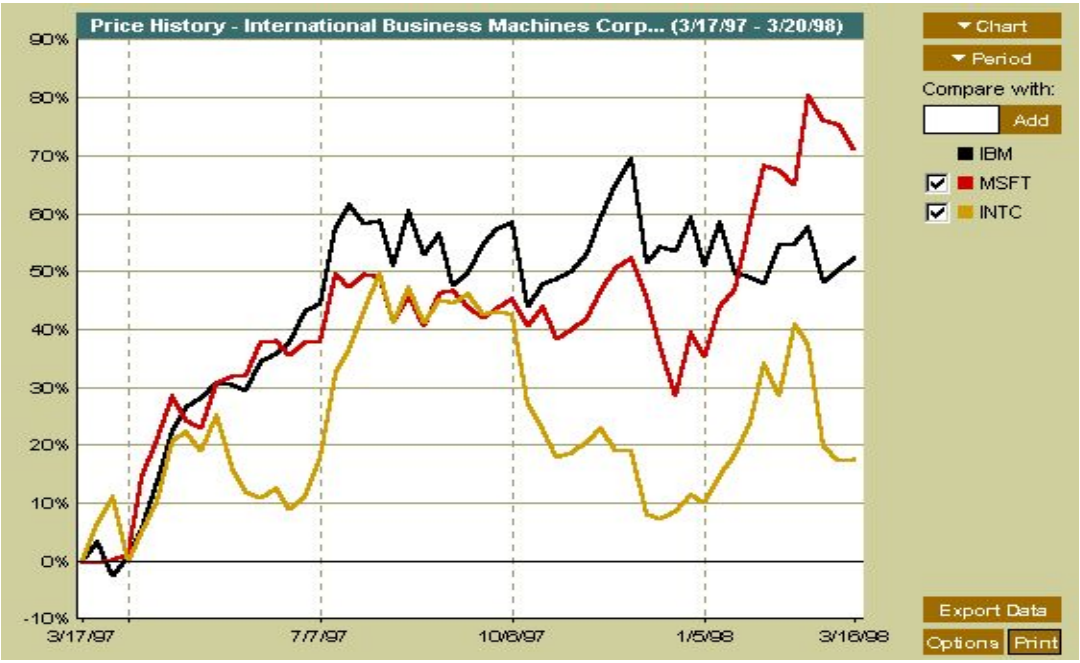
Mining Time-Series Data

Time-Series Database

- Consists of sequences of values or events obtained over repeated measurements of time (weekly, hourly…)
  - Stock market analysis, economic and sales forecasting, scientific and engineering experiments, medical treatments etc.
- Can also be considered as a Sequence database
  - A sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time.
  - Examples : Web page Traversal, Customer shopping transaction sequences
- Time-Series data can be analyzed to:
  - Identify correlations within time-series data
  - Analyze huge data to find similar / regular patterns, trends, outliers, bursts

Trend Analysis



Time Series involving a variable Y can be represented as a function of time t, Y = F(t)

Goals of Time-Series Analysis

- Modeling time series - To gain insight into the mechanism or underlying forces that generate the time series
- Forecasting time series - To predict the future values of the time-series variables

Trend Analysis

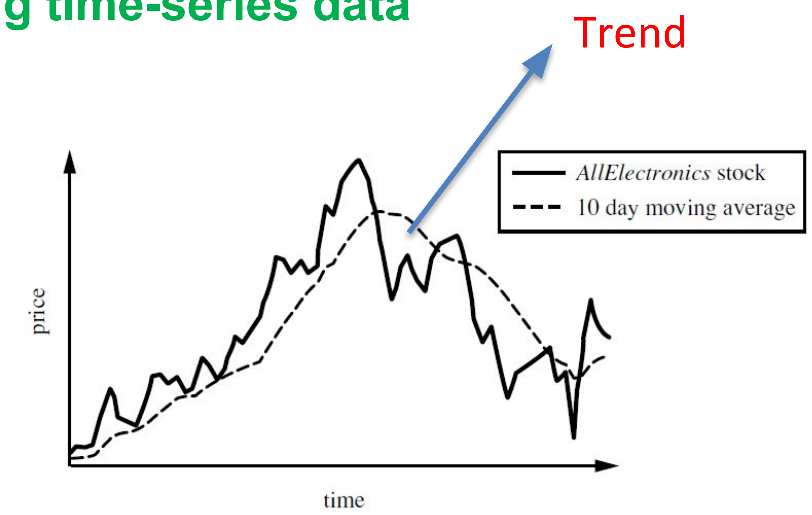
- Trend Analysis Major Components / Movements for Characterizing time-series data

