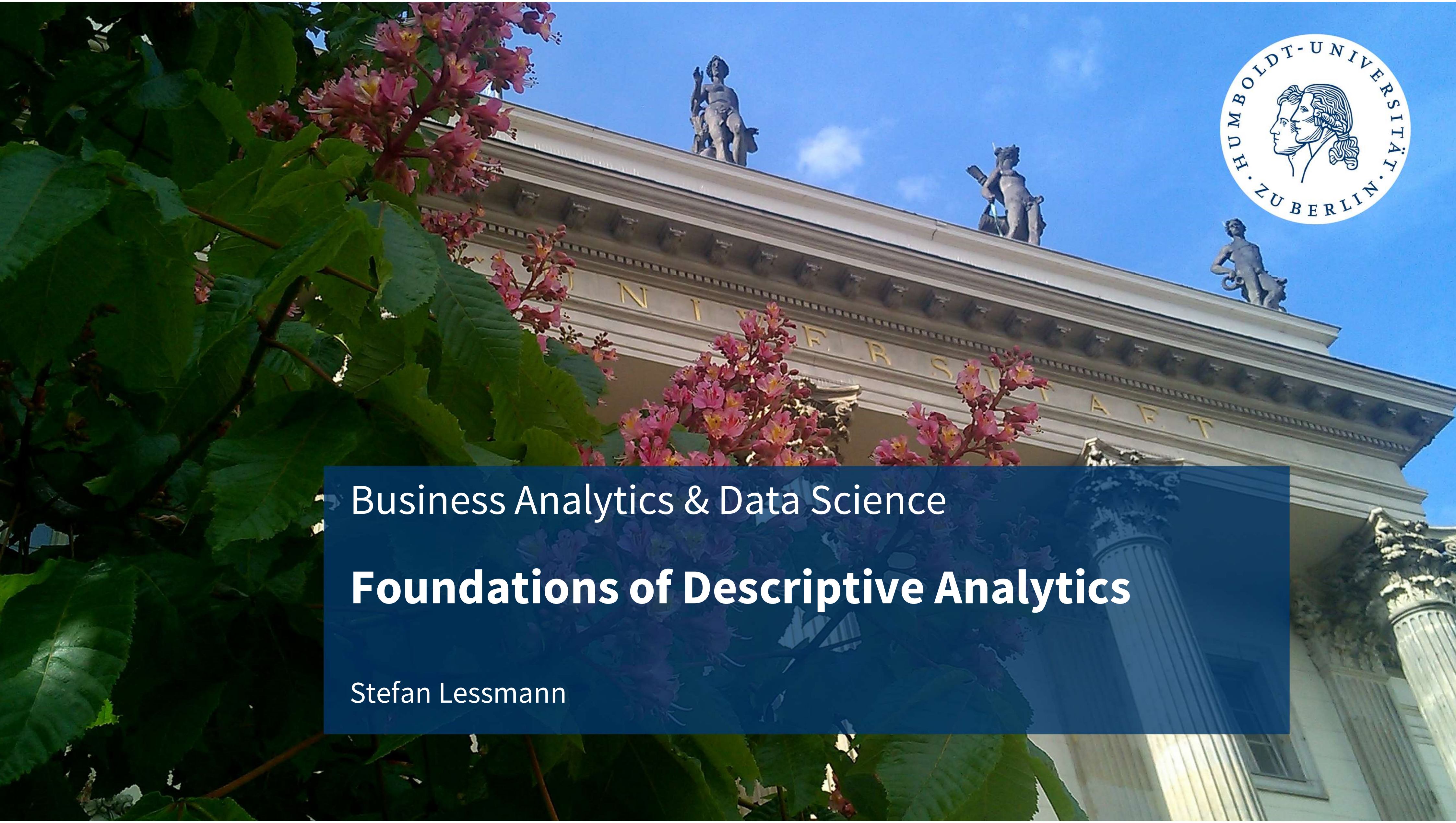


Business Analytics & Data Science

Foundations of Descriptive Analytics

Stefan Lessmann

Agenda

Business Analytics Revisited
Definitions, analytics process model, data structure

Descriptive Analytics in a Nutshell
Scope and flavors, business applications

Cluster Analysis Case Study

Cluster Analysis Methods
Approaches toward clustering, distance metrics, k-means algorithm

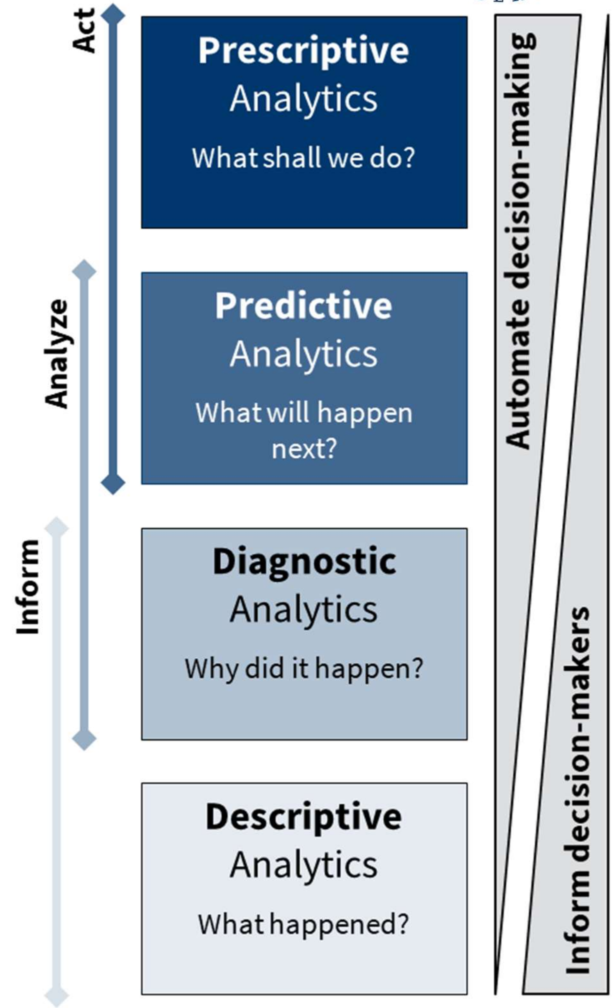
Summary

Business Analytics Revisited

Definitions, analytics process model, data structure

Recap: The Scope of Business Analytics

- **Descriptive analytics**
  - Use data to understand the past
  - Aggregation, clustering, unsupervised machine learning
- **Diagnostic analytics**
  - Depict data to maximize insight and minimize cognitive effort
  - Nontrivial for complex data
- **Predictive analytics**
  - Use historic data to detect generalizable patterns for anticipating what will happen in the future
  - Supervised machine learning, deep learning, forecasting
- **Prescriptive analytics**
  - Use forecasts and other information to recommend specific actions
  - Optimization, treatment effects, reinforcement learning



Recap: Data Science Models and Algorithms

Modeling approaches

Predictive

Regression

- Linear Regression
- Neural network
- CART
- …

Classification

- Logistic regression
- Nearest neighbor
- Boosting
- …

Descriptive

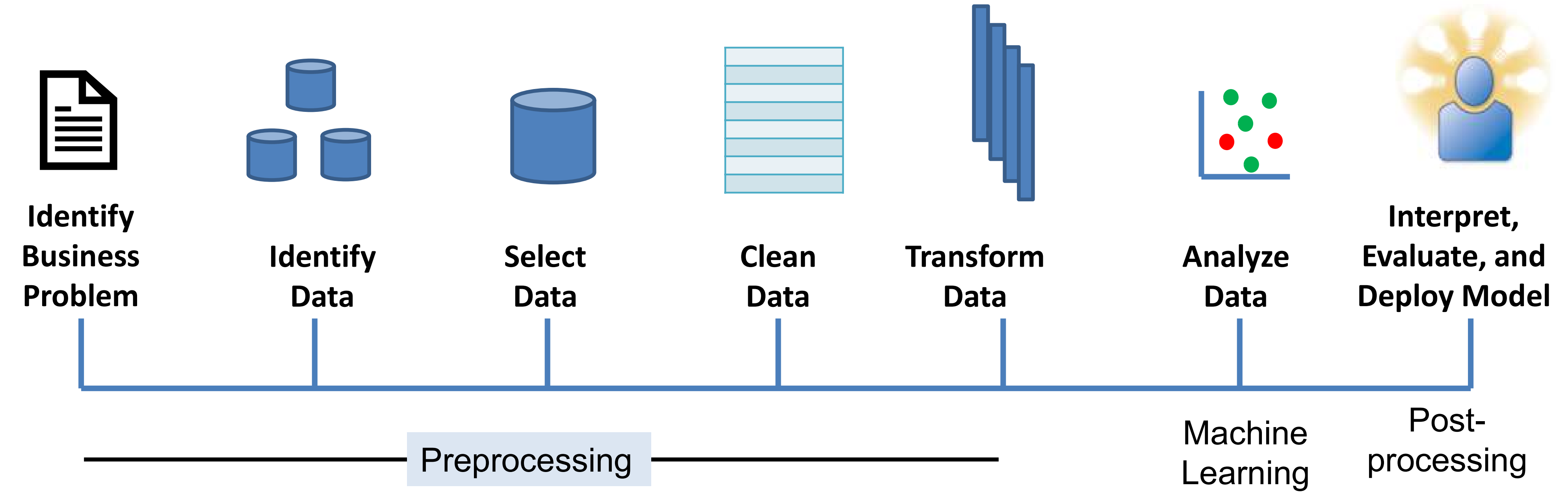
Segmentation

- Hierarchical
- K-means
- DBSCAN
- …

Rule/sequence mining

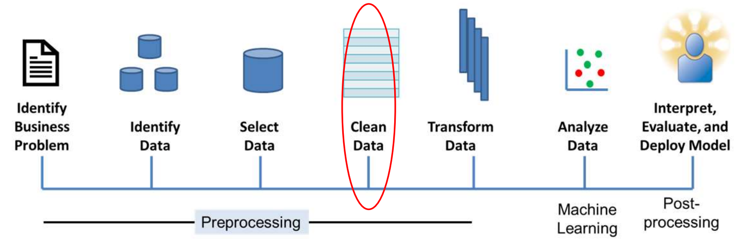
- Apriori-algorithm
- FP-growth
- PrefixSpan
- …

