Clustering

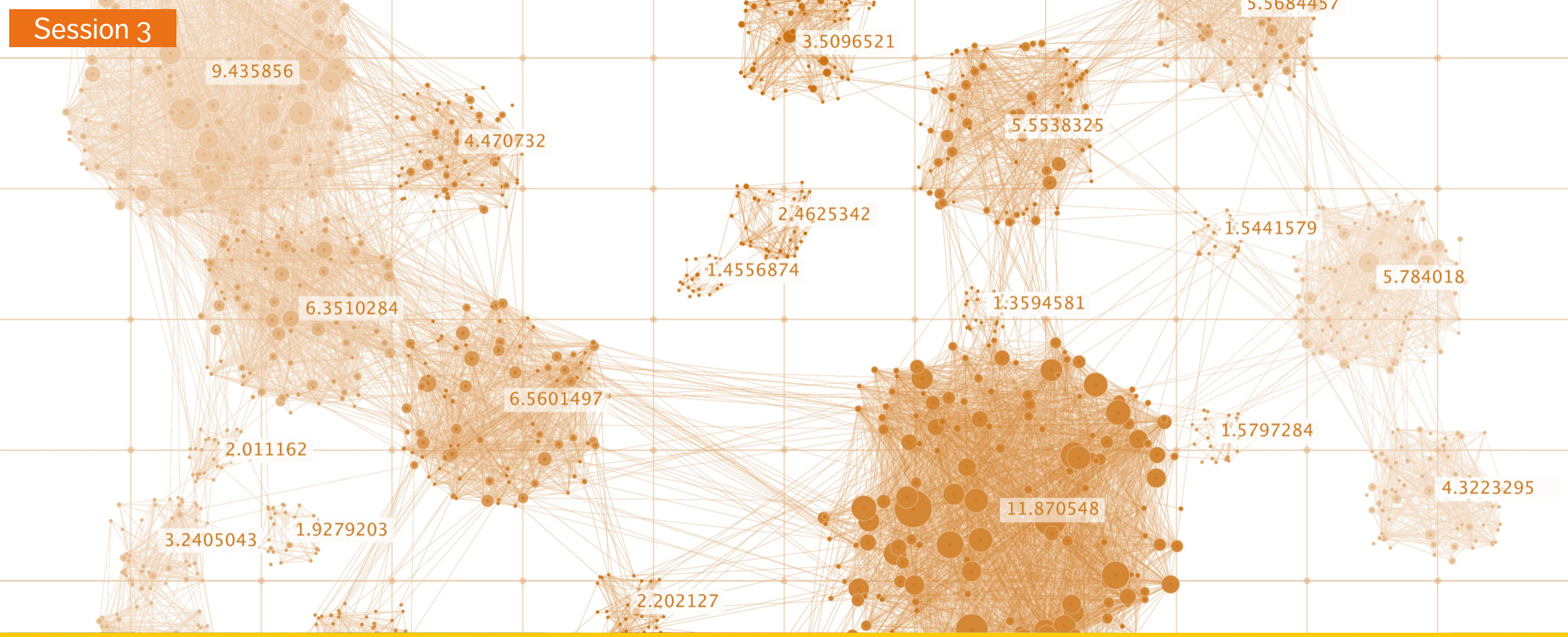
INTRODUCTION TO MACHINE LEARNING

Clustering is in the eye of the beholder, and as such, researchers have proposed many induction principles and models whose corresponding optimisation problem can only be approximately solved by an even larger number of algorithms.

[V. Estivill-Castro, *Why So Many Clustering Algorithms?*]

Woes clusters. Rare are solitary woes; they love a train, they tread each other's heel.

[E. Young]



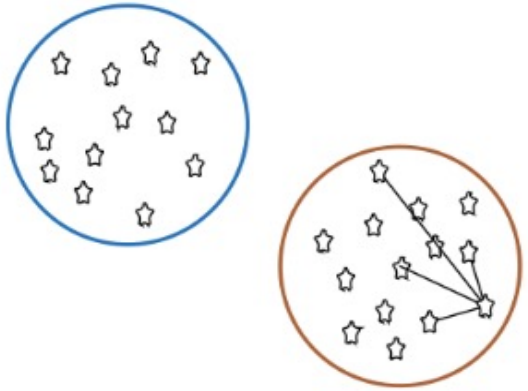
7. Clustering Overview

Overview

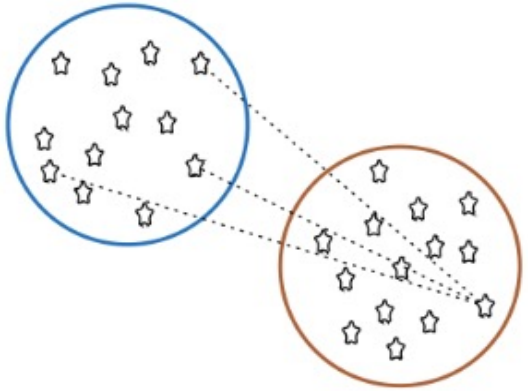
In **clustering**, the data is divided into **naturally occurring groups**. Within each group, the data points are **similar**; from group to group, they are **dissimilar**.

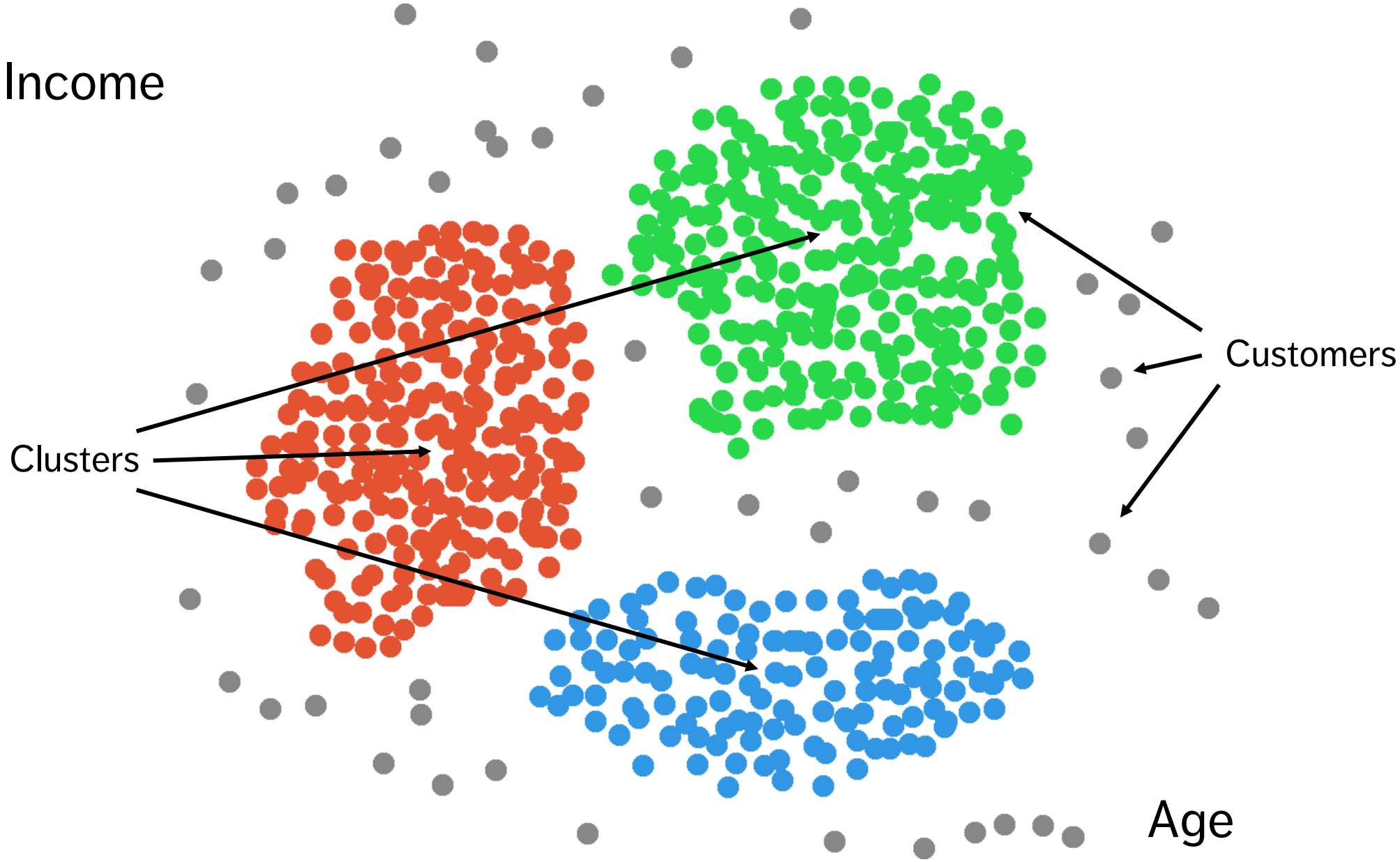
The grouping labels are not determined ahead of time, so clustering is an example of **unsupervised** learning.

average distance to points in own cluster (**low is good**)



average distance to points in neighbouring cluster (**high is good**)




Income
Customers
Clusters
Age

Overview

Clustering algorithms can be **complex** and **non-intuitive**, based on varying notions of similarities between observations.

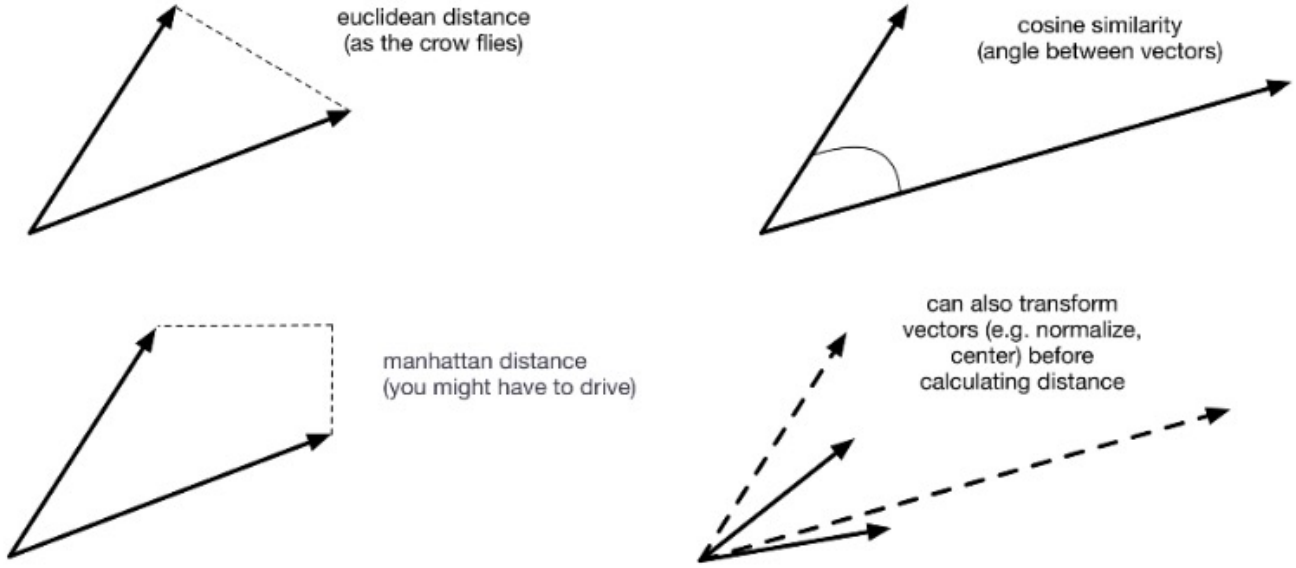
- in spite of that, the temptation to explain clusters *a posteriori* is **strong**

They are also (typically) **non-deterministic:**

- the same algorithm, applied twice (or more) to the same dataset, can discover completely different clusters
- the order in which the data is presented can play a role
- so can starting configurations

Clustering Requirement

A measure of **similarity** w (or a distance d) between observations:



IMPORTANT: data must be scaled before it is fed into clustering algorithms.

Typically, $w \to 1$ as $d \to 0$, and $w \to 0$ as $d \to \infty$.

Distance Measures (Metrics)

Categorical Variables*

- Hamming distance
- Russel/Rao index
- Jaccard
- Dice's coefficient
- etc.

