
Data Science
For Social Good
Statistics Without Borders

Who is Statistics Without Borders?

Statistics Without Borders Helps UNICEF

1 JULY 2016 809 VIEWS NO COMMENT

Stephanie Eckman, Monica Dashen, Aliou Diouf Mballo, and Robert Johnston

Experts Leverage Statistical Methods to Investigate Human Trafficking

Use of GPS-Enabled Mobile Devices to Conduct Health Surveys: Child Mortality in Sierra Leone

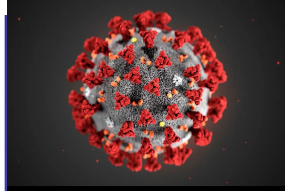
• Columns,Here's to Your Health

Haiti after the earthquake

Statistics Without Borders

When a major disaster strikes, urgent needs may be food, water, shelter, medicines – and data. Unless you know the numbers of people involved and how their lives have been affected, giving efficient help is impossible. Statistics Without Borders tries to provide the data. The team that worked on a project in Haiti describe one effort.

Our Work



COVID-19

California COVID-19 By The Numbers

July 27, 2020

Numbers as of July 26, 2020

USA

4,280,135

TOTAL CASES

CDC | Updated: Jul 28 2020 2:46PM

USA

147,672

TOTAL DEATHS

CDC | Updated: Jul 28 2020 2:46PM

USA

1,306

Cases per 100,000 People

CDC | Updated: Jul 28 2020 2:46PM

CALIFORNIA COVID-19 SPREAD

460,550 (+6,891)

TOTAL CASES

Ages of Confirmed Cases
- 0-17: 41,148
- 18-49: 278,295
- 50-64: 88,800
- 65+: 51,772
- Unknown/Missing: 535

Gender of Confirmed Cases
- Female: 230,423
- Male: 227,545
- Unknown/Missing: 2,582

8,445 (+29)

Fatalities

Hospitalizations*

Confirmed COVID-19: **6,935/2,012** Hospitalized/in ICU

Suspected COVID-19: **1,484/209** Hospitalized/in ICU

For county-level hospital data: bit.ly/hospitalsca

Your actions save lives.

For county-level data: data.chhs.ca.gov

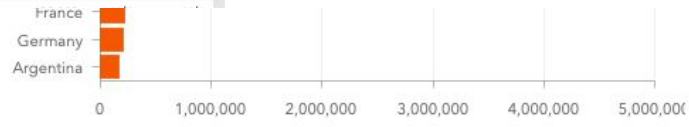
covid19.ca.gov

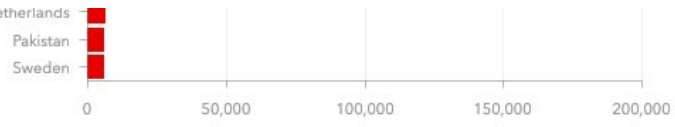
Total Cases by State/Territory

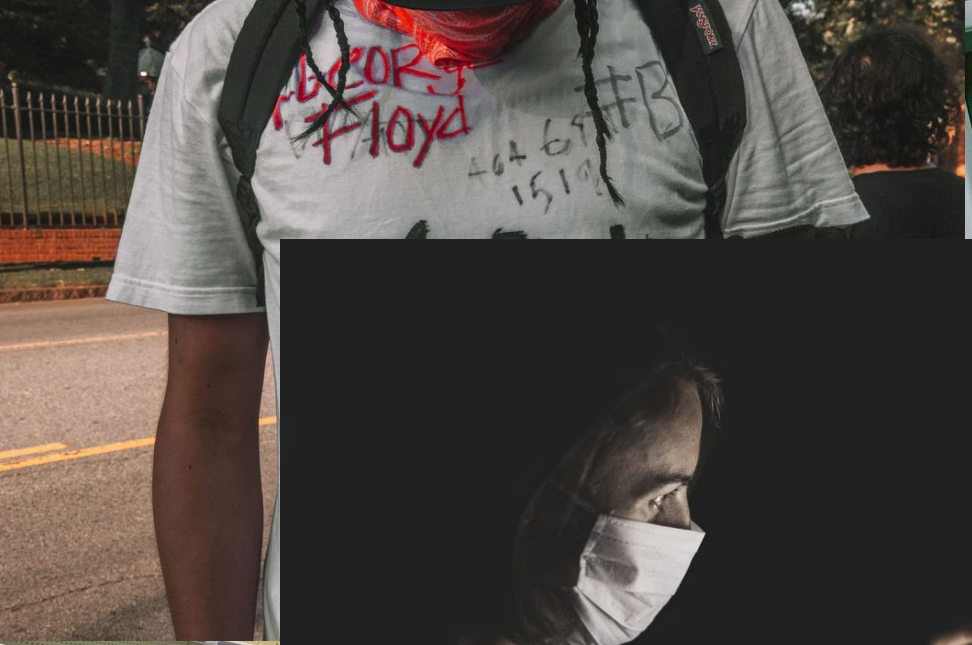
State/Territory	Total Cases	Confirmed	Probable
California	460,550	N/A	N/A
Florida	427,698	N/A	N/A
Texas	385,923	N/A	N/A
New York City*	225,541	220,907	4,634

Territories

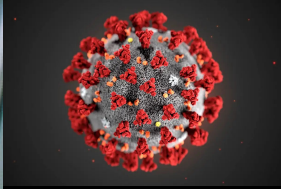
AS | FSM | GU | MP | PR | PW | RMI | VI



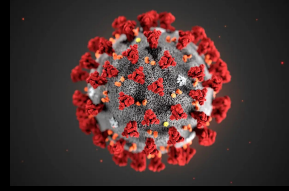



#Floyd


Working 24/7 for your journey
Working 24/7 for your journey



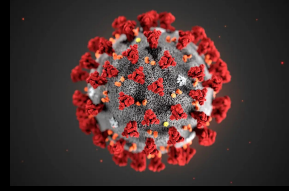
COVID-19



SWB + MCCERT

According to OMNICORE (2020):

- > 330 million active Twitter users
- > 500 million tweets are posted per day
- 71% of users say they use Twitter to get their news (Pew Research Center, 2019)



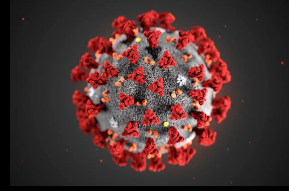
SWB + MCCERT

Client Organization: Montgomery County, Community Emergency Response Team

- MCCERT asked SWB to try this methodology in an independent geographic area around Palo Alto, California

- **Steve Peterson**
  - Virtual Emergency Response Team

    - Steve developed a framework (Peterson et al., 2019) to utilize Twitter data to inform emergency response in the National Capital Region using George Mason University's streaming analytics system, Citizen Helper.



