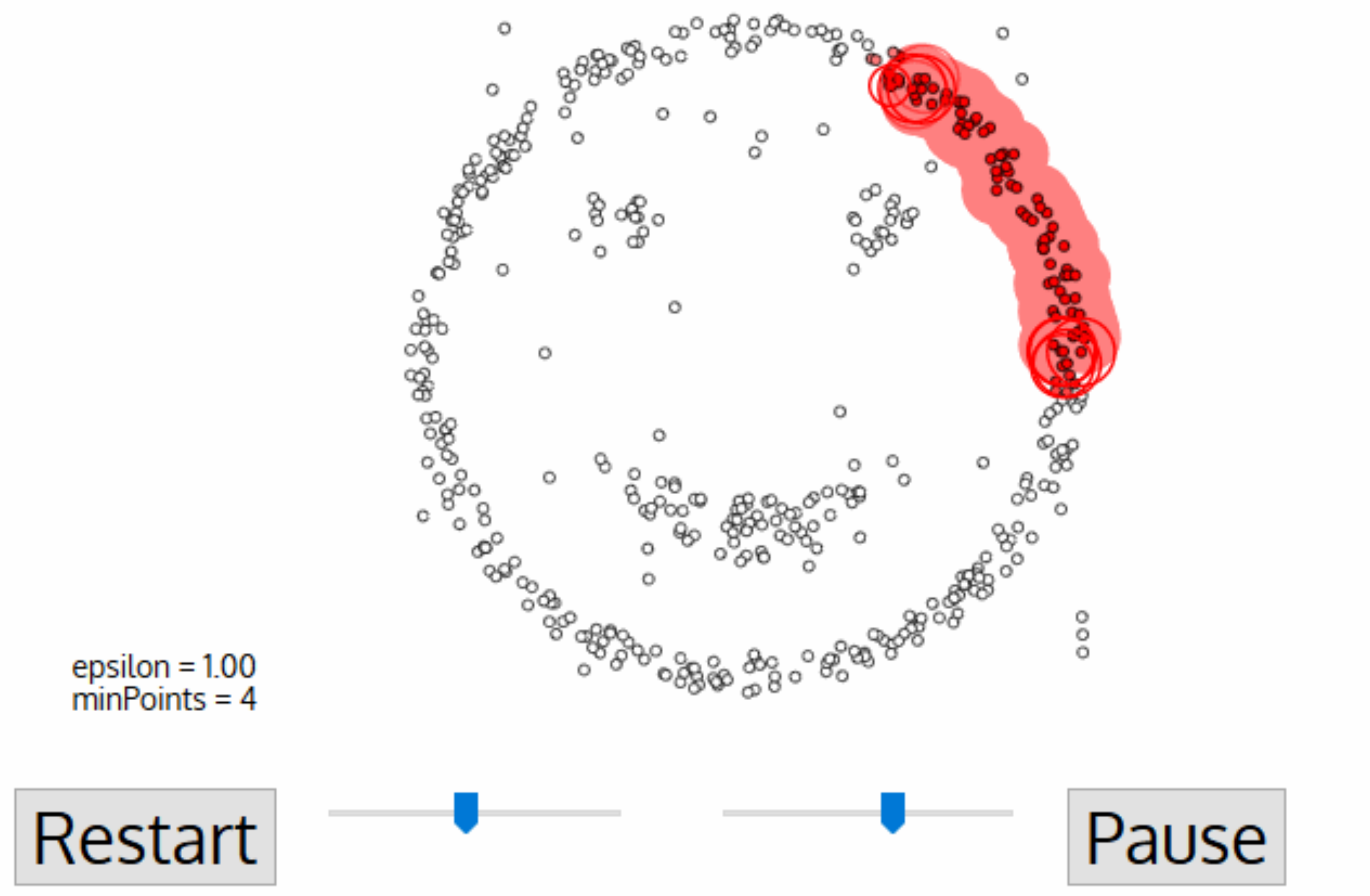
# Learning from the data

# Pattern recognition & data mining

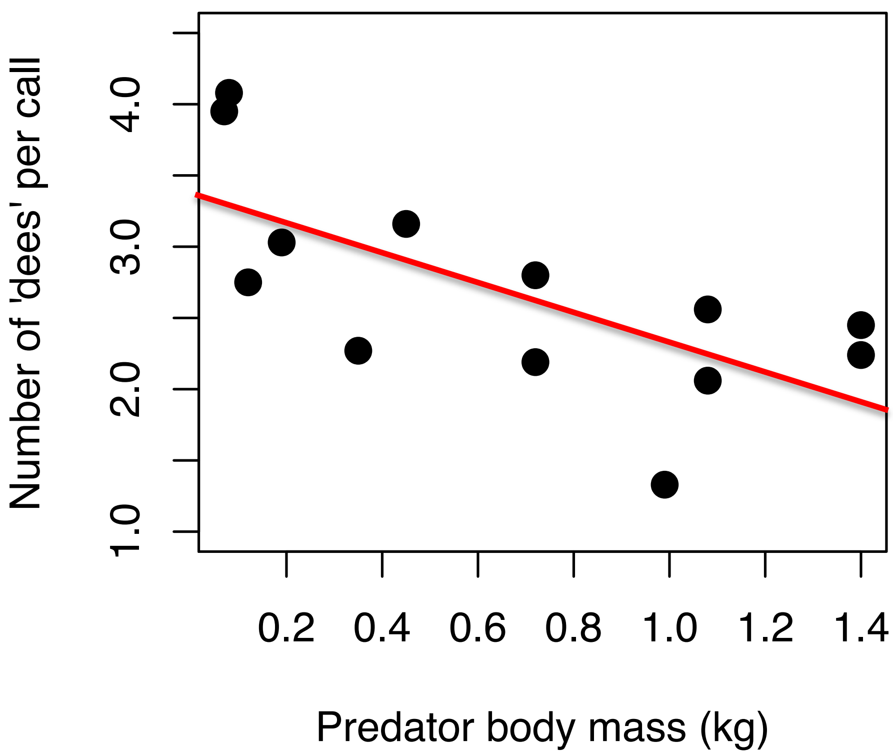
What is machine learning?
by Sabine Hauert, University of Bristol
(for the Royal Society)

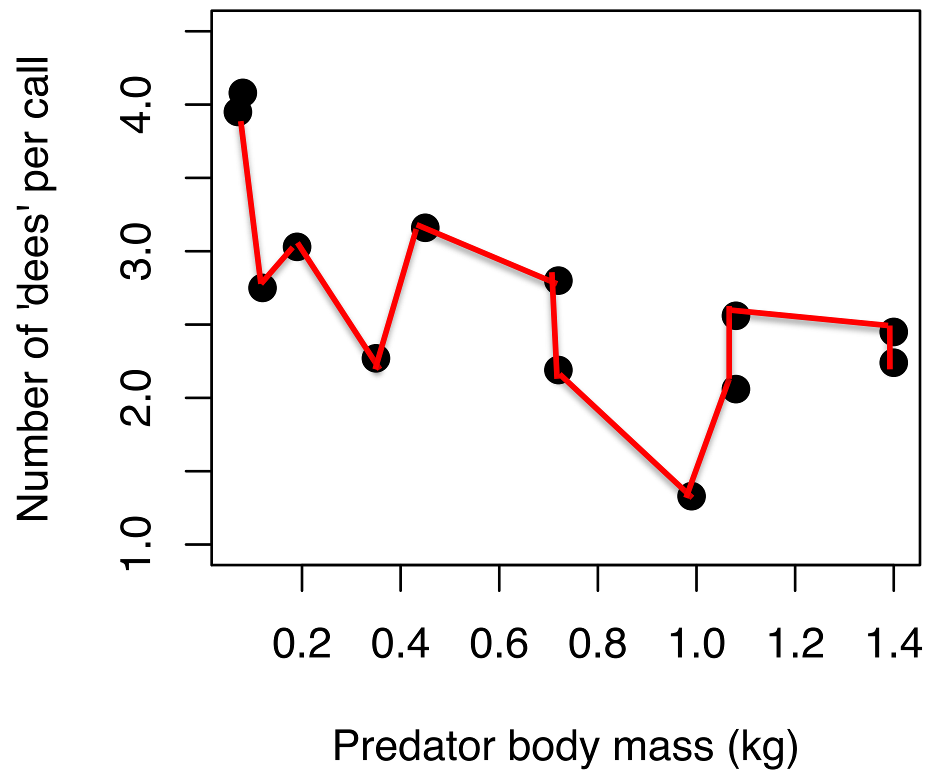
https://www.youtube.com/embed/F1wlCerC40E




https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

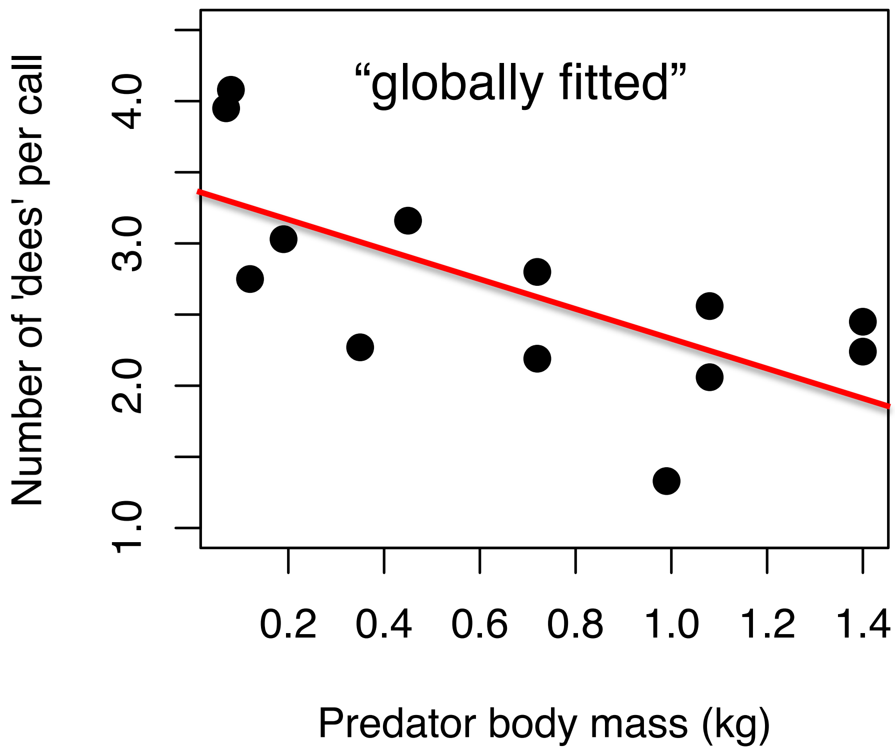
# Learning from the data

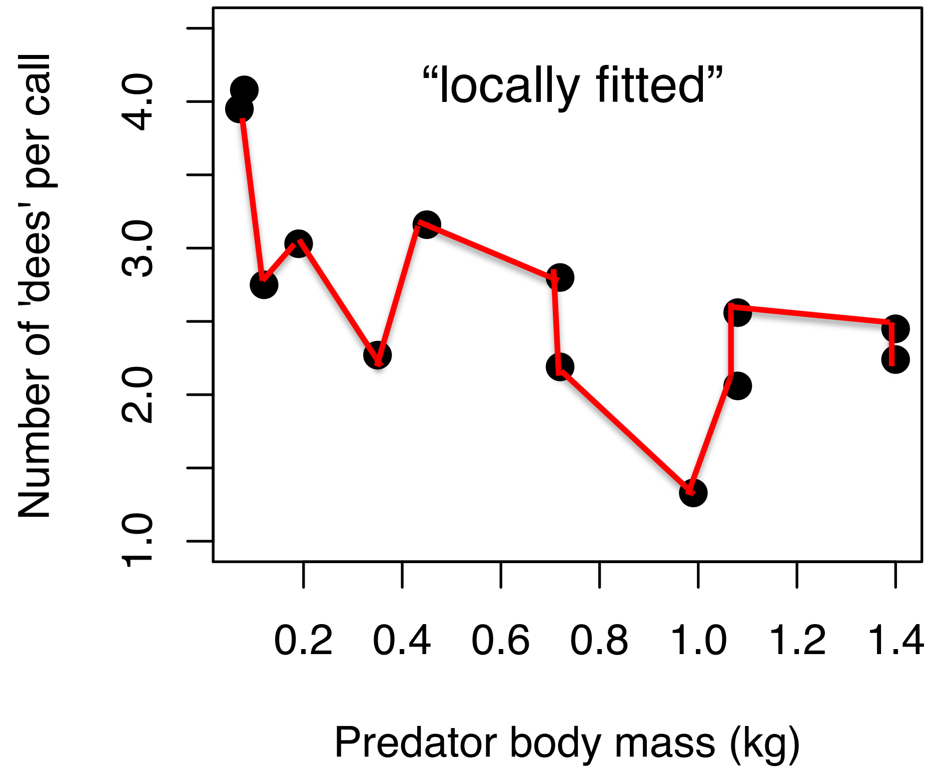



# What model do you prefer? Why? Which model does better?

# Learning from the data




## What model do you prefer? Why?

"Intelligence is 10 million rules" (Doug Lenat)….but Rules are meant to be generalizable

# Learning from the data - Machine learning algorithms

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).

# Learning from the data - Machine learning algorithms

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).

- **Machine learning** focuses on the development of computer algorithms that can change when exposed to new data. The process of **machine learning** is similar to that of data mining. The process is not strictly static following programming instructions; instead, they make data driven decisions (adapted from Wikipedia).

# Learning from the data - Machine learning algorithms

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).

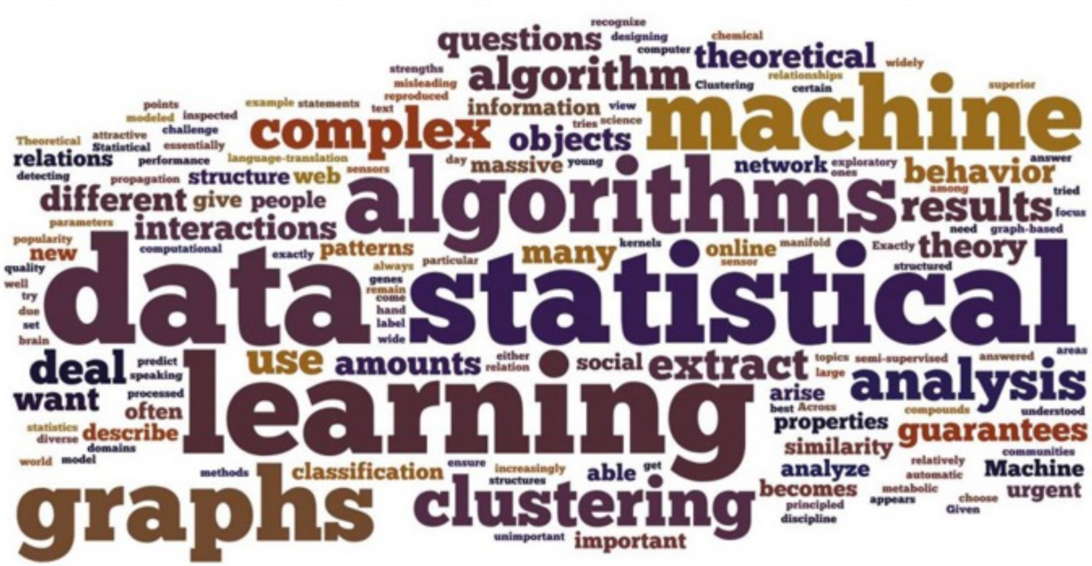
- **Machine learning** focuses on the development of computer algorithms that can change when exposed to new data. The process of **machine learning** is similar to that of data mining. The process is not strictly static following programming instructions; instead, they make data driven decisions (adapted from Wikipedia).

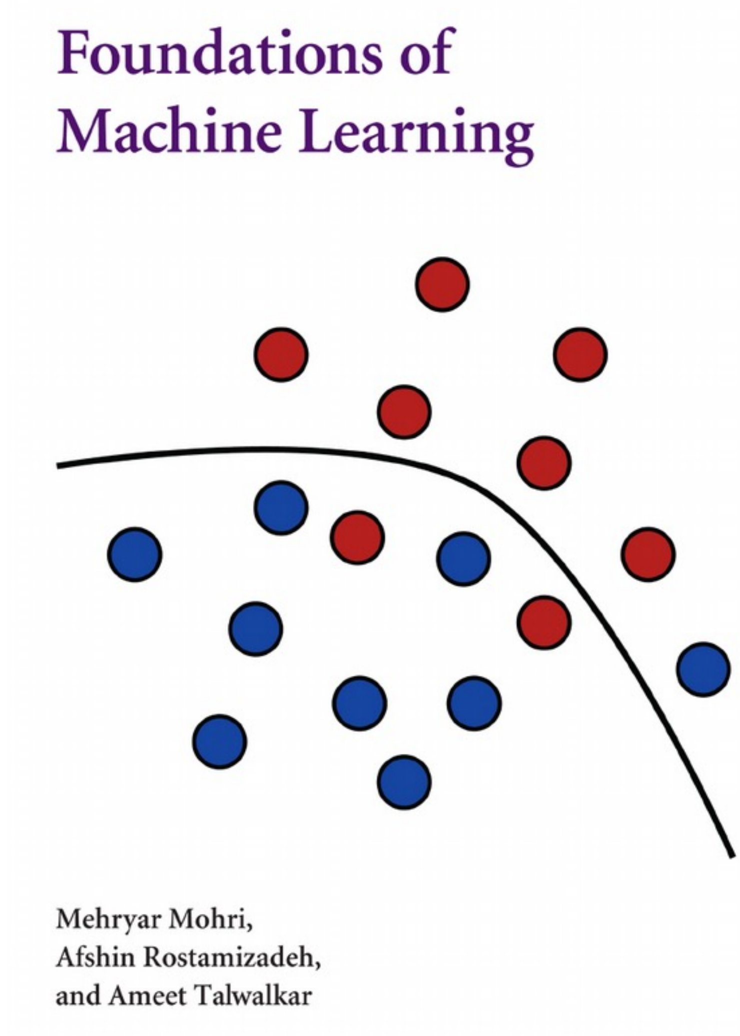
- Analysis based on **machine learning** may change when the learning process algorithm is run on the same data multiple times.

