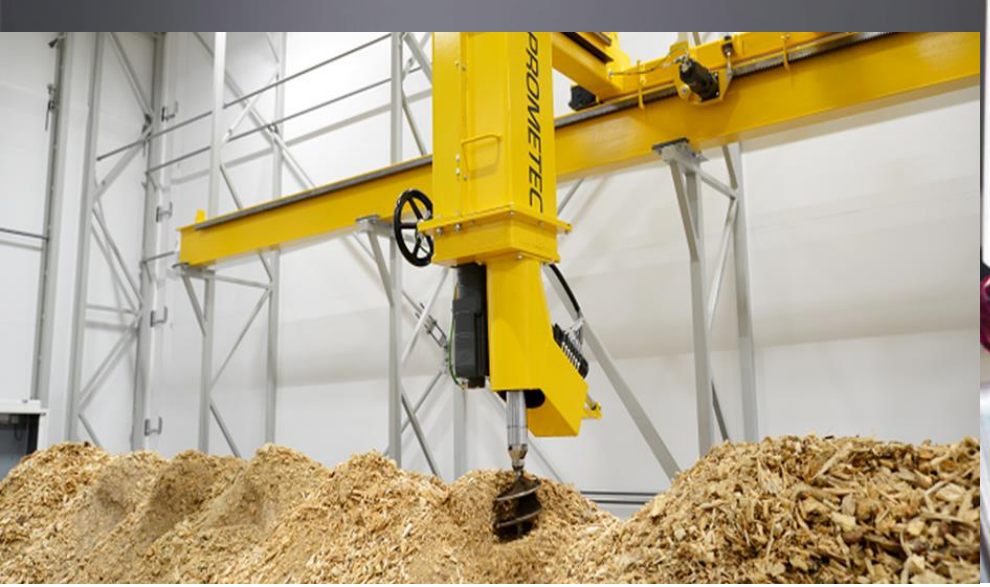


Data mining and its applicability to biofuel measurement data

Michal Voldán, ČMI

Prometec

TÉCNICO LISBOA

DANISH TECHNOLOGICAL INSTITUTE

VERDO

CETIAT
ensemble, innover et valider

PTB
Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin

INSTITUT FRESNEL

TÜBİTAK UME

IMBIH

ČESKÝ METROLOGICKÝ INSTITUT

biofMET

EMPIR

EURAMET

The EMPIR initiative is co-funded by the European Union's Horizon 2020 research and innovation programme and the EMPIR Participating States

Stakeholder workshop
28.-29. March 2023

WP3 – Industrial validation

- Task 3.3: Data mining and data-science techniques

The aim of this task is to use **advanced data-science techniques** to improve reproducibility and repeatability of moisture content measurements by:

- **Development of a method for the analysis of moisture content measerement results** (machine learning, artificial intelligence, deep learning)

- **Refining existing analyses by applying data science methods** such as machine learning and artificial intelligence on the now fully digitalised measurements (correlations arising from the industrial environment - temperature, reflections, humidity, dust, vibrations etc.)