Recap: The Analytics Process Model



Data Structure for Business Analytics

Data is typically brought into a tabular format



Cases / observations / examples

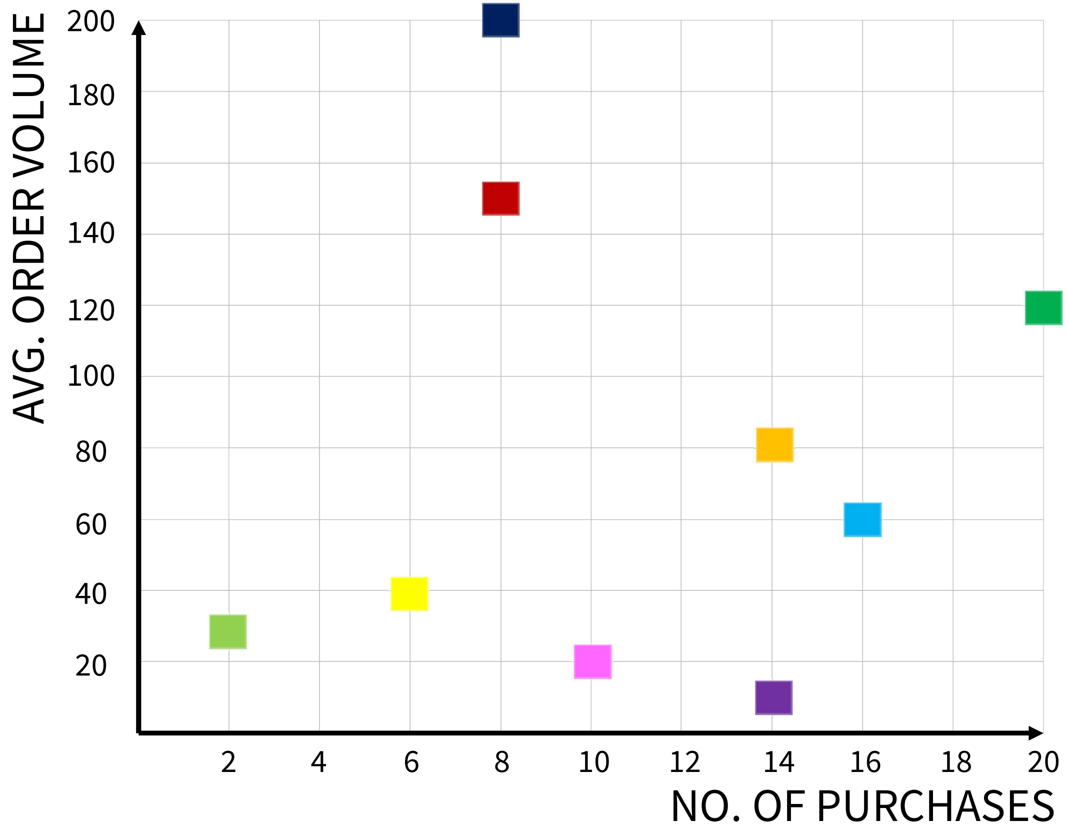
Age group	Gender	No. of orders	No. of returns	Days since last order	Total purchases	…
<18	M	3	1	7	€150	…
18-29	M	1	0	13	€75	…
<18	F	5	2	5	€33	…
30-50	M	2	0	2	€24	…
>50	F	1	0	25	€120	…
19-29	F	3	1	17	€41	…
>50	F	9	1	9	€284	…
18-29	M	2	2	14	€10	…
<18	F	1	0	11	€18	…

Variables/ characteristics / attributes/ features / predictors/ covariates

Graphical Interpretation of Tabular Data

An observation equates to a data point in a multi-dimensional feature space

	NO PURCHASES	AVG. ORDER VOLUME	…
	8	€150	…
	14	€80	…
	6	€40	…
	2	€30	…
	20	€120	…
	16	€60	…
	8	€200	…
	14	€10	…
	10	€20	…



Descriptive Analytics in a Nutshell

Scope and flavors, business applications

Descriptive Analytics
Employs algorithms from the field of unsupervised learning

- **Data set with several features and no target variable**
- **Find structure / patterns in the data**
- **Multiple forms**
  - Clustering
  - Dimensionality reduction
  - Association rule mining
  - Sequence rule mining
- **Widely applicable as plain (i.e. unlabeled) data is easily available**
- **Hard to formally evaluate model outputs as ground truth data is not available**
- **Often hard to ensure that detected patters are relevant to the business**
- **Inform decision-making but do not recommend concrete actions**

Association and Sequence Rule Mining

Findings co-occurrences of items in transactional databases

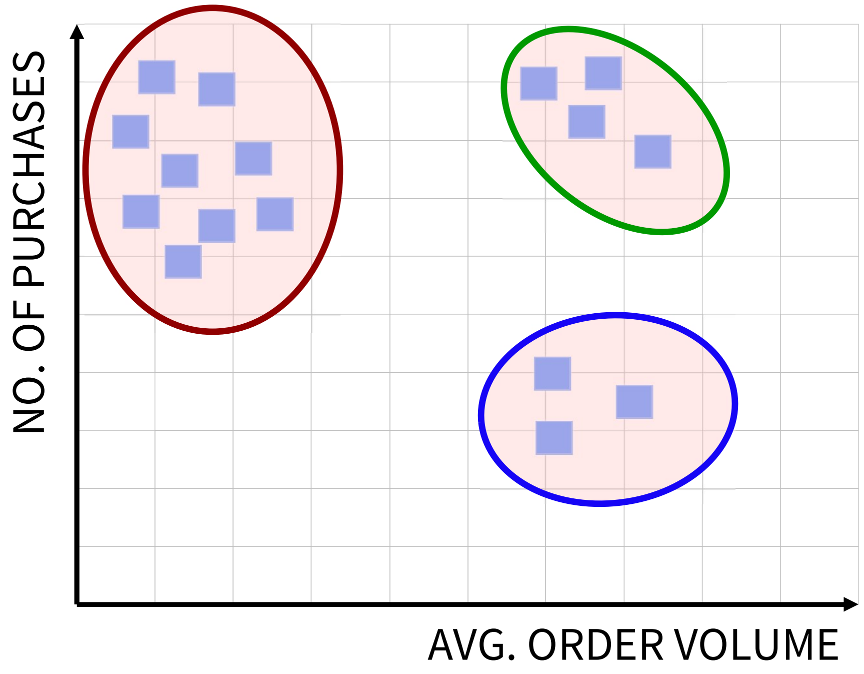
- **Search for rules of the form *If A then B* ($A \Rightarrow B$)**
- **Data structure**
  - Tupels of items
  - With or w/o ordering
- **Computational challenges**
- **Business applications**
  - Market basket analysis
  - Clickstream analysis
  - Fraud analytics
  - Financial market modeling

Trans-action	Items (e.g., products in shopping basket)
1.	Beer, milk, diapers, baby food
2.	Coke, beer, diapers
3.	Cigarettes, diapers, baby food
4.	Chocolates, diapers, milk, apples
5.	Tomatoes, water, apples, beer
6.	Spaghetti, diapers, baby food, beer
7.	Water, beer, baby food
8.	Diapers, baby food, spaghetti
9.	Baby food, beer, diapers, milk
10.	Apples, wine, baby food

Session ID	Web page	Sequence
1	A	1
1	B	2
1	C	3
2	B	1
2	C	2
3	A	1
3	C	2
3	D	3
4	A	1
4	B	2
4	D	3
5	D	1
5	C	2
5	A	3

Cluster Analysis
Findings sub-groups of observations that display similarity

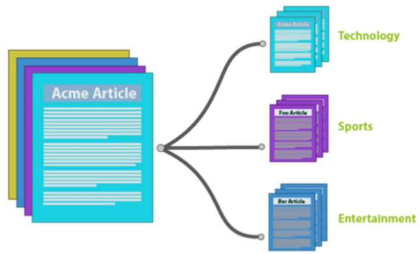
- **Goal is to describe the inherent structure of a data set**
  - Homogeneous subgroups
  - Summarization
- **Search for similarity**
  - Cases within a cluster as homogeneous as possible
  - Cases of other clusters as different as possible
  - Requires some measure of similarity
  - Number of clusters needs to be determined



Business Applications of Cluster Analysis

- **Marketing**
  - Targeted marketing (mass customization)
  - Identifying need for new products / services
  - Differentiating between brands in a portfolio

- **Text analytics (e.g., document clustering)**
  - Document clustering
  - Development of document / web-page taxonomy



- **Fraud / anomaly detection**
  - Identify “unusual” cases
  - Card transactions, phone calls, network traffic, etc.

Cluster Analysis Case Study

Cluster Analysis Case Study
Targeting Leaflets

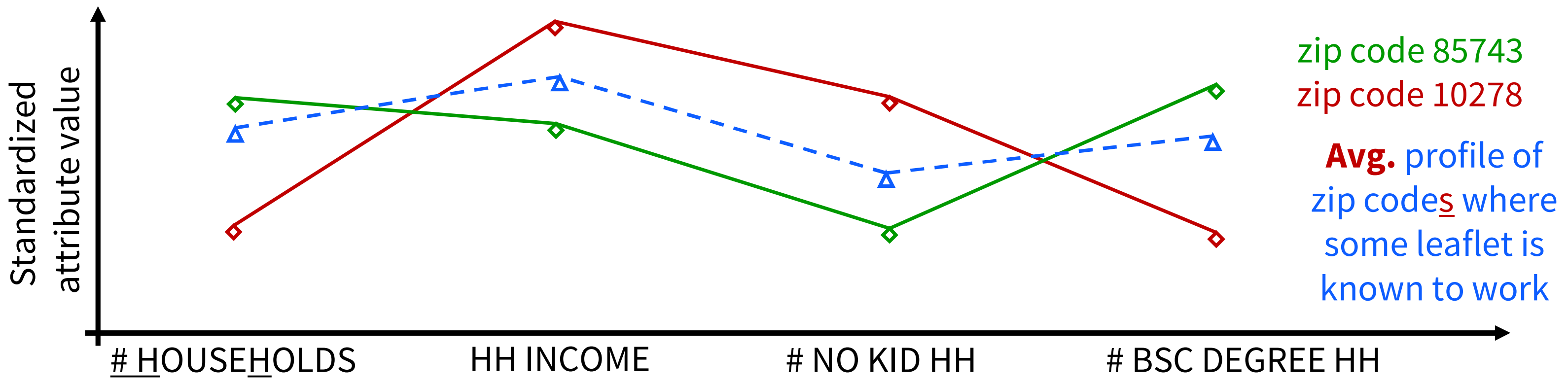
- **Leading DIY company**
  - Mainly operating in Germany, Austria, and Switzerland
  - About €7.5 bln annual turnover
- **Leaflets as major advertising channel**
  - Advertising special offers and campaigns
  - Multiple types of leaflets
  - Distribution via partners (eg newspaper)
- **Project goals**
  - Improve targeting
  - Raise return on advertising

