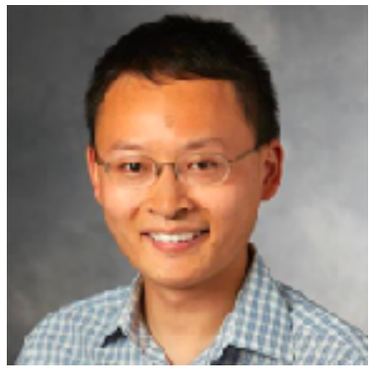
BIODS215

Topics in Biomedical Data Science: Large-scale inference

Spring Quarter 2017

Course Instructors

Prof. Manuel A. Rivas
365 Lasuen Street
Littlefield Room 337
mrivas@stanford.edu
rivaslab.stanford.edu



Prof. James Zou
Littlefield Room 334
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/

Lecture structure

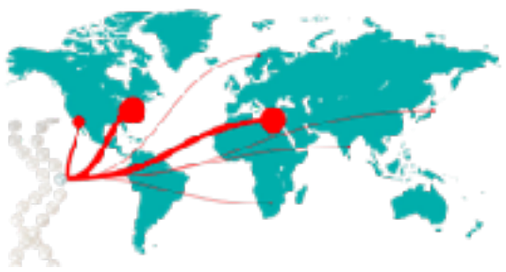
~20-45 minutes motivating biomedical example

~30-55 minutes statistical inference concept lecture

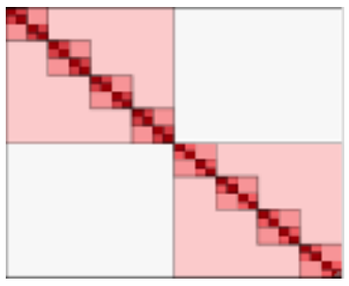
5-7 minute break in the middle

RIVASLAB

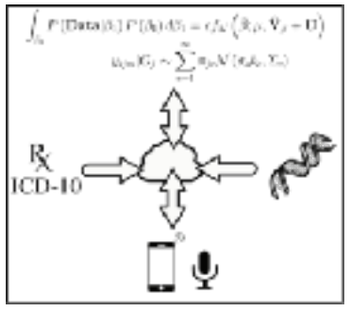
Generating effective therapeutic hypotheses



Genetic epidemiology



High-dimensional methods development and optimization



Technology development

Rivas Lab develops statistical models and computational tools for population-scale studies using genomic and phenotype data.

Zou Lab

Rigorous adaptive data analysis

Deep learning for genomics and biomedical imaging

Leveraging side information and randomization in hypothesis testing

Geometry of neural networks

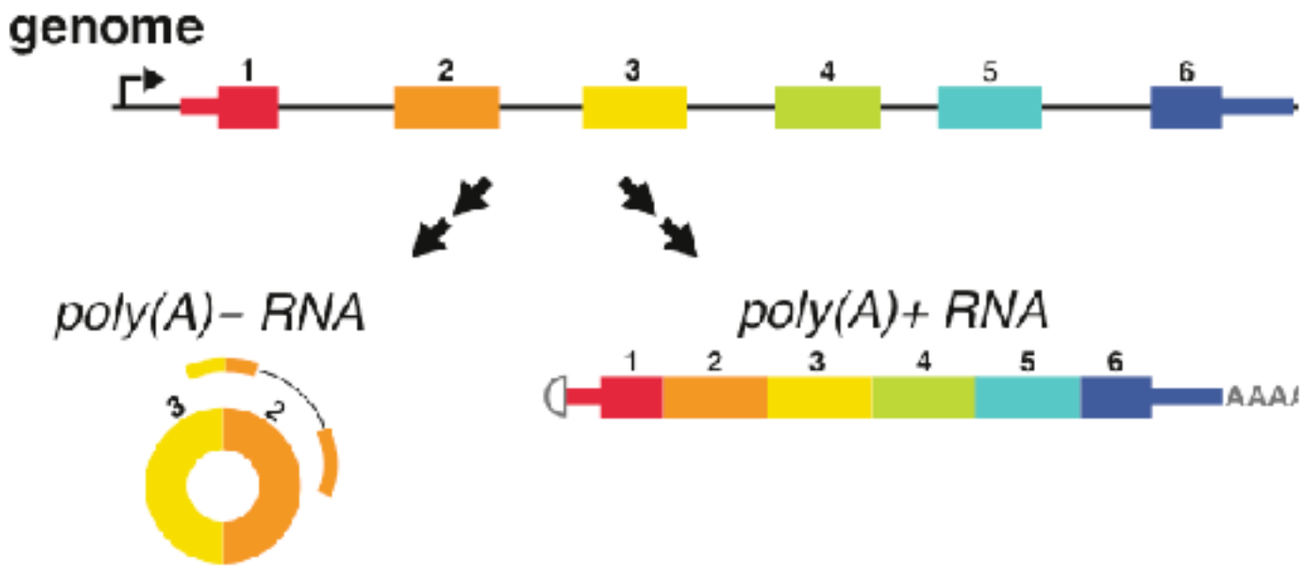
Machine learning for synthetic biology
Contrastive learning

