Foundational Models for Forecasting: Some Fundamental Limitations

Daniel F. Schmidt
Joint work with Christoph Bergmeir

Department of Data Science and AI
Monash University

Machine Learning Coffee Seminar
Finnish Centre for AI (FCAI)
June 2nd, 2025

Outline

About Monash University

- Largest university in Australia
- I am from the Department of Data Science and AI in the Faculty of IT
- Department has researchers working in:
  - Time series (classification, forecasting)
  - Bayesian machine learning (deep learning)
  - Vision and language
  - Bioinformatics and medical health
- Most famous staff member is Prof. Chris Wallace, inventor of minimum message length
- Department always looking for collaborations, and able to host visitors
- A little advertising for our group: we have a brand new repository and "bake-off' for large time series for classification
  - "MONSTER: Monash Scalable Time Series Evaluation Repository", Dempster et al., https://arxiv.org/abs/2502.15122

Outline

Forecasting

- We will be examining forecasters
- In our context, a forecaster:
  - takes an input series, say $\mathbf{x} = (x_1, \ldots, x_n)$;
  - produces an estimate of the future h values of the series, i.e., $\hat{\mathbf{x}} = (\hat{x}_{n+1}, \ldots, \hat{x}_{n+h})$
- The aim is to build forecasters that produce forecasts that are close (in some sense) to future realisations of the process we are forecasting
- To keep things simple we acknowledge the existence of other modes of operation, i.e., exogenous predictors, etc. but will stick to this setup to keep things simple
  - It is also a very common scenario

Local and Global Models (1)

- Local models (ARIMA, ETS, etc.)
  - Traditionally, one time series is seen as a dataset
  - One model is built per time series
- Global models
  - A set of time series is a dataset (e.g., a set of series from retail, smart meters, etc.)
  - Build a model across the series

  $\Longrightarrow$ now enough data to make ML methods more competitive
- A foundation model for forecasting is essentially a type of high complexity global model, usually a deep neural network of some kind, usually trained on a "diverse" set of time series

Local and Global Models (2)

- Global models have shown success to a surprising degree
- Idea originally was that the series have to be in some way "related/similar" so that we can learn something useful across them
- Montero-Manso and Hyndman (2020):
  - Global model can produce the same forecasts as local models, without any assumptions about similarity
  - The series don't have to be "related", *per se*
  - Series are related through the evaluation regime: we evaluate as an average (aggregate) error over all of them
  - Similar ideas are present in the statistical theory: Stein's paradox

Outline

A Simple Forecasting Problem (1)

- Consider the following simple forecasting problem
- We observe a single observation, say y, from a Gaussian process
  - We are asked to forecast a future value y_f
- Let us assume stationarity
  - We can then write $y_f = \mu + \varepsilon_f$, where $\varepsilon_f \sim N(0, \sigma^2)$
- Given a forecast $\hat{y}_f$, our squared-error forecast risk is

$$\begin{aligned}\mathbb{E}\left[(y_f - \hat{y}_f)^2\right] &= \mathbb{E}\left[(\mu + \varepsilon_f - \hat{y}_f)^2\right] \\ &= \mathbb{E}\left[(\mu - \hat{y}_f)^2\right] + \sigma^2\end{aligned}$$

- The second term doesn't depend on our forecast (irreducible)
  $\Longrightarrow$ so the problem is determined by how close $\hat{y}_f$ is to μ

A Simple Forecasting Problem (2)

- Forecast error would be minimized by $\hat{y}_f = \mu$, i.e., the oracle
- Obviously, we don't know μ but can estimate it via ML/LS $\implies$ use $\hat{y}_f \equiv \hat{\mu}_{\mathrm{LS}} = y$ as our forecast
- The squared-error risk would then be

$$\mathbb{E}\left[(\mu - \hat{\mu}_{\mathrm{LS}})^2\right] = \mathbb{E}\left[(\mu - \mu + \varepsilon)^2\right] = \sigma^2$$

  where we note that $y = \mu + \varepsilon$
- This is in fact the optimal forecast for this problem
  - Uniquely minimax and admissable (cannot be dominated)