Cluster Analysis Case Study

Using cluster analysis to develop customer profiles

- **Individual-level targeting infeasible**
- **Decision making based on areal units (zip codes)**
- **What is the "profile" of a zip code?**

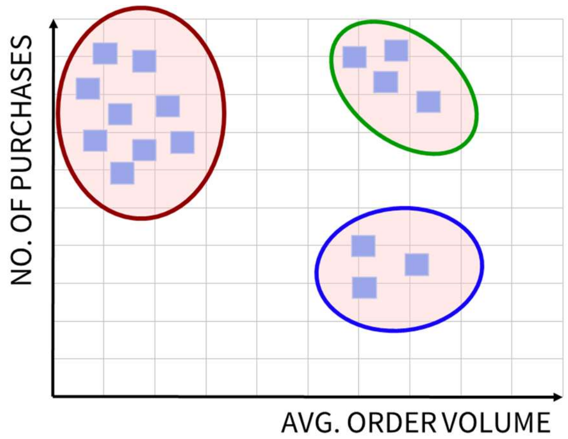
# HOUSEHOLDS	HH-INCOME	# NO KID / HH	# BSC DEGREE / HH	…
15000	35000	4500	2750	…
5700	67000	3125	1300	…
…	…	…	…	…

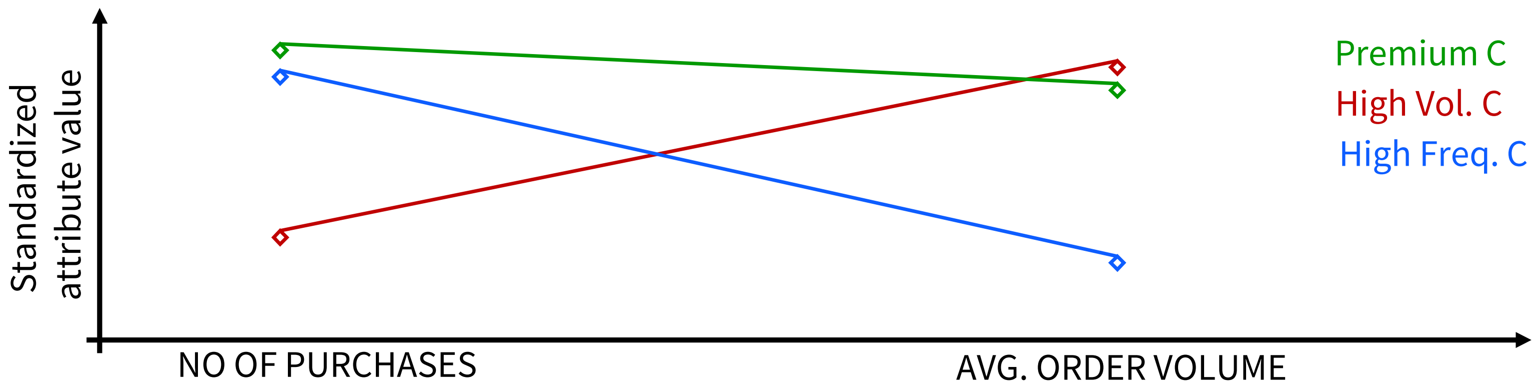


Cluster Analysis Case Study

Using cluster analysis to develop customer profiles

- **Clustering algorithm finds an allocation of objects to clusters**
- **Cluster centroids give a cluster signature with respect to attribute values (= profile)**
- **Can compare this profile to the signature of a new object**



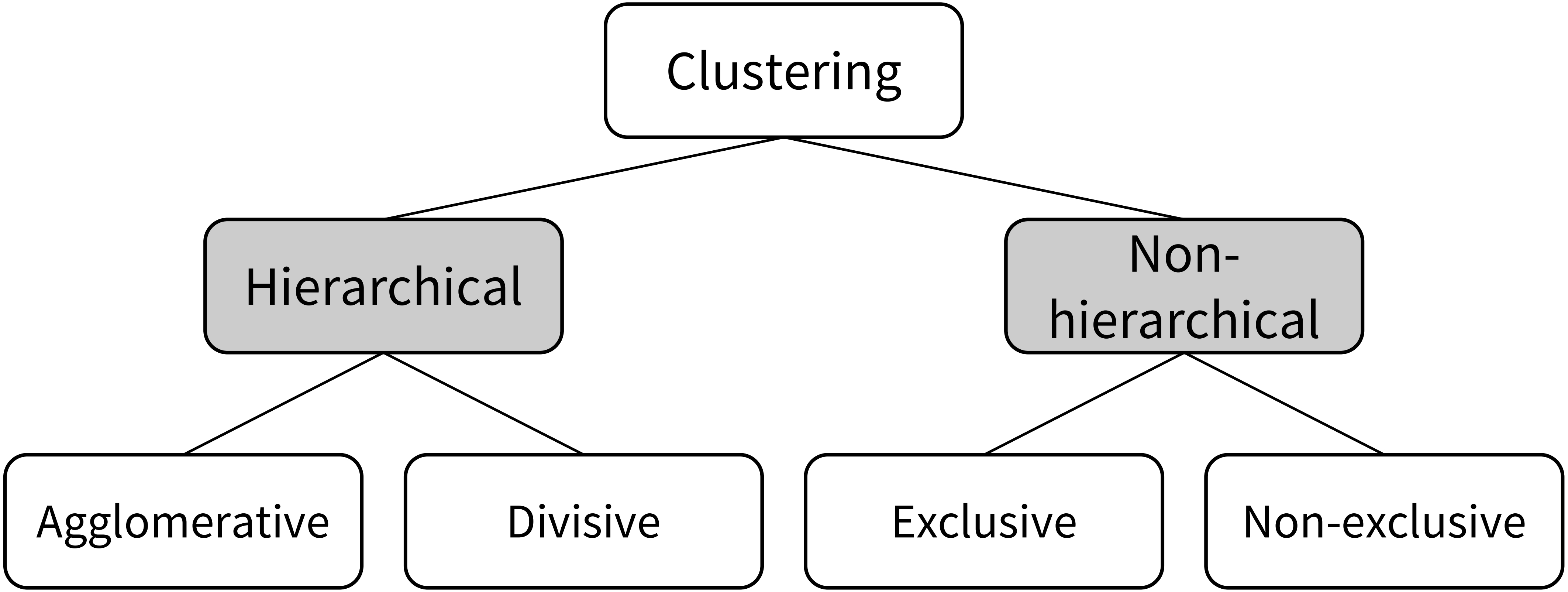


Cluster Analysis Methods

Approaches toward clustering, distance metrics, k-means algorithm

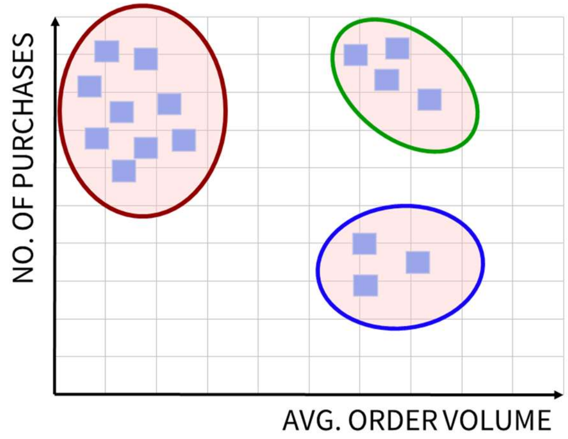
Approaches Toward Cluster Analysis

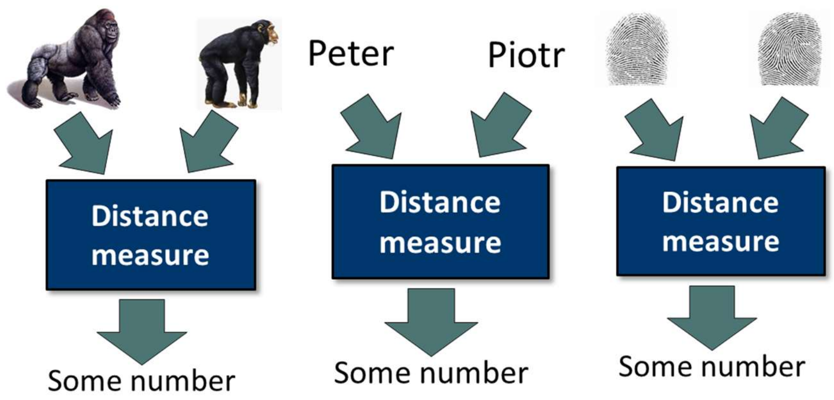
Cluster analysis is based on the distance and/or similarity between objects



