Enhancing cluster analysis with explainable AI and multidimensional cluster prototypes

03.03.2022

Michał Kuk - *AGH University of Science and Technology,* Kraków, Poland
Szymon Bobek - *Jagiellonian University*, Kraków, Poland
Maciej Szelążek - *AGH University of Science and Technology,* Kraków, Poland
Grzegorz J. Nalepa- *Jagiellonian University*, Kraków, Poland

Presentation plan

1. Introduction into explainable artificial intelligence (XAI)
   - Reasons for XAI application
   - XAI methods
   - Challenges

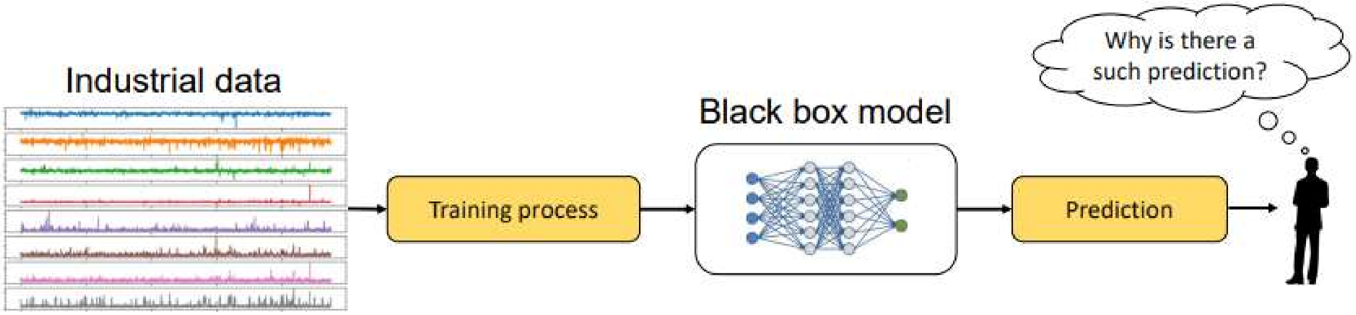
2. Developed methodology – Cluster Analysis with Multidimensional Prototypes (**CIAMP**)

3. Applications
   - Artificial datasets
   - Industrial case – Hot rolling process
   - Industrial case (preliminary study) – oil & gas well production management

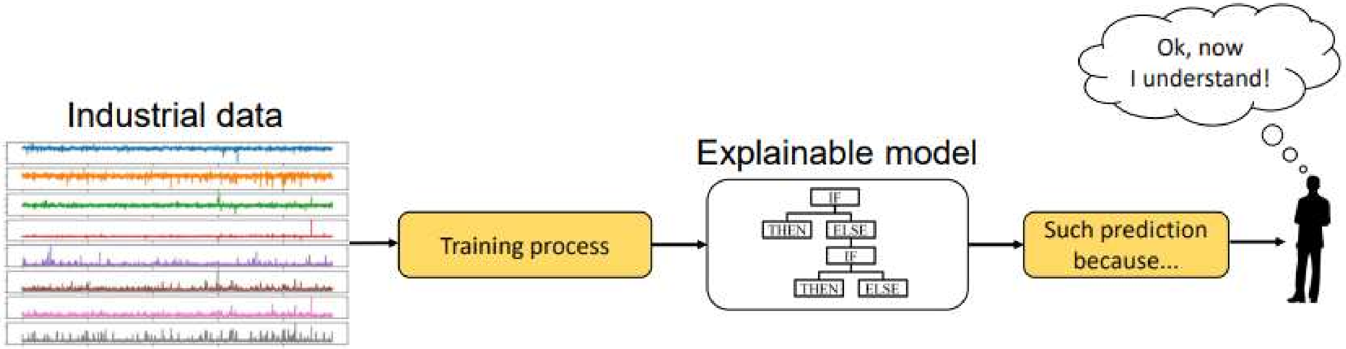
4. Summary

Introduction into explainable artificial intelligence (XAI)

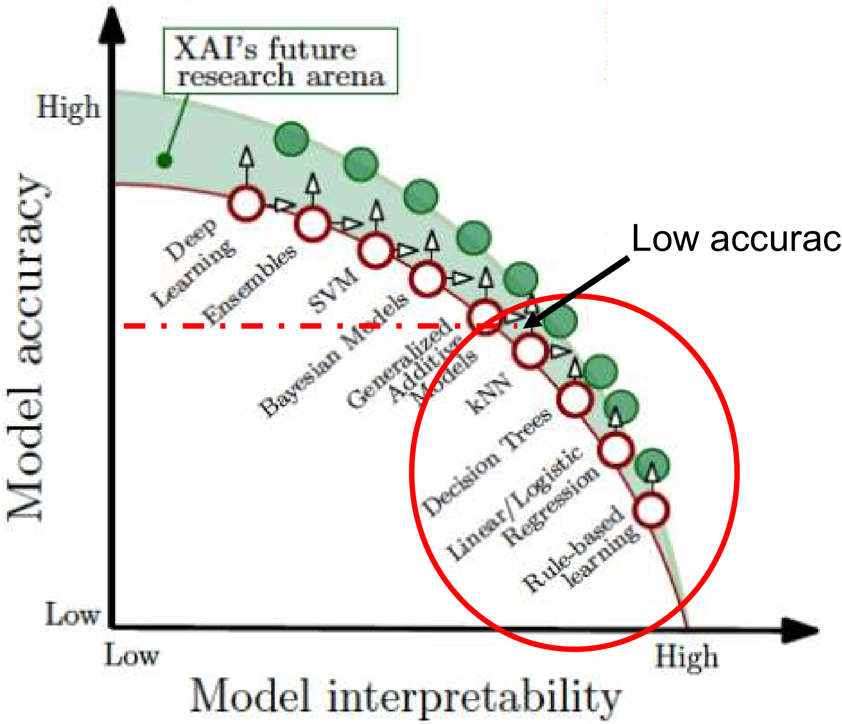
Questions:
Why?
Why not?
When you succeed?
When you fail?
How?



Answers:
I understand why
I understand why not
I know when you succeed
I know when you fail
I know how to correct



Introduction into explainable artificial intelligence (XAI)



Low accuracy

Source: Trade-off between model interpretability and accuracy, (Arrieta, Del Ser et al; 2019)

Learnings techniques examples:

1. Glass box
   - Decision Trees
   - Linear Logistic Regression
   - Rule-based learning
2. Black box
   - Deep learning
   - SVM
   - Generalized Additive Models (GAN)

