Machine Learning, forecasting, and explainability

Affan Shoukat
Assistant Professor of Data Science
University of Regina

University of Regina

Forecasting for decision makers

Surge capacity

Beds, ICU and staffing over the next 1–4 weeks

Vaccines & supplies

Timing campaigns, antivirals and stock

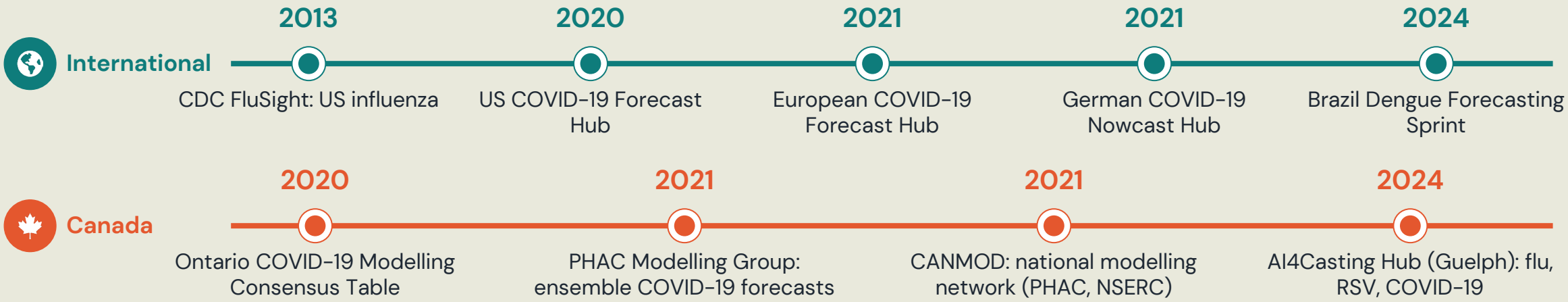
Vector control

Acting on dengue before the peak, not after

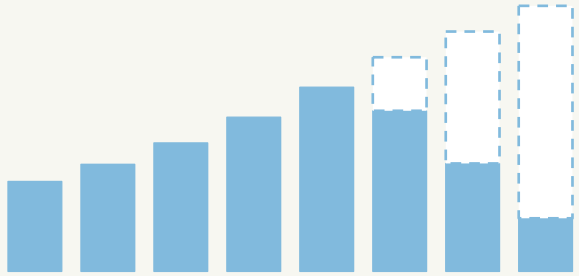
Public messaging

Telling people what to expect, honestly

Collaborative forecasting hubs and networks



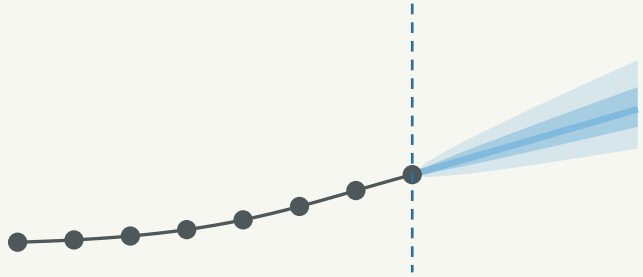
Nowcast, forecast, or projection?



Nowcast

What is happening right now?

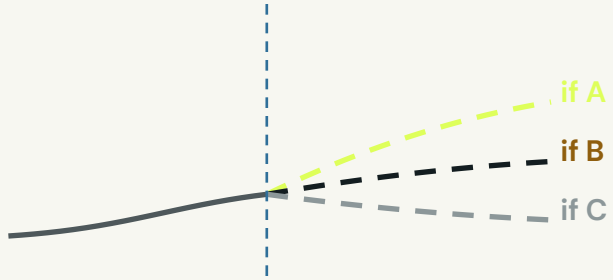
Fills in cases that have not been reported yet, fix reporting errors



Forecast

What will happen in the next 1–4 weeks?

