Clustering

Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
@TWlyY29

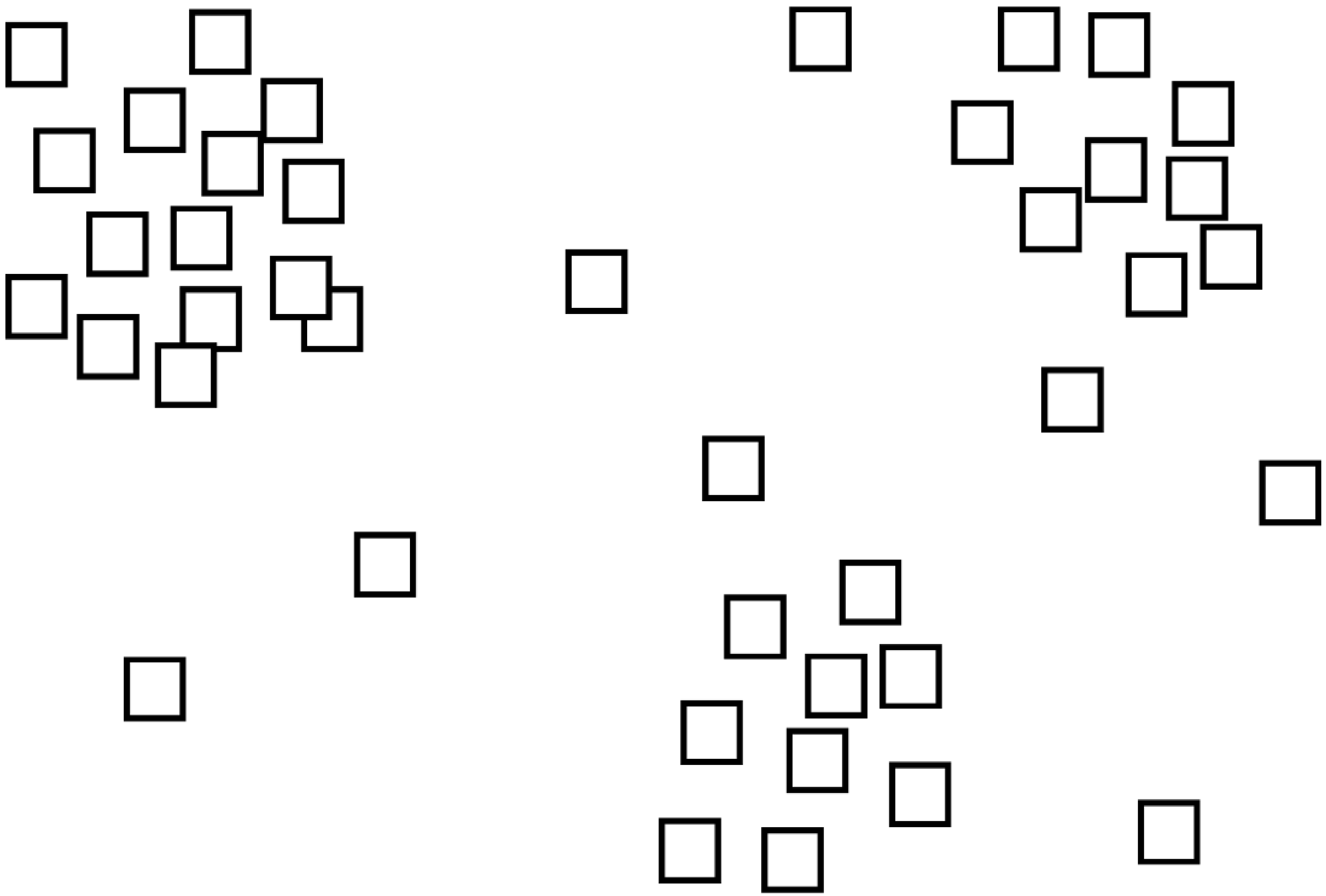
Ursus Wehrli, https://www.kunstaufraeumen.ch

UNIVERSITÄT BAYREUTH

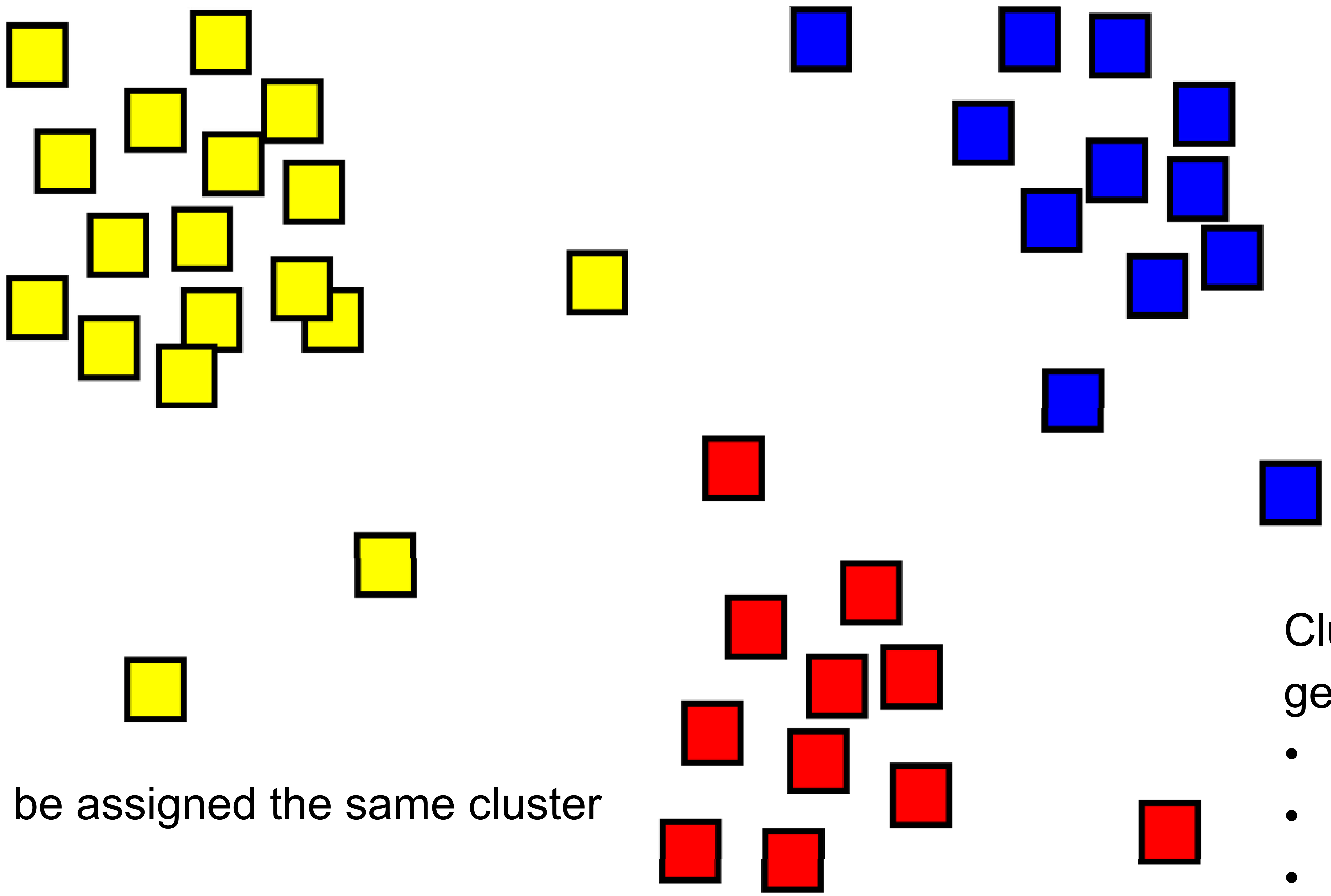
UNIVERSITÄT BAYREUTH

Clustering is always *subjective*

Clustering is hard!



Clustering is fuzzy



Similar objects should be assigned the same cluster

Dissimilar objects should end up in different clusters

Clusters aren‘t pre-defined

Clusters should have a few geometric characteristics:

- Connected
- Separated
- Low variance
- Higher density than surrounding

Why is it hard and fuzzy?

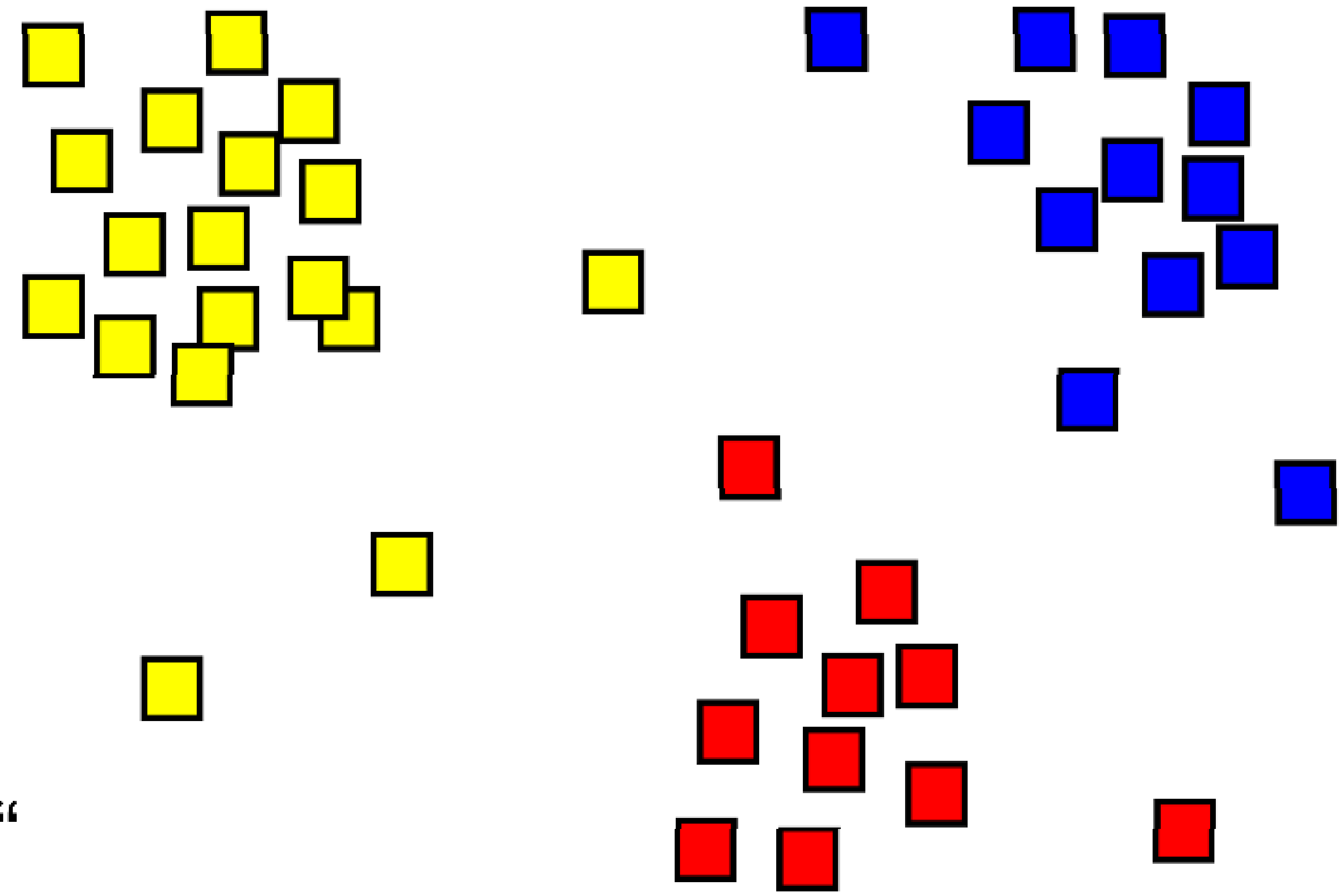
Many applications involve several hundred or several thousand dimensions

High-dimensional spaces look different
(Pairs of points are hard to distinguish)

No precise definition of „clusters“

No precise definition of „validity“ of clusters

Subjective results, no specific definition seems „best“
in the general case



Clustering Problems

Marketing: discover groups of purchasing activities

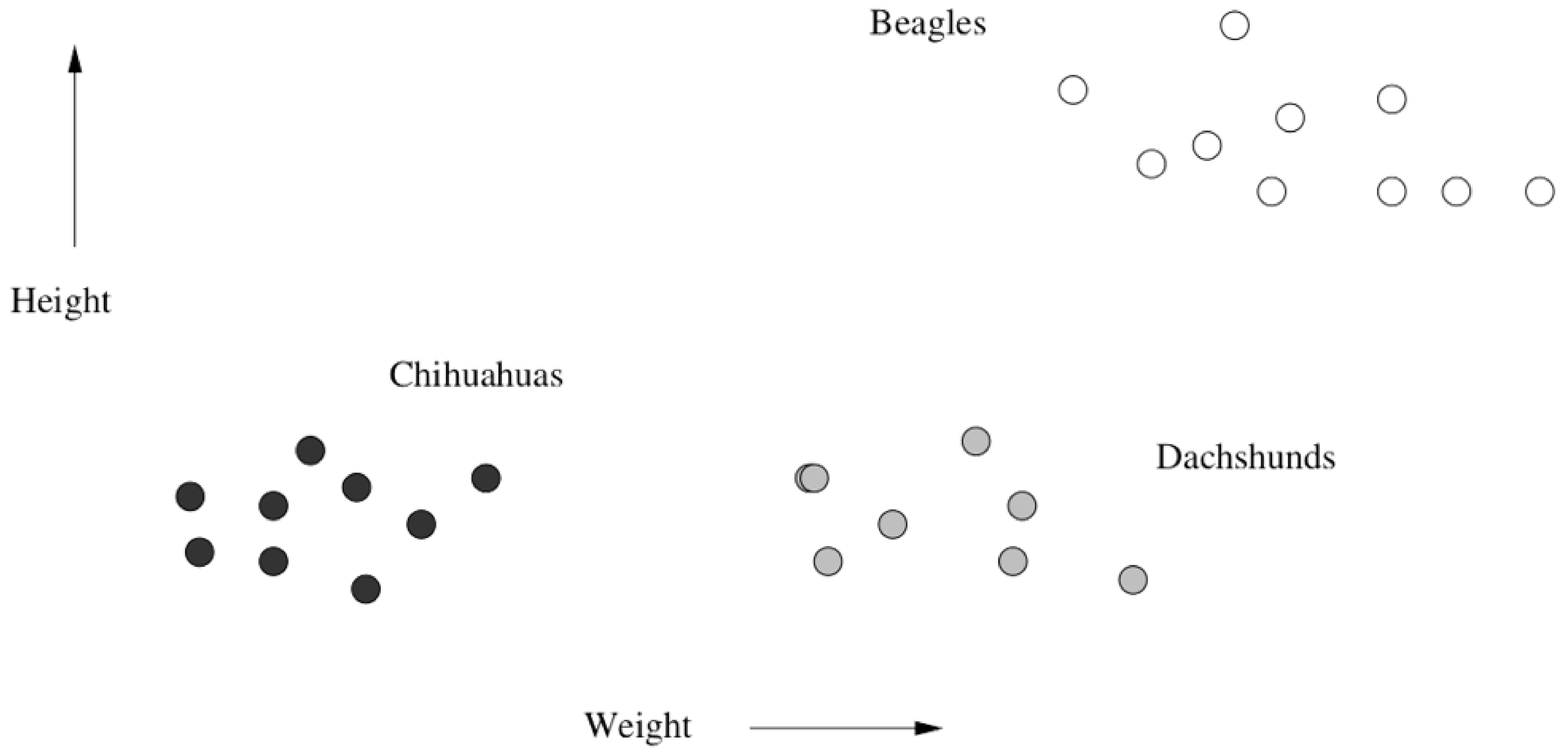
Climate: patterns of atmospheric phenomena help understand Earth climate

Economics: market research

Information Science: Clustering documents according to their topic

Requirements for Clustering

A dataset which is a collection of *points*
which belong to some *space*
which allows to measure *distance*.



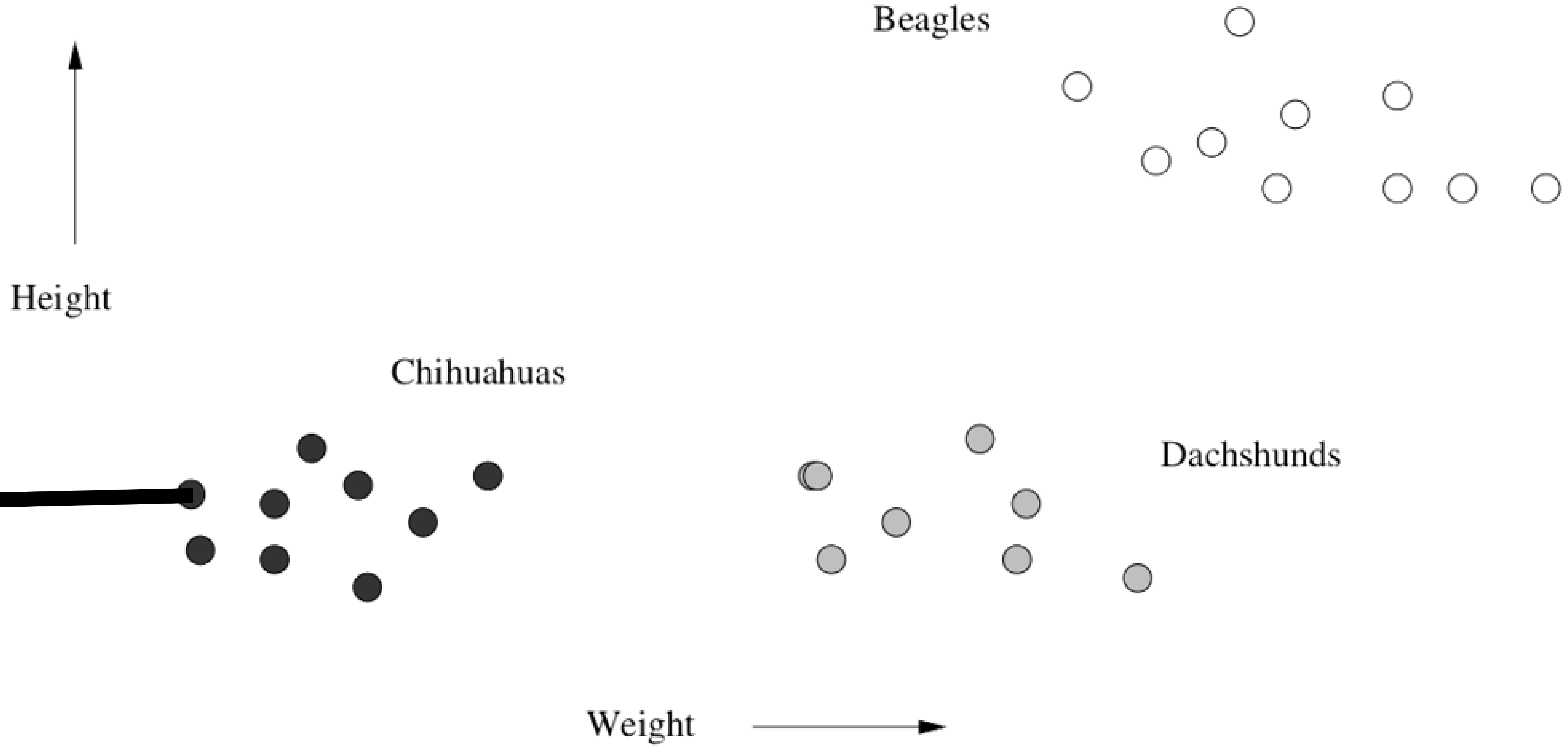
Leskovec, J., Rajaraman, A. and Ullman, J.D., 2020. *Mining of massive data sets*. Cambridge university press.