A More Complex Problem (1)

- Now let us generalize the problem a little
- Imagine that we have p stationary, independent Gaussian processes
- We observe a single observation from each, i.e., $\mathbf{y} = (y_1, \ldots, y_p)$
  - We now need to forecast p future values, one for each process
- We can assess the performance based on aggregate squared-error risk

$$\begin{aligned} R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathrm{LS}}) &= \mathbb{E}\left[||\mathbf{y}_f - \hat{\mathbf{y}}_f||^2\right] \\ &= \mathbb{E}\left[||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}||^2\right] + p\sigma^2 \end{aligned}$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ is the vector of unconditional means for each of the p processes
- An example of the fundamental "multiple-means" problem
- Such problems are not uncommon; imagine we are trying to forecast primary votes for p different parties in an election based on a survey

A More Complex Problem (2)

- We need to estimate $\boldsymbol{\mu}$
- Again, we could use least-squares, i.e., $\hat{\boldsymbol{\mu}}_{\mathrm{LS}} = \mathbf{y}$
  - This is a type of local model
  - Each "series" used independently to make predictions
- The aggregate squared-error is then

$$\mathbb{E}\left[||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathrm{LS}}||^2\right] = \sum_j \mathbb{E}\left[(\mu_j - y_j)^2\right] = p\sigma^2$$

- The problems are independent, and the least-squares rule is optimal for each coordinate, so this must be the optimal rule, right?

James-Stein Estimation (1)

- Remarkably, the answer is, not necessarily
- Stein (1956), and James and Stein (1961) proved the following facts:
  - ✓ The least-squares estimator is minimax
  - ✓ The least-squares estimator, for $p < 3$, is admissable
  - ✗ The least-squares estimator, for $p \geq 3$, is not admissable!
- The last fact stunned the statistics community
- Its implication is that there exist estimators that dominate least-squares in this problem
- James and Stein went a step further: they provided one such estimator!

James-Stein Estimator (2)

- They proposed to use the following estimator

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}} = c(\mathbf{y})\,\mathbf{y}$$

  where

$$c(\mathbf{y}) = 1 - \frac{\sigma^2(p-2)}{||\mathbf{y}||^2}$$

  is the shrinkage factor, and $||\mathbf{y}||^2 = \sum_{j=1}^{p} y_j^2$.
- This is a type of global model
  $\Longrightarrow$ it uses information from all series to make forecasts
- This estimator dominates least-squares for $p \geq 3$, i.e.,

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathrm{JS}}) = \mathbb{E}\left[||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathrm{JS}}||^2\right] < p\sigma^2$$

  for all possible configurations values of $\boldsymbol{\mu}$
- The James-Stein estimator seems to do the impossible
  - combine unrelated problems together and improve on all of them

James-Stein Estimator (2)

- They proposed to use the following estimator

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}} = c(\mathbf{y})\,\mathbf{y}$$

where

$$c(\mathbf{y}) = 1 - \frac{\sigma^2(p-2)}{||\mathbf{y}||^2}$$

is the shrinkage factor, and $||\mathbf{y}||^2 = \sum_{j=1}^{p} y_j^2$.

- This is a type of global model
  $\Longrightarrow$ it uses information from all series to make forecasts
- This estimator dominates least-squares for $p \geq 3$, i.e.,

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathrm{JS}}) = \mathbb{E}\left[||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathrm{JS}}||^2\right] < p\sigma^2$$

for all possible configurations values of $\boldsymbol{\mu}$

- The James-Stein estimator seems to do the impossible
  - combine unrelated problems together and improve on all of them

James-Stein Estimation (3)

- *"Do you mean that if I want to estimate tea consumption in Japan I will do better to estimate simultaneously the speed of light and weight of hogs in Montana?"* (Efron and Morris, 1973)
- To answer the question, consider the properties of the risk function

$$\begin{aligned} R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}^{JS}) &< R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \text{ for all } \boldsymbol{\mu} \in \mathbb{R}^p \\ \max_{\mu_j} R(\mu_j, \hat{\mu}_j^{JS}) &\approx \sqrt{p}\sigma^2 \end{aligned}$$

Recall that $R(\mu, \hat{\mu}_{\text{LS}}) = \sigma^2$
$\Longrightarrow$ So the JS estimator does better than LS at estimating the whole of $\boldsymbol{\mu}$ at the expense of any individual μ_j

- So, the answer is yes, but only if what you will evaluate in the end is average performance across all three problems!

