Knowledge discovery process: Self-tuning strategies for a deeper understanding of real industrial data

Tania CERQUITELLI
Department of Control and Computer engineering, Politecnico di Torino, Italy

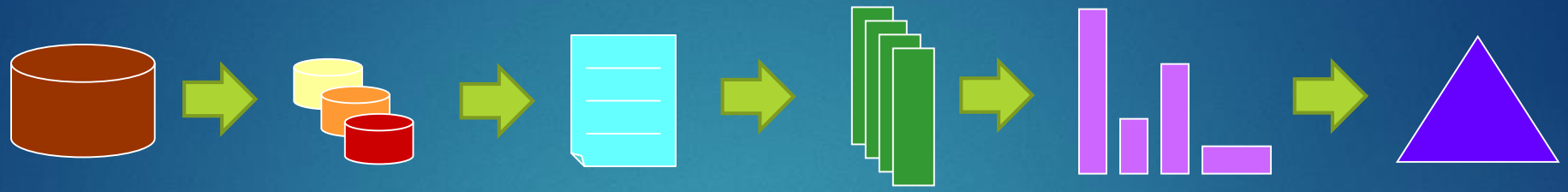
Outline

- Knowledge discovery process in the context of predictive maintenance in Industry 4.0
- Self-tuning methodologies for predictive maintenance
  - Two real use cases
- Semi-supervised data labelling
- Concept drift management
- Technological aspects
- Open issues

Knowledge discovery process in the context of predictive maintenance in Industry 4.0

Data Base and Data Mining Group of Politecnico di Torino

Knowledge extraction process



Data

Selection

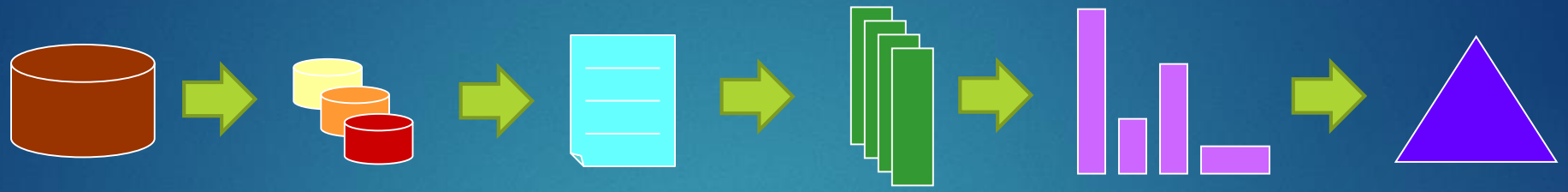
Preprocessing

Transformation

Knowledge extraction

Visualization Interpretation

Knowledge extraction process



Data

Selection

Preprocessing

Transformation

Knowledge extraction

Visualization Interpretation

Knowledge

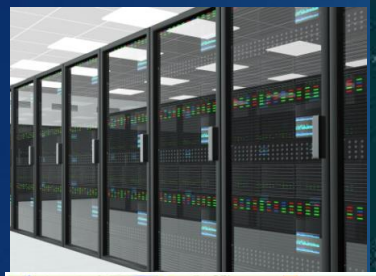
Value

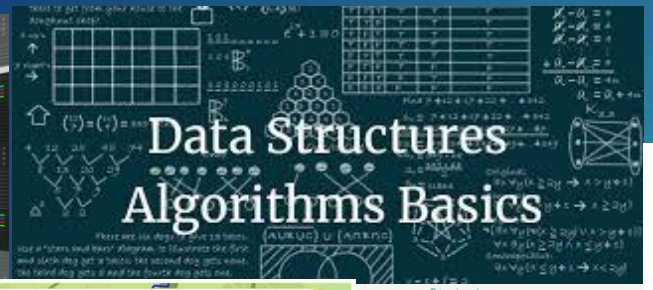
KDD

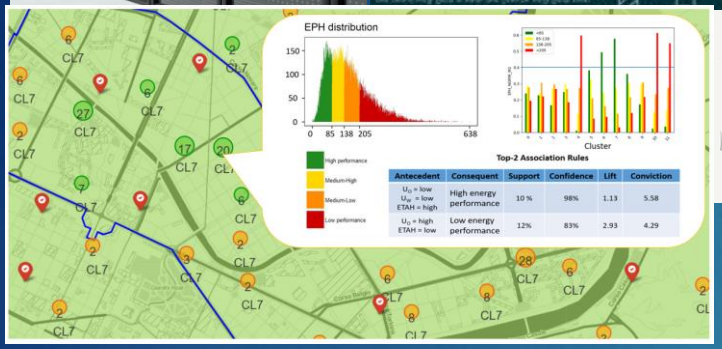
Knowledge Discovery from Data

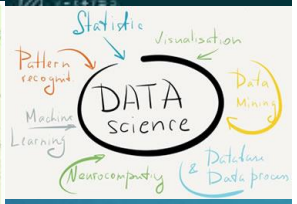
Data Base and Data Mining Group of Politecnico di Torino

KDD from Industrial data: two key roles









DATA SCIENTIST

- Design **innovative and efficient algorithms**
- Select the **optimal techniques** to address the challenges of the analysis
- Identify the best **trade-off** between knowledge quality and execution time

DOMAIN EXPERT

- Support **the data pre-processing** phase
- **Assess** extracted **knowledge**
- Strong involvement in the algorithm definition phase which should **respect/include physical laws** and correctly **model physical events**

Data Base and Data Mining Group of Politecnico di Torino

Predictive maintenance in Industry 4.0

- **Intelligent techniques** to identify **symptoms of imminent machine failure** before their actual occurrence
  - It combines **physical models** of complex devices (machines, robots, conveyors, etc.) together with **data-driven algorithms** to effectively support smart predictive diagnostics (prognostics)
- It anticipates failures and estimates the **Remaining Useful Lifetime** (RUL)
  - It exploits innovative **analytics methods** to forecast the **future evolution of machine degradation**
- **On-line data** collected in the factory characterize the current dynamics of the process/machine from the

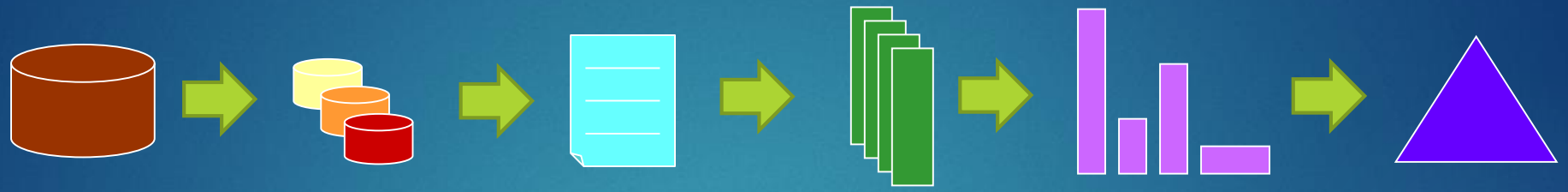
- Some of the **most common needs of manufacturing enterprises**
  - compatibility with both the on-premises and the in-the-cloud environments
  - exploitation of reliable and largely supported Big Data platforms
  - easy deployment through containerized software modules
  - virtually unlimited horizontal scalability
  - fault-tolerant self-reconfiguration

Data Base and Data Mining Group of Politecnico di Torino

Main issues

- **Tailoring the KDD process to predictive maintenance requires a lot of expertise**
- **Identifying the best data preprocessing approach is very challenging**
  - Summarize long time-dependent series through ad-doc statistics capturing the main key features for the prediction
  - Use-case dependent
- **A variety of state-of-the-art algorithms is available**
  - Data driven methodologies possibly enriched with physical models
- **Each algorithm is characterized by many different input-parameters**

Knowledge extraction process: Innovations

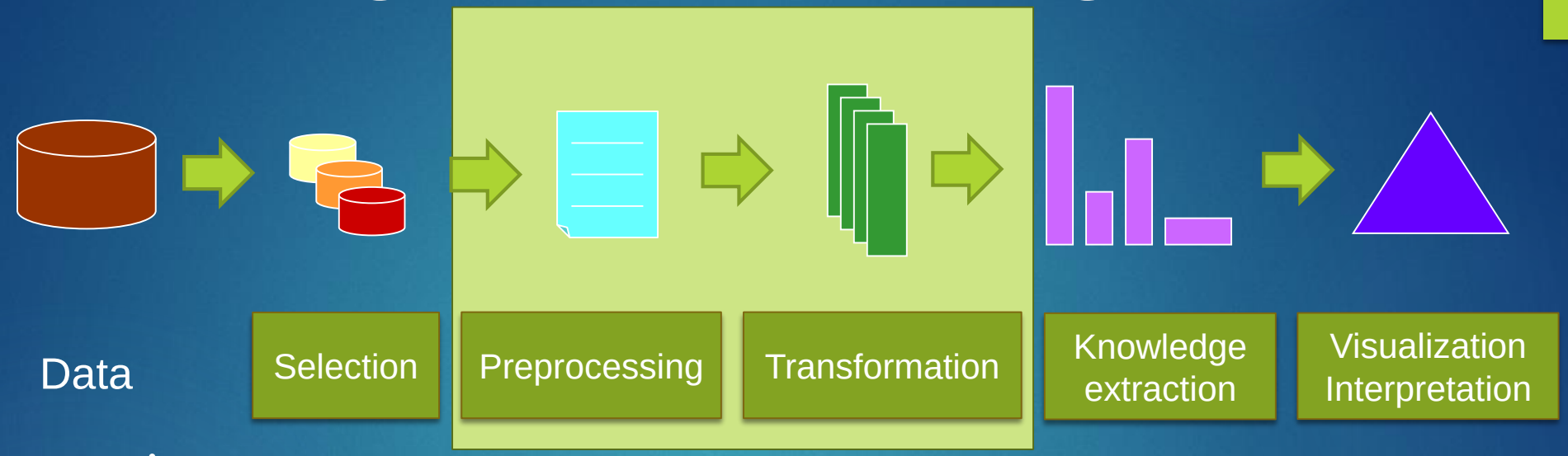


Data → Selection → Preprocessing → Transformation → Knowledge extraction → Visualization Interpretation

Innovations in the data analytics process

- **Tailor** the **analytic** steps to the different key aspects of **industrial data**
- **Design ad-hoc data transformation strategies to capture different facets of data**
- **Self-tuning strategies to offload the data scientist from algorithm configuration**
- Design **informative dashboards** to support the translation of the extracted knowledge into effective actions

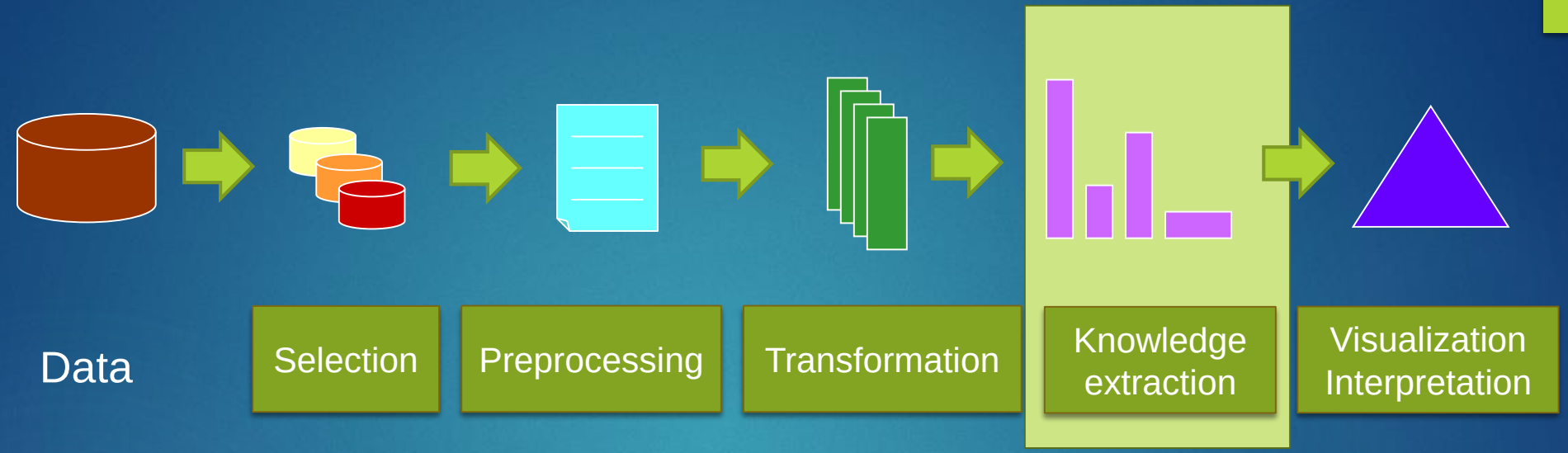
Knowledge extraction algorithms



- **In the preprocessing step**
  - **A time series alignment technique might be required**
    - **e.g., padding technique**
  - **In case of multivariate problem, correlated time series should be identified and removed**
    - **Correlation-based approach**
    - **Domain-driven knowledge**
    - **Mixed approach**
- **Transformation**
  - **Feature engineering**

