MOLGROUP

Complex Systems Monitoring
Research Group

Neurális hálózat alapú megoldások az iparban

a sikeres alkalmazások feltételei

Hanzelik Pál, Kummer Alex, **Abonyi János**

janos@abonyilab.com

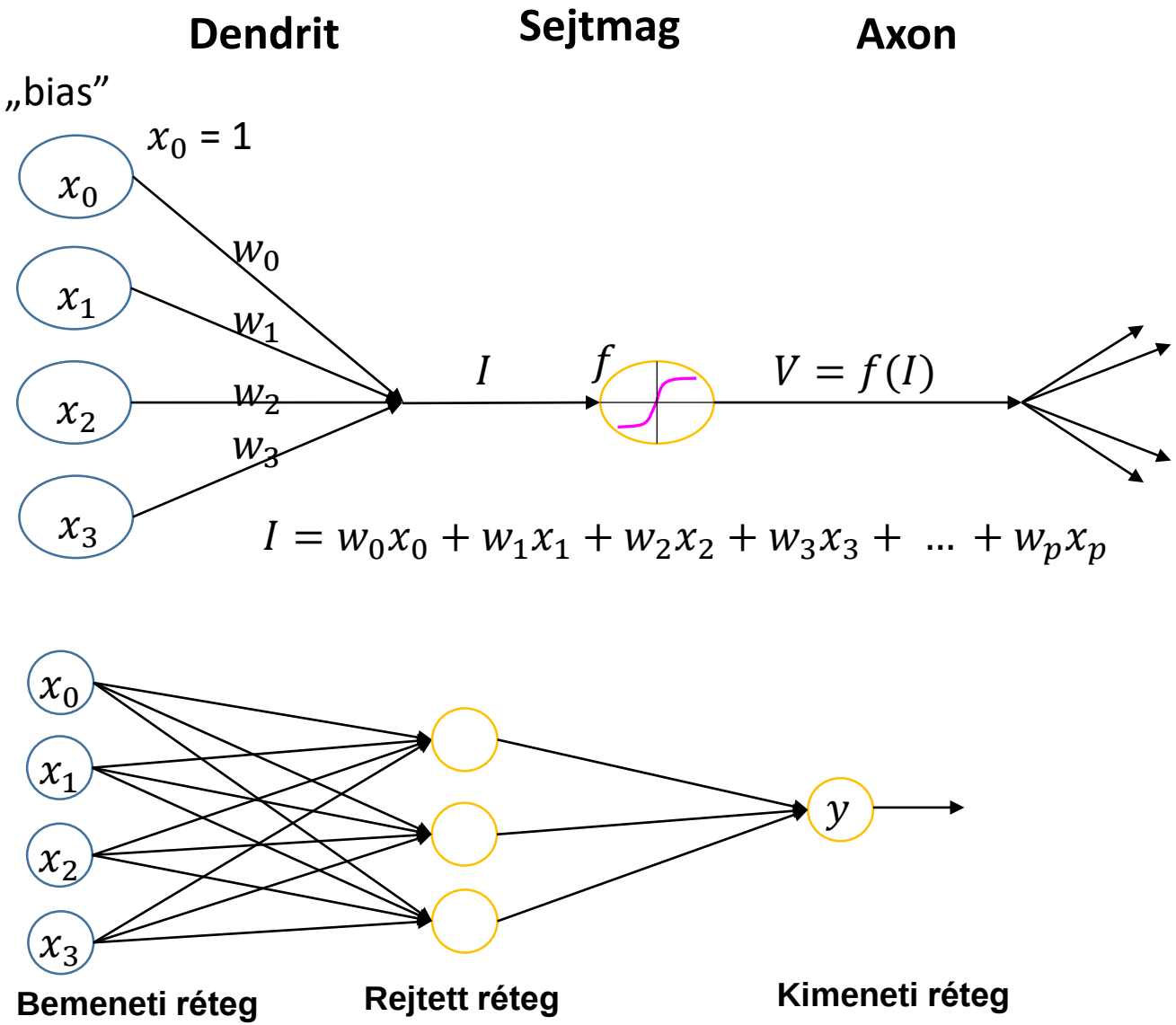
dataSTREAM 2023 - Networking

Pannon Egyetem
University of Pannonia

P. P. Hanzelik, A. Kummer, J. Abonyi (2022): Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development. *Sensors*. 22. 4268.

Egyszerű és nagyszerű eszköz …

Már csak ennyi? Mi a kihívás?



Write a python code for training a neural network with one hidden layer with 15 neurons. The training data should be loaded from an excel file.

Sure! To train a neural network with one hidden layer and 15 neurons using training data loaded from an Excel file, you can use the `pandas` library to read the data from the Excel file and the `scikit-learn` library to build and train the neural network. Here's an example code:

python — Copy code

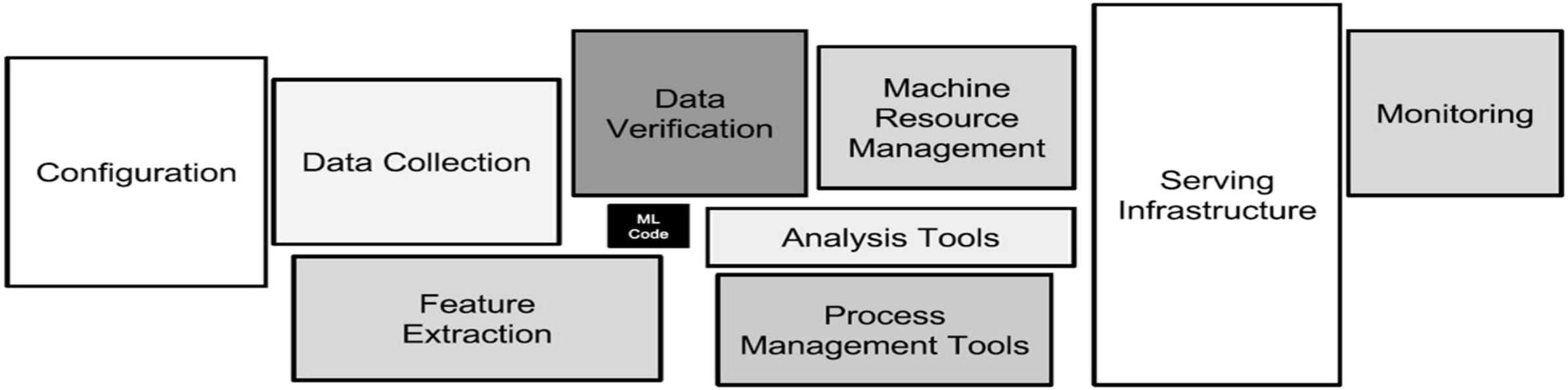
```python
import pandas as pd
from sklearn.neural_network import MLPClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Load the data from the Excel file
data = pd.read_excel('training_data.xlsx')

# Split the data into features (X) and labels (y)
X = data.drop('label_column_name', axis=1)  # Replace 'label_column_name' wi
y = data['label_column_name']  # Replace 'label_column_name' with the actual

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, ran

# Standardize the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Create and train the neural network
model = MLPClassifier(hidden_layer_sizes=(15,), random_state=42)  # One hidd
model.fit(X_train_scaled, y_train)

# Evaluate the model
accuracy = model.score(X_test_scaled, y_test)
print("Accuracy:", accuracy)
```

