Predicting COVID rates in municipalities

John Caskey, 02-24-2021

Roadmap

- Setup
- Overview of XGBoost
- Impact of COVID-19
- Format/transform data
- Build and test model

- Need to have python installed, or RStudio installed
- Need to have an internet connection to download packages and data
- All code, media, and data are available on GitHub

Python

- xgboost
- pandas
- numpy
- scikit-learn

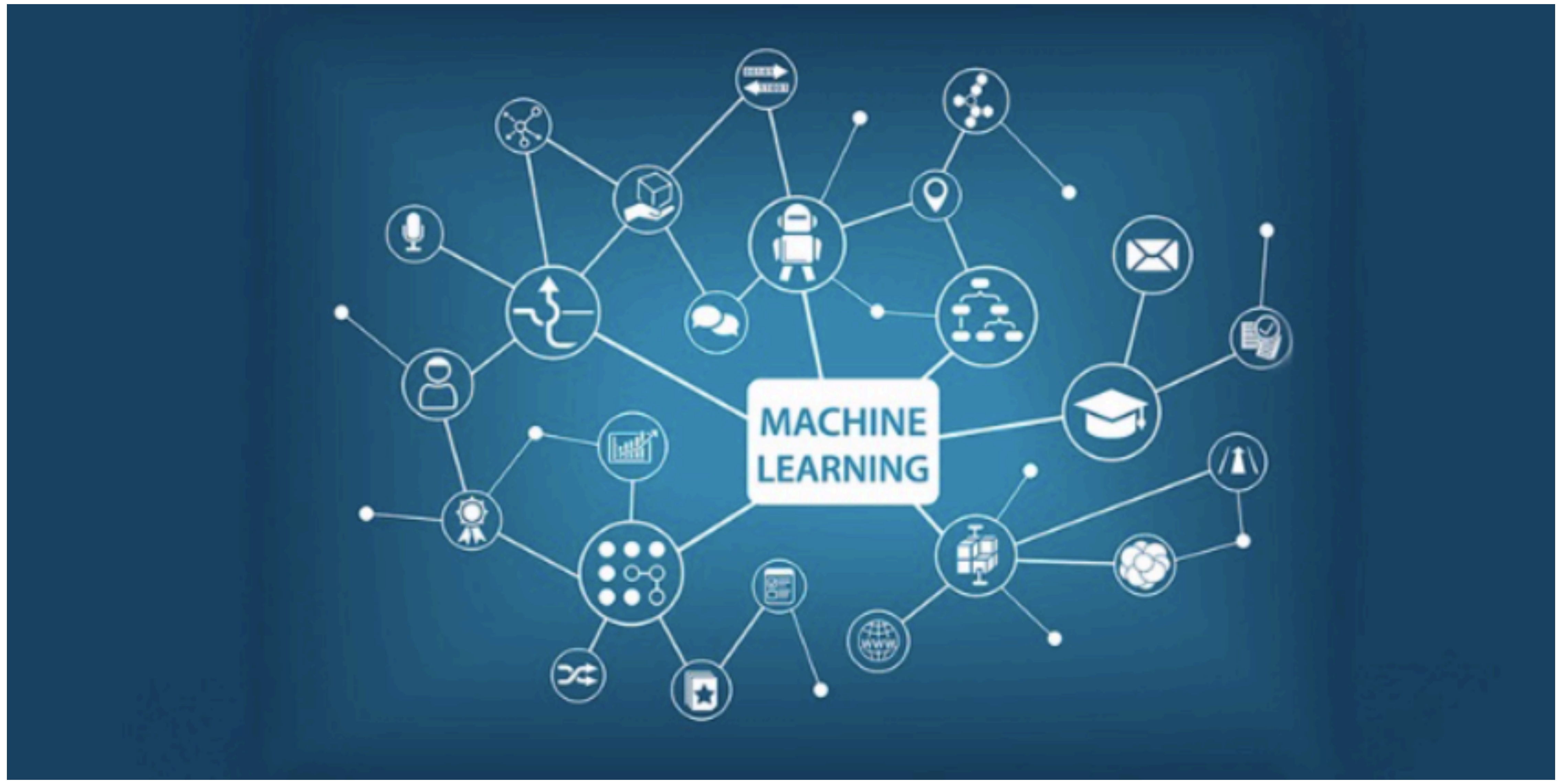
RStudio

- caret
- plyr
- dplyr
- pROC
- ROCR
- xgboost

Questions?

Machine Learning

- Programming computers to make predictions about an input data set
- More efficient and accurate than rules and filters (usually)
- Requires lots of data to be accurate

