

Data Mining

Munawar, PhD

4. Mining The Data





Agenda

01 Mining Sequence Patterns

02 Mining Time Series Data

03 QA?



Mining Sequence Pattern



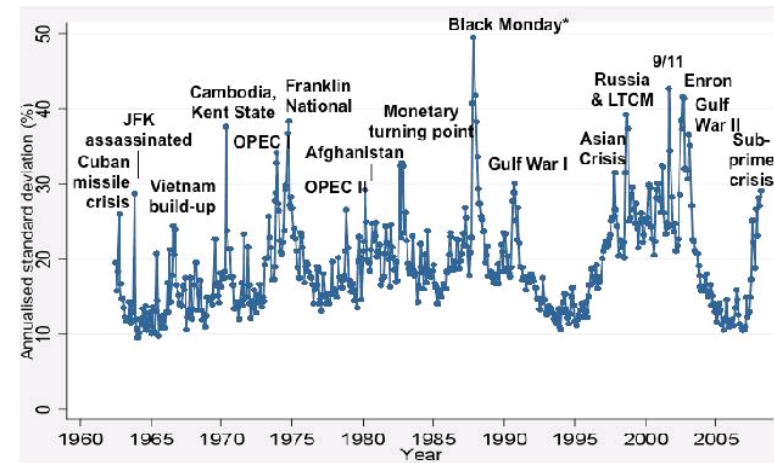
Mining Sequence Patterns

- Sequential pattern mining
 - Mining of frequently occurring ordered events or subsequences as patterns
 - Example
 - Customers who buy helicopter models in some on-line store receive e-mail promotions
 - Regarding batteries
 - Then after a while regarding rotor wings, since most of them will break



Mining Sequence Patterns

- Sequential pattern mining
 - Applications
 - Customer retention, targeted marketing
 - Ranging from disasters (e.g. earthquakes, wars) to market prediction



Mining Sequence Patterns

- Mining sequence patterns, vocabulary
 - Let $I = \{I_1, I_2, \dots, I_p\}$ be the set of all items
 - An **itemset** is a nonempty set of items from I
 - A **sequence S** is an ordered list of events
 - Denoted $\langle e_1 e_2 e_3 \dots e_k \rangle$, where event e_1 occurs before e_2 etc.
 - An **event** is an itemset, i.e. an unordered list of items
 - E.g., $(I_2 I_1 I_3)$, where $I_1, I_2, I_3 \in I$

Mining Sequence Patterns

- E.g., a customer bought items (abc) at a store. This is an event e_1 . Now if later he buys another itemset (ade), representing a second event e_2 , we obtain a shopping sequence s
 - $e_1=(abc), e_2=(ade)$
 - $s=\langle e_1e_2 \rangle = \langle (abc)(ade) \rangle$
- The number of instances of items in a sequence is called the **length** of the sequence
 - Length of s is 6
- A sequence with length k is called a **k -sequence**

Mining Sequence Patterns

– Subsequence & supersequence

- A sequence $\alpha = \langle a_1 a_2 \dots a_n \rangle$ is called a **subsequence** of another sequence $\beta = \langle b_1 b_2 \dots b_m \rangle$ denoted $\alpha \sqsubseteq \beta$ (β is called **supersequence** of α) if there exist integers $1 \leq j_1 < j_2 < \dots < j_n \leq m$ such that
$$a_1 \subseteq b_{j_1}, a_2 \subseteq b_{j_2}, \dots, a_n \subseteq b_{j_n}$$
- E.g., if $\alpha = \langle (ab)d \rangle$ and $\beta = \langle (abc)(de) \rangle$ then $\alpha \sqsubseteq \beta$

– Sequence database

- A sequence database S is a set of tuples $\langle \text{SID}, s \rangle$
- E.g., contains the sequences for all customers of the store

Mining Sequence Patterns

– **Support of a sequence** in a sequence database

- The support of α in S is the number of tuples in S , containing α
- $\text{sup}_S(\alpha) = |\{\langle \text{SID}, s \rangle \mid (\langle \text{SID}, s \rangle \in S) \wedge (\alpha \sqsubseteq s)\}|$

– **Frequent sequence**

- α is a frequent sequence if $\text{sup}_S(\alpha) \geq \text{min_sup}$, where min_sup is the **minimum support threshold**

– A frequent sequence is called a **sequence pattern**

- A sequence pattern of length k is called an **k-pattern**

Mining Sequence Patterns

- Sequence patterns, example

- Given

- $I = \{a, b, c, d, e, f, g\}$, $\text{min_sup} = 2$
and the sequence table

- Length of $\langle a(abc)(ac)d(cf) \rangle$ is 9 and although there are 3 'a' items in the first 3 events from record 1, it contributes to the $\text{sup}(a)$ with just 1

SID	Sequence
1	$\langle a(abc)(ac)d(cf) \rangle$
2	$\langle (ad)c(bc)(ae) \rangle$
3	$\langle (ef)(ab)(df)cb \rangle$
4	$\langle eg(af)cbc \rangle$

Mining Sequence Patterns

- Sequence patterns, example

- $\langle a(bc)df \rangle$ is a subsequence of the first record

- $\langle a(bc)df \rangle \sqsubseteq \langle a(abc)(ac)d(cf) \rangle$

- $\text{sup}(\langle (ab)c \rangle) = 2$

- $\langle (ab)c \rangle \sqsubseteq \langle a(abc)(ac)d(cf) \rangle$ and
 $\langle (ab)c \rangle \sqsubseteq \langle (ef)(ab)(df)cb \rangle$

- If $\text{min_sup} = 50\%$, $\langle (ab)c \rangle$ is a **sequential pattern** or a **3-pattern** (i.e. it has length 3)

