



# Unsupervised Learning

Mirco Schönfeld  
[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)





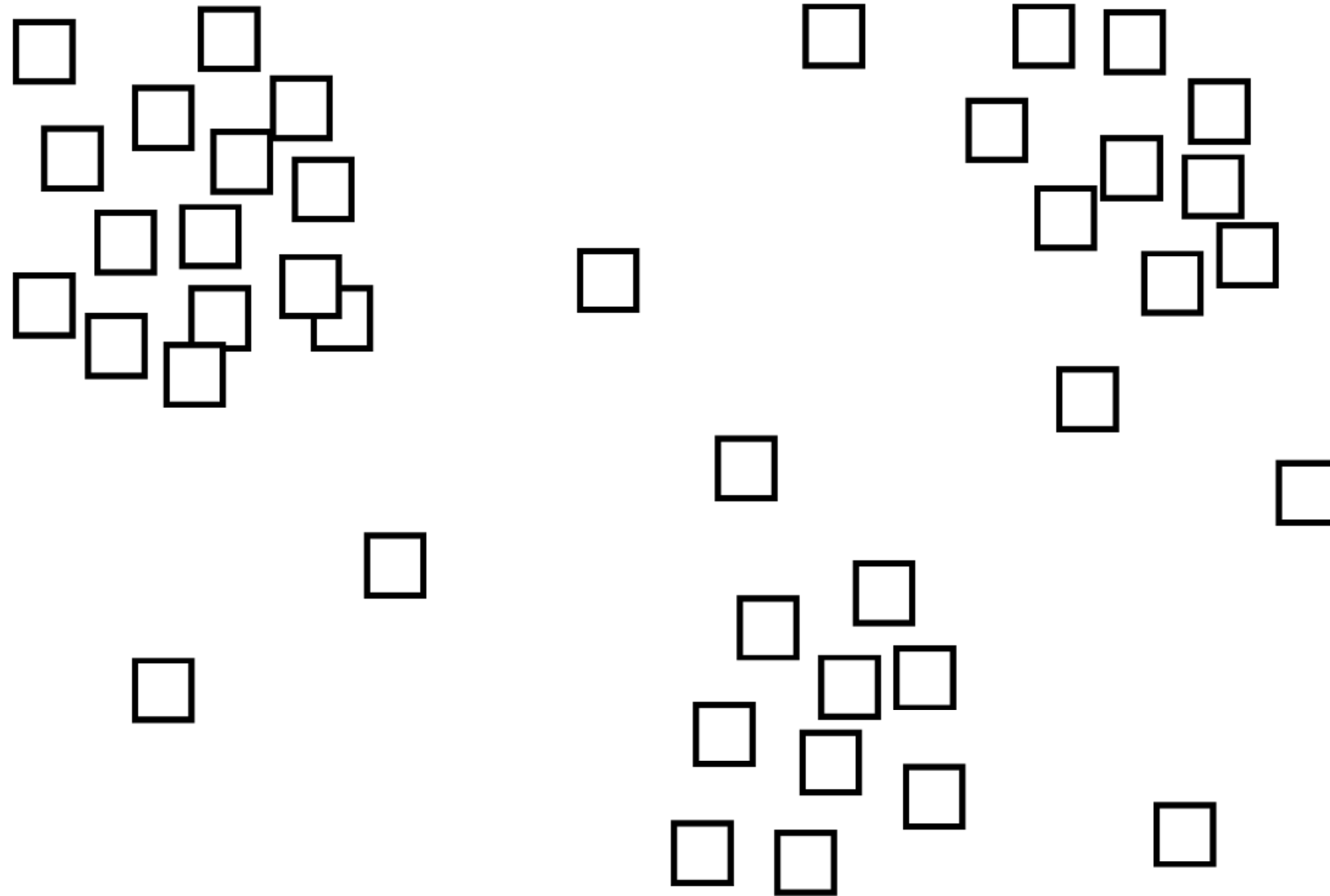






