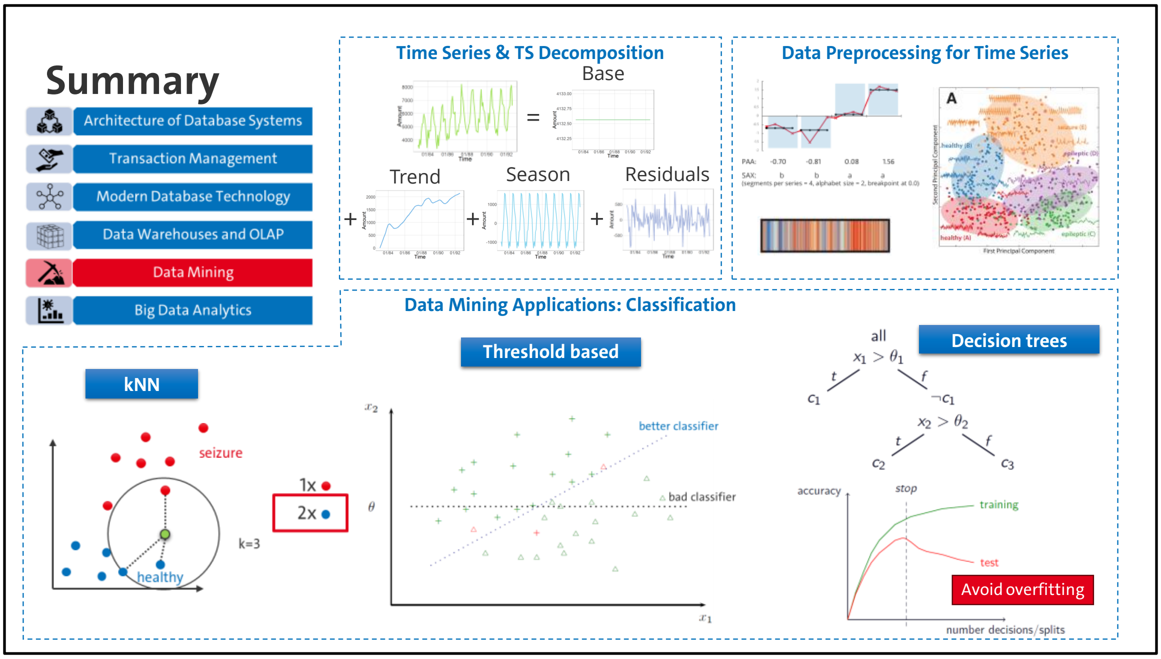
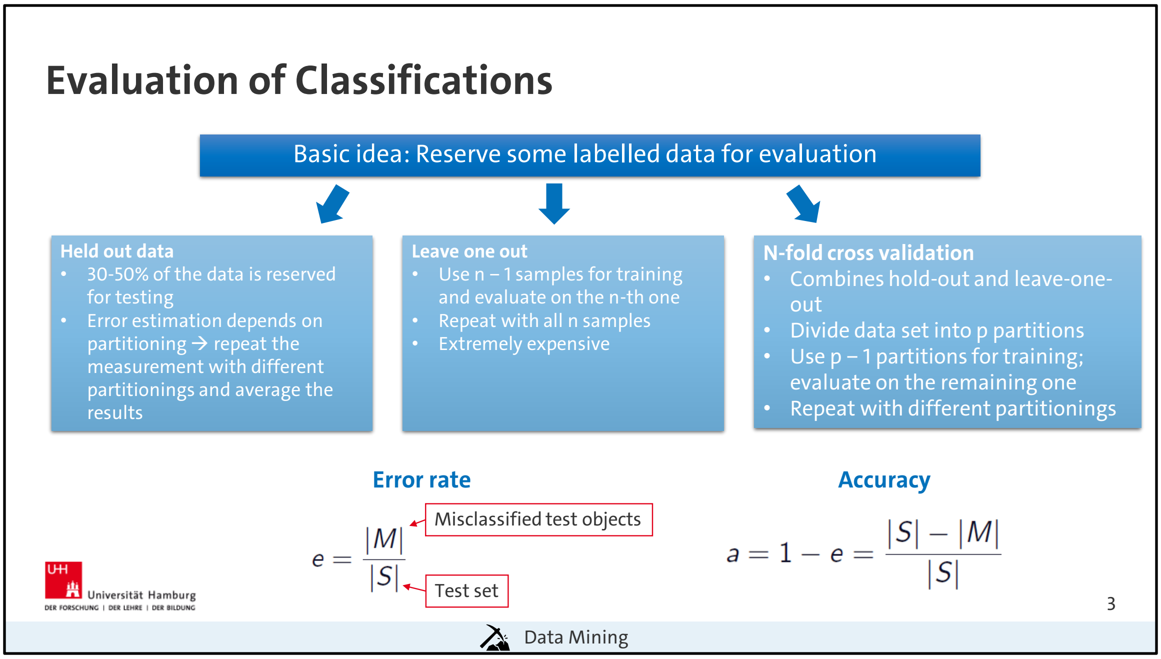

