Forecasting at Scale

SDSU Data Science Club

Rami Krispin, Oct 26, 2023

Intro

- Data Science and Eng Manager
- Forecasting
- MLOps
- Open Source
- Author

Agenda

Forecast Single Series

Horse Racing

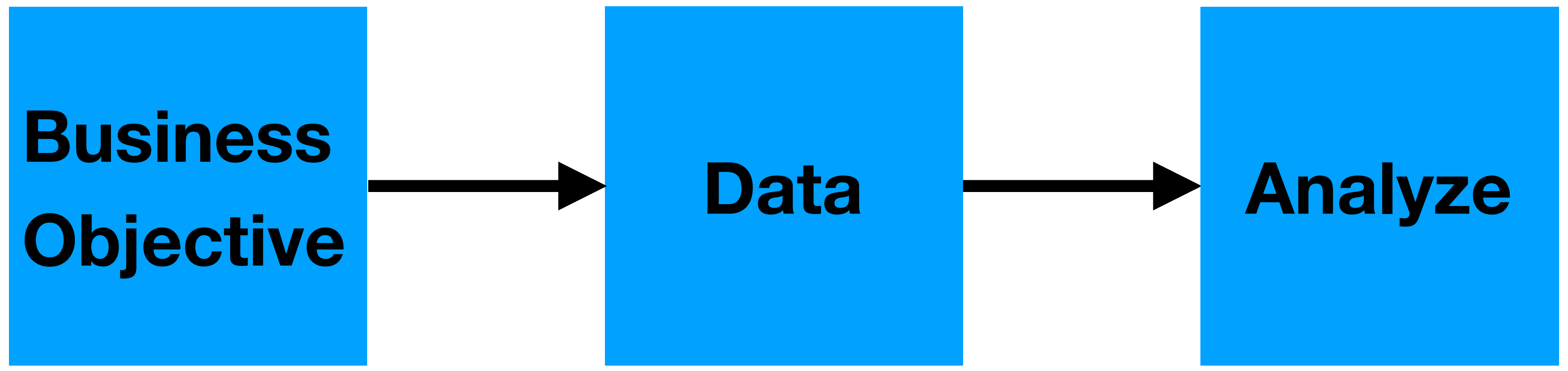
Analyze Multiple Series

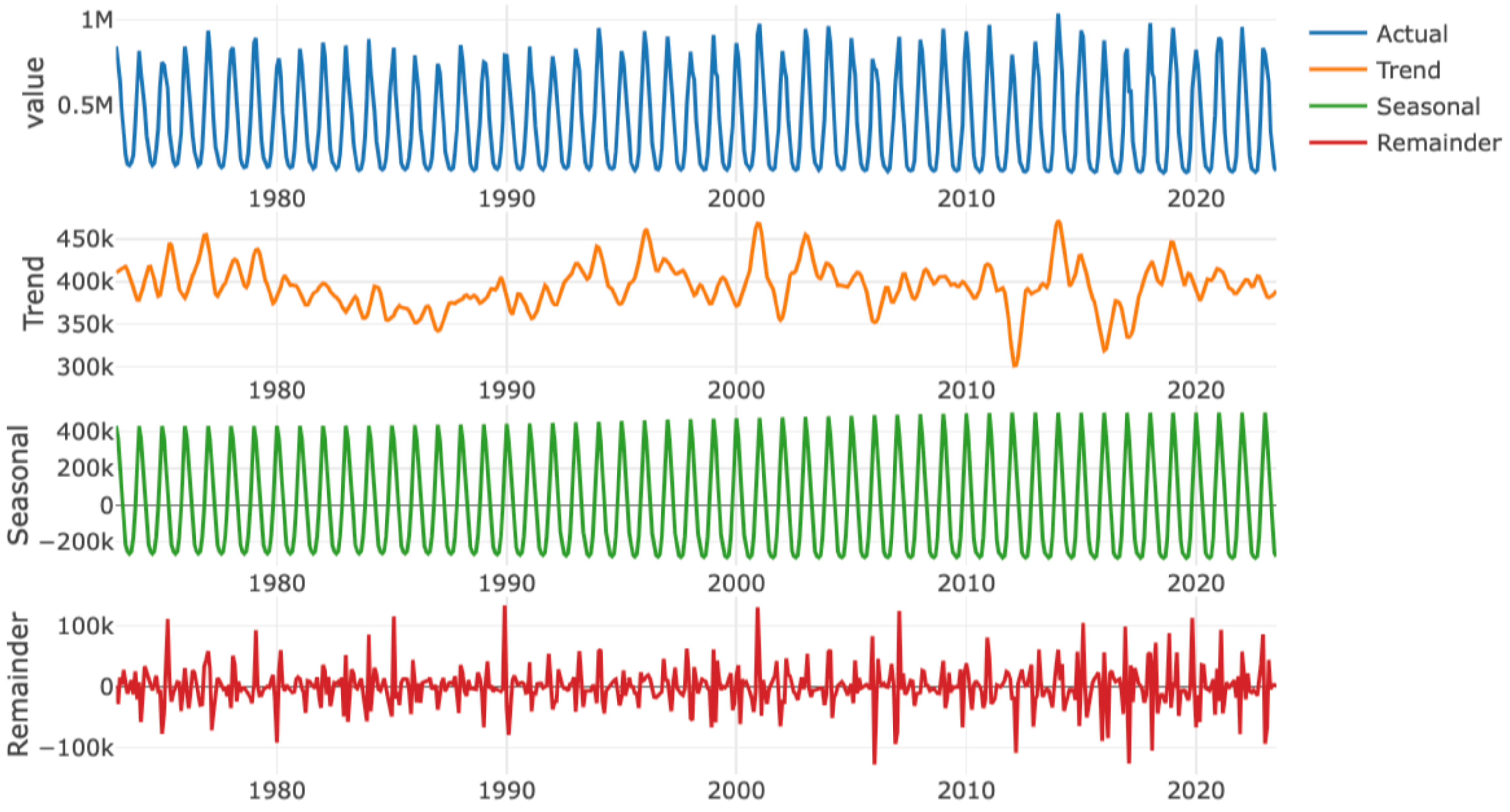