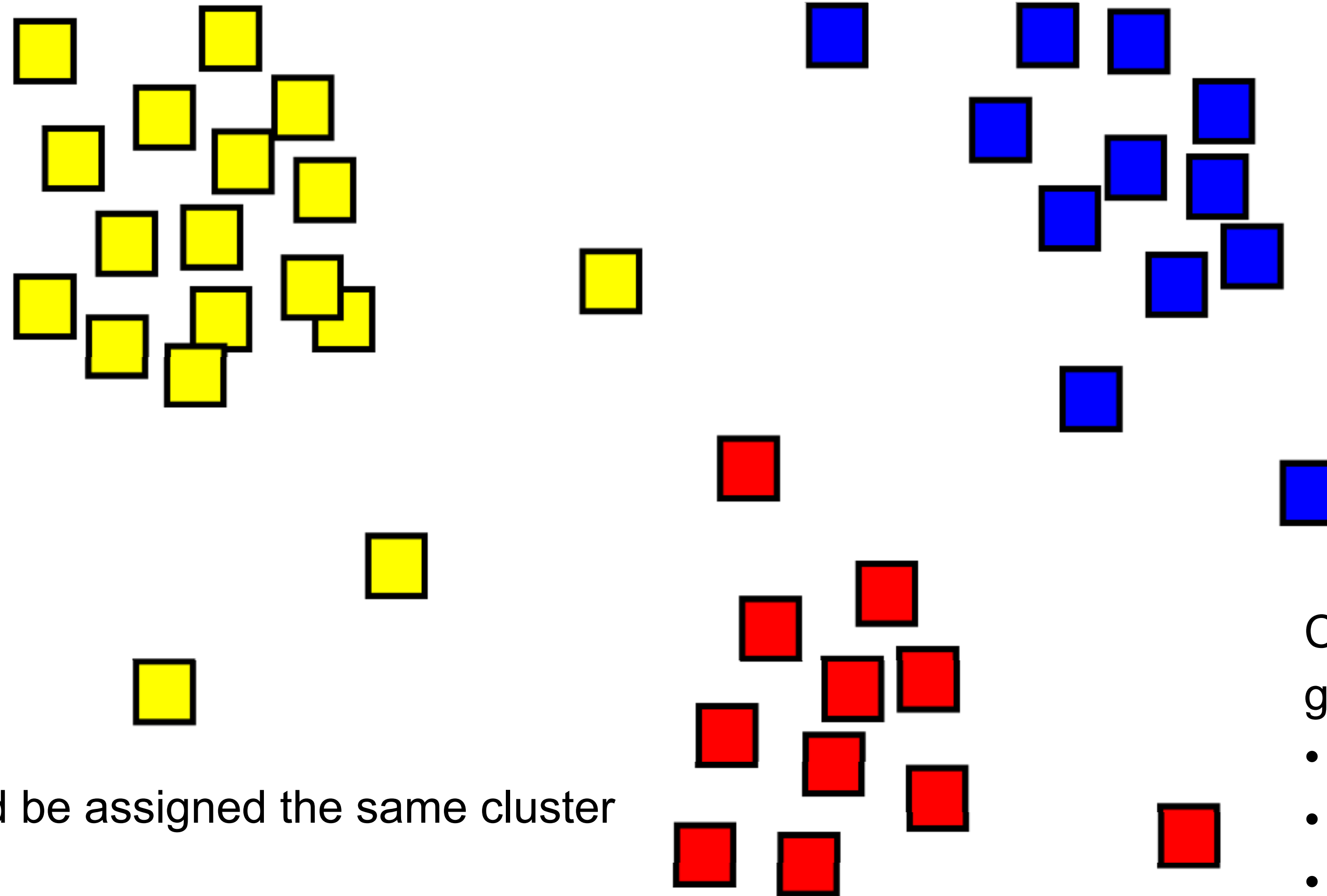
Clustering is always  
*subjective*

# 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

