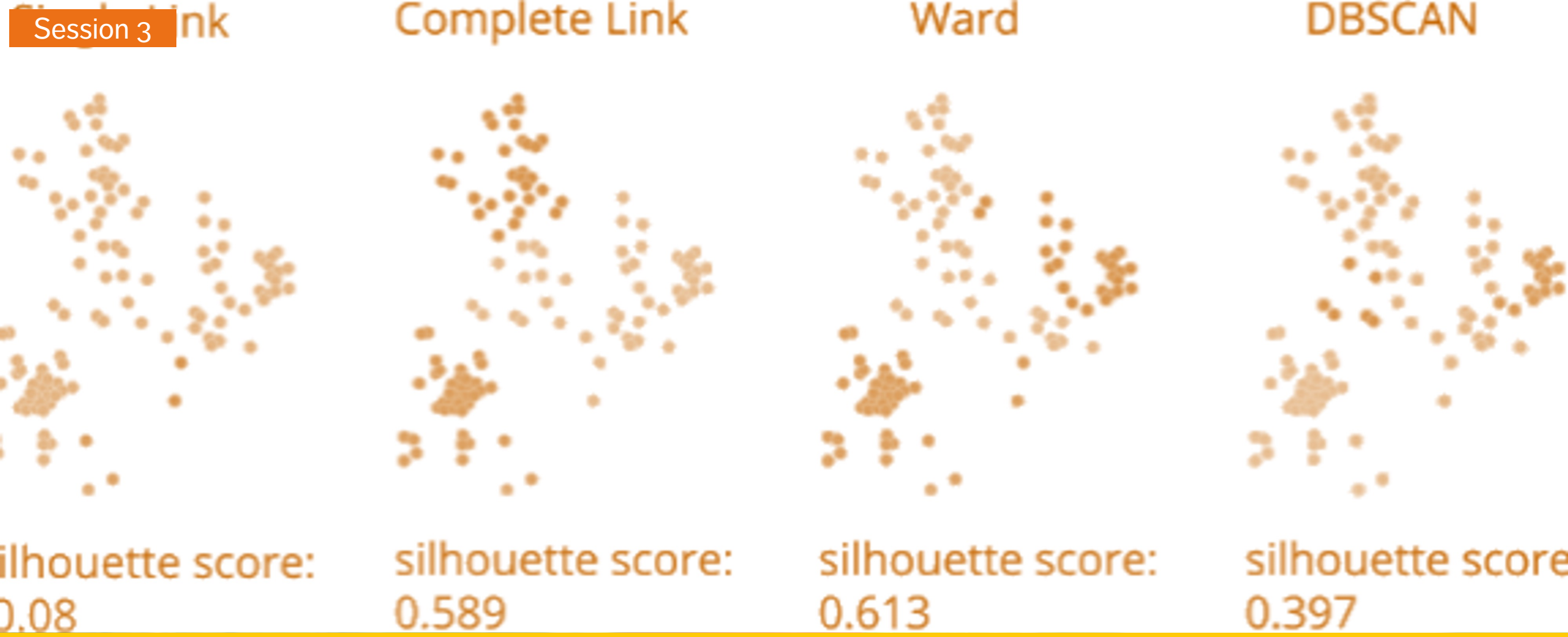


9. Validation and Notes

Clustering Validation

What does it mean for a clustering scheme to be **better** than another?

What does it mean for a clustering scheme to be **valid**?

What does it mean for a single cluster to be **good**?

How many clusters are there in the data, really?

Right vs. wrong is meaningless: seek **optimal vs. sub-optimal**.

Clustering Validation

Optimal clustering scheme:

- maximal separation between clusters
- maximal similarity within groups
- agrees with human eye test
- useful at achieving its goals

Validation types

- **external** (uses additional information)
- **internal** (uses only the clustering results)
- **relative** (compares across clustering attempts)

Clustering Validation

Clustering involves two main activities:

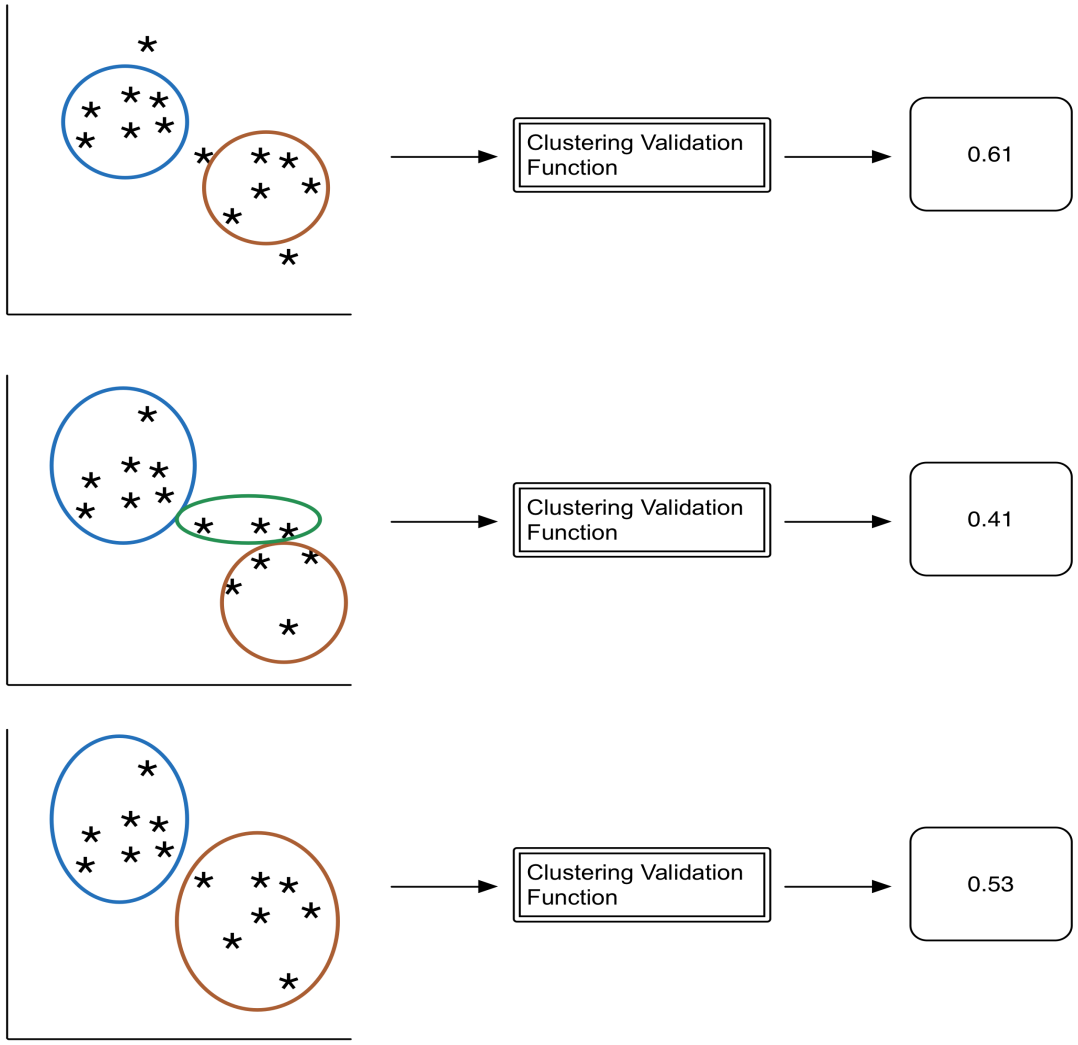
- creating clusters
- **assessing cluster quality**