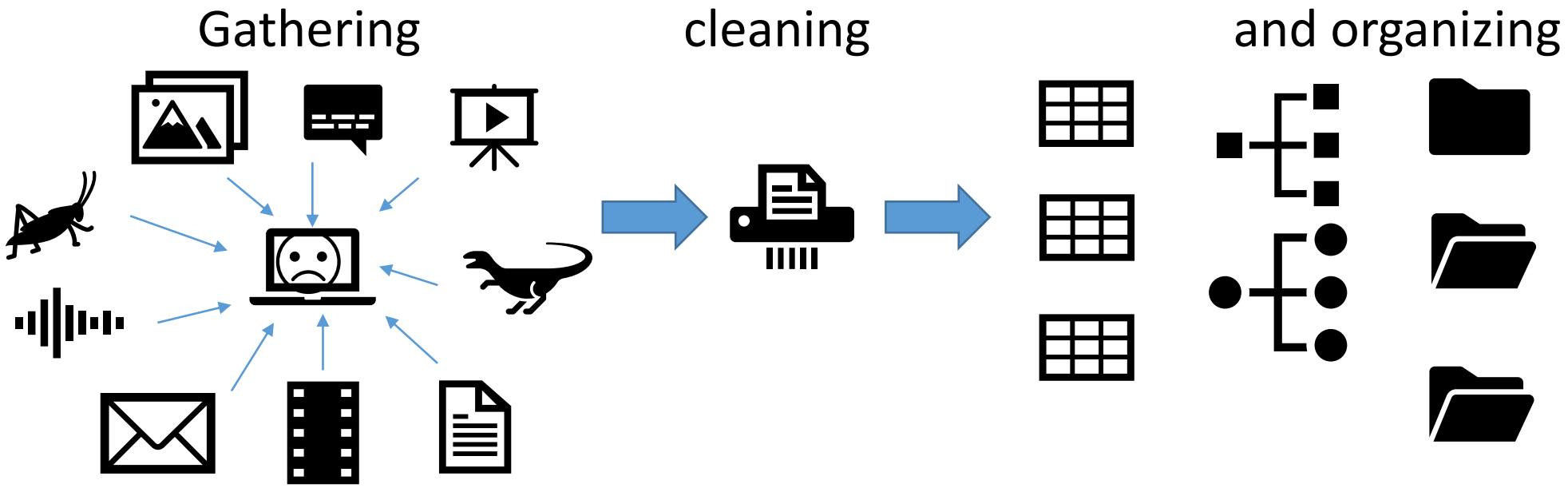
Data mining and data-science techniques

- Data mining = the process of discovering patterns, trends, and correlations from large amounts of data.
- It involves using statistical and computational techniques
- The goal is to extract meaningful information from raw data