Data Base and Data Mining Group of Politecnico di Torino

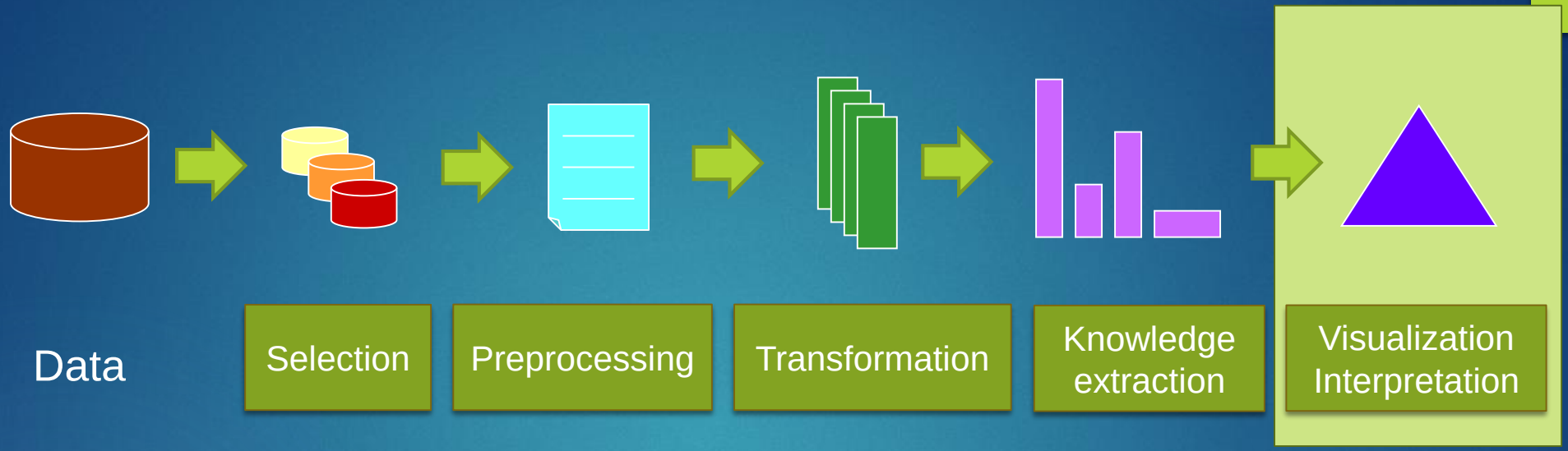
Knowledge extraction algorithms



- **Predictive algorithms**
  - **Classification task to classify production cycles**
  - **Regression task to estimate RUL (Remain Useful Life)**
- **Clustering algorithms**
  - **Data labeling to support predictive analytics**
- **Concept drift detection**
- **Anomaly detection techniques**

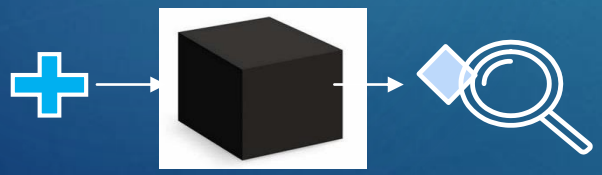
DBMG
Data Base and Data Mining Group of Politecnico di Torino

Knowledge visualization and interpretation



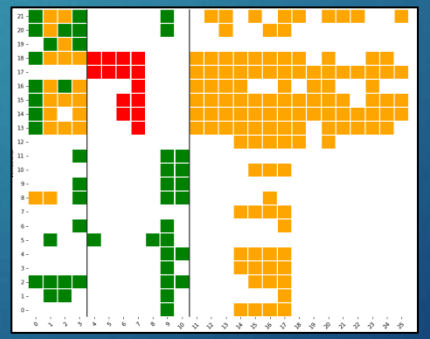
Interpretation of models and results

- Identification of algorithms and methodologies for semantic transparency of Machine Learning models
- Explanation methods for explaining individual predictions of black box models



Informative dashboard

- Visualization methodologies



Data Base and Data Mining Group of Politecnico di Torino

Self-tuning methodologies for predictive maintenance

Predictive maintenance

- **The identification of machine failures before their actual occurrence.**
  - It may combines **physical models** of complex devices together with **innovative analytics methods**
  - to forecast the **future evolution of machine degradation** from current on-line data collected in factories
- **Two real industrial cases**
  - Degradation over time of the belt tensioning in a robot (Robotics use case)
  - Alarm prediction in slowly-degrading multi-cycle industrial processes (White goods use case)

Versatile plug-and-play platform enabling remote predictive maintenance

Robotics use case: Belt tensioning

- The tensioning of the belt is necessary to assure the correct functioning of the robot
  - Low tension causes slippage, overheating and premature wear of the belt and pulley.
  - Too much tension leads to excessive strain on belts, bearings and shafts.
- There is a wide range of tensions that guaranties a good functioning
- The loss of tension occurs on all the belts
  - Loss of tension goes from 50% to 70% wrt the original tension.
- Use case
  - The tensioning of the belt is measured by the number of washers used to tension it.
  - The effect of different belt tensions can be extracted from the current (Ampere)
  - KDD objective: to predict the number of washers cycle by cycle.

White goods use case: a foaming process

- Foaming process, a slowly-degrading industrial process
  - During each production cycle, a nozzle is used to inject an isolating foam composed of two reacting chemicals
  - Several sensors measure different properties of the foaming process
    - Temperature of the chemicals involved
    - Pressure of the liquids before the injection
    - Injection timing and quantity, ratio of the injected chemicals, etc.
- Production data gathered from industrial foaming machines
  - KDD Objective: to monitor and predict the degradation of the equipment, and so promptly trigger the maintenance interventions.
  - Degradation of the equipment measured in terms of alarm conditions
  - A high number of alarms would bring to machine faults and production interruptions.

A self-tuning pipeline for Pred. Maintenance

Data

Selection

Preprocessing

Transformation

Knowledge extraction

Visualization Interpretation

Preprocessing

- Outlier and noise detection
- Signal alignment

Smart Data computation

- **Time domain feature computation**
- **Frequency domain feature computation**
- Feature selection

Data aggregation

- **Time Domain**
- **Frequency Domain**

Predictive analytics

- **Self-tuning Model building**
- Predictive model
- Real time prediction

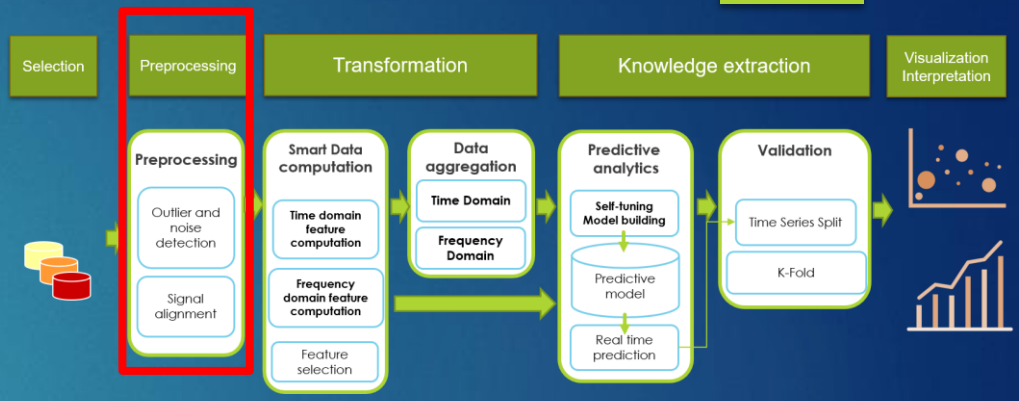
Validation

- Time Series Split
- K-Fold

Data Base and Data Mining Group of Politecnico di Torino

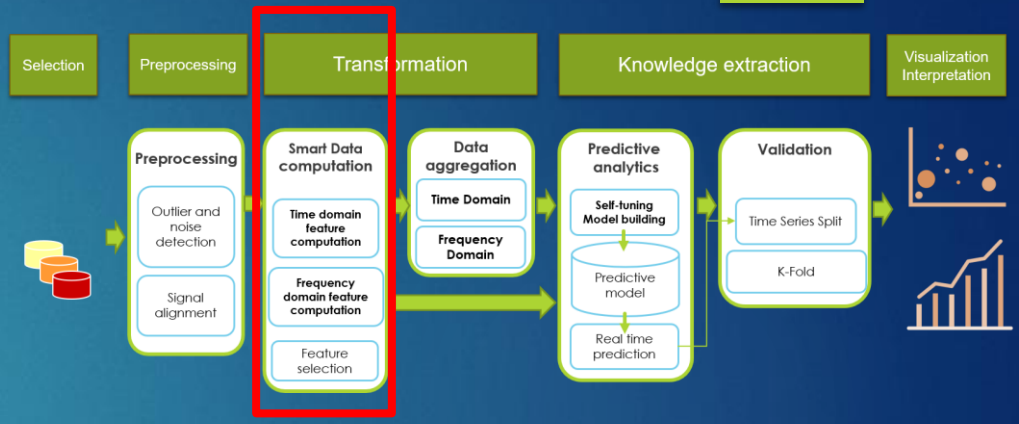
Data preprocessing

- **Outlier and noise detection**
  - Outliers: extreme observations which deviates too much from other observations (e.g., the overall pattern of the data collection).
  - Noise: statistical noise is unexplained variability within a data sample. In this context, noise identifies unwanted background values that affect the signals and generally lower the quality of the data.
- **Signal alignment**
  - Adjusting data observations such that all the points in the collection have the same shape may be required by the algorithms exploited in the analytics process.



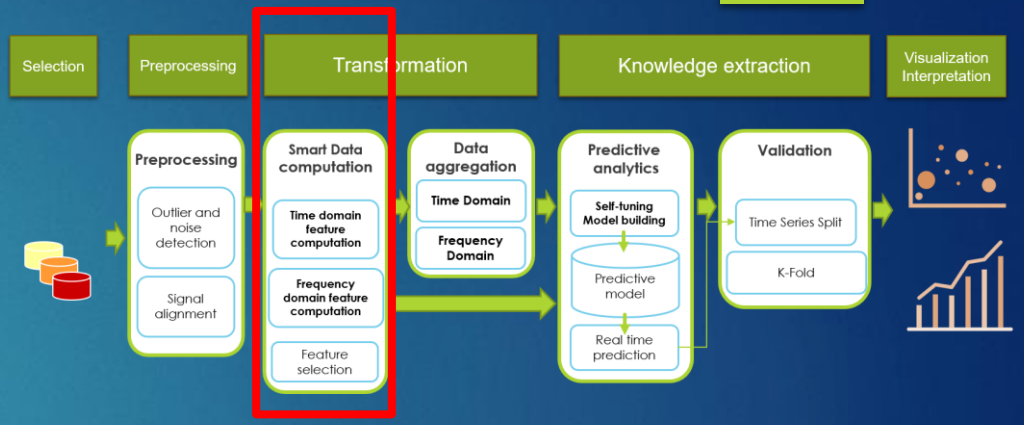
Smart data computation

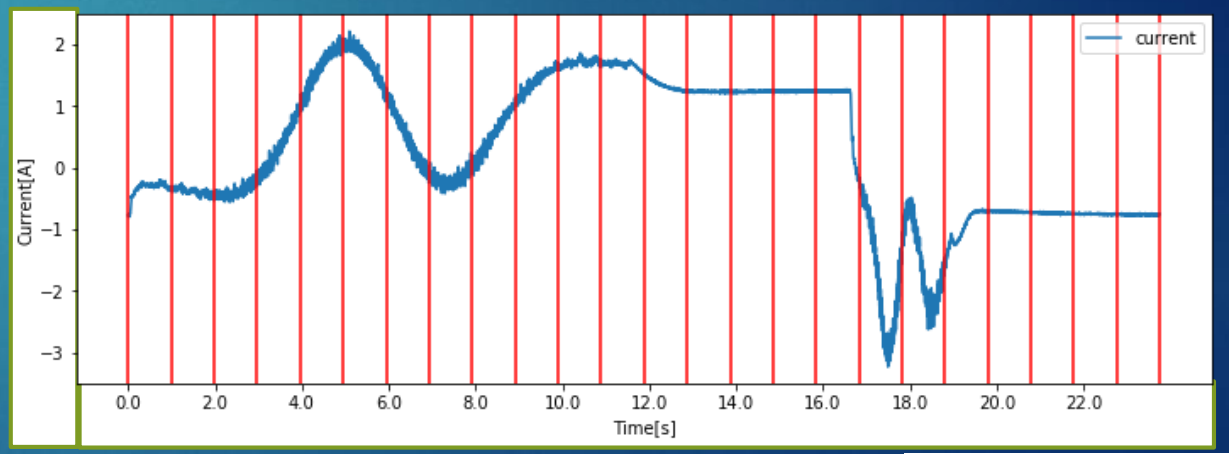
- The core component of the data analytics workflow
- Computation of the **most relevant features** that well describe and represent the datasets under analysis
  - a data-driven methodology to extract key features has been proposed
  - Domain knowledge can also guide the features extraction process.
- Three main phases:
  - Features computation
    - Time domain feature extraction
    - Frequency domain feature extraction
  - Features selection



Smart data computation

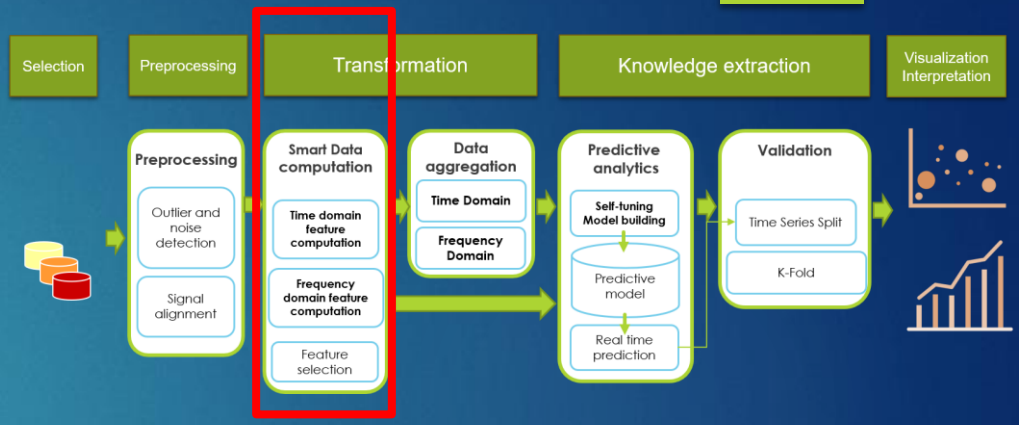
- **Time domain feature** computation
  - Signals are divided in **splits**, to better capture the signals variability
  - Features capturing the signal characteristics and variability in the time domain are computed in each split separately.
    - Mean, Standard deviation,
    - Minimum and Maximum values,
    - Kurtosis and Skewness,
    - RMS, sum of absolute values, # of elements over the mean, absolute energy, mean absolute change, etc.