Clustering functions

- input: instances (vectors)
- output: cluster assignment to each instance

Assessing cluster quality

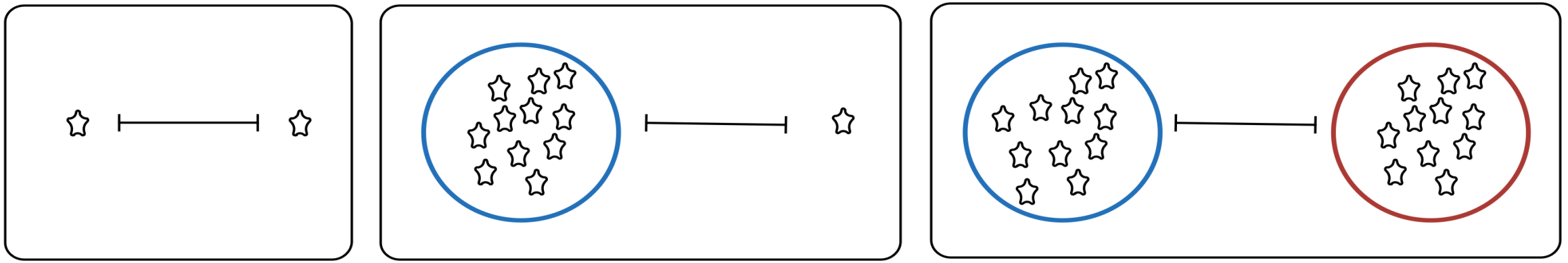
- input: instances + cluster assignments (+ similarity matrix, usually)
- output: a numeric value



Function Components

There are many clustering and cluster validation functions, but they are all built out of basic measures relating to instance or cluster properties:

- **instance properties**
- **cluster properties**
- **instance – instance relationship properties**
- **cluster – instance relationship properties**
- **cluster – cluster relationship properties**



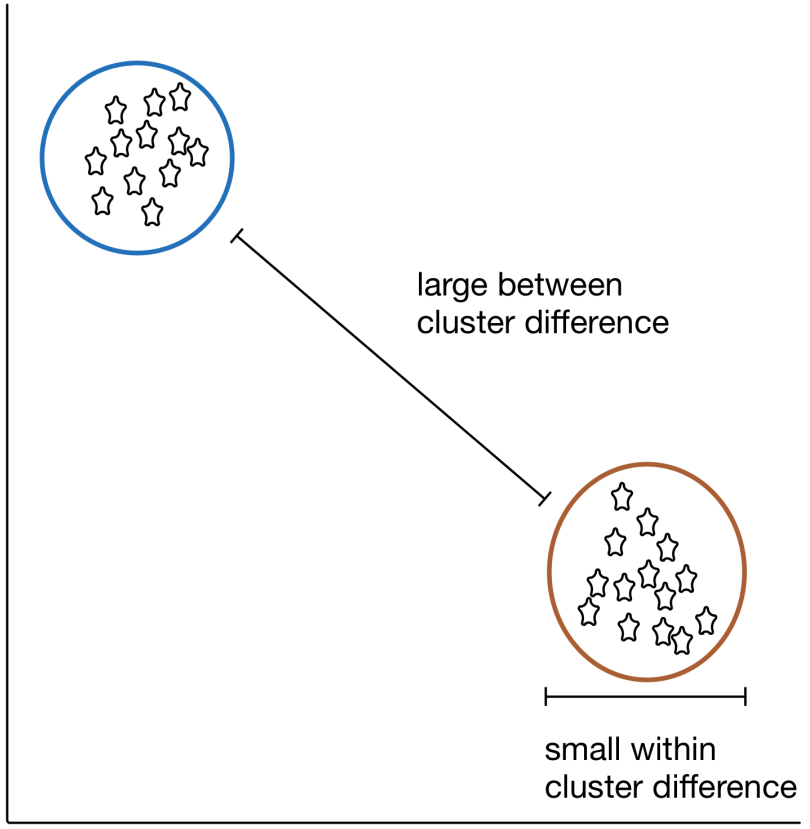
Internal Validation Goals

Within clusters, everything is very similar. Between clusters, there is a lot of difference.

The problem: there are many ways for clusters to deviate from this ideal.

How do we weigh the good aspects (e.g., high **within-cluster similarity**) relative to the bad (e.g., low **between-cluster separation**).

Thus, the large # of **cluster quality metrics** (CQM).



