic Association Results

Note: 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.

Browseable phenotypes

- [Cancer](#)
- [Family History](#)
- [Verbal and ICD disease grouping](#)
- [Quantitative measures](#)



Up next

August 10-18, 2017

- Imputation results upload with ([Neale Lab](#)).

Recent News

September 1, 2017

- Check out quantitative measures II.

August 24, 2017

- GBE Power App available.

August 18, 2017

- Support for gene-based results: top five phenotypes per gene (independent effects model; more documentation soon).

August 13, 2017



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

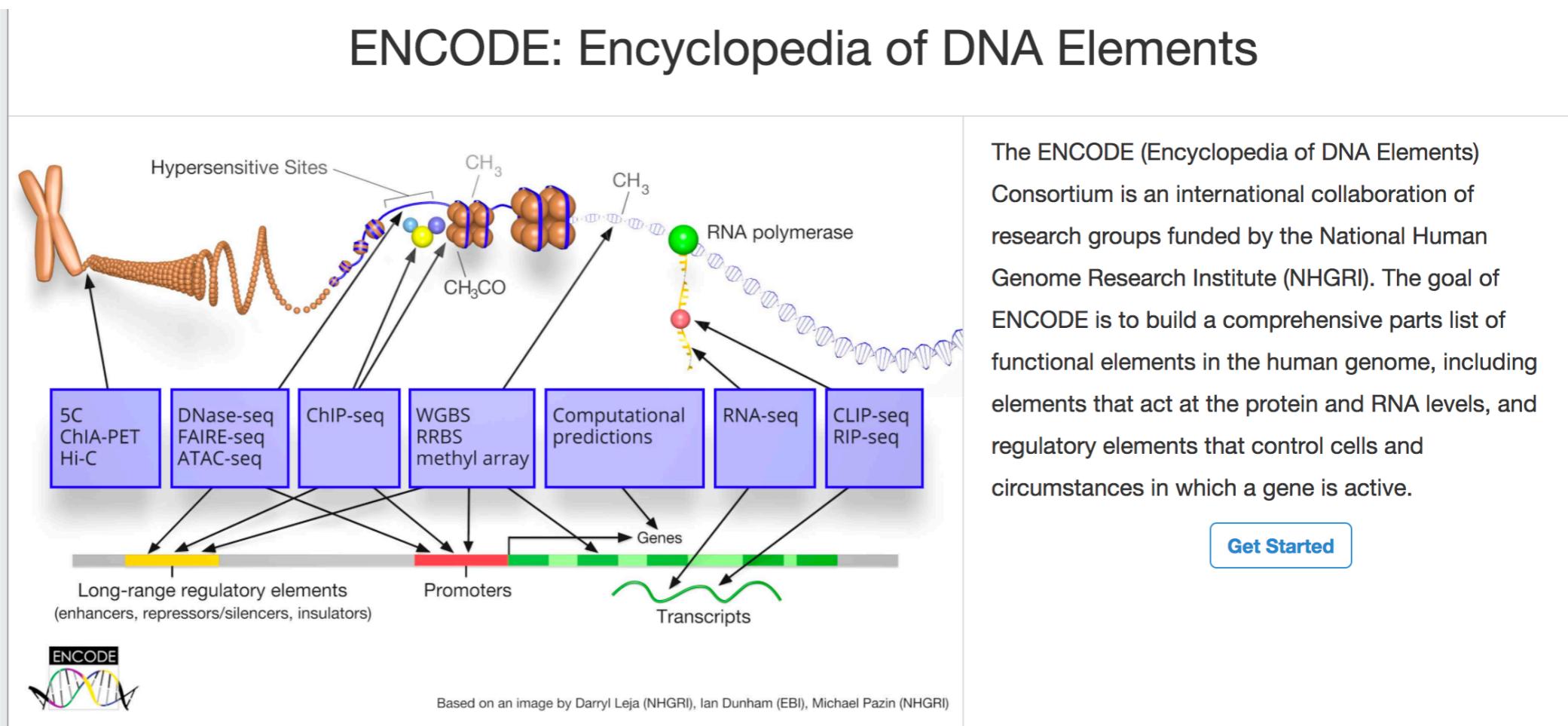
2. ENCODE

Data available in the course

1. UK Biobank

Summary statistic data for over 100 diseases

2. ENCODE



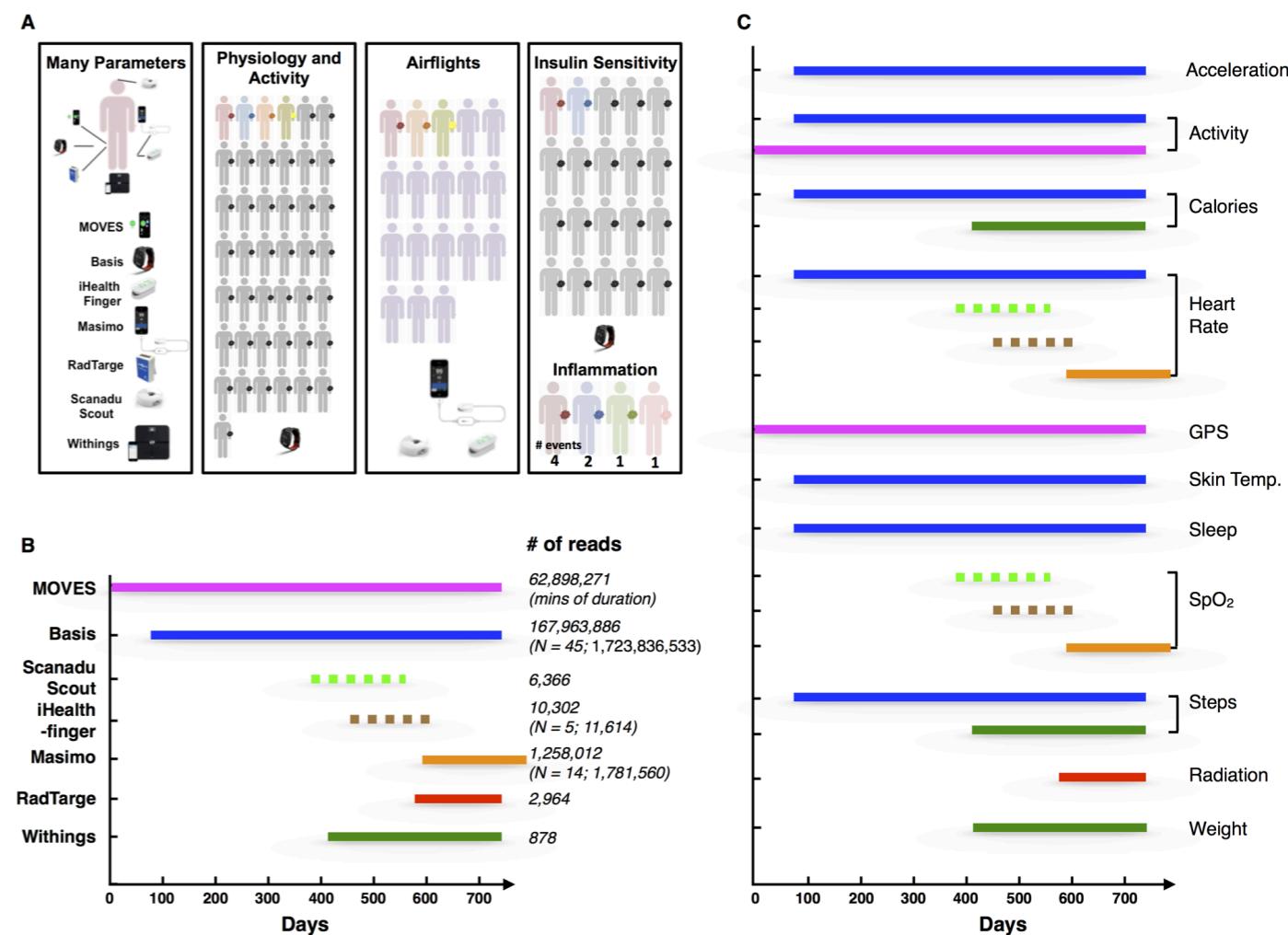
Data available in the course

1. UK Biobank

Summary statistic data for over 100 diseases

