# A critical evaluation of COVID-19 pandemic forecasts

Nicholas G. Reich

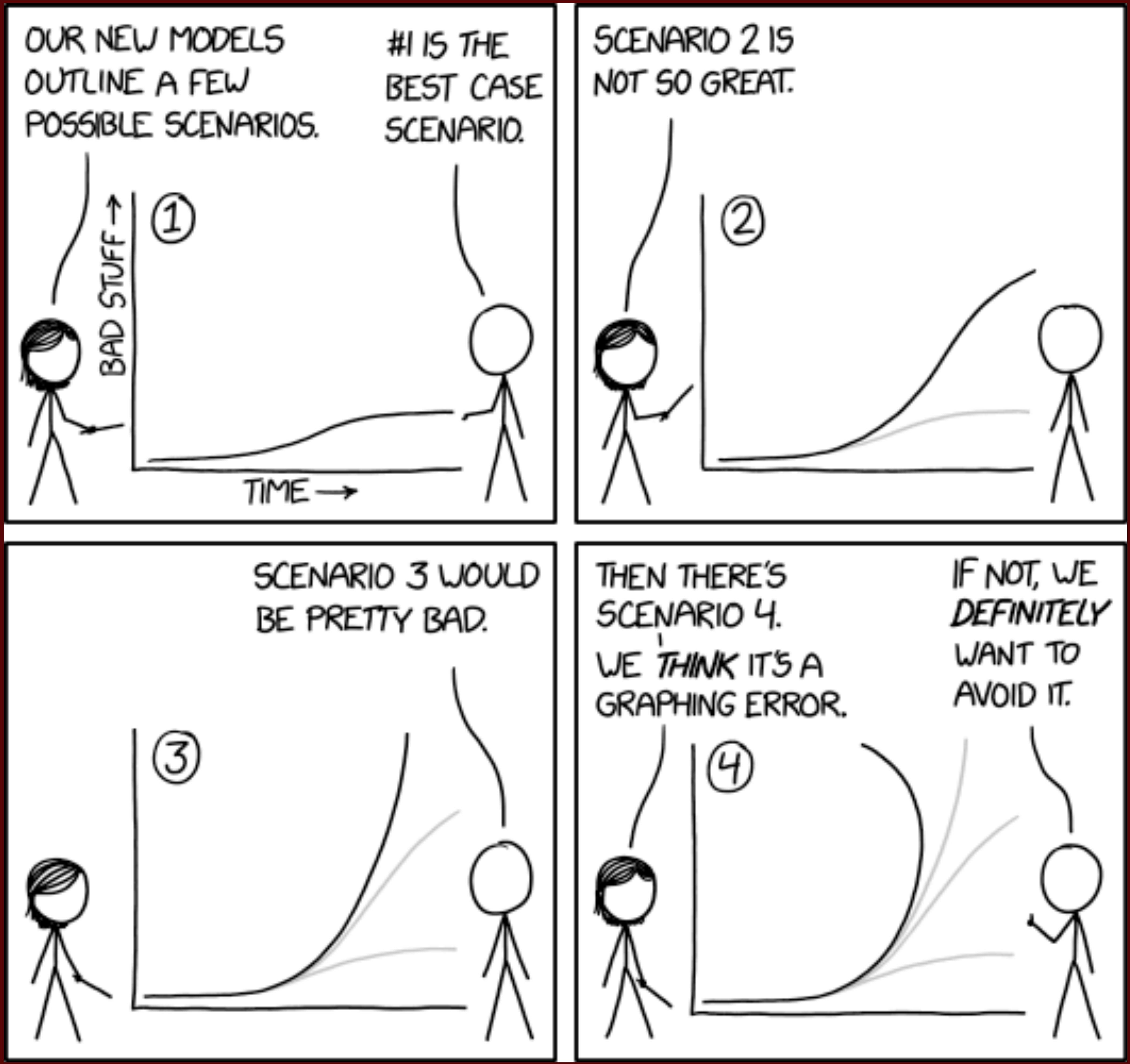
ACM SIGSPATIAL COVID-2020 Workshop
3 November 2020

reichlab.io
reichlab

Reich Lab | AT UMASS AMHERST

These slides are available for download at: https://covid19forecasthub.org/doc/talks/

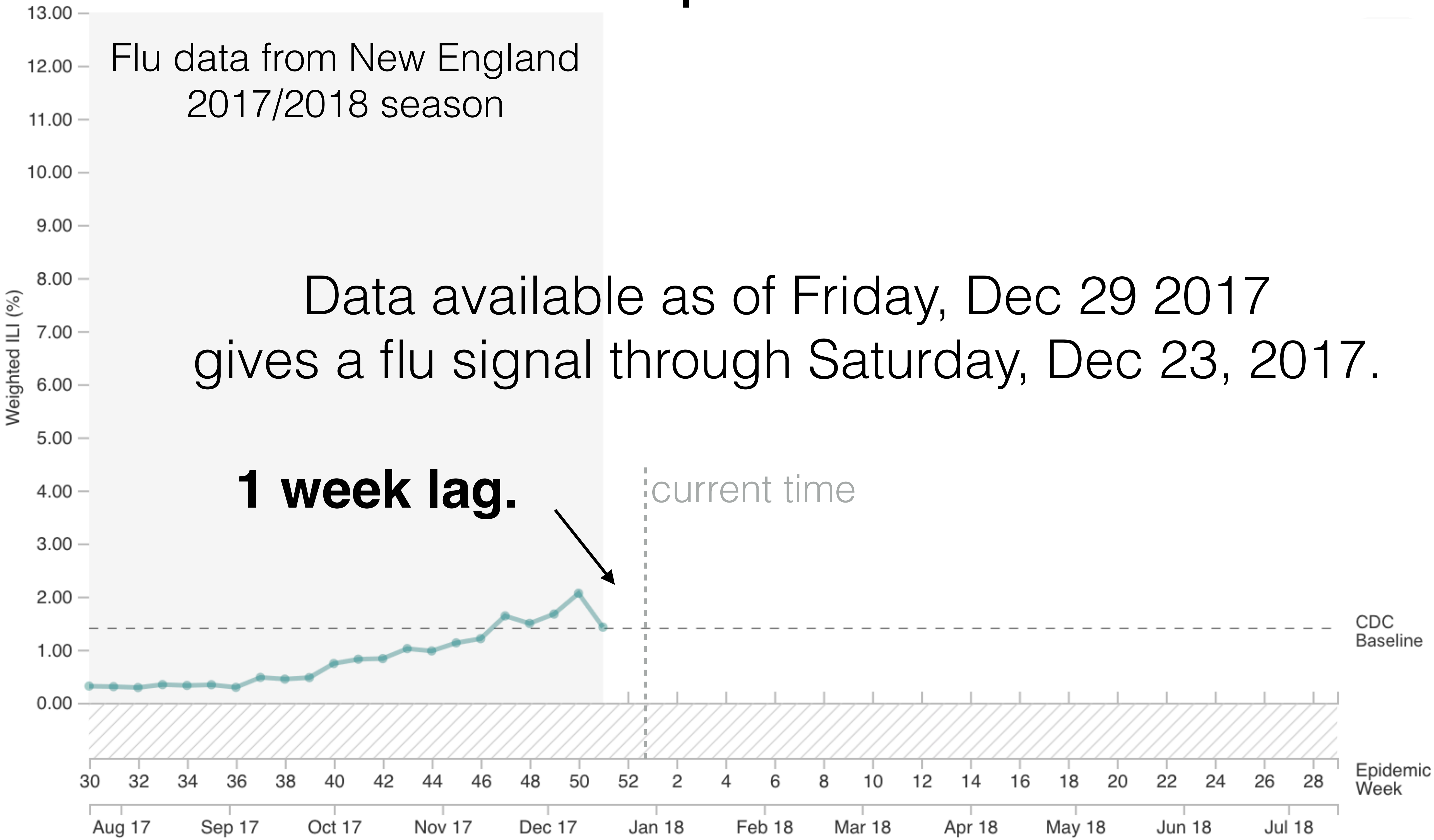
This work has been supported by the National Institutes of General Medical Sciences (R35GM119582) and the Centers for Disease Control and Prevention (1U01IP001122). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS, the National Institutes of Health, or CDC.



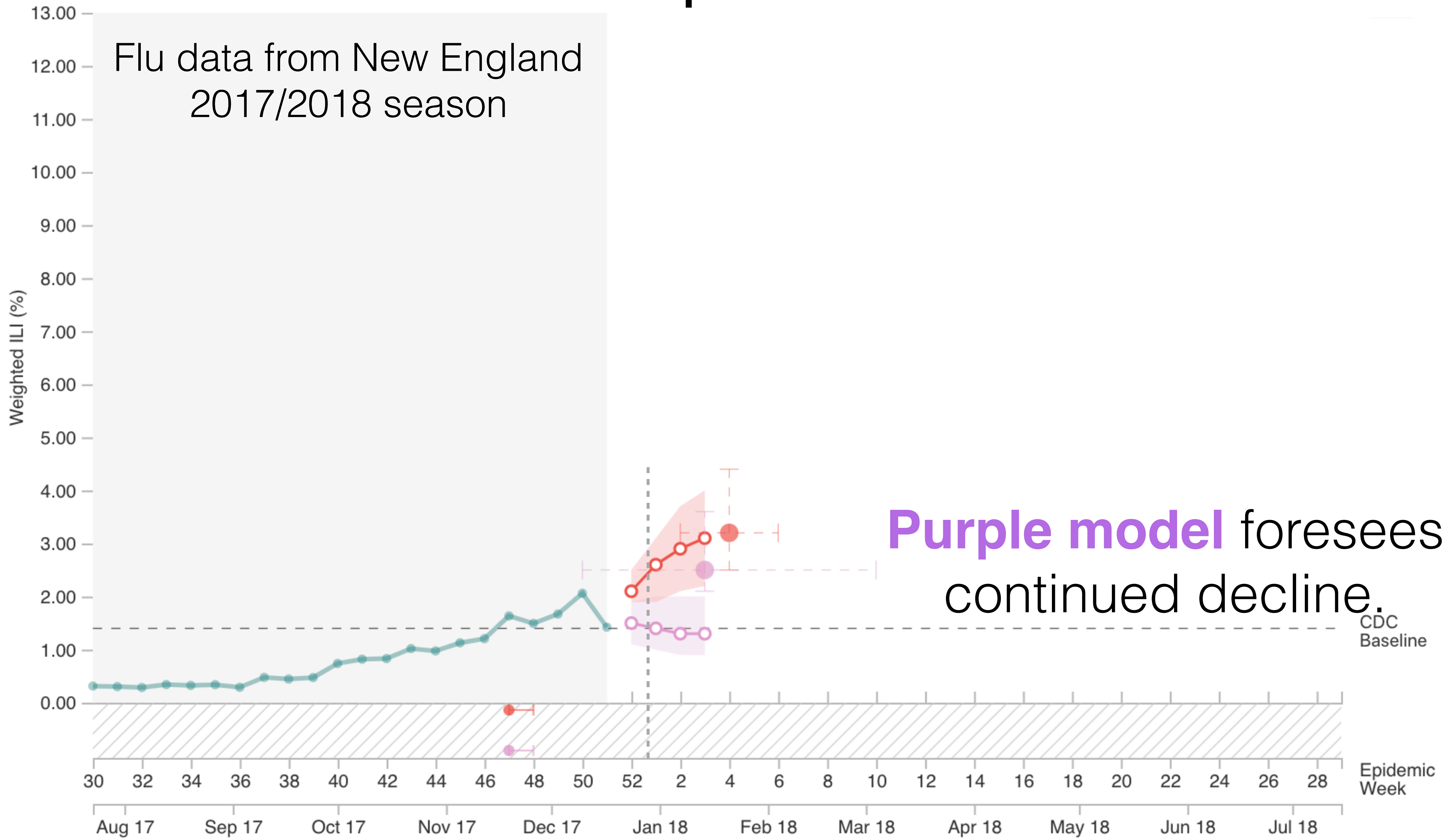

https://xkcd.com/2289/

# Why model?

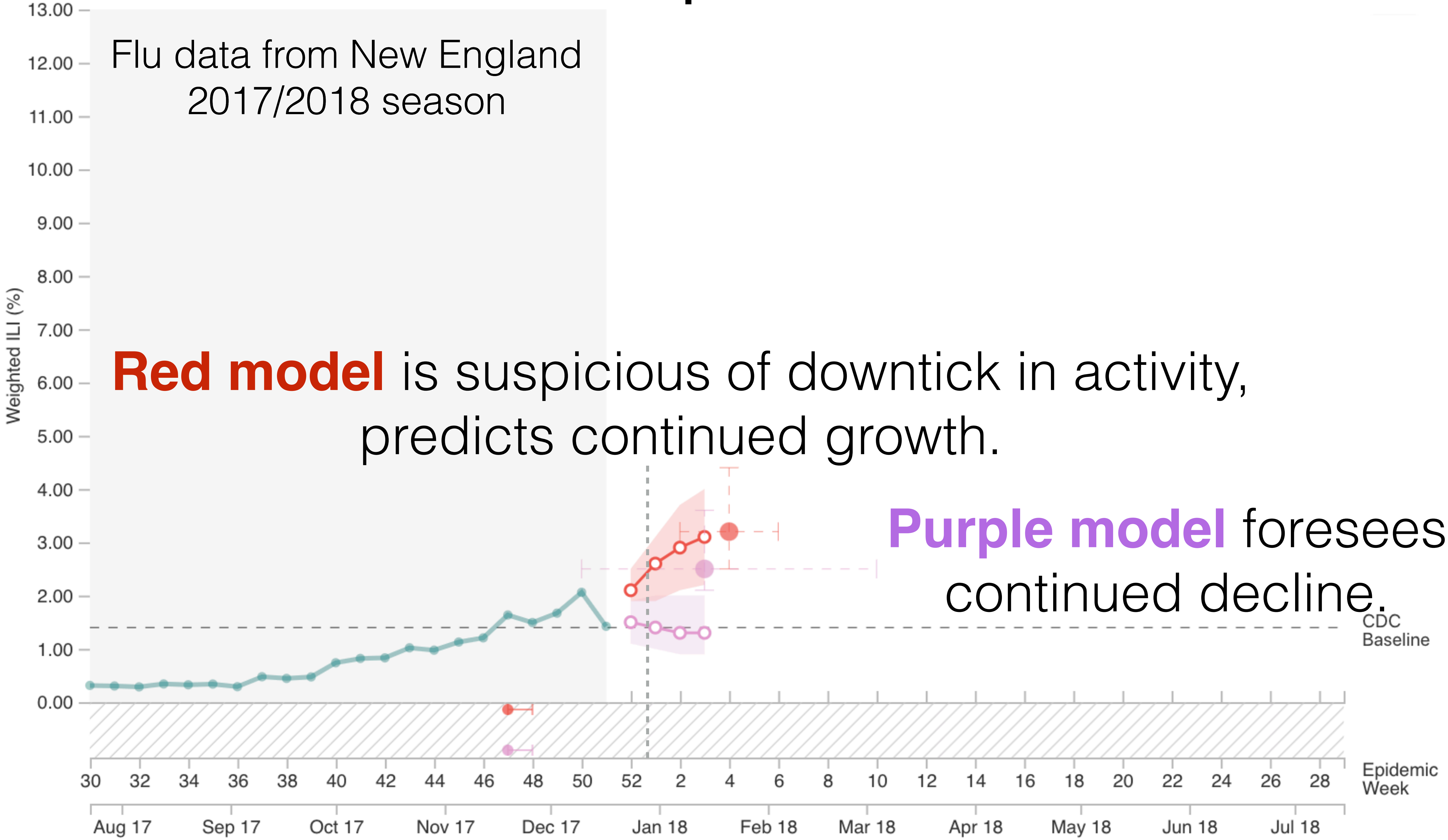
# Real-time public health data is imperfect



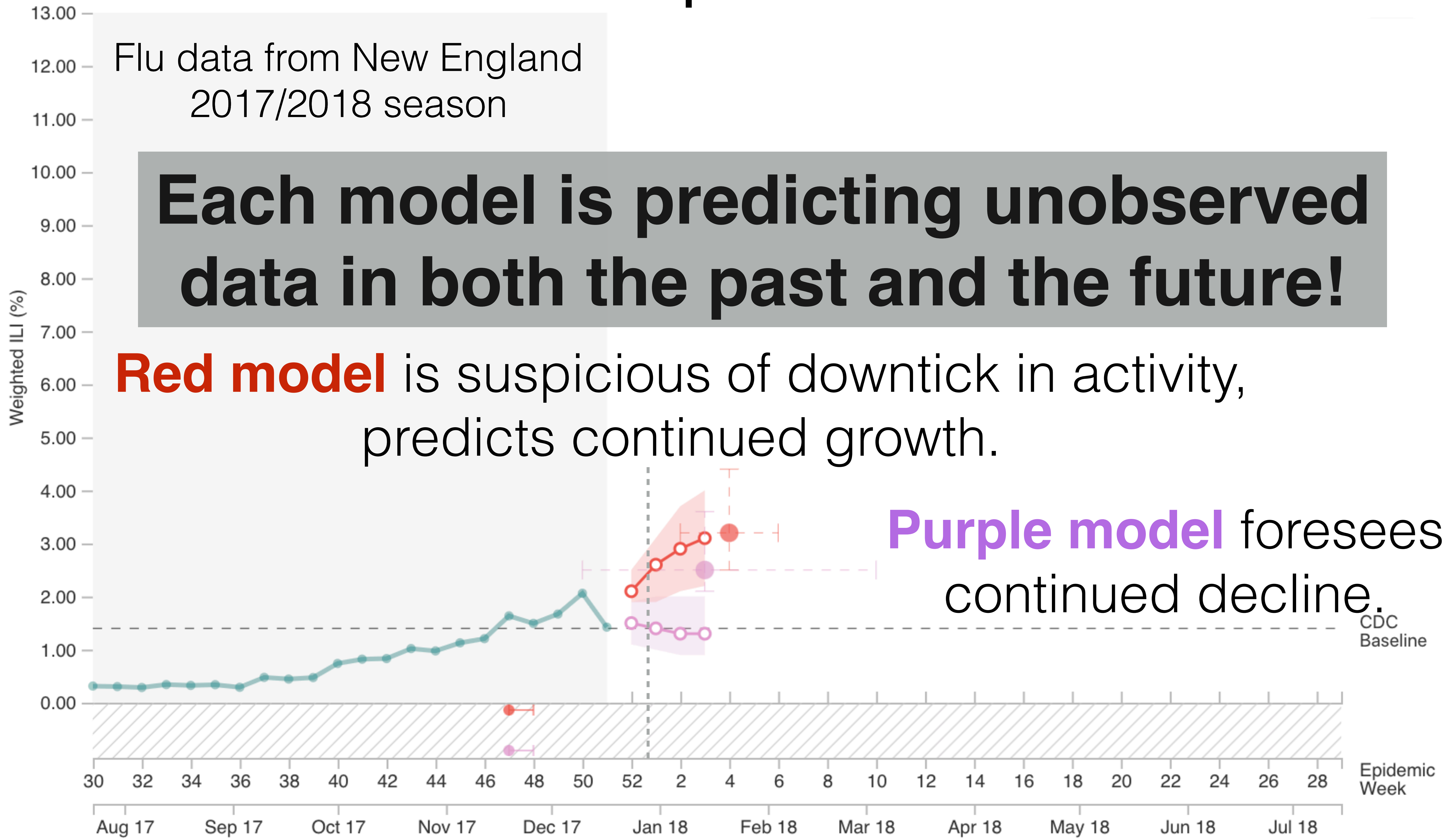
# Real-time public health data is imperfect



