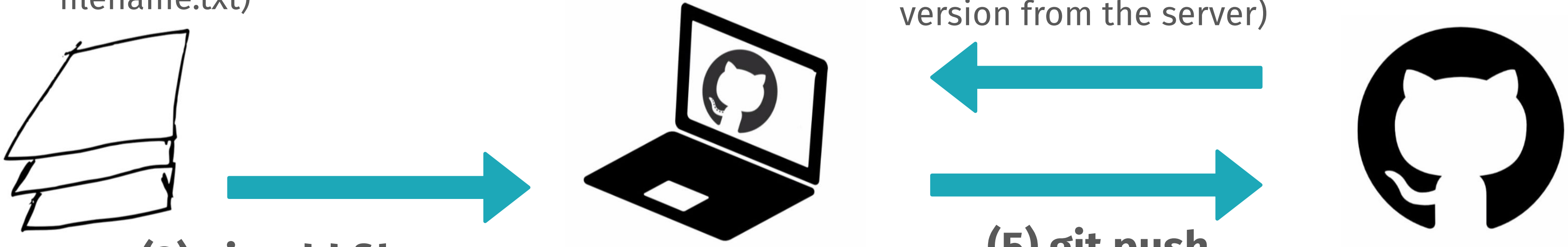
Announcements

- Lab 0 solutions have been released on Canvas/GitHub
- Lab 1 released today; due Friday February 6 at 6pm ET
- Office Hours:
  - Wednesdays 9-10am (starting next week)
  - Fridays 2:30-3:30pm

Review of Last Time

Review of Last Time: Typical Git/GitHub Pipeline



(2) make local changes

(e.g., create file called filename.txt)

(3) git add filename.txt

(changes are staged/waiting to be committed)

(4) git commit -m "[description of changes]"

(commit when you have made some changes and want to be able to save your current checkpoint as a snapshot)

LOCAL REPOSITORY

(1) git pull

(to retrieve the most recent version from the server)

REMOTE REPOSITORY

(5) git push

(make changes available to everyone with access to the repo)

Warning: remember to "git pull" before "git push" to mitigate potential merge conflicts

Student GitHub repositories that I have access to:

- ajluy
- apham2-del
- BaimatNiiazaliev
- bbaron26
- cquirk2
- gprofy2
- isabelhenderson
- kangjdh
- kschilz
- Lynn58259
- Xhershey90
- Zongyu-Li

