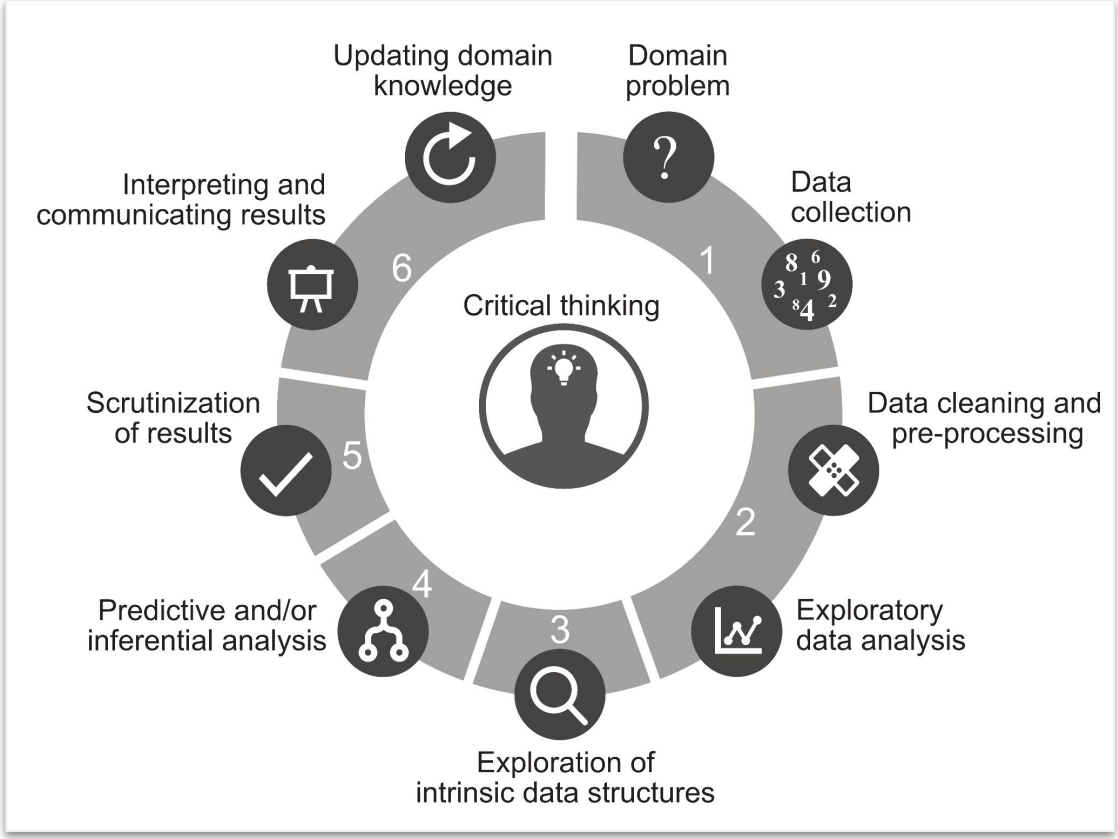
Beginning your Data Science Life Cycle

Problem Formulation & Data Collection

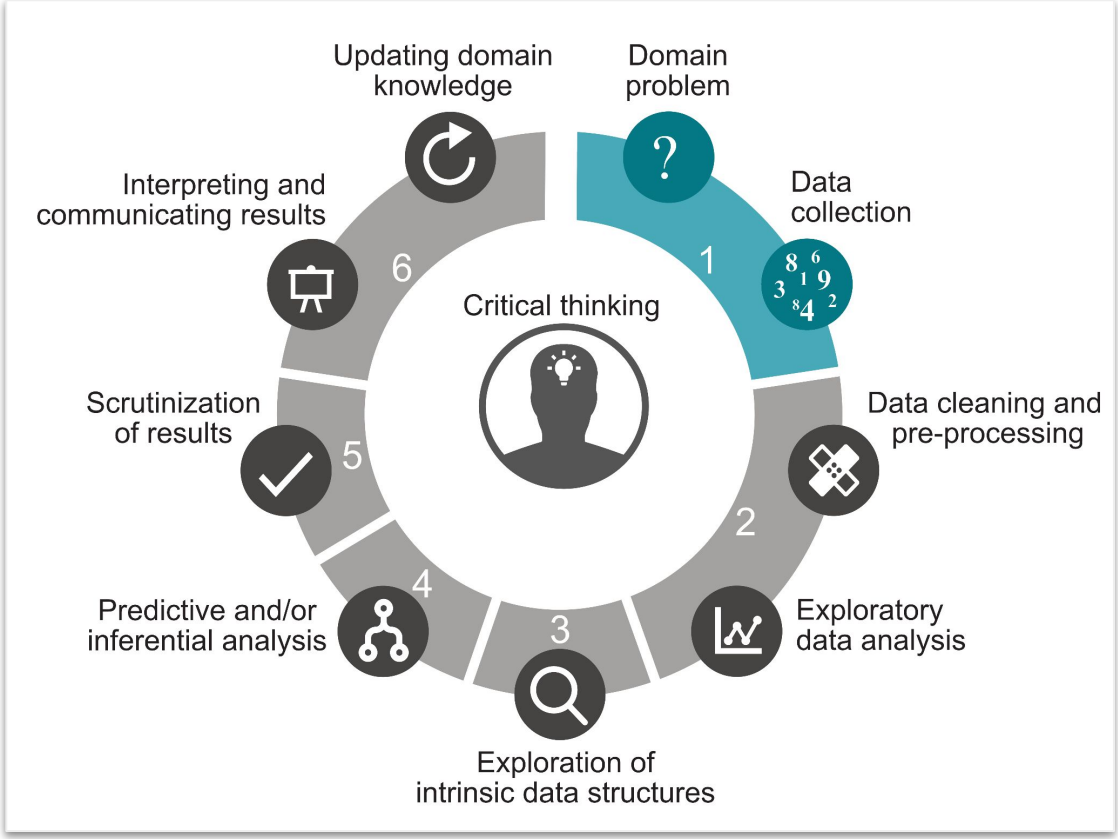
January 15, 2025

The Big Picture: Data Science Life Cycle



Credits: Veridical Data Science (VDS) Textbook

The Big Picture: Data Science Life Cycle



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**

Goal for today: create a minimum **checklist** for the problem formulation and data collection stages

Case Study 1: Cardiovascular Genomics

Case Study: Cardiovascular Genomics

Euan Ashley, MD, PhD (Stanford University)

Which gene interactions are important drivers of heart disease?

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)

Which gene interactions are important drivers of heart disease? We can run experiments to validate these genes in the wet-lab.

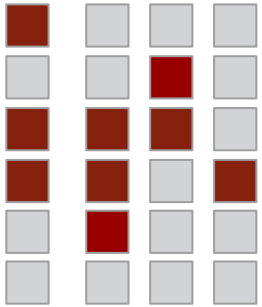
Data: UK Biobank

Patients (n ~ 500,000)

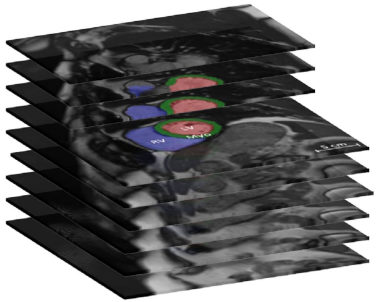
Single Nucleotide Variants
(p ~ 15 million)

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

ICD10 Diagnosis Codes
(many many diseases)



Cardiac MRIs



(n ~ 50,000)