Transfer Learning

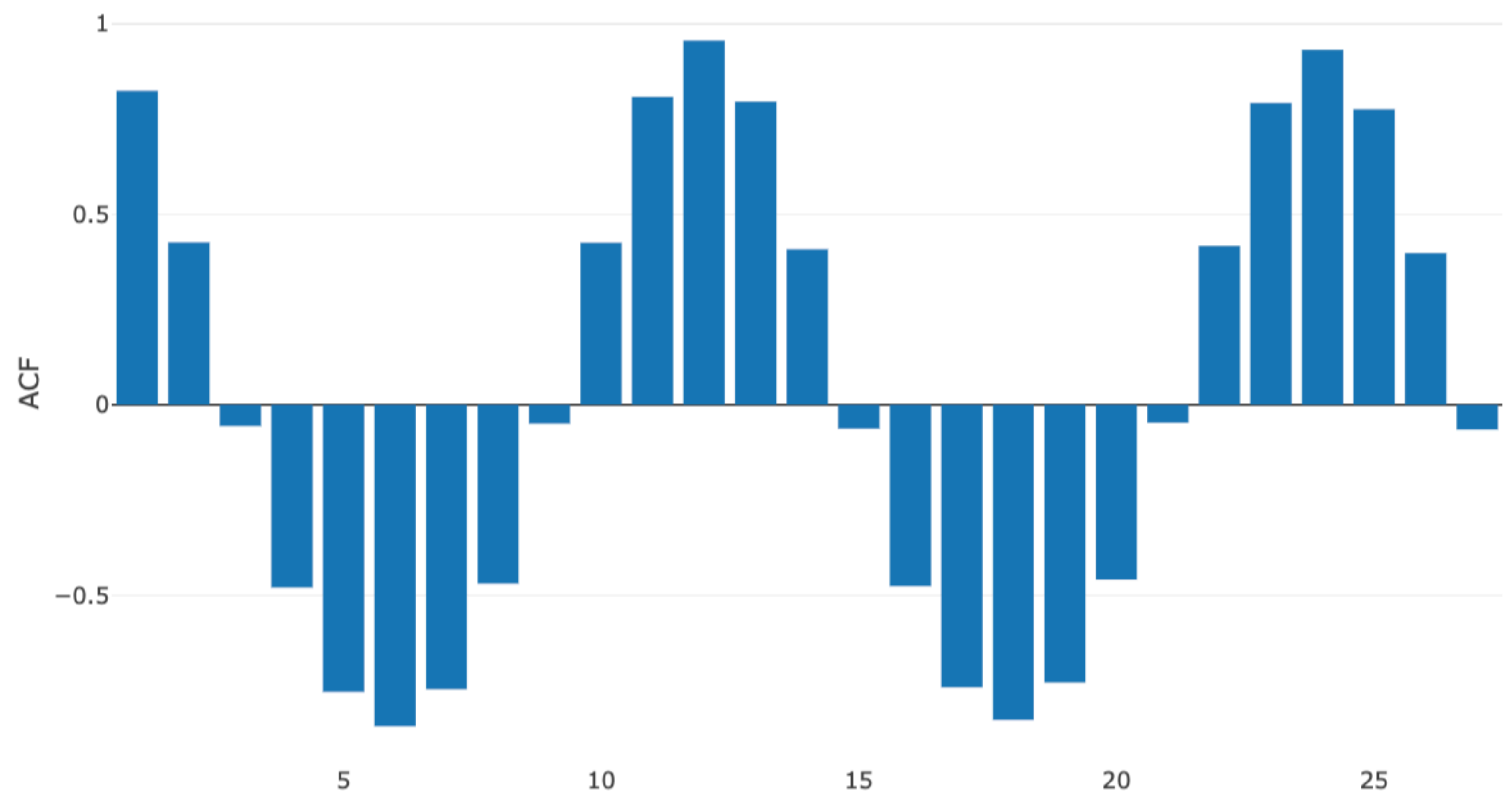
Monitoring

Forecast Single Series

Traditional Approach

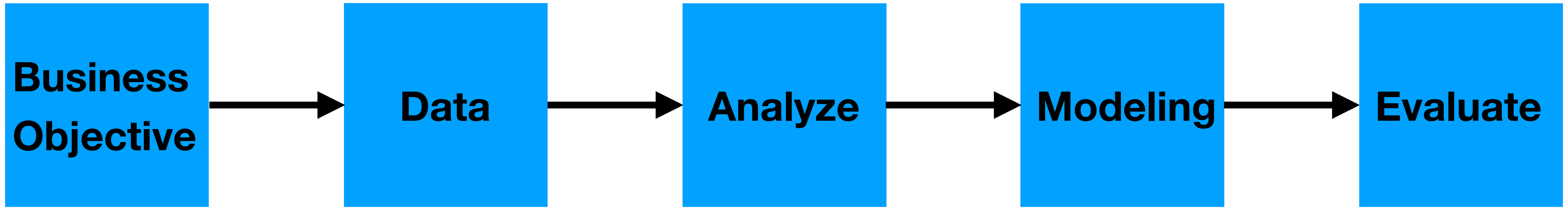


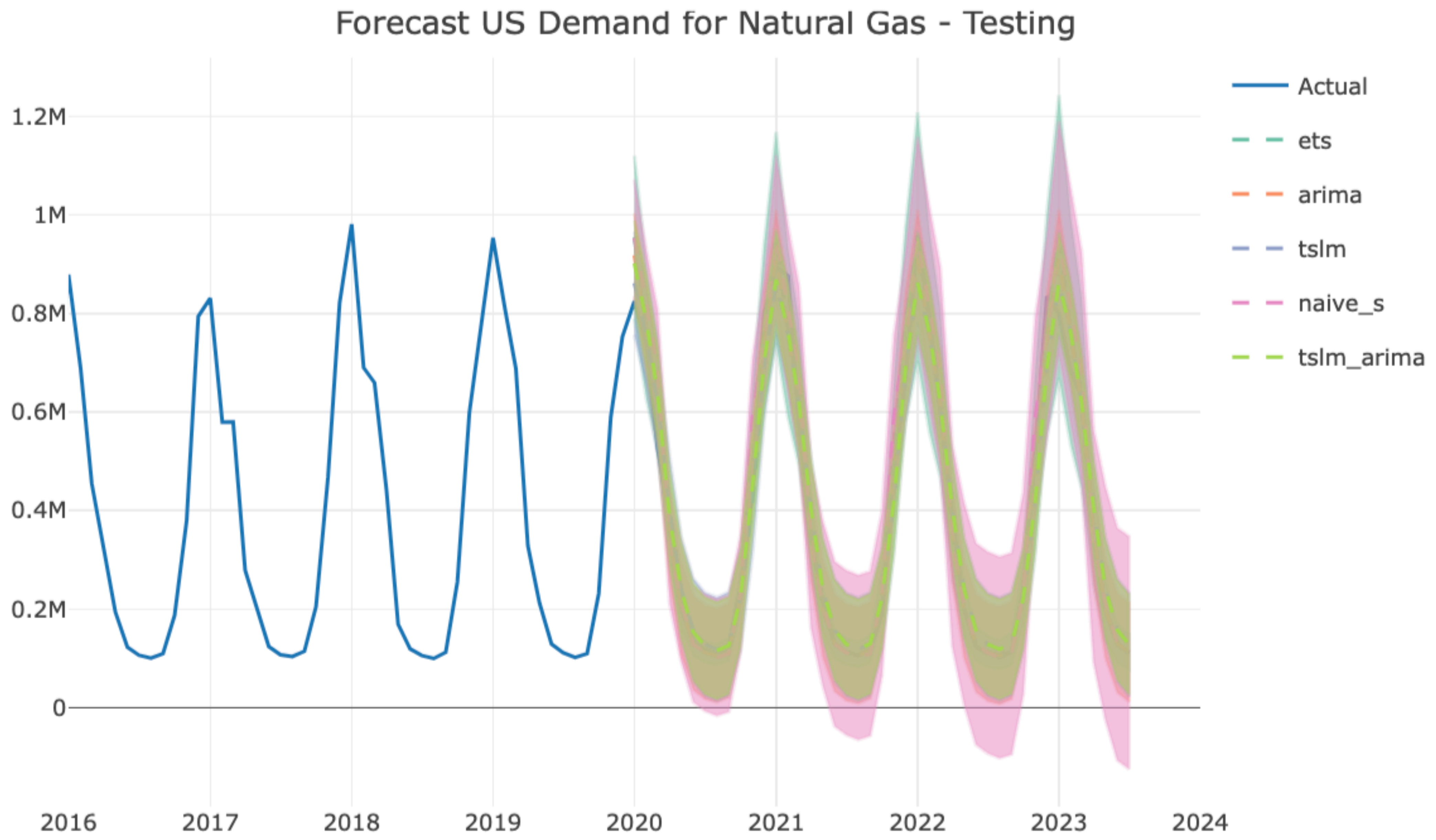




Forecast Single Series

Traditional Approach