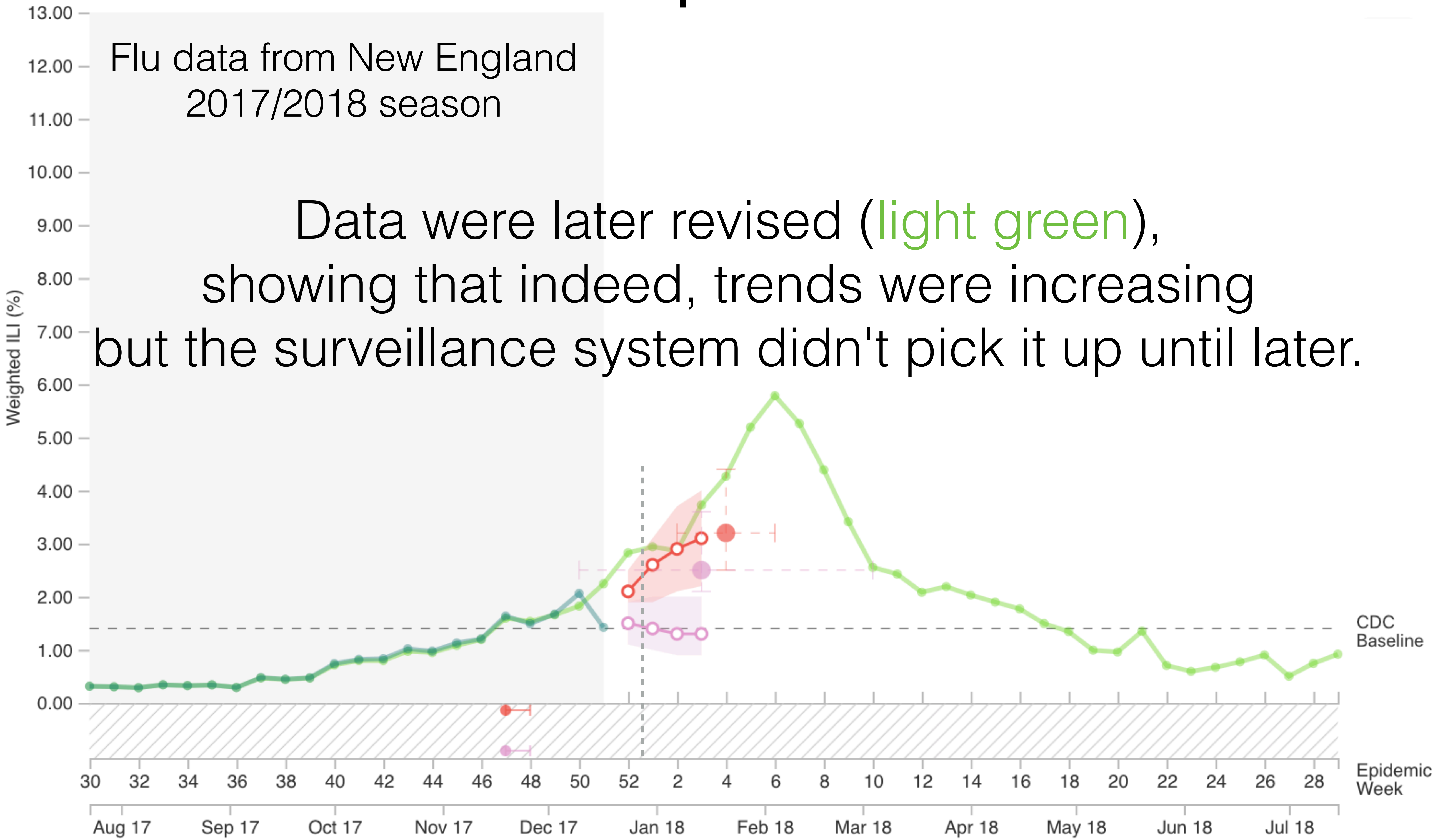
# Real-time public health data is imperfect



# Real-time public health data is imperfect



# Real-time public health data is imperfect



# Good models might...

- Anticipate and adjust for data quality issues.
- Infer what is happening right now.
- Forecast what will be observed in the near future.
- Project hypothetical outcomes in the distant future.

# Good models might…

- Anticipate and adjust for data quality issues.
- Infer what is happening right now.
- Forecast what will be observed in the near future.
- Project hypothetical outcomes in the distant future.

**Don't expect a single model to do all of these things well!**

# COVID-19 example

## California COVID Assessment Tool

https://calcat.covid19.ca.gov/cacovidmodels/

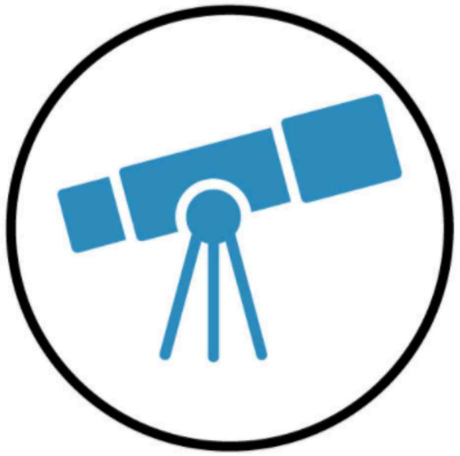
### Nowcasts

How fast is COVID-19 spreading right now?

### Forecasts

What can we expect in the next 2-4 weeks?

### Scenarios

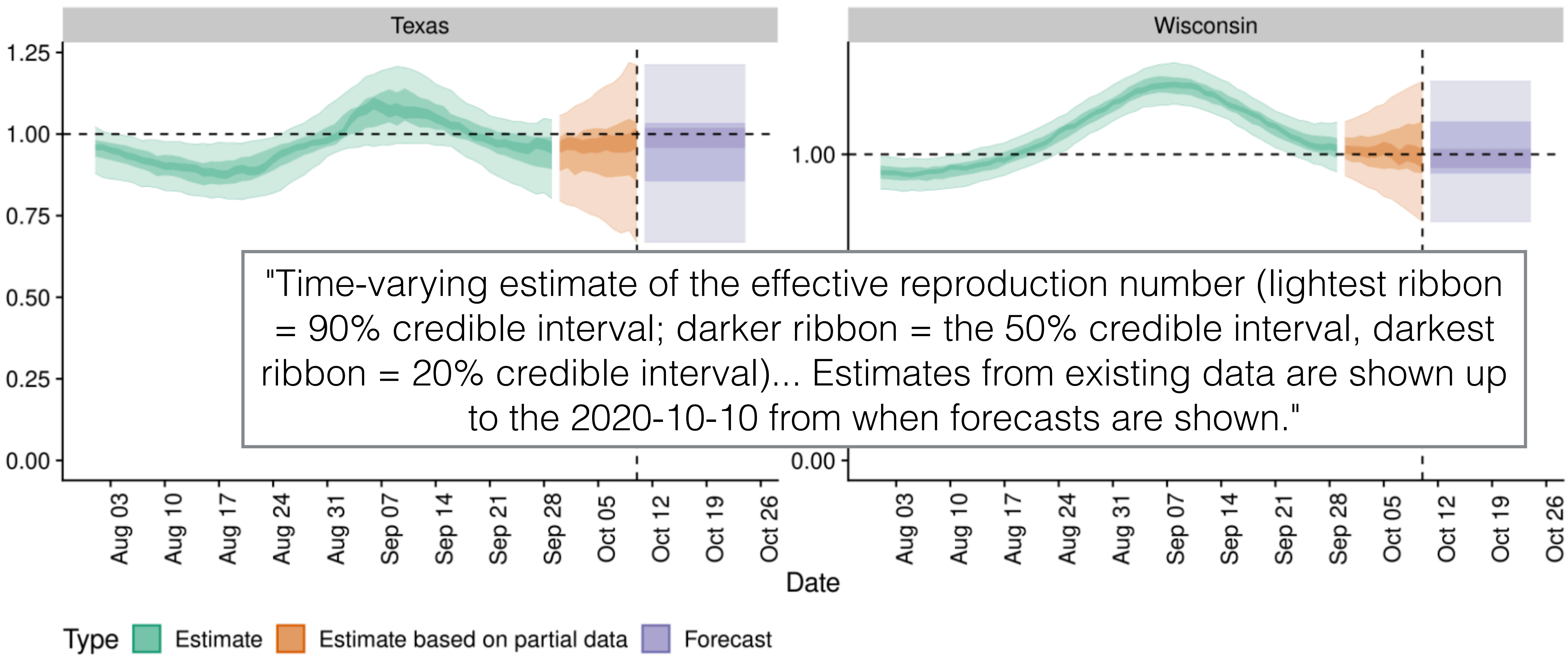


What are the long-term impacts under different scenarios?

How fast is COVID-19 spreading right now?

# Nowcasting

Not as agreed upon definition, but I'd vote for "building a model that draws inference about trends the recent past."



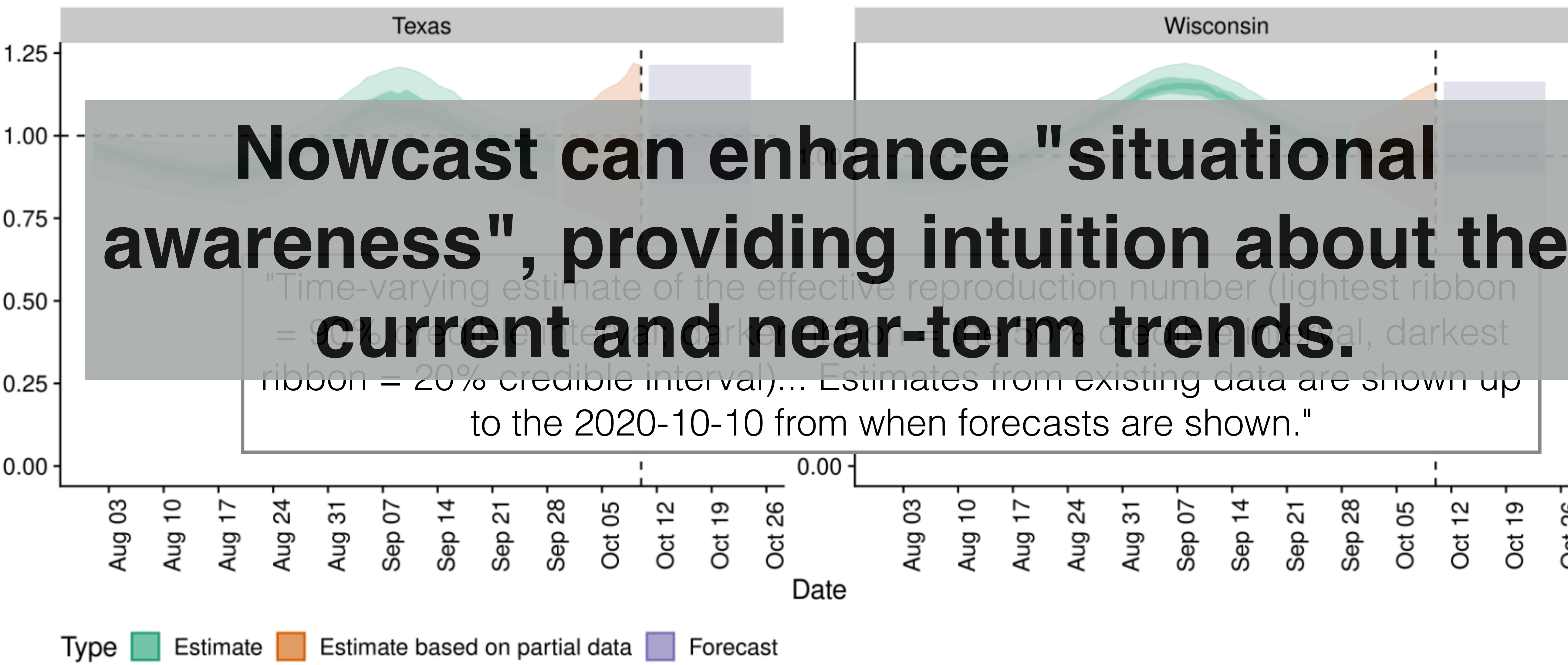

"Time-varying estimate of the effective reproduction number (lightest ribbon = 90% credible interval; darker ribbon = the 50% credible interval, darkest ribbon = 20% credible interval)... Estimates from existing data are shown up to the 2020-10-10 from when forecasts are shown."

https://epiforecasts.io/covid/posts/national/united-states/

How fast is COVID-19 spreading right now?

# Nowcasting

Not as agreed upon definition, but I'd vote for "building a model that draws inference about trends the recent past."




**Nowcast can enhance "situational awareness", providing intuition about the current and near-term trends.**

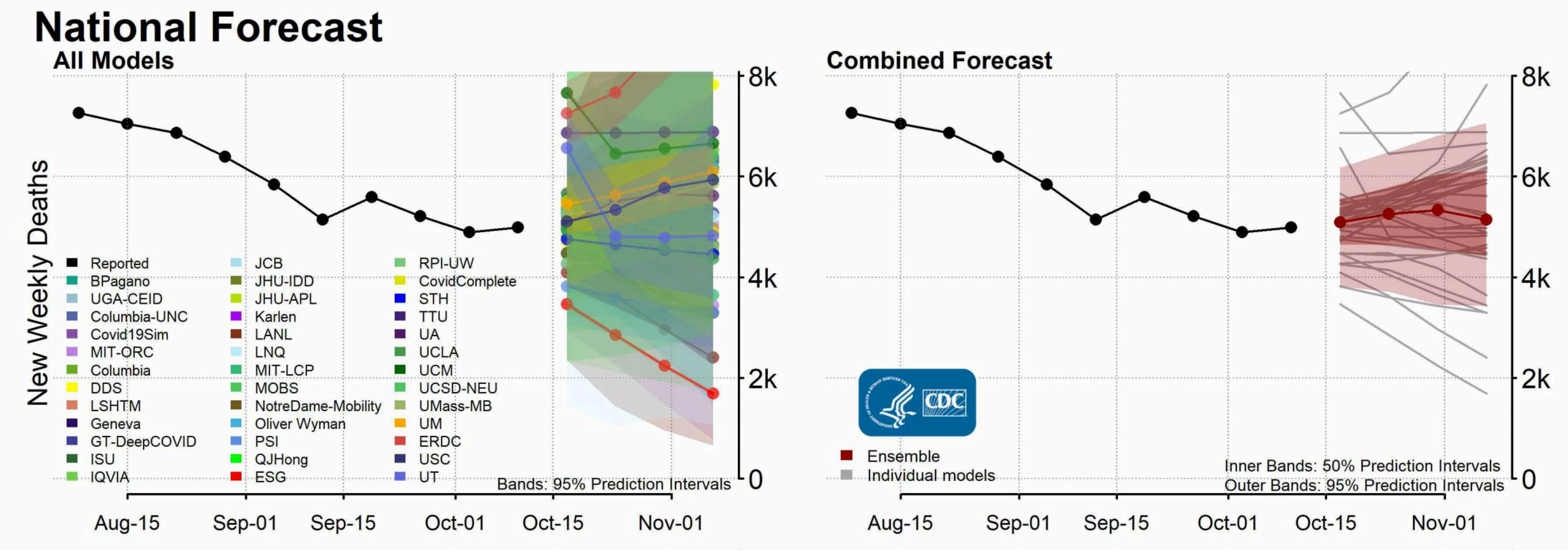
"Time-varying estimate of the effective reproduction number (lightest ribbon = 90% credible interval; darker ribbon = the 50% credible interval, darkest ribbon = 20% credible interval)... Estimates from existing data are shown up to the 2020-10-10 from when forecasts are shown."

https://epiforecasts.io/covid/posts/national/united-states/

# Short-term Forecasting

What can we expect in the next 2-4 weeks?

Making **falsifiable, evaluable** predictions of observable future quantities.

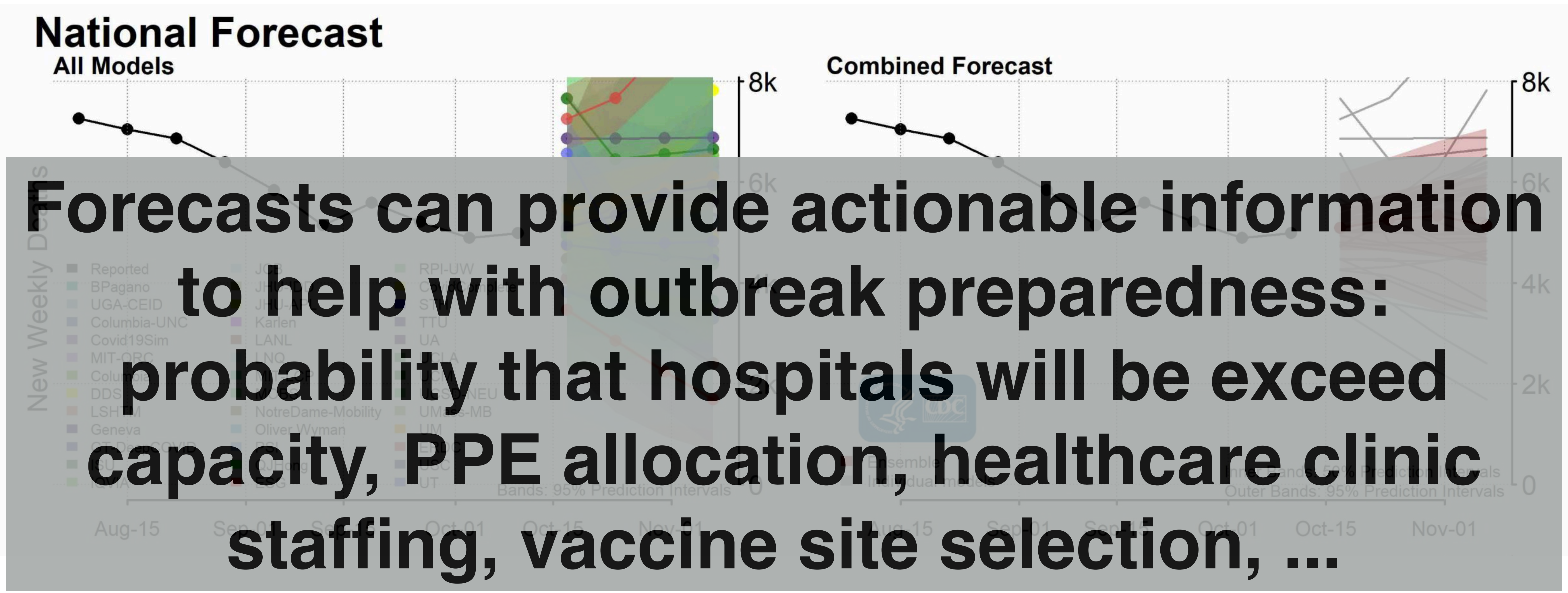



https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

# Short-term Forecasting

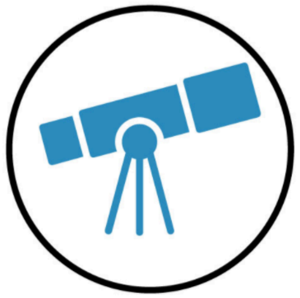
What can we expect in the next 2-4 weeks?

Making **falsifiable, evaluable** predictions of observable future quantities.




**Forecasts can provide actionable information to help with outbreak preparedness: probability that hospitals will be exceed capacity, PPE allocation, healthcare clinic staffing, vaccine site selection, ...**

https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html



What are the long-term impacts under different scenarios?

# Long-term Scenarios

Projections based on specific assumptions.

