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

NSF PREVENT Symposium
Feb. 22, 2021

PI: Bin Yu

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

B. Park

C. Singh (Student Lead)

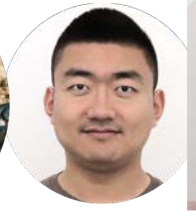
Y. Tan

T. Tang

Y. Wang

A.Agarwal

M. Shen



C. Zhang

D. Wang.

P. Norvig

Google collaborators

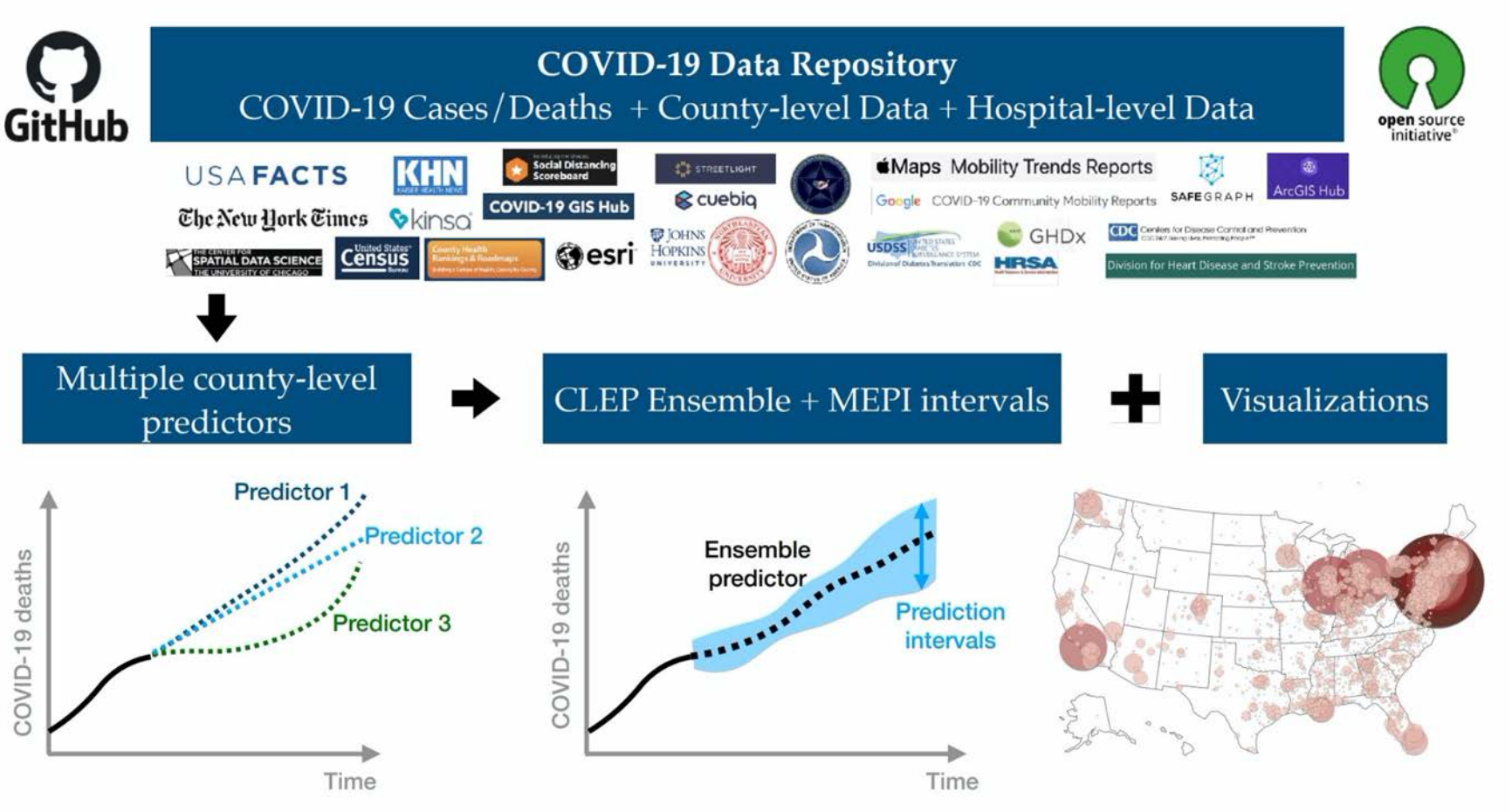
Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...

