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

IAS Virtual Event Series
June 25, 2020

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

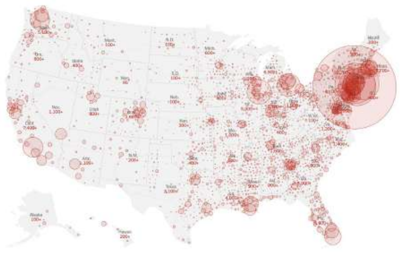
M. Shen

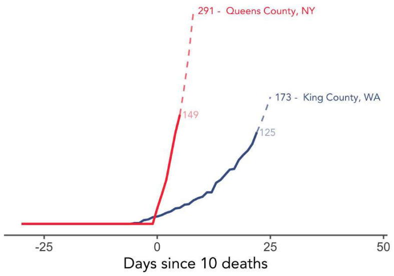
C. Zhang

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

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

Initial Goal: Help Aid Resource Allocation





Data Curation

- Hospital data
- County data

Modeling

- County-level 7-day severity prediction
- hospital demand prediction

Evaluation / Visualization

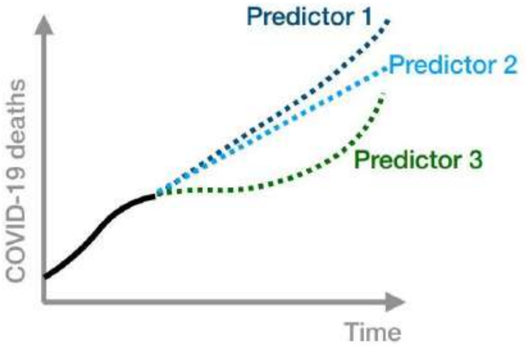
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts

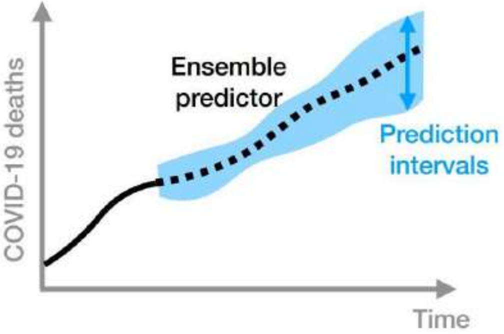
Overview: Current Data Repository & Prediction Pipeline (Open Source)

COVID-19 Data Repository

COVID-19 Cases / Deaths + County-level Data + Hospital-level Data

Multiple county-level predictors → CLEP Ensemble + MEPI intervals + Visualizations



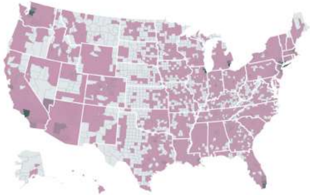


Curating a COVID-19 Data Repository

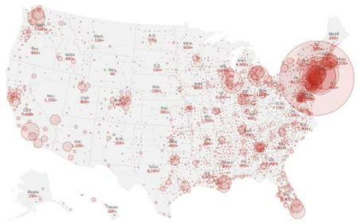
Data curation: scraped from a variety of sources

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

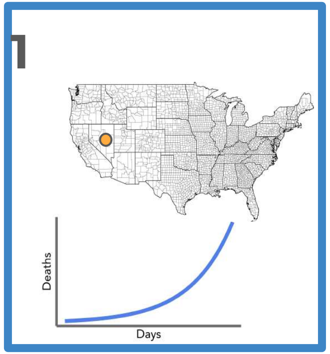
Forecasting county death counts

Curses and blessings

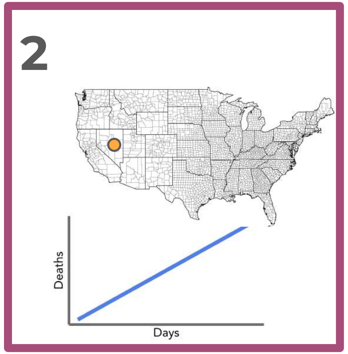
- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

- 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)

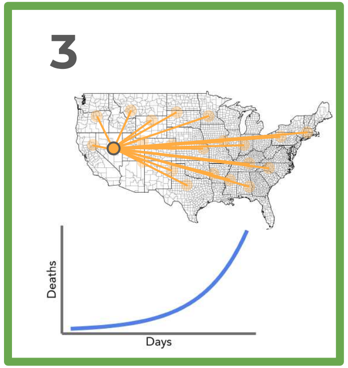
Combined Linear and Exponential Predictors (CLEP)



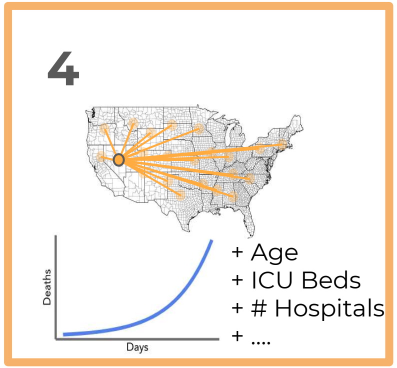
Separate-county exponential predictor



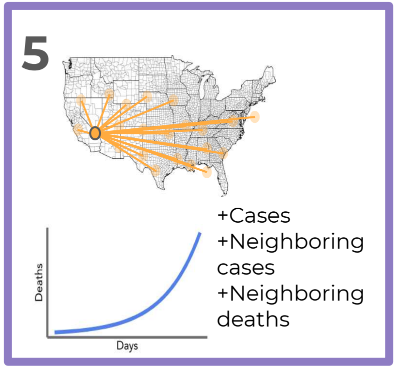
Separate-county linear predictor



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]

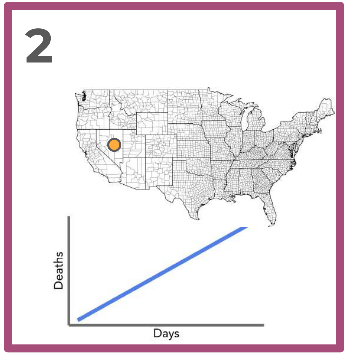
$$w_t^m \propto \exp\left(-c(1-\mu)\sum_{i=t_0}^{t-1}\mu^{t-i}\ell(\widehat{y}_i^m, y_i)\right)$$

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.

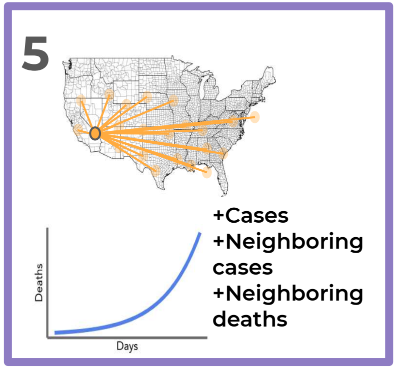
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two predictors performs well



Separate-county linear predictor

+



Expanded Shared-county

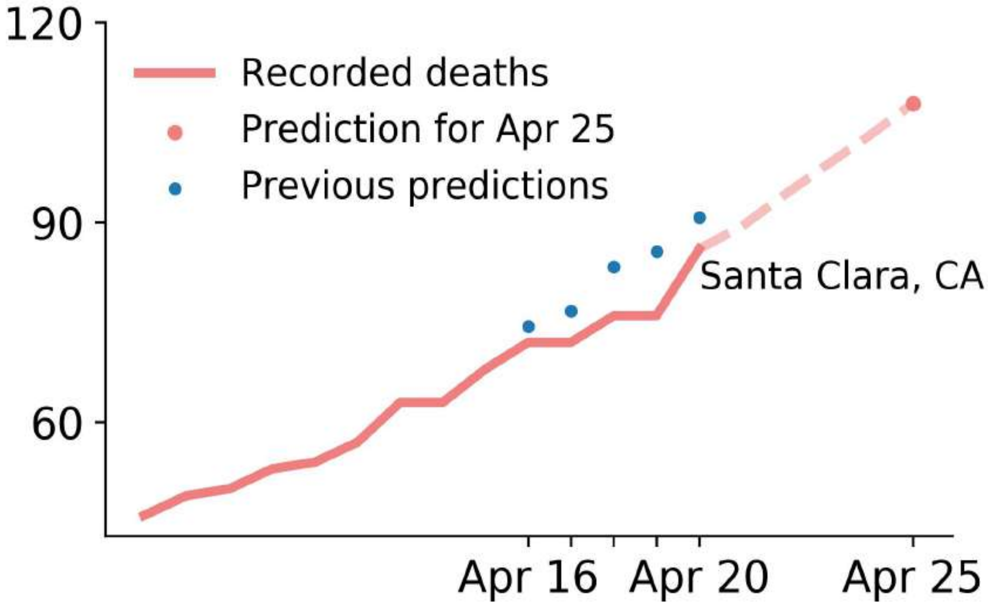
k=7 for 7-day prediction

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

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.

Prediction Intervals based on conformal prediction[2]



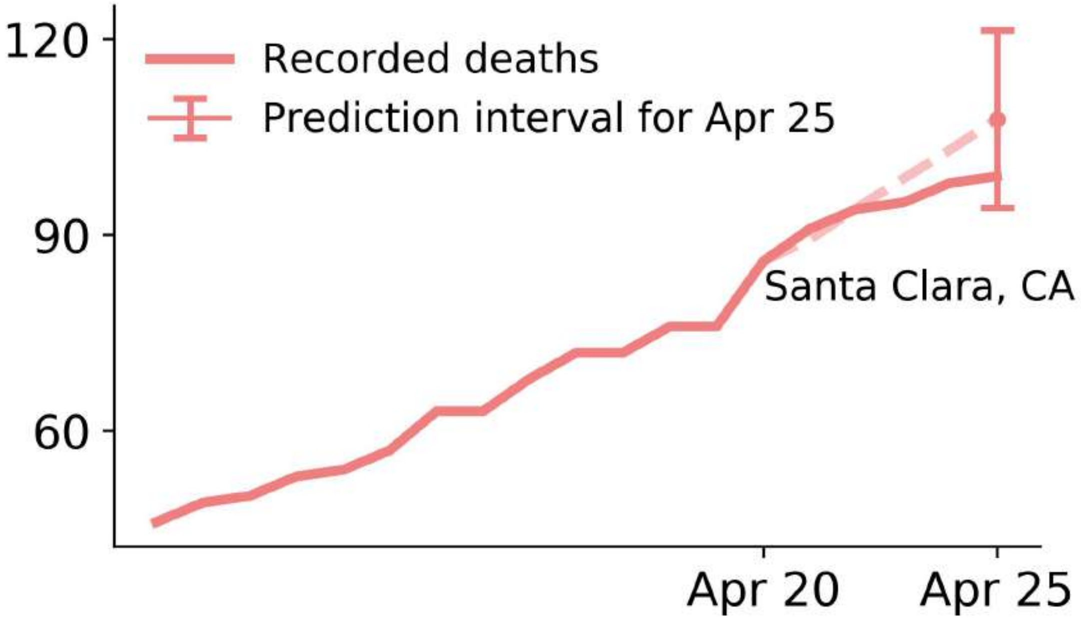
Previous 5-day-ahead rel. prediction errors (%)

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:



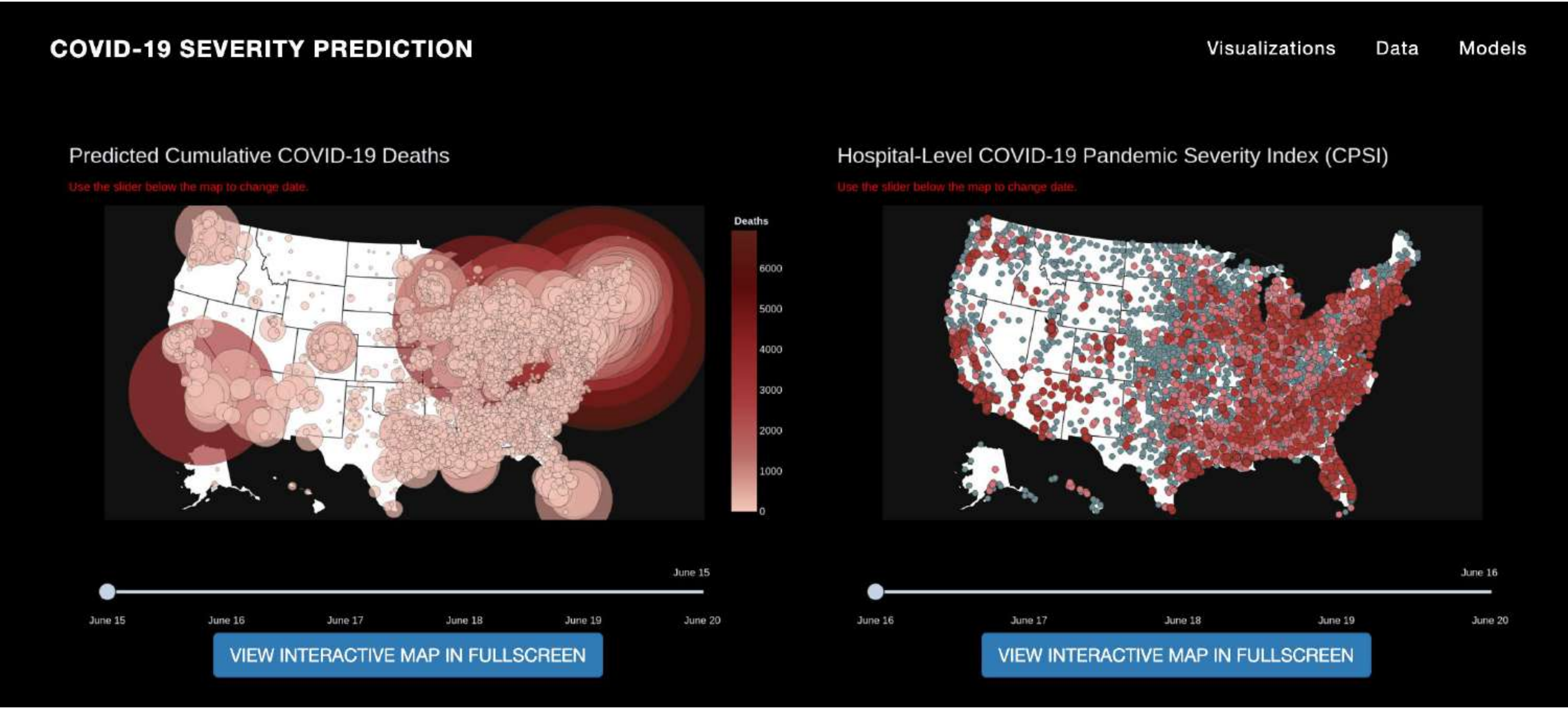
Predicted range of error

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

Actual error:

Apr 25 8.8%

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



Covidseverity.com is an automated AI system

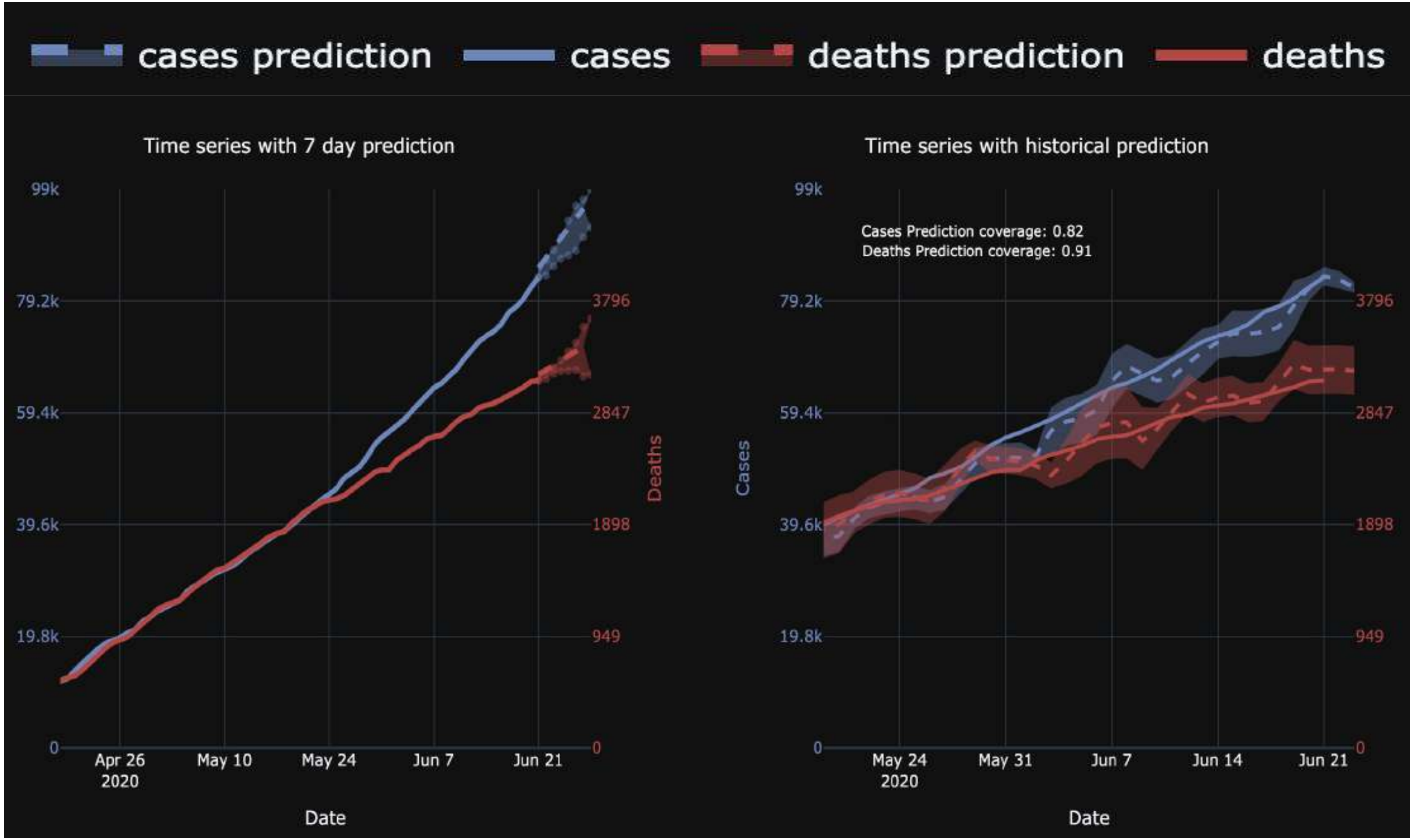
1. Data (daily county case and death numbers) from USAFacts is scrapped automatically to our AWS instance
2. Our CLEP prediction algorithm runs on updated data on AWS automatically (Thanks to AWS and NSF)
3. Predictions, prediction intervals, plots, and maps are generated and displayed automatically

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.

7-day prediction: LA county (new county search function)

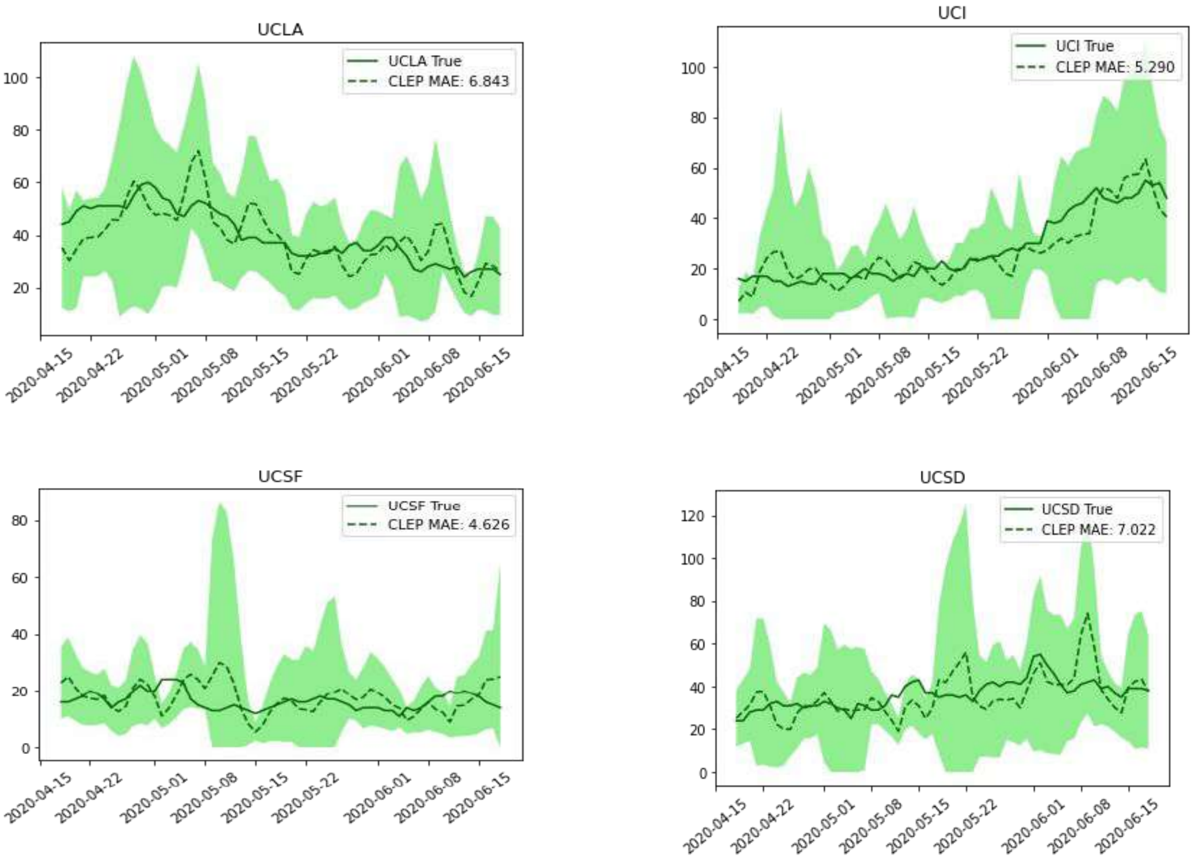


D. Wang

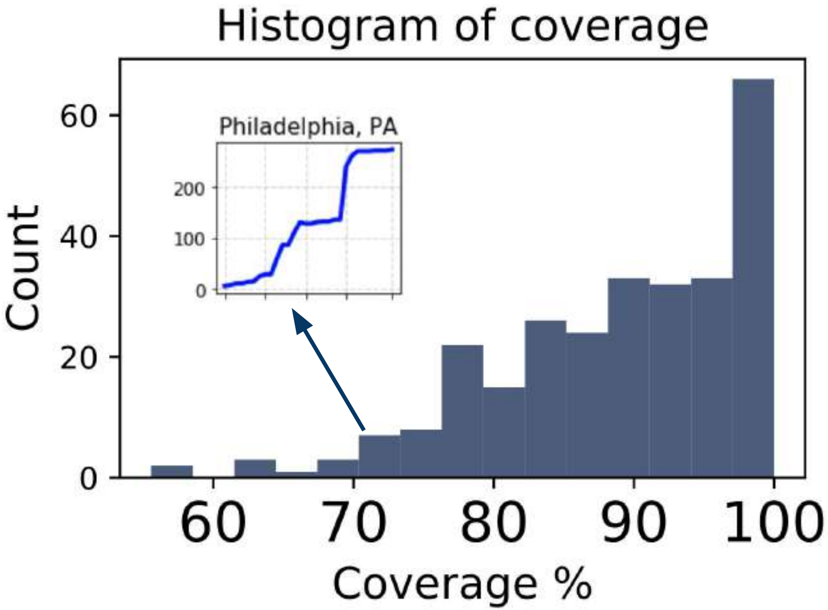
P. Norvig

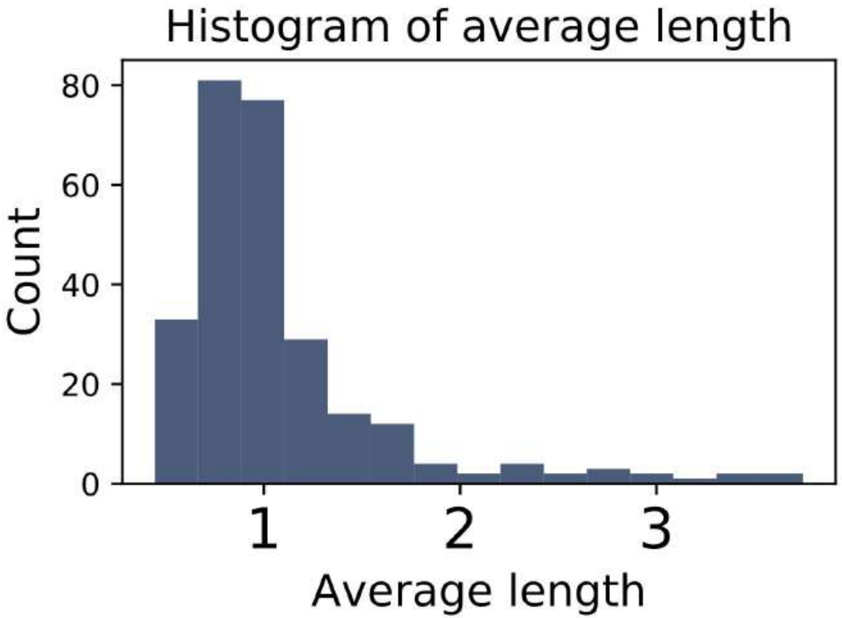
Thanks to Google

CLEP works also for predicting hospitalization for UC hospitals



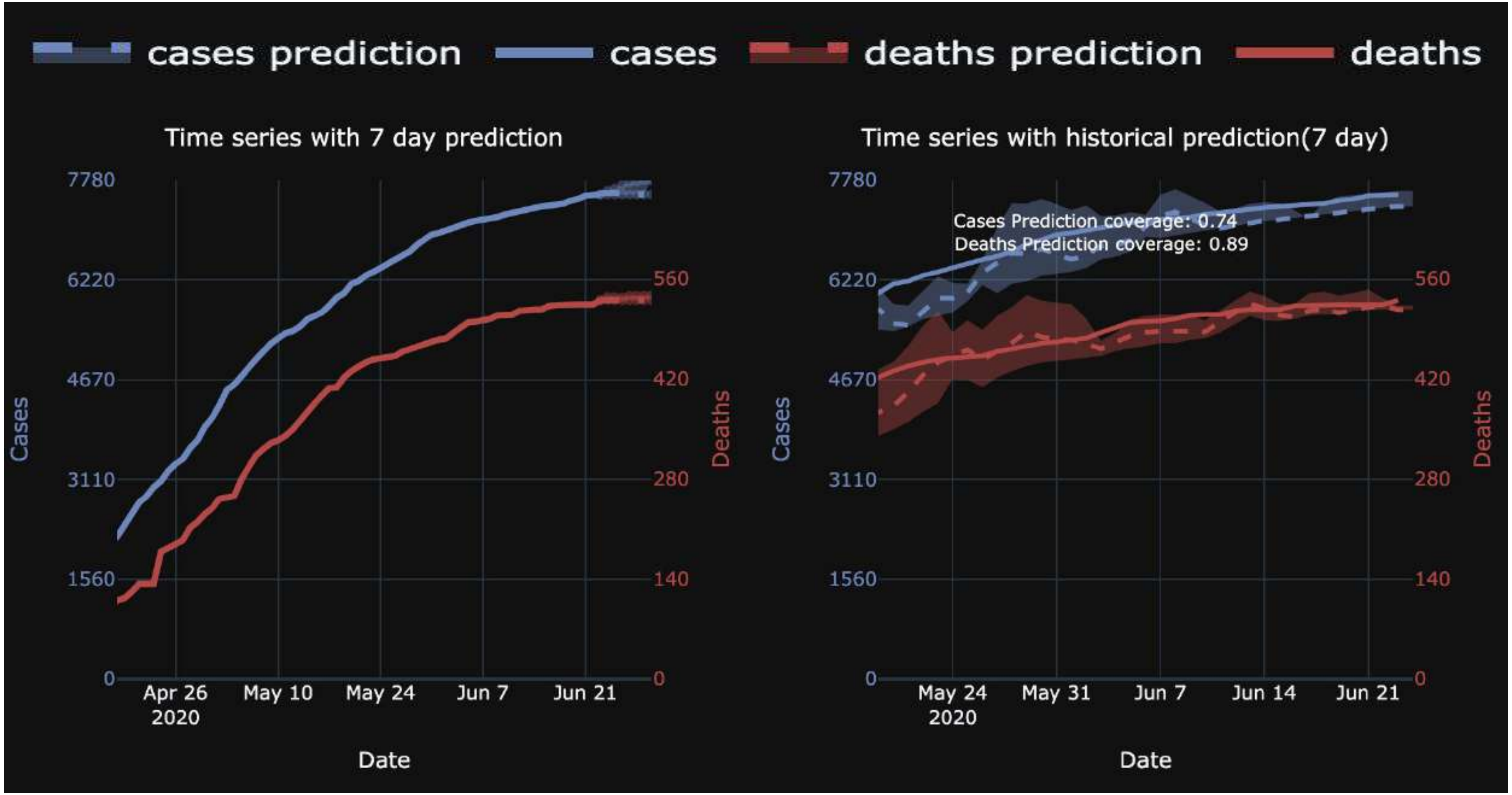
Empirical performance of MEPI for death counts



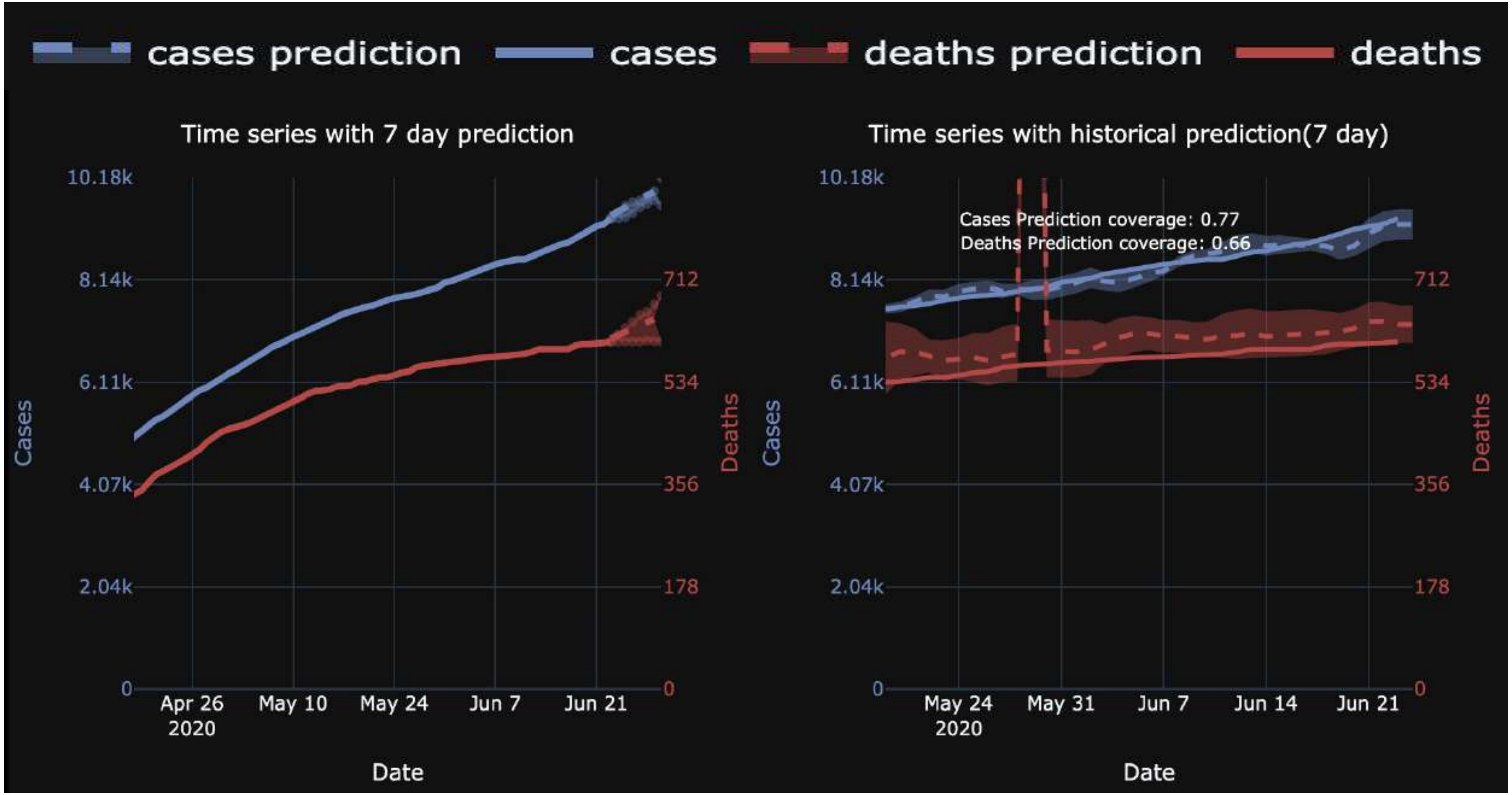


Evaluation period: March 28--April 27. Only include days since the county has 10 deaths.
Having a normalized length of 0.8 means the PI is roughly (0.6 $\widehat{y}_{t+k}$, 1.4 $\widehat{y}_{t+k}$).