SID	Sequence
1	$\langle a(abc)(ac)d(cf) \rangle$
2	$\langle (ad)c(bc)(ae) \rangle$
3	$\langle (ef)(ab)(df)cb \rangle$
4	$\langle eg(af)cbc \rangle$

Mining Sequence Patterns

- **Challenges** of sequence pattern mining
 - A huge number of possible sequential patterns are hidden in databases
 - A mining algorithm should
 - Find the **complete set of patterns**, when possible, satisfying the minimum support threshold
 - Be **highly efficient**, scalable, involving only a small number of database scans
 - Be able to incorporate various kinds of user-specific constraints

Mining Sequence Patterns

- **Algorithms**

- Apriori-based method

- Generalized Sequential Patterns (GSP)

- Pattern-growth methods

- FreeSpan & PrefixSpan

- Vertical format-based mining

- Sequential Pattern Discovery using Equivalent classes (SPADE)

- Mining closed sequential patterns

- CloSpan



Mining Sequence Patterns

- **Generalized Sequential Patterns (GSP)**

- Based on the Apriori property of sequential patterns

- Downward closure: If a sequence s is not frequent then **none** of its super-sequences can be frequent
- E.g., let $\text{min_sup}=2$; if $\langle hb \rangle$ is infrequent then $\langle hab \rangle$ and $\langle (ah)b \rangle$ are also infrequent!

$\langle hb \rangle$ is a subset of only record 3



SID	Sequence
1	$\langle (bd)cb(ac) \rangle$
2	$\langle (bf)(ce)b(fg) \rangle$
3	$\langle (ah)(bf)abf \rangle$
4	$\langle (be)(ce)d \rangle$
5	$\langle a(bd)bcb(ade) \rangle$

GSP

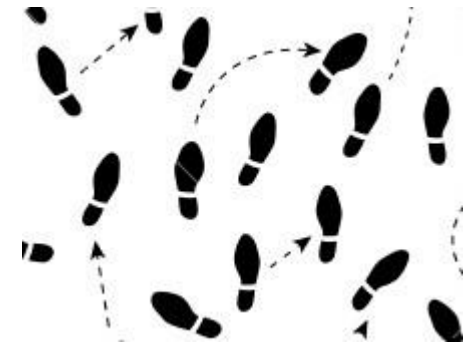
- GSP algorithm, 2 step description

- Initial step

- Every item in the sequence database is a candidate of length 1

- Generalization

- Scan database to collect support count for each k length, candidate sequence, and establish the k-patterns
- Generate candidate sequences of length (k+1) from k-patterns using the Apriori property
- Repeat this generalization step until no more candidates can be found e.g., there are no more k length frequent sequences



GSP

– Initial step

- All singleton sequences are $\langle a \rangle$, $\langle b \rangle$, $\langle c \rangle$, $\langle d \rangle$, $\langle e \rangle$, $\langle f \rangle$, $\langle g \rangle$, $\langle h \rangle$

– General step, $k = 1$

- Scan database once, count support for candidates
- $\langle g \rangle$ and $\langle h \rangle$ are not 1-patterns since $\text{sup}(\langle g \rangle) = 1 < \text{min_sup} = 2$ $\text{sup}(\langle h \rangle) = 1 < \text{min_sup} = 2$
- According to the Apriori property: since $\langle g \rangle$ and $\langle h \rangle$ are not 1-patterns, they can't form any 2-patterns. So they can be removed!

min_sup = 2

SID	Sequence
1	$\langle (bd)cb(ac) \rangle$
2	$\langle (bf)(ce)b(fg) \rangle$
3	$\langle (ah)(bf)abf \rangle$
4	$\langle (be)(ce)d \rangle$
5	$\langle a(bd)bcb(ade) \rangle$

Cand	Support
$\langle a \rangle$	3
$\langle b \rangle$	5
$\langle c \rangle$	4
$\langle d \rangle$	3
$\langle e \rangle$	3
$\langle f \rangle$	2
$\langle g \rangle$	1
$\langle h \rangle$	1



– General step, $k = 1$, generate length 2 candidates

- First generate 2 event candidates

– $6*6 = 36$
candidates

	<a>		<c>	<d>	<e>	<f>
<a>	<aa>	<ab>	<ac>	<ad>	<ae>	<af>
	<ba>	<bb>	<bc>	<bd>	<be>	<bf>
<c>	<ca>	<cb>	<cc>	<cd>	<ce>	<cf>
<d>	<da>	<db>	<dc>	<dd>	<de>	<df>
<e>	<ea>	<eb>	<ec>	<ed>	<ee>	<ef>
<f>	<fa>	<fb>	<fc>	<fd>	<fe>	<ff>

- Then generate 1 event candidates, each with 2 items

– $6*5/2 = 15$
candidates

	<a>		<c>	<d>	<e>	<f>
<a>		<(ab)>	<(ac)>	<(ad)>	<(ae)>	<(af)>
			<(bc)>	<(bd)>	<(be)>	<(bf)>
<c>				<(cd)>	<(ce)>	<(cf)>
<d>					<(de)>	<(df)>
<e>						<(ef)>
<f>						

GSP

– $k = 2$, we have 51 2-length candidates

- After the second table scan we remain with 19 2-patterns
- Then we generate candidates for length 3, and so on...
- $\langle (bd)cba \rangle$ is a 5-pattern, meaning that events (bd), c, b and a were frequent in the table, in this order