Long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time

Typical methods for determining a trend curve or trend line include the weighted moving average method and the least squares method

Cyclic movements or cycle variations: long term oscillations about a trend line or curve
e.g., business cycles, may or may not be periodic

The cycles need not necessarily follow exactly similar patterns after equal intervals of time.



Trend Analysis

- **Trend Analysis Major Components / Movements for Characterizing time-series data**

  Seasonal movements or seasonal variations

  i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  Ex: sudden increase in sales of department store items before Christmas.

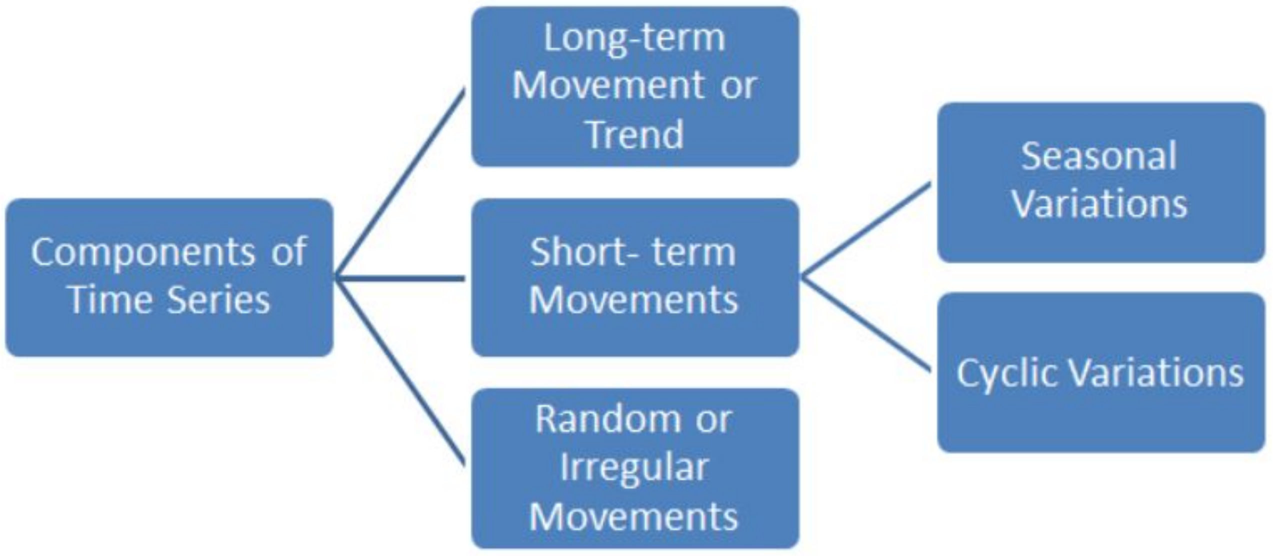
  Irregular or random movements - labor disputes, floods, or announced personnel changes within companies

