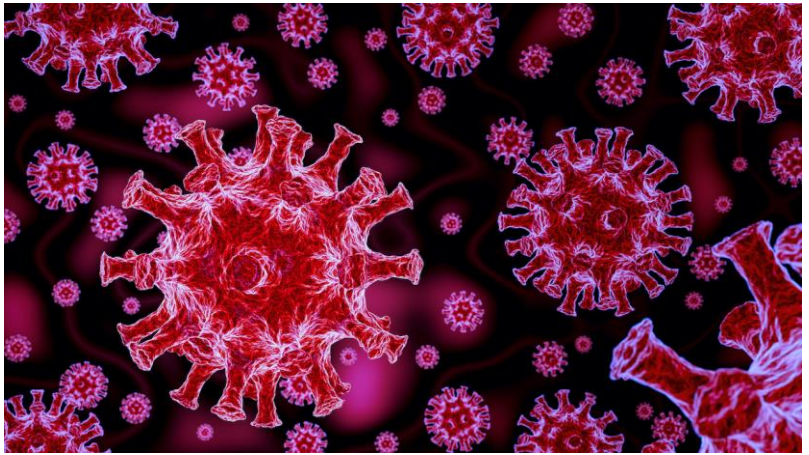
# D3M Research Project

Student: Hshmat Sahak
Supervisor: Prof. Scott Sanner

UNIVERSITY OF TORONTO

# Project Description

- Twitter is a great sensor of human attitudes towards worldwide events, including the pandemic.
- Tracking Twitter behaviour is useful as it provides a huge corpus of data, it is an always-on data source and is nonreactive
- My project over the summer was investigating the usefulness of Twitter as an indicator of Covid metrics- specifically cases & deaths, as well as general sentiments and how they evolved throughout the pandemic.
- Essentially: what can we measure on Twitter that seems to be indicative of COVID waves?



# Understanding the Data

tweet.txt

- id: the tweets id
- author_id: the id of the user who posted the tweet
- created_at: creation date of the tweet
- text: the content of the tweet
- public_metrics: public metrics like retweet count, reply count, like count, and quote count
- entities: entities mentioned in the text including mentions and hashtags
- geo: which gives the id of the place that can be retrieved from 'places.txt'.

users.txt

- id: id of the user
- username:
- name: the name that appears on the profile.]
- entities: entities mentioned such as a url
- created_at: profile creation date
- description: description given by the user
- location: location of the user
- public_metrics: public metrics such as followers/following count, tweet count, and listed count.

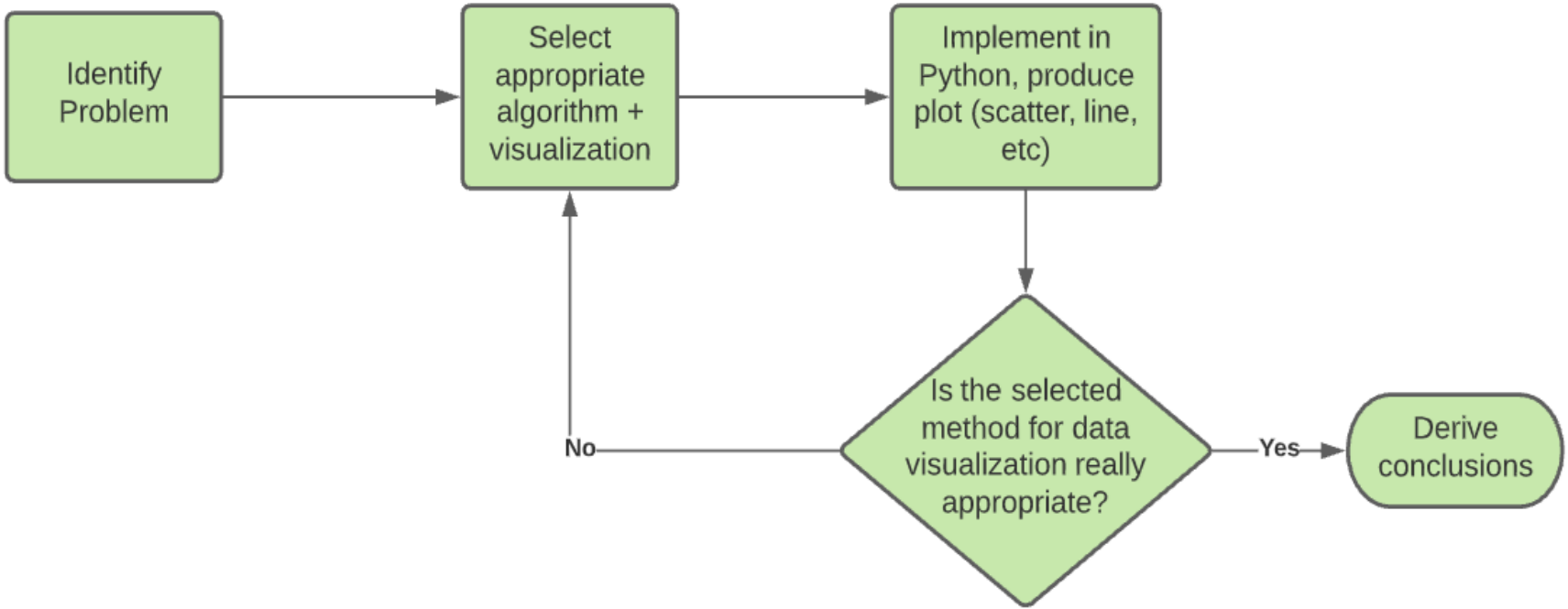
places.txt

- id: id of the place
- full_name: name of the place
- country: name of the coutry of the place (in french for unknown reasons)
- country_code: code of the country