5000 PPEs for Temple Univ Hospital on May 8, 2020

Don Landwirth.

Rick Brennan

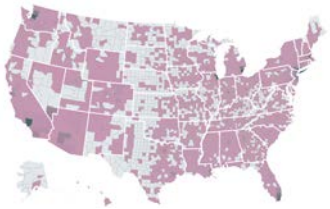
Overview: Current Data Repository & Prediction Pipeline (Open Source)



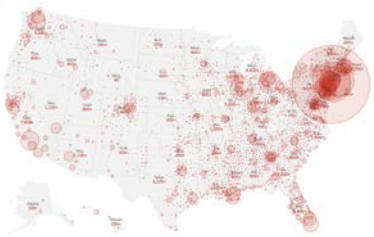
Data curation: scraped from a variety of sources

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

JOHNS HOPKINS UNIVERSITY

Maps Mobility Trends Reports

Google COVID-19 Community Mobility Reports

Hospital-level Data

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

HRSA
Health Resources & Services Administration

ArcGIS Hub

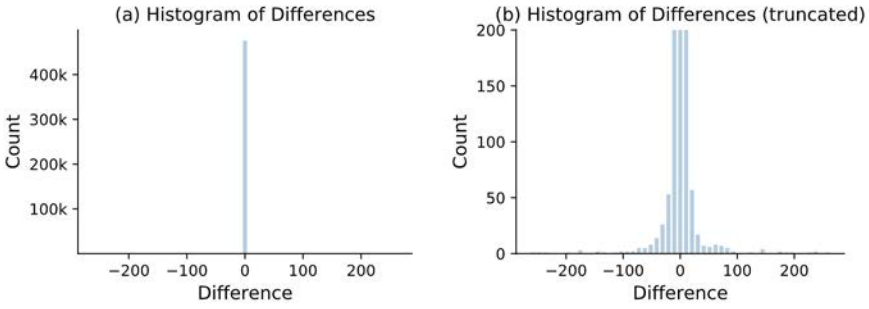
HOMELAND INFRASTRUCTURE FOUNDATION-LEVEL DATA SUB COMMITTEE

Samuel Scarpino

NORTHEASTERN UNIVERSITY

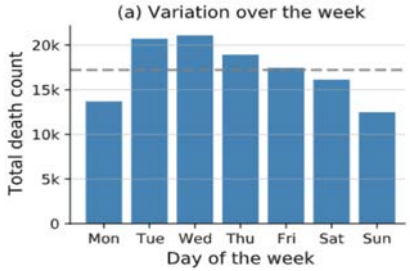
Data quality issues about death counts

1. Undercount problems
2. USAFacts and NYT data come from the same sources, but do not always agree
3. Weekdays are different from weekends
4. Historical data revisions

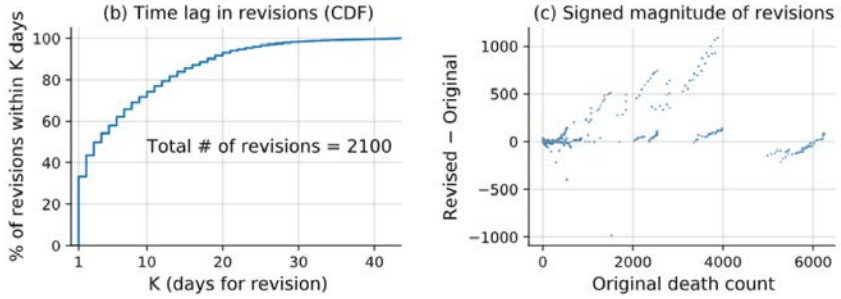


(2)