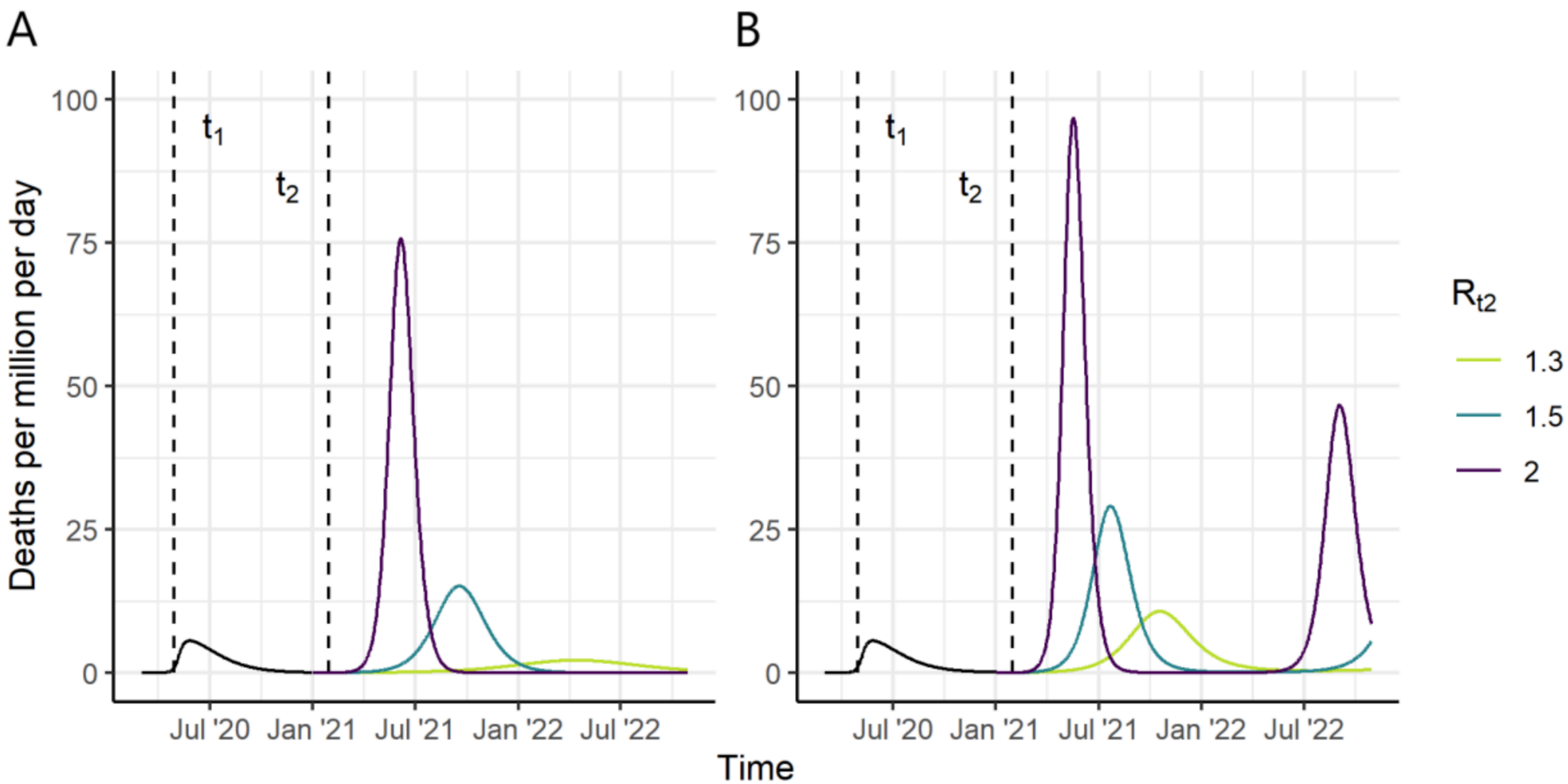


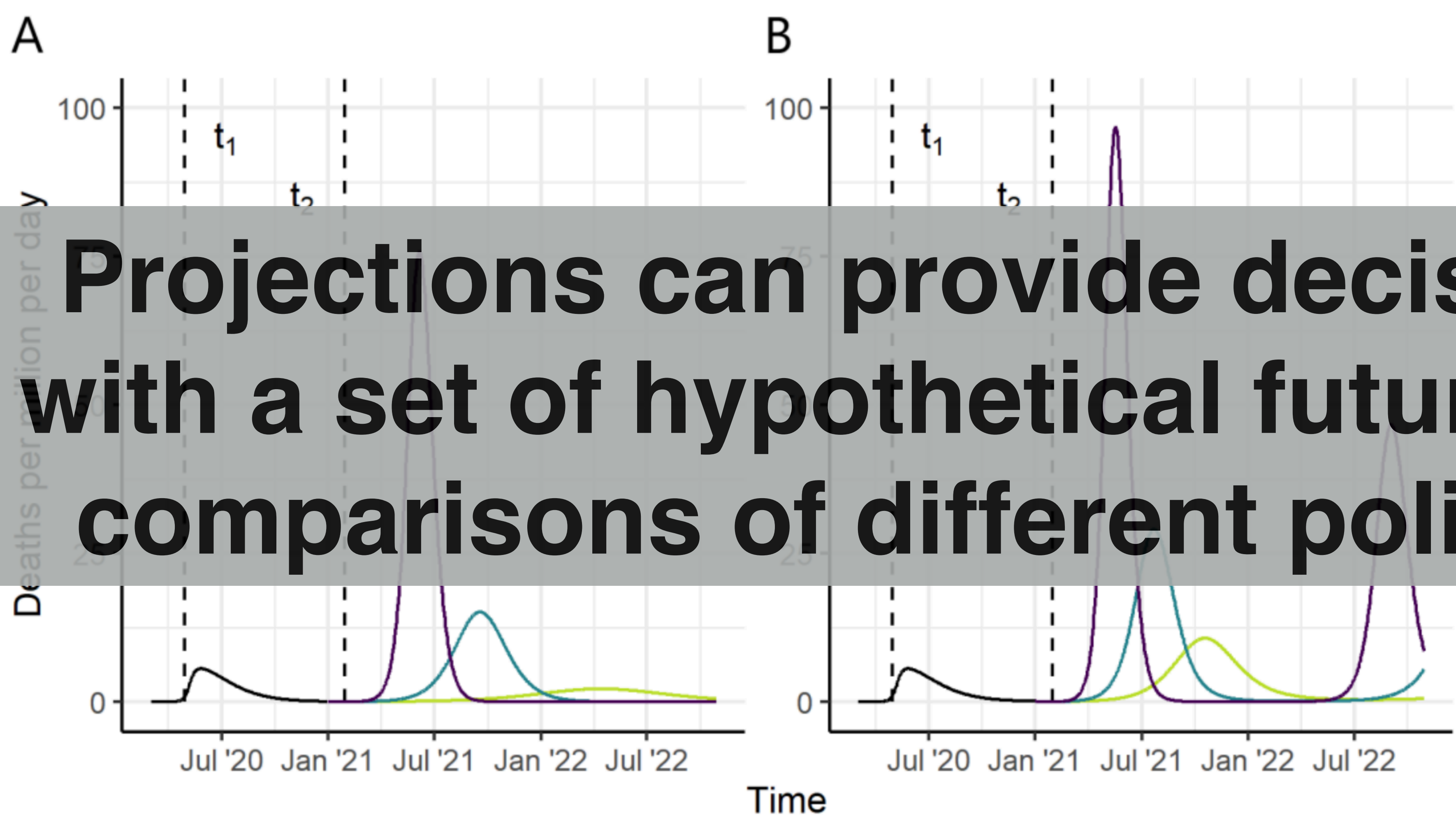

**Figure 1: Scenarios for the Course of the Epidemic from 2020–2022, for a High-Income Country Setting, in the Absence of a Vaccine (counterfactual scenarios).** (A) Assuming "long immunity" and (B) assuming an average duration of naturally acquired immunity of 1 year. We assume that $R_0$=2.5 up to time $t_1$ (May 2020) and that $R_{t1}$

https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-09-25-COVID19-Report-33.pdf

# Long-term Scenarios

What are the long-term impacts under different scenarios?

## Projections based on specific assumptions.




**Projections can provide decision-makers with a set of hypothetical futures based on comparisons of different policy choices.**

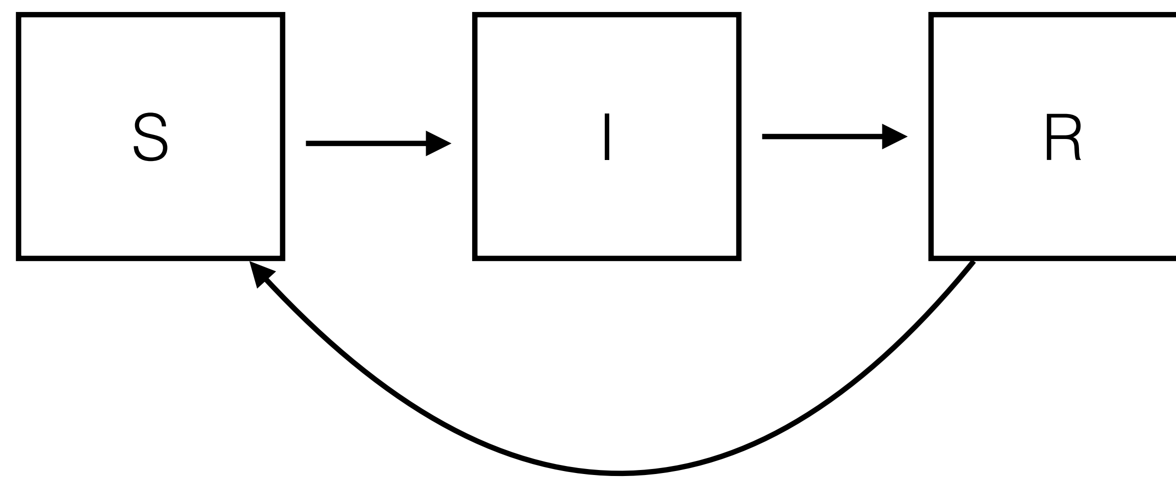
**Figure 1: Scenarios for the Course of the Epidemic from 2020–2022, for a High-Income Country Setting, in the Absence of a Vaccine (counterfactual scenarios).** (A) Assuming "long immunity" and (B) assuming an average duration of naturally acquired immunity of 1 year. We assume that $R_0$=2.5 up to time $t_1$ (May 2020) and that $R_{t1}$

https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-09-25-COVID19-Report-33.pdf

# A Brief History of Epidemic Modeling

# Modeling disease transmission

Susceptible-Infectious-Recovered (SIR) epidemiological models encode a mechanistic understanding of the biological transmission of disease.


Model theory has been developed for over 100 years.

Google: "epidemic theory youtube" or https://www.youtube.com/watch?v=CmhL4rVLwn0

“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, ... for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.”

– Pierre-Simon LaPlace (1825)

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, ... for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

– Pierre-Simon LaPlace (1825)

**i.e. knowledge of the "mechanism" is crucial**

“The world is woven from billions of lives, every strand crossing every other. What we call premonition is just movement of the web. <u>If you could attenuate to every strand of quivering data the future would be entirely calculable.</u> As inevitable as mathematics.”

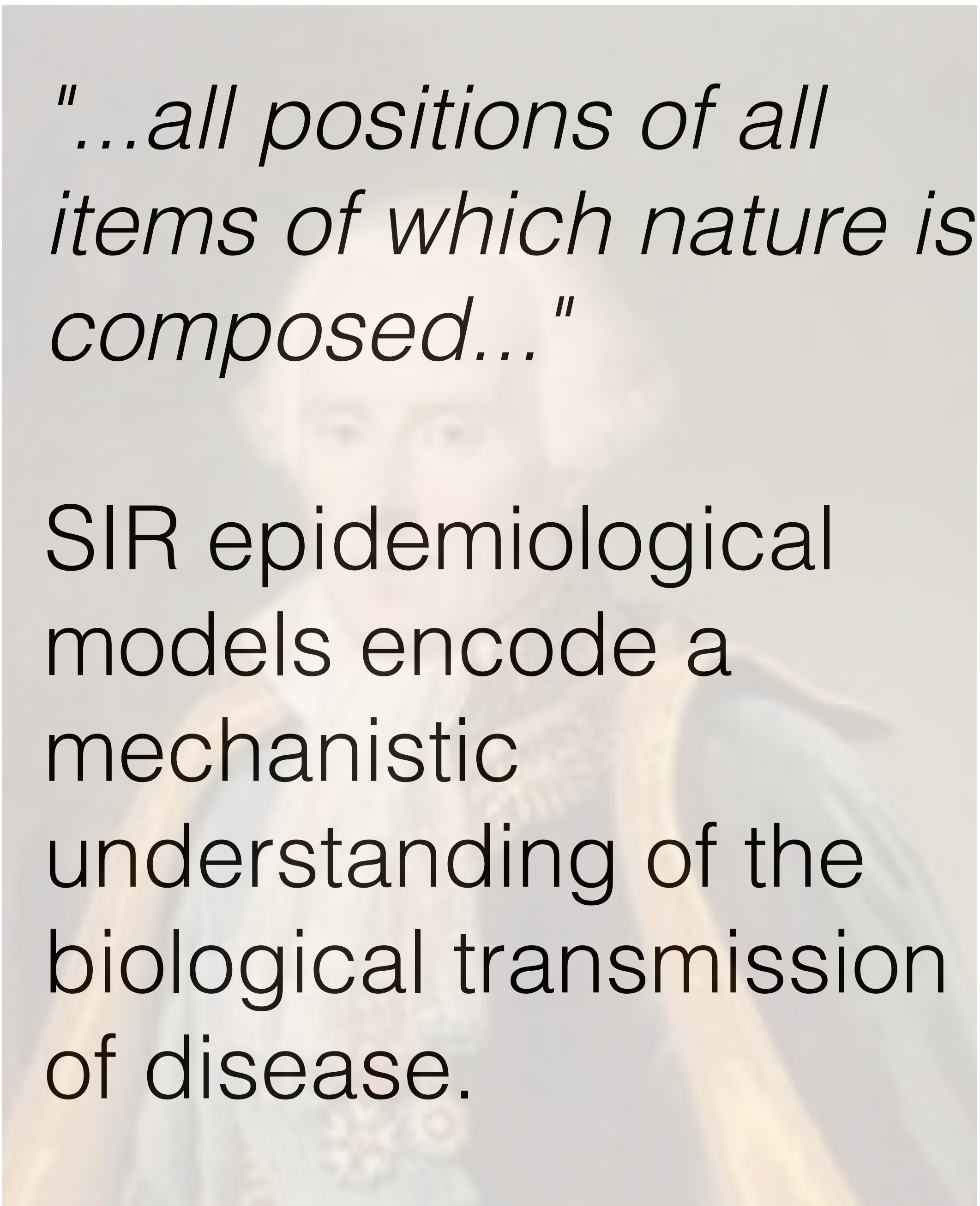
– Sherlock (2017)

“The world is woven from billions of lives, every strand crossing every other. What we call premonition is just movement of the web. If you could attenuate to every strand of quivering data the future would be entirely calculable. As inevitable as mathematics.”

– Sherlock (2017)

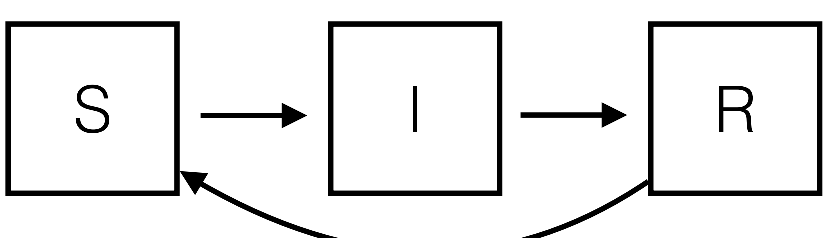

**i.e. data is all you need**

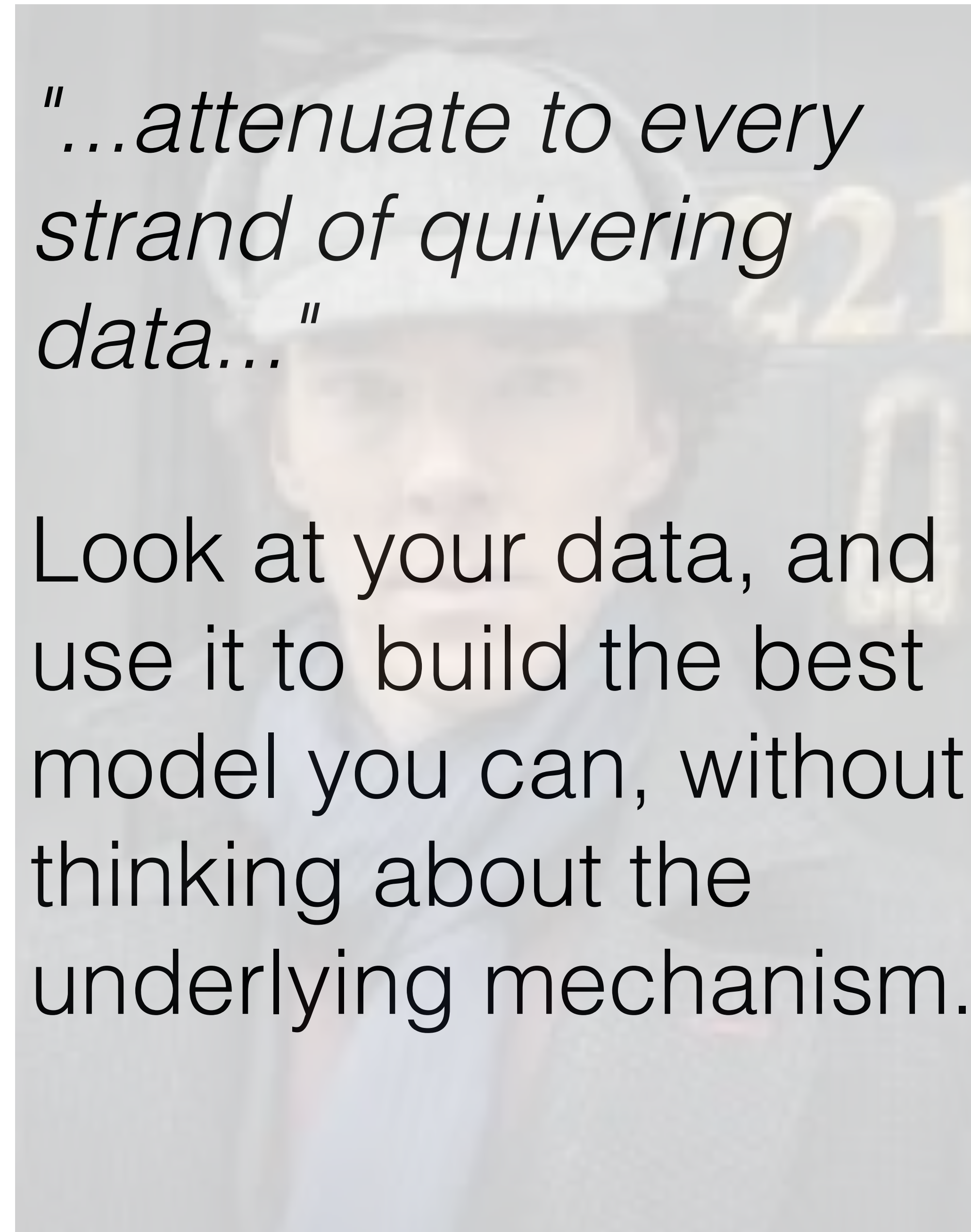
# Model Taxonomy: Mechanism vs. Phenomenon



*"...all positions of all items of which nature is composed..."*

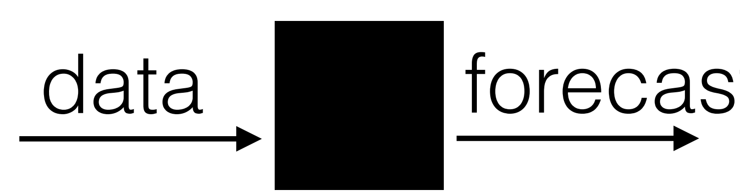
SIR epidemiological models encode a mechanistic understanding of the biological transmission of disease.



