

CTIS: The COVID-19 Trends and Impact Survey

Tracking a Pandemic in Real Time

Alex Reinhart

Delphi Group, Carnegie Mellon University

December 7, 2021

Delphi

- Since 2012, Delphi has developed "the theory and practice of epidemic forecasting, and its role in decision making"
- Led by Roni Rosenfeld and Ryan Tibshirani, with several participating faculty and graduate students
- Participated in annual CDC flu forecasting challenges, won several
- Named an Influenza Forecasting Center of Excellence by the CDC in 2019
- Published open code and data, including numerous influenza surveillance streams
- [I joined in April 2020]

<https://delphi.cmu.edu/>

Delphi's COVID-19 Response

March 2020 saw a rapid expansion in Delphi and a change in goals

Now, with **over 70 members**, Delphi develops COVIDcast: data sources, maps, surveys, and code to support researchers, plus COVID forecasting

not everyone! →



Today's Focus

This presentation is the reverse of most applied statistics talks.

1. The COVIDcast project
2. The COVID-19 Trends and Impact Survey (CTIS)
3. Statistical challenges and results
4. All the *other* challenges and results
5. How you can get involved

COVIDcast

Motivation

Imagine yourself in March 2020. How do you help public health officials make decisions when little data is available?

A hierarchy of data types:

1. Deaths – publicly available
2. ICU use – not consistently available
3. Hospitalization – harder to get (EMR, insurance claims...)
4. Case ascertainment – publicly available (lab tests)
5. Outpatient visits – harder to get (EMR, insurance claims...)
6. Symptoms – ???
7. Infections – only via seroprevalence surveys
8. General population – mobility data

We want to fill out the hierarchy of public data.

COVIDcast

The COVIDcast project has many parts:

1. Code and infrastructure to obtain "indicators" daily—each indicator measures some signal relevant to COVID-19
2. Unique relationships with healthcare and tech partners granting us access to indicators
3. A historical database of all indicators, including revision tracking, with >2.3b observations
4. An open API for requesting this data, with R and Python packages for easy access
5. An interactive visualization, built on the API, at delphi.cmu.edu/covidcast/
6. Forecasting and modeling work building on the data and API

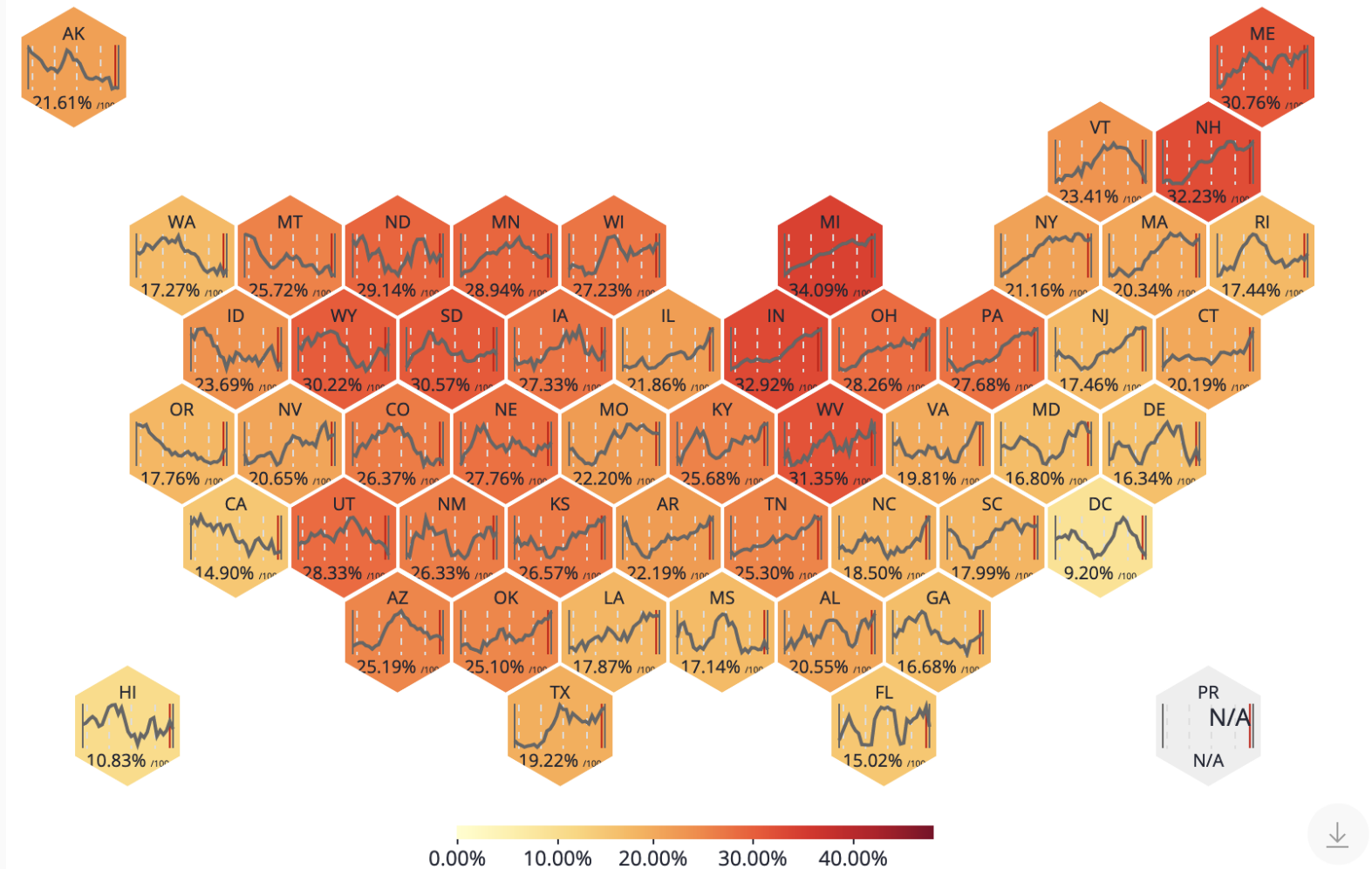
COVIDcast Indicators

Freely available through the COVIDcast API, updated daily:

1. **Deaths** – from public reports
2. ICU use – not yet
3. **Hospitalization** – from claims data and HHS data
4. **Case ascertainment** – from public reports, Quidel antigen tests, CTIS
5. **Outpatient visits** – from claims data
6. **Symptoms** – from CTIS, Google Search Trends
7. Infections – not yet
8. **General population** – SafeGraph mobility, plus CTIS

Most at the county level!

