Master Seminar

Time Series Data Mining

ANGEWANDTE INFORMATIK IV
UNIVERSITÄT BAYREUTH

Andreas Braun 1200197
Master Applied Computer Science
INF302: Master-Seminar (5LP)

Motivation: Time Series + Data Mining

Sensors are part of the IoT

Important role of time dimension in data warehouse analyses

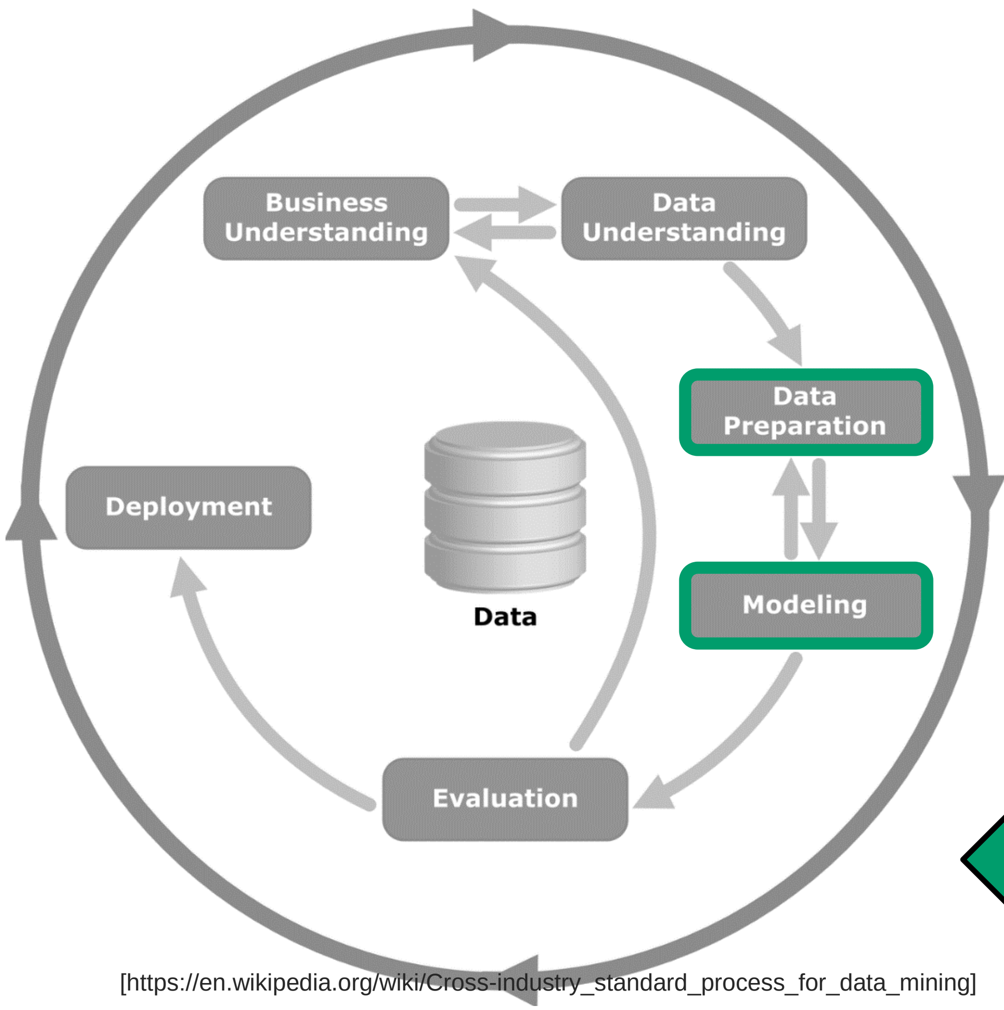
Data Mining Expertise		Application Domain Expertise
Data Mining (DM)		
Statistics	Machine Learning	Visual Analytics

Application domain examples:

- **Sensor data**
- **Medical devices**
- **Financial market data**

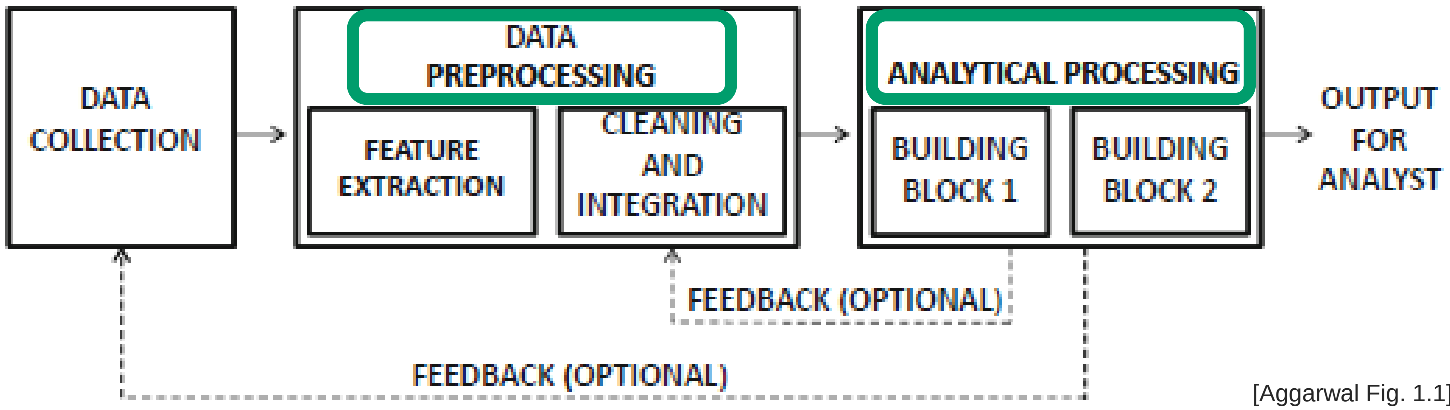
Motivation: Time Series + Data Mining

Cross-Industry Standard Process for data mining (CRISP-DM)



[https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining]

Data processing pipeline

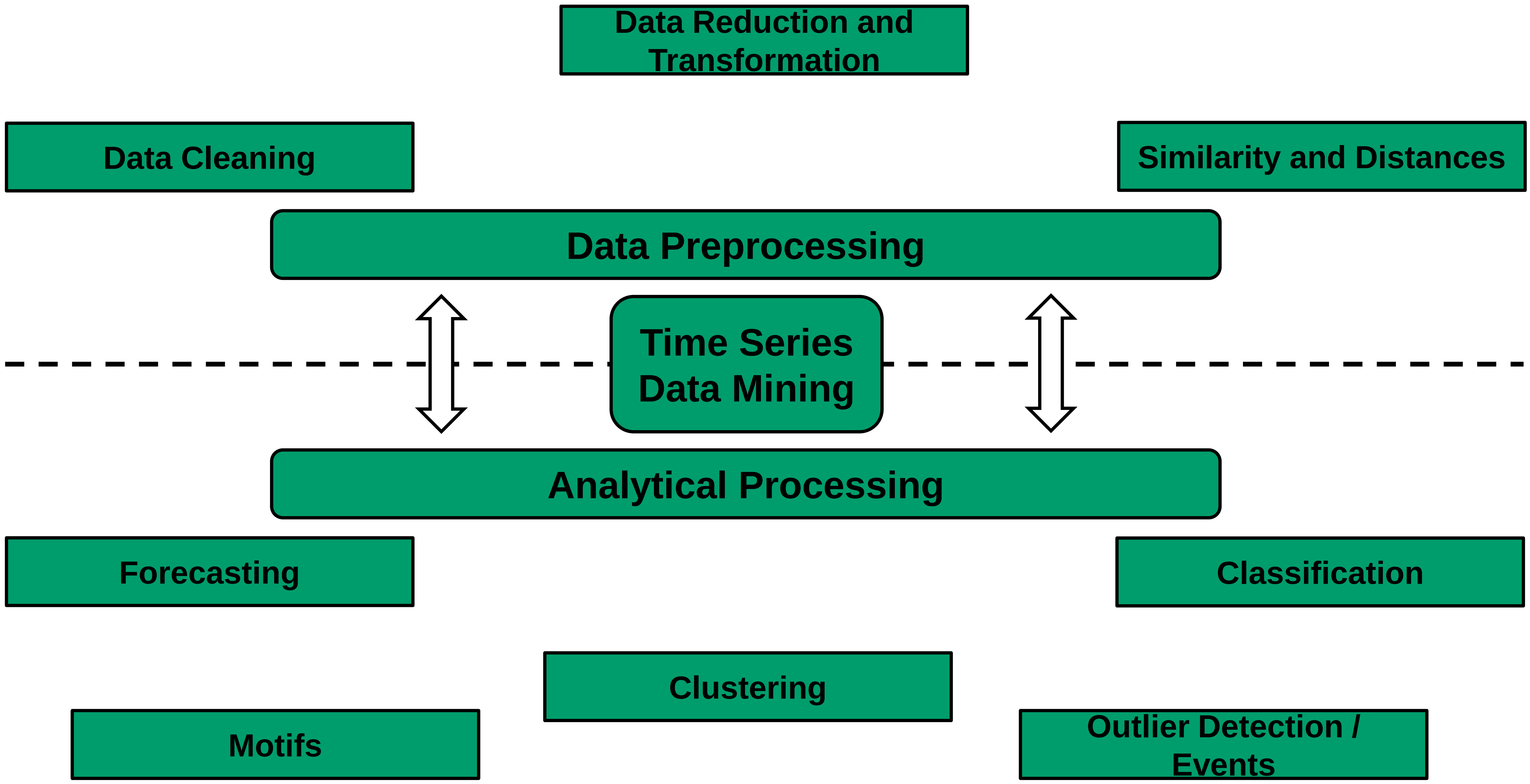


[Aggarwal Fig. 1.1]

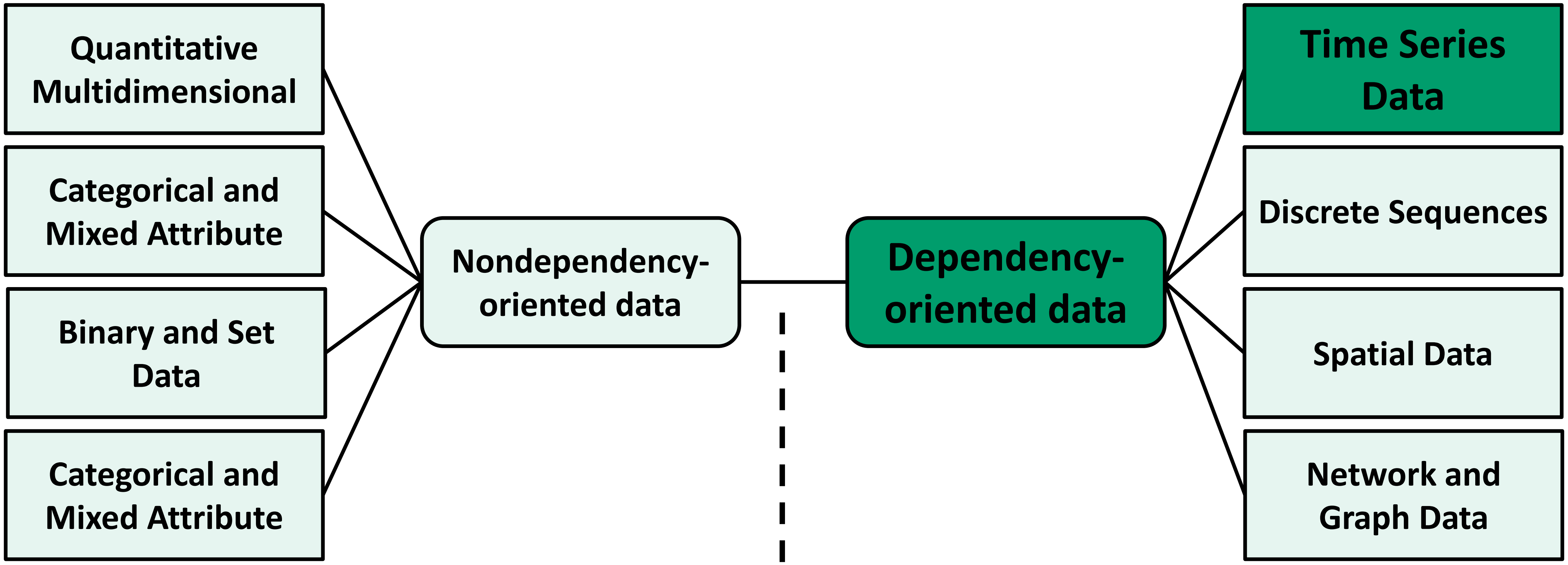
Data Preparation == Data Preprocessing

Modelling == Analytical Processing

Overview


Data Reduction and Transformation
Data Cleaning
Similarity and Distances
Data Preprocessing
Time Series Data Mining
Analytical Processing
Forecasting
Classification
Clustering
Motifs
Outlier Detection / Events

Data Preprocessing – Basic Data Types



- **Simplest and most common**
- **No specified dependencies between the data items or attributes**

