

Analyzing Time Series at Scale with Cluster Analysis in R

Workshop for Ukraine

Rami Krispin, Oct 17, 2024

About Me

- Data Science and Engineering manager
- Forecasting
- MLOps
- Docker Captain
- Open Source
- Author



Agenda

- Problem Statement
- Forecasting at Scale
- Feature-based Time Series Analysis
- Demo

[https://github.com/RamiKrispin/
ts-cluster-analysis-r](https://github.com/RamiKrispin/ts-cluster-analysis-r)

Poll

Are you familiar with the following?

- R
- Time series analysis
- Forecasting
- Cluster analysis

Definitions

- **Time series** - a vector of data points measured over time
- **Regular time series** - the data measurement is equally spaced (i.e., hour, day, month, etc.)
- **Irregular time series** - the data measurement is not equally spaced
- **Forecast** - a prediction of observations in the future with a time dimension

Definitions

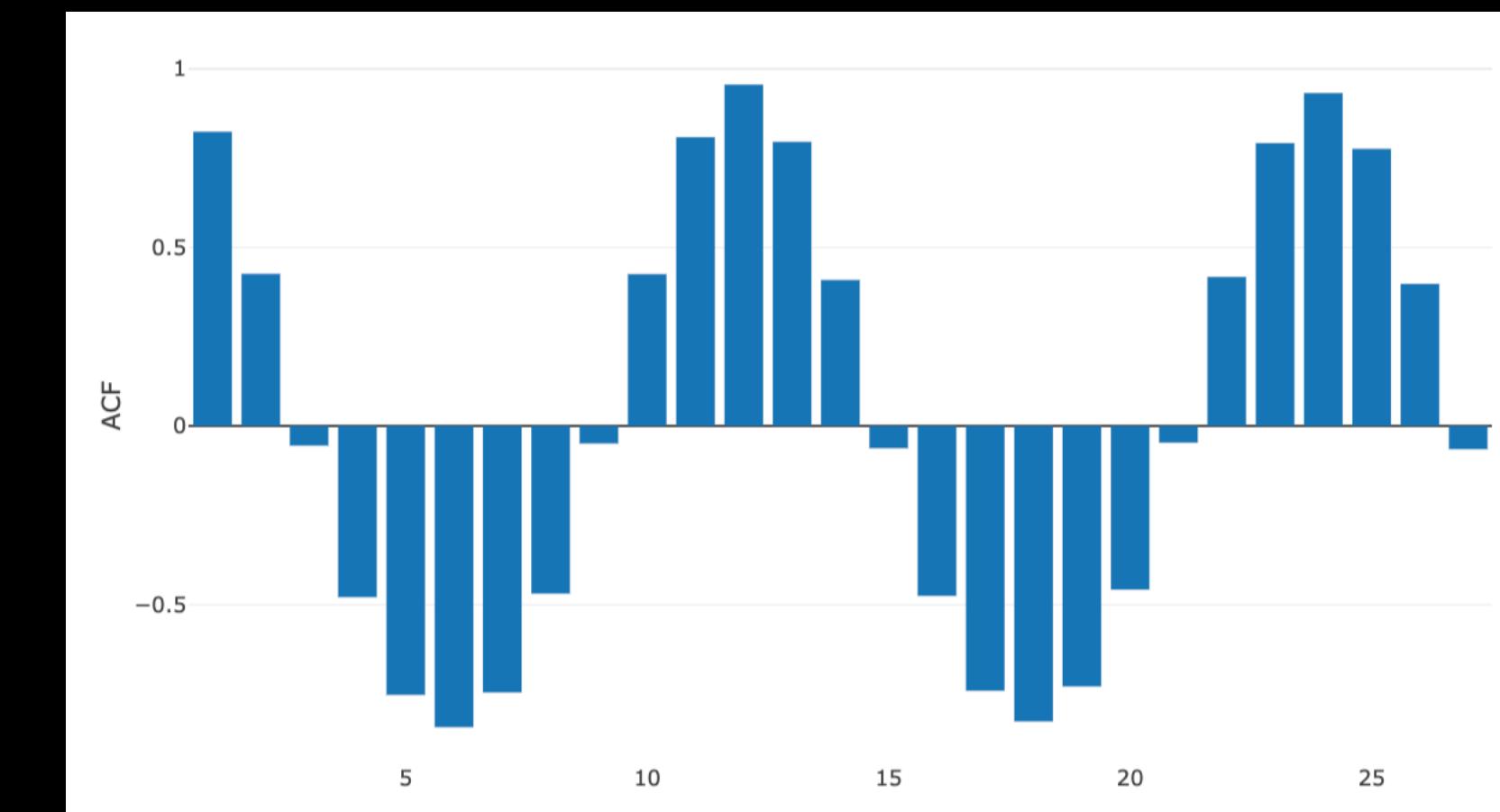
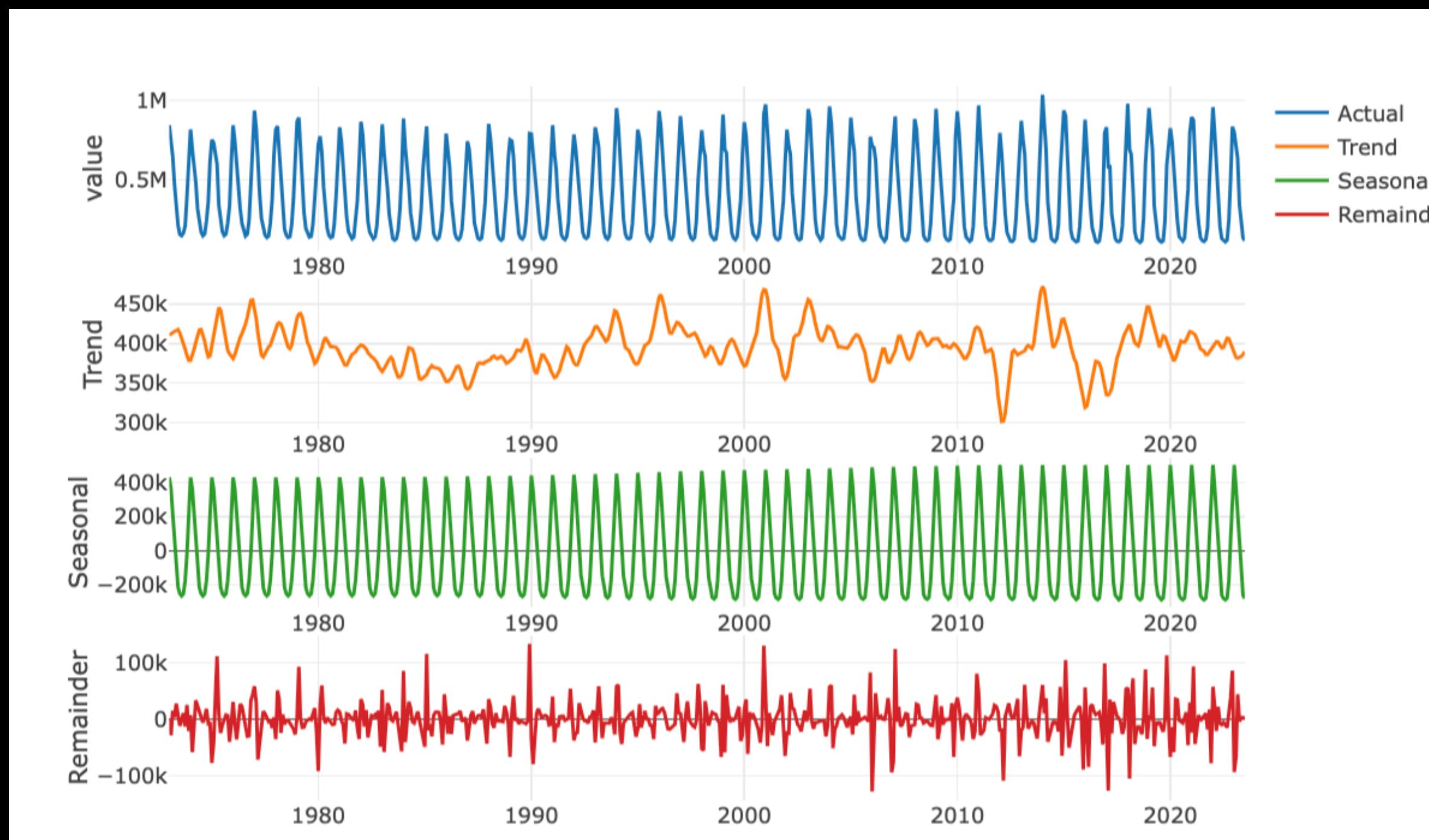
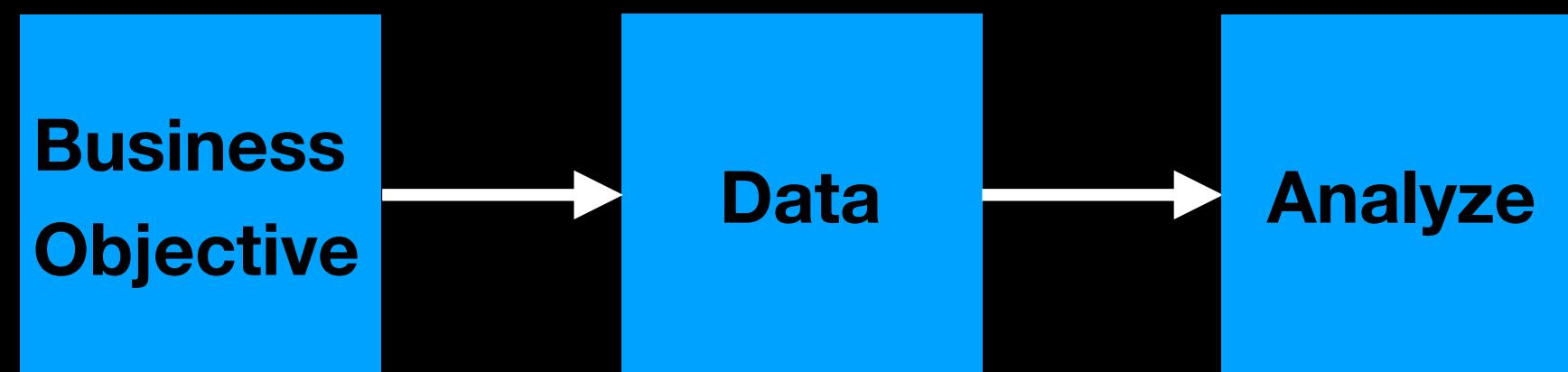
- **Time series** - a vector of data points measured over time
- **Regular time series** - the data measurement is equally spaced (i.e., hour, day, month, etc.)
- **Irregular time series** - the data measurement is not equally spaced
- **Forecast** - a prediction of observations in the future with a time dimension

