kaggle days

14-15 NOVEMBER 2022
PARIS

Sales Forecasting and Fraud Detection

Jean-Francois Puget (CPMP)

Organizers

LOGIC AI kaggle LVMH

SEPHORA LOUIS VUITTON Dior Christian Dior PARFUMS

CELINE GIVENCHY BVLGARI MoëtHennessy TIFFANY & CO.

Sales Forecasting and Fraud Detection

- Sales forecasting
  - Cast as regression
  - Time based validation
  - Recency and seasonality bias
  - Sequence to sequence
  - More data
- Recommender systems
- Fraud Detection
  - Common traps

About me

CPMP

RAPIDS and deep learning at NVIDIA

France

Joined 10 years ago · last seen in the past day

https://rapids.ai/

Followers 6124

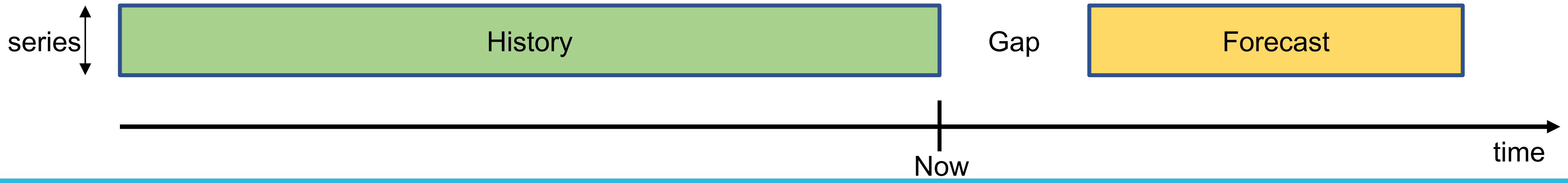
Competitions Grandmaster

Home | Competitions (84) | Datasets (15) | Code (932) | Discussion (9,790) | Followers (6,124) | Notifications | Account | Edit Public Profile

	Competitions Grandmaster	Datasets Expert	Notebooks Grandmaster	Discussion Grandmaster
Current Rank	10 of 193,900	607 of 73,487	164 of 243,462	4 of 313,583
Highest Rank	6	94	12	1
Gold	24	0	15	447
Silver	26	0	11	577
Bronze	4	6	18	4489

Sales forecasting

- Predict sales in the future
- From Historical data
  - Several (many) time series, per product, region, and aggregated
- Mind the gap!
  - There is a delay between when training data is collected and prepared and when the model can be put in production



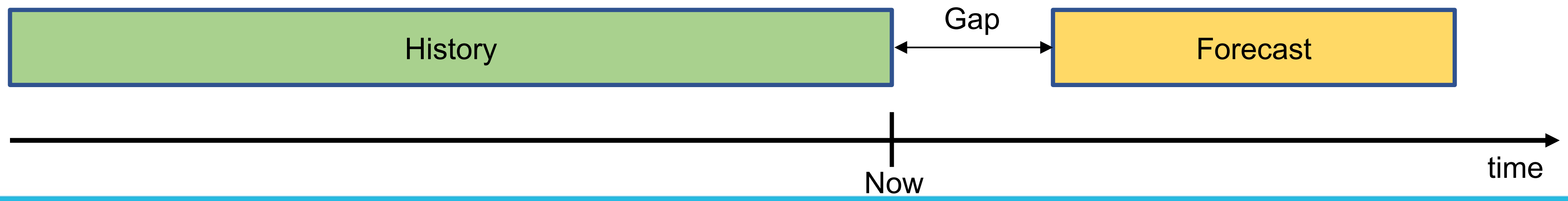
Many models

- Facebook Prophet
- VAR
- ARIMA
- TFT (best deep learning model?)
- …

Auto regressive models and sequence to sequence models, usually for single series

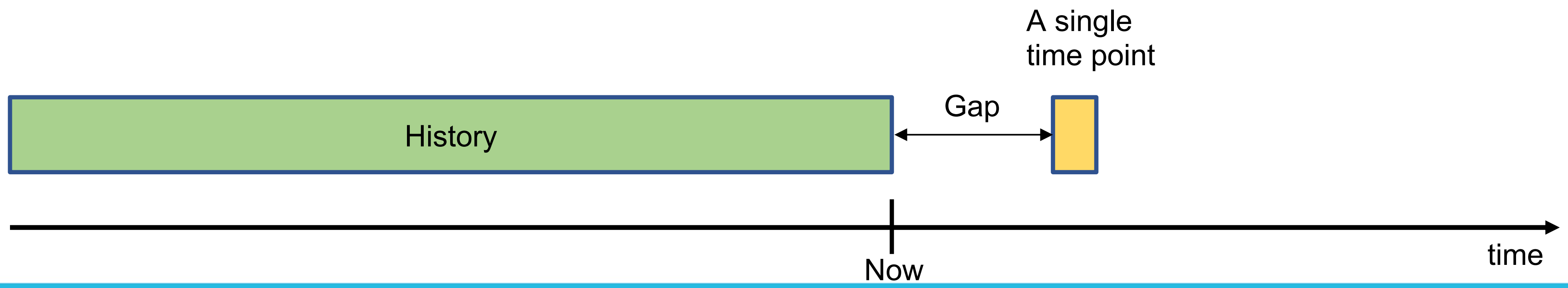
There is another way!