Csak ennyi! A kihívás máshol van. De ott temérdek …

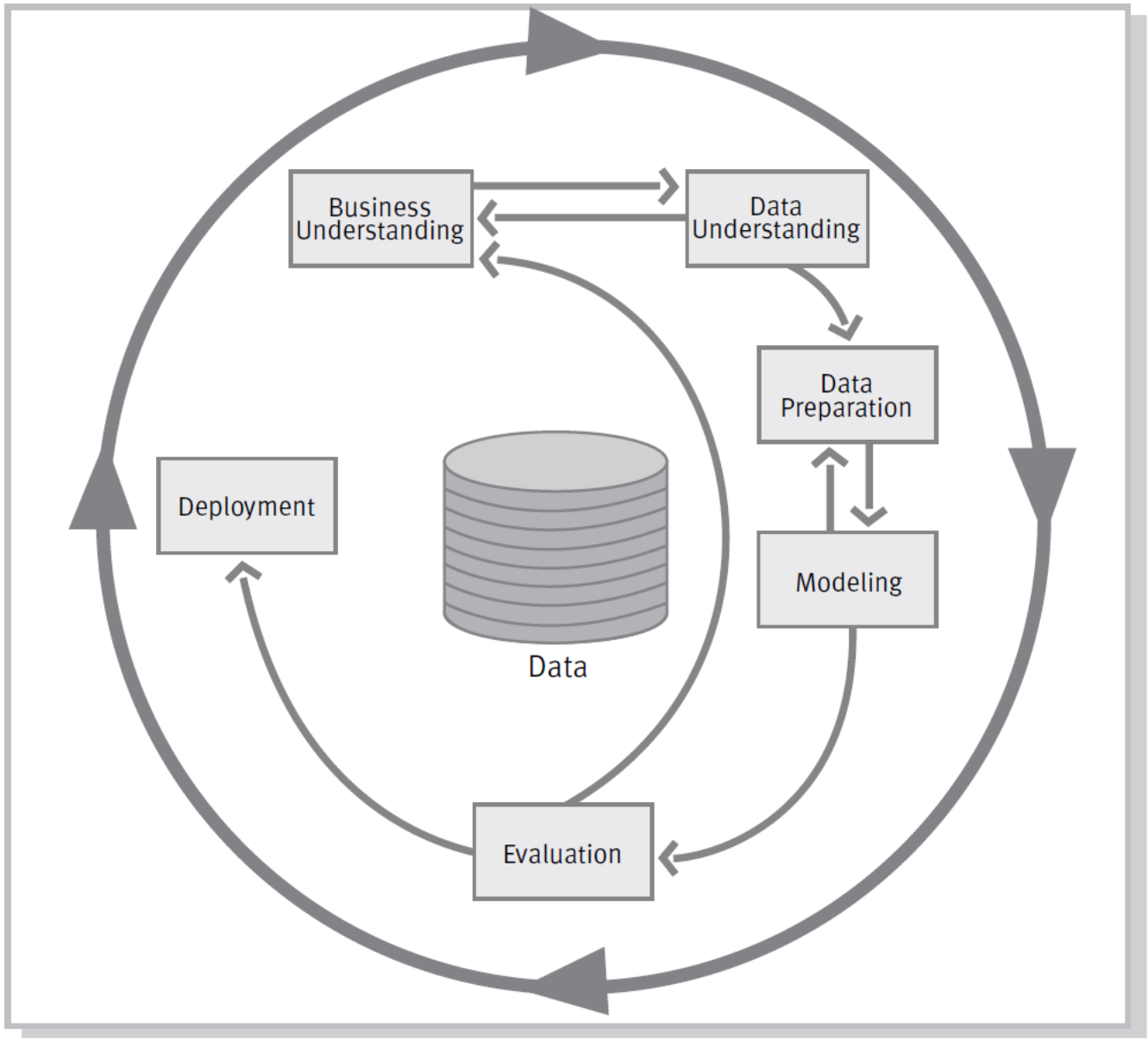
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

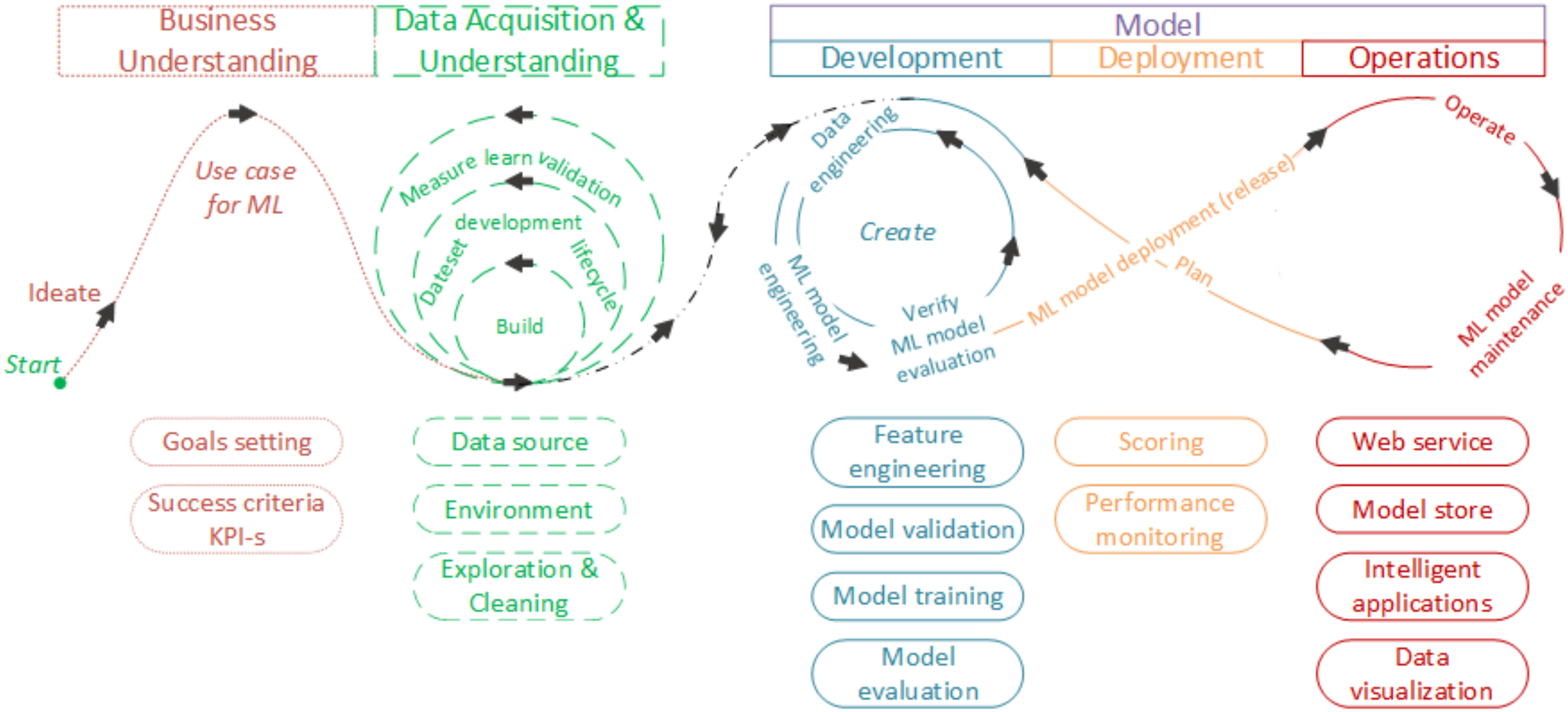


A kihívás visszavezethető a modellezési feladatra

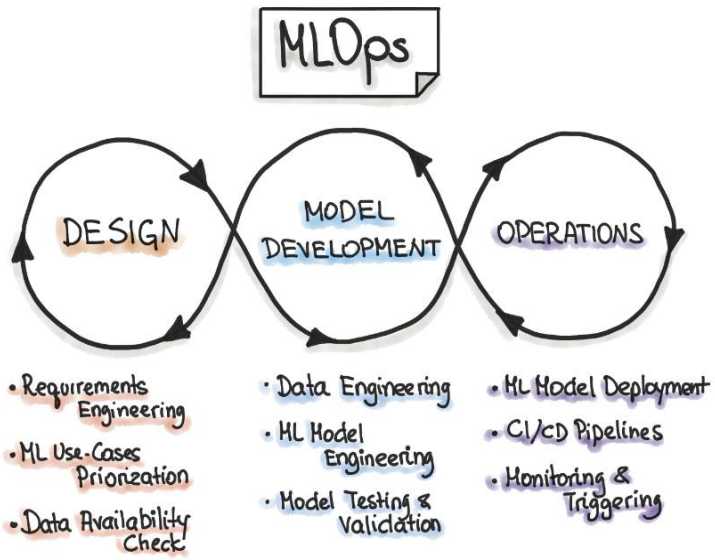


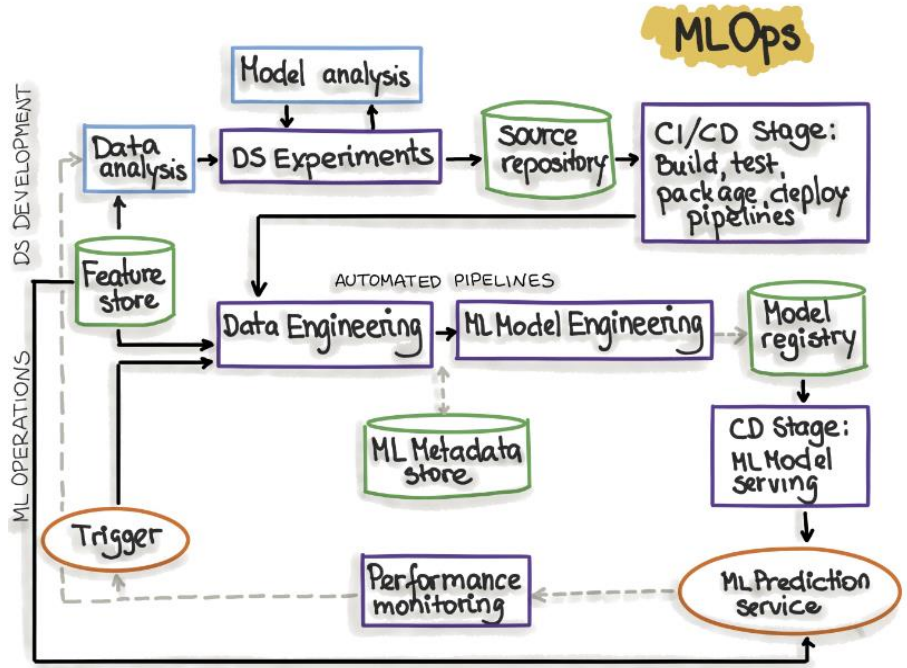
CRISP-DM (Cross Industry Strandard Process for Data Mining) was conceived in late 1996 by three “veterans” of the young and immature data mining market. DaimlerChrysler (then Daimler-Benz) was already ahead of most industrial and commercial organizations in applying data mining in its business operations. SPSS (then ISL) had been providing services based on data mining since 1990 and had launched the first commercial data mining workbench—**Clementine®**—in 1994