7-day prediction: Mercer county (Princeton), NJ

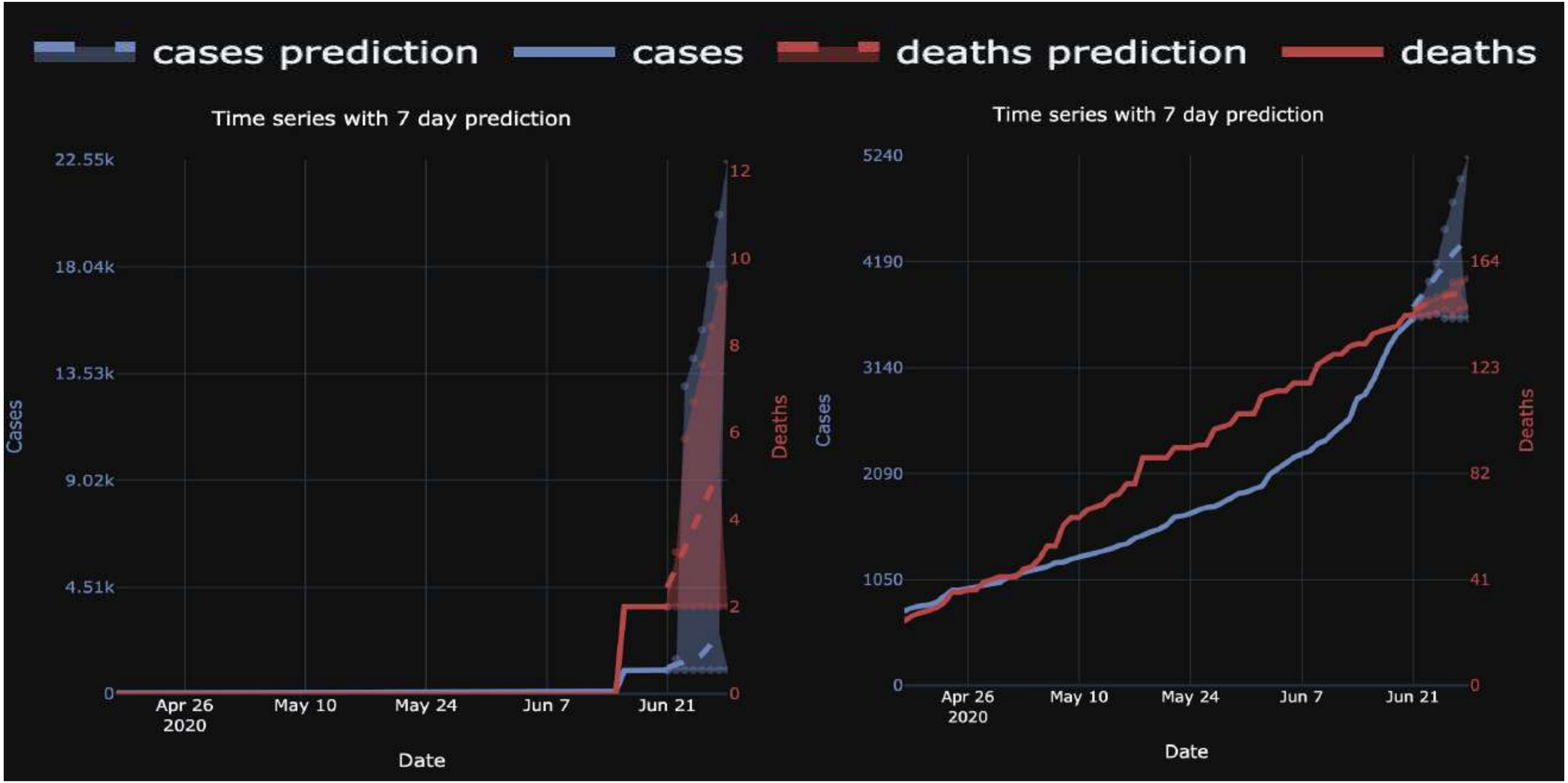


7-day prediction: King county (Seattle), WA



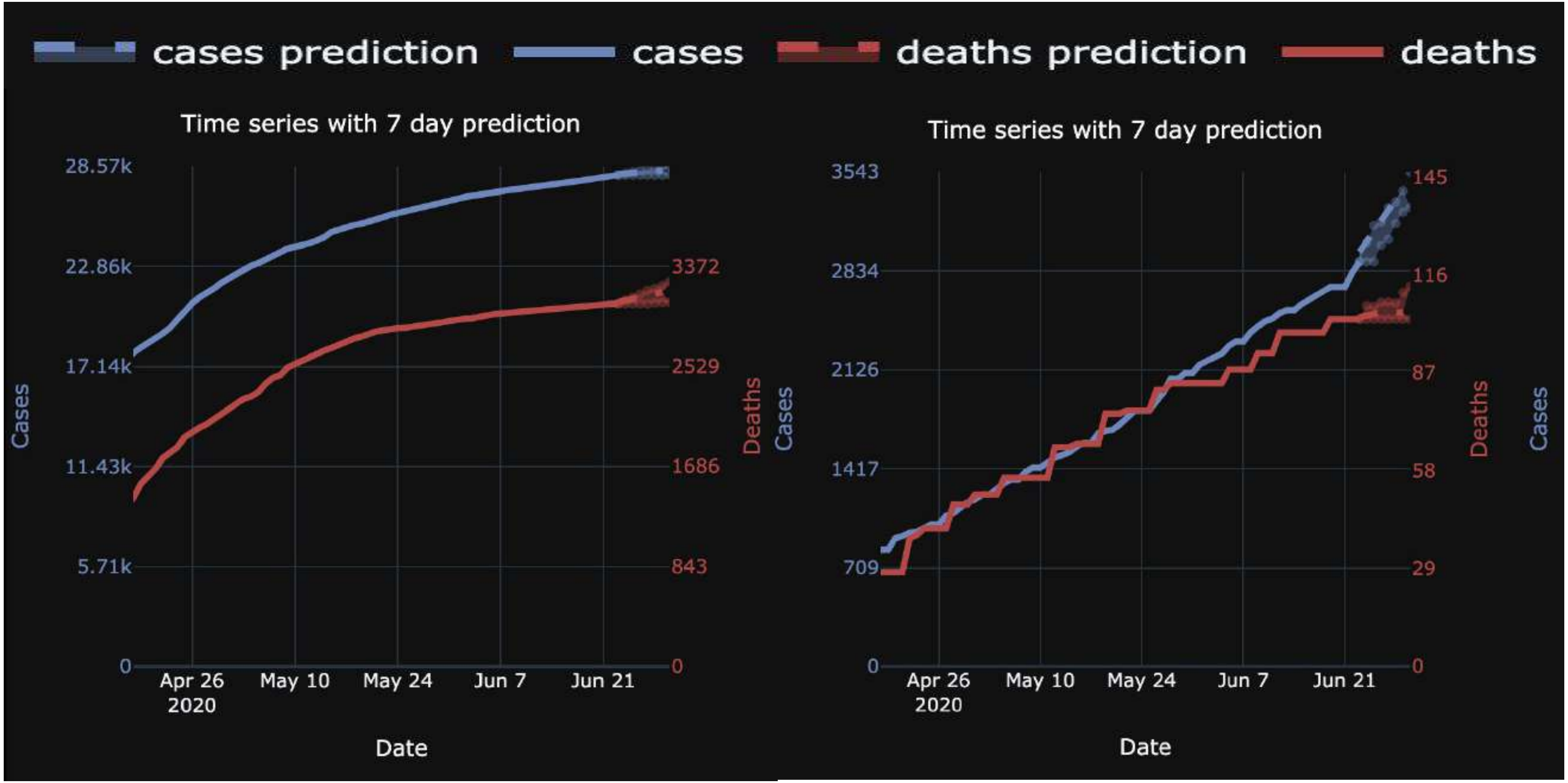
High case growth
Anderson County in TX

High death growth
Lee County in FL



Manhattan

San Mateo, CA



Covidseverity.com is an automated AI system

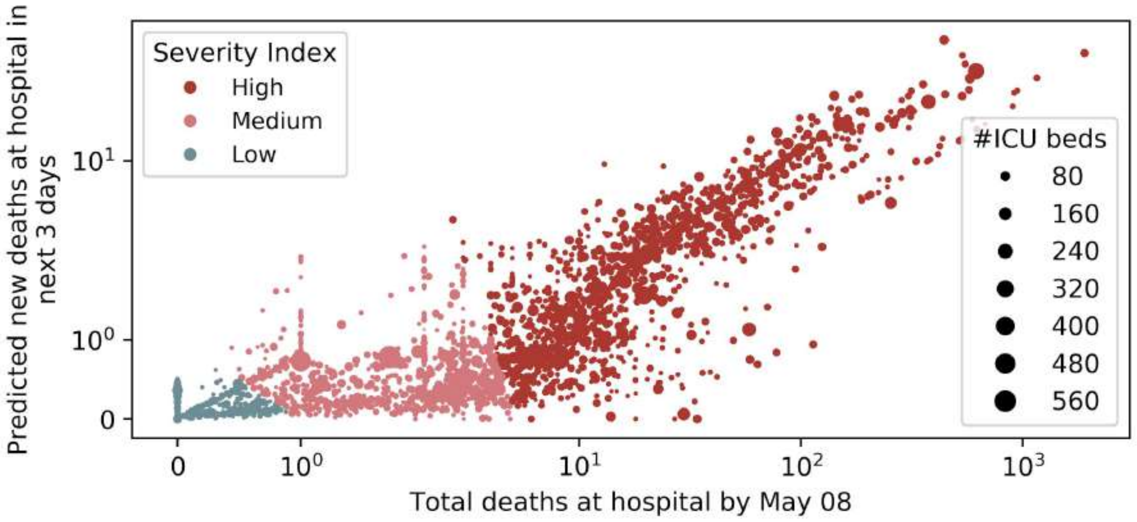
1. Data (daily county case and death numbers) from USAFacts is scraped automatically to our AWS instance
2. Our CLEP prediction algorithm runs on updated data on AWS automatically (Thanks to AWS and NSF)
3. Predictions, prediction intervals, plots, and maps are generated and displayed automatically

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.

Severity Index at covidseverity.com



A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

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

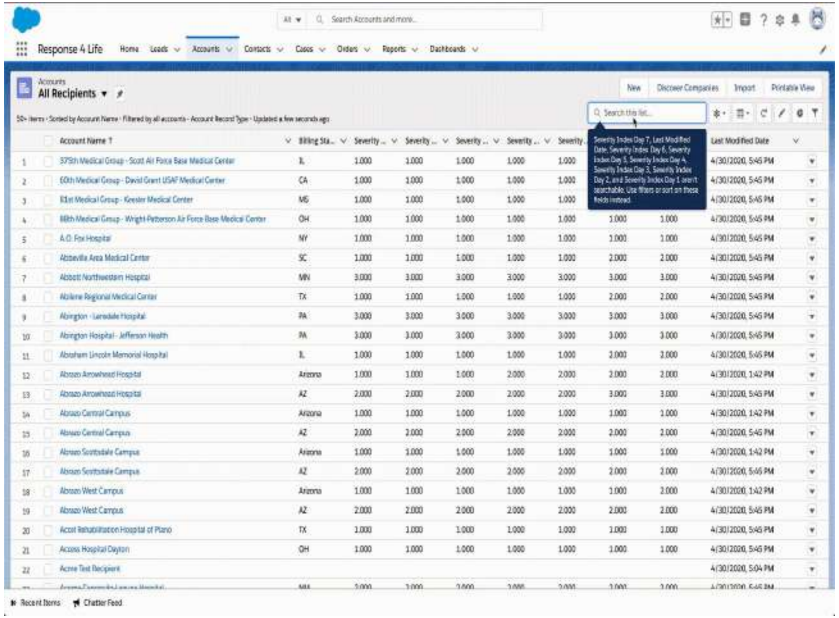
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L

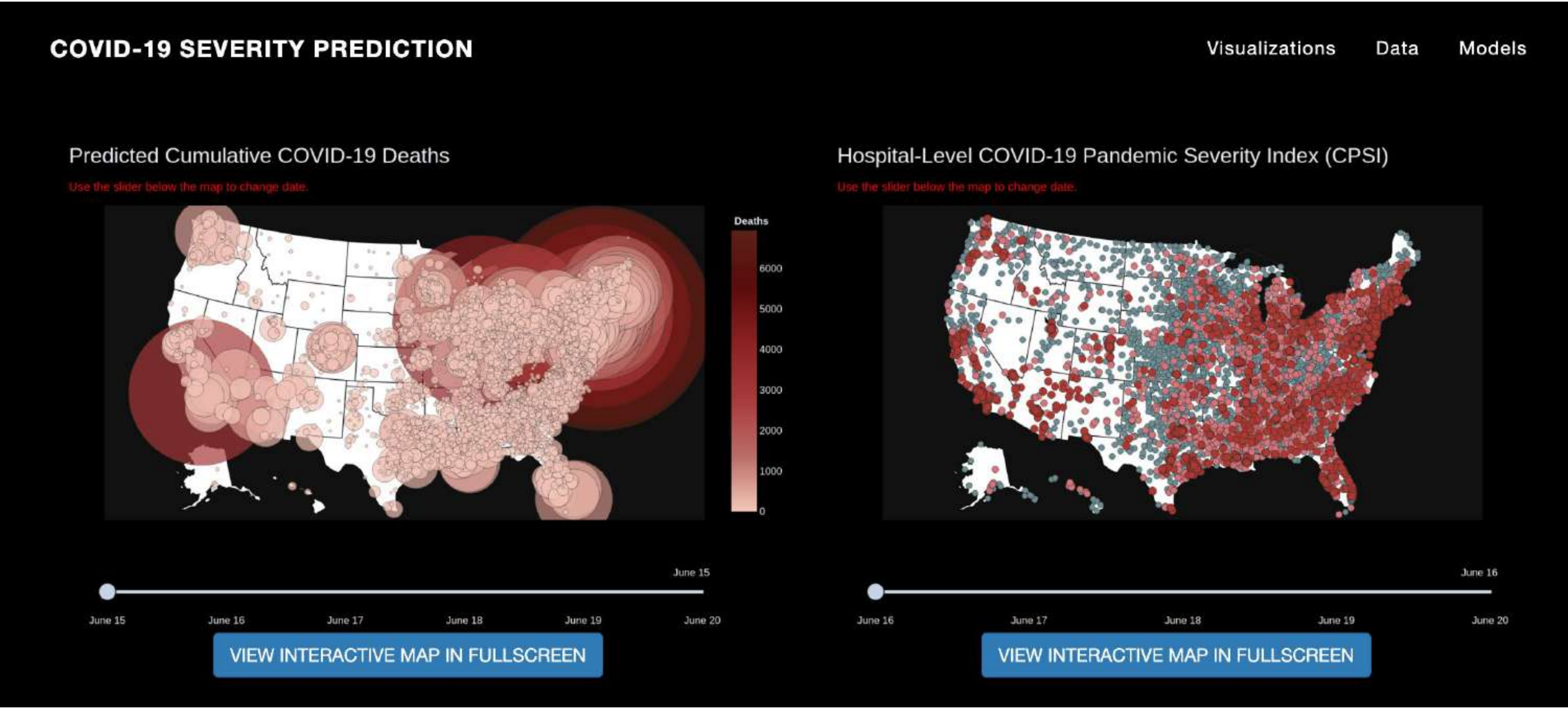
Impacts through Response4life

- 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[6], Raaz Dwivedi[2], Karl Kumbier[3], Xiao Li[1], Robert Netzorg[2], Briton Park[1], Chandan Singh*[2], Yan Shuo Tan[1], Tiffany Tang[1], Yu Wang[1], Bin Yu*[1, 2, 4, 5, 6]

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

May 19, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

Current directions

- 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

- Continued support to Response4Life
- Results and blog on CSDS atlas at Univ of Chicago
- **Hospitalization prediction** in collaboration with google (and possible collaboration with California Department of Public Health and Microsoft)
- **Causal investigation (e.g. impact of social distancing; matching of counties)** (beginning)

Thank you!

Any questions?

Please visit covidseverity.com

COVID-19 Data Repository and County Death Count Prediction

Bin Yu
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

Incremental Causal Effects

Dominik Rothenhaeusler
Stanford Statistics

ONR PI Meeting
June 24, 2020

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. Shenl

C. Zhang

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

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

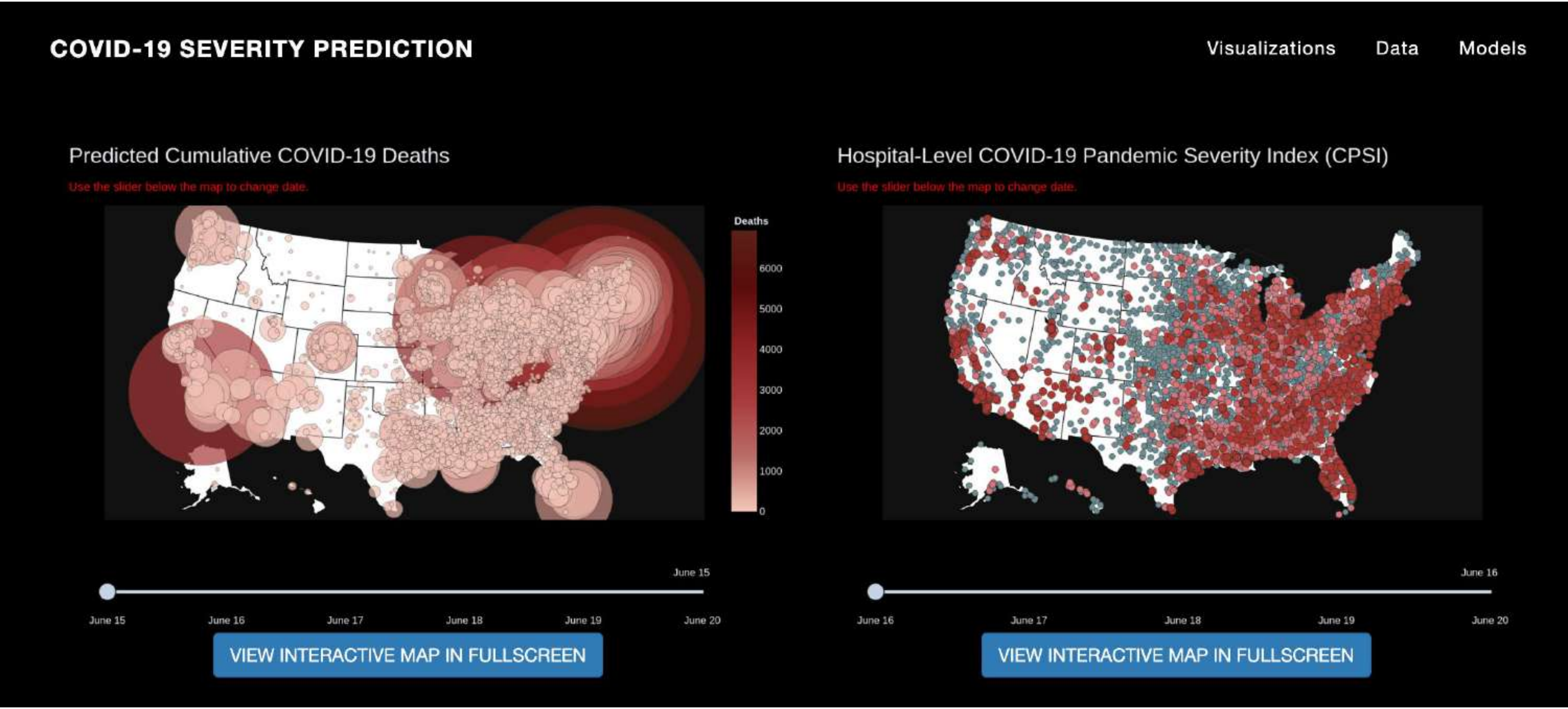
Nick Altieri[1, †], Rebecca L Barter[1], James Duncan[6], Raaz Dwivedi[2], Karl Kumbier[3], Xiao Li[1], Robert Netzorg[2], Briton Park[1], Chandan Singh*[2], Yan Shuo Tan[1], Tiffany Tang[1], Yu Wang[1], Bin Yu*[1, 2, 4, 5, 6]

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

May 19, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

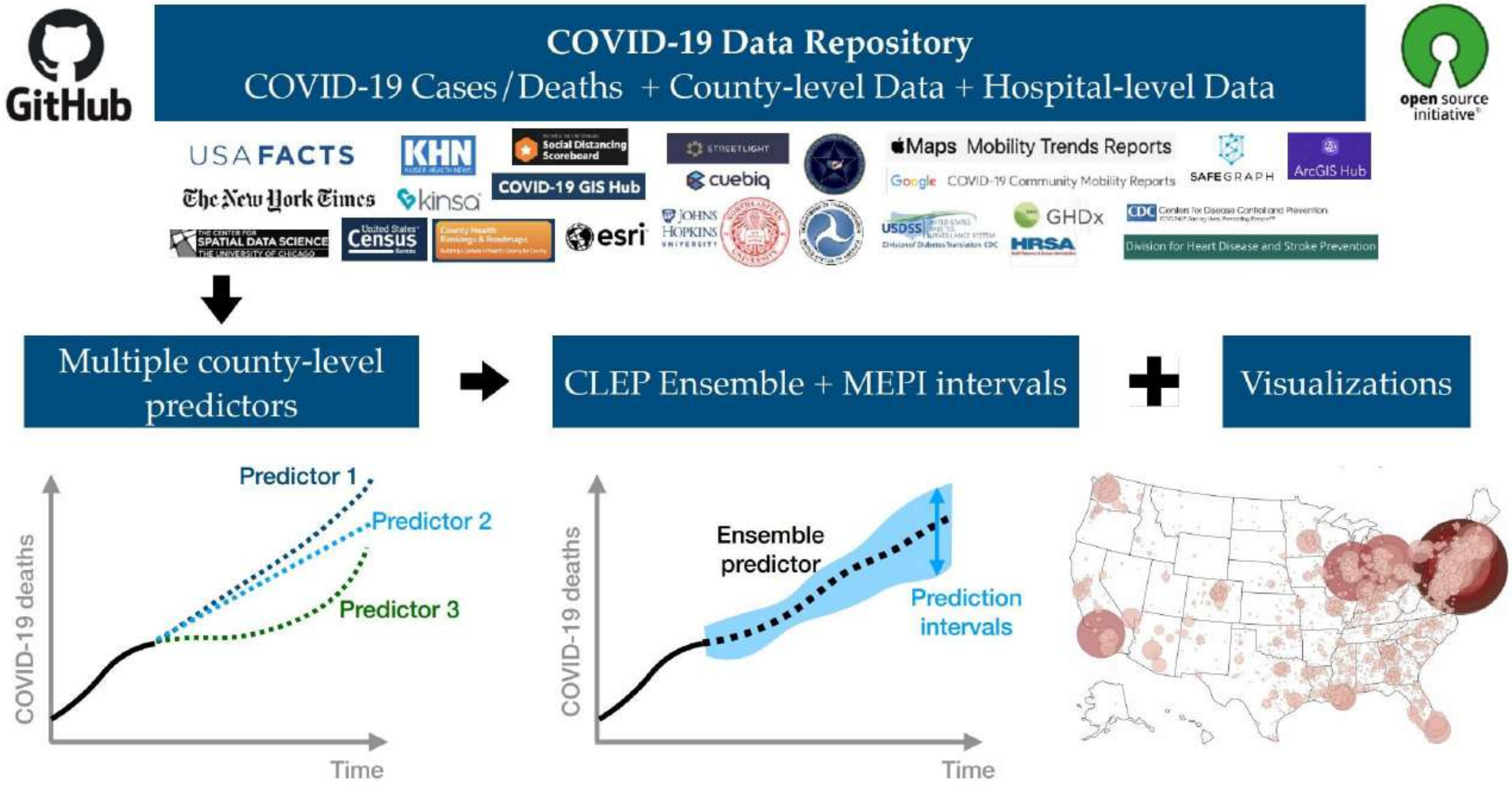
Data and code at **covidseverity.com (searchable by county)**



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

Initial Goal: Help Aid Resource Allocation

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

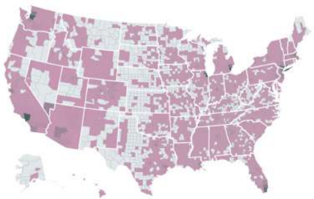


Curating a COVID-19 Data Repository

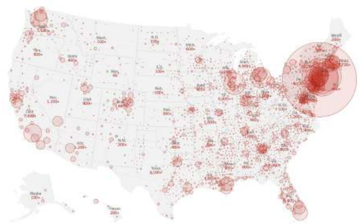
Data curation: scraped from a variety of sources

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

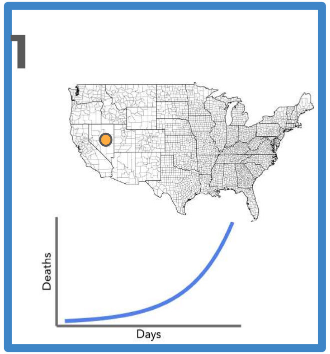
Forecasting county death counts

Curses and blessings

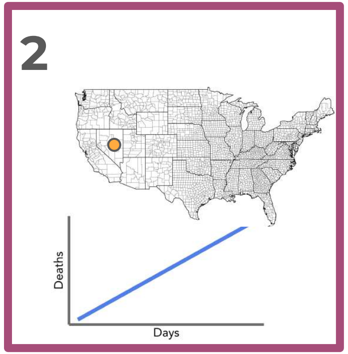
- Very dynamic data
- Long-term predictions have to deal with feedback
- We want to predict for all 7000 counties in the US because of R4L

- 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)

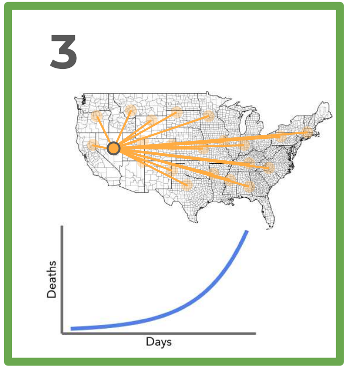
Combined Linear and Exponential Predictors (CLEP)



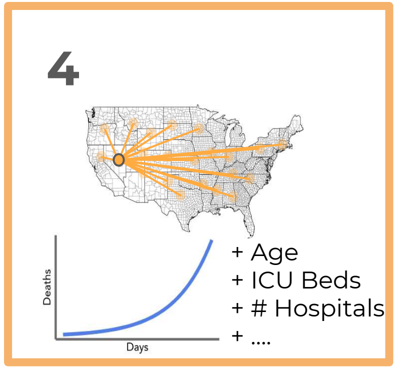
Separate-county exponential predictor



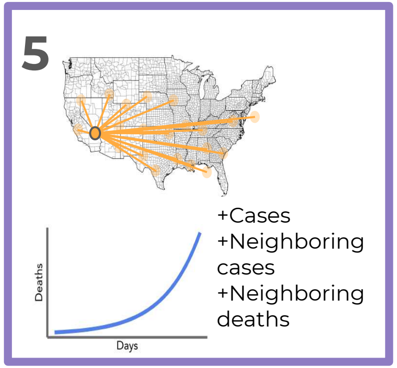
Separate-county linear predictor



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]

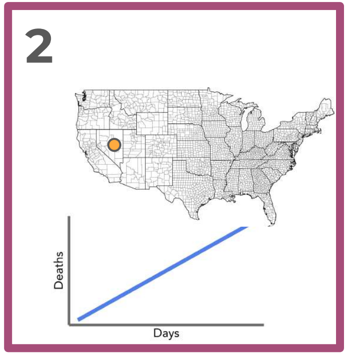
$$w_t^m \propto \exp\left(-c(1-\mu)\sum_{i=t_0}^{t-1}\mu^{t-i}\ell(\widehat{y}_i^m, y_i)\right)$$

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.

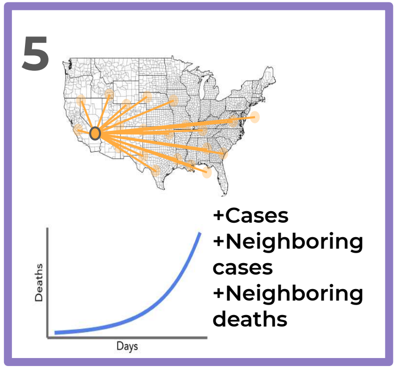
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two predictors performs well



Separate-county linear predictor

+



Expanded Shared-county

Elastic-net regularized Poisson GLM

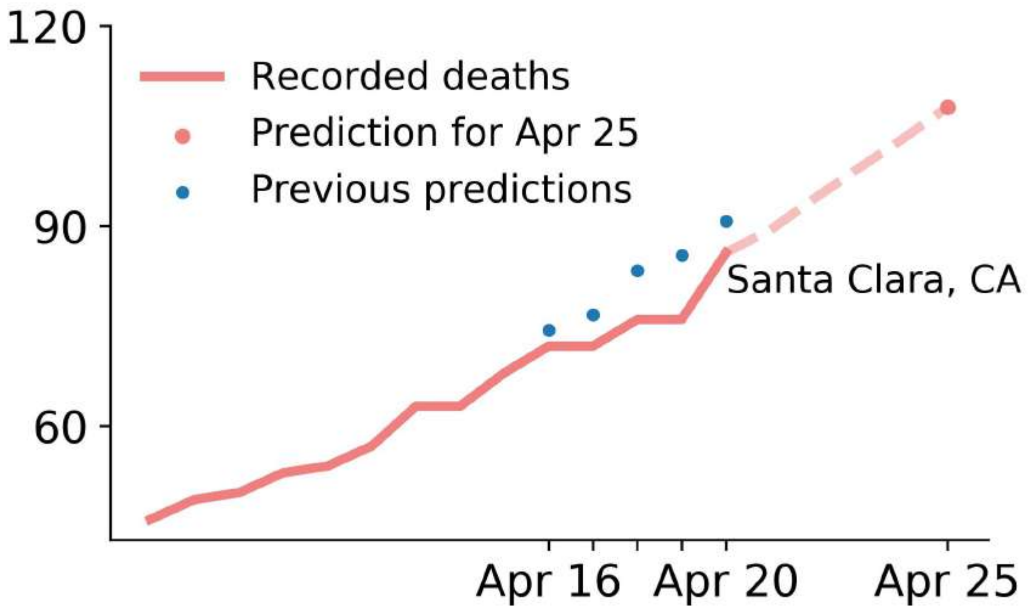
k=7 for 7-day prediction

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

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.

Prediction Intervals based on conformal prediction[2]



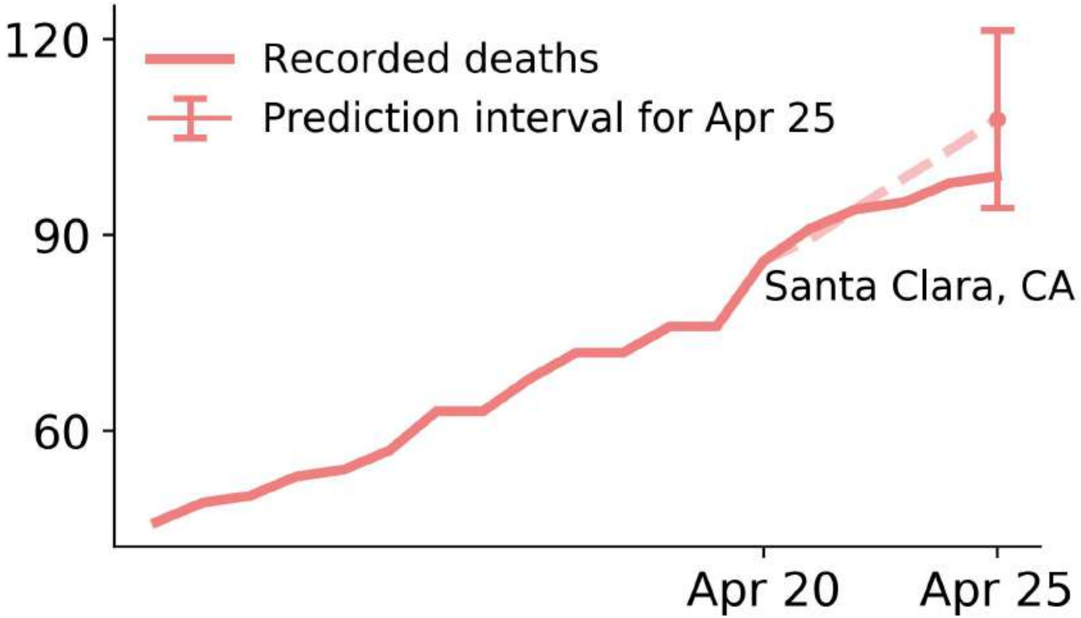
Previous 5-day-ahead prediction errors (%)

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%

Maximum (absolute) error prediction intervals (MEPI)

Step 1

Find normalized error of our predictor in the past. $\Delta_\tau := |y_\tau - \widehat{y}_\tau| / |\widehat{y}_\tau|.$

Step 2

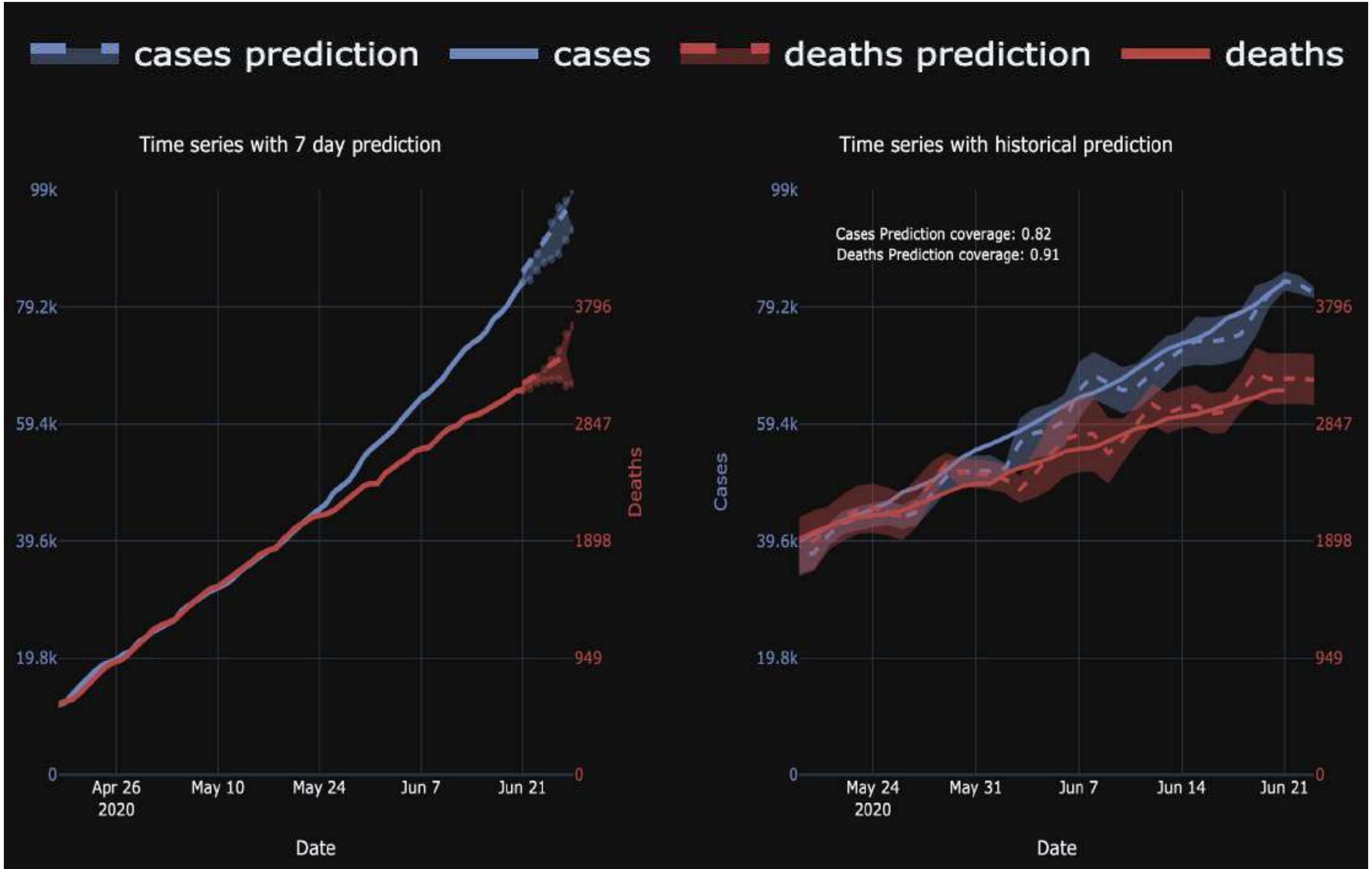
Find maximum error of past 5 days. $\Delta_{\max} := \max_{0 \le j \le 4} \Delta_{t-j}.$

Step 3

$$\widehat{\mathrm{PI}}_{t+k} := \left[\max\left\{\widehat{y}_{t+k}(1-\Delta_{\max}), y_t\right\},\ \widehat{y}_{t+k}(1+\Delta_{\max})\right]$$

Can be applied to any ML model, and it works well under **exchangeability** condition on the errors.