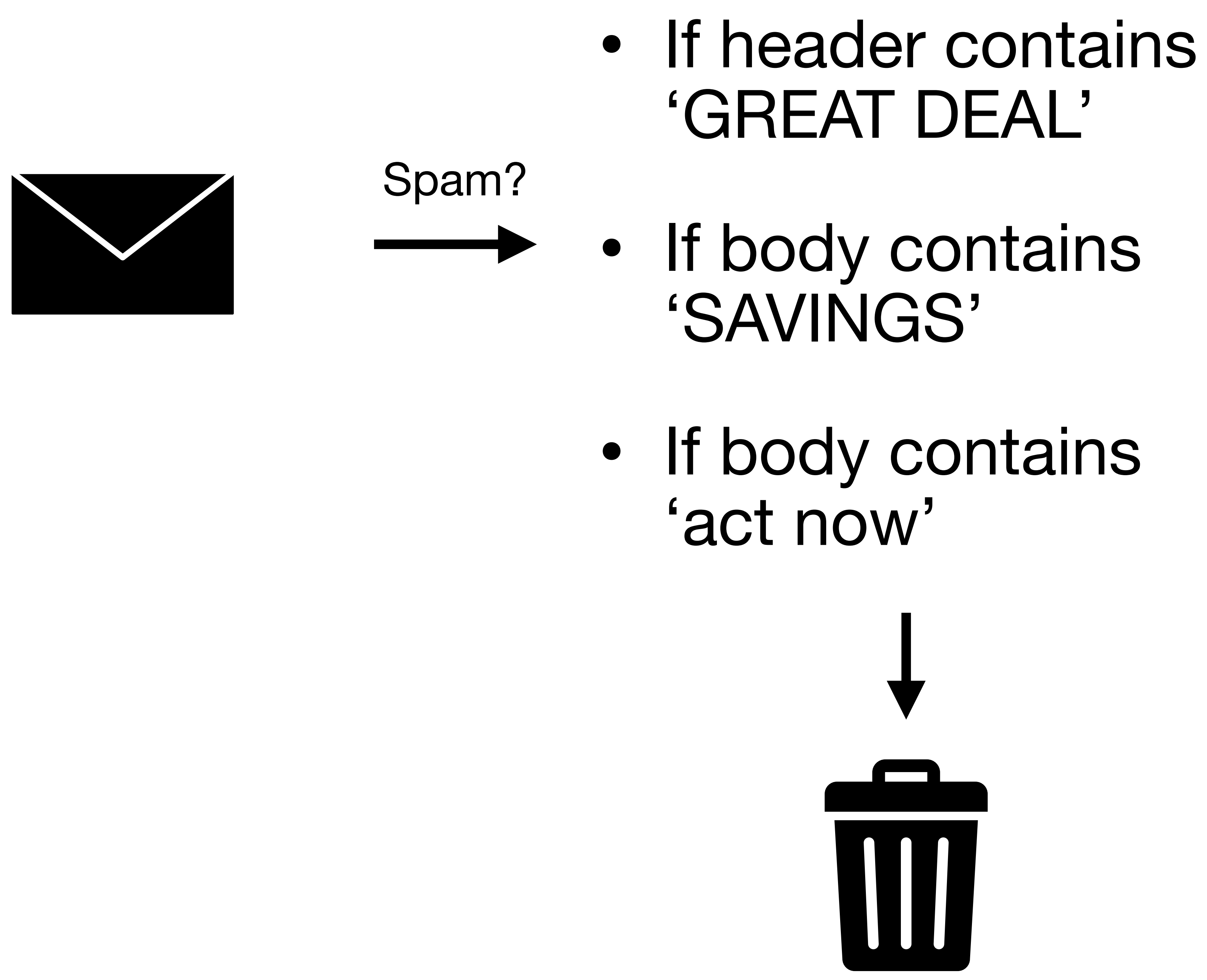


Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*

Image Credit: Towards Data Science https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

Machine Learning

- Programming computers to make predictions about an input data set
- More efficient and accurate than rules and filters (usually)
- Requires lots of data to be accurate



Spam?

- If header contains ‘GREAT DEAL’
- If body contains ‘SAVINGS’
- If body contains ‘act now’

Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*

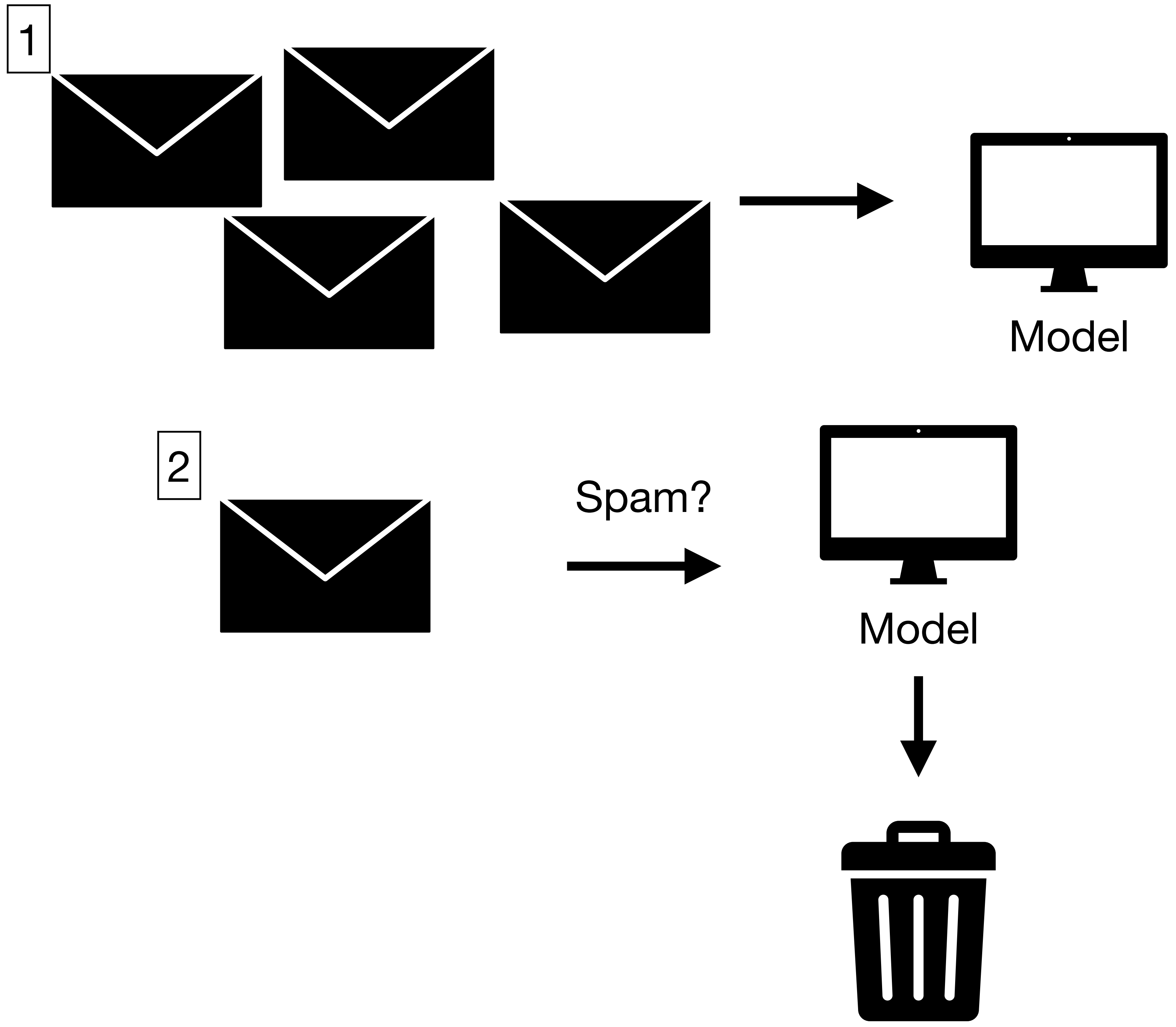
Image Credit: Towards Data Science https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

Machine Learning

- Programming computers to make predictions about an input data set

- More efficient and accurate than rules and filters (usually)

