Covid-19 HASTE Workshop

April 15, 2020

Hacking the New York Times Covid-19 Dataset

John Lafferty

Department of Statistics & Data Science

Yale University

This pdf includes animations, and is best viewed with Adobe Acrobat.

Outline

- Motivation
- The New York Times dataset
- Some visualizations
- A hierarchical Bayesian model of infection
- Opportunities

YData: Computational and Inferential Thinking

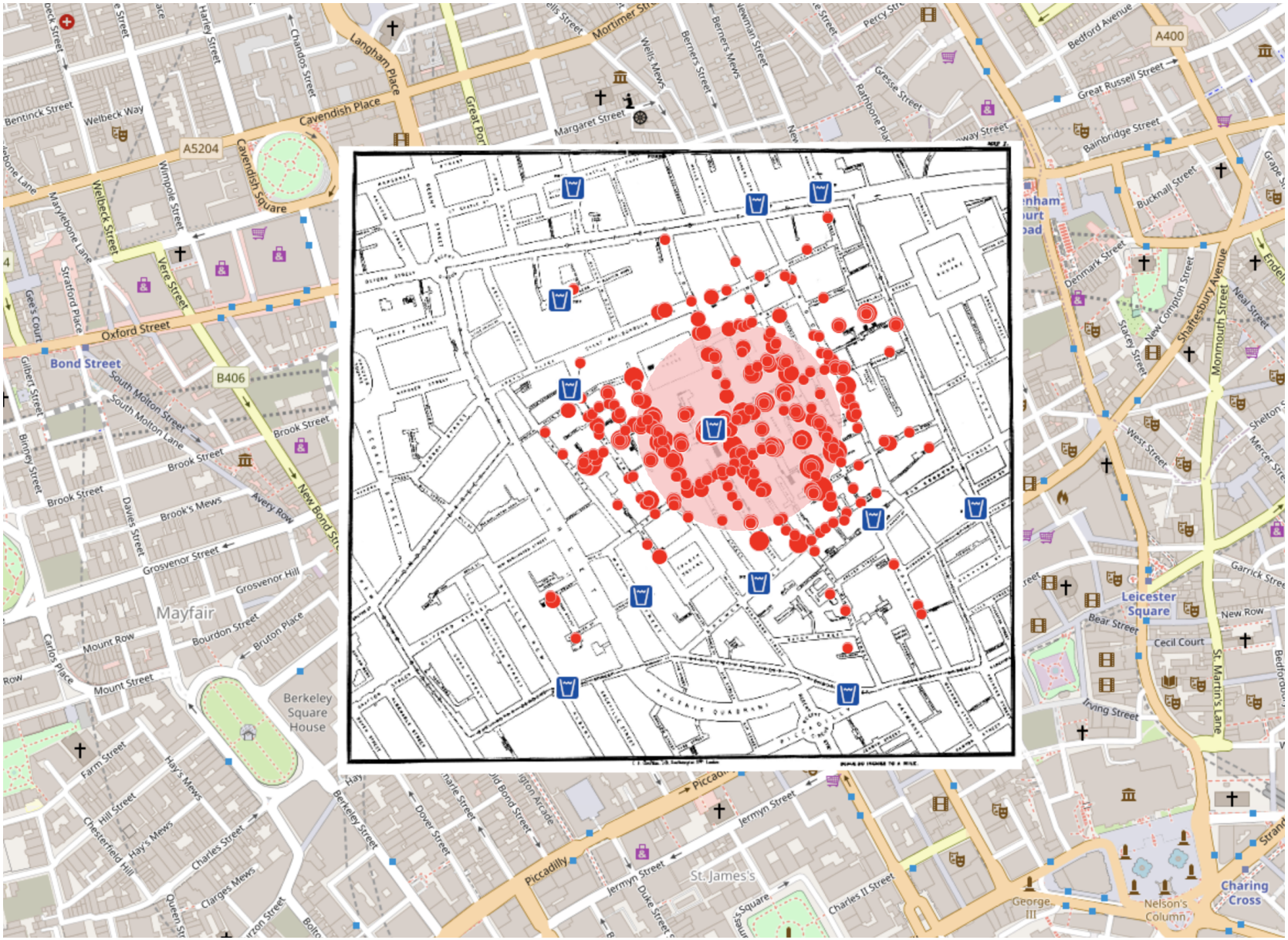
- Programming in Python: Dataframes, Tables, ...
- Visualization, sampling and empirical distributions, ...
- Testing hypotheses, simulation, ...
- Causality and natural experiments

A spinoff of Berkeley's hugely popular "Data8" course.

John Snow and the Broad Street Pump

Yale Maps: https://bit.ly/3elJBkL

John Snow and the Broad Street Pump



Yale Maps: https://bit.ly/3elJBkL

Habits of mind

"All Yale College students should develop the habits of mind that will enable them to identify the strengths and weaknesses in empirical evidence, ask probing questions about empirical claims, and use quantitative evidence wisely in forming opinions and making decisions."