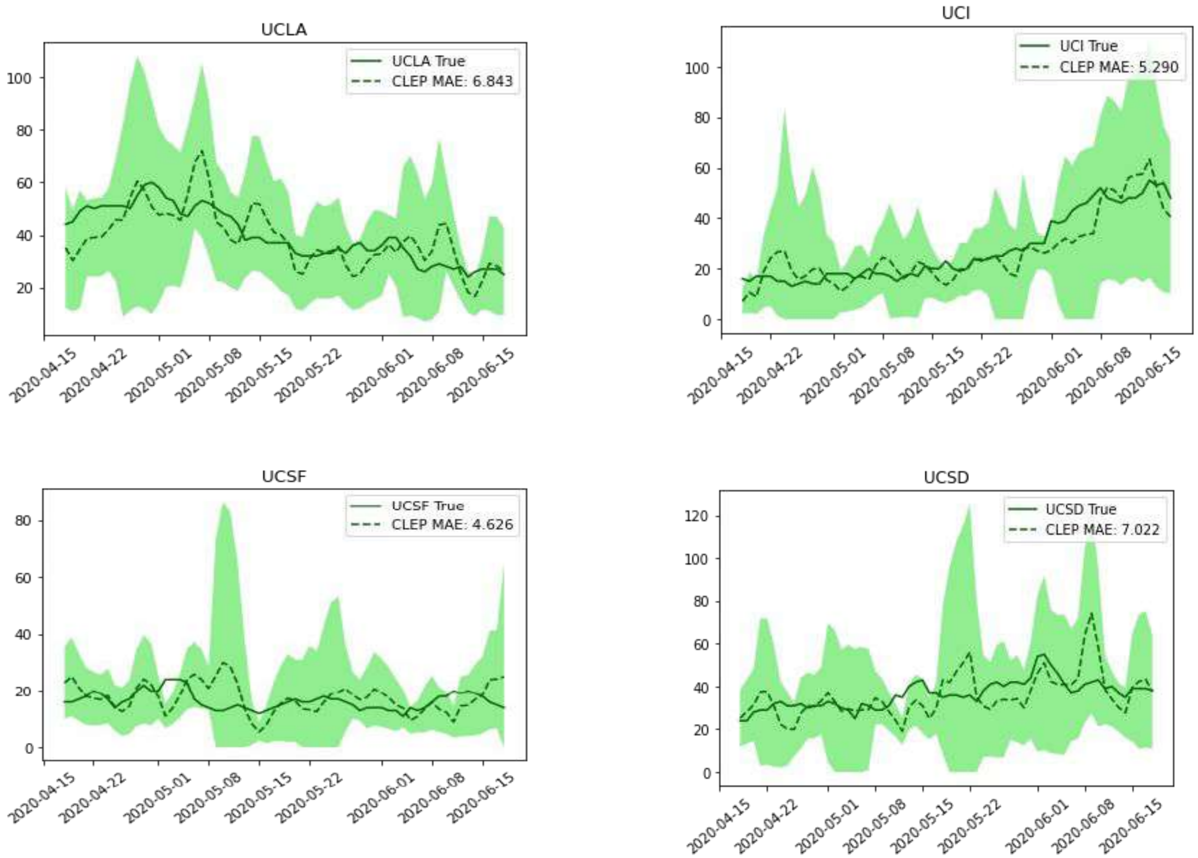
7-day prediction: LA county (new at covidseverity.com)



D. Wang

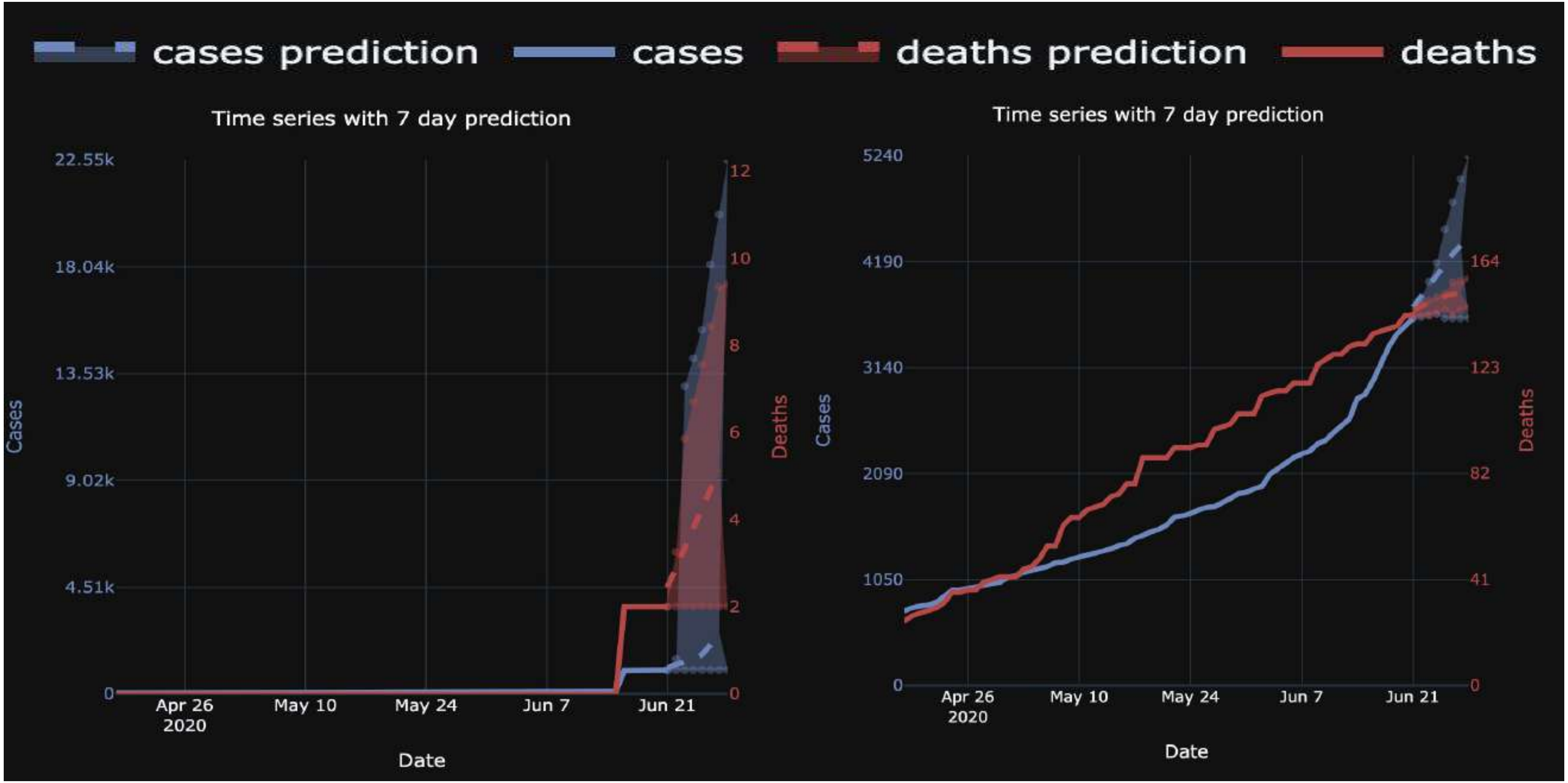
P. Norvig

CLEP works also for predicting hospitalization for UC hospitals

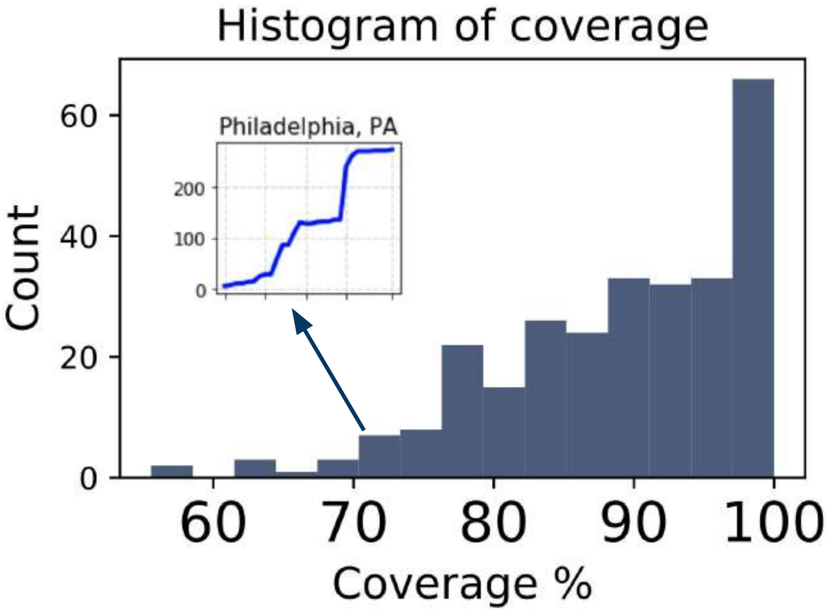


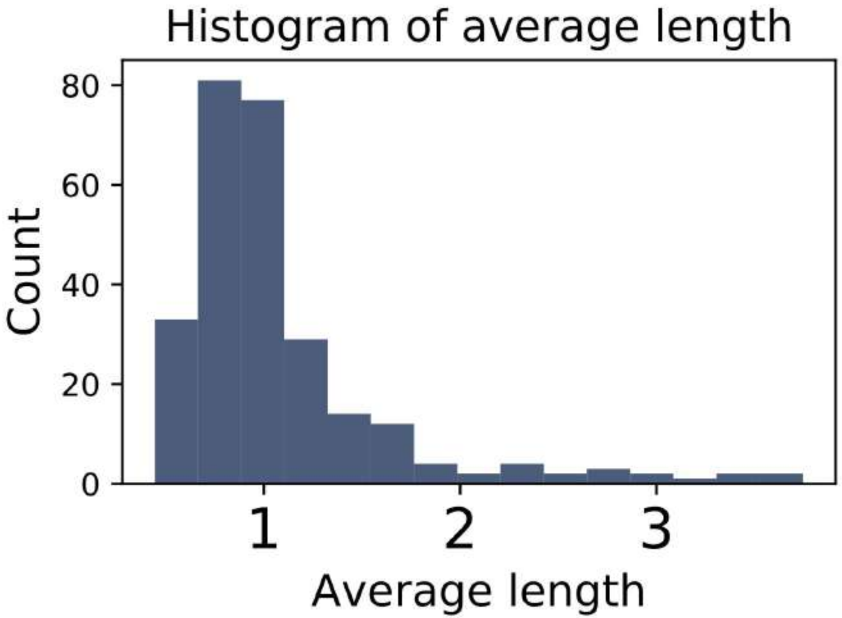
High case growth Anderson County in TX

High death growth Lee County in FL



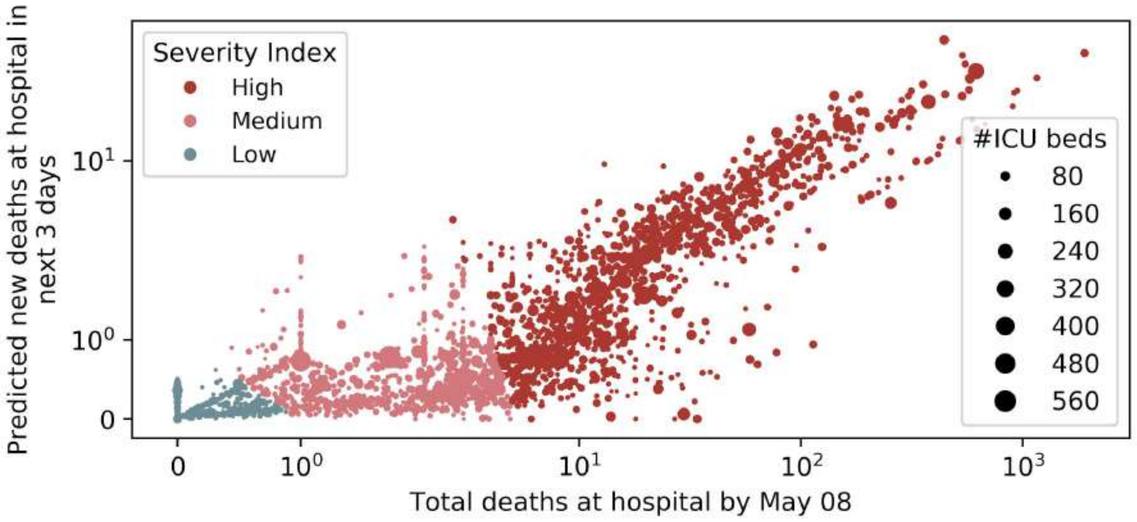
Empirical performance of MEPI for death counts





Evaluation period: March 28--April 27. Only include days since the county has 10 deaths.
Having a normalized length of 0.8 means the PI is roughly (0.6 $\widehat{y}_{t+k}$, 1.4 $\widehat{y}_{t+k}$).

Severity Index



A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

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

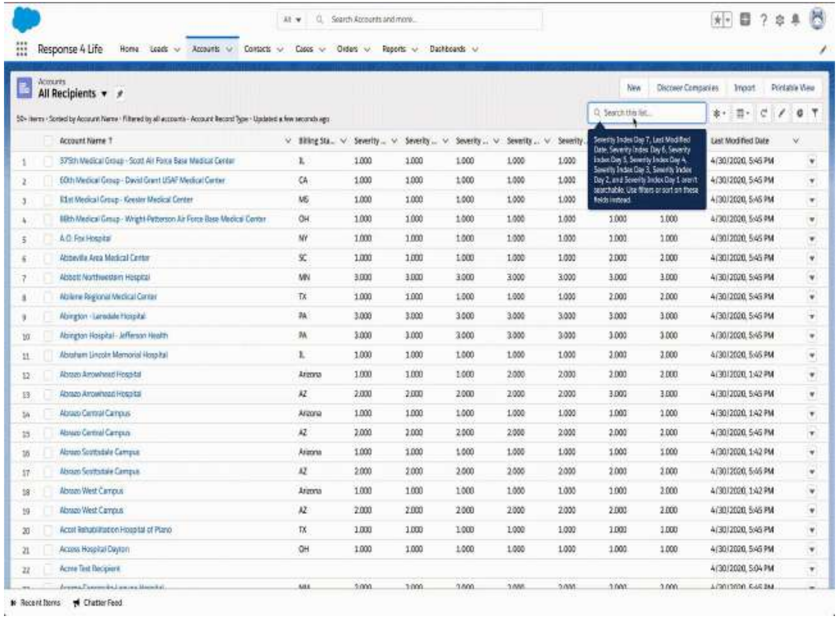
5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, 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**



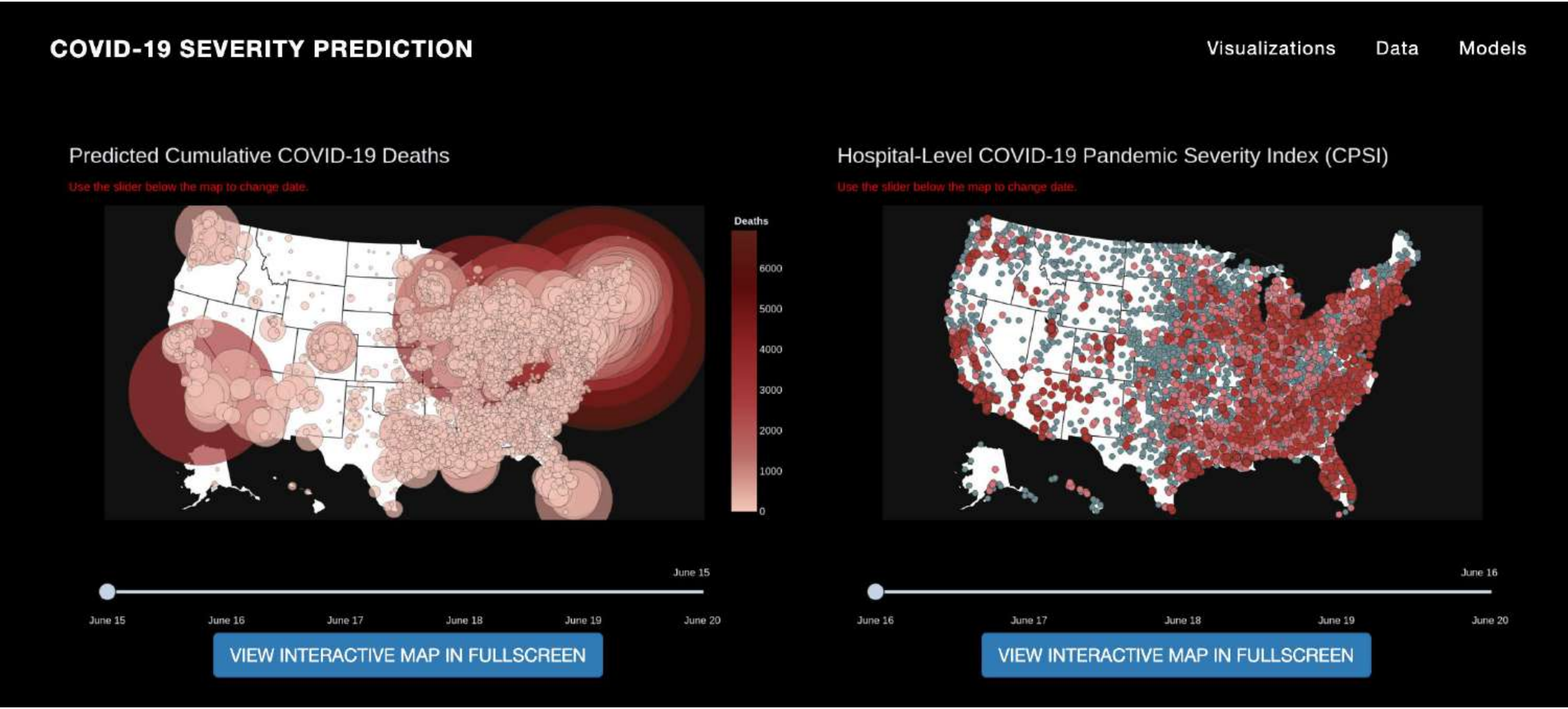
Impact of our work beyond R4L

- 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

- Results and blog on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at UC Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- **Hospitalization prediction** in collaboration with google (and possible collaboration with California Department of Public Health and Microsoft)
- **Causal investigation (e.g. impact of social distancing; matching of counties)** (beginning)

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



Incremental Causal Effects

Dominik Rothenhäusler[1] and Bin Yu[2]

[1]Department of Statistics, Stanford University
[2]Department of Electrical Engineering and Computer Science, and
Department of Statistics, University of California, Berkeley

https://arxiv.org/abs/1907.13258

Supported by ONR Grant

Incremental causal effects (Rothenhaeusler and Yu, 2019)

Causal inference from observational data is challenging

Problems with confounding, overlap, weak instruments,...

Incremental causal effects (Rothenhaeusler and Yu, 2020)

Causal inference from observational data is challenging

Problems with confounding, overlap, weak instruments,...

An important motivation for causal inference is evidence to act. Action decision might need weaker evidence than a positive average treatment effect (ATE) (e.g. whether to increase exercise time).

Moving the goalpost from ATE to other estimands can help:

- Local Average Treatment Effects (Imbens and Angrist, 1994)
- Weighted ATEs (Crump et al., 2006)
- Incremental propensity score interventions (Kennedy, 2019)
- ...

"I'm not cheating, I'm game-changing."

Incremental causal effects: looking for gradient effect

For a continuous treatment T and smooth potential outcomes Y(t) define the incremental causal effect

$$\tau_{\text{incr}} = \mathbb{E}[\partial_t Y(T)]$$

This corresponds to the average change in outcome if slightly increasing the treatment for every unit in the population.

It is often estimated via the average derivative $\mathbb{E}[\partial_t \mathbb{E}[Y|X,T]]$ under appropriate assumptions. Such estimands have appeared in the econ literature (Powell et al., 1989, Newey & Stoker 2003, Banerjee, 2007,...) but have received relatively little attention.

Incremental causal effects - our contributions

- **Incremental causal effects are identified under weaker assumptions** (a **local ignorability** and **local overlap** assumption)

  Conditionally on covariates, units only have to be comparable locally at current treatment t, not necessarily globally across all t

- Incremental causal effects can be estimated with **lower or equal variance** than ATE E[Y(t+1)] - E[Y(t)] if the treatment distribution is Gaussian

- In high-dimensional settings, we use orthogonalization to **transform** the problem of estimation and inference of incremental effects to estimation and inference of a coefficient in a **standard regression model**

  We can use the desparsified Lasso for estimation and inference of incremental causal effects

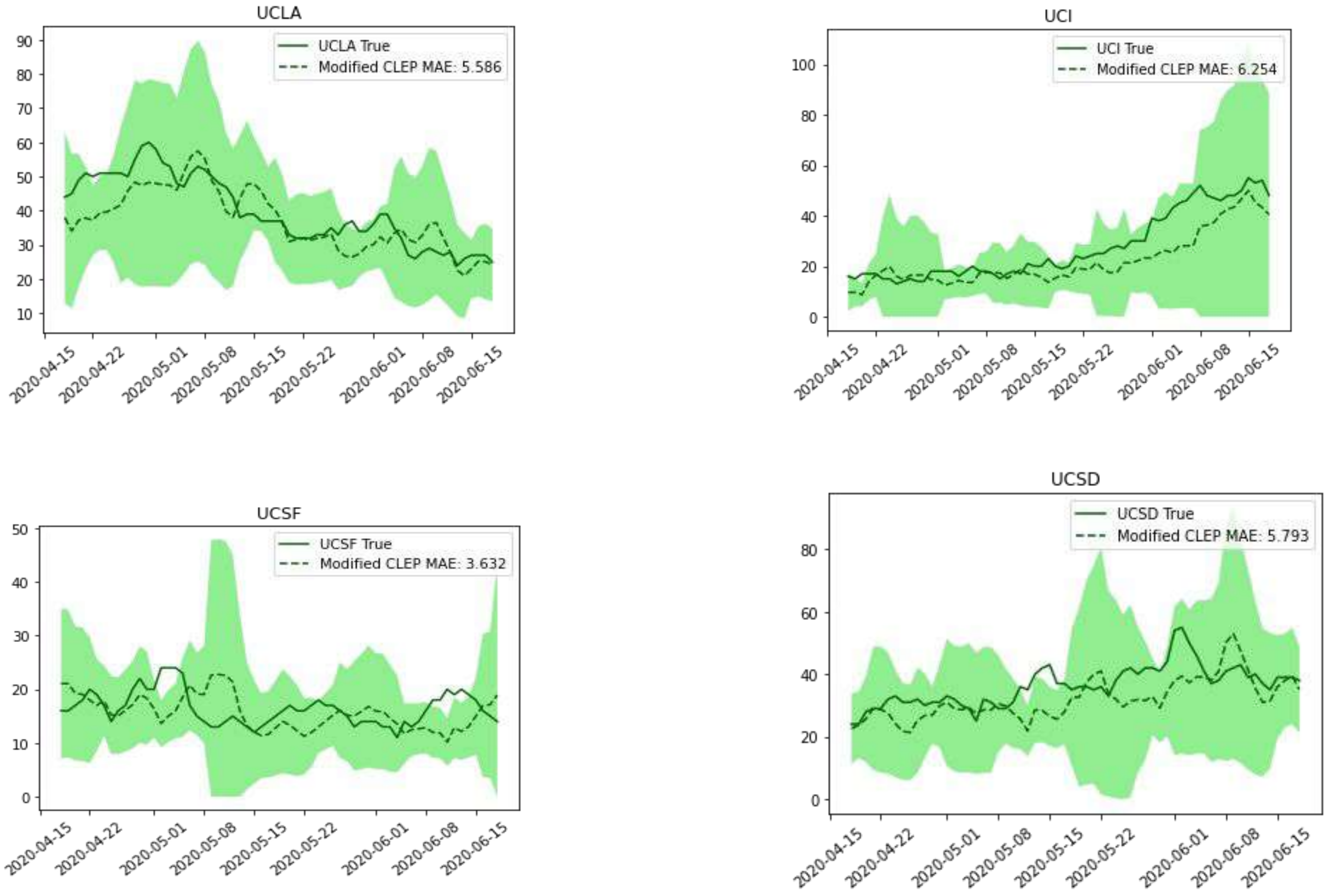
  Paper available at https://arxiv.org/abs/1907.13258

Future work on “weak causality”

- So far: change type of intervention or target population
- Interpolate between effects that are easy to estimate and the ATE. What’s the right way to interpolate?
- Aggregate weak causal evidence across data sets
- Investigate “relaxed causal invariance constraints”

Thank you!

covidseverity.com


UCLA
UCLA True
Modified CLEP MAE: 5.586
UCI
UCI True
Modified CLEP MAE: 6.254
UCSF
UCSF True
Modified CLEP MAE: 3.632
UCSD
UCSD True
Modified CLEP MAE: 5.793
2020-04-15
2020-04-22
2020-05-01
2020-05-08
2020-05-15
2020-05-22
2020-06-01
2020-06-08
2020-06-15

In partnership with

COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

IAS Virtual Event Series
June 25, 2020

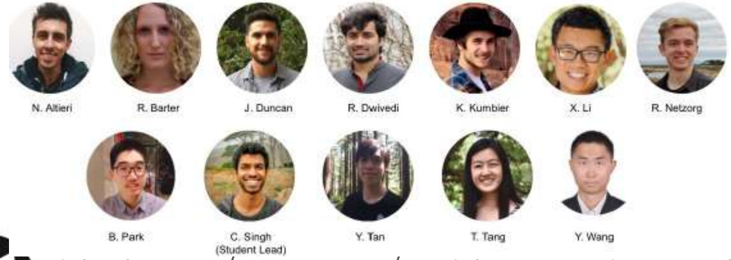
In partnership with

COVID-19 Data Repository and Severity Prediction

Yu Group

UC Berkeley Statistics, EECS, CCB

PI: Bin Yu



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

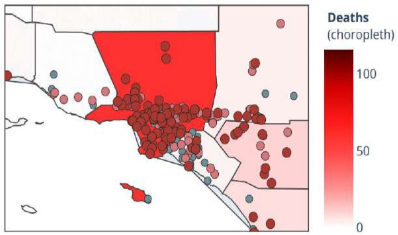
Website: covidseverity.com

5000 Face Shields arrived at Temple Univ Hospital on May 8

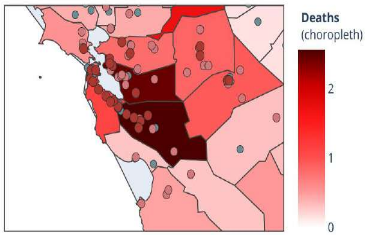
Don Landwirth, R4L

- Curated data repository
- Developed ensemble prediction algorithm at county level for death counts, 7-days ahead
- Designed covid severity index at hospital level for a Salesforce logistics system by R4L

Predicted New Deaths for 2020-05-10



Predicted New Deaths for 2020-05-10



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

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

Our team

from UC Berkeley Statistics/EECS and UCSF

PI: B. Yu

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

B. Park

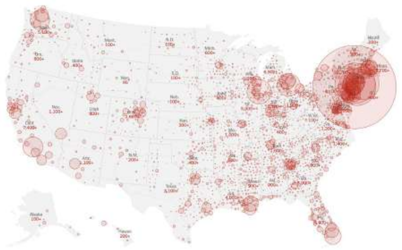
C. Singh
(Student Lead)

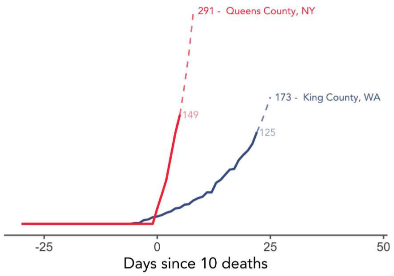
Y. Tan

T. Tang

Y. Wang

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





Data Curation

- Hospital data
- County data

Modeling

- County-level 7-day severity prediction
- hospital demand prediction

Evaluation / Visualization

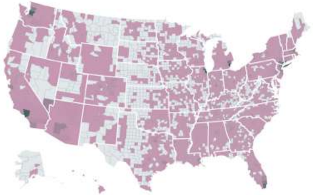
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts

Curating a COVID-19 Data Repository

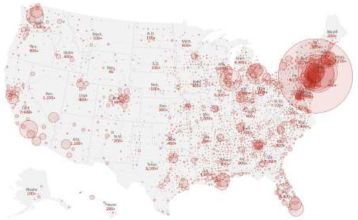
Data curation: scraped from a variety of sources

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

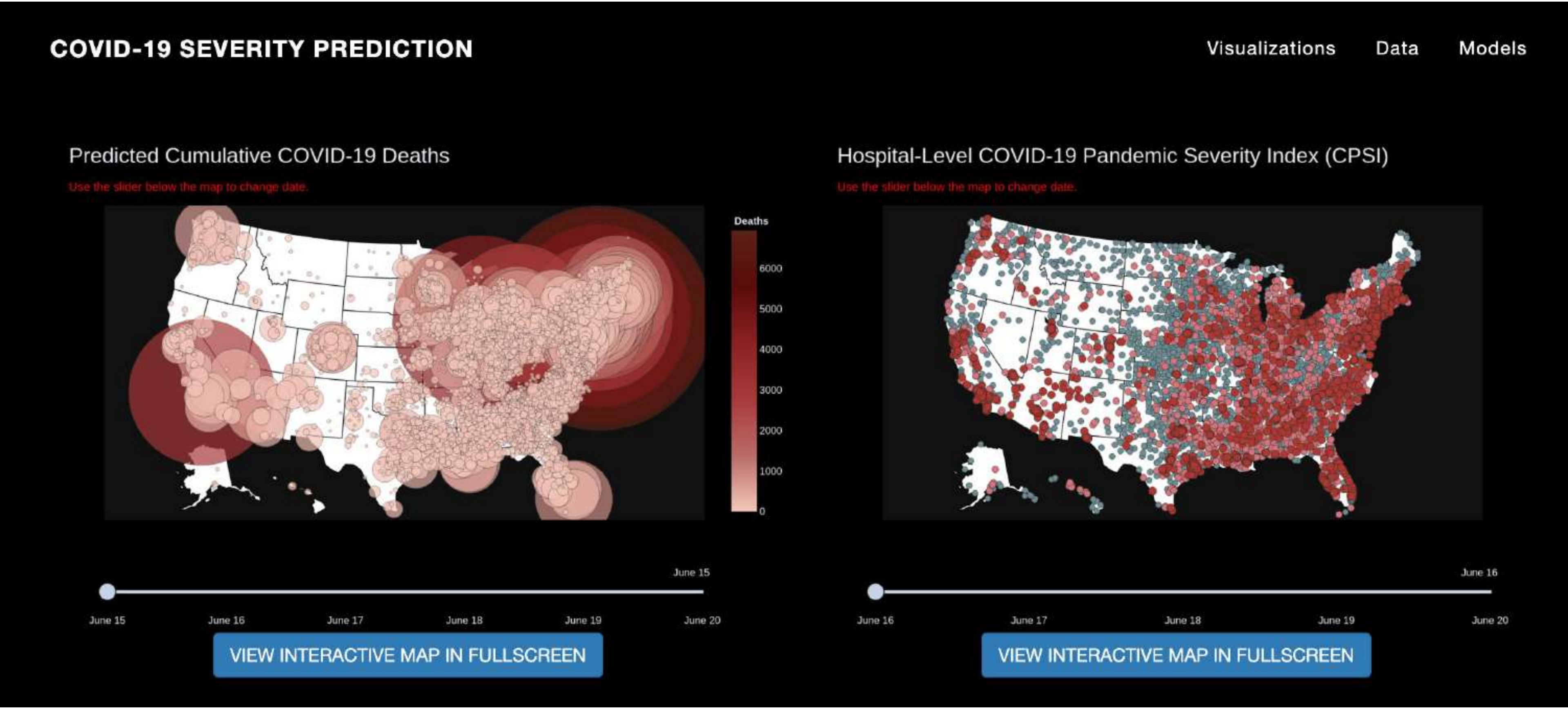
- COVID-19 cases and deaths (NYT and USAFacts)
- Demographics
  - Population, population density, age structure
- Health risk factors
  - Heart disease, stroke, respiratory disease, smoking, diabetes, overall mortality
- Socioeconomic risk factors
  - Social vulnerability index, unemployment, poverty, 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 Repository Traffic & Users (Last 2 weeks)

Estimated total views: ~18K

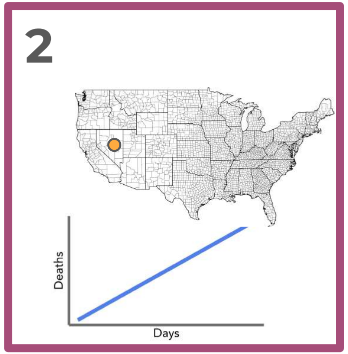
Forecasting county death counts

Website: **covidseverity.com**



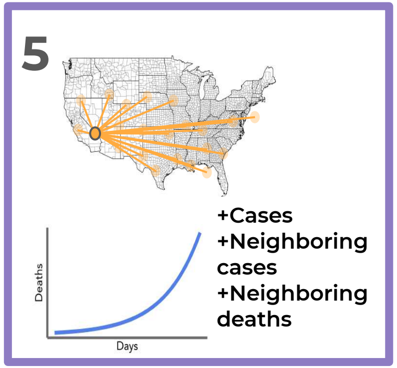
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two models performs well



Separate-county linear predictor

+



Expanded Shared-county exponential predictor

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

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

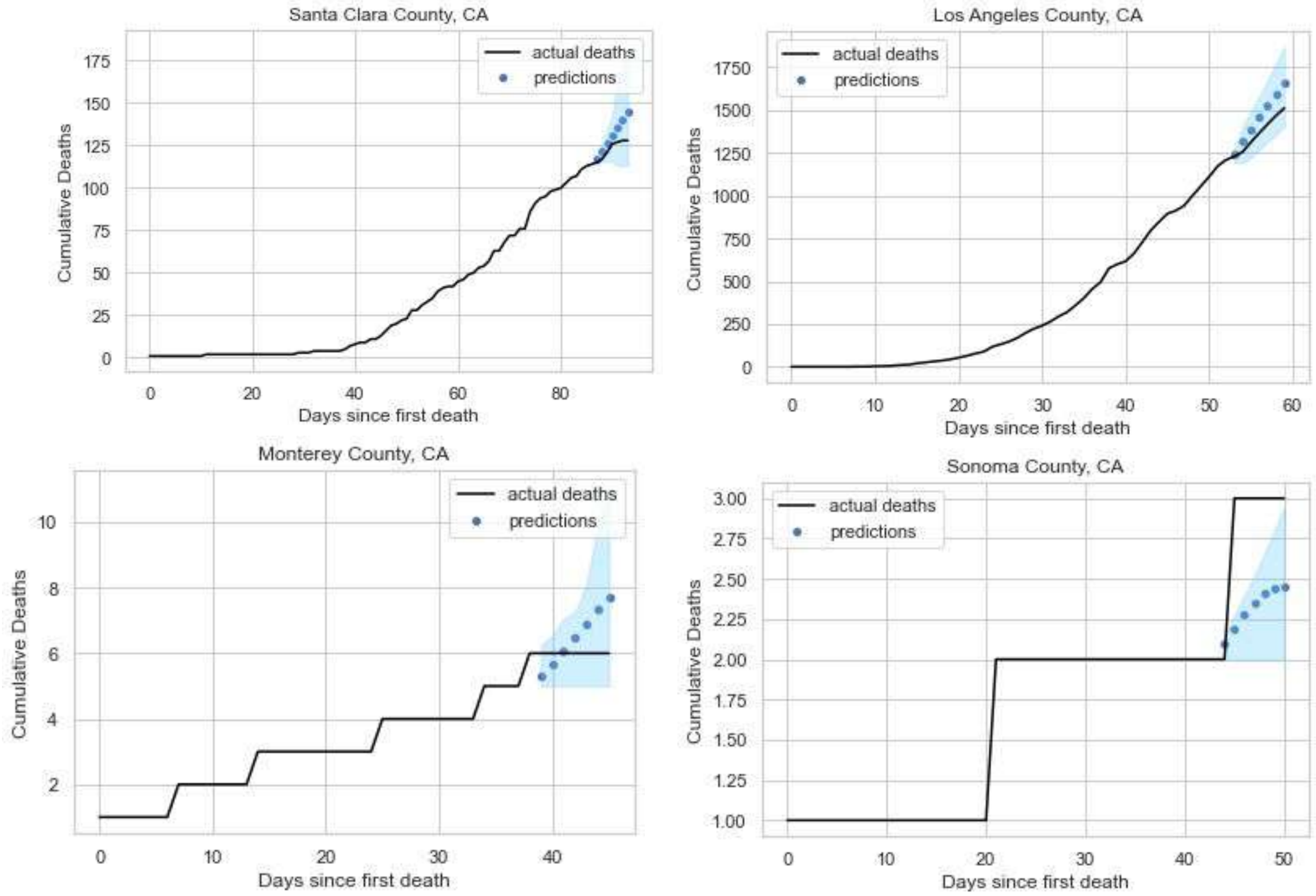
Combined Linear and Exponential Predictors (CLEP)

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

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

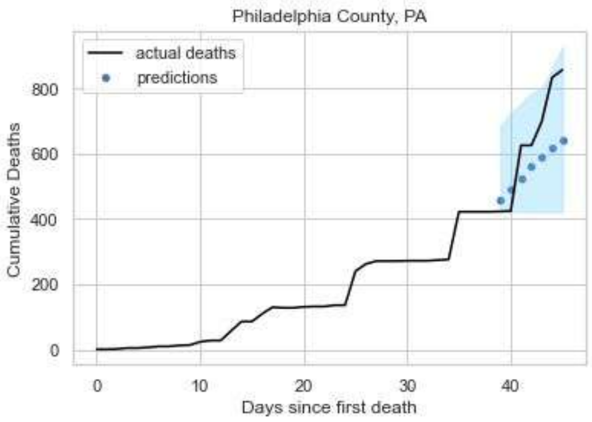
[1]. Schuller, Gerald DT, et al. "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

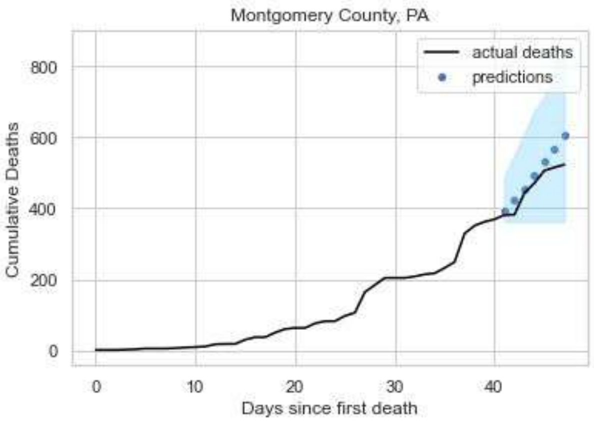
Our county-level 7-day predictive performance



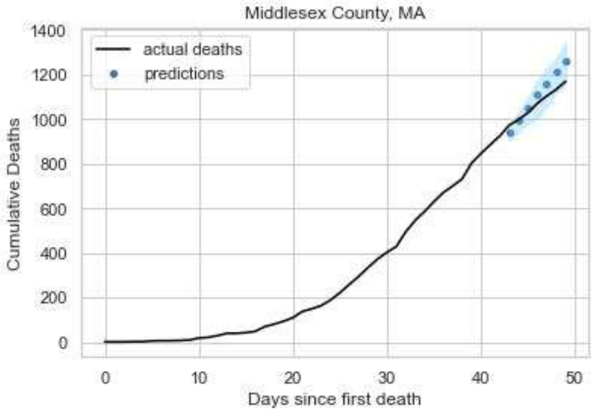
Selected CA counties

Our county-level 7-day predictive performance

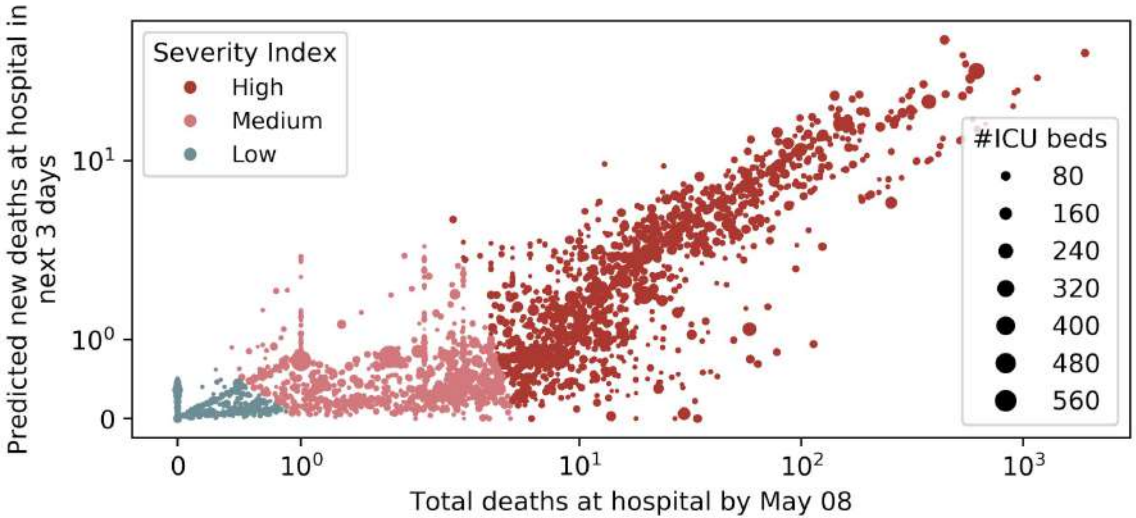




Rapidly Growing Counties



Severity Index

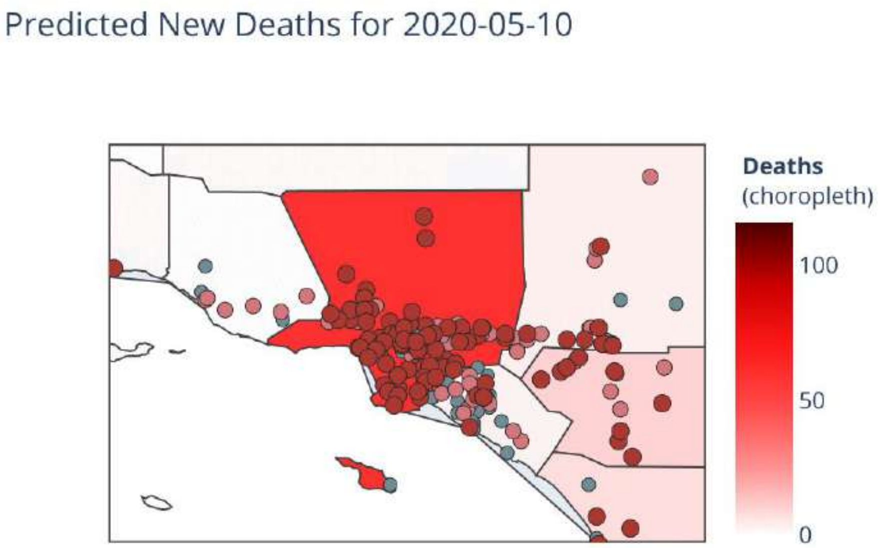


A score* for each hospital based on:

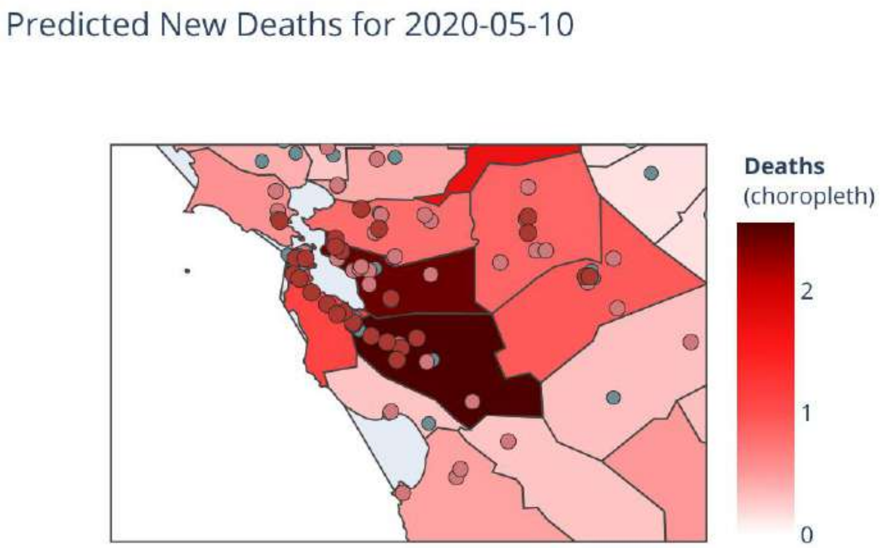
1. Predicted cumulative deaths
2. Predicted daily deaths

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

Mapping Deaths and the Hospital Severity Index Over Time



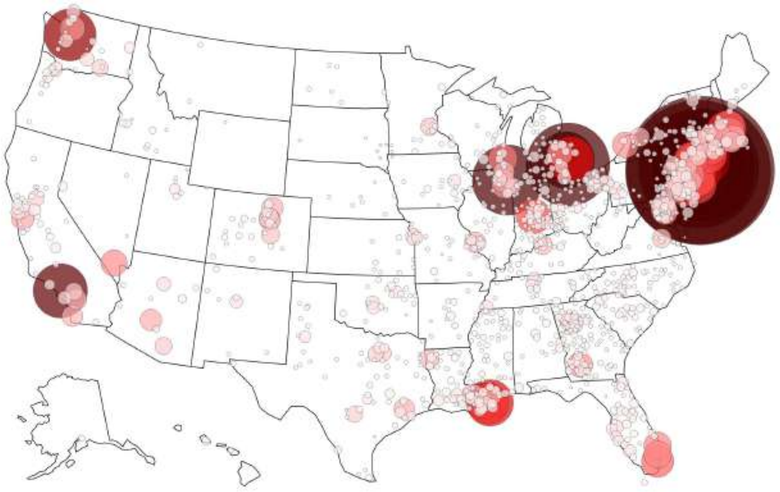
Los Angeles



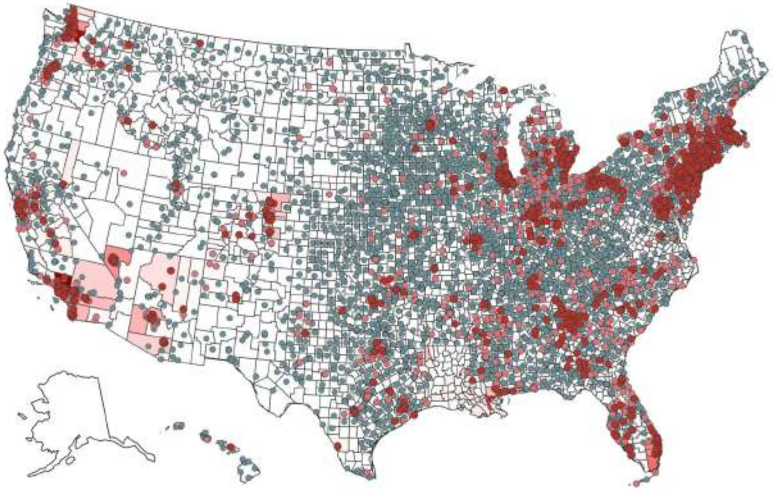
Bay Area

(Interactive) map visualizations

County-level predicted cumulative # of deaths*



Hospital severity index*



*Maps for 04/15

THE CENTER FOR SPATIAL DATA SCIENCE THE UNIVERSITY OF CHICAGO

Collaborating with the Center for Spatial Data Science (**CSDS**) at **University of Chicago** to add our predictions and severity index to the U.S. COVID-19 Atlas.

5000 Face Shields arrived at Temple Univ Hospital on May 8

Don Landwirth, R4L

5000 Face Shields arrived at Temple Univ Hospital on May 8

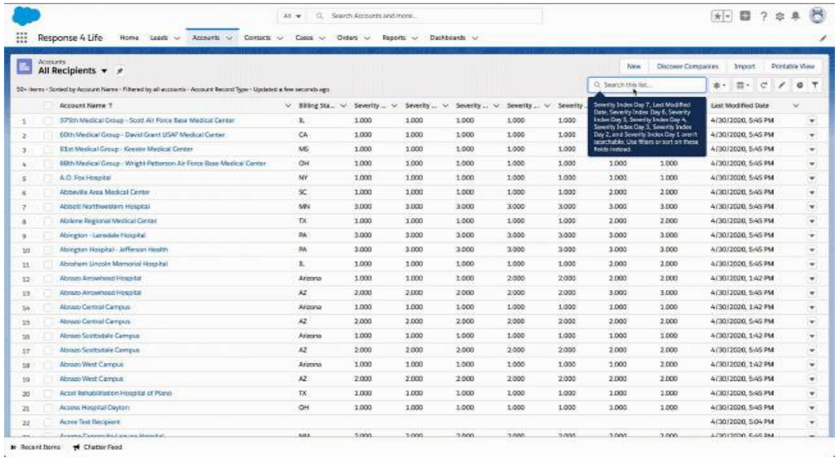
Don Landwirth, R4L

Impacts through Response4life

500k face shields in US by the end of may

- 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 in 2 weeks**
- **+500k outside the US**

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



Impact of our work beyond R4L

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

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

- Results on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at UC Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Possible collaboration with California Department of Public Health
- Possible causal inference through matching of counties

Paper available at tinyurl.com/yugroup-covid19 and at Bin Yu's webiste

Curating a COVID-19 data repository and forecasting county-level death counts in the United States

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

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

April 29, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors

This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

In partnership with

COVID-19 Data Repository and County Death Count Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

PI: Bin Yu

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

B. Park

C. Singh
(Student Lead)

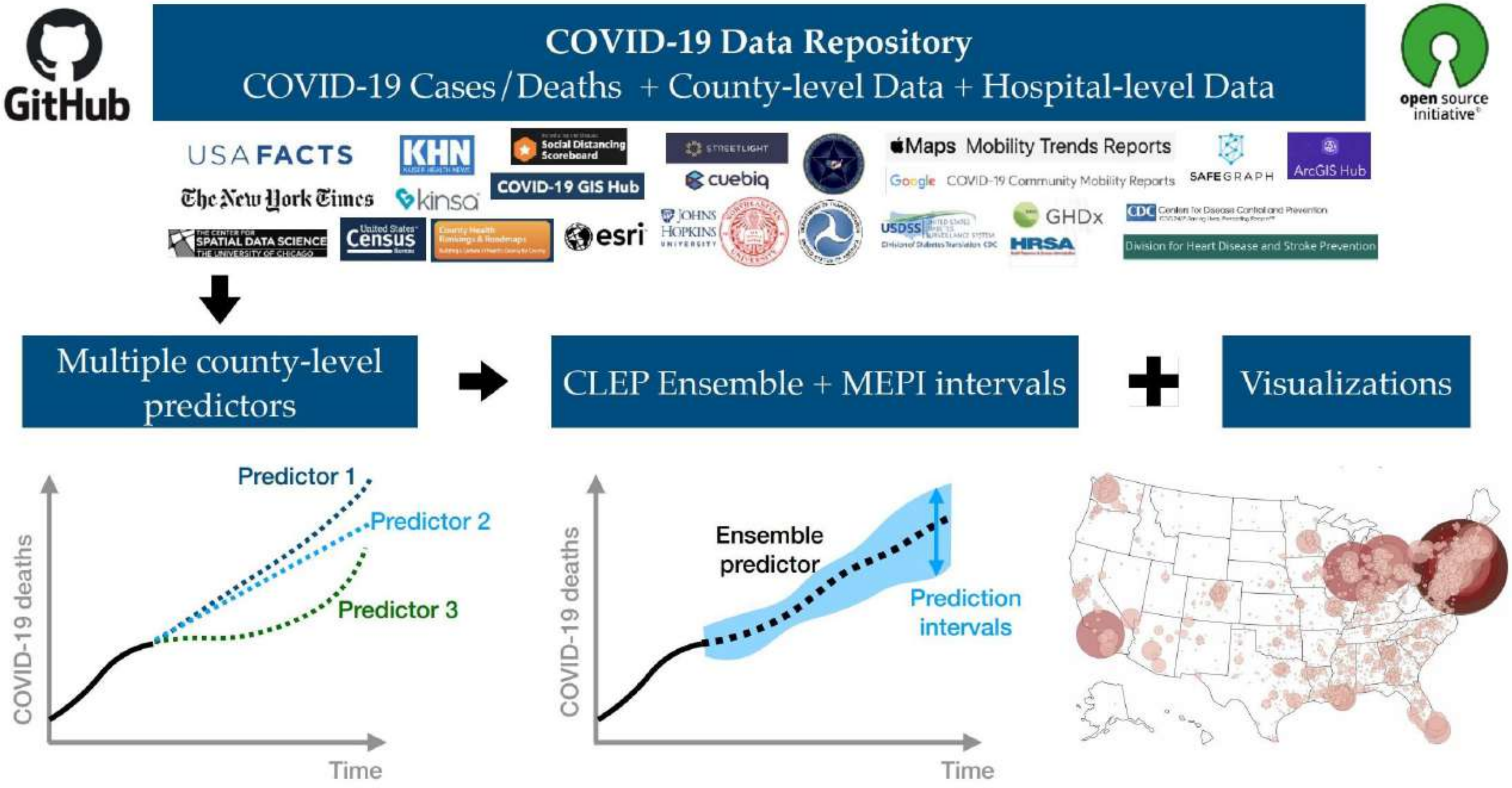
Y. Tan

T. Tang

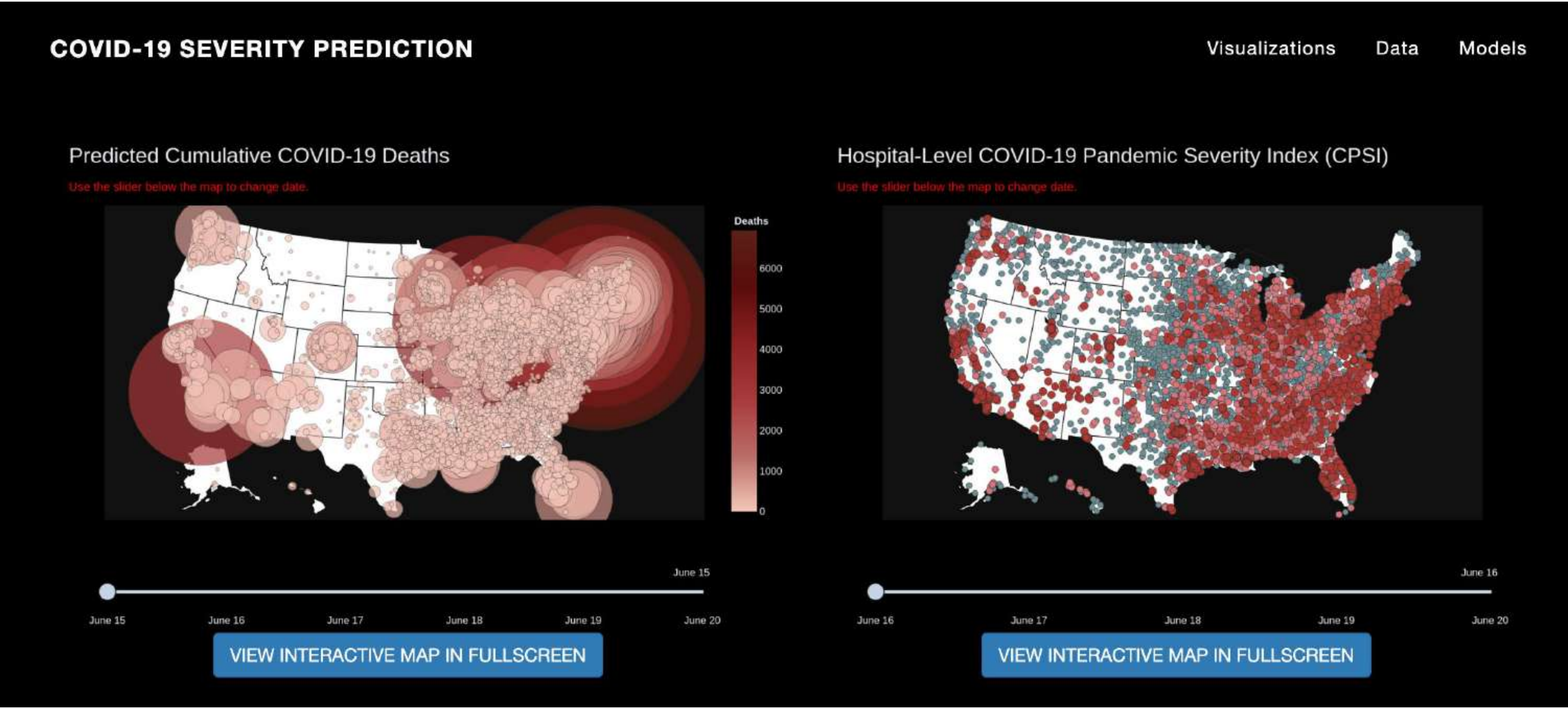
Y. Wang

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

Overview: Current Data Repository & Prediction Pipeline

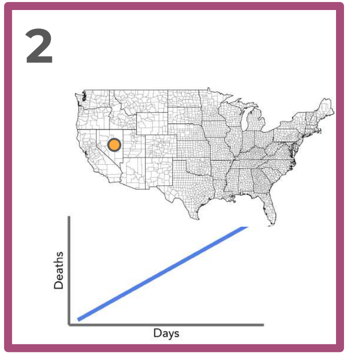


Website: **covidseverity.com**



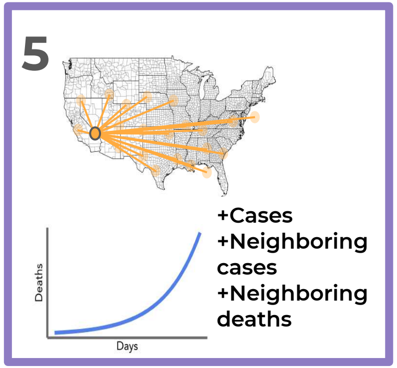
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two models performs well



Separate-county linear predictor

+

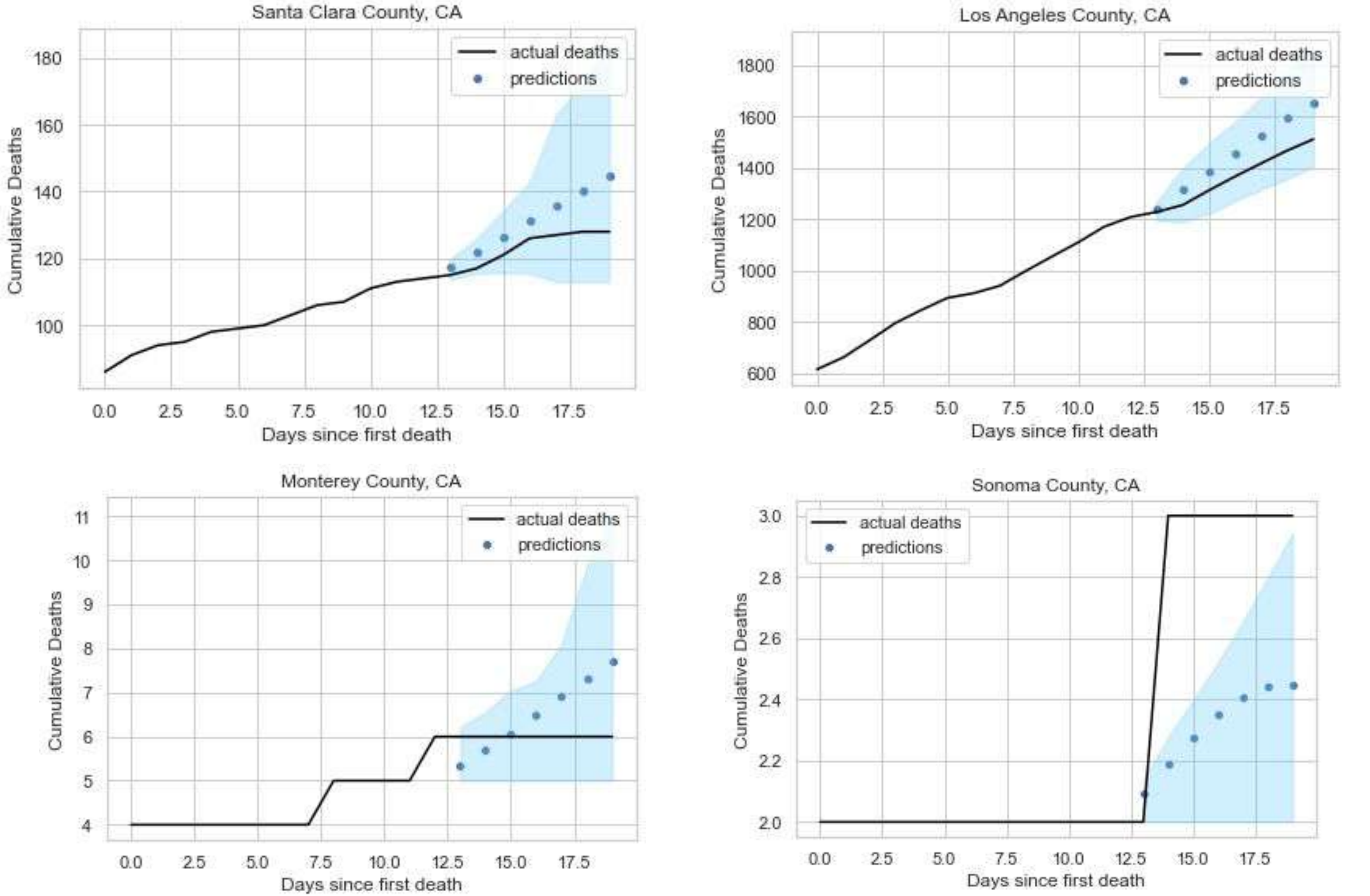


Expanded Shared-county exponential predictor

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

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

Death count prediction results: 4/20-5/10



Selected CA counties

Data Repository

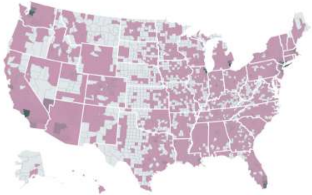
Overview of sources (county/hospital)

- pipelines/processes
- Current users
- Current efforts

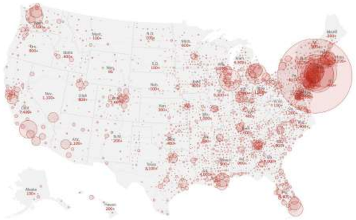
Data: scraped from a variety of sources

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

esri

COVID-19 GIS Hub

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

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

IHME GHDx

Introducing the Unacast Social Distancing Scoreboard

USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

JOHNS HOPKINS UNIVERSITY

CMS.gov
Centers for Medicare & Medicaid Services

United States® Census Bureau

STREETLIGHT

KHN KAISER HEALTH NEWS

SAFEGRAPH

cuebiq

kinsa®

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 Repository Traffic & Users (Last 2 weeks)

Estimated total views: ~18K

Impact: 5000 Face Shields arrived at Temple Univ Hospital on May 8

RESPONSE 4LIFE

Don Landwirth, R4L

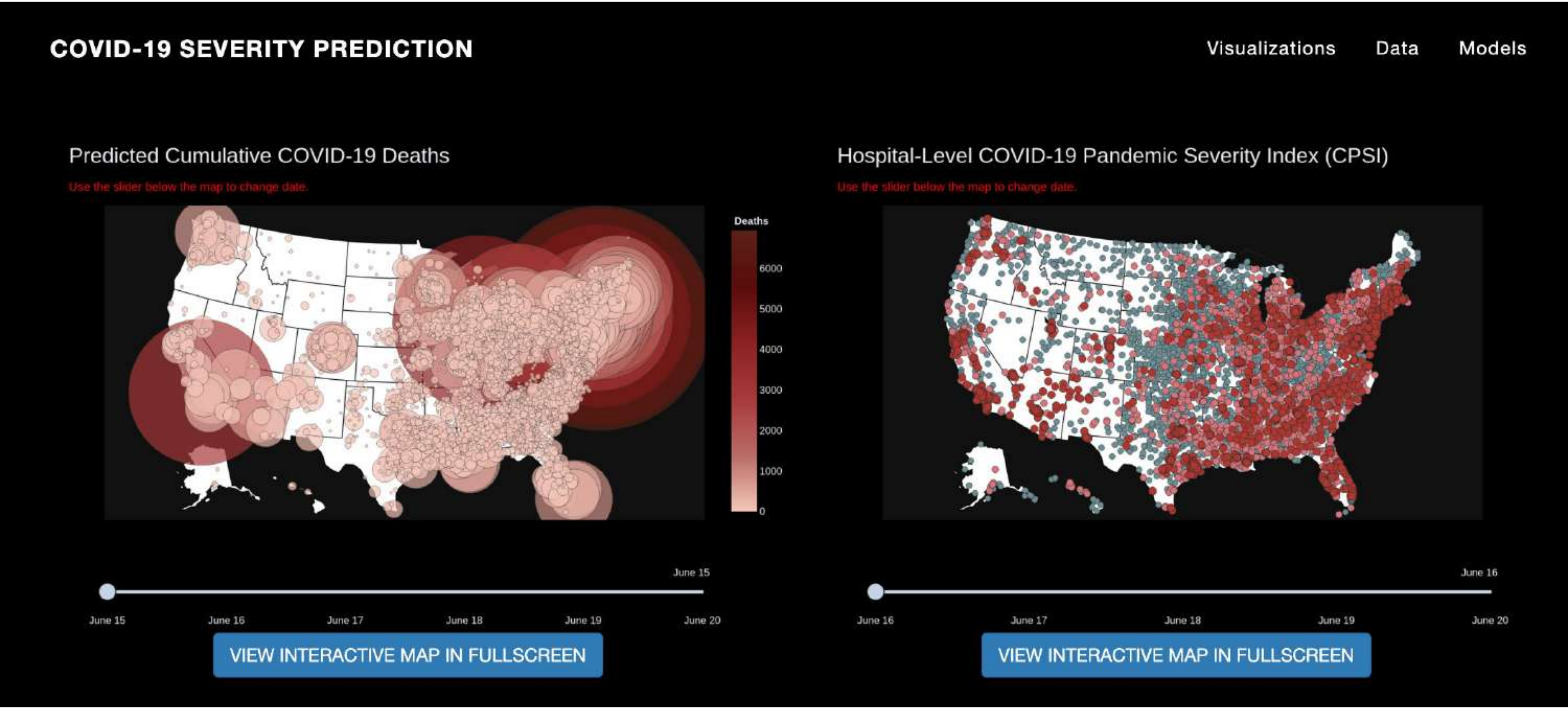
Other Impacts of Our Data Repository

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

  In last two weeks, 2.9K visits with 394 unique visitors;

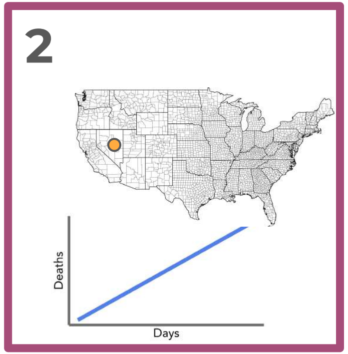
  153 clones with 102 unique cloners
- Results on CSDS atlas at University of Chicago
- Final project option for DS100 at UC Berkeley (> 1000 students) and Stat542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Collaboration with Google OpenSource, Microsoft's AI for Good, on hospitalization need prediction (on-going)
- Possible collaboration with with California Department of Public Health
- Exploratory causal inference through matching of counties (on-going)

Website: **covidseverity.com**



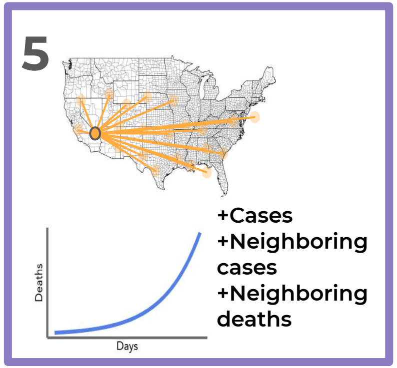
Combined Linear and Exponential **Predictor** (**CLEP**)

A combination of two models performs well



Separate-county linear predictor

+

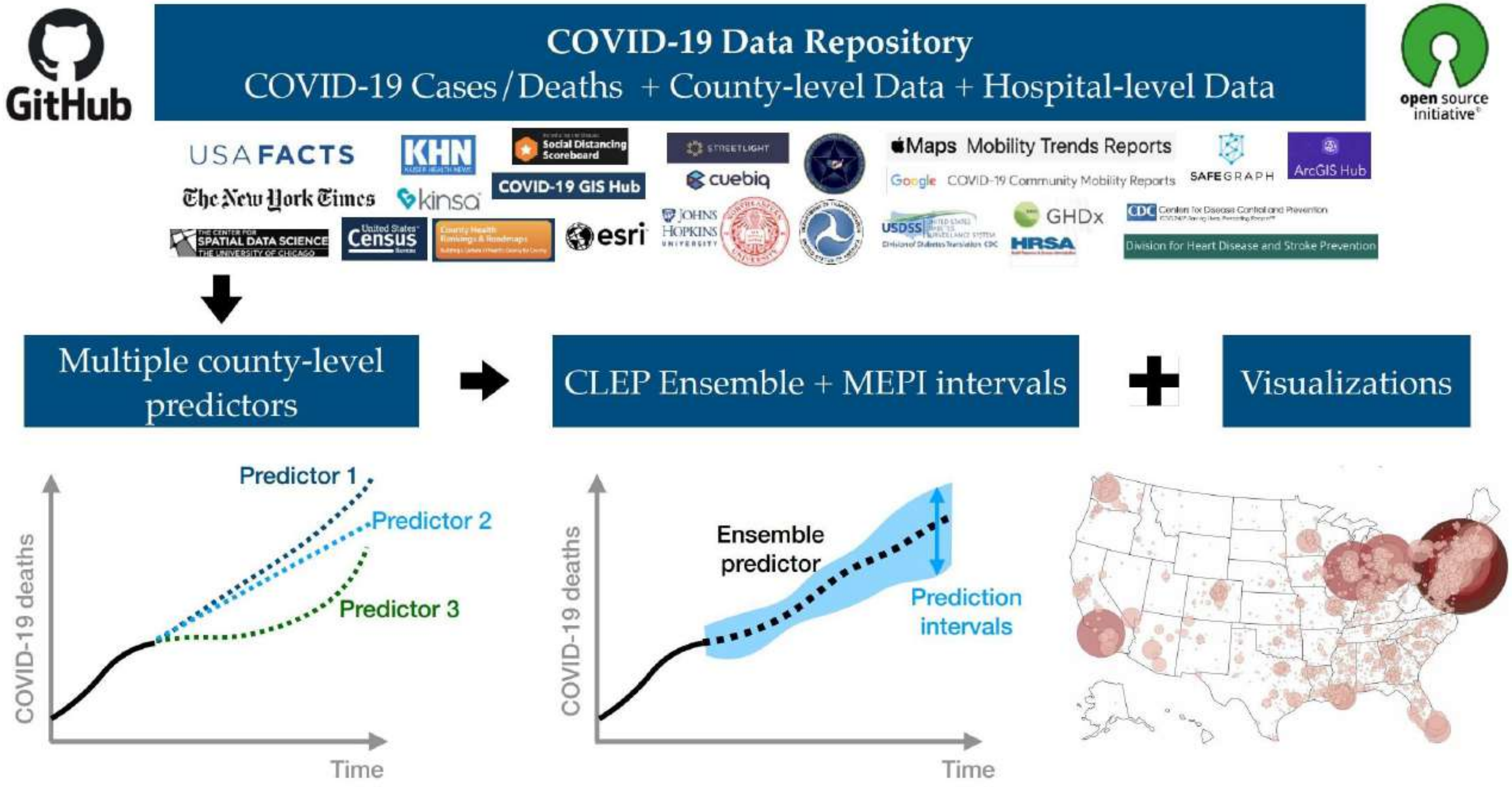


Expanded Shared-county exponential predictor

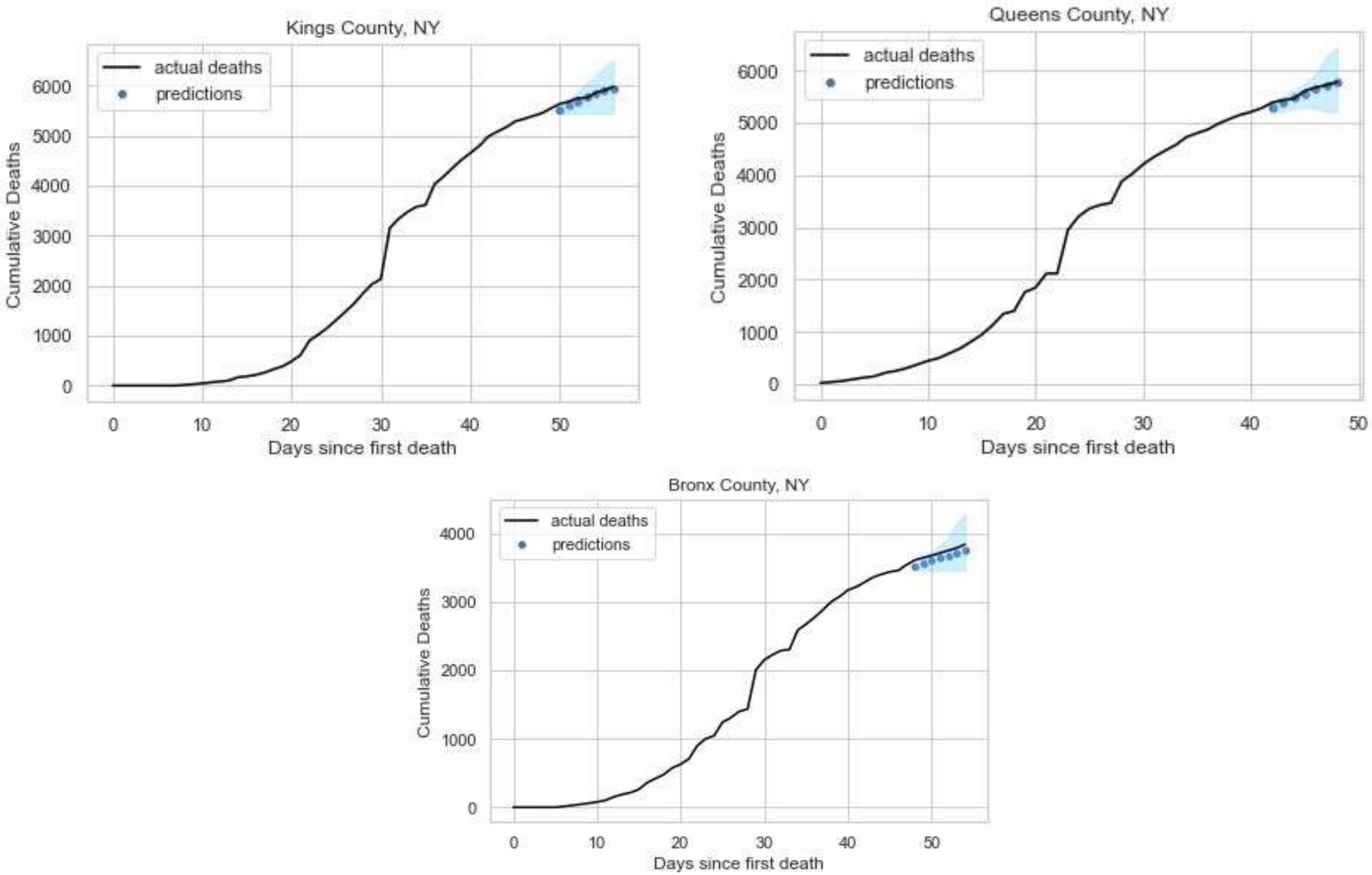
Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[1]

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

Current Data Repository & Pipeline(alternative to page 2)



Our county-level 7-day predictive performance

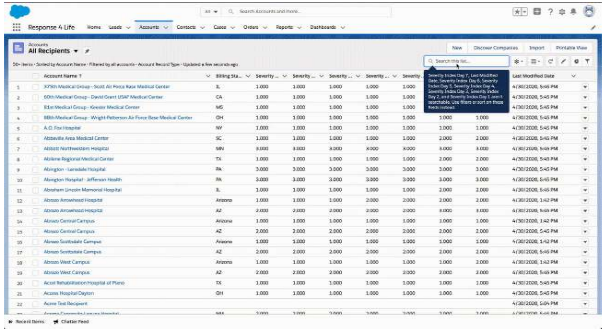


Worst Affected Counties

Impacts through Response4life

500k face shields in US by the end of may

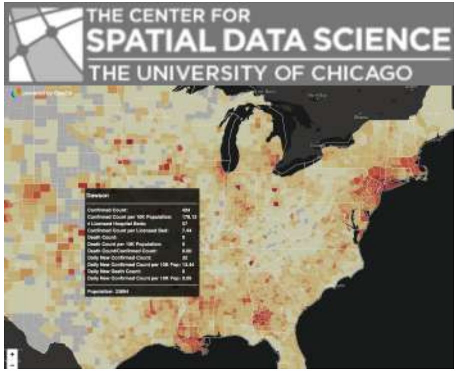
- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space,
- **+65k to 25 recipients in 15 states in 2 weeks**
- **+500k outside the US**