Points in Euclidean Space

Clustering performs best in low-dimensional Euclidean spaces:

- Every point is a vector of real numbers
- The length of the vector is the number of dimensions
- Components of vector are coordinates of points

chihuahua_3: <2.53, 21.2>
Weight: 2.53 kg
Height: 21.20 cm



Leskovec, J., Rajaraman, A. and Ullman, J.D., 2020. *Mining of massive data sets*. Cambridge university press.
https://en.wikipedia.org/wiki/File:Chihuahua1_bvdb.jpg

Points in Non-Euclidean Space

Example: a text document is described by occurring words

One axis represents one word, values of 0 or 1 only indicating the presence of a word

The „space“ consists of all axes describing all words of a dictionary (i.e. the set of selected words)

„The internet is a network of computers. In this network, a lot of data is transmitted.“

Vector representation: <0,1,0,0,1,0,0,0,1>

Words:

1. Social
2. Network
3. Computer
4. Media
5. Internet
6. Meme
7. Machine
8. Learning
9. Data

Measuring Distance

A distance measure is a function $d(x, y)$ that produces a real number, to which arguments x and y are points in space

Important properties:

- No negative distances:

$$d(x, y) \geq 0$$

- Zero-distances only for the distance from a point to itself

$$d(x, y) = 0 \; if \; and \; only \; if \; x = y$$

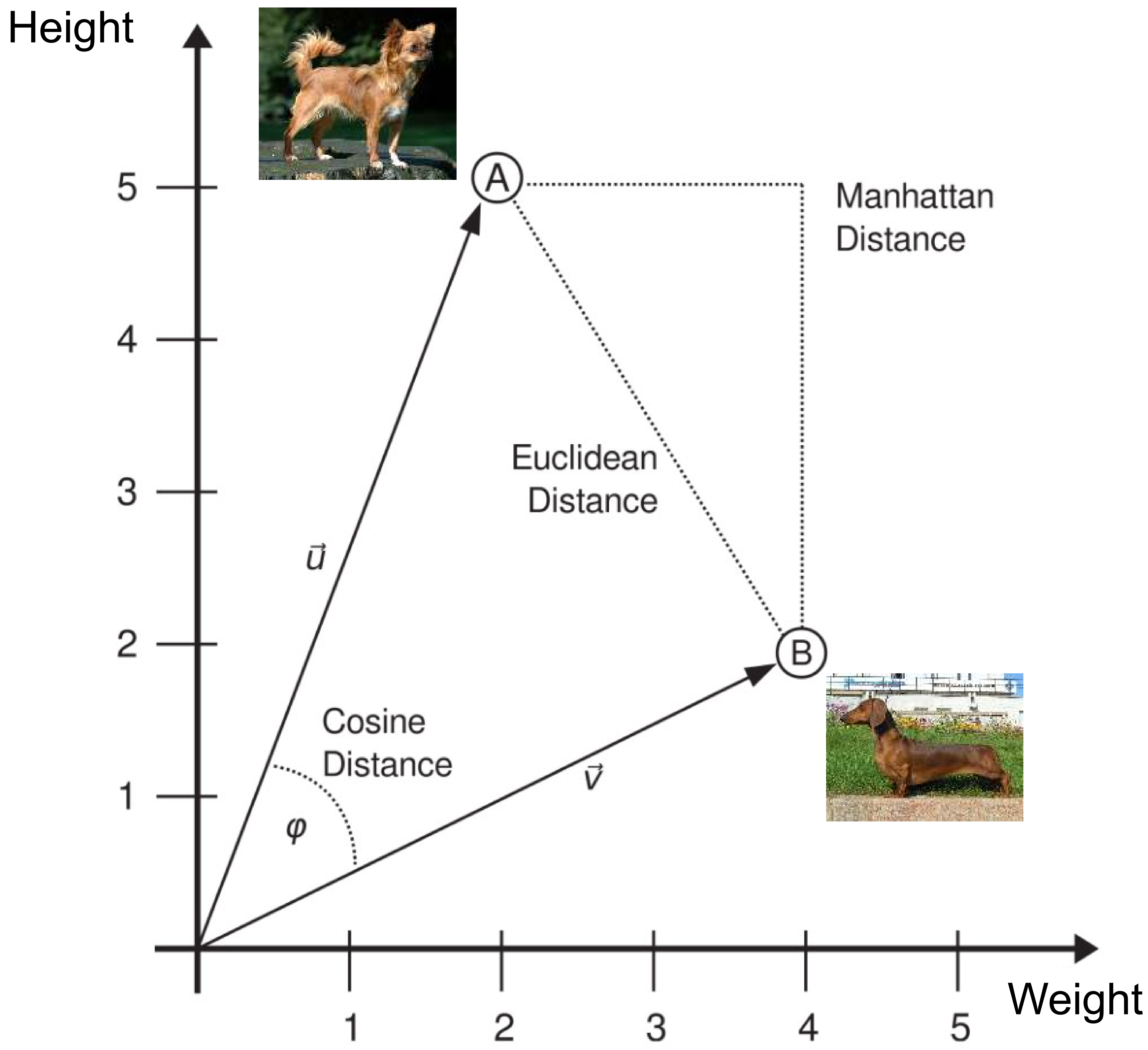
- Distances are symmetric

$$d(x, y) = d(y, x)$$

- Triangle inequality

$$d(x, y) \leq d(x, z) + d(z, y)$$

Well-Known Distance Metrics



Euclidean space:
- Euclidean distance
- Mahalanobis distance
- Manhattan distance
- Cosine distance

Non-Euclidean space:
- Jaccard distance
- Hamming distance
- Gower‘s distance

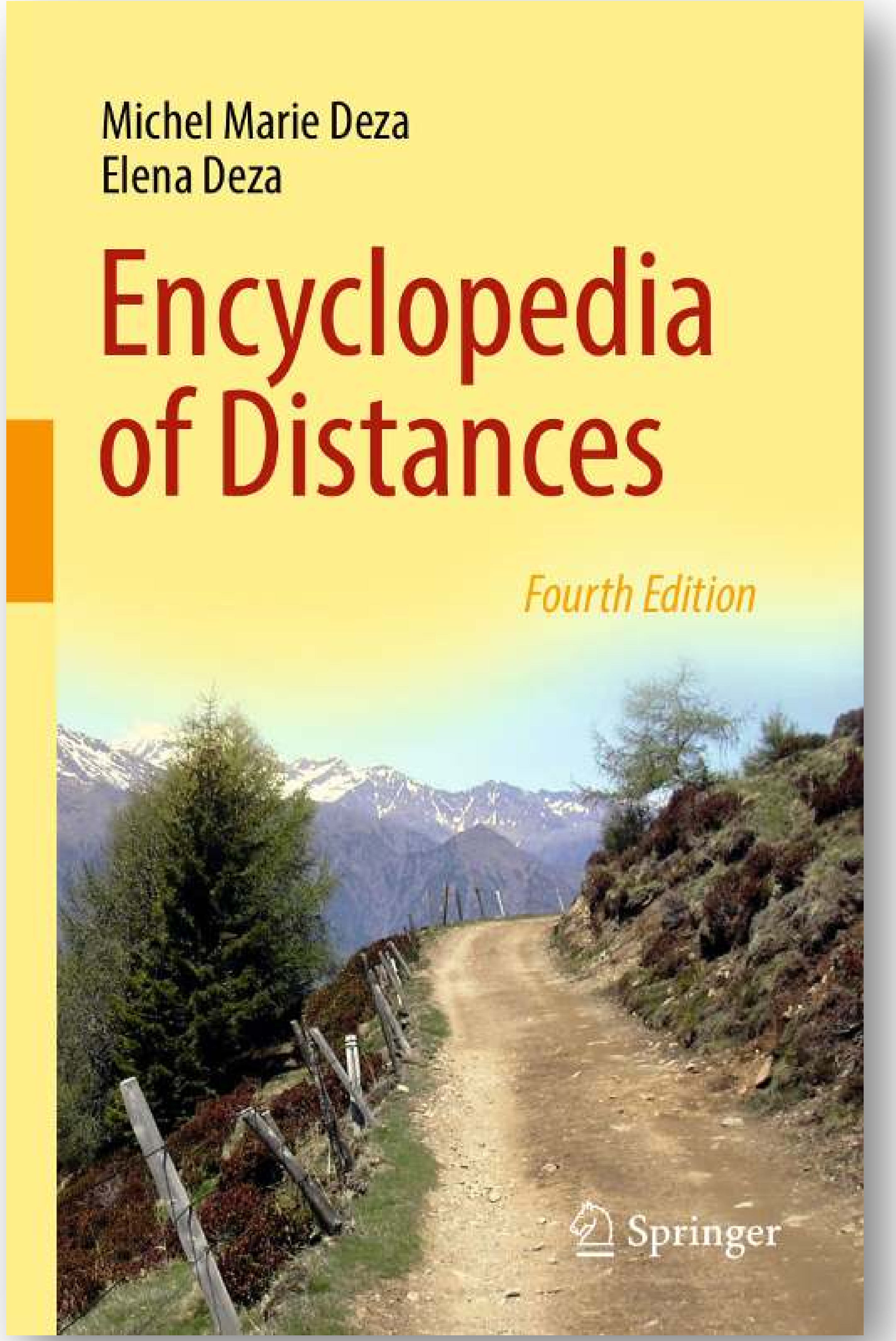
Schöch, Christof. "Quantitative Analyse." Digital Humanities. JB Metzler, Stuttgart, 2017. 279-298.

More Distance Metrics

There are a lot more distances!

Every data type needs their own distance metric, for example:

- distances between geographic coordinates
- distances between text documents
- distances between graphs or nodes in graphs
- …



Strategies of Clustering

Hierarchical Agglomerative Clustering	Point assignment
Each point is in its own cluster	Initial clusters are estimated
Clusters are combined based on their "closeness"	Points are considered in some order
Combination stops when undesirable clusters occur	Points are assigned to clusters into which they best fit

Leskovec, J., Rajaraman, A. and Ullman, J.D., 2020. *Mining of massive data sets*. Cambridge university press.

Examples: Hierarchical Clustering

```
WHILE more than one cluster left
DO
      pick the best two clusters to merge
      combine those two clusters into one cluster
END
```

Examples: Hierarchical Clustering

```
WHILE more than one cluster left
DO
      pick the best two clusters to merge
      combine those two clusters into one cluster
END
```

How will clusters be represented?

How will we choose which clusters to merge?

This is the agglomerative approach (bottom up).
A divisive approach exists as well which starts
with one cluster that is recursively split

Hierarchical Clustering: Represent Clusters

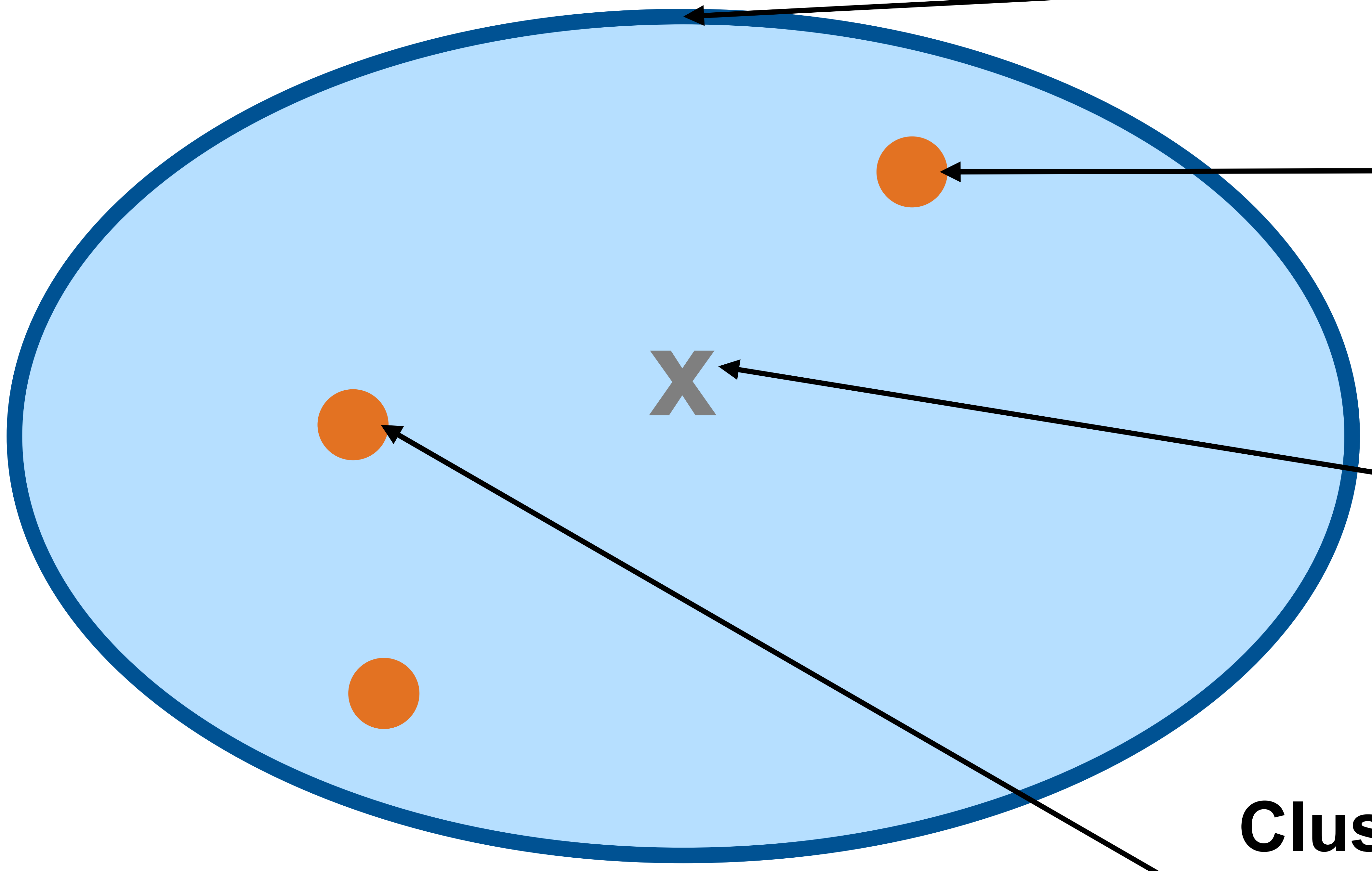
We need to combine nearest/closest clusters.

Key question: how to represent the „location" of each cluster
to tell which pair of clusters is closest?

In Euclidean spaces: each cluster has an average of its points – the **centroid**

In Non-Euclidean spaces:
Only „locations" are the points themselves
We do not have an average of points
Choose a **clustroid** which is a point closest to other points

Centroids and Clustroids



Cluster on 3 points

(Data)Point

Centroid
The average of all points in the cluster. It is an **artificial** point.

Clustroid
An **existing** point that is closest to all other points in the cluster.

Determining the clustroid, i.e. the point being closest to all other points:
- Point with smallest maximum distance to other points
- Point with smallest average distance to other points
- More complicated notions

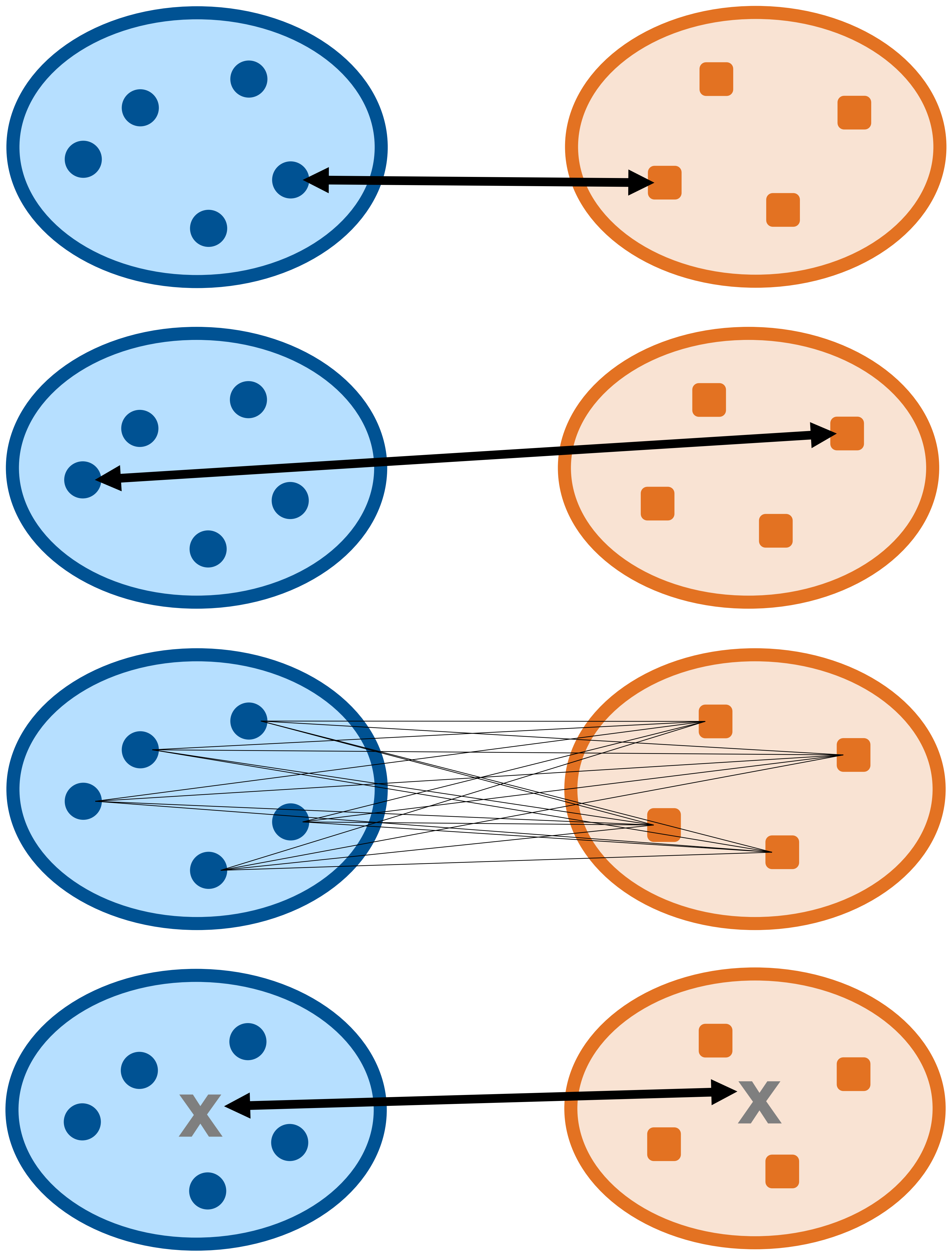
Hierarchical Clustering: Compare Clusters

Single-linkage:
Minimum distance (roughly maximum similarity)

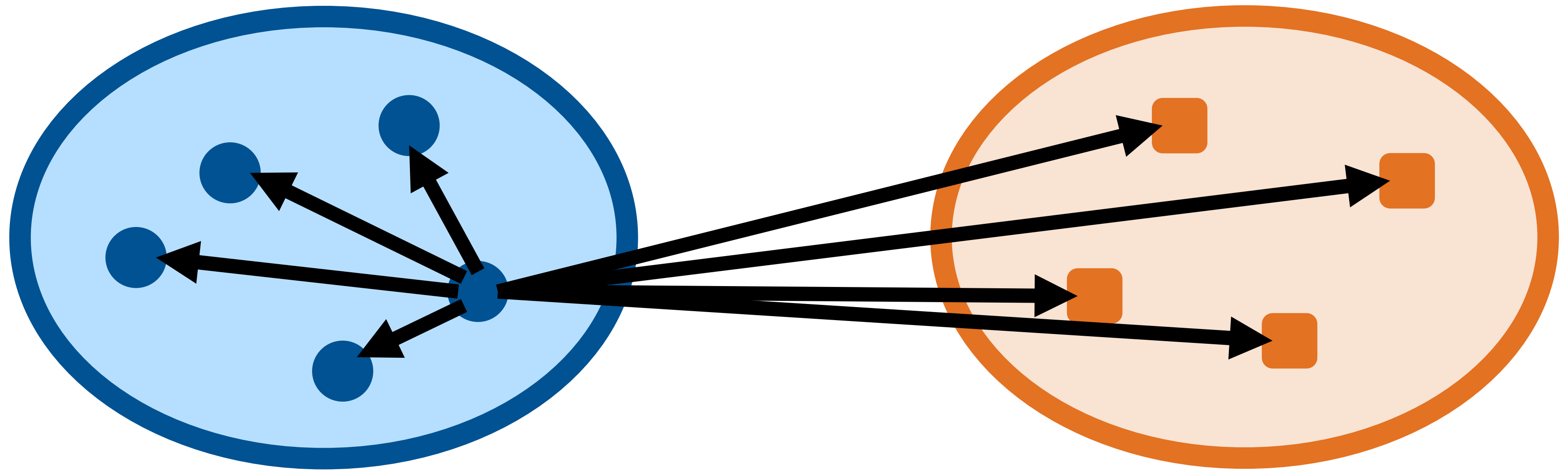
Complete-linkage:
Maximum distance (roughly minimum similarity)

Average-linkage:
Average distance

Centroid-linkage:
Distance between cluster centroids.
Only for Euclidean spaces.

