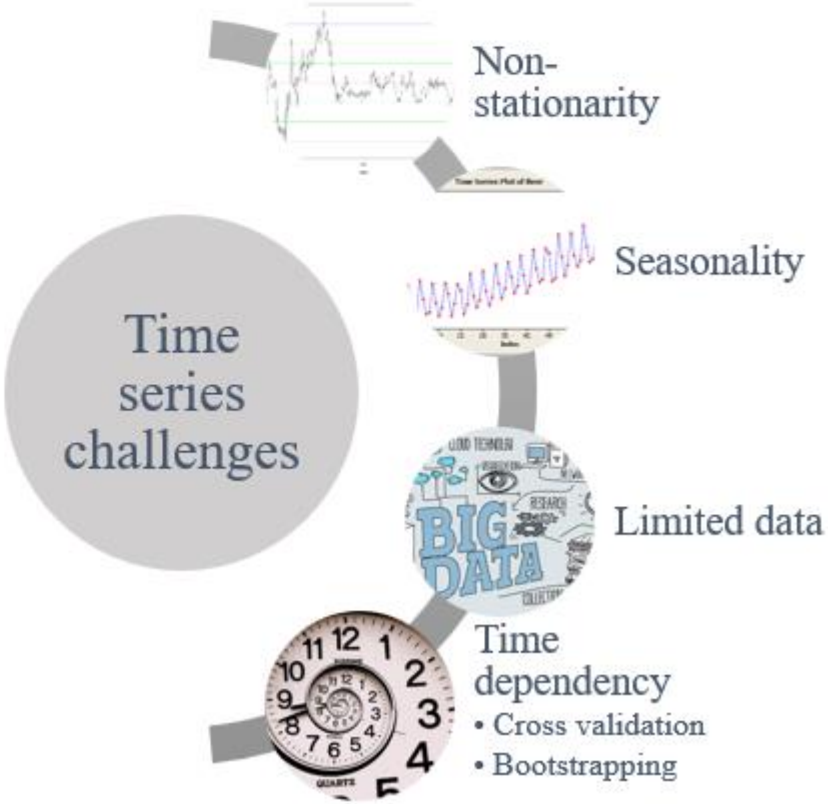
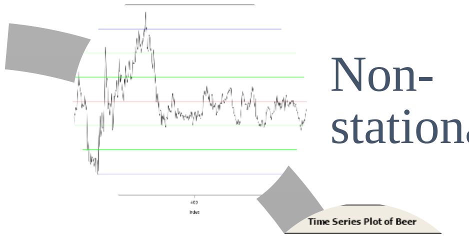
Module 10 – Part III
Challenges in Time Series Machine Learning



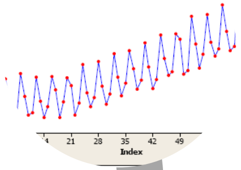
Jon M. Huntsman School of Business
Utah State University
Pedram Jahangiry