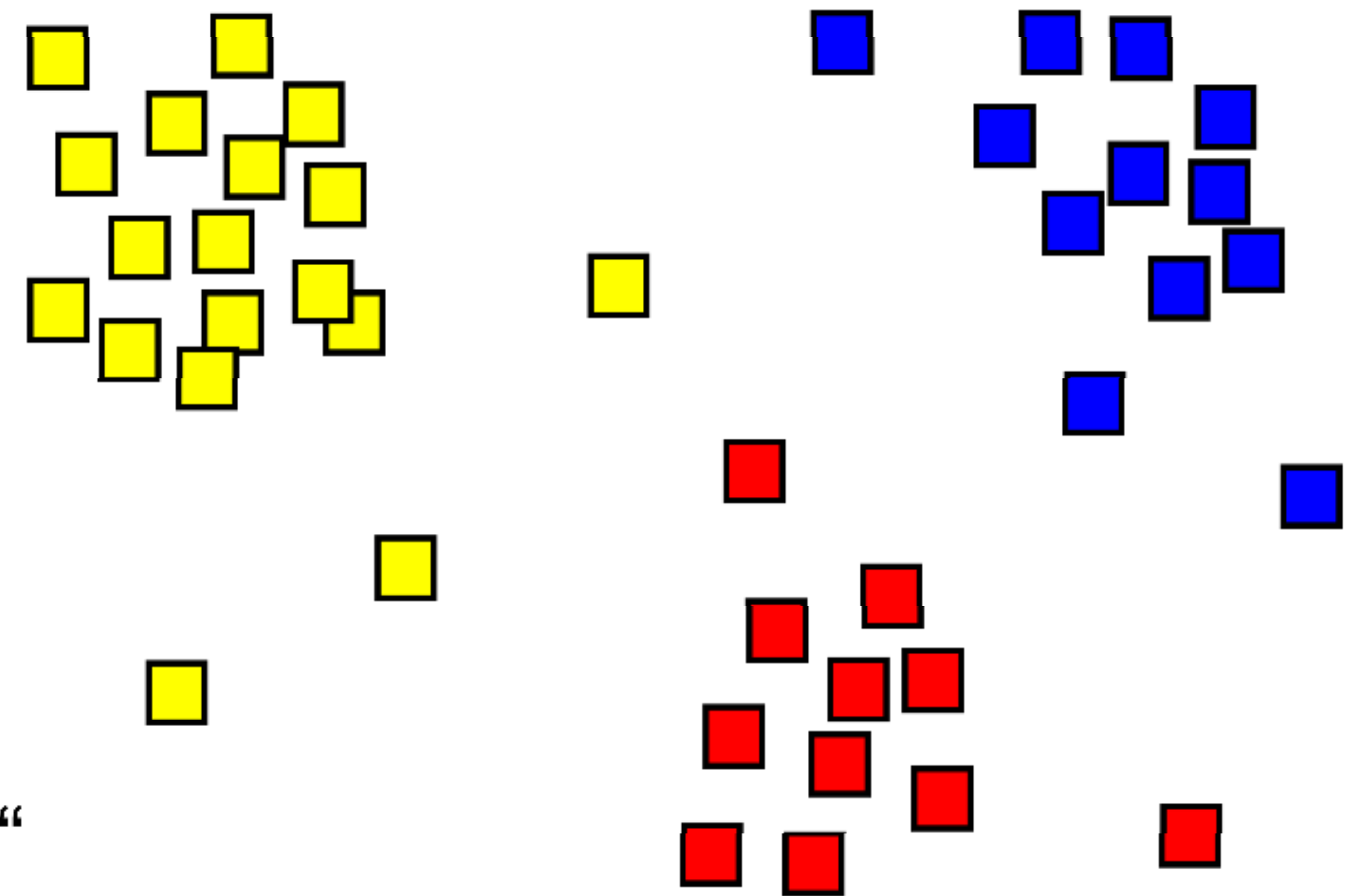
High-dimensional spaces look different

(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

