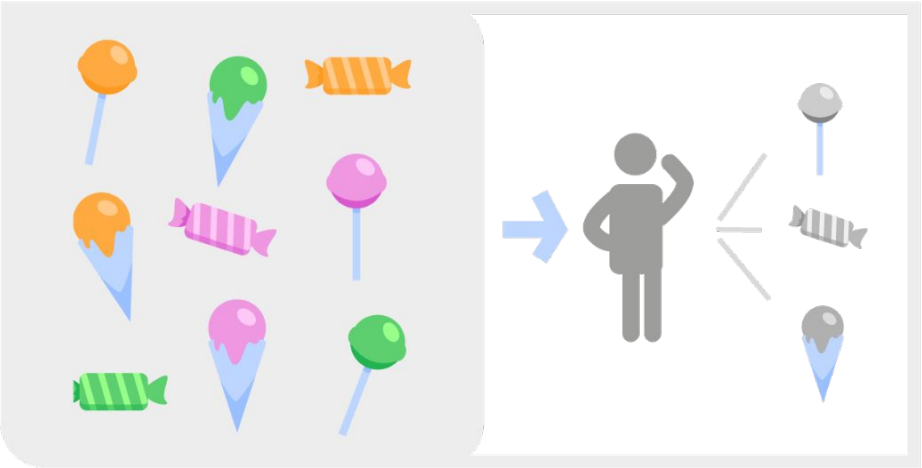
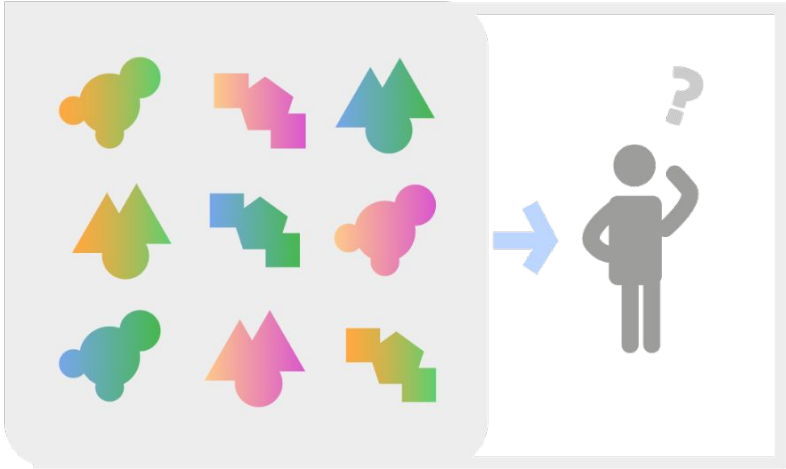
Constrained Clustering
Challenges and Frontiers

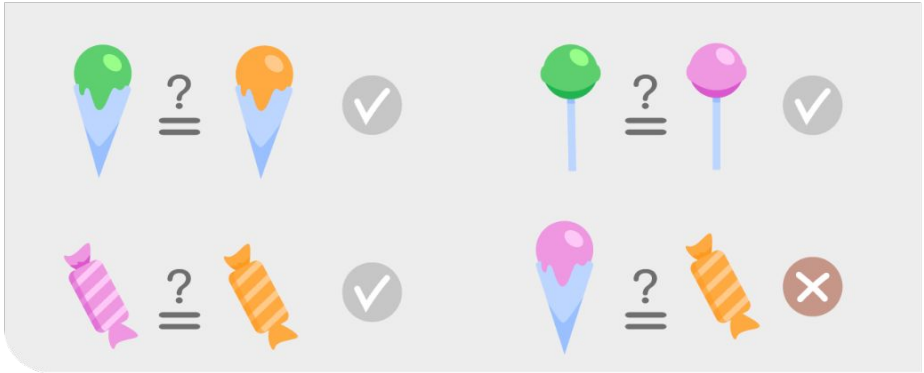
Dr. Germán González Almagro

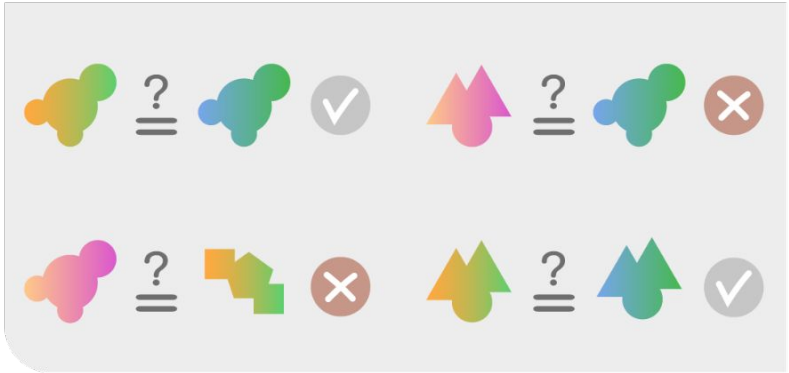
Categorical Info





Relational Info





Semi-Supervised Learning

Semi-Supervised Clustering

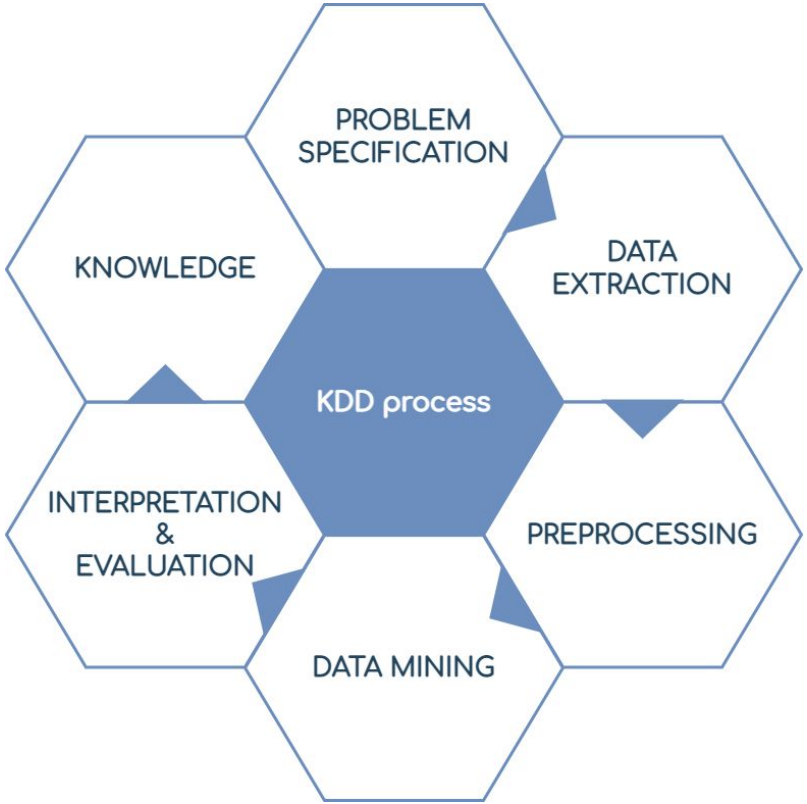
Constrained Clustering

Hybrid Models for CC

New Paradigms in CC

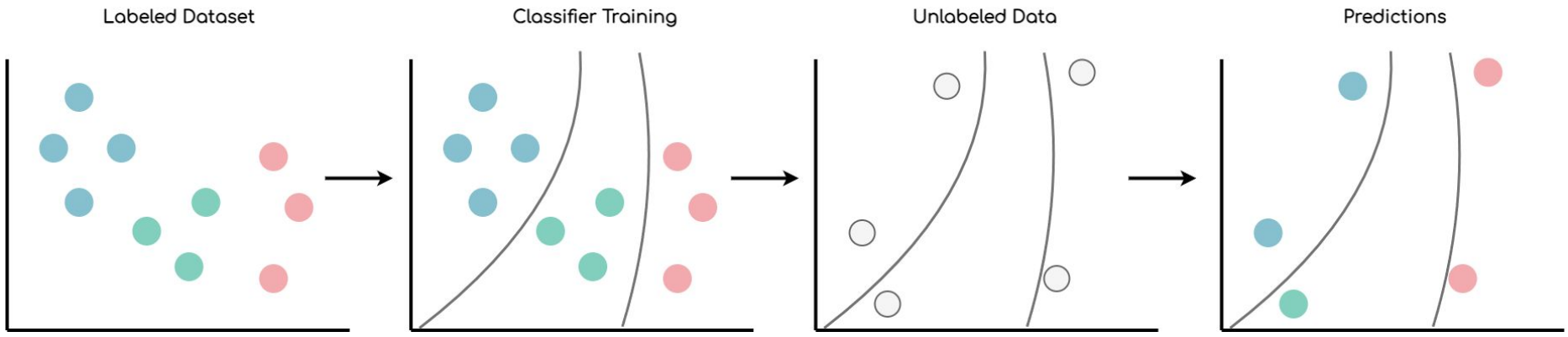
Challenges and Frontiers

Knowledge Discovery in Databases (KDD)

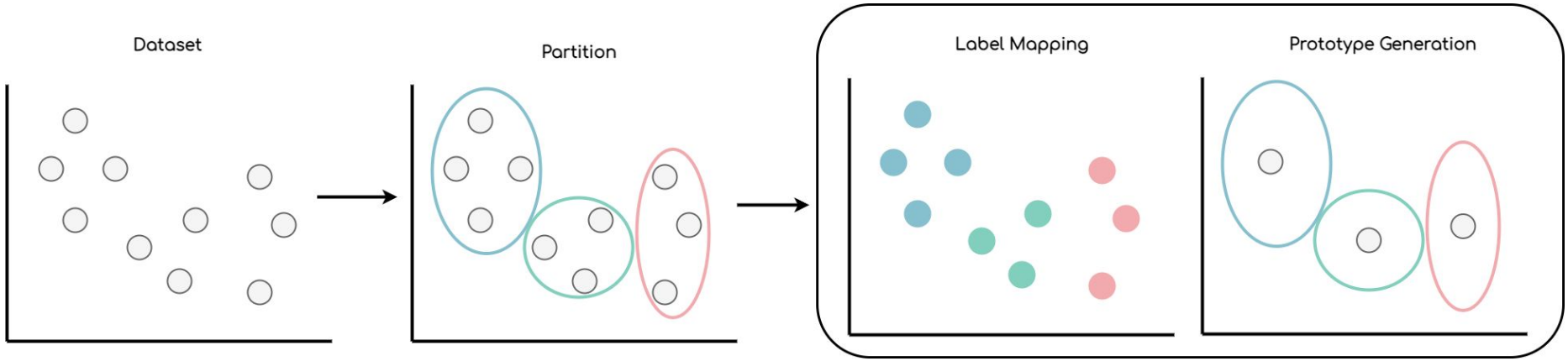


KDD: a set of stages that enable the identification of valuable patterns and relationships within the data. **Our goal is to extract meaningful insights from the data.**

Supervised Learning: the goal is to build a classifier or regressor that, trained with a set of examples and their corresponding output values, can **predict the value of unseen inputs**.

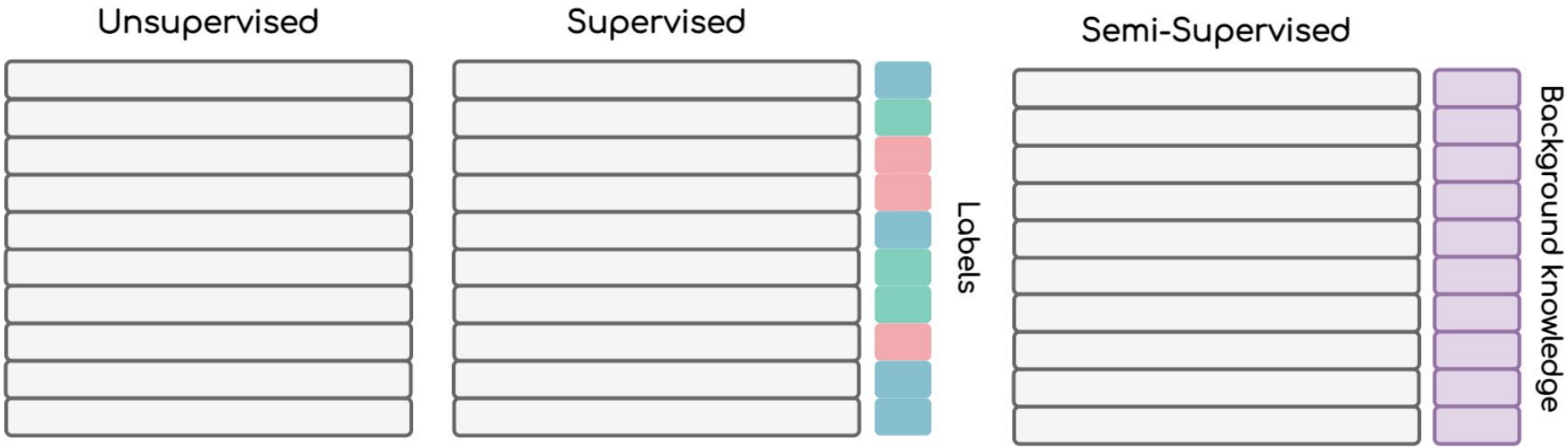


Unsupervised Learning: only the set of examples is available, and no output value is provided. The goal here is to **discover some underlying structure in the data.**



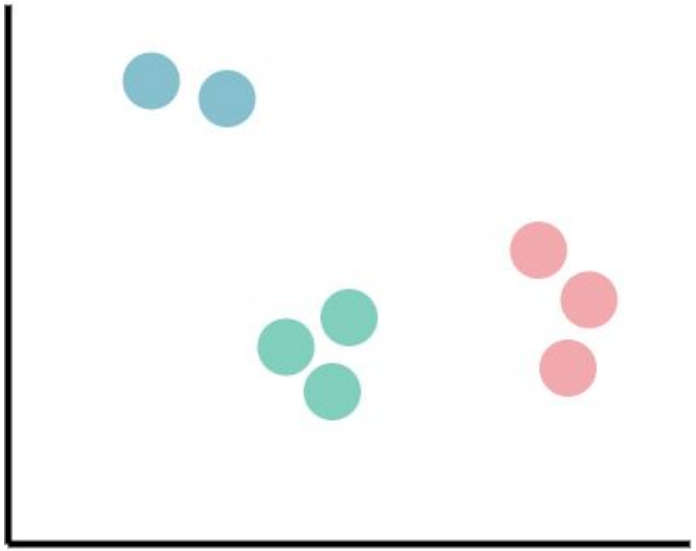
Semi-Supervised Learning (SSL)

Semi-Supervised Learning: tries to combine the benefits of supervised and unsupervised learning by making use of both labeled and unlabeled data, or **other kinds of expert knowledge.**

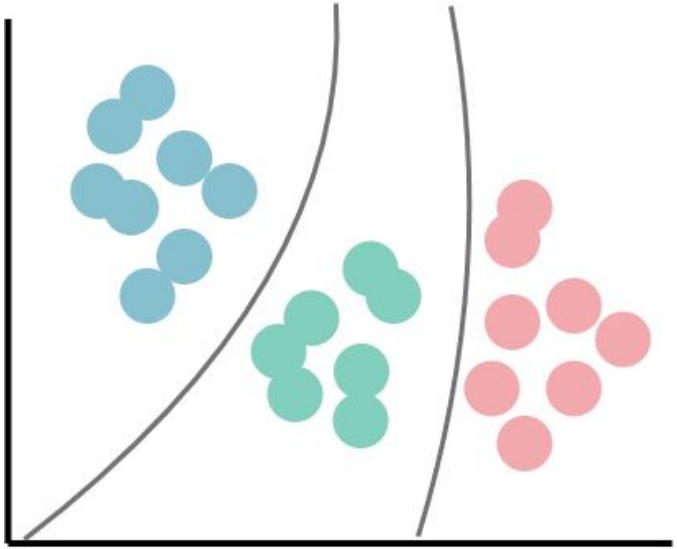


Assumptions

Smoothness Assumption.
Two instances that are close in the input space should have the same label.