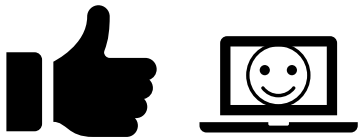



The process of data mining

Data Preparation – it includes



the data from various sources into a format that can be used for analysis

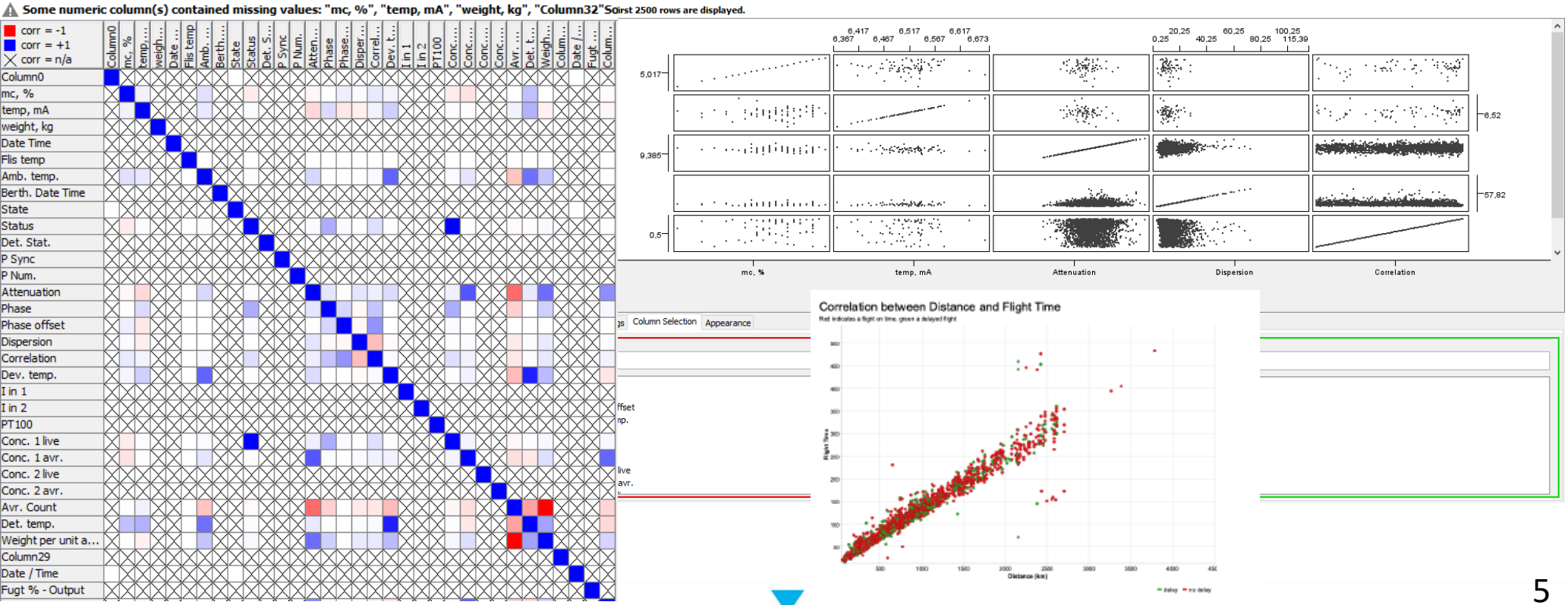


This step is crucial as the quality of the data will impact the accuracy of the results obtained from data mining.

The process of data mining

Data Exploration –to identify any patterns or trends within the data

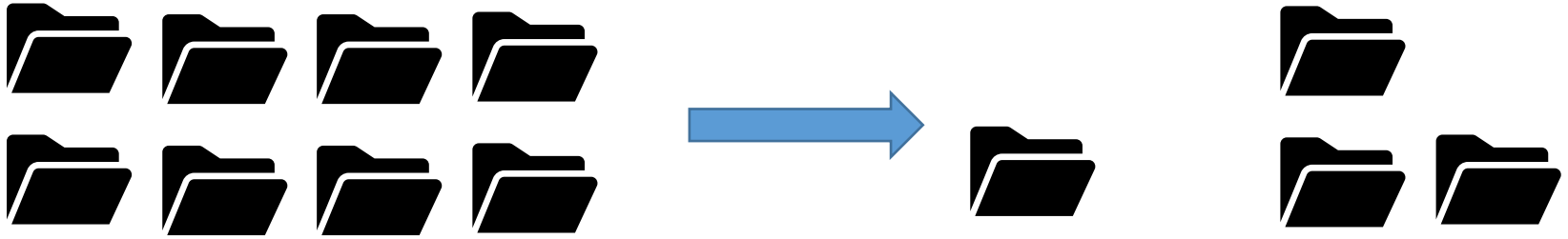
- can be done using various visualization techniques such as scatter plots, histograms, and box plots



The process of data mining

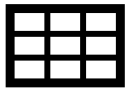
Data Reduction

- reducing the amount of data that needs to be analysed



- removing irrelevant data (e.g., duplicate or irrelevant records)

- unifying the data to make it more manageable

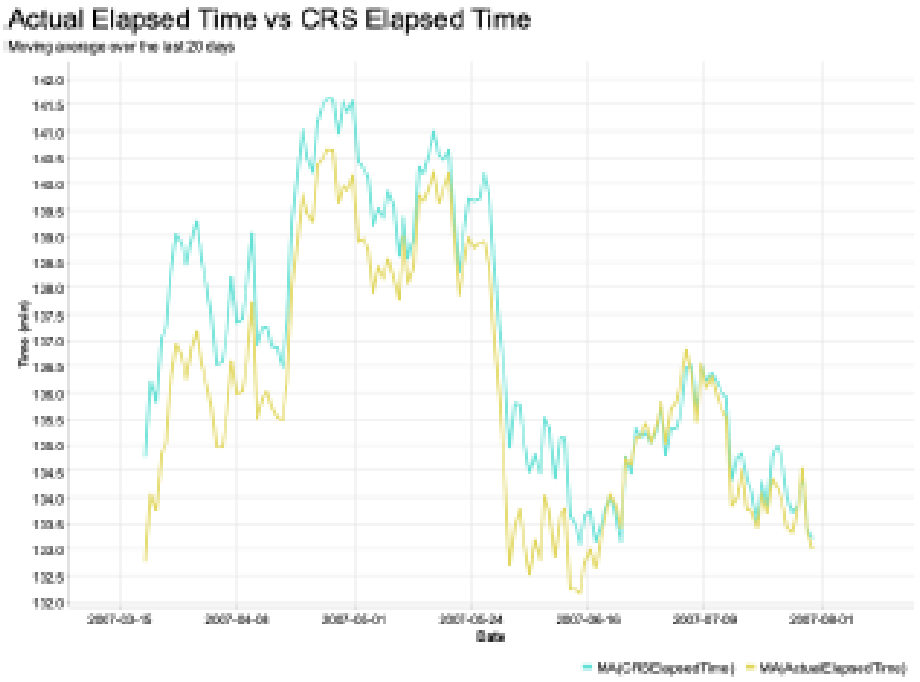


The process of data mining

Model Building

- selecting the appropriate data mining algorithm
- building a model that can identify patterns and relationships within the data