High accuracy

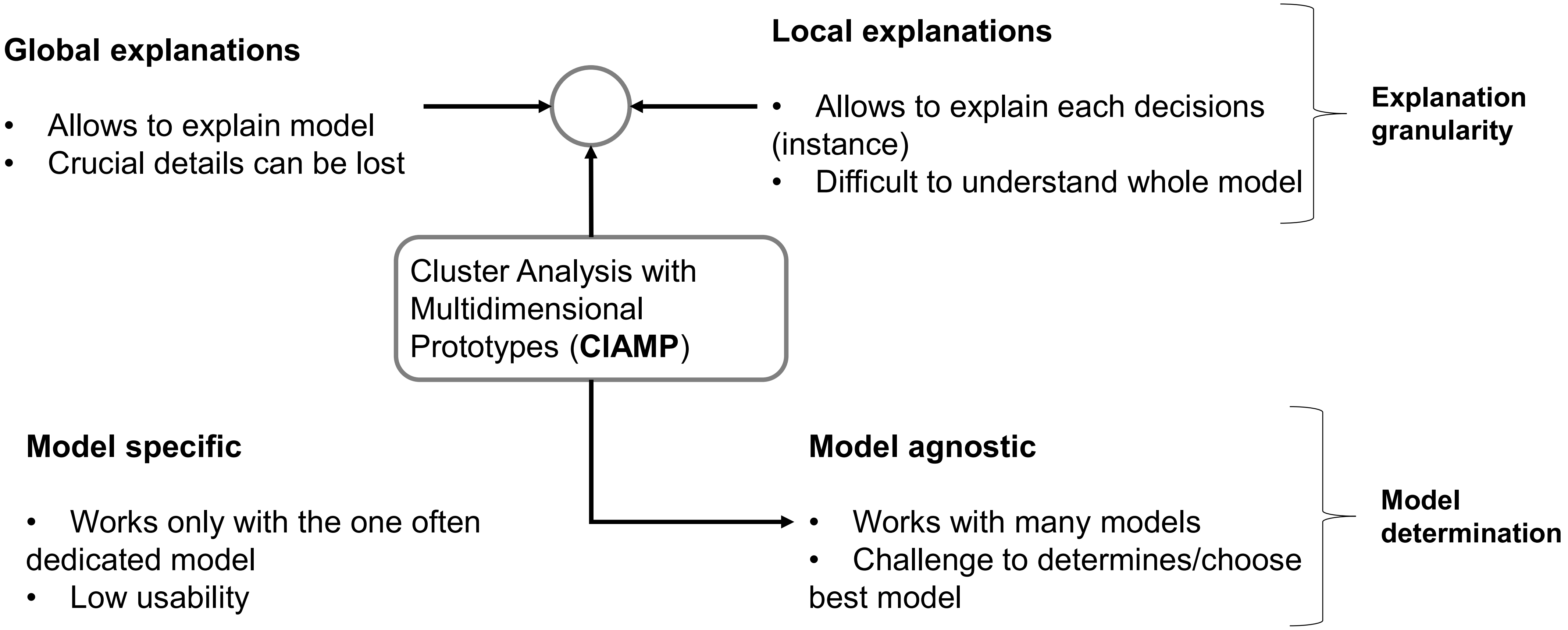
Post-hoc explanation methods:

1. Lore
2. Anchor
3. Lux
4. Shap
5. Lime

Rule form

Importance form

Introduction into explainable artificial intelligence (XAI)



Global explanations

- Allows to explain model
- Crucial details can be lost

Local explanations

- Allows to explain each decisions (instance)
- Difficult to understand whole model

Explanation granularity

Cluster Analysis with Multidimensional Prototypes (**CIAMP**)

Model specific

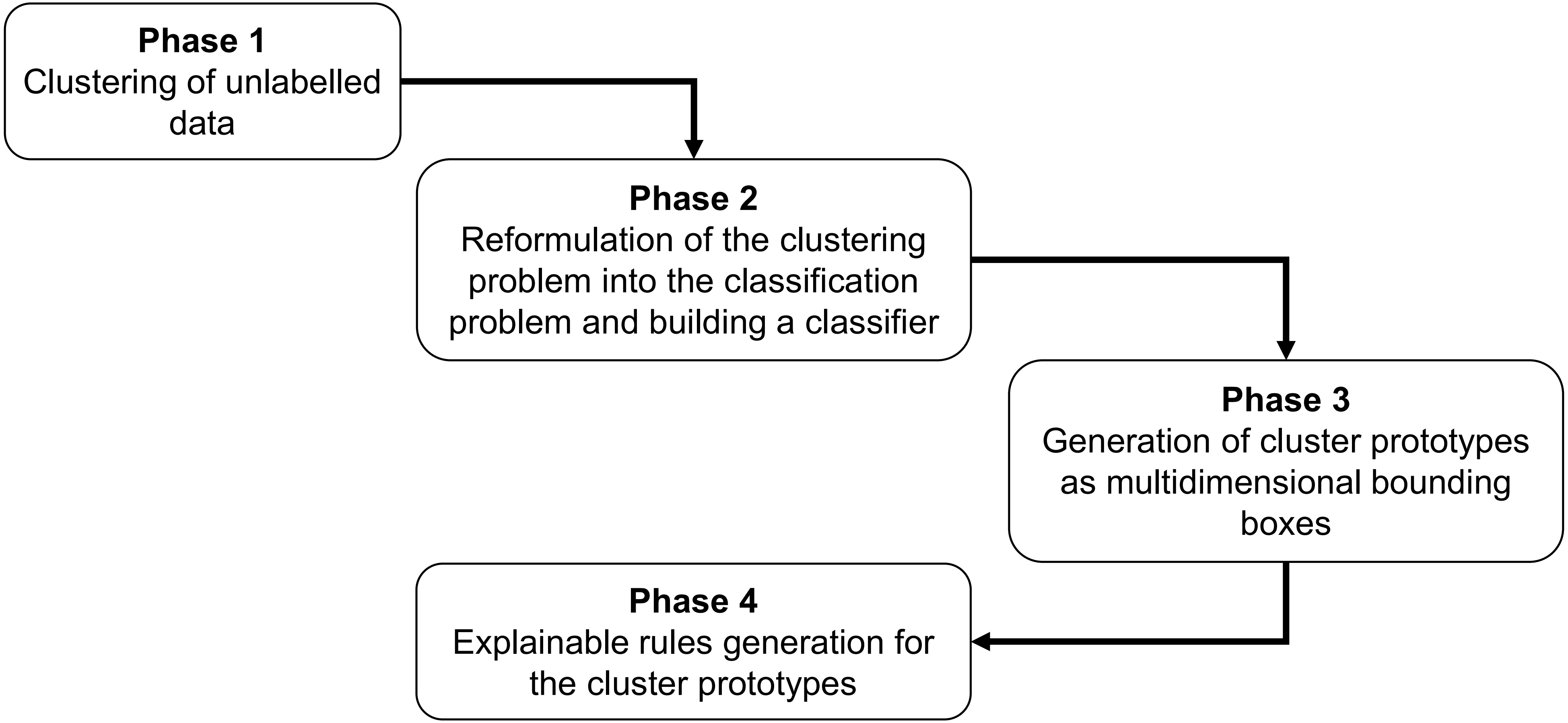
- Works only with the one often dedicated model
- Low usability

Model agnostic

- Works with many models
- Challenge to determines/choose best model

Model determination

Developed methodology – Cluster Analysis with Multidimensional Prototypes (ClAMP)