Internal Validation CQM

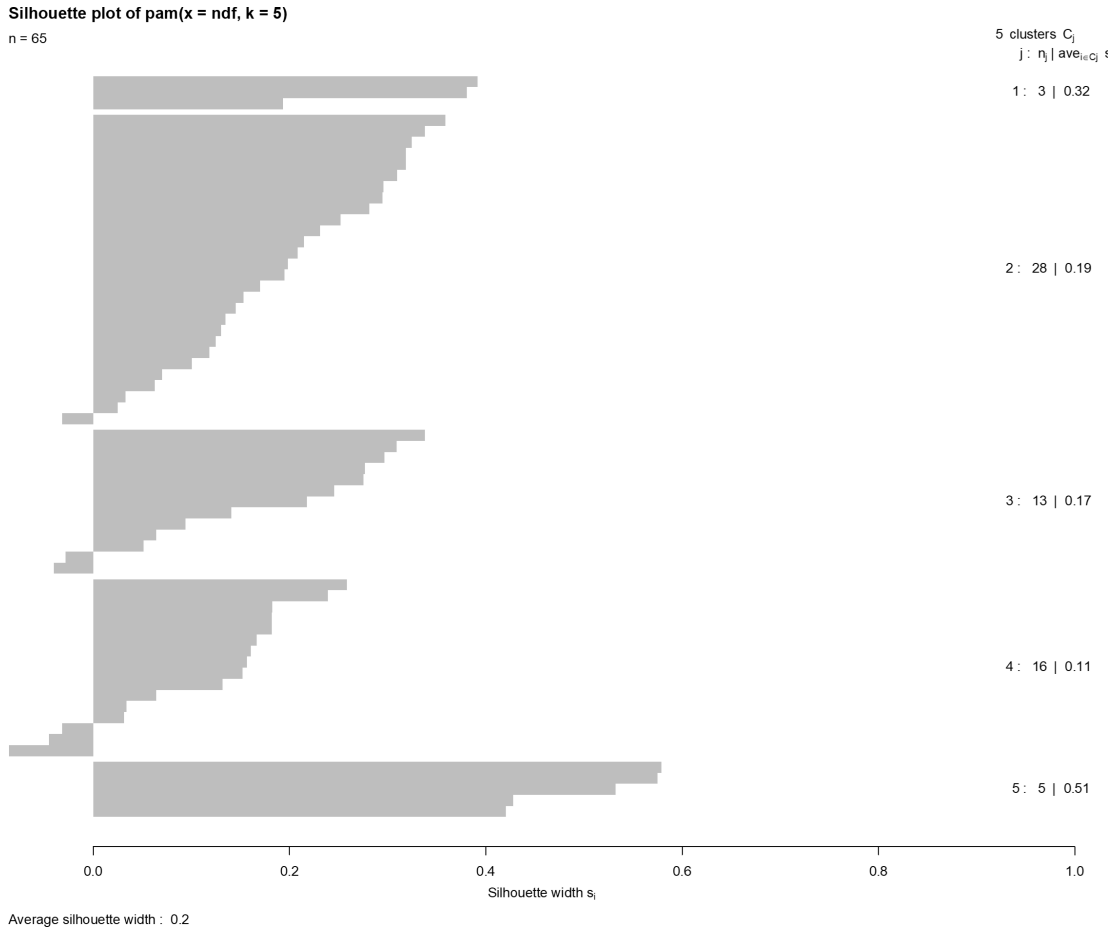
Davies-Bouldin index

Dunn's index

Silhouette metric

Within Sum of Squares

etc. (there are tons!)

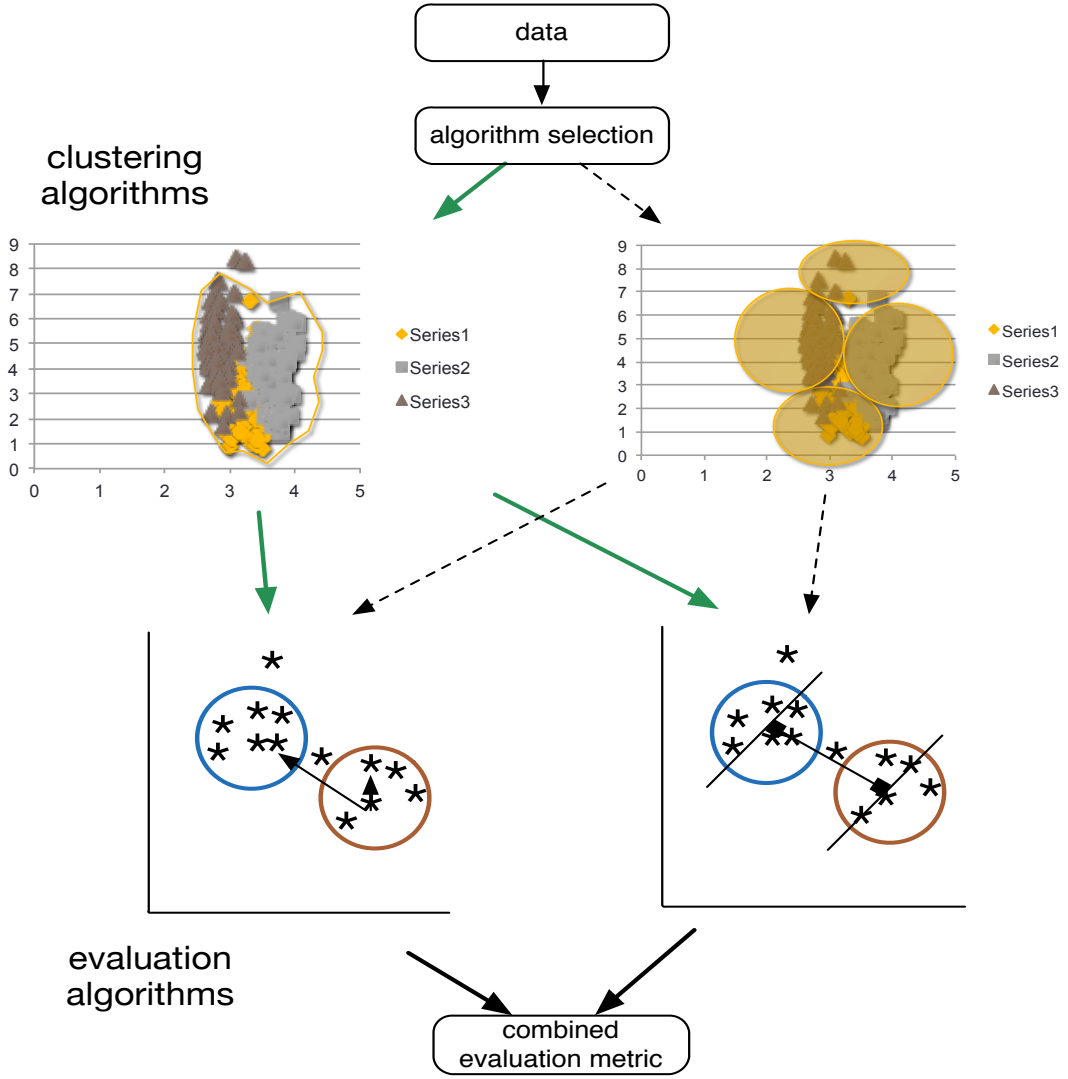


Relative Validation

Getting a single validation measure for a single clustering is not that useful – could the results be better? Is this the best we can hope for?