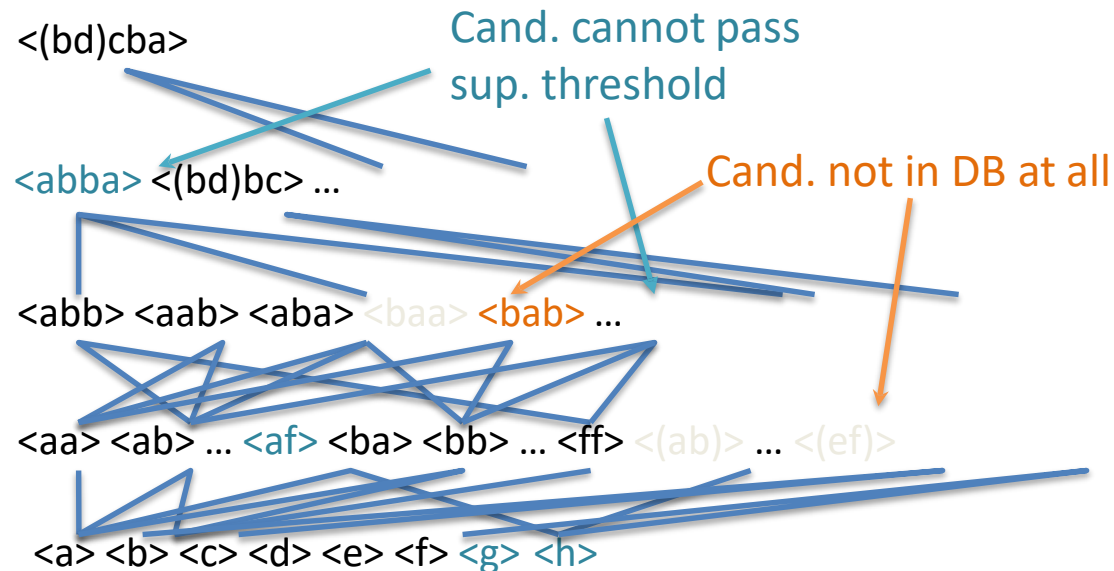
5th scan: 1 cand. 1 5-pattern

4th scan: 8 cand. 6 4-patterns

3rd scan: 47 cand. 19 3-patterns,
20 cand. not in DB at all

2nd scan: 51 cand. 19 2-patterns
10 cand. not in DB at all

1st scan: 8 cand. 6 1-patterns





- Drawbacks of GSP
 - A **huge set** of candidate sequences generated
 - Especially 2-item candidate sequence
 - Multiple scans of database needed
 - The length of each candidate grows by one for each database scan
 - Inefficient for mining long sequential patterns
 - Long patterns grow from short patterns
 - The number of short patterns is exponential in the length of mined patterns



Time Series Data

Time Series Data

- Sequence patterns mining
 - Are ordered events
 - No **concrete notion of time**
- Combining sequences of events with repeated measurements of **time** (at equal time intervals) we obtain **time-series** data



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**



Time Series Data

- Applications

- Financial

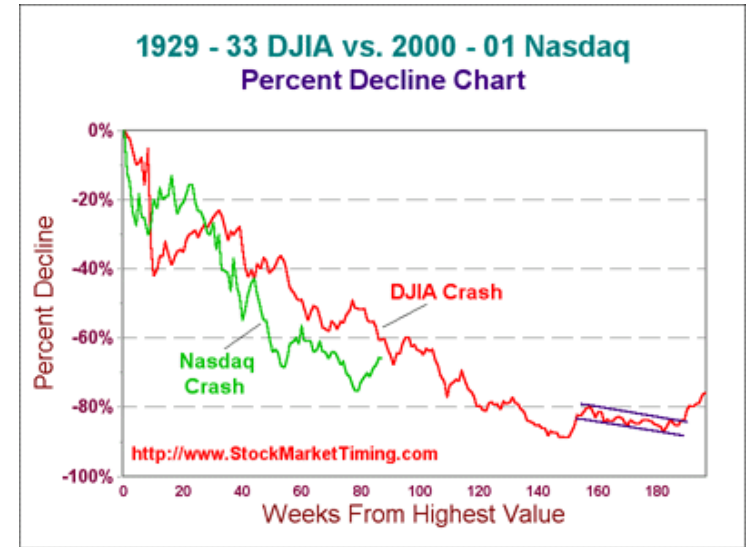
- Stock market, sales forecasting, inflation

- Industry

- Power consumption, workload projections, process and quality control

- Meteorological

- Observation of natural phenomena such as precipitation, temperature, wind, earthquakes



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



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

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

Regression Analysis

- RA, example: determine appropriate **levels of advertising** for a particular market segment
 - Consider the problem of managing sales of coffee 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

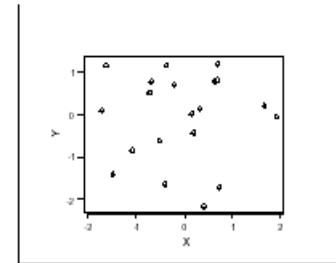
Regression Analysis

- Set up a model of the following type:
 - $\text{sales} = b_0 + b_1(\text{print budget}) + b_2(\text{radio budget}) + b_3(\text{sports promo budget}) + b_4(\text{other promo}) + \text{error}$
- This model is called **linear regression analysis**
 - $Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$
 - 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

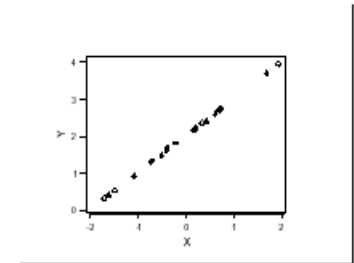
Regression Analysis

– Correlation (noted R)

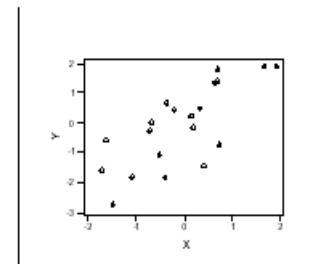
- Refers to the interdependence or co-relationship of variables
- Reflects the accuracy of the linear relationship between X and Y
- Lies between -1 and 1 with:
 - 1 is anti-correlated
 - 1 is independent
 - 2 is linearly correlated