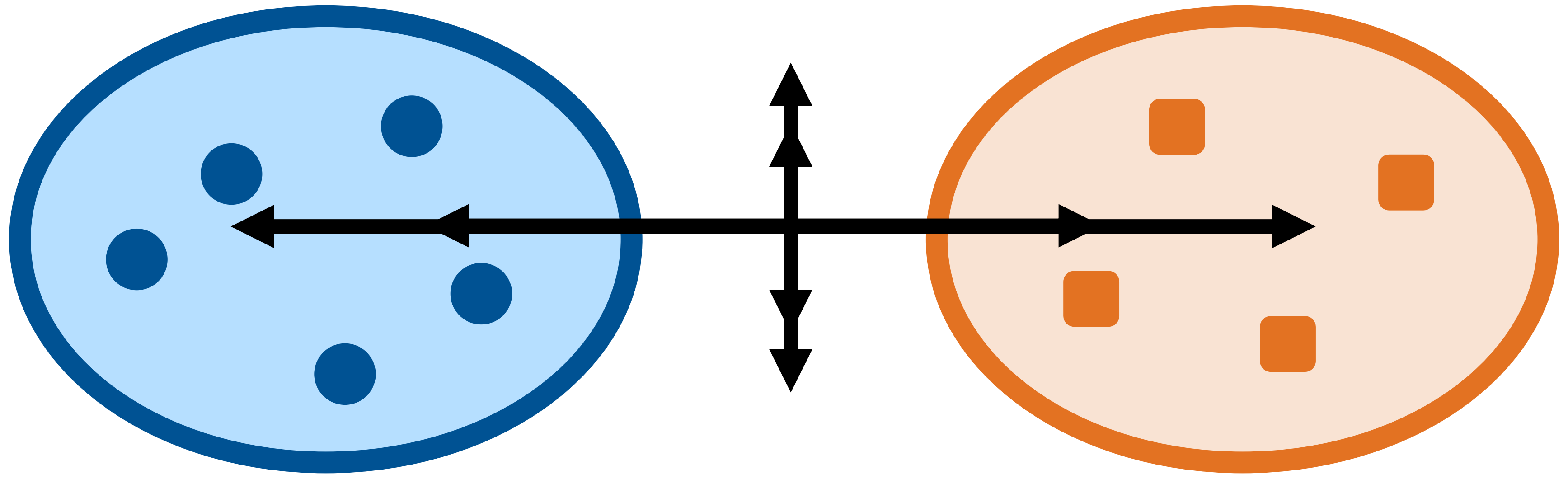


Hierarchical Clustering: Compare Clusters



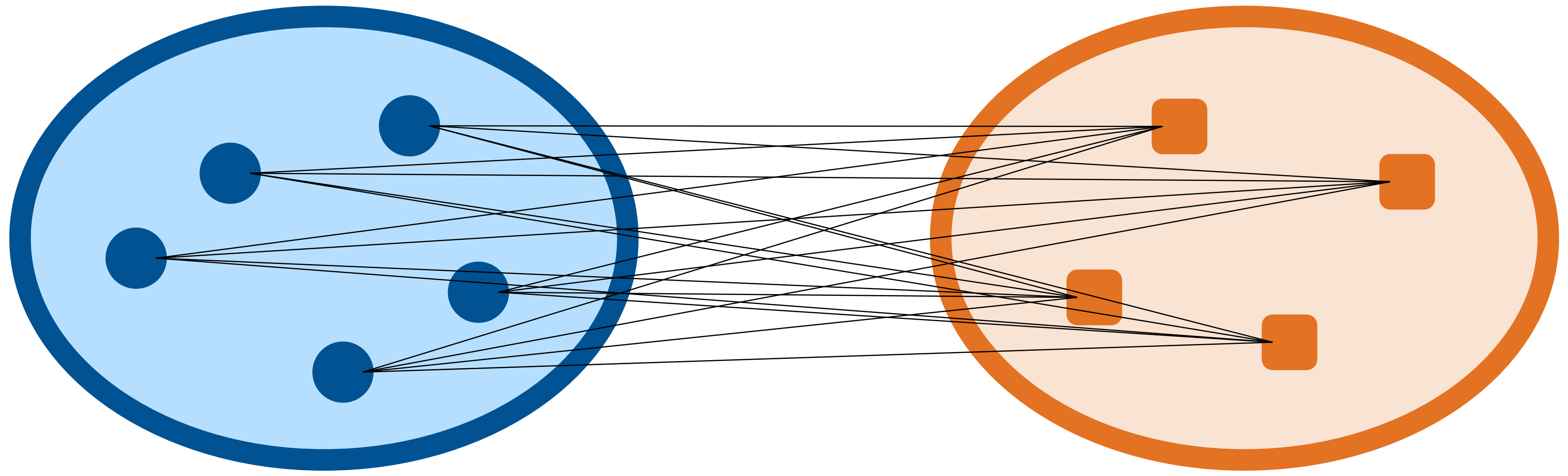
Min-Max-linkage:

Best maximum distance (best minimum similarity)



Ward-linkage:

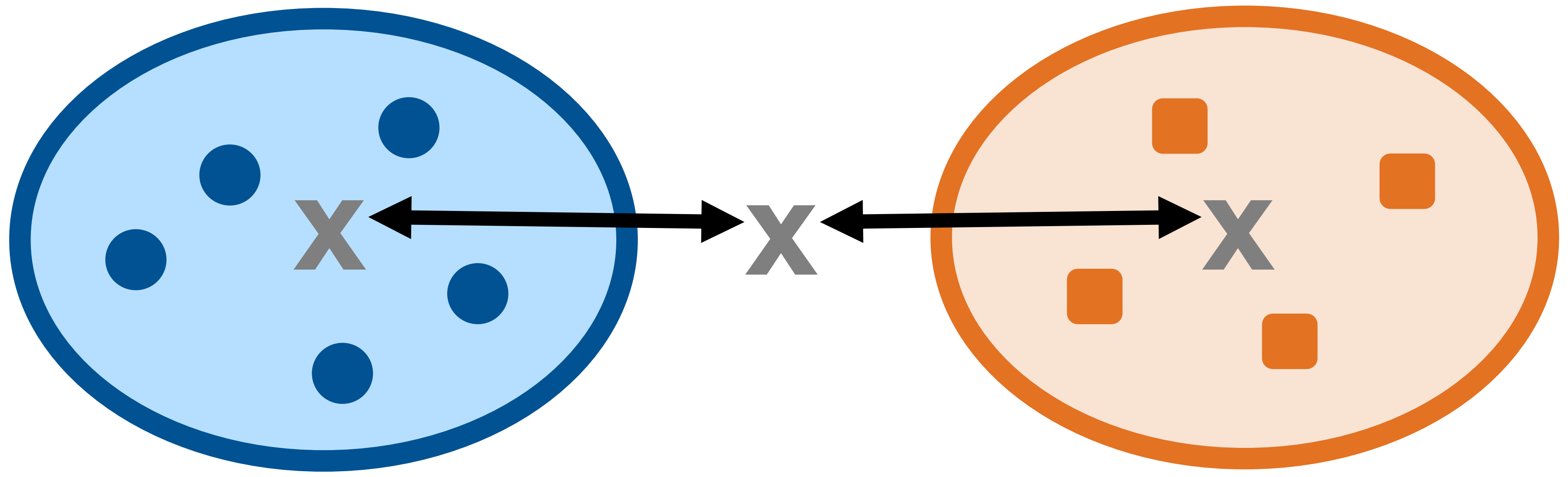
Minimum increase of squared error



McQuitty (WPGMA):

Average distance to the previous two clusters.

Recursive definition

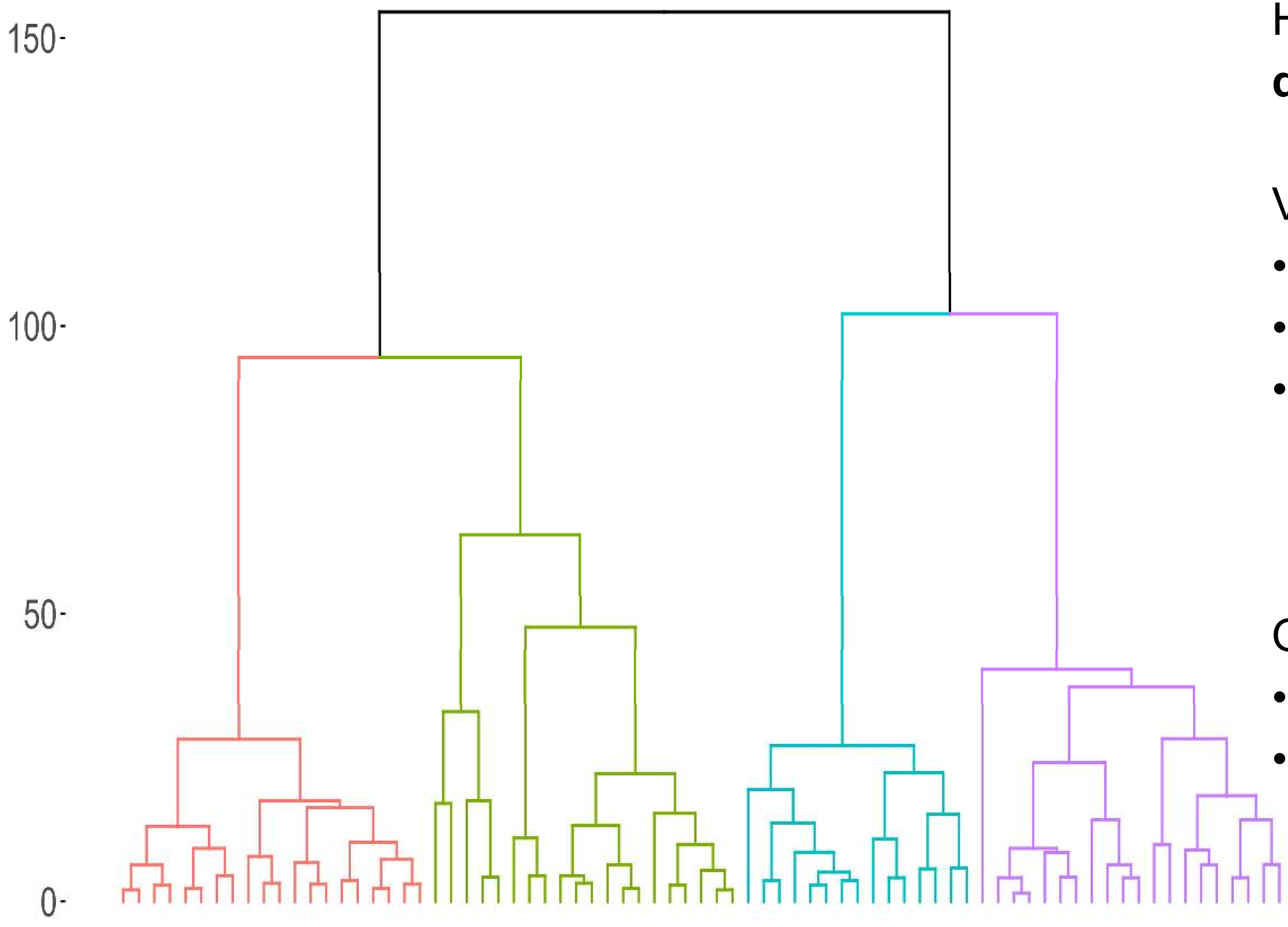


Median-linkage:

Distance between cluster midpoints.

Recursive definition

From Dendrograms to Clusters



Hierarchical clustering outputs a **dendrogram**, but not „clusters“

Various strategies to select a clustering:

- Choose visually interesting branches
- Cut tree horizontally
- Other scientific approaches using cluster distances, densities, sizes, clustered objects, ….

Questions:

- Are clusters allowed to overlap?
- How to handle outliers?

Hierarchical Clustering: why and why not?

Pro:

- Very general. Supports any distance metric
- Number of clusters doesn‘t need to be known beforehand

Contra:

- Unbalanced cluster sizes
- Outliers
- Slow for large datasets

Examples: k-Means Clustering

```
Place each point in the cluster whose current centroid is the nearest
WHILE points are moving between clusters and centroids not stabilized
DO
      Update locations of centroids of k clusters
      Reassign all points to their closest centroid
END
```

Disclaimer:
This is the standard k-means algorithm proposed by Lloyd (1982)
It is, however, not the most efficient variant..

Lloyd, Stuart. "Least squares quantization in PCM." IEEE transactions on information theory 28.2 (1982): 129-137.

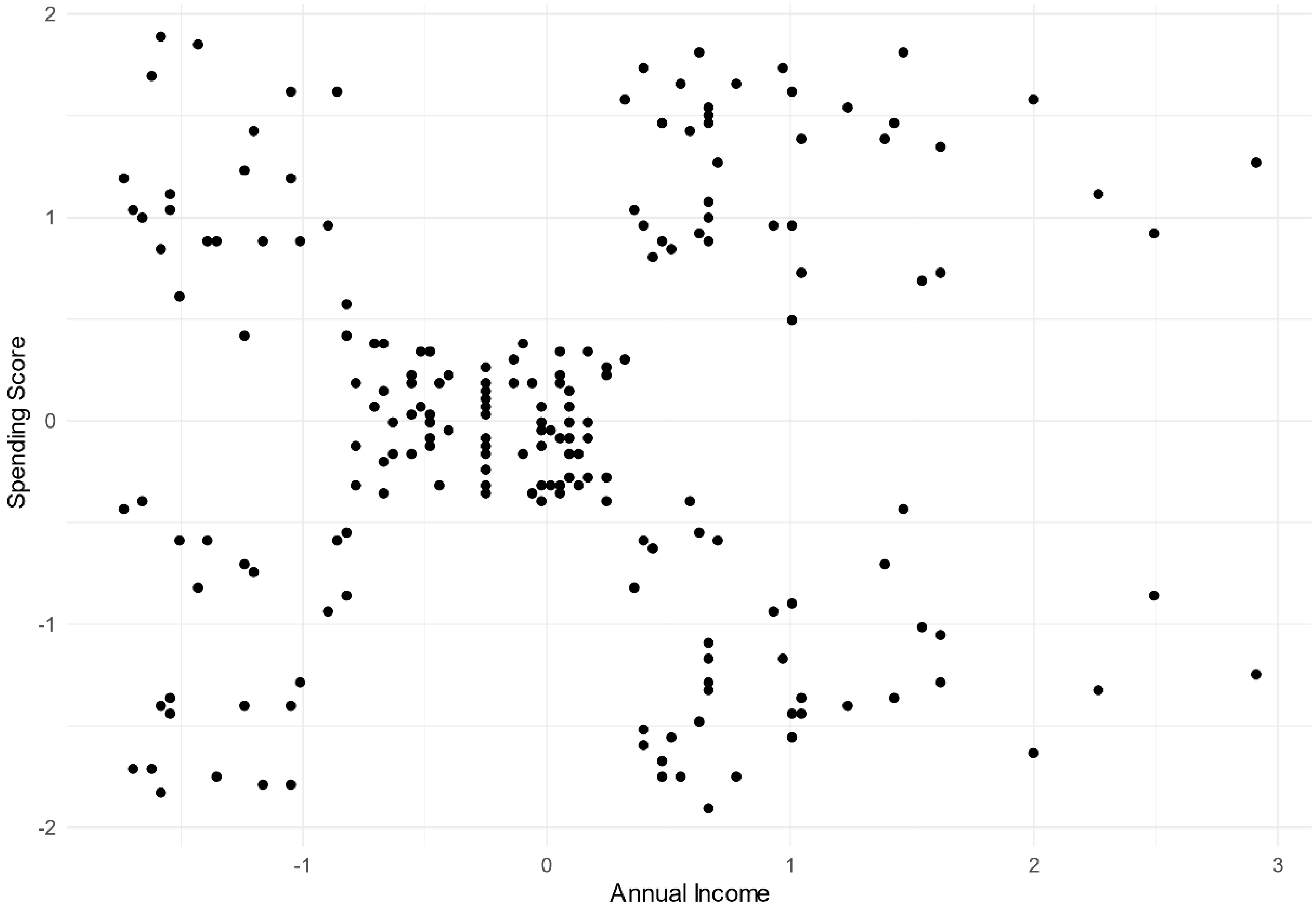
Examples: k-Means Clustering

Clusters represented by their arithmetic mean

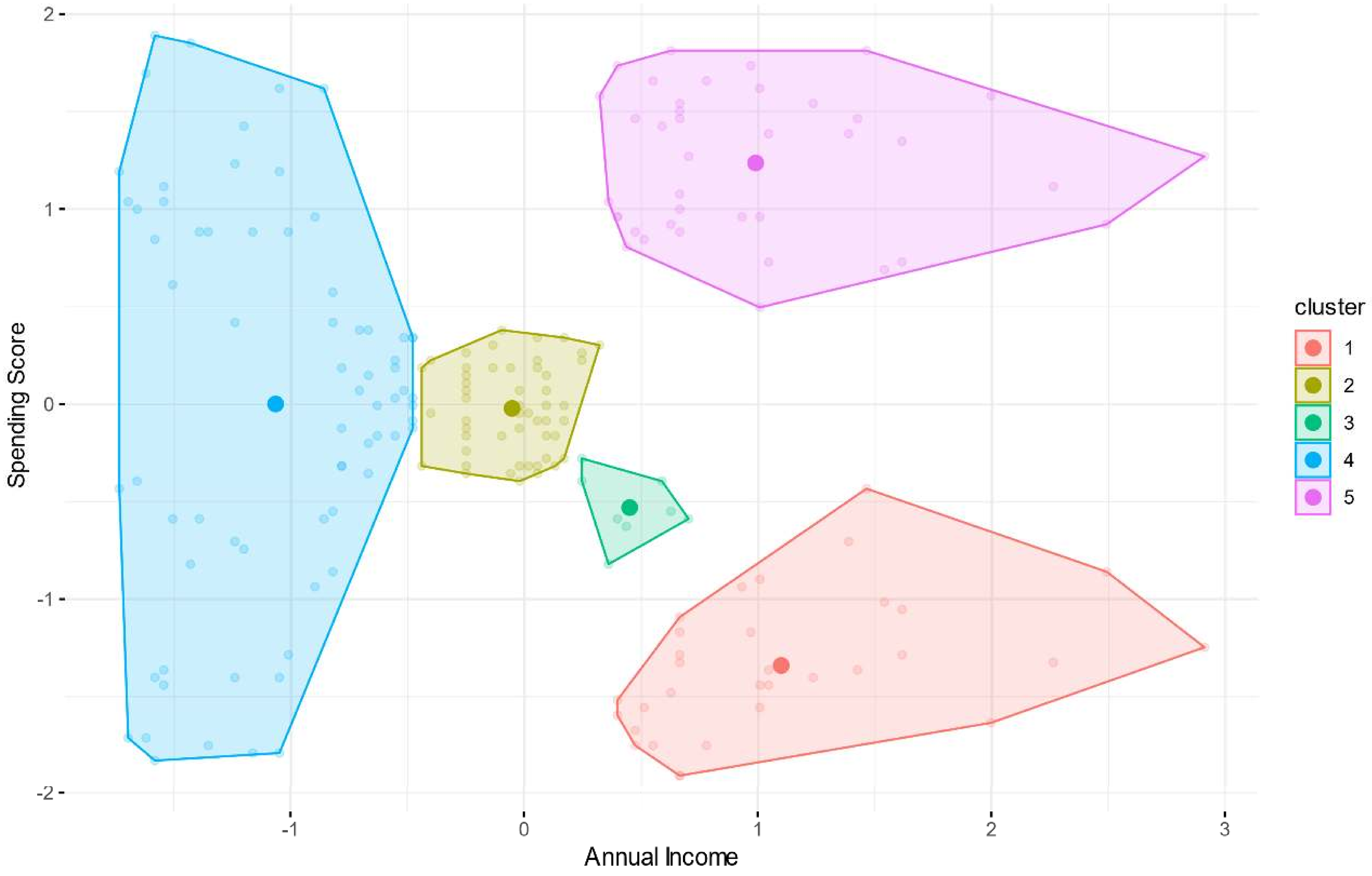
Optimizes „least squared errors“, i.e. minimizes distance of points from centroids