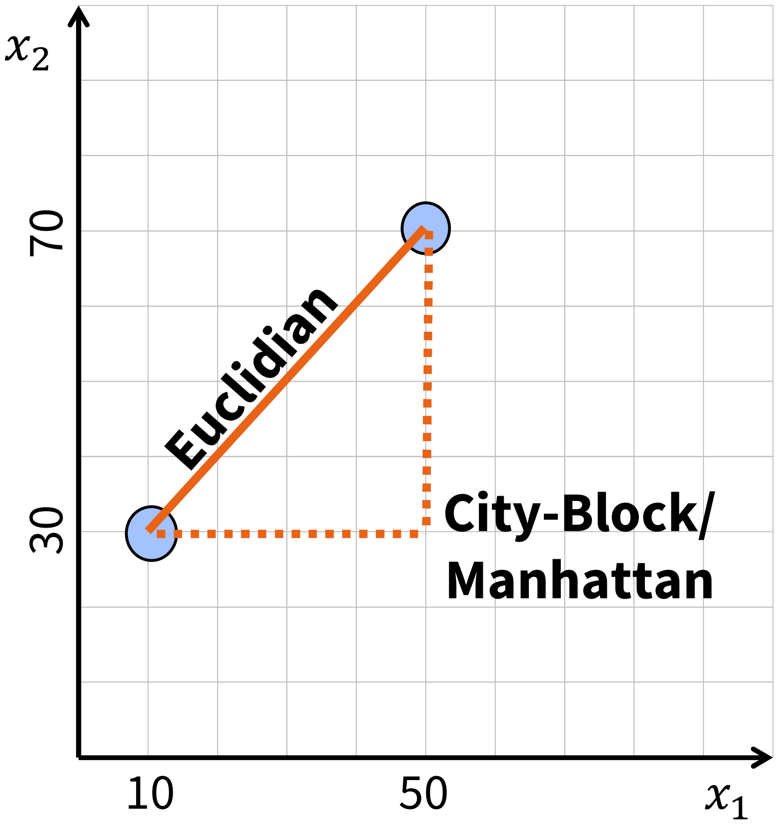
Distance and Distance Measurement

- **Aim of clustering**
  - Maximize intra-cluster homogeneity
  - Maximize inter-cluster heterogeneity
- **Distance measures**
  - Formal way to quantify (dis)similarity between objects/clusters
    - Homogeneity: average of pairwise distances of objects in the same cluster
    - Heterogeneity: distance(s) between objects of different clusters
  - Distance between two objects is a real number
  - Properties of a distance measure
    - Function of two inputs
    - Producing one output





Distance Measures for Numeric Data



Euclidian: $\sqrt{(50-10)^2+(70-30)^2} = 56.57$

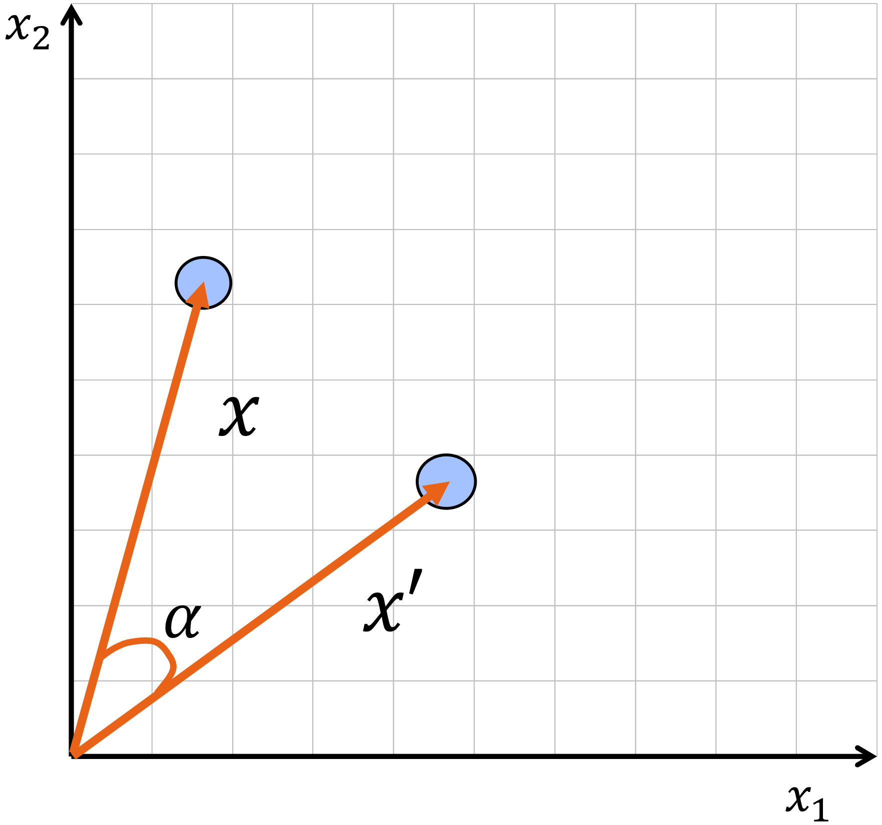
Manhatten: $|50-10| + |70-30| = 80$

Generalization

$$\text{Lp} - \text{Metric}: d_{ij} = \left(\sum_{k=i}^{m} \left|x_{ik} - x_{jk}\right|^p\right)^{1/p}$$

Distance Measures for Numeric Data (cont.)

Cosine similarity



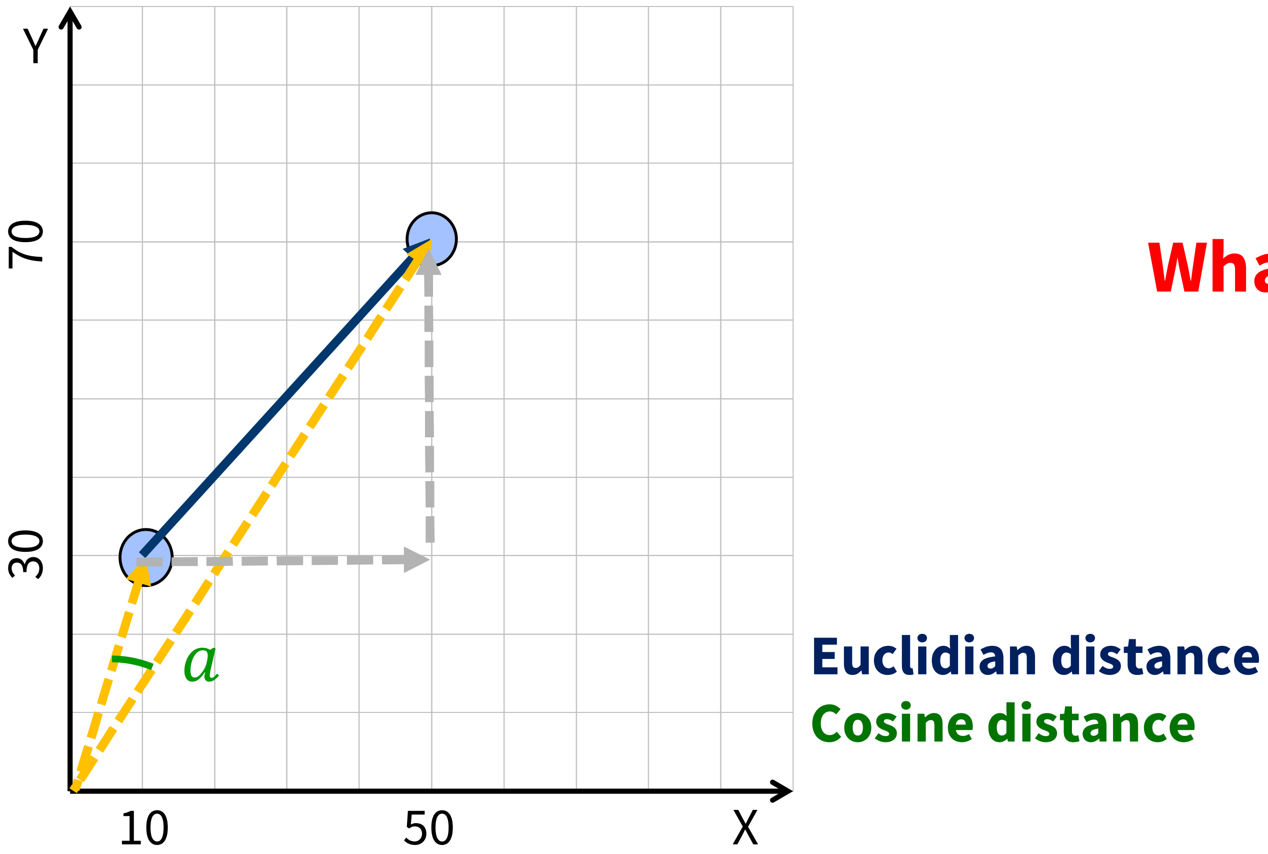
Angle between vectors:
Direction of the vector captures the relationship between variable values.

$$\boldsymbol{x} = (x_1, x_2)$$
$$\boldsymbol{x}' = (x'_1, x'_2)$$

$$\cos\alpha = \frac{\boldsymbol{x} \cdot \boldsymbol{x}'}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{x}'\|} = \frac{x_2 \cdot x'_2 + x_1 \cdot x'_1}{\sqrt{x_1^2 + x_2^2} \cdot \sqrt{x_1'^2 + x_2'^2}}$$

Distance Measures for Numeric Data (cont.)

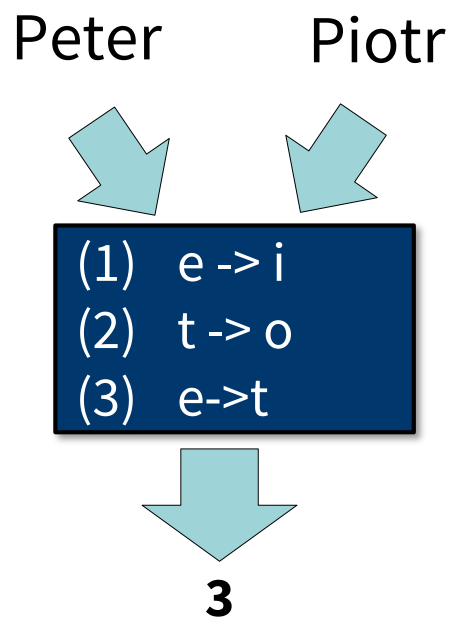
The choice of the distance measure impacts the cluster solution



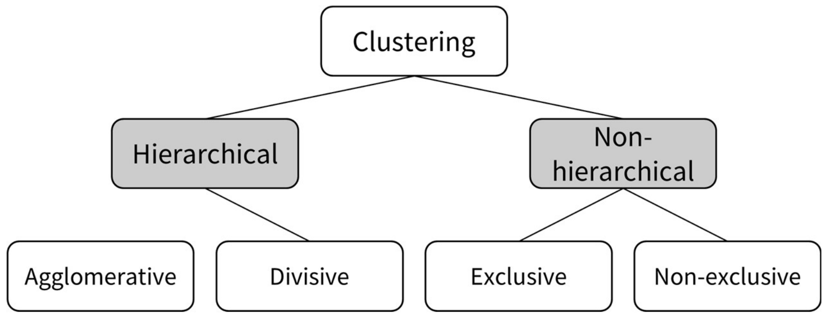
What's a good measure for a given application?