Checked against data as it arrives



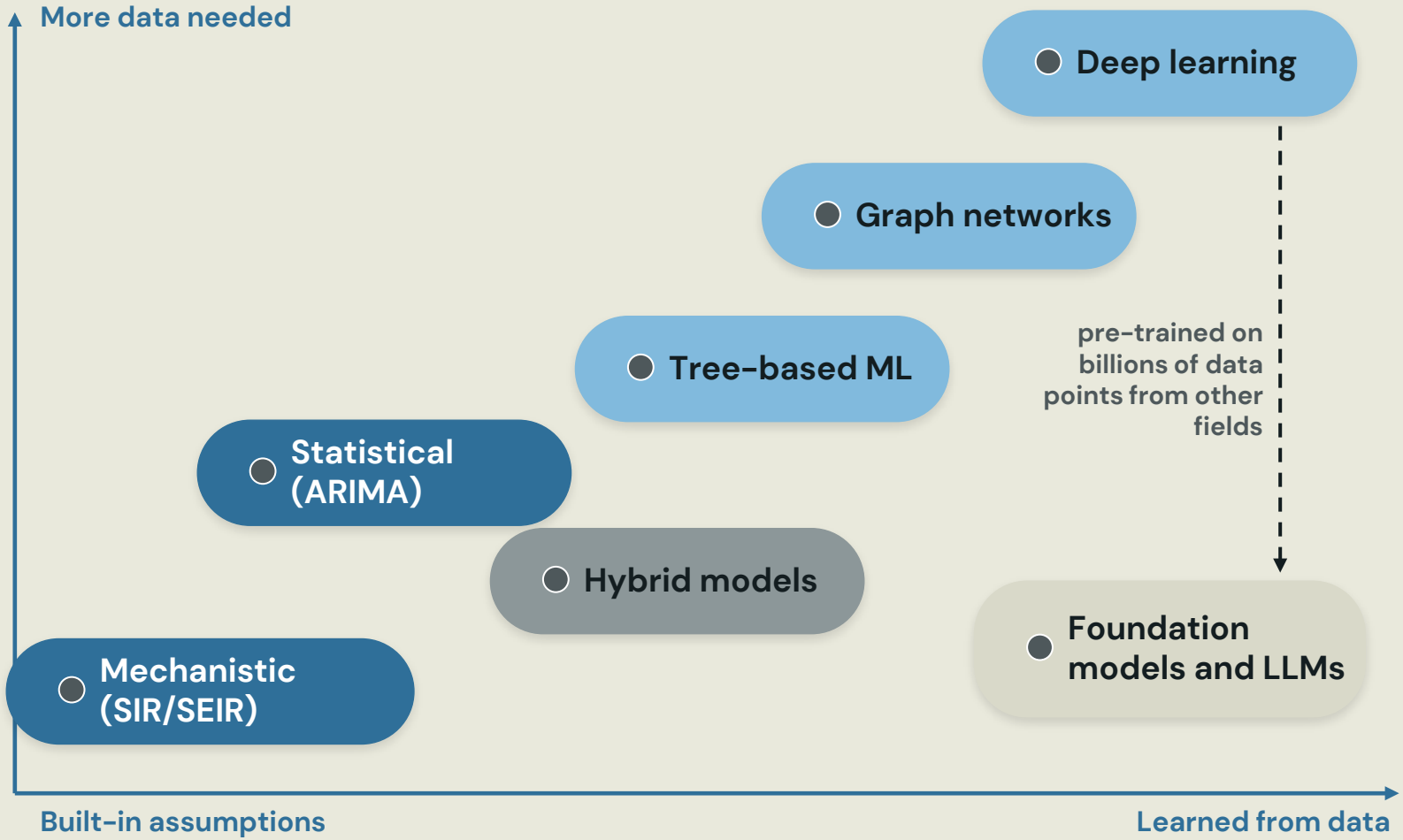
Projection

What would happen if…?

Compares options; counterfactual evaluations

Most AI forecasting research answers the middle question. Keep scenario projections separate when briefing decision-makers.