# Learning from the data - Machine learning algorithms

**Machine learning** mixes computer sciences and statistics and relaxes assumptions ("sometimes").

# Learning from the data - Machine learning algorithms

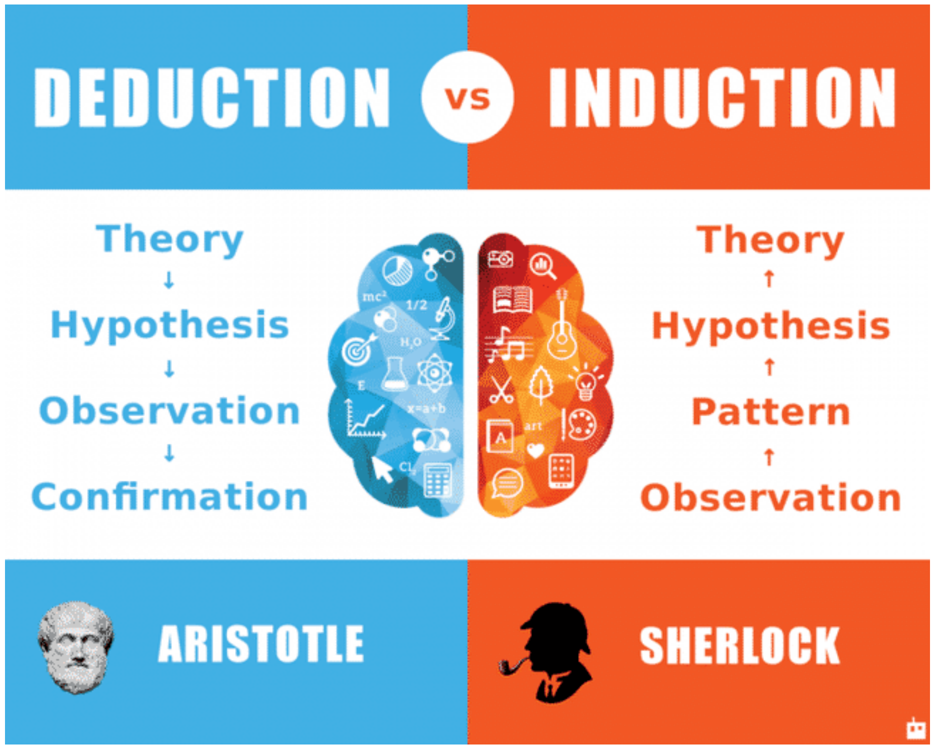


# Machine learning as an inductive process

General

Specific




General

Specific

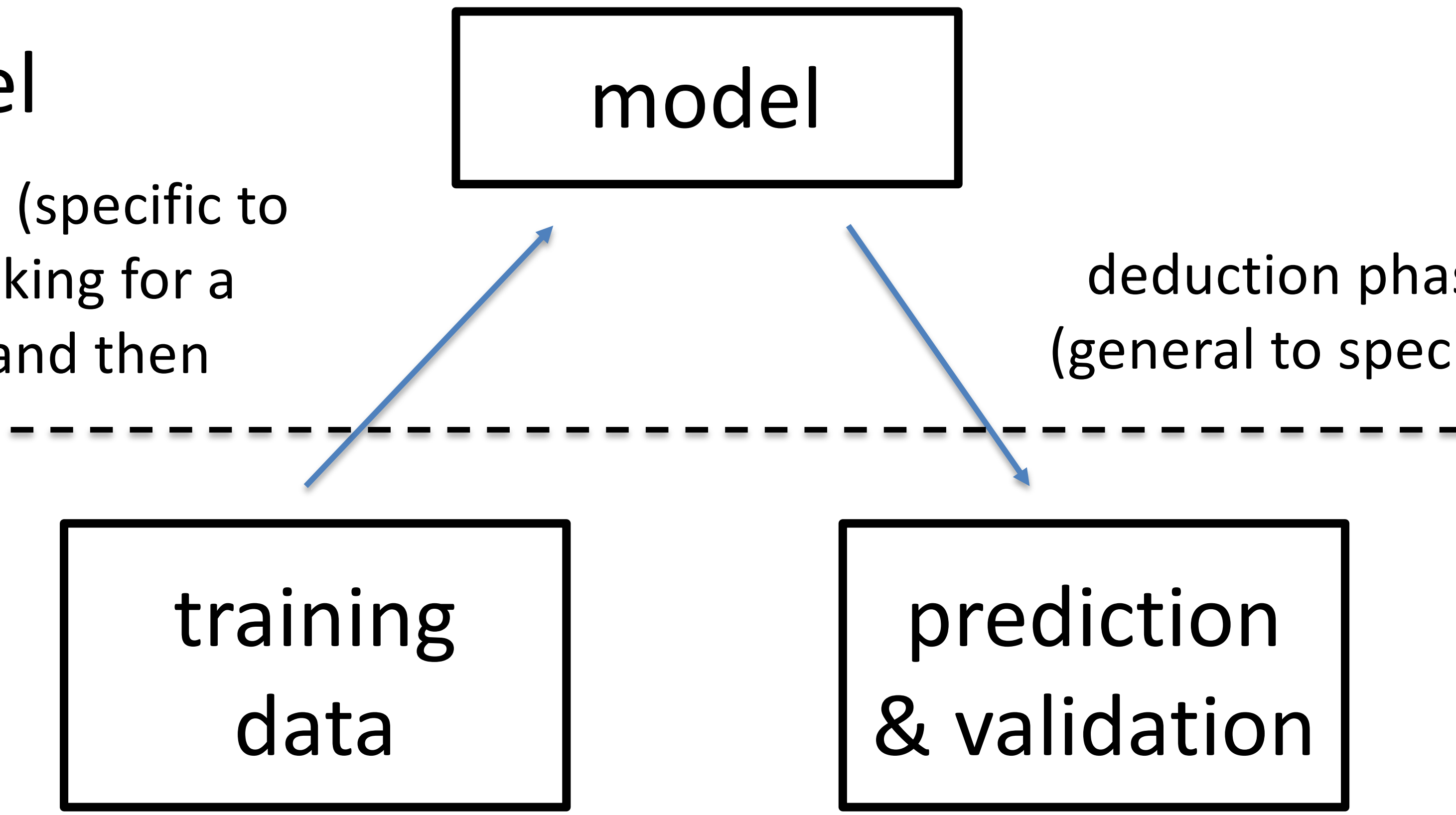
https://danielmiessler.com/blog/the-difference-between-deductive-and-inductive-reasoning/

# Learning from the data - Machine learning algorithms

Model level

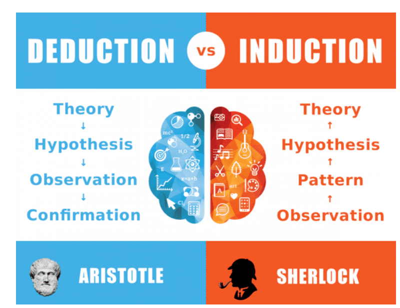
Induction phase (specific to general; i.e., looking for a pattern in data and then generalize it)

deduction phase (general to specific)




Data level

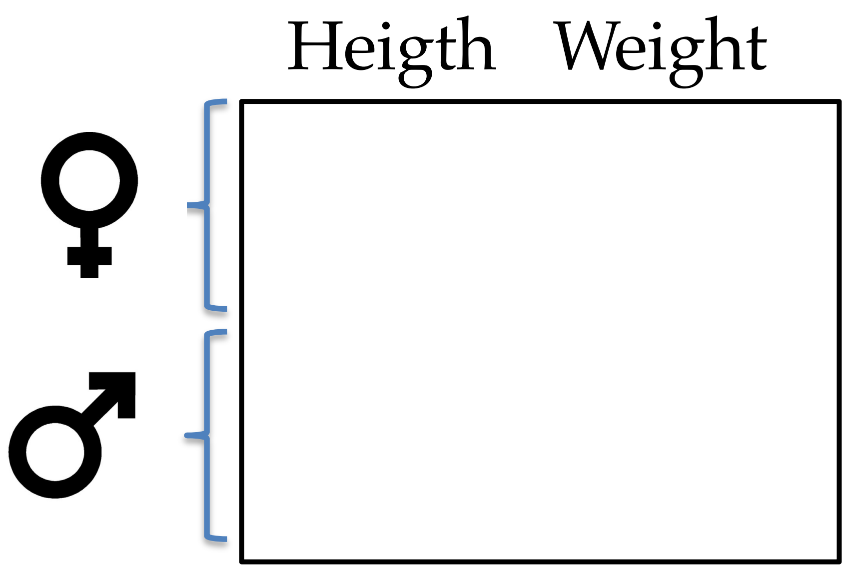
Modified from
http://www.cs.joensuu.fi/~whamalai/skc/ml.html



# Learning from the data
# Machine learning algorithms - Two main types

Labeled data → **Supervised Learning Algorithm** → Prediction based on knowing the label of observations

Label = gender



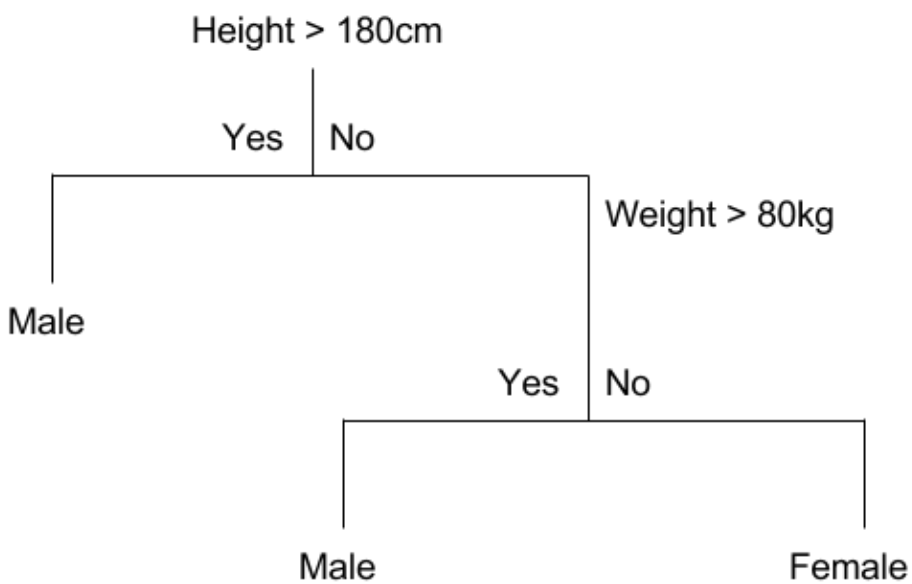
# Learning from the data
# Machine learning algorithms - Two main types

Labeled data

**Supervised Learning Algorithm**

Prediction based on knowing the label

Predicting gender on the basis of Height and Weight



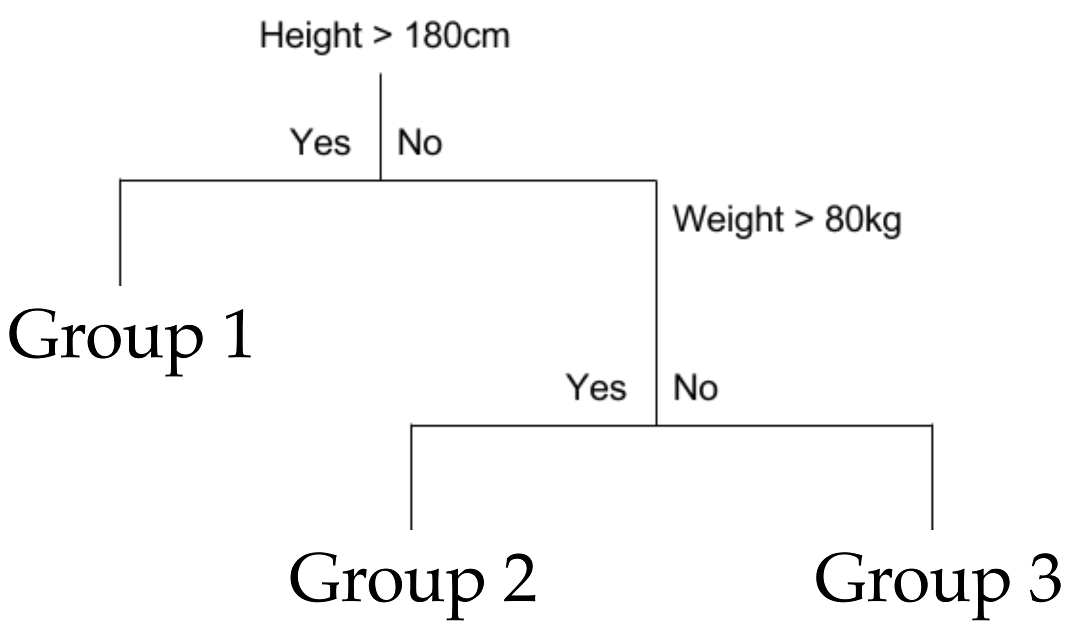

Label = gender

# Learning from the data [TODAY]
# Machine learning algorithms - Two main types

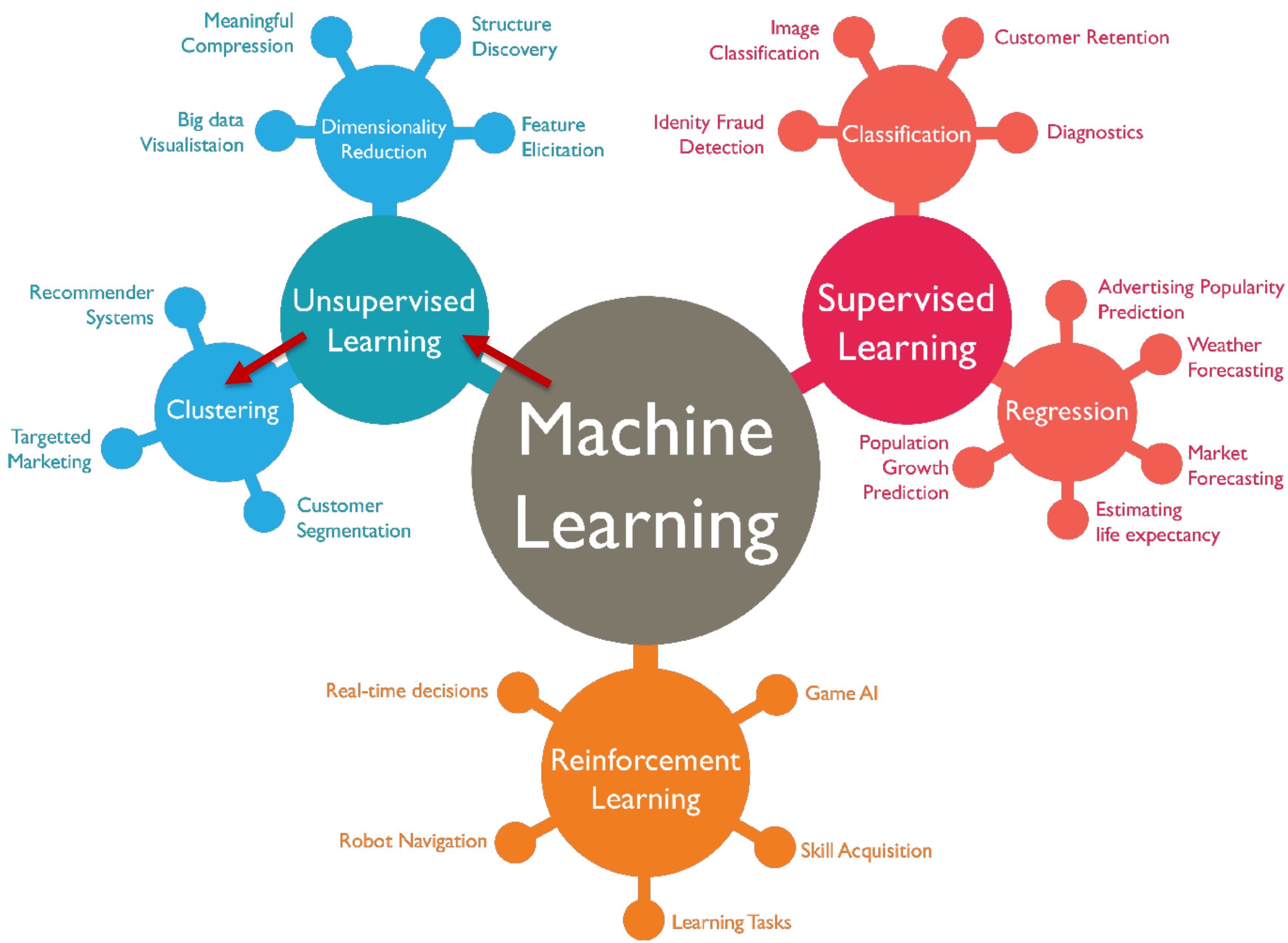
Unlabeled data

**Unsupervised Learning Algorithm**

Prediction based on finding patterns in the data




e.g., Finding number of groups in data and ways to classify (predict) observations based on their characteristics (height/weight)




https://medium.com/marketing-and-entrepreneurship/10-companies-using-machine-learning-in-cool-ways-887c25f913c3

The k-means clustering algorithm

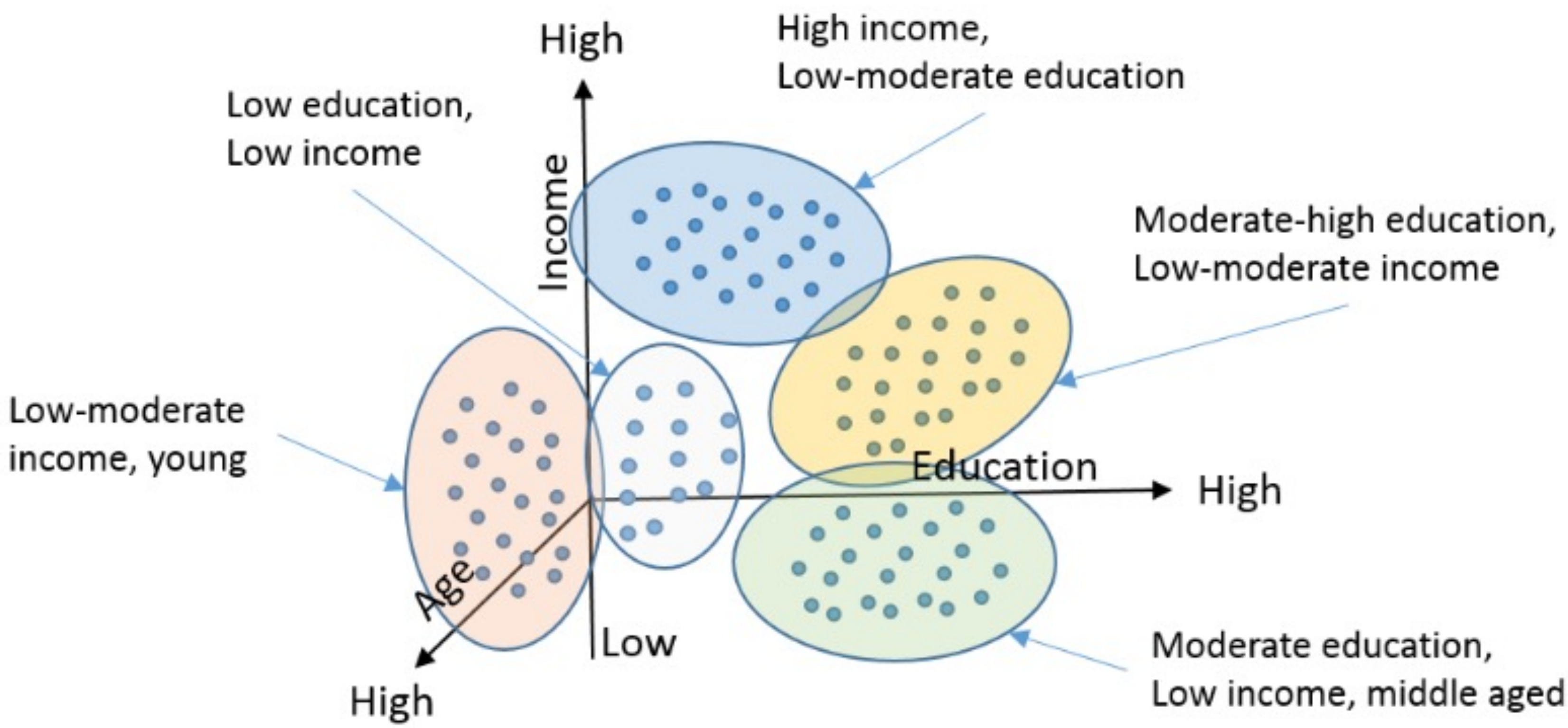
Easy to see what it does (video)

https://www.youtube.com/watch?v=4b5d3muPQmA

Learning from the data - Machine learning algorithms

# K – means clustering method (unsupervised algorithm)

(an example outside of biology)




