Backtesting and error metrics for modern time series forecasting

PyData London 2024
June 2024
Kishan Manani

About me

- Machine learning and data science consultant
- PhD Physics: Modelling and large scale time series analysis abnormal heart rhythms
- Online course developer, see courses: trainindata.com/p/forecasting-specialization

Kishan Manani, PhD

KishManani

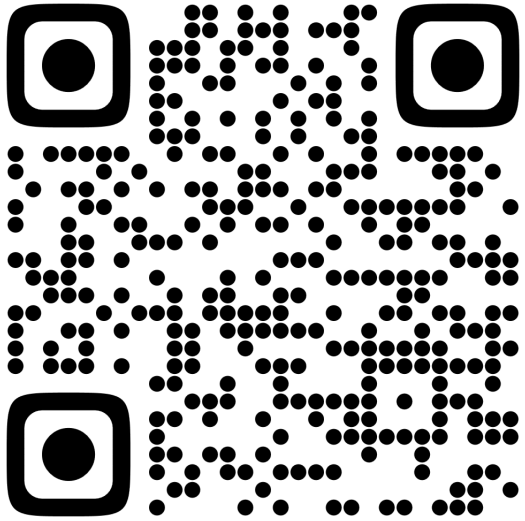
@KishManani

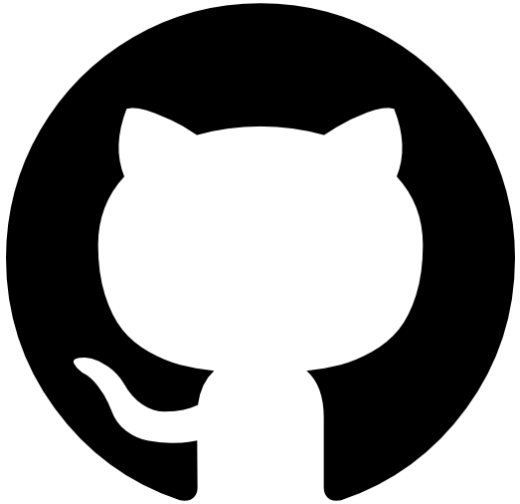
In/kishanmanani

medium.com/@kish.manani

Slides

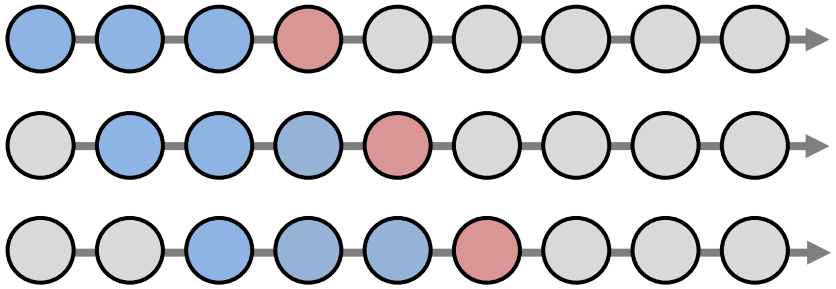
https://github.com/KishManani/PyDataLondon2024





What is this talk about?

Backtesting

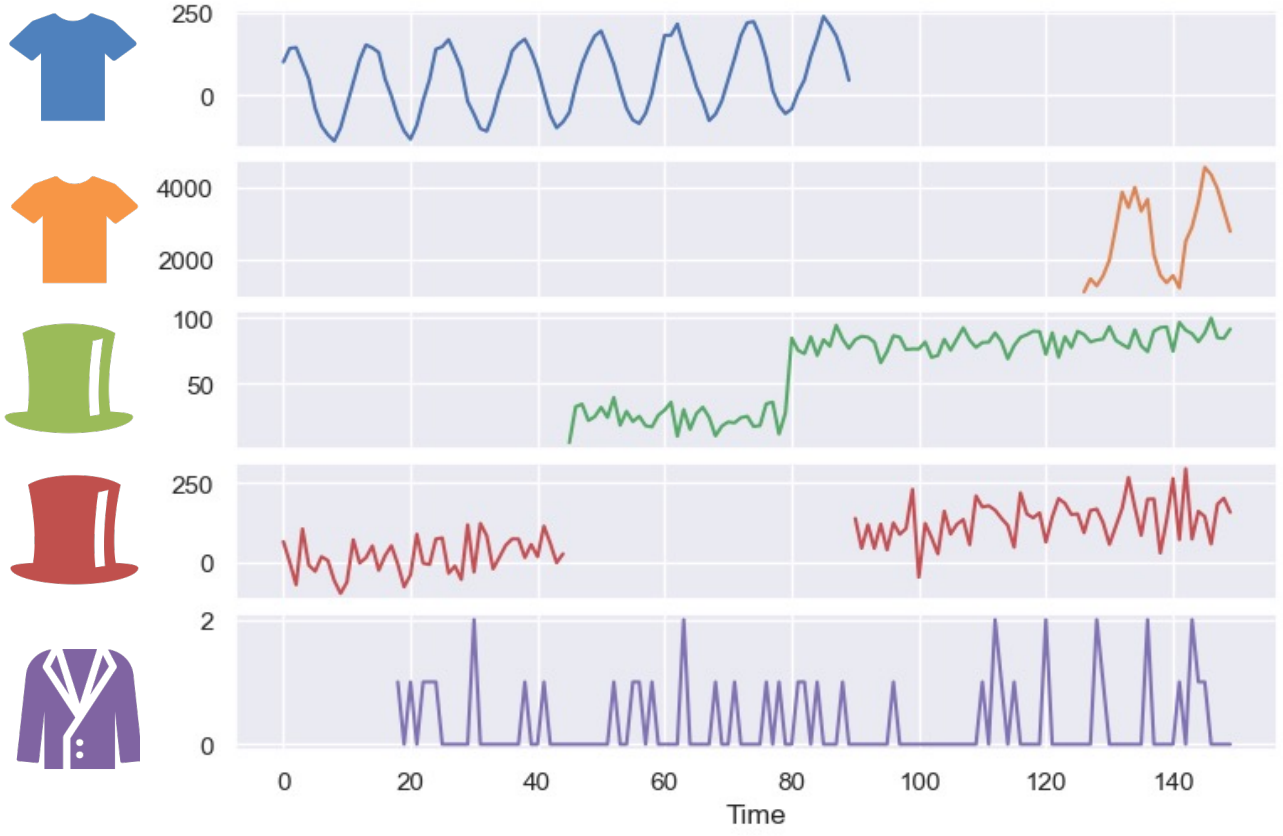


Error metrics

$$MAE = \frac{1}{N}\sum_i |e_i|$$

$$WAPE = \frac{1}{\sum_t w_t}\sum_t w_t |p_t|$$

Number of products sold



Contents

Modern time series forecasting

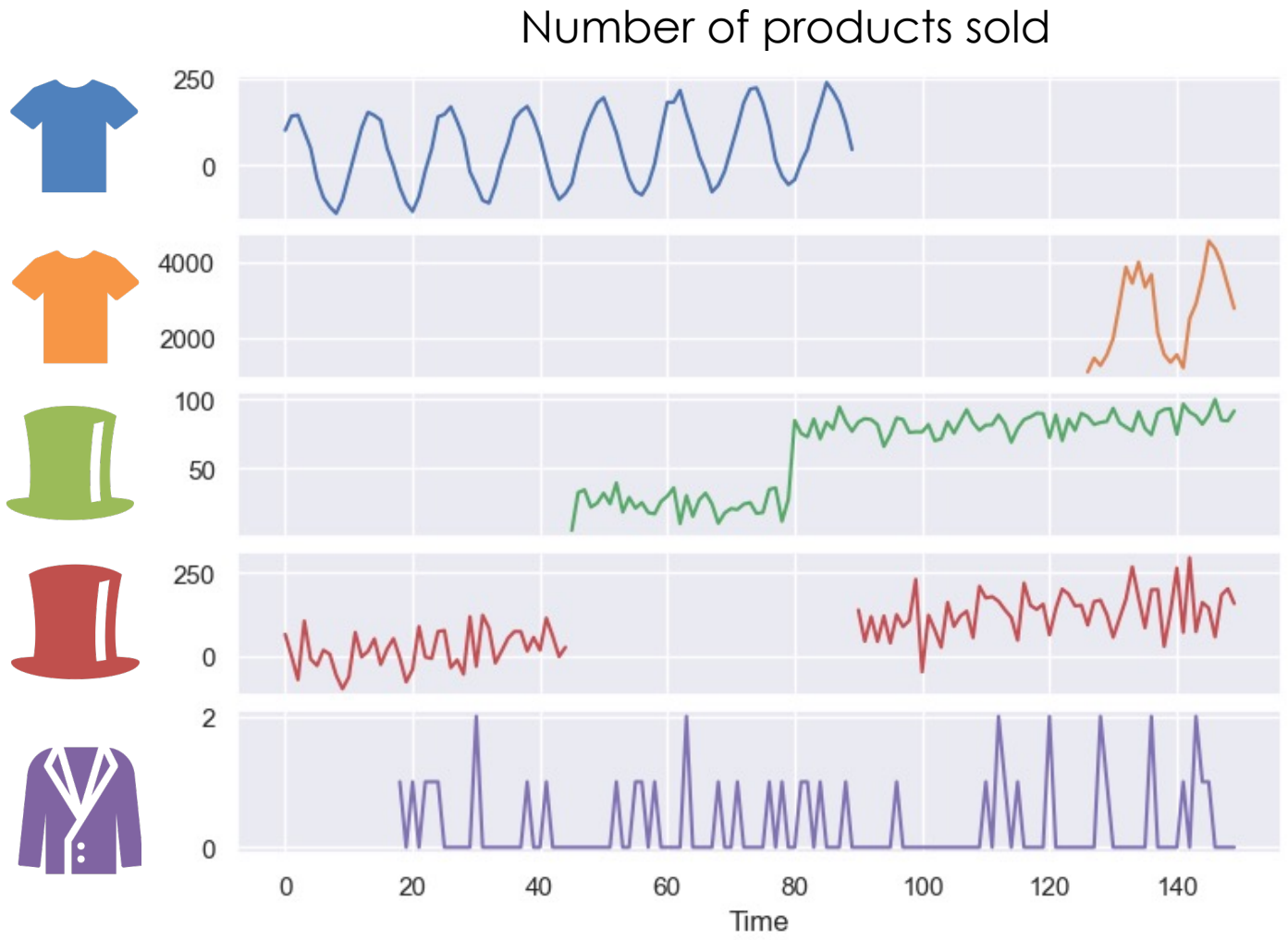
Backtesting

Error metrics

Modern time series forecasting

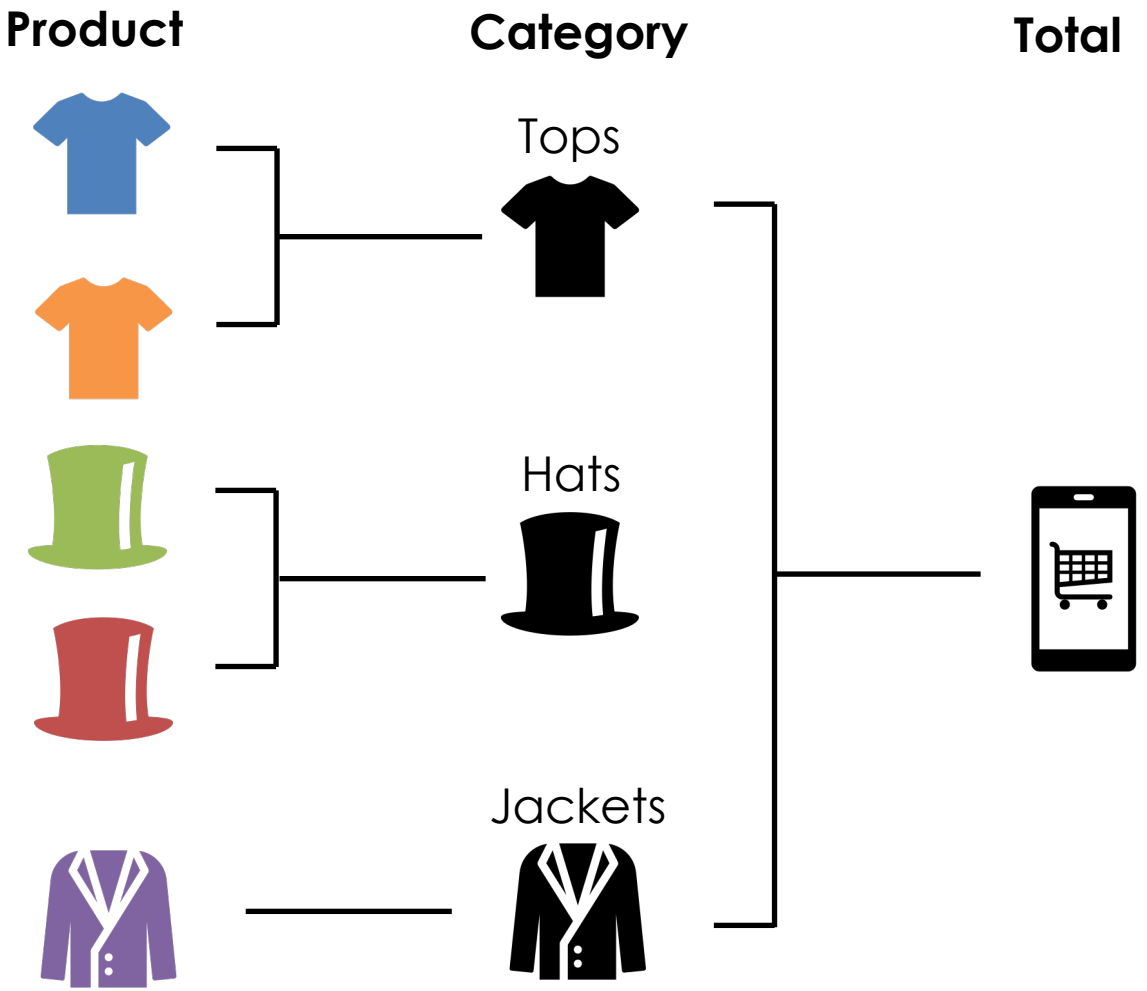
- Large number of related time series

- Multiple characteristics
  - Trend
  - Seasonality
  - Outliers
  - Intermittency
  - Missing data
  - Hierarchical

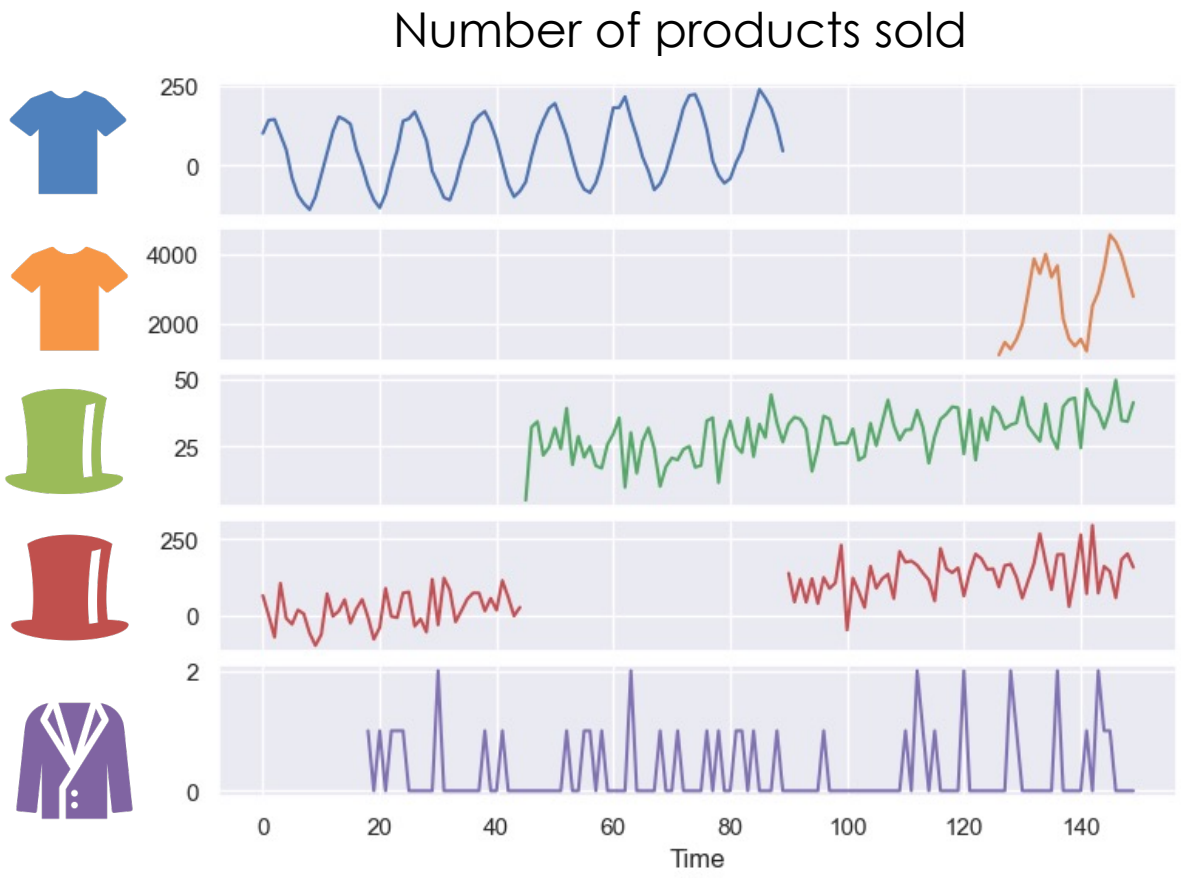


Modern time series forecasting

- Large number of related time series
- Multiple characteristics
  - Trend
  - Seasonality
  - Outliers
  - Intermittency
  - Missing data
  - Hierarchical

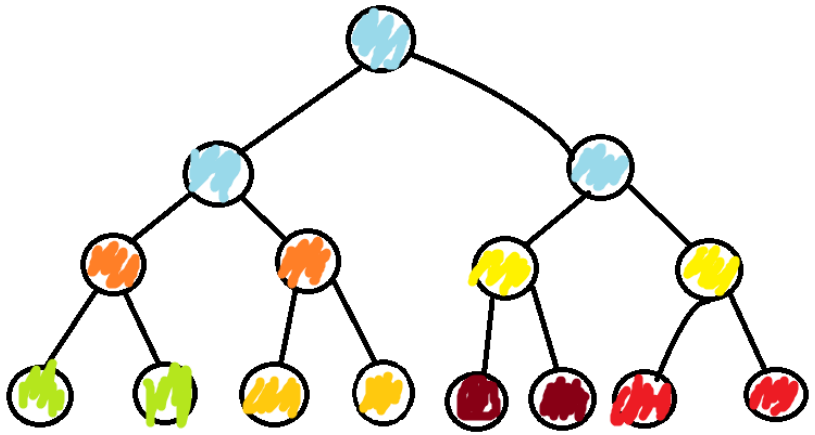


Modern time series forecasting

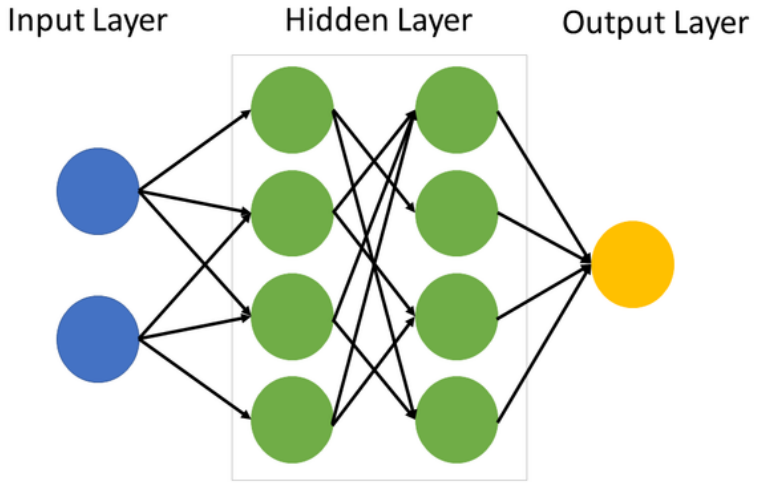


Time	Product ID	Sales
…	…	…
1	1	10
2	1	12
3	1	343
4	1	54
…	…	…
140	2	545
141	2	53
142	2	32
…	…	…

