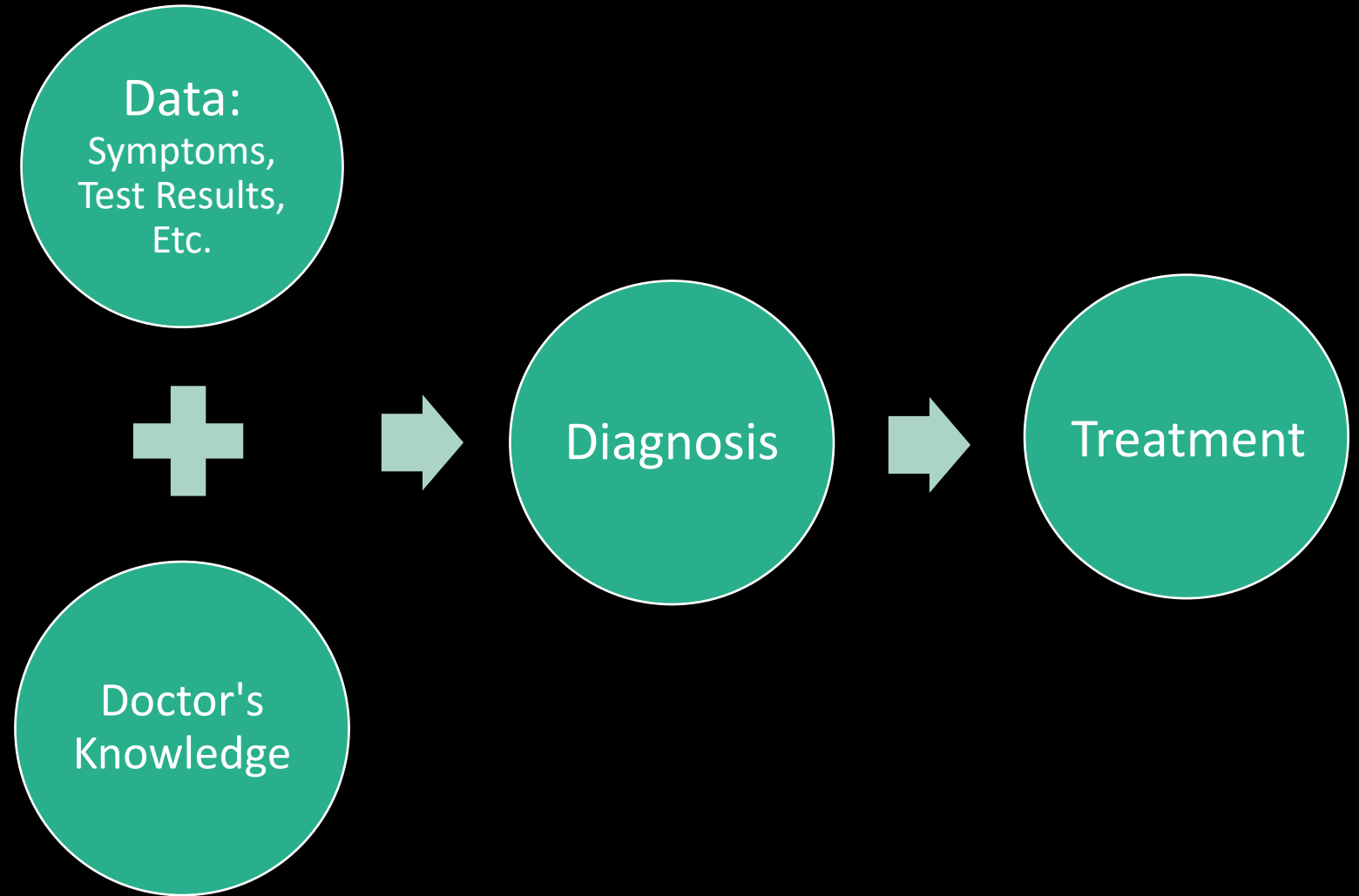




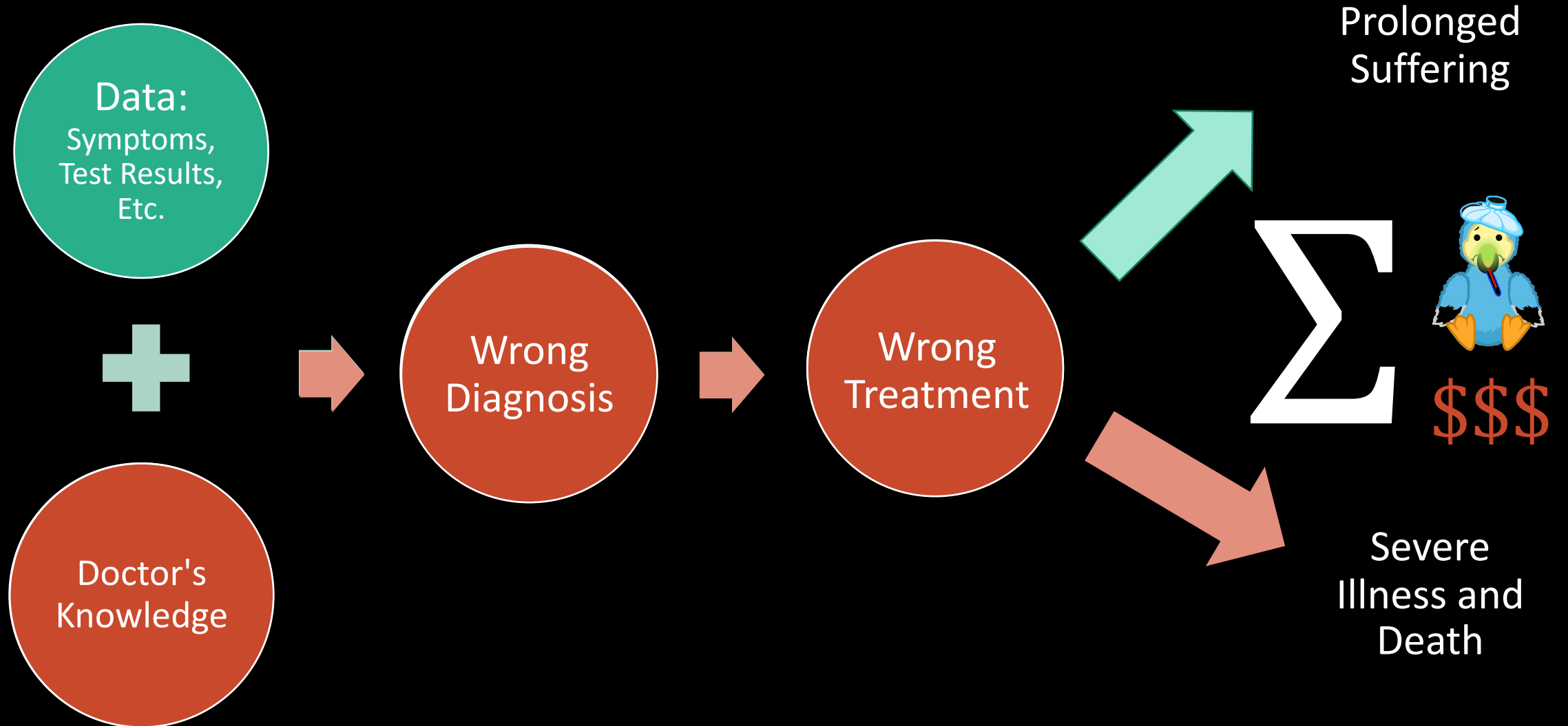
# PATIENT DIAGNOSIS WITH MACHINE LEARNING

Caitlin Ortega Ruble

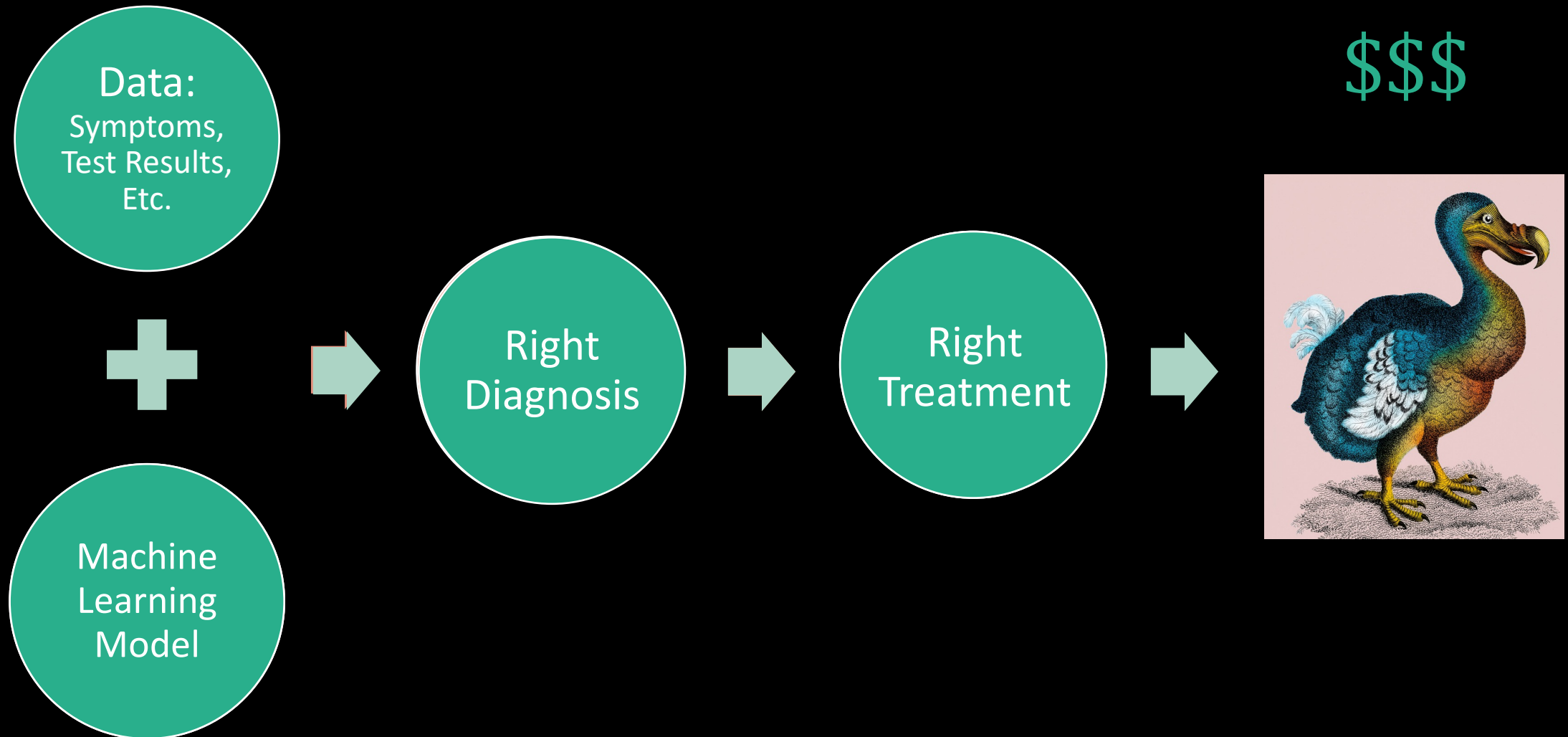
# How diagnosis is done



# How can it go wrong?



# What can we do?



Goal:

build a multi-class classifier that can accurately classify a patient's diagnosis from based on their presenting symptoms, with  $>95\%$  accuracy.

The data: ["Disease Prediction Using Machine Learning"](#) from Kaggle

### Data:

- 4962 patient observations
- 131 symptoms
- 41 diseases

Some light tidying to get the data ready for analysis and modeling

## Data Handling

```
graph TD; A[Data Handling] --> B[Duplicate Features]; A --> C[Formatting]; A --> D[Label Encoding];
```

### Duplicate Features

- One symptom had two associated columns with nearly identical names
- One column was empty
- Empty column dropped, other column renamed

