

BIG DATA TO APPROXIMATE MENTAL HEALTH INDICATORS

[by RINJANI]

BIG DATA FOR OFFICIAL STATISTICS













- **Official statistics** are usually produced based on **surveys and/or census** which are held **at most every year**.
- Between surveys and/or censuses there is an **information gap**.
- Along with the enhancement of technology, this gap could be filled with **real-time/big data** to give early warning to stakeholders about certain issues of interest.

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security


Pedro J. Zufiria , David Pastor-Escuredo , Luis Úbeda-Medina , Miguel A. Hernandez-Medina , Iker Barriales-Valbuena , Alfredo J. Morales , Damien C. Jacques , Wilfred Nkwambi , M. Bamba Diop , John Quinn , Paula Hidalgo-Sanchis , Miguel Luengo-Oroz 

Published: April 26, 2018 • <https://doi.org/10.1371/journal.pone.0195714>

THE LANCET
Diabetes & Endocrinology

COMMENT | VOLUME 6, ISSUE 8, P595-598, AUGUST 01, 2018

Using big data for non-communicable disease surveillance


Ran D Balicer  • Miguel Luengo-Oroz • Chandra Cohen-Stavi • Enrique Loyola • Frederiek Mantingh • Liudmyla Romanoff • et al. [Show all authors](#)

Published: November 13, 2017 • DOI: [https://doi.org/10.1016/S2213-8587\(17\)30372-8](https://doi.org/10.1016/S2213-8587(17)30372-8) • [Check for updates](#)

 Springer Link

Published: 18 June 2020

Estimating city-level poverty rate based on e-commerce data with machine learning

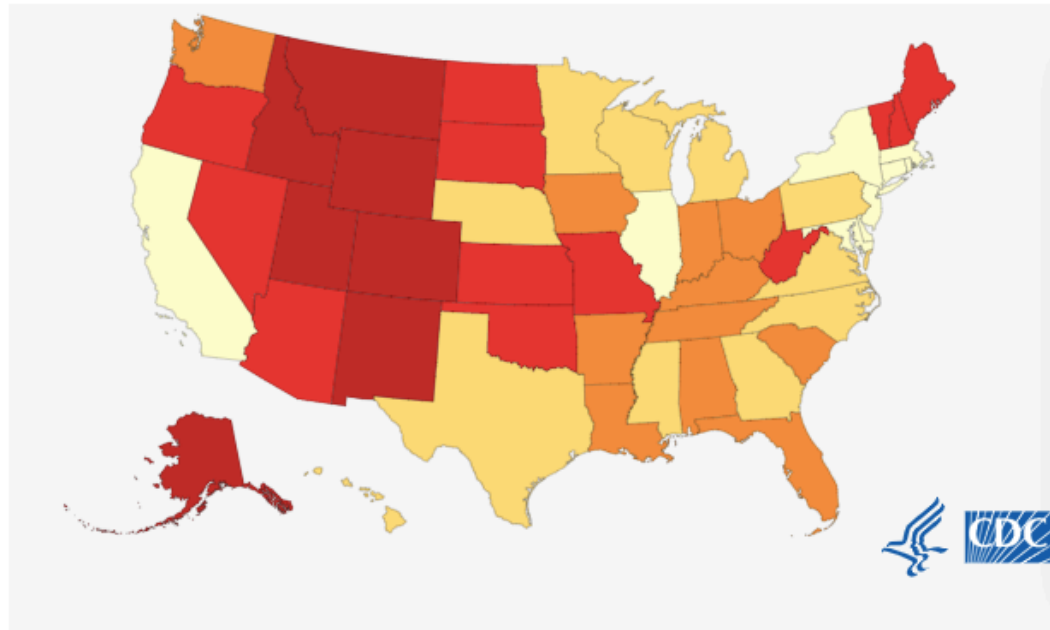
Dedy Rahman Wijaya , Ni Luh Putu Satyaning Pradnya Paramita, Ana Uluwiyah, Muhammad Rheza, Annisa Zahara & Dwi Rani Puspita

[Electronic Commerce Research](#) (2020) | [Cite this article](#)

131 Accesses | [Metrics](#)

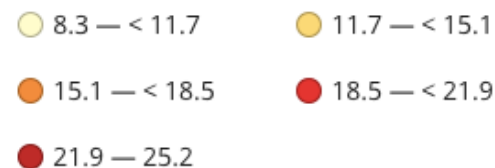
MENTAL HEALTH IN A TIME OF CRISIS

Suicide Mortality by State



The latest suicide rate available on CDC/WPR web page, as of March 2020, reported in 2018.