Modern time series forecasting

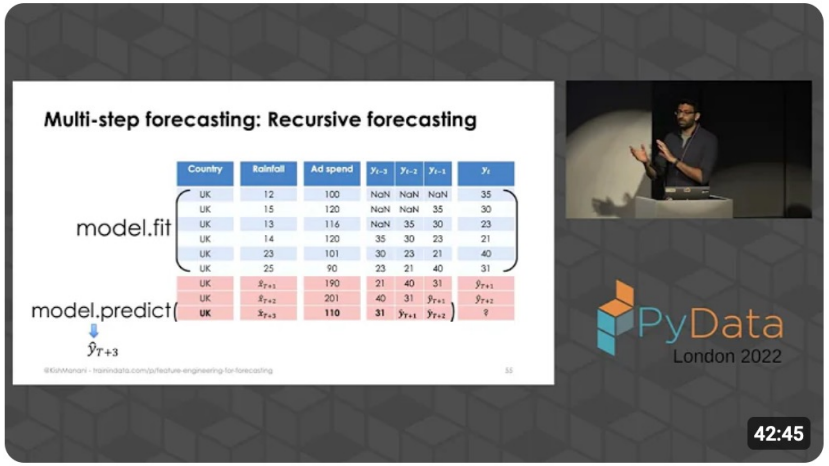


Traditional ML models



Deep learning

Modern time series forecasting



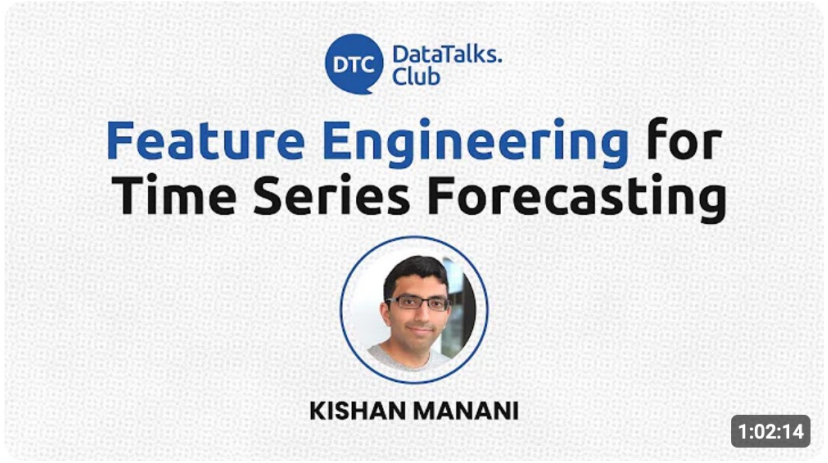
Kishan Manani - Feature Engineering for Time Series Forecasting | PyData London 2022

71K views • 1 year ago

PyData

Kishan Manani present: **Feature Engineering for Time Series Forecasting** To use our favourite supervised learning models…

Intro | About this talk | Why use machine learning for forecasting? | Don't neglect simpl… **22 chapters**



Feature Engineering for Time Series Forecasting - Kishan Manani

23K views • Streamed 1 year ago

DataTalksClub

In this podcast episode, we talked with Kishan Manani about **feature engineering for time series forecasting**. 0:00 Introdu…

Introduction and Welcome | Speaker Introduction | Topic Introduction: Feature… **19 chapters**

Contents

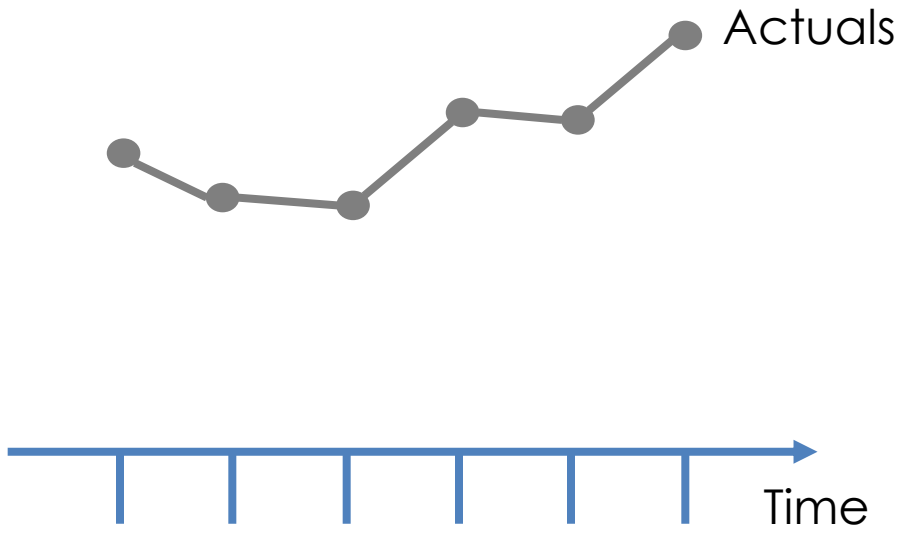
Modern time series forecasting

Backtesting

Error metrics

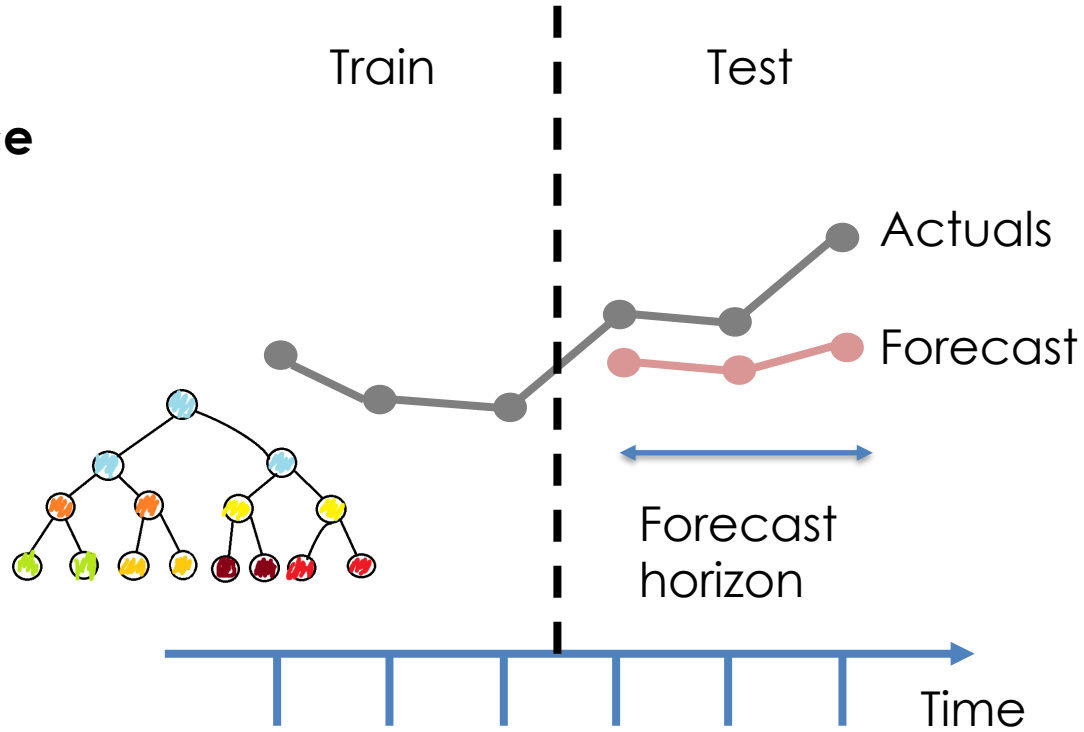
Backtesting

- Backtesting is where we **measure the performance** of a **forecasting model** on **historic data**.
- We want to **estimate the performance** of the model **on future data**.



Backtesting

- Backtesting is where we **measure the performance** of a **forecasting model** on **historic data**.
- We want to **estimate the performance** of the model **on future data**.



Backtesting

- Backtesting is where we **measure the performance** of a **forecasting model** on **historic data**.
- We want to **estimate the performance** of the model **on future data**.
- It is common to randomly shuffle data into train and test sets.

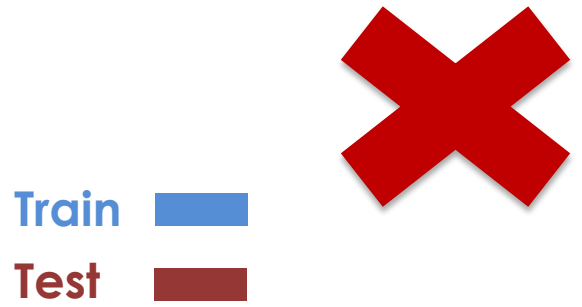
Train

Test

x1	x2	y
…	…	0
…	…	1
…	…	0
…	…	0
…	…	0
…	…	1
…	…	0

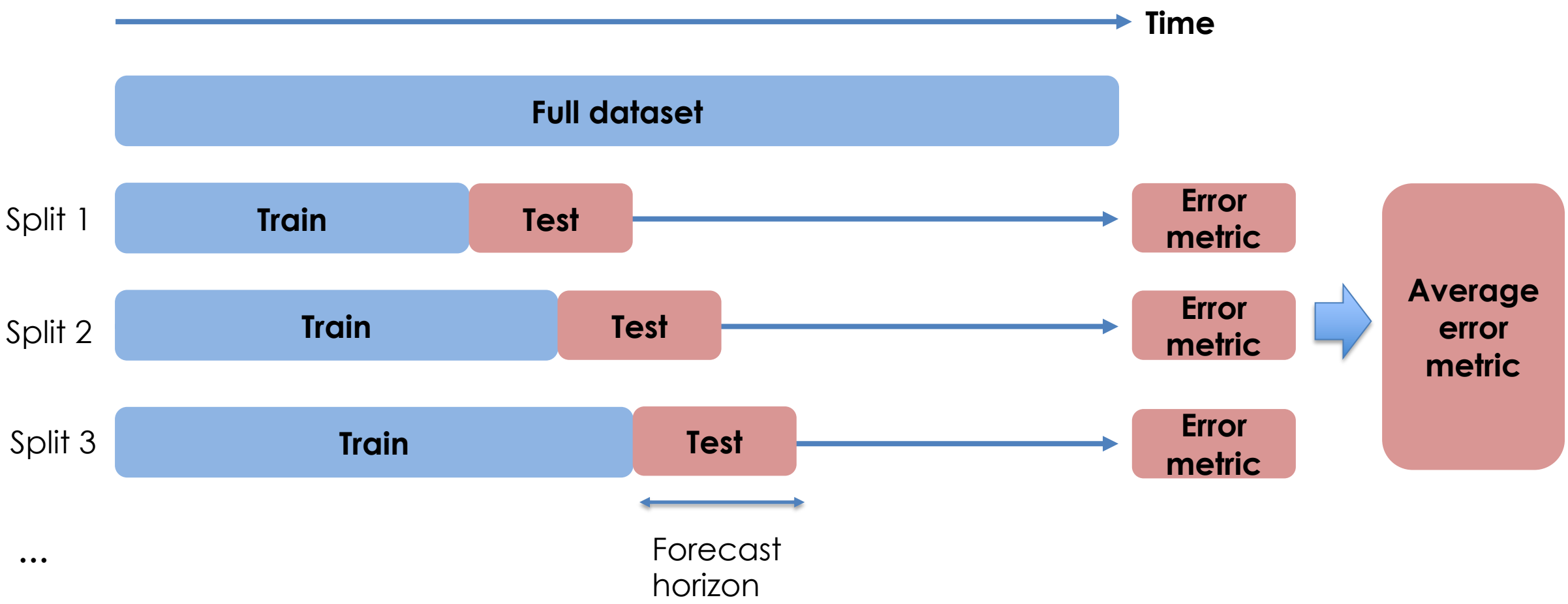
Backtesting

- Backtesting is where we **measure the performance** of a **forecasting model** on **historic data**.
- We want to **estimate the performance** of the model **on future data**.
- It is common to randomly shuffle data into train and test sets.
- We **can't randomly shuffle** the data into train and test sets when working **with time series**.



time	x1	x2	y
1	…	…	0
2	…	…	1
3	…	…	0
4	…	…	0
5	…	…	0
6	…	…	1
7	…	…	0

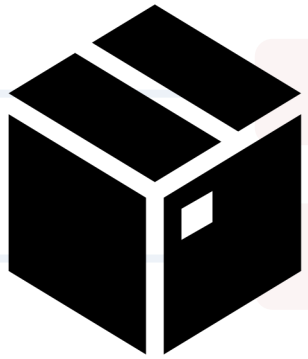
Backtesting



Backtesting

Large dataset with multiple time series

Time consuming



Memory consuming

Error prone

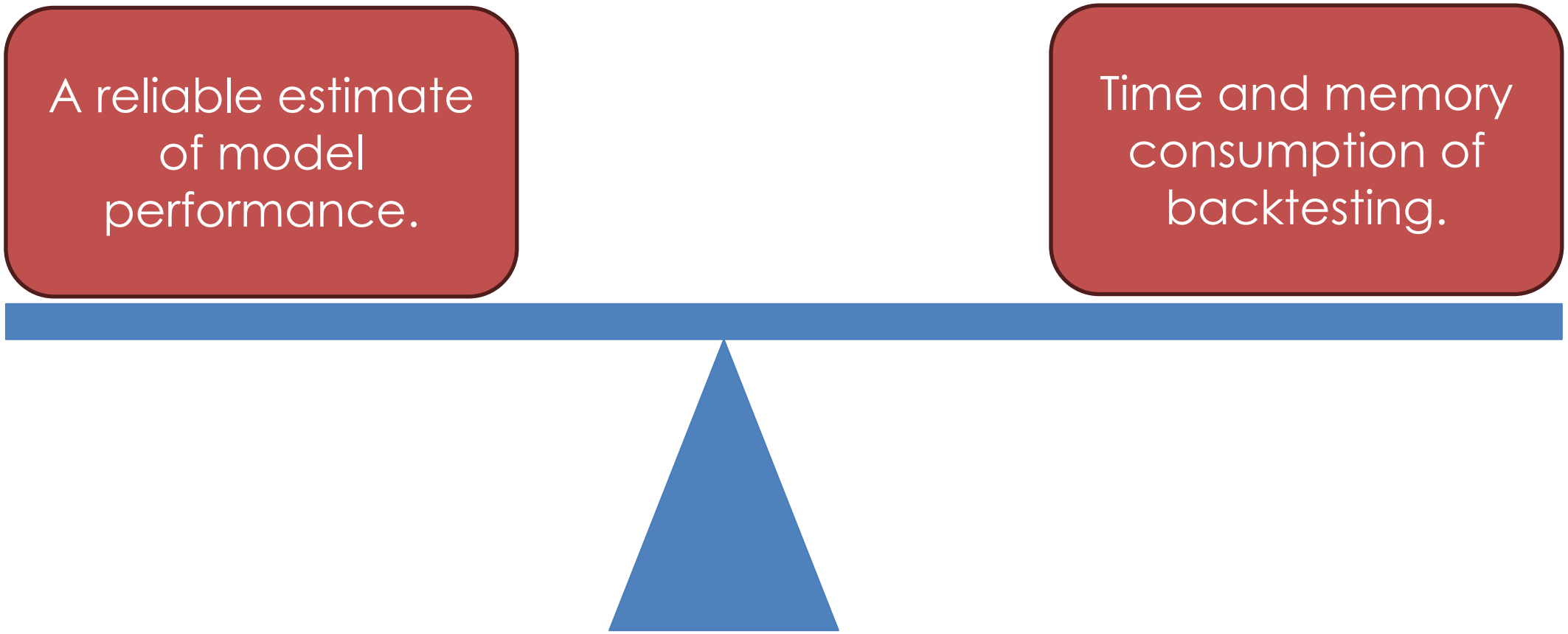
Backtesting parameters

Training window type

Step size

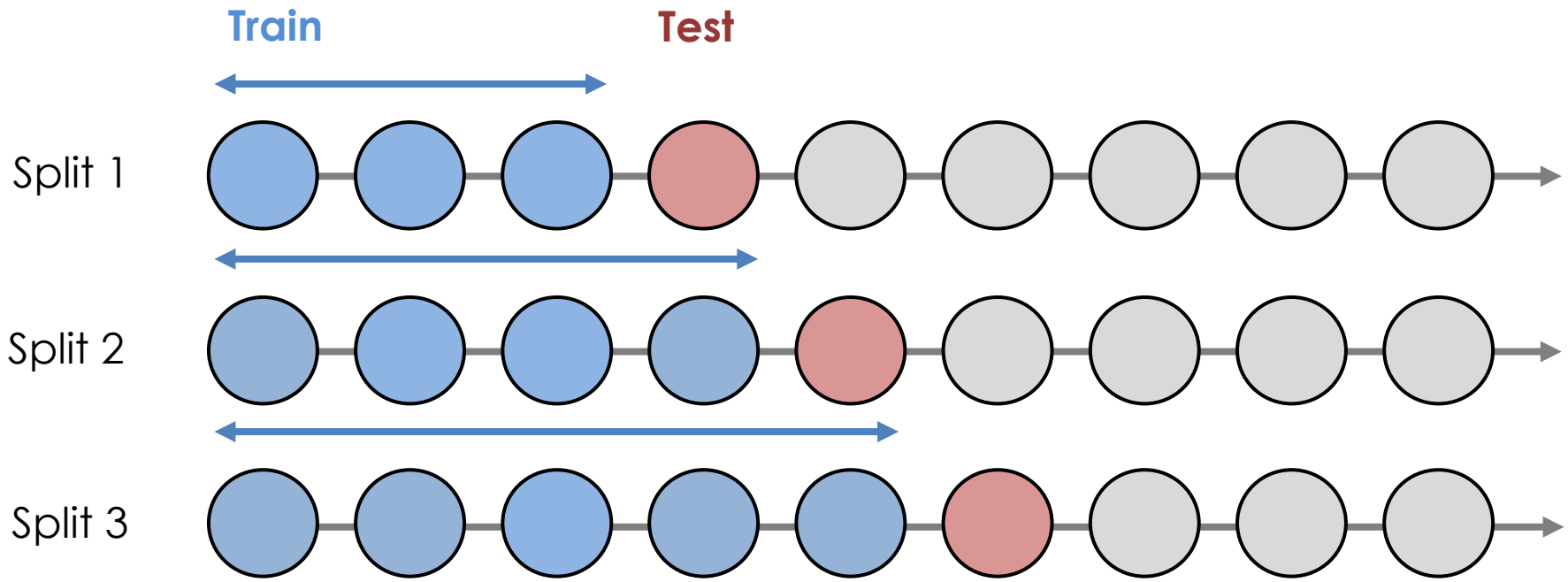
Refitting frequency

Choosing backtesting parameters is a trade-off



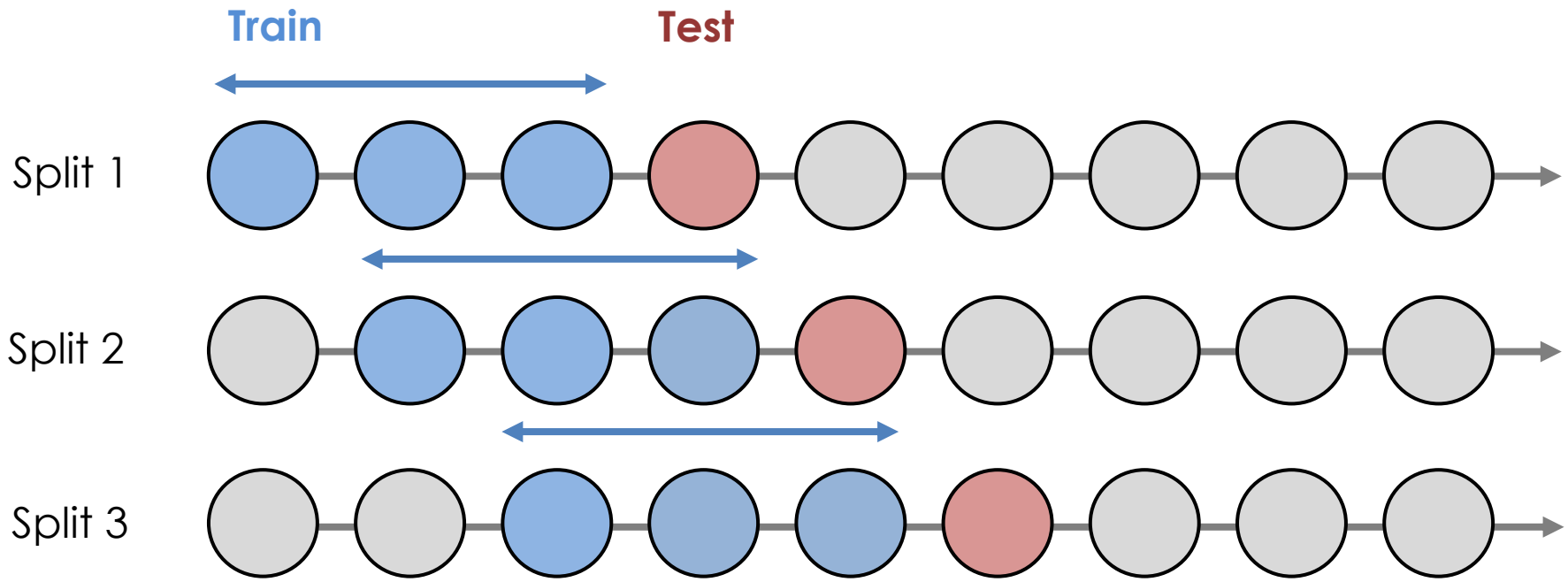
Backtesting parameters: Window type

- **Expanding window**



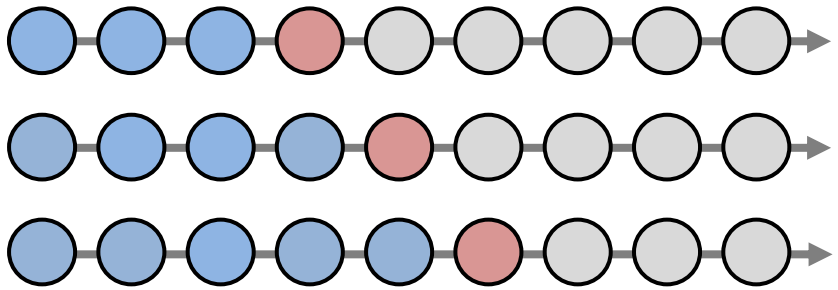
Backtesting parameters: Window type

- **Rolling window (window size is fixed)**



Backtesting parameters: Window type

Expanding window



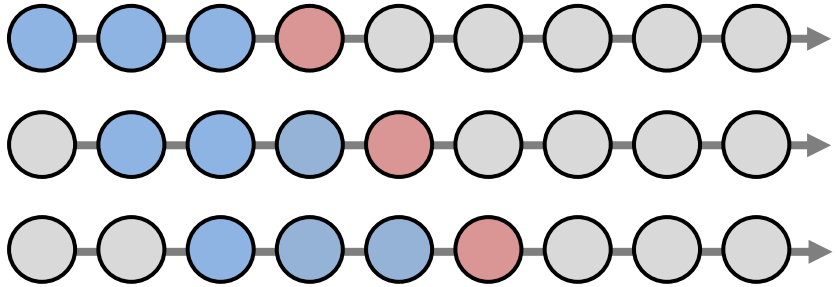
Pros
Uses all the training data.