- Cast as regression



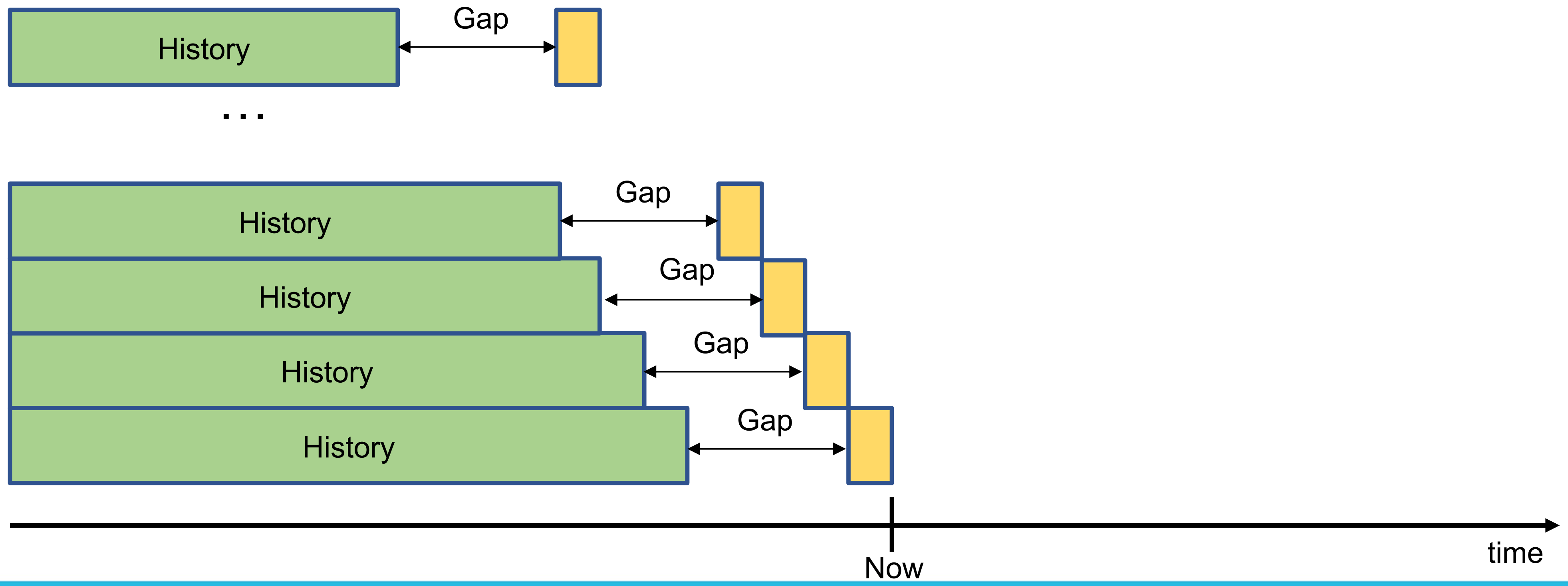
Forecasting as Regression

- Sequence to value: single prediction
  - We need to forecast one value
- Generalize to Sequence prediction
  - Iterate single prediction
  - Train several models
  - Train one model



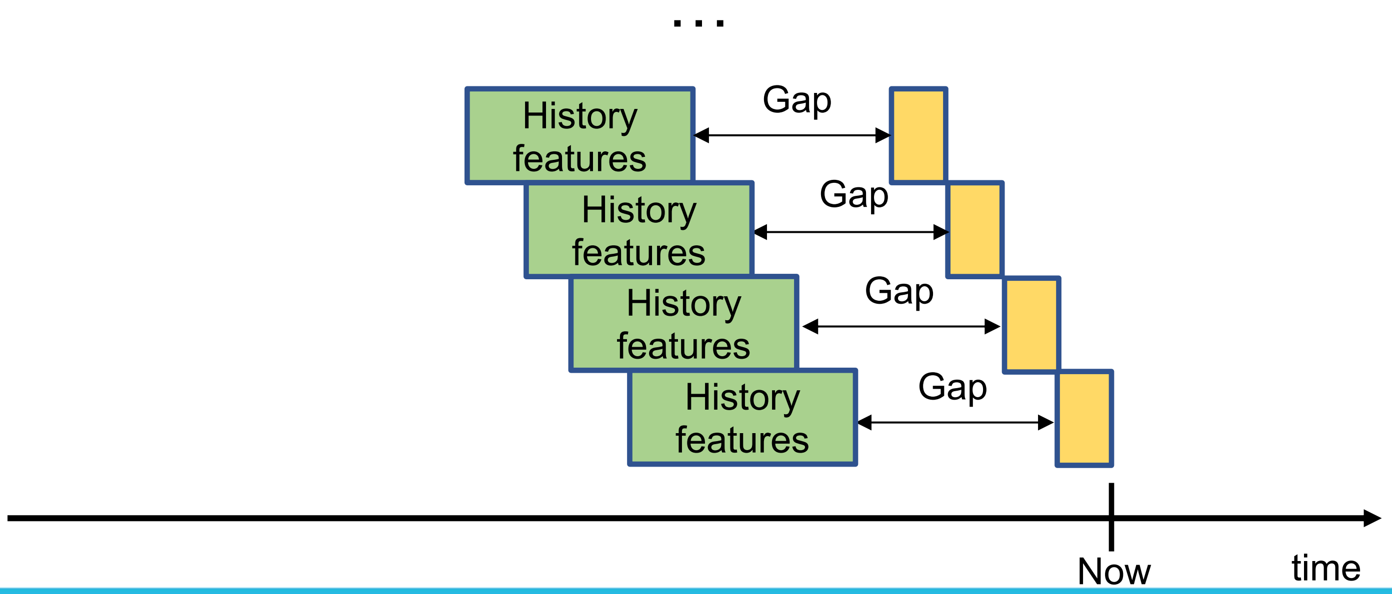
Single regression

- Create training instances by shifting train end date and target date: keep the gap constant!
- Features are aggregates of the history data