Age-Adjusted Death Rates¹



- **Suicide** is one of the big issues in **mental health** field.
- Referring to CDC/WPR web page, suicide rates by state are produced every year.
- Statistics for year 2019 and 2020 are not available. It is suggested to refer to the latest rate that is available → 2018 rate.
- During the time of crisis, like **current COVID-19, being informed about the suicide mortality is crucial**, since it's important to know if there's certain increase during this period.

WHAT WE AIM TO SOLVE?

PROBLEMS ARISE:

1. We're **lack of information** on suicide rate in 2019 and 2020 (although 2020 data might be available later end of year)
2. Even if we had yearly suicide rates, we still experience **information gaps** between those "end of years" when the rates are likely to be available.
3. If we want to **improve the disease surveillance** and help the department of health to develop public health interventions **in real-time** (or at least near real-time), we need to **fill this gap**, which will be **useful during the time of crisis like the current COVID-19**.



APPROACH PROPOSED:

"HARNESSING DIFFERENT SOURCES OF BIG DATA TO APPROXIMATE SUICIDE RATES BY STATE"

Why big data?

Because it's real-time!

In **every second**, there's always people:

- updating their **social media**
- using **search engines** to find information online
- **move** from one place to another

Note:

This approximation is **NOT** to **replace surveys/censuses**, but to **complement them** and **fill the information gap** between surveys/census.

BIG DATA SOURCE



We pulled tweets with **#covid19**, **#covid_19**, **#coronavirus** **#coronavirusoutbreak** **#coronavirusPandemic**, which are the same hashtags included in the dataset available on XPRIZE. (we don't use Twitter data on XPRIZE because the data doesn't have longitude-latitude information, which we really need in order to map all tweets by state)

- (1) We create new 121 dummy variables to encode if the tweets contain **121 words/lexicons related to suicide/depression**:
 - 112 lexicons taken from https://github.com/gamallo/depression_classification
 - 9 additional lexicons based on different sources
 - (2) We calculate negative-positive **sentiment score** (ranging from -1 to 1) for each tweet using the "sentimentR" package in R
 - (3) We do the **emotion analysis** based on **eight emotion axes**, i.e. anger, anticipation, disgust, fear, joy, sadness, surprise, trust, for each tweet using the "syuzhet" package in R
- => Total we have **131 variables corresponding to each tweet**, as well as **longitude-latitude mapped to state** from where the tweets are posted



32319 tweets to be aggregated by "state" → (1) summation, (2) summation and average, (3) summation and average → 141 features extracted from tweets data → 108 features to be included in the model