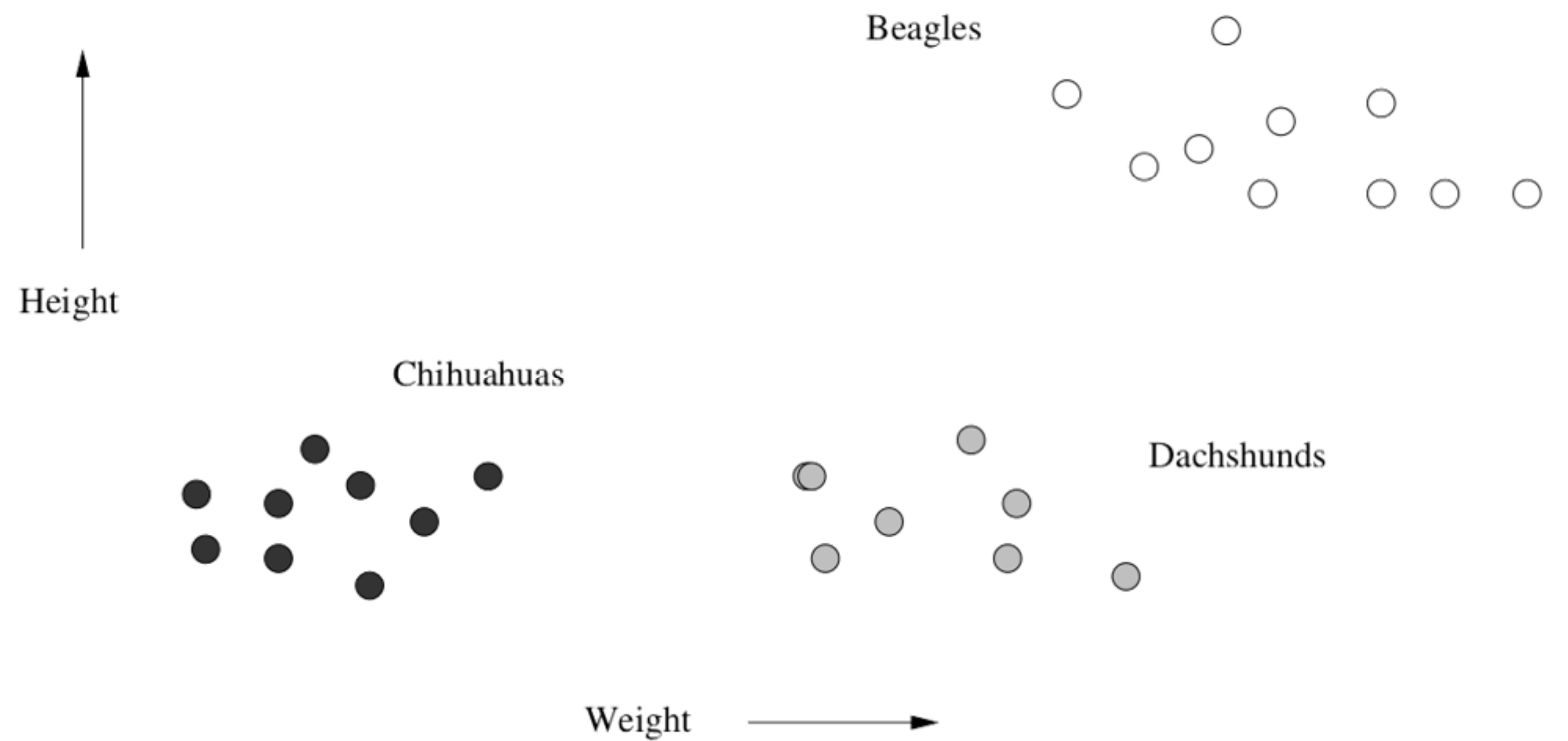
Economics: market research

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*.