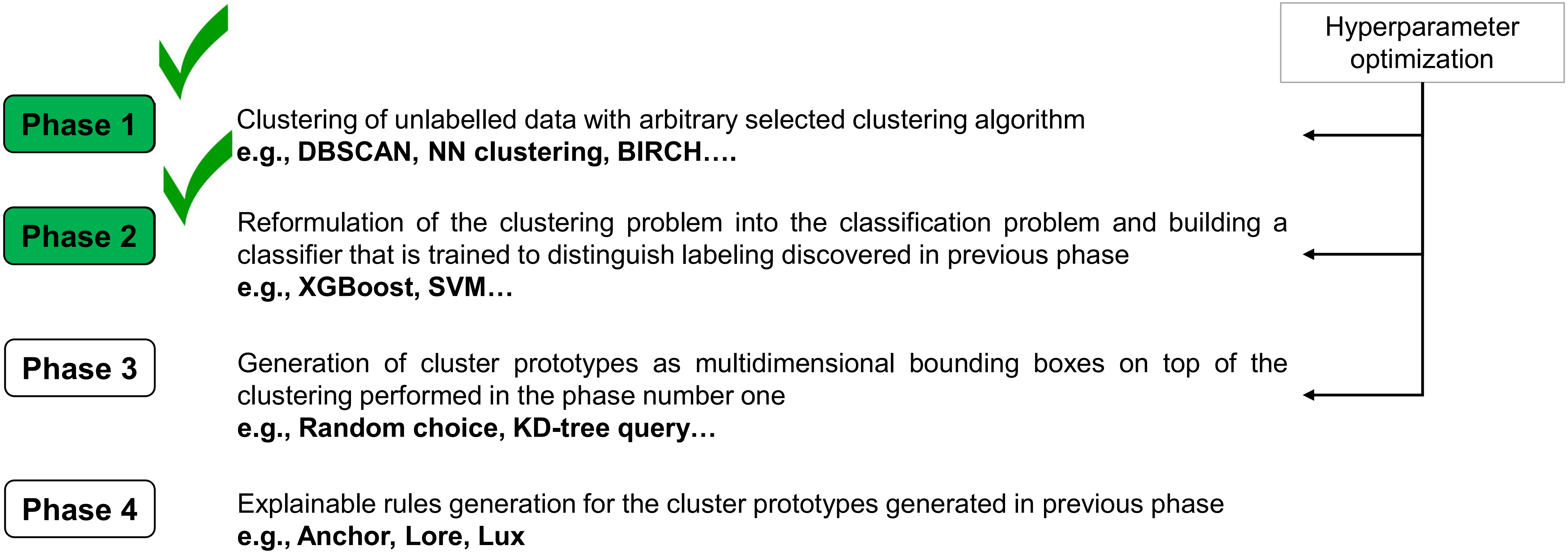
Phase 1
Clustering of unlabelled data

Phase 2
Reformulation of the clustering problem into the classification problem and building a classifier

Phase 3
Generation of cluster prototypes as multidimensional bounding boxes

Phase 4
Explainable rules generation for the cluster prototypes

Developed methodology – Cluster Analysis with Multidimensional Prototypes (ClAMP)



Hyperparameter optimization

Phase 1 ✓

Clustering of unlabelled data with arbitrary selected clustering algorithm
e.g., DBSCAN, NN clustering, BIRCH….

Phase 2 ✓

Reformulation of the clustering problem into the classification problem and building a classifier that is trained to distinguish labeling discovered in previous phase
e.g., XGBoost, SVM…

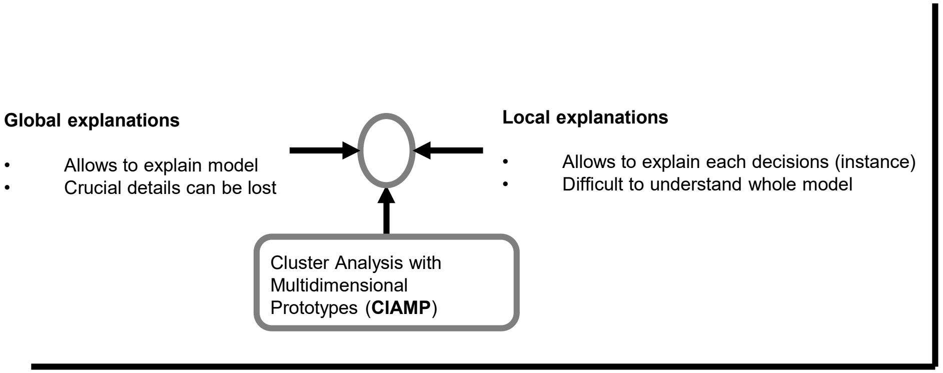
Phase 3

Generation of cluster prototypes as multidimensional bounding boxes on top of the clustering performed in the phase number one
e.g., Random choice, KD-tree query…

Phase 4

Explainable rules generation for the cluster prototypes generated in previous phase
e.g., Anchor, Lore, Lux

Developed methodology – Cluster Analysis with Multidimensional Prototypes (CIAMP)



Global explanations

- Allows to explain model
- Crucial details can be lost

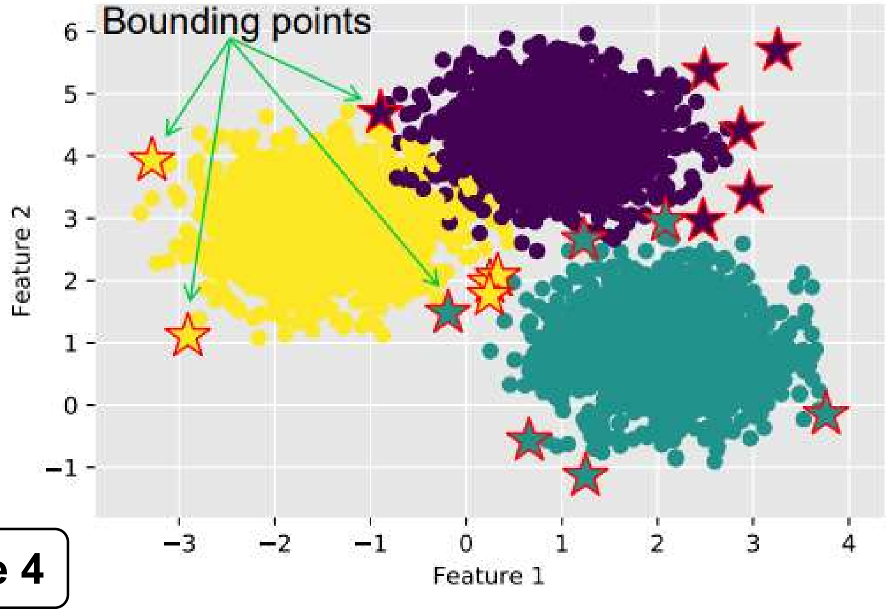
Local explanations

- Allows to explain each decisions (instance)
- Difficult to understand whole model

Cluster Analysis with Multidimensional Prototypes (**CIAMP**)

Main GOAL

Formulating an XAI methodology that allows to balance a trade-off between granularity of global explanations and complexity of instance-based explanation for the cluster of arbitrary shape and dimensionality.