A map of forecasting models



Not a model family, but…

Ensembles

Combine forecasts from any of these families

Early warning

Flag that a change is coming, without predicting counts

Different Model Types

Traditional

Simulate transmission (SIR/SEIR) or extend past patterns (ARIMA)

Needs: One series; biology for mechanistic

Still the standard, and hard to beat 1–4 weeks out

Tree-based ML

Hundreds of small if-then rules (e.g. gradient boosting)

Needs: Many locations; lagged inputs

Top FluSight model in 2023/24 (Flusion)

Deep learning

Neural networks that learn long, non-linear lags (LSTMs)

Needs: Long histories or many series; GPUs

Strong on climate-driven dengue; overfits small data

Graph networks

Regions linked by travel; forecasts "listen" to neighbours

Needs: Regional series plus mobility data

Modest gains; mobility data hard to get

Hybrids

Epidemic equations plus machine learning for unknown parts

Needs: Counts, covariates, modelling expertise

Interpretable outputs; a top US COVID-19 hub model

Foundation models

Pre-trained on billions of series; no disease-specific training

Needs: Little local history; GPU or cloud

Best in 30 of 36 benchmark settings, but overconfident

Ensembles

Combine many models' forecasts, usually by the median

Needs: Shared targets and many teams (a hub)

Most consistent and best calibrated

No family wins everywhere

What matters is how a model scores against a baseline, in real time.

Next: scoring →

What makes a good forecast

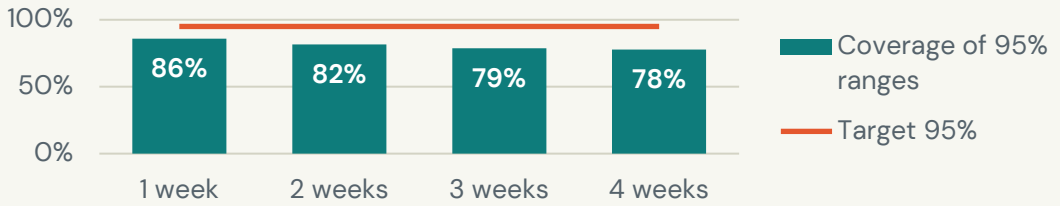
Coverage

Can the forecast's ranges be trusted?

How it's calculated

1. Collect many forecasts, each with a 95% range
2. When the real numbers arrive, count the hits
3. Coverage = hits ÷ forecasts; **target 95%**

Real example: CDC flu ensemble, 2022–23



Below target: overconfident (ranges too narrow). **Above:** too cautious.

Weighted interval score (WIS)

How good is the forecast overall? Lower is better

How it's calculated: for each forecast, add up

1. **Distance** from the forecast's middle to what happened
2. **Width** of the ranges (narrow is rewarded)
3. **Penalty** if the truth falls outside, bigger the further out