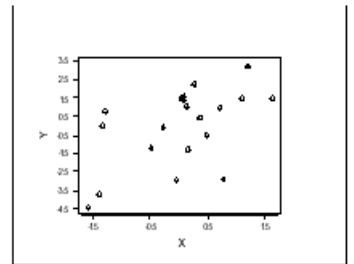
$r=0$



$r=1$



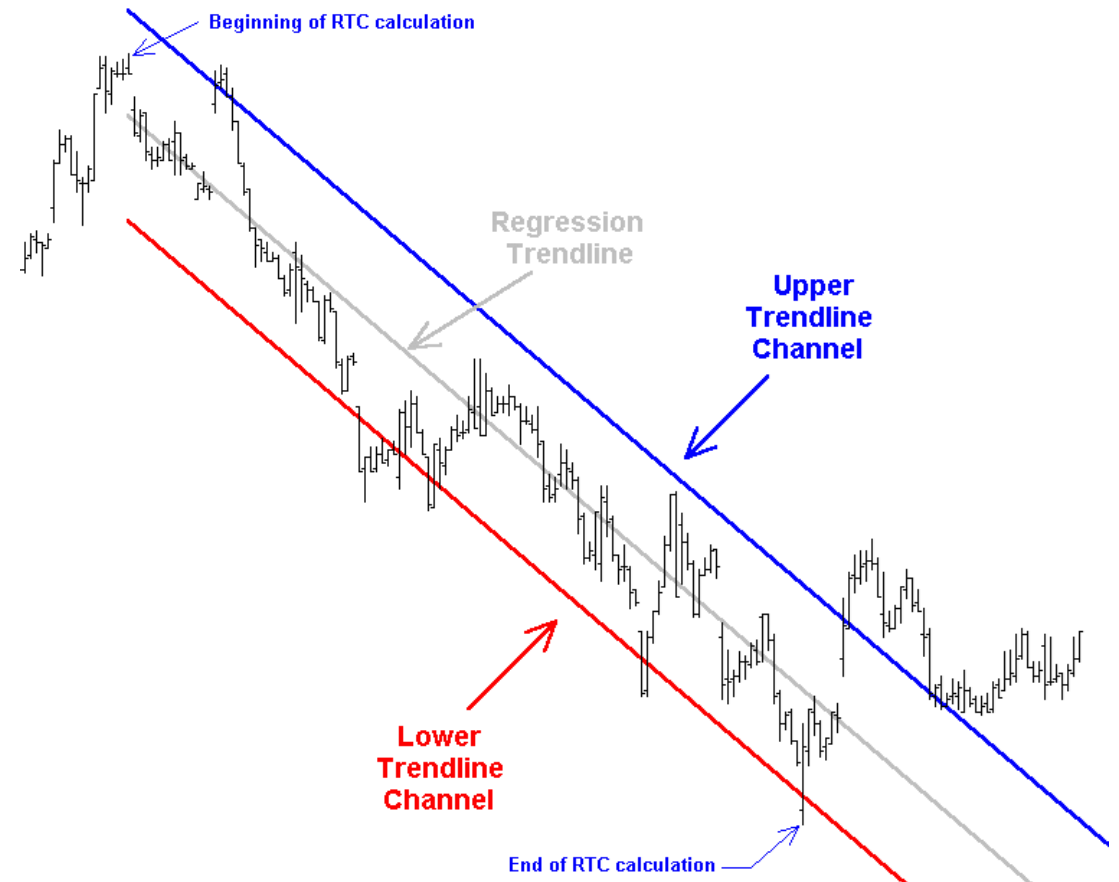
$r=.75$



$r=.5$

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?



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

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**

Trend Analysis

- Basic movements are characteristic **time-series movements** (often called **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

Trend Analysis

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

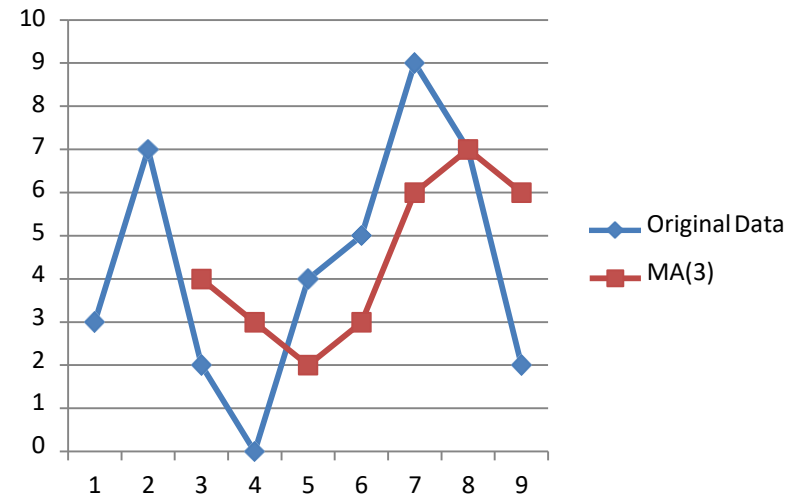
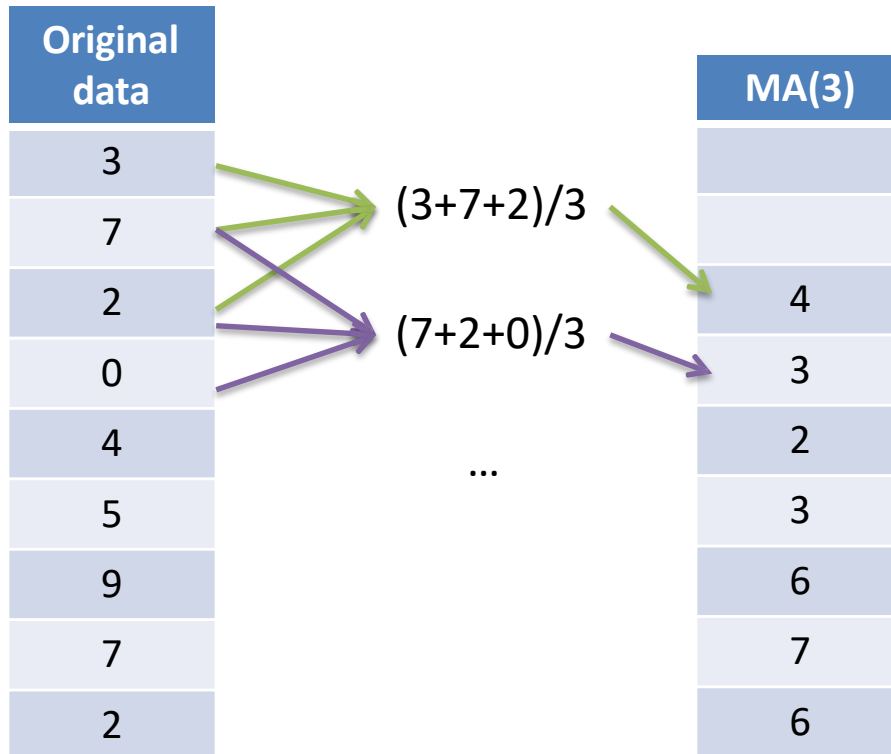
Trend Analysis