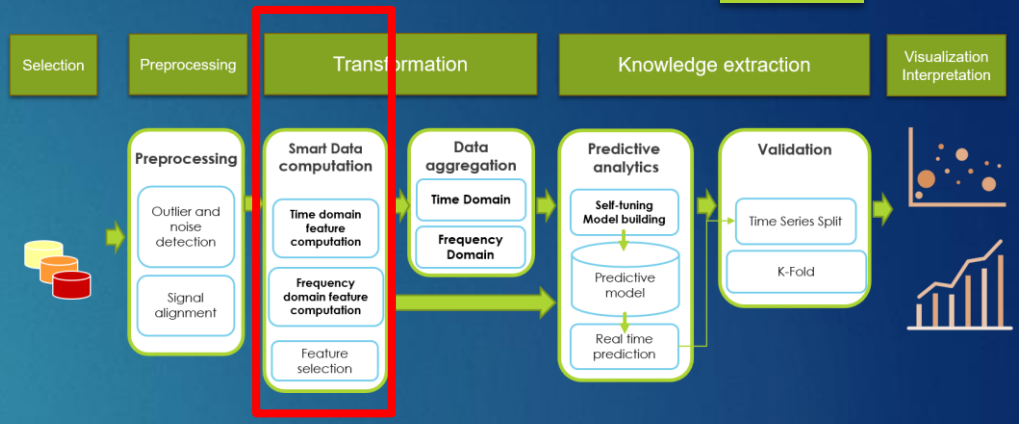
Smart data computation

- **Frequency domain feature** computation (**work-in progress**)
  - The time series are described as a sum of sinusoidal components (harmonics), e.g., using the Fourier Transform.
  - The most significant frequencies are kept into account for the analysis.
- The computed features can be directly used to train/test a predictive model, or a further transformation step may be required.



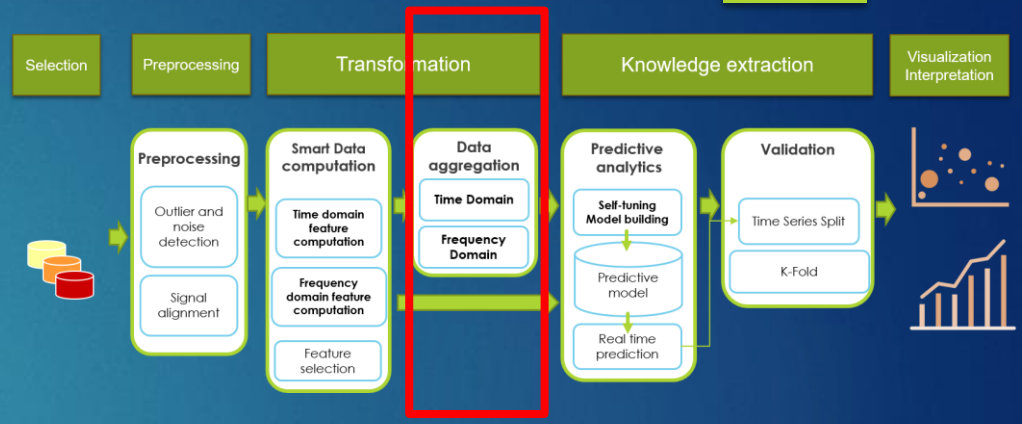
Smart data computation: Feature selection

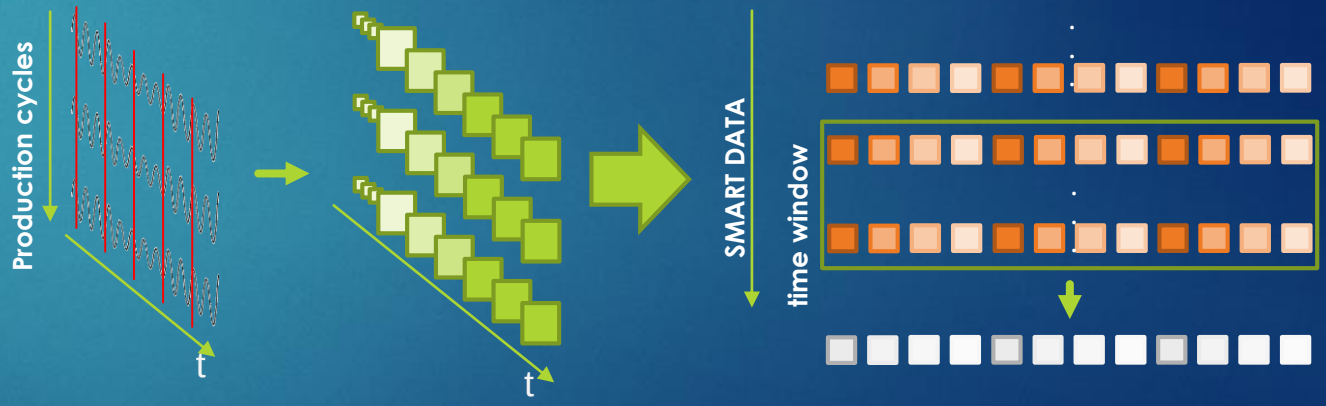
- **Removing correlated features, beside simplifying the model computation, can improve the model performance**
- **Multicollinearity**
  - Variables that can be predicted from the others with a substantial degree of accuracy using a multiple regression model could be discarded from the analysis.
- **Correlation Test**
  - Features highly correlated with other attributes (i.e., having dependence or association in any statistical relationship, whether causal or not) could be discarded from the analysis.



Data aggregation

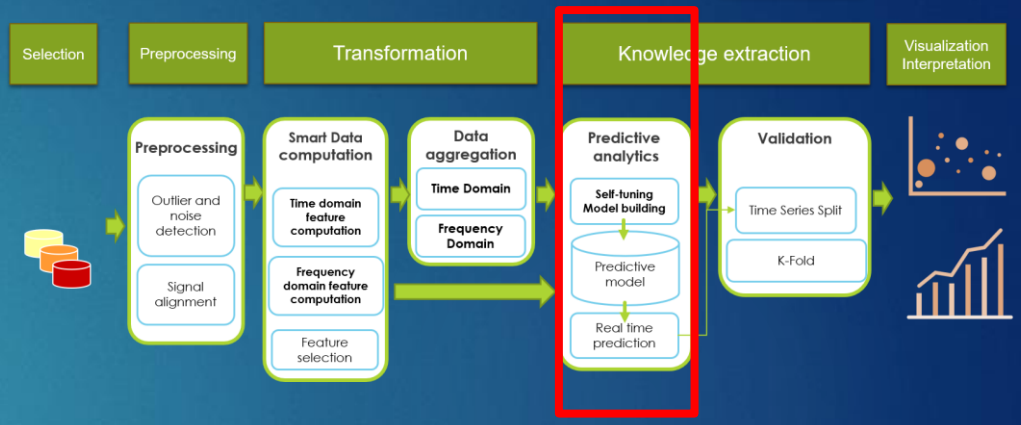
- Depending on the data under analysis, a data aggregation phase could be needed
- Prediction may not target a single production cycle, but a longer time horizon
  - Slowly-degrading multi-cycle industrial processes
- This step **reduces the data dimension**
  - **by aggregating smart data over time** and computing different sets of features to describe the behavior of the aggregated smart data





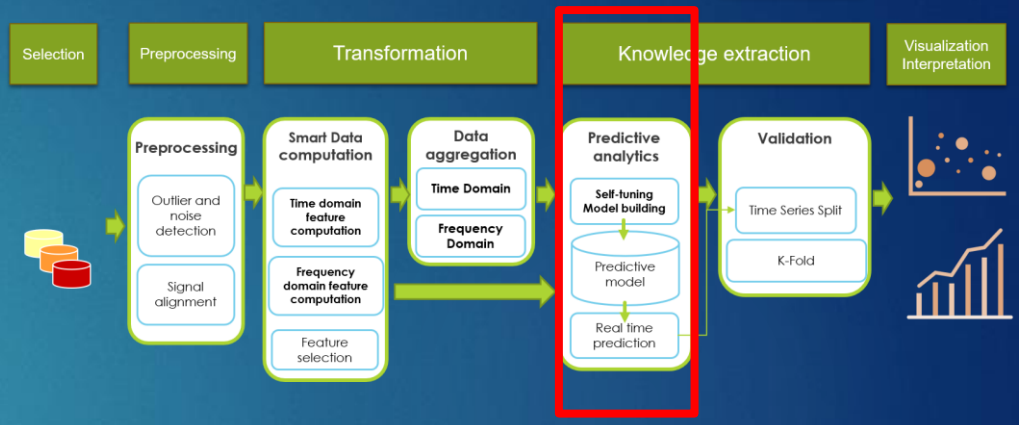
Predictive analytics

- Model building
  - This step requires **a set of historical labelled training data**.
  - The model is then tested over a part of the dataset.
- Real Time predictions
  - Once created, the model is applied in real time to the new signals received from the industrial processes.
- **Self-tuning strategy** to offload the data scientist to manually
  - set the specific algorithm parameters included in the proposed approach
  - identify the best algorithm to perform the prediction



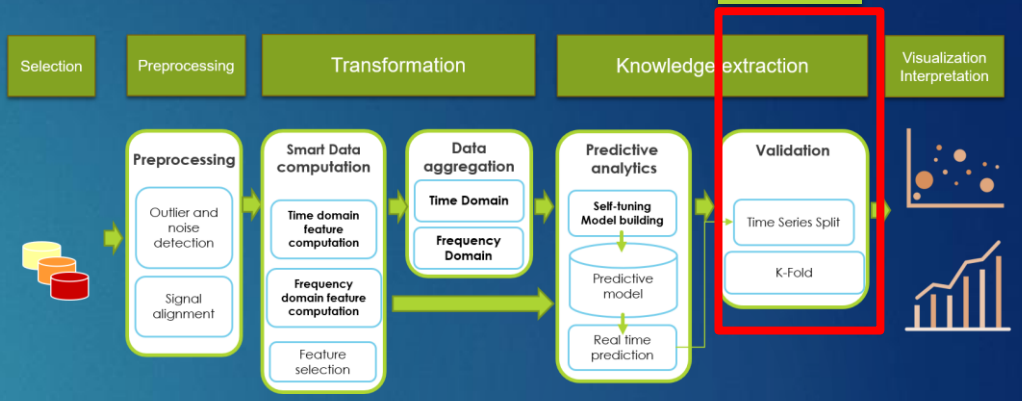
Predictive analytics

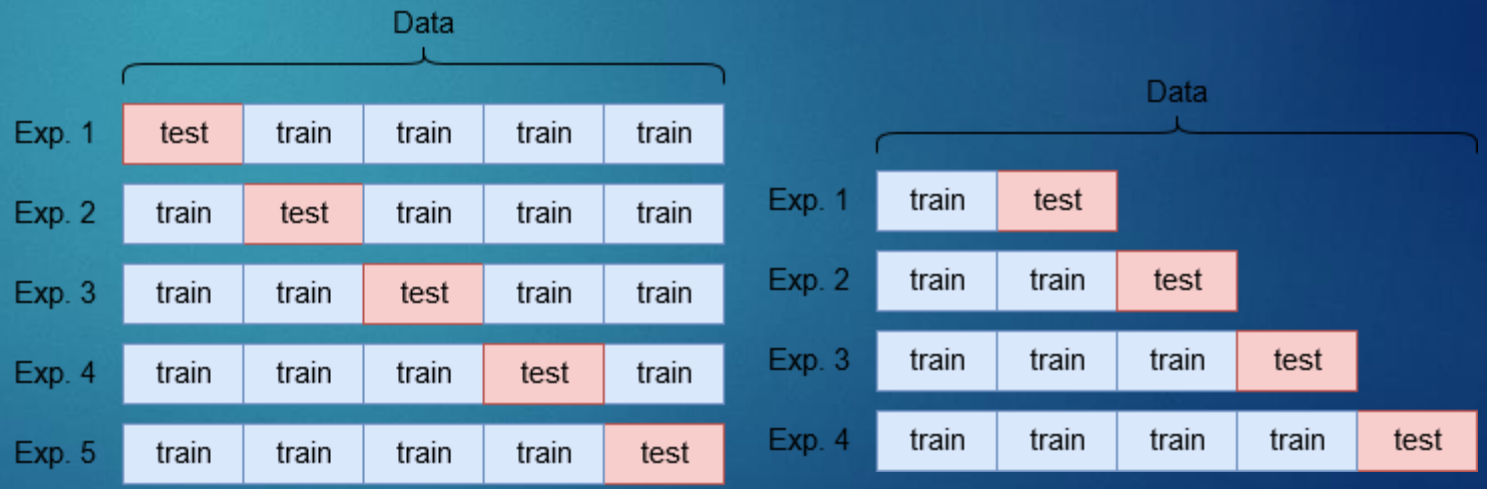
- Transparent mining (i.e., native interpretable models) allow the final user to know why **prediction outcomes** have been taken
  - Needed to target specific problems in the production processes and trigger precise corrective actions.
  - Algorithms used in the frameworks are: AdaBoost, Gradient Boosted Tree, Random Forest, K-Nearest Neighbors
- Black box models: very accurate models
  - They do not allow the final user to have a deep understanding the causes of the prediction



Validation

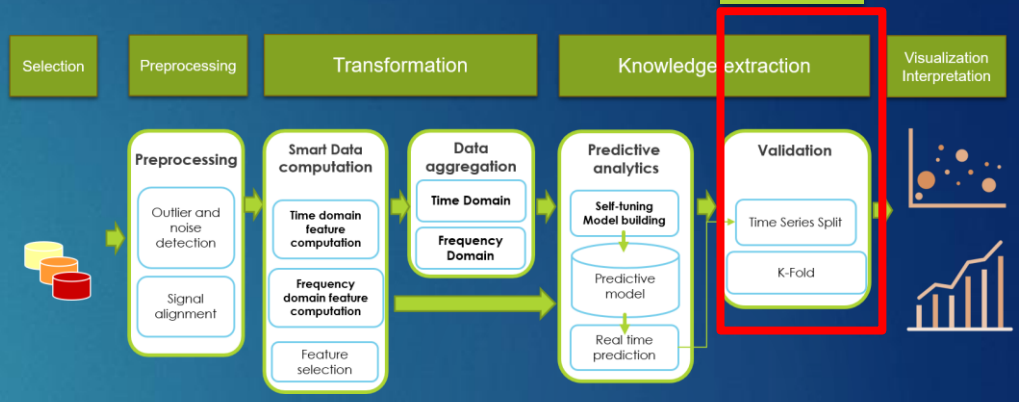
- Several partitioning techniques can be exploited to select the test set:
  - **K-Fold Stratified Cross Validation**
    - the dataset is split in K parts, each one used as test set alternatively
  - **TimeSeries Split Validation**
    - The training set grows at each iteration, following the time evolution