That‘s why k-means is bound to Euclidean distance in Euclidean spaces

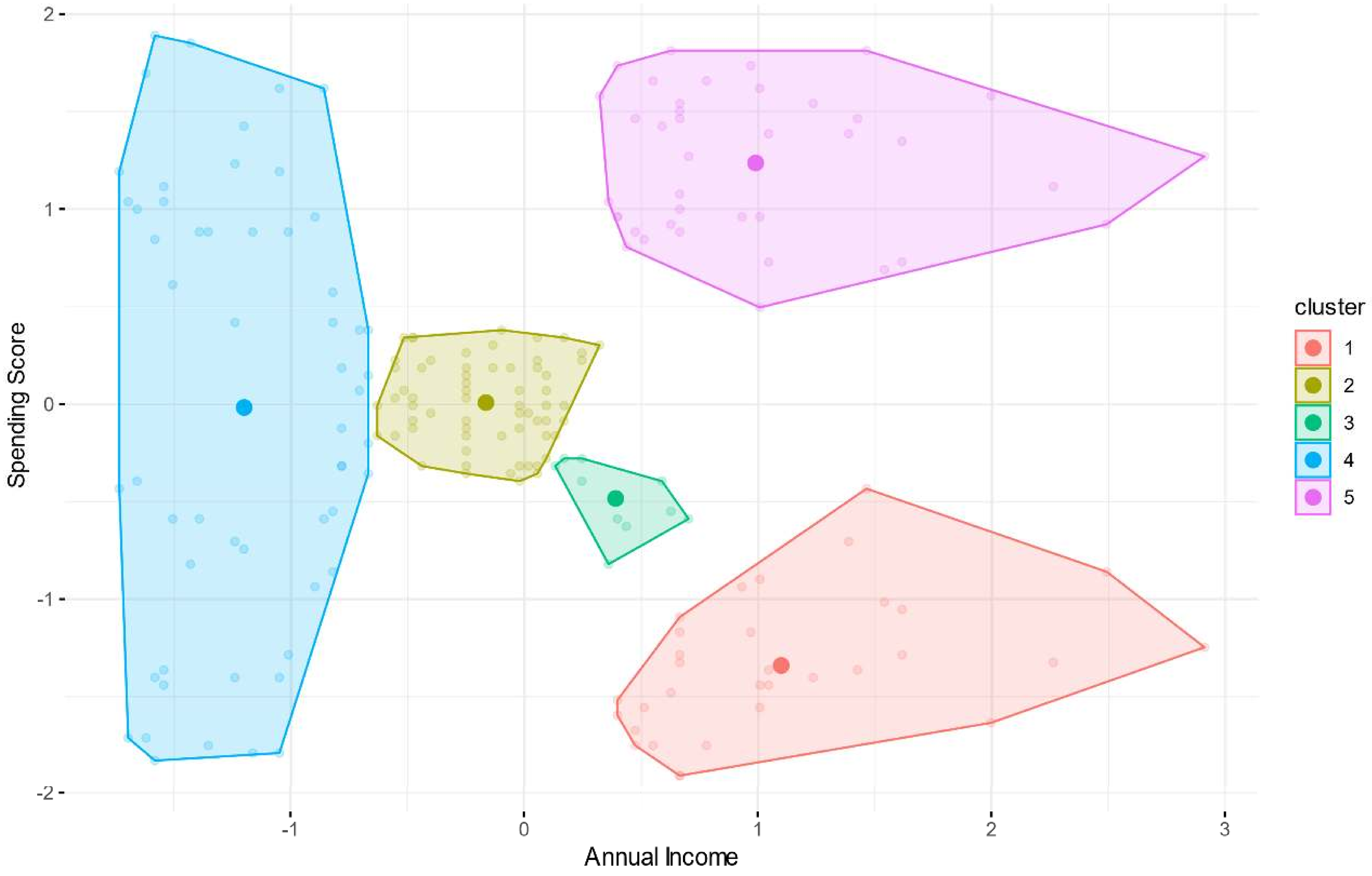
Examples: k-Means Clustering



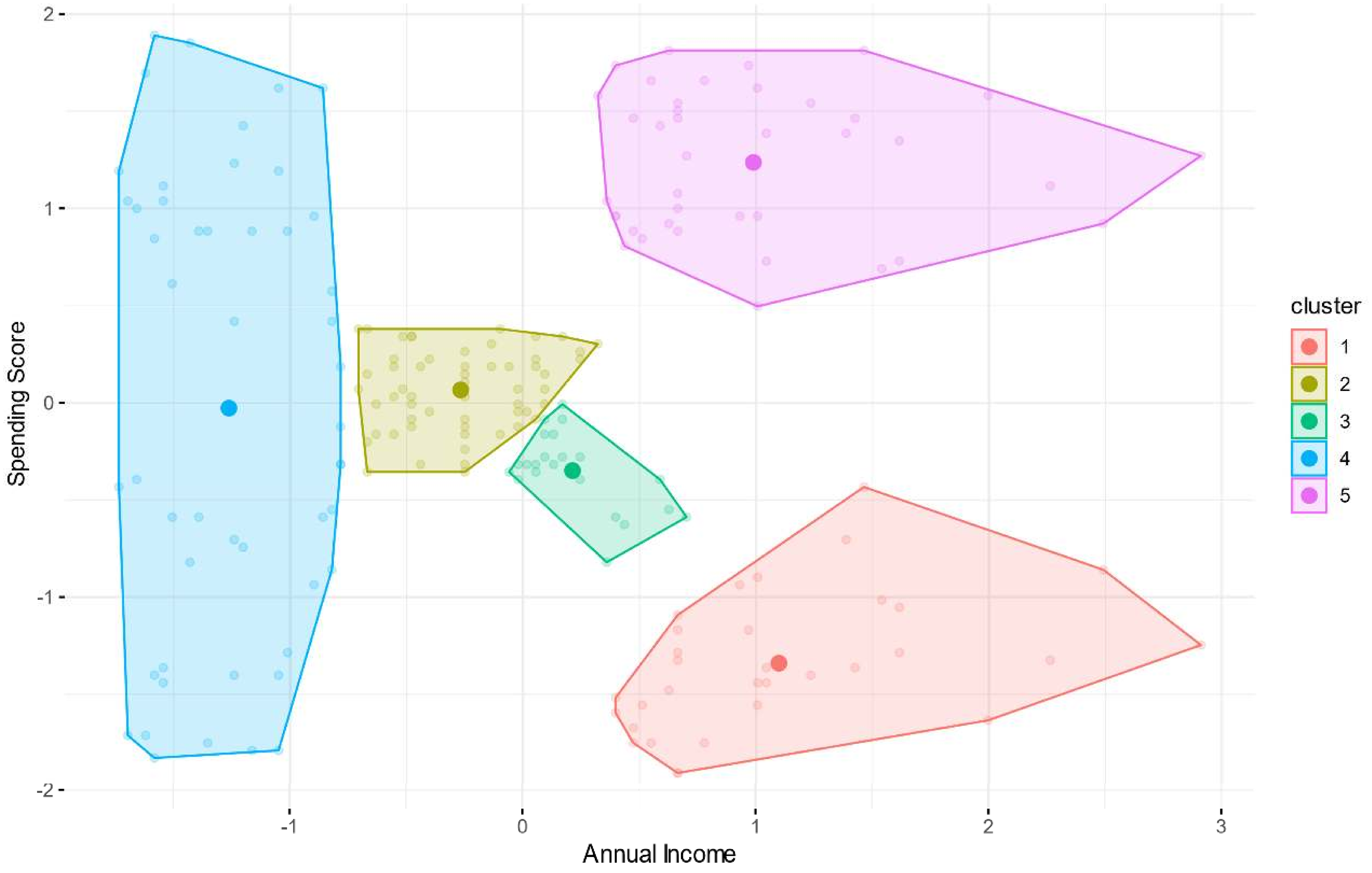
Examples: k-Means Clustering



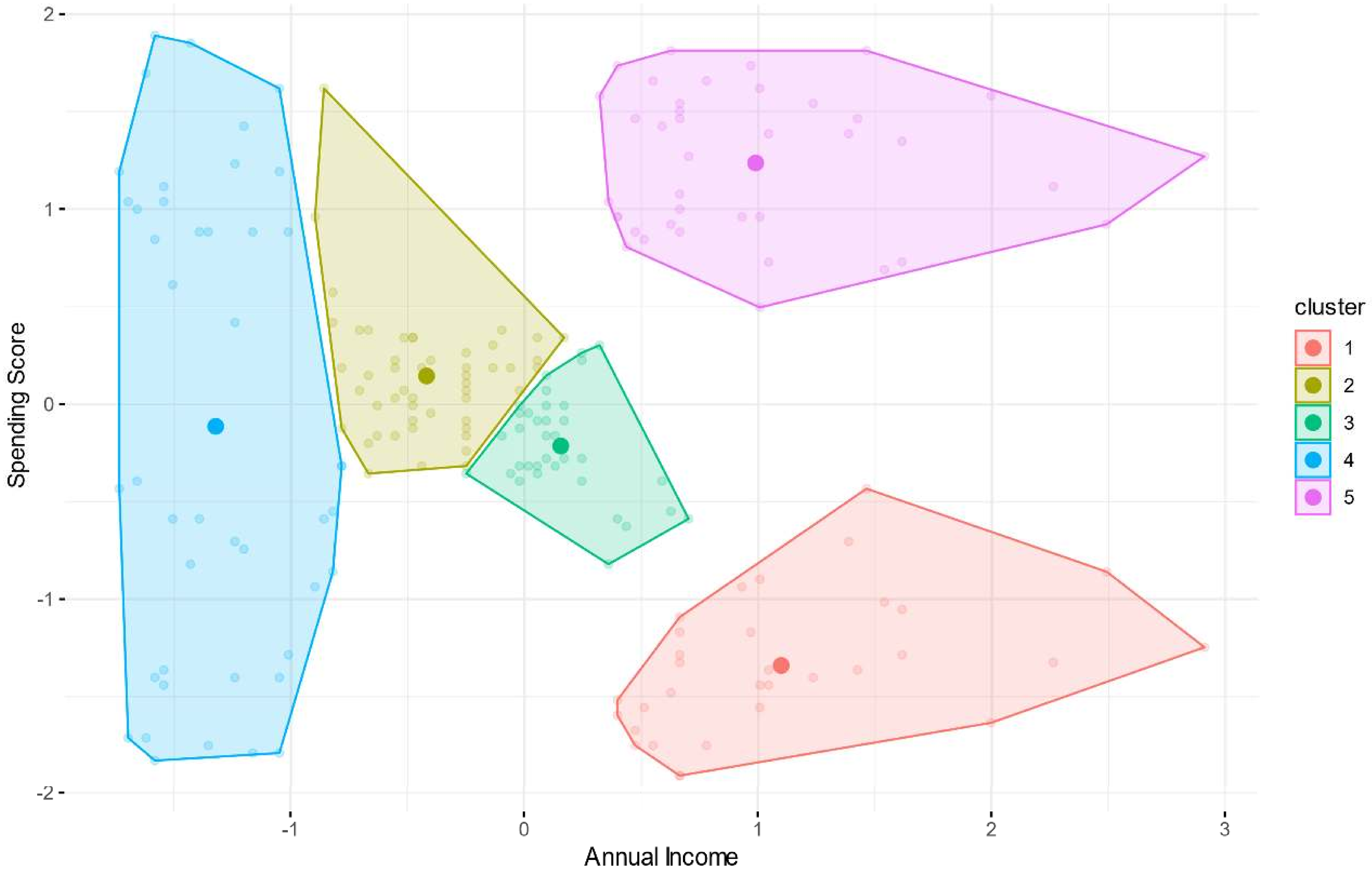
Examples: k-Means Clustering



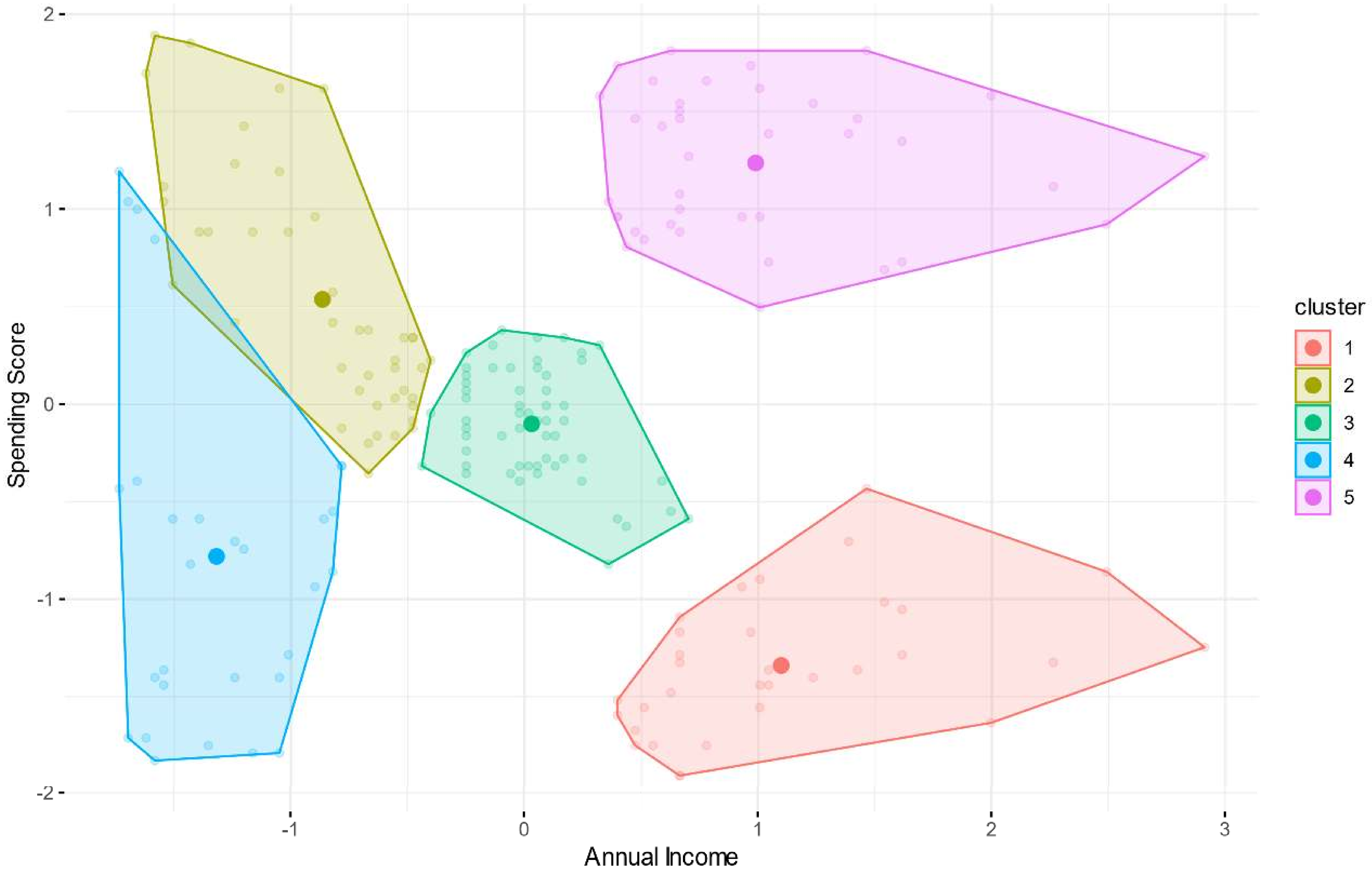
Examples: k-Means Clustering



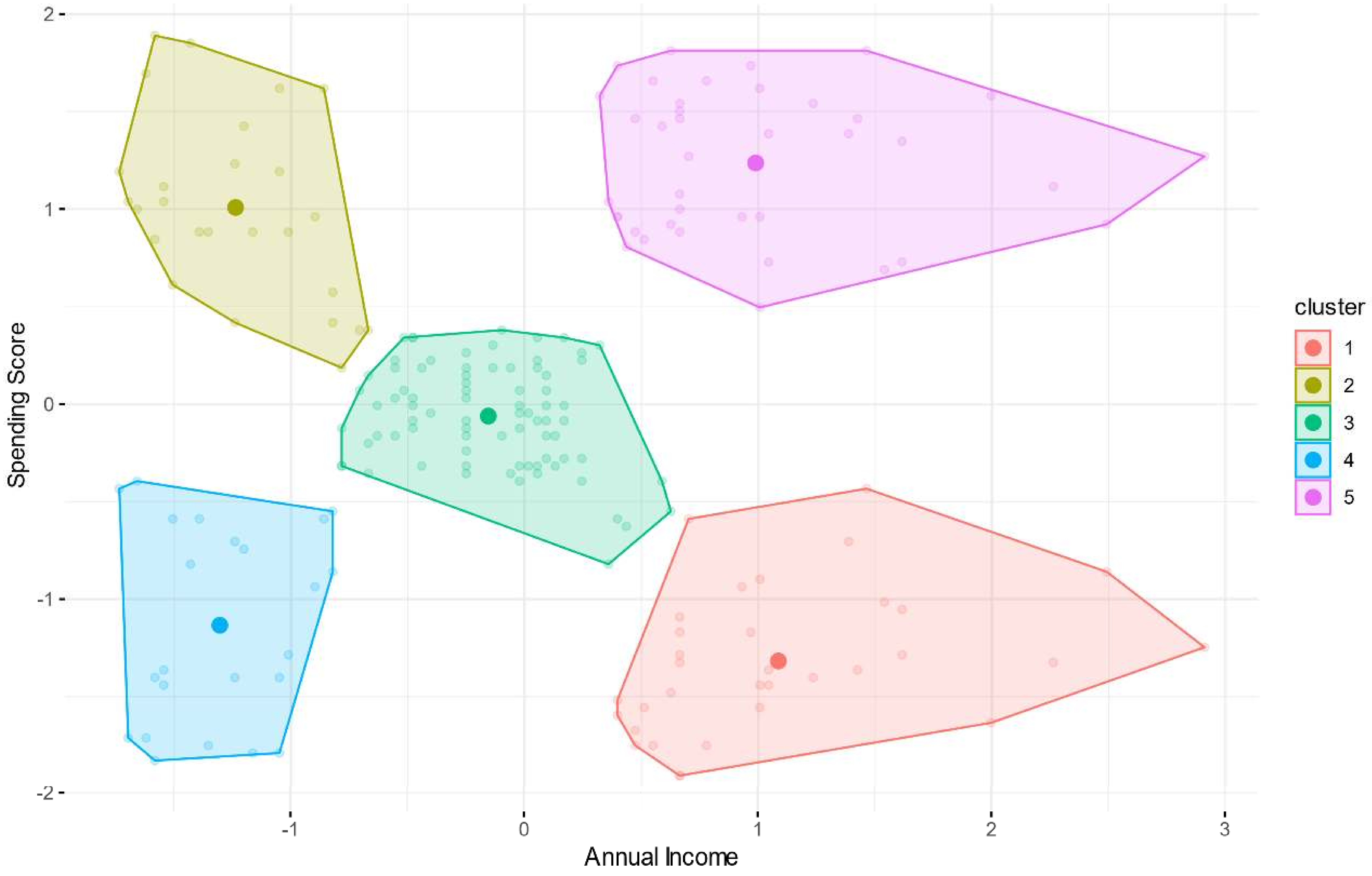
Examples: k-Means Clustering



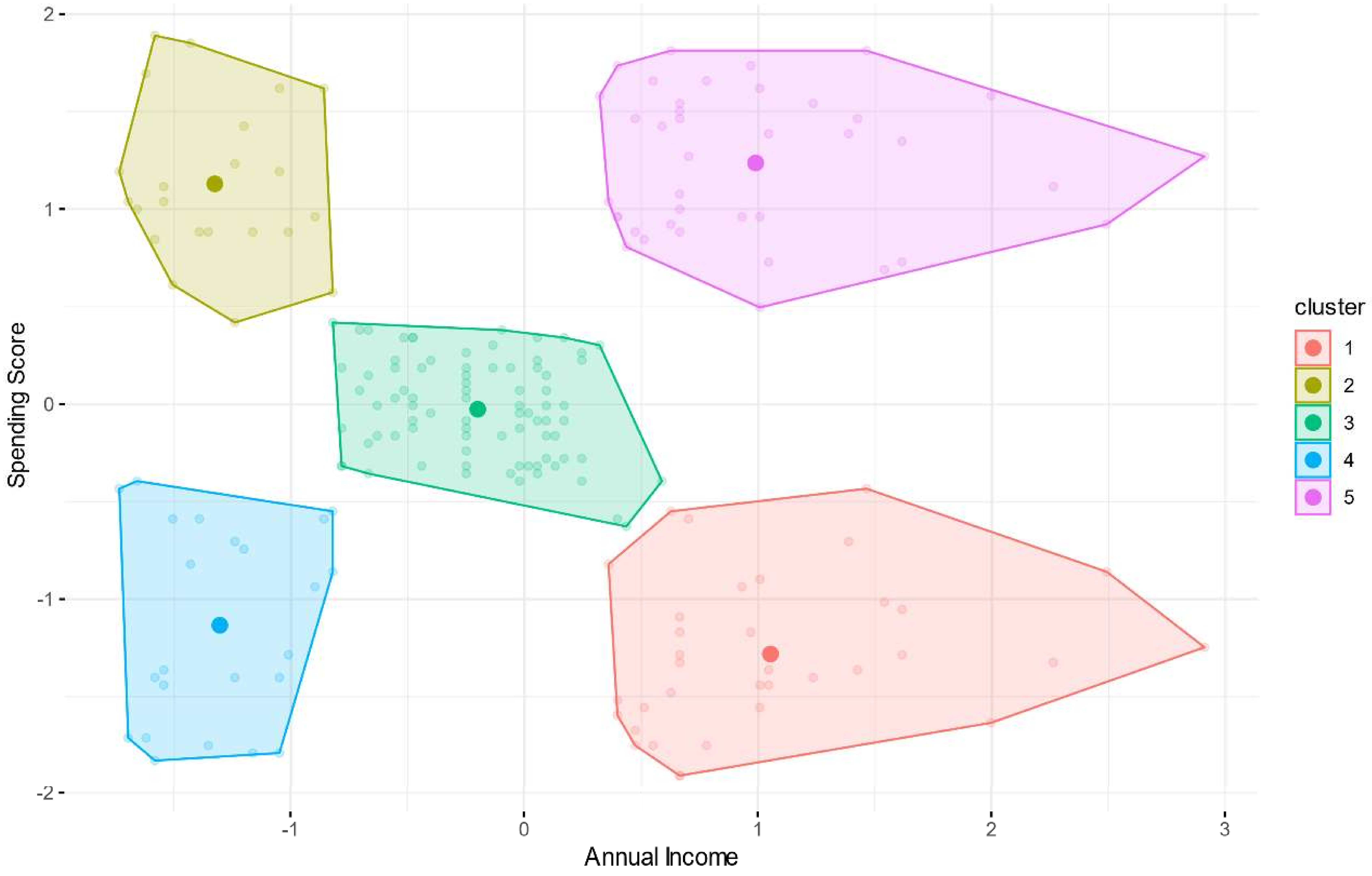
Examples: k-Means Clustering



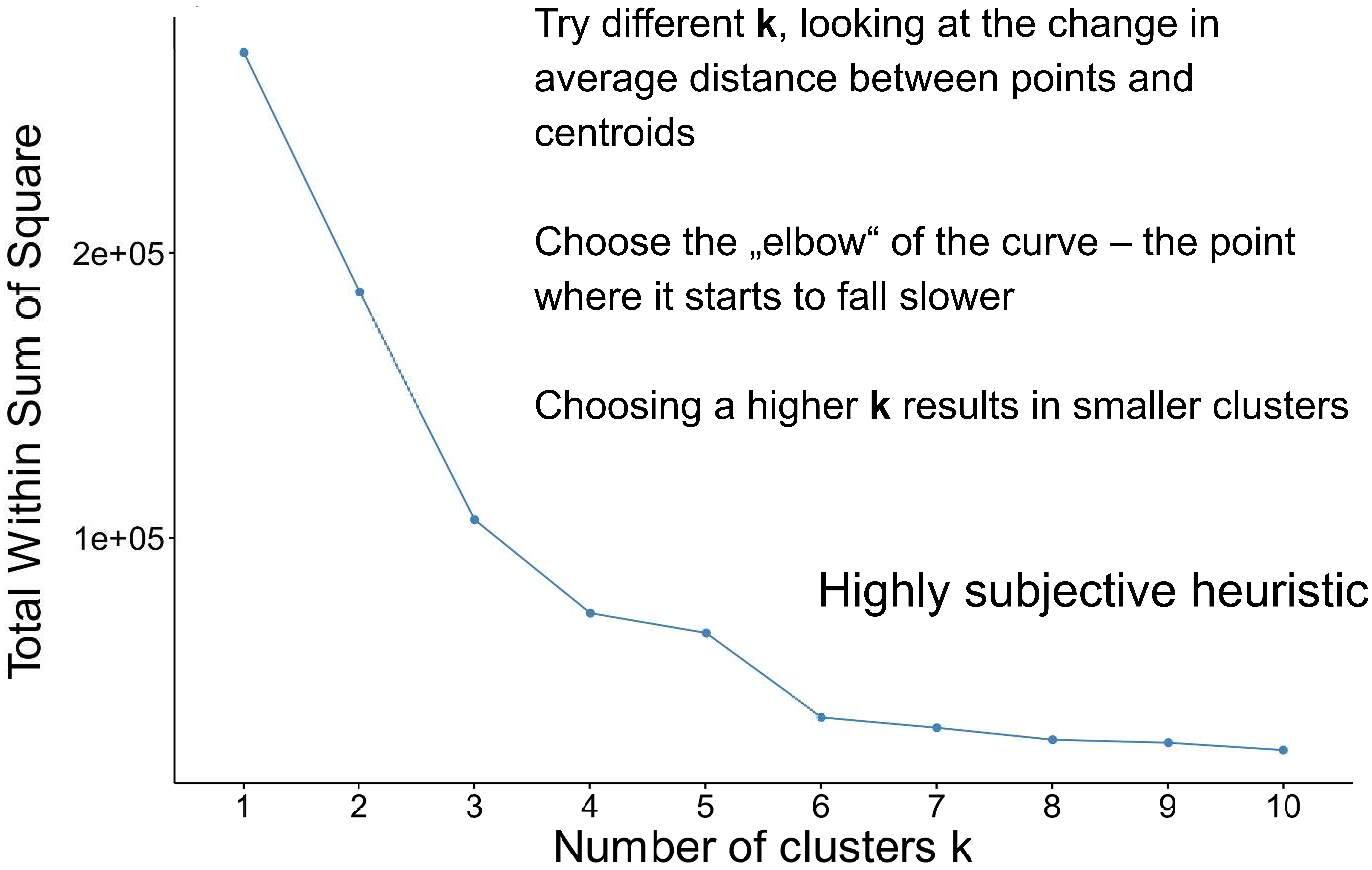
Examples: k-Means Clustering



Examples: k-Means Clustering



How to choose *k*: Elbow Criterion



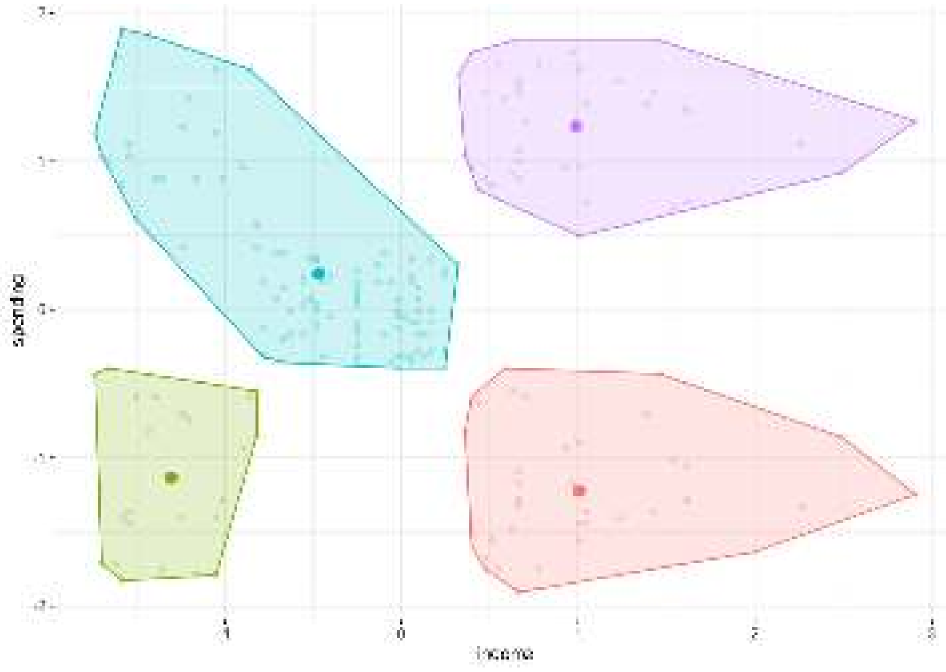