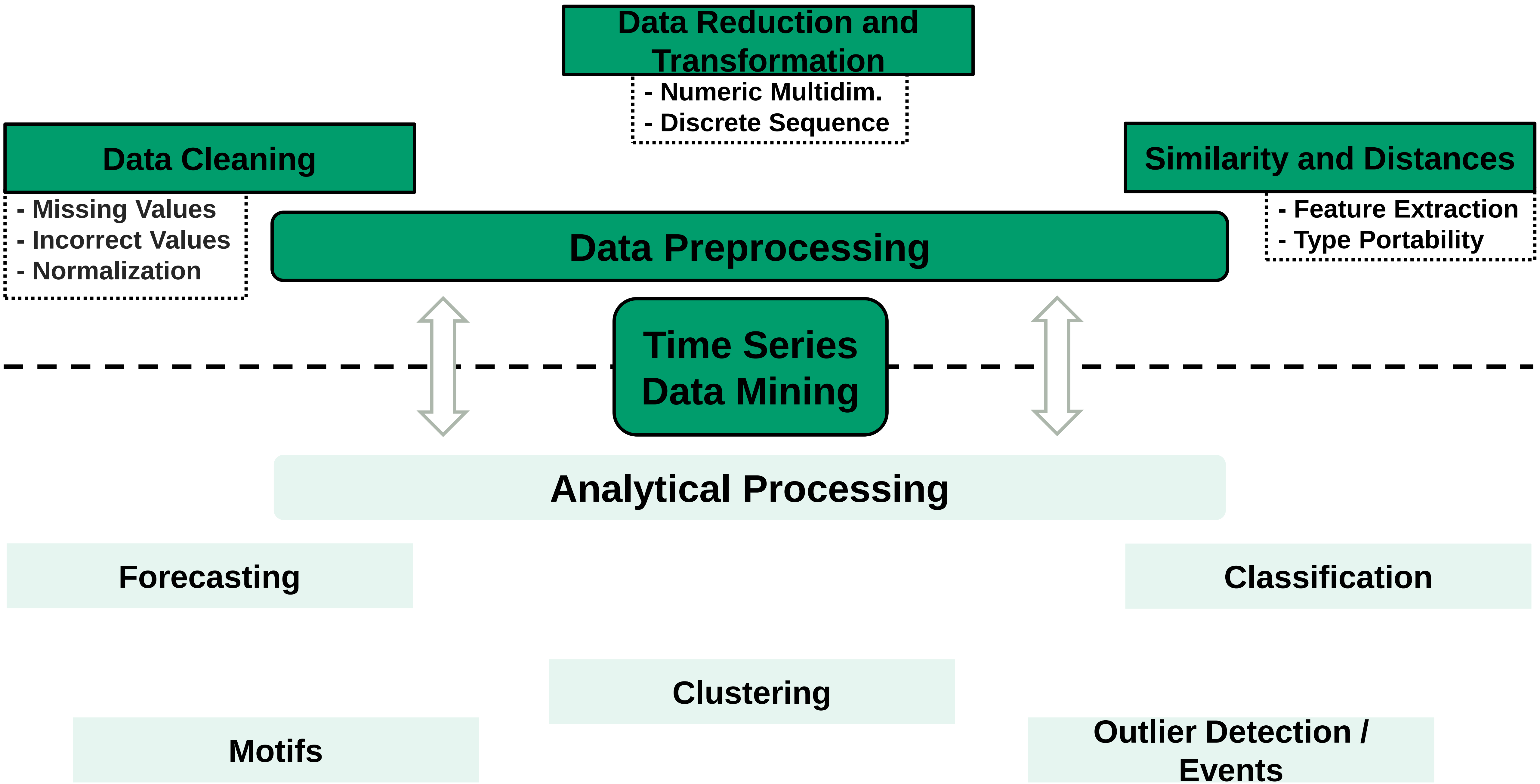
- **Implicit or explicit relationship exists**
- **TS has an implicit one via the contextual time attribute (t_i-1, t_i)**

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

[Aggarwal Tab. 1.1]

[www.google.de]

Overview


Data Reduction and Transformation
- Numeric Multidim.
- Discrete Sequence
Data Cleaning
- Missing Values
- Incorrect Values
- Normalization
Similarity and Distances
- Feature Extraction
- Type Portability
Data Preprocessing
Time Series Data Mining
Analytical Processing
Forecasting
Classification
Clustering
Motifs
Outlier Detection / Events

Data Preprocessing – Data Cleaning (Missing Values)

Data Reduction and Transformation

Data Cleaning

Similarity and Distances

Data Preprocessing

Missing Values examples:
- Hardware failure
- Clock synchronization
- Empty fields

Common approaches:

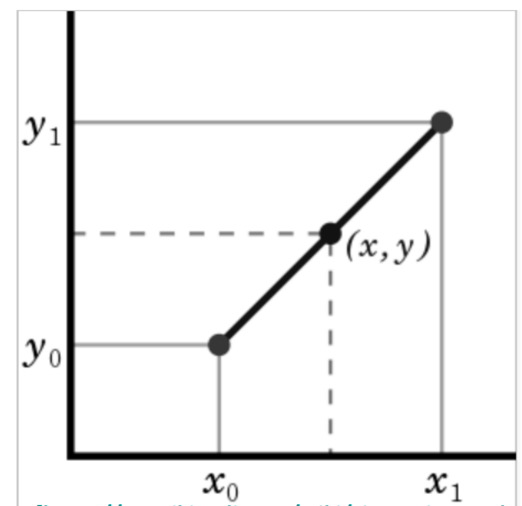
Elimination	Estimation	Robust
– #Samples	e.g. via classification – Error Imputation	Inherent robust DM algorithm + desirable, as no bias

Time Series approaches:
- Implicit dependency allows simpler estimation via contextually nearby records

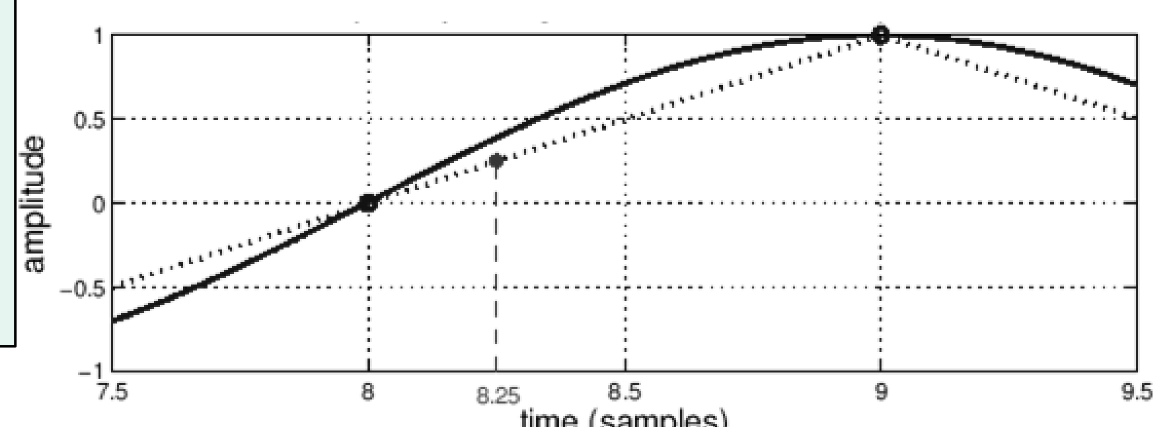
Linear Interpolation:

$$y = y_i + \left(\frac{t - t_i}{t_j - t_i}\right) \cdot (y_j - y_i)$$

→ equally spaced and synchronized values across the different behavioral attributes
+ no significantly superior results via more complex (interpolation) methods

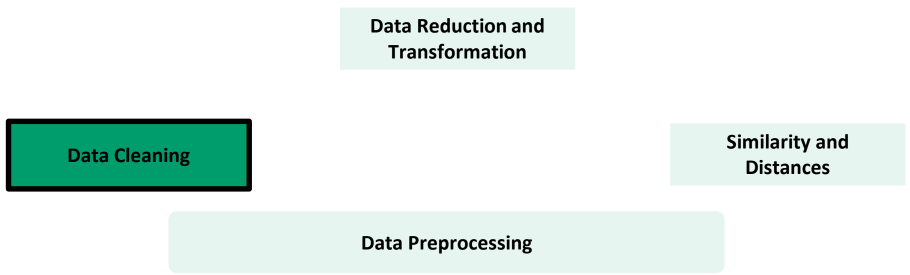


[https://en.wikipedia.org/wiki/Linear_interpolation]



[http://musicweb.ucsd.edu/~trsmyth/filtersDelayII/img15.png]

Data Preprocessing – Data Cleaning (Incorrect Values)



Incorrect Values examples:
- Inaccurate sensors
- Intentional
- Manual errors

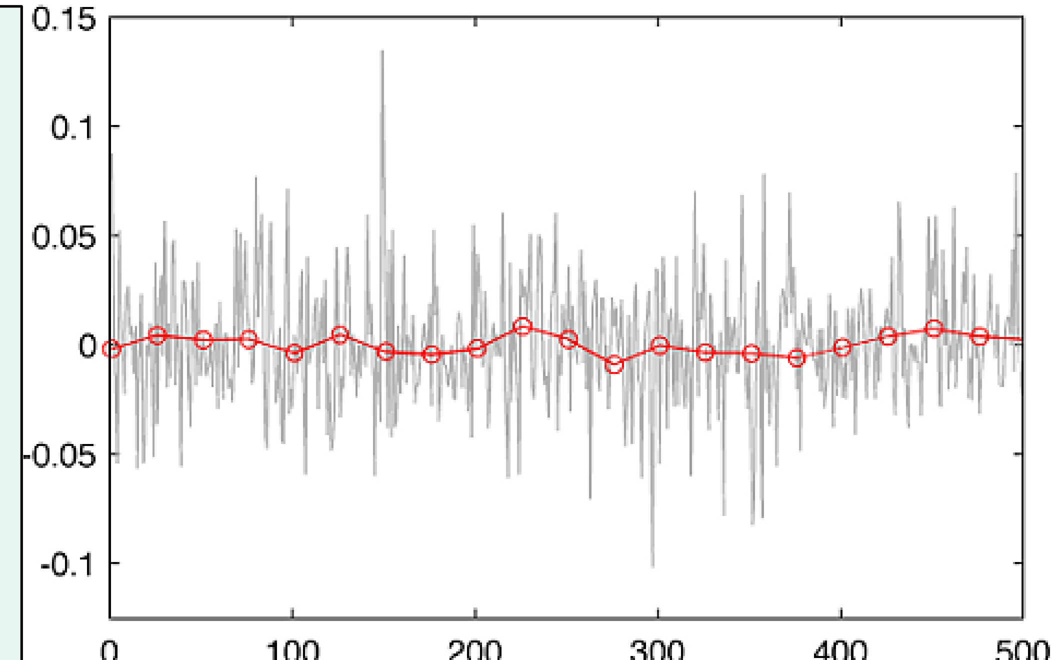
Common approaches:

Inconsistency detection	Domain knowledge e.g. ranges	Data-centric methods e.g. statistics

Time Series approaches:
- Noise vs. Outlier, i.e. interesting fluctuation (e.g. event) → Cleaning and smoothing not generally applicable
- Noise prone sensors → Remove short-term fluctuations
- Methods:
  1) Binning
  2) Moving-Average smoothing 3) Exponential smoothing

1) Binning
- Assumptions: Equally spacing, bins of the same size
- Median better than Average: + robust; e.g. [1,1,2,4,37] → 2

– lossy for large bins
+ compressed representation, e.g. fast distance computation



[http://www.quantatrisk.com/2013/03/22/rebinning-of-financial-time-series/#jp-carousel-929]

Data Preprocessing – Data Cleaning (Incorrect Values)

2) Moving-Average Smoothing
- **Is a rolling average (=overlapping bins)**

+ lesser loss

– lagging, loss of points in the beginning

o larger bin size

→ greater smoothing and lag

→ loss of short term trends possible

→ risk of misinterpretation, i.e. downtrends where there are peaks and vice versa

3) Exponential Smoothing
- **Can be expressed as an exponentially decayed sum of the series values:**

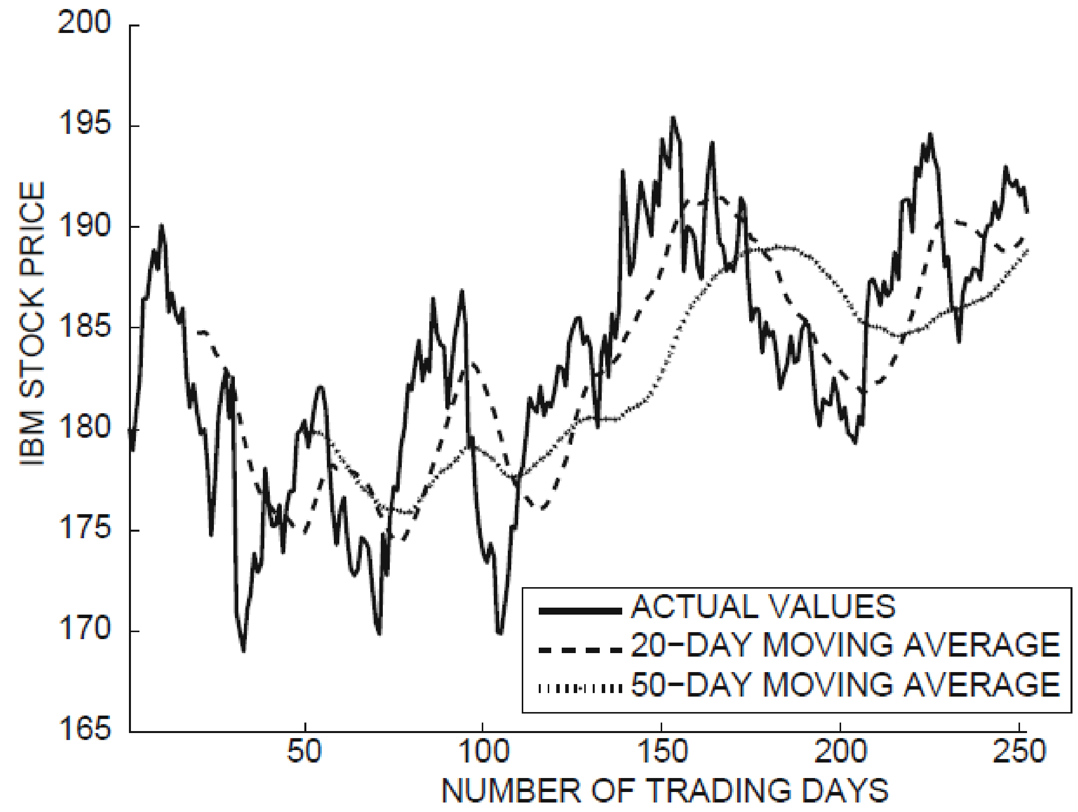
$$y'_i = (1-\alpha)^i \cdot y'_0 + \alpha \cdot \sum_{j=1}^{i} y_j \cdot (1-\alpha)^{i-j}$$