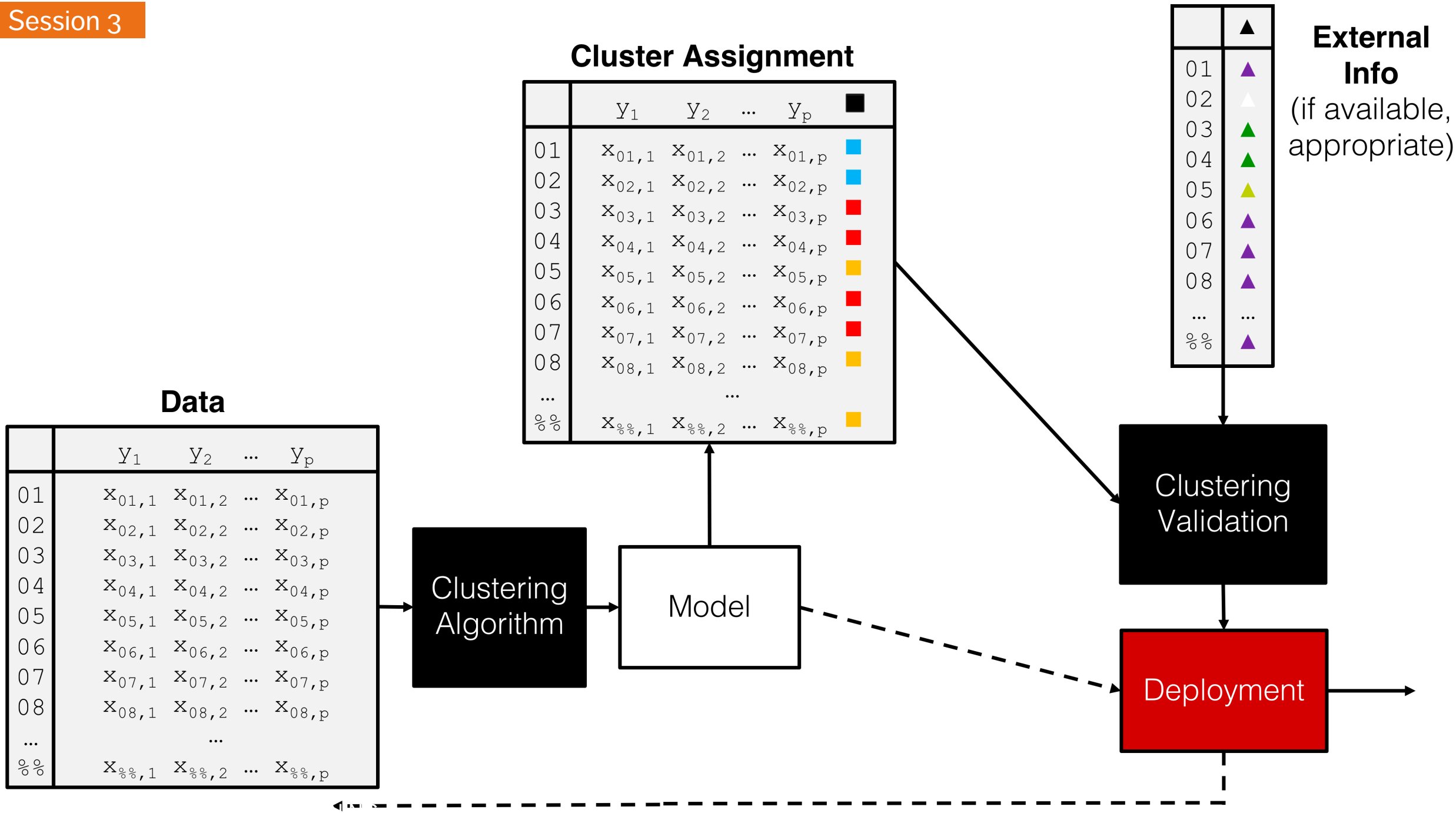
No steadfast rule to determine which distance to use; competing schemes are often produced with diff. metrics.

Numerical Variables

- Euclidean
- Manhattan
- correlation
- cosine
- etc.

We may need to create hybrid metrics for dataset with both categorical and numerical variables.

Session 3


Cluster Assignment
External Info (if available, appropriate)
Data
Clustering Algorithm
Model
Clustering Validation
Deployment

Applications

Text Documents

- grouping similar documents according to their topics, based on the patterns of common and unusual words

Product Recommendations

- grouping online purchasers based on the products they have viewed, purchased, liked, or disliked
- grouping products based on customer reviews

Marketing and Business

- grouping client profiles based on their demographics and preferences

Applications

Dividing a larger group (or area, or category) into **smaller** groups, with members of the smaller groups guaranteed to have similarities of some kind.

- tasks may then be solved separately for each of the smaller groups
- this may lead to increased accuracy once the separate results are aggregated

Creating taxonomies **on the fly**, as new items are added to a group of items

- this would allow for easier product navigation on a site like Netflix, for instance

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Objective

When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city?

The researchers aims to draw the boundaries of **livehoods**, areas of similar character within a city, by using clustering models. Unlike **static** administrative neighborhoods, the livehoods are defined based on the **habits** of their inhabitants.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Methodology

The authors use **spectral clustering** to discover **distinct geographic areas** of the city based on collective **movement patterns**.

Livehood clusters are built as follows:

1. a **geographic distance** is computed based on pairs of check-in venues' coordinates;
2. a **social similarity** is computed between each pair of **venues** using cosine measurements;
3. spectral clustering produces **candidate livehoods**;
4. interviews are conducted with residents in order to **explore**, **label**, and **validate** the clusters discovered by the algorithm.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Data

The data comes from two sources, combining approximately 11 million check-ins from the dataset of Chen et al. (a recommendation site for venues based on users' experiences) and a new dataset of 7 million Twitter check-ins downloaded between June and December of 2011.

For each check-in, the data consists of the **user ID**, the **time**, the **latitude and longitude**, the **name of the venue**, and its **category**.

In this case study, data from the city of Pittsburgh, Pennsylvania, is examined *via* 42,787 check-ins of 3840 users at 5349 venues.

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Strengths and Limitations of the Approach

- The technique used in this study is **agnostic** towards the particular source of the data: it is not dependent on meta-knowledge about the data.
- The algorithm may be prone to "majority" bias, possibly misrepresenting/hiding minority behaviours.
- The dataset is built from a **limited** sample of check-ins shared on Twitter and are therefore biased towards the types of visits/locations that people typically want to share **publicly**.
- Tuning the clusters is non-trivial: experimenter bias may combine with "confirmation bias" of the interviewees in the validation stage – if the researchers are residents of Pittsburgh, will they see clusters when there were none?

Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

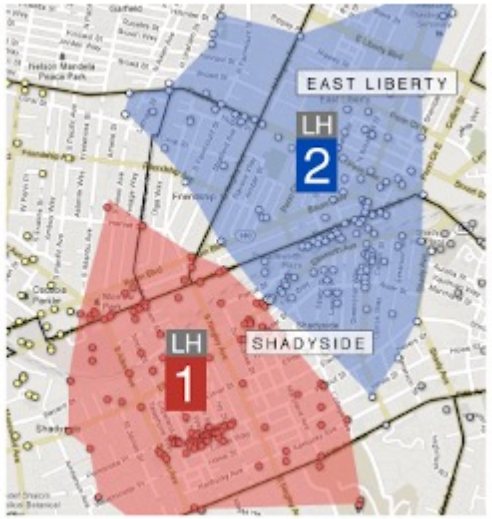
Results, Evaluation, and Validation

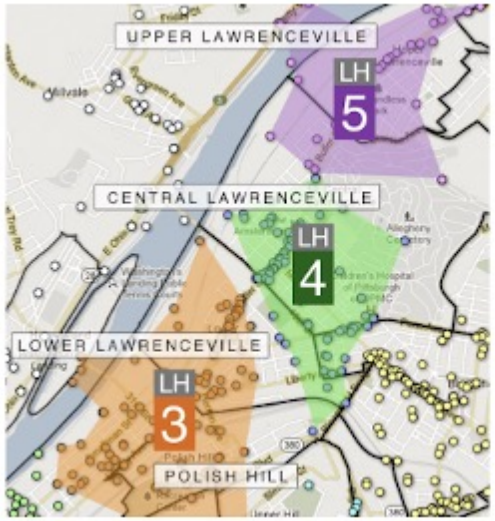
Over 3 areas of the city, 9 livehoods have been identified and validated by 27 Pittsburgh residents

- **Municipal Neighborhoods Borders:** livehoods are dynamic, and evolve as people's behaviours change, unlike fixed neighbourhoods set by the city government.
- **Demographics:** the interviews displayed strong evidence that the demographics of the residents and visitors of an area play a strong role in explaining the livehood divisions.
- **Development and Resources:** economic development can affect the character of an area. Similarly, the resources provided by a region has a strong influence on the people that visit it, and hence its resulting character.

Case Study

Livehoods

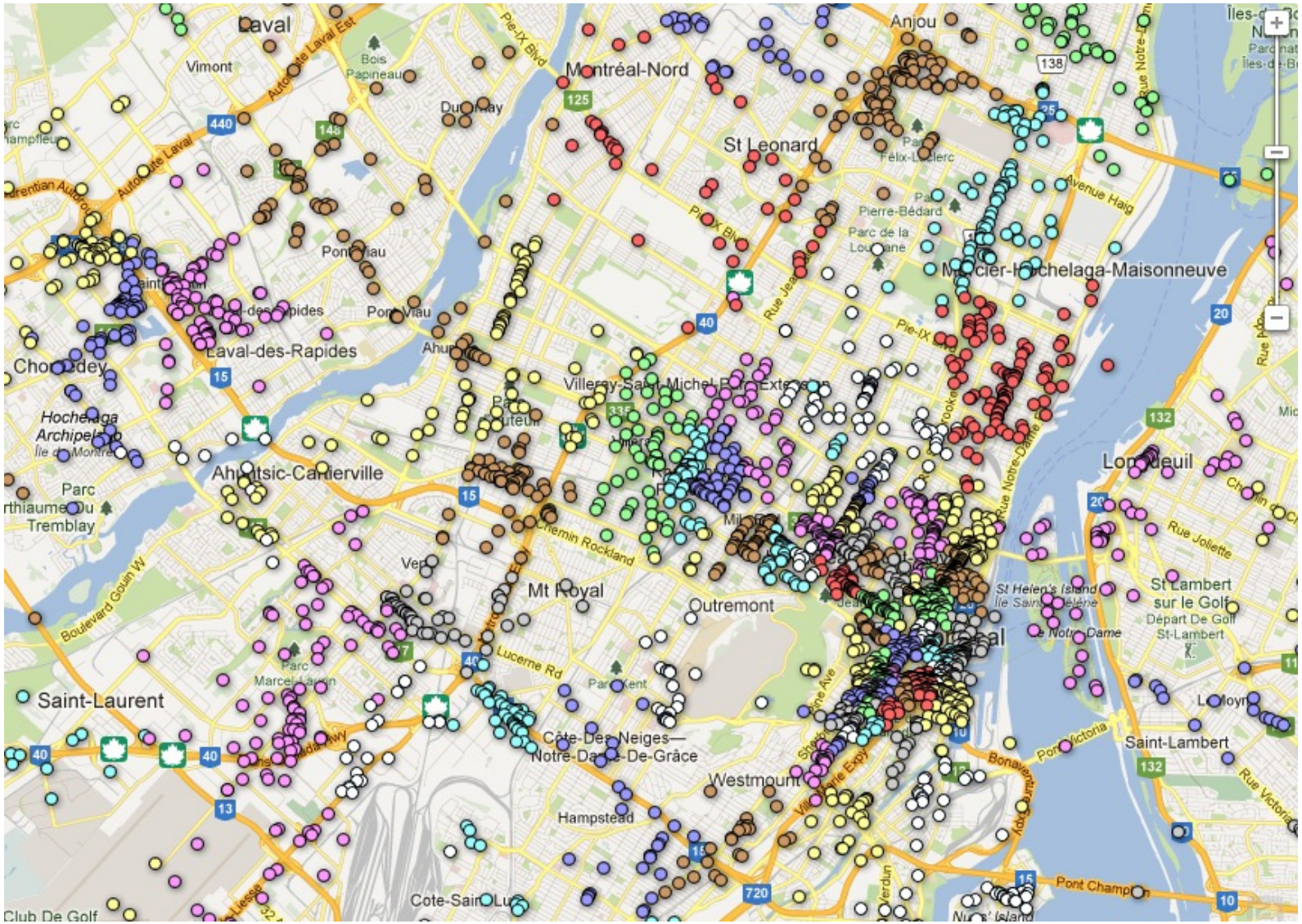




Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Case Study

Livehoods



Cranshaw *et al.*
[The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City](#)
ICWSM, 2012

General Remarks

Clustering is a relatively **intuitive** concept for human beings as our brains do it unconsciously:

- facial recognition
- searching for patterns, etc.

In general, people are very good at **messy** data, but computers and algorithms have a harder time.

