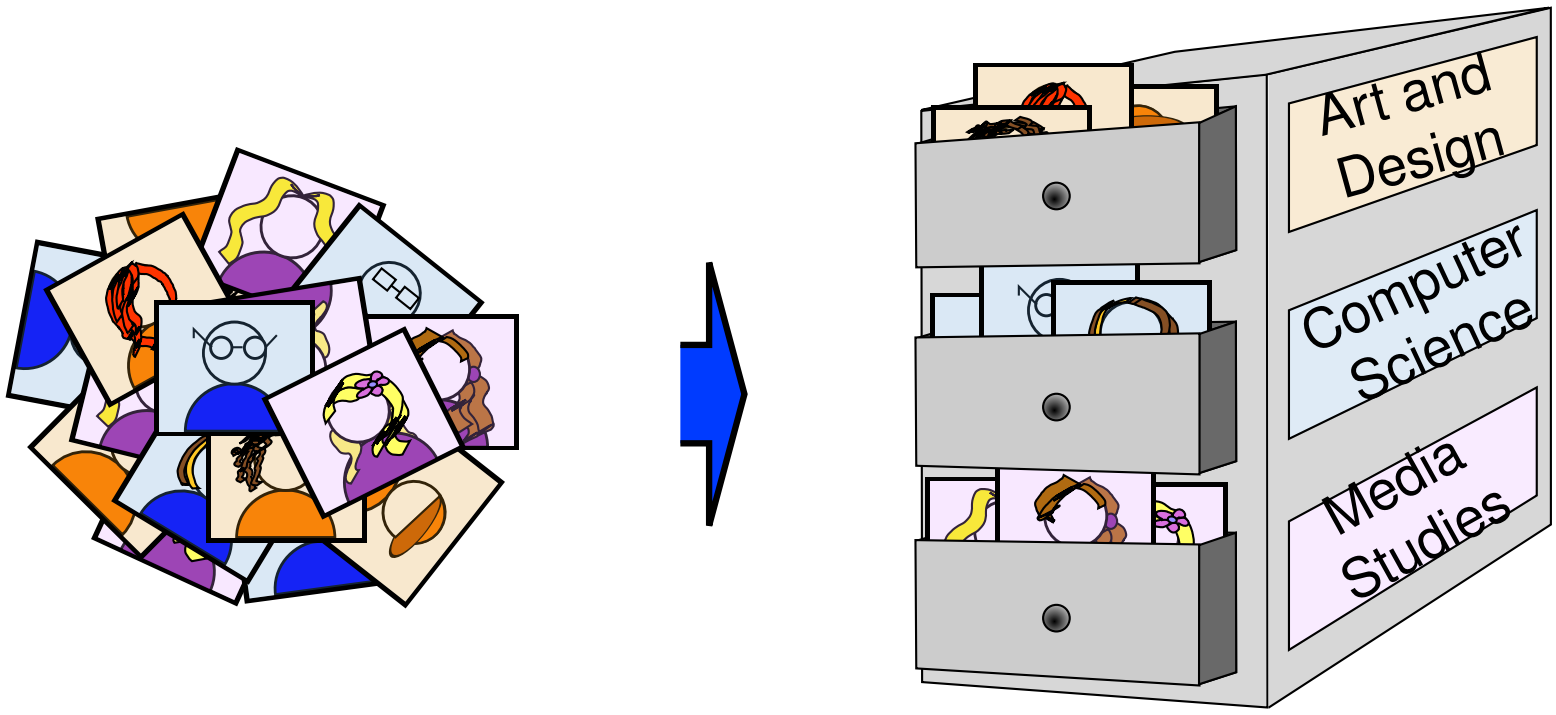
Clusterability in Model Selection

Johannes Kiesel

Bauhaus-Universität Weimar

28th May, 2014

Cluster Analysis: Motivation



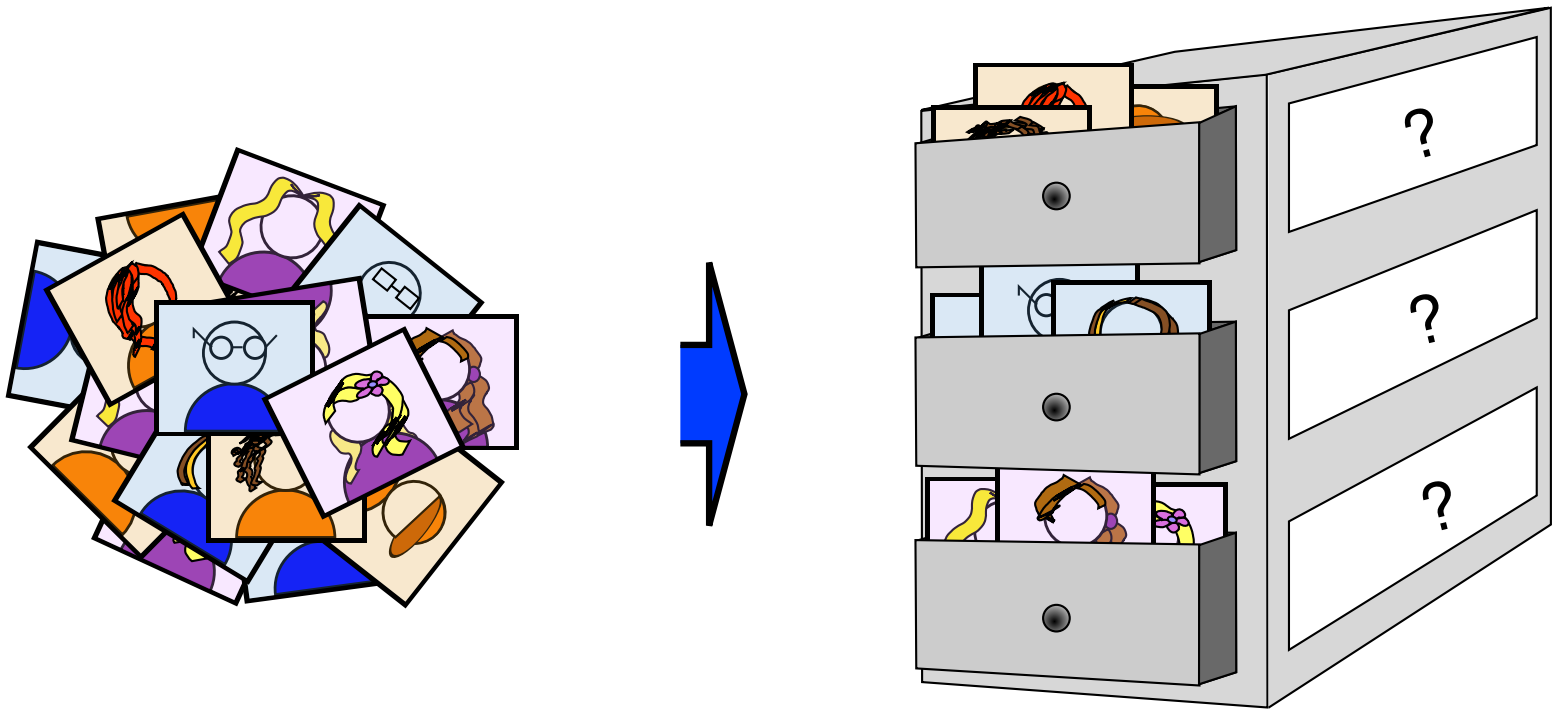
Data

Categorization

Given data (a set of comparable *entities* or *objects*)

Find a categorization of it

Cluster Analysis: Motivation



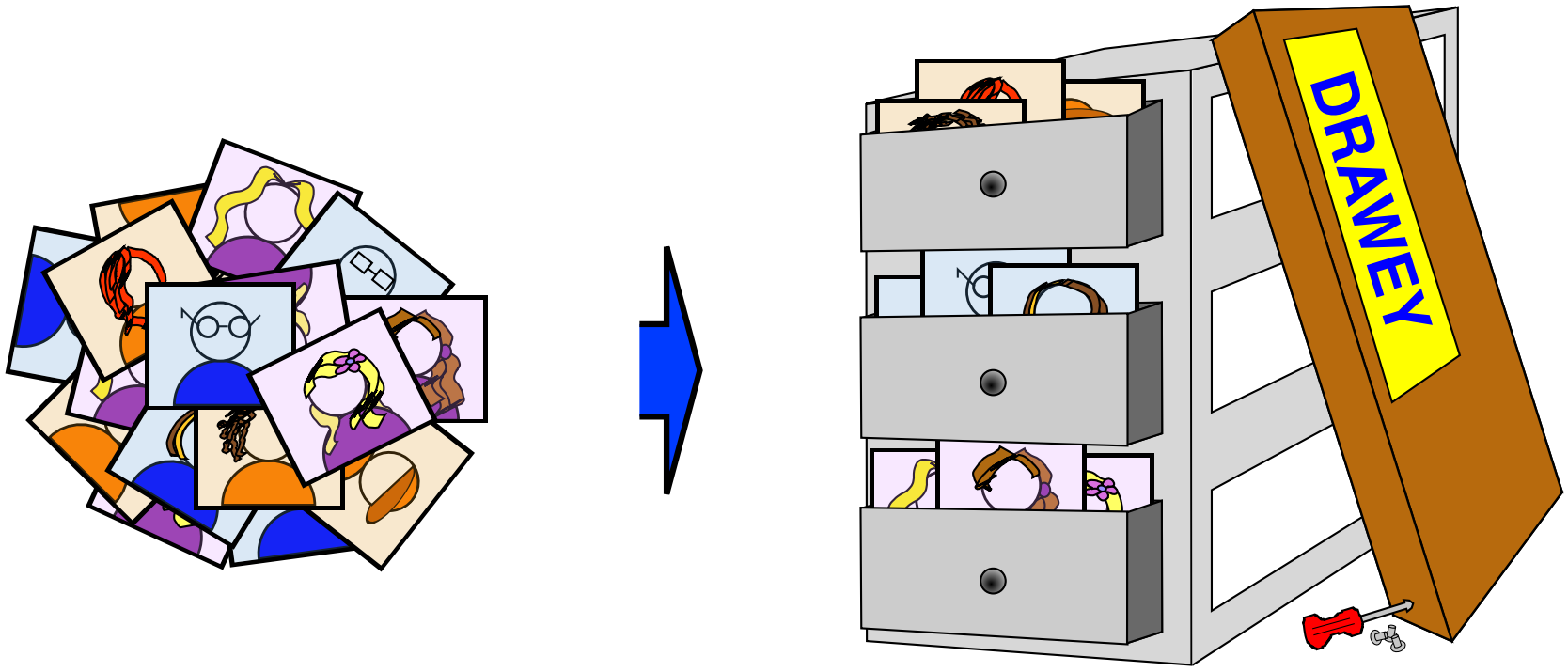
Data

Categorization

Given data (a set of comparable *entities* or *objects*)

Find a categorization of it (without labels)

Cluster Analysis: Motivation



Data

Categorization

Given data (a set of comparable *entities* or *objects*)

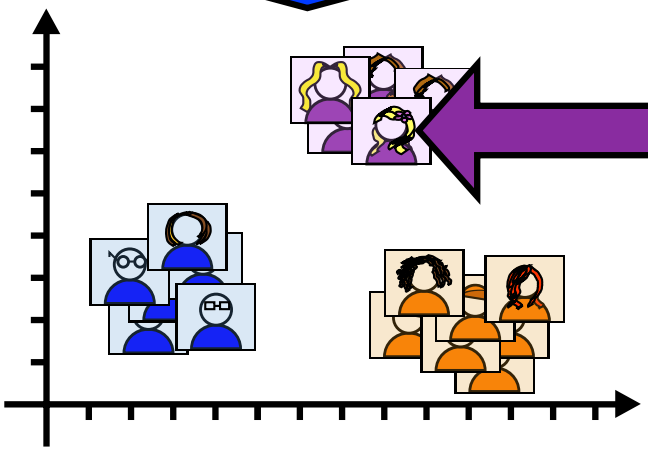
Find a categorization of it (without labels)