James-Stein Estimation (3)

- *"Do you mean that if I want to estimate tea consumption in Japan I will do better to estimate simultaneously the speed of light and weight of hogs in Montana?"* (Efron and Morris, 1973)
- To answer the question, consider the properties of the risk function

$$\begin{aligned} R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}^{JS}) &< R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \text{ for all } \boldsymbol{\mu} \in \mathbb{R}^p \\ \max_{\mu_j} R(\mu_j, \hat{\mu}_j^{JS}) &\approx \sqrt{p}\sigma^2 \end{aligned}$$

  Recall that $R(\mu, \hat{\mu}_{\text{LS}}) = \sigma^2$
  $\Longrightarrow$ So the JS estimator does better than LS at estimating the whole of $\boldsymbol{\mu}$ at the expense of any individual μ_j
- So, the answer is yes, but only if what you will evaluate in the end is average performance across all three problems!

James-Stein Estimation (4)

- Sometimes this is exactly what you plan to do
- For example, linear regression

$$Y = \sum_j \beta_j X_j + \varepsilon$$

  - We are interested in predicting the overall output, which is a weighted sum of coefficients times input data
  - We are not (necessarily) interested in estimating the individual coefficients well
  - JS: Can predict the overall output better at the expense of the estimation of the coefficients

  $\Longrightarrow$ this is the theoretical foundation for modern regularisation
- Important to stress these results do not depend crucially on our simple model setup, but are much more general!

Stein's Paradox and Foundational Forecasting Models

- And sometimes it is exactly not what you want to do
- Aggregate risk reduction is a very egalitarian thing to do
  - The individual's interests are subordinated to the group's interests
- But in practice, we are all a bit selfish :)
- If I have a series I want to forecast, I don't care how well the forecaster does on an average over a body of different series, I care about how well it will do on my particular problem
- Foundational (global) models trained and evaluated on a wide range of series might "win" on a particular repository – this is great for authors of the paper
- But in practice the user is interested in their specific problem, not an (arbitrary) collection of problems. So not so good for the user?

Outline

Just Get More Data ...

- We are in the era of deep learning
- So the answer to this problem must clearly be: collect an even more diverse body of data to train our foundational model and the neural network will just "figure it out"
- And this might be a good answer, if our time series are long and our problems with high signal-to-noise ratio, so that the series itself provides good contextual clues about the types of futures that are plausible
- But an enormous range of important, real-world forecasting problems do not meet these conditions

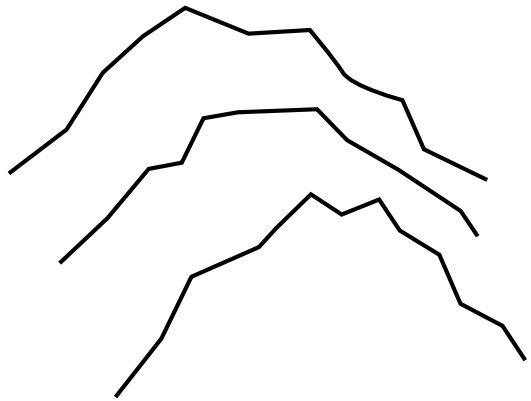
Low-granularity series are more common than you think

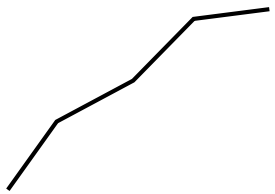
- Most supply chains run on monthly planning cycles
- Most macro-economic series have slow dynamics
  - "Let's forecast the 5-minutely unemployment rate"
- E.g., Federal Reserve Economic Data (FRED) database: "825,000 US and international time series from 114 sources"

Frequency	**Number**
Annual	$440,000+$
Monthy	$190,000+$
Quarterly	$100,000+$
Weekly	$2,400+$
Semiannual	$2,000+$
Daily	$1,500+$
5-yearly	$450+$

