universität innsbruck
Department of Computer Science

STI INNSBRUCK

K-means Clustering

Ass.-Prof. Dr.rer.nat Anna Fensel

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

What you should be able to do after this lecture?

- Understand the concept of clustering, and particularly k-means clustering
- Explain the k-means clustering algorithm
- Provide diverse usage examples of the k-means clustering algorithm
- Understand different challenges in the use of the k-means clustering algorithm and its extensions/variations

Motivation: Why clustering?

What is clustering?

Motivation: Why clustering?

What is clustering?

» Finding “natural” groupings between objects

» We want to find similar objects (f.e. documents) to treat them in the same way

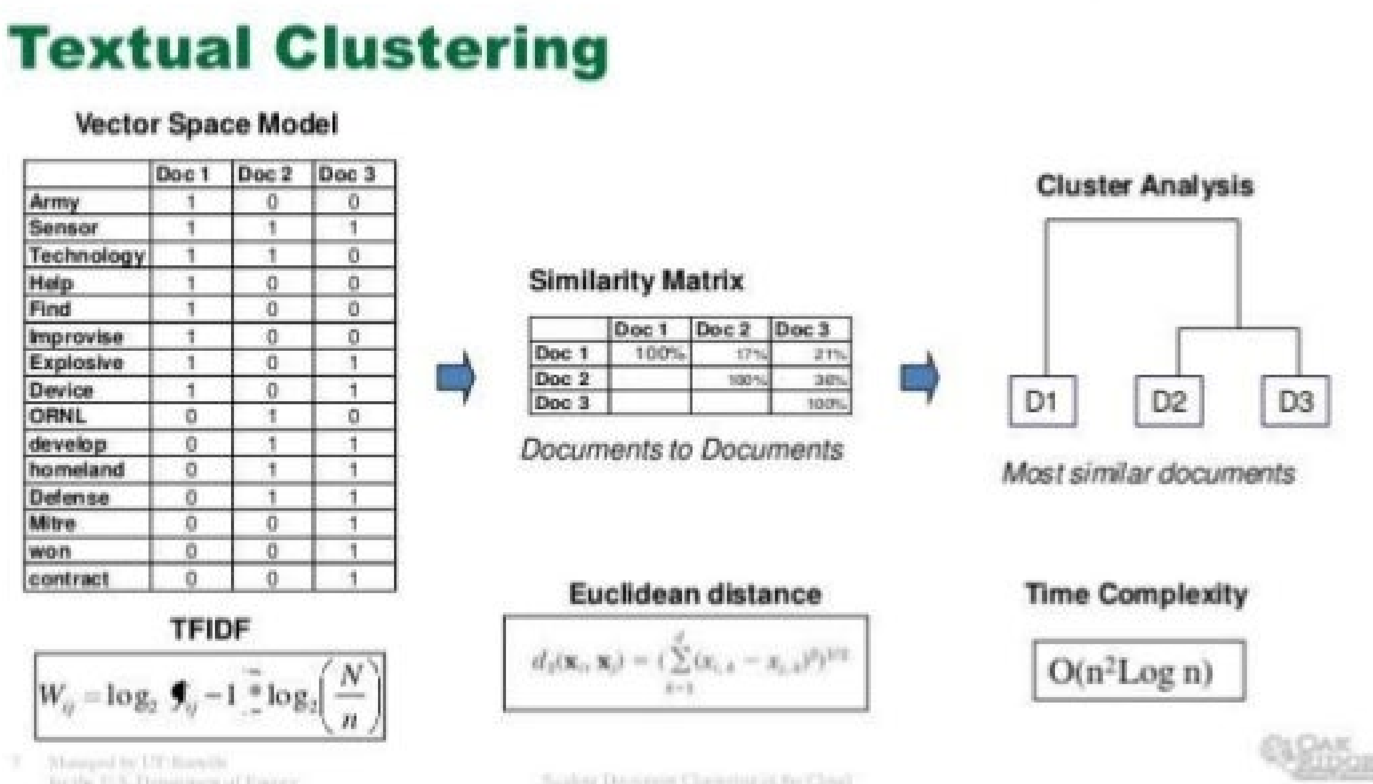
We aim at:

» High intra-cluster similarity

» Low inter-cluster similarity

Motivating example: Web document search

- A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information.
- Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.



Is clustering typically ...?

A. Supervised
B. Unsupervised

Is clustering typically ...?

