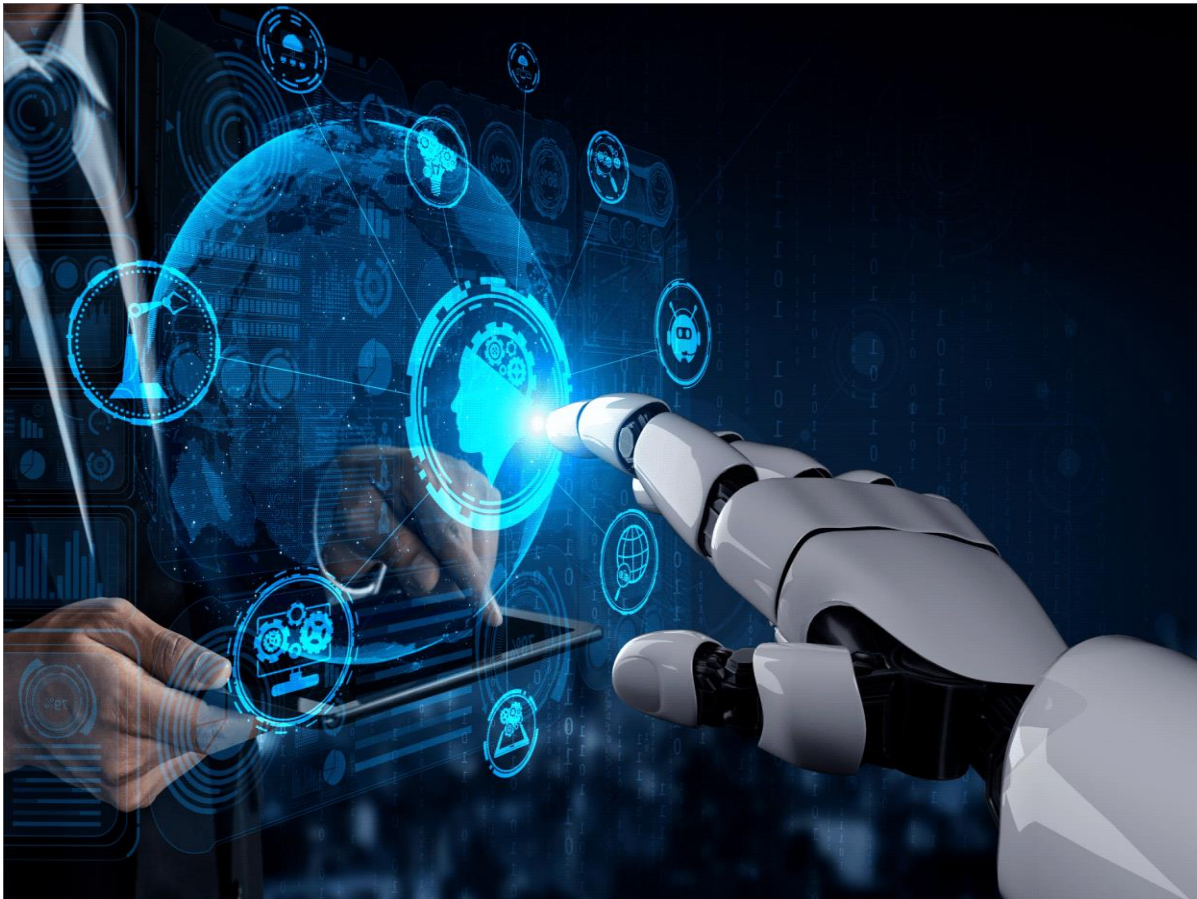# Forecasting 2.0 Accelerator

Topics

Why a new accelerator?
Who are the target personas?
What has been developed?
Profiling
Notebooks
Conceptual process flow
Functions
Improvements

Why a new forecasting accelerator?

- We focus on updating existing forecasting repository to generate our deliverables more quickly

- Completeness of forecasting algorithms is not practical:
  - [There exist many public algorithms for forecasting](#).
  - [Forecasting AutoML in Azure ML](#) is already published.

- **MCS opportunity** is to leverage the project exposure to a wide range of sectors to share the experience of industry specificities (finance, manufacturing, energy, etc).

