In partnership with

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

Yu Group

UC Berkeley Statistics, EECS, CCB

Presenter: Tiffany Tang

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

Website: covidseverity.com

Our Team

PI: Bin Yu

N. Altieri

R. Barter

J. Duncan

R. Dwivedi

K. Kumbier

X. Li

R. Netzorg

D. Wang

B. Park

C. Singh (Student Lead)

Y. Tan

T. Tang (Data Team)

Y. Wang (Data Team)

A.Agarwal

M. Shen

C. Zhang

P. Norvig

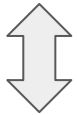
Many others at UC Berkeley, UCSF, Stanford, Northeastern, Univ. of Chicago, UW-Madison, ...

Goal: Help Aid Resource Allocation

RESPONSE 4LIFE

Want to predict...

~~hospital PPE/supply need~~



~~number of COVID-19 hospitalizations~~

number of COVID-19 deaths at the county-level

Overview: Current Data Repository & Prediction Pipeline

GitHub

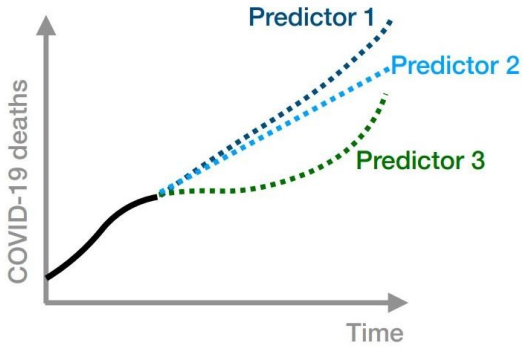
COVID-19 Data Repository

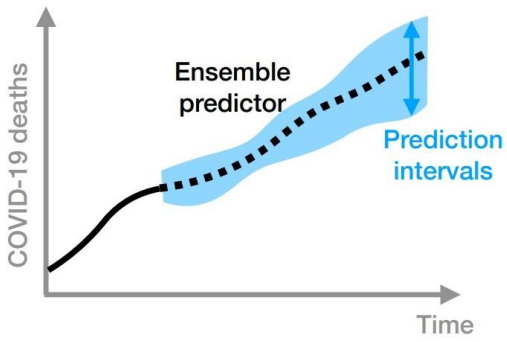
COVID-19 Cases/Deaths + County-level Data + Hospital-level Data

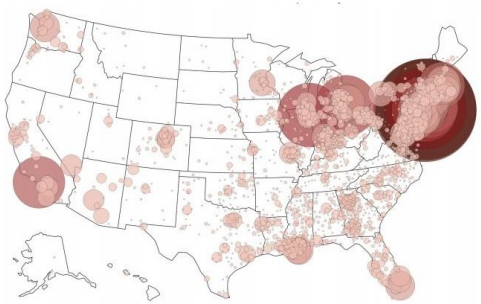
open source initiative

USAFACTS · KHN · Social Distancing Scoreboard · COVID-19 GIS Hub · STREETLIGHT · cuebiq · Maps Mobility Trends Reports · Google COVID-19 Community Mobility Reports · SAFEGRAPH · ArcGIS Hub · The New York Times · kinsa · JOHNS HOPKINS UNIVERSITY · USDSS · GHDx · HRSA · CDC Centers for Disease Control and Prevention · Division for Heart Disease and Stroke Prevention · THE CENTER FOR SPATIAL DATA SCIENCE THE UNIVERSITY OF CHICAGO · United States Census Bureau · County Health Rankings & Roadmaps · esri

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







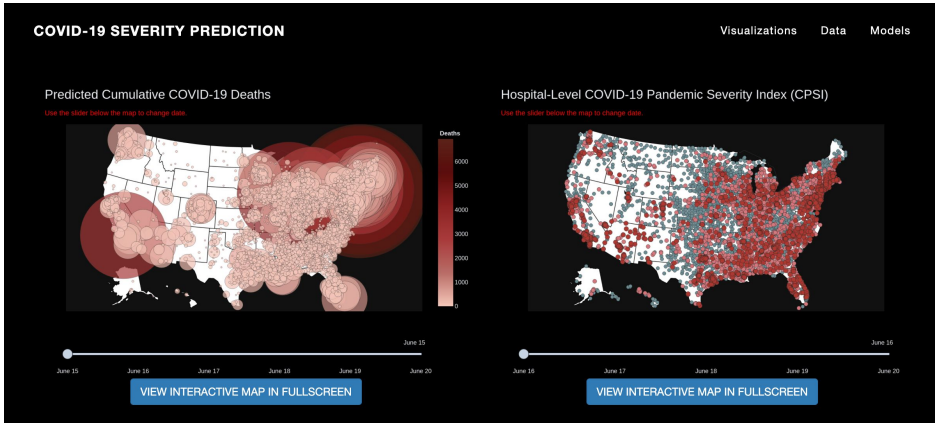
Altieri, Nick, et al. "Curating a COVID-19 data repository and forecasting county-level death counts in the United States." *arXiv preprint arXiv:2005.07882* (2020).