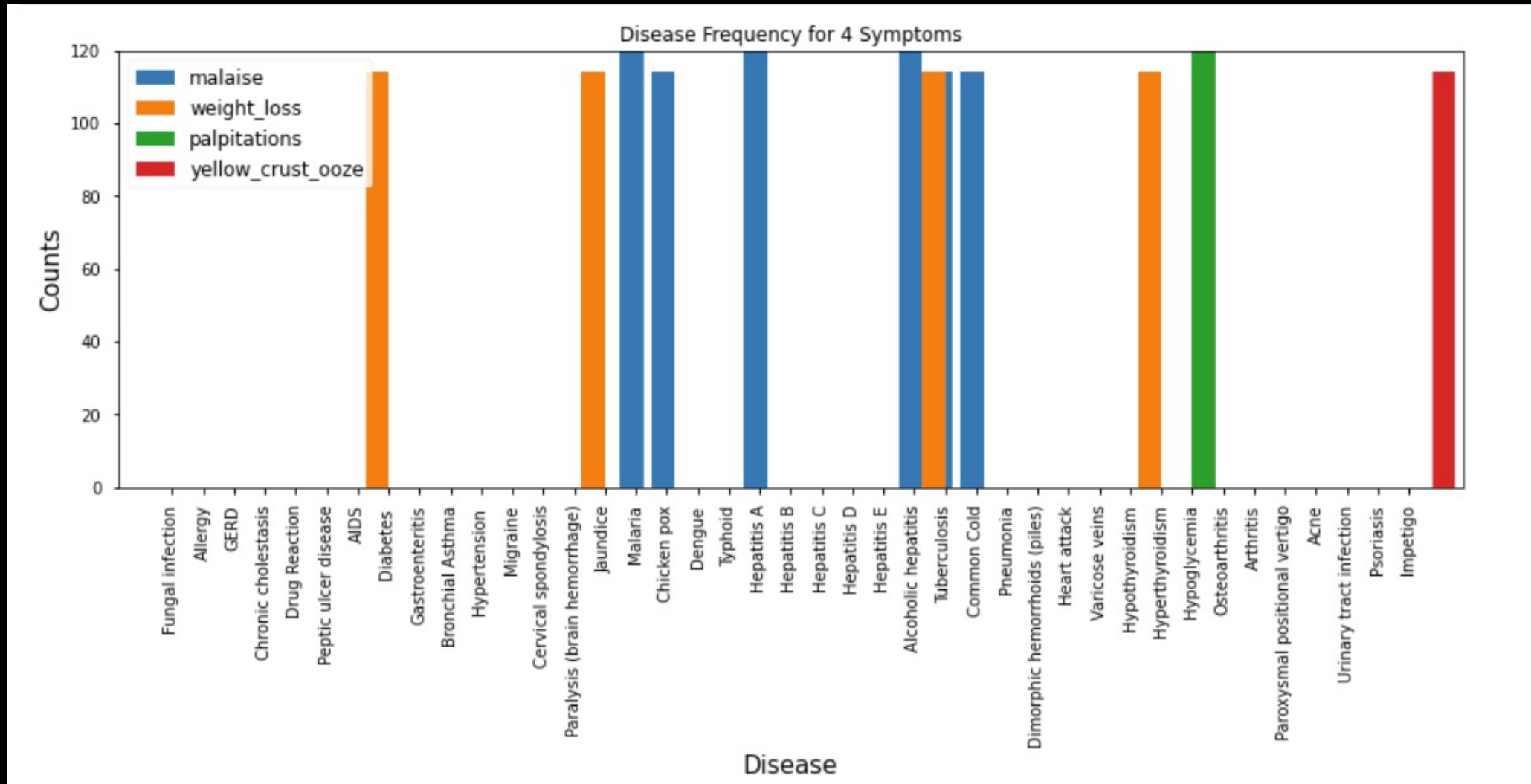
### Formatting

- Several disease names had simple misspellings
  - Replaced with correct
- Some disease names had inconsistent formatting
  - Capitalized first letter