Cluster Analysis: In the Beginning was the Data

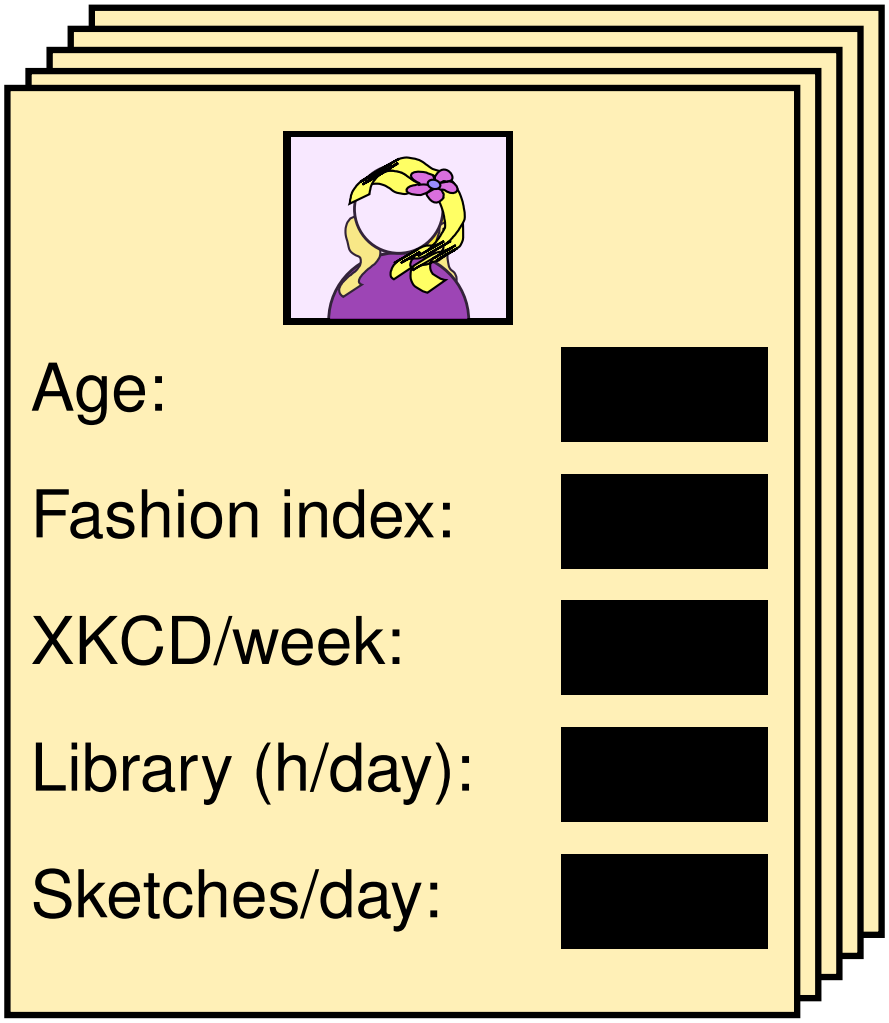
Data

Cluster Analysis: Modeling

Data

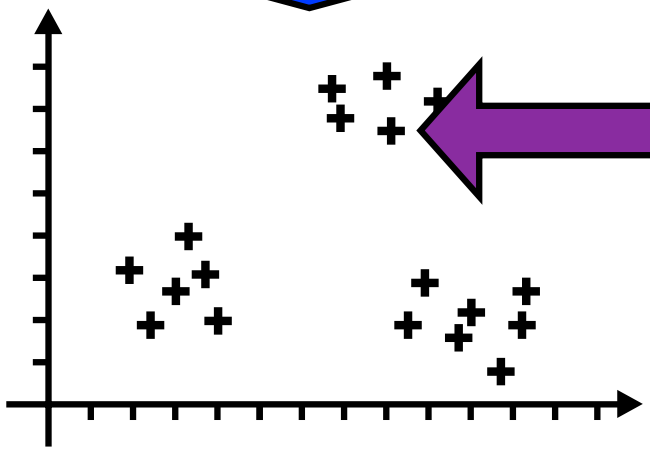


Model

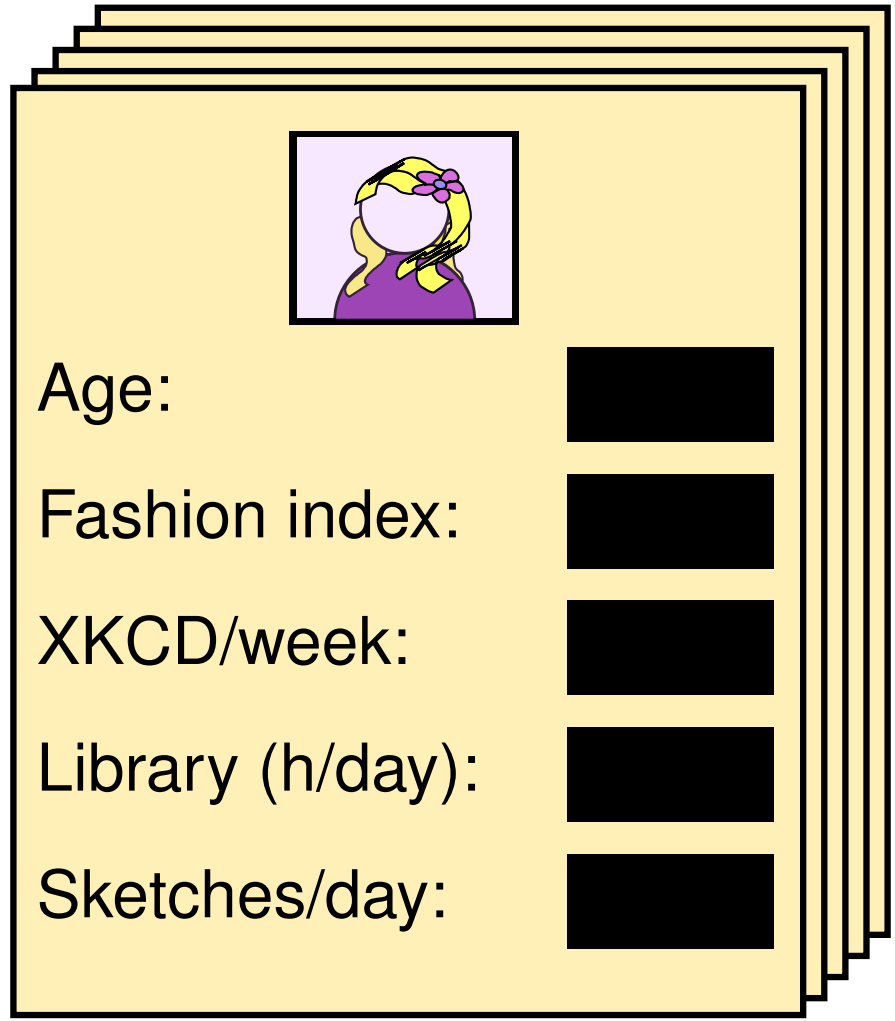


Cluster Analysis: Modeling

Data

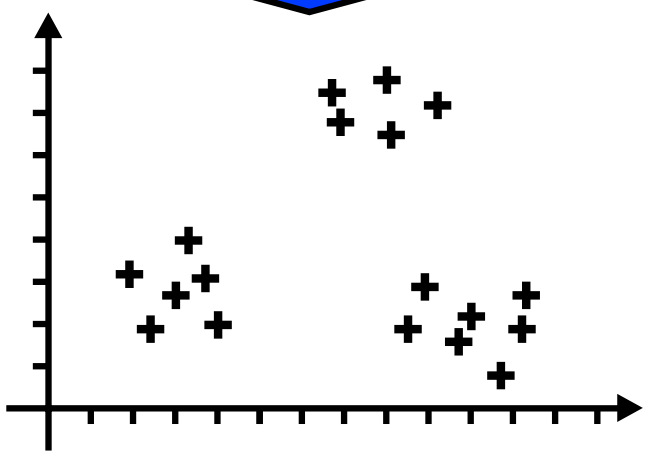


Model



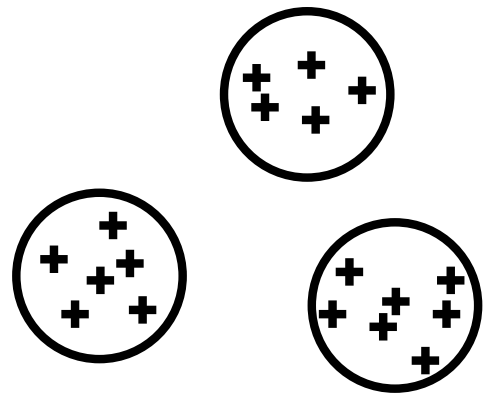
Cluster Analysis: Clustering

Data



Model

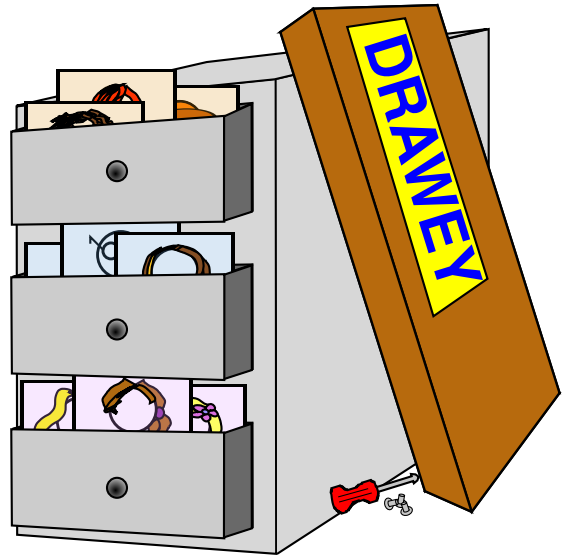
Clustering algorithm



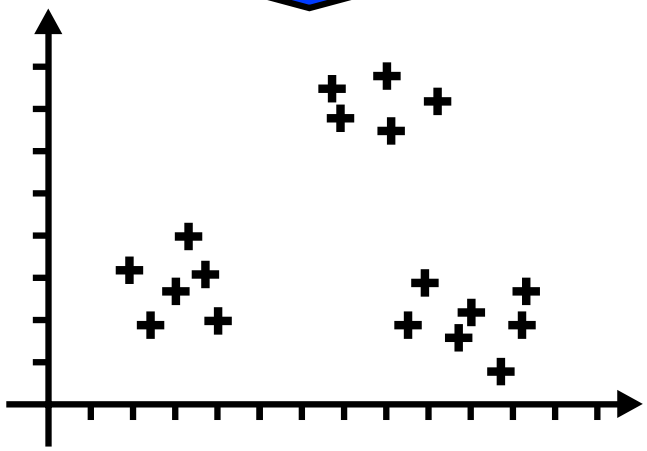
Clustering

Cluster Analysis: Clustering

Data

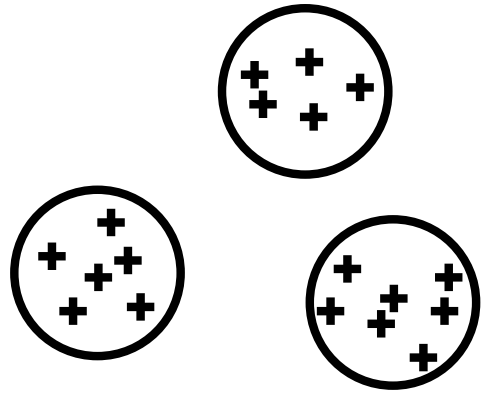


Categorization



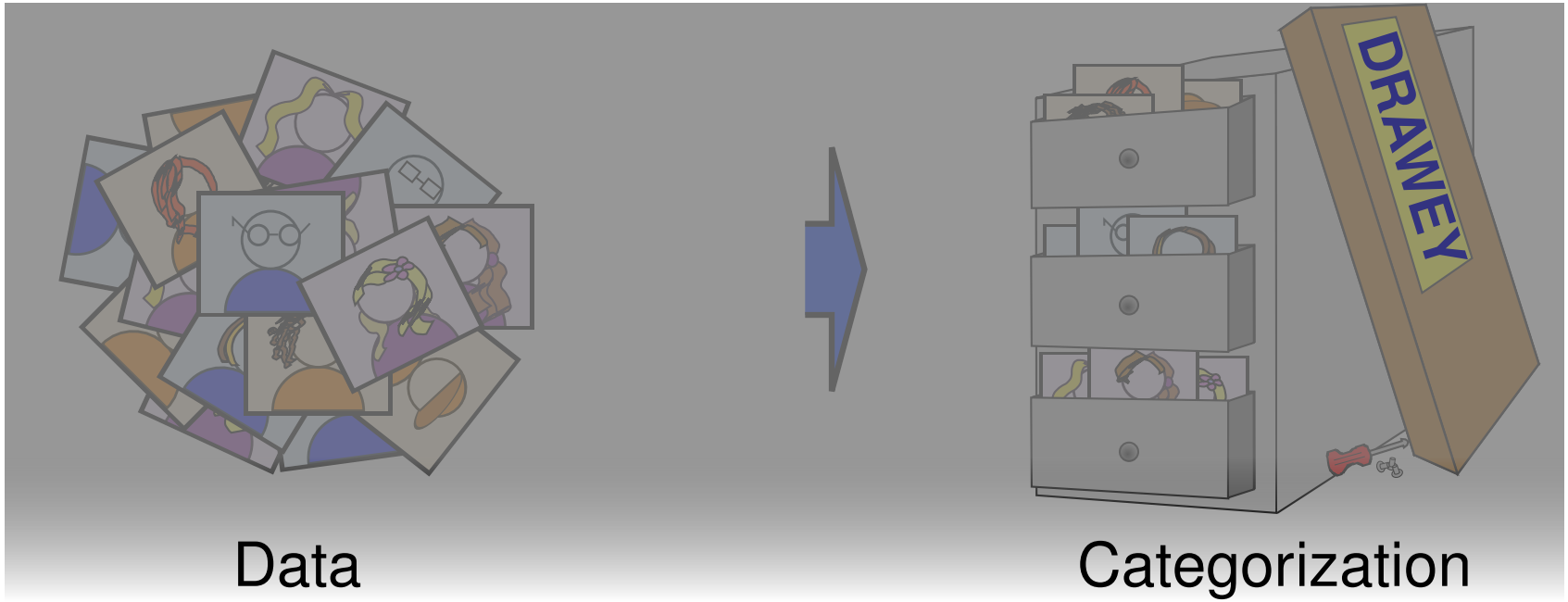
Model

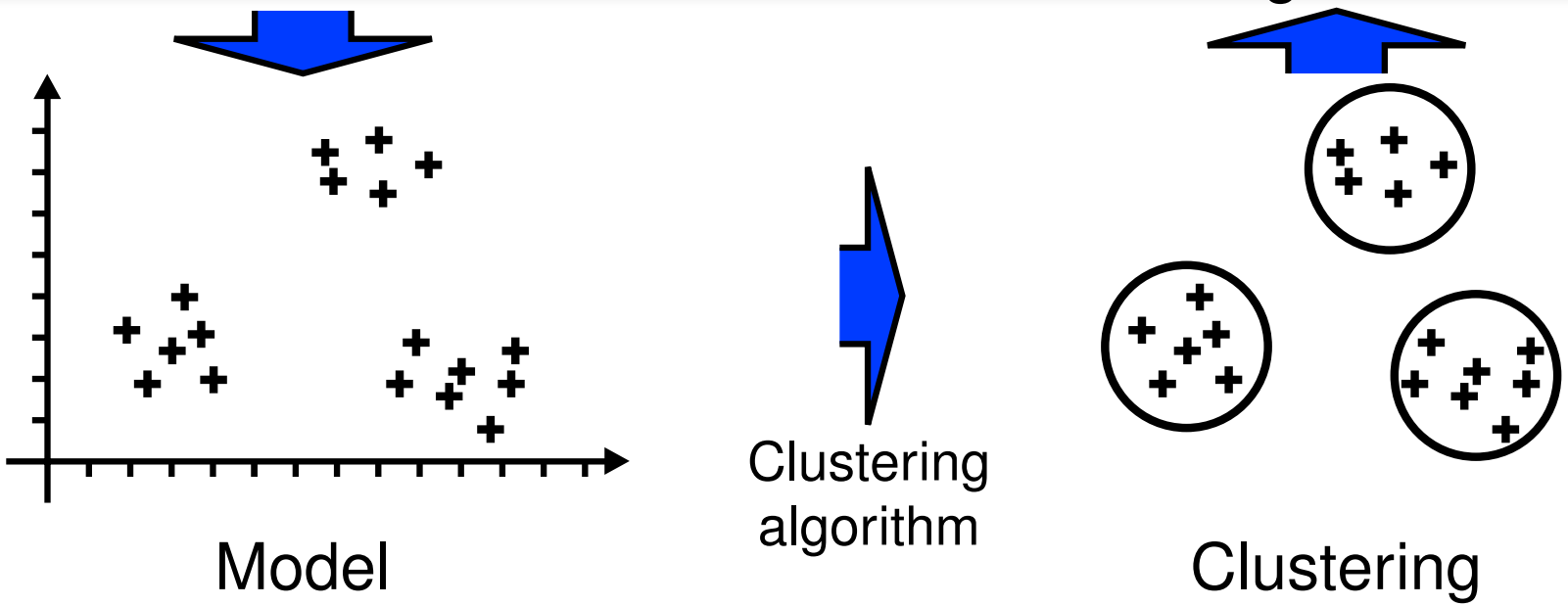
Clustering algorithm



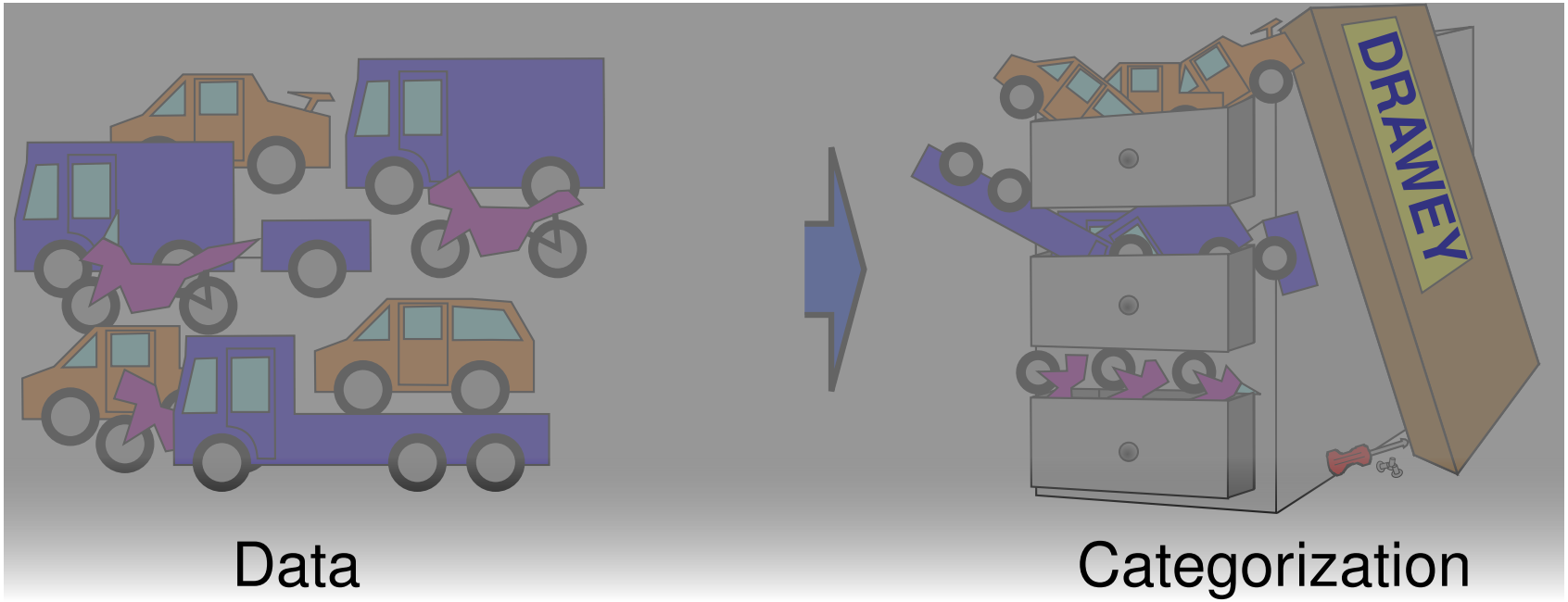
Clustering

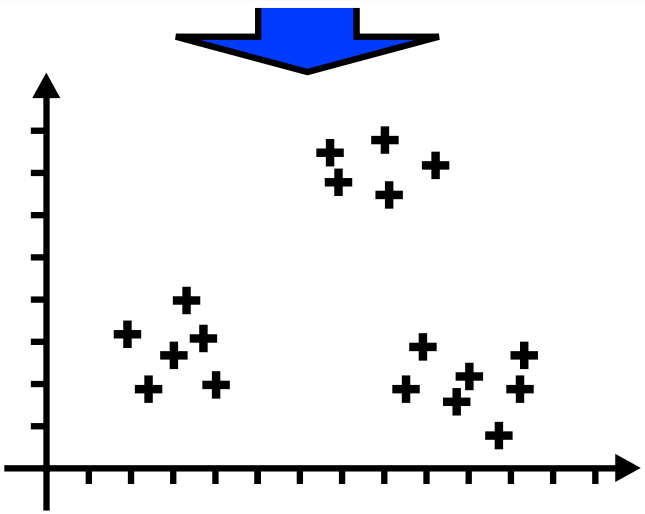
Cluster Analysis: Clustering





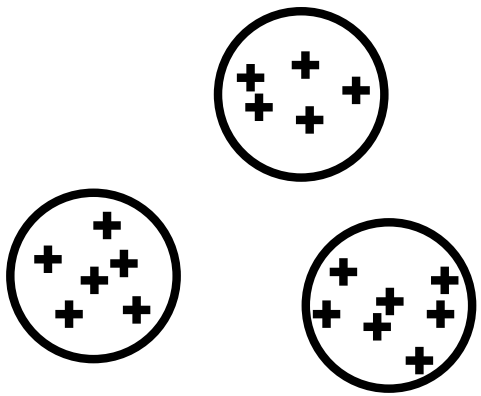