Cons
More memory used.

Longer to train model.

Very old data may not be that useful.

Rolling window



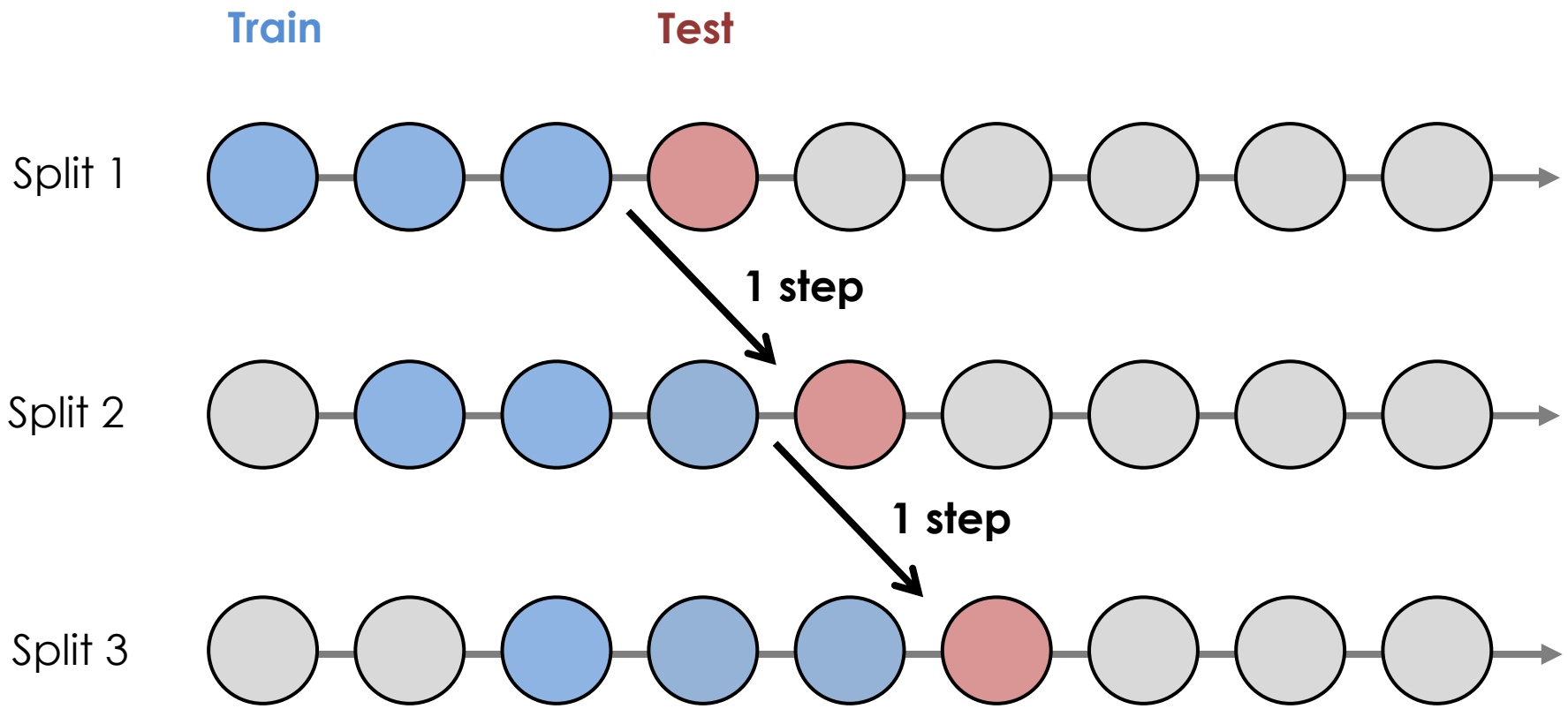
Pros
Less memory used.

Faster to train model.

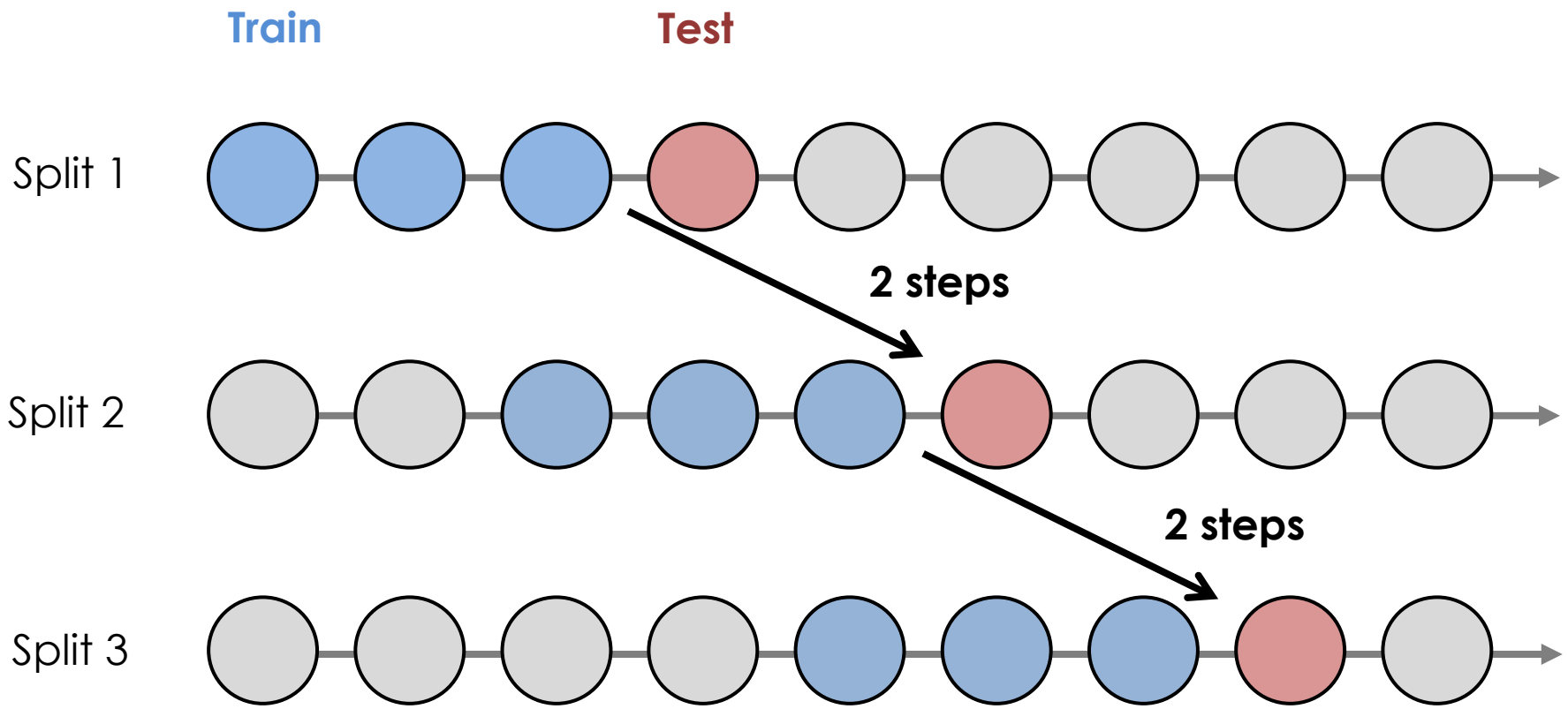
Focuses more on recent data.

Cons
Less training data.

Backtesting parameters: Step size

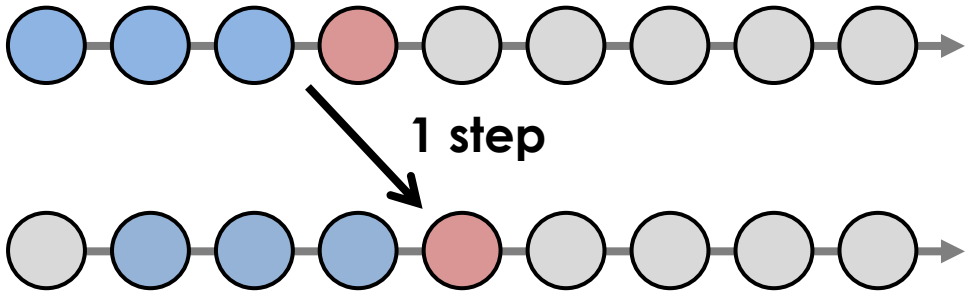


Backtesting parameters: Step size



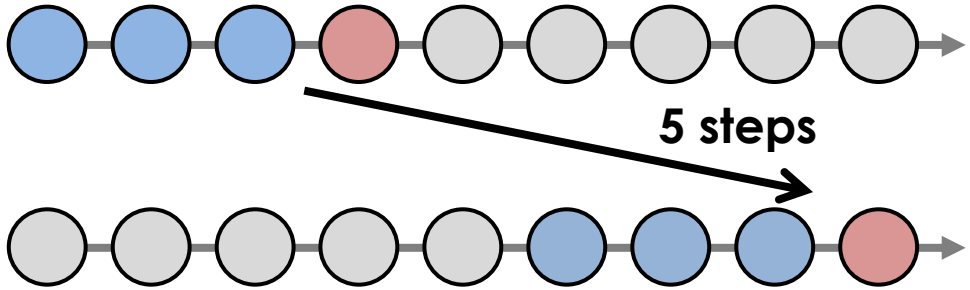
Backtesting parameters: Step size

Small step size



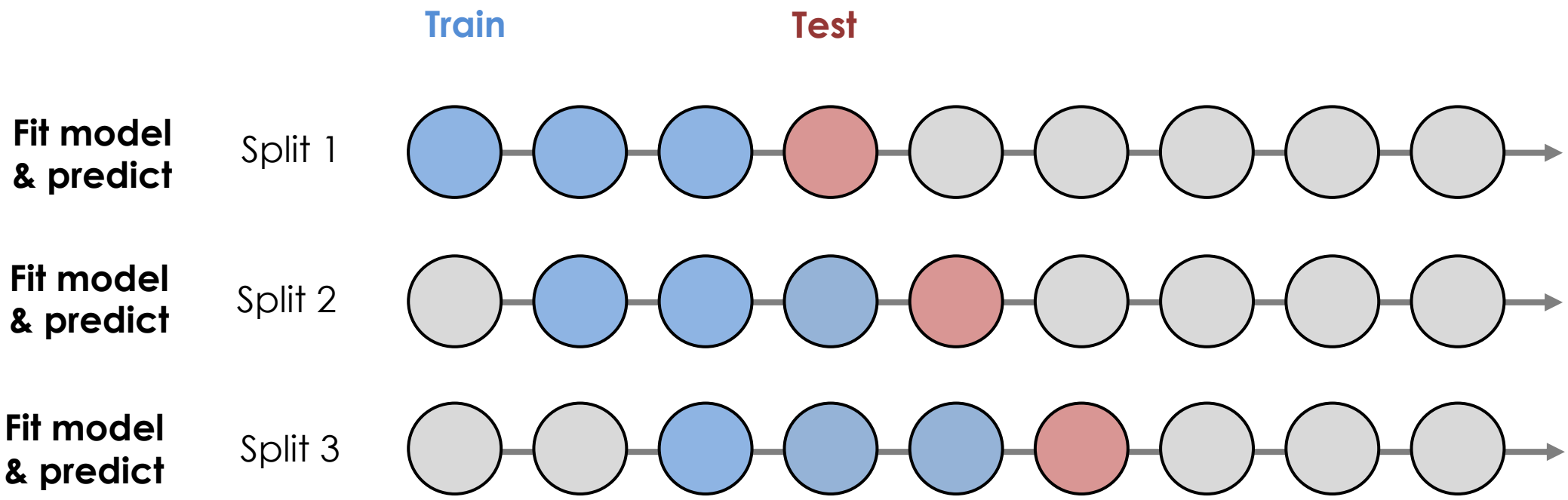
Pros	Cons
Larger # of splits gives better error estimate.	More time and memory used.

Large step size

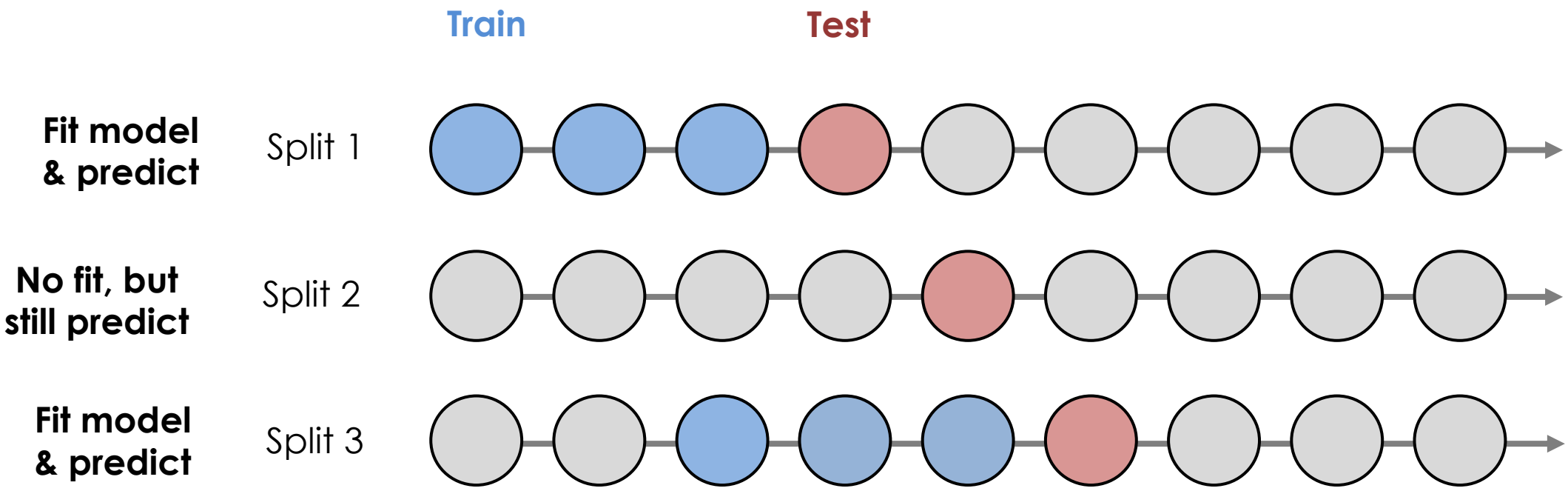


Pros	Cons
Less time and memory used.	Fewer # of splits gives worse error estimate.

Backtesting parameters: Refit frequency

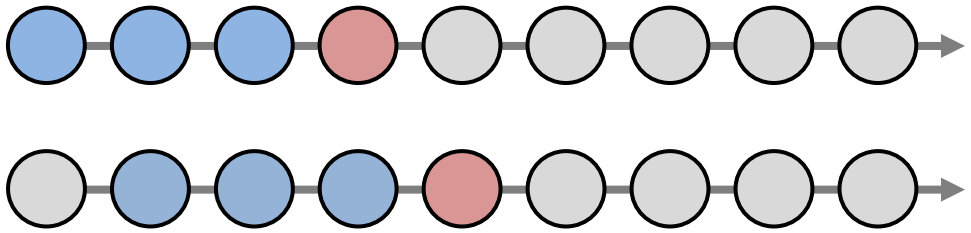


Backtesting parameters: Refit frequency



Backtesting parameters: Refit frequency

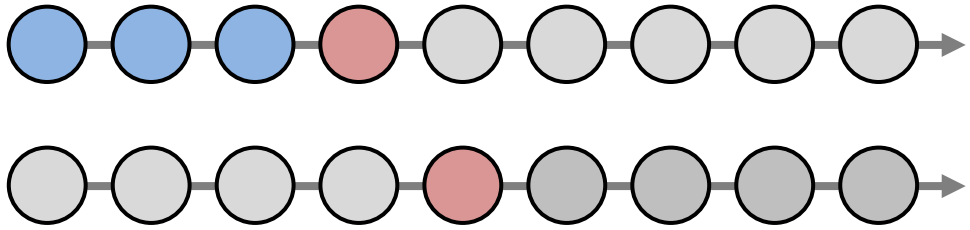
Refit every step



Pros
Better estimate of model performance.

Cons
More time.

Refit periodically

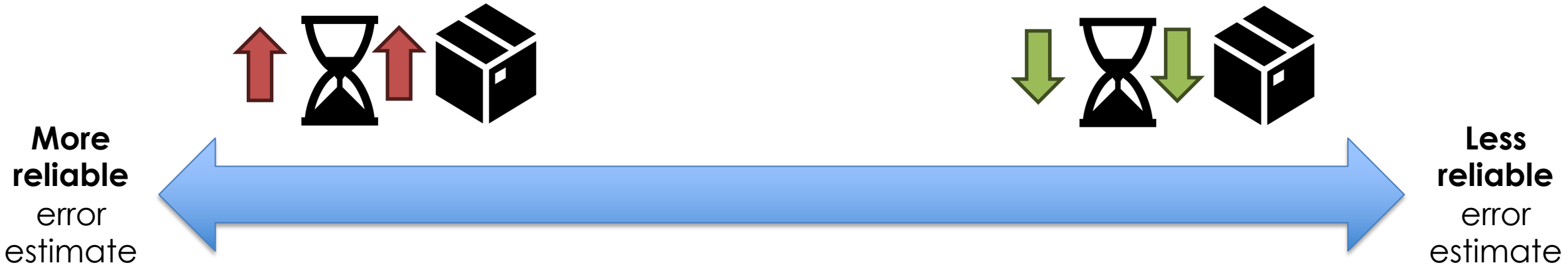


Pros
Less time.

Cons
Worse estimate of model performance.

Choosing backtesting parameters is a trade-off

- We trade-off **reliability** against **time & memory**.
- The trade-off varies for **model selection**, **hyperparameter tuning**, and **error measurement**.



Measure forecasting error accurately.