Normal Heart

Hypertrophic Cardiomyopathy

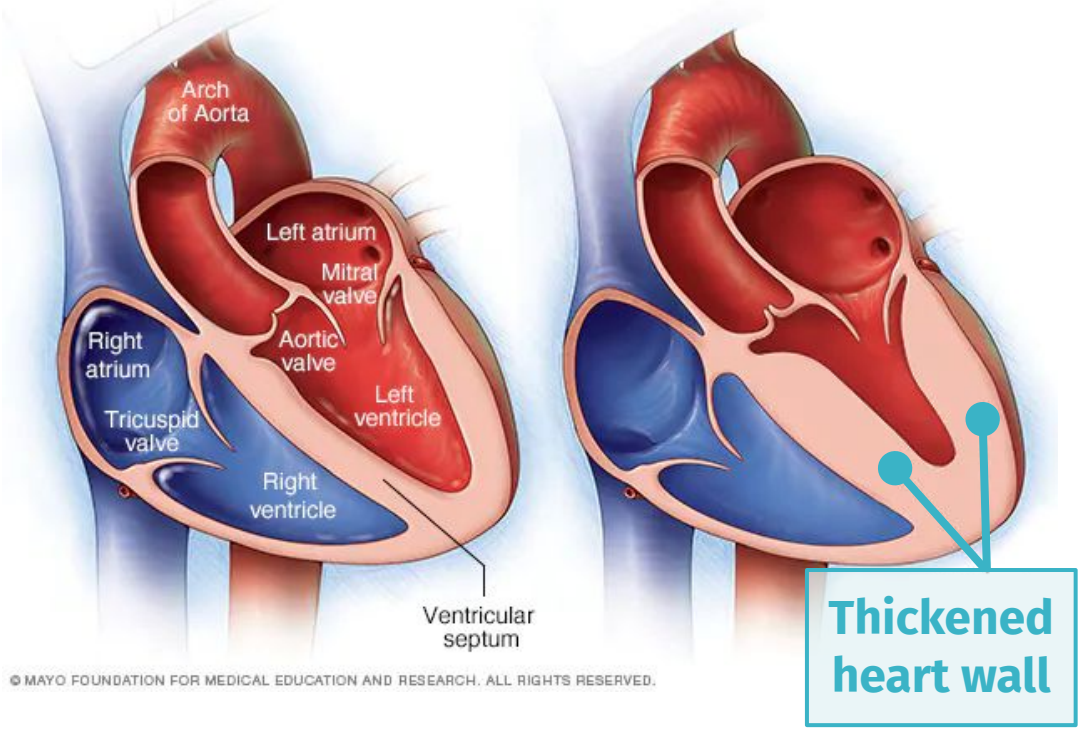


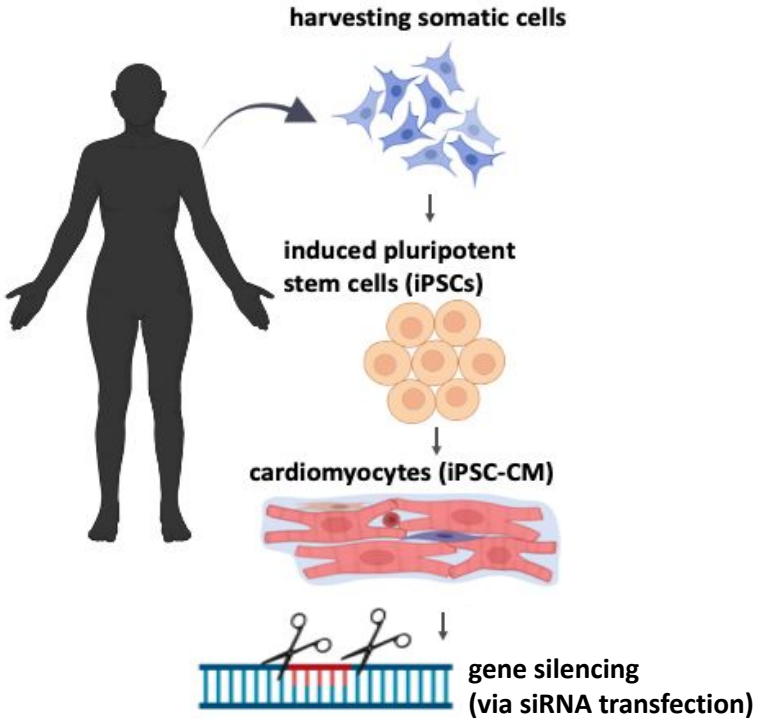
Image Source: https://www.mayoclinic.org/diseases-conditions/left-ventricular-hypertrophy/symptoms-causes/syc-20374314