SWB + MCCERT

3-Month Project

8 SWB Volunteers

Julia Reid, PCM

Keri Wheatley

Heli Vora

Satyajeet Pradhan

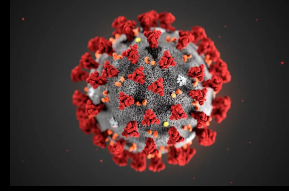
Rachel Doehr

Qingyuan Wang

Harshit Sharma

Lena Lickteig, DQA

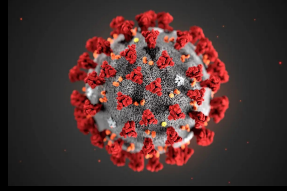
Collaboration by video, email, phone, text, and chat



SWB + MCCERT

Objective:

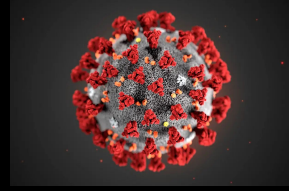
- Gather tweets and sort by relevance to COVID-19
  - Specific locations of interest
  - Specific terms of interest
    - Keywords associated with:
      - Prevention, Symptoms, etc.



SWB + MCCERT

The Process

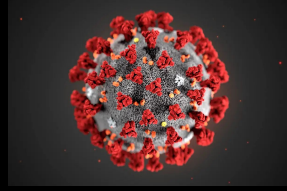
- Establish a flow of targeted tweets
  - Web scraping and data engineering



SWB + MCCERT

The Process

- Establish methods for predicting the relevance of Tweets for emergency response
  - Data wrangling, conditional statements
  - Natural Language Processing, Pre-processing
  - Modeling Approaches: Supervised, Unsupervised



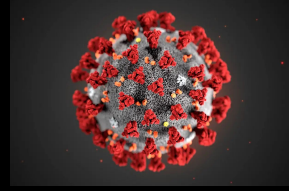
SWB + MCCERT

Tweet

This stresses me out. Why? Because my clinic has a shortage of supplies as well. We also had to lock up supplies because people are stealing them. PPE like gloves & masks are vital! #seattlecovid19

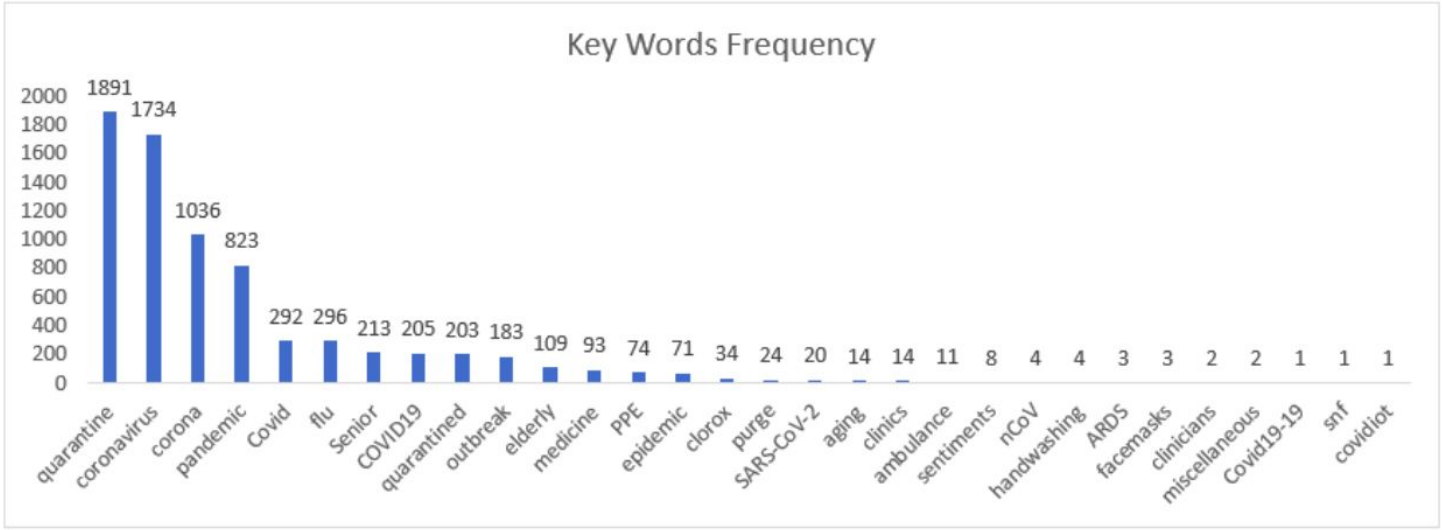
> · Mar 13
>
> A hospital in Seattle area has sent out a note to staff, shared with me, suspending elective surgery and warning that "our local COVID-19 trajectory is likely to be similar to that of Northern Italy." The hospital is down to a four-day supply of gloves.

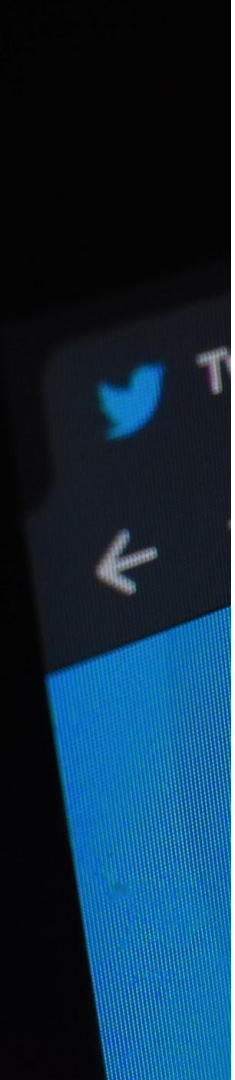
6:46 PM · Mar 13, 2020 · Twitter for iPhone