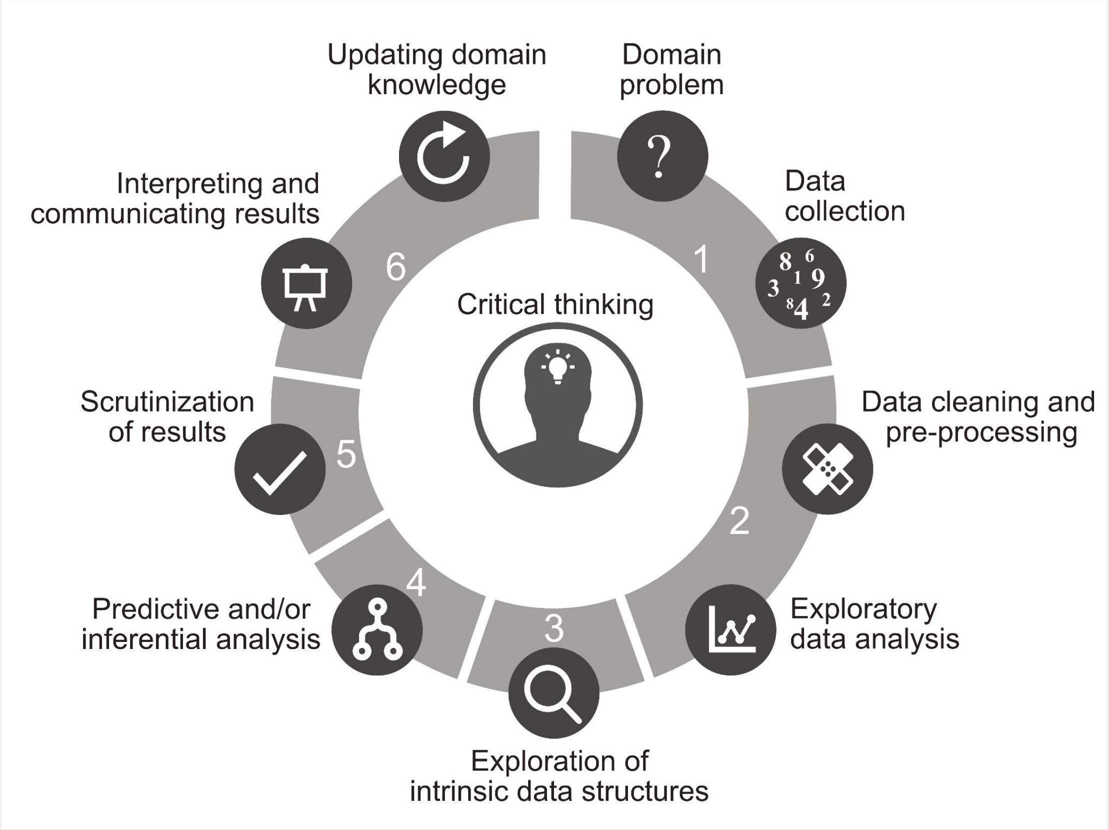
Today

Beginning your Data Science Life Cycle

Problem Formulation & Data Collection

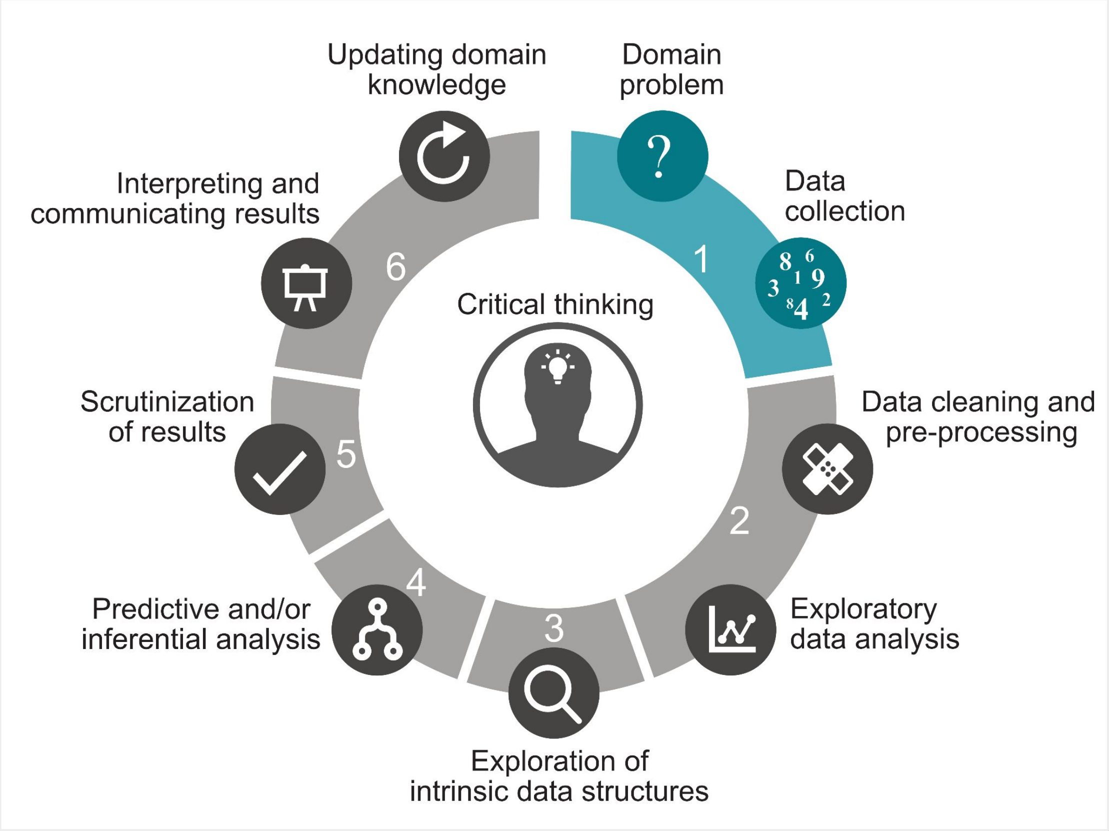
January 20, 2026

The Big Picture: Data Science Life Cycle


Updating domain knowledge
Domain problem
Data collection
Interpreting and communicating results
1
6
Critical thinking
8 6 1 9 3 8 4 2
?
Scrutinization of results
5
Data cleaning and pre-processing
2
Predictive and/or inferential analysis
4
3
Exploratory data analysis
Exploration of intrinsic data structures

Credits: Veridical Data Science (VDS) Textbook

The Big Picture: Data Science Life Cycle


Updating domain knowledge
Domain problem
Data collection
Interpreting and communicating results
Critical thinking
Scrutinization of results
Data cleaning and pre-processing
Predictive and/or inferential analysis
Exploratory data analysis
Exploration of intrinsic data structures
1
2
3
4
5
6
8 6 1 9 3 8 4 2

Credits: Veridical Data Science (VDS) Textbook

Plan for today

What do **problem formulation** and **data collection** look like in reality?

1. **Case Study 1: Cardiovascular Genomics**
2. **Case Study 2: COVID-19 PPE Resource Allocation**
3. **Case Study 3: A Day in the Life of a Redwood Tree (Lab 1)**

Case Study 1: Cardiovascular Genomics

Case Study: Cardiovascular Genomics

Euan Ashley, MD, PhD (Stanford University)

Imagine that you are in your initial intake meeting with Dr. Ashley.
What follow-up questions would you like to ask?