Model selection

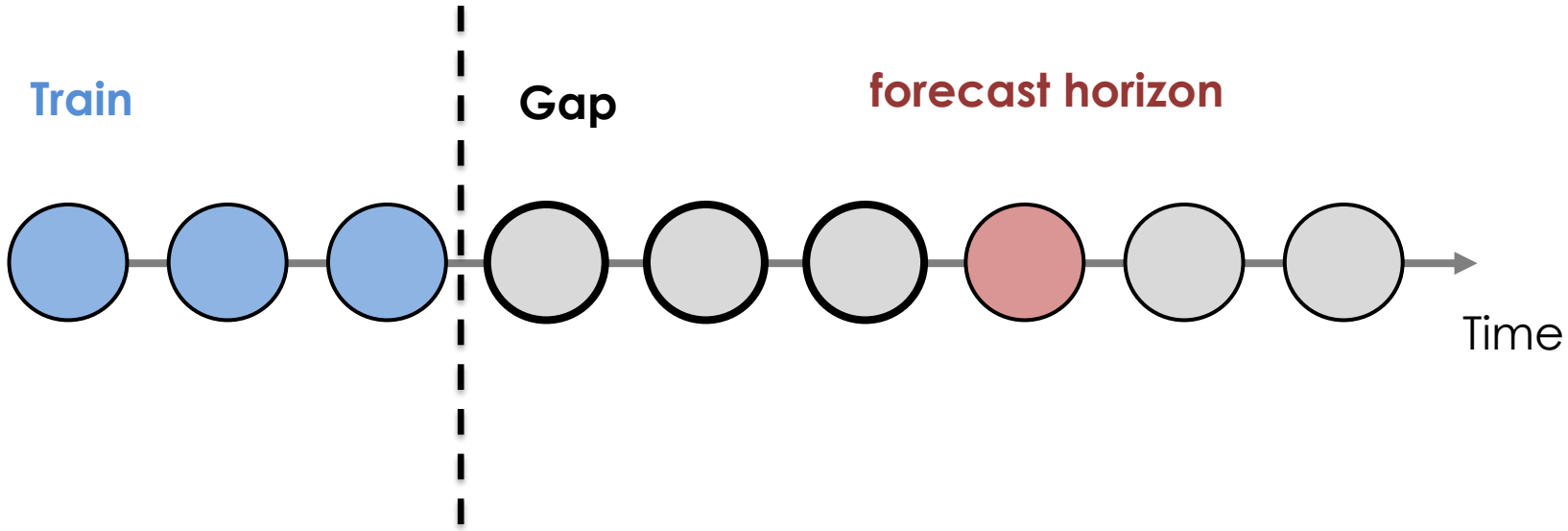
Longer time series

Many time series

Hyperparameter tuning

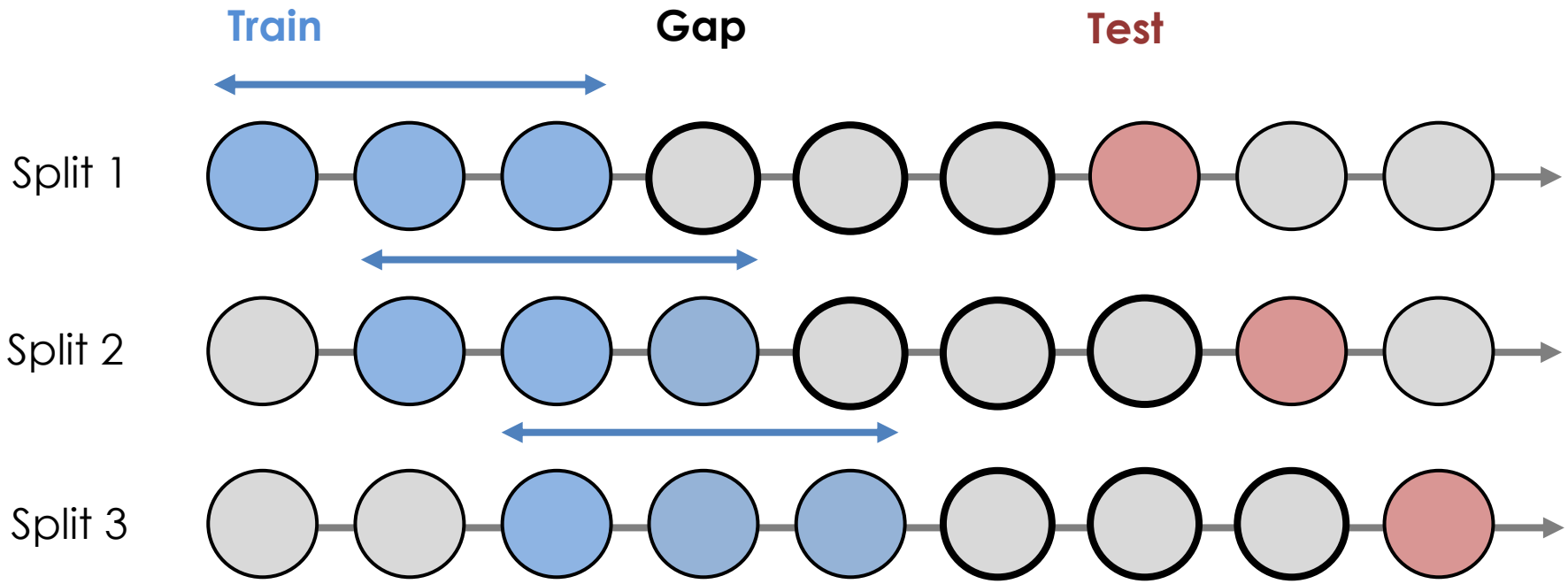
Backtesting: Try to reflect process in production

Example: There is a gap between when the forecast is created relative to when it is used. This gap should be reflected in backtesting.



Backtesting: Try to reflect process in production

Example: There is a gap between when the forecast is created relative to when it is used. This gap should be reflected in backtesting.



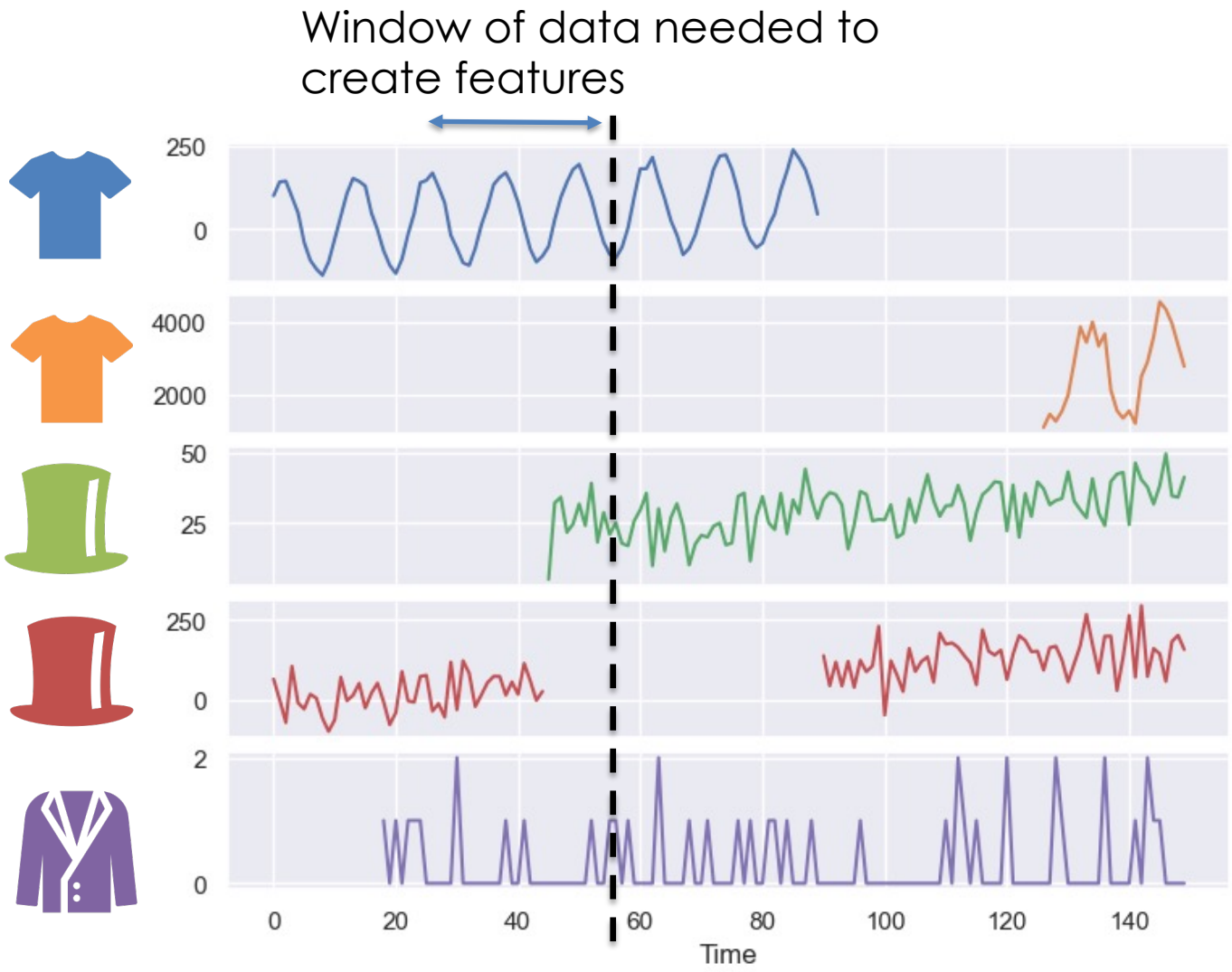
Backtesting: Edge cases for multiple series

Product doesn't exist at train time but does exist at in forecast horizon.

Product does exist at train time but not in the the forecast horizon.

Backtesting: Edge cases for multiple series

Not enough historic data to create features.



Open source libraries make this easier

```
forecaster = ForecasterAutoreg(
                regressor = RandomForestRegressor(random_state=123),
                lags      = 15
             )

metric, predictions = backtesting_forecaster(
                          forecaster            = forecaster,
                          y                     = data['y'],
                          steps                 = 10,
                          metric                = 'mean_squared_error',
                          initial_train_size    = 100,
                          fixed_train_size      = False,
                          gap                   = 0,
                          allow_incomplete_fold = True,
                          refit                 = True,
                          n_jobs                = 'auto',
                          verbose               = True,
                          show_progress         = True
                      )
```

Open source libraries make this easier

```python
from functime.forecasting import linear_model
from functime.metrics import mase

forecaster = linear_model(lags=24, fit_intercept=False, freq="1mo")

y_preds, y_resids = forecaster.backtest(
    y=y_train,
    X=X_train,
    test_size=6,
    step_size=1,
    n_splits=3,
    window_size=1,
    strategy="expanding",
    drop_short=True,
)
```

functime

Forecasting with machine learning in Python

Contents

Modern time series forecasting

Backtesting

Error metrics

There are a lot of forecasting error metrics

Table 8 Error measure definitions in the forecasting literature

Category	Error measure	Definition
Scale-Dependent Measures	Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(e_t^2)}$
	Root Median Squared Error (RMdSE)	$RMdSE = \sqrt{median(e_t^2)}$
	Median Absolute Error (MdAE)	$MdAE = median(\lvert e_t \rvert)$
	Geometric Root Mean Squared Error (GRMSE, Syntetos and Boylan 2005)	$GRMSE = \sqrt[2n]{\prod_{t=1}^{n} e_t^2}$
	Geometric Mean Absolute Error (GMAE)	$GMAE = \sqrt[n]{\prod_{t=1}^{n} \lvert e_t \rvert}$
Measures based on Percentage Errors	Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{n}\sum_{t=1}^{n}(\lvert p_t \rvert)$

Hewamalage, Hansika, Klaus Ackermann, and Christoph Bergmeir. "Forecast evaluation for data scientists: common pitfalls and best practices." *Data Mining and Knowledge Discovery* 37.2 (2023): 788-832.

This review paper documents over 40 metrics.

Why are there so many error metrics?

How do we pick the right error metric(s)?

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.

Mean absolute error

$$MAE = mean(|e_t|)$$

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.
- Each forecast, $\hat{y}_t$, is associated with a **base error**.

Mean absolute error

$$MAE = mean(|e_t|)$$

Scale-dependent base error:

$$e_t = y_t - \hat{y}_t$$

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.
- Each forecast, $\hat{y}_t$, is associated with a **base error**.
- The base error is **transformed** to be **positive**.

Mean absolute error

$$MAE = mean(|e_t|)$$

Scale-dependent base error:

$$e_t = y_t - \hat{y}_t$$

Positive transform:

Absolute value: |x|

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.
- Each forecast, $\hat{y}_t$, is associated with a **base error**.
- The base error is **transformed** to be **positive**.
- The base errors are **aggregated** with a **summary operator.**

Mean absolute error

$$MAE = \boldsymbol{mean}(|e_t|)$$

Scale-dependent base error:

$$e_t = y_t - \hat{y}_t$$

Positive transform:

Absolute value: |x|

Summary operator:

mean

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.
- Each forecast, $\hat{y}_t$, is associated with a **base error**.
- The base error is **transformed** to be **positive**.
- The base errors are **aggregated** with a **summary operator.**

Mean absolute percentage error

$$MAPE = mean(|p_t|)$$

Percentage error:

$$p_t = 100\frac{e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$

Positive transform:

Absolute value: |x|

Summary operator:

mean

Structure of most error metrics

- The properties of an error metric depends on three main components that make a metric.
- Each forecast, $\hat{y}_t$, is associated with a **base error**.
- The base error is **transformed** to be **positive**.
- The base errors are **aggregated** with a **summary operator.**
- Not all error metrics follow this pattern.

Mean absolute percentage error

$$MAPE = mean(|p_t|)$$

Percentage error:

$$p_t = 100\frac{y_t - \hat{y}_t}{y_t}$$

Positive transform:

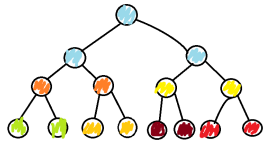
Absolute value: |x|

Summary operator:

mean

How to pick error metrics for forecasting?

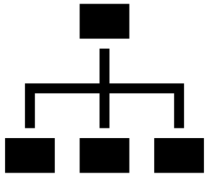
Use case & properties of the time series



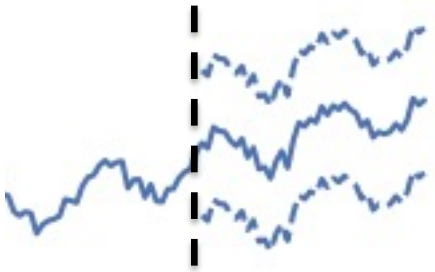
Modelling

Reporting

Does scale matter?



Hierarchy



Over vs under forecasting

Level shifts

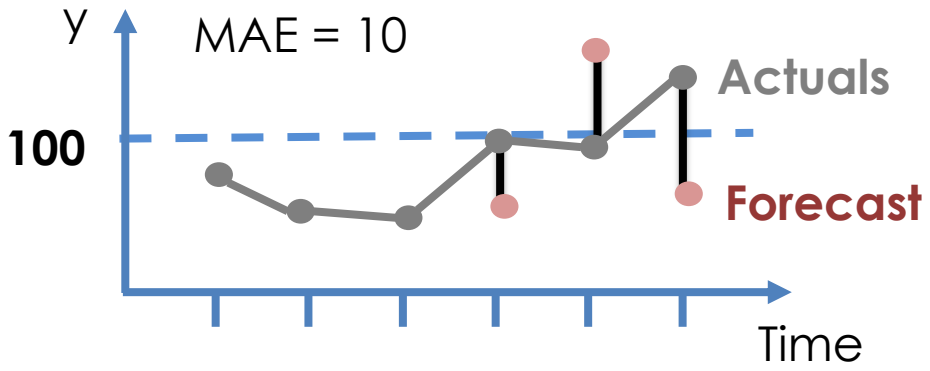
Intermittency

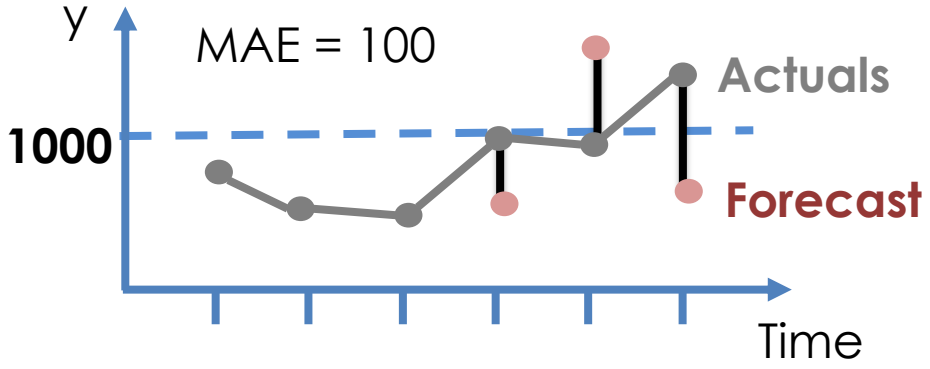
Outliers

Properties of error metrics

$$E(y_t, \hat{y}_t)$$

How to pick error metrics for forecasting?



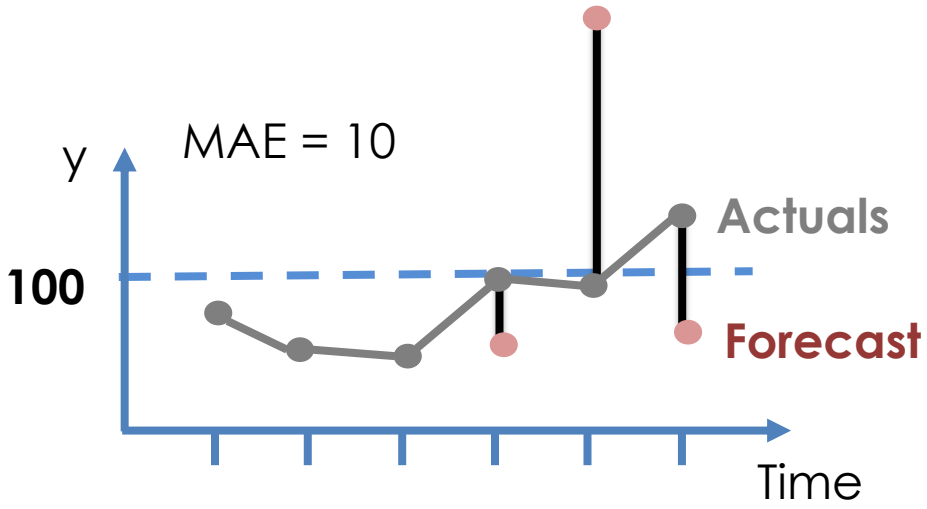


Properties of error metrics

$$E(y_t, \hat{y}_t)$$

- **Scale dependence.**

How to pick error metrics for forecasting?

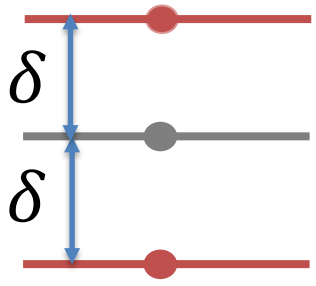


Properties of error metrics

$$E(y_t, \hat{y}_t)$$

- Scale dependence.
- **Sensitive to outliers.**

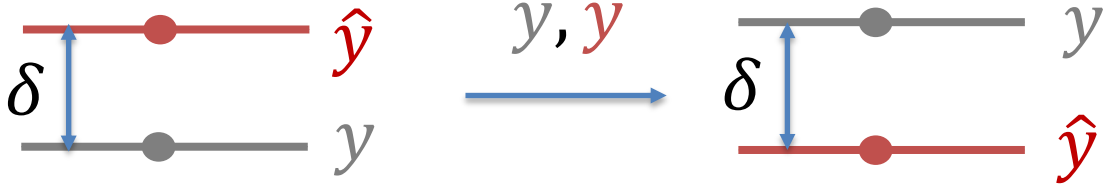
How to pick error metrics for forecasting?



$\hat{y} = y + \delta$ Over forecast

y Actual

$\hat{y} = y - \delta$ Under forecast



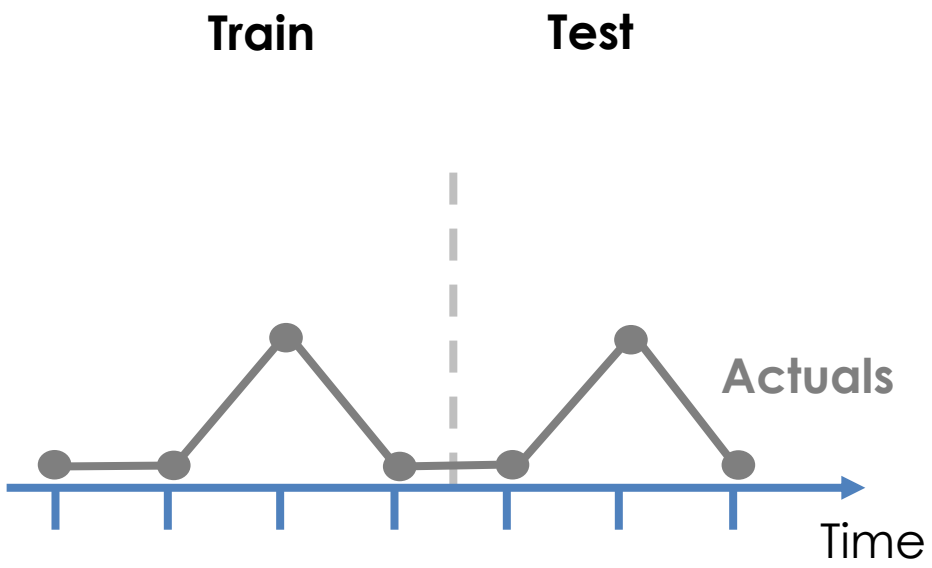
Properties of error metrics

$$E(y_t, \hat{y}_t)$$

- Scale dependence.
- Sensitive to outliers.
- **Symmetric to over and under forecasting.**

How to pick error metrics for forecasting?

Intermittent time series



Properties of error metrics

$$E(y_t, \hat{y}_t)$$

- Scale dependence.
- Sensitive to outliers.
- Symmetric to over and under forecasting.
- **Can it handle zeros?**

How to pick error metrics for forecasting?