Problem Statement

Scaling Limitation

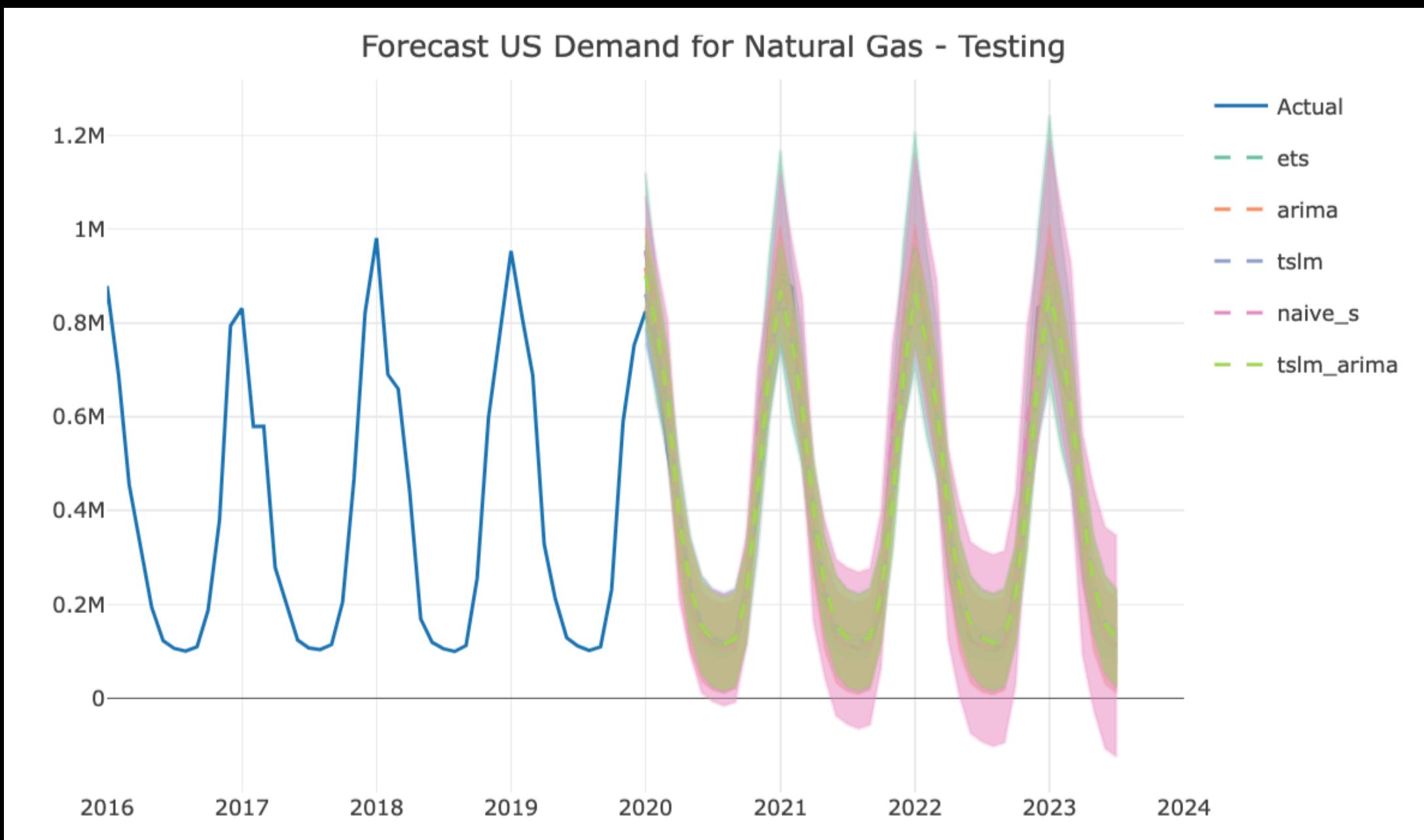
Forecasting A Single Series

Traditional Approach



Forecasting A Single Series

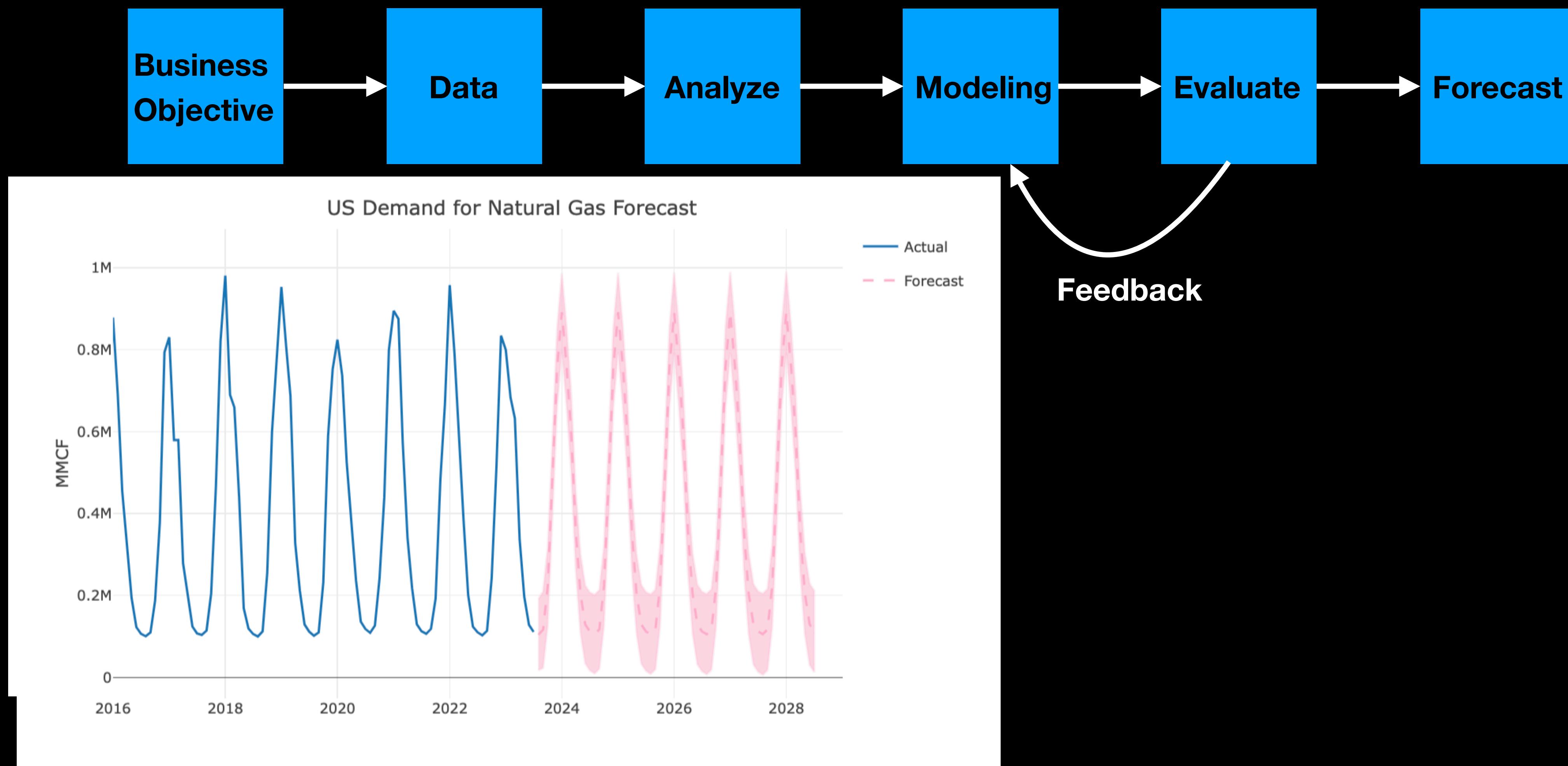
Traditional Approach



model	mape	rmse
arima	0.0612	3.76e18
ets	0.0681	7.42e18
naive_s	0.0928	1.88e19
tslm_arima	0.115	7.89e18
tslm	0.115	7.11e18