Cluster Analysis: Clustering





Model

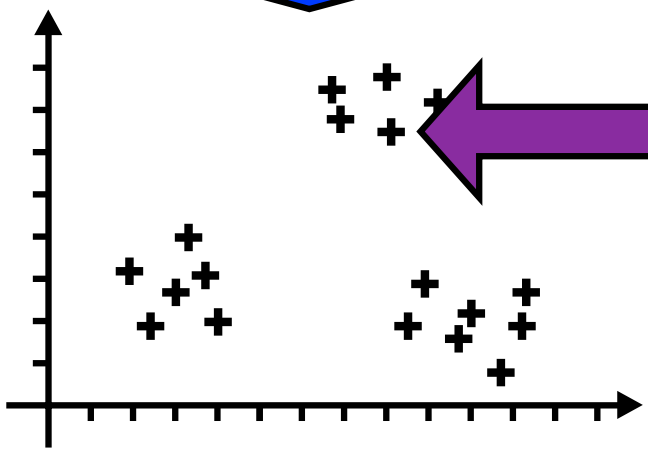
Clustering algorithm



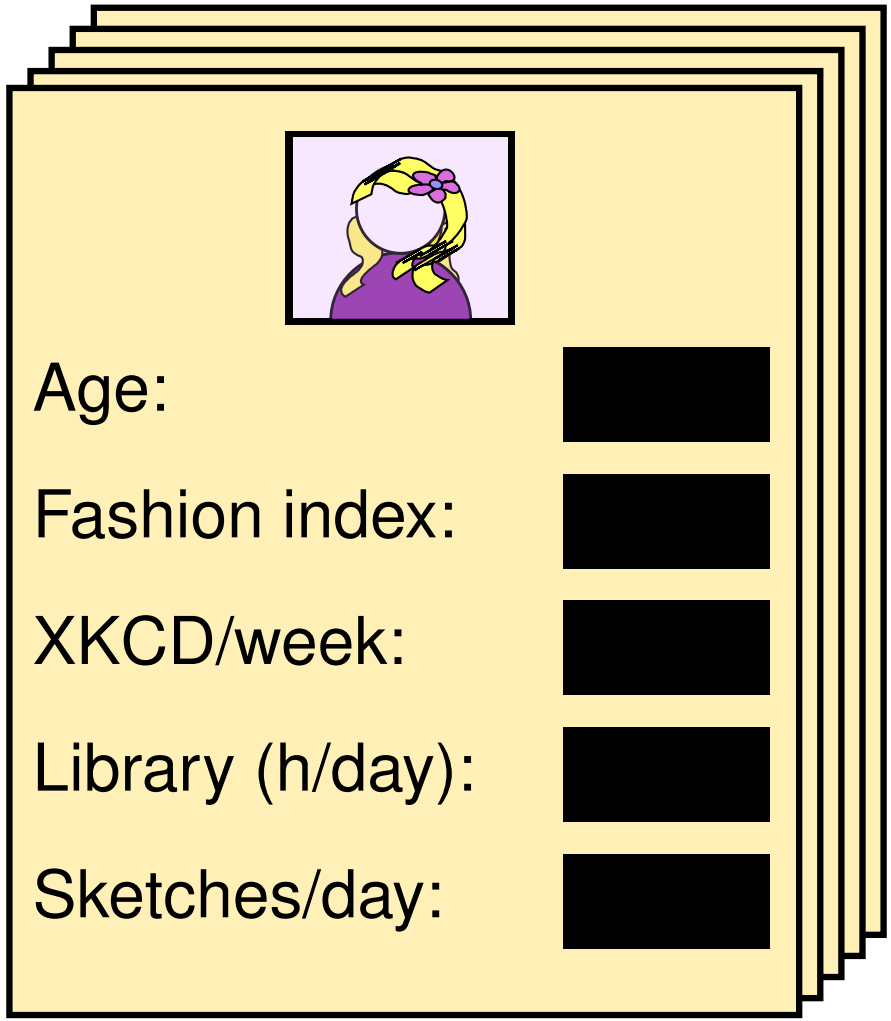
Clustering

Cluster Analysis: Modeling II

Data

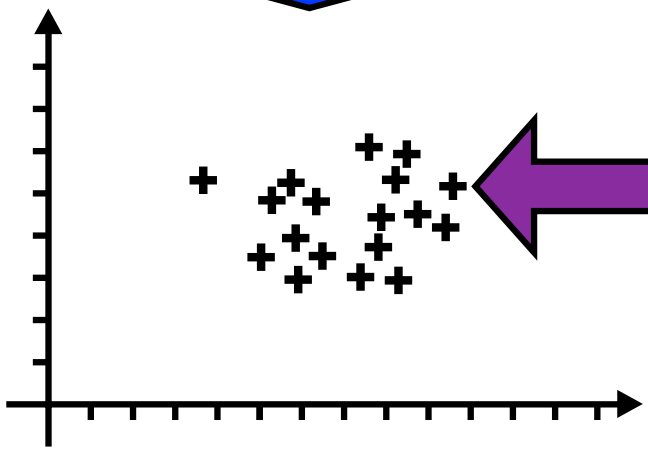


Model

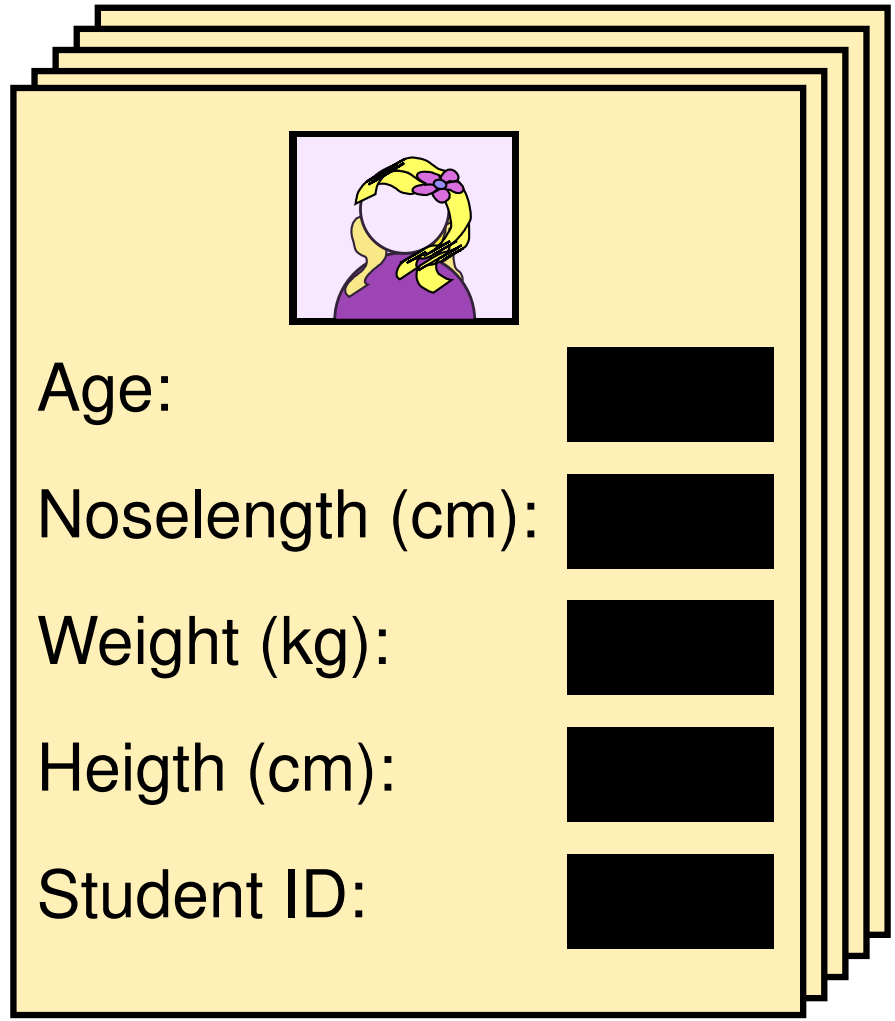


Cluster Analysis: Modeling II

Data

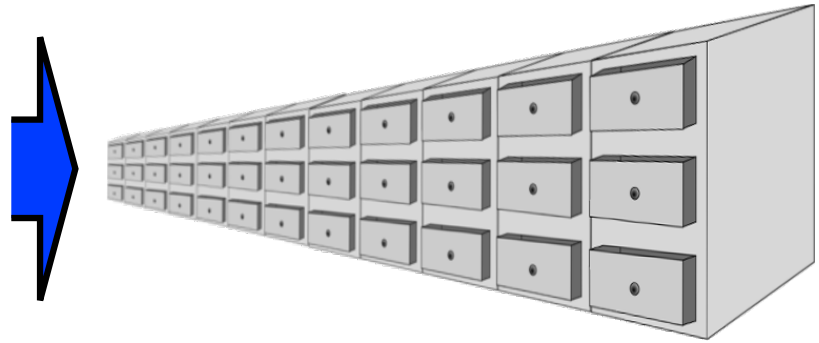


Model

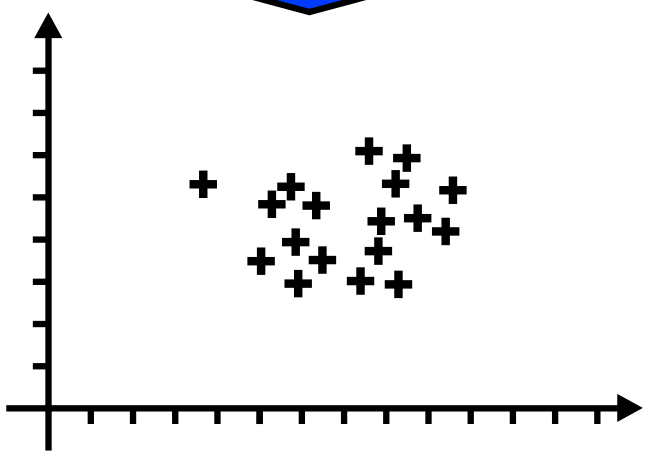


Cluster Analysis: Modeling II

Data

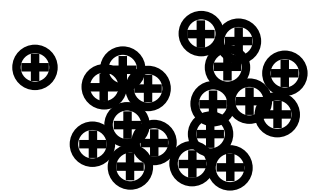


Categorization



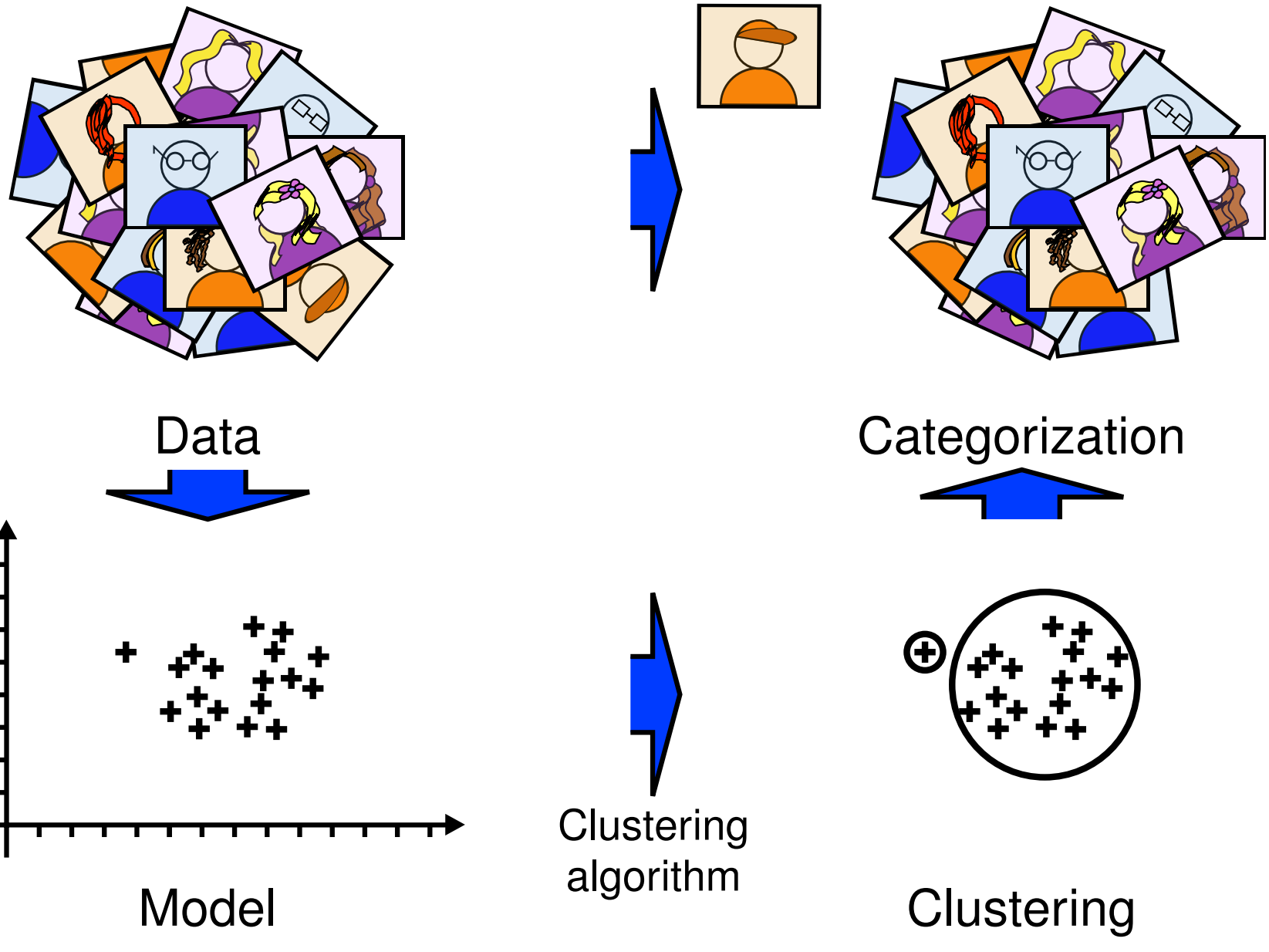
Model

Clustering algorithm

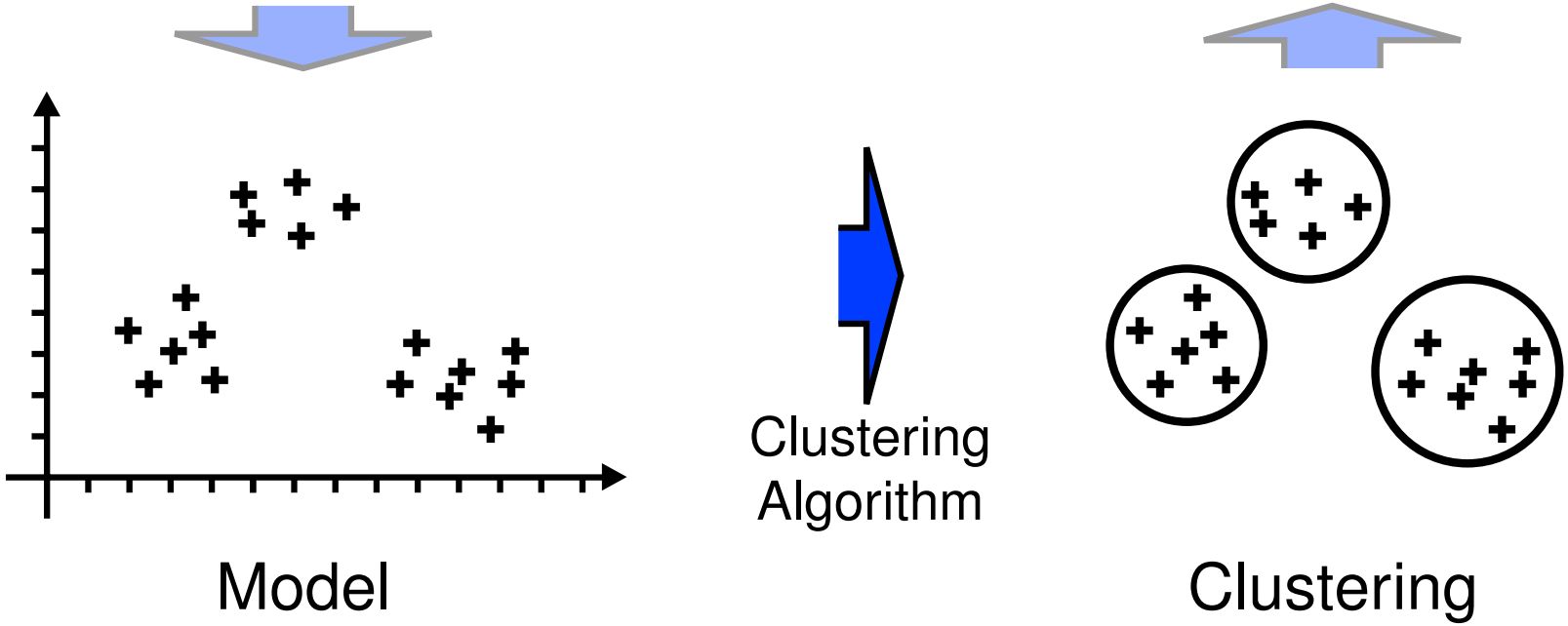


Clustering

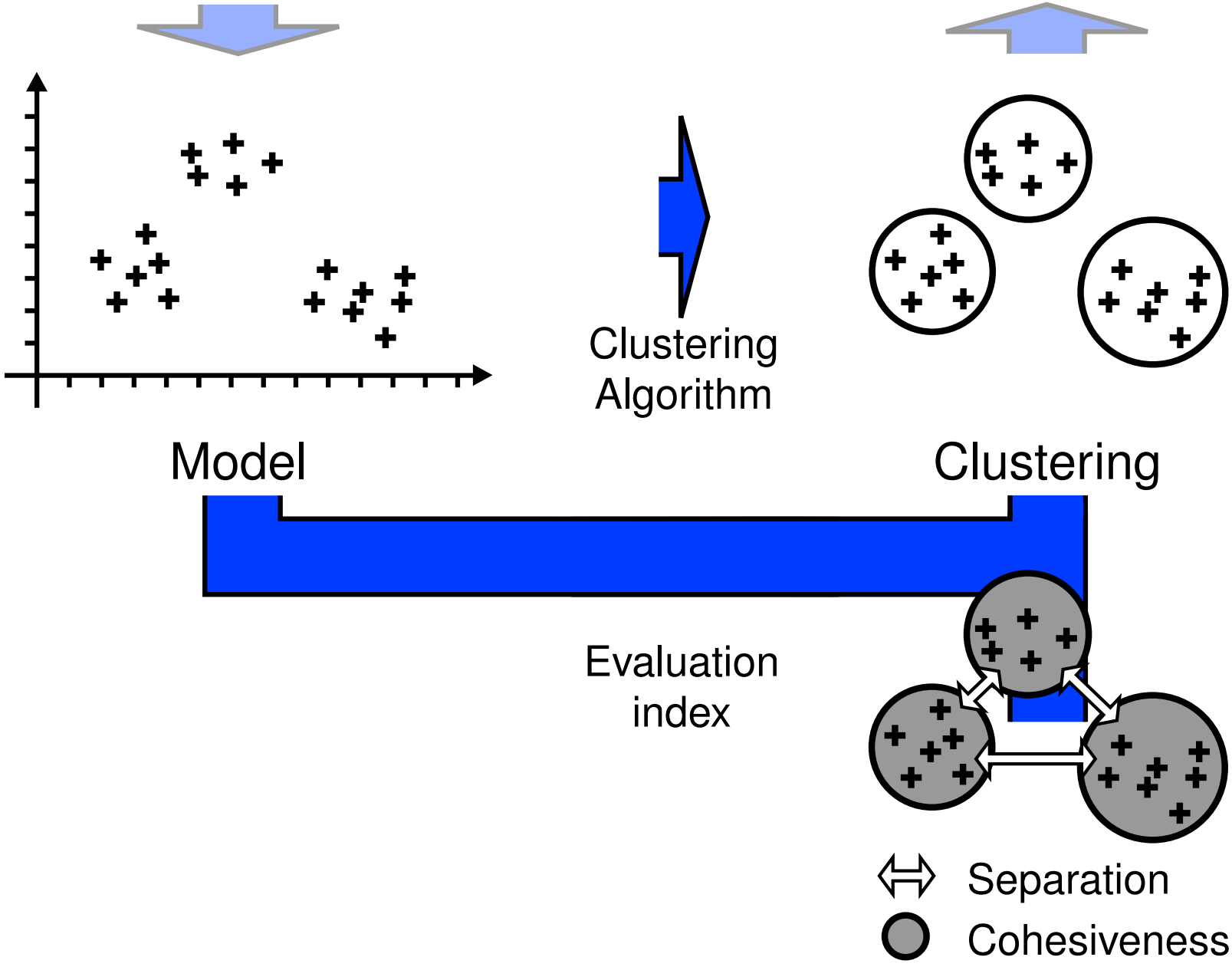
Cluster Analysis: Modeling II



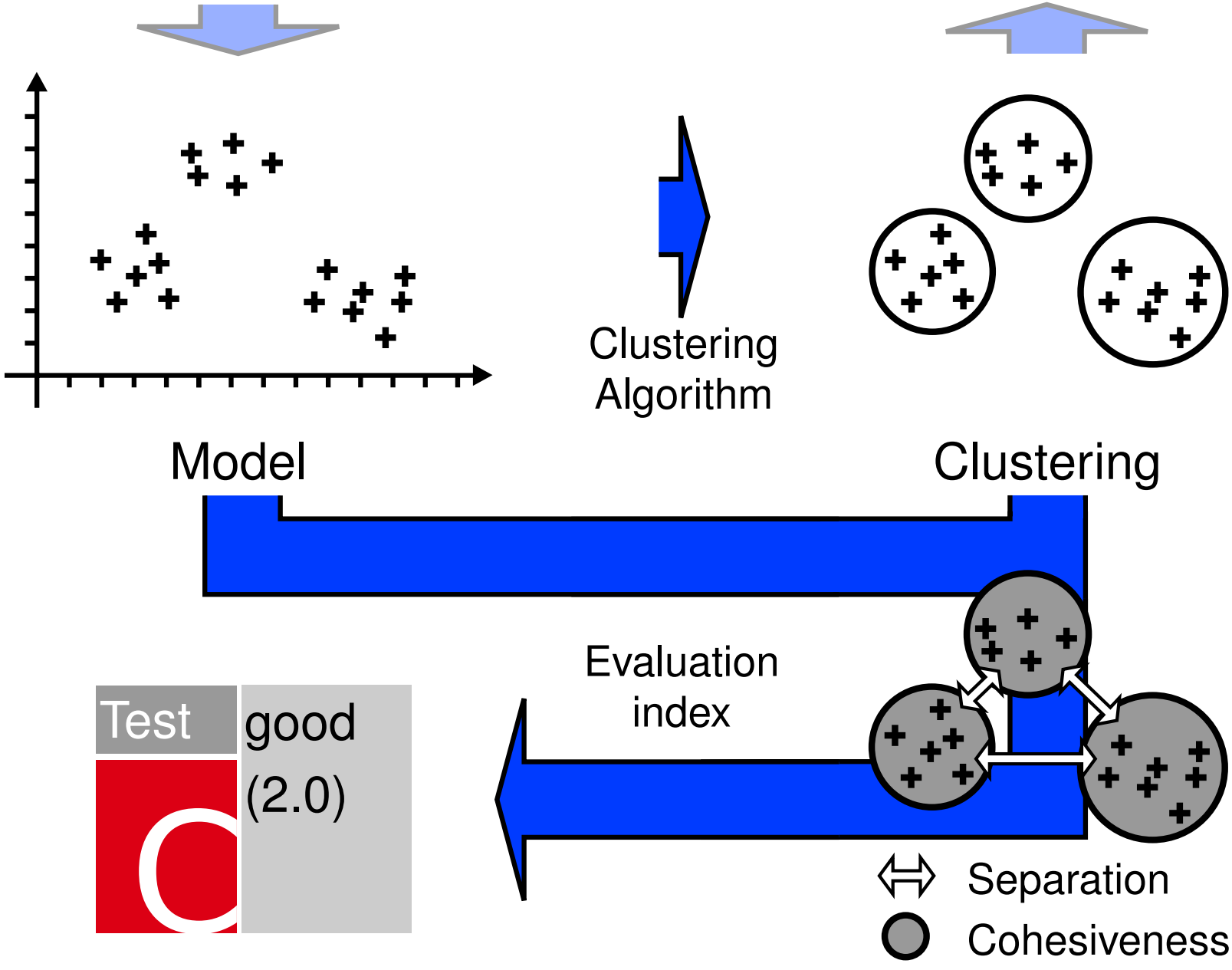
Cluster Analysis: Cluster Evaluation



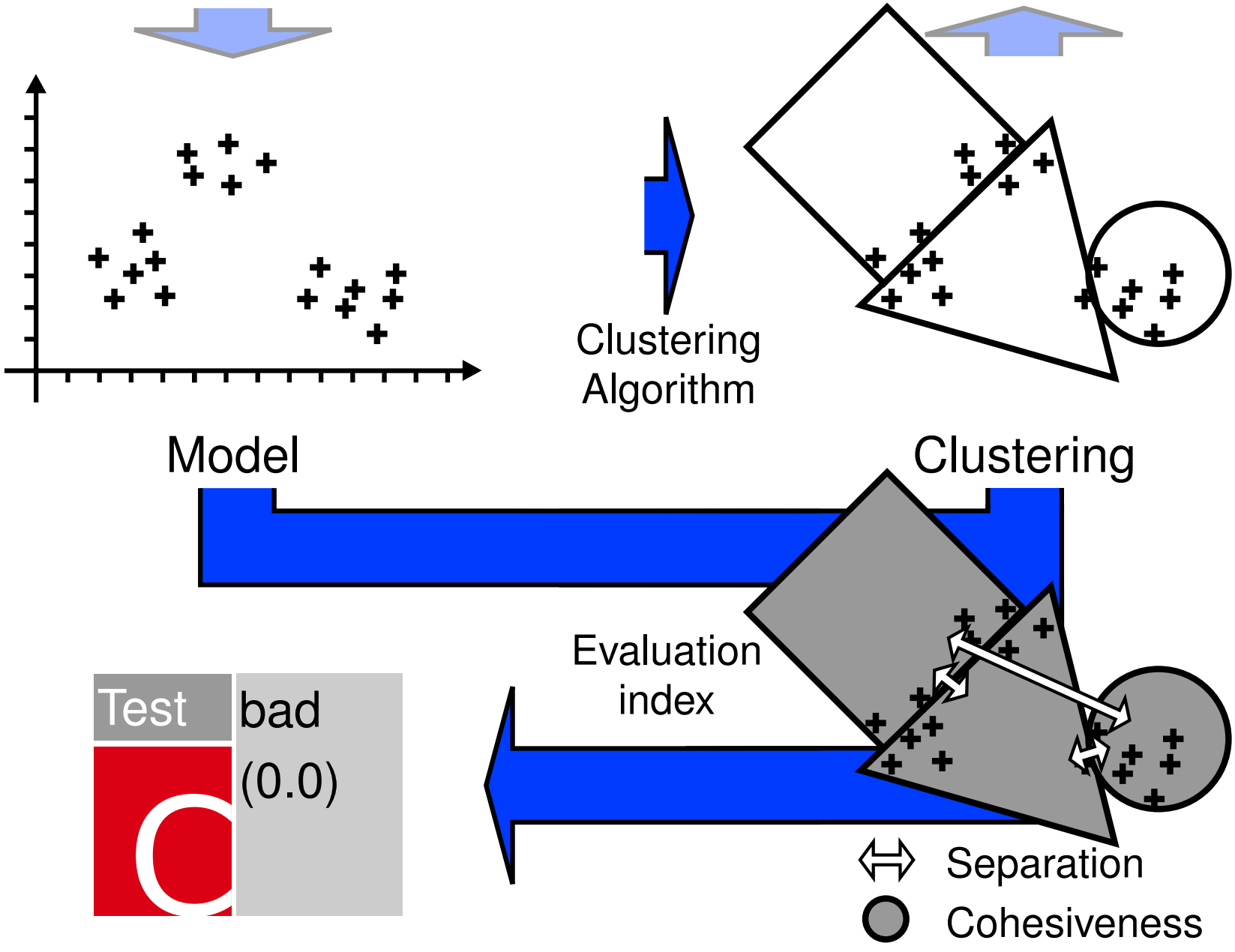
Cluster Analysis: Cluster Evaluation



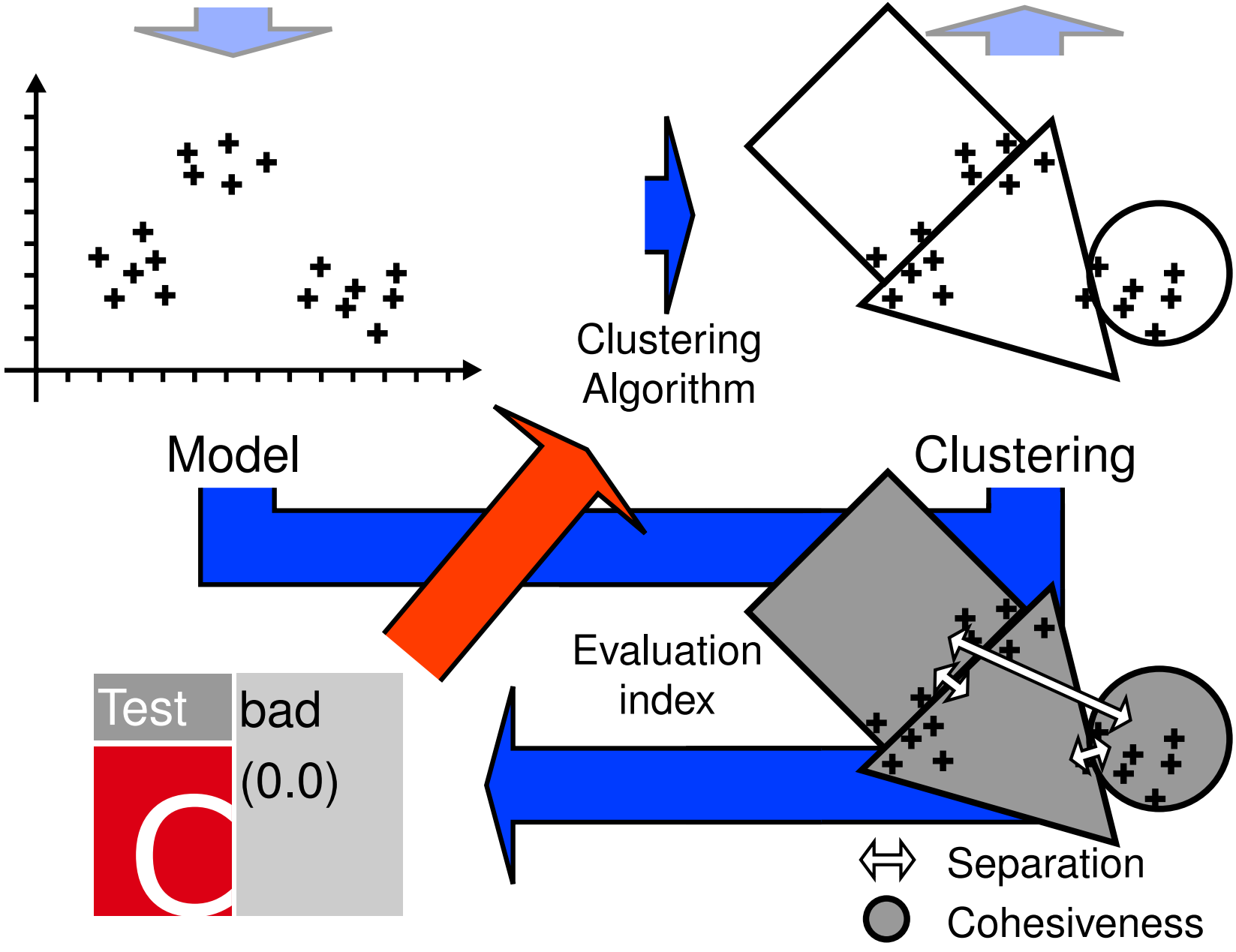