- The **values** are:
  1. demonstrate delivery experience during customer talks
  2. acceleration of delivery through focused discussion and data requirements

Leverage ISD's experience in a wide range of sectors/industries to:
(1) **understand fundamental** and common **challenges**
(2) **standardize and scale** forecasting solutions

Why this accelerator might be useful for you
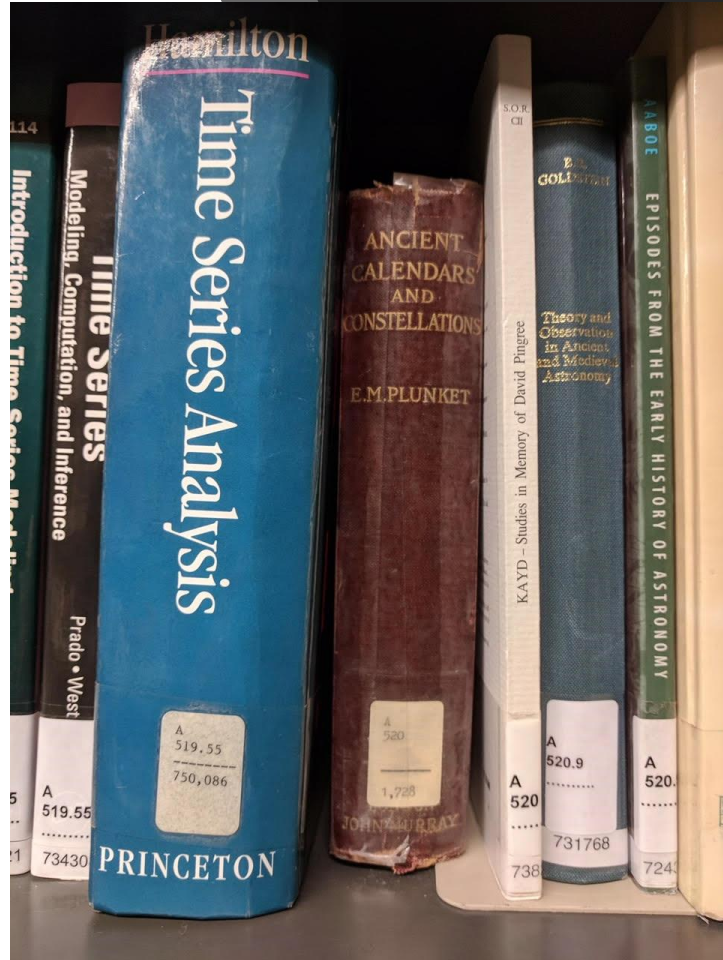
1. It provides you with guidelines Notebooks that can help you taking into account all necessary steps in order to perform a good data preparation, which is crucial in forecasting

2. It provides you with a bunch of useful functions you might need when dealing with demand forecasting

3. If you have several time series to forecast, thanks to the Profiling module it allows you to quickly understand how "difficult" to forecast are the time series you are dealing with by classifying time series as intermittent or regular.

Who are the target personas?

- Data scientist / Consultants

- Data & AI practitioners searching for industry specifics use-cases

When to use

1. When dealing with 1 or many time series, you may want to quickly characterize time-series the data (continuous 0 values, high variances, etc.) We provide some indicators to seize such characteristics and assign best approach in generating forecasting models.

2. When you want to quickly gain insights of a specific sector

3. When you want to explore an alternative approach to Azure AutoML which tries to generate models by brute-force manners.

What has been developed?

- The goal is to share industry knowledge: use-cases, challenges, data requirements

- Create code template to analyze time series data applicable to any industry and provide recommendations on best approach

- Add-ons for specific challenges

- The **forecasting 2.0** contains the **following parts:**

  1. Data schema (how to organize data ingestion flow) and data pre-processing scripts
     - Target: Architects / Consultants

  2. **Time-series profiling scripts** (understand type of series and recommend models)
     - Target: Data scientists

  3. Documentation on industries specificities/requirements

     - Target: Digital Advisor, Consultants, Data scientists, …

  4. **Add-ons** for specific scenarios (many-models, forecasting in finance, sol-acc, etc)

Time-series profiling scripts

- **Goal:** identify consumption patterns that are similar to each other in order to assign the optimal model in terms of min KPI (MAE or MSE)

- **How to:** identify those series that are classified as "intermittent" with respect to those "regular"