Overview: Experimental Workflow


Qianru Wang


Nate Youlton

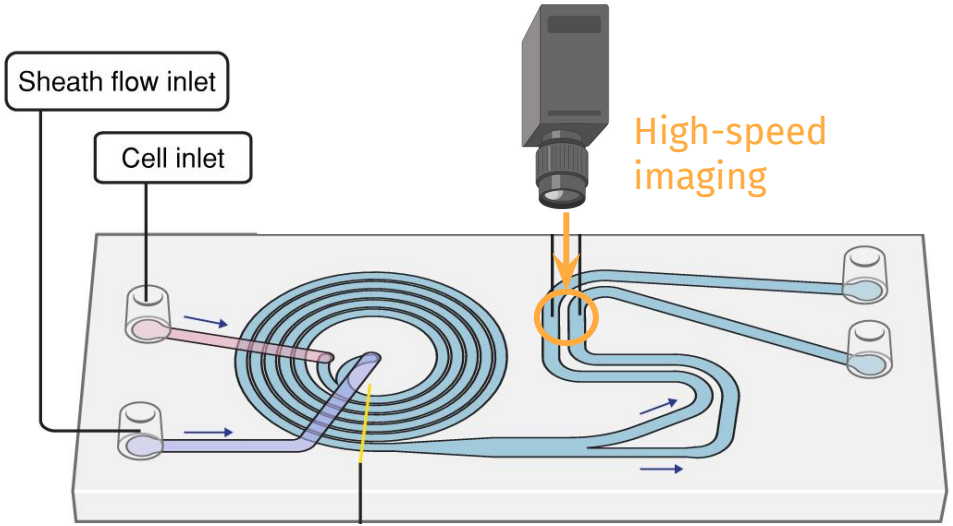
How do the size of heart cells change when we silence a gene or pair of genes?

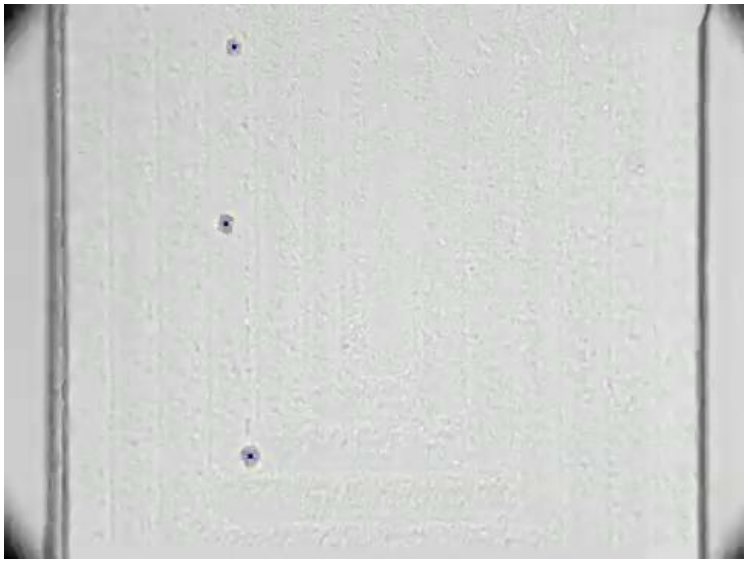


1. Silence Gene A
   → evaluate cell size
2. Silence Gene B
   → evaluate cell size
3. Silence Gene A and B
   → evaluate cell size

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

High-throughput microfluidics + image processing





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 problem interesting or relevant? And **to whom** is this relevant?
- What is already known about this substantive domain? Any relevant **background information**?
- What are the **current challenges** that make solving this problem difficult?
- What is the specific aim or **contribution** of this present work? What is the **end goal**?

Our aim

To develop an **end-to-end pipeline** for identifying **genes** and **gene-gene interactions** that affect **hypertrophic cardiomyopathy**

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)

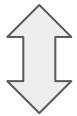
❗ **Setting:** beginning of March 2020

We have PPE to donate to hospitals. Which hospitals are in most need of the PPE so that we can send it to them?

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

Want to predict...