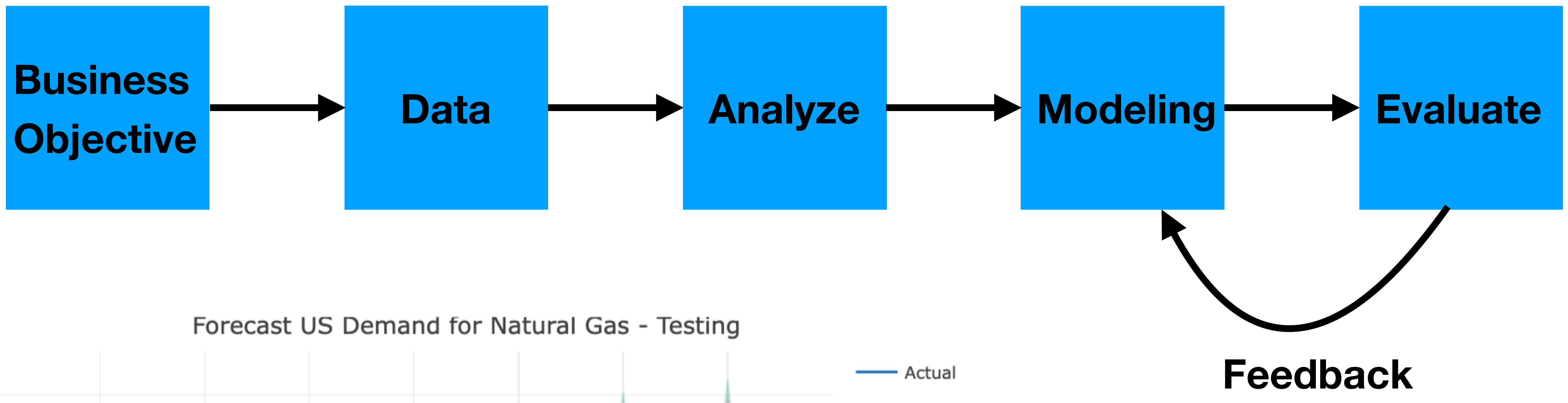


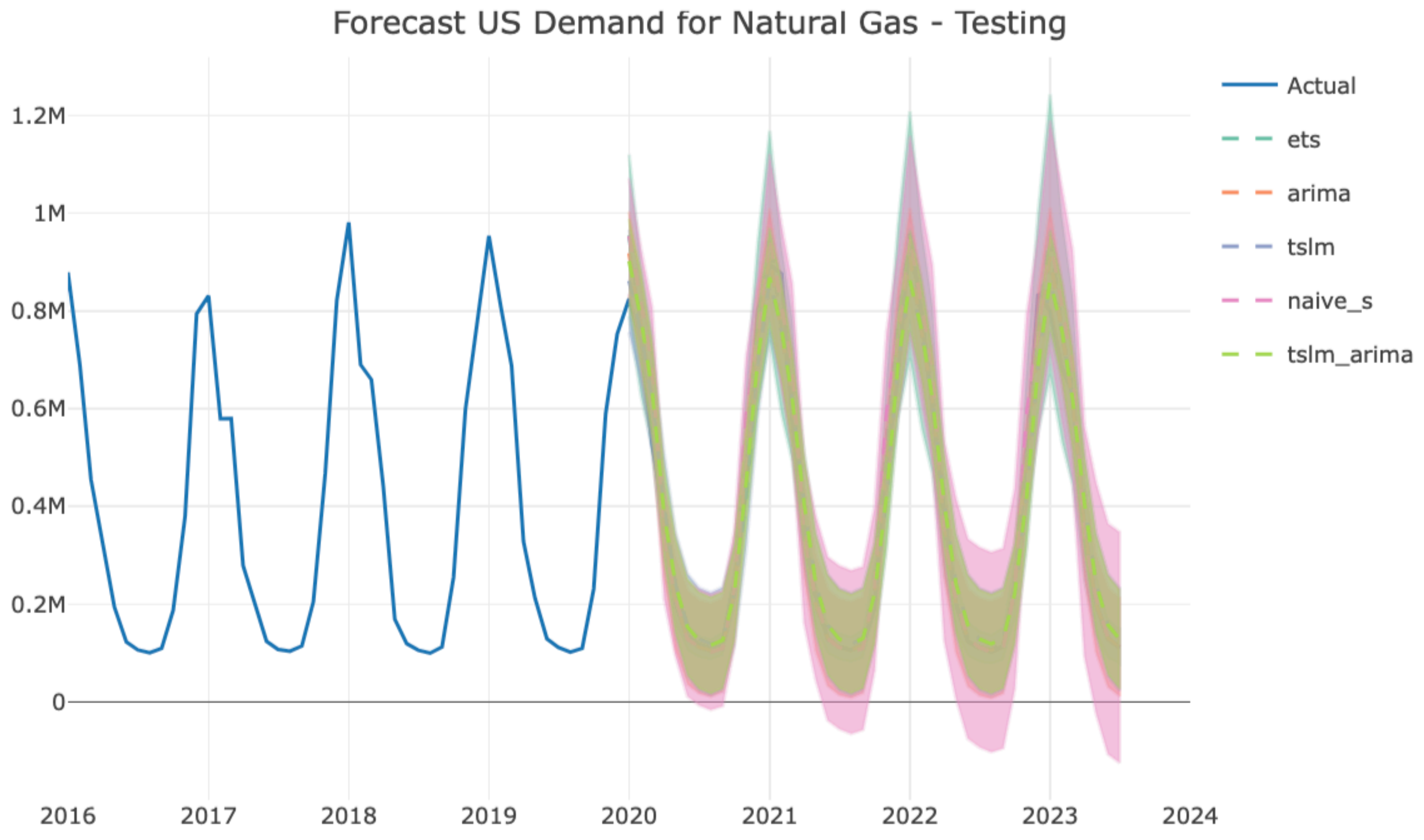



```
# A tibble: 5 × 3
  model        mape     rmse
  <chr>       <dbl>    <dbl>
1 arima      0.0612 3.76e18
2 ets        0.0681 7.42e18
3 naive_s    0.0928 1.88e19
4 tslm_arima 0.115  7.89e18
5 tslm       0.115  7.11e18
```