The **RMSE**:

$$\sqrt{\frac{1}{N}\sum_{t}(y_t - \hat{y}_t)^2}$$

is **minimized by the mean**:

$$\hat{y}_t = mean(y_t)$$

The **MAE**:

$$\frac{1}{N}\sum_{t}|y_t - \hat{y}_t|$$

is **minimized by the median**:

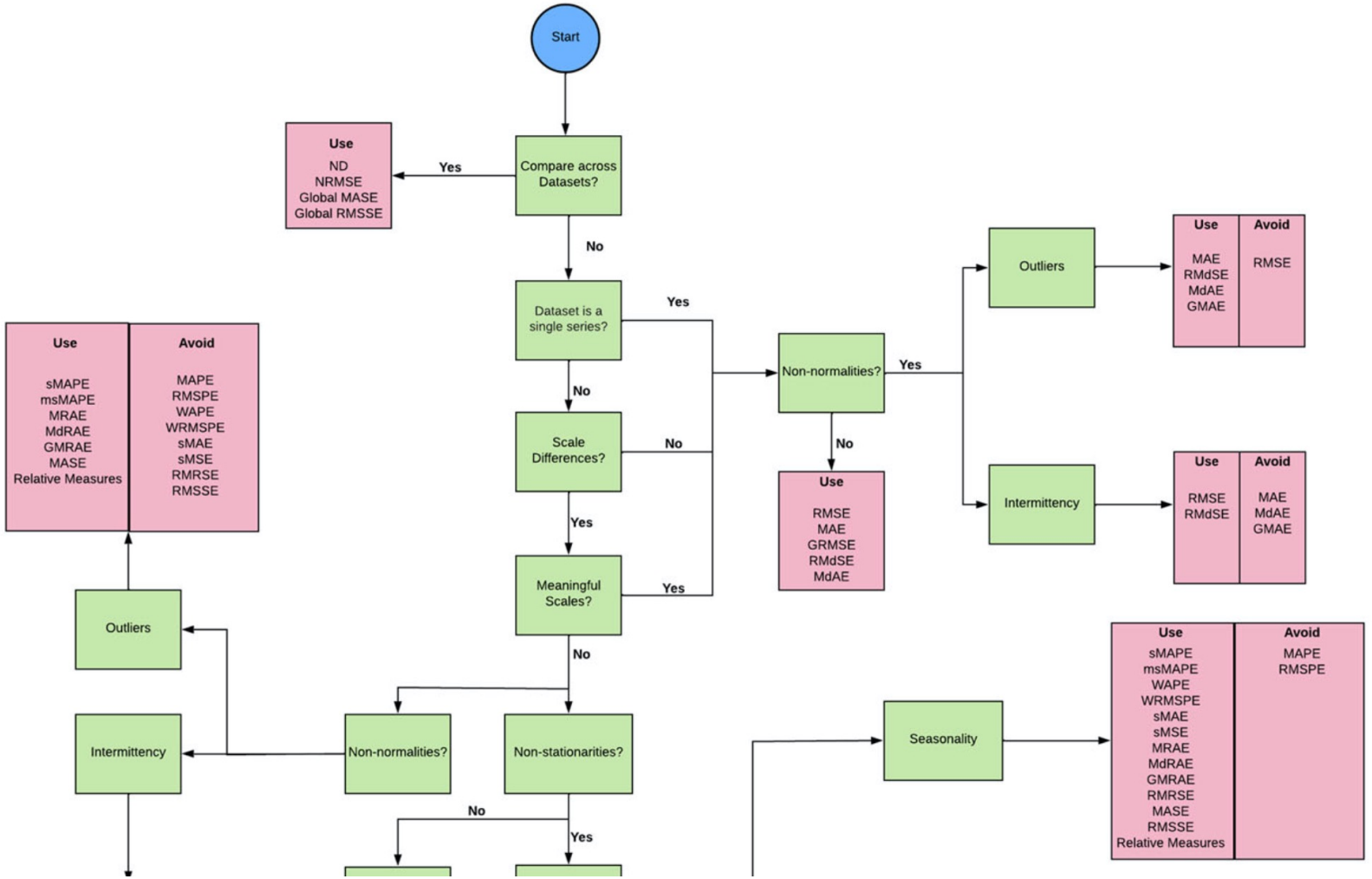
$$\hat{y}_t = median(y_t)$$

Properties of error metrics

$$E(y_t, \hat{y}_t)$$

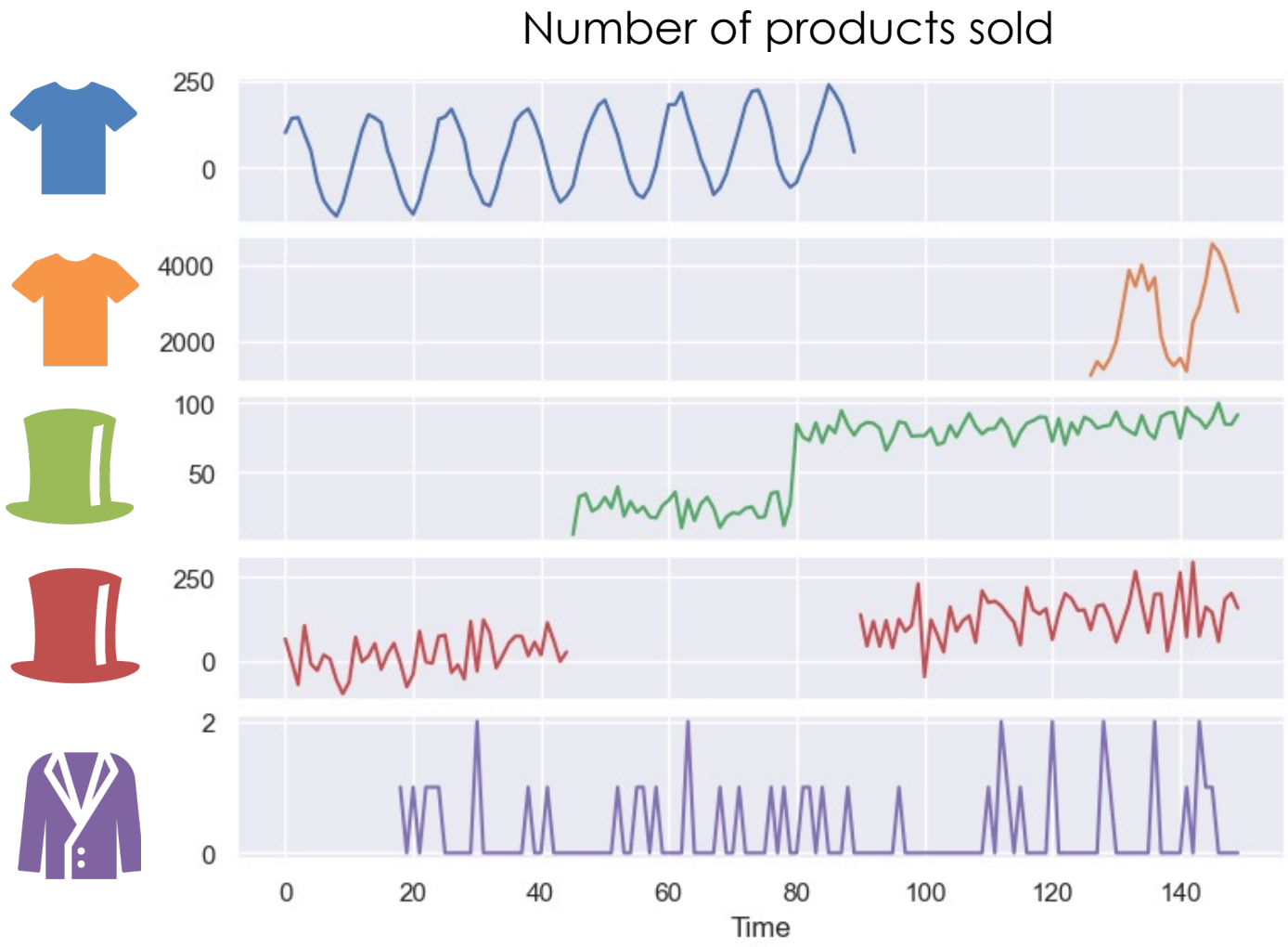
- Scale dependence.
- Sensitive to outliers.
- Symmetric to over and under forecasting.
- Can it handle zeros?
- **Optimise for mean, median, etc.**

Guidelines



Hewamalage, Hansika, Klaus Ackermann, and Christoph Bergmeir. "Forecast evaluation for data scientists: common pitfalls and best practices." *Data Mining and Knowledge Discovery* 37.2 (2023): 788-832.

Multiple time series where the scale matters



We want:

- Larger time series to contribute more to the error (i.e., scale dependence).
- Symmetry to over/under forecasting.
- Tolerate zeros.
- To be able to compare errors for different subsets of timeseries at different scales.
- RMSE and MAE cover **almost** all these requirements.

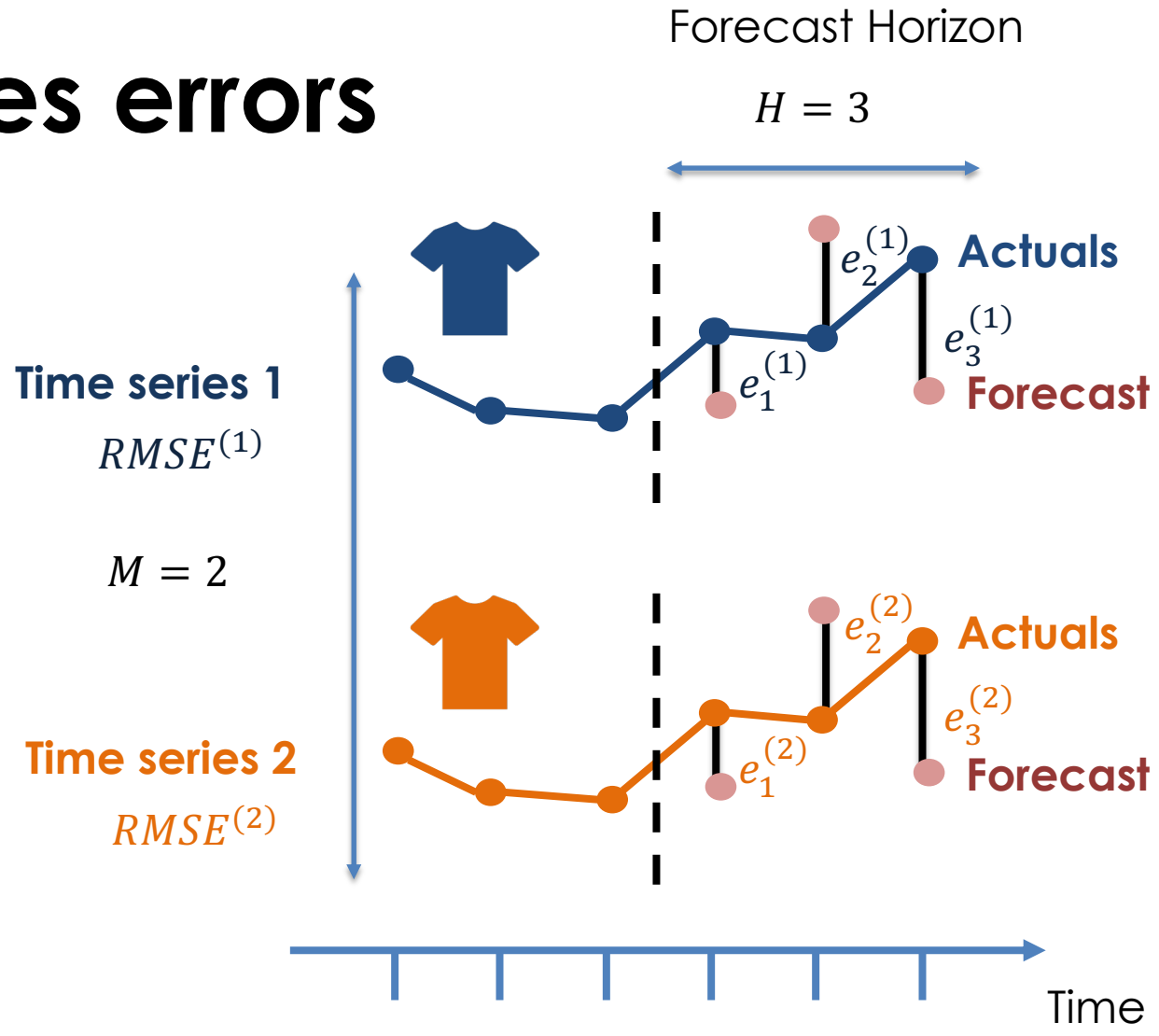
Multiple time series errors

Averaging

$$mean\ RMSE = \frac{1}{M}\sum_{i=1}^{M} RMSE^{(i)}$$

Pooling

$$RMSE = \sqrt{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}\left(e_t^{(i)}\right)^2}$$



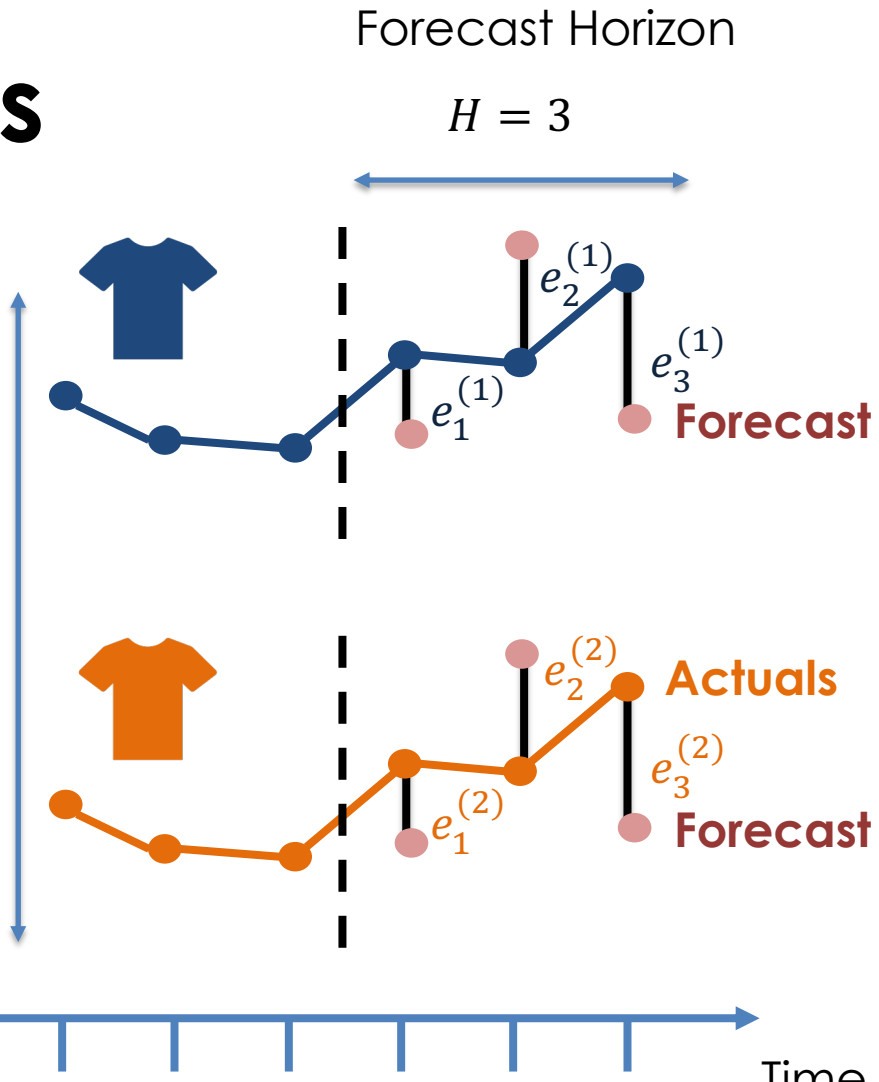
Multiple time series errors

Averaging

$$mean\ RMSE = \frac{1}{2}(RMSE^{(1)} + RMSE^{(2)})$$

Pooling

$$RMSE = \sqrt{\frac{1}{2 * 3}(e_1^{(1)^2} + e_2^{(1)^2} + e_3^{(1)^2} + e_1^{(2)^2} + e_2^{(2)^2} + e_3^{(2)^2})}$$



Multiple time series errors

Pool scale-dependent errors over multiple time series and **scale by mean of all the time series**.

Normalised RMSE

$$NRMSE = \frac{\sqrt{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}(e_t^{(i)})^2}}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{RMSE}{\overline{|y|}}$$

Normalised Deviation
(aka normalized MAE)

$$ND = \frac{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|e_t^{(i)}|}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{MAE}{\overline{|y|}}$$

Multiple time series errors

Larger time series contribute more to the error than smaller ones.

Normalised RMSE

$$NRMSE = \frac{\sqrt{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}(e_t^{(i)})^2}}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{RMSE}{\overline{|y|}}$$

Normalised Deviation
(aka normalized MAE)

$$ND = \frac{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|e_t^{(i)}|}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{MAE}{\overline{|y|}}$$

Multiple time series errors

Symmetric to over/under forecasting and can tolerate zeros.

Normalised RMSE

$$NRMSE = \frac{\sqrt{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}(e_t^{(i)})^2}}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{RMSE}{\overline{|y|}}$$

Normalised Deviation
(aka normalized MAE)

$$ND = \frac{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|e_t^{(i)}|}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{MAE}{\overline{|y|}}$$

Multiple time series errors

These metrics can be used to **compare errors between different subsets/datasets** of **time series**.

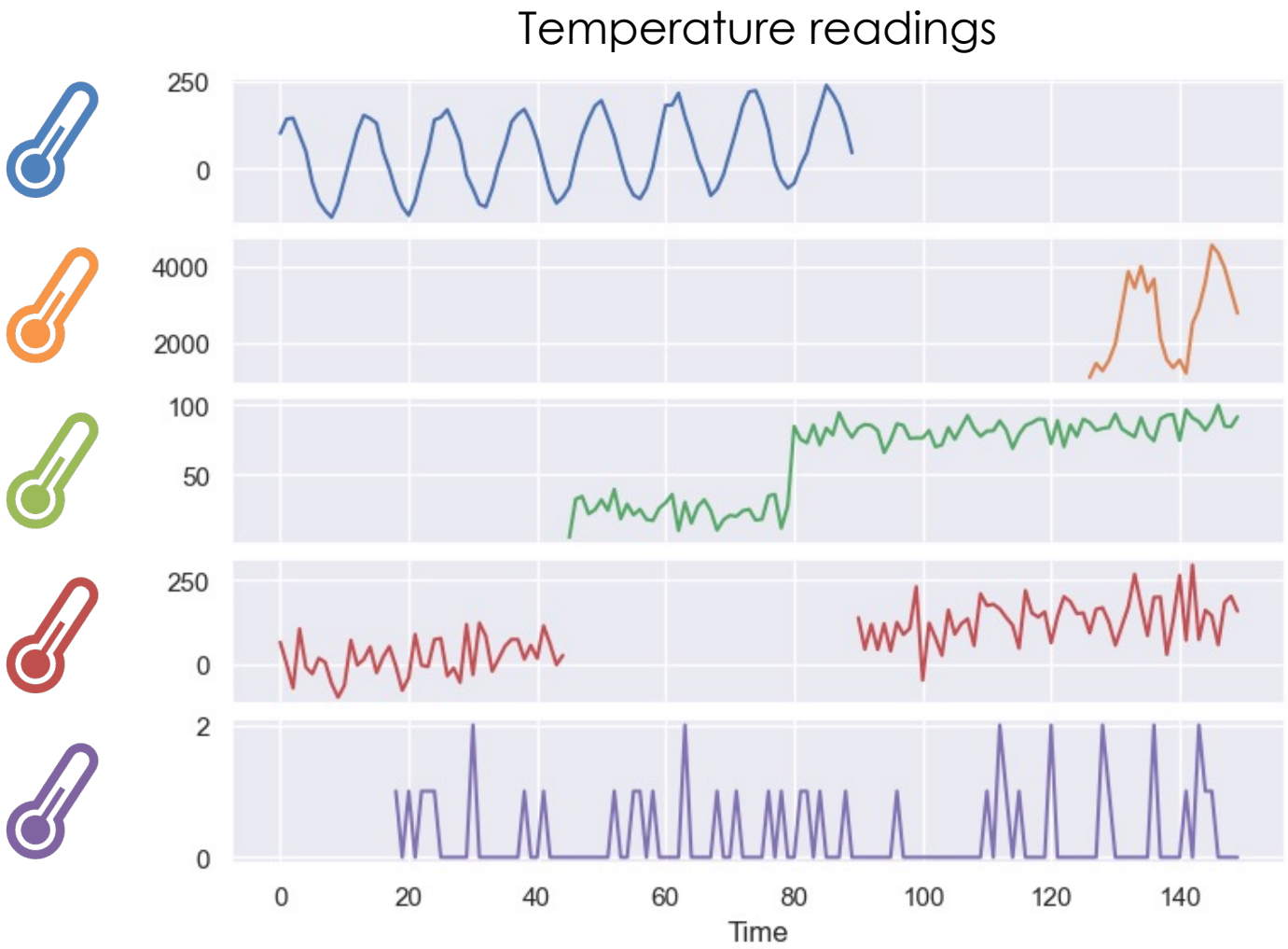
Normalised RMSE

$$NRMSE = \frac{\sqrt{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}(e_t^{(i)})^2}}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{RMSE}{\overline{|y|}}$$

Normalised Deviation
(aka normalized MAE)

$$ND = \frac{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|e_t^{(i)}|}{\frac{1}{MH}\sum_{i=1}^{M}\sum_{t=1}^{H}|y_t^{(i)}|} = \frac{MAE}{\overline{|y|}}$$

Multiple time series where the scale is arbitrary

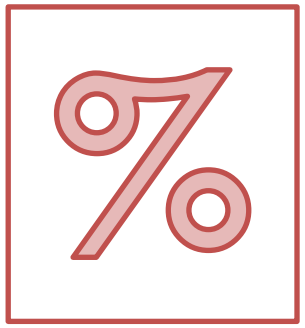


We want:

- Each time series to contribute equally to the error regardless of scale.
- Symmetry to over/under forecasting.
- Tolerate zeros.
- To be able to compare errors for different subsets of timeseries.
- **We need to use scale-independent error metrics.**

