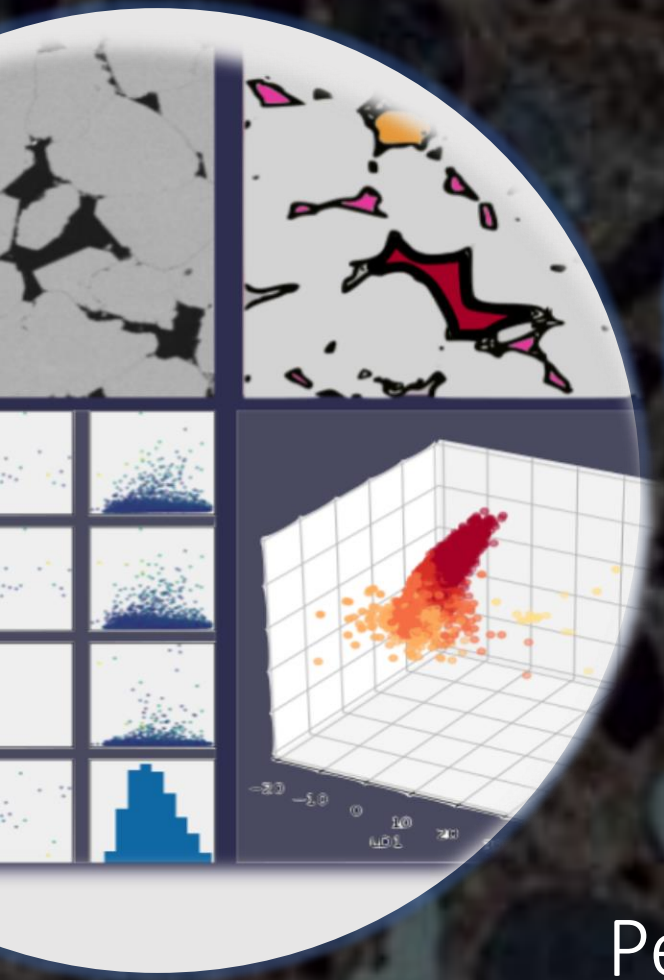
POLITECNICO
DI TORINO

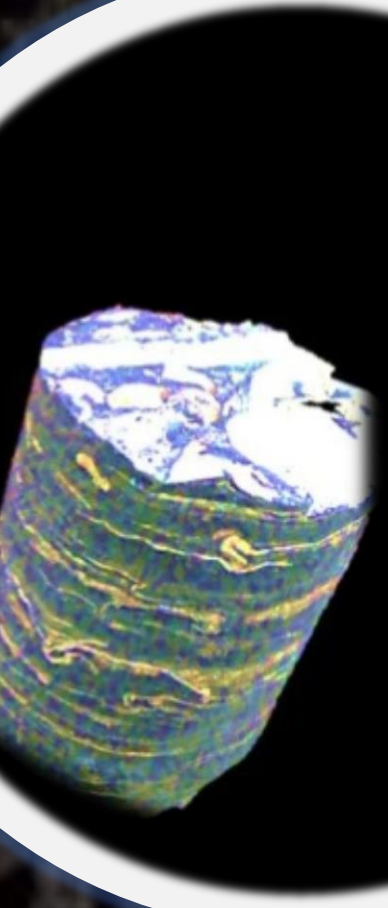
Data Science Lab

Semi-supervised clustering of geological data

DataBase and Data Mining Group

Andrea Pasini, Elena Baralis



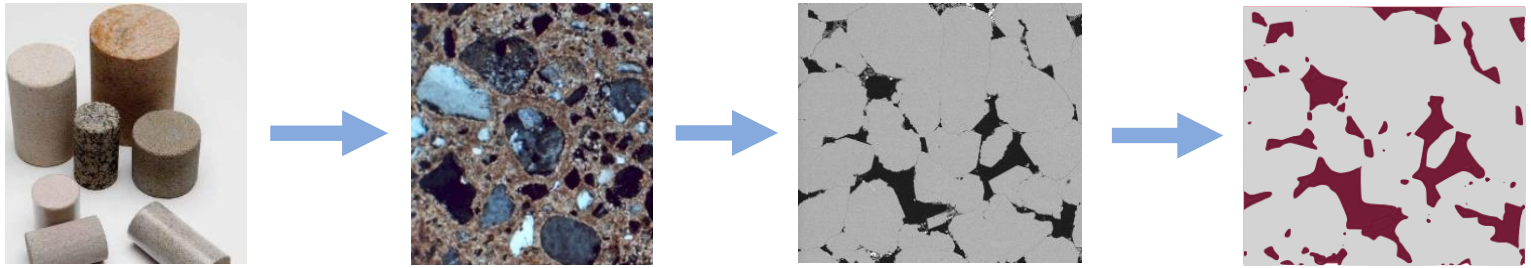


Petrographic Image Analysis

Andrea Pasini, Elena Baralis, Paolo Garza

Petrographic Image Analysis

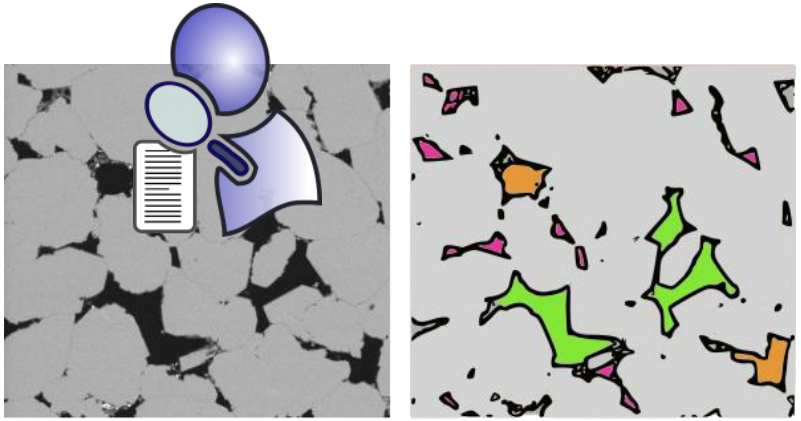
- **Rock core**: small cylinder extracted from a rock sample
- **Thin section**: slice of a core to be analyzed with scanning electronic microscope (SEM)
- **Pore**: cavity among mineral grains of the thin section



Rock cores | Thin sections | Microscope scanning | Extracted pores

Petrographic Image Analysis

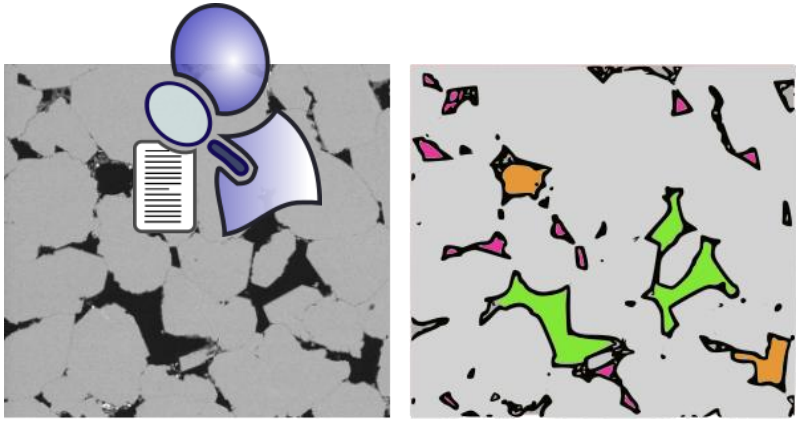
- Pore Typing
  - Each thin section contains **millions of pores**
  - **Experts manually analyze** a small subsample of pores
  - **Grouping these pores** according to their nature allows characterizing the rock sample
    - **structure**
    - **permeability**



a) SEM image b) Pore clustering

Petrographic Image Analysis

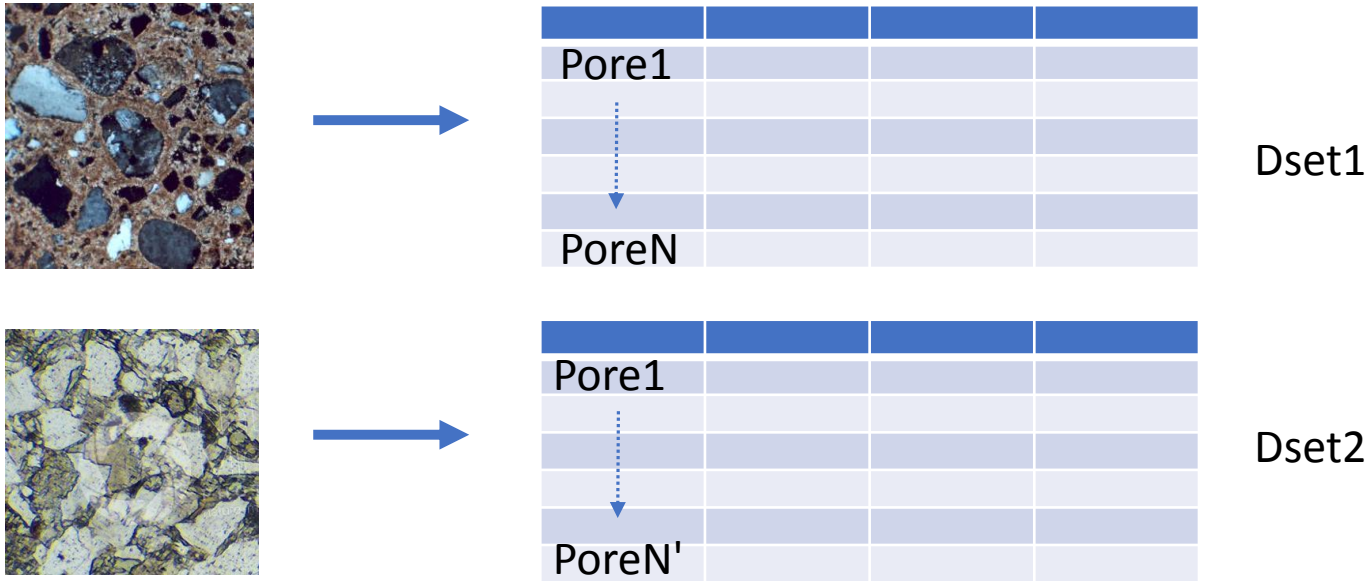
- **Research objectives**:
  - Reduce the **manual effort** for categorizing pores
  - Improve categorization **accuracy** by extending the analysis to the whole amount of pores



a) SEM image b) Pore clustering

Problem setting

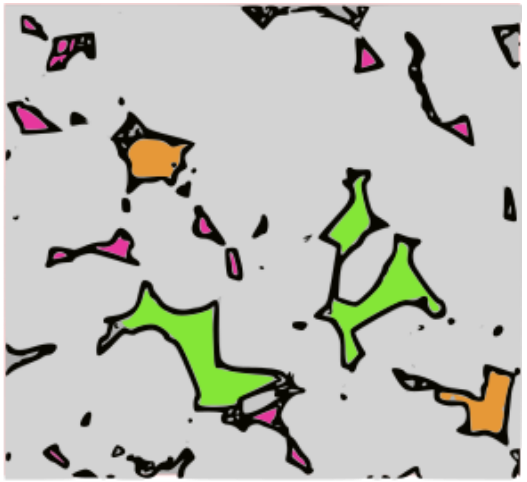
- Input data:
  - About 30 different datasets from different **rock types** (e.g. sandstone, quartz)
  - Each dataset contains pores from a single **thin section**



Problem setting

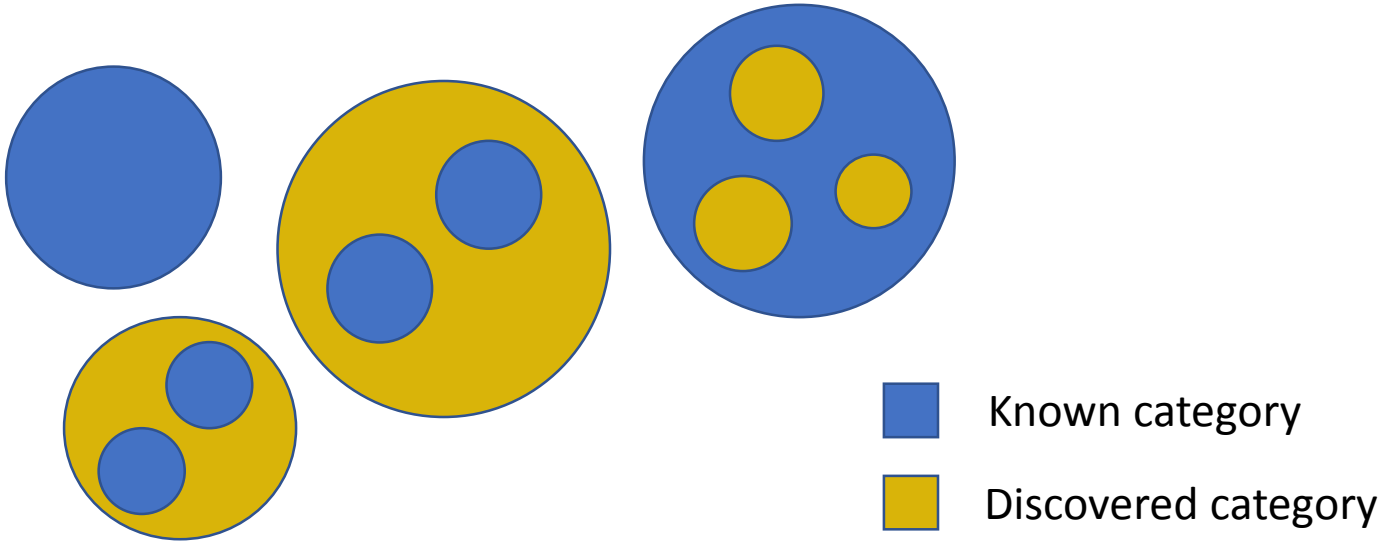
- Analysis output
  - Pore categorization
    - 1 label for each pore in the thin section

