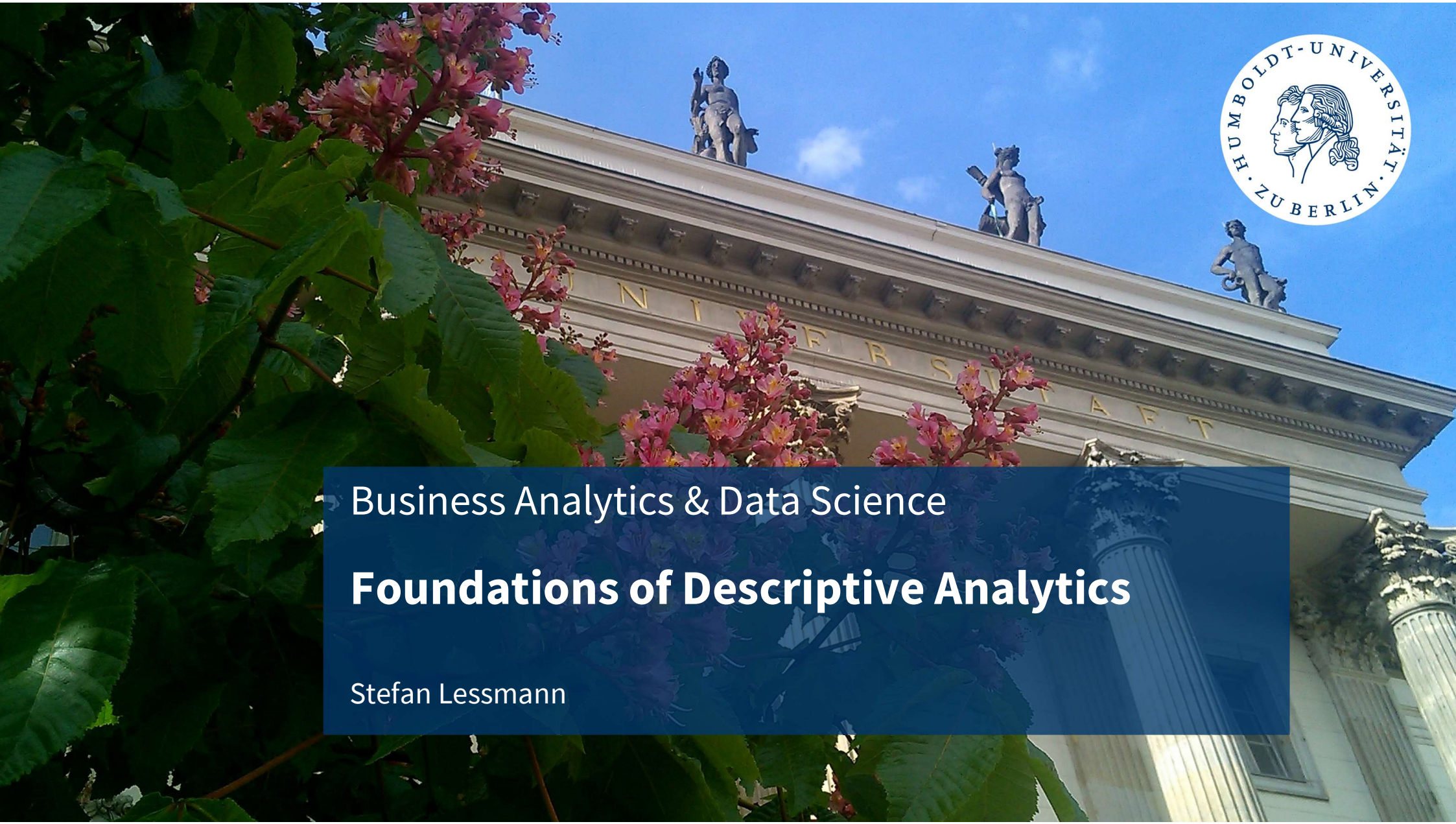
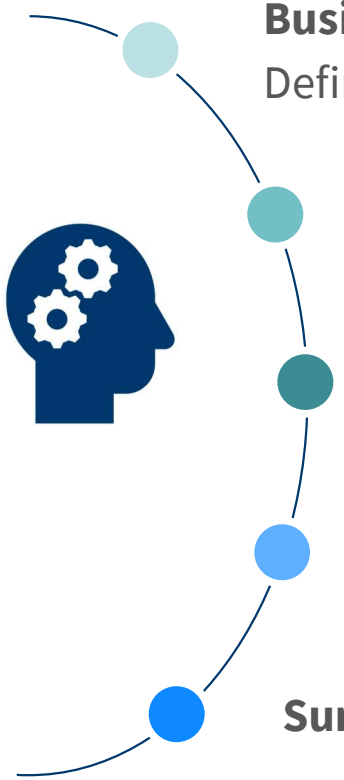


Business Analytics & Data Science

Foundations of Descriptive Analytics

Stefan Lessmann

Agenda



Business Analytics Revisited
Definitions, analytics process model

It is All About Data
Types of data, terminology, and a bit of formalism

Descriptive Analytics in a Nutshell
Scope and flavors, business applications

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

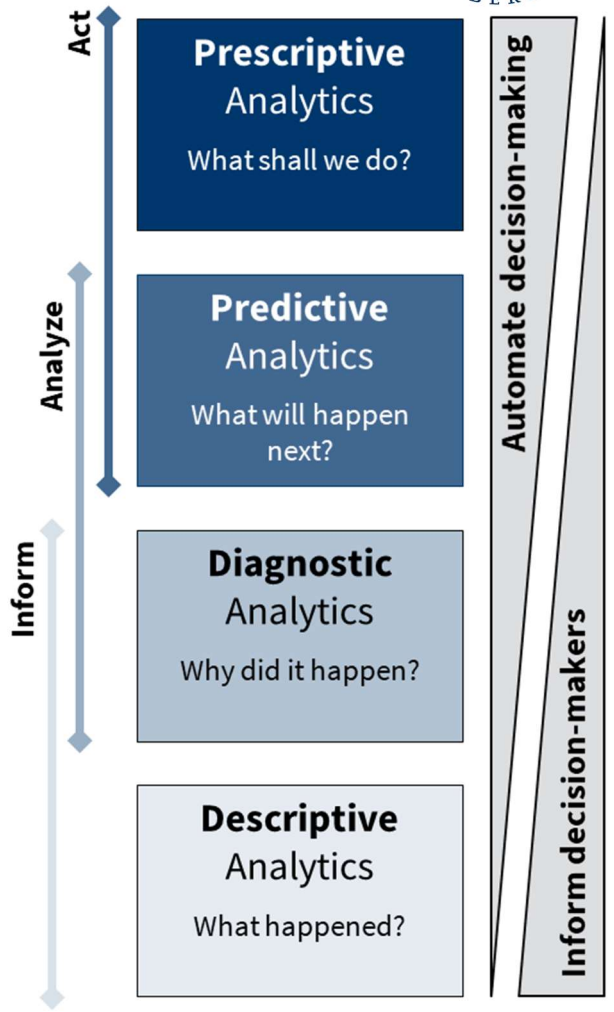
Summary

Business Analytics Revisited

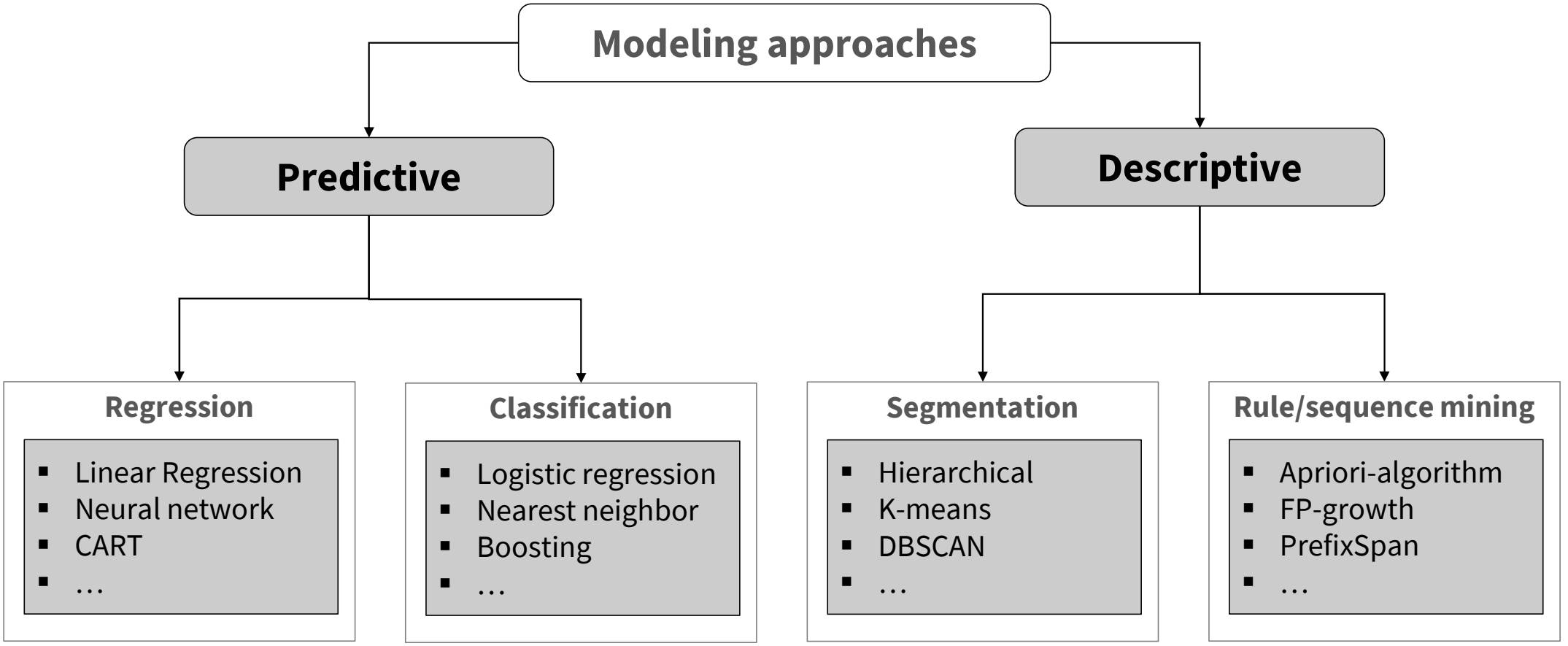
Definitions, analytics process model, data structure

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



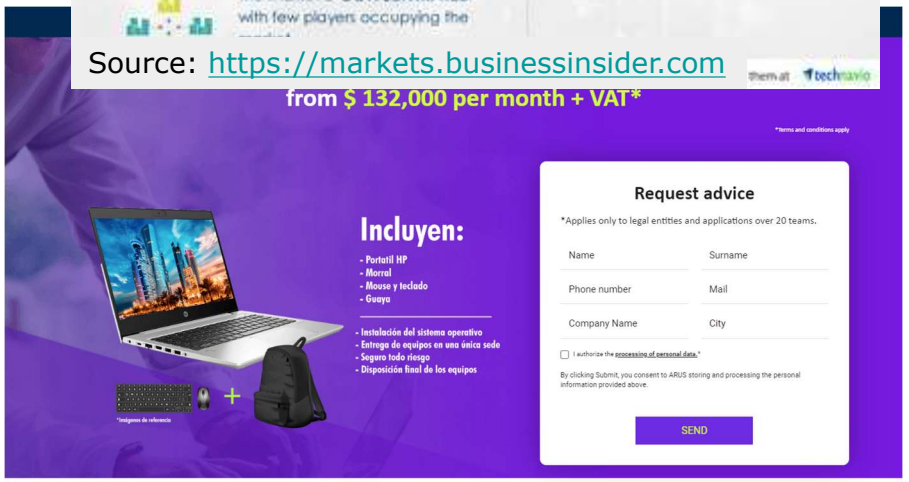
Recap: Business Use Case II: Leasing Industry

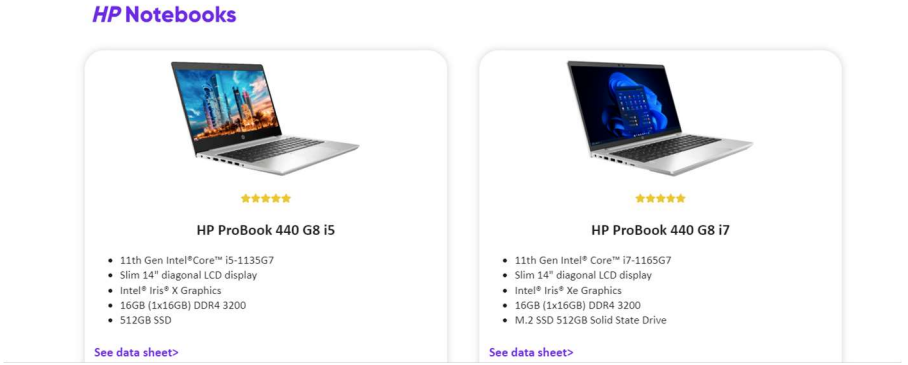
- **Important channel to market durables**
  - Most prominently cars
  - Machinery, IT equipment, etc.
- **Typical setup**
  - Clients lease equipment for a given period
  - Provider receives monthly fee
  - Client returns the item when contract expires
  - Provides resales the used item in the second-hand market
- **Business question:**

  how to price a leasing contracts?