Phase 1 **Phase 2** **Phase 3** **Phase 4**

Developed methodology – Cluster Analysis with Multidimensional Prototypes (ClAMP)

Multidimensional Prototypes

1. Random selection

- generates a randomly selected set of points belonging to each cluster
- the number of points to be selected from each cluster is treated as a hyperparameter which should be optimized

2. K-D tree

- generates the most outer points – boundaries of each cluster
- method’s parameters are treated as a hyperparameters e.g., metric

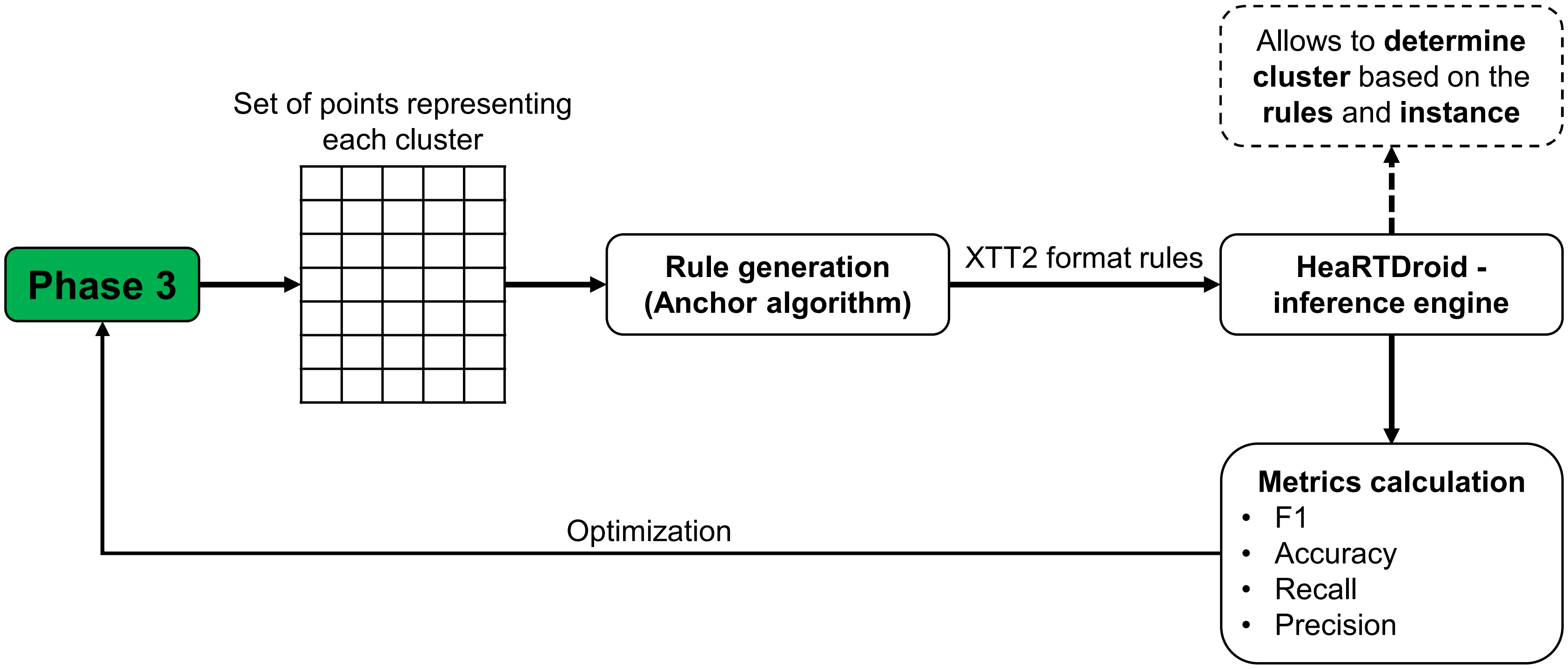
3. Isolation forest

- of the ways to execute outlier detection in high-dimensional datasets
- method’s parameters are treated as a hyperparameters e.g., contamination

Phase 1 Phase 2 Phase 3 Phase 4

Developed methodology – Cluster Analysis with Multidimensional Prototypes (ClAMP)

Anchor Model-Agnostic explainer



Set of points representing each cluster

Phase 3

Rule generation (Anchor algorithm)

XTT2 format rules

HeaRTDroid - inference engine

Allows to **determine cluster** based on the **rules** and **instance**

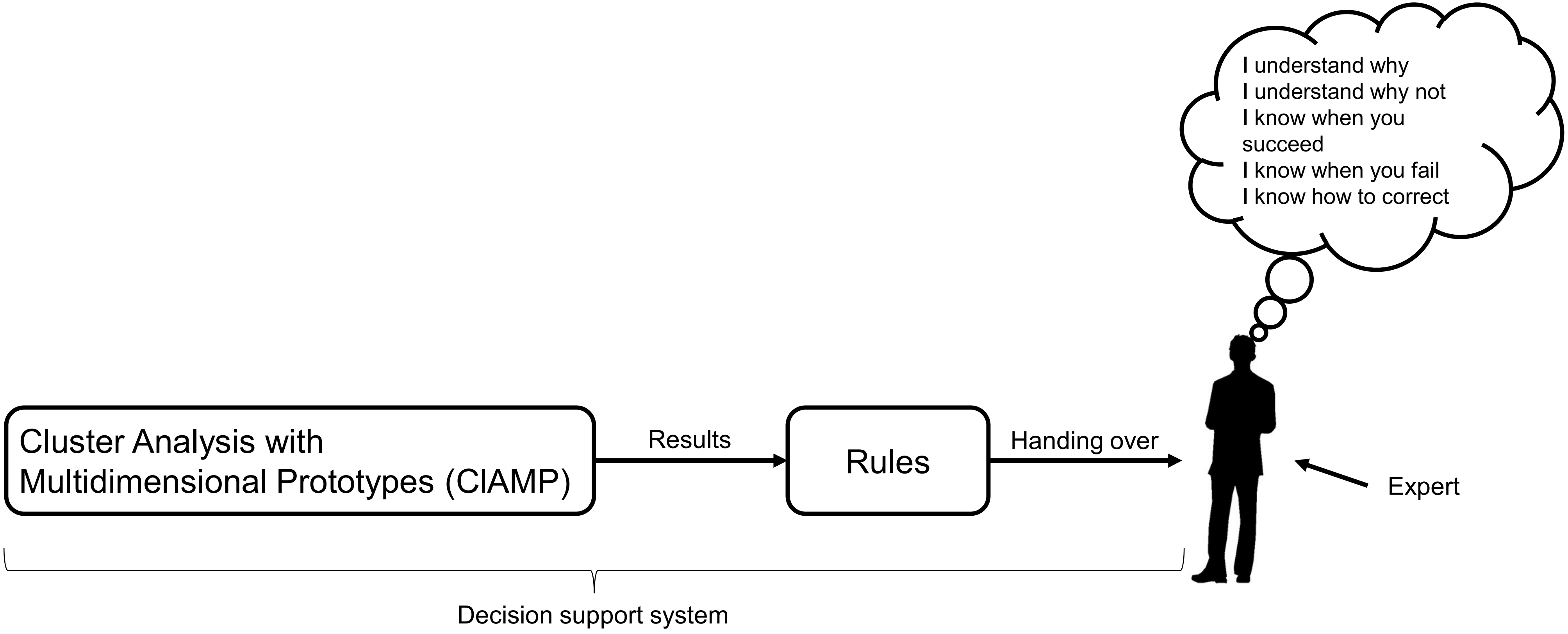
Metrics calculation

- F1
- Accuracy
- Recall
- Precision

Optimization

Phase 1 | Phase 2 | Phase 3 | Phase 4

Developed methodology – Cluster Analysis with Multidimensional Prototypes (ClAMP)

