In partnership with

COVID-19 Data Repository and County-level Death Count Prediction in the US

Bin Yu

UC Berkeley Statistics, EECS, CCB

github.com/Yu-Group/covid19-severity-prediction

Website: covidseverity.com

MSRI Workshop on Mathematical Models for Prediction and Control of Epidemics

August 14, 2020

On March 22, we responded to a call for data science expertise by Response4Life...

Initial Goal: Help Aid Resource Allocation

PI: Bin Yu


N. Altieri


R. Barter


J. Duncan


R. Dwivedi


K. Kumbier


X. Li

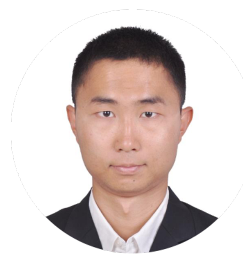

R. Netzorg


B. Park


C. Singh (Student Lead)

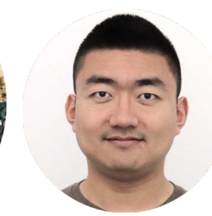

Y. Tan


T. Tang


Y. Wang

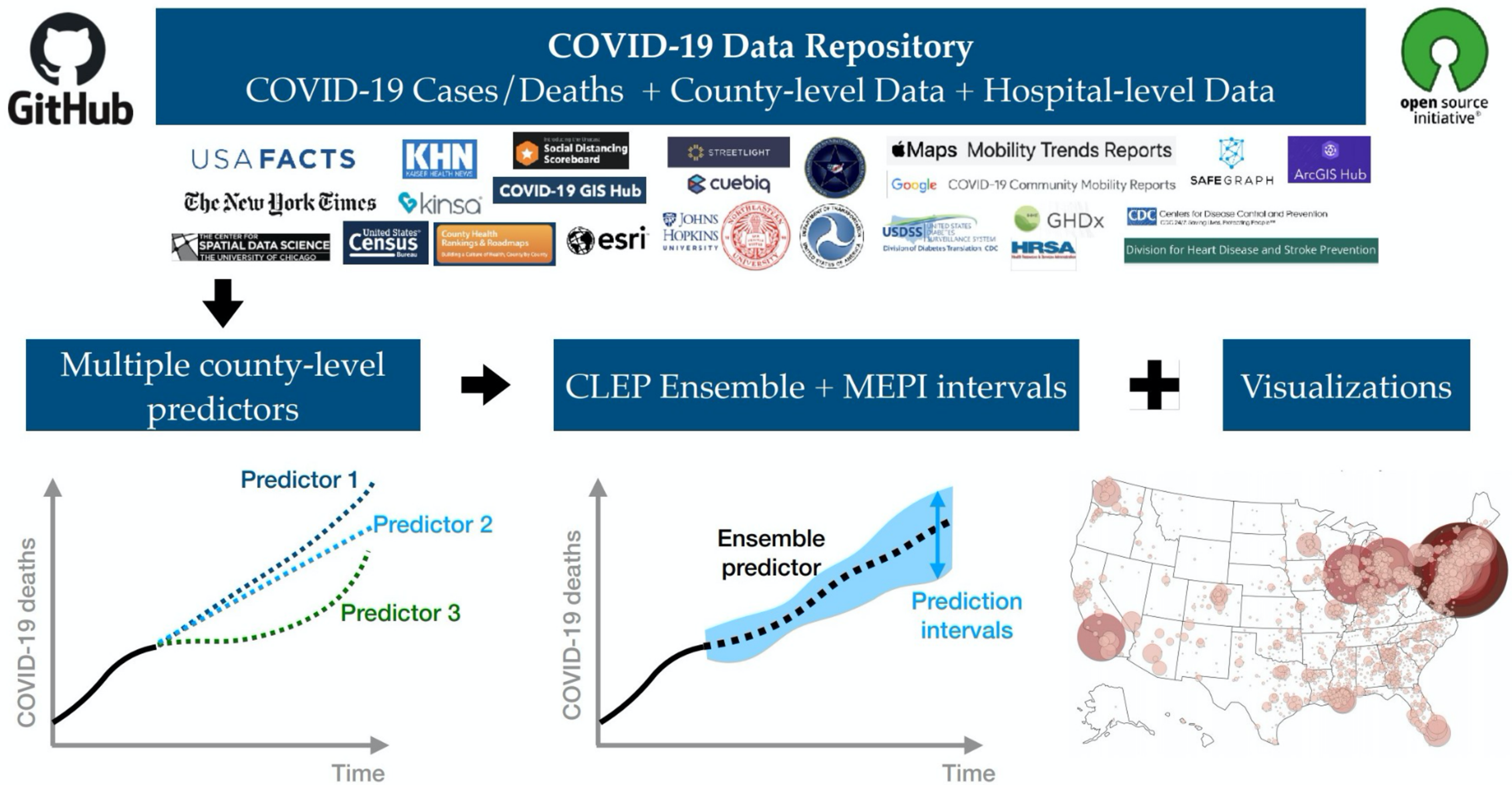

A.Agarwal


M. Shen


C. Zhang

Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...

Overview: Current Data Repository & Prediction Pipeline (Open Source)

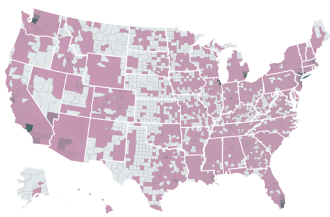


Curating a COVID-19 Data Repository

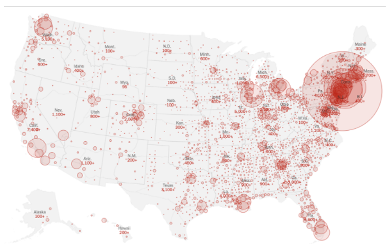
Data curation: scraped from a variety of sources

COVID-19 Cases/Deaths

USAFACTS



The New York Times



THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

County-level Data
(Risk Factors, Demographics, Social Mobility)

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™
Division for Heart Disease and Stroke Prevention

Department of Transportation, United States of America

esri
COVID-19 GIS Hub

IHME GHDx

County Health Rankings & Roadmaps
Building a Culture of Health, County by County

USDSS United States Diabetes Surveillance System
Division of Diabetes Translation, CDC

CMS.gov
Centers for Medicare & Medicaid Services

United States® Census Bureau

SAFEGRAPH

kinsa®

STREETLIGHT

Introducing the Unacast
Social Distancing Scoreboard

KHN
KAISER HEALTH NEWS

cuebiq

Maps Mobility Trends Reports

Johns Hopkins University

Google COVID-19 Community Mobility Reports

Hospital-level Data
(e.g., #ICU beds, staff)

HRSA
Health Resources & Services Administration

ArcGIS Hub

Homeland Infrastructure Foundation-Level Data Subcommittee

Samuel Scarpino

Northeastern University

A bird's-eye view of the **hospital-level & county-level data**

- ~7000 hospitals in US
- ~200 features:
    - Geographical identifiers: address, lat/long, county
    - Type of facility (e.g., short term acute care, critical access)
    - Urban/rural
    - # total beds, # Med-Surg beds, # ICU beds
    - ICU Occupancy rate
    - #Employees, #RNs
    - Total discharges, average length of stay, average daily census
    - Hospital overall rating

- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
    - Population, population density, age structure
- Health risk factors
    - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
    - Social vulnerability index, unemployment, poverty, education, severe housing
- Social distancing and mobility
    - County-to-county work commute, change in distance traveled, government orders
- Other relevant data
    - Sample of flight itineraries in 2019, Kinsa temperature data, voting data

Data quality issues about death counts

- Undercount problems (after April 14, counts include probable deaths)

Data quality issues about death counts

- Undercount problems (after April 14, counts include probable deaths)
- USAFacts and NYT data come from the same sources, but do not always agree

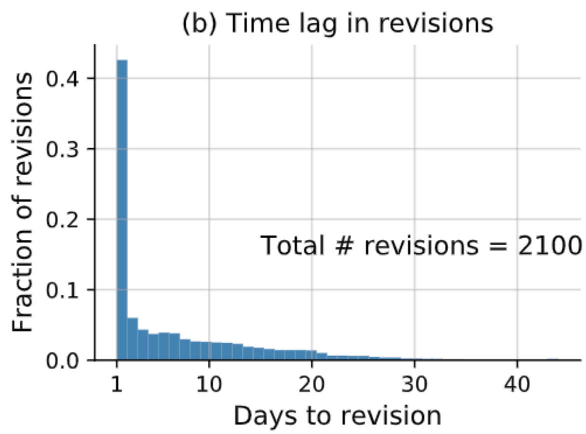
We use USAFacts data because it does not lump NYC counties together

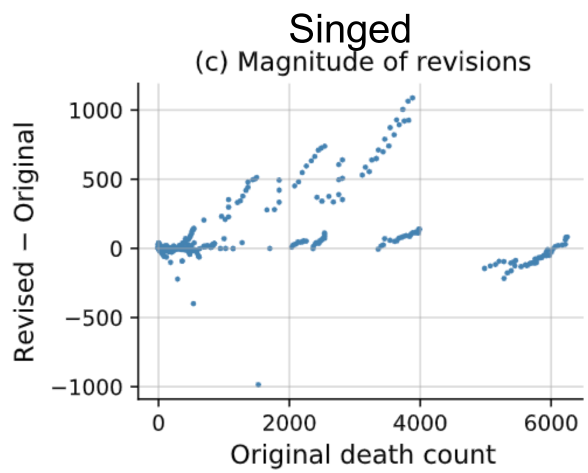
Data quality issues about death counts

- Undercount problems
- USAFacts and NYT data come from the same sources, but do not always agree
- Weekdays are different from weekends