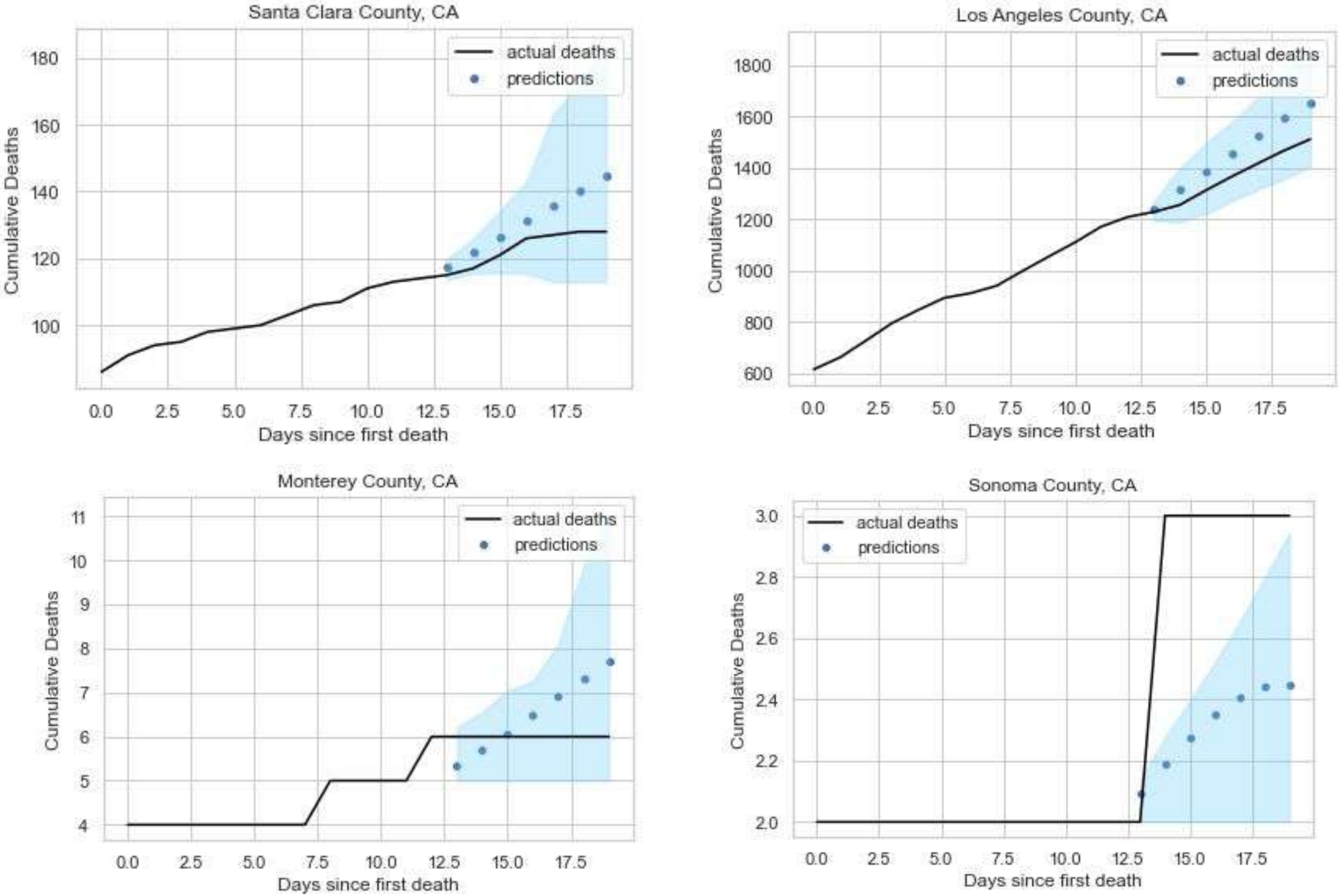
Curating data repository

Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

Visualizations

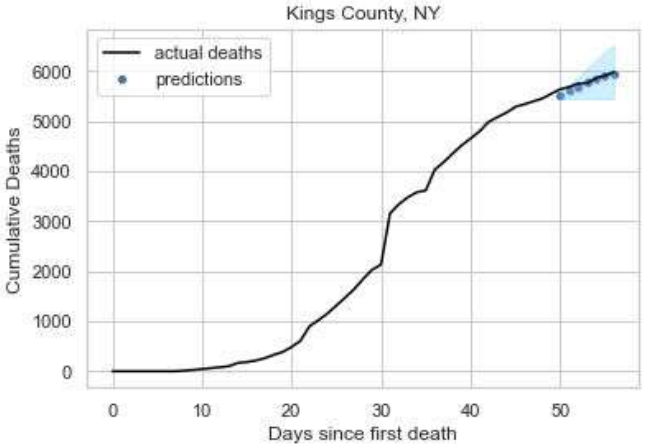


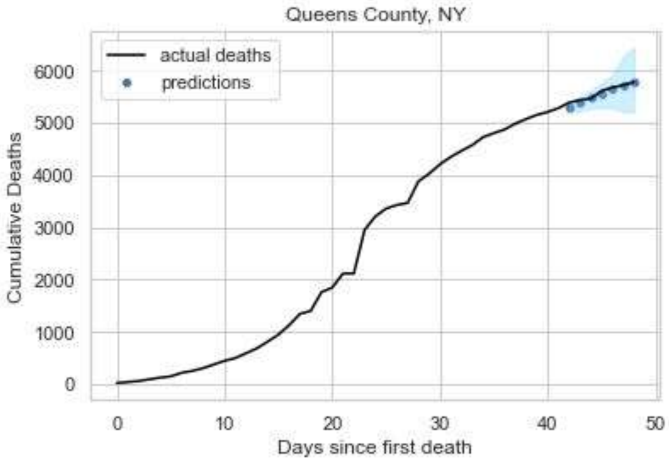
Most recent 20 days zoom in



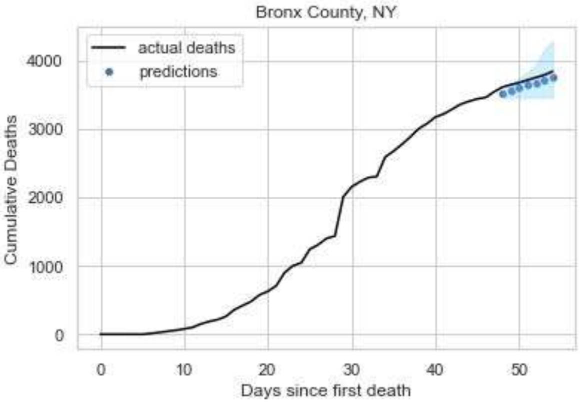
Selected CA counties

Our county-level 7-day predictive performance

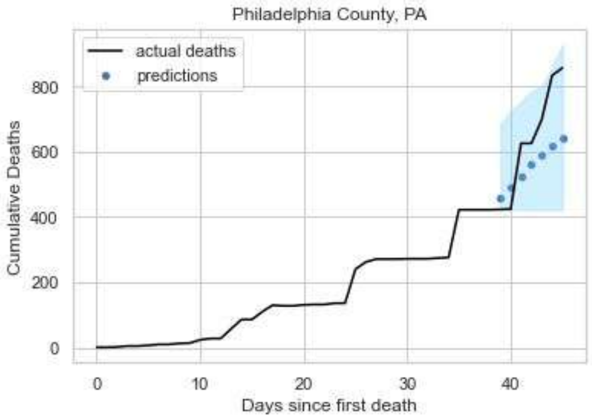


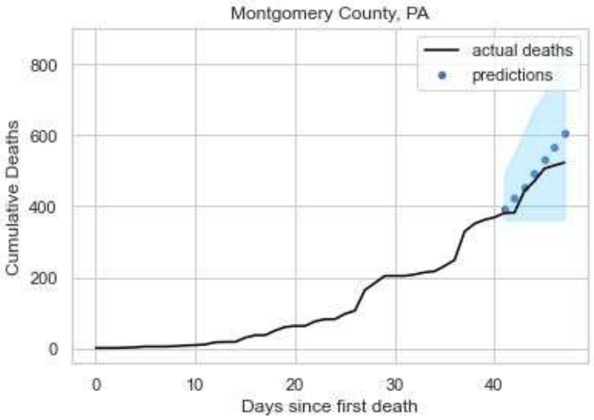


Worst Affected Counties

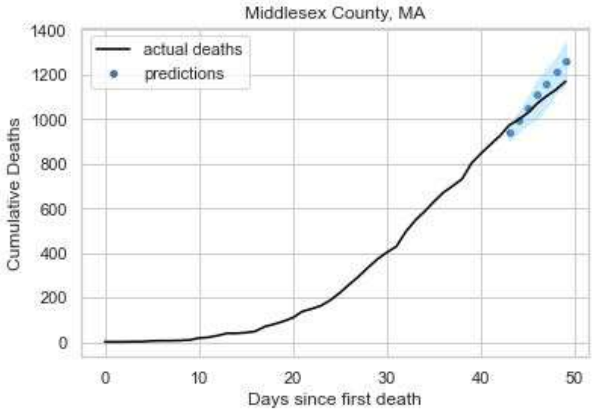


Our county-level 7-day predictive performance





Rapidly Growing Counties



In partnership with

COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

California Department of Public Health, Modeling Group
May 11, 2020

Goal: Help Aid Resource Allocation

Our team

from UC Berkeley Statistics/EECS and UCSF

PI: B. Yu

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

B. Park

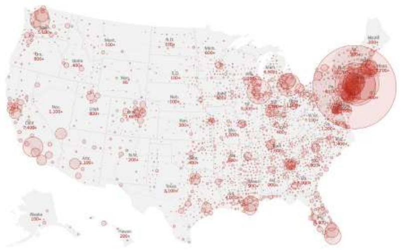
C. Singh
(Student Lead)

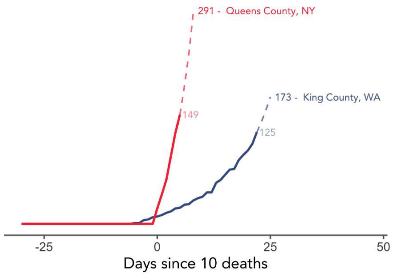
Y. Tan

T. Tang

Y. Wang

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





Data Curation

- Hospital data
- County data

Modeling

- County-level 7-day severity prediction
- hospital demand prediction

Evaluation / Visualization

- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts

Curating a COVID-19 Data Repository

Data Processing Pipeline

Data Scraping → **Data Cleaning** → **Data Validity**

Data Scraping

- **Collect 1M records from 10+ data sources**
- **Monitor data changes 24/7 powered by AWS**

Data Cleaning

- **Handling missing and erratic entries**
- **Automated python script**

Data Validity

- **Compare data across different sources to ensure data validity**
- **Search for emerging data sources**

For almost a month, 2 full-time students, and on-going with 1 full-time student

Amazon EC2

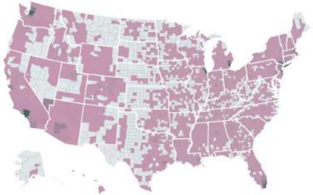
Data and code available: https://github.com/Yu-Group/covid19-severity-prediction

★ Being used by multiple research groups across the country

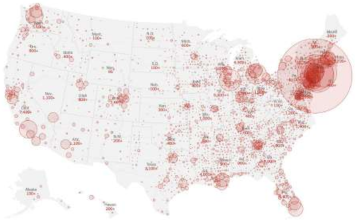
Data: scraped from a variety of sources

COVID-19 Cases/Deaths

USA FACTS



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

esri

COVID-19 GIS Hub

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

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

IHME GHDx

Introducing the Unacast
Social Distancing Scoreboard

USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

JOHNS HOPKINS UNIVERSITY

CMS.gov
Centers for Medicare & Medicaid Services

United States® Census Bureau

STREETLIGHT

KHN KAISER HEALTH NEWS

SAFEGRAPH

cuebiq

kinsa®

Hospital-level Data

(e.g., #ICU beds, staff)

HRSA
Health Resources & Services Administration

ArcGIS Hub

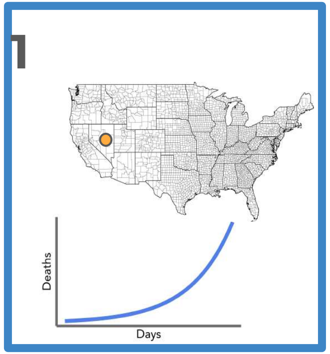
HOMELAND INFRASTRUCTURE FOUNDATION-LEVEL DATA SUBCOMMITTEE

Samuel Scarpino

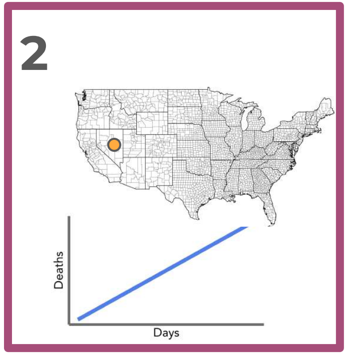
NORTHEASTERN UNIVERSITY

Forecasting county death counts

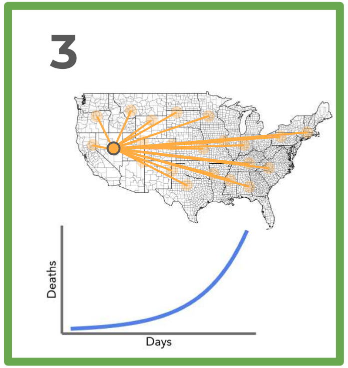
Combined Linear and Exponential Predictors (CLEP)



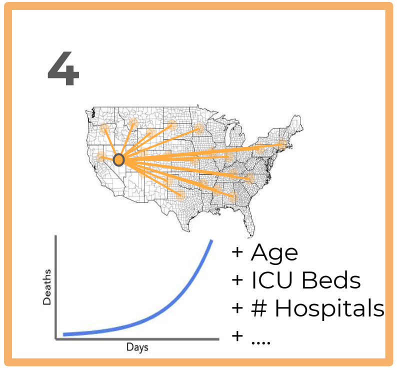
Separate-county exponential predictor



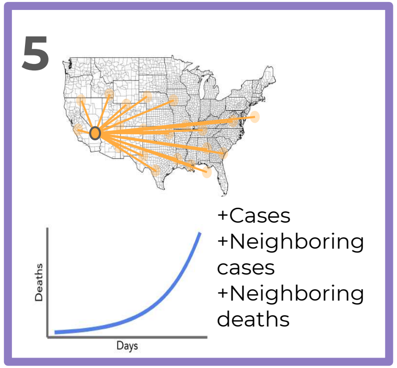
Separate-county linear predictor



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 historical performance[2]

[2]. Schuller, Gerald DT, et al. "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)

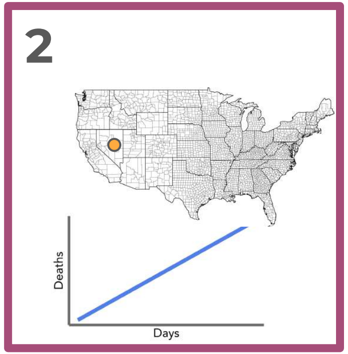
Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

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

[2]. Schuller, Gerald DT, et al. "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

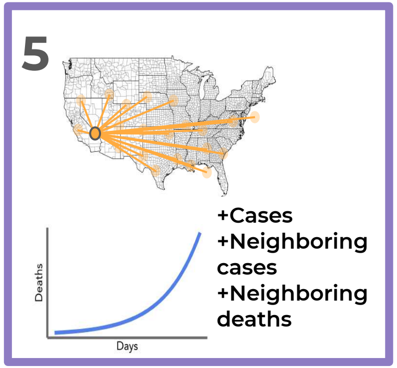
Combined Linear and Exponential Predictors (CLEP)

A smaller combination performed better



Separate-county linear predictor

+

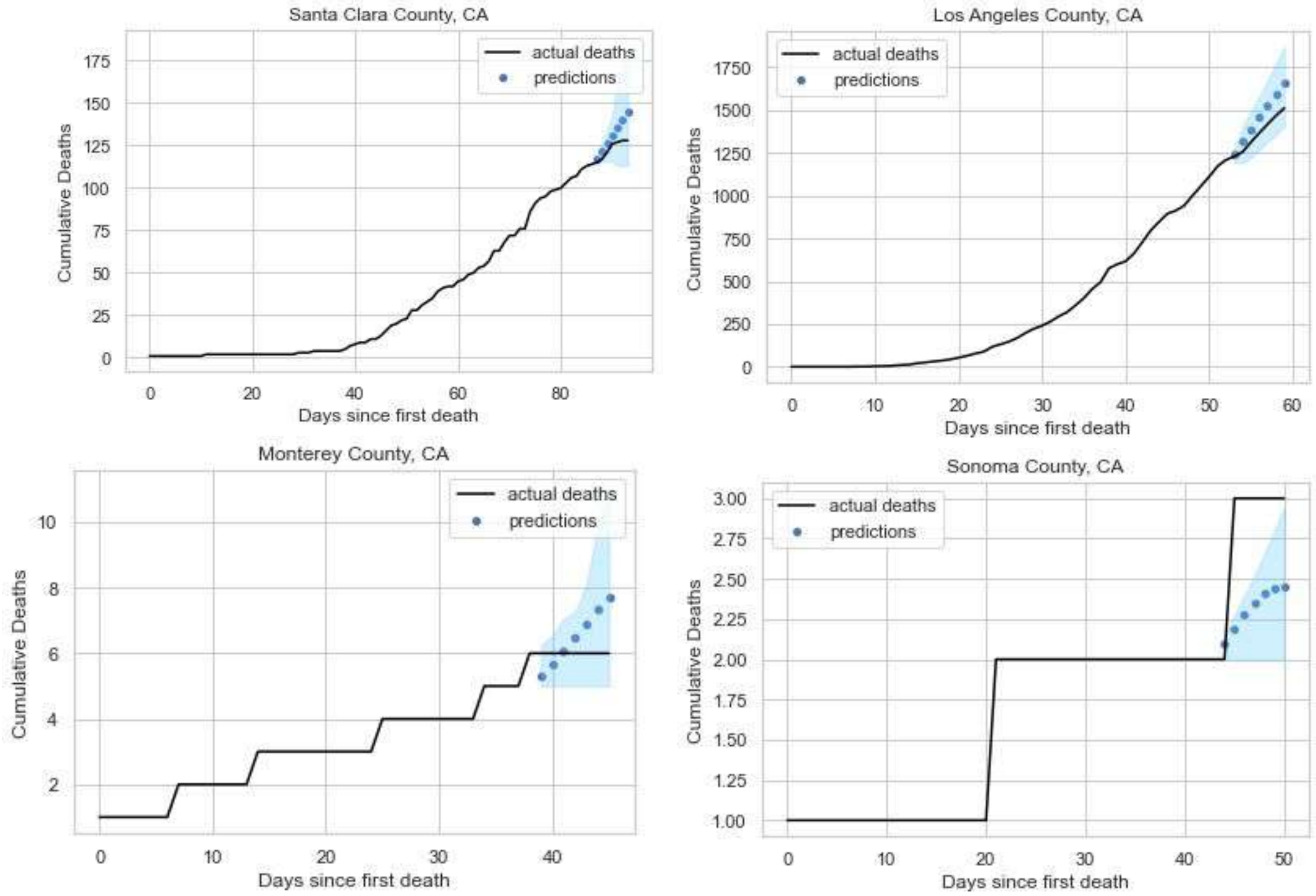


Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

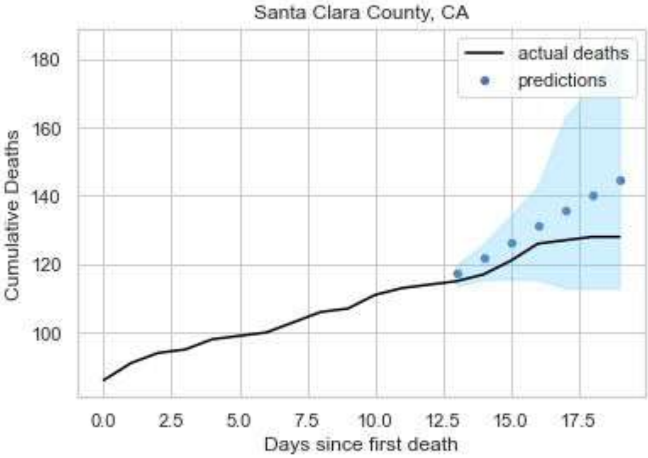
[2]. Schuller, Gerald DT, et al. "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

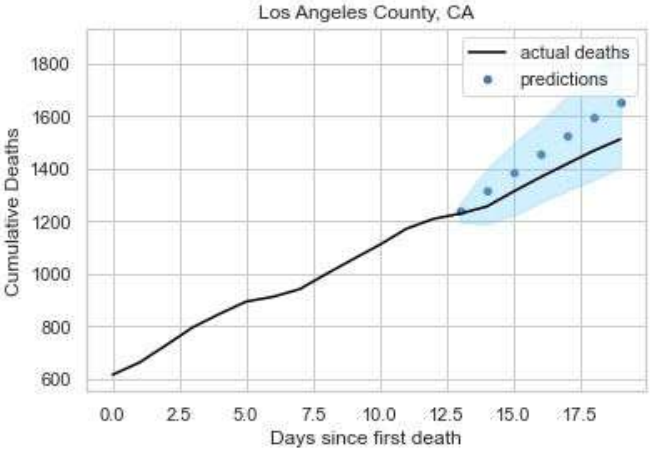
Our county-level 7-day predictive performance



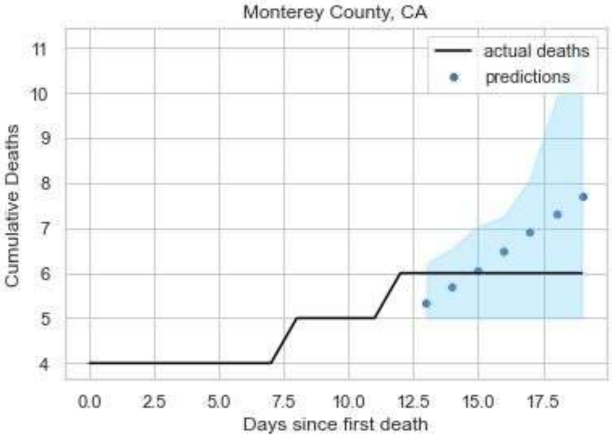
Selected CA counties

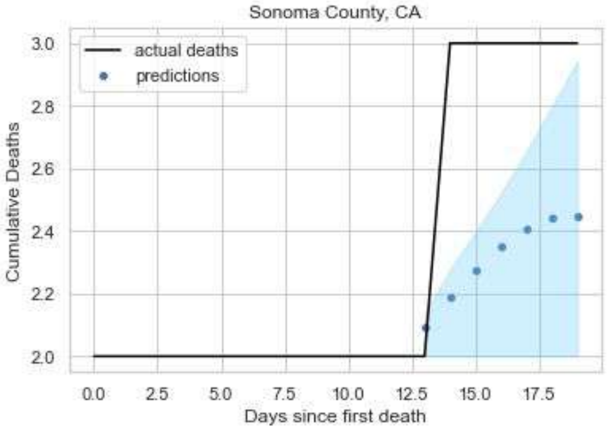
Most recent 20 days zoom in



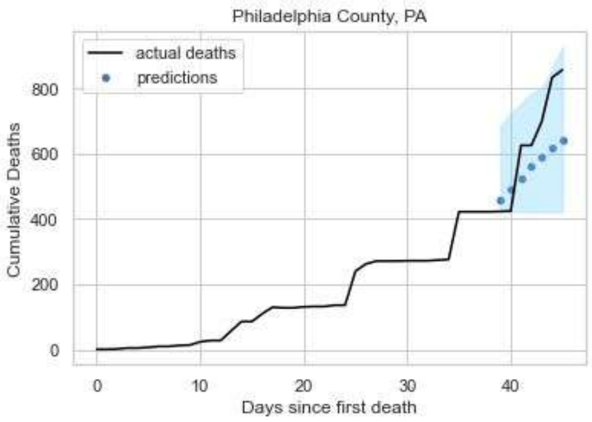


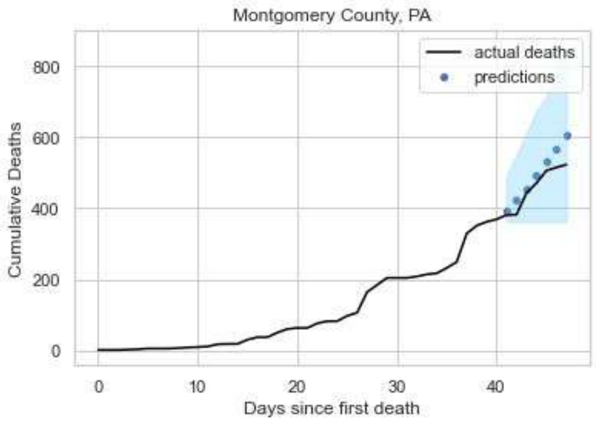
Selected CA counties



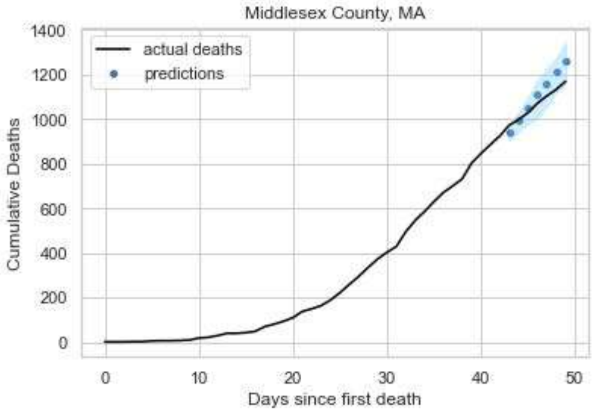


Our county-level 7-day predictive performance

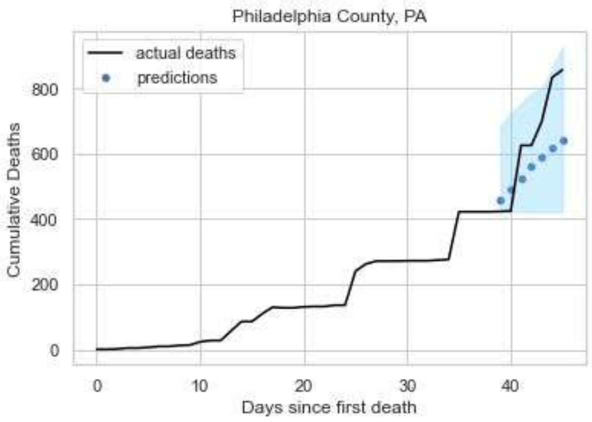


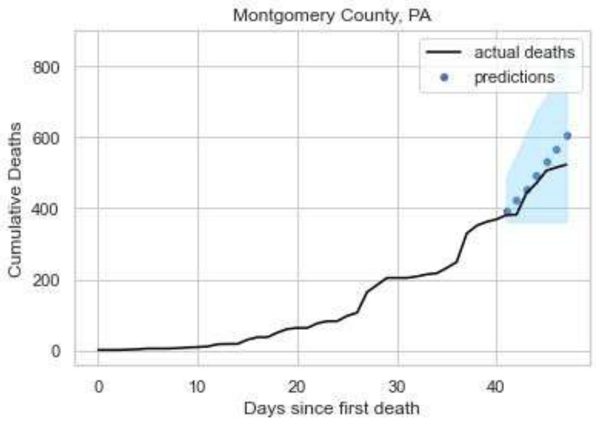


Rapidly Growing Counties

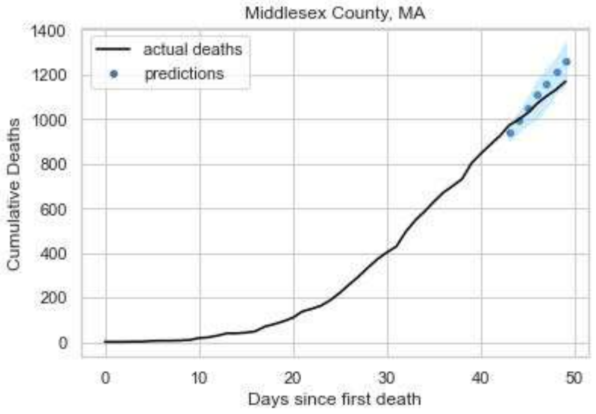


Our county-level 7-day predictive performance

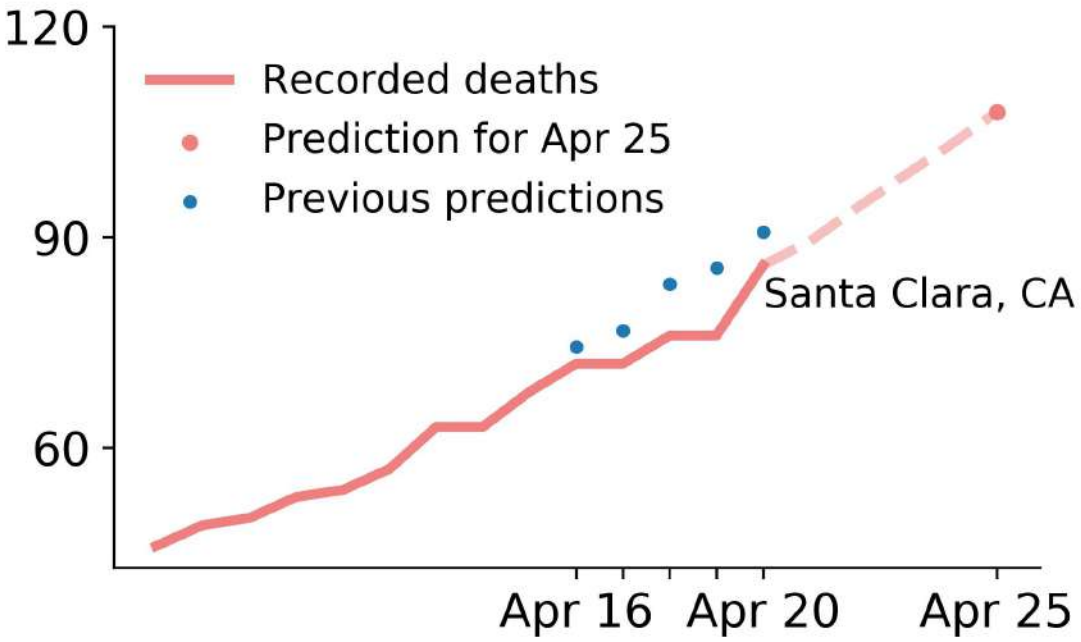




Rapidly Growing Counties



Prediction Intervals:

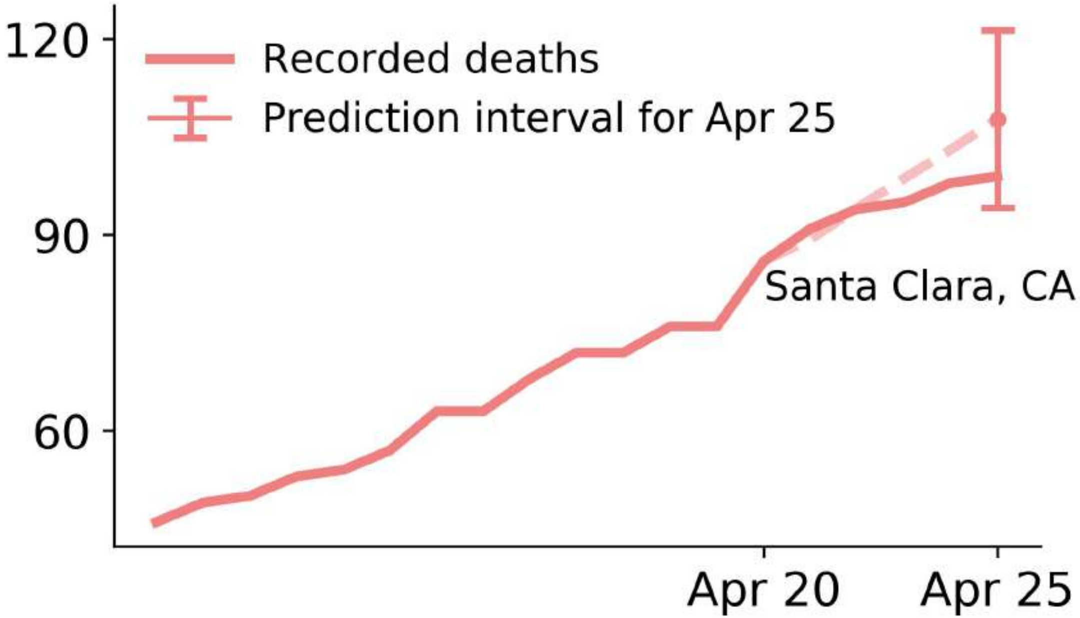


Previous 5-day-ahead prediction errors (%)

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

Prediction Intervals:



Predicted range of error

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

Actual error:

Apr 25 8.8%

Maximum (absolute) error prediction intervals (MEPI)

Step 1

Find normalized error of our predictor in the past.

$$\Delta_\tau := |y_\tau - \widehat{y}_\tau| / |\widehat{y}_\tau|.$$

Step 2

Find maximum error of past 5 days.

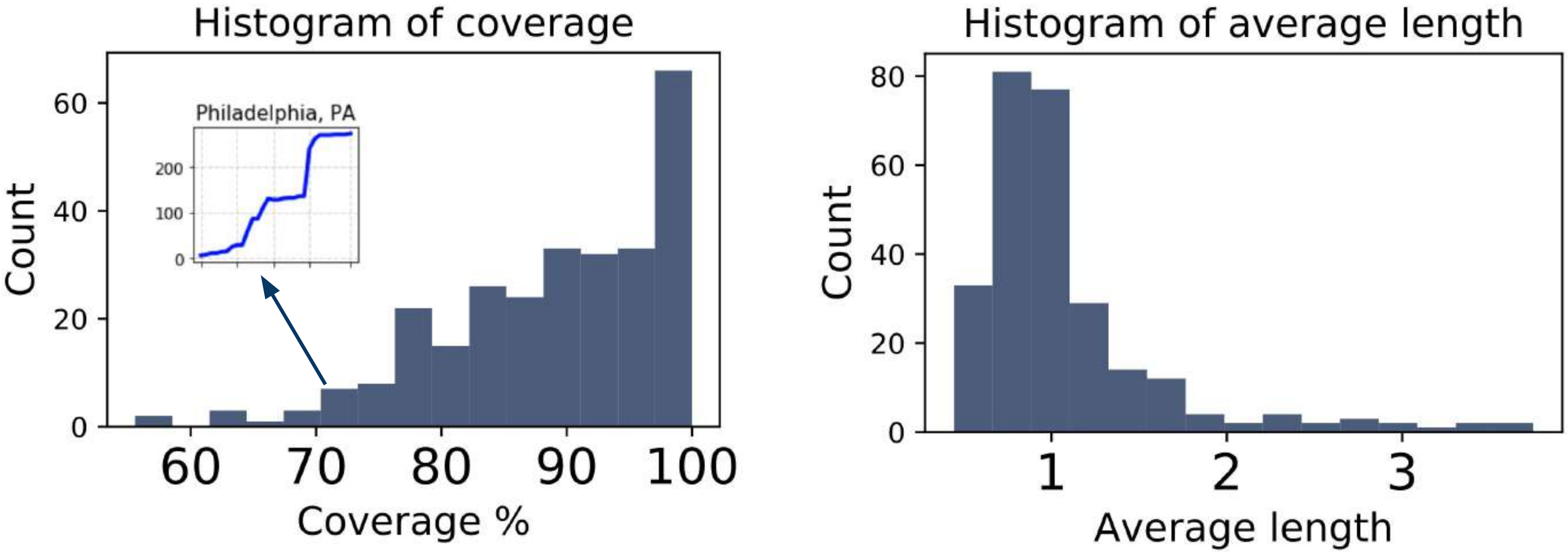
$$\Delta_{\max} := \max_{0 \le j \le 4} \Delta_{t-j}.$$

Step 3

$$\widehat{\mathrm{PI}}_{t+k} := \left[\max\left\{\widehat{y}_{t+k}(1 - \Delta_{\max}), y_t\right\},\ \widehat{y}_{t+k}(1 + \Delta_{\max})\right]$$

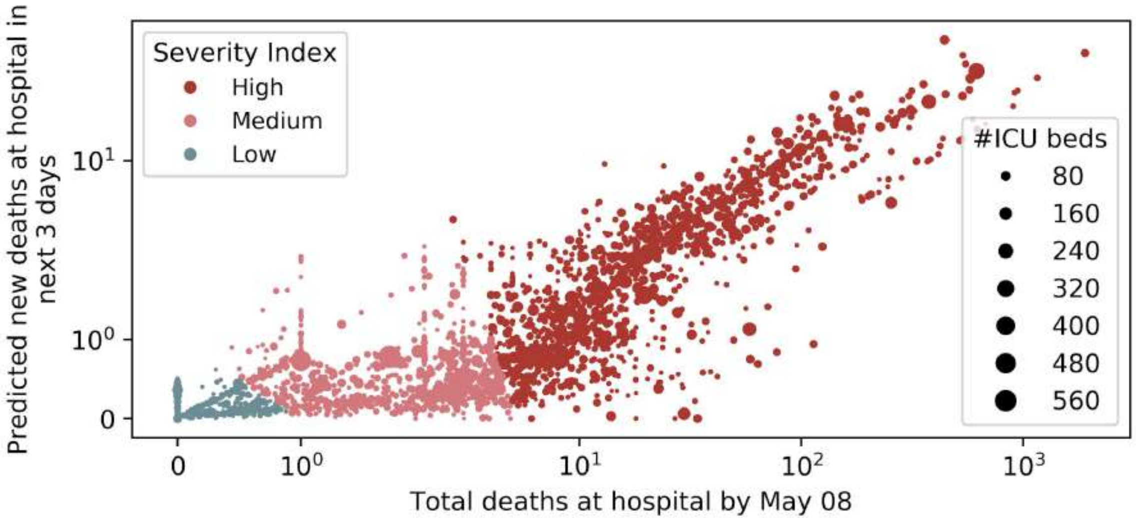
Can be applied to any ML model, and it works well under **exchangeability** condition on the errors.

Empirical performance of MEPI



Evaluation period: March 28--April 27. Only include days since the county has 10 deaths.
Having a normalized length of 0.8 means the PI is roughly (0.6 $\widehat{y}_{t+k}$, 1.4 $\widehat{y}_{t+k}$).