BIG DATA SOURCE

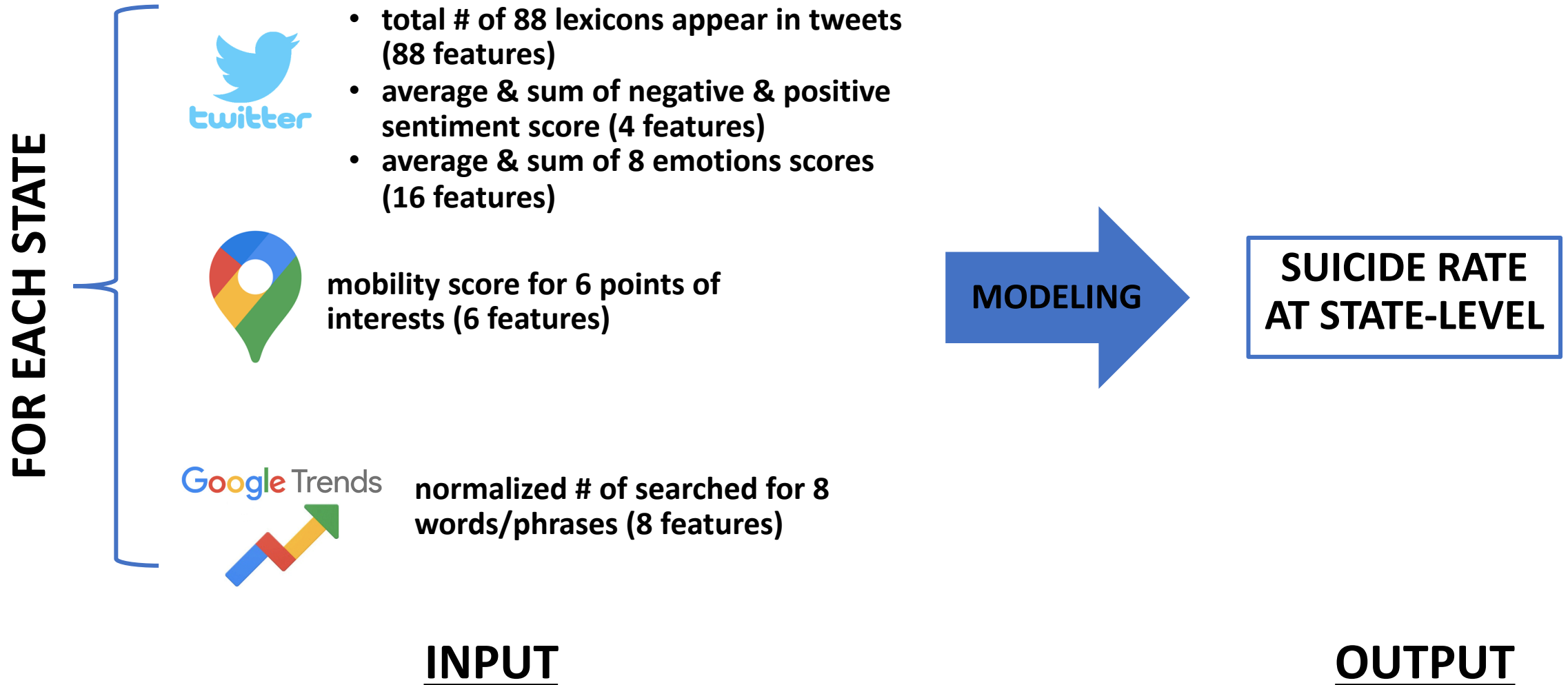


We use **Google mobility** data available on XPRIZE which contains the % of increase/decrease average mobility of people to certain points of interests by state: **Retail_recreation, Grocery_pharmacy, Parks, Transit_stations, Workplaces, Residential**



We use **Google Trend** to track the number of the following words/phrases searched in each state: **Antidepressant, Psychiatrist, Psychiatric hospital, Drugs, Suicide, I hate my life, Life sucks, I wanna die**

FRAMEWORK



METHODS: L1, L2 REGULARIZATION

To deal with **collinearity** and **high-dimensional data**, we fit the data with L1 and L2 regularization, i.e. adding penalty term in the SSE minimization.

- In L1 regularization (Lasso regression), we minimize:

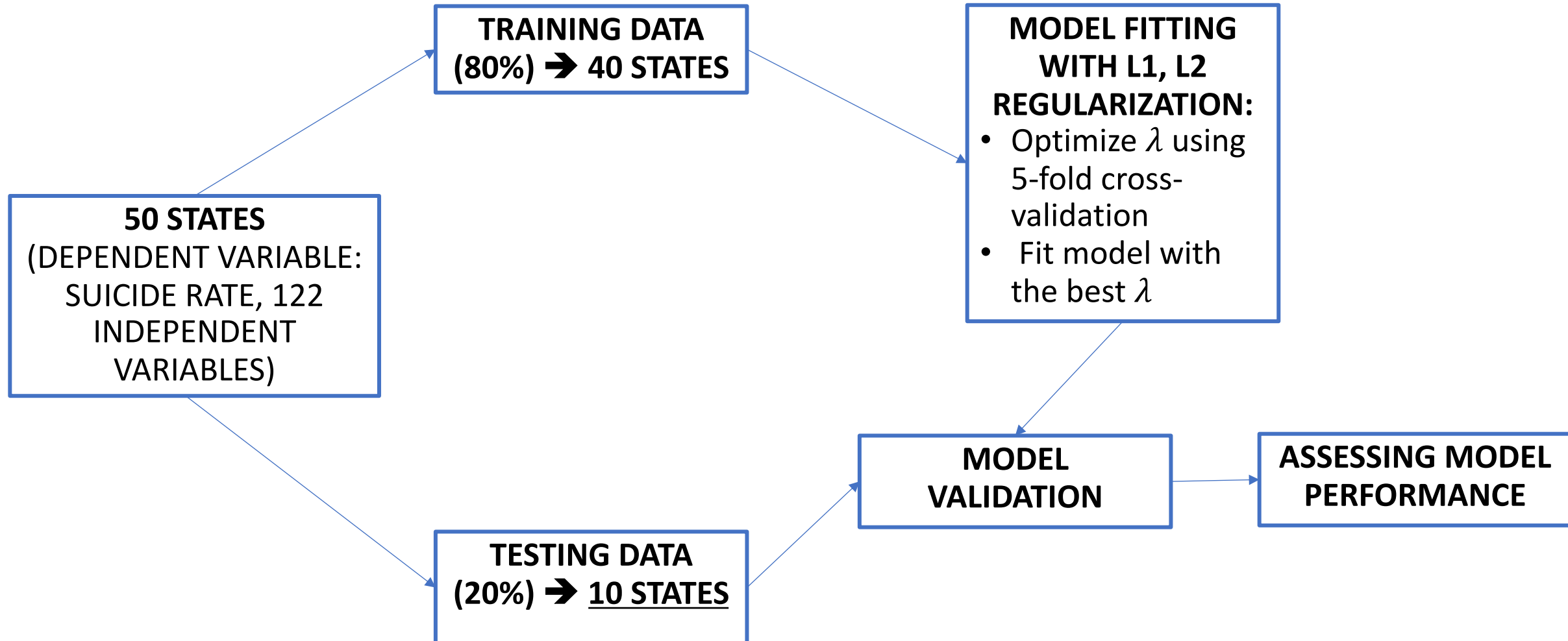
$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- In L2 regularization (Ridge regression), we minimize:

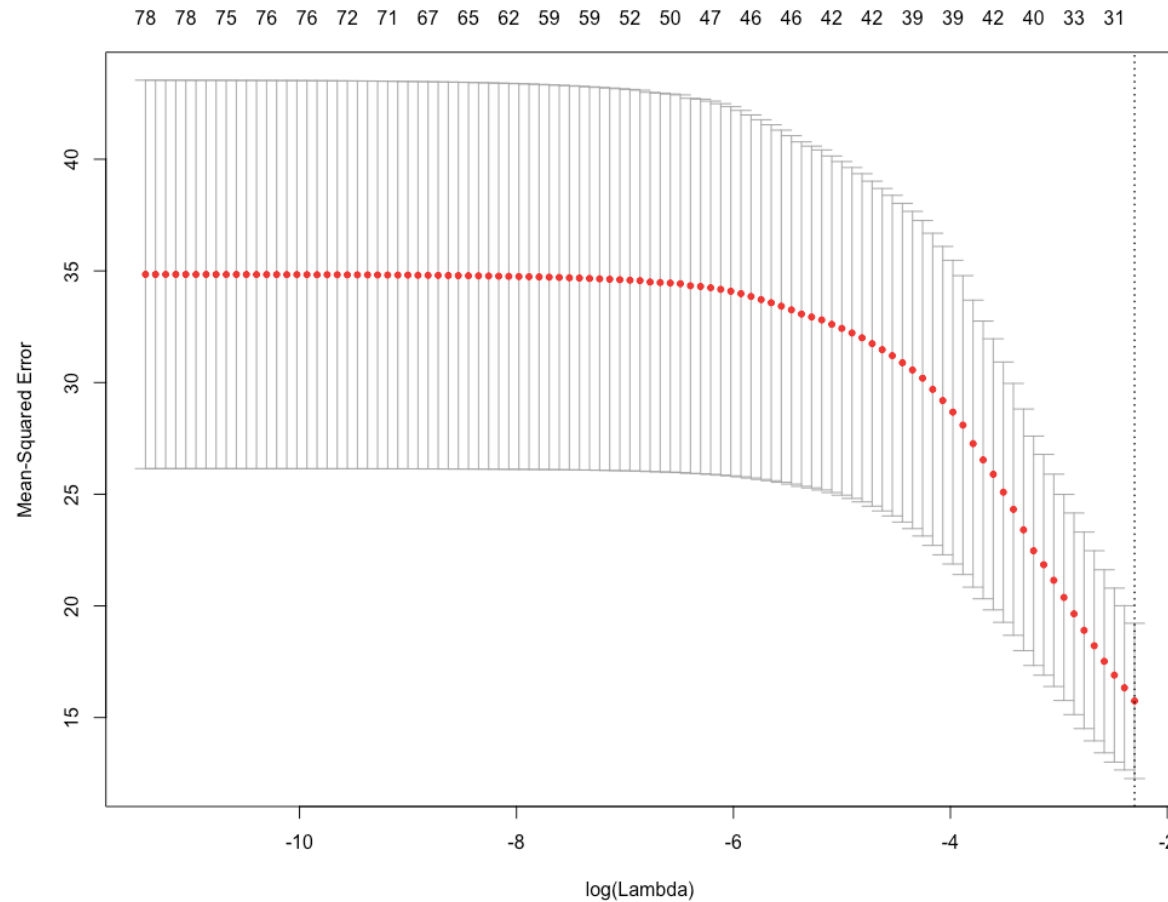
$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