- **Expected output:** label each time series as intermittent with respect to regular
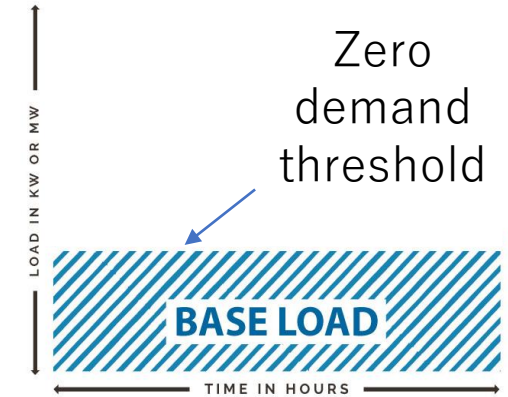
Intermittent time series
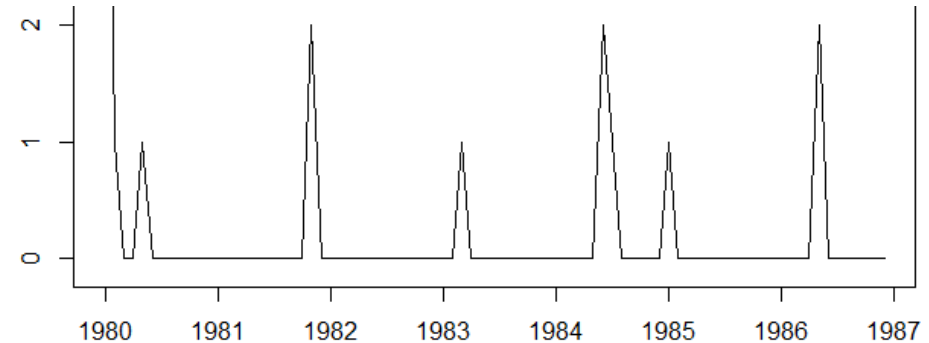
Regular time series (shows seasonality)
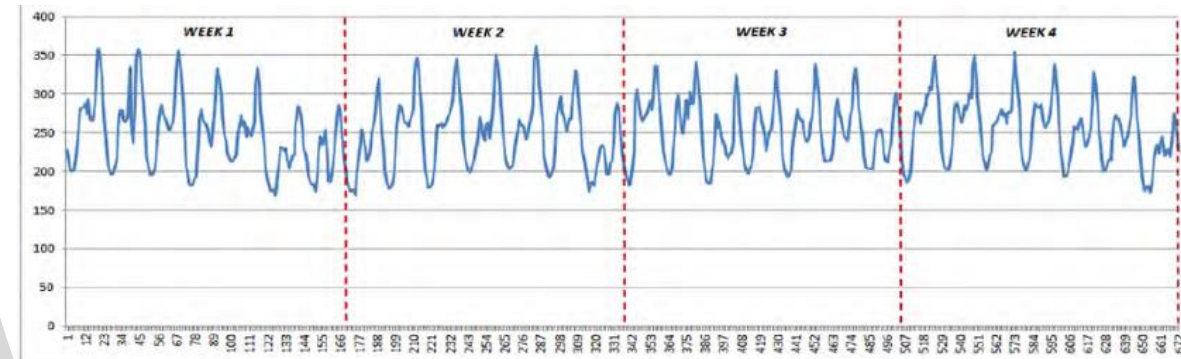
Identifying intermittent time series

- Intermittent time series or demand comes about when a product or a time series experiences several periods of zero demand. Often in these situation, when demand occurs it is small, and sometimes highly variable in size

- To identify intermittent time series, compute the following indicators:
  - <u>Zero demand threshold</u>, this parameter identifies what is considered to be constant at zero demand (e.g. baseload in case of energy) and constant demand
  - <u>Average Demand Interval (ADI)</u>, this parameter is period based which is calculated as average interval time between two demand occurrences
  - <u>SD Demand Interval (SDDI)</u>, this parameter is period based which is calculated as the standard deviation of interval time between two demand occurrences
  - <u>Coefficient of Variation Squared (CV2)</u>, is the ratio of the standard deviation to the mean and shows the extent of variability when demand occurs. The higher the CV, the greater the dispersion. The squared coefficient of variation represents variability of demand size.