A. Supervised

B. **Unsupervised**

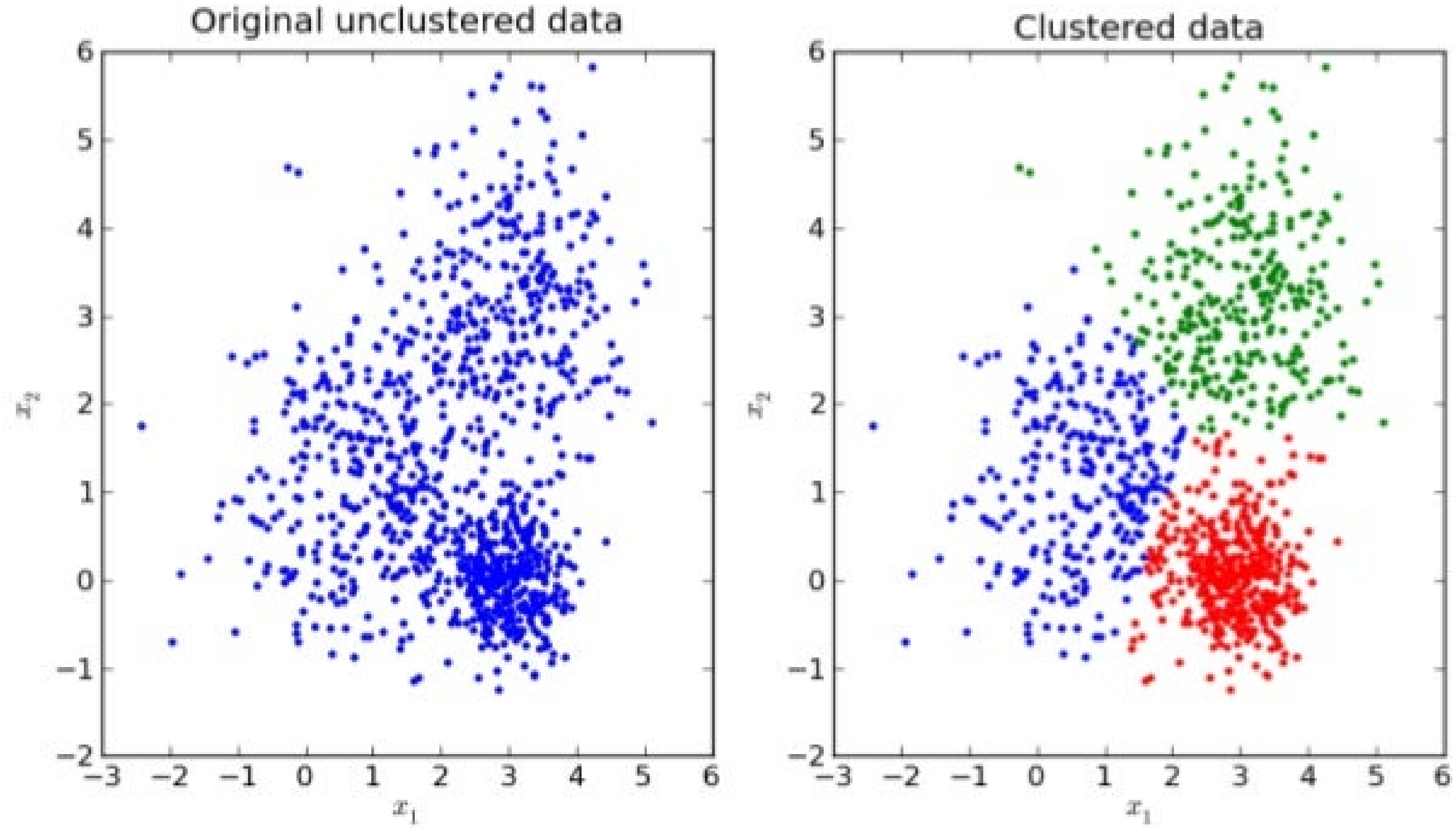
Supervised Classification	Unsupervised Clustering
• known number of classes	• unknown number of classes
• based on a training set	• no prior knowledge
• used to classify future observations	• used to understand (explore) data

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  - definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

What is K-means clustering?

- *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean



- Works for n-dimensional spaces as well

How do we measure similarity?
Give examples of similarity measures.

- » …
- » …

How do we measure similarity? Give examples of similarity measures.

- Similarity is subjective
- Its measure therefore depends on the data, the use case, the users
- In practice it is not always straightforward which metrics work well - then "trial and error" can be followed
- Examples of similarity measures: Euclidean, Manhattan, cosine distance

Mathematically, Euclidean distance between two n-dimensional vectors

$(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ is:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Manhattan distance between two n-dimensional vectors