Distance Measures for Non-Numeric Data

- **Nominal variables**
  - Hamming distance
  - Jaccard similarity coefficient
- **Text**
  - Hamming distance (for texts of equal length)
  - Levenshtein distance (for texts of unequal length)
- **Graphs, time series, gene strings, streams, etc.**
- **General notion of similarity/distance**
  - No. of edit operations to transform one object into another
  - Transformation-specific costs



The k-Means Algorithm



- **Iterative algorithm based on centroids**
  - Define number of clusters, K
  - Randomly guess cluster centers (→ centroid)
  - Assign cases to nearest centroid to obtain initial cluster solution
  - Update centroids, assuming correctness of the current solution
  - Repeat until cluster assignment stops changing
- **Objective: Minimize intra-cluster variance**

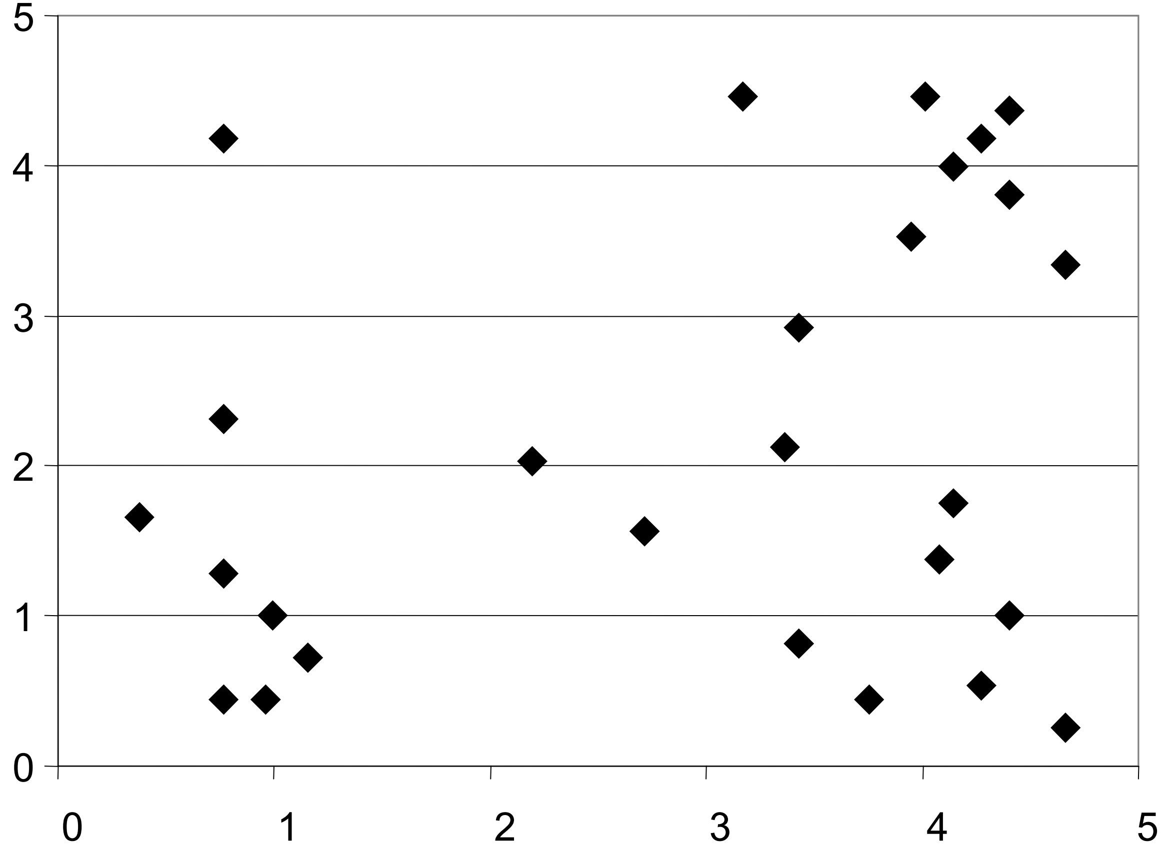
$$\mathrm{C}^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k\|^2$$

Mean vector of cluster k

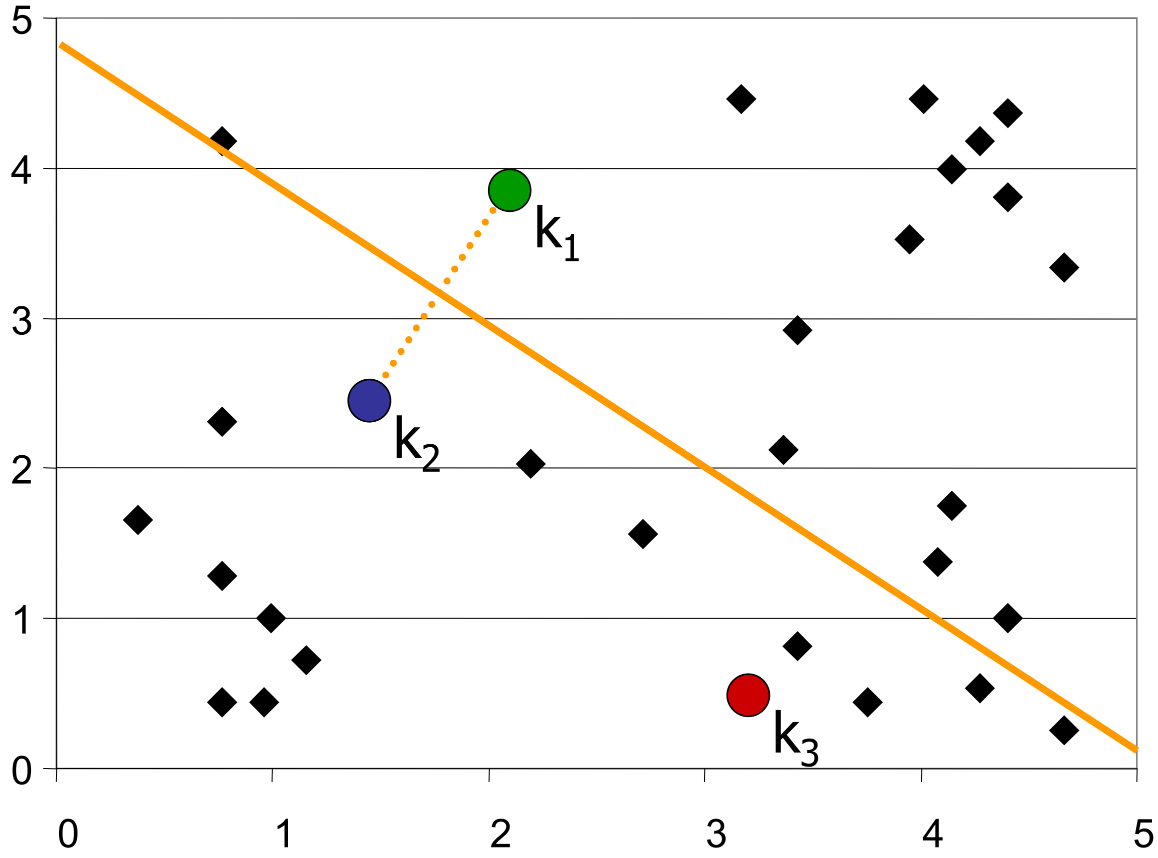
"Optimal" cluster assignment

Number of cases in cluster k

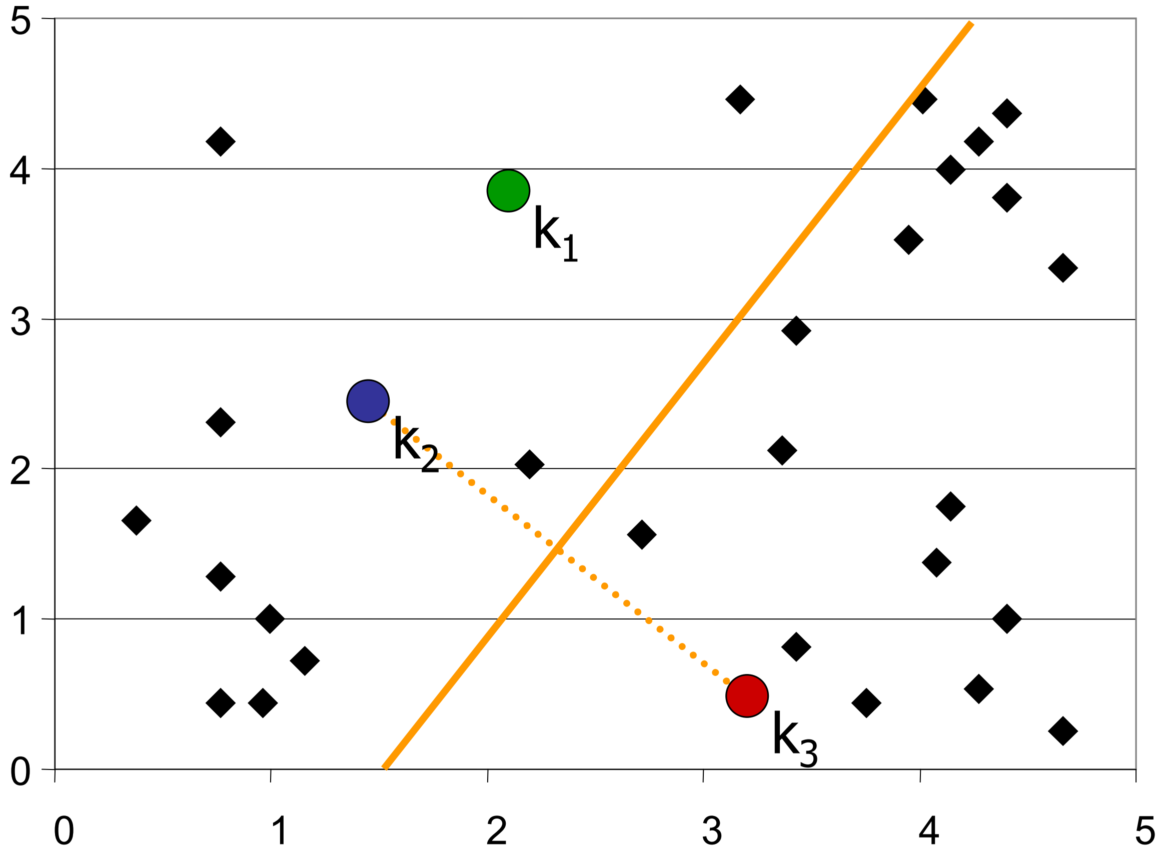
The k-Means Algorithm
A two-dimensional example