- **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

Trend Analysis

– Moving average (MA) of order n

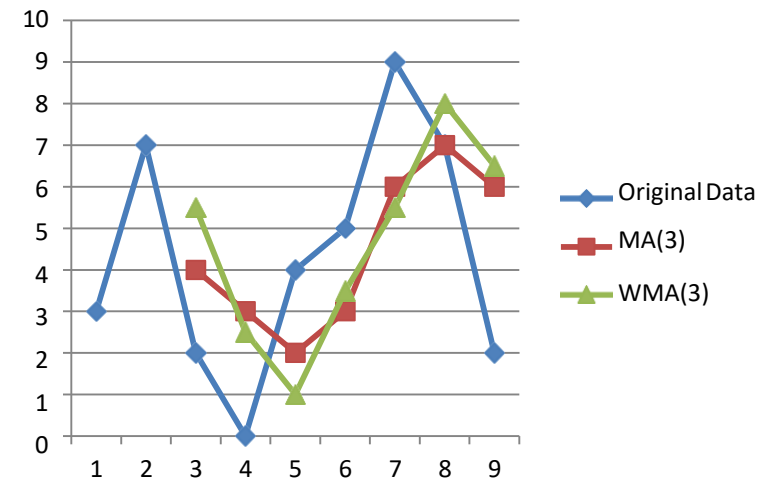
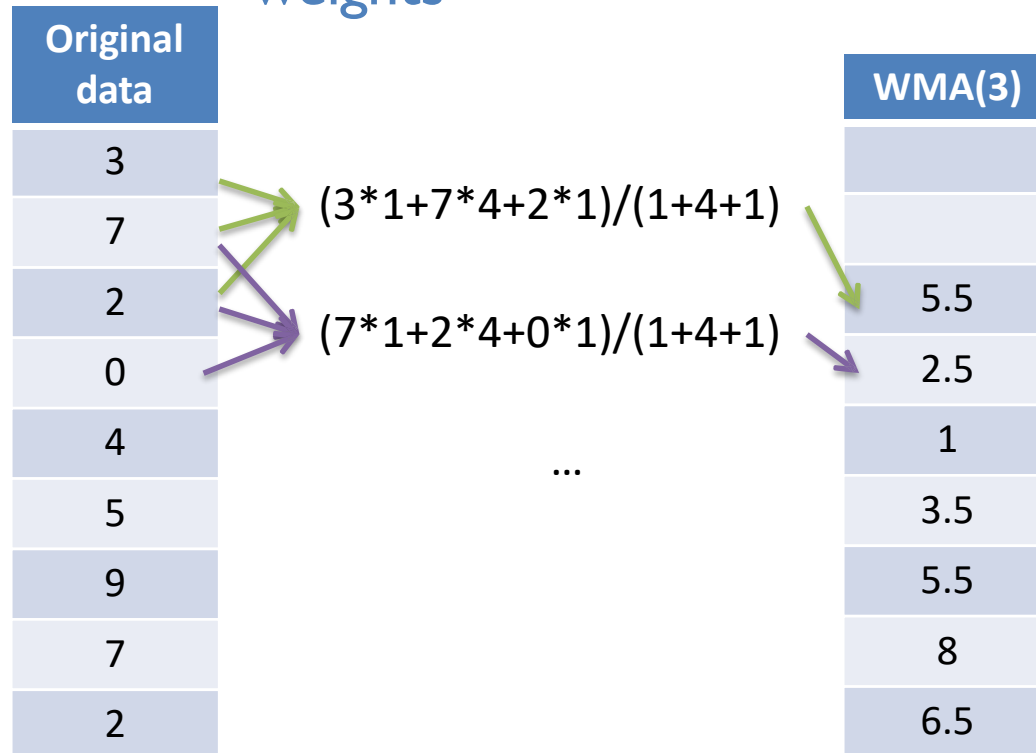
- E.g., $\frac{y_1 + y_2 + \dots + y_n}{n}$, $\frac{y_2 + y_3 + \dots + y_{n+1}}{n}$, $\frac{y_3 + y_4 + \dots + y_{n+2}}{n}$, ...