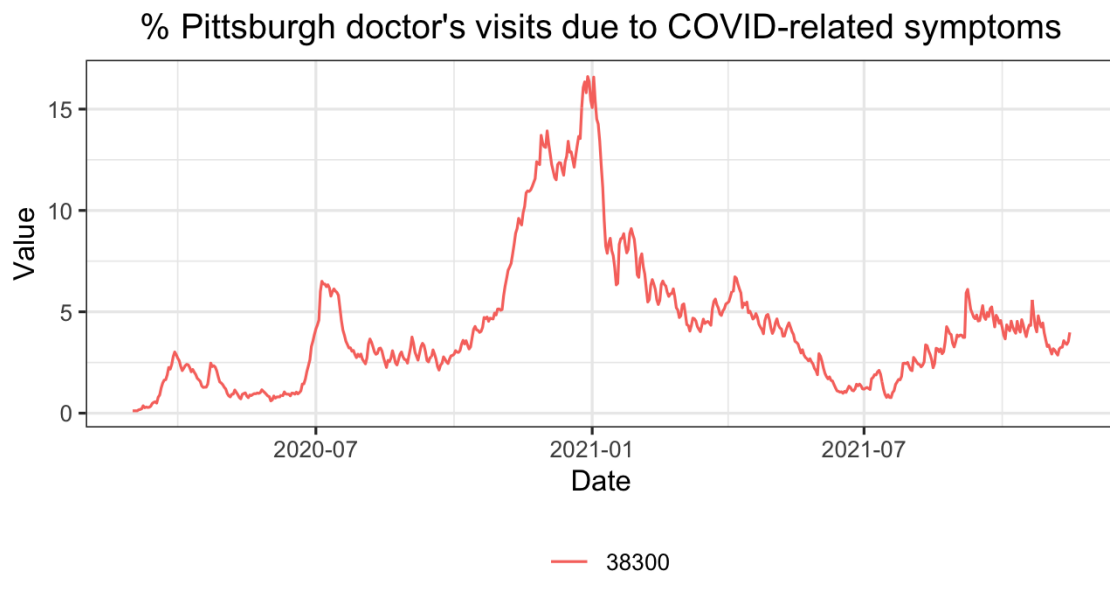
Full list: https://cmu-delphi.github.io/delphi-epidata/api/covidcast_signals.html



COVIDcast API Access

All indicators are freely available through our API:

```
library(covidcast)
dv_pitt ← covidcast_signal(
  "doctor-visits", "smoothed_adj_cli",
  start_day = "2020-03-01", end_day = "2021-11-15",
  geo_type = "msa", geo_values = name_to_cbsa("Pittsburgh"))
plot(dv_pitt, plot_type = "line") +
  labs(title = "% Pittsburgh doctor's visits due to COVID-related symptoms")
```



Alternate Data Sources



Severity Pyramid

A hierarchy of data types:

1. Deaths
2. ICU use
3. Hospitalization
4. **Case ascertainment**
5. Outpatient visits
6. **Symptoms**
7. Infections
8. **General population**

Through a recruitment partnership with Facebook, Delphi surveys **40,000 people daily** (25 million since April 2020) in the United States about

- symptoms they are currently experiencing
- COVID testing
- COVID vaccination uptake, acceptance, and obstacles
- mask wearing and social distancing
- mental health
- demographics

A parallel effort by the University of Maryland reaches 100+ countries globally.

Alex, Take a COVID-19 Survey From Carnegie Mellon University

Even if you feel well, your survey participation may help health researchers predict the spread of COVID-19. Could you take a few minutes to answer a short survey from Carnegie Mellon University?

[View Survey](#)

Not Now

- Survey implemented on Qualtrics, managed by CMU
- Facebook does **not** receive individual responses
- Designed to take about 10 minutes; about 35 questions
- Questions selected for relevance for forecasting but also for research and public health
- Facebook calculates survey weights designed to demographically match US state age & gender while accounting for non-response bias
- Respondents provide their ZIP code
- Individual response files shared with 60+ research groups

This is the largest non-Census research survey ever conducted (that we know of).

<https://delphi.cmu.edu/covid19/ctis/>

Details: Salomon, Reinhart, Bilinski, Chua, La Motte-Kerr, Rönn, Reitsma, Morris, LaRocca, Farag, Kreuter, Rosenfeld, and Tibshirani (2021). [The U.S. COVID-19 Trends and Impact Survey, 2020-2021: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing and vaccination](#). *Proceedings of the National Academy of Sciences*, in press.

Access to Survey Data

County-level aggregates available in the COVIDcast API:

- COVID vaccine uptake and acceptance; reasons for vaccine hesitancy
- estimated population percentage with COVID-like symptoms
- percentage who know someone who is currently sick
- percentage wearing masks regularly
- working outside home, going to bars and restaurants indoors
- testing rates and test positivity

Over 100 signals in total: <https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/fb-survey.html>

Detailed demographic breakdowns also available: <https://cmu-delphi.github.io/delphi-epidata/symptom-survey/contingency-tables.html>

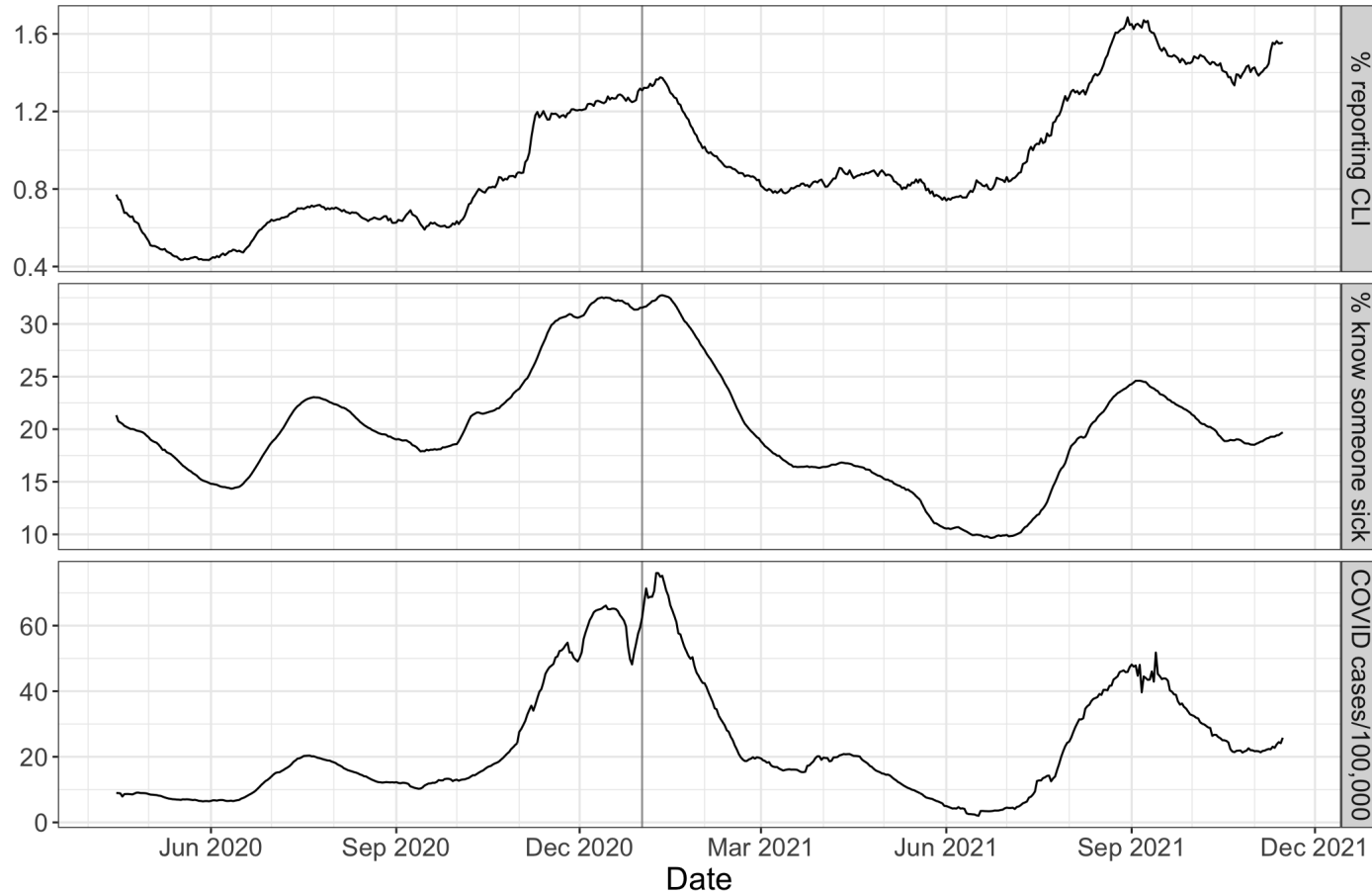
```
library(covidcast)
d ← covidcast_signal(
  "fb-survey", "smoothed_wwearing_mask_7d",
  "2021-12-01", "2021-12-01",
  geo_values = name_to_cbsa("Philadelphia"),
  geo_type = "msa")
d$value
```

```
## [1] 61.72201
```