The k-Means Algorithm
A two-dimensional example

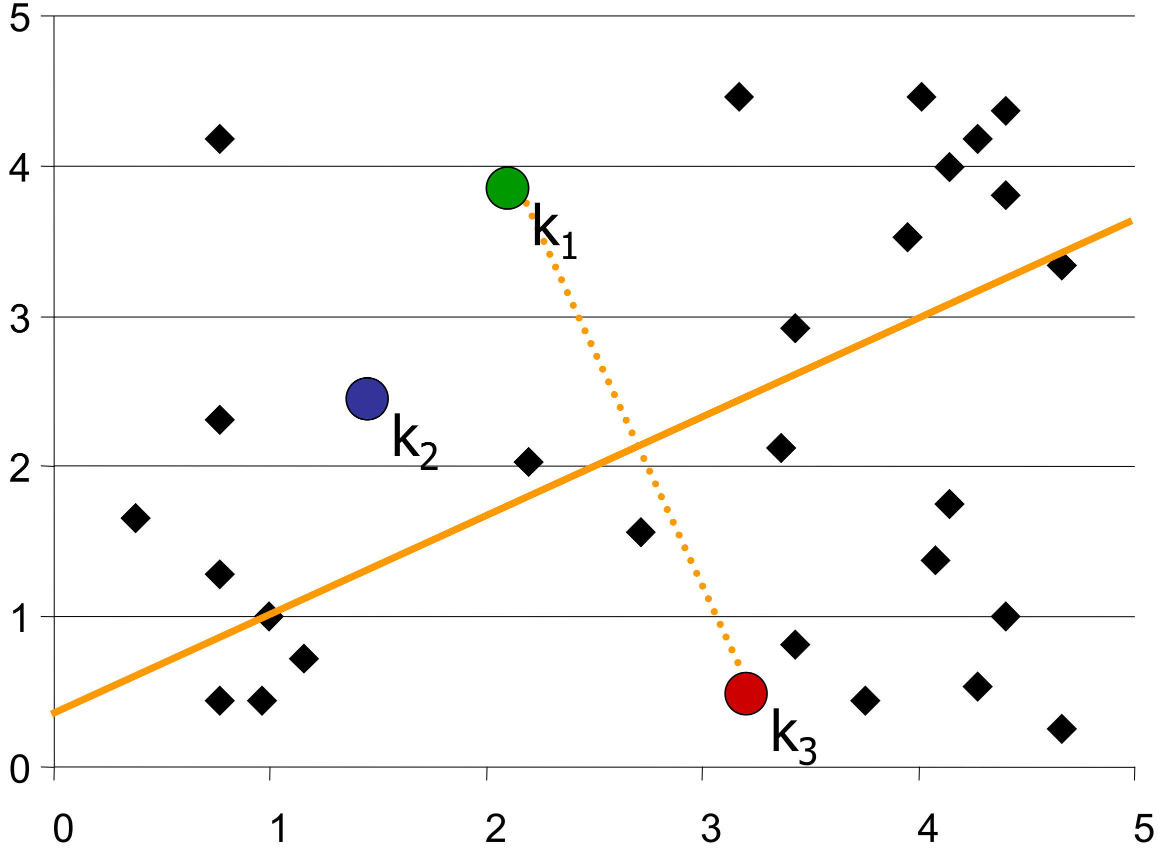


The k-Means Algorithm
A two-dimensional example

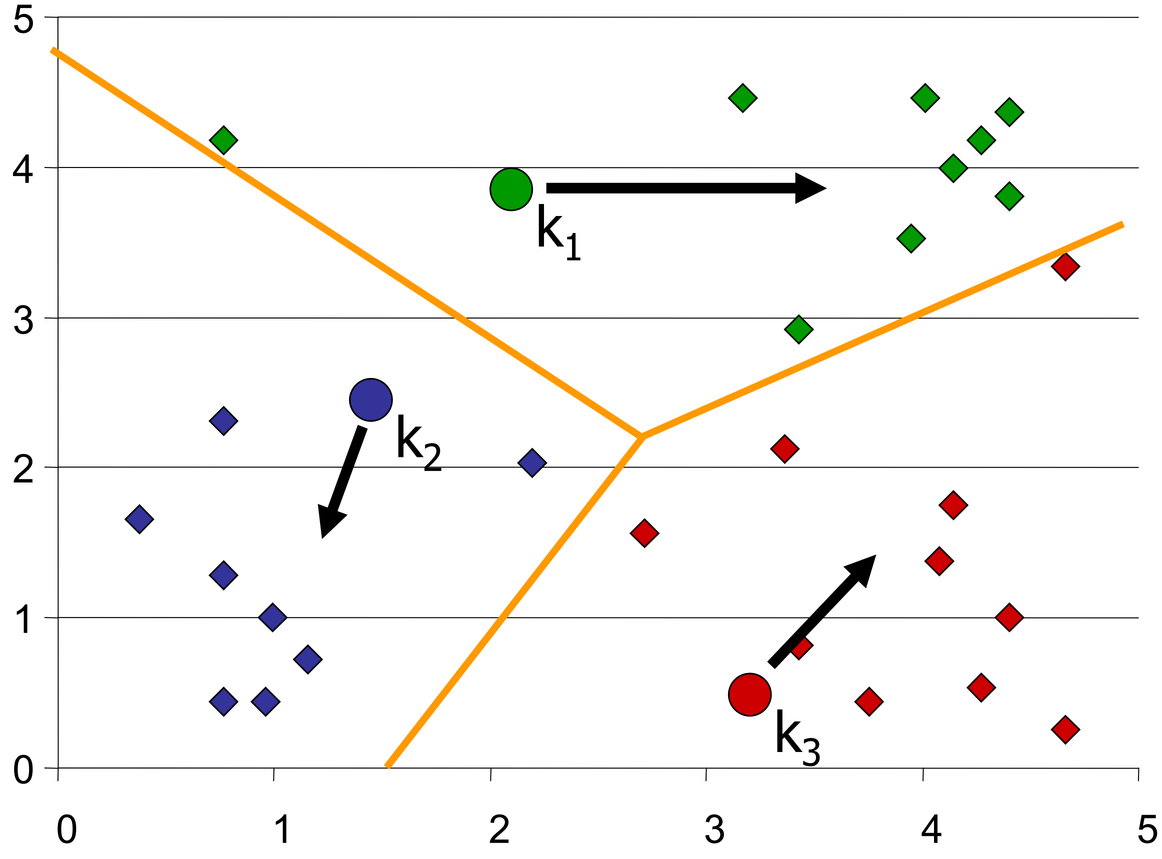


The k-Means Algorithm

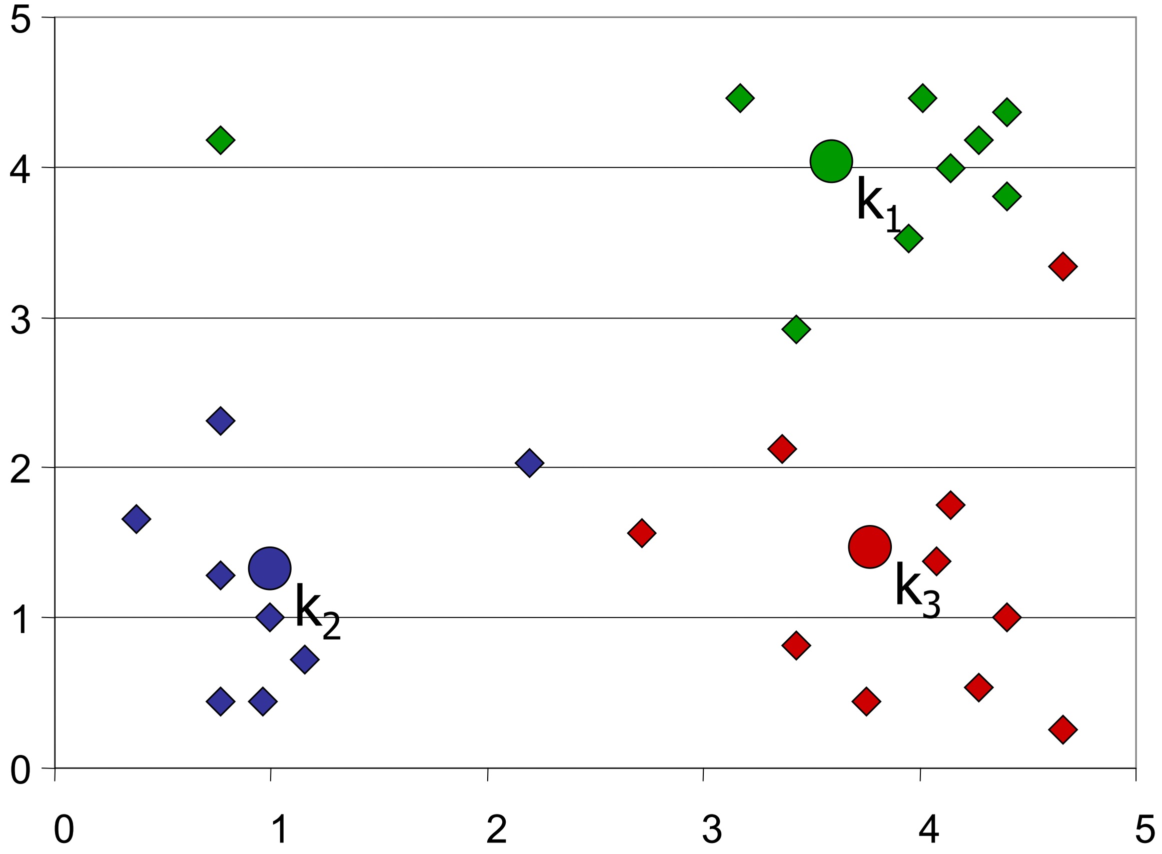
A two-dimensional example



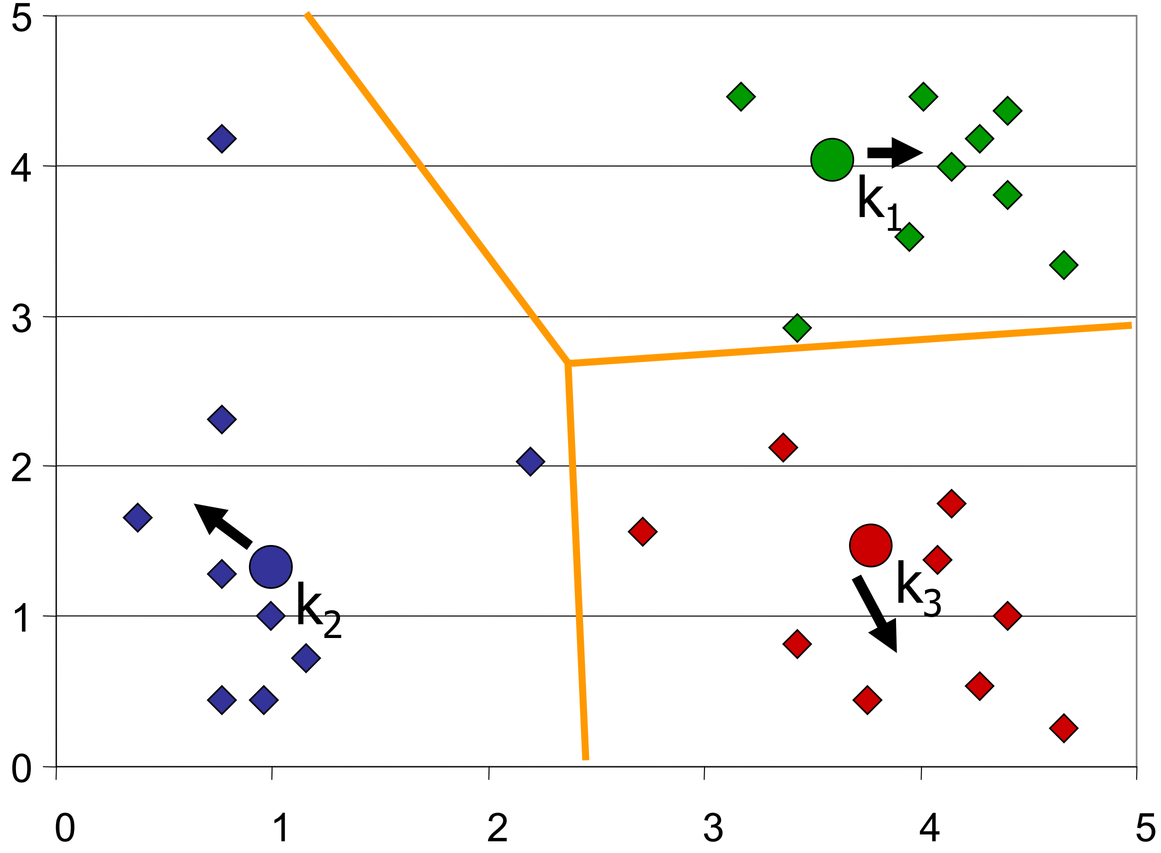
The k-Means Algorithm
A two-dimensional example



The k-Means Algorithm
A two-dimensional example

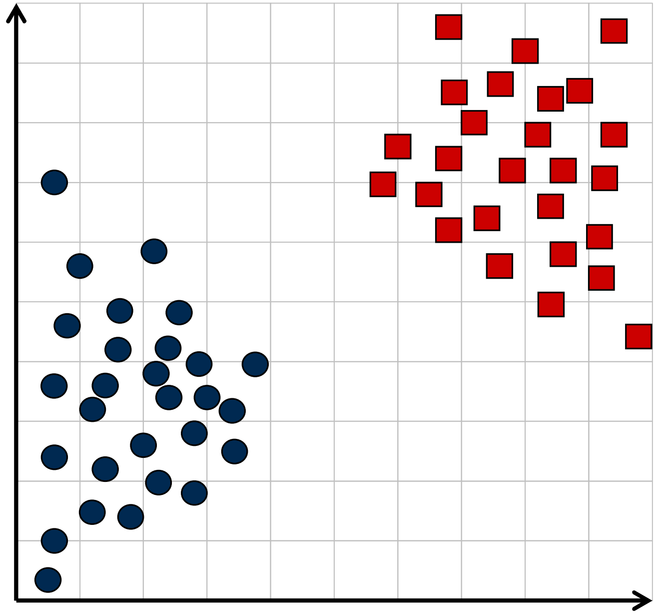


The k-Means Algorithm
A two-dimensional example



The k-Means Algorithm
Determining the number of clusters

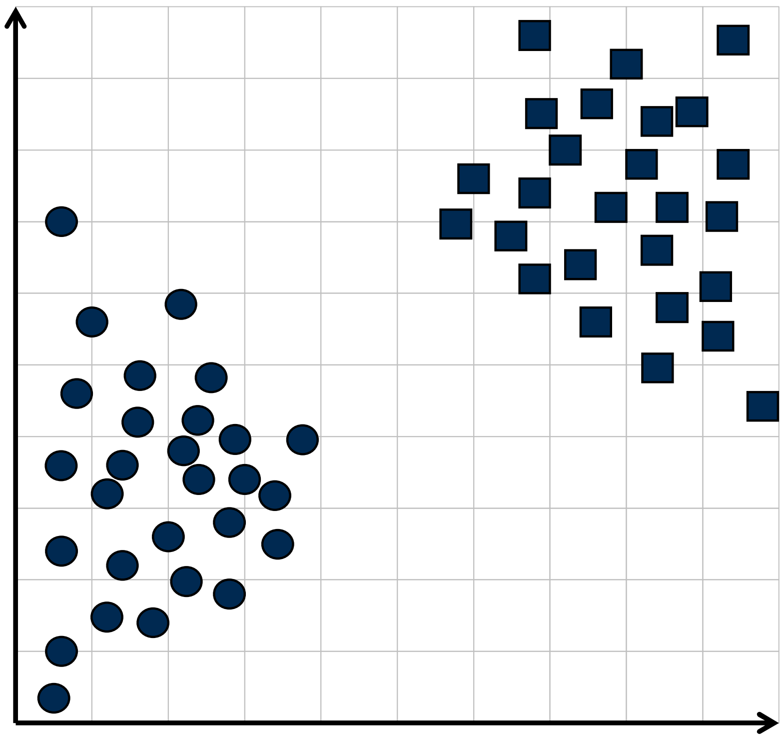
- No “oracle solution”
- Based on domain knowledge
- Heuristic approaches