Salzman Lab

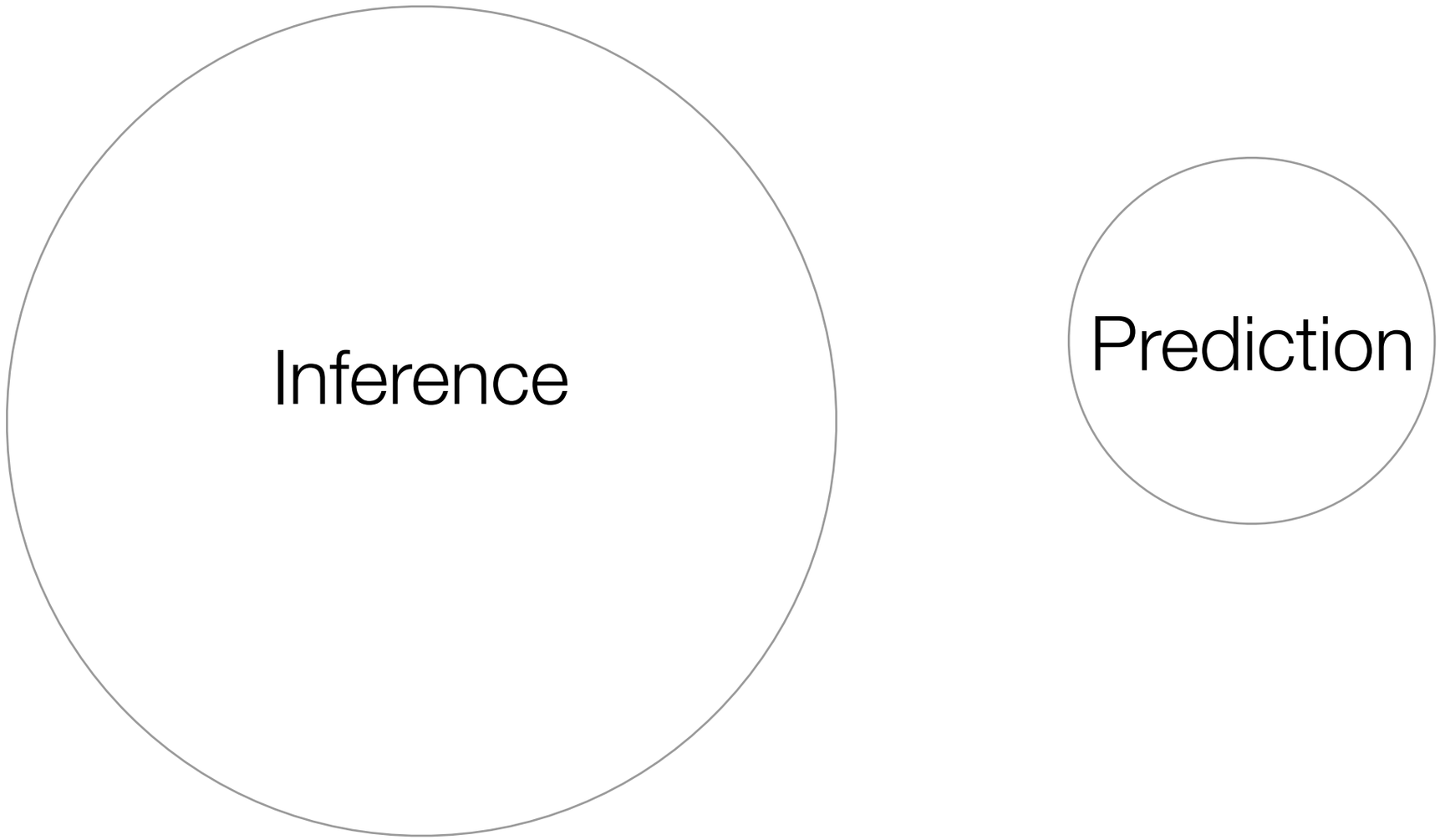
Circular RNA

Structural Variation in Human Cancer

Statistical Approaches for Next-Generation Sequencing Data



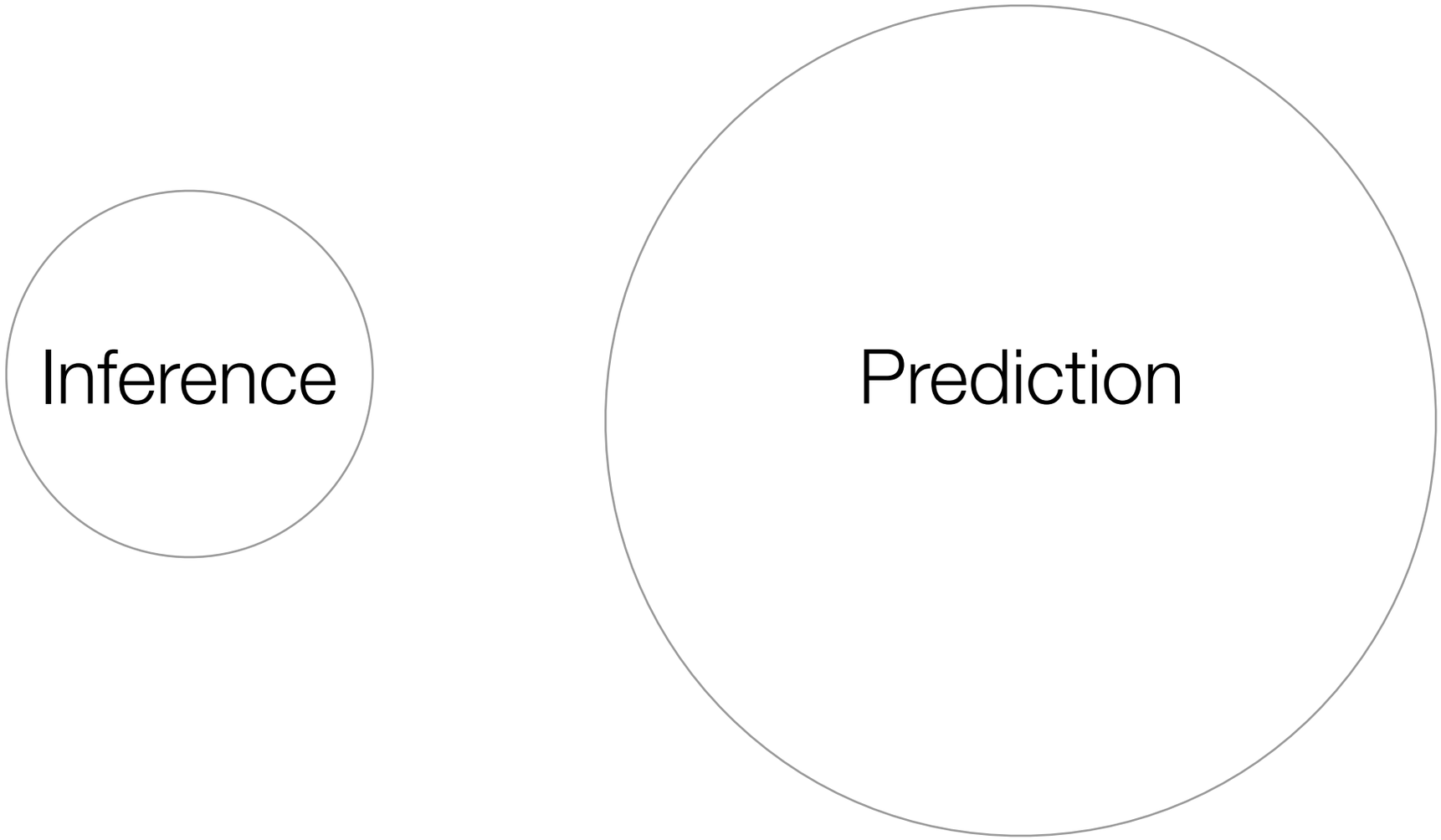
Data explosion and worldview across fields



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

Statisticians, Biomedicine (therapeutics)

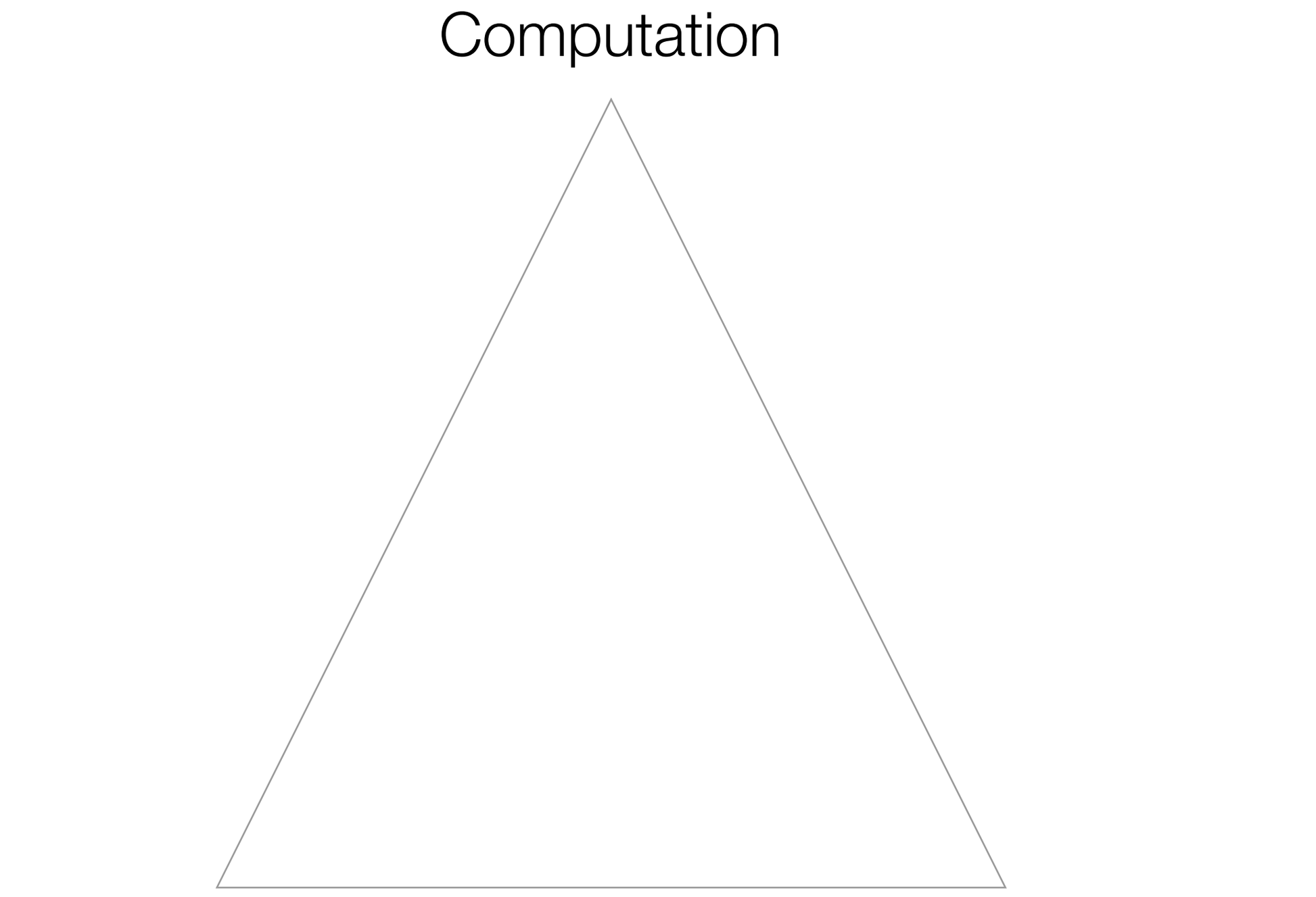
Data explosion and worldview across fields



