Unsupervised Learning: Wrap-up

Mirco Schönfeld
mirco.schoenfeld@uni-bayreuth.de

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!

Assessing Quality of Clustering

Meaningful clusters are highly subjective

Also, data is never exact or complete

Optimal results are maybe not the most useful

Feature Selection

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:

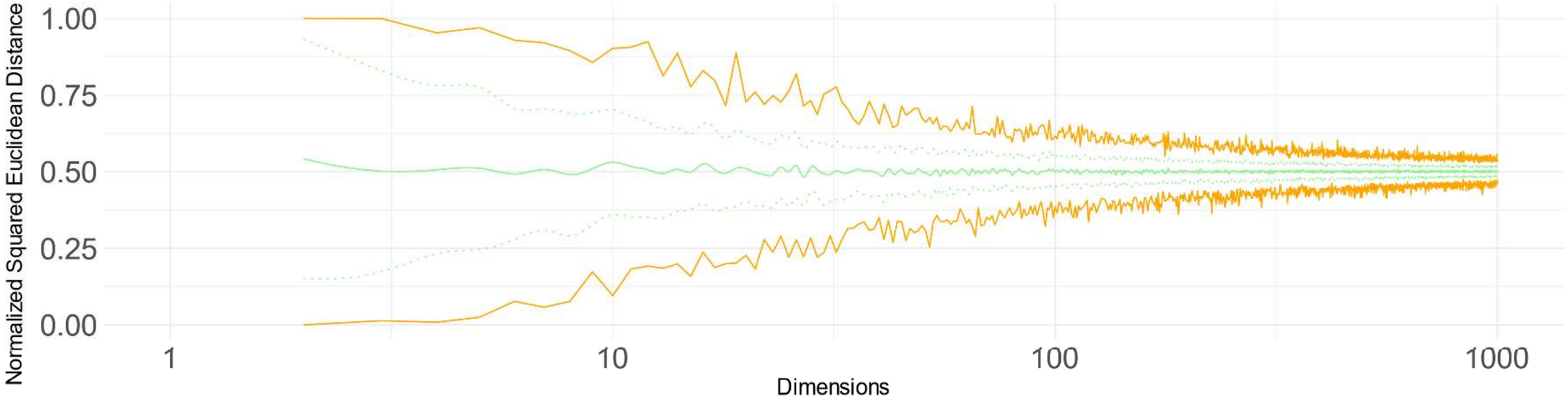
Including more features will improve classification *conceptually* but will render computation increasingly difficult.

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)

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

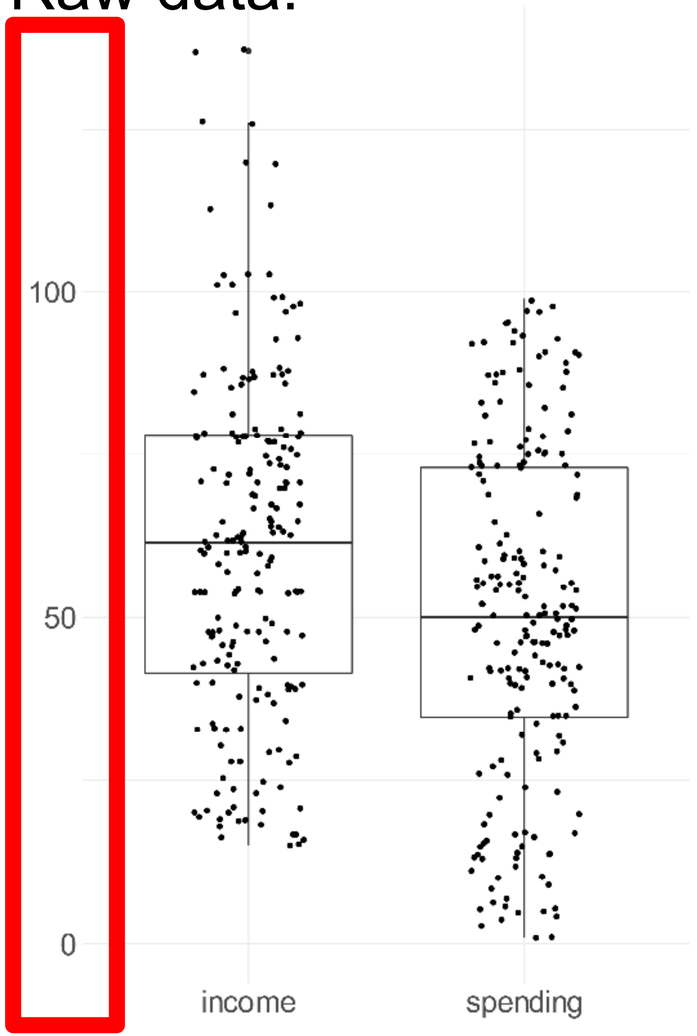
$\mu(X_i)$ is the arithmetic mean of variable X_i
$\sigma(X_i)$ is the standard deviation of variable X_i