Try different **k**, looking at the change in average distance between points and centroids

Choose the „elbow“ of the curve – the point where it starts to fall slower

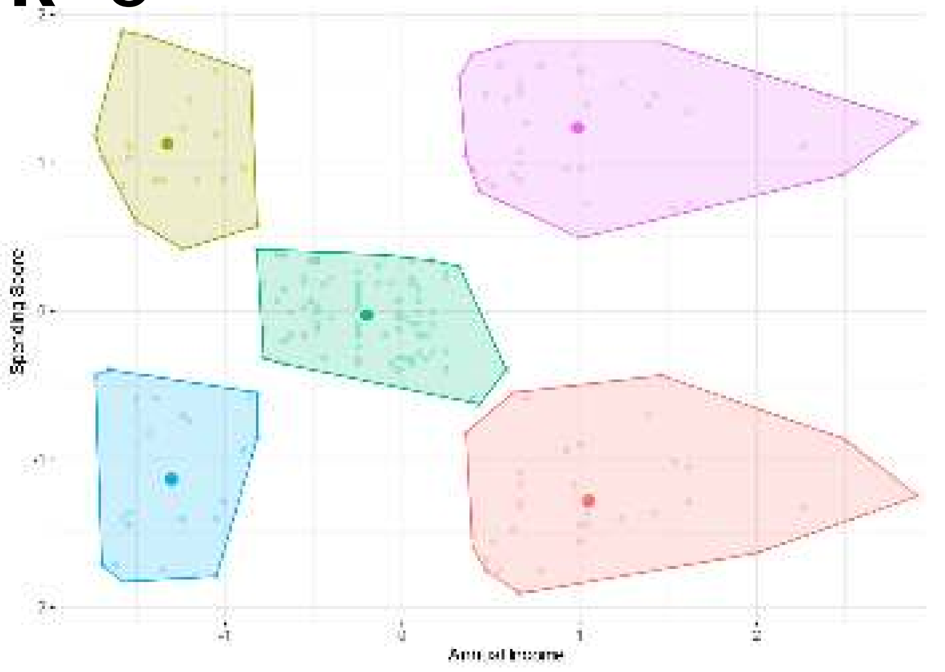
Choosing a higher **k** results in smaller clusters

Highly subjective heuristic

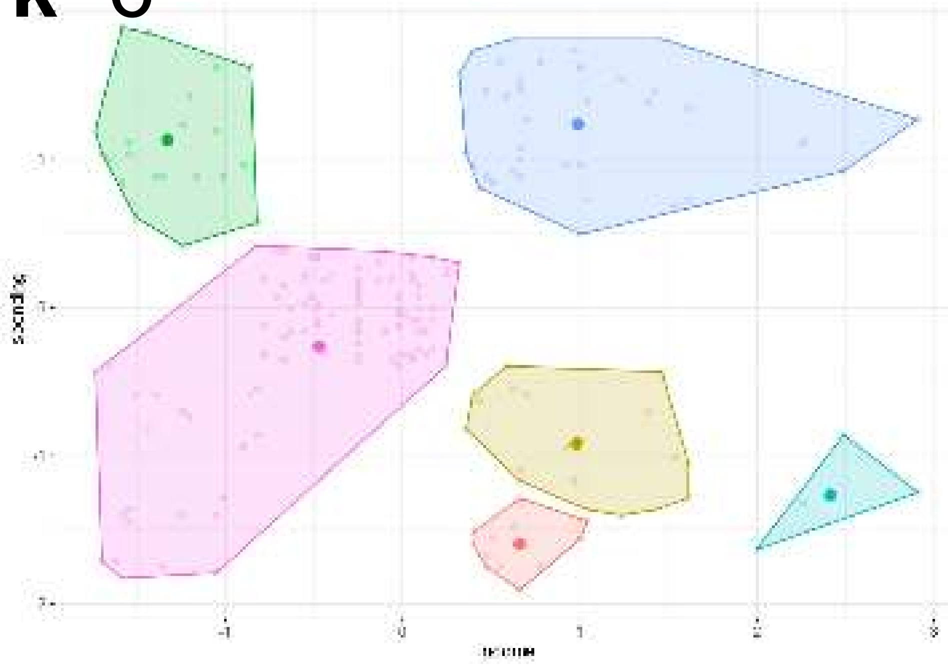
k=4



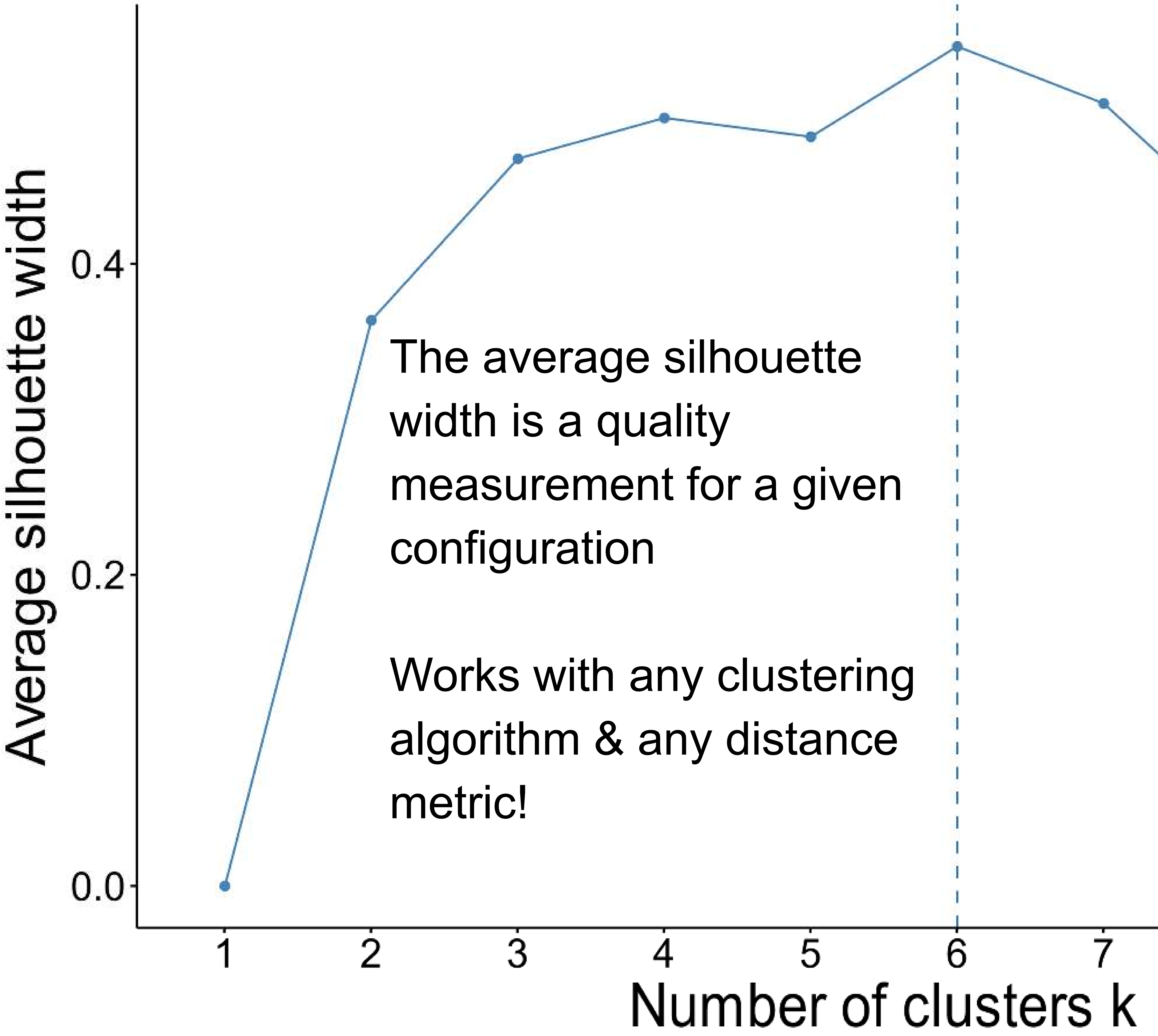
k=5



k=6



How to choose *k*: Silhouettes

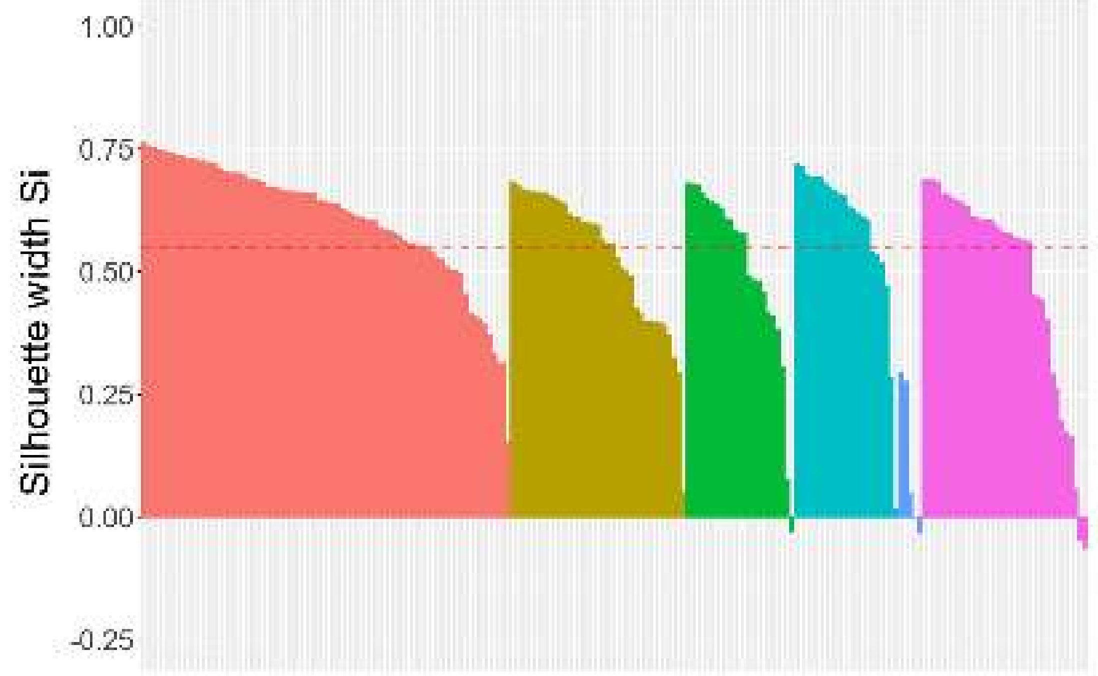


The average silhouette width is a quality measurement for a given configuration

Works with any clustering algorithm & any distance metric!

k=5

k=6



The silhouette value indicates how similar a point is to its own cluster compared to other clusters.