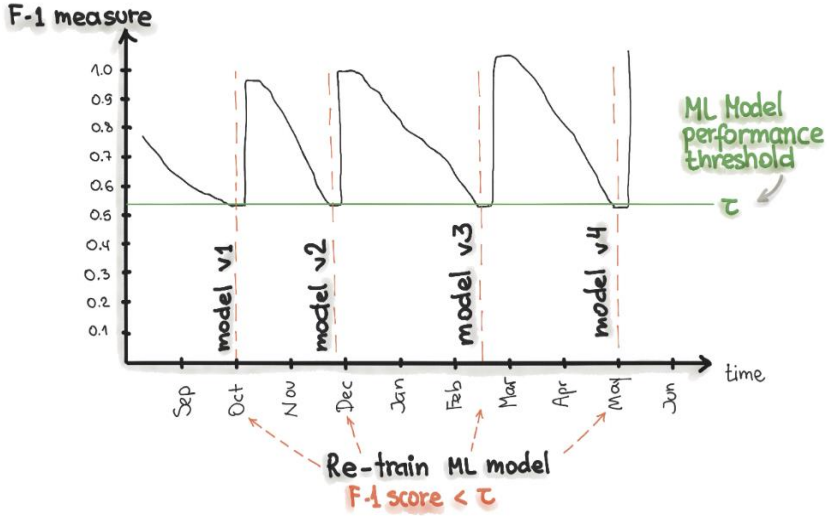
Amely feladat összetettsége egyre nyilvánvalóbb … (lásd CRISP-ML)



Annyira, hogy szabványosítani és automatizálni érdemes … (MLOps)