### Label Encoding

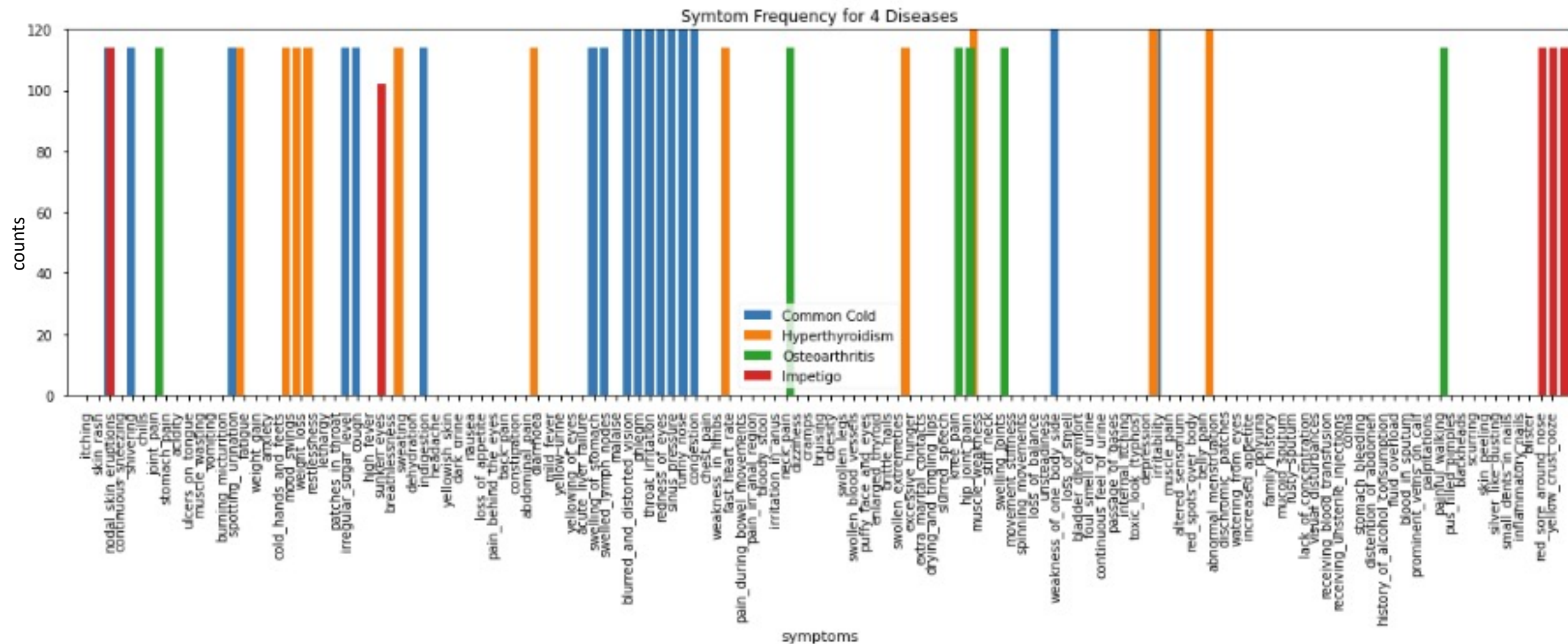
- Disease names were label encoded prior to modeling