Summary
Architecture of Database Systems
Transaction Management
Modern Database Technology
Data Warehouses and OLAP
Data Mining
Big Data Analytics
Time Series & TS Decomposition
Base
Trend
Season
Residuals
Data Preprocessing for Time Series
Data Mining Applications: Classification
kNN
seizure
healthy
k=3
1x
2x
Threshold based
better classifier
bad classifier
Decision trees
all
accuracy
stop
training
test
Avoid overfitting
number decisions/splits



Contrastive analysis: comparison with a baseline case

- Absolute improvement/degradation:

$\Delta_{abs}a = a_n - a_{n-1}$

$\Delta_{abs}e = e_n - e_{n-1}$

- Relative improvement/degradation:

$\Delta_{rel}a = (a_n - a_{n-1}) / a_{n-1}$

$\Delta_{rel}\, e = (e_n - e_{n-1})/e_{n-1}$

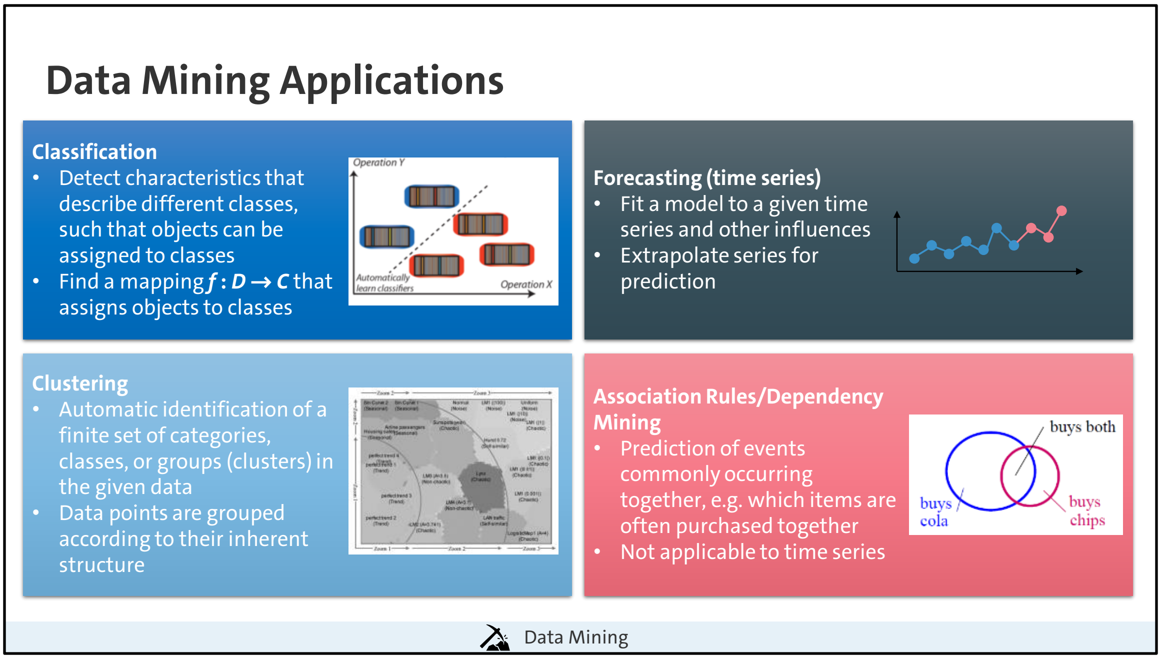
- Essential difference between absolute and relative improvement/degradation
- Often used manipulation:
  - Relative improvement for positive effects
  - Absolute improvement for negative (adverse) effects
- There are more error measures, e.g. weighted error measures

Further important quality aspects

- Compactness, e.g. size of the decision tree (size determines classification time)
- Interpretability (how many insights can be imparted by the classifier to the user?)
- Efficiency of the construction process and application of the constructed

classifier

- Scalability with respect to large data sets (construction as well as application)
- Robustness against noise and missing values



