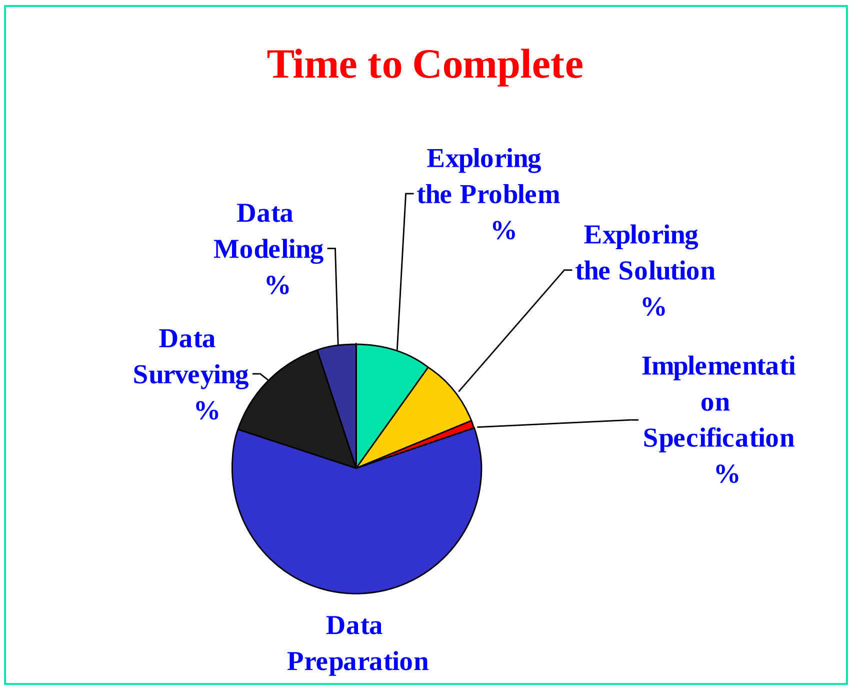
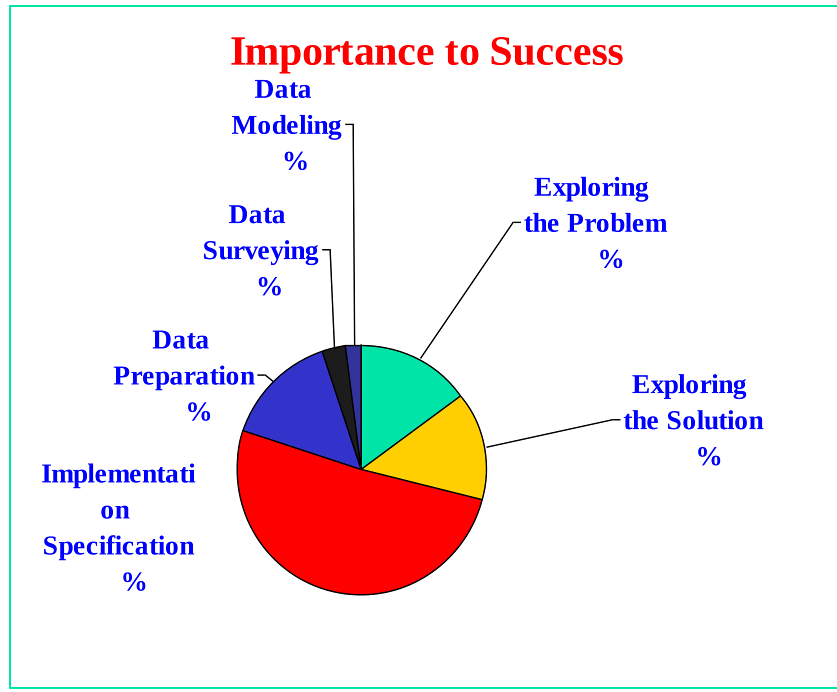
Lesson 2- Data Preparation and Data Engineering

- Stages of a Knowledge Discovery Project
- Data in Reality
- Problems in Data Accessibility
- Data Pre-processing
- Data Cleaning
- Preparation of Time Series Data

Stages of a Knowledge Discovery Project
(based on Pyle, 1999)





Data in Reality

- What DM Tools Need?
  - Data Availability
  - One static data table
  - Clear meaning of each attribute
  - Well-defined domain of each attribute
  - Values of all attributes
  - Data reliability
  - No duplicate information
  - Data consistency

- What we have?
  - Data is not readily accessible
  - Several tables / databases / data streams
  - Missing metadata
  - Out-of-range values
  - Missing values
  - Noisy data
  - Redundant information
  - Inconsistent data