Silhouette values range from -1 to +1. High values mean that points are well matched to their own clusters and poorly matched to neighboring clusters.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Assessing Quality of Clustering

Meaningful clusters are highly subjective

Also, data is never exact or complete

Optimal results are maybe not the most useful

Modeling Decisions

No single „best“ setting for the general case

Expert decisions required on

- Feature selection
- The choice of clustering algorithm
- Parameters of algorithm
- Preprocessing and optimization techniques applied to data
- Distance measure suitable for the scenario
- Cluster quality criterion

Every configuration might yield different results!

Feature Selection

UNIVERSITÄT BAYREUTH

Describing objects is a careful process called *feature selection*

Information needs to be selected that describe the objects best for the task of interest

Producing redundancy in features should be avoided!

Feature Selection

Formulate characteristics that help distinguishing objects.

For spam-detection: find words or combinations of words that indicate a mail being spam.

Spam: Wholesale Fashion Watches -57% today. Designer watches for cheap ...
Spam: You can buy ViagraFr$1.85 All Medications at unbeatable prices! ...
Spam: WE CAN TREAT ANYTHING YOU SUFFER FROM JUST TRUST US ...
Spam: Sta.rt earn*ing the salary yo,u d-eserve by o’btaining the prope,r crede’ntials!

Ham: The practical significance of hypertree width in identifying more ...
Ham: Abstract: We will motivate the problem of social identity clustering: ...
Ham: Good to see you my friend. Hey Peter, It was good to hear from you. ...
Ham: PDS implies convexity of the resulting optimization problem (Kernel Ridge ...

Curse of Dimensionality:

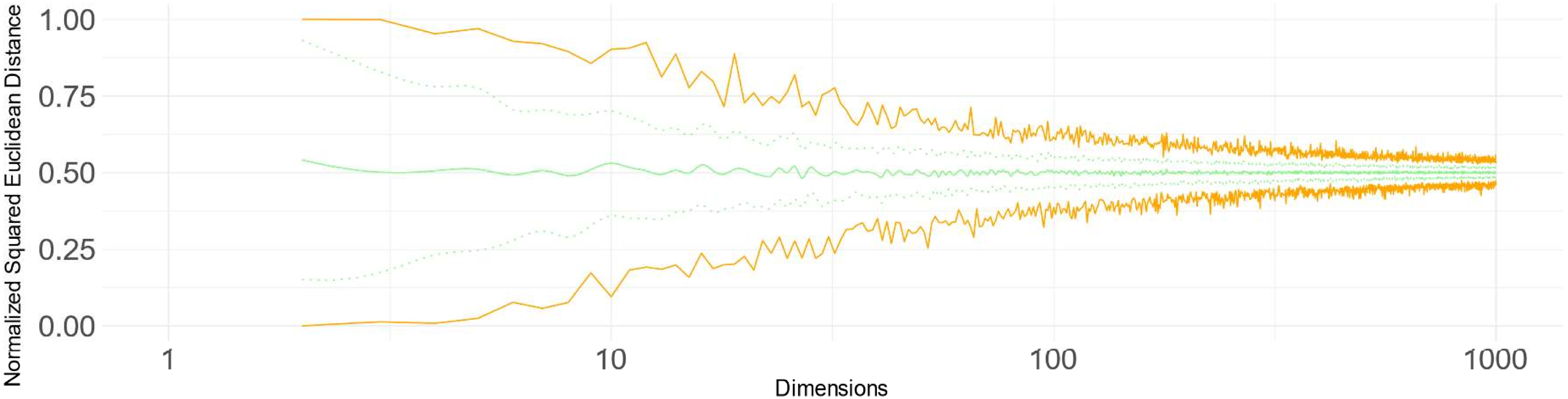
Including more features will improve classification *conceptually* but will render computation increasingly difficult.

Curse of Dimensionality

In high-dimensional spaces

- …almost all pairs of points are equally far away from one another
- …almost any two vectors are almost orthogonal

Variance in distances shrink



It will be hard to build clusters if there are almost no differences in distances

Normalization and Standardization

Normalizing variables means mapping values into a new interval, usually [0,1]

$$x'_i = \frac{x_i - \min(X_i)}{\max(X_i) - \min(X_i)}$$

Standardizing variables means to transform values to z-scores indicating divergence from mean (unit: standard variance)

$$x'_i = \frac{x_i - \mu(X_i)}{\sigma(X_i)}$$

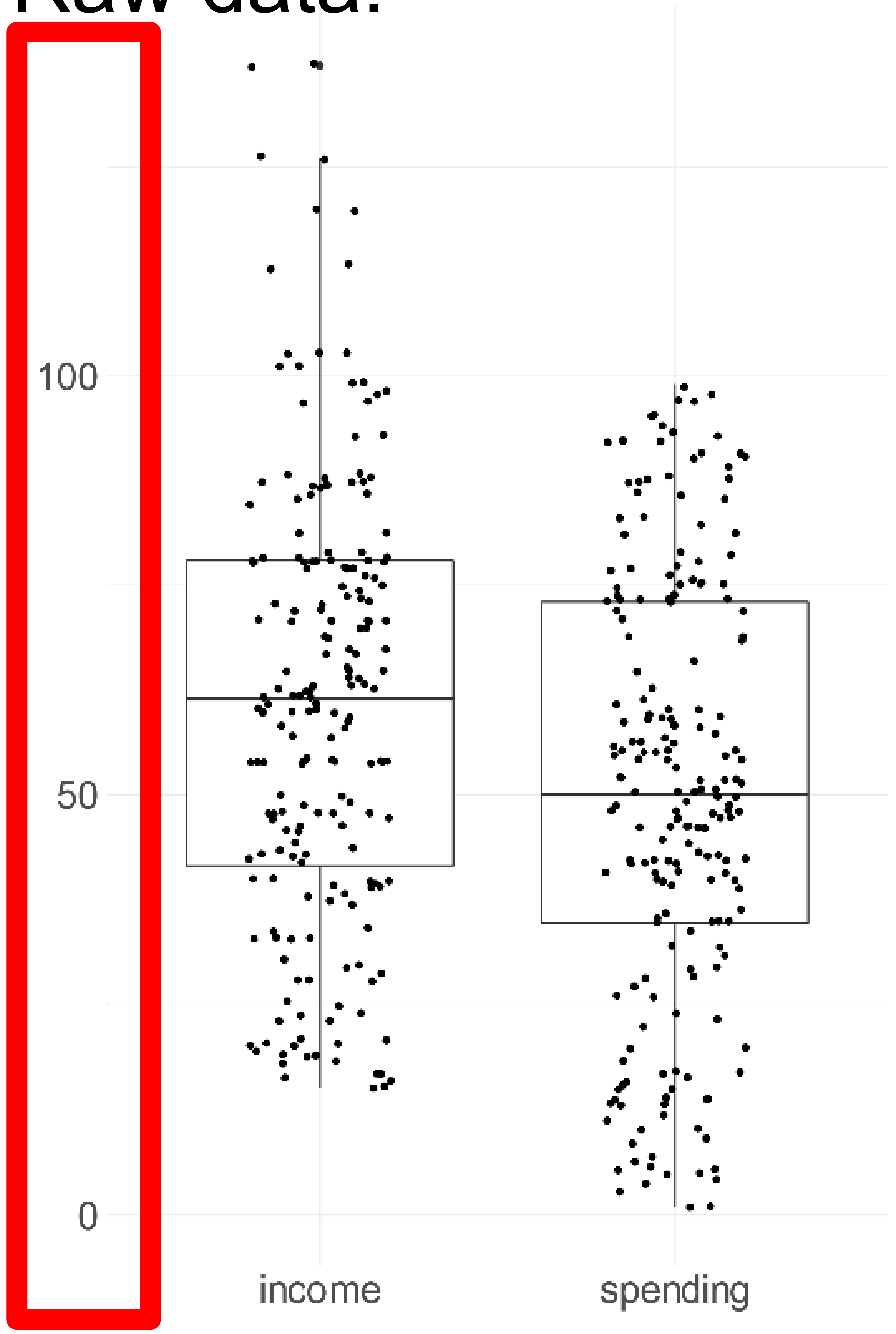
$\mu(X_i)$ is the arithmetic mean of variable X_i
$\sigma(X_i)$ is the standard deviation of variable X_i

Often required to be able to compare features.
Other (non-linear) transformations possible – e.g. to deal with skewness of variables

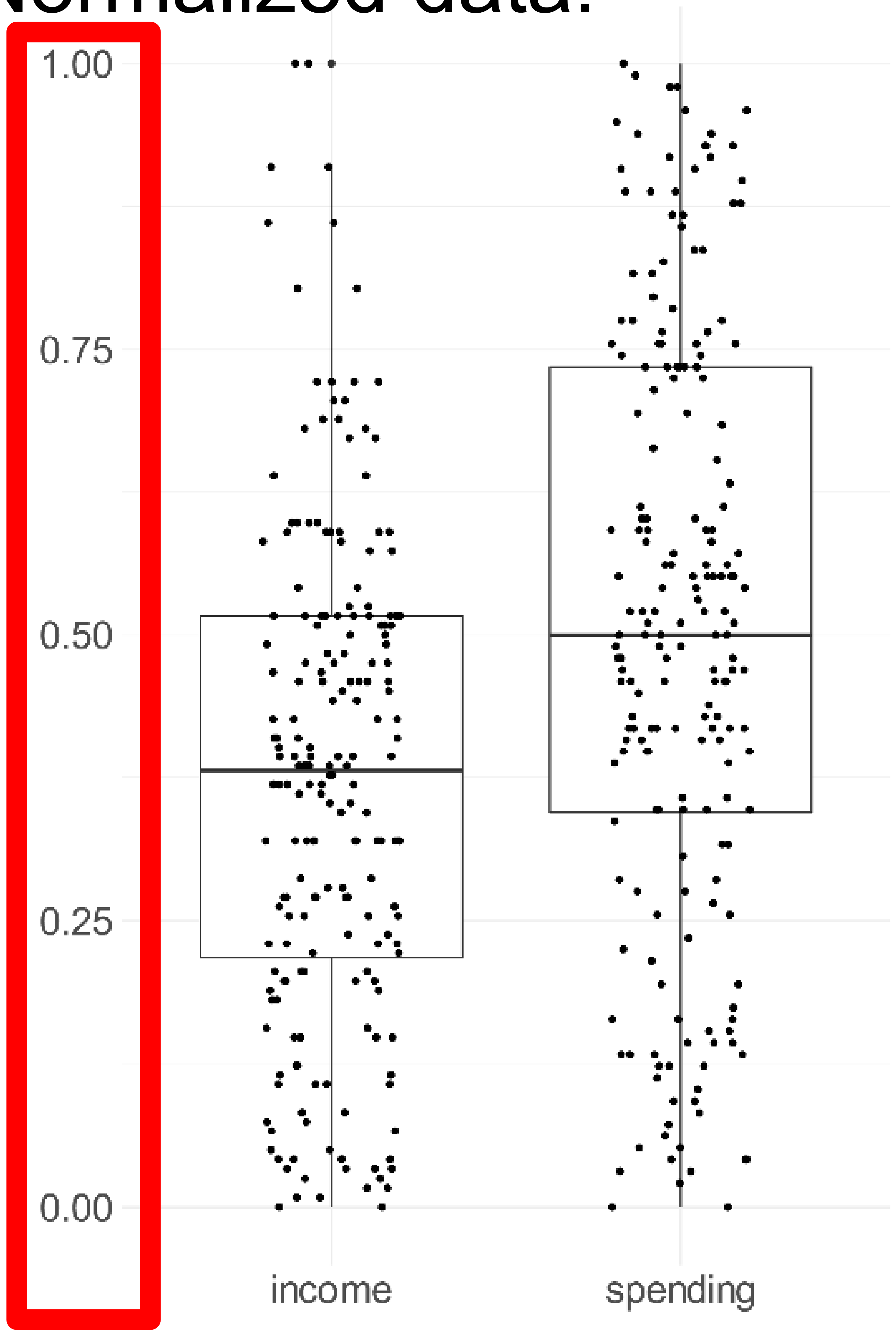
Normalized features matter „the same amount“

Normalization and Standardization

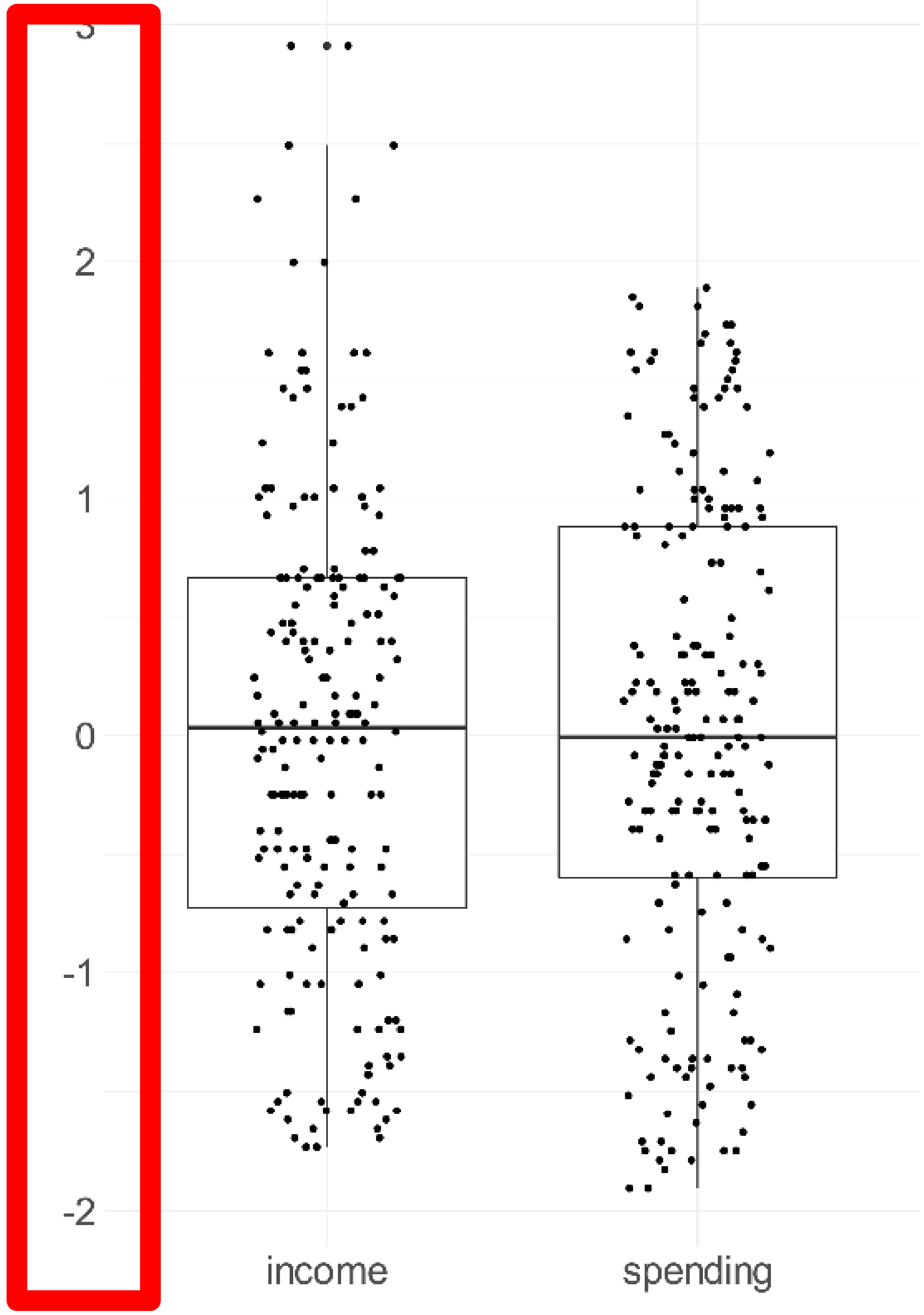
Raw data:



Normalized data:



Standardized data:



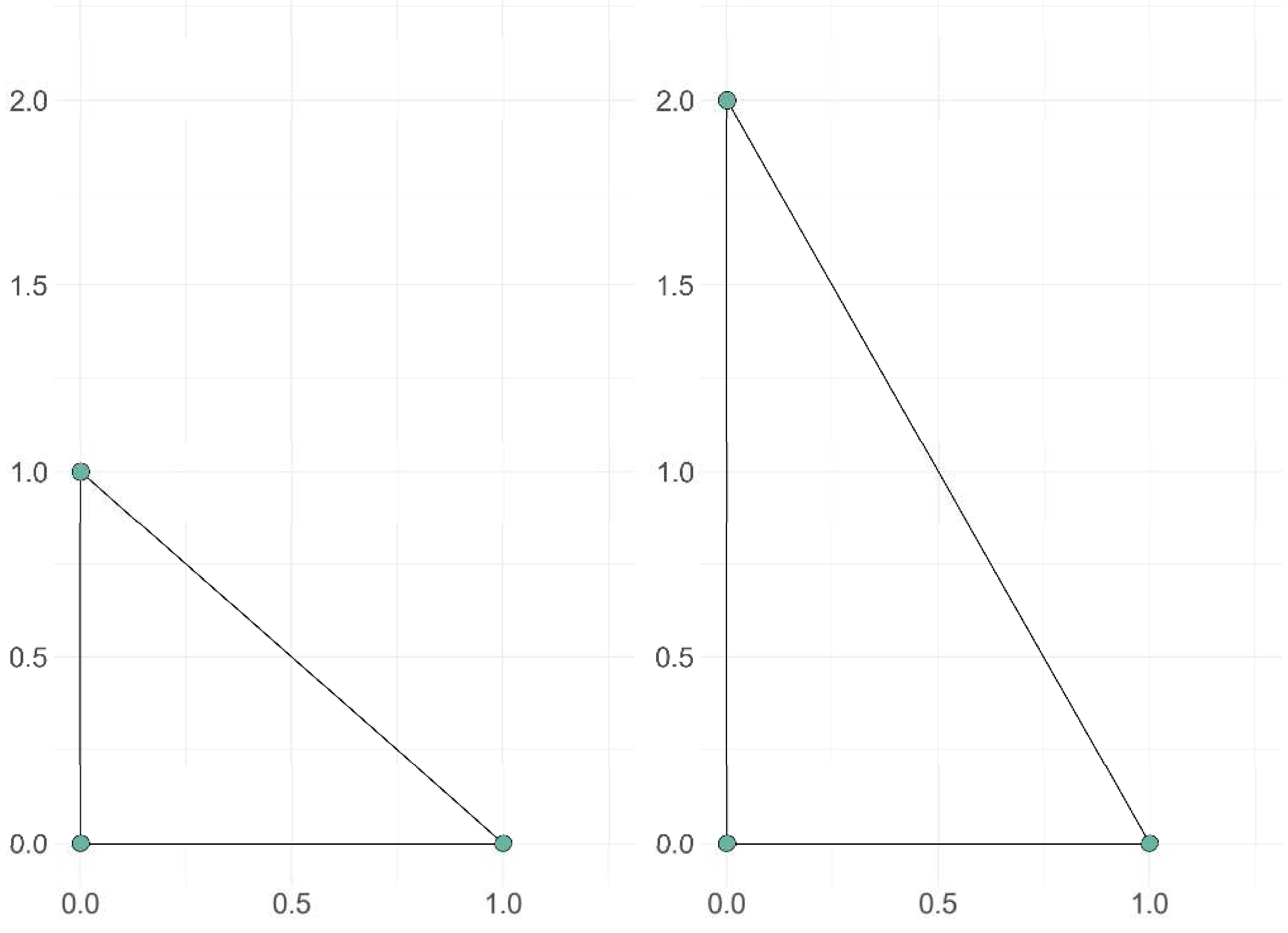
Scaling

Scaling means to transform values of certain features

Scaling effects distances between points, i.e. it allows to influence the „relevance“ of certain features (weighting)