Uncertainty in (implicit) model selection (1)

Training data for foundational model

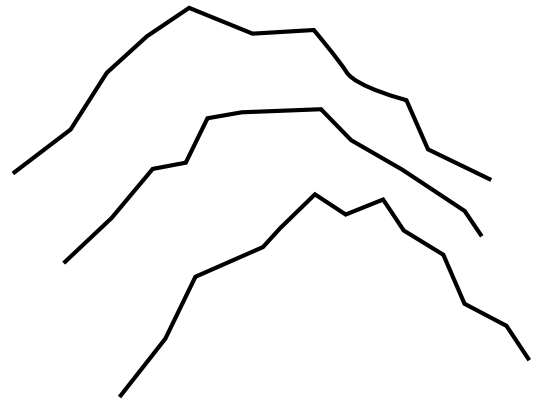


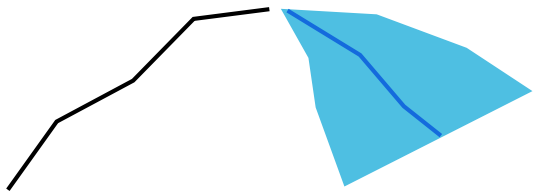


Forecast?

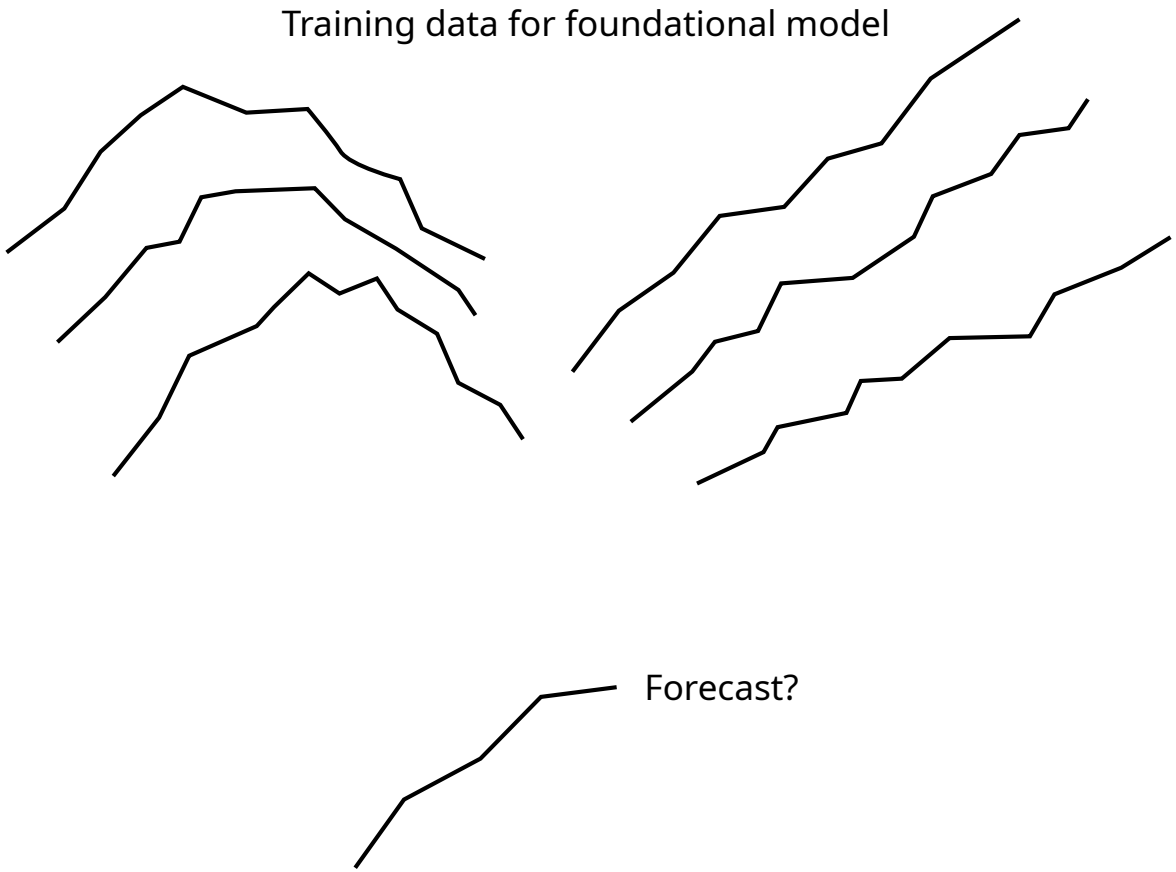
Uncertainty in (implicit) model selection (2)