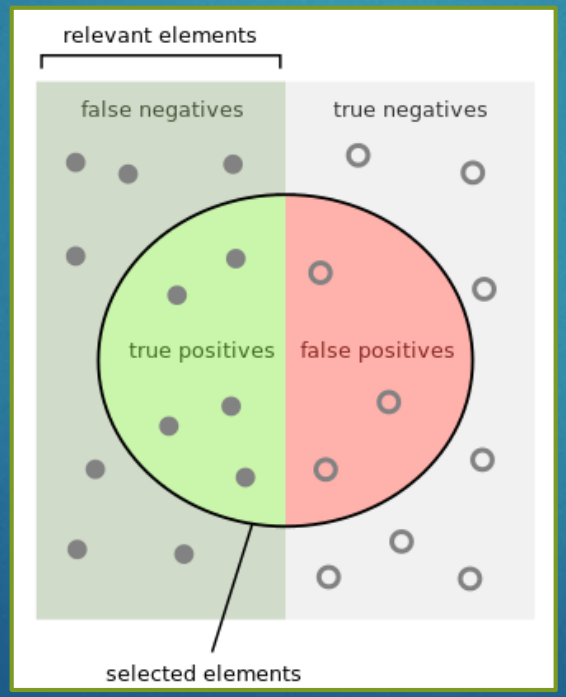


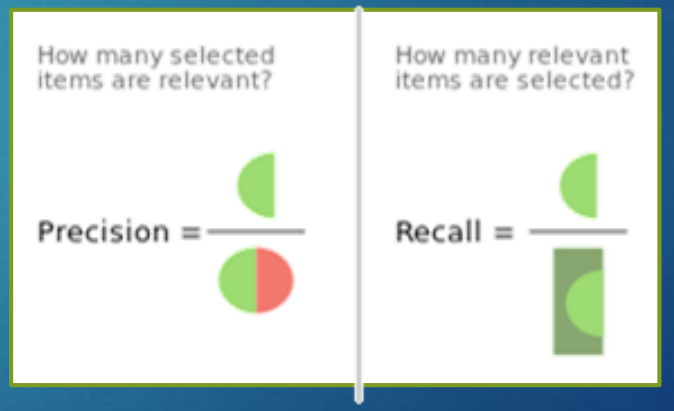


Validation

- The evaluation is performed by means on metrics
  - *F1-score: harmonic mean between Precision and Recall*
  - *Precision*
  - *Recall*







Experimental results: Robotics use case

Robotics use case

Data

Selection

Preprocessing

Transformation

Knowledge extraction

Visualization
Interpretation

Preprocessing

- Outlier and noise detection
- Signal alignment

Smart Data computation

- **Time domain feature computation**
- ~~**Frequency domain feature computation**~~
- Feature selection

Data aggregation

- ~~**Time Domain**~~
- ~~**Frequency Domain**~~

Predictive analytics

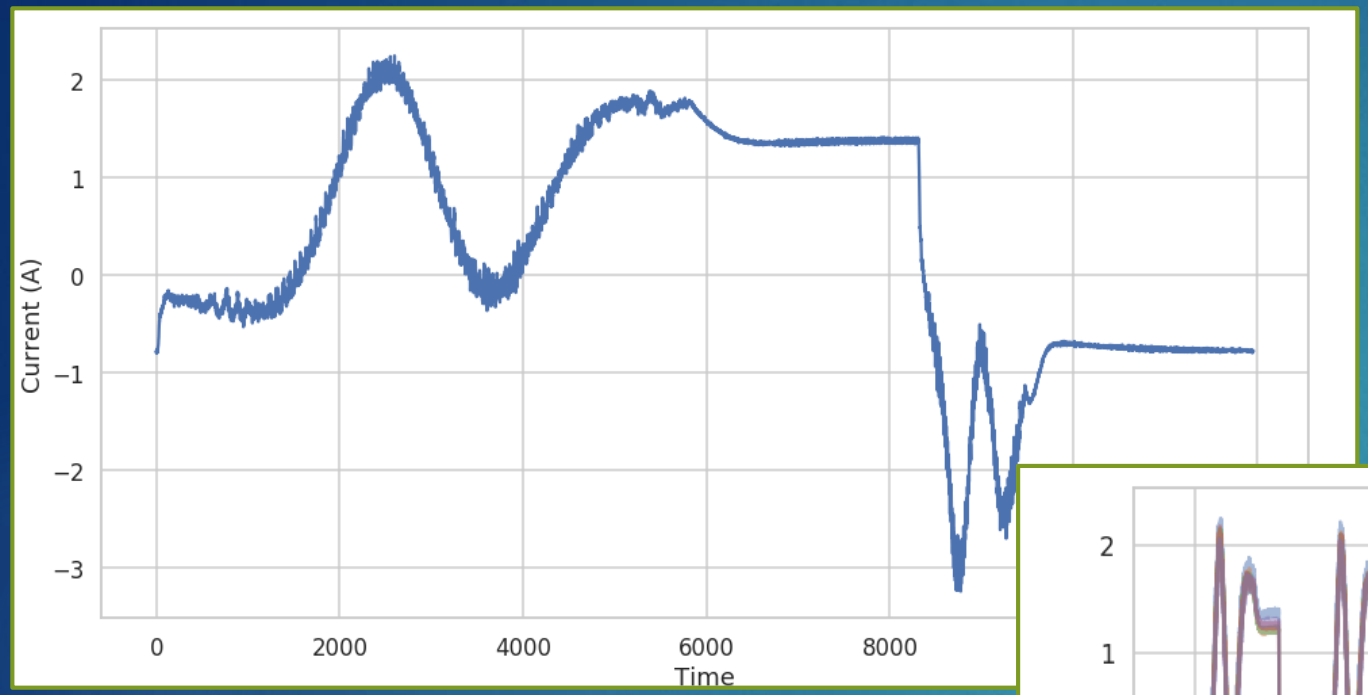
- **Self-tuning Model building**
- Predictive model
- Real time prediction

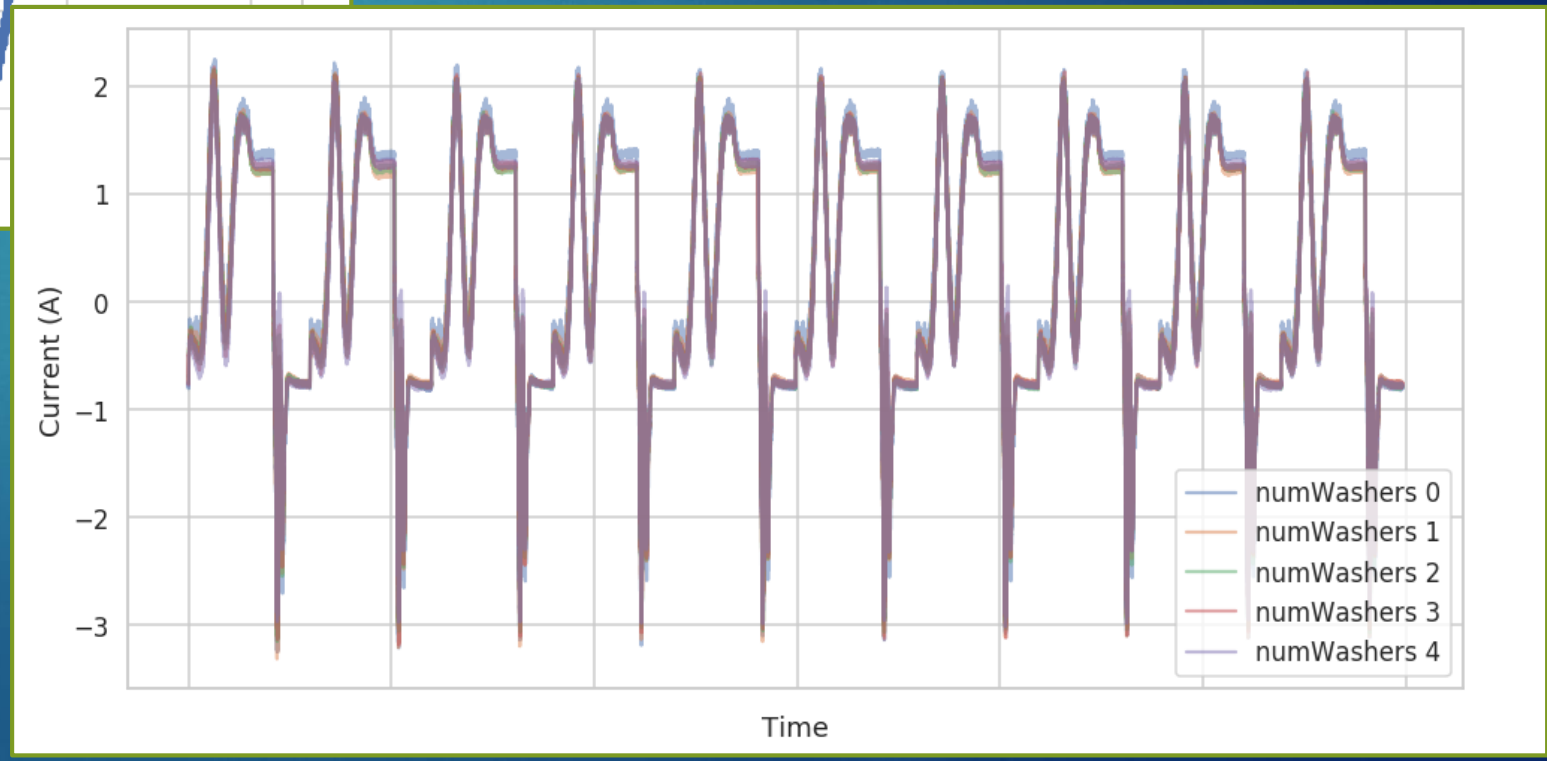
Validation

- Time Series Split
- K-Fold

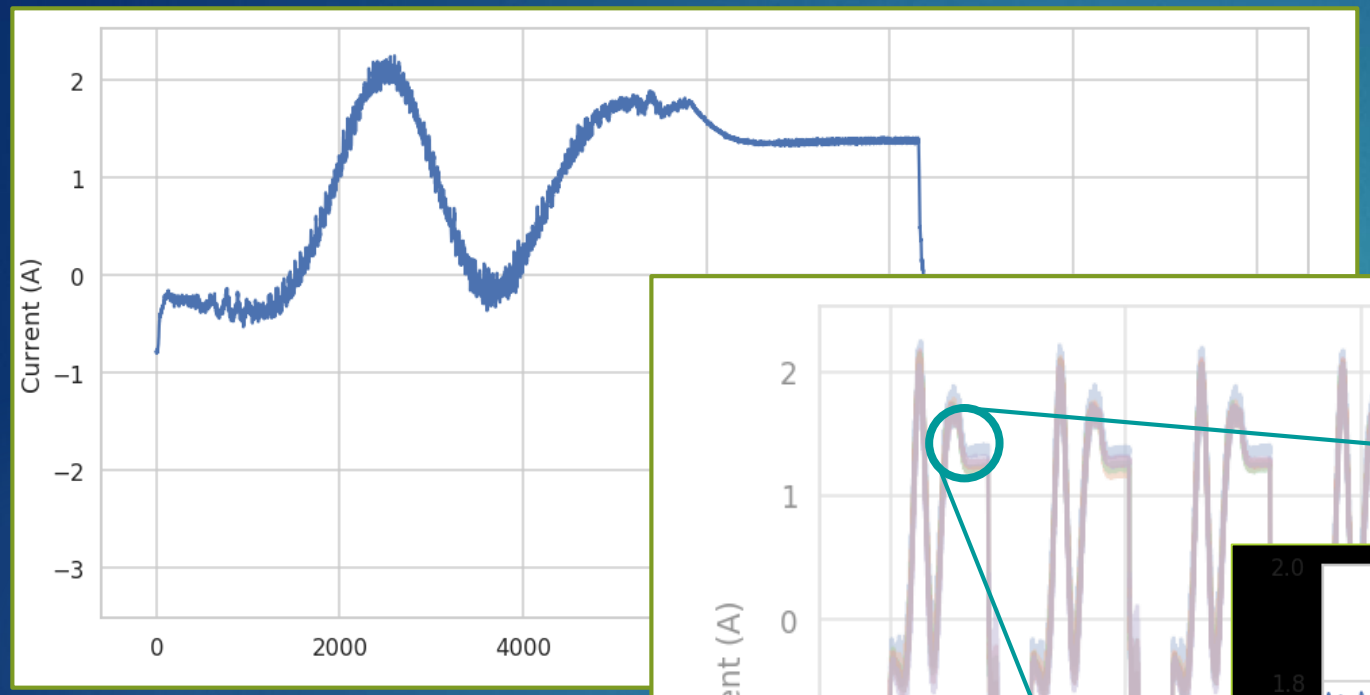
Data Base and Data Mining Group of Politecnico di Torino