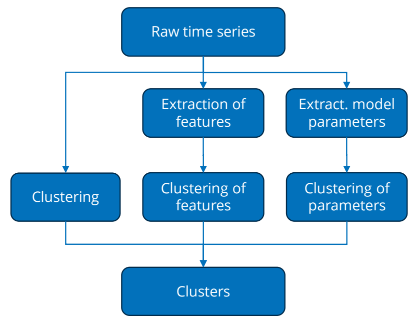
- Definition of the optimal clustering depends on
  - Application domain (i.e. structure/semantics of the data, meaning of distance)
  - Data mining target (i.e. object correlations that should be detected)
- Computing the optimal clustering is computationally infeasible → greedy, sub-optimal approaches
- Different clustering algorithms might lead to different clustering results

Clustering of Time Series

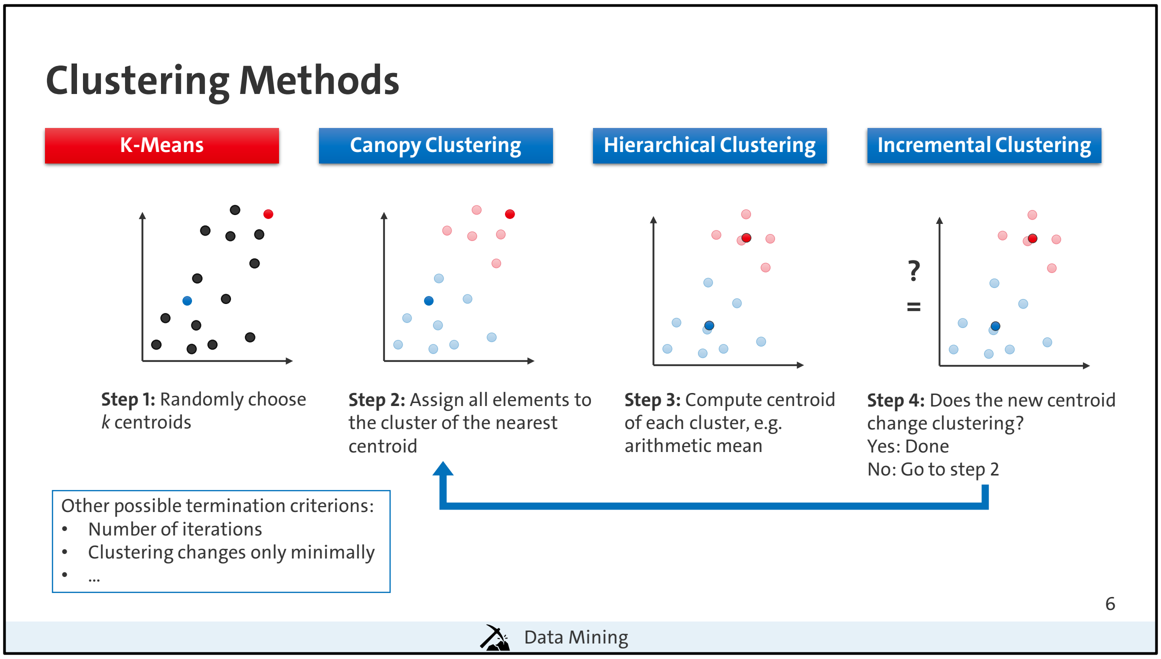
Goal
Partition a set of time series into groups in such a way that homogenous time series are grouped together

Challenges
- Often too large to fit in main memory
- Highly dimensional and it is difficult for clustering algorithms to handle them
- Homogeneity strongly depends on the selected similarity measure

Applications
- Which patterns appear frequently?
- Which patterns appear surprisingly?



- Raw-value-based approach
    - Raw time series is matched as well as possible
    - Usage of stretching and contraction
- Feature-based approach
    - Reduce time series to feature vector
    - Similarity of vector, usually by Euclidean distance
- Model-based approach
    - Fit a model to each given time series
    - Time series distances are calculated based on model parameters