2. ENCODE

3. Wearable Biosensor



Data available in the course

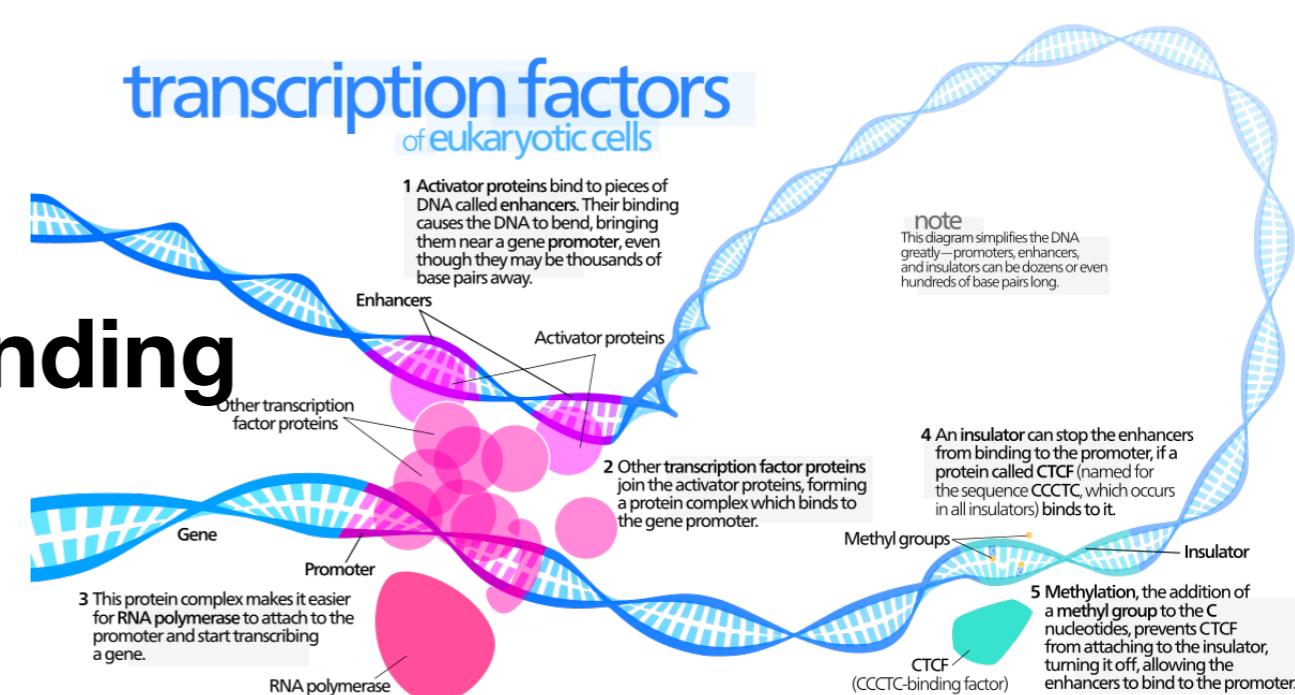
1. UK Biobank

Summary statistic data for over 100 diseases

2. ENCODE

3. Wearable Biosensor

4. Transcription factor binding



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

Chambers: more emphasis on data preparation and presentation

50 years of Data Science

Breiman: more emphasis on prediction

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

The Six Divisions of Greater Data Science

1. Data exploration and preparation

The Six Divisions of Greater Data Science

1. Data exploration and preparation
2. Data representation and transformation

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

The Six Divisions of Greater Data Science

1. Data exploration and preparation
2. Data representation and transformation
3. Computing with data
4. Data modeling

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

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

Poisson Models

Rank based methods - permutation

Monte Carlo, Markov Chains

Splines, Fourier analysis, PCA

Bayesian multilevel modeling

Bayesian mixture models

Bayesian prediction

Distributed computing with privacy preserving schemes

Convolution Neural Networks, LSTM

Course website

Syllabus

Reading materials

Lecture notes

<https://canvas.stanford.edu/courses/66507>

Problem Sets

Data links

Sample STAN programs

Application to data

<https://biods215.github.io/>

Github repository

Github repository for the course



Welcome to BIODS215 Topics in Biomedical Data Science: Large-scale inference

This page will be used to host Github repositories for the course.

Course Instructors

[Manuel A. Rivas](#)

[Julia Salzman](#)

[James Zou](#)

<https://biods215.github.io/>