Note: we only look at US tweets that have valid geolocation attribute

# Exploratory Data Analysis

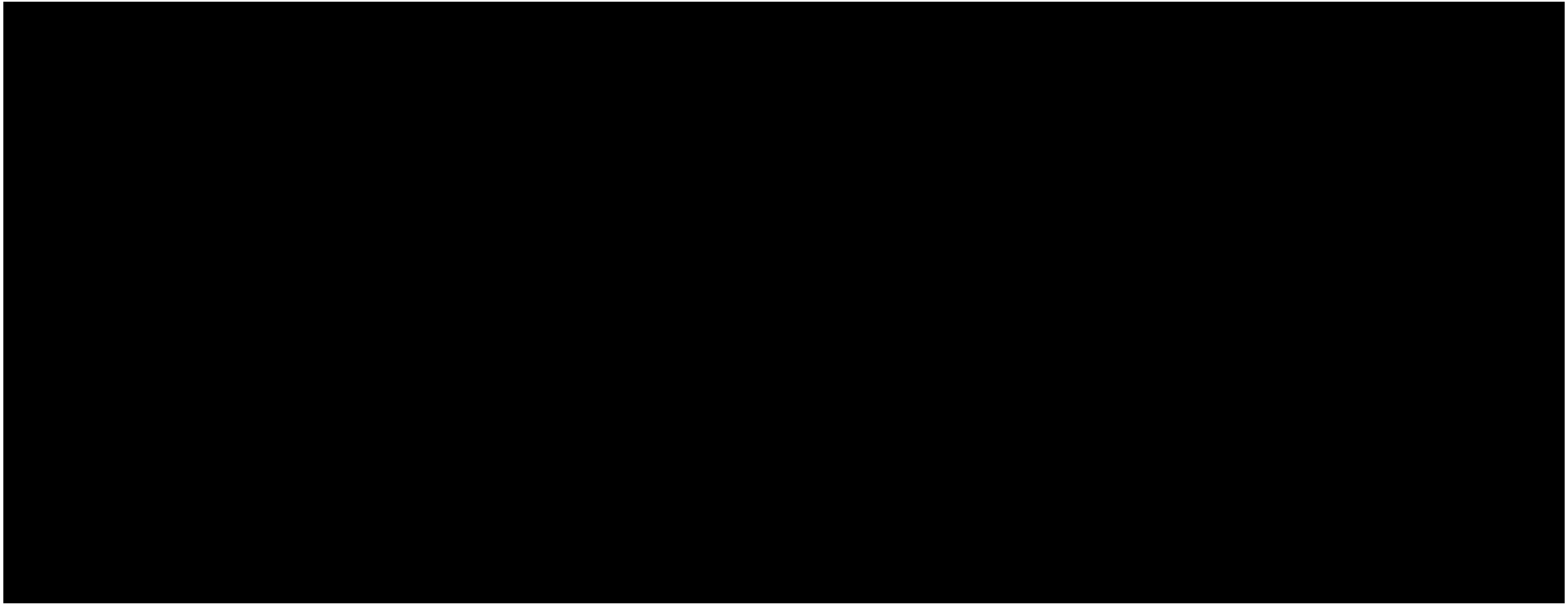
- Majority of time was spent here



# Tweet Distribution over Space and Time

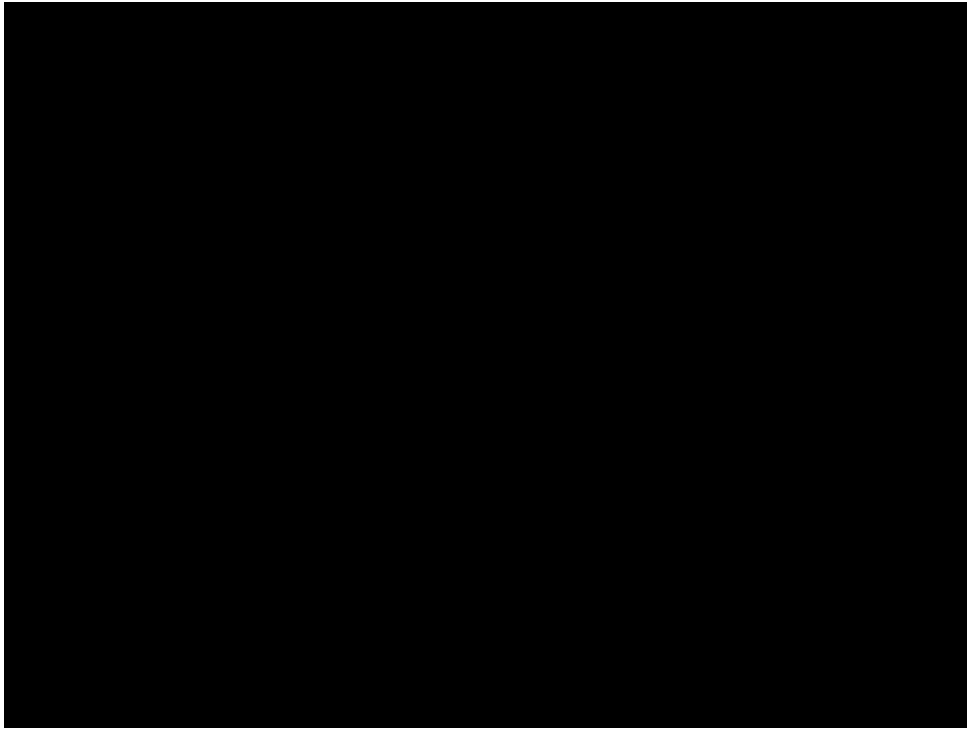
Objective: Find out how tweet count varies with time over the US, ask if that matches general trends in covid cases/deaths
Approach: Heatmap with timeslider. Each frame is one week
Result: Good for visualization, but want more concrete divisions of area. Grid would address box nature of geolocation

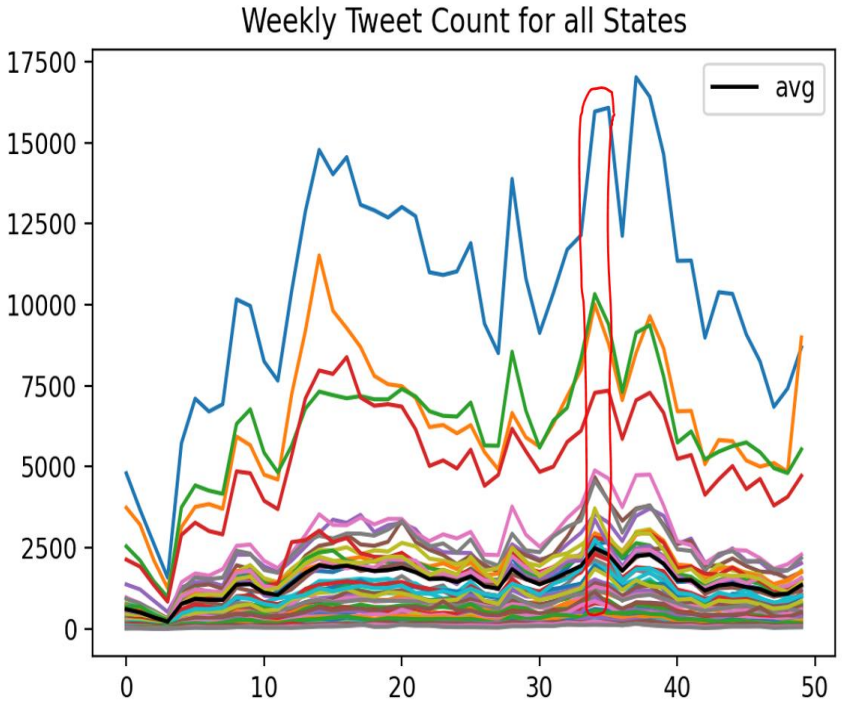
# Tweet Distribution over Space and Time



Objective: Find out how tweet count varies with time over the US, ask if that matches general trends in covid cases/deaths
Approach: Divide US as rectangular grid of arbitrary size; display time slider grid map where each frame is one week
Result: Need means of normalization. We will discuss this shortly. Also, switch from grid to choropleth.

# Tweet Count over time in each State





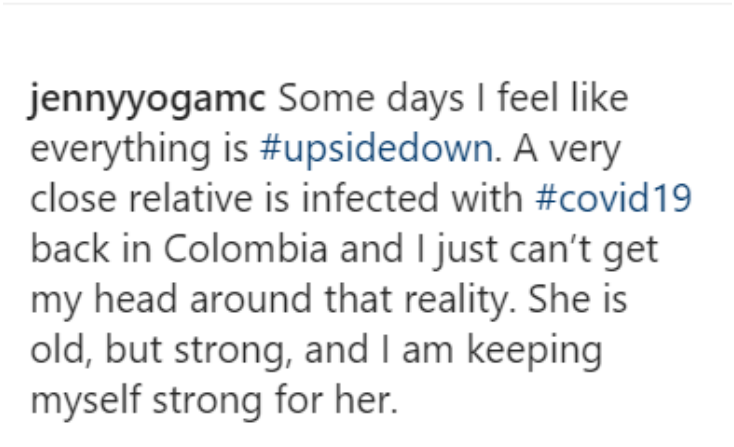

Objective: Find out how tweet count varies with time over each state
Approach: Choropleth map with time slider. Each frame is one week
Result: Peak in number of tweets coincides with election week. So filter out election-related tweets by searching for:
Trump|Biden|Election|democratic|republican|party|President|campaign|elector|candidate

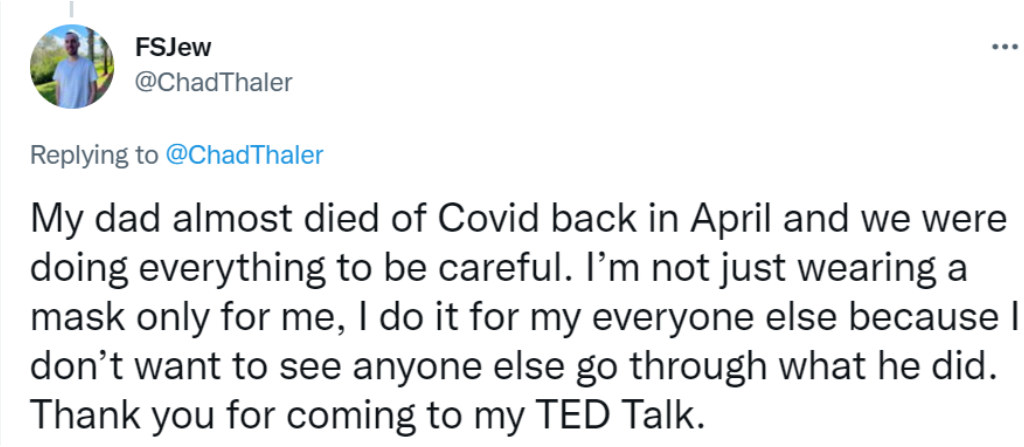
# Identifying Covid-related Tweets

Filtering tweets by election-related keywords helped get rid of a lot of unrelated tweets. These were included in the original dataset by nature of how twitter users use common hashtags or the strong tendency to politicize opinions about mask and vaccine attitude.

We can certainly do better. How do we extract tweets that refer directly to Covid19?

An important observation is that almost all tweets that refer to family/relative/etc talk about a loved one who is directly affected by covid! Another is that people who are mentally affected by covid policies like lockdown tend to tweet in a more “depressing” tone.