Training data for foundational model

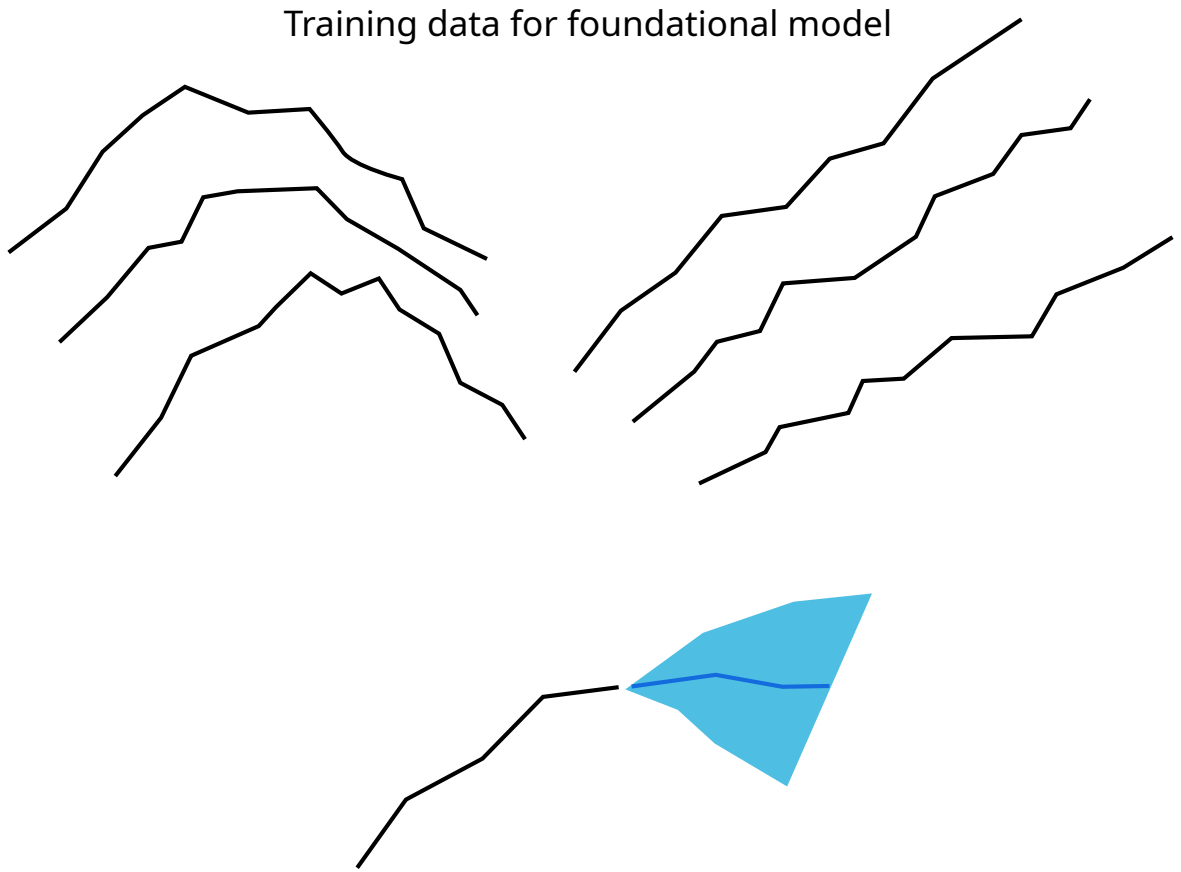




Uncertainty in (implicit) model selection (3)