Cluster Analysis: Cluster Evaluation



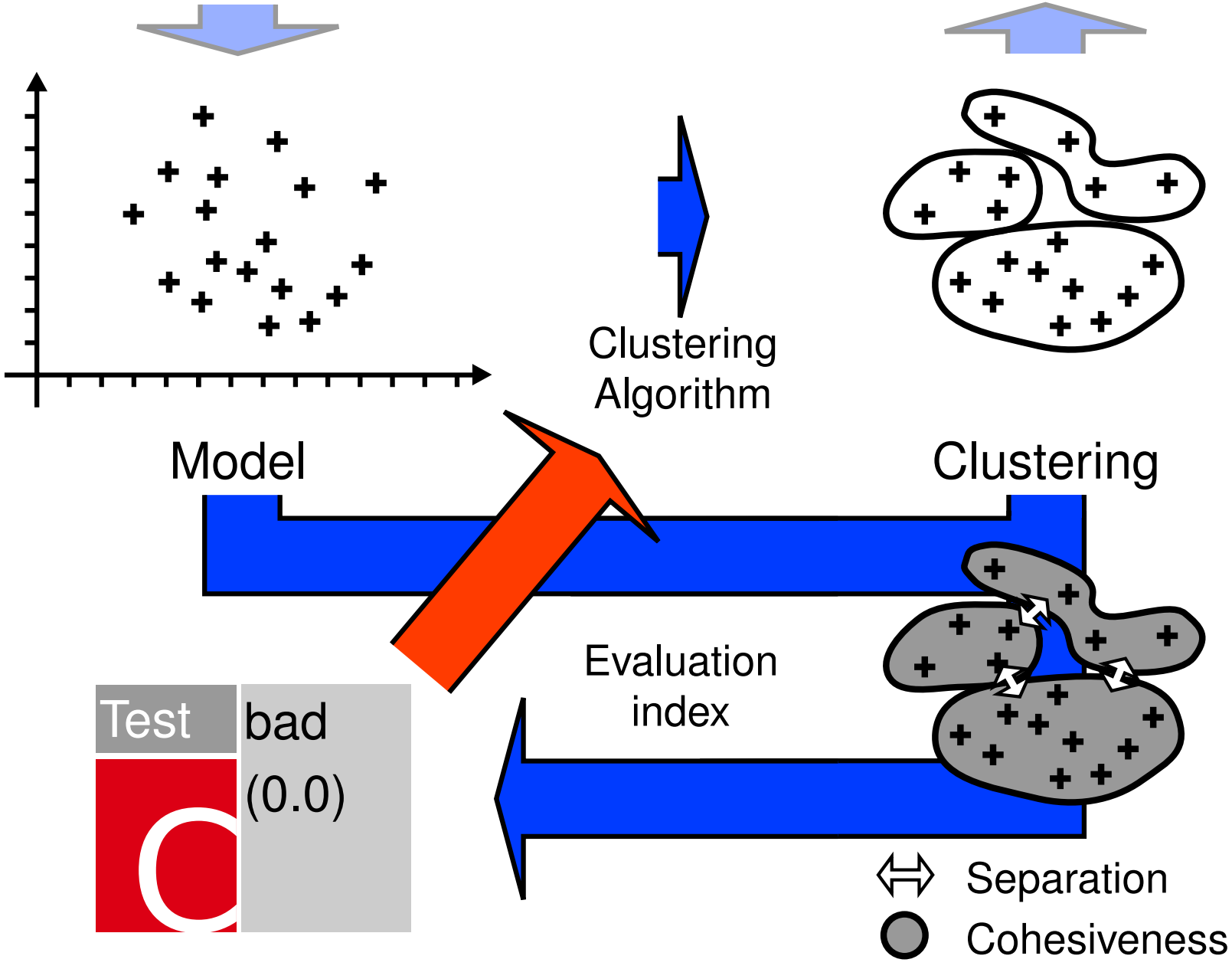
Cluster Analysis: Cluster Evaluation



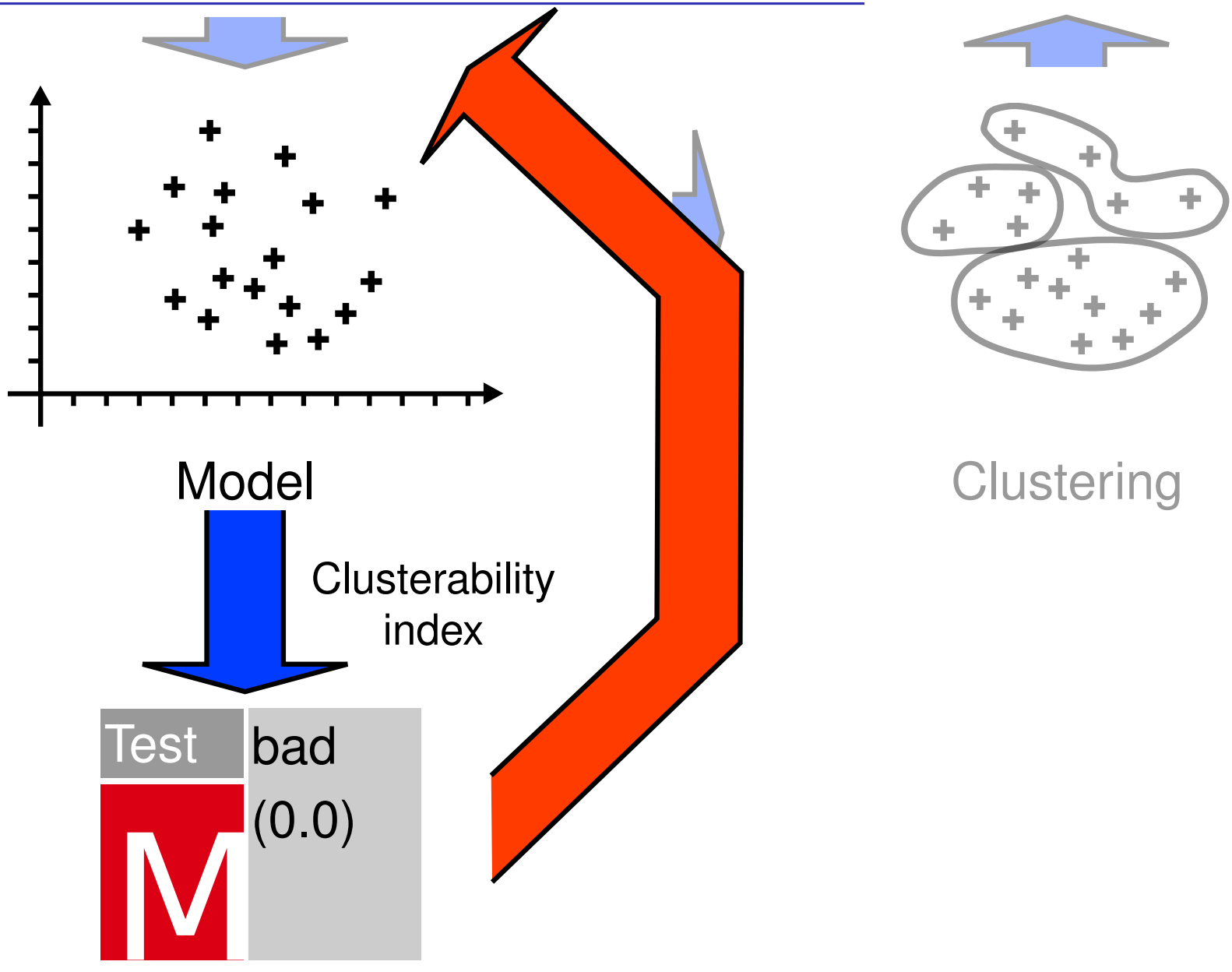
Cluster Analysis: Cluster Evaluation



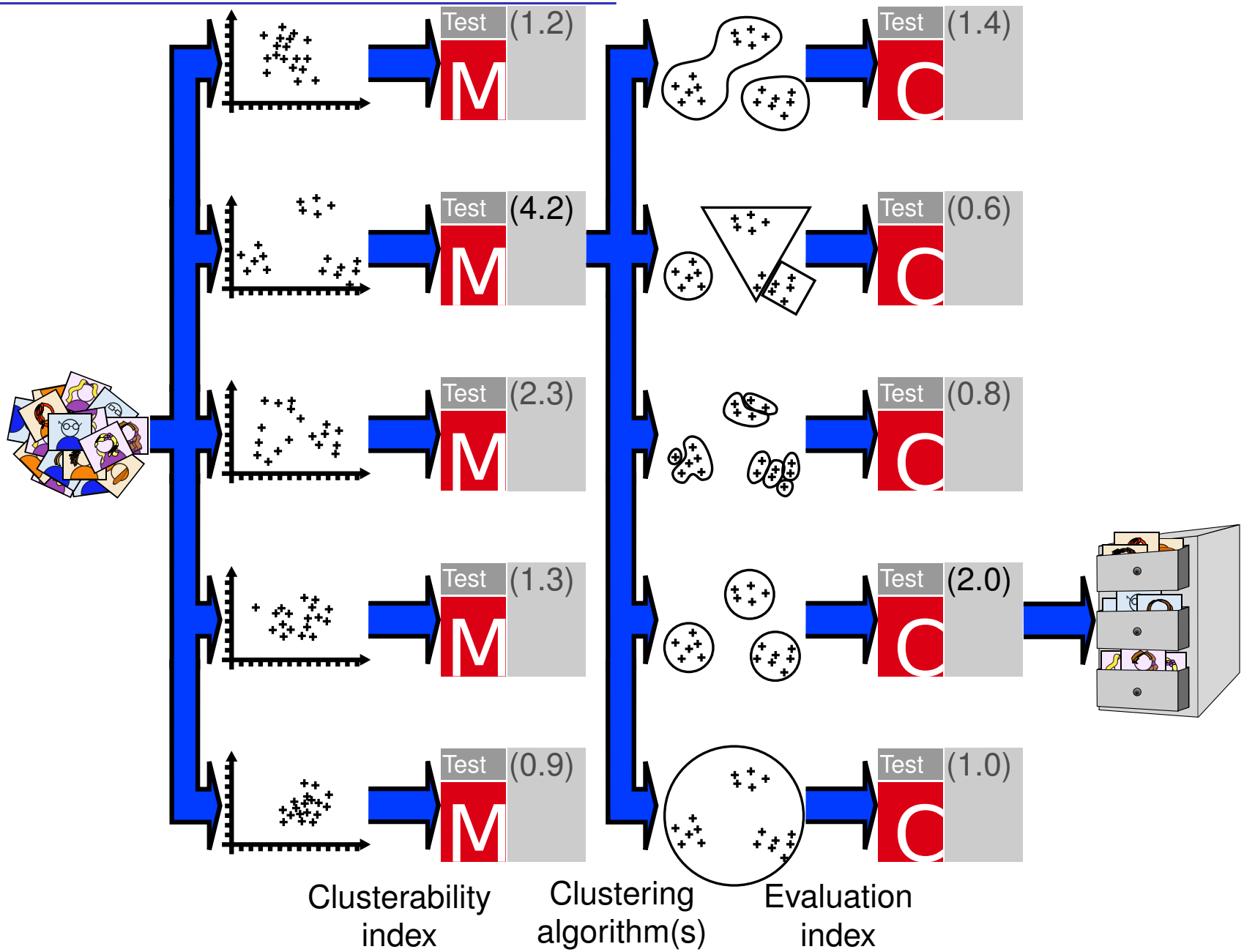
Cluster Analysis: Cluster Evaluation



Cluster Analysis: Model Evaluation

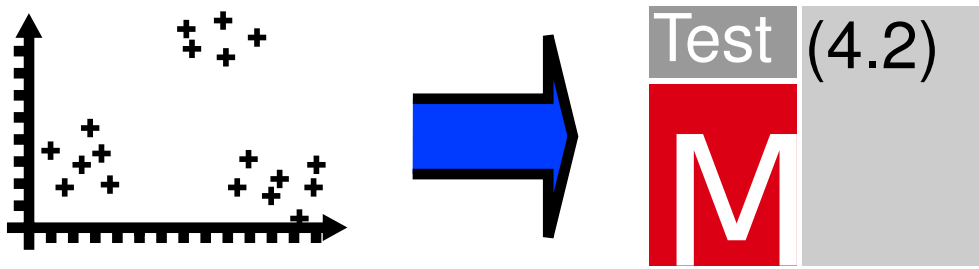


Cluster Analysis: Overview



Clusterability

- Task: calculate a score for a model
- Has to be comparable at least among similar models (same number of objects)

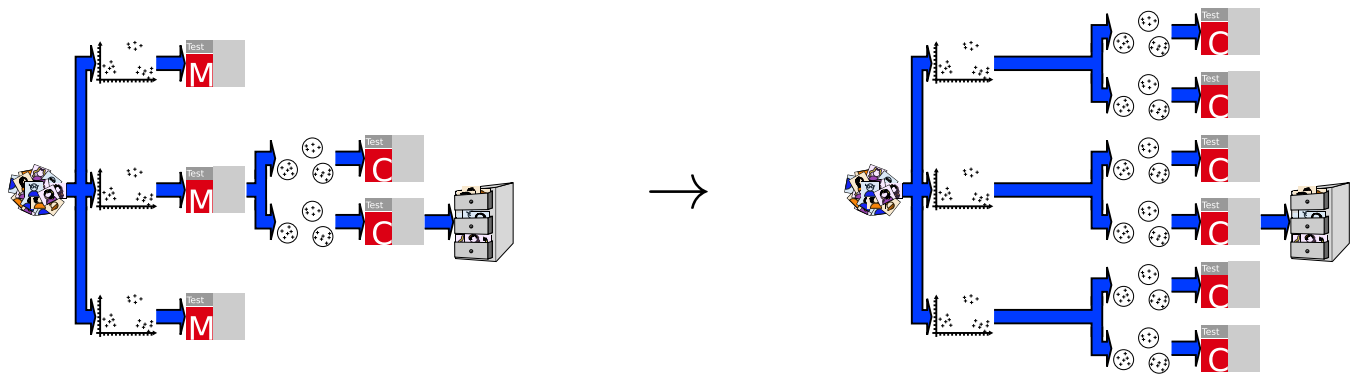


- A clusterable model (high score) has a dominant structure of mutually **separated** parts that are **cohesive** groups of objects.

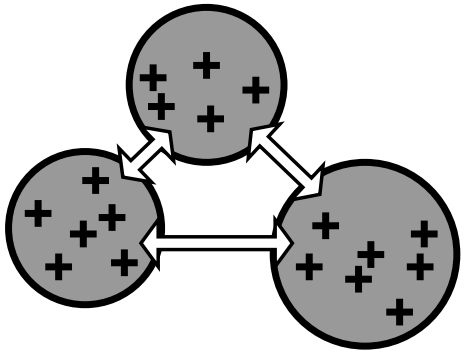
Clusterability I: Salient Clustering

Idea Model selection by cluster evaluation (“one-step”)

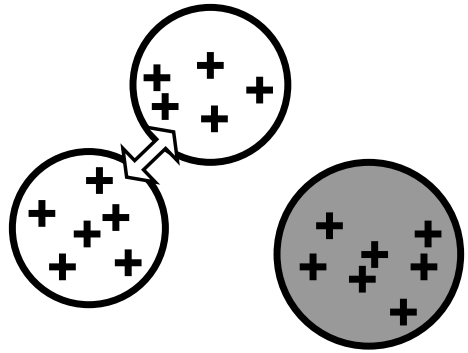