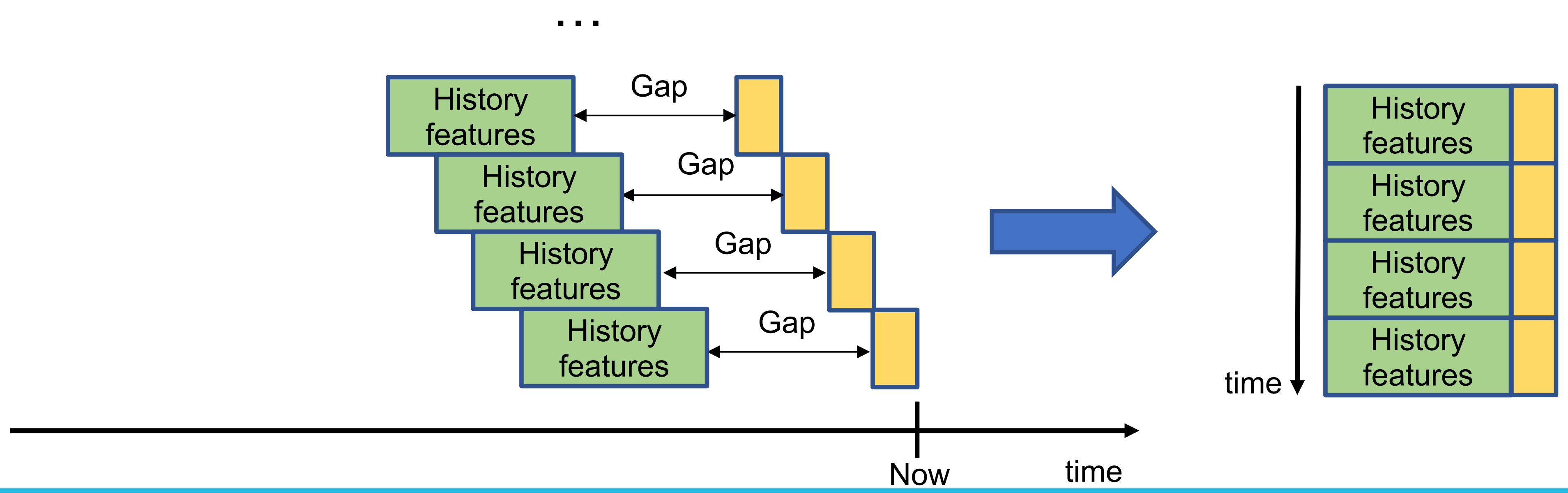
Single regression

- Create training instances by shifting train end date and target date: keep the gap constant!
- Features are aggregates of the history data
  - Last value
  - Min, Max, Median, Mean, Std of last week
  - Min, Max, Median, Mean, Std of last month
  - …



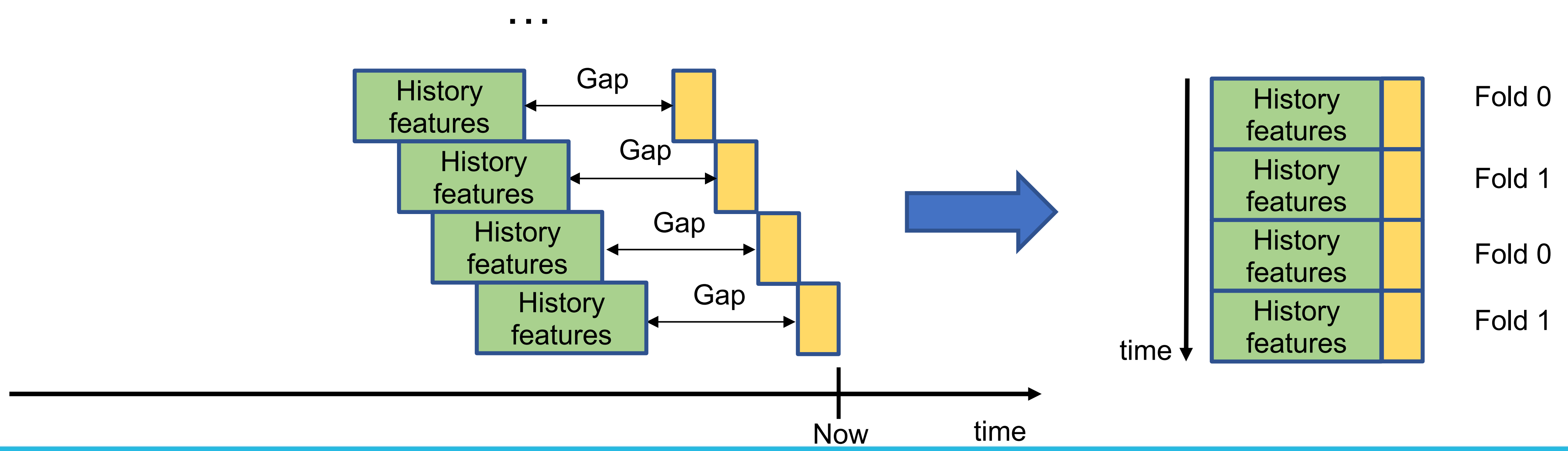
Single regression

- We then get a regression problem
- Features can come from more than one time series -> multivariate time series



