

Forecasting at Scale

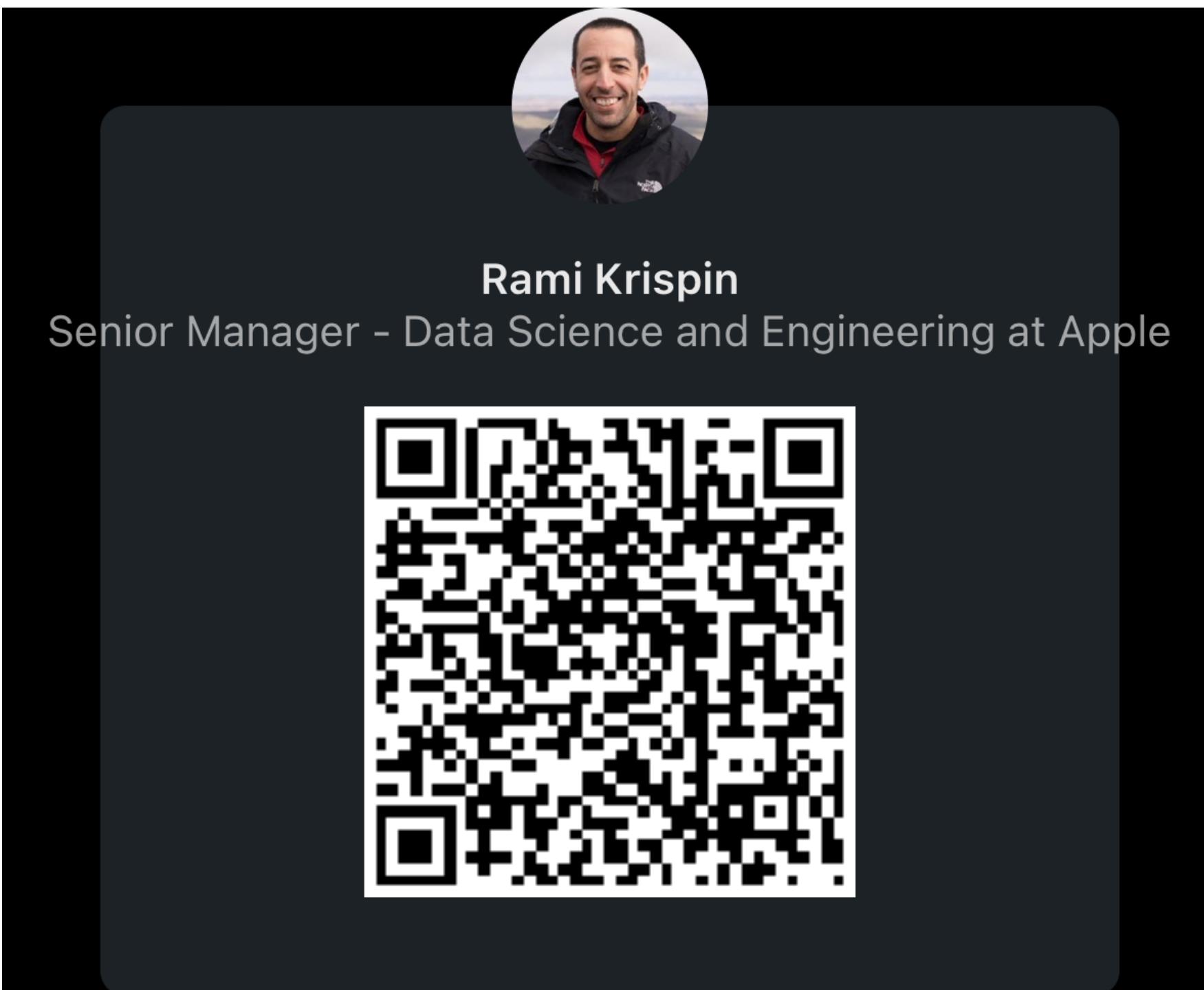
Øredev 2023

Rami Krispin, Nov 9, 2023

About Me

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

Get in Touch



Agenda

The Forecasting Process

General Architecture

Backtesting

Feature Engineering with Cluster Analysis

Definitions

Forecasting

- **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** - prediction of observations with time dimension

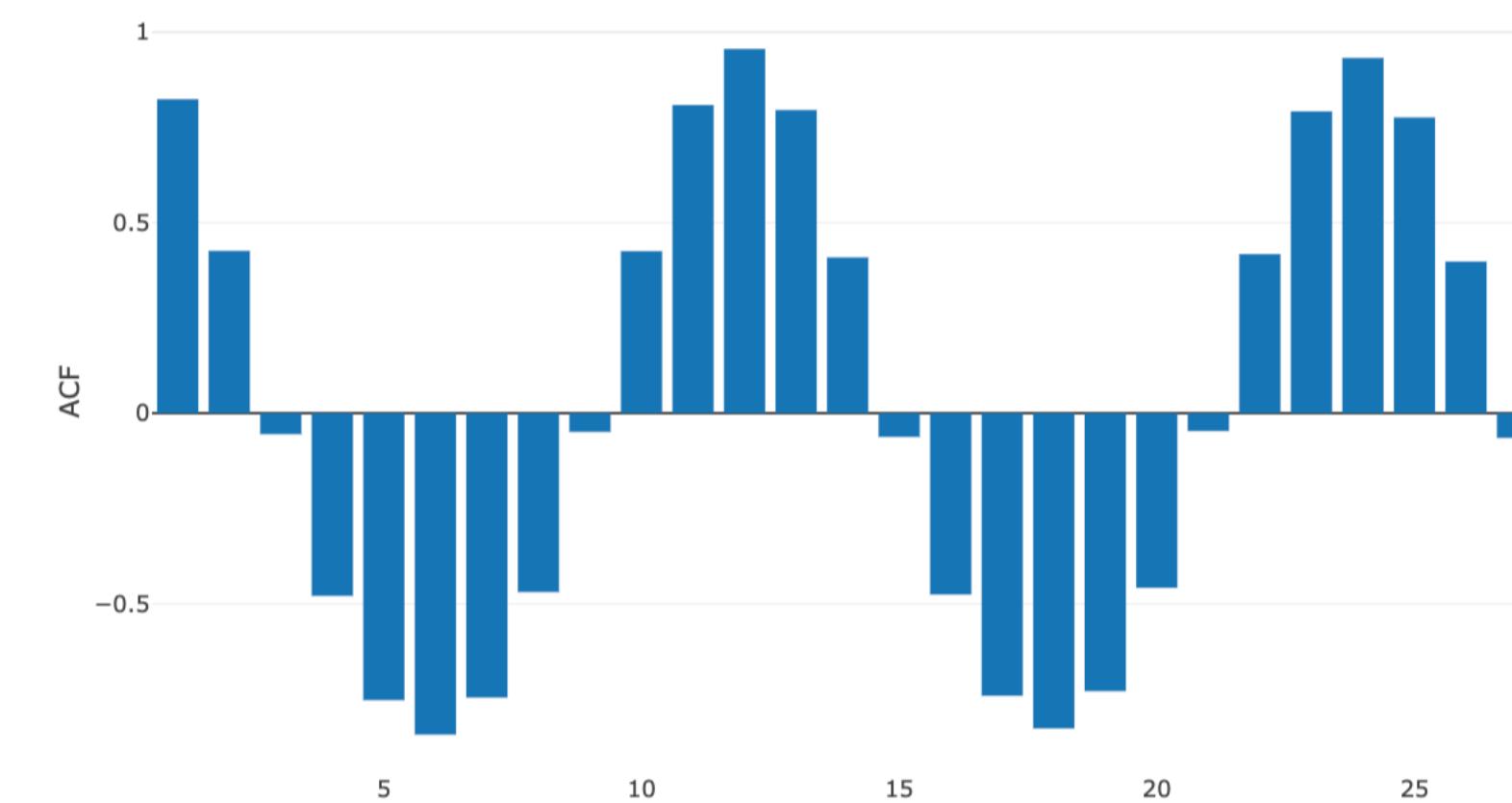
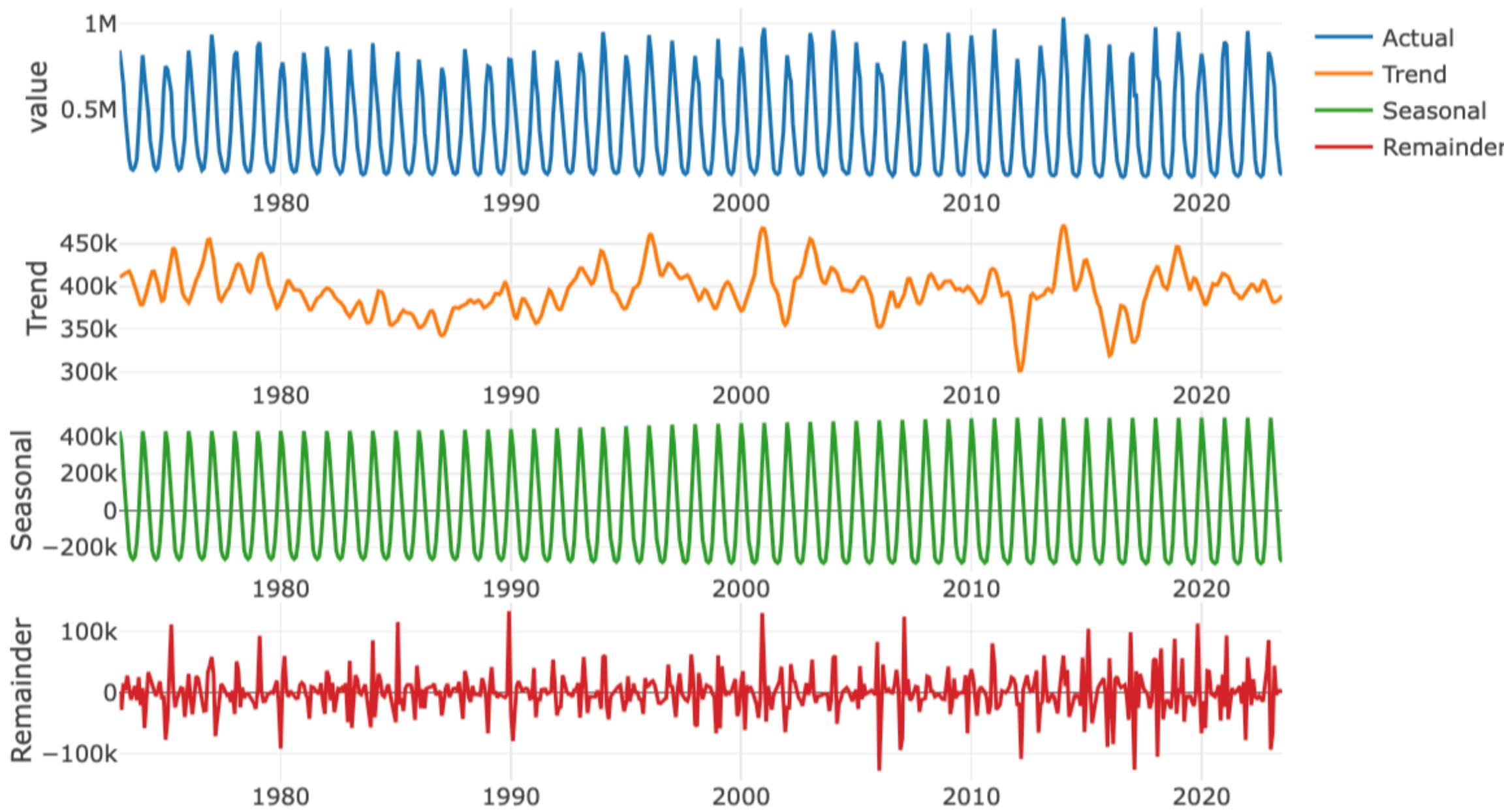
Definitions

Forecast

- **Time series** - a vector of data point 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** - prediction of observations with time dimension

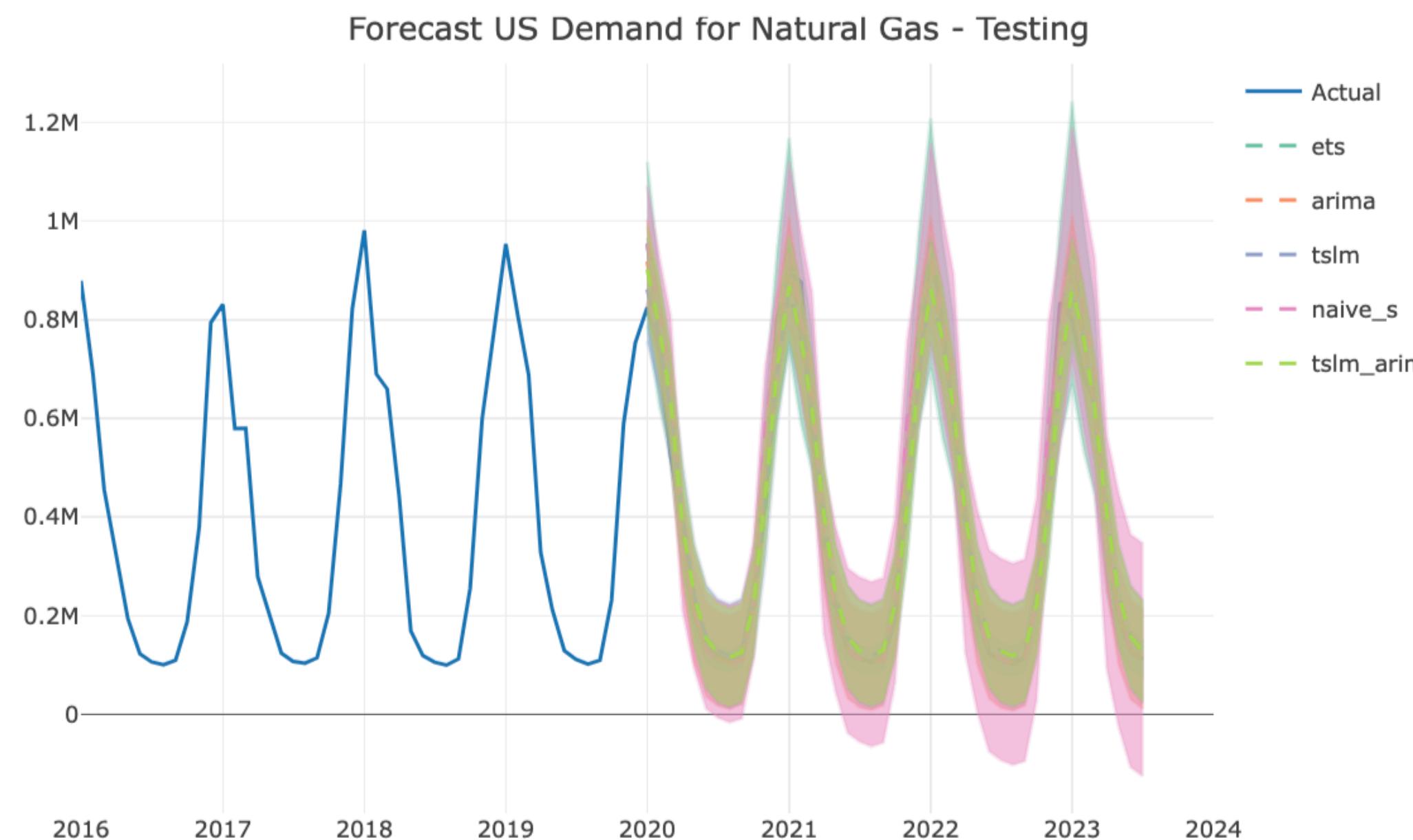
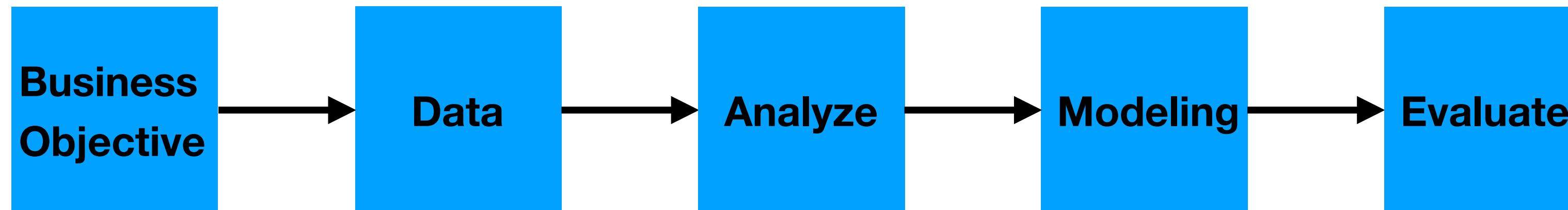
Forecast Single Series

