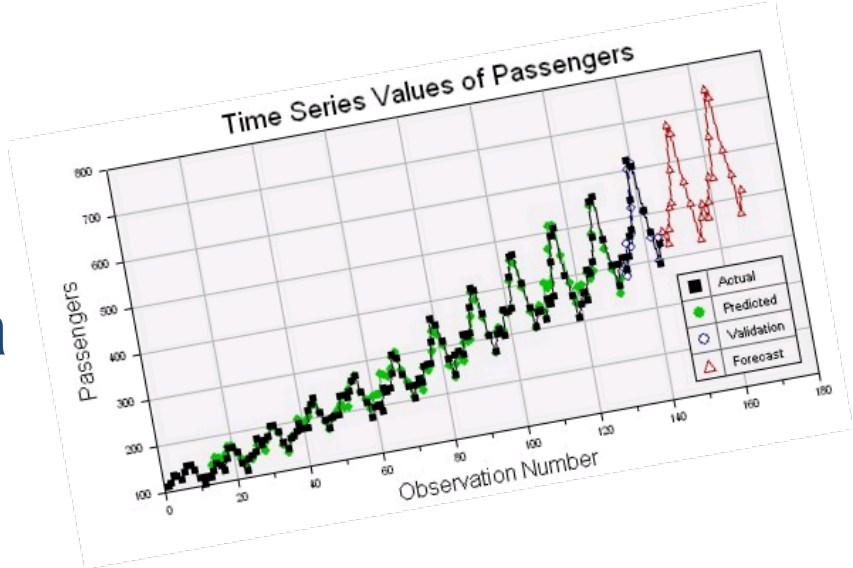
Welcome to Lecture 08

AI503: Advanced Machine Learning

Summary – last week

- Last week:
  - Mining Sequence Patterns

- This week:
  - Time Series Data



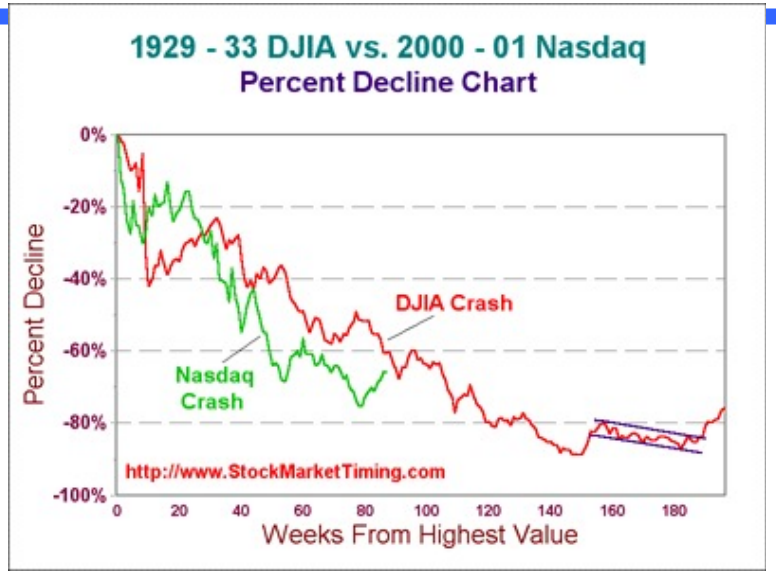
Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Time-Series Data

- **Time-series databases**
  - Time series reveal **temporal behavior** of the underlying mechanism that produced the data
  - Consists of sequences of values or events **changing with time**
  - Data is recorded at **regular intervals**

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Time-Series Data



- Applications
  - Financial
    - Stock market, sales forecasting, inflation
  - Industry
    - Power consumption, workload projections, process and quality control
  - Meteorological
    - Observation of natural phenomena such as precipitation, temperature, wind, earthquakes

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Time-Series Data

- **Goals** of time-series data analysis
  - **Modeling** time-series
    - Get insight into the mechanisms or underlying forces that generate the time series
  - **Forecasting** time-series
    - Predict the future values of the time-series variables
- Methods
  - Trend analysis
  - Similarity search