**6 groups seem to describe the data quite well**

# Parallel *k*-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets

Jitendra Kumar[a,1], Richard T. Mills[a], Forrest M. Hoffman[a], William W. Hargrove[b]

[a]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[b]Southern Research Station, USDA Forest Service, Asheville, NC, USA

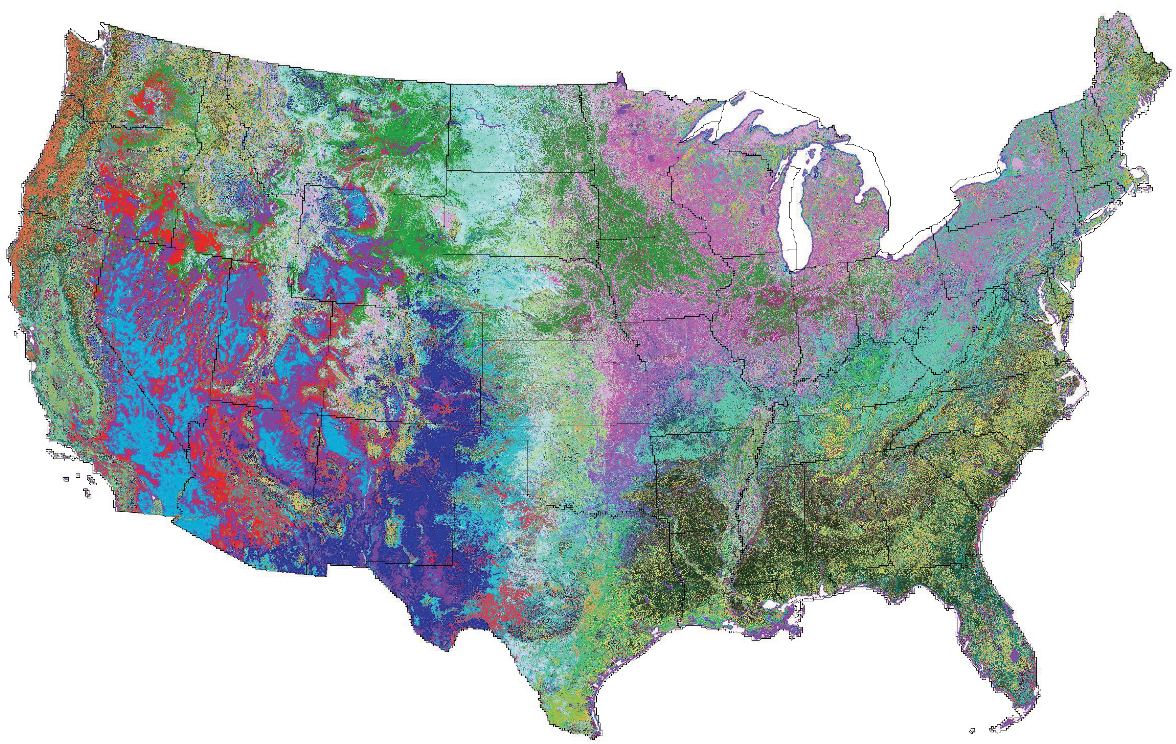


Figure 6: The 2008 map of 50 phenoregions defined for the CONUS derived from cluster analysis of Phenology data

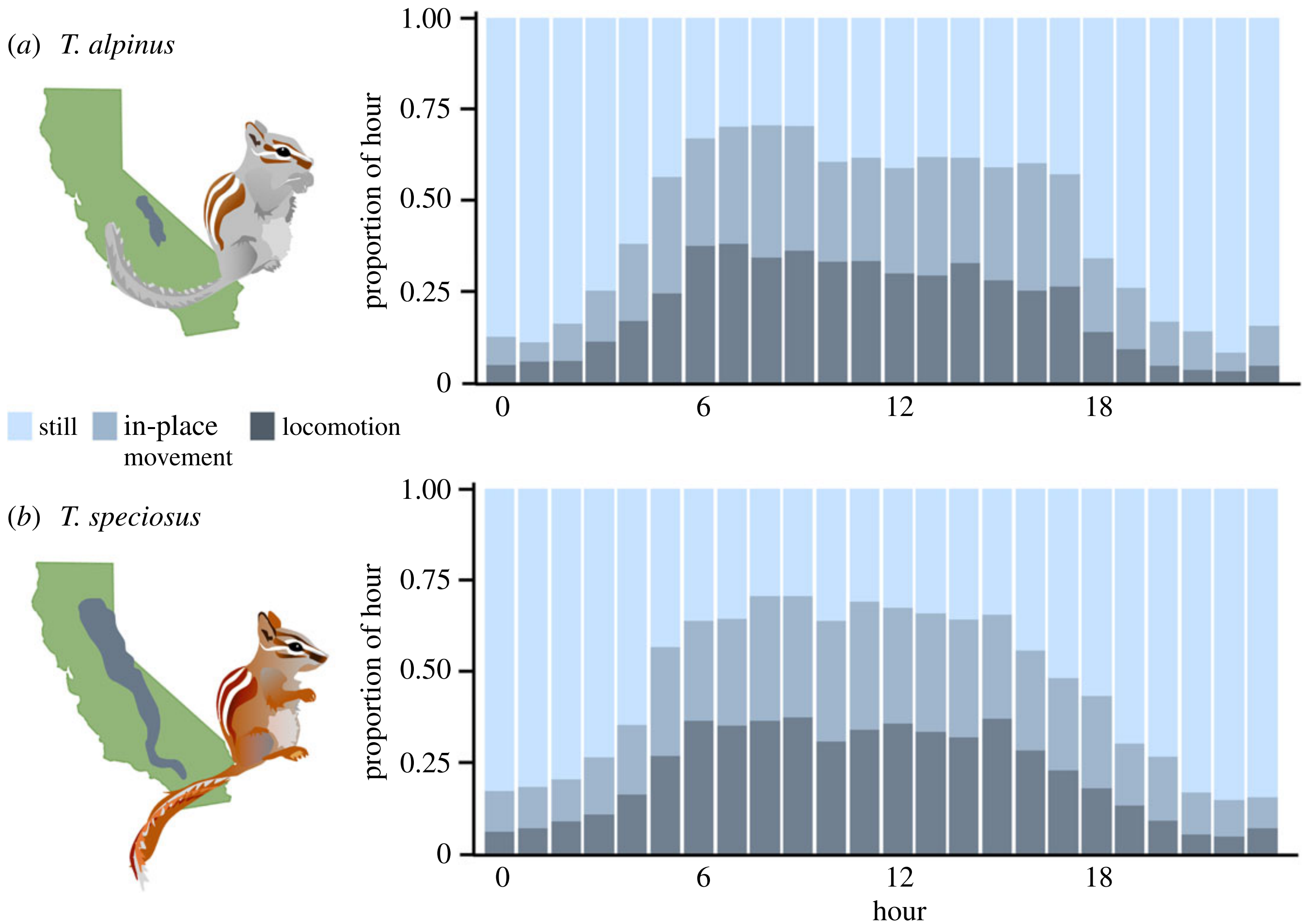



**Figure 1.** Daily activity budgets for *T. alpinus* (*a*) and *T. speciosus* (*b*). Mean proportion of each hour spent still (light shading), moving in place (medium shading) or in locomotion (dark shading) is shown; no significant differences in activity were found between species. Species distributions are shown on the left. (Online version in colour.)


BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Animal behaviour


# Ecological specialization, variability in activity patterns and response to environmental change


Talisin T. Hammond[1,2], Rupert Palme[3] and Eileen A. Lacey[1]

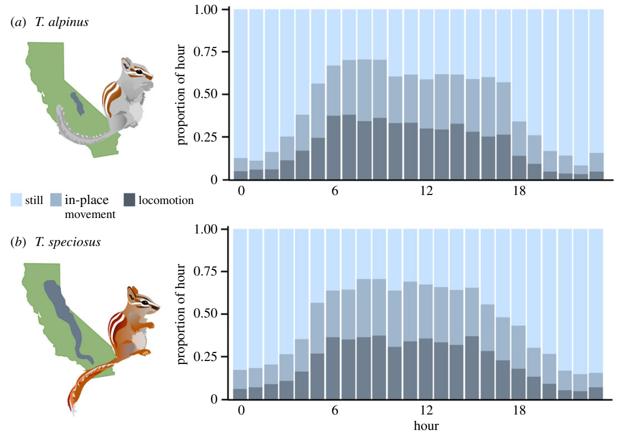
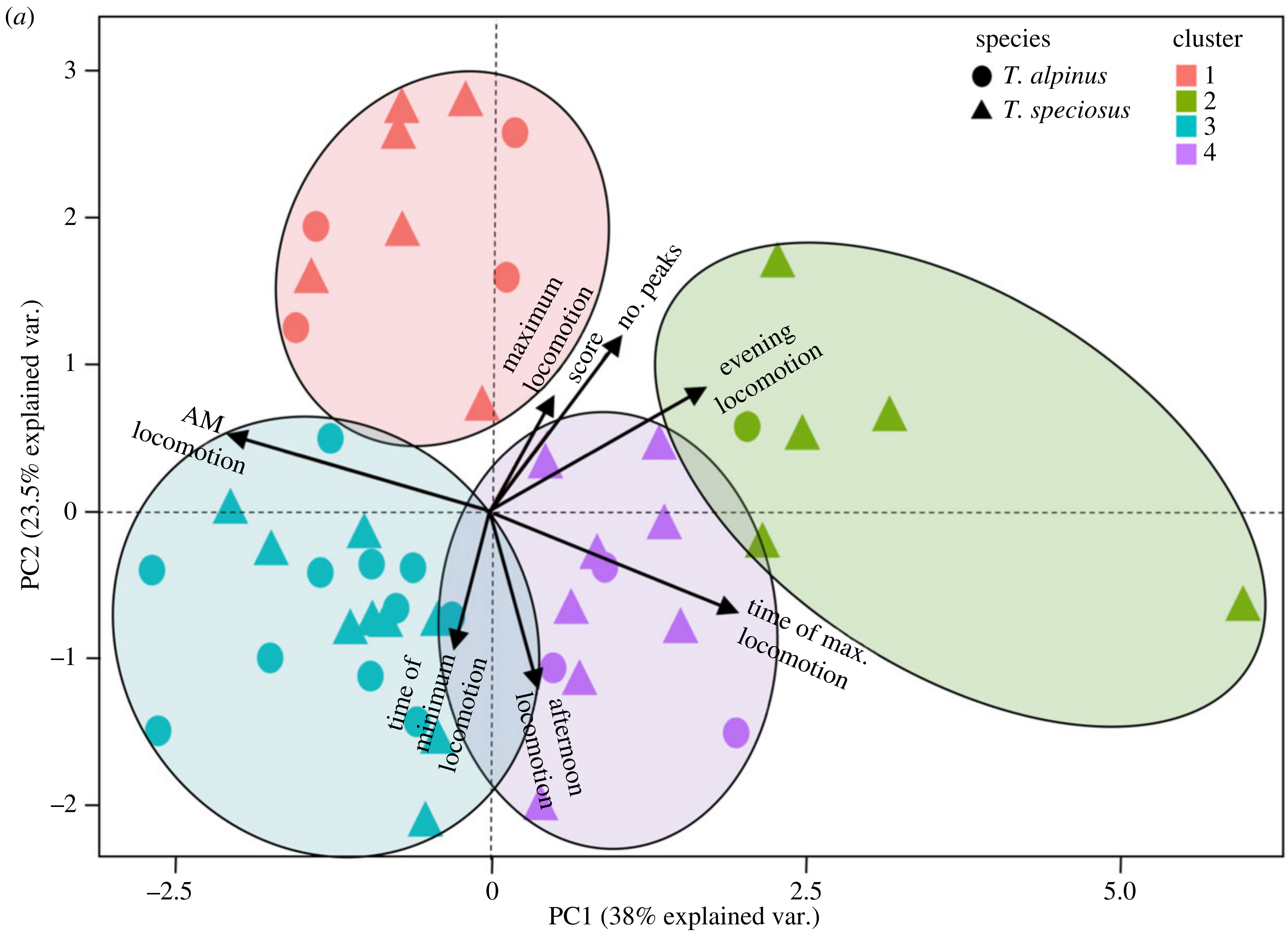



**Figure 1.** Daily activity budgets for *T. alpinus* (*a*) and *T. speciosus* (*b*). Mean proportion of each hour spent still (light shading), moving in place (medium shading) or in locomotion (dark shading) is shown; no significant differences in activity were found between species. Species distributions are shown on the left. (Online version in colour.)

From following individual activity, 7 variables were generated per individual

| feature | description |
|---|---|
| maximum | magnitude of maximum locomotion |
| time of maximum | time of maximum locomotion |
| time of minimum | time of minimum locomotion |
| afternoon (∼10:45 – 15:15) locomotion | area under the curve (AUC) of afternoon hours divided by AUC of daylight hours |
| morning (∼06:30 – 10:45) locomotion | AUC of morning hours divided by AUC of daylight hours |
| evening (∼15:15 – 19:30) locomotion | AUC of evening hours divided by AUC of daylight hours |
| no. of peaks | modality of locomotion curve (e.g. bimodal = 2) |




Clustering of activity patterns. A biplot showing clustering of activity data along the first two PCs of locomotion-based features

# The basis of k-means

- Partition *n* points (observations) across multiple variables into k groups.

- The goal is to minimize an objective function (here the sum-of-squares of multivariate distances (Euclidean) within groups).

number of clusters

number of cases

centroid for cluster *j*

case *i*

$$\text{objective function} \leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

# Learning from the data – Machine learning algorithms: k – means

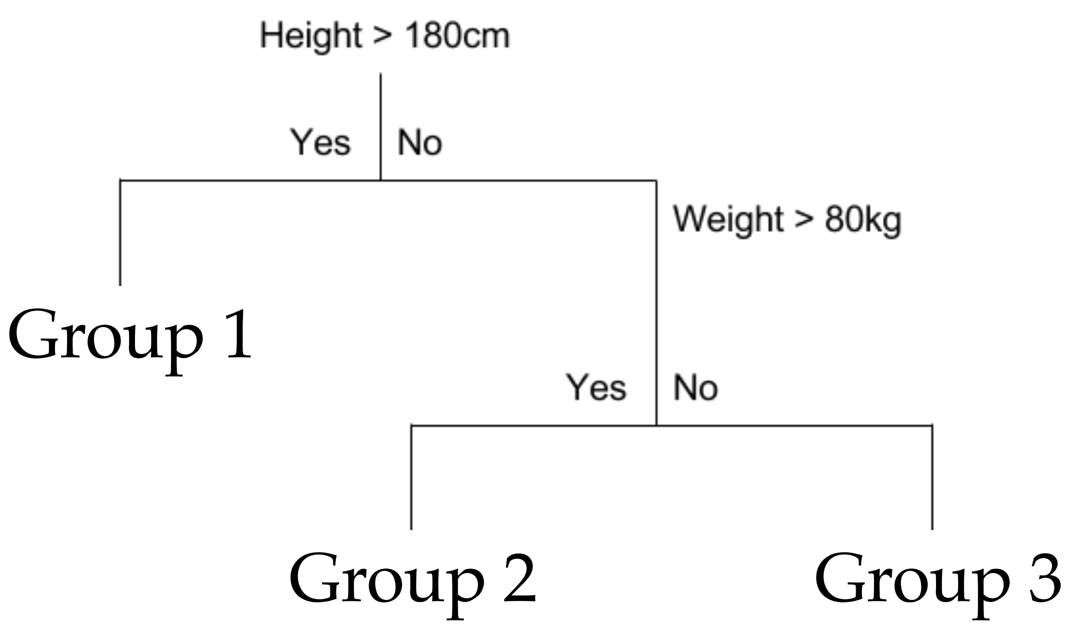
We will consider only two dimensions here for visual simplicity (Height and Weight)

Unlabeled data

**Unsupervised Learning Algorithm**

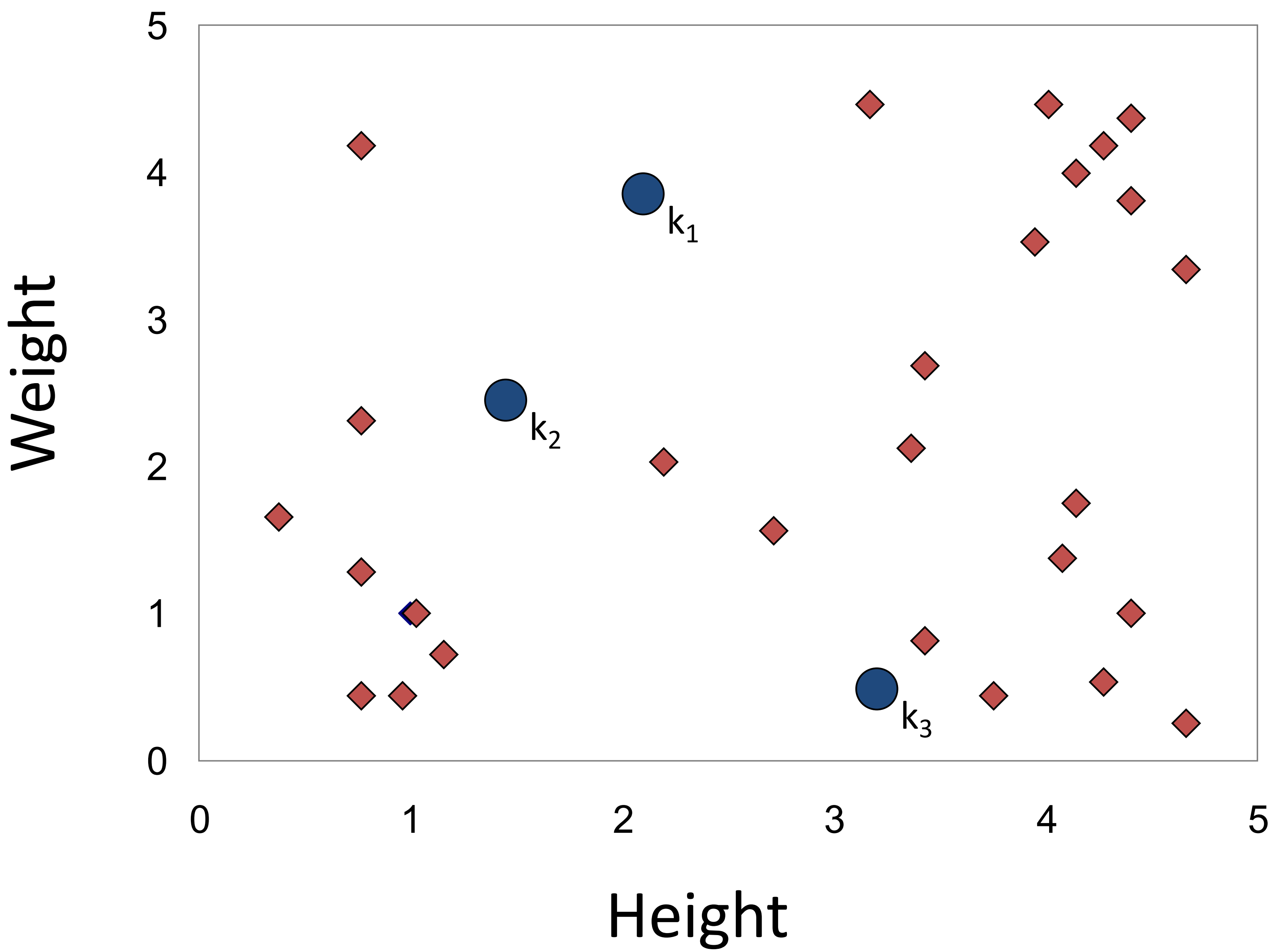
Prediction based on finding patterns in the data

e.g., Finding number of groups in data and ways to classify (predict) observations based on their characteristics (height/weight)