Zero demand threshold

BASE LOAD

LOAD IN KW OR MW
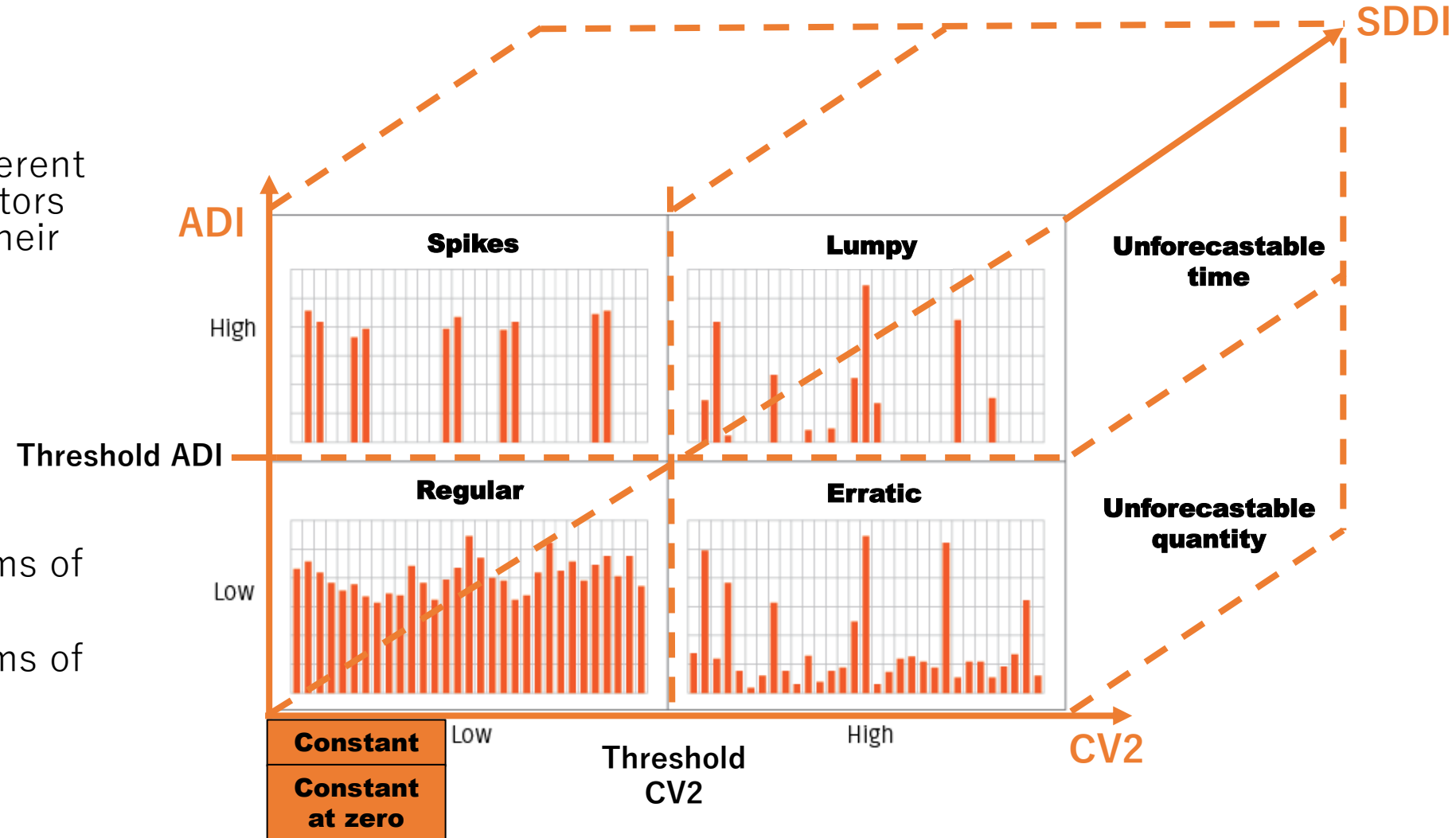
TIME IN HOURS

$$ADI = \frac{\sum_{i=1}^{n} t_i}{N}$$

$$SDDI = \frac{\sqrt{\sum_{i=1}^{n}(t_i - t)^2}}{N}$$

$$CV^2 = \left[\frac{\frac{\sqrt{\sum_{i=1}^{n}(\varepsilon_i - \varepsilon)^2}}{N}}{\varepsilon}\right]^2$$

Intermittent indicators parameters

The combination of different thresholds of the indicators defines time series by their profile:

- spikes
- lumpy
- erratic
- unforecastable in terms of time volatility
- unforecastable in terms of quantity volatility
- constant
- constant at zero
- regular time series

Methods to forecast _intermittent_ time series

- **Croston's method**
  - In 1972, J.D. Croston published "Forecasting and Stock Control for Intermittent Demands," an article introducing a new technique to forecast products with intermittent demand
- **References**
  - [Lancaster Centre For Marketing Analytics and Forecasting](#)
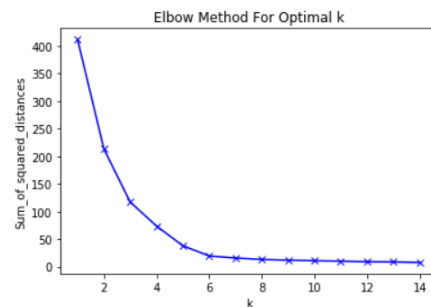 - [Methods for Intermittent Demand Forecasting](#)

Intermittent demand item

Clustering regular time series

fig 2: After applying K-means clustering

Clustering profiles

- Clustering regular time series using K-Means flat or hierarchical

Choose the optimal number of clusters

- As a method to choose the optimal number of cluster, use max explained variance at the minimum number of cluster -> _Elbow Method_

