Beginning your Data Science Life Cycle

Problem Formulation, Data Collection, & Data Cleaning

January 22, 2025

Today's plan

1. **Review: Problem Formulation + Data Collection**

2. **Data Cleaning**

3. **Introduction to Lab 1**

4. **Setting up Git/GitHub**

Review: Problem Formulation + Data Collection

Review of last time

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**

Case Study: Cardiovascular Genomics

Euan Ashley, MD, PhD (Stanford University)

> Which gene interactions are important drivers of heart disease?

+ How do you define "interaction"?

+ What type of data do we have? Any limitations with the data?

+ Which heart disease? How do you quantify that heart disease?

+ Can you tell me more about the particular type of heart disease that you want to study?

+ Why do you want to find these interactions? How many genes?

+ ...

🔑 Communicate, communicate, communicate across the full scientific team!
+ Learn about the **data** and the **science**

Case Study: COVID-19 PPE Resource Allocation

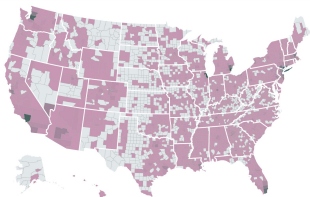**Don Landwirth (Response4Life)**   ❗ **Setting:** beginning of March 2020

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

+ What types of PPE?

+ Where are the hospital regions of interest?

+ How long does it take to deliver the PPE to various places?

+ Loads of questions about PPE availability and the available hospital data...

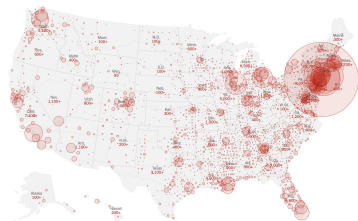🔑 No data → modify the problem formulation + **justify** your modifications

Data Collection (from 20+ sources)

COVID-19 Cases/Deaths

County-level Data
(Risk Factors, Demographics, SES, Social Mobility)

Hospital-level Data
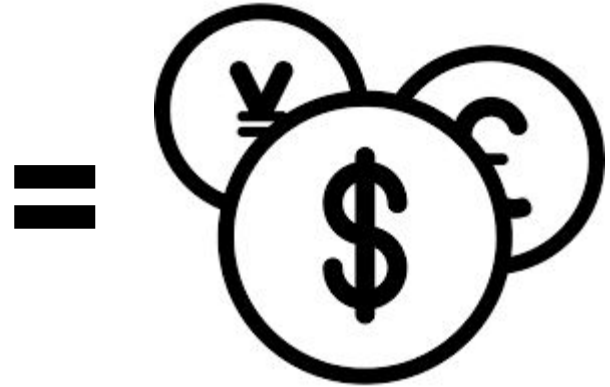(e.g., #ICU beds, staff)

Samuel Scarpino
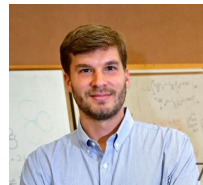
Data is the real currency

Image credit: Dall-E

Lack of representation in training data reinforces societal biases

+ **Google Photos tagging**: found to be racially biased
    + Mistakenly tagged Black people as gorillas [Forbes report (2015)]
    + "Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's." [NYT (2023)]

+ **Facial Recognition:** typically higher error rates for women and people of color

+ **Algorithmic biases in healthcare:**
    + Many dermatological AI tools are trained predominantly on images of lighter skin tones
    + Pulse oximeters are known to be less accurate in individuals with darker skin
        + During COVID-19 pandemic, this led to under-detection of hypoxemia in Black patients
    + Very classical heart disease diagnosis algorithms were developed in trials with majority white males → misdiagnosis and/or delayed treatment for minority populations