$$d = |a1 - b1| + |a2 - b2| + \ldots + |an - bn|$$

The formula for the cosine distance between n-dimensional vectors

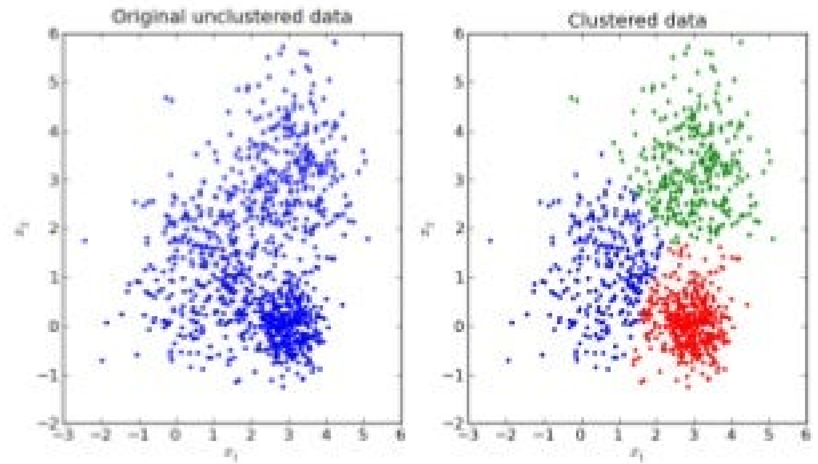
$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{(\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)}\sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)})}$$

How does K-means do the clustering?

K-means is a very important/basic flat clustering algorithm.

Its objective is to minimize the average squared Euclidean distance of values from their cluster centers where a cluster center is defined as the mean or *centroid* $\vec{\mu}$ of the values in a cluster ω:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$



Selection of centroids

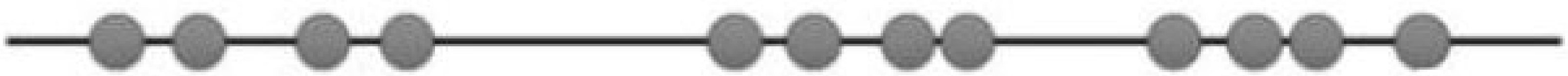
- The first step of K-means is to select as initial cluster centers K randomly selected documents, the *seeds*.
- The algorithm then moves the cluster centers around in space in order to minimize RSS (the function that defines how central the centroids are).

K-means illustration step-by-step (1 dimensional)

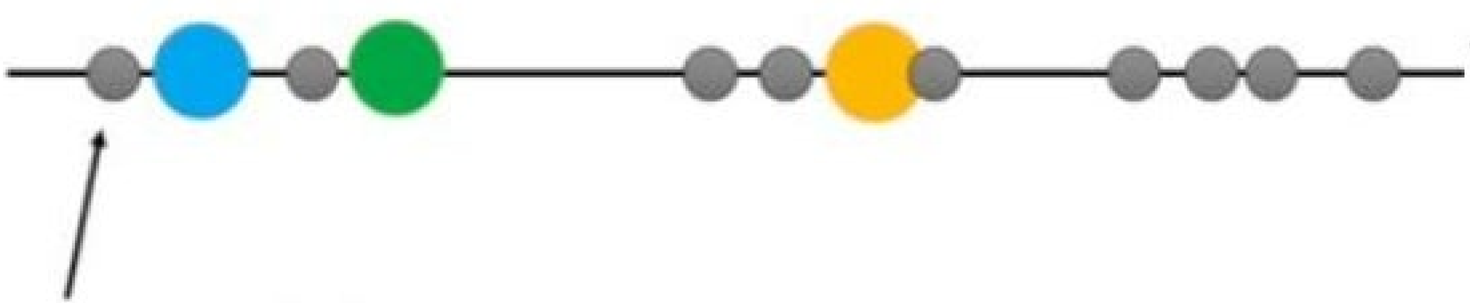
From: StatQuest: K-means clustering: https://www.youtube.com/watch?v=4b5d3muPQmA

Step 1: Select the number of clusters you want to identify in your data. This is the "K" in "K-means clustering".

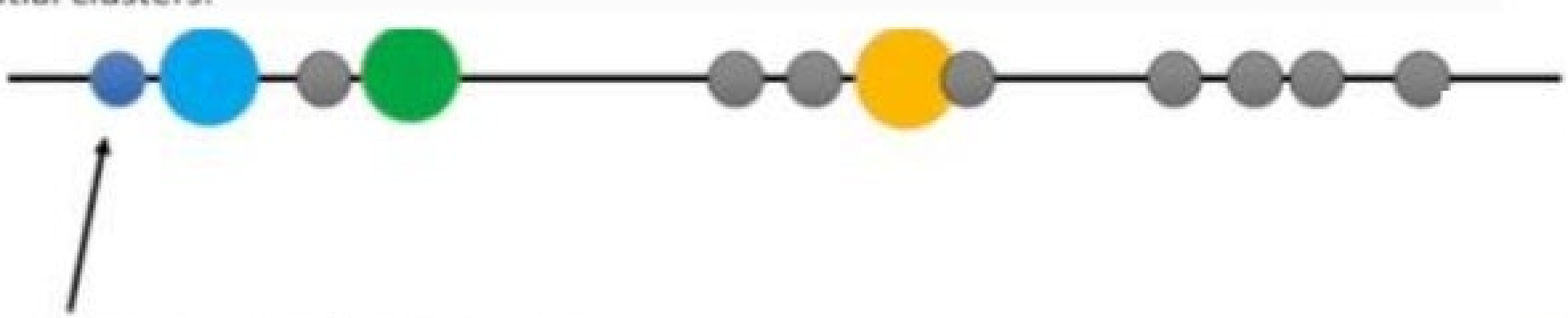
In this case, we'll select K=3. That is to say, we want to identify 3 clusters.