- **Machine Learning support?**

Source: https://markets.businessinsider.com

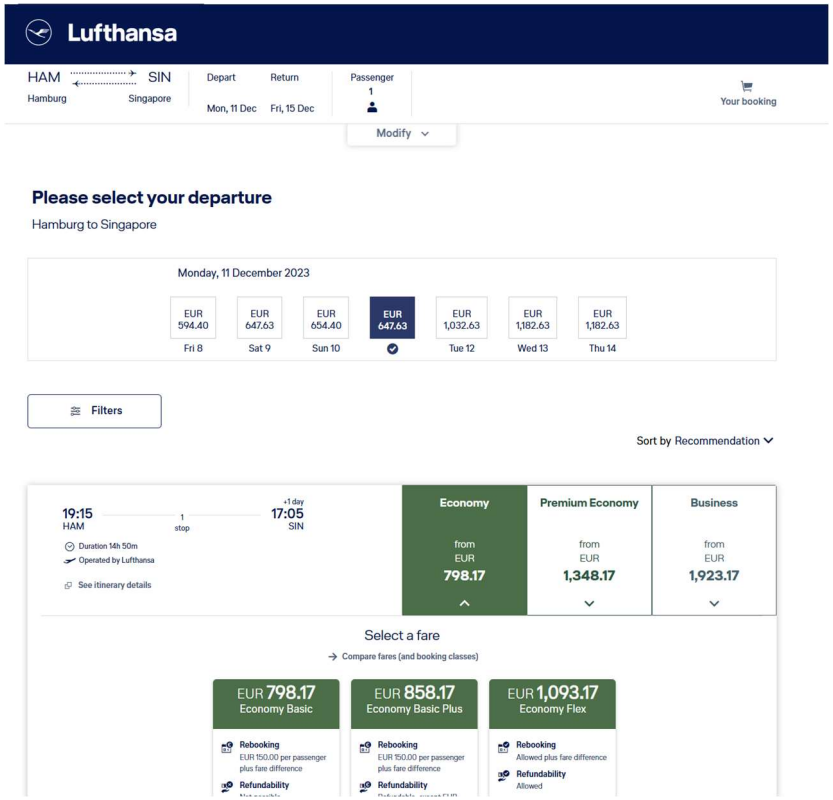


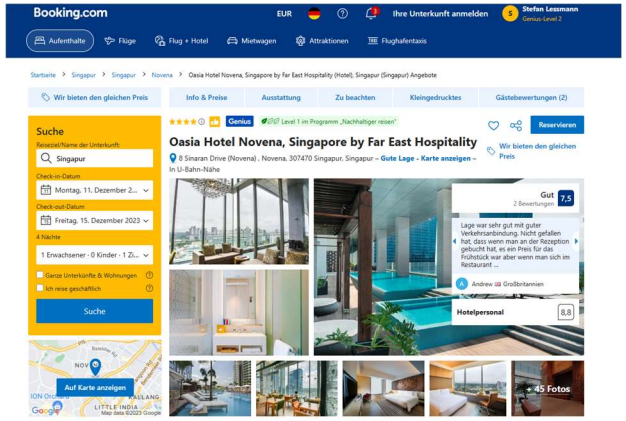


Recap: Business Use Case III: Tariff Design

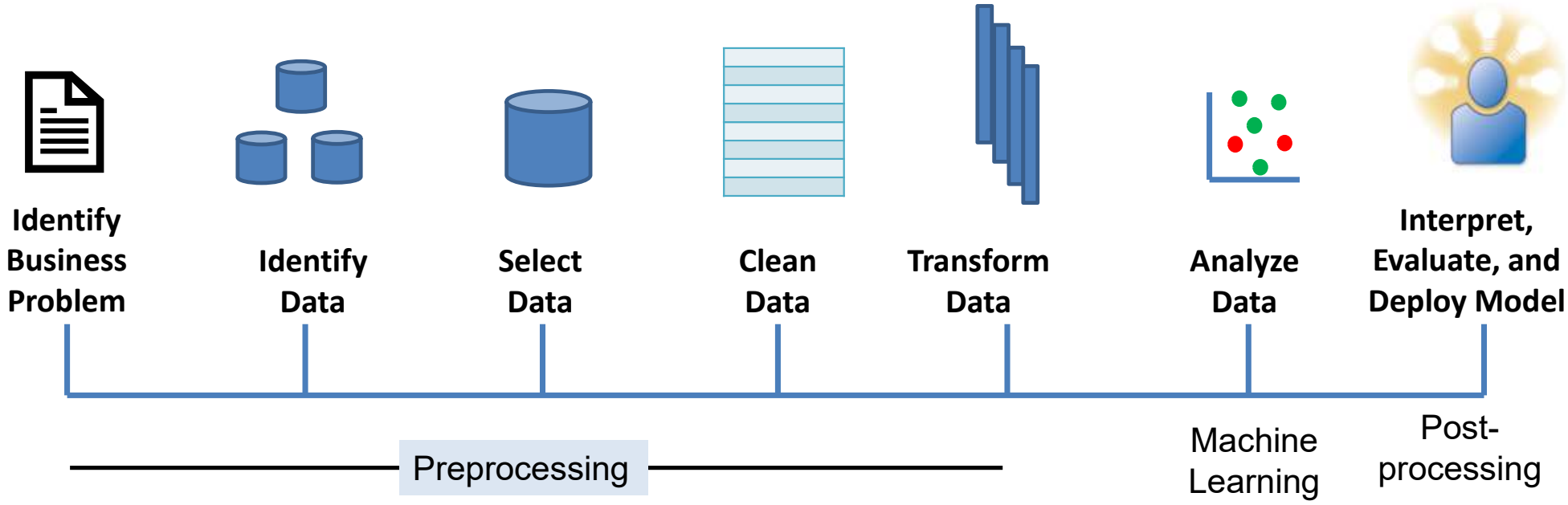
- **Travel & hotel industry, transportation, app industry, etc.**
- **Much literature on price discrimination & dynamic pricing**
  - How supplier can maximize revenue
  - How to avoid premium customers moving to lower quality channel
- **Business question:**

  how many different groups of customers exist?
- **Machine Learning support?**





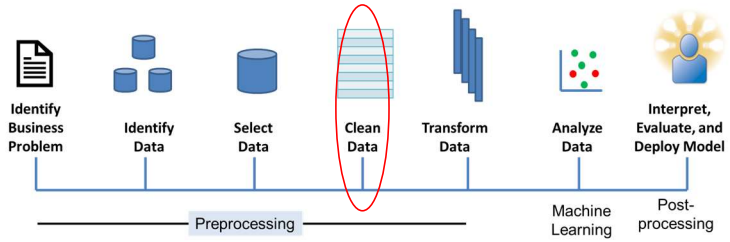
Recap: The Analytics Process Model



It is All About Data

Types of data, terminology, and a bit of formalism

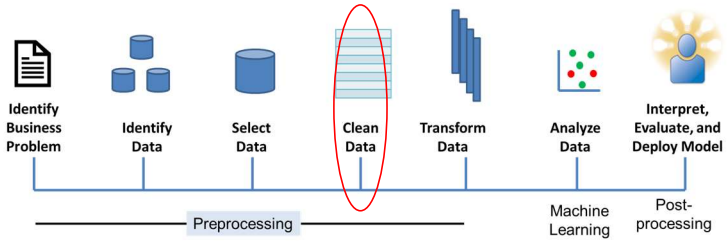
Resale Price Forecasting Data



- **Data describing leasing items along a range of attributes**
- **True value of the (eventually!) observed resale price**
  - Setup assumes that the actual target of the forecasting exercise is known
  - Day-to-day business operations will provide this kind of data

PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…	**OBSERVED RESALE PRICE [$]**
Dell XPS 15’	2,500	36	Mining	…	347
Dell XPS 15’	2,500	24	Health	…	416
Dell XPS 17’	3,000	36	Manufacturing	…	538
HP Envy 17’	1,300	24	Office	…	121
HP EliteBook 850	1,900	36	Manufacturing	…	172
Lenovo Yoga 11’	799	12	Office	…	88
Lenovo Yoga 13’	1,100	12	Office	…	266
…	…	…	…	…	…

Tariff Design Data



- **Data describing individual customers along a range of attributes**
  - Demographic information (age, gender, ...)
  - Behavioral information (e.g., web site usage: time on web page, return frequency, no. of searchers, …)
  - Transactional information (e.g., no. previous trips, destinations, lengths, …)

CLIENT	AGE [YEARS]	WEBSITE VISITS/MONTH	AVG. SPENDING AMOUNT	TOTAL NO. OF TICKETS	LAST DESTINATION	MODE DEPARTURE	…
Mary	61	36	375 €	43	Virginia	Hamburg	…
Peter	53	24	125 €	3	Warsaw	Munich	…
Ying	18	36	2500 €	7	Singapore	Berlin	…
Carlos	36	24	1500 €	18	Havana	Munich	…
Emre	21	36	980 €	27	Frankfurt	Frankfurt	…
Güley	48	12	650 €	215	Berlin	Frankfurt	…
…	…	…		…	…	…	…

Data Science Lingo

Target, outcome, label, response (variable), dependent (variable)

Features, attributes, characteristics, covariates, predictors, (independent) variables

Observations, cases, examples, data items, subjects

PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…	**OBSERVED RESALE PRICE [$]**
Dell XPS 15’	2,500	36	Mining	…	347
Dell XPS 15’	2,500	24	Health	…	416
Dell XPS 17’	3,000	36	Manufacturing	…	538
HP Envy 17’	1,300	24	Office	…	121
HP EliteBook 850	1,900	36	Manufacturing	…	172
Lenovo Yoga 11’	799	12	Office	…	88
Lenovo Yoga 13’	1,100	12	Office	…	266
…	…	…	…	…	…

A More Formal Perspective on Data

Note how mathematical notation aids abstraction

$\boldsymbol{X} = (X_1, X_2, \ldots, X_m) \in \mathbb{R}^m$ $\quad Y \in \{0,1\}$

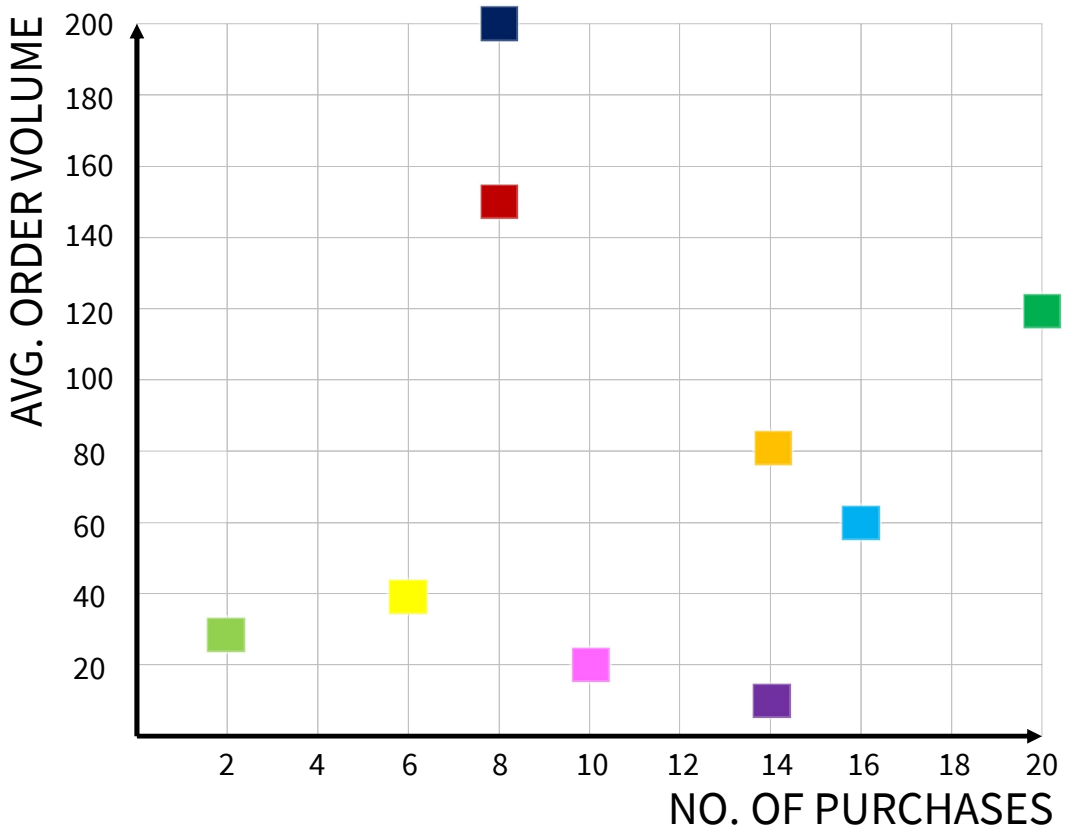
$\mathcal{D} = \{Y_i, \boldsymbol{X}_i\}_{i=1}^n$