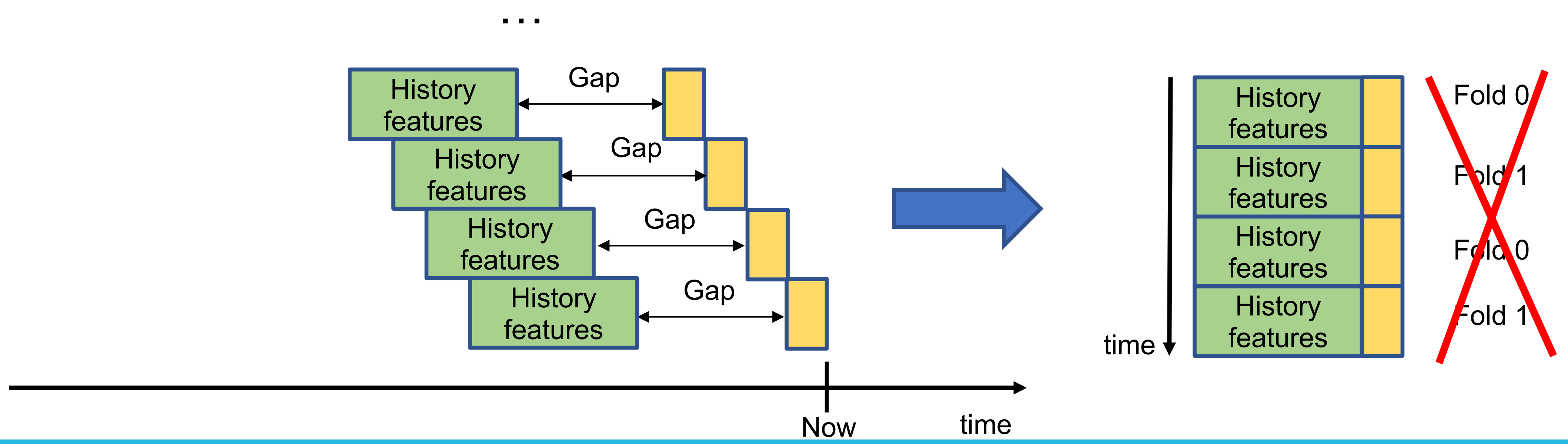
Single regression

- Cross validation
- Never ever use for training data which is in the future of validation
- Example with 2 folds



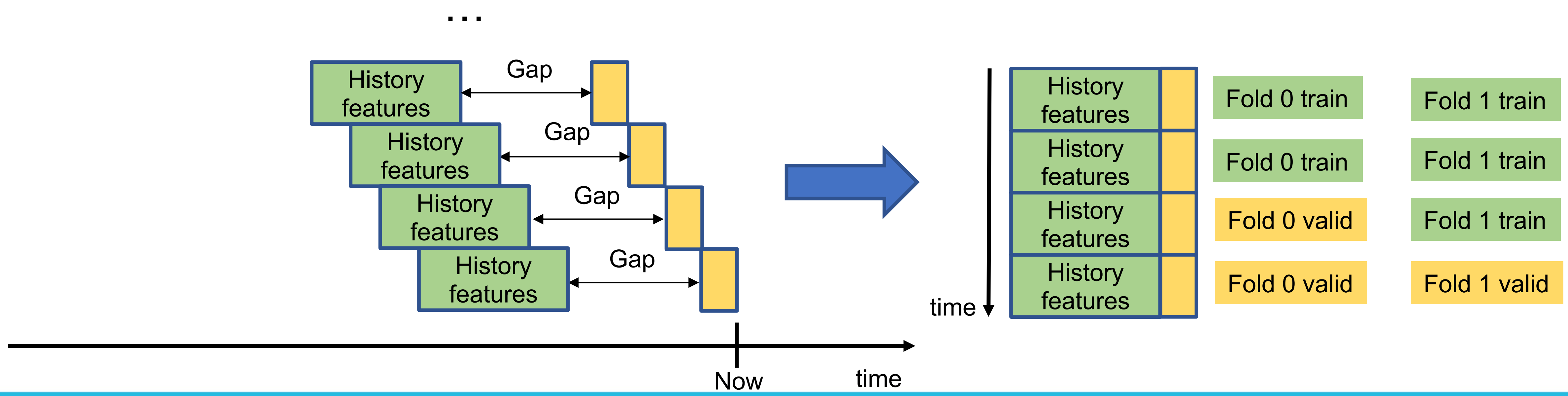
Single regression

- Cross validation
- Never ever use for training data which is in the future of validation
- Example with 2 folds



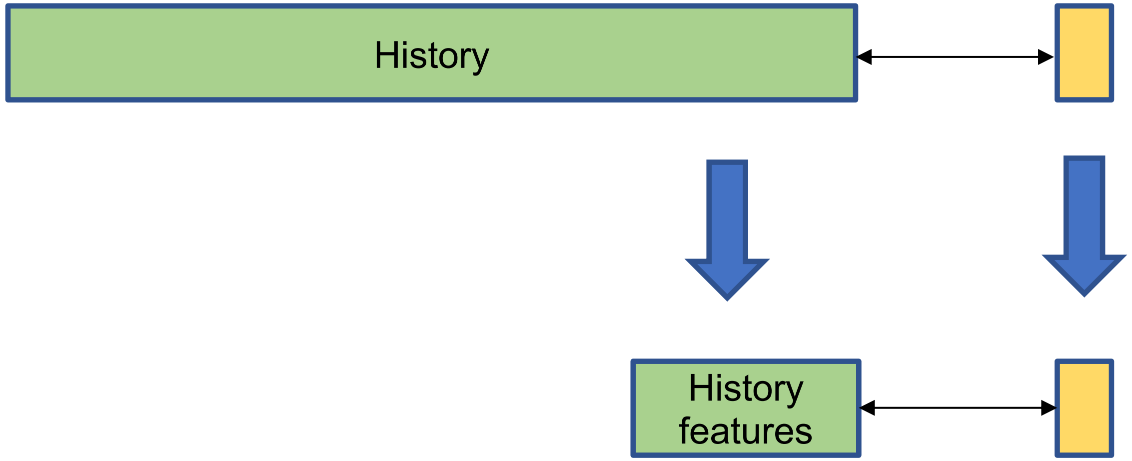
Single regression

- Cross validation
- Never ever use for training data which is in the future of validation
- Example with 2 folds