Forecasting A Single Series

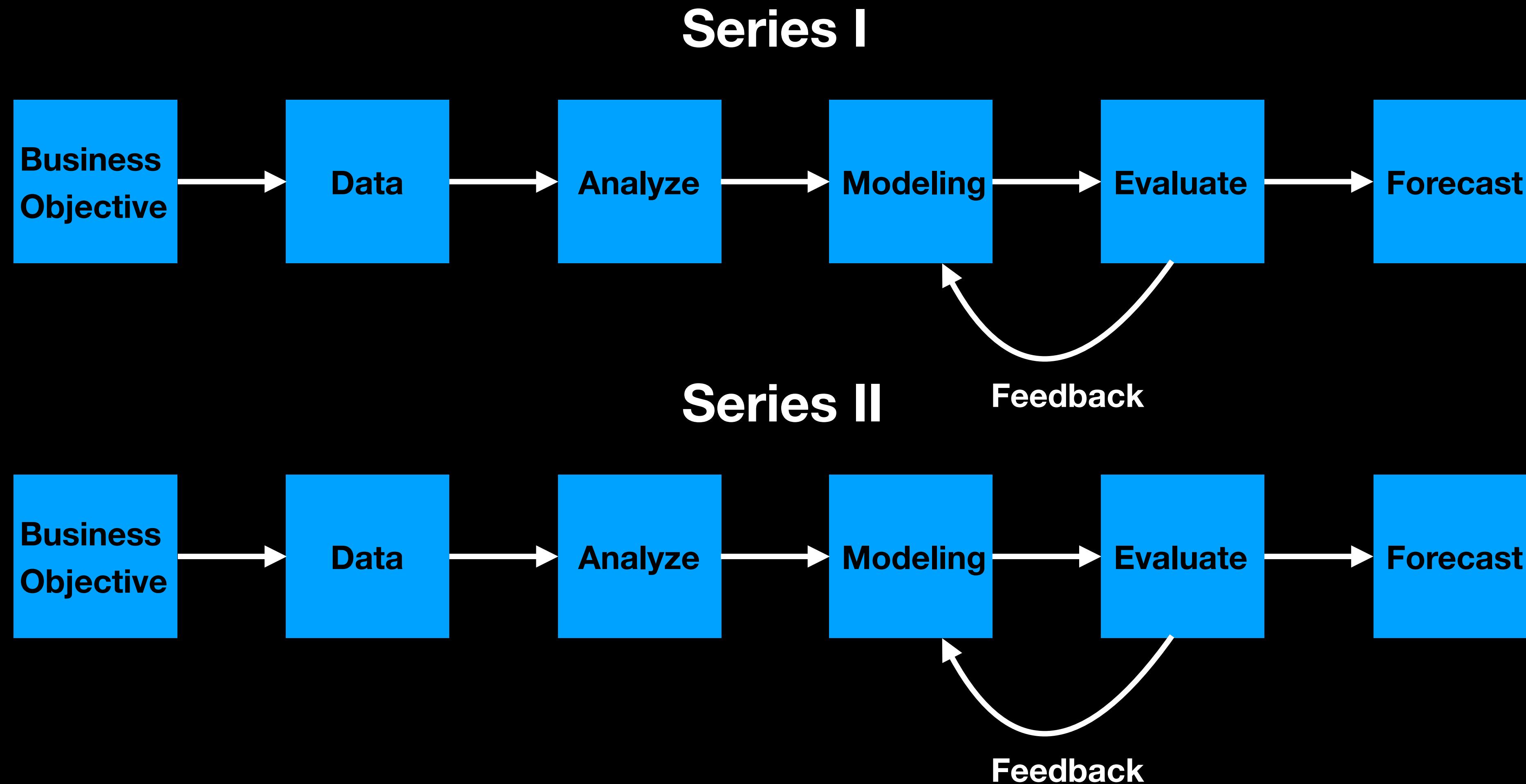
Traditional Approach



What if you have two series to forecast?

Forecasting A Single Series

Traditional Approach



What if you have two **hundreds**
series to forecast?



Forecasting at Scale

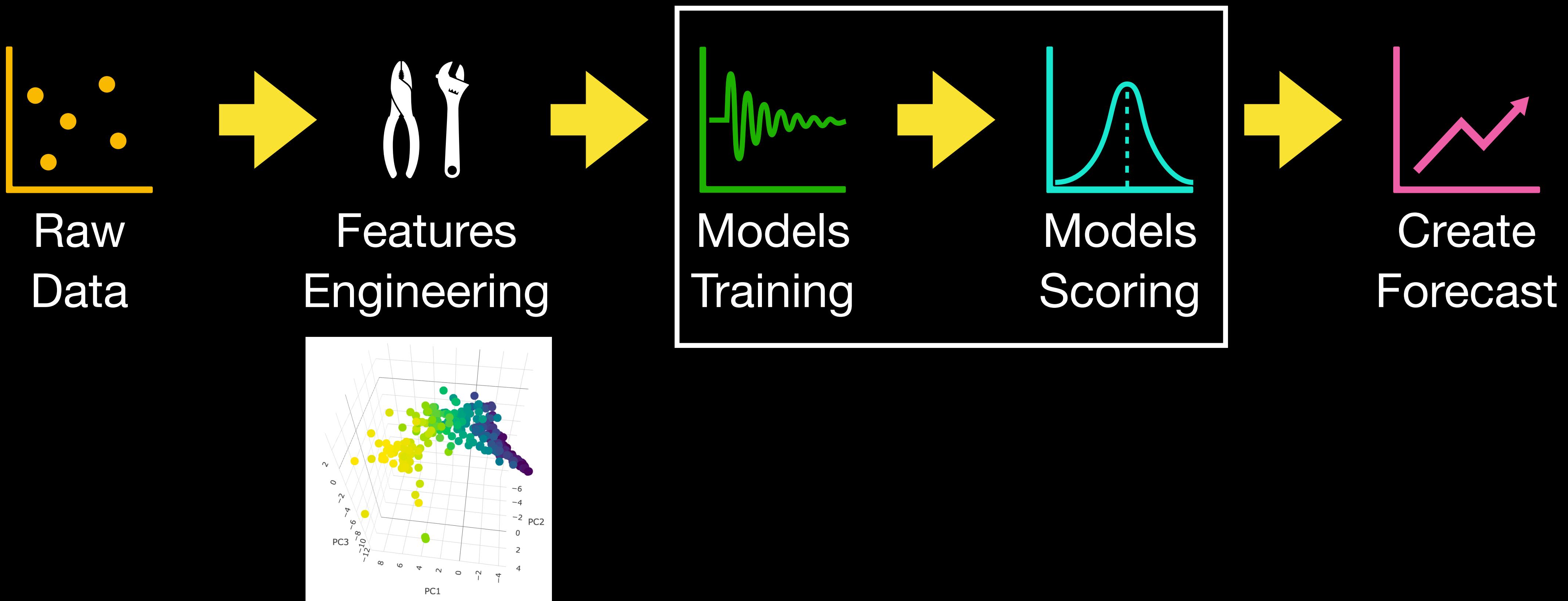
Forecasting at Scale

Definition

- The level of effort of adding additional series is non linear with marginal decay
- Modeling approaches
- Infrastructure dependency
- Trade-off - potential drop in accuracy