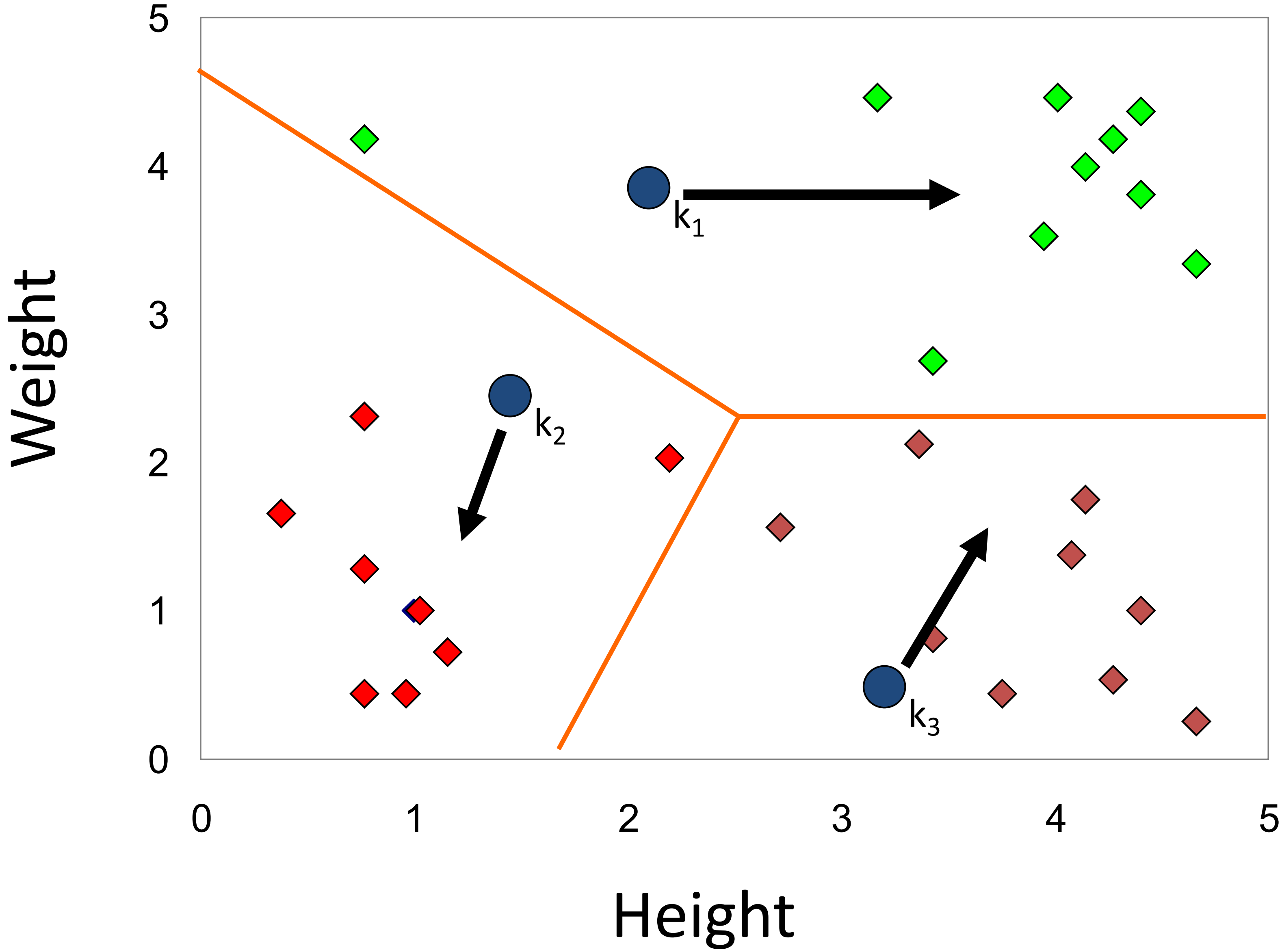
# K-means Clustering – steps 1 & 2

1. Clusters the data into $k$ groups where $k$ is predefined.
2. Select $k$ points at <u>*random*</u> as cluster centers.



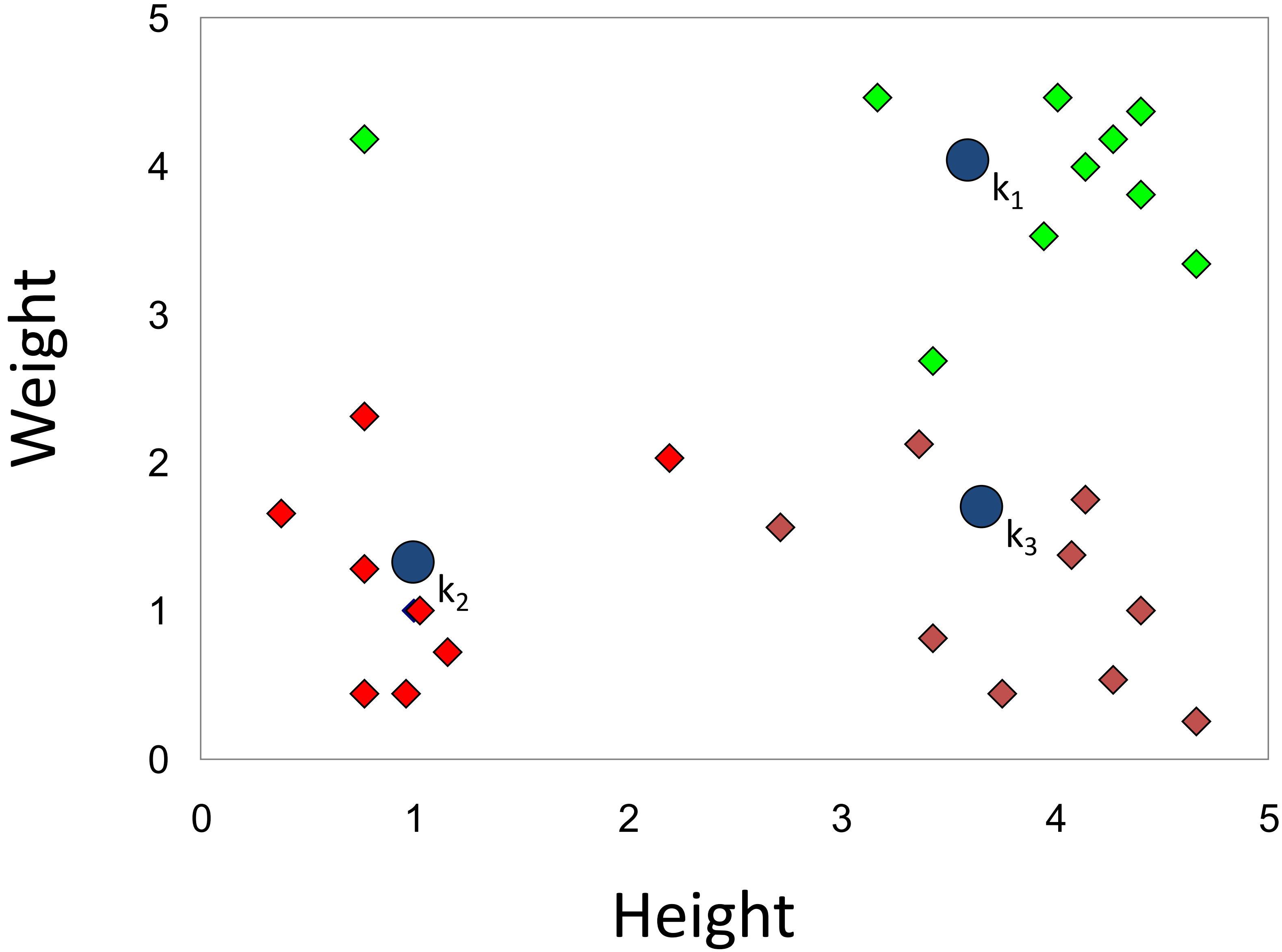
# K-means Clustering – step 3

Assign objects to their closest cluster center according to the *Euclidean distance* function.



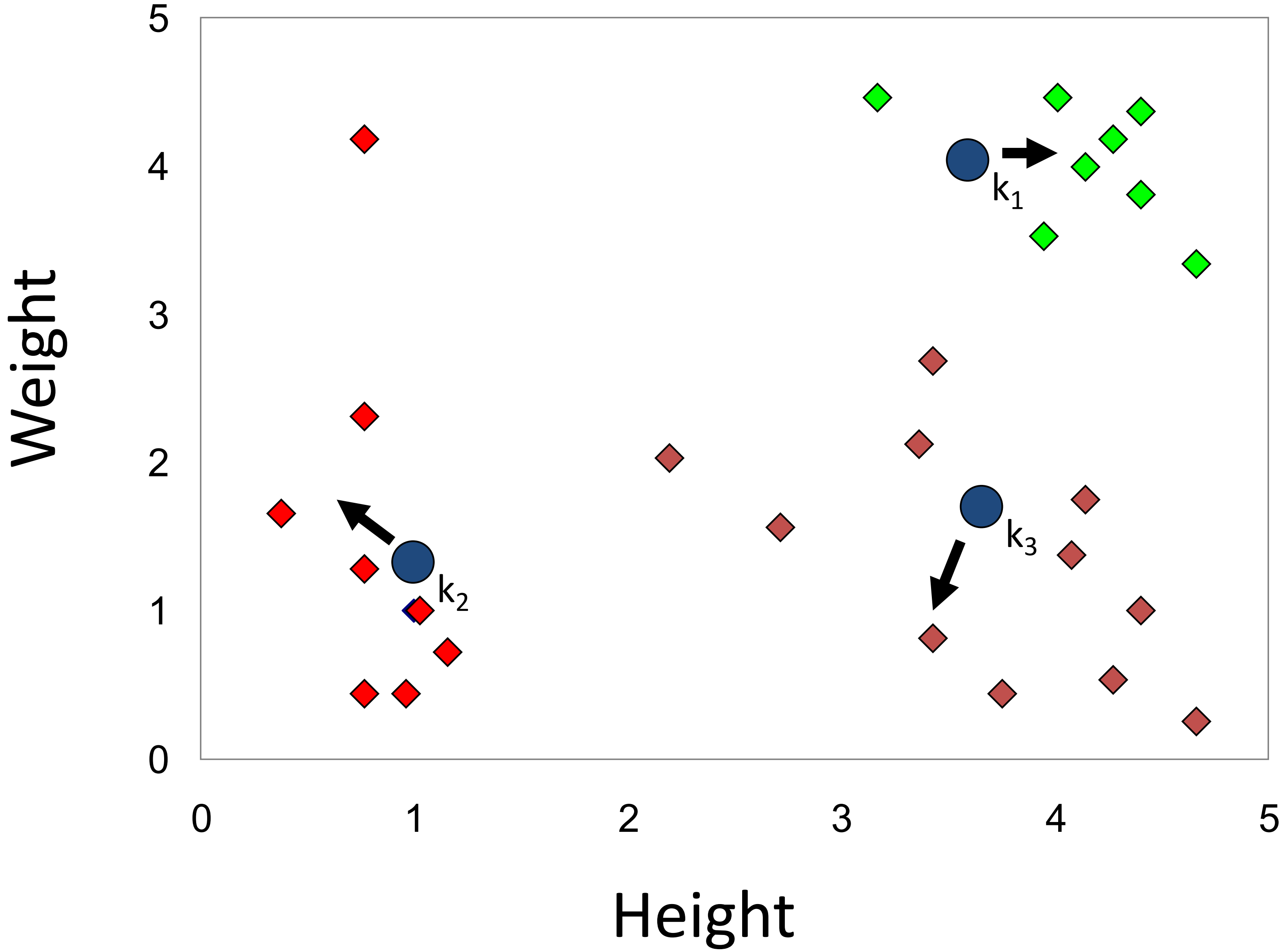
# K-means Clustering – step 4

Calculate the centroid or mean of all objects in each cluster.

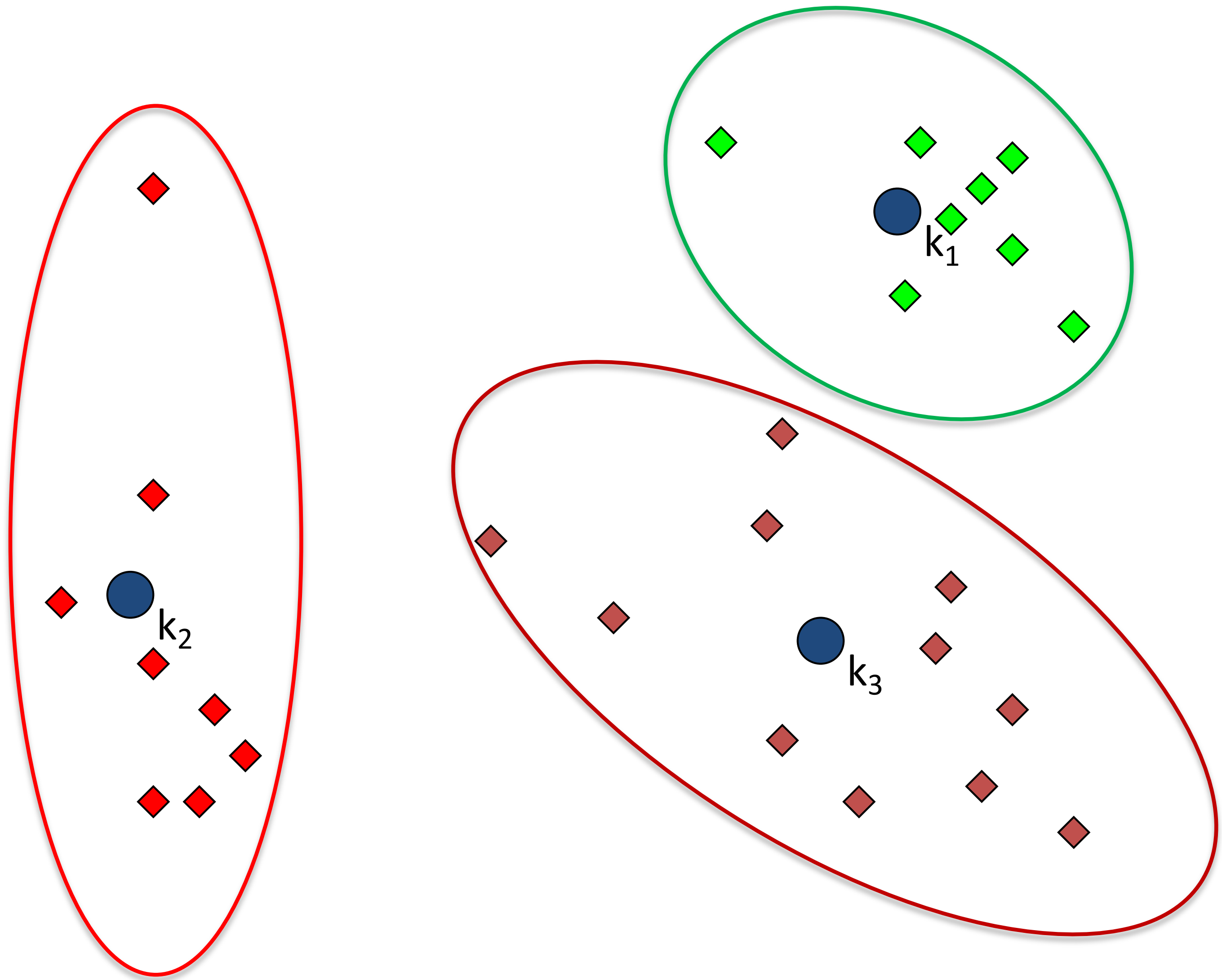


# K-means Clustering – step 5

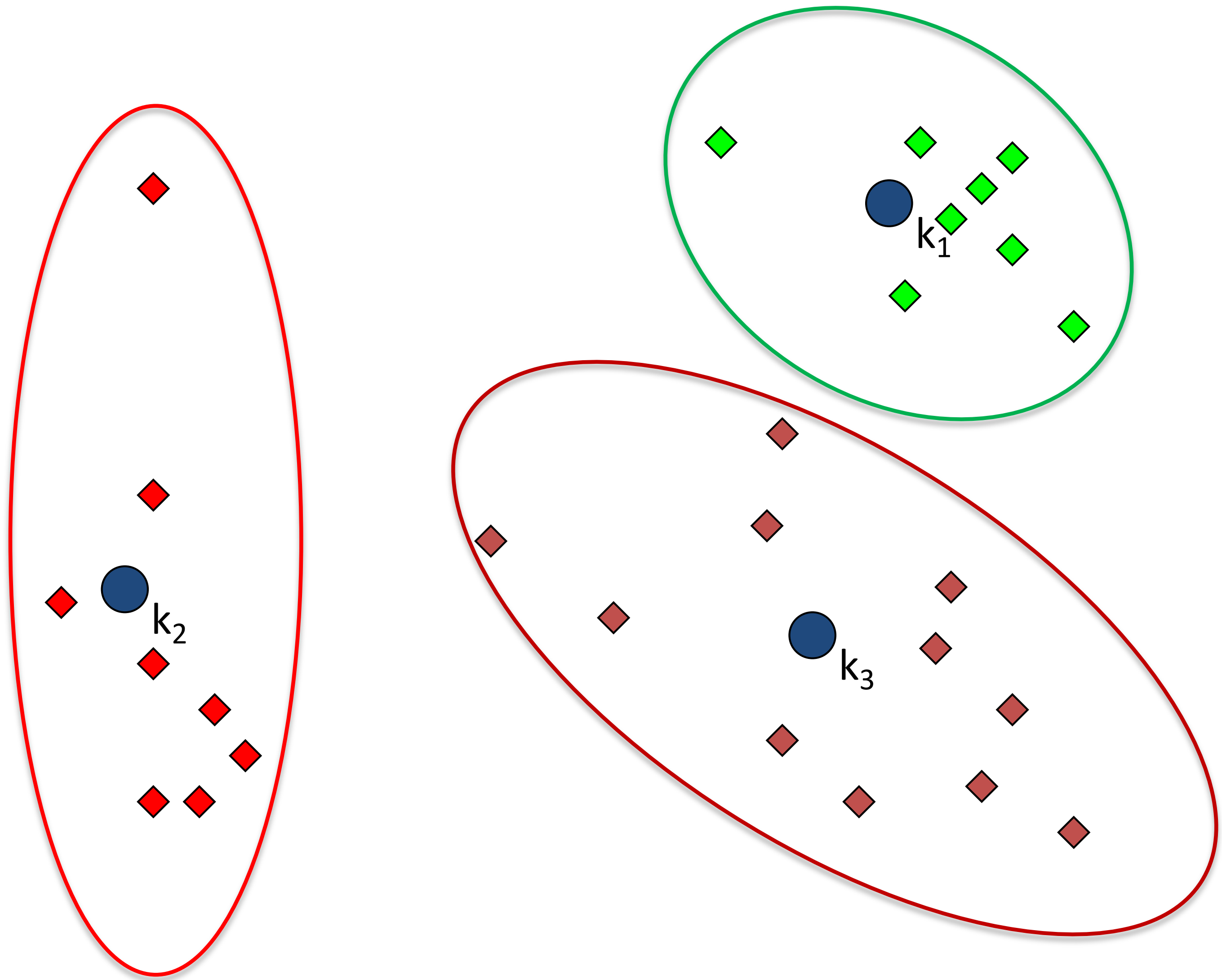
Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.