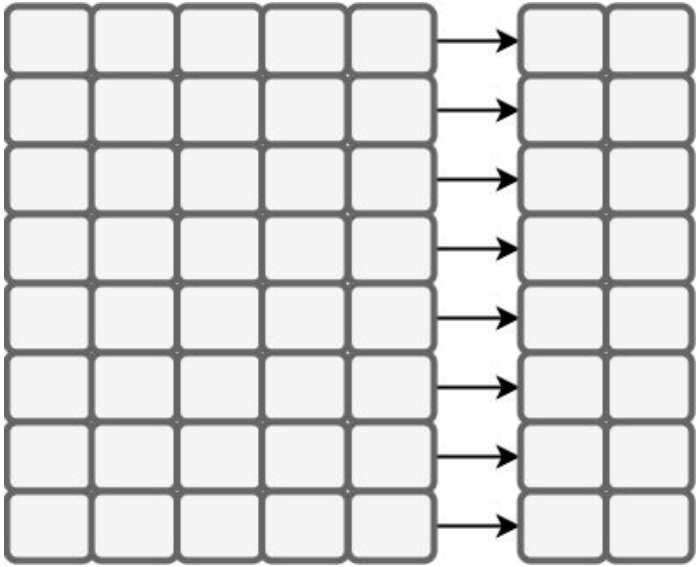
Low-density Assumption.
Decision boundaries should preferably pass through low-density spatial regions.



Assumptions

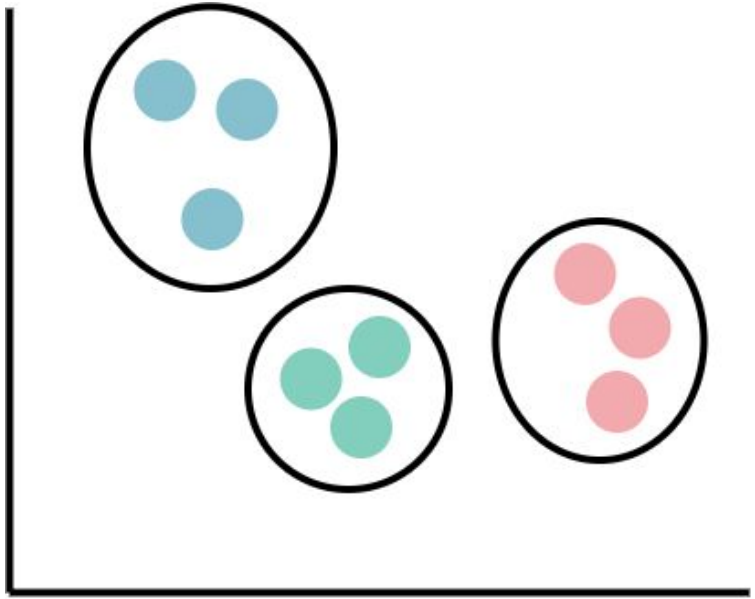
Manifold Assumption

Instances in a high-dimensional input space are usually gathered along lower-dimensional structures.



Cluster Assumption

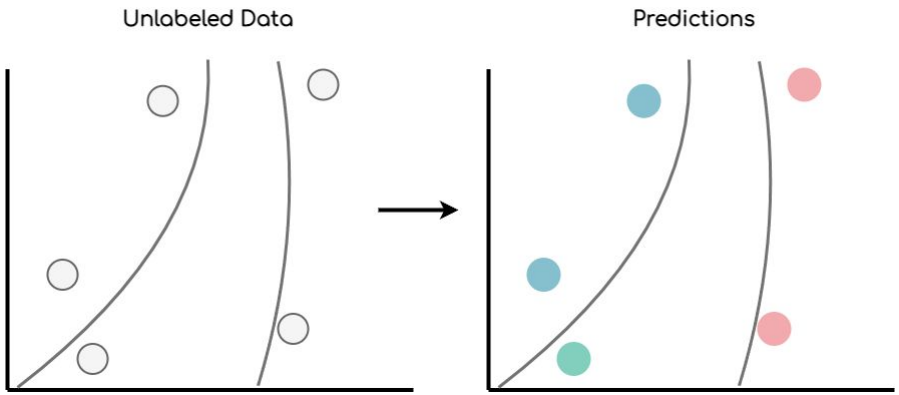
Instances which belong to the same cluster also belong to the same class.

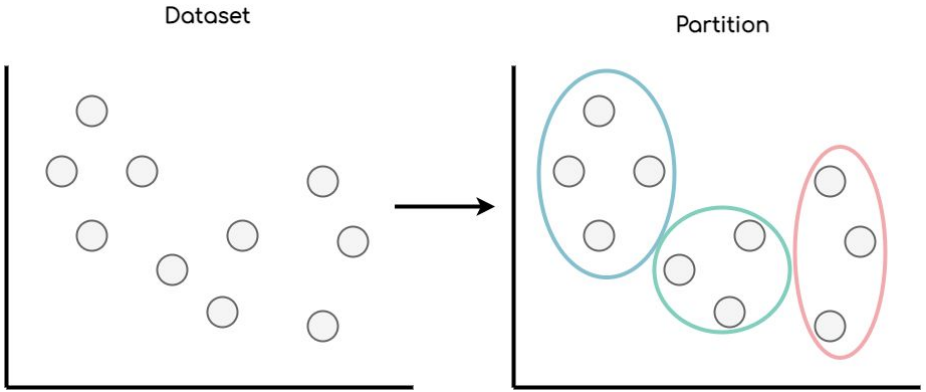


Dichotomy

Inductive Methods: aim to build a classifier capable of outputting a label for any instance in the input space. The predictions for unseen instances are independent. **Inductive methods in SSL are categorized as semi-supervised classification.**

Transductive Methods: their predictions are limited to the data used during the training phase. Transductive methods do not have separated training and testing phases. **Transductive methods in SSL are categorized as semi-supervised clustering.**





Semi-Supervised Learning

Semi-Supervised Clustering

Constrained Clustering

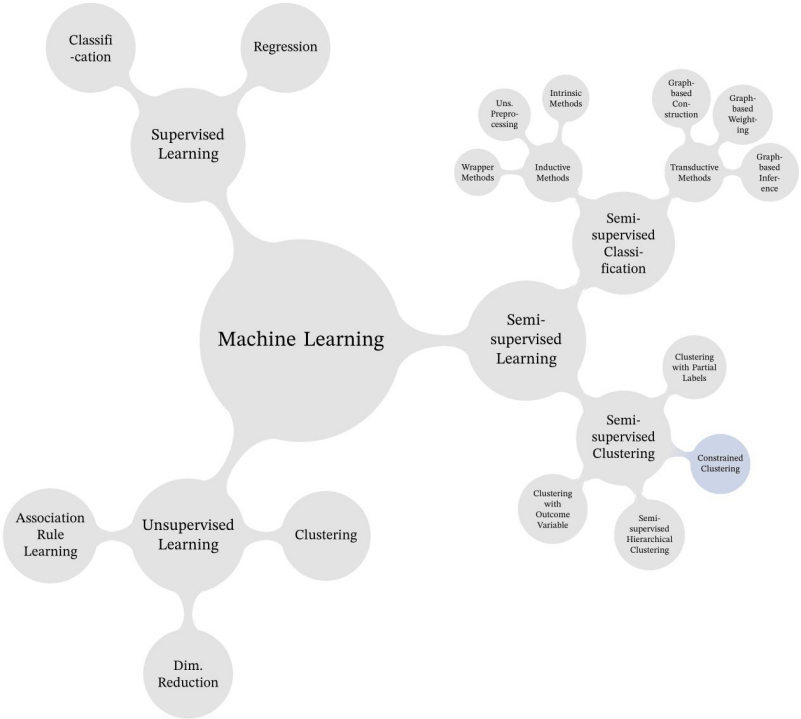
Hybrid Models for CC

New Paradigms in CC

Challenges and Frontiers

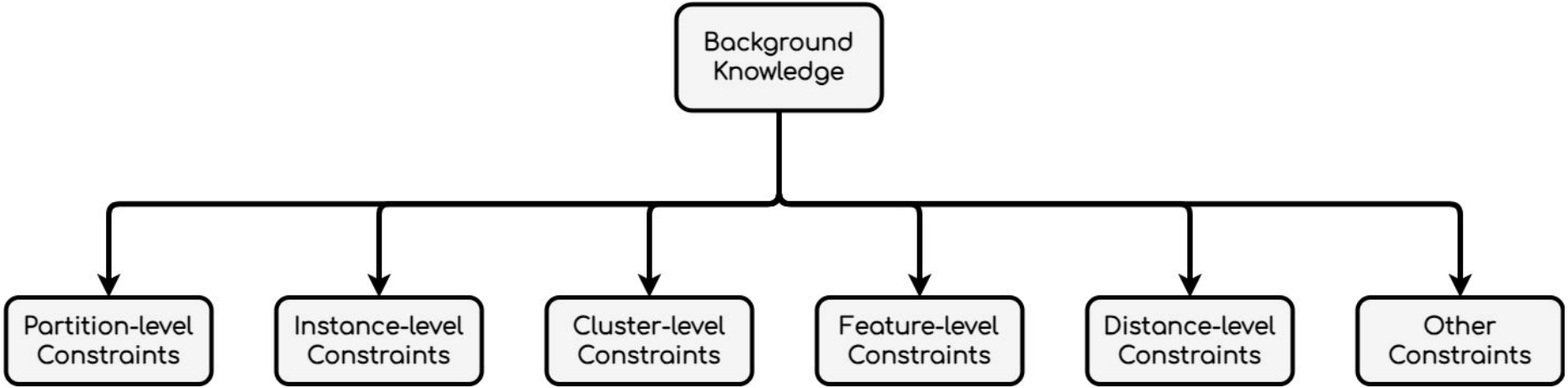
Semi-Supervised Clustering

Semi-Supervised Learning: tries to combine the benefits of supervised and unsupervised learning by making use of both labeled and unlabeled data, or **other kinds of expert knowledge.**



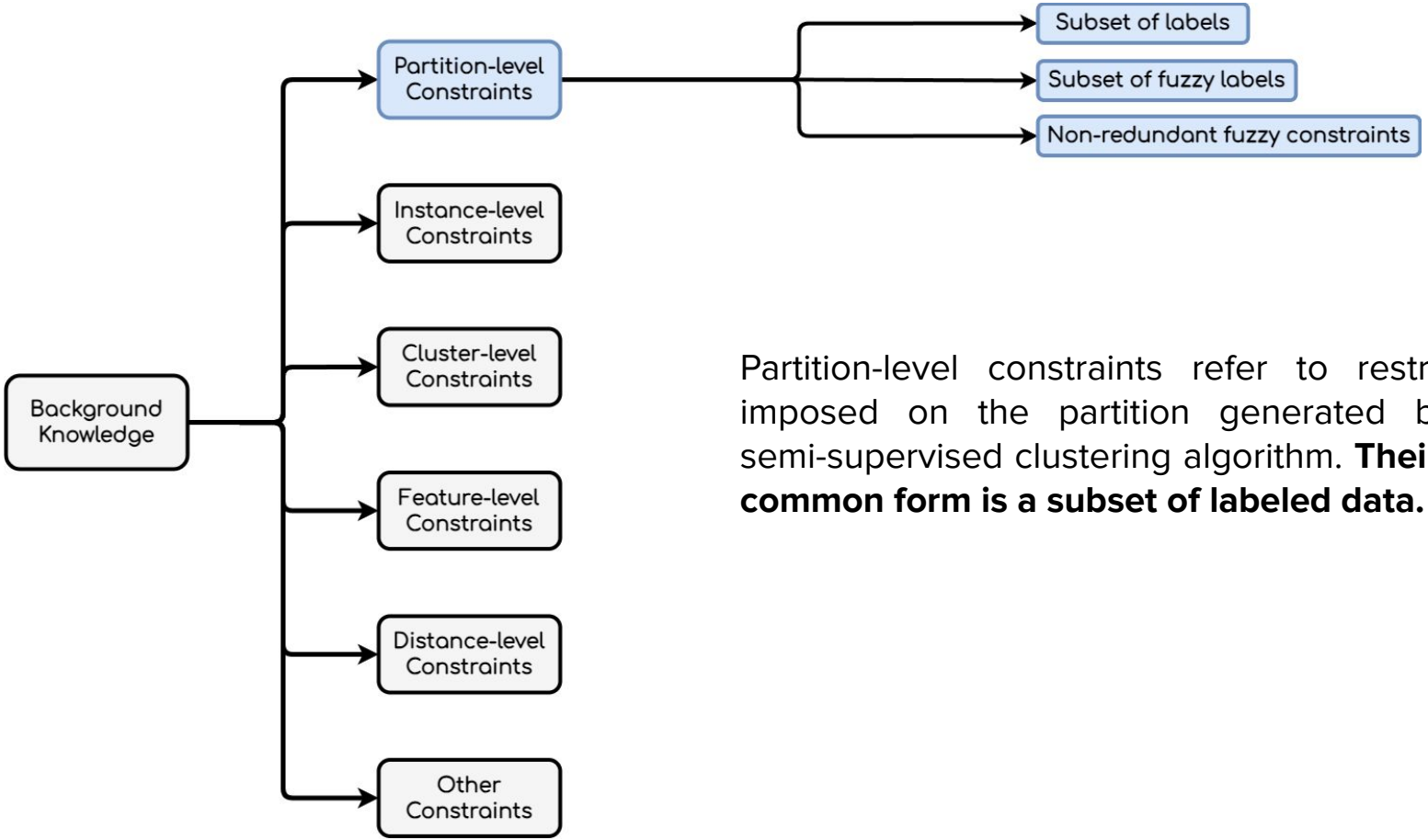
Semi-Supervised Clustering: in addition to the unlabeled dataset, background knowledge is given to perform clustering. When the **background knowledge is given in the form of constraints**, the resulting clustering paradigm is known as **Constrained Clustering (CC).**

Background Knowledge



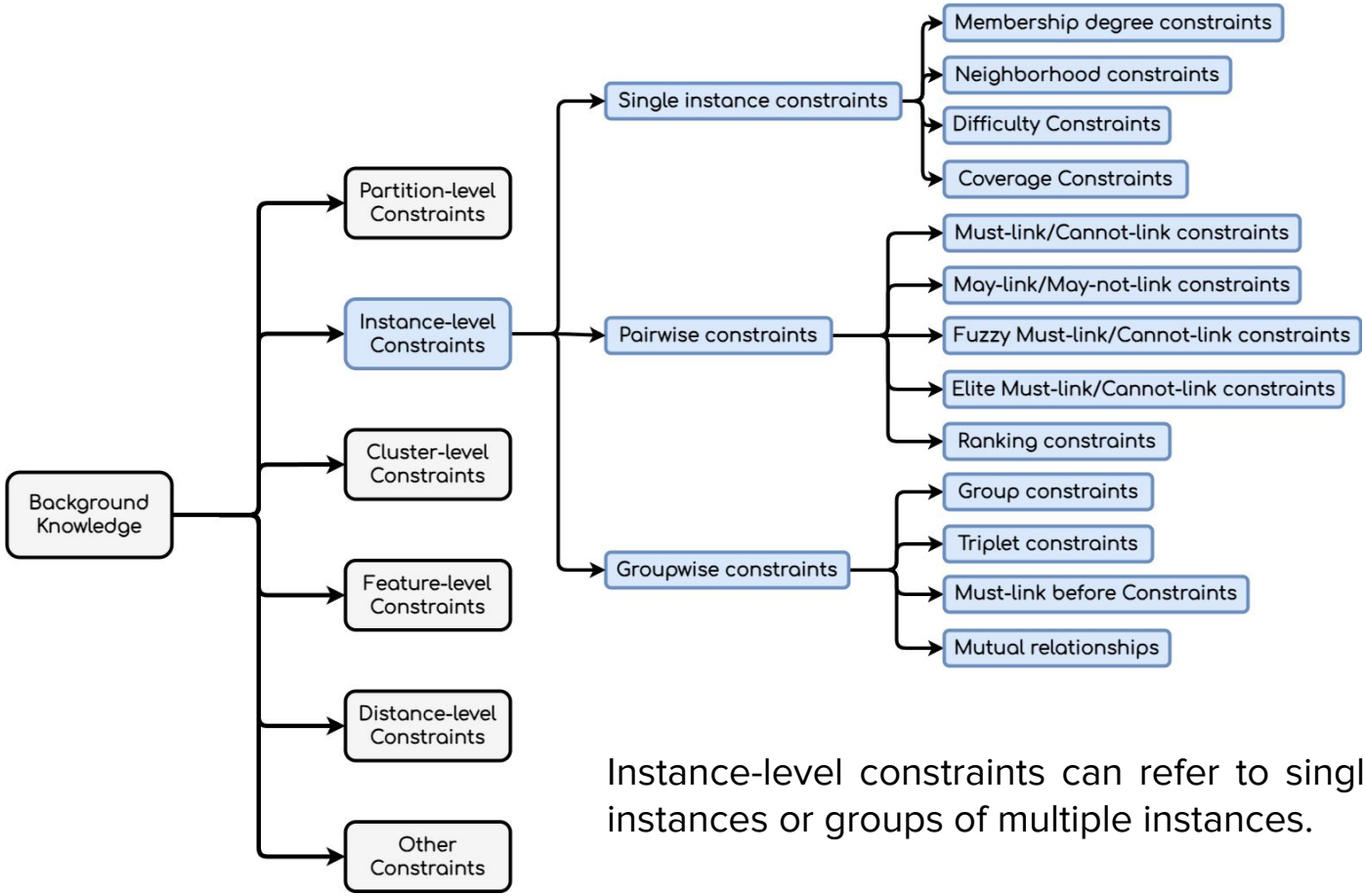
Background Knowledge can be found in many forms, which can be categorized as follows.

