# Semi-supervised clustering: a review

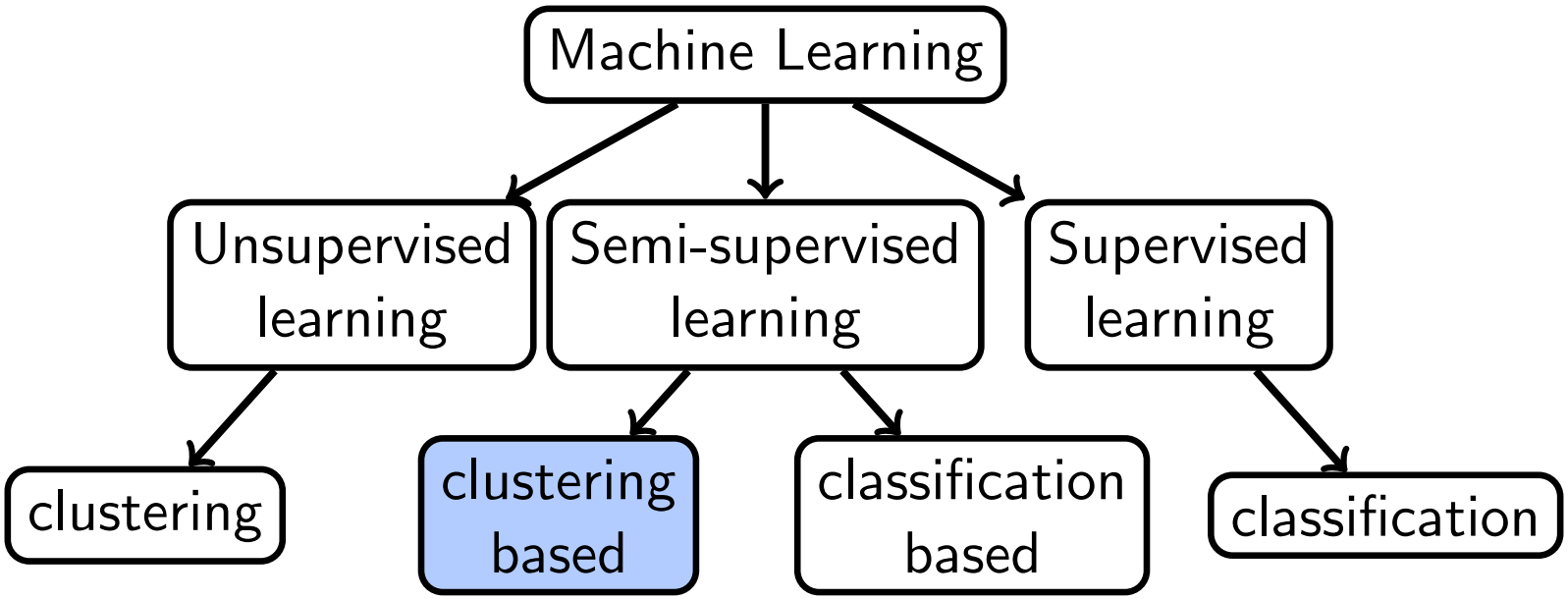
V. Antoine

Clermont-Auvergne INP, LIMOS, UMR CNRS 6158, France