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

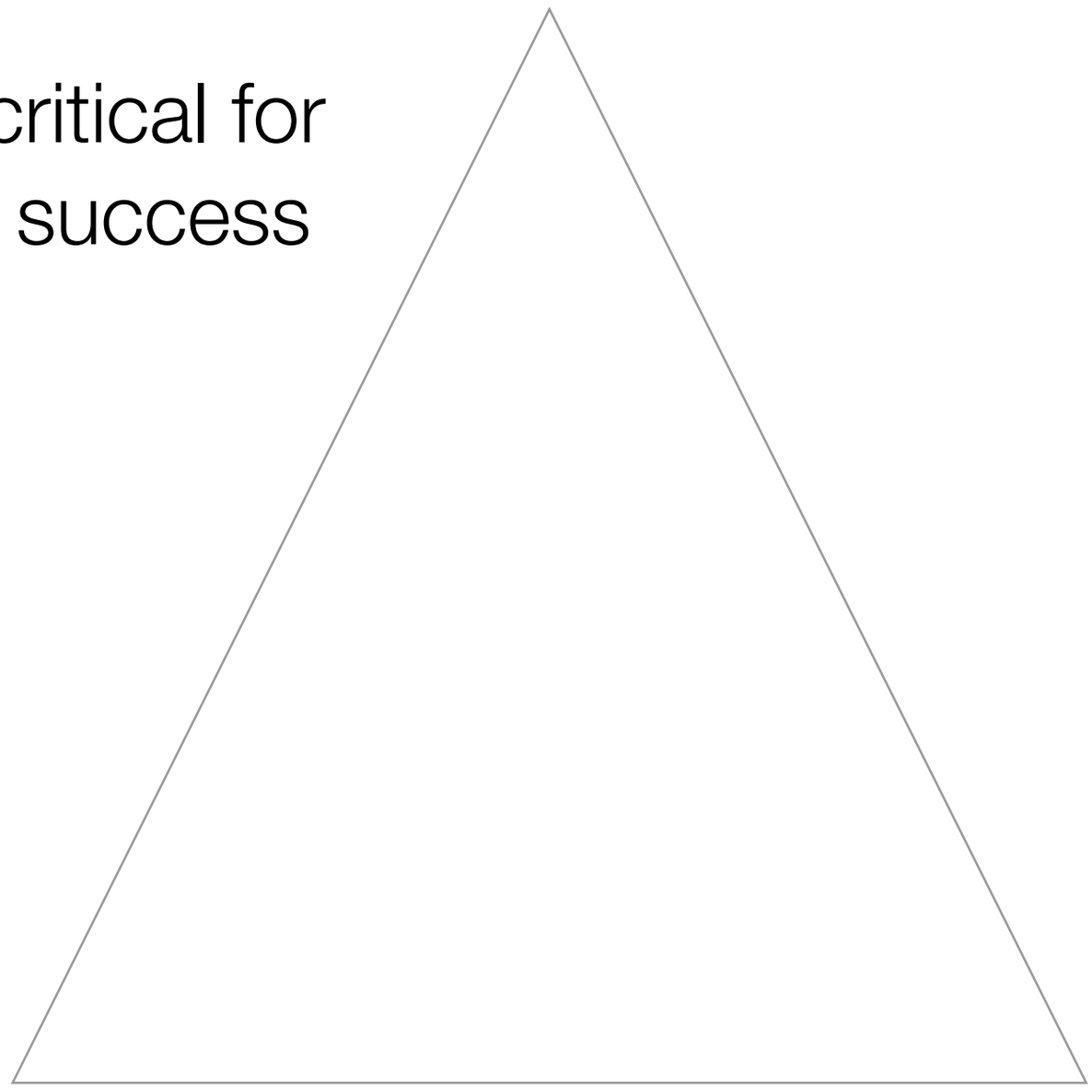
Machine Learning, Computer science

Learning objectives



Learning objectives

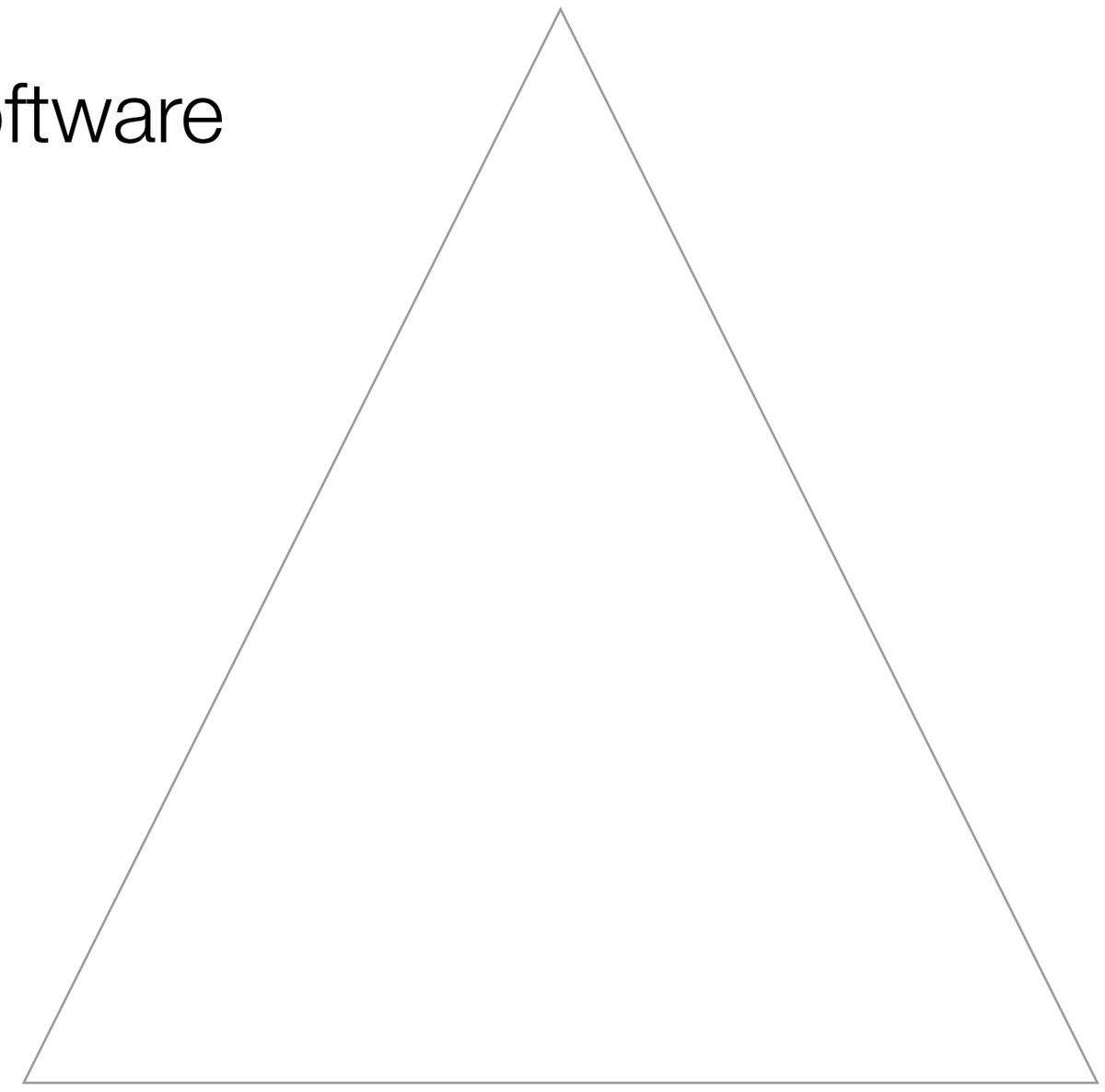
Not fun, but critical for downstream success