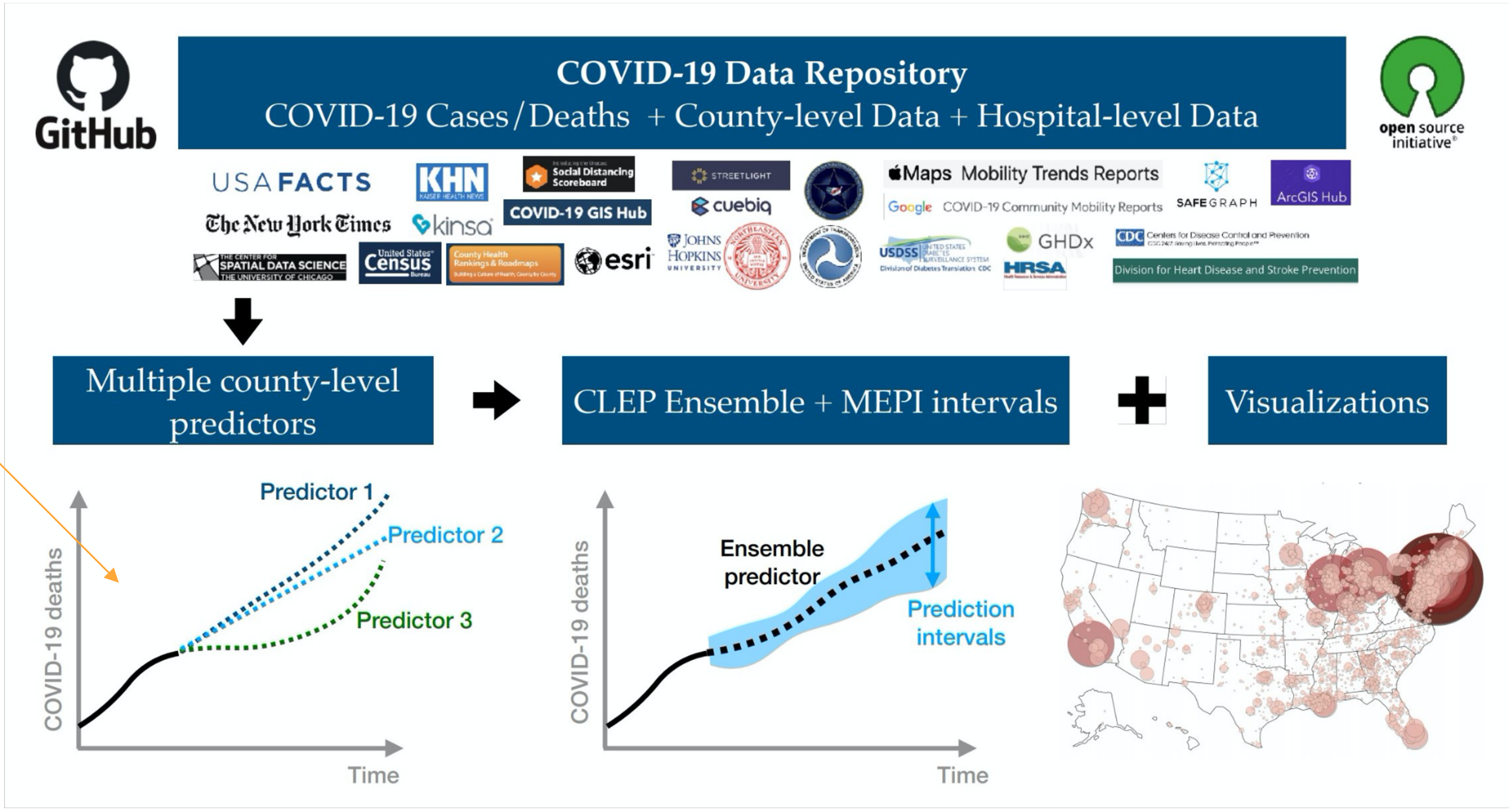
Data quality issues about death counts

- Undercount problems
- USAFacts and NYT data come from the same sources, but do not always agree
- Weekdays are different from weekends
- Historical data revisions





Overview: Current Data Repository & Prediction Pipeline (Open Source)



Forecasting county death counts

Curses

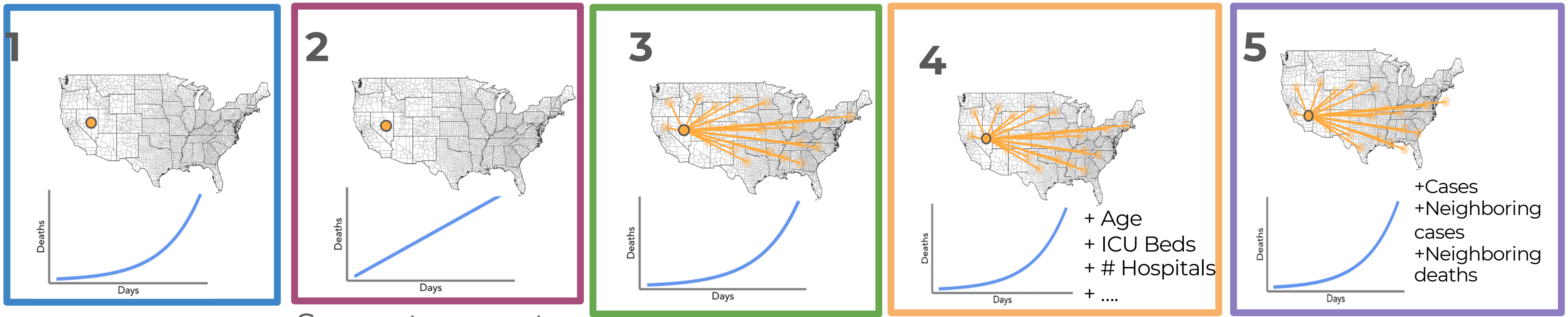
- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

Curses and blessings

- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

- Everyday, we get new observed data to measure our predictions against -- great reality check and keeps one honest
- For PPE supplies, one week prediction is adequate (we can actually do 14 day reasonably well)

Individual Linear and Exponential Predictors



Separate-county exponential predictor

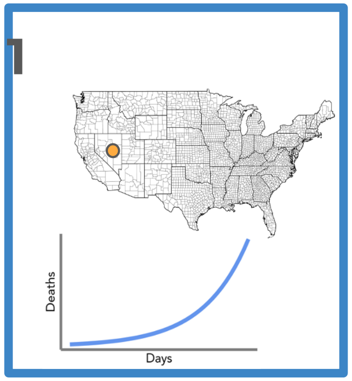
Separate-county linear predictor

Shared-county exponential predictor

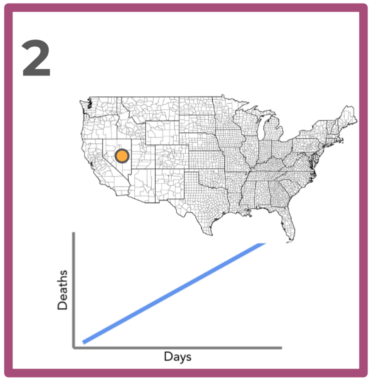
Shared-county exponential predictor + demographics

Expanded Shared-county exponential predictor

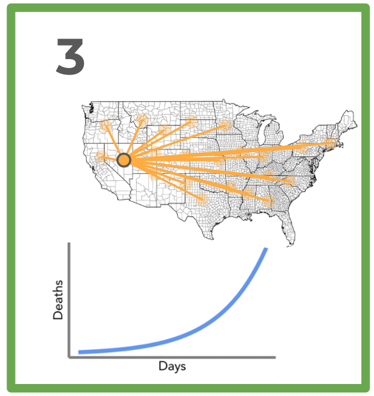
Combined Linear and Exponential Predictors (CLEP)



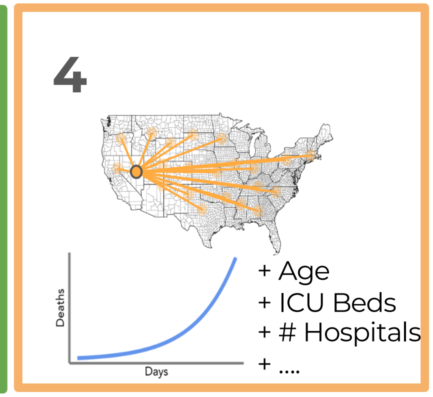
Separate-county exponential predictor



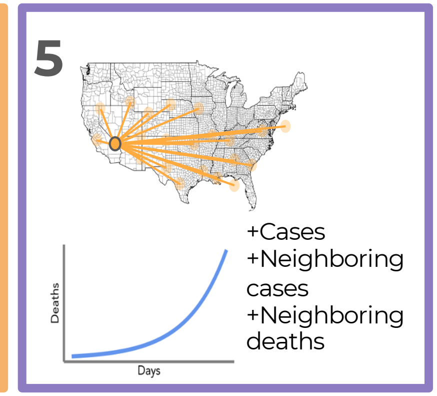
Separate-county linear predictor



Shared-county exponential predictor



Shared-county exponential predictor + demographics



Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better (recent) historical performance[1]

[1]. Schuller-Yu-Huang-Edler "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

Combined Linear and Exponential Predictors (CLEP)

Calculate a weighted average of the predictions: higher weight to the models with better (recent) historical performance[1]

$$w_t^m \propto \exp\left(-c(1-\mu)\sum_{i=t_0}^{t-1}\mu^{t-i}\ell(\widehat{y}_i^m, y_i)\right)$$

Without μ , the weights are well motivated through Rissanen's predictive MDL (Minimum Description Length) principle , and μ in (0,1) allows adaptation to changing dynamics.