Forecast Single Series

Traditional Approach

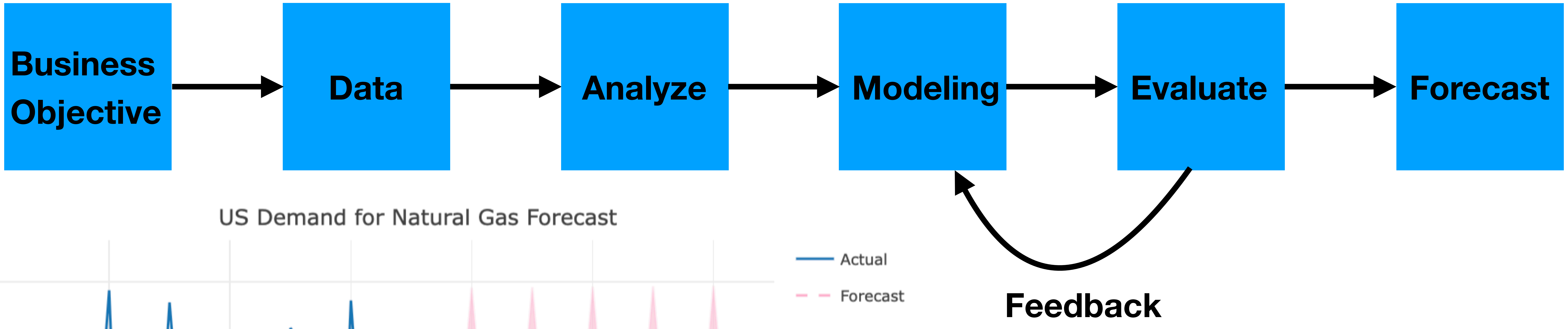


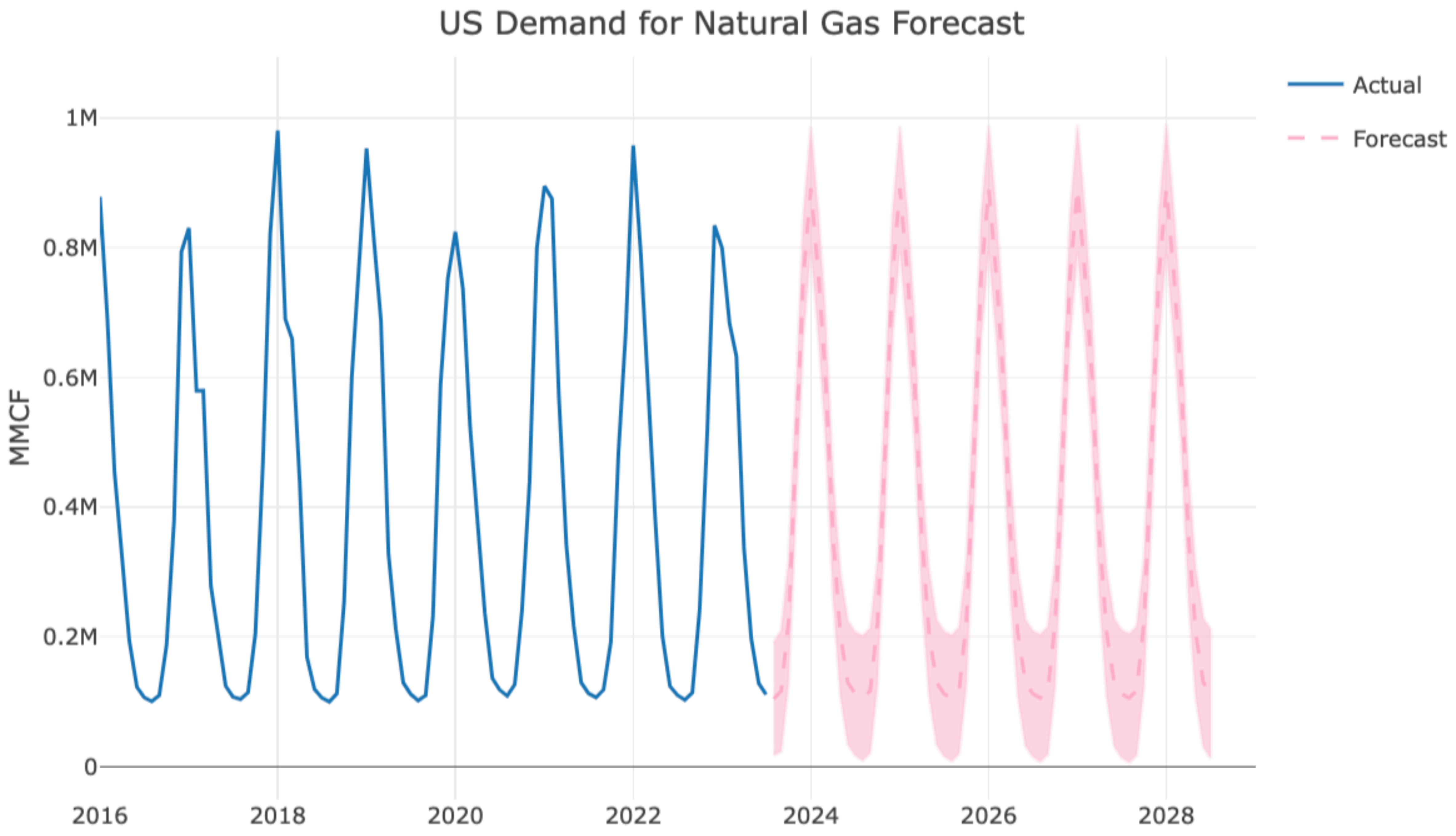



```
# A tibble: 5 × 3
  model           mape     rmse
  <chr>          <dbl>    <dbl>
1 arima         0.0612 3.76e18
2 ets           0.0681 7.42e18
3 naive_s       0.0928 1.88e19
4 tslm_arima    0.115  7.89e18
5 tslm          0.115  7.11e18
```