Learning objectives

Computation

Algorithms: software and hardware



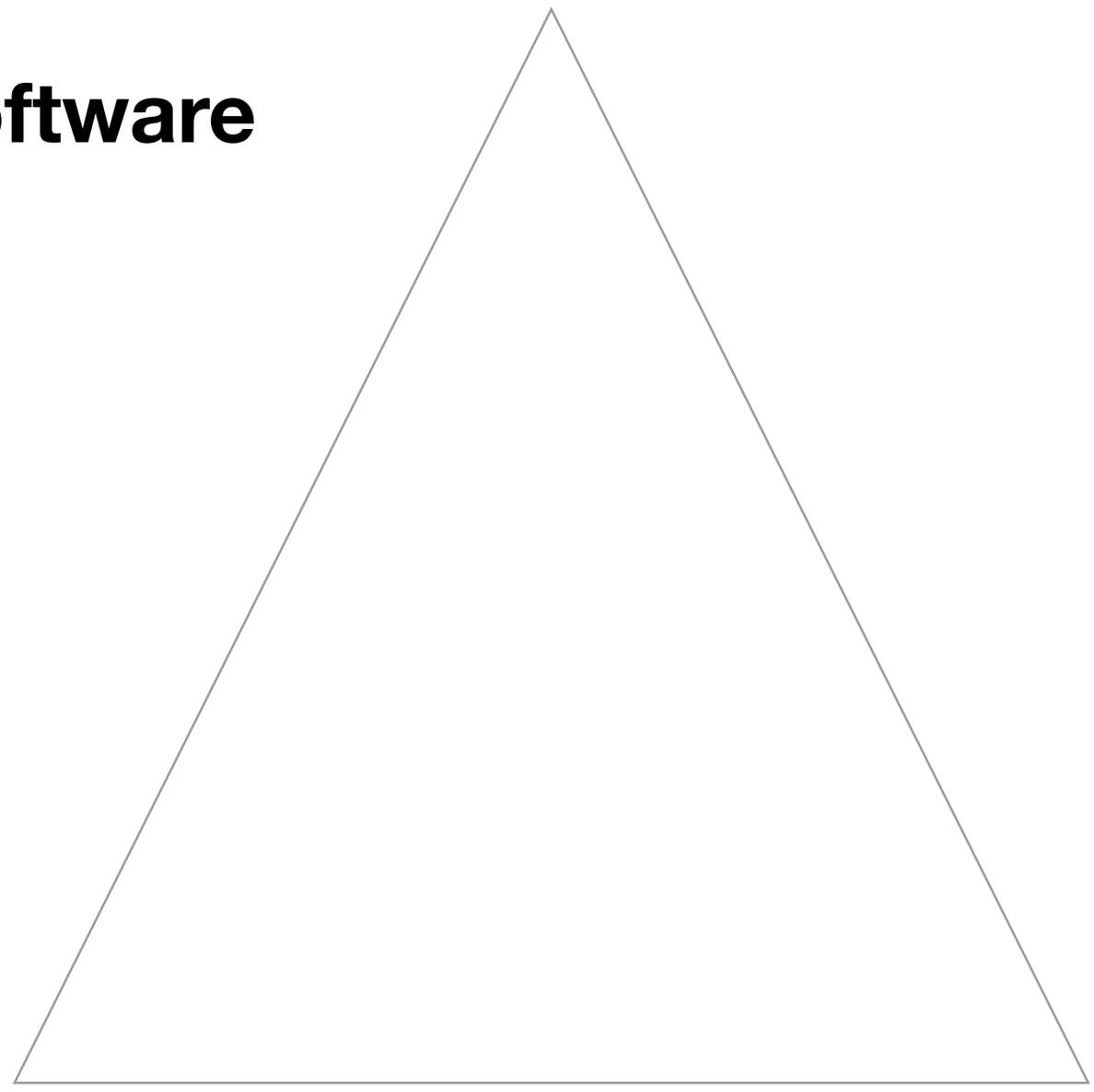
Data cleaning

Modeling

Learning objectives

Computation

Algorithms: **software** and hardware

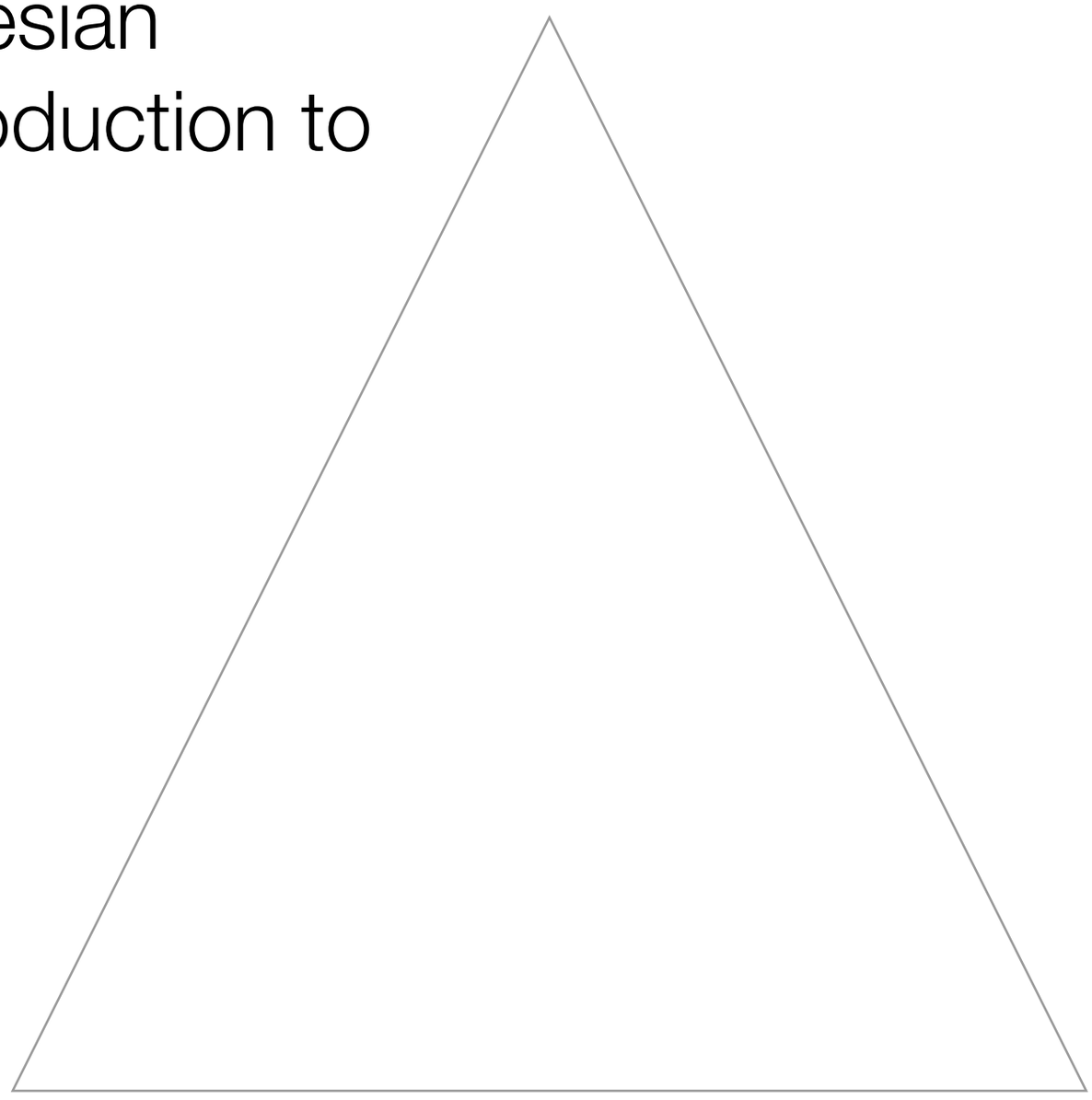


Data cleaning

Modeling

Learning objectives

Focus on Bayesian modeling, introduction to deep learning



Course requirements and grading

Two homework assignments (40%)

Final project (50%)

Class participation (10%)

Syllabus

Date	Lecturer	Topic
4/4	Manuel	Intro to biomedical data, computing, and data repositories
4/6	Manuel	Model comparison and hypothesis testing
4/11	Manuel	Meta-analysis and variance-components
4/13	Manuel	Regression - The Linear model and discrete data models

Week 1 and 2

Syllabus

Date	Lecturer	Topic
4/18	James Z.	Sampling algorithms for inference
4/20	Julia S.	Permutation testing and monte carlo for p-value and FDR computation
4/25	Julia S	Poisson Arrivals
4/27	Julia S	Martingales

Week 3 and 4

Syllabus

Date	Lecturer	Topic
5/2	Manuel	Hierarchical modeling
5/4	Manuel	Mixture model case studies and computation
5/9	Manuel	High-dimensional inference methods on summary statistics
5/11	Manuel, James, Julia S	Students present project plans

Week 5 and 6

Syllabus - students present project plans

Date	Lecturer	Topic
5/2	Manuel	Hierarchical modeling
5/4	Manuel	Mixture model case studies and computation
5/9	Manuel	High-dimensional inference methods on summary statistics
5/11	Manuel, James, Julia S	Students present project plans

Week 5 and 6

Syllabus - guest lecture by Prof. Julia Palacios

Date	Lecturer	Topic
	Guest lecture	
5/16	Prof. Julia Palacios	Gaussian Process regression
	Guest Lecture	
5/18	Prof. Julia Palacios	Dirichlet Process

Week 7

Department of Statistics
Biomedical Data Science
http://juliapalacios.github.io/

Syllabus

Date	Lecturer	Topic
5/23	Julia S.	Advanced modern topics in statistical cancer genomics (with normal population variation as controls)
5/25	James Z.	Search engine data for public health

Week 8

Syllabus

Date	Lecturer	Topic
5/30	James Z.	Deep learning
6/1	Manuel	Inference, prediction, and risk modeling from biomedical data repositories

Week 9

Syllabus

Date	Lecturer	Topic
6/6	Manuel, James, Julia S	Final Project presentations
		Final project due.