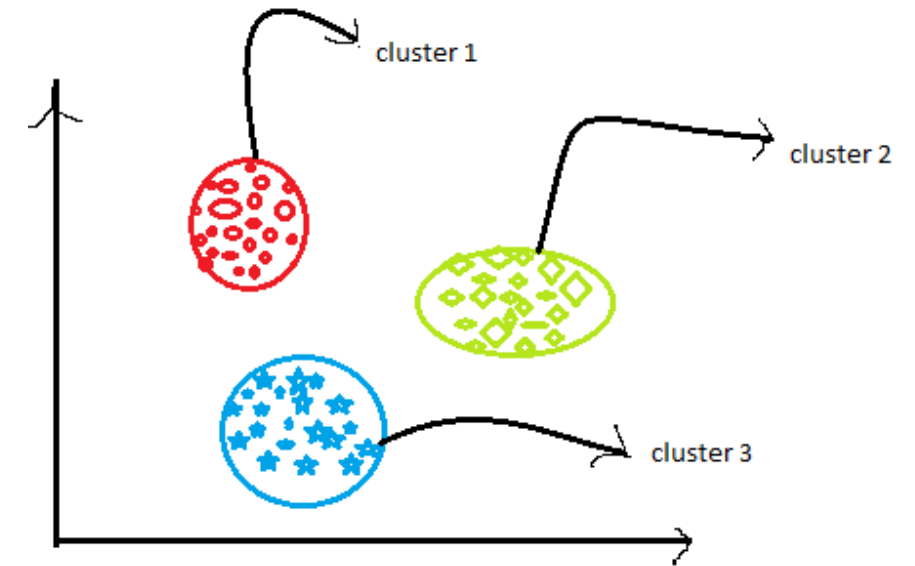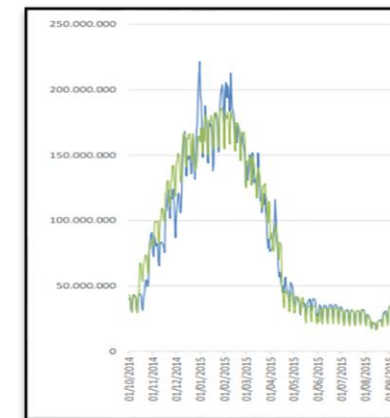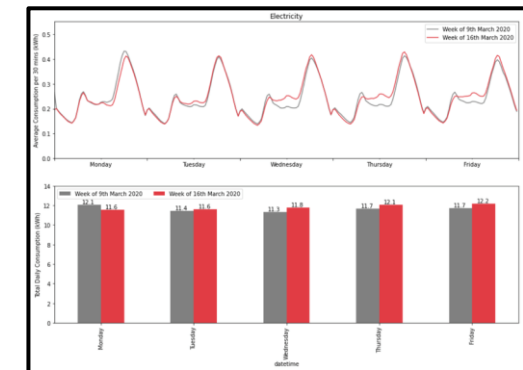- Check whether identified profiles have a business meaning

**Cluster 1: thermal consumption, use temperatures**

Cluster 2: industrial consumption, use calendar variables

Methods to forecast _regular_ time series

Treat regular time series with "traditional" machine learning methods, some have been implemented in the accelerator, for additional reference check the **GitHub Repo**