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

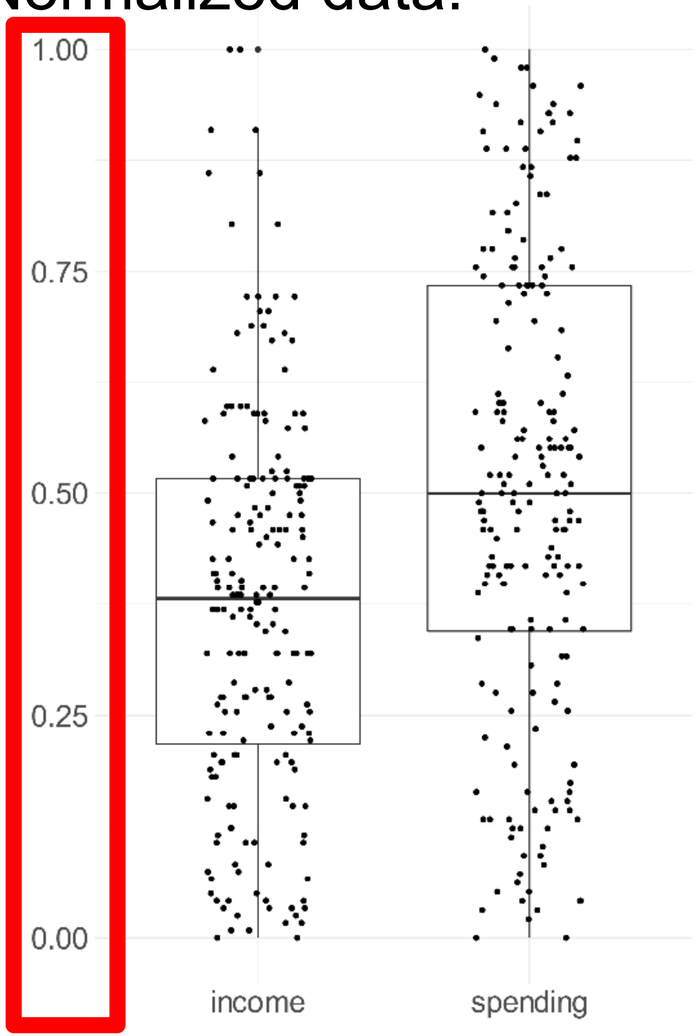
Normalized features matter „the same amount“