Background Knowledge



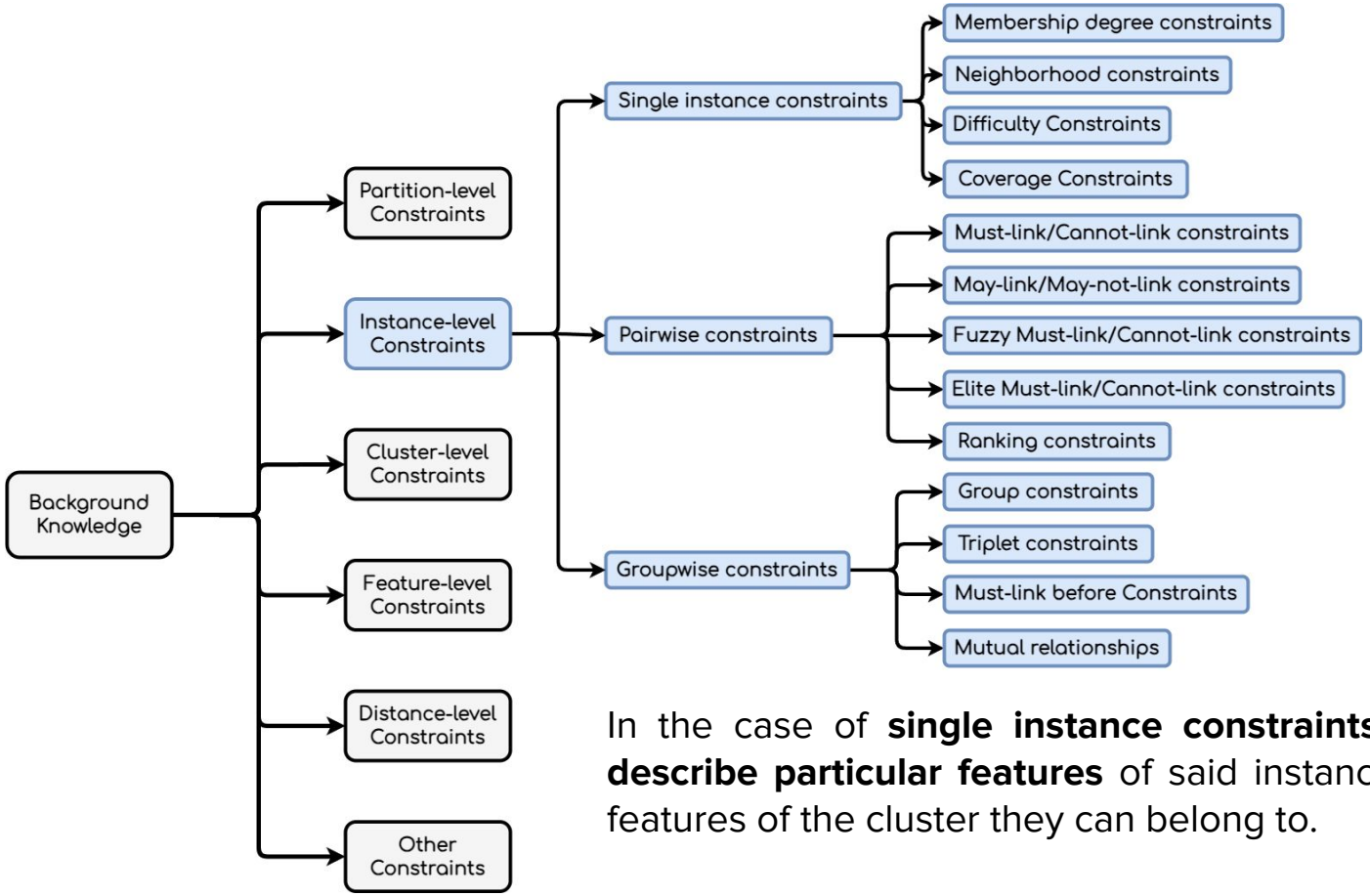
Partition-level constraints refer to restrictions imposed on the partition generated by the semi-supervised clustering algorithm. **Their most common form is a subset of labeled data.**

Background Knowledge



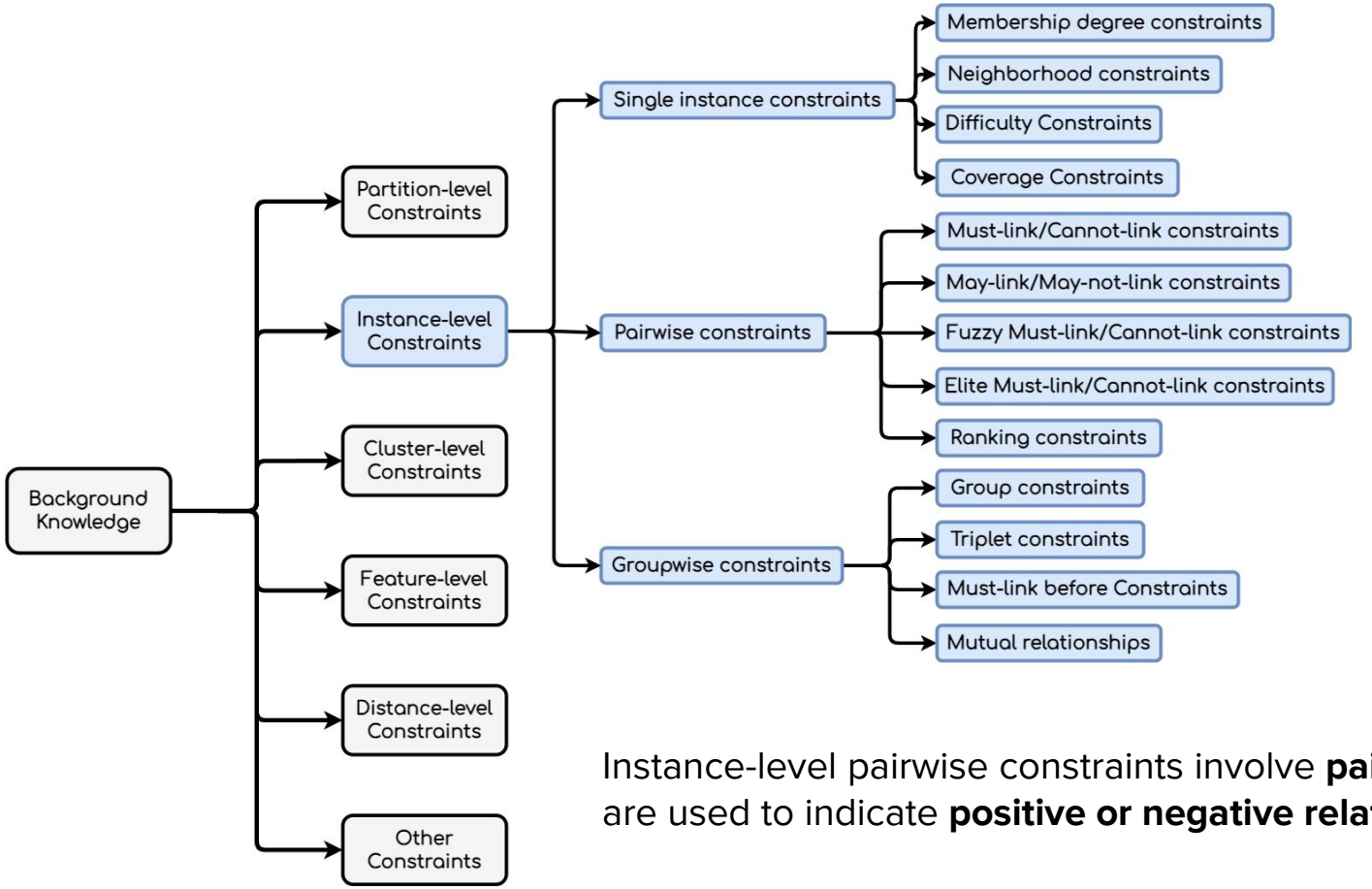
Instance-level constraints can refer to single instances, pairs of instances or groups of multiple instances.

Background Knowledge



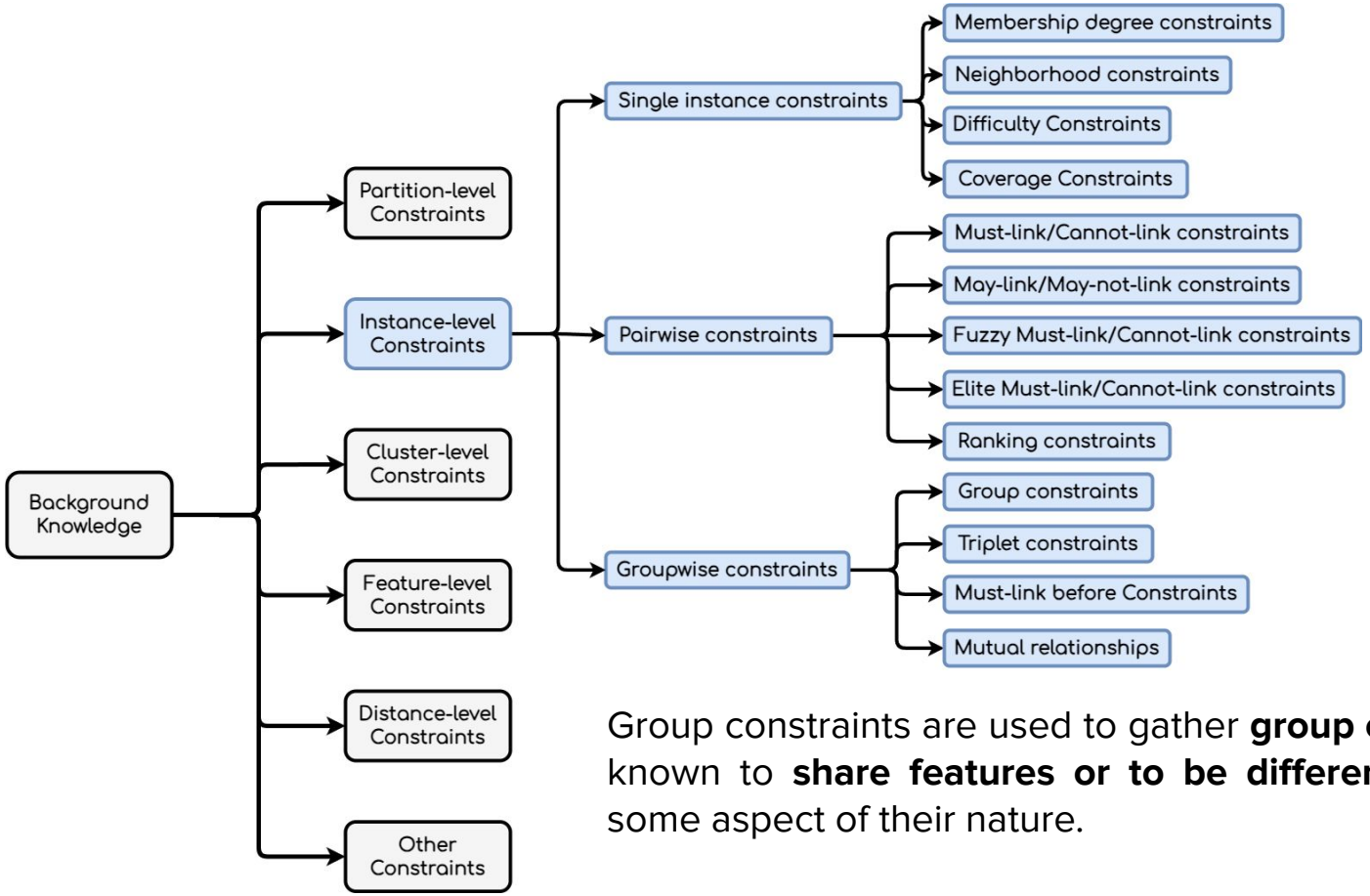
In the case of **single instance constraints**, they are used to **describe particular features** of said instances or to restrict the features of the cluster they can belong to.

Background Knowledge



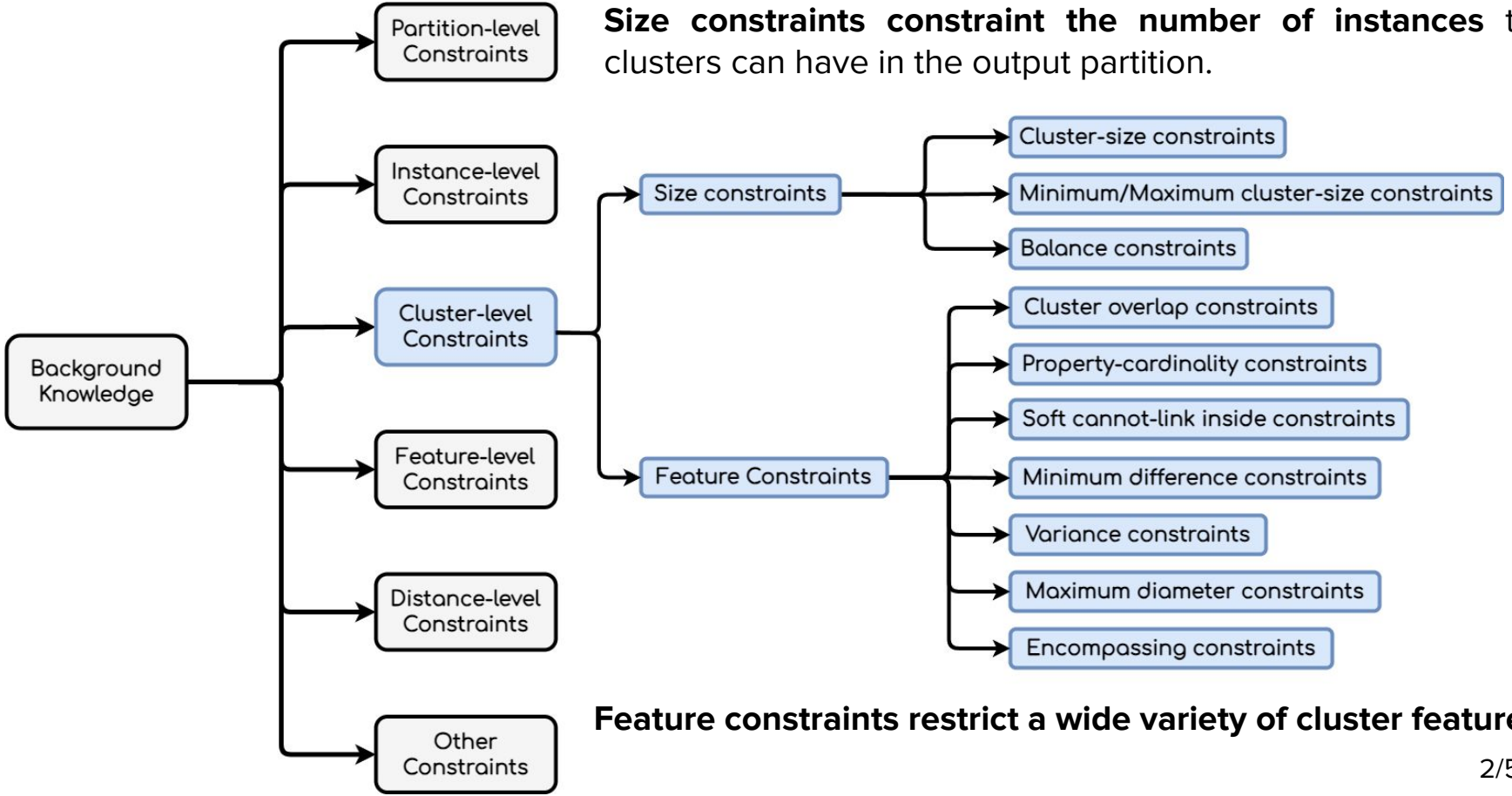
Instance-level pairwise constraints involve **pairs of instances** and are used to indicate **positive or negative relationships**.