- Requires lots of data to be accurate

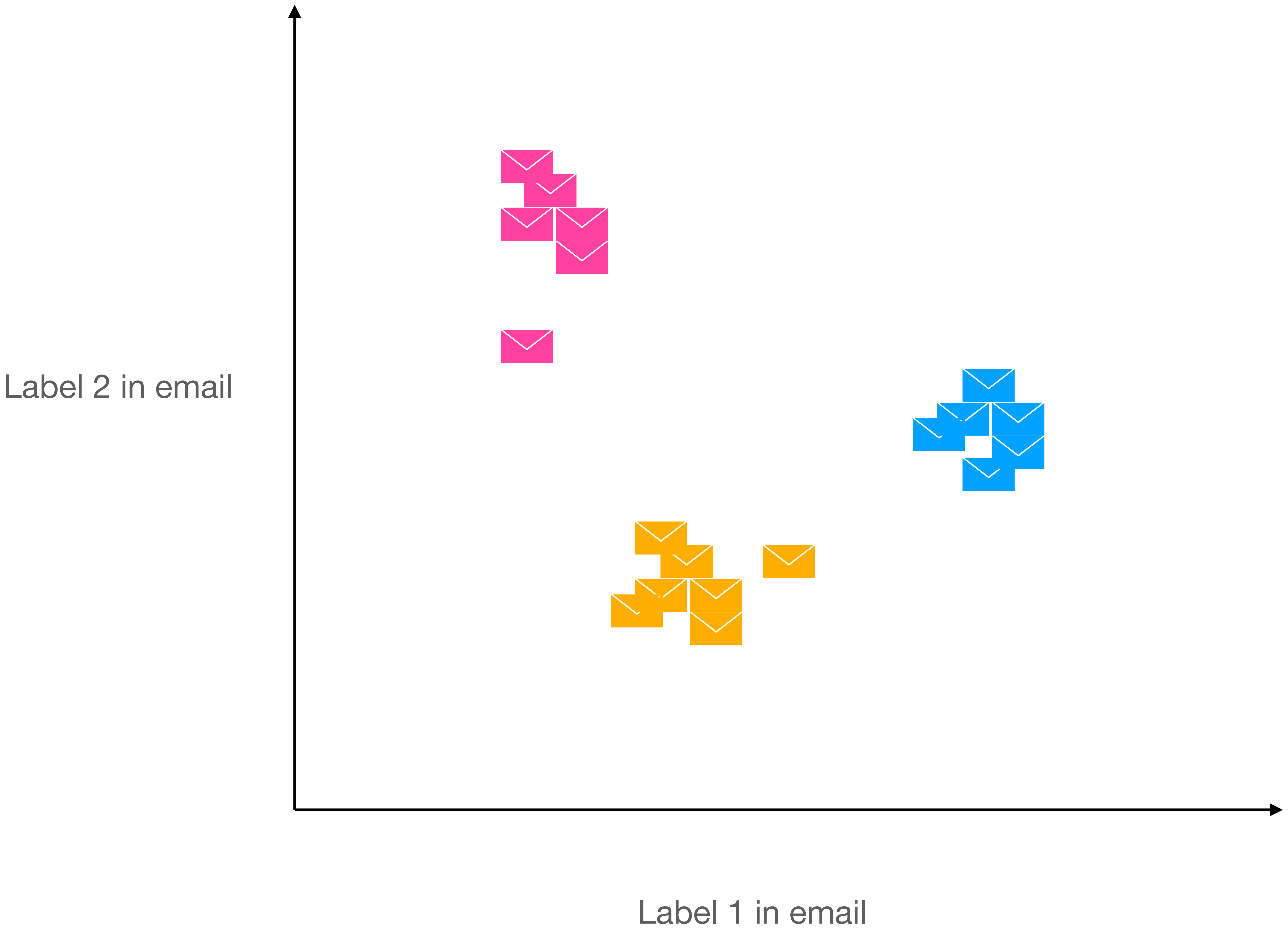


Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*

Image Credit: Towards Data Science https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

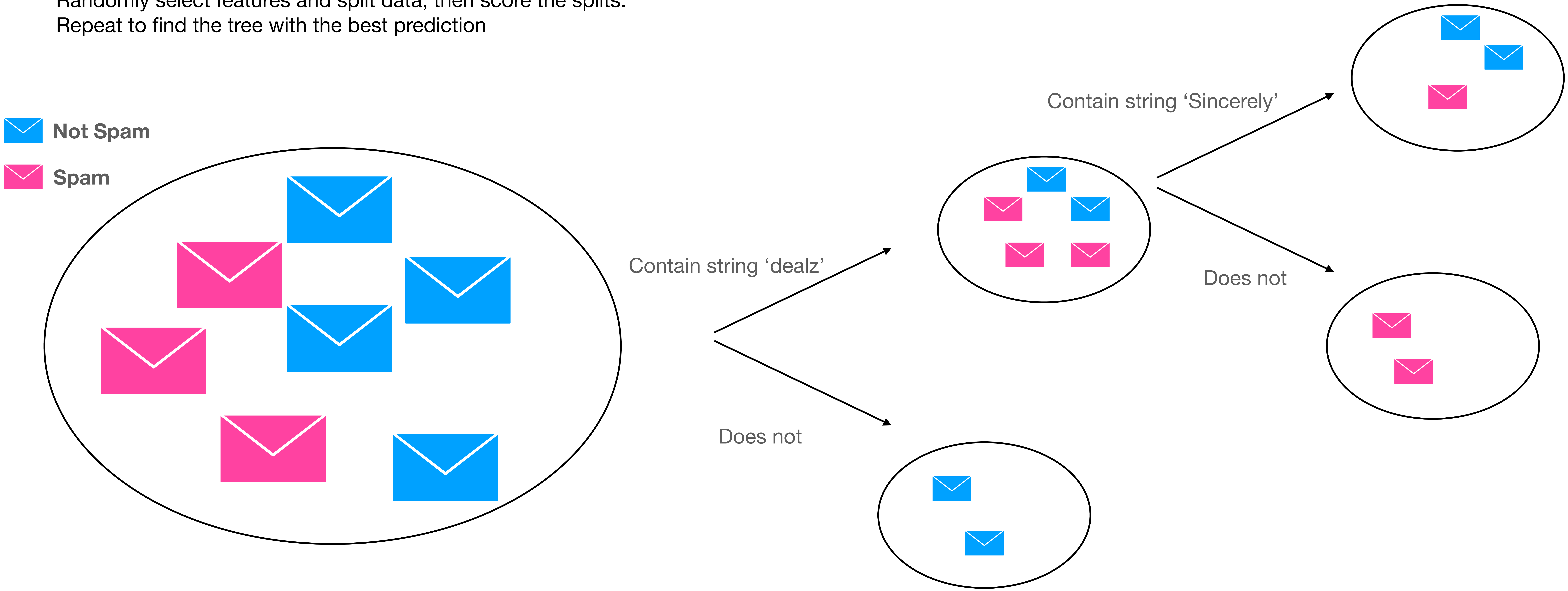
Unsupervised Learning: k means

Among all emails, find 3 clusters with similar characteristics



Supervised Learning: Decision Tree Ensemble

Randomly select features and split data, then score the splits.
Repeat to find the tree with the best prediction



Not Spam

Spam

Contain string ‘dealz’

Does not

Contain string ‘Sincerely’