Week 10

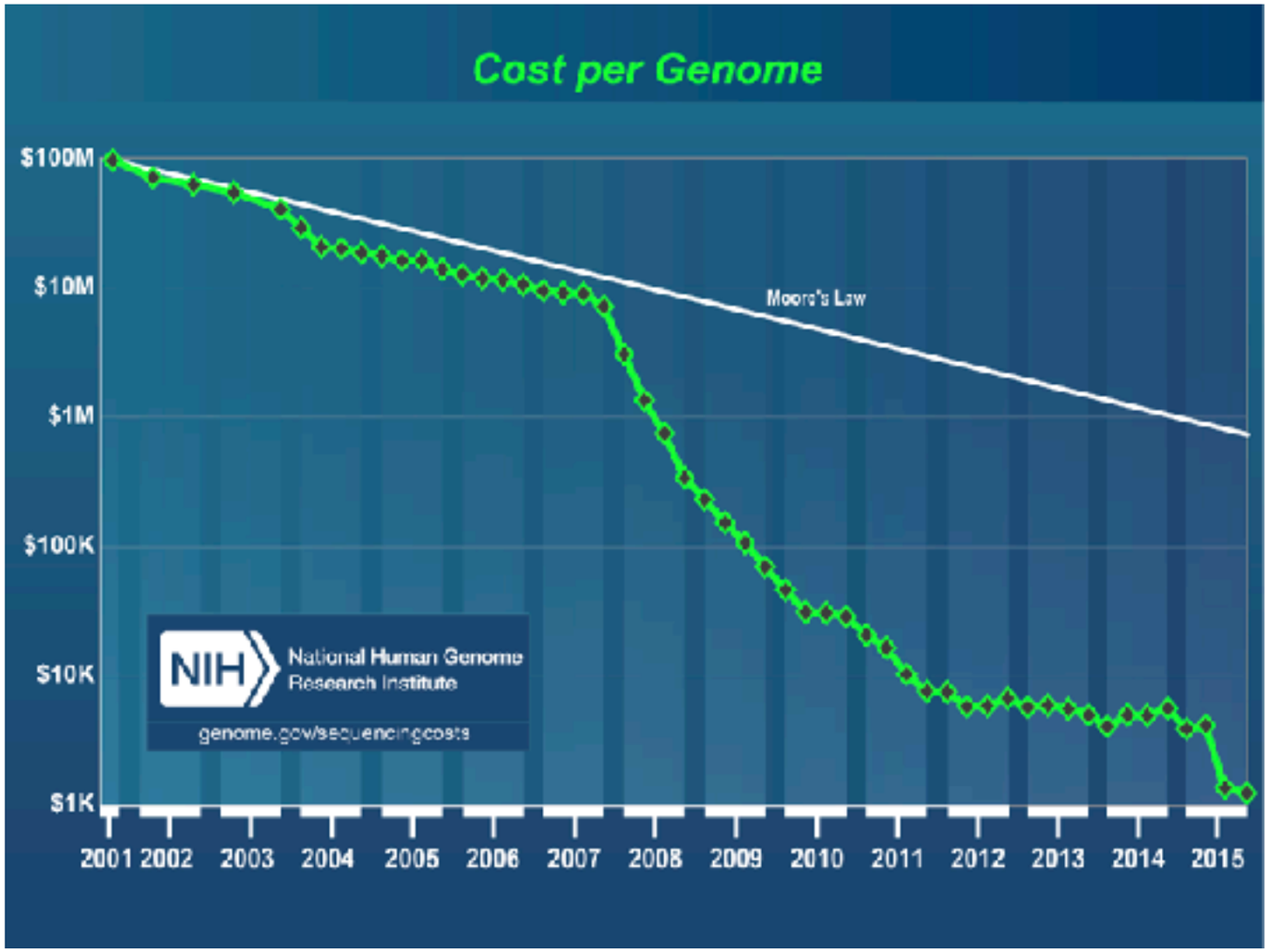
Transformation of many industries

Transformation of many industries

What is missing?

Technologies transforming biomedicine

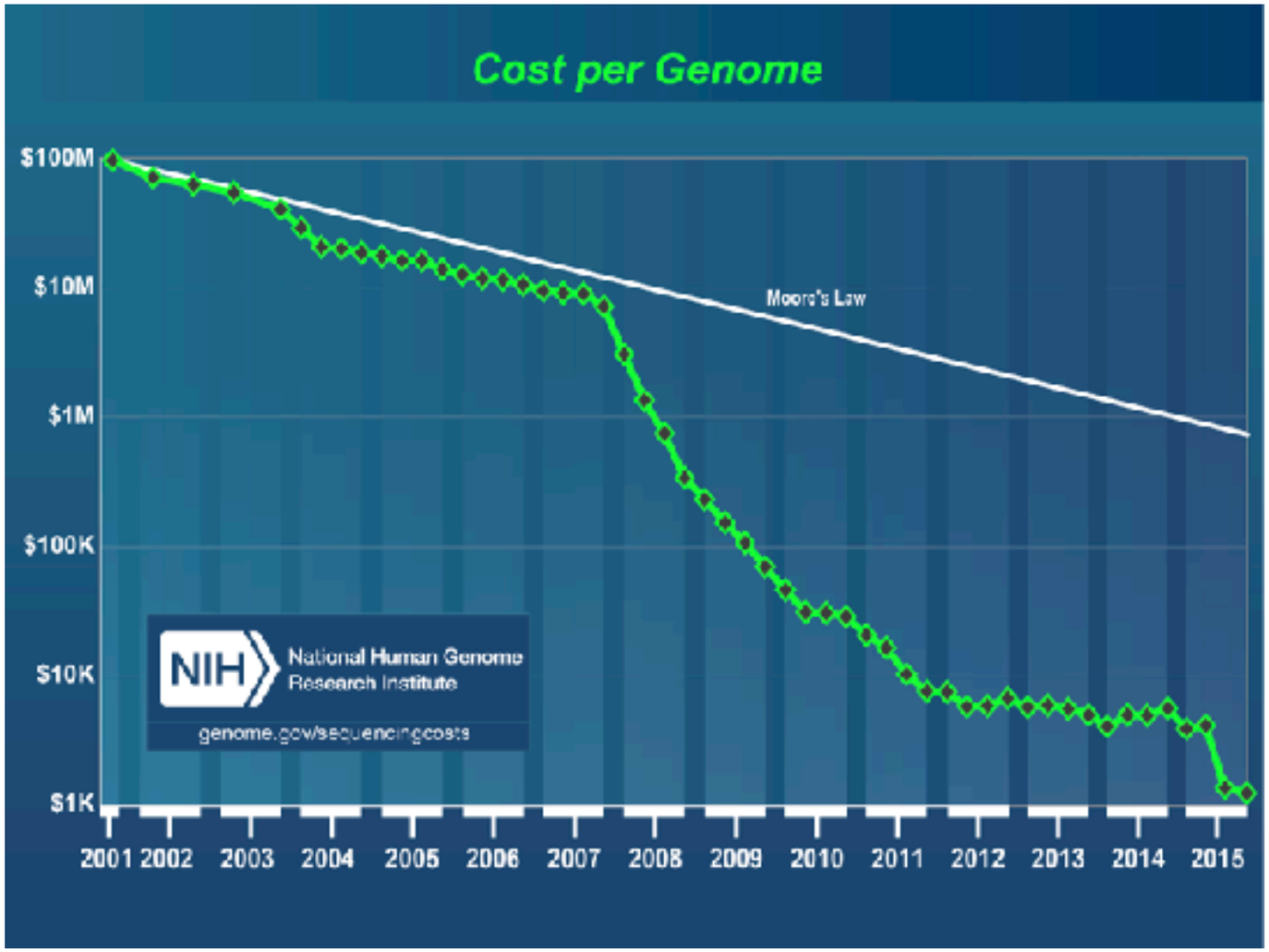
Cost of sequencing has plummeted over the past 15 years