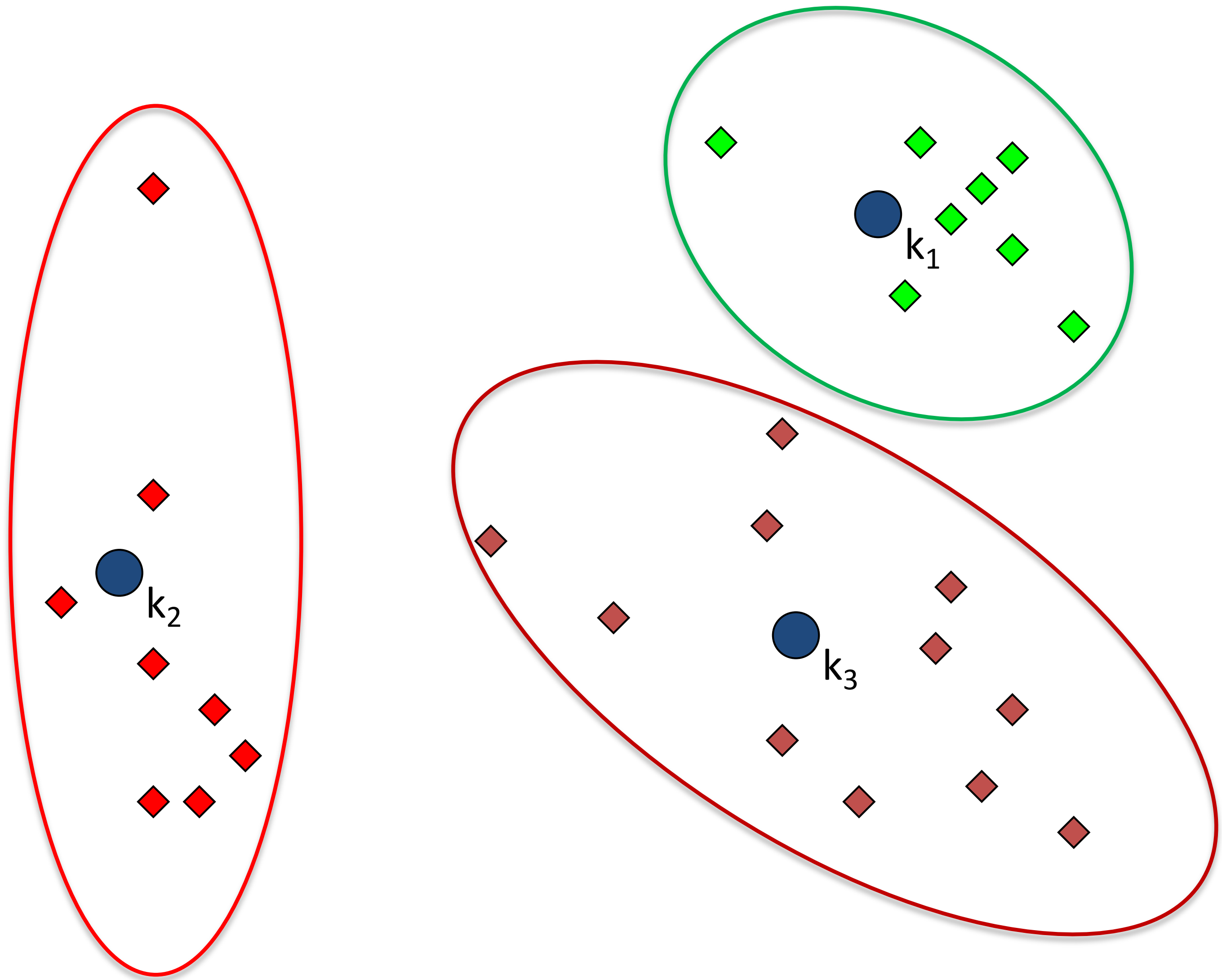
K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



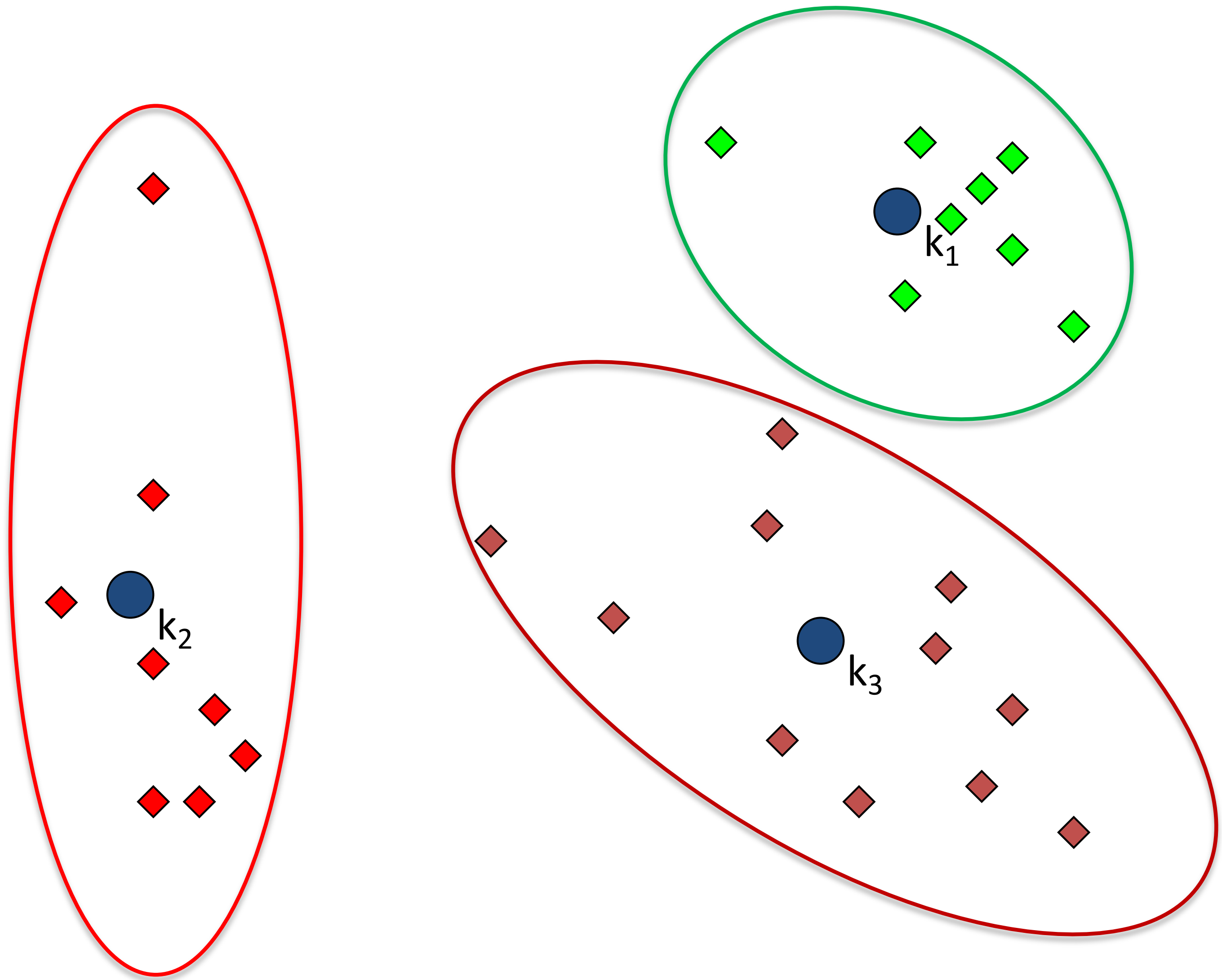
K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



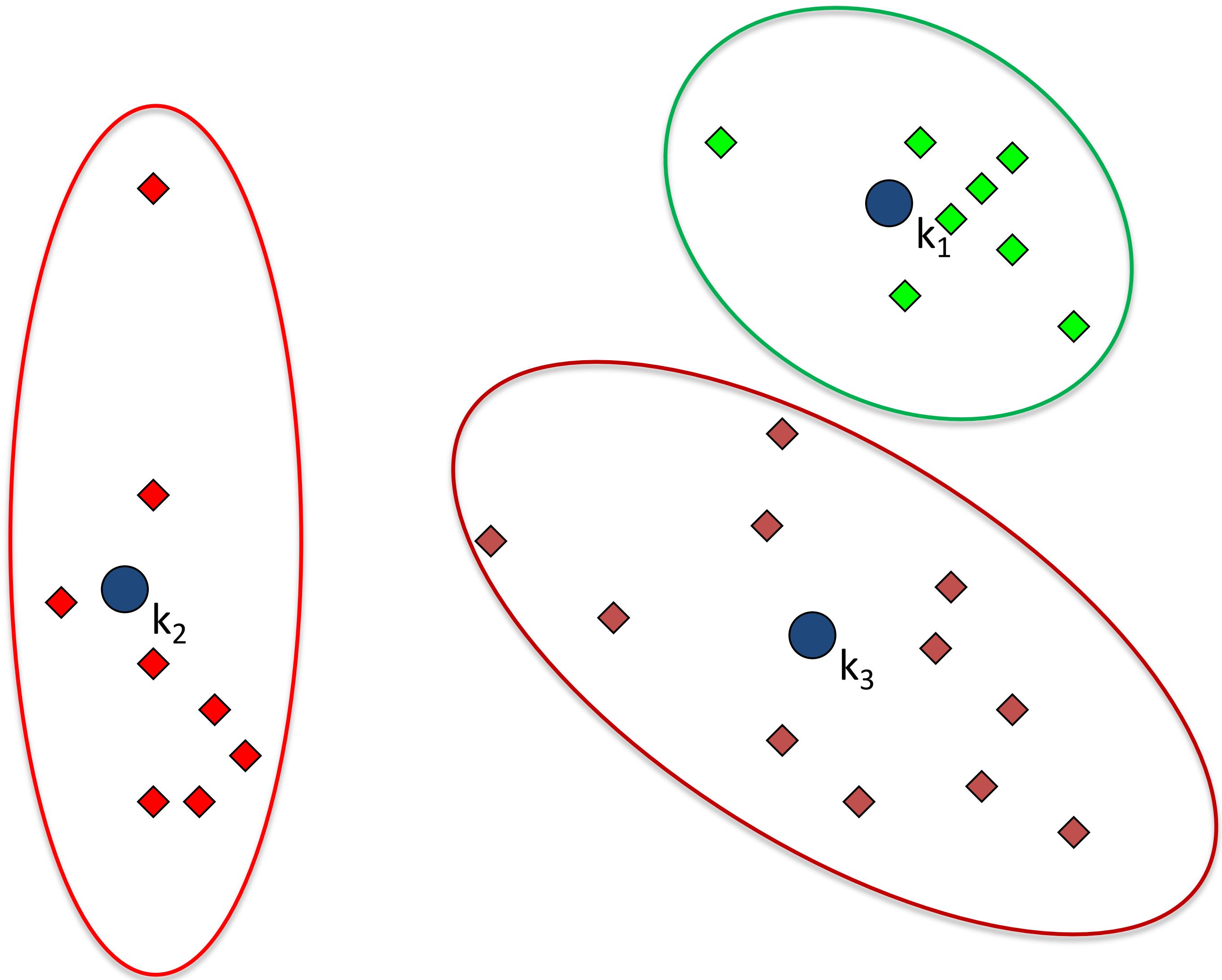
K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



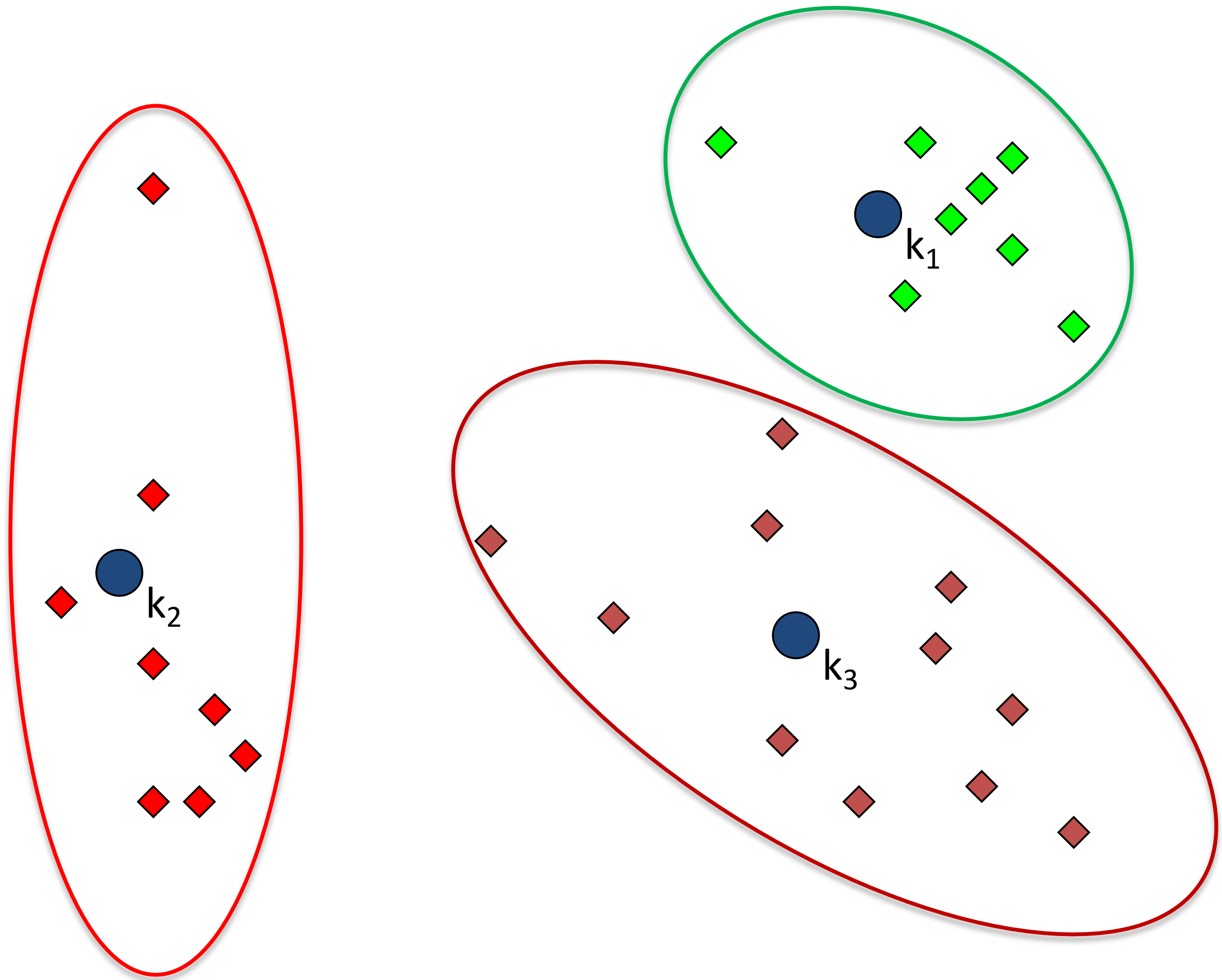
K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



# K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



# K-means Clustering: final cluster – no more movements that greater improve the fit (threshold = 0.00001) are possible



# The (iterative) k-means algorithm (summary of "general" algorithm – there are others)

The number of clusters, k, is decided first; the iterative steps are then:

1) Generate an initial set of k points as the first estimate of the cluster points (random seed points).

2) Loop over all observations reassigning them to the group with the closest mean value.

3) Re-compute the mean of each group.

Iterate steps 2 and 3 until convergence (i.e., the mean distance of each object to its group mean does not change according to a very small threshold (e.g., 0.000001).

# The (iterative) k-means algorithm (summary of “general” algorithm – there are others)

The number of clusters, k, is decided first; the iterative steps are then:

1) Generate an initial set of k points as the first estimate of the cluster points (random seed points).

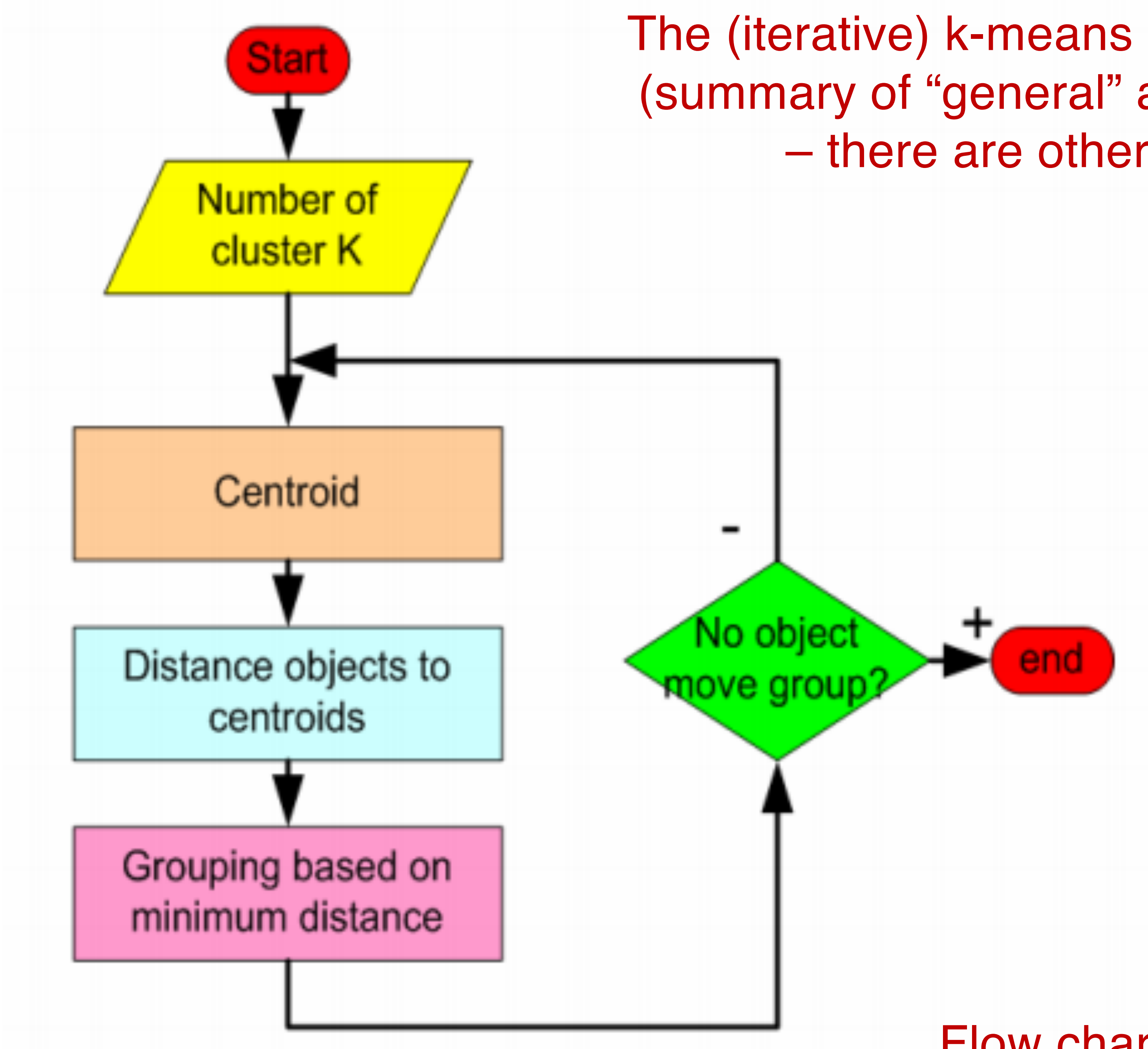
2) Loop over all observations reassigning them to the group with the closest mean value. Assign objects to their closest cluster center according to the *Euclidean distance* function.