Does not

https://xgboost.readthedocs.io/en/latest/tutorials/model.html

https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/

https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9

Key Terms

Train model with best parameters Θ to fit training data *x* and labels *y*

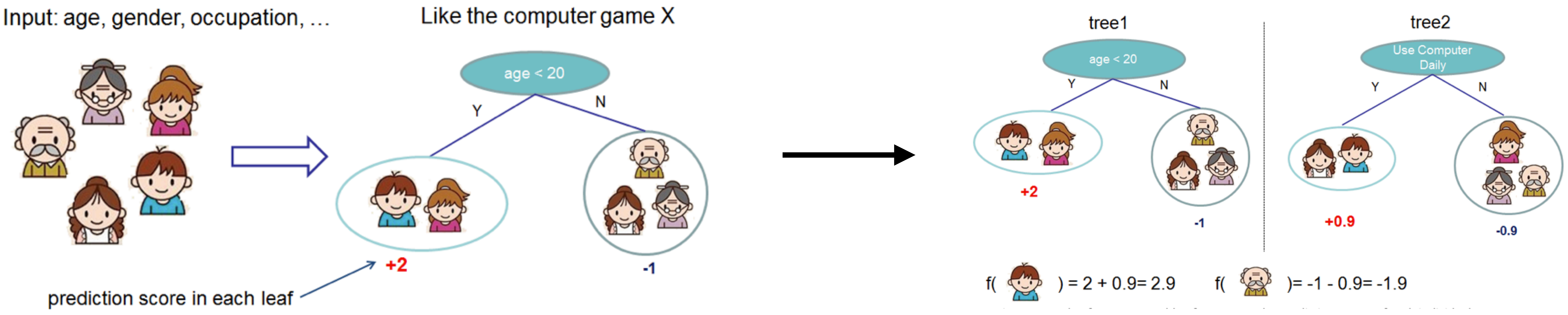
- objective function: measures how well the model fits the training data
- training loss: measure of how predictive the model is with respect to the training data, or how accurate the model's prediction was on data *x* with *y*
- regularization term: a penalty for overfitting the model to data, or too closely fitting the model to *x*

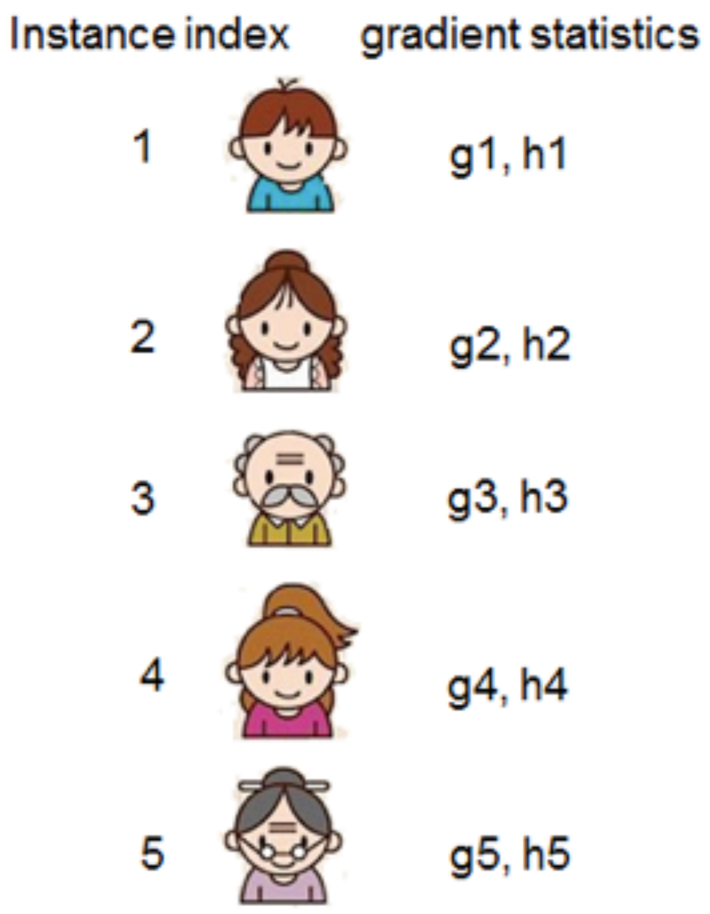
https://xgboost.readthedocs.io/en/latest/tutorials/model.html

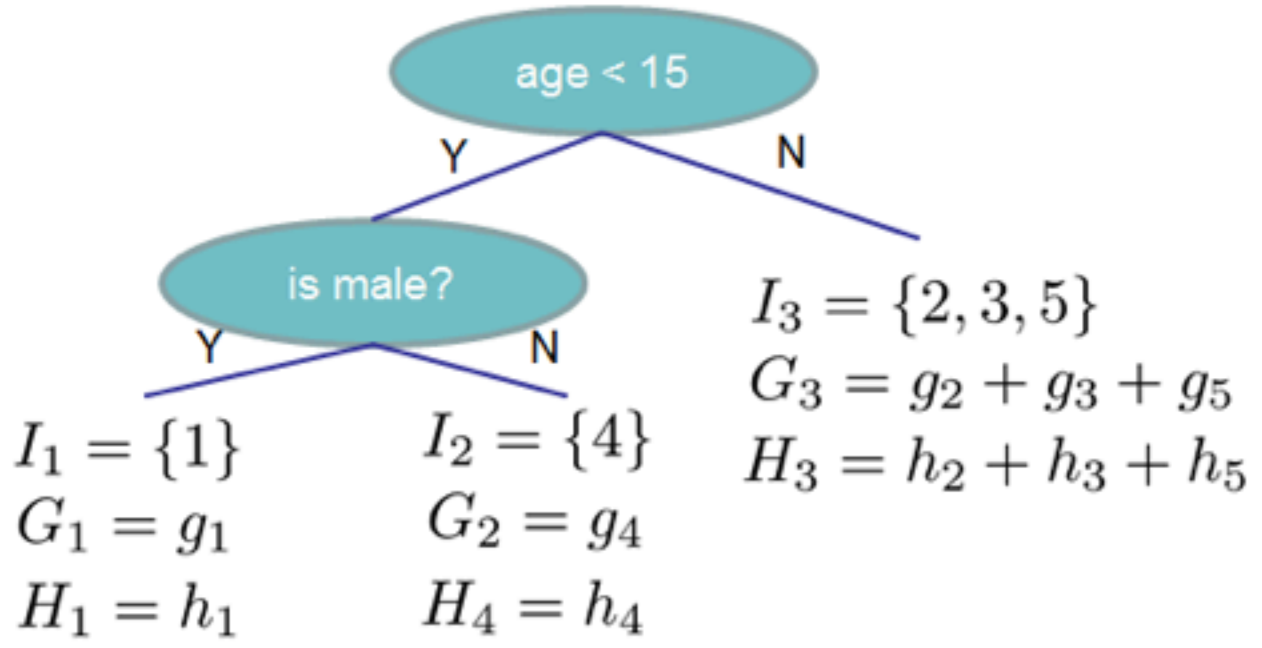
https://towardsdatascience.com/over-fitting-and-regularization-64d16100f45c