We could **compare results** across runs or parameter settings.

The main difficulty is to determine how to compare results of **individual runs**.



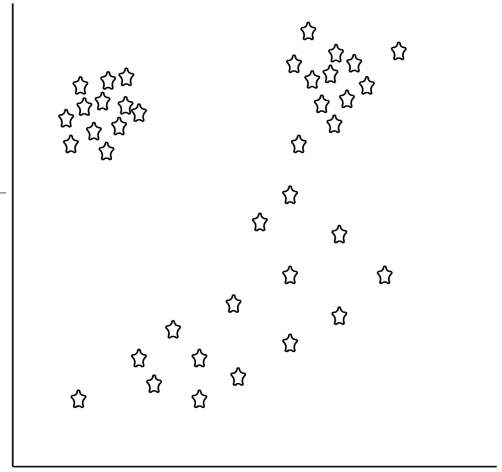
Ensemble Methods

Some options:

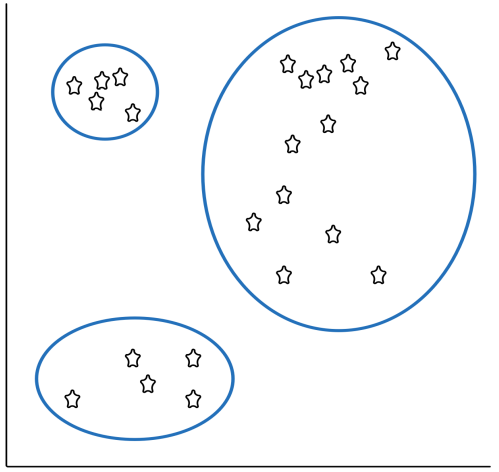
- multiple samples from the same source
- different subsets of columns are used
- different algorithms are used

The **similarity** of the clustering results is measured.

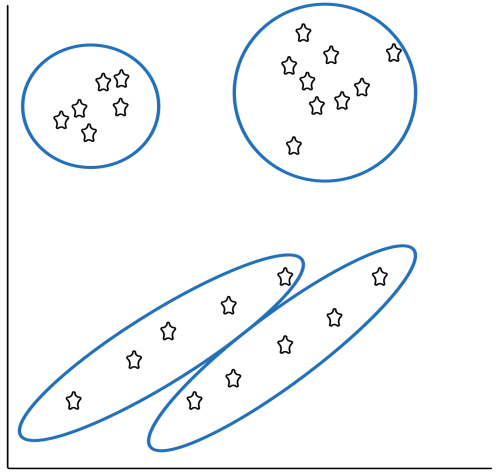
If the results are **not stable** across the clustering outcomes, more investigation is required.



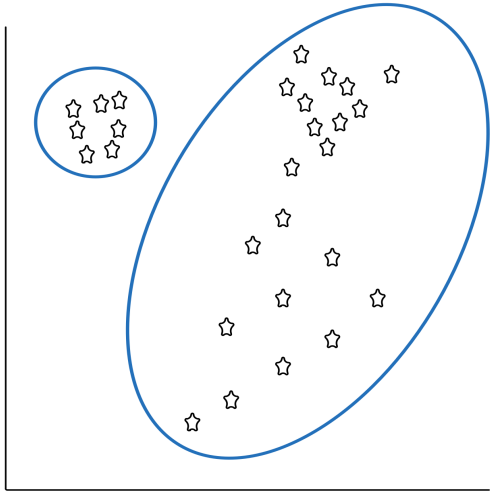
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

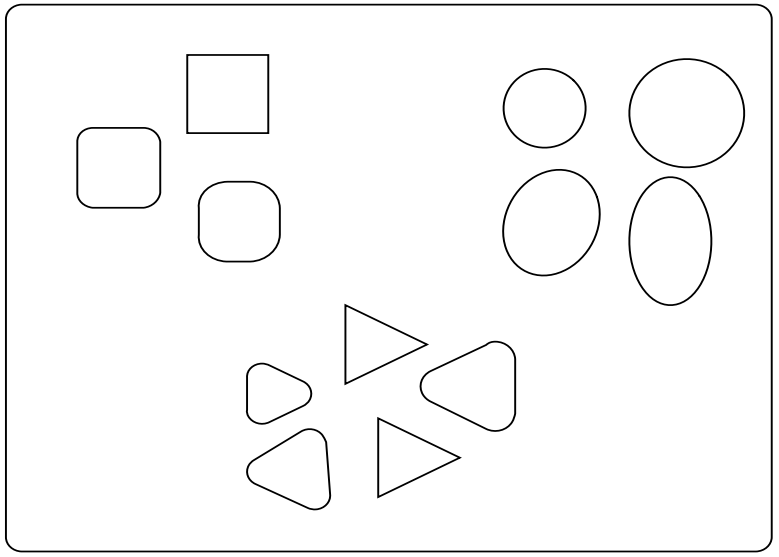
External Validation

Brings in outside info. to **evaluate** the clusters.

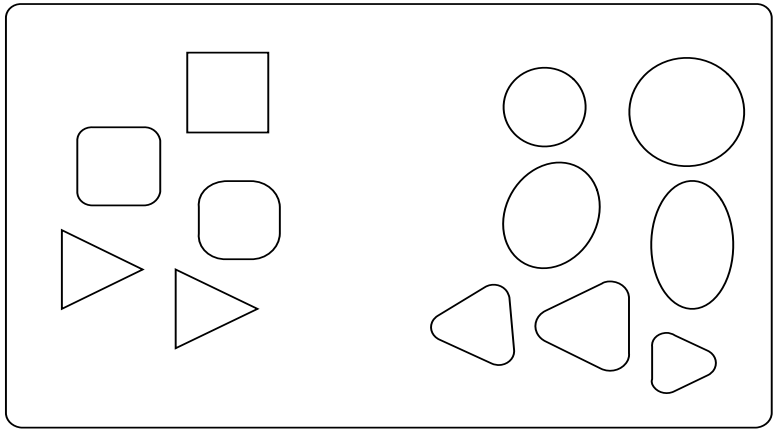
Outside information is typically the 'correct' class.

How is this different from classification then?

Often used to build confidence in the overall approach, based on preliminary or sample results.