Symptoms can be indicated in one or more diseases

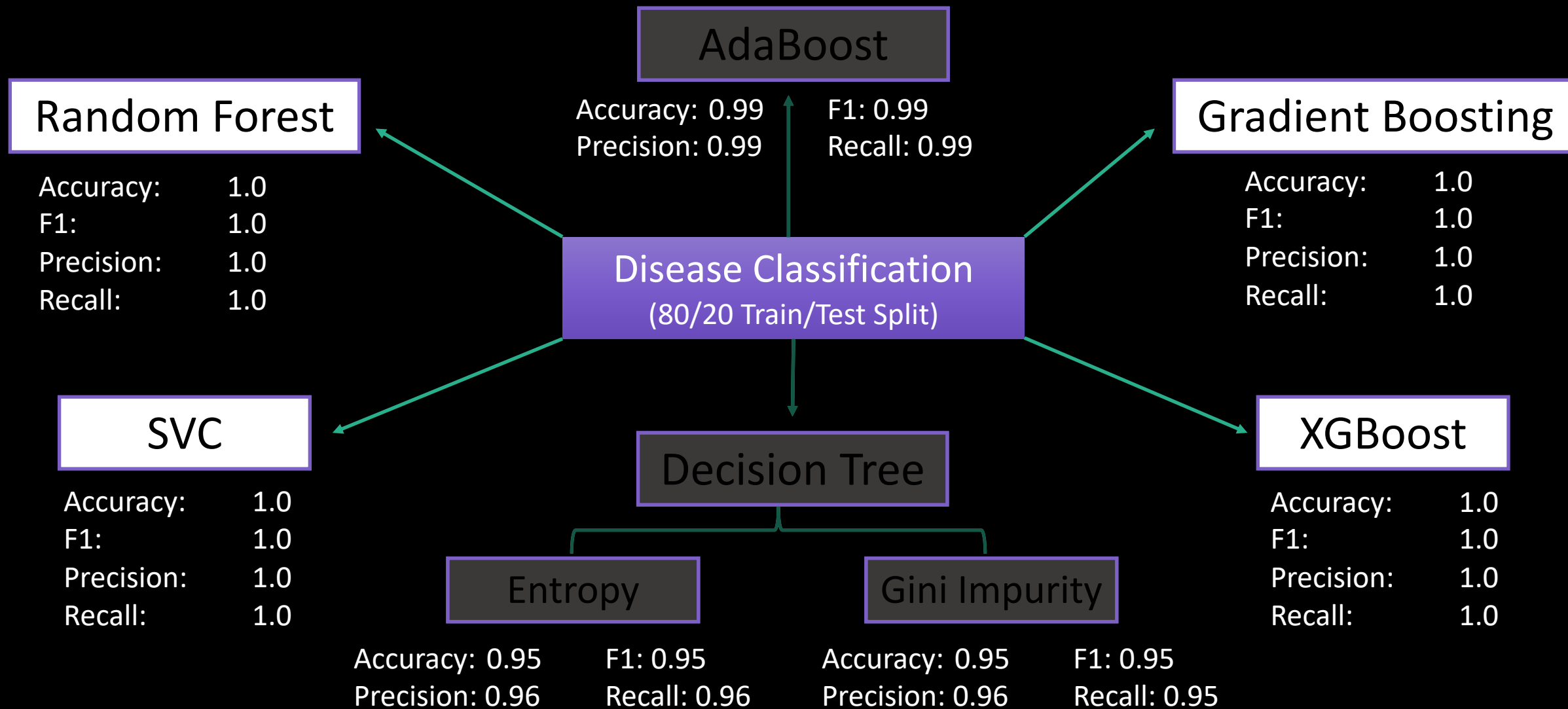




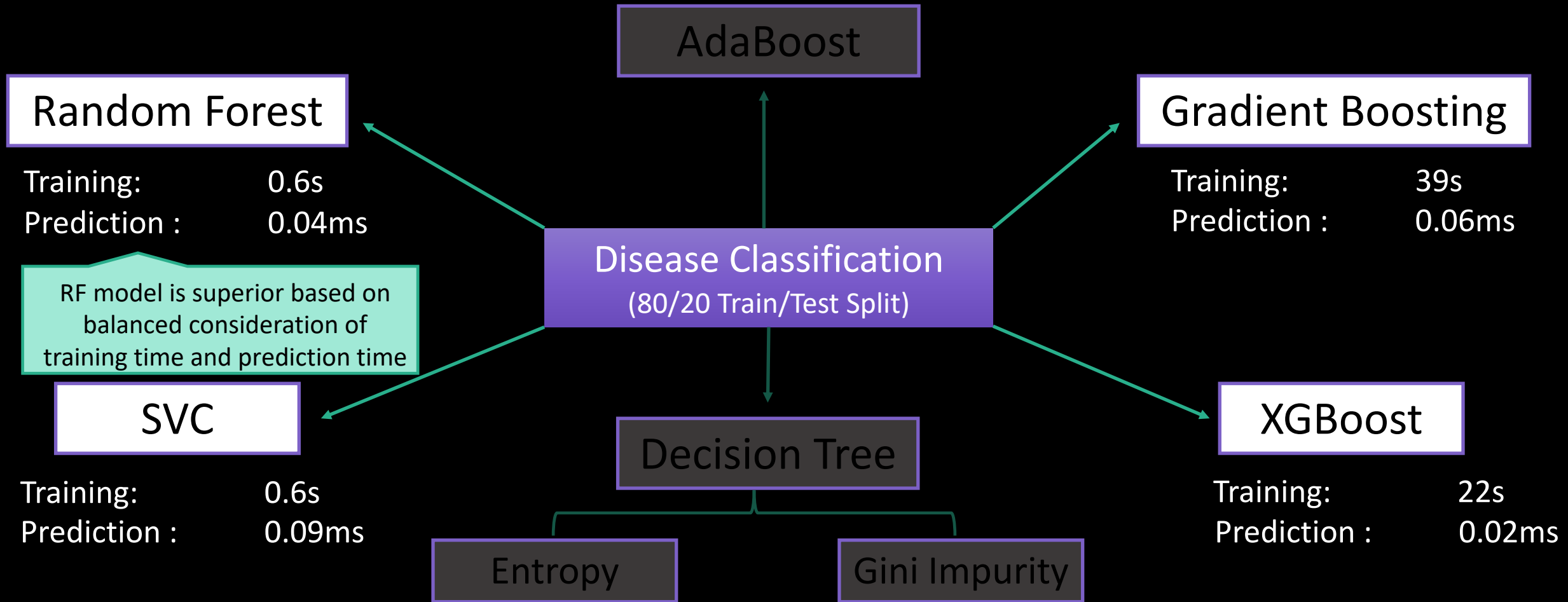
