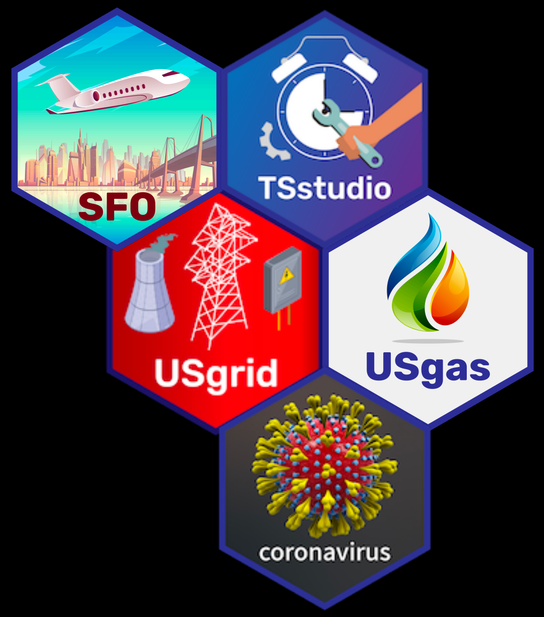
Analyzing Time Series at Scale with Cluster Analysis in R

Workshop for Ukraine

Rami Krispin, Oct 17, 2024

About Me

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



Agenda

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

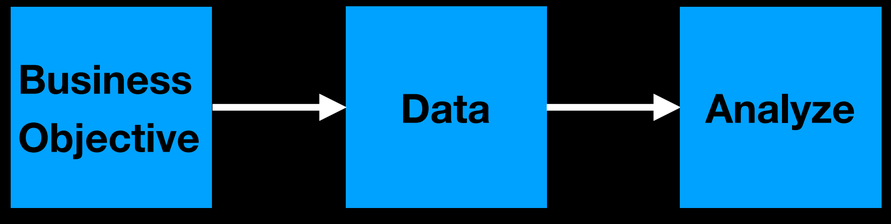
- **Time series -** a vector of data points measured over time
- **Regular time series -** the data measurement is equally spaced (i.e., hour, day, month, etc.)
- **Irregular time series -** the data measurement is not equally spaced
- **Forecast -** a prediction of observations in the future with a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