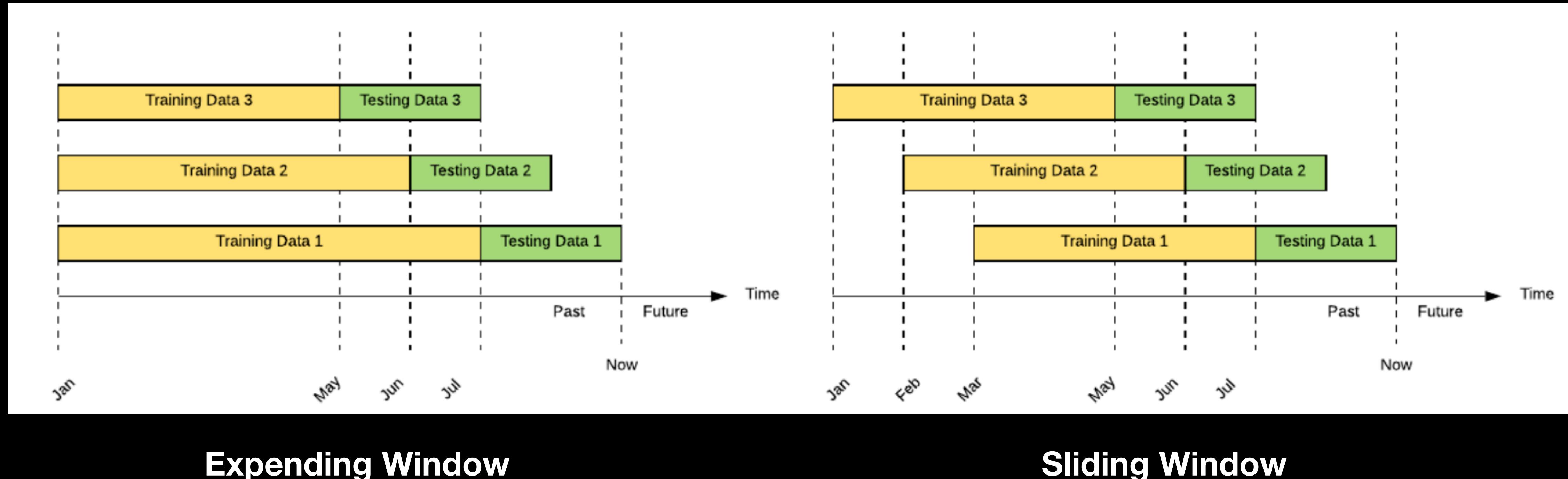
Forecasting at Scale

In a Nutshell



Backtesting in a nutshell is the time series equivalent to ML cross validation

Backtesting Approach



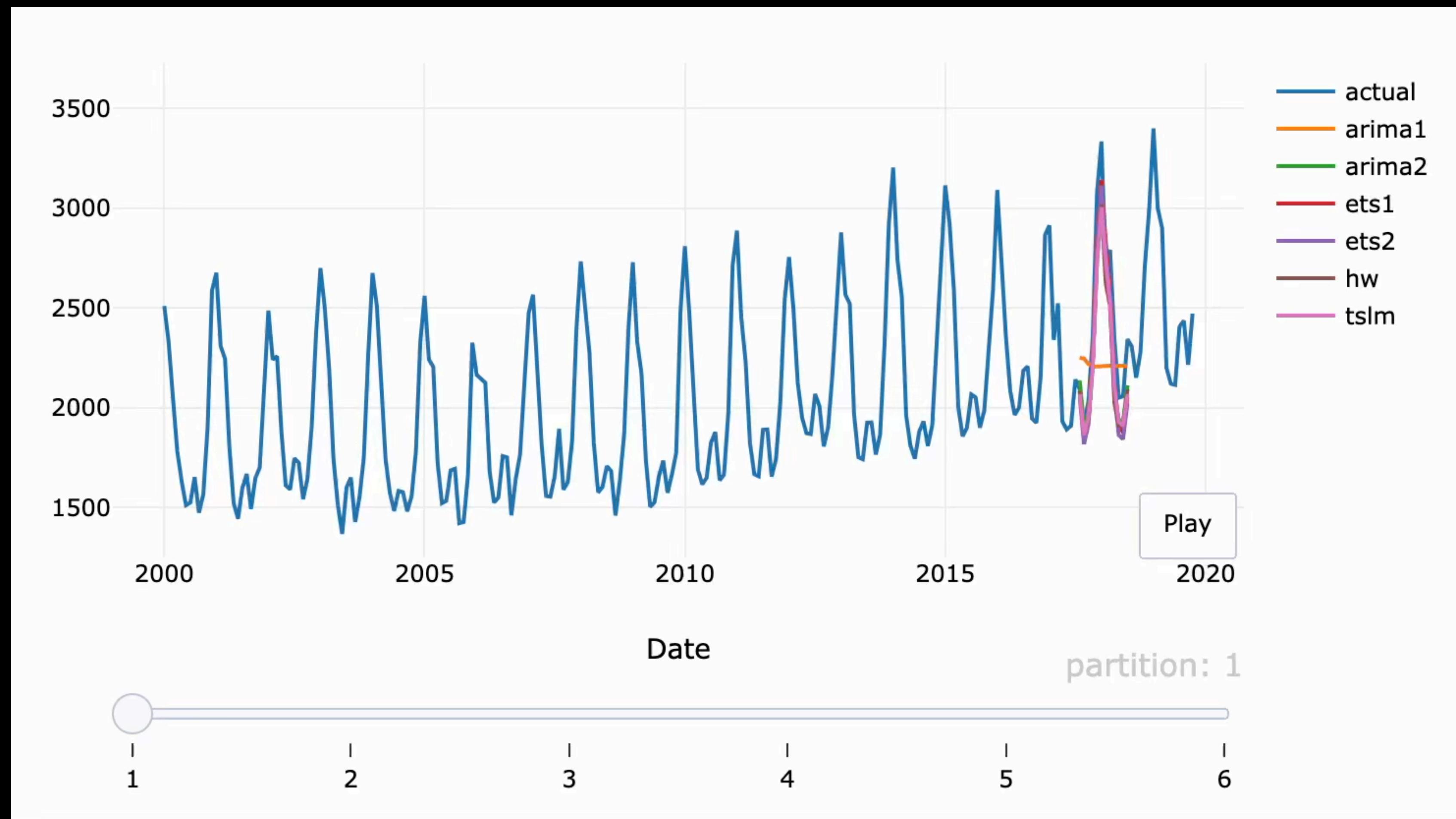
Horse Racing



Source: Wikipedia https://en.wikipedia.org/wiki/Horse_racing

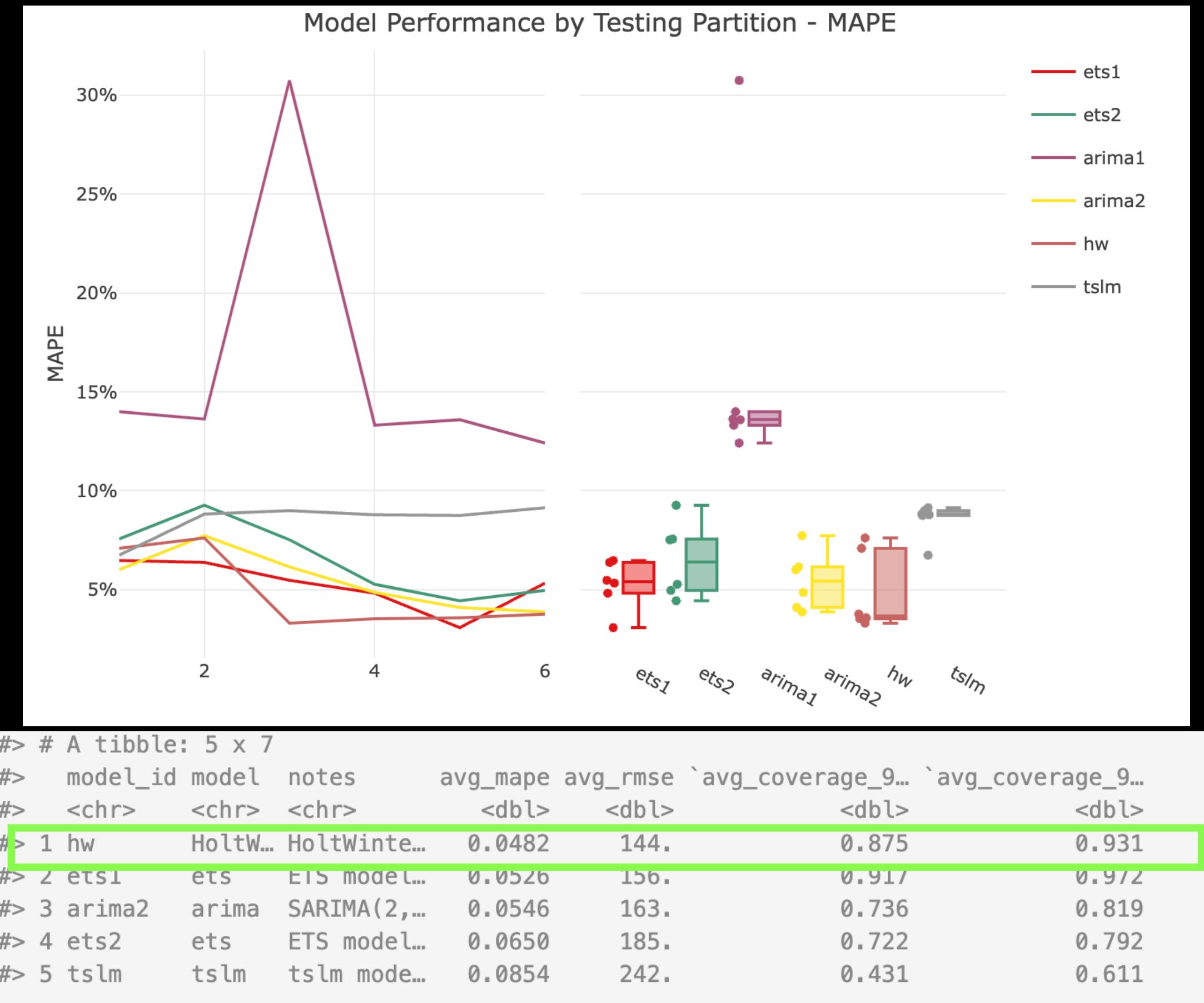
Horse Racing with Backtesting

Training



Horse Racing with Backtesting

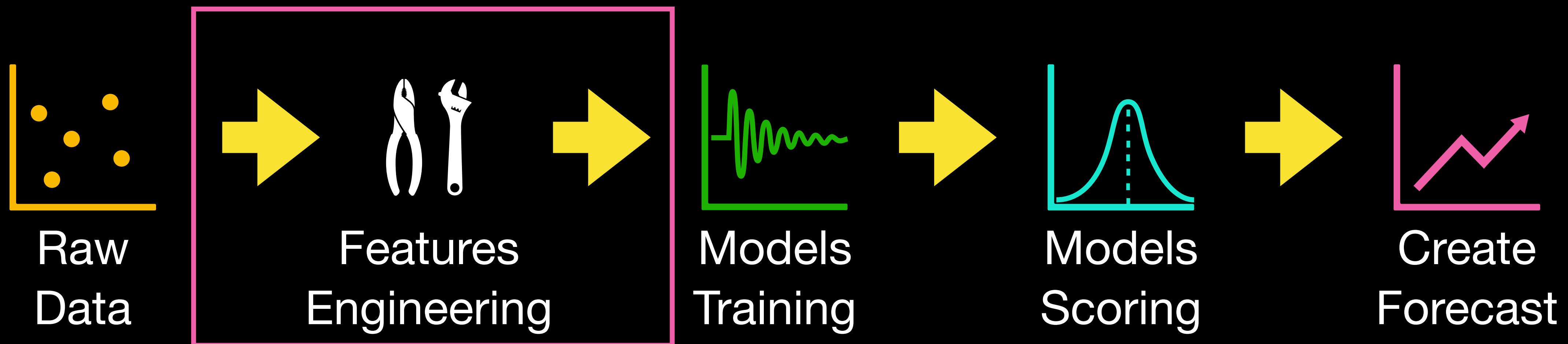
Models Scoring and Ranking



Backtesting + Horse Racing =
AutoML

Forecasting at Scale

In a Nutshell



Main Challenges

How to identify the right
modeling approaches?

How to identify the right features
to use?

Feature-based Time Series Analysis

Feature-based Time Series Analysis

Feature-based time series analysis

DATE

16 September 2019

TOPICS

TIME SERIES GRAPHICS STATISTICS R TIDYVERTS ANOMALIES
DATA SCIENCE

In my [last post](#), I showed how the `feasts` package can be used to produce various time series graphics.