- Cluster the model with different algorithms and/or parameter settings
- Evaluate all clusterings
- Choose best combination of model & clustering



Clusterability I: Dunn Index

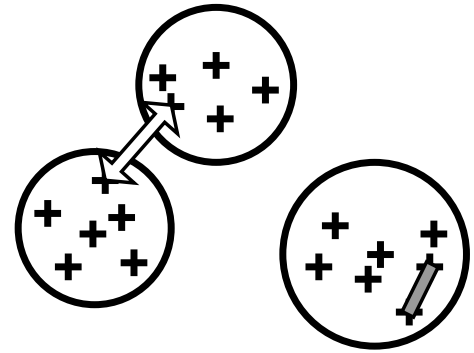


Evaluation index



Dunn index family

$\min(\Leftrightarrow)/\max(1/\bullet)$



Dunn MST

Minimum spanning tree Dunn index (Dunn MST)

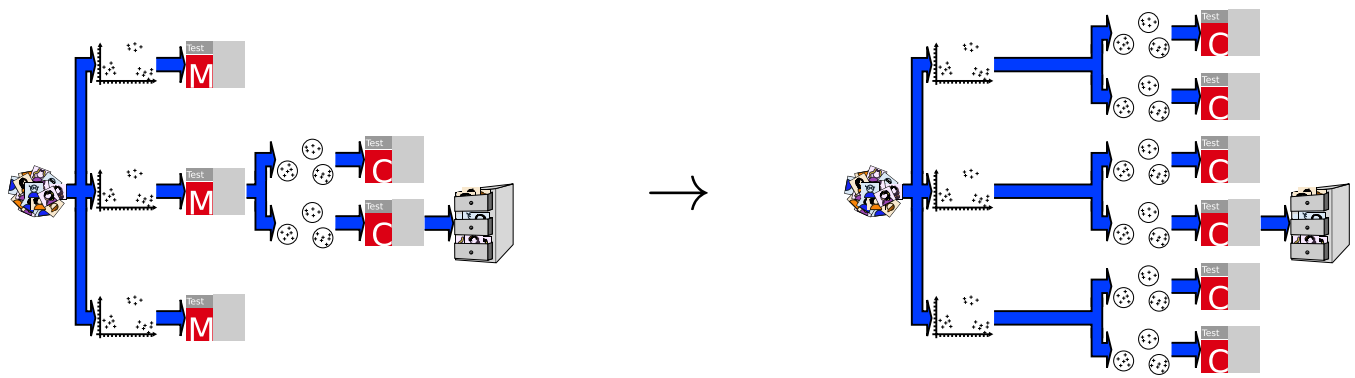
$1/\bullet$ Largest edge length in the minimum spanning tree of the cluster

$\Leftrightarrow$ Smallest dissimilarity of objects from different clusters

Optimum clustering is feasibly computable
(no other clustering algorithm necessary)

Clusterability I: Salient Clustering

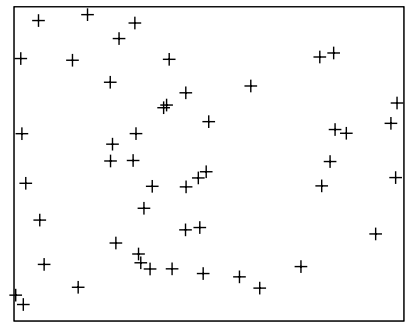
+	-
+ Needs no additional clusterability index	- Most evaluation indices require local optimization
+ Evaluation indices are better understood	- Not all evaluation indices can compare clusterings of different models