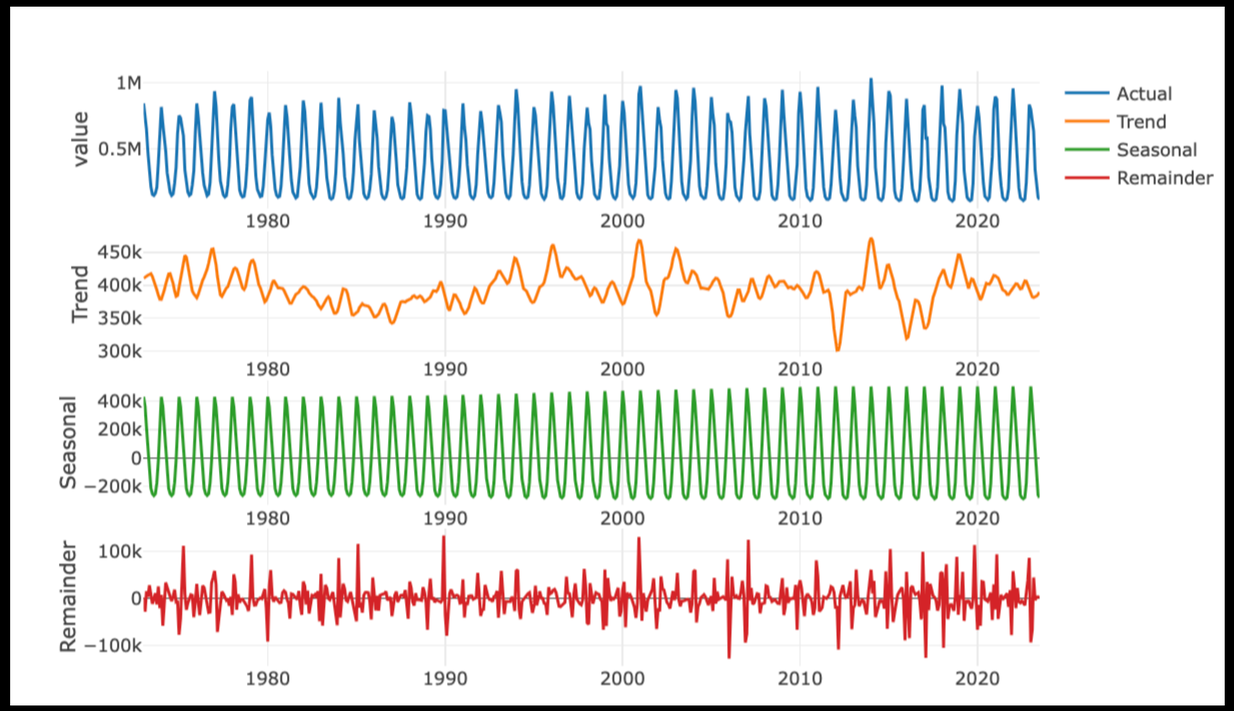
Problem Statement

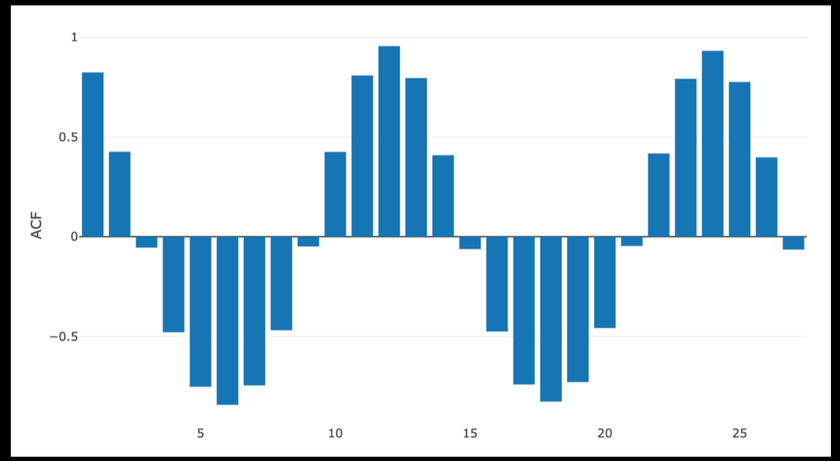
Scaling Limitation

Forecasting A Single Series

Traditional Approach

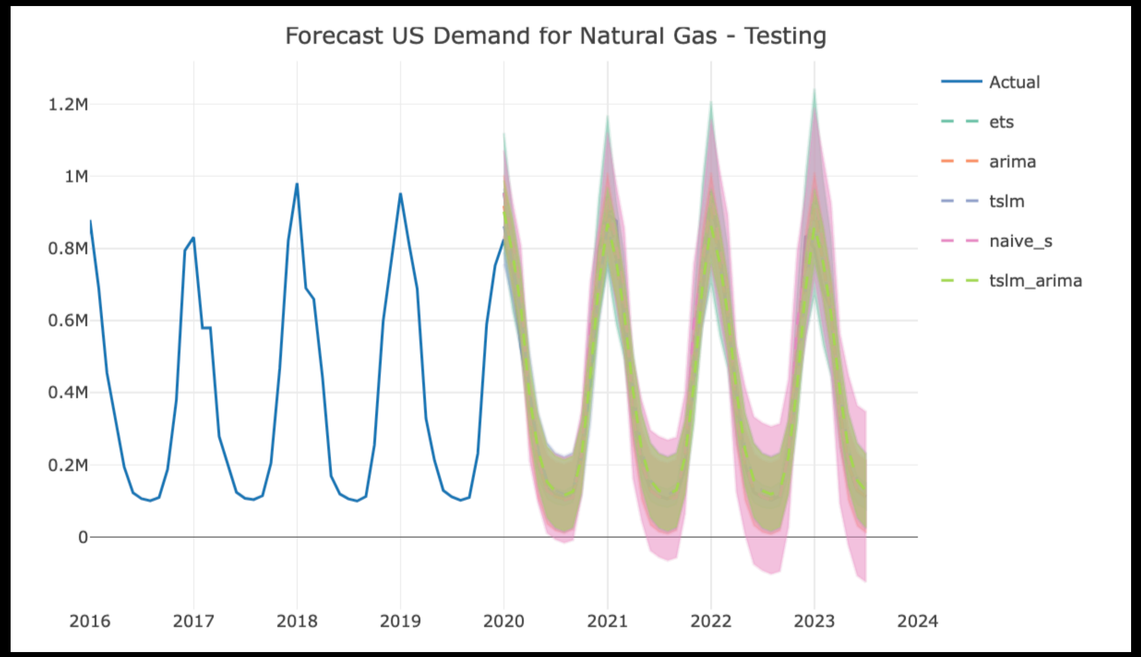






Forecasting A Single Series

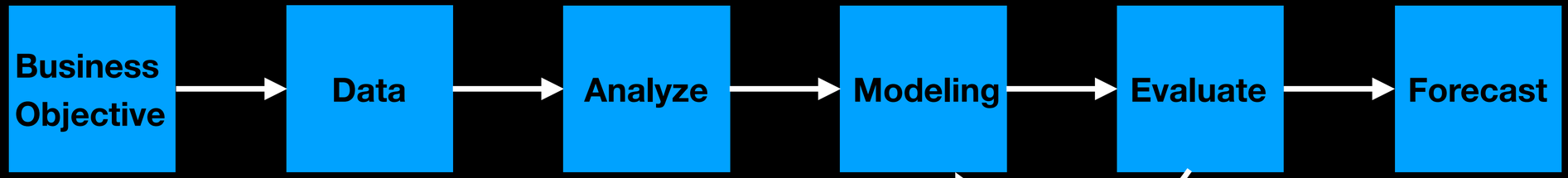