Impacts

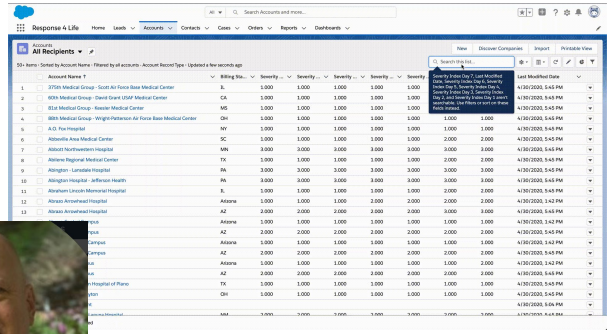
Sending PPE through R4L

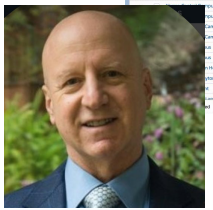
- >5000 face shields to Santa Clara + Temple University Med Center in Philadelphia in collaboration with GetUsPPE, AeroBridge, Maker Nexus, Synergy Mill maker space, R4L
- +65k to 25 recipients in 15

Visualizations (covidseverity.com)



Salesforce system





Don Landwirth, R4L

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

Curating a COVID-19 Data Repository

Outline

- Our data curation pipeline
- Overview of the data
- A closer look at the COVID-19 daily cases/deaths data
- Navigating the data repo on GitHub

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 20+ data sources**
- **Monitor data changes 24/7 powered by AWS**
- **Handling missing and erratic entries**
- **Automated python script**
- **Compare data across different sources**
- **Search for emerging data sources**

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

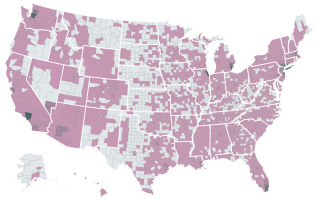
Data and code available: https://github.com/Yu-Group/covid19-severity-prediction

★ Being used by multiple research groups across the country

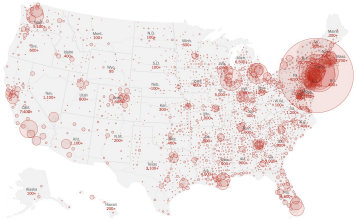
Our Data Repository: scraped from **20+ sources**