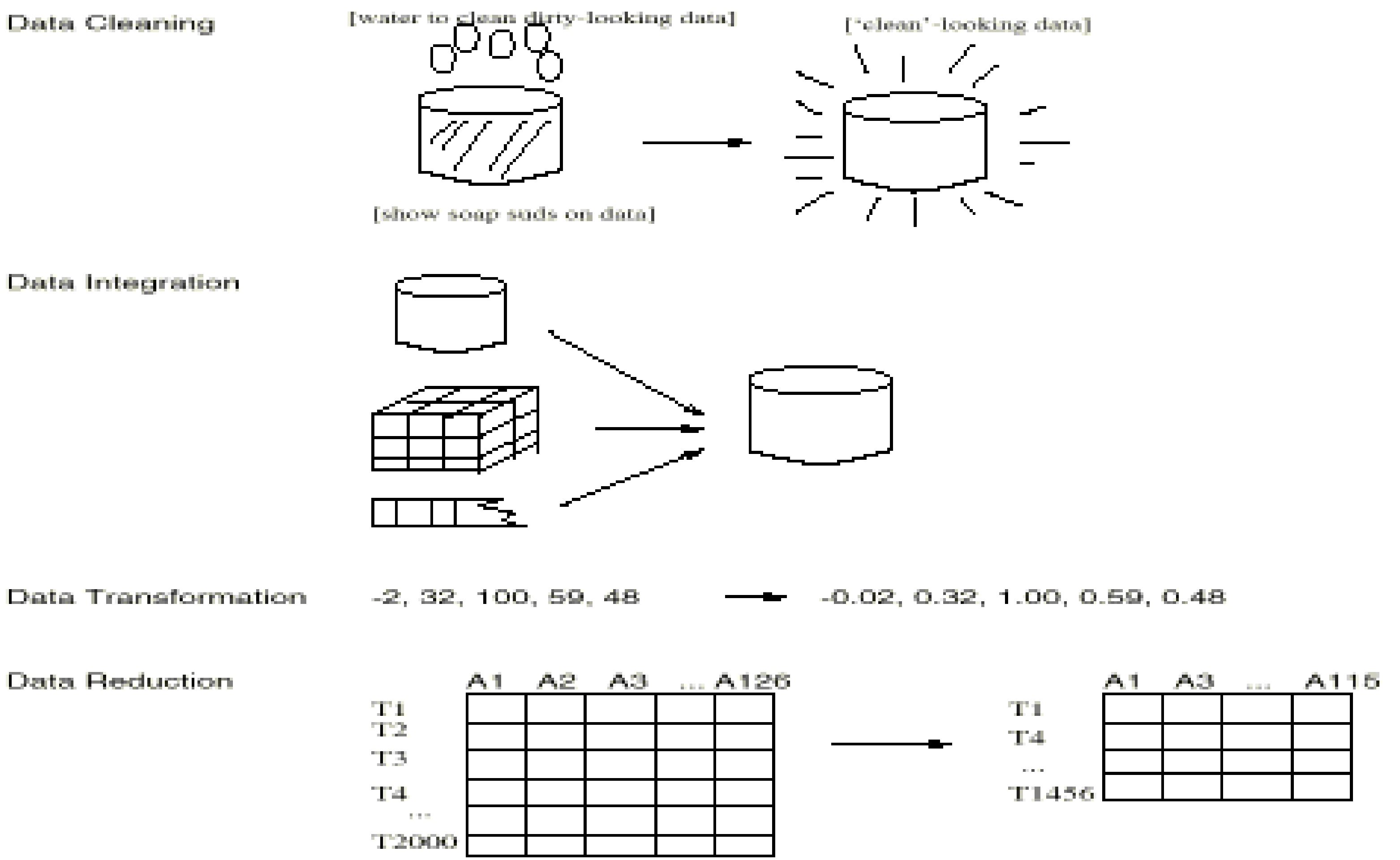
Data Accessibility Problems

- Legal Issues
  - Information Privacy
  - Information Security
- Departmental Access
- Political Reasons
- Data Format
- Architectural Reasons
- Timing

Why Is Data Dirty?

- Incomplete data comes from
  - n/a data value when collected
  - Different consideration between data collection and data analysis.
  - Human/hardware/software problems (e.g., earthquake catalog)
- Noisy data comes from the process of data
  - Collection (e.g., vital signs)
  - Entry
  - Transmission
- Inconsistent data comes from
  - Different data sources
  - Functional dependency violation
  - Business process changes (e.g., wine quality)

Forms of data preprocessing

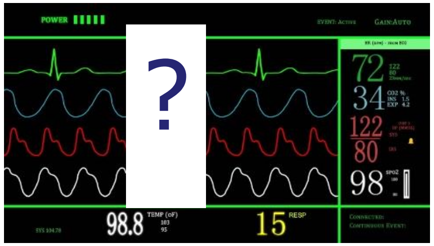


Data Cleaning Tasks

- Handle missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

How to Handle Missing Data?