[1]. Schuller-Yu-Huang-Edler "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

CLEP details with M predictors for k day (ahead) prediction

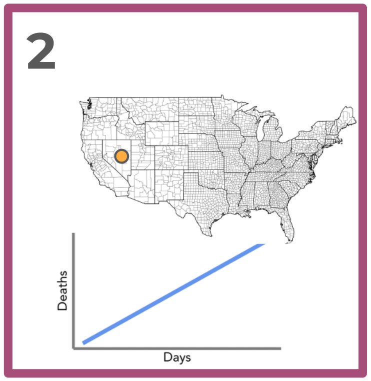
$$\widehat{y}_{t+k-1}^{\text{CLEP}} = \sum_{m=1}^{M} w_t^m \widehat{y}_{t+k-1}^{\text{m}}.$$

$$w_t^m \propto \exp\left(-0.5 \sum_{i=t-7}^{t-1} (0.5)^{t-i-1} \left|\sqrt{\widehat{y}_i^m} - \sqrt{y_i}\right|\right)$$

using the past 7 day errors for each predictor and forgetting factor 0.5

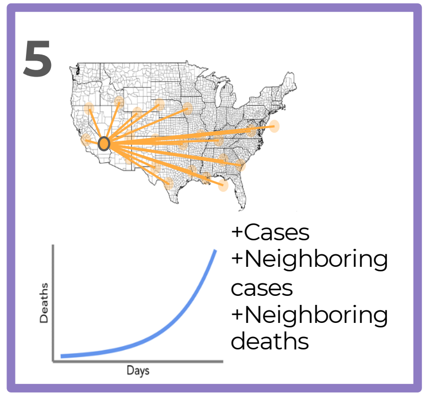
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two predictors performs well



Separate-county linear predictor

+



Expanded Shared-county exponential

k=7 for 7-day prediction

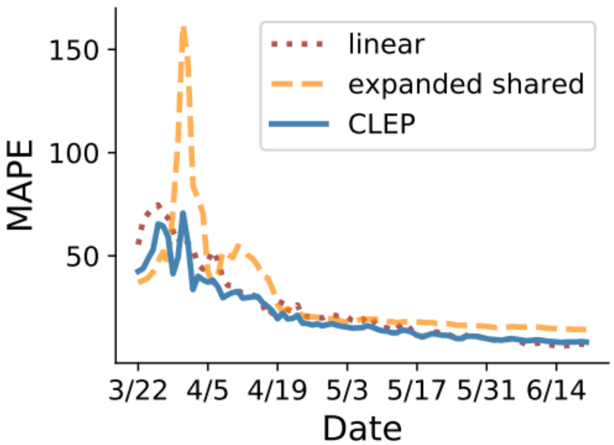
$$\mathrm{E}[\text{deaths}_t|t] = \exp\Big(\beta_0 + \beta_1\log(\text{deaths}_{t-1}+1) + \beta_2\log(\text{cases}_{t-k}+1) + \beta_3\log(\text{neigh\_deaths}_{t-k}+1) + \beta_4\log(\text{neigh\_cases}_{t-k}+1)\Big)$$

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[1]

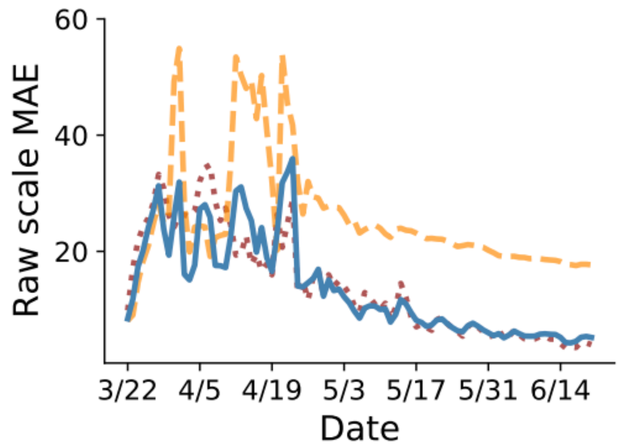
[1].Schuller-Yu-Huang-Edler . "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

Absolute error results over March 22 – June 20 (7-day prediction)

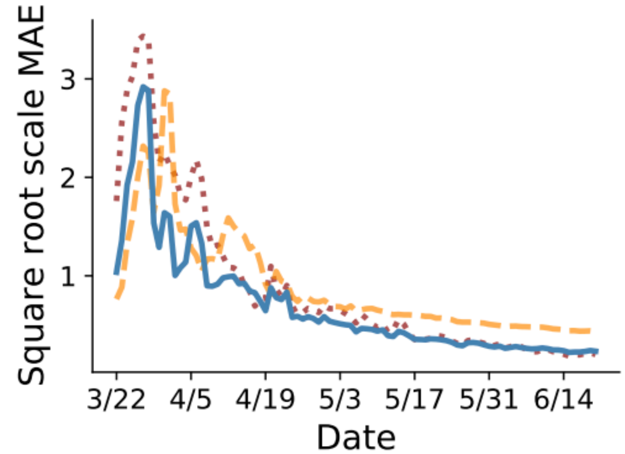
CLEP is combining linear and expanded shared (with monotonicity)



(a) MAPE



(b) Raw-scale MAE



(c) Square-root-scale MAE

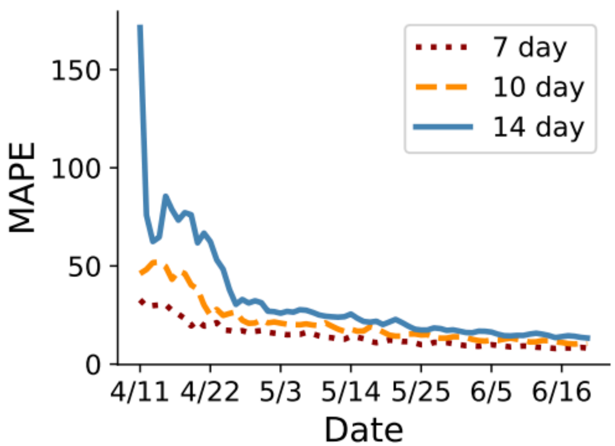
$$\text{Raw-scale MAE}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}|\widehat{y}_t^c - y_t^c|$$

$$\text{MAPE}_t(\%\text{ error}) = 100\times\frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\frac{|\widehat{y}_t^c - y_t^c|}{y_t^c}$$

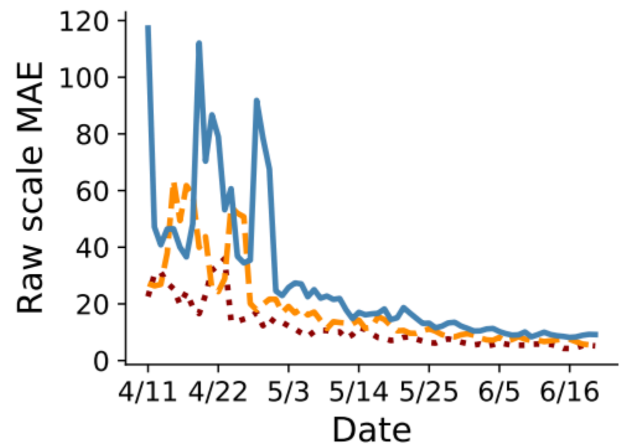
$$\text{Sqrt-scale MAE}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\left|\sqrt{\widehat{y}_t^c} - \sqrt{y_t^c}\right|$$

$\mathcal{C}_t$ contains counties with at least 10 deaths on day t

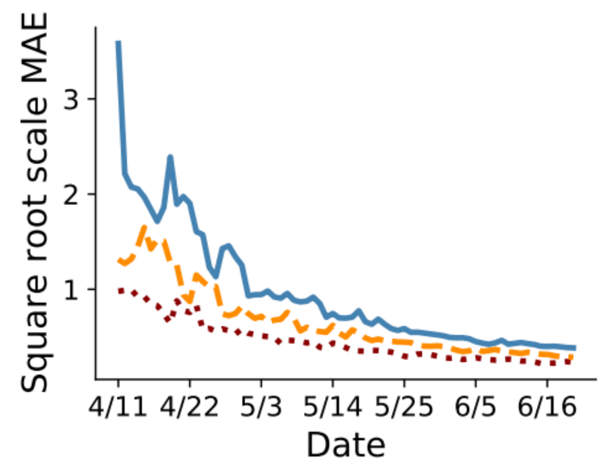
Absolute error results over March 22 – June 20 (7-, 10-, 14- day ahead)