Case Study: Cardiovascular Genomics

Euan Ashley, MD, PhD (Stanford University)

Data: UK Biobank

Single Nucleotide Variants

(p ~ 15 million)

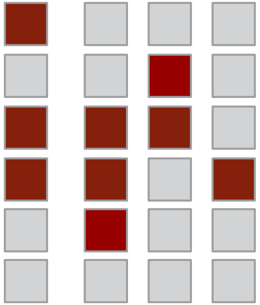
Patients

(n ~ 500,000)

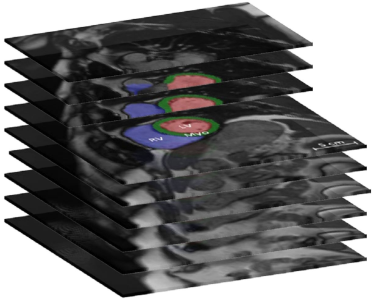
$$\begin{bmatrix} 0 & 2 & 1 & 2 & 0 & 2 & \dots & 2 & 2 \\ 2 & 0 & 0 & 2 & 1 & 0 & \dots & 2 & 2 \\ 0 & 1 & 0 & 2 & 0 & 0 & \dots & 2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & 2 & 0 & 2 & 1 & \dots & 0 & 0 \end{bmatrix}$$

ICD10 Diagnosis Codes

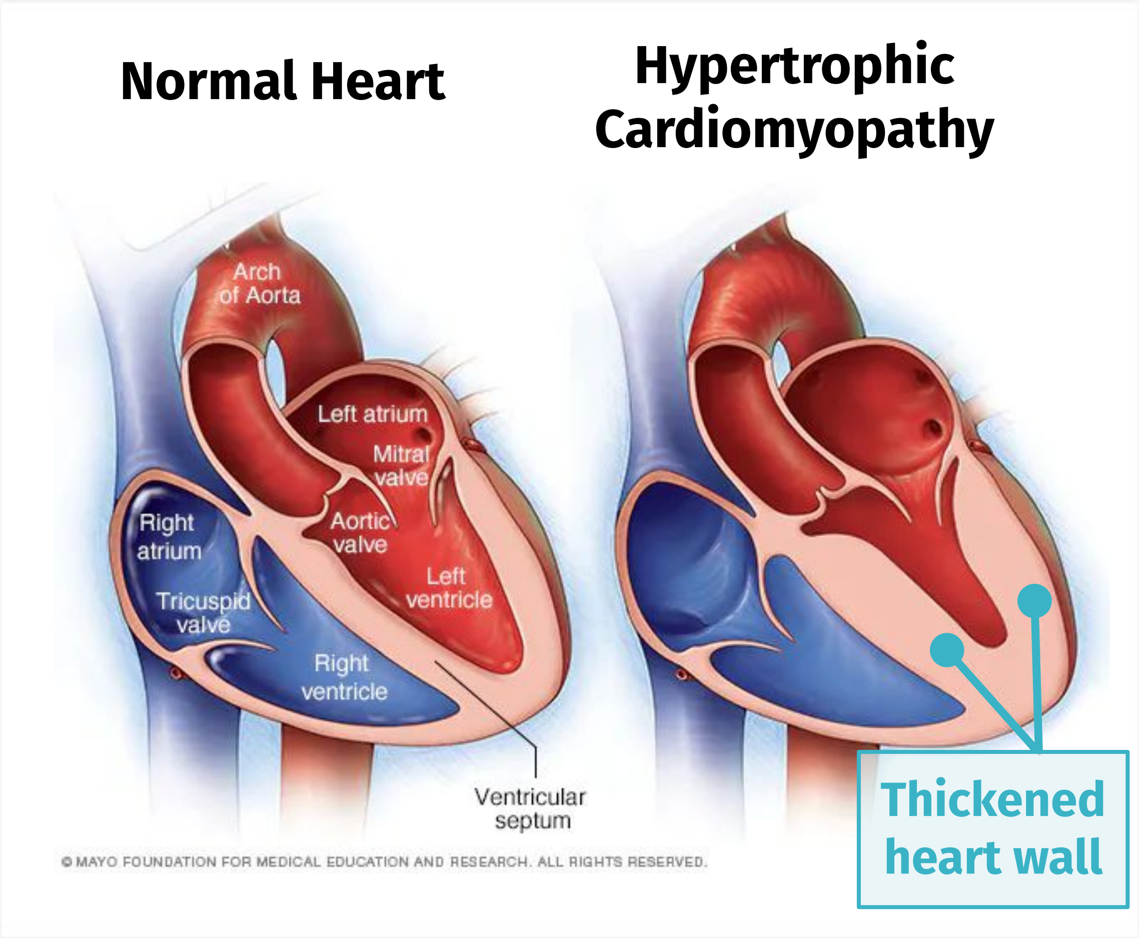
(many many diseases)