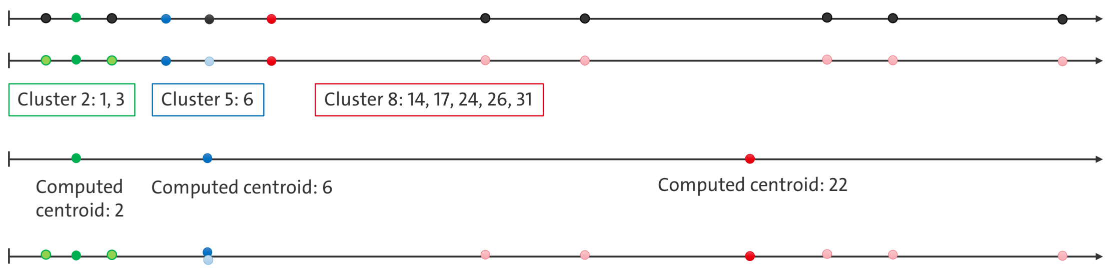
- Also works with 1 dimension or more than 2 dimensions (if an according distance measure exists)

Problems:

- Converges to local minima, i.e. the clustering is not necessarily optimal
- Workaround: Start the algorithm several times
- Relatively high effort to compute distances and cluster centroids

Example K-Means

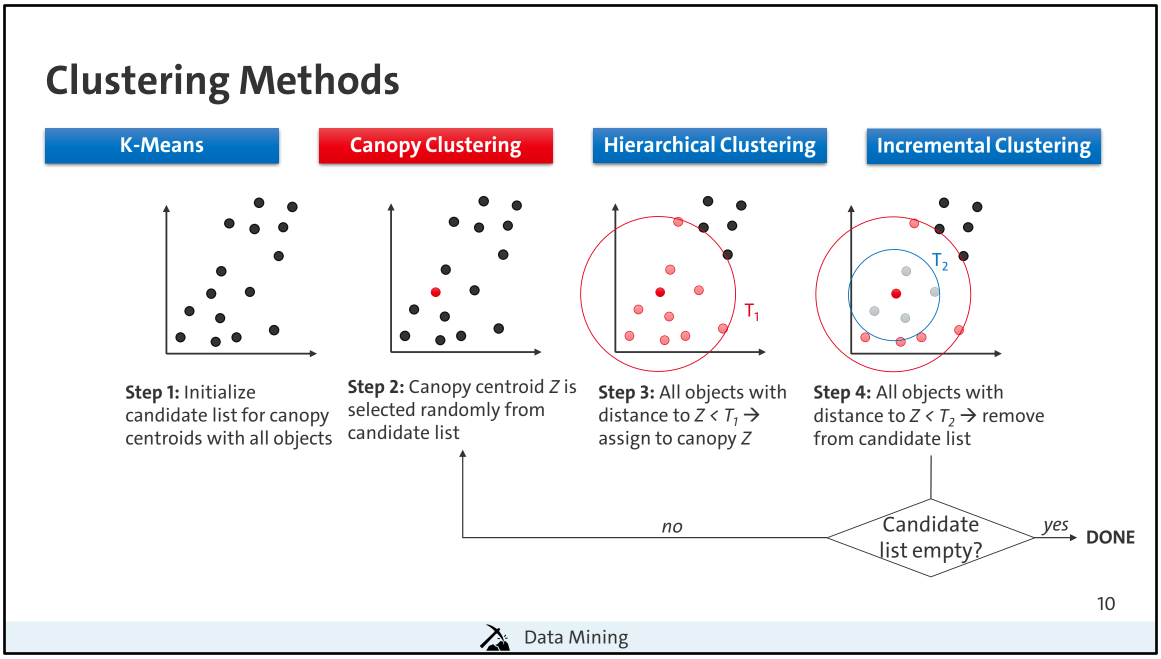
1, 3, 6, 14, 17, 24, 26, 31
K=3
Randomly chosen centroids: 2, 5, 8



Clustering did not change → Done

Exercise K-Means

- Clustering of the numbers 1, 3, 6, 14, 17, 24, 26, 31
- k=3
- Centroids randomly chosen in step 1: 10, 21, 29

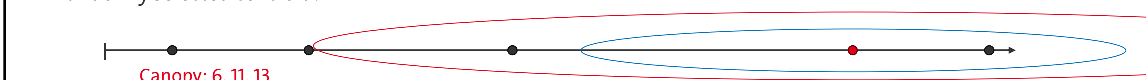


Construction of overlapping clusters (canopies)

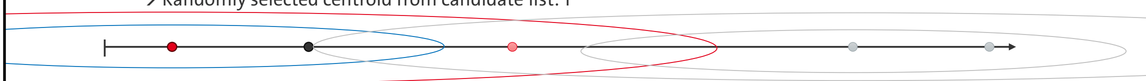
- Non-Partitioning Clustering
- Often used as a first step in a multi-step analysis approach
- Scalable for large data sets
- Can be used for string data
- Requires distance function and 2 thresholds T_1 and T_2 ($T_1 > T_2$)