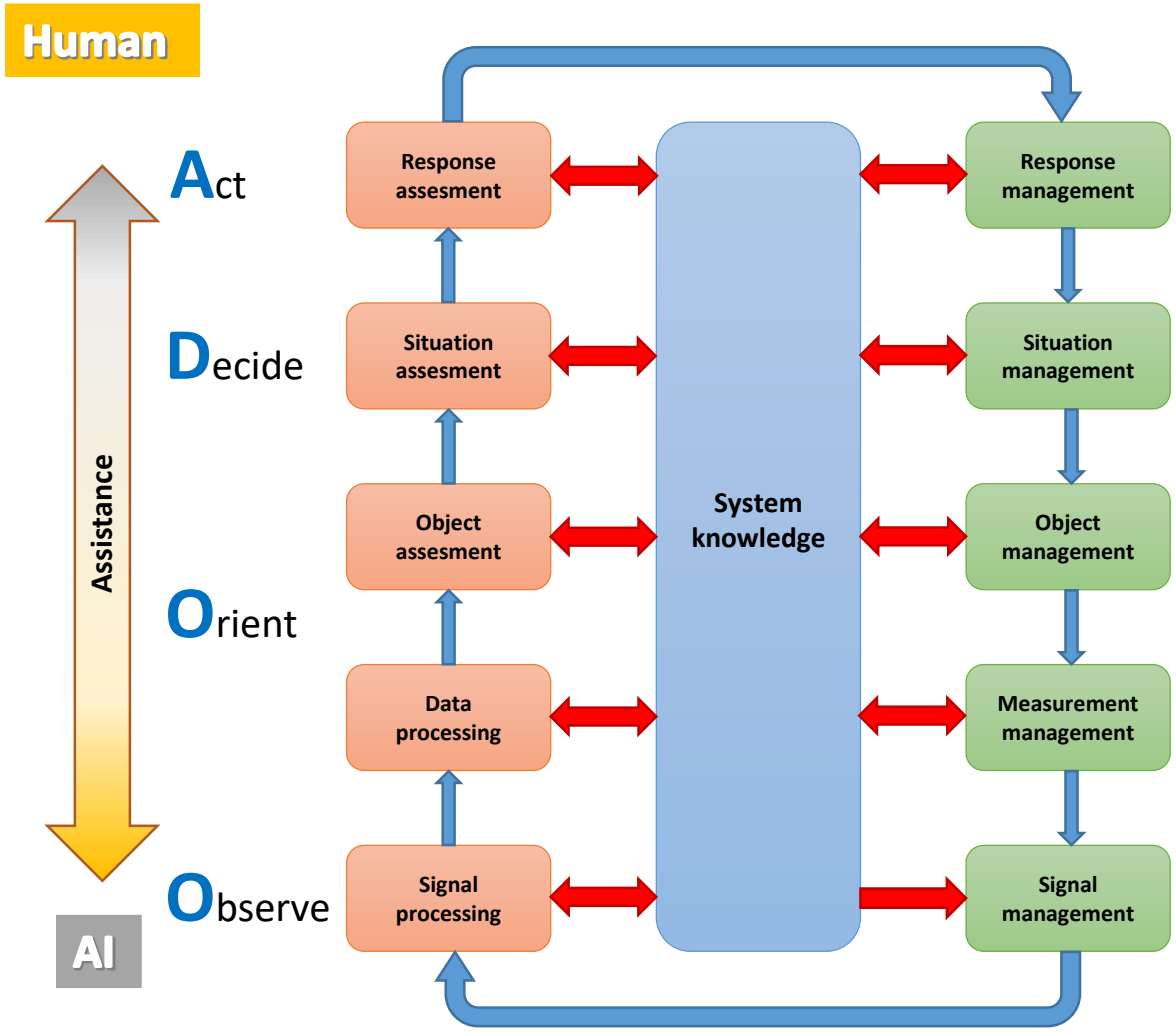


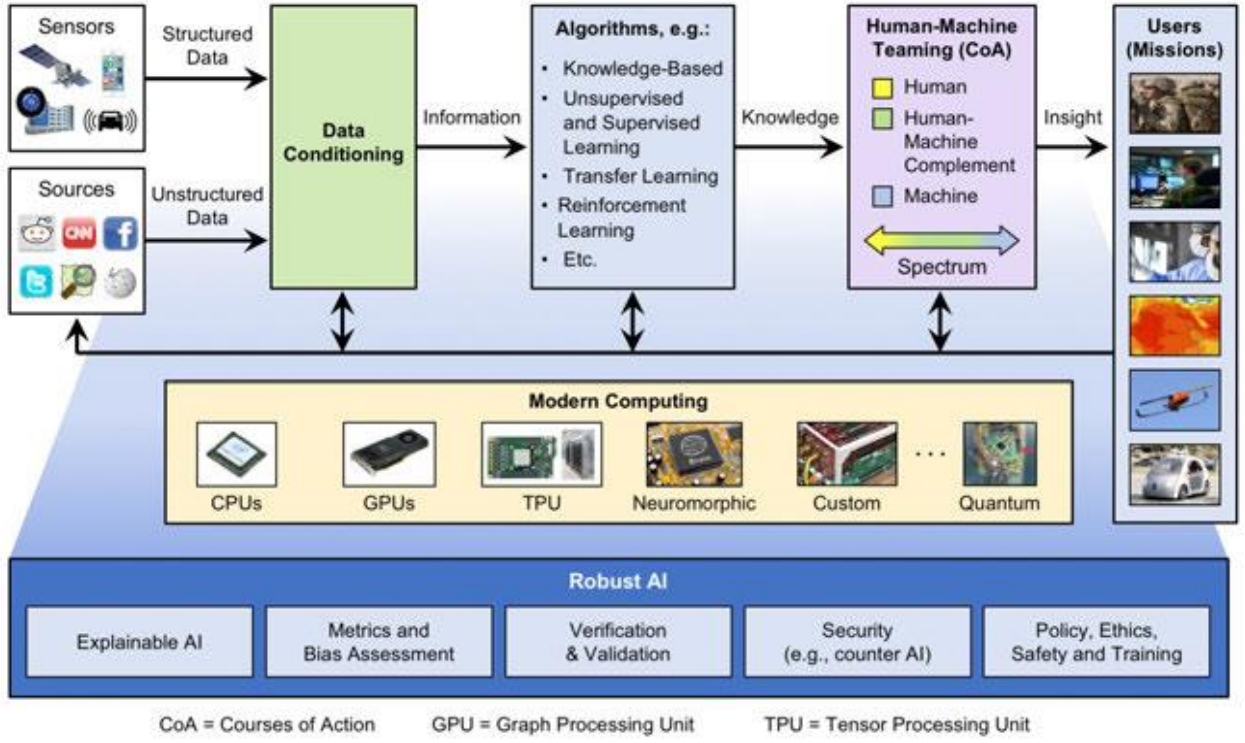




A kihívás az alkalmazás körülményeire is visszavezethető

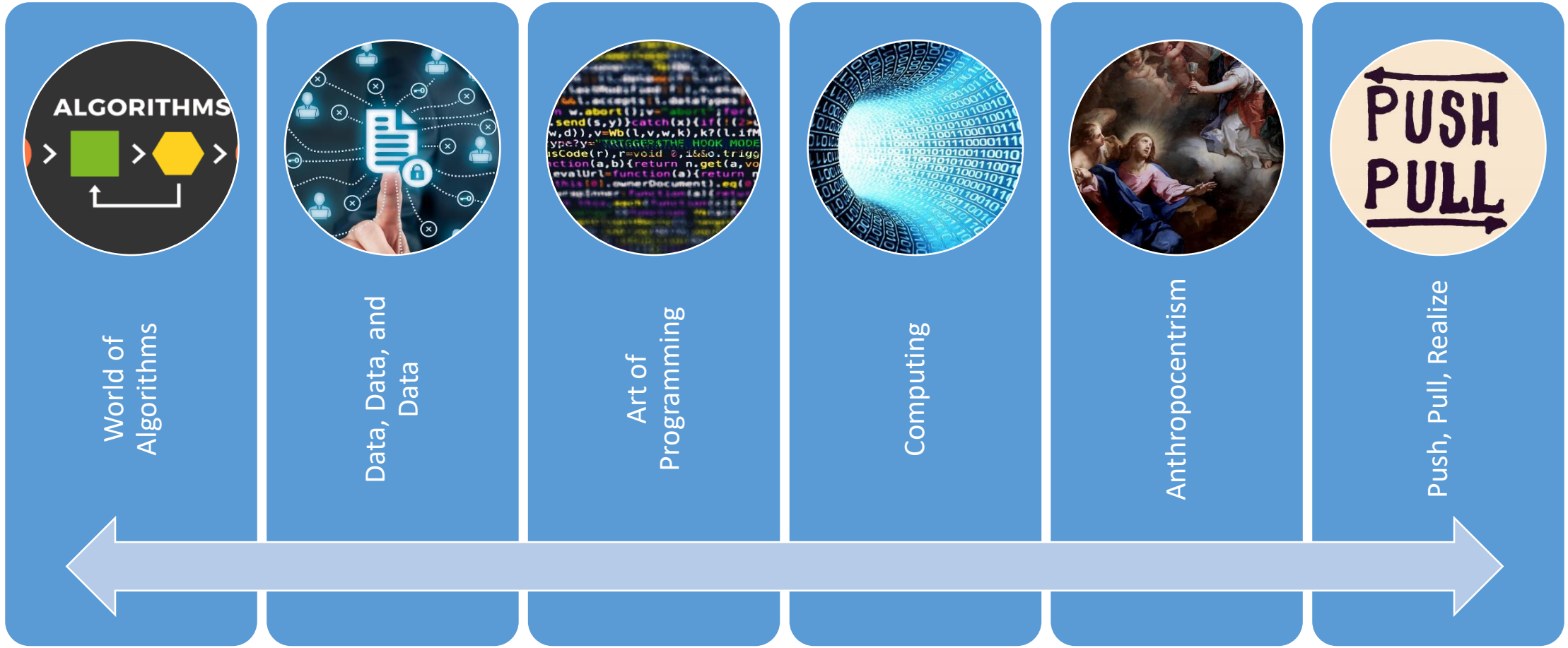
Az AI/ML alkalmazásai tipizálhatók … így az algoritmusok környezete is.





MI a mi megközelítésünkben

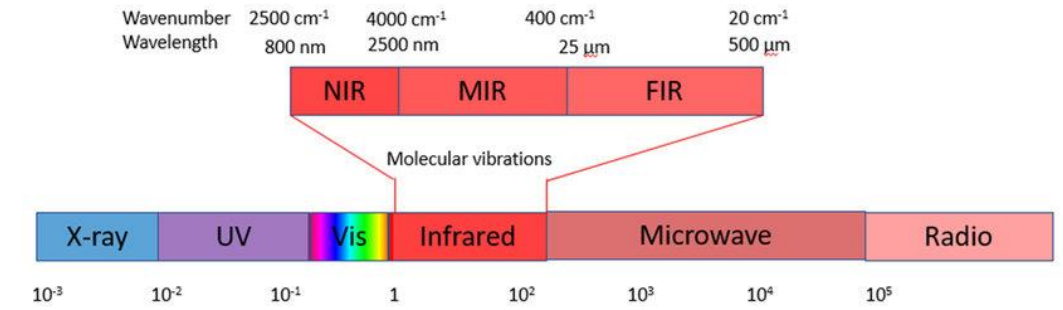
MI megoldásokat fejlesztő szakmérnök: www.ai-academy.hu
Ipar 4.0 megoldásokat fejlesztő szakmérnök: www.ipar4.org

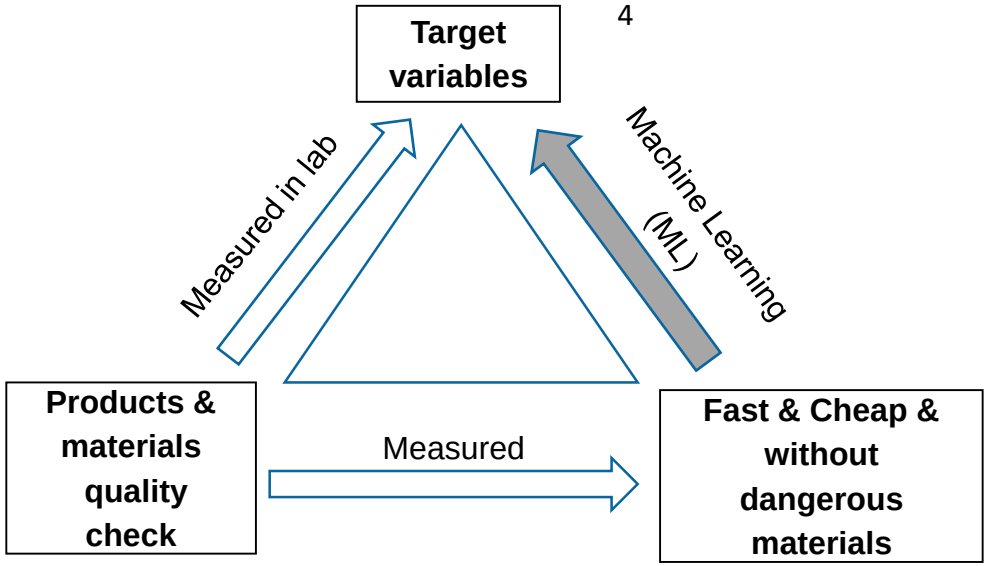


Esettanulmány … mintarendszer a MOL-ból

The qualification of each material is carried out by determining several essential parameters. Unfortunately, traditional measurements are slow, expensive and contain dangerous substances.