Relative WIS = model's average WIS ÷ baseline's average WIS

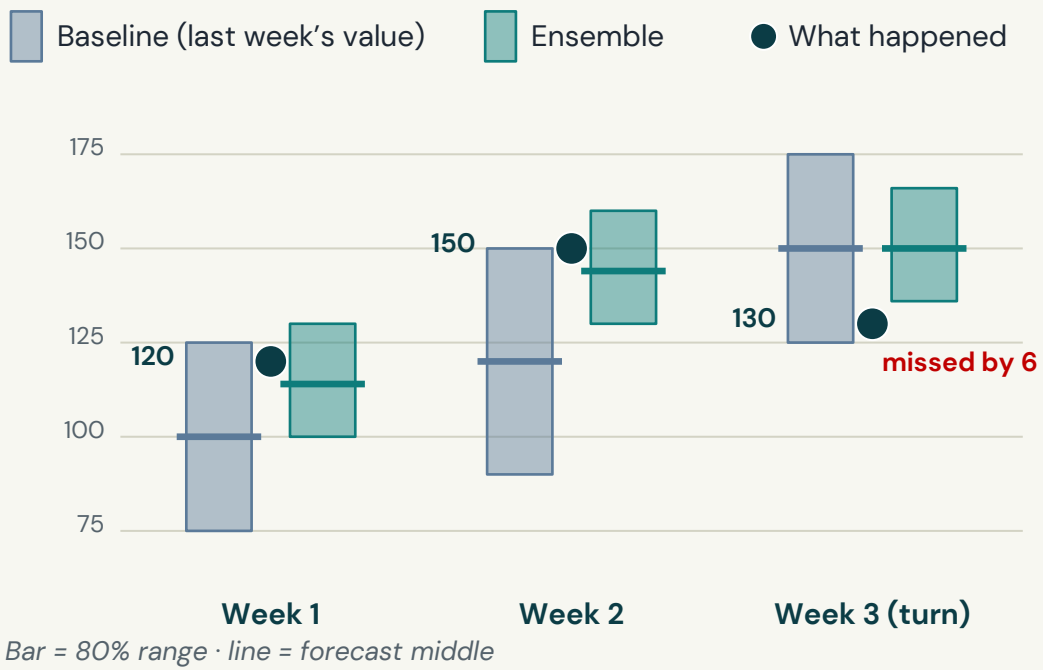
Baseline: "next week looks like this week"



Both are worked out after the real numbers arrive, across many forecasts. The next slide walks through a small example.

Example of Coverage and WIS

Weekly deaths: forecasts and what happened



Bar = 80% range · line = forecast middle

WIS: error score (lower is better)

Week	Baseline	Ensemble
1	10.0	4.0
2	14.0	4.0
3	10.0	**12.7**
Average	**11.3**	**6.9**

Relative WIS

0.61

6.9 ÷ 11.3: about 39% less error than the baseline

Coverage: did the 80% range catch the truth?

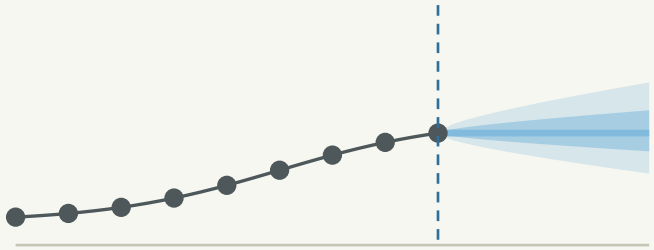
	Wk 1	Wk 2	Wk 3	Coverage	Target 80%
Baseline	✓	✓	✓	**3 of 3 = 100%**	Too cautious
Ensemble	✓	✓	✗	**2 of 3 = 67%**	Overconfident

WIS asks how good the forecast is overall and coverage asks whether its ranges can be trusted.

Beating the baseline score

Naive baseline

Next week looks like this week

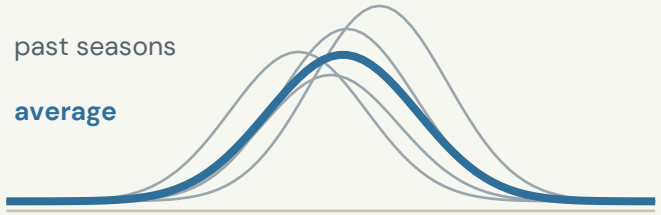


Takes the latest value as the forecast, with uncertainty from past week-to-week changes.