Natural Groupings



Clustering Results

Purity

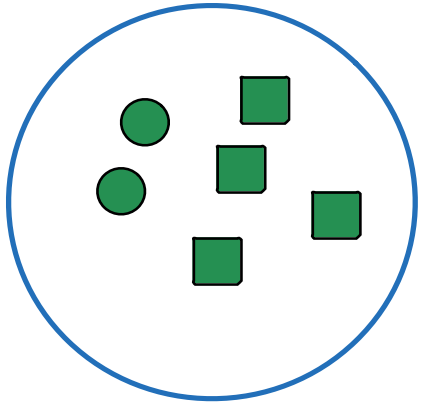
For this external validation metric, each cluster is assigned to the class which is **most frequent** in the cluster.

We calculate the **purity** as follows: number of correctly assigned points / number of points in the cluster.

Some other options: **precision**, **recall**.

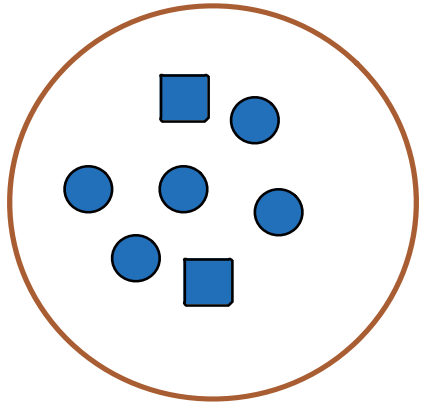
Assuming we are interested in shape…

SQUARE CLUSTER



purity = 66%

CIRCLE CLUSTER



purity = 71%

Clustering Challenges

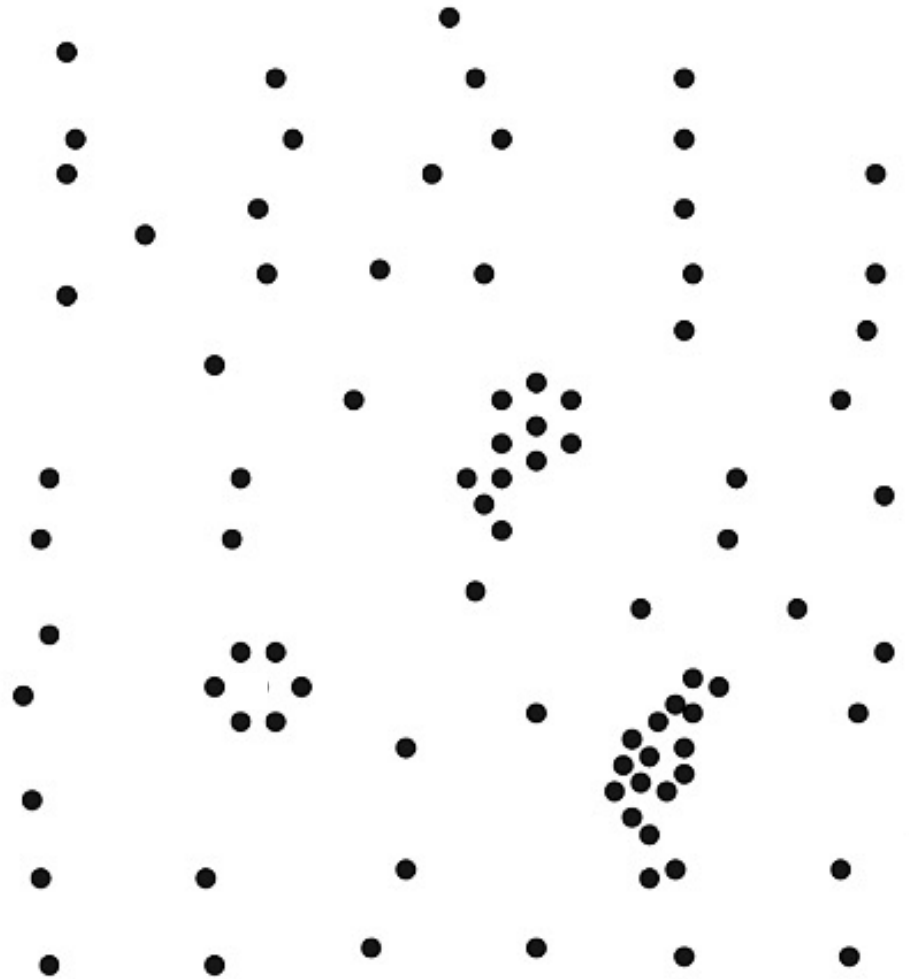
Automation
relatively intuitive for humans, but harder for machines

Lack of a clear-cut definition
no universal agreement as to what constitutes a cluster

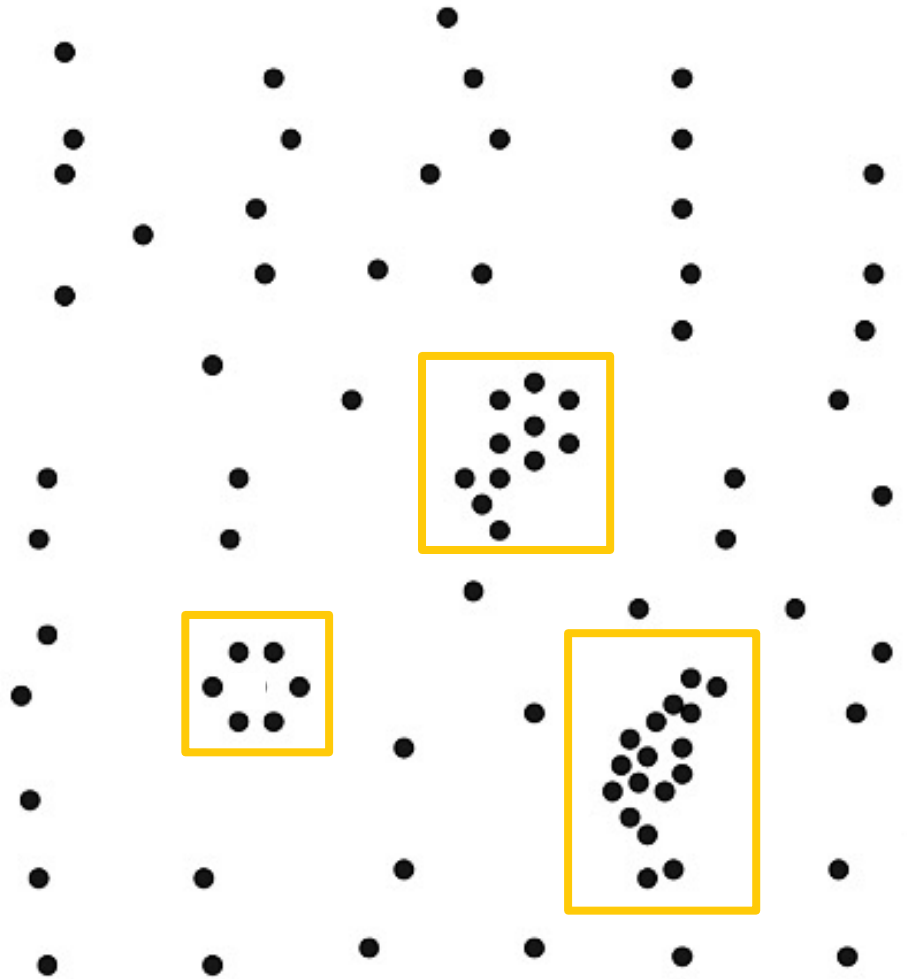
Lack of repeatability
non-deterministic: the same algorithm, applied twice to the same dataset can discover completely different clusters