January 2023

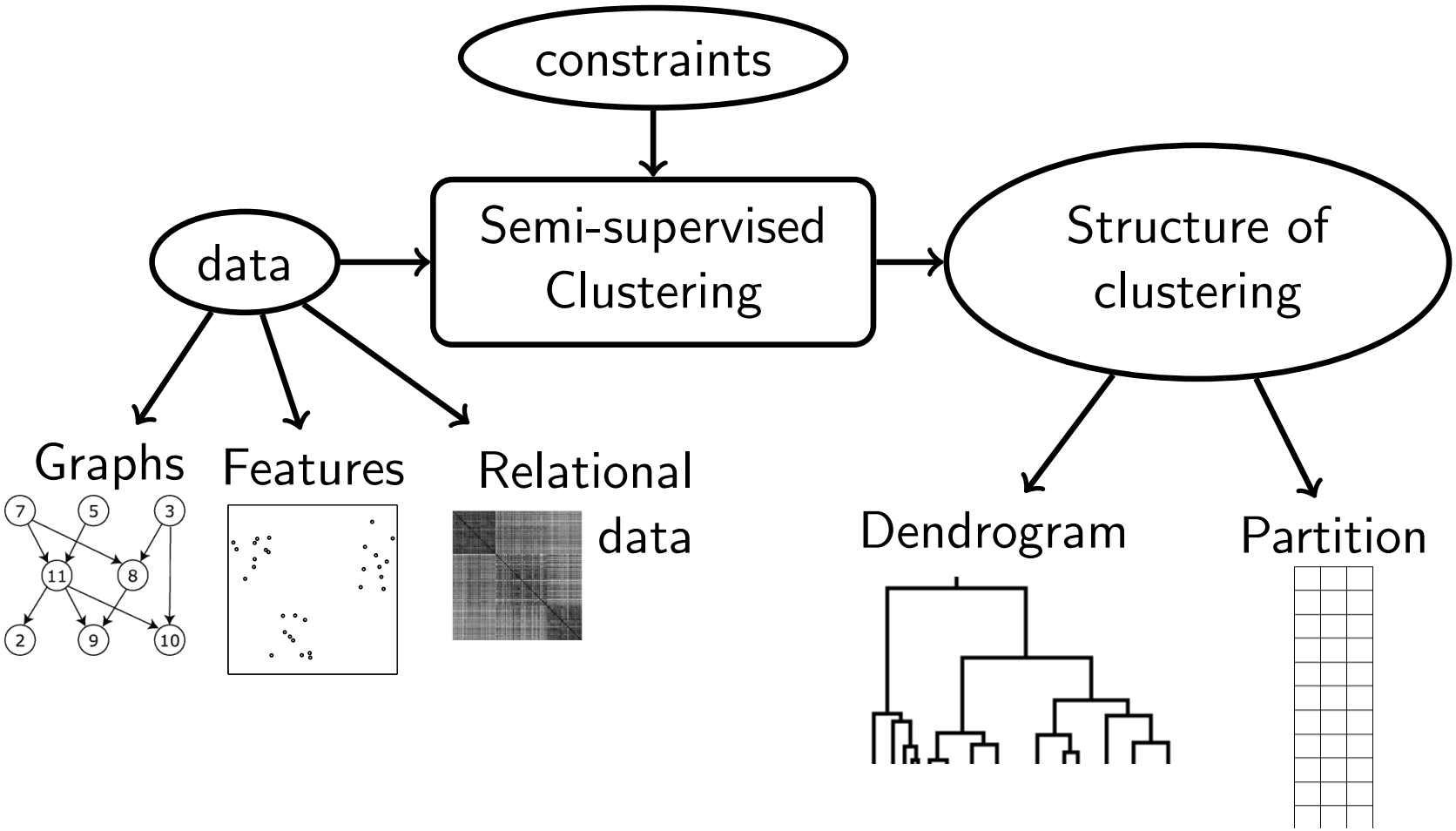


# Semi-supervised clustering : a Machine Learning technique