Severity Index

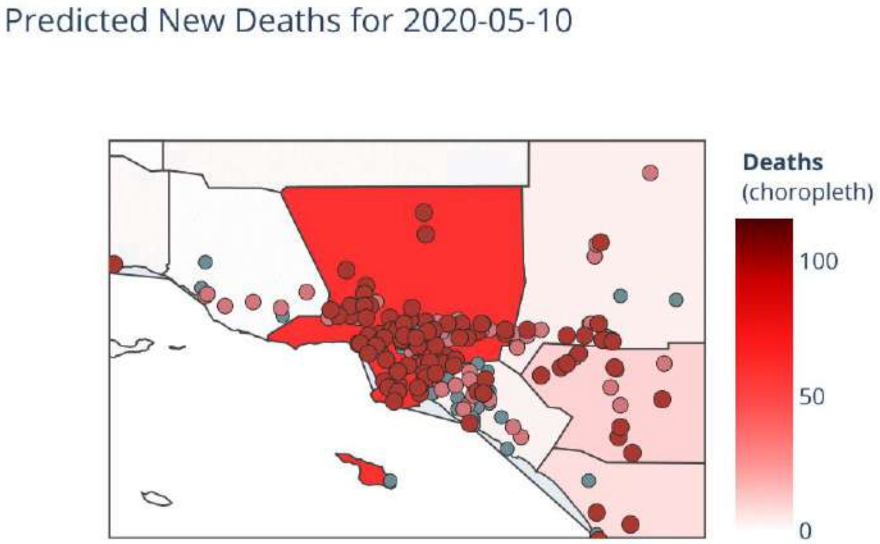


A score* for each hospital based on:

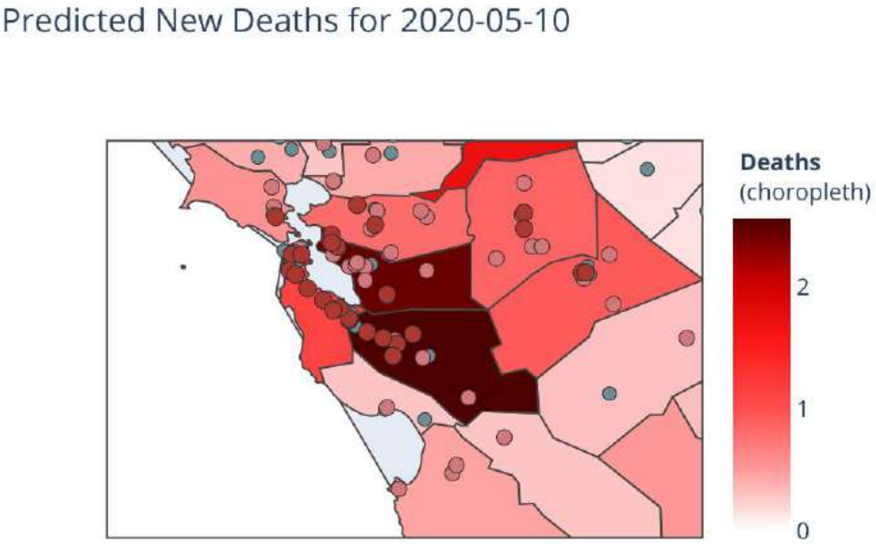
1. Predicted cumulative deaths
2. Predicted daily deaths

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

Mapping Deaths and the Hospital Severity Index Over Time



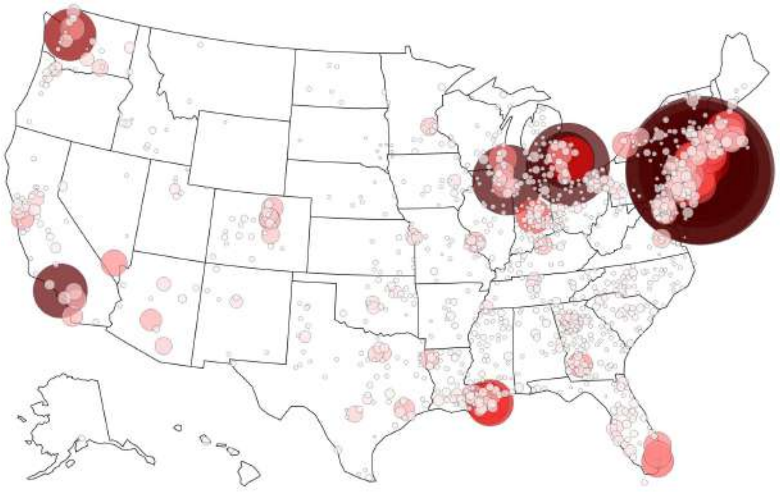
Los Angeles



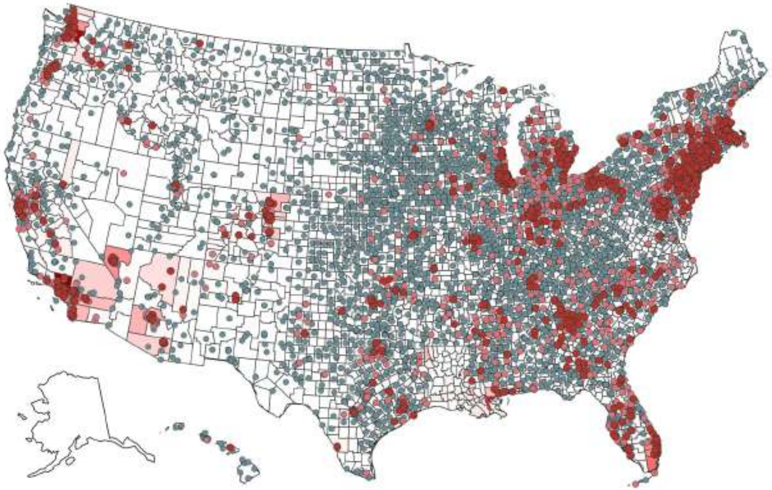
Bay Area

(Interactive) map visualizations

County-level predicted cumulative # of deaths*



Hospital severity index*



*Maps for 04/15

THE CENTER FOR SPATIAL DATA SCIENCE THE UNIVERSITY OF CHICAGO

Collaborating with the Center for Spatial Data Science (**CSDS**) at **University of Chicago** to add our predictions and severity index to the U.S. COVID-19 Atlas.

5000 Face Shields arrived at Temple Univ Hospital on May 8

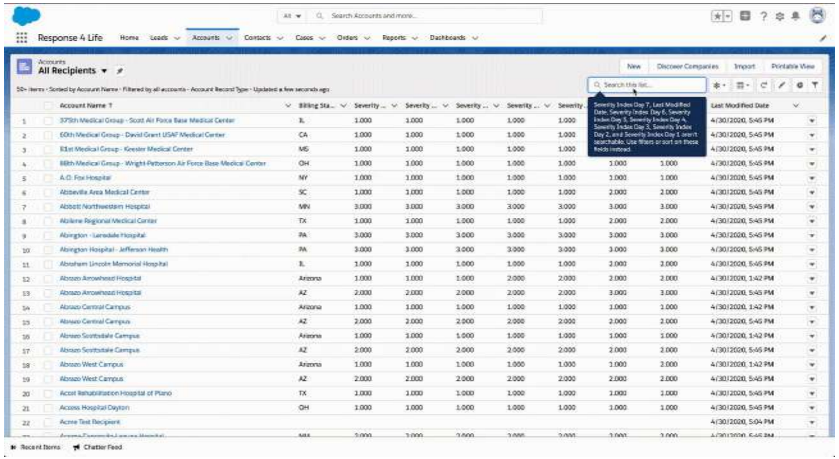
Don Landwirth, R4L

Impacts through Response4life

500k face shields in US by the end of may

- 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 in 2 weeks**
- **+500k outside the US**

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



Impact of our work beyond R4L

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

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

- Results on CSDS atlas at Univ of Chicago
- Final project option for DS 100 at Berkeley (> 1000 students) and Stat 542 at University of Illinois Urbana-Champaign (graduate stat-ml course)
- Possible causal inference through matching of counties
- Possible collaboration with California Department of Public Health (?)

Paper available at tinyurl.com/yugroup-covid19 and at

Curating a COVID-19 data repository and forecasting county-level death counts in the United States

Curating a COVID-19 data repository and forecasting county-level death counts in the United States

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

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

April 29, 2020

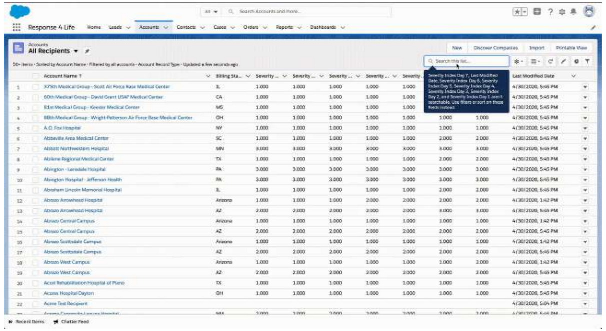
†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

Impacts

500k face shields in US by mid-may

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- **+65k to 25 recipients in 15 states in 2 weeks**
- **+500k outside the US**

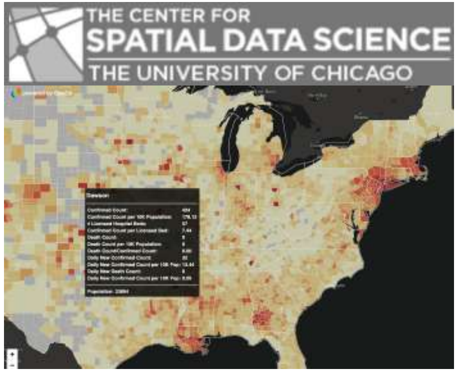
Salesforce system



Curating data repository

Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

Visualizations

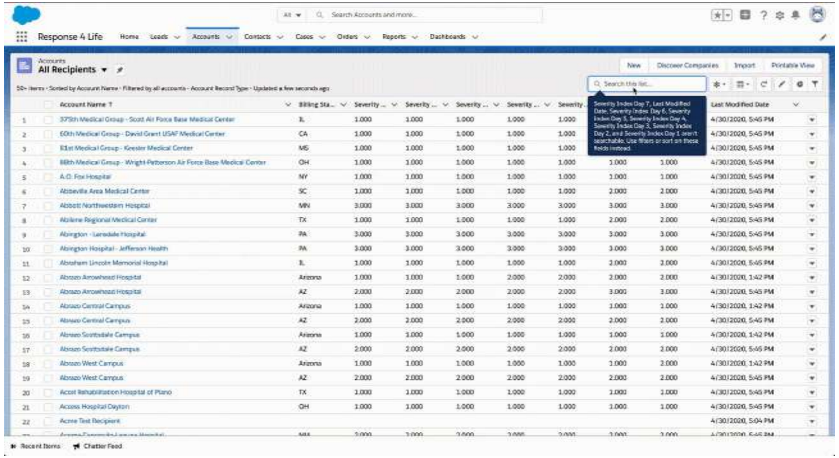


Impacts through Response4life

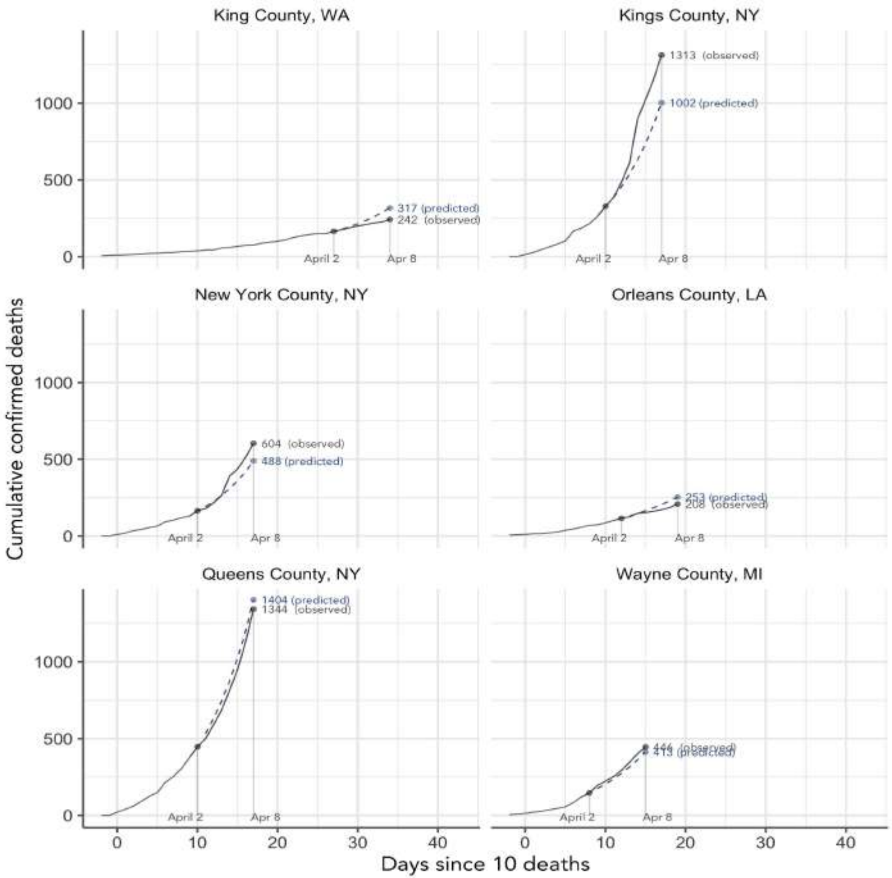
500k face shields in US by the end of may

- 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 in 2 weeks**
- **+500k outside the US**

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



Our county-level 7-day predictive performance



Focusing on 6 of the worst-affected counties

*Based on 4/8 data

In partnership with

COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

Goal: Help Aid Resource Allocation

SICK DOCTORS, NURSES AND NOT ENOUGH EQUIPMENT: NYC HEALTH CARE WORKERS ON THE FIGHT AGAINST THE CORONAVIRUS

Health officials warn US government does not have enough stockpiled medical equipment to deal with coronavirus

RESPONSE 4LIFE

Our team

from UC Berkeley Statistics/EECS and UCSF

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

B. Park

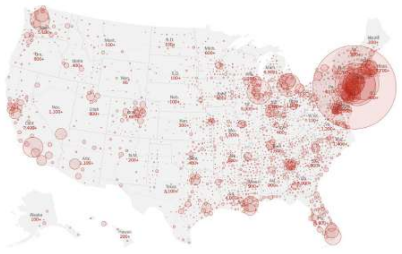
C. Singh
(Student Lead)

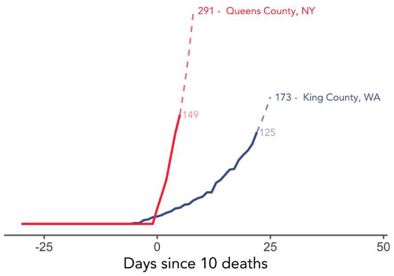
Y. Tan

T. Tang

Y. Wang

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





Data Curation

- Hospital data
- County data

Modeling

- County-level 7-day severity prediction
- hospital demand prediction

Evaluation / Visualization

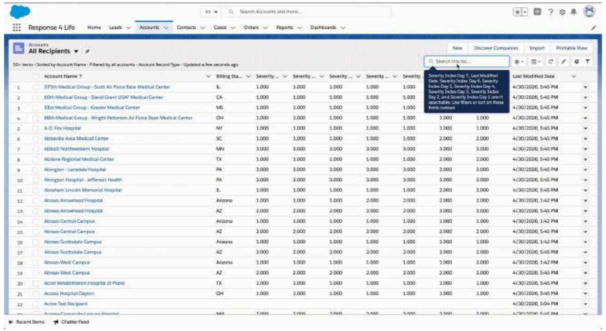
- Identify hotspots and risk factors via news articles
- Visualization
- Validate forecasts

Impacts

500k face shields in US by mid-may

- Santa Clara + Temple University Med Center in Philadelphia
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- **+65k to 25 recipients in 15 states in 2 weeks**
- **+500k outside the US**

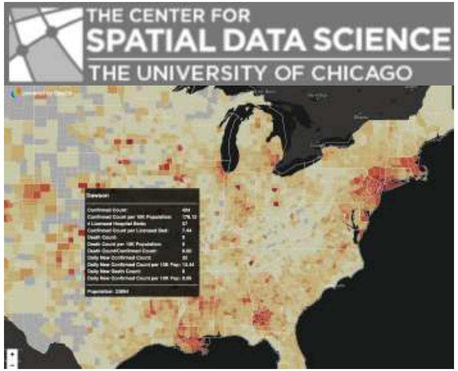
Salesforce system



Curating data repository

Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

Visualizations



Part I: Curating a COVID-19 Data Repository

Outline of Part I: Data Curation

- Our data processing pipeline
- Overview of the data
- Frequently overlooked aspects and challenges
- Some useful tools

Image Source: https://info.aldensys.com/joint-use/how-the-iot-will-move-companies-from-data-collection-to-data-driven

Data Processing Pipeline

Data Scraping	Data Cleaning	Data Validity
• Collect 1M records from 10+ data sources • Monitor data changes 24/7 powered by AWS	• Handling missing and erratic entries • Automated python script	• Compare data across different sources to ensure data validity • Search for emerging data sources

For almost a month, 2 full-time students, and on-going with 1 full-time student

Amazon EC2

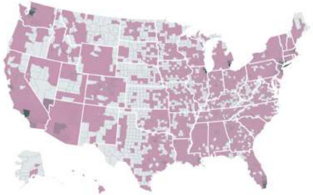
Data and code available: https://github.com/Yu-Group/covid19-severity-prediction

★ Being used by multiple research groups across the country

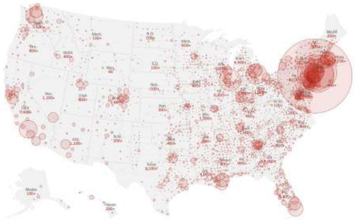
Data: scraped from a variety of sources

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

esri

COVID-19 GIS Hub

DEPARTMENT OF TRANSPORTATION UNITED STATES OF AMERICA

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

IHME GHDx

Introducing the Unacast
Social Distancing Scoreboard

USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

JOHNS HOPKINS UNIVERSITY

CMS.gov
Centers for Medicare & Medicaid Services

United States® Census Bureau

STREETLIGHT

KHN
KAISER HEALTH NEWS

SAFEGRAPH

cuebiq

kinsa®

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

HRSA
Health Resources & Services Administration

ArcGIS Hub

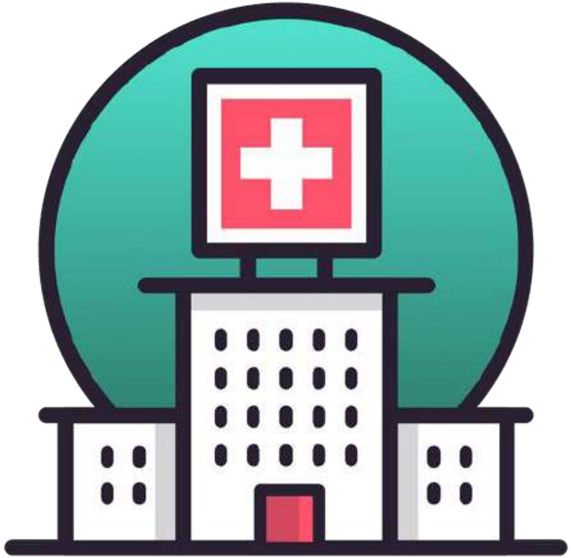
HOMELAND INFRASTRUCTURE FOUNDATION-LEVEL DATA SUBCOMMITTEE

Samuel Scarpino

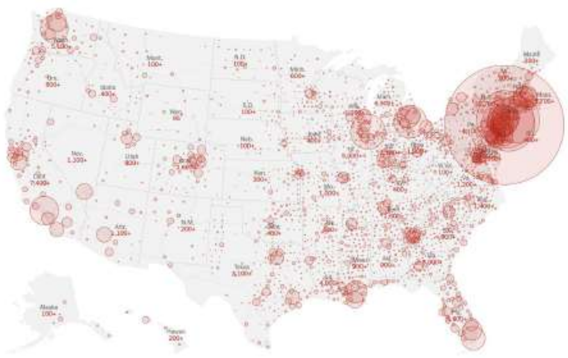
NORTHEASTERN UNIVERSITY

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



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



- 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

Our **data repository** can be found at the following link:

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

Now a little journey through cleaning the USAFacts COVID-19 cases/deaths data…

A journey through cleaning the USAFacts COVID-19 cases/deaths data

Got the data from website

countyFIPS	County Nar	State	stateFIPS	1/22/2020	1/23/2020
0	Statewide U	AL	1	0	0
1001	Autauga Cc	AL	1	0	0
1003	Baldwin Co	AL	1	0	0
1005	Barbour Co	AL	1	0	0
1007	Bibb Count	AL	1	0	0
1009	Blount Cou	AL	1	0	0
1011	Bullock Cou	AL	1	0	0
1013	Butler Cour	AL	1	0	0
1015	Calhoun Co	AL	1	0	0
1017	Chambers C	AL	1	0	0
1019	Cherokee C	AL	1	0	0
1021	Chilton Cou	AL	1	0	0
1023	Choctaw Cc	AL	1	0	0
1025	Clarke Cour	AL	1	0	0
1027	Clay County	AL	1	0	0
1029	Cleburne C	AL	1	0	0
1031	Coffee Cou	AL	1	0	0
1033	Colbert Cou	AL	1	0	0
1035	Conecuh Cc	AL	1	0	0
1037	Coosa Cour	AL	1	0	0
1039	Covington (	AL	1	0	0
1041	Crenshaw C	AL	1	0	0
1043	Cullman Co	AL	1	0	0
1045	Dale Count	AL	1	0	0
1047	Dallas Cour	AL	1	0	0
1049	DeKalb Cou	AL	1	0	0
1051	Elmore Cou	AL	1	0	0

Some counties are duplicated.

Some cases/deaths cannot be allocated to a county.

Cumulative deaths counts sometimes decrease.

Are we done? 🤔

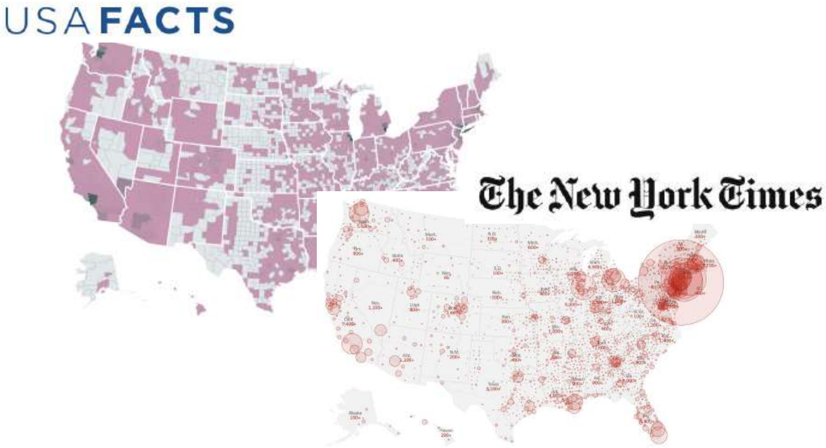
A journey through cleaning the USAFacts COVID-19 cases/deaths data

Things got interesting when multiple data sources are available.

Got the data from nytimes! → **different counties?** → **USAFacts do not cover all the counties.**

→ **Some counties changed their countyFIPS code.**

→ **NYTimes aggregated some counties together.**



A journey through cleaning the USAFacts COVID-19 cases/deaths data

Things got interesting when multiple data sources are available.

Got the data from nytimes!

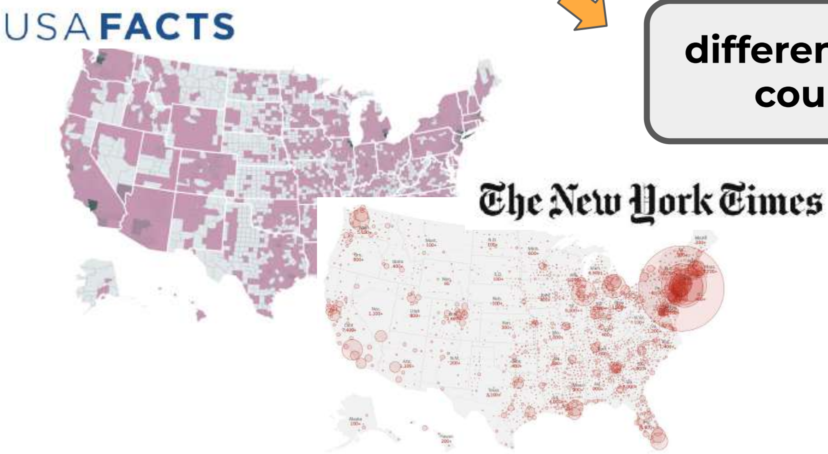
different counties?

different death counts?

Different time cut-off: CDC as of 4 PM ET

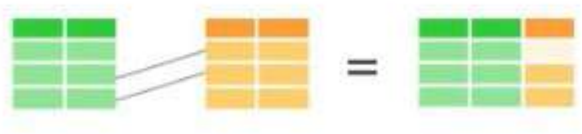
How to deal with probable deaths?

How to count one person who lives at A, tested at B, go to hospital at C? Flight travels?



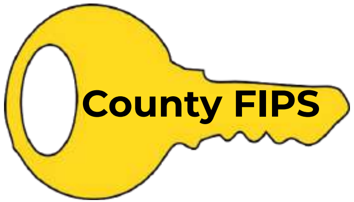
Multiple data sources give us insights into the caveats of the data.

Additional challenges in data cleaning



- What is a “primary key”?
  - Use primary key to merge different sources of data together.
  - Ideally, key should be stable over time and no duplicates.

For county-level data



County FIPS can change over time (though this is rare)

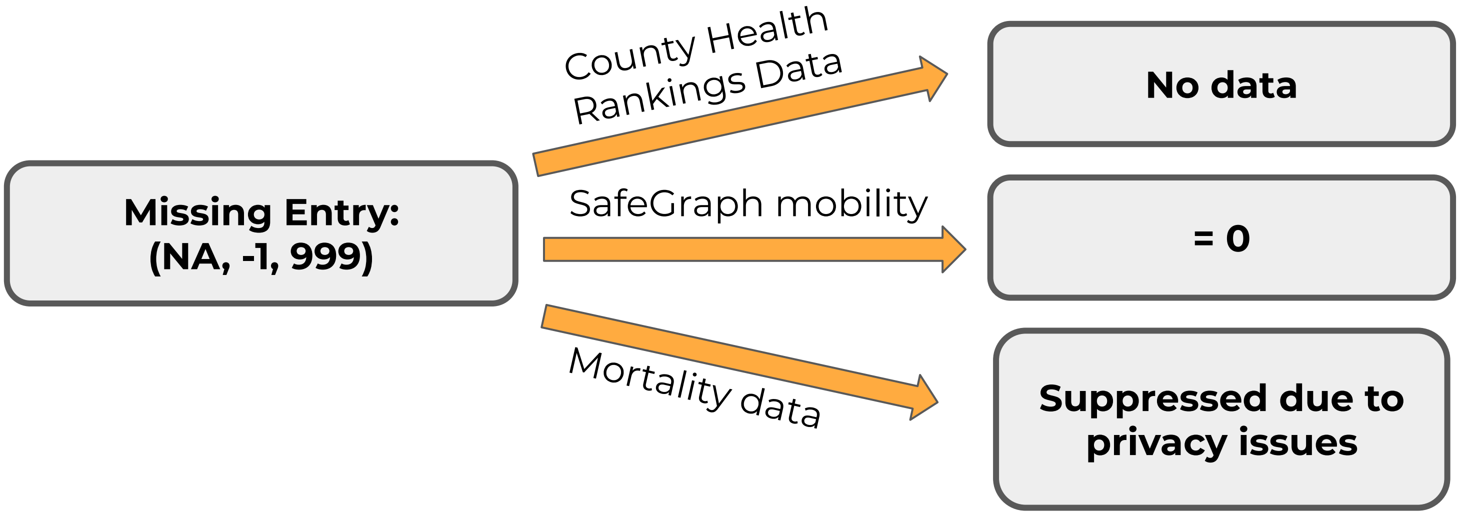
For hospital-level data

Not all the hospitals have this number (e.g., Indian reservation hospitals)

For commute and county adjacency data

Additional challenges in data cleaning

- Missing data entries
  - Encoded as NAs, -1, 999, and more...
  - Meaning can depend on the data set



Frequently overlooked questions

- Who is the audience or end user?
  - How to present the data to make it easily accessible by our modeling team, visualization team, and other researchers in the broader community
    - Clear documentation
    - Abridged version and unabridged version of the county-level data

Image Source: https://medium.com/@manjunath2137/how-to-create-the-right-target-audience-for-your-facebook-ad-9b5ed562b35f

Frequently overlooked questions

- Who is the audience or end user?
  - How to present the data to make it easily accessible by our modeling team, visualization team, and other researchers in the broader community
    - Clear documentation
    - Abridged version and unabridged version of the county-level data

readme.md

Interactive Atlas of Heart Disease and Stroke - All Strokes (2014-2016)

- **Data source**: https://www.cdc.gov/dhdsp/maps/atlas/index.htm
- **Last downloaded**: 04/02/2020
- **Data description**: county-level estimates of mortality rates per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Known data quality issues**: Data values within the table of "-1" or "-9999" indicate "Insufficient Data."
- **Short list of data columns**:
  - **countyFIPS**: county FIPS
  - **StrokeMortality**: estimate of mortality rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016) from all strokes (ICD10 codes: I60-I69)
- **Notes**:
  - Data downloaded from the Interactive Atlas of Heart Disease and Stroke, a website developed by the Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention. http://nccd.cdc.gov/DHDSPAtlas.

271 lines (250 sloc) | 33 KB

List of columns - county level

Identifying variables

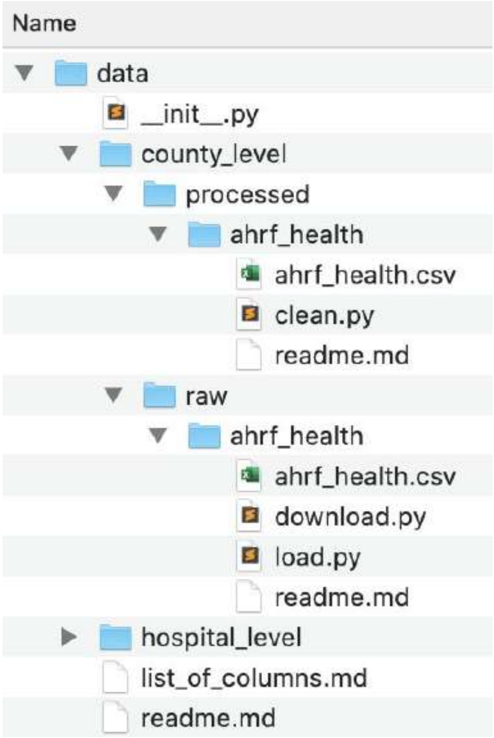
Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

Data variables

Geographical identifiers

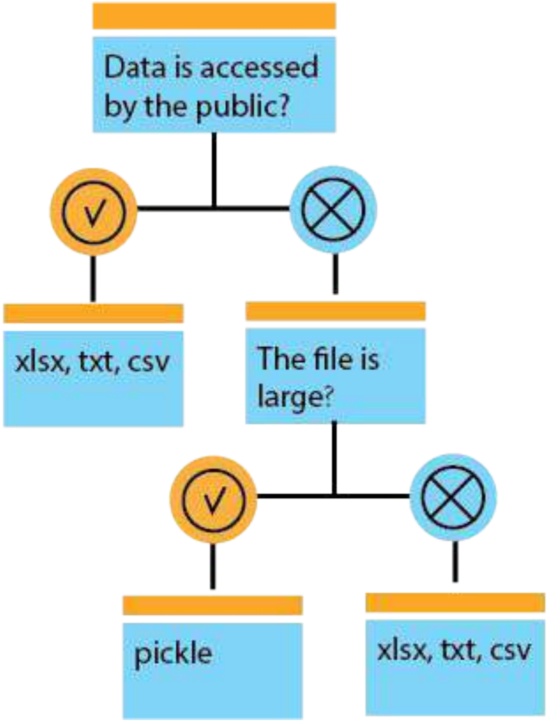
Frequently overlooked questions

- What are the naming conventions and organization structure for data storage and preprocessing?
  - Improves accessibility for end users
  - Necessary to quickly integrate new members and volunteers
  - Best to set standards at the beginning
  - But this is very challenging because:
    - A good convention depends on the data we collect but we don't know what data will be there.
    - Some data sets might change over time



Frequently overlooked questions

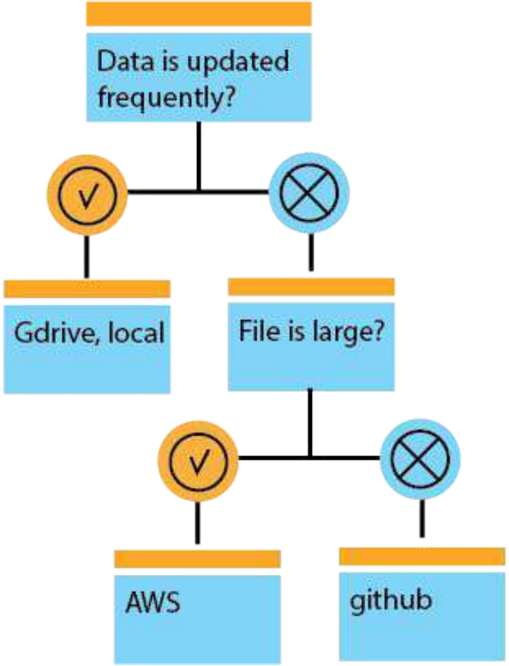
- Which file format?
  - txt, csv, pickle, xlsx
  - compressed versions



Frequently overlooked questions

- Which file format?
  - txt, csv, pickle, xlsx
  - compressed versions

- How to store the data?
  - Locally
  - GitHub
  - AWS
  - Google drive



The data team is at its best when working closely alongside everyone on the team

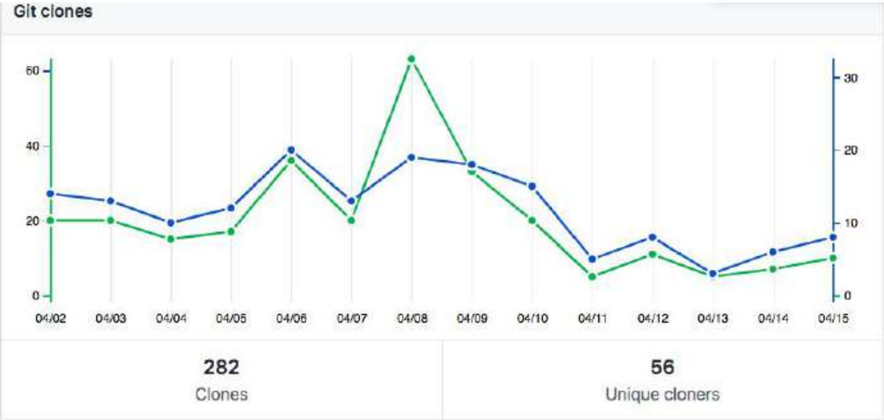
- In particular, modeling team depends on data team AND data team depends on modeling team
    - Determine what are relevant data sets
    - Iterative process between two teams to figure out how to clean the data

Overview of some useful tools

- Git commands: pull, push, merge conflicts
- Linux commands
  - shell commands
  - wget
    - Can easily download data from online source (including google drive)
  - cron jobs
    - To automatically update data, predictions, and visualizations daily
- AWS package
  - S3 buckets
- Google cloud package (update google sheet)

Summary: Data and code

Data repository:
https://github.com/Yu-Group/covid19-severity-prediction



Summary: Paper

Paper available:

Curating a COVID-19 data repository and forecasting county-level death counts in the United States

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

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

April 29, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

In partnership with

COVID-19 Data Repository and Severity Prediction

Yu Group
UC Berkeley Statistics, EECS, CCB

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

Website: covidseverity.com

Goal: Help Aid Resource Allocation

RESPONSE 4LIFE

Our team

from UC Berkeley Statistics/EECS and UCSF

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

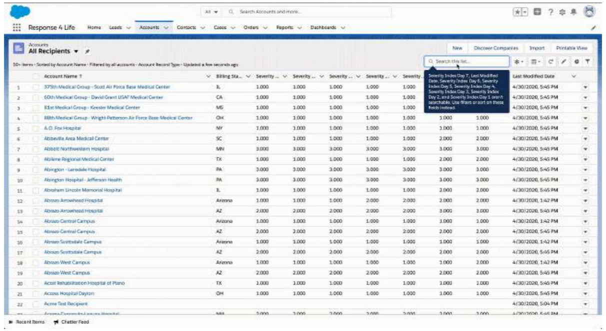
Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...