Cardiac MRIs



(n ~ 50,000)



Normal Heart

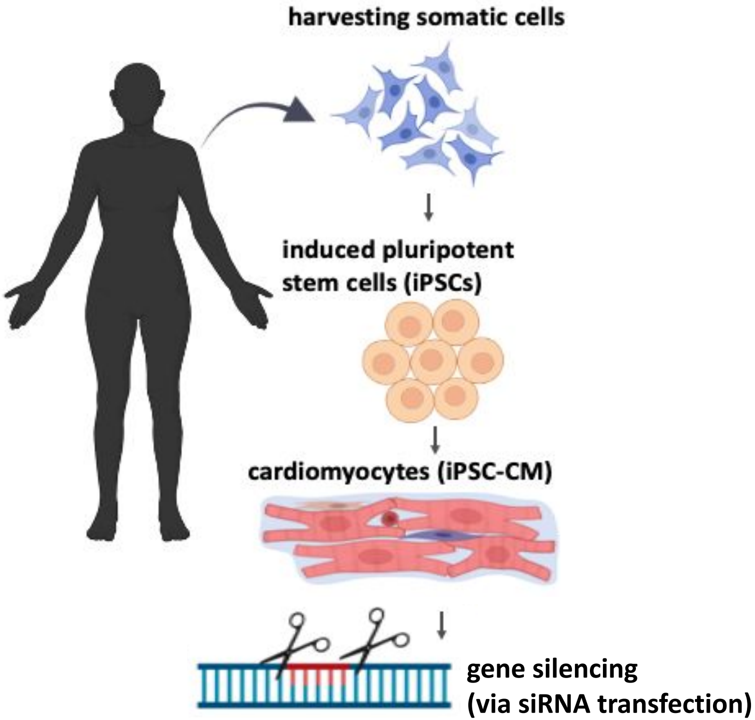
Hypertrophic Cardiomyopathy

© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Image Source: https://www.mayoclinic.org/diseases-conditions/left-ventricular-hypertrophy/symptoms-causes/syc-20374314

Overview: Experimental Workflow

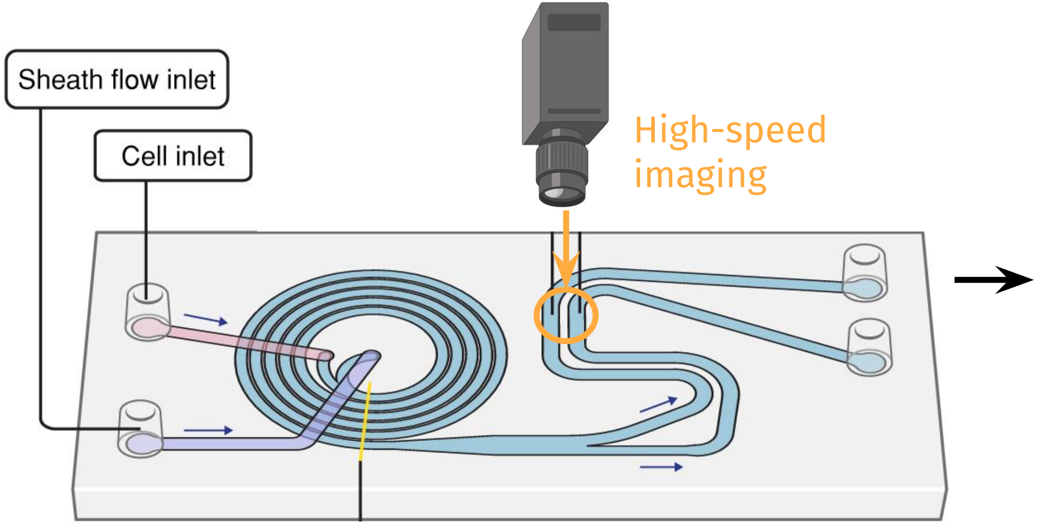
How do the size of heart cells change when we silence a gene or pair of genes?


harvesting somatic cells
induced pluripotent stem cells (iPSCs)
cardiomyocytes (iPSC-CM)
gene silencing (via siRNA transfection)