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- **Trend analysis**
  - Application of **statistical techniques** e.g., regression analysis, to make and justify statements about trends in the data
  - Construct a **model**, independent of anything known about the physics of the process, to explain the behavior of the measurement
    - E.g., increasing or decreasing trend, that can be statistically distinguished from random behavior: take daily average temperatures at a given location, from winter to summer

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- **Regression analysis** (RA)
  - Popular tool for modeling time series, finding trends and outliers in data sets
  - Analysis of numerical data consisting of values of a **dependent variable** (also called a response variable) and of one or more **independent variables**
    - The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants“) and an error term

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Regression Analysis

- RA, example: determine appropriate **levels of advertising** for a particular market segment
  - Consider the problem of managing sales of Coca Cola at large college campuses
    - Sales over one semester might be influenced by ads in the college paper, ads on the campus radio station, sponsorship of sports-related events, sponsorship of contests, etc.
    - Use data on advertising and promotional expenditures at many different campuses to extract the marginal value of dollars spent in each category

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

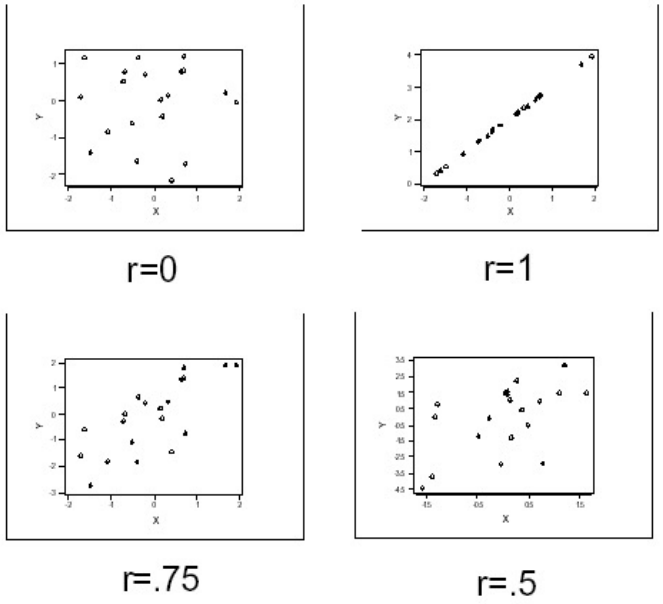
Regression Analysis

- Set up a model of the following type:
  - sales = b_0 + b_1(print budget) + b_2(radio budget) + b_3(sports promo budget) + b_4(other promo) + error
- This model is called **linear regression analysis**
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$
  - Y = predicted score
  - b_0 = intercept/origin of regression line
  - b_i = regression coefficient representing unit of change in dependent variable with the increase in 1 unit on X variable. The values of these coefficient can be calculated using Ordinary Least Square method.