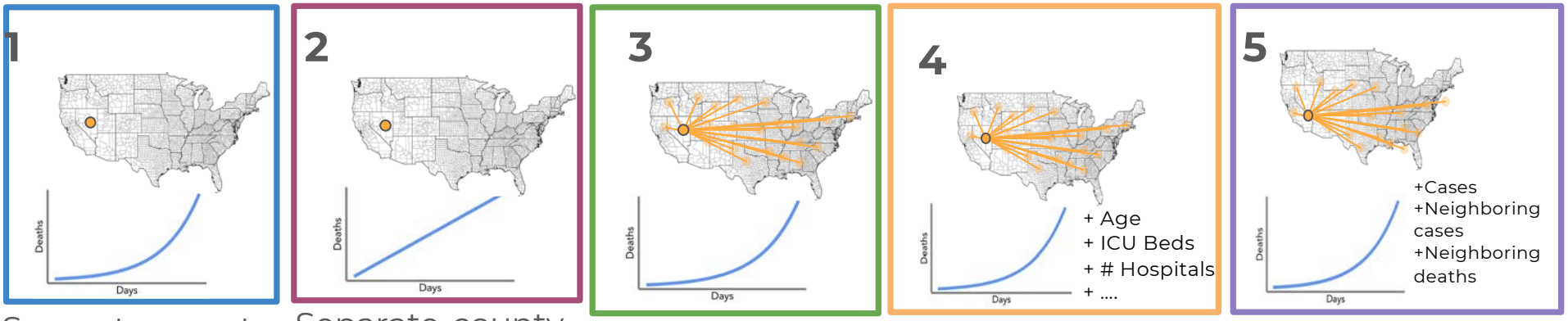
(3)



(4)



Development of many transparent predictors



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]

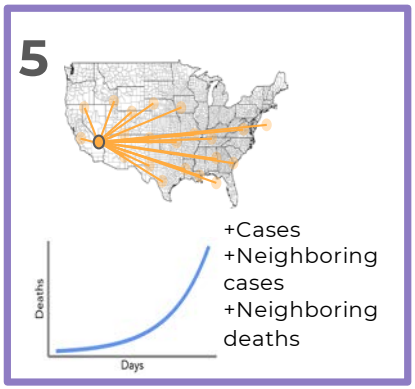
[1]. Schuller-Yu-Huang-Edler "Perceptual audio coding using adaptive pre-and post-filters and lossless compression." *IEEE Transactions on Speech and Audio Processing* 10.6 (2002): 379-390.