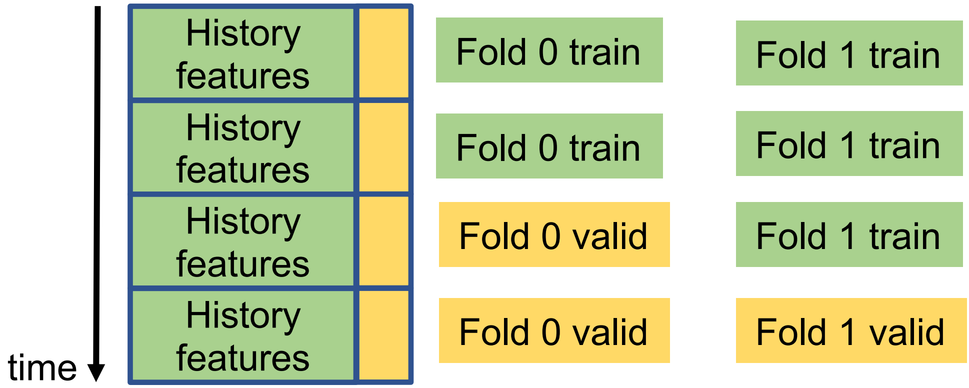
De-trending

- Gradient boosting (e.g. XGBoost) cannot extrapolate!
  - Often heard to dismiss it for forecasting
- Predict the residual of the trend
  - Modify the target!
- New target is original target minus
  - Mean value over last month
  - Same value a year ago (yearly seasonality)
  - Same value last week (weekly seasonality)
  - A linear baseline that combines the above
- Baseline can be even more complex!
  - I won a 2nd prize using a sequence to sequence NN model as baseline
  - XGBoost to predict the residuals



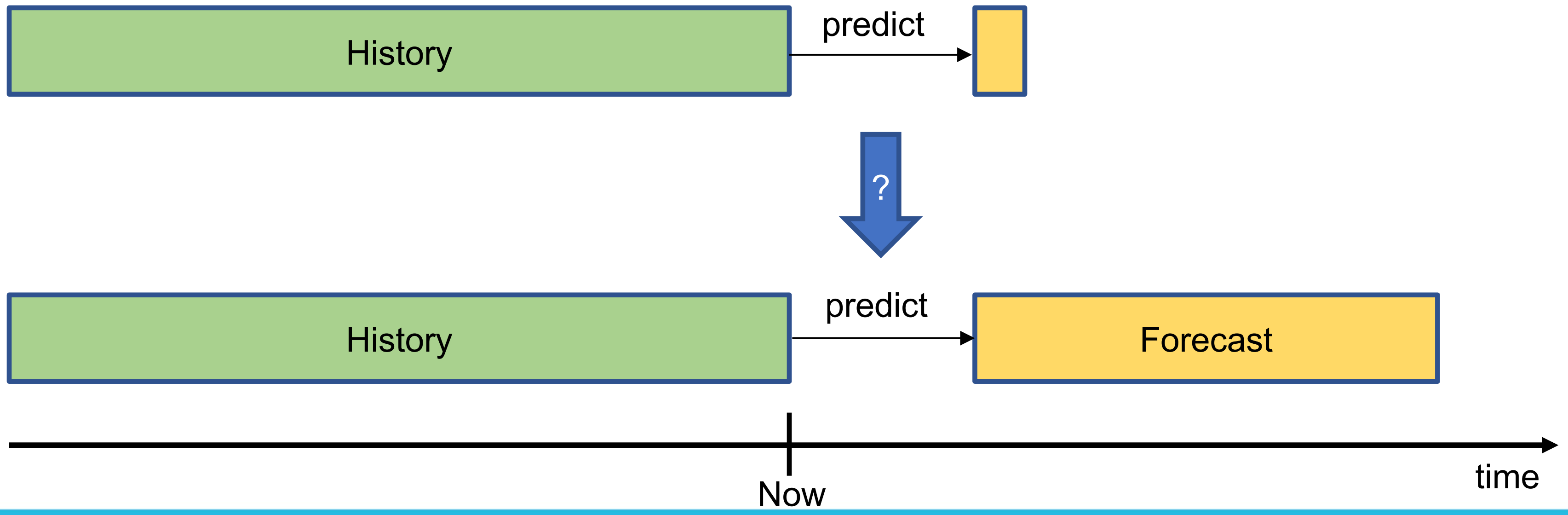
Recency bias

- Last known value is a great predictor
- More recent samples are more important
- Weight samples by recency when training
- Weight last year same period more If there is a yearly seasonality



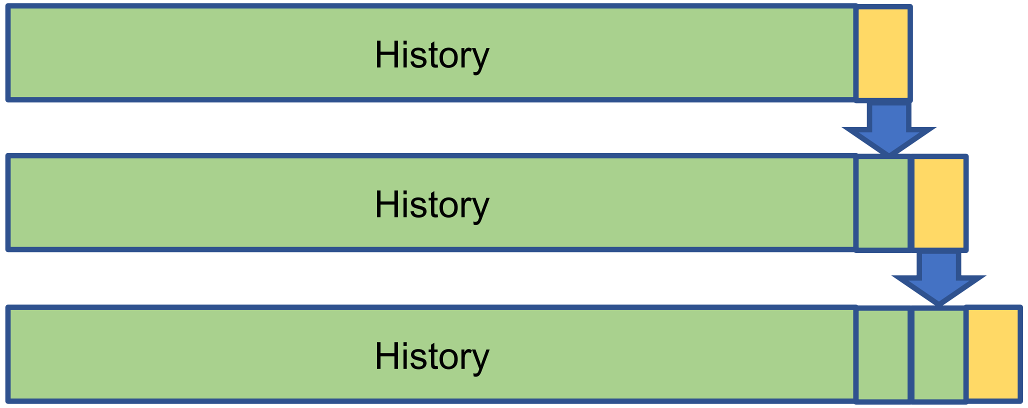
Sequence to sequence

- We have shown how to train a model to predict one time period in the future
- How do we predict a sequence in the future with it?
- XGBoost cannot predict sequences!