Part of the difficulty is that there is **no agreed-upon definition of what constitutes a cluster:**

- “I may not be able to define what it is, but I know one when I see one”

Suggested Reading

Clustering Overview

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.5 Clustering
    - Overview
    - Case Study: Livehoods

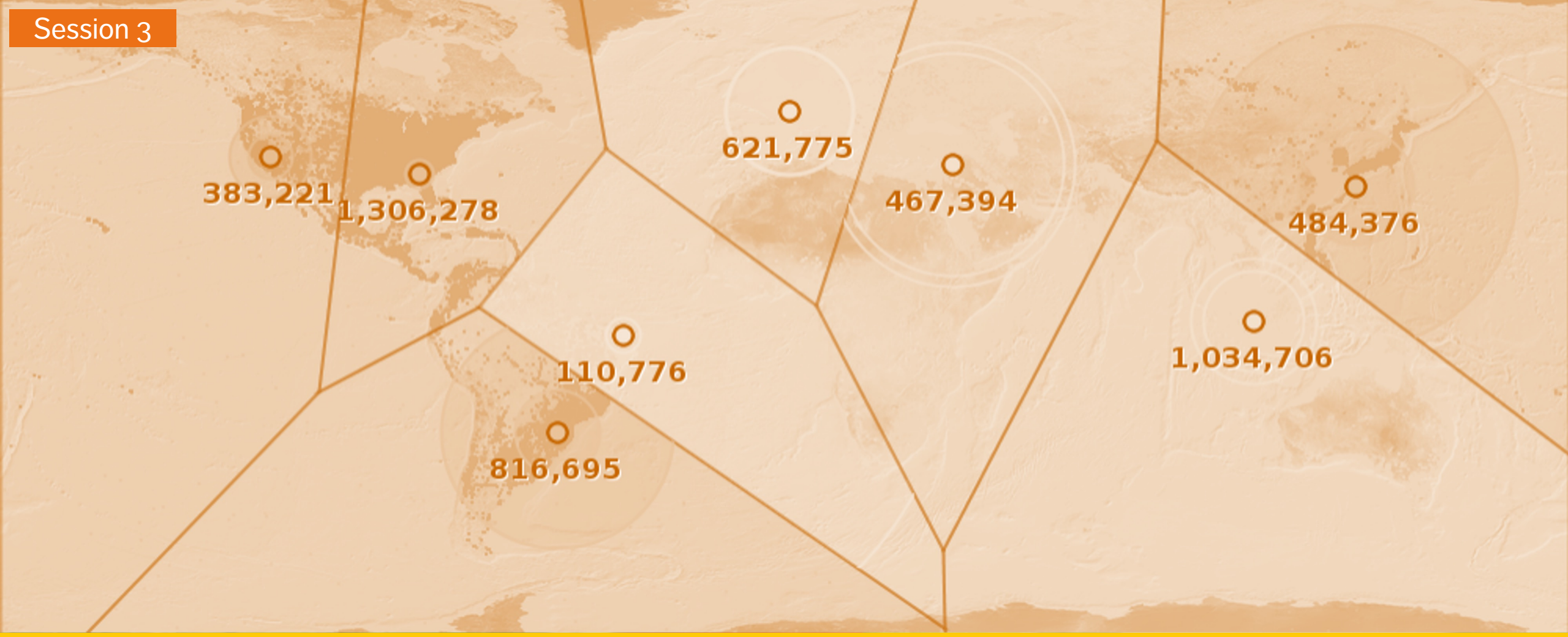
22. Focus on Clustering
    22.1 Overview

Exercises

Clustering Overview

1. What does the (potential) non-replicability of clustering imply for validation? For client and/or stakeholder buy-in?

2. Identify scenarios and questions that could use clustering in your every day work activities.



8. k-Means and Other Algorithms

Clustering Algorithms

k-Means

- classical (and over-used) model
- assumptions made about the shape of clusters

Hierarchical Clustering

- easy to interpret, deterministic

Cluster Ensembles

Latent Dirichlet Allocation

- used for topic modeling

Expectation Maximization

Clustering Algorithms

Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

- graph-based

Affinity Propagation

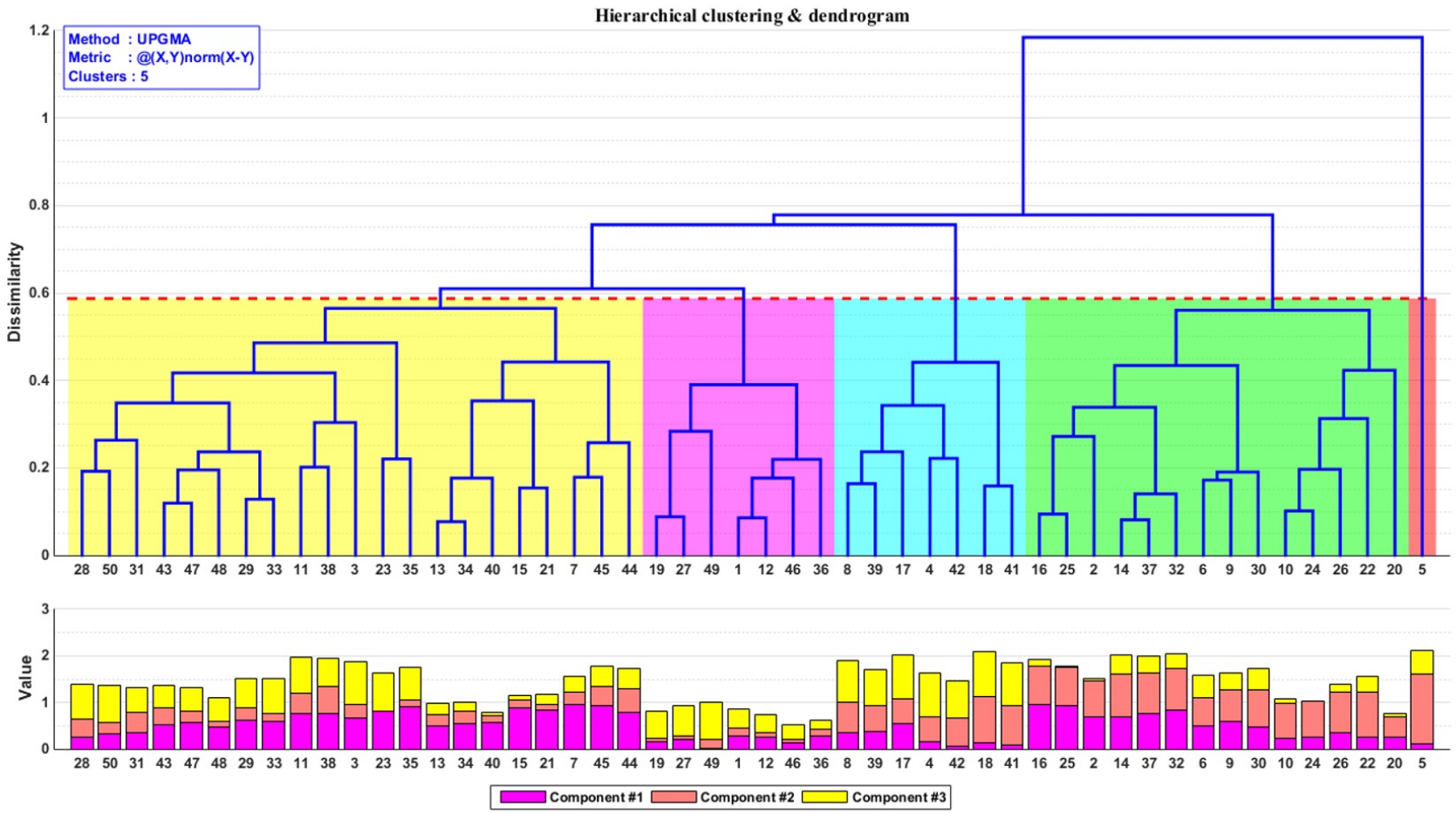
- selects the optimal number of clusters automatically

Spectral Clustering

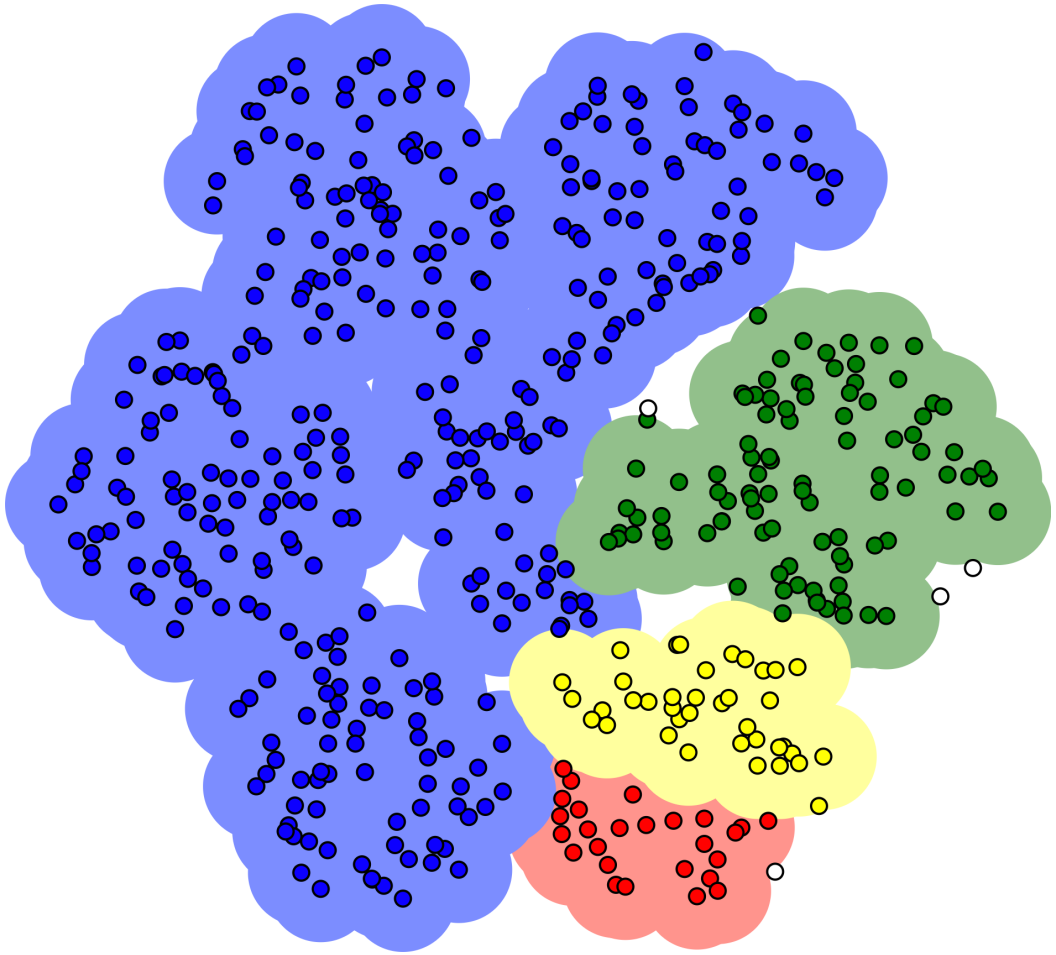
- recognizes non-blob clusters

Fuzzy Clustering

Hierarchical Clustering

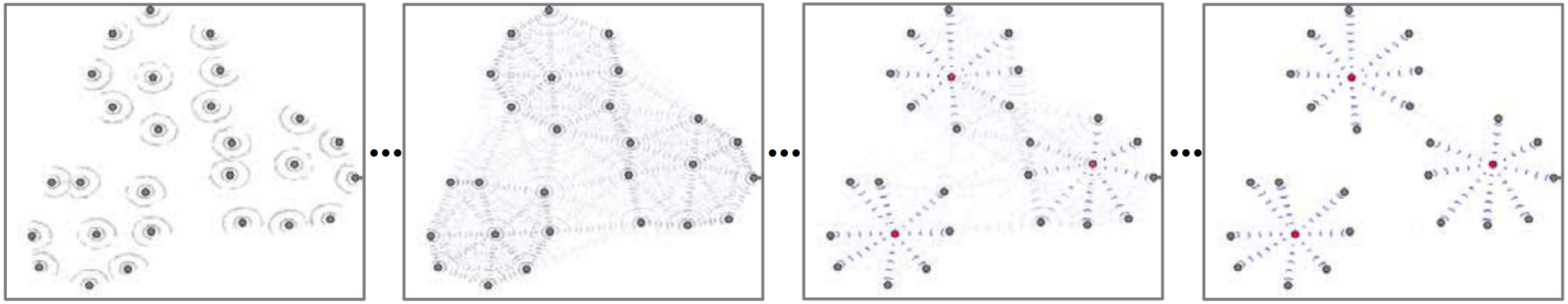


DBSCAN



[D. Dueck]