Challenges in Time Series Machine Learning

Time series challenges



Non-stationarity



Seasonality

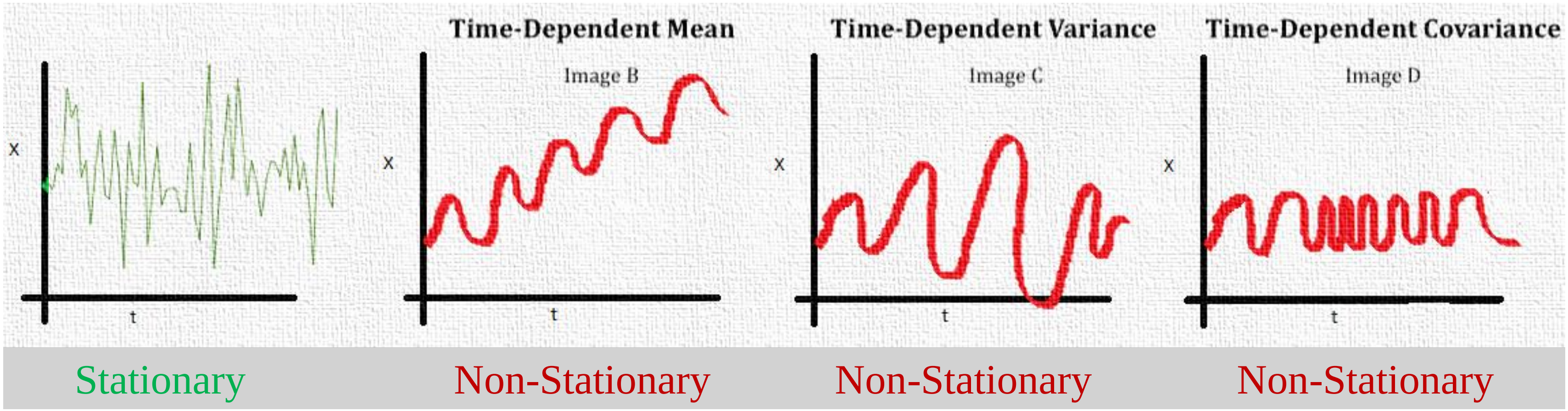
Limited data

Time dependency

- Cross validation
- Bootstrapping

Stationarity

- Stationary vs Non-Stationary Data. What makes a data set Stationary?
- In a stationary timeseries, the statistical properties do not depend on the time



- Data with trend and seasonality are NOT stationary!

Time Series Cross Validation

- With time series data, we cannot shuffle the data! TS data is not IID.
- We also need to avoid data leakage!

- The main time series CV methods are:
  1) Purged K-Fold CV
  2) Walk forward rolling / expanding window
  3) Combinatorial purged CV

JON M. HUNTSMAN SCHOOL OF BUSINESS UtahStateUniversity

Pedram Jahangiry

Purged K-Fold CV

- Leakage takes place when the training set contains information that also appears in the testing set.
- Leakage will enhance the model performance
- Solution: Purging and Embargoing

- Purged K-Fold CV: Adding purging and embargoing whenever we produce a train/test split in K-Fold CV.

Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

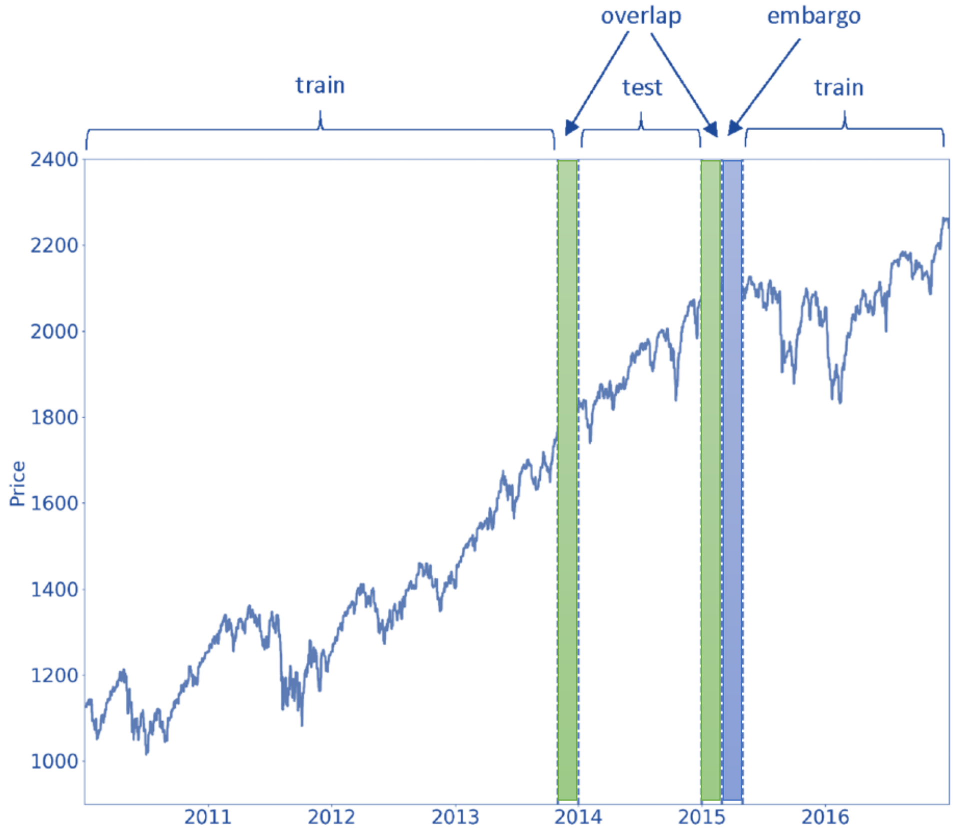
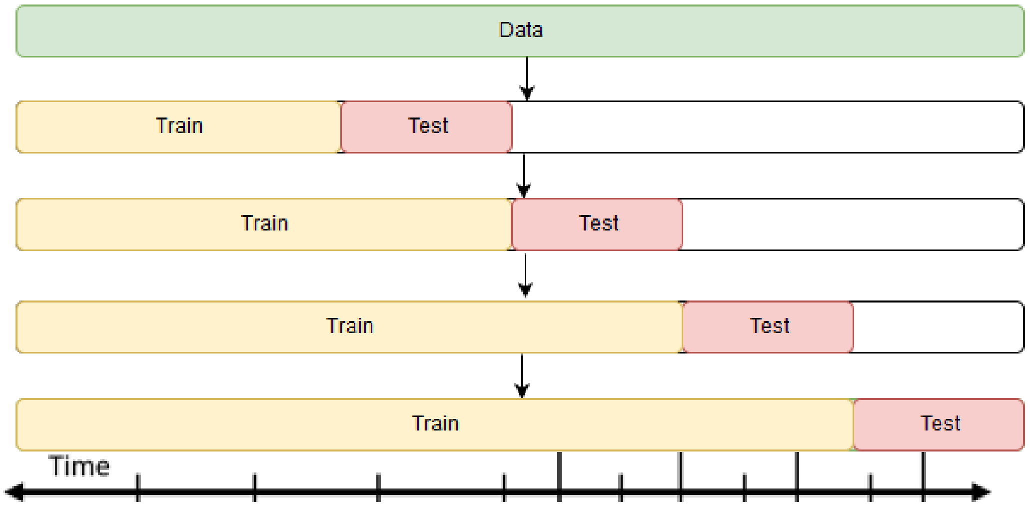


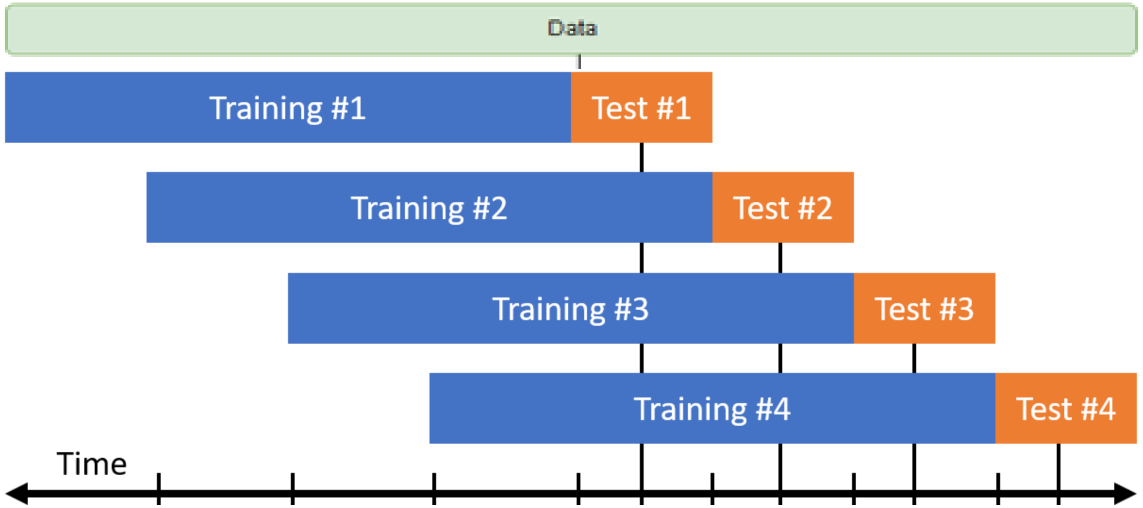
FIGURE 7.3 Embargo of post-test train observations