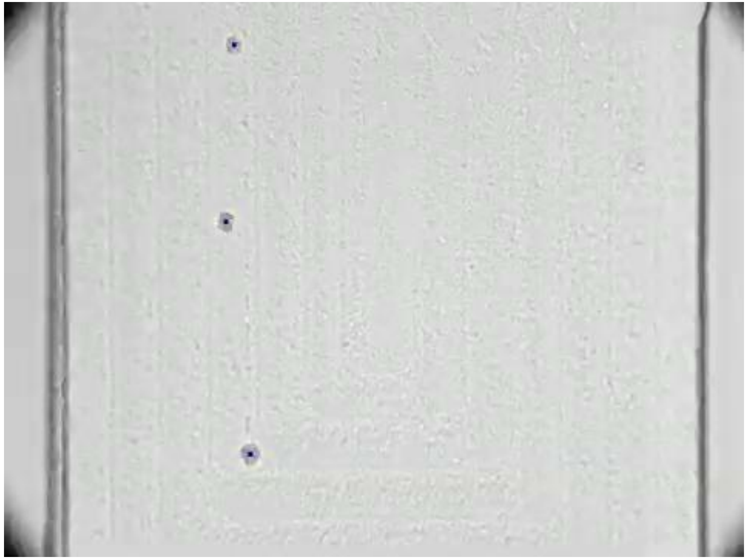
1. Silence Gene A
   → evaluate cell size
2. Silence Gene B
   → evaluate cell size
3. Silence Gene A and B
   → evaluate cell size

Finally, compare cell sizes and assess whether there is an interaction

High-throughput microfluidics + image processing



High-speed imaging



Problem Formulation: A Checklist

Problem formulation is not just about formulating the **statistical** problem, but also formulating the **substantive** problem

- What is the **big-picture substantive question/problem**?
- **Why** is this scientific problem interesting or relevant to researchers in the field and/or general members of society?
- What is already known about this substantive problem? Any relevant **background information**?
- What are the **existing limitations/challenges** that make solving this problem difficult?
- What is the specific aim or **contribution** of this present work? What is your **end goal**?

Our aim

end-to-end pipeline

Gene / interaction recommendation system

+

Wet-lab experimental validation

A Bird’s Eye View of What Really Happened

In-person visit to Ashley and Priest Labs at Stanford

↳ Many discussions about which **heart disease** phenotype to study

Proceeded to study **Hypertrophic Cardiomyopathy (HCM)** due to

- High prevalence (~1 in 500)
- Team’s clinical expertise
- Experimental capabilities for measuring cell size

Hit a roadblock with **HCM:**

- **~50%** balanced classification accuracy
- (Typically) driven by rare variants
- Under-diagnosis and noisy labels

Left Ventricular Mass (LVM)

Case Study 2: COVID-19 PPE Resource Allocation

Case Study: COVID-19 PPE Resource Allocation

Don Landwirth (Response4Life)

❗ **Setting:** beginning of March 2020

Imagine that you are in your initial intake meeting with Response4Life. What follow-up questions would you like to ask?

Want to predict...

~~hospital PPE/supply need~~

↕

~~number of COVID-19 hospitalizations~~

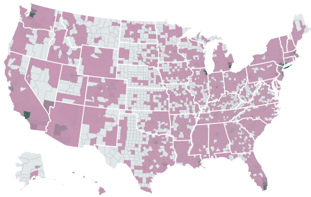
↕

number of COVID-19 deaths at the county-level

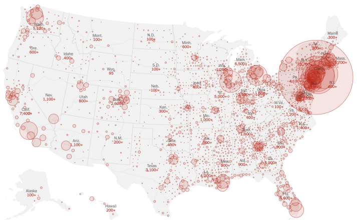
Our Data Repository: scraped from 20+ sources

COVID-19 Cases/Deaths

USAFACTS



The New York Times



County-level Data

(Risk Factors, Demographics, SES, Social Mobility)

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Division for Heart Disease and Stroke Prevention

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

esri COVID-19 GIS Hub

IHME GHDx

County Health Rankings & Roadmaps
Building a Culture of Health, County by County

USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

CMS.gov
Centers for Medicare & Medicaid Services

United States® Census Bureau

SAFEGRAPH

kinsa®

STREETLIGHT

cuebiq

Introducing the Unacast
Social Distancing Scoreboard

KHN KAISER HEALTH NEWS

Maps Mobility Trends Reports

JOHNS HOPKINS UNIVERSITY

Google COVID-19 Community Mobility Reports

Hospital-level Data

(e.g., #ICU beds, staff)