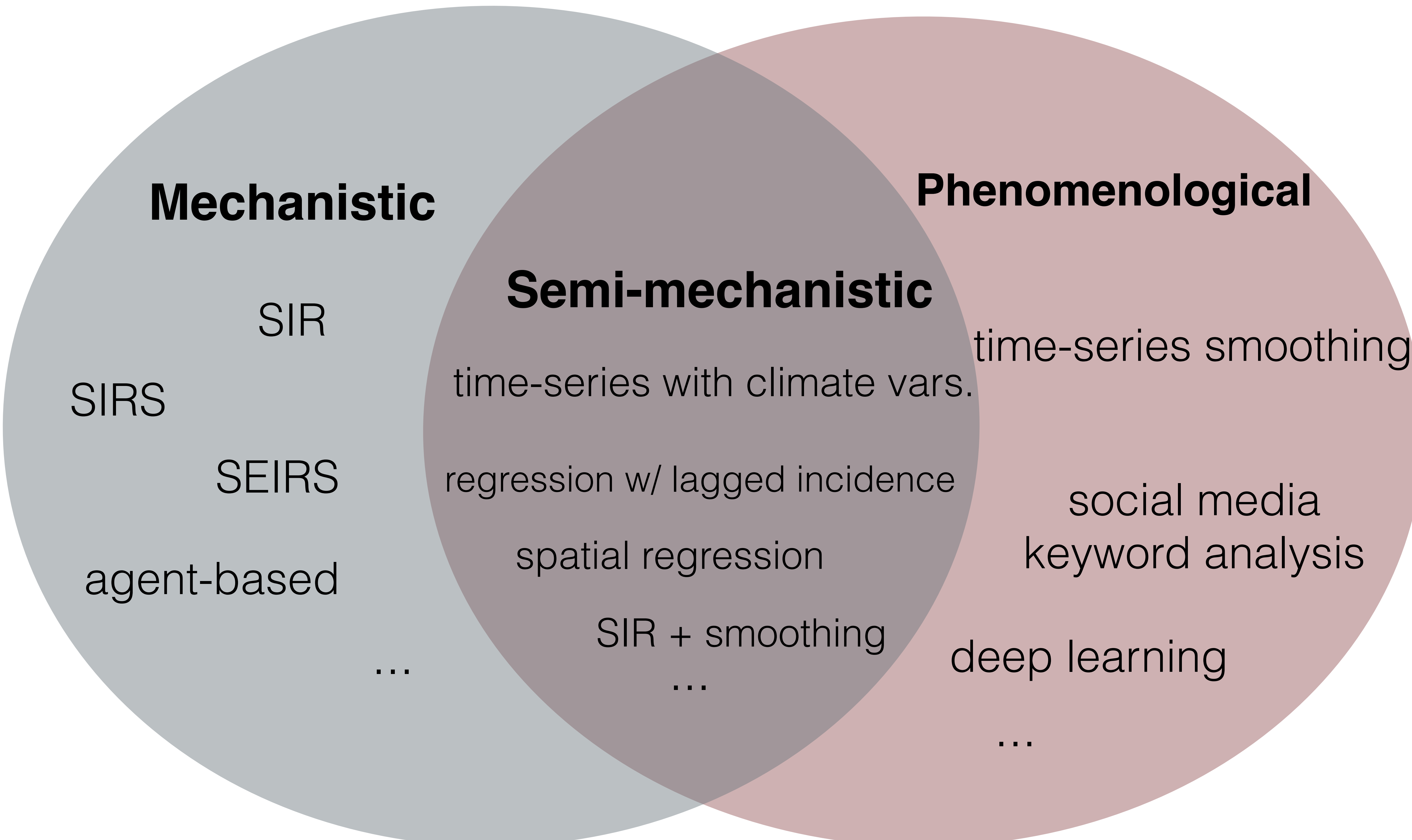



*"...attenuate to every strand of quivering data..."*

Look at your data, and use it to build the best model you can, without thinking about the underlying mechanism.



# Infectious Disease Model Taxonomy

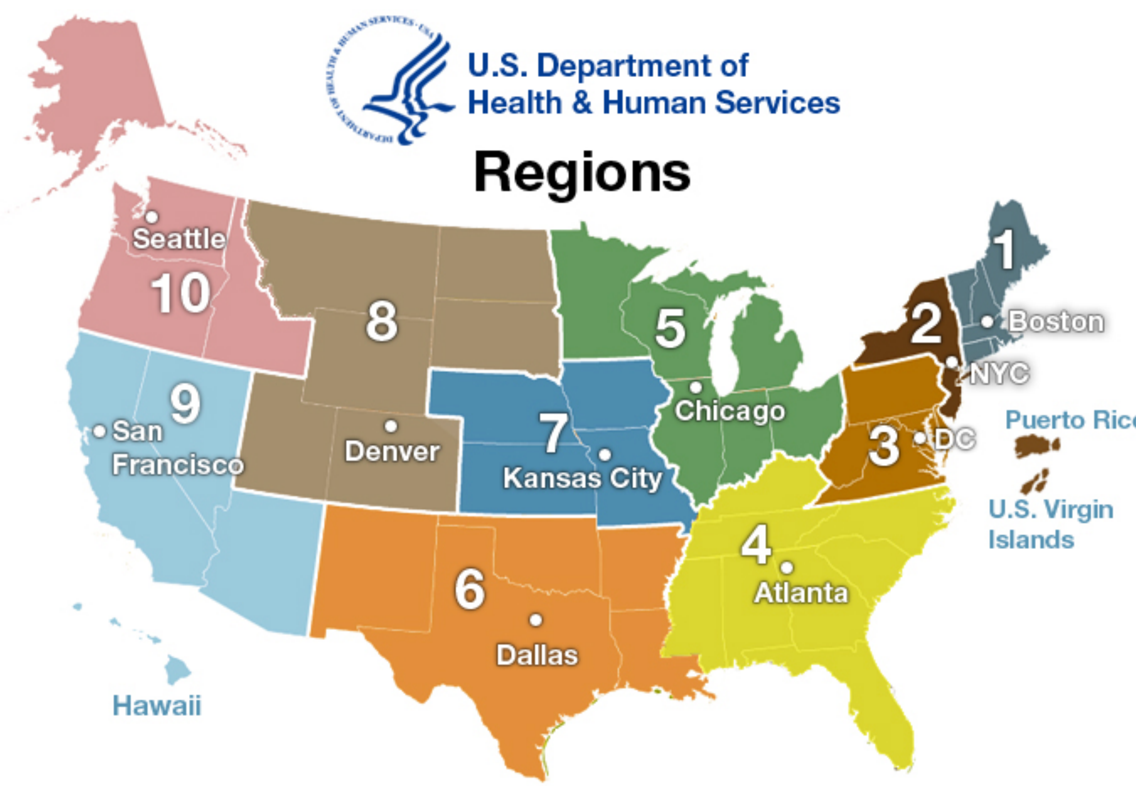


# Lessons from flu forecasting

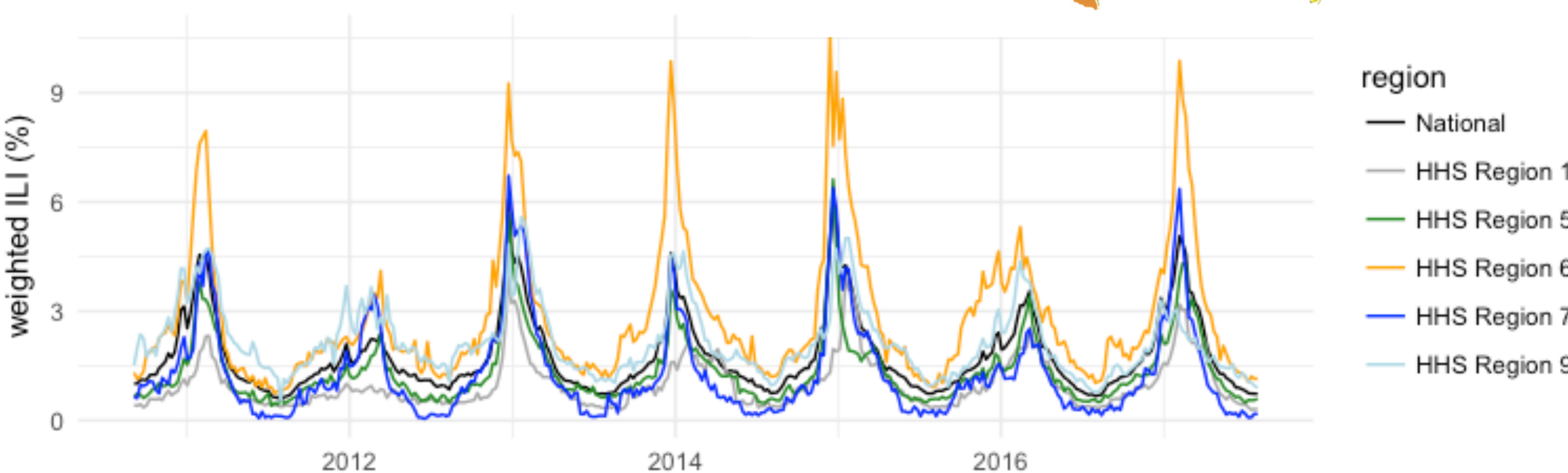

Reich et al. 2019, *PNAS*. https://doi.org/10.1073/pnas.1812594116
Reich et al. 2019, *PLOS Comp Bio*. https://doi.org/10.1371/journal.pcbi.1007486
McGowan et al. 2019, *Sci Rep*. https://doi.org/10.1038/s41598-018-36361-9

# Forecasting Seasonal Flu

CDC FluSight challenges: U.S. national, regional, state forecasts

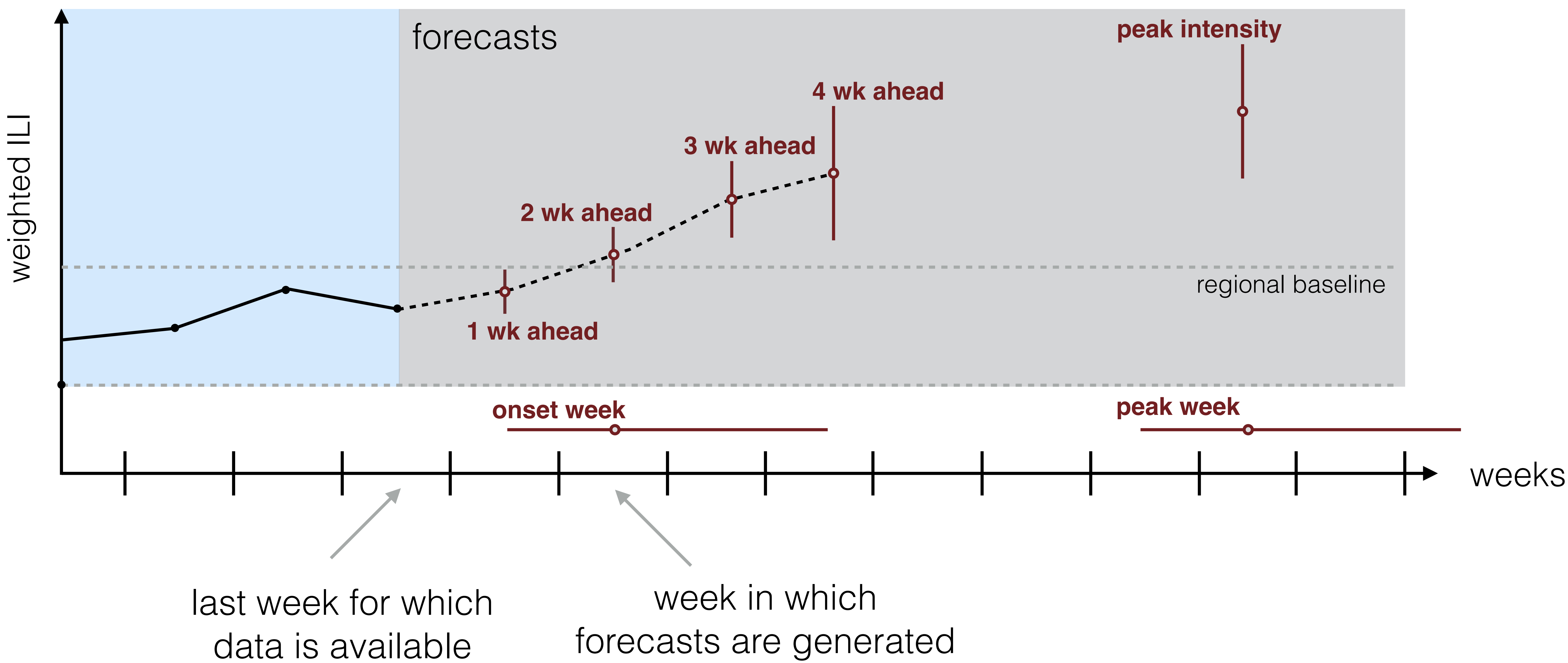



**Target variable "weighted ILI":**
The % of all outpatient visits with primary complaint of influenza-like illness (ILI), weighted by state population.

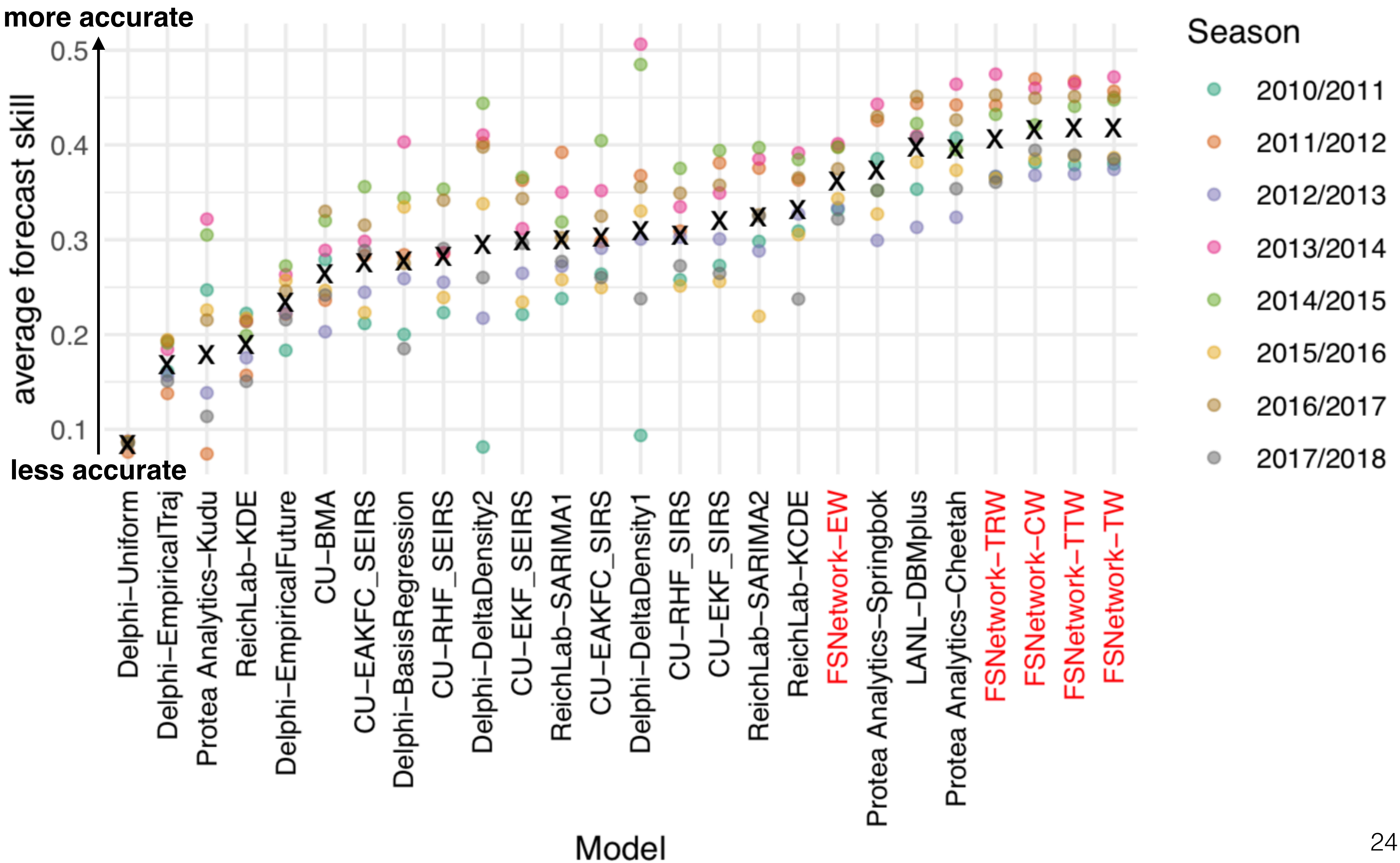


# Targets with Public Health Relevance

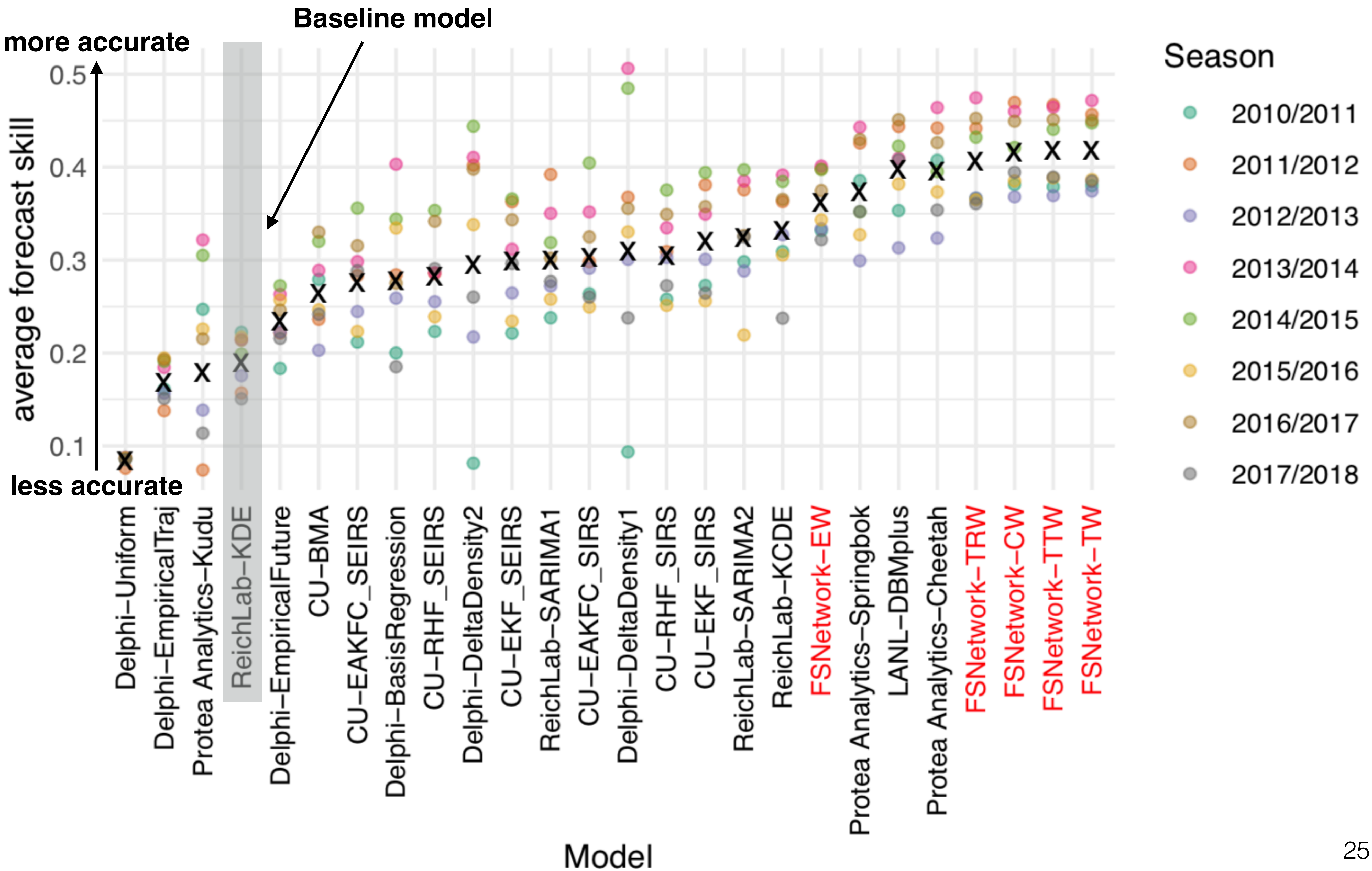
## based on annual CDC FluSight forecasting challenge