Background Knowledge



Group constraints are used to gather **group of instances** that are known to **share features or to be different to each other** in some aspect of their nature.

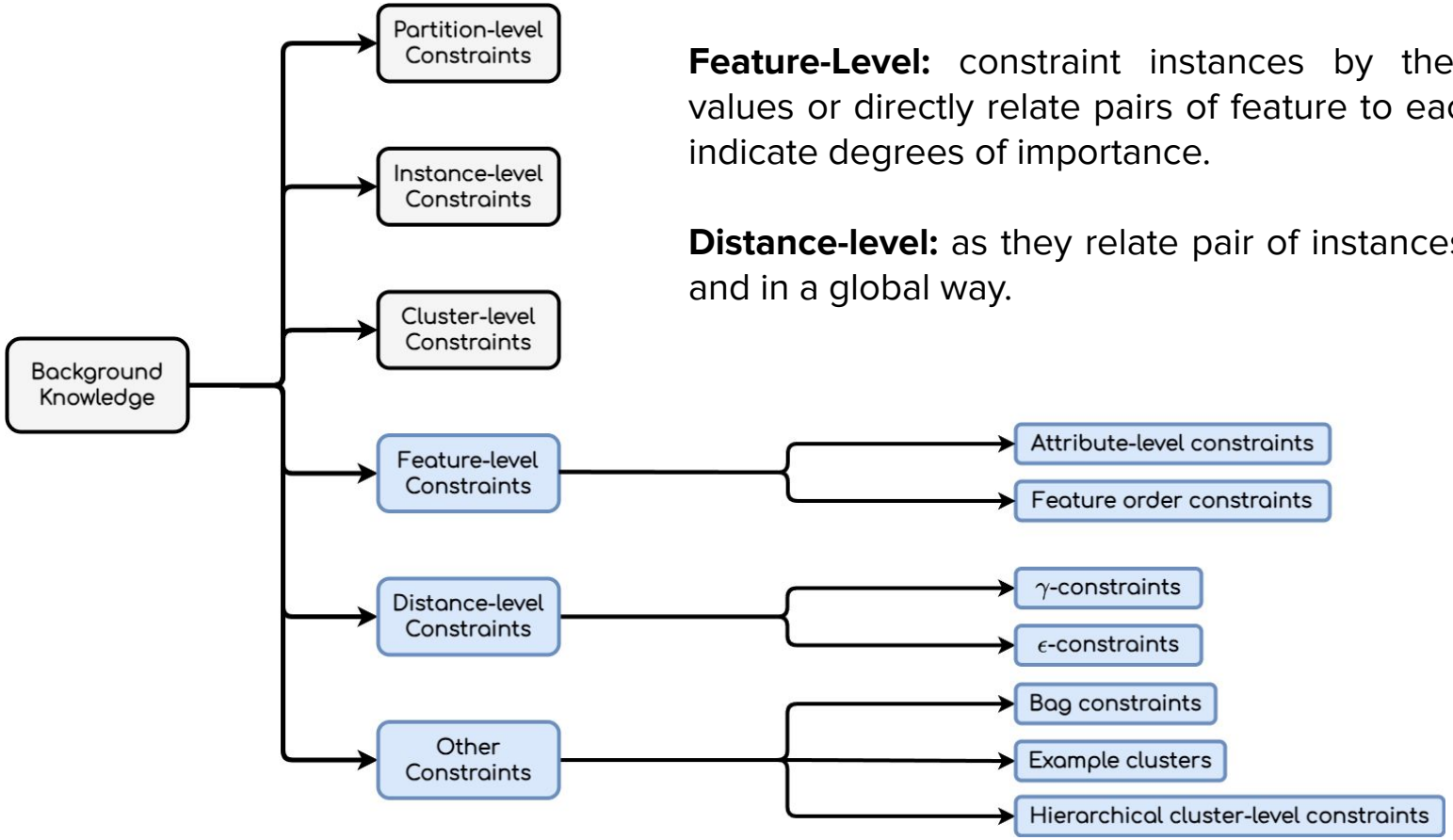
Background Knowledge



Size constraints constraint the number of instances that clusters can have in the output partition.

Feature constraints restrict a wide variety of cluster features.

Background Knowledge

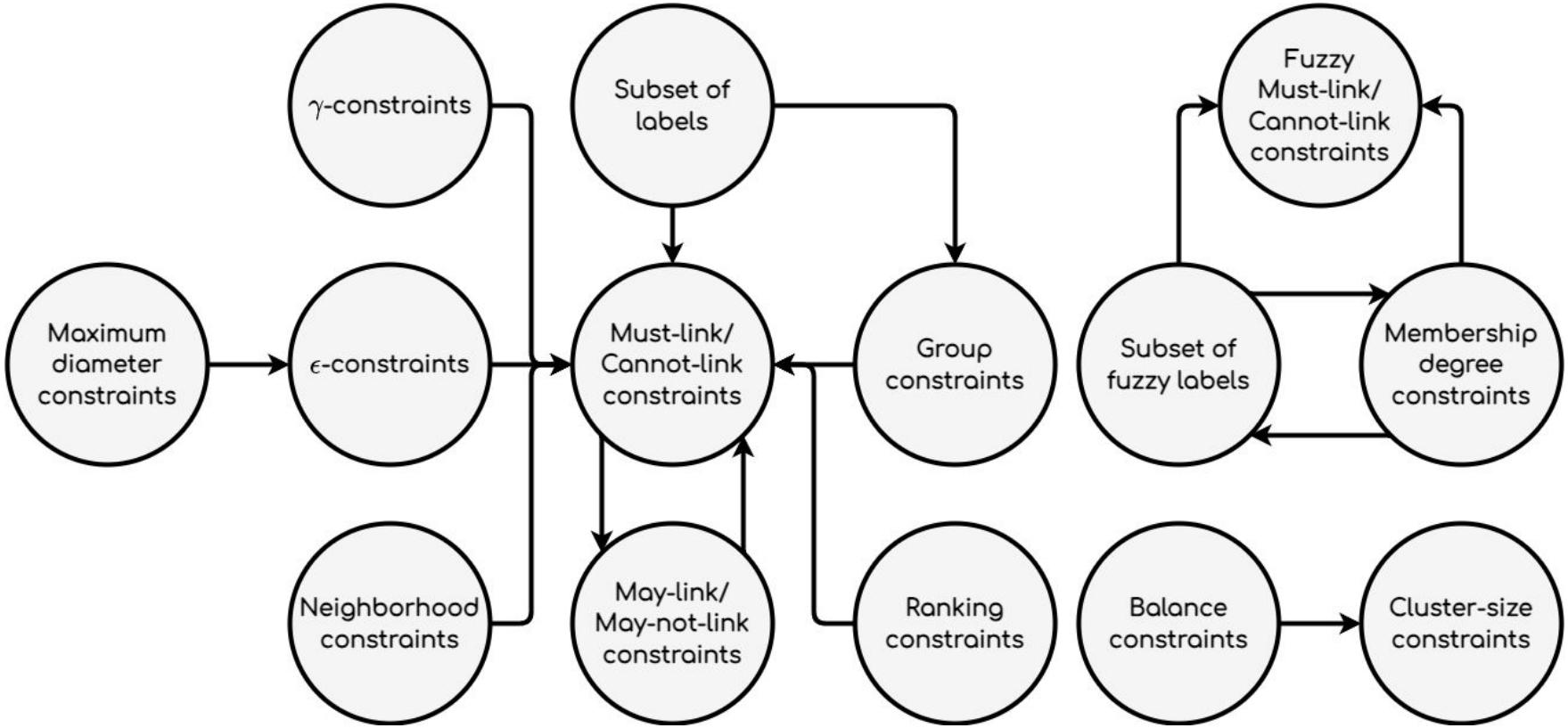


Feature-Level: constraint instances by their feature values or directly relate pairs of feature to each other to indicate degrees of importance.

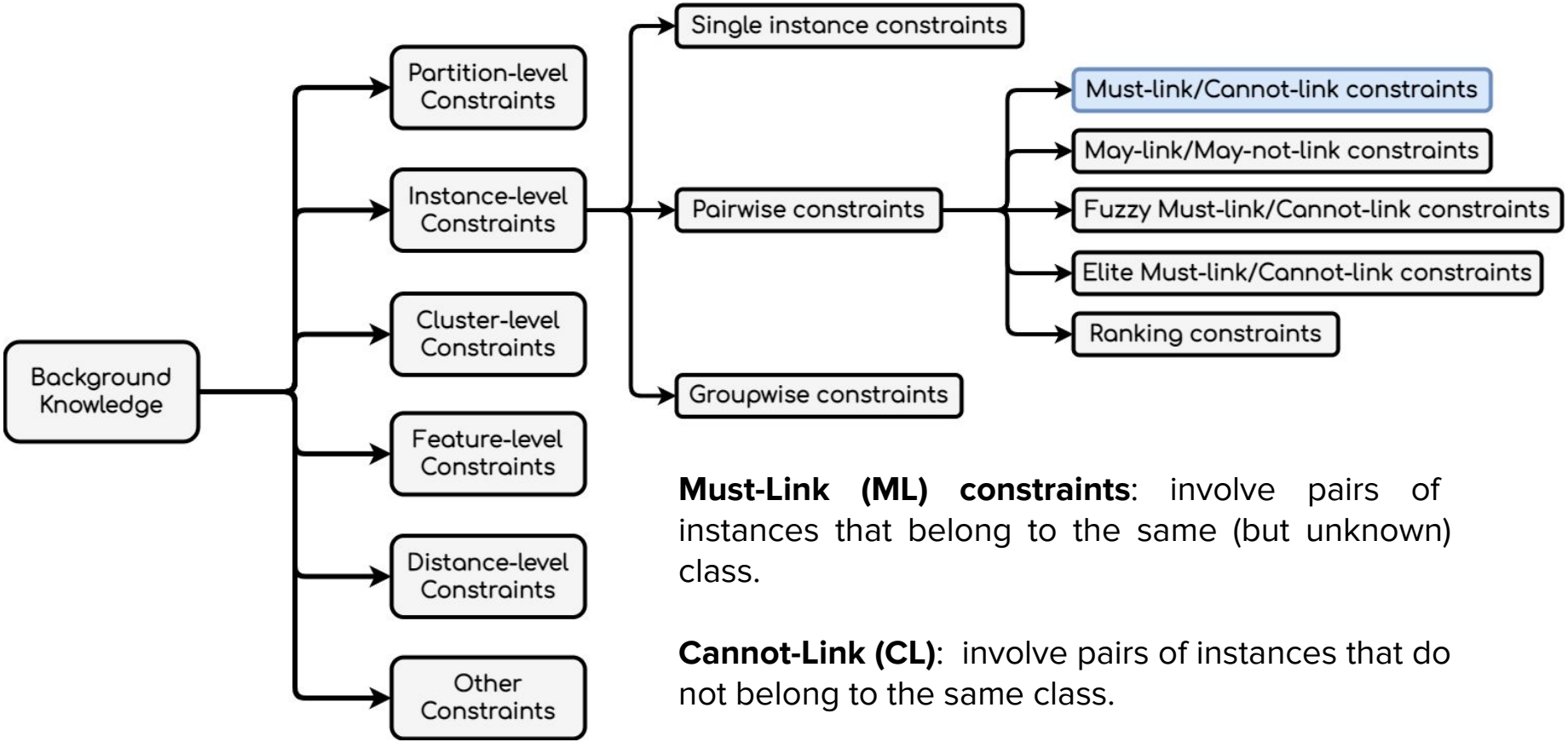
Distance-level: as they relate pair of instances indirectly and in a global way.

Semi-Supervised Clustering

Equivalencies



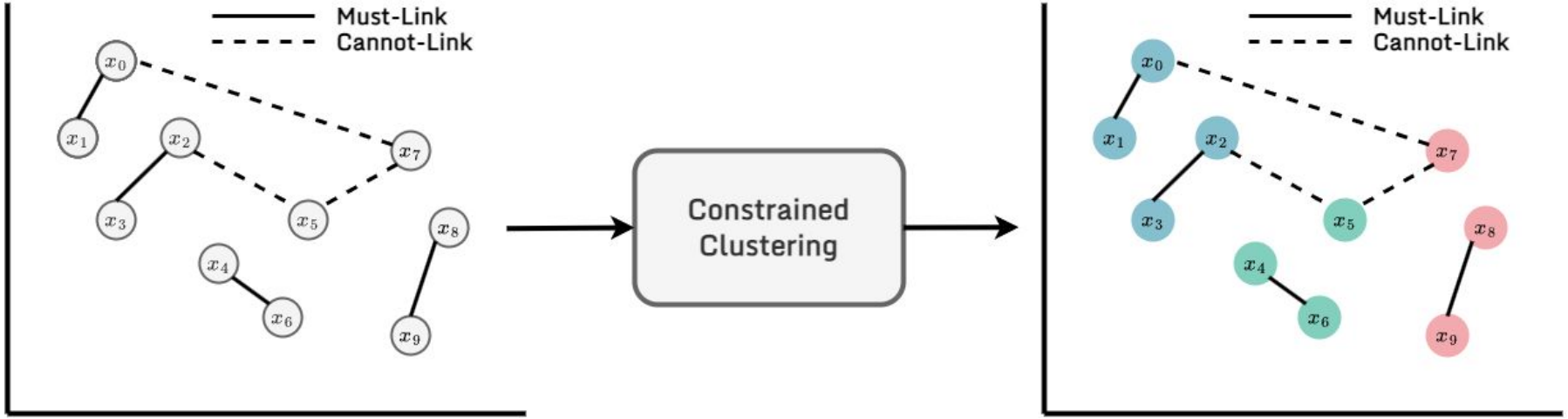
Must-link & Cannot-link



Must-Link (ML) constraints: involve pairs of instances that belong to the same (but unknown) class.

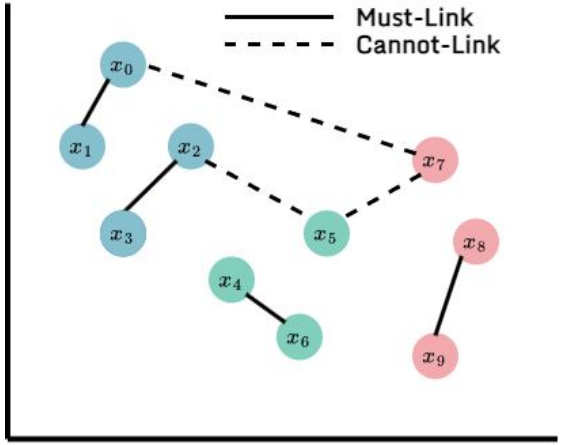
Cannot-Link (CL): involve pairs of instances that do not belong to the same class.

The goal of constrained clustering is to **find a partition of the dataset that ideally meets all constraints** in the union of both constraint sets.