Scale-independent error metrics

Percentage error



$$p_t = \frac{e_t}{y_t} = \frac{y_t - \hat{y}_t}{y_t}$$

Weighted mean absolute percentage error

$$WAPE = \frac{1}{\sum_t w_t} \sum_{t=1}^{H} w_t |p_t|$$

$$\xrightarrow{w_t = |y_t|} = \frac{1}{\sum_t |y_t|} \sum_{t=1}^{H} |e_t|$$

$$= \frac{MAE}{mean(|y_t|)}$$

Scale independent when comparing series. ✅

Interpretable. ✅

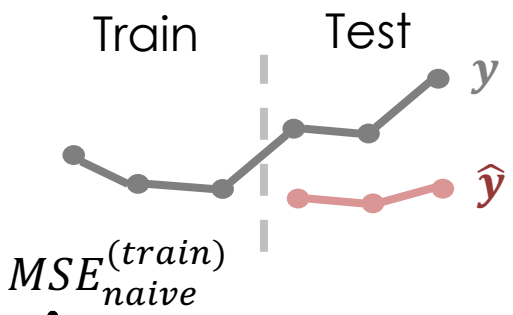
Symmetric to over/under forecasting ✅

Well defined when $\hat{y}_t = 0$ or some $y_t = 0$. ✅

Not well defined when all $y_t = 0$ in horizon. ❌

Scale-independent error metrics

Scaled error



$$q_t = \frac{e_t}{S} = \frac{y_t - \hat{y}_t}{S}$$

Mean absolute scaled error

$$MASE = mean(|q_t|)$$

$$S = MAE_{naive}^{(train)}$$

Root mean squared scaled error

$$RMSSE = \sqrt{mean(q_t^2)}$$

$$S = MSE_{naive}^{(train)}$$

Scale independent. ✅

Symmetric to over/under forecasting ✅

Well defined when $\hat{y}_t = 0$ or $y_t = 0$. ✅

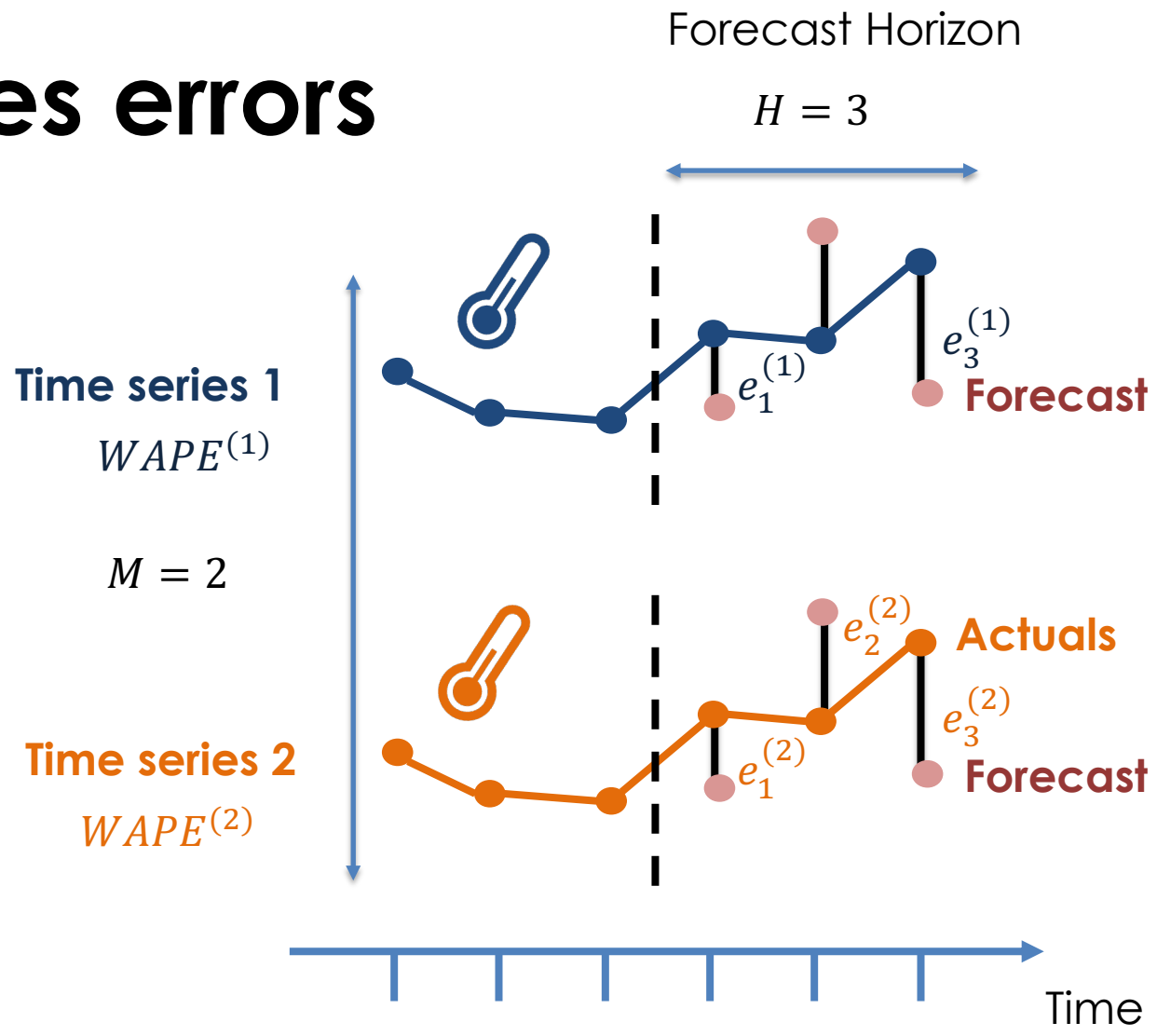
Not well defined when all $y_t = 0$ in train set. ❌

Multiple time series errors

Average the scale-independent error metrics

$$mean\ WAPE = \frac{\sum_{i=1}^{M} WAPE^{(i)}}{M}$$

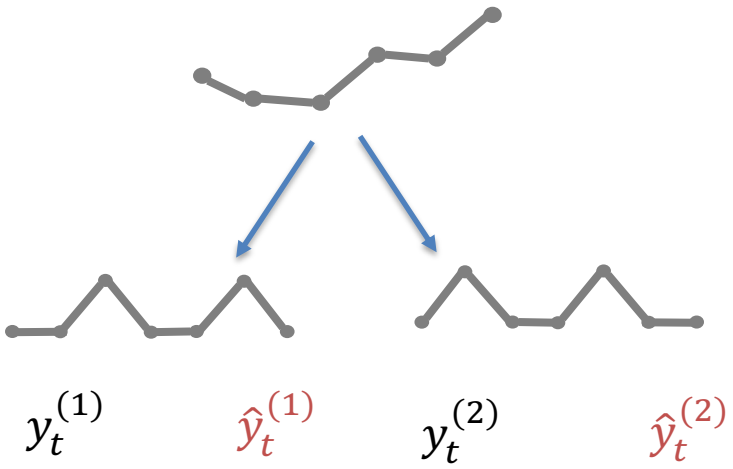
$$= \frac{1}{2}(WAPE^{(1)} + WAPE^{(2)})$$



Hierarchical forecasts

Coherent forecast

$z_t = y_t^{(1)} + y_t^{(2)}$ $\quad \hat{z}_t = \hat{y}_t^{(1)} + \hat{y}_t^{(2)}$



We want the most accurate forecasts at all levels.

We pick an error metric E.

We create a forecast to minimise $E(y_t^{(1)}, \hat{y}_t^{(1)})$ and $E(y_t^{(2)}, \hat{y}_t^{(2)})$.

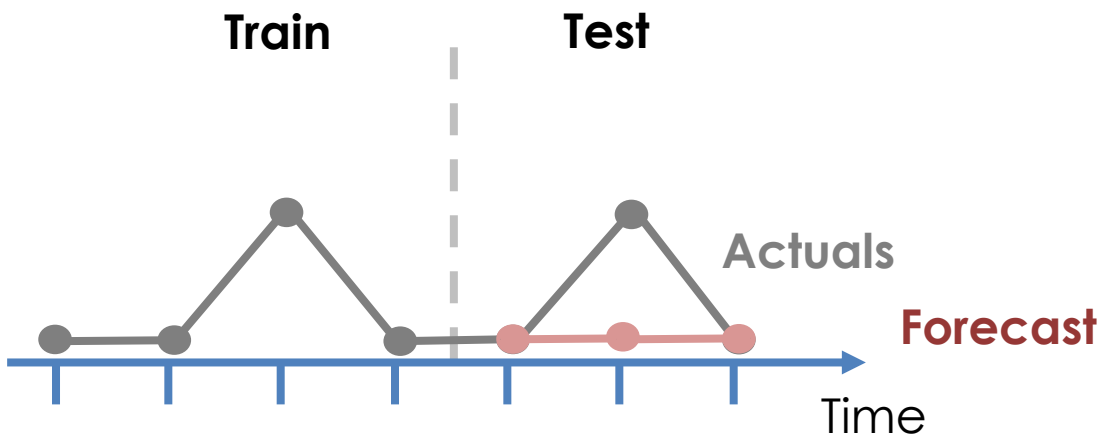
Does this also minimize $E(z_t, \hat{z}_t)$?

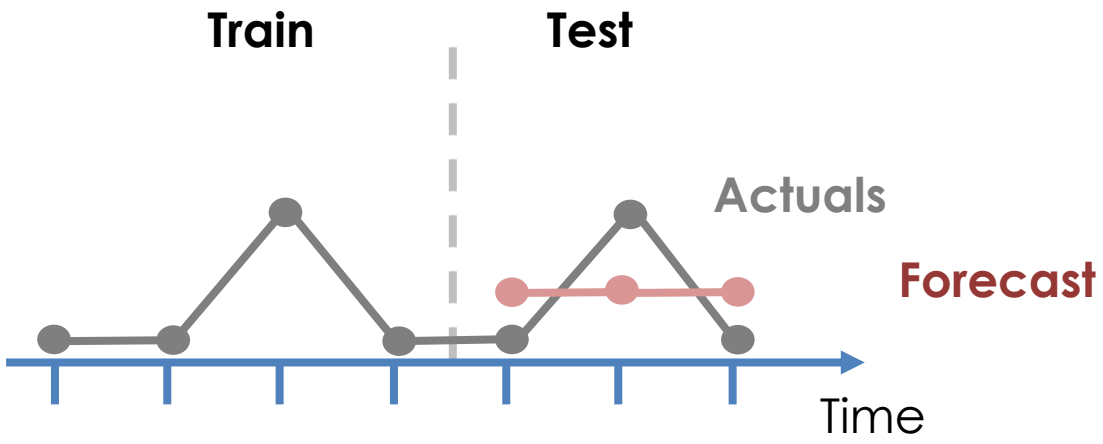
No! Only if the error metric is a **function of e_t^2**.

Examples: RMSE, RMSSE

Kolassa, S., 2023. Do we want coherent hierarchical forecasts, or minimal MAPEs or MAEs?(We won't get both!). *International Journal of Forecasting*, 39(4), pp.1512-1517.

Intermittent time series





- MAE is minimised by the median.
- "Best" forecast is zero. MAE is bad metric for intermittent data.

- RMSE is minimised by the mean.
- "Best" forecast is a non-zero constant: "avg. units sold over time".

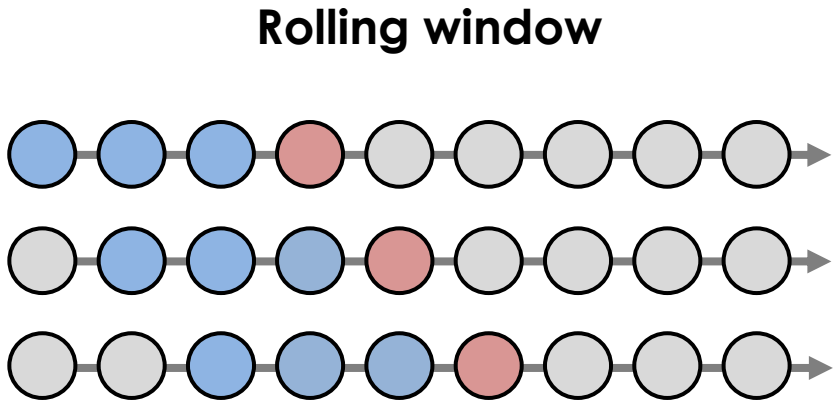
Conclusions

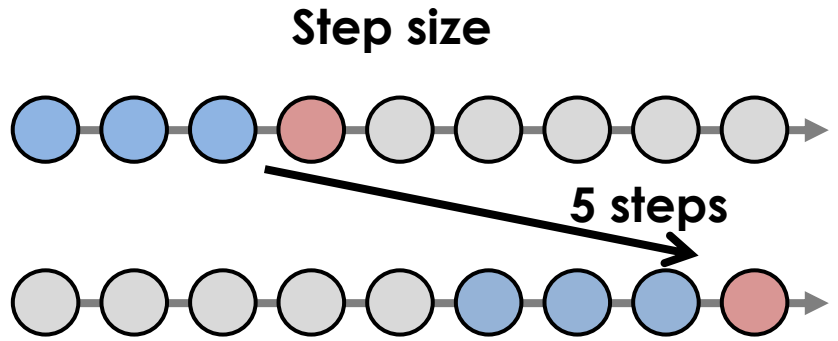
1. **Modern time series forecasting** involves many related time series.

Conclusions

1. **Modern time series forecasting** involves many related time series.

2. Selecting **backtesting** parameters is a trade-off between reliability and speed.



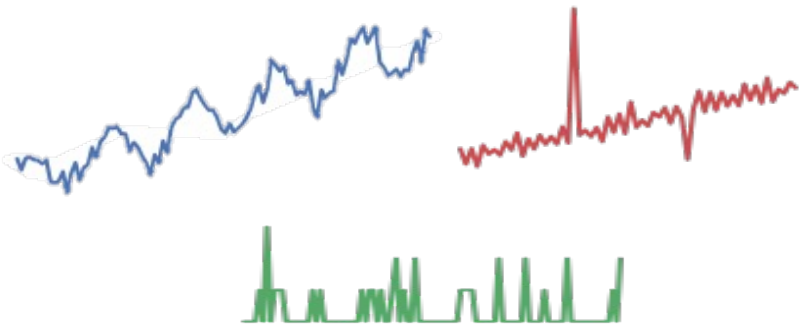


Conclusions

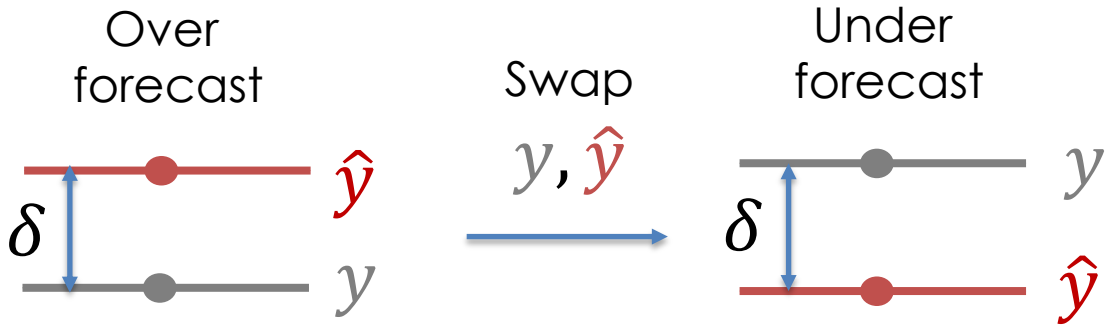
1. **Modern time series forecasting** involves many related time series.

2. Selecting **backtesting** parameters is a trade-off between reliability and speed.

3. Selecting **error metrics** requires pairing properties of our data and error metrics.



$$MAE = \frac{1}{N}\sum_{i} |e_i|$$



Conclusions

1. **Modern time series forecasting** involves many related time series.

2. Selecting **backtesting** parameters is a trade-off between reliability and speed.

3. Selecting **error metrics** requires pairing properties of our data and error metrics.

4. Many **python libraries** are available to support with forecasting.

functime

NIXTLA

Darts

SKTIME

If you'd like to learn more …

trainindata.com/p/forecasting-specialization

In/kishanmanani

@KishManani

In/soledad-galli

@Soledad_Galli

References

[1] Hewamalage, H., Ackermann, K. and Bergmeir, C., 2023. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*(2), pp.788-832.

[2] Kolassa, S., 2023. Do we want coherent hierarchical forecasts, or minimal MAPEs or MAEs?(We won't get both!). *International Journal of Forecasting*, *39*(4), pp.1512-1517.

[3] Hewamalage, Hansika, Christoph Bergmeir, and Kasun Bandara. "Global models for time series forecasting: A simulation study." *Pattern Recognition* 124 (2022): 108441.

[4] Salinas, David, et al. "DeepAR: Probabilistic forecasting with autoregressive recurrent networks." *International Journal of Forecasting* 36.3 (2020): 1181-1191.

[5] Hyndman, Rob J., and Anne B. Koehler. "Another look at measures of forecast accuracy." *International journal of forecasting* 22.4 (2006): 679-688.

Any questions?

Slides

https://github.com/KishManani/PyDataLondon2024

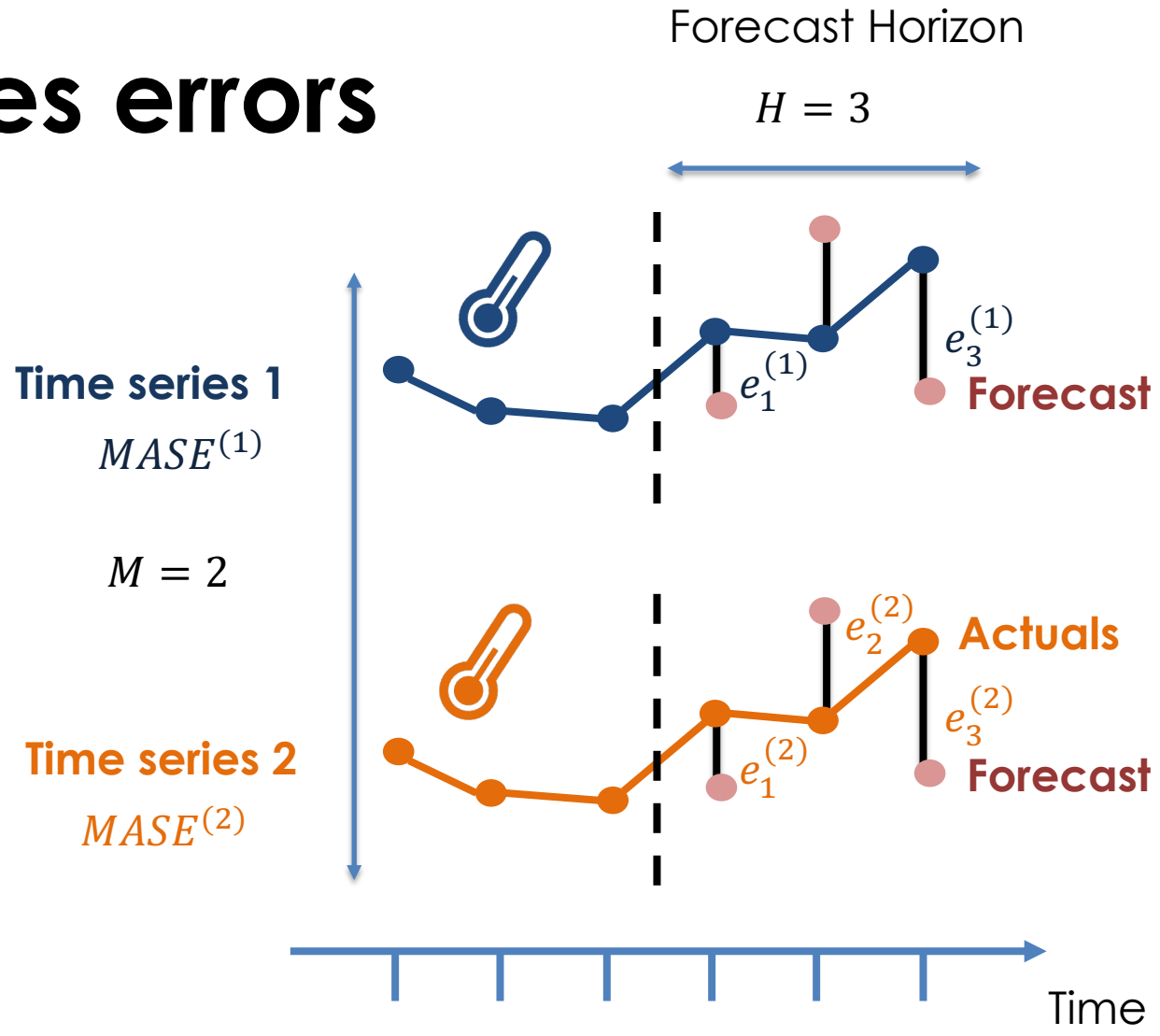
Appendix

Multiple time series errors

Average the scale-independent error metrics

$$mean\ MASE = \frac{\sum_{i=1}^{M} MASE^{(i)}}{M}$$

$$= \frac{1}{2}(MASE^{(1)} + MASE^{(2)})$$



Multiple time series error metrics

Take the **average** of a **scale independent** error metric.

$$\text{weighted mean } RMSSE = \frac{\sum_{i=1}^{M} w_i RMSSE^{(i)}}{\sum_{i=1}^{M} w_i}$$

Average over each time series

$$\text{weighted mean } WAPE = \frac{\sum_{i=1}^{M} w_i WAPE^{(i)}}{\sum_{i=1}^{M} w_i}$$

Average over each time series

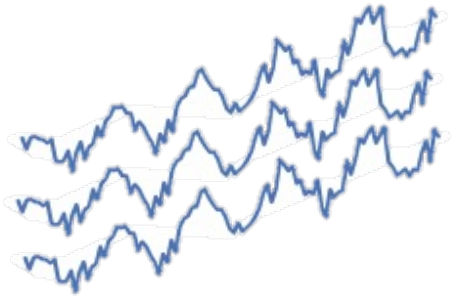
A scale independent metric is computed for each time series.