# Semi-supervised clustering

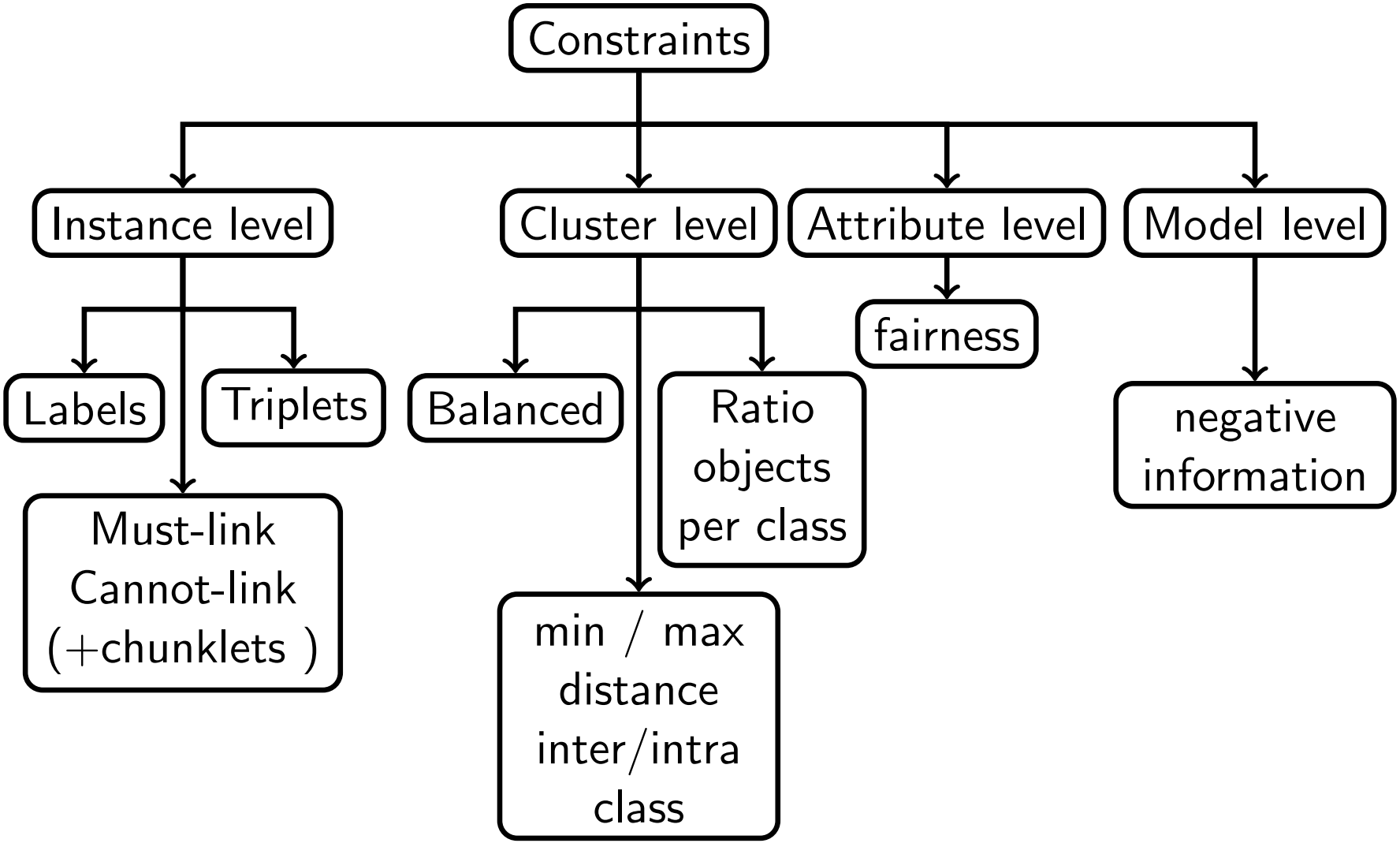
Determines groups of objects using a data set and some constraints