XGBoost: Extreme Gradient Boosting

Example: Will a person like computer game x?




Instance index
gradient statistics
1
g1, h1
2
g2, h2
3
g3, h3
4
g4, h4
5
g5, h5



$$Obj = -\sum_j \frac{G_j^2}{H_j+\lambda} + 3\gamma$$

The smaller the score is, the better the structure is

https://xgboost.readthedocs.io/en/latest/tutorials/model.html

Questions?

COVID-19

- Pandemic affecting the world
- Caused by SARS-CoV-2 virus
- 28,828,370 cases reported (as of 02/23/2021)
- 495,270 deaths (as of 02/23/2021)

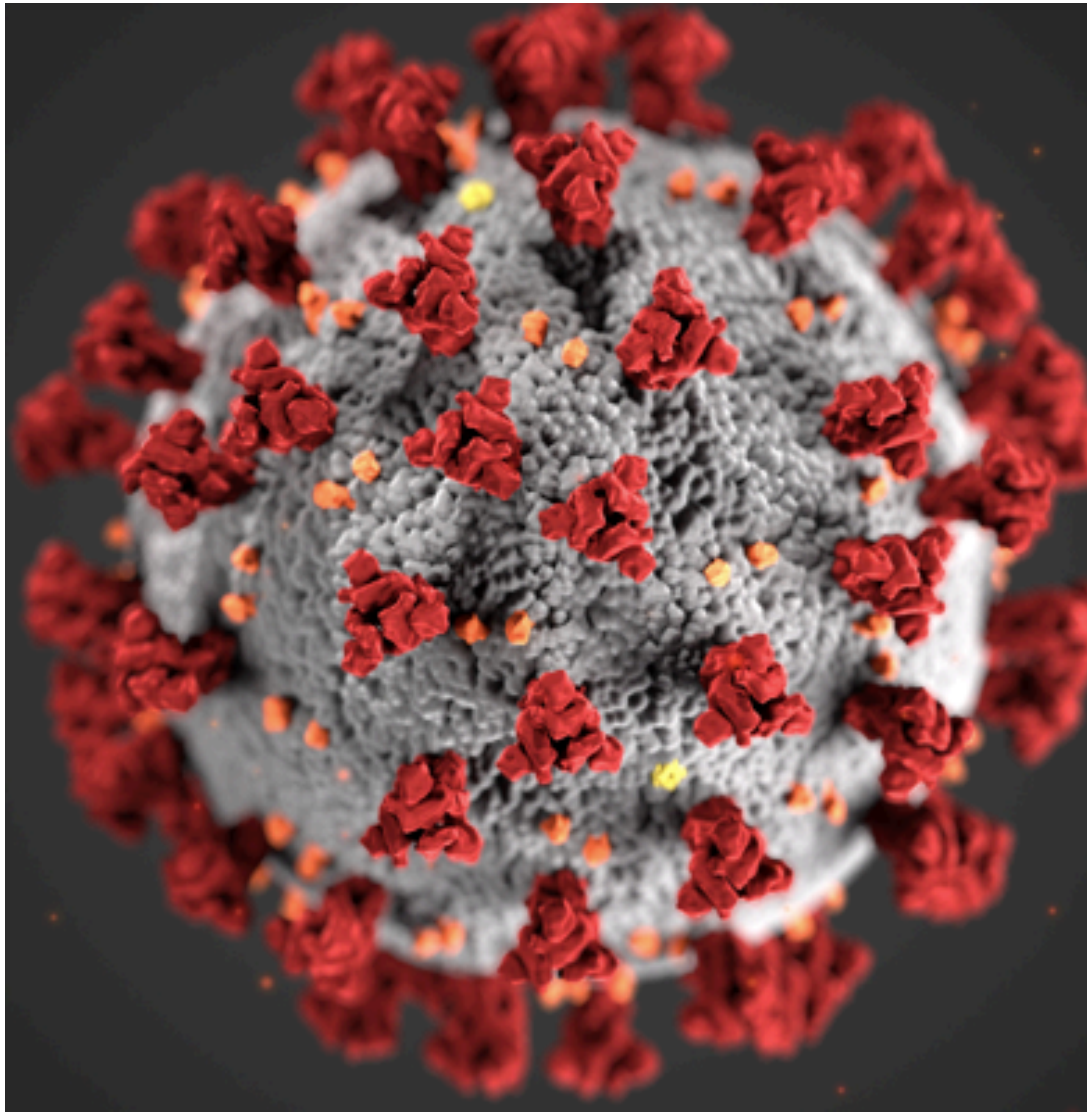
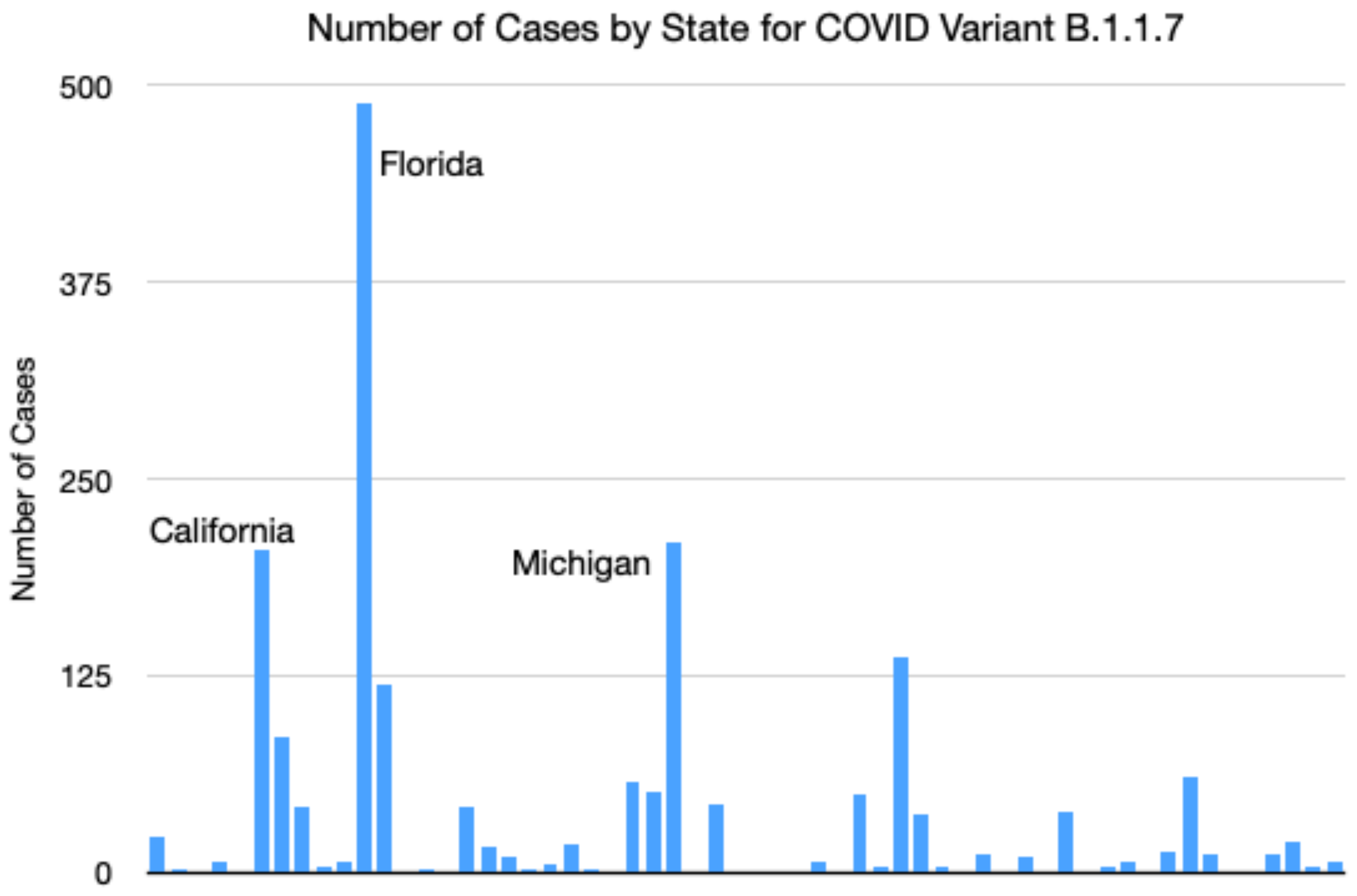


Image: https://www.fda.gov/food/food-safety-during-emergencies/food-safety-and-coronavirus-disease-2019-covid-19

https://covid.cdc.gov/covid-data-tracker/#global-counts-rates

Mask up!

- Vaccination efforts are underway
- Variants are emerging
- Transmission prevention efforts are critical as well as vaccination