We've got the data. What's next?
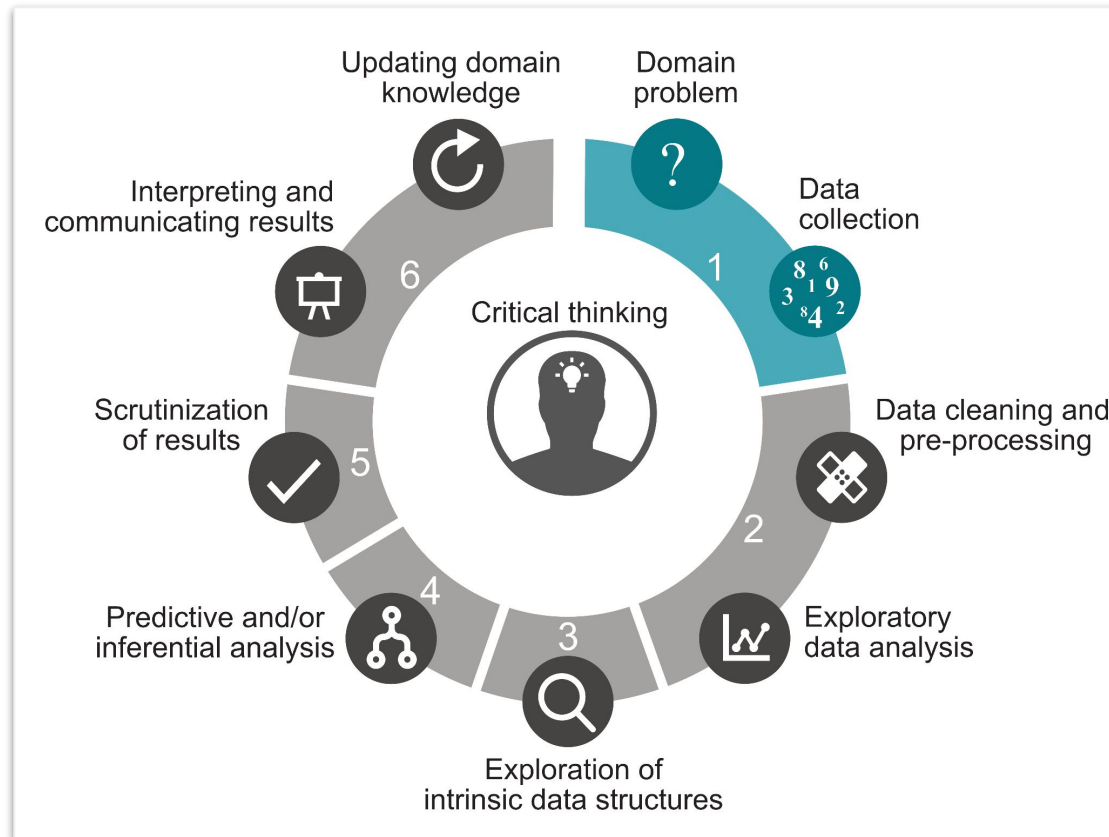
We've got the data. What's next?
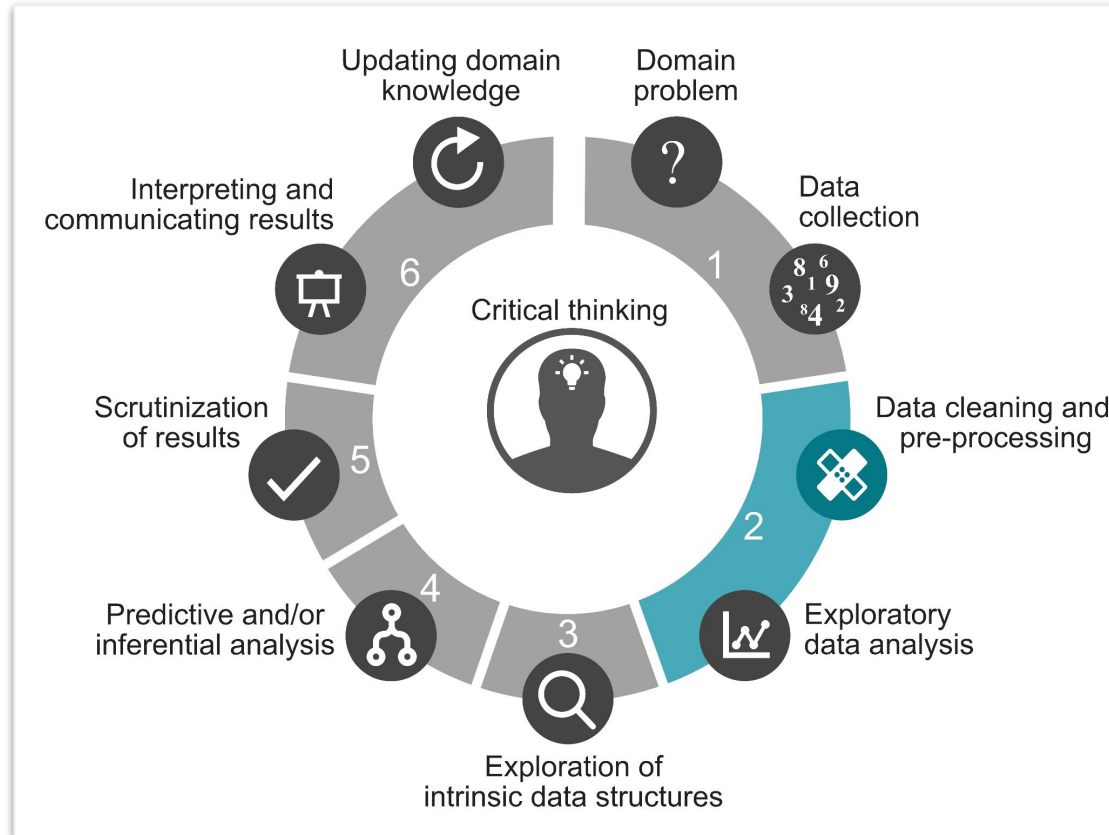
Data Cleaning