predict



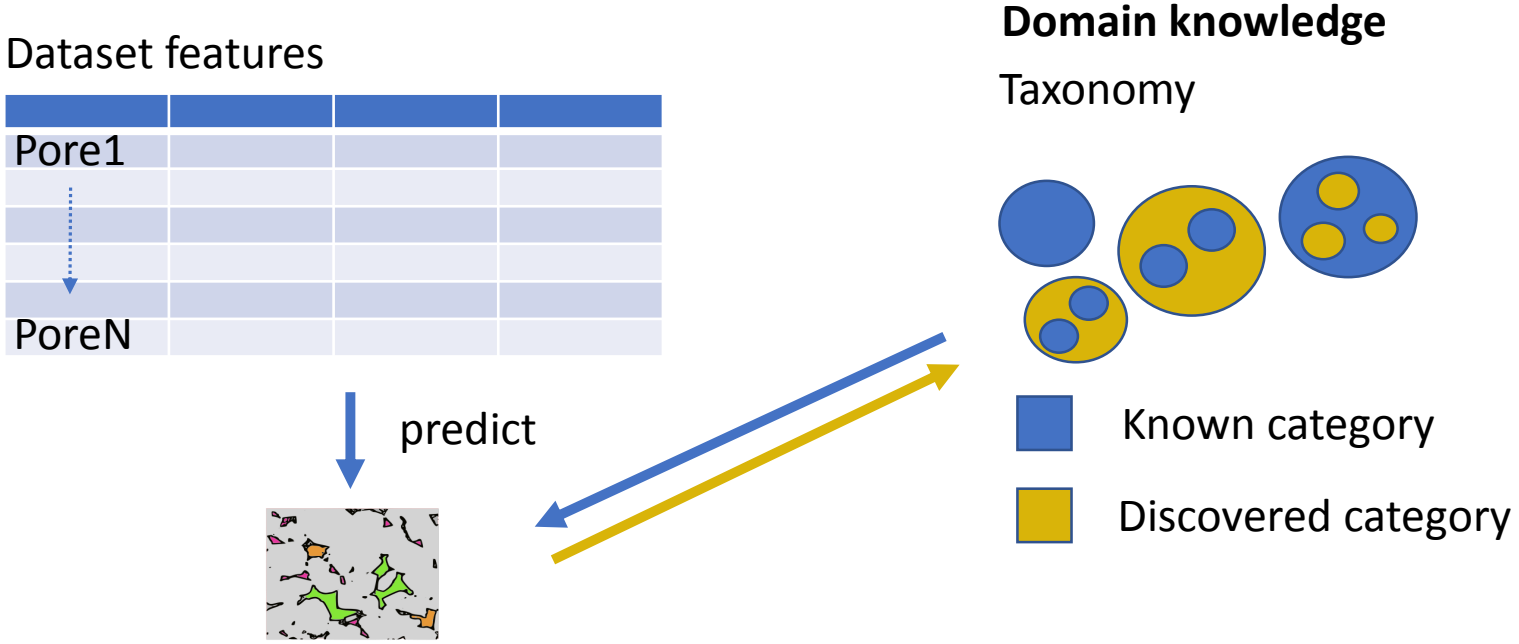
Problem setting

- Pore categories are **only partially known**
  - Geologists defined a **taxonomy** with **known categories**
  - However they want to complete the taxonomy by **discovering** new, unknown, pore categories



Problem setting

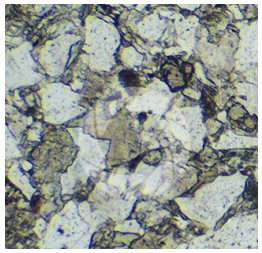
- Provided datasets do not have training labels
  - Labels are missing even for known categories
- **Unsupervised** or **semi-supervised** approaches should be used



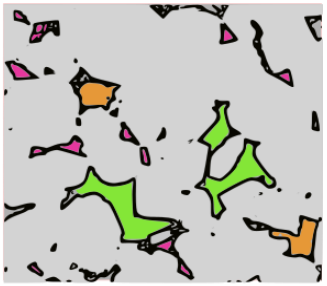
Problem setting

- Different **rock types** (e.g. sandstone, quartz) possibly present different pore types
  - Probably this implies training different models

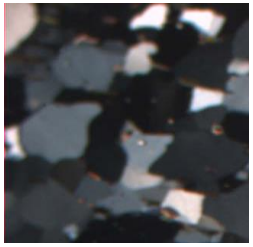
Dset1
Type = Sandstone



Model a



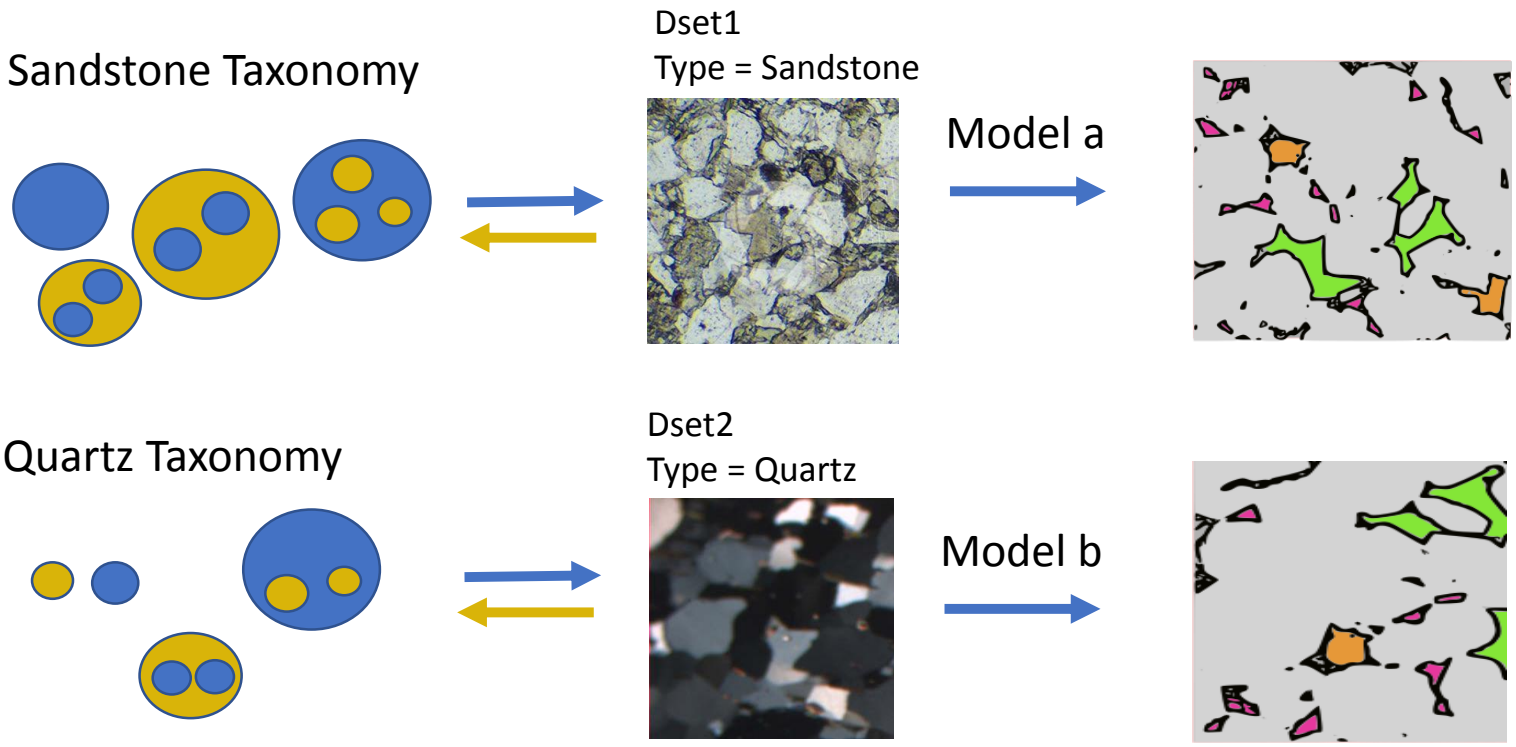
Dset2
Type = Quartz



Model b

Problem setting

- Different **rock types** (e.g. sandstone, quartz) possibly present different pore types
  - Different taxonomies

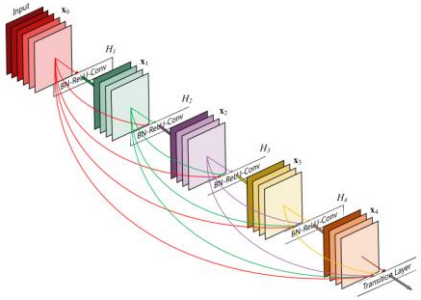


Design of the analysis workflow

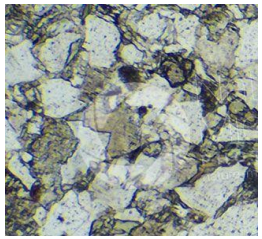
- Pore **description**
  - What are meaningful features to describe data?
- **Algorithm** for predictions
  - Unsupervised vs supervised models
- **Result** description and evaluation
  - We aim at interpretable results
  - How to evaluate the quality of the results?

Design of the analysis workflow

- Pore description
  - **Image** analysis with **deep learning** techniques (e.g. CNNS)
    - However we do not have class labels for training supervised models

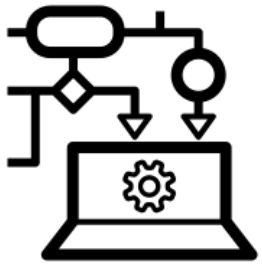


- **Customized** feature extraction techniques
  - Exploit **domain knowledge** to transform pore images to numerical features
  - This is the most natural choice given our problem setting



Design of the analysis workflow

- Algorithm for predictions
  - Unsupervised vs supervised models



- Apply **simpler** techniques first
  - Then inspect more complex techniques to improve quality of the predictions
- **Clustering** techniques can be easily applied to pores
  - **Requirement**: represent pores with numerical features

Design of the analysis workflow

- Result description and evaluation
  - Interpretable results:
    - E.g. cluster description for geologist inspection

  - How to evaluate the quality of the results?
    - **No class labels**:
      - Cannot use external metrics (e.g. ARS)
    - **Geometrical** indexes (e.g. SSE, silhouette)
      - May not be aligned with the **semantic taxonomy**
    - Domain expert **inspection** by means of cluster descriptions

Design of the analysis workflow

- The proposed workflow evolved during the different **interactions** with geologists

Data science experts

interactions

Domain experts

time

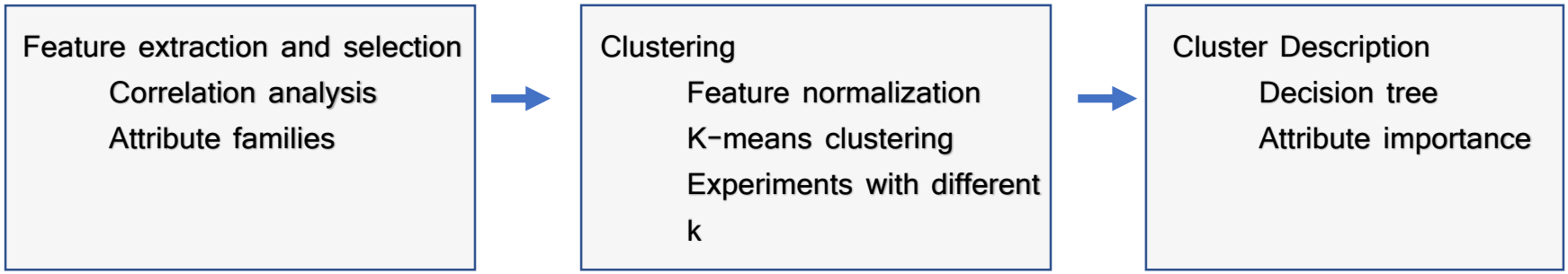
Baseline (standard clustering techniques)

Adaptive Hierarchical Clustering (AHC)

AdaPro (semi-supervised)

Baseline workflow

- Pore categorization by means of standard **clustering** techniques



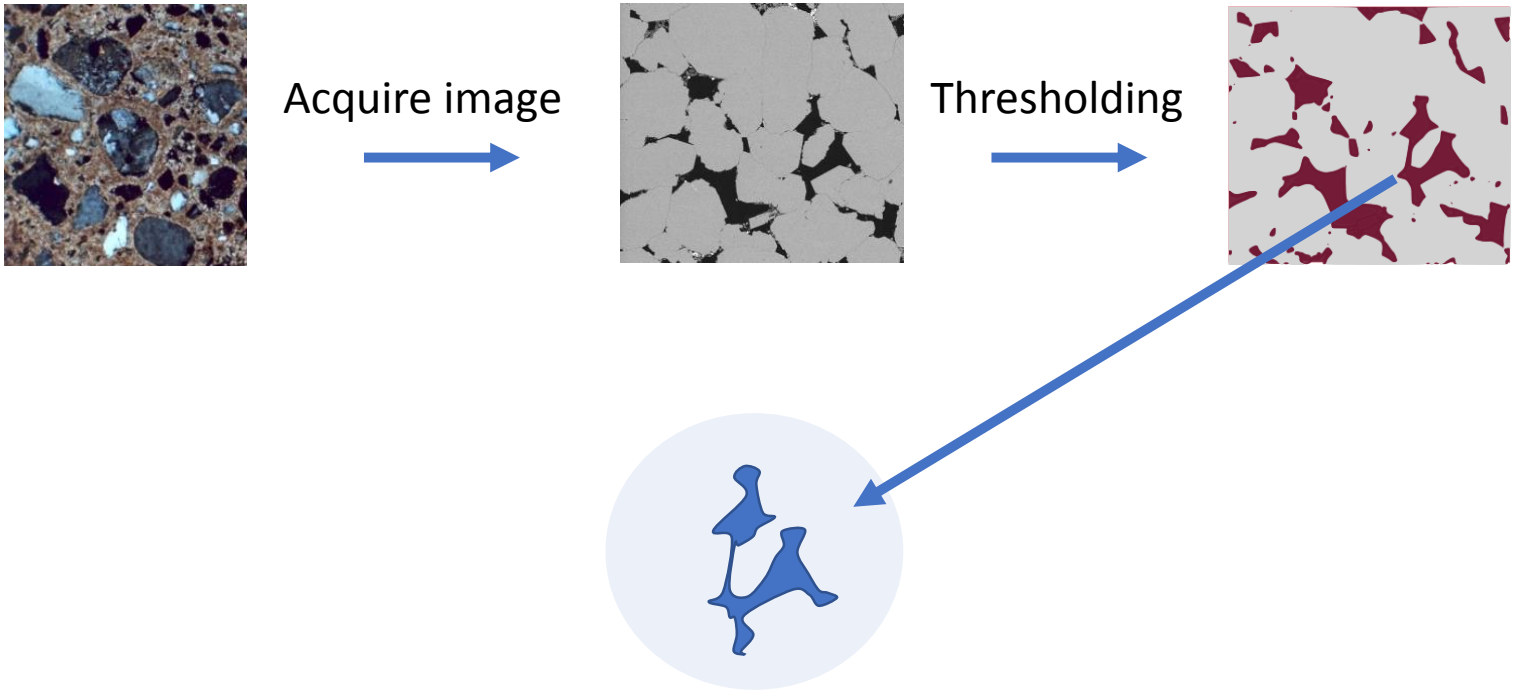
Baseline workflow

- Feature extraction
  - Each **pore** of a thin section is considered a **sample**
  - **Samples** should be described with numerical features