Clusterability II: Statistical Tests on Structure

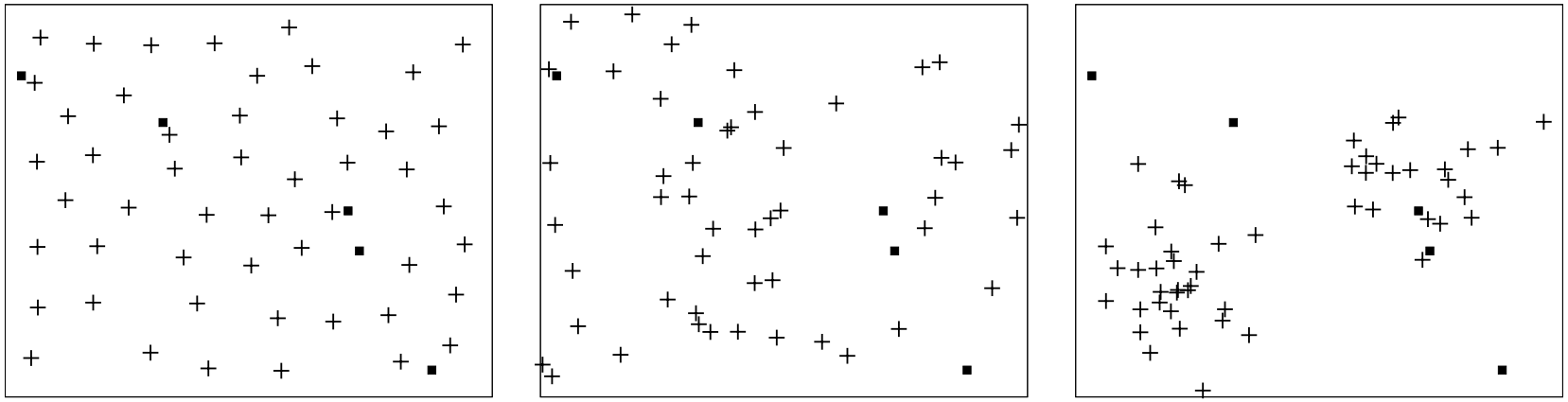
Idea Use a statistical test for unstructured models

- Null hypothesis: model generated from a model distribution that generates non-clusterable models (e.g., uniform distribution)



- Calculate a test statistic with known distribution under the null hypothesis
- Use the probability that a similar large value occurs under the null hypothesis for the clusterability assessment

Clusterability II: Hopkins and Skellam Statistic



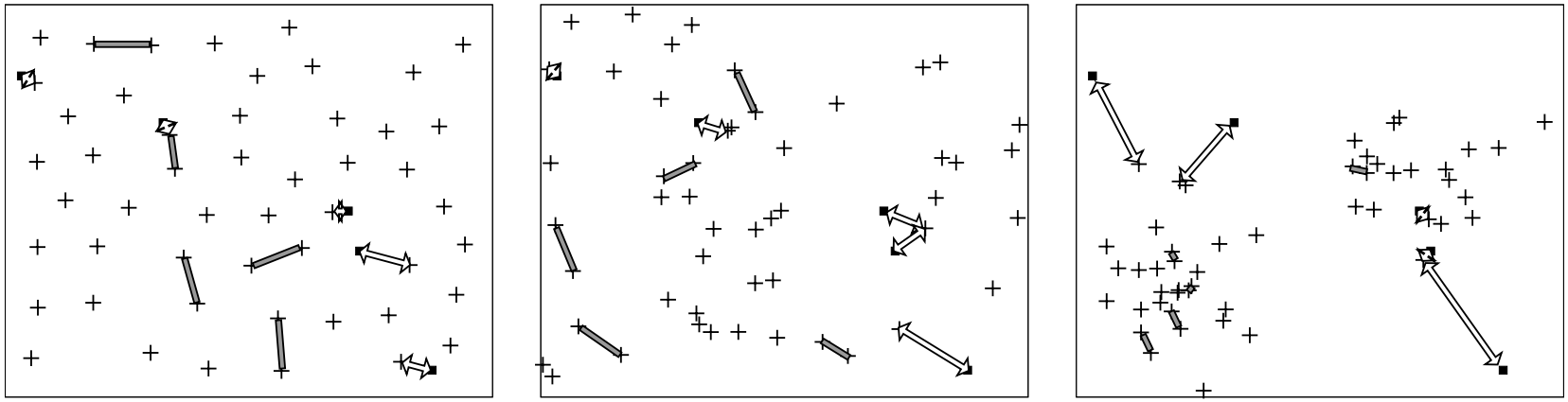
+ $\mathbf{x}$ ▪ $\mathbf{x}^0$

spaced uniform clustered

Compare distribution of original objects ($\mathbf{x}$) and r uniformly sampled $\mathbf{x}^0$ (null hypothesis)

[Hopkins and Skellam. A New Method for Determining the Type of Distribution of Plant Individuals. 1954]

Clusterability II: Hopkins and Skellam Statistic



$+\ \mathbf{x}$ $\blacksquare\ \mathbf{x}^0$ $\psi_{nn}(\mathbf{x})$ $\Leftrightarrow \psi_{nn}(\mathbf{x}^0)$

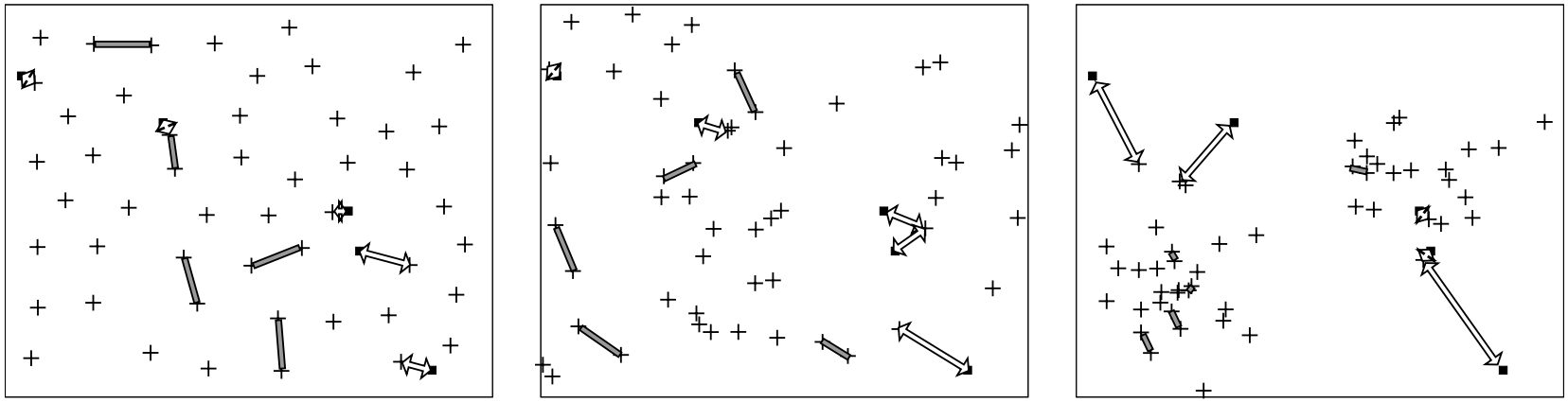
$\mathrm{H}_r \to 0$ $\mathrm{H}_r \approx 0.5$ $\mathrm{H}_r \to 1$

Compare distribution of original objects ($\mathbf{x}$) and r uniformly sampled $\mathbf{x}^0$ (null hypothesis)

$\psi_{nn}(\mathbf{x})$ Dissimilarity of $\mathbf{x}$ to its nearest neighbor

[Hopkins and Skellam. A New Method for Determining the Type of Distribution of Plant Individuals. 1954]

Clusterability II: Hopkins and Skellam Statistic



$+\ \mathbf{x}$ $\blacksquare\ \mathbf{x}^0$ $\psi_{nn}(\mathbf{x})$ $\Leftrightarrow \psi_{nn}(\mathbf{x}^0)$

$\mathrm{H}_r \to 0$ $\mathrm{H}_r \approx 0.5$ $\mathrm{H}_r \to 1$

Compare distribution of original objects ($\mathbf{x}$) and r uniformly sampled $\mathbf{x}^0$ (null hypothesis)

$$\mathrm{H}_r = \frac{\sum_{i=1}^{r} (\psi_{nn}(\mathbf{x}_i^0))^m}{\sum_{i=1}^{r} (\psi_{nn}(\mathbf{x}_i^0))^m + \sum_{i=1}^{r} (\psi_{nn}(\mathbf{x}_{\pi(i)}))^m}$$

$\psi_{nn}(\mathbf{x})$ Dissimilarity of $\mathbf{x}$ to its nearest neighbor

m Number of dimensions

[Hopkins and Skellam. A New Method for Determining the Type of Distribution of Plant Individuals. 1954]

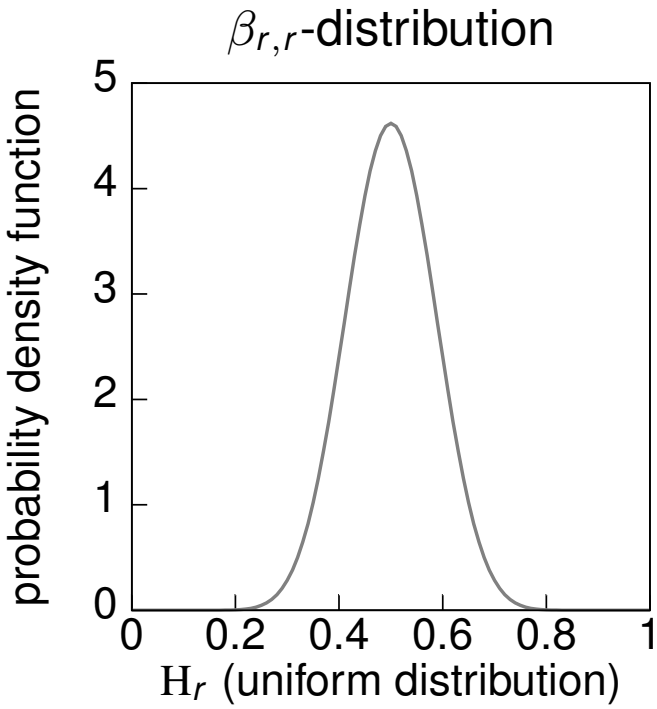
Clusterability II: Statistical Tests on Structure

+

+ The distribution under the null hypothesis allows for an interpretation of the score
+ Often requires only a sample

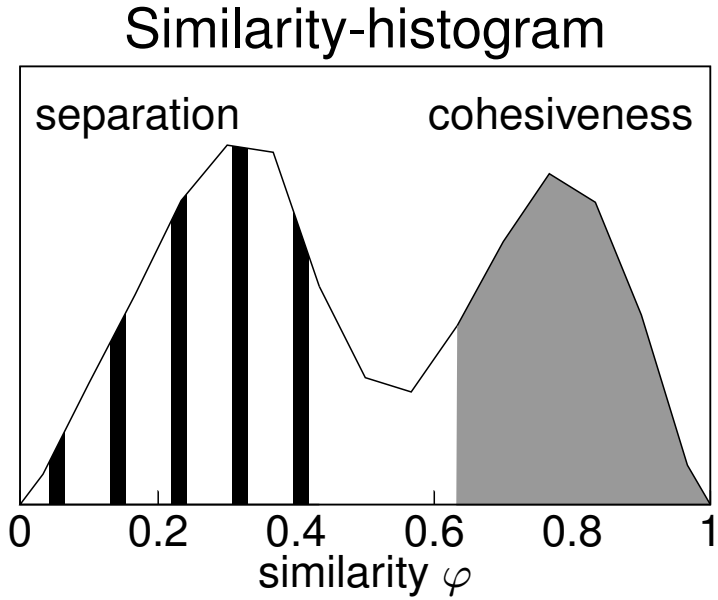
-

- Depends heavily on the null hypothesis
- Adjustment of statistics is not trivial



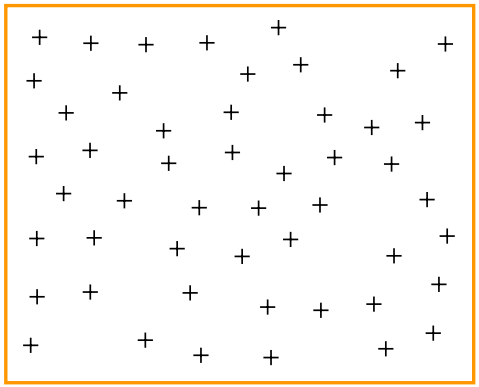
Clusterability III: Concentration of Dissimilarities

Idea In a clusterable model most object pairs should be either very dissimilar (different clusters) or very similar (same clusters)

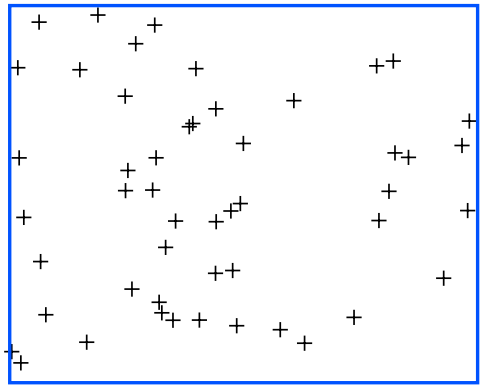


- Test if relatively few dissimilarities are of average size

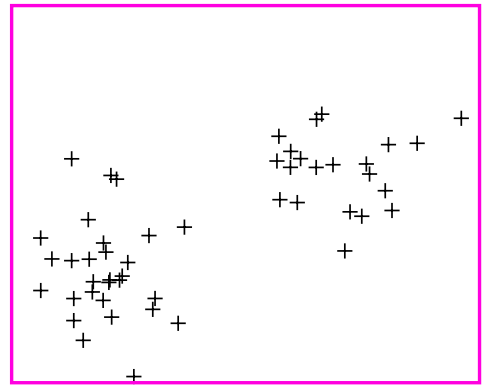
Clusterability III: Dash et al. score



spaced

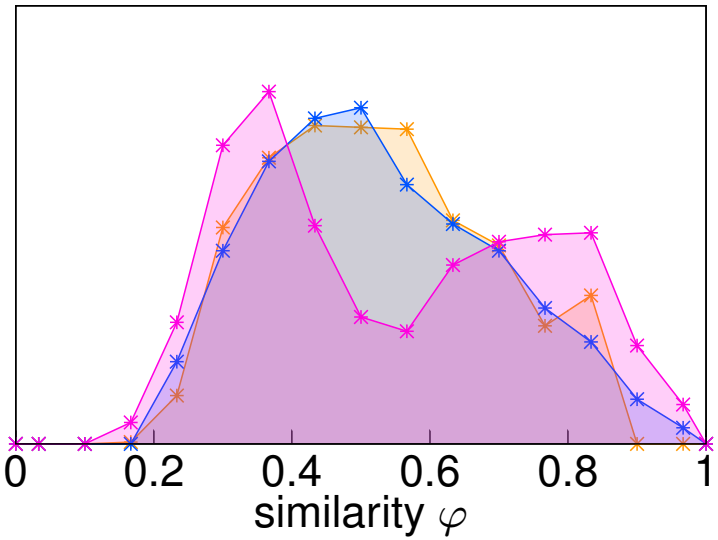


uniform



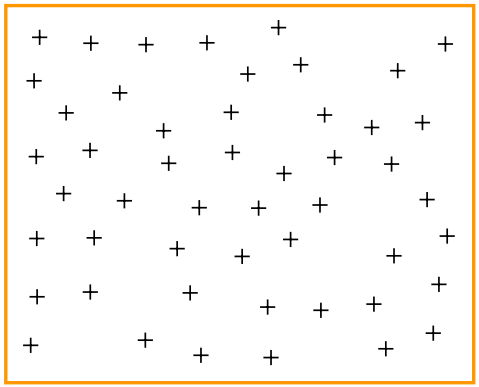
clustered

Similarity-histogram

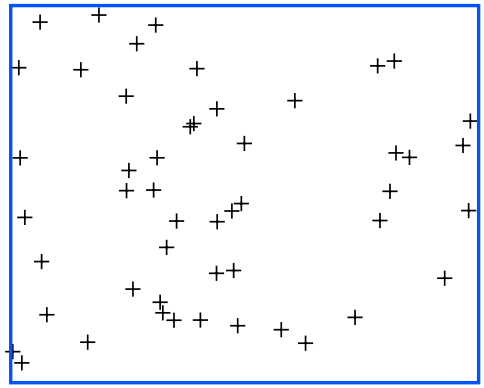


[Dash et al. Dimensionality Reduction for Unsupervised Data. 1997]