COVID-19 Cases/Deaths

USAFACTS



The New York Times



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

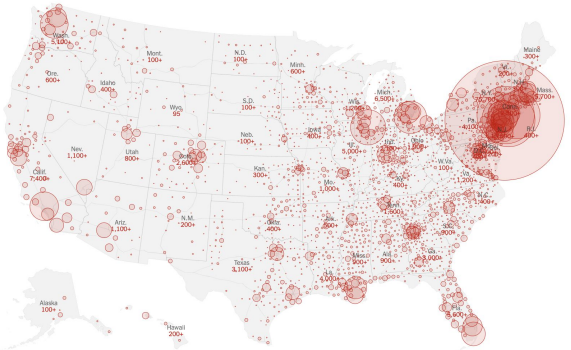
ArcGIS Hub

HOSPITAL INFRASTRUCTURE FOUNDATION-LEVEL DATA SUBCOMMITTEE

Samuel Scarpino

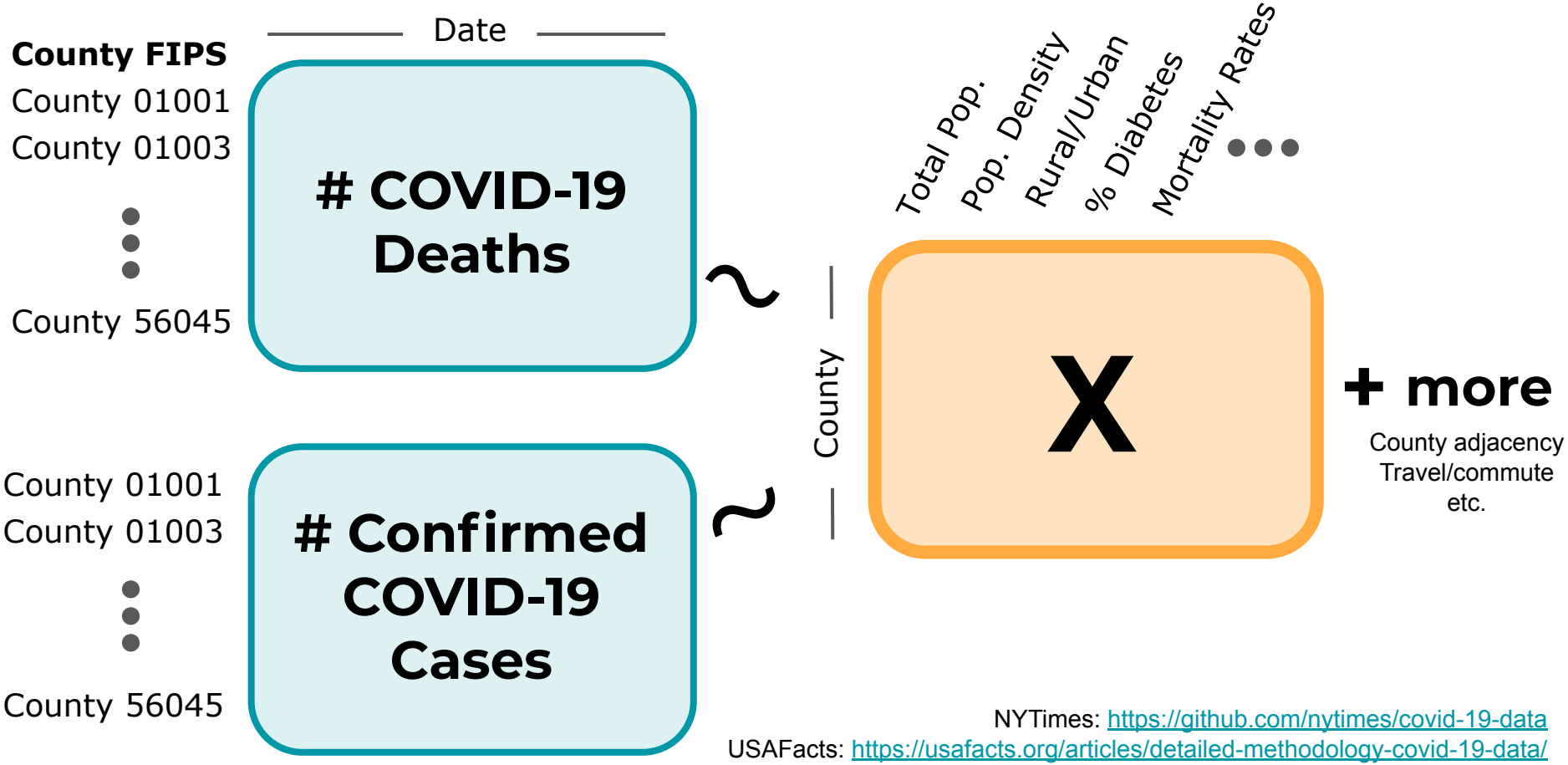
NORTHEASTERN UNIVERSITY

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

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

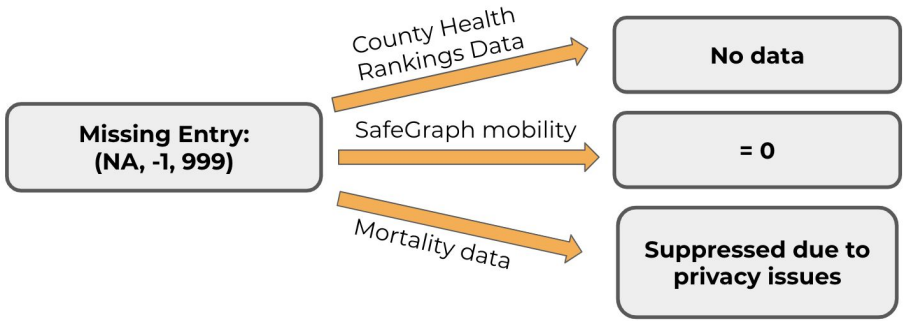


NYTimes: https://github.com/nytimes/covid-19-data
USAFacts: https://usafacts.org/articles/detailed-methodology-covid-19-data/

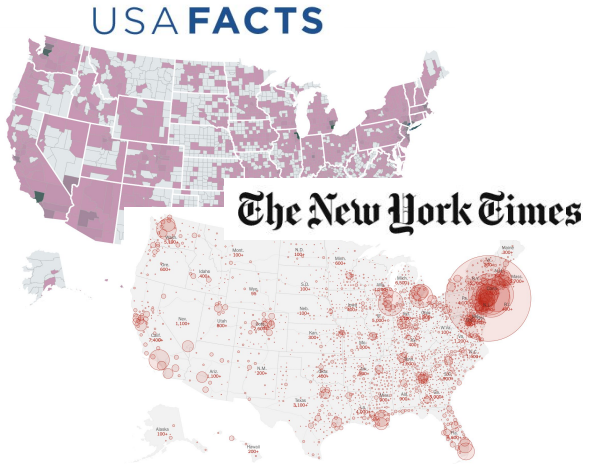
A closer look at the USAFacts and NYTimes COVID-19 cases/deaths data...