Affinity Propagation



initialization

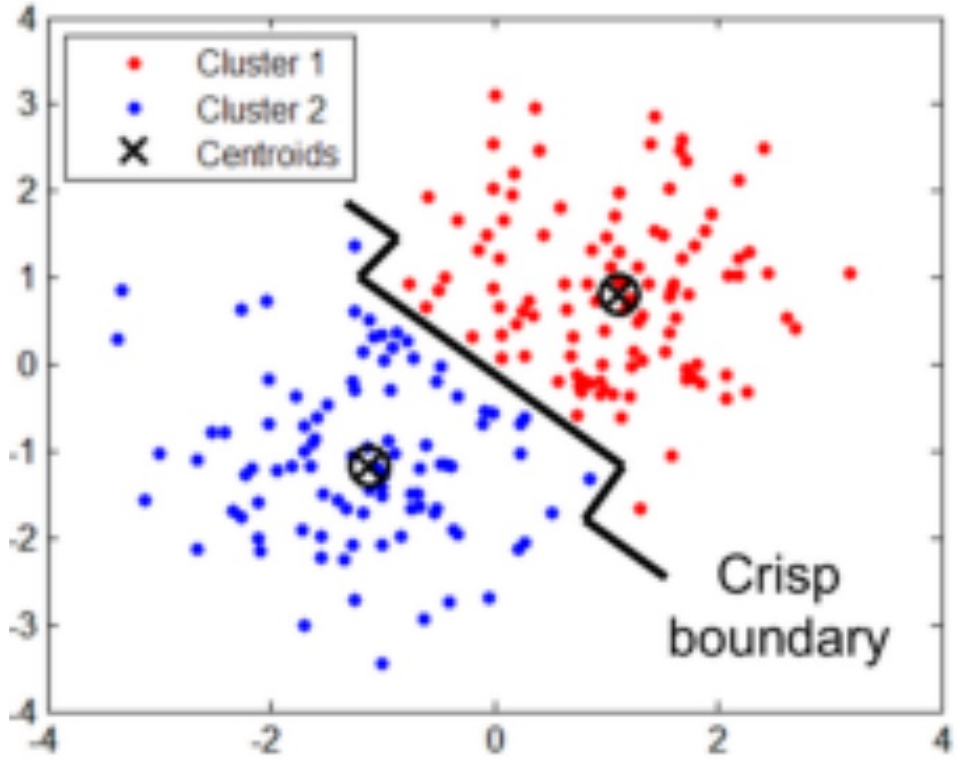
iteration X
(uncertainty)

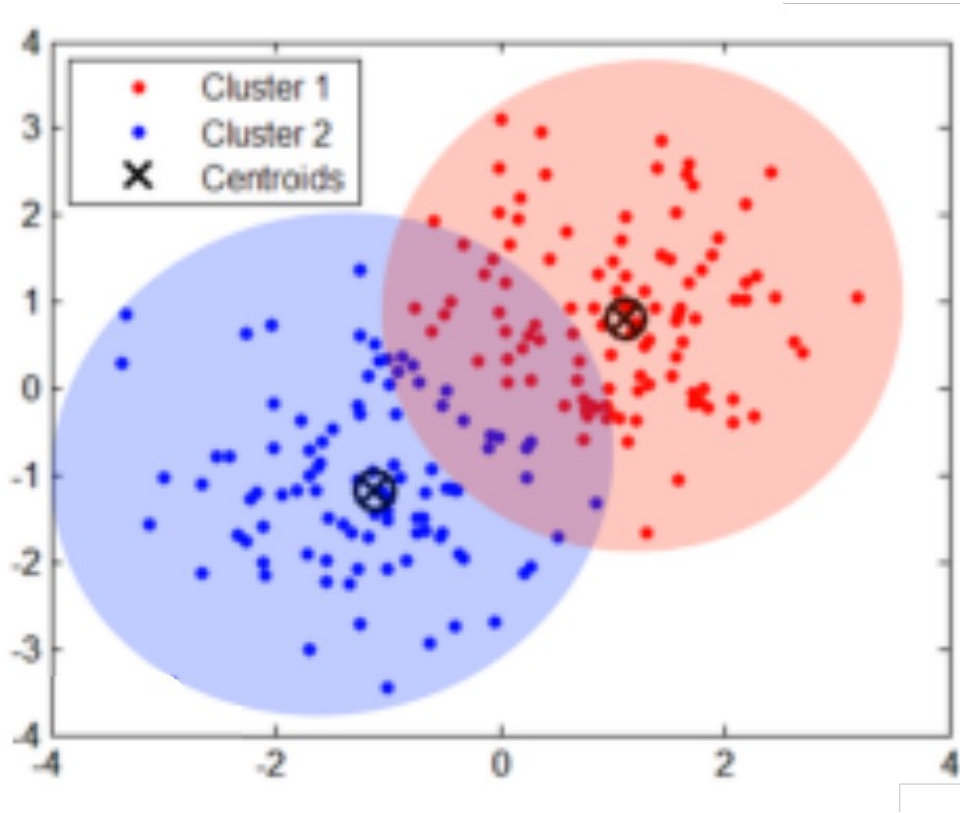
iteration Y
(emergence)

final iteration
(convergence)

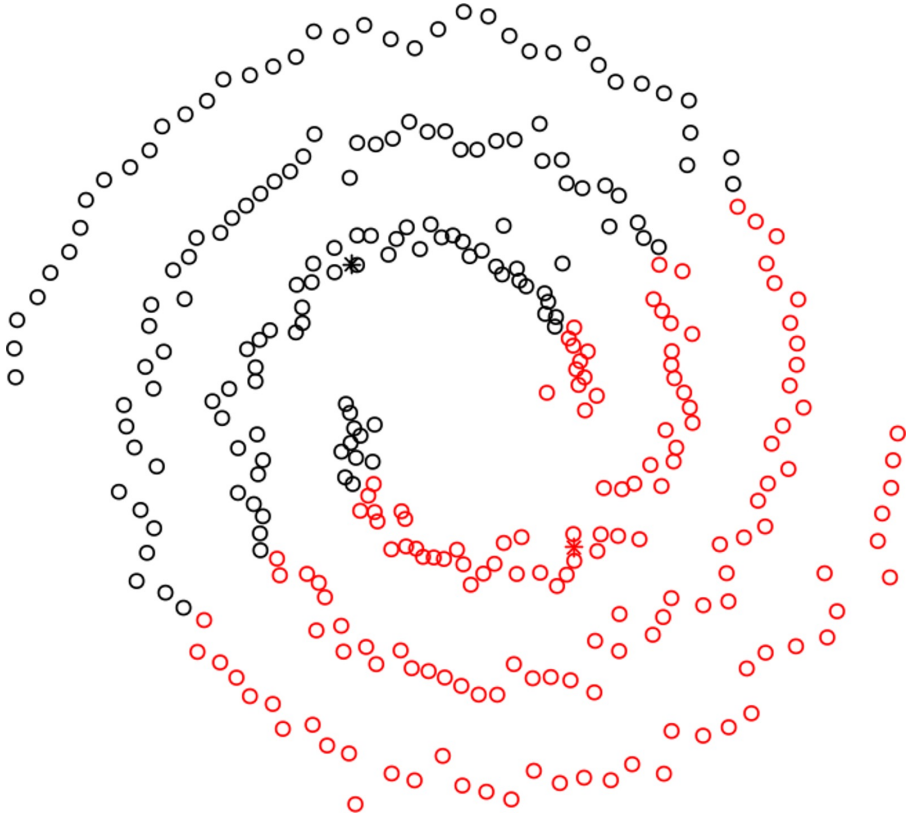
[A. Bagnaro, F. Baltar, and G. Brownstein]

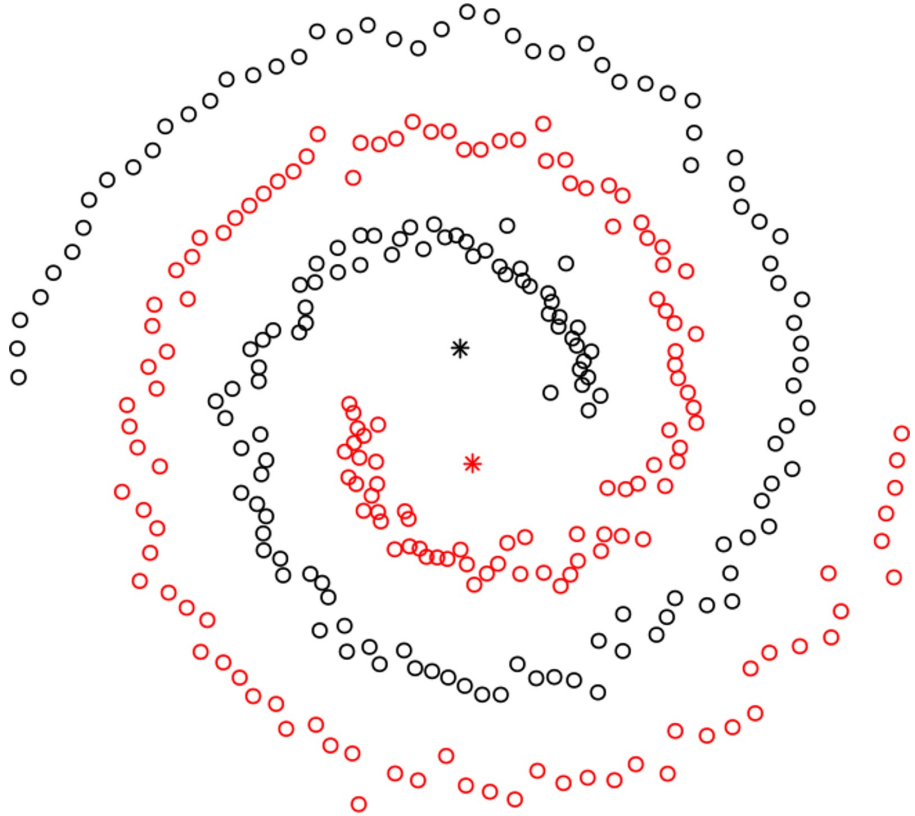
k-Means and Fuzzy c-Means





k-Means and Spectral Clustering

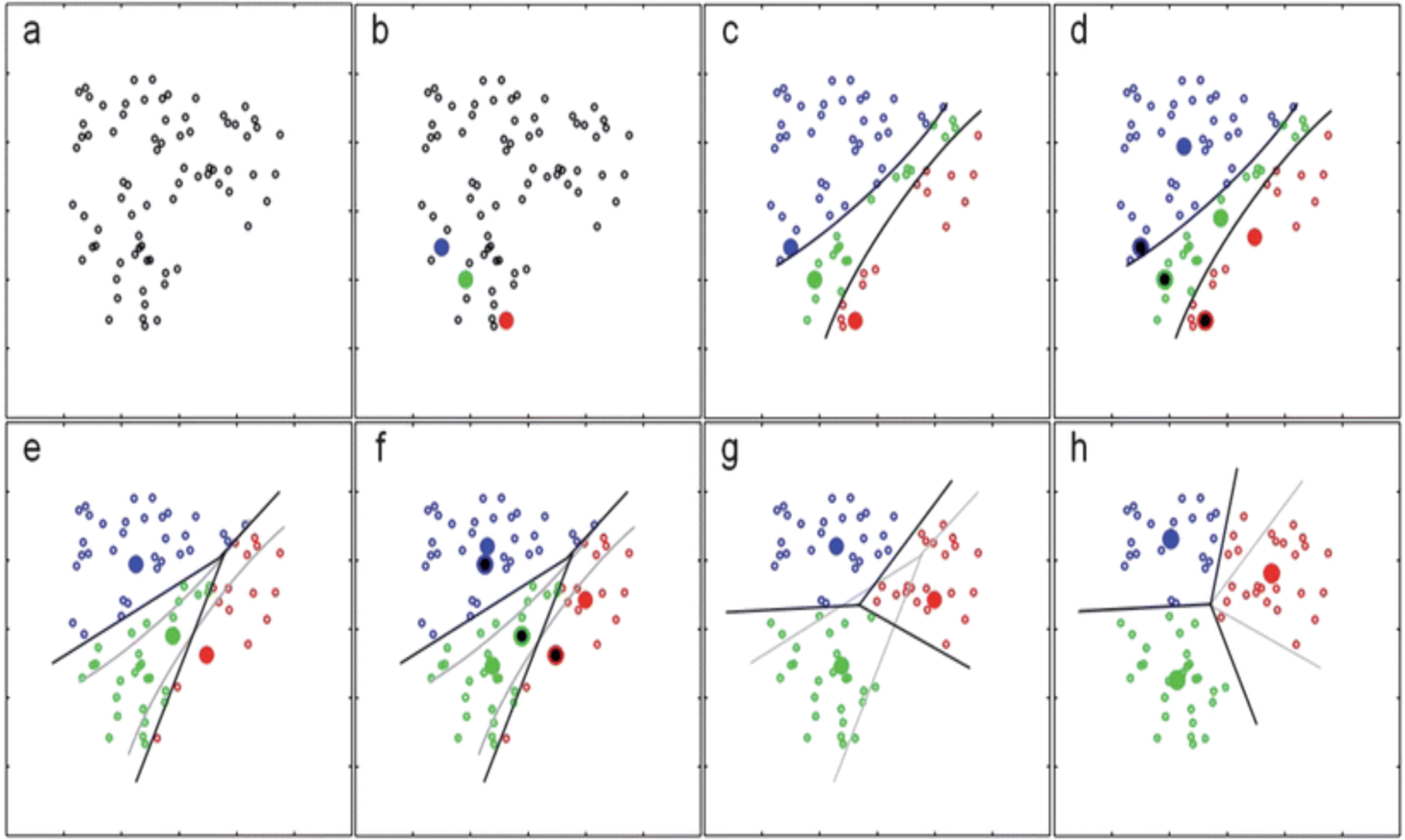




k-Means Algorithm

1. Select the desired **number of clusters**, say k
2. Randomly choose k instances as initial **cluster centres**
3. Calculate the **distance** from each observation to each centre
4. Place each instance in the cluster whose centre it is **nearest** to
5. Compute the **centroid** for each cluster
6. Repeat steps 3 – 5 with the new centroids
7. Repeat step 6 until the clusters are **stable**

[A.H. Du]



k-Means Strengths

Easy to implement (without having to actually compute pairwise distances).