HRSA
Health Resources & Services Administration

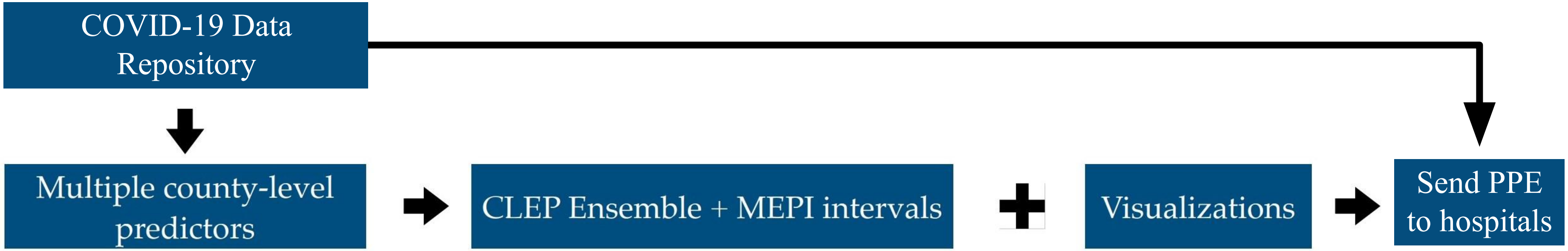
ArcGIS Hub

HOMELAND INFRASTRUCTURE FOUNDATION-LEVEL DATA SUBCOMMITTEE

Samuel Scarpino

NORTHEASTERN UNIVERSITY 18 98 LVX VERITAS VIRTVS

Overview of Modeling Pipeline



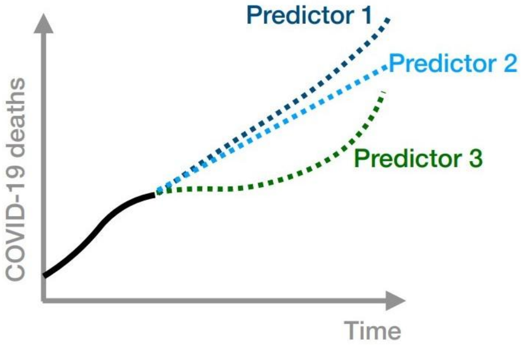
COVID-19 Data Repository

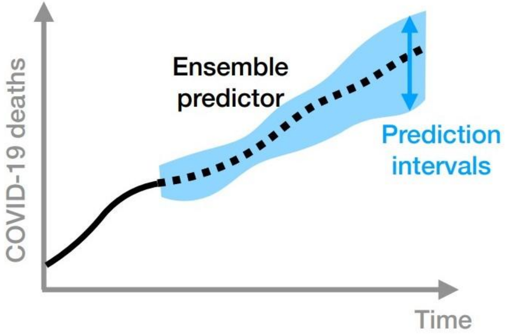
Multiple county-level predictors

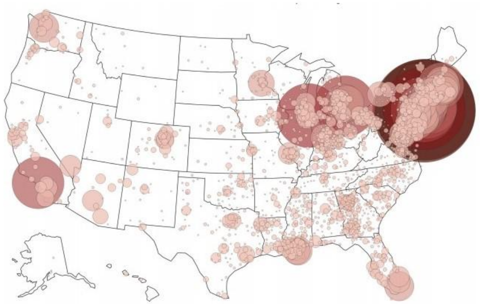
CLEP Ensemble + MEPI intervals

Visualizations

Send PPE to hospitals







Q: What types of data would you like to have to solve this problem?

Altieri, N., et al. (2021). Curating a COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States. Harvard Data Science Review. https://doi.org/10.1162/99608f92.1d4e0dae

Data Collection: A Checklist

BE TRANSPARENT AND DOCUMENT!

- What data is available?
  - What are the "samples" and the "variables" in the data?
- How was the data collected or generated? (Who collected the data?)
  - Describe how the variables were measured
- Why was the data collected?
  - Describe why these variables are important
- How does your data connect to the scientific question?
  - Any special properties of the data/data collection that make it uniquely suited to answer your question?
- Are there any limitations or words of caution when using the data to answer the domain problem of interest? **Garbage in, garbage out!**

We’ve got the data. What’s next?

Let the **data preprocessing/cleaning** journey begin… (next class)