PRODUCT	LIST PRICE [$]	AGE [month]	CLIENT INDUSTRY	…	**OBSERVED RESALE PRICE [$]**
Dell XPS 15’	2,500	36	Mining	…	347
Dell XPS 15’	2,500	24	Health	…	416
Dell XPS 17’	3,000	36	Manufacturing	…	538
HP Envy 17’	1,300	24	Office	…	121
HP EliteBook 850	1,900	36	Manufacturing	…	172
Lenovo Yoga 11’	799	12	Office	…	88
Lenovo Yoga 13’	1,100	12	Office	…	266
…	…	…	…	…	…

A Visual Perspective on Tabular Data

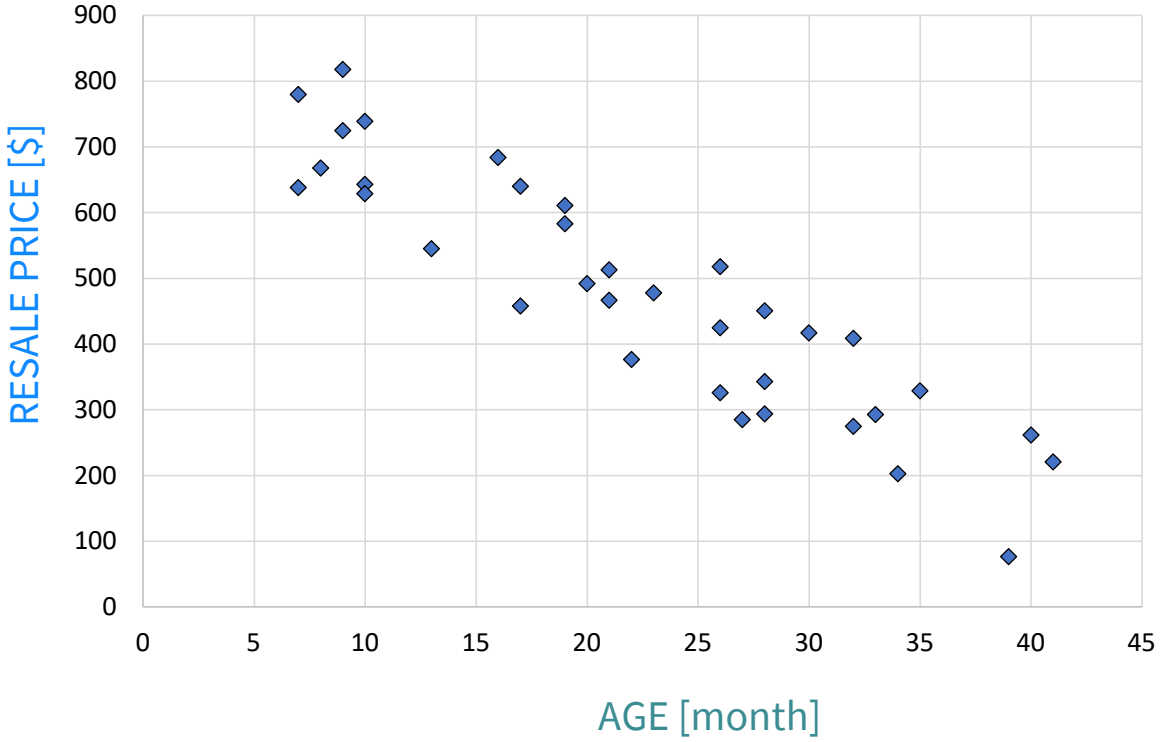
An observation equates to a data point in a multi-dimensional feature space

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



A Visual Perspective on Data

Resale price forecasting example



PRODUCT	LIST PRICE	AGE	…	RESALE PRICE
Dell XPS 15’	2,500	36	…	347
Dell XPS 15’	2,500	24	…	416
Dell XPS 17’	3,000	36	…	538
HP Envy 17’	1,300	24	…	121
HP EliteBook	1,900	36	…	172
Lenovo Yoga 11’	799	12	…	88
Lenovo Yoga 13’	1,100	12	…	266
…	…	…	…	…

Many Other Exciting Forms of Data Exist

- **Other structured, tabular data**
  - Panel time series, etc.
  - See next session
- **Unstructured data**
  - Text, images, audio, video, etc.
  - Older approaches mapped such data into a tabular format
  - Advanced approaches use specialized components for different data types
  - Most recently, we see a convergence toward one type of approach, which works well with all sorts of data (see deep learning session)
- **Multimodal models**
  - Integrate multiple diverse types of data
  - GPT-4, for example, accepts text and image inputs, and outputs text

This is a piece of text. There is no a priori defined structure in this text. Sentences can be long or short. The building block of text data is a word (or character). We can interpret text as a sequence of words. A learning algorithm for text data needs to understand the meaning of words. NLP is the discipline concerned with crafting such algorithms. Much corporate data is available in textual form, which makes NLP a mega-topic on managers' agenda.

Obtain a *tabular** representation of the unstructured data				
X_1	X_2	…	X_m	Y

*Shameless simplification! Modern AI methods represent unstructured data as vectors, matrices, and/or tensors.

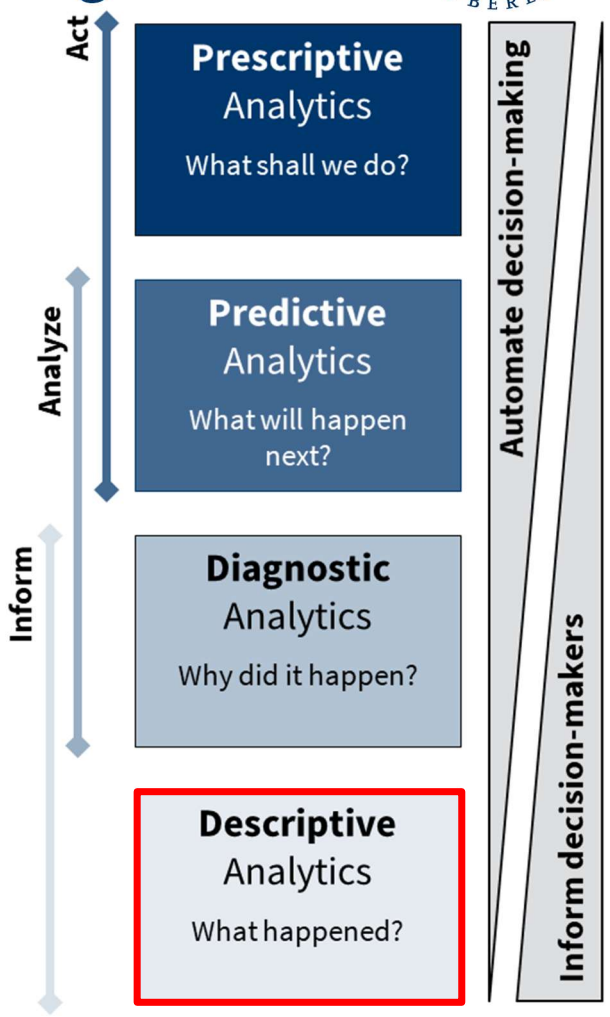
Descriptive Analytics in a Nutshell

Scope and flavors, business applications

Descriptive Analytics

Employs algorithms from the field of unsupervised learning

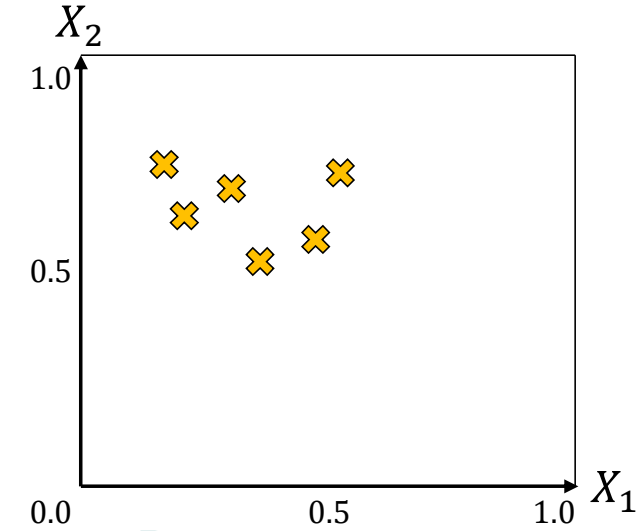
- **Data set with several features and no target variable**
- **Find structure / patterns in the data**
- **Multiple forms**
  - Dimensionality reduction
  - Clustering Association rule / Sequence mining
- **Widely applicable and versatile**
  - Plain (i.e. unlabeled) data is easily available
  - Corporate databases, spreadsheets, documents, web pages, images,…
- **Often hard to evaluate descriptive models**
  - Evaluation is a comparative exercise
  - Compare expectations/targets to what has really happened or a basline
  - How to evaluate when there is no ground truth data is available?
- **Typical business use case it to inform decision-making**



Unsupervised Machine Learning

Discover patterns in data

- **Multiple forms**
  - Dimensionality reduction, clustering, rule mining
  - Look for a specific type of “pattern”
- **Example: dimensionality reduction**
  - Remove features while sustaining their informational value
  - Extract latent features (e.g., principal components)



Observable features

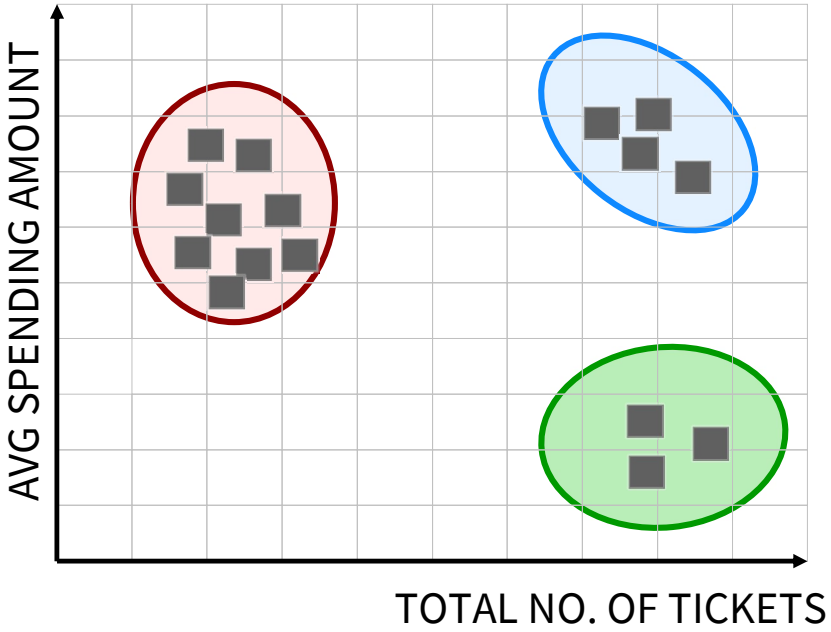
CLIENT	AGE [YEARS]	WEBSITE VISITS/MONTH	AVG. SPENDING AMOUNT	TOTAL NO. OF TICKETS	LAST DESTINATION	MODE DEPARTURE	…
Mary	61	36	375 €	43	Virginia	Hamburg	…
Peter	53	24	125 €	3	Warsaw	Munich	…
Ying	18	36	2500 €	7	Singapore	Berlin	…
Carlos	36	24	1500 €	18	Havana	Munich	…
Emre	21	36	980 €	27	Frankfurt	Frankfurt	…
Güley	48	12	650 €	215	Berlin	Frankfurt	…
…	…	…		…	…	…	…