- extremely common as a consequence
- elegant and simple

In many contexts, k-means is a **natural** way to look at grouping observations.

Helps provide a **basic understanding of the data structure** in a first pass.

k-Means Limitations

Data points can only be assigned to **one** cluster

- this can lead to overfitting
- robust solution: consider the probability of belonging to each cluster

Underlying clusters are assumed to be **blob-shaped**

- k-means will fail to produce useful clusters if that assumption is not met in practice

Clusters are assumed to be separate (discrete)

- k-means does not allow for **overlapping** or **hierarchical** groupings

k-Means Limitations

There are many ways to pick the **optimal number** of clusters k.

One problem is that the algorithm is stochastic: different initial configurations may yield **different outcomes**, which may yield a different optimal number.

It may also depend on the **size** of data, the choice of **distance**, the choice of **cluster quality metric**, etc.

Suggested Reading

k-Means and Other Algorithms

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.5 Clustering
    - Clustering Algorithms
    - k-Means
    - Toy Example: Iris Dataset

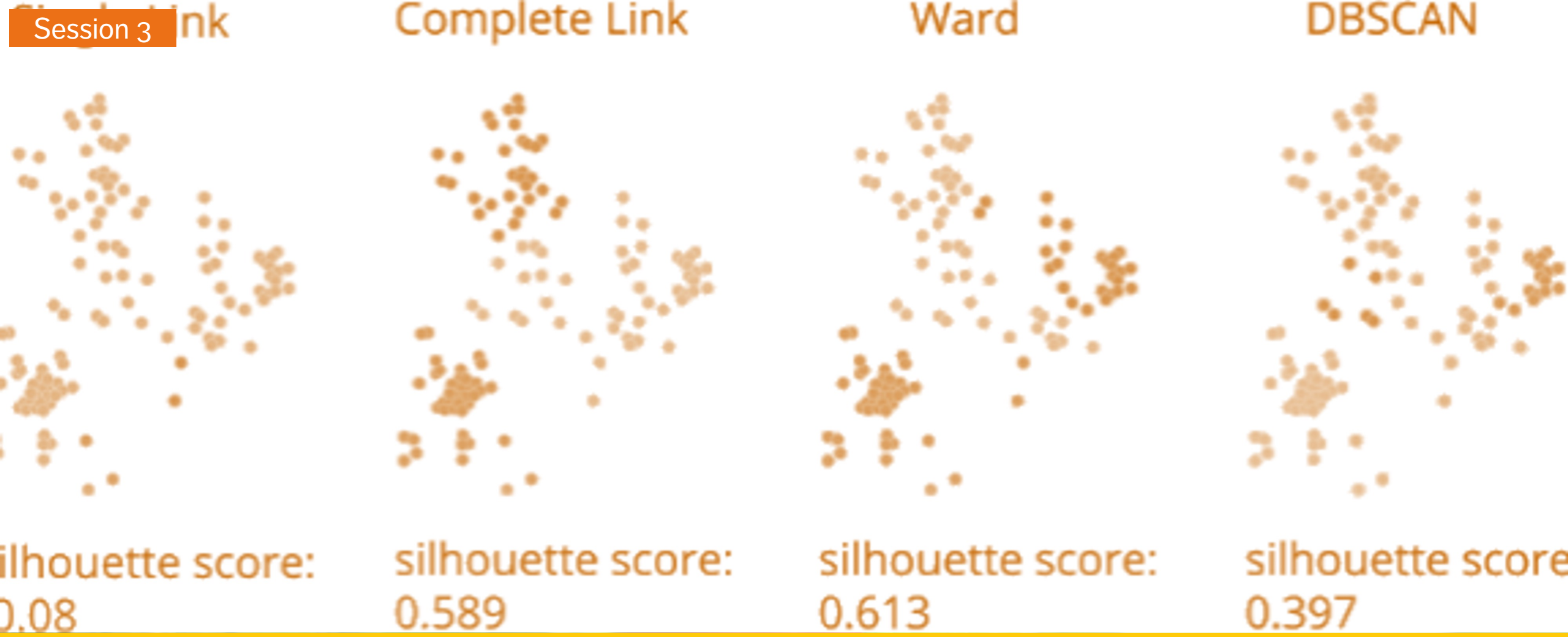
    19.7 R Examples
    - Clustering: Iris Dataset

22. Focus on Clustering
    22.2 Simple Clustering Methods
    22.4 Advanced Clustering Approaches

Exercises

k-Means and Other Algorithms

1. Go over the iris clustering example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank dataset (you may wish to visualize the dataset first) in order to build a clustering scheme. Determine the optimal number of clusters using the Davies-Bouldin index.



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:

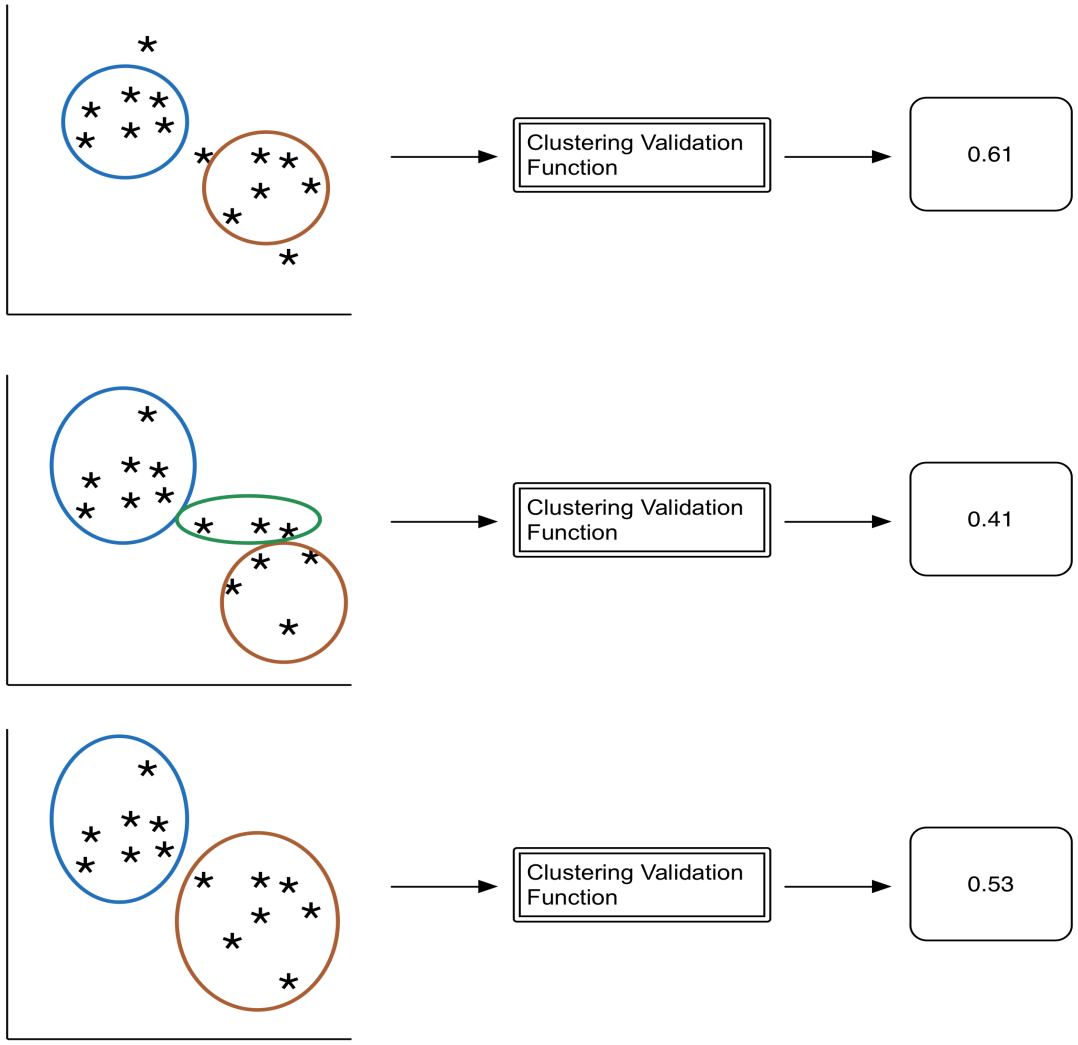
- creating clusters
- **assessing cluster quality**

Clustering functions

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

Assessing cluster quality

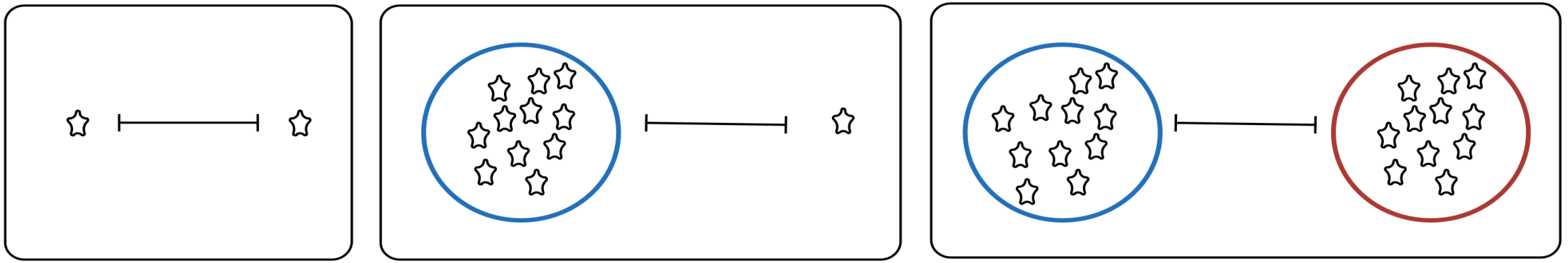
- input: instances + cluster assignments (+ similarity matrix, usually)
- output: a numeric value