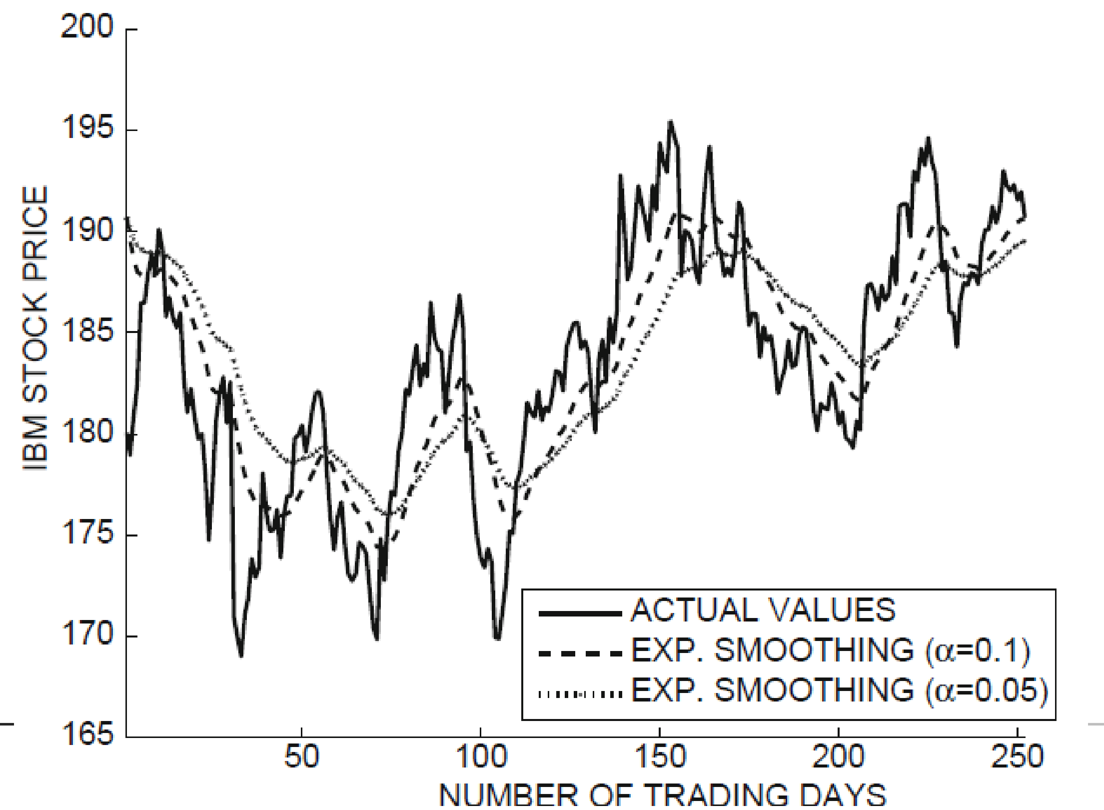
- **Smoothing parameter / decay factor alpha [0,1]**

o Generally slightly better smoothing for lower lag

+ Emphasis on more recent data points

+ No loss of data points at the beginning



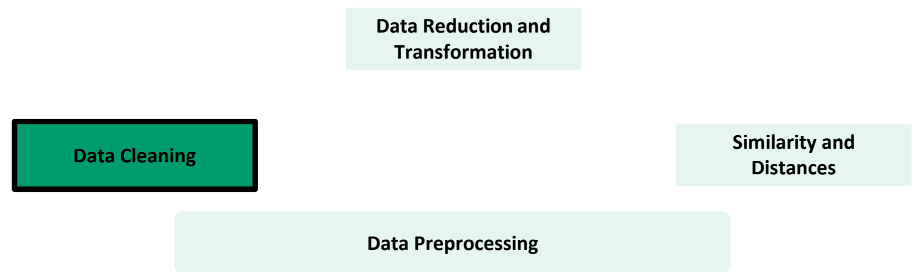


[Aggarwal Fig. 14.1]

Data Preprocessing – Data Cleaning (Normalization)

Normalization examples:
- Domination of one attribute over another (i.a. ranges)
- Ignoring of relevant features

j … Attribute j
i … ith record of the time series



Common approaches:

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$ **Standardization**

Min-Max Scaling $$y_i^j = \frac{x_i^j - min_j}{max_j - min_j}$$

- **Standardization: z is typically in the range of [-3,3]**
- **Min-Max Scaling: y is in range of [0,1]; – extreme outliers**

Time series approach:

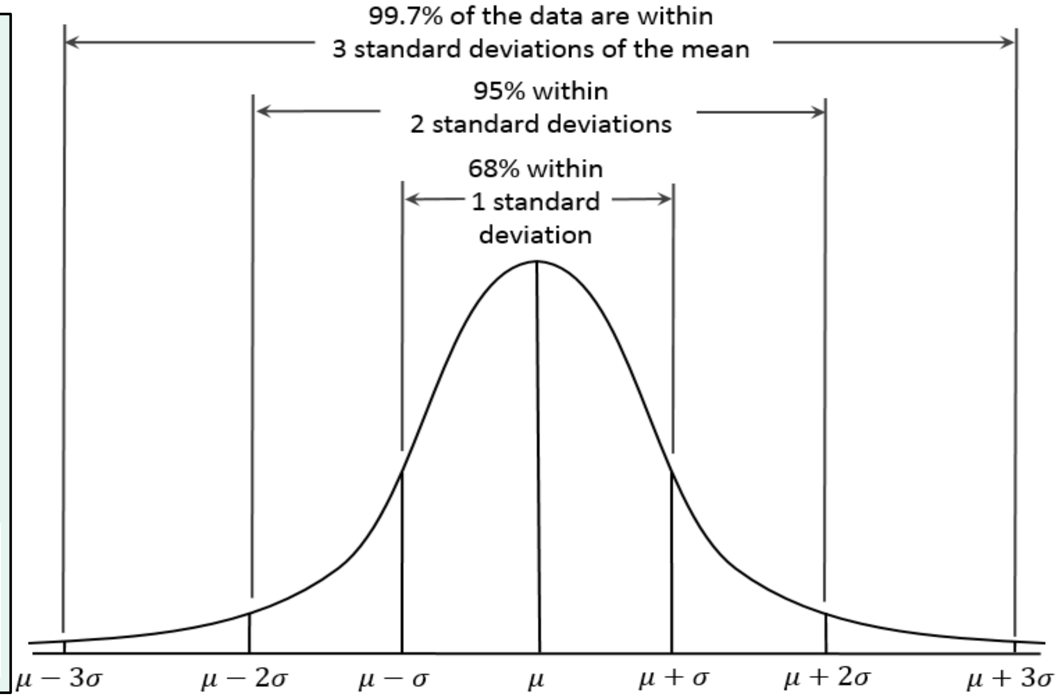
Multiple time series, e.g. temperature T and pressure p with different scales. T in 10^3[K] and p is in 10^6[Pa]

- **Standardization:**
  Z-value mapping of the time series
  + preferred method
  o no guarantee to a specific range

$$z_i = \frac{y_i - \mu}{\sigma}$$

- **Range-Based: y'_i in the range [0,1]**

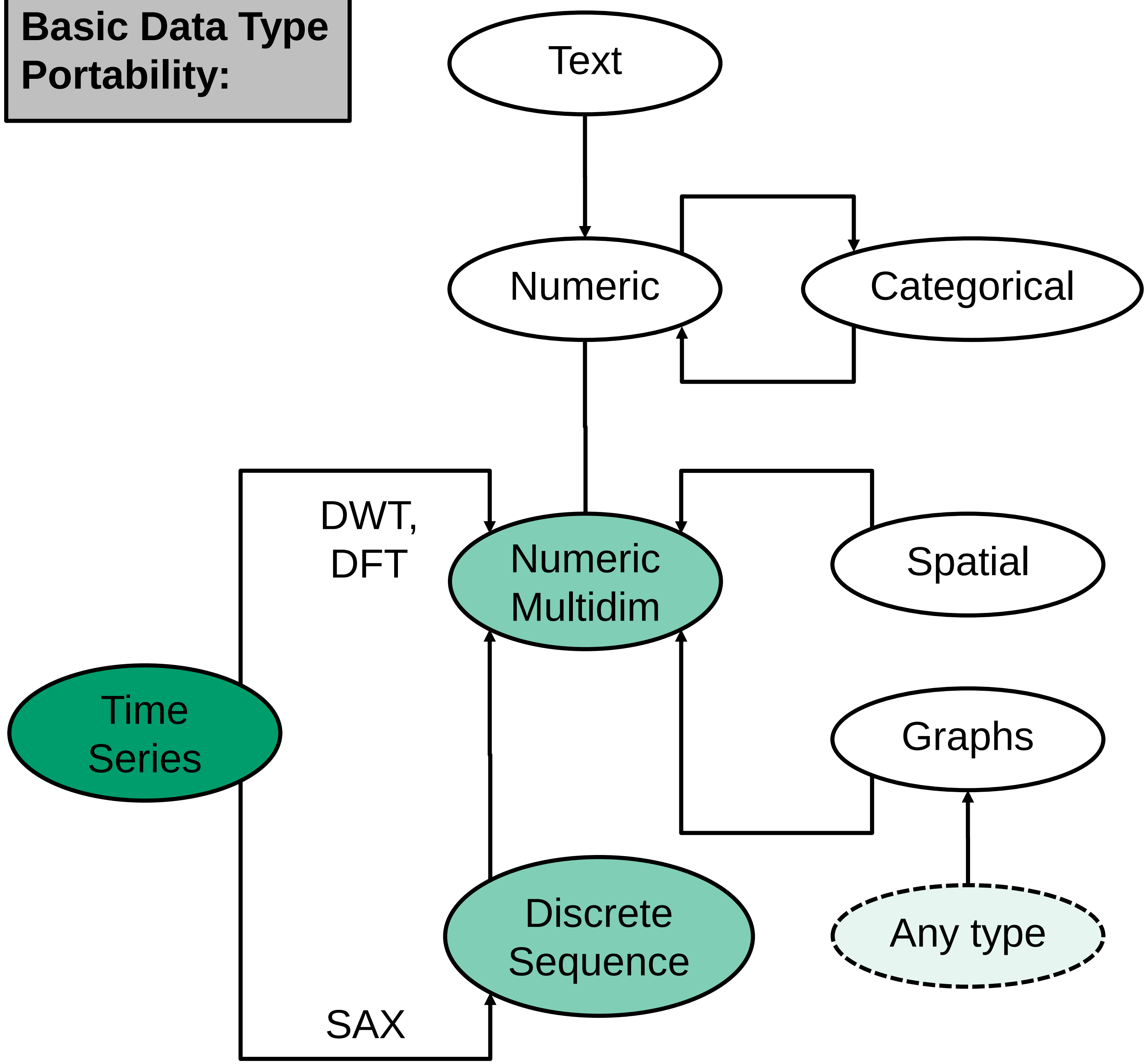
$$y_i' = \frac{y_i - min}{max - min}$$

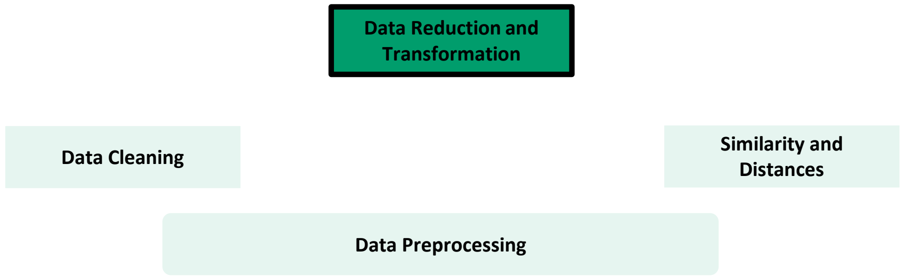


[https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Empirical_Rule.PNG]

Data Preprocessing – Data Transformation

Basic Data Type Portability:



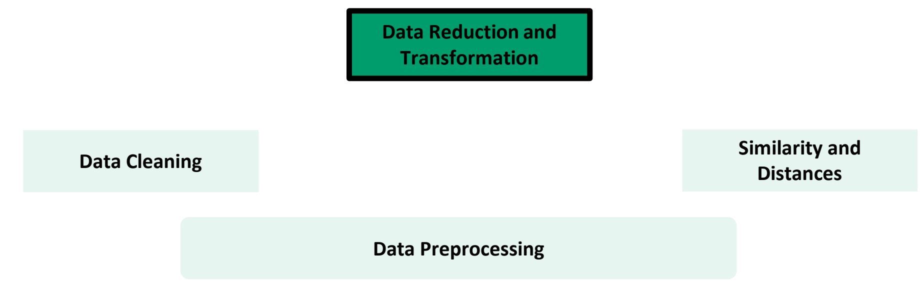


Data type portability allows:

- \+ **Data mining algorithms available in other data type domains allow more diverse data exploration and interpretation**
- \+ **Smaller size and complexity of the data set**

Data Preprocessing – Data Reduction

Data Transformation without information loss, but with quantity reduction.