Syndromic Surveillance

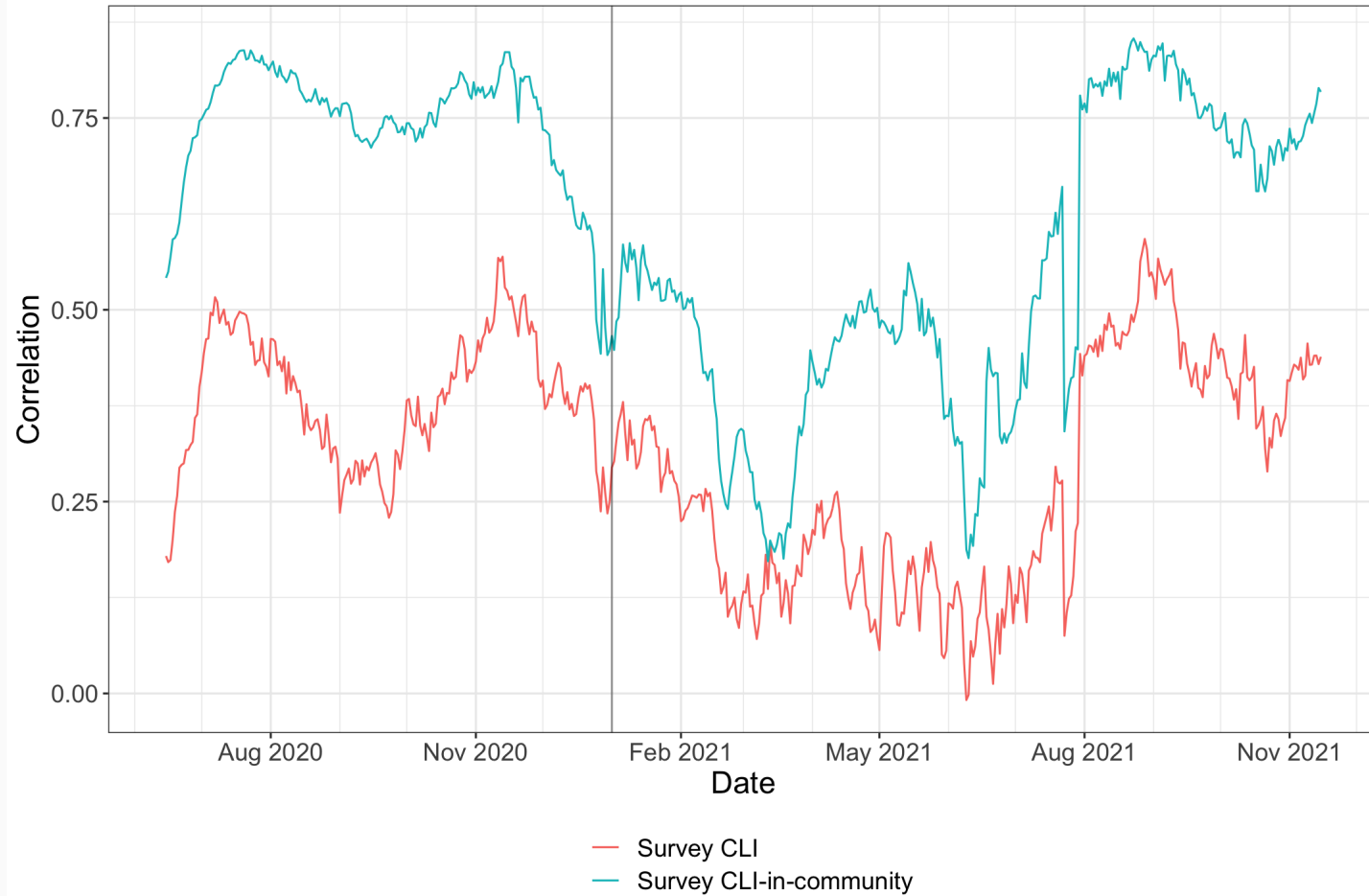
National comparison of survey data to COVID case rates