# Hashtag Lists for Sentiment Classification

Strategy:

Remove election-related tweets:

- Trump|Biden|Election|democratic|republican|party|President|campaign|elector|candidate

Then divide into:

Family-related Tweets

- sister|brother|mother|father|grandpa|grandma|grandparents|grandmother|grandfather|cousin|dad|mom

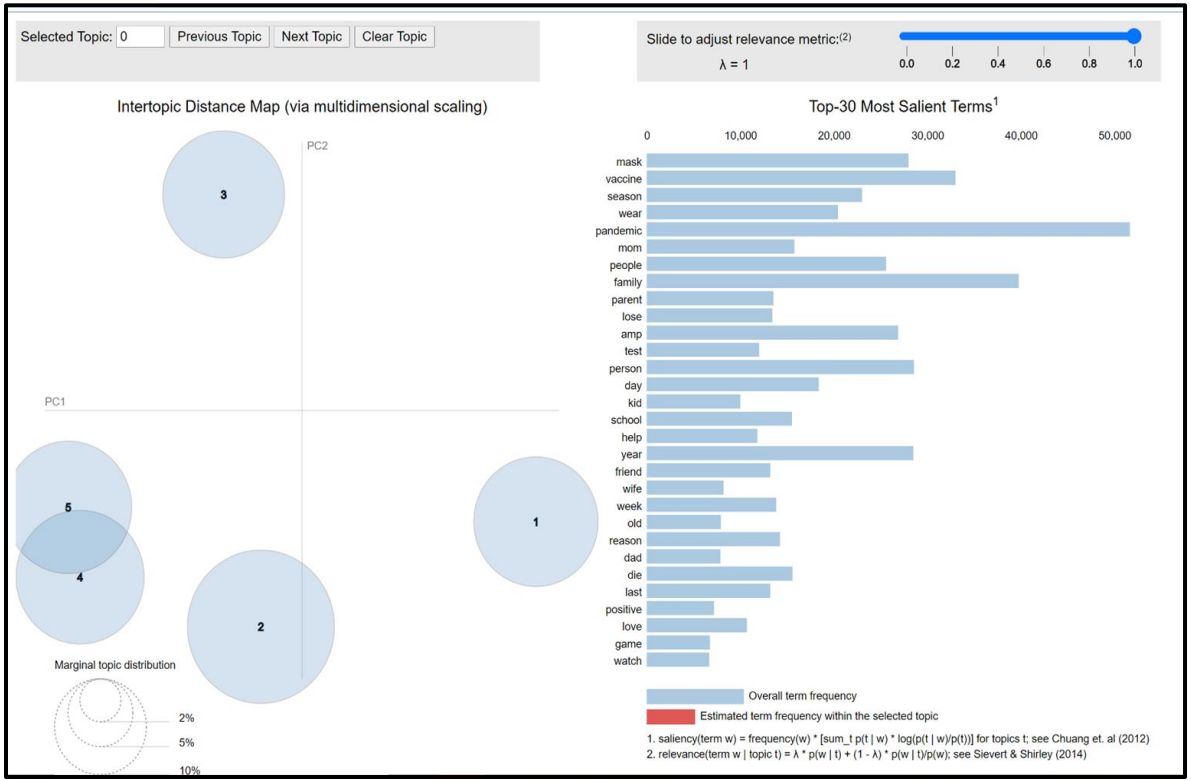
Depression-related Tweets

- overwhelmed|exhausted|distressed|anxiety|anxious|tired|low|depression|discouraged|desperate|demotivated|insomnia|cry|nervous|worried|lonely|sad|empty|suicide|antidepressant|hopeless

Strict Depression-related Tweets

- overwhelmed|exhausted|distressed|anxiety|anxious|depression|discouraged|demotivated|insomnia|lonely|empty|suicide|antidepressant|hopeless

# Family Dataframe: Topic Modelling

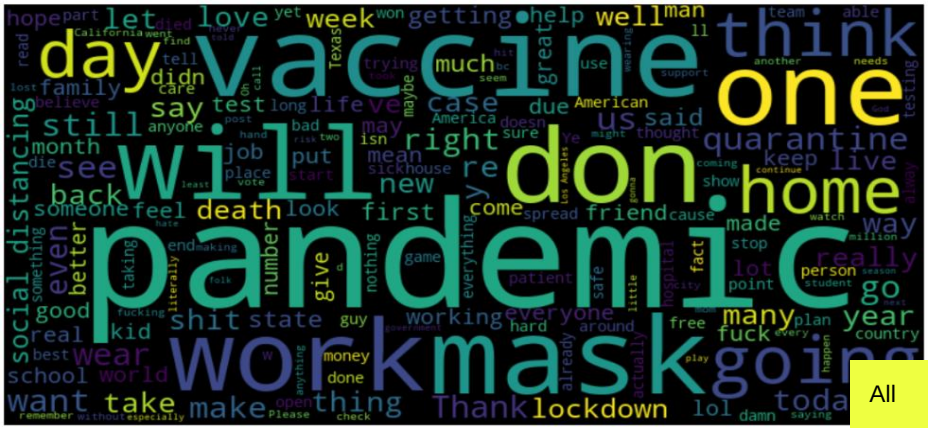


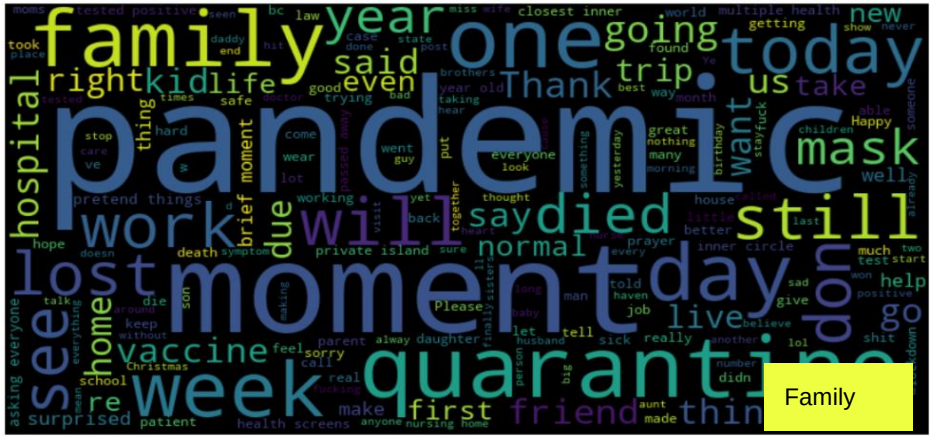

```
[(0,
  '0.045*"covid" + 0.030*"get" + 0.024*"mom" +
0.016*"go" + 0.013*"dad" + '
  '0.012*"day" + 0.011*"year" + 0.011*"die" +
0.011*"mother" + '
  '0.010*"brother"'),
 (1,
  '0.046*"pandemic" + 0.026*"moment" +
0.019*"need" + 0.018*"life" + '
  '0.016*"live" + 0.011*"start" + 0.008*"pass" +
0.008*"ago" + 0.007*"state" + '
  '0.006*"hit"'),
 (2,
  '0.030*"time" + 0.025*"week" + 0.025*"take" +
0.017*"thing" + 0.016*"back" + '
  '0.015*"come" + 0.015*"test" + 0.012*"mask"
+ 0.011*"last" + 0.011*"old"')]
```

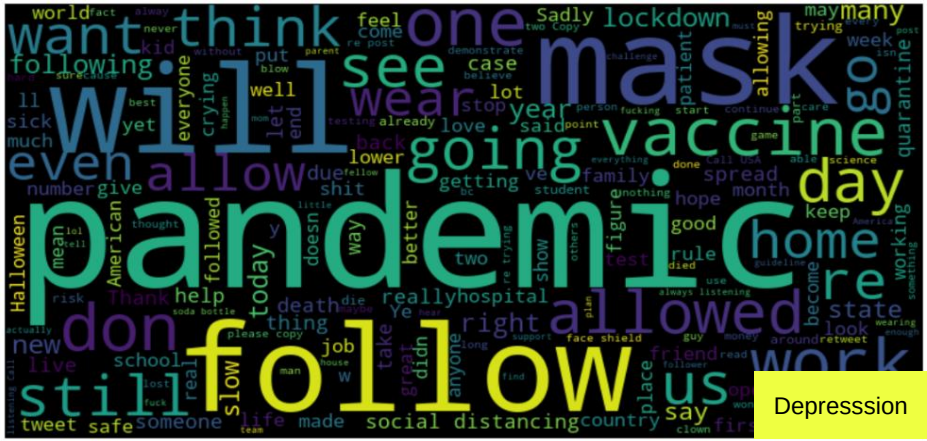
Objective: Verify whether family dataframe can serve as dataframe for tweets indicative of covid waves or policy
Approach: Topic Modelling using the Gensim Library
Result: All topics seems directly related to covid cases. This is exactly what we hoped for.