Step 2: Randomly select 3 distinct data points.

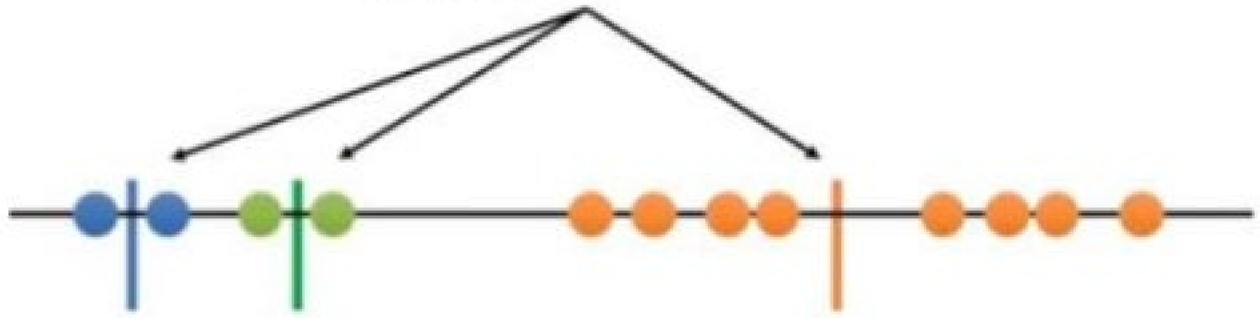


Step 3: Measure the distance between the 1st point and the three initial clusters.

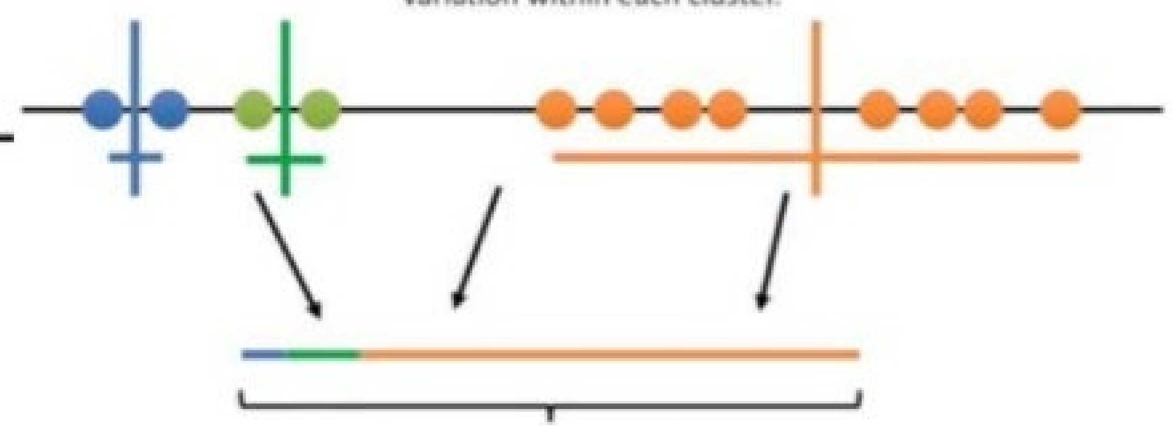


Step 4: Assign the 1st point to the nearest cluster. In this case, the nearest cluster is the **blue** cluster.

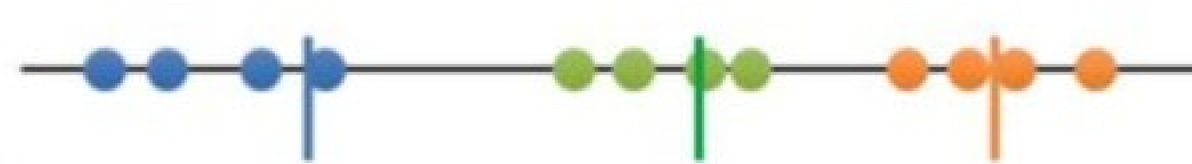
Step 5: calculate the mean of each cluster.



We can assess the quality of the clustering by adding up the variation within each cluster.



...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.



Calculating the “centrality” of the centroids
[Manning & Schütze, 2008]

A measure of how well the centroids represent the members of their clusters is the *residual sum of squares* or *RSS*,

$$\mathrm{RSS}_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2$$

the squared distance of each vector from its centroid summed over all vectors:

$$\mathrm{RSS} = \sum_{k=1}^{K} \mathrm{RSS}_k$$

Our goal is to minimize RSS (i.e. the average squared distance) till it is possible.