7-day averages



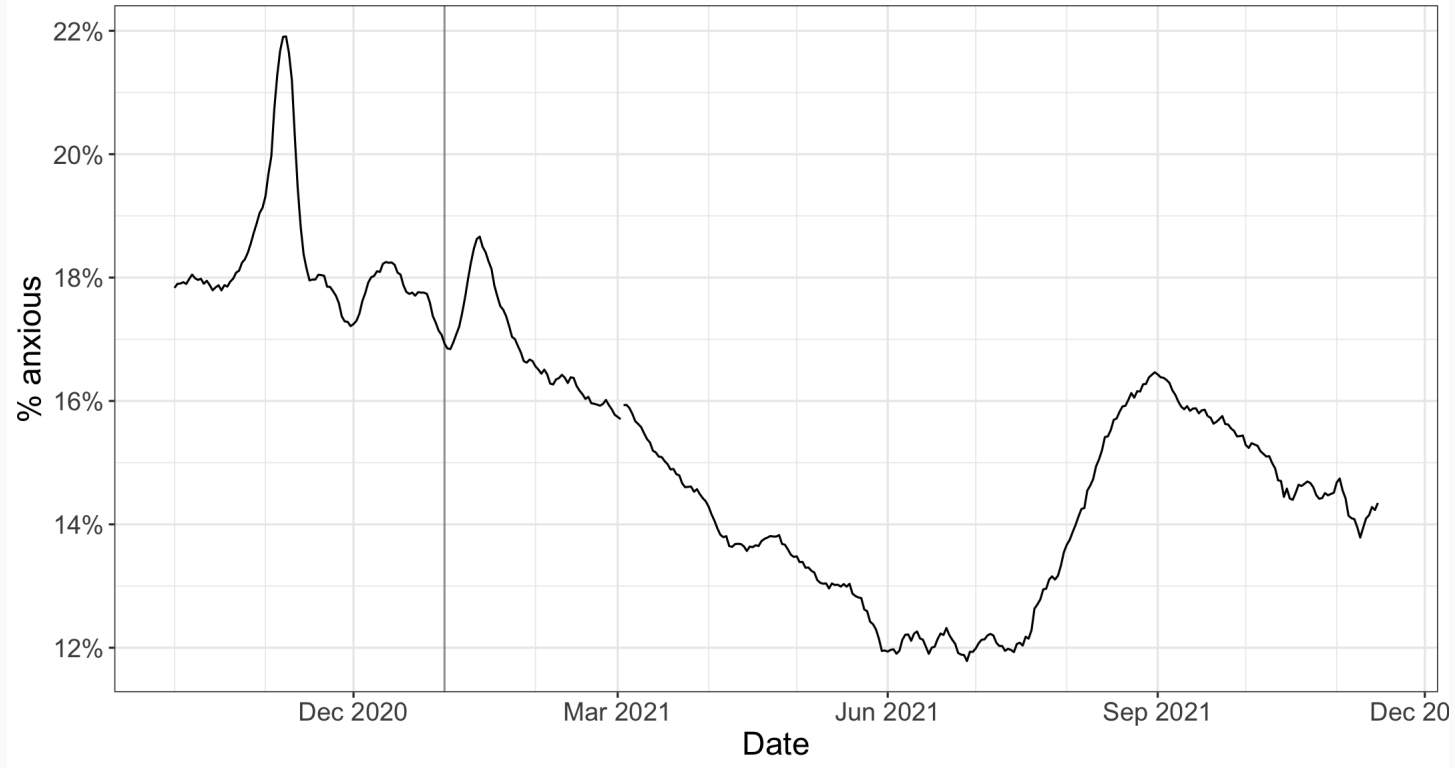
Syndromic Surveillance

Geo-wise correlation between signals and case rates



Not Just Symptoms

Respondents anxious most or all of the time
7-day averages



Survey Revisions

The survey has gone through 10 versions since April 2020:

- Added mask-wearing, more social distancing
- New schooling questions
- COVID vaccination questions constantly changing to capture hesitancy and barriers
- New items on knowledge and beliefs about COVID

Revisions require extensive collaboration:

- IHME, White House, CDC, NIOSH, Johns Hopkins, and others have all had input on priorities and item design
- Researchers see this as a free opportunity to get a huge sample size
- ...but even minor changes can cause trend breaks that harm analysis
- Good statistics and survey design sometimes bow to other priorities

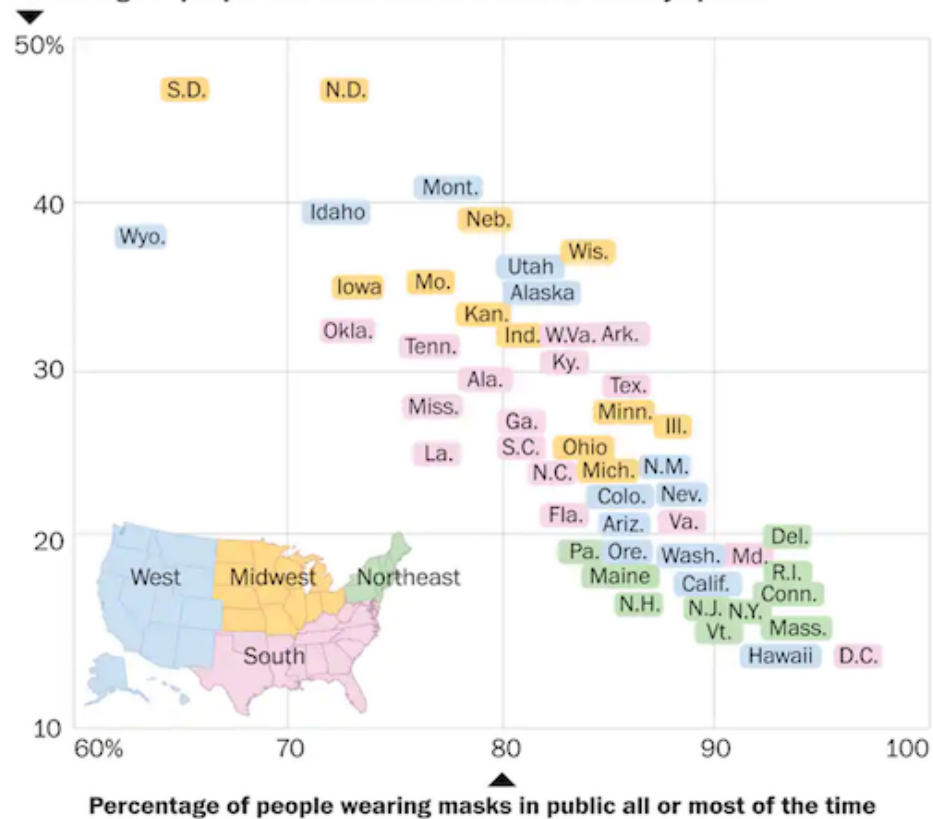
Research and Policy Results

Mask-Wearing

Masking up

Fewer covid-19 symptoms reported in states with higher rates of mask use.