Common approaches:

- **Sampling**
- **Feature Subset Selection**
- **Dim. Reduct. with Axis Rot.**
- **Dim. Reduct. With Type Transf.**

Time series approach:

1) Time Series to Numeric Multidimensional Data

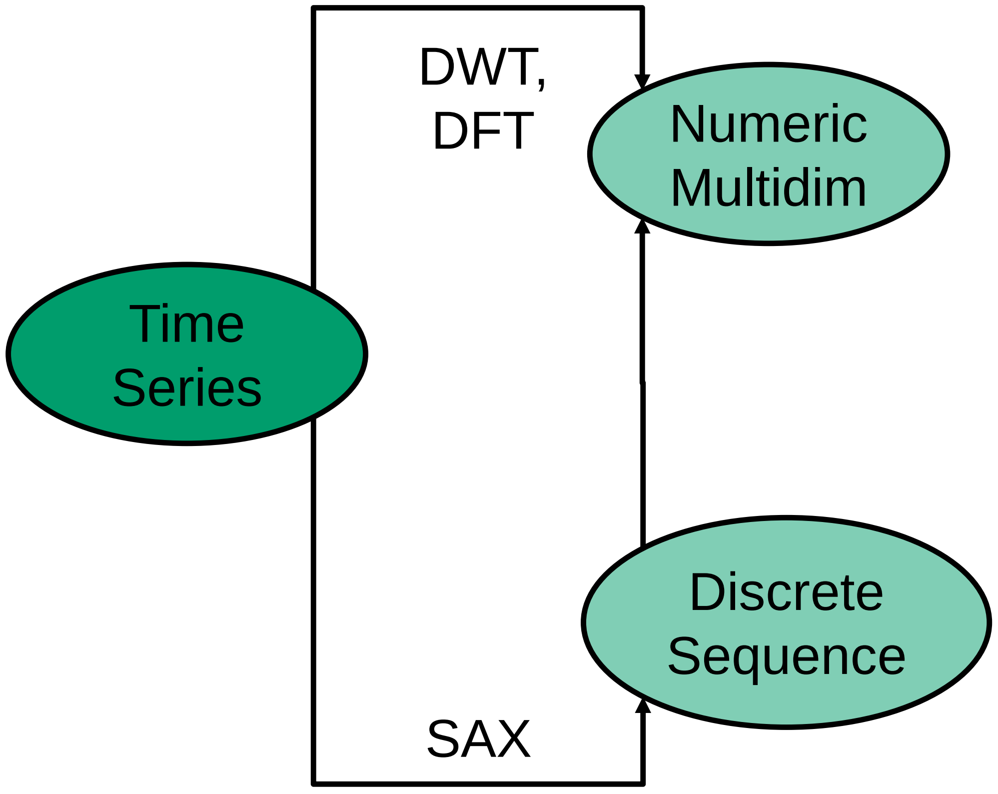
- **Via Discrete Wavelet Transformation (DWT) and**
- **Via Discrete Fourier Transformation (DFT)**

– loss of implicit dependency

2) Time Series to Discrete Sequence Data

- **Via Symbolic Aggregate Approximate (SAX)**

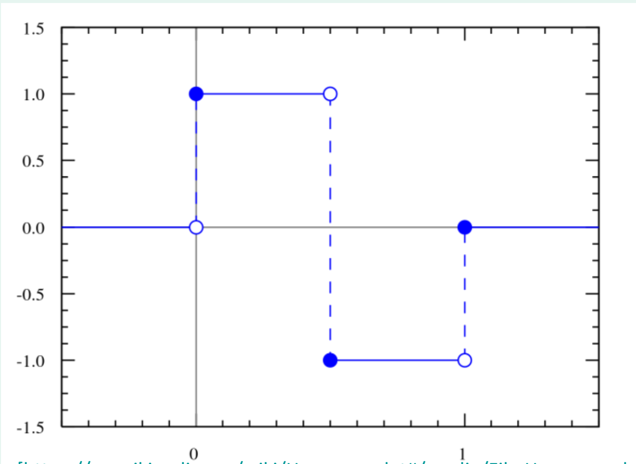
+ Rich set of algorithms in the field of discrete sequence data can be used



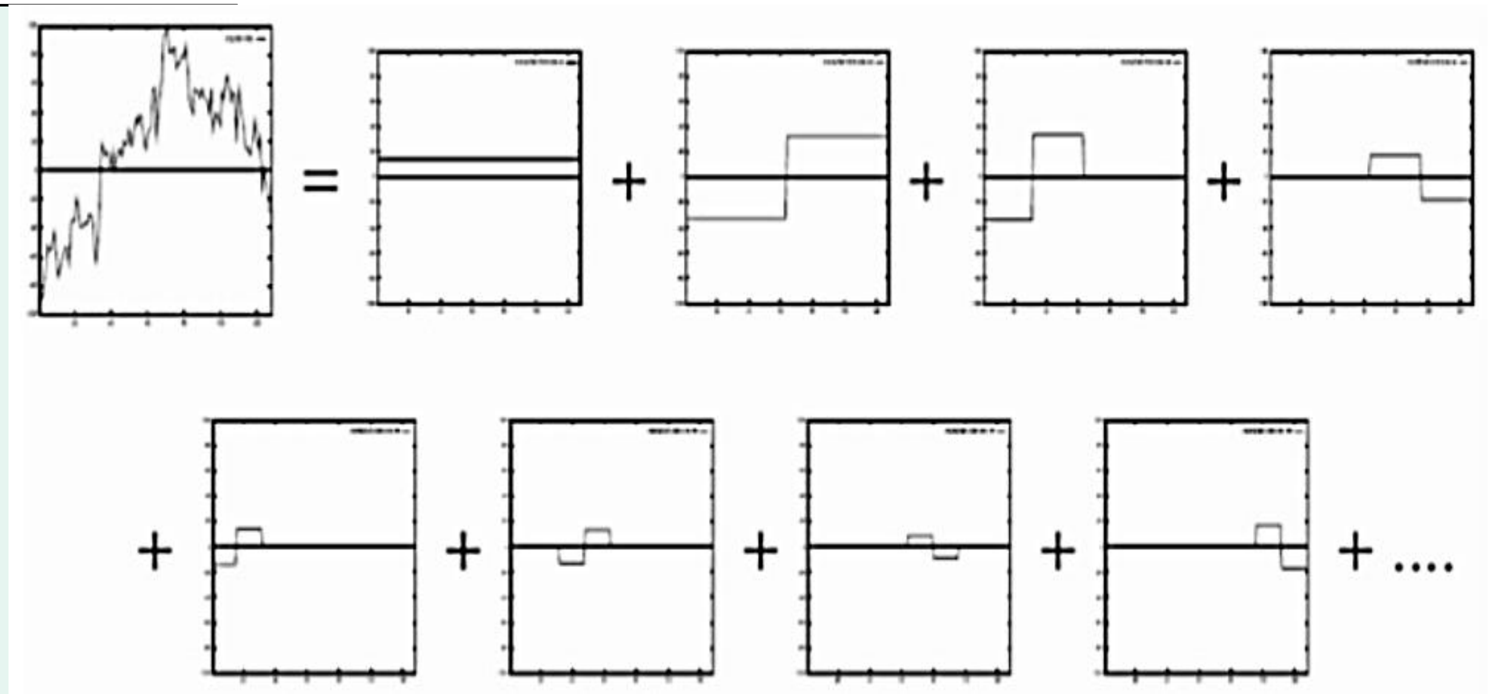
Data Preprocessing – Data Reduction

1) Time Series to Numeric Multidim:

- **Discrete Wavelet Transformation (DWT) = Linear Combination of Wavelets, here Haar wavelet:**



[https://en.wikipedia.org/wiki/Haar_wavelet#/media/File:Haar_wavelet.svg]

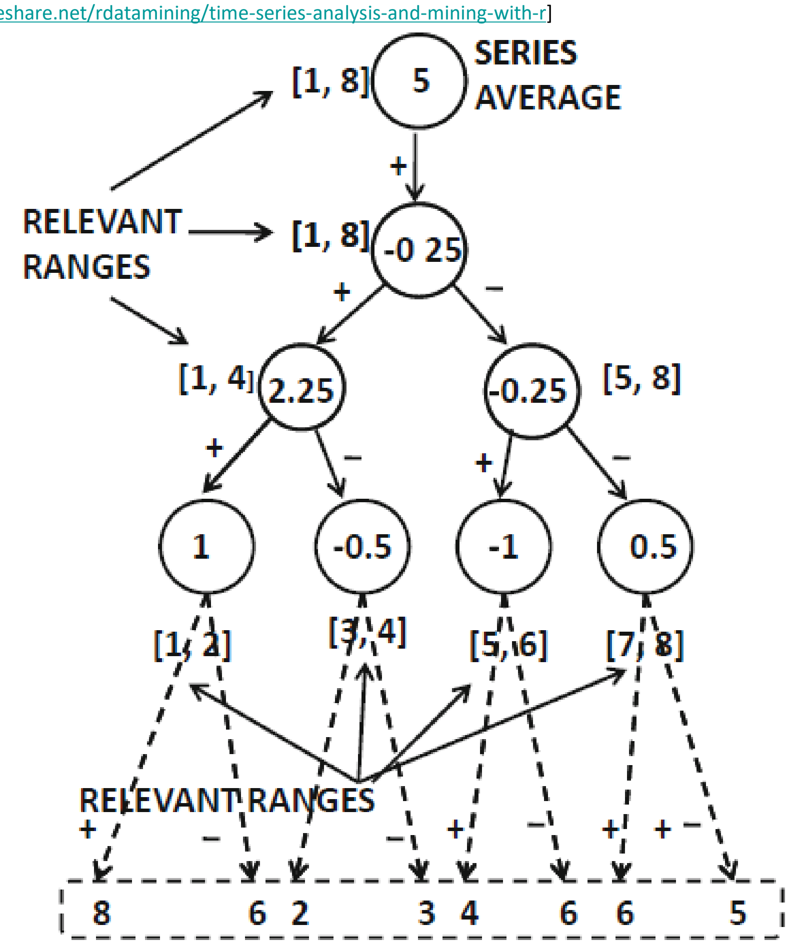


[https://www.slideshare.net/rdatamining/time-series-analysis-and-mining-with-r]

- **Schematic idea:**
  Store series average -- store average of the halfes -- store average of the quarters -- … (Recursively apply) … -- store single measurement

\+ **multigranularity decomposition and summarization**

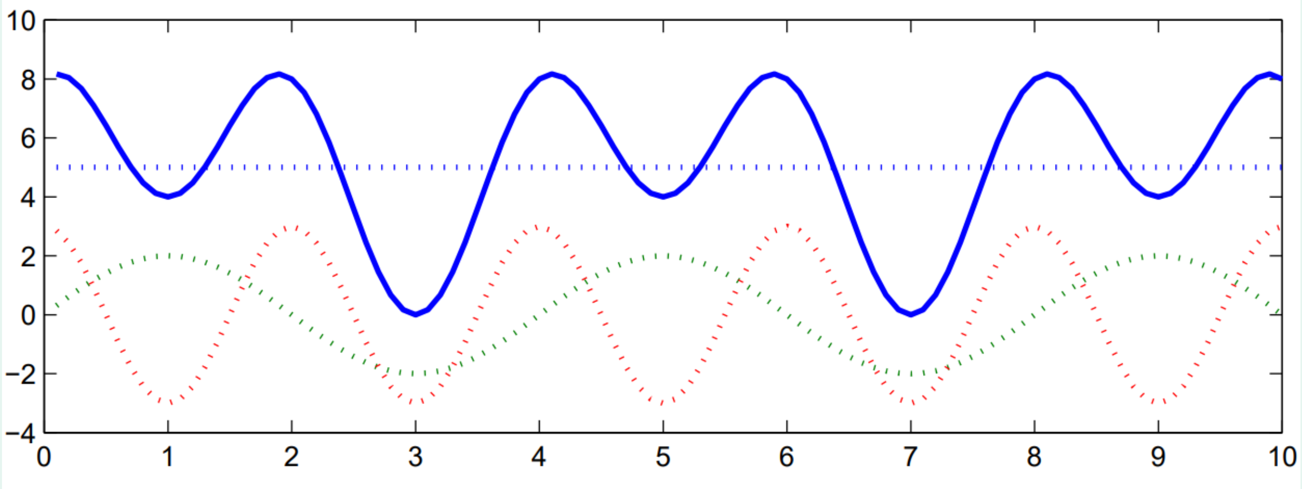
- **High order coefficients correspond to large ranges**
  → represent the broad trends
- **Low order coefficients → capture localized trends**



[Aggarwal Fig. 2.6]

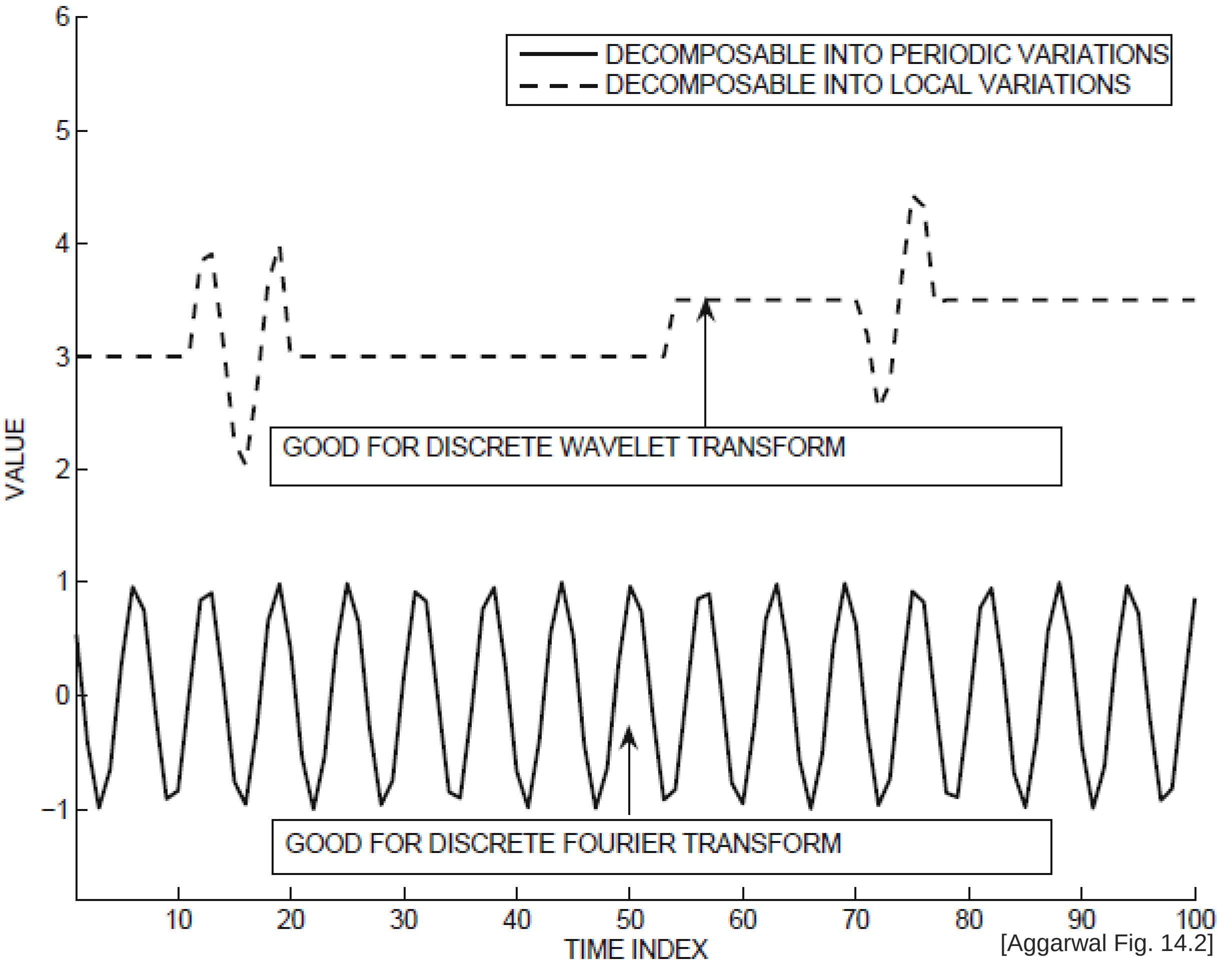
Data Preprocessing – Data Reduction

1) Time Series to Numeric Multidim:
- **Discrete Fourier Transformation (DWT)**
  = Linear combination of smooth periodic sinusoidals