Sequence to sequence

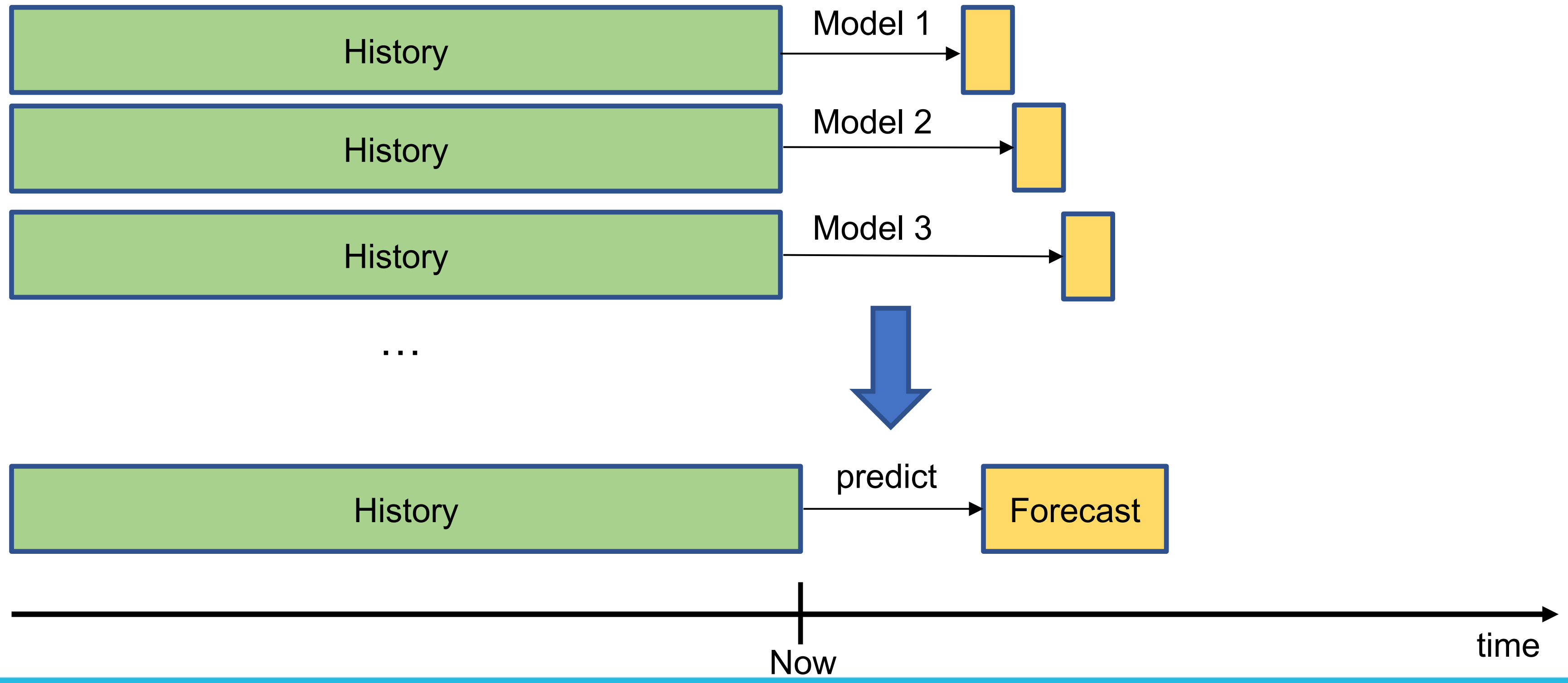
- If gap is negligible, predict next period, then use the prediction as additional history to predict one period more, etc



- Main disadvantage: prediction errors accumulate, and compound effect can be really bad