Traditional Approach



Forecast Single Series

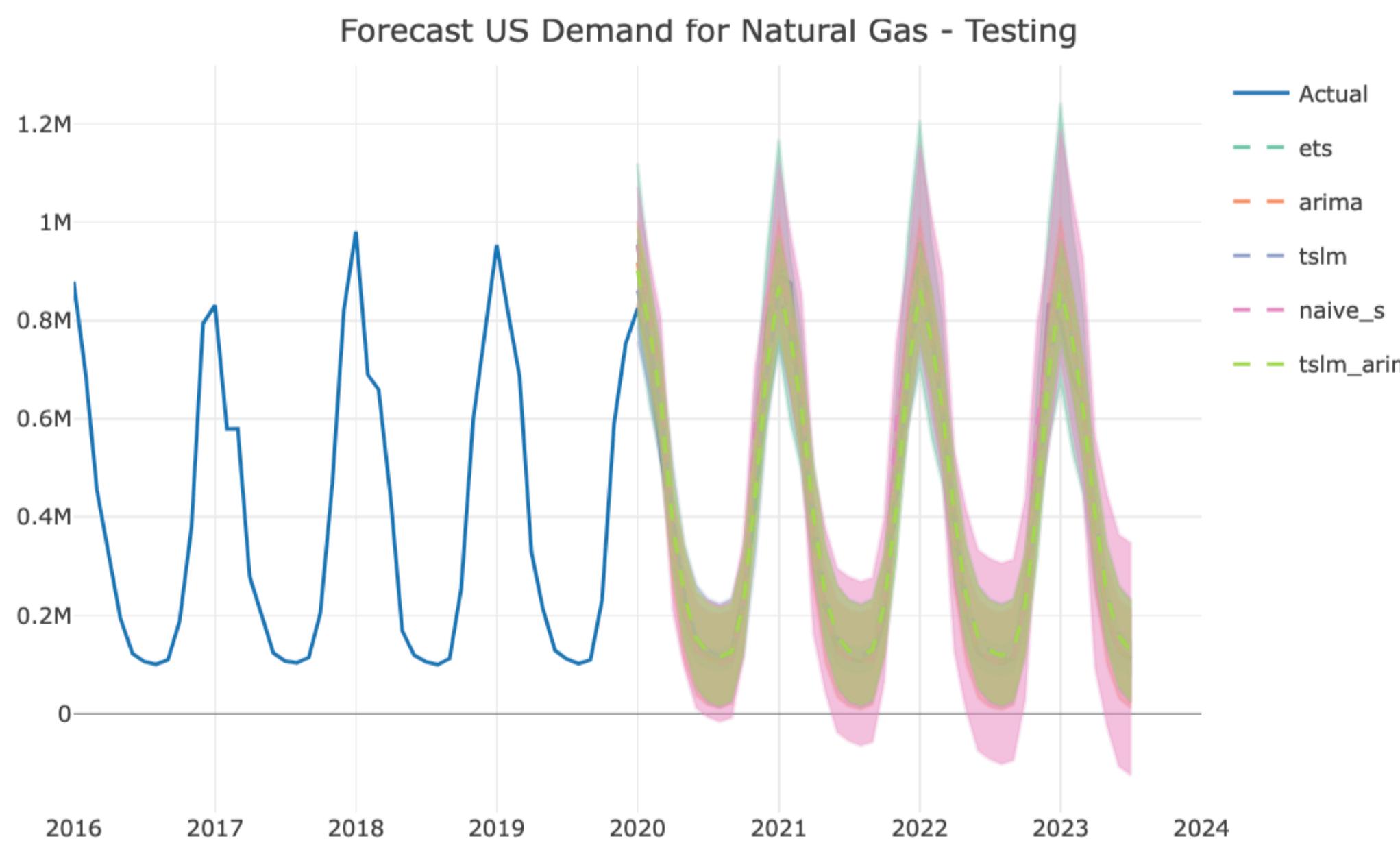
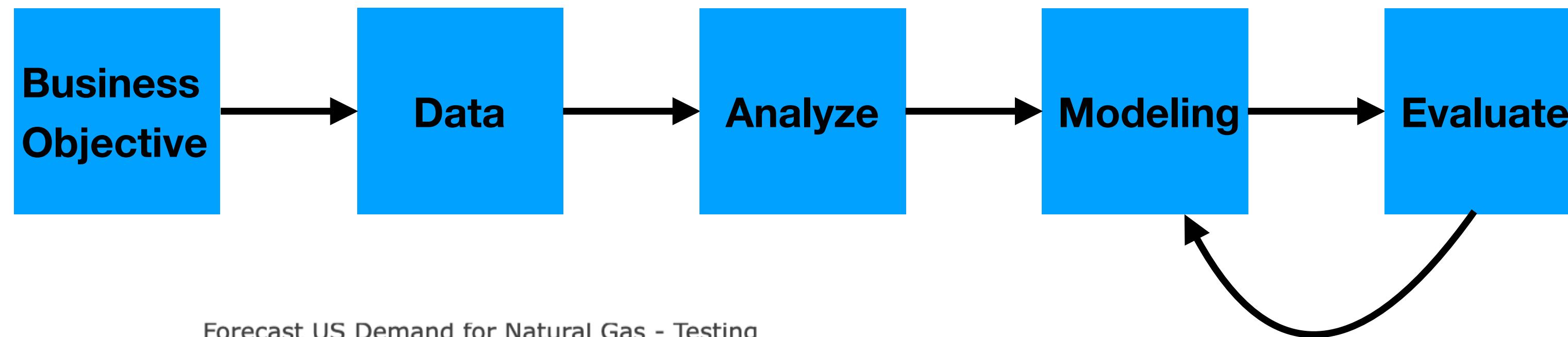
Traditional Approach



```
# 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

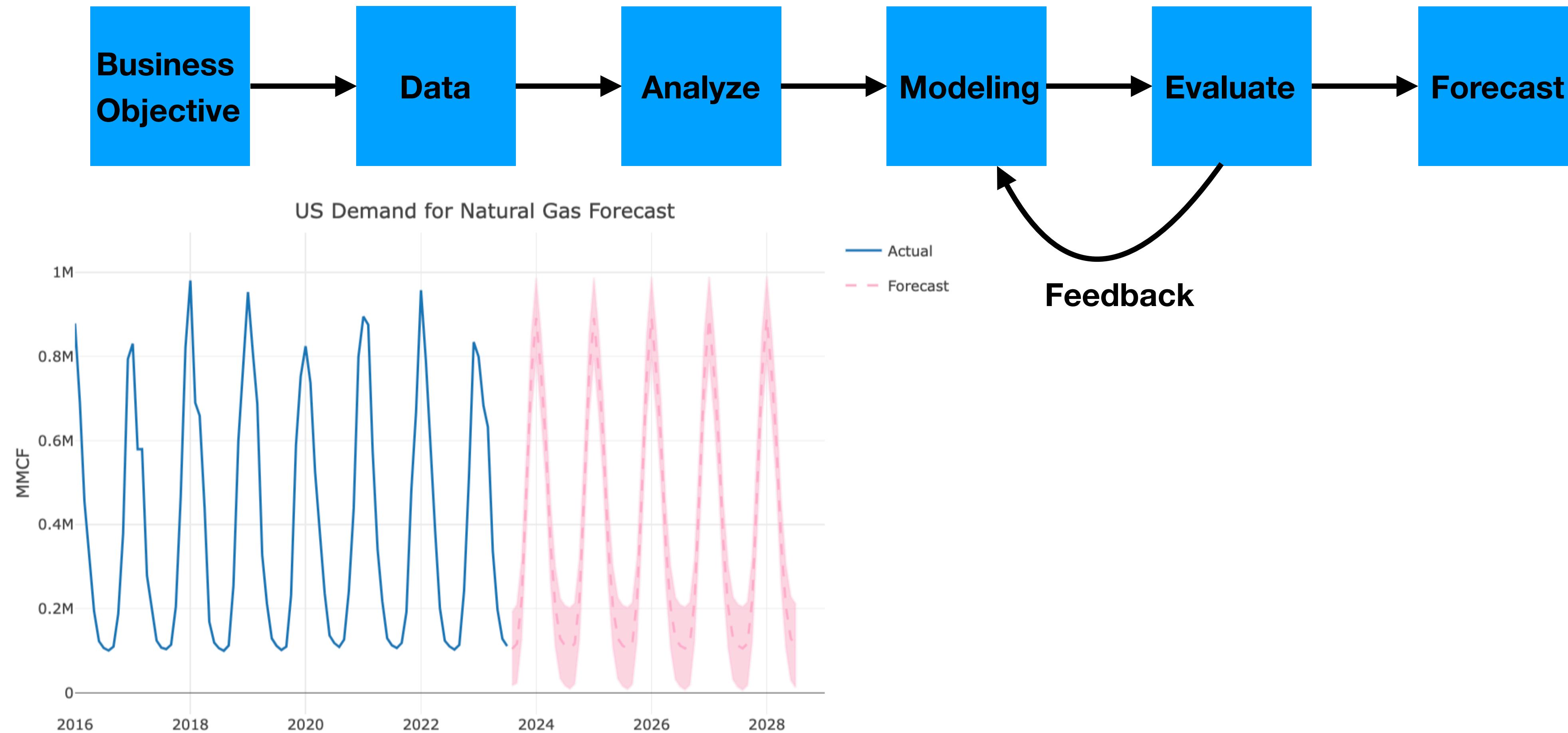
Traditional Approach



# 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

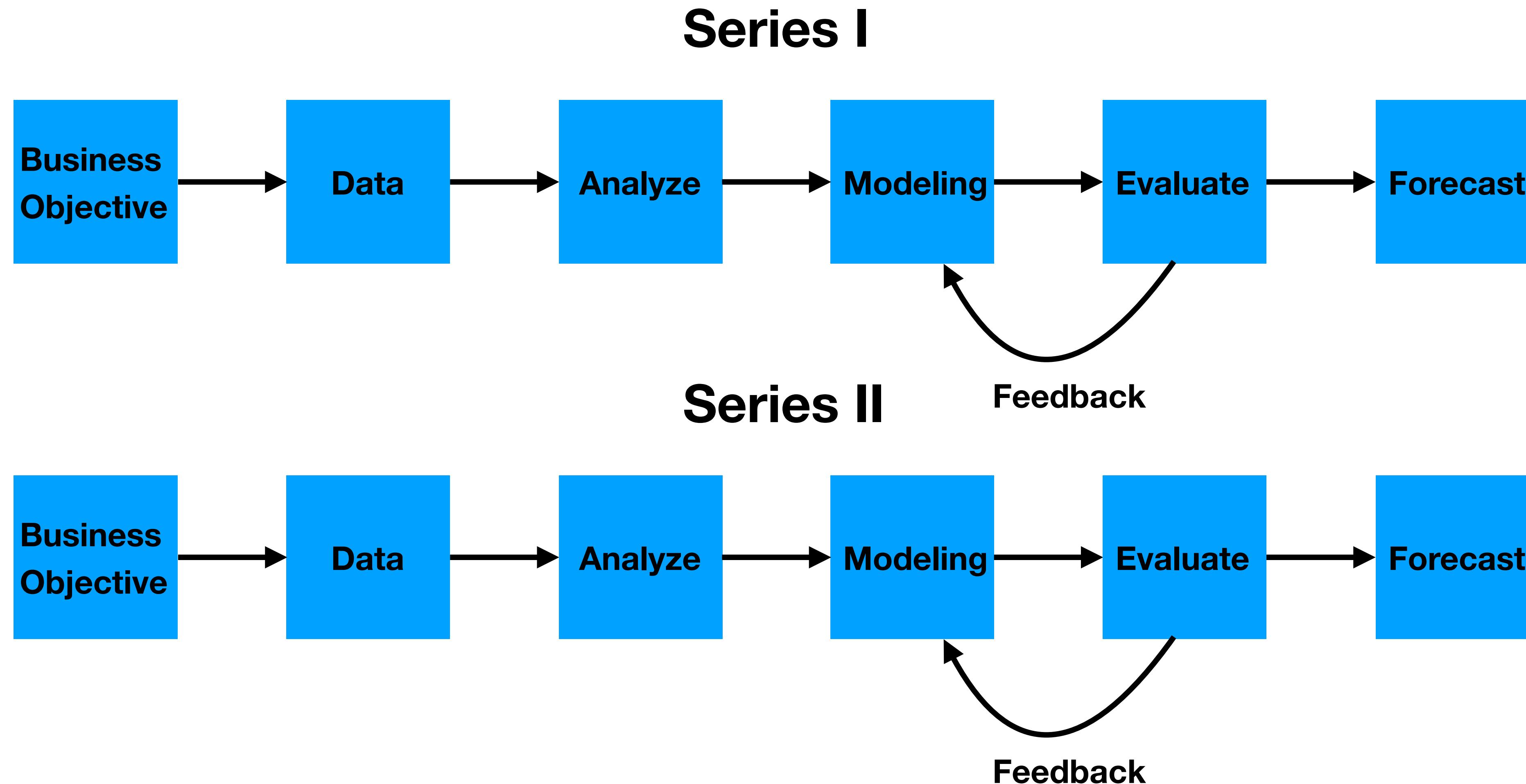
Traditional Approach



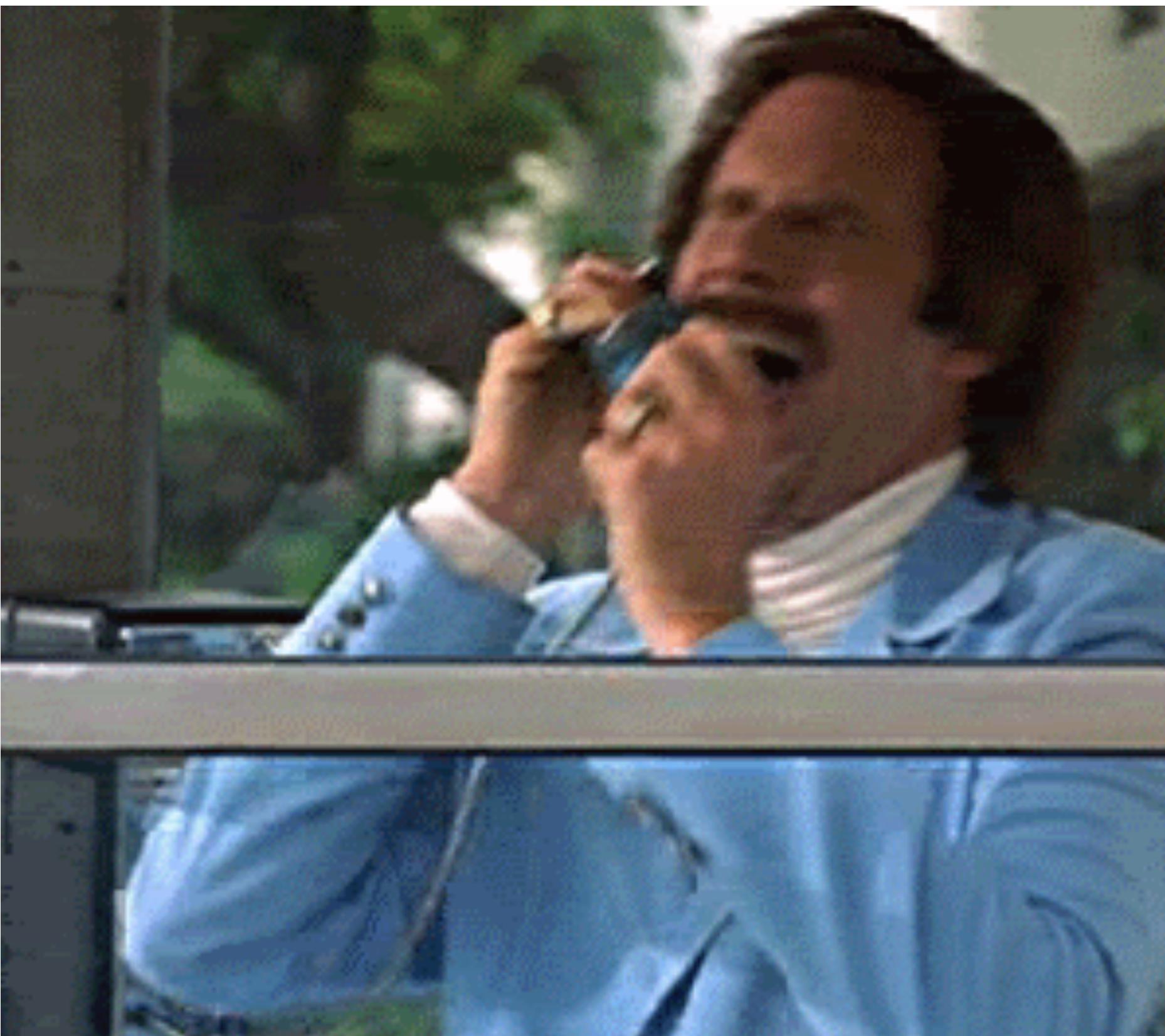
**What if you have two series to
forecast?**

Forecast Single Series

Traditional Approach



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