[http://www.robots.ox.ac.uk/~sjrob/Teaching/SP/l7.pdf Fig 7.2]

DWT vs DFT:
- **Most variation in specific local regions → DWT**
- **Most variation is periodical → DFT**



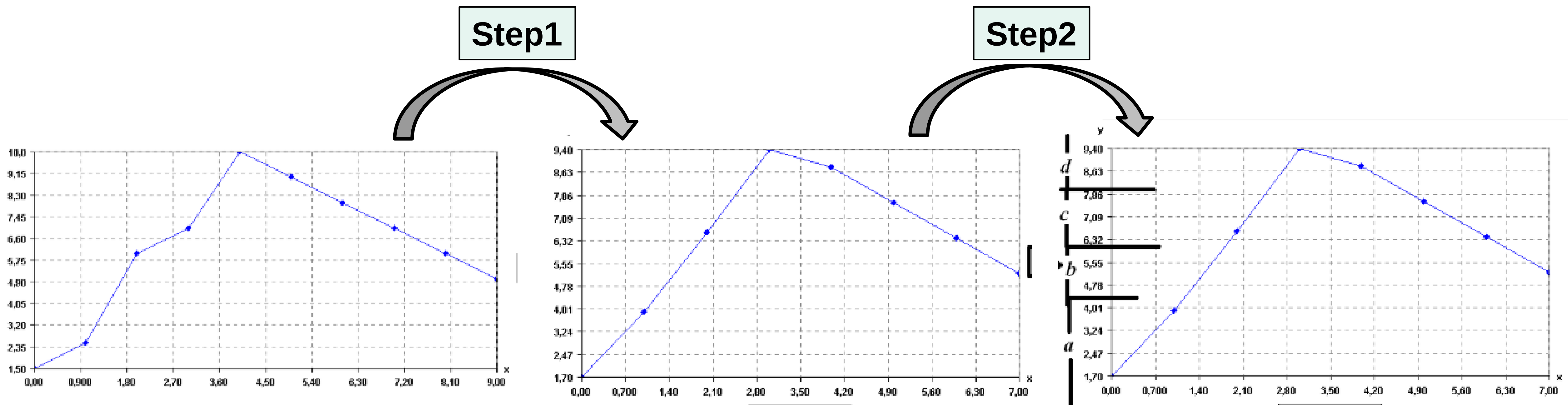
[Aggarwal Fig. 14.2]

Data Preprocessing – Data Reduction

2) Time Series to Discrete Sequence Data: Symbolic Aggregate Approximate (SAX)

Step1: Window-based averaging, i.e. use window and compute average in it = Piecewise aggregate approximation (PAA)

Step2: Value-based discretization, i.e. discretize into smaller number of approximately equi-depth intervals. Assumption of normal distribution → values (= symbols a,b,c,d) are approximately equally distributed.



Original time series (11 data points)

Piecewise aggregate approximation (8 data points)

Piecewise aggregate approximation (8 data points)

SAX representation of the time series

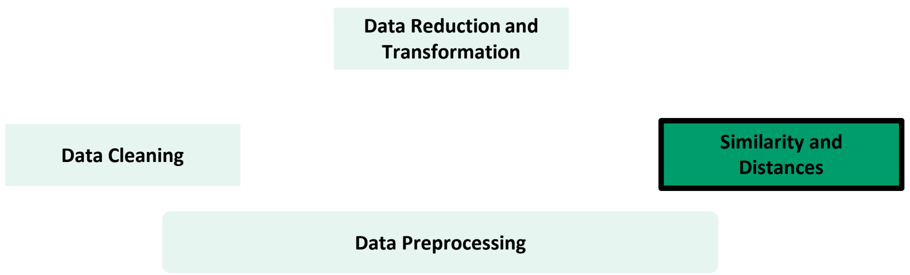
a, a, c, d, d, c, c, b

[http://data-mining.philippe-fournier-viger.com/introduction-time-series-mining-spmf/]

Data Preprocessing – Similarity and Distances

Distances importance e.g.:
- **Clustering**
- **Classification**

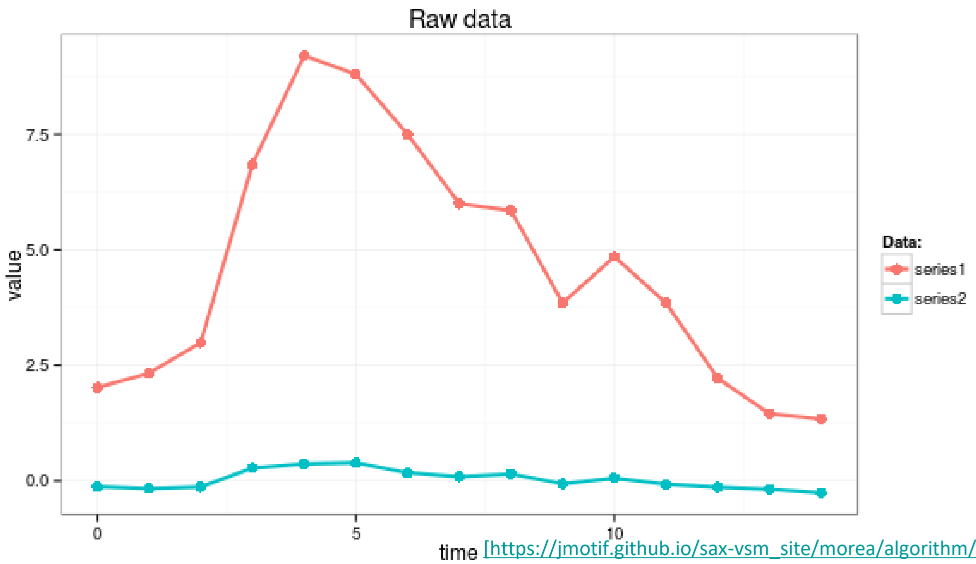
Aspects: Dimensionality, Data Distribution, Data Type.



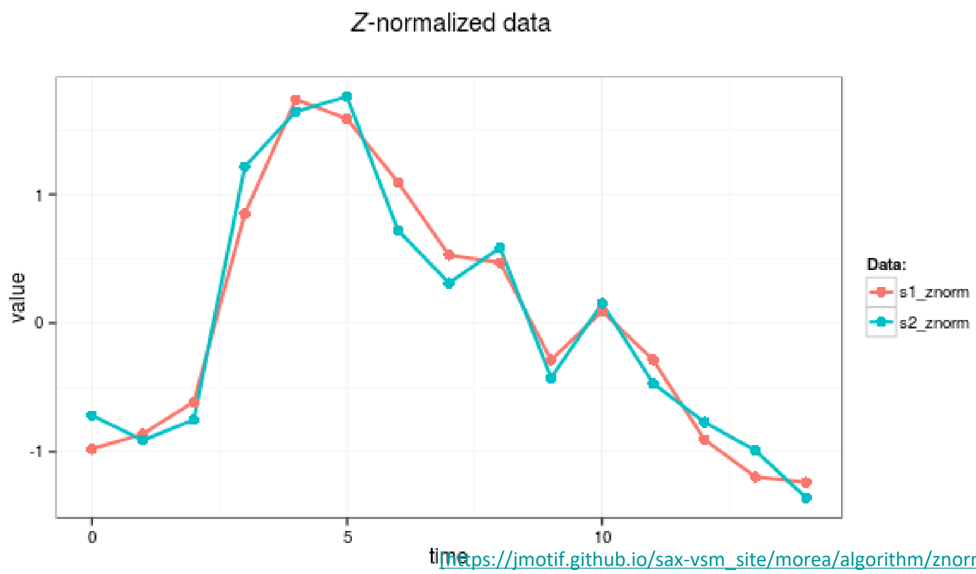
<u>Time Series approach:</u>
- **Distance described algorithmically**
- **Behavioral Attribute: Scaling and translation addressed by normalization. (Done only if needed)**

Methods:
1) L_p-Norm
2) Dynamic Time Warping Distance (DTW)

1) L_p Norm / Euclidean distance (p=2)
- **TS of same length**
- **1:1 correspondence**
- **Computation of distance at each timestamp**



[https://jmotif.github.io/sax-vsm_site/morea/algorithm/znorm.html]



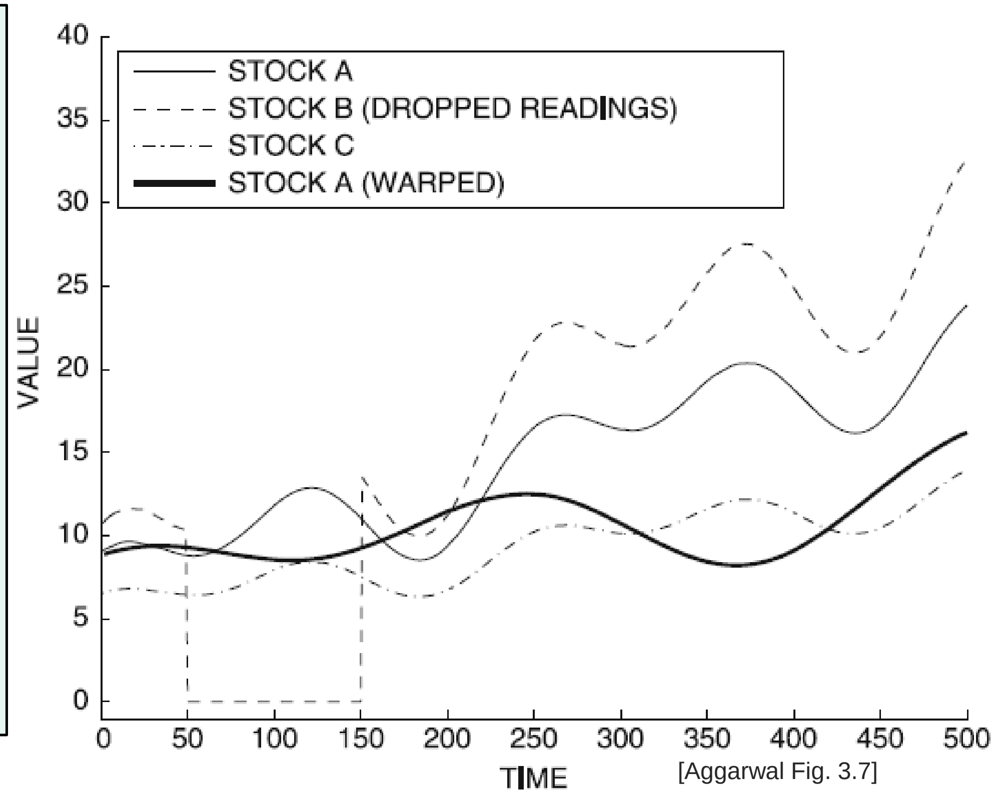
[https://jmotif.github.io/sax-vsm_site/morea/algorithm/znorm.html]

Data Preprocessing – Similarity and Distances

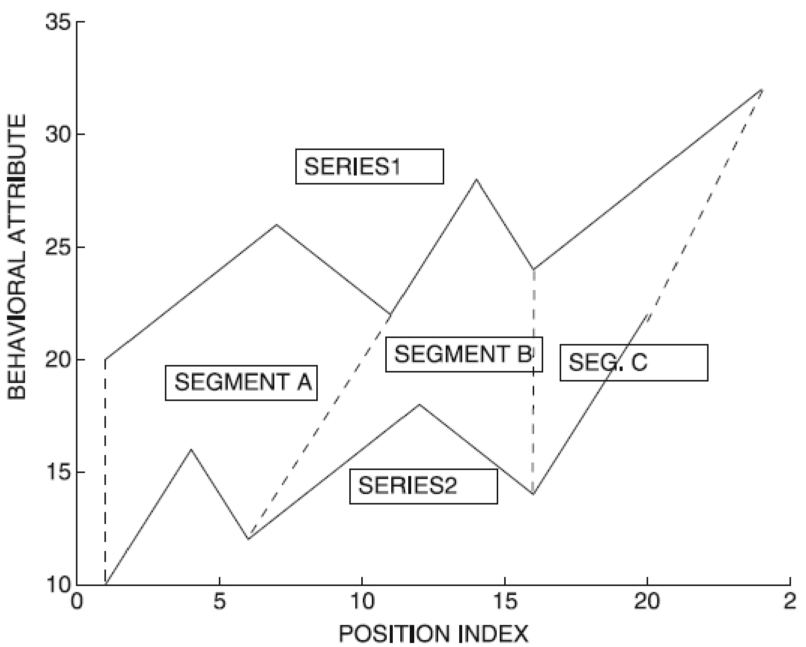
- **Contextual attribute distortion factors:**
  Scaling and noncontiguity (e.g. dropped readings)

2) DTW
- **Optimal matching via streching and shrinking of the time dimension in different portions**
- **Application example: speech recognition, i.e. different speeds**

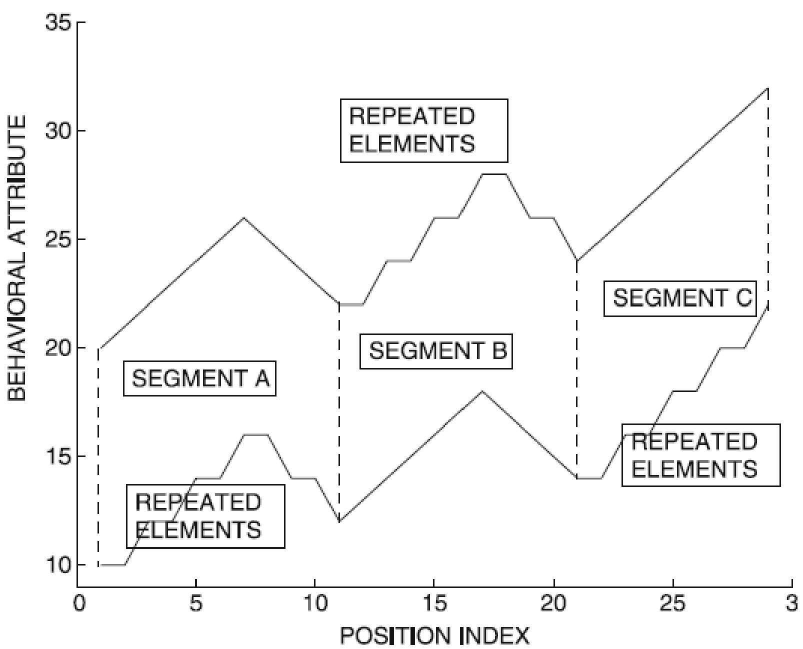
\+ **Adresses the issue of contextual attribute scaling**
\+ **unrelated to the nature of the behavioral attribute**
\+ **Allows 1:n mapping**