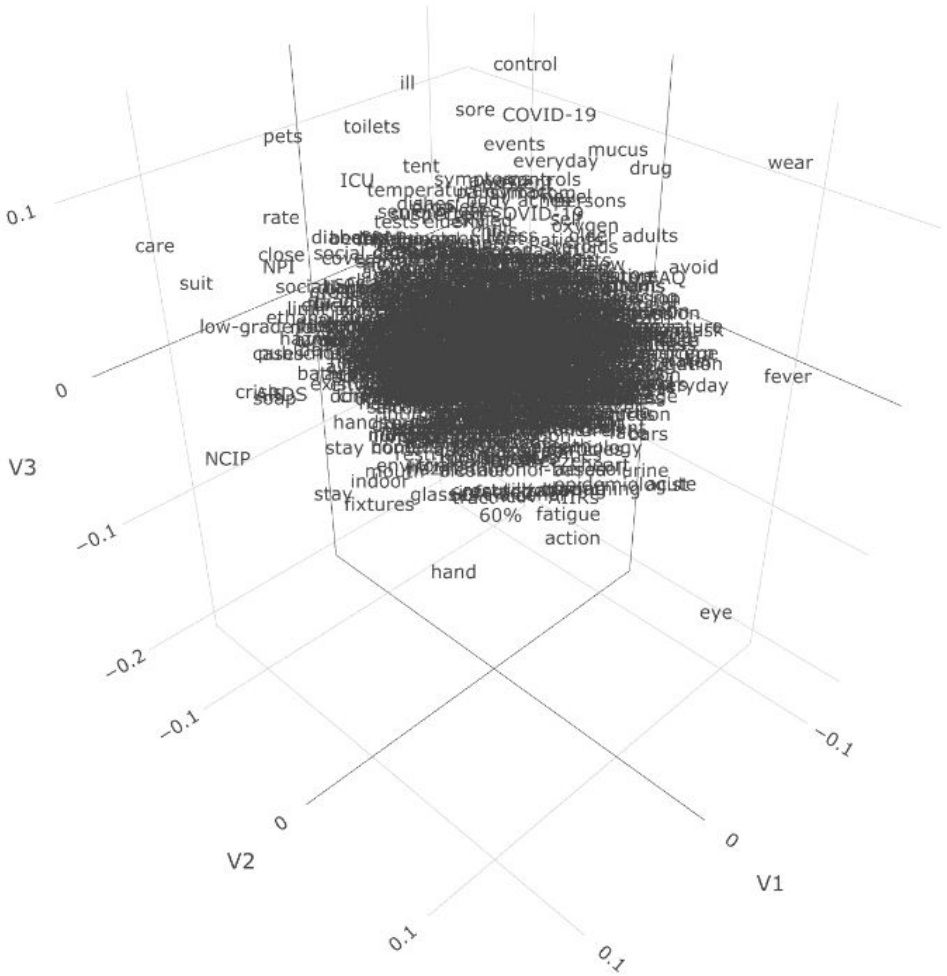


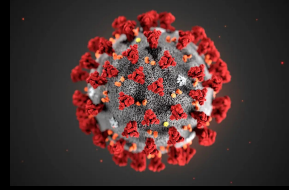
SWB + MCCERT

Case 1: Exact Matches



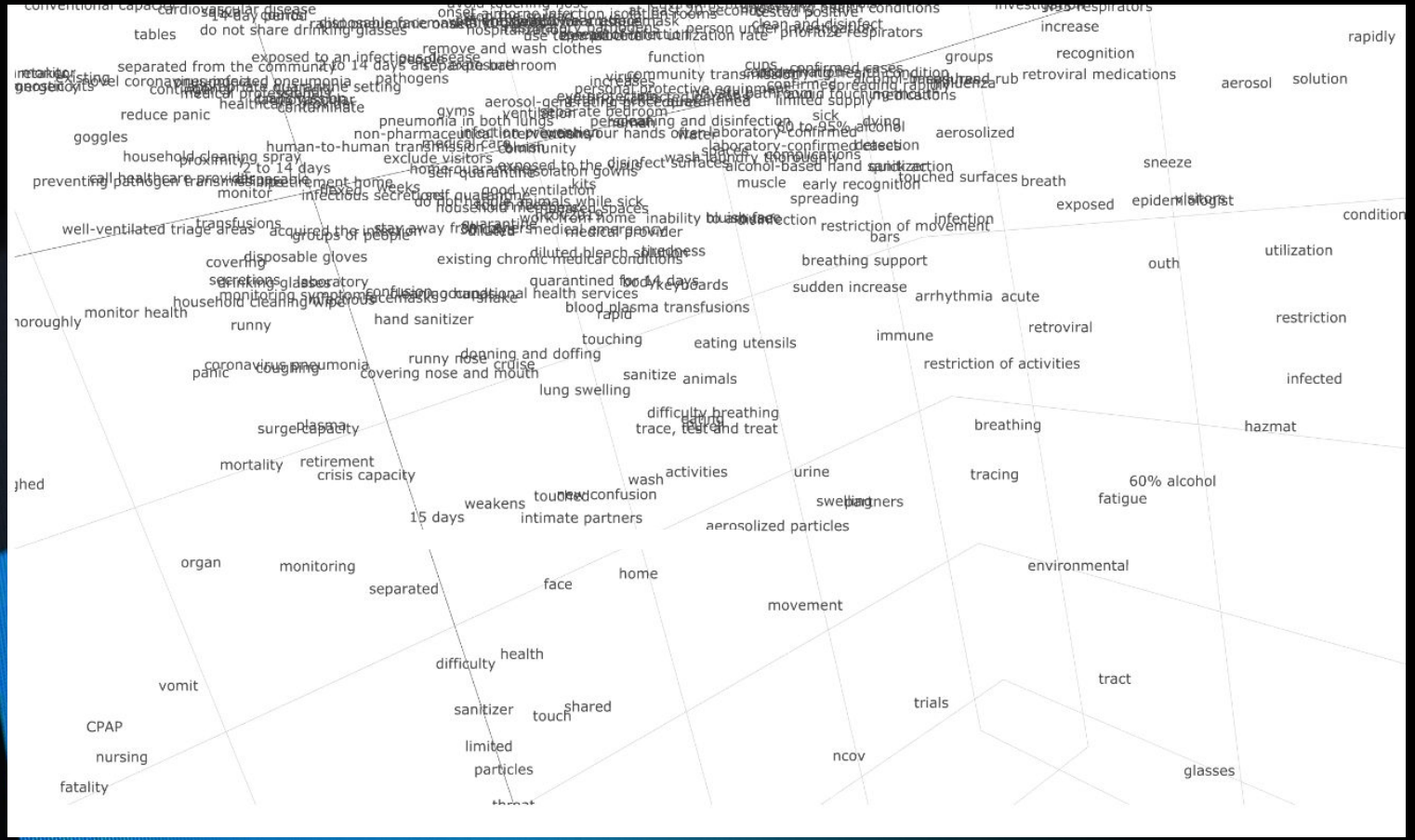


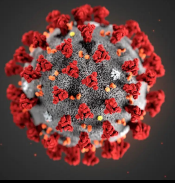




SWB + MCCERT

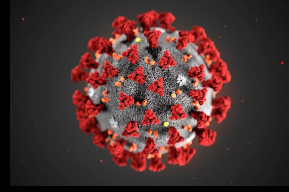
COMMUNITY EMERGENCY RESPONSE TEAM





SWB + MCCERT

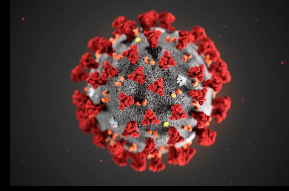
COMMUNITY EMERGENCY RESPONSE TEAM



SWB + MCCERT

Example of Methods and Models Applied

- Word Embeddings: TF-IDF, Word2vec, GLOVE, fastText
- Unsupervised Learning: K-Means Clustering, DBSCAN, Latent Dirichlet Allocation (Topic Modeling)
- Upsampling the minority class: SMOTE
- Transfer Learning: MERS → COVID-19
- Supervised Learning:
  - Naive Bayes, Logistic Regression, GLMNET, Support Vector Machines, ULMFIT, and XGBOOST,