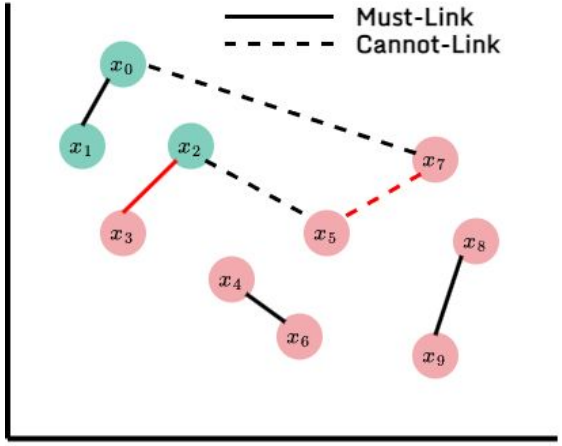


The Infeasibility

The infeasibility refers to **the number of constraints broken** by a given partition.



$\mathtt{Infs}(C, CS) = 0$

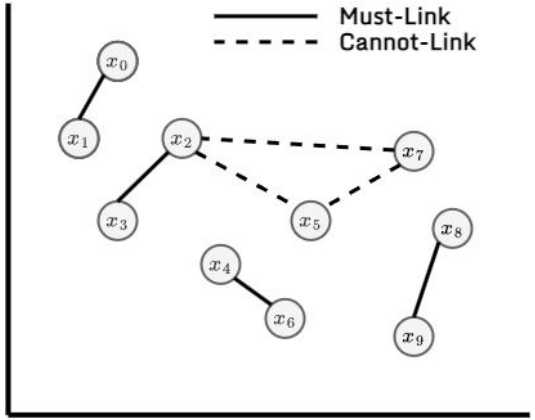


$\mathtt{Infs}(C, CS) = 2$

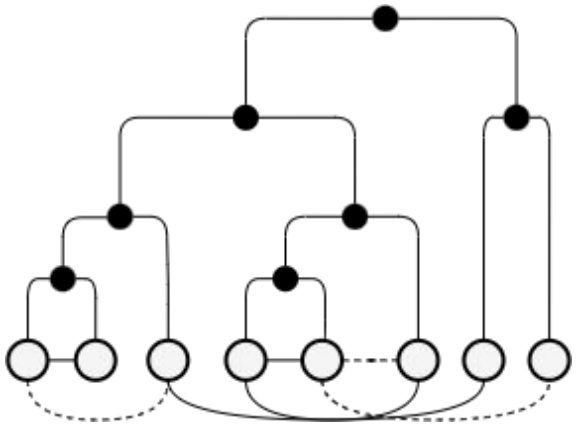
$$\mathrm{Infs}(C, CS) = \sum_{C_=(x_i,x_j)\in CS} \mathbb{1}[\![l_i^C \neq l_j^C]\!] + \sum_{C_\neq(x_i,x_j)\in CS} \mathbb{1}[\![l_i^C = l_j^C]\!]$$

How do constraints affect the complexity of clustering? Intuitively, the clustering problem goes from its classic formulation “find the best partition for a given dataset” to its constrained form **“find the best partition for a given dataset satisfying all constraints in the constraint set”**.

In **partitional CC** the number of clusters is bounded and fixed.



In **hierarchical CC** the full dendrogram is available.



The Complexity of the Feasibility Problem

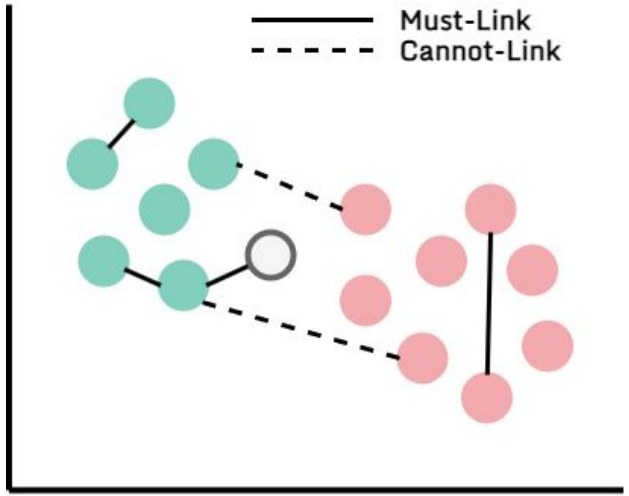
- **Partitional CC is NP-Complete** – It can be reduced from the Graph K-Colorability problem.
- **Hierarchical CC is NP-Complete** – It can be reduced from the One-in-three 3SAT with positive literals problem.

Constraints	Partitional CC	Hierarchical CC	Dead Ends?
ML	**P**	**P**	No
CL	**NP**-complete	**NP**-complete	Yes
ML and CL	**NP**-complete	**NP**-complete	Yes

Knowing that a feasible solution exists does not help us find it.

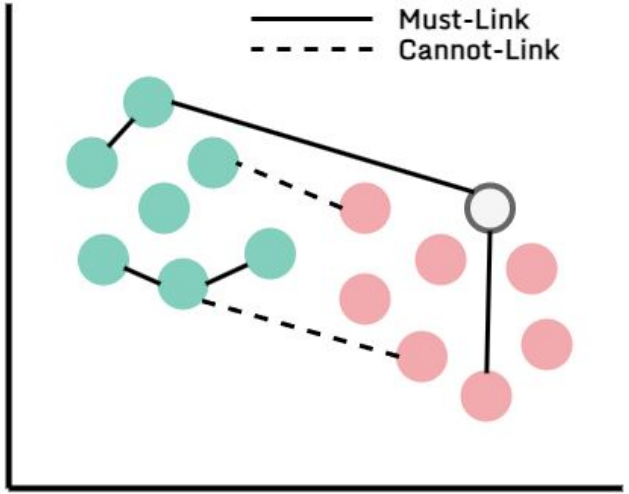
Hard Constraints VS. Soft Constraints

Hard Constraints



¿$\mathrm{Infs}(C, CS) = 0$?

Soft Constraints



¿$\mathrm{Infs}(C, CS) = 0$? ¿$\mathrm{Infs}(C, CS) = 1$?

Soft constraints are **resilient to noise**, and allow for **flexibility** in the cost/objective function and in the optimization procedure.

Semi-Supervised Learning

Semi-Supervised Clustering

Constrained Clustering

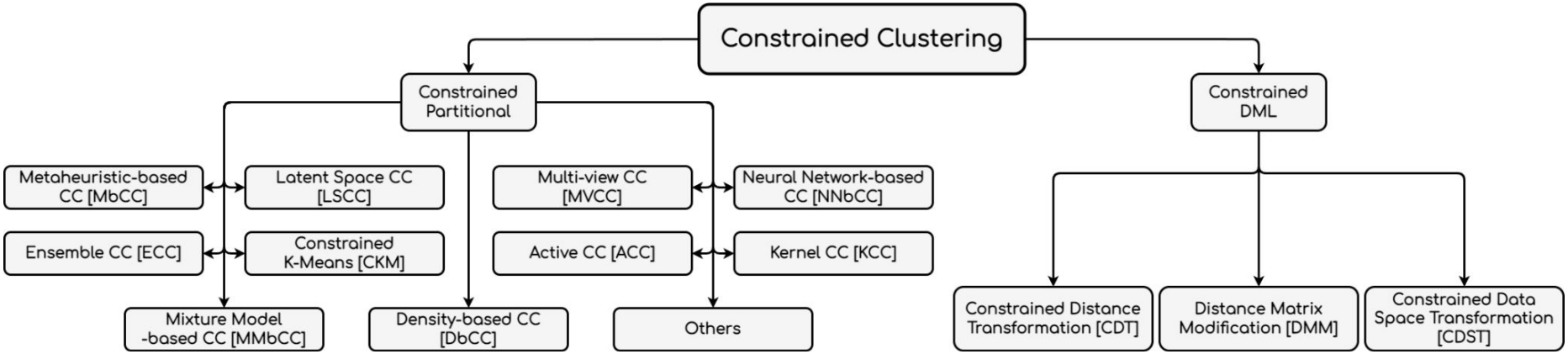
Hybrid Models for CC

New Paradigms in CC

Challenges and Frontiers

Constrained Clustering

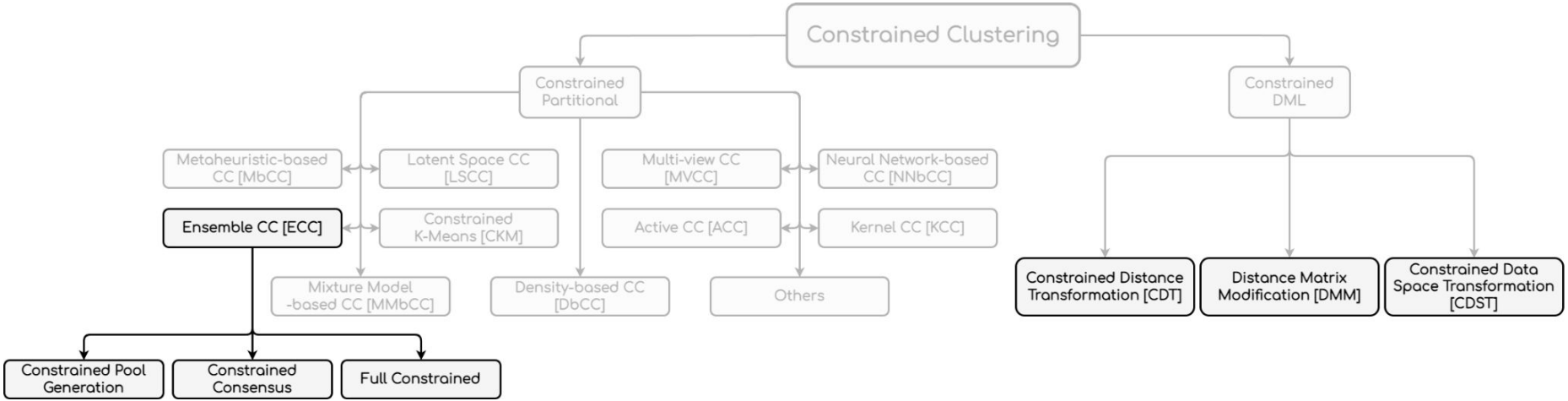
Taxonomic Tree



- **2 Major Categories:** Constrained Partitional & Constrained DML.
- **17 Subcategories:** Constrained K-Means, Ensemble CC, Active CC, etc.
- **29 Final Categories.**
- **307 methods reviewed.**

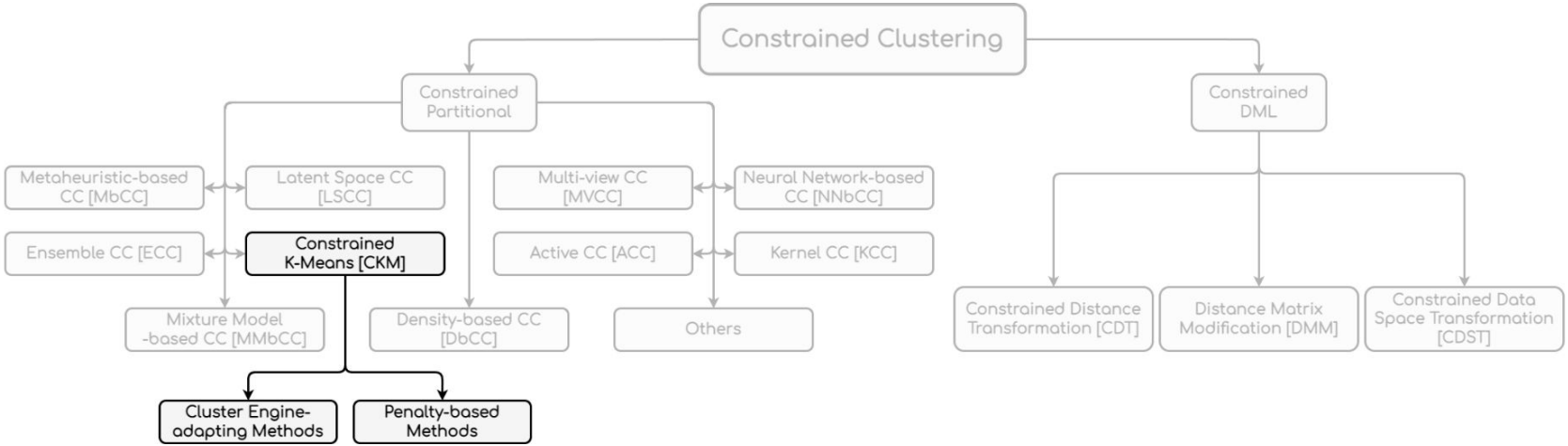
Constrained Clustering

Taxonomic Tree



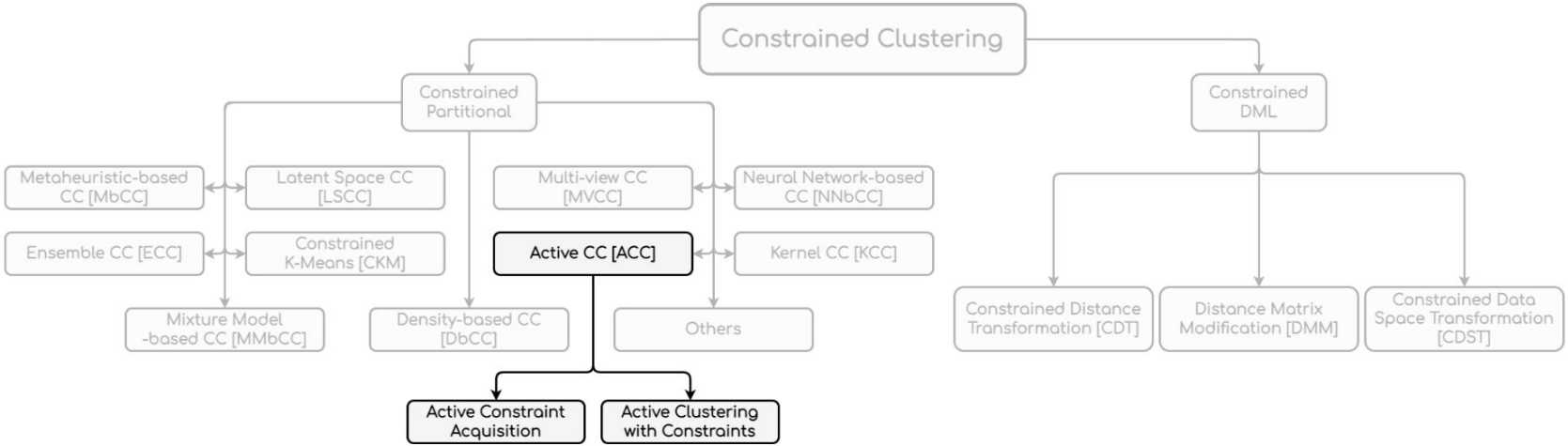
Some categories are exclusive to CC and are proposed for the first time in this taxonomization.

Taxonomic Tree



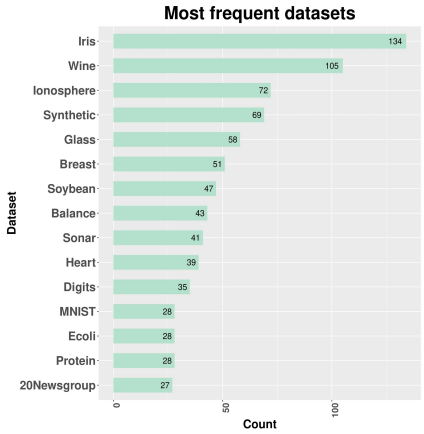
Other categories are exclusive to CC and existed before this taxonomization.