# Outline : semi-supervised clustering

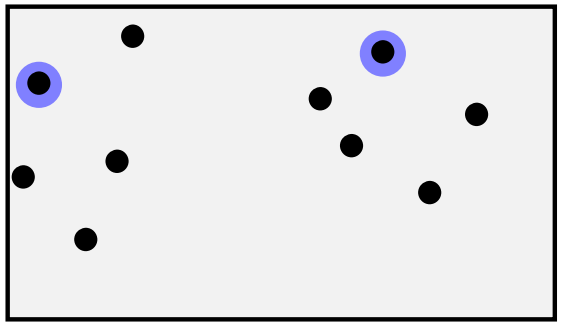
# Outline

# Constraint types



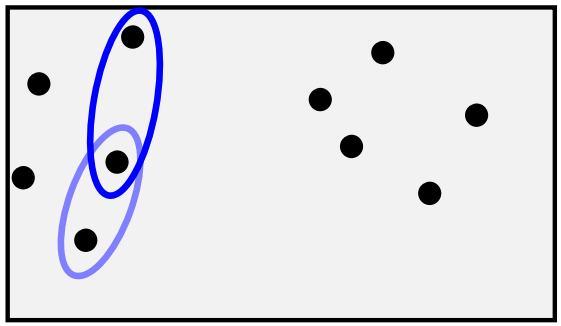
# Instance level constraints

Informativeness



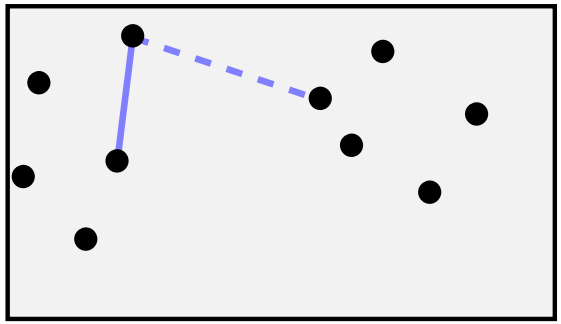
## Label

An object belong to a class: $x_i \in \mathcal{L}$



## Triplet

$x_a$ is closer to $x_b$ than to $x_c$: $d(x_a, x_b) < d(x_a, x_c)$



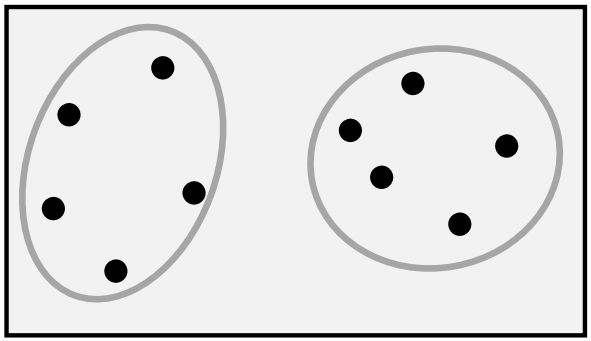
## Must-Link / Cannot-Link:

Two objects are in the same/different class:

$(x_i, x_j) \in \mathcal{M}$ / $(x_i, x_j) \in \mathcal{C}$

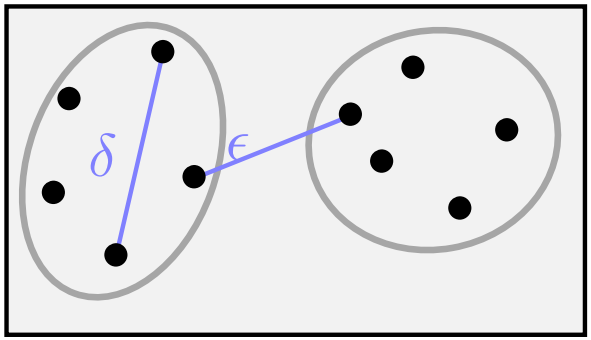
Chunklets: set of objects in the same class

# Cluster level constraints



Ratio objects per class / balanced clusters :

$\frac{n_k}{n} = \tau_k$ / $\frac{n_k}{n} = \frac{1}{c}$




Min / Max

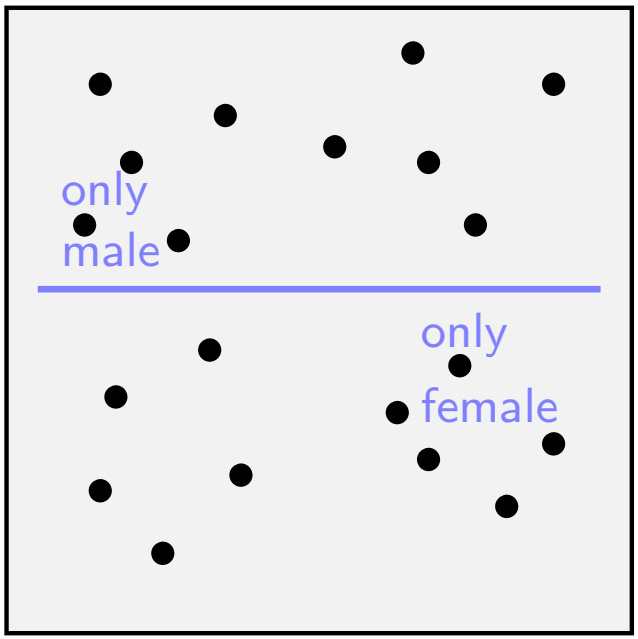
distance inter-class ($\delta$)

distance intra-class ($\epsilon$)

# Attribute level constraints

## Caution [2]

Protected attributes can not be explicitly used in decision making !