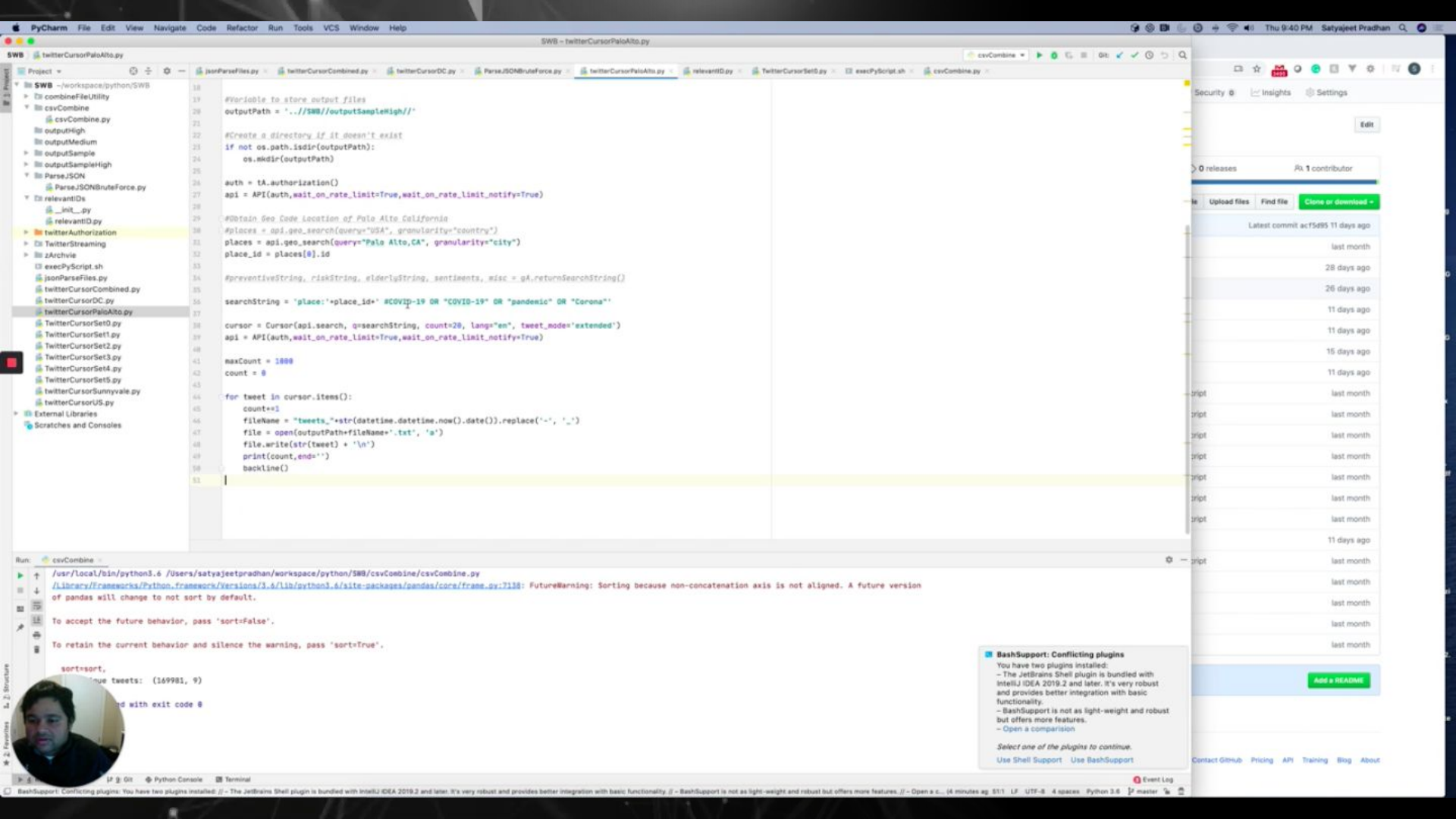
SWB + MCCERT

Deliverables to MMCERT

- Data acquisition pipeline
- Text preprocessing scripts
- Auditable model pipelines
- A collection of tweets over the course of the beginning of widespread awareness of the COVID-19 epidemic with emergency response relevance predictions

Data Science For Social Good

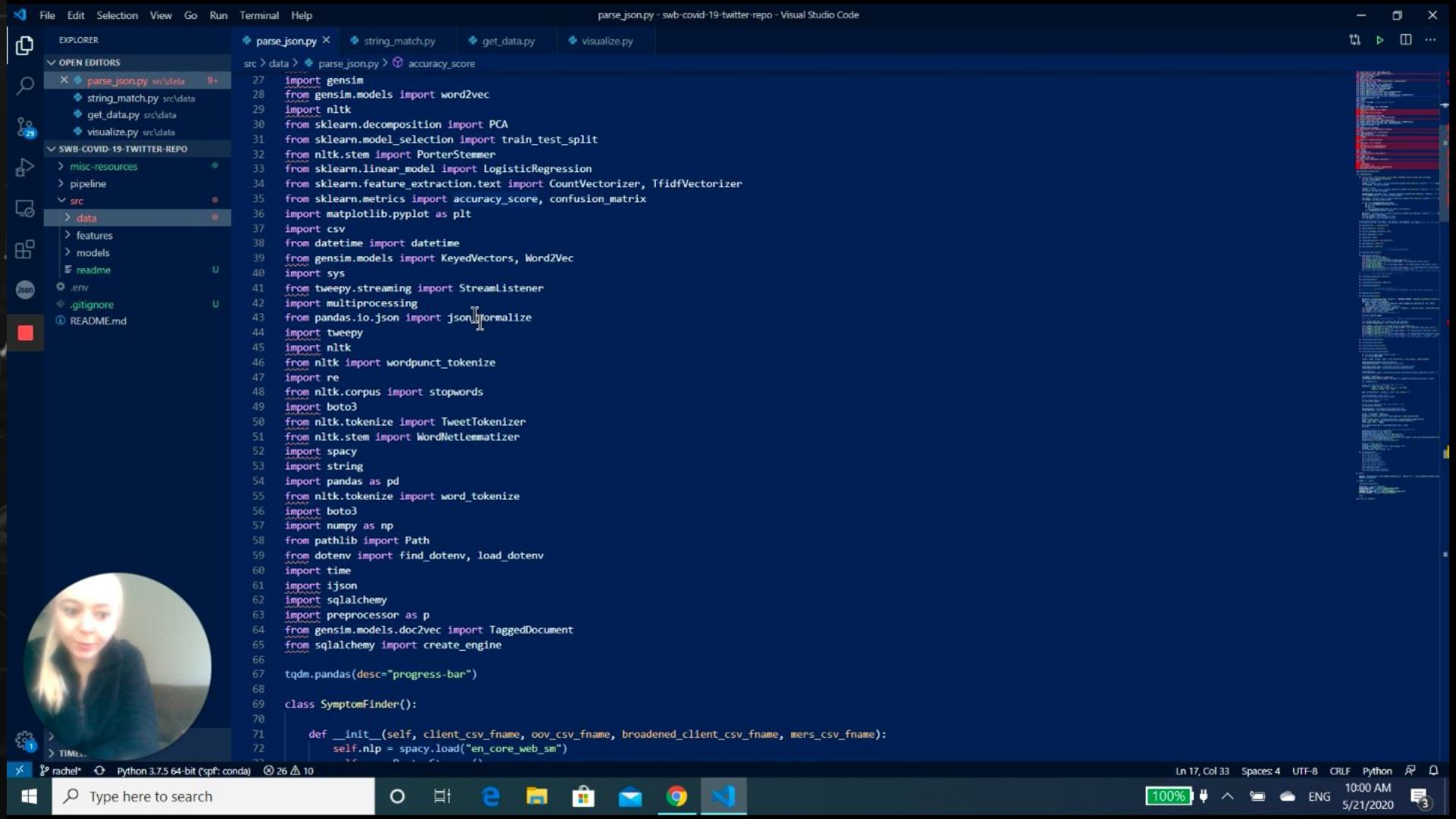
Wrangling
Visualizing
Modeling
Explaining
Communicating