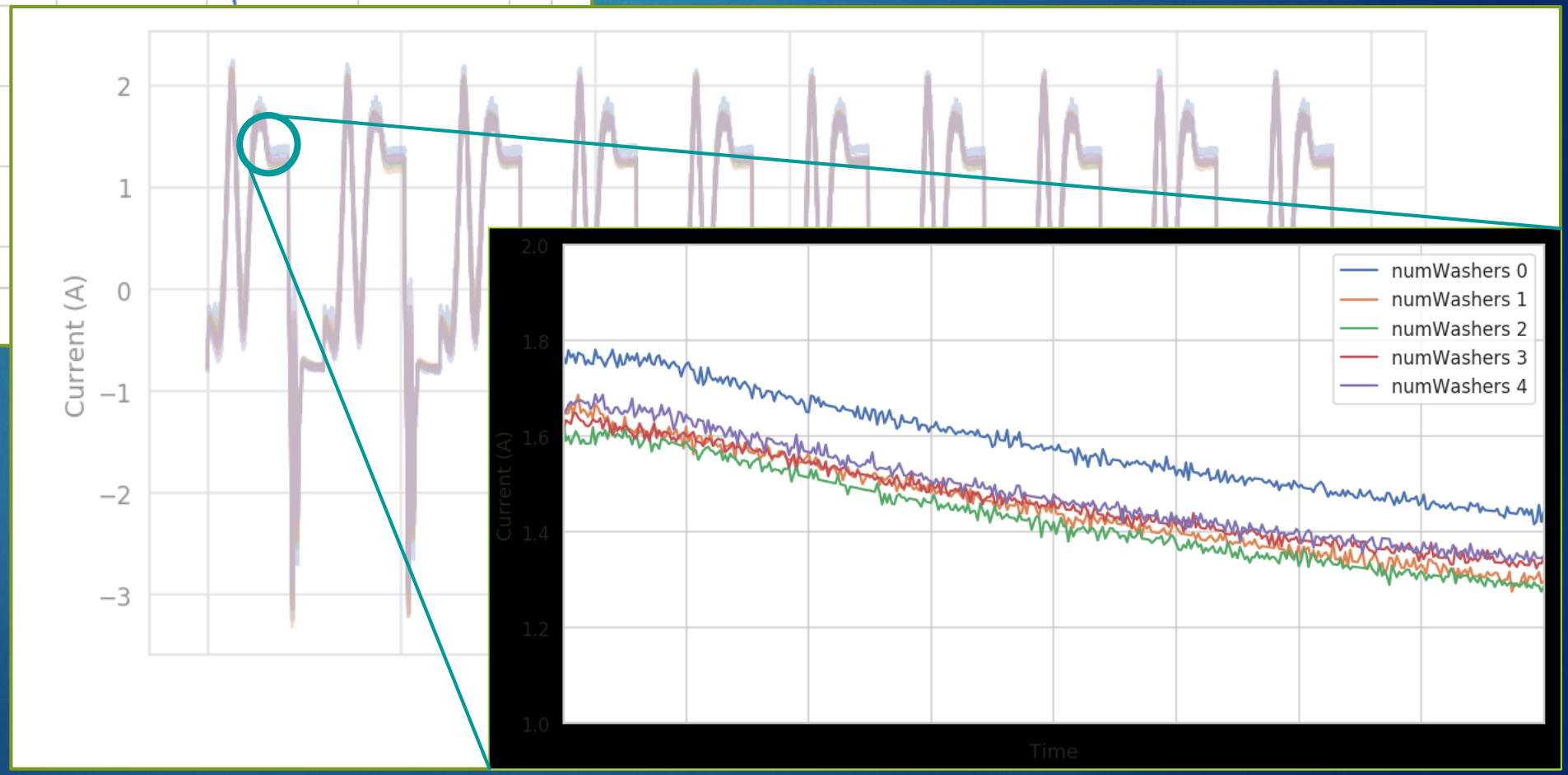
Belt tensioning data





Belt tensioning data

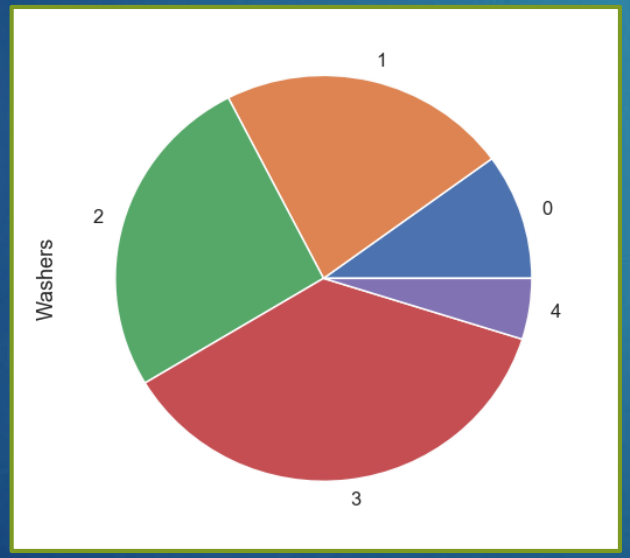


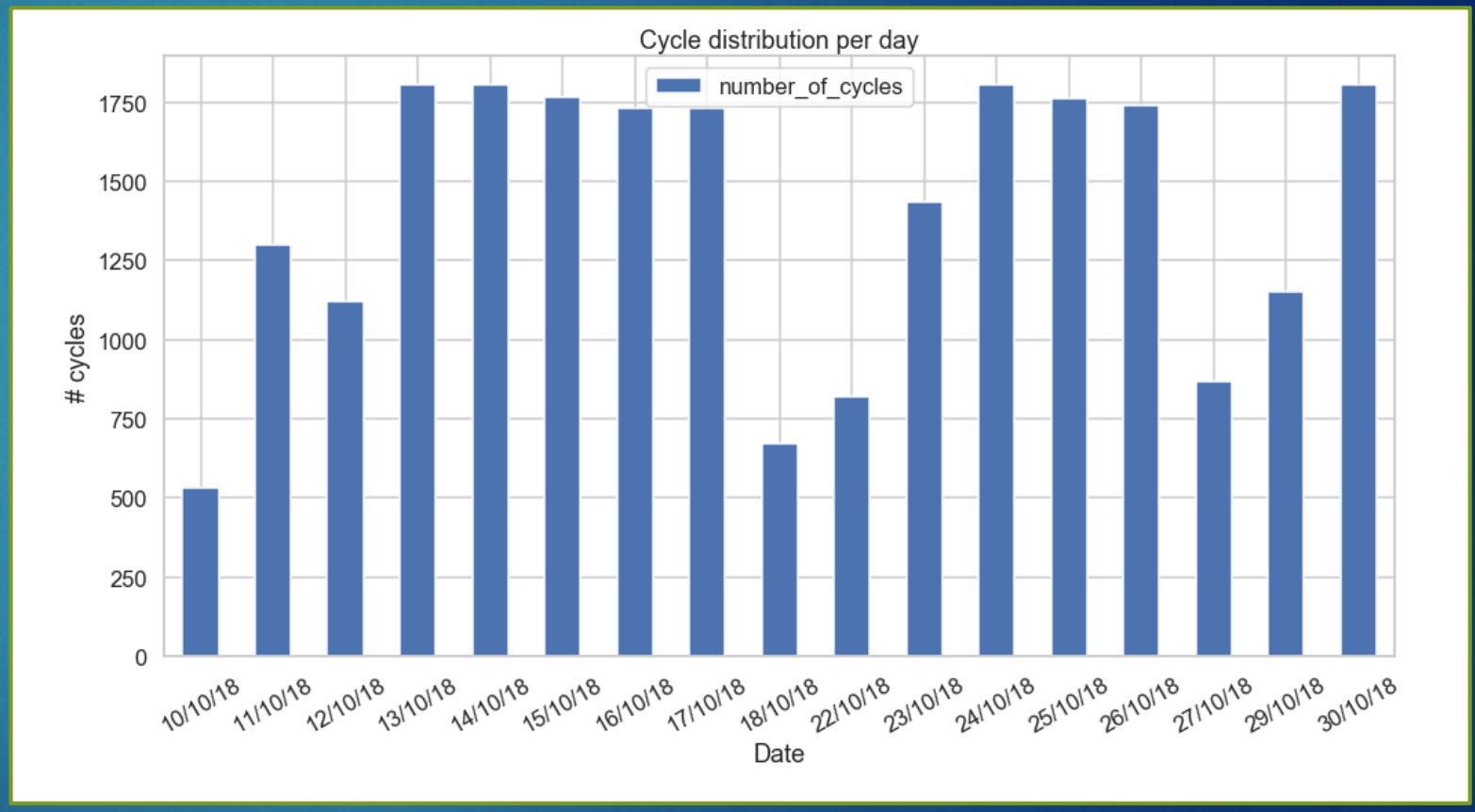


Belt tensioning data

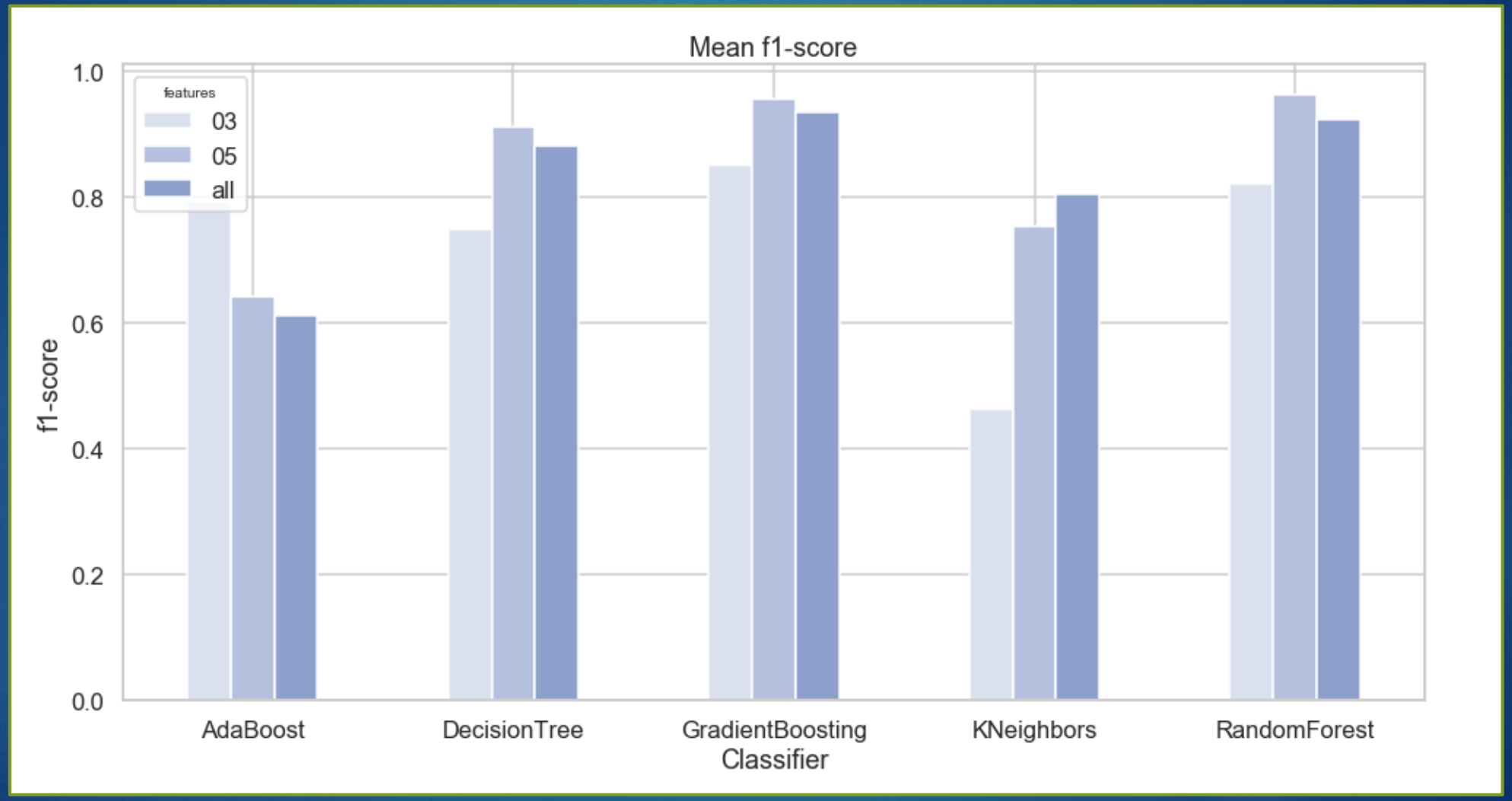
Number of cycles per class (NumWashers)

NumWashers	Number of cycles	% over the complete dataset
0	2,392	10.03
1	5,367	22.52
2	6,212	26.06
3	8,707	36.53
4	1,155	4.85
	23,833	

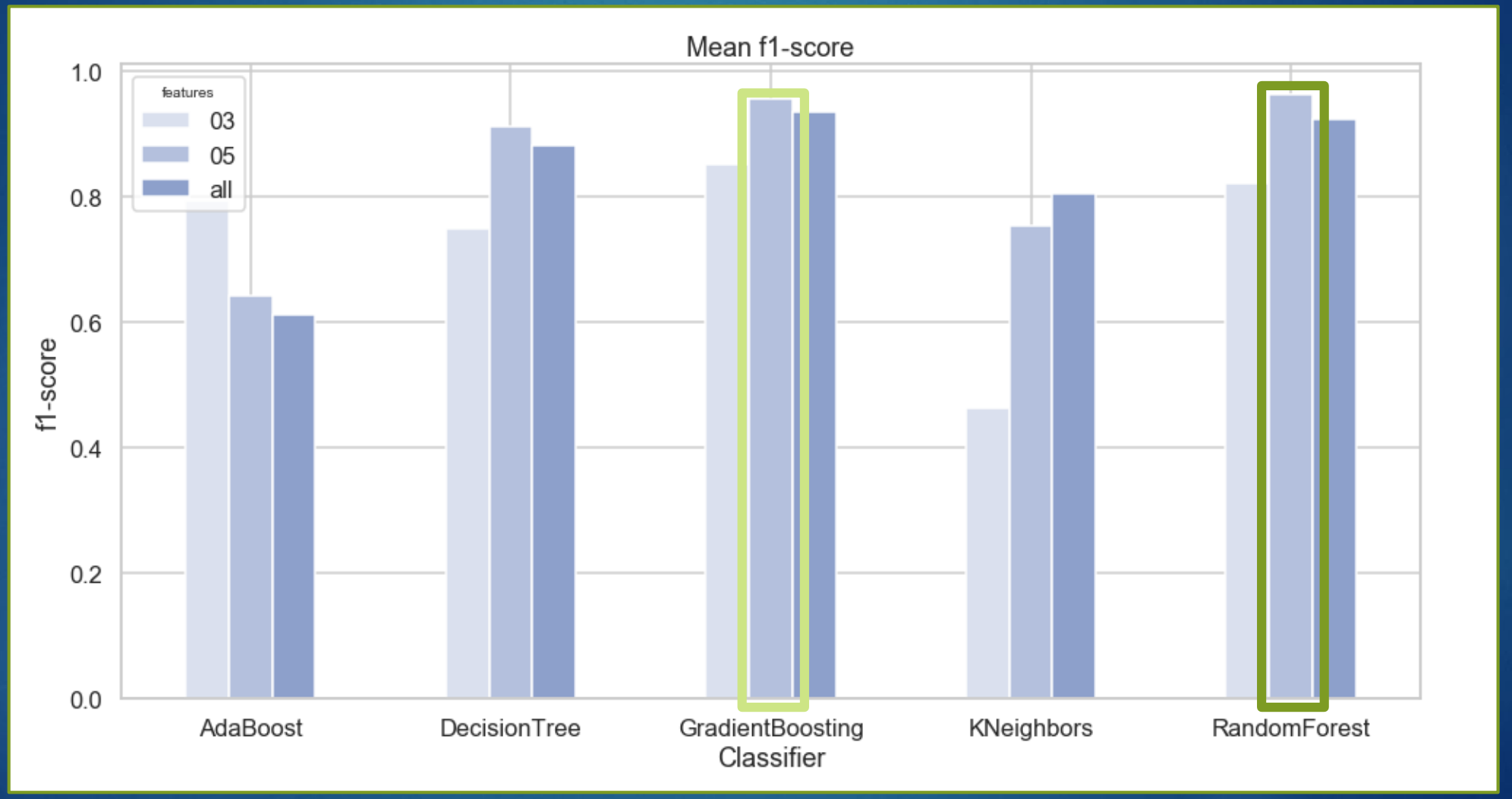




Predictive analytics



Predictive analytics



Experimental results: White goods use case

White goods use case

Data

Selection

Preprocessing

Transformation

Knowledge extraction

Visualization
Interpretation

Preprocessing

- Outlier and noise detection
- Signal alignment

Smart Data computation

- **Time domain feature computation**
- ~~**Frequency domain feature computation**~~
- Feature selection