e.g. Infrared spectroscopy is fast, one of the cheapest and chemical-free





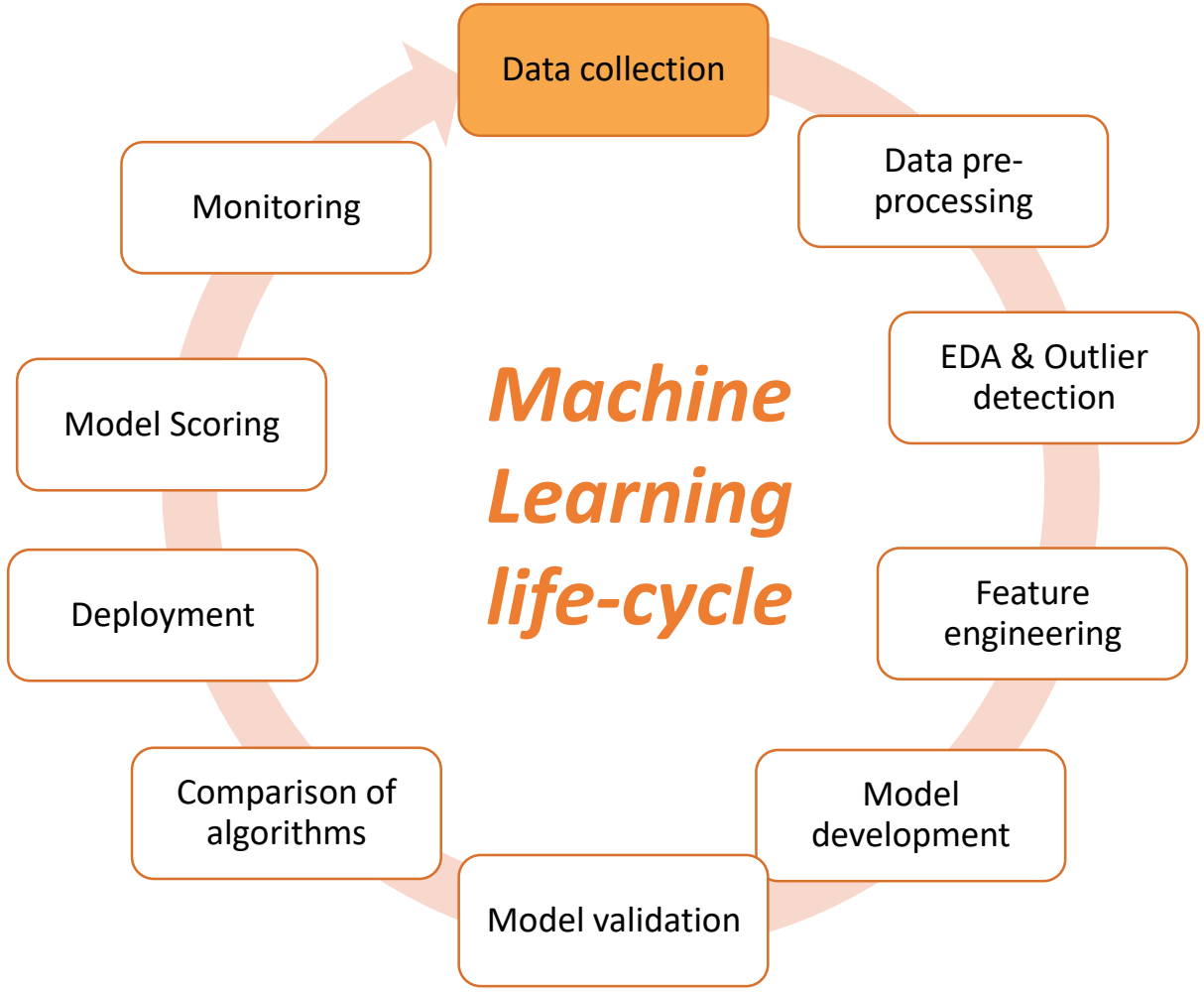
Data collection

Collection of laboratory data

- Use of island-like devices and edge computing devices
- Collection through 4G network, storage in a private company cloud

Collecting huge amount of data quickly

- Continuous data connection with the laboratory information system (LIMS)
- Development of reports according to the needs of operators, engineers and managers (decision support)

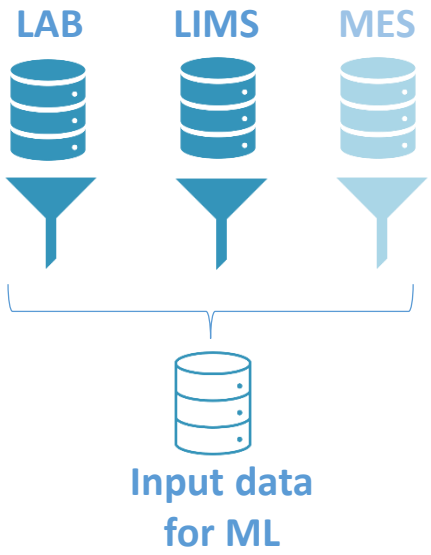


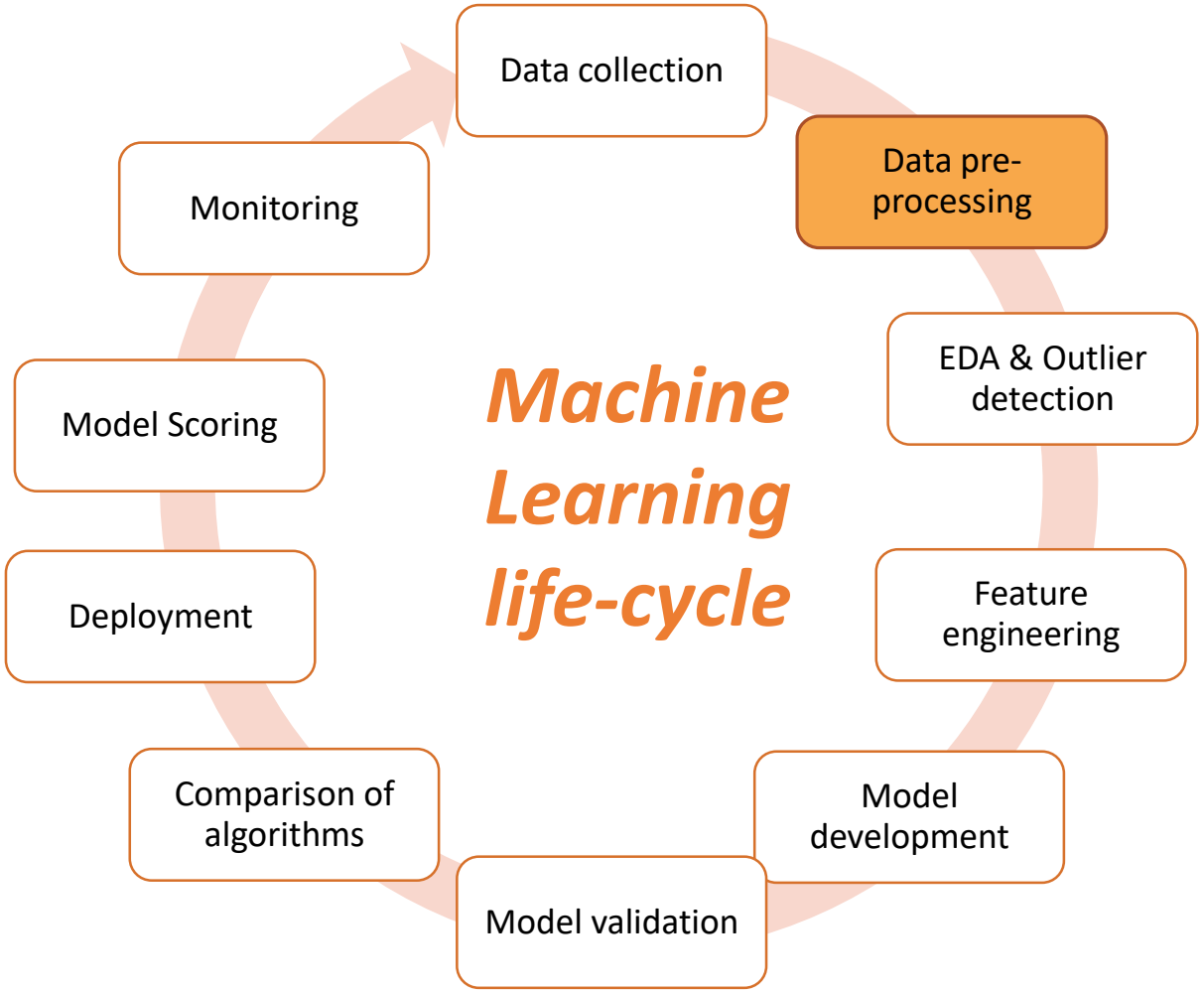
Data pre-processing

Data cleaning
- Missing data
- Noisy data

Data integration
- Laboratory raw data
- LIMS data
- (MES data)

Data transformation
- Normalisation
- Smoothing
- Derivation





EDA & Outlier detection & Feature engineering

Exploratory Data Analysis

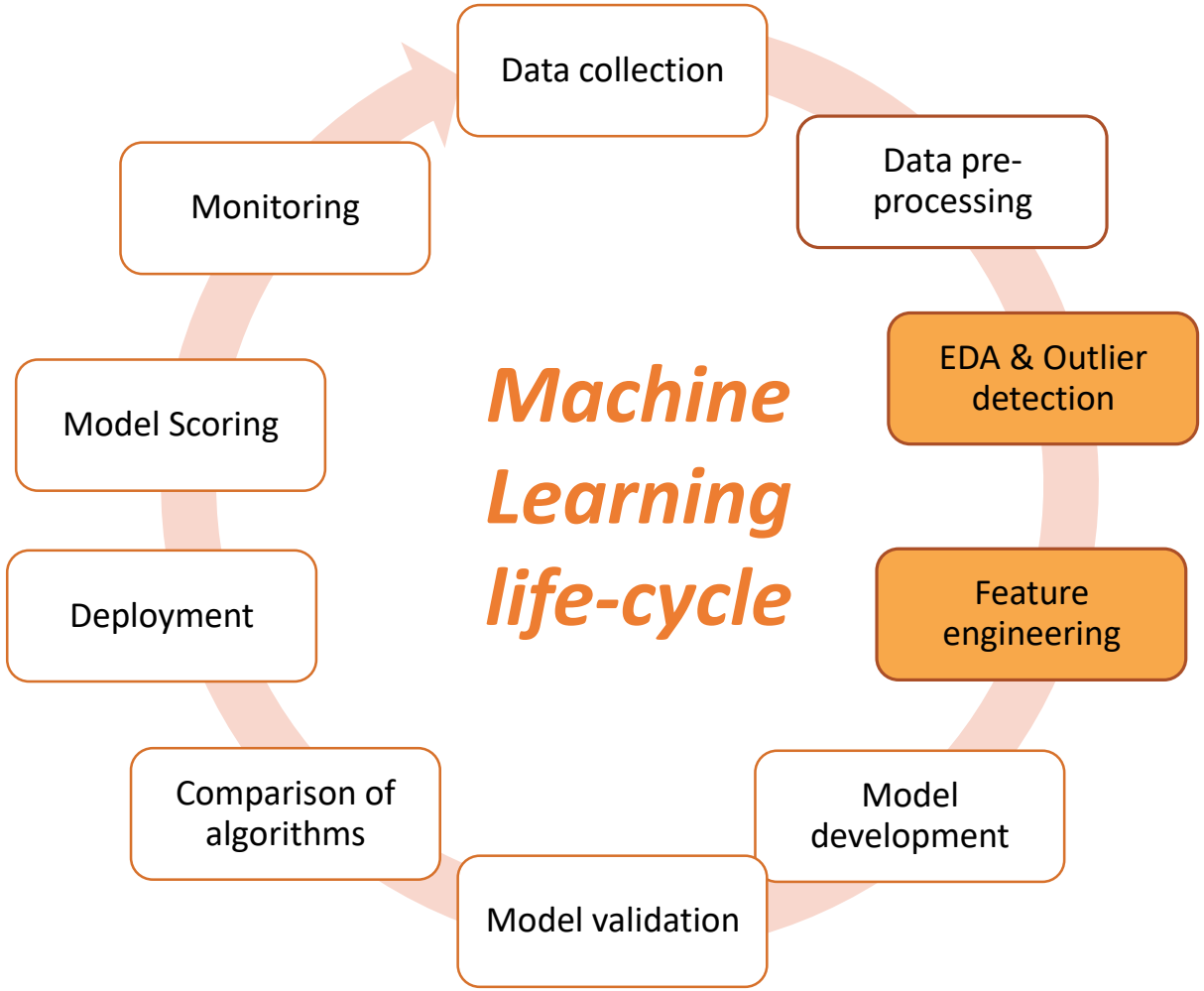
- Preparation
- Base statistic analysis
- Plot generation