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - a global constant
  - the attribute mean
  - the attribute mean for all samples of the same class: smarter
  - the most probable value: inference-based

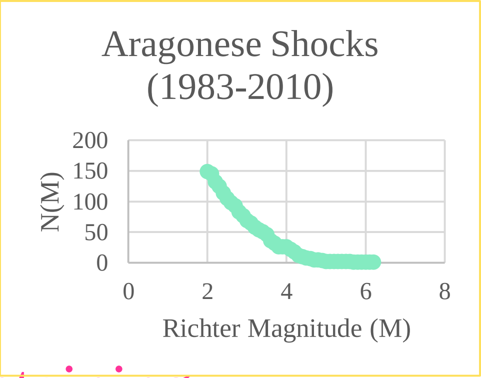


How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)
- Regression
  - smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.



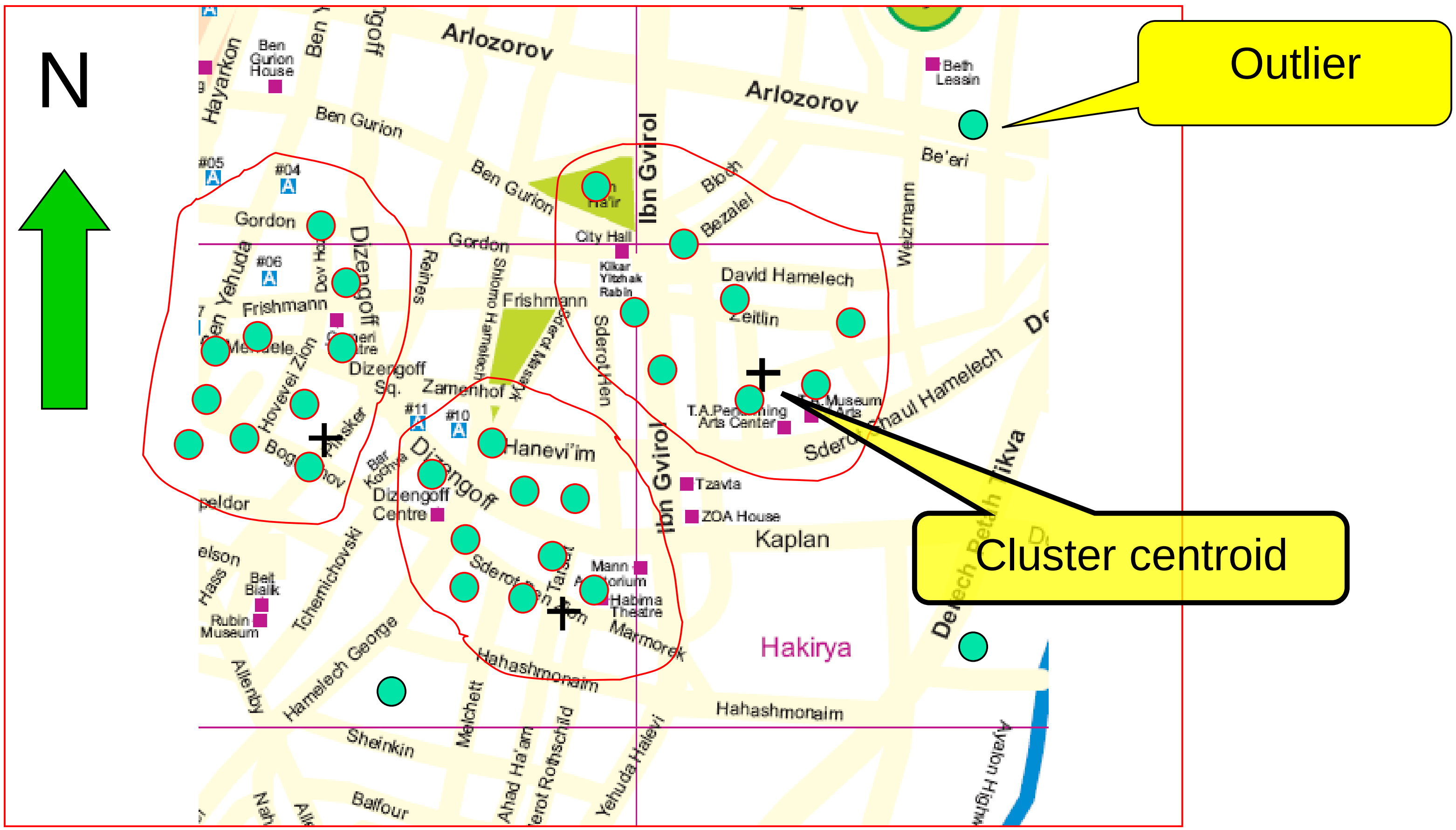
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Examples