Forecast Single Series

Traditional Approach

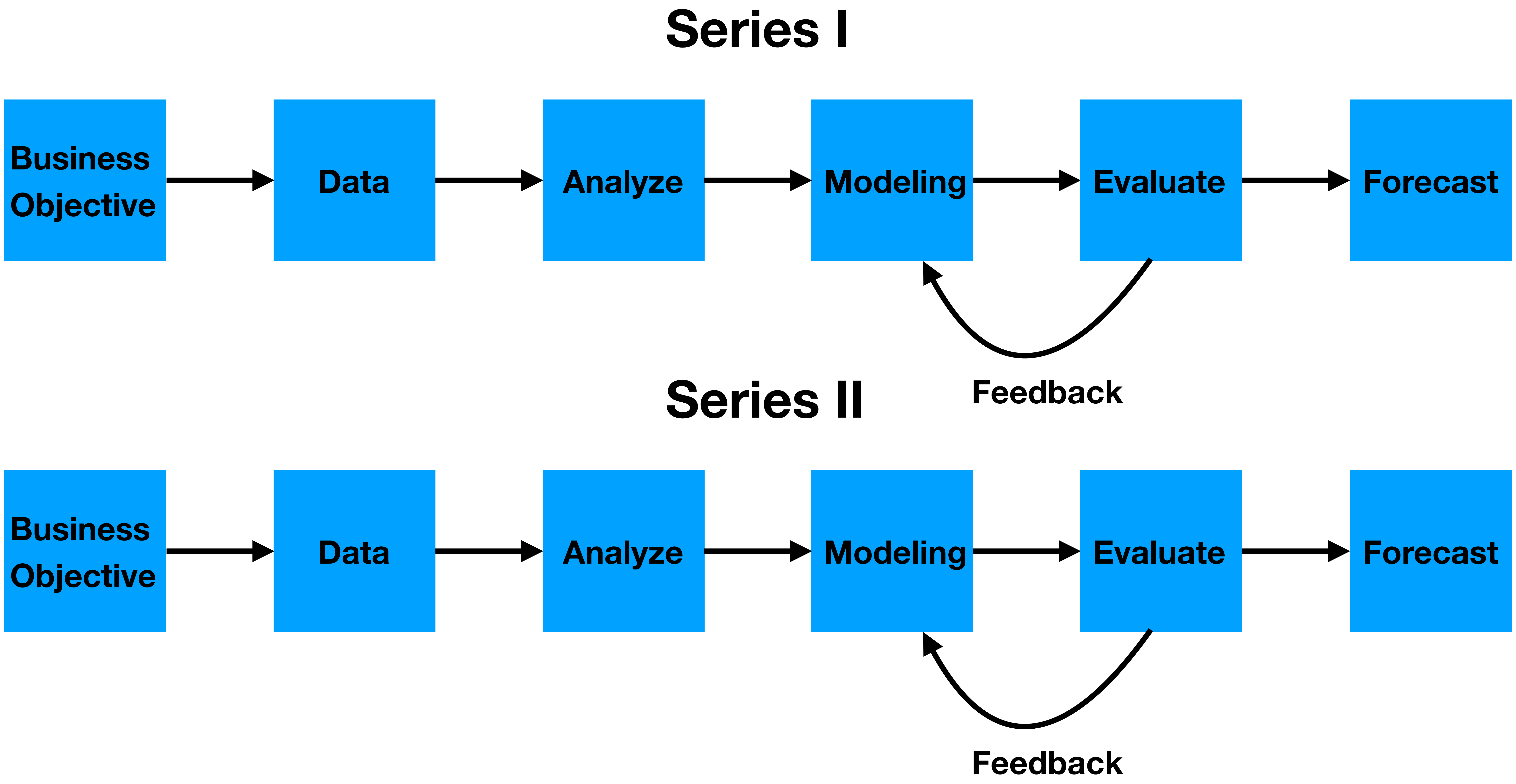




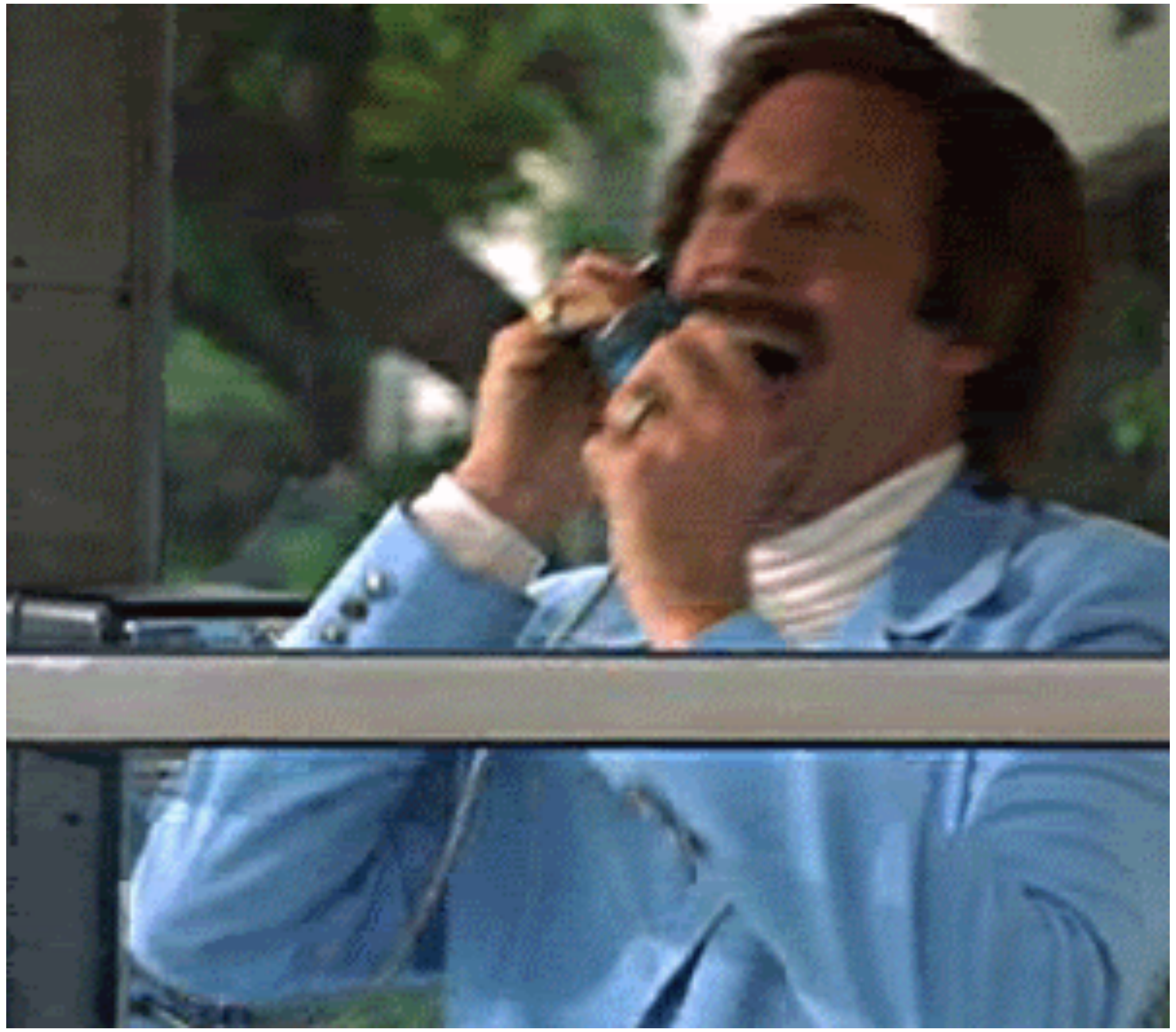
What if you have two series to forecast?

Forecast Single Series

Traditional Approach



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



Forecasting at Scale

Definition

- The level of effort of adding additional series is non linear or with marginal decay
- A function of forecasting approach
- Infrastructure dependency
- Come with the cost of potential lower accuracy

Forecasting at Scale

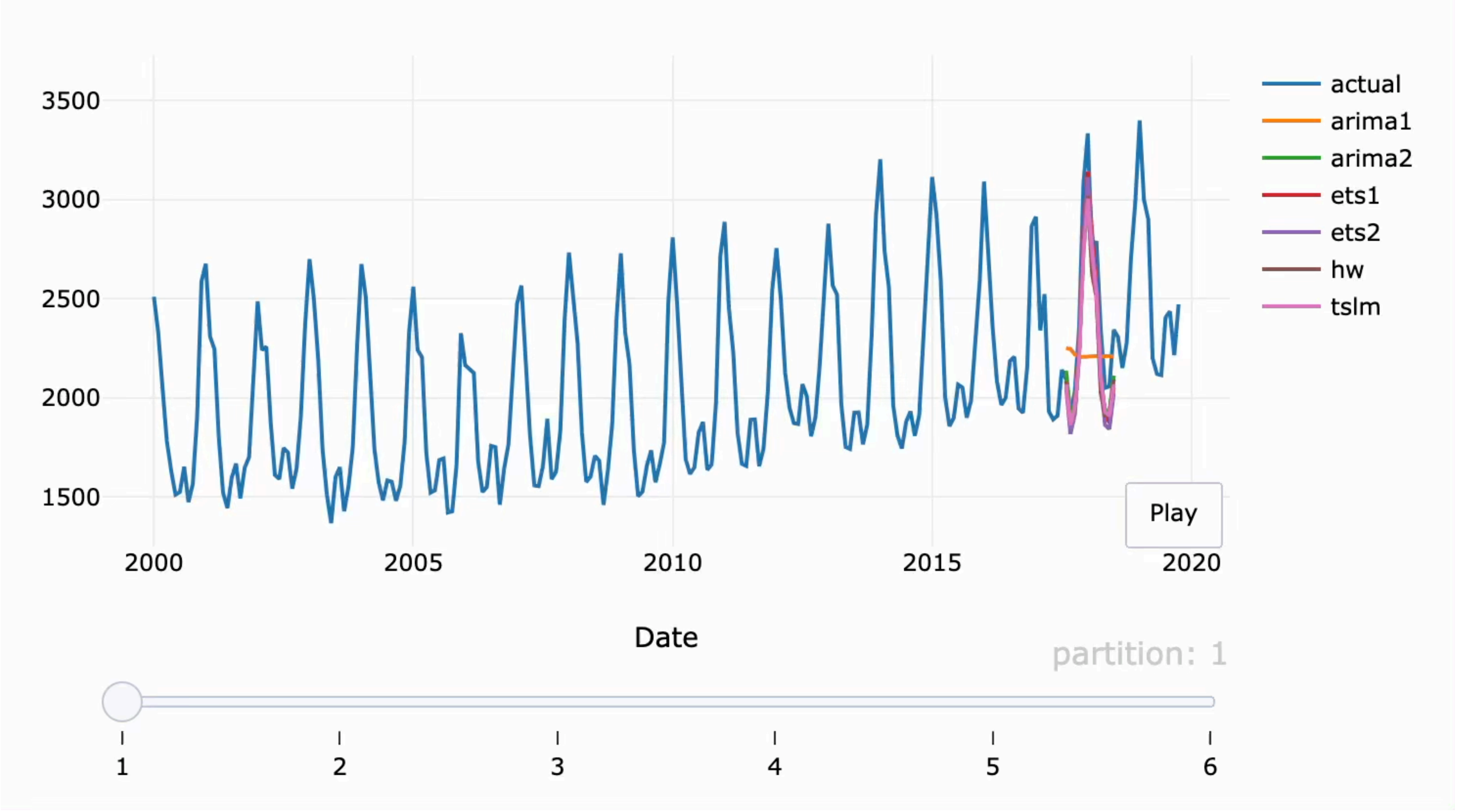
General Approaches

- Horse Racing
- Feature-based Time Series Analysis
- Transfer Learning
- Monitoring