Latent features

X_1	X_2
0.5	1.0
0.1	0.3
1.0	0.4
0.4	0.7
0.4	0.2
0.1	0.1
…	…

Unsupervised Machine Learning

Discover patterns in data



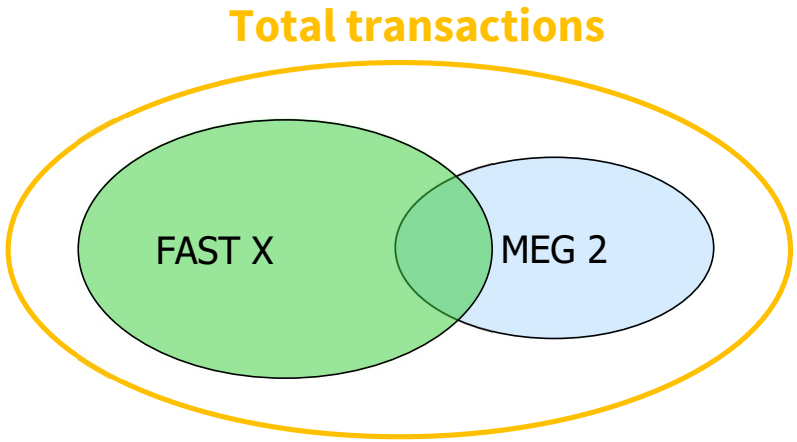
- **Multiple forms**
  - Dimensionality reduction, clustering, rule mining
  - Look for a specific type of “pattern”
- **Example: clustering**
  - Detect homogeneous sub-groups in the data
  - Use distance functions to measure homogeneity/heterogeneity

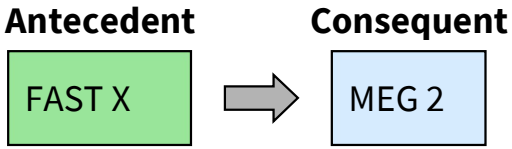
CLIENT	AGE [YEARS]	WEBSITE VISITS/MONTH	AVG. SPENDING AMOUNT	TOTAL NO. OF TICKETS	LAST DESTINATION	MODE DEPARTURE	…
Mary	61	36	375 €	43	Virginia	Hamburg	…
Peter	53	24	125 €	3	Warsaw	Munich	…
Ying	18	36	2500 €	7	Singapore	Berlin	…
Carlos	36	24	1500 €	18	Havana	Munich	…
Emre	21	36	980 €	27	Frankfurt	Frankfurt	…
Güley	48	12	650 €	215	Berlin	Frankfurt	…
…	…	…		…	…	…	…

Unsupervised Machine Learning

Discover patterns in data

- **Multiple forms**
  - Dimensionality reduction, clustering, rule mining
  - Look for a specific type of “pattern”
- **Example: association rules**
  - Detect co-occurrences that appear *unusually often*
  - Market basket analysis, path analysis, etc.
  - Main challenge is computational complexity in large assortments





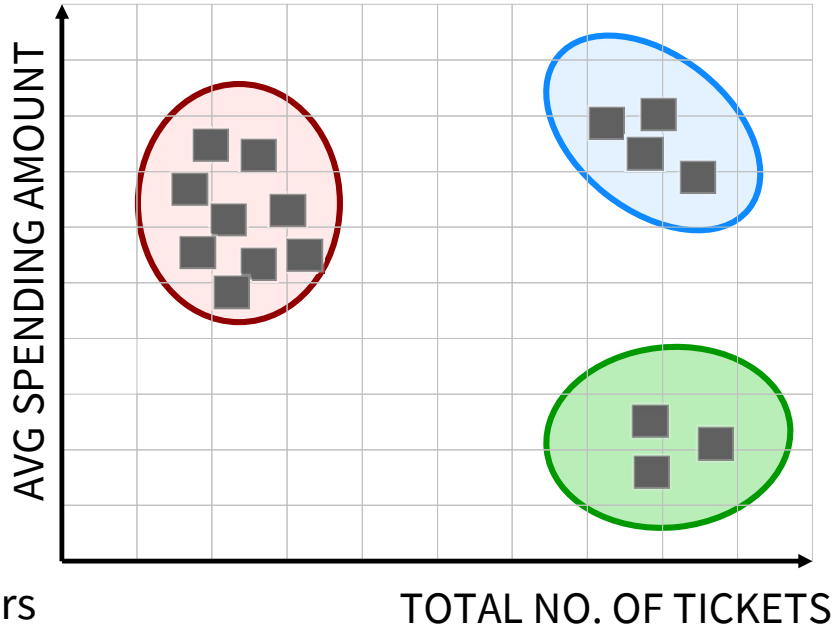
If a person watched FAST X
How likely will s/he also watch MEG 2?

	AVATAR WOW	FAST X	BARBIE	JOHN WICK CH4	MEG 2	OPPENHEIMER	…
Mary	X			X		X	…
Peter		X					…
Ying	X					X	…
Carlos				X	X		…
Emre			X				…
Güley		X		X			…
…							…

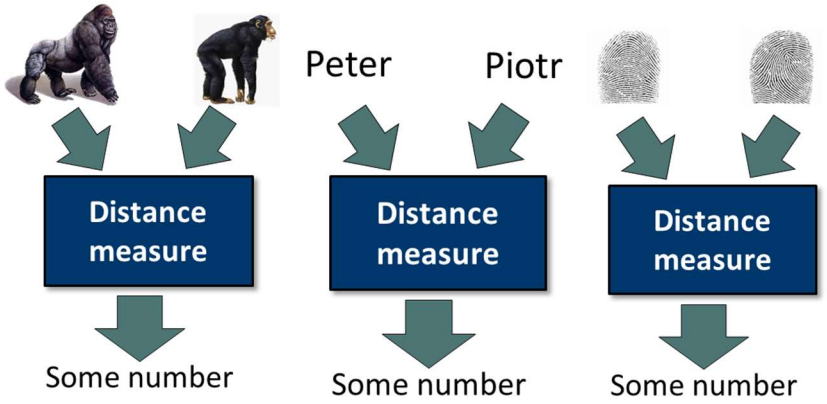
Cluster Analysis Methods

Goal formalization, similarity measures, the k-means algorithm

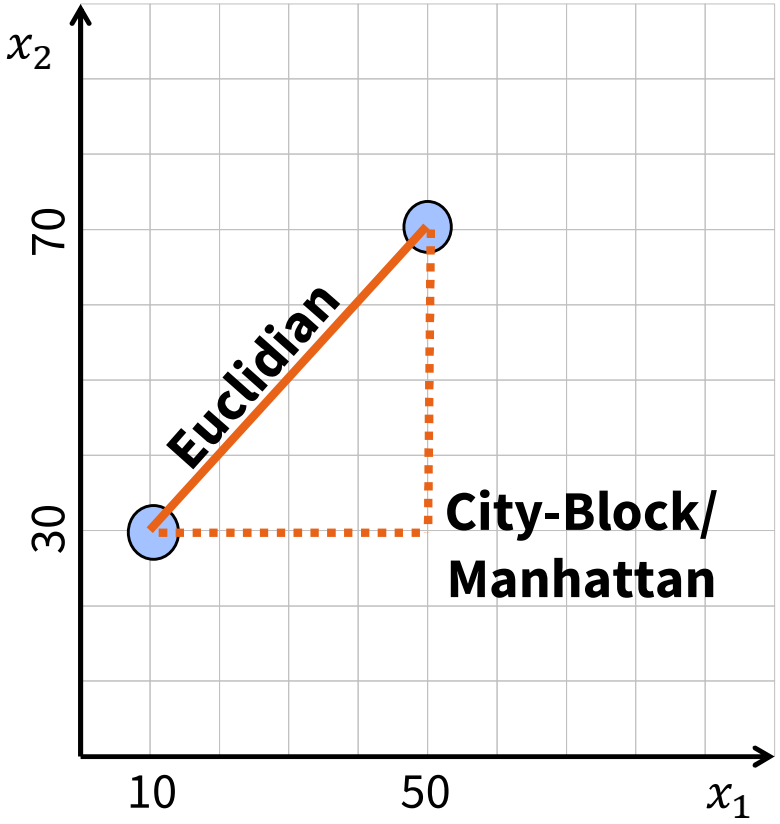
Distance and Distance Measurement



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

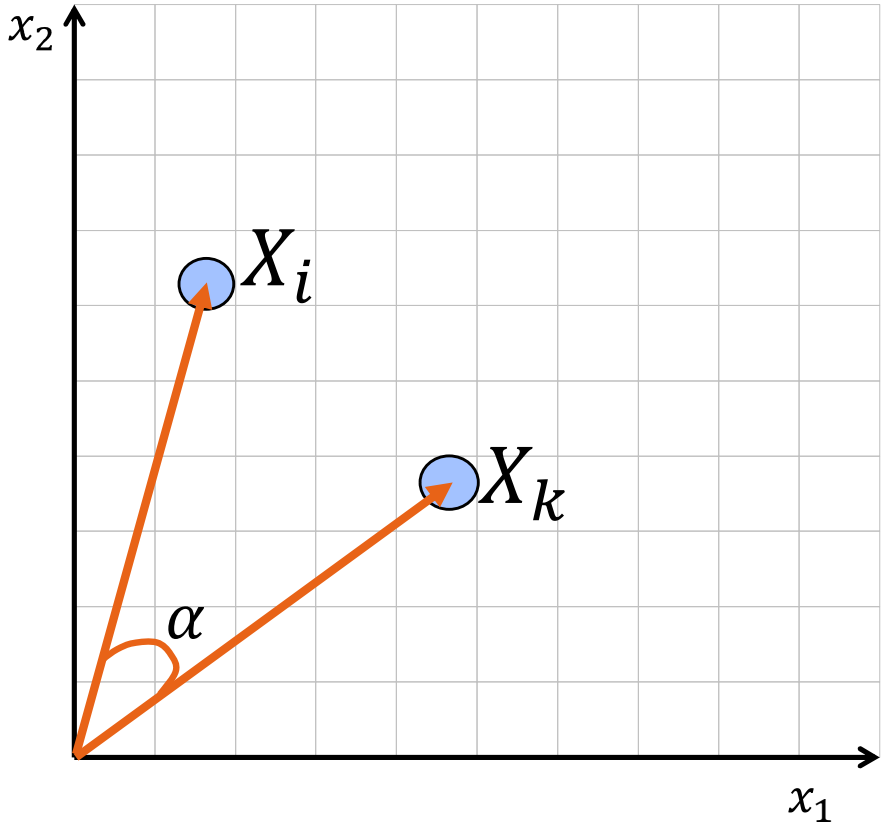
Abstraction & generalization

$$X = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$$

$$\text{Lp} - \text{Metric: } d(X_i, X_k) = \left(\sum_{j=1}^{m} |x_{ij} - x_{kj}|^p \right)^{1/p}$$

Distance Measures for Numeric Data (cont.)

Cosine similarity



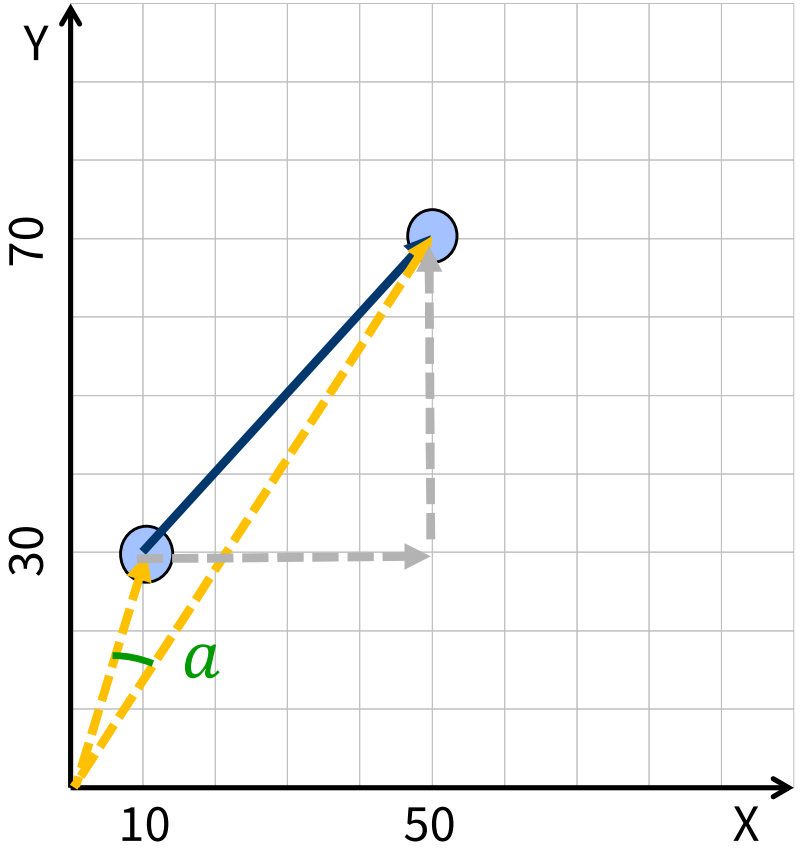
Angle between vectors:
Direction of the vector captures the relationship between variable values.

$$X_i = (x_{i1}, x_{i2})$$
$$X_k = (x_{k1}, x_{k2})$$

$$\cos\alpha = \frac{X_i \cdot X_k}{\|X_i\| \cdot \|X_k\|}$$

Distance Measures for Numeric Data (cont.)