(a) MAPE



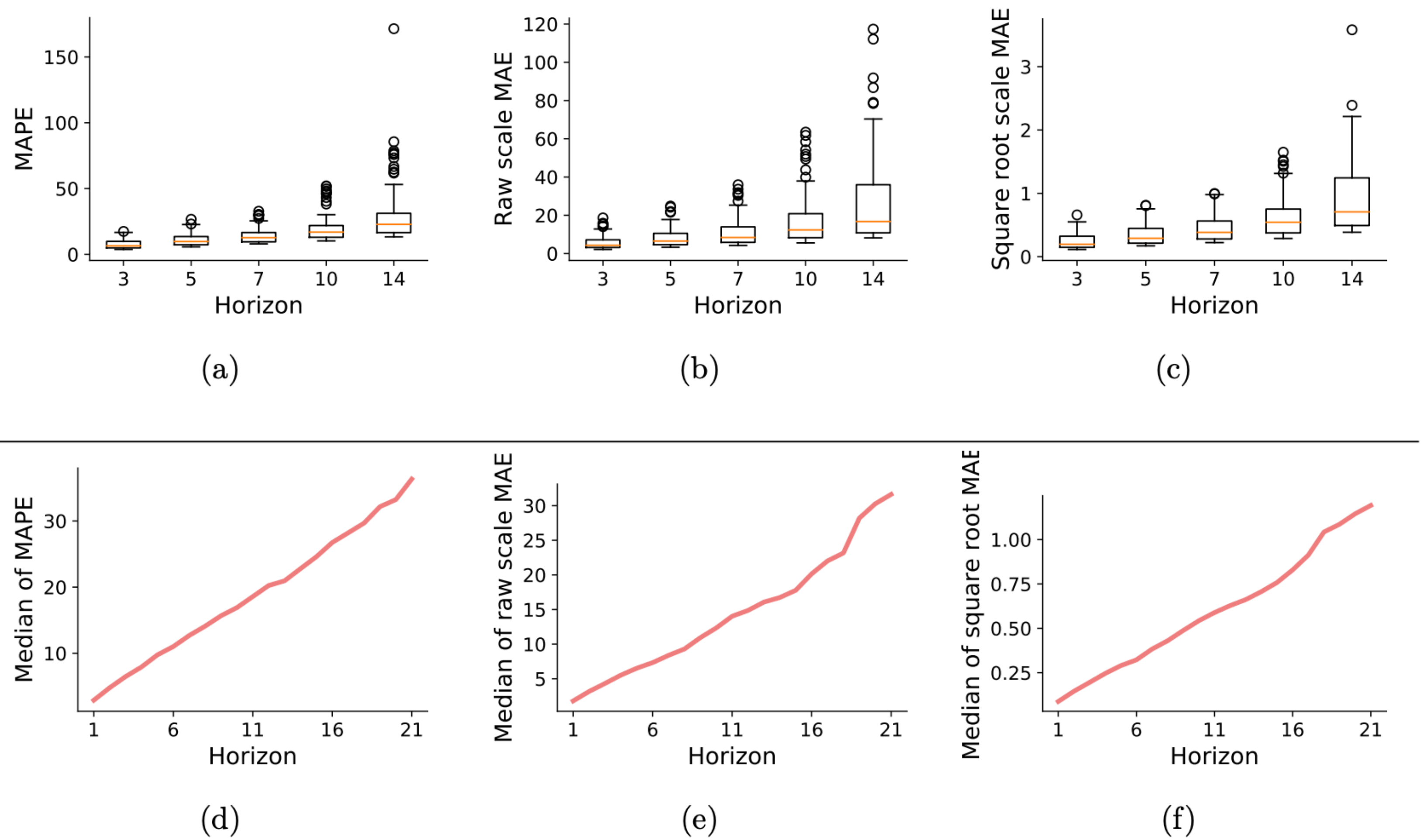
(b) Raw-scale MAE



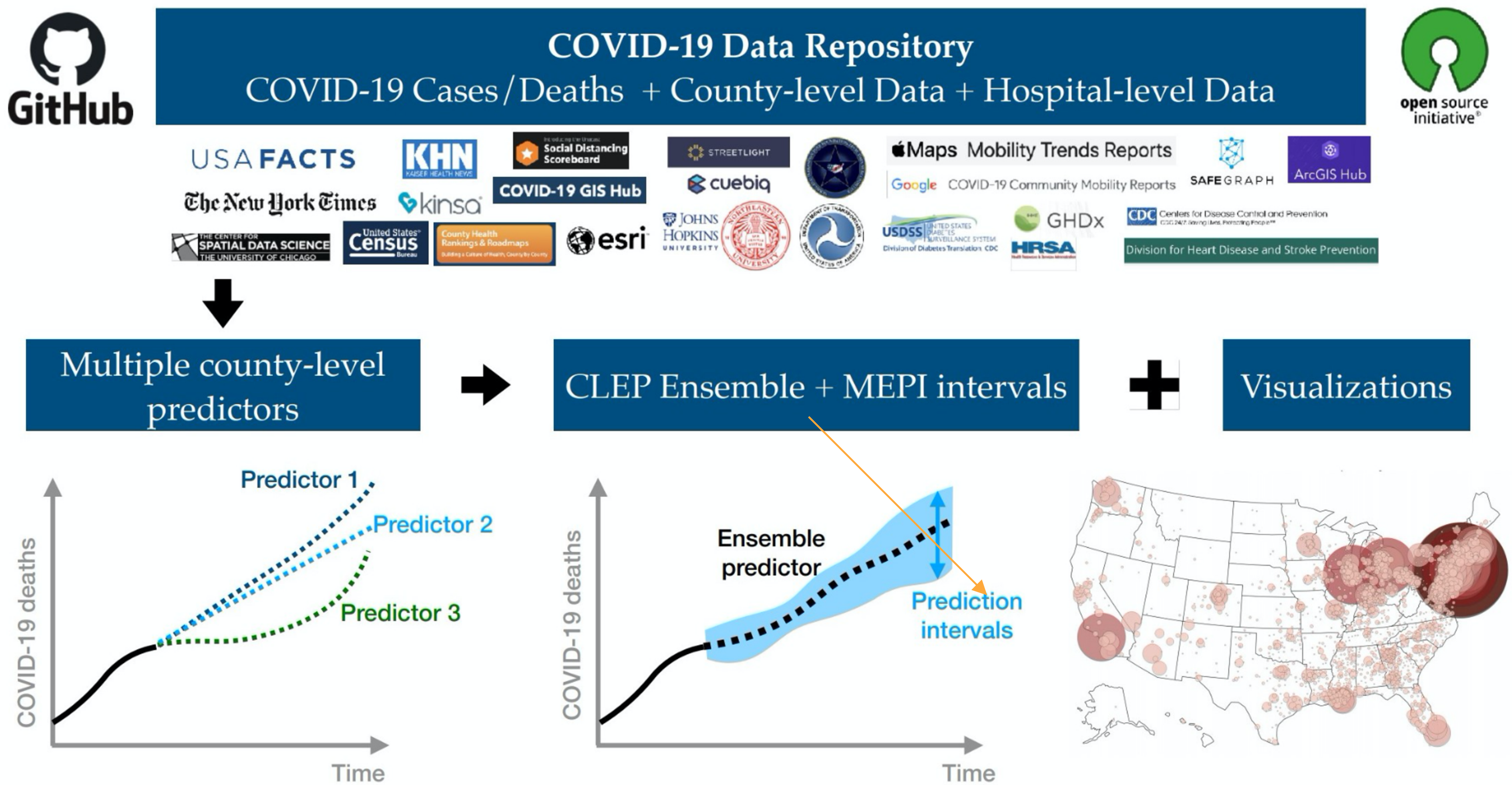
(c) Square-root-scale MAE

$$\text{Raw-scale MAE}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}|\widehat{y}_t^c - y_t^c| \qquad \text{MAPE}_t(\%\text{ error}) = 100\times\frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\frac{|\widehat{y}_t^c - y_t^c|}{y_t^c} \qquad \text{Sqrt-scale MAE}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t}\left|\sqrt{\widehat{y}_t^c} - \sqrt{y_t^c}\right|$$

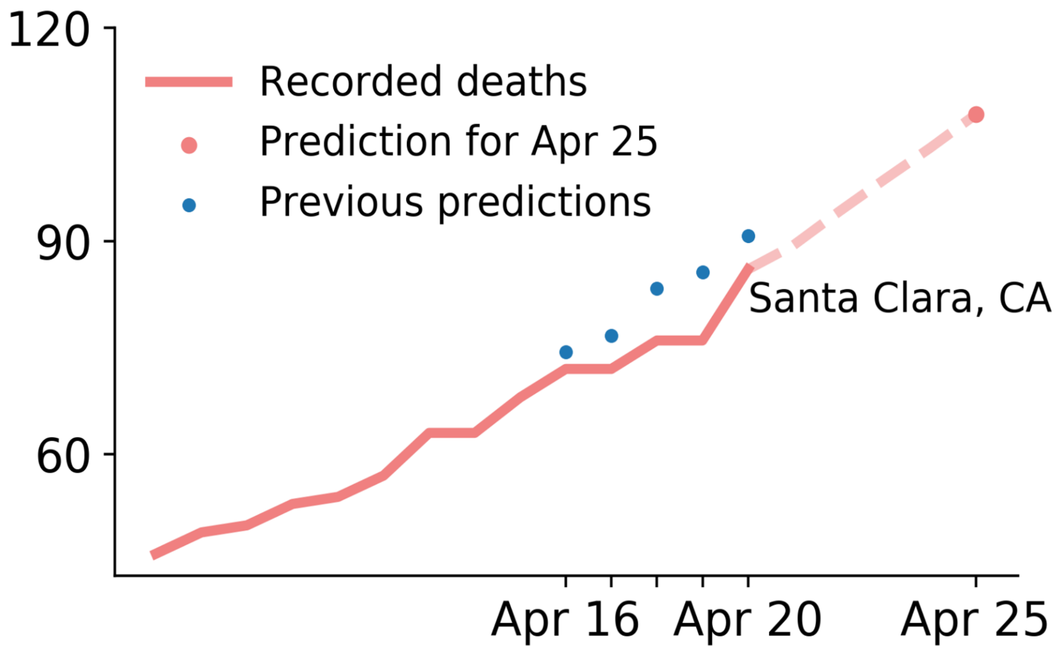
The further into the future, the lager the prediction error



Overview: Current Data Repository & Prediction Pipeline (Open Source)



Prediction Intervals based on conformal prediction[2]