# Each disease has a unique combination of symptoms



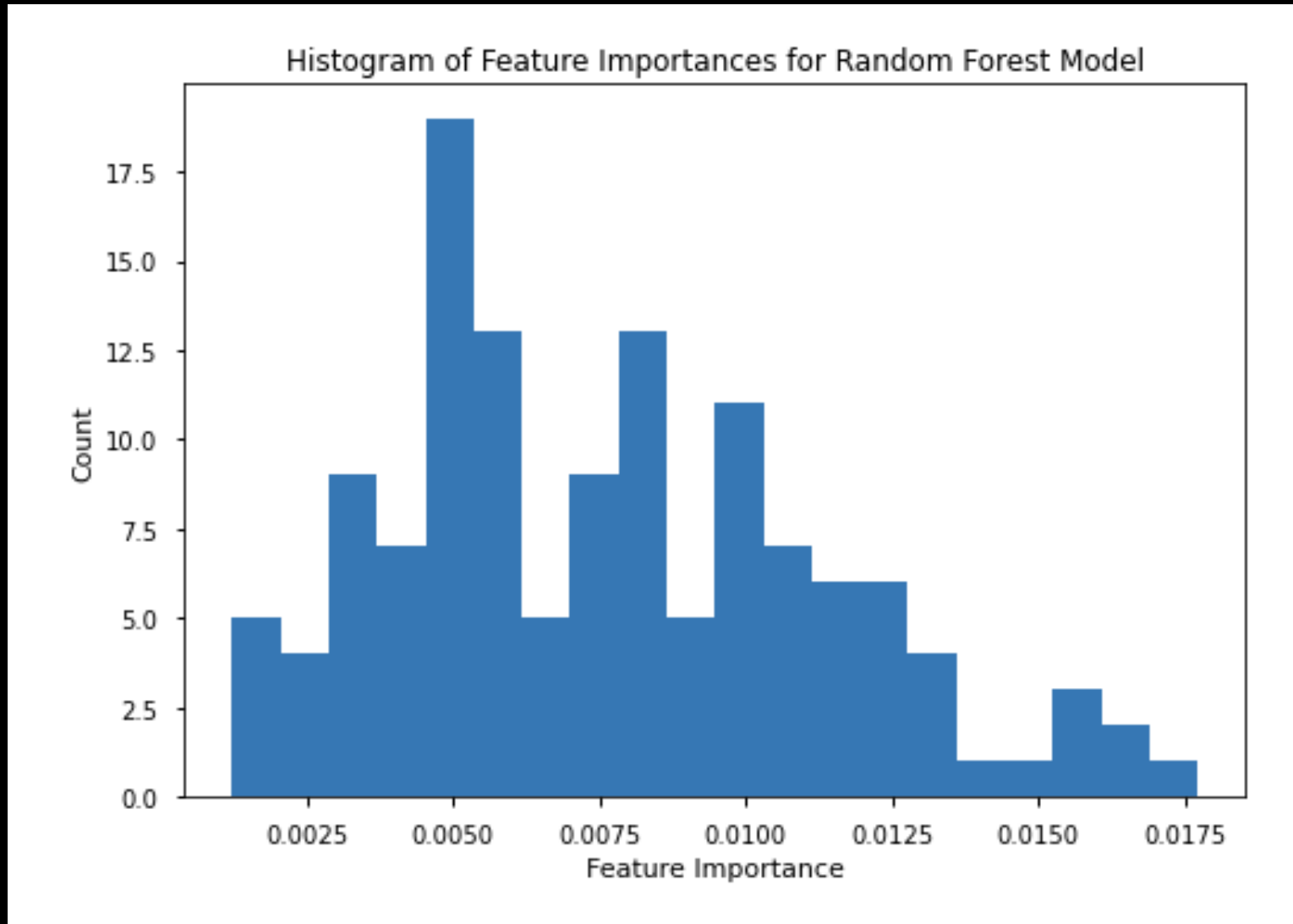
7 classification models were tested and 4 had perfect test accuracy scores