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

A combination of two predictors performs well

Separate-county linear predictor

+



Expanded Shared-county exponential

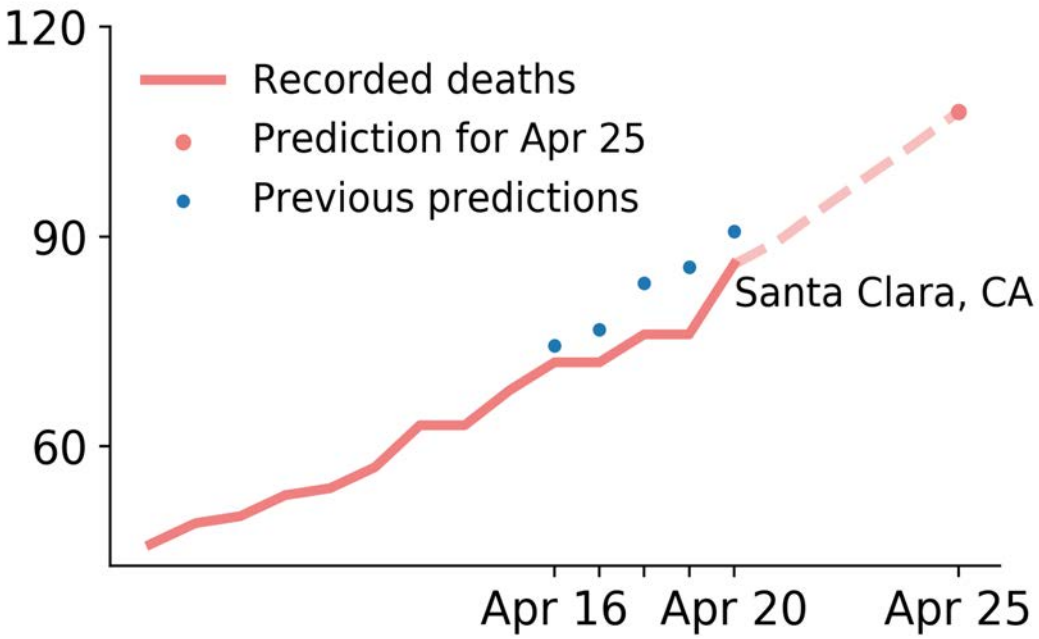
k=7 for 7-day prediction

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

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]



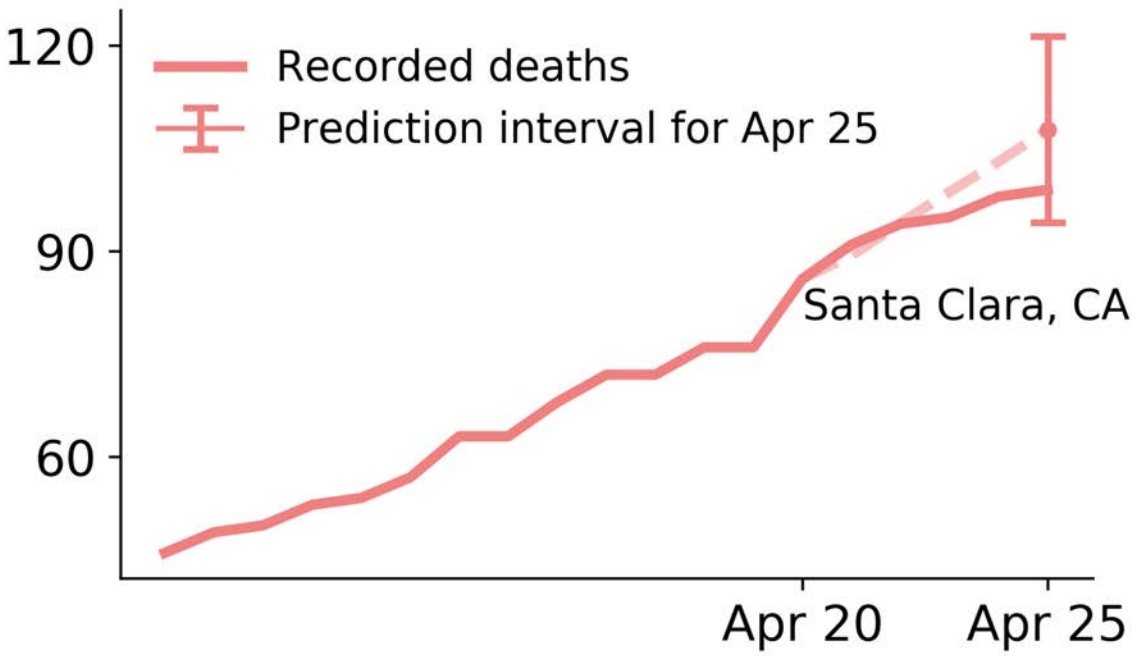
Previous 5-day-ahead rel. prediction errors (%)

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: Max. Error Prediction Interval (**MEPI**)