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Regression Analysis

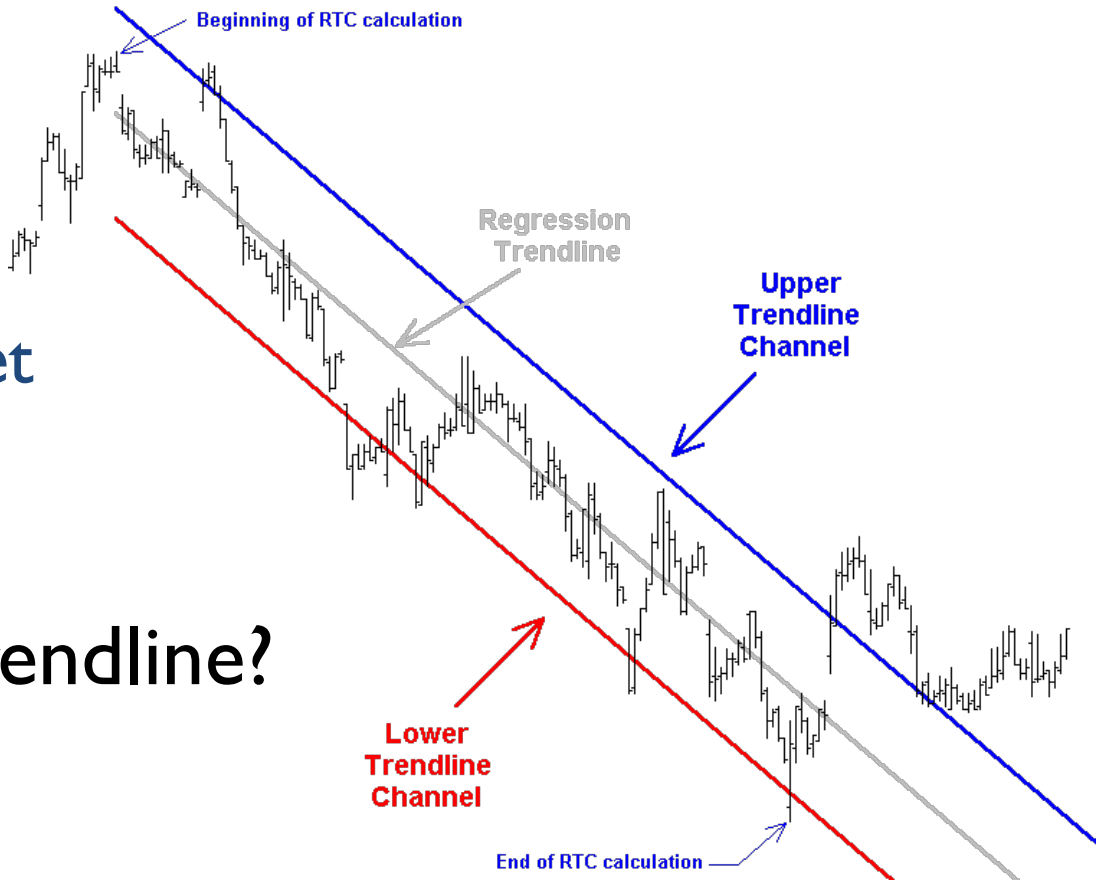
- Correlation (noted R)
  - Refers to the interdependence or co-relationship of Variables
  - The correlation is denoted by Pearson Correlation Coefficient (PCC).
  - Reflects the closeness of the linear relationship between X andY
  - Lies between -1 and 1 with



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Regression Analysis

- Regression trend channels (RTC)
  - Very useful in defining and containing the trend of the market
  - When the prices break a well established trend channel, the market usually changes trend
- Upper & Lower trendline?



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Regression Analysis

- What is RTC?
  - The mathematical **standard deviation** of the linear
  - regression
  - Basically it is made up of three parallel lines
    - The center line is the linear regression line
    - This center line is bracketed by two additional lines that represent the +/- standard deviation of the linear regression data

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Regression Analysis

- The linear regression model is the most simple
- model, but there are others
  - **Nonlinear** regression (the model function is not linear in the parameters), **Bayesian** methods, etc.
- Regression analysis can't capture all trend movements that occur in real-world applications
  - The solution is to **decompose** time-series into **basic movements**
- **Basic movements?**

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Characteristic **time-series movements** (components)
  - Trend (T)
    - Reflects the long term progression of the series
  - Seasonal (S)
    - Seasonal fluctuations i.e., almost identical patterns that a time series appears to follow during corresponding months of successive years
  - Cycle (C)
    - Describes regular fluctuations caused by the economic cycle e.g., business cycles
  - Irregular (I)
    - Describes random, irregular influences

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Time-series **decomposition**
  - Additive Modal
    - Time-series =T + C + S + I
  - Multiplicative Modal
    - Time-series =T × C × S × I
- To perform decomposition we must identify each of the 4 movements in the time-series

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- **Trend analysis (T)**, methods
  - The **freehand** method
    - Fit the curve by looking at the graph
    - **Costly** and barely reliable for large-scaled data mining
  - The **least-square** method
    - Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points
  - The **moving-average** method
    - Eliminates cyclic, seasonal and irregular patterns
    - Loss of end data
    - Sensitive to outliers