# Training time and prediction time per patient to differentiate models



An analysis of feature importances reassures that the RF model is attributing similar weights to each symptom and that the model is valid



# However...

An investigation of duplicate values at this point told us the data set was almost entirely composed of duplicates.

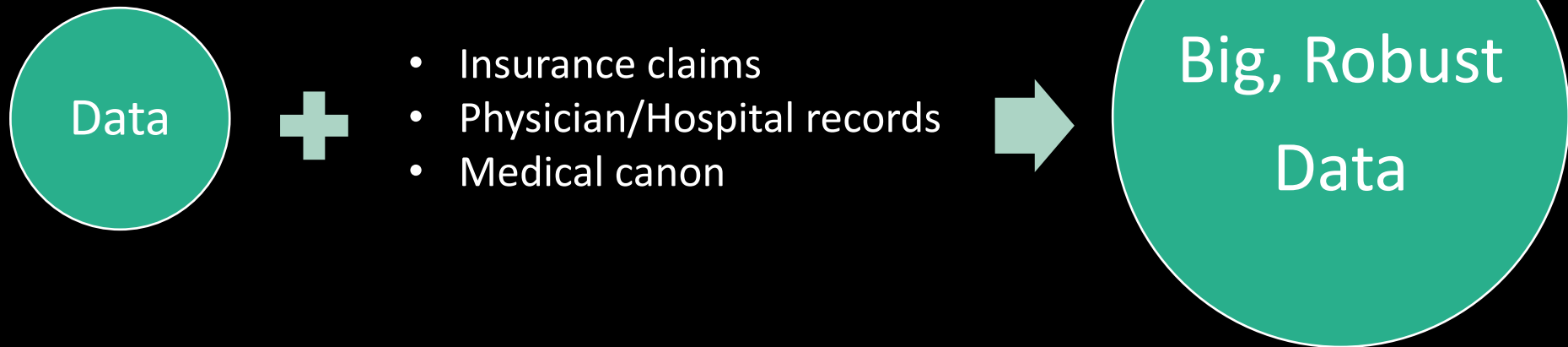
We concluded that the data was most likely a poorly-generated synthetic dataset with no modeling value whatsoever!

A difficult lesson was learned—you ***must*** ensure your data quality! “Garbage in, garbage out.”

The data must be replaced with a sufficient quantity of high-quality, real patient data before the true value of the models can be assessed. Fortunately, the code is ready to go for when the data comes in!

## Where to go from here

- Proof of concept; needs further testing
- Need more volume, variety and quality of data:



- Recommendation System

## Contact Info

For questions, suggestions, feedback and collaboration please email me at:

[caitlinoruble@gmail.com](mailto:caitlinoruble@gmail.com)