Canopy Clustering Example

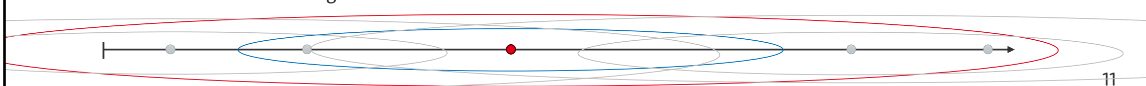
Clustering of the numbers 1, 3, 6, 11, 13
$T_1 = 8$, $T_2 = 4$
Randomly selected centroid: 11



Canopy: 6, 11, 13
Remove 11 and 13 from the candidate list → candidate list: 1, 3, 6
→ Randomly selected centroid from candidate list: 1



Canopy: 1, 3, 6
Remove 1 and 3 from the candidate list → candidate list: 6
→ Select last remaining centroid from candidate list: 6

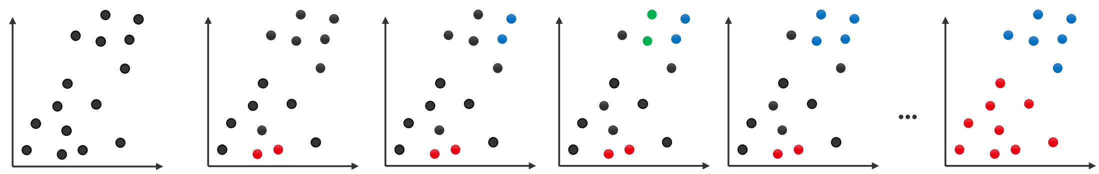


11

Canopy: 1, 3, 6, 11, 13 → No candidates left → Done

Clustering Methods

K-Means | Canopy Clustering | **Hierarchical Clustering** | Incremental Clustering



Step 1: Each data point is an individual cluster

Step 2: Merge least distant clusters into a new cluster

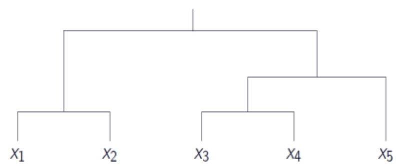
Step 3: Repeat

Step 4: Stop if all clusters have been merged or the minimal cluster distance exceeds a predefined threshold

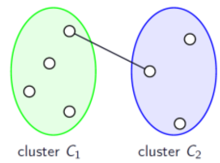
Clustering Methods

K-Means | Canopy Clustering | **Hierarchical Clustering** | Incremental Clustering

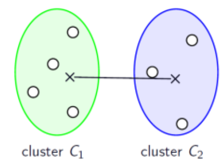
Results can be displayed as a dendrogram



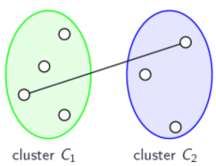
Distance measures for clusters



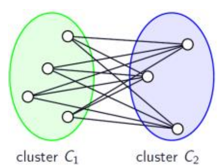
Single link
Minimal distance between two data points