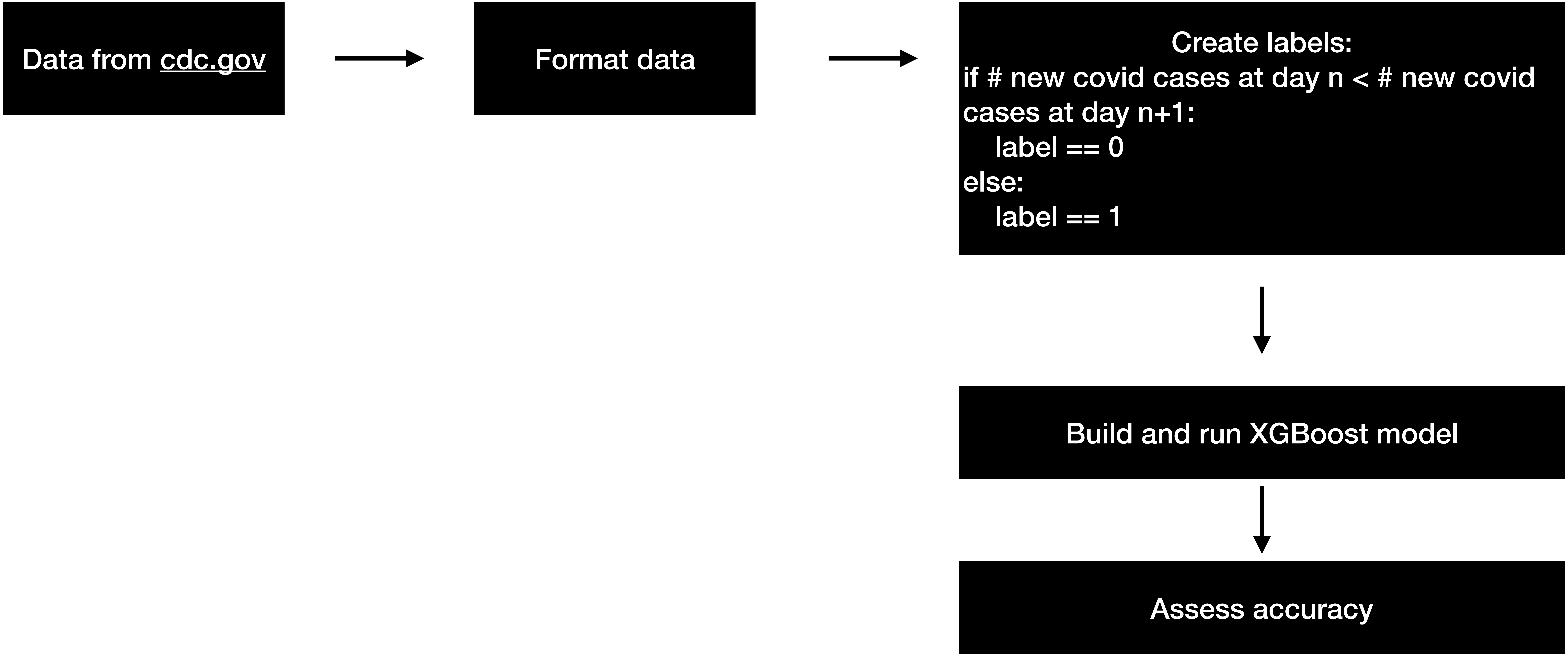


https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant-cases.html
Image: https://www.fda.gov/food/food-safety-during-emergencies/food-safety-and-coronavirus-disease-2019-covid-19

- Is there a way to predict new cases?
    - Yes, several models and predictors exist. The models use numerous parameters, such as how well prevention guideline like social distancing and masking are followed
    - https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html
- Can we use existing data to build a simplified model to predict if there will be an increase in COVID cases the next day? 🤔

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7190554/

The number of hospitalizations and covid deaths can predict whether there will be an increase in COVID-19 cases the following day



Data from cdc.gov

Format data

Create labels:

```
if # new covid cases at day n < # new covid cases at day n+1:
    label == 0
else:
    label == 1
```

Build and run XGBoost model

Assess accuracy

- Download dataset from https://data.dhsgis.wi.gov/datasets/covid-19-historical-data-by-state/data?orderBy=DATE
- Descriptions on the columns is available here:
    - https://www.dhs.wisconsin.gov/publications/p02677.pdf
- Import into Python pandas, RStudio (or Excel)
- Dataset is also available at GitHub

Python RStudio

- Import csv data
- Order by date
- Create columns with counts of new hospitalizations, positive male cases, positive female cases, negative cases
- Create label column, then trim first 20 rows

Import data, sort by date

Python

RStudio

Import data, sort by date

Python

```
df = pd.read_csv('COVID-19_Historical_Data_by_State.orig.csv')

df = df.sort_values('DATE')
df = df.reset_index(drop=True)
```

RStudio

```
df = read.csv('COVID-19_Historical_Data_by_State.orig.csv')
df = df[order(df$DATE),]
row.names(df) <- NULL
```

Create formatted columns and labels

Python

RStudio

Create formatted columns and labels

Python

```
# repeat for the values "POS_FEM_NEW", "POS_MALE_NEW", "HOSP_UNK_NEW",
df['HOSP_NEW'] = df['HOSP_YES'].diff()

# setup labels
df['INCREASE'] = -1
df.loc[df['SCAN'] < 0, ['INCREASE']] = 1
df.loc[df['SCAN'] >= 0, ['INCREASE']] = 0

df = df.iloc[20:]
df_filtered.drop(df.tail(1).index, inplace=True)
```

RStudio

```
# add HOSP_NEW values as difference between current day and prev day
df = df %>% mutate(HOSP_NEW = HOSP_YES - lag(HOSP_YES, default = HOSP_YES[1]))

# repeat for the values: HOSP_NO, HOSP_UNK, POS_FEM, POS_MALE

df = df %>% mutate(SCAN = POS_NEW - lead(POS_NEW, default = POS_NEW))
df$INCREASE = ifelse(df$SCAN < 0, 1.0, 0.0)
```

Build and Run Model!