The Gerber Report

February 24, 2020

`https://bit.ly/3cghfa0`

The New York Times Covid-19 Database

- County-level database of confirmed cases and deaths
- Compiled from state and local governments and health departments
- Initial release: Thursday, March 26, 2020; updated daily
- Has fueled many articles and graphics by The Times
- https://github.com/nytimes/covid-19-data

Mitch Smith, Karen Yourish, Sarah Almukhtar, Keith Collins, Danielle Ivory and Amy Harmon have been leading our U.S. data collection efforts. Data has compiled by Jordan Allen, Jeff Arnold, Aliza Aufrichtig, Mike Baker, Robin Berjon, Matthew Bloch, Nicholas Bogel-Burroughs, Maddie Burakoff, Christopher Calabrese, Andrew Chavez, Robert Chiarito, Carmen Cincotti, Alastair Coote, Matt Craig, John Eligon, Tiff Fehr, Andrew Fischer, Matt Furber, Rich Harris, Lauryn Higgins, Jake Holland, Will Houp, Jon Huang, Danya Issawi, Jacob LaGesse, Hugh Mandeville, Patricia Mazzei, Allison McCann, Jesse McKinley, Miles McKinley, Sarah Mervosh, Andrea Michelson, Blacki Migliozzi, Steven Moity, Richard A. Oppel Jr., Jugal K. Patel, Nina Pavlich, Azi Paybarah, Sean Plambeck, Carrie Price, Scott Reinhard, Thomas Rivas, Michael Robles, Alison Saldanha, Alex Schwartz, Libby Seline, Shelly Seroussi, Rachel Shorey, Anjali Singhvi, Charlie Smart, Ben Smithgall, Steven Speicher, Michael Strickland, Albert Sun, Thu Trinh, Tracey Tully, Maura Turcotte, Miles Watkins, Jeremy White, Josh Williams and Jin Wu.

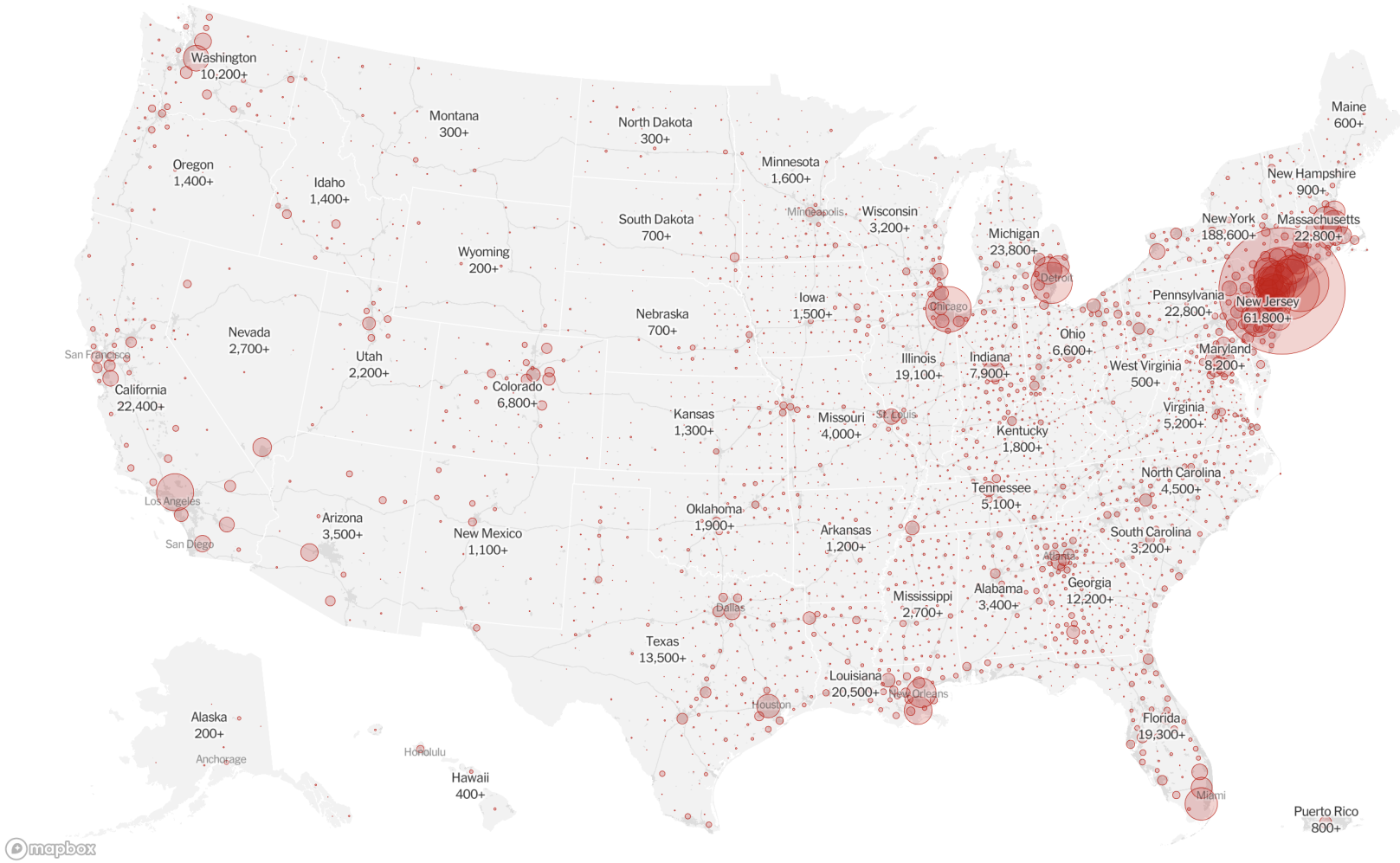
Hack #1: Reverse-engineering the Times maps