Then a weight is used to give more importance to larger scale time series.

We can give equal weight if we care about each series equally.

Guidelines: Multiple time series & scale independence

Well-behaved

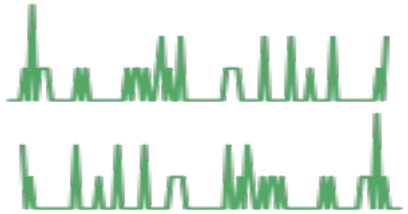


✅

Average of a scale-independent error

RMSSE, MASE, WAPE

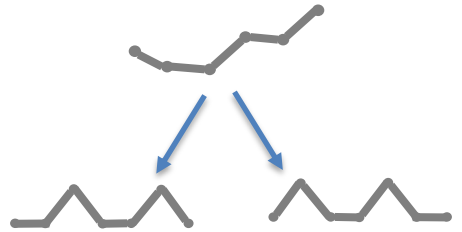
Intermittent and/or hierarchical



✅

Squared errors

- RMSSE

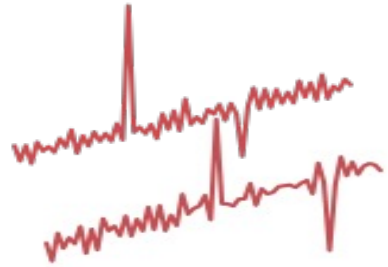


❌

Absolute errors & percentage errors

- WAPE
- MASE

Outliers



✅

Absolute errors

- WAPE
- MASE

❌

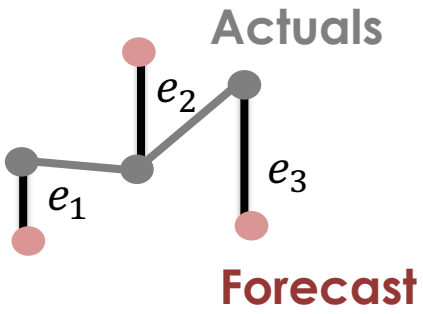
Squared errors

- RMSSE

Base errors

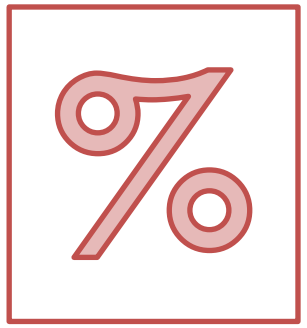
Scale-independent base errors

Scale-dependent base error



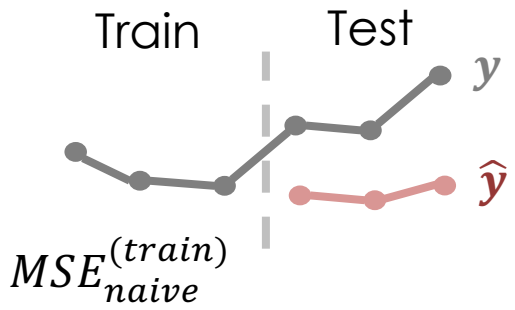
$$e_t = y_t - \hat{y}_t$$

Percentage error



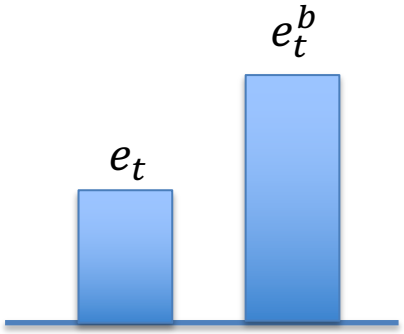
$$p_t$$

Scaled error



$$q_t$$

Relative error

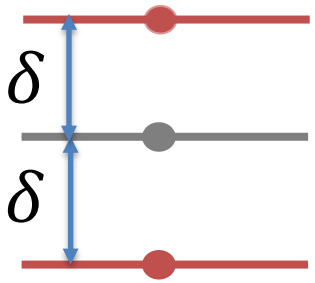


$$r_t$$

Base errors: scale-dependent base error

- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅

$$e_t = y_t - \hat{y}_t$$



$\hat{y} = y + \delta$ Over forecast

y Actual

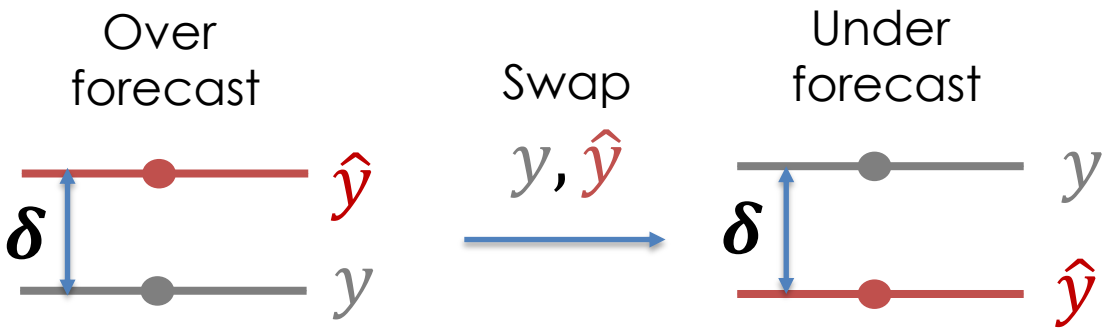
$\hat{y} = y - \delta$ Under forecast

$$|e(y, \hat{y} + \delta)| = |e(y, \hat{y} - \delta)|$$

Base errors: scale-dependent base error

- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅

$$e_t = y_t - \hat{y}_t$$

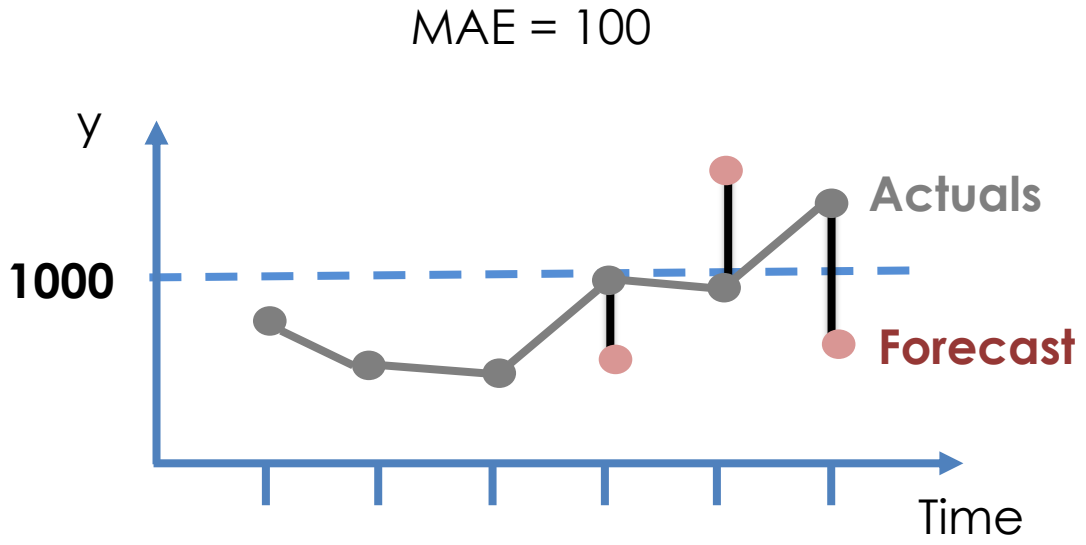


$$|e(y, \hat{y})| = |e(\hat{y}, y)|$$

Base errors: scale-dependent base error

- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅
- No issues when $y_t = 0$ or $\hat{y}_t = 0$. ✅
- Scale-dependent. ❗

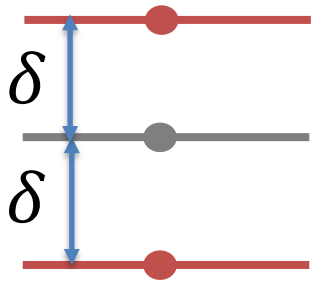
$$e_t = y_t - \hat{y}_t$$



Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅

$$p_t = \frac{100 e_t}{y_t} = 100 \frac{y_t - \hat{y}_t}{y_t}$$



$\hat{y} = y + \delta$ Over forecast

y Actual

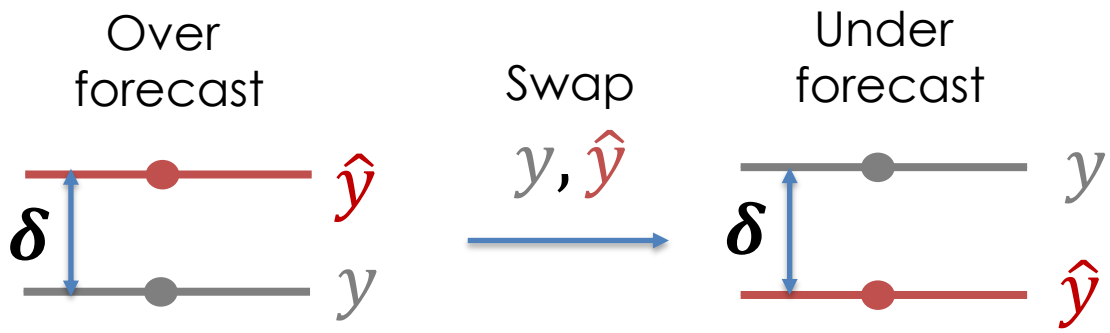
$\hat{y} = y - \delta$ Under forecast

$$e(y, \hat{y} + \delta) = e(y, \hat{y} - \delta)$$

Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Not symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ❌

$$p_t = \frac{100e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$



$$e(y, \hat{y}) \neq e(\hat{y}, y)$$

Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Not symmetric to over/under forecasting when $\hat{y} \leftrightarrow \mathrm{y}$. ❌
- Unstable when y_t is small. ❌

$$p_t = \frac{100e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$

Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Not symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ❌
- Unstable when y_t is small. ❌
- When $y_t = 0$ p_t is undefined. ❌

$$p_t = \frac{100e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$

$$p_t = 100\frac{0 - \hat{y}_t}{0} = \infty; \quad y_t = 0$$

Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Not symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ❌
- Unstable when y_t is small. ❌
- When $y_t = 0$ p_t is undefined. ❌
- When $\hat{y}_t = 0$ p_t is always 100%, regardless of e_t. ❌

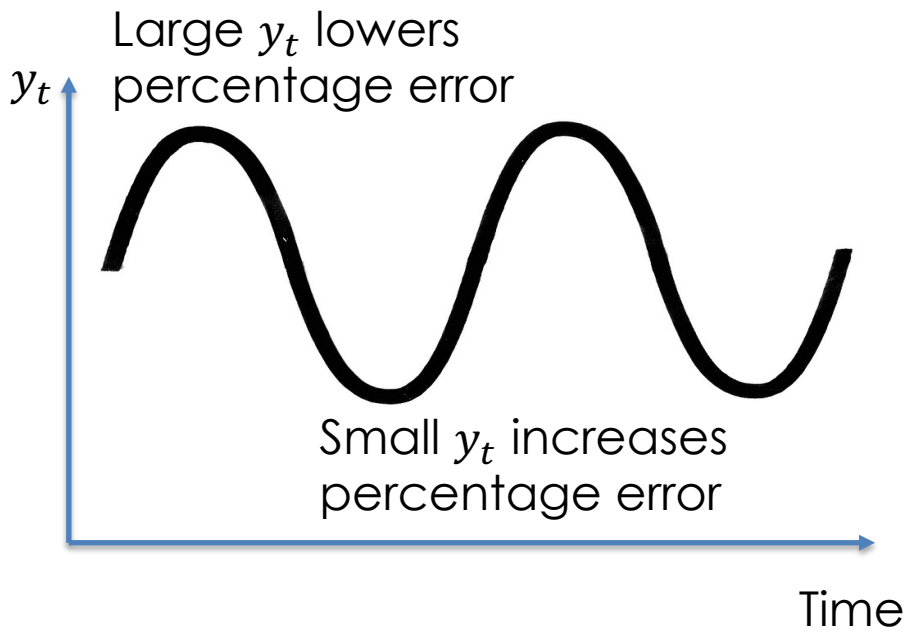
$$p_t = \frac{100e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$

$$p_t = 100\frac{y_t - 0}{y_t} = 100;\ \hat{y}_t = 0$$

Base errors: percentage error

- Scale independent. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Not symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ❌
- Unstable when y_t is small. ❌
- When $y_t = 0$ p_t is undefined. ❌
- When $\hat{y}_t = 0$ p_t is always 100%, regardless of e_t. ❌
- Seasonality and outliers can inflate and deflate p_t. ❌

$$p_t = \frac{100e_t}{y_t} = 100\frac{y_t - \hat{y}_t}{y_t}$$



Base errors: symmetric percentage error

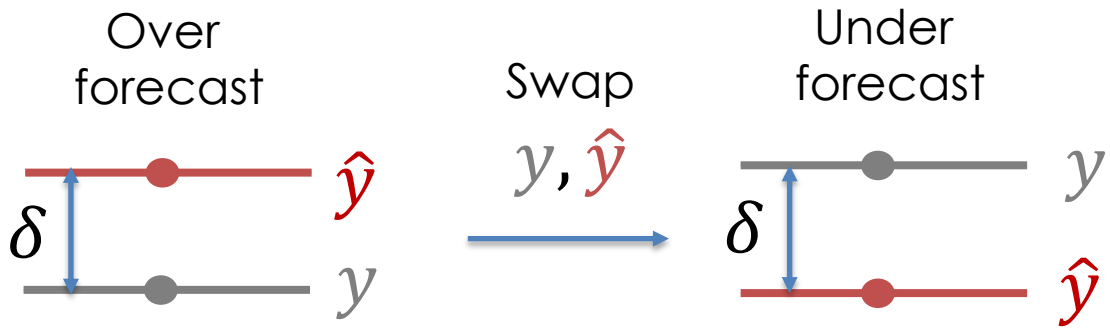
- Scale independent. ✅
- Less interpretable. ❌

$$p_t^* = \frac{100e_t}{\frac{1}{2}(|y_t| + |\hat{y}_t|)}$$

Base errors: symmetric percentage error

- Scale independent. ✅
- Less interpretable. ❌
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅

$$p_t^* = \frac{100e_t}{\frac{1}{2}(|y_t| + |\hat{y}_t|)}$$

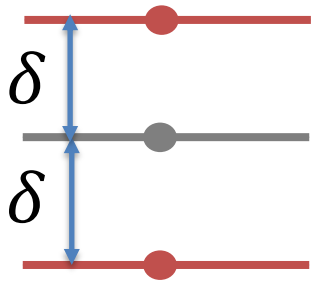


$$e(y, \hat{y}) = e(\hat{y}, y)$$

Base errors: symmetric percentage error

- Scale independent. ✅
- Less interpretable. ❌
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅
- Not symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ❌

$$p_t^* = \frac{100e_t}{\frac{1}{2}(|y_t| + |\hat{y}_t|)}$$



$\hat{y} = y + \delta$ Over forecast

y Actual

$\hat{y} = y - \delta$ Under forecast

$$e(y, \hat{y} + \delta) \neq e(y, \hat{y} - \delta)$$

Base errors: symmetric percentage error

- Scale independent. ✅
- Less interpretable. ❌
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅
- Not symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ❌
- If either $y_t = 0$ or $\hat{y}_t = 0$ then p_t^* is 200%, regardless of e_t. ❌

$$p_t^* = \frac{100 e_t}{\frac{1}{2}(|y_t| + |\hat{y}_t|)}$$

$$p_t^* = 200;\ \hat{y}_t = 0\ or\ y_t = 0$$

Exercise caution or avoid using MAPE and sMAPE

Mean absolute percentage error

$$MAPE = mean(|p_t|)$$

Symmetric mean absolute percentage error

$$sMAPE = mean(|p_t^*|)$$

WAPE as an alternative

- Weighted mean absolute percentage error (WAPE) can be helpful overcoming weaknesses.
- Scale independent when comparing series. ✅
- Interpretable. ✅
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Symmetric to over/under forecasting when $\hat{y} \leftrightarrow y$. ✅
- Well defined when $\hat{y}_t = 0$. ✅
- Well defined when some $y_t = 0$ in horizon. ✅
- Not well defined when all $y_t = 0$ in horizon. ❌
- Outliers can inflate/deflate the error. ❌

Weighted mean absolute percentage error

$$WAPE = \frac{1}{\sum_t w_t} \sum_{t=1}^{H} w_t |p_t|$$

$$\xrightarrow{w_t = |y_t|} = \frac{1}{\sum_t |y_t|} \sum_{t=1}^{H} |e_t|$$

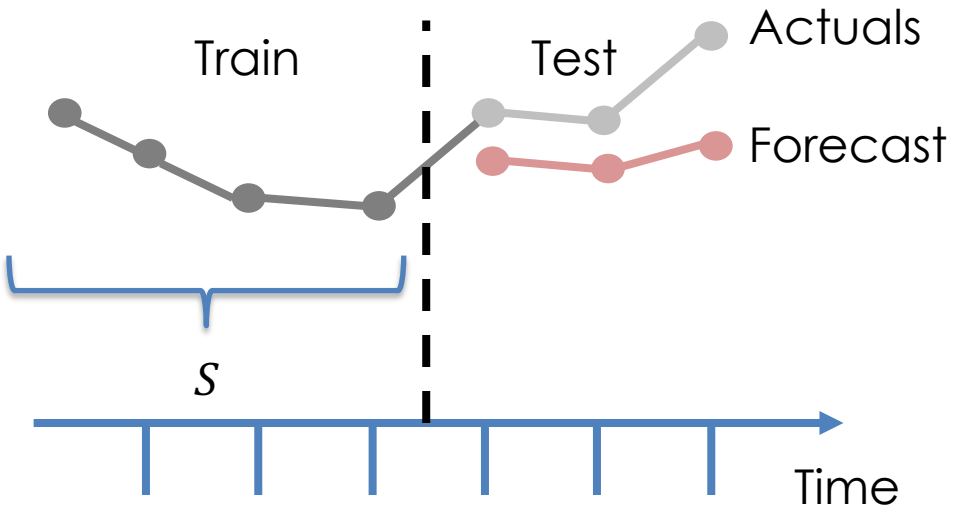
$$= \frac{MAE}{mean(|y_t|)}$$

Base errors: scaled error

- Introduced in Hyndman & Koehler 2005

$$q_t = \frac{e_t}{S}$$

MAE or MSE of **1-step or seasonal naïve forecast on the training set**



$$S = MAE_{naive} = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|$$

$$S = MSE_{naive} = \frac{1}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})^2$$

Base errors: scaled error

$$q_t = \frac{e_t}{S}$$

- Scale independent. ✅
- Interpretable. ❓
- Symmetric to over/under forecasting when $\hat{y} = y \pm \delta$. ✅
- Symmetric to under/over forecasting when $\hat{y} \leftrightarrow y$. ✅.
- Well defined when $\hat{y}_t = 0$ or $y_t = 0$. ✅
- Not well defined when y_t is constant in train set. ❌
- Scale of the errors in the training set may not be the same in the future. ❌

Mean absolute scaled error

$$MASE = mean(|q_t|)$$

$$S = MAE_{naive}^{(train)}$$

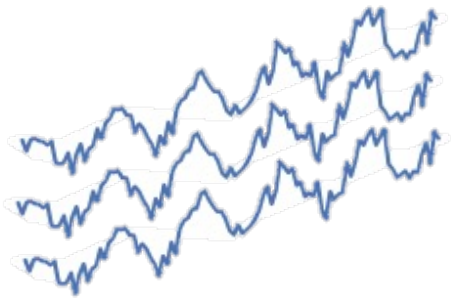
Root mean squared scaled error

$$RMSSE = \sqrt{mean(q_t^2)}$$

$$S = MSE_{naive}^{(train)}$$

Guidelines: Multiple time series & scale dependence

Well-behaved



✅

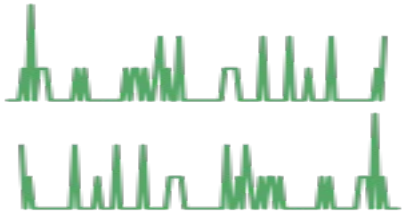
Pooled scale-dependent error

NRMSE, ND

Weighted average of a scale-independent error

W-RMSSE, W-MASE, W-WAPE

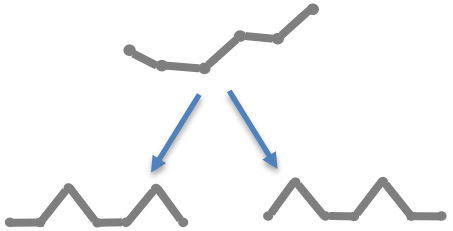
Intermittent and/or hierarchical



✅

Squared errors

- NRMSE
- W-RMSSE

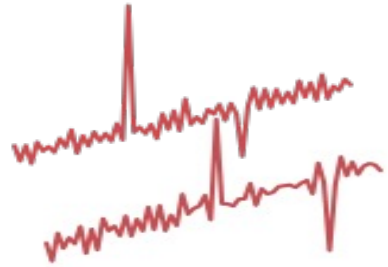


❌

Absolute errors & percentage errors

- ND
- W-WAPE
- W-MASE

Outliers



✅

Absolute errors

- ND
- W-MASE

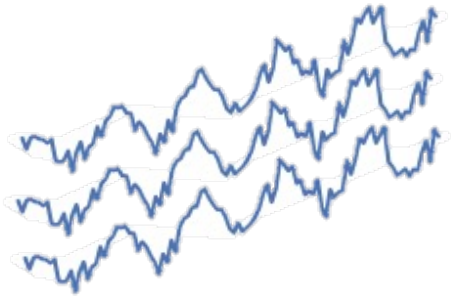
❌

Squared errors

- NRMSE
- W-RMSSE

Guidelines: Multiple time series & scale independence

Well-behaved

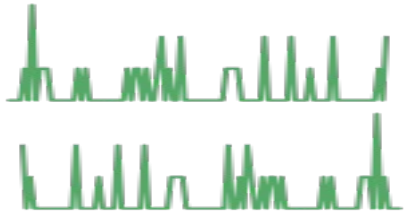


✅

Average of a scale-independent error

RMSSE, MASE, WAPE

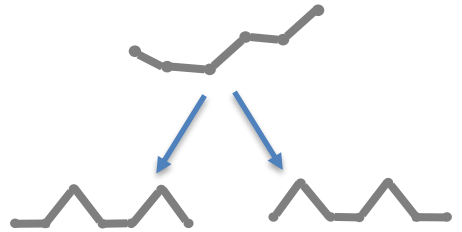
Intermittent and/or hierarchical



✅

Squared errors

- RMSSE

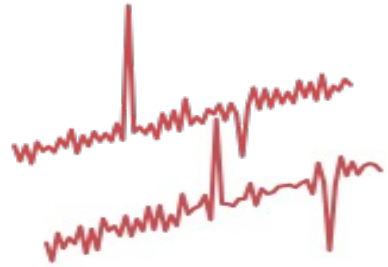


❌

Absolute errors & percentage errors

- WAPE
- MASE

Outliers



✅

Absolute errors

- WAPE
- MASE

❌

Squared errors

- RMSSE

Summary operators

- **Mean**
  - Sensitive to outliers.
- **Weighted mean**
  - Sensitive to outliers.
  - Allows us to give more importance to specific time points or values.
- **Median**
  - Outlier robust.