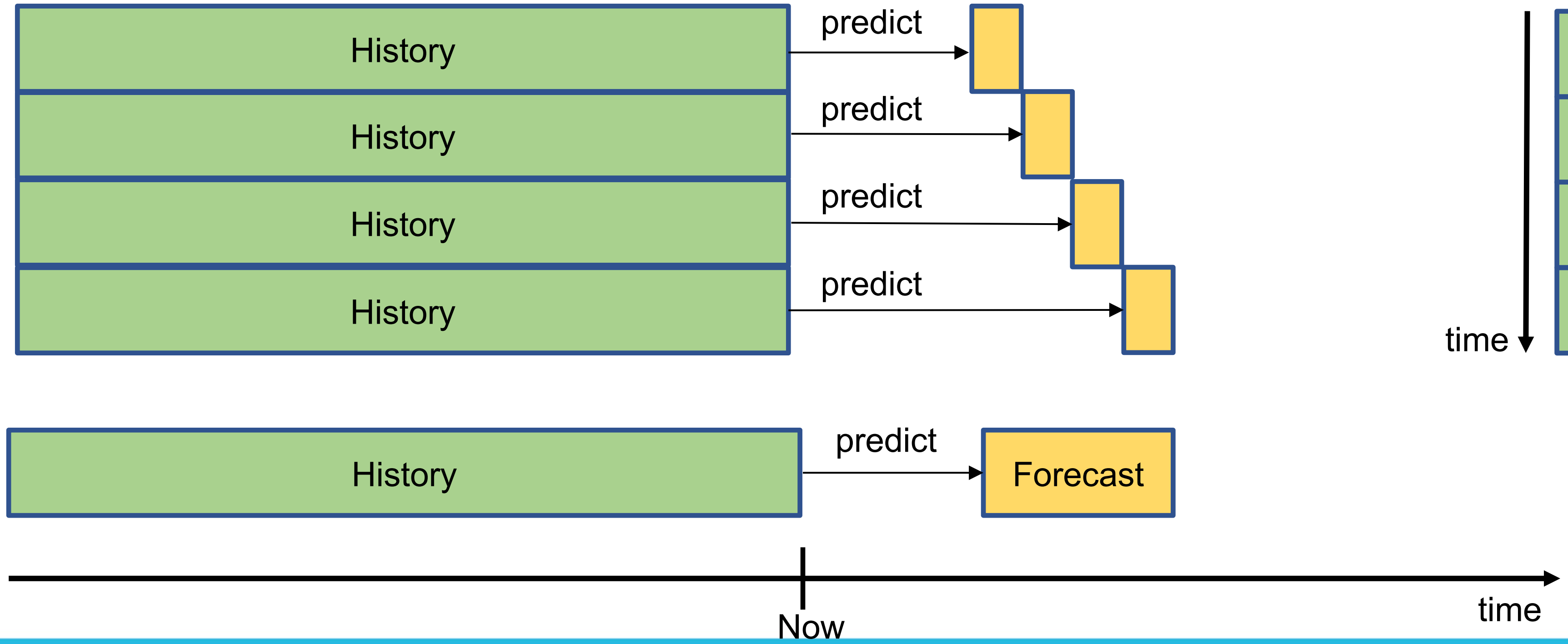
Sequence to sequence

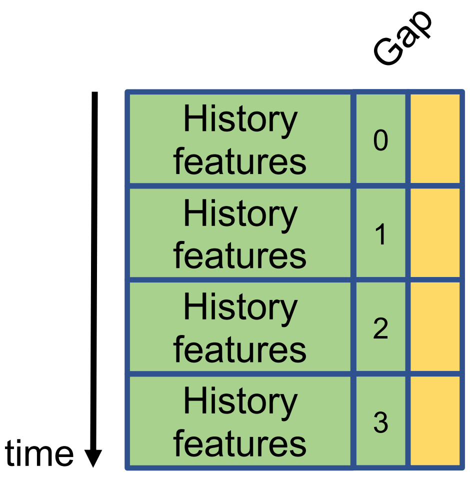
- Train N models where N is the length of the sequence to predict, with increasing gaps



Sequence to sequence

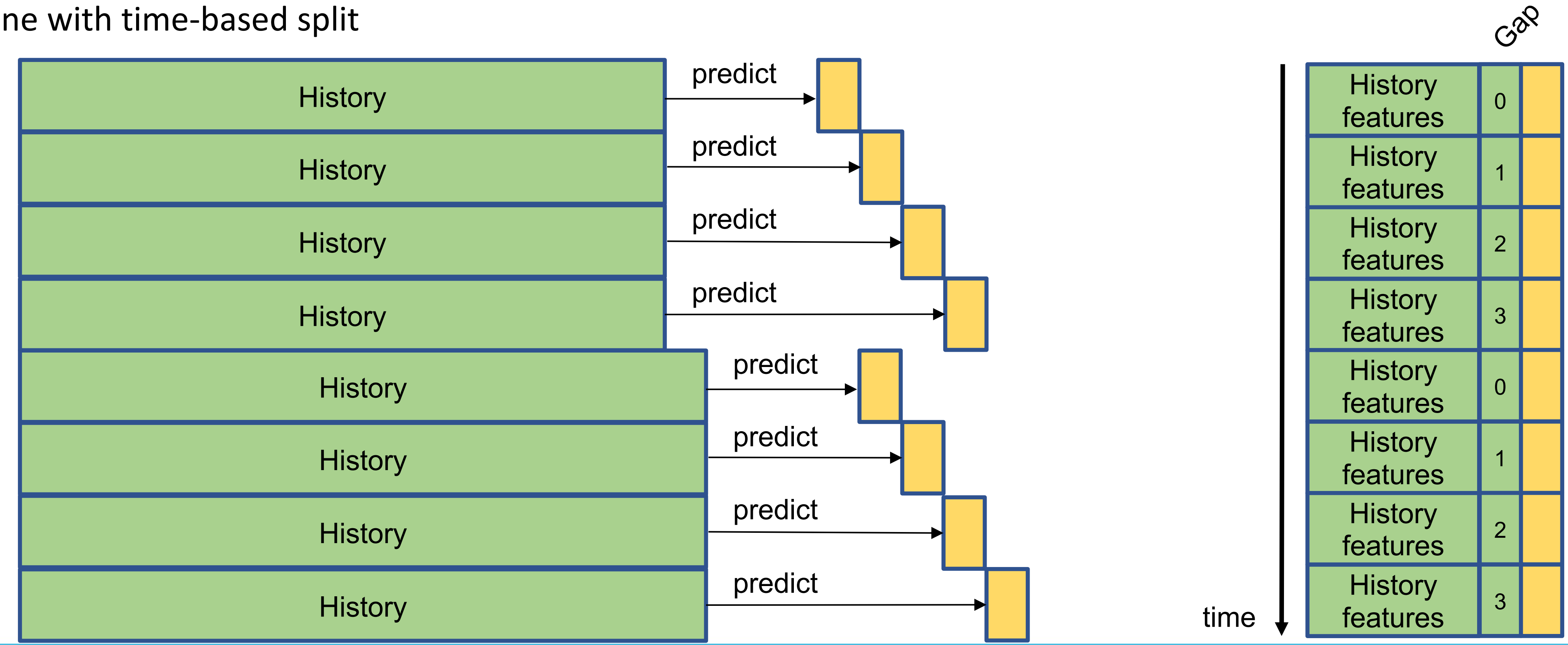
- Train one model
- Each sample becomes N samples, with increasing gap values