Assume we do not know that our data sets exhibits two clusters. We can experiment with different values of K and see what happens.

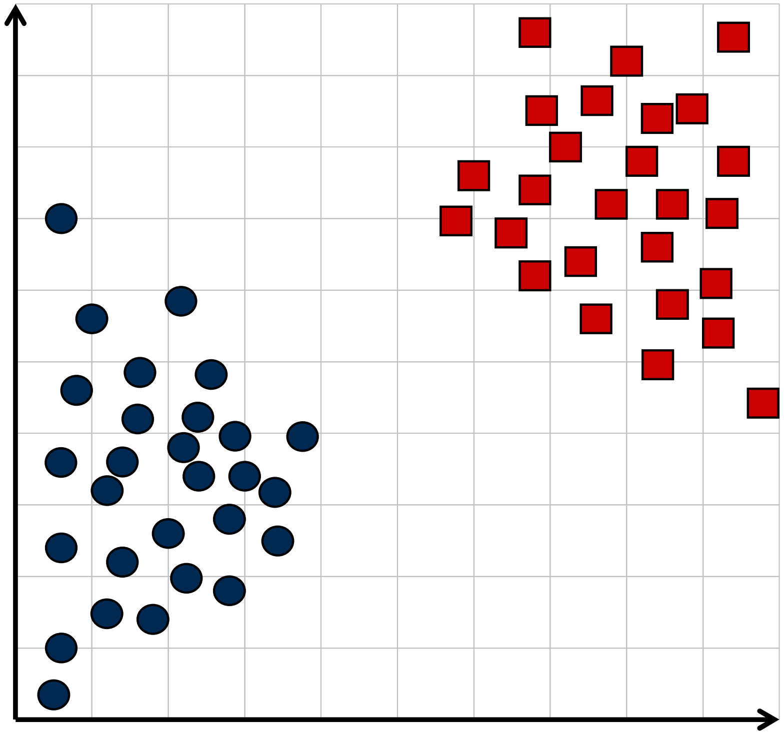
The k-Means Algorithm
Solution with K=1

- **K-Means objective**
  - Minimize intra-cluster variance
  - Sum of squared differences between members and the cluster center (i.e., centroid)
- **Assume the objective value for K=1 is 873**



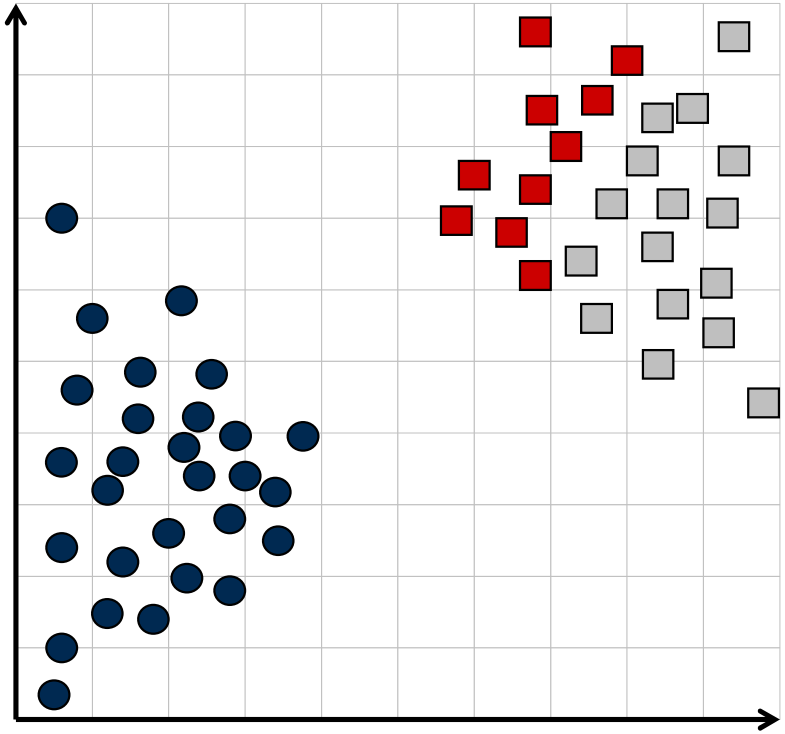
The k-Means Algorithm
Solution with K=2

- **K-Means objective**
  - Minimize intra-cluster variance
  - Sum of squared differences between members and the cluster center (i.e., centroid)
- **Assume the objective value for K=2 is 123**