The choice of the distance measure impacts the cluster solution



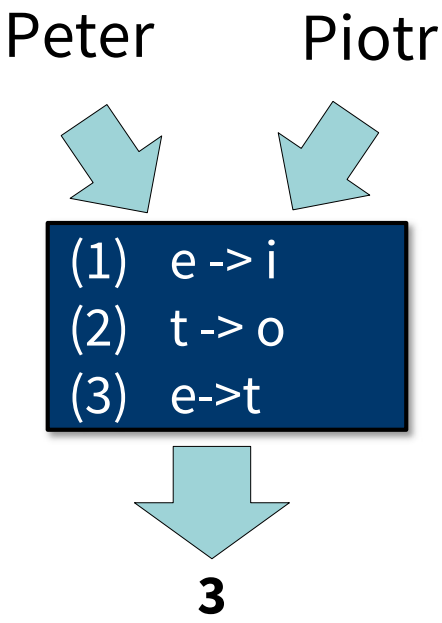
What's a good measure for a given application?

Euclidian distance

Cosine distance

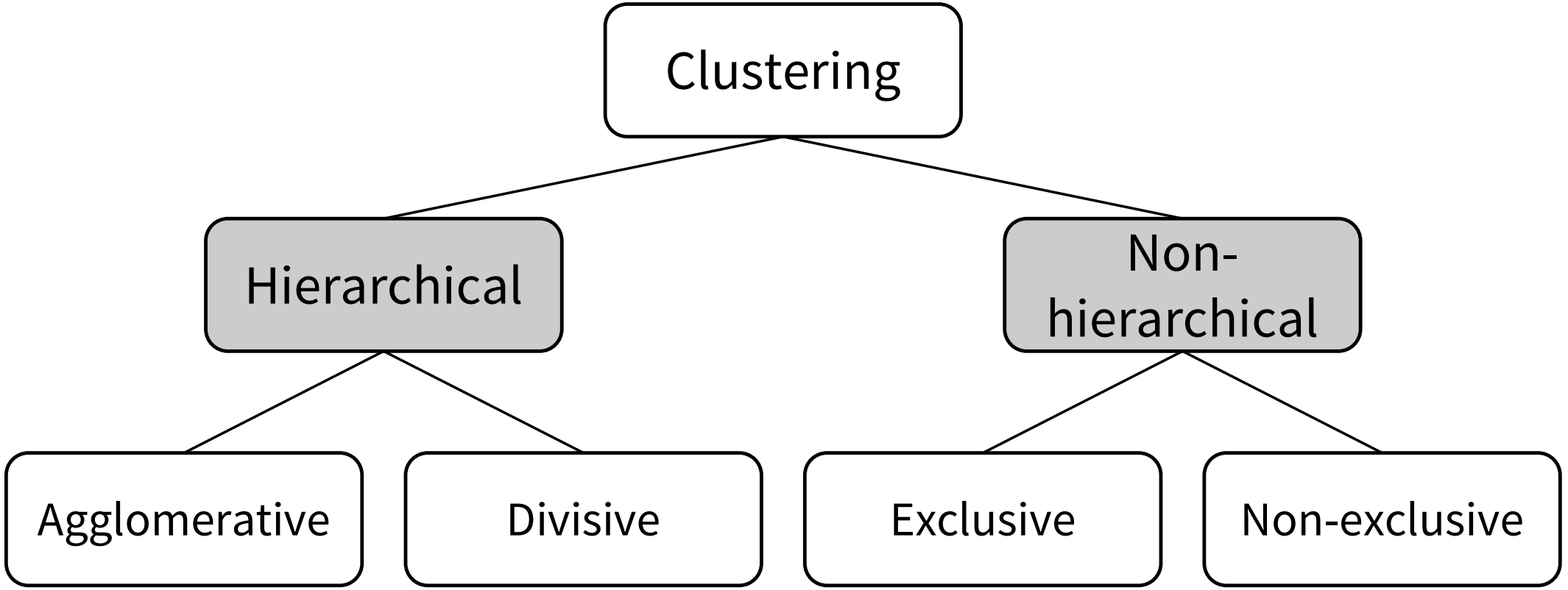
Distance Measures for Non-Numeric Data

- **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



Approaches Toward Cluster Analysis

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



The k-Means Algorithm
Widely used non-hierarchical clustering method

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

$$C^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|X_i - \bar{X}_k\|^2$$

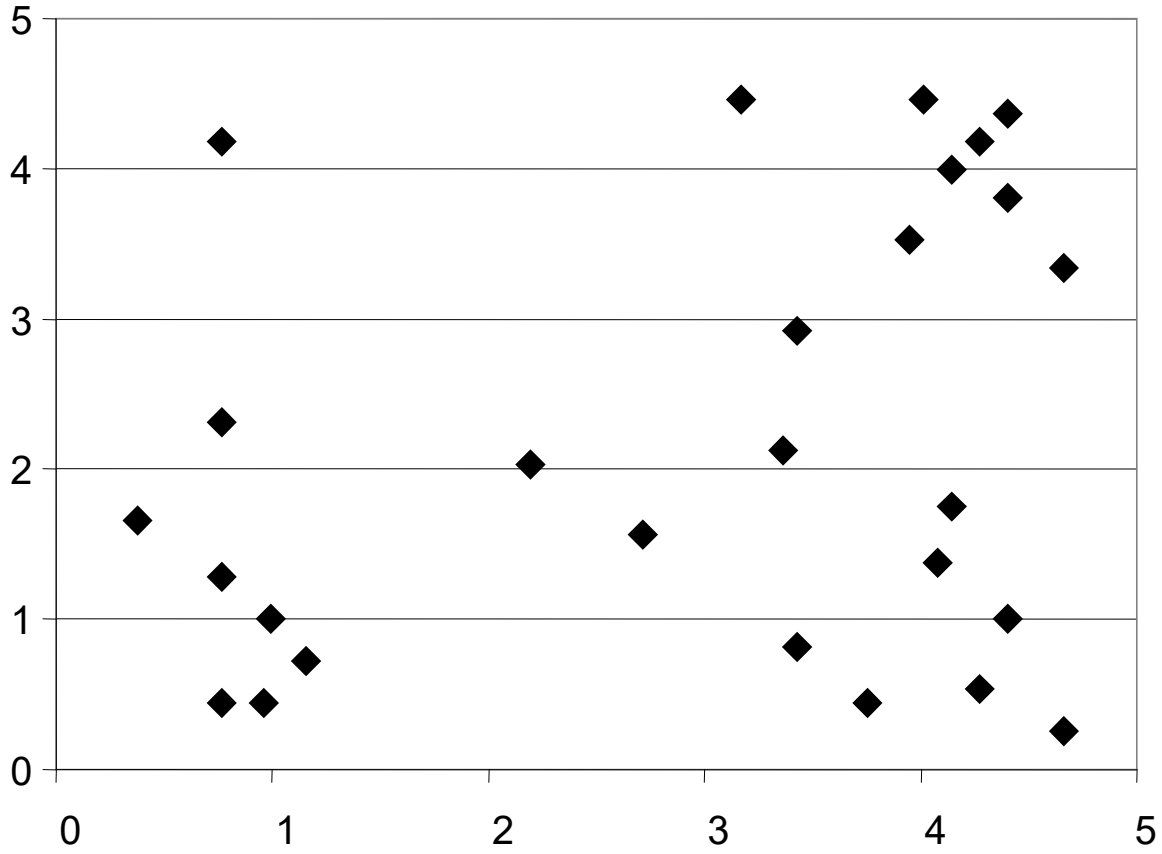
Centroid of cluster k

"Optimal" cluster assignment