Impact

500k face shields in US by mid-may

- Santa Clara + Temple University Med
- in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- +65k to 25 recipients in 15 states in 2 weeks
- many more expected

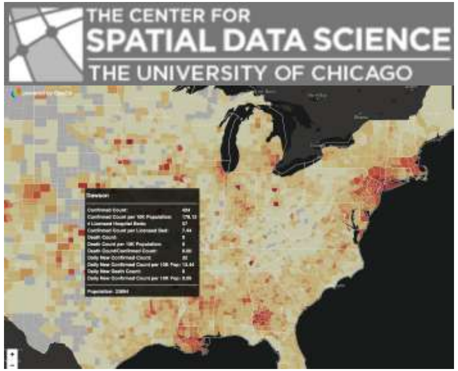
Salesforce system



Data Repository and Code Base

Data variable	Description	Source data set
countyFIPS	state-county FIPS Code	county_fips
STATEFP	state FIPS Code	county_popcenters
COUNTYFP	county FIPS Code	county_popcenters
CountyName	county name	county_fips
StateName	state abbreviation	county_fips
State	state name	county_latlong

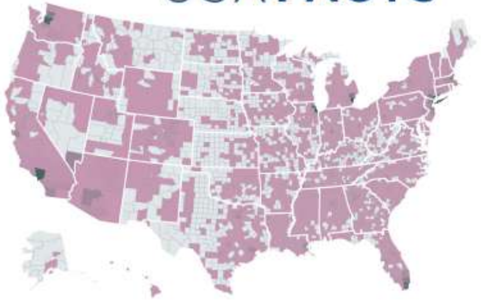
Visualizations



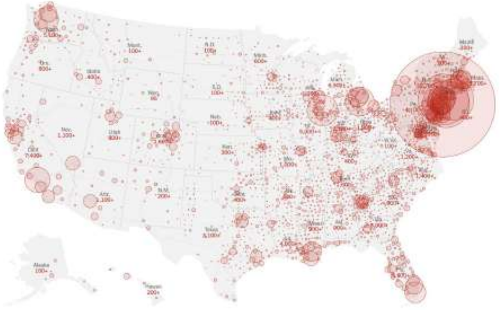
Last Week: Curating a COVID-19 Data Repository

Covid 19 Cases/Deaths

USAFACTS



The New York Times



Risk Factors, Demographics, County-level Data

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

esri COVID-19 GIS Hub

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

USDSS UNITED STATES DIABETES SURVEILLANCE SYSTEM
Division of Diabetes Translation, CDC

IHME GHDx

Introducing the Unacast
Social Distancing Scoreboard

United States® Census Bureau

CMS.gov
Centers for Medicare & Medicaid Services

kinsa®

KHN KAISER HEALTH NEWS

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

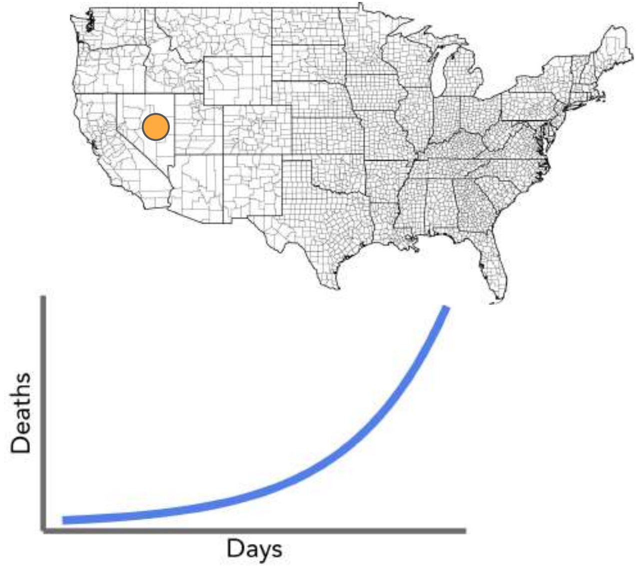
This Week: Forecasting death counts

Ensemble different predictors

We combined many different prediction approaches

Ensemble predictors

1. Separate-county **exponential** model[1]

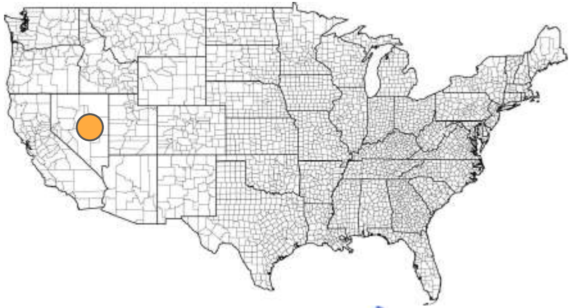


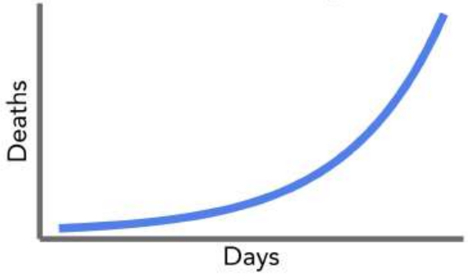
[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Ensemble predictors

We combined many different model approaches

1. Separate-county **exponential** model[1]





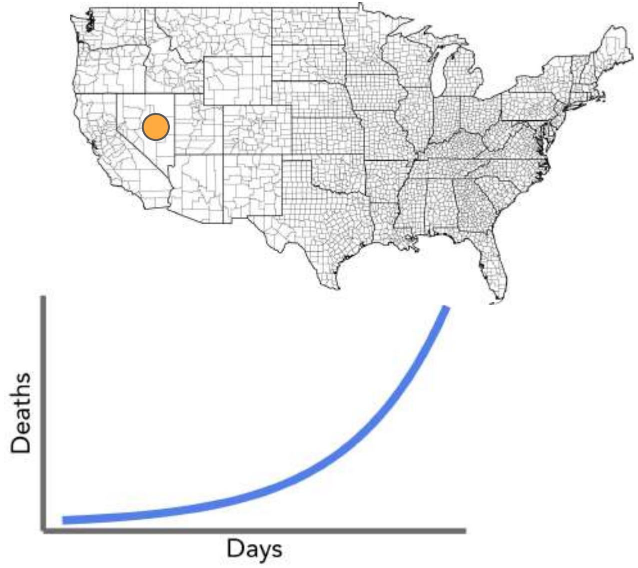
$$\mathrm{E}(\text{deaths}_t \mid t) = e^{\beta_0 + \beta_1 t}$$

[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

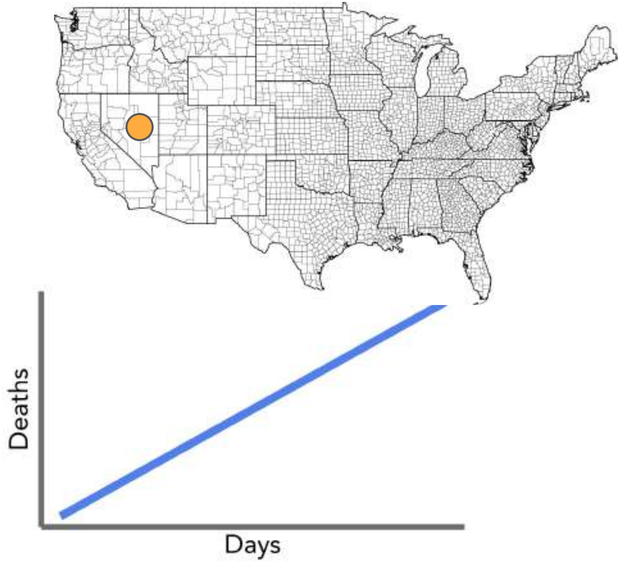
Ensemble predictors

We combined many different model approaches

1. Separate-county **exponential** model[1]



2. Separate-county **linear** model

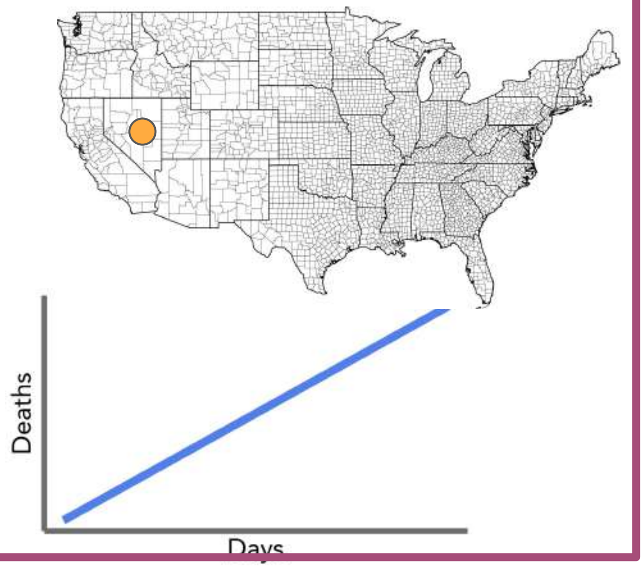


[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Ensemble predictors

We combined many different model approaches

2. Separate-county **linear** model



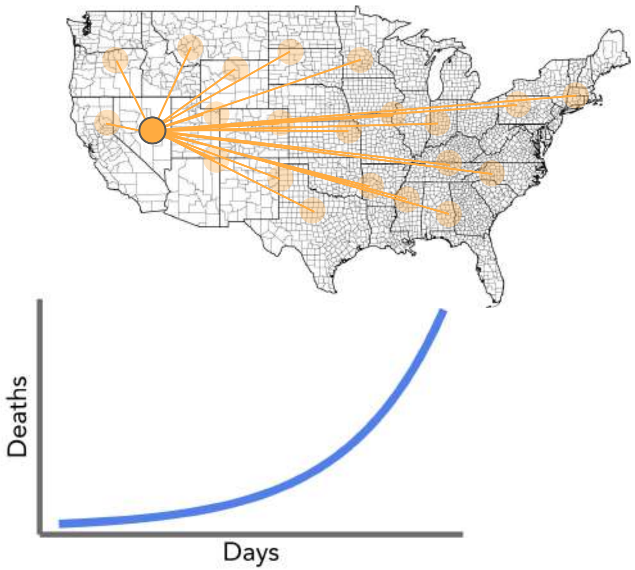
$$\mathrm{E}[\mathrm{deaths}_t | t] = \beta_0 + \beta_1 t$$

[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Ensemble predictors

We combined many different prediction approaches

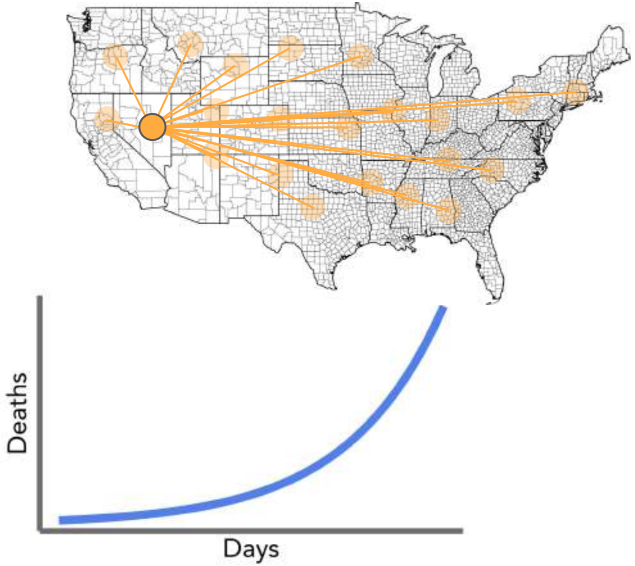
3. **<u>Shared</u>**-county exponential model



Ensemble predictors

We combined many different prediction approaches

3. **Shared**-county exponential model

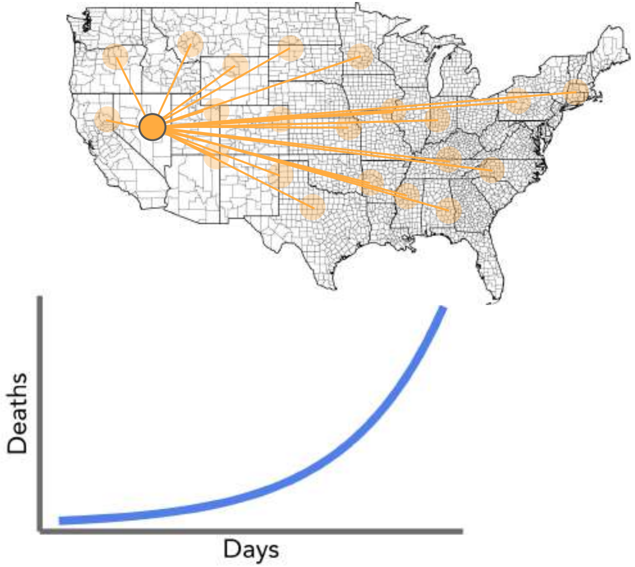


$$\mathrm{E}(\mathrm{deaths}_t \mid t) = e^{\beta_0 + \beta_1 \log(\mathrm{deaths}_{t-1} + 1)}$$

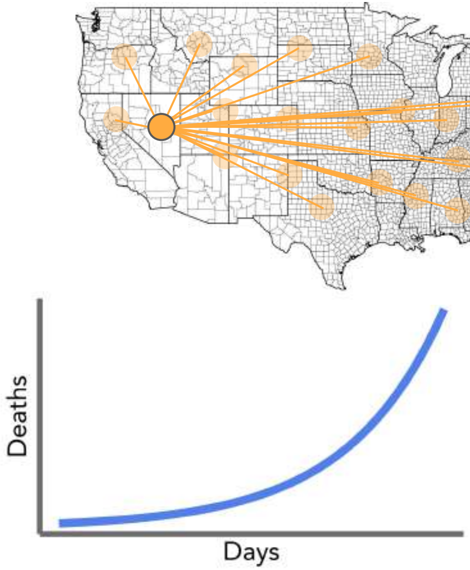
Ensemble predictors

We combined many different prediction approaches

3. **Shared**-county exponential model



4. **Shared**-county exponential + **demographics** model

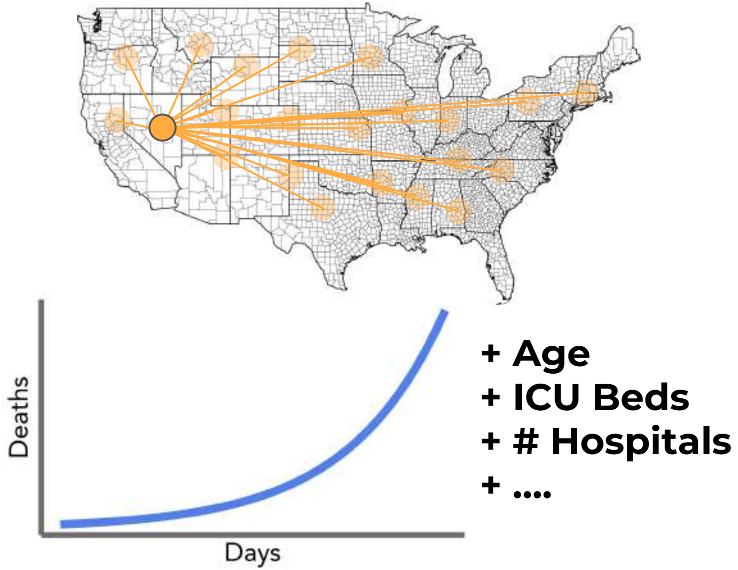


+ Age
+ ICU Beds
+ # Hospitals
+ ....

Ensemble predictors

We combined many different prediction approaches

4. **Shared**-county exponential + **demographics** model

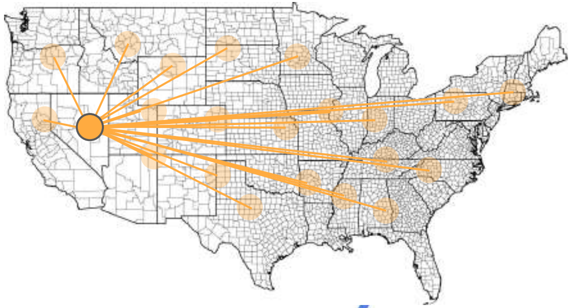


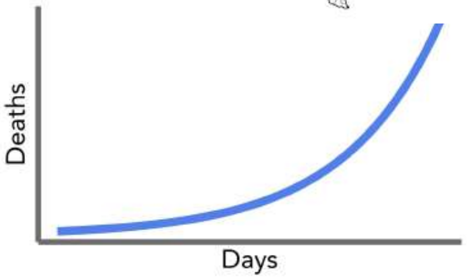
- County density and size
- County healthcare resources
- Demographic information

Ensemble predictors

We combined many different prediction approaches

5. **Expanded Shared**-county exponential model

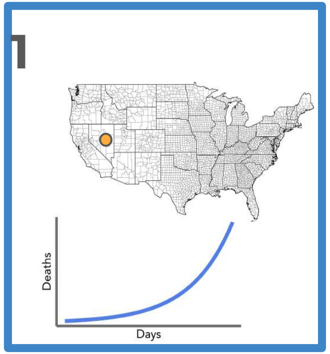




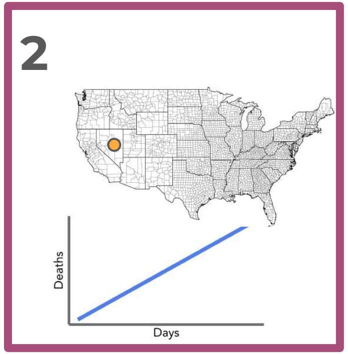
- \+ **Cases**
- \+ **Neighboring cases**
- \+ **Neighboring deaths**

- log(Cases)
- log(Cases in Neighboring counties)
- log(Deaths in neighboring counties)

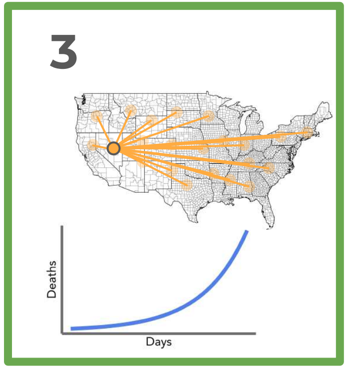
Combined Linear and Exponential Predictors (CLEP)



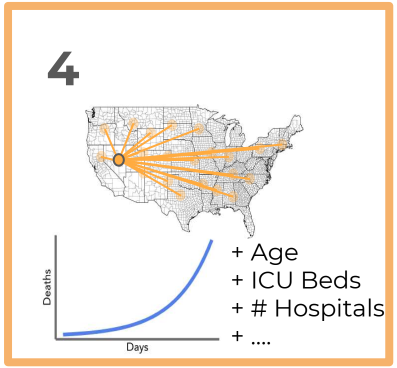
Separate-county exponential predictor



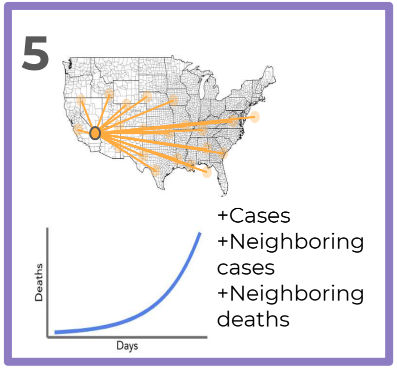
Separate-county linear predictor



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 historical performance[2]

[2]. Schuller, Gerald DT, et al. "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 historical performance[2]

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

[2]. Schuller, Gerald DT, et al. "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)

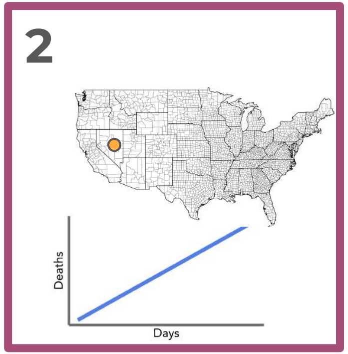
Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

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

[2]. Schuller, Gerald DT, et al. "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

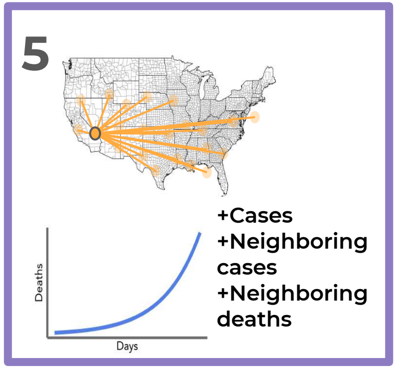
Combined Linear and Exponential Predictors (CLEP)

A smaller combination performed better in practice:



Separate-county linear predictor

+

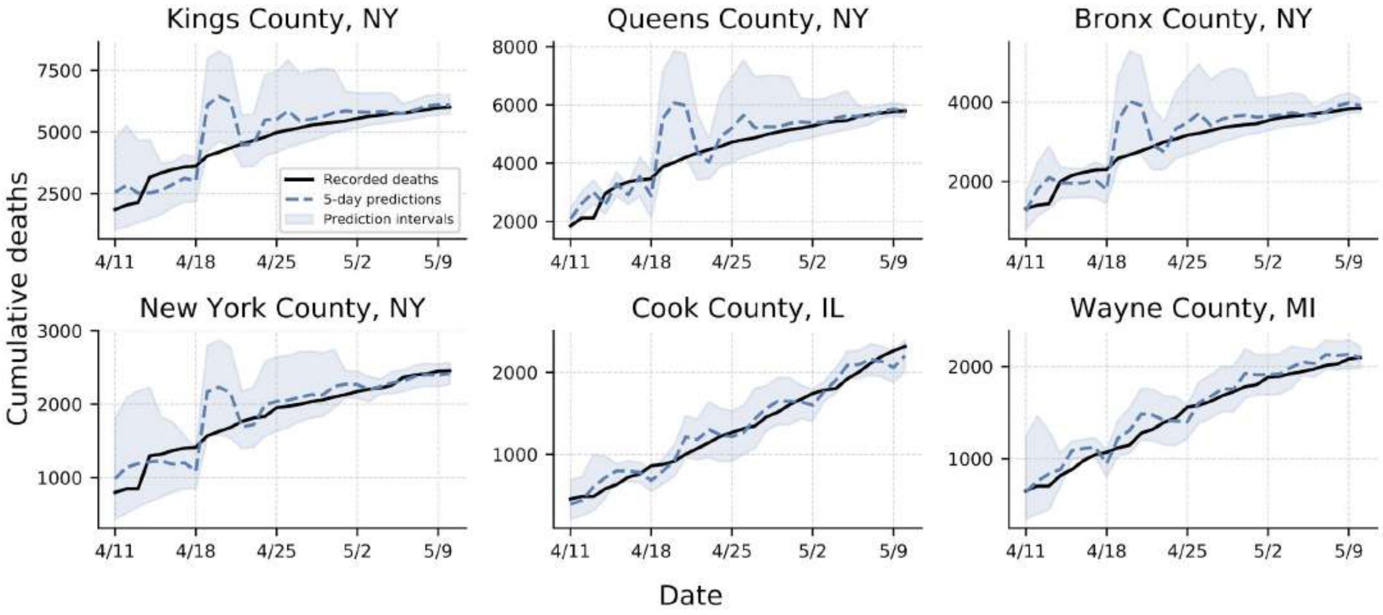


Expanded Shared-county exponential predictor

Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

[2]. Schuller, Gerald DT, et al. "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

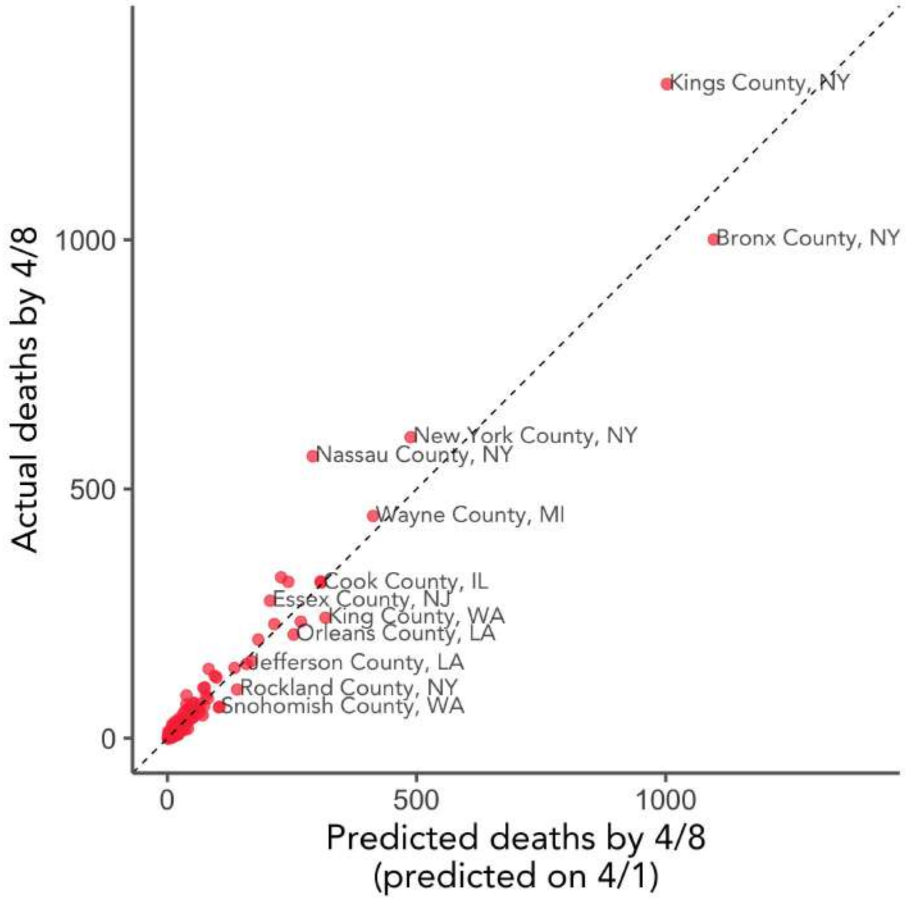
Our county-level 7-day predictive performance



Focusing on 6 of the worst-affected counties

*Based on 4/8 data

Our county-level 7-day predictive performance



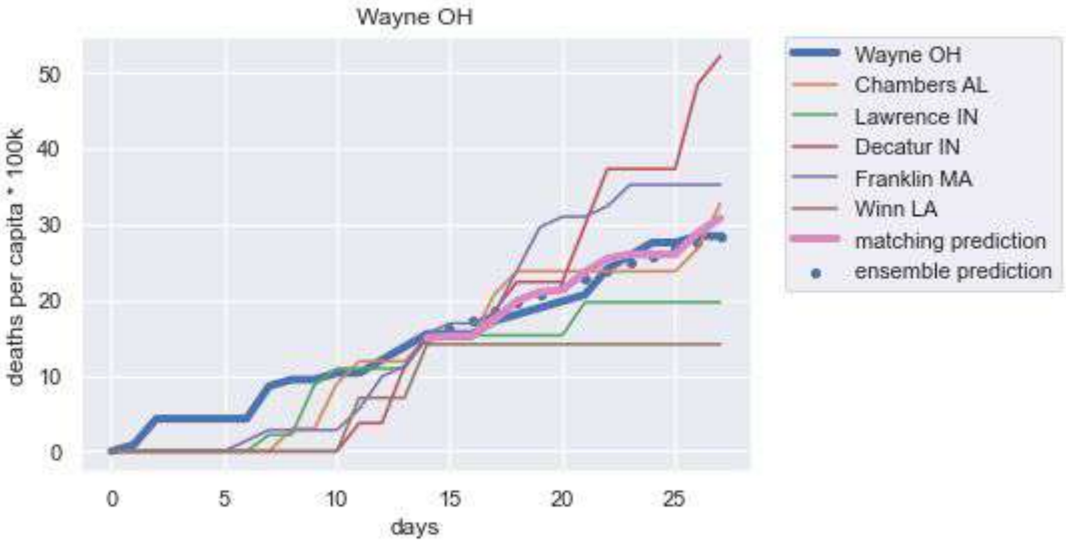
Takeaway:
The 7-day forecasted predictions are fairly accurate

"Actual deaths" : recorded deaths by a given day

Ongoing Work:

County Matching

Matching Counties:

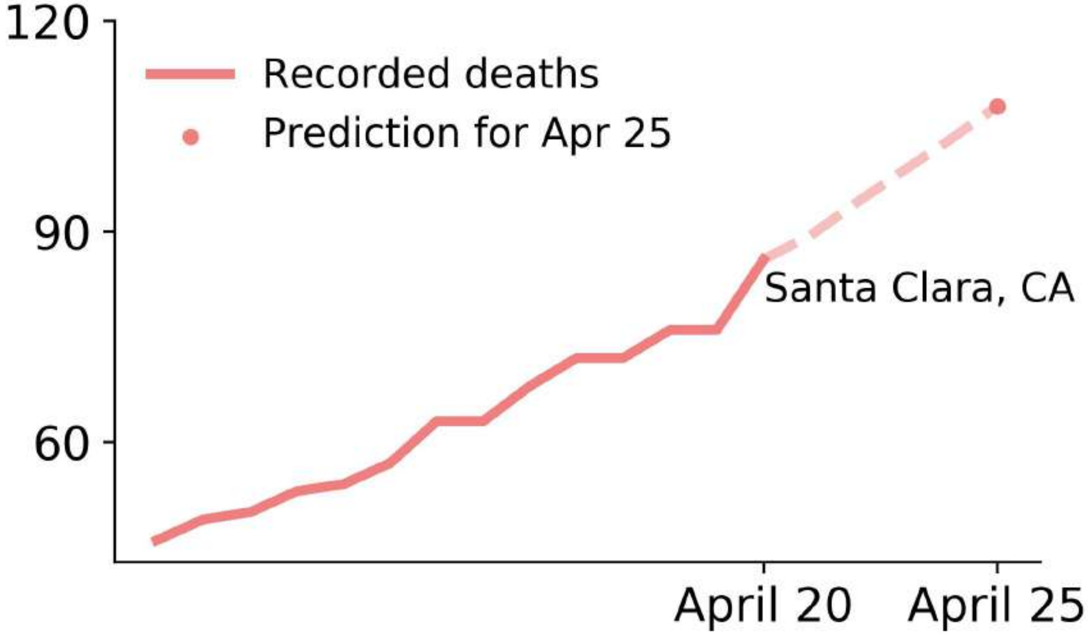


Find similar counties and use these to predict trajectory

Prediction Intervals:

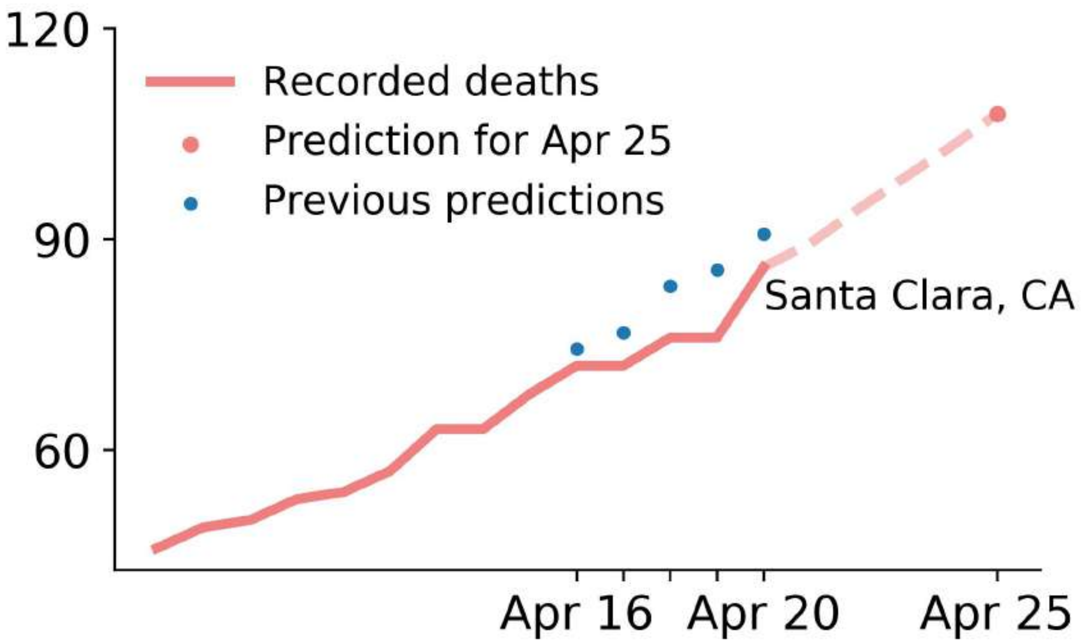
How confident should we be about our predictions?

Prediction Intervals:



How confident are we with the prediction for April 25?

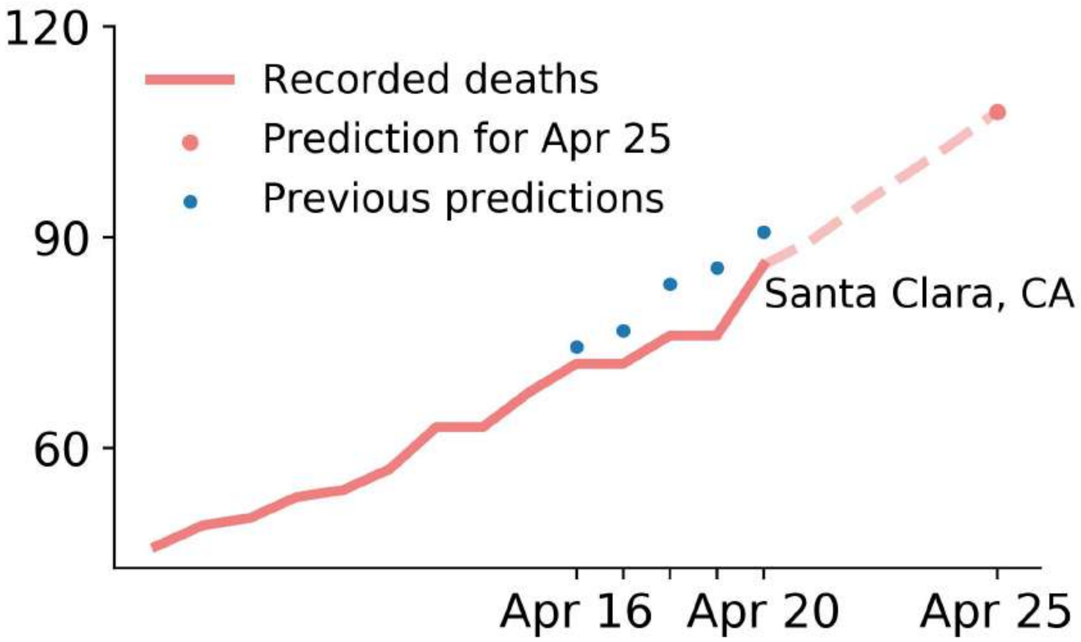
Prediction Intervals:



How confident are we with this prediction for April 25?

Use **past experience** to determine confidence in new predictions.

Prediction Intervals:

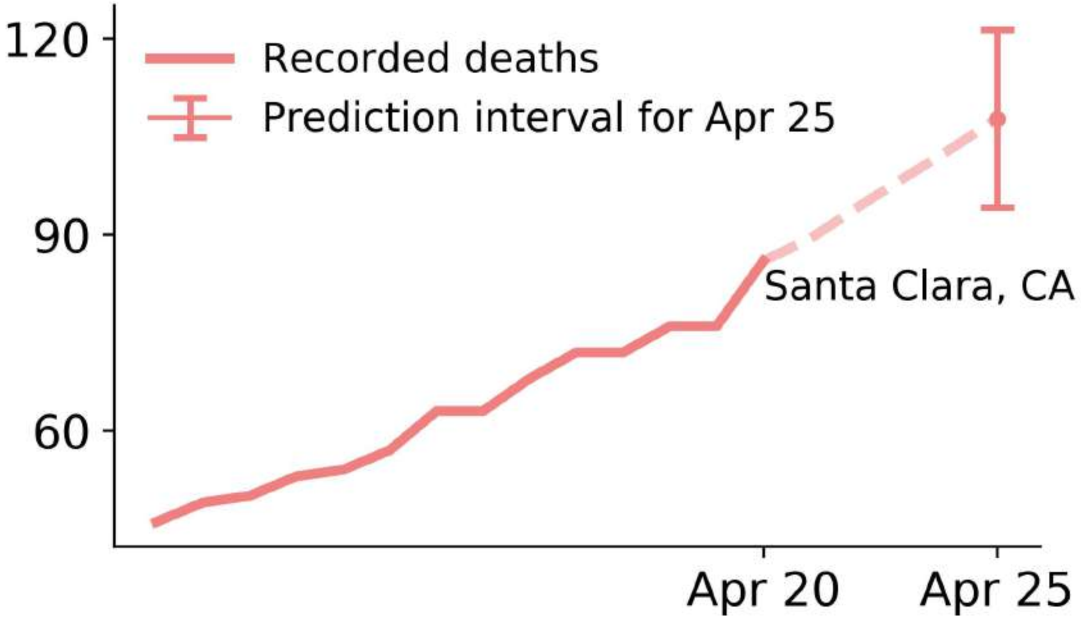


Previous 5-day-ahead prediction errors (%)

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

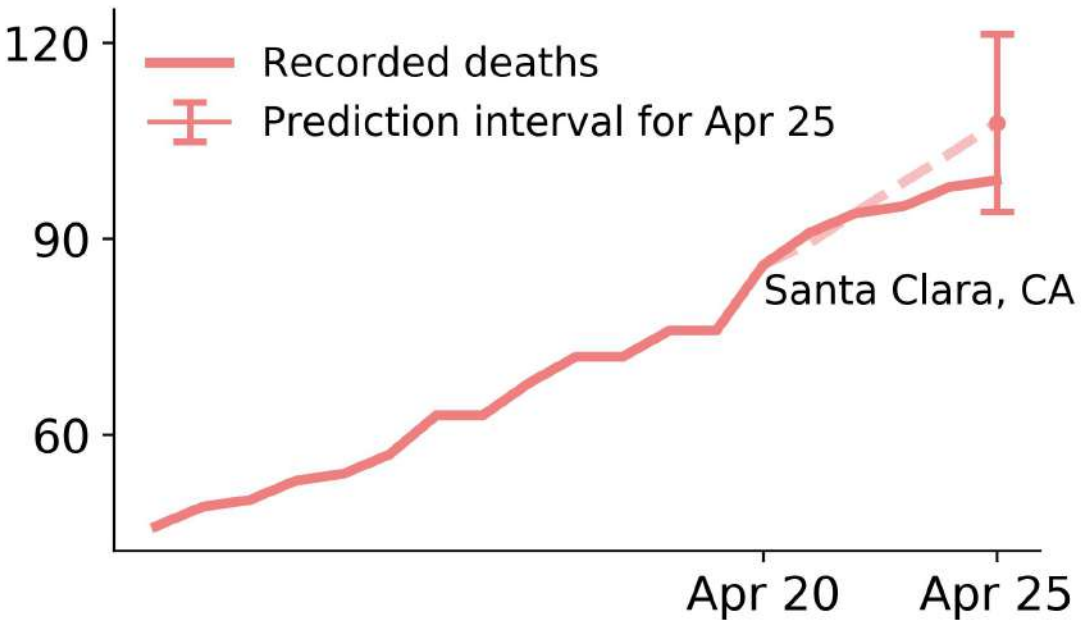
Prediction Intervals:



Predicted range of error

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

Prediction Intervals:



Predicted range of error

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

Actual error:

Apr 25 8.8%

Maximum (absolute) error prediction intervals (MEPI)

Step 1

Find normalized error of our predictor in the past.

$$\Delta_\tau := |y_\tau - \widehat{y}_\tau| / |\widehat{y}_\tau|.$$

Step 2

Find maximum error of past 5 days.

$$\Delta_{\max} := \max_{0 \le j \le 4} \Delta_{t-j}.$$

Step 3

$$\widehat{\mathrm{PI}}_{t+k} := \left[\max\left\{\widehat{y}_{t+k}(1 - \Delta_{\max}), y_t\right\},\ \widehat{y}_{t+k}(1 + \Delta_{\max})\right]$$

Can be applied to any ML model!