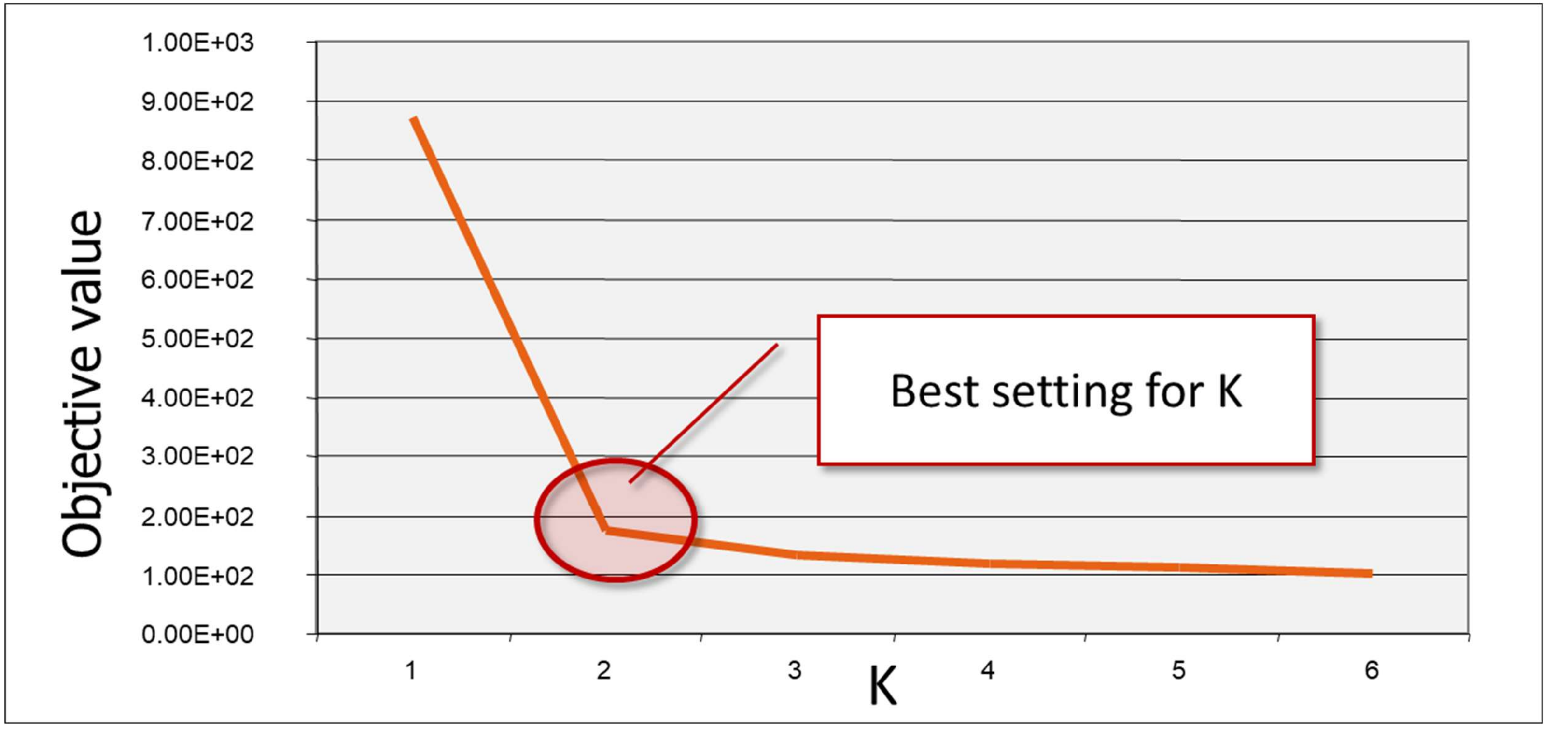
The k-Means Algorithm
Solution with K=3

- **K-Means objective**
  - Minimize intra-cluster variance
  - Sum of squared differences between members and the cluster center (i.e., centroid)
- **Assume the objective value for K=3 is 115**

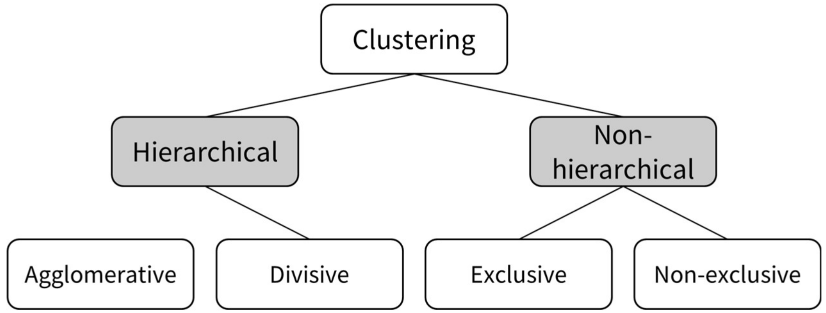


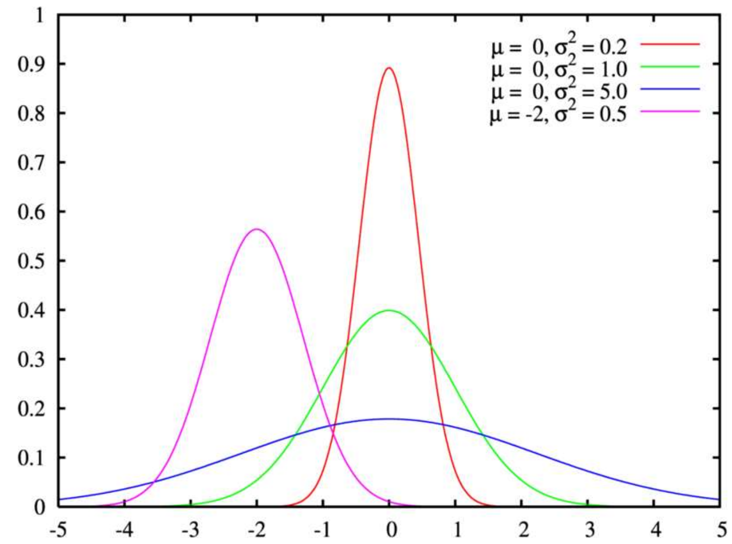
The k-Means Algorithm
Graphical heuristic to decide on K

- **Plot objective value against K**
- **Elbow-spotting**



Extensions of K-means
Exclusive vs. Non-Exclusive Clustering





- **K-Means assigns every case to exactly one cluster**
- **Gaussian mixture models (GMM)**
  - Soft-cluster assignment via cluster-membership probabilities
  - More robust toward outliers & better for overlapping distributions
- **GMM in a nutshell**
  - Model data using mixture of k Gaussians
  - Each mixture component influences every observation
    - Strong influence if a case is close to mean vector
    - Small influence otherwise
- **Estimate using E(xpectation)-M(aximization) algorithm**
  - Start from random solution and iterate between E- and M-step
  - **E**-step: for each case, compute association to mixture components (called “responsibility”), given mixture parameter (i.e., mean vector and covariance matrix)
  - **M**-step: re-compute parameters based on responsibilities

Evaluation of Clustering Solutions

Silhouette score

- **Measure of cluster cohesion versus separation**
- **How similar is an object is to its own cluster compared to other clusters?**
- **Values range from $[-1, +1]$ for an individual data point**
- **High scores indicate that data point is well matched to its own cluster and poorly to neighboring clusters**
- **Averaging scores across data points and clusters provides a global score of the quality of the clustering**

Evaluation of Clustering Solutions
Silhouette score calculation

- **Mean similarity (e.g., distance) between x_i and other points in the same cluster $\mathcal{C}_I$**

$$a(\boldsymbol{x}_i) = \frac{1}{|\mathcal{C}_I| - 1} \sum_{j \in \mathcal{C}_I, i \neq j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- **Minimal mean dissimilarity between x_i and data points of other clusters $\mathcal{C}_J$ where $J \neq I$**

$$b(\boldsymbol{x}_i) = \min_{J \neq I} \frac{1}{|\mathcal{C}_J|} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  - Cluster with minimal mean dissimilarity is the next best fit for data point $\boldsymbol{x}_i$
  - Also called neighboring cluster

- **Silhouette of x_i**

$$s(\boldsymbol{x}_i) = \frac{b(\boldsymbol{x}_i) - a(\boldsymbol{x}_i)}{\max(a(\boldsymbol{x}_i), b(\boldsymbol{x}_i))}$$ provided $|\mathcal{C}_I| > 1$, and 0 otherwise

- **Finally, we obtain**

$$SC = \max_k \bar{s}_k$$

  - Where $\bar{s}_k$ represents the mean $s(\boldsymbol{x}_i)$ over all data points for a specific number of clusters k

Evaluation of Clustering Solutions
Other criteria

- **Silhouette score is one of only few measures that are truly unsupervised**
- **Several other options**
  - Davies-Bouldin score (average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances)
  - Calinski and Harabasz score / Variance ratio criterion (ratio of the sum of between-cluster dispersion and of within-cluster dispersion)
  - Rand index (computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings)
  - For further examples/more information, see, e.g.,
    - https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics.cluster
    - https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6
    - https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/

Summary

Summary

Learning goals

- Forms of descriptive analytics
- Functioning of selected methods

Findings

- Flavors: segmentation, rule mining, dim. reduction
- Business apps & use cases of cluster analysis
- Cluster analysis is all about the similarity of objects, which we measure using distances
- Functioning of kMeans

What next

- Foundations of predictive analytics
- Business applications and algorithms

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093. 99540
Fax. +49.30.2093. 99541

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de