Outlier detection

- Dimension reduction with Principal Component Analysis

Feature engineering

- Spectrum range selection
- New features generation



ML development & validation

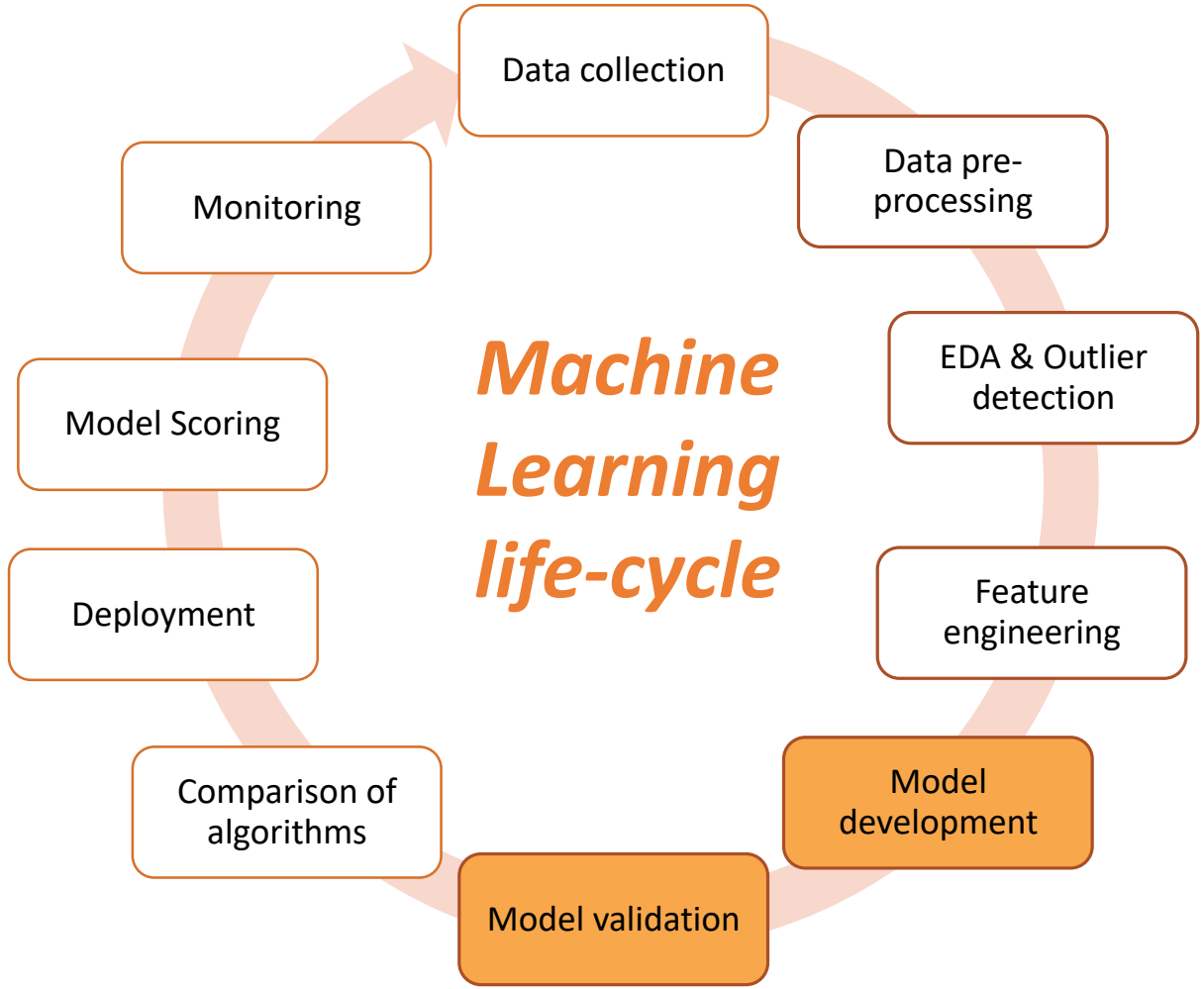
Supervised learning:

- PLSR (Partial Least Squares Regression)
- RFR (Random Forest Regression)
- XGBR (Extreme Gradient Boosting Regression)