- **Automatic tool for**
  - **Acquiring** grayscale SEM image from a thin section
  - **Extracting** pore pixels
  - **Computing** features to describe each pore

Baseline workflow

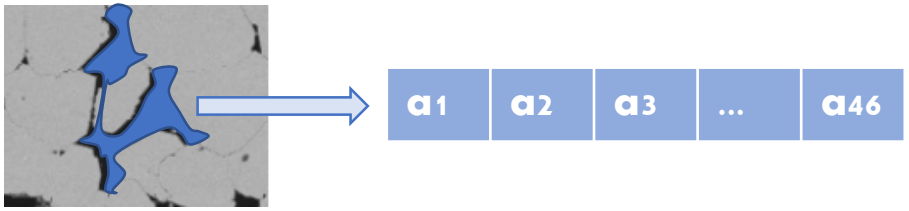
- Pore extraction procedure



Example of extracted pore

Baseline workflow

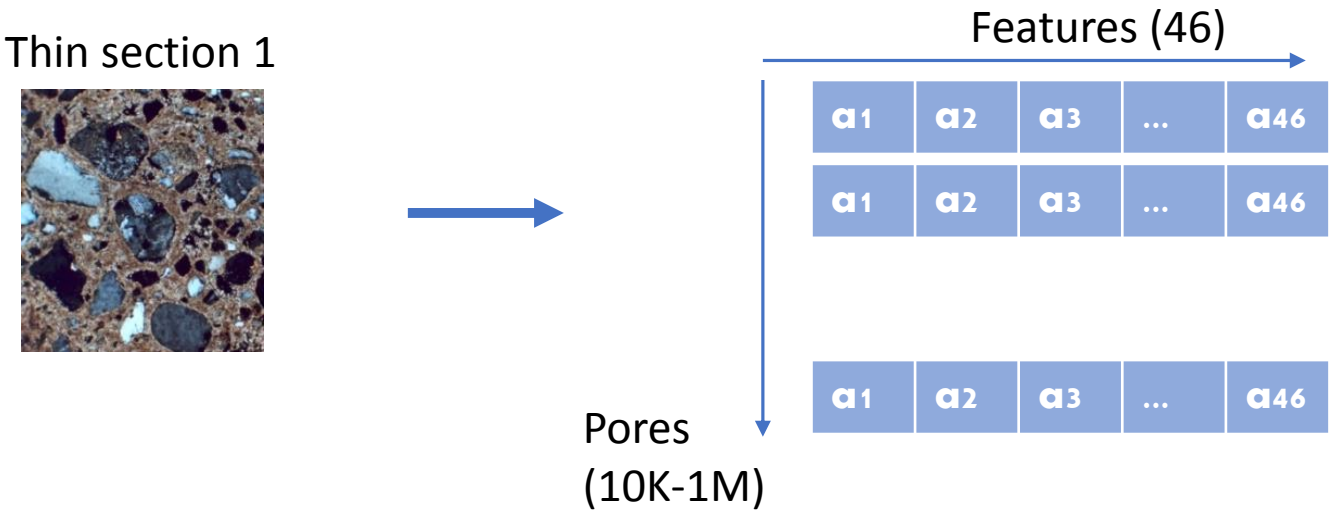
- Pore extraction procedure



  - The acquisition tool used by domain experts extracts **46** numerical features
  - Pores must be characterized according to both **shape** and **size** features
  - Pore shape:
    - e.g. aspect, stretching and irregularity of the pore outline
  - Pore size:
    - e.g. area, diameter

Baseline workflow

- Dataset generation:
  - Each thin section is converted to a tabular dataset by collecting features of all its pores



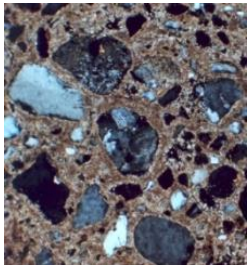
Baseline workflow

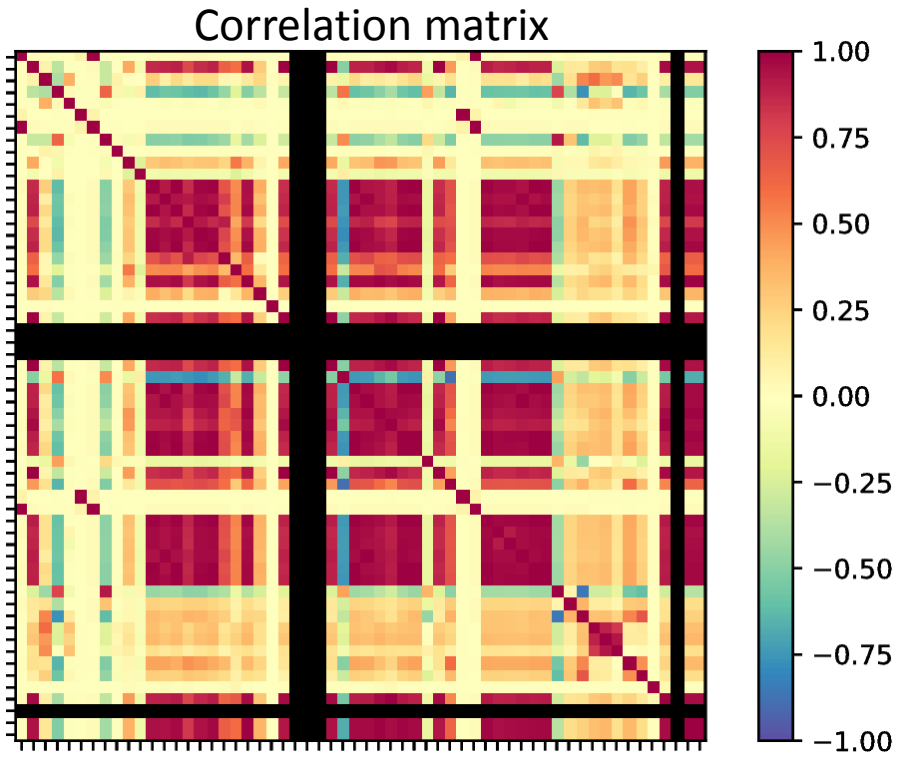
- Feature selection

Baseline workflow

- **Pearson** correlation to inspect relationships
  - Black cells correspond to the attributes that have been removed

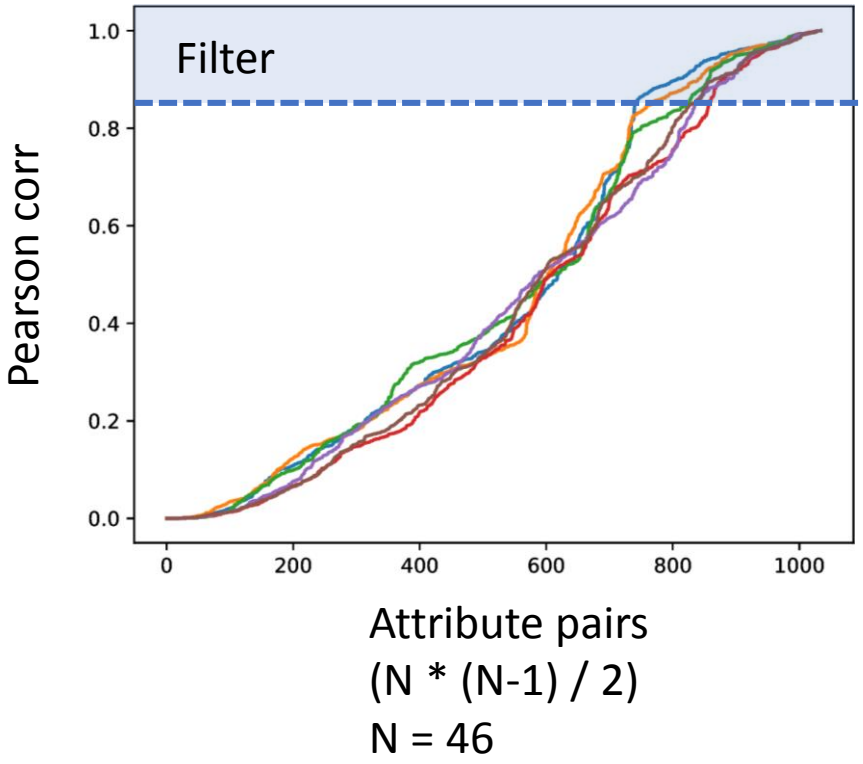
Thin section 1





Baseline workflow

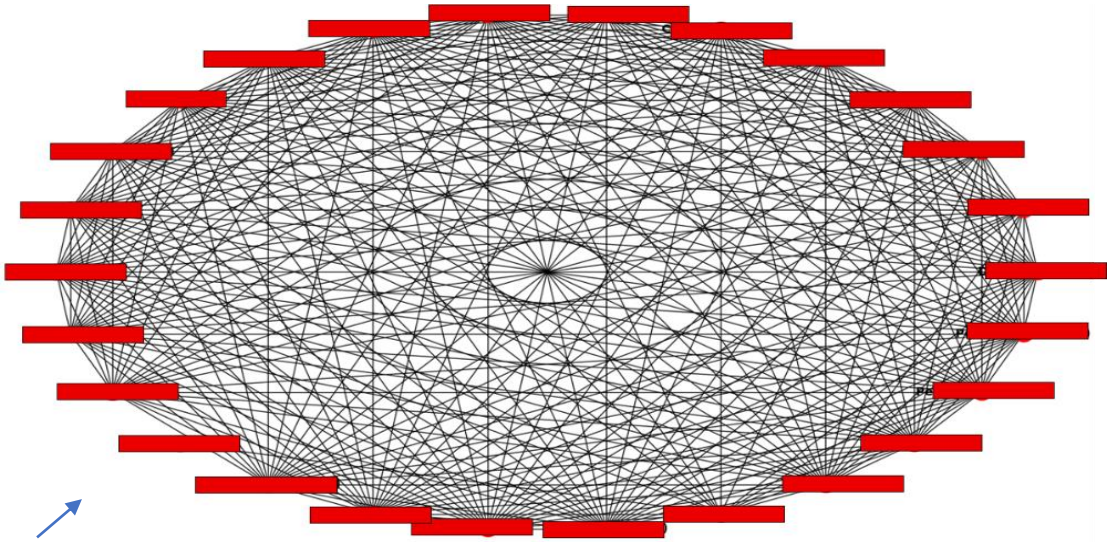
- **Pearson** correlation to inspect relationships



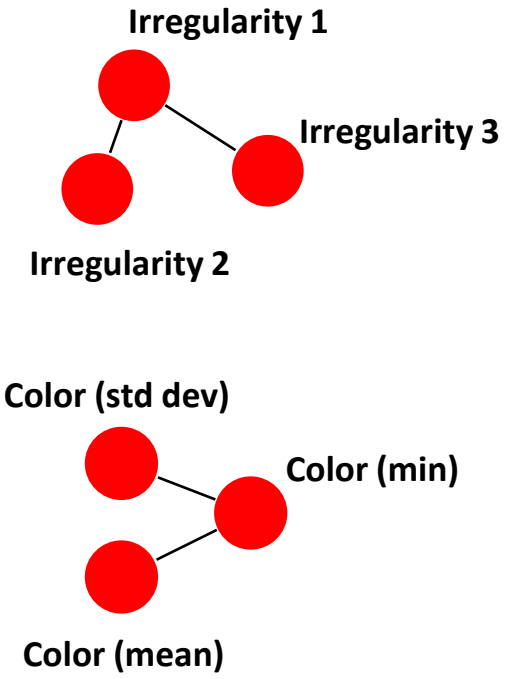
- For each **pair of attributes** plot the **module** of Pearson correlation
- **Sort** attribute pairs by increasing Pearson values
- Different colors correspond to **different thin sections**

Baseline workflow

- Describe correlated attributes with a **graph**
  - Draw an **edge** between attribute pairs with abs(pearson) > 0.85
  - Extract connected components to detect **attribute families**
- Example on dataset Dset1
  - 26 + 3 + 3 = **32 correlated** attributes, **14 uncorrelated** ones (singleton)

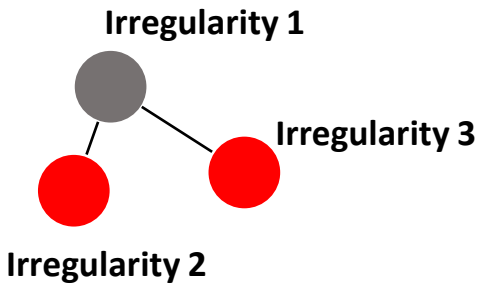


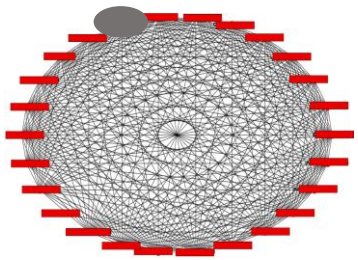
26 attributes (e.g. radius, diameter, ...)