Technologies transforming biomedicine

Cost of sequencing has plummeted over the past 15 years

~$1500 cost point projected for 17/18



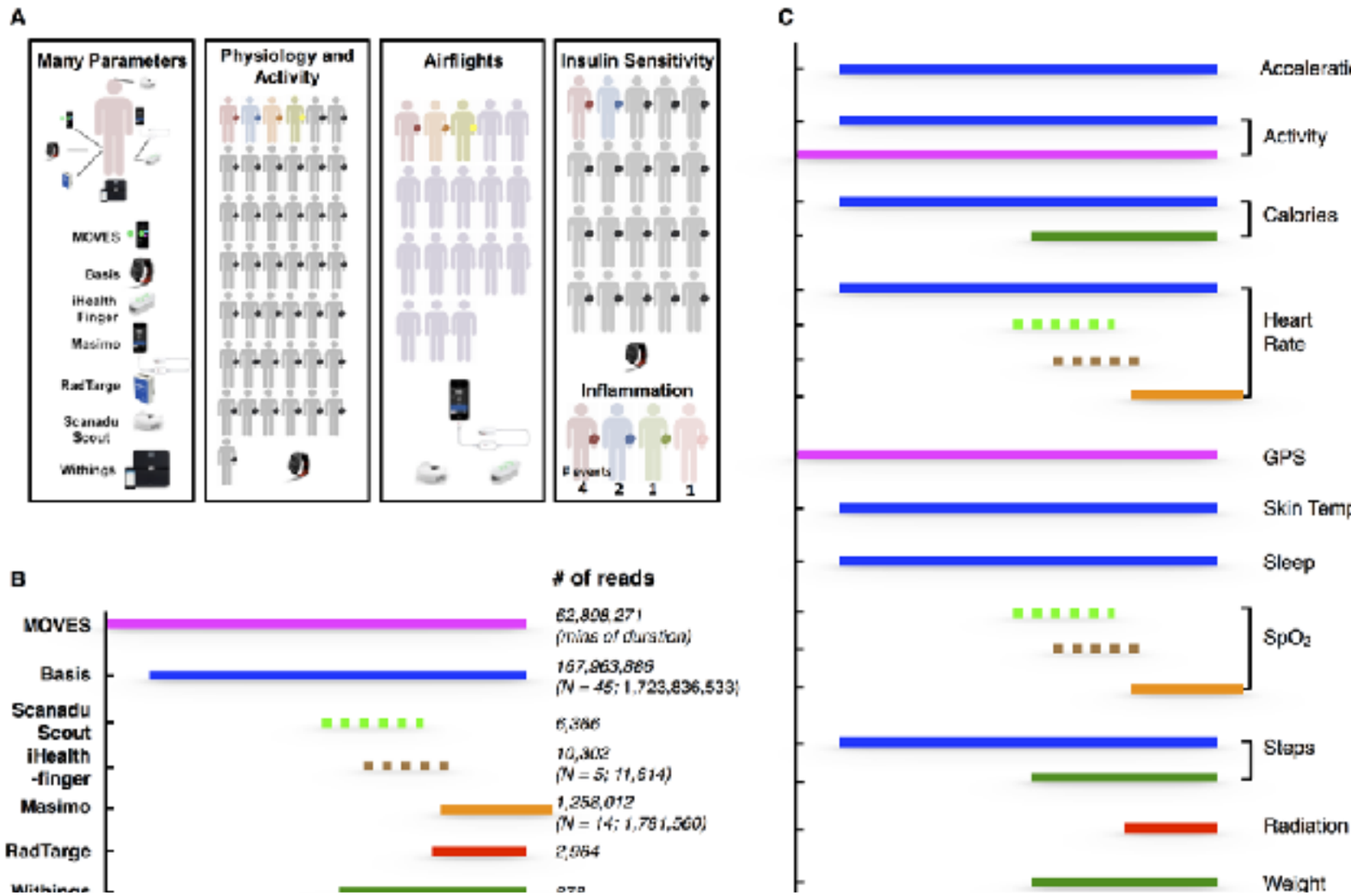
Technologies transforming biomedicine

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

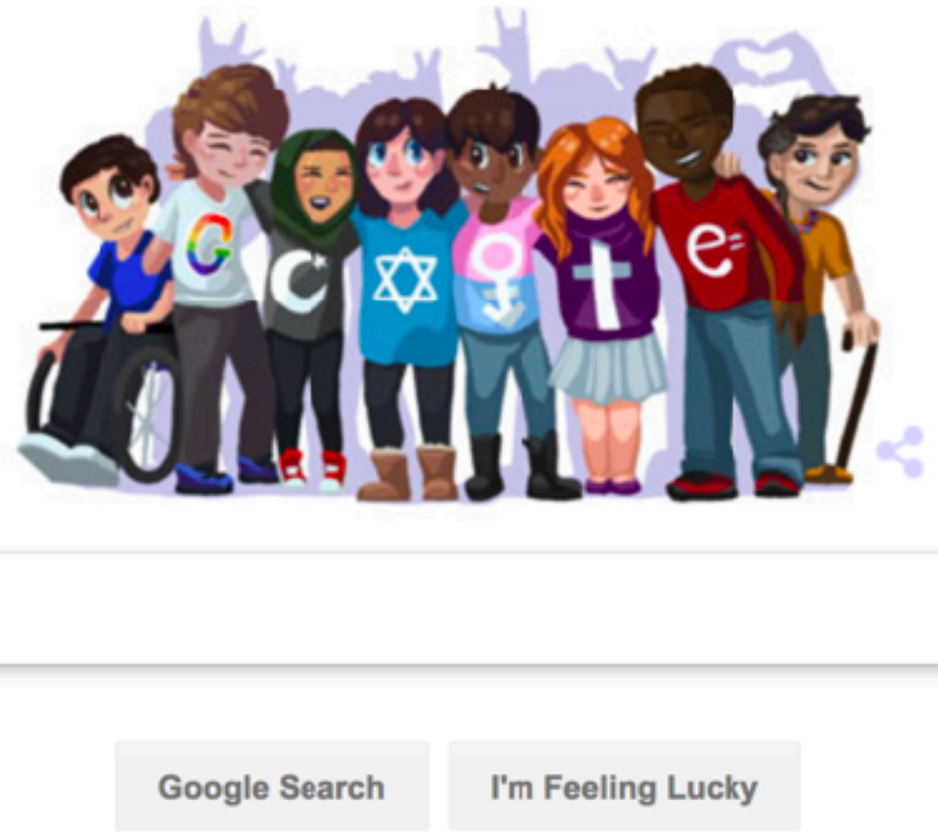
Technologies transforming biomedicine

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]

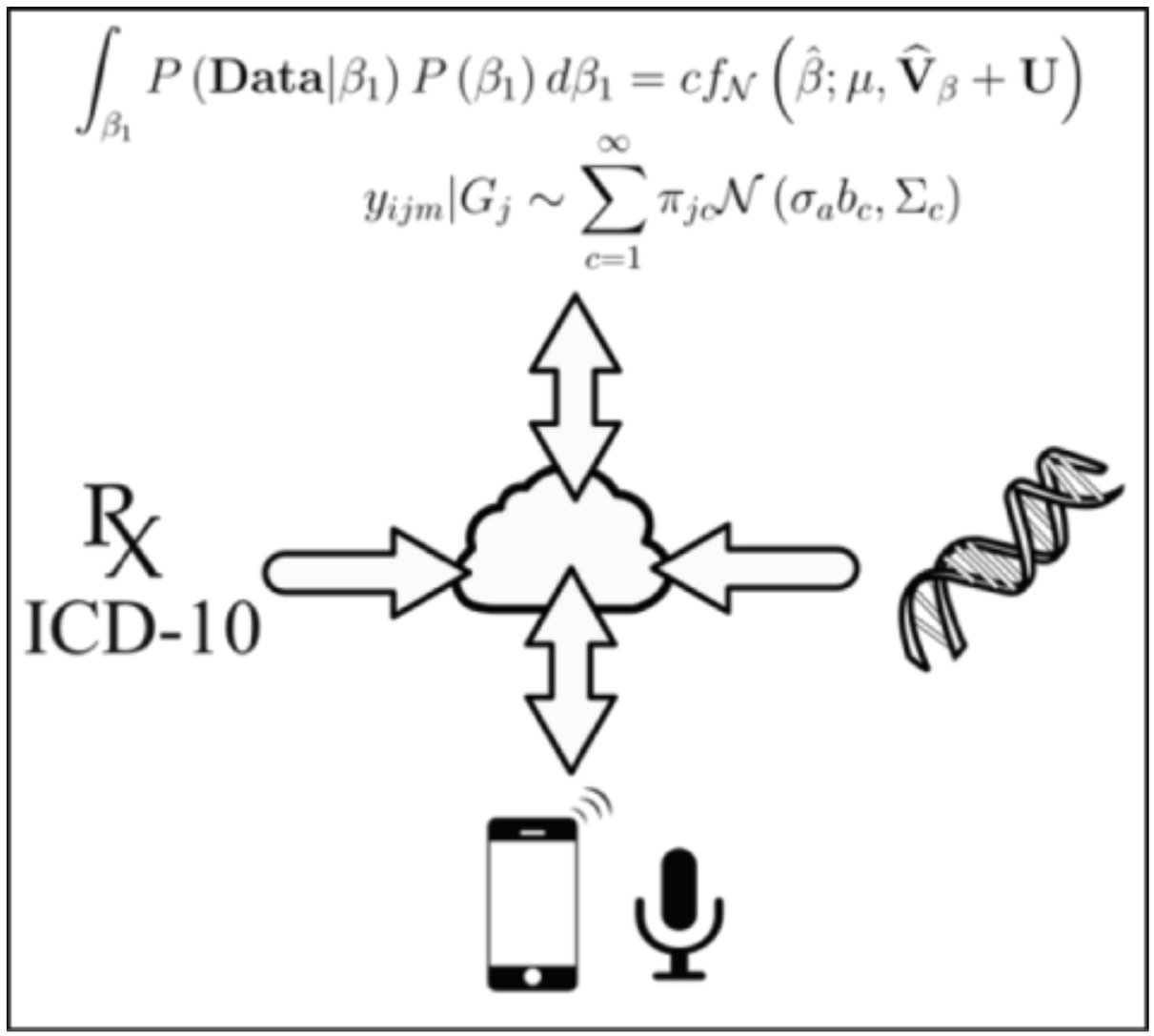
[1]Google Inc. [2]Centers for Disease Control and Prevention

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?