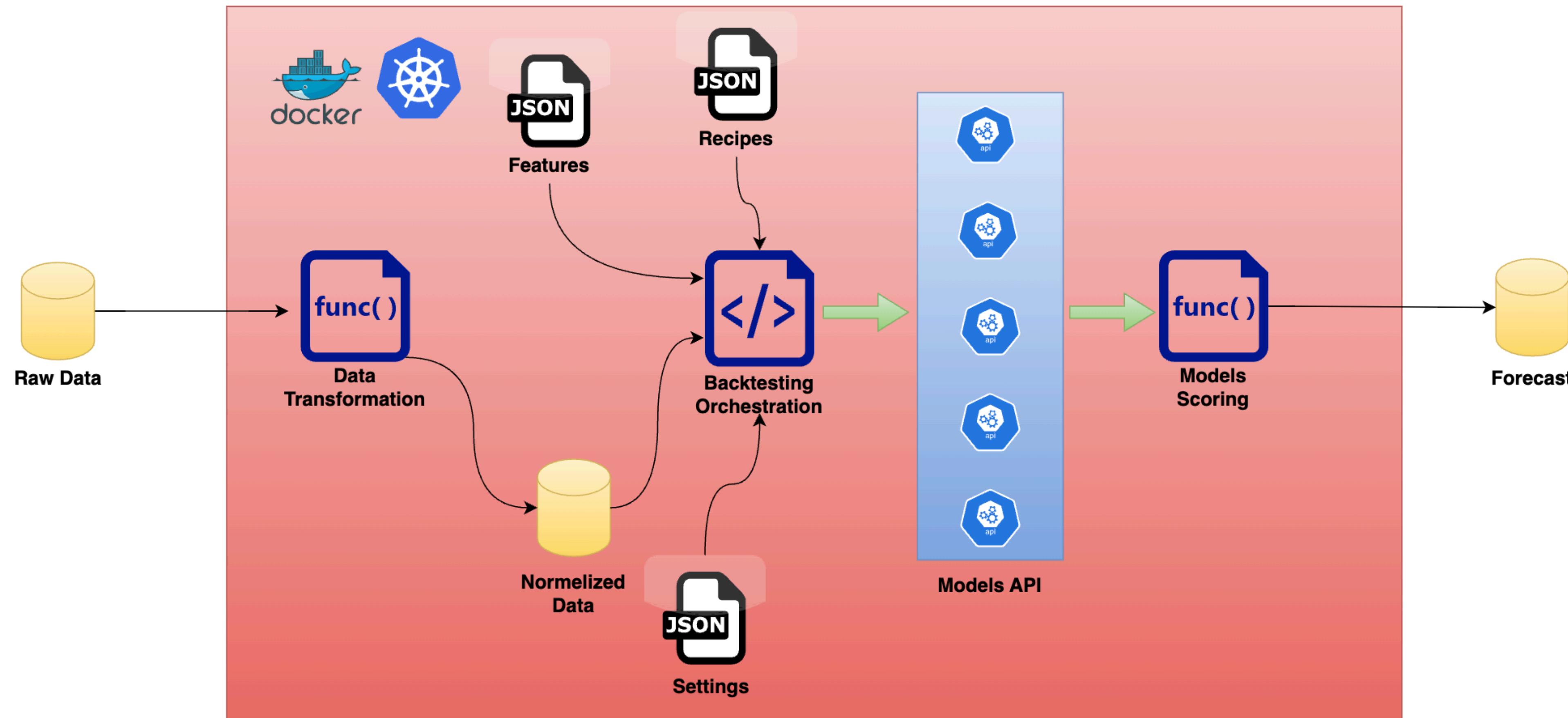
Forecasting at Scale

Forecasting at Scale

Definition

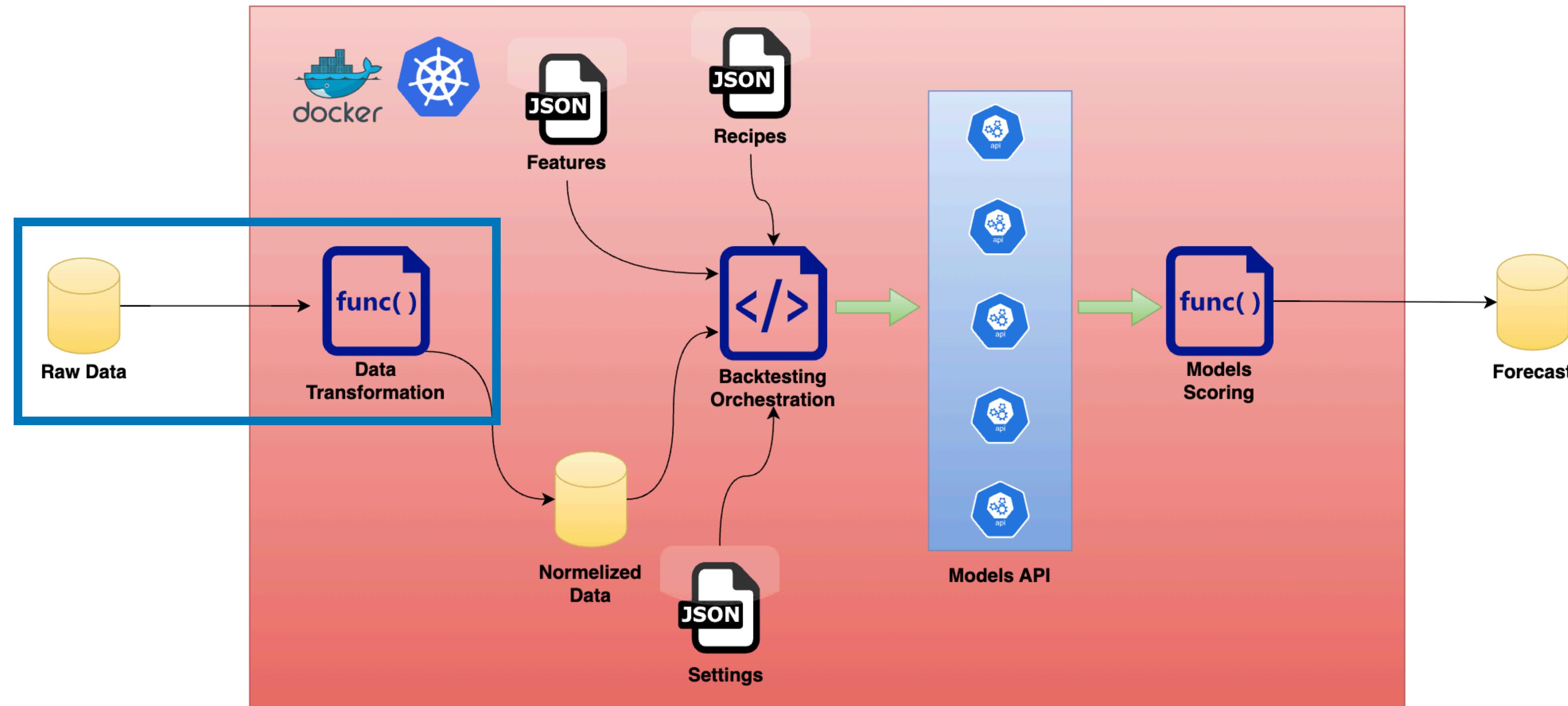
- The level of effort of adding additional series is non linear with marginal decay
- Modeling approaches
- Infrastructure dependency
- Trade-off - potential drop in accuracy

General Architecture



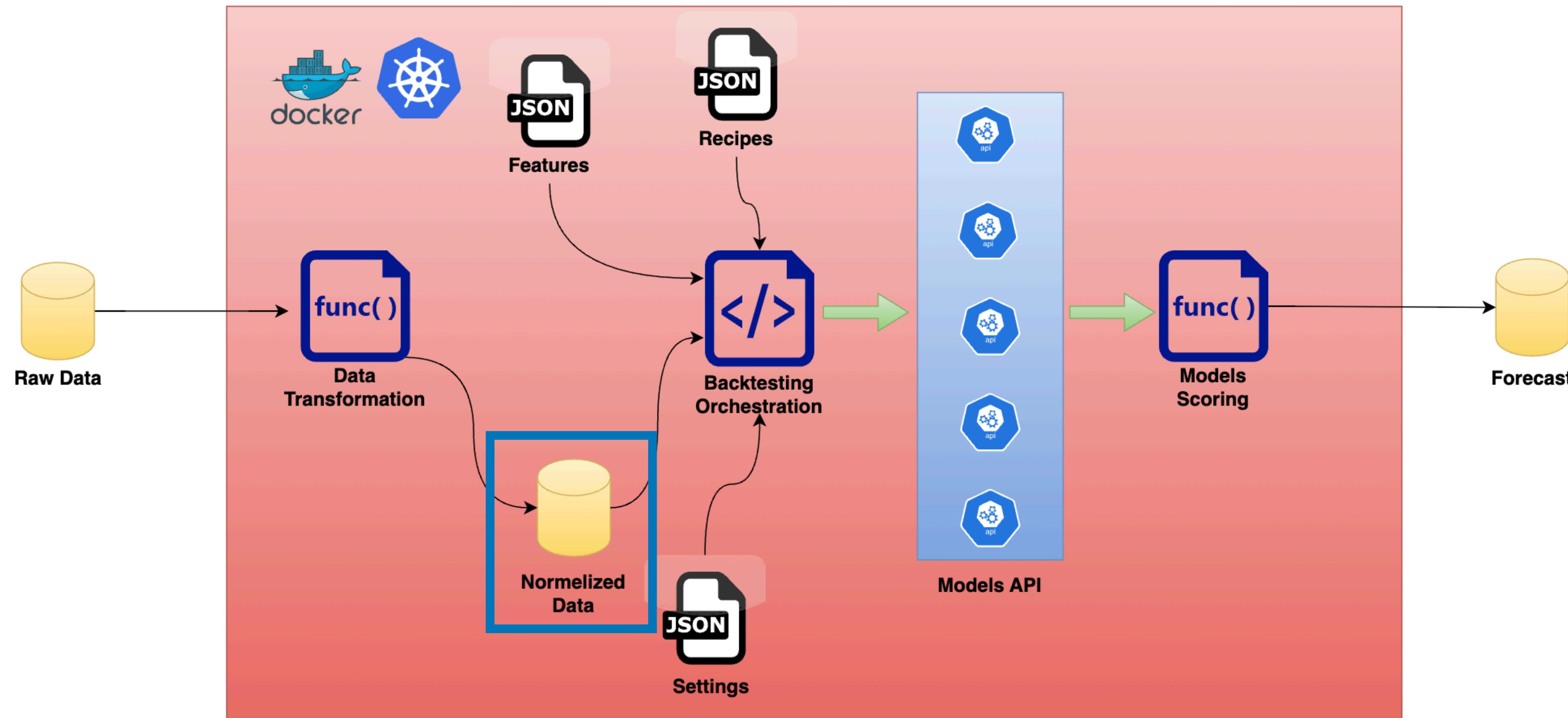
General Architecture

Data Transformation



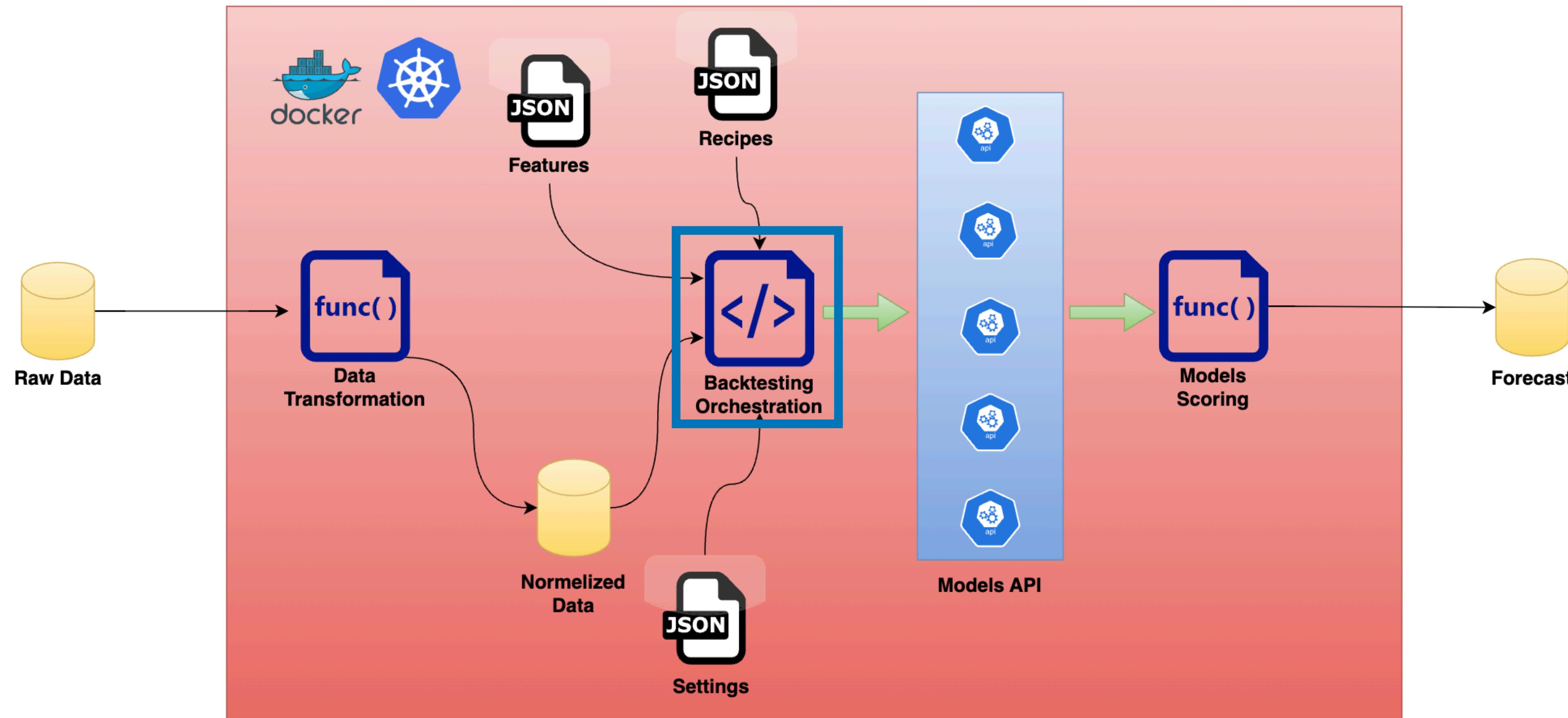
General Architecture

Data Transformation



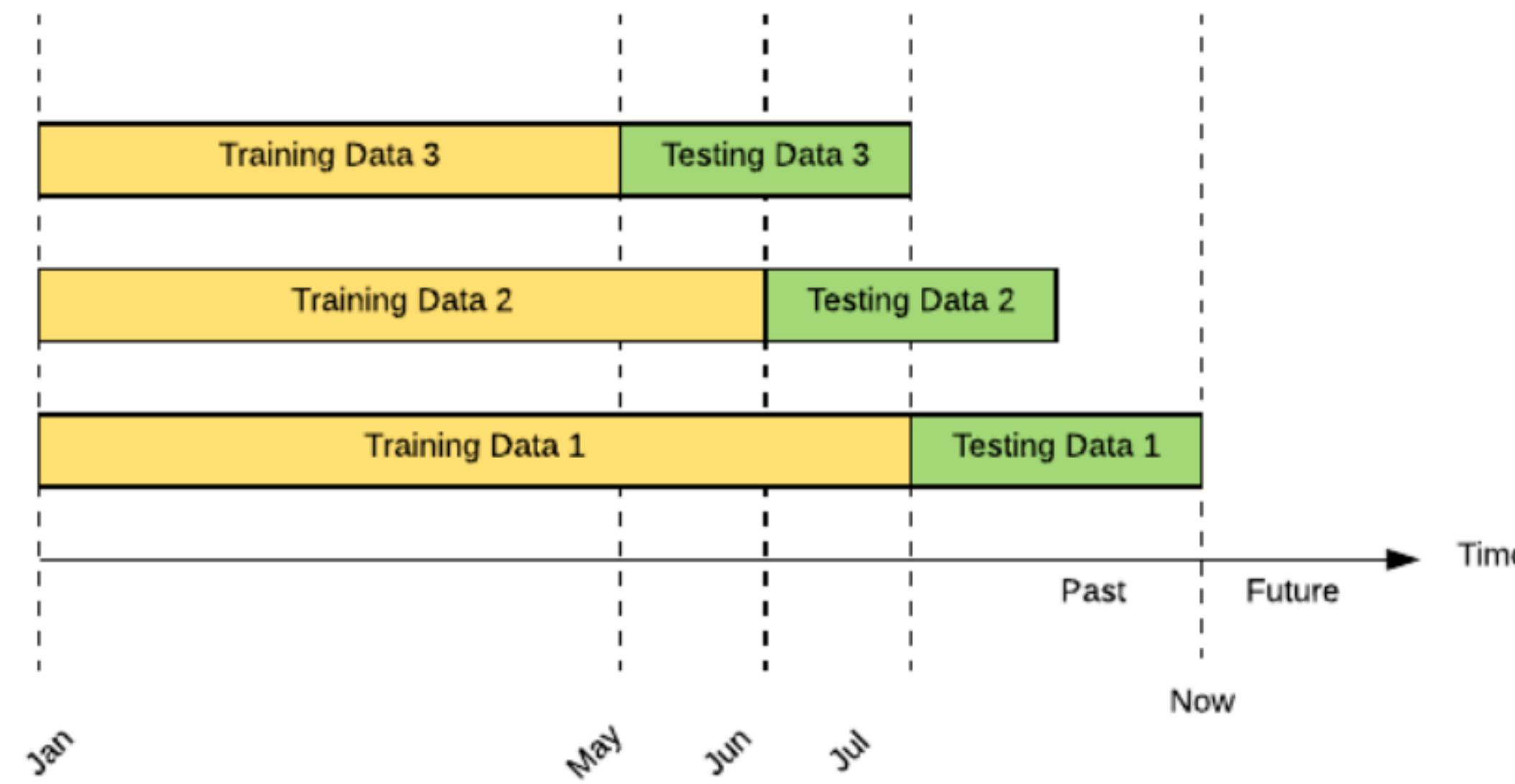