K-means algorithm summary

[Manning & Schütze, 2008]

$\text{K-MEANS}(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$

$(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow \text{SELECTRANDOMSEEDS}(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$

for $k \leftarrow 1$ **to** K

do $\vec{\mu}_k \leftarrow \vec{s}_k$

while stopping criterion has not been met

do for $k \leftarrow 1$ **to** K

  do $\omega_k \leftarrow \{\}$

  for $n \leftarrow 1$ **to** N

  do $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$

    $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*

  for $k \leftarrow 1$ **to** K

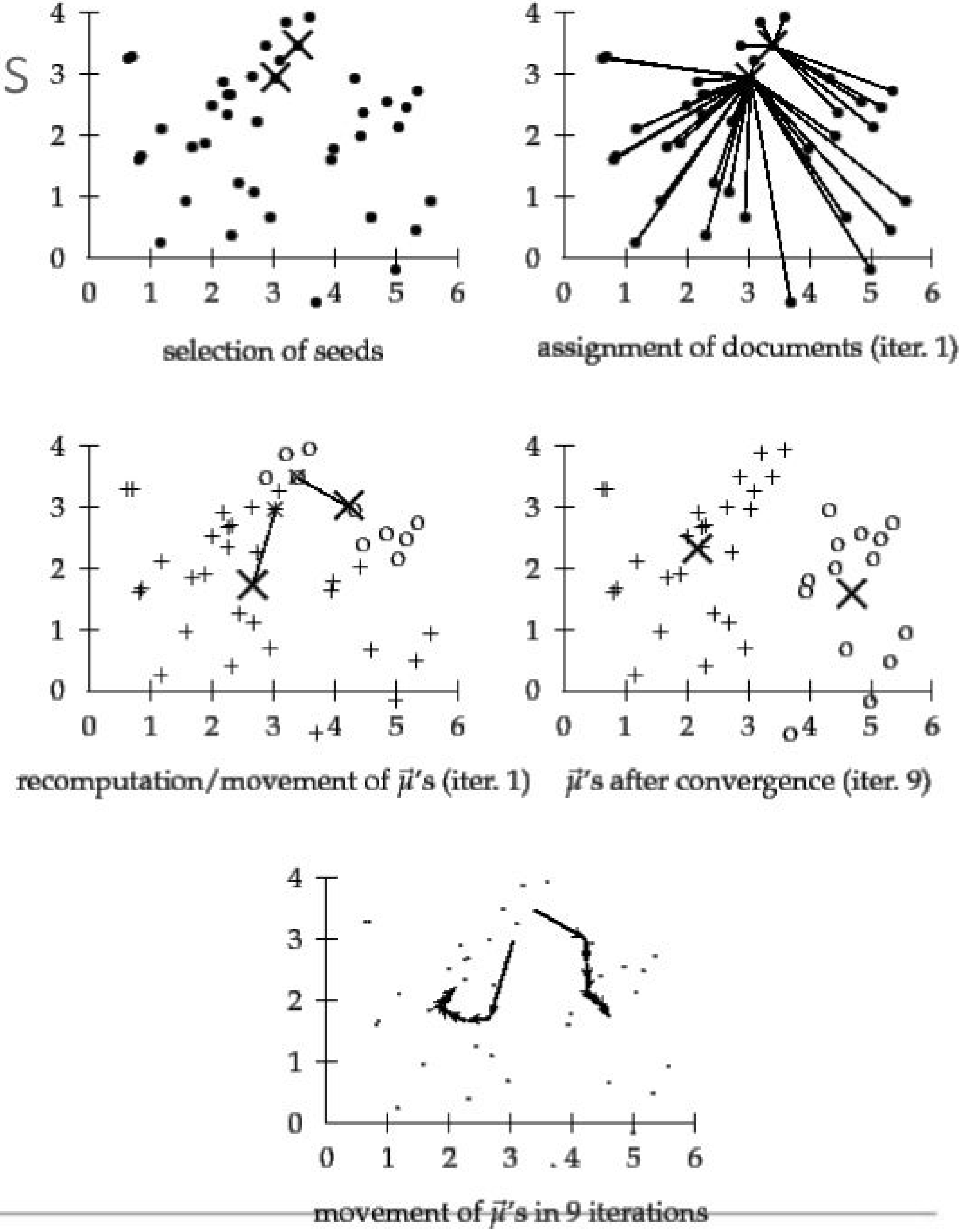
  do $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*

return $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

► **Figure 16.2** The K-means algorithm. For most IR applications, the vectors $\vec{x}_n \in \mathbb{R}^M$ should be length-normalized. Alternative methods of seed selection and initialization are discussed on page 16.4 .

K-means – iteration process
[Manning & Schütze, 2008]

- reassigning documents to the cluster with the closest centroid,
- recomputing each centroid based on the current members of its cluster.