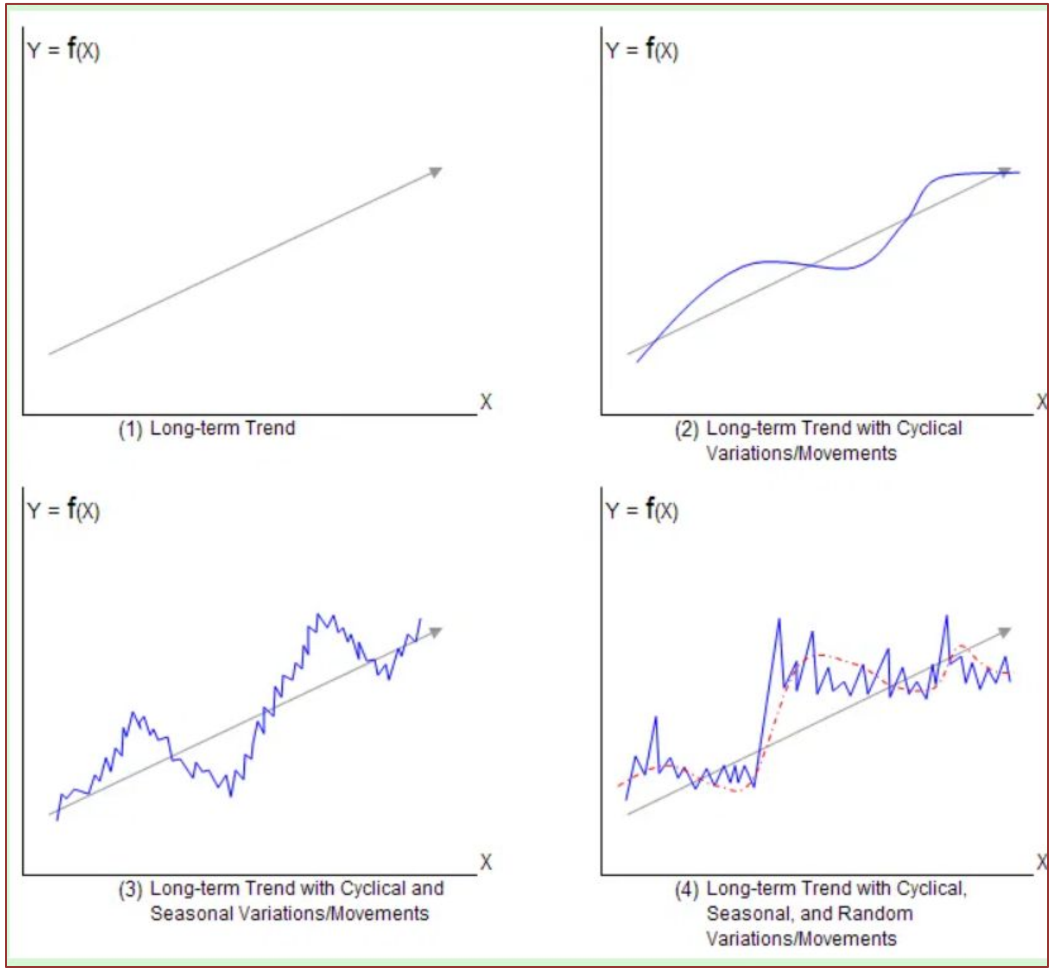
- Time series analysis: decomposition of a time series into these four basic movements