Traditional Approach

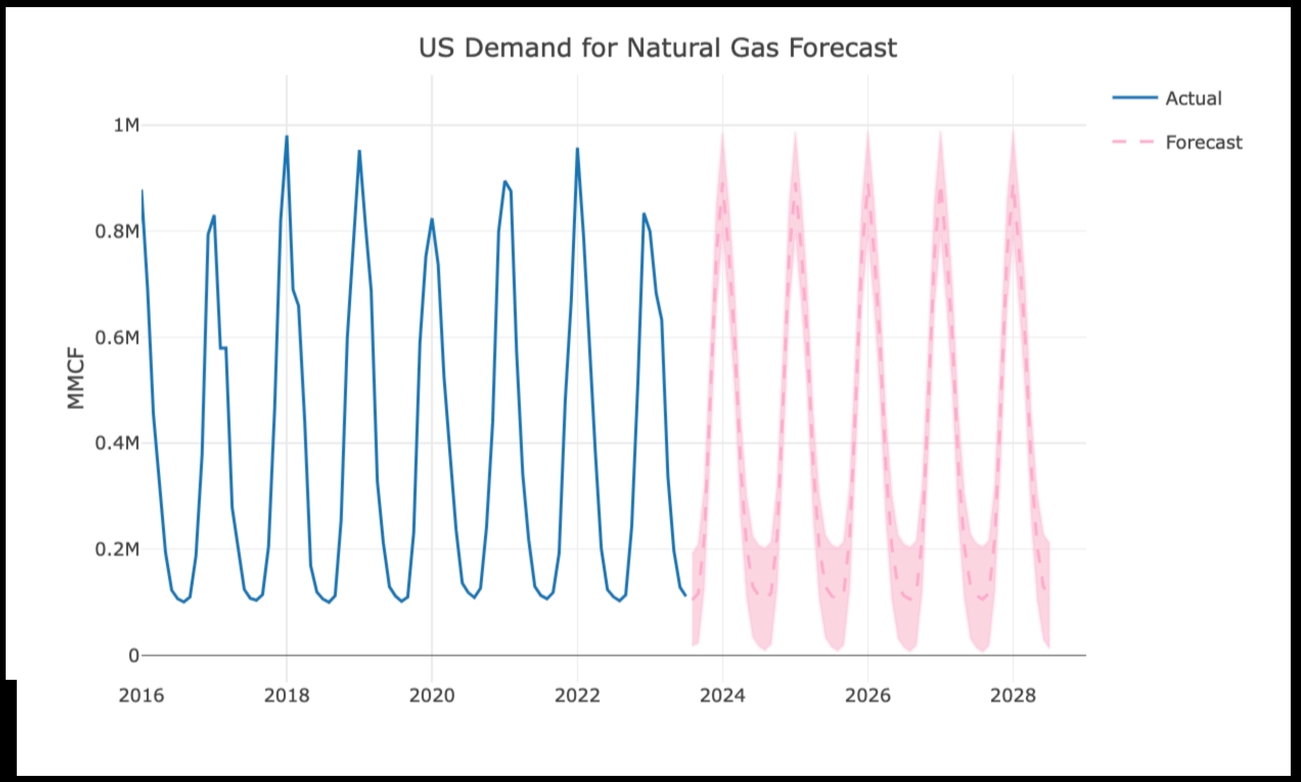



```
# A tibble: 5 × 3
  model          mape     rmse
  <chr>         <dbl>    <dbl>
1 arima        0.0612 3.76e18
2 ets          0.0681 7.42e18
3 naive_s      0.0928 1.88e19
4 tslm_arima   0.115  7.89e18
5 tslm         0.115  7.11e18
```