The `feasts` package also includes functions for computing FEatures And Statistics from Time Series (hence the name). In this post I will give three examples of how these might be used.

```
library(tidyverse)
library(tsibble)
library(feasts)
```

Exploring Australian tourism data

I used this example in [my talk at useR!2019 in Toulouse](#), and it is also the basis of [a vignette in the package](#), and a recent [blog post by Mitchell O'Hara-Wild](#). The data set contains domestic tourist visitor nights in Australia, disaggregated by State, Region and Purpose.

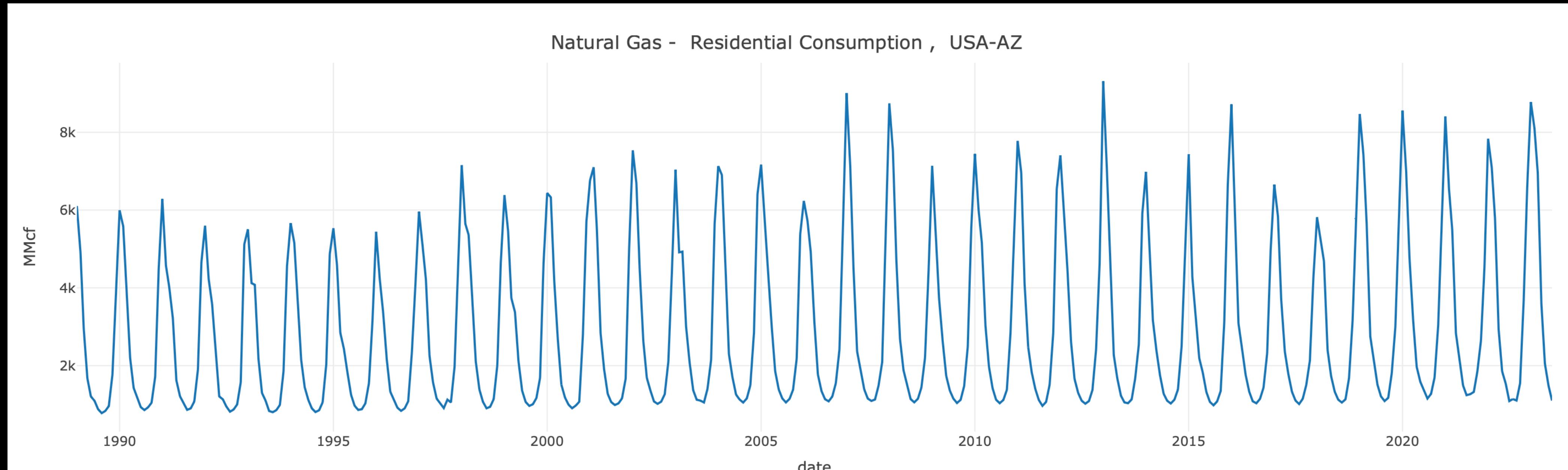
Cluster Analysis

Definition

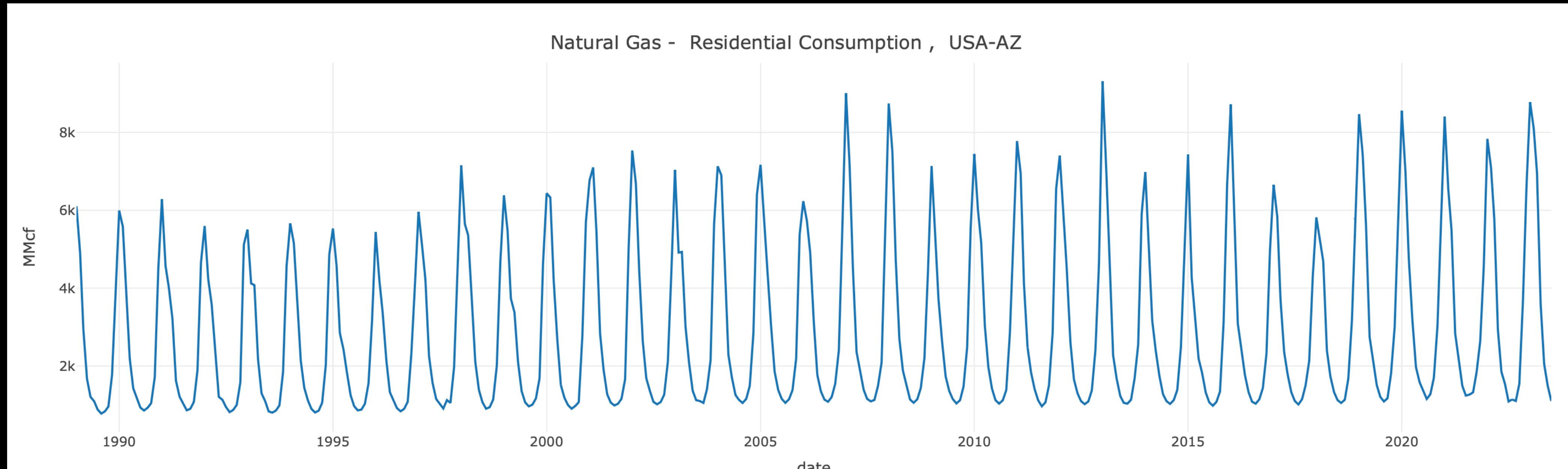
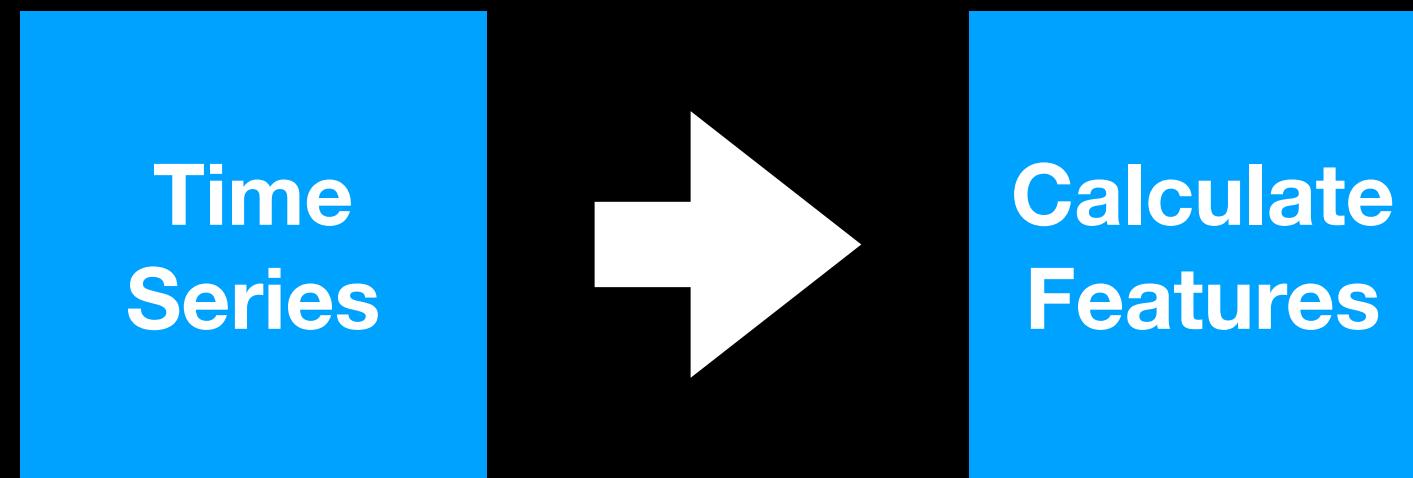
- **Cluster analysis** - is the task of group observations that share similar characteristics based on some criteria. It is a unsupervised leaning method
- **Principal components analysis (PCA)** - is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.
- **K-means clustering** - is a method of vector quantization that aims to partition n observations into k clusters