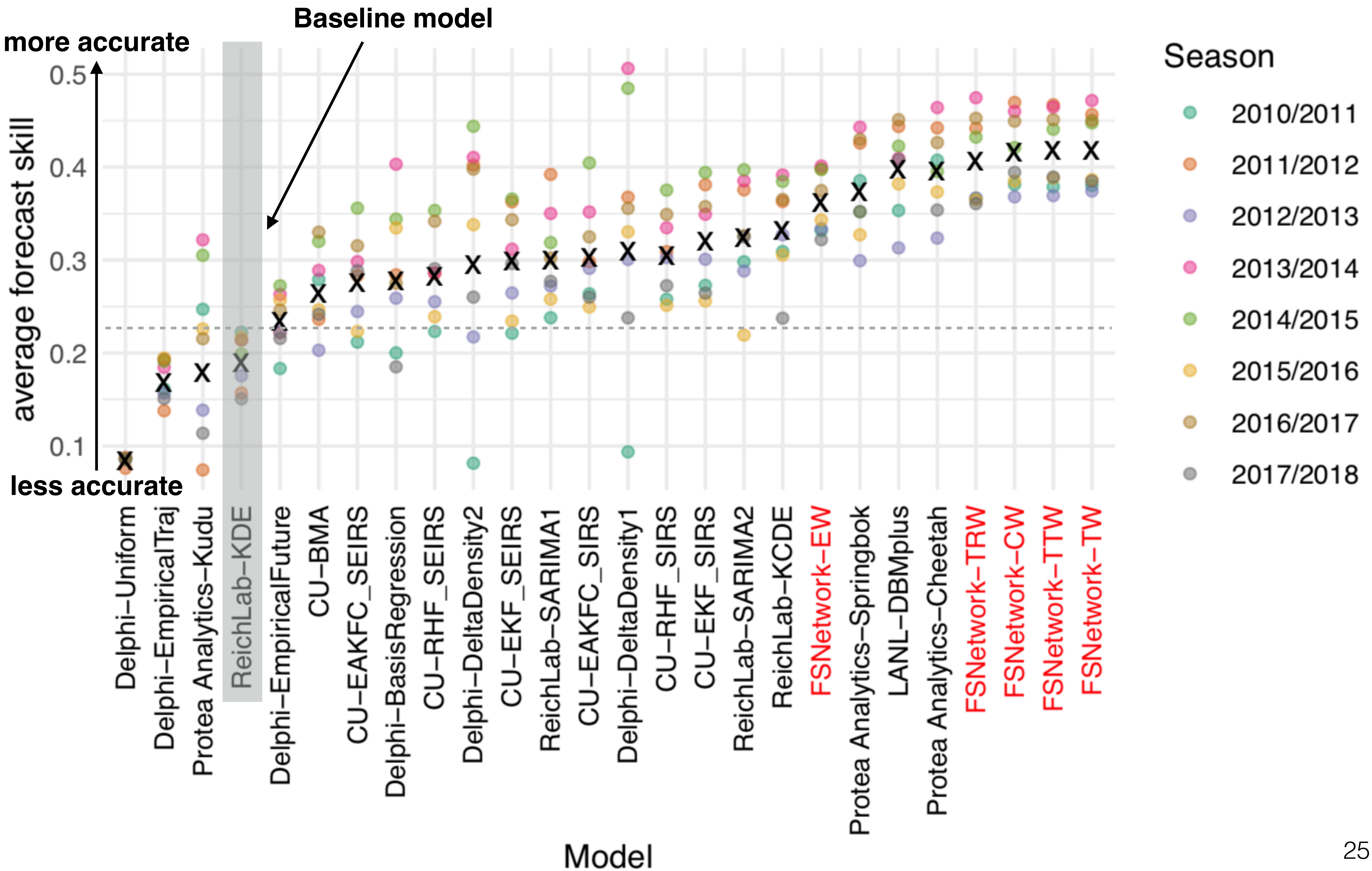
# Evaluating seasonal flu forecasts



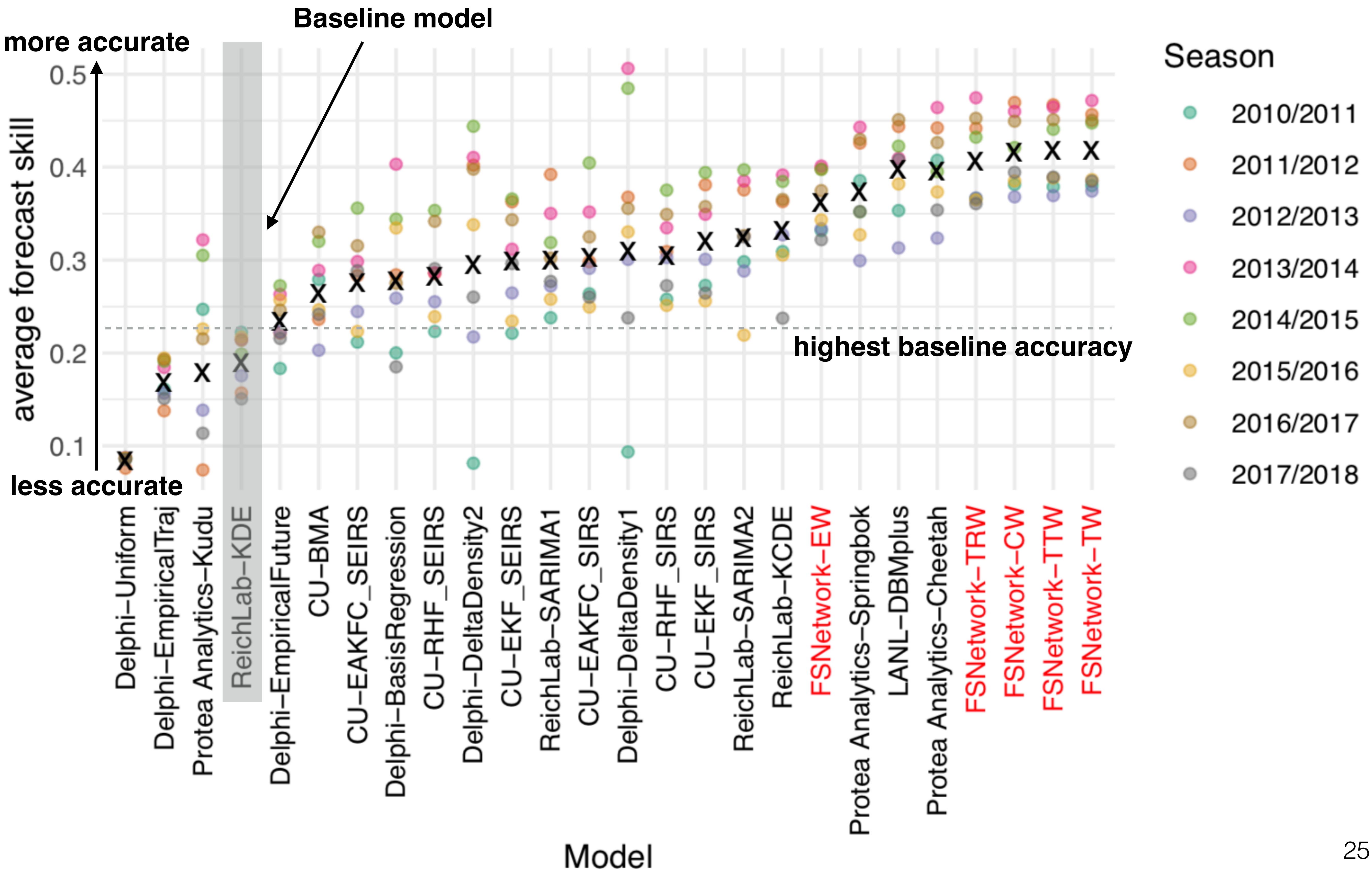
# Many models outperform baseline



# Many models outperform baseline

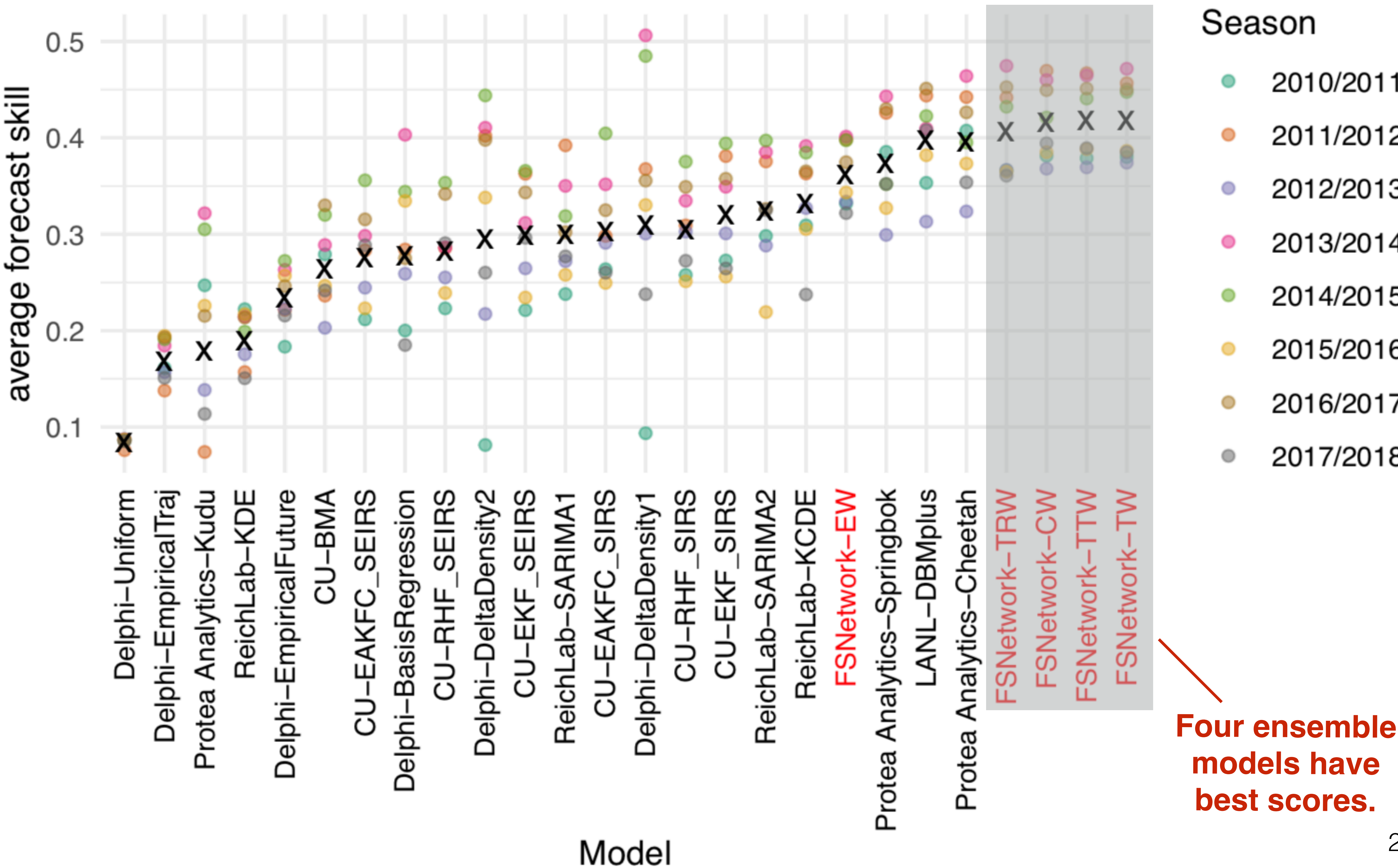


# Many models outperform baseline



# Ensembles have best scores

Models with “FSNetwork” prefix are different versions of the ensemble models.




**Four ensemble models have best scores.**

# Forecasting COVID-19

Ray et al, 2020, *medrxiv*. https://doi.org/10.1101/2020.08.19.20177493
Bracher et al, 2020, arxiv. https://arxiv.org/abs/2005.12881
Brooks, Ray et al, 2020, IIF blog.
https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/

COVID-19 ForecastHub

https://covid19forecasthub.org/

**Team:** Martha Zorn, Nutcha Wattanachit, Serena Wang, Ariane Stark, **Nicholas Reich**, Evan Ray, Jarad Niemi, Khoa Le, Abdul Kanji, Dasuni Jayawardena, Yuxin Huang, Katie House, Estee Cramer, Matt Cornell, Andrea Brennen, Johannes Bracher

* underline denotes ensemble contributor

**CDC Collaborators**: Michael Johansson, Matthew Biggerstaff, Jo Walker, Velma Lopez, Rachel Slayton

**Ensemble “advisors"**: Jacob Bien, Logan Brooks, Sebastian Funk, Tilmann Gneiting, Anja Muhlemann, Aaron Rumack, Ryan Tibshirani

**Modeling groups:** Over 50 groups at various institutions have contributed forecasts to the hub

COVID-19 ForecastHub

# Background

- Each week the Hub receives forecasts of weekly incident and cumulative deaths and incident cases in the US due to COVID-19 from over 50 teams.

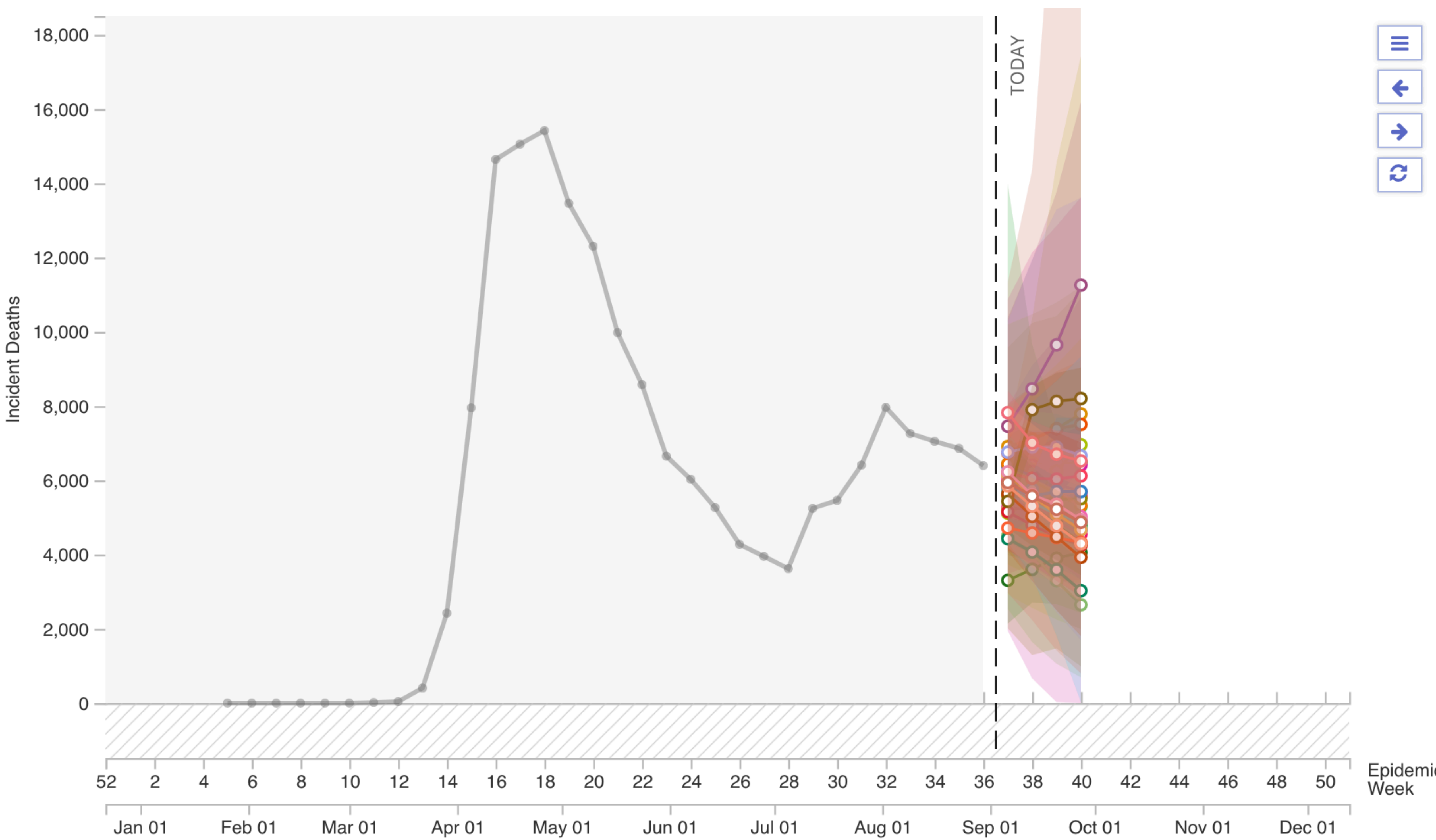
- The Hub builds an ensemble that combines predictions from these models for 1 through 4 week ahead forecasts.

# Modeling approaches vary

- YYG-ParamSearch: "**machine learning** techniques on top of a **classic infectious disease model** to make projections for infections and deaths."
- UMass-MechBayes: "**classical compartmental models from epidemiology**, prior distributions on parameters, models for time-varying dynamics, models for partial/noisy observations of confirmed cases and deaths."
- UCLA-SuEIR: "an improved **SEIR model** for predicting the dynamics among the cumulative confirmed cases and death of COVID-19"
- IHME-CurveFit: "**hybrid modeling approach** to generate our forecasts, which incorporates elements of statistical and disease transmission models."
- MOBS-GLEAM_COVID: "The GLEAM framework is based on **a metapopulation approach** in which the world is divided into geographical subpopulations. Human **mobility between subpopulations is represented on a network**."
- UT-Mobility: "For each US state, **we use local data from mobile-phone GPS traces** made available by [SafeGraph] to quantify the changing impact of social-distancing measures on 'flattening the curve.' "
- GT-DeepCOVID: "This **data-driven deep learning model** learns the dependence of hospitalization and mortality rate on various detailed syndromic, demographic, mobility and clinical data."