PyCharm File Edit View Navigate Code Refactor Run Tools VCS Window Help
SWB – twitterCursorPaloAlto.py
Project
SWB ~/workspace/python/SWB
combineFileUtility
csvCombine
csvCombine.py
outputHigh
outputMedium
outputSample
outputSampleHigh
ParseJSON
ParseJSONBruteForce.py
relevantIDs
__init__.py
relevantID.py
twitterAuthorization
TwitterStreaming
zArchvie
execPyScript.sh
jsonParseFiles.py
twitterCursorCombined.py
twitterCursorDC.py
twitterCursorPaloAlto.py
TwitterCursorSet0.py
TwitterCursorSet1.py
TwitterCursorSet2.py
TwitterCursorSet3.py
TwitterCursorSet4.py
TwitterCursorSet5.py
twitterCursorSunnyvale.py
twitterCursorUS.py
External Libraries
Scratches and Consoles
#Variable to store output files
outputPath = '..//SWB//outputSampleHigh//'
#Create a directory if it doesn't exist
if not os.path.isdir(outputPath):
os.mkdir(outputPath)
auth = tA.authorization()
api = API(auth,wait_on_rate_limit=True,wait_on_rate_limit_notify=True)
#Obtain Geo Code Location of Palo Alto California
#places = api.geo_search(query="USA", granularity="country")
places = api.geo_search(query="Palo Alto,CA", granularity="city")
place_id = places[0].id
#preventiveString, riskString, elderlyString, sentiments, misc = gA.returnSearchString()
searchString = 'place:'+place_id+' #COVID-19 OR "COVID-19" OR "pandemic" OR "Corona"'
cursor = Cursor(api.search, q=searchString, count=20, lang="en", tweet_mode='extended')
api = API(auth,wait_on_rate_limit=True,wait_on_rate_limit_notify=True)
maxCount = 1000
count = 0
for tweet in cursor.items():
count+=1
fileName = "tweets_"+str(datetime.datetime.now().date()).replace('-', '_')
file = open(outputPath+fileName+'.txt', 'a')
file.write(str(tweet) + '\n')
print(count,end='')
backline()
Run: csvCombine
/usr/local/bin/python3.6 /Users/satyajeetpradhan/workspace/python/SWB/csvCombine/csvCombine.py
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-packages/pandas/core/frame.py:7138: FutureWarning: Sorting because non-concatenation axis is not aligned. A future version
of pandas will change to not sort by default.
To accept the future behavior, pass 'sort=False'.
To retain the current behavior and silence the warning, pass 'sort=True'.
sort=sort,
(169981, 9)
with exit code 0
BashSupport: Conflicting plugins
You have two plugins installed:
– The JetBrains Shell plugin is bundled with IntelliJ IDEA 2019.2 and later. It's very robust and provides better integration with basic functionality.
– BashSupport is not as light-weight and robust but offers more features.
– Open a comparison
Select one of the plugins to continue.
Use Shell Support Use BashSupport
Python Console
Terminal
Event Log
Security
Insights
Settings
Edit
releases
contributor
Upload files
Find file
Clone or download
Latest commit acf5d95 11 days ago
last month
28 days ago
26 days ago
11 days ago
11 days ago
15 days ago
11 days ago
Add a README
Contact GitHub
Pricing
API
Training
Blog
About

```
import pandas as pd

from pandas.io.json import json_normalize
from sklearn.metrics.pairwise import linear_kernel

pd.set_option('display.max_colwidth', -1)
pd.set_option('display.float_format', lambda x: '%f' % x)
```

In [526]:

```
# Normalize JSON data to get the original urls
def normalize_url(data):

    import datetime

    print('normalize_url Start:', datetime.datetime.now())

    data['Ext URL'] = data['Ext URL'].fillna('[]').map(lambda x: x.strip())

    norm_url = pd.DataFrame()

    for i, row in data.iterrows():
        ext_url = row['Ext URL']
        if ext_url != '[]':
            ext_url = ext_url.replace("'url': '",'"url": "')\
                             .replace("', 'expanded_url': '",'", "expanded_url": "') \
                             .replace("', 'expanded_url':",'", "expanded_url":') \
                             .replace("', 'display_url': '",'", "display_url": "') \
                             .replace(", 'display_url': '",', "display_url": "') \
                             .replace(", 'display_url': ",', "display_url": ') \
                             .replace("', 'indices'",'", "indices"') \
                             .replace("'indices'",'"indices"')
            new_row = pd.DataFrame(json_normalize(json.loads(ext_url)))
            new_row['TweetID'] = row['TweetID']
            norm_url = norm_url.append(new_row, sort=False)

    orm_url['indices']
    _url.drop_duplicates(inplace=True)

    print('normalize_url End:', datetime.datetime.now())
```


File Edit Selection View Go Run Terminal Help
parse_json.py - swb-covid-19-twitter-repo - Visual Studio Code
EXPLORER
OPEN EDITORS
parse_json.py src\data 9+
string_match.py src\data
get_data.py src\data
visualize.py src\data
SWB-COVID-19-TWITTER-REPO
misc-resources
pipeline
src
data
features
models
readme U
.env
.gitignore U
README.md
parse_json.py
string_match.py
get_data.py
visualize.py
src > data > parse_json.py > accuracy_score
import gensim
from gensim.models import word2vec
import nltk
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from nltk.stem import PorterStemmer
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
import csv
from datetime import datetime
from gensim.models import KeyedVectors, Word2Vec
import sys
from tweepy.streaming import StreamListener
import multiprocessing
from pandas.io.json import json_normalize
import tweepy
import nltk
from nltk import wordpunct_tokenize
import re
from nltk.corpus import stopwords
import boto3
from nltk.tokenize import TweetTokenizer
from nltk.stem import WordNetLemmatizer
import spacy
import string
import pandas as pd
from nltk.tokenize import word_tokenize
import boto3
import numpy as np
from pathlib import Path
from dotenv import find_dotenv, load_dotenv
import time
import ijson
import sqlalchemy
import preprocessor as p
from gensim.models.doc2vec import TaggedDocument
from sqlalchemy import create_engine
tqdm.pandas(desc="progress-bar")
class SymptomFinder():
def __init__(self, client_csv_fname, oov_csv_fname, broadened_client_csv_fname, mers_csv_fname):
self.nlp = spacy.load("en_core_web_sm")
rachel* Python 3.7.5 64-bit ('spf': conda) 26 10
Ln 17, Col 33 Spaces: 4 UTF-8 CRLF Python
Type here to search
100%
ENG
10:00 AM
5/21/2020

colab.research.google.com/drive/1IJZt6YUkn5k_c5nT2kd-NAJPwS-7SV1Q#scrollTo=BBBh1ucBGfEd