Used by: US and European COVID-19 hubs, FluSight today

Historical baseline

What usually happens this week

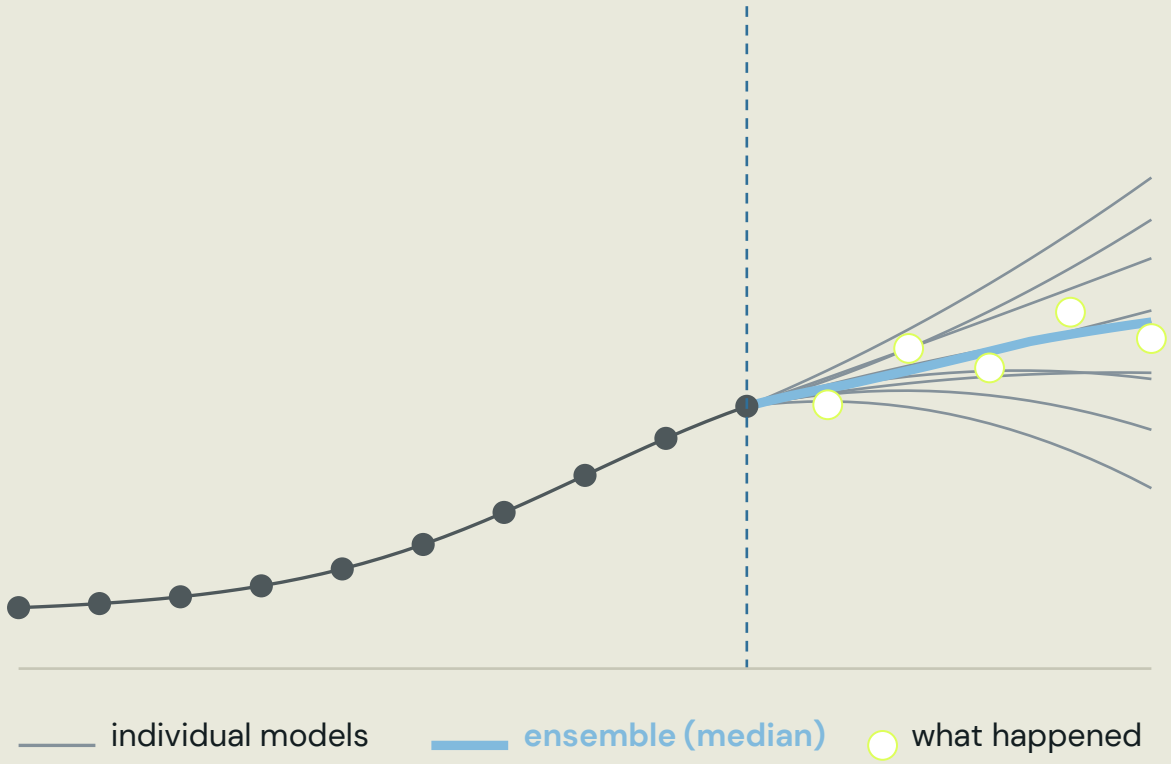


Forecasts the average for that week of the year from past seasons; ignores this season's data.

Used in: CDC's evaluation of flu forecasts, 2010–17

A forecast is only as impressive as the baseline it beats. Relative WIS below 1 means better than the baseline; always ask which baseline.

Ensemble Methods



0.61 Relative WIS for US COVID-19 deaths: about 39% less error than the baseline

91% of individual model forecasts outscored by the European ensemble (deaths, 32 countries)

All US locations: the ensemble beat the baseline everywhere, the only model that did

The top single model keeps changing; the ensemble stays near the top

2010–17: mechanistic + data assimilation

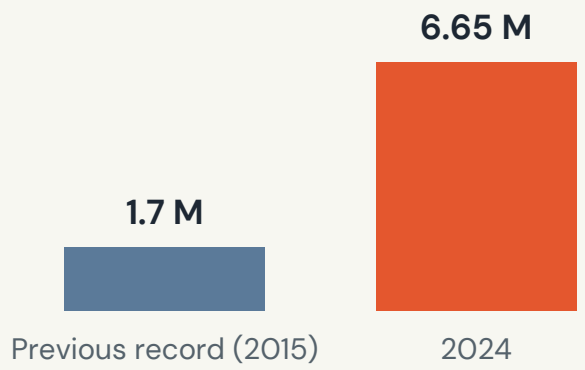
2023/24: gradient boosting

2024/25: diffusion AI model (preprint)