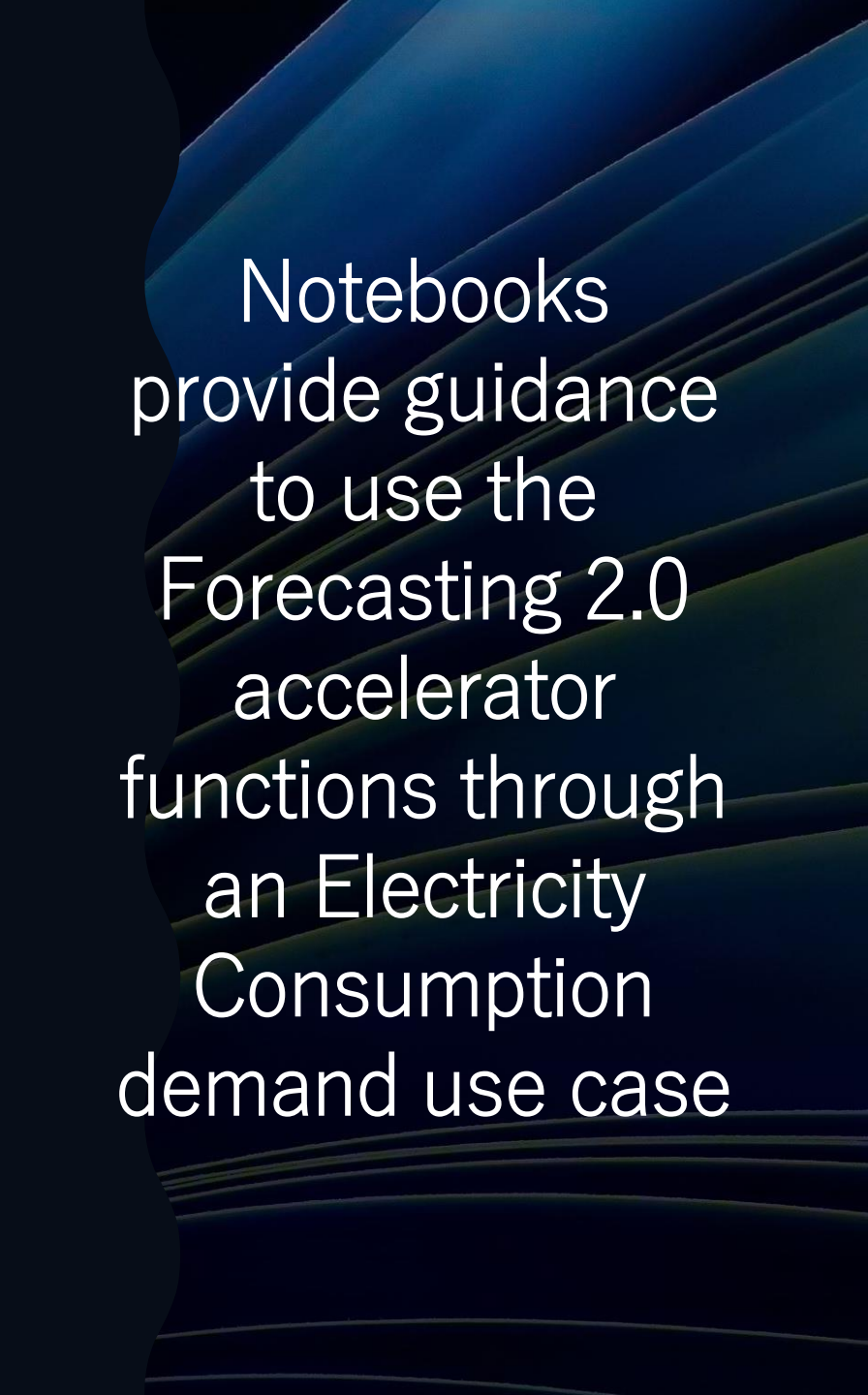| Model	Library	Status
Linear regression	statsmodel	Implemented in Forecasting 2.0
Gradient boosting	xgboost	Implemented in Forecasting 2.0
Random forest	statsmodel	Implemented in Forecasting 2.0
Prophet	Prophet	Not yet implemented
Neural networks	Neural prophet	Not yet implemented
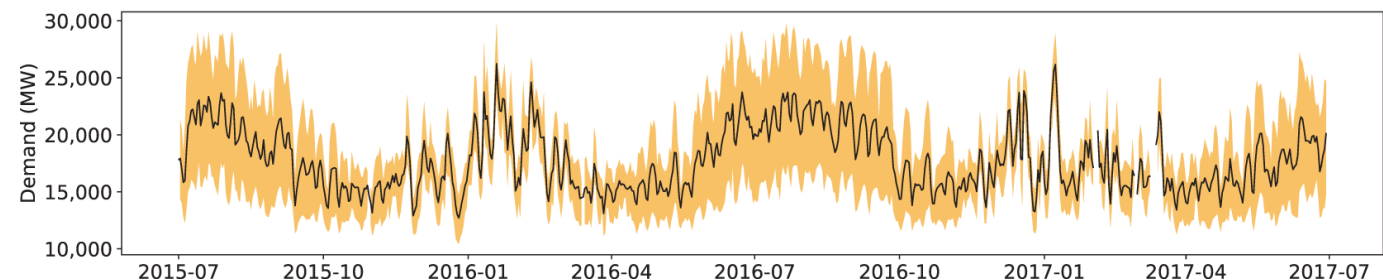
Notebooks provide guidance to use the Forecasting 2.0 accelerator functions through an Electricity Consumption demand use case

This accelerator deals with panel data. In statistics and econometrics, panel data or longitudinal data is data that contains observations about different cross sections (groups or ids) across time. Examples of groups that may make up panel data series include countries, firms, individuals, or demographic groups.