3) Re-compute the mean (multivariate centroids) of each group.

Iterate steps 2 and 3 until convergence (i.e., the mean distance of each object to its group mean does not change according to a very small threshold (e.g., 0.000001).

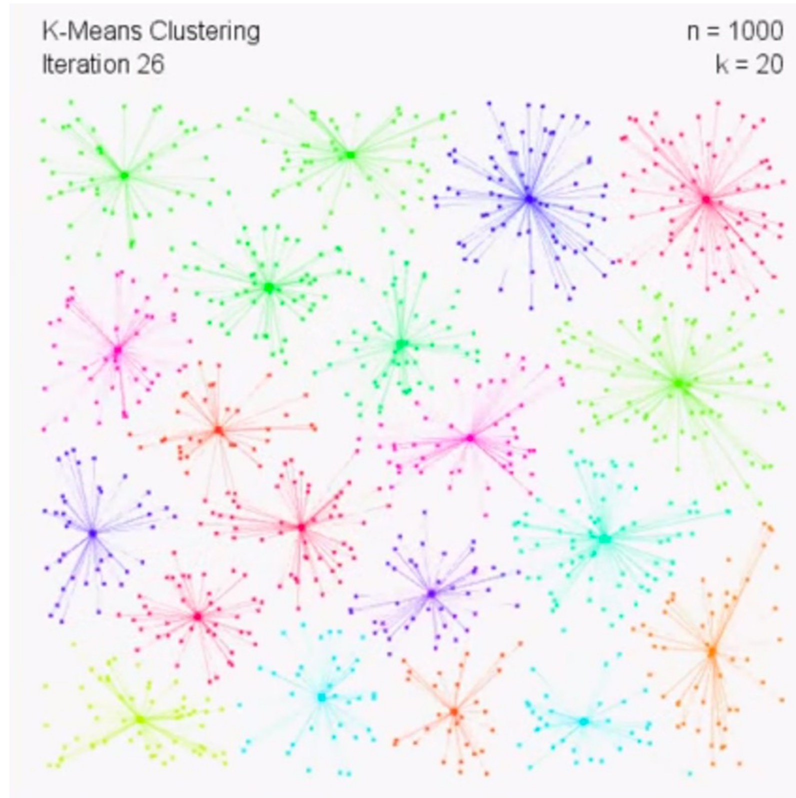
An **iterative method** is called convergent if the corresponding sequence converges regardless of the initial approximations (random seed points).

The (iterative) k-means algorithm
(summary of “general” algorithm
– there are others)




Flow chart version

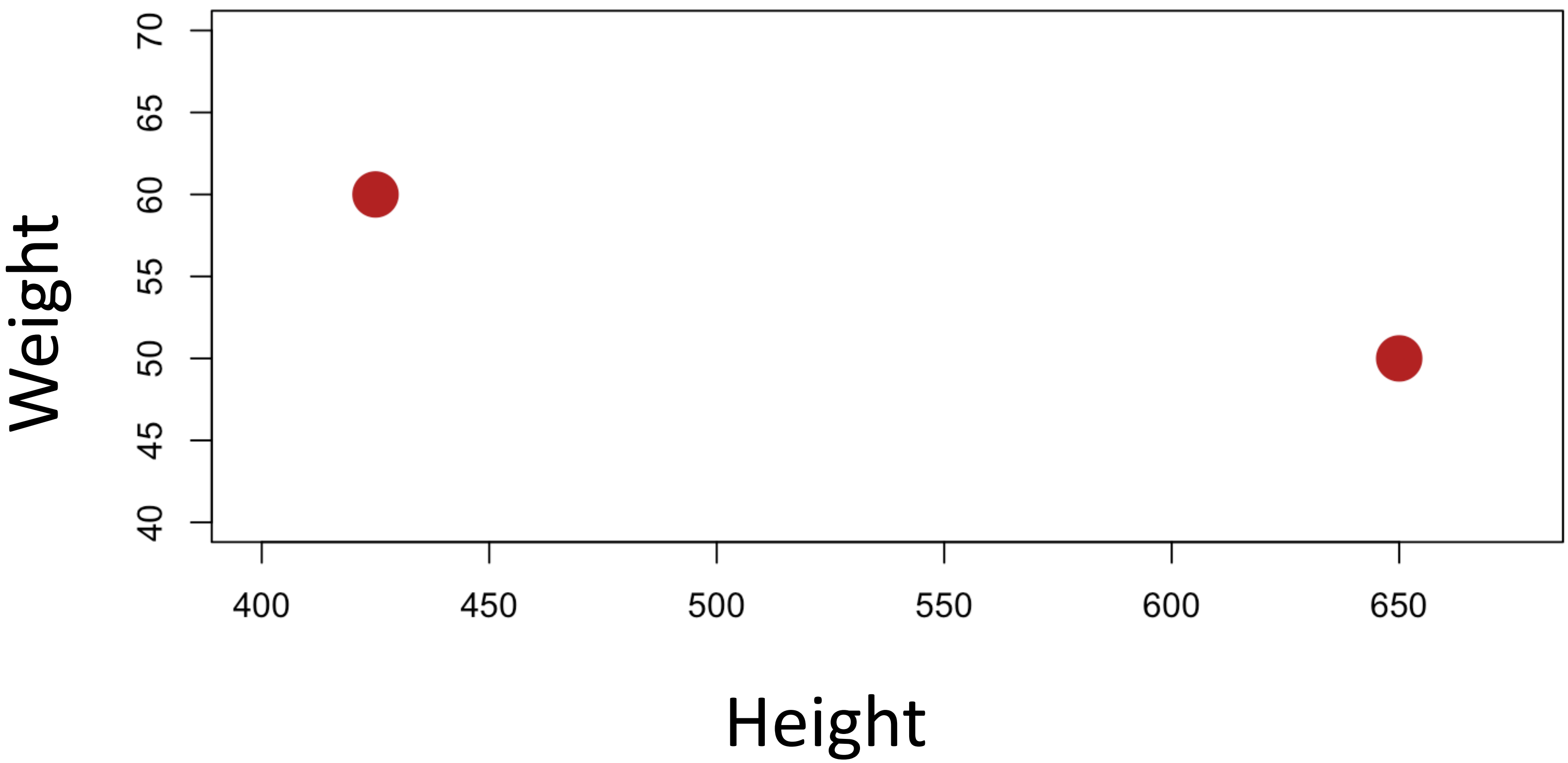
# K-means Clustering: number of groups (k) and number of iterations (moving objects) (n)

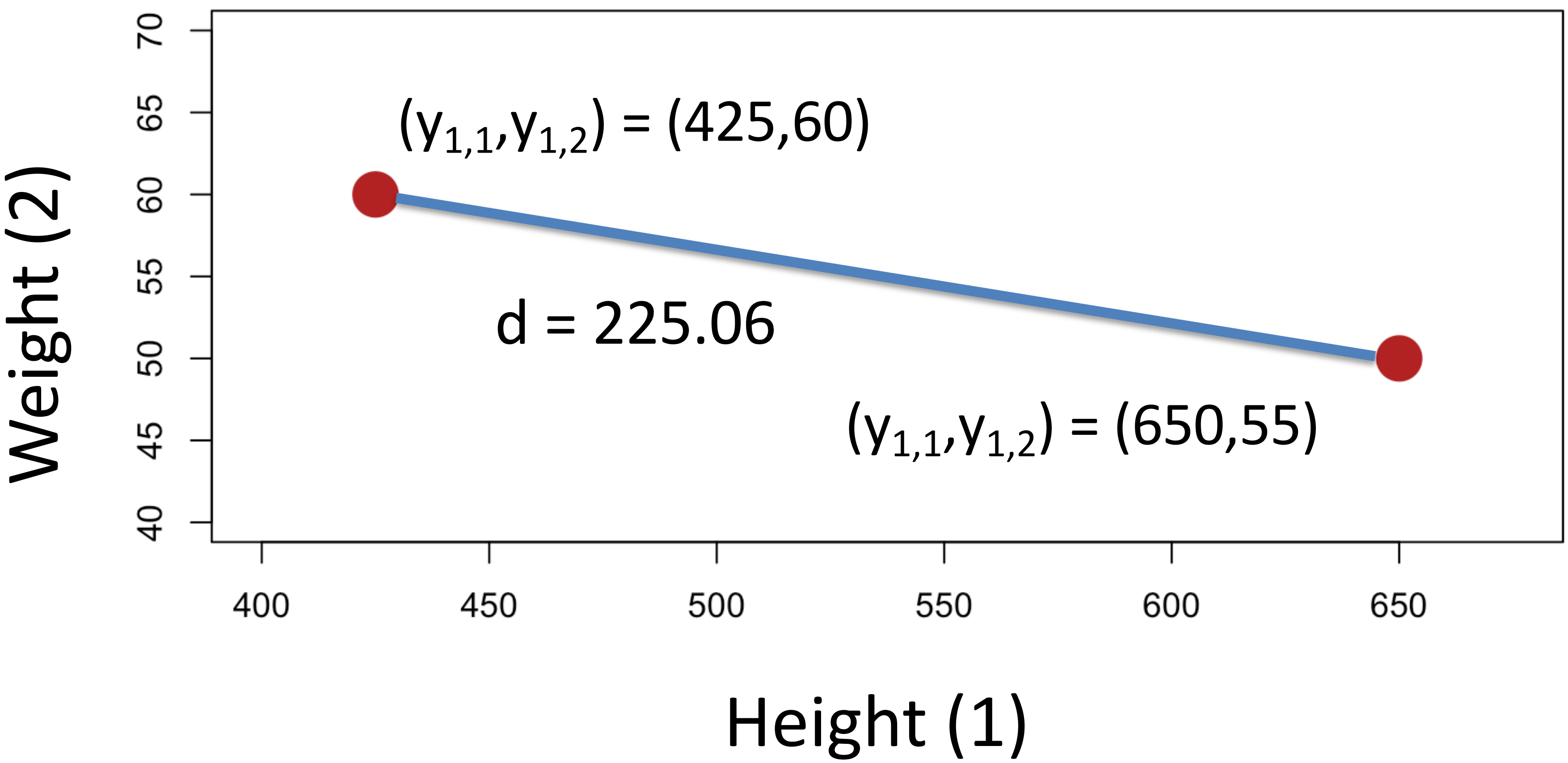

https://www.youtube.com/watch?v=BVFG7fd1H30

# Measuring fit of the k-means clustering

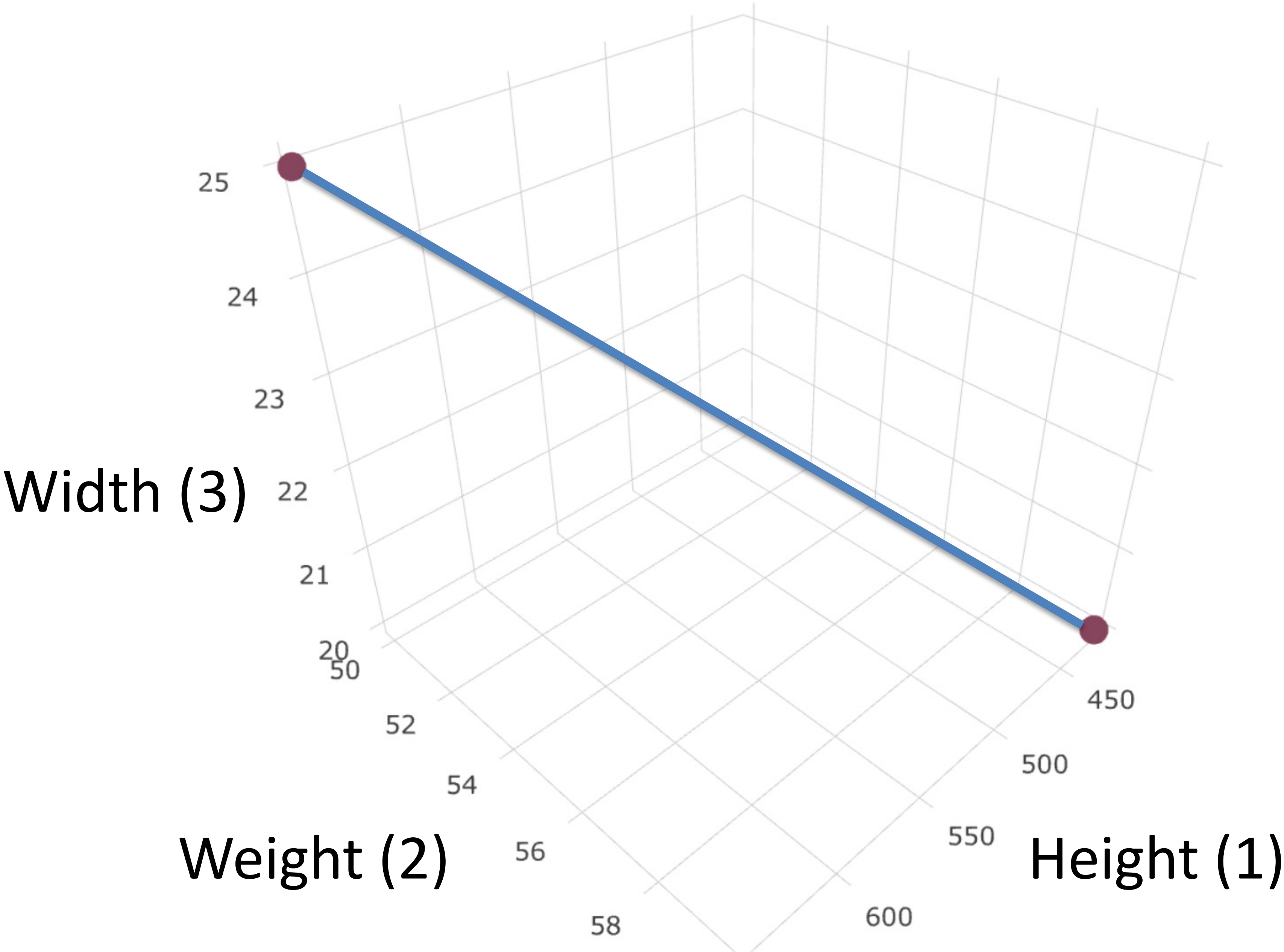
# Fit metrics are based on Euclidean distances! How are they calculated? 2 dimensions



# How are Euclidean distances calculated? 2 dimensions



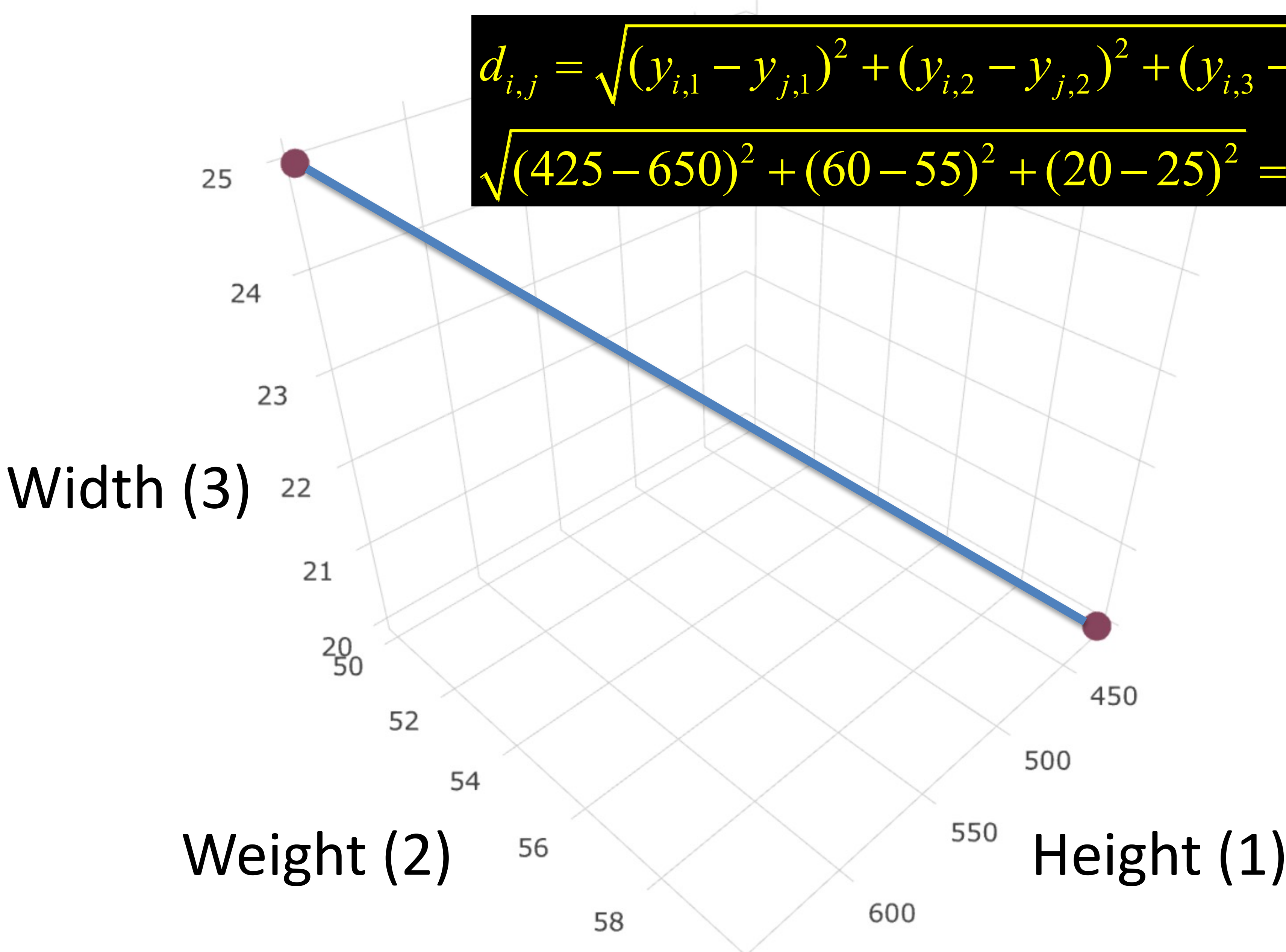

$$d_{i,j} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2} = \sqrt{(425 - 650)^2 + (60 - 55)^2} = 225.06$$