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

– **Moving average** (MA) of order n

- E.g.,

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

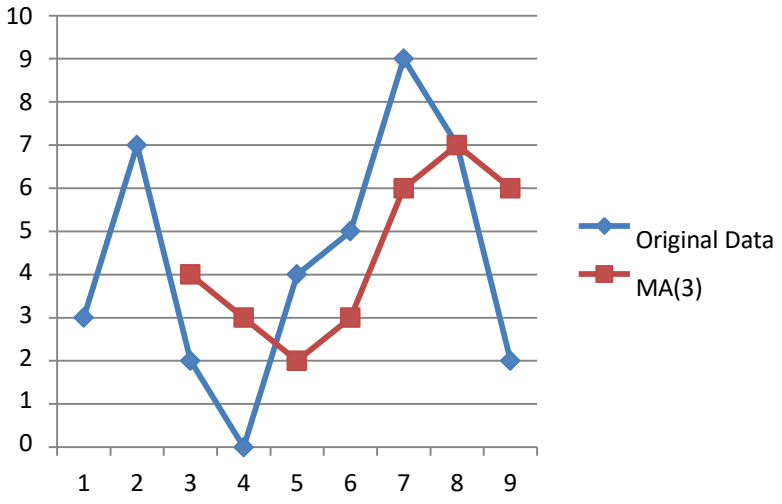
Original data
3
7
2
0
4
5
9
7
2

(3+7+2)/3

(7+2+0)/3

...

MA(3)
4
3
2
3
6
7
6



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Moving average

– Influence of extreme values can be reduced with **weighted moving average** (WMA)

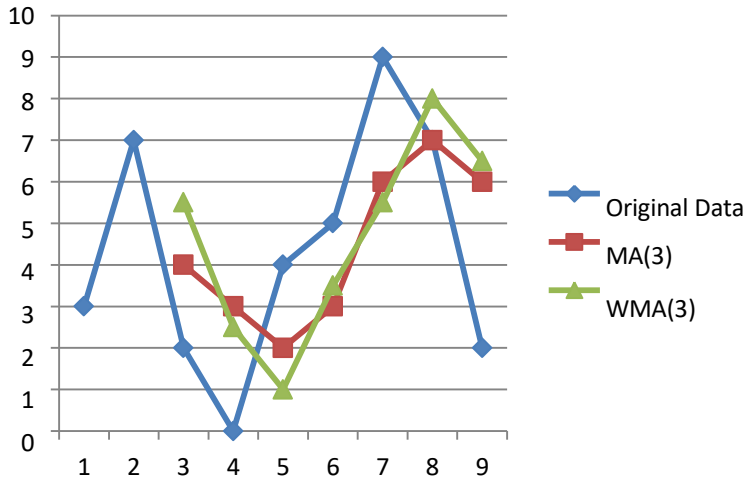
- WMA is MA with weights e.g., WMA(3) with (1,4,1) as weights

Original data	WMA(3)
3	
7	
2	5.5
0	2.5
4	1
5	3.5
9	5.5
7	8
2	6.5

(3*1+7*4+2*1)/(1+4+1)

(7*1+2*7+0*1)/(1+4+1)

...



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Moving average

- Other forms of MA
  - **Cumulative moving average** (CA), also called long running average

$$CA_i = \frac{x_1 + \cdots + x_i}{i}.$$

$$CA_{i+1} = CA_i + \frac{x_{i+1} - CA_i}{i+1}.$$

  - **Exponential weighted moving average** (EWMA), applies weighting factors which decrease exponentially
    - Gives much more importance to recent observations while still not discarding older observations entirely