Data Cleaning!!

- Duplicated counties
- Some cases/deaths cannot be allocated to a county
- Cumulative deaths counts sometimes decrease
- Aside: data cleaning beyond USAFacts/NYTimes
  - Some counties changed their countyFIPS code
  - Missing data entries encoded as NAs, -1, 999, and more...
    - Meaning can depend on the data set



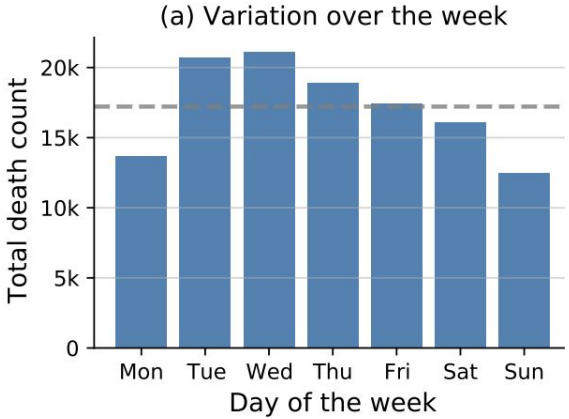
Which to choose: NYTimes vs USAFacts?

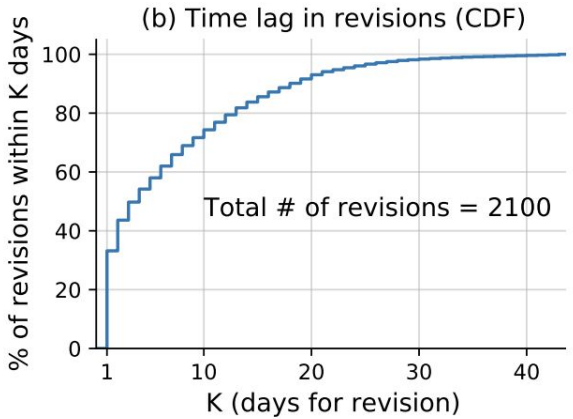


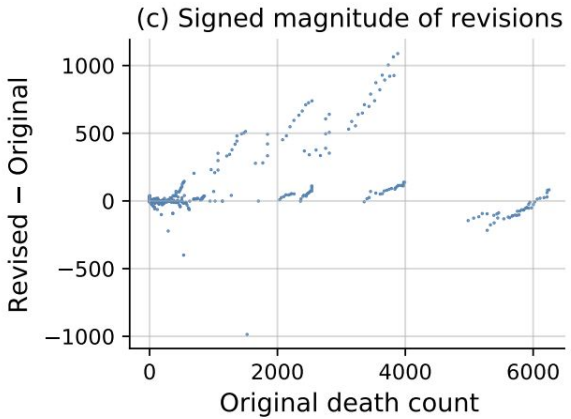
- Out of all possible (county, day) combinations, **98.4%** of the COVID-19 case and death counts in USAFacts and NYT datasets are identical
- Possible reasons for the slight discrepancies
  - Different time to update data
  - How to deal with probable deaths?
  - Different data processing pipelines
- NYTimes aggregates cases/deaths over the NYC counties while USAFacts does not
- Multiple data sources give us insights into the caveats of the data.

Some interesting data biases from the USAFacts data

(a) Weekly patterns: cumulative death counts per weekday
(b) Historical data revision: around 2100 revisions until June 21
(c) Magnitude of the revisions are mostly upward.