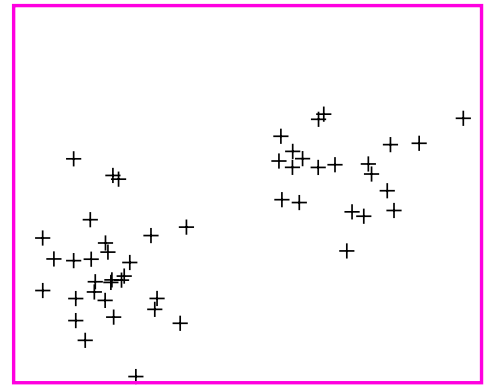
Clusterability III: Dash et al. score



spaced

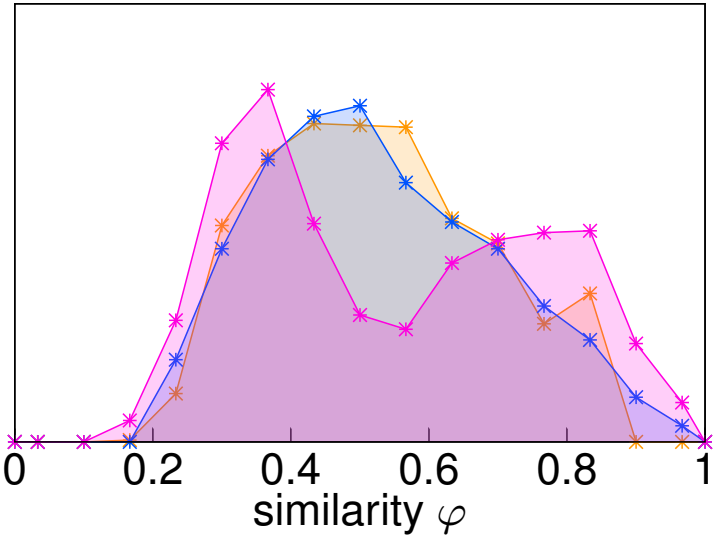


uniform

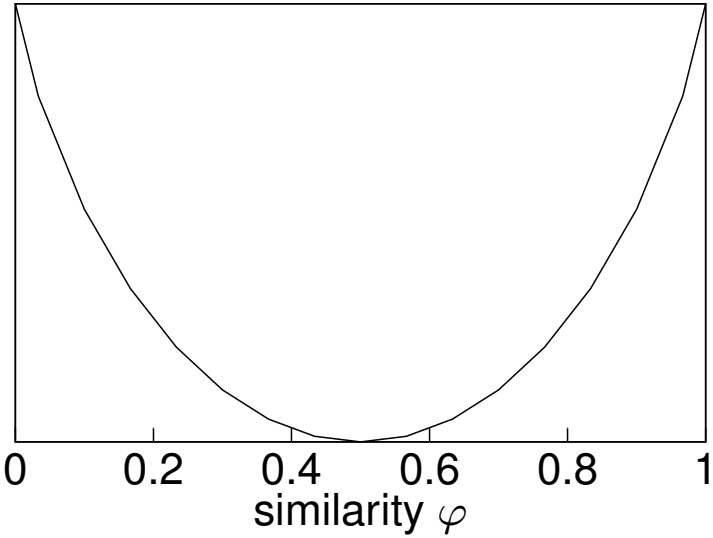


clustered

Similarity-histogram



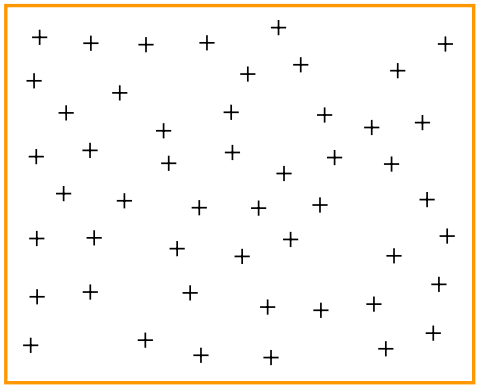
Weighting-function



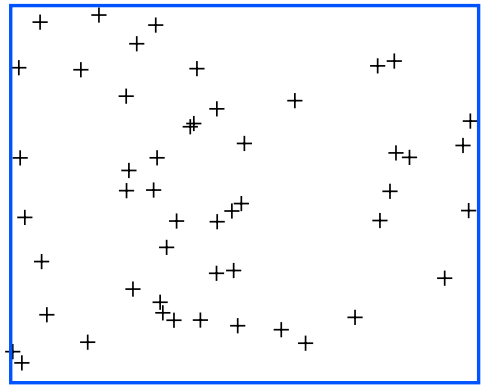
$$1 - (\varphi \cdot \log_2(\varphi) + (1 - \varphi) \cdot \log_2(1 - \varphi))$$

[Dash et al. Dimensionality Reduction for Unsupervised Data. 1997]

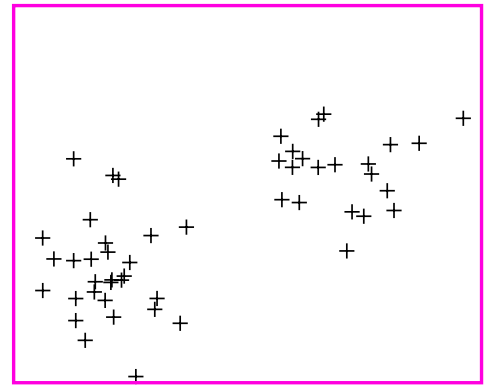
Clusterability III: Dash et al. score



spaced

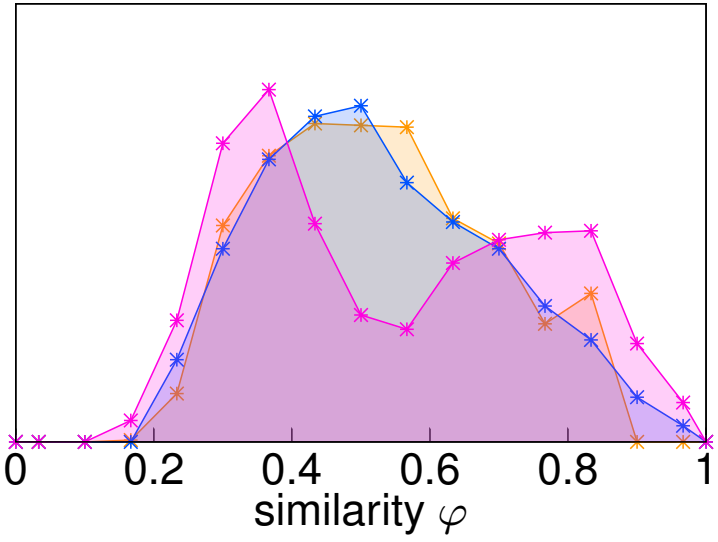


uniform

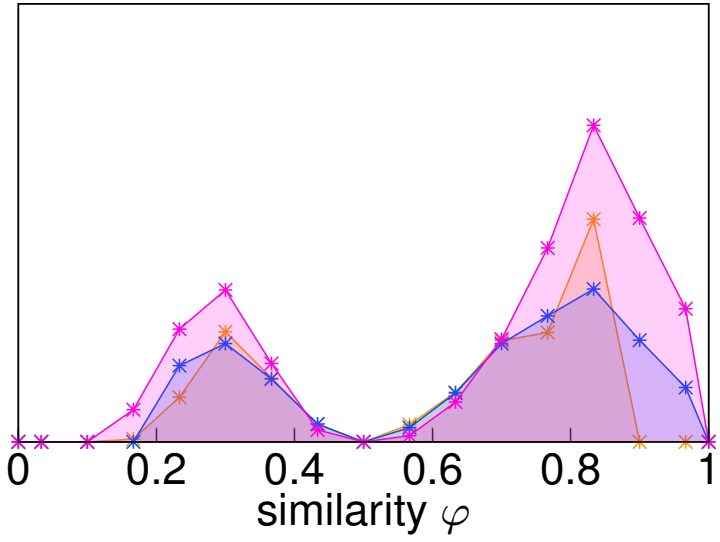


clustered

Similarity-histogram



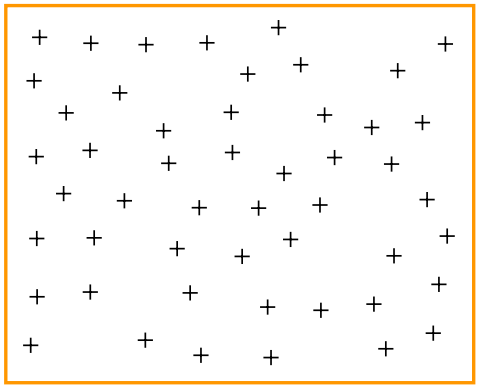
Weighted similarity-histogram



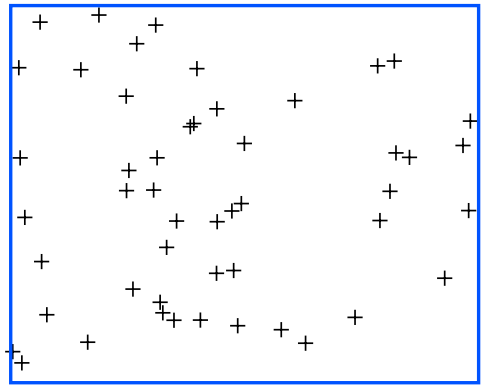
$$1 - (\varphi \cdot \log_2(\varphi) + (1 - \varphi) \cdot \log_2(1 - \varphi))$$

[Dash et al. Dimensionality Reduction for Unsupervised Data. 1997]

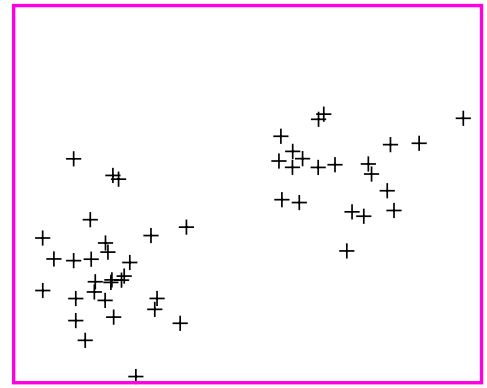
Clusterability III: Dash et al. score



spaced

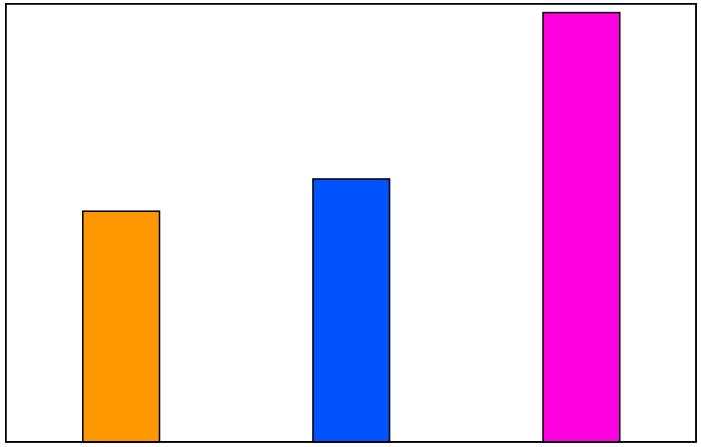


uniform

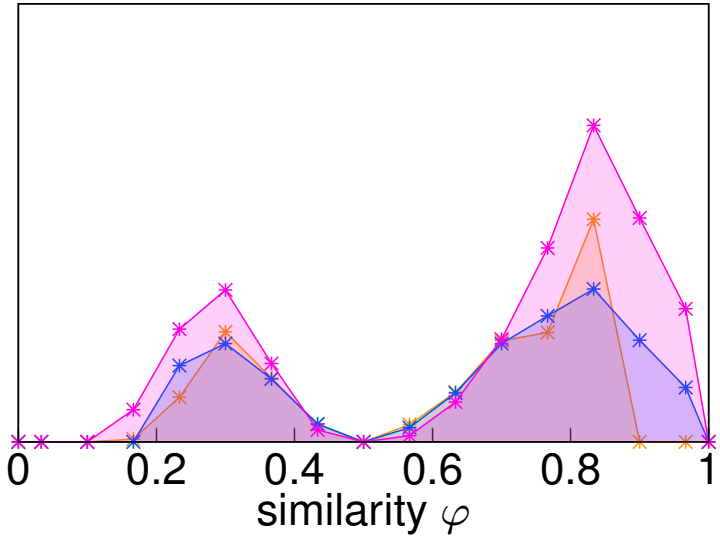


clustered

Clusterability-score



Weighted similarity-histogram



$$1 - (\varphi \cdot \log_2(\varphi) + (1 - \varphi) \cdot \log_2(1 - \varphi))$$

[Dash et al. Dimensionality Reduction for Unsupervised Data. 1997]

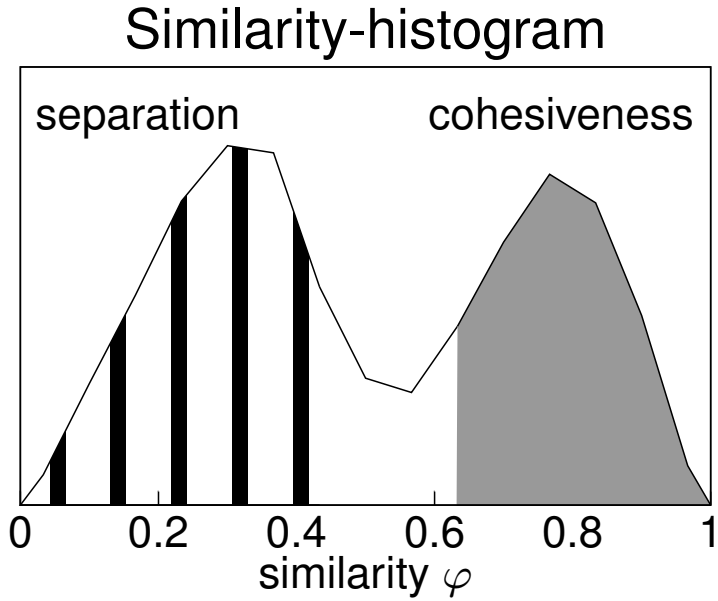
Clusterability III: Concentration of Dissimilarities

\+

+ Very general idea
+ Related to the concept of intrinsic dimensionality

\-

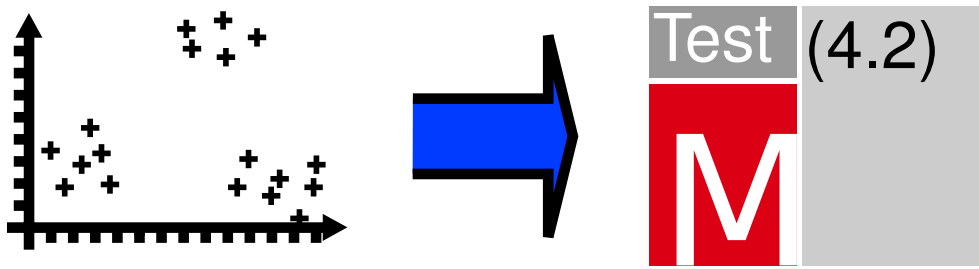
- Not clear when the used heuristic (see right figure) applies
- Lacks the interpretability of statistical tests



[Dash et al. Dimensionality Reduction for Unsupervised Data. 1997]

Clusterability: Overview

- A clusterable model has a dominant structure of mutually **separated** parts that are **cohesive** groups of objects.



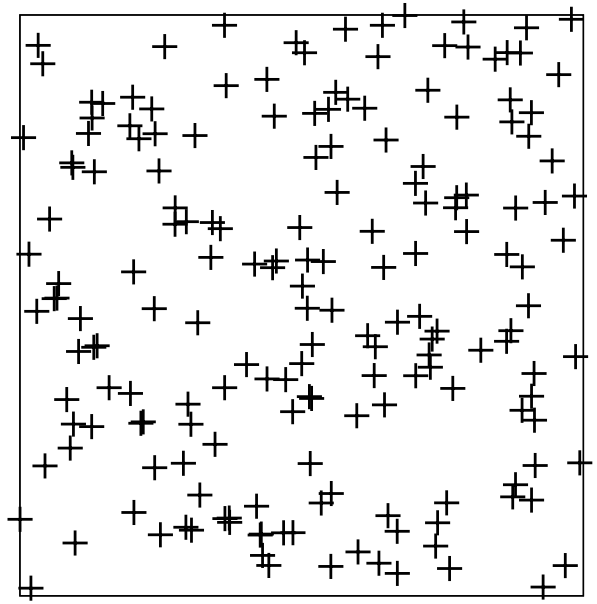
- Clusterability is related to various other topics in data analysis
  - Evaluation indices (Dunn)
  - Tests on model distributions (Hopkins and Skellam)
  - Methods of unsupervised feature selection (Dash et al.)
  - Estimators of intrinsic dimensionality
  - . . . ?

Experiment: Synthetic Models

Can the clusterability indices identify clusterable models?