- the model is typically tested using a subset of the data to validate its accuracy and reliability



Data mining techniques (the most commonly used)

Classification:

- to classify data into predefined categories or groups
- based on specific attributes or characteristics
- can be used to predict the category of a new instance based on its features
- **classification model is trained on a labeled training dataset, containing examples of the different classes or categories.**

Data mining techniques (the most commonly used)

Classification:

- **The model learns the relationships between the input variables and the output classes, and then applies this knowledge to new, unseen data to predict the class or category to which it belongs.**

Algorithms to build classification model

Decision trees:

- Creates tree-like model of decisions and possible consequences

- most popular
- Each node in the tree represents a decision based on one of the input properties
- each branch represents a possible value of that feature
- leaves of the tree represent the predicted class or category.

Algorithms to build classification model

Logistic regression:

- uses a logistic function to model the probability of a binary outcome (e.g., true/false)
- often used in binary classification problems

Naive Bayes:

- calculates the probability of a data instance belonging to a particular class based on the values of its attributes
- It assumes that the attributes are independent of each other, which is often not true in real-world data.

Algorithms to build classification model

SVMs (support vector machines):

- powerful classification algorithm

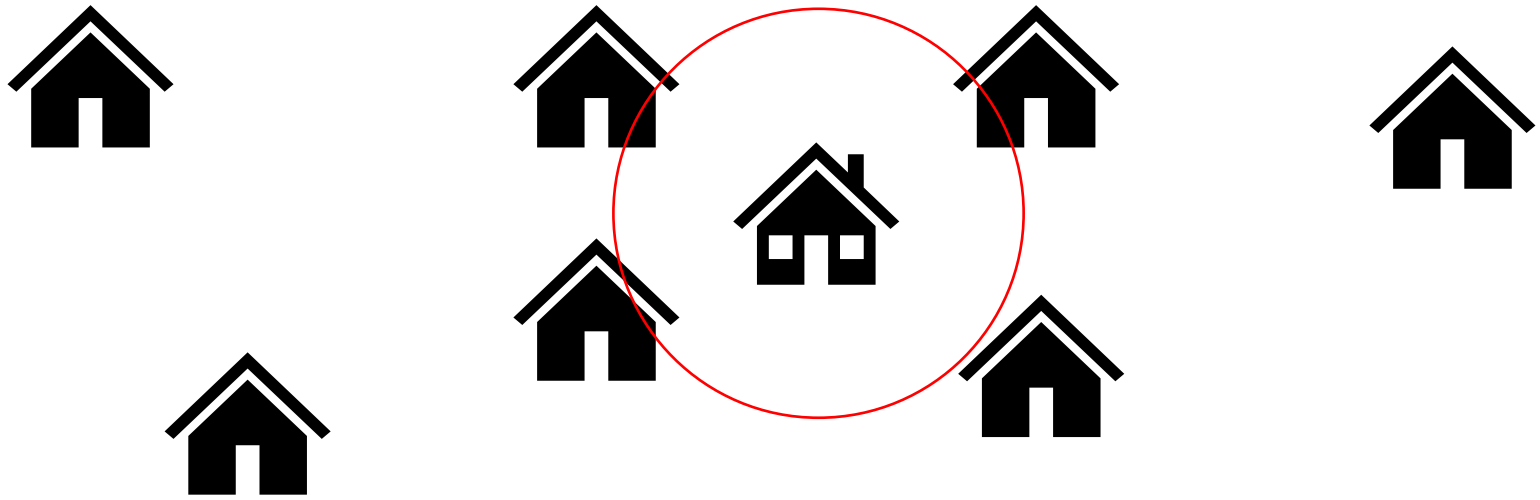
- can handle non-linear decision boundaries by transforming the input data into a higher-dimensional space

- They work by finding the hyperplane that maximizes the margin between the classes.