Normalization and Standardization

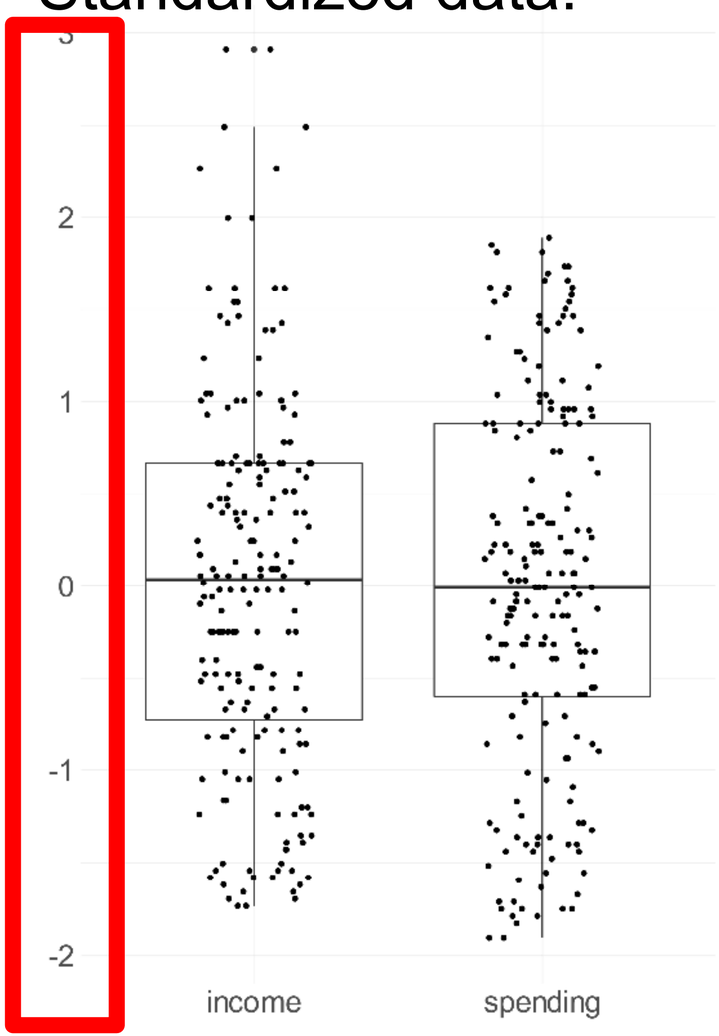
Raw data:



Normalized data:



Standardized data:



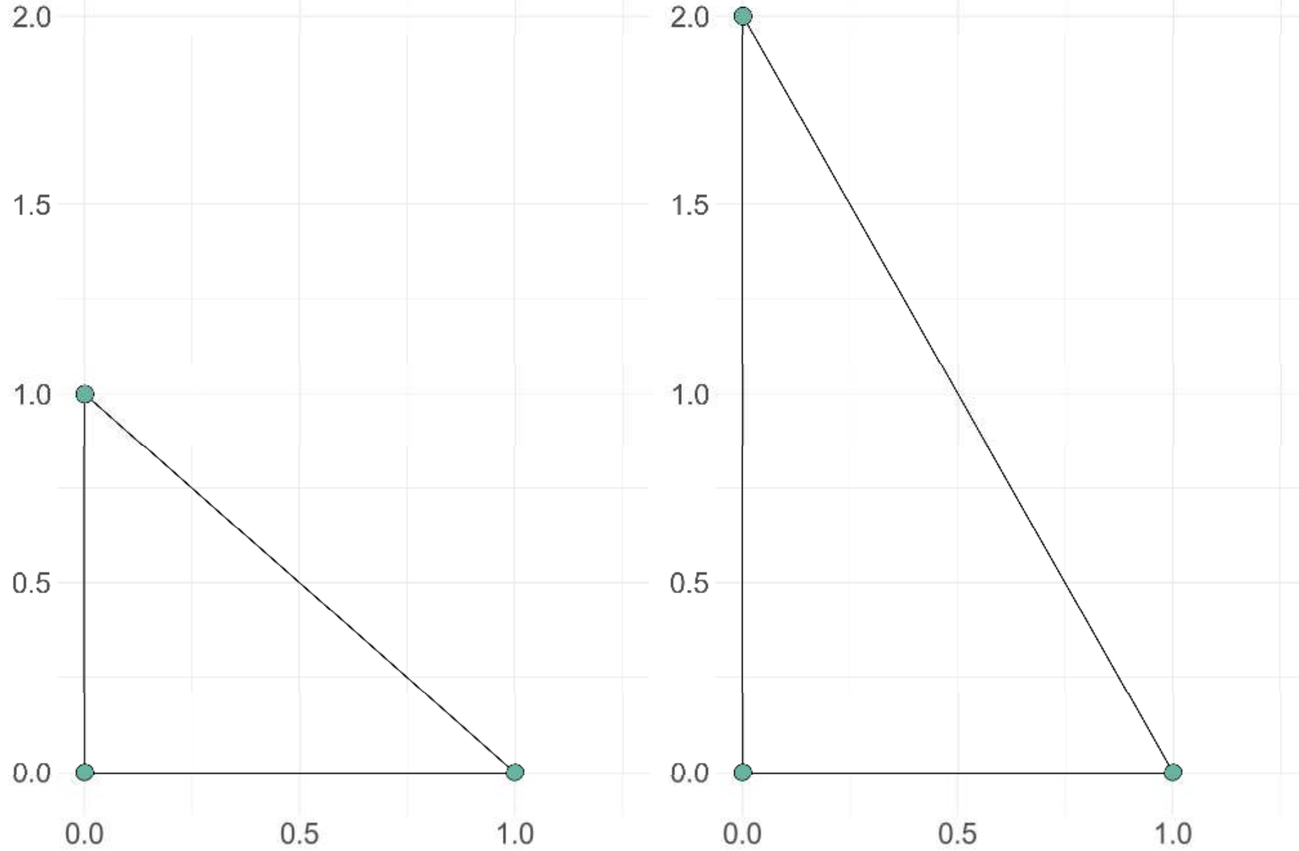
Scaling

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)

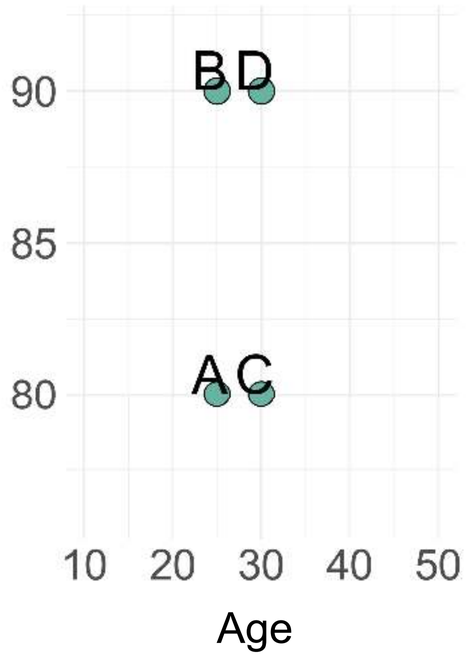
Suppose some data with 2 features

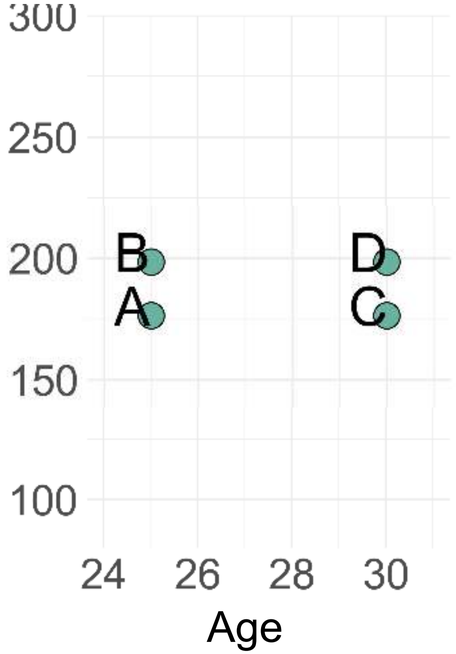
Multiplying the second feature by 2 influences the distance to other points