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**



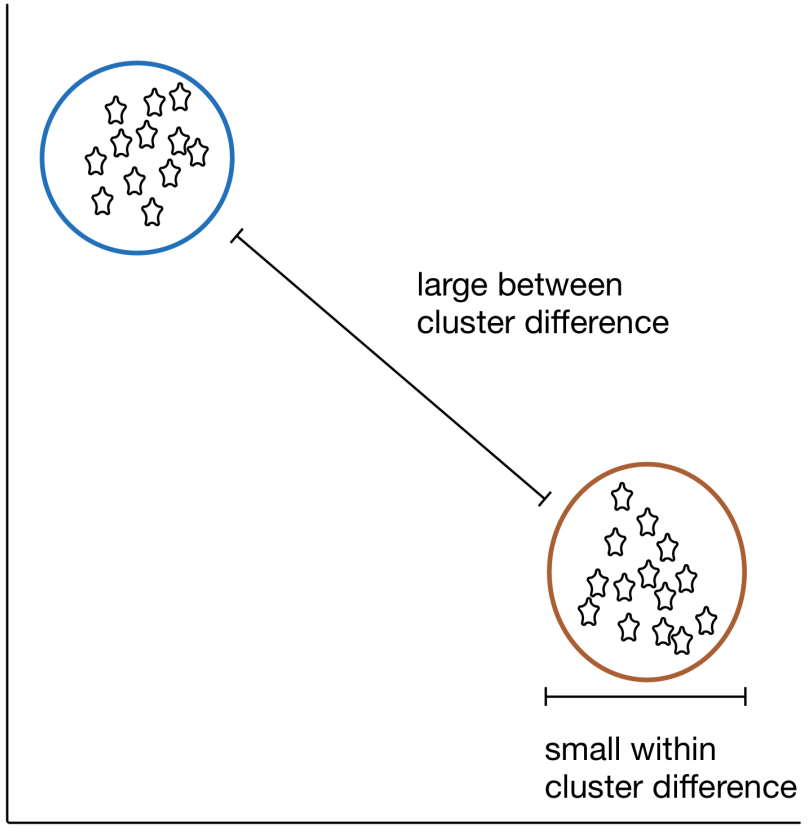
Internal Validation Goals

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

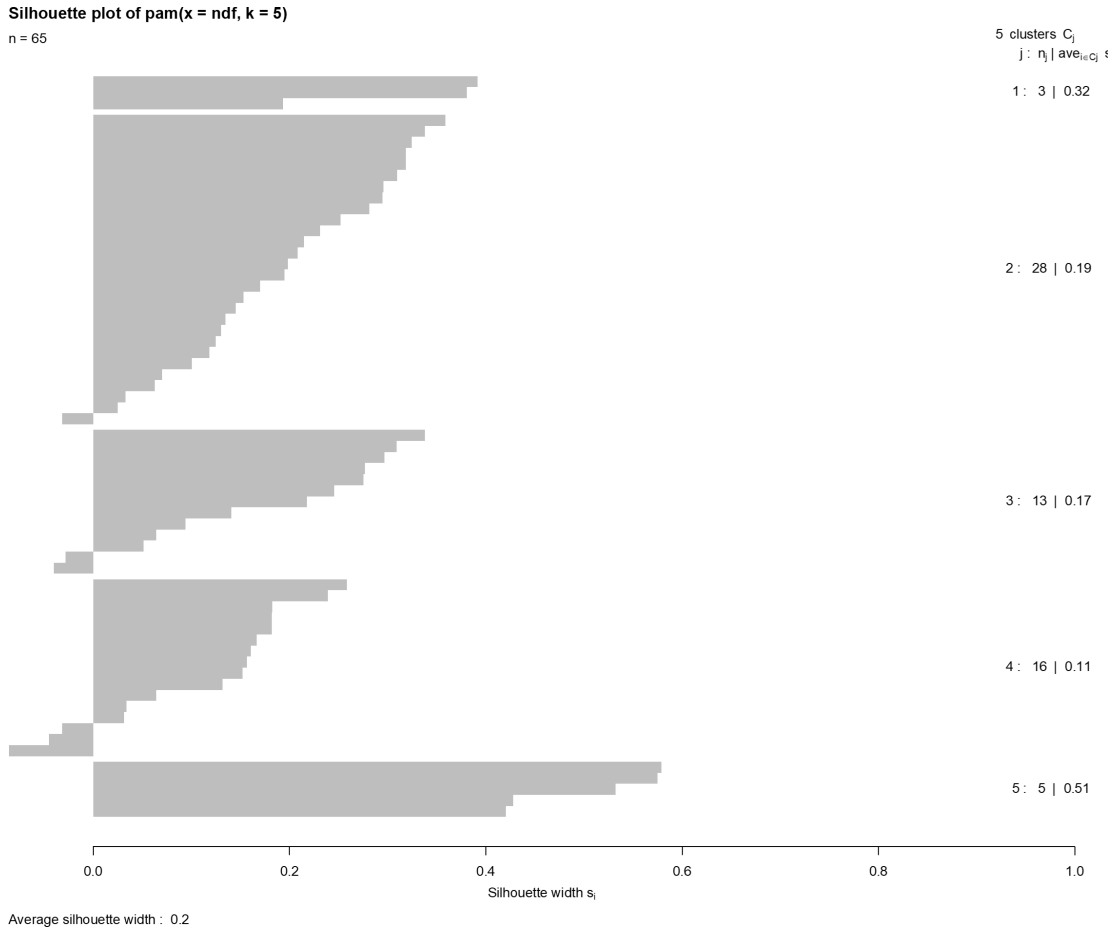
Davies-Bouldin index

Dunn's index

Silhouette metric

Within Sum of Squares

etc. (there are tons!)

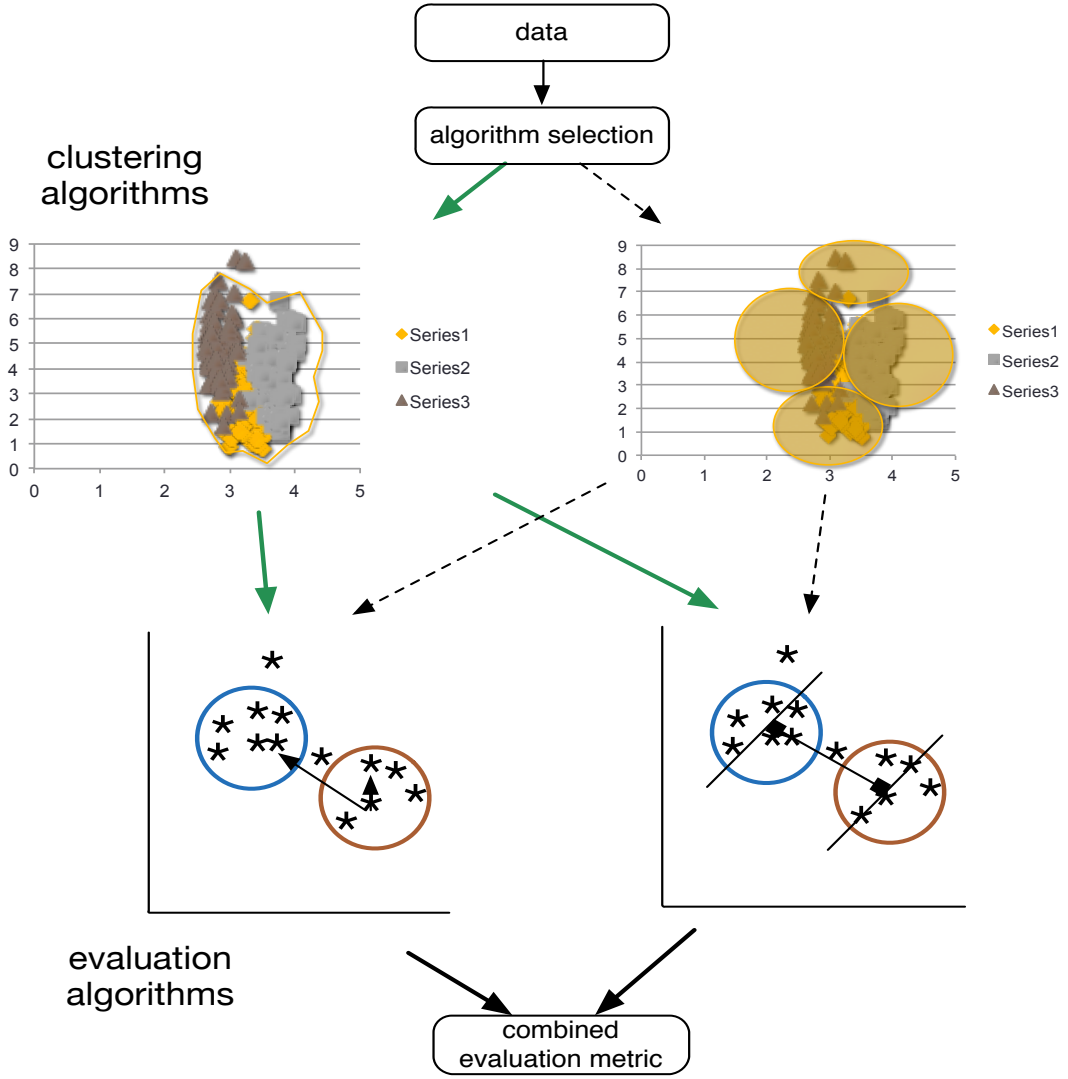


Relative Validation

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**.



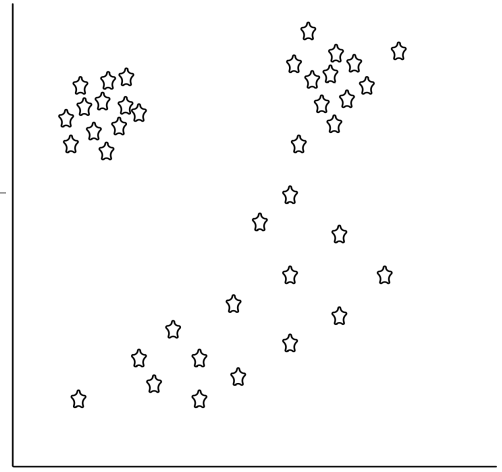
Ensemble Methods

Some options:

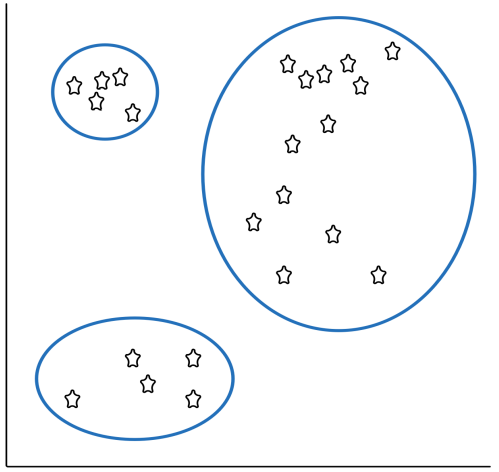
- multiple samples from the same source
- different subsets of columns are used
- different algorithms are used

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

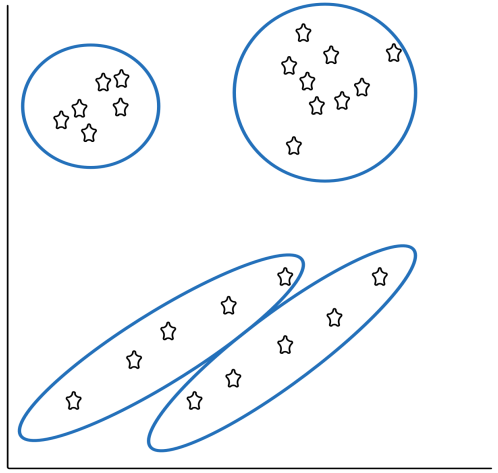
If the results are **not stable** across the clustering outcomes, more investigation is required.



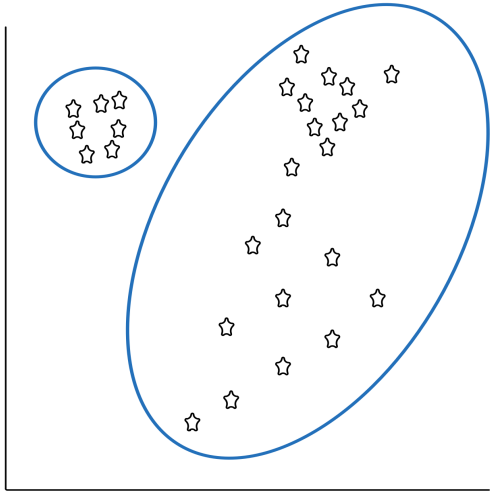
Full dataset



Sample 1 clustering



Sample 2 clustering



Sample 3 clustering

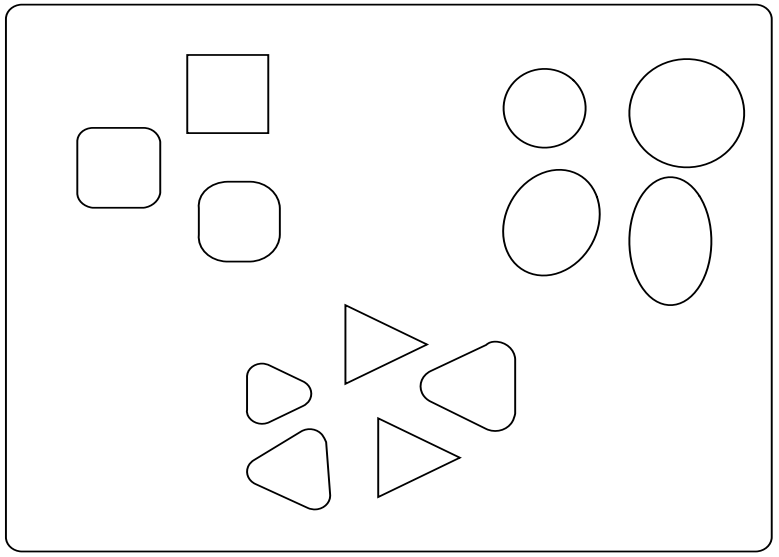
External Validation

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

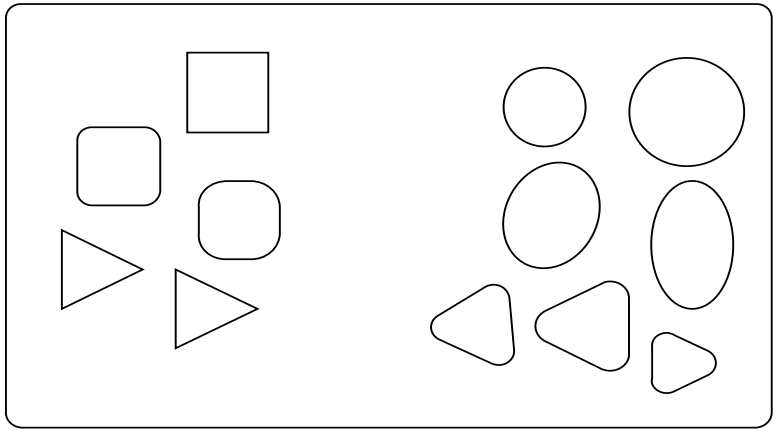
Outside information is typically the 'correct' class.