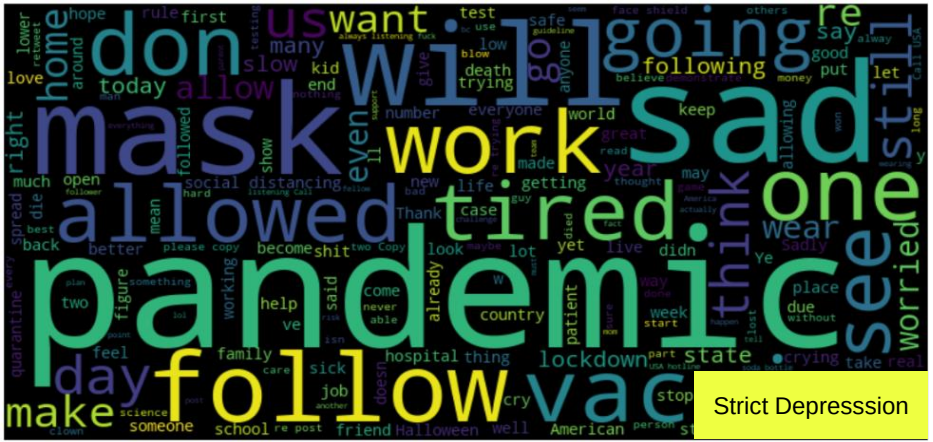
# Word Clouds for Different Datasets

Same Objective: determine if these datasets are a good way to filter for covid-specific tweets



# Family and Depression-related Tweets Choropleth

Family-related tweets

Depression-related tweets

Based on feedback from Prof. Sanner, family-related tweets did the best job of following Covid19 metrics.The initial hypothesis that mentions of relative were directly related to cases of covid is supported, but note that this dataset is not big

# Investigating Relationship between Mask Attitude and Covid Metrics

We choose to focus on masks. Goal is to use twitter to detect attitude towards masks (positive, negative, or neutral).

We need a way to detect pro or anti- mask sentiment.

We tried out 4 methods:

- Vader
- Hashtag
- Regex
- ML Model

**BoogieSnacks** ☕
@Boogie_Snacks

The amount of parties happening rn is absurd. Please be safe, wear a mask, and read one piece.

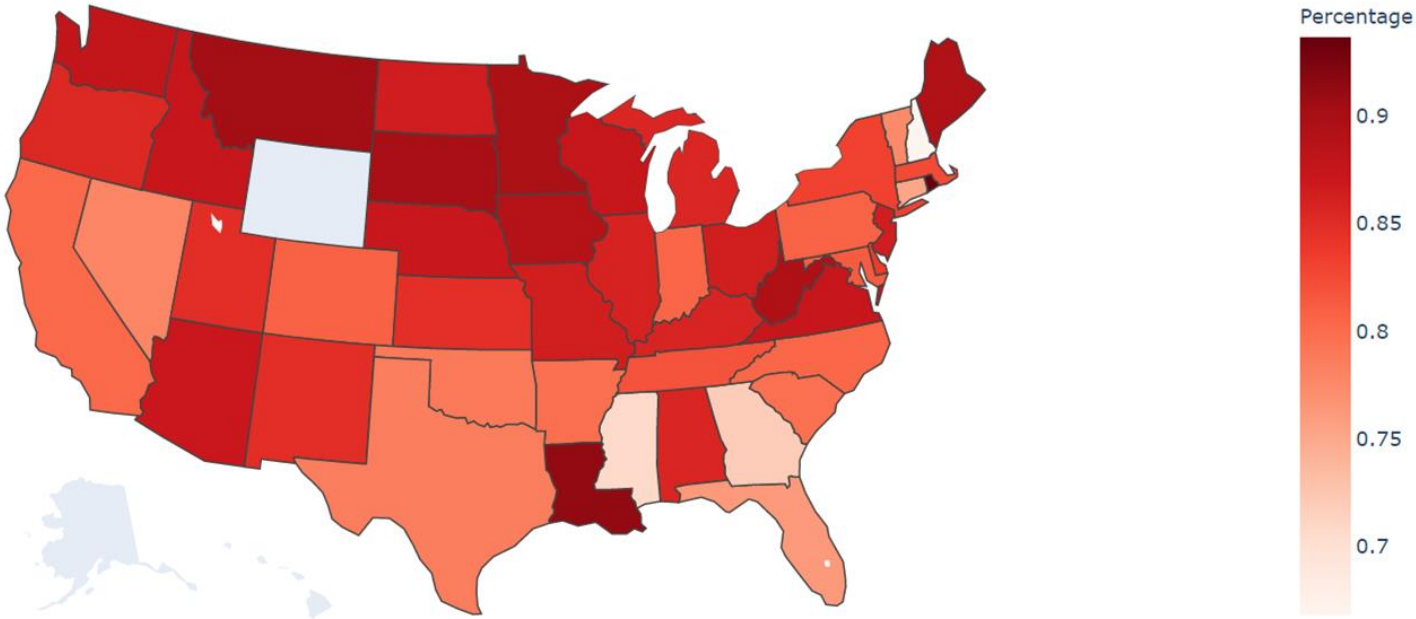
**Low-Rentz**
@YiddishSteel

Replying to @pdx_mavs and @TheCensoredRock

I don't wear a mask because I'm not an insufferable twat.

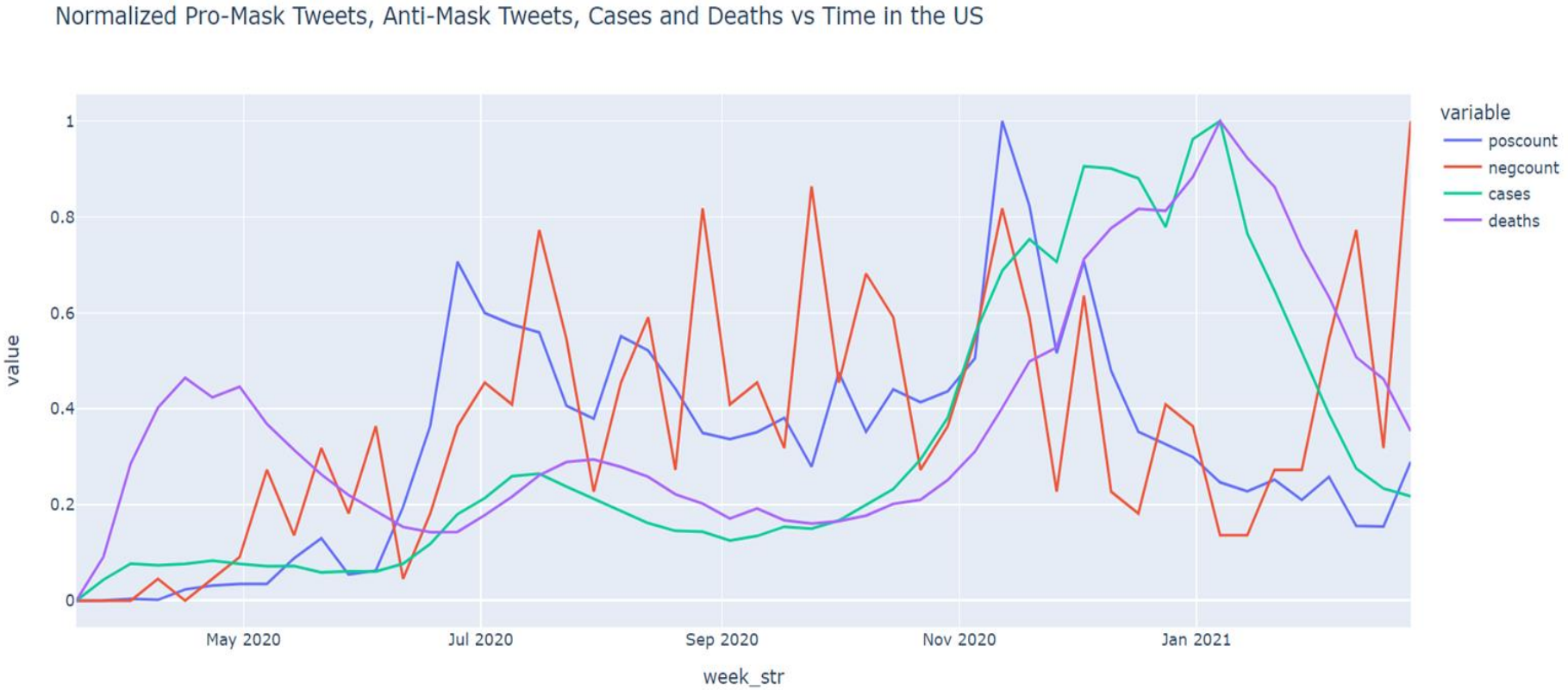
# Hashtag Results

Percentage of Tweets in Year that are Pro-Mask by State (excluding WY as it is highest by far)