General Architecture

Backtesting Orchestration

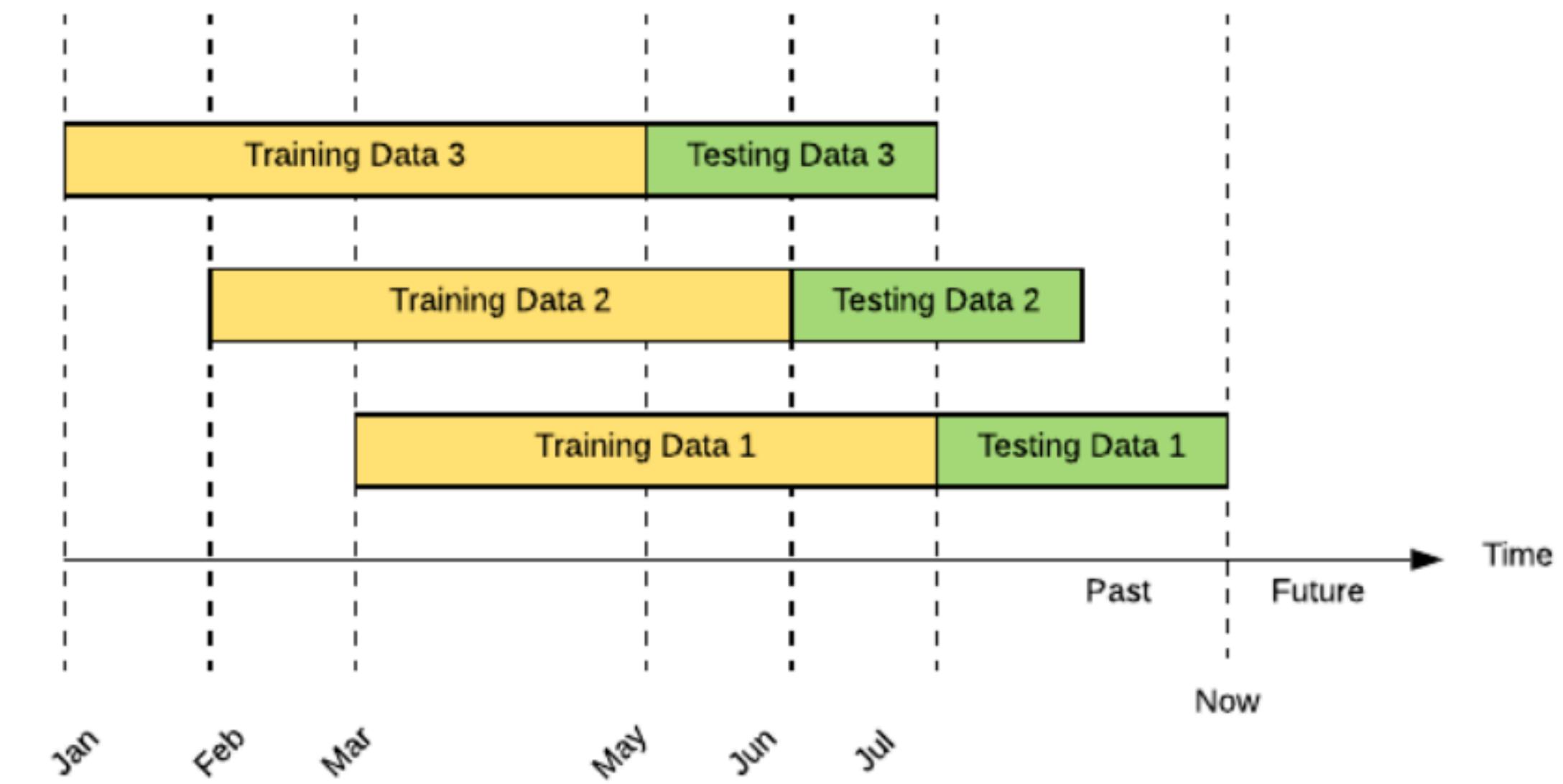


Backtesting in a nutshell is the time series equivalent to cross validation

Backtesting Approach



Expanding Window



Sliding Window

Horse Racing

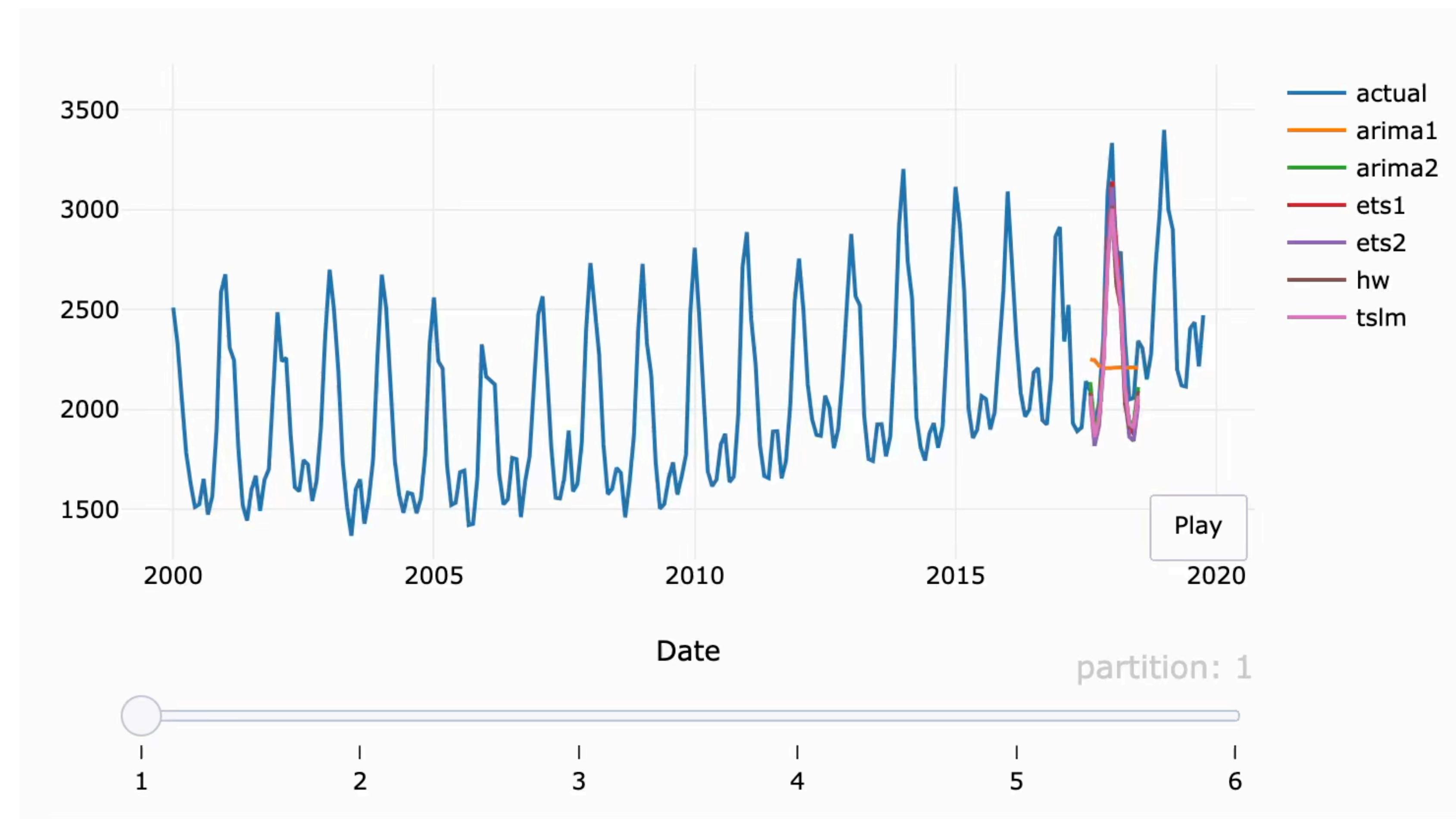


Source: Wikipedia https://en.wikipedia.org/wiki/Horse_racing

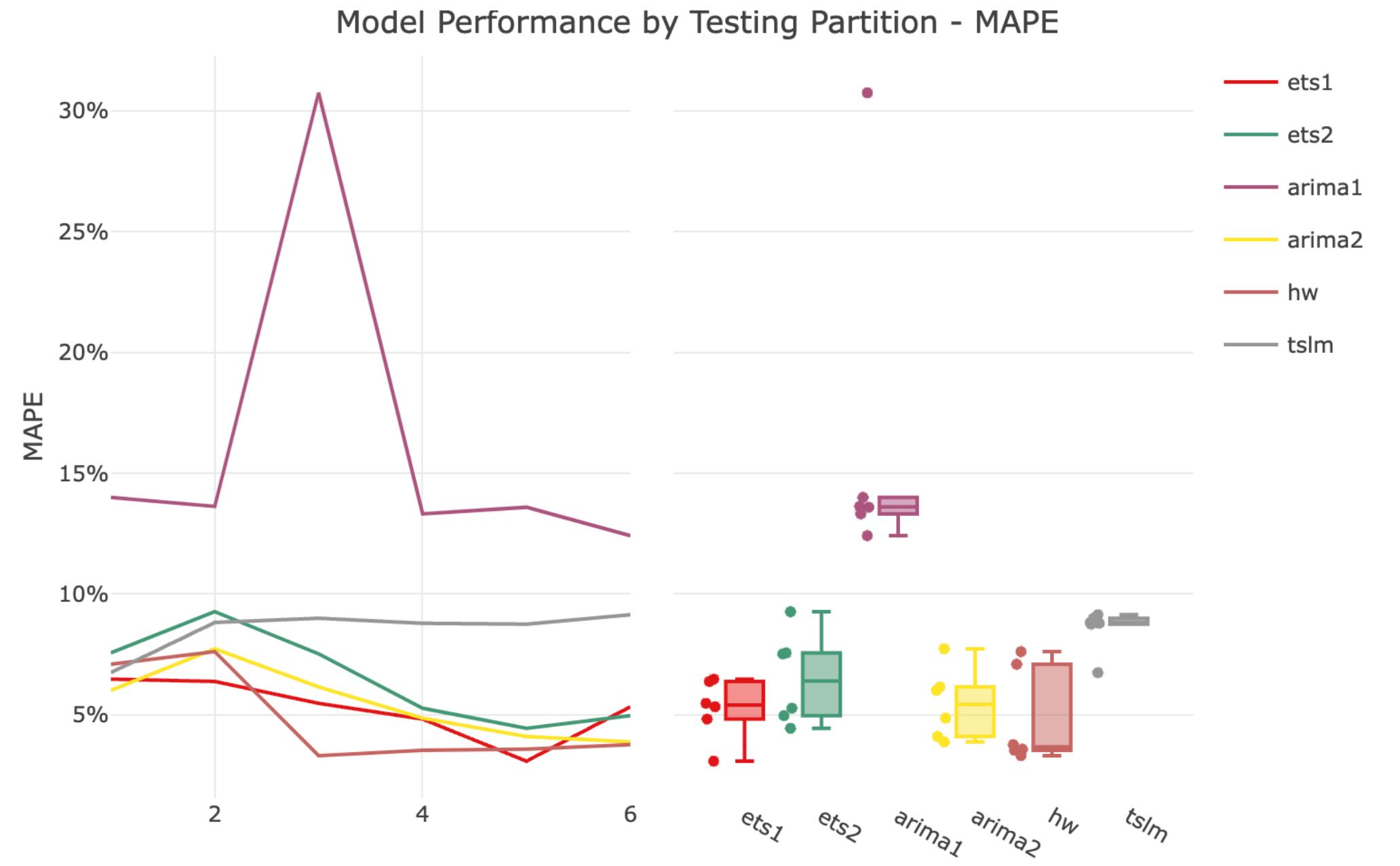
**Backtesting + Horse Racing =
AutoML**

Horse Racing with Backtesting

Training



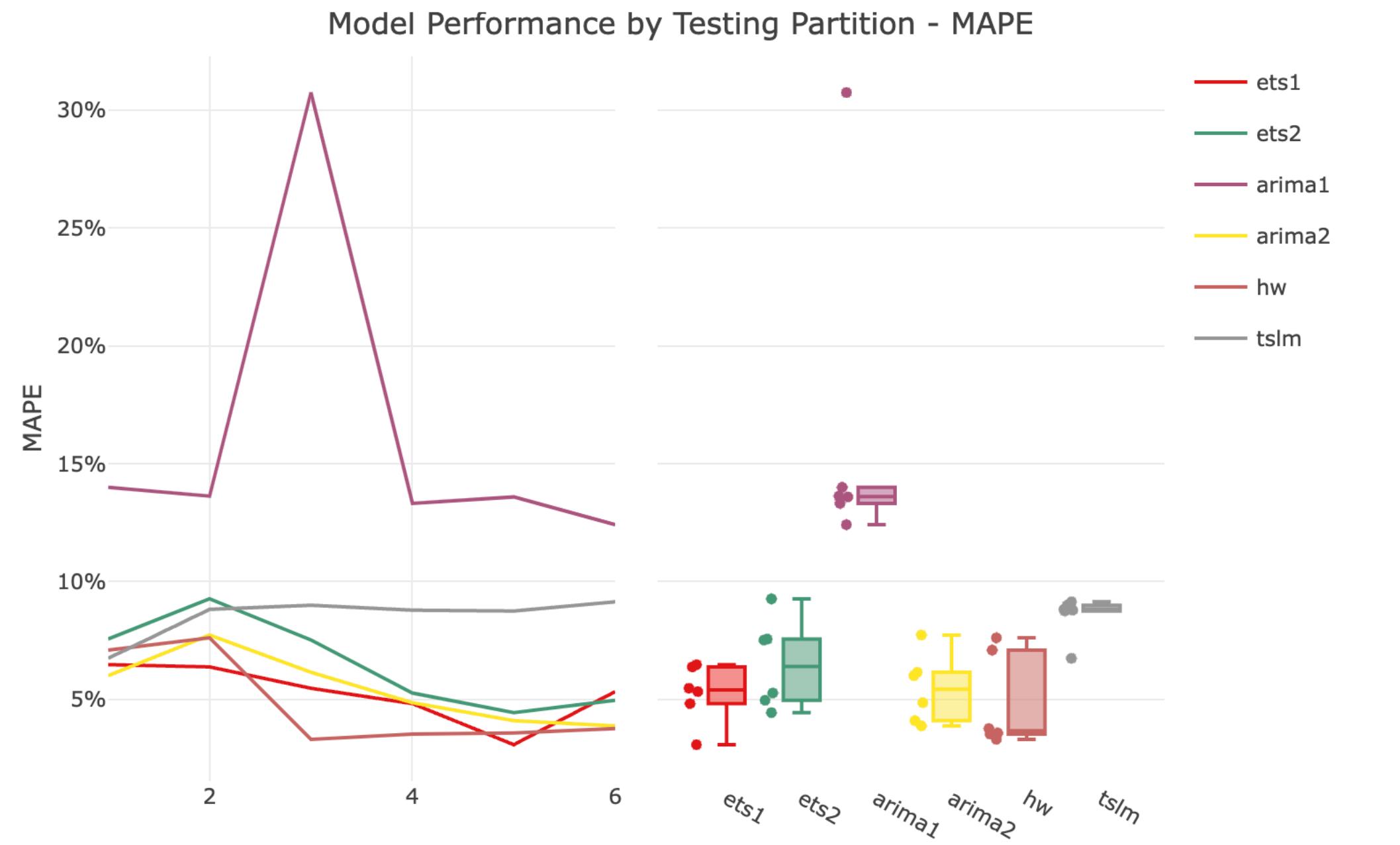
Horse Racing with Backtesting Score