Horse Racing

Approach

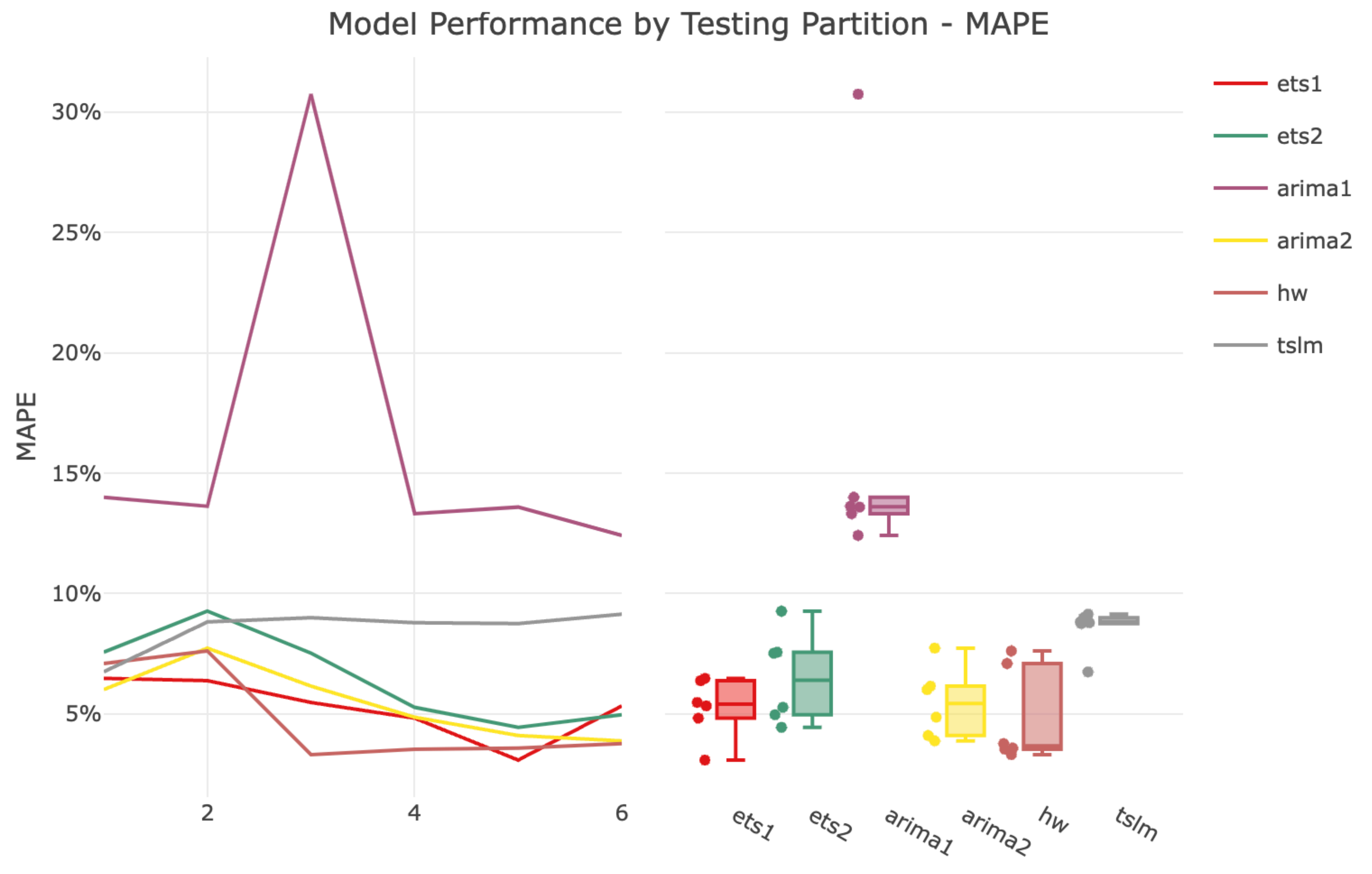
- Train multiple models



Horse Racing

Approach

- Train multiple models
- Evaluate their performance

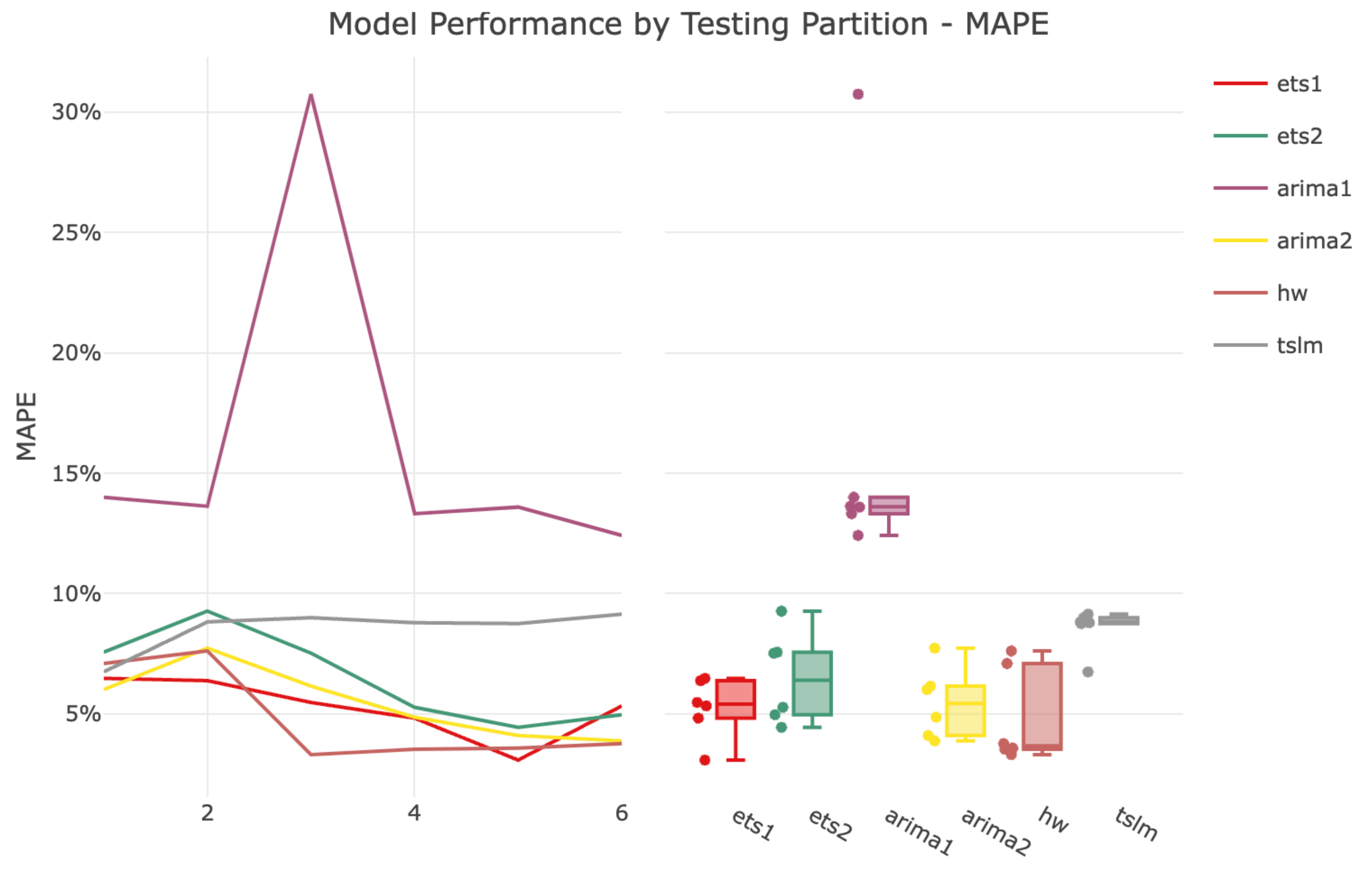



```
#> # A tibble: 5 x 7
#>   model_id model  notes      avg_mape avg_rmse `avg_coverage_9… `avg_coverage_9…
#>   <chr>    <chr>  <chr>         <dbl>    <dbl>            <dbl>            <dbl>
#> 1 hw       HoltW… HoltWinte…   0.0482     144.            0.875            0.931
#> 2 ets1     ets    ETS model…   0.0526     156.            0.917            0.972
#> 3 arima2   arima  SARIMA(2,…   0.0546     163.            0.736            0.819
#> 4 ets2     ets    ETS model…   0.0650     185.            0.722            0.792
#> 5 tslm     tslm   tslm mode…   0.0854     242.            0.431            0.611
```

Horse Racing

Approach

- Train multiple models
- Evaluate their performance
- Select the one that perform best on the testing partition




```
#> # A tibble: 5 x 7
#>   model_id model  notes      avg_mape avg_rmse `avg_coverage_9… `avg_coverage_9…
#>   <chr>    <chr>  <chr>         <dbl>    <dbl>            <dbl>            <dbl>
#> 1 hw       HoltW… HoltWinte…   0.0482     144.            0.875            0.931
#> 2 ets1     ets    ETS model…   0.0526     156.            0.917            0.972
#> 3 arima2   arima  SARIMA(2,…   0.0546     163.            0.736            0.819
#> 4 ets2     ets    ETS model…   0.0650     185.            0.722            0.792
#> 5 tslm     tslm   tslm mode…   0.0854     242.            0.431            0.611
```

Horse Racing

Pitfalls

- Similar to AutoML
  - Work well on structural data
  - Less with unstructured data
- Using backtesting might fail miserably if unexpected changes occur during the last testing period (e.g., COVID-19 impact, etc.)