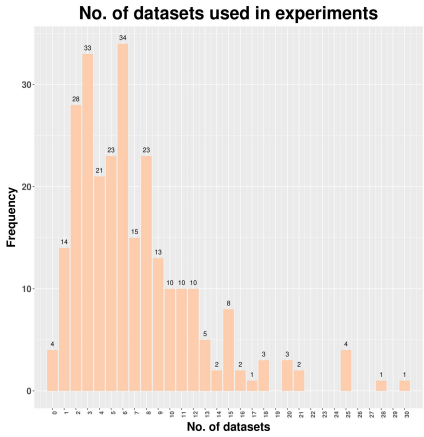
Taxonomic Tree

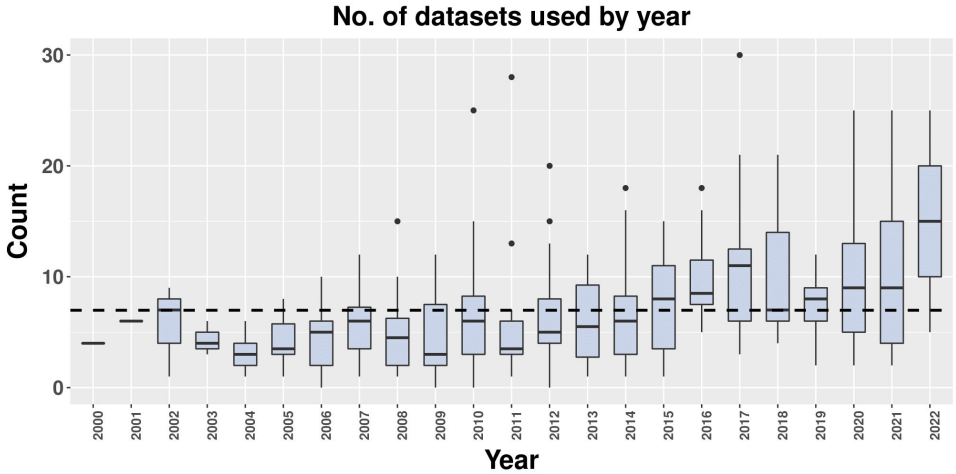


Other categories are exclusive to CC and existed before this taxonomization.

Experimental elements – Datasets



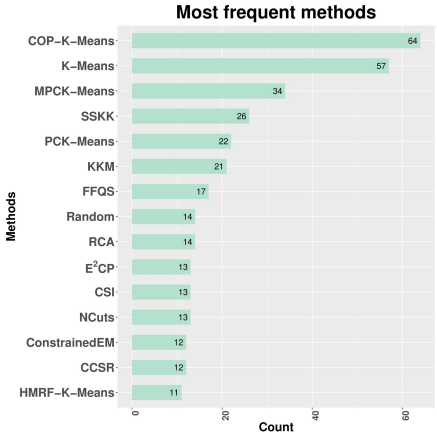


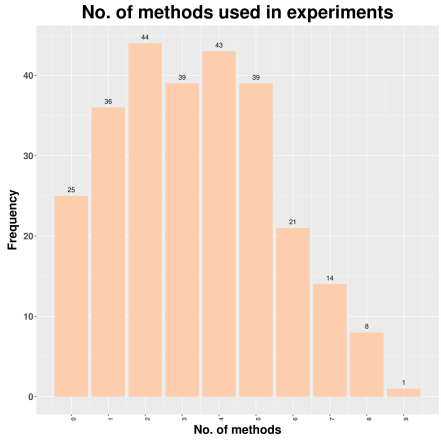


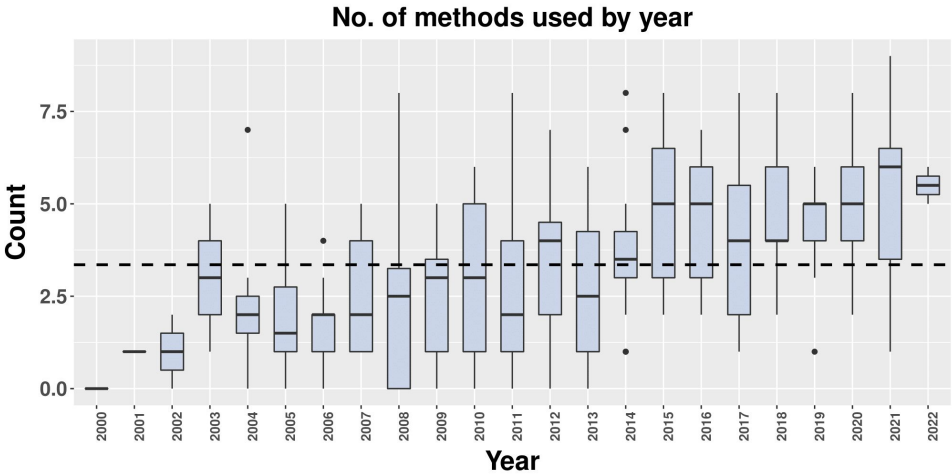
- Classification datasets are used as benchmarks, using the **labels as an oracle** to generate the constraint sets.
- Most studies use between 1 and 10 datasets in their experiments.
- The number of datasets used in experiments shows an increasing tendency.