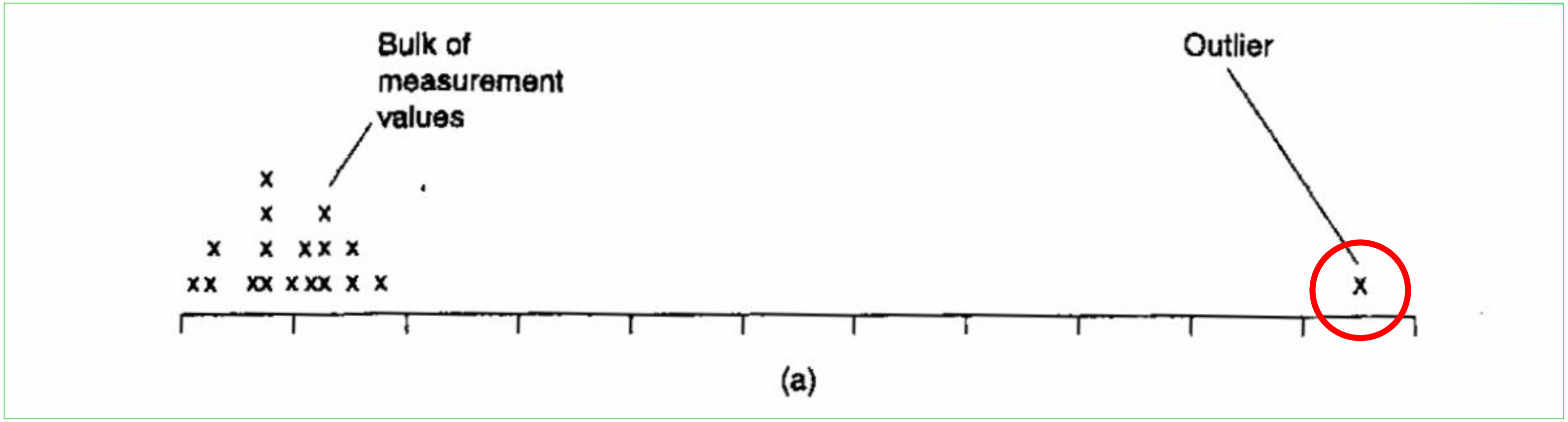
* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 (11 distinct values)
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means: (3 distinct values)
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries: (6 distinct values)
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

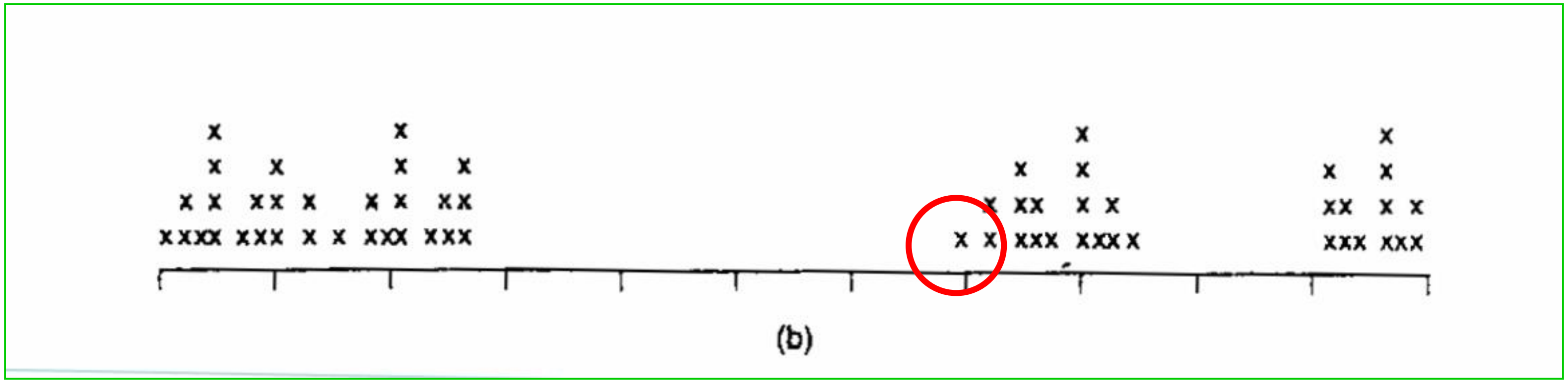
Cluster Analysis

Example: customer locations in a city



Outliers (Source: Pyle, 1999)





Conclusion: an outlier may be an element of another cluster

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., patient age
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Why reduce/avoid redundancies and inconsistencies before mining the data?