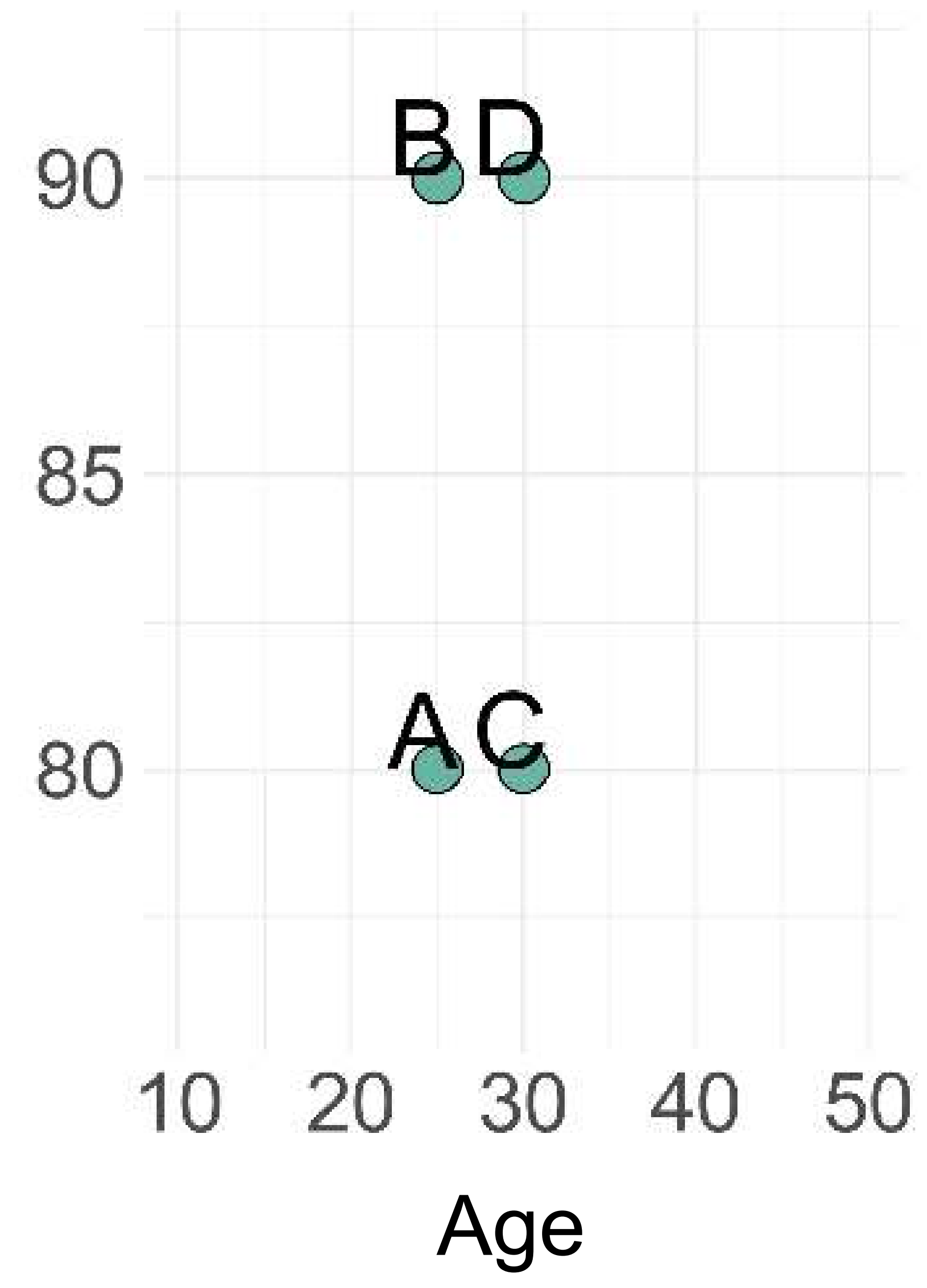
Suppose some data with 2 features

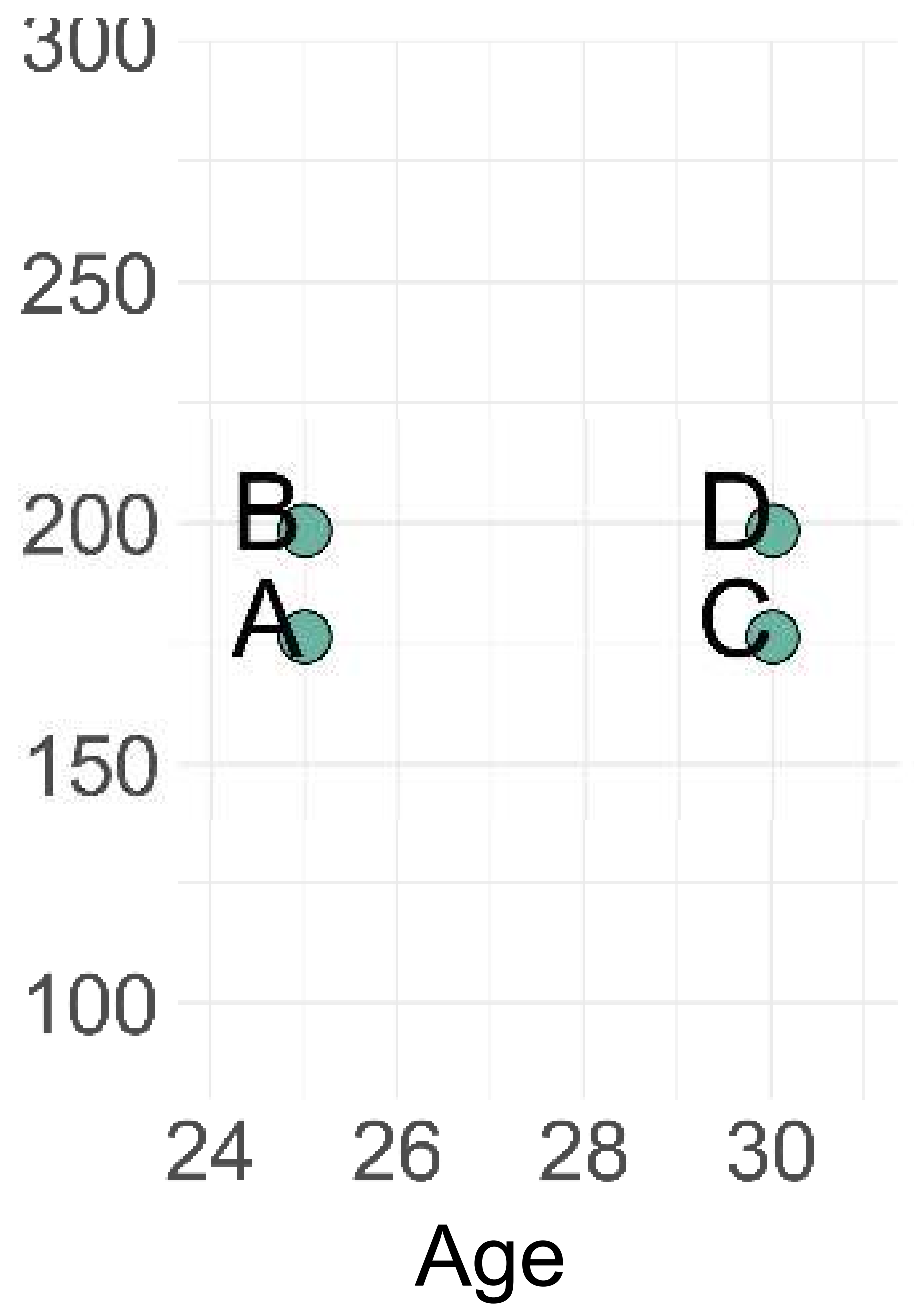
Multiplying the second feature by 2 influences the distance to other points



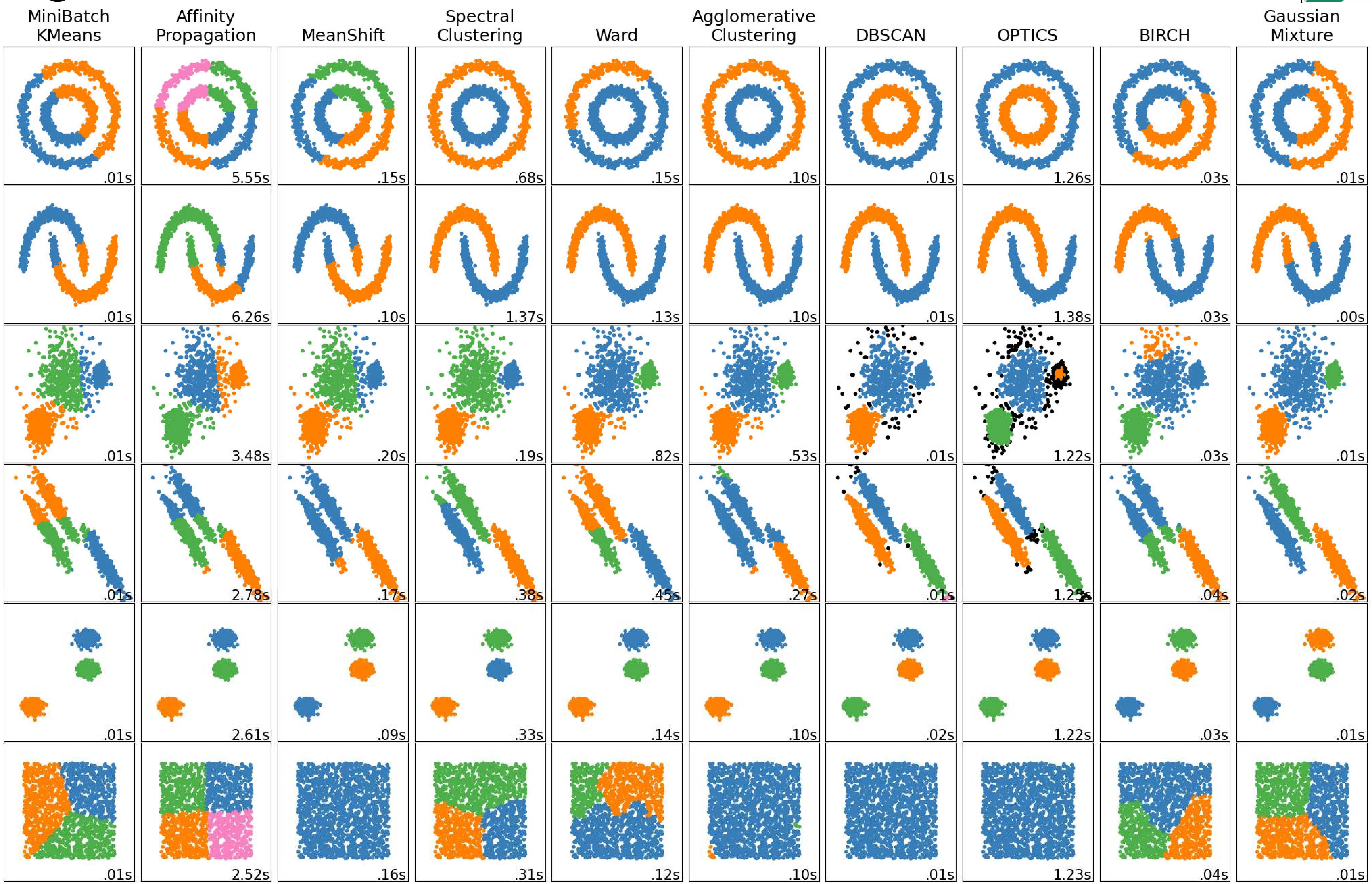
When to scale?

Name	Age	Weight (kg)	Weight (lbs)
A	25	80	176.37
B	25	90	198.42
C	30	80	176.37
D	30	90	198.42