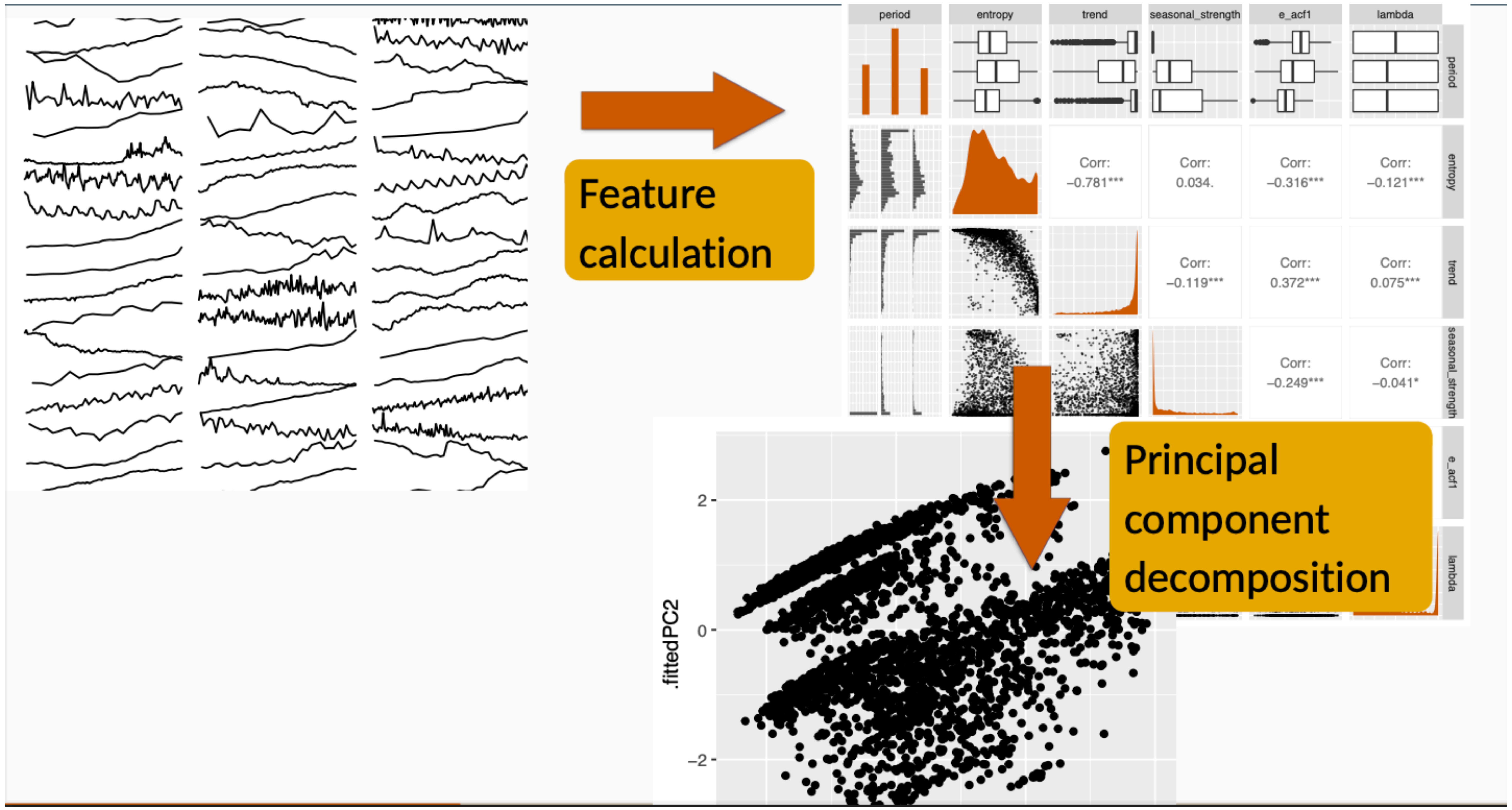
Feature-based Time Series Analysis

Approach

- Collapse each time series to row of features (correlation, seasonality, trend, distribution, etc.)
- Conduct unsupervised learning method (e.g., PCA, etc.)
- Identify clusters and extract insights