Navigating the COVID-19 Data Repository

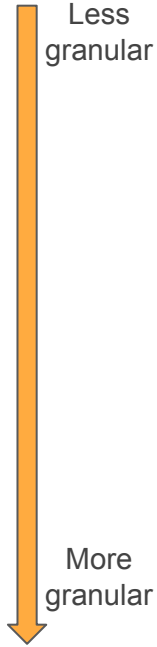
Our **data repository** can be found at the following link:

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

- Monitored and updated daily, powered by AWS
- Can easily download, clean, and merge (abridged or unabridged) data in one line of code
- Lots of documentation

A Hierarchy of Resources

- Overview of “main” features and datasets in Tables 1 and 3 in arXiv paper
  - Data Readme on GitHub with *all* data sets + brief descriptions
    - List of columns in county-level data set (contains descriptions for *most* county-level features)
      - Individual data folder readmes on GitHub at data/county_level/<raw or processed>/<source>/



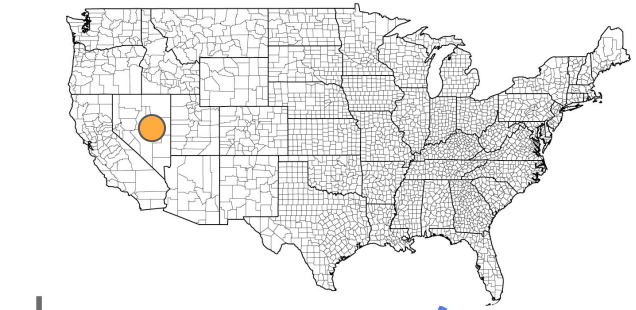
Thank you!

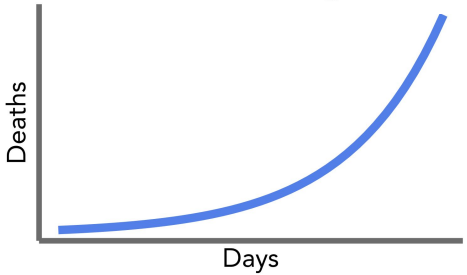
Ensemble different predictors

We combined many different prediction approaches

Ensemble predictors

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



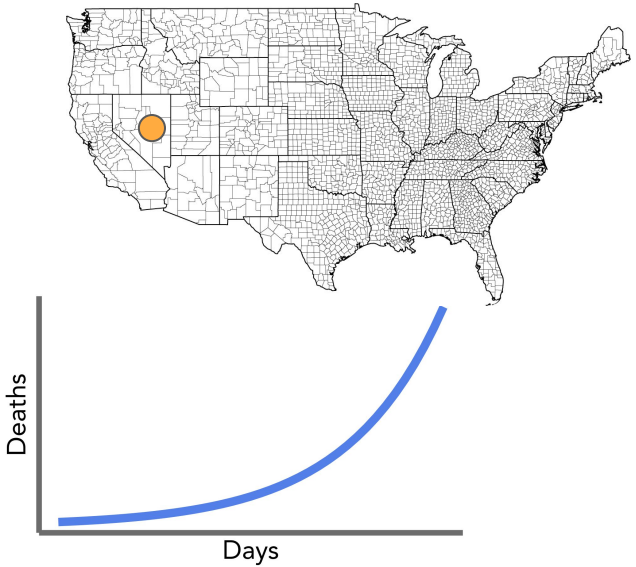


[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Ensemble predictors

We combined many different model approaches

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



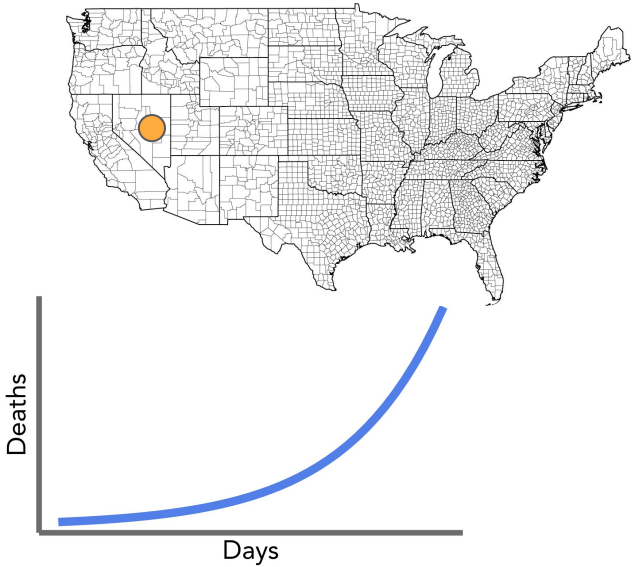
$$\mathrm{E}(\mathrm{deaths}_t \mid t) = e^{\beta_0 + \beta_1 t}$$