[Aggarwal Fig. 3.7]

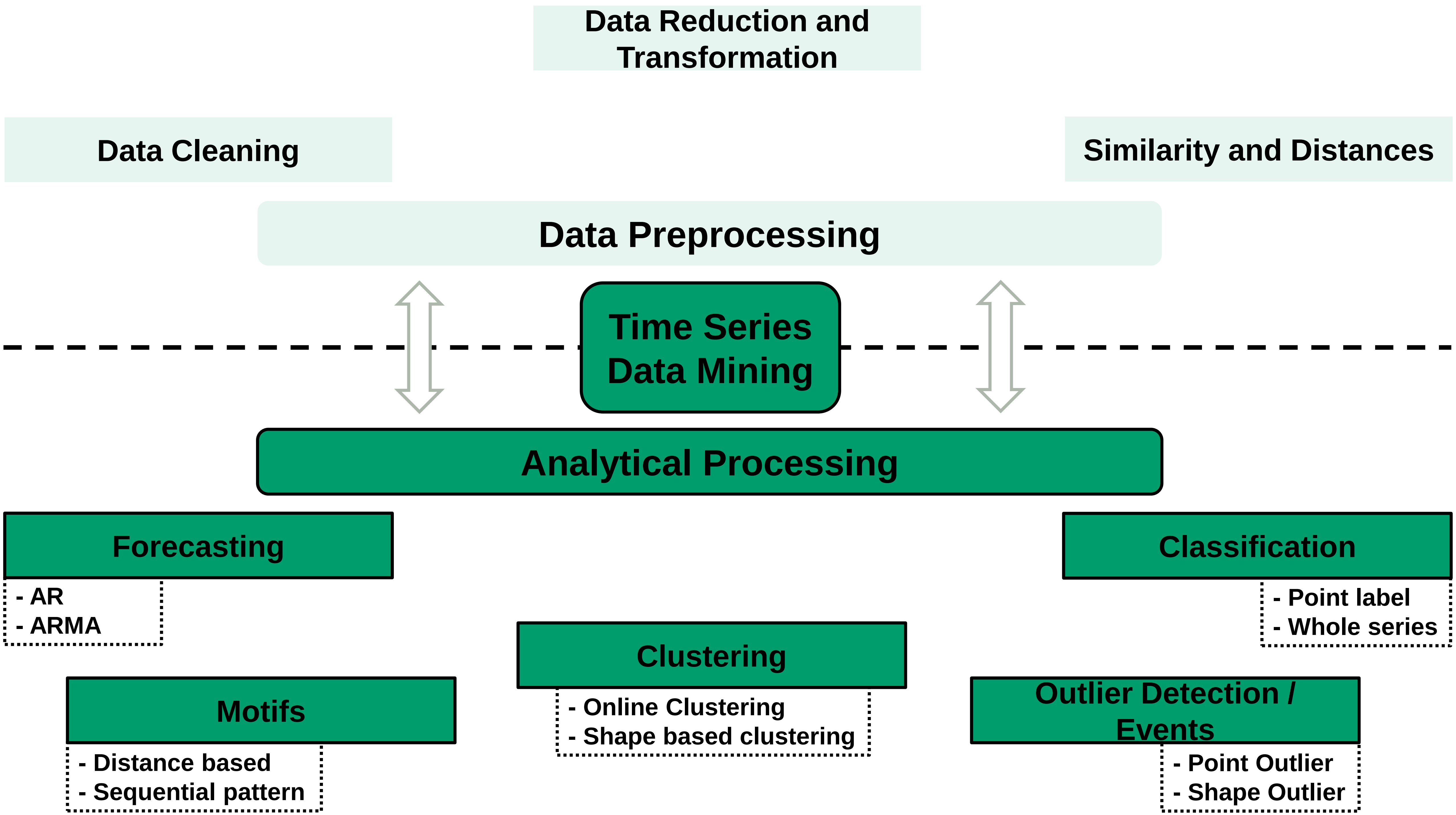


(a) Original series [Aggarwal Fig. 3.8]



(b) Warped series [Aggarwal Fig. 3.8]

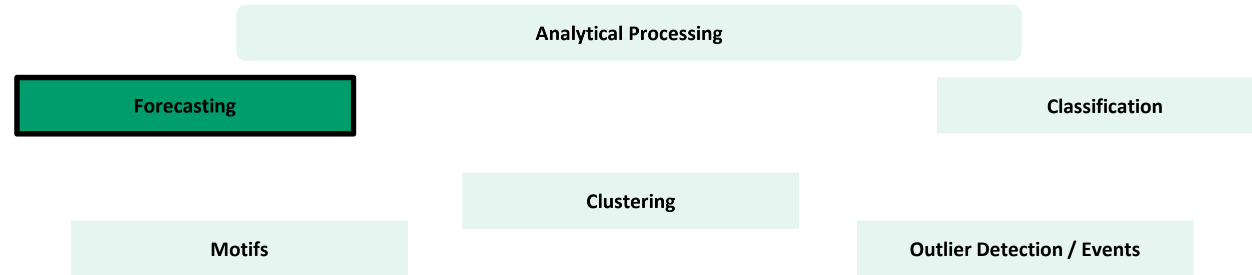
Overview

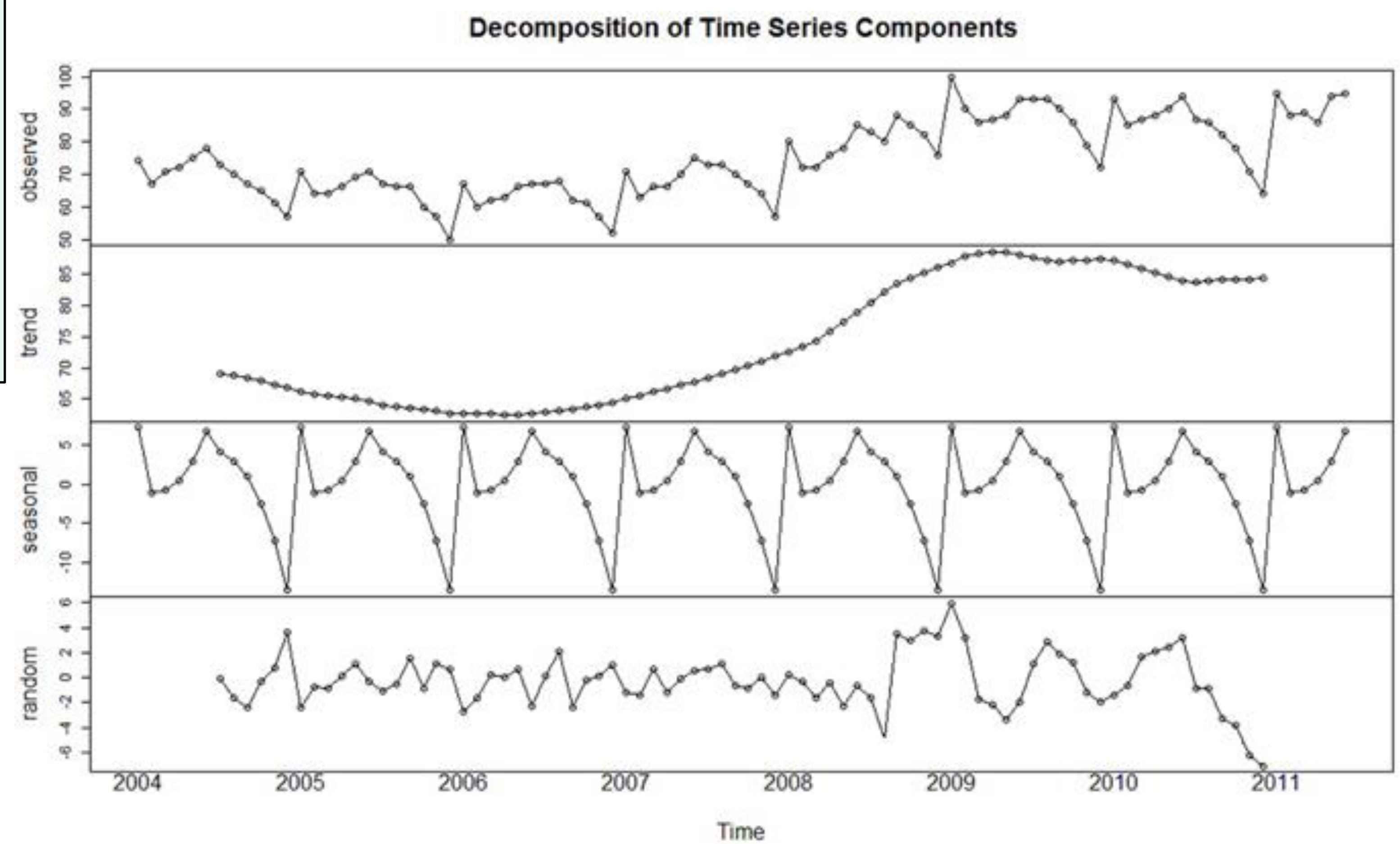

Data Reduction and Transformation
Data Cleaning
Similarity and Distances
Data Preprocessing
Time Series Data Mining
Analytical Processing
Forecasting
- AR
- ARMA
Classification
- Point label
- Whole series
Clustering
- Online Clustering
- Shape based clustering
Motifs
- Distance based
- Sequential pattern
Outlier Detection / Events
- Point Outlier
- Shape Outlier

Analytical Processing – Forecasting

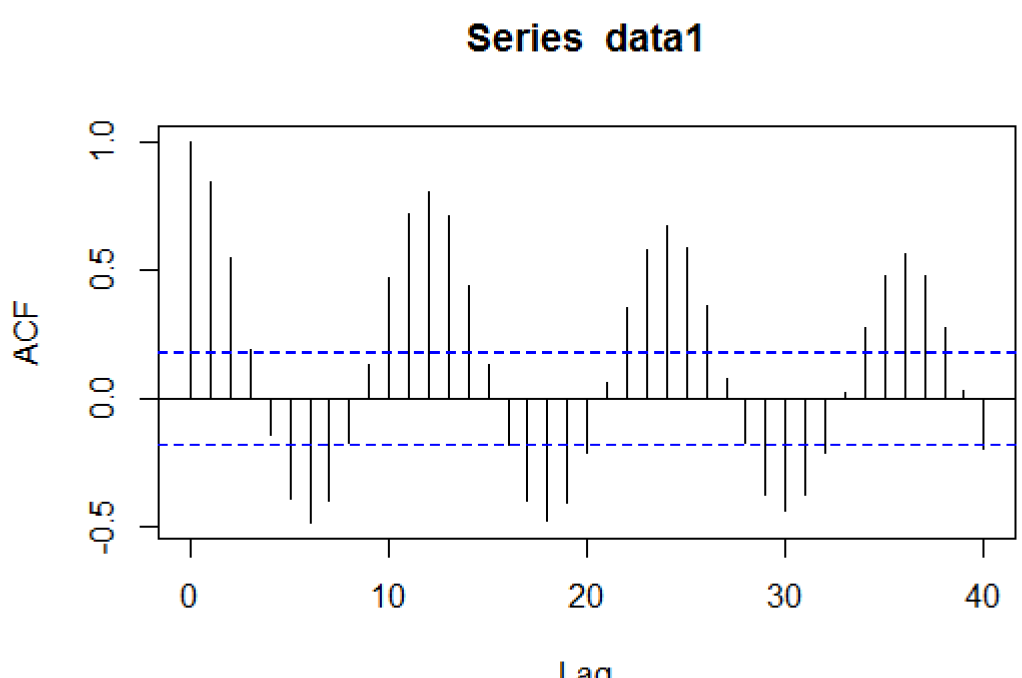
Forecasting

- Stationary series via differencing, detrending, no seasonality
- → ARIMA model used for prediction
- Order of AR and MA mostly of low order (high order = overfitting)
- Autocorrelation = correlation of a signal with itself
- Partial Autocorrelation, i.e. no more periodic correlations





[https://www.researchgate.net/figure/The-original-time-series-decomposed-into-its-trend-seasonal-and-irregular-ie_fig2_279249485]



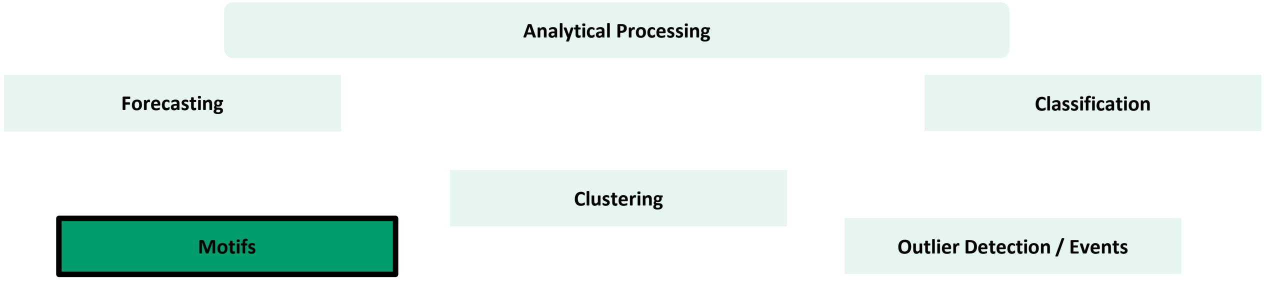
[https://stats.stackexchange.com/questions/138555/interpretation-of-the-autocorrelation-plot?rq=1]

Analytical Processing – Motifs

<u>Motif</u> = frequently occuring pattern or shape

Nature of Motifs (application specific)

- **Single series vs Multi series**
- **Contiguous vs non Contiguous**
- **Multigranularity Motifs**