Experiment setup:

- 10 model distributions of varying intuitive clusterability

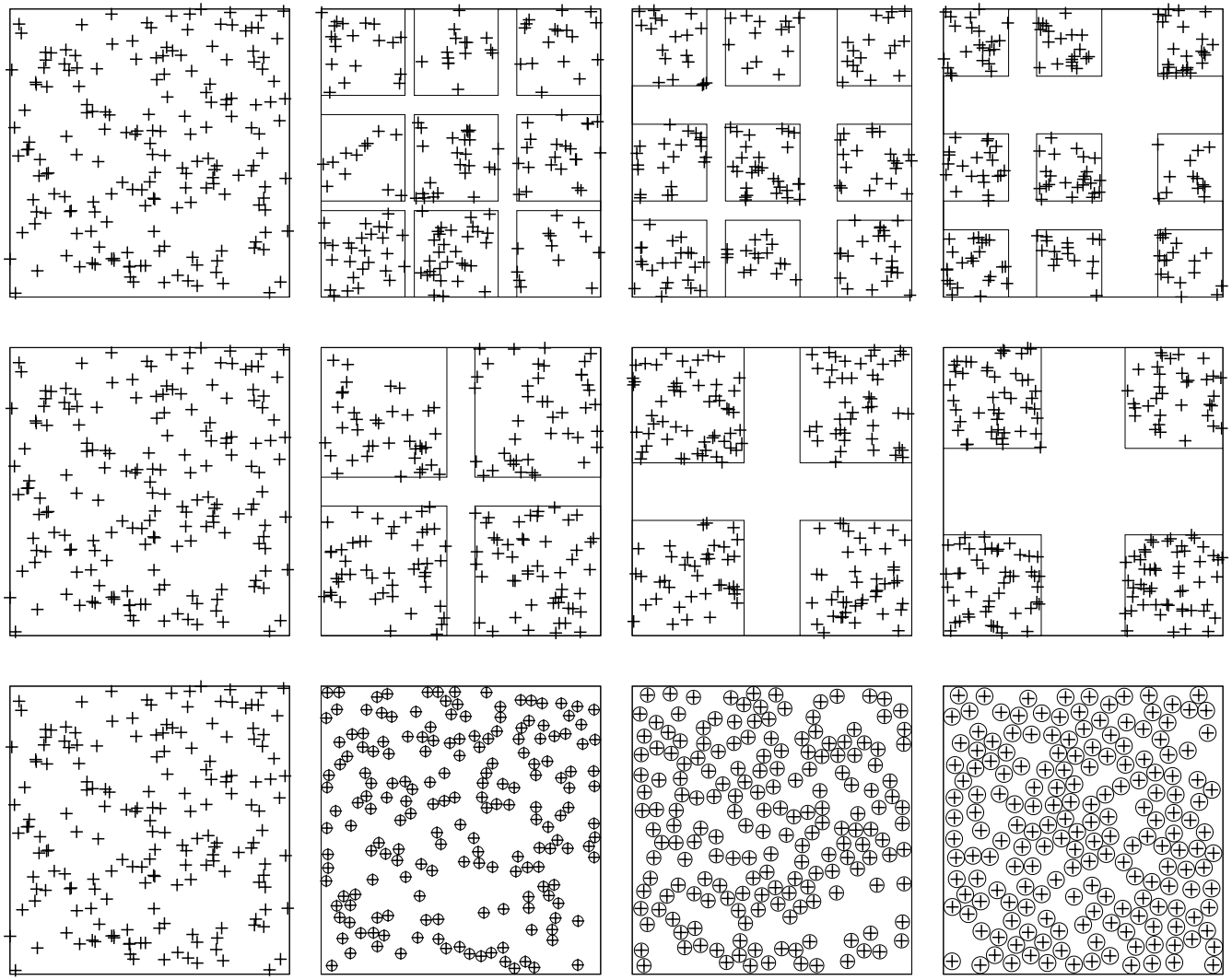


1 model from the uniform distribution

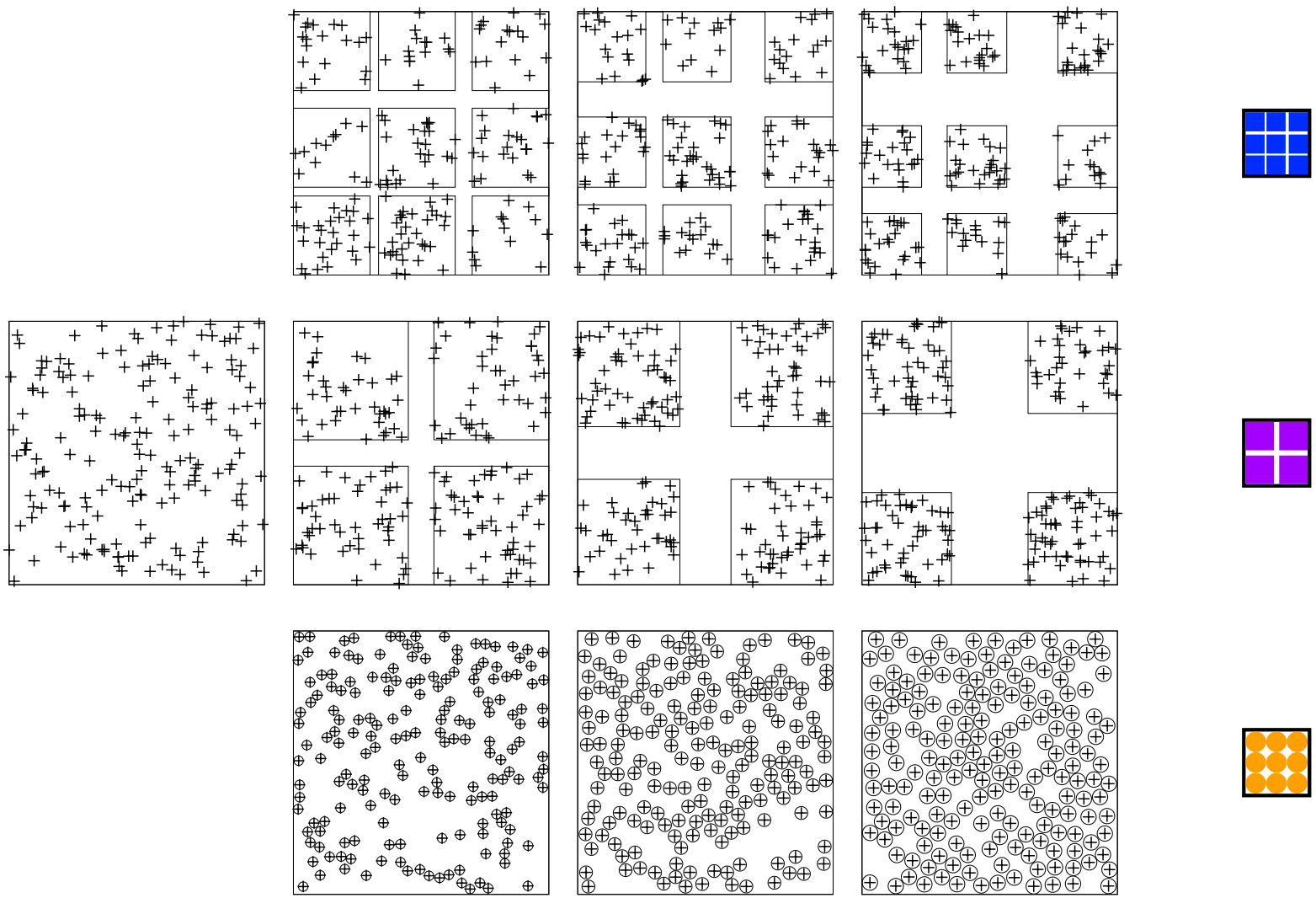
- 1 000 models per distribution (results are means)
- 180 2-dimensional objects per model

Experiment: Synthetic Models

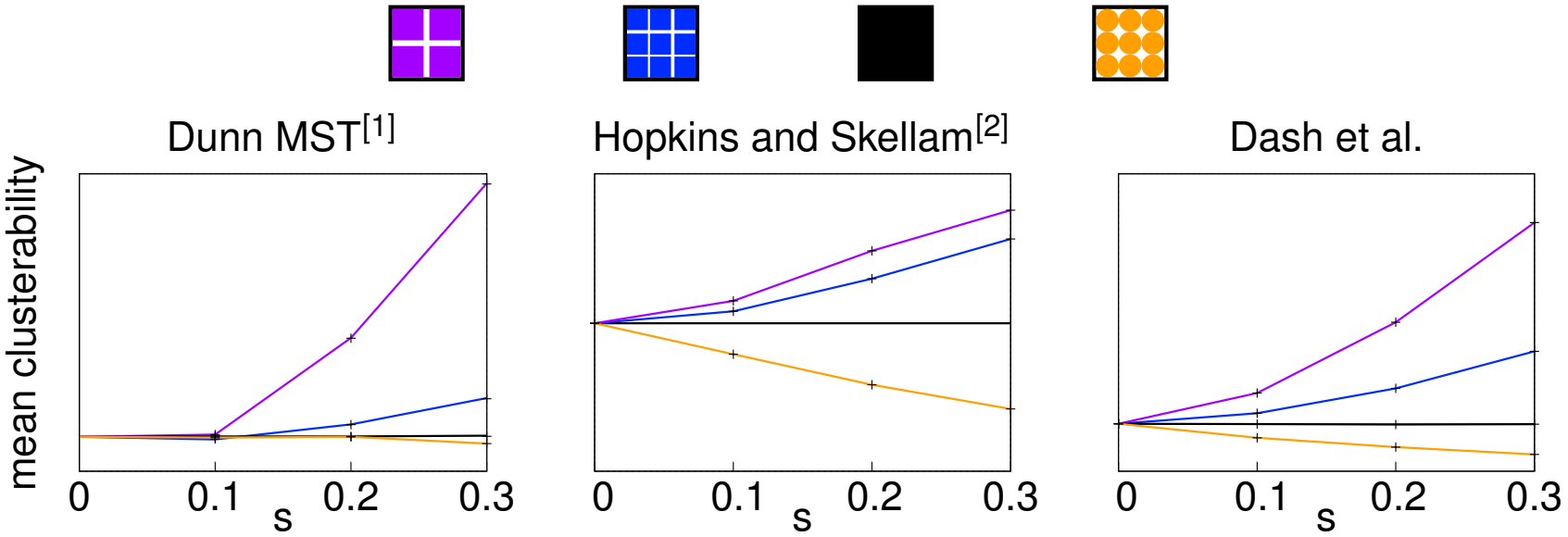
$s = 0$	$s = 0.1$	$s = 0.2$	$s = 0.3$



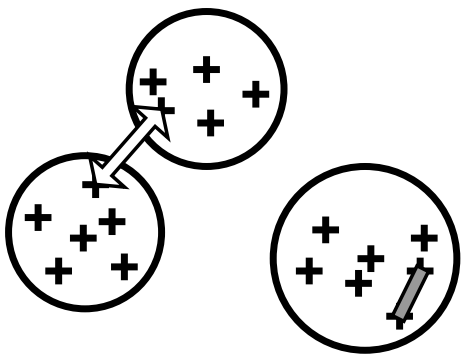
Experiment: Synthetic Models

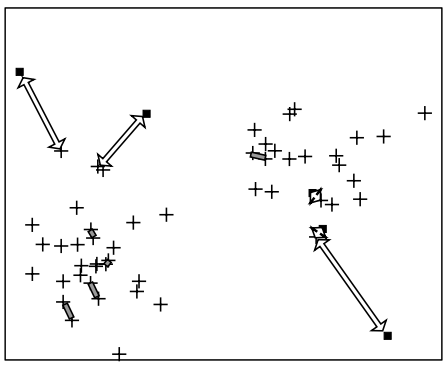


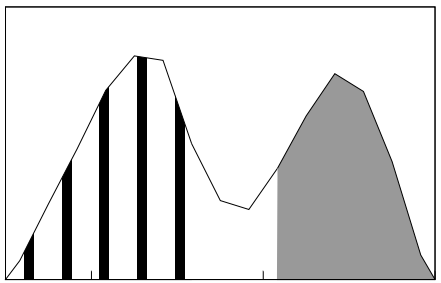
Experiment: Synthetic Models



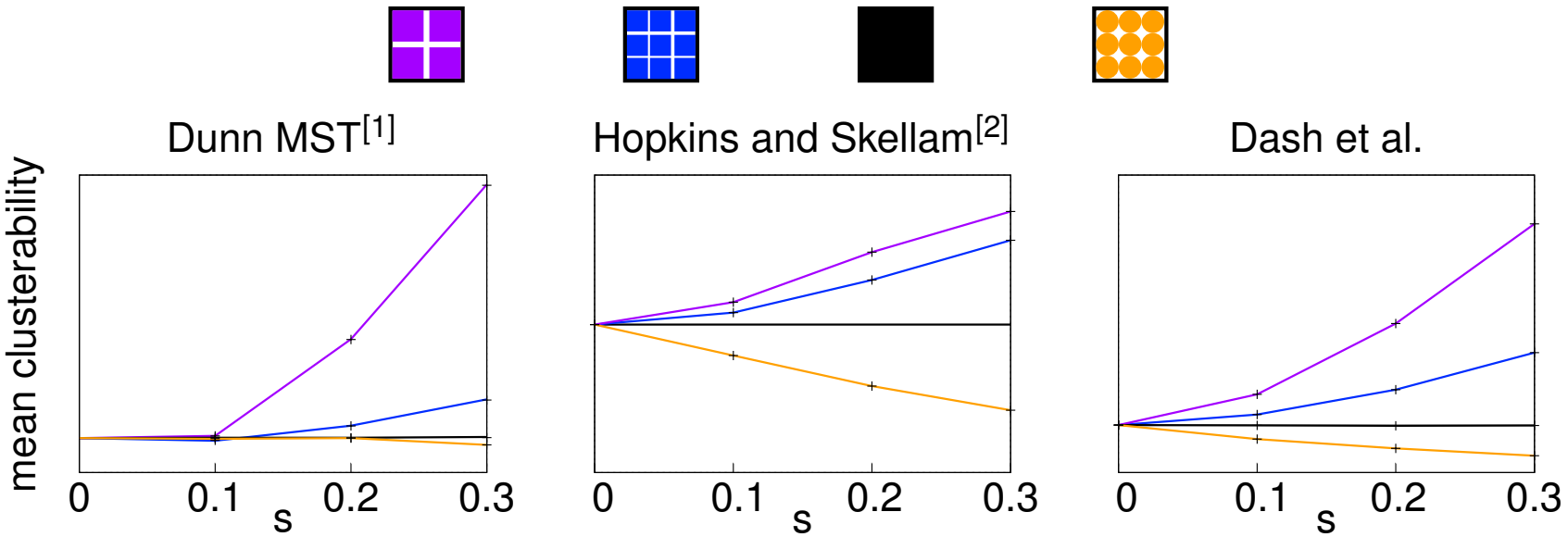
[1] Limited to clusterings with 13 or less clusters
[2] Mean of 1 000 applications per model



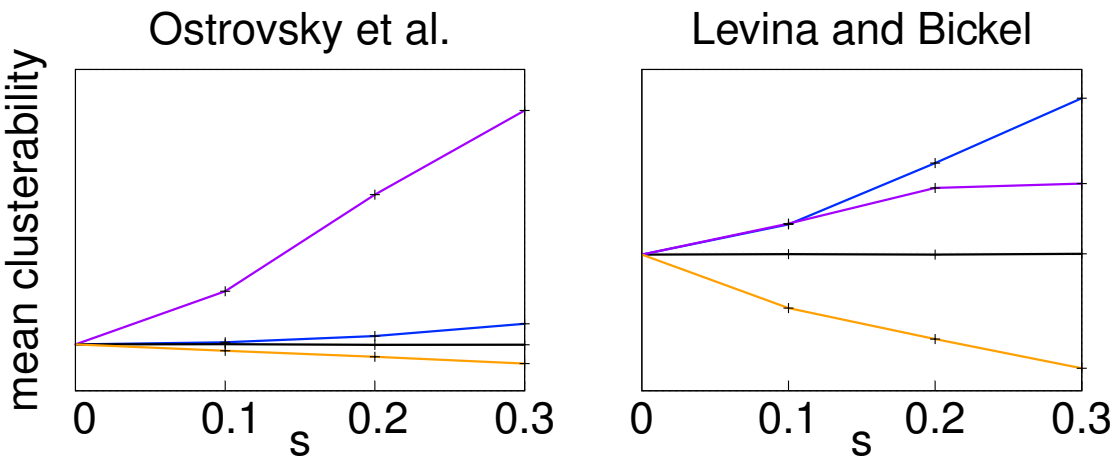




Experiment: Synthetic Models



[1] Limited to clusterings with 13 or less clusters
[2] Mean of 1 000 applications per model



Contributions

A clusterable model has a dominant structure of mutually **separated** parts that are **cohesive** groups of objects.

- Clusterability indices can be used for model selection
- The indices differ, among others, with respect to their preference for fine or coarse structure
- If models are (somewhat) meaningful for a dataset, the more clusterable models are assumed to be also the more meaningful
- Clusterability can incorporate ideas from various related topics (especially clustering evaluation)
- Formal properties of clustering evaluation indices can be converted to properties of clusterability indices

Future Work

- Further formalization of clusterability indices
- Application to large datasets
- Application to high-dimensional problems
- Relation to cluster stability
- Incorporation of additional knowledge (constraint clustering)

Thank you
for your attention.