Connection to conformal inference[1], [2]

General conformal inference recipe: **95% percentile** of **all past errors**

MEPI: **max** of **past 5 errors**

[1] G. Shafer and V. Vovk. A tutorial on conformal prediction. Journal of Machine Learning Research, 9(Mar):371–421, 2008.

[2] V. Vovk, A. Gammerman, and G. Shafer. Algorithmic learning in a random world. Springer Science & Business Media, 2005

Connection to conformal inference[1], [2]

General conformal inference recipe: **95% percentile** of **all past errors**

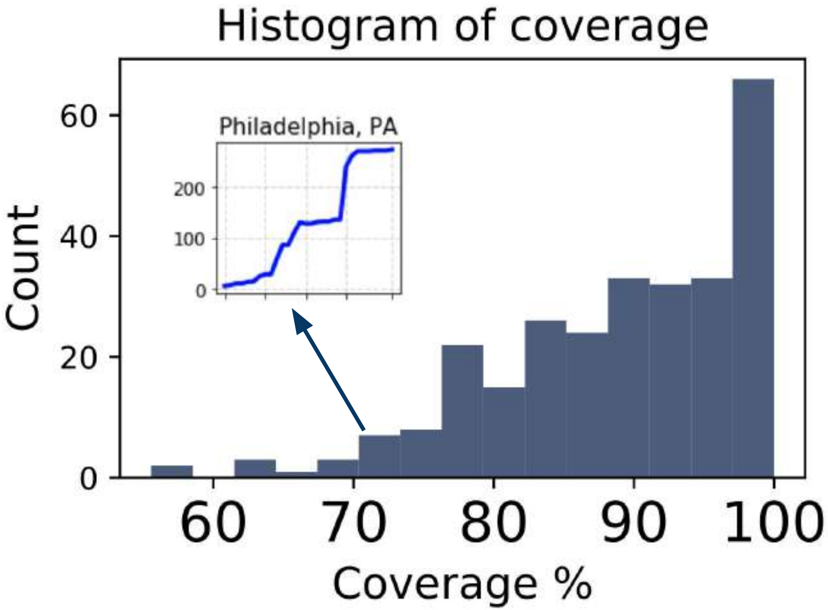
MEPI: **max** of **past 5 errors**

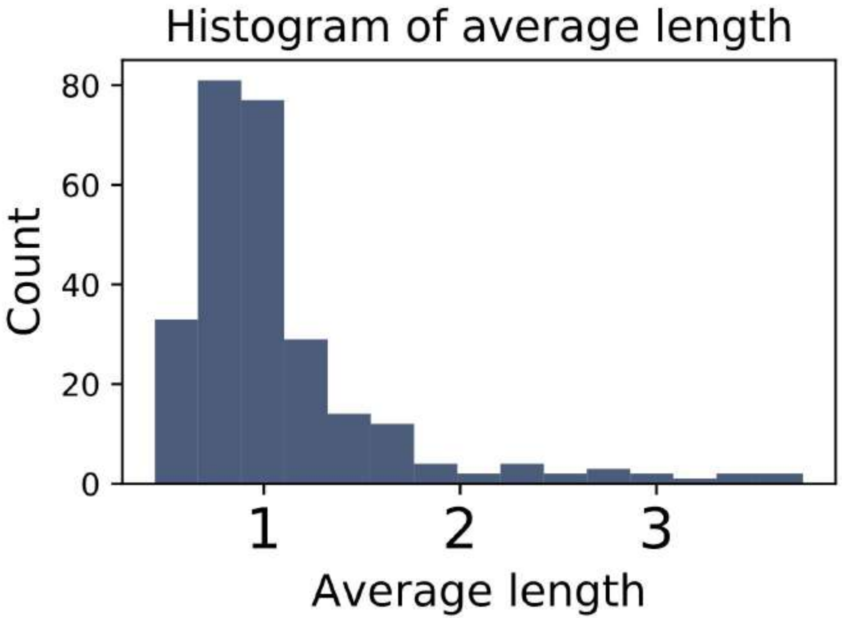
If the errors $\{\Delta_{t+k}, \Delta_t, \Delta_{t-1}, \Delta_{t-2}, \Delta_{t-3}, \Delta_{t-4}\}$ are **exchangeable**, then

$$\mathbb{P}\left(y_{t+k} \in \widehat{\mathrm{PI}}_{t+k}\right) = \mathbb{P}\left(\Delta_{t+k} < \Delta_{\max}\right) = 1 - \mathbb{P}\left(\Delta_{t+k} = \Delta_{\max}\right) = \frac{5}{6} \approx 0.83.$$

[1] G. Shafer and V. Vovk. A tutorial on conformal prediction. Journal of Machine Learning Research, 9(Mar):371–421, 2008.
[2] V. Vovk, A. Gammerman, and G. Shafer. Algorithmic learning in a random world. Springer Science & Business Media, 2005

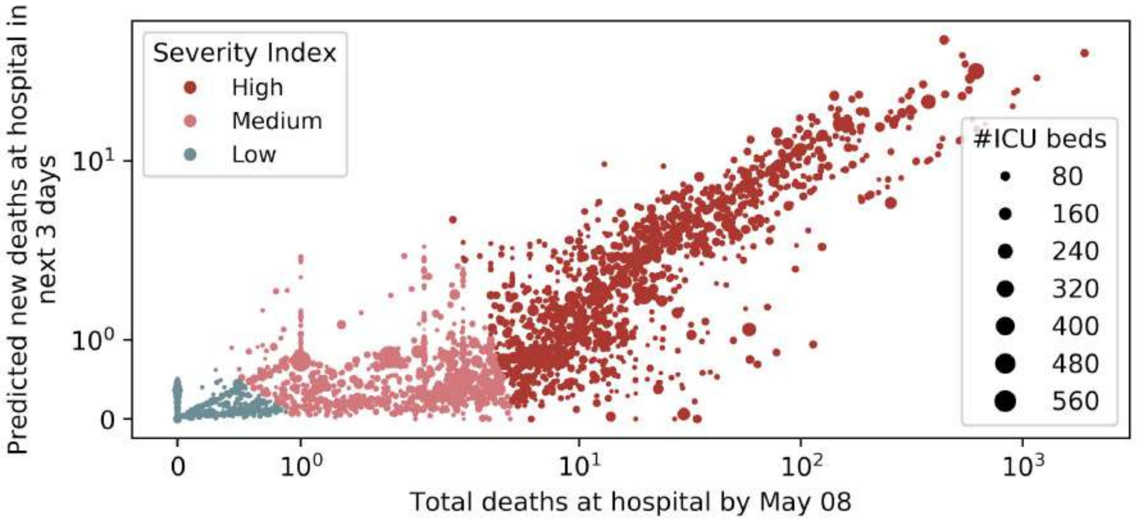
Empirical performance of MEPI





Evaluation period: March 28--April 27. Only include days since the county has 10 deaths.
Having a normalized length of 0.8 means the PI is roughly (0.6 $\widehat{y}_{t+k}$, 1.4 $\widehat{y}_{t+k}$).

Severity Index



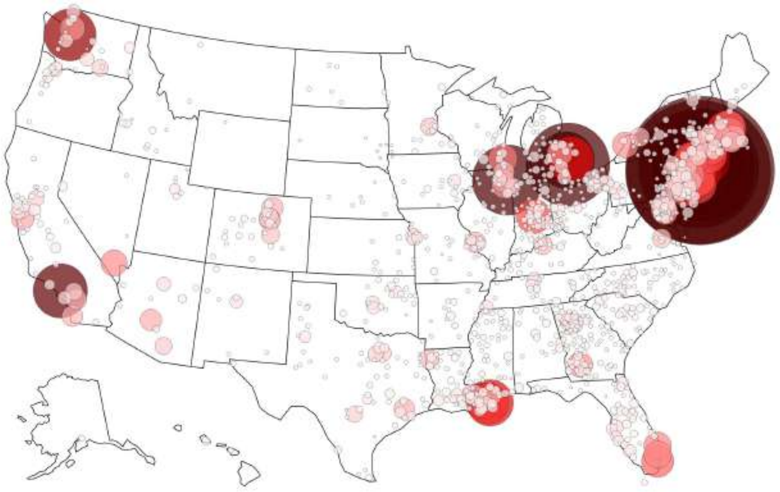
A score* for each hospital based on:

1. Predicted cumulative deaths
2. Predicted daily deaths

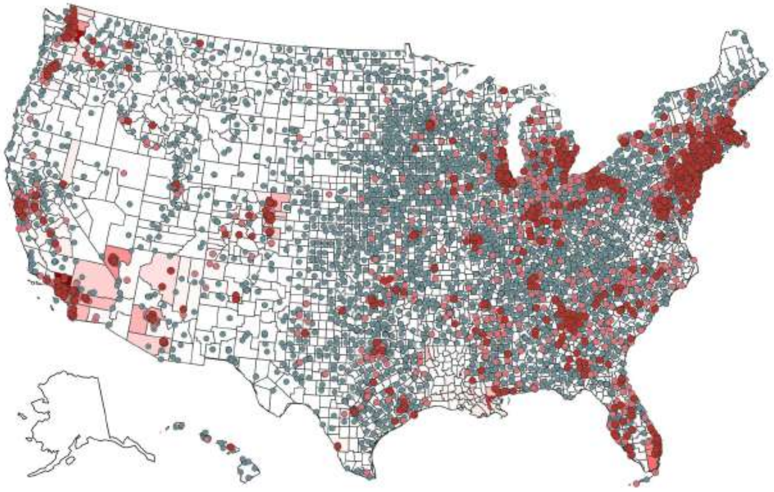
* county level predicted deaths are distributed to hospitals proportional to #employees

(Interactive) map visualizations

County-level predicted cumulative # of deaths*



Hospital severity index*



*Maps for 04/15

Collaborating with the Center for Spatial Data Science (**CSDS**) at **University of Chicago** to add our predictions and severity index to the U.S. COVID-19 Atlas.

Paper available at tinyurl.com/yugroup-covid19

Curating a COVID-19 data repository and forecasting county-level death counts in the United States

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

[1]Department of Statistics, University of California, Berkeley
[2]Department of EECS, University of California, Berkeley
[3]Department of Pharmaceutical Chemistry, University of California, San Francisco
[4]Chan Zuckerberg Biohub, San Francisco
[5]Center for Computational Biology, University of California, Berkeley
[6]Division of Biostatistics, University of California, Berkeley

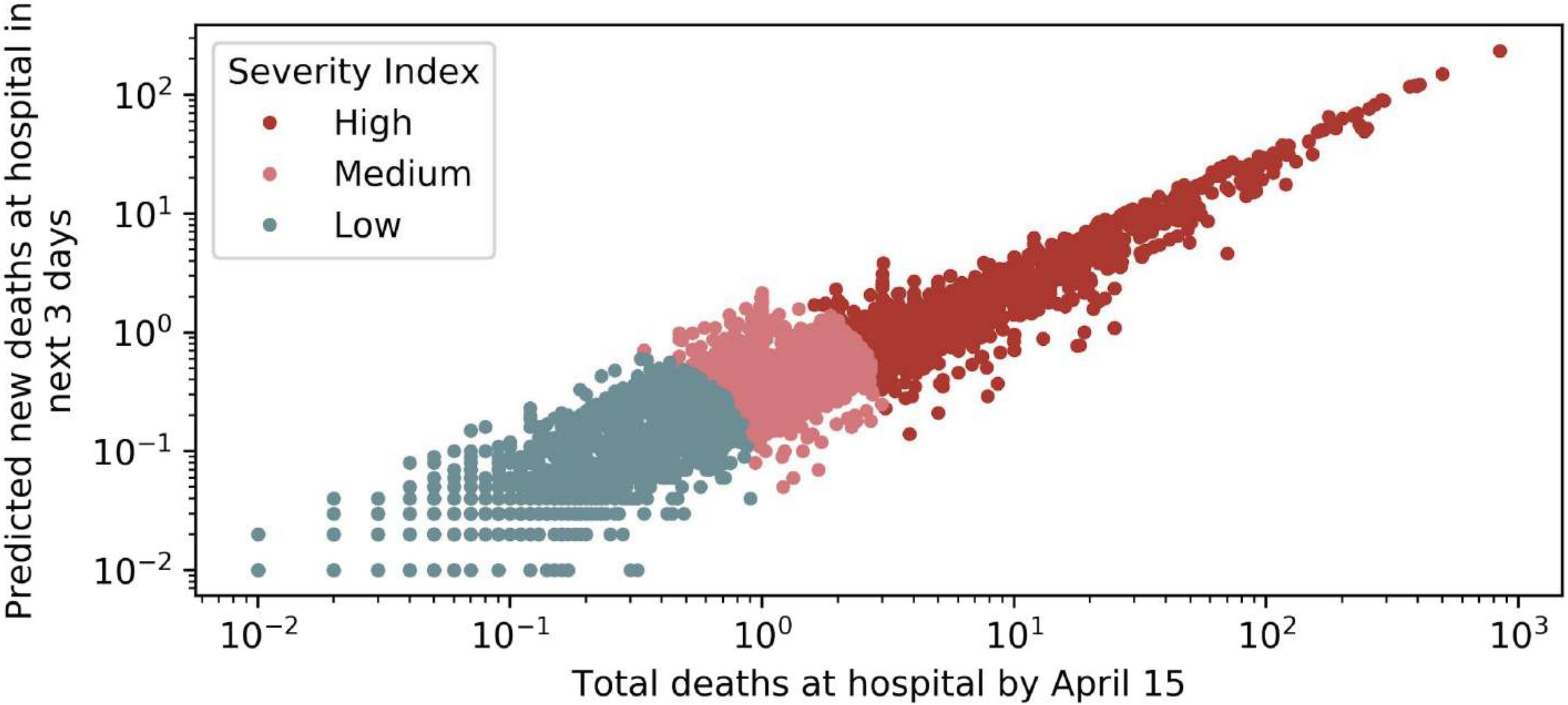
April 29, 2020

†Authors ordered alphabetically. All authors contributed significantly to this work.
*Corresponding authors
This project was initiated on March 21, 2020, with the goal of helping aid the allocation of supplies to different hospitals in the U.S., in partnership with the non-profit Response4Life.

Thank you!

Misc

Assign severity index to hospital based on predicted cumulative deaths



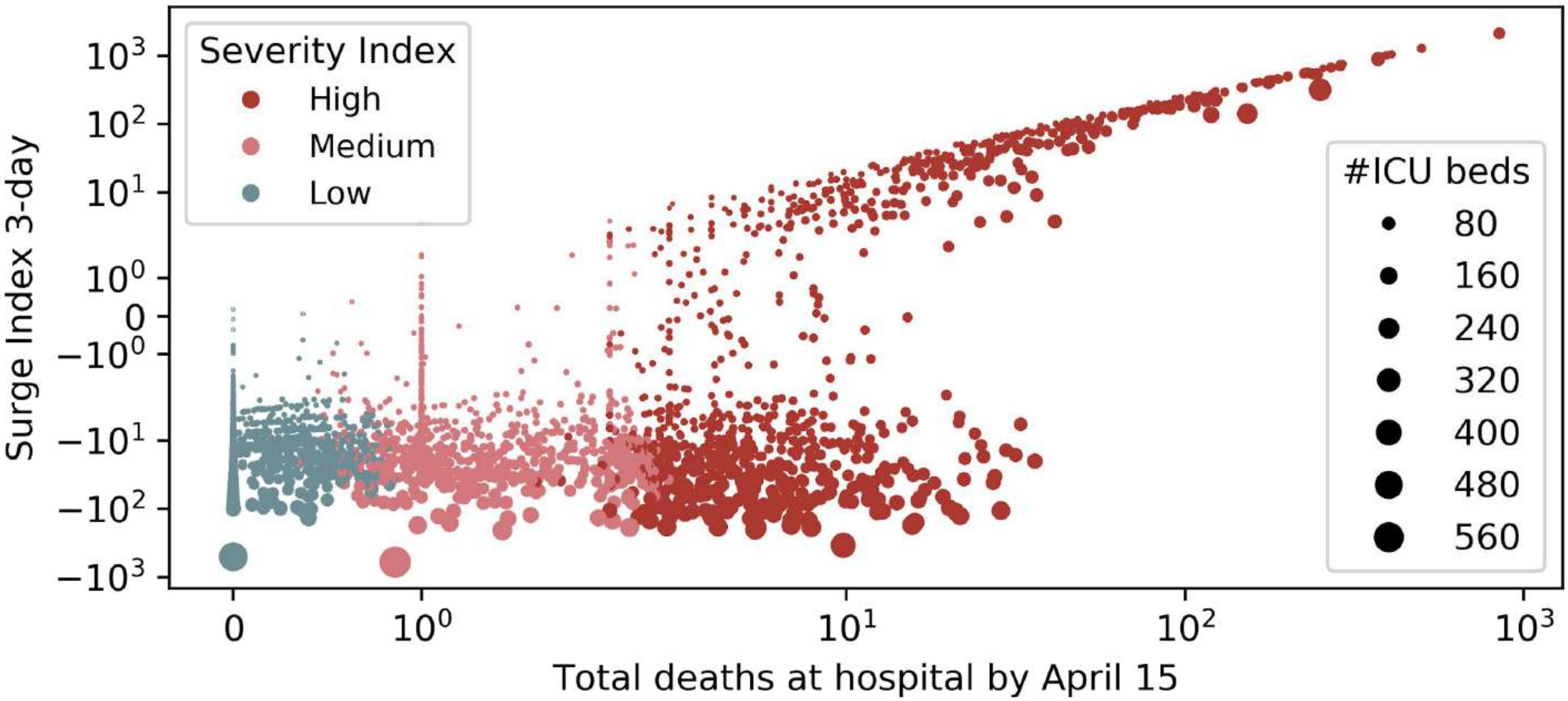
Surge Index

A score for each hospital based on:

(Estimated # ICU beds needed*) - (# ICU beds available)

*2x predicted cumulative number of deaths

Assign surge index based on #ICU beds



Volunteer Team: Local News and Emerging Hotspots

Hospital	Severity	Deaths as of April 10
Beaumont Health (Ohio County) Michigan	3	328

10-12 volunteers find local news and gather hard to find on-the-ground data

Compare collected data against predicted severity

Other works -- at state or country level

Curve fitting epi. Modeling (e.g. IHME -- dominant in the US)

Compartment epi. modeling (e.g. ICL -- dominant in UK and Europe)

Both have parameters that are tuned based on data mostly from other countries

No comparisons yet on prediction with US data …

Current & Future Directions

Continue to update predictors

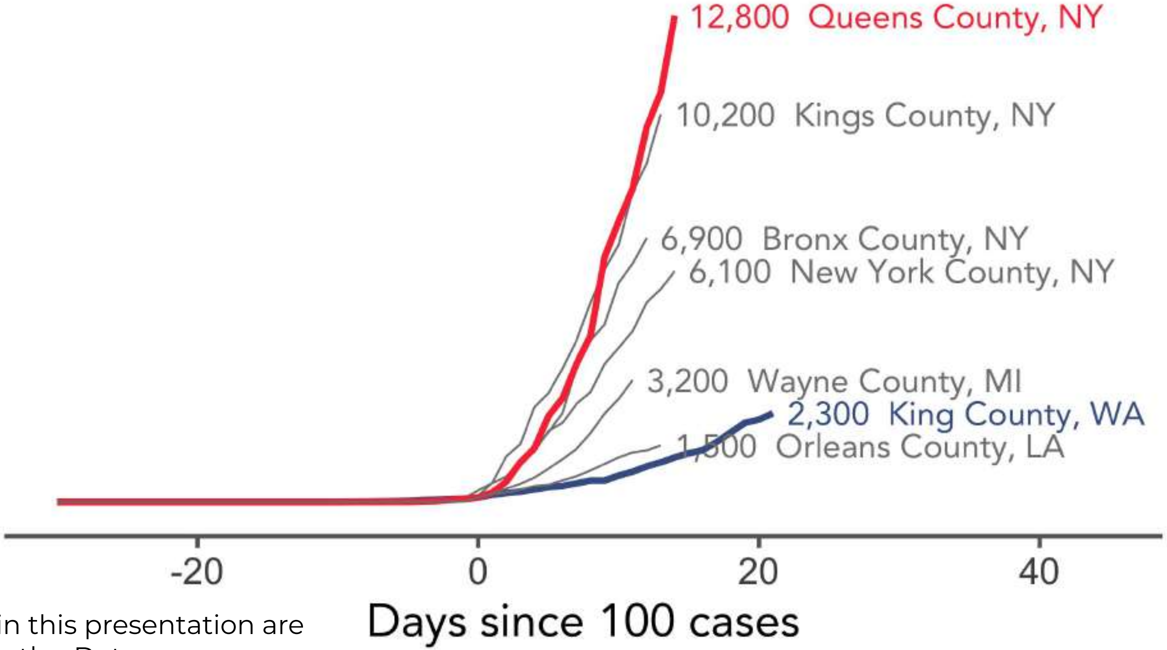
Look at long-term trajectories

Incorporate epidemiology models

Concentrate on rural areas

Current situation: Exponential growth of COVID-19

Cumulative number of *cases* by county



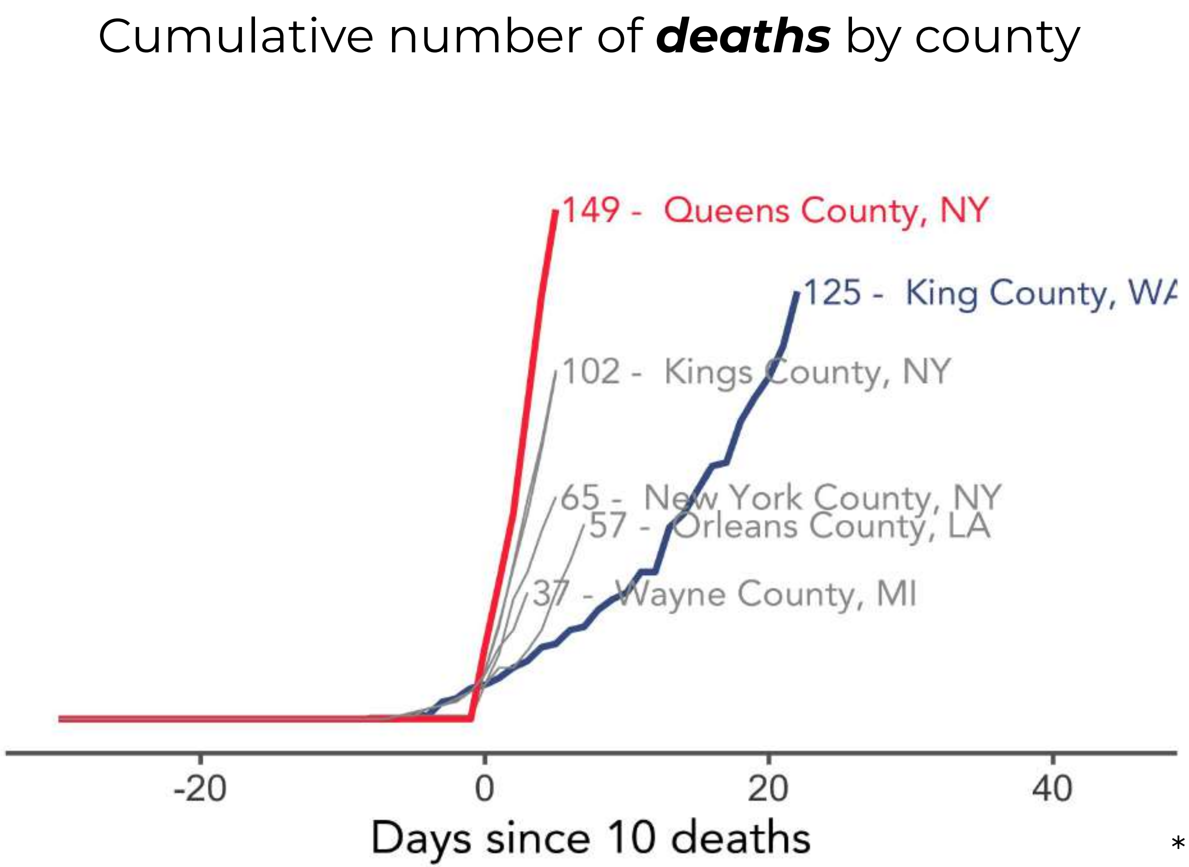
Focusing on 6 of the worst-affected counties

*"Cases" and "deaths" in this presentation are recorded cases and deaths. Data source: https://usafacts.org.

*Based on 3/30 data

Current situation: Exponential growth of COVID-19

Focusing on 6 of the worst-affected counties

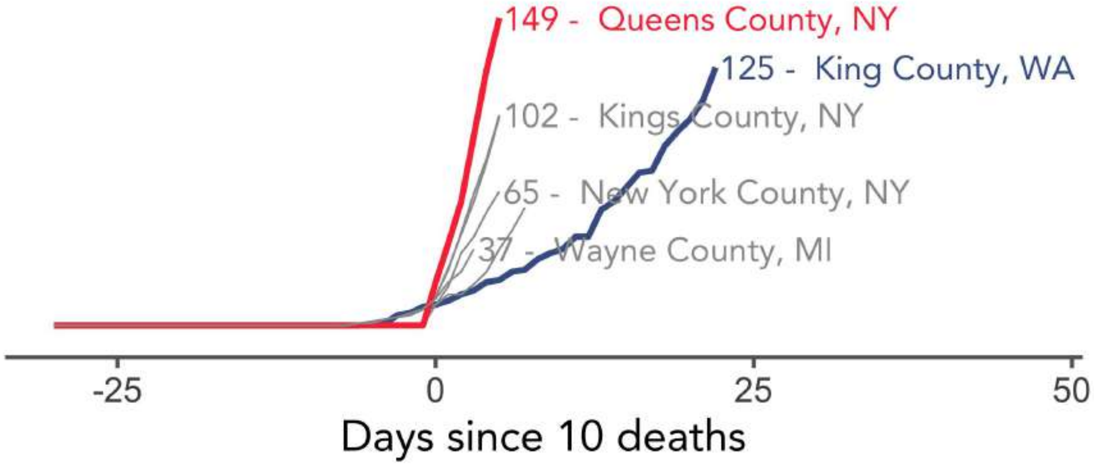


*Based on 3/27 data

Goal: **Predict** COVID-19 at the county level

Our goal: <u>**predict**</u> COVID-19 at the county level

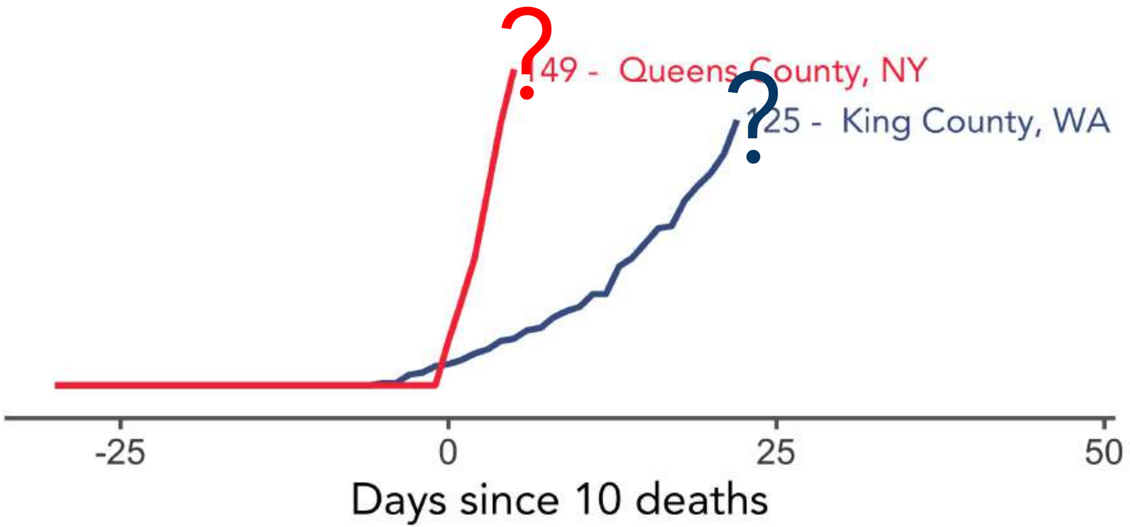
Cumulative number of ***deaths*** by county



*Based on 3/27 data

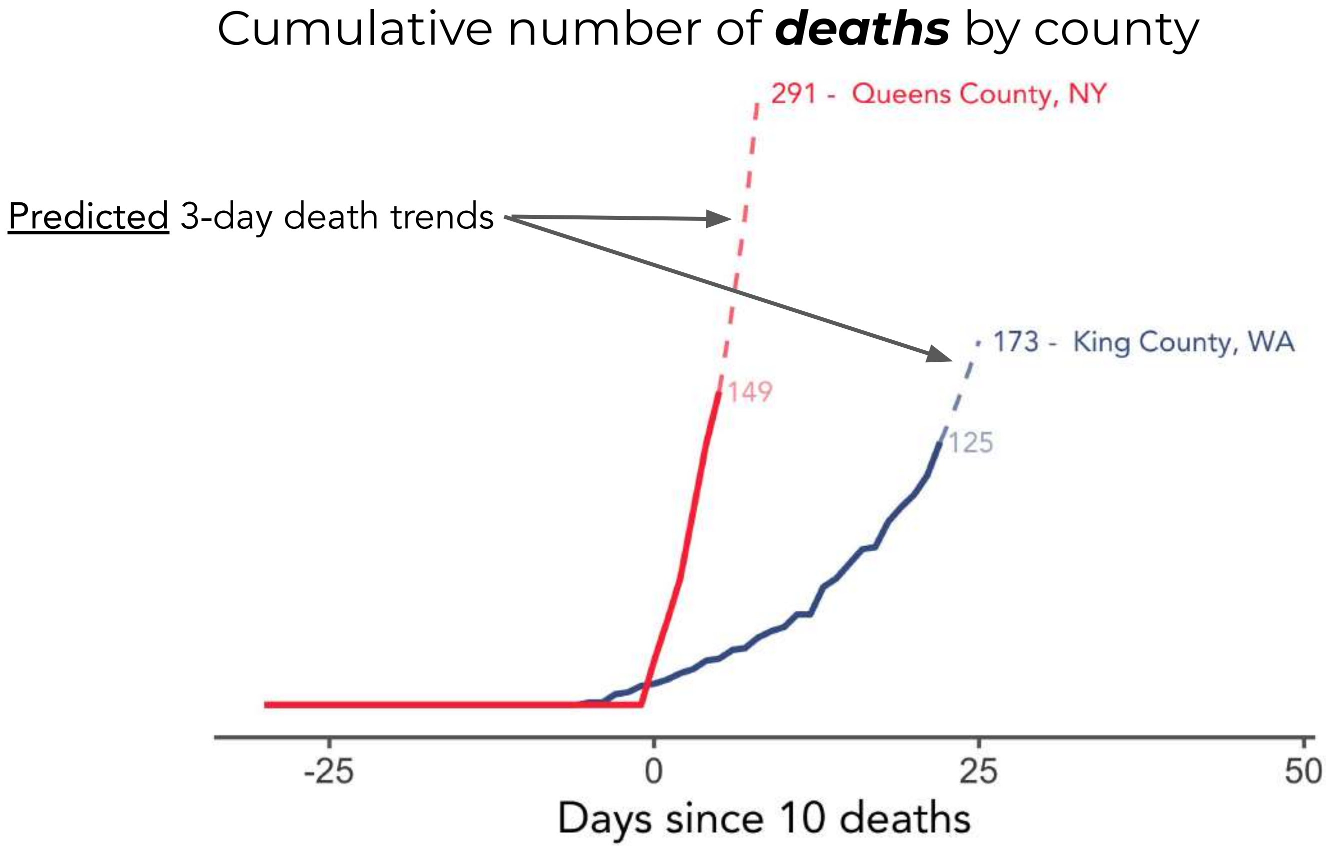
Goal: **Predict** COVID-19 at the county level

Cumulative number of ***deaths*** by county



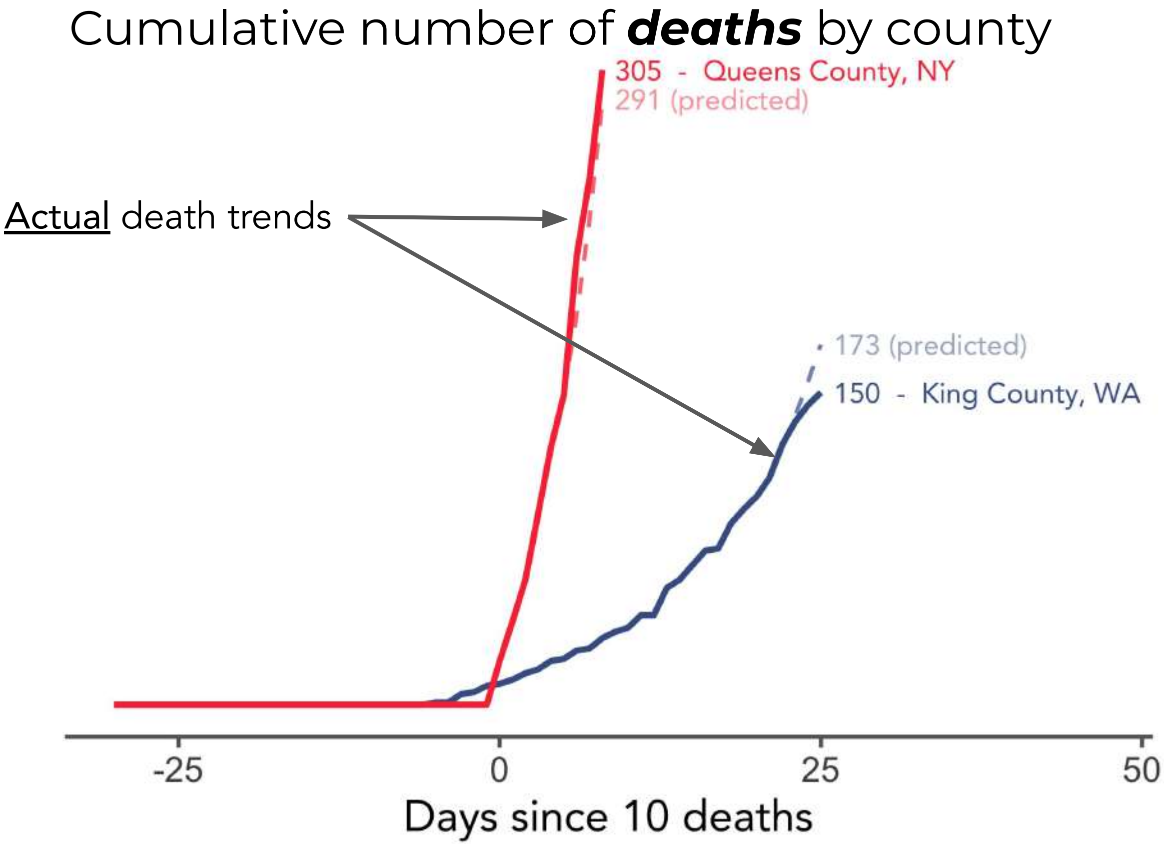
*Based on 3/27 data

Goal: **Predict** COVID-19 at the county level



*Based on 3/27 data

Goal: **Predict** COVID-19 at the county level



*Based on 3/30 data