Extreme events, where models struggle

Record seasons

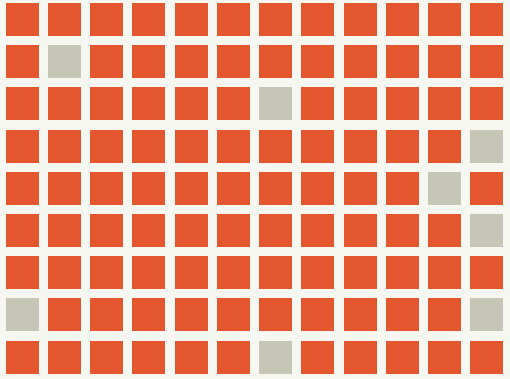
Brazil, probable dengue cases (millions)



No individual or ensemble model forecast the extreme 2024 season. Ensembles still inform Brazil's 2025 planning because they reduce uncertainty.

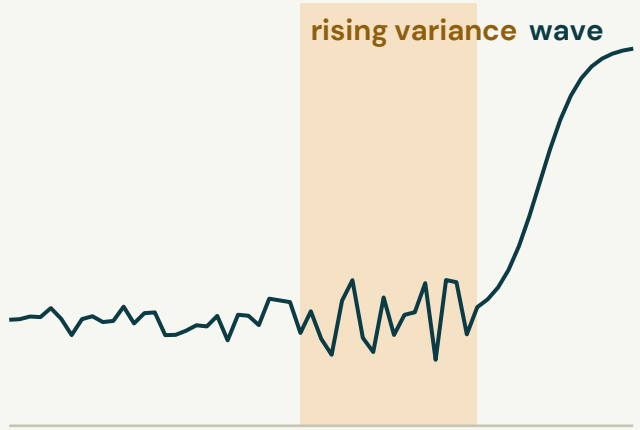
Big data hubris

Google Flu Trends vs CDC, from Aug 2011



Overestimated flu in 100 of 108 weeks. Fix: combine search data with surveillance (ARGO).

Early warning signals



UK first COVID-19 wave: signals rose 14–29 days ahead. Europe's second wave: they did not reliably rise.

Explaining a forecast

Why explain a forecast?

Accountability
Decision-makers have to defend what they act on.

Catch spurious learning
Google Flu Trends' top searches included "high school basketball", which peaks in flu season.

Learn about drivers
Which climate signals or data streams come first?

Two kinds of tools

Interpretable by design
Regression coefficients; parameters of mechanistic and hybrid models such as R(t) or the reporting rate.

Explained after the fact
SHAP or permutation importance: how much worse does the forecast get if we scramble an input?

Three things to look out for

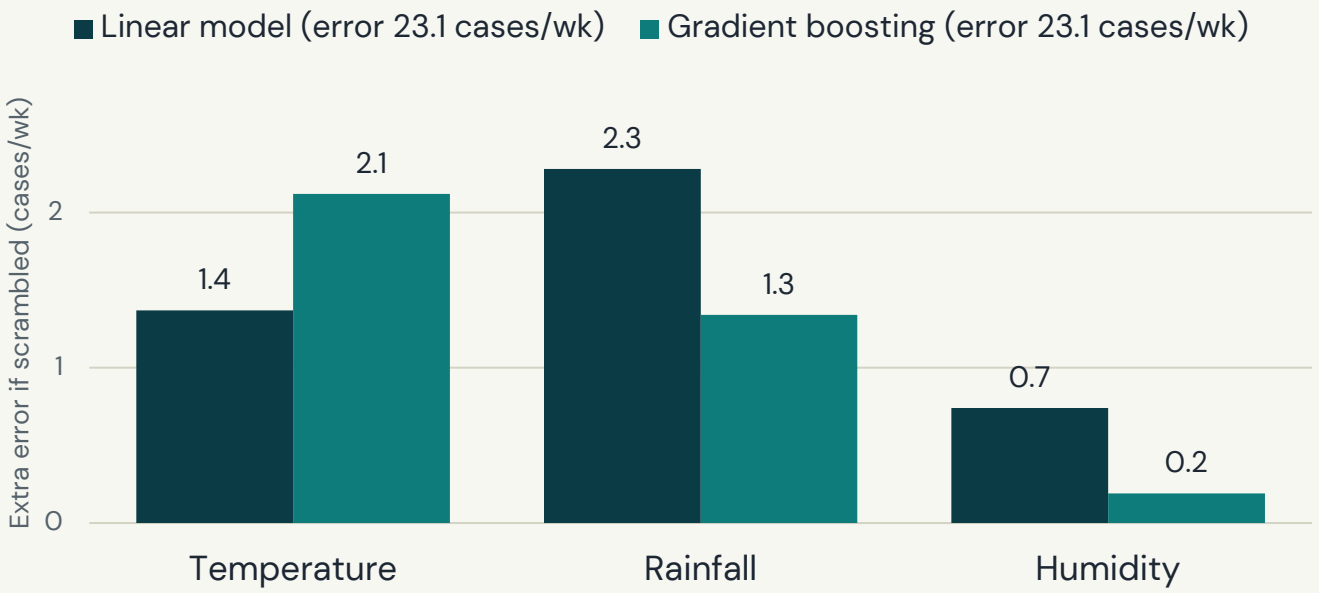
1. Equally accurate models can tell different stories
2. Interpretable doesn't mean identifiable
3. An explanation isn't a causal effect