Forecasting A Single Series

Traditional Approach



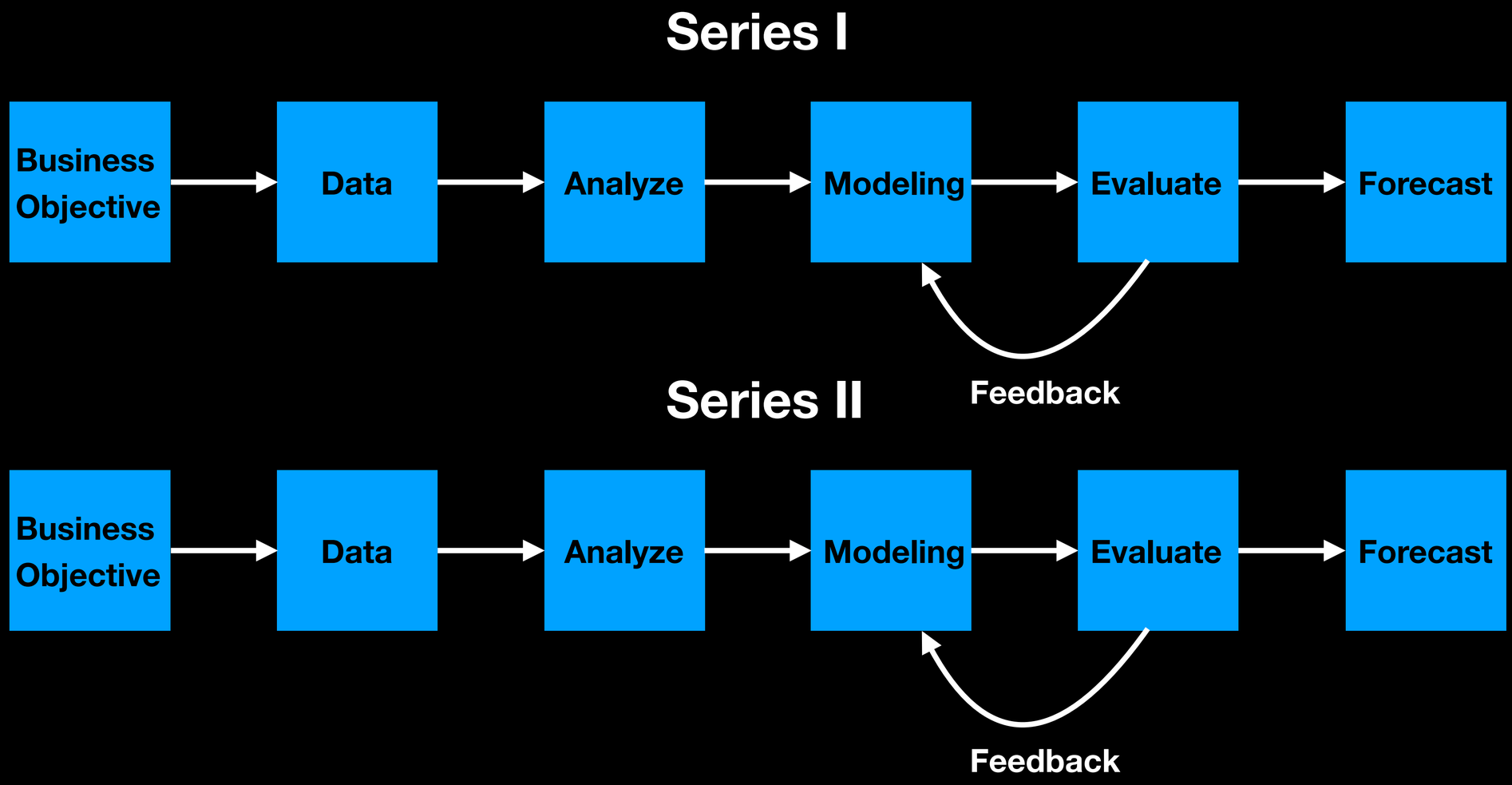


Feedback

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

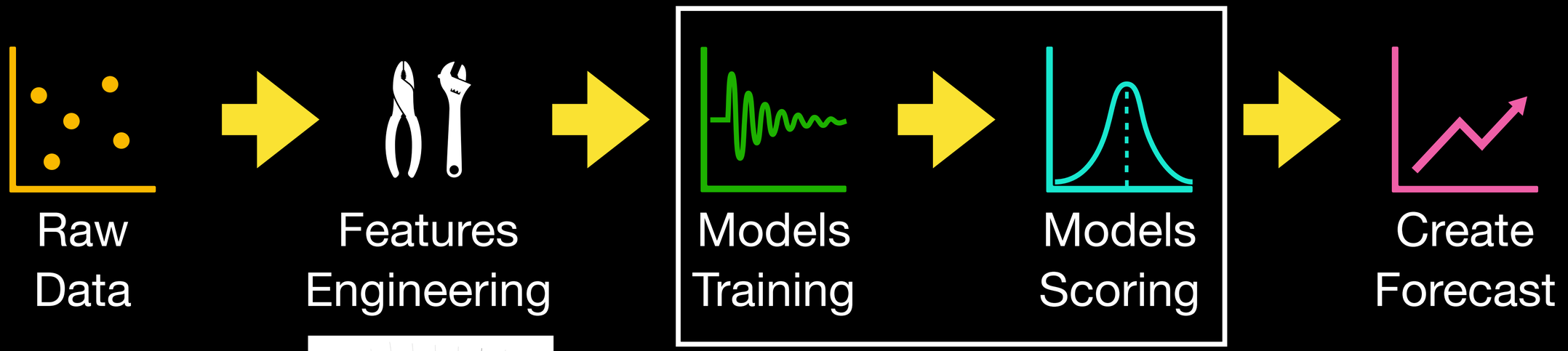
Forecasting at Scale

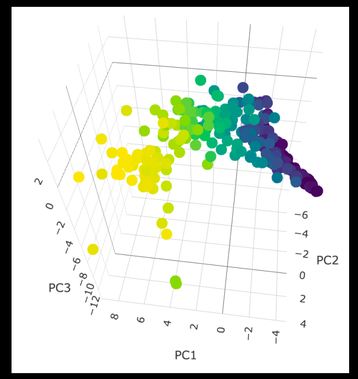
Definition

- 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

Expending Window

Sliding Window

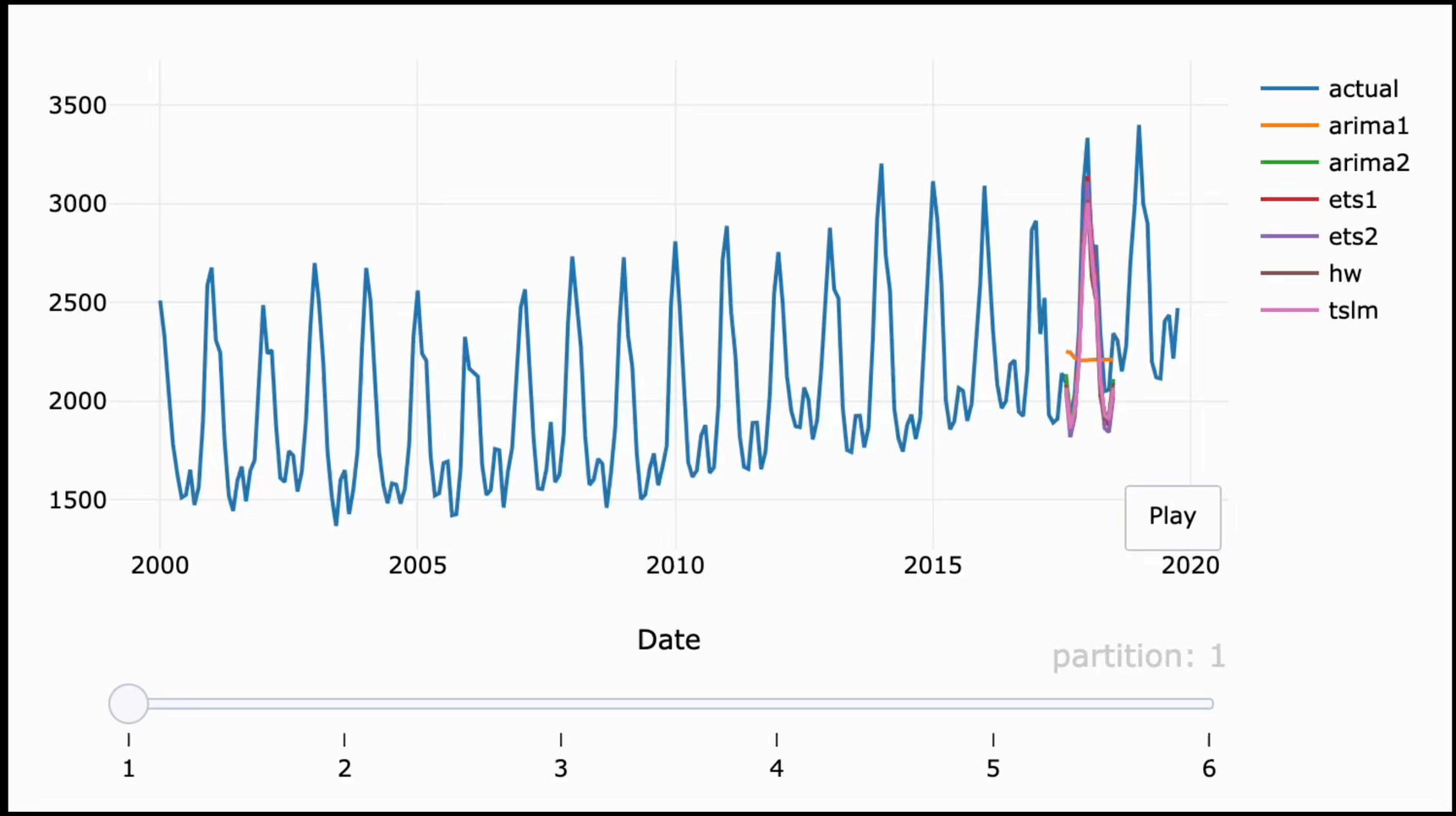
Source: Uber Engineering https://www.uber.com/en-DE/blog/backtesting-at-scale/