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

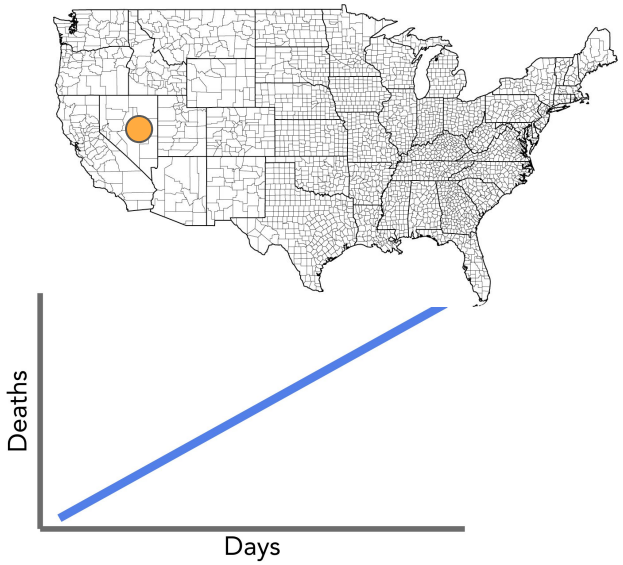
Ensemble predictors

We combined many different model approaches

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



2. Separate-county **linear** model

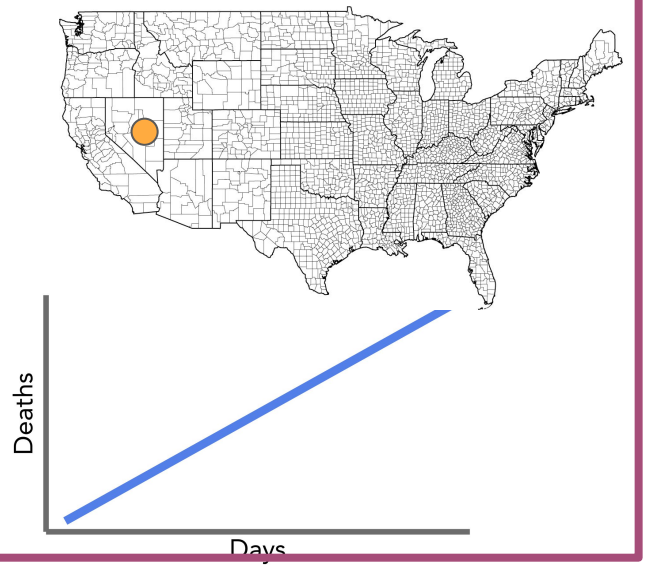


[1] Anderson, Roy M., B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

Ensemble predictors

We combined many different model approaches

2. Separate-county **linear** model



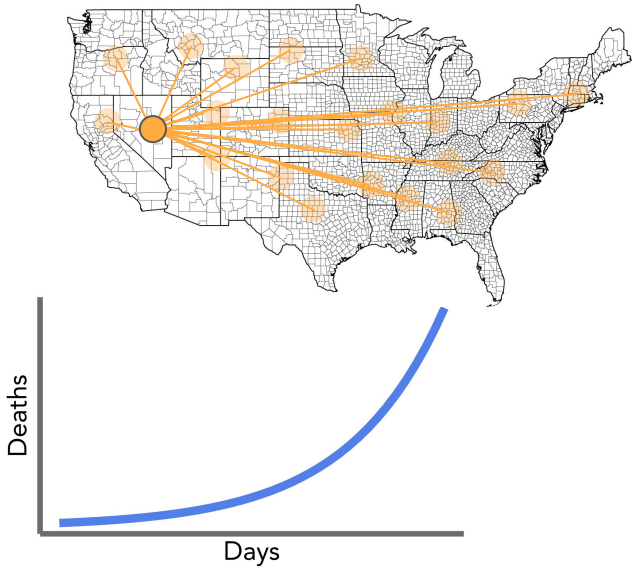
$$\mathrm{E}[\mathrm{deaths}_t | t] = \beta_0 + \beta_1 t$$