<u>Dicovery of Motifs</u>

1) Distance based motifs

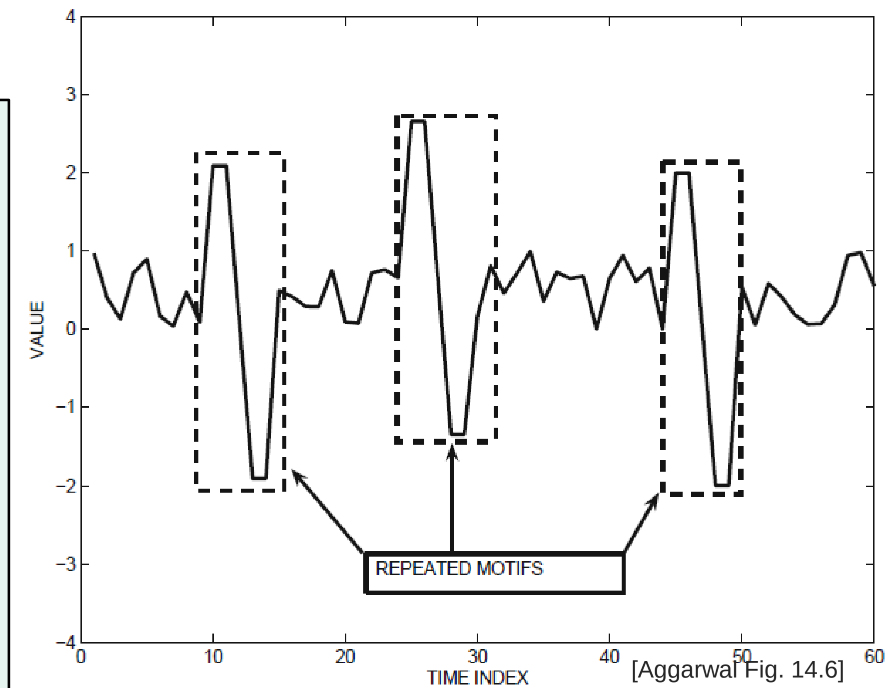
- **Distance thresholding with a contiguous segment**
- **Recap. Distance: Euclidean or DTW**
- **Only use most frequent motifs.**

2) Transformation to sequential pattern mining

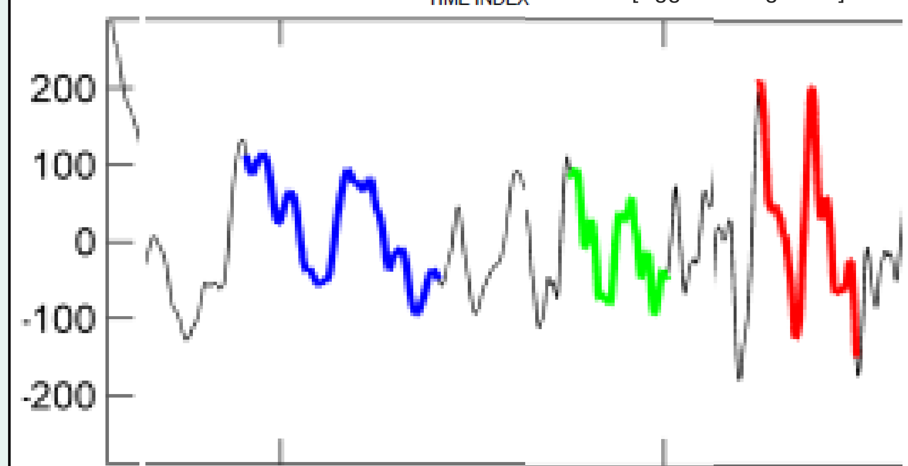
- **Now Motif = discrete subsequence of the sequence**
- **Behavioral attributes are now categorical values**
- **Robust sequence representation via binning**

\+ **Discover noncontiguous patterns as no contiguity is assumed by default**

- **Multiresolution pattern via the DWT coefficients (includes local patterns)**
- **Periodic pattern via the DFT coefficients**



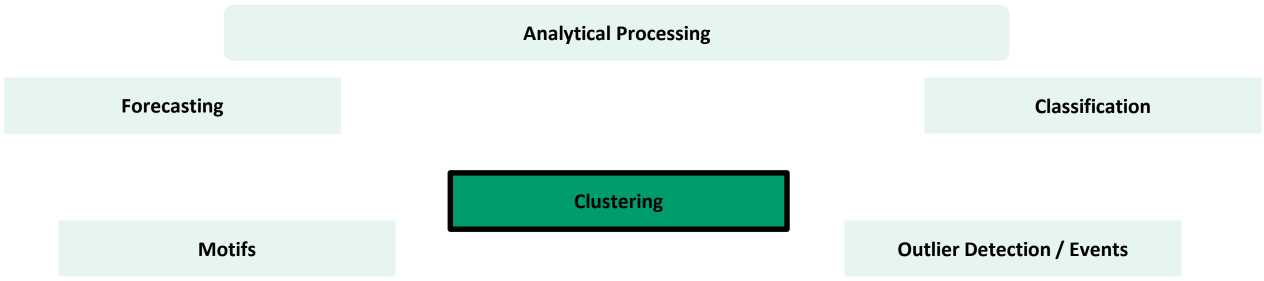
[Aggarwal Fig. 14.6]



[http://alfa.di.uminho.pt/~castro/mrmotif/]

Analytical Processing – Clustering

Clustering scenarios
1) Online Clustering (real time)
2) Shape based Clustering within a DB

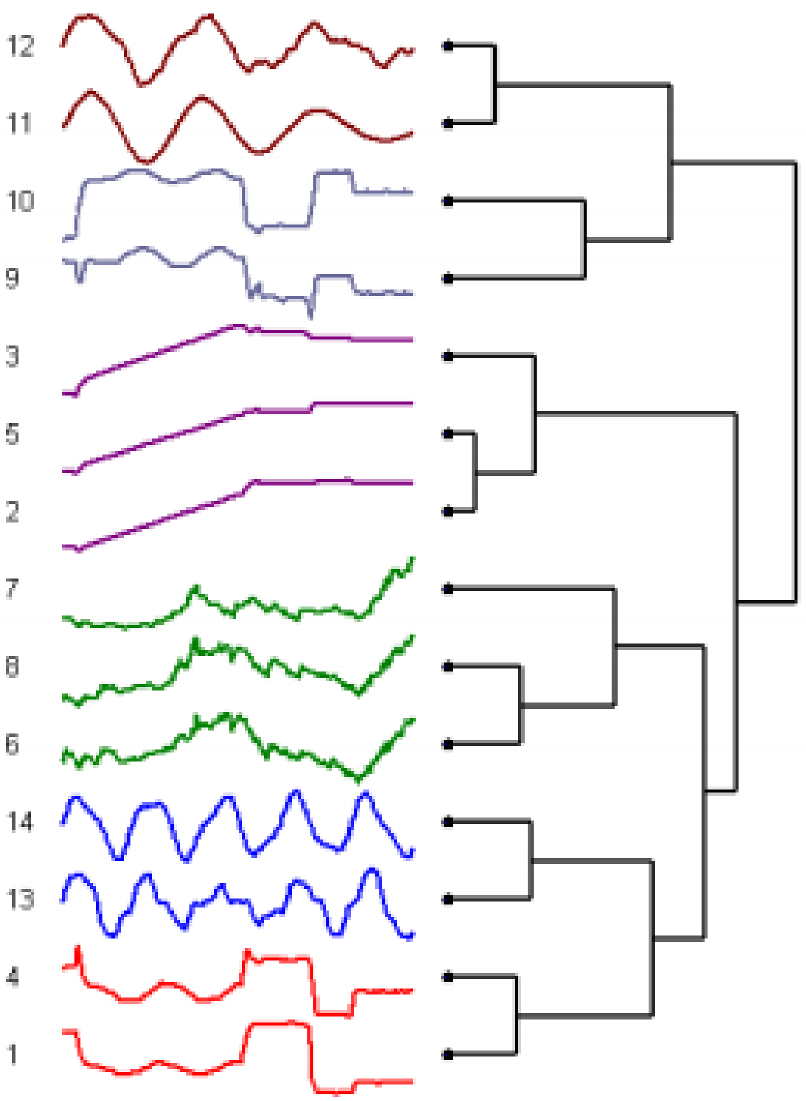


1) Online Clustering (real time)

- **Receiving the TS simultaneously**
- **Regression based similarity functions to compute the similarities between the different streams (forecasting)**

2) Shape based Clustering within a DB

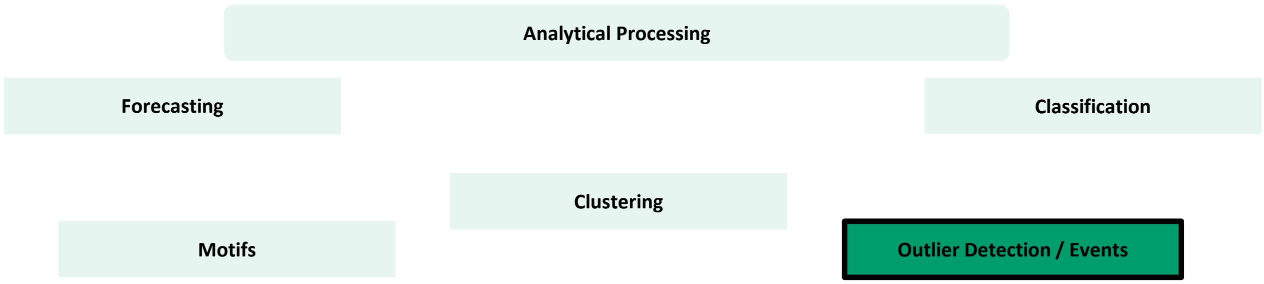
- **Adaption of common clustering methods via different similarity functions (e.g. DTW)**
- **K-medoids: Existing data point is the mean. Arbitrary distance function → DTW aplicable**
- **Hierarchical: Use if the number of TS is small. Distance functions between all pairs of the TS needed – Expensive**



[http://www.hanselsolutions.com/img/blog/dendogram-matlab.png]

Analytical Processing – Outlier Detection

Outlier is a significant deviation from the expected forecasted value
1) Point Outlier
2) Shape Outlier

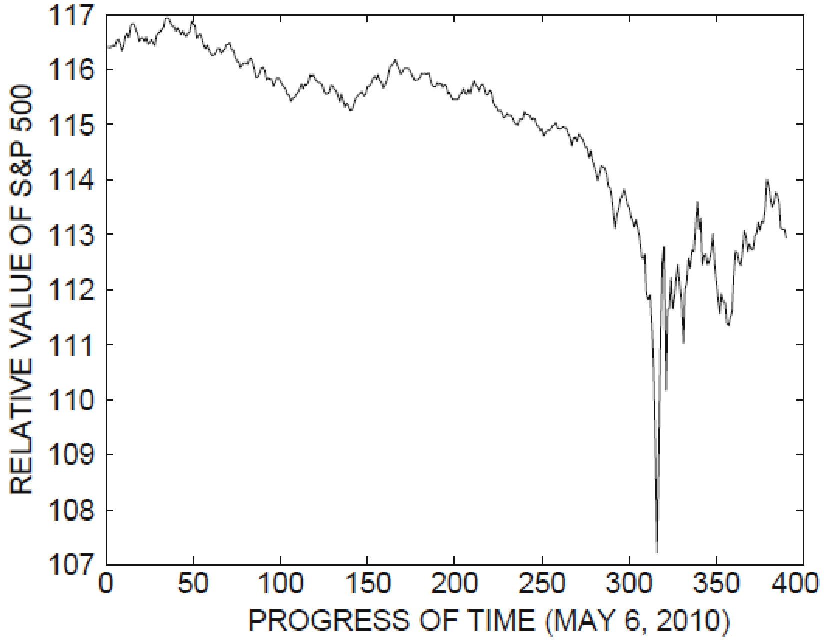


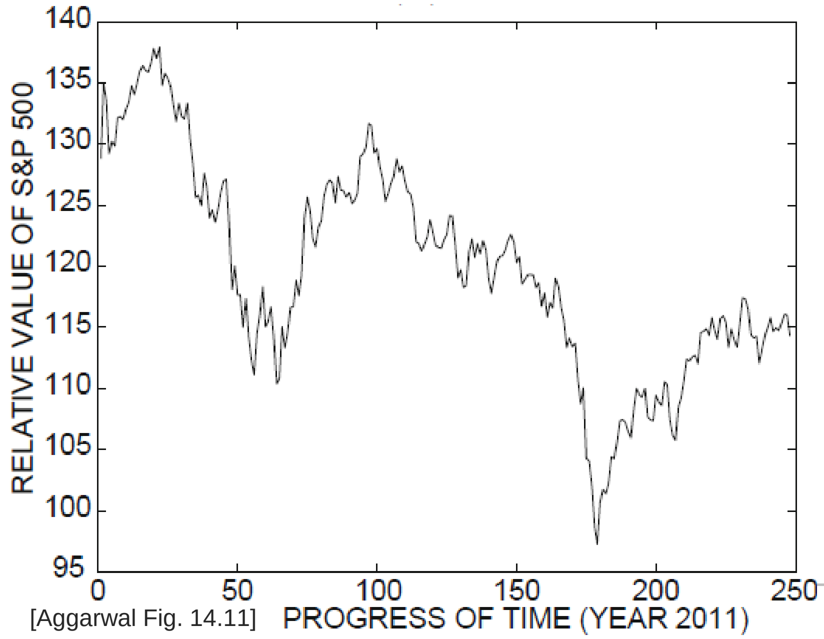
Time Series:

1) Point Outlier

- **Contextual outliers regarding time**
- **Event detection = outlier detection performed in real time**
- **Steps for one time series:**
  - **a) Determine the forecasted value of the TS at each timestamp**
  - **b) Compute deviations at each timestamp between the predicted and actual value**
  - **c) Compute mean and deviation value of the deviations**
  - **d) Compute the normalized deviations (essentially equal to the Z-value of a normal distribution)**
- **Outlier = if threshold >3, is sufficient**

- **Outlier ensemble analysis: unified alarm level of deviation scores for many TS.**





[Aggarwal Fig. 14.11]