Horse Racing

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

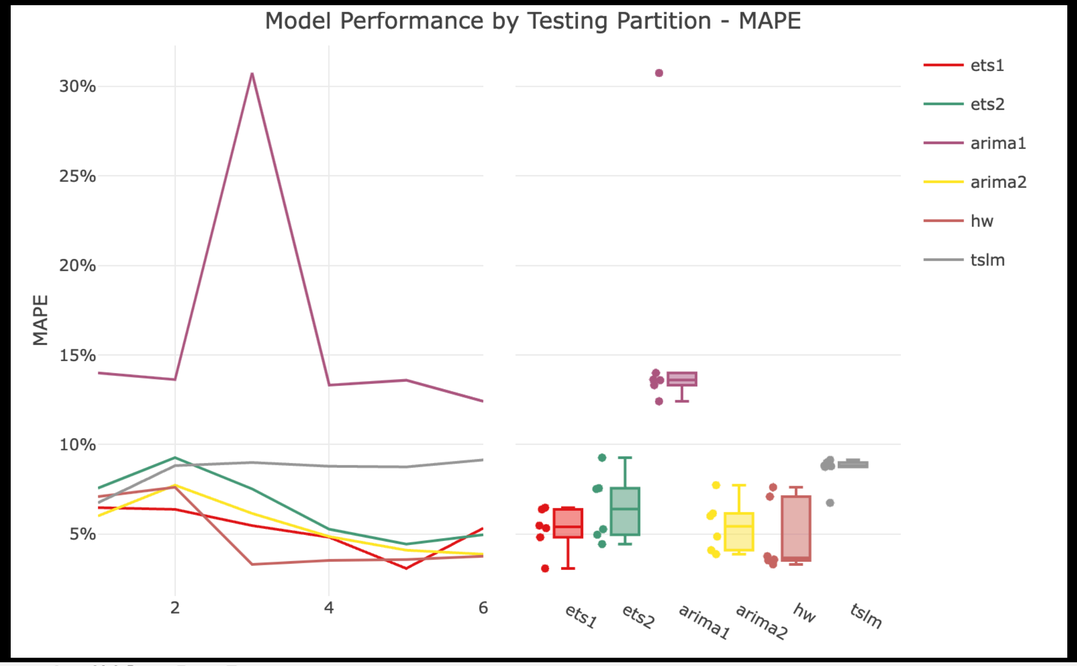
Horse Racing with Backtesting

Training



Horse Racing with Backtesting

Models Scoring and Ranking

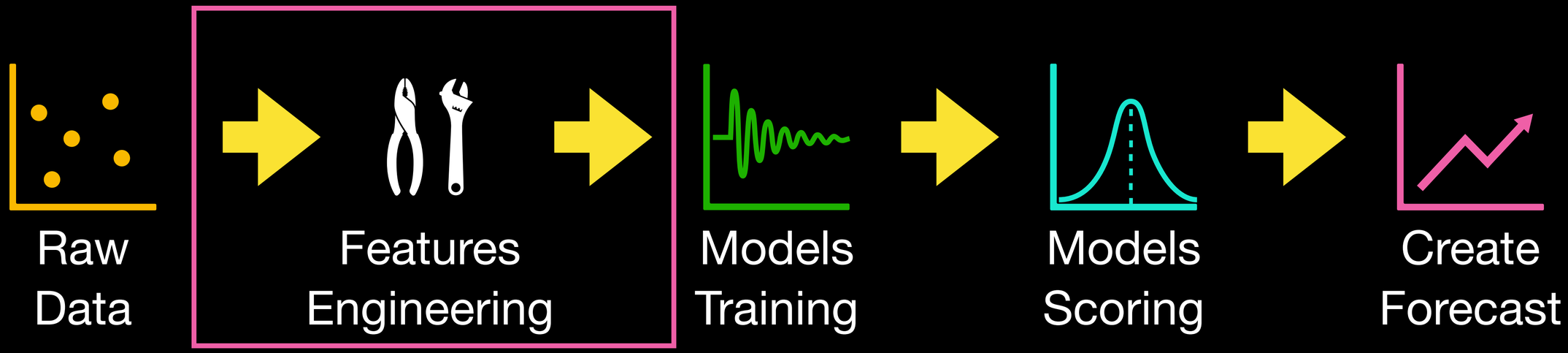



```
#> # A tibble: 5 x 7
#>   model_id model  notes      avg_mape avg_rmse `avg_coverage_9… `avg_coverage_9…
#>   <chr>    <chr>  <chr>         <dbl>    <dbl>            <dbl>            <dbl>
#> 1 hw       HoltW… HoltWinte…   0.0482     144.            0.875            0.931
#> 2 ets1     ets    ETS 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
```

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.

Source: Rob J Hyndman, Feature-based time series analysis, https://robjhyndman.com/hyndsight/fbtsa/