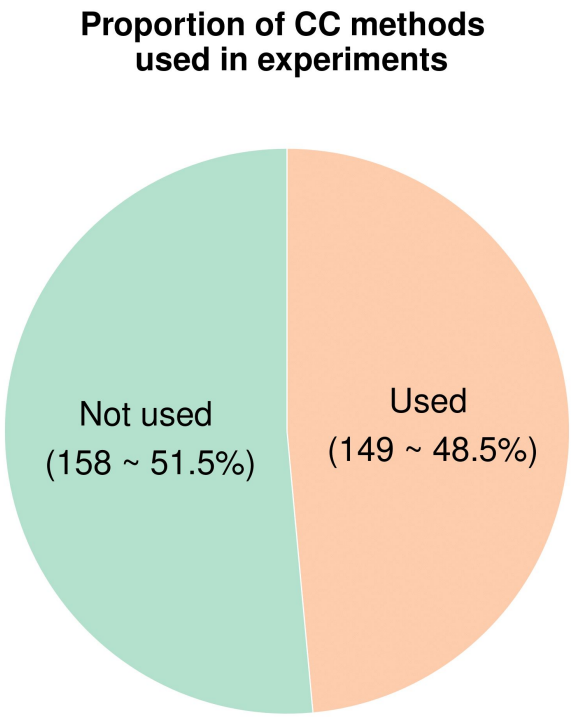
Constrained Clustering

Experimental elements – Methods



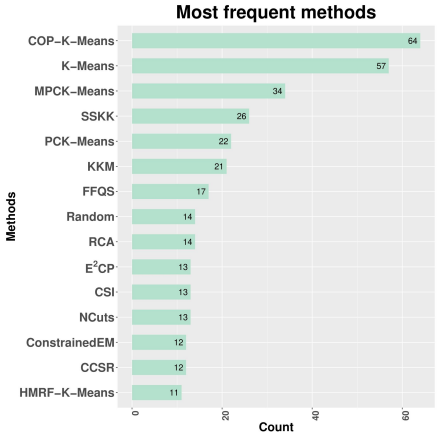


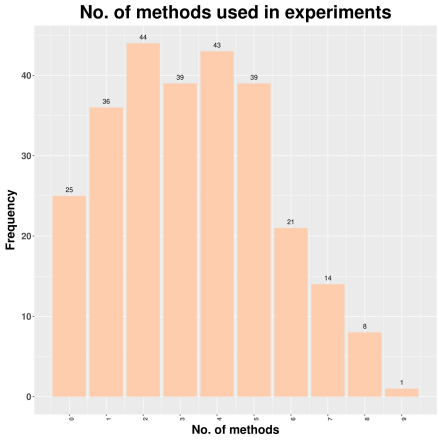


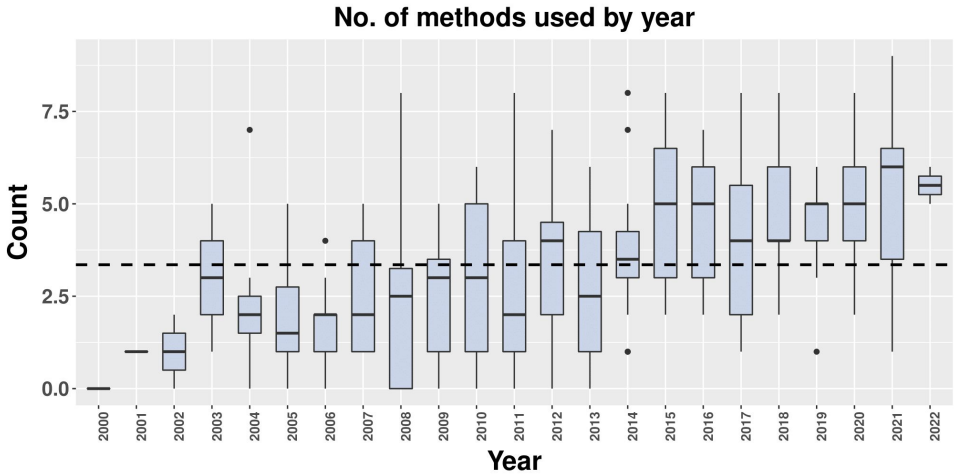


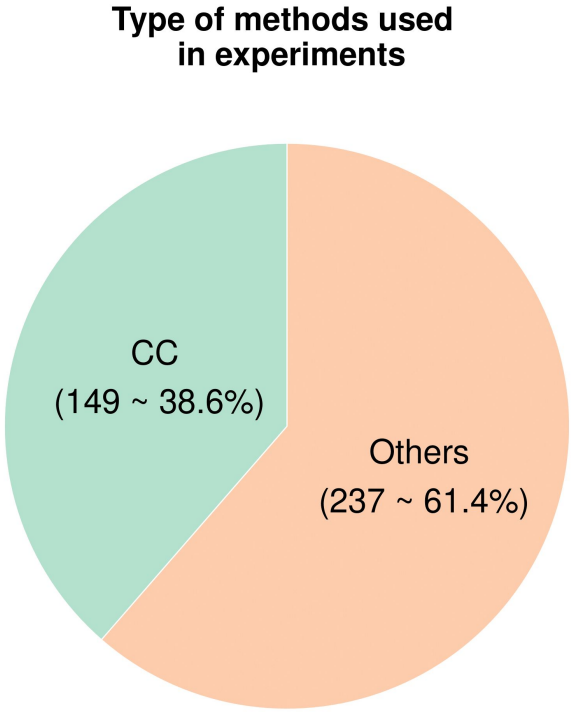
Constrained Clustering

Experimental elements – Methods



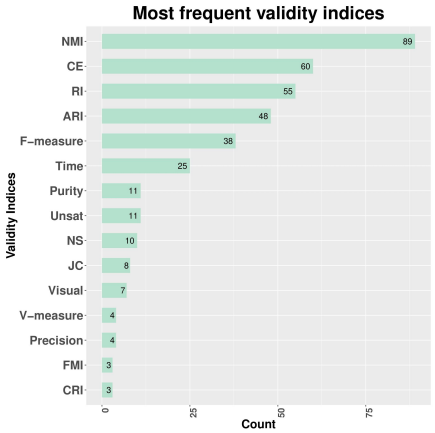


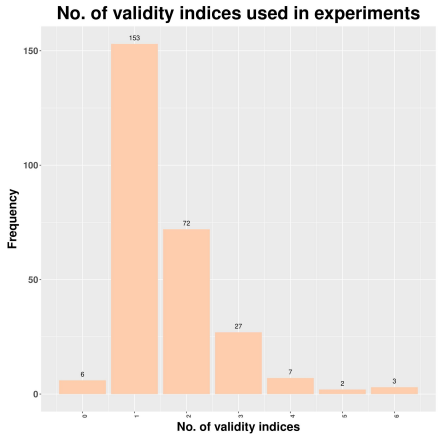


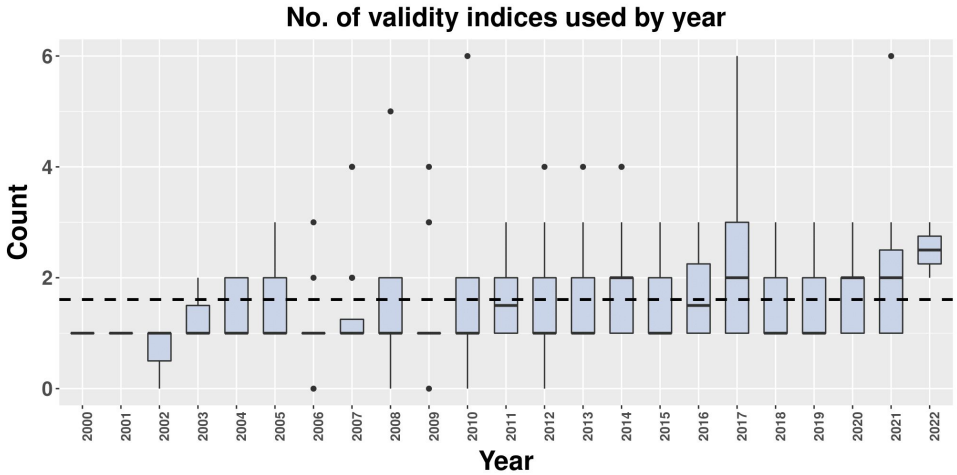


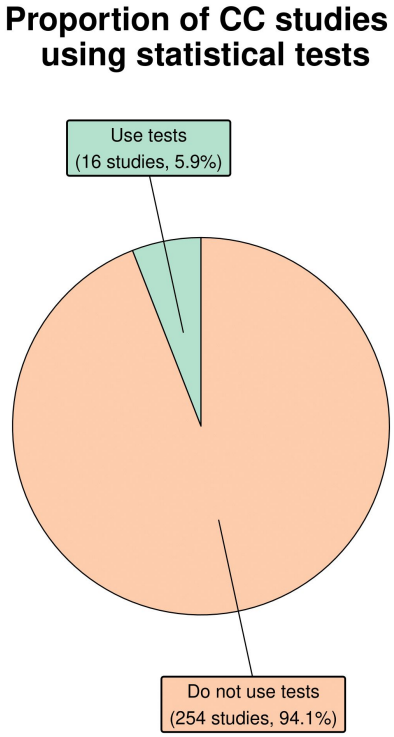
Constrained Clustering

Experimental elements – Validity Indices



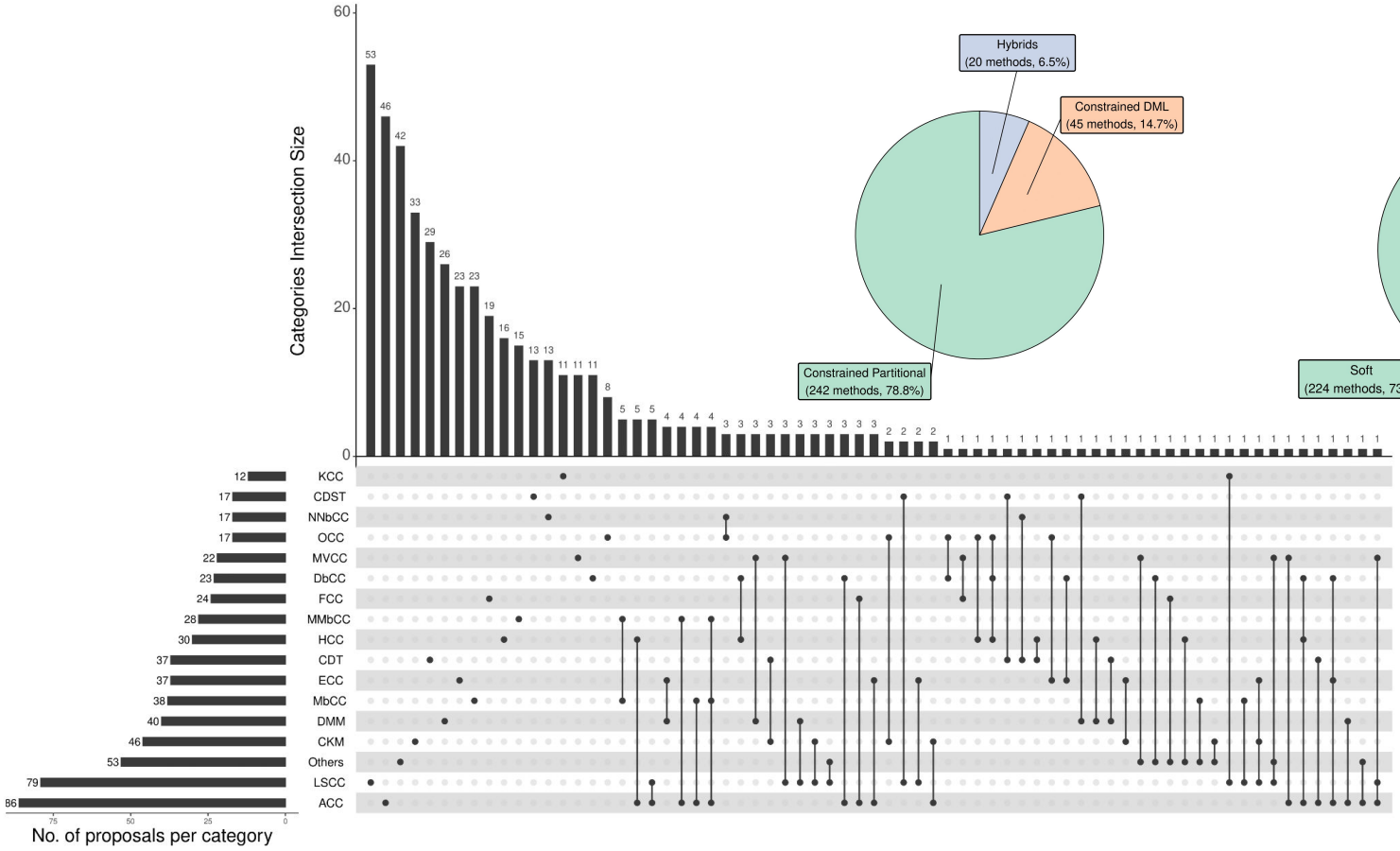






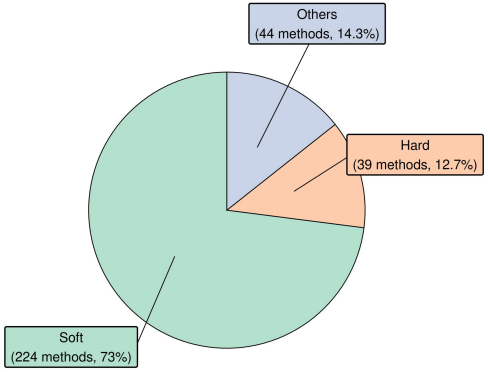
Constrained Clustering

Summary



Proportion of partitional and DML methods in CC studies

Proportion of the types of constraints used in CC methods



Semi-Supervised Learning

Semi-Supervised Clustering

Constrained Clustering

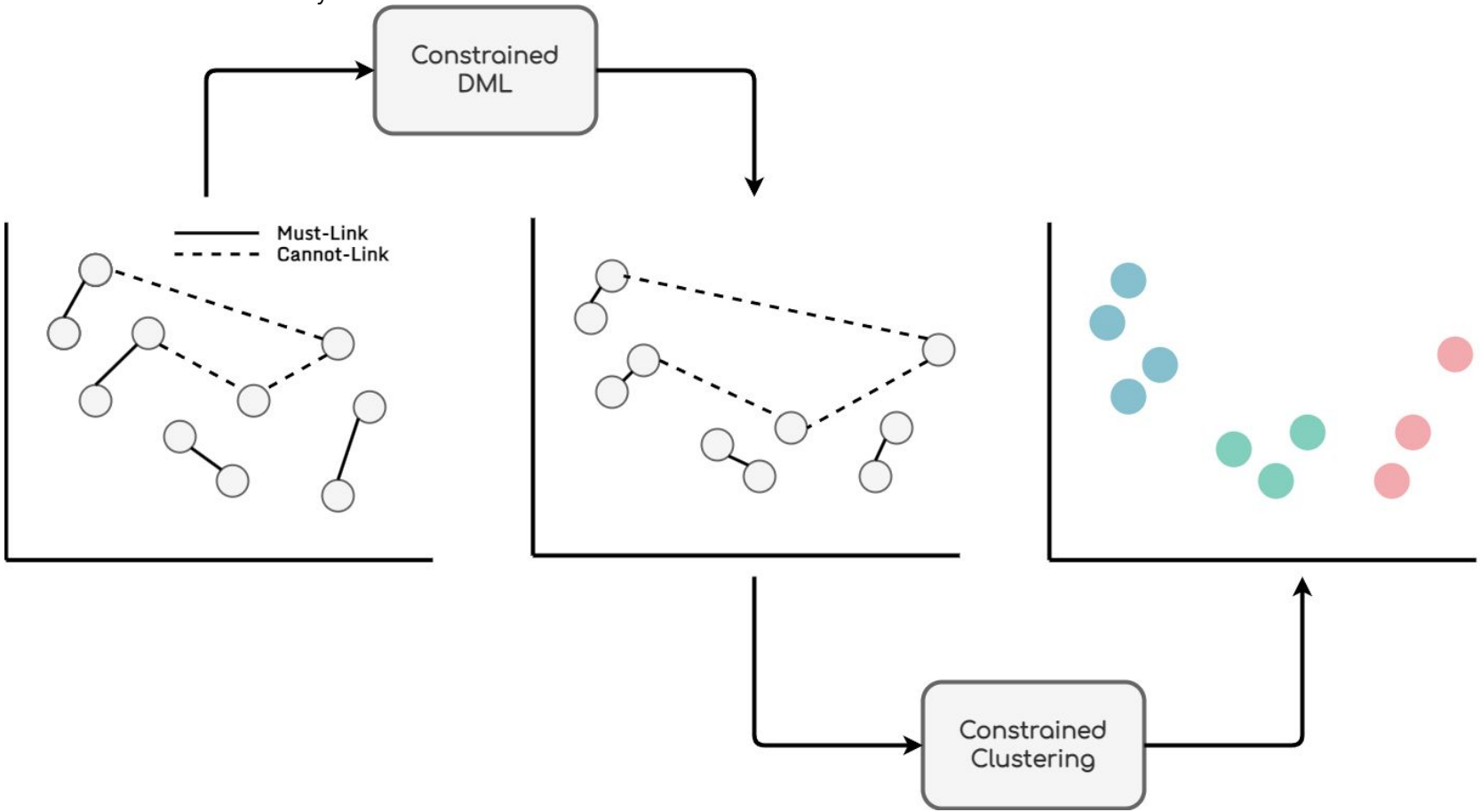
Hybrid Models for CC

New Paradigms in CC

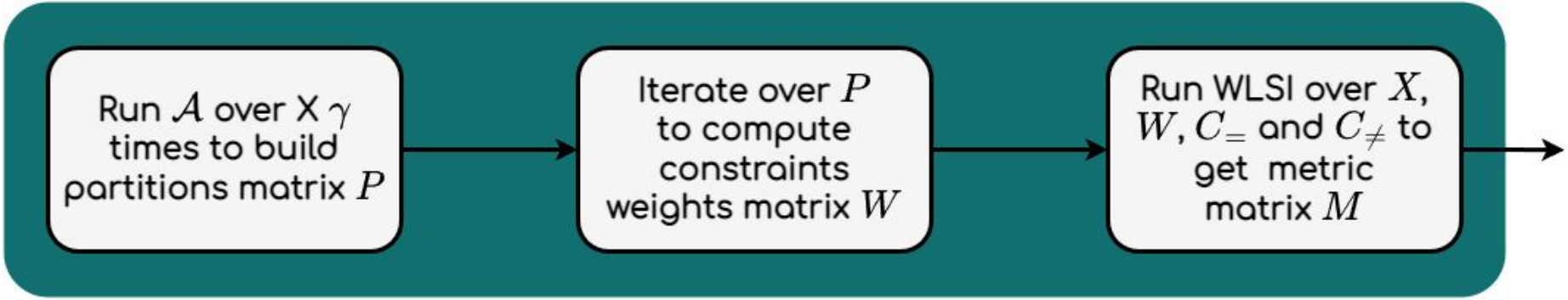
Challenges and Frontiers

Hybrid Models for CC

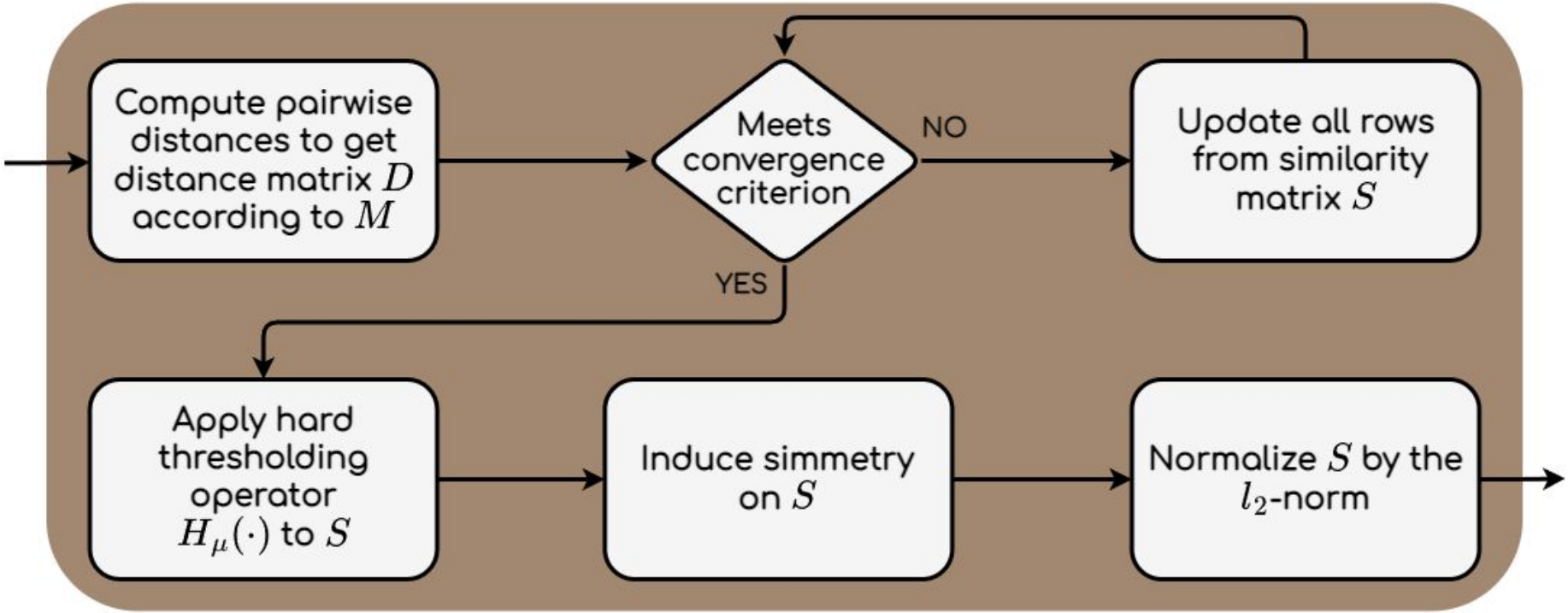
Constraints Redundancy



3SHACC – First Stage

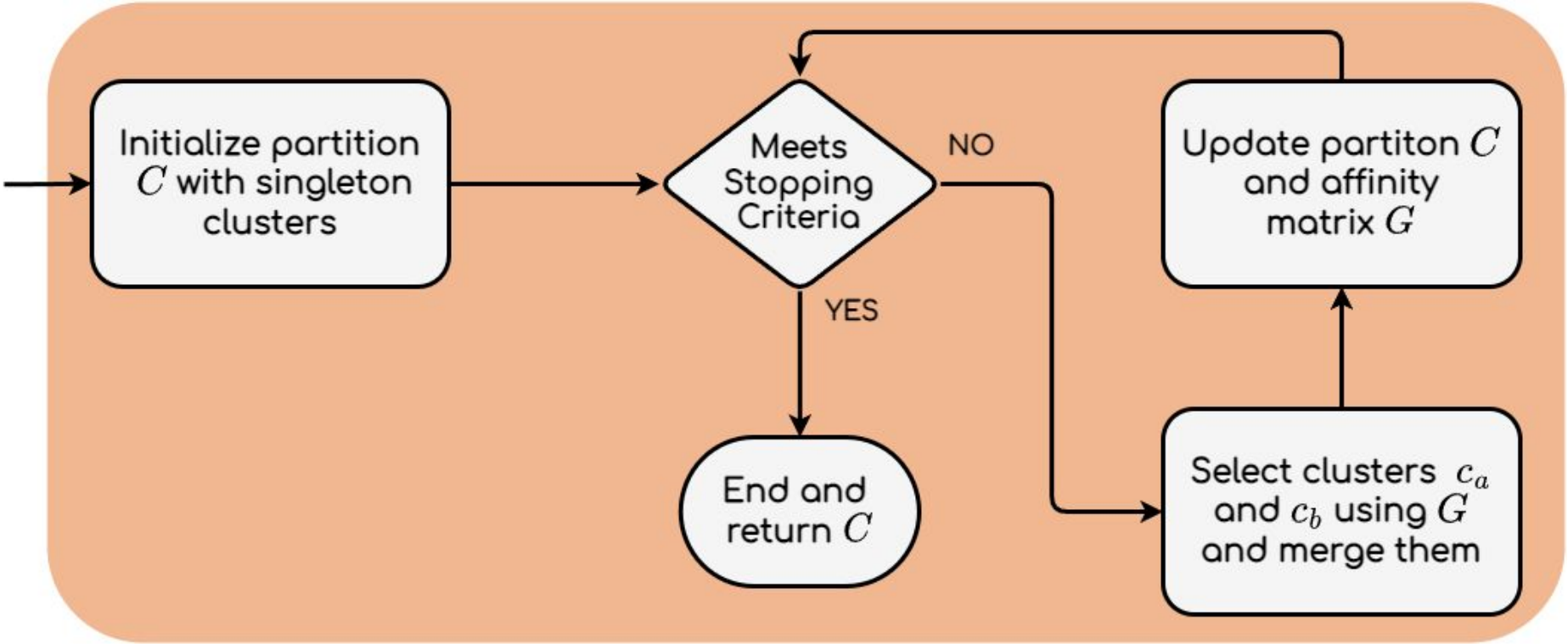


3SHACC – Second Stage



The similarity computation is based on the reconstruction coefficient, which computes similarities in terms of how much an instance is explained by others.