where $\lambda \geq 0$.

METHODS: MODELING FLOW



RESULTS: L1 REGULARIZATION

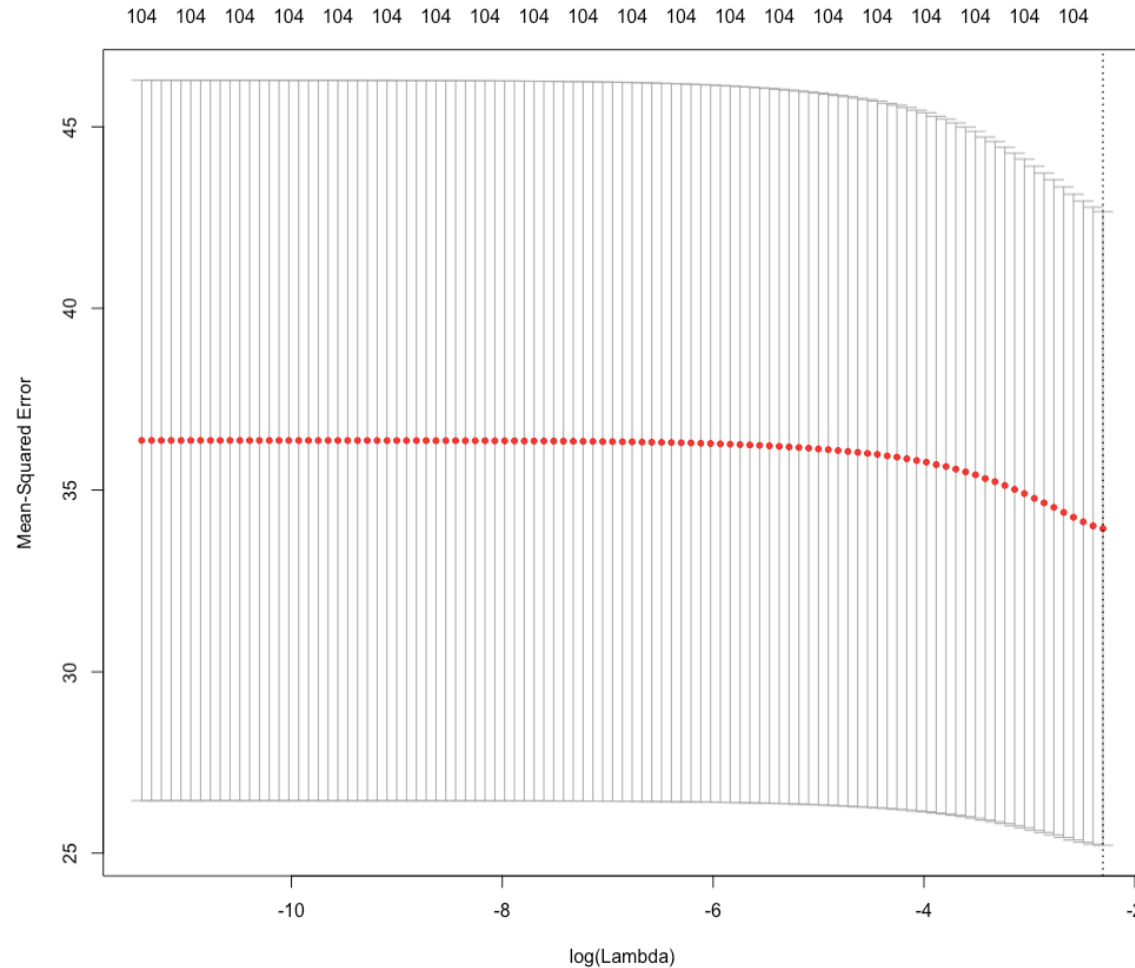


Number of features selected = 30

nrc.anger.x	sexual
nrc.anticipation.x	boy
nrc.joy.x	music
nrc.sadness.x	religion
nrc.surprise.x	father
sentiment_neg.x	depress
sentiment_pos.x	fuck
hate	support
save	suicide
amazing	workplaces
diagnosis	residential
woman	psych.hos
him	suic
girl	phrase1
Game	phrase2