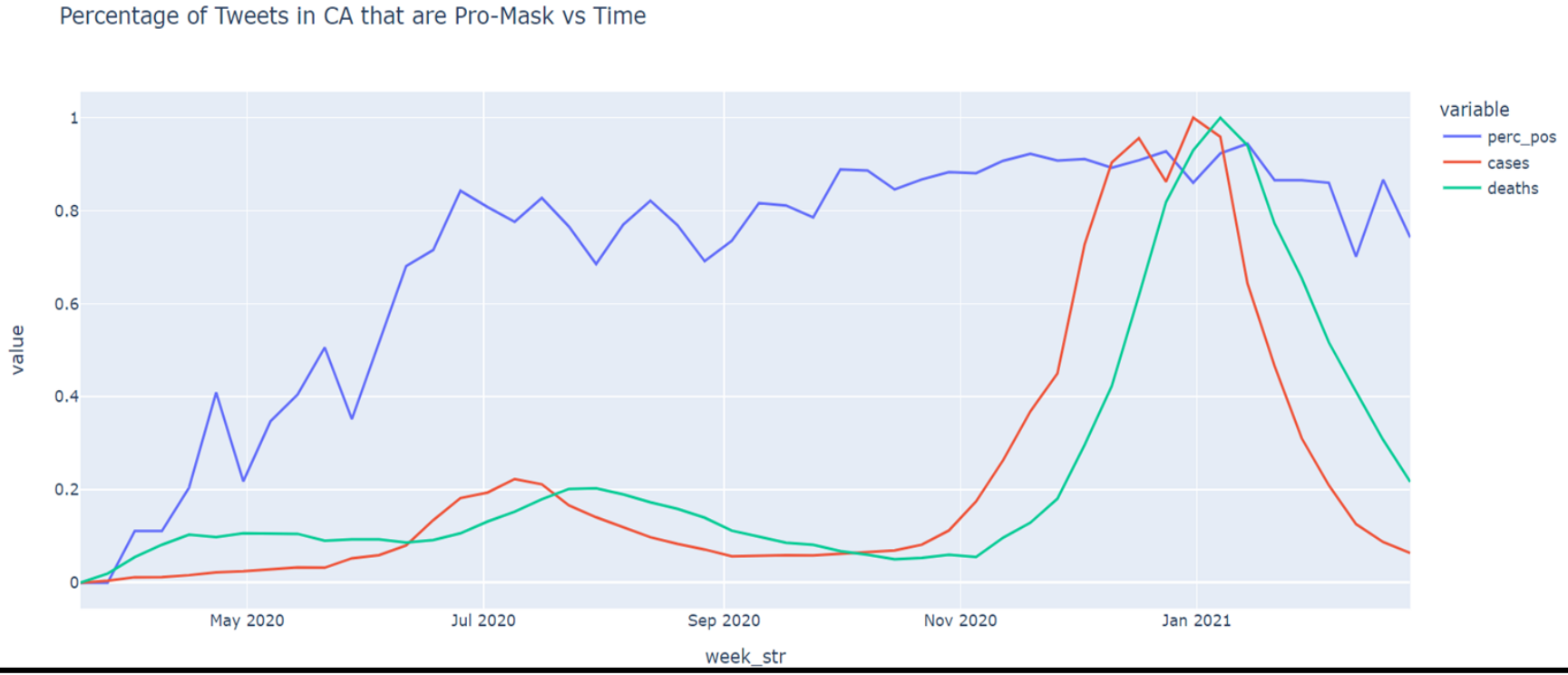

Objective: Visualize how different states vary in mask support, as well as compare # pro-mask vs # anti-mask
Approach: Choropleth, color intensity depends on fraction of total tweets that are positive
Result: Each state has more positive than negative, many states have much more positive than negative

# Hashtag Results (Continued)




Due to small number of tweets in dataframe, negative counts are not useful. However, positive tweets do loosely follow covid case count. We also see a time lag between cases and deaths, which is not surprising, but useful to validate data

# Hashtag Results (Continued)




Objective: Check if relationship between positive tweet counts and cases/deaths holds for each state.
Approach: Create individual line graph for all 50 states
Result: Some states follow the trend, others don't. At this stage, we aren't even confident that hashtag classifier is doing a good job, so no results were presented as conclusions

# Hashtag Results (Continued)


Objective: Is there a strong relation between positive tweet count and number of cases that is state-independent?
Approach: Scatter plot
Result: Kind of (R^2=0.77), but population effects serve as confounder here. Higher population leads to both higher poscount and higher cases. We need to figure out a way to get rid of population effects.
NORMALIZATION APPROACH: Normalize cases by population, poscount by number of twitter users in the state.

# Additional Commentary on Normalization

It's important to mention that if we normalize points by their population on both axes, I don't think we don't change the correlation, and even when normalizing by the population on the y-axis and unique Twitter users per state on the x-axis, I'm not sure we do much better (especially if unique Twitter users per state is the population times a fixed percentage)!  I believe this is actually the well-known problem of spurious correlation of ratios:

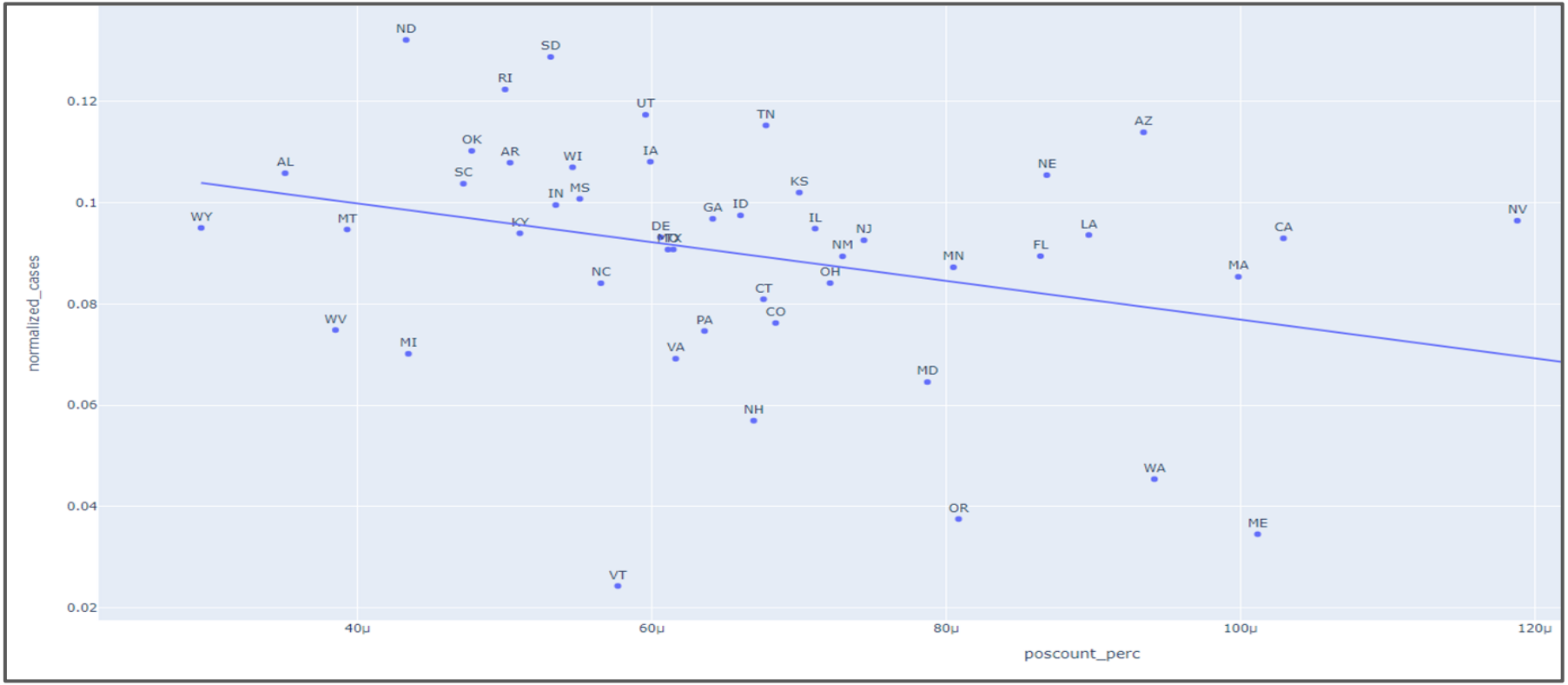
https://en.wikipedia.org/wiki/Spurious_correlation_of_ratios

which I believe can be addressed via John Aitchison's "Compositional Data Analysis" by taking a log (or other transform):

https://en.wikipedia.org/wiki/Compositional_data
https://en.wikipedia.org/wiki/Compositional_data#Linear_transformations