Hyper-parameter tuning

Testing with optimized parameters

Testing the model on unknown samples



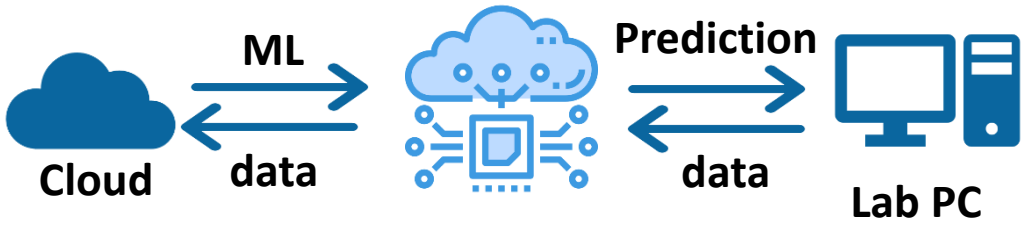
ML algorithms development

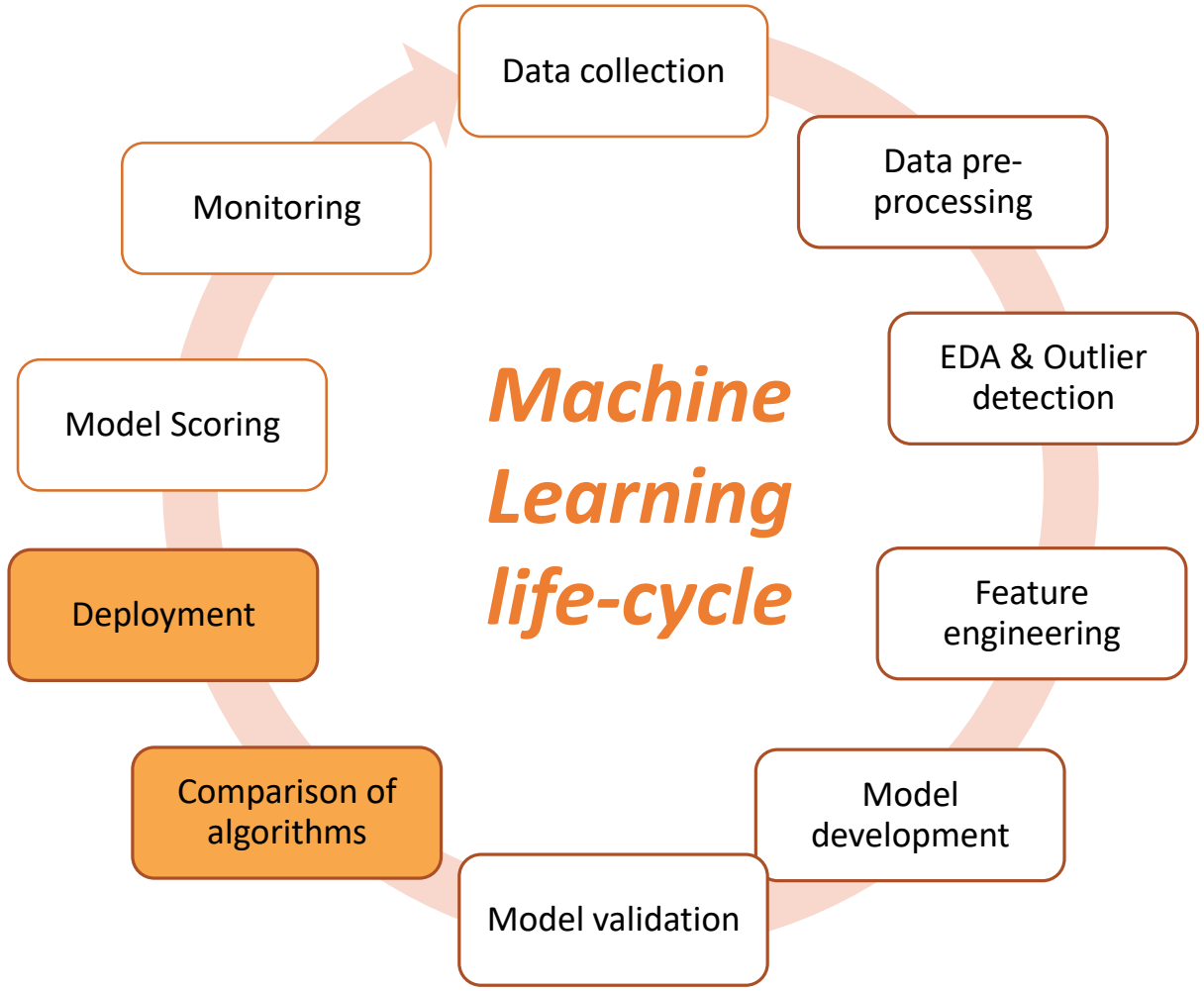
Continuous ranking of ML models

Displacement of the best model

Target variables	RMSE		R^2		RPD	
	10-cv	perf.	10-cv	perf.	10-cv	perf.
PLSR	0.010	0.035	0.999	0.975	57.73	6.36
RFR	0.089	0.084	0.972	0.929	5.98	3.77
XGBR	0.005	0.112	0.999	0.747	31.62	1.98

10-cv: 10-fold cross-validation
perf.: results of the performance dataset (unknown samples)



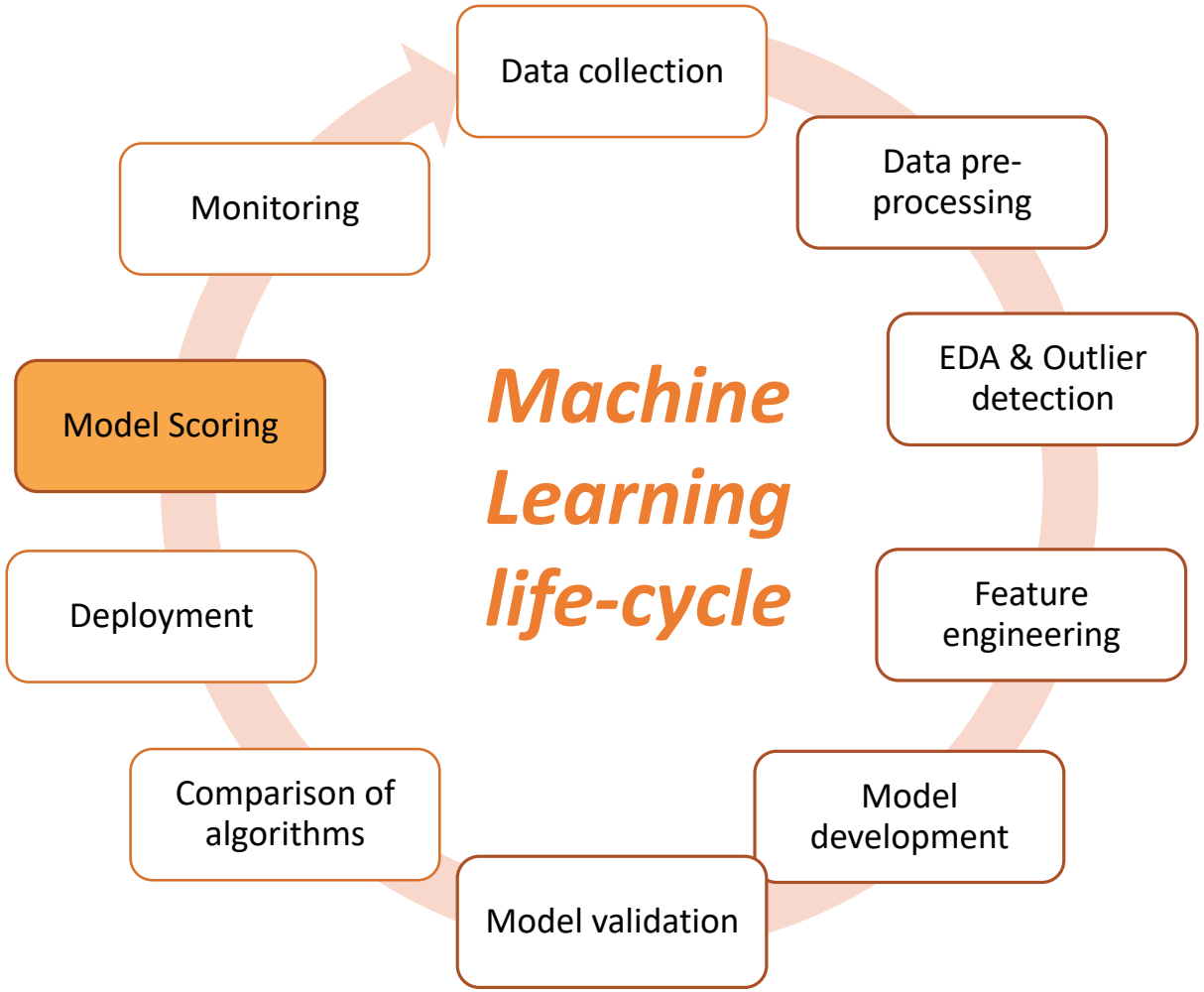


ML performance metrics

Several indicators must be continuously monitored to ensure the good performance of the models

Potential KPIs:

- RMSE (root mean squared error)
- RPIQ (ratio of performance to interquartile)
- R^2 (coefficient of determination)
- RPD (residual prediction deviation)
- NSE (Nash-Sutcliffe efficiency)