Algorithms to build classification model

k-NN (k-nearest neighbors):

- lazy learning algorithm that classifies data based on the classes of its k nearest neighbors in the training dataset



Data mining techniques (the most commonly used)

Clustering:

- **used to group similar data points or objects together based on their attributes or characteristics.**
- **The goal is to identify meaningful patterns or relationships in the data that may not be immediately apparent.**

- data points are grouped together based on their similarity or proximity to each other.
- The similarity measure depends on the type of data
- Metrics like Euclidean distance, Manhattan distance, cosine similarity, and Jaccard similarity.

Algorithms to build clustering model

Hierarchical clustering:

- starts with each data point as a separate cluster and iteratively merges them until all data points belong to the same cluster

K-means clustering

Density-based clustering:

- groups data points based on their density in a given area.

Data mining techniques (the most commonly used)

Regression:

- **to predict the relationship between a dependent variable and one or more independent variables**
- **supervised learning method**

requires labeled training data to build a model that can predict the values of the dependent variable based on the values of the independent variables

Algorithms to build regression model

Linear regression:

- fitting a straight line to the data points to model the relationship between the independent and dependent variables. It is often used in simple regression problems where there is only one independent variable.

Logistic regression:

- uses a logistic function to model the probability of a binary outcome (e.g., true/false).
- It is often used in binary classification problems

Algorithms to build regression model

Polynomial regression:

- fits a polynomial function to the data points to model the relationship between the independent and dependent variables. It can handle non-linear relationships between the variables.

Decision tree regression:

- fitting a decision tree to the data points to model the relationship between the independent and dependent variables. It is often used in problems where the data has a hierarchical structure

Data mining techniques (the most commonly used)

Association rule mining:

- **aims to discover the co-occurrence patterns and relationships between different items in a dataset**

- **finding the itemsets that occur frequently in the dataset and then generating rules that describe the relationships between the items.**

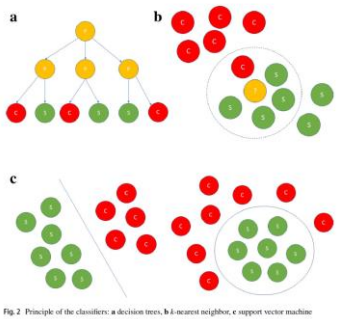
Data mining techniques (the most commonly used)

Outlier detection:

- **process of identifying data points that are significantly different from the rest of the data set**

Text mining:

- **analysing unstructured text data to identify patterns and relationships**



Neural networks:

- used to model complex relationships between variables

Applicability of data mining techniques

Classification: Customer segmentation, Fraud detection, Medical diagnosis, Image recognition

Clustering: Customer segmentation, Anomaly detection, Image segmentation, Recommendation systems, Social network analysis

Regression: Data forecasting, Financial modeling, Risk analysis, Marketing analysis, Environmental modeling

Association rule: Market basket analysis, Customer behavior analysis, Fraud detection, Healthcare analysis, Recommendation systems

Outlier detection: Fraud detection, Quality control, Anomaly detection, Network security, Medical diagnosis

Text mining: Sentiment analysis, Customer service, Market research, Fraud detection, Healthcare analysis, Content recommendation

Neural networks: Image and speech recognition, Financial forecasting, Medical diagnosis, Fraud detection, Customer behavior analysis, Autonomous vehicles

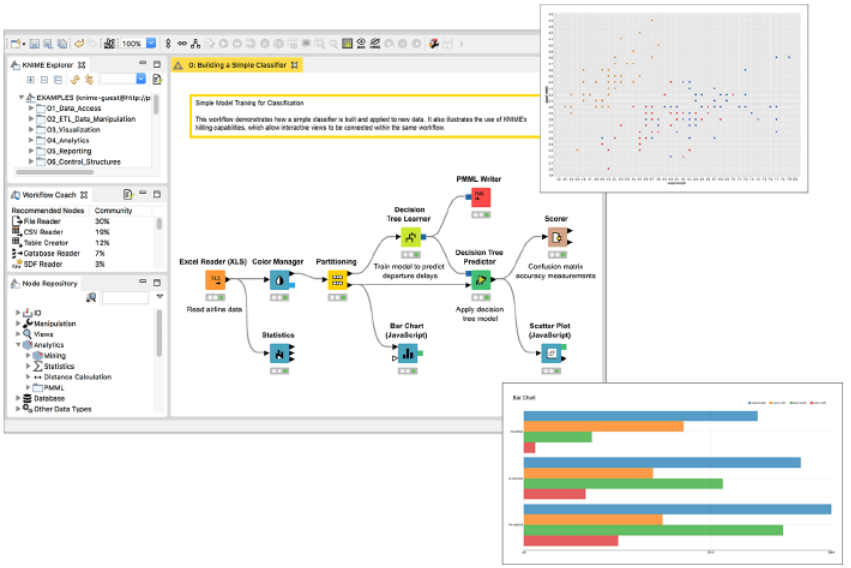
Data science tool for advance AI/ML models