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature extraction
  - New attributes constructed from the given ones (e.g., DNN)

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

$$\frac{73,600}{100,000} = 0.736$$

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature extraction
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Dimensionality Reduction

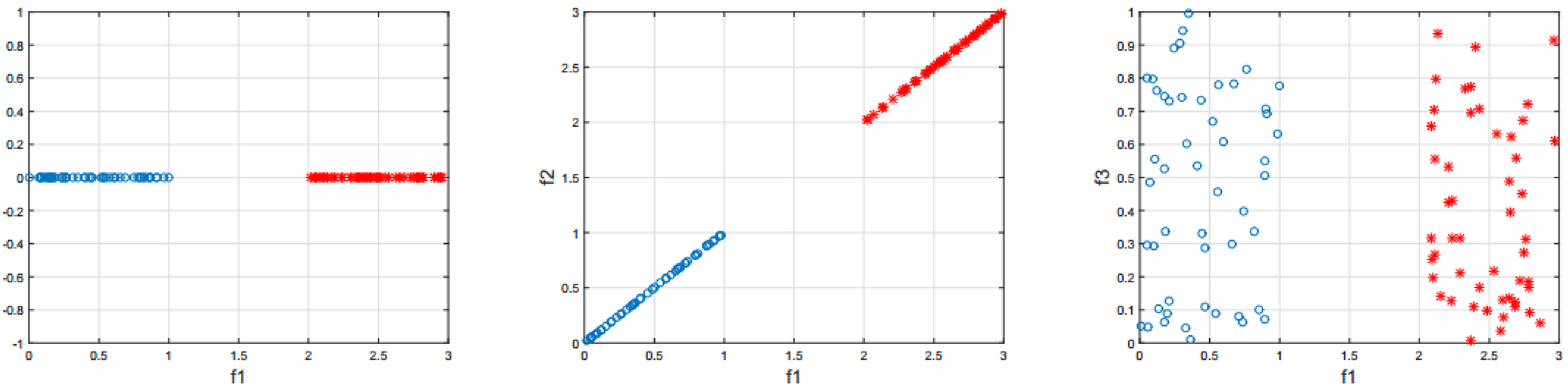
- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier interpretability and visualization of DM results
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Relevant, Redundant and Irrelevant Features

- Feature selection retains relevant features for learning and removes redundant or irrelevant ones
- For a binary classification task below, f_1 is relevant, f_2 is redundant given f_1, and f_3 is irrelevant