Number of clusters
optimal number of clusters difficult to determine

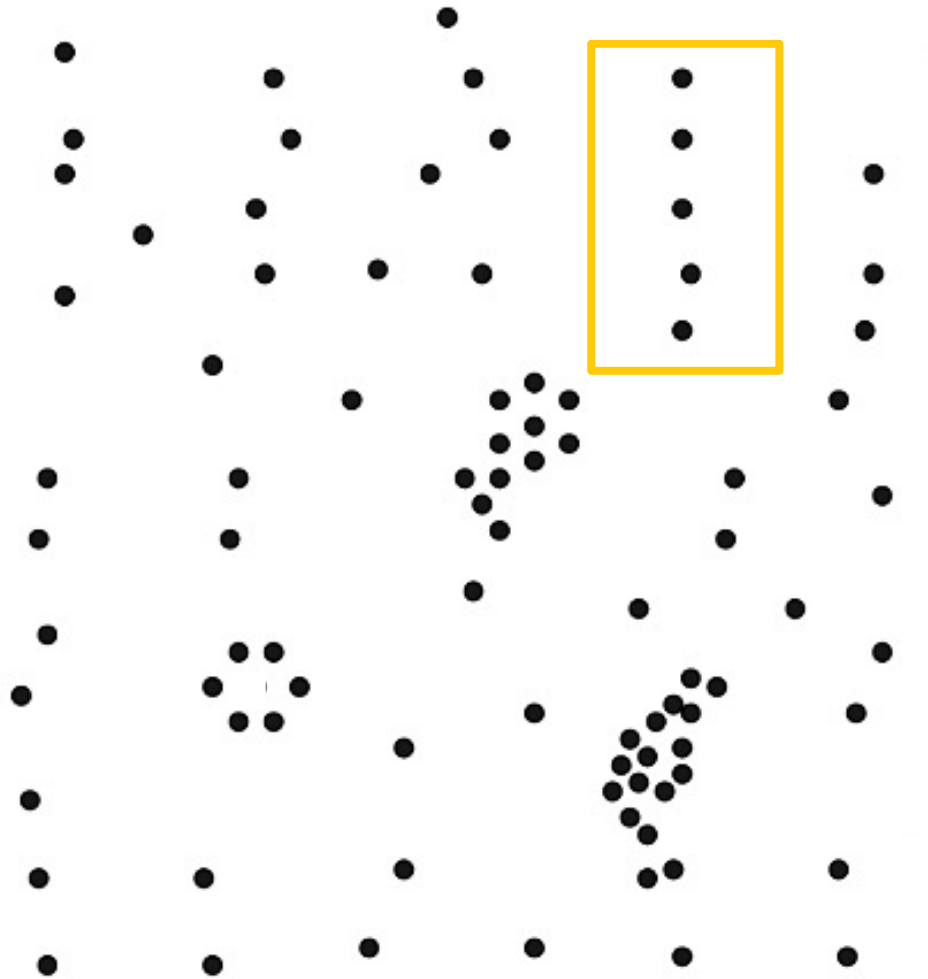
Session 3



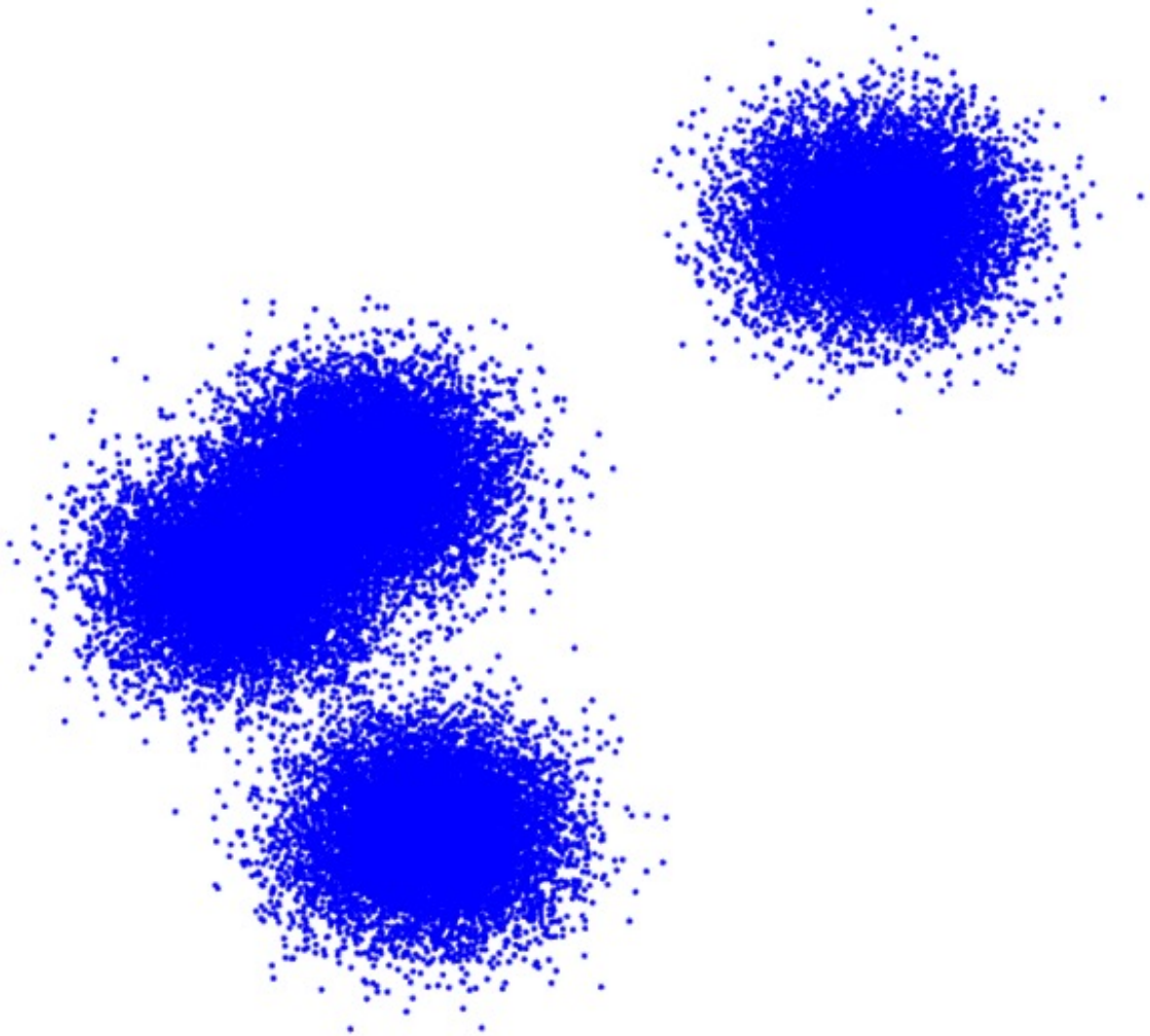
Session 3



Session 3



Session 3



Clustering Challenges

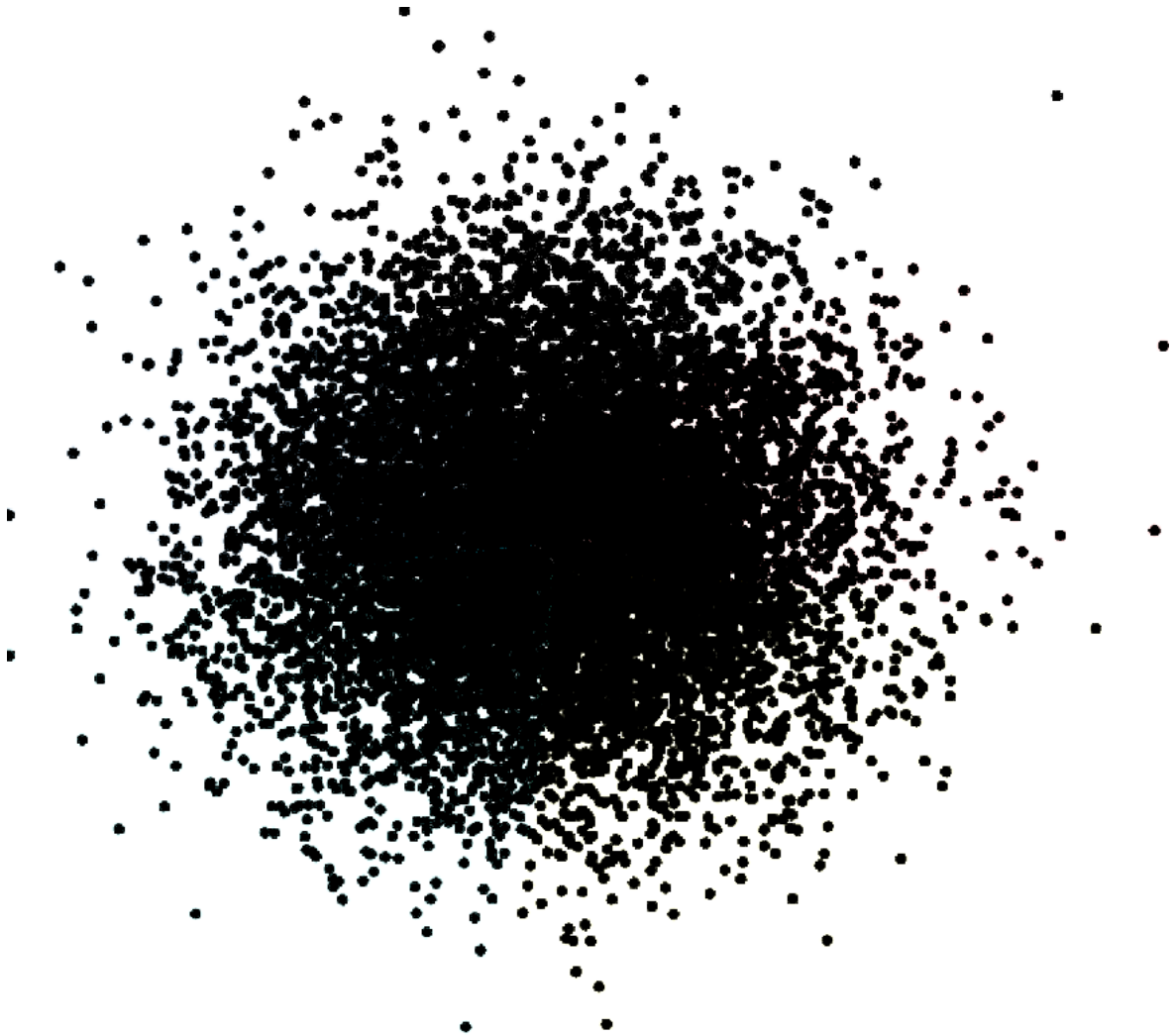
Cluster description
should clusters be described using representative instances or average values?

Model validation
no true clustering information against which to contrast the clustering scheme, so how do we determine if it is appropriate?