# 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:  $\langle 2.53, 21.2 \rangle$

Weight: 2.53 kg

Height: 21.20 cm

Height ↑

Chihuahuas

Beagles

Dachshunds

Weight →



# 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:  $\langle 0, 1, 0, 0, 1, 0, 0, 0, 1 \rangle$

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

$$d(x, y) = 0 \text{ if and only if } x = y$$

- Distances are symmetric

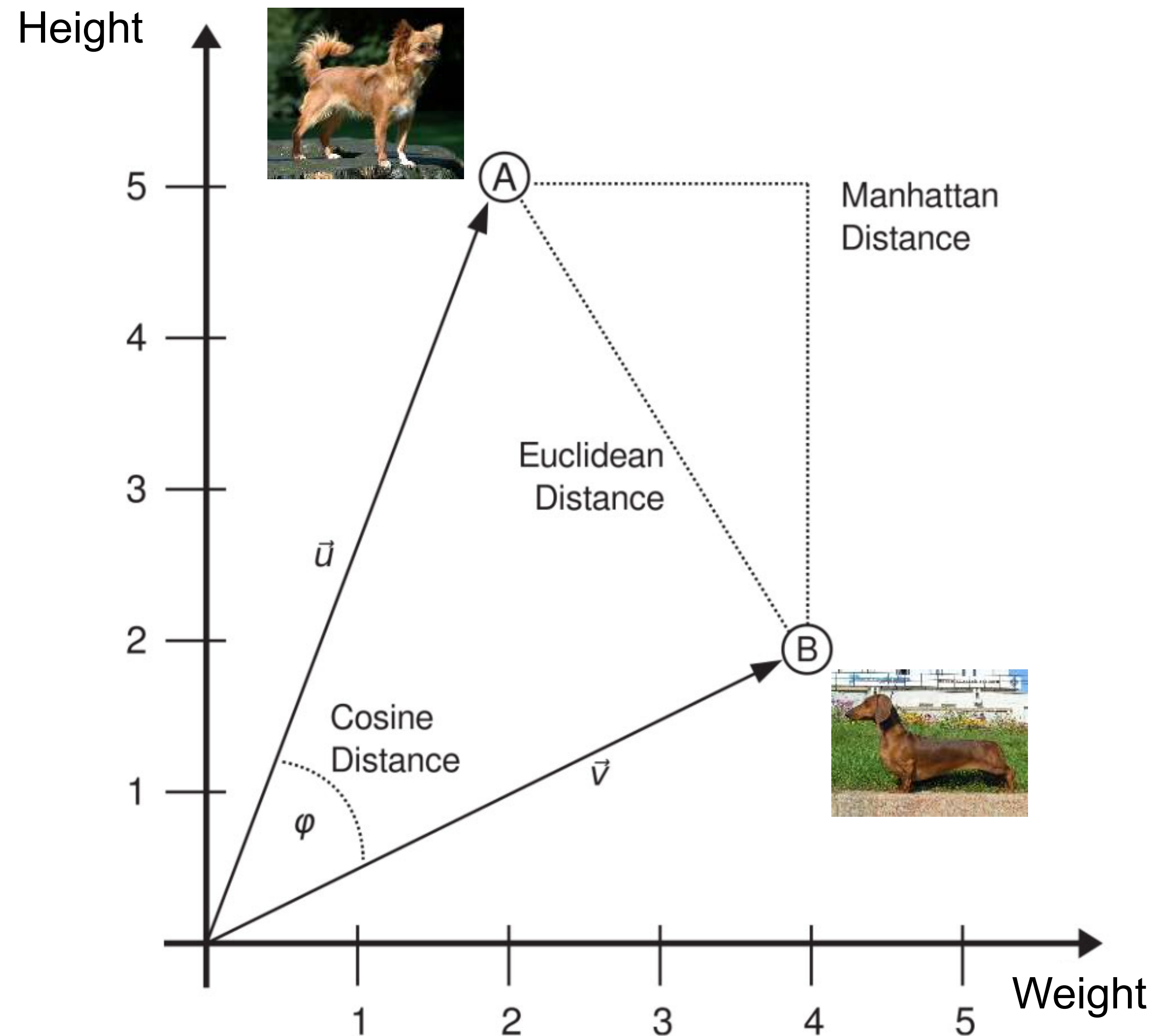
$$d(x, y) = d(y, x)$$