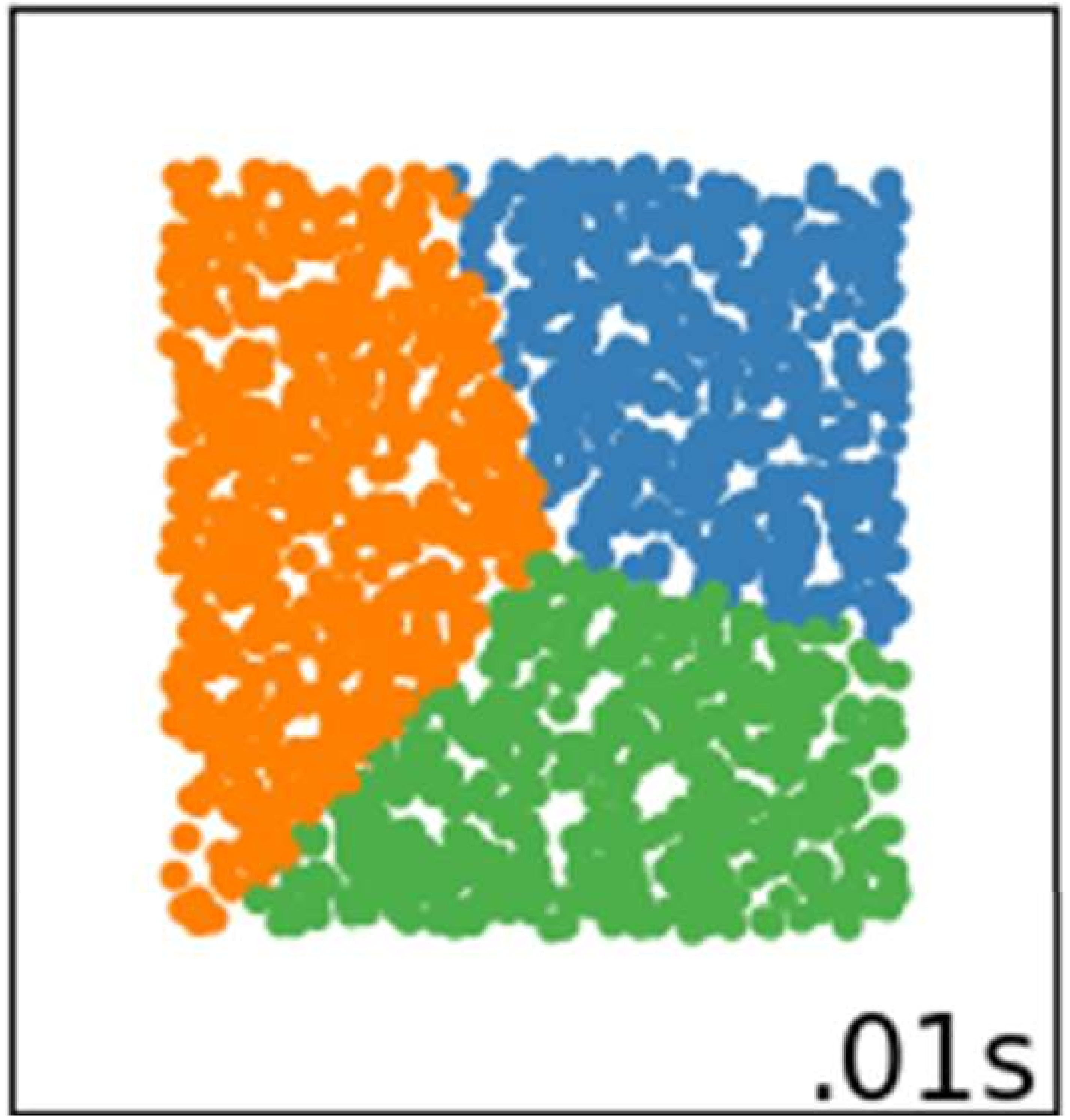


Data and Algorithms

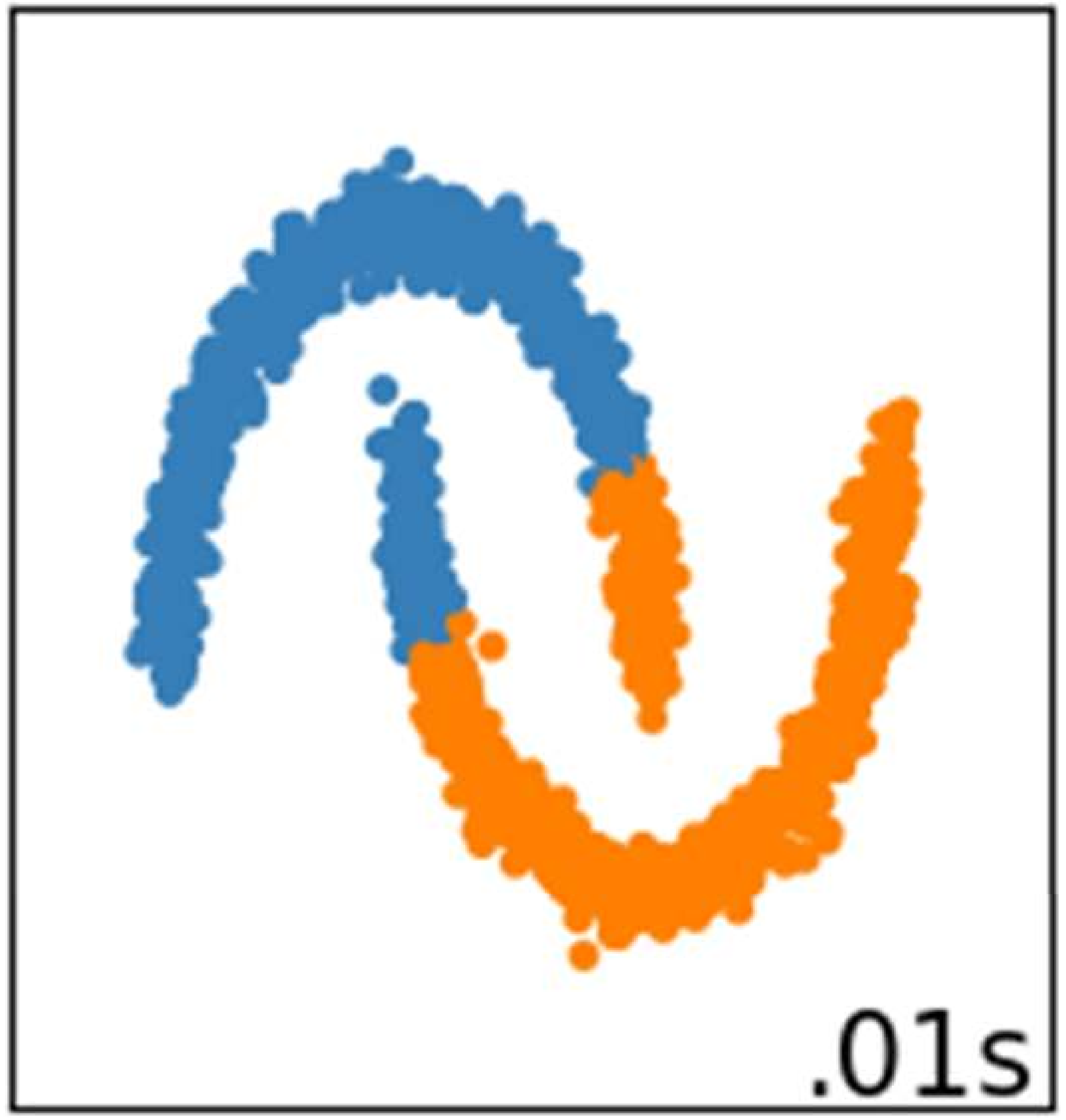


https://scikit-learn.org/stable/modules/clustering.html

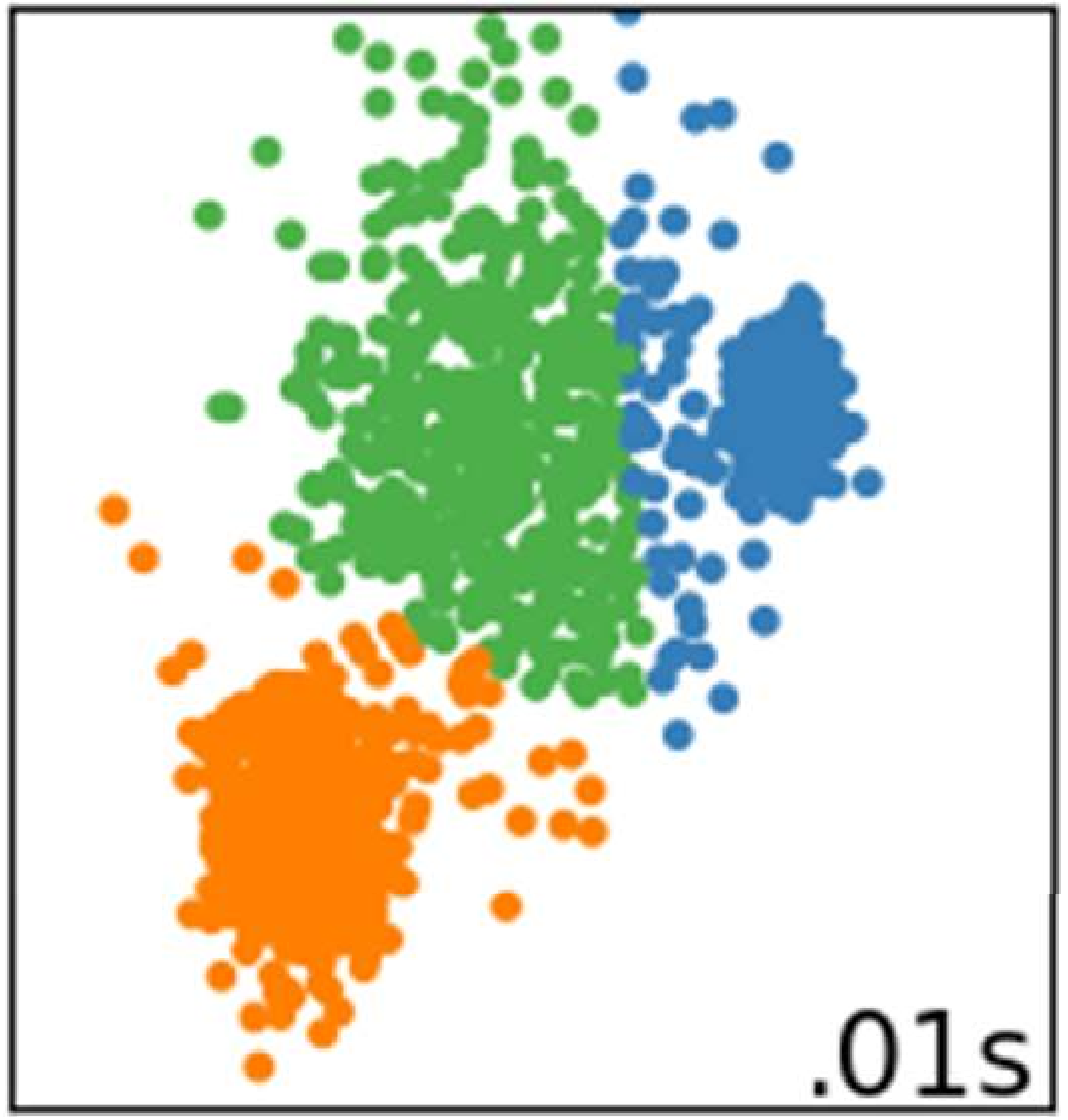
Noteworthy 1: k-means



Badly chosen *k*

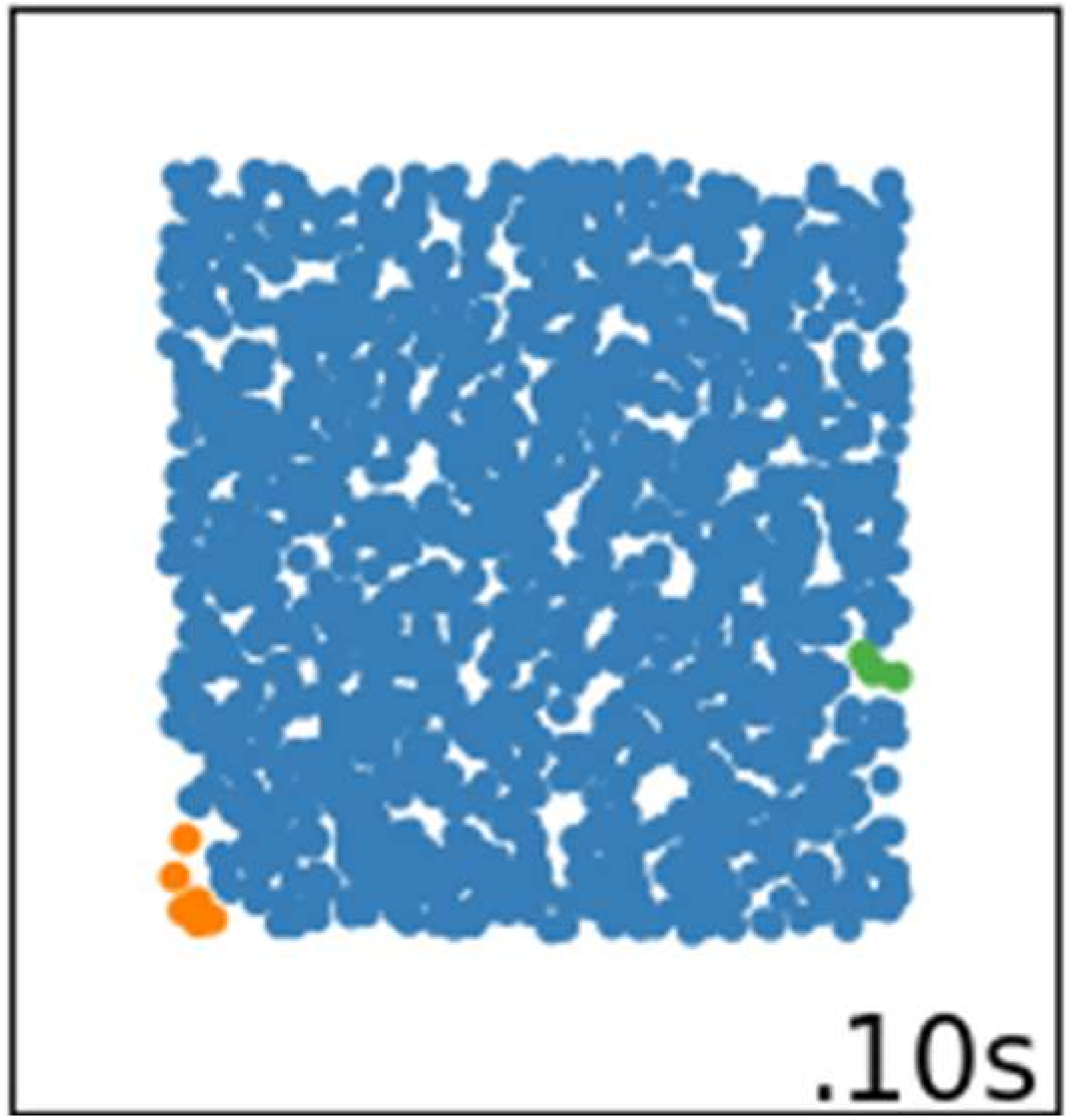


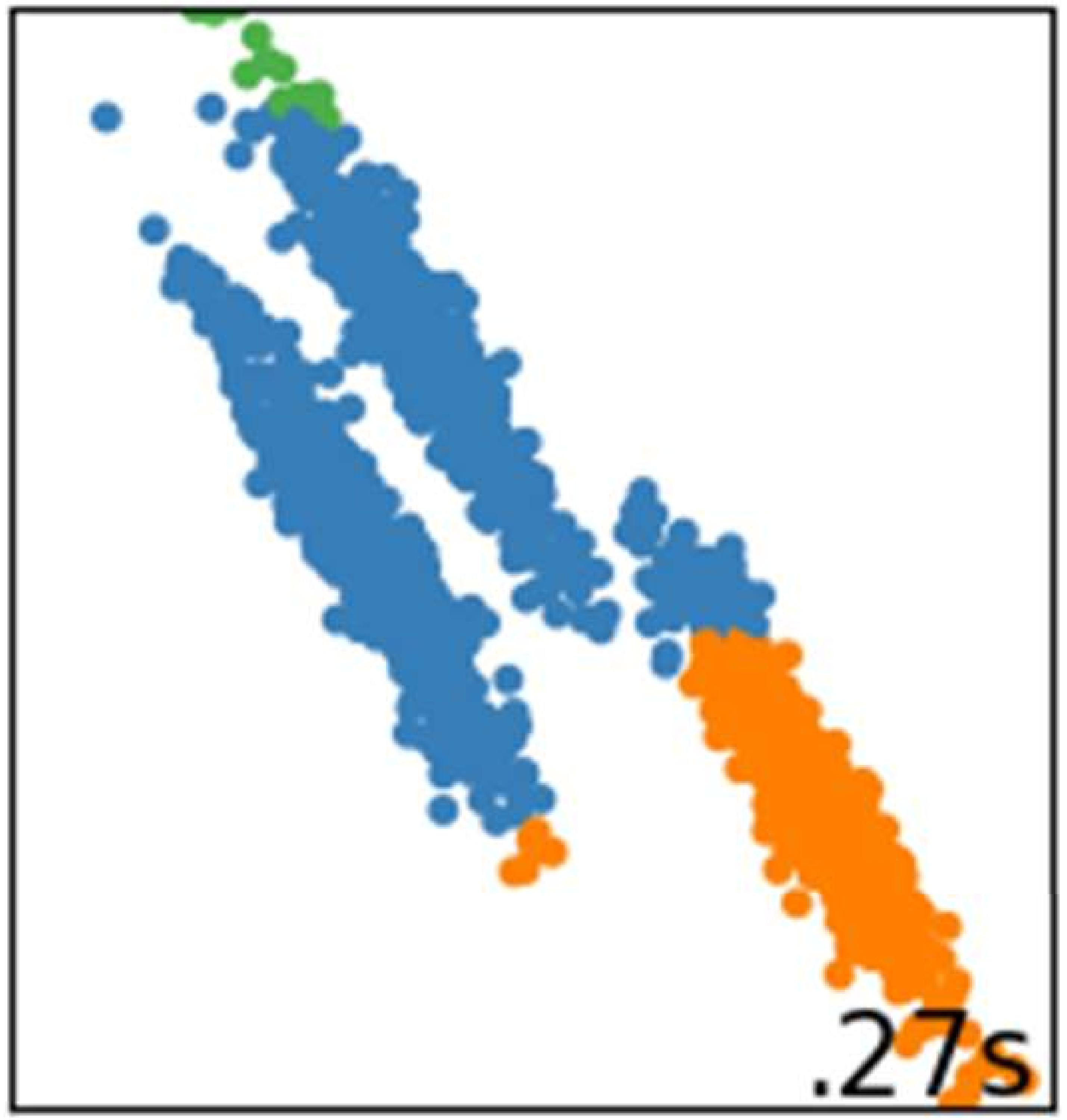
Non-spherical cluster shapes



Different cluster diameter
& different cluster densities

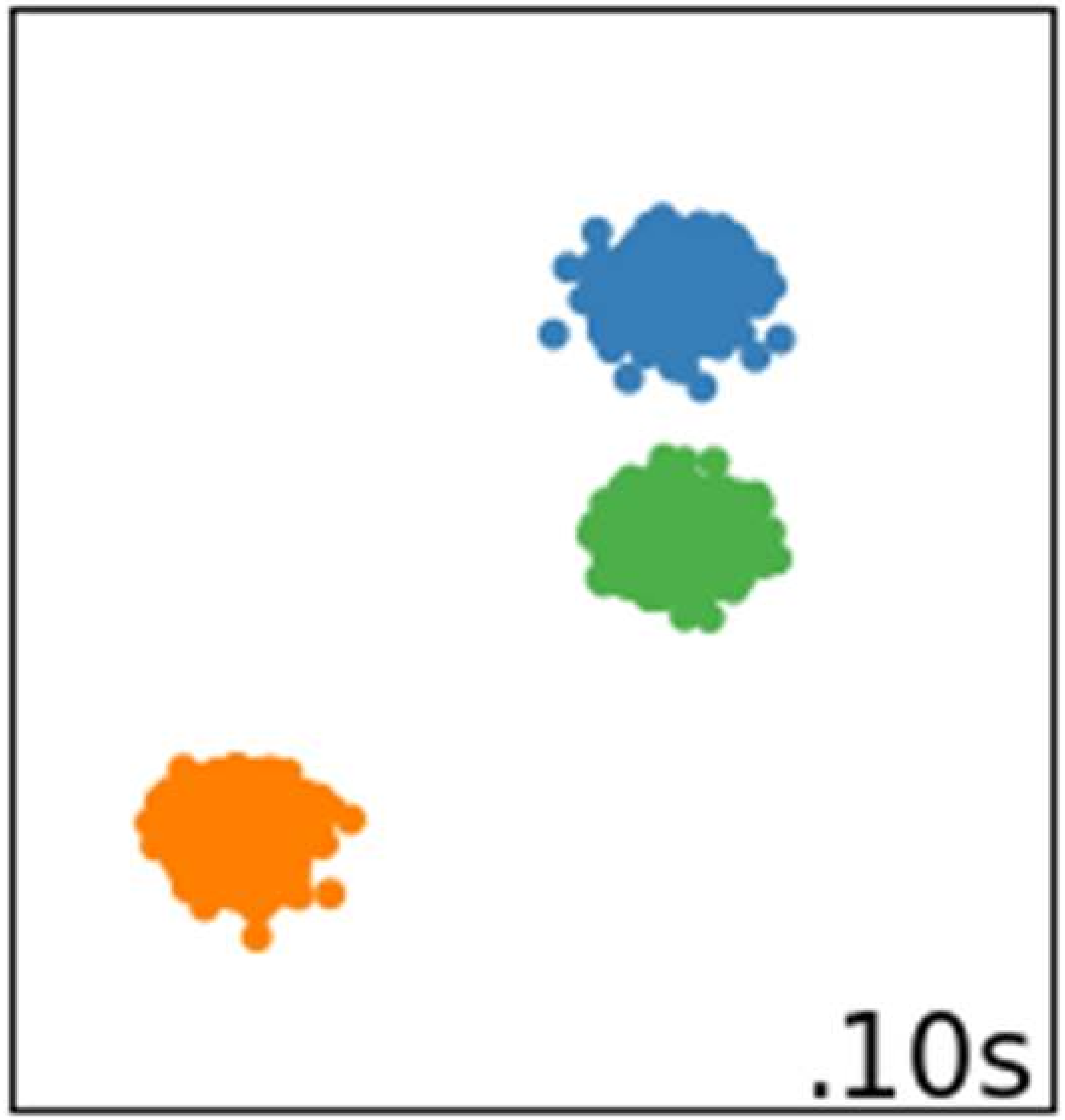
Noteworthy 2: Hierarchical clustering

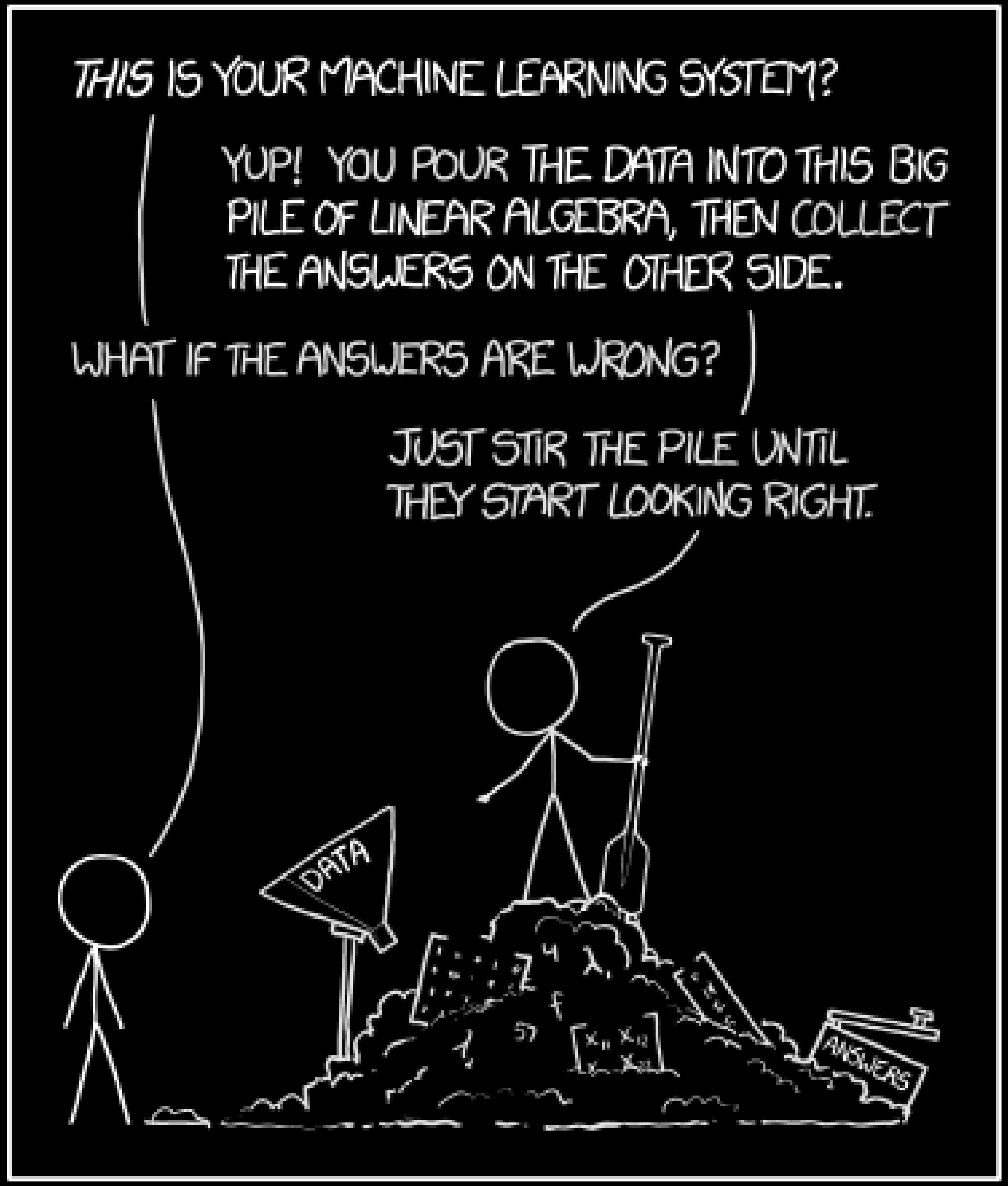




Not easy to specify both the distance metric *and* the linkage criteria

When does your data look like this?





https://xkcd.com/1838/

Thanks.

mirco.schoenfeld@uni-bayreuth.de