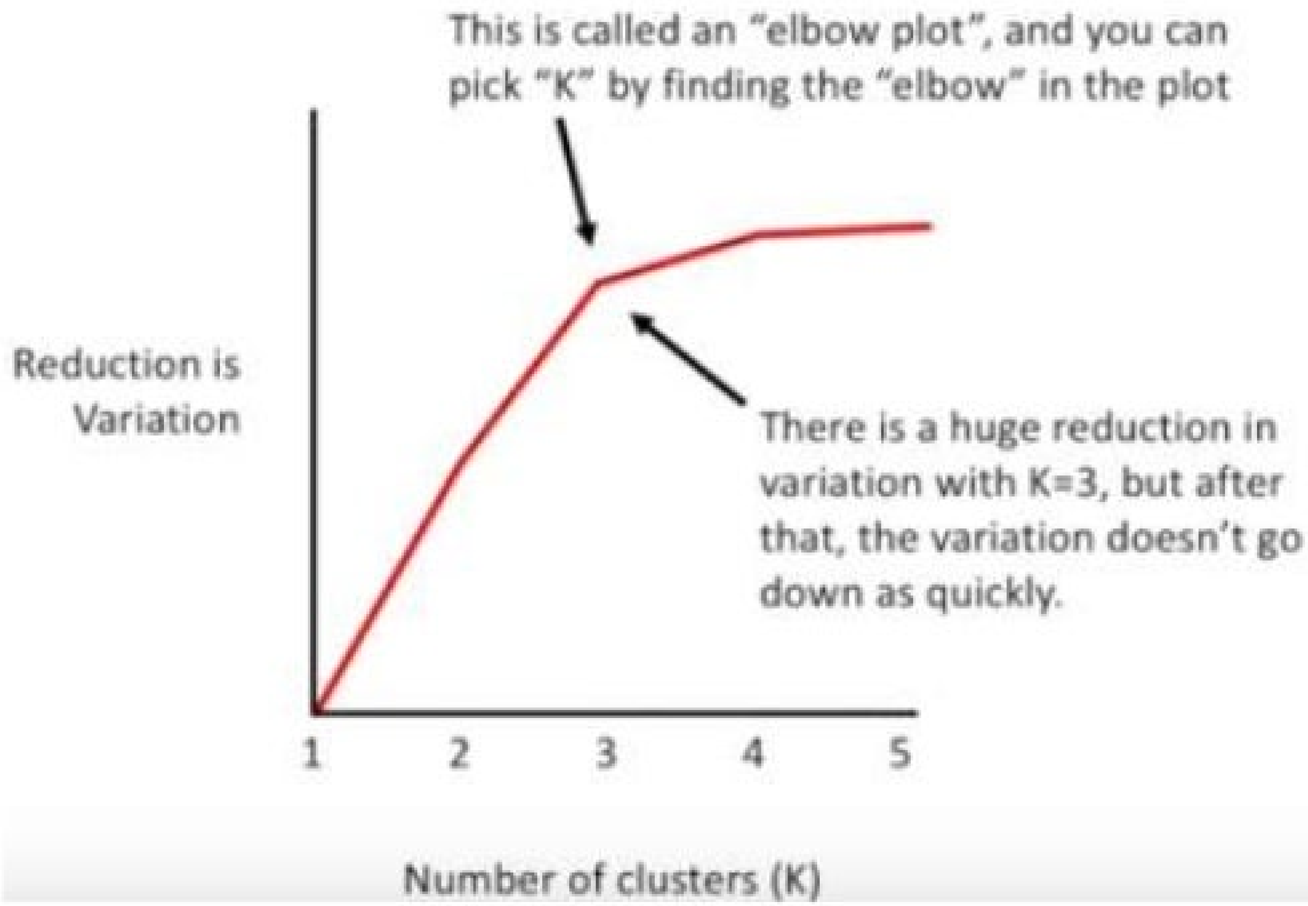
▶ **Figure 16.3** A K-means example for $K = 2$ in $\mathbb{R}^2$. The position of the two centroids ($\vec{\mu}$'s shown as X's in the top four panels) converges after nine iterations.

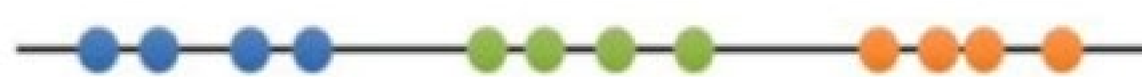
How to set where to terminate the algorithm? [Manning & Schütze, 2008]

- A fixed number of iterations I has been completed.
  - This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
- Assignment of documents to clusters (the partitioning function γ) does not change between iterations.
  - Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
- Centroids $\vec{\mu}_k$ do not change between iterations.
  - This is equivalent to γ not changing.
- Terminate when RSS falls below a threshold.
  - This criterion ensures that the clustering is of a desired quality after termination. In practice, we need to combine it with a bound on the number of iterations to guarantee termination.
- Terminate when the decrease in RSS falls below a threshold θ .
  - For small θ , this indicates that we are close to convergence. Again, we need to combine it with a bound on the number of iterations to prevent very long runtimes.