# Demo Visualization

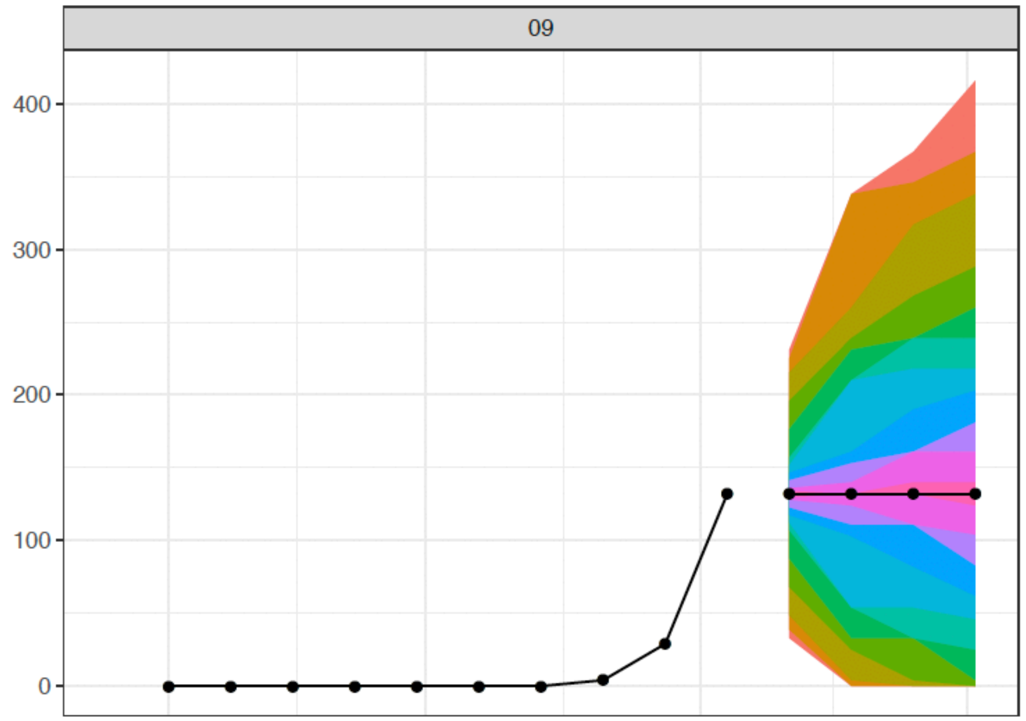
**https://viz.covid19forecasthub.org/**



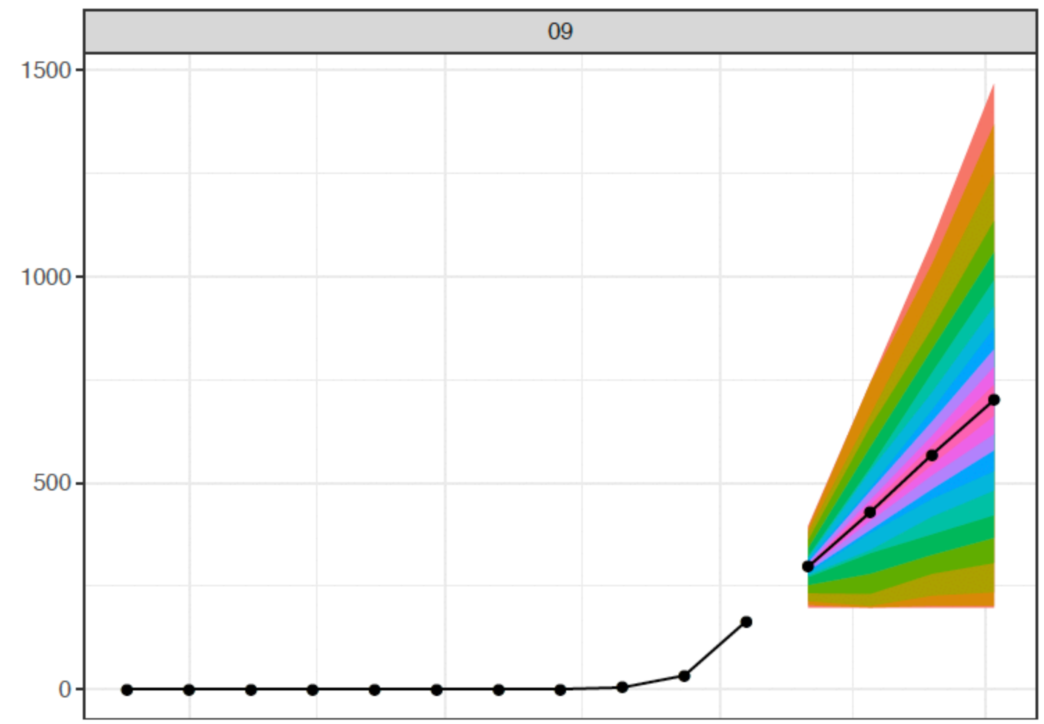
# Baseline Model

- Different from flu forecasting baseline model! Not "seasonally" driven.
- Acknowledgment: idea adapted from a suggestion by Ryan Tibshirani (CMU).
- Goal: Median predicted incidence is most recent observed incidence.
- Predictions of cumulative deaths derived from predictions of incident deaths.

**Incident Deaths**



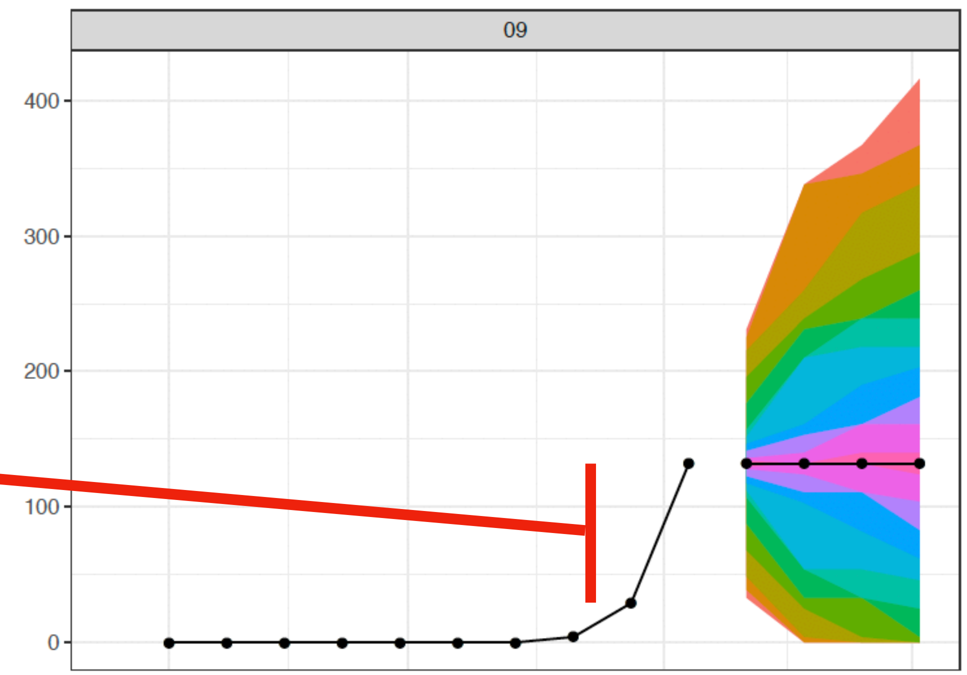

**Cumulative Deaths**



# Baseline Model

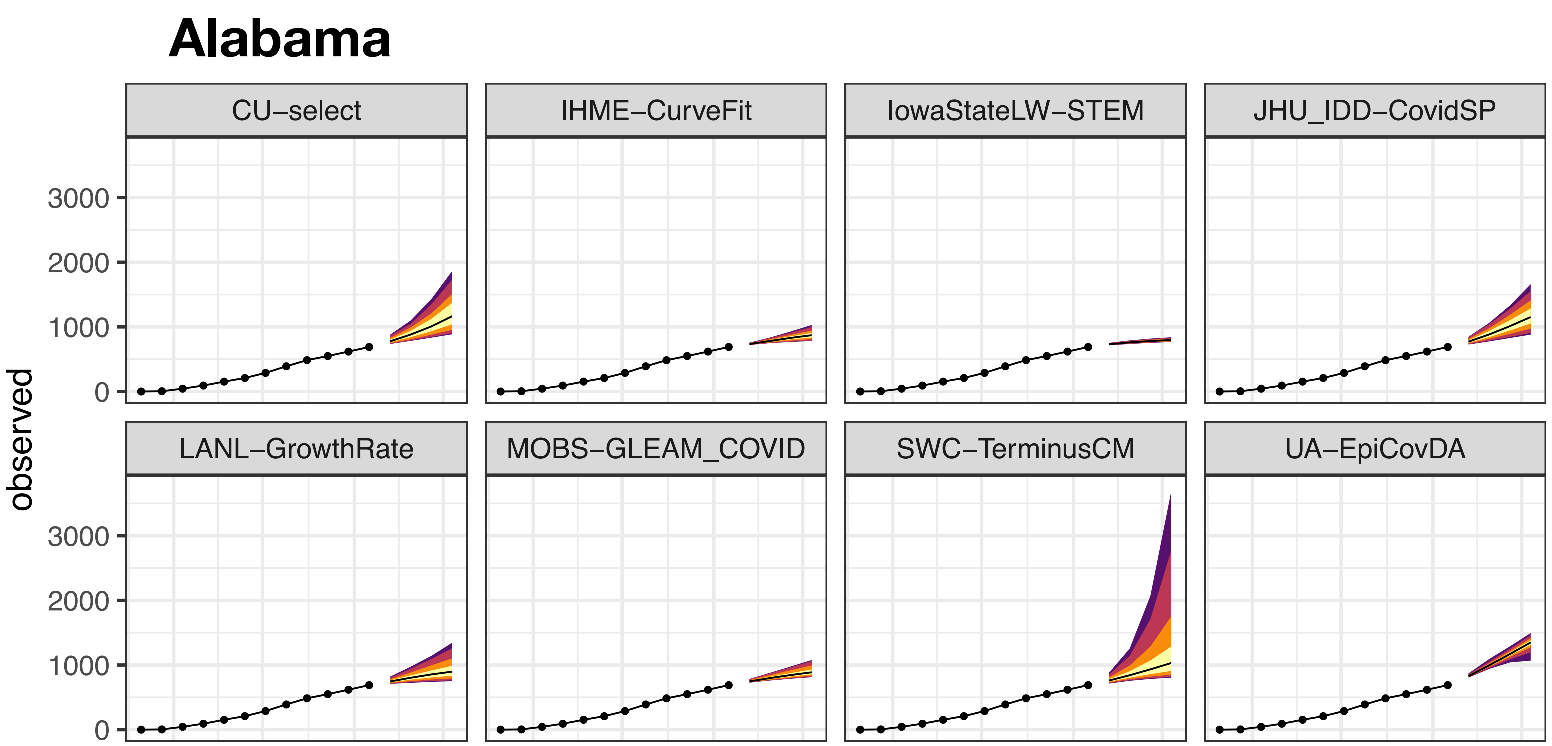
- Procedure:
  - Compute first differences of historical incidence:

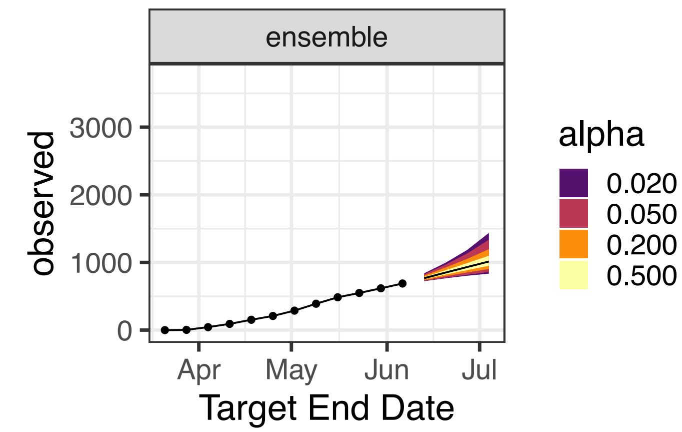
$$d_t = y_t - y_{t-1}$$




  - Collect first differences and their negatives
  - Sample first differences and add to last observed incidence; take quantiles of the resulting distribution
  - Iterate for horizons > 1
  - Adjustments for "niceness":
    - Force median = last observed incidence
    - Truncate at 0

# Building the Ensemble: View 1




- For each combination of spatial unit s, time point t, and forecast horizon h, teams are required to submit K=23 quantiles of a predictive distribution:

$$\widehat{P}(Y \leq q^m_{s,t,h,1}) = 0.01,\ \widehat{P}(Y \leq q^m_{s,t,h,2}) = 0.025,\ \ldots,\ \widehat{P}(Y \leq q^m_{s,t,h,12}) = 0.5,\ \ldots,\ \widehat{P}(Y \leq q^m_{s,t,h,23}) = 0.99$$

The predictive median

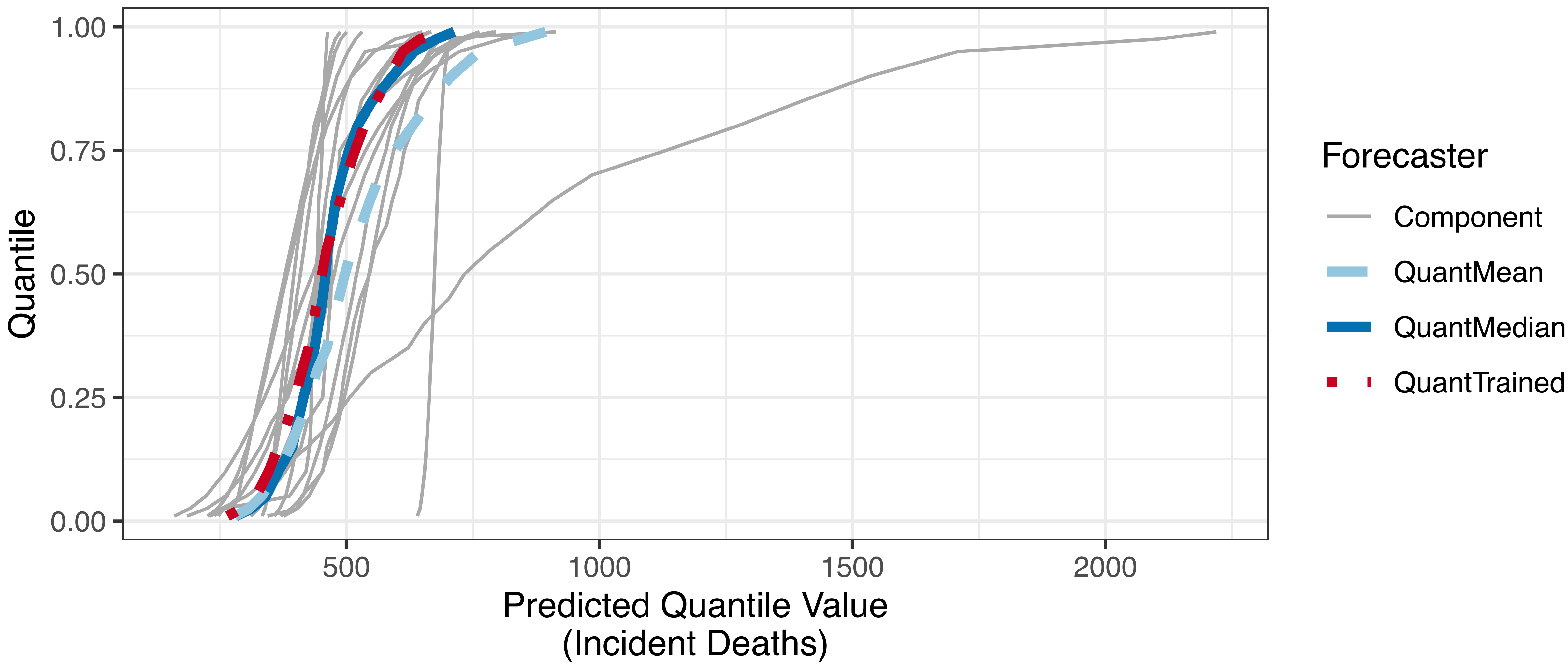
Limits of a 98% prediction interval

- The predictive quantiles for the ensemble are a combination of component predictions at each quantile level:

$$q_{s,t,h,k} = f(q^1_{s,t,h,k}, \ldots, q^M_{s,t,h,k}) \textbf{ for each } k = 1,\ldots,23$$

# Building an Ensemble: View 2

- The pairs $\left(q^m_{s,t,h,k}, \widehat{P}(Y^m_{s,t,h} \leq q^m_{s,t,h,k})\right)$ fall along the predictive CDF for model m



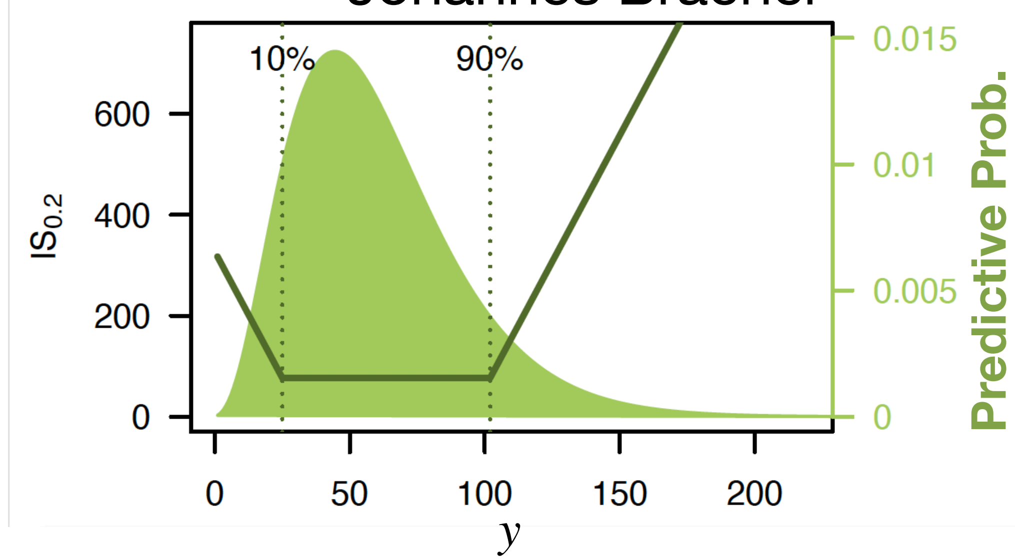

- Three options for the combination function f:
  - QuantMean: $q_{s,t,h,k} = \frac{1}{M}\sum_{m=1}^{M} q^m_{s,t,h,k}$ **Used through July 21, 2020**
  - QuantMedian: $q_{s,t,h,k} = \textbf{median}(q^1_{s,t,h,k}, \ldots, q^M_{s,t,h,k})$ **Used starting July 28, 2020**
  - QuantTrained: $q_{s,t,h,k} = \beta^0_{t,h,k} + \sum_{m=1}^{M} \beta^m_{t,h,k} \cdot q^m_{s,t,h,k}$ **Evaluated, not released each week**

# Forecast Skill: Weighted Interval Score

- Consider a single $(1-\alpha)\times 100\,\%$ predictive interval [l, u] for the observed response y. The interval score is:

$$\mathbf{IS}_\alpha(F,y) = \underbrace{(u-l)}_{\text{Width of interval}} + \underbrace{\frac{2}{\alpha}\cdot(l-y)\cdot\mathbf{1}(y<l)}_{\text{Penalty if interval is too high}} + \underbrace{\frac{2}{\alpha}\cdot(y-u)\cdot\mathbf{1}(y>u)}_{\text{Penalty if interval is too low}},$$