Feature-based Time Series Analysis

Example

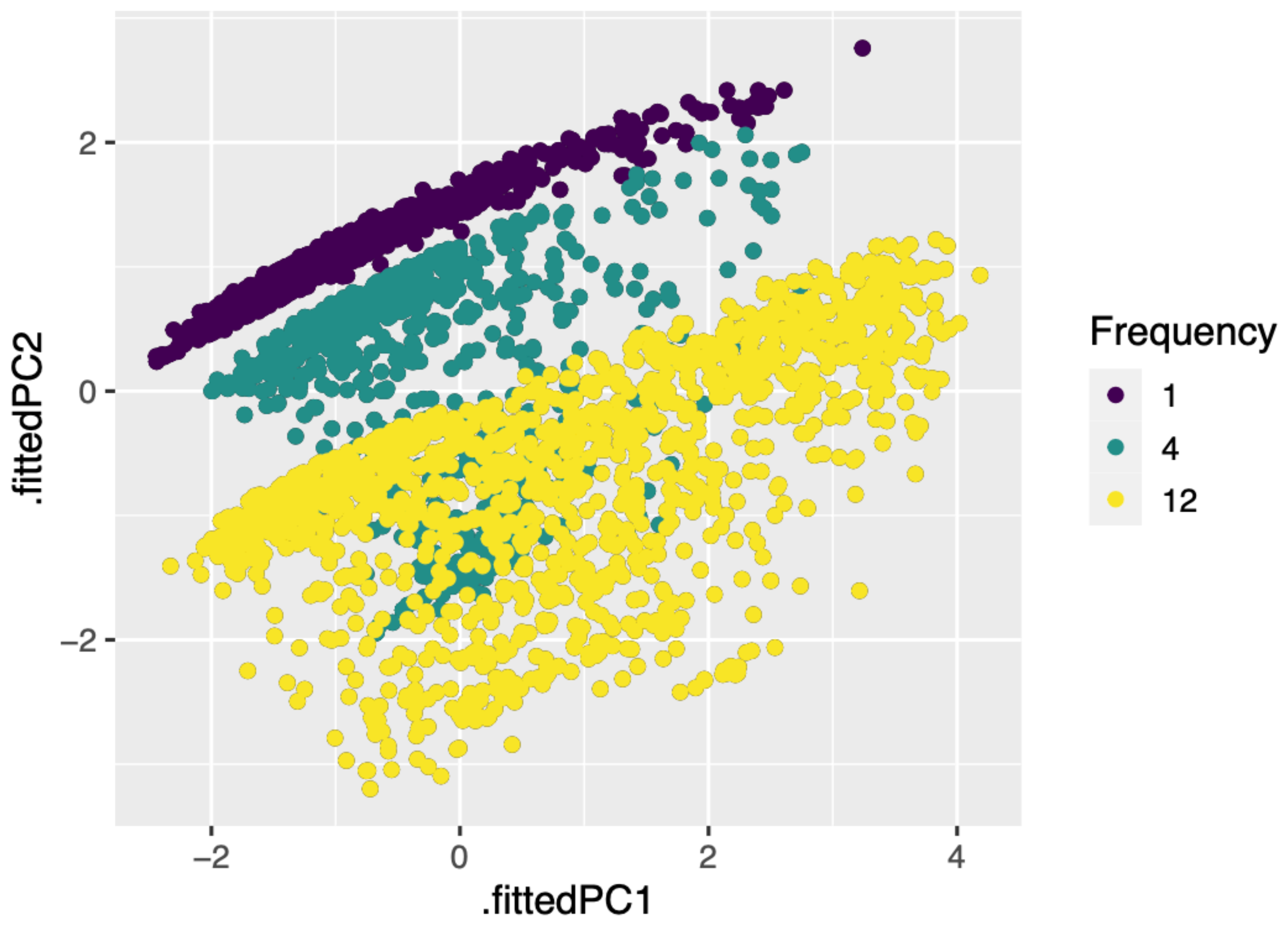


Source: Rob J Hyndman, Feature-based time series analysis

https://robjhyndman.com/seminars/fbtsa-ssc/

Feature-based Time Series Analysis

Example

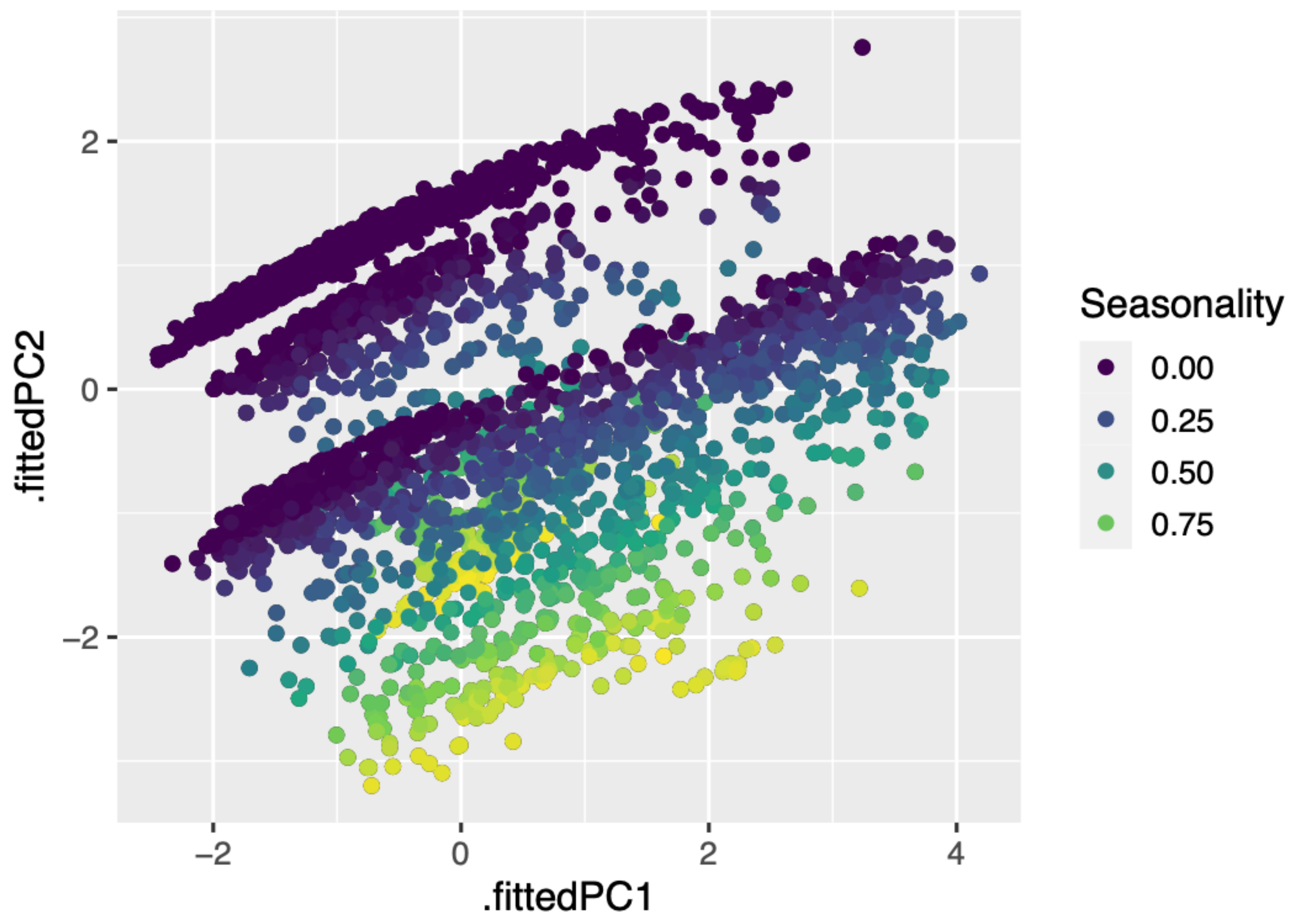


Source: Rob J Hyndman, Feature-based time series analysis

https://robjhyndman.com/seminars/fbtsa-ssc/

Feature-based Time Series Analysis

Example



Source: Rob J Hyndman, Feature-based time series analysis

https://robjhyndman.com/seminars/fbtsa-ssc/

Feature-based Time Series Analysis

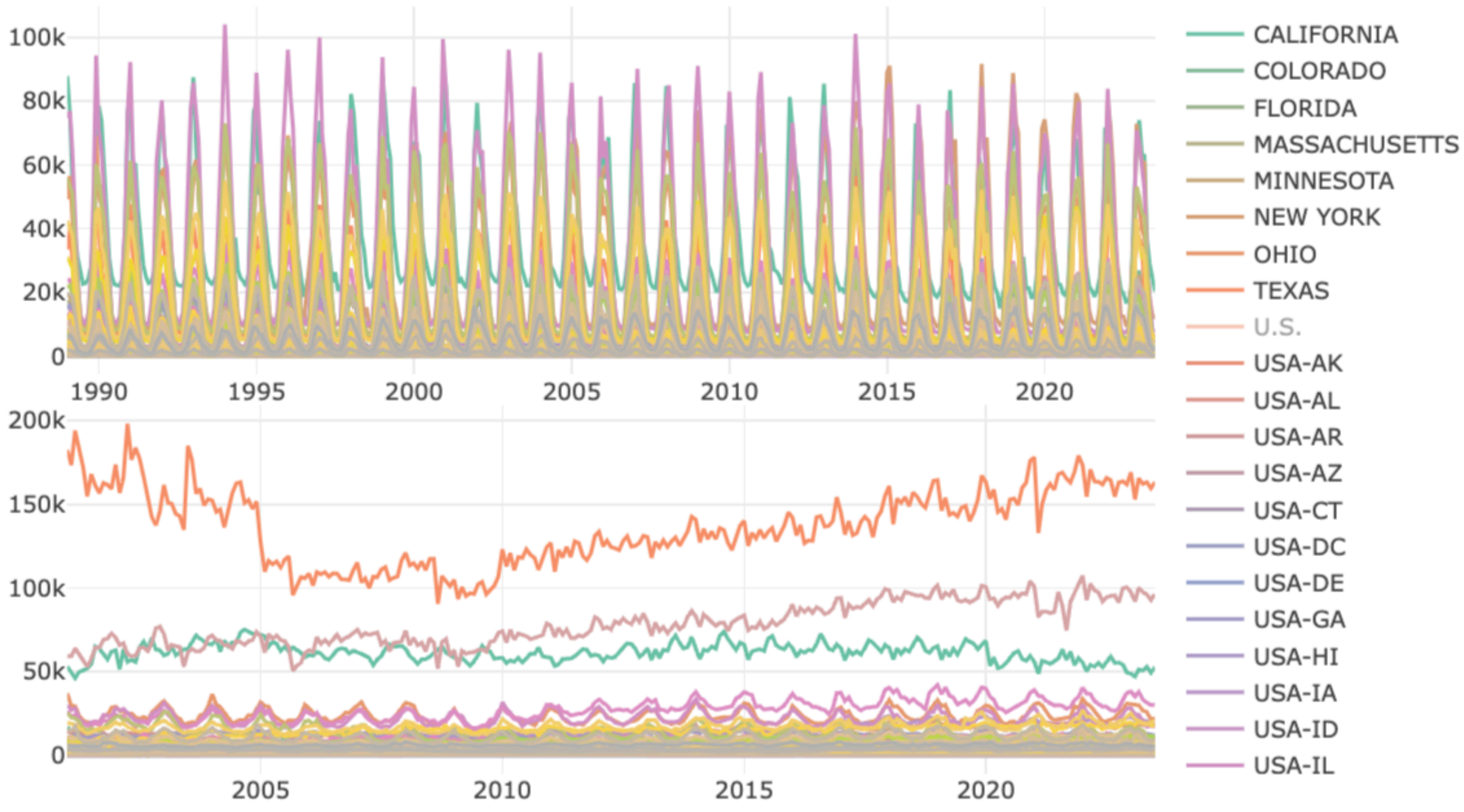
Pitfalls

- Required to identify features
- Unsupervised

Transfer Learning

Approach

- Identify cluster
- Select representative series
- Analyze, identify features
- Apply with on the rest of the cluster



Transfer Learning

Pitfalls

- Identify the series that represent best the cluster
- Most likely won’t a 100% representation of the cluster

“Statisticians are sad people. They know from the first place they are wrong, so they go and measure it…”

My Stats Professor

Monitoring

Performance Improvement Mitigating Risks

- Monitor the forecast performance
- Identify outliers and performance drift
- Post mortem analysis
- Apply back feedback

Questions?

Thank You!