Number of cases in cluster k

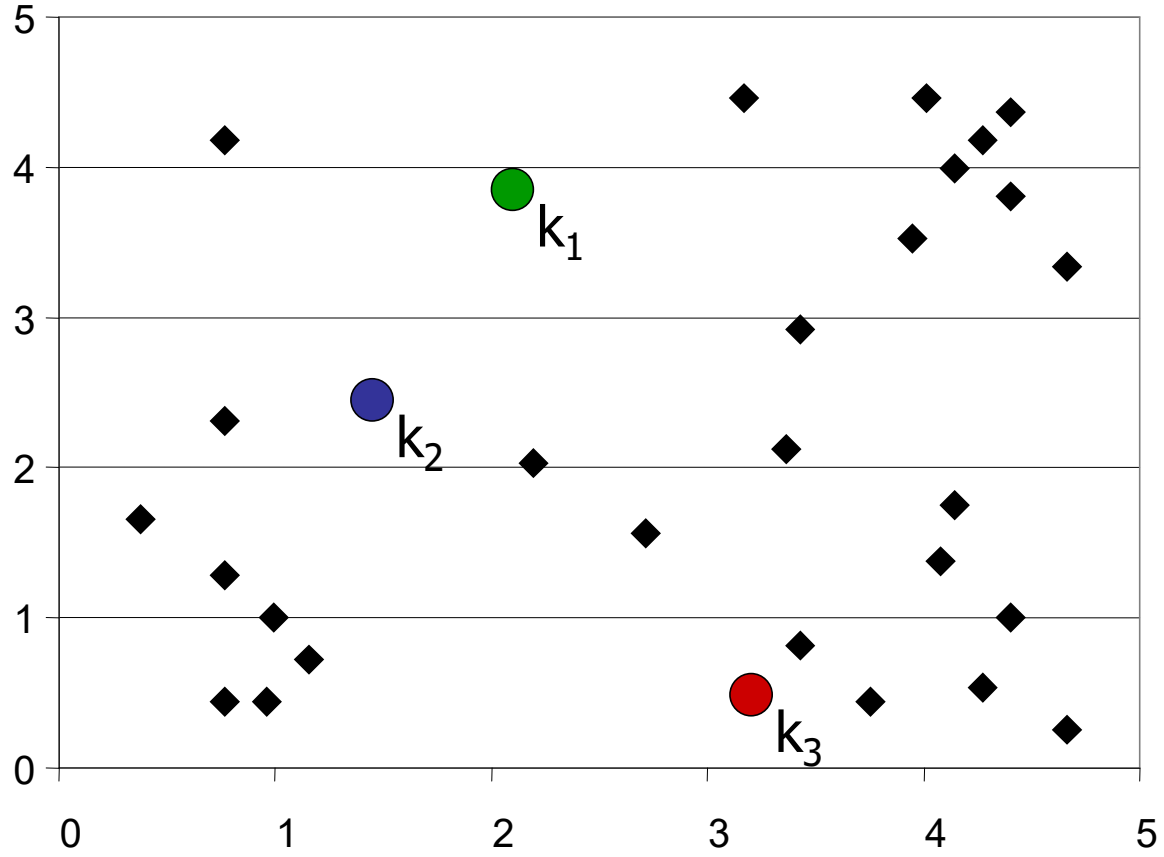
The k-Means Algorithm

A two-dimensional example with unlabeled data



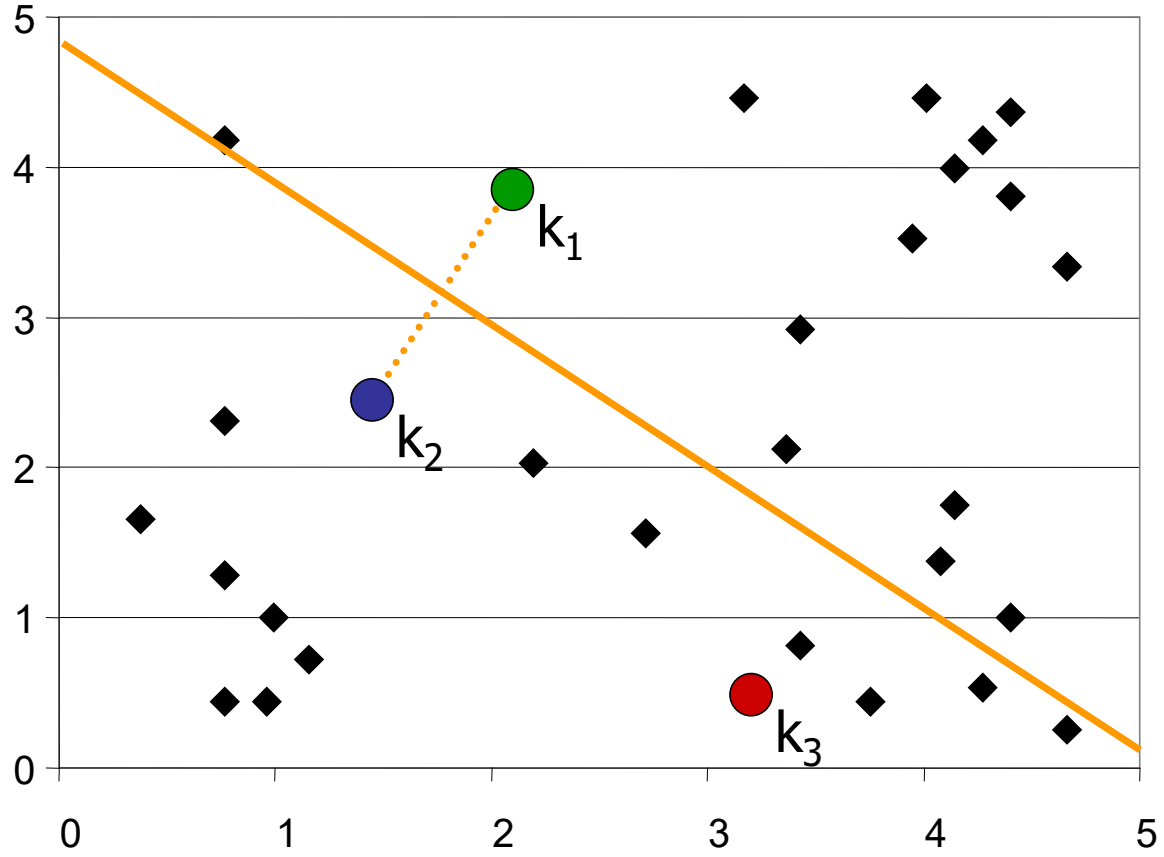
The k-Means Algorithm

Say we set K=3 and randomly sample three centroids



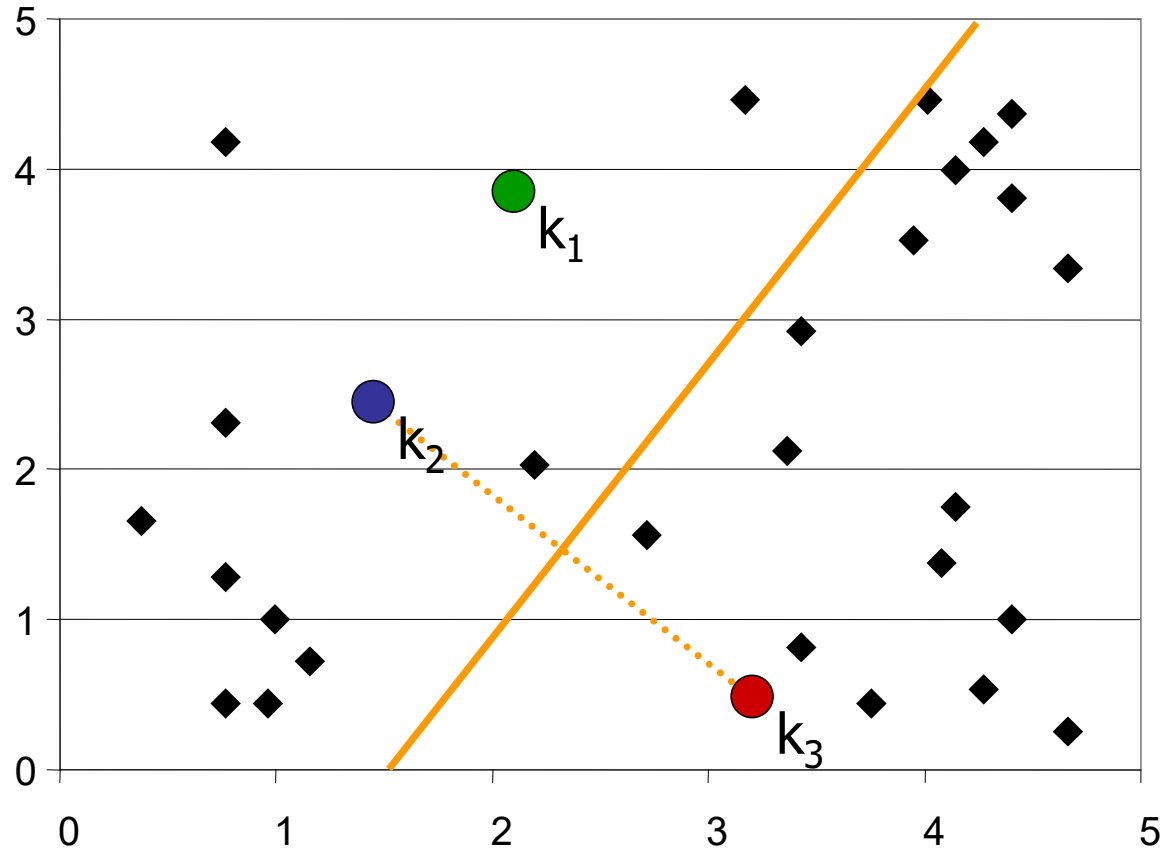
The k-Means Algorithm

The centroids partition the space into 3 clusters



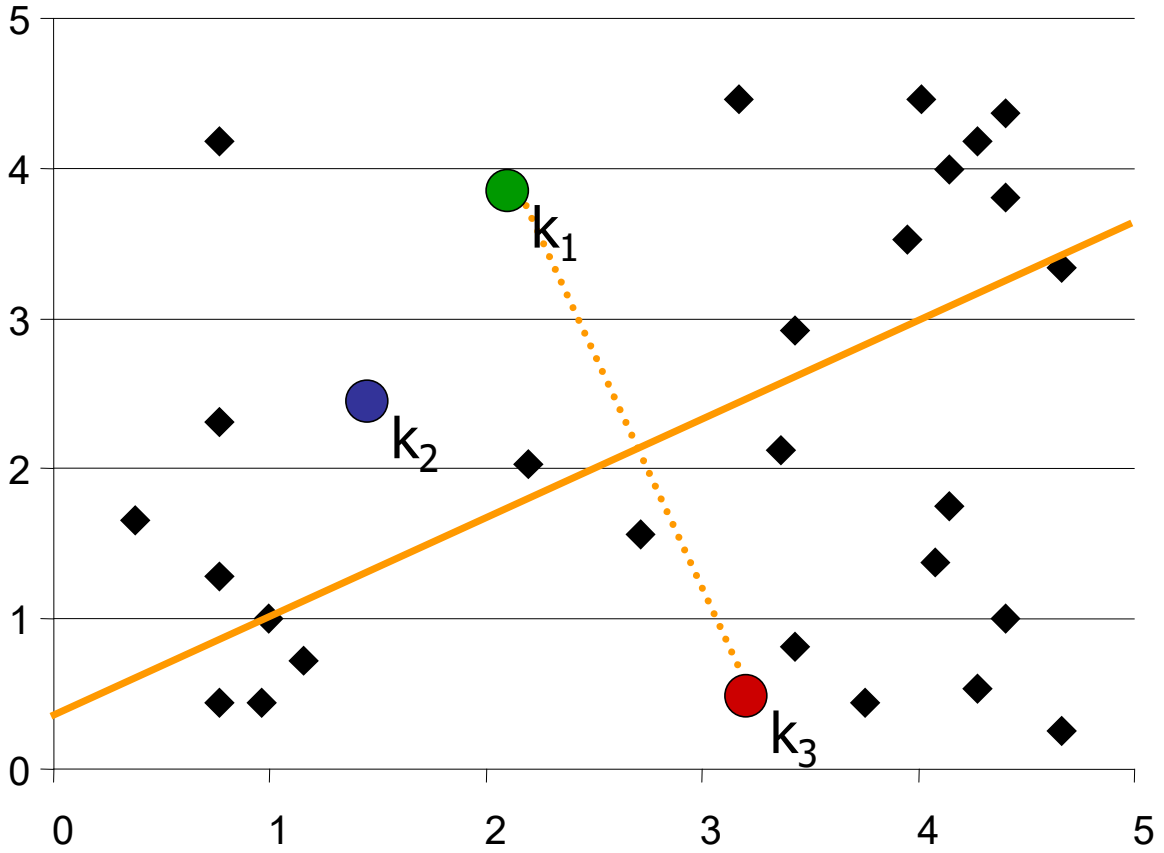
The k-Means Algorithm

The centroids partition the space into 3 clusters



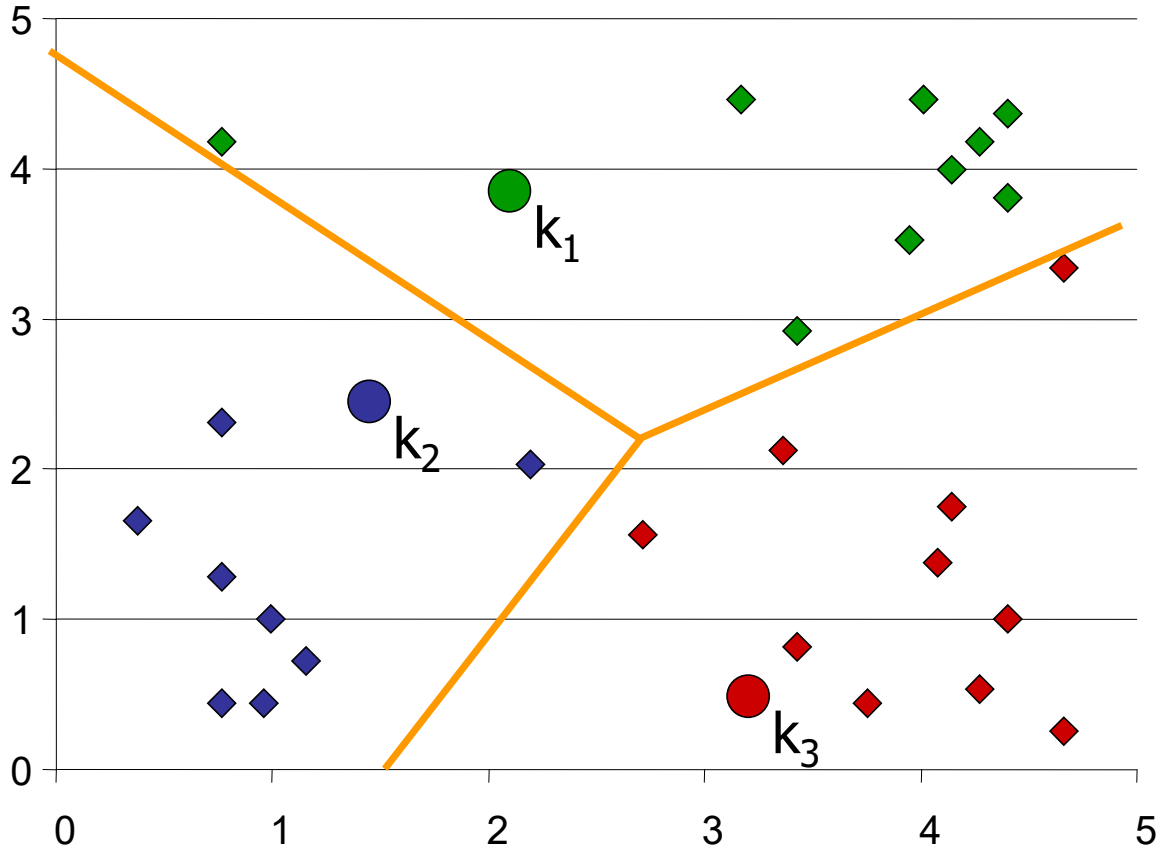
The k-Means Algorithm

The centroids partition the space into 3 clusters



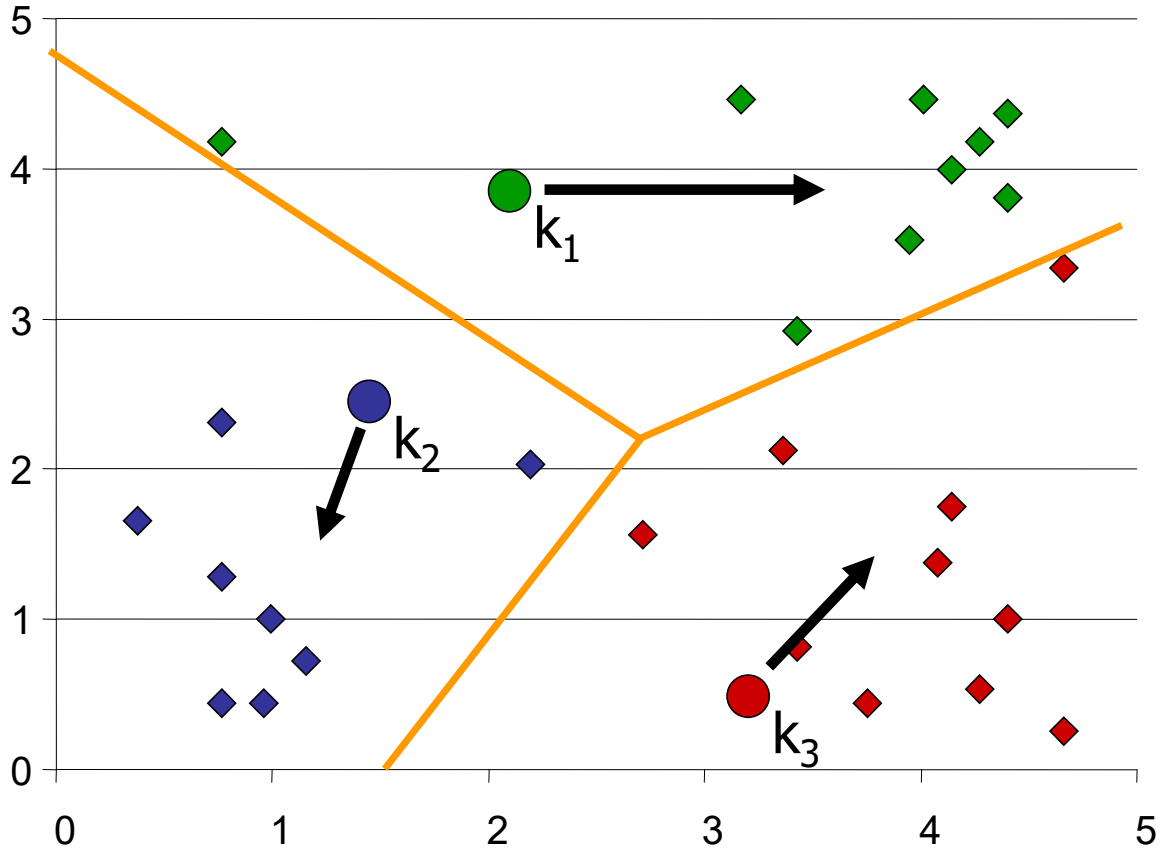
The k-Means Algorithm

We assign every data point to the cluster defined by the nearest centroid