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

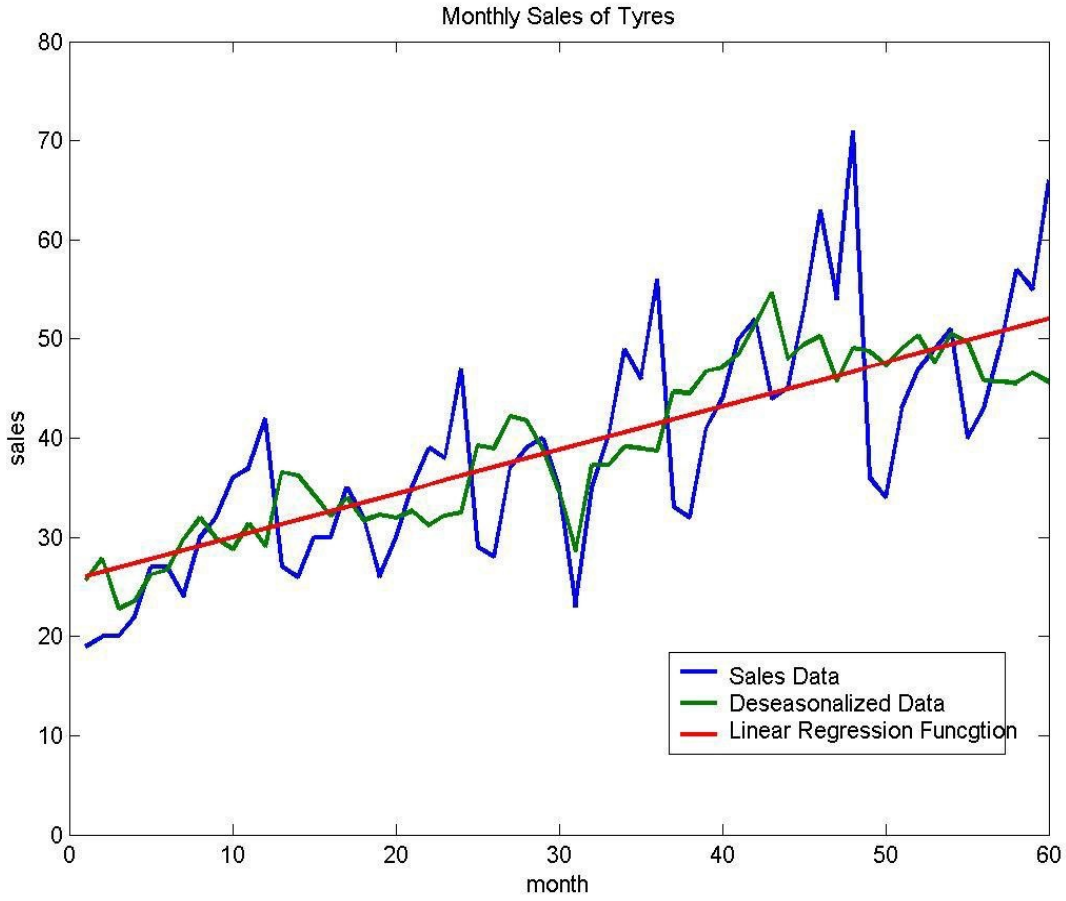
Trend Analysis

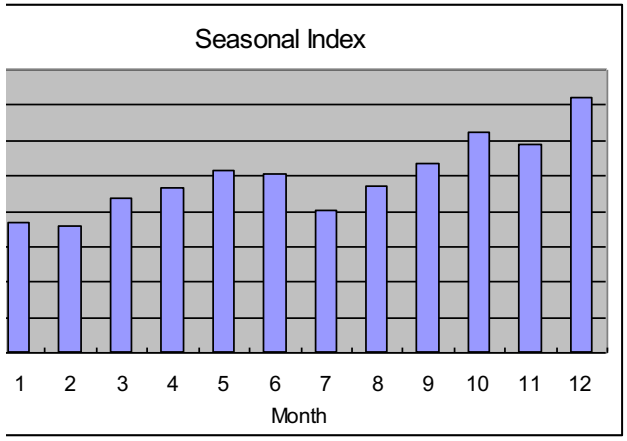
- Estimation of **seasonal variations (S)**
  - Seasonal index
    - Set of numbers showing the relative values of a variable during the months of the year
    - E.g., if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months
  - Deseasonalized data
    - Data adjusted for seasonal variations
    - E.g., **divide** the original monthly data by the seasonal index numbers for the corresponding months