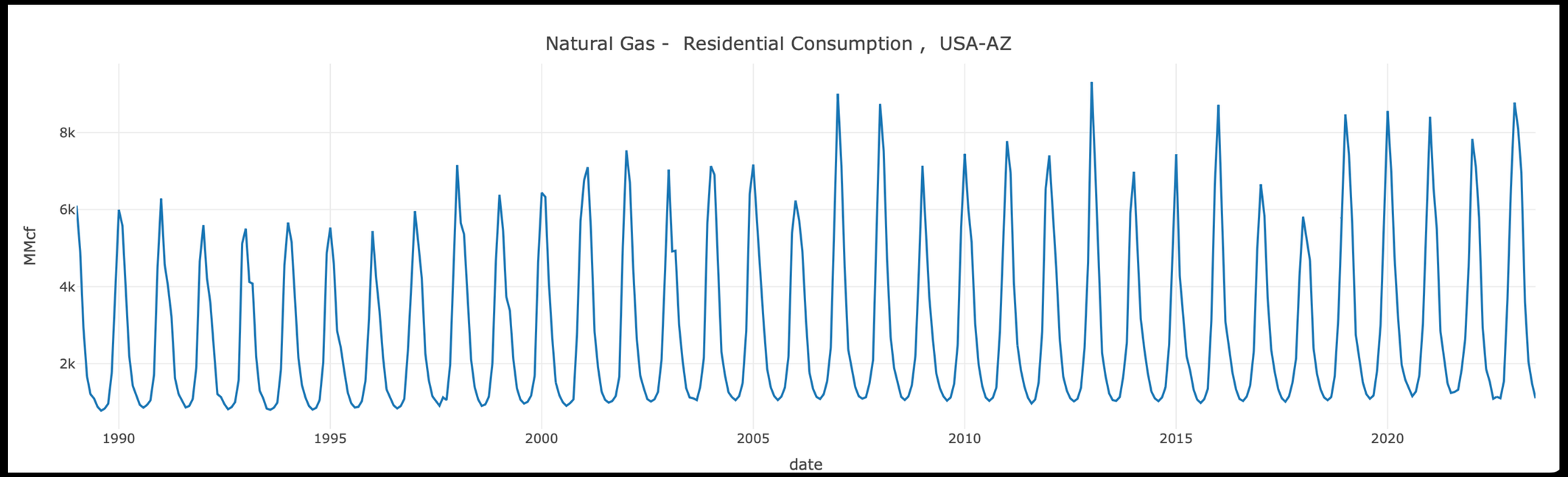
Cluster Analysis

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

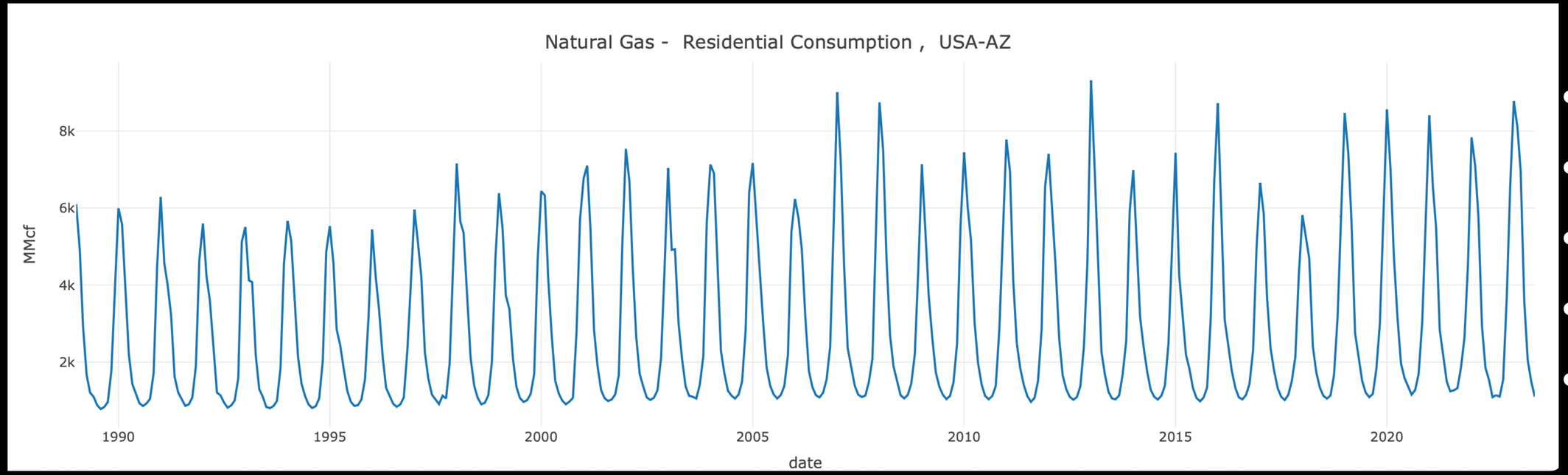
Feature-based Time Series Analysis

Time Series



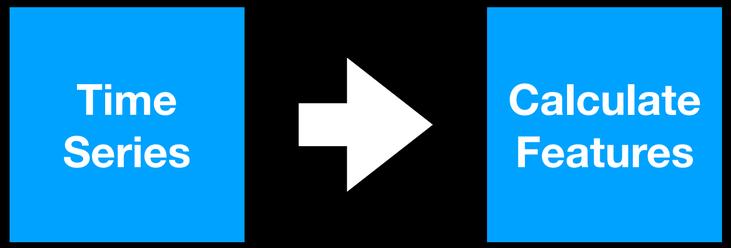
Feature-based Time Series Analysis

Time Series → **Calculate Features**



- Trend
- Seasonality
- Correlation
- Linearity
- Etc.