3SHACC – Third Stage



Semi-Supervised Learning

Semi-Supervised Clustering

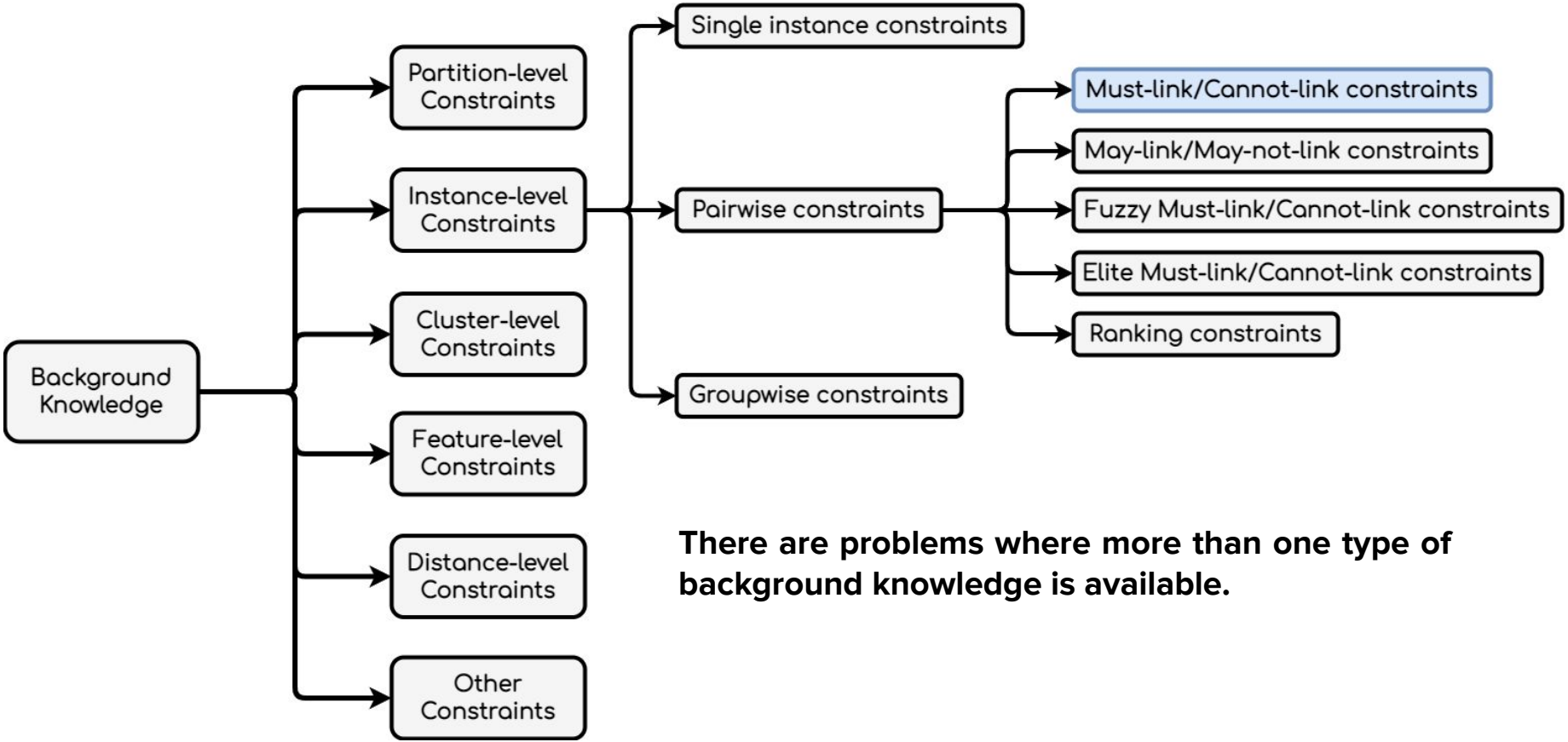
Constrained Clustering

Hybrid Models for CC

New Paradigms in CC

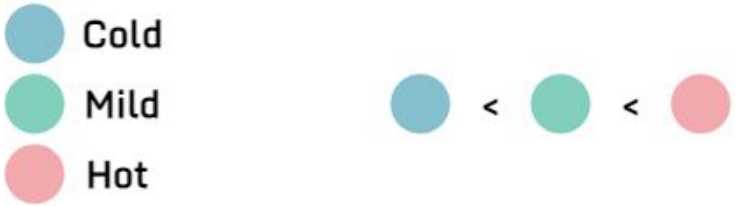
Challenges and Frontiers

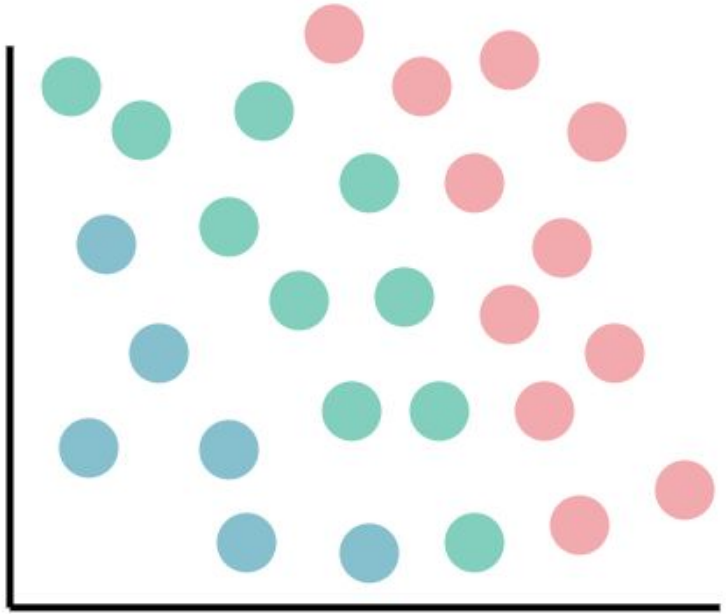
Types of Background Knowledge



There are problems where more than one type of background knowledge is available.

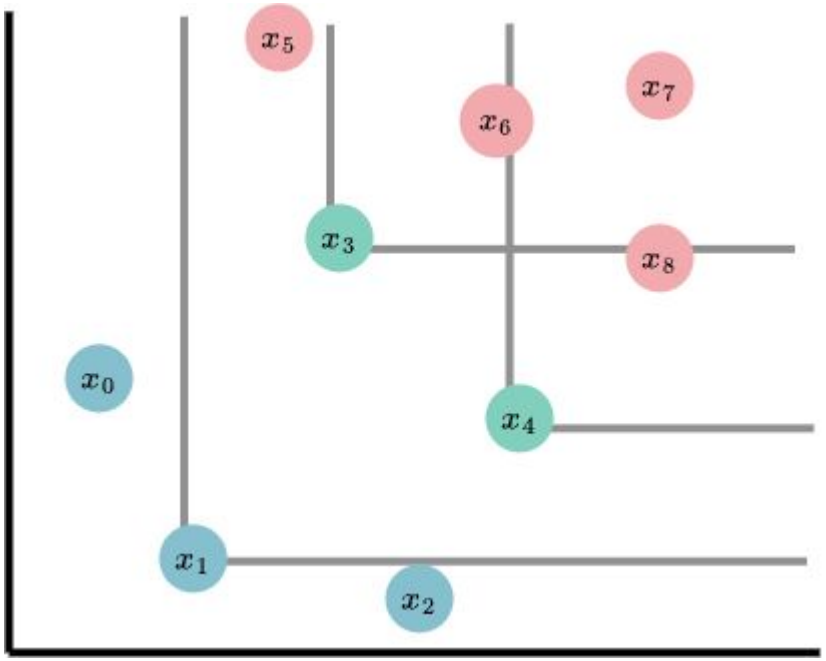
Monotonicity in Classification





- There is an order relationship between classes.
- Our goal is to minimize the number of misclassifications regarding the class order.
- The costs of misclassifications are different for every class.

MCDA introduces the concept of preference. The **preference** quantifies the addition of differences between the features of two instances, **limited to the features in which one of them is strictly better than the other**.



$$r(x_i, x_j) = \sum_{d : x_{[i,d]} > x_{[j,d]}}^{u} w_d x_{[i,d]} - w_d x_{[j,d]}.$$

$$L_1(x_i, x_j) = r(x_i, x_j) + r(x_j, x_i).$$

PCKM-Mono – Objective Function

An instances is assigned to a cluster if **its centroid is the most similar to it in terms of preference**, taking the number of violated constraints into account.

$$J_{PCKMM} = \frac{1}{K}\sum_{k=1}^{K}\sum_{x_i \in c_k} \boxed{|(r(x_i,\mu_k) - r(\mu_k,x_i)|} + \sum_{(x_i,x_j)\in C_=} \mathbb{1}[\![l_i \neq l_j]\!] + \sum_{(x_i,x_j)\in C_\neq} \mathbb{1}[\![l_i = l_j]\!]$$

PCKM-Mono – Objective Function

An instances is assigned to a cluster if **its centroid is the most similar to it in terms of preference**, taking the number of violated constraints into account.

$$J_{PCKMM} = \frac{1}{K}\sum_{k=1}^{K}\sum_{x_i \in c_k} |(r(x_i,\mu_k) - r(\mu_k, x_i)| + \boxed{\sum_{(x_i,x_j)\in C_=} \mathbb{1}[\![l_i \neq l_j]\!] + \sum_{(x_i,x_j)\in C_\neq} \mathbb{1}[\![l_i = l_j]\!]}$$

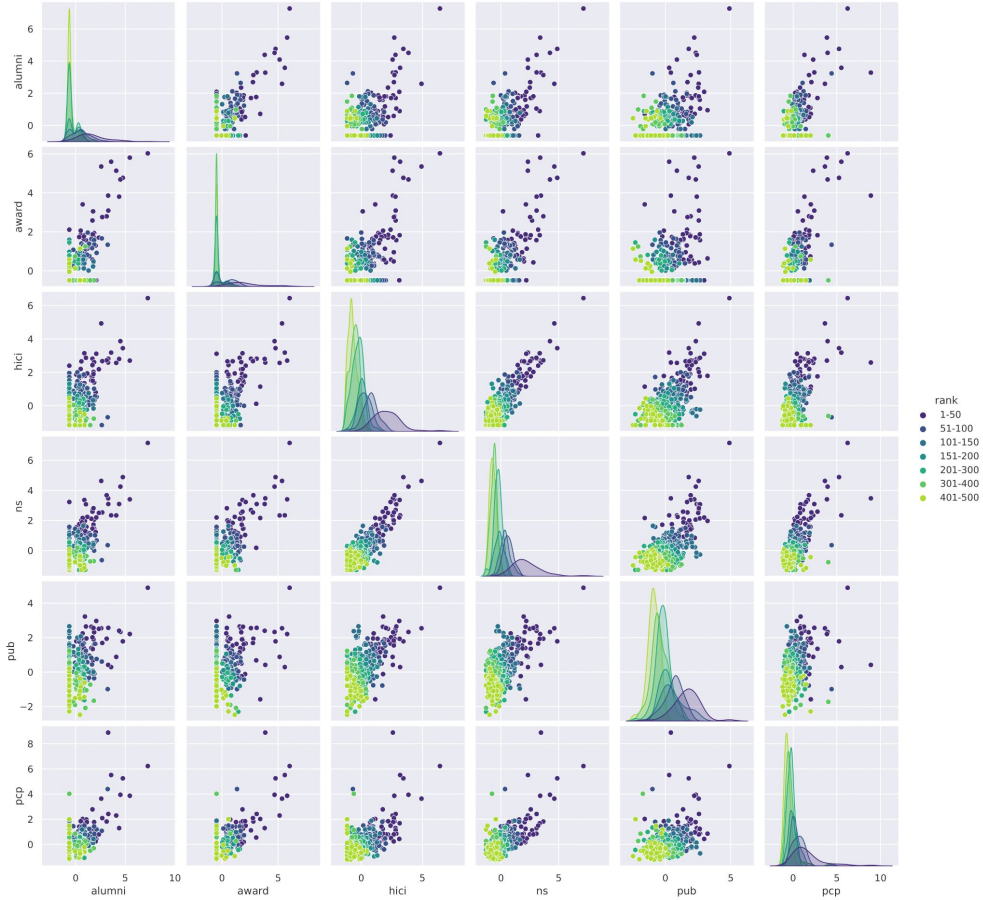
PCKM-Mono – Expectation-Minimization

An instances is assigned to a cluster if its centroid is the most similar to it in terms of preference, taking the number of violated constraints into account.

$$\mathtt{argmin}_h \Big(\big| \textstyle\sum_{j=1}^{u} (x_{[i,j]} - \mu_{[h,j]}) \big| + \\ \textstyle\sum_{x_j : (x_i, x_j) \in C_=} \mathbb{1}[\![l(c_h) \neq l_j]\!] + \sum_{x_j : (x_i, x_j) \in C_\neq} \mathbb{1}[\![l(c_h) = l_j]\!] \Big)$$

The **centroids must be neutral** in terms of preference with respect to all instances in the cluster.

$$\mu_i = \frac{1}{|c_i|} \sum_{x_i \in c_i} x_i$$



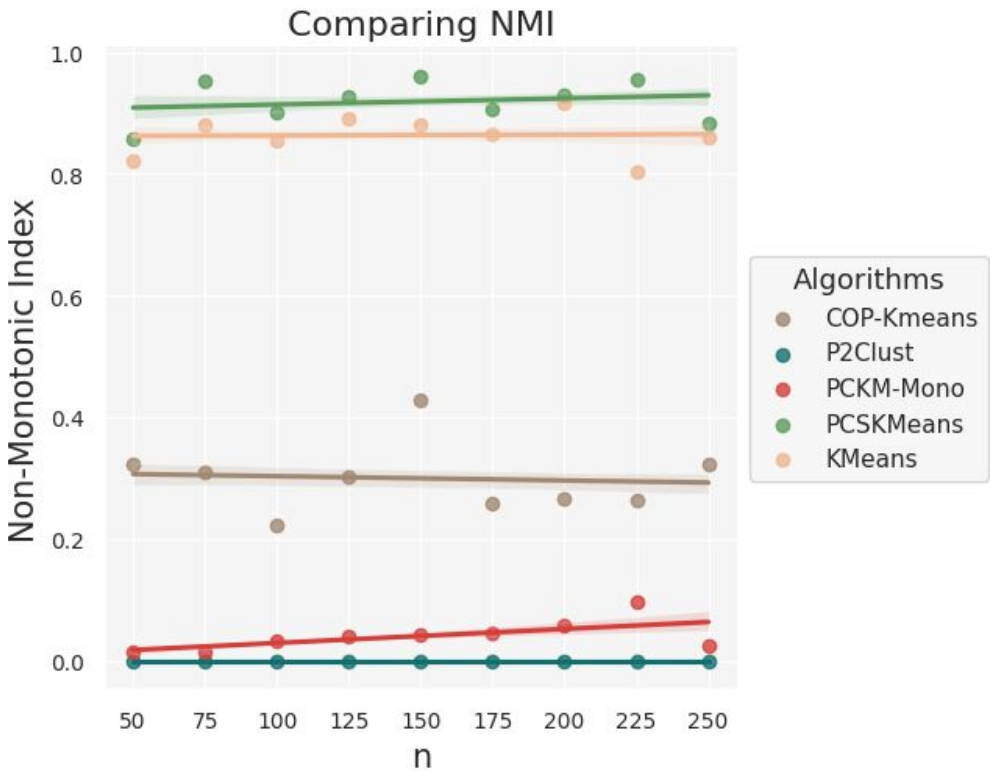
Shanghai Ranking of World Universities

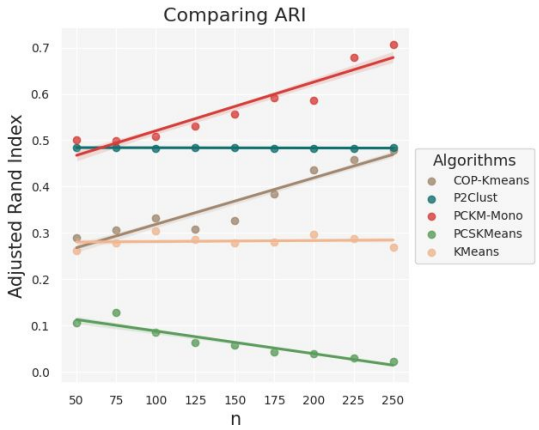
Clustering problem with both monotonicity constraints and instance-level pairwise constraints.

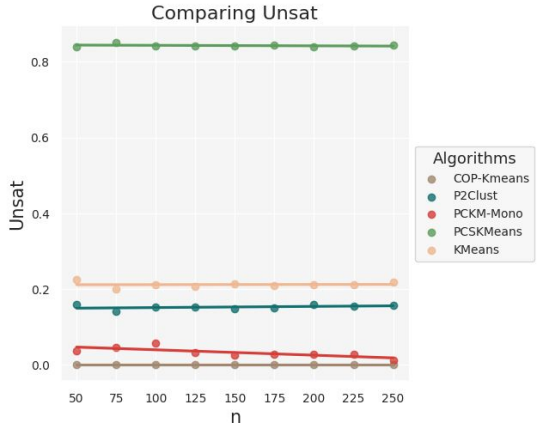
Features monotonicity faults in 7% of its instances.

New Paradigms in CC

Case of Study







Semi-Supervised Learning

Semi-Supervised Clustering

Constrained Clustering

Hybrid Models for CC

New Paradigms in CC

Challenges and Frontiers

Future Work

- **Creation of a free CC library**. The creation of an open-access library specialized in CC methods would greatly stimulate research in the area.
- **Constraint-based preprocessing**. We argue that constraints can be beneficial in preprocessing procedures.
- **Preprocessing the constraint set**. As the constraint set can be considered a dataset itself, it can suffer from the same imperfections as traditional datasets, such as: missing values, noise or redundancies.
- **New combinations of types of background knowledge**. Another potential research direction is to investigate how to automatically identify the best combination of background knowledge for a given problem, hence keeping low human effort and cost.

Thanks for your attention

Proportion of proposals by field of application