  Additive Model: TS = T + C + S + I

  Multiplicative Model: TS = T × C × S × I

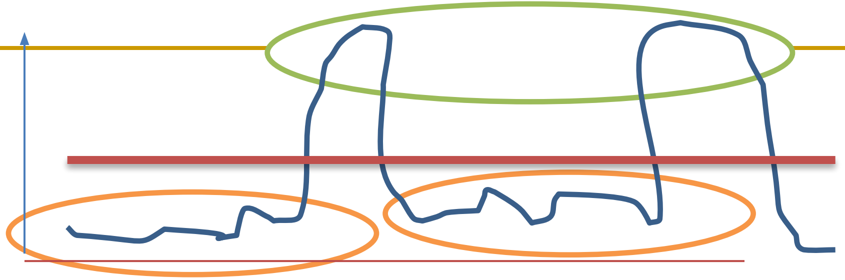
Trend Analysis



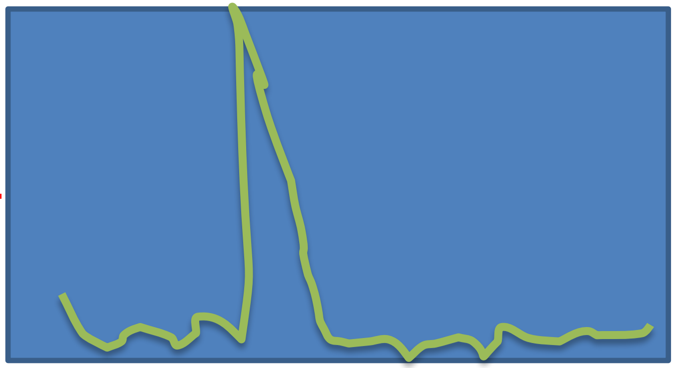


https://itfeature.com/time-series-analysis-and-forecasting/components-of-time-series

Trend Analysis



- Adjusting Seasonal fluctuations
  - Given a series of measurements y_1, $y_{,2}$, y_3… influences of the data that are systematic / calendar related must be removed
    - Fluctuations conceal true underlying movement of the series and non-seasonal characteristics
    - De-seasonalize the data (or *adjusted for seasonal variations*)
  - Seasonal Index – set of numbers showing the relative values of a variable during the months of a year
    - Sales during Oct, Nov, Dec – 80%, 120% and 140% of average monthly sales – Seasonal index – 80, 120, 140
    - Dividing original monthly data by seasonal index – De-seasonalizes data
- Auto-Correlation Analysis
  - To detect correlations between ith element and (i-k)th element k-
  - can be used between $<y_1, y_2, \ldots y_{N-k}>$ and $<y_{k+1}, y_{k+2}, \ldots y_N>$



Trend Analysis

Are there other ways to estimate the trend?

- Estimating Trend Curves
  - The freehand method
    - An approximate curve or line is drawn to fit a set of data based on the user's own judgement
    - Costly and barely reliable for large-scaled data mining
  - The least-square method
  - The moving-average method

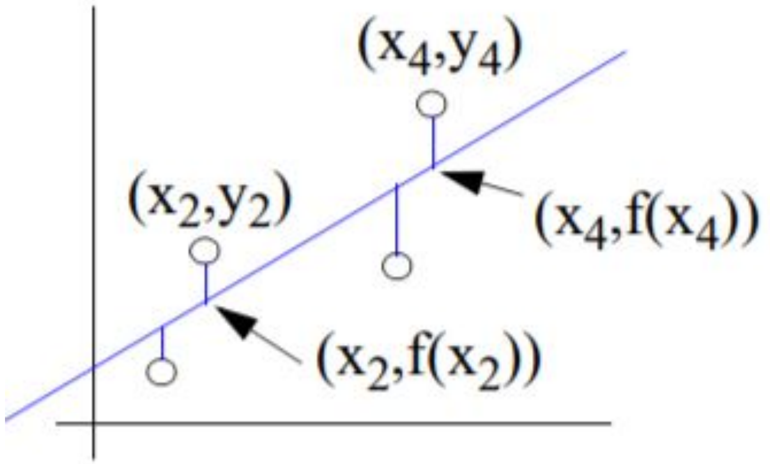
Trend Analysis

Are there other ways to estimate the trend?

- **Estimating Trend Curves**
  - The least-square method
    - Find the curve minimizing the sum of the squares of the deviation d_i of points y_i on the curve from the corresponding data points

$$\sum_{i=1}^{n} d_i^2$$

$$err = \sum_{i=1}^{\#\,\text{data points}} (y_i - f(x_i))^2 = \sum_{i=1}^{\#\,\text{data points}} (y_i - (ax_i + b))^2$$



Estimate the Temp (10 Years) given precipitation
Model is represented using Equation Temp
New input data (Precipitation) Temperature

Trend Analysis

- Moving Average Method

"How can we determine the trend of the data?"