Feature-based Time Series Analysis

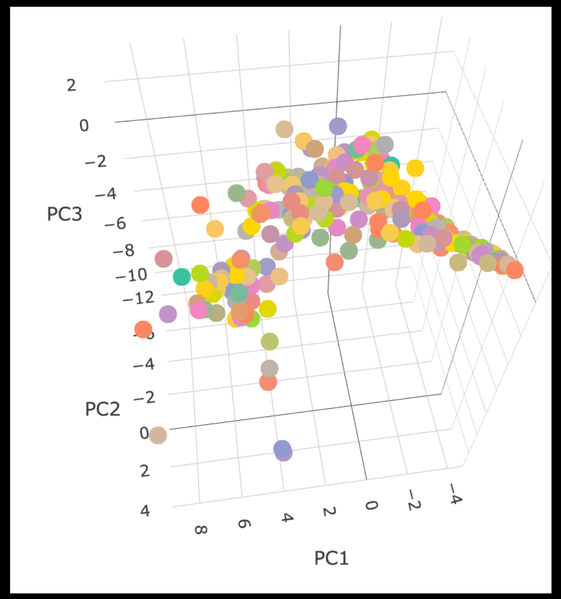


	trend	spike	linearity	curvature	e_acf1	e_acf10	entropy	x_acf1	x_acf10	diff1_acf1	diff1_acf10	diff2_acf1	diff2_acf10	arch_stat	embed2_incircle_1	embed2_incircle_2	ac_9	firstmin_ac	trev_num
1	-0.9899970	0.08520655	-1.183406	0.4899137	0.2654095	0.07020989	1.2657207	-1.744773	-0.9311406	-0.2875607	-0.8877382	-0.4798144	-0.42517492	-2.3968238	-0.1184621	-0.1404637	-0.9512467	-0.3619961	-0.07831742
2	1.5287094	-0.71898122	1.020666	-0.5438890	-2.8443028	-1.41022073	-1.7599324	1.587047	2.1003062	-1.9980183	-0.4293386	-1.7462309	2.44634510	0.6708338	-0.1184621	-0.1404637	1.6454182	2.4397117	-0.07829846
3	-0.5445202	0.74495879	-1.598277	-0.7016720	0.1331232	-0.57729071	1.2577619	-1.341148	-1.3087329	-0.0656362	-0.8982515	0.2356164	-0.65610120	-1.1324845	-0.1184621	-0.1404637	-0.8841765	-0.3619961	-0.07010398
4	-0.7918566	1.25295007	-1.512500	-0.6056961	-0.1025825	-0.72580295	0.7538760	-2.347966	-1.3904018	-0.1003487	-0.7297453	0.3232403	-0.02048141	-1.0394197	-0.1184621	-0.1404637	-1.1599089	-0.4329255	-0.06964442
5	0.4195154	-0.25756724	-1.216838	-1.8302126	-0.9065856	-0.90543203	1.2827370	-1.084730	-0.7186413	-1.3741011	-0.7631615	-1.1529199	1.27965529	-1.1551813	-0.1184621	-0.1404637	0.1989565	-0.3619961	-0.07835162
6	-1.0693536	-0.05563009	-1.041192	-0.3659868	0.7317786	0.73997355	0.6918166	-0.699501	-0.5053118	0.9388589	0.2423670	1.2366727	-0.97456956	-0.2943584	-0.1184621	-0.1404637	-1.1070054	-0.3619961	-0.07193407

	motiftwo_entro3	walker_propcross	nonlinearity	x_pacf5	diff1x_pacf5	diff2x_pacf5	area_name	process
1	1.0258784	0.07861881	-0.01608262	-1.3055130	-1.1980991	-0.1372165	CALIFORNIA	VCS
2	-2.1467869	-0.84211928	1.20489193	0.3075486	0.0253915	2.0971878	CALIFORNIA	VDV
3	1.1624061	0.34585743	-0.64158567	-1.6341821	-1.0236839	-0.4650336	CALIFORNIA	VEU
4	1.3838888	1.37888065	-0.58014154	-1.9294899	-0.7969625	-0.1657227	CALIFORNIA	VGT
5	1.0794332	0.81071788	-0.64167888	-1.7177817	-0.9460363	1.3830305	CALIFORNIA	VIN
6	0.3632215	-0.39297875	-0.22348312	0.4894870	0.3698136	-1.0249950	CALIFORNIA	VRS

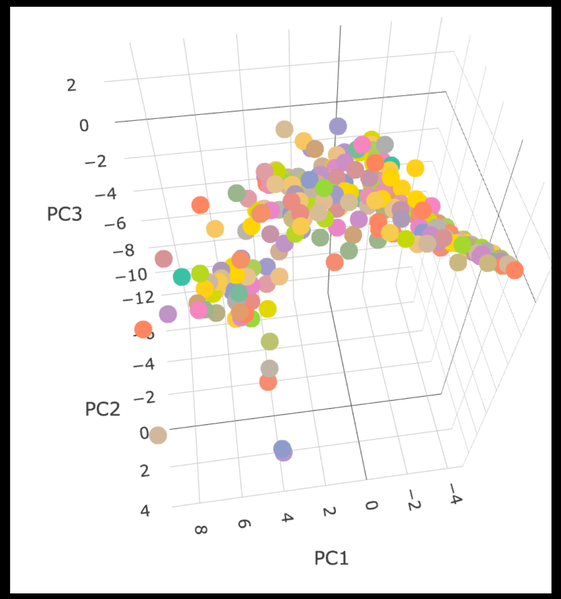
Feature-based Time Series Analysis

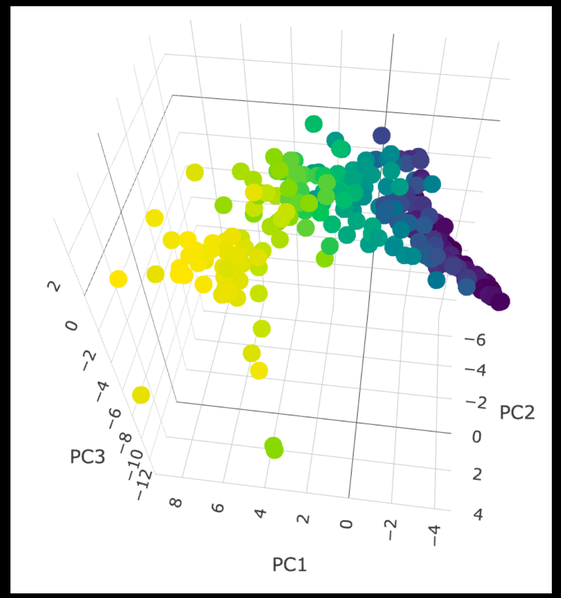
Time Series → Calculate Features → Dimension Reduction