Data aggregation

- **Time Domain**
- ~~**Frequency Domain**~~

Predictive analytics

- **Self-tuning Model building**
- Predictive model
- Real time prediction

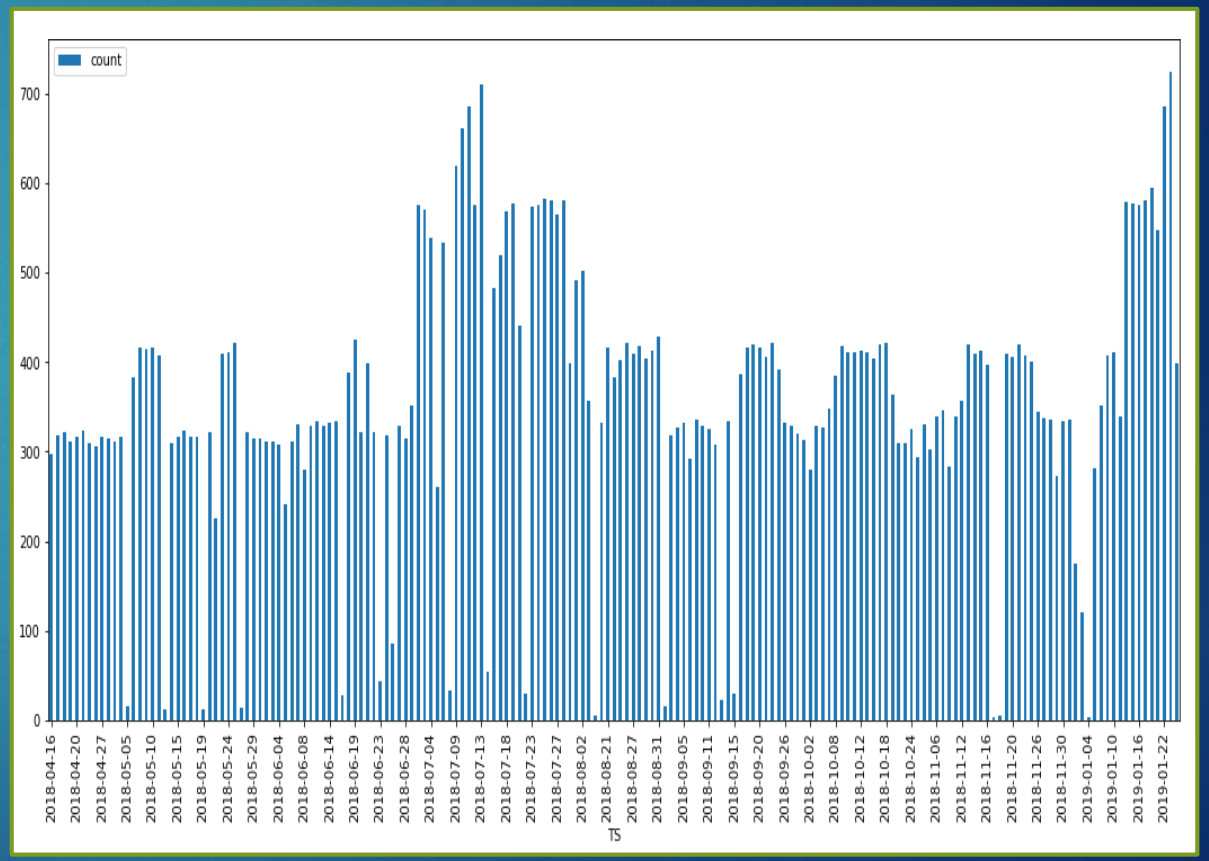
Validation

- Time Series Split
- K-Fold

White goods use case: dataset

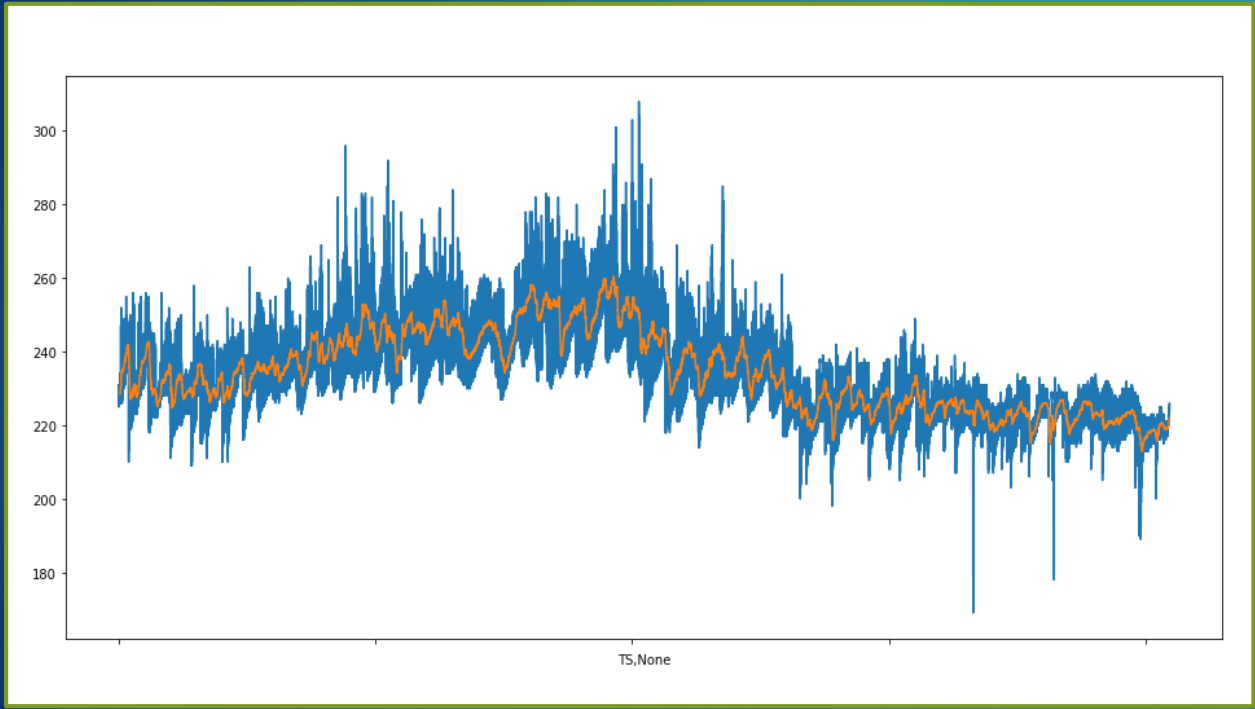
- # production cycles: 65,986
  - Days of analysis: 183
  - Start date: 2018-04-16 08:17:22
  - End date: 2019-01-30 08:59:47
  - Number of signals: 10 signals for each production cycle

- Number of cycles for each day

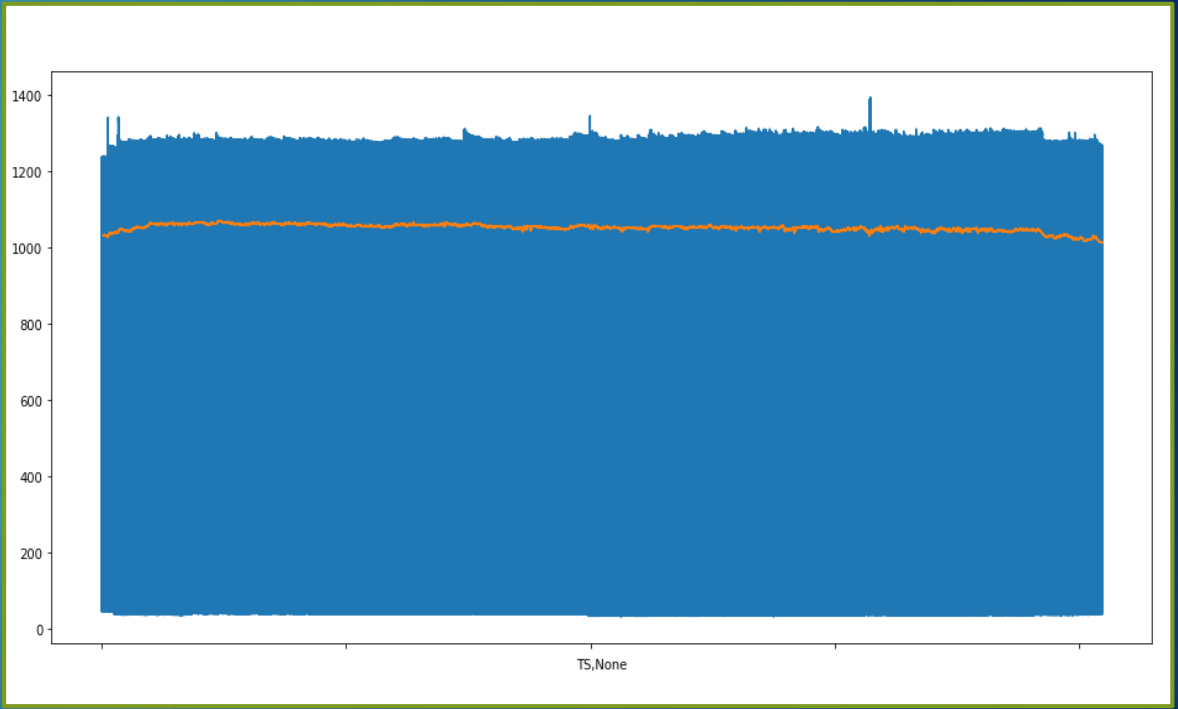


White good dataset

- Signal 2 distribution
  - The orange line represents the variable trends over the time axis



- Signal 1 distribution
  - The orange line represents the variable trends over the time axis



Data aggregation and labelling

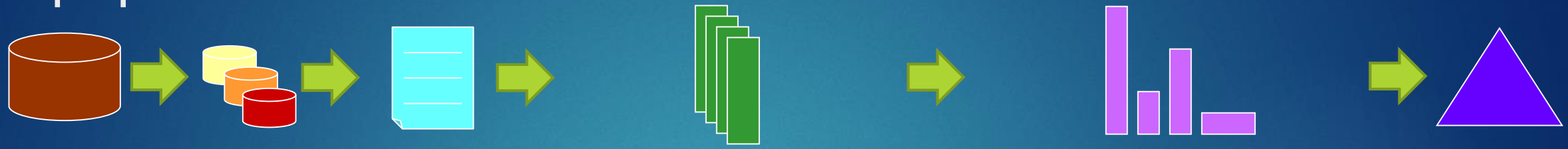
- Signal aggregation performed in the time domain
  - Aggregating time windows
    - 4 Hours
    - 8 Hours
    - 1 Day
- Labels are assigned accordingly to the time windows.
  - Labelling process requires domain expertise to be correctly carried out.
  - All the cycles in the window have been labelled as bringing to a failure if during the window an alarm occurred.