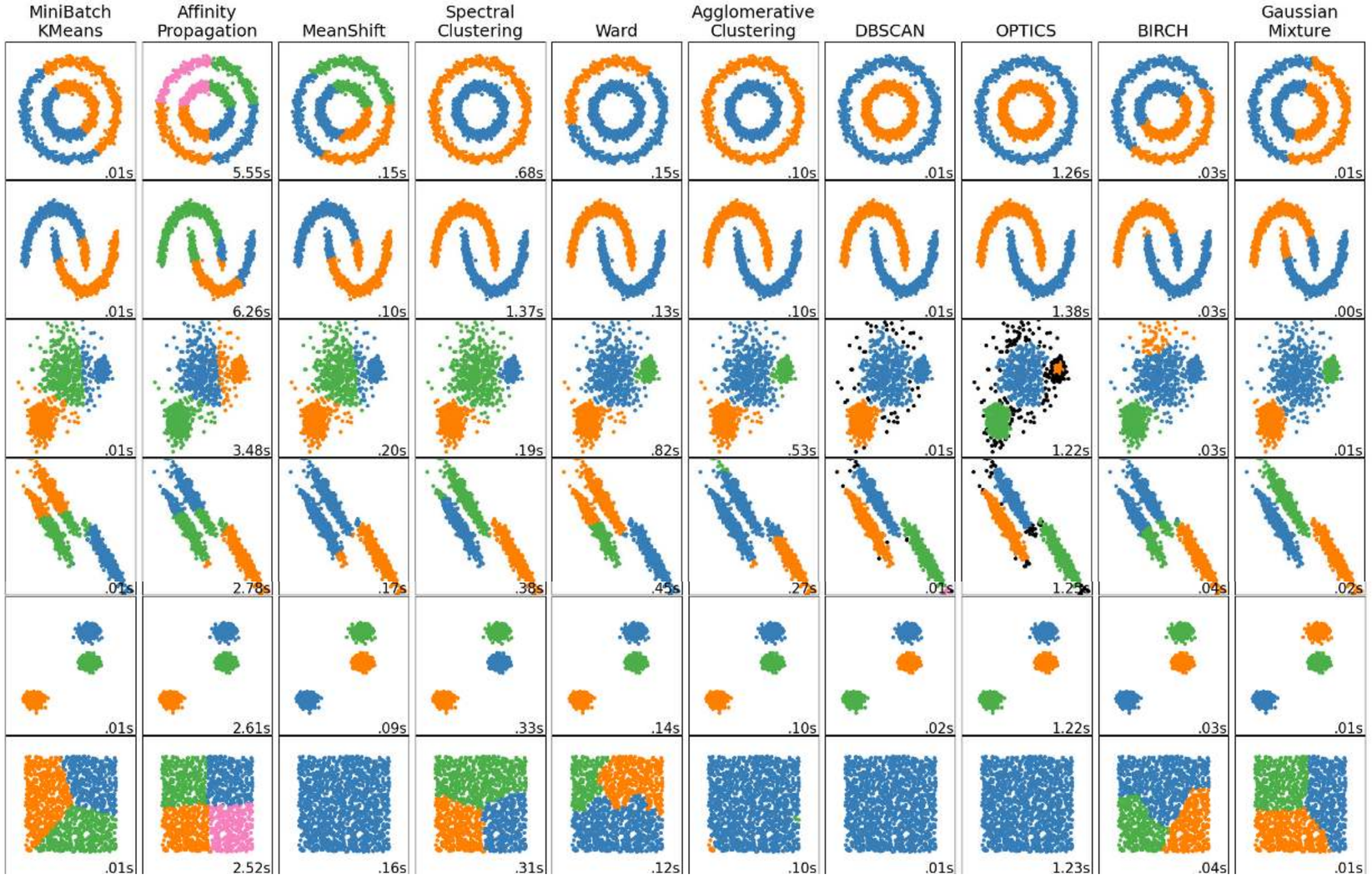
When to scale?