A variation: How do you know how many clusters you should make?

- It is possible to try different cluster numbers.
- And check where the variation stabilizes to decide on the number of clusters.





From: StatQuest: K-means clustering:
https://www.youtube.com/watch?v=4b5d3muPQmA

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

K-means computational aspects

- K-means converges, but there is unfortunately no guarantee that a global minimum in the objective function will be reached.
  - This is a particular problem if a document set contains many outliers , documents that are far from any other documents and therefore do not fit well into any cluster. We may end up with a singleton cluster (a cluster with only one document) even though there is probably a clustering with lower RSS.

What is the time complexity K-means?
[Manning & Schütze, 2008]

- Most of the time is spent on computing vector distances. One such operation costs $\Theta(M)$.
- The reassignment step computes *KN* distances, so its overall complexity is $\Theta(KNM)$.
- In the recomputation step, each vector gets added to a centroid once, so the complexity of this step is $\Theta(NM)$.
- For a fixed number of iterations *I*, the overall complexity is therefore $\Theta(IKNM)$.

Extensions

There are numerous extensions to the K-means clustering f.e.

- K-means clustering can be generalized e.g. into a Gaussian mixture model.
- Efficiency problem can be addressed e.g. by K-medoids , a variant of K-means that computes medoids instead of centroids as cluster centers.

  The medoid of a cluster as the value that is closest to the centroid. Distance computations are faster in this case.

Tool support

A number of tools implementing k-means clustering are available:

- Open source e.g. Apache Spark Torch, R, and
- Proprietary e.g. MATLAB, Mathematica, SAP HANA

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

Application example: Image compression

- Aim: compress an image in size
- Question: with how many dimensional space we are working here?