Classification and validation

Configuration	Classifier	Label	Precision	Recall	F1-Score
1d_all	**RandomForestClassifier**	**1**	**0.84**	**0.85**	**0.83**
1d_all	GradientBoostingClassifier	1	0.81	0.81	0.79
1d_all	AdaBoostClassifier	1	0.77	0.75	0.74
1d_correlation	***RandomForestClassifier***	***1***	***0.66***	***0.67***	***0.65***
1d_correlation	GradientBoostingClassifier	1	0.63	0.65	0.62
1d_multicollineariry	AdaBoostClassifier	1	0.58	0.64	0.59
1d_multicollineariry	GradientBoostingClassifier	1	0.62	0.58	0.59
1d_multicollineariry	***RandomForestClassifier***	***1***	***0.58***	***0.55***	***0.55***
1d_correlation	AdaBoostClassifier	1	0.59	0.53	0.52
4H_all	RandomForestClassifier	1	0.63	0.35	0.42
4H_all	GradientBoostingClassifier	1	0.45	0.32	0.36
8H_all	RandomForestClassifier	1	0.51	0.30	0.35
8H_all	GradientBoostingClassifier	1	0.47	0.30	0.33
8H_all	AdaBoostClassifier	1	0.39	0.30	0.32
4H_all	AdaBoostClassifier	1	0.33	0.26	0.29
8H_correlation	AdaBoostClassifier	1	0.37	0.26	0.27
8H_multicollineariry	AdaBoostClassifier	1	0.33	0.28	0.27
8H_multicollineariry	GradientBoostingClassifier	1	0.34	0.25	0.27
8H_correlation	GradientBoostingClassifier	1	0.31	0.22	0.23
4H_correlation	AdaBoostClassifier	1	0.27	0.22	0.23
4H_correlation	GradientBoostingClassifier	1	0.26	0.19	0.21
4H_multicollineariry	RandomForestClassifier	1	0.37	0.14	0.20
4H_multicollineariry	GradientBoostingClassifier	1	0.28	0.15	0.20
8H_multicollineariry	RandomForestClassifier	1	0.30	0.14	0.19
8H_correlation	RandomForestClassifier	1	0.26	0.14	0.18
4H_multicollineariry	AdaBoostClassifier	1	0.20	0.16	0.18
4H_correlation	RandomForestClassifier	1	0.33	0.11	0.17

Semi-supervised data labelling

A semi-supervised data labelling pipeline



Data

Selection

Preprocessing

Transformation

Knowledge extraction

Visualization Interpretation

Preprocessing

- Outlier and noise detection
- Cycle alignment

Smart Data computation

- Time series feature computation
- Feature selection

Data aggregation

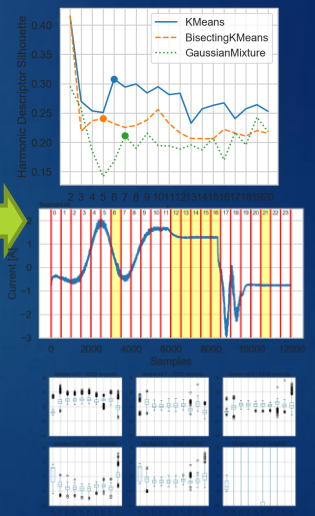
- Multi-cycle time-based aggregation

Self-tuning cluster analysis

- Unsupervised data partitioning
- Self-tuning strategy
- Unsupervised data models

Semi-supervised data labelling

- Top 10 features selection
- Data characterization
- Labelled data

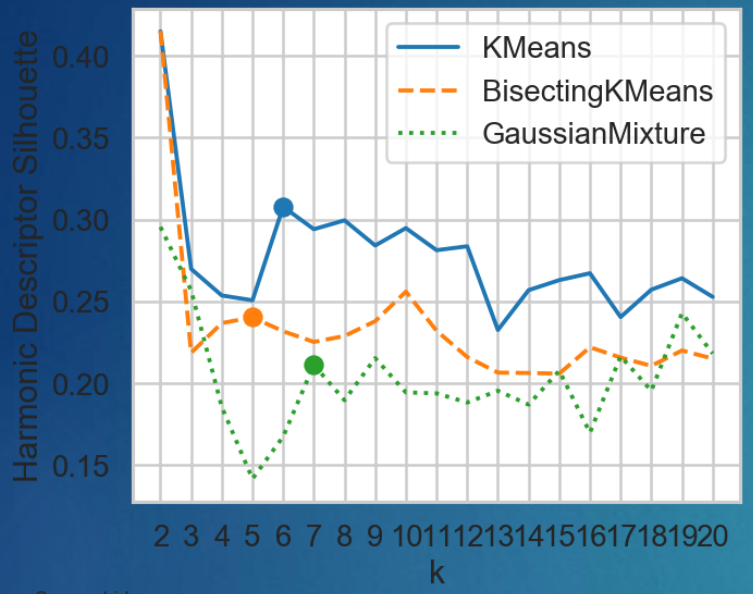


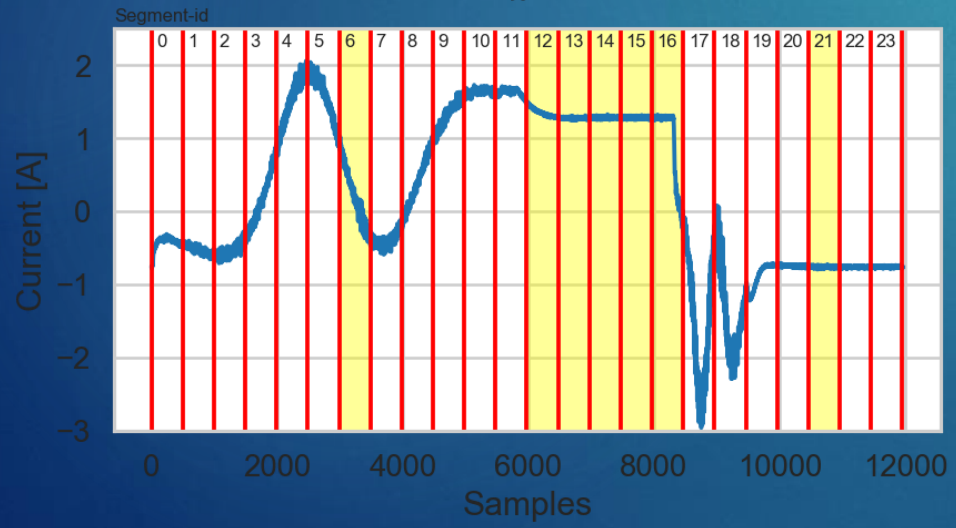
Data Base and Data Mining Group of Politecnico di Torino

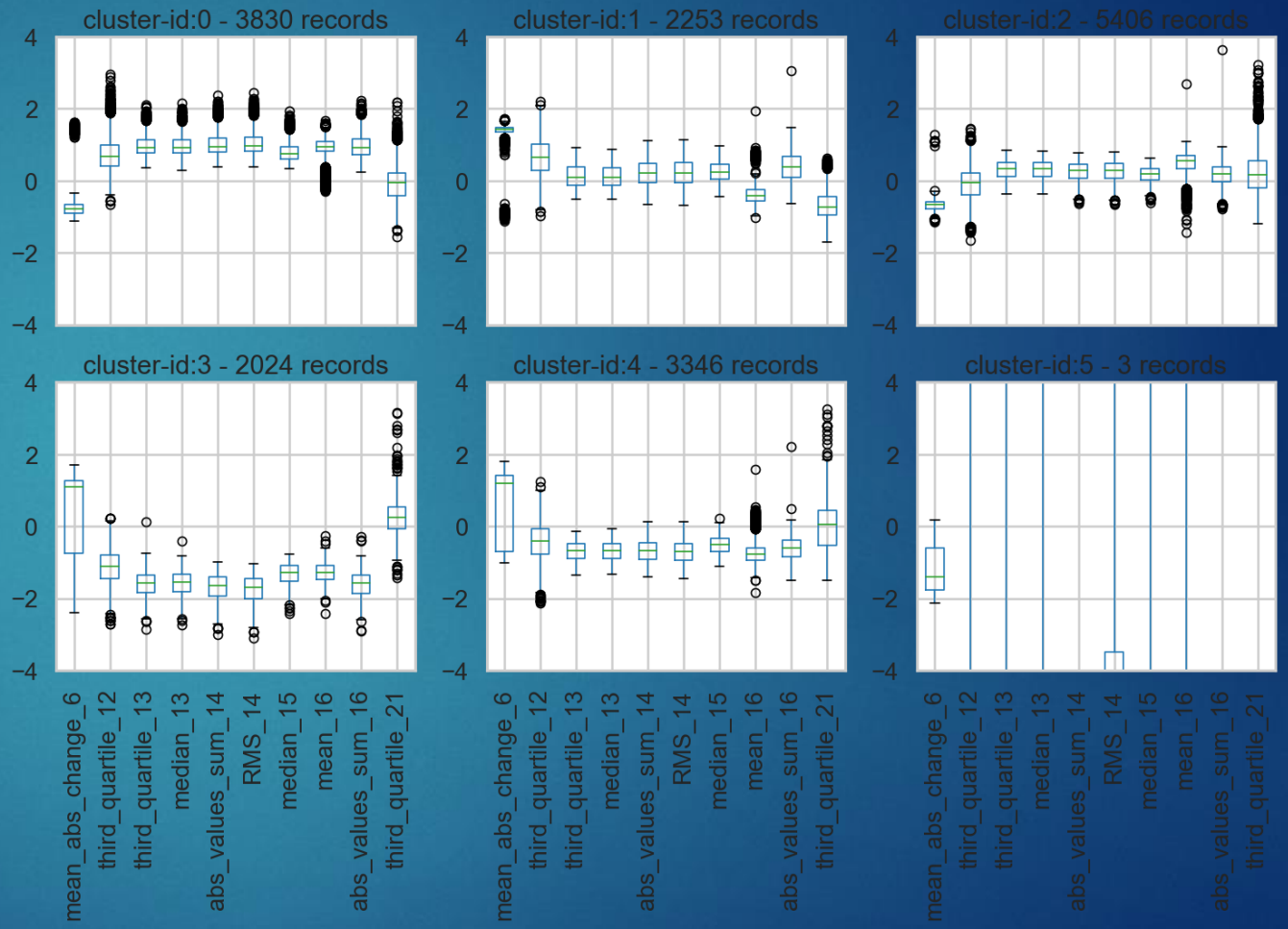
Key components

- **Self-tuning strategy to offload the data scientist from**
  - setting the specific algorithm parameters
  - selecting the best algorithm
- **To help the domain expert to easily define the data labelling to production cycles/machines in the plants**
  - The top-k smart data features to focus only on the most relevant features
  - Boxplot distribution for the top-k features
  - Few representative samples for each cluster are manually inspected
  - Most relevant sub-cycles are highlighted

Preliminary results: Robotics use case



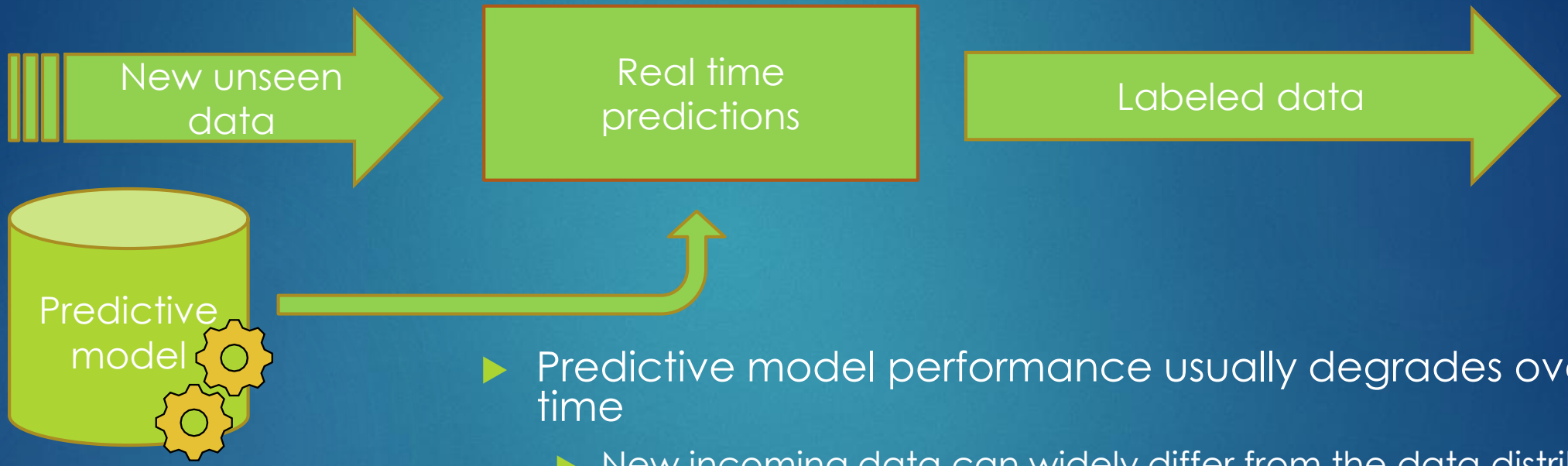




Data Base and Data Mining Group of Politecnico di Torino

Concept drift management

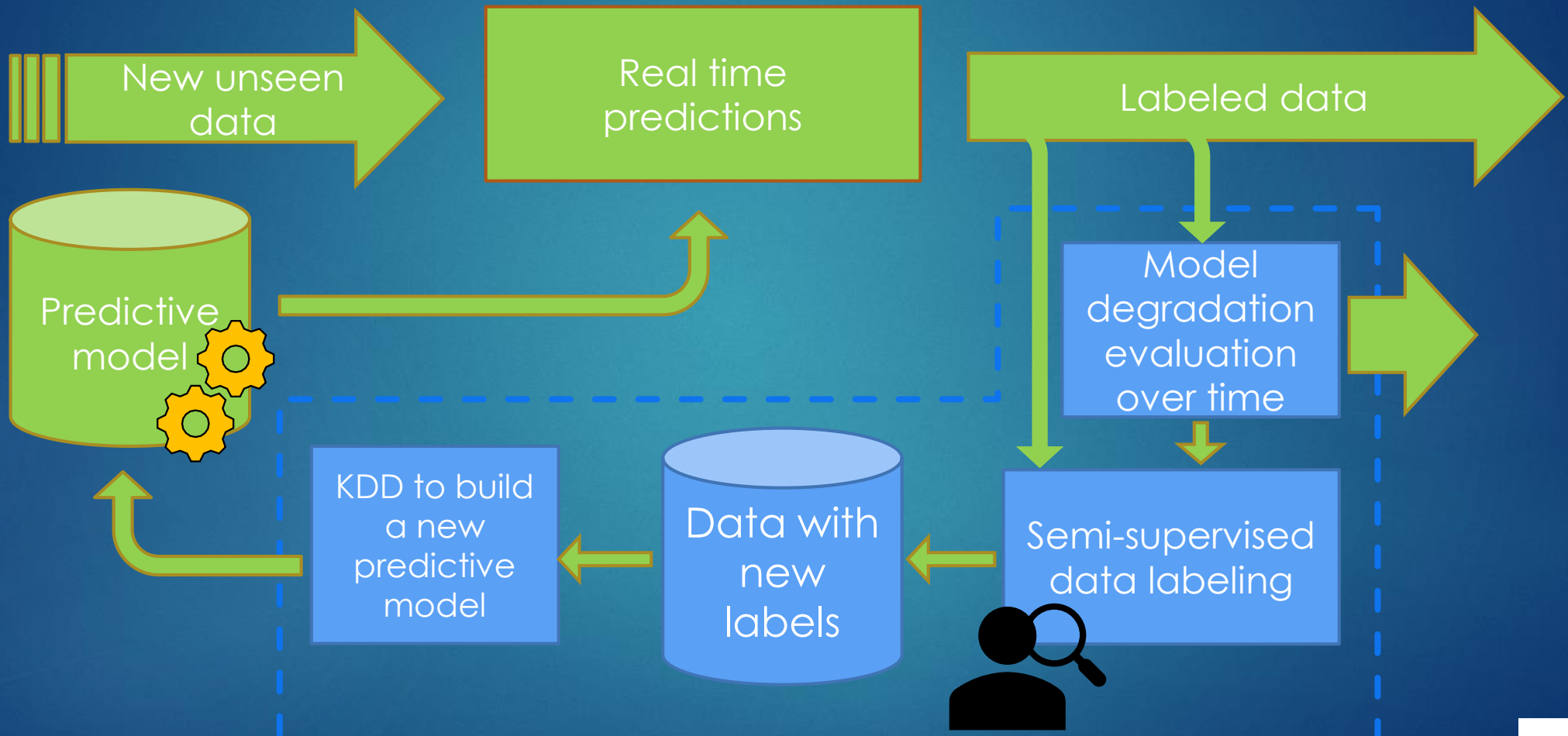
Concept drift management



- Predictive model performance usually degrades over time
  - New incoming data can widely differ from the data distribution on which the model was trained
  - Not all possible classes (labels) are known at training time
  - Real time predictions performed on new unseen data may be misleading or totally wrong