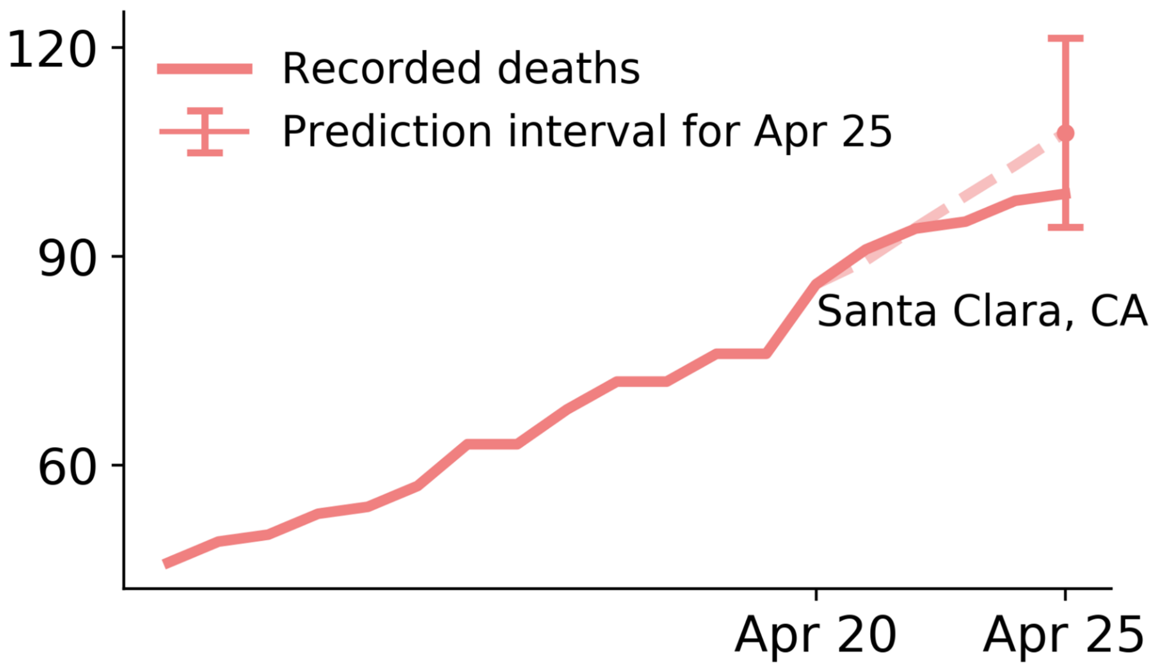
Previous 5-day-ahead rel. prediction errors (%)

Apr 16	*3.3%*
Apr 17	*6.5%*
Apr 18	*9.6%*
Apr 19	*12.6%*
Apr 20	*5.5%*
Apr 25	**?**

Take the max

[2]. G. Shafer and V. Vovk "A tutorial on conformal prediction." *JMLR* (2008): 371-421.

Prediction Intervals:



Predicted range of error

Apr 25 **[-12.6%, 12.6%]**

Actual error:

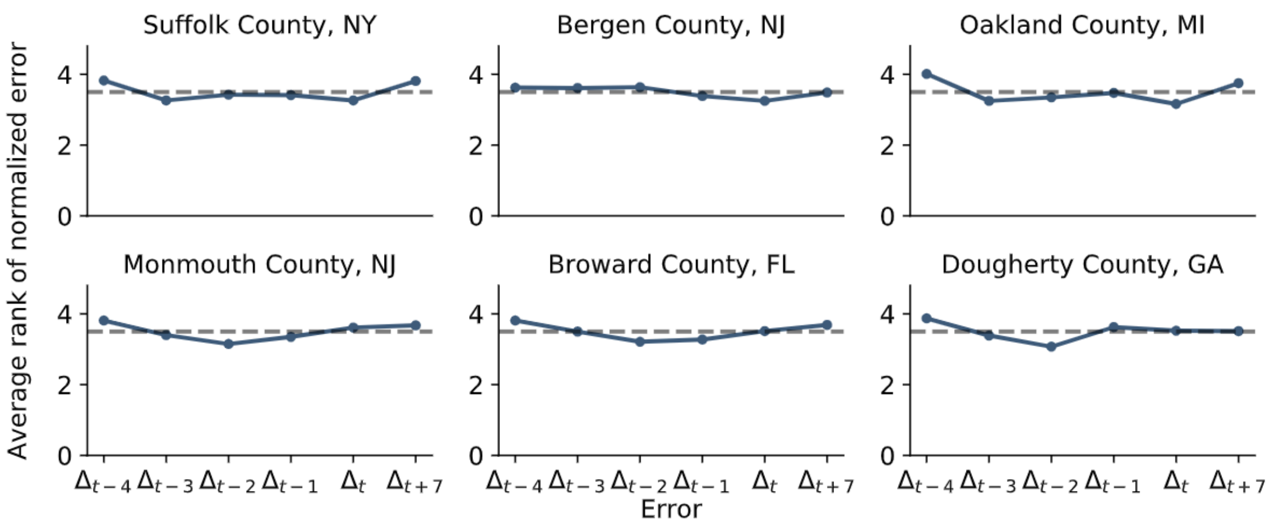
Apr 25 8.8%

Exchangeability assumption on normalized prediction errors

- If the normalized prediction errors are exchangeable, then the MEPI coverage is 5/6=83%