Accuracy is not an explanation

https://arxiv.org/html/2606.06056v1

Which climate factor drives dengue in San Juan?

Two models forecasting cases 8 weeks ahead, with near-identical accuracy



Why it happens

Temperature, humidity and rainfall rise and fall together with the seasons, so models can swap credit between them.

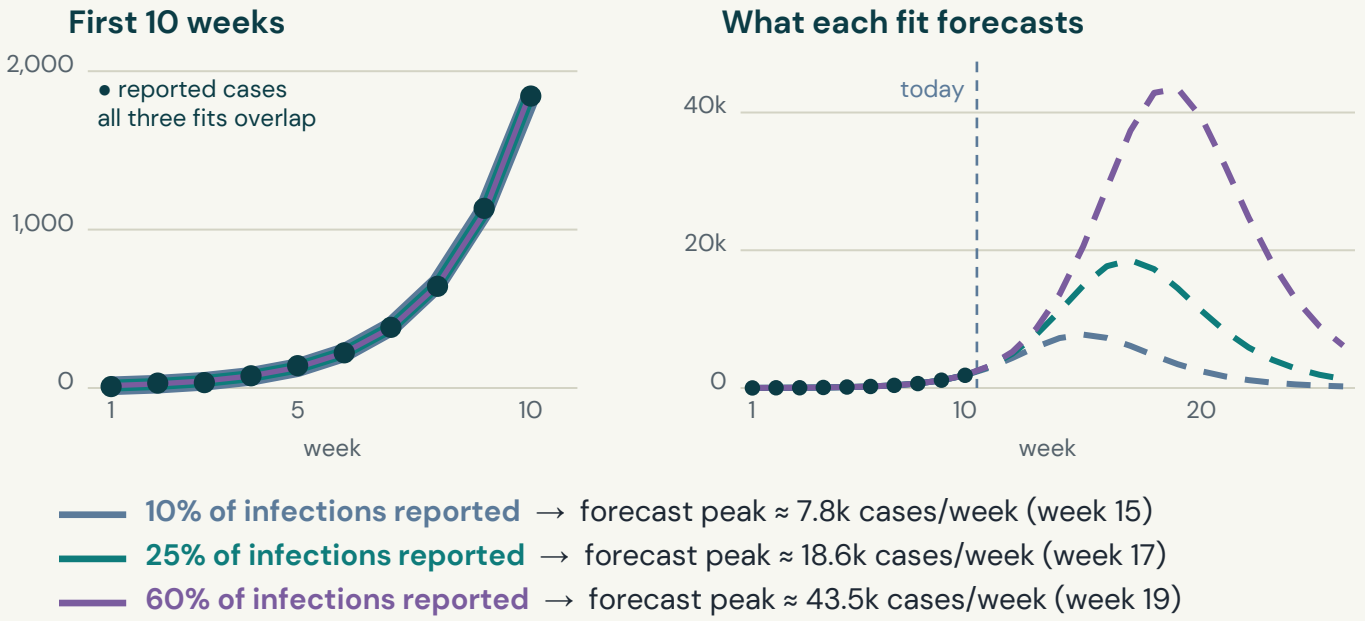
What to do: before drawing conclusions about drivers, check explanations across many good models, not just one.