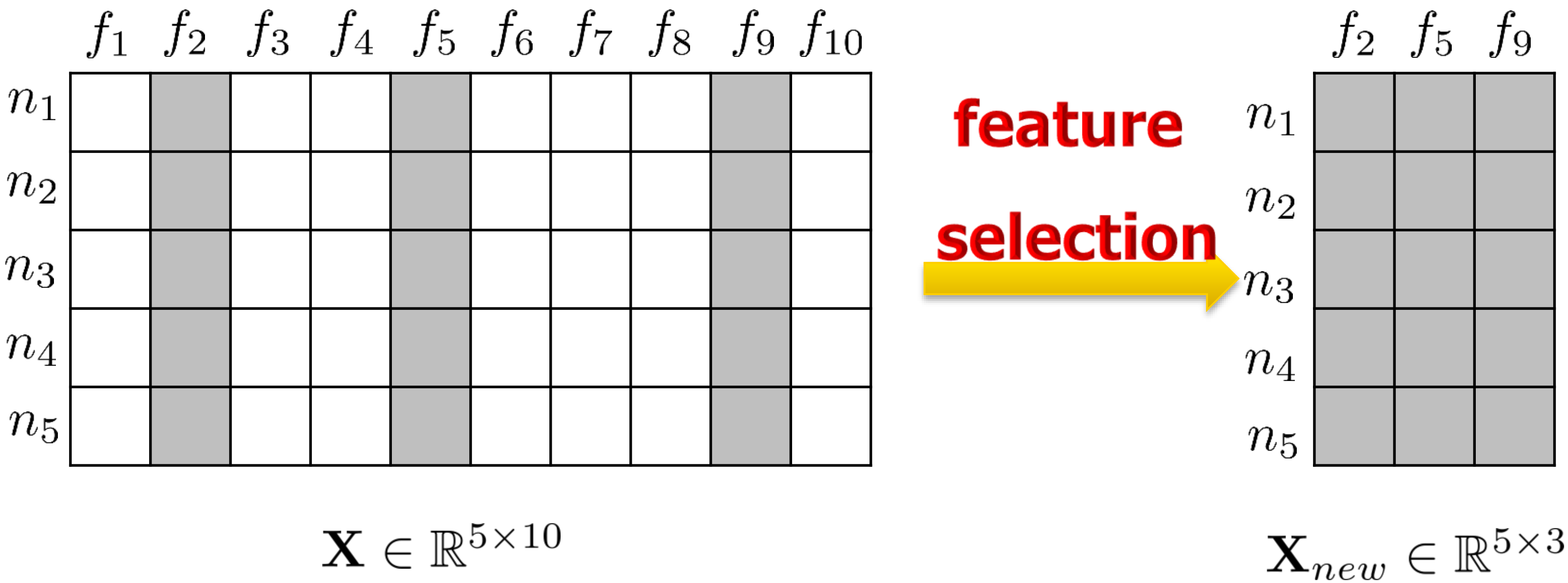


(a) relevant feature f_1 (b) redundant feature f_2 (c) irrelevant feature f_3

Source: Huan Liu

Feature Selection

Feature selection selects an ‘optimal’ subset of relevant features from the original high-dimensional data given a certain criterion



Source: Huan Liu

Sampling / Record Selection

- Simple random sampling
- Sampling without replacement
- Sampling with replacement
- Stratified sampling
- Active sampling / learning

Attribute Creation (Feature Extraction)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space
    - E.g., Fourier transformation, wavelet transformation, manifold approaches, DNN
  - Attribute construction
    - Combining features
    - Data discretization

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data—For numeric data, use discretization methods
  - Nominal data example: *street* < *city* < *state* < *country*

Example: Medical Records
(Source: Mortality Records, Israeli Ministry of Health)

- Input Attributes
  - Age
  - Date of Death
  - Gender
  - Area of Residence (about 30 areas)
  - Religion (14 codes)
  - Country of Birth

- Target Attribute
  - Medical Diagnosis (6-digit ICD-9 code)
- Additional data tables
  - Areas (נפה)
  - Regions (אזור)
  - Religions
  - Countries
  - Places of Birth (Continents)

Example: Medical Records (cont.) Data Pre-Processing

- Generalizing diagnoses to 36 groups
- Decoding age codes
- Generalizing areas (נפות) to regions (אזורים)
- Generalizing country of birth to continent of birth

Sample Medical Records

Raw Data

ID	סיבת מוות	גיל	נפה	ארץ לידה
100	428000	403	51	400
200	496000	373	53	110
300	799900	202	51	900
400	745200	108	11	900

Generalizing Medical Diagnoses

(Based on first 3 digits of ICD-9-CM Codes)

Code	Intervals	Diagnosis
0	0, 209	Other
1	1 - 139	Infectious and Parasitic Diseases
2	140-152, 155-161, 163-173, 175-184, 186-203	Other Malignant Neoplasms
3	153-154	Malignant Neoplasm of colon-rectum
4	162	Malignant Neoplasm of trachea etc.
5	174	Malignant Neoplasm of female breast
6	185	Prostate
7	204-208	Leukaemia
8	210-239	Non-Malignant Neoplasms
9	240-249, 251-279	Other Endocrine Diseases
10	250	Diabetes
11	280-289	Diseases of Blood
12	290-319	Mental Disorders
13	320-389	Diseases of the Nervous System
14	390-409	Other Diseases of the Circlulatory System
15	410-414	Ischaemic heart disease
16	415-429	Diseases of pulmonary circulation
17	430-438	Cerebrovascular disease
18	439-459	Other Diseases of the Circlulatory System

Code	Intervals	Diagnosis
19	460-479, 488-489, 497-519	Diseases of the Respiratory System
20	480-487	Pneumonia and Influenza
21	490-496	Chronic obstructive pulmonary disease
22	520-579	Diseases of the Digestive System
23	580-599	Diseases of the Urinary System
24	600-629	Diseases of the Genital Organs
25	630-639	Abortion
26	640-679	Pregnancy etc.
27	680-709	Diseases of the Skin
28	710-739	Diseases of the Muscoloskeletal System
29	740-759	Congenital Anomalies
30	760-779	Perinatal period
31	780-799	Symptoms and Ill-Defined Conditions
32	800-809, 820-949, 970-999	Other Accidents
33	810-819	Motor Vehicle Traffic Accidents
34	950-959	Suicide and Self-inflicted injuries
35	960-969	Homicide

Generalizing Medical Diagnoses (cont.)