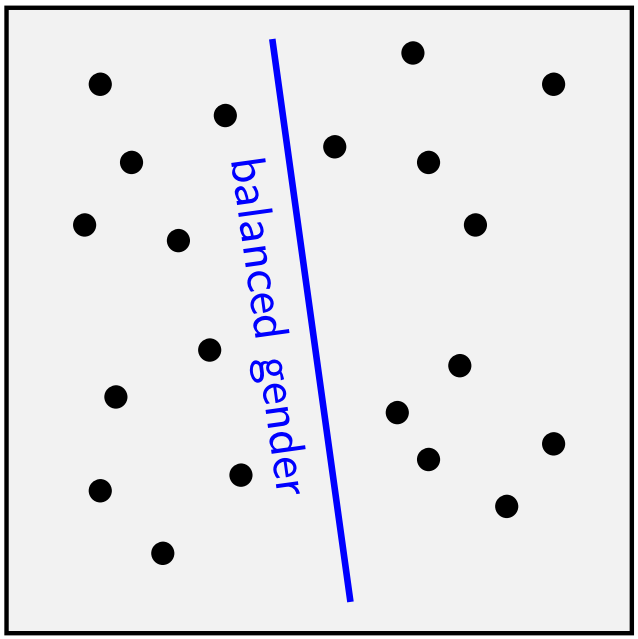

### Fairness

Percentage $p$%-rule on the modality of an attribute

# Attribute level constraints

## Caution [2]

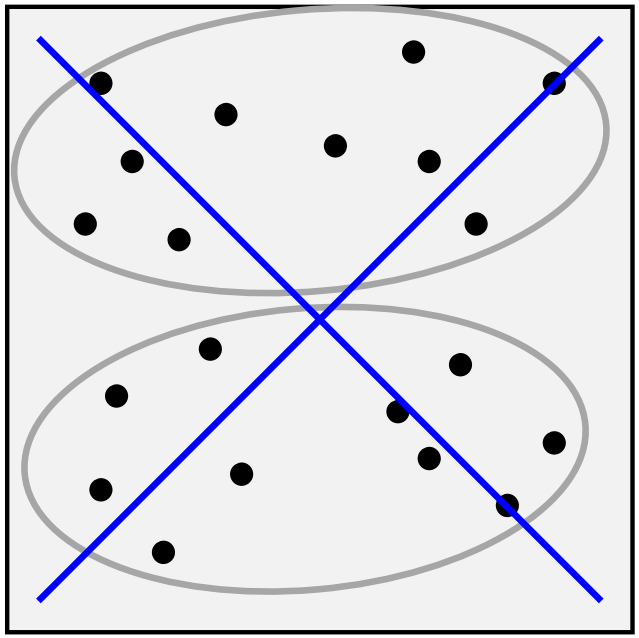
Protected attributes can not be explicitly used in decision making !