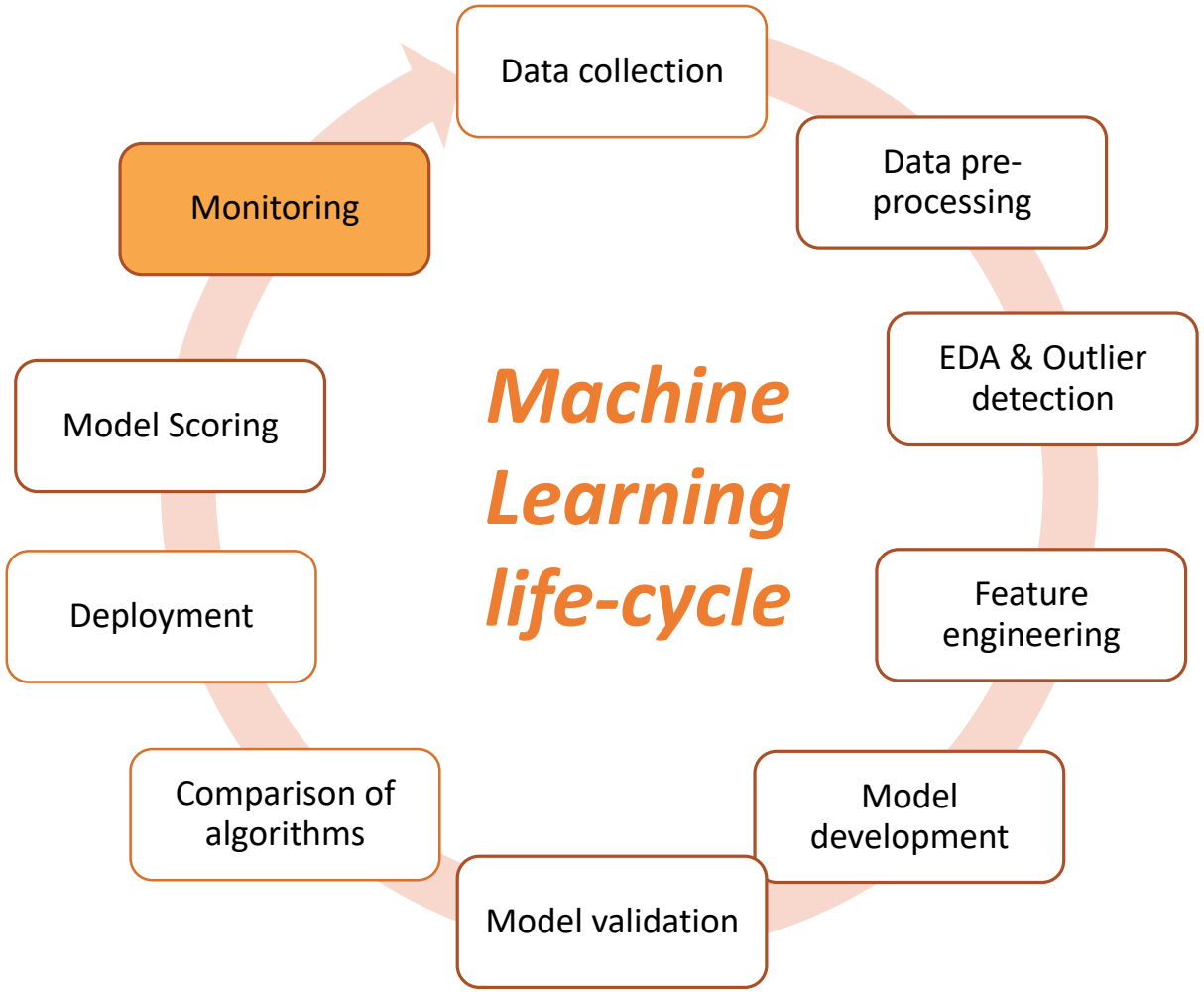
Performance monitoring of ML

The performance of ML models may deteriorate

The cause of the deterioration can change over time

Time constant of the performance change

- Fast vs. slow change of the performance refers to different root causes



Performance monitoring of ML

Performance indicators take a different value from the usual values

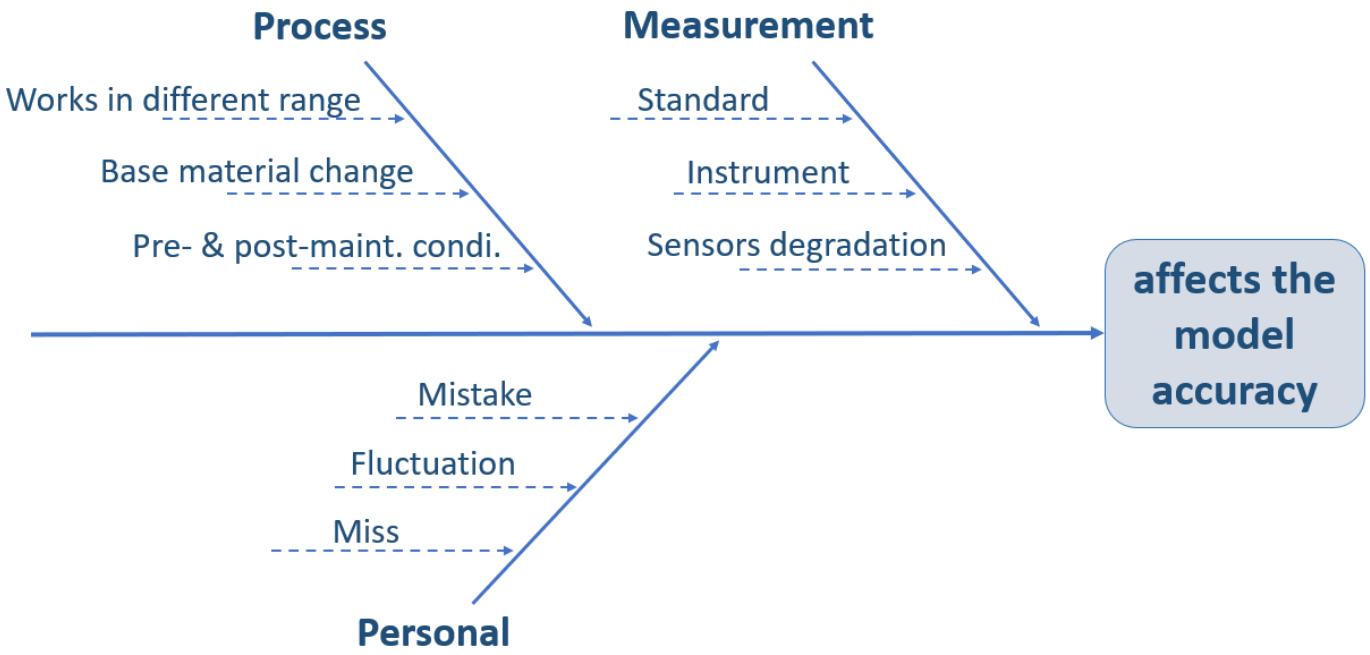
- Only one indicator changes
- All indicators change continuously over time, the values gradually deteriorate, and it is no longer suitable after a particular time.

When do we classify a model as no longer suitable?

- What do we do with it?
- What work order do we issue for it?
- If the model no longer works, what options are there to manage the model

Development of Statistical Process Control (SPC) processes - business process development

- Active learning
- Augmented data generation



Visualization of ML models

We need to visualize metrics and model performance over time.

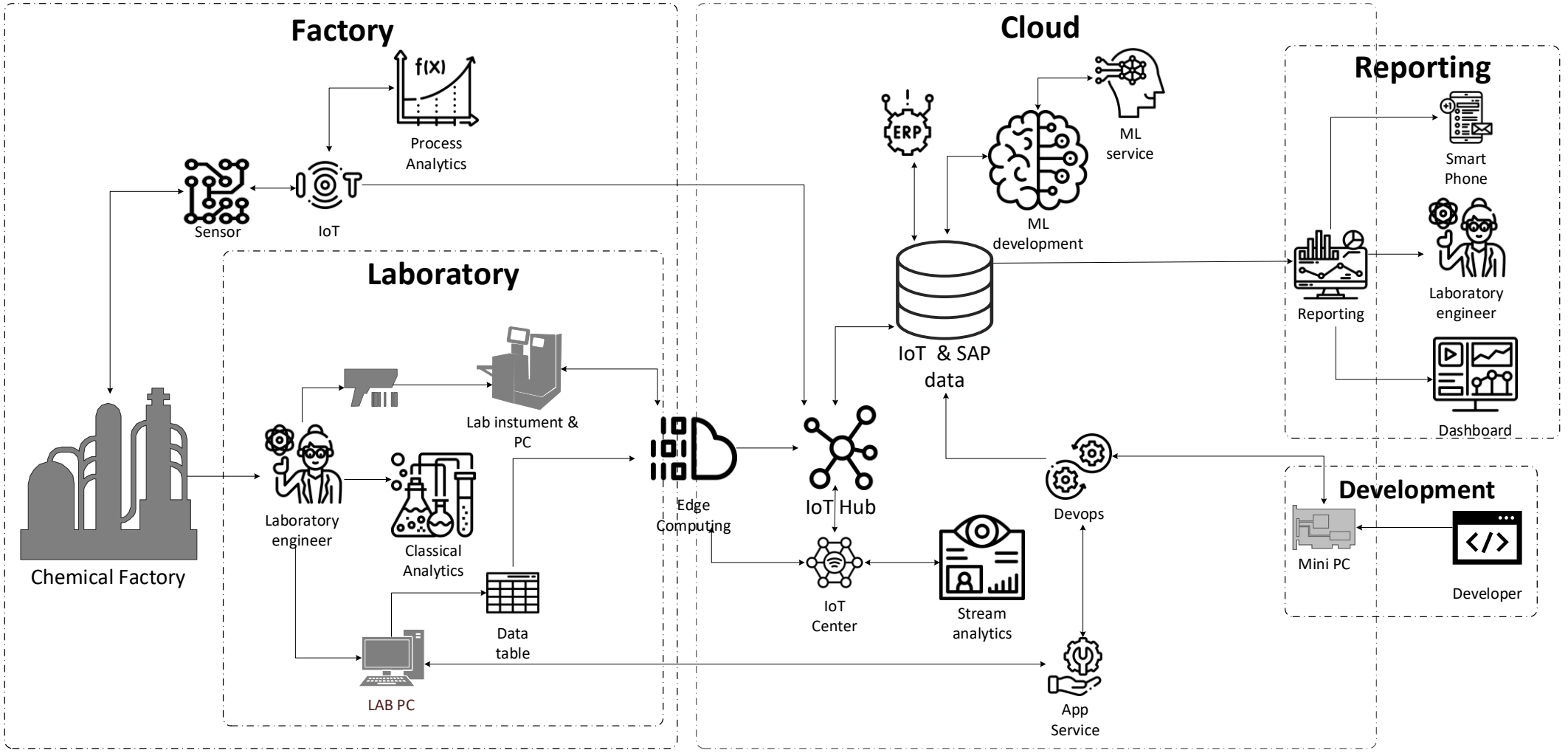
The performance of ML models can decrease over time, which can be caused by several things.

Changes in performance must be detected in time, and production must be intervened if necessary.

It is essential to consider several indicators and perform a sensitivity test.

The aim is to reveal which production change has what effect and to what extent each indicator.

Industrial architecture for ML application



[3] P. P. Hanzelik, A. Kummer, J. Abonyi (2022): Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development. *Sensors*. 22. 4268.

Lessons learned

The development of ML algorithms and their operation is a matter of research at MOL

Development project, development of appropriate methodology

- Indicators should be monitored
- Maintenance of models should be continuous
- It must be managed at the system level
- It requires expertise
- Compliance must be qualified
- Facilitating decision support

Let's face it bravely

Complex Systems Monitoring
Research Group

MOLGROUP

Thank you very much for your kind attention!