Feature-based Time Series Analysis

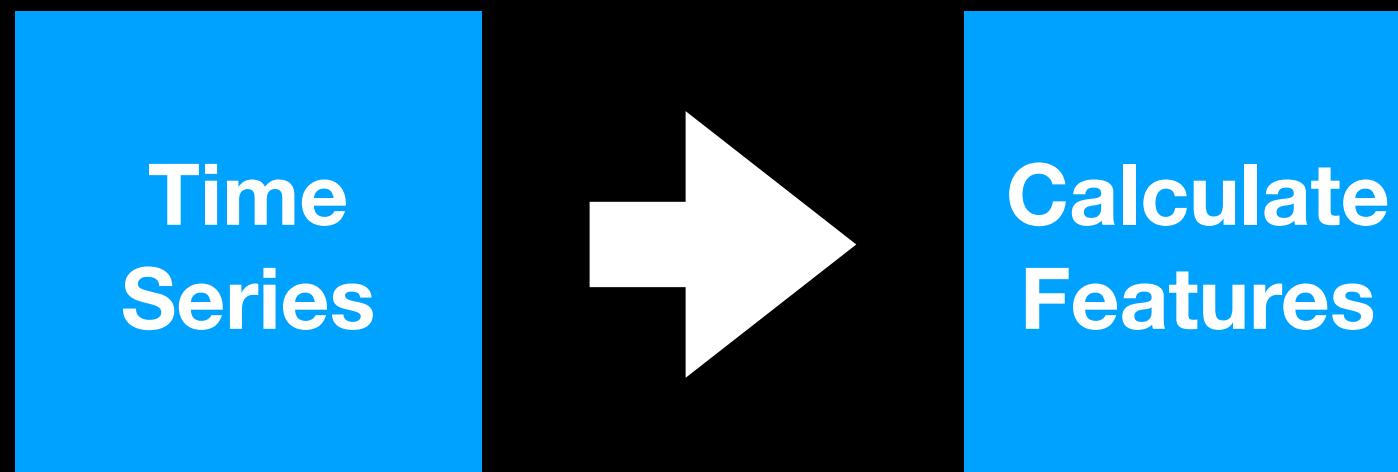
Time
Series



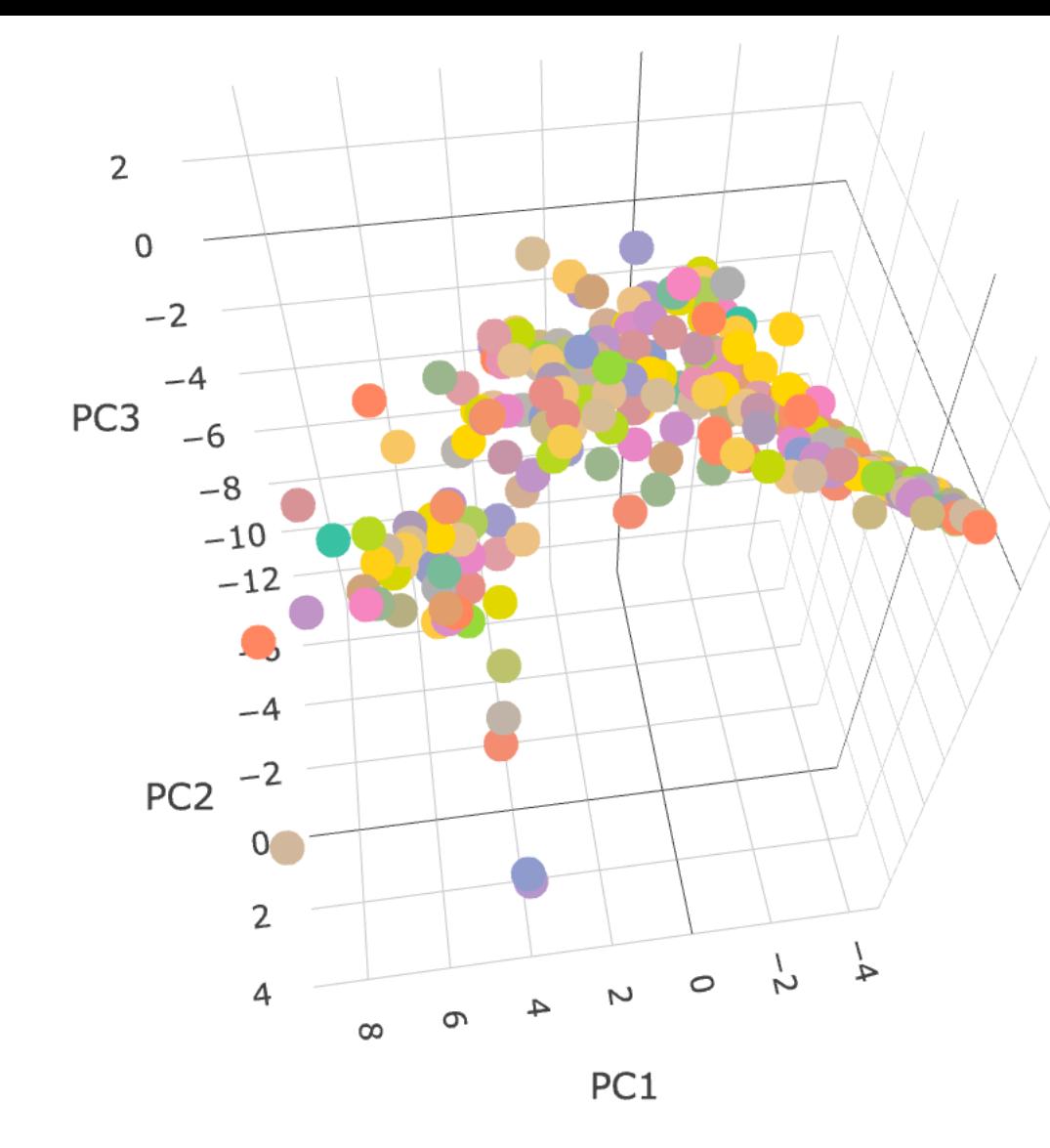
Feature-based Time Series Analysis