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

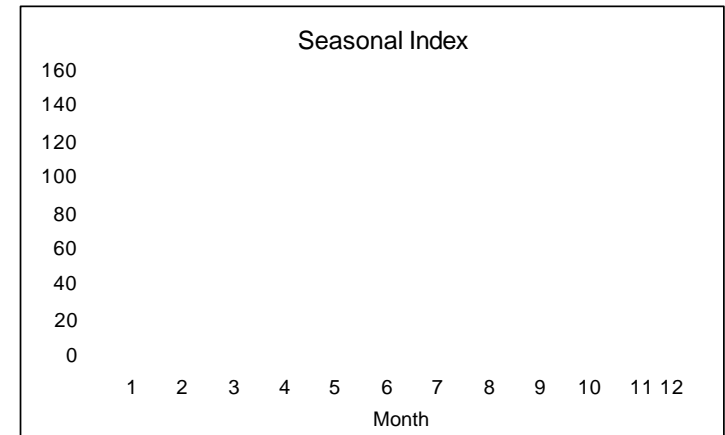
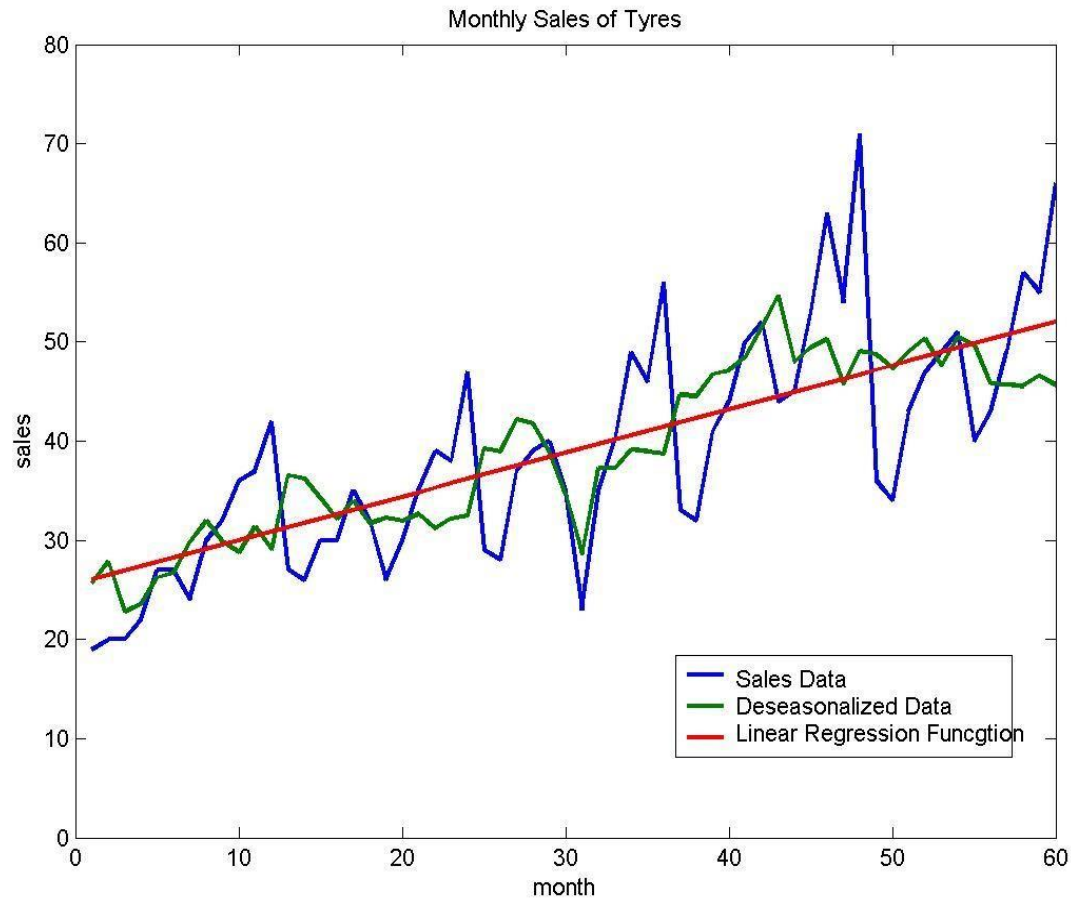


Trend Analysis

- 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 or subtract the original monthly data by the seasonal index numbers for the corresponding months

Trend Analysis

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



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

Trend Analysis

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



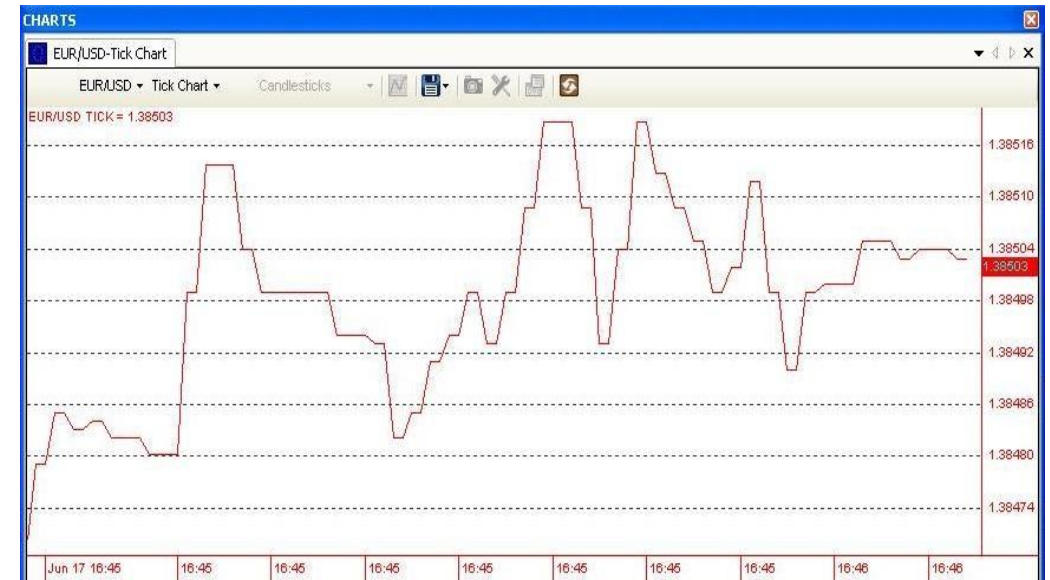
Trend Analysis

- Applications of trend analysis: large corporations selling their products world-wide
 - Products are sold in different countries with different currencies
 - Currency has to be exchanged back and forth
 - The cost of the currency exchange has to be kept under control!
 - Timing is everything in foreign exchange