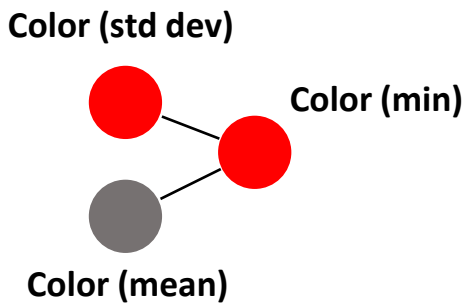


Baseline workflow

- Family leader
  - Select from a family the attribute that is:
    - The most **correlated** with other attributes **within** the family
    - The most **uncorrelated** from attributes **outside** the family

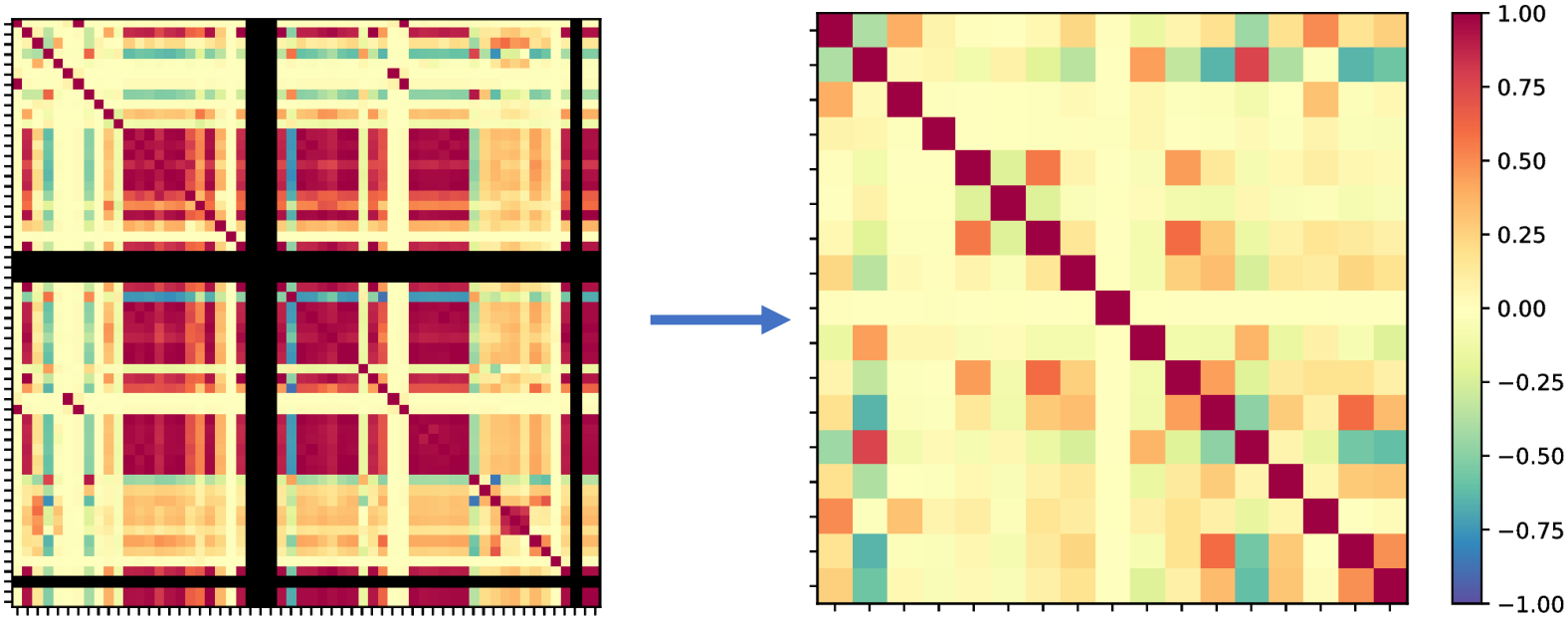






Baseline workflow

- Select 1 **family-leader** from each family of correlated attributes
  - Replace the attributes of the same family with the family leader only

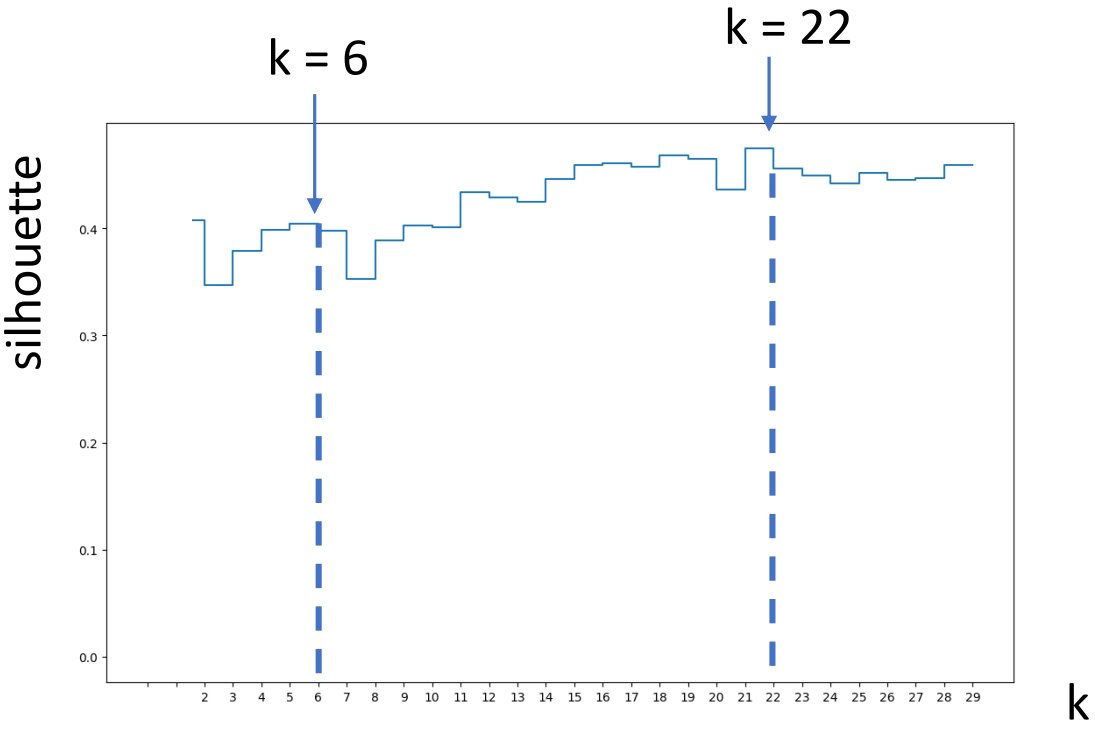


Baseline workflow

- Pore clustering

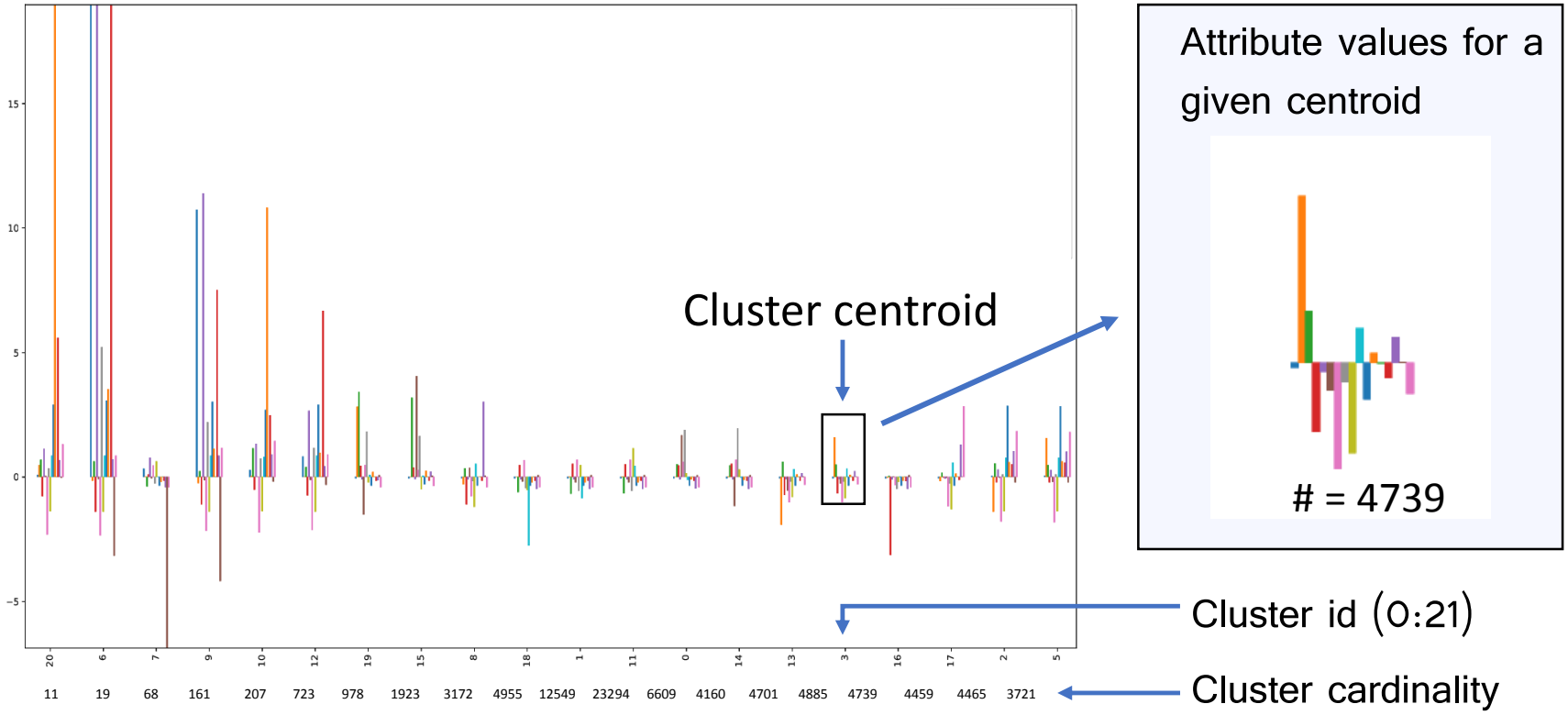
Baseline workflow

- First clustering attempt
  - **KMeans** with different values of k



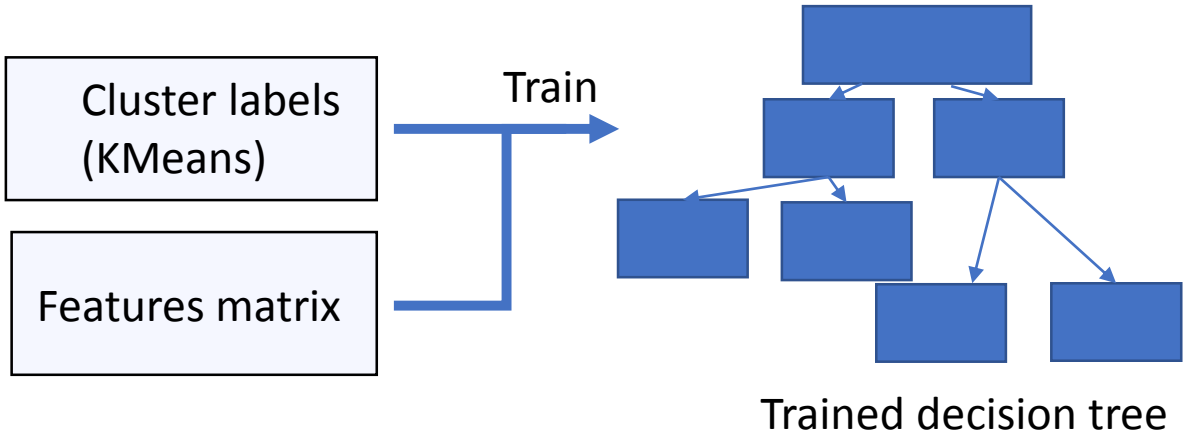
Baseline workflow

- Cluster **description** k = 22
  - Show attribute values for cluster **centroids**



Baseline workflow

- Cluster **description (method 2)**
  - Train a **decision tree classifier** to predict the cluster labels obtained with KMeans
  - Inspecting the generated tree should provide **descriptions** of the cluster shape

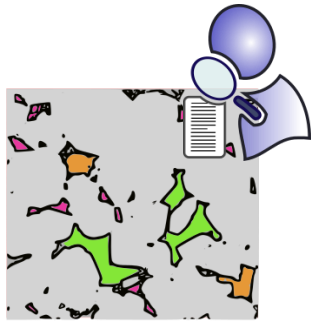


Baseline workflow

- Cluster **description (method 2)**
  - Compute a decision tree for each dataset (thin section)
  - Sort attributes by **feature importance** (decision tree metric)
    - Top-5 important attributes
      - Effectively useful for pore categorization, according to geologists

Baseline workflow

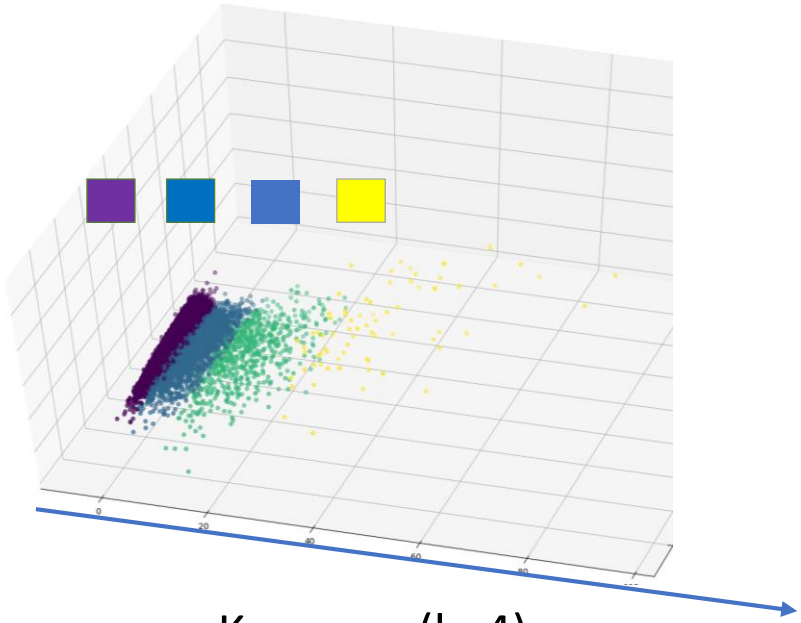
- Discussion of the results
  - Geologist **inspect microscope images** where clustered pores are depicted with different colors