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Estimation of **seasonal variations (S)**





Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Estimation of **cyclic variations (C)**
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes
- Estimation of **irregular variations (I)**
  - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make **long- or short-term predictions** (time-series forecasting) with reasonable quality

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Time-series **forecasting**
  - Finds a mathematical formula that will approximately generate the historical patterns
  - Forecasting models: most popular, **auto-regressive integrated moving average** (ARIMA)
    - ARIMA can be applied in cases where data show evidence of non-stationarity

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

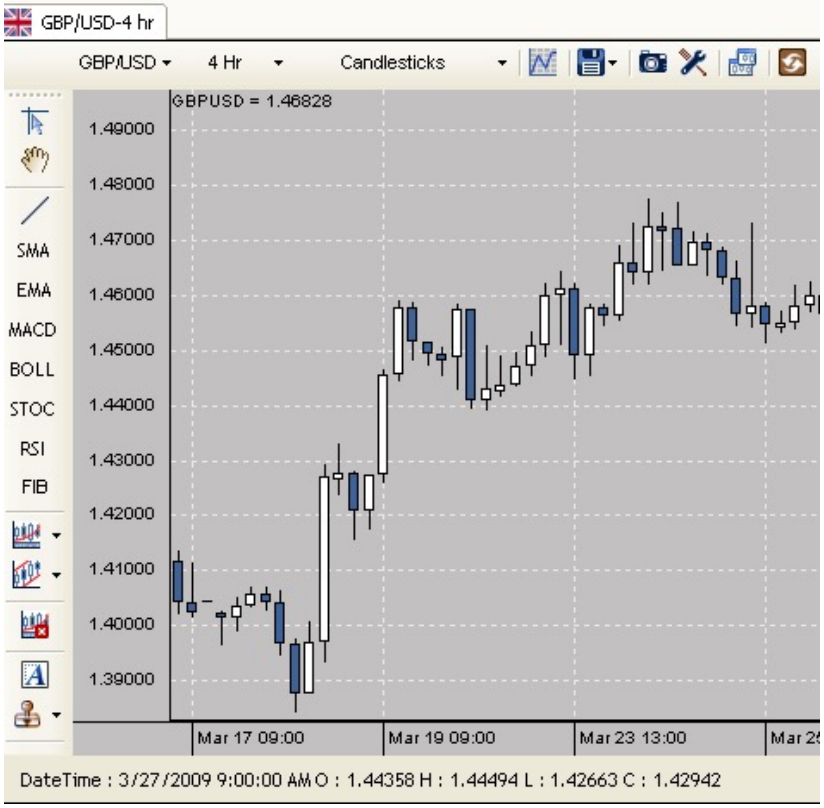
Trend Analysis

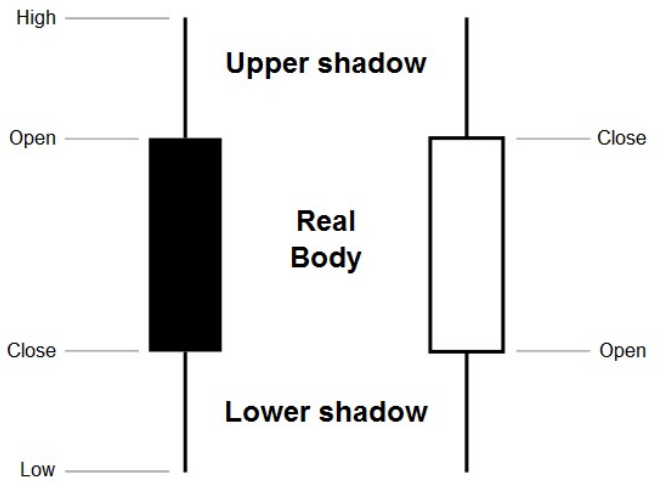
- Applications of trend analysis
  - Foreign exchange market (FOREX)
    - High data volume
    - Small granularity
  - Transform data to an adequate granularity e.g., 4 hours a candle for FOREX E.g. Walmart Currency change