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

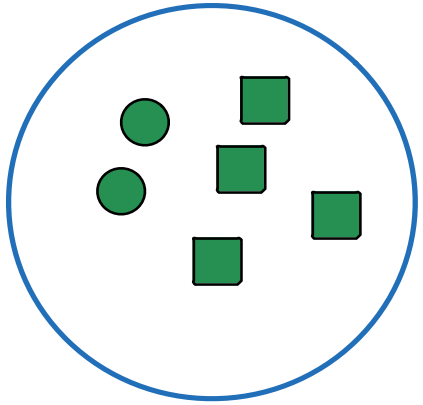
For this external validation metric, each cluster is assigned to the class which is **most frequent** in the cluster.

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

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

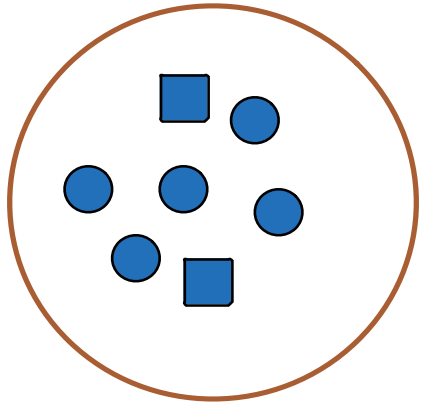
Assuming we are interested in shape…

SQUARE CLUSTER



purity = 66%

CIRCLE CLUSTER



purity = 71%

Clustering Challenges

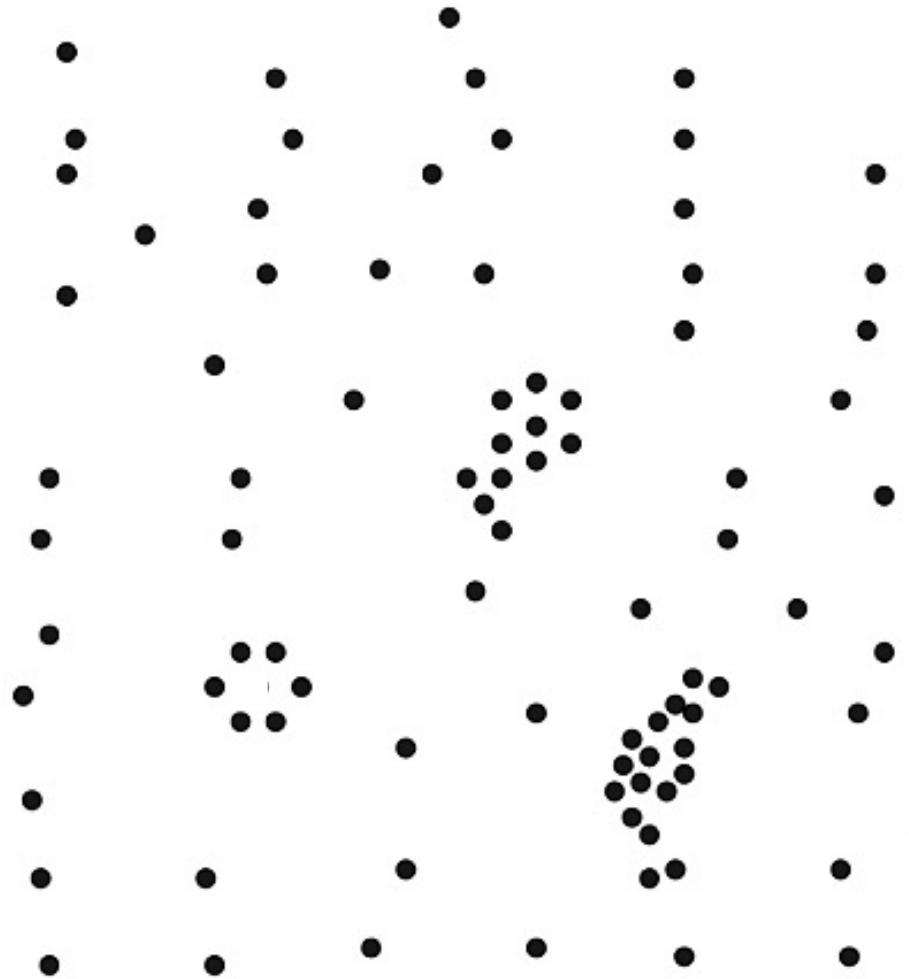
Automation
relatively intuitive for humans, but harder for machines

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

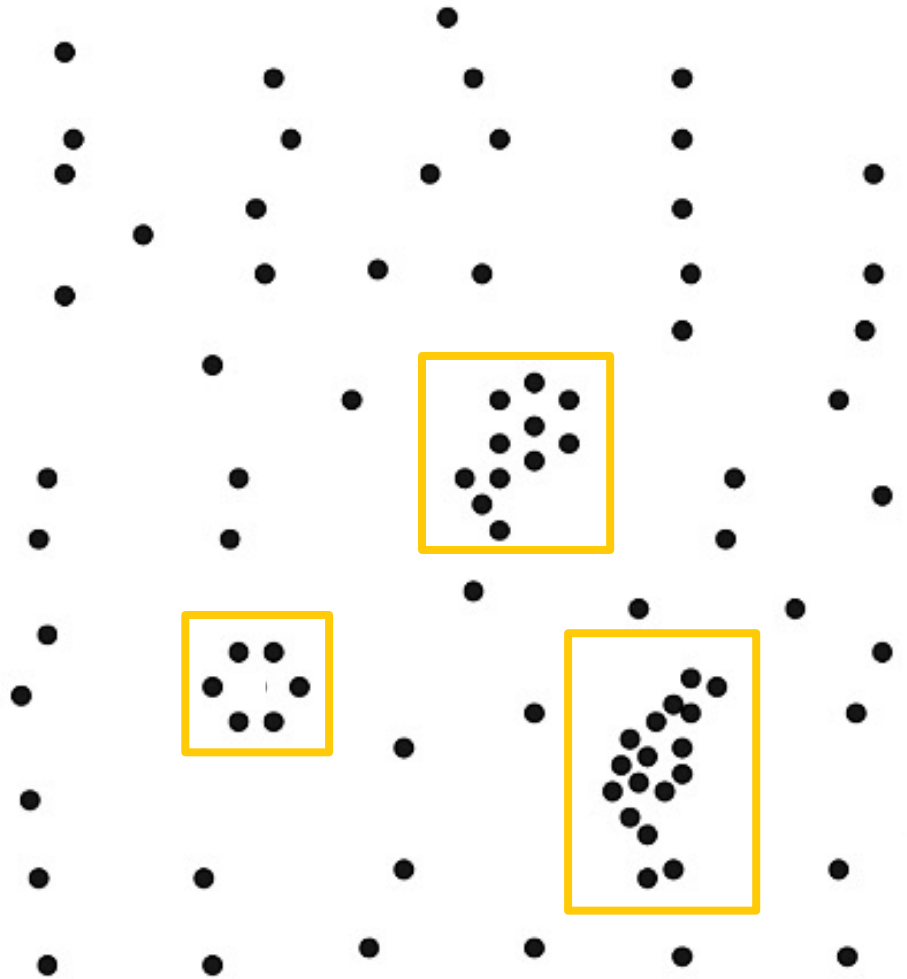
Lack of repeatability
non-deterministic: the same algorithm, applied twice to the same dataset can discover completely different clusters