Challenge for this generation

Precision medicine and biobank initiatives

Precision medicine and biobank initiatives

UK Biobank

China Kadoorie Biobank

Precision Medicine Initiative

Introduction to the UK Biobank project

Major source of data for this course

About the UK Biobank

National and international health resource

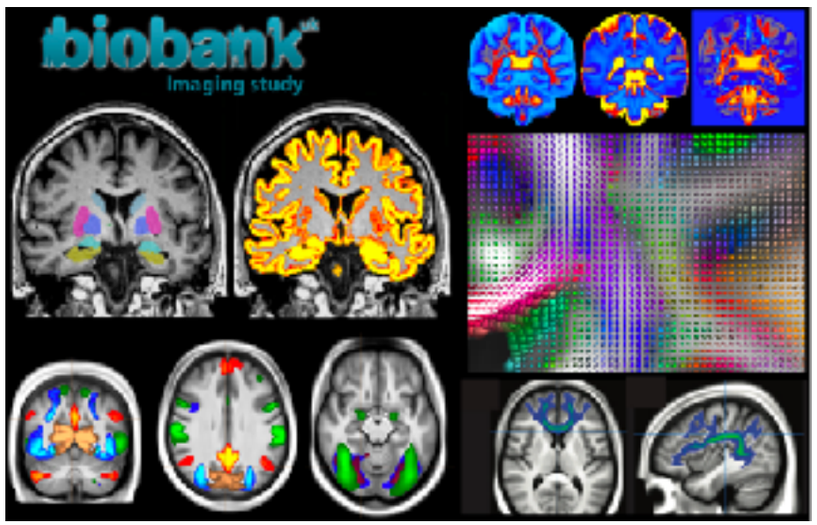
About the UK Biobank

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

~5,000 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

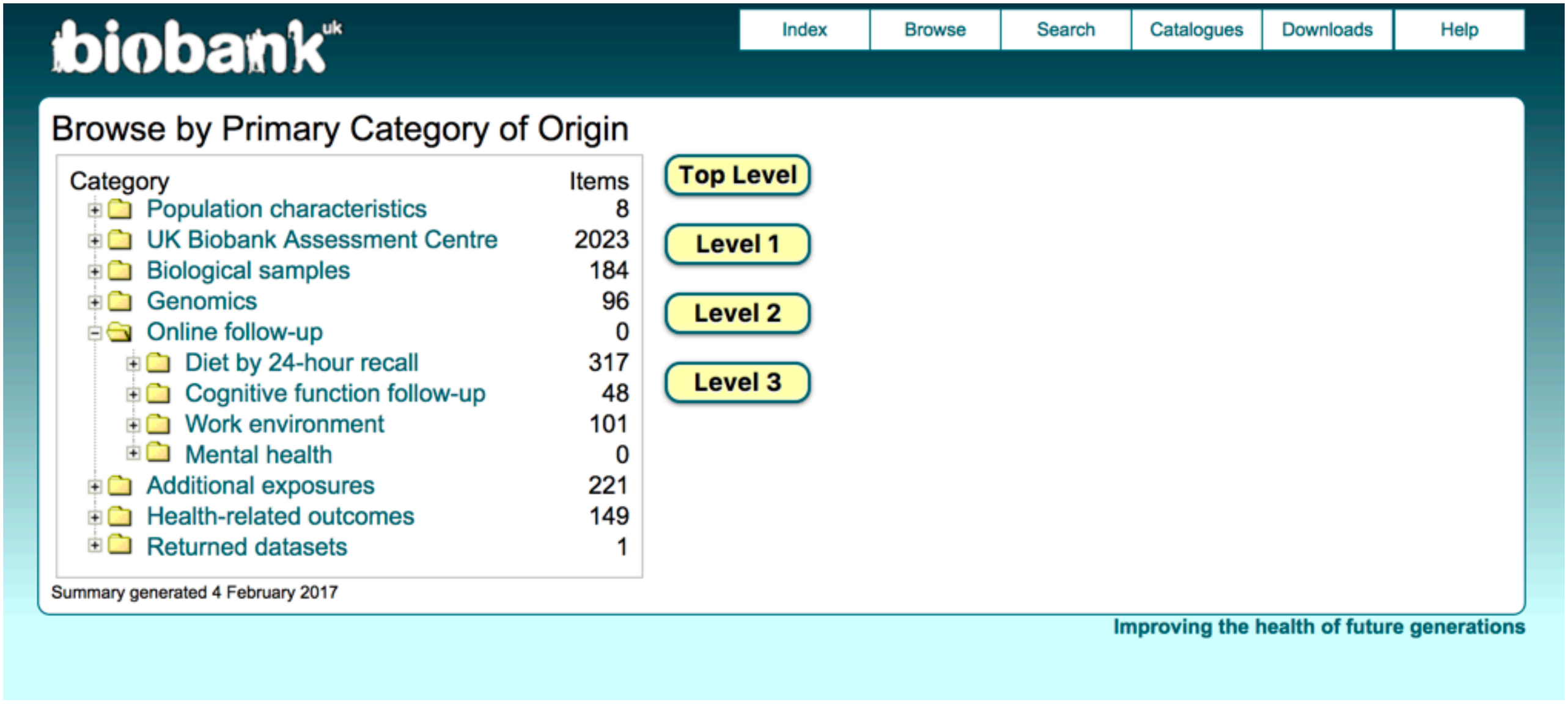
Level 2

Level 3

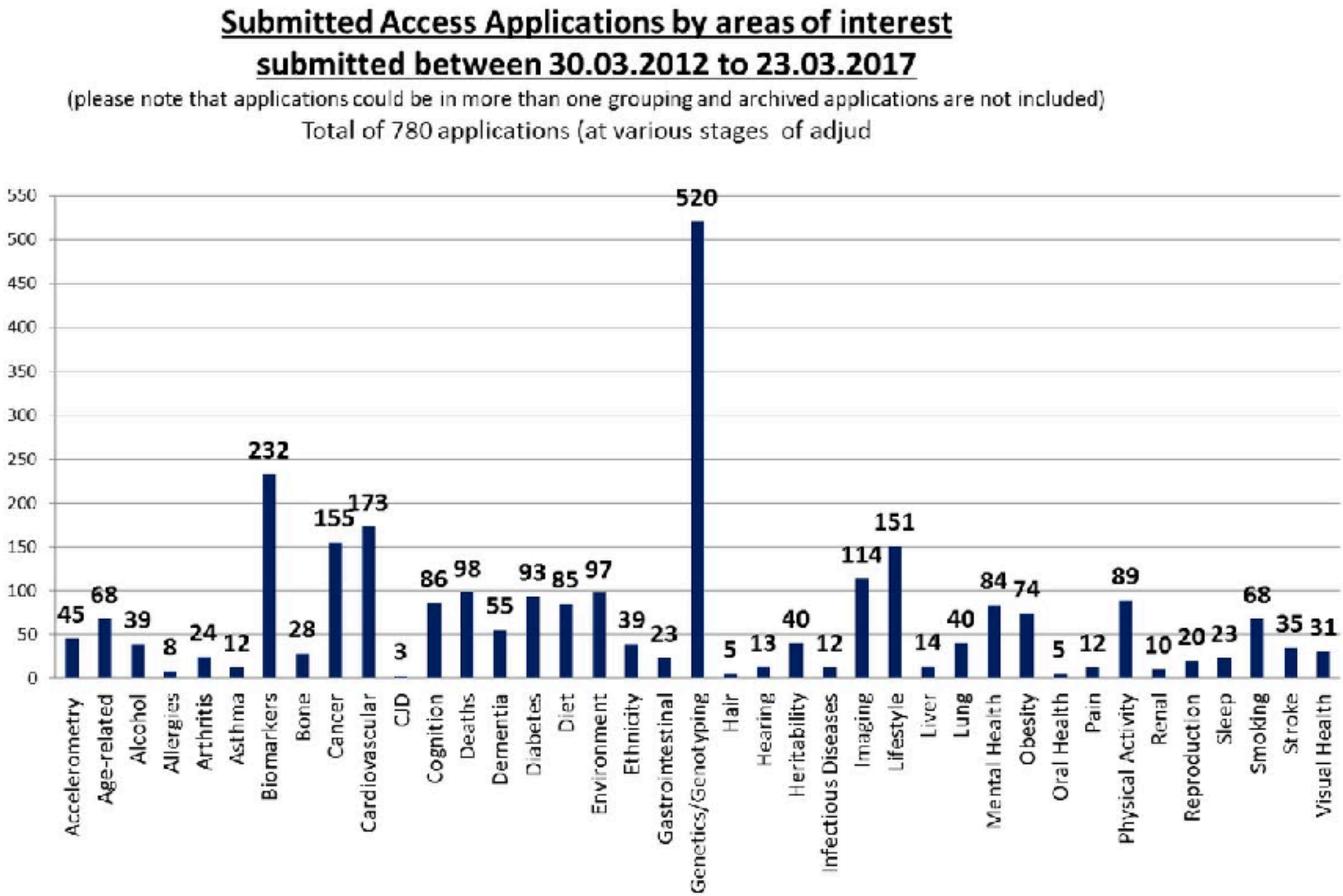
Summary gererated 4 February 2017

Improving the health of future generations

UK Biobank data showcase webpage



Data access summary



Course projects

Topics will be proposed during next week's lecture

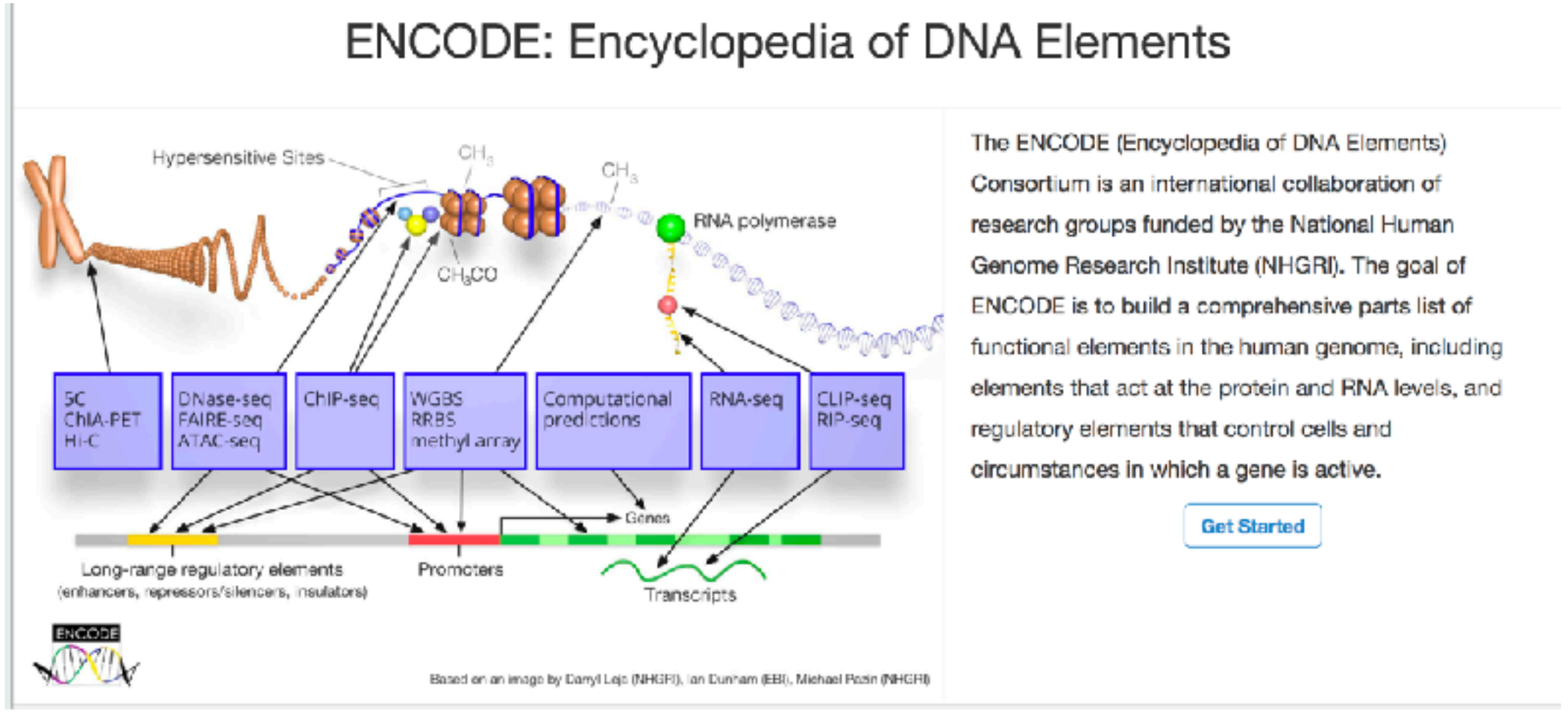
Data are organized in a Github repository