## Euclidean distance – 3 dimensions

$$d_{i,j} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2 + (y_{i,3} - y_{j,3})^2} =$$

$$\sqrt{(425 - 650)^2 + (60 - 55)^2 + (20 - 25)^2} = 225.11$$




Euclidean distance – 3 dimensions

# Euclidean distance in *p* (here 5) dimensions

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| | | | | |
| 0.086 | 0.465 | 0.144 | 0.760 | 0.229 |
| 0.651 | 0.790 | 0.982 | 0.844 | 0.413 |
| 0.791 | 0.730 | 0.178 | 0.282 | 0.805 |
| 0.409 | 0.637 | 0.119 | 0.468 | 0.364 |
| 0.984 | 0.701 | 0.879 | 0.570 | 0.098 |
| 0.093 | 0.268 | 0.115 | 0.357 | 0.104 |
| 0.164 | 0.294 | 0.143 | 0.028 | 0.044 |
| 0.623 | 0.879 | 0.329 | 0.217 | 0.139 |
| 0.668 | 0.651 | 0.048 | 0.179 | 0.987 |
| 0.071 | 0.846 | 0.715 | 0.909 | 0.653 |
| 0.659 | 0.432 | 0.595 | 0.523 | 0.241 |
| 0.928 | 0.274 | 0.344 | 0.189 | 0.634 |
| 0.877 | 0.451 | 0.223 | 0.517 | 0.872 |
| 0.281 | 0.836 | 0.172 | 0.349 | 0.179 |
| 0.373 | 0.773 | 0.050 | 0.439 | 0.924 |

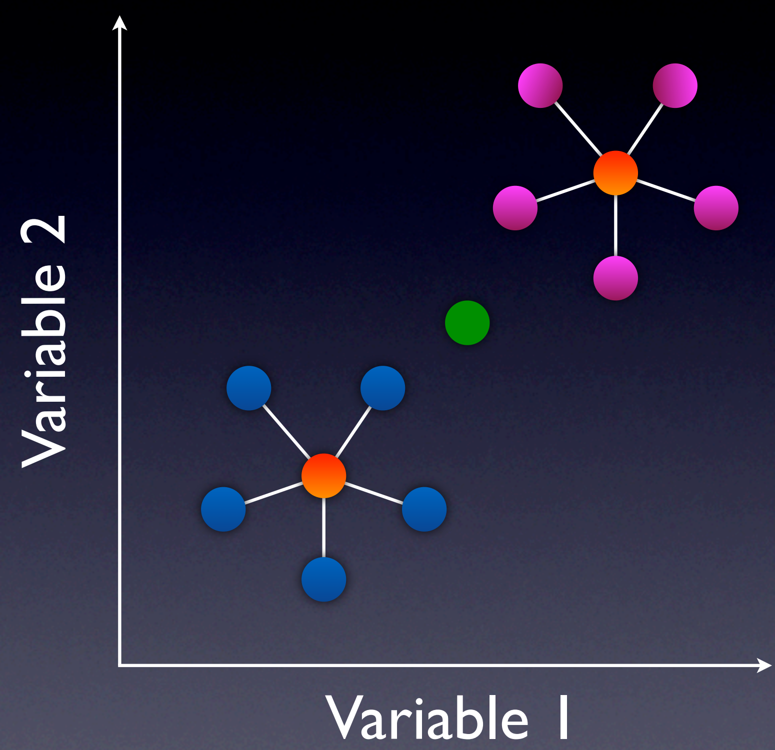
$$d_{12} = \sqrt{\begin{array}{l}(0.086 - 0.651)^2 + \\ (0.465 - 0.790)^2 + \\ (0.144 - 0.982)^2 + \\ (0.760 - 0.844)^2 + \\ (0.229 - 0.413)^2\end{array}}$$

# K – means clustering method
# Assessing quality of the clustering in *k* groups
# (minimize distances of points within clusters)

## The within-groups sum-of-squares (SSw)

sum of squared dissimilarities between objects within each group, summed over the groups

$$SS_w = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d^2{}_{ij\in ij}$$

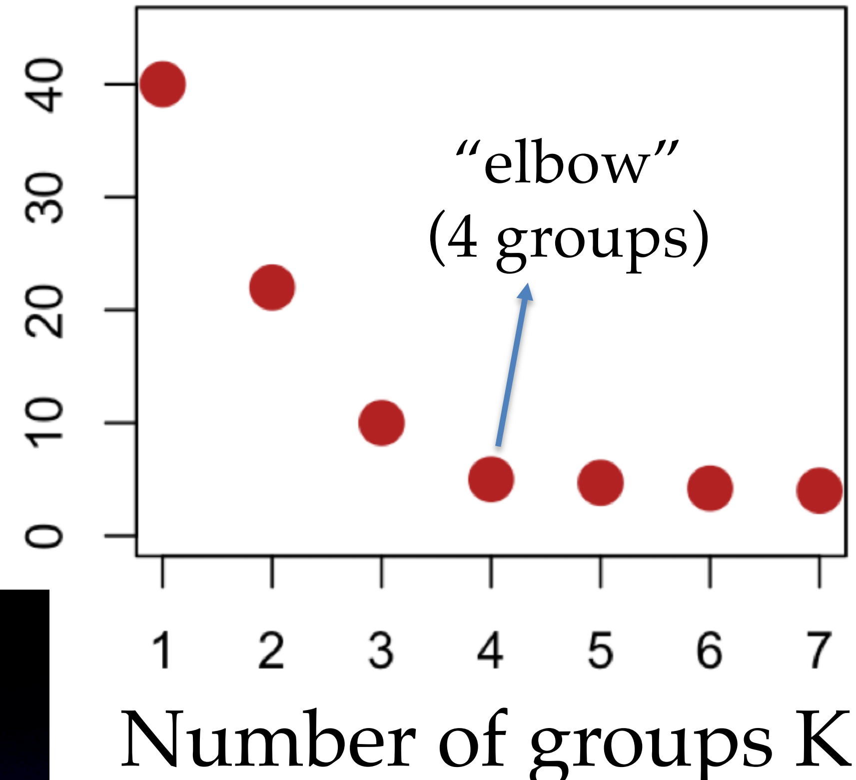


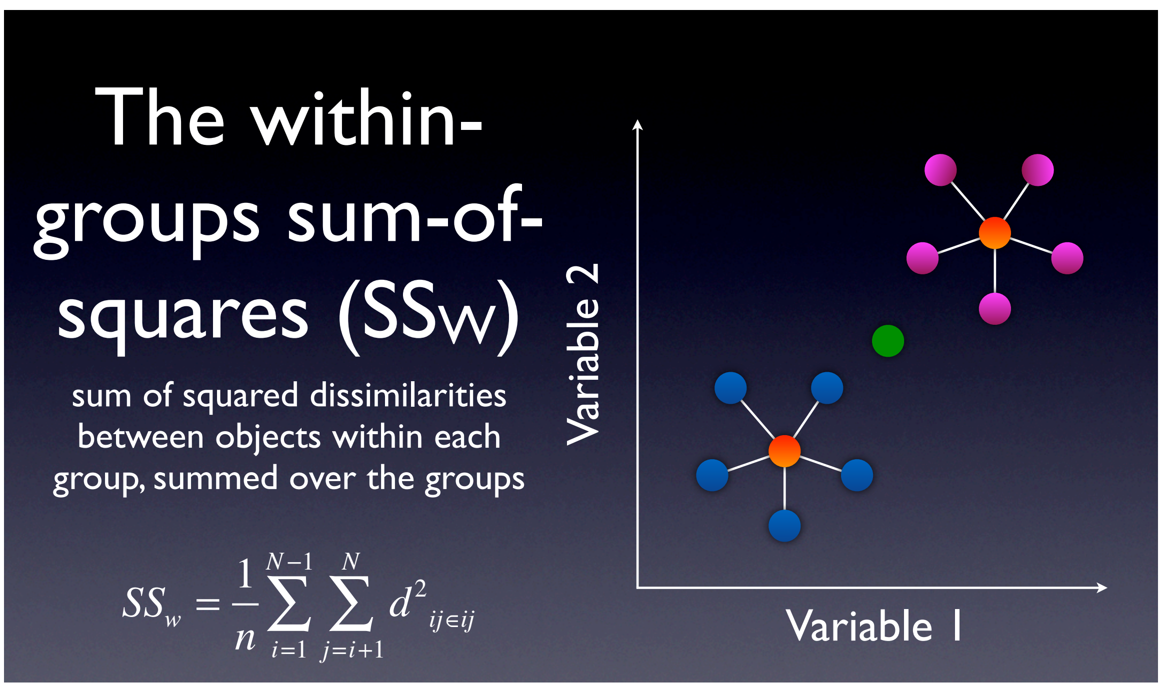

We would like to produce clusters with the smallest possible SSw.

From Alejandro Ordonez

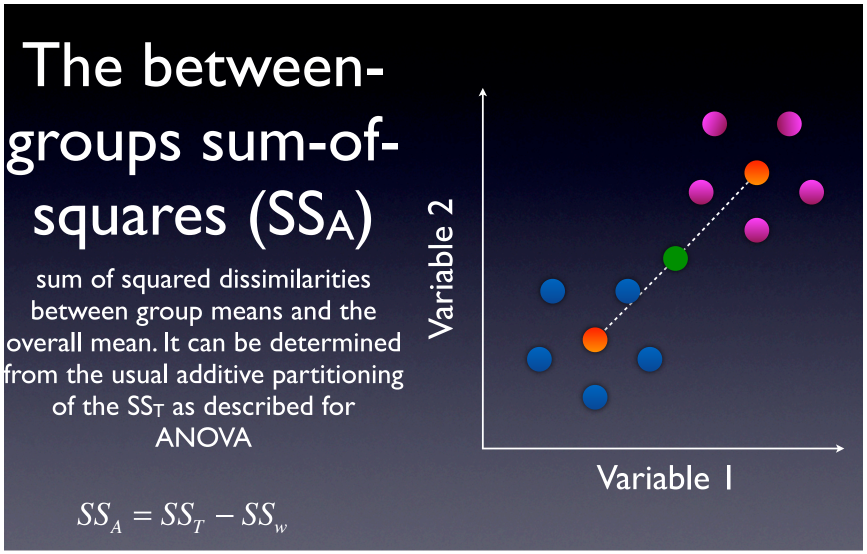
What is the optimal number of group?
lots of methods, e.g., the "elbow method"

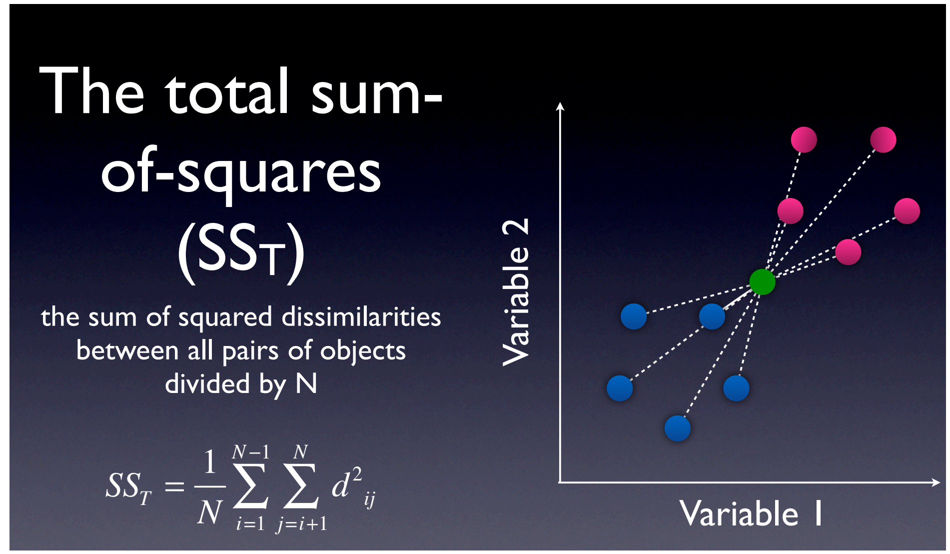
SSw = Average within cluster distance to centroid



# K – means clustering method
# Quality of the clustering in k groups : $SS_A / SS_T$

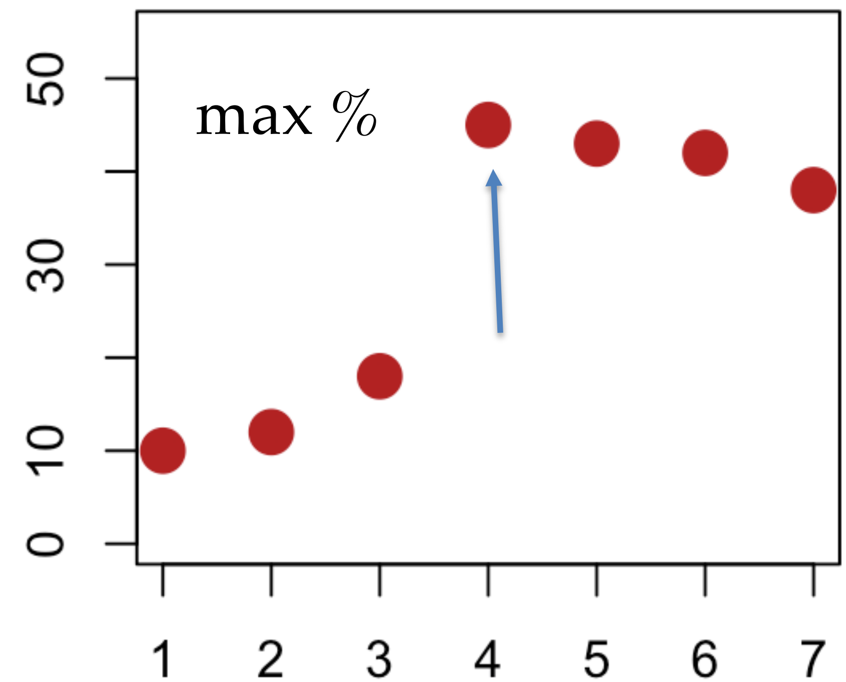



The $SS_A / SS_T$ % is a measure of the total variance in the data set that is explained by the clustering. k-means minimize the within group dispersion and maximize the between-group dispersion. By assigning the samples to k clusters rather than n (number of samples) clusters achieved a reduction in sums of squares of $SS_A / SS_T$ %.

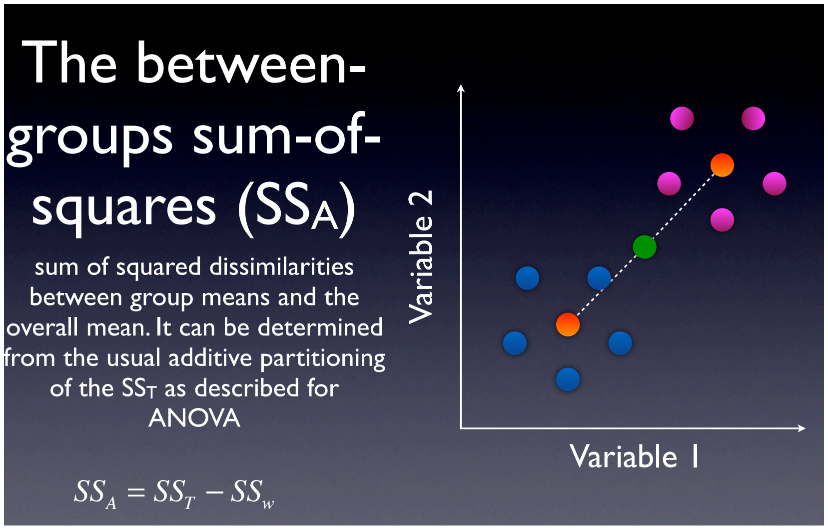
What is the optimal number of group?
lots of methods, e.g., total variance explained