```
#> # A tibble: 5 x 7
#>   model_id model  notes      avg_mape  avg_rmse `avg_coverage_95%` `avg_coverage_99%
#>   <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 with Backtesting

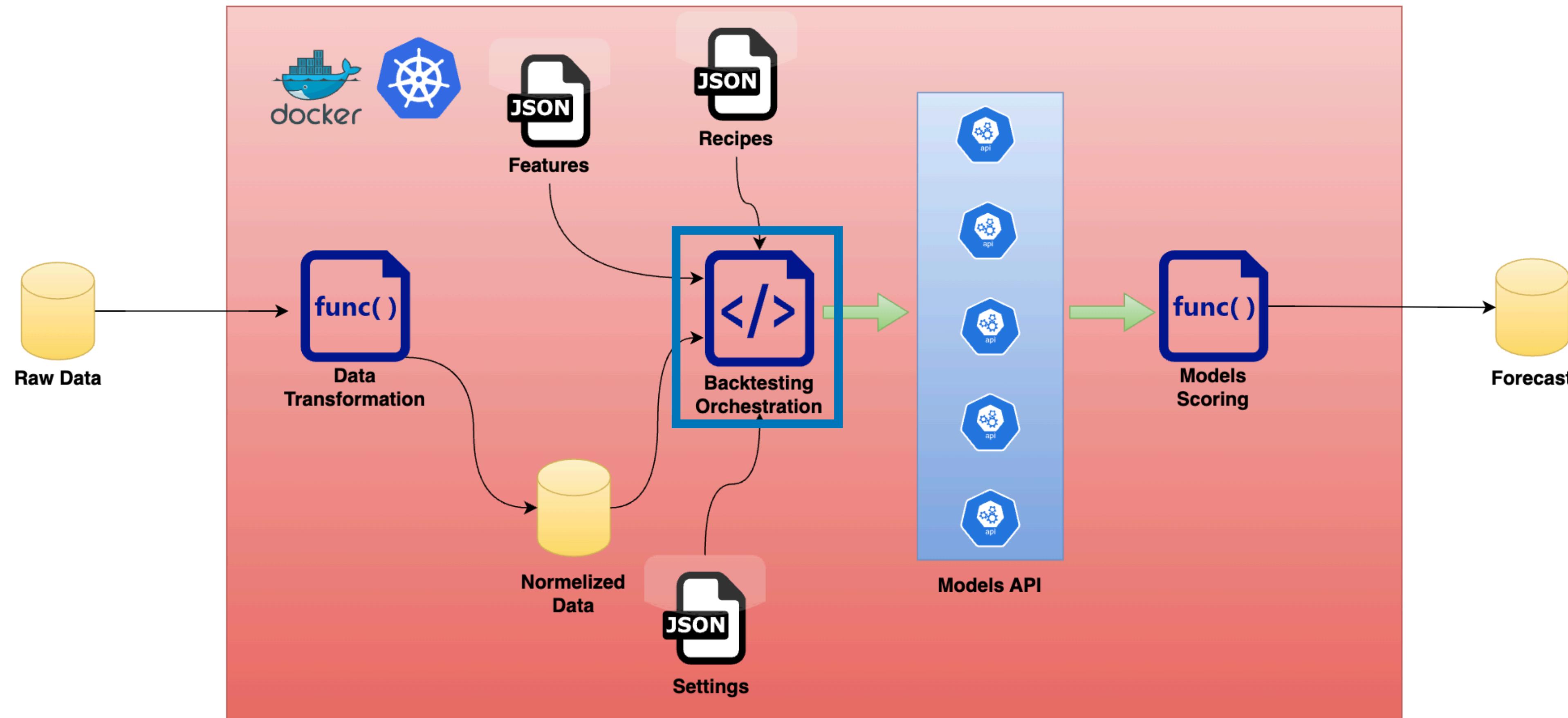
Rank



```
#> # 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
```