- Weekend of March 27
- Python interface to Plotly
- “Bubble maps” showing cases by county
- Merged with population data from US Census
- Added geocoding interface to input addresses
- Closely matched The Times maps published online

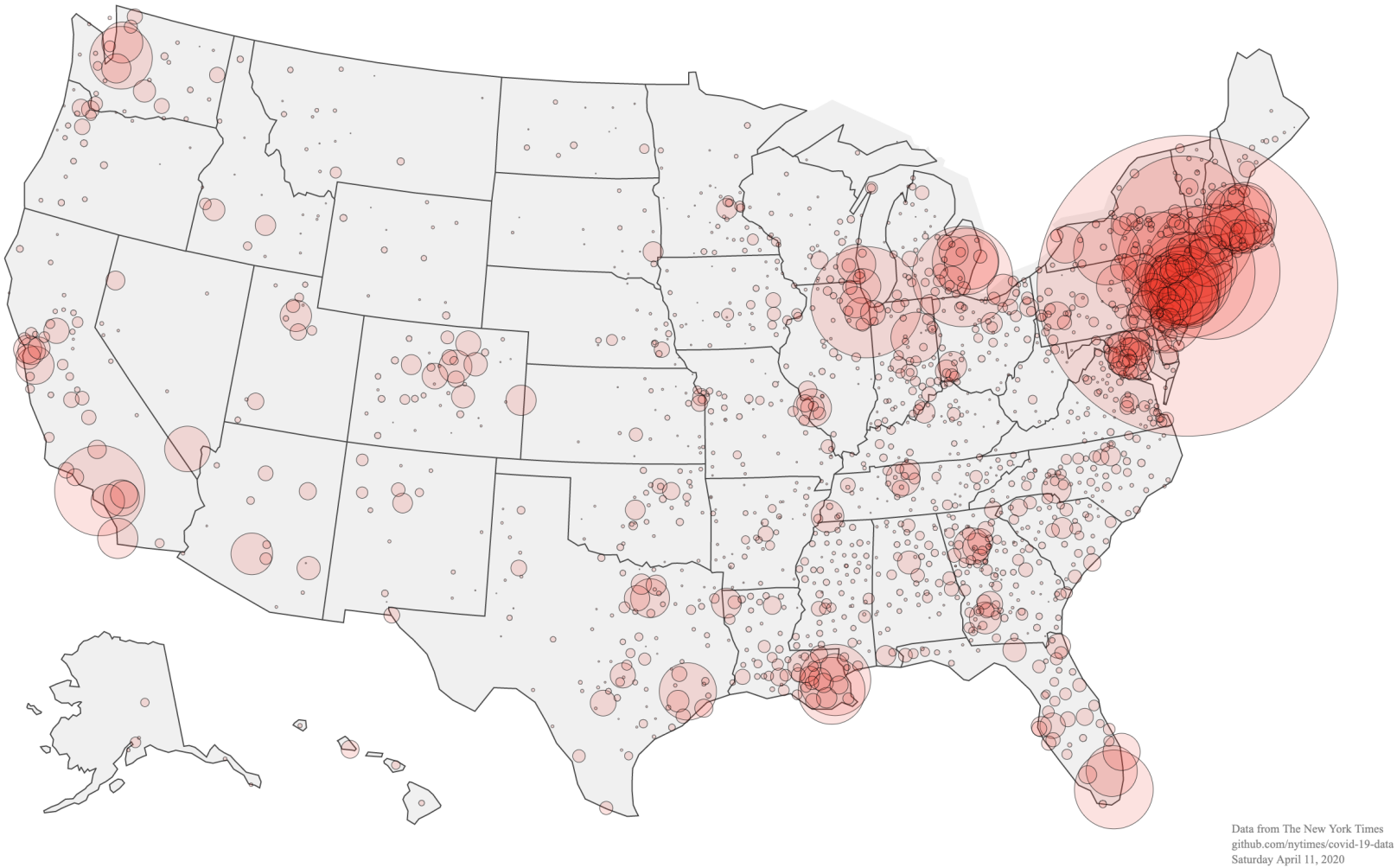
---

http://covid-19.ydata123.org

April 12


Washington
10,200+
Oregon
1,400+
California
22,400+
Nevada
2,700+
Idaho
1,400+
Montana
300+
Wyoming
200+
Utah
2,200+
Arizona
3,500+
Colorado
6,800+
New Mexico
1,100+
North Dakota
300+
South Dakota
700+
Nebraska
700+
Kansas
1,300+
Oklahoma
1,900+
Texas
13,500+
Minnesota
1,600+
Iowa
1,500+
Missouri
4,000+
Arkansas
1,200+
Louisiana
20,500+
Wisconsin
3,200+
Illinois
19,100+
Michigan
23,800+
Indiana
7,900+
Ohio
6,600+
Kentucky
1,800+
Tennessee
5,100+
Mississippi
2,700+
Alabama
3,400+
Georgia
12,200+
Florida
19,300+
South Carolina
3,200+
North Carolina
4,500+
Virginia
5,200+
West Virginia
500+
Pennsylvania
22,800+
New York
188,600+
Maryland
8,200+
New Jersey
61,800+
Massachusetts
22,800+
New Hampshire
900+
Maine
600+
Alaska
200+
Hawaii
400+
Puerto Rico
800+
Minneapolis
Chicago
Detroit
St. Louis
Atlanta
Dallas
Houston
New Orleans
Miami
San Francisco
Los Angeles
San Diego
Anchorage
Honolulu
mapbox