Percentage of people who know someone with covid-19 symptoms



Data as of Oct. 19

Source: Delphi CovidCast, Carnegie Mellon University

THE WASHINGTON POST

Vaccine Hesitancy



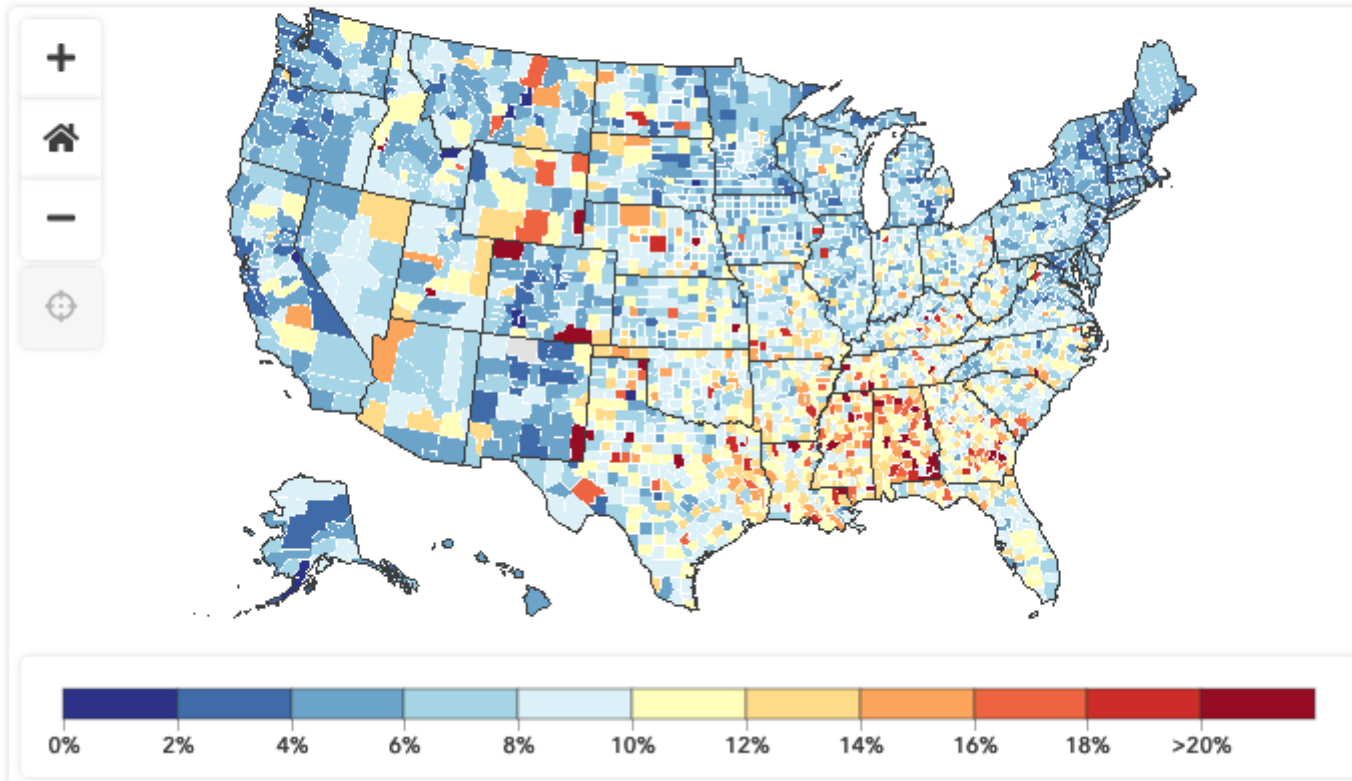
IHME | COVID Collaborative

GHDX VIZ HUB



Vaccine hesitancy by county

Jul 09, 2021 - Jul 15, 2021



Research Results

The comprehensive nature of the survey has facilitated many studies outside Delphi:

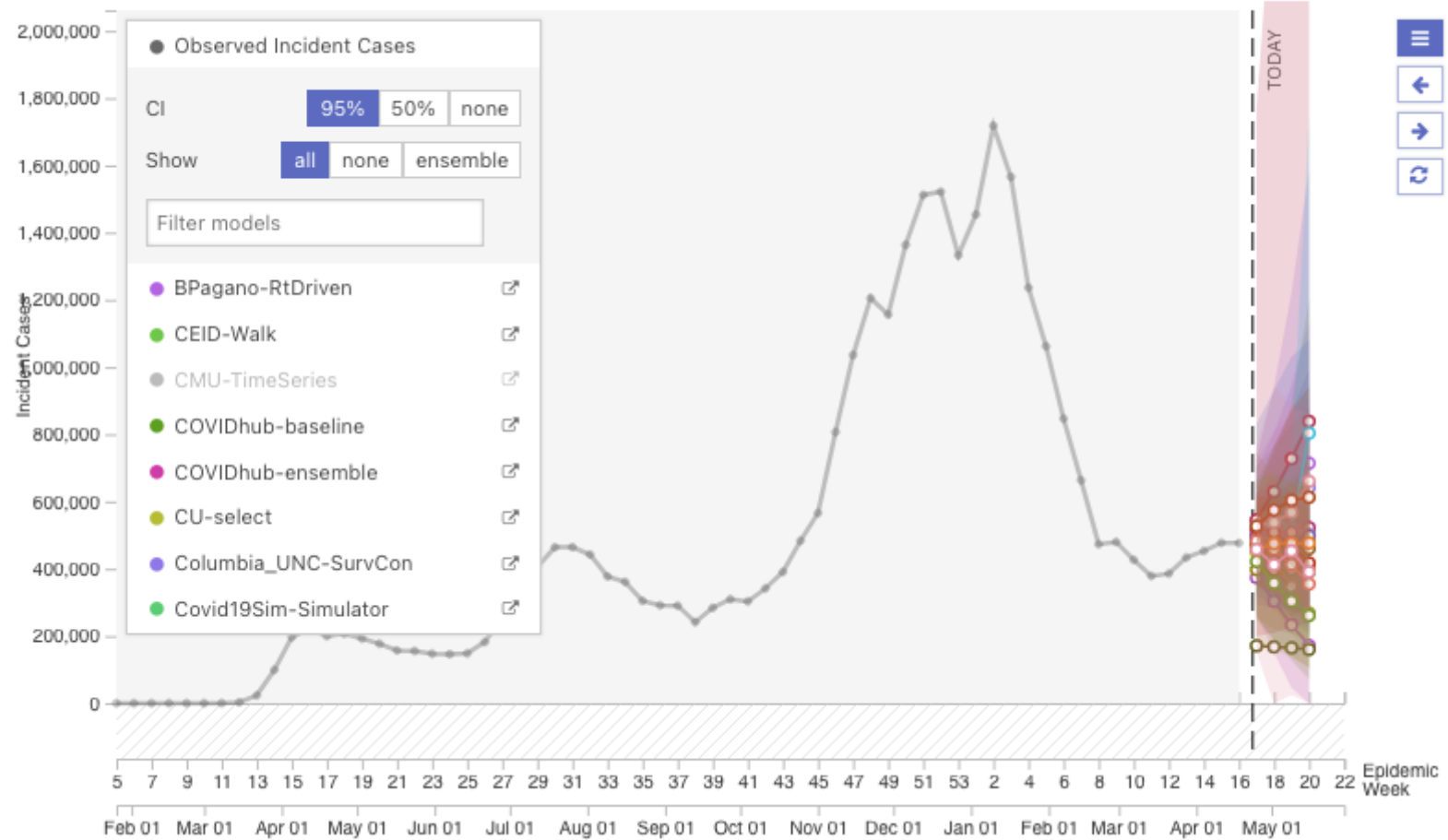
- Anosmia is the symptom most strongly associated with testing positive for COVID (Sudre et al, 2020)
- People change their behavior in reaction to cases, but possibly too late (Bilinski et al, 2021)
- In-person schooling with sufficient precautions may not be a COVID transmission risk (Lessler et al, 2021)
- Healthcare workers show lower COVID incidence than other occupations (Flaxman et al, 2020)
- Vaccine hesitancy varies widely by occupation (King et al, 2021)

and more.