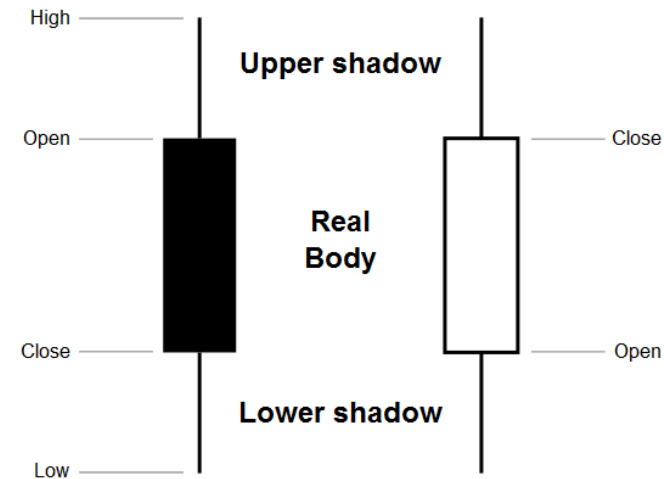
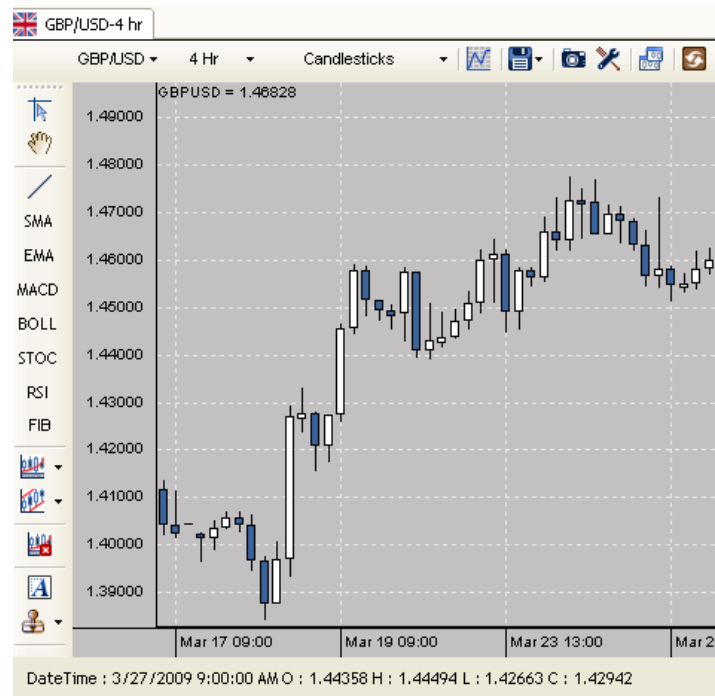
Trend Analysis

- Foreign exchange market (FOREX)
 - High data volume
 - Small granularity – ticks milliseconds away
- Transform data to an adequate granularity
e.g., 4 hours a candle for FOREX



Trend Analysis

- Granularity change
 - Use Japanese candlesticks (developed in the 16th century by Japanese rice traders) for data charting



Trend Analysis

- When trading the goal is to buy low and sell high!
 - **Use trends to trade!**



Trend Analysis

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



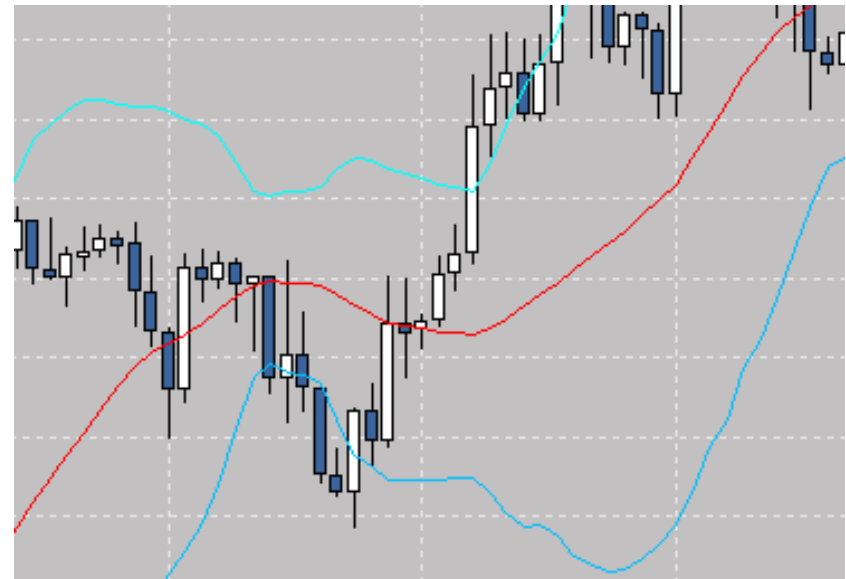
Trend Analysis

- Perform smoothing with simple moving average
 - E.g., SMA with window size of 21 bars
- **Trend:** k consecutive points on the SMA show constant increase or decrease on Y-axis