  - KMeans, K=6
    - Categories are mainly divided by **pore size**
    - Complex geological categories are not recognized
  - KMeans, K=22
    - Pores of different geological categories are mostly divided
    - However there is a high **oversplitting** of pore categories

Baseline workflow

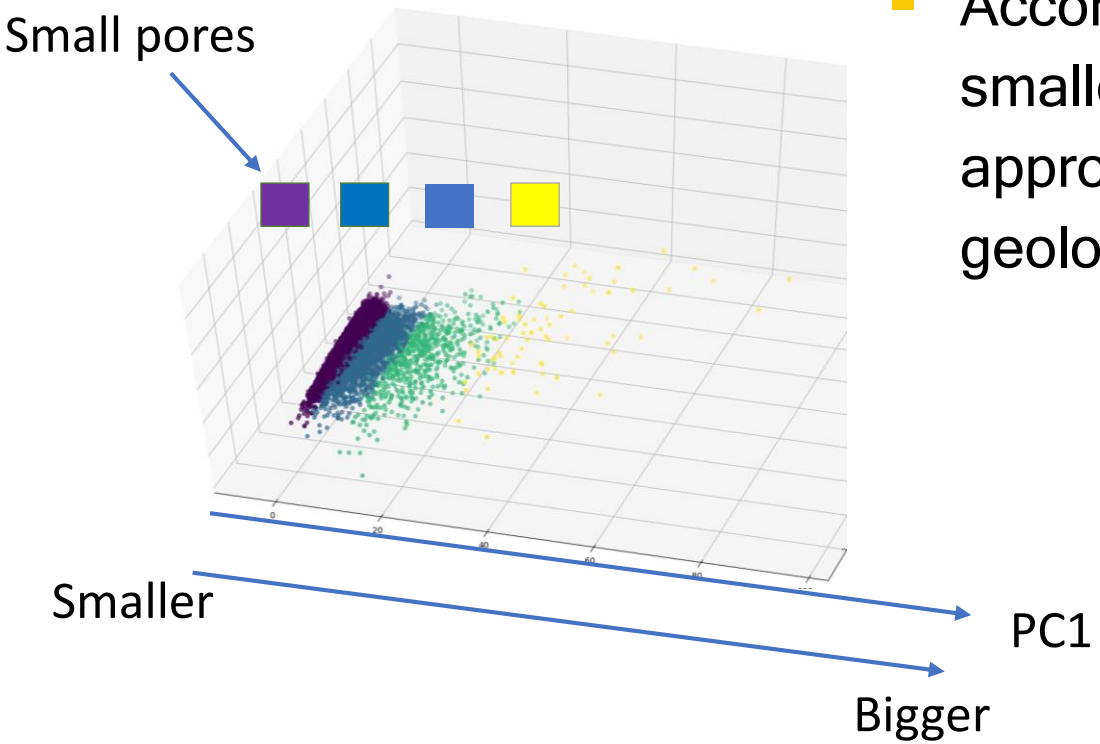
- Verify that **pore size** is mainly driving cluster separation with low values of k
- PCA representation, K = 4



- Principal component PC1 is parallel to many attributes related to pore size (e.g. diameter, area, radius, …)

Baseline workflow

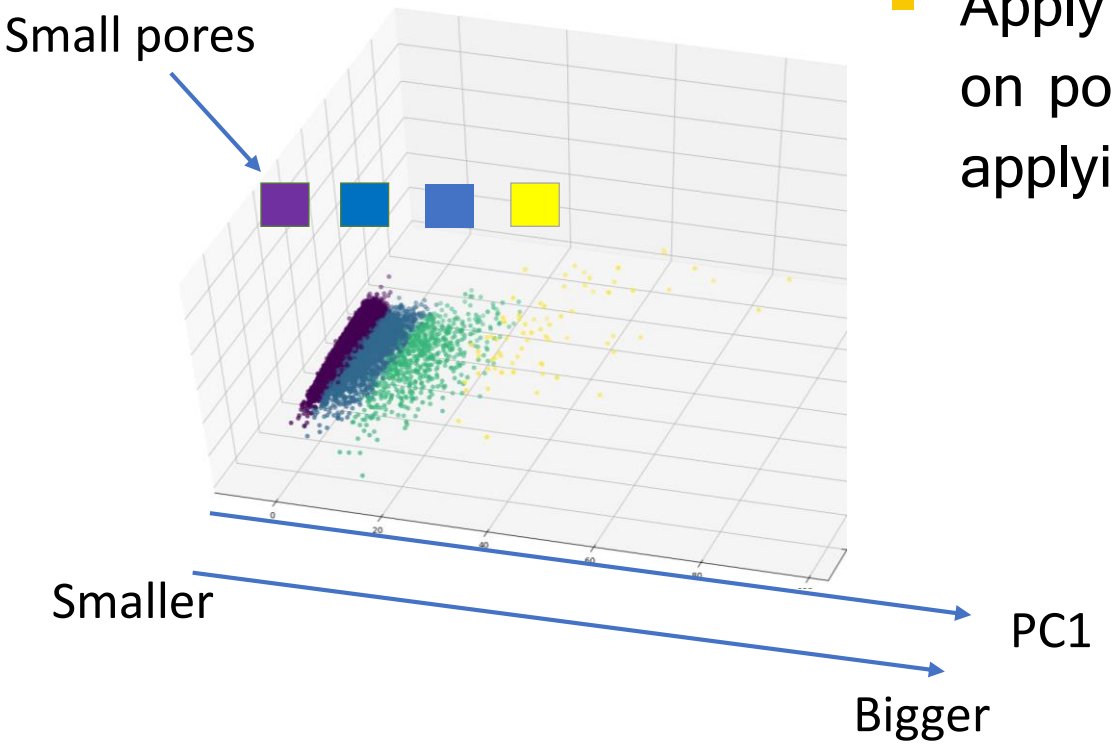
- Small pores are the majority of our dataset
- They form a **dense cluster**



- According to geologists, smaller pores have all approximately the same geological characteristics

Baseline workflow

- The dense cluster of small pores can be **removed** from the analysis to facilitate the recognition of other important pore categories

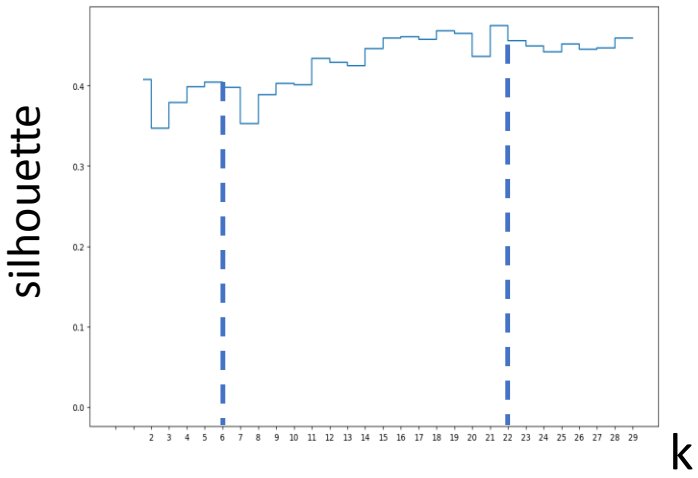


- Apply a **filter threshold** on pore diameter before applying Kmeans

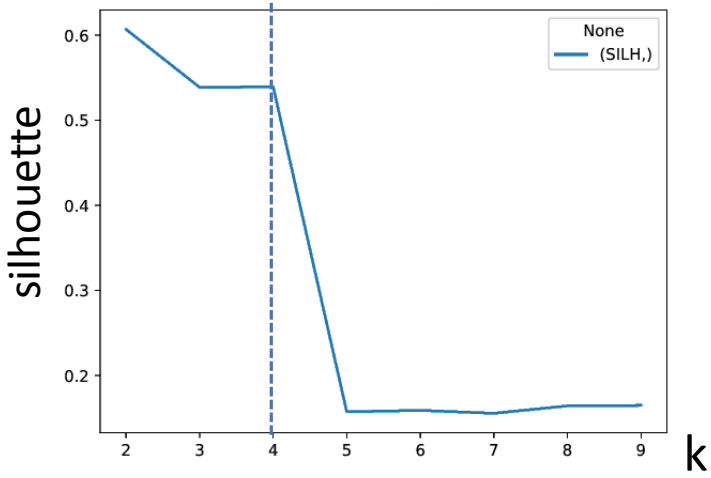
Baseline workflow

- Second clustering attempt
  - **KMeans** after **filtering** small pores

Without filtering



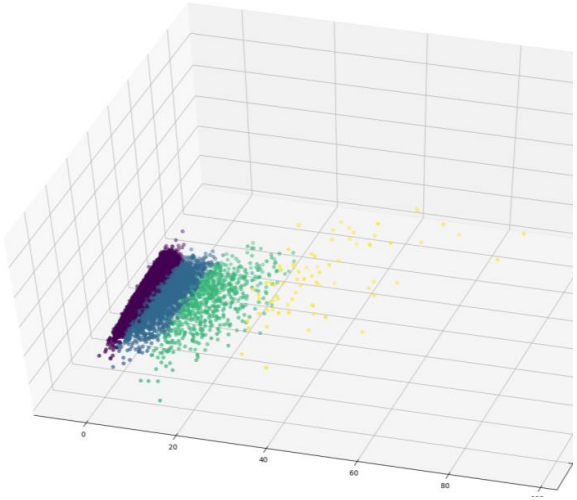
With filtering



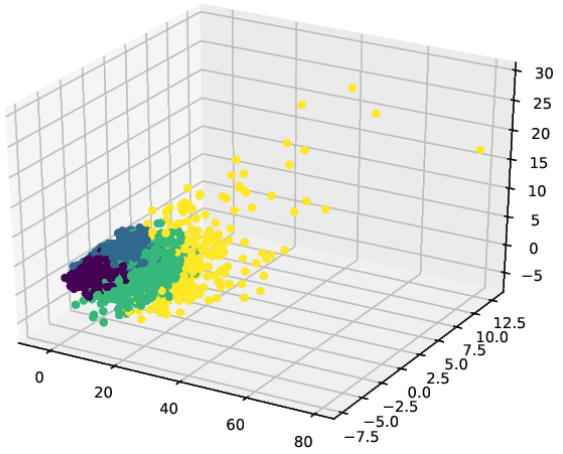
Baseline workflow

- Second clustering attempt
  - **KMeans** after filtering small pores

Without filtering



With filtering

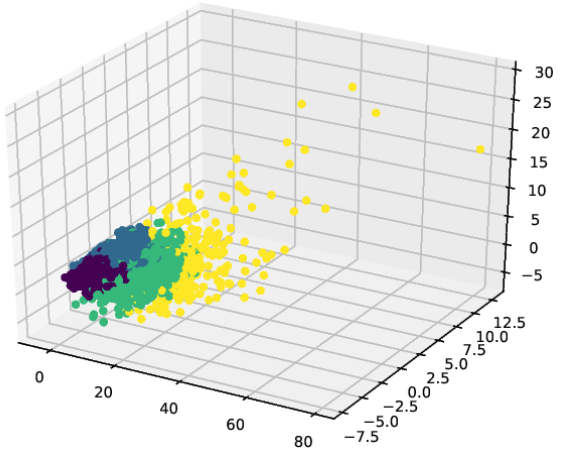


Baseline workflow

- According to domain experts, the new obtained clusters (K=4) start to recognize important geological groups among pores
  - However some of the detected groups should be **further divided** into sub-clusters

Good cluster

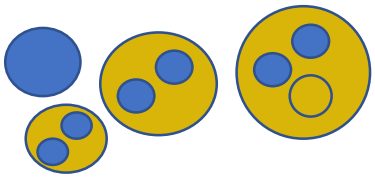
Need for sub-clusters



Baseline workflow

- A hierarchy with super- and sub-clusters may help to create the pore categorization **taxonomy**

Domain knowledge
Taxonomy

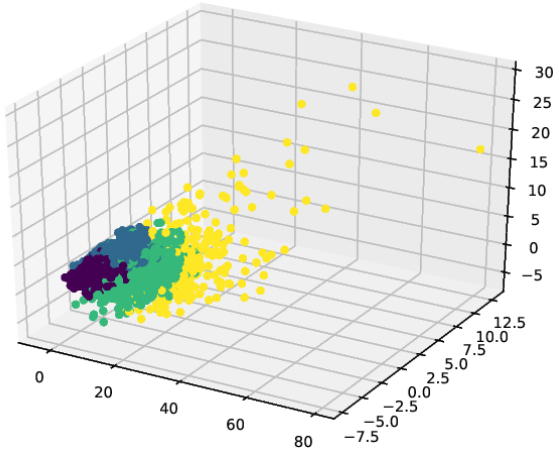


Known category

Discovered category

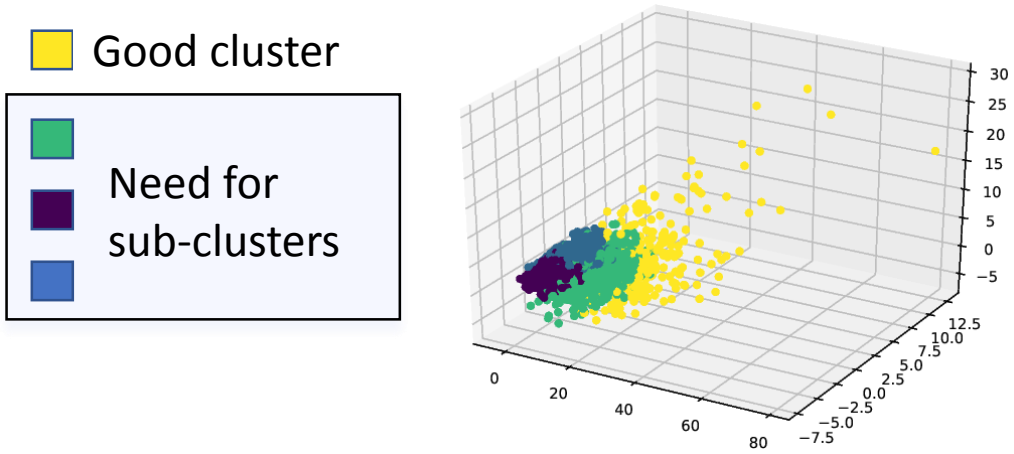
Good cluster

Need for sub-clusters to obtain **known categories**



Baseline workflow

- Solutions
  - Run KMeans on the clusters that need further division
    - Higher computational time, did not give good results
  - **Exploit a hierarchical clustering technique to find multiple levels of clusters**