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



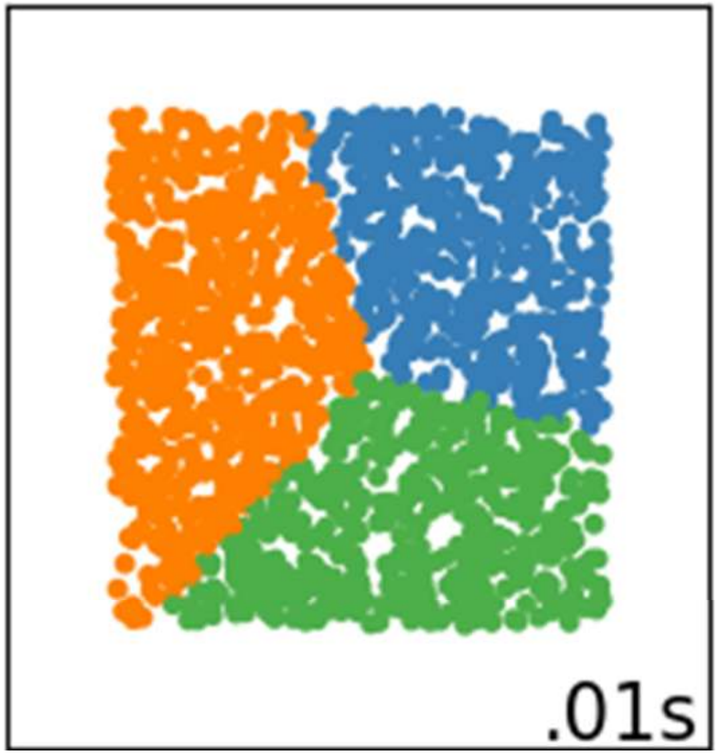


Data and Algorithms

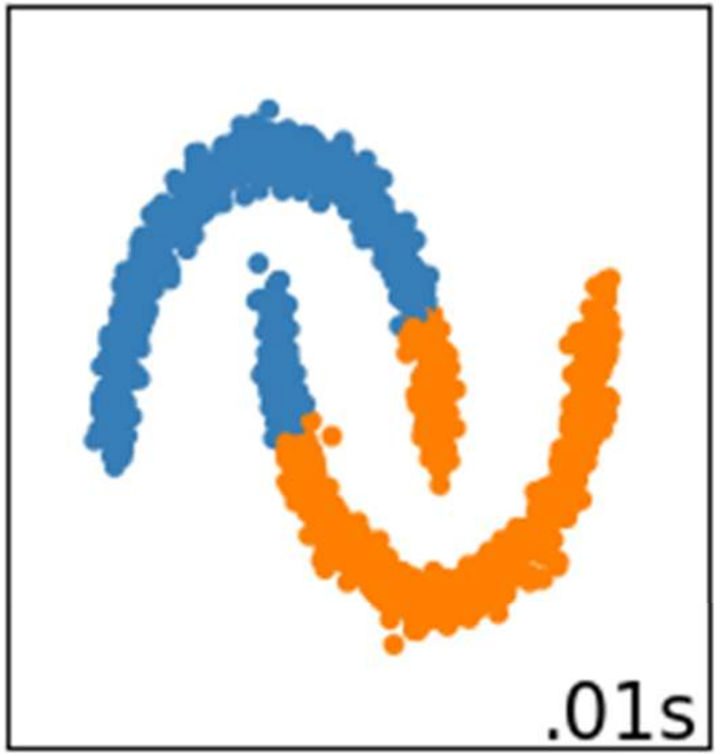


https://scikit-learn.org/stable/modules/clustering.html

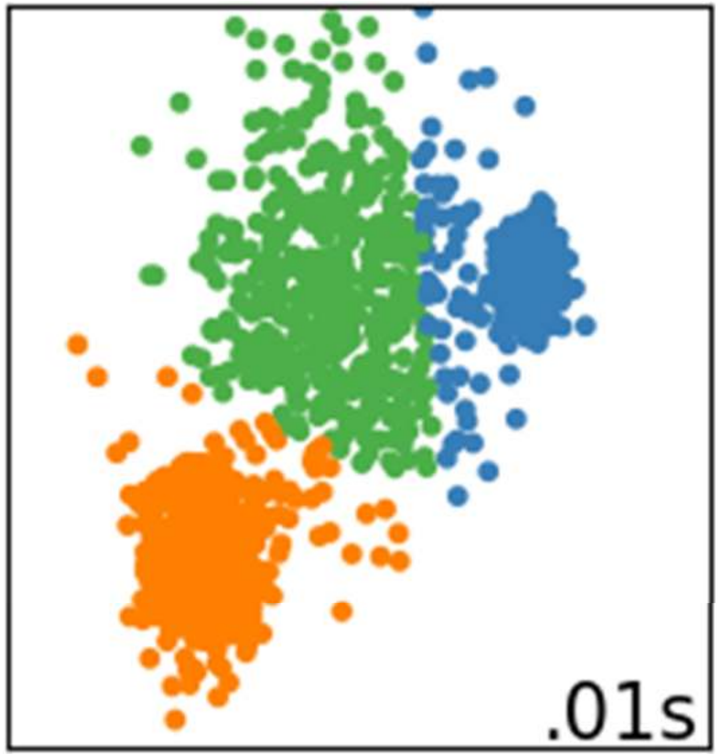