- Checking this assumption using observed normalized prediction errors

Average rankings around 3.5 as expected under assumption



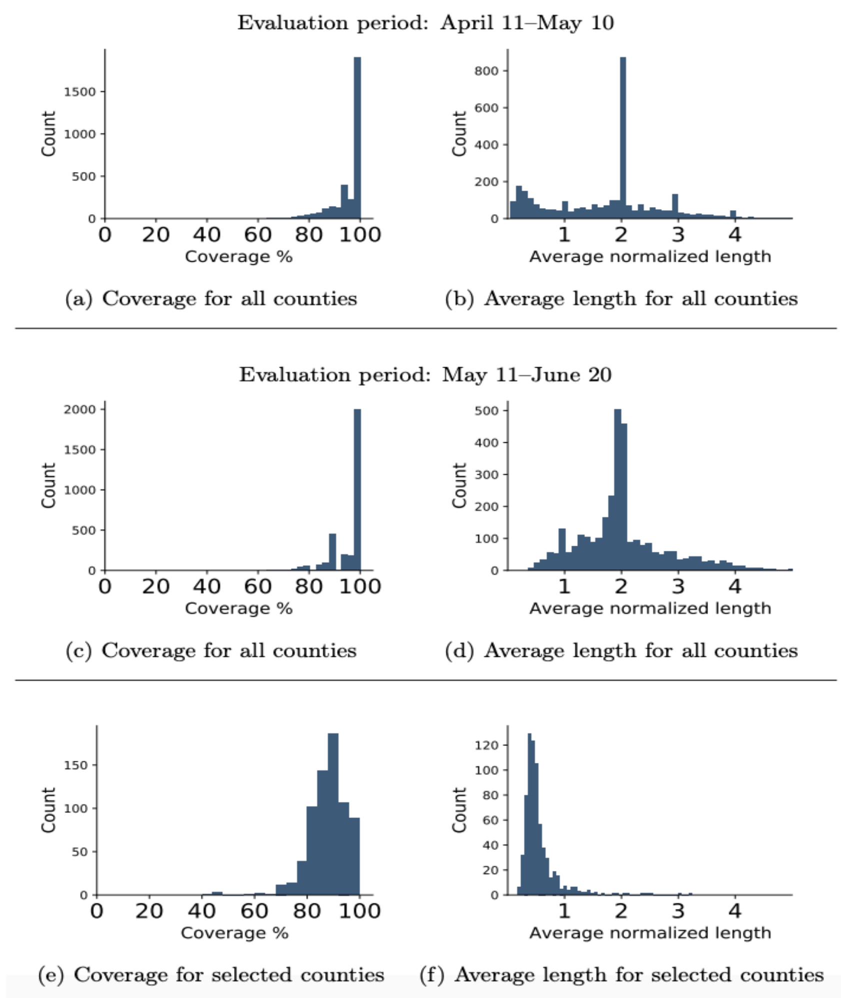
(b) Six randomly-selected counties

Empirical evaluation of coverage of prediction intervals

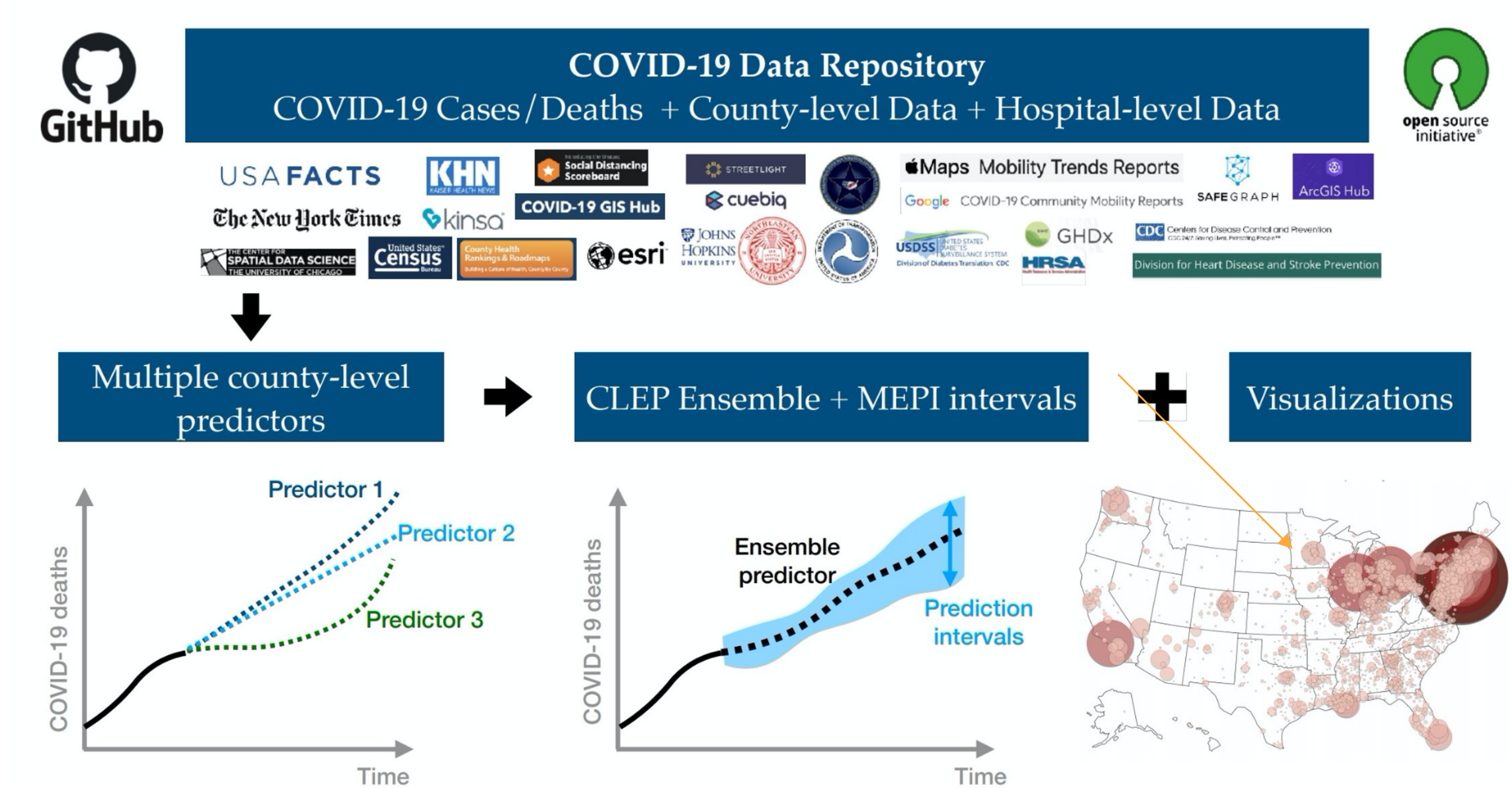
- April 11- May 10

- May 11- June 20

- April 11 - June 20
  (over selected days with deaths>10)



Overview: Current Data Repository & Prediction Pipeline (Open Source)



Covidseverity.com is an automated AI system

1. Data (daily county case and death numbers) from USAFacts is scrapped automatically to our AWS instance
2. Our CLEP prediction algorithm runs on updated data on AWS automatically (Thanks to AWS and NSF)
3. Predictions, prediction intervals, plots, and maps are generated and displayed automatically

This AI system could not spot that “1525” on May 21 for King County, WA was an error. Humans in the loop would be better.

Future of AI should be human-machine collaboration

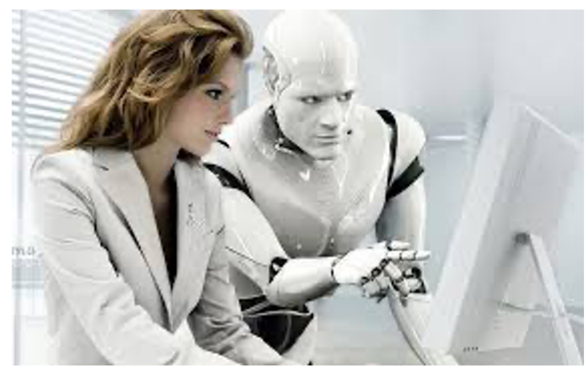
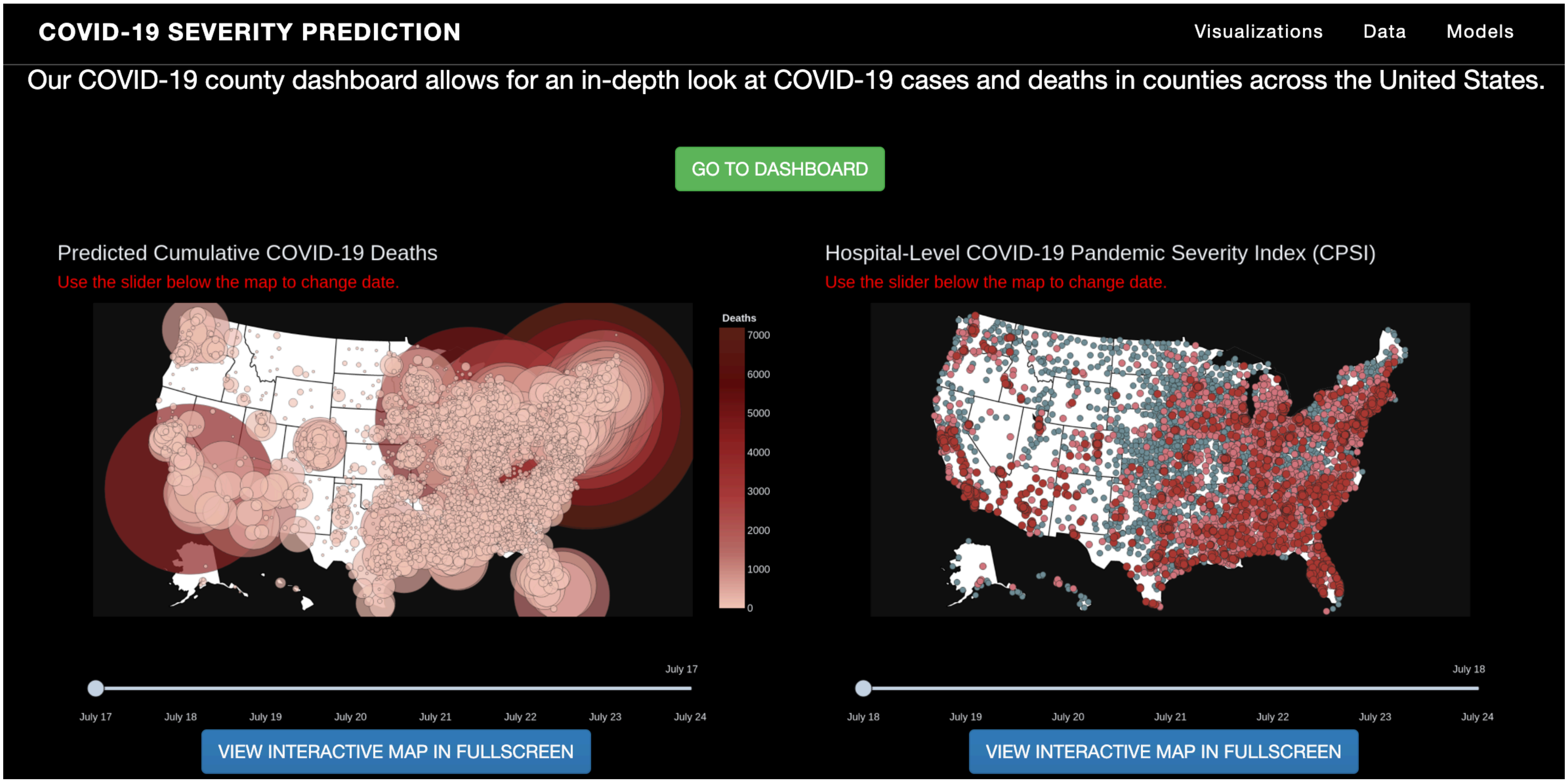


Image credit: trademed.com.

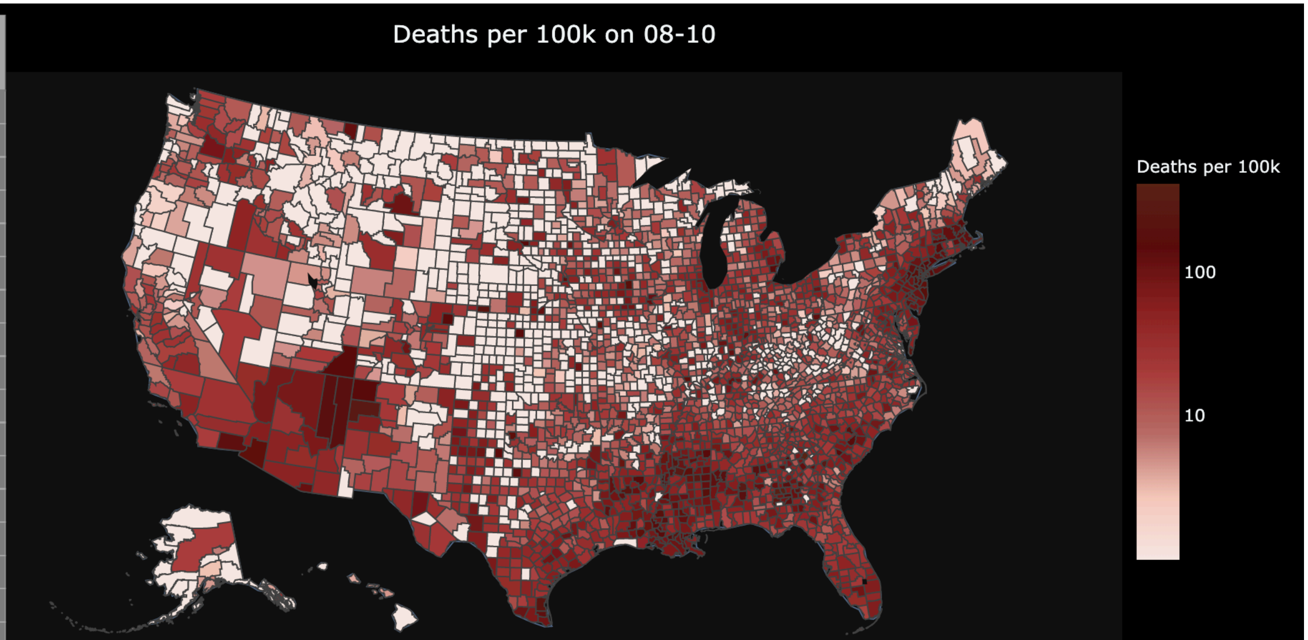
Data and code at **covidseverity.com (searchable by county)**