Analytical Processing – Outlier Detection

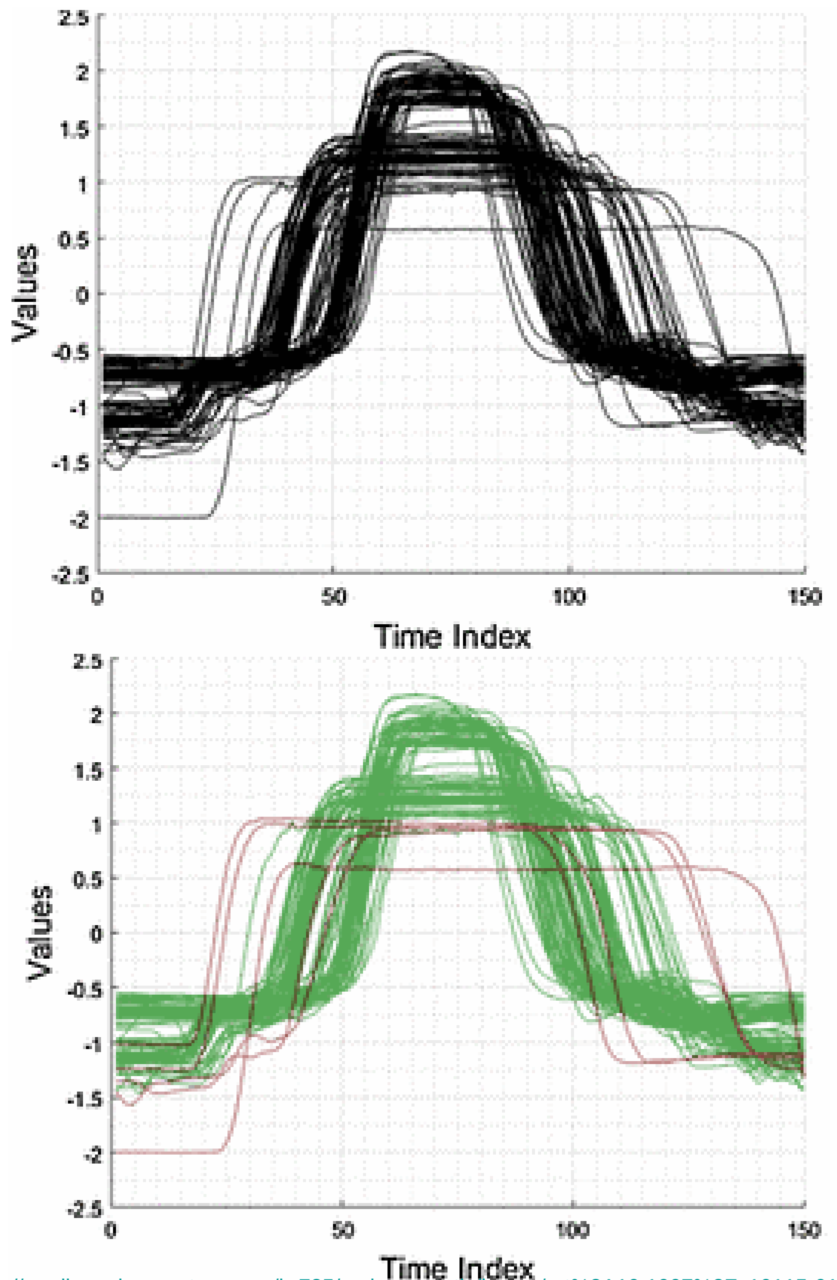
2) Shape Outlier

- **Pattern of data points within a contiguous window**
- **No individual point is considered an anomaly**
- **E.g. Irregular heartbeat of a patient**
- **Hotsax approach → windows of unusual shapes from a TS. Steps:**

  a) extract windows via sliding window approach

  b) for each extracted window, compute the euclidean distance to the other nonoverlapping windows (less trivial matches)

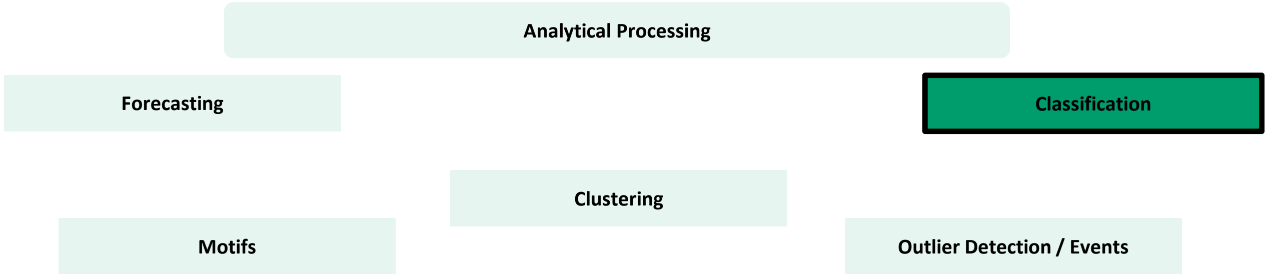
  c) windows with the highest k-nearest neighbor distance are reported as outliers



https://media.springernature.com/lw785/springer-static/image/art%3A10.1007%2Fs10115-017-1067-8/MediaObjects/10115_2017_1067_Fig3_HTML.gif

Analytical Processing – Classification

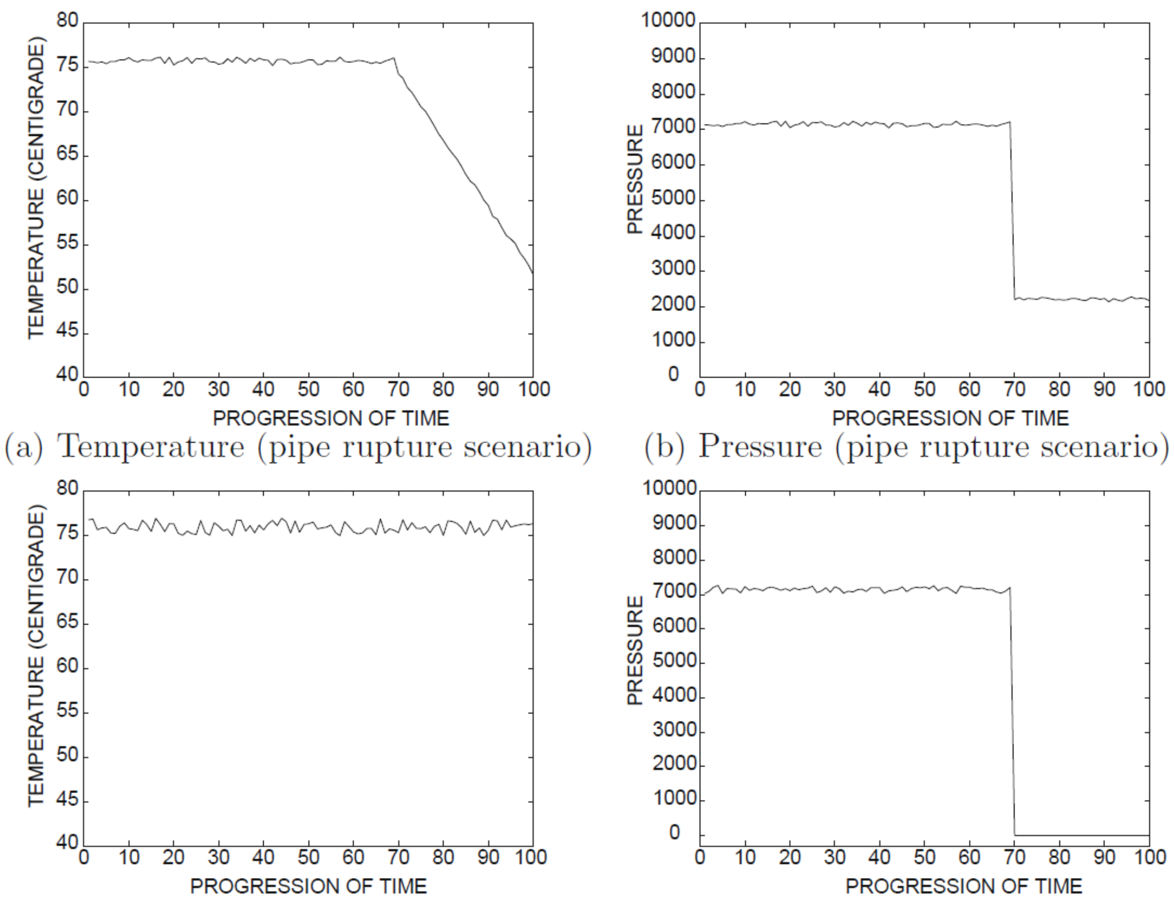
<u>Classification</u> is the association of an label.
1) Point label
2) Whole series label



1) Point label classification (timestamp)

- **refferred as supervised event detection (with labels)**
- **Few rare class labels = events**
- **E.g. malfunction of the machine with unsual sensor reading**

- **Supervised method helps to remove the cause of the bad events**
- **Anomaly noise vs anomaly of interest. Differentiate among the deviations of the different behavioral attributes**



(a) Temperature (pipe rupture scenario) (b) Pressure (pipe rupture scenario)
(c) Temperature (sensor failure scenario) (d) Pressure (sensor failure scenario)

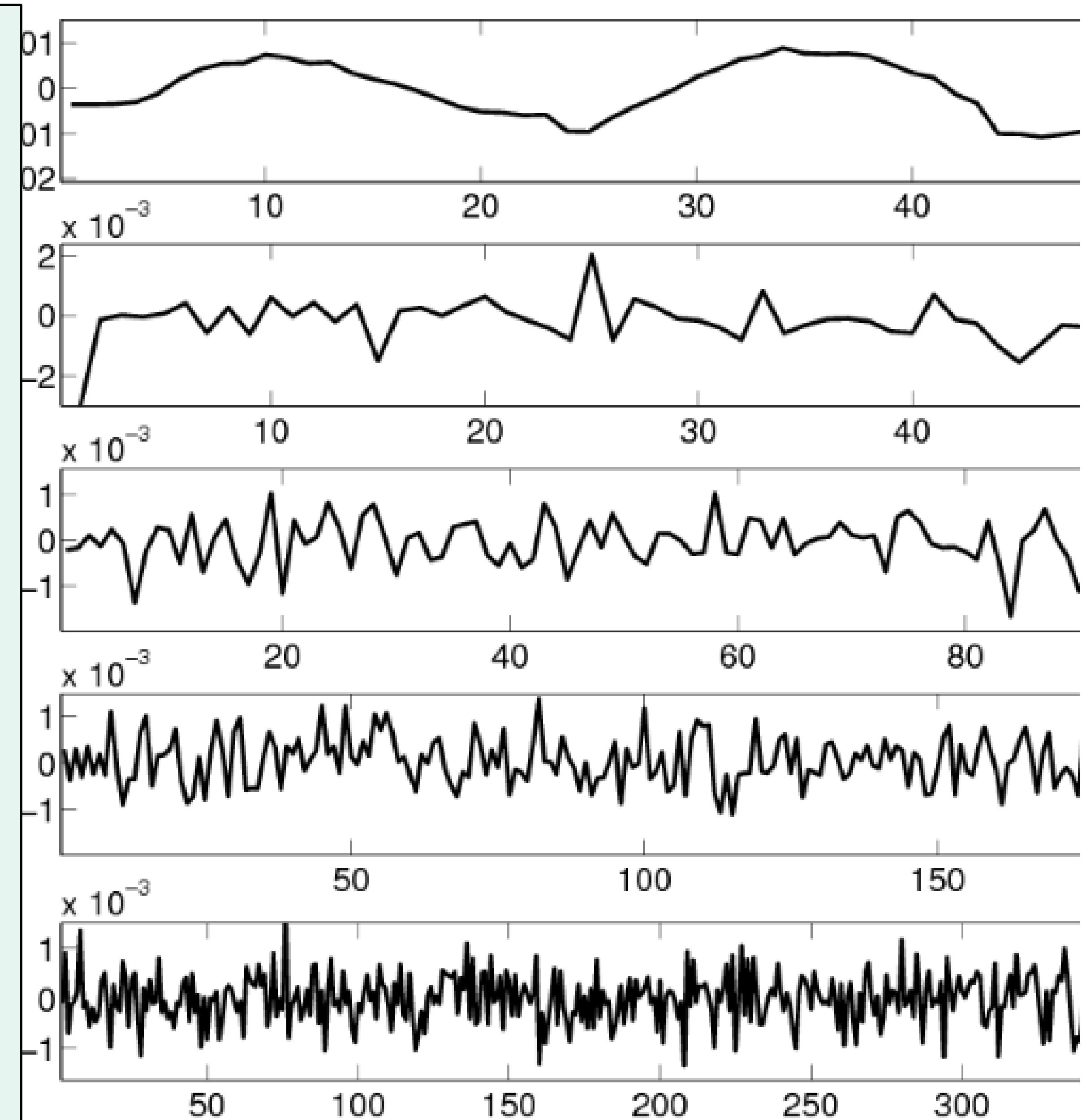
[Aggarwal Fig. 14.12]

Analytical Processing – Classification

2) Whole-series classification:
- **Shape based classification**
- **Use of distance based classifier**

a) Wavelet-based rules
- **→ multigranularity frequent trends**
- **Step1: Generate wavelet representation (or DFT)**
- **Step2: Discretize representation into a categorical representation**
- **Step3: Generate rule set using a rule based classifier method.**
- **Categorical values correspond to the signature shapes in the TS that are relevant to classification**

b) Nearest Neighbour Classifier
- **NNC with appropriate distance function (Euclidean, distance, DTW)**



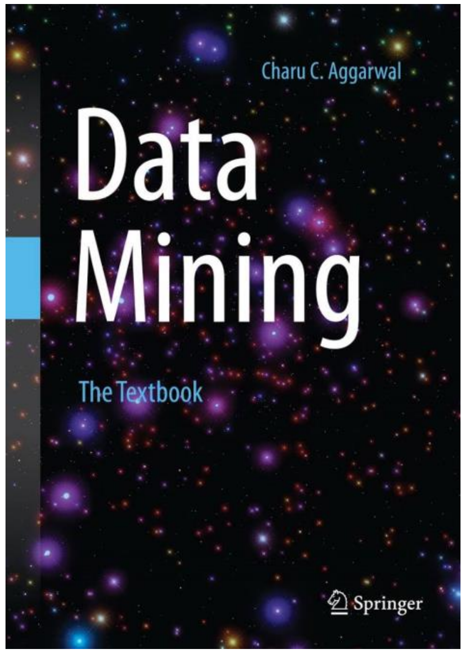
[http://inspirehep.net/record/1221474/plots]

Of DWT

Literature

The content and structure of this seminar presentation is based on:

C. Aggarwal. Data Mining The Textbook. Springer, 2015.
Chapters 1, 2, 3 and 14.



For additional reading:

T. Mitsa. Temporal Data Mining. CRC Press, 2010.

Thank you for your attention.

Now Q & A.