Forecasting

- The CDC sponsors, though UMass Amherst, the **COVID-19 Forecast Hub**
- Dozens of teams submit standardized hospitalization, case, and death **distribution** forecasts for counties and states in the US:

The **ensemble** forecast is a multi-model ensemble developed and published weekly in real-time that combines models with varied approaches, data sources, and assumptions.



Forecasting

Prediction is very difficult, ~~especially if it's about the future~~ even if it's about the present.

- Case and death data is reported with variable lag and varied definitions
- States routinely correct data or post batches of backdated cases
- Case ascertainment varies in space and time

Survey data is (mostly) not subject to these problems, and can be a useful covariate.

- Delphi's Forecast Hub submissions incorporate COVID-like illness estimates from the survey
- Georgia Tech's [DeepCOVID](#) does as well
- IHME incorporates survey data in their COVID scenario modeling & policy briefings
- Survey symptom data appears to allow forecasting several days farther in advance

"However, forecasts of cases and hospitalizations showed repeated, sustained lapses in accuracy for longer-term forecasts, especially at key points during some the larger pandemic waves." ([Reich et al, 2021](#))

Statistics and Public Health

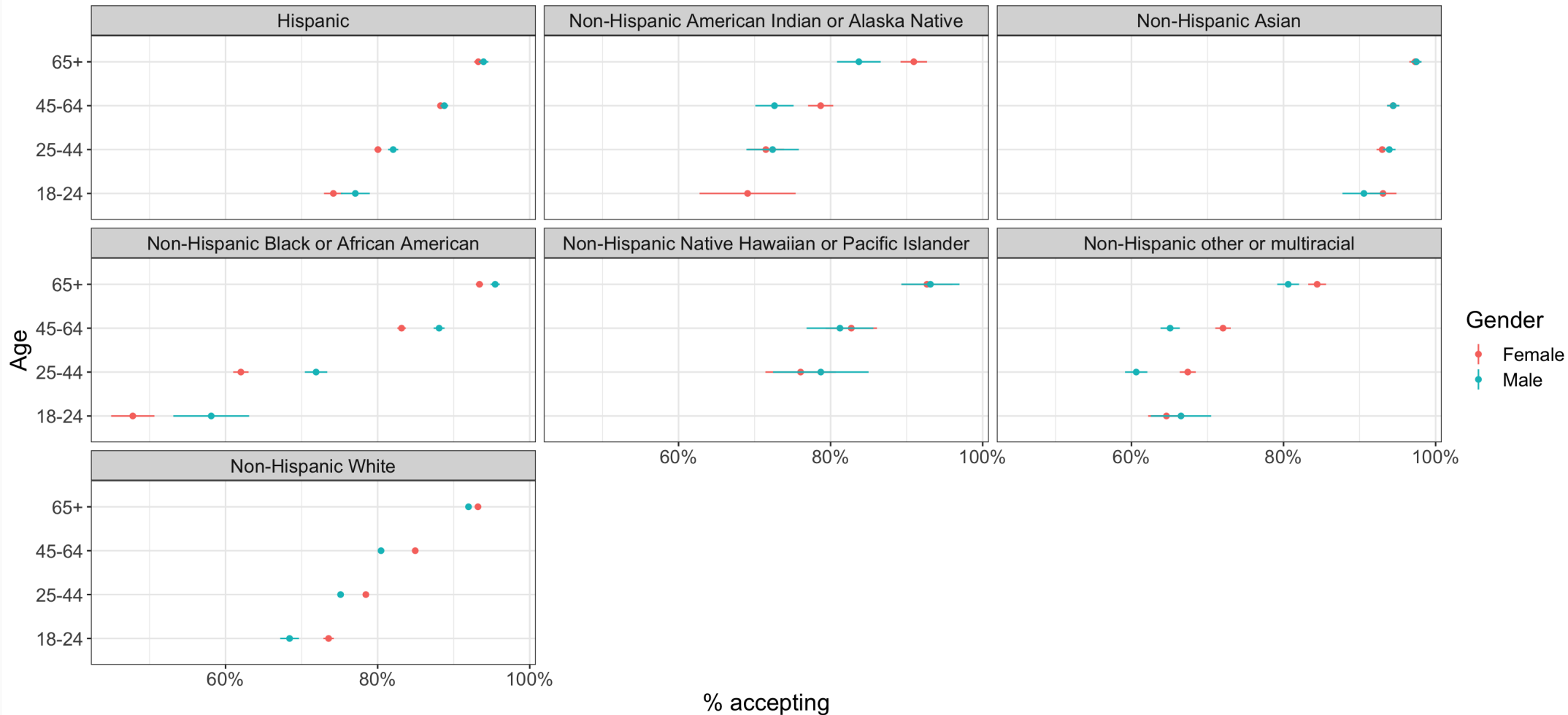
What do Policymakers Need?

- There are numerous interesting statistical problems to solve using this data
 - What symptoms best predict cases? Can we build a symptom-based risk score?
 - Can symptoms and testing data tell us where cases are underreported?
 - Are there combinations of metrics that can predict sudden case increases ("hotspots") a week in advance?
 - Can survey estimates of social behavior and mask use inform epidemic models?
- Public health officials are a little busy.
- They don't need numbers, models, or dashboards; they need *conclusions*
- Common questions:
 - Where should we direct our messaging about the vaccines? What should it say?
 - Are people wearing masks? Should we continue our campaign?
 - How many hospital beds will we need next week?