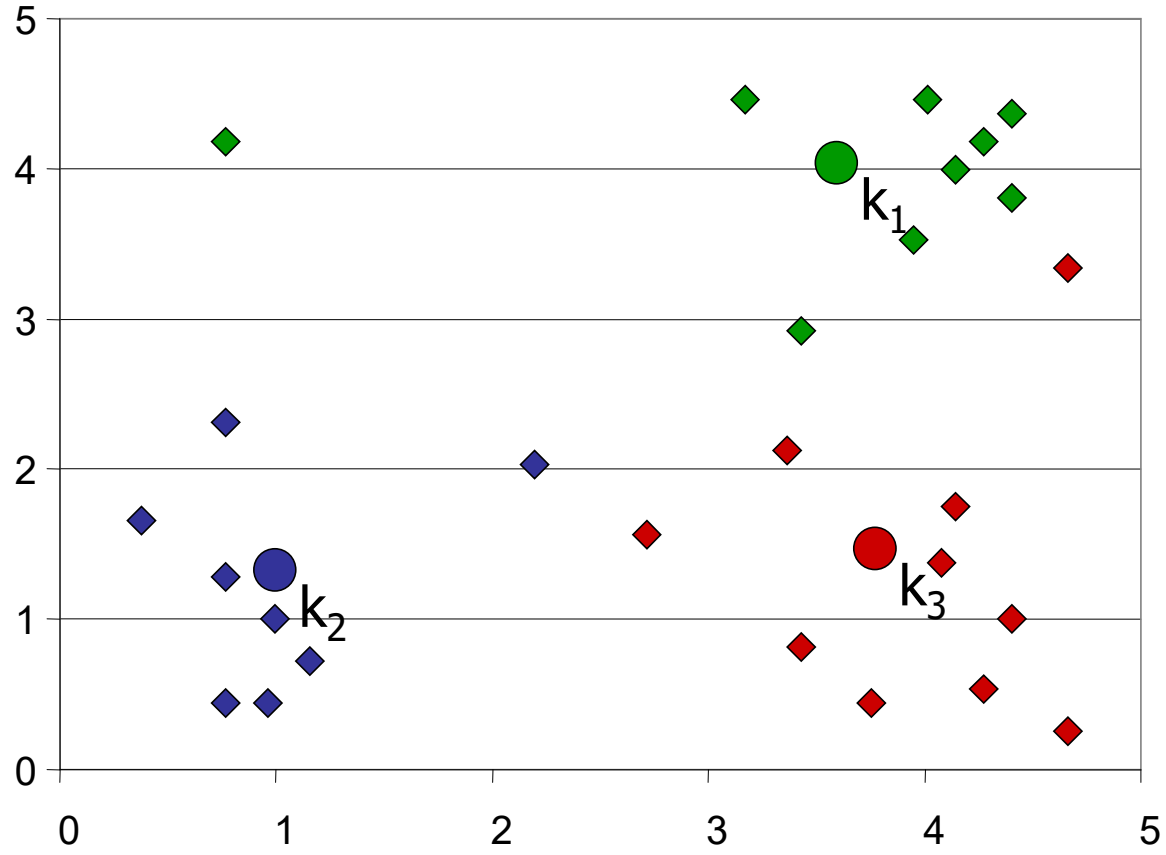
The k-Means Algorithm

Next, we update the centroids by calculating the midpoint of each cluster



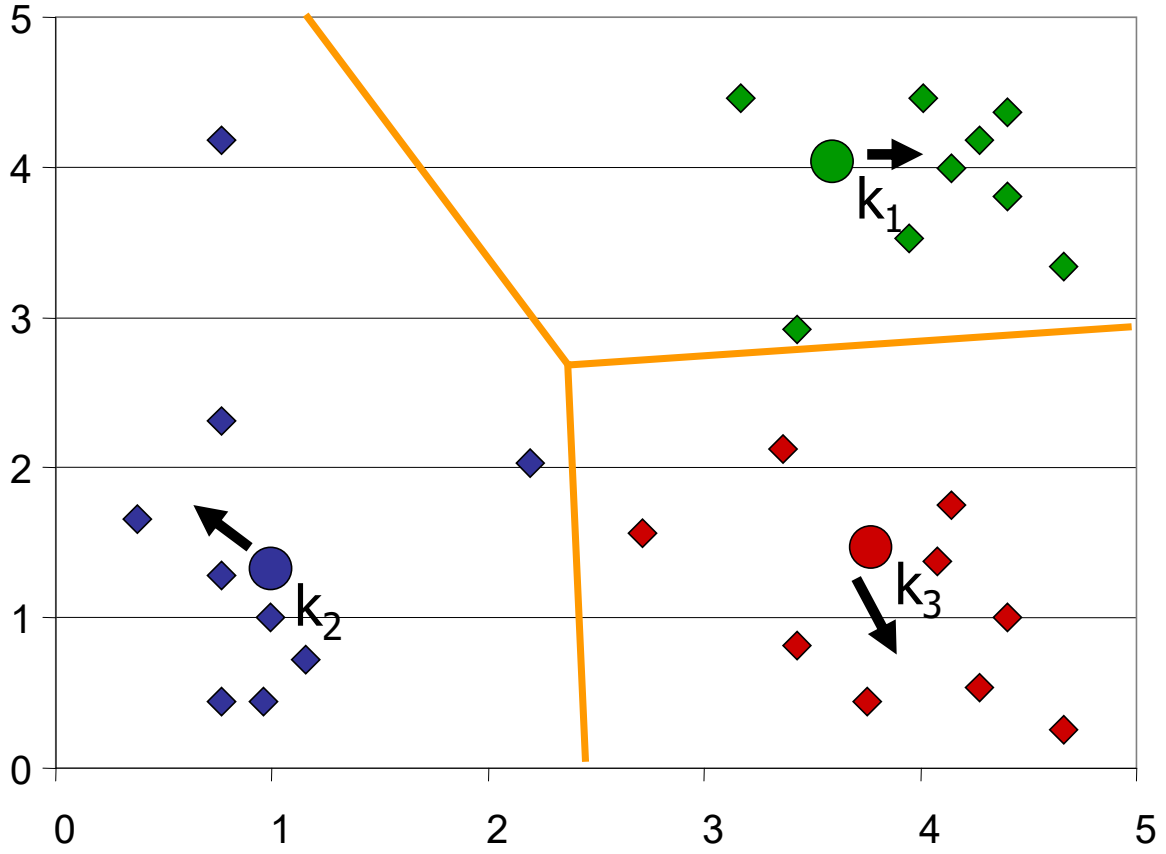
The k-Means Algorithm

Having obtained the new/updated centroids we repeat the whole process



The k-Means Algorithm

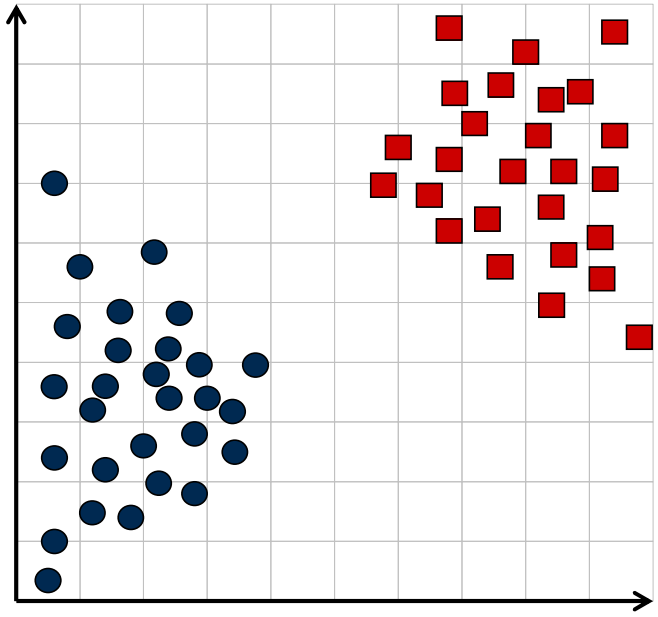
And continue until the cluster solution stops to changing



The k-Means Algorithm

Determining the number of clusters

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

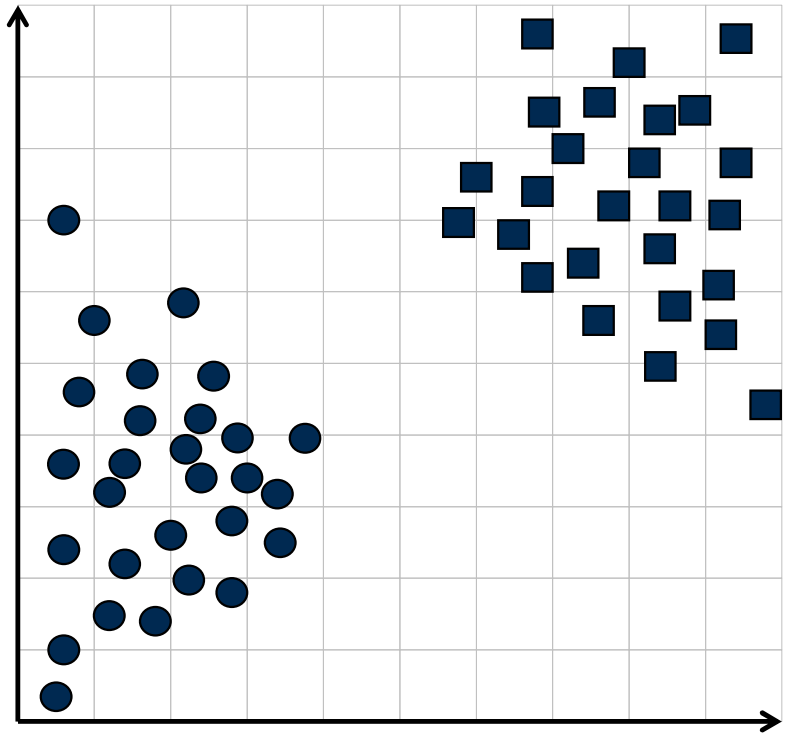


Assume we do not know that our data sets exhibits two clusters. We can experiment with different values of K and see what happens.