Open for Innovation KNIME

Swiss army knife for data

What is KNIME Analytics Platform?

- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Many integrations, such as Java, R, Python, Weka, Keras, Plotly, H2O, etc.

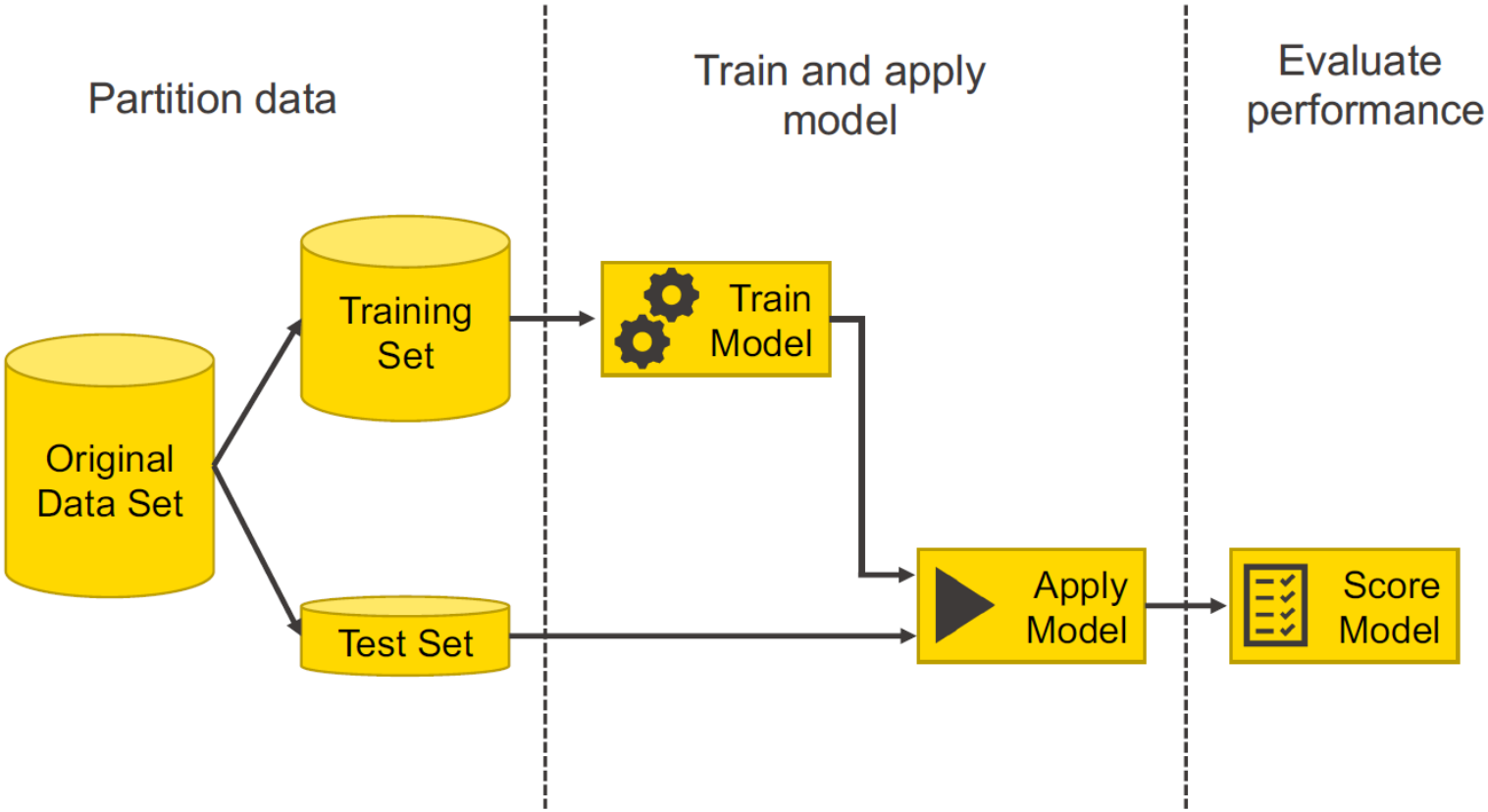


Data science tool for advance AI/ML models

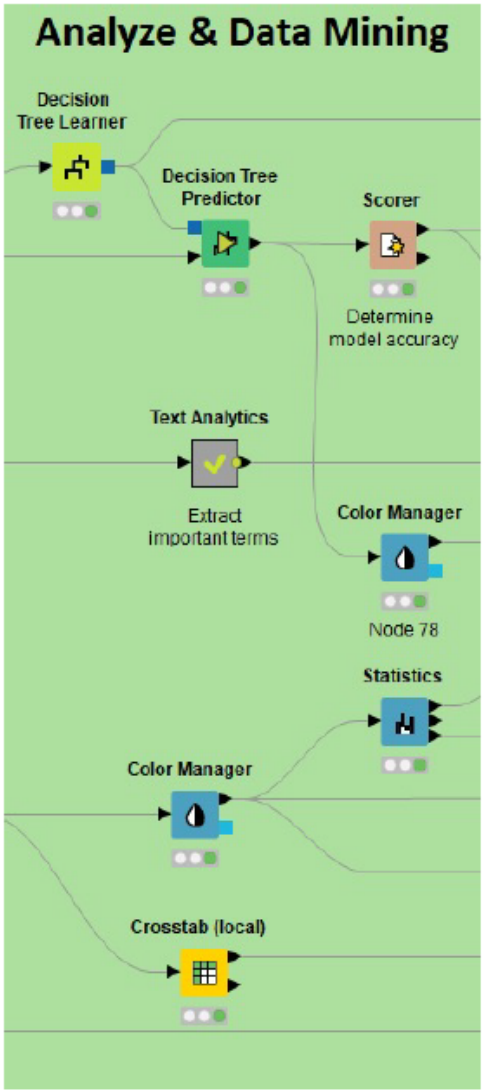
Open for Innovation KNIME

Swiss army knife for data

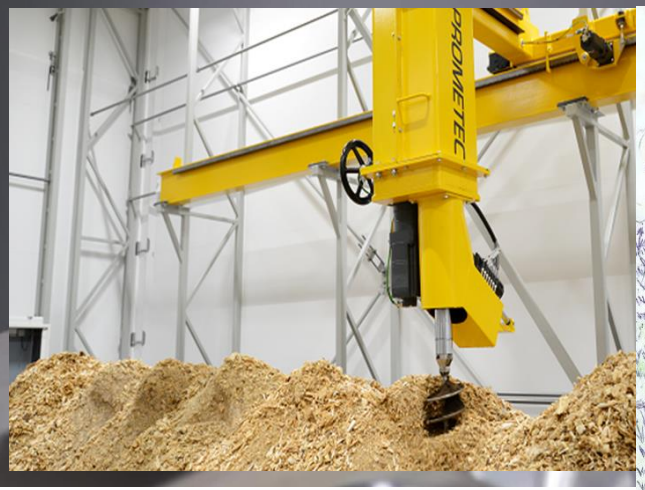
Data Mining: Process Overview



Data science tool for advance AI/ML models



- Regression
  - Linear, logistic
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Deep Learning
  - Keras, DL4J
- External
  - R, Python, Weka, H2O, Keras



Thank you for your attention

Prometec

TÉCNICO LISBOA

DANISH TECHNOLOGICAL INSTITUTE

VERDO

TÜBİTAK UME

IMBIH

CETIAT
ensemble, innover et valider

PTB
Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin

THE FRESNEL COMPANIES

ČESKÝ METROLOGICKÝ INSTITUT

biofMET

EMPIR

EURAMET

The EMPIR initiative is co-funded by the European Union's Horizon 2020 research and innovation programme and the EMPIR Participating States

Stakeholder workshop
28.-29. March 2023