Sequence to sequence

- Combine with time-based split



Sales forecasting

- Target is not symmetric, over predicting maybe fine
  - Unsold products can be sold later
  - Missing products cannot be sold
- Not true for some products
  - Limited shelf life, e.g. dairy
  - Deadline, e.g. Christmas tree
- Quantile regression
  - Weight over prediction error differently than under prediction error

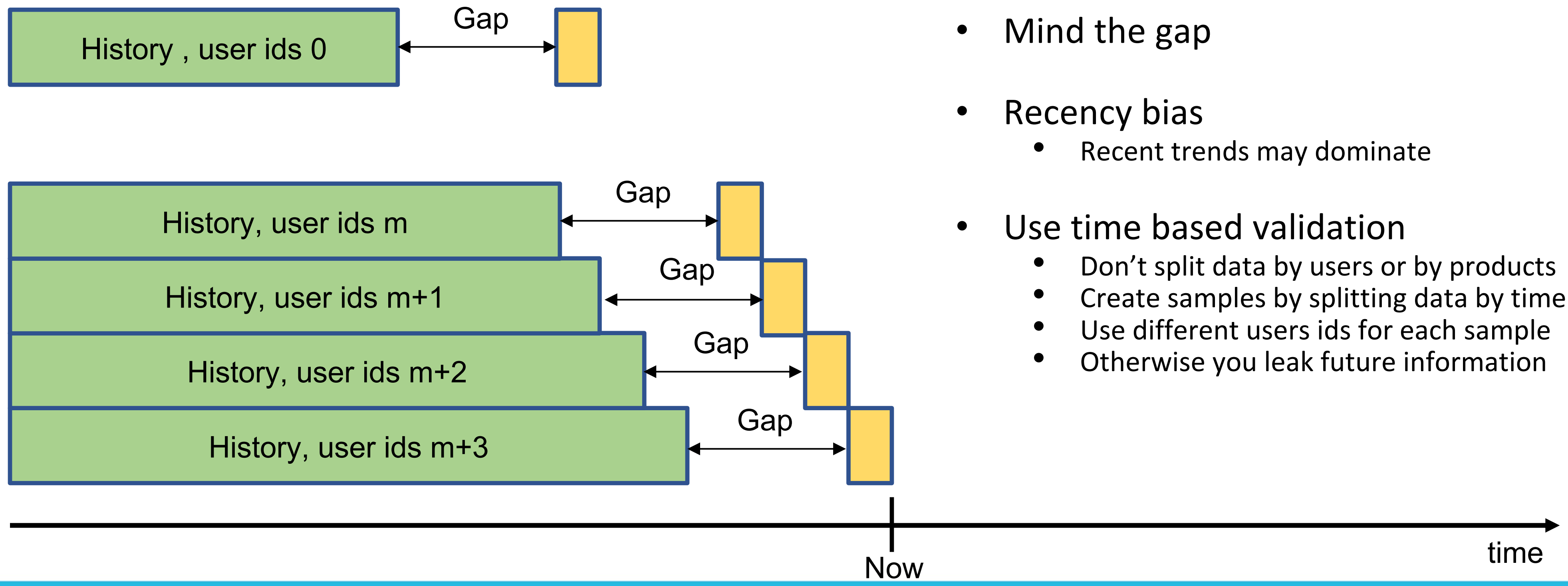
More Data Than Past Sales

- History data
  - Actual sales, not demand
  - Sales may have been limited by inventory
  - Sales maybe impacted by pricing, marketing campaigns, holidays, events

- Include inventory information in model input
  - Estimate unmet demand in the past
  - Product cannibalization

- Include pricing, campaigns, holidays, events
  - Multivariate time series

Recommender Systems



Most of what precedes applies to recsys!

- Mind the gap
- Recency bias
  - Recent trends may dominate
- Use time based validation
  - Don't split data by users or by products
  - Create samples by splitting data by time
  - Use different users ids for each sample
  - Otherwise you leak future information

Fraud detection

- Can be cast as a binary forecasting
- Given some history, is the current tentative payment a fraud?

Many issues

- Very imbalanced (fortunately)
- Ground truth is known with some delay
  - Large gap
- Little relevant history (when a fraud pattern is detected then fraudsters adapt)

Maybe better to see it as anomaly detection
- Model usual behavior (amounts, frequency, locations, etc)
- Flag large deviations

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”
- The file had 1800 transactions and 6 frauds among them

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”
- The file had 1800 transactions and 6 frauds among them
- The bank should spend time to label more data instead of exhausting all ML vendors…

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”
- The file had 1800 transactions and 6 frauds among them
- The bank should spend time to label more data instead of exhausting all ML vendors…

- In another case I was asked to help on a difficult wire transfer fraud
- Great data
- All labelled as fraud/non fraud!

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”
- The file had 1800 transactions and 6 frauds among them
- The bank should spend time to label more data instead of exhausting all ML vendors…

- In another case I was asked to help on a difficult wire transfer fraud
- Great data
- All labelled as fraud/non fraud!
- I got suspicious as this NEVER is the case
- Got assurance that this was real ground truth

Fraud detection Ground Truth

- Hard to get

- In one case I was asked to build a credit card fraud detection
- I got access to a 300M credit card transactions
- ”The labels are in this other file”
- The file had 1800 transactions and 6 frauds among them
- The bank should spend time to label more data instead of exhausting all ML vendors…

- In another case I was asked to help on a difficult wire transfer fraud
- Great data
- All labelled as fraud/non fraud!
- I got suspicious as this NEVER is the case
- Got assurance that this was real ground truth
- Once I delivered a model I learned that the GT was in fact the output of a legacy pipeline the customer wanted to replace…

Online Commerce Topics

- Decompose the problem into a regression
- Proper sampling for training
- Avoid future leaking into training
- Sales forecasting is more than univariate time series forecasting
- Lessons relevant to recys
- Fraud detection deserves a full presentation

kaggle™
days