Code	Intervals	Diagnosis
16	415-429	Diseases of pulmonary circulation
21	490-496	Chronic obstructive pulmonary disease
31	780-799	Symptoms and Ill-Defined Conditions
29	740-759	Congenital Anomalies

ID	סיבת מוות	Reason
100	428000	16
200	496000	21
300	799900	31
400	745200	29

Decoding Age Codes

- Age code: XXX (3 digits)
- First digit
  - 1 – days
  - 2 – months
  - 3 – years (1-99)
  - 4 – years (100-)

ID	גיל	Age (years)
100	403	103
200	373	73
300	202	0
400	108	0

Calculating region (אזור) from area (נפה)

קוד נפה	שם נפה	קוד אזור	שם אזור
11	ירושלים	1	ירושלים
51	תל-אביב	5	ת"א
53	חולון	5	ת"א

ID	נפה	Region_Code
100	51	5
200	53	5
300	51	5
400	11	1

Calculating place of birth (continent) from country of birth

Country	Continent_Name	Continent_Code
110	אסיה	0
400	אירופה אמריקה	1
900	ישראל	3

ID	ארץ לידה	Cont_Birth
100	400	1
200	110	0
300	900	3
400	900	3

Transformation Results

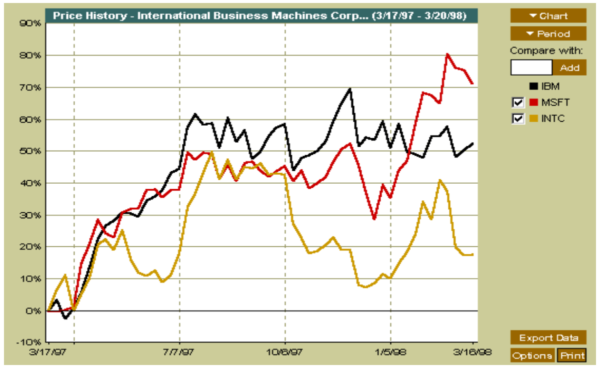
Raw Data

ID	סיבת מוות	גיל	נפה	ארץ לידה
100	428000	403	51	400
200	496000	373	53	110
300	799900	202	51	900
400	745200	108	11	900

Final Data

ID	Reason	Age (years)	Region_Code	Cont_B
100	16	103	5	1
200	21	73	5	0
300	31	0	5	3
400	29	0	1	3

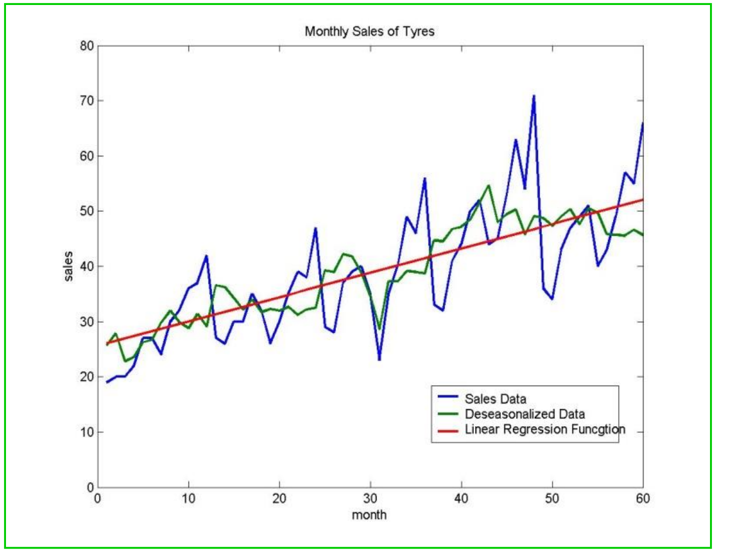
Preparation of Time Series Data



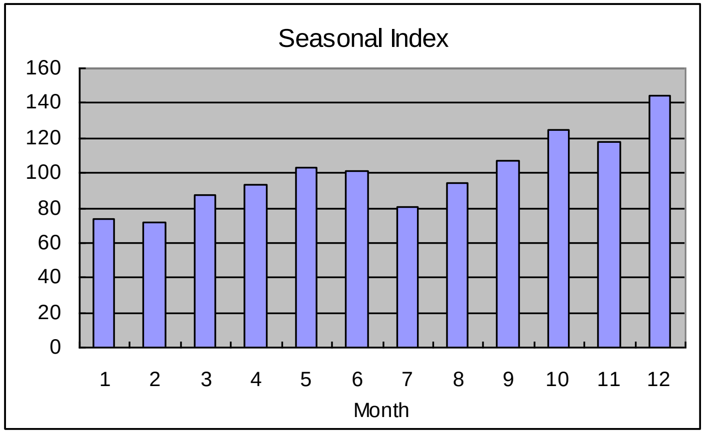
- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, noise
- Time series data mining tasks
  - Finding *clusters* of similar time series
  - Detecting *events* (change points) in time series
  - Predicting future values of time series
  - etc.