Walk Forward Cross Validation

Walk forward cross validation
Expanding windows

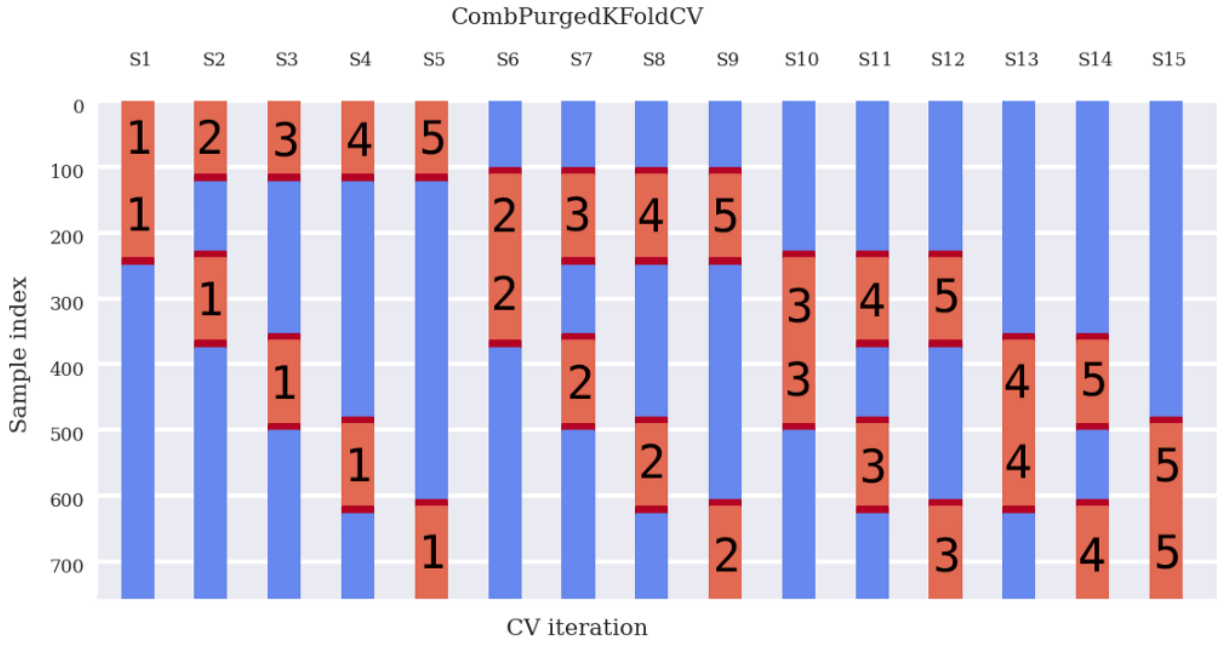


Walk forward cross validation
Rolling windows



Combinatorial Purged Cross Validation (CPCV)

- The goal is to generate multiple unique back-test path that span the entire data set.
- In each path, we can look at the model's OOS performance for the entire time period.



Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

Time Series Bootstrapping

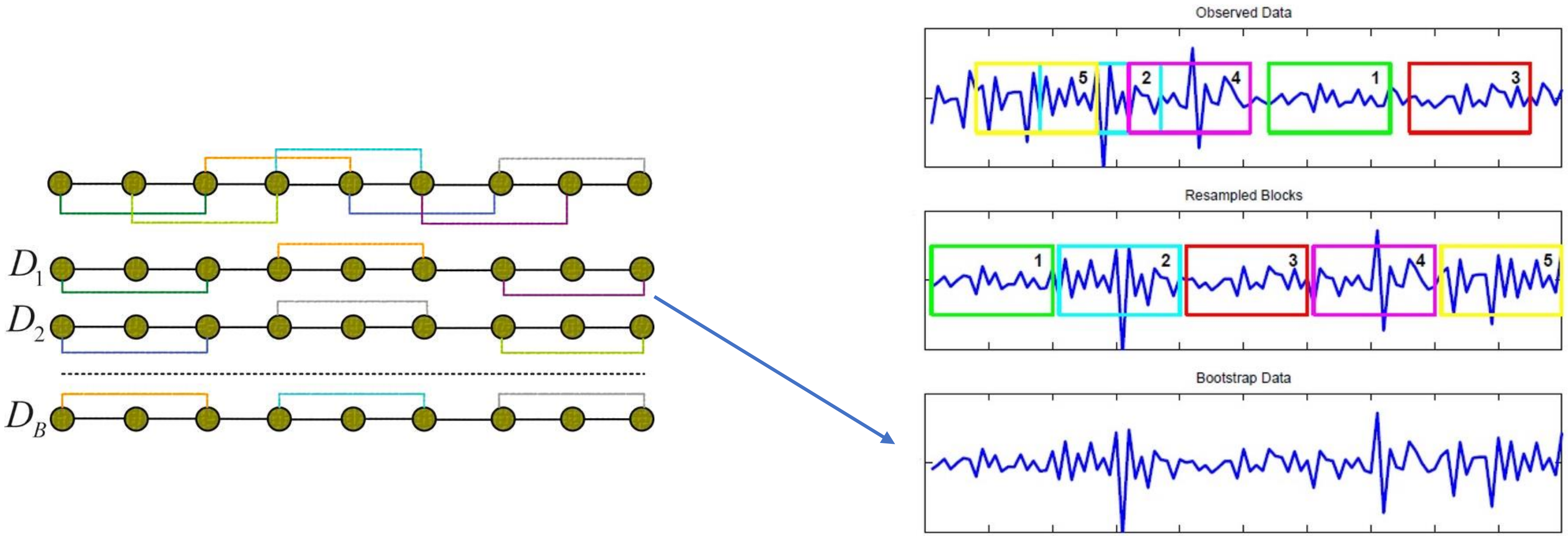
- IID bootstrapping (random sample with replacement) does not work for time series data with temporal dependency.
- Time series Bootstrapping methods:
  - Parametric (based on models with iid residuals and resampling from residuals. Example: ARIMA bootstrap)
  - Non-parametric block bootstrap (data is directly resampled. Assumption: blocks can be samples so that they are approximately iid)
    - Moving Block Bootstrap (MBB)
    - Circular Block Bootstrap (CBB)
    - Stationary Bootstrap (SB)

JON M. HUNTSMAN SCHOOL OF BUSINESS UtahStateUniversity

Pedram Jahangiry

Moving Block Bootstrap (MBB)

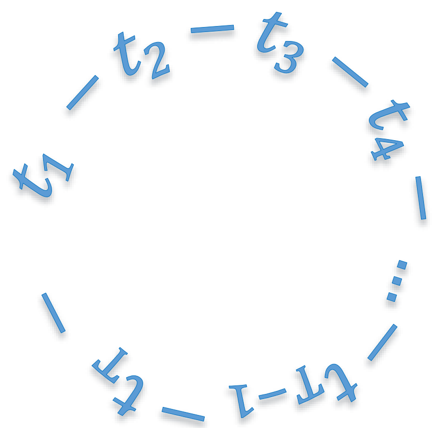
- Moving Block Bootstrap, samples overlapping fixed size blocks of m consecutive observations.
- Blocks starts at indices 1, … , T-m+1



Circular Block Bootstrap (CBB)

- CBB is a simple extension of MBB which assumes the data live on a circle so that $y_{T+1} = y_1, \; y_{T+2} = y_2$, etc.

- CBB has better finite sample properties since all data points get sampled with equal probability.



Stationary Bootstrap (SB)

- In SB, the block size is no longer fixed.
- Chooses an average block size of m rather than an exact block size.
- Popularity of SB stems from difficulty in determining optimal m
- Once applied to stationary data, the resampled pseudo time series by SB are stationary. This is not the case for MBB and CBB.