Bookmarks | TA-BigQuery | TA-GCP Bucket | Investopedia - Educ... | Brain Pickings | Learn to code | Cod... | Aerin Kim – Tow... | dbamman/anlp19:... | Other bookmarks

COVID_19_Tweets_Categorization_ULMFit.ipynb

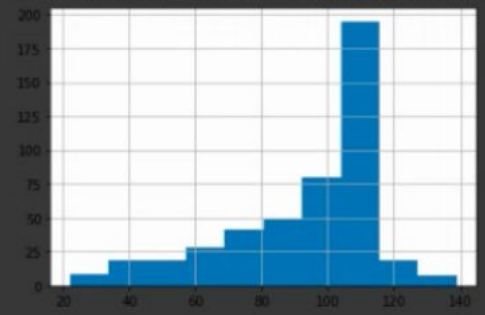
File Edit View Insert Runtime Tools Help All changes saved

Comment Share H

Files

Connecting to a runtime to enable file browsing.

+ Code + Text Reconnect Editing

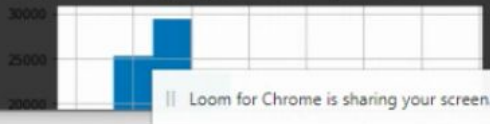



```
# checking for shortened tweet_text
dat_[dat_.preprocessed_text.apply(lambda x: '\.\.' in x)]
```

	id	FINAL_LABEL	status_obj	text	hashtags	url	expanded_url	label_mod	preprocessed_text

```
palo_tweets.Processed_Text.apply(len).hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f90b8a91a58>
```



Loom for Chrome is sharing your screen. Stop sharing Hide

100%

Type here to search

ENG INTL 12:19 AM 5/30/2020

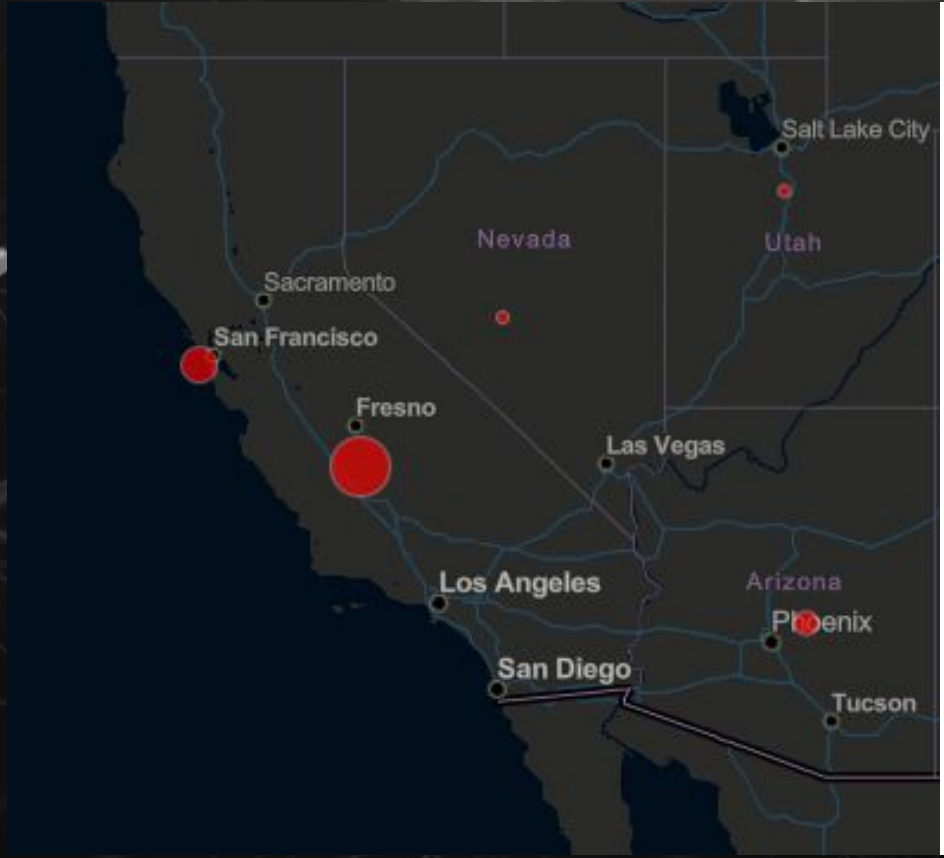
"Not only does data wrangling consume most of an analyst's workday, it also represents much of the analyst's professional process: it captures activities like understanding what data is available; **choosing** what data to use and at what level of detail; **understanding** how to meaningfully **combine(ing)** multiple sources of data; and **deciding** how to **distill(ing)** the results to a size and shape that can **drive(ing)** downstream analysis."

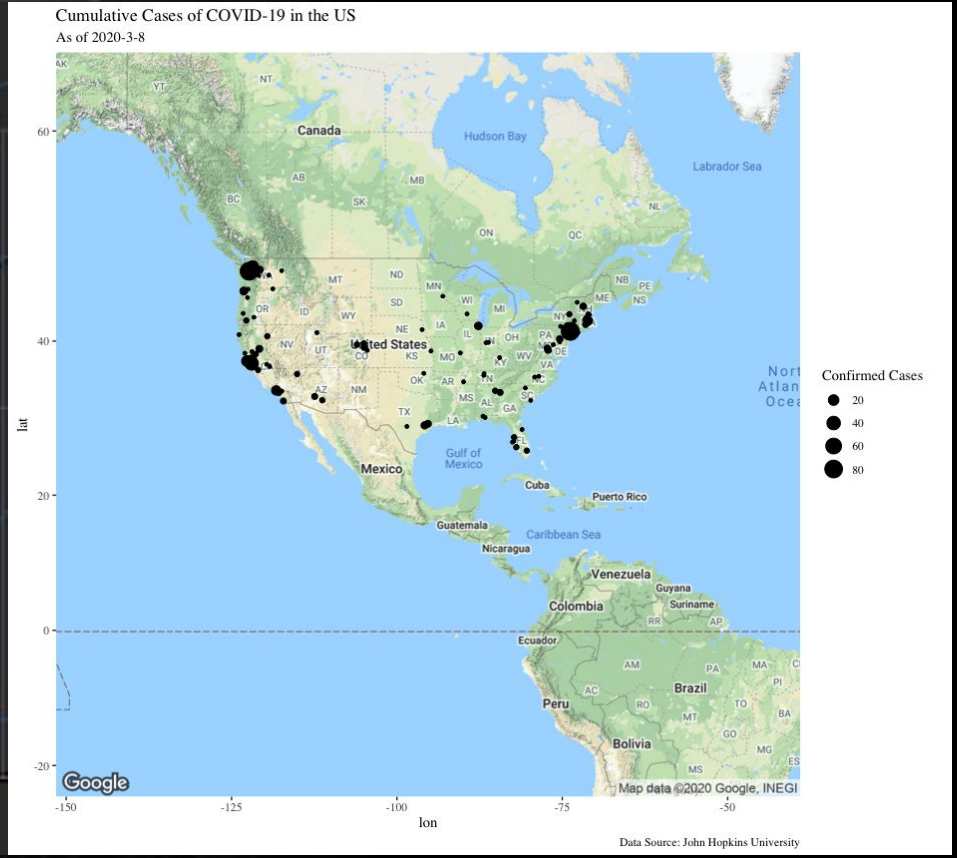
- Principles of Data Wrangling
  Rattenbury et al. (2017)