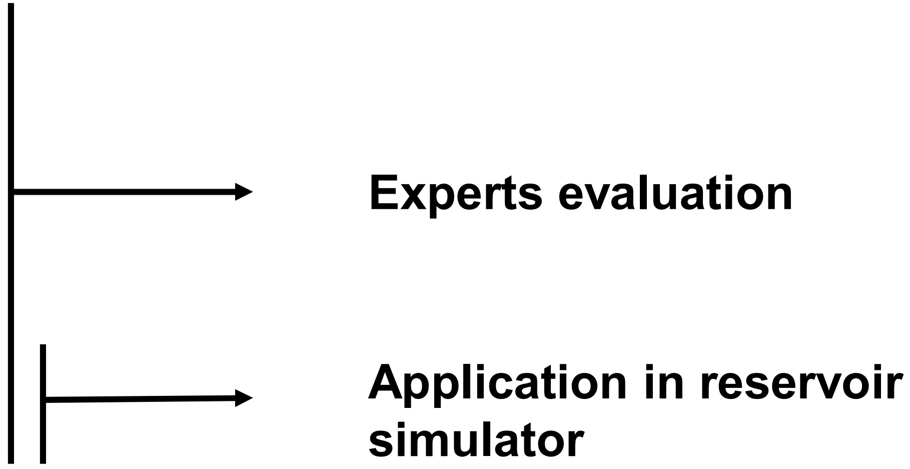


Phase 1 Phase 2 Phase 3 Phase 4

CIAMP applications

Considered applications:

- Artificial datasets
  - ✓ Make blobs in 2D
  - ✓ Make blobs in 3D
  - ✓ Random values
  - ✓ Iris
- Industrial case – hot rolling process
- Industrial case (preliminary study) – oil&gas well production optimization



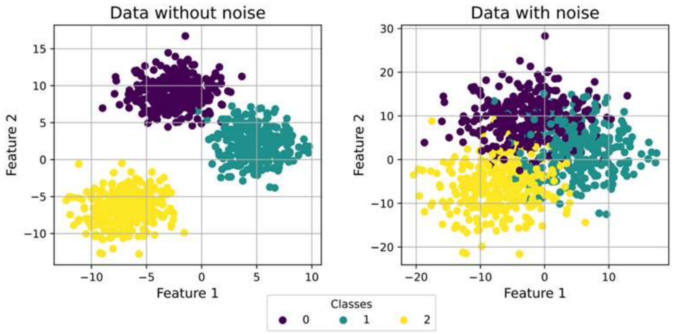
Experts evaluation

Application in reservoir simulator

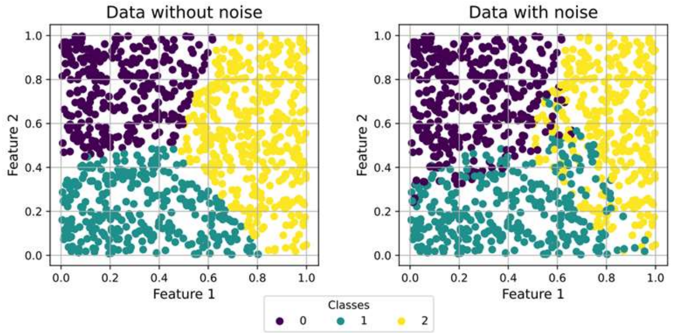
Artificial datasets

Make blobs in 2D, Make blobs in 3D, Random values, Iris

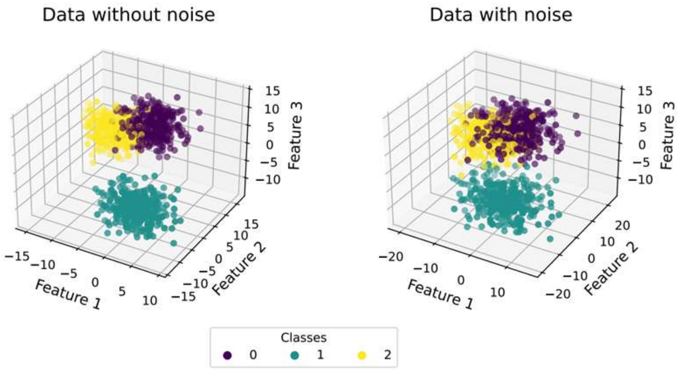
Make blobs 2D



Make blobs 3D



Random values



- DSAA 2021 conference:
  - ✓ **2 describing methods,**
  - ✓ **no hyperparameter optimization,**
  - ✓ **no expert evaluation**

- Current state:
  - ✓ **3 describing methods,**
  - ✓ **hyperparameter optimization,**
  - ✓ **expert evaluation**

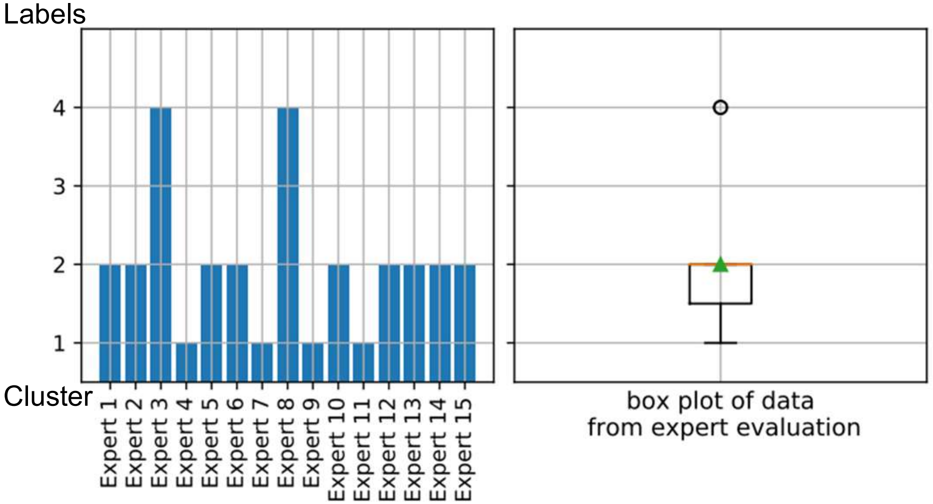
Dataset	**Bounding box method**	
	K-D tree	**Isolation forest**
Make blobs 3D	0.97	0.94
Random values	0.82	0.88

Artificial datasets

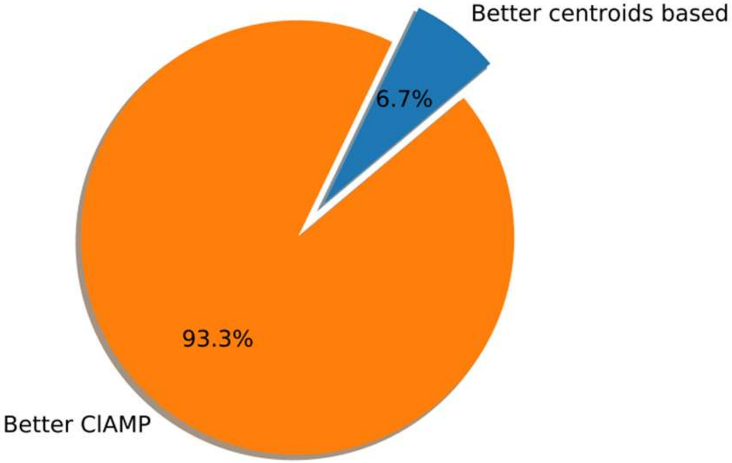
Make blobs in 2D, Make blobs in 3D, Random values, Iris

Artificial datases summary:

Are the rules adequate to explain a given cluster or more individual instances in the cluster?