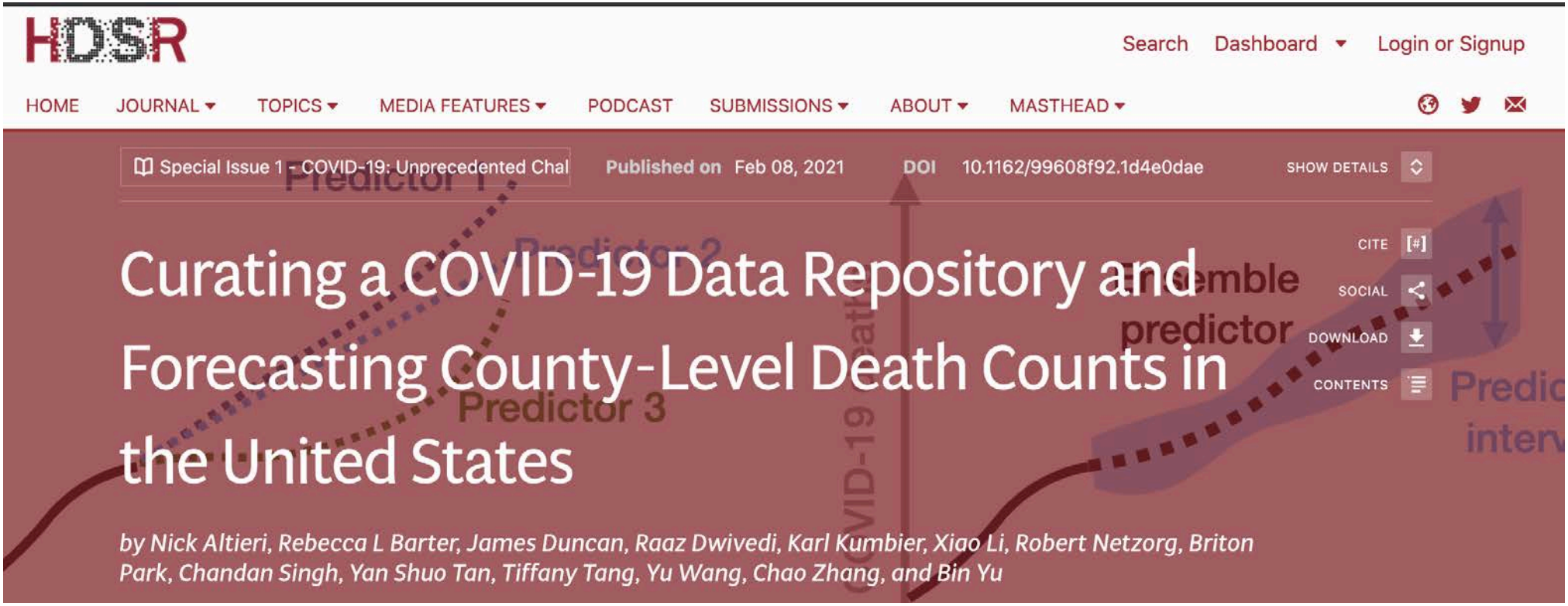
Predicted range of error