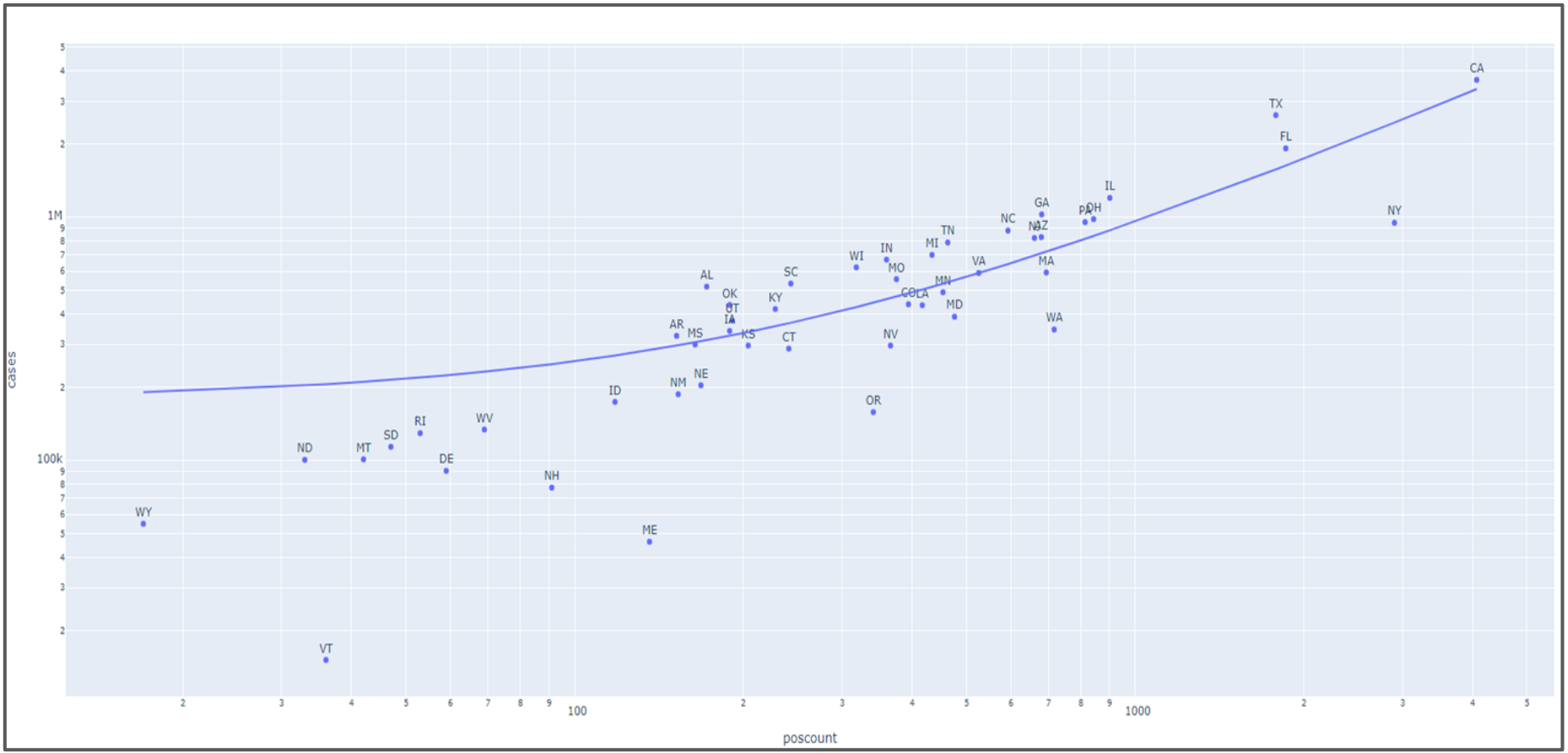
*Note: This is copied directly from email exchange with Prof. Sanner*

# Normalize x and y axis by State Population




R^2 = 0.135, proving linear relation previously is influenced by population effects..

# Unnormalized, log-log plot



# Mask Hashtag Frequencies

```
In [20]: positive_mask_hashtag_freqs
```

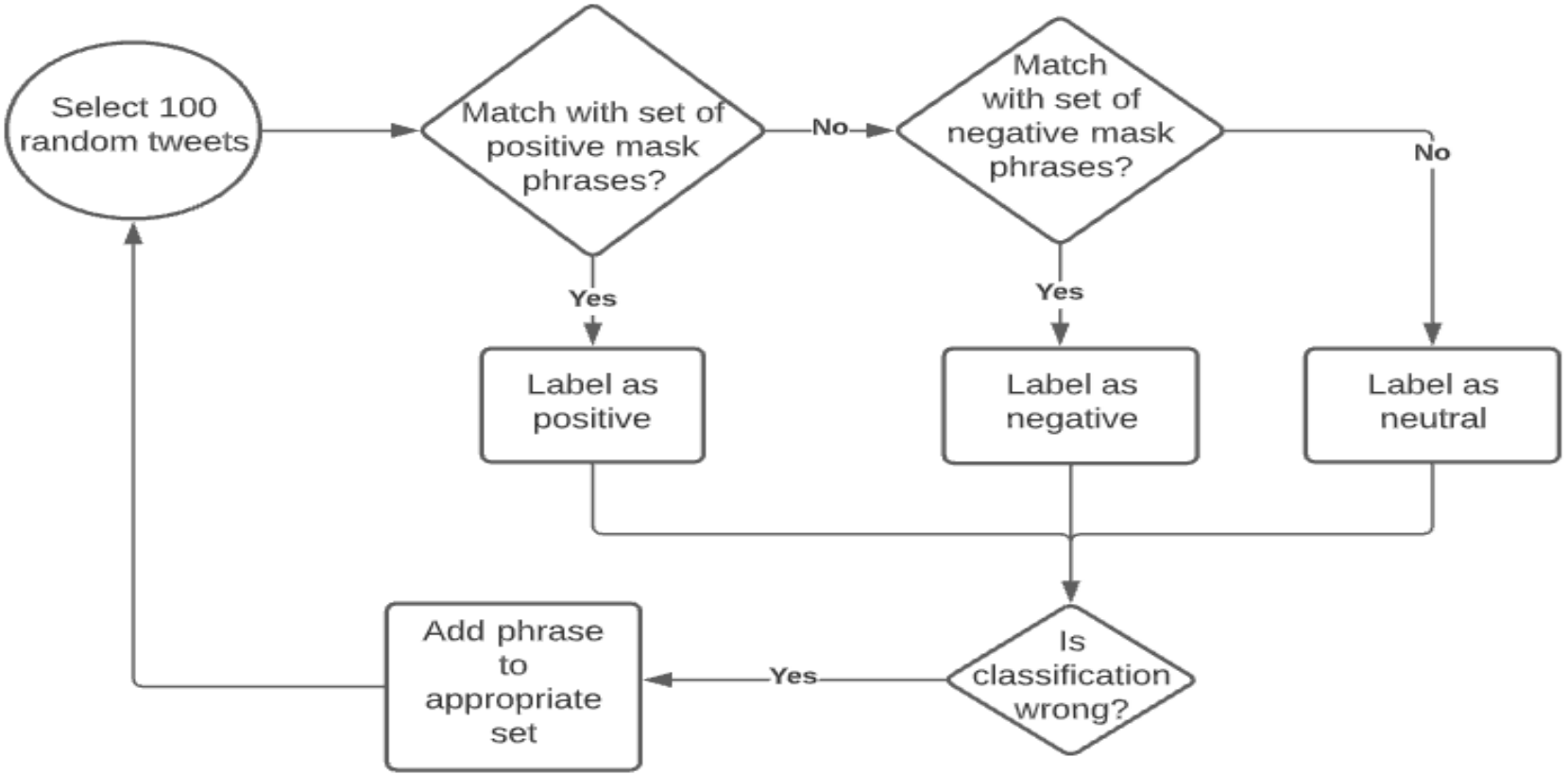
```
Out[20]: {'wearamask': 23348,
          'maskup': 5731,
          'wearadamnmask': 3180,
          'masks4all': 997,
          'wearamasksavealife': 786,
          'wearyourmask': 636,
          'maskssavelives': 588,
          'maskupamerica': 427,
          'maskupaz': 352,
          'wearthemask': 281,
          'wearamask.': 270,
          'maskitorcasket': 259,
          'wearthedamnmask': 202,
          'maskupnola': 199,
          'wearamaskplease': 180,
          'maskupmichigan': 156,
          'maskupmn': 149,
          'maskson': 144,
          'wearamask,': 143,
```

```
In [21]: negative_mask_hashtag_freqs
```

```
Out[21]: {'masksdontwork': 88,
          'unmaskamerica': 67,
          'masksoff': 56,
          'masksoffamerica': 33,
          'nomoremasks': 23,
          'unmaskcops': 9,
          'unmask': 8,
          'burnyourmask': 6,
          'unmaskarizona': 6,
          'fuckyourmask': 5,
          'unmasknj': 5,
          'nomaskrequired': 4,
          'takeoffthemask': 4,
          'masksareuseless': 4,
          'nomaskforme': 3,
          'nomoremask': 3,
          'fuckthemask': 3,
          'maskdontwork': 3,
          'fuckmasks': 3,
```