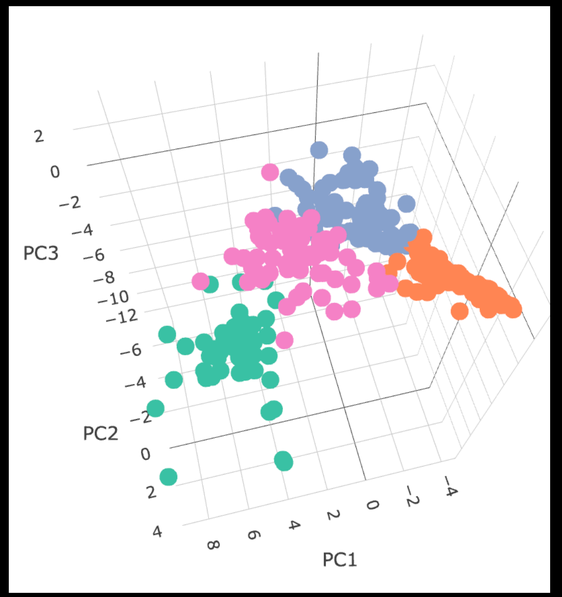


Feature-based Time Series Analysis

Time Series → Calculate Features → Dimension Reduction → Cluster

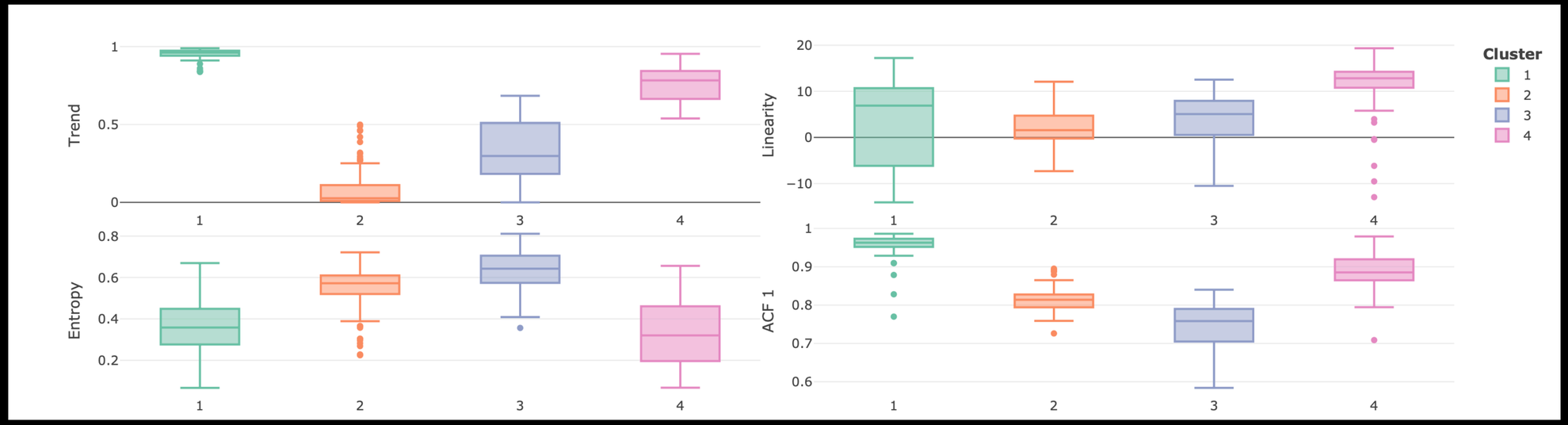






Feature-based Time Series Analysis

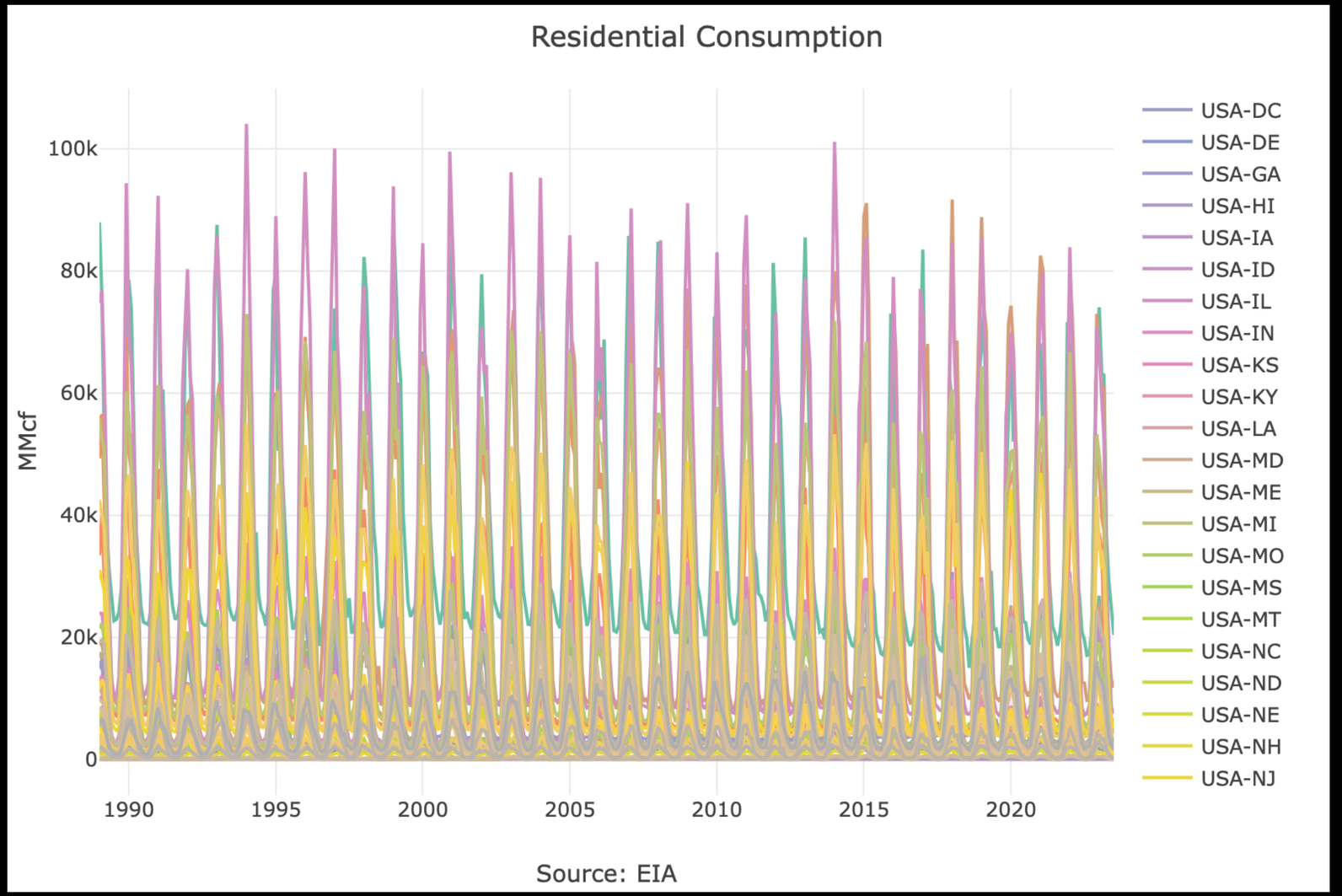
Time Series → Calculate Features → Dimension Reduction → Cluster → Common Patterns



Demo

Feature-based Time Series Analysis

- 314 series
- 5 categories
- Monthly



Feature-based Time Series Analysis

Tools

- **Data -** EIAapi, 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-of
- 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/