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

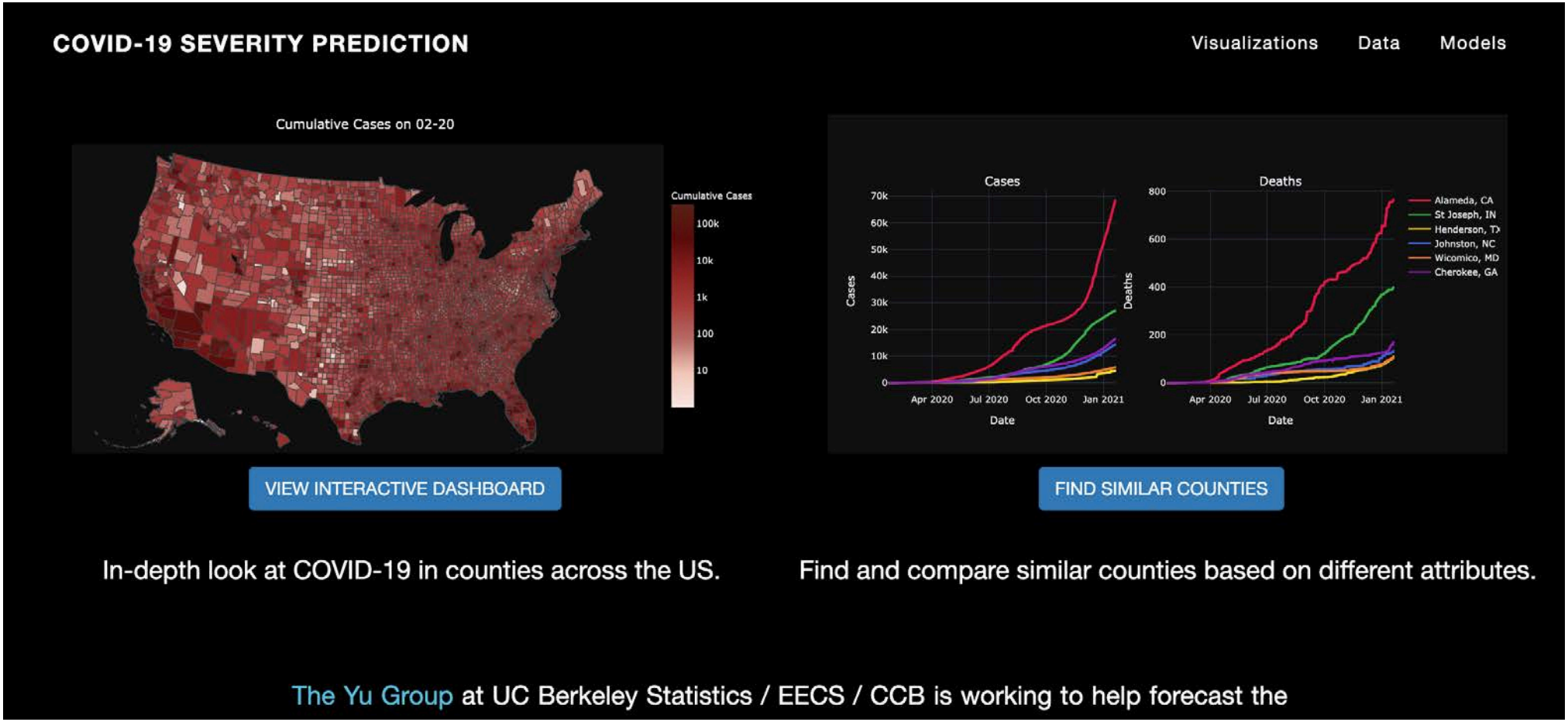
Actual error:

Apr 25 8.8%

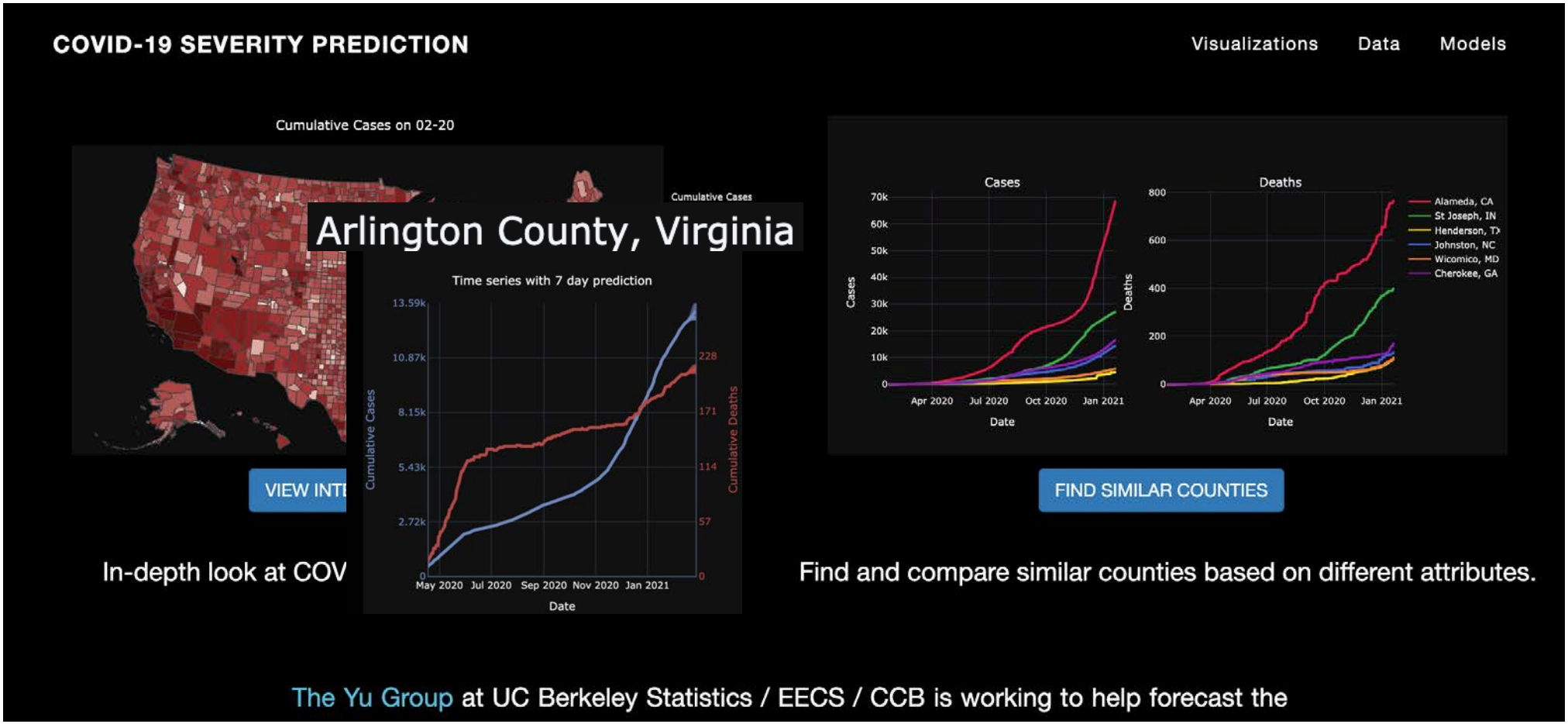
https://hdsr.mitpress.mit.edu/pub/p6isyf0g/release/1 **Harvard Data Science Review (HDSR)**



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



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.

AI: human-machine collaboration

Image credit: trademed.com.

Summary and Current Directions

covidseverity.com

- Data repository and open-source code on github
- Our CLEP/MEPI: simple, transparent, interpretable, and fast, and comparable short-term prediction performance as agent-based models
- Hospitalization prediction (on-going)
- Causal investigation (on-going)

Challenges and Opportunities

A nimble, national and international, surveillance and intervention network system

Complete with manufacture, supply chain logistics planning and stakeholder coordination

Collecting, cleaning/curating data at multi-scale with **data quality control**

Powered by a responsible, trustworthy, and reproducible AI system with humans in the loop