Data Base and Data Mining Group of Politecnico di Torino

Automated concept drift management



Key components

- **Automatic triggering of the predictive model retraining only when necessary**
  - Towards real-time evaluation
- **Unsupervised approach, without the ground-truth labels for the newly classified samples**
  - Different (scalable) quality metrics for drift detection
    - e.g., the Silhouette score

The Silhouette score

- It is a quality measure of how similar an object is to its own cluster/group (**cohesion**) compared to other clusters/groups (**separation**).
- The silhouette ranges from -1 to +1
  - a **high value** indicates that the object is **well matched** to its own cluster and poorly matched to neighbouring cluster.
- For each record in the dataset, the silhouette is defined as:

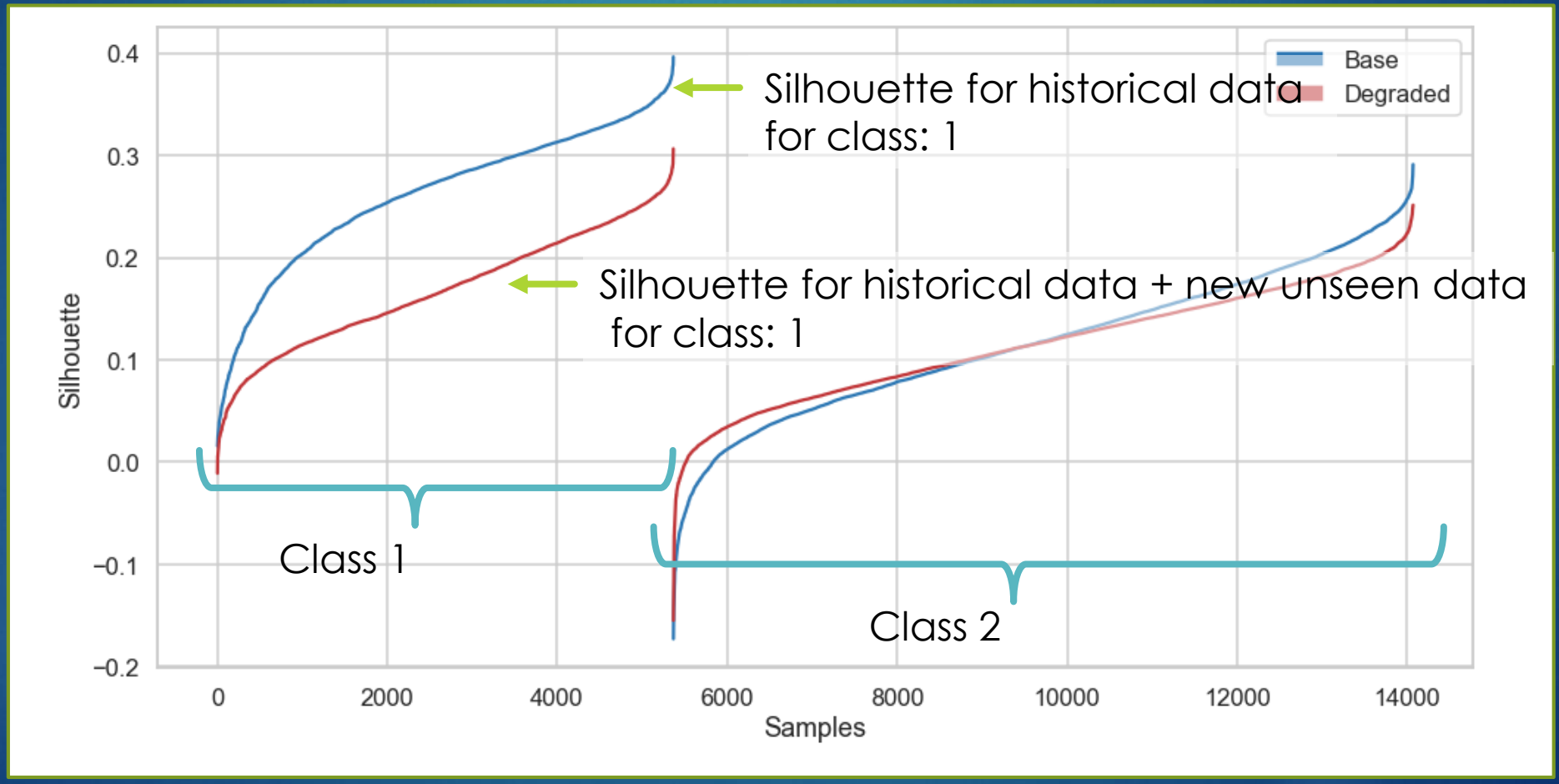
$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

where:

- a_i is the **average distance** between i and the other records in the same cluster,
- b_i is the **lowest average distance** between the record i and each one of the other clusters (not containing the record i)

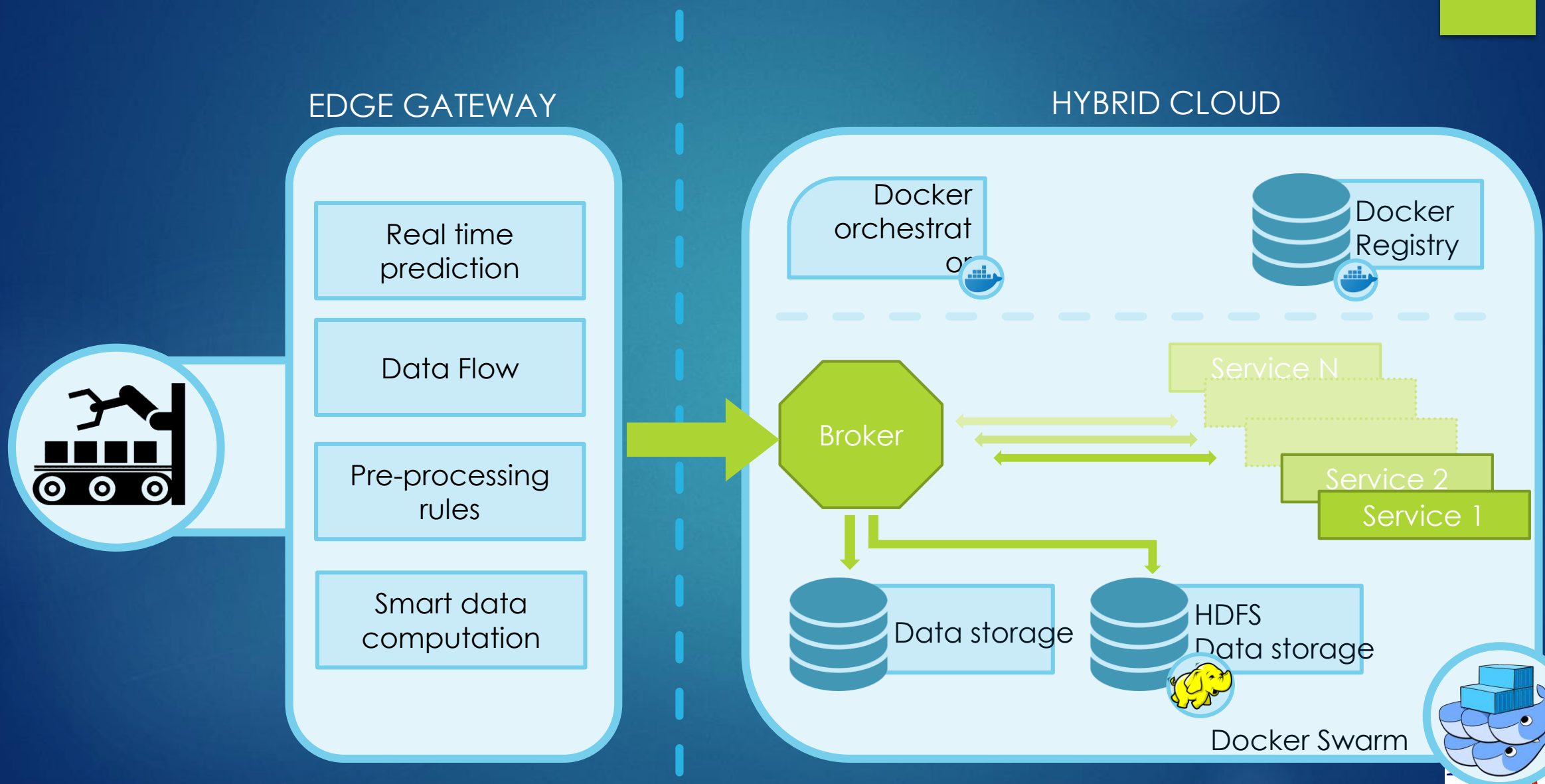
Concept drift detection

Silhouette score for each cycle before and after the prediction of the unseen data

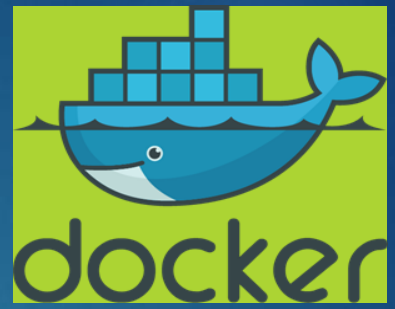


Technological aspects

A cloud-to-edge architecture



- Fast and general engine for **large-scale data processing**
  - Batch processing
  - Streaming processing
  - 10 to 100 times faster than Hadoop MapReduce
- **Advanced Analytics**
  - Map-Reduce, SparkSQL, MLlib **(machine learning)**
  - Spark **Streaming**, Spark Graphix (graph support)
- **Unified Engine**, Spark can run on top of
  - Hadoop, Cassandra, Amazon S3
  - Mesos, Standalone
- **In-memory data sharing**
  - Good for iterative, interactive and event stream processing tasks.
- Active, expanding user **community**



- Containerization platform
- Microservices architecture deployment
- Faster application development and delivery
- It improves modularity
  - applications easier to understand
- Container isolation makes application portable to any infrastructure
- Open source technology and a modular design
  - easy to integrate into your existing environment

Open issues

- **The automated predictive analytics pipeline is still a dream**
  - Lack of a **general-purpose** data-driven methodology
  - **Different** data-analytics **solutions** might be required to correctly address **a specific use-case**
  - Specific aspects in the Industry 4.0 context requires specific algorithms
  - The data-driven methodologies might be **context-dependent**

Publications

- Tania Cerquitelli, Stefano Proto, Francesco Ventura, Daniele Apiletti, Elena Baralis: **Towards a real-time unsupervised estimation of predictive model degradation.** BIRTE 2019: 5:1-5:6
- Daniele Apiletti, Claudia Barberis, Tania Cerquitelli, Alberto Macii, Enrico Macii, Massimo Poncino, Francesco Ventura: **iSTEP, an Integrated Self-Tuning Engine for Predictive Maintenance in Industry 4.0.** In IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, 2018, Melbourne, Australia, December 11-13, 2018. IEEE 2018, ISBN 978-1-7281-1141-4. pp 924-931 – Best paper Award
- Stefano Proto, Francesco Ventura, Daniele Apiletti, Tania Cerquitelli, Elena Baralis, Enrico Macii, Alberto Macii: **PREMISES, a Scalable Data-Driven Service to Predict Alarms in Slowly-Degrading Multi-Cycle Industrial Processes**. 2019 IEEE International Congress on Big Data, BigData Congress 2019, Milan, Italy, July 8-13, 2019. IEEE 2019, ISBN 978-1-7281-2772-9: pp. 139-143
- Francesco Ventura, Stefano Proto, Daniele Apiletti, Tania Cerquitelli, Simone Panicucci, Elena Baralis, Enrico Macii, Alberto Macii: **A New Unsupervised Predictive-Model Self-Assessment Approach that SCALEs**. 2019 IEEE International Congress on Big Data, BigData Congress 2019, Milan, Italy, July 8-13, 2019. IEEE 2019, ISBN 978-1-7281-2772-9: pp. 144-148
- Tania Cerquitelli, David Bowden, Angelo Marguglio, Lucrezia Morabito, Chiara Napione, Simone Panicucci, Nikolaos Nikolakis, Sotiris Makris, Guido Coppo, Salvatore Andolina, Alberto Macii, Enrico Macii, Niamh O'Mahony, Paul Becker, Sven Jung: **A Fog Computing Approach for Predictive Maintenance**. CAiSE Workshops 2019: 139-147

Data Base and Data Mining Group of Politecnico di Torino

... questions?

Tania CERQUITELLI
tania.cerquitelli@polito.it

Data Base and Data Mining Group of Politecnico di Torino