The process of replacing the time series by its moving average eliminates unwanted fluctuations - referred to as the smoothing of time series

A common method for determining trend is to calculate a moving average of order *n* as

$$\frac{y_1 + y_2 + \cdots + y_n}{n}, \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \cdots$$

Temp
Jan 1 2, 3, Feb 1, 2…
Jan 1, 2,3,4,5 6
Jan 2, 3,4,5,6 7
Jan 3, 4,5,6,7 8

A moving average tends to reduce the amount of variations present in the dat

It smoothes the data
Eliminates cyclic, seasonal and irregular movements

Loses the data at the beginning or end of a series

If weighted arithmetic means are used, the resulting sequence is called a weighted moving average of order *n*

Sensitive to outliers (can be reduced by Weighted Moving Average)

Assigns greater weight to center elements to eliminate smoothing effects

Trend Analysis

"How can we determine the trend of the data?"

- Weighted Moving Average Method

If weighted arithmetic means are used, the resulting sequence is called a weighted moving average of order n

- Loses the data at the beginning or end of a series
- Sometimes generate cycles or other movements that are not present in the original data; and
- may be strongly affected by the presence of extreme values
- The influence of extreme values can be reduced by employing a weighted moving average with appropriate Weights
- An appropriate moving average can help smooth out irregular variations in the data

Trend Analysis

"How can we determine the trend of the data?"

- Moving Average Method – Example

$$\frac{y_1 + y_2 + \cdots + y_n}{n}, \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \cdots$$

Given a sequence of nine values, we can compute its moving average of order 3, and its weighted moving average of order 3 using the weights (1, 4, 1).

Original data:	3	7	2	0	4	5	9	7	2
Moving average of order 3:		4	3	2	3	6	7	6	
Weighted (1, 4, 1) moving average of order 3:		5.5	2.5	1	3.5	5.5	8	6.5	

The weighted average typically assigns greater weights to the central elements in order to offset the smoothing effect

Trend Analysis

- An appropriate moving average will smooth out the irregular variations. This leaves us with only cyclic variations for further analysis
- Once trends are detected data can be divided by corresponding trend values
- Cyclic Variations can be handled using Cyclic Indexes

Time-Series Forecasting

- Finds a mathematical formula that will approximately generate the historical patterns in a time series
- Used to make Long term / Short term predictions of future values
- Several models are available for forecasting:

Popular Method : ARIMA – Auto-Regressive Integrated Moving Average (also known as the Box-Jenkins method)

- Powerful, complex, quality of results depends on the User's level of experience

Similarity Search

Input S	Test x
TS1	Subsequences
TS2	

- Normal database query finds exact match
  - Similarity search finds data sequences that differ only slightly from the given query sequence

Two categories of similarity queries

Whole matching: find a sequence that is similar to the query sequence

Subsequence matching: find all pairs of similar sequences

Given a set of time-series sequences, *S*,
there are two types of similarity searches: *subsequence matching* and *whole sequence matching*.

Subsequence matching finds the sequences in *S* that contain ***subsequences*** that are similar to a given query sequence *x*, while whole sequence matching finds a set of sequences in *S* that are similar to each other (as a whole).

Similarity Search

- Typical Applications
  - Financial market
  - Market basket data analysis
  - Scientific databases
  - Medical diagnosis

Data Reduction and Transformation

- Time Series data – high-dimensional data – each point of time can be viewed as a dimension
- Dimensionality Reduction techniques
  - Signal Processing techniques
    - Discrete Fourier Transform
    - Discrete Wavelet Transform
    - Singular Value Decomposition based on PCA
  - Random projection-based Sketches
  - Time Series data is transformed and strongest coefficients – features
  - Techniques may require values in Frequency domain
    - Distance preserving Ortho-normal transformations
    - The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain

1. Due to the tremendous size and high-dimensionality of time-series data, *data reduction* often serves as the first step in time-series analysis.
2. Data reduction leads to not only much smaller storage space but also much faster processing