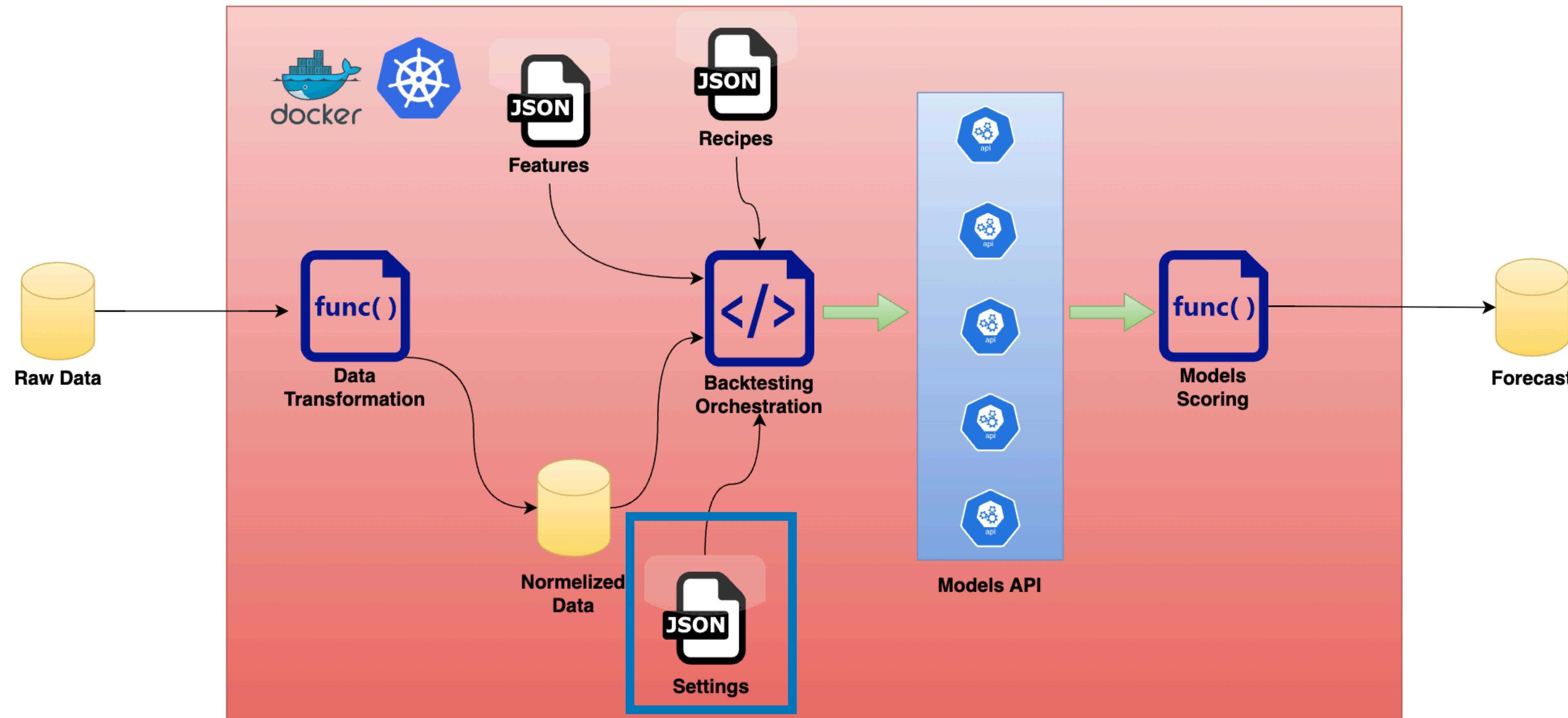
General Architecture

Backtesting Orchestration



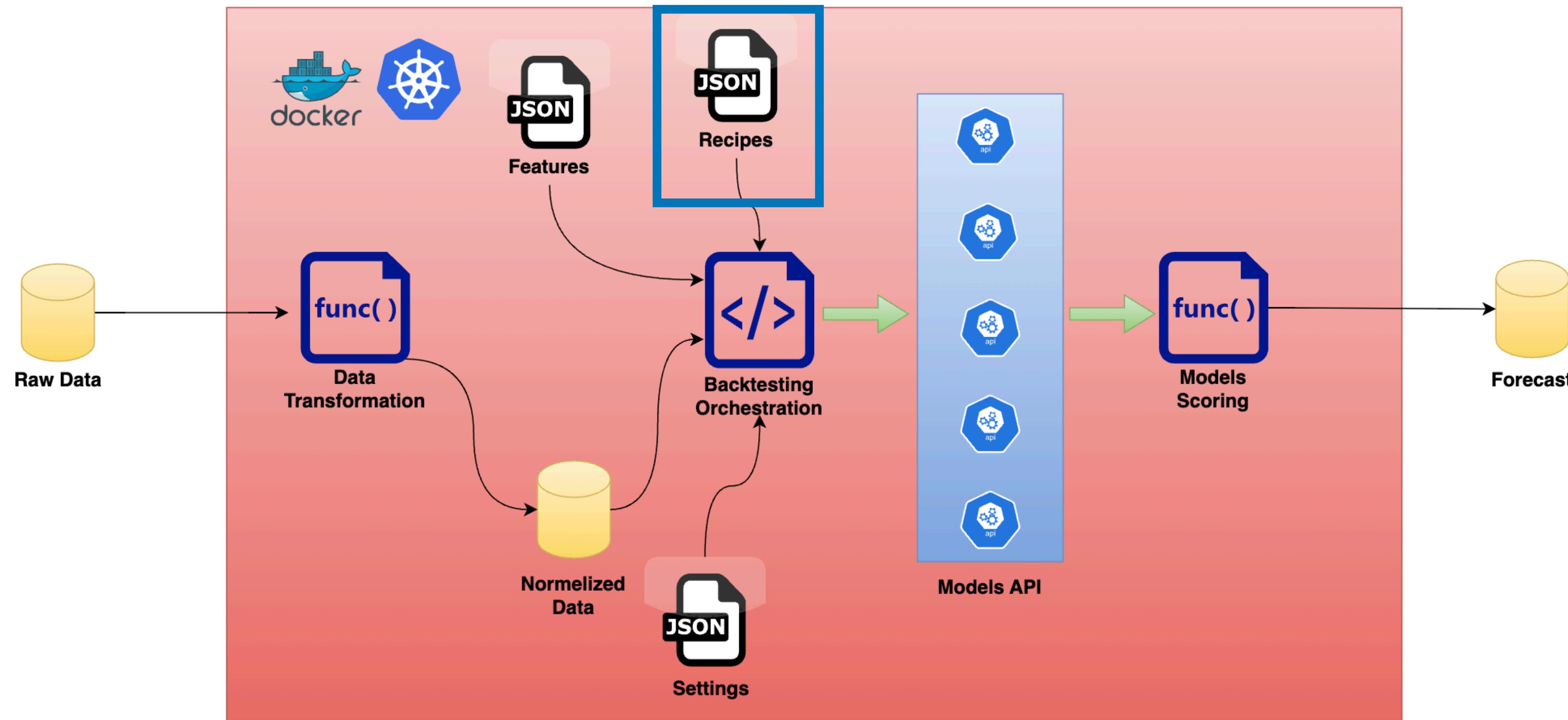
General Architecture

Backtesting Orchestration



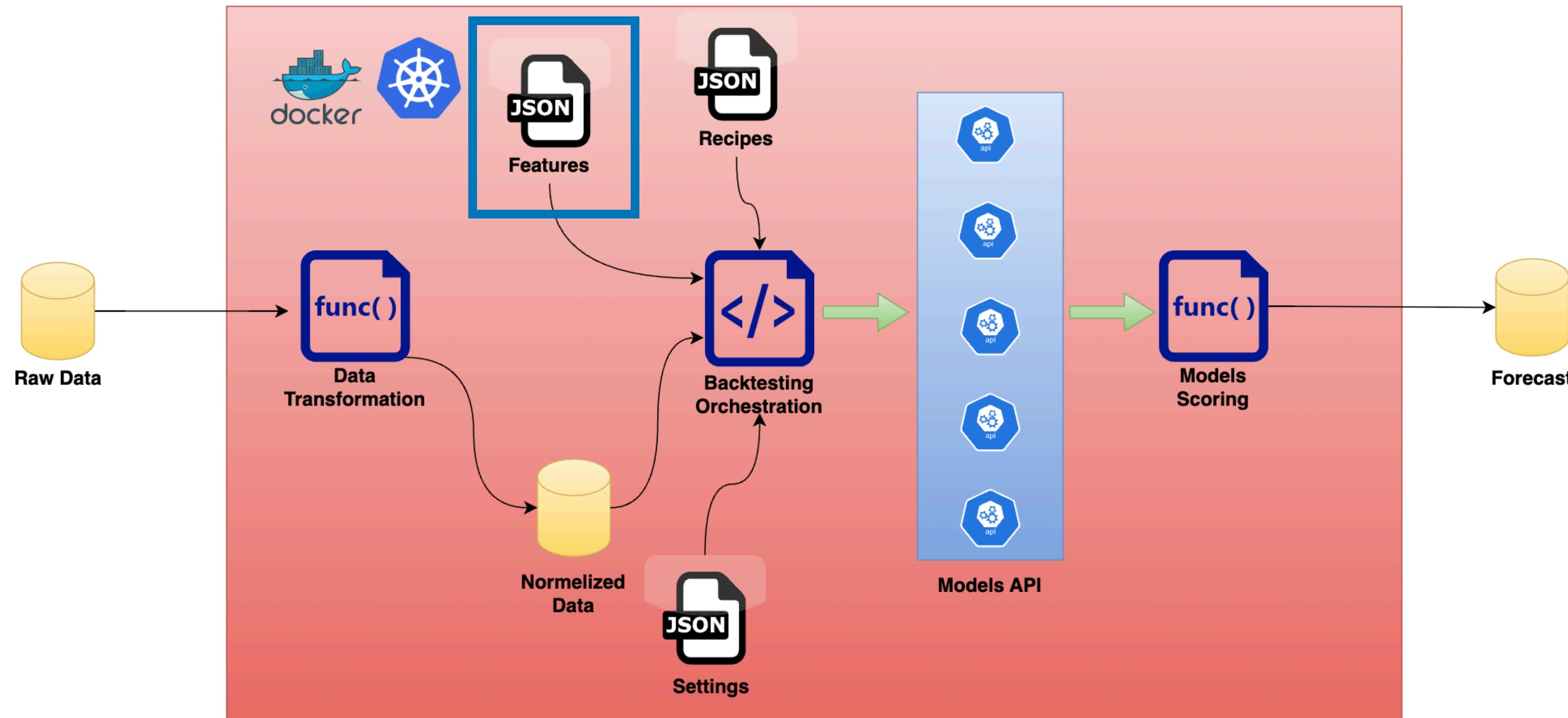
General Architecture

Backtesting Orchestration



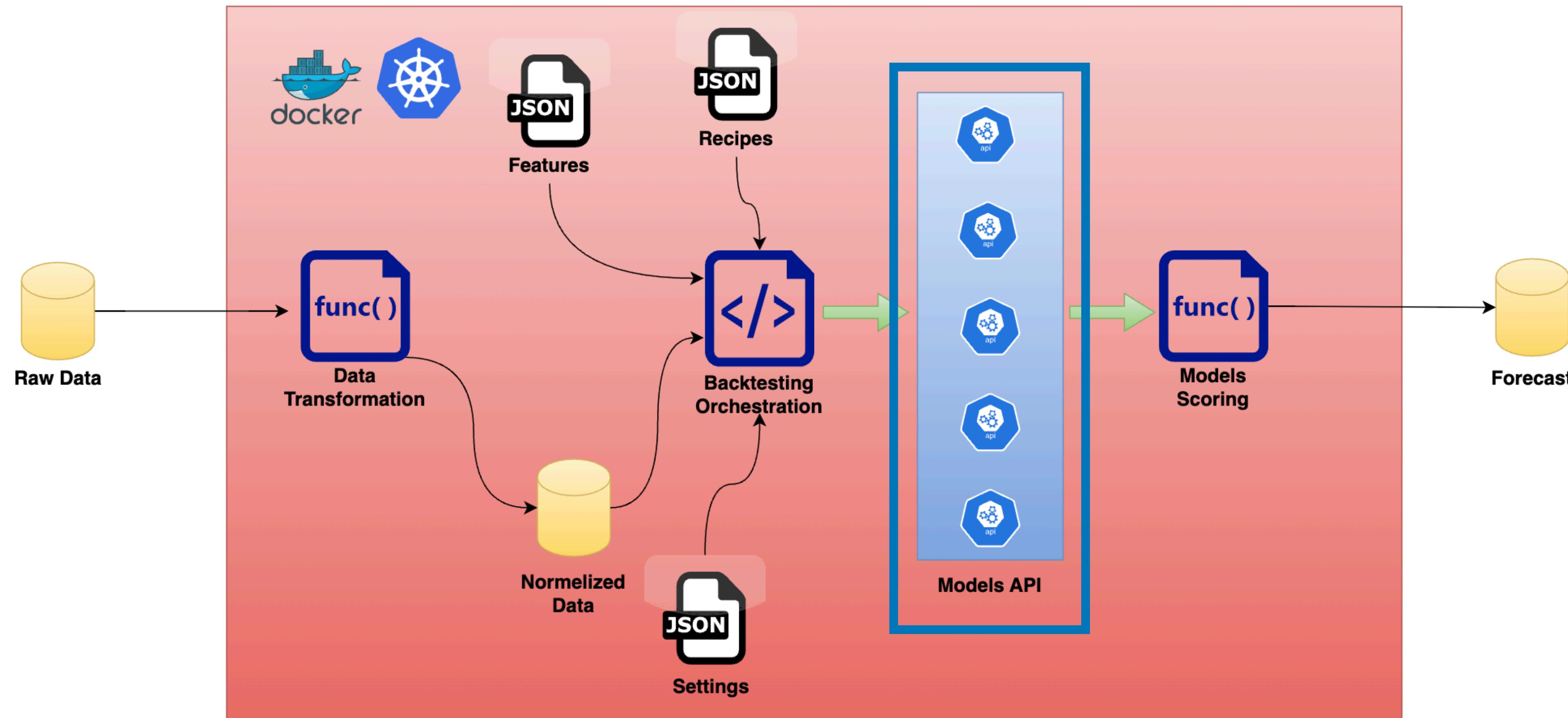
General Architecture

Backtesting Orchestration



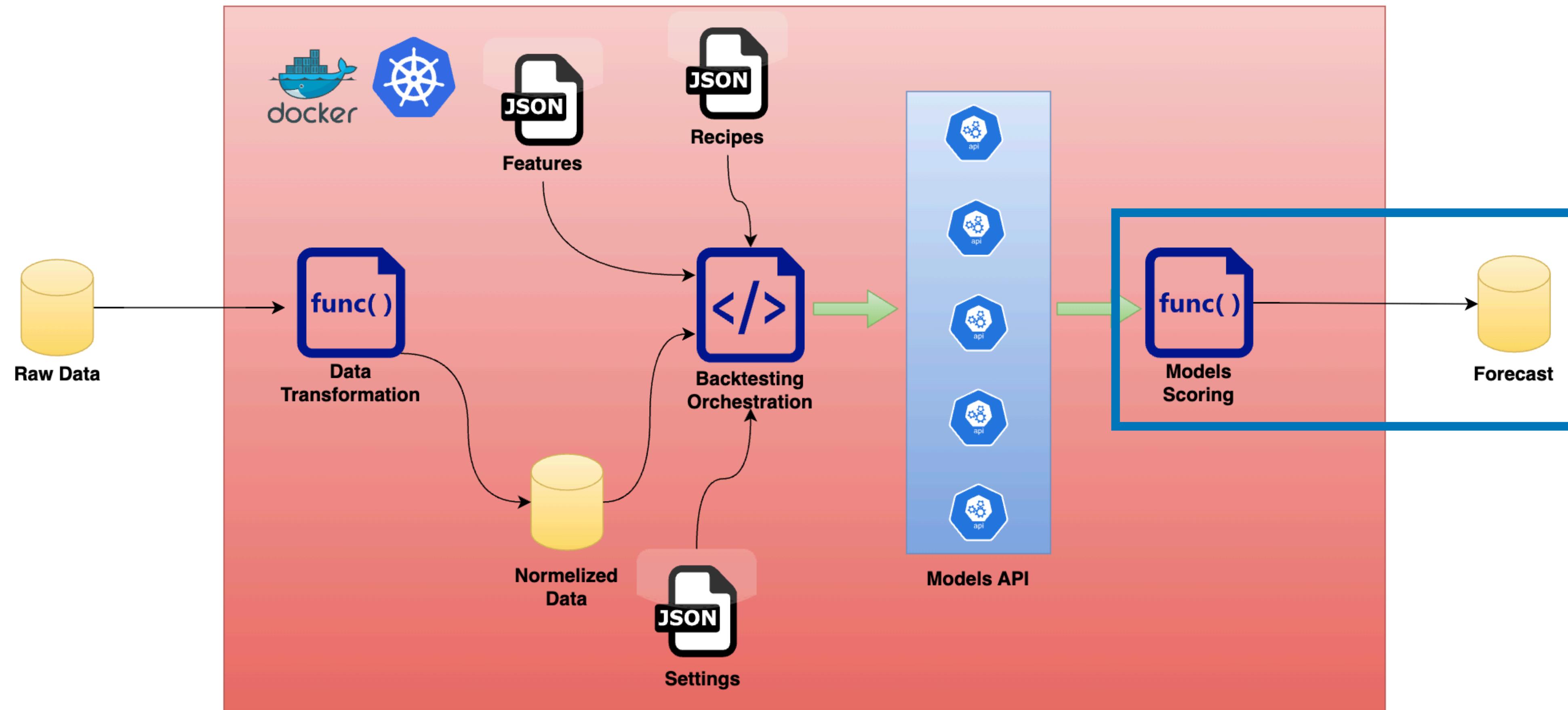
General Architecture

Backtesting Orchestration



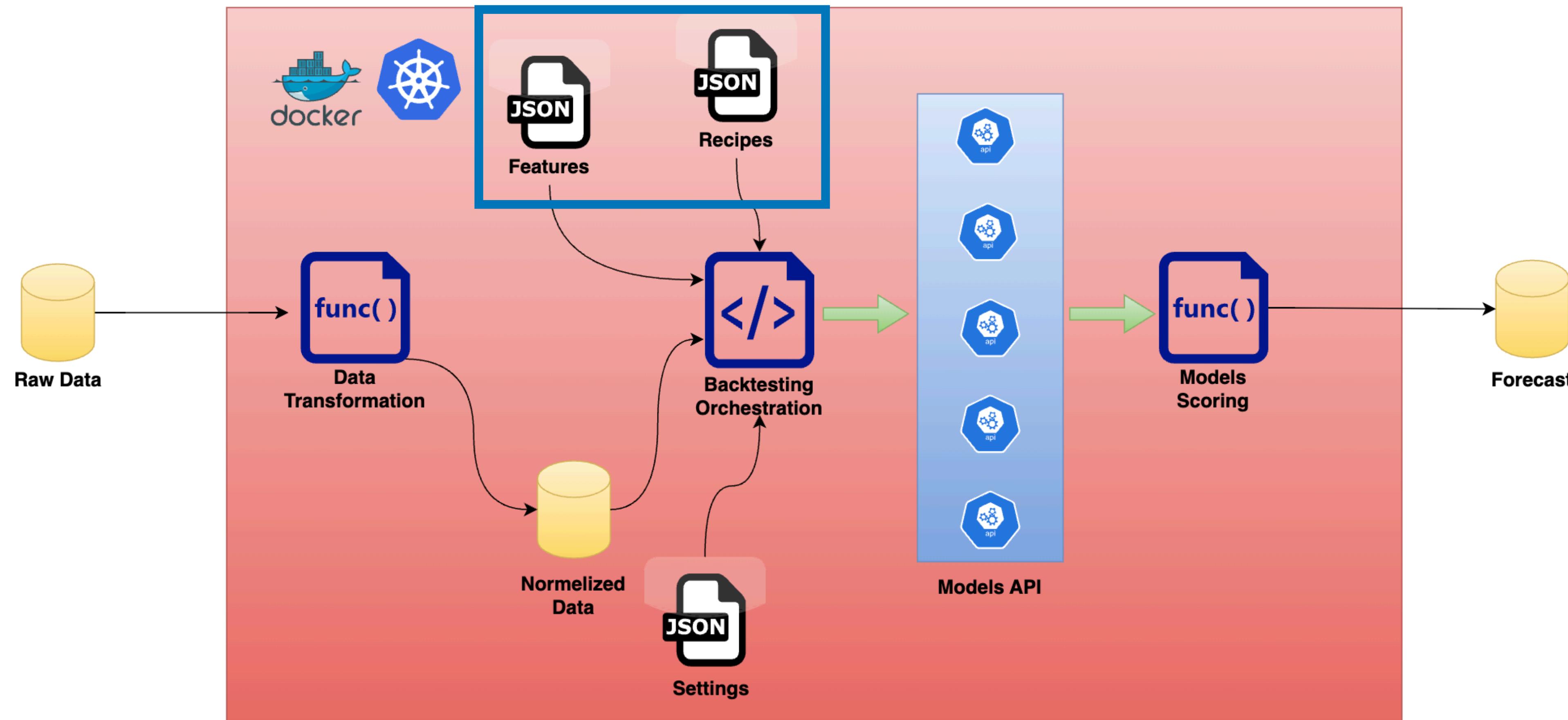
General Architecture

Backtesting Orchestration



General Architecture

Backtesting Orchestration



Main Challenges

How to identify the right model?

How to identify the right features?

Feature-based Time Series Analysis

Feature-based Time Series Analysis

In a Nutshell

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.

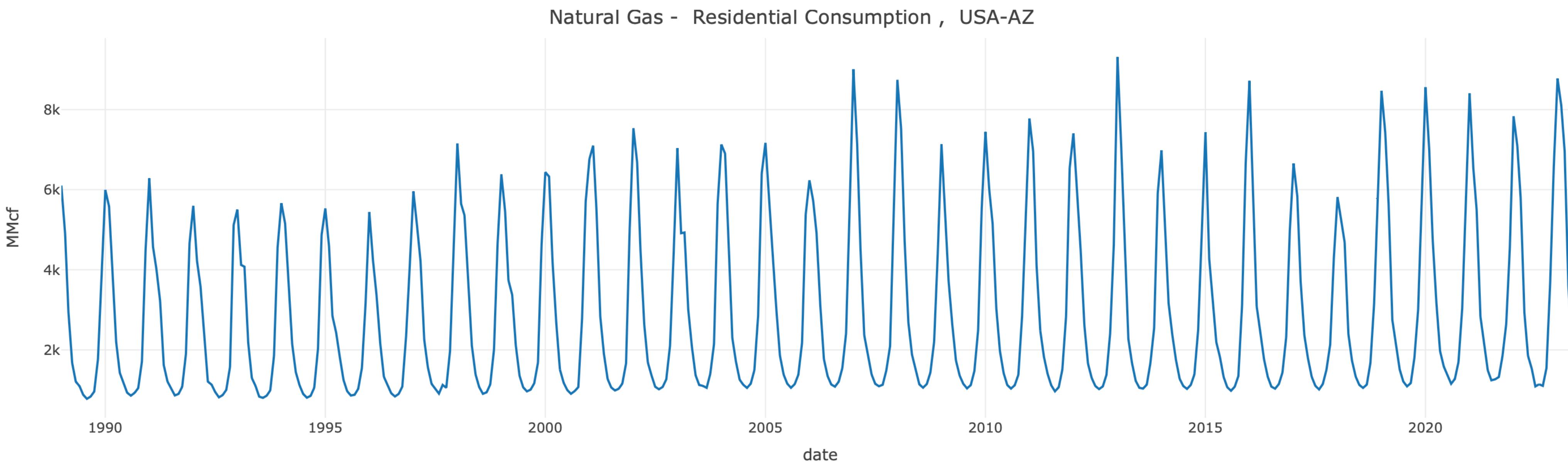
```
library(tidyverse)
library(tsibble)
library(feasts)
```

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.