Smallest train MSE = 15.74537, obtained at lambda = 0.1

RESULTS: L2 REGULARIZATION

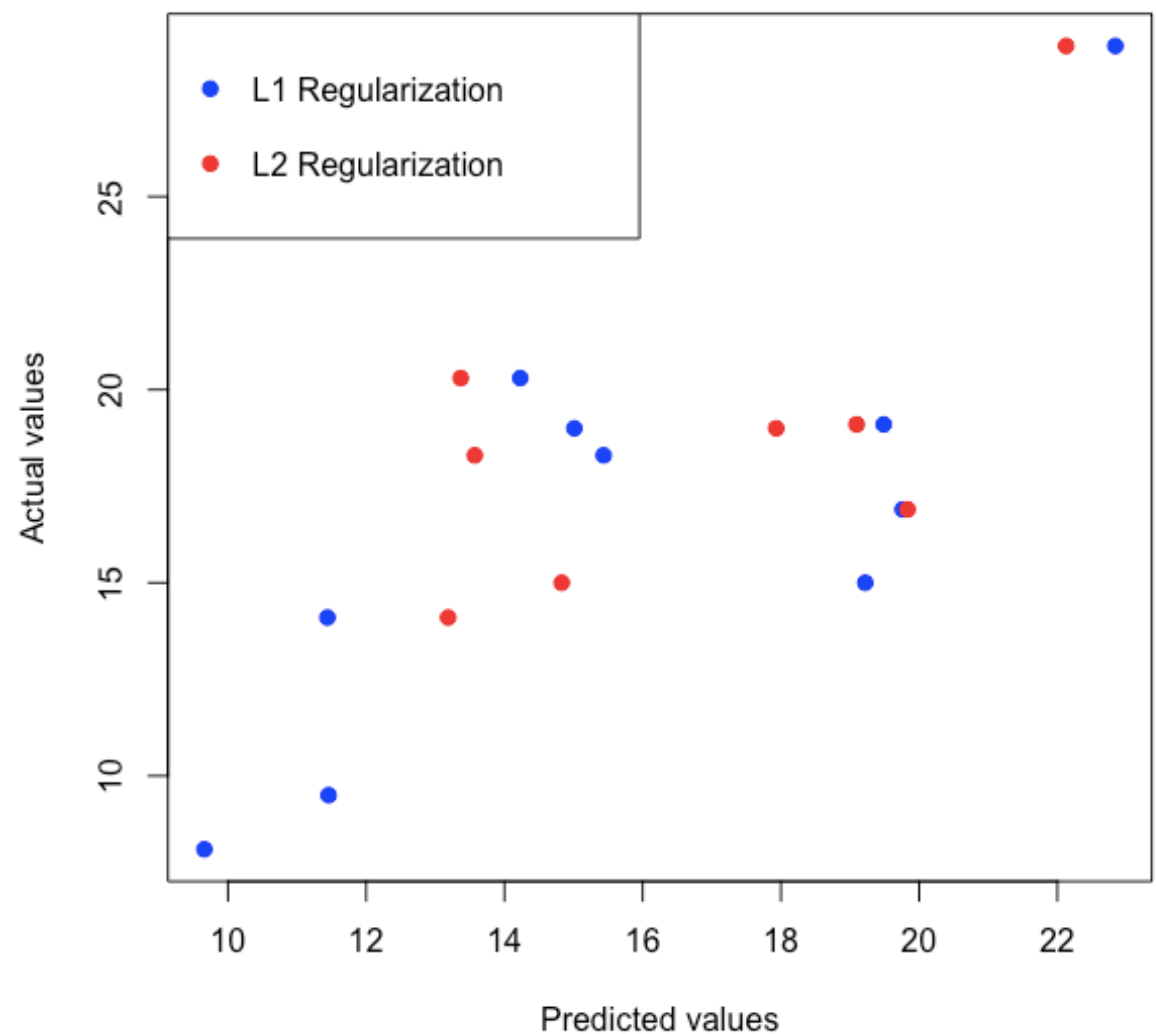


**Number of features selected = 104,
eliminating the following 18 features:**

withdrawal	side_effects
nausea	stimulant
episodes	psychotherapy
irritability	pills
headache	toxicity
imbalance	X40mg
drowsy	jesus
tolerance	bible
helpful	heaven