Uncertainty in (implicit) model selection (4)



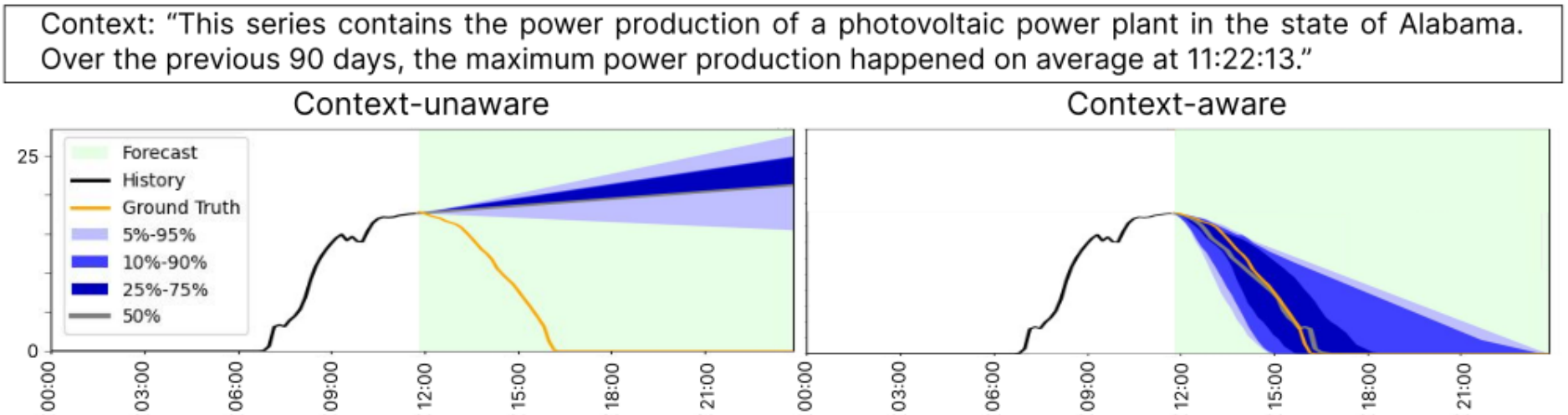
Uncertainty in (implicit) model selection (5)

- Low-granularity time series are frequently noisy, may have very limited endogenous contextual information
- Foundational-style models will perform some sort of implicit model averaging over all implicit sub-models they have built for series sub-groups "
  $\Longrightarrow$ This is precisely how they minimize aggregate error
- Compare the forecast to one made by a (local) linear trend
  - For the first class of data the linear forecaster will be great; for the second class very poor
- The foundation model would have better aggregate error but it is somewhat misleading
  - It will certainly be better from point of view of a publication ...
  - But the linear trend, when used appropriately is much better

The Solution is Context (1)

- To address this, we need to add context
  - So if we want to train on ever more data, we may need to tell the algorithm what every time series is
  - Then, we can include this description in the prompt of the model
  - Context is not just something nice to have for better accuracy
- The fundamental way to get foundational models that are truly foundational
- Not performing better on something on the expense of something else

The Solution is Context (2)



Source: Williams, A. R., Ashok, A., Marcotte, É., Zantedeschi, V., Subramanian, J., Riachi, R., ... & Drouin, A. (2024). "Context is Key: A Benchmark for Forecasting with Essential Textual Information." arXiv preprint arXiv:2410.18959.

Conclusion

- It is always possible to improve aggregate error at the expense of individual error
- As a result, it depends on the training dataset if a particular foundational model will give good results for your use case
- More data will not necessarily make foundational forecasting models better
- Inclusion of context promises a way forward

Thank You!

- This is a cut-down/remixed version of a longer talk given by my colleague Christoph Bergmeir
  - "Fundamental limitations of foundational forecasting models: The need for multimodality and rigorous evaluation",
- Invited talk given at the 2024 NeurIPS workshop "Time Series in the Age of Large Models".
- Has a lot of additional material focusing on evaluation in practice (and how it has affected the literature)
- URL: https://cbergmeir.com/talks/neurips2024/