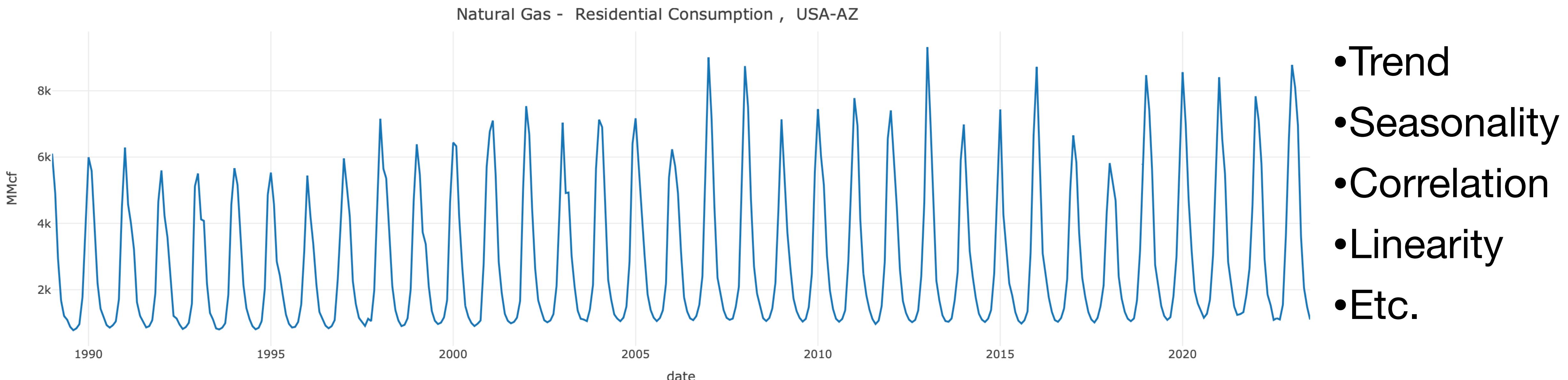
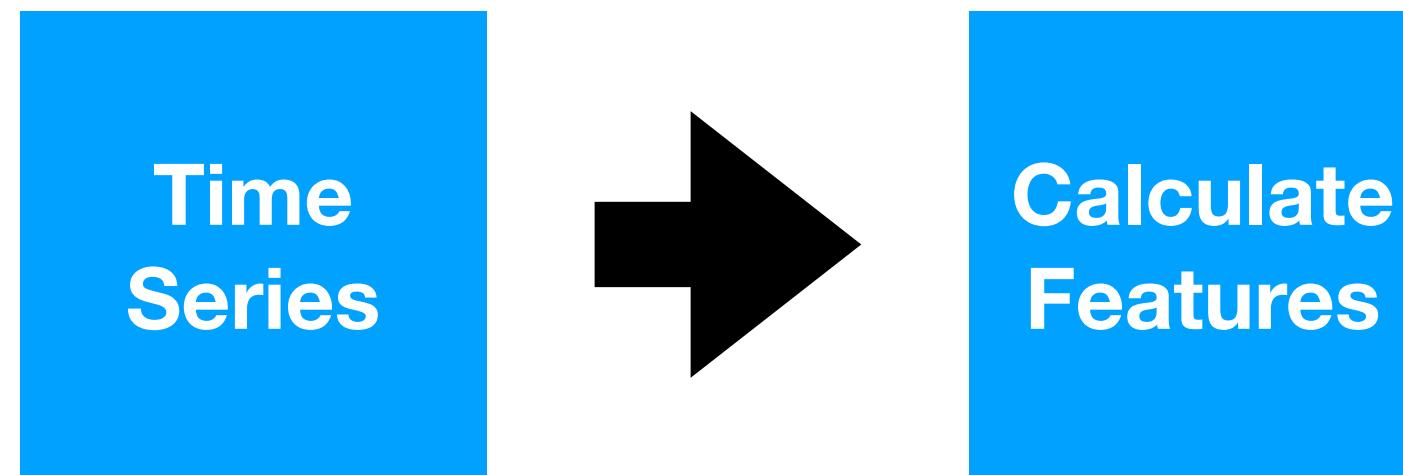
Feature-based Time Series Analysis

In a Nutshell

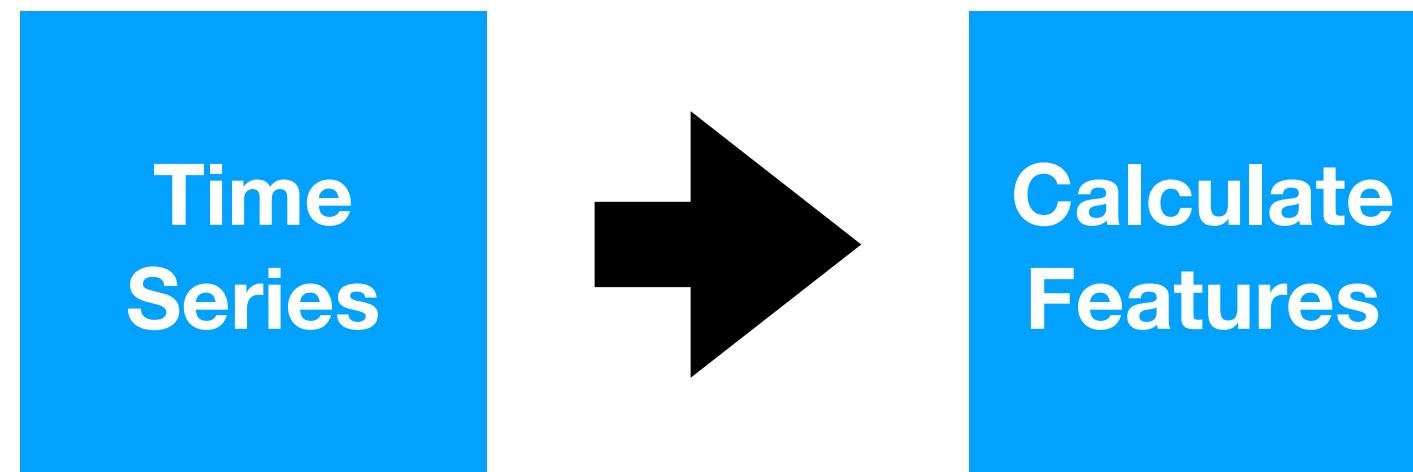


Feature-based Time Series Analysis

In a Nutshell

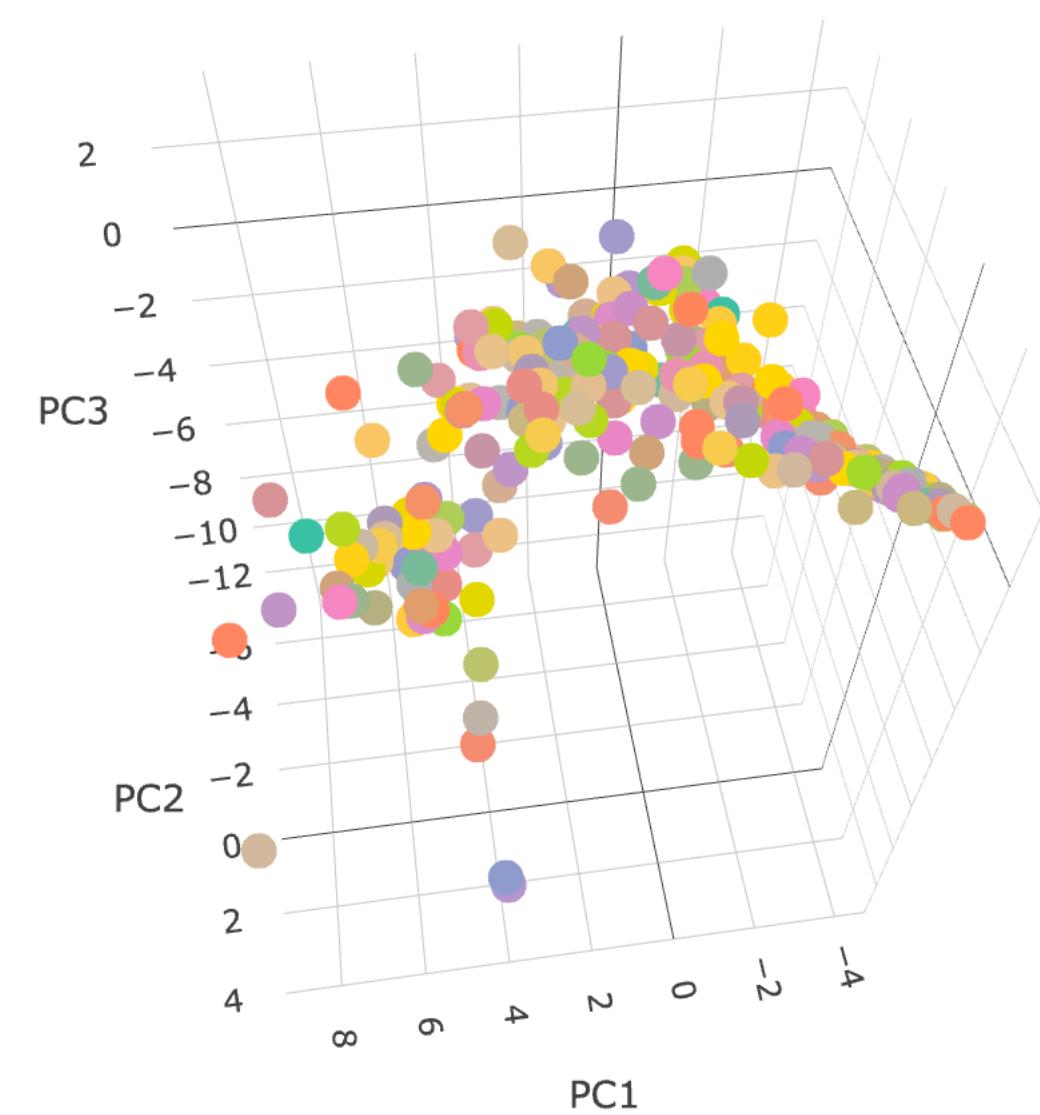
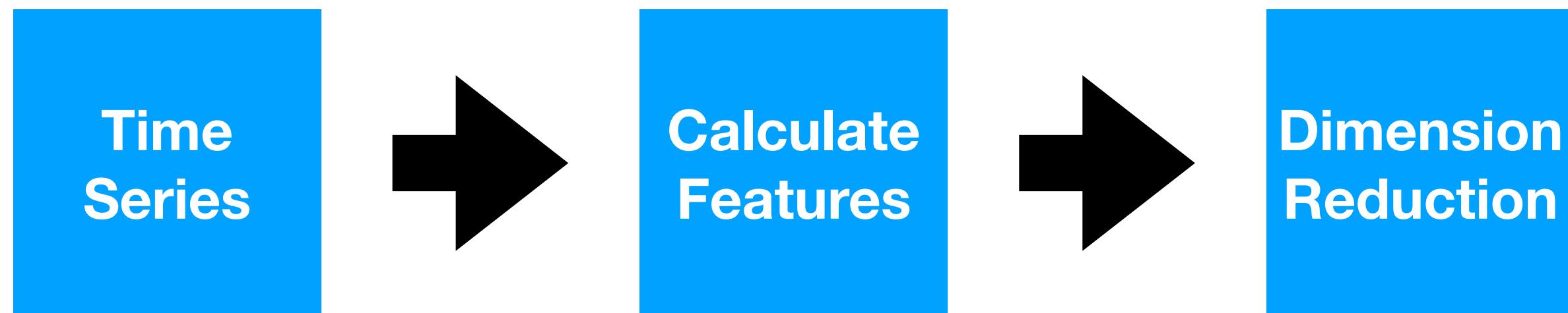