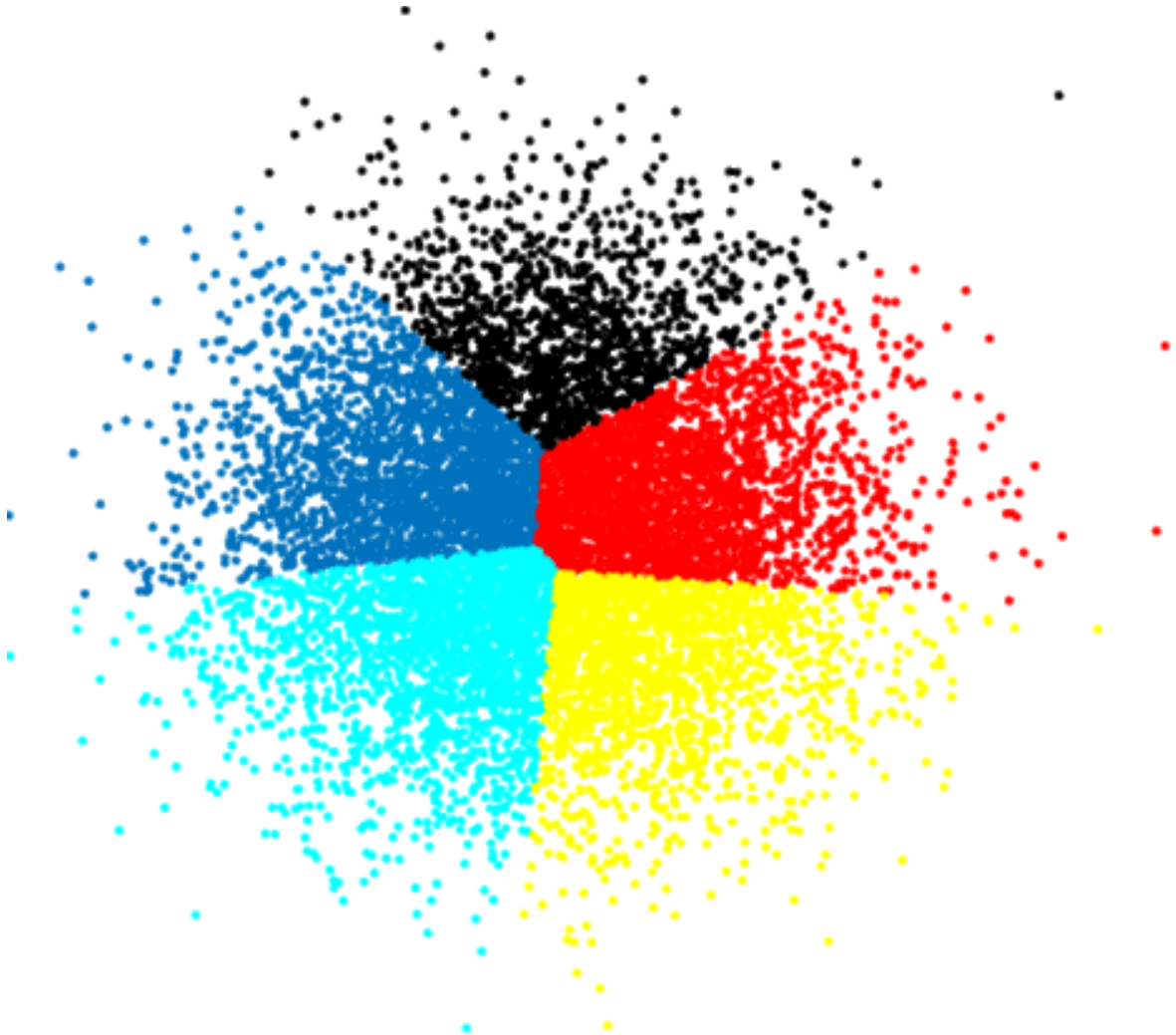
Ghost clustering
most methods will find clusters even if there are none in the data

***A posteriori* rationalization**
once clusters have been found, it is tempting to try to "explain" them ...

Session 3



Session 3



Suggested Reading

Validation and Notes

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Clustering Validation

Spotlight on Clustering

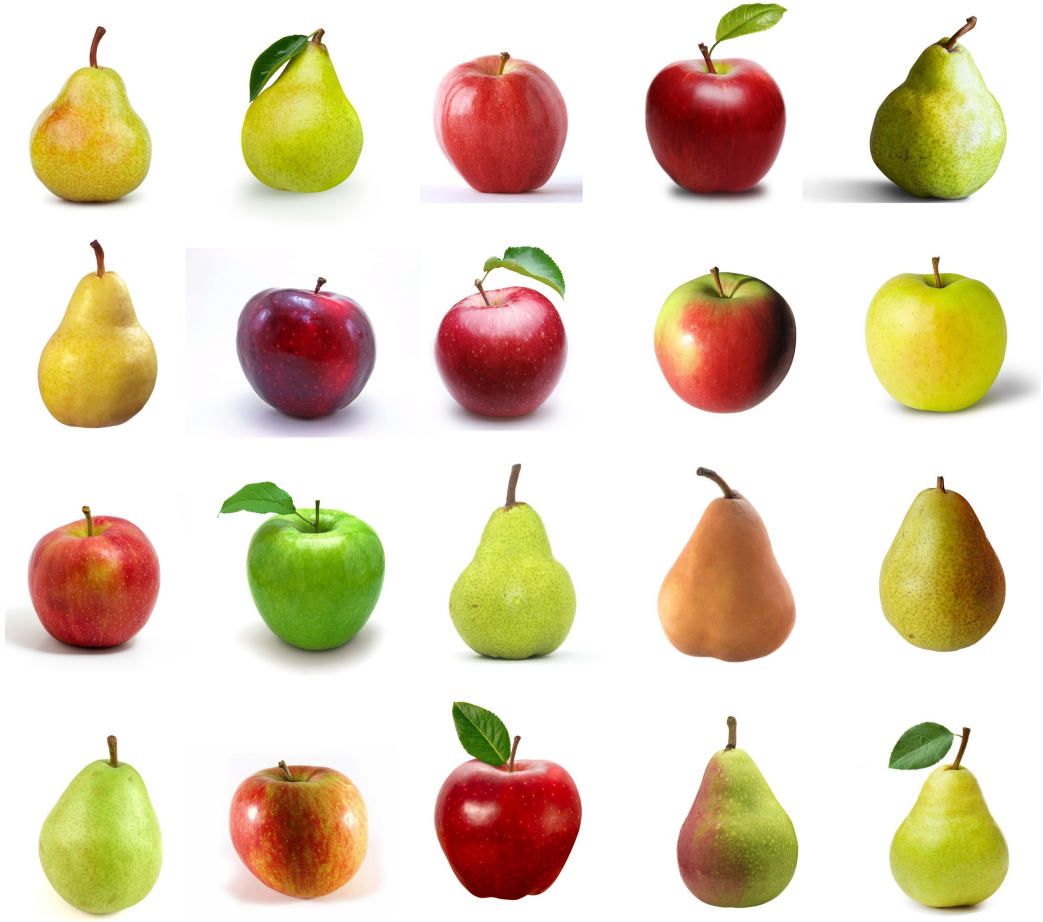
*Clustering Evaluation (advanced)

- Clustering Assessment
- Model Selection

Exercises

Validation and Notes

Consider the fruit image dataset below.



Provide a few clustering schemes for the data, and discuss how you would validate them.