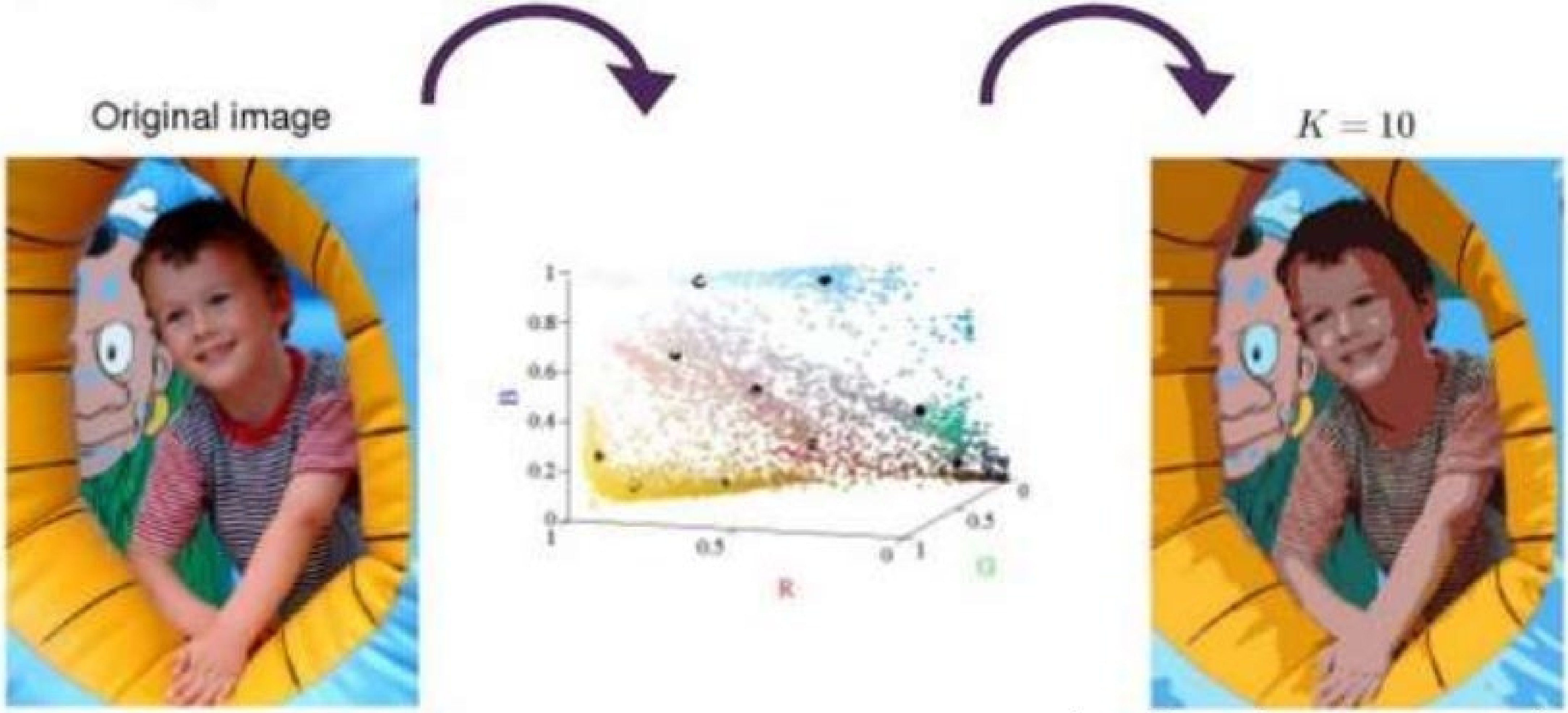
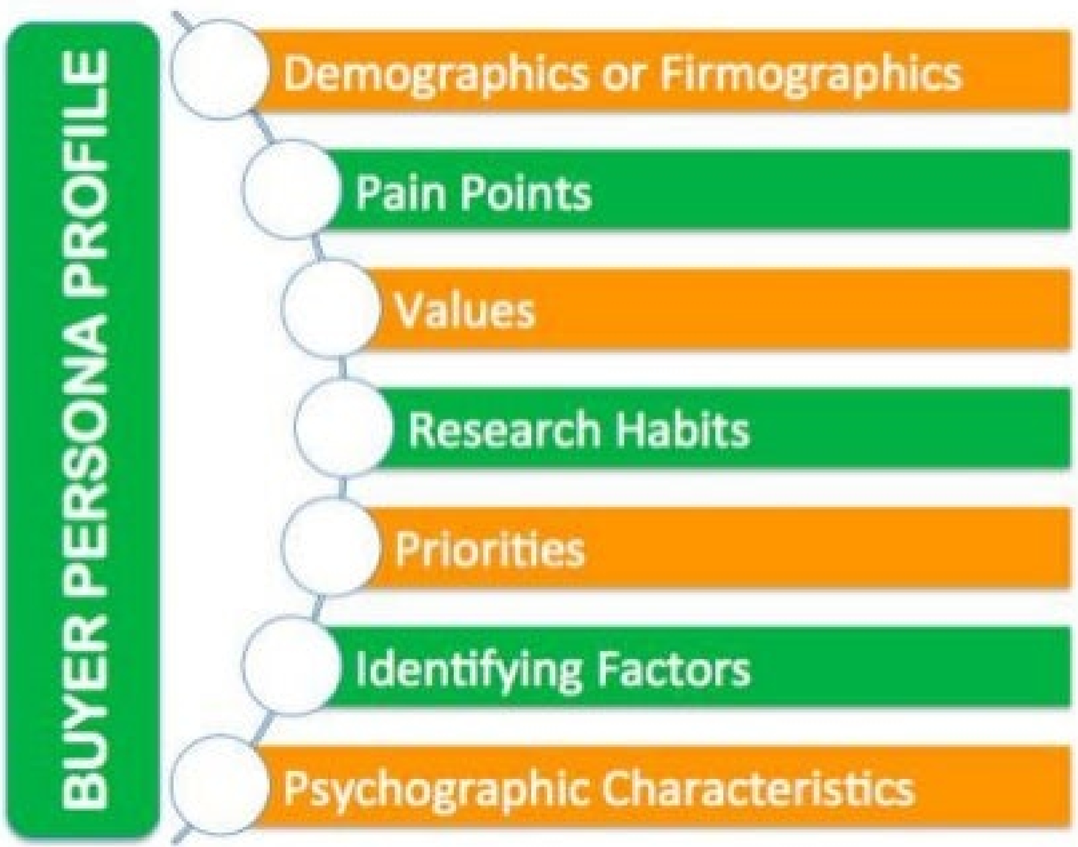


Image credit: Shinichi Tamura @ Slideshare

Application example: Retail – recommendation and yield management

- User profiles/personas: similar purchase behavior…



- Product profiles: similar selling patterns
- Deciding when to discount product groups

Question: with how many dimensional space we are working here?

Outline

- Introduction, learning goals
- Motivation and example
- Clustering
- K-means clustering algorithm
  definition, functions, iteration process, pseudocode
- Computational complexity
- Extensions
- Tools
- Application examples
- Conclusions
- References

Summary

- Provide 5 most important points you have learnt from today's lecture.
- ...
- ...
- ...
- ...
- ...
- (and let's compare the points)

References

- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.
- Manning C, R. P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
- K-Means Clustering at Wikipedia: https://en.wikipedia.org/wiki/K-means_clustering
- StatQuest: K-means clustering: https://www.youtube.com/watch?v=4b5d3muPQmA
- Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 16.1. Gaussian Mixture Models and k-Means Clustering". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8.

Thank you for attention.
Questions?

www.uibk.ac.at/informatik
www.sti-innsbruck.at