PCA

Image Dataset = 10 x 10
Dim = 100 points (real numbers)
PCA
Input Eigen values
Pick only the highest eigen values
Ex: 50 eigen values (50 dim)

100 50 dimensions
Used as input features

Attribute Selection

Dataset = Temp, outlook, Humidity, Windy (4)

Approach – only the relevant and important attributes
Final = Outlook, Windy (2)

4 2 dimensions
Used as input attributes

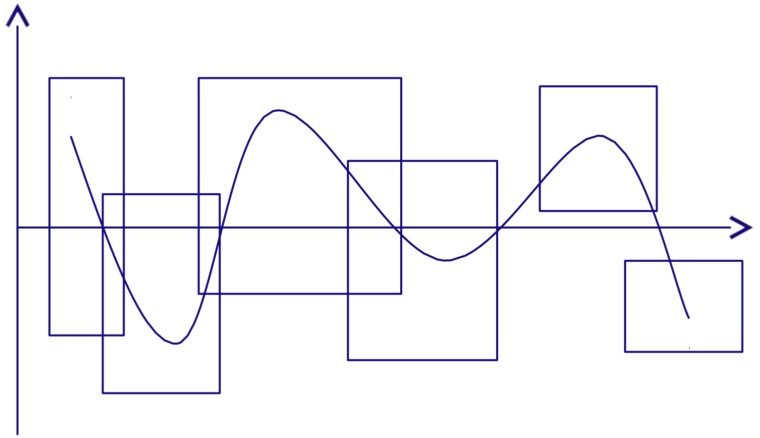
Indexing methods for Similarity Search

"Once the data are transformed by DFT, how can we provide support for efficient search in time-series data?"

- Multi-dimensional index – can be constructed using the first few Fourier coefficients
  - Use the index to retrieve the sequences that are at most a certain small distance away from the query sequence
  - Perform post-processing by computing the actual distance between sequences in the time domain and discard any false matches
- Indexing techniques
  - R-trees, R*-trees, Suffix trees etc