- **Adaptive Hierarchical Clustering (AHC)**

"Adaptive Hierarchical Clustering for Petrographic Image Analysis"

Andrea Pasini, Elena Baralis, Paolo Garza et al. DARLI-AP 2019, Lisbon

AHC workflow

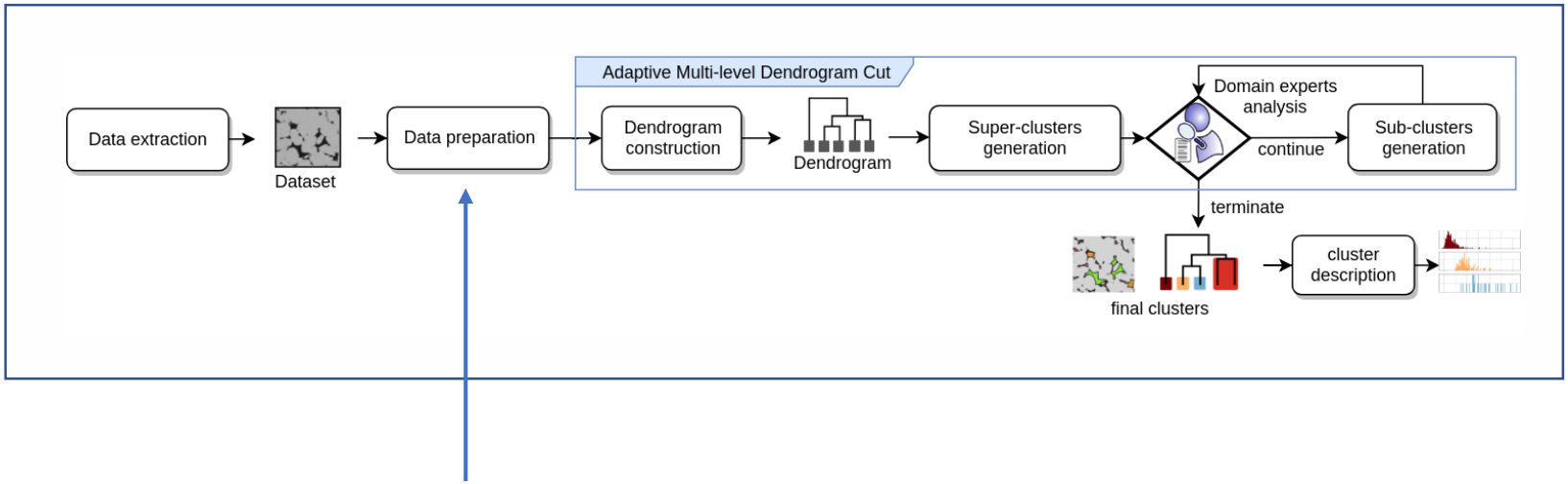
- **Adaptive Hierarchical Clustering (AHC)**
  - We propose a **novel** algorithm that exploits hierarchical clustering dendrograms to inspect multiple level clusters
  - Super-clusters:
    - Few **macro-categories** similar to the ones obtained with K=4 and the baseline workflow
  - Sub-clusters
    - **Further subdivision** of the super-clusters that are not pure according to domain experts

AHC workflow

- **Adaptive Hierarchical Clustering (AHC)**
  - We use Hierarchical clustering with **Ward's linkage**
    - Allows obtaining similar clusters to the ones generated by KMeans
  - The **dendrogram hierarchy** can be exploited to obtain the super- and sub-cluster levels without running multiple times a clustering algorithm

AHC workflow

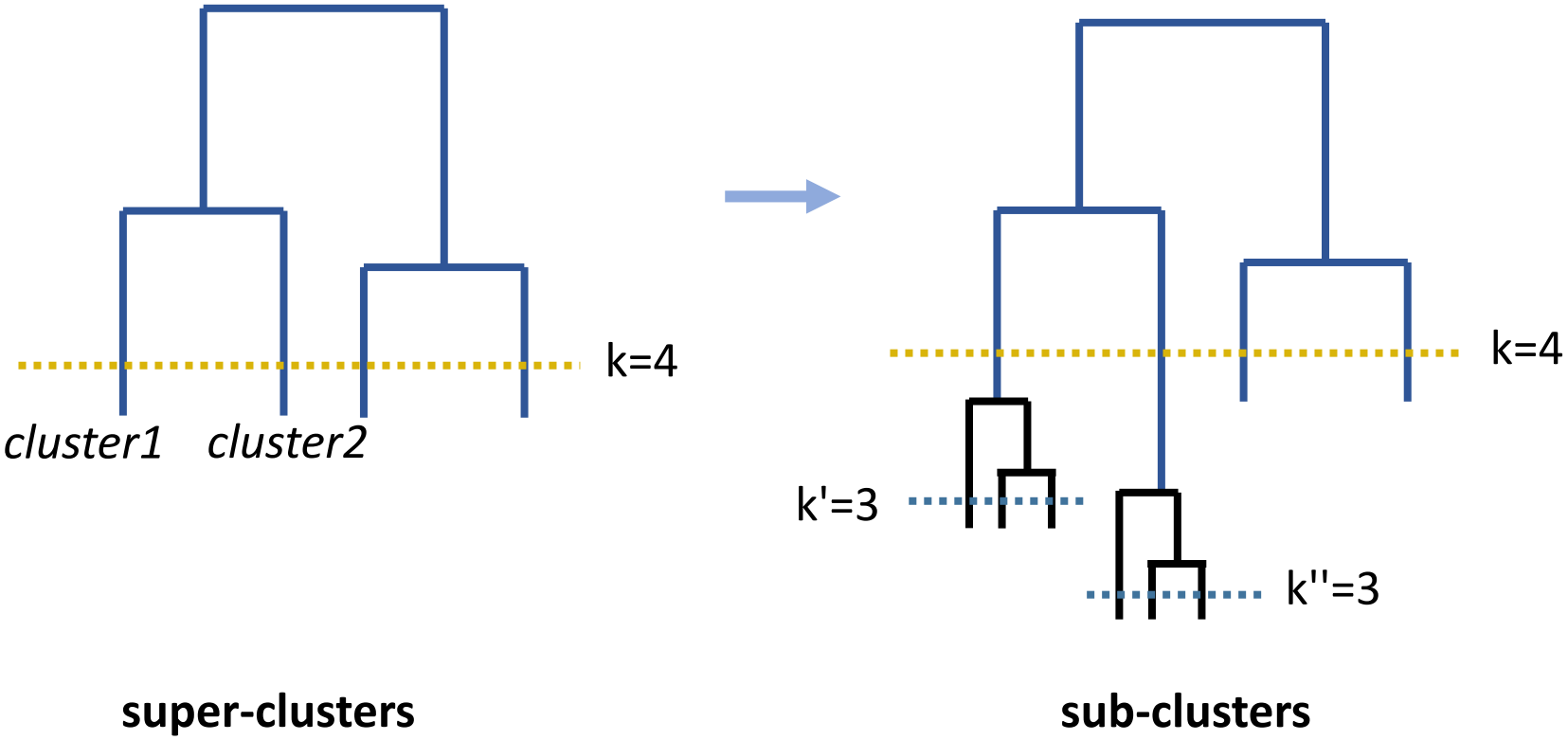
- **Adaptive Hierarchical Clustering (AHC)**



- Feature selection
- Normalization
- Pore filtering by size

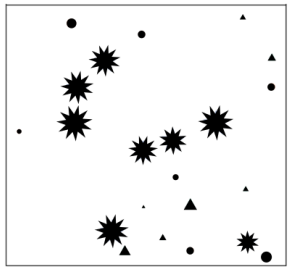
AHC workflow

- **Adaptive Hierarchical Clustering (AHC)**

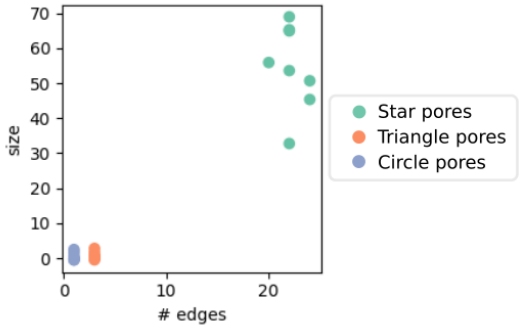


AHC workflow

Artificially generated thin section

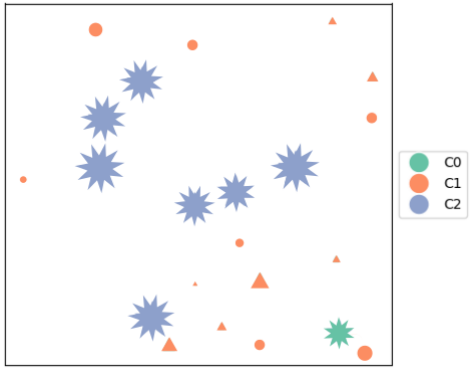


Data distribution



Why standard algorithms do not suit our needs

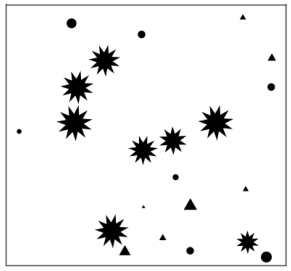
Hierarchical clustering
k=3



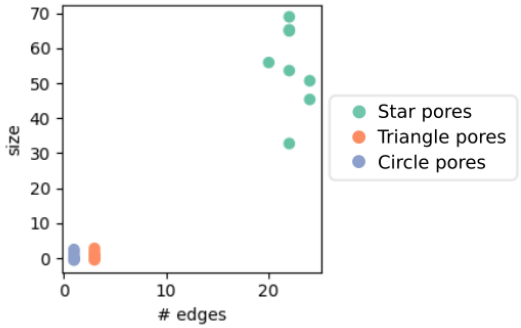
Triangles and **circles** should lie in different clusters

AHC workflow

Artificially generated thin section

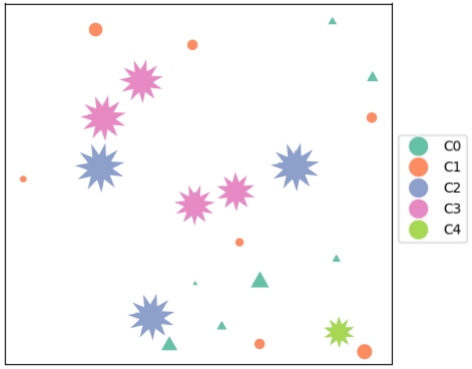


Data distribution



Why standard algorithms do not suit our needs

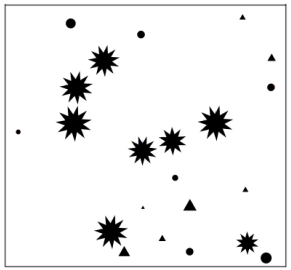
Hierarchical clustering
k=5



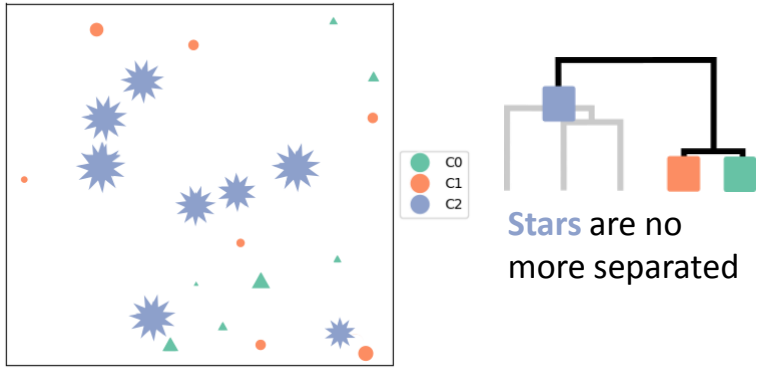
Stars should **not** be separated

AHC workflow

Artificially generated thin section



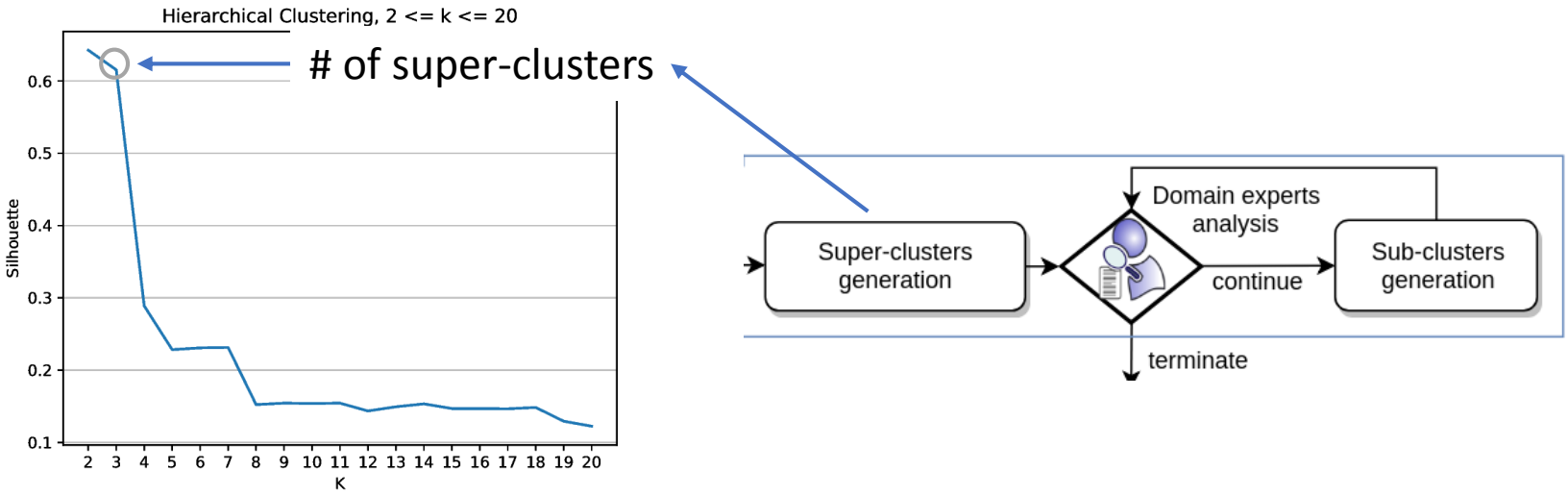
Our technique cuts dendrogram at different heights