Reconstruction: April 12



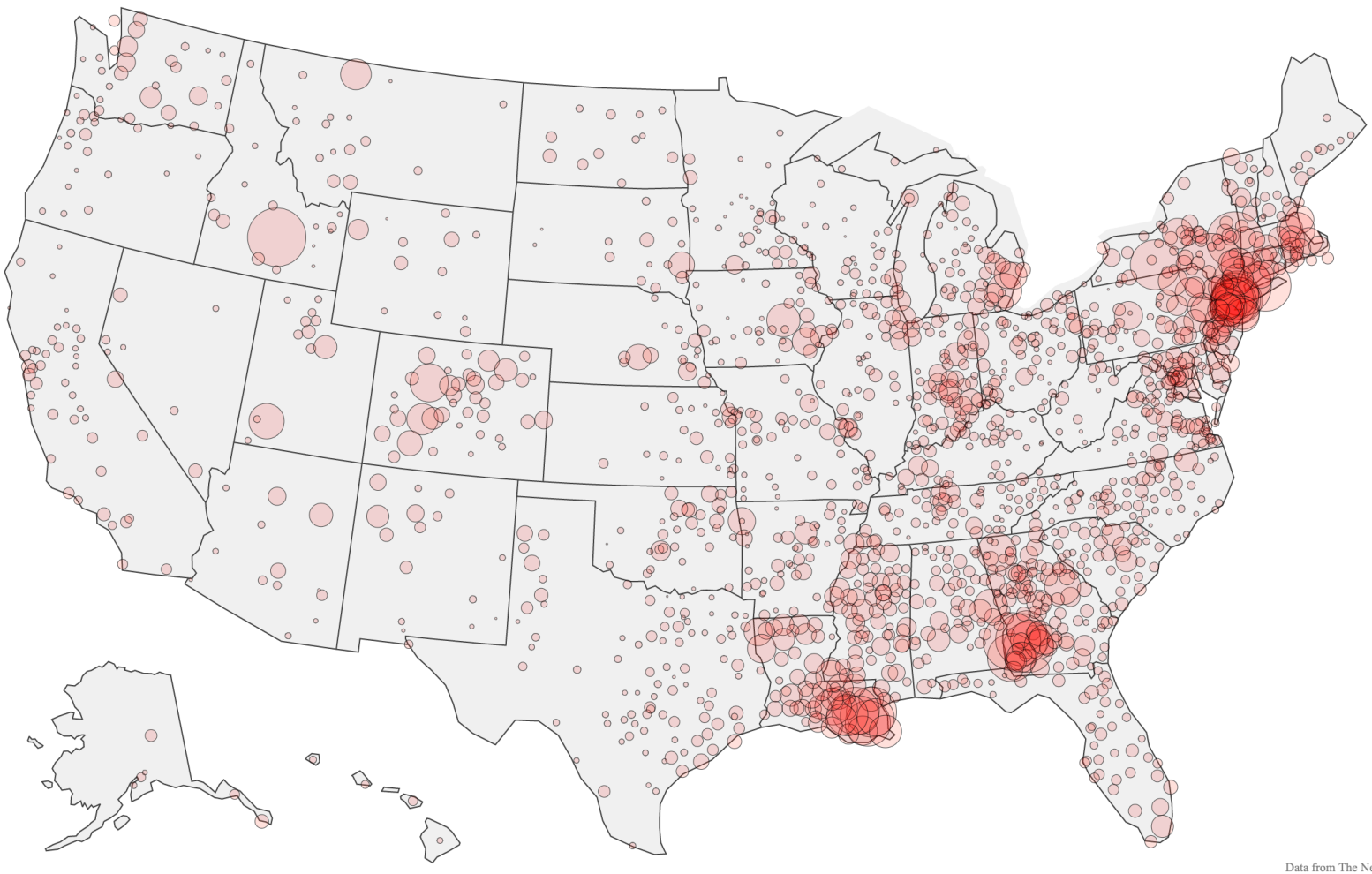
“Sustainability” issues with Times visualizations

- Scaling factor is arbitrary
- Initial emotional response: “A New York problem?”
- Area scales linearly with cases
- Not “sustainable” with exponential growth
- Not normalized by population
- Similar issues for JHU maps and ubiquitous political maps (Gerber, Huber, Nordhaus...)