Complete link
Maximal distance between two data points

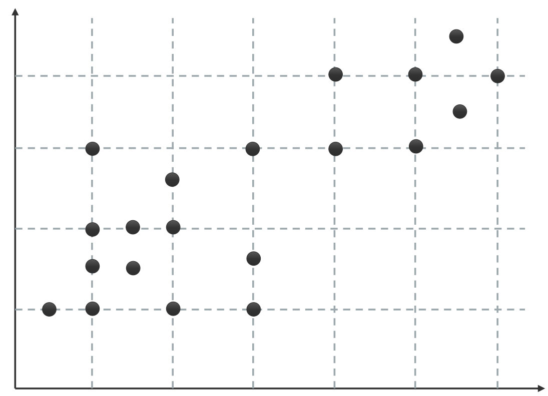


Average link
Average distance between two data points

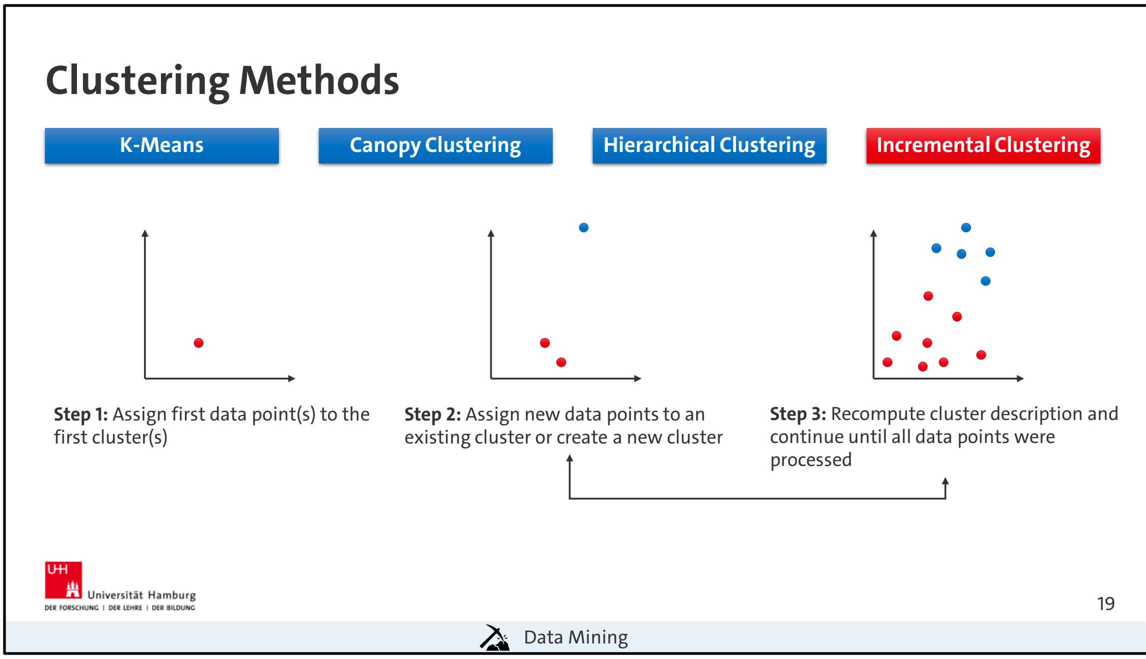


Canonical entity
Distance between two cluster representatives
(e.g. the centroids)

Exercise: Hierarchical Clustering



- Distance: Complete link
- Stop when minimal cluster distance > 4

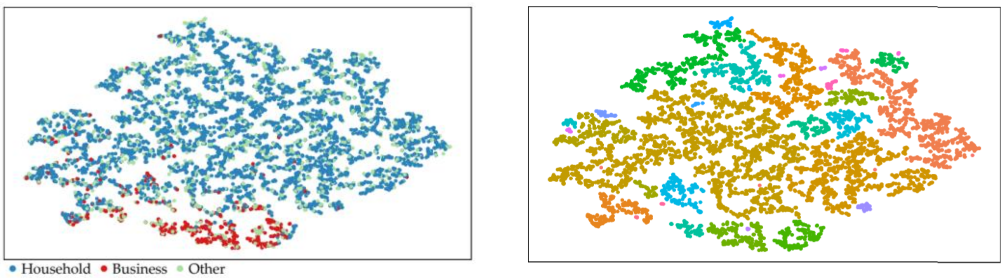


- Is a concept to enable the use of clustering algorithms, e.g. k-means, with large data sets, i.e. data sets which do not fit into main memory
- Problem: Result depends on the order in which data points are processed

Cluster description consists at least of

- Centroid
- Number of data points in the cluster
- “Radius” of the cluster (e.g. the maximal distance of a point to the centroid)