### Fairness

Percentage $p$%-rule on the modality of an attribute

# Model level constraints



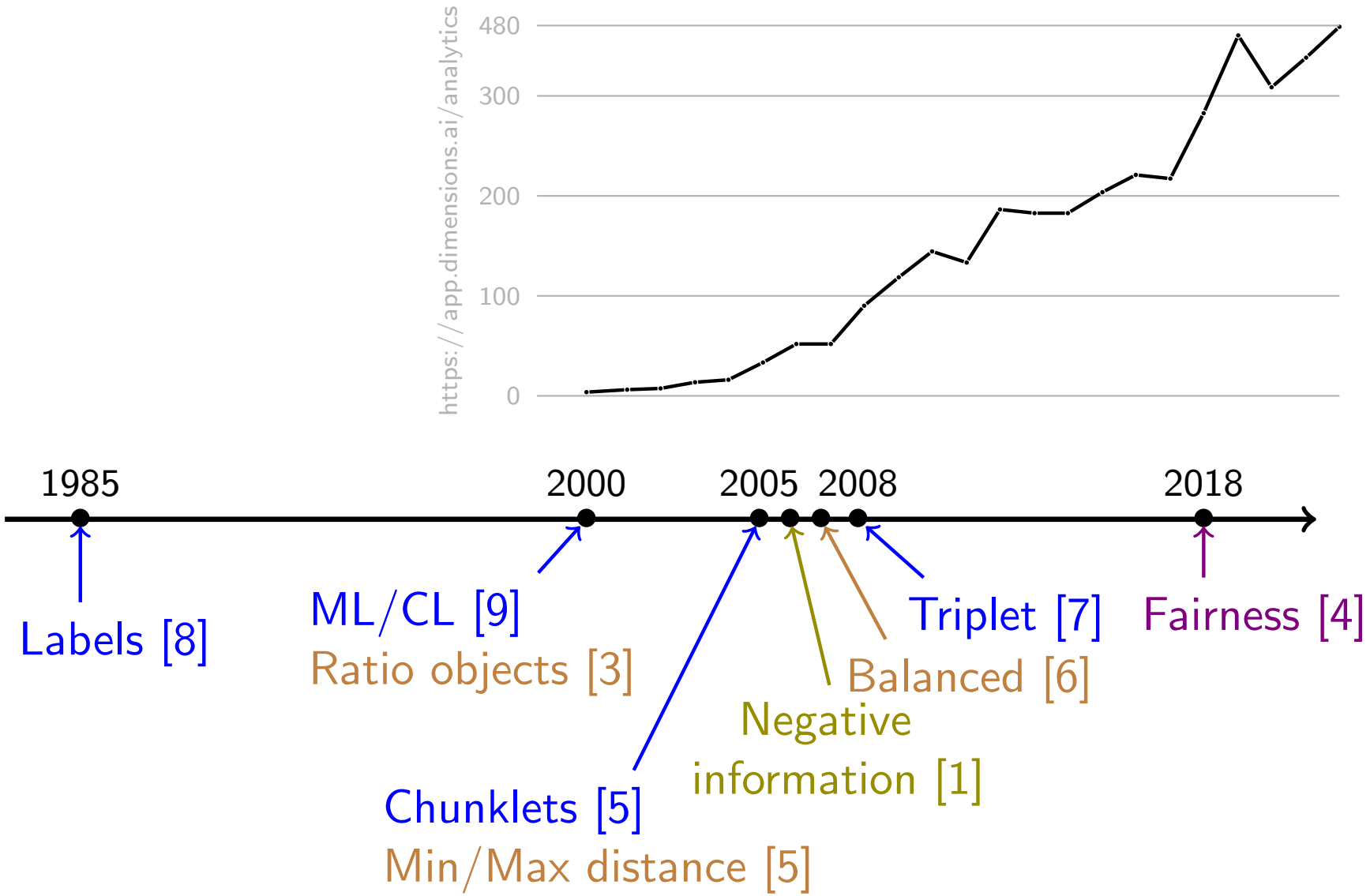
## Negative information

A specific partition should not be the final solution

**Remark**

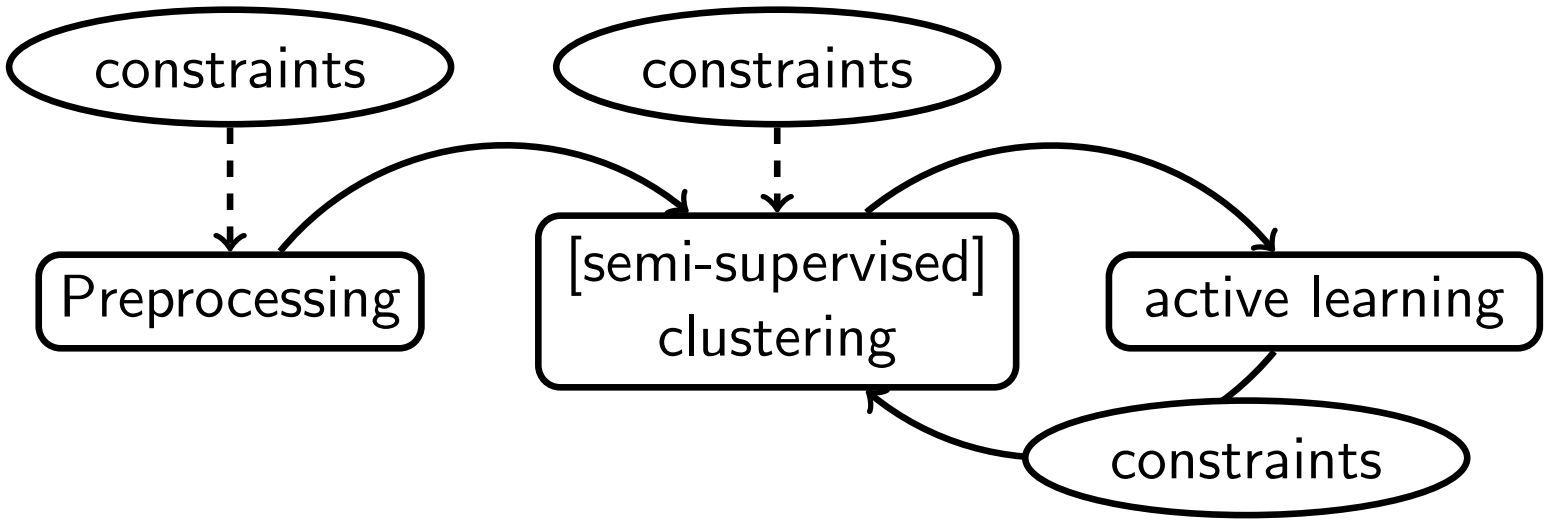
Close to singular alternative clustering problem