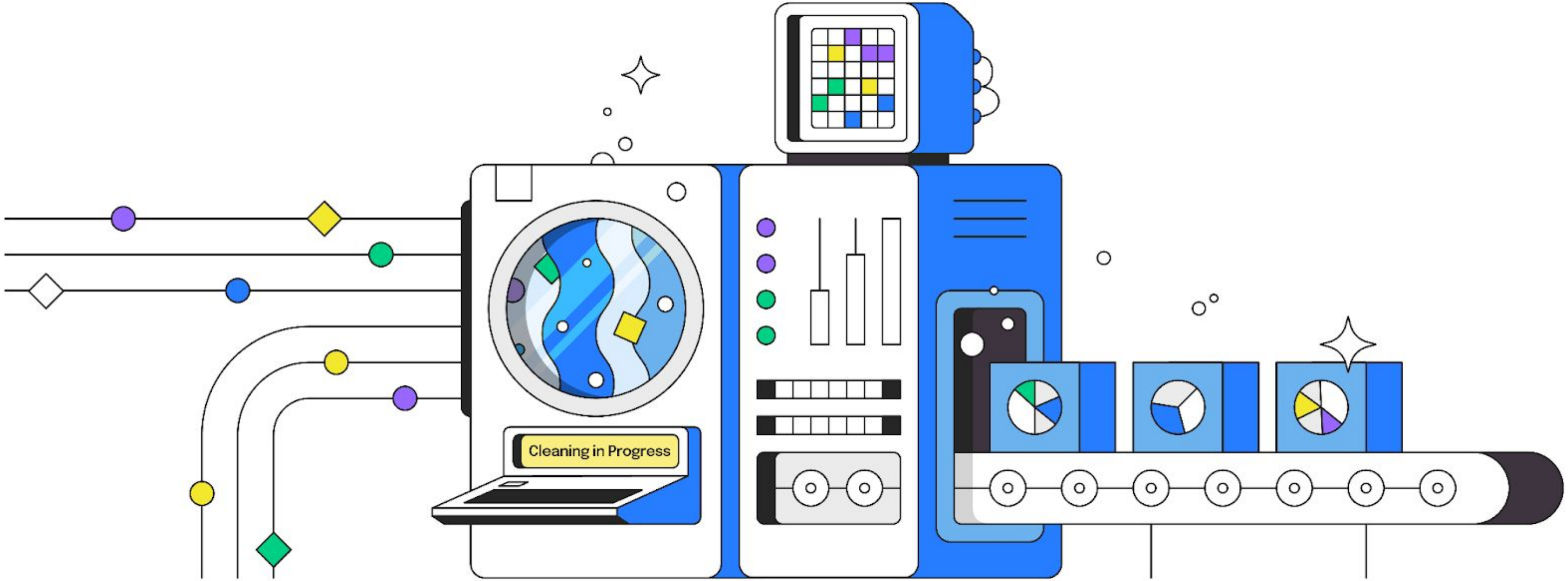
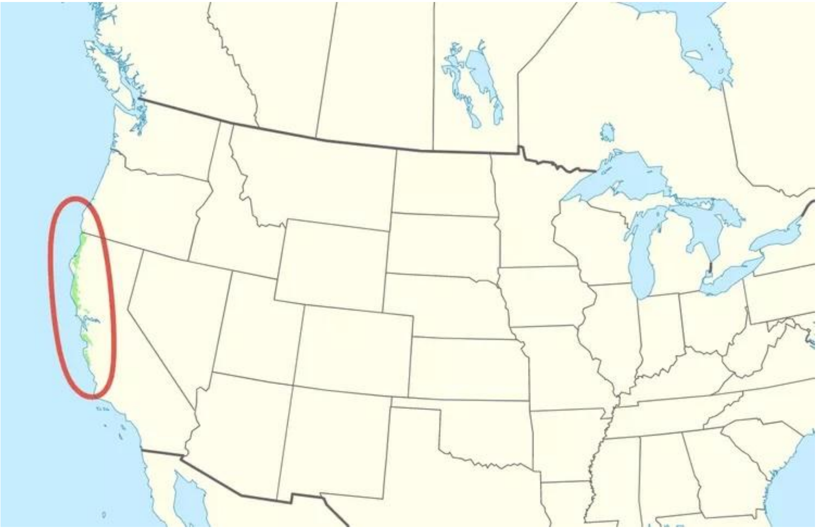


Image Credits: obviously.ai

Case Study 3: A day in the life of a redwood tree (Lab 1)

A Macroscope in the Redwoods [Tolle et al. (2005)]

- Coastal redwood trees
  - Tallest trees in the world (>350ft or 115m)
  - Incredibly old species (pre-dating humans, spiders, and flowers, first appearing over 240 million years ago during the time of the dinosaurs)



A Macroscope in the Redwoods [Tolle et al. (2005)]

- 44-day study in Sonoma, California
  (April 27, 2004 5:10pm - June 10, 2004 2pm)

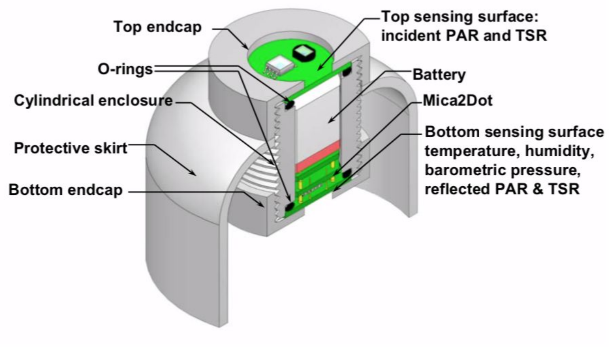

Top endcap
O-rings
Cylindrical enclosure
Protective skirt
Bottom endcap
Top sensing surface: incident PAR and TSR
Battery
Mica2Dot
Bottom sensing surface temperature, humidity, barometric pressure, reflected PAR & TSR