Figure due to Johannes Bracher



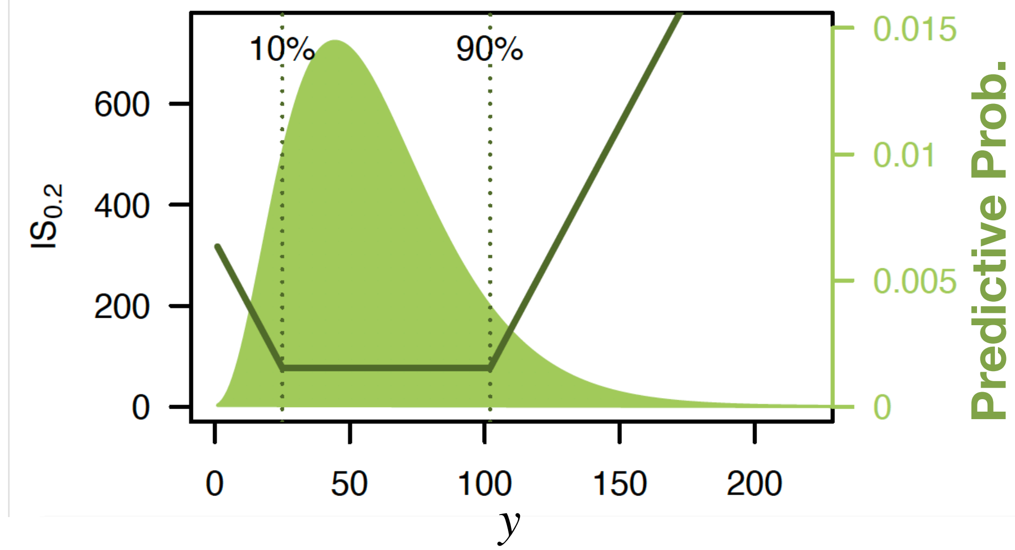

- Smaller $\mathbf{IS}_\alpha$ is better

# Forecast Skill: Weighted Interval Score

- Consider a single $(1-\alpha)\times 100\,\%$ predictive interval [l, u] for the observed response y. The interval score is:

$$\mathbf{IS}_\alpha(F,y) = \underbrace{(u-l)}_{\text{Width of interval}} + \underbrace{\frac{2}{\alpha}\cdot(l-y)\cdot\mathbf{1}(y<l)}_{\text{Penalty if interval is too high}} + \underbrace{\frac{2}{\alpha}\cdot(y-u)\cdot\mathbf{1}(y>u)}_{\text{Penalty if interval is too low}},$$

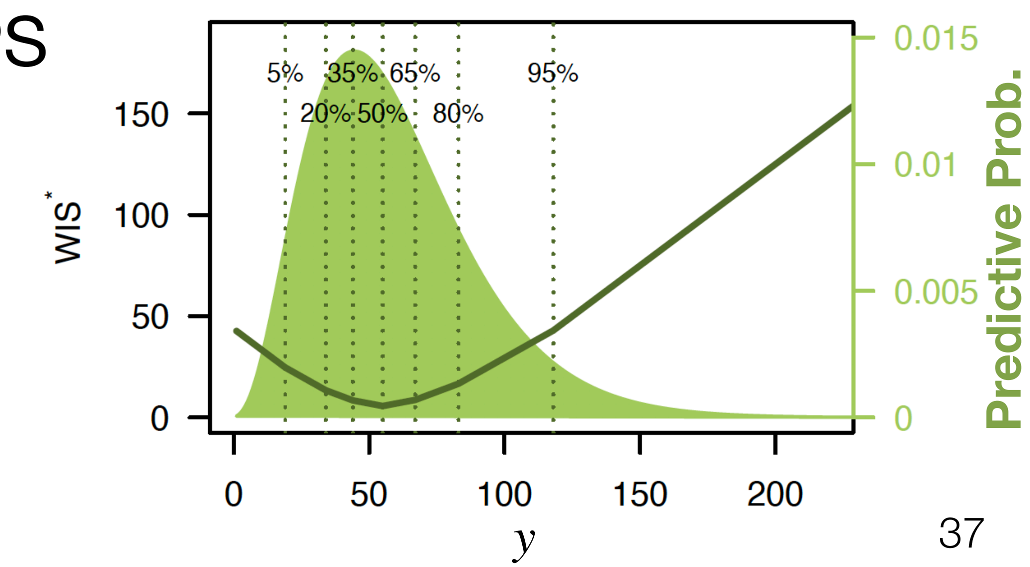
Figure due to Johannes Bracher




- Smaller $\mathbf{IS}_\alpha$ is better
- For multiple predictive intervals, we compute a weighted average of $\mathbf{IS}_\alpha$

$$\mathbf{WIS}_{\alpha_{0:K}}(F,y) = \frac{1}{K+1}\times\left(w_0\times 2\times|y-m| + \sum_{k=1}^{K}\left(w_k\times\mathbf{IS}_{\alpha_k}(F,y)\right)\right).$$

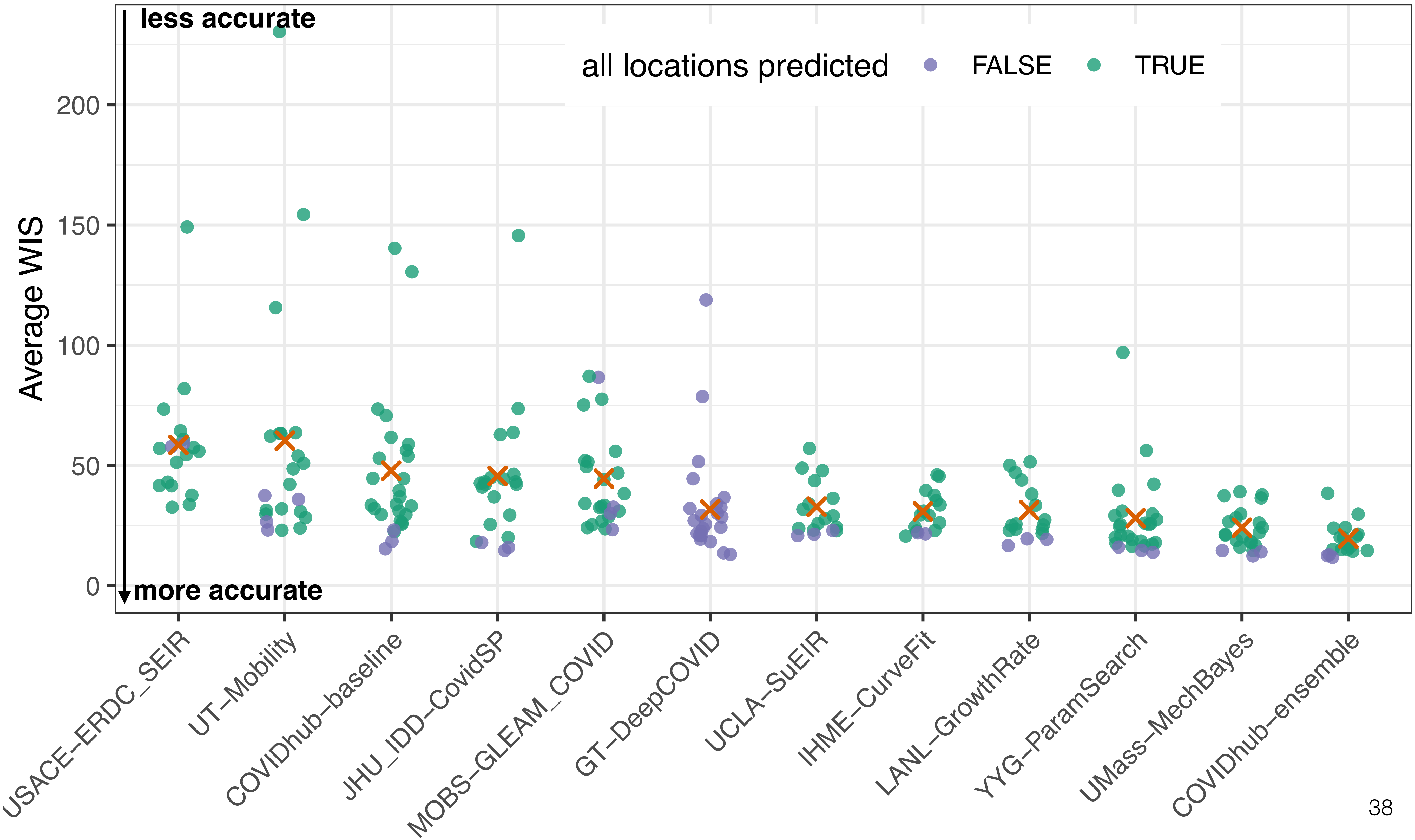
- We use weights $w_i = \dfrac{\alpha_i}{2}$, in which case WIS ≈ CRPS (continuous ranked probability score)
- The resulting score is **proper**: in expectation, it is minimized by the true predictive distribution.
- See Bracher et al. (2020) for more: https://arxiv.org/abs/2005.12881



# Evaluation: Ensemble vs Components (WIS)

Forecasts evaluated from teams submitting weekly between May and Sept.

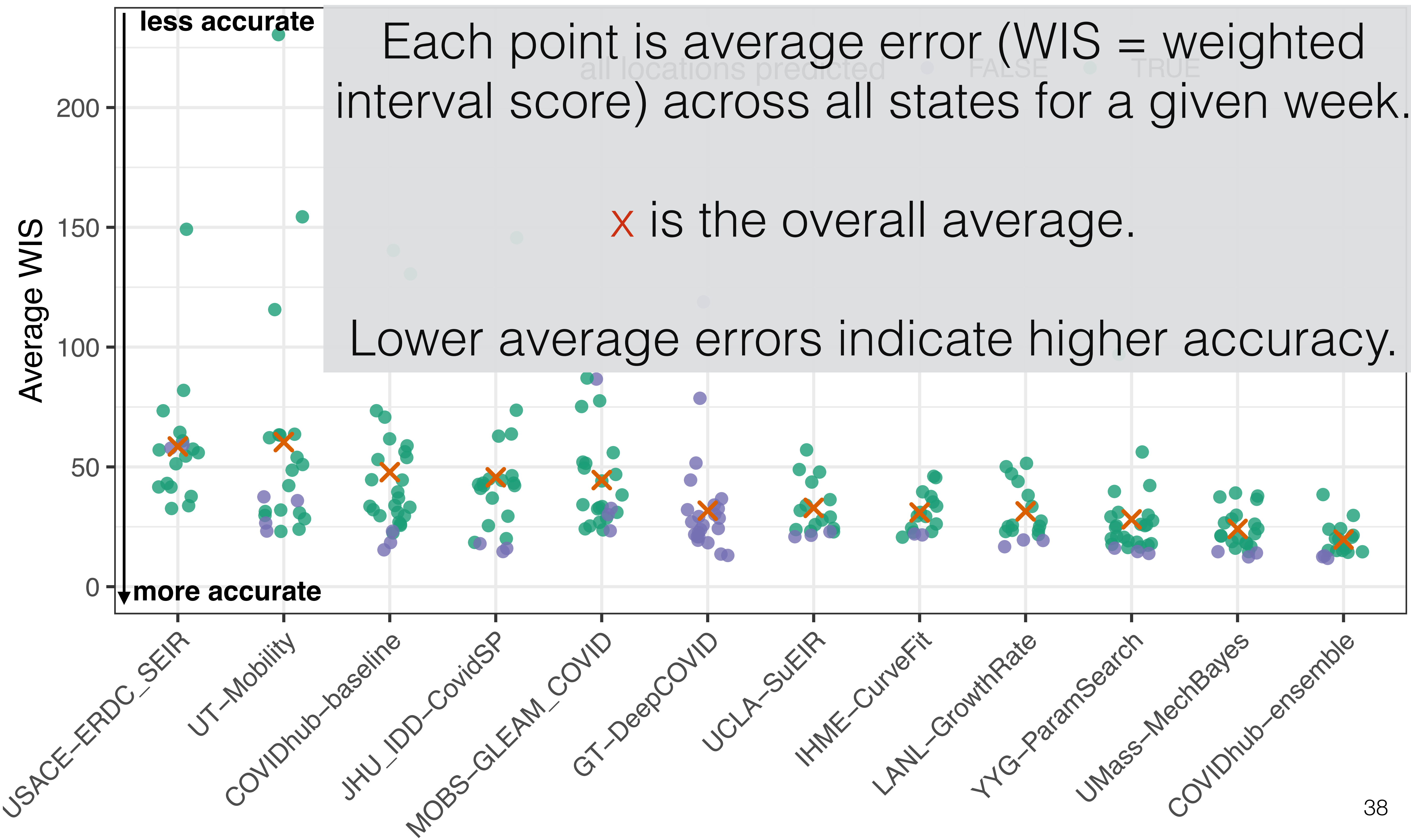
**Mean WIS across all forecasted locations, one point per week and model**



# Evaluation: Ensemble vs Components (WIS)

Forecasts evaluated from teams submitting weekly between May and Sept.

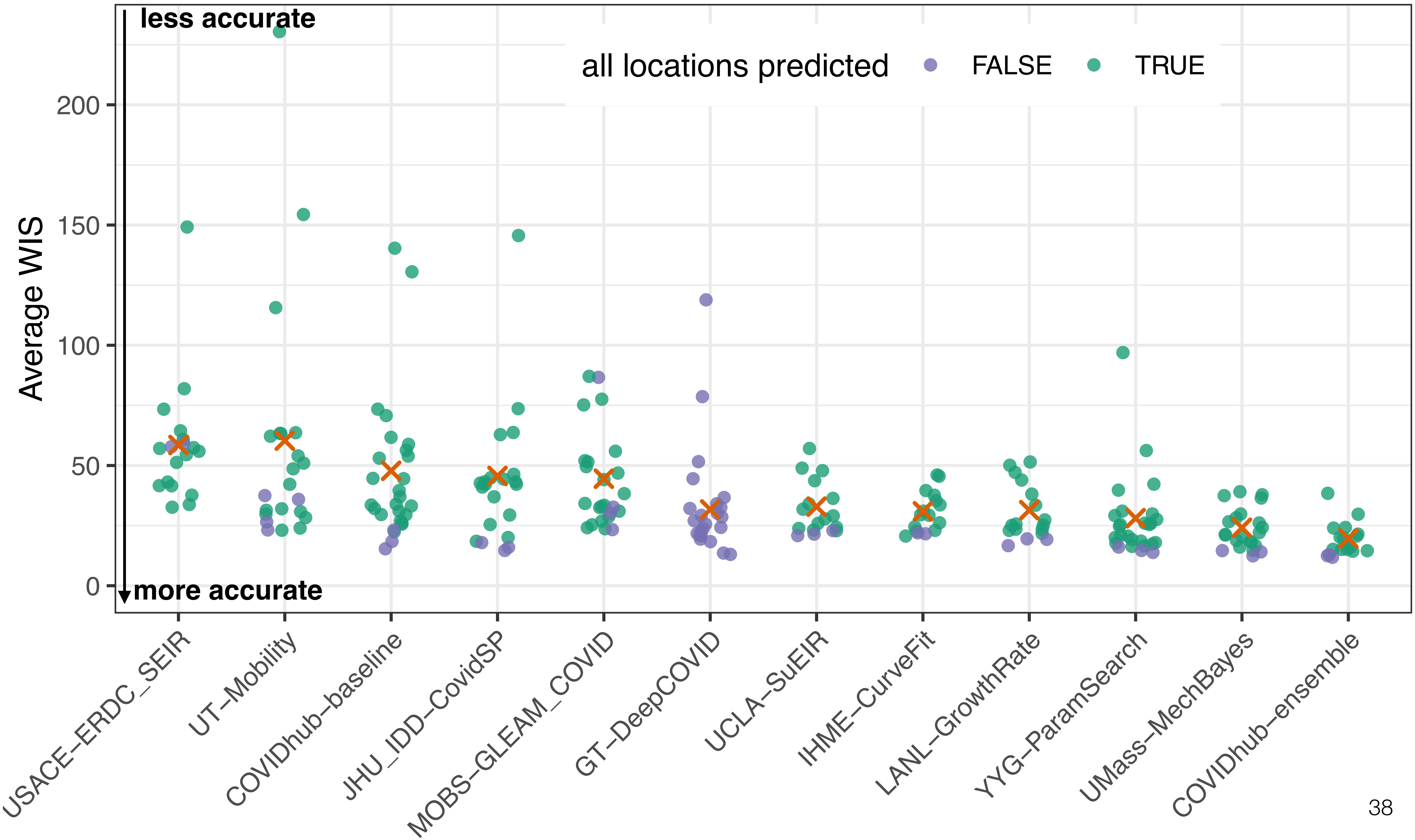
**Mean WIS across all forecasted locations, one point per week and model**



# Evaluation: Ensemble vs Components (WIS)

Forecasts evaluated from teams submitting weekly between May and Sept.

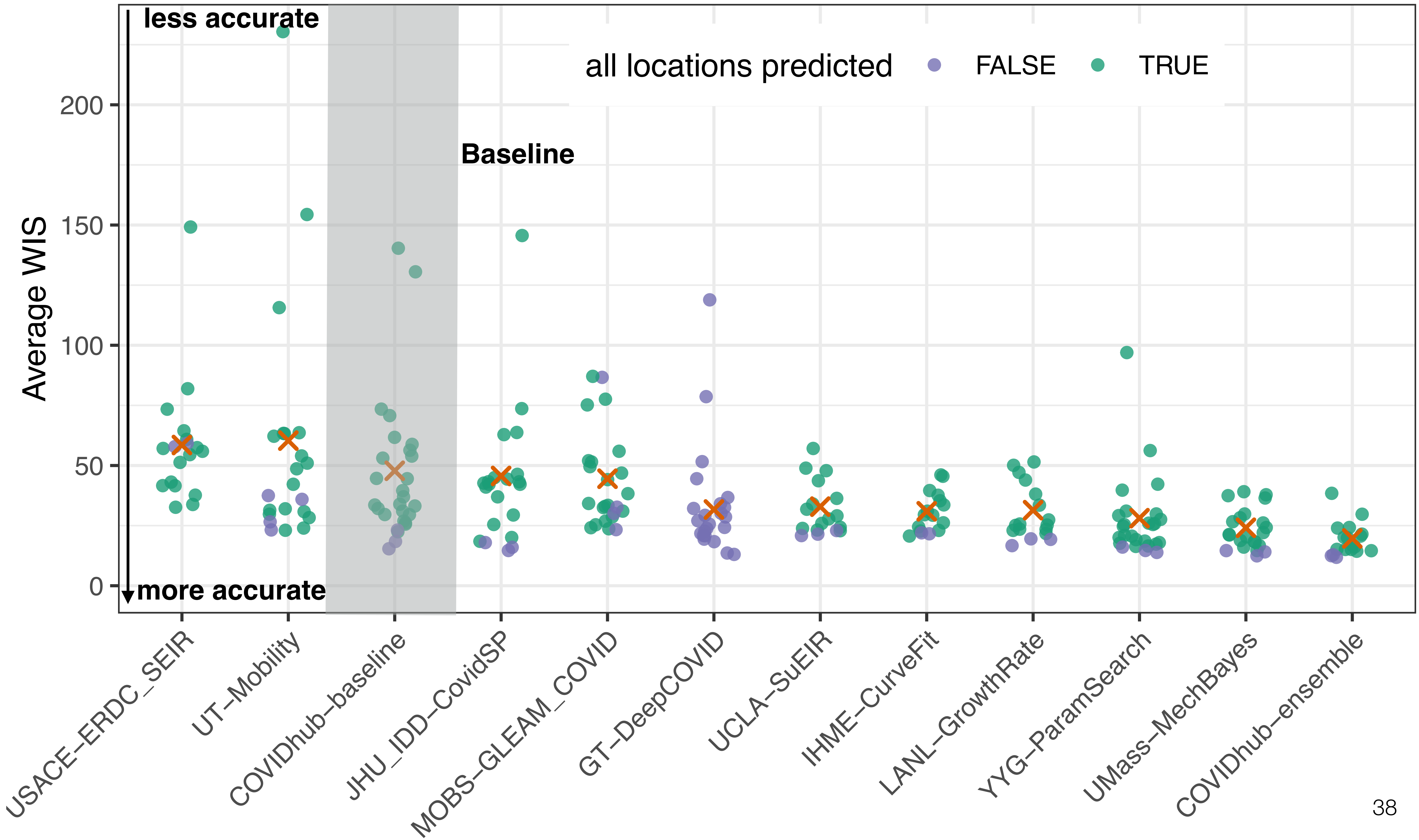
**Mean WIS across all forecasted locations, one point per week and model**



# Evaluation: Ensemble vs Components (WIS)

Forecasts evaluated from teams submitting weekly between May and Sept.