Clustering Time Series



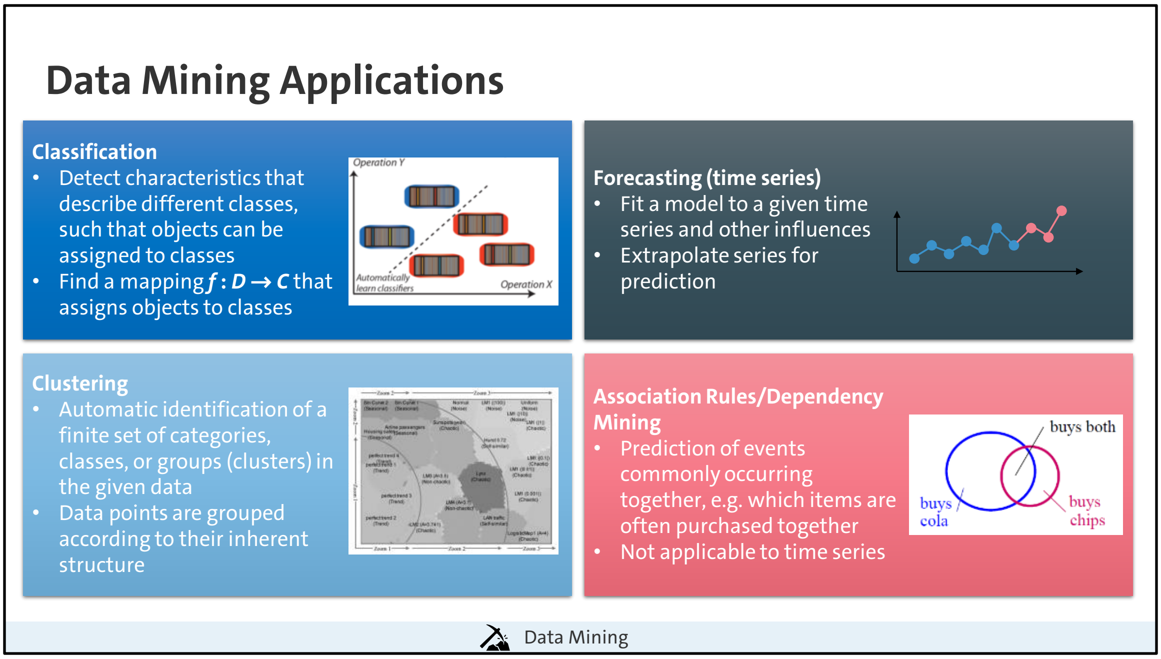
(a) Class-based

(b) Feature-based [1]

Metering Dataset

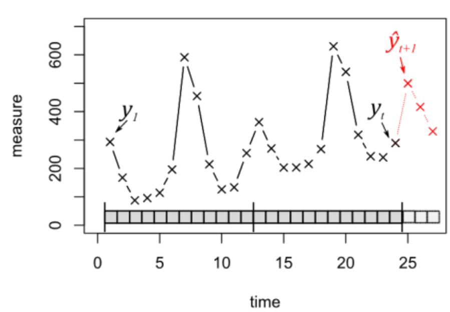
[1] Source: Feature-based Time Series Analytics, Lars Kegel, 2020

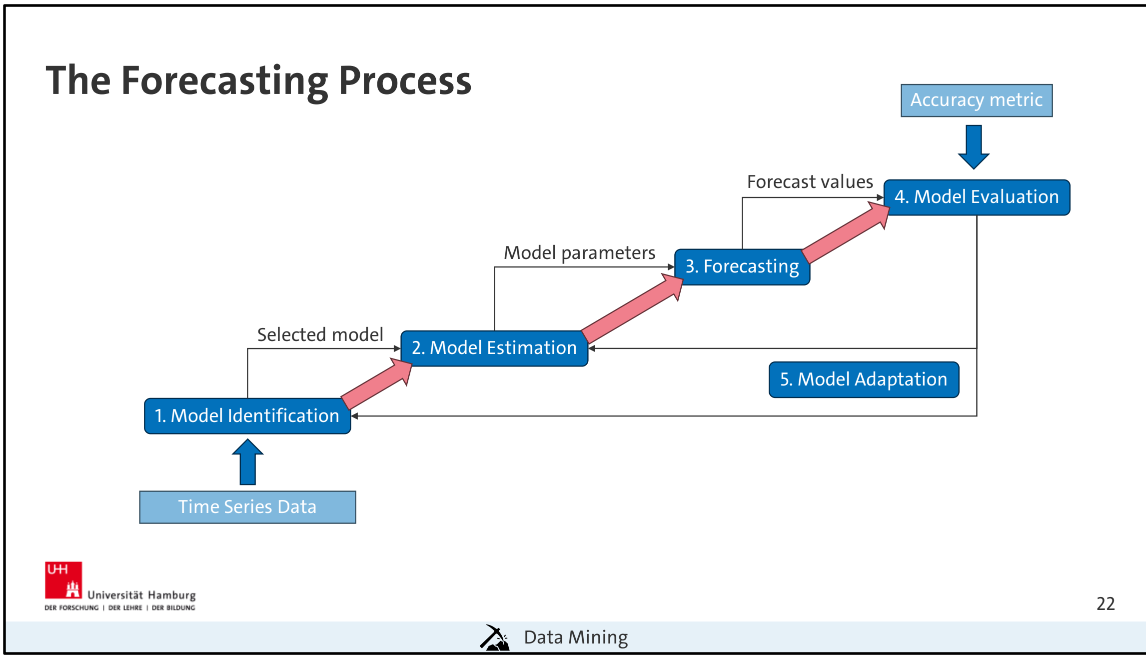
Class-based: Dataset provides class labels