In comparison to benchmark (centroids based) are CIAMP results better?

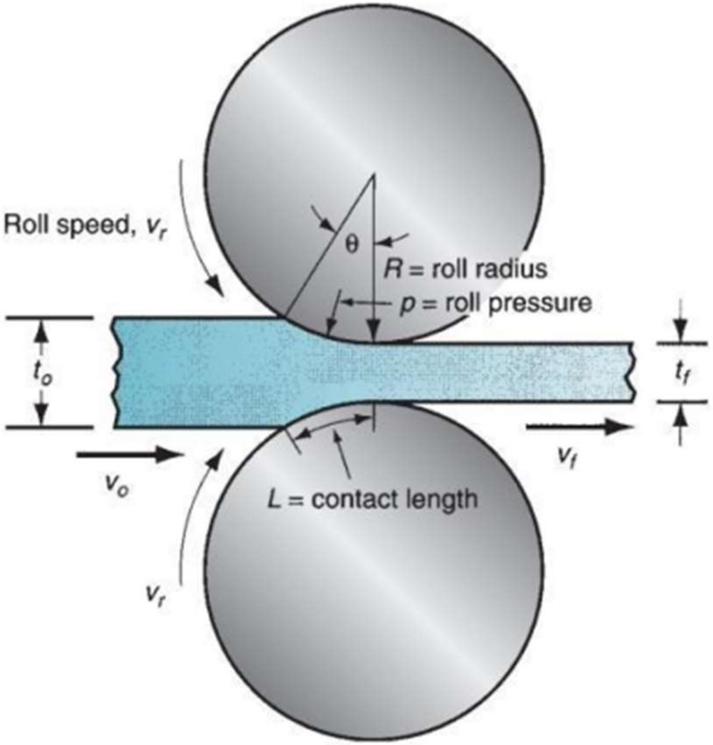


Industrial case – hot rolling process

Hot rolling process

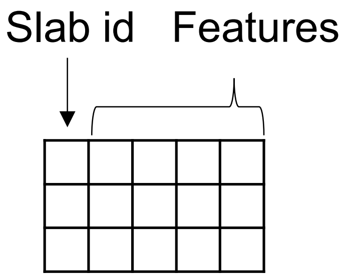
Considered slabs parameter:

- Width
- Profile
- Exit temperature
- Coil temperature



Input to CIAMP

- Standard deviation
- Average

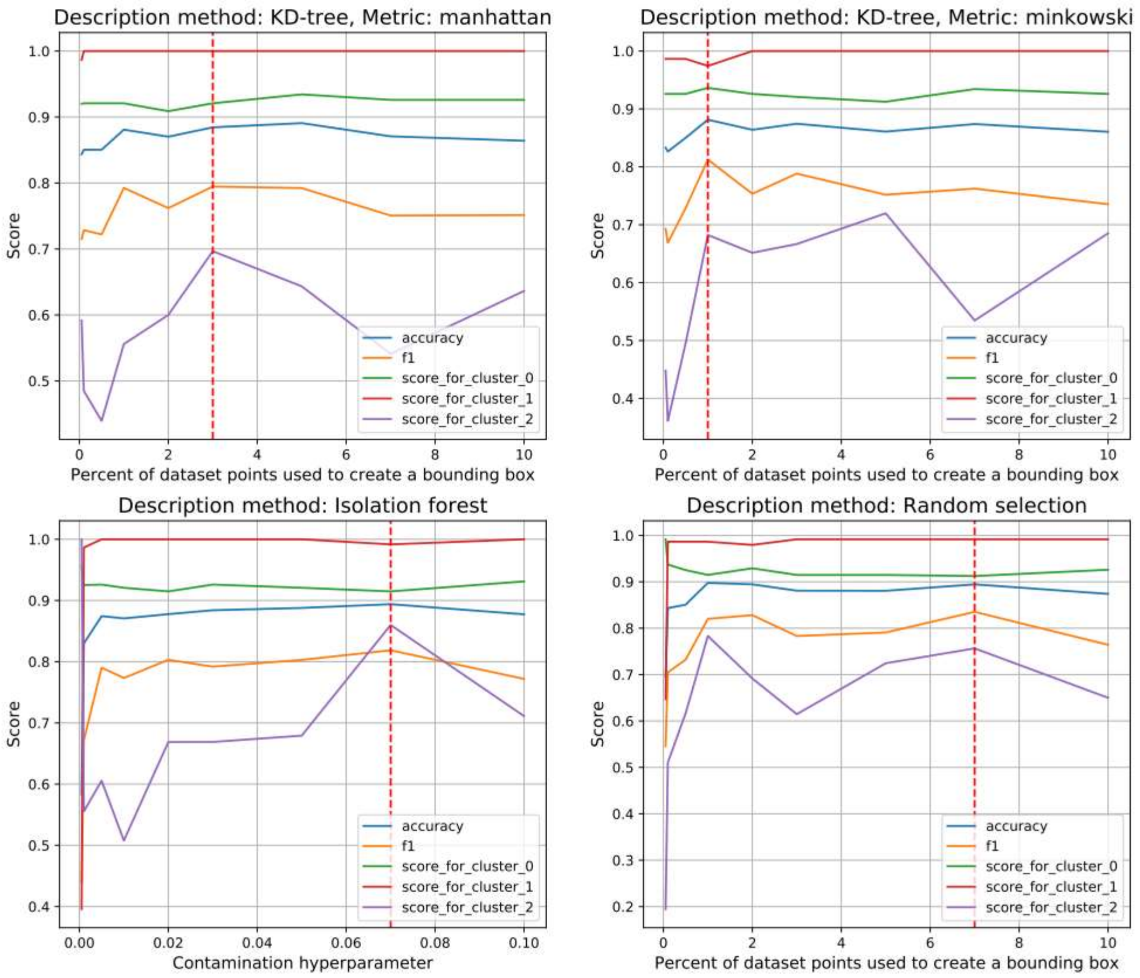


Slab id Features

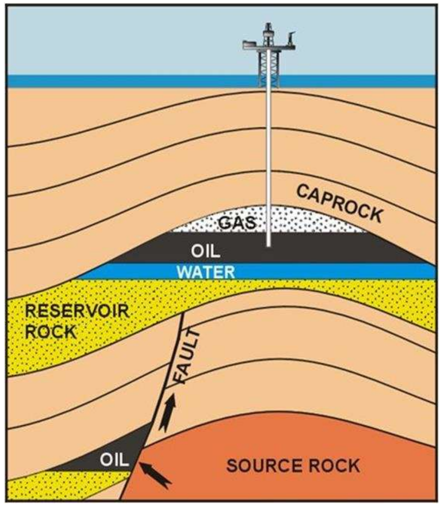
Industrial case – hot rolling process

Hot rolling case summary:

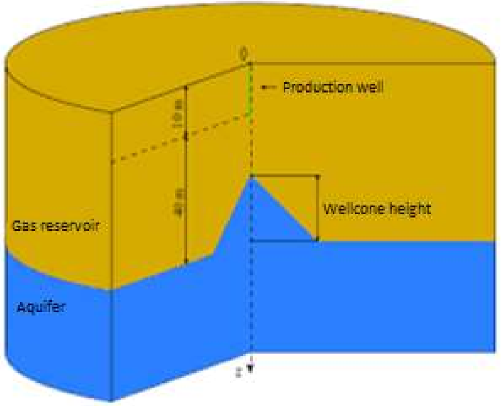
- Dataset with slabs has been divided into 3 **clusters** (groups)
- The best bounding box method was **Random Selection**
- To generate the bounding box, we used **1%** of dataset points in each cluster – which provided the generation of **28 rules**
- We used HeaRTDroid to predict clusters labels based on the generated rules - obtained scores of about **0.8**.