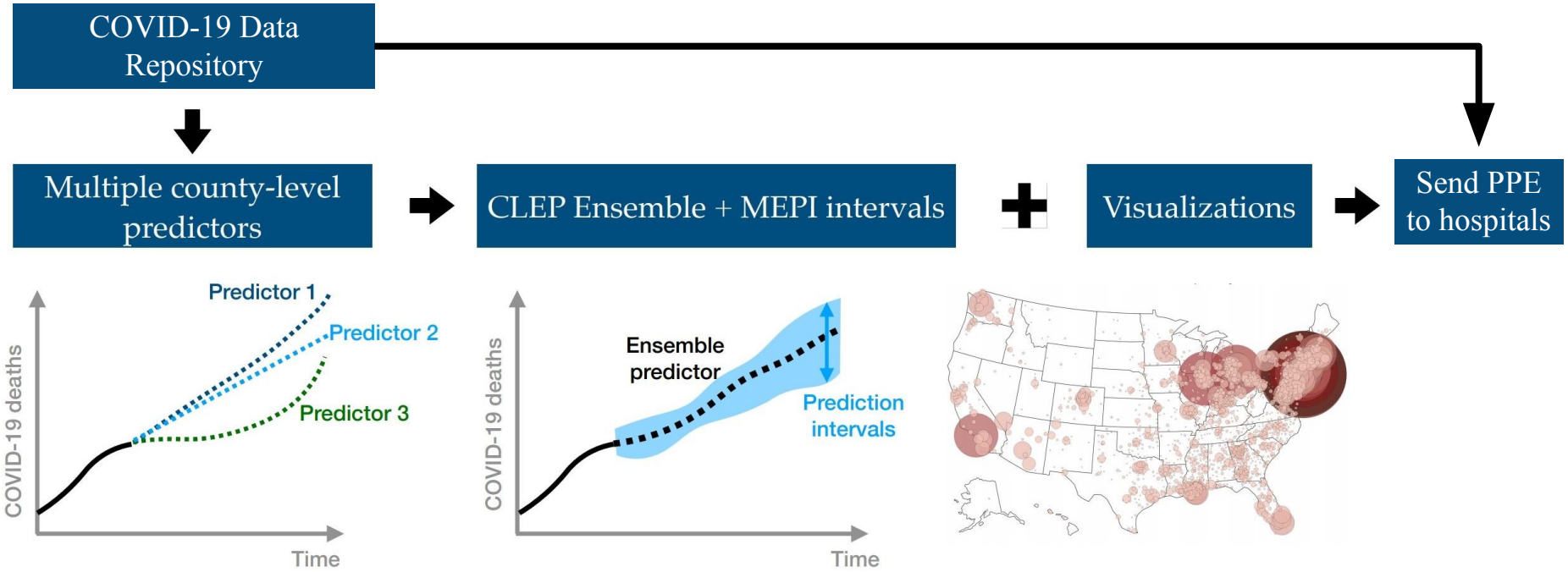
~~hospital PPE/supply need~~



~~number of COVID-19 hospitalizations~~

number of COVID-19 deaths at the county-level

Overview of Modeling Pipeline



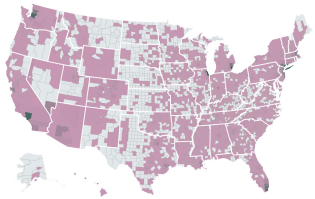
Q: What types of data would you like to have to solve this problem?

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

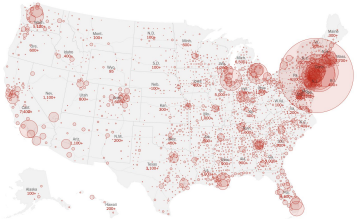
Our Data Repository: scraped from **20+ sources**

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

Introducing the Unacast
Social Distancing Scoreboard

KHN KAISER HEALTH NEWS

cuebiq

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

Some Highlights of COVID-19 Data Repository

- Monitored and updated daily
- Lots and lots of documentation
- Organized, consistent file structure for easy navigation
- Hosted on GitHub: https://github.com/Yu-Group/covid19-severity-prediction/tree/master/data

readme.md

Interactive Atlas of Heart Disease and Stroke - All Strokes (2014-2016)

- **Data source**: https://www.cdc.gov/dhdsp/maps/atlas/index.htm
- **Last downloaded**: 04/02/2020
- **Data description**: county-level estimates of mortality rates per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Known data quality issues**: Data values within the table of "-1" or "-9999" indicate "Insufficient Data."
- **Short list of data columns**:
  - **countyFIPS**: county FIPS
  - **StrokeMortality**: estimate of mortality rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Notes**:
  - Data downloaded from the Interactive Atlas of Heart Disease and Stroke, a website developed by the Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention. http://nccd.cdc.gov/DHDSPAtlas.

271 lines (250 sloc) | 33 KB

List of columns - county level

Identifying variables

Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

Data variables

Geographical identifiers

Data Collection: Checklist

BE TRANSPARENT AND DOCUMENT!

- What data is available?
  - Describe what variables were collected
- 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...