Noteworthy 1: k-means



Badly chosen *k*

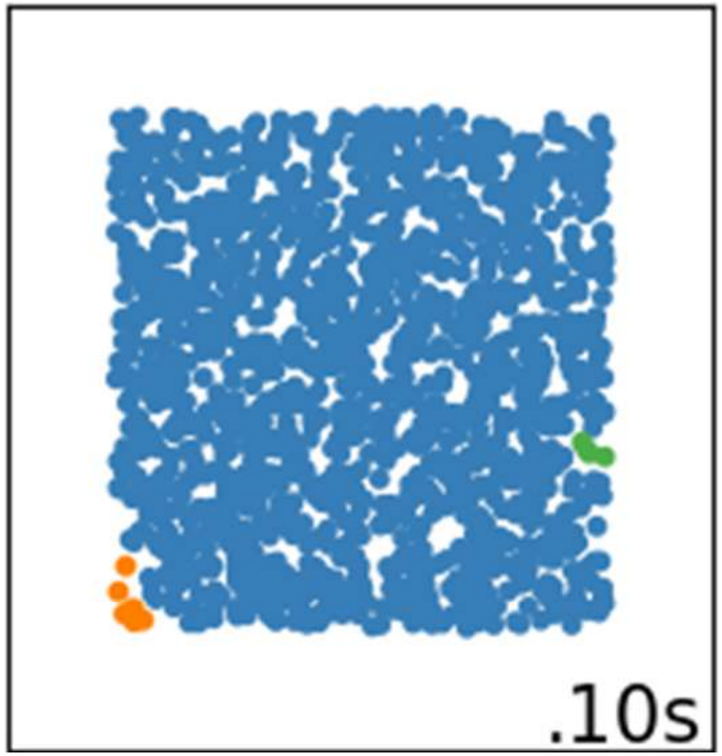


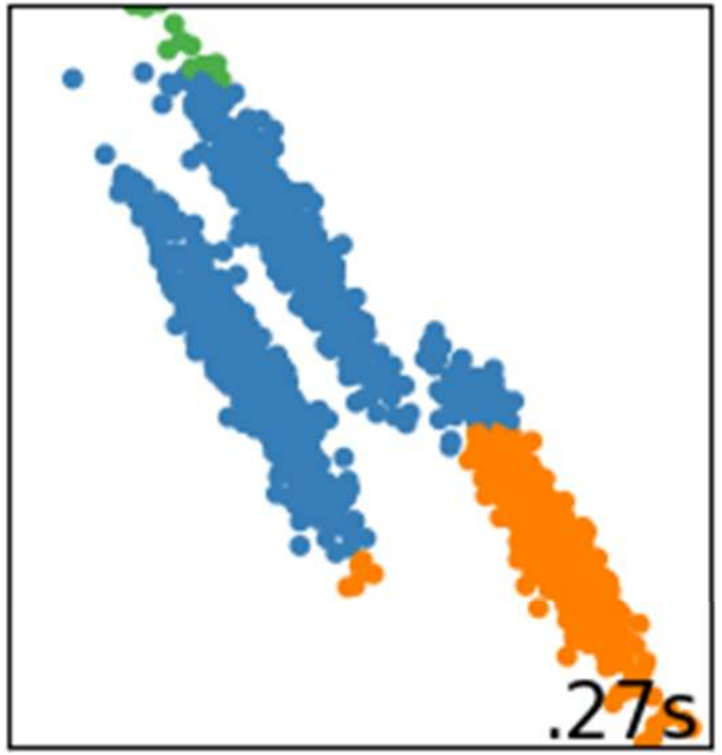
Non-spherical cluster shapes



Different cluster diameter
& different cluster densities

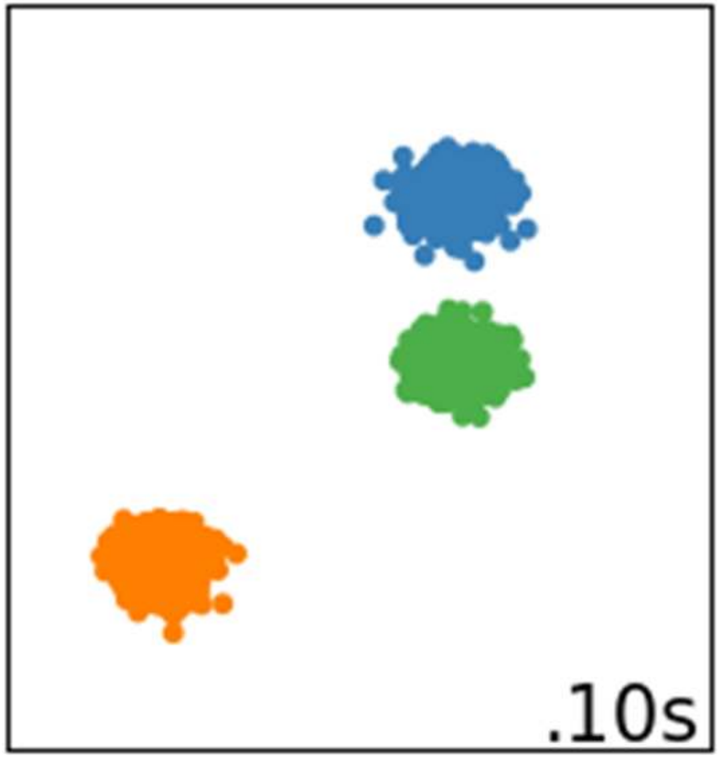
Noteworthy 2: Hierarchical clustering

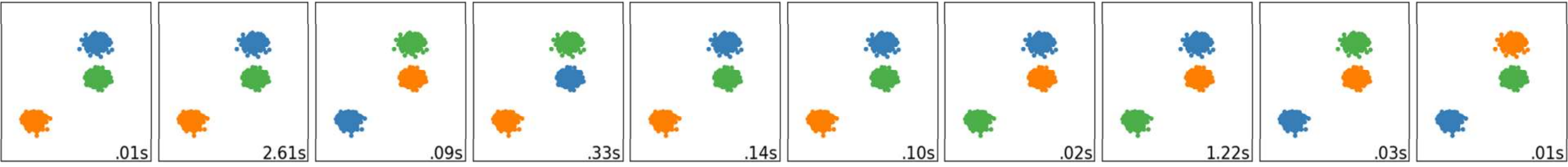




Not easy to specify both the distance metric *and* the linkage criteria

When does your data look like this?





https://xkcd.com/1838/

Thanks.

mirco.schoenfeld@uni-bayreuth.de