Industrial case (preliminary study) – oil & gas well production management



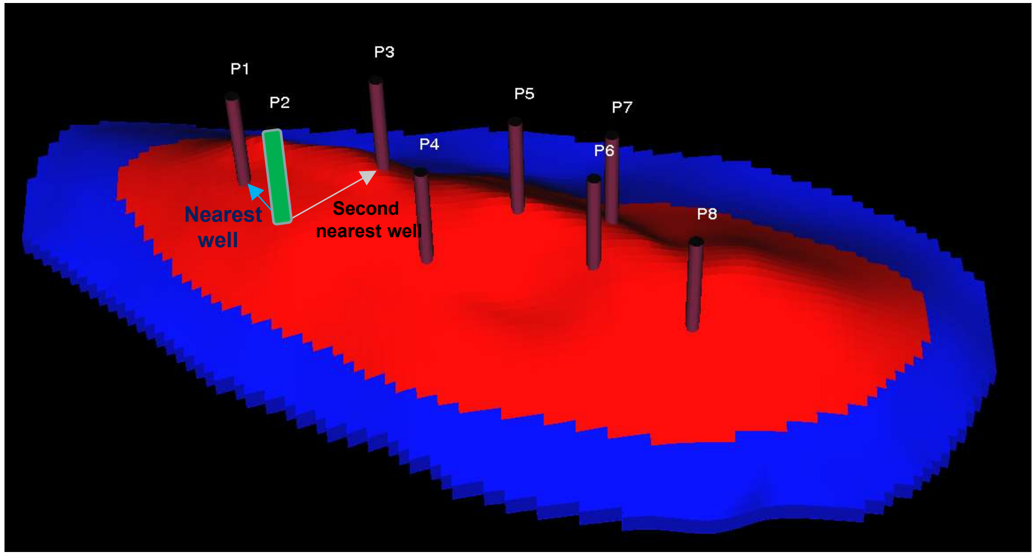
Cost of the one well ~ a few million $



Water conning problem

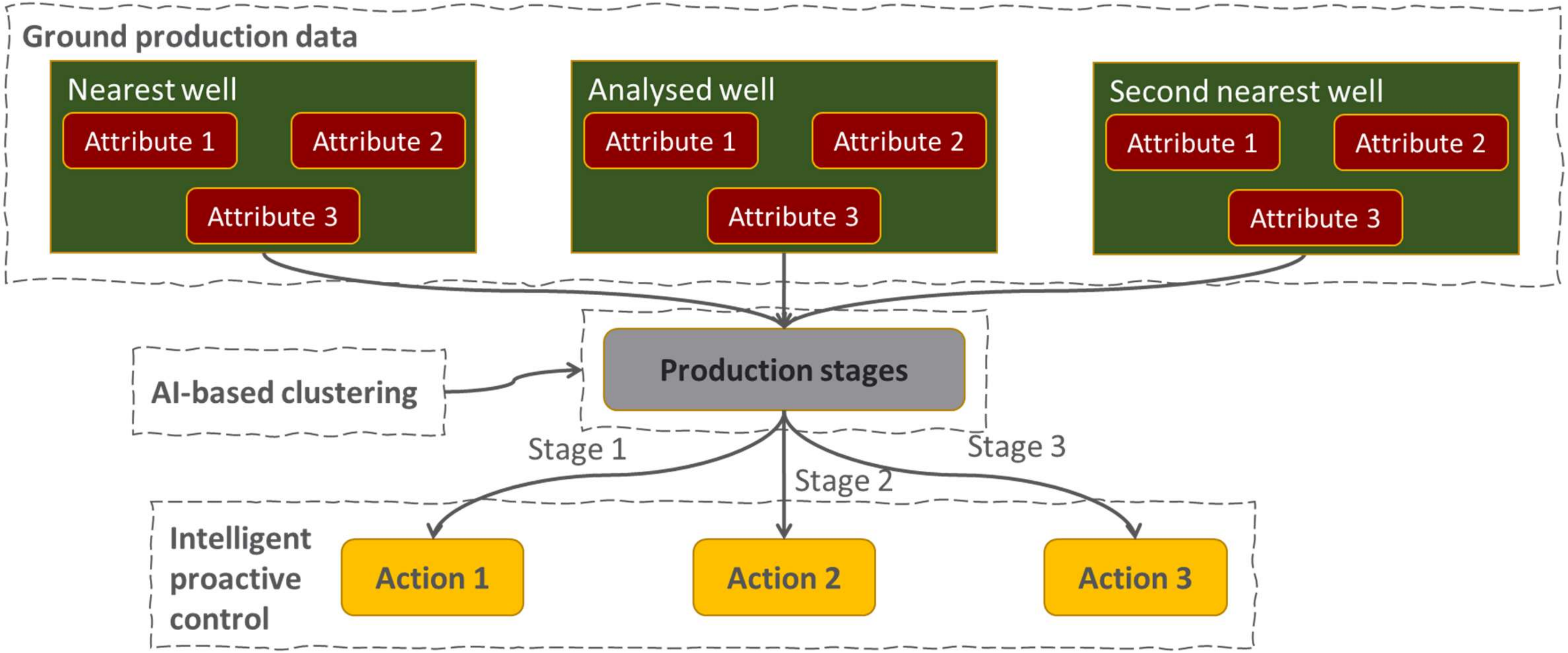
Challenges:

- Maximize oil & gas production from each well
- Minimize operational cost (e.g., production water)