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	AL	1	0	0
1001	Autauga Co	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 C	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

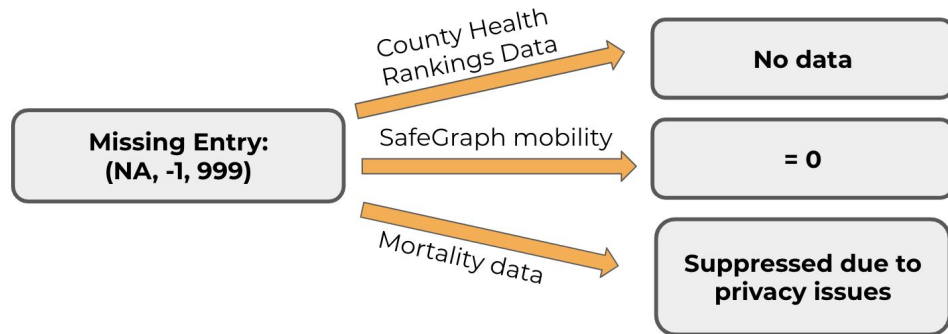# USAFacts COVID-19 County-level Case & Death Counts Data

Issues found from an **examination** of the data + (basic) **domain knowledge**:

- **Irrelevant entries:** remove state entry if interested in counties only

- **Duplicates:** multiple entries for the same county

- **Invalid values:** *cumulative* deaths counts sometimes decrease

- **Missingness:** some cases/deaths cannot be allocated to a county

- **Data revisions:** some case/death counts are revised days after posted
    - Also a major change in the reporting "probable" cases/deaths ~ April 14

Some issues are *fixable* while some aren't, but we do what we can and **document** as much as possible.

More Data Cleaning!

Improperly coded missing values

Missing Entry: (NA, -1, 999)		
County Health Rankings Data →		No data
SafeGraph mobility →		= 0
Mortality data →		Suppressed due to privacy issues

⚠️ **Read all available documentation!**

More Data Cleaning!

Different ways of encoding the same text

- IN = Indiana = indiana

- St. Joseph = St. Joseph County = St Joseph County = st joseph county

- "01234" zip = 1234 zip

- If there are two Joe Smith's, are they the same person or different?

Data encoding choices for non-numerical data

- Categorical variables → dummy or one-hot encoding

- Text → text embeddings (e.g., word2vec)

- Graph → adjacency matrix or graph embeddings

- Images/videos → matrix or tensor of pixel/RGB values

More Data Cleaning!

Data transformations

- Using log-, sqrt-, or Box-Cox-transformations to transform highly-skewed data to more Gaussian-like data

- Should we center and scale the data so that each feature's values have mean 0 and variance 1?

Row/column filtering

- Feature selection: which features do we keep?

- Outlier removal: which samples do we remove?

Feature engineering

- Can we create new features that might be helpful for our task (maybe based on domain knowledge)?

Tips + Takeaways from data cleaning

"The data cleaning procedure will involve learning about the data collection process, obtaining domain knowledge, examining the data, asking and answering questions, checking assumptions, and identifying and documenting any issues and inconsistencies…"

- Veridical Data Science (VDS) Textbook (ch 4)

It's an iterative and interactive process!

Summary of the data science life cycle so far

+ **Problem formulation:** includes formulation of both the *statistical* and the *substantive* problem

+ **Data collection:** garbage in, garbage out

+ **Data cleaning:** a highly iterative process

    + Choices must be made. Justify them whenever possible.

+ Ask questions, use common sense, and **document** everything