Stars are no more separated

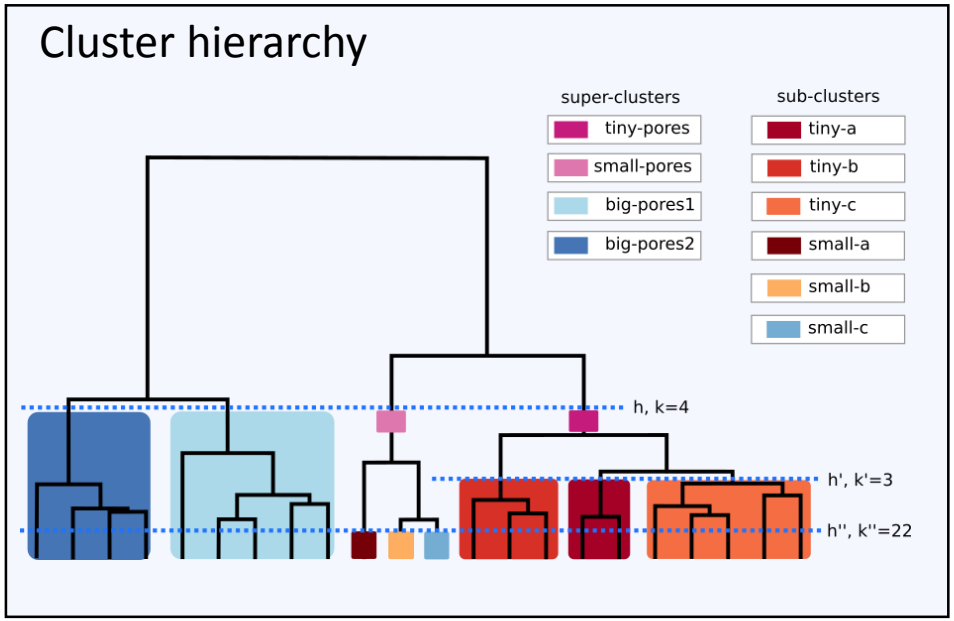
AHC workflow

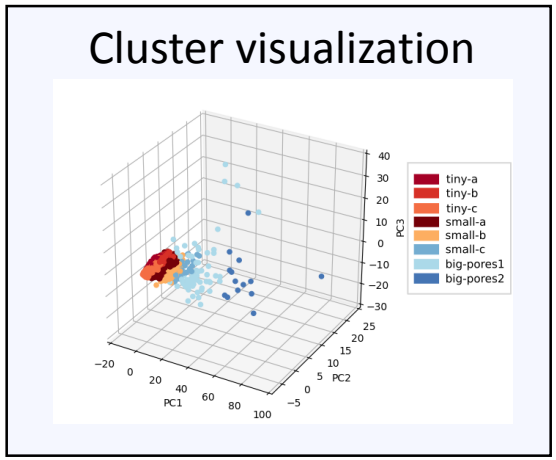
- **super-clusters**
  - Number of clusters decided by means of **silhouette**
- **sub-clusters**
  - Their number is decided by domain experts after SEM images **inspection**

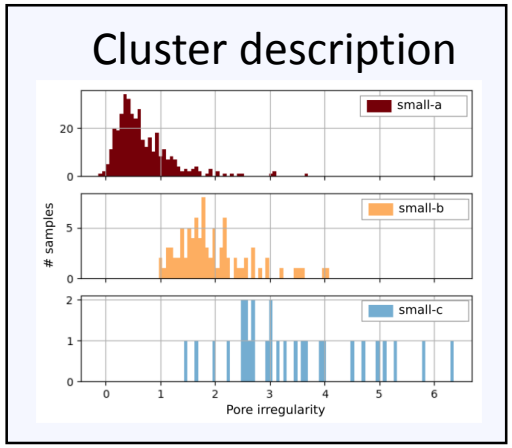


AHC workflow

- Results and discussion







AHC workflow

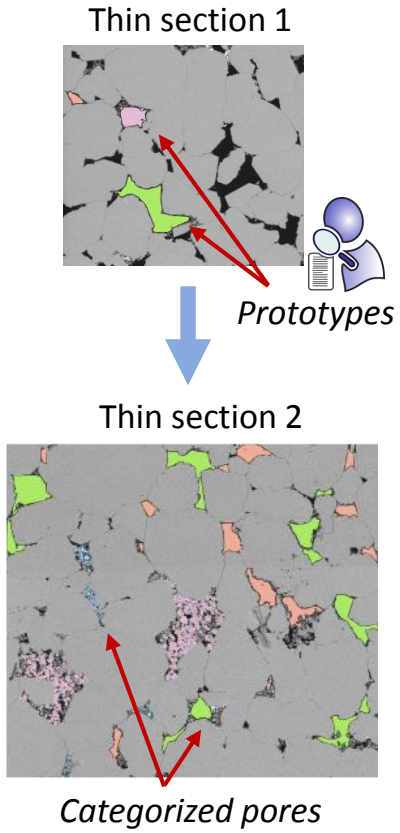
- Results and discussion
  - The proposed pipeline is still characterized by a **human-in-the-loop** technique
    - Number of super-clusters and sub-clusters must be decided **for each thin section** to be analyzed
  - Can we automate the technique?

- **Adaptive hierarchical clustering by means of Prototypes (AdaPro)**

Andrea Pasini, Elena Baralis, Paolo Garza

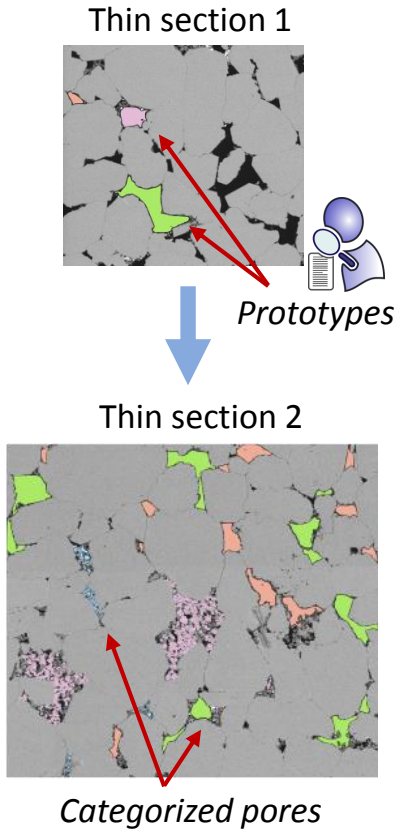
AdaPro workflow

- Geologists provide few labeled pores (**prototypes**)
  - Prototypes extracted from a given thin section should be reused for clustering others
    - Reduced manual effort

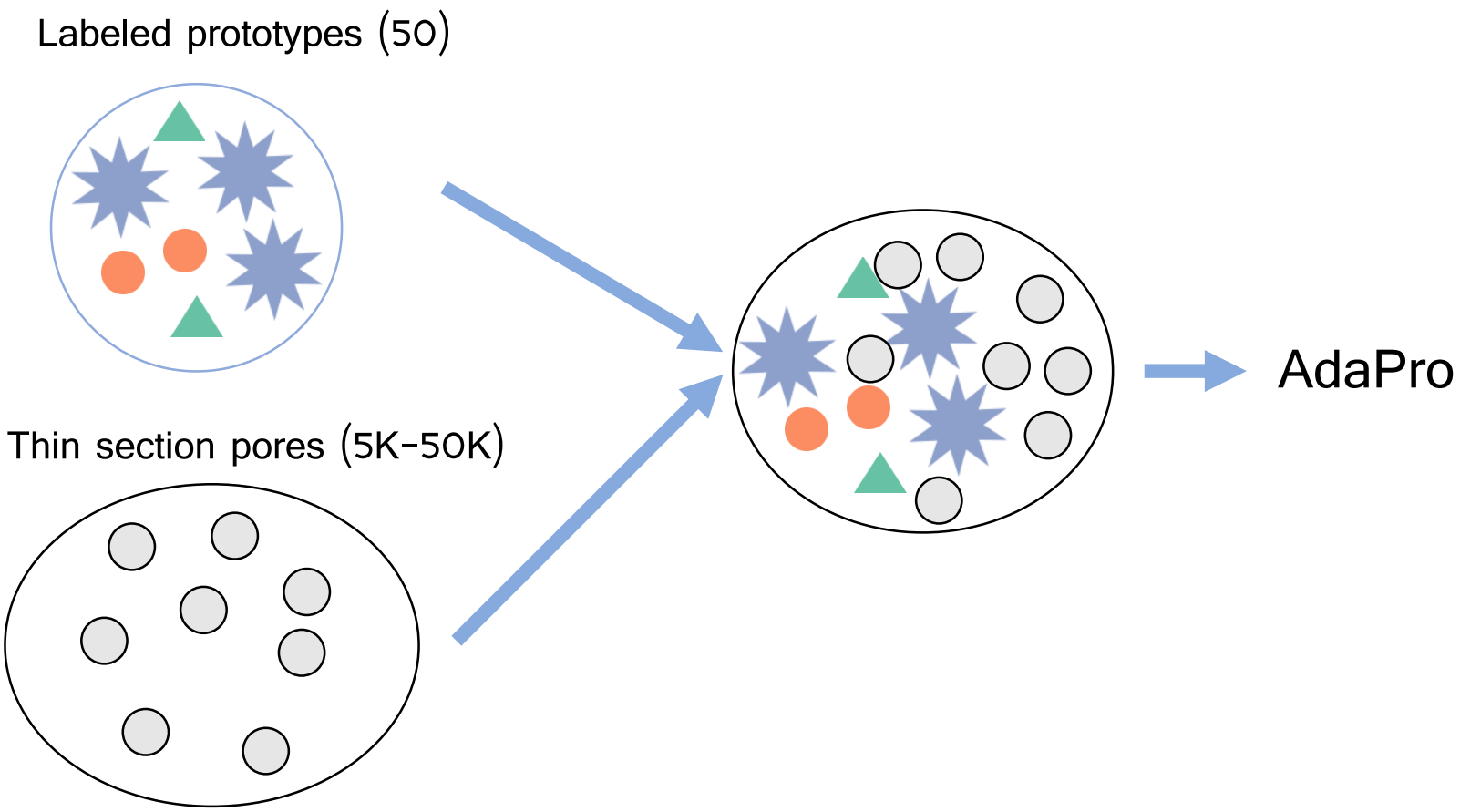


AdaPro workflow

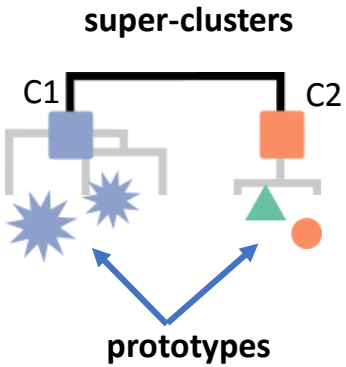
- Prototypes are **too limited** for classification
  - Approximately 50 prototypes (0.07% of the total)
    - 5 pore categories
- Standard clustering approaches cannot exploit labeled data



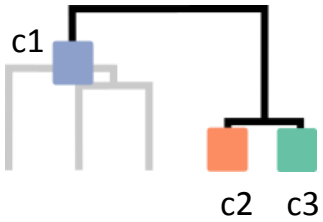
AdaPro workflow



AdaPro workflow



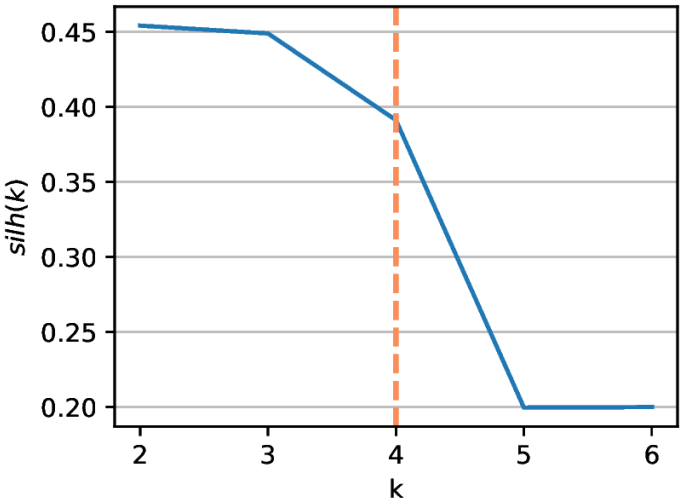
Exploit the **prototype** distribution in each super-cluster to decide which of them should be further divided



Super-clusters that are not pure are split in **sub-clusters**

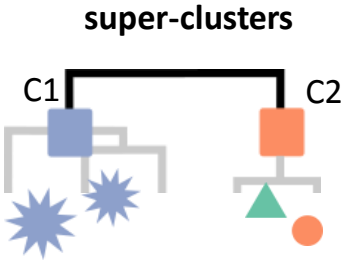
AdaPro workflow

- **Automatic** discovery of super-clusters by means of **silhouette**
  - Allows discovering unknown **coarse-grained** categories



$$\hat{k} = \underset{k}{\operatorname{argmin}}(silh(k+1) - silh(k)), \quad 2 <= k <= 6$$

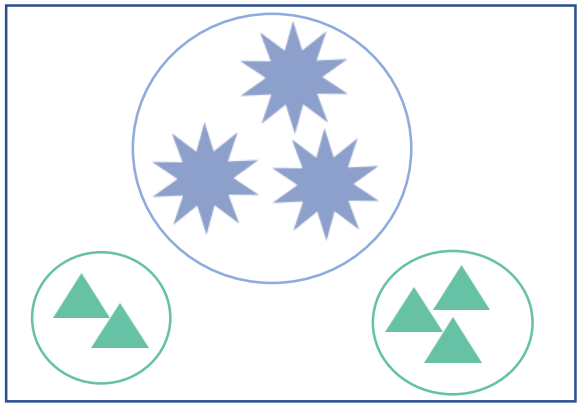
AdaPro workflow



- **Automatic** discovery of the number of sub-clusters
  - Divide super-clusters to obtain a pure subdivision