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Granularity change
  - Use Japanese candlesticks for data representation





Trend Analysis

- Simple moving average for trend analysis
  - E.g., SMA with window size of 21 bars

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Why do we need trends?
  - Once we have found a trend, we can trade
    - **Open position when in the trend** (buy if it will go up, or sell if it will go down)

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- **Close position on trend turns:** detect turns with Bollinger bands, resistance lines, etc.

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- Bollinger bands
  - Calculated based on the moving average
    - N standard deviations up, N down
  - Useful for detection of over-buy and over-sell

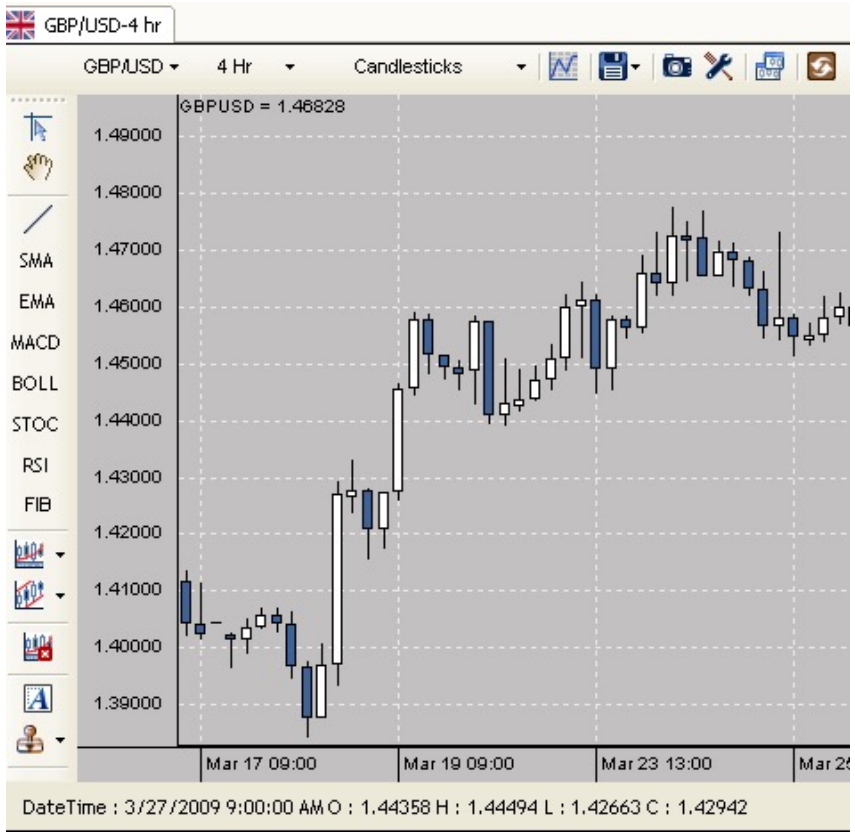
Trend Analysis

- Psychological pressure of the market
  - **Resistance lines** are determined by the reaction of the market participants to the previous evolution of the data

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Trend Analysis

- And there are many more indicators for **in the trend** and **on trend turns**
  - E.g., **momentum** analysis



Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Similarity Search

- **Similarity search**
  - Normal database queries find exact matches
  - Similarity search finds data sequences that **differ only slightly** from the given query sequence
- Problem: given a time-series database, identify all the sequences that are **similar** to one another

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Similarity Search

- Typical **applications**
  - Financial market
    - Finding stock items with similar trends
  - Market basket
    - Finding products with similar sales trends
  - Scientific databases
    - Finding periods with similar temperature patterns, finding persons with similar voice clips

Acknowledgment - Thanks to Wolf-Tilo Balke and Silviu Homoceanu - TU Braunschweig for the slides

Summary

- Time Series Data

Next week

- Classification