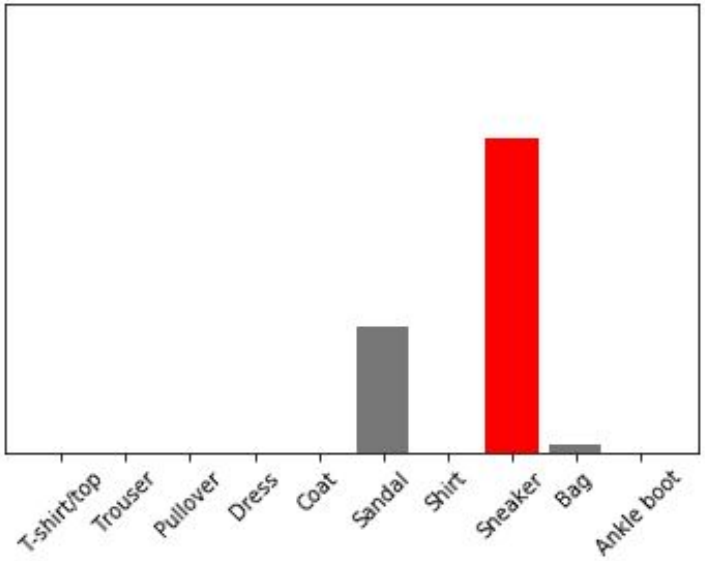
"Not only does data wrangling consume most of an analyst's workday, it also represents much of the analyst's professional process: it captures activities like understanding what data is available; choosing what data to use and at what level of detail; understanding how to meaningfully combine multiple sources of data; and deciding how to distill the results to a size and shape that can drive downstream analysis."

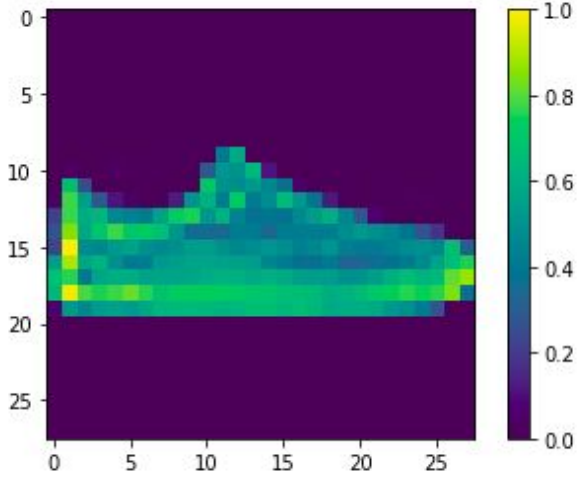
- Principles of Data Wrangling
  Rattenbury et al. (2017)

JHU Public Dashboard vs. Using JHU Data (2020-3-8)