Examples

$$MSE = \frac{1}{N}\sum_t e_t^2$$

$$WAPE = \frac{1}{\sum_t w_t}\sum_t w_t|p_t|$$

$$RMdSE = \sqrt{median(e_t^2)}$$

Transforms

- **Absolute value:** $|x|$
  - Less sensitive to outliers.
  - Optimises for the median.

- **Square:** x^2
  - Sensitive to outliers.
  - Optimises for the mean.
  - Better for intermittent data.
  - Better for coherent hierarchical forecasts.

Examples

$$MAE = \frac{1}{N}\sum_{t}|e_t|$$

$$RMSE = \sqrt{mean(e_t^2)}$$

Modified percentage errors

- Exercise caution and/or avoid using MAPE and sMAPE.
- Consider using **modified** MAPE or sMAPE to handle instability when denominator is small.
- Still asymmetric to under and over forecasting. ❌
- $\hat{y}_t = 0$ and $y_t = 0$ still an issue. ❌

Modified MAPE

$$mMAPE = \frac{1}{H}\sum_{t=1}^{H}\frac{100|e_t|}{|y_t| + \epsilon}$$

$\epsilon = 1$ in Bandera et al. 2019

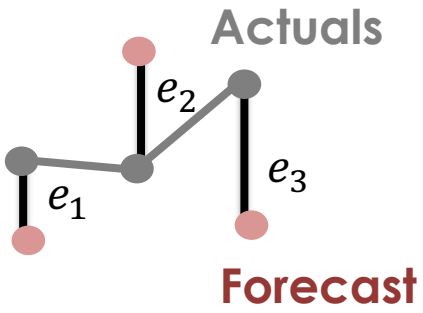
Modified sMAPE

$$msMAPE = \frac{1}{H}\sum_{t=1}^{H}\frac{100|e_t|}{\frac{1}{2}\max(|y_t| + |\hat{y}_t| + \epsilon, 0.5 + \epsilon)}$$

$\epsilon = 0.1$ in Suilin 2017

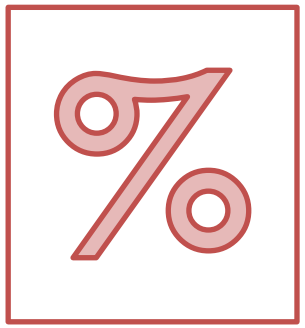
Base errors

Scale-dependent base error



$$e_t = y_t - \hat{y}_t$$

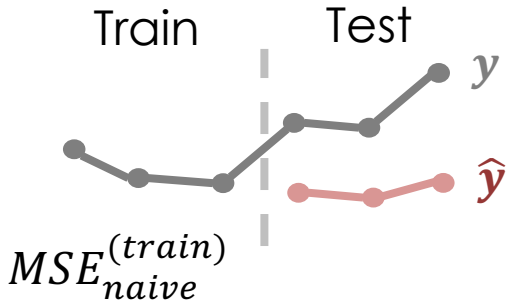
Percentage error



$$p_t = \frac{e_t}{y_t}$$

$$p_t^* = \frac{e_t}{\frac{1}{2}(|y_t| + |\hat{y}_t|)}$$

Scaled error

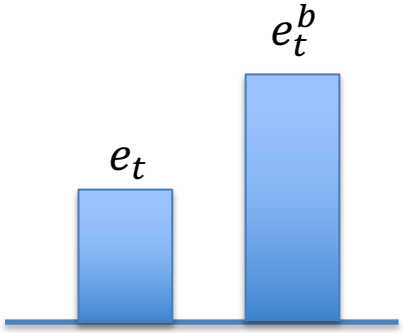


$$q_t = \frac{e_t}{S}$$

$$S = MSE_{naive}^{(train)}$$

$$S = MAE_{naive}^{(train)}$$

Relative error



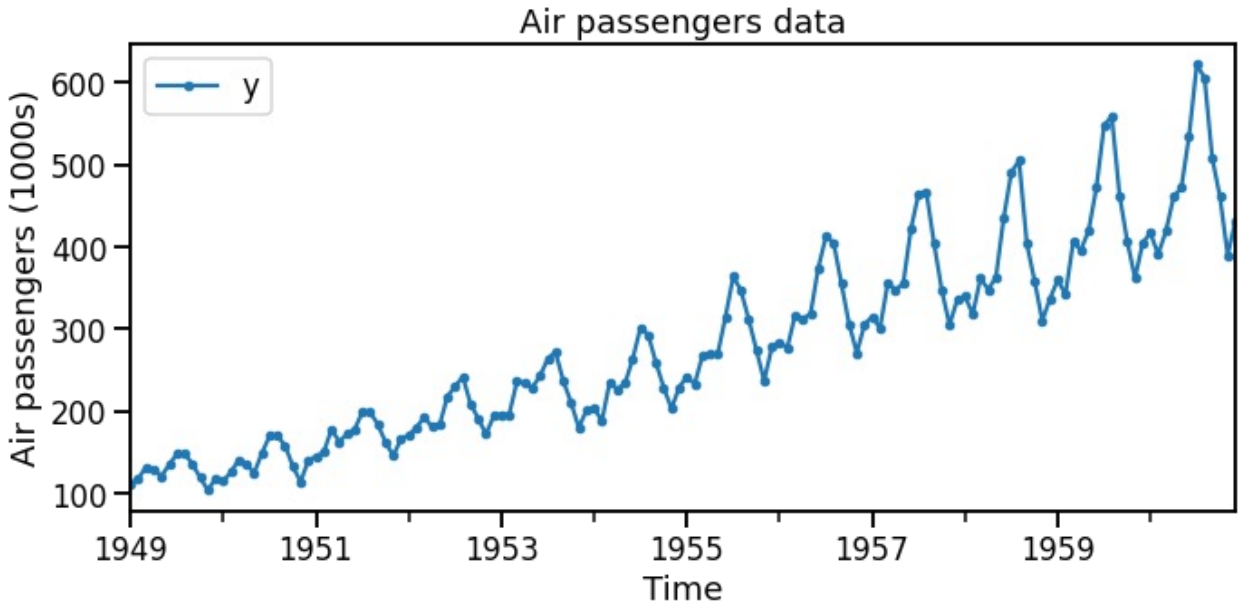
$$r_t = \frac{e_t}{e_t^b}$$

$$e_t^b = y_t - \hat{y}_t^b$$

Use simpler methods for “easy” time series

Time series characteristics

- Strong seasonality and/or trend.
- Small number of time series.
- Uncorrelated time series.
- No sparsity or intermittency.
- Few or no exogenous features.

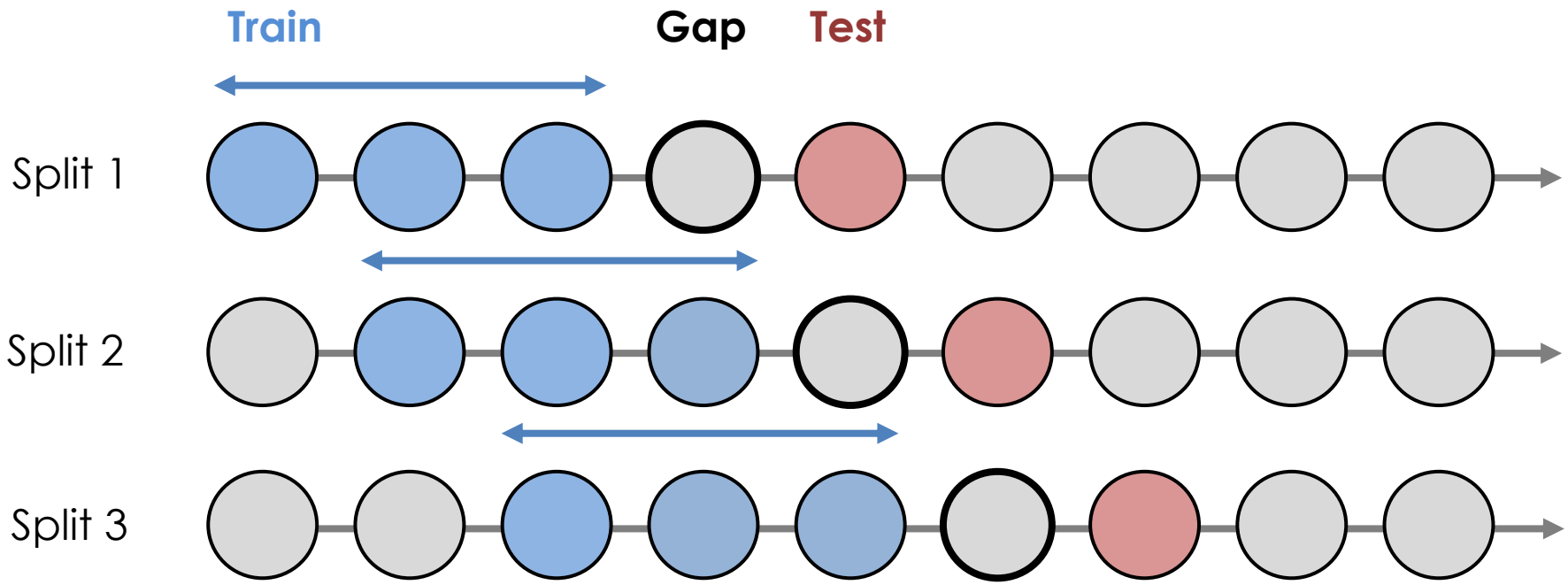


Methods

- Naïve and rules based forecasts
- ARIMA
- ETS
- Prophet

Backtesting: Try to reflect process in production

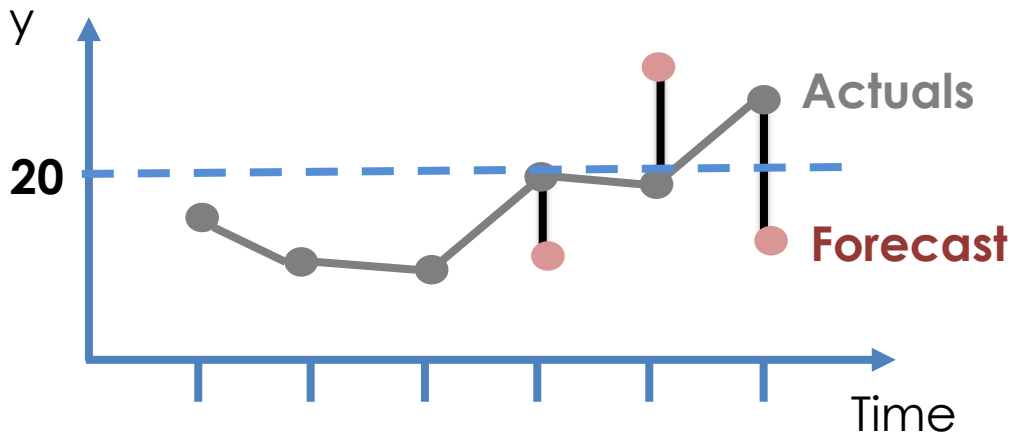
Example: There is a gap between when the forecast is created relative to when it is used. This gap should be reflected in backtesting.



Which forecast is better?

Time series 1

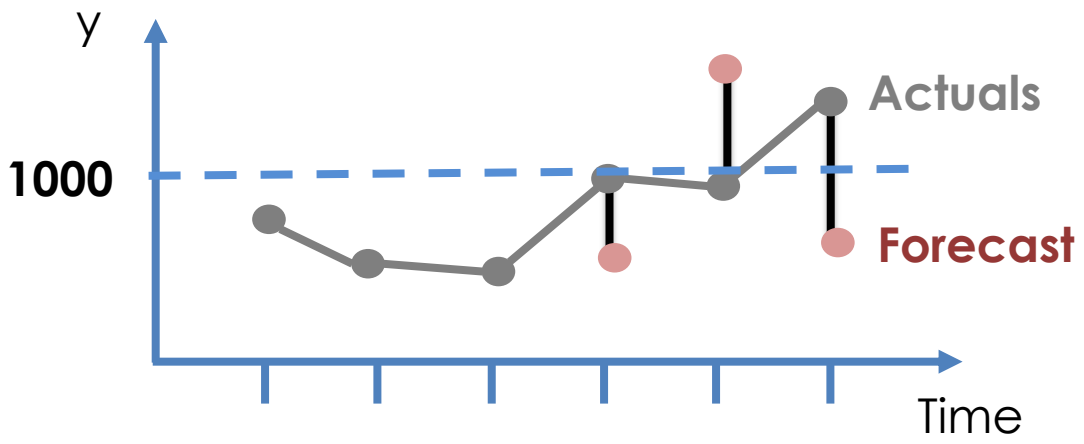
MAE = 10



Error is ~ 50%

Time series 2

MAE = 100

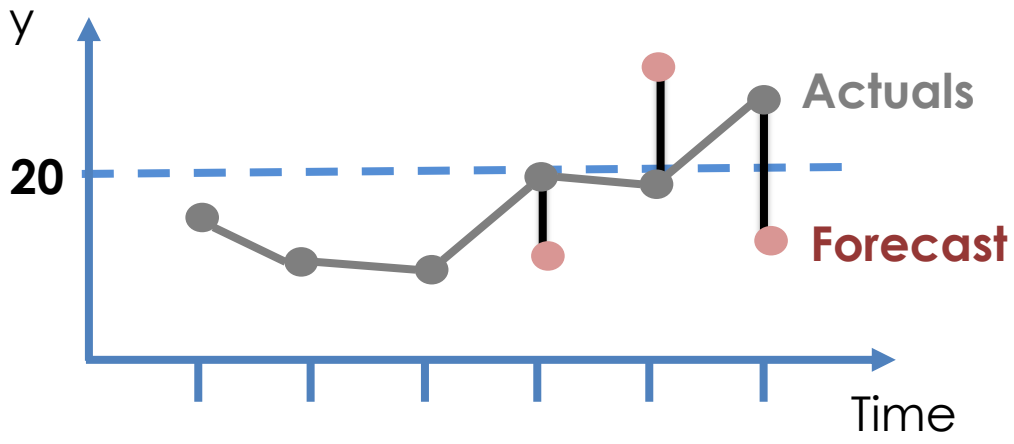


Error is ~ 10%

Which forecast is better?

Time series 1

MAE = 10



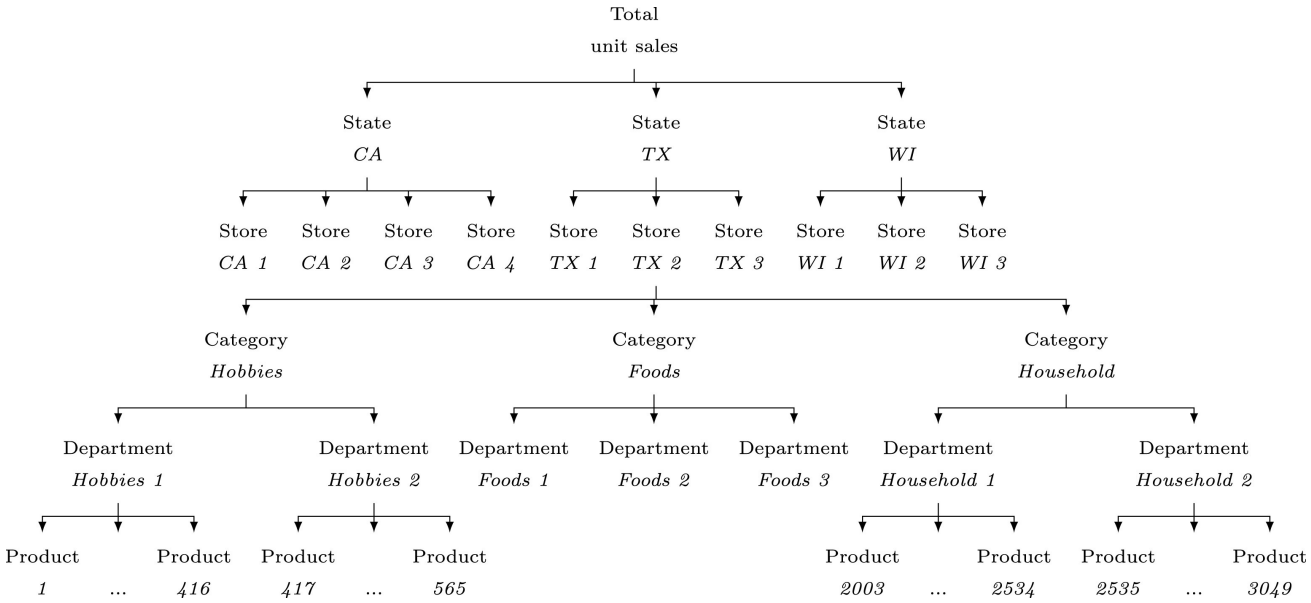
Error is ~ 50%

The **scale** of the time series matters.

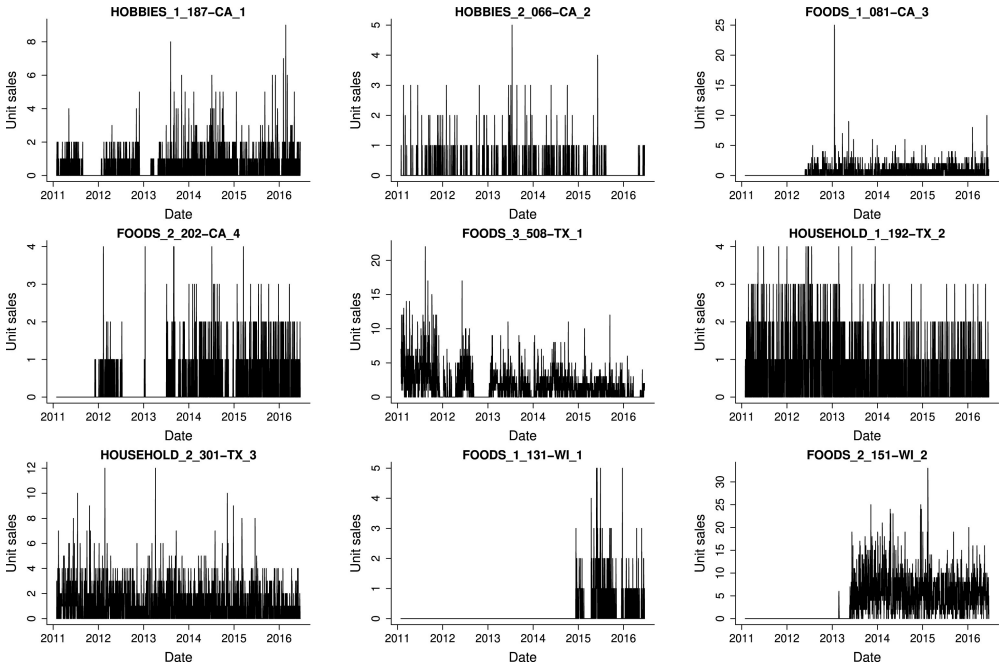
This is important when comparing errors across **multiple time series**, across **time periods** within a series, and if there are **outliers**.

Use ML methods for “hard” time series

- Large number of correlated time series (30,490)
- Hierarchical structure
- Varying length for each time series



- High sparsity & intermittency
- Exogenous variables (price, promos, etc.)
- Multiple seasonal patterns



[1] Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. "The M5 competition: Background, organization, and implementation." *International Journal of Forecasting* (2021).