- 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

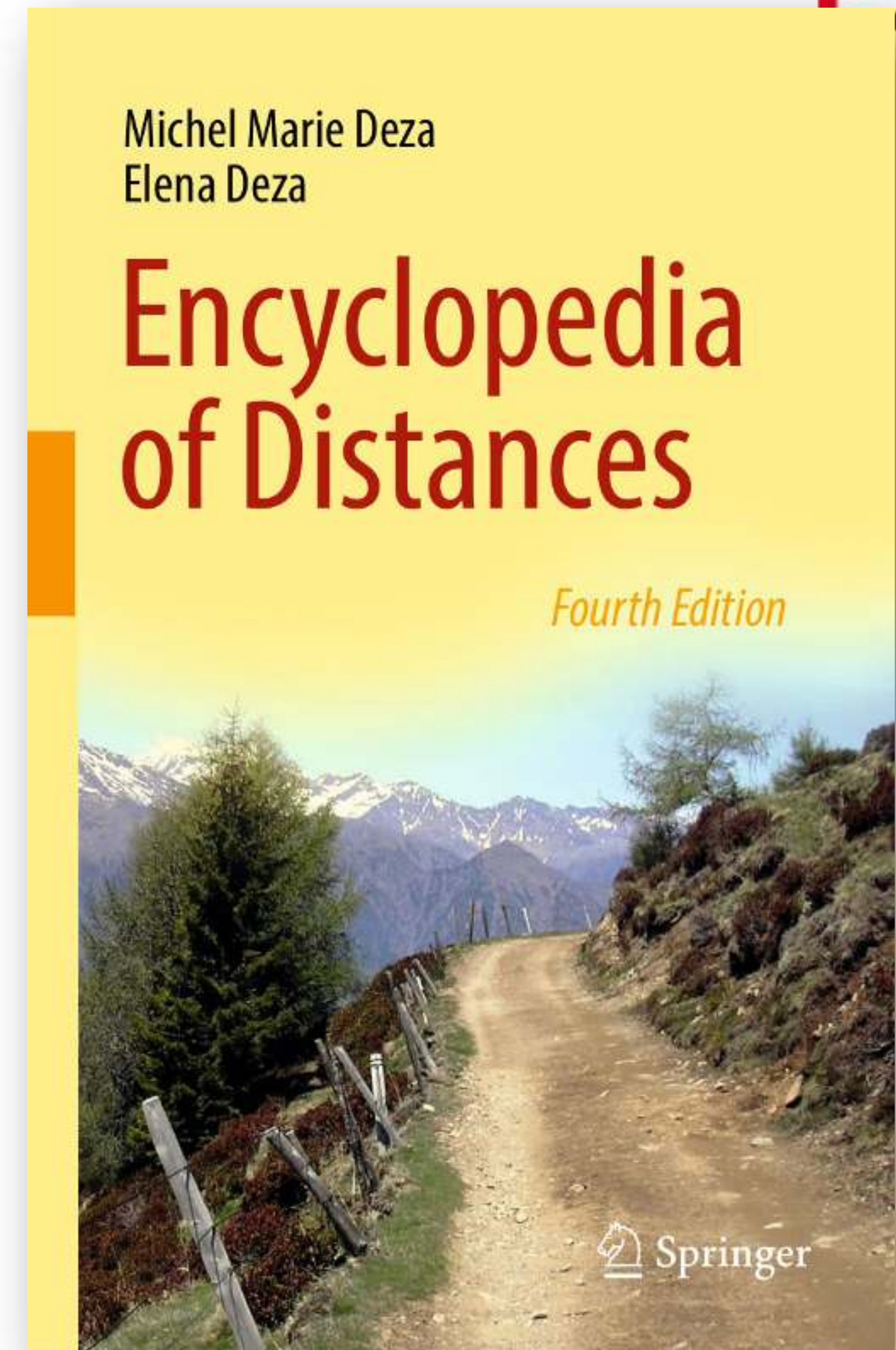


# More Distance Metrics

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

Each point is in its own cluster

Clusters are combined based on their “closeness”

Combination stops when undesirable clusters occur

## Point assignment

Initial clusters are estimated

Points are considered in some order

Points are assigned to clusters into which they best fit