Key observations:
MUCH more positive tweets than negative tweets (39942 vs 405) !
Many users use common hashtags, but tweet not related to the pandemic
Suspect that mask hashtags not picking up on all negative tweets

# Revisiting Intent Classification: Regex Matching by Keywords



# Regex (keyword matches) not sufficient !

Tweets very personalized, expected angry mob of twitter users to write short, declarative sentences like “Wear a mask”, but tweets can be personal and generally hard to classify.

Holds mask as if it's a 4 x 6 card with his notes.
Claims America is "fighting" COVID.
This PSA Is a HOAX https://t.co/hVvQ41V07B

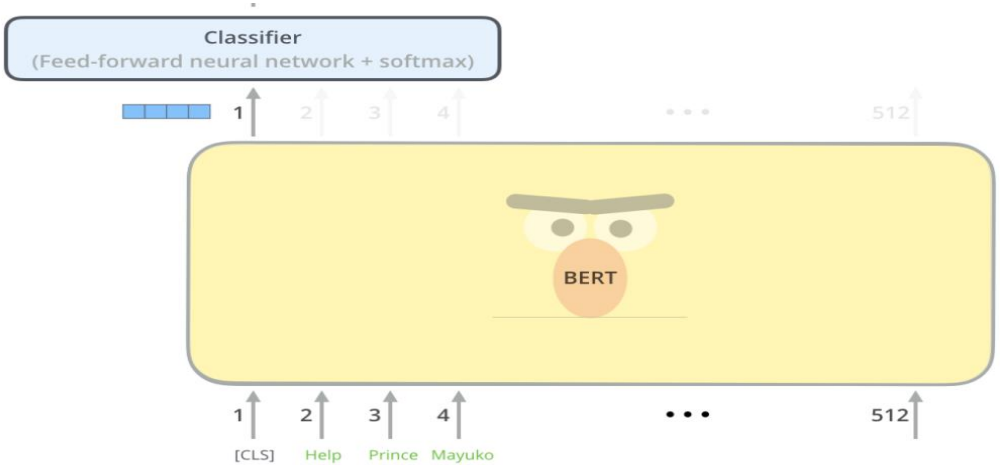
@chasestrangio Thank you for this. After just starting to pass as male in public before covid, I’m now back to only being read as female with a mask on and it’s been really hard on me. It helps to hear I’m not alone in it.

@kimtopher22 Nope. Still can’t wear a mask.

I am going to wear a mask outside. We live on an ½ acre so when washing the car or cutting grass raking leaves OR WEARING A MASK WHILE DRIVING IN MY CAR. SORRY I DONT WANT TO BE MISTAKEN AS A BIDEN VOTER I have more SELF RESPECT than that. Wolfs a good little minion a follower https://t.co/yJj0abGT2M

# Investigating BERT to use as Classifier

- Chose bert due its recent successes at NLP tasks (Question Answer, Intent Recognition, Sentiment Analysis, etc)
- Fine-Tune Pre-Trained model by using 1000 hand-labelled, randomly selected tweets
  - {0: “positive”, 1: “negative”, 2: “neutral”}
- Very few anti-mask tweets, so tried to search for anti-mask tweets by hashtags identified previously
- Learning Bert: https://jalammar.github.io/illustrated-bert/
- Implementation Guide: https://www.youtube.com/watch?v=gE-95nFF4Cc



# BERT Model Architecture

Followed guide on online tutorial for intent recognition with BERT, which in turn followed recommendations by the original paper.

```
model.summary()
```