Two Groups from a Panel
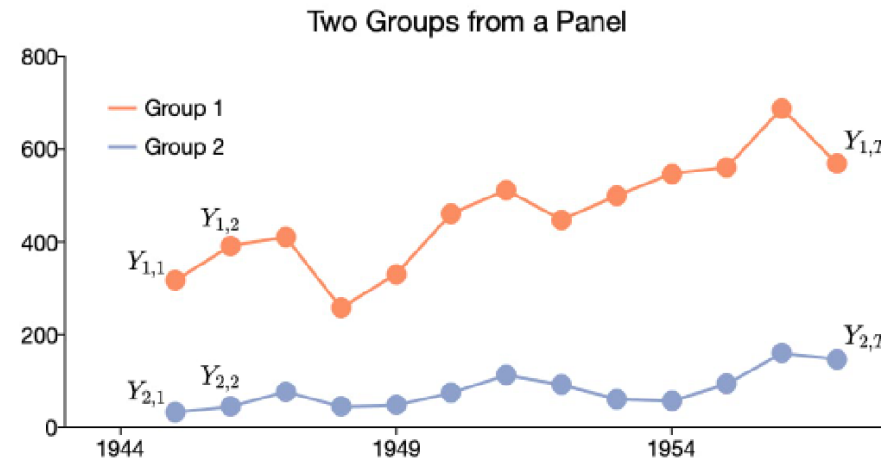
Group or Id	Time period	Notation
1	1	Y_{11}
1	2	Y_{12}
1	T	Y_{1T}
$\vdots$	$\vdots$	$\vdots$
N	1	Y_{N1}
N	2	Y_{N2}
N	T	Y_{NT}

Notebooks provide guidance to use the Forecasting 2.0 accelerator functions through an Electricity Consumption demand use case

EnergyDataExploration

- A notebook that provides an exploratory data analysis in order to understand the type of time series you are dealing with

EnergyPredictionDataPreparation

- A notebook that helps with Time Series Data Preparation, in particular how to deal with NAs, how to aggregate time series and how to add create useful regressors (e.g. calendar variables)

EnergyProfilingIntermittent

- A notebook that profiles time series by classify them among regular, intermittent, lumpy, erratic, unforecastable in terms of time, unforecastable in terms of quantity, constant and constant at zero

EnergyClusteringRegular

- A notebook that performs a k-means flat cluster analysis on those time series that were classified as regular

EnergyPredictionScoring

- A notebook that helps you produce a forecast, plotting the results and compute KPIs on a panel dataframe, where you have multiple timeseries identified by a given group or id (e.g. multiple sensors time series, multiple plants or site-id energy consumption, etc)

Conceptual process flow in generating forecasting models
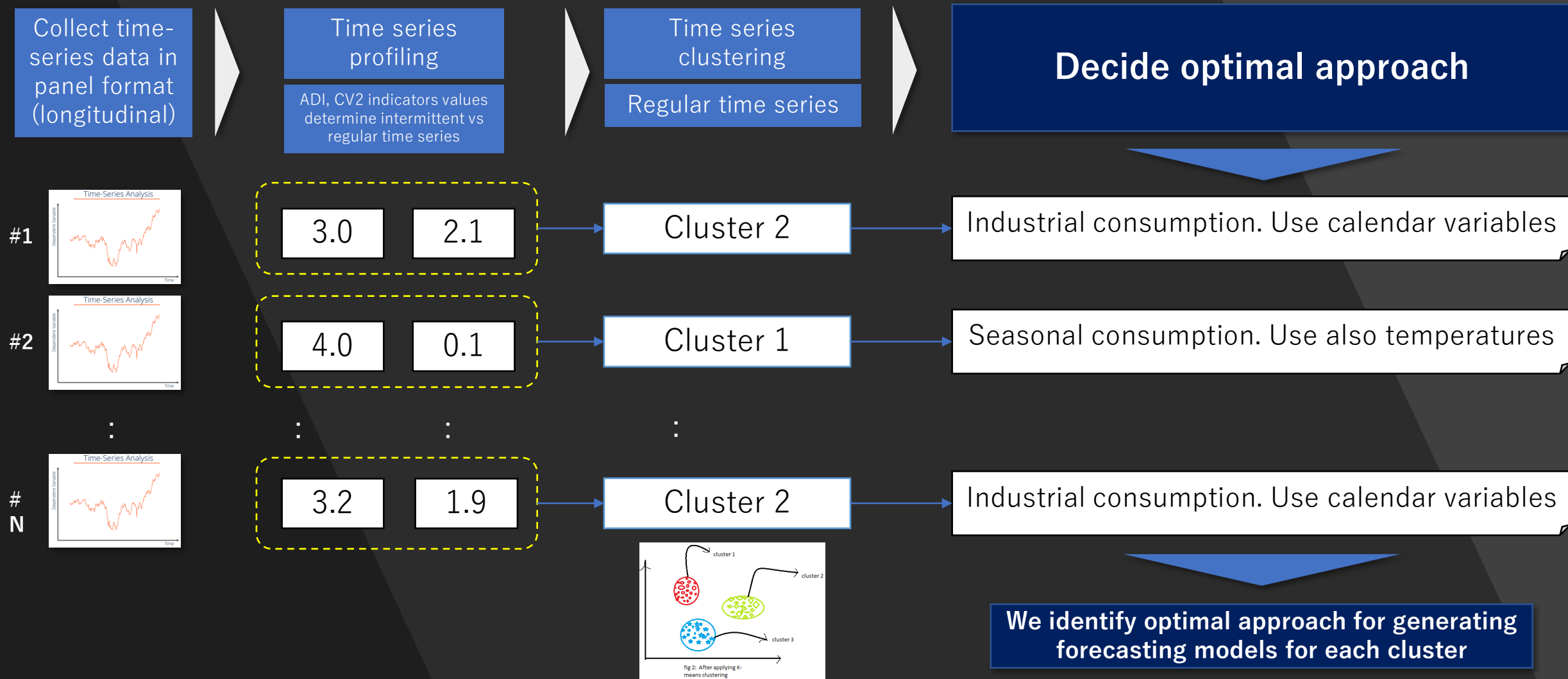
Useful functions