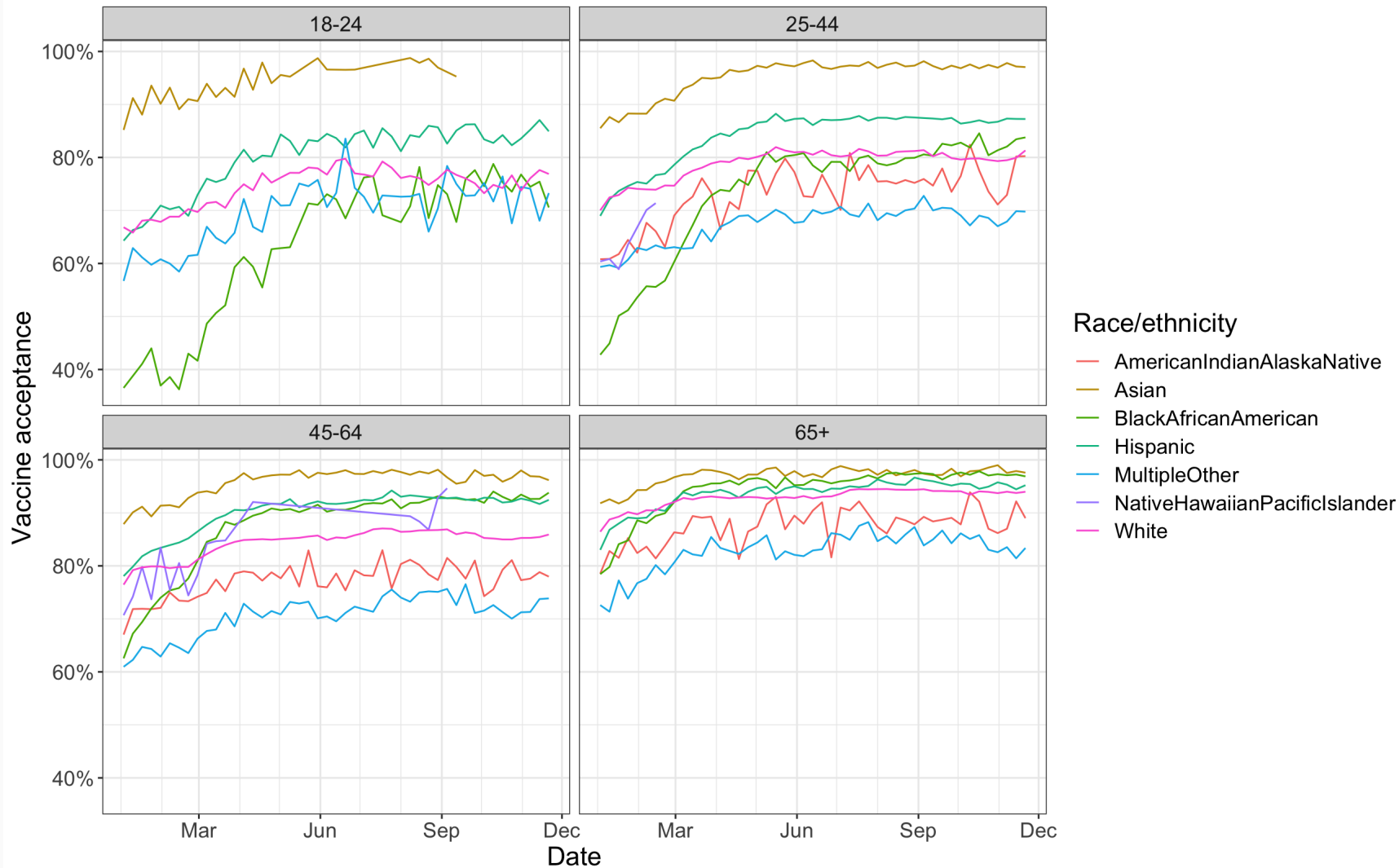
Tracking Vaccine Acceptance

Respondents who are already vaccinated or would definitely/probably get vaccinated

March 2021, N = 917,204



Vaccine Acceptance Over Time



Vaccine Acceptance by Occupation

In November 2021:

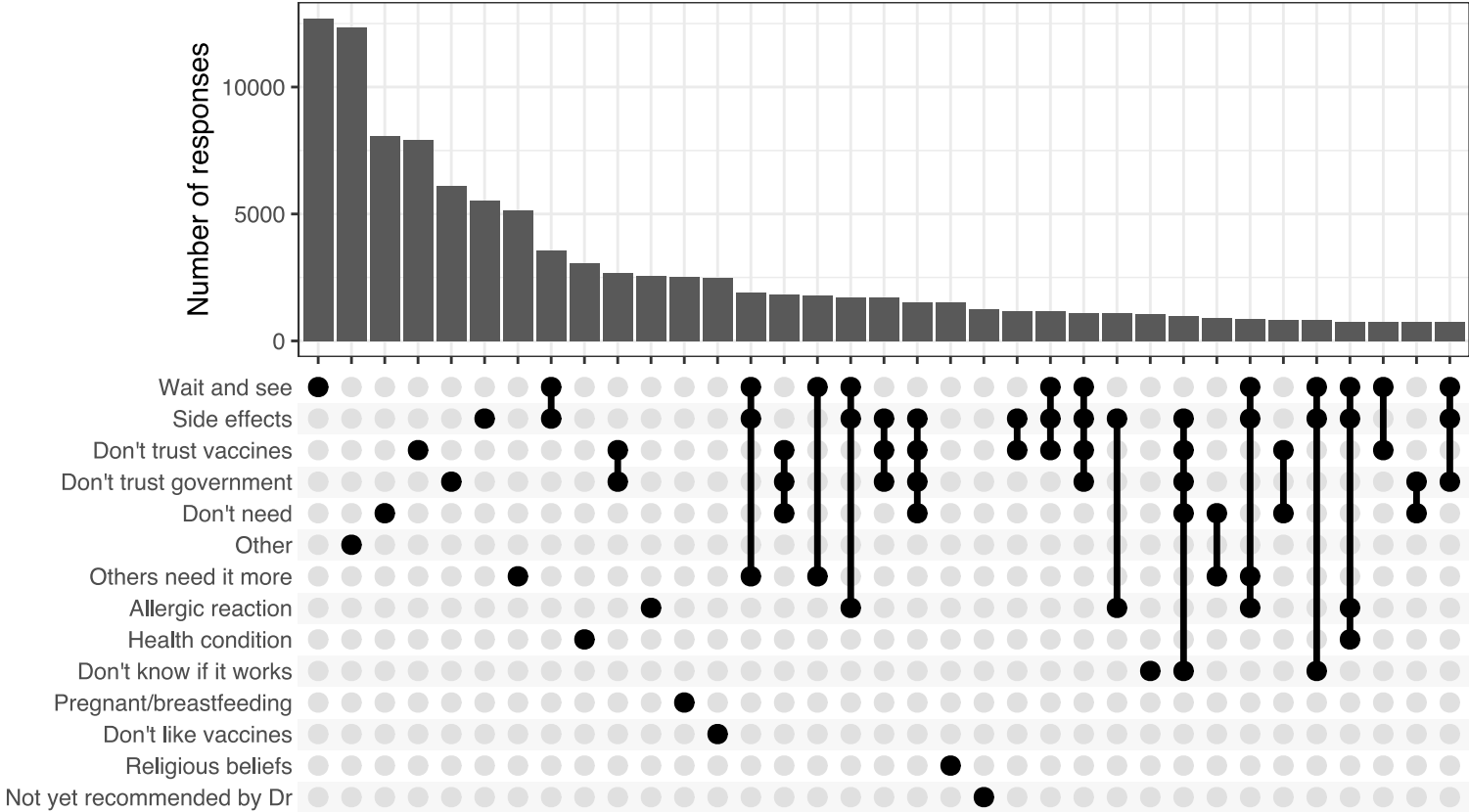
Occupation	% accepting	N
Education	92.4	29669
Arts	91.1	10794
SocialService	90.5	13923
HealthcareSupport	89.3	17475
HealthcarePractitioner	89.2	26303
Office	88.9	37820
PersonalCare	82.5	6545
Other	82.3	76254