Feature-based Time Series Analysis In a Nutshell



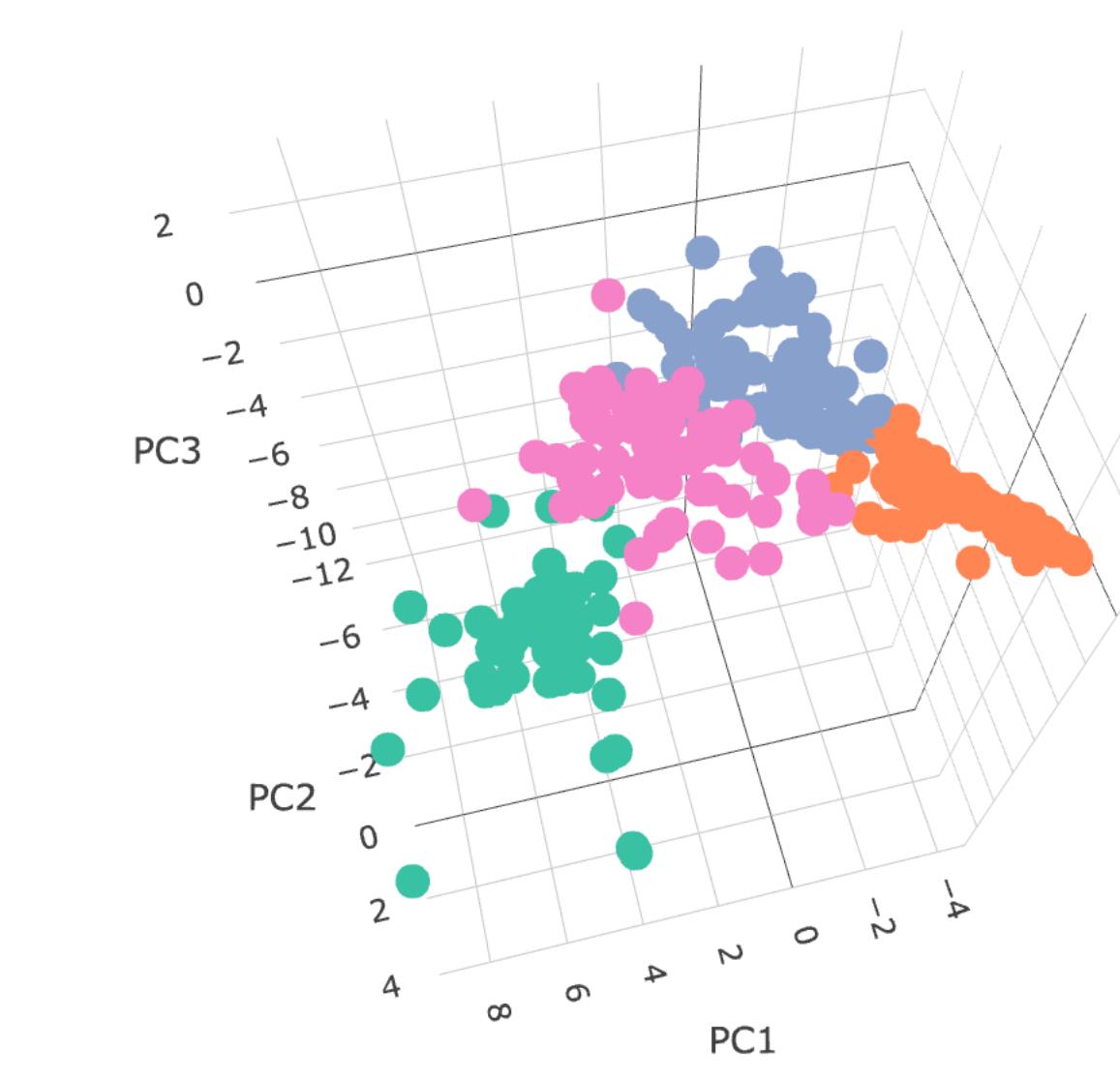
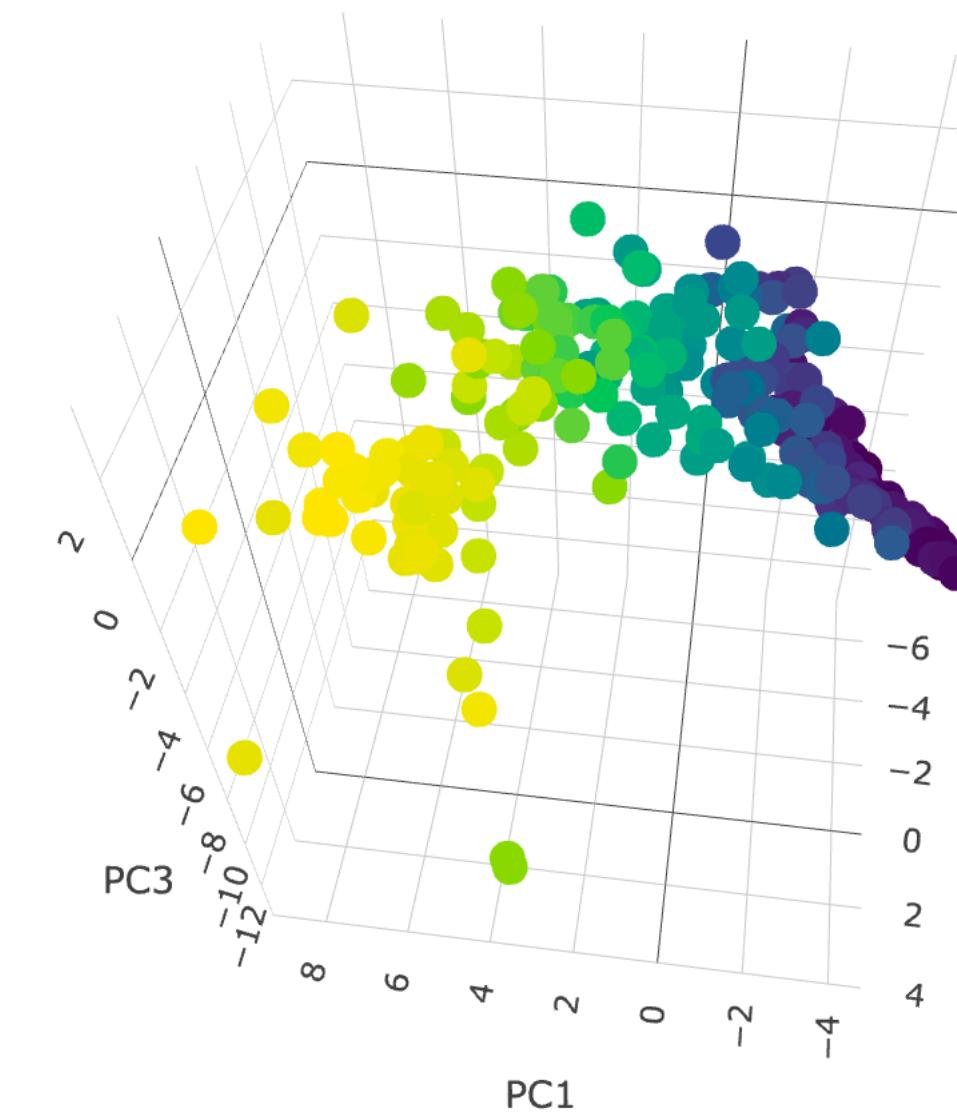
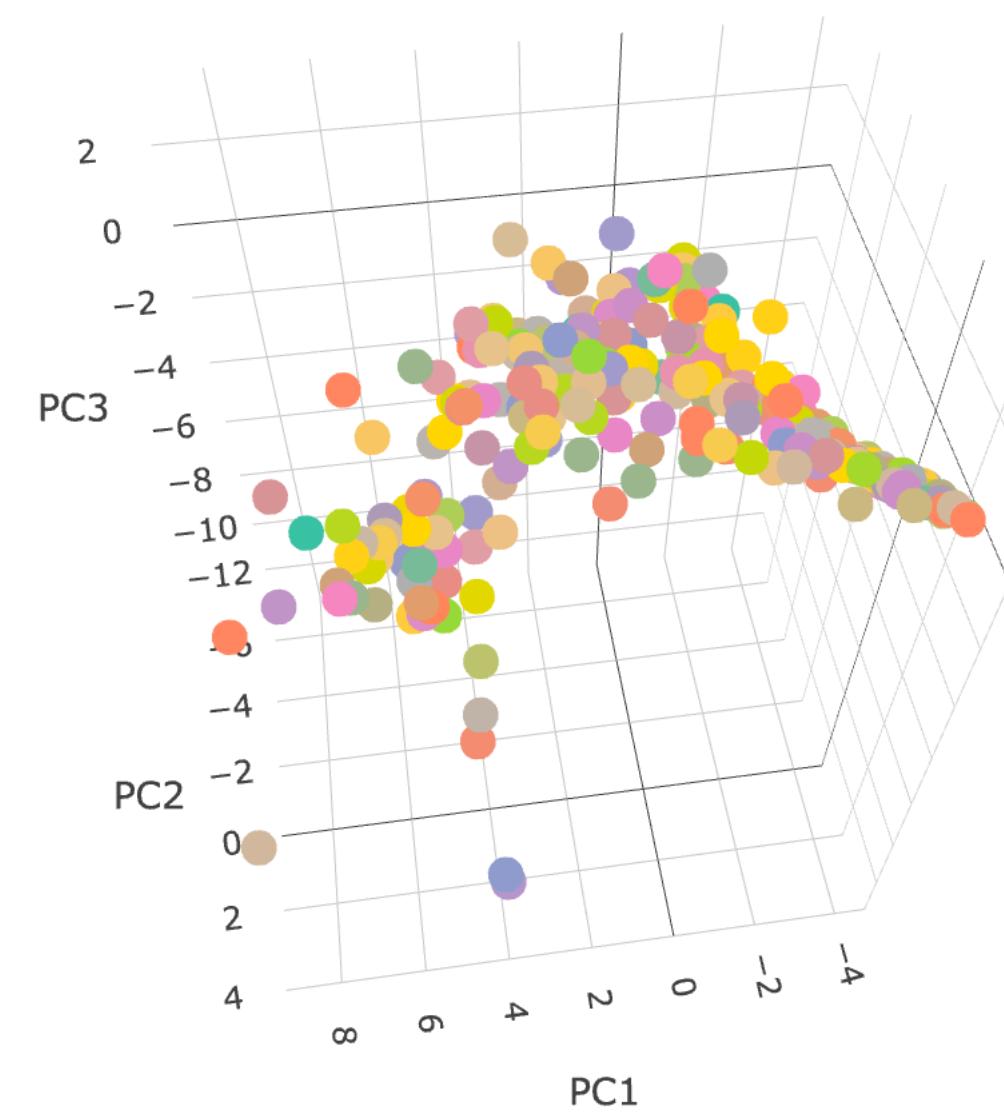
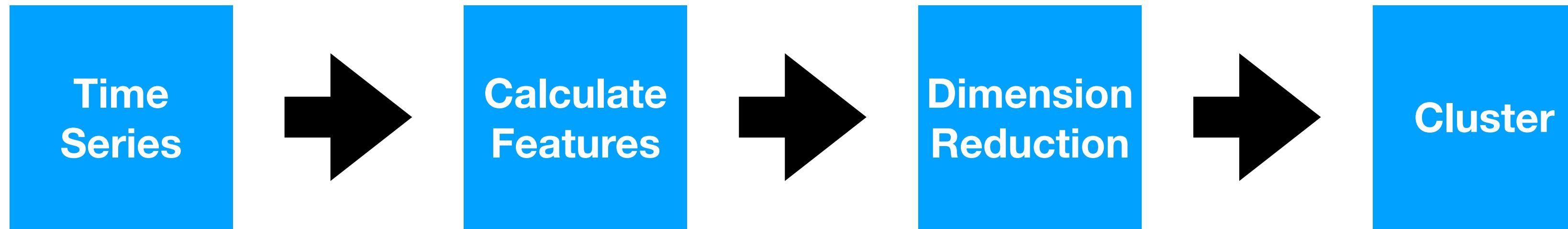
Feature-based Time Series Analysis

In a Nutshell



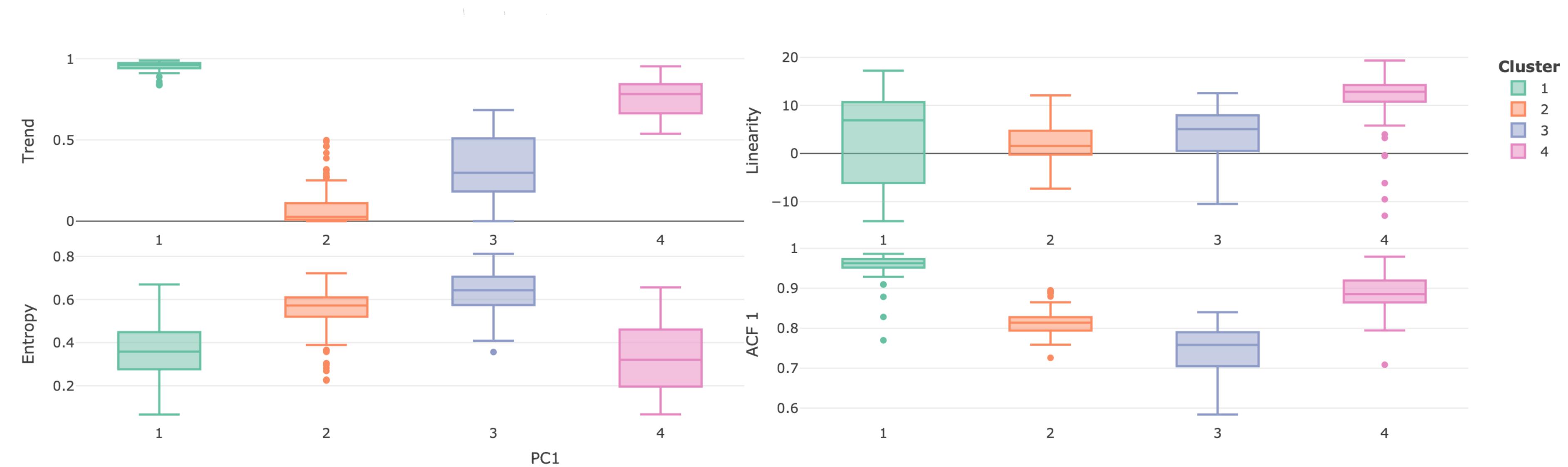
Feature-based Time Series Analysis

In a Nutshell



Feature-based Time Series Analysis

In a Nutshell

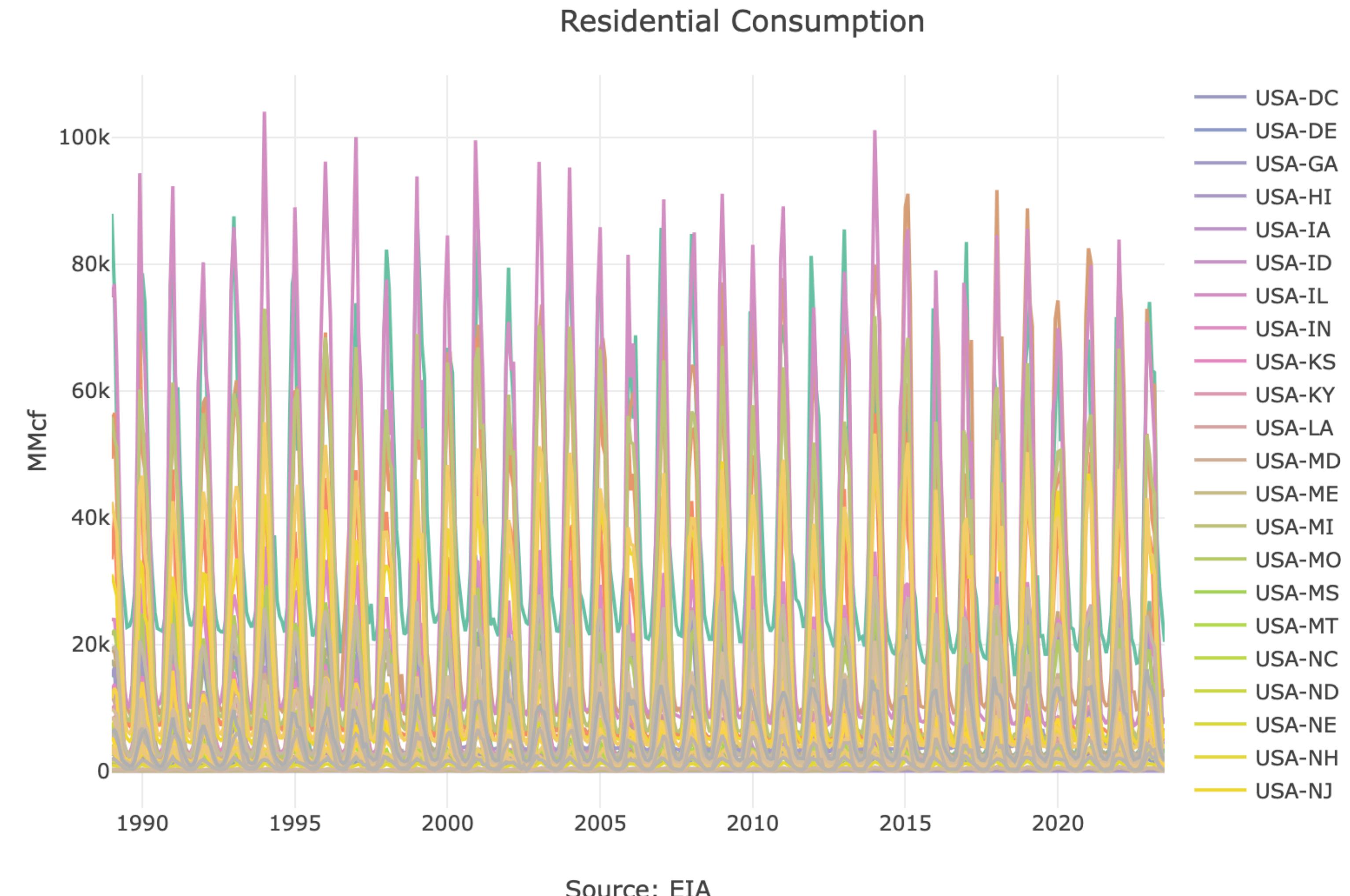


Demo

Feature-based Time Series Analysis

The Demand for Natural Gas in the USA

- 250 series
- 5 categories
- Monthly



Summary

- Scaling depending on infra and models
- Backtesting is a great method for automation
- Features engineering important but time consuming
- Features based time series analysis can help with features engineering

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/forecasting-at-scale>
- Uber Engineering: <https://www.uber.com/en-DE/blog/backtesting-at-scale/>

Questions?

Thank You!