---

https://www.natureindex.com/news-blog/behind-johns-hopkins-university-coronavirus-dashboard (April 7),
https://arstechnica.com/science/2020/04/new-covid-19-dashboard-just-for-the-us-offers-rich-county-level-data/ (April 13)

Cases per 100,000 people



Data from The New York Times
github.com/nytimes/covid-19-data
Tuesday April 14, 2020

Input address Go ☑ per 100,000 people

A simple hierarchical model

For each county c,

$$P_c \sim F$$
$$Y_c \,|\, P_c = p_c \sim \text{Binomial}(n_c, p_c)$$

Transform by $\psi_c \equiv \log\left(\frac{p_c}{1-p_c}\right)$ to use Gaussian approximation:

$$\psi_c \sim N(\mu, \tau^2)$$
$$Z_c \,|\, \psi_c \sim N(\psi_c, \sigma_c^2)$$

Run Gibbs sampling to carry out posterior inference

A simple hierarchical model

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code Dashboard View:

In [1]:
```
import covid19 as cvd
import covid19_predict as cvd_predict
```

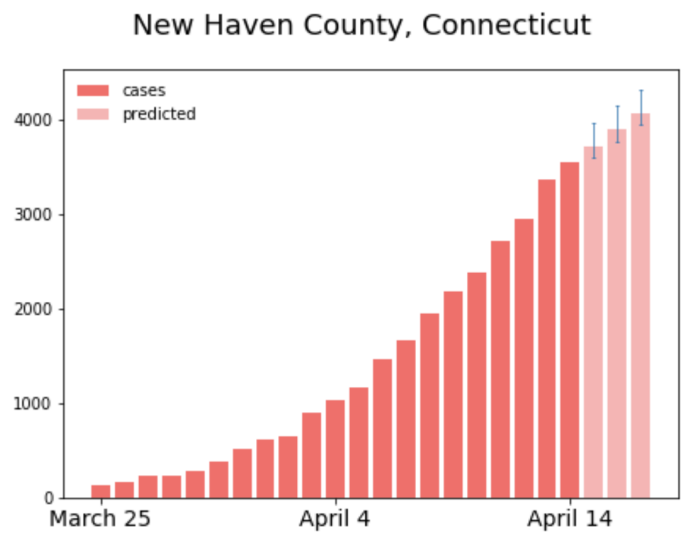
```
covid19: Most recent NY Times data: Tuesday April 14, 2020
covid19_predict: initializing for simulation
covid19_predict: running Gibbs sampler: n=3193, B=10000
covid19_predict: making predictions
```

In [2]:
```
cvd_predict.df[cvd_predict.df['county']=='New Haven']
```

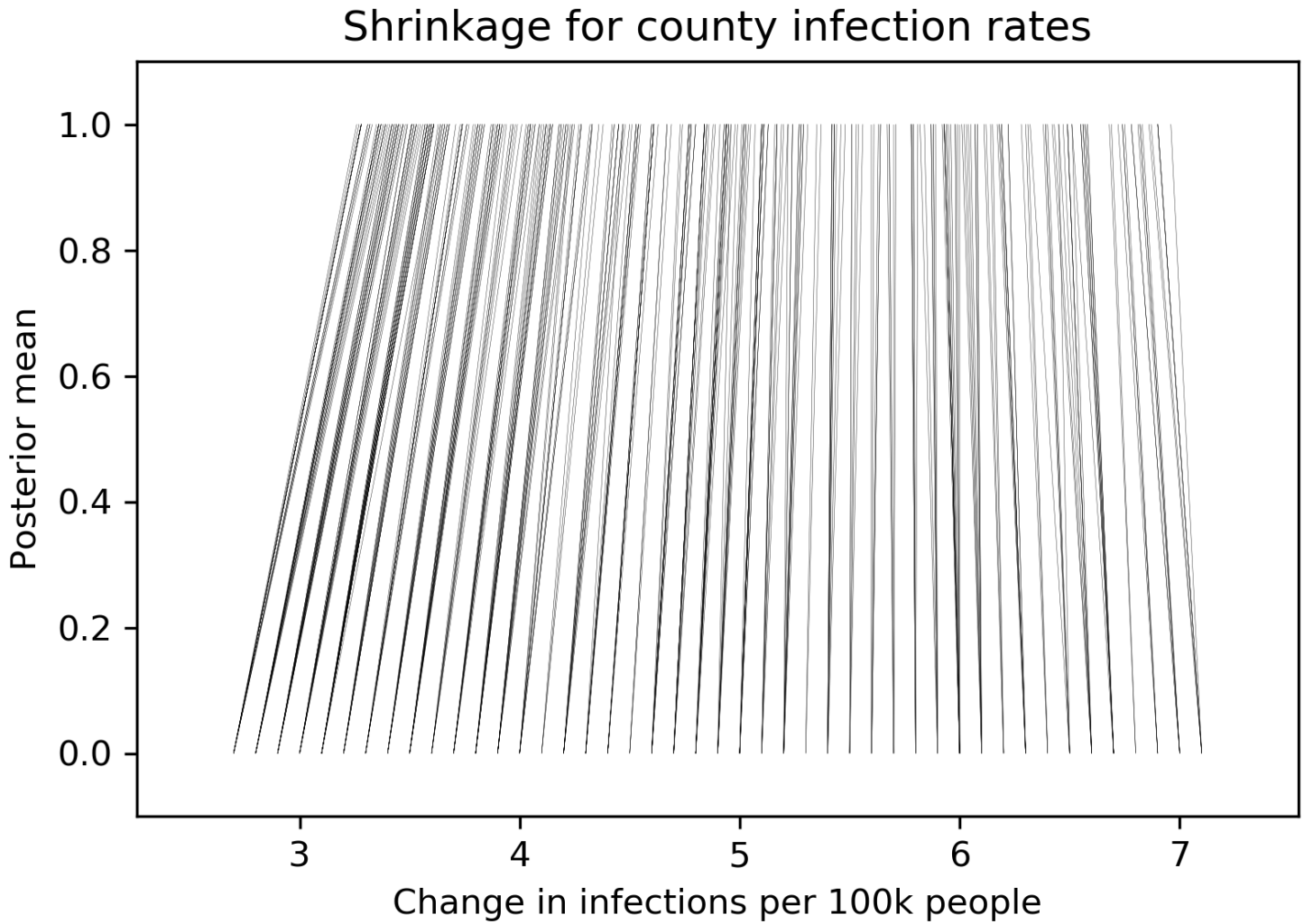
Out[2]:

	date	county	state	cases	deaths	population	cases_per_100k	delta	delta_bar	delta_95	delta_05	cases_predicted	cases_95_credible
83	2020-04-14	New Haven	Connecticut	3543	151	854757	414.5	21.6	20.36	28.79	14.42	3716	3789

In [3]:
```
_ = cvd_predict.plot_predictions_for_addr('New Haven, CT', show=True)
```

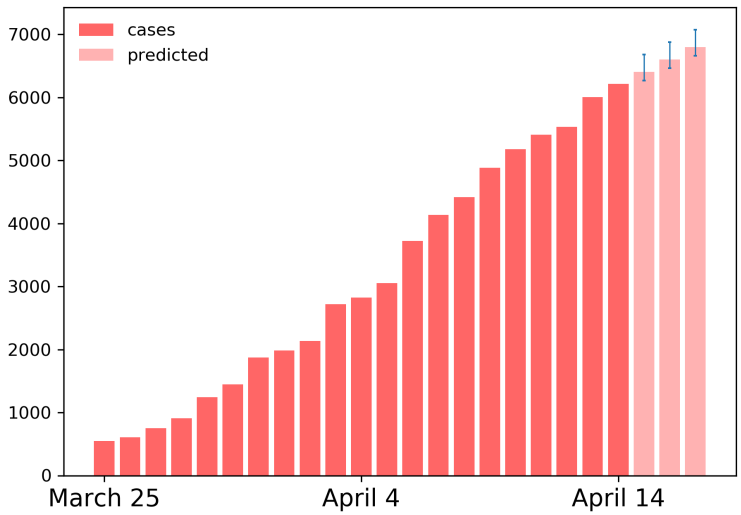


A simple hierarchical model

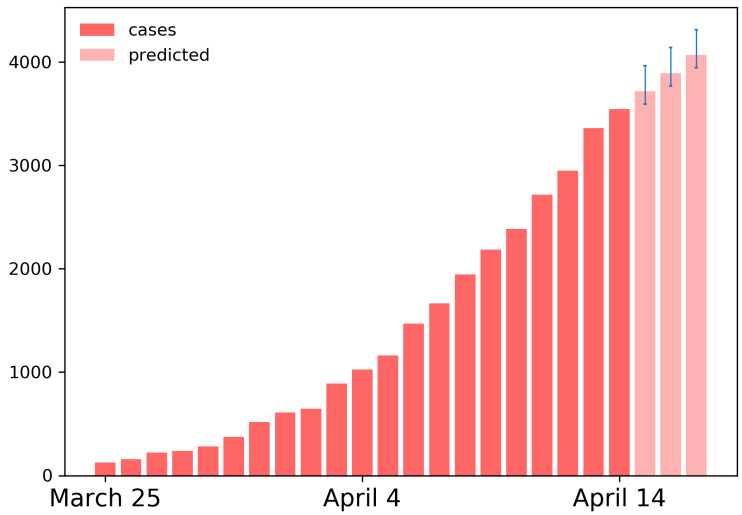


Examples

Fairfield County, Connecticut

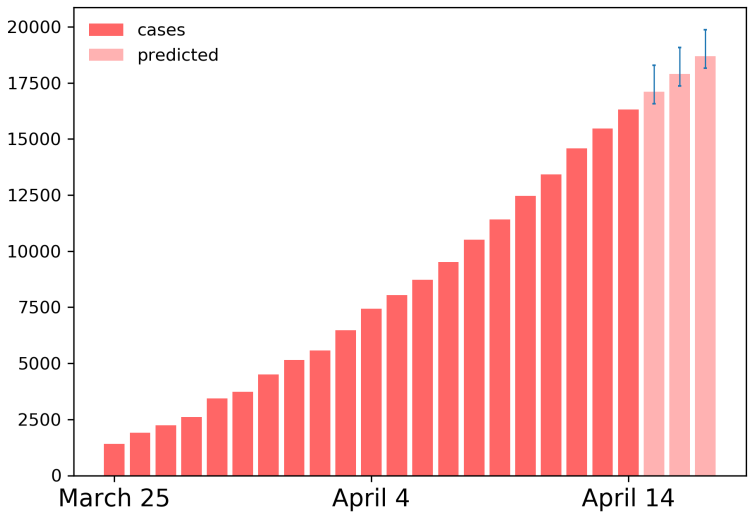