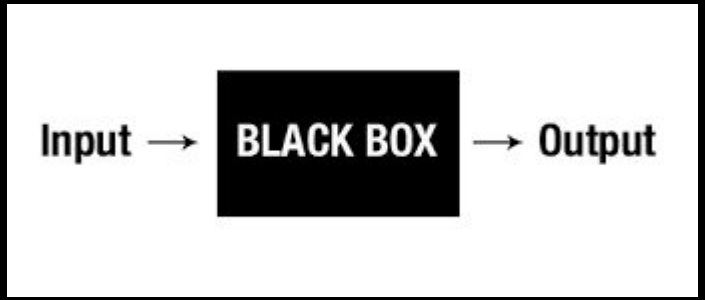


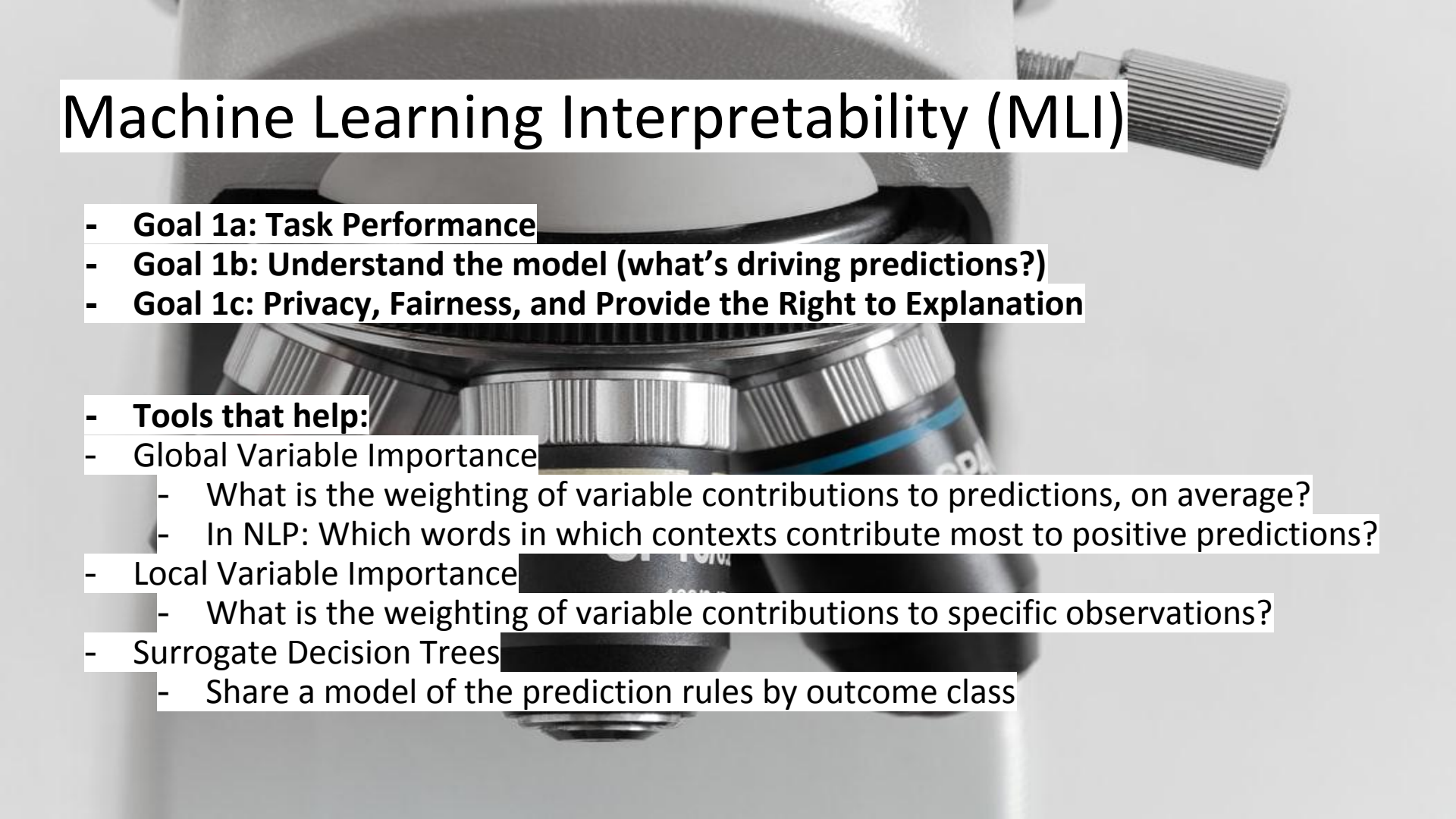
kinestry.io

ML/AI predictions are probabilistic.

Machine Learning Interpretability (MLI)

More complex models take more work to explain, but **may no longer be** “black boxes”.





Machine Learning Interpretability (MLI)

- **Goal 1a: Task Performance**
- **Goal 1b: Understand the model (what's driving predictions?)**
- **Goal 1c: Privacy, Fairness, and Provide the Right to Explanation**

- **Tools that help:**
- Global Variable Importance
    - What is the weighting of variable contributions to predictions, on average?
    - In NLP: Which words in which contexts contribute most to positive predictions?
- Local Variable Importance
    - What is the weighting of variable contributions to specific observations?
- Surrogate Decision Trees
    - Share a model of the prediction rules by outcome class

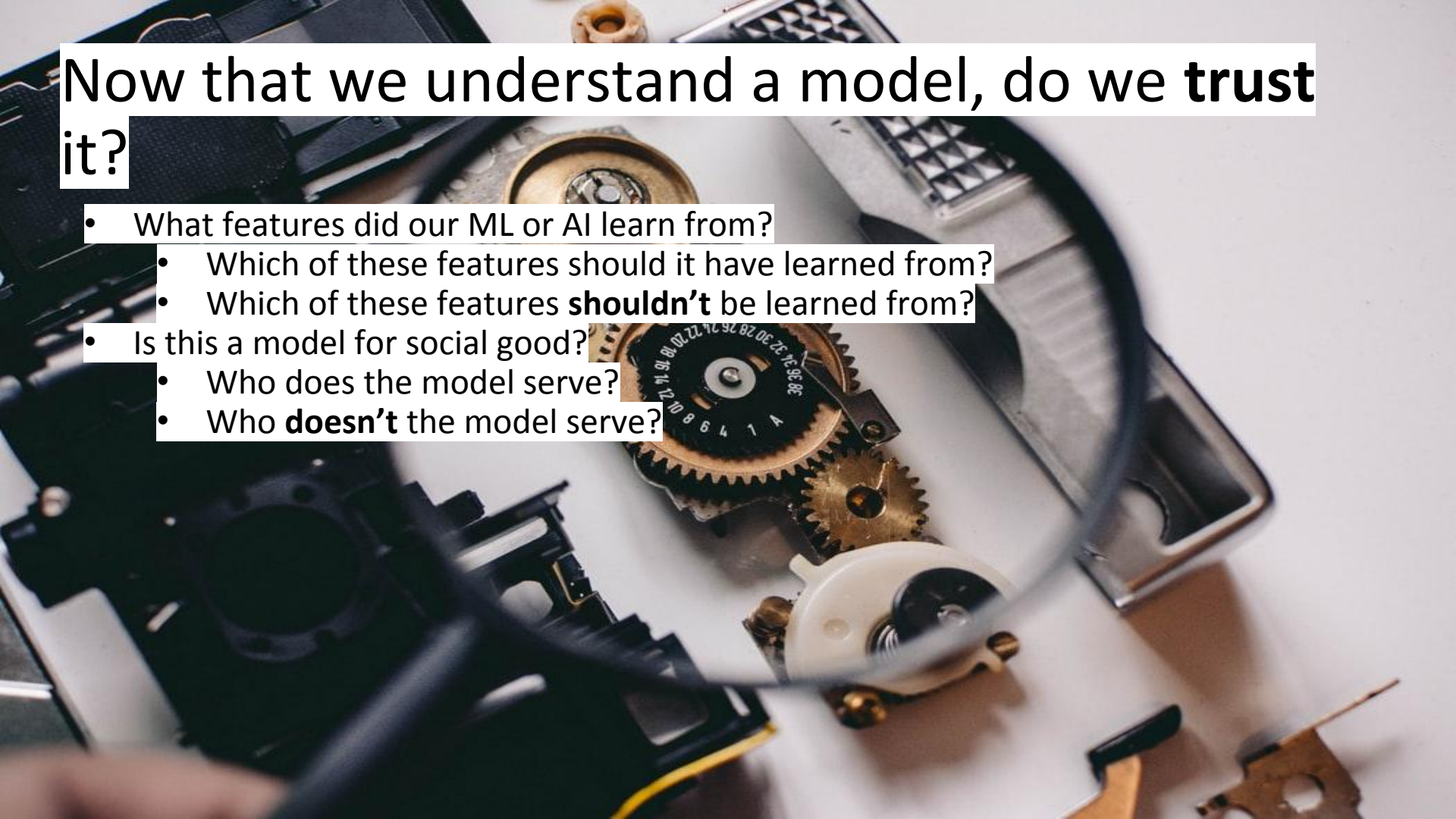
Machine Learning Interpretability (MLI)

- **Sensitivity Analysis**
  - Vary the inputs; make small changes
  - How does this influence predictions?
  - What small changes would “push observations (or people) over the threshold”?
  - This may inform subsequent iterations in data collection

Fairness

- "...unfairness and discrimination are pervasive when decisions are being made by humans, which, unfortunately, are not automatically solved, and can even be amplified, when machines are put in control." - Zhang and Bareinboim (2017)
- "Fairness in machine learning is an emerging topic with the overarching aim to critically assess algorithms (predictive and classification models) whether their results reinforce existing social biases." - Kozodoi and Varga (2020)

- **General Approaches**
- Disparate Impact Analysis
  - ex) Accuracy Parity… [performance metric] by group relative to the reference group
- Root Cause Analysis
  - Do we know whether **protected features** influenced the prediction?



Now that we understand a model, do we **trust** it?

- What features did our ML or AI learn from?
  - Which of these features should it have learned from?
  - Which of these features **shouldn't** be learned from?
- Is this a model for social good?
  - Who does the model serve?
  - Who **doesn't** the model serve?

Ultimately, our goal is to do data science for social good.

This is why solving problems in ways that we can explain and trust is essential.

SWB Team

Julia Reid, PCM
Keri Wheatley
Heli Vora
Satyajeet Pradhan
Rachel Doehr
Qingyuan Wang
Harshit Sharma
Lena Lickteig, DQA