# Semi-supervised clustering history



# Outline

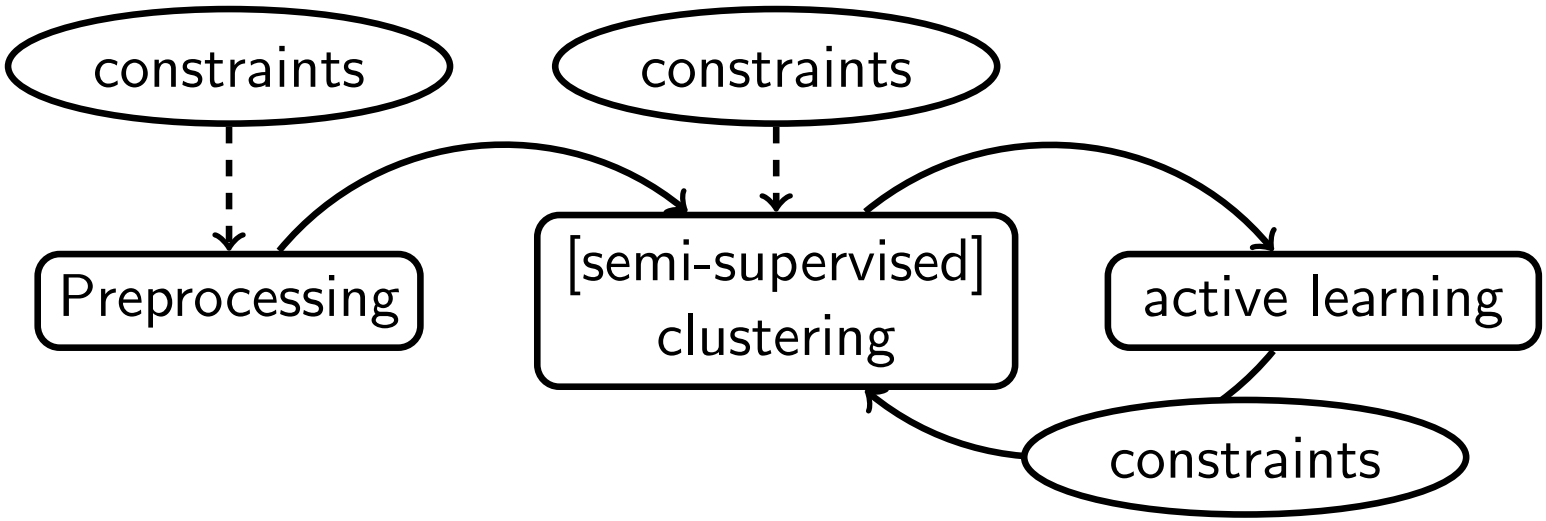
# Adding constraints in unsupervised learning




## Preprocessing

- learning step
  - centroids, distance,...
- constraints influence
  - feasability solution w.r.t constraints ?
  - augmenting constraints
  - reducing constraints to informative ones