Describing a Series

- Trend
  - A non-cyclic, monotonically increasing or decreasing component of the waveform
- Cycle
  - A trend over one period may be a cycle over a different period
- Seasonality
  - Certain seasons (e.g., X-mas) are inherently different disregarding any other trend, cycle, or noise
- Noise
  - The component left after the trend, cyclic, and seasonal components have been extracted

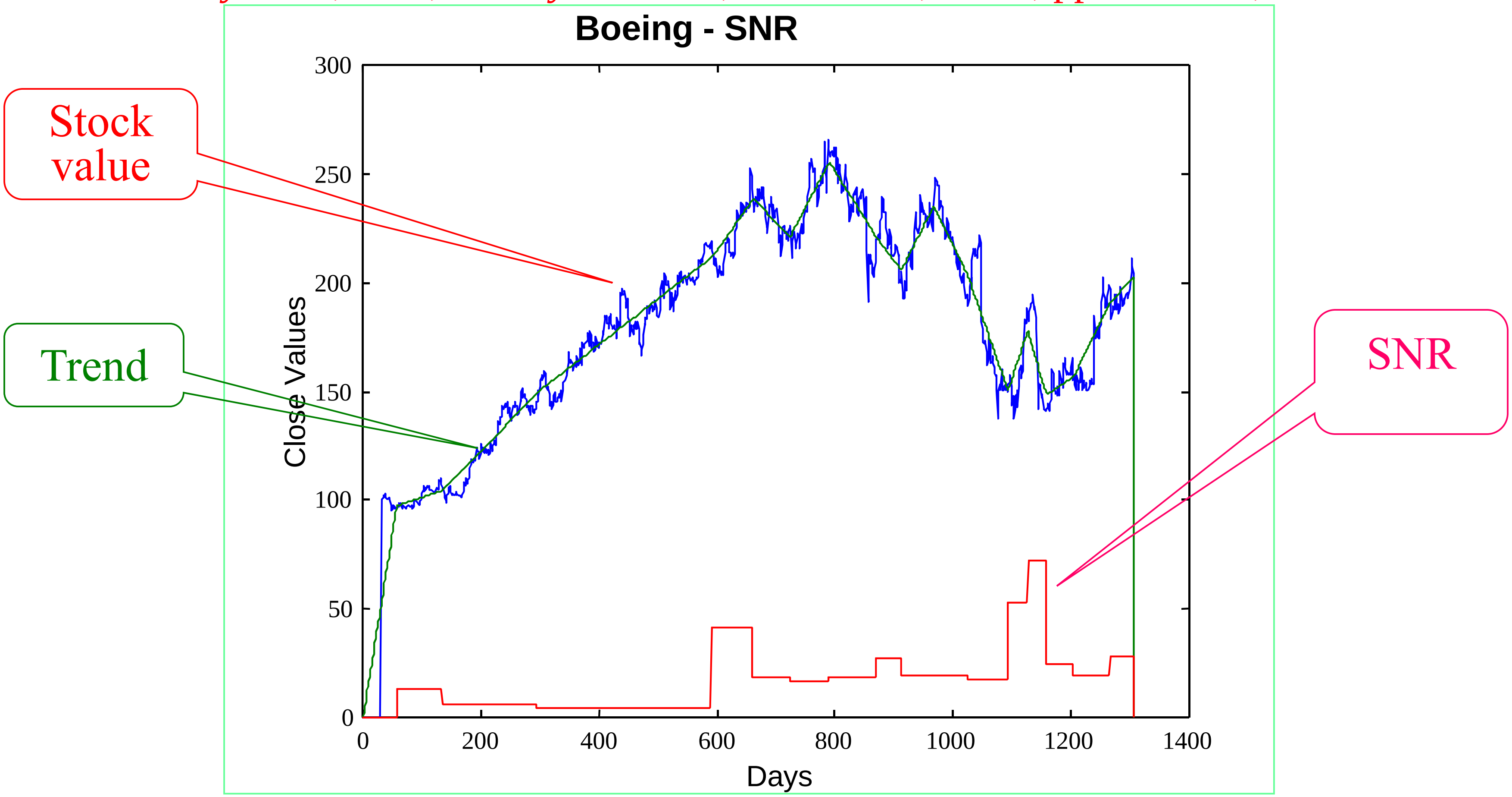


Raw data from
http://www.bbk.ac.uk/manop/man/docs/QII_2_2003%20Time%20series.pdf



Preparation of Time Series Data: Example

Based on: Last, Klein, Kandel, Knowledge Discovery in Time Series Databases, *IEEE Transactions on Systems, Man, and Cybernetics*, 31: Part B, No. 1, pp. 160-169, Feb. 2001



Moving Average Methods

- Goal
  - Determining the trend of time series data
- Most common methods
  - Simple Moving Average
  - Weighted Moving Average
  - Exponential Moving Average

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt

Simple Moving Average

- The forecast is simply the average of the most recent *k* observations:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-k+1}}{k}$$

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt