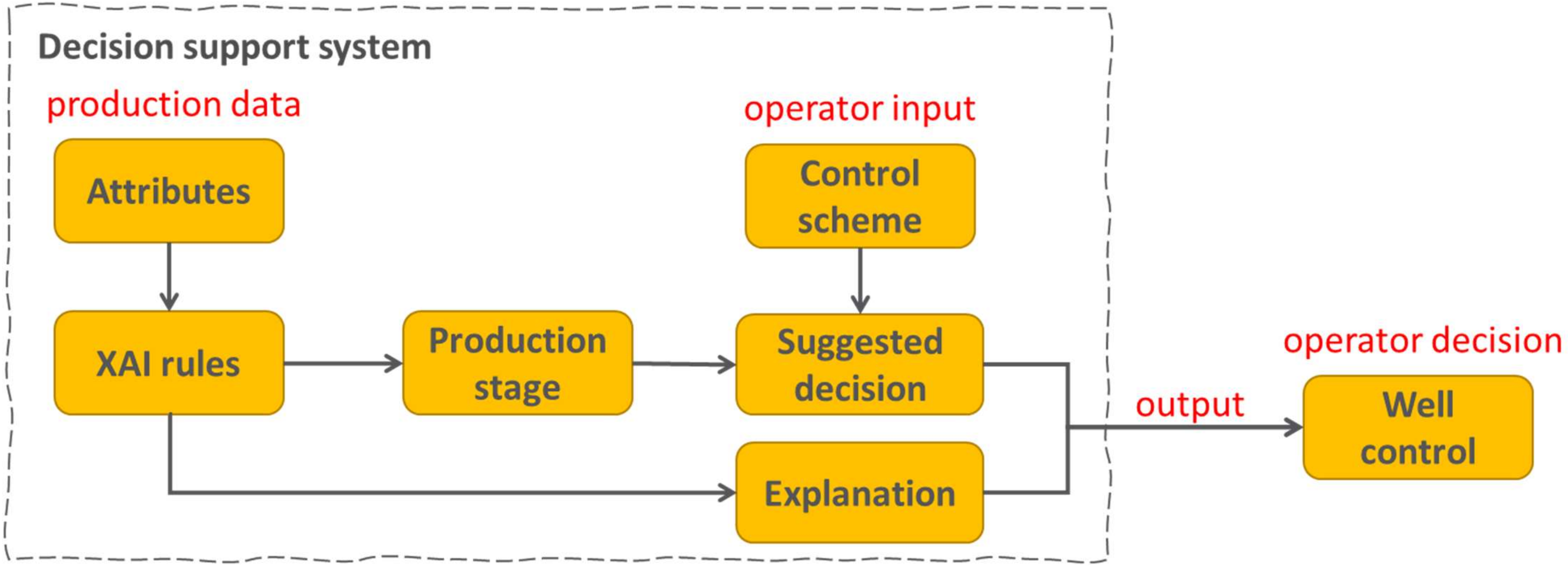


Industrial case (preliminary study) – oil & gas well production management

Well control approach



Explainable intelligent well control algorithm

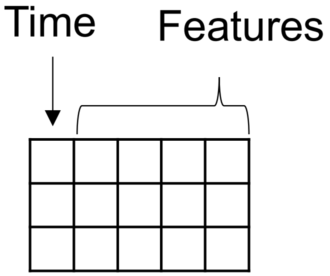


Provided decision support system can be used in **real-time reservoir management**

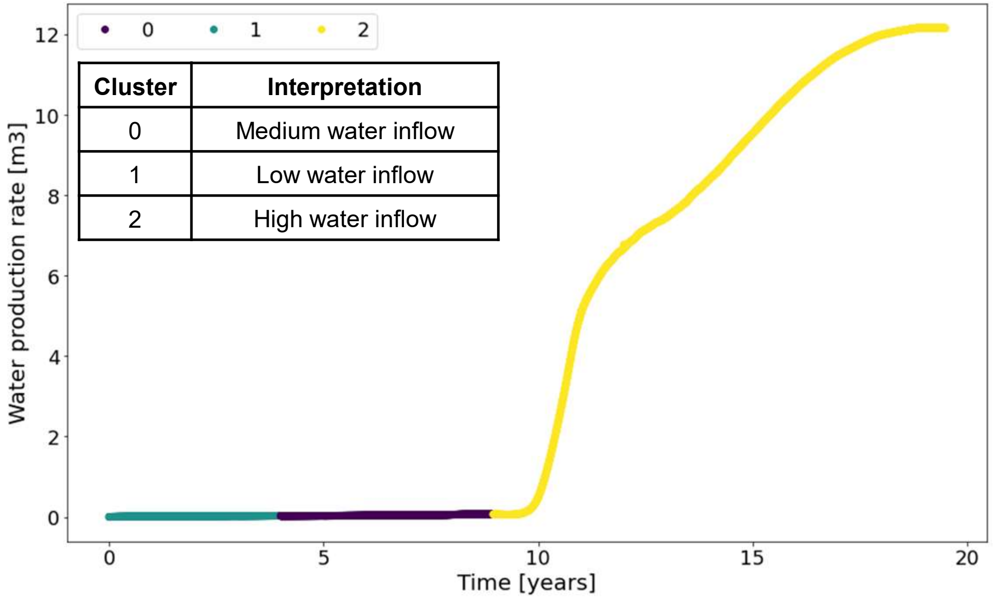
Industrial case (preliminary study) – oil & gas well production management

Input to CIAMP

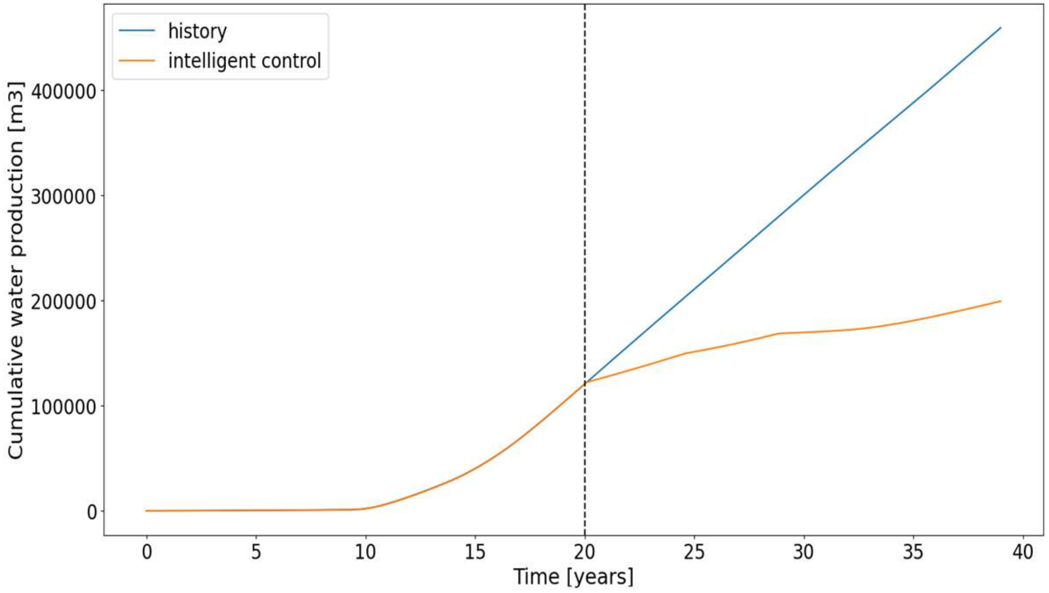
- Real production data
- 9 features, ~25 000 instances



Developed decision support system allowed **the total water production to be reduced by 56%** comparing with historical data.



Cluster	Interpretation
0	Medium water inflow
1	Low water inflow
2	High water inflow



Industrial case (preliminary study) – oil & gas well production management

Reservoir management summary:

- Available historical data was divided into **3 clusters** (groups)
- To generate the bounding box, we used **0.4%** of dataset points in each cluster – which provided the generation of **12 rules**
- Thanks to explainable algorithms We were able to distinguish one cluster which is not obvious
- It helps to better understand fluid behaviors and allows to determine required steps
- Application of generated rules in Eclipse reservoir simulator allows to decrease water production by about 60%.

Summary

- Based on the obtained results, research shows that there is the possibility to apply the ClAMP methodology to the real industrial cases
- The ClAMP allows gaining information about discovered patterns during clustering
- Hyperparameters optimization allows increasing the chance to obtain higher scores and more precise rules
- Considering obtain results and comments from experts it is important to prepare data that could be understandable for the experts

Thank You for Your attention ☺