Ranking counties using 8 metrics

Cumulative Cases | Cumulative Deaths | New Cases | New Deaths | Cases per 100k | **Deaths per 100k** | New Cases per 100k | New Deaths per 100k

	County	Deaths per 100k
1	Hancock, GA	419.26
2	Randolph, GA	380.51
3	Galax City, VA	373.66
4	Terrell, GA	348.39
5	Bronx, NY	339.91
6	Neshoba, MS	315.88
7	Queens, NY	315.77
8	McKinley, NM	312.63
9	Early, GA	312.29
10	Emporia City, VA	292.91
11	Kings, NY	281.13
12	Holmes, MS	278.06
13	Jenkins, GA	276.4
14	Essex, NJ	263.95
15	Lowndes, AL	260.68
16	Northampton, VA	247.12
17	Passaic, NJ	246.77
18	Union, NJ	241.91
19	Perkins, NE	238.99
20	Turner, GA	227.5



State: Georgia

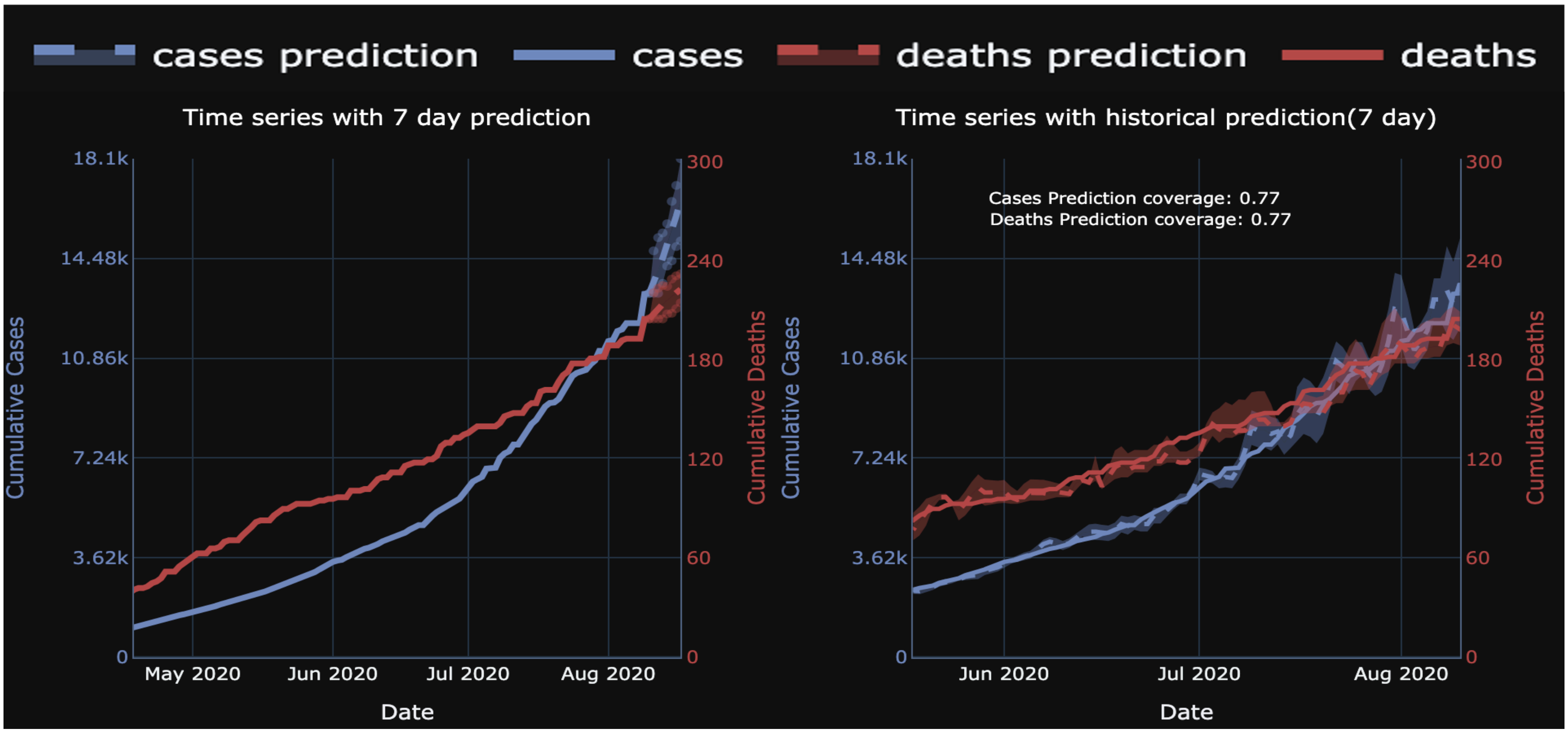
County: Hancock County

D. Wang

P. Norvig

Thanks to Google

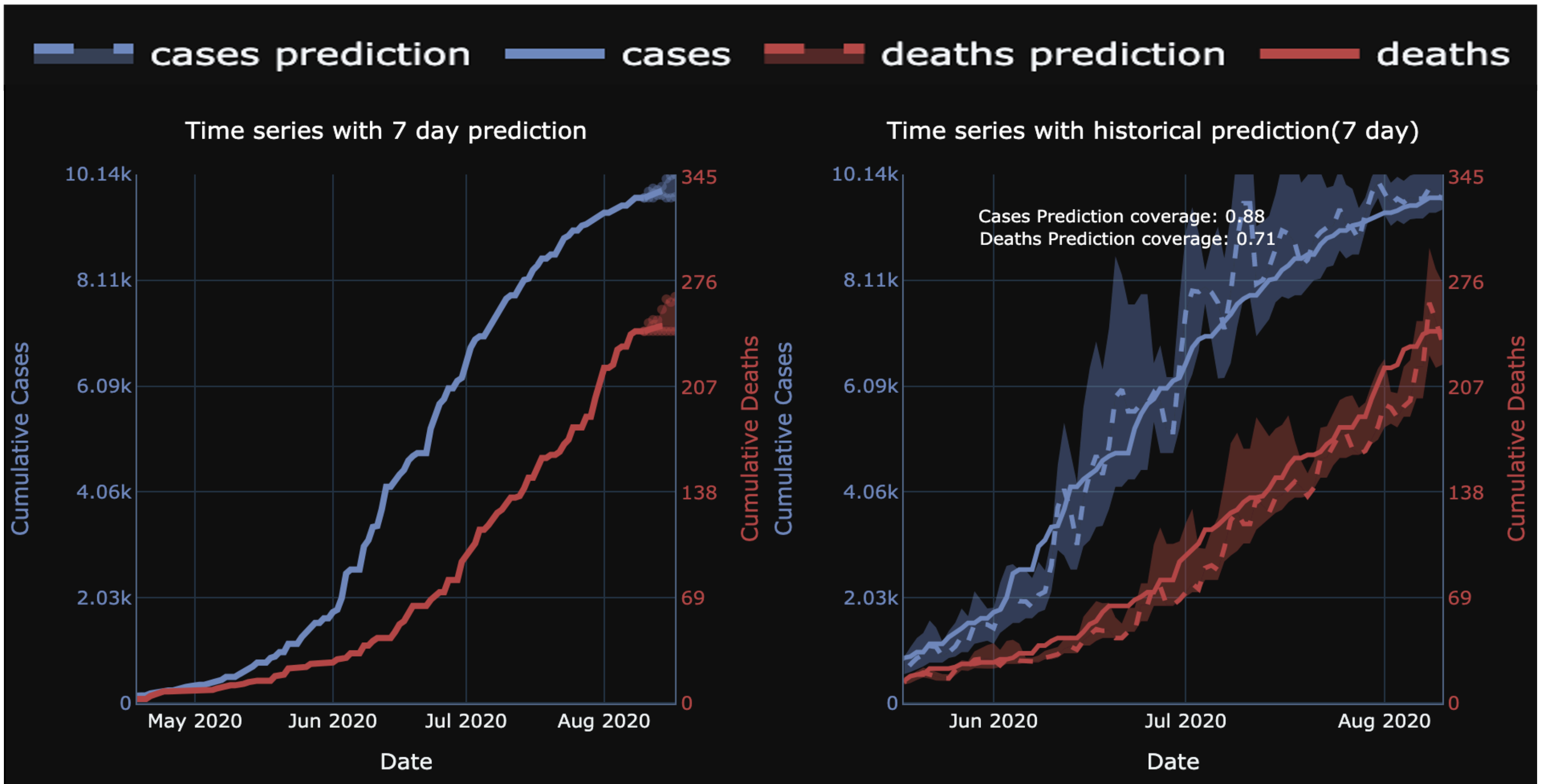
7-day prediction: Alameda County, CA (county search)



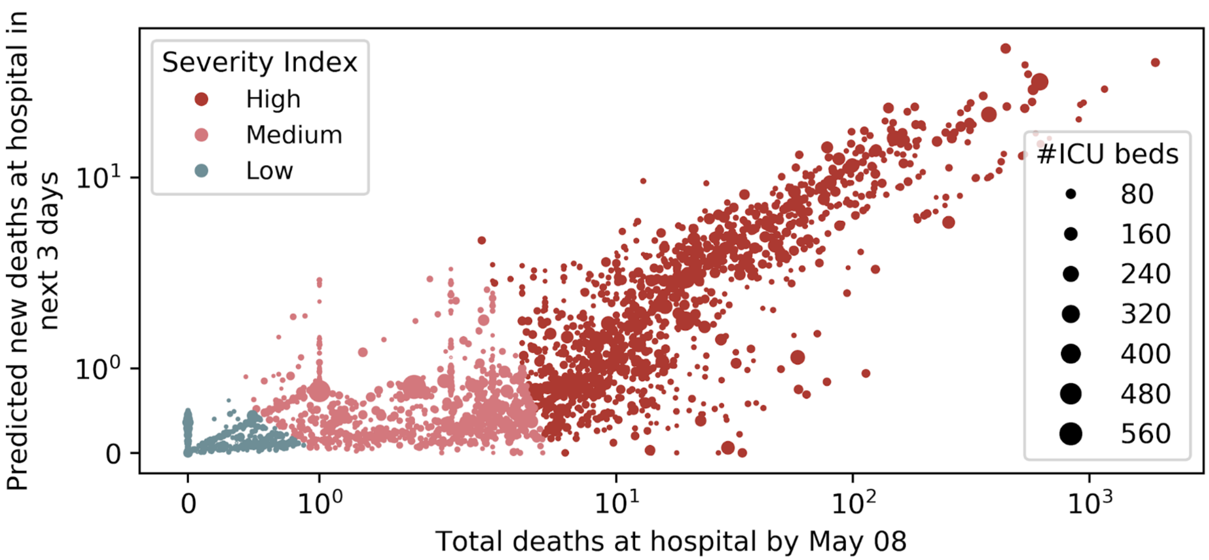
Cases/deaths

New cases/deaths

7-day prediction: Imperial County, CA (county search)