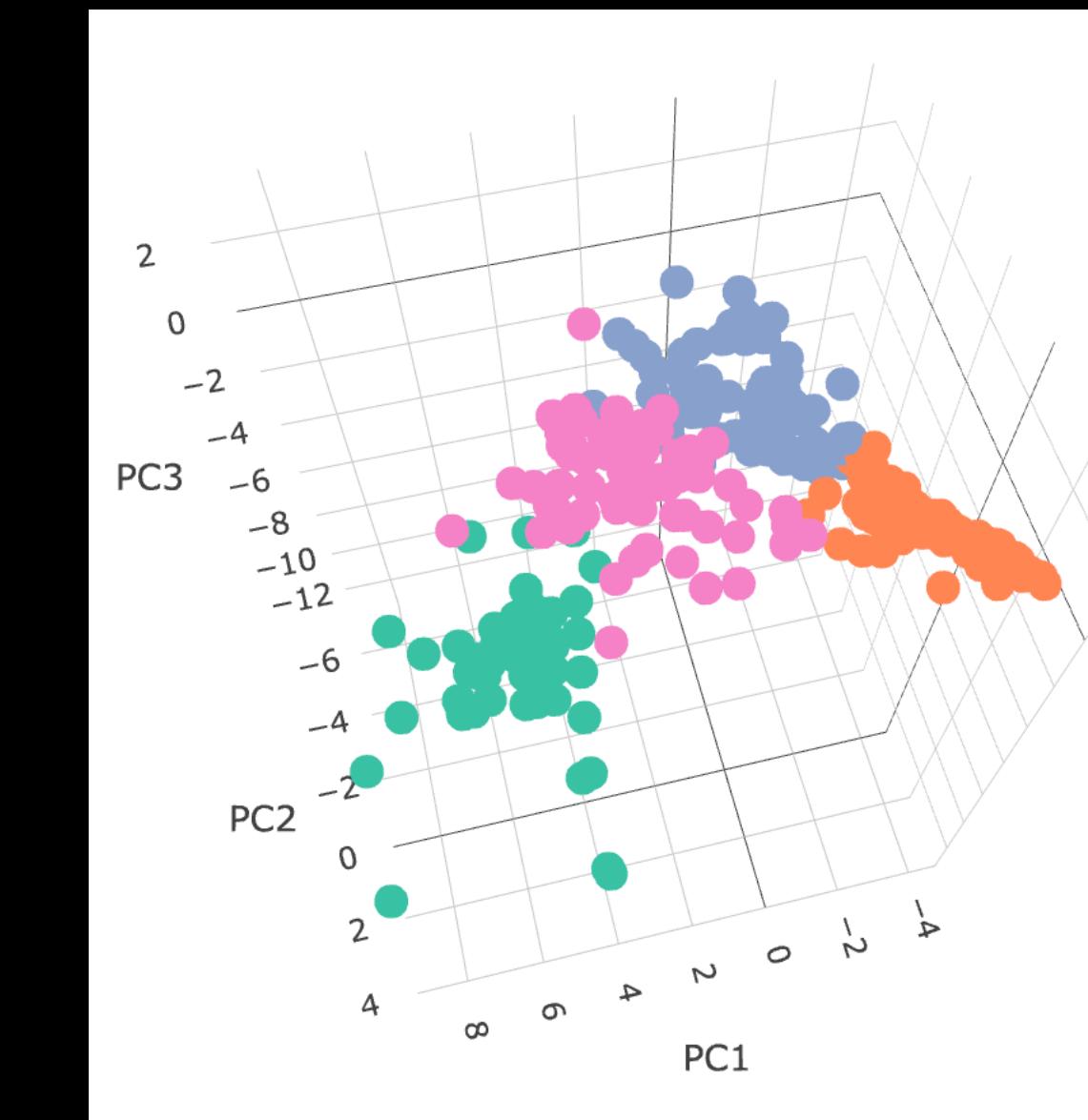
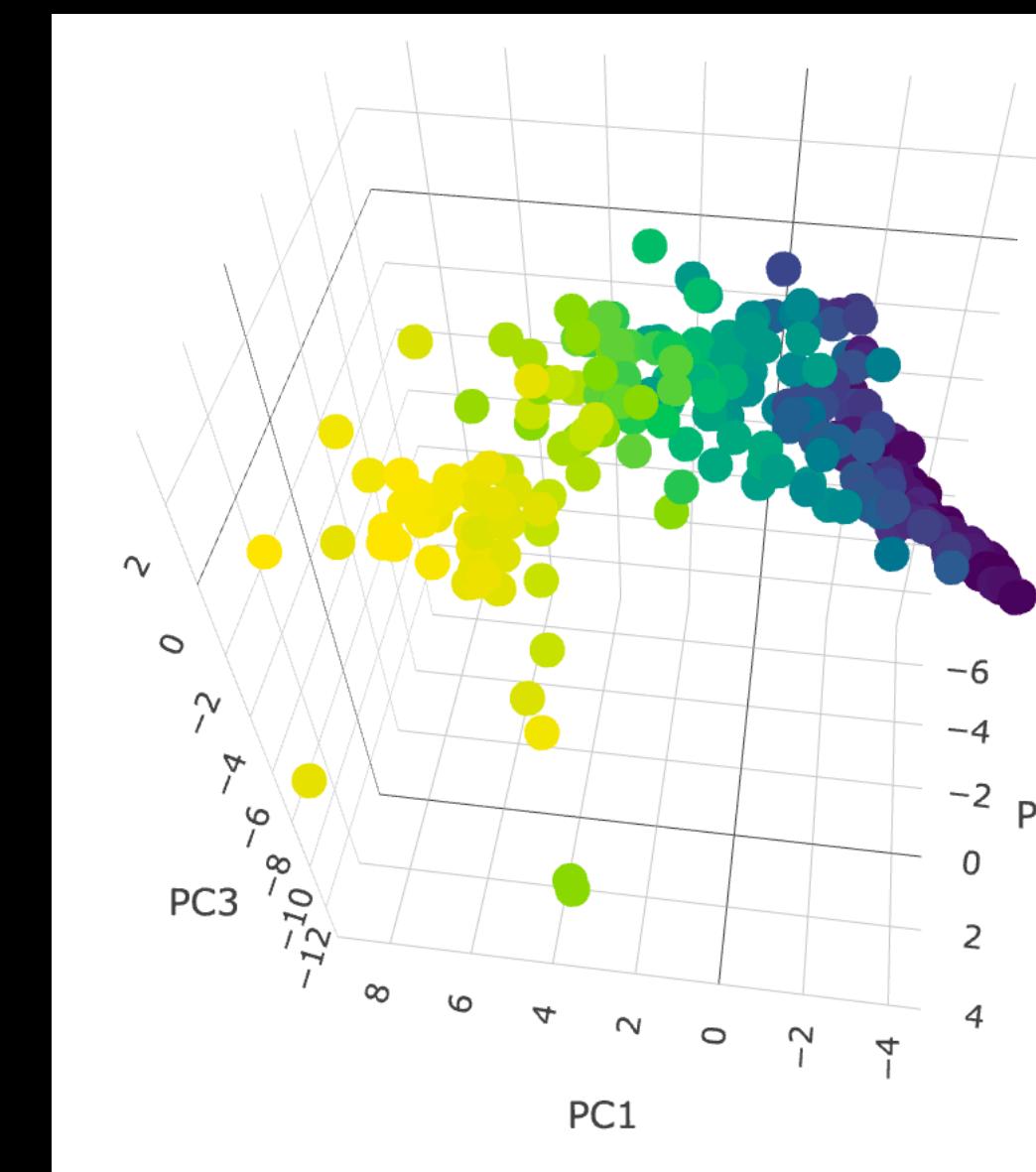
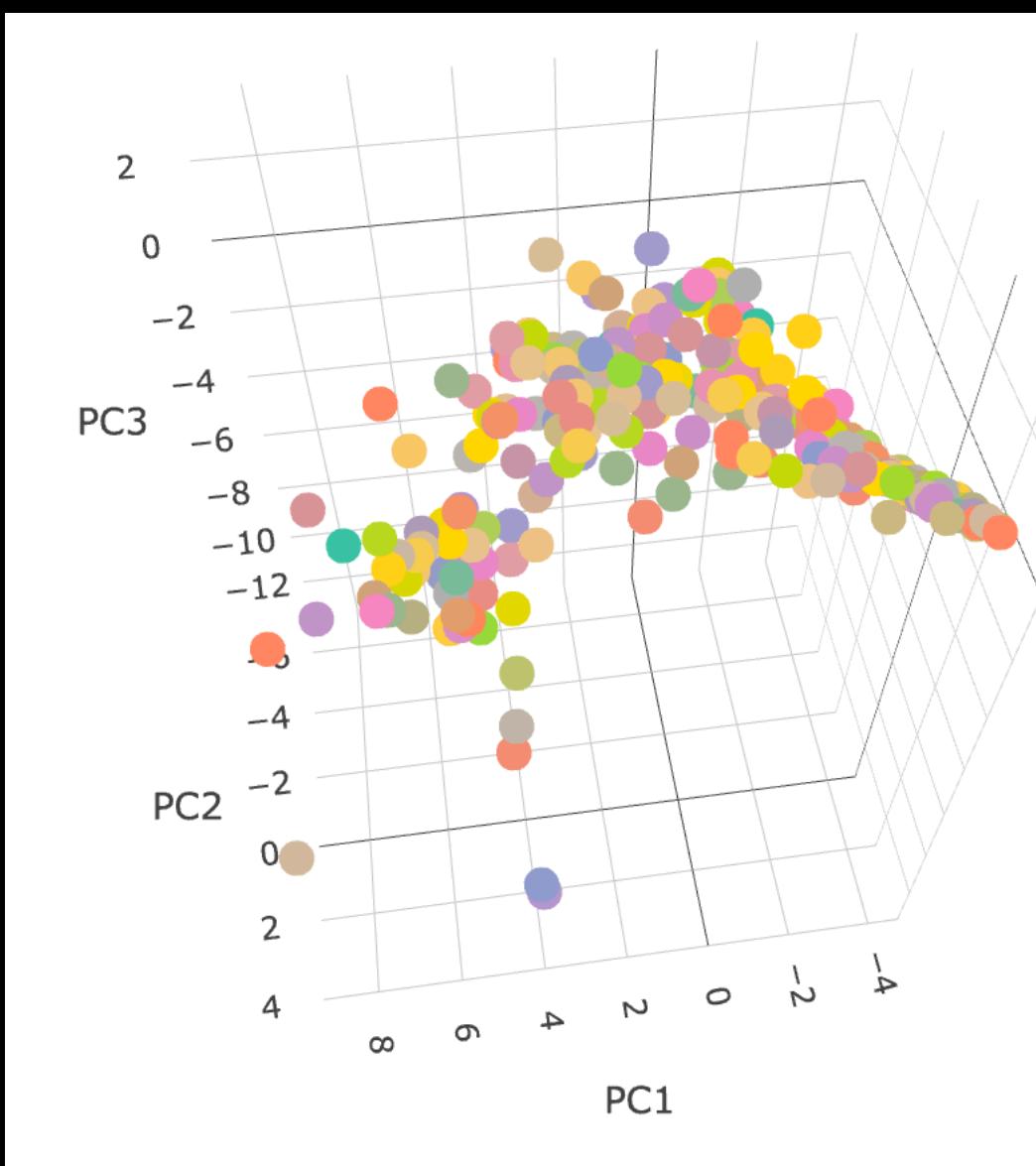
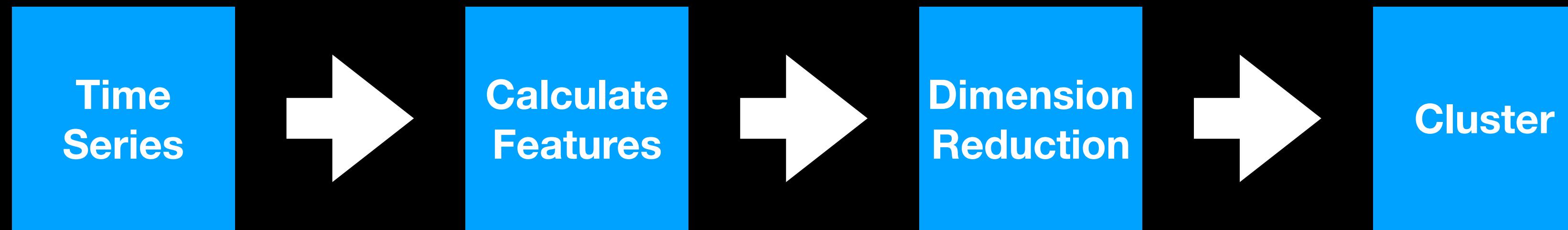
Feature-based Time Series Analysis