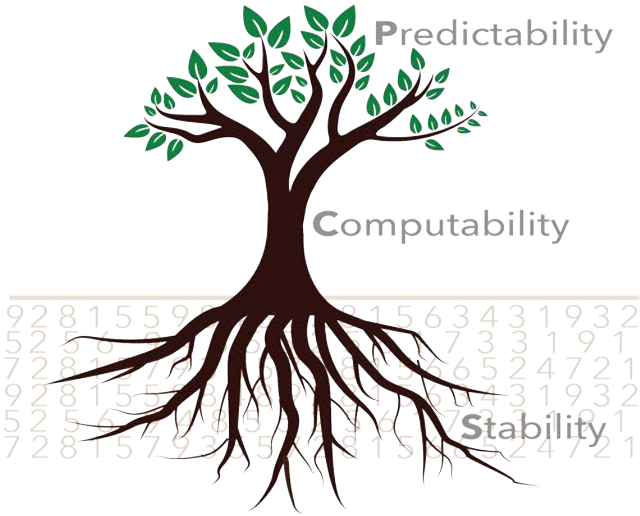
Challenges and Opportunities

Trans-disciplinary framework

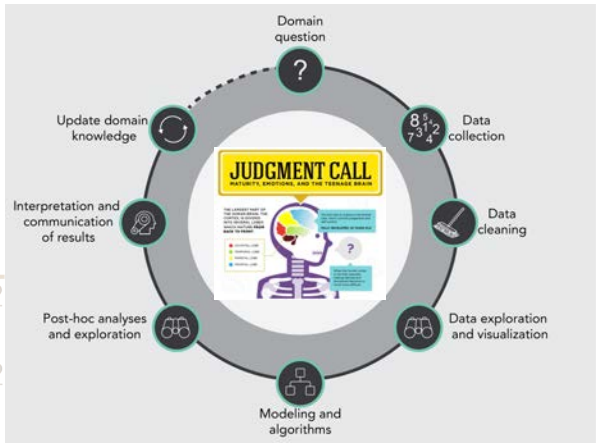
- coherent and unified technical and accessible terminology
- an integrated and distributed computing platform (e.g. SPARK or DataBricks)
- a combination of diverse data/science driven prediction approaches (agent-based, ML/Stats/SP) and with uncertainty assessments
- validating and reality checking policy decisions with multi-objectives (health. economy, fairness)
- **agile** to adapt to fast dynamics due to climate changes and unpredictable human events

Our PCS framework
for responsible data analysis and decision making

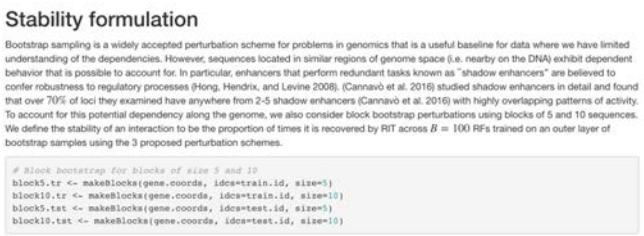
Veridical Data Science (Yu and Kumbier, PNAS, 2020)



PCS



Data science life cycle



Stability formulation

Bootstrap sampling is a widely accepted perturbation scheme for problems in genomics that is a useful baseline for data where we have limited understanding of the dependencies. However, sequences located in similar regions of genome space (i.e. nearby on the DNA) exhibit dependent behavior that is possible to account for. In particular, enhancers that perform redundant tasks known as "shadow enhancers" are believed to confer robustness to regulatory processes (Hong, Hendrix, and Levine 2006). (Cannavò et al. 2016) studied shadow enhancers in detail and found that over 70% of loci they examined have anywhere from 2-5 shadow enhancers (Cannavò et al. 2016) with highly overlapping patterns of activity. To account for this potential dependency along the genome, we also consider block bootstrap perturbations using blocks of 5 and 10 sequences. We define the stability of an interaction to be the proportion of times it is recovered by RIT across $B = 100$ RFs trained on an outer layer of bootstrap samples using the 3 proposed perturbation schemes.

```
# Block bootstrap for blocks of size 5 and 10
block5.tr <- makeBlocks(gene.coords, idcs=train.id, size=5)
block10.tr <- makeBlocks(gene.coords, idcs=train.id, size=10)
block5.tst <- makeBlocks(gene.coords, idcs=test.id, size=5)
block10.tst <- makeBlocks(gene.coords, idcs=test.id, size=10)
```

PCS documentation

Human is at the center

- Transdisciplinary team: **collaborative, not winner-takes-all, culture**
- Visionary and fair Incentives and rewards (grant, authorship, promotion, award)

NSF – not enough to support transformative research, but also support **verification and replication** for trustworthy and reliable research

- Human-machine collaboration in AI system
- Transdisciplinary education, including communication and interpersonal skills

Thank you!