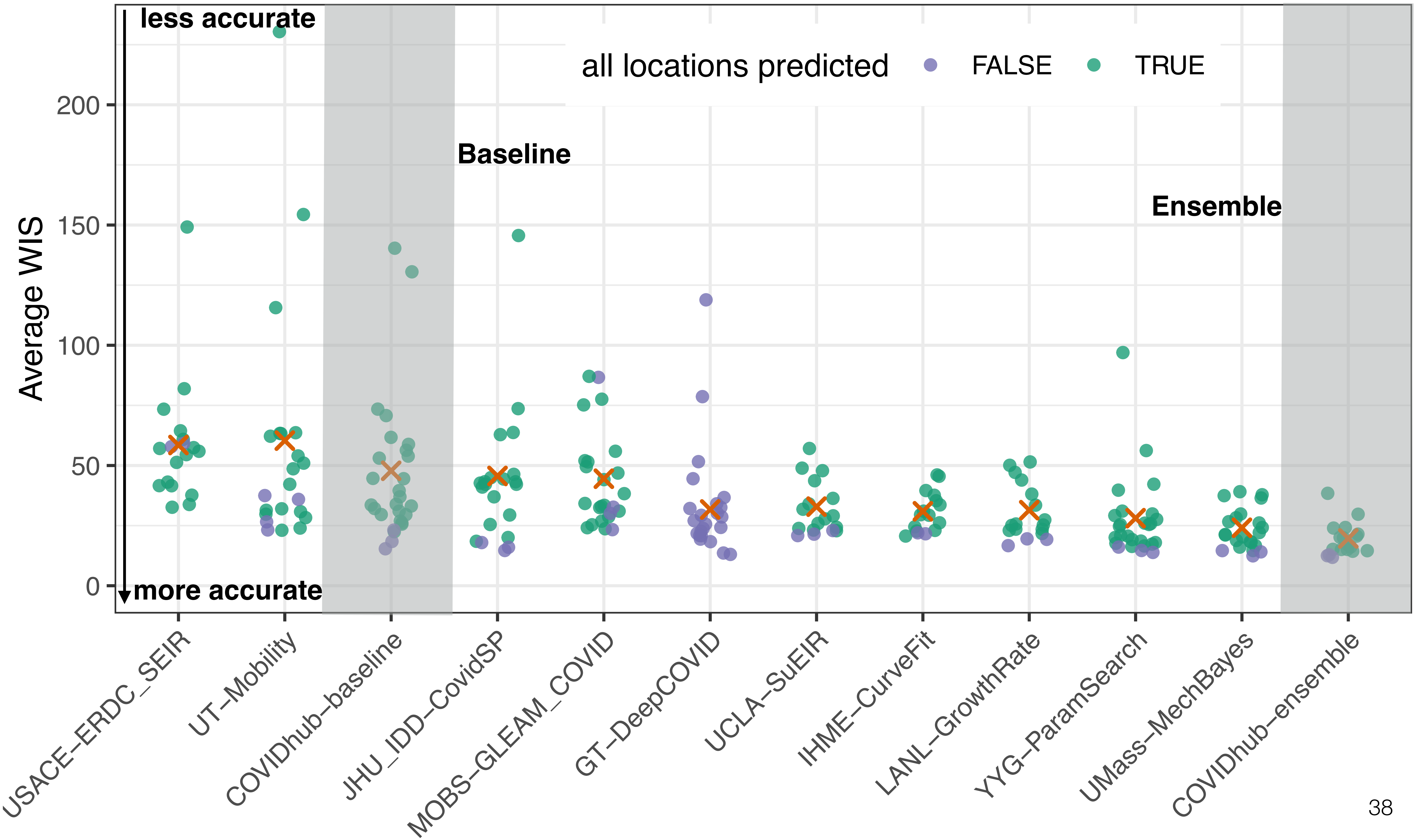
**Mean WIS across all forecasted locations, one point per week and model**



# Evaluation: Ensemble vs Components (WIS)

Forecasts evaluated from teams submitting weekly between May and Sept.

**Mean WIS across all forecasted locations, one point per week and model**



# Ensemble coverage rates

Observed prediction interval (PI) coverage rates are close to nominal levels. Below numbers calculated across all 1-4 week ahead incident death ensemble forecasts from June through October where observed data is available.

| interval level | empirical coverage rate |
|---|---|
| **50% PI** | 54% |
| **80% PI** | 79% |
| **95% PI** | 90% |

# Evaluation: Ensembles Compared

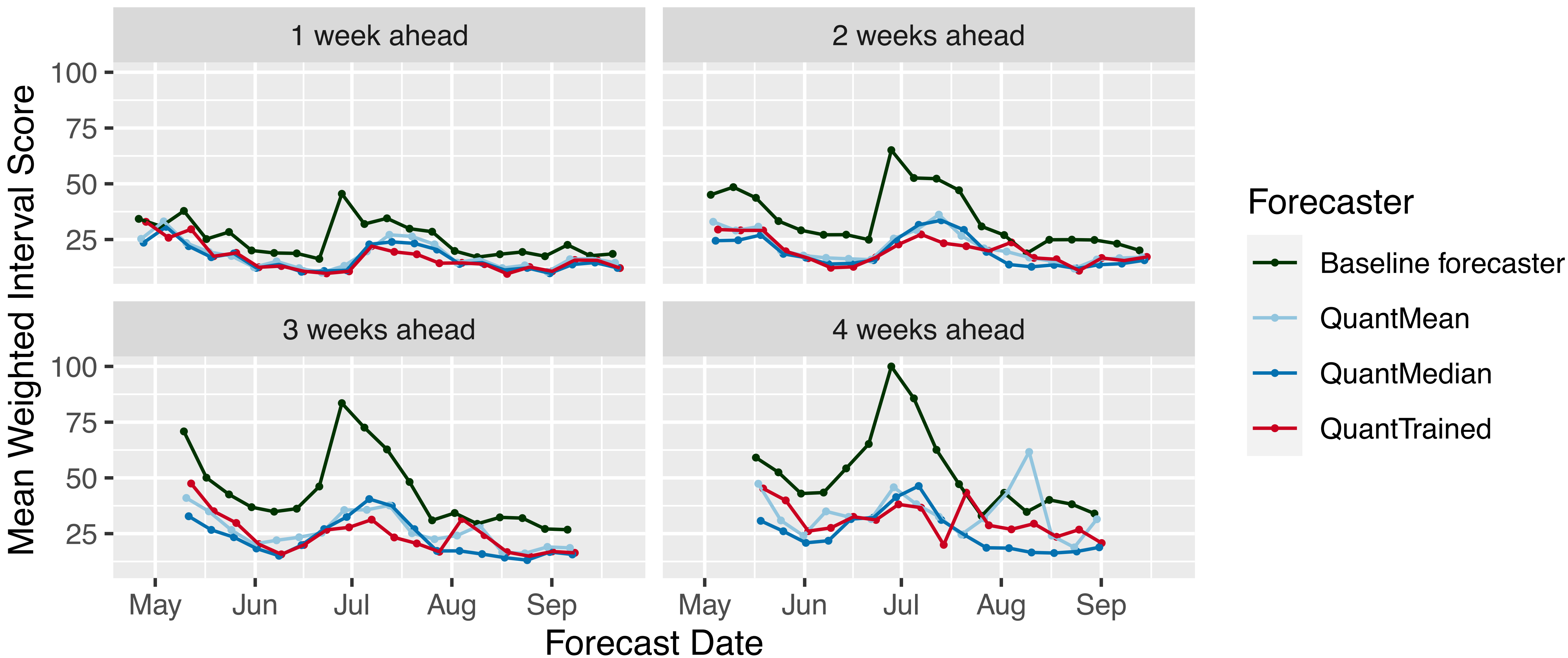



Figure credit: Logan Brooks (CMU)

Summary:

Baseline < QuantMean < QuantTrained <= QuantMedian

https://forecasters.org/blog/2020/10/28/comparing-ensemble-approaches-for-short-term-probabilistic-covid-19-forecasts-in-the-u-s/

# Lessons from COVID-19 forecasts

- **COVID-19 is "less predictable" than flu**
  - with flu, we have 10 years of training data
  - harder to beat the baseline model
  - model performance week-to-week varies
  - not a big sample size to work with!

- **(Simple) ensemble forecasts add value**
  - more accurate than any single model
  - add'l complexity doesn't improve ensembles

# Infectious Disease Forecasting: ongoing challenges

# Challenge 1: Data sparsity

(infectious disease dynamics cannot be observed like the weather)

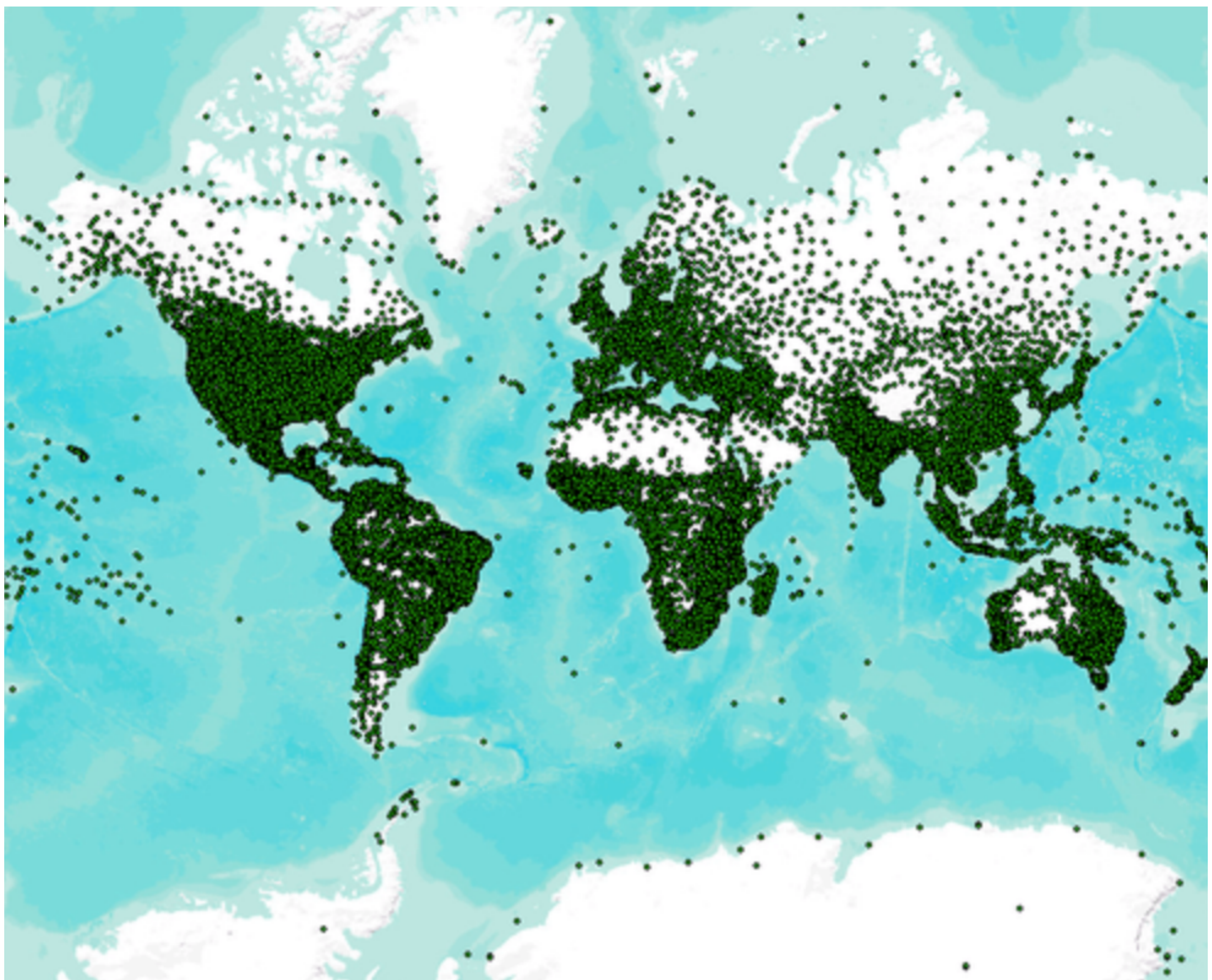


image credit: https://goo.gl/images/CSSQRv

Each dot represents a weather station whose data was used to create the WorldClim dataset.

image credit: https://databasin.org/datasets/15a31dec689b4c958ee491ff30fcce75

# Challenge 2: Feedback loop

- Weather forecasts can't change the weather.
- An outbreak forecast could change an outbreak.


US military troops heading to Liberia to assist with Ebola outbreak.
image: defense.gov

Images of vector-control activities to control dengue in Thailand
courtesy of Sopon Iamsirithaworn, Thailand Department of Disease Control

# Challenge 3: Translation into action

Dan Jernigan, Director of Influenza Division, CDC
September 2018

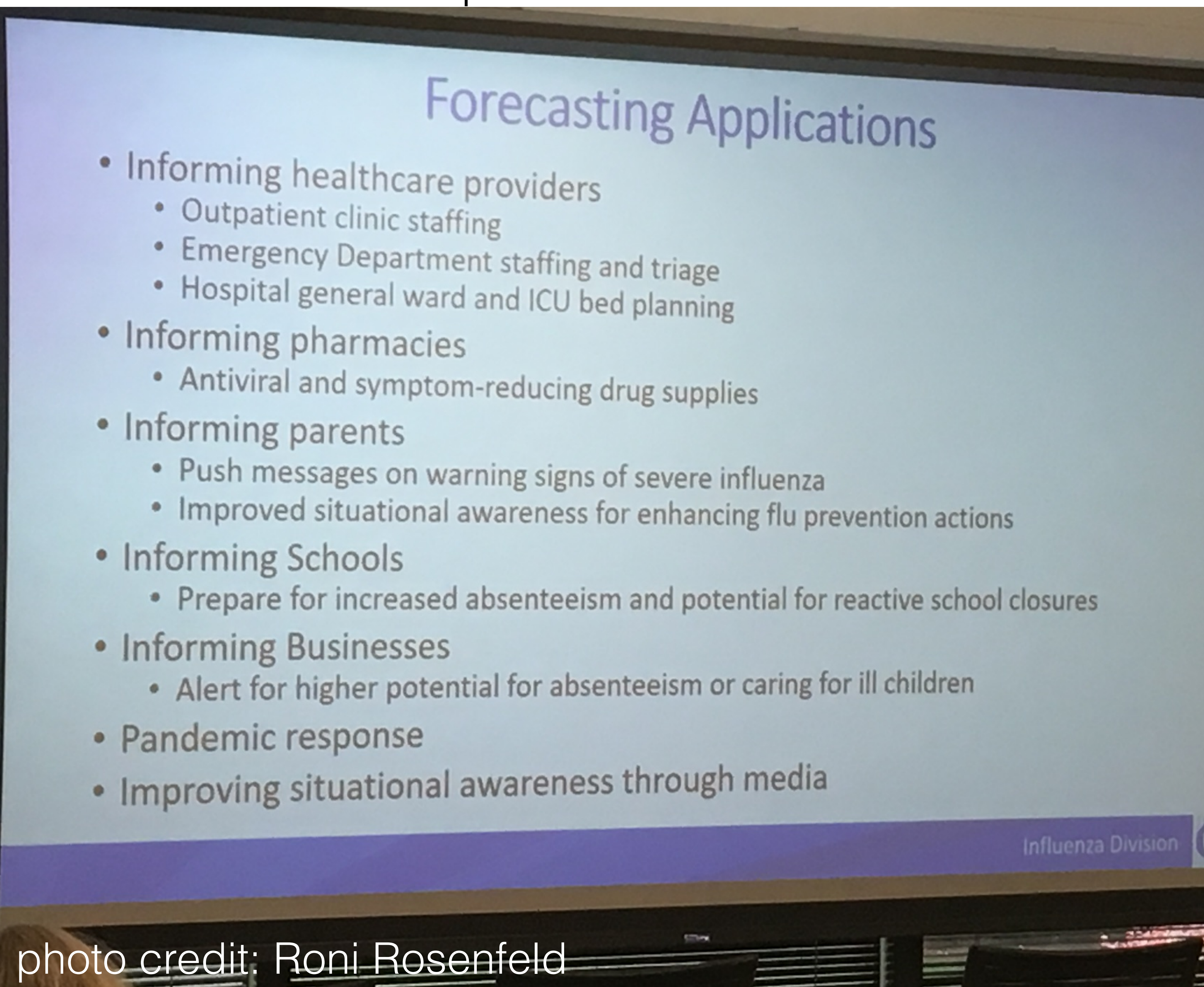



photo credit: Roni Rosenfeld

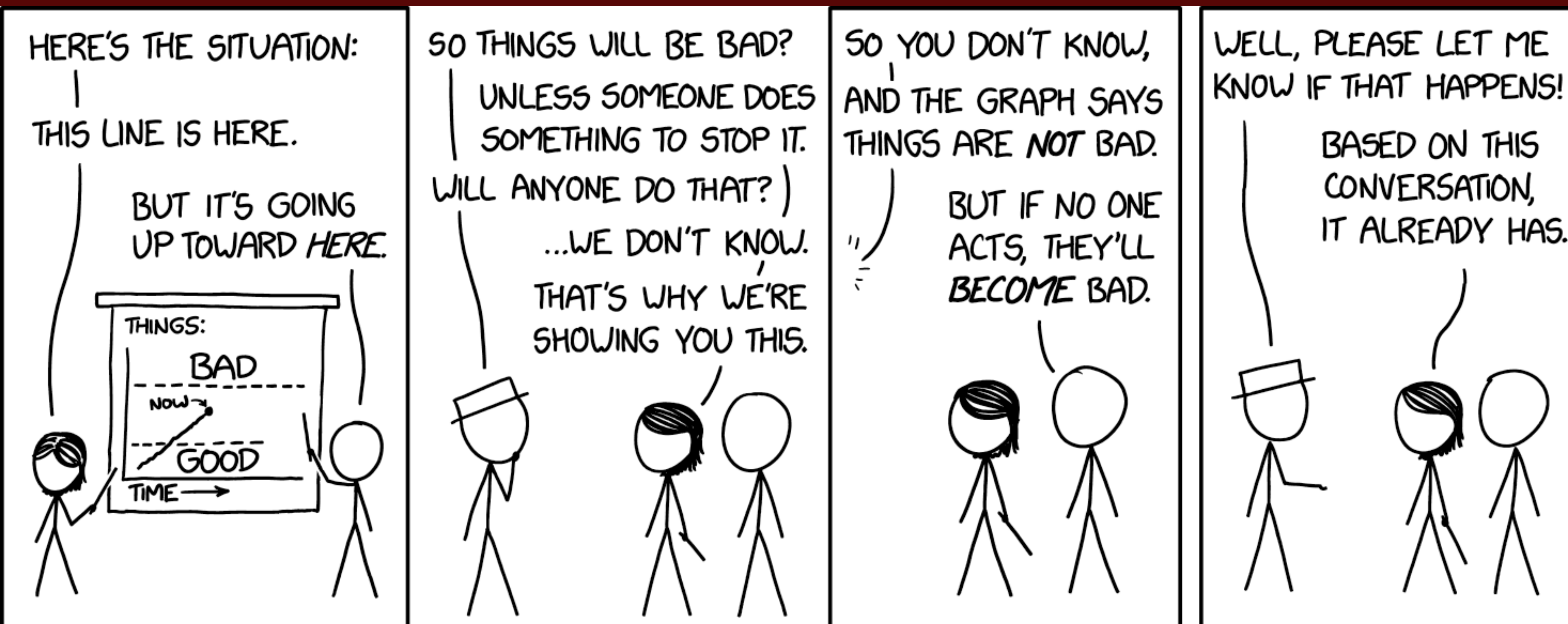
# Challenge 3: Translation into action

COVID-19 ForecastHub

# Thank you!

(we're hiring a post-doc, link on Hub website)

Reich Lab | AT UMASS AMHERST




https://xkcd.com/2278/