[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

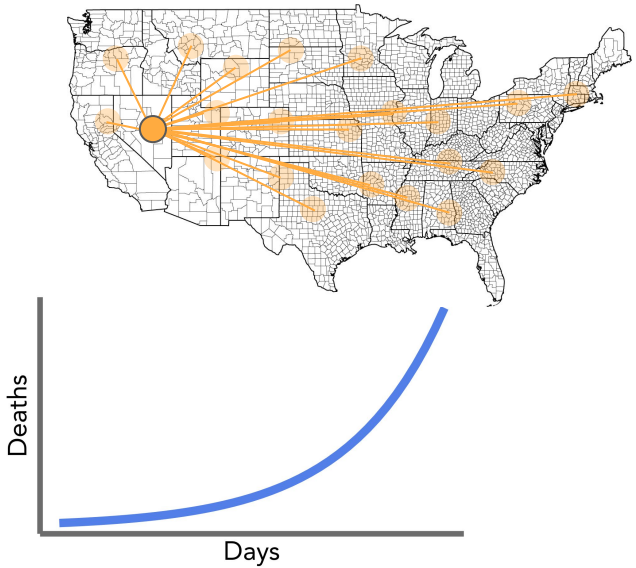
3. **Shared**-county exponential model



Ensemble predictors

We combined many different prediction approaches

3. **Shared**-county exponential model

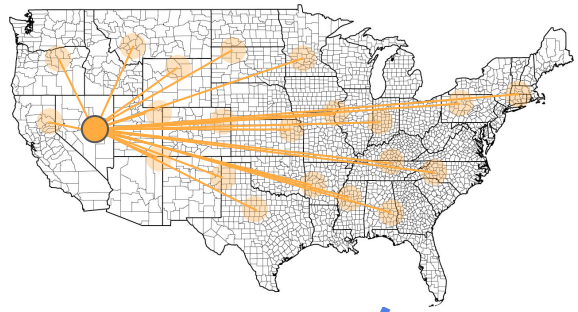


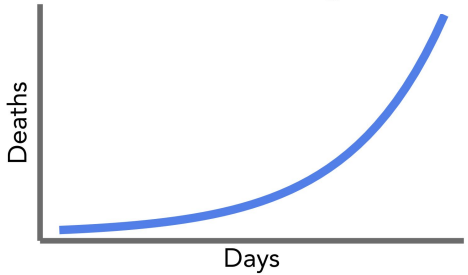
$$\mathrm{E}(\mathrm{deaths}_t \mid t) = e^{\beta_0 + \beta_1 \log(\mathrm{deaths}_{t-1} + 1)}$$

Ensemble predictors

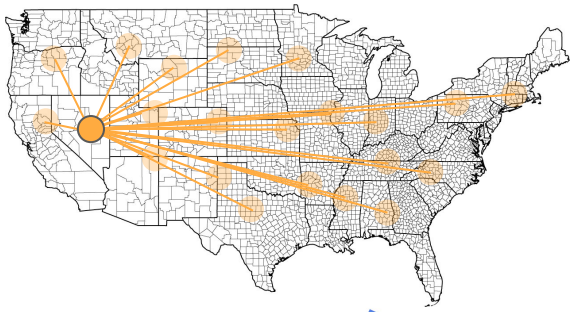
We combined many different prediction approaches

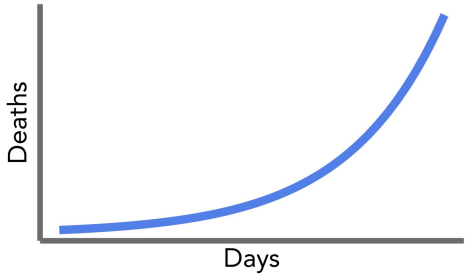
3. **Shared**-county exponential model





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



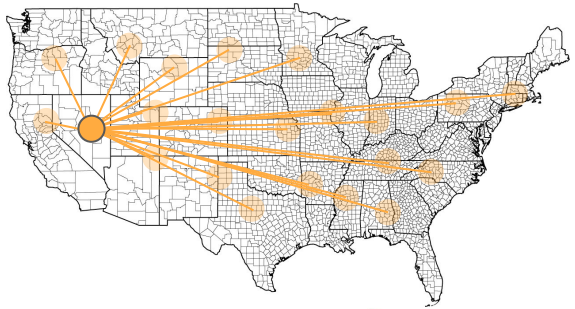


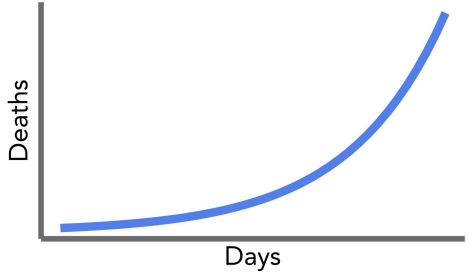
+ Age
+ ICU Beds
+ # Hospitals
+ ....

Ensemble predictors

We combined many different prediction approaches

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





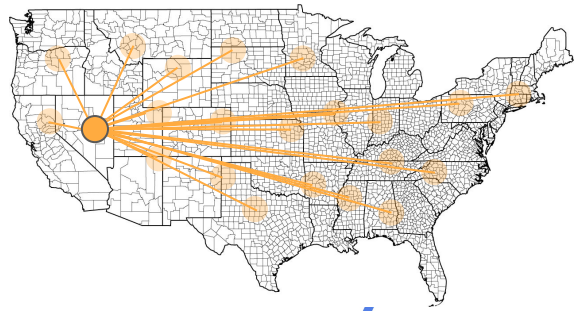
+ Age
+ ICU Beds
+ # Hospitals
+ ....