Cool sliding plots to visualize series and their forecasts

Adding **holidays** by country or other regressors such as months, weekdays and interaction terms

Creating normal **temperature future scenarios** to generate years-ahead forecasts

Filling missing data using similar days or similar weeks values

Compute errors like mean absolute error and mean absolute percentage error

Wrap up results in tidy Excel or csv files

Time series slider
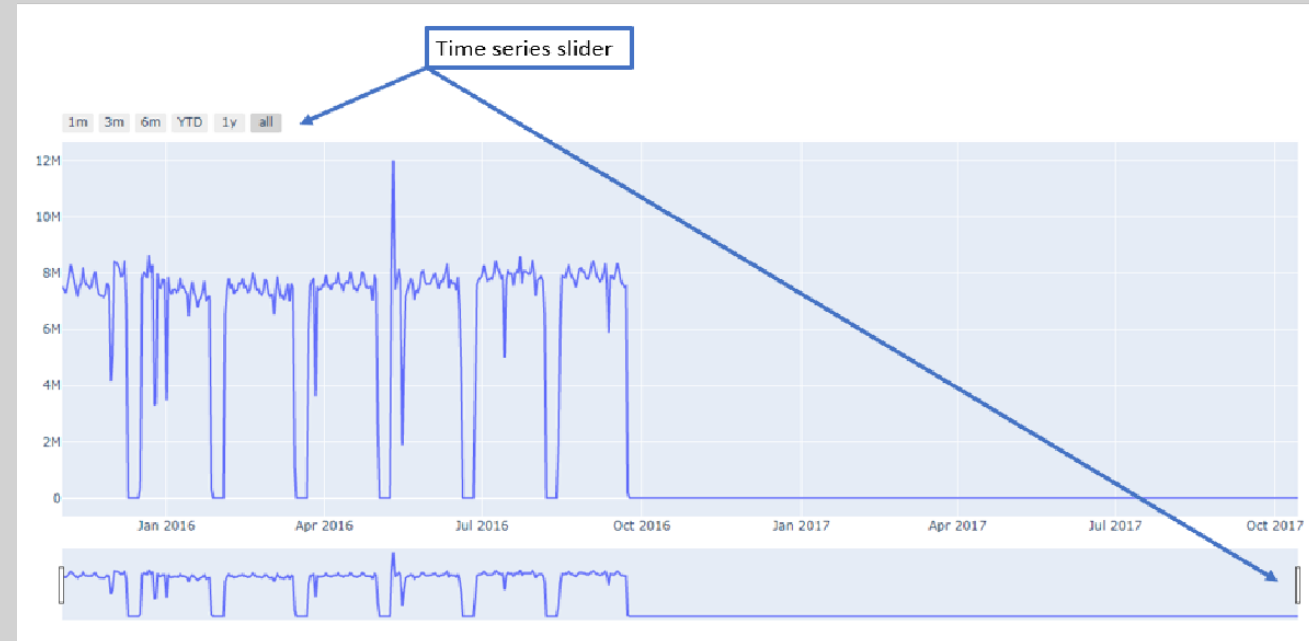
As data scientist, how can I contribute?

How to contribute to profiling?

- What needs to be done is to test and define intermittent indicators for other types of data than electricity consumption

How to contribute to data preparation and scoring?

- What needs to be done is the improve the code to make it scalable and more efficient when working with big datasets (e.g. more than 100 id)

For more details, go to
[Forecasting 2.0 GitHub Repo](#)