Number of clusters
optimal number of clusters difficult to determine

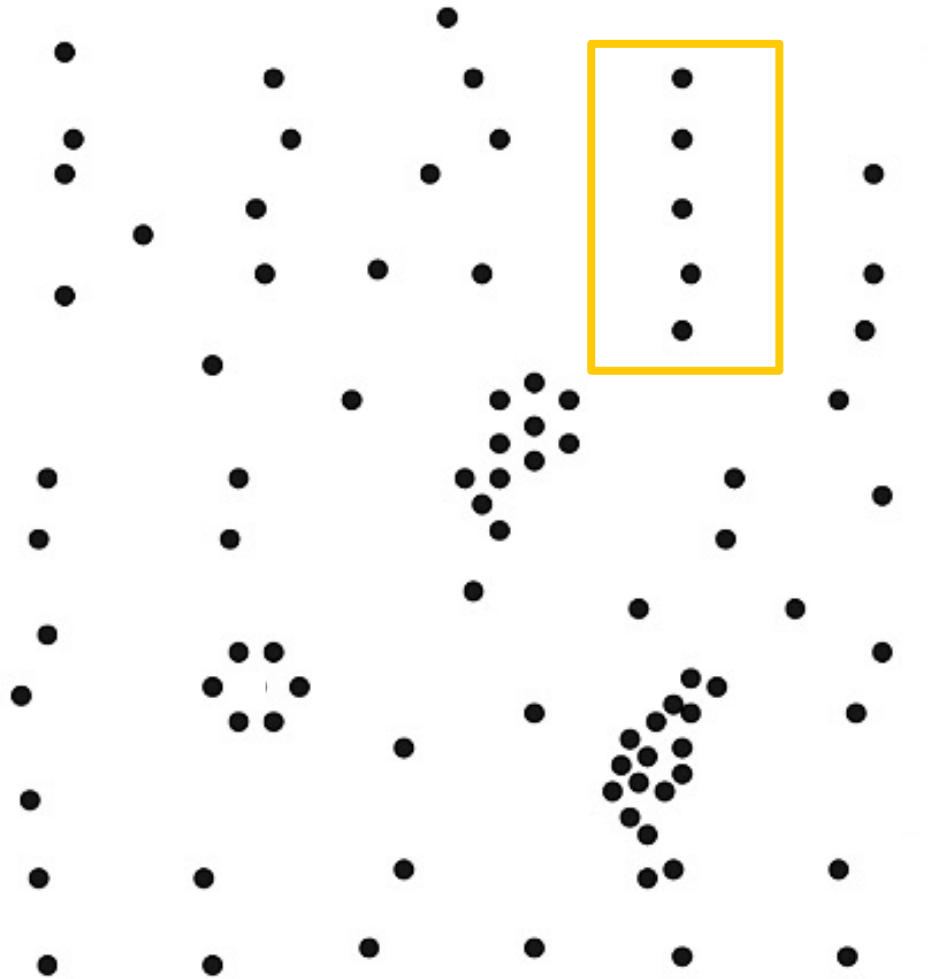
Session 3



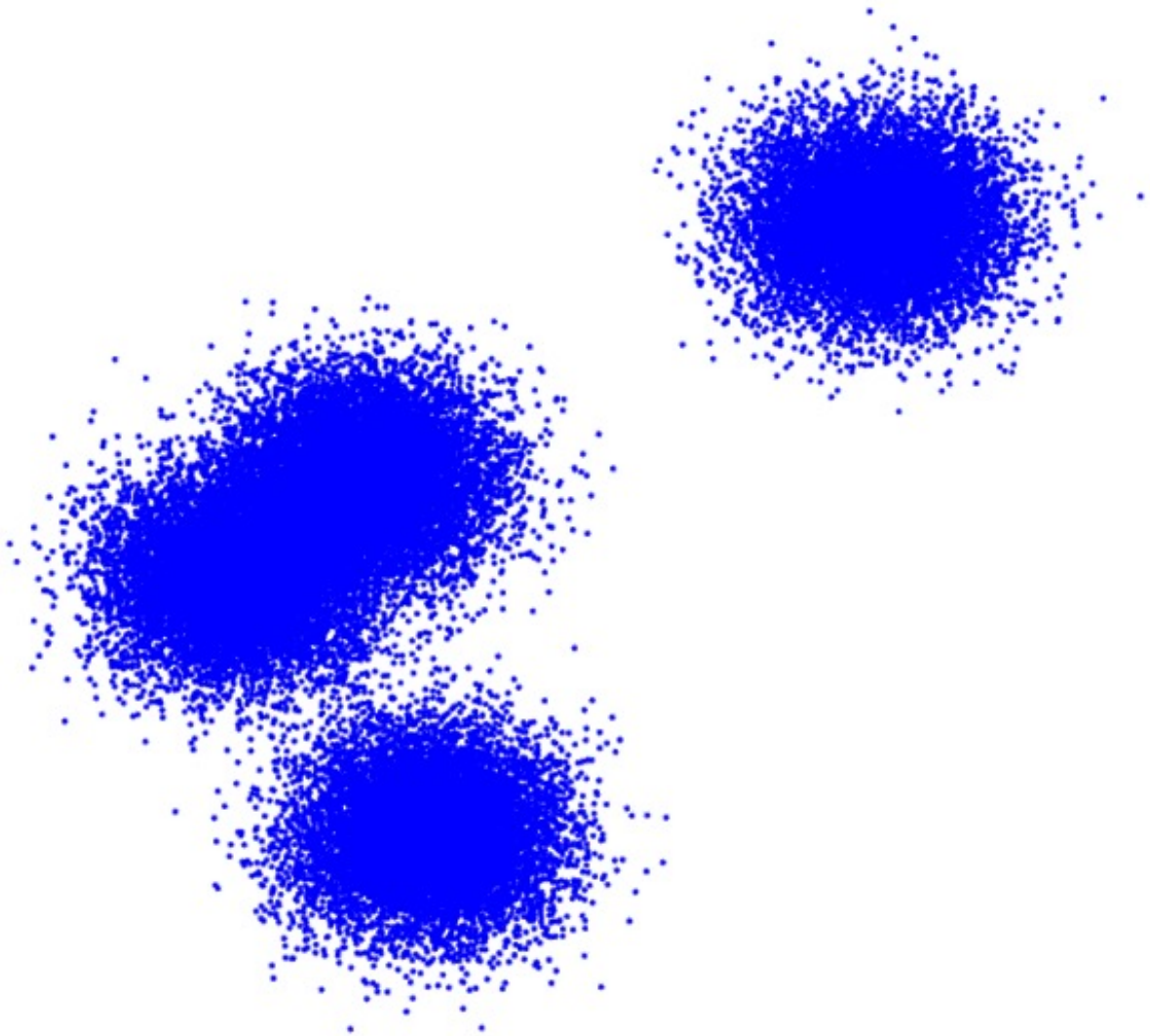
Session 3



Session 3



Session 3



Clustering Challenges

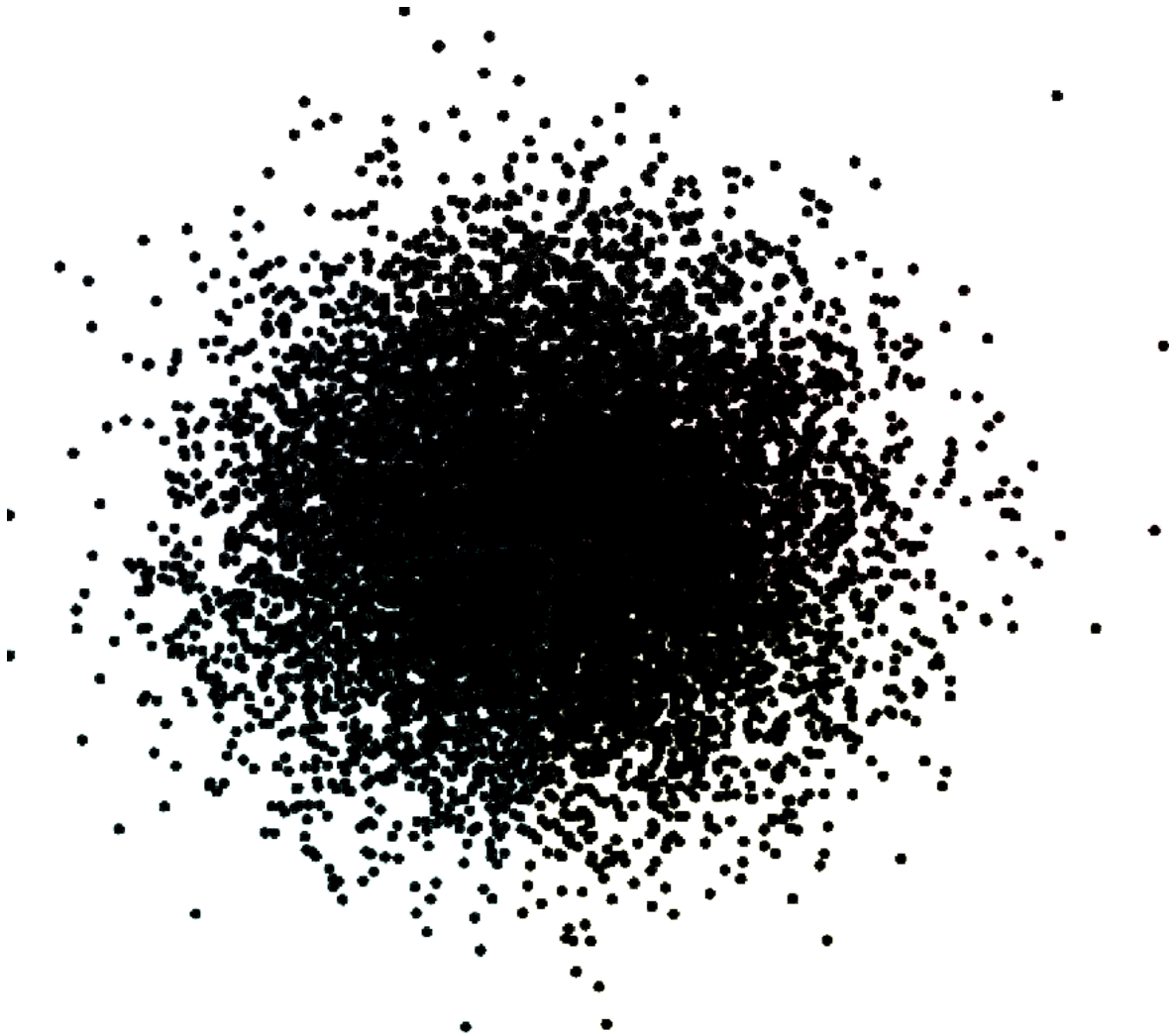
Cluster description
should clusters be described using representative instances or average values?

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

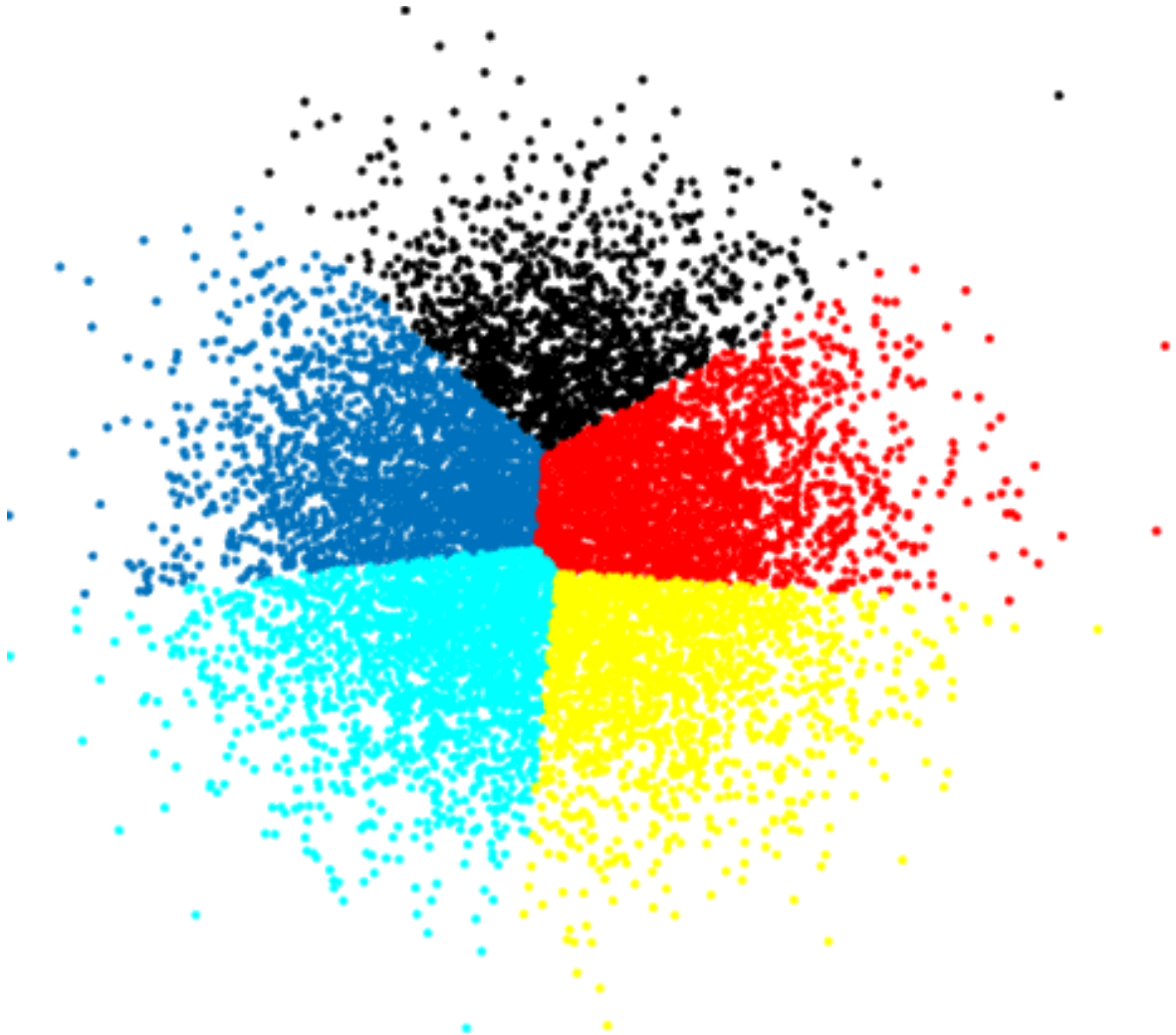
Ghost clustering
most methods will find clusters even if there are none in the data

***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
Volume 3: Spotlight on Machine Learning

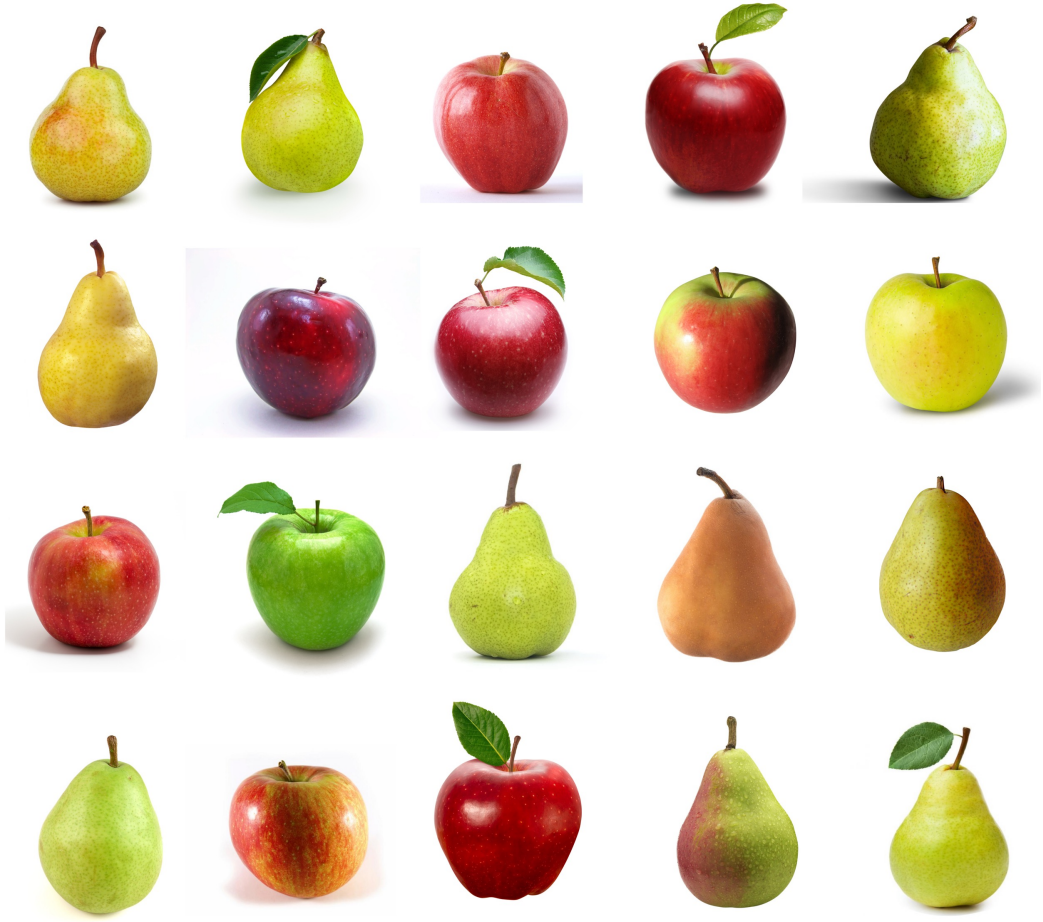
19. Introduction to Machine Learning
    19.5 Clustering
    - Validation

22. Focus on Clustering
    22.3 Clustering Evaluation

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.