The k-Means Algorithm

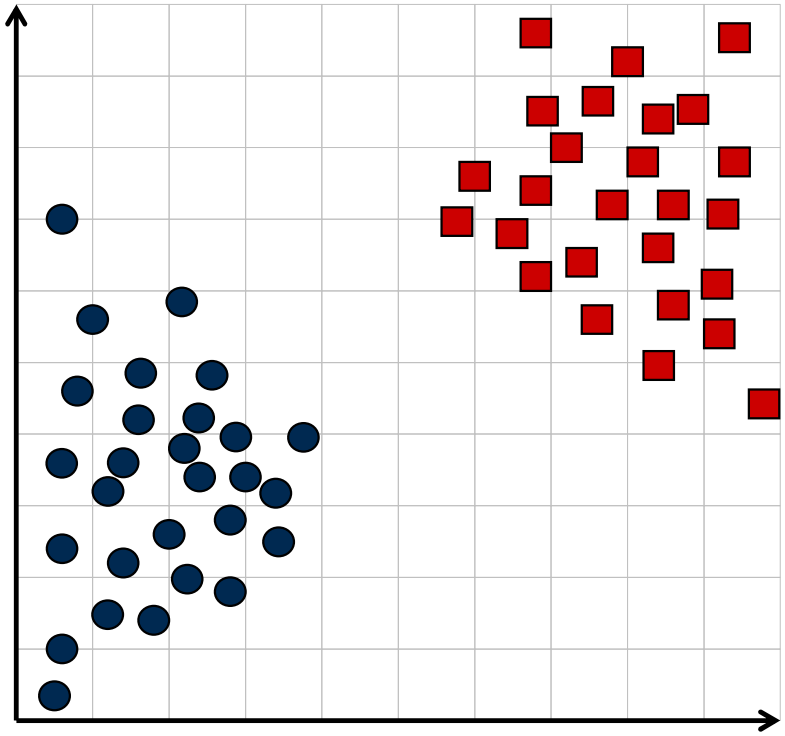
Solution with K=1

- **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

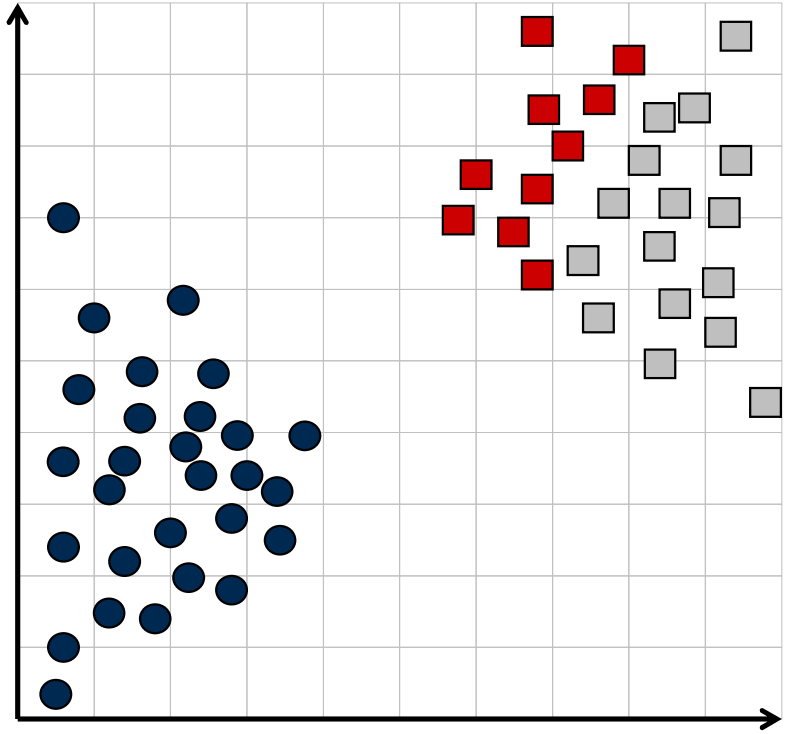
- **K-Means objective**
  - Minimize intra-cluster variance
  - Sum of squared differences between members and the cluster center (i.e., centroid)
- **Assume the objective value for K=2 is 123**



The k-Means Algorithm

Solution with K=3

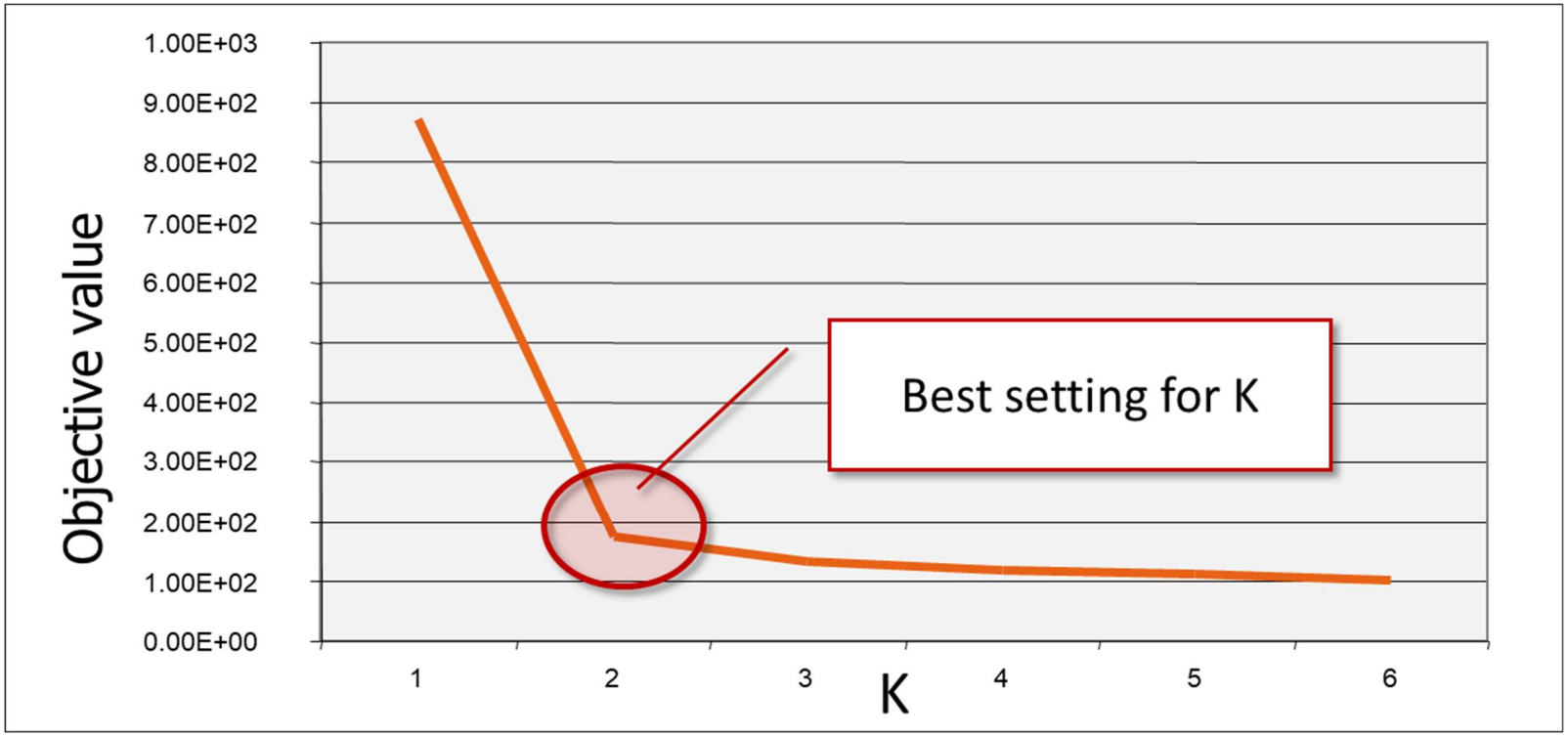
- **K-Means objective**
  - Minimize intra-cluster variance
  - Sum of squared differences between members and the cluster center (i.e., centroid)
- **Assume the objective value for K=3 is 115**



The k-Means Algorithm

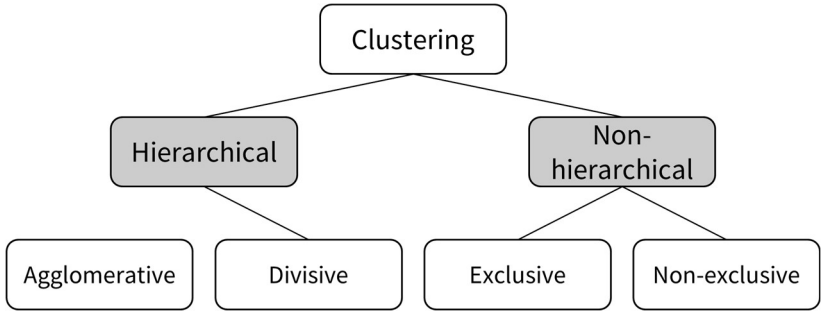
Graphical heuristic to decide on K

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

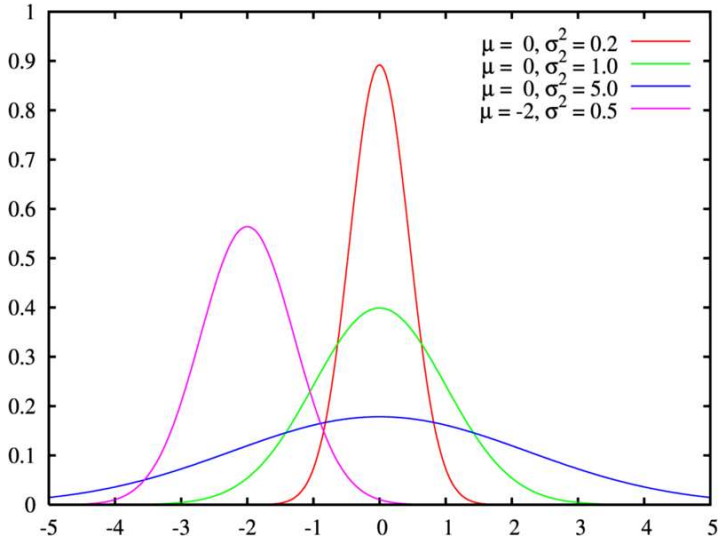


Extensions of K-means

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

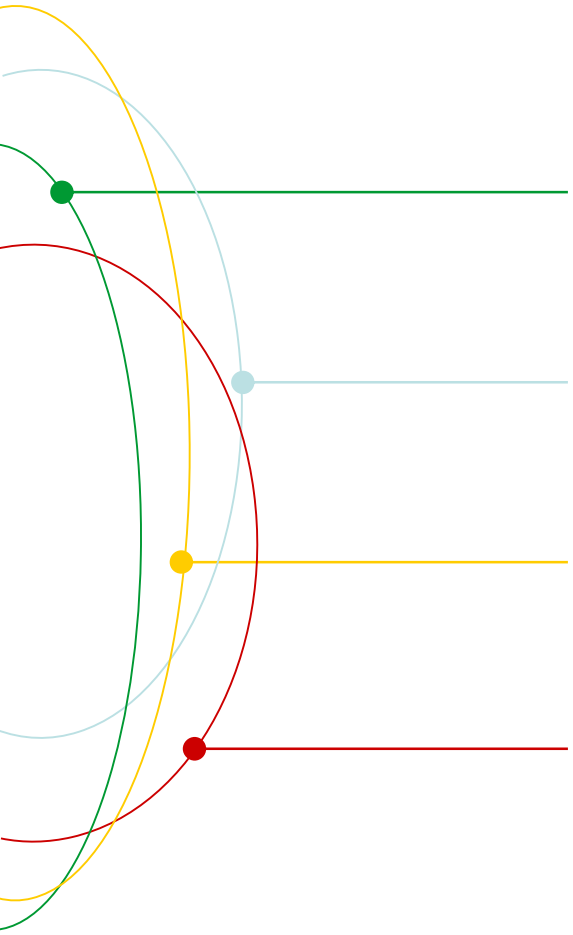
Silhouette score

- **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
- **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

Photo: Heike Zappe