Subsequence Matching

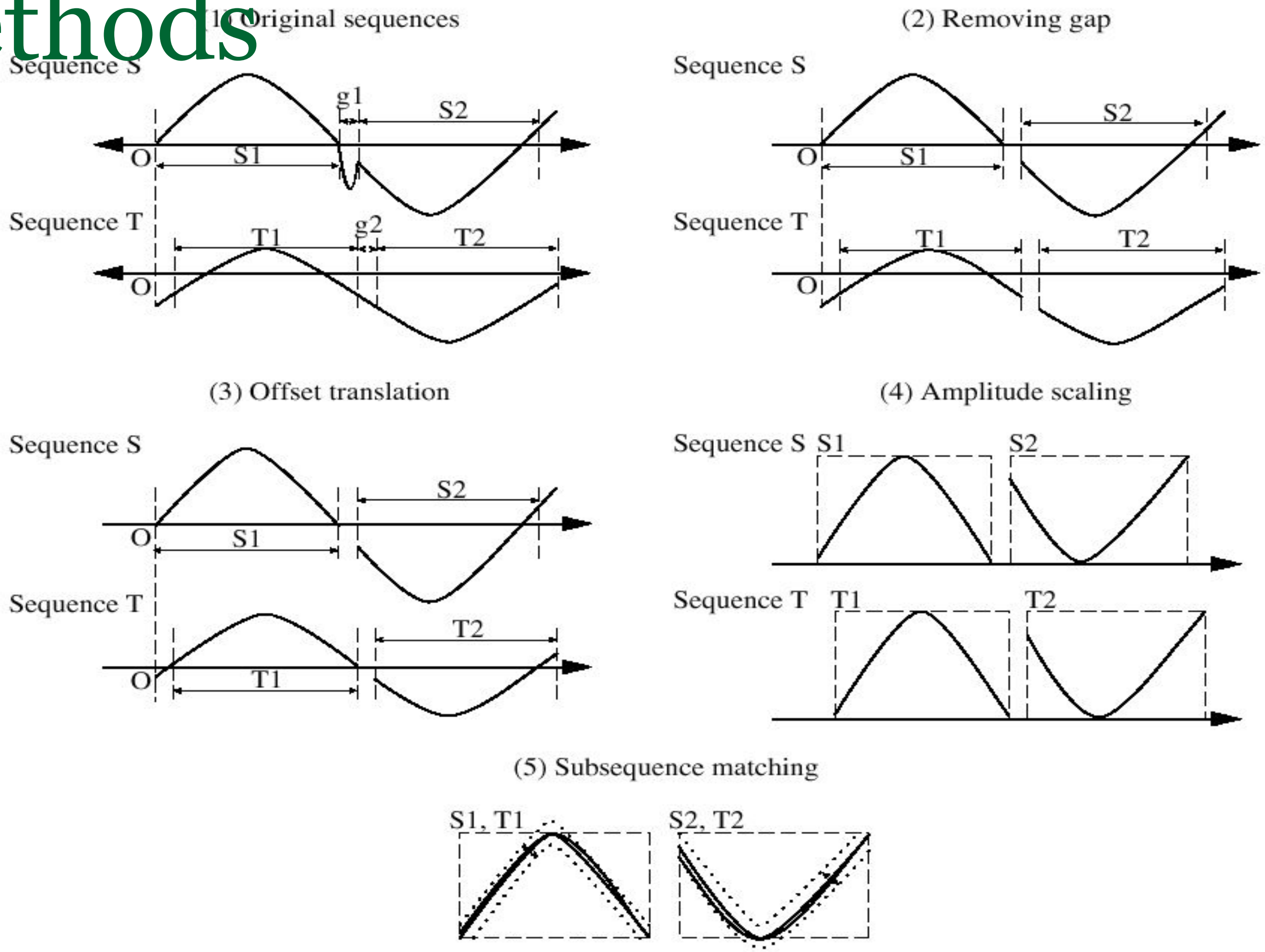
- Break each sequence into a set of pieces of window with length *w*
- Extract the features of the subsequence inside the window
- Map each sequence to a "trail" in the feature space
- Divide the trail of each sequence into "subtrails" and represent each of them with minimum bounding rectangle
- Use a multi-piece assembly algorithm to search for longer sequence matches
- Uses Euclidean distance (Sensitive to outliers)



Similarity Search Methods

- Practically there maybe differences in the baseline and scale
  - Distance from one baseline to another – offset
  - Data has to be normalized
  - Sequence X = $<x_1, x_2, ..x_n>$ can be replaced by X' = $<x_1', x_2', \ldots x_n'>$ where $x_i' = x_i - \mu / \sigma$
  - Two subsequences are considered similar if one lies within an envelope of ε width around the other, ignoring outliers
  - Two sequences are said to be similar if they have enough non- overlapping time-ordered pairs of similar subsequences Parameters specified by a user or
  - expert: sliding window size, width of an envelope for similarity, maximum gap, and matching fraction

Similarity Search Methods



Similarity Search Method

- Atomic matching
  - Find all pairs of gap-free windows of a small length that are similar
- Window stitching
  - Stitch similar windows to form pairs of large similar subsequences allowing gaps between atomic matches
- Subsequence Ordering
  - Linearly order the subsequence matches to determine whether enough similar pieces exist

Query Languages for Time Sequence

- Time-sequence query language
  - Should be able to specify sophisticated queries like
    - Find all of the sequences that are similar to some sequence in class *A*, but not similar to any sequence in class *B*
  - Should be able to support various kinds of queries: range queries, all- pair queries, and nearest neighbor queries

- Shape definition language
  - Allows users to define and query the overall shape of time
  - sequences Uses human readable series of sequence transitions or
  - macros Ignores the specific details
    - E.g., the pattern up, Up, UP can be used to describe increasing degrees of rising slopes
    - Macros: spike, valley, etc.