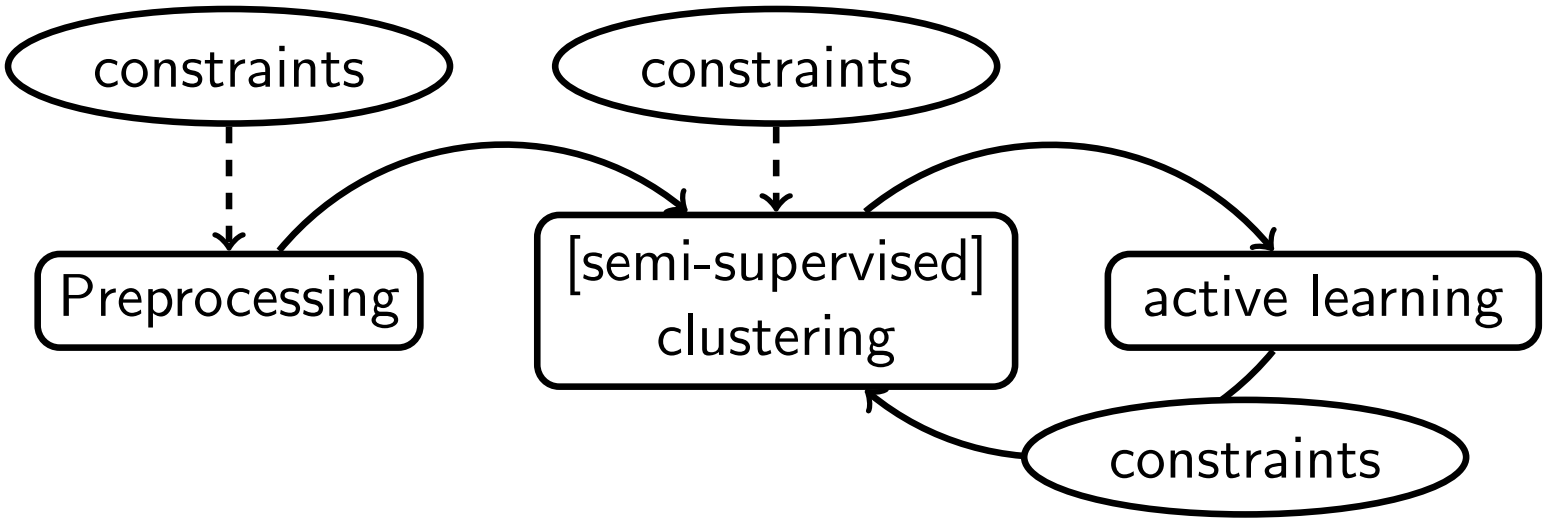
# Adding constraints in unsupervised learning




## Semi-supervised clustering

- learning step
  - partition, distance
- constraints
  - hard / soft respect of the constraints
  - noisy constraints ?

# Adding constraints in unsupervised learning




## Active learning scheme

- Originally from semi-supervised learning
- Ask expert few selected constraints
  - exploration / consolidation

# Conclusion

## Semi-supervised clustering

- ✓ adding constraints improves the partition accuracy
- ✓ several methods have been already implemented

## Perspectives

- Real application
- Use of soft constraints

# References I


E. Bae and J. Bailey.

Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity.
In *Sixth International Conference on Data Mining (ICDM'06)*, pages 53–62. IEEE, 2006.

D. Biddle.

*Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*.
Routledge, 2017.

P. S. Bradley, K. P. Bennett, and A. Demiriz.

Constrained k-means clustering.
*Microsoft Research, Redmond*, 20(0):0, 2000.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii.

Fair clustering through fairlets.
*Advances in Neural Information Processing Systems*, 30, 2017.

I. Davidson and S. Ravi.

Clustering with constraints: Feasibility issues and the k-means algorithm.
In *Proceedings of the 2005 SIAM international conference on data mining*, pages 138–149. SIAM, 2005.

R. Ge, M. Ester, W. Jin, and I. Davidson.

Constraint-driven clustering.
In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 320–329, 2007.

N. Kumar and K. Kummamuru.

Semisupervised clustering with metric learning using relative comparisons.
*IEEE Transactions on Knowledge and Data Engineering*, 20(4):496–503, 2008.

# References II

W. Pedrycz.

Algorithms of fuzzy clustering with partial supervision.
*Pattern recognition letters*, 3(1):13–20, 1985.

K. Wagstaff and C. Cardie.

Clustering with instance-level constraints.
*AAAI/IAAI*, 1097:577–584, 2000.

Thank you