Smallest train MSE = 33.93968, obtained at lambda = 0.1

RESULTS: VALIDATION



STATE	ACTUAL	PREDICTED	
		L1 REG	L2 REG
COLORADO	20.3	14.23	13.37
KANSAS	19.1	19.49	19.09
KENTUCKY	16.9	19.76	19.83
MASSACHUSETTS	9.5	11.45	8.99
MICHIGAN	14.1	11.44	13.18
MISSISSIPPI	15.0	19.22	14.82
MONTANA	28.9	22.84	22.12
NEW YORK	8.1	9.66	3.77
OREGON	19.0	15.01	17.93
VERMONT	18.3	15.44	13.57

DISCUSSION: MODEL PERFORMANCE

- For both models we assess the performance by looking at MSE and R^2 .

	L1 REGULARIZATION	L2 REGULARIZATION
MSE	13.72	14.58
R^2	0.60	0.70

- Both models are actually promising, account for 60 – 70% variance in the data and pretty small errors.
- We recommend to use **L1 regularization (Lasso regression)** model because
 - (i) has the **smaller error**
 - (ii) more **interpretable** with only 30 features included in the model

DISCUSSION: INTERPRETATION

(Intercept)	2.001684e+01	
nrc.anger.x	6.685433e+00	extracted from tweets
nrc.anticipation.x	-9.100540e-03	
nrc.joy.x	-2.290914e+00	
nrc.sadness.x	-8.020730e+00	
nrc.surprise.x	1.934552e+00	
sentiment_neg.x	8.664771e+00	
sentiment_pos.x	7.276492e+00	
hate	1.320996e-02	
save	-1.555107e-01	
amazing	1.141553e-02	
diagnosis	4.013131e-01	
woman	2.661923e-01	
him	-2.392646e-01	
girl	2.232640e+00	
game	-2.220934e-01	
sexual	3.769089e+00	Google mobility
boy	2.111139e-01	
music	-8.751156e-01	
religion	3.726142e-04	Google trend
father	-2.904836e+00	
depress	1.557927e-07	
fuck	-6.570366e-01	
support	-1.976194e-01	
suicide	6.047503e-01	
workplaces	-9.085011e-02	
residential	-1.401861e+00	
psych.hos	-1.496544e+02	
suic	2.213104e+01	
phrase1	-1.119414e-02	
phrase2	-2.160927e+00	

L1 REGULARIZATION MODEL:

- Suicide rate can be approximated with **30 features** extracted from **tweets, Google mobility, and Google trend.**
- **15 features affect suicide rate negatively** (the ones with negative coefficient estimates)
- **15 features affects suicide rate positively** (the ones with positive coefficient estimates)

LIMITATIONS & FUTURE WORKS

LIMITATION

In general, the limitations of this project is that the available data are not in the same time frame.

1. The latest suicide rates available on WPR web page, as of March 2020, were reported in 2018.
2. We did not use tweets data available on the XPRIZE web page since this dataset does not have longitude and latitude information for the tweets. We needed longitude and latitude information to map each tweet to the states. Thus, we pulled tweets data on our own. We were only able to pull last week's tweets (pulled on September 17th, 2020, so we had September 10th - 17th 2020 tweets) as we used the free Twitter API.
3. The Google mobility data available on XPRIZE were last updated on March 30th, 2020.
4. The implications of (2) is that we pulled Google trend data during the period of September 10th - 17th 2020 to at least match the time frame of Twitter data.

POTENTIAL FUTURE WORKS

We could scale-up this work with two different scenarios:

1. Perform **similar cross-section analysis but using one-year** tweets (paid API is needed) and Google mobility for more representativeness, since suicide rates data are also available yearly
2. Perform **longitudinal/time-series analysis** to look at the change/trend overtime – this one would need **at least 5 years** of suicide rates, tweets, Google mobility, and Google trend.

It is also worth to try experimenting with other modeling approaches.

THANK YOU!