Identifiability of variables

Interpretable isn't the same as identifiable

Same data, three assumed reporting rates

SIR model fitted to the first 10 weeks of reported cases



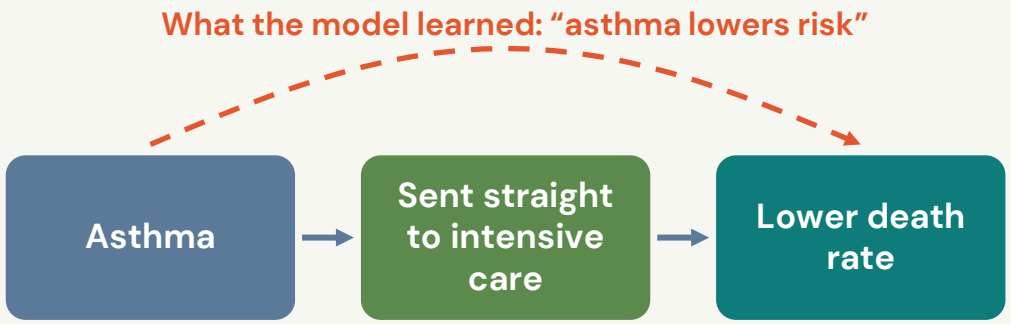
- Interpretable: means the parameter has a meaning or a definition.
  - For example, the reporting rate.
- Identifiable means the data contains enough information to work out its value.

What to do: check what the data can pin down before interpreting a parameter, and anchor it with independent data such as serosurveys.

An explanation is not a causal effect

Pneumonia risk model

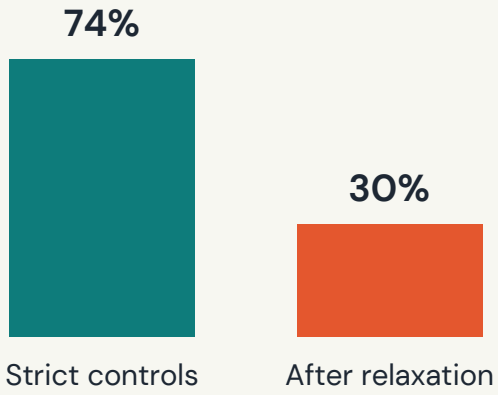
Predicting death among patients with pneumonia



- The model learned the care pathway
- Acting on it, by treating asthma patients less aggressively, would put them at greater risk.

Mobility and COVID-19, 52 countries

How well mobility explained transmission (median R^2)



80%

of countries saw transmission decouple from mobility once controls were relaxed

A strong predictor in spring 2020 stopped tracking transmission once behaviour changed. A predictor isn't a policy lever.

A forecast can use any signal that predicts. "What happens if we act?" needs a causal model, or a scenario projection with explicit assumptions.

A few take-home messages

Use multiple models (ensembles).

Clearly state your baseline score and evaluate forecasts against it.

Are there other models that produce the results?

Explanations describe the model, not the epidemic or a policy effect

Five questions for any AI forecast

1. Was it tested in real time on data it had never seen?
2. Does it beat a simple baseline, with ranges scored by WIS and coverage?
3. Did it work at the turning points?
4. Do other equally accurate models give the same explanation?
5. Can the data pin down the parameters being interpreted?

Thank you for attending

KEEP IN TOUCH


ai4ph.dlsph@utoronto.ca


@ai4ph.bsky.social


/ai4ph-health-research-training-platform


AI4PH Slack workspace



AI4PH is supported by the Canadian Institutes of Health Research (CIHR).