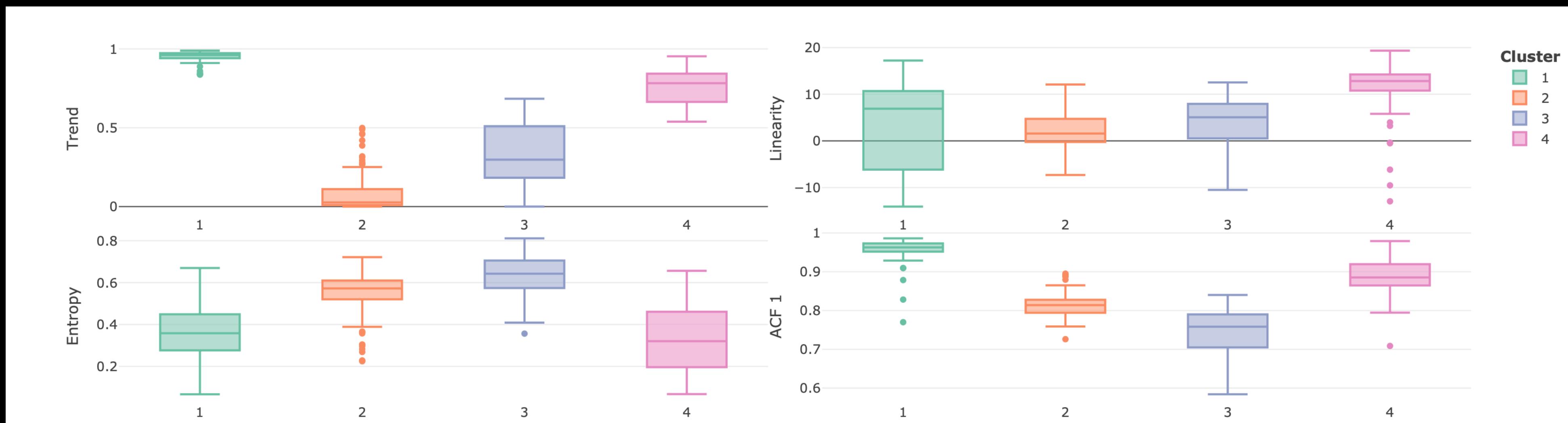
Feature-based Time Series Analysis



Feature-based Time Series Analysis



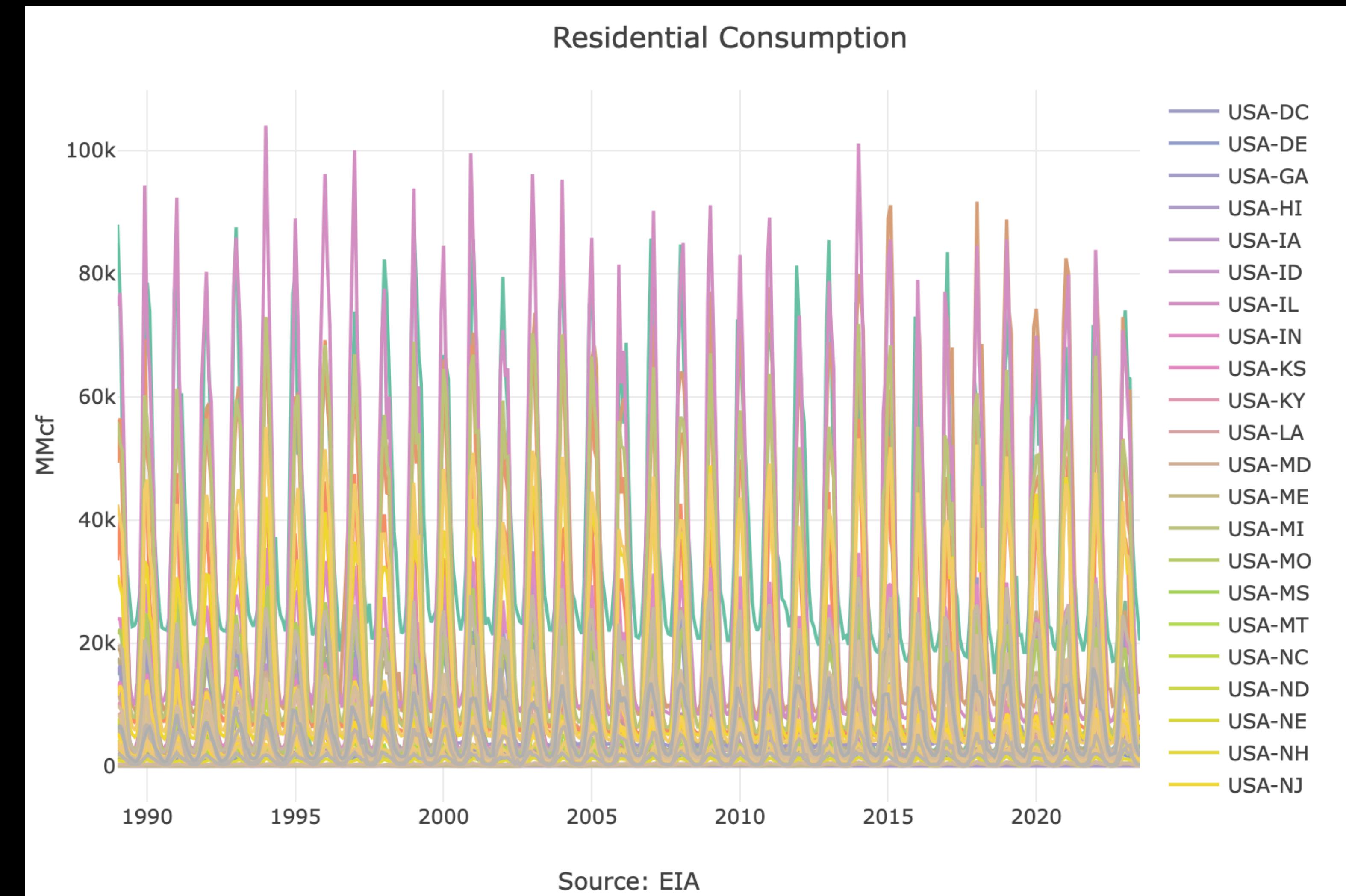
Feature-based Time Series Analysis



Demo

Feature-based Time Series Analysis

- 314 series
- 5 categories
- Monthly



Feature-based Time Series Analysis

Tools

- **Data** - ElAapi, dplyr, tsibble
- **Features engineering** - tsfeatures
- **Cluster analysis** - stats
- **Data visualization** - plotly, shiny

Summary

- Clusters analysis methods enable to produce insights at scale
- Enables to identify required features and define models strategy at the cluster level
- Potential features automation
- Accuracy trade-off
- Monitoring

Resources

- Feature-based time series analysis, Rob Hyndman:
 - Blog post: <https://robjhyndman.com/hyndsight/fbtsa/>
 - Seminar: <https://robjhyndman.com/seminars/fbtsa-ssc/>
- Code: <https://github.com/RamiKrispin/ts-cluster-analysis-r>
- Uber Engineering: <https://www.uber.com/en-DE/blog/backtesting-at-scale/>