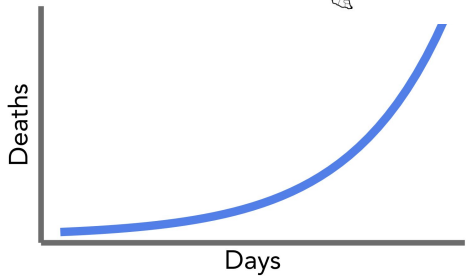
- County density and size
- County healthcare resources
- Demographic information

Ensemble predictors

We combined many different prediction approaches

5. **<u>Expanded Shared</u>**-county exponential model

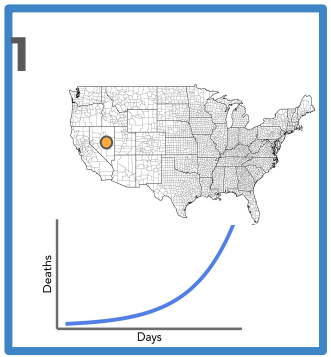




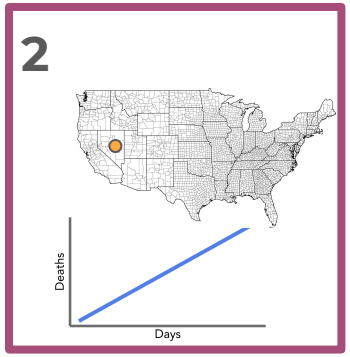
+ Cases
+ Neighboring cases
+ Neighboring deaths

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

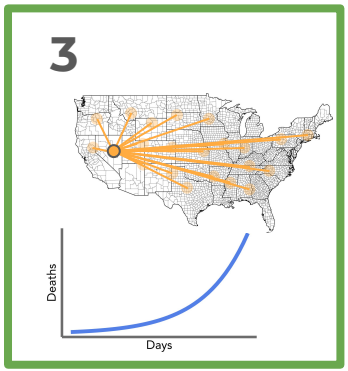
Combined Linear and Exponential Predictors (CLEP)



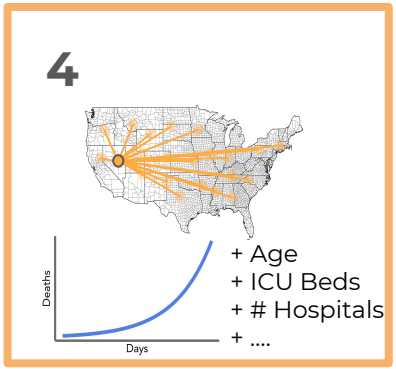
Separate-county exponential predictor



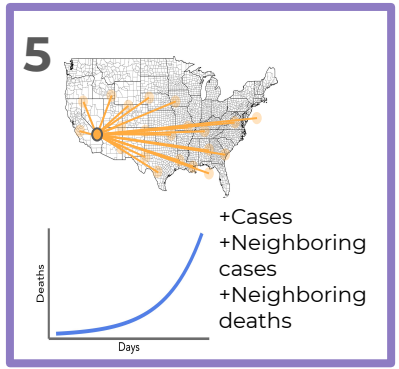
Separate-county linear predictor



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)

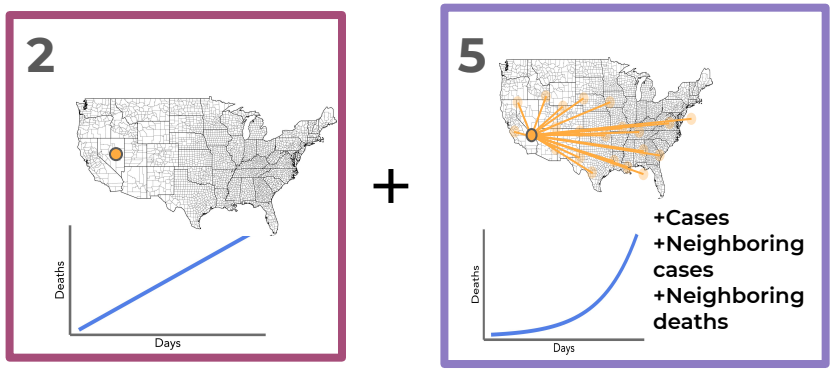
Calculate a **weighted average of the predictions**: higher weight to the models with better historical performance[2]

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

[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.