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_ids (InputLayer)       [(None, 300)]             0
_________________________________________________________________
bert (BertModelLayer)        (None, 300, 768)          108890112
_________________________________________________________________
lambda_1 (Lambda)            (None, 768)               0
_________________________________________________________________
dropout_2 (Dropout)          (None, 768)               0
_________________________________________________________________
dense_2 (Dense)              (None, 768)               590592
_________________________________________________________________
dropout_3 (Dropout)          (None, 768)               0
_________________________________________________________________
dense_3 (Dense)              (None, 3)                 2307
=================================================================
Total params: 109,483,011
Trainable params: 109,483,011
Non-trainable params: 0
_________________________________________________________________
```

# BERT Model: Compile and Fit

## Compile Parameters:

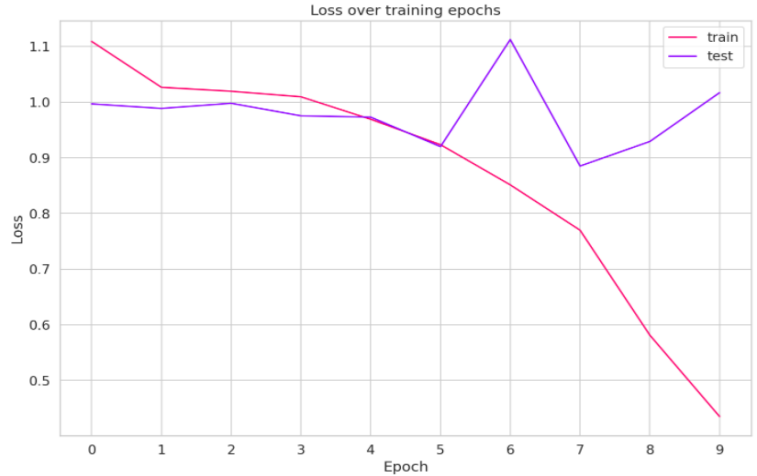
- Optimizer: Adam
- Loss Function: Sparse Categorical Cross entropy

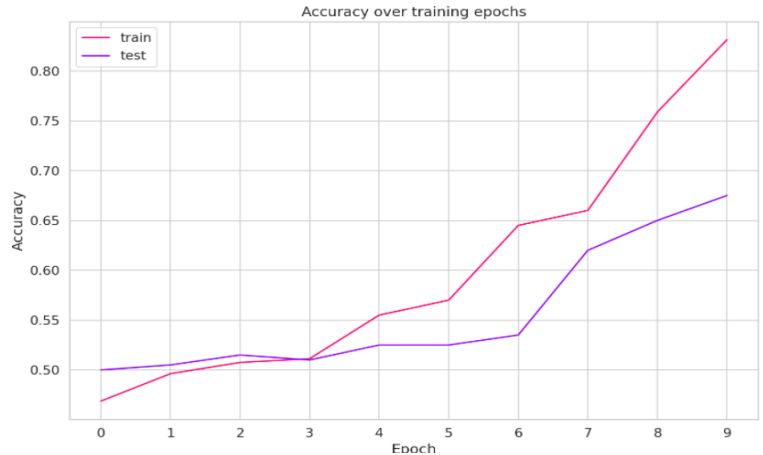
## Fit Parameters

- Validation_split = 0.2
- Batch size = 16
- Epochs = 10

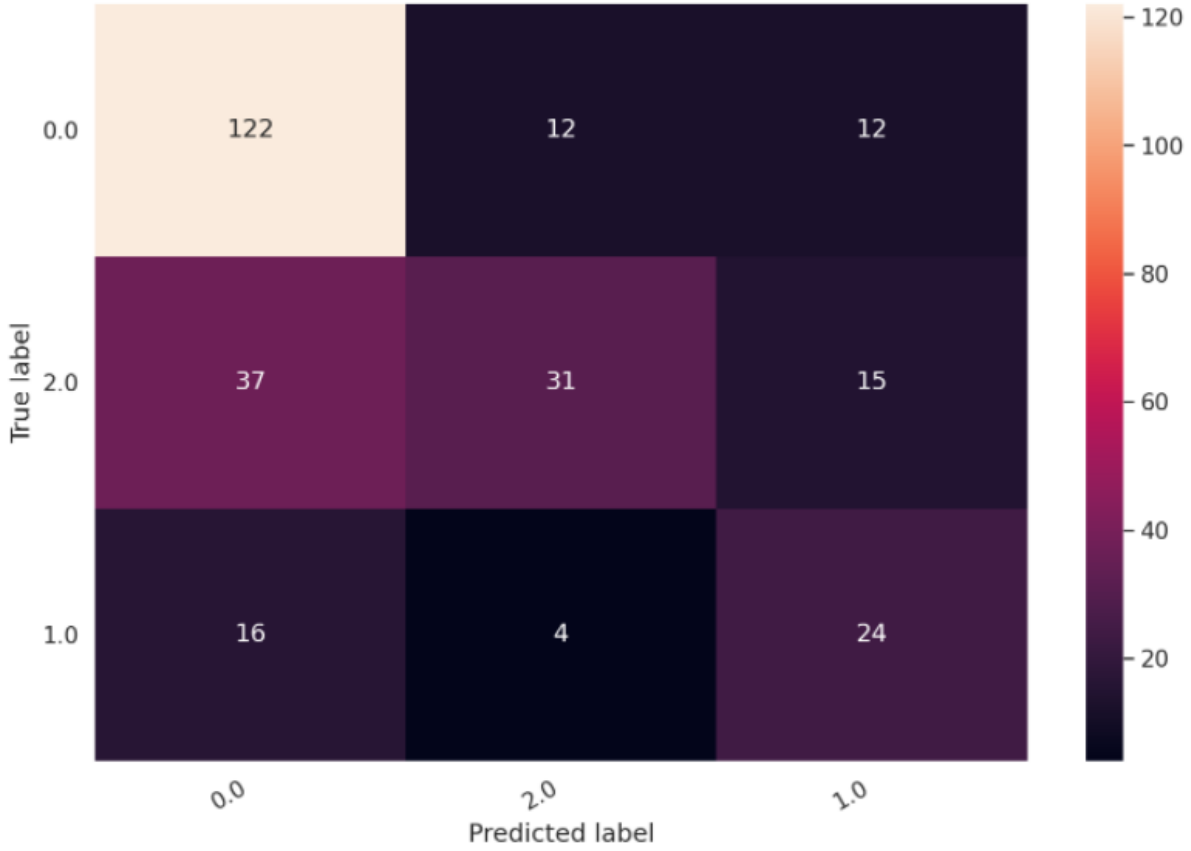
## Results:

- train acc 0.9010000228881836
- test acc 0.6483516693115234



# BERT Model: Confusion Matrix of Test Set

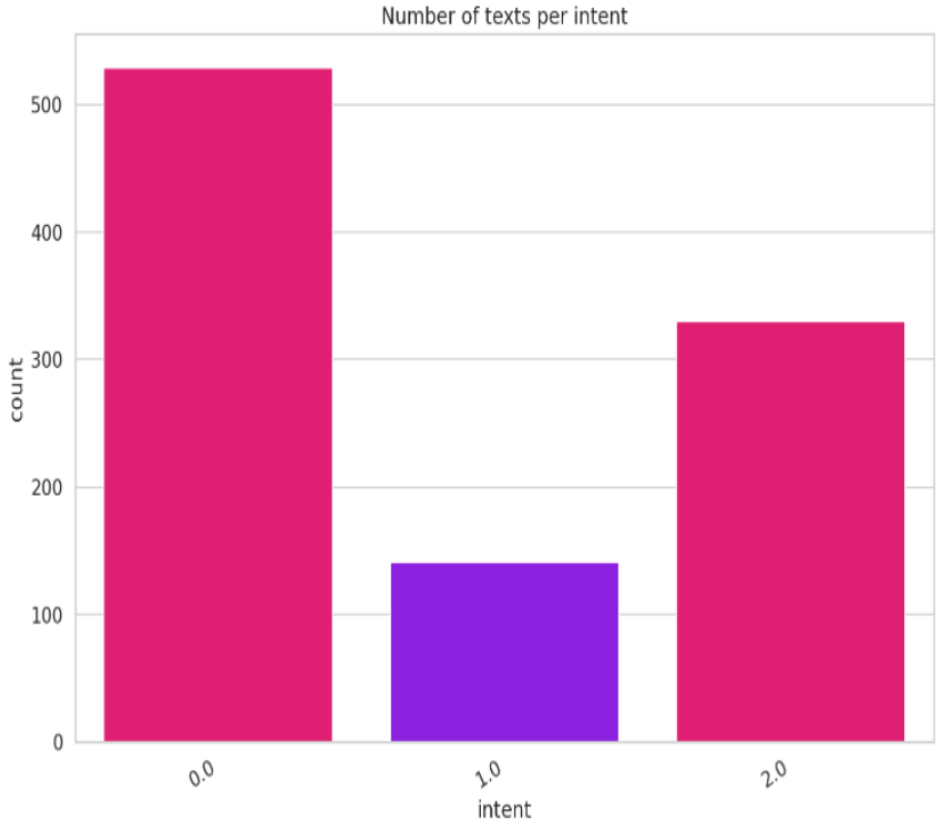



**Legend:**
0: pro-mask
1: anti-mask
2: neutral

# Explaining Poor Performance of BERT

Results are clearly not good enough to be used for classification. This makes sense for the following reasons:

- Only ~1100 labelled tweets
- Uneven class labelling even after searching for negative tweets
- Difficult to hand label in the first place. Many tweets are either hard to classify between a) pro-mask and neutral and b) anti-mask and neutral
- Only 1 human annotator. Prefer to use 2 and compute a Kappa score
- Another architecture is probably better. Having a single densely-connected layer with softmax reduces the number of parameters-> less overfit in training set
- Only played with epoch, not batch size etc



# Future Work

- Continue work on BERT. Classification of tweets turns out to be a very hard but interesting problem.
  - Label more tweets. Add more anti-mask tweets to balance the 3 sentiment classes
  - Due to small number of anti-mask tweets, focusing on 2 classes: pro-mask vs non-promask may be more useful.
  - Adjust model architecture, play with compilation and fit parameters
- Use classifier to reconstruct visualizations done with hashtag-based classifier
  - Number of pro-mask and anti-mask tweets classified by our model may follow covid cases/deaths or not.
  - Address normalization
- Investigate vaccine hesitancy
  - Initial investigation suggests this suffers from the same classification problem as masks. This is expected
- Covid metrics investigated are very simple: cases & deaths. Expand study to investigate whether tweet attitudes (towards masks, vaccines, etc) reflect changes in policy- if so, we can introduce a tool that uses Twitter to measure impact of policy on a population, spatially and temporally, allowing adequate and timely resource allocation (i.e: mental health support during quarantine)

# **Conclusions:** Key Insights we Gained from the Analysis of Data in this Project

- One clever approach: focus on indirect indications of "being affected by COVID" like mentions of relatives, or depression-related keywords
- Accurately classifying intent from tweet is a sub-research problem of its own
- Intent recognition (specifically, pro- or anti-mask classification) is hard!
- There are much more pro-mask tweets than anti-mask tweets, which makes training a classifier impractical with the dataset we were working with
- We theorize that BERT should perform well with more data
- More generally, there is a lot of noise in Twitter data (eg Election-related, protests-related, bots, etc)
- Twitter data can actually be useful datasets for such models, contrary to initial belief:
    - *“I did not expect mask attitude detection to be such a challenging task. These tweets are very personal, very contextual, and often quite indirect in their expression. ...I expected an angry mob of Twitter users spewing simple catchphrases and that's not at all what we're seeing here.” - Prof. Sanner*

# **Conclusions:** Data Challenges that Prevented us from Other Goals

We have covered these throughout the presentation, but to summarize:

- Small number of anti-mask tweets
  - Pro-mask tweets are retweeted more as well, so this movement has more momentum
- Tweets related to Covid19 are very personal, regex-based classifiers do not perform well.
- Hashtags are very often used just to relate an event to current time. Many tweets with certain hashtags are often not related to Covid19 at all.
- Sarcastic tweets are hard to detect.
- To our knowledge, no existing dataset with tweets and mask classification. Also, human annotation is time-expensive and requires 2 people to compute a Kappa score, but only 1 person did the annotations

# Acknowledgements

Prof. Scott Sanner

Prof. Reda Bouadjenek

Shreyas Sekar

Samin Aref

Amy Su

John Zhou