New Haven County, Connecticut

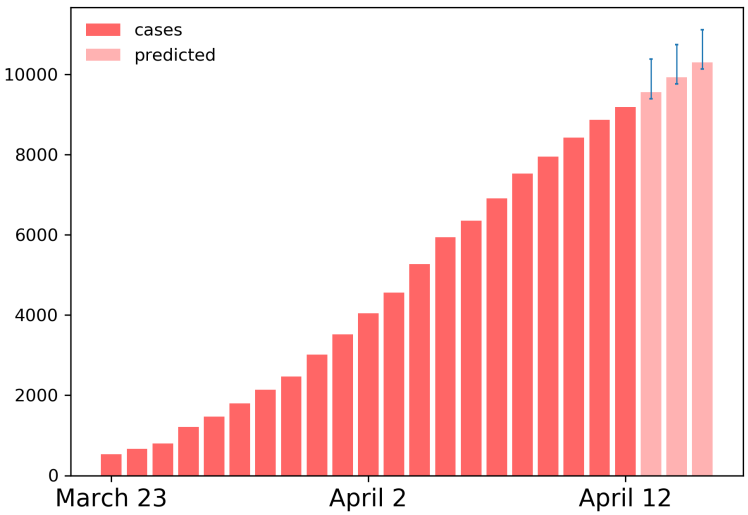


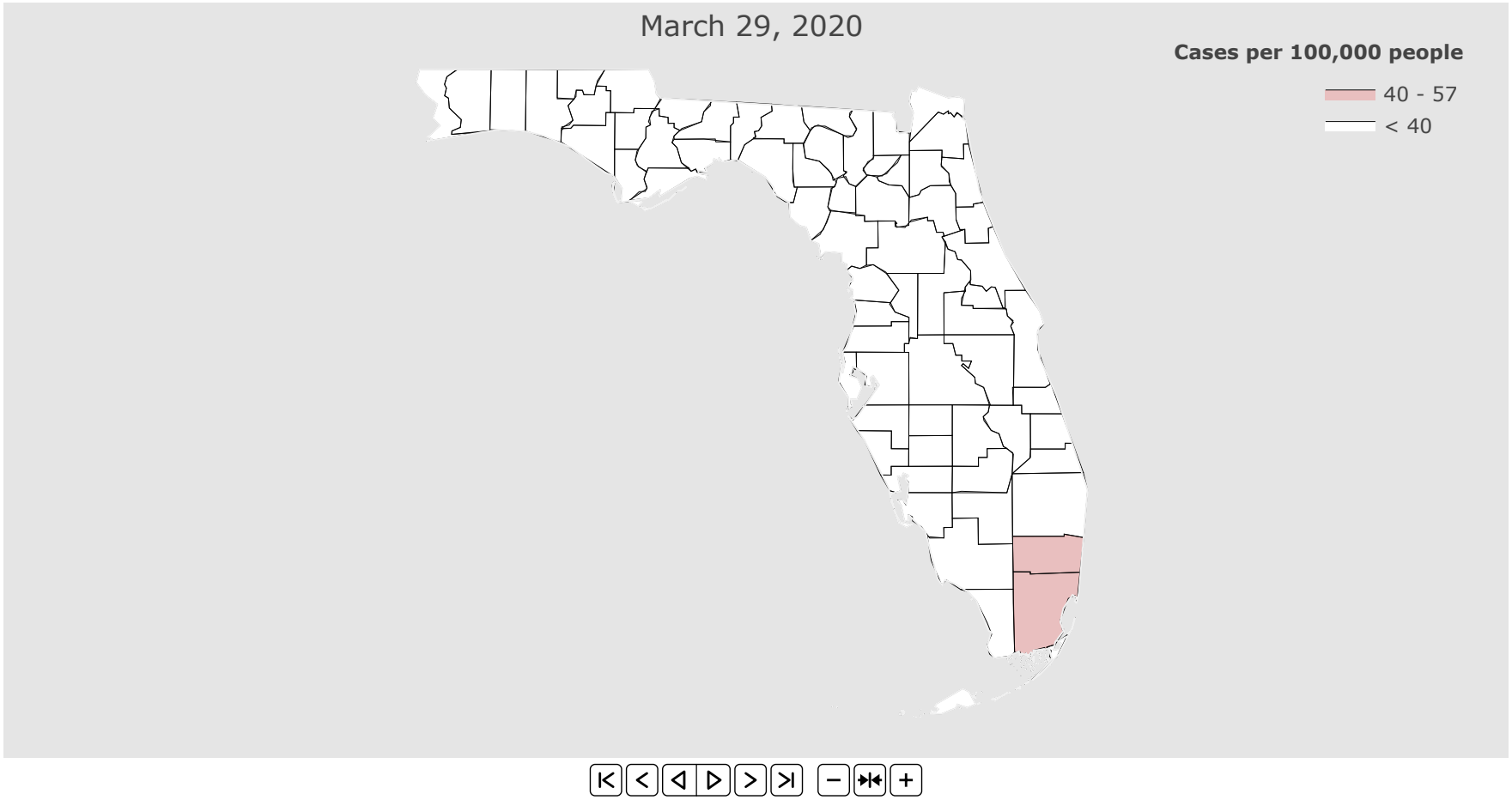
Examples

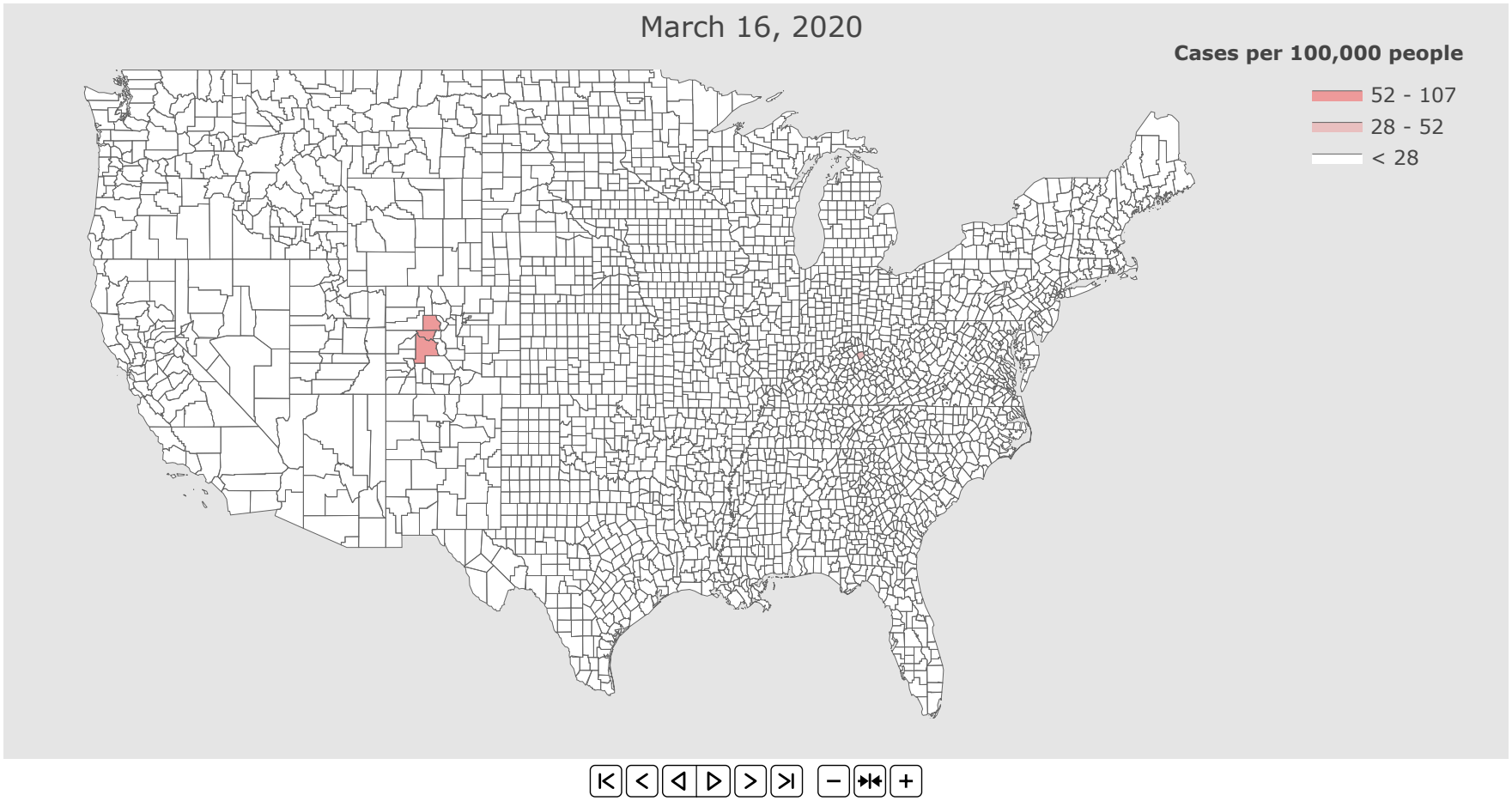
Cook County, Illinois



Los Angeles County, California




March 29, 2020
Cases per 100,000 people
40 - 57
< 40


March 16, 2020
Cases per 100,000 people
52 - 107
28 - 52
< 28

Opportunities

- Engage Yale community in brainstorming, tool-development
- Merge in diverse data — population density, demographics, economic indicators, ...
- Create Python package to allow rapid exploration, testing
- Extend simple statistical models
- Look for natural experiments (à la Broad Street Pump map)
- `https://covid.yale.edu/innovation/mapping/`

---

https://github.com/jdlafferty/covid-19

Gerber report redux

We—the Yale data science community—have a responsibility to help others identify the strengths and weaknesses in empirical evidence, ask probing questions about empirical claims, and use quantitative evidence wisely in forming opinions and making decisions.