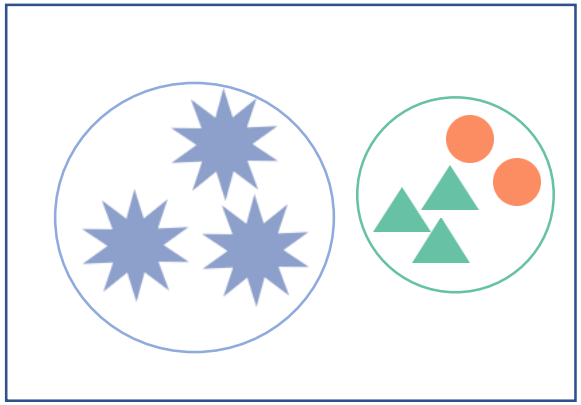
- Homogeneity and completeness scores of **prototypes** inside sub-clusters

Homogeneous but not complete



Over-splitting

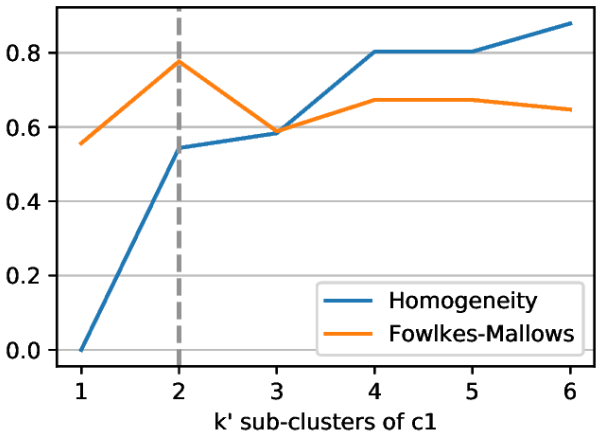
Complete but not homogeneous

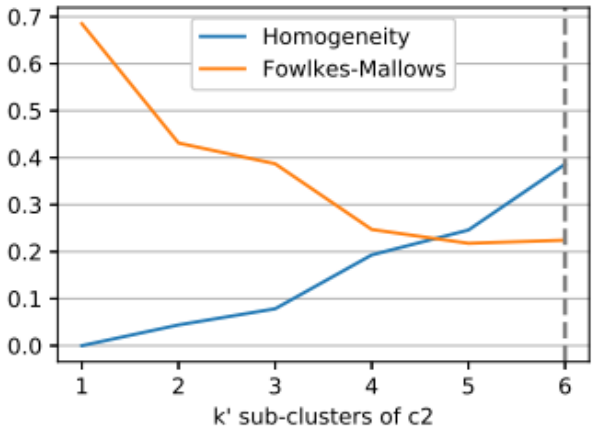


Under-splitting

AdaPro workflow

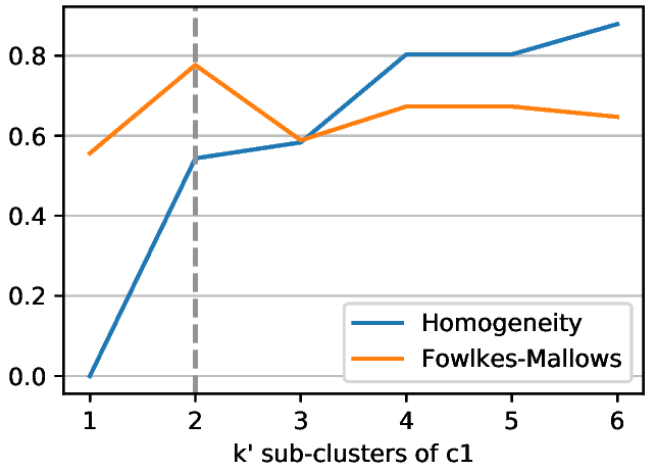
- **Homogeneity** and **completeness** scores
  - These scores should be optimized with sub-clusters
- **Fowlkes-Mallows (FM)** score averages the effects of homogeneity and completeness
  - Compute FM for each super-cluster and for different values of k' (number of sub-clusters)





AdaPro workflow

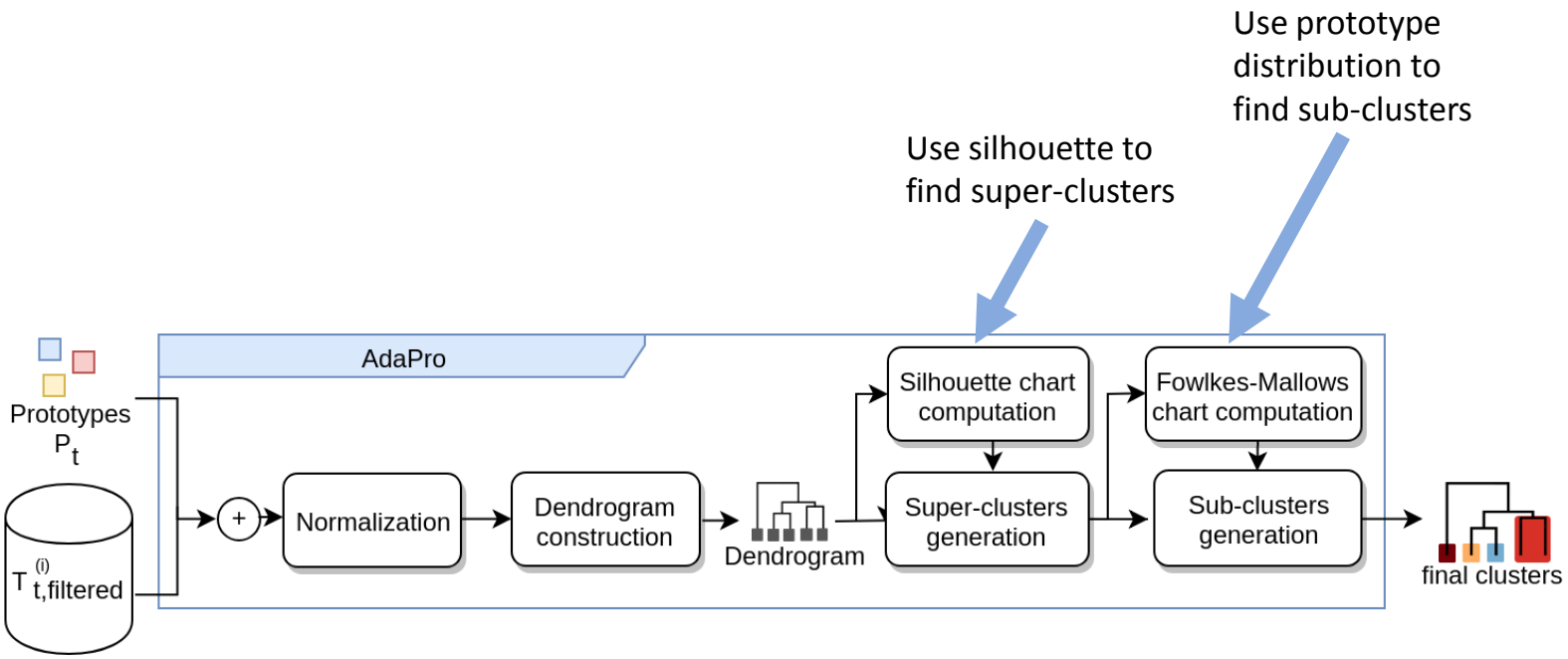
- **Automatic** discovery of sub-clusters
  - Find regions where FM increases with k'



$$K' = \{fowlkes(k'-1) - fowlkes(k') > 0\}$$

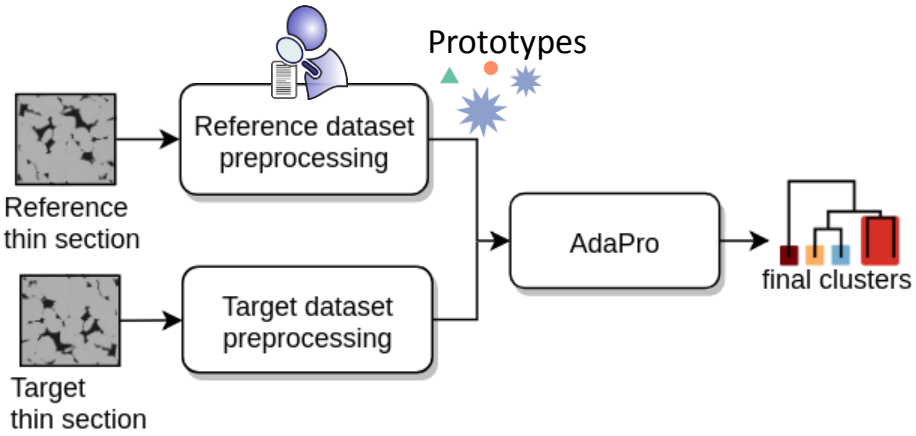
$$\hat{k'} = \underset{k'}{\operatorname{argmax}}(fowlkes(k'), k' \in K')$$

AdaPro workflow



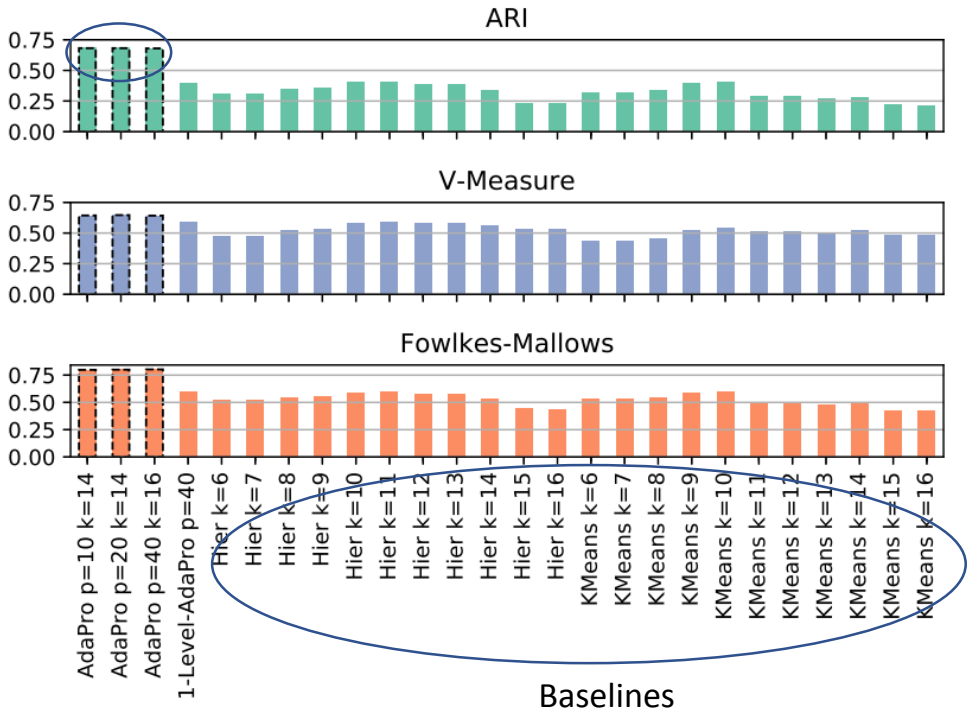
AdaPro workflow

- AdaPro – Use case
  - Experts manually label a small subsample of prototypes in a **reference** thin section
  - AdaPro takes as input the prototypes and a **target** thin section
    - **Advantage**: prototypes are **reusable** for different thin sections



AdaPro workflow

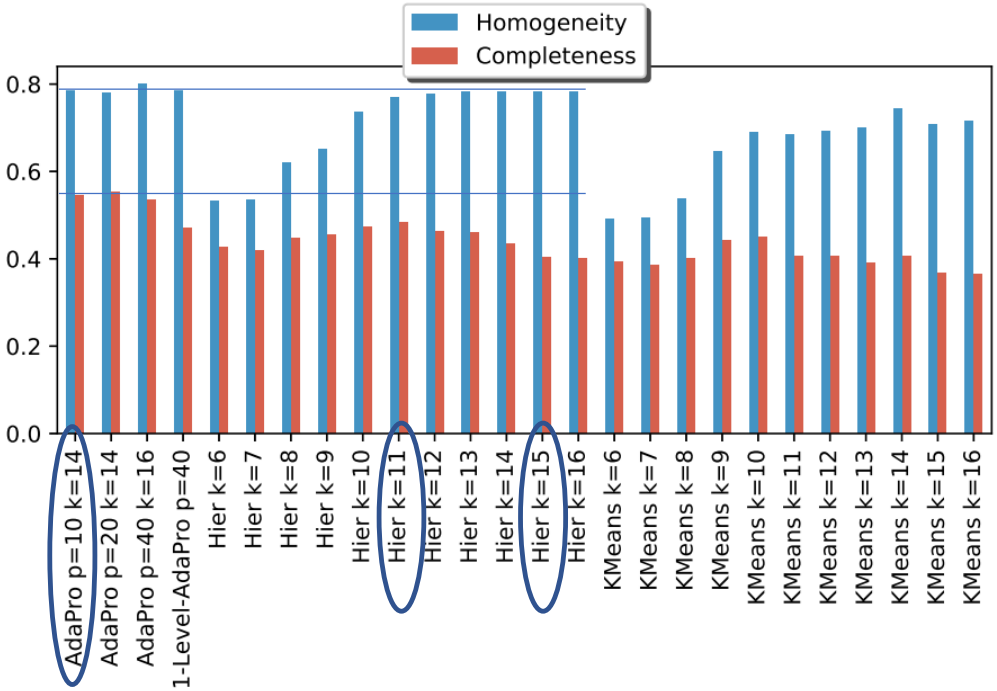
- Evaluation on a UCI dataset (non-geological domain)



- For this dataset ground truth labels are available
  - ARI, V-Measure, Fowlkes-Mallows measure the similarity between clustering results and ground truth
- Best performances for AdaPro

AdaPro workflow

- Evaluation on a UCI dataset (non-geological domain)



- Hierarchical clustering:
  - k=15 homogeneous but not complete
  - k=11 complete but not homogeneous
- AdaPro:
  - Reaches high homogeneity whilst keeping high completeness

Thank you for the attention
Any questions?