Severity Index to help PPE distribution at covidseverity.com



A score* for each hospital based on:

1. Predicted cumulative deaths

1. Predicted daily deaths

* county level predicted deaths are distributed to hospitals proportional to #employees

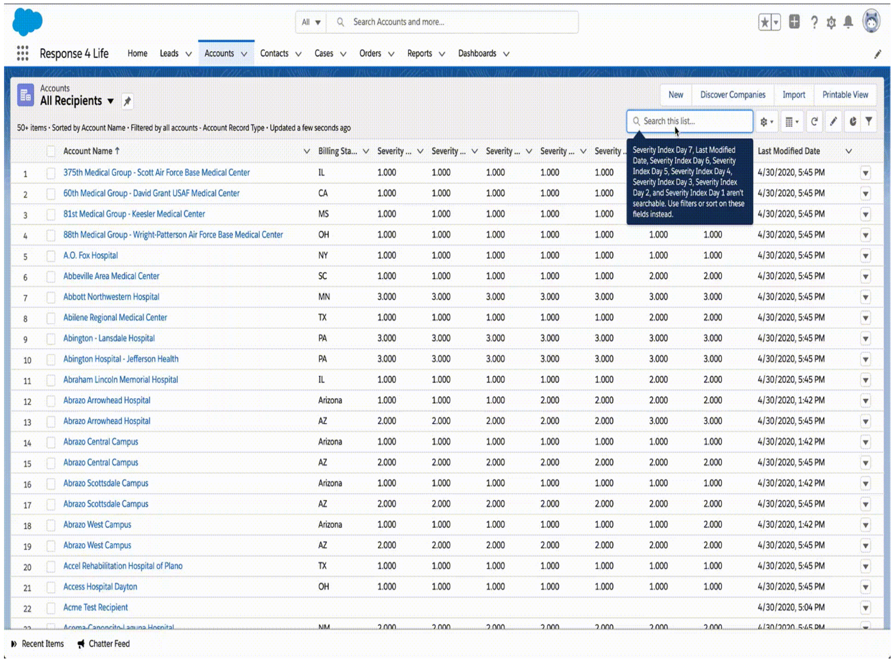
5000 Face Shields arrived at Temple Univ Hospital on May 8

D. Landwirth R. Brenan (both from R4L)

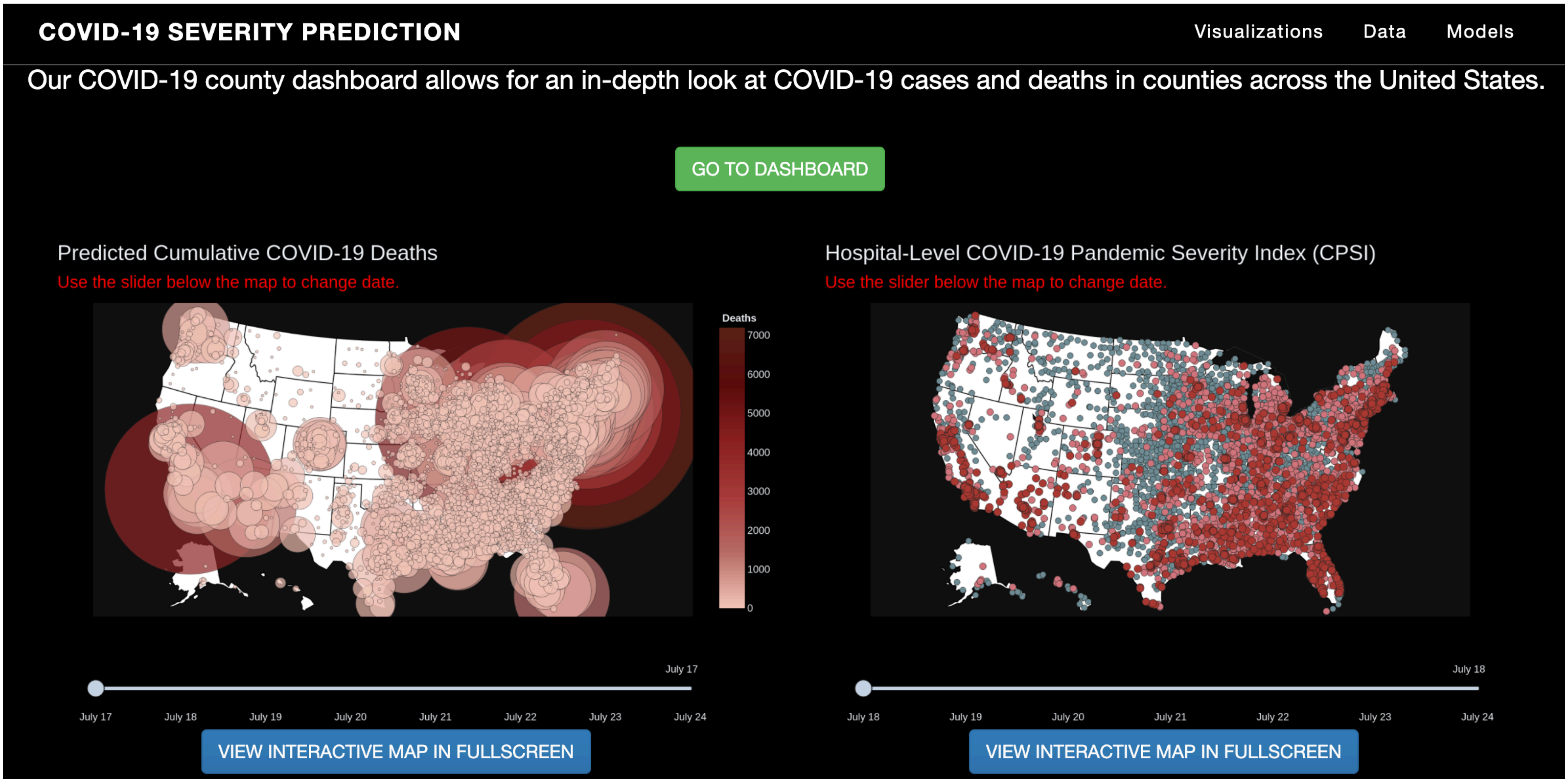
Impacts through Response4life

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- **+65k to 25 recipients in 15 states**

R4L is building a salesforce logistics system for supply chain that uses our **severity index**



Data and code at **covidseverity.com (searchable by county)**



Paper at https://arxiv.org/abs/2005.07882 and under revision for Harvard Data Science Review (HDSR)

CURATING A COVID-19 DATA REPOSITORY AND FORECASTING COUNTY-LEVEL DEATH COUNTS IN THE UNITED STATES

Nick Altieri[1], Rebecca L Barter[1], James Duncan[4], Raaz Dwivedi[2], Karl Kumbier[6], Xiao Li[1], Robert Netzorg[2], Briton Park[1], Chandan Singh[2, *], Yan Shuo Tan[1], Tiffany Tang[1], Yu Wang[1], Chao Zhang[3], Bin Yu[1, 2, 4, 5, 7, *]

[1] Department of Statistics, [2] Department of EECS, [3] Department of IEOR
[4] Division of Biostatistics, [5] Center for Computational Biology
University of California, Berkeley

[6] Department of Pharmaceutical Chemistry
University of California, San Francisco

[7] Chan Zuckerberg Biohub, San Francisco

CLEP and MEPI ideas are generally applicable

- CLEP weighting can be used to combine other predictors including those from agent based models.
- MEPI is agnostic to predictors (under e exchangeability)
- They can be applied to other time series data such as hospitalization

Summary

- Data repository a popular resource for other covid-19 activities

  In a period of two weeks, 12K visits with 1.1K unique visitors; 108 clones with 53 unique cloners

- CLEP and MEPI: transparent, and fast, generally applicable to other series (under exchangeability of recent prediction errors for MEPI)

- Continued support to Response4Life

- Results and blog on CSDS atlas at Univ of Chicago

Current directions

- **Our CLEP is at CDC forecast hub** https://covid19forecasthub.org/

- **Hospitalization prediction**)

- **Adaptive tuning** of CLEP

- **Causal investigation (e.g. impact of social distancing; matching of counties)**

Question for the experts

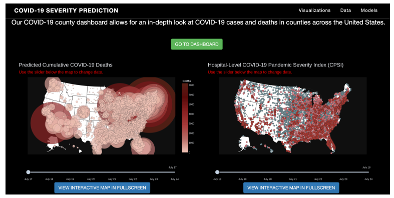
How can ML short-term predictions help with agent-based models?

Thank you!

Data and code at

github.com/Yu-Group/covid19-severity-prediction

Visualization at covidseverity.com



Paper at https://arxiv.org/abs/2005.07882