Forecast Models

- One model represents one time series
- Trend (long-term changes)
- Season (regular reoccurring changes)
- Accuracy is the only requirement





1. **Model Identification** – Choose the optimal model type
2. **Model Estimation** – Instantiation of the model by training its model parameters
3. **Forecasting** – Usage of the model to calculate the next time series values
4. **Model Evaluation** – Compare the model's results with real time series values using an error measure
5. **Model Adaption** – Adaption of the model parameters or the model type

Model Types

Domain Specific Extensions
White-Box

HWT
(Energy Single Equation)
EGRV
(Energy, Multi Equation)
IFS
(Weather)

Base Forecast Models

(Auto)Regression | **Exponential Smoothing**
Gray-Box

AR
MA
ARMA
ARIMA
SARIMA
ARMAX
MLR (Multiple Linear Regression)
VAR (Vector Autoregression)
MARS (Multivariate Adaptive Regression Splines)
GBM (Gradient Boosting Machine)

SES
(Single Exponential Smoothing)
DES
(Double Exponential Smoothing)
TES/HoltWinters
(Triple Exponential Smoothing)

Machine Learning
Black-Box

BN
(Bayesian Networks)
SVR
(Support Vector Regression)
ANN
(Artificial Neural Networks)

Exponential Smoothing

Weight time series values with smoothing factor

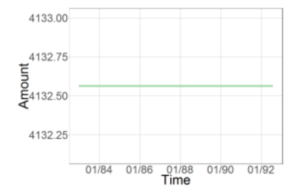
- Declines exponentially for older time series values

Model variants

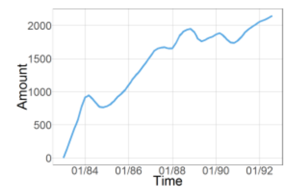
- Simple Exponential Smoothing (SESM): Base
- Double Exponential Smoothing (DESM): Base, Trend
- Triple Exponential Smoothing (TESM): Base, Trend, Season Holt-Winters-Model

Example Base:

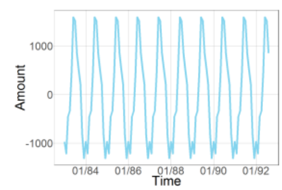
$$\begin{aligned} x_1^* &= x_1 \\ x_t^* &= \alpha \cdot x_t + (1-\alpha) \cdot x_{t-1}^* \\ \hat{x}_{t+1} &= x_t^* \end{aligned}$$



Base



Trend



Season

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Further Reasing

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1).

Exponential Smoothing

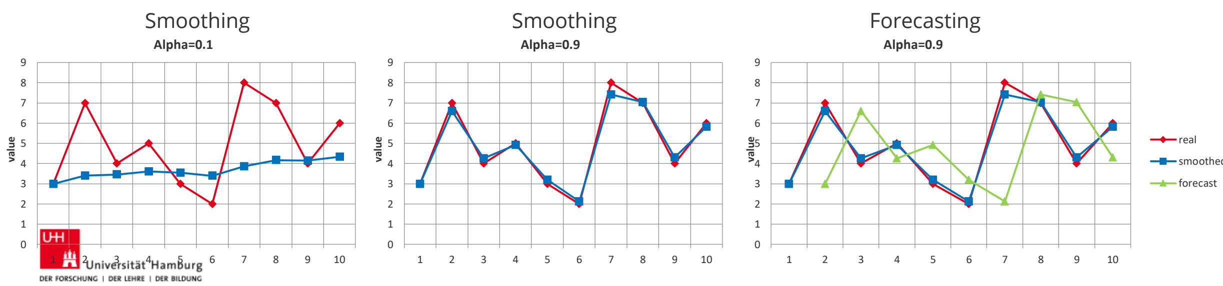
Example Single Exponential Smoothing

Base: $x_1^* \;= x_1$

$x_t^* \;= \alpha \cdot x_t + (1-\alpha) \cdot x_{t-1}^*$ Smoothing

$\hat{x}_{t+1} = \; x_t^*$ Forecasting

t	1	2	3	4	5	6	7	8	9	10
x_t	3	7	4	5	3	2	8	7	4	6
x_t^* (α=0.9)	3	6.6	4.26	4.926	3.19	2.12	7.412	7.041	4.304	5.83



- Often bad results for data with a trend → double/triple exponential smoothing
- Double exponential smoothing: Include the difference between data points and smooth it
- Triple exponential smoothing: Include and smooth seasonal correction factors, different smoothing factors for base, trend, and season possible