Trend Analysis

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



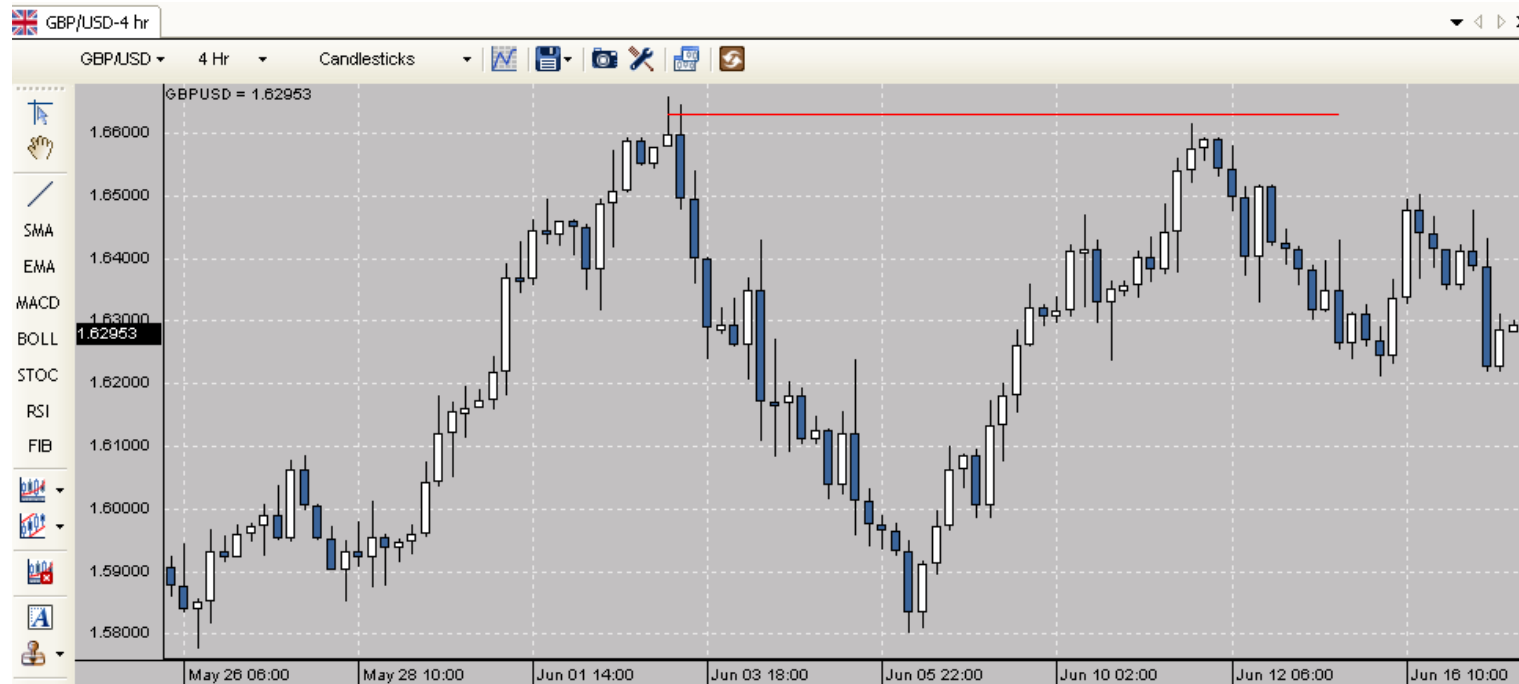
Trend Analysis

- Transactions...
 - between the lower band and the SMA show signs of **over-sell** and transactions
 - between SMA and upper band – **over-buy**
 - outside the Bollinger bands – **trend turn**



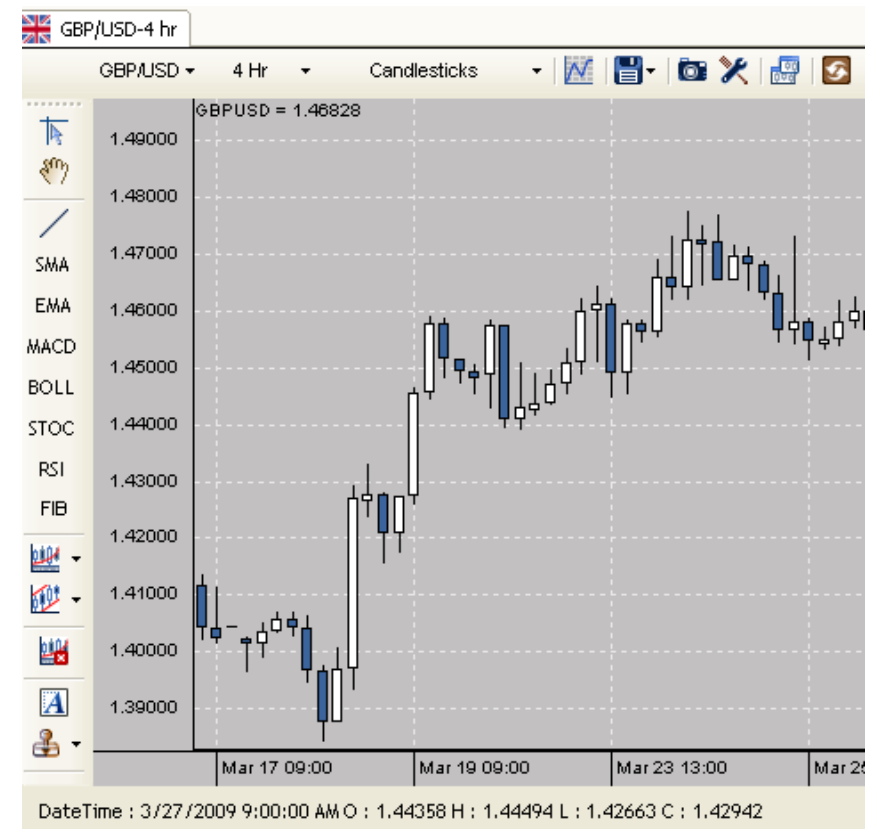
Trend Analysis

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



Trend Analysis

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





QA ???



THANK YOU

Munawar, PhD – moenawar@gmail.com – www.moenawar.web.id