Occupation	% accepting	N
FoodService	82.1	16410
Sales	80.1	25580
BuildingMaintenance	77.5	6144
Production	74.0	7597
ProtectiveService	73.1	3973
Transportation	71.2	12515
Maintenance	63.9	8651
Construction	58.1	5712

Hesitancy Reasons

Common vaccine hesitancy reasons

March 2021, N = 206,655



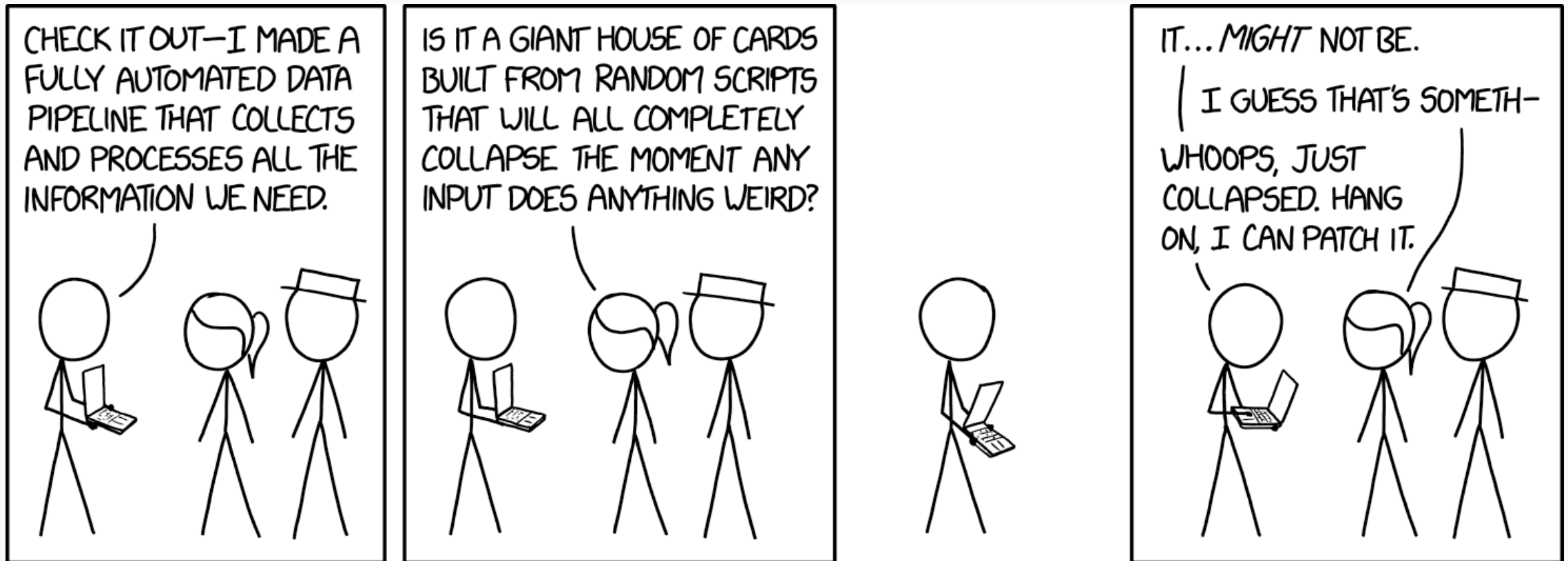
Delivering Results

- Academic publication does not meet the needs of policymakers—at least not for the kinds of localized and specific questions that our data can answer
- But direct engagement with individual agencies requires huge staffing
- ...and many policy questions have *no definite answers*
- Several approaches:
 - Facebook's policy team connects directly with the CDC, which already engages with numerous individual agencies
 - Publishing data directly, with documentation, allows other companies/academics to work with their local agencies
 - Additional staff/consulting helps us deliver data and insight more rapidly

Practicing Statistics in a Pandemic

Software Engineering

- "Statistical computing" usually means "learning R" or "learning MCMC and optimization"
- It almost **never** means "building automated systems with backups, logging, and alerting" or "deploying code automatically on merge" or...



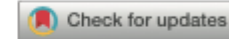
Software Engineering

- Delphi was fortunate to poach skilled technical staff to help manage these processes
- ...and to borrow 13 engineers, designers, and managers from Google.org
- Nonetheless, the survey pipeline was a challenge:
 - Just enough statistical detail to require careful review (weights, importance sampling)
 - Data volume meant that naive approaches could take more than 24 hours to calculate each day's aggregates
 - Survey is an airplane built in flight—so maintainability is key
 - Over 6,000 lines of code to build our data products
 - The naive way (giant R script full of dplyr operations) is a route to madness

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2021, VOL. 29, NO. 51, 57–515
<https://doi.org/10.1080/10691898.2020.1845109>



 OPEN ACCESS



Expanding the Scope of Statistical Computing: Training Statisticians to Be Software Engineers

Alex Reinhart  and Christopher R. Genovese

Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA

- Statisticians & data scientists are now often asked to make *products*, not analyses
- Working as a "data engineer" is much different than working on a data analysis report
- What we're doing isn't too different from what data scientists do in industry all day
- But software engineers are trained to do it and we aren't. Courses need:
 - Realistic long-term projects
 - Topics like unit testing, version control, and software design
 - Basics of algorithms and data structures

Wrapping Up

What a Weird Year

- Surveys on social media can be a surprisingly good tool to track a pandemic
- But delivering them presents numerous challenges:
 - How do you design a practical survey for such a wide audience?
 - How do you rapidly process the data?
 - How do you deliver useful insights fast enough for them to get used?
- Not every interesting statistical question is a *useful* statistical question
- We weren't trained for any of this, so everything had to happen on the fly

(I had never designed or run a survey before July 2020!)

Access to Survey Microdata

Want to study a problem that can be answered with 25 million US survey responses since April 2020? Possible topics:

- Reasons for vaccine hesitancy among specific demographic groups
- Symptoms reported by people testing positive, stratified by chronic conditions, age, etc.
- Test rates and availability by employment and occupation
- Mental health impacts of interventions
- Disparate impacts on minorities and disadvantaged groups
- ...anything you can think of

Raw response data is freely available to researchers who can sign DUAs to protect confidentiality of responses.

We're building a network of academic and non-profit researchers to learn from the survey.

<https://cmu-delphi.github.io/delphi-epidata/symptom-survey/>

Thank you

Thank you all for attending, and many thanks to

- the [entire Delphi team](#)
- CMU Legal, Sponsored Programs, Communications, IT, and numerous staff
- Facebook, Google, and Amazon Web Services
- Quidel
- Change Healthcare
- Qualtrics
- Centers for Disease Control and Prevention

Contact: <https://delphi.cmu.edu>, areinhar@stat.cmu.edu