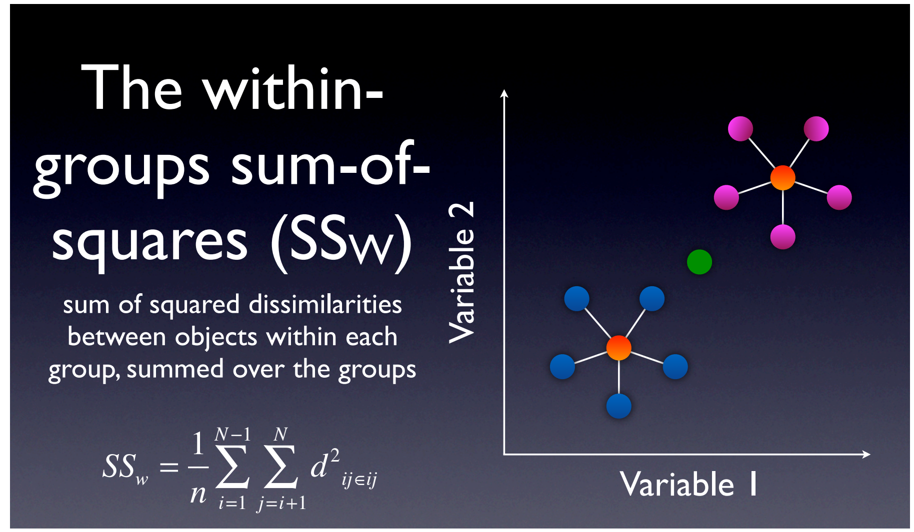
total variance explained
$(SS_A/SS_T)$




Number of groups K

# K – means clustering method
# Quality of the clustering in k groups : SSI

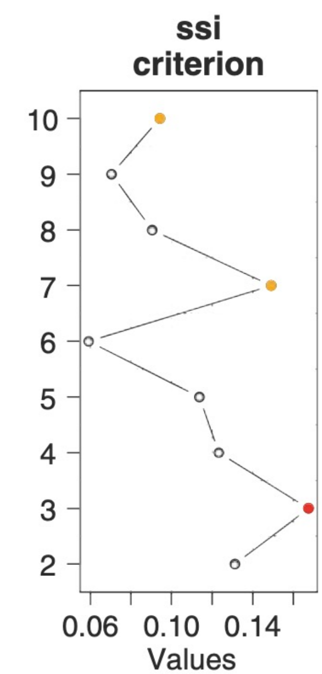


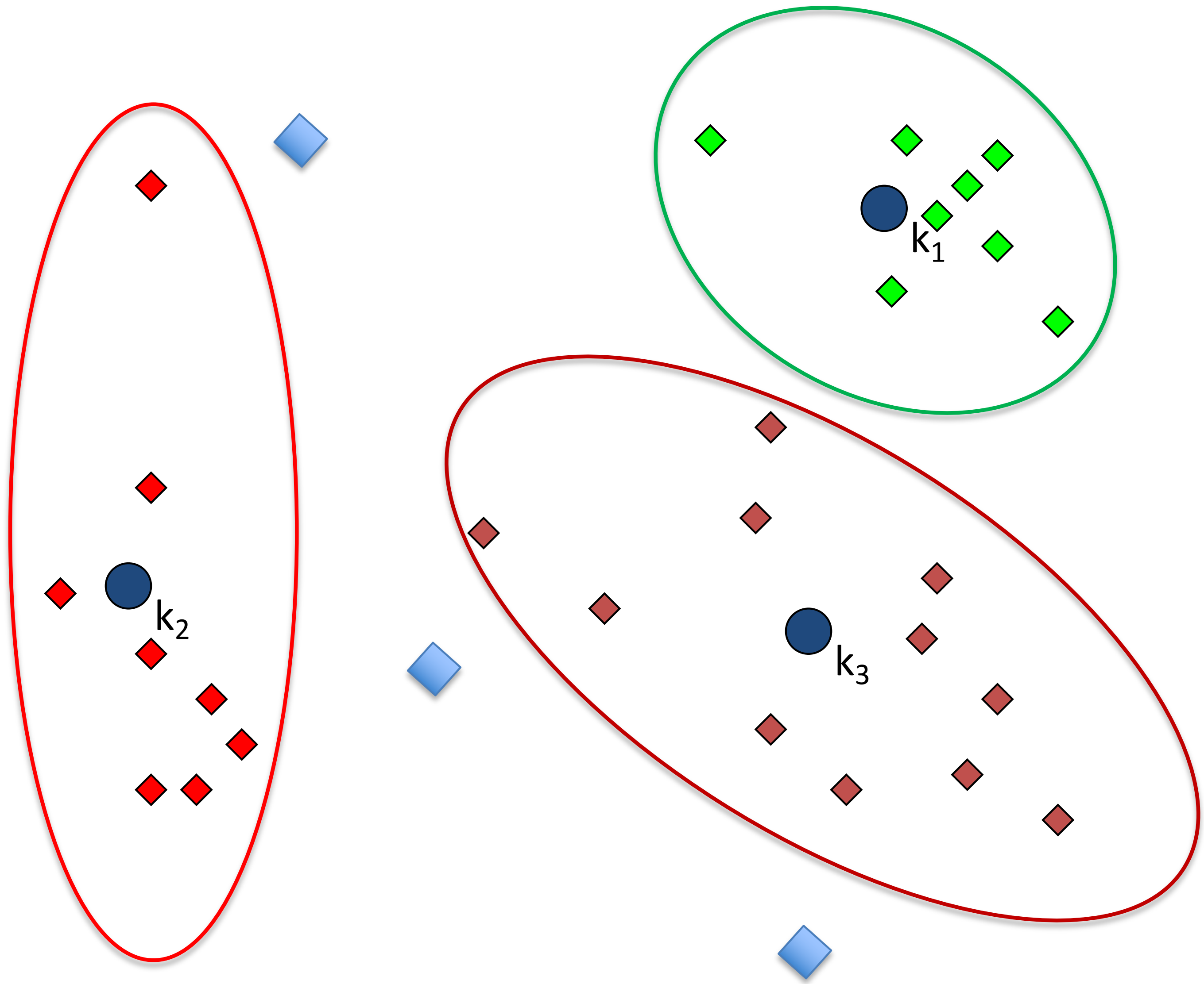

simple structure index (SSI) =

$$(SS_A/(K-1))/(SS_W/(n-K))$$

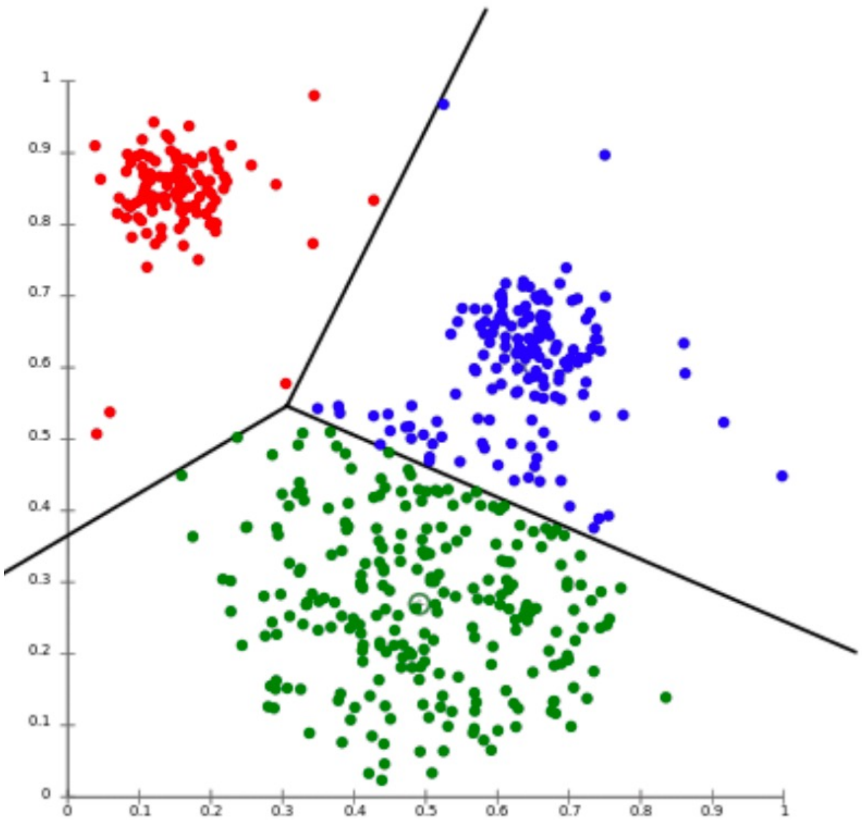
n = number of objects (observations, data points); k = number of groups

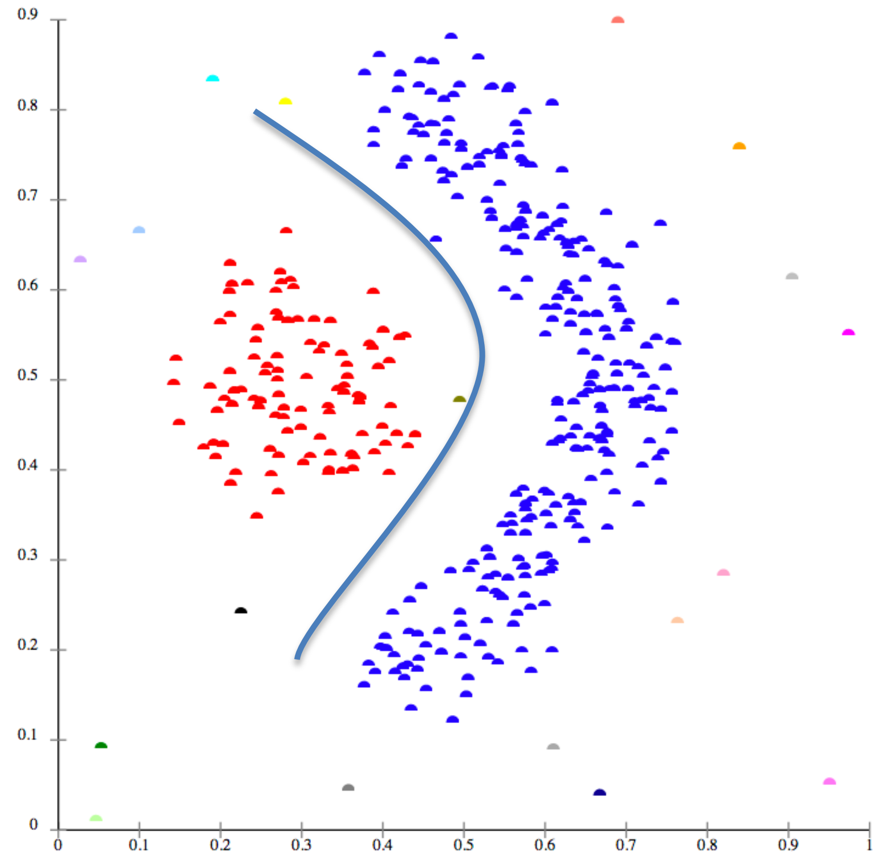


K-means as a predictive model: for each of new observations we can estimate its probability to belonging to a particular group (cluster)



# Linear versus non-linear group partitioning



## K-means is used in a variety of problems

IJCAT International Journal of Computing and Technology, Volume 1, Issue 4, May 2014
ISSN : 2348 - 6090
**www.IJCAT.org**

# Human Genome Data Clustering Using K-Means Algorithm

**[1] Amrita A. Kulkarni, [2] Prof. Deepak Kapgate**

[1] Department of C.S.E., GHRAET, Nagpur University,
Nagpur, Maharashtra, India

[2] Department of C.S.E., GHRAET, Nagpur University,
Nagpur, Maharashtra, India

# K-means is used in a variety of problems

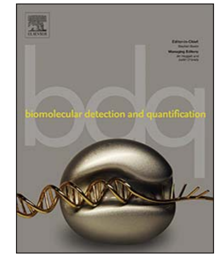
Biomolecular Detection and Quantification 13 (2017) 7–31

ELSEVIER

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq




Research paper

## *K-means and cluster models for cancer signatures

CrossMark

Zura Kakushadze[a,b,1,*], Willie Yu[c]

COMPUTATIONAL GENOMICS APPROACHES TO PRECISION MEDICINE

# K-means is used in a variety of problems

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY


Volume 4, Issue 6
June 2013
Pages 542-551



Research Article | Free Access

## Spherical k-means clustering is good for interpreting multivariate species occurrence data

Mark. O. Hill, Colin A. Harrower, Christopher D. Preston

First published: 2 April 2013 | https://doi.org/10.1111/2041-210X.12038 | Cited by:3


Frontiers of Environmental Science & Engineering
February 2014, Volume 8, Issue 1, pp 117–127 | Cite as

## Application of *k*-means clustering to environmental risk zoning of the chemical industrial area

Authors | Authors and affiliations

Weifang Shi, Weihua Zeng

# K-means is used in a variety of problems

## A protocol for classifying ecologically relevant marine zones, a statistical approach

Els Verfaillie [a,b,*], Steven Degraer [c,d], Kristien Schelfaut [a,e], Wouter Willems [c], Vera Van Lancker [a,d]

Estuarine, Coastal and Shelf Science 83 (2009) 175–185

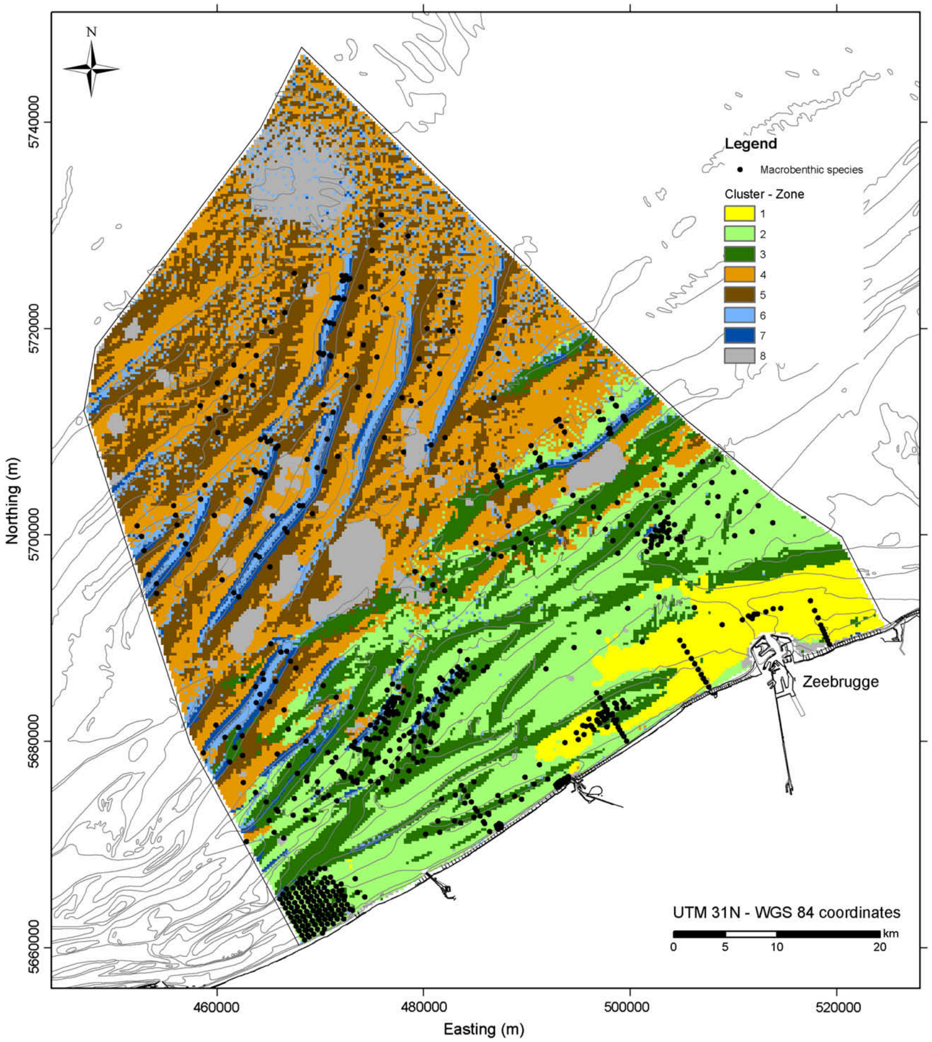
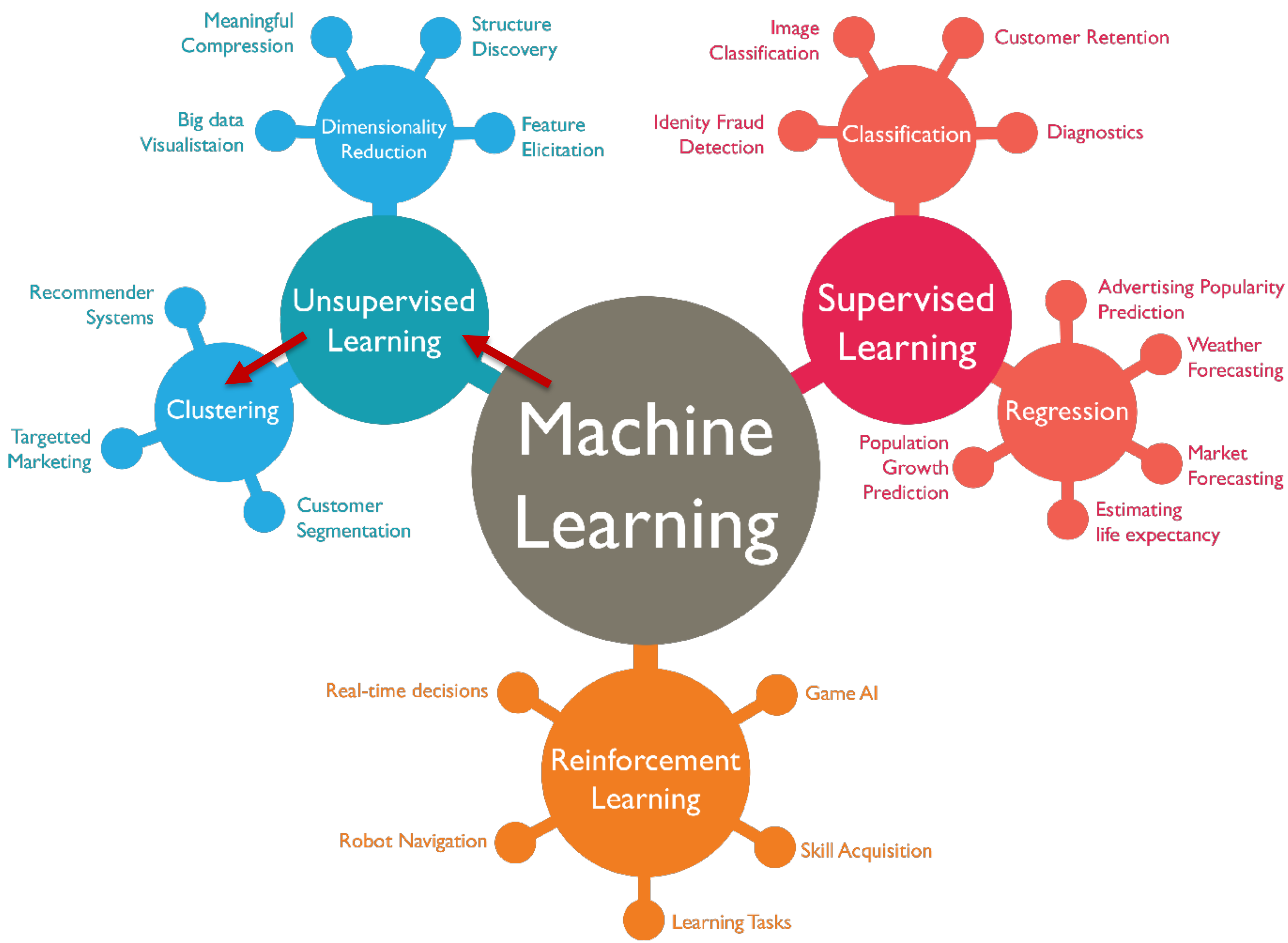



**Fig. 2.** Belgian part of the North Sea with 8 clusters or zones. The location of macrobenthic community samples is plotted for validation. Important patterns of the original abiotic variables are clearly visible on the map: e.g. high silt–clay % in cluster 1, alternation of sandbanks and flats–depressions in clusters 2, 3, 4, 5, 6 and 7; patches of gravel and shell fragments in cluster 8.

# Abiotic variables

**Table 1**
Abiotic variables as input for the PCA and cluster analysis (–, No unit).

| Abiotic variable | Unit | Reference |
|---|---|---|
| *Sedimentology* | | |
| | | All based on a sedimentological database ('sedisurf@') hosted at Ghent University, Renard Centre of Marine Geology. |
| Median grain-size of sand fraction (63–2000 μm) or $d_s50$ | μm | Verfaillie et al. (2006) |
| Silt–clay percentage (0–63 μm) | % | Van Lancker et al. (2007) |
| Sand percentage (63–2000 μm) | % | Van Lancker et al. (2007) |
| Gravel percentage (>2000 μm) | % | Van Lancker et al. (2007) |
| *Topography* | | |
| Digital terrain model (DTM) of bathymetry | m | Flemish Authorities, Agency for Maritime and Coastal Services, Flemish Hydrography<br>All other topographic variables are derived from the DTM |
| Slope = a first derivative of the DTM | ° | Evans (1980), Wilson et al. (2007) |
| Aspect = a first derivative of the DTM | | Hirzel et al. (2002), Wilson et al. (2007) |
| Indices of northness and eastness provide continuous measures (−1 to +1) describing orientation of the slopes. | | |
| Eastness = sin (aspect) | – | |
| Northness = cos (aspect) | – | |
| Rugosity = ratio of the surface area to the planar area across the neighborhood of the central pixel | – | Jenness (2002), Lundblad et al. (2006), Wilson et al. (2007) |
| Bathymetric Position Index (BPI) = measure of where a location, with a defined elevation, is relative to the overall landscape | | Lundblad et al. (2006), Wilson et al. (2007) |
| BPI (broad-scale) | – | |
| BPI (fine-scale) | – | |
| *Hydrodynamics* | | |
| Maximum bottom shear stress = frictional force exerted by the flow per unit area of the seabed | N/m$^2$ | Management Unit of the North Sea Mathematical Models and the Scheldt estuary |
| Maximum current velocity | m/s | |
| *Satellite derived variables* | | |
| Maximum Chlorophyll *a* (Chl *a*) concentration over a 2-year period (2003–2004) | mg/m$^3$ | MERIS satellite datasets; compiled by European Space Agency and Management Unit of the North Sea Mathematical Models and the Scheldt estuary |
| Maximum Total Suspended Matter (TSM): measure for turbidity over a 2-year period (2003–2004) | mg/l | |
| Distance to coast | m | Computed in GIS |




https://medium.com/marketing-and-entrepreneurship/10-companies-using-machine-learning-in-cool-ways-887c25f913c3