Figure 2: Sensor node and packaging

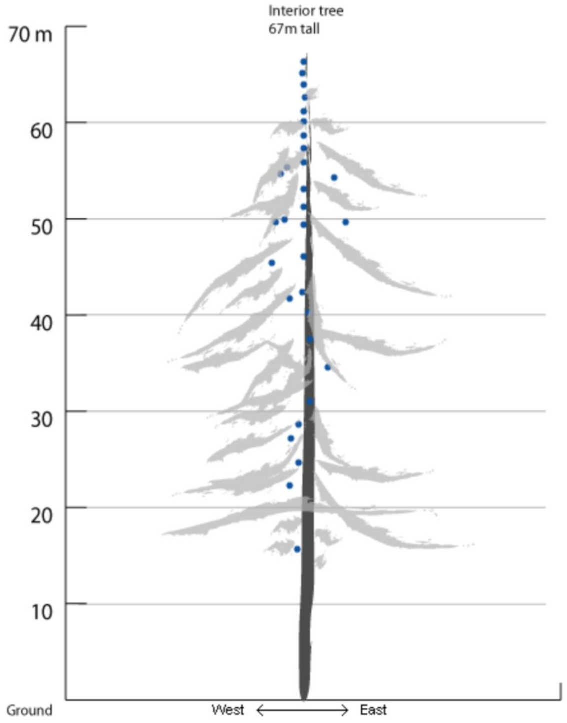
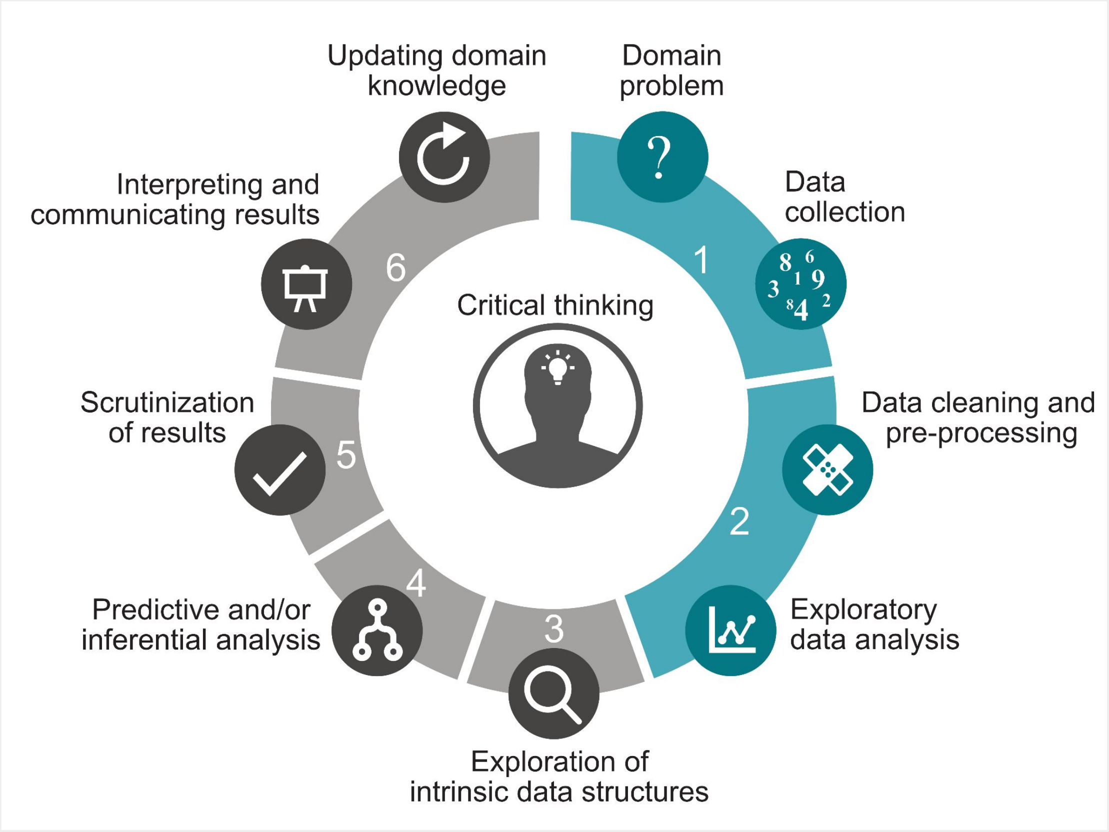


Figure 1: The placement of nodes within the tree

Lab 1: Redwood Trees



Updating domain knowledge
Domain problem
Data collection
Interpreting and communicating results
Critical thinking
Data cleaning and pre-processing
Scrutinization of results
Exploratory data analysis
Predictive and/or inferential analysis
Exploration of intrinsic data structures
1
2
3
4
5
6
?
8 6 1 9 3 8 4 2

Summary of Today

- **Problem formulation** and **data collection** are crucial stages when beginning your data science life cycle.
  - **Problem formulation:** includes formulation of both the *statistical* and the *substantive* problem
    - Learn about the *data* and the *science*
  - **Data collection:** garbage in, garbage out
    - Data is the real currency in modeling
- Ask questions, use common sense, and document everything

Next time: reproducible project workflows (lab 1) + data cleaning