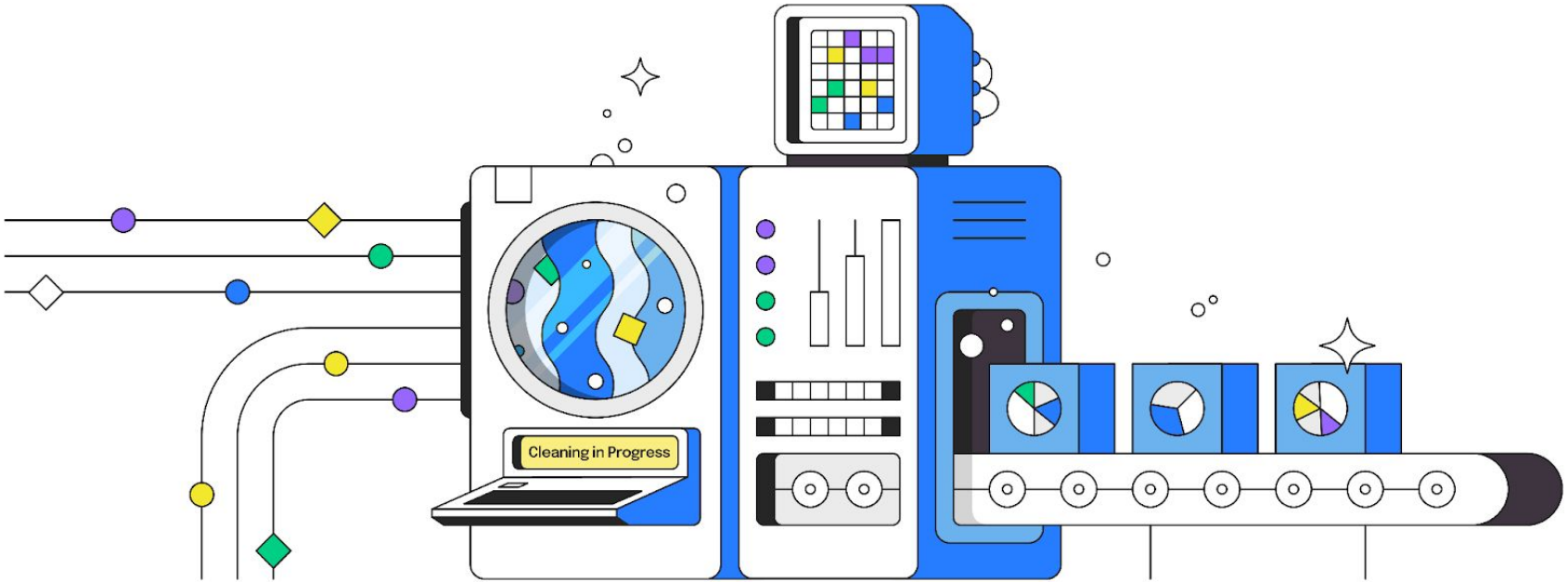


Image Credits: obviously.ai

Data Curation Pipeline

For almost a month, 2 full-time students + others part-time

Data and code available: https://github.com/Yu-Group/covid19-severity-prediction

USAFacts COVID-19 County-level Case & Death Counts Data

Got the data from website

Q: What are potential data issues to look out for?

countyFIPS	County Nar	State	stateFIPS	1/22/2020	1/23/2020
0	Statewide I	AL	1	0	0
1001	Autauga Cc	AL	1	0	0
1003	Baldwin Co	AL	1	0	0
1005	Barbour Co	AL	1	0	0
1007	Bibb Count	AL	1	0	0
1009	Blount Cou	AL	1	0	0
1011	Bullock Cou	AL	1	0	0
1013	Butler Cour	AL	1	0	0
1015	Calhoun Co	AL	1	0	0
1017	Chambers (	AL	1	0	0
1019	Cherokee C	AL	1	0	0
1021	Chilton Cou	AL	1	0	0
1023	Choctaw Cc	AL	1	0	0
1025	Clarke Cour	AL	1	0	0
1027	Clay County	AL	1	0	0
1029	Cleburne Cı	AL	1	0	0
1031	Coffee Cou	AL	1	0	0
1033	Colbert Cou	AL	1	0	0
1035	Conecuh Cc	AL	1	0	0
1037	Coosa Cour	AL	1	0	0
1039	Covington (	AL	1	0	0
1041	Crenshaw (	AL	1	0	0
1043	Cullman Co	AL	1	0	0
1045	Dale Count	AL	1	0	0
1047	Dallas Cour	AL	1	0	0
1049	DeKalb Cou	AL	1	0	0
1051	Elmore Cou	AL	1	0	0

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
  - **Data collection:** garbage in, garbage out
- Ask questions, use common sense, and document everything