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

Data available in the course

1. UK Biobank

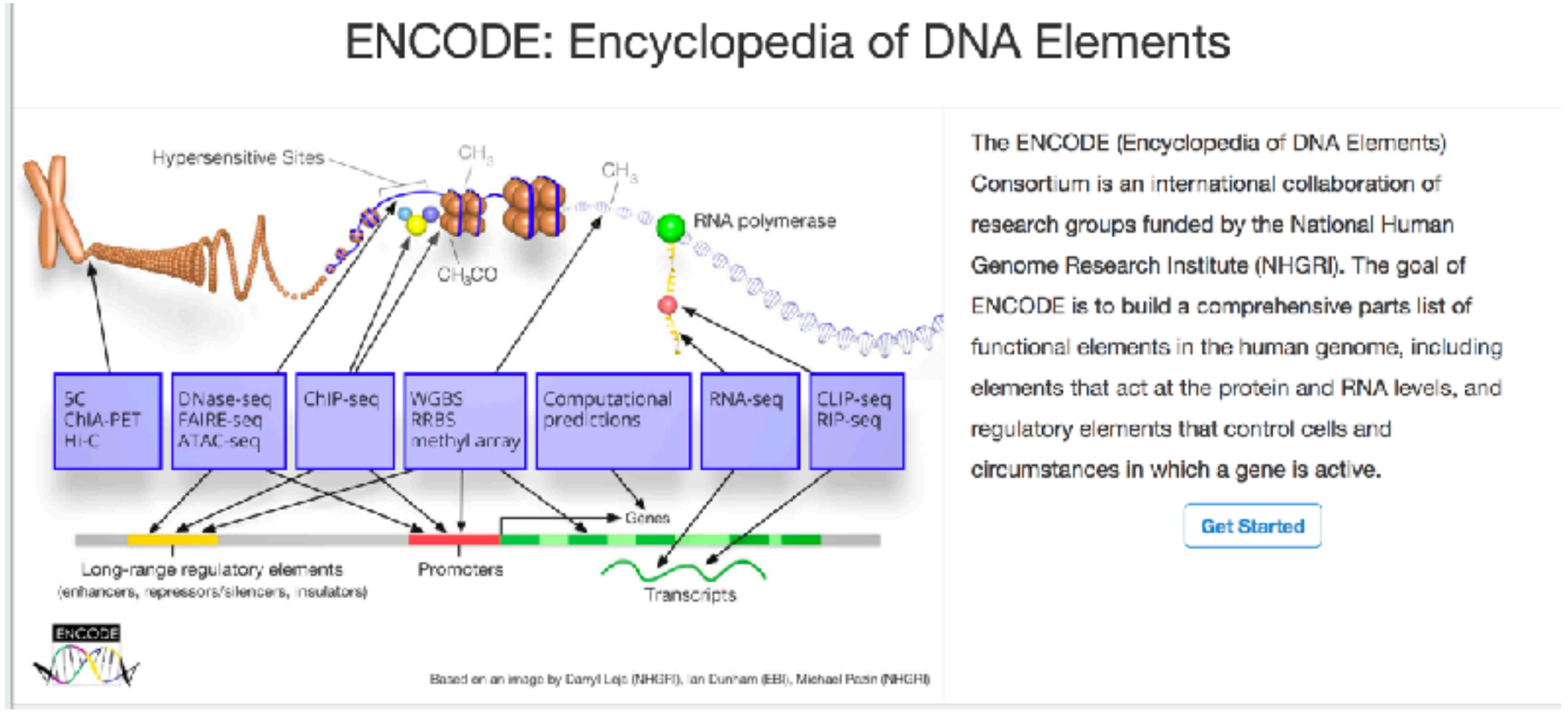
2. ENCODE



Data available in the course

1. UK Biobank

2. ENCODE

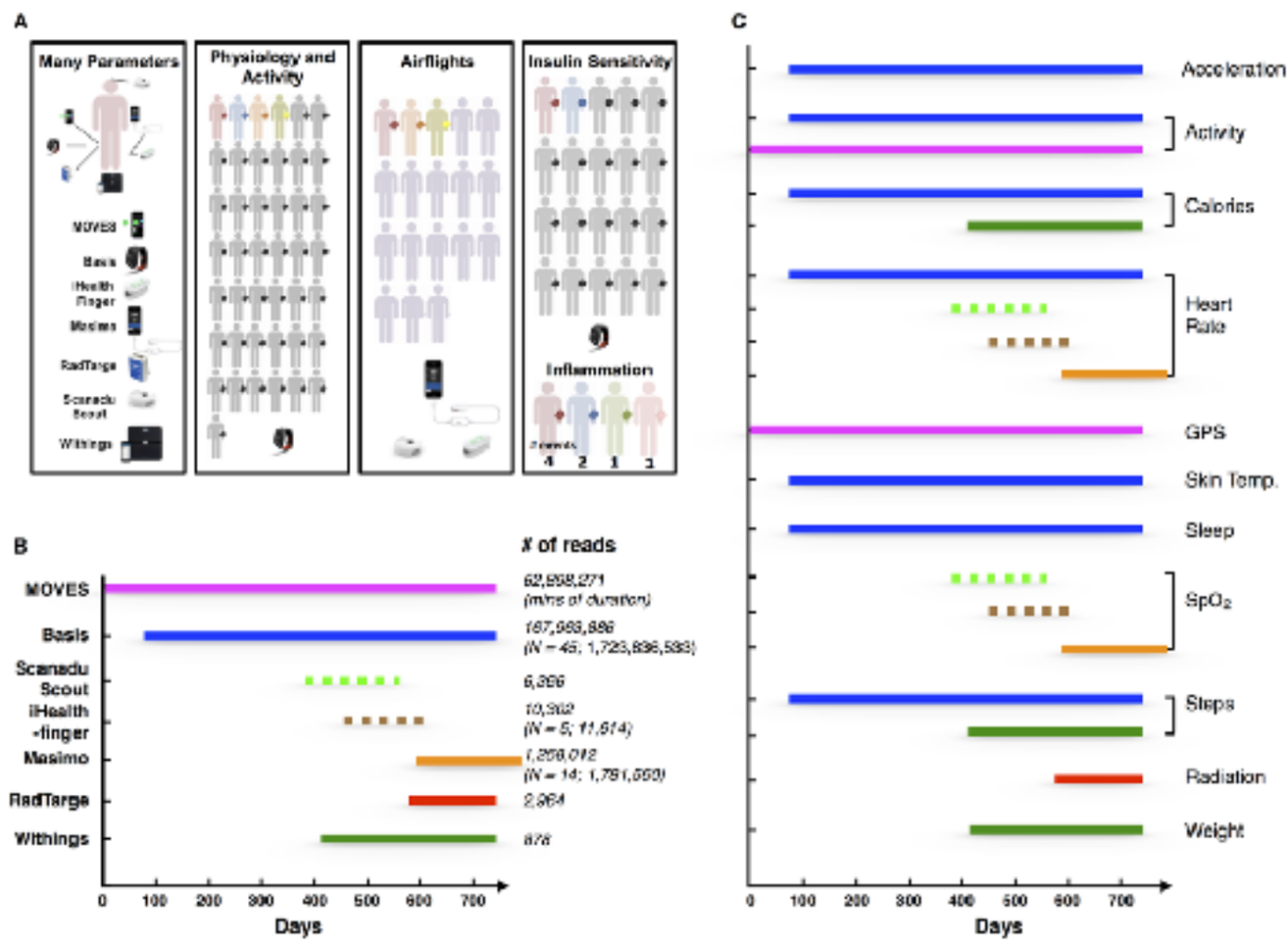


Data available in the course

1. UK Biobank

2. ENCODE

3. Wearable Biosensor



STAN - probabilistic programming language

Stan

http://mc-stan.org/

STAN - probabilistic programming language

A probabilistic programming language for statistical inference written in C++.

STAN - probabilistic programming language

A probabilistic programming language for statistical inference written in C++.

The Stan language is used to specify a (Bayesian) statistical model with an imperative program calculating the log probability density function.

STAN - probabilistic programming language

A probabilistic programming language for statistical inference written in C++.

The Stan language is used to specify a (Bayesian) statistical model with an imperative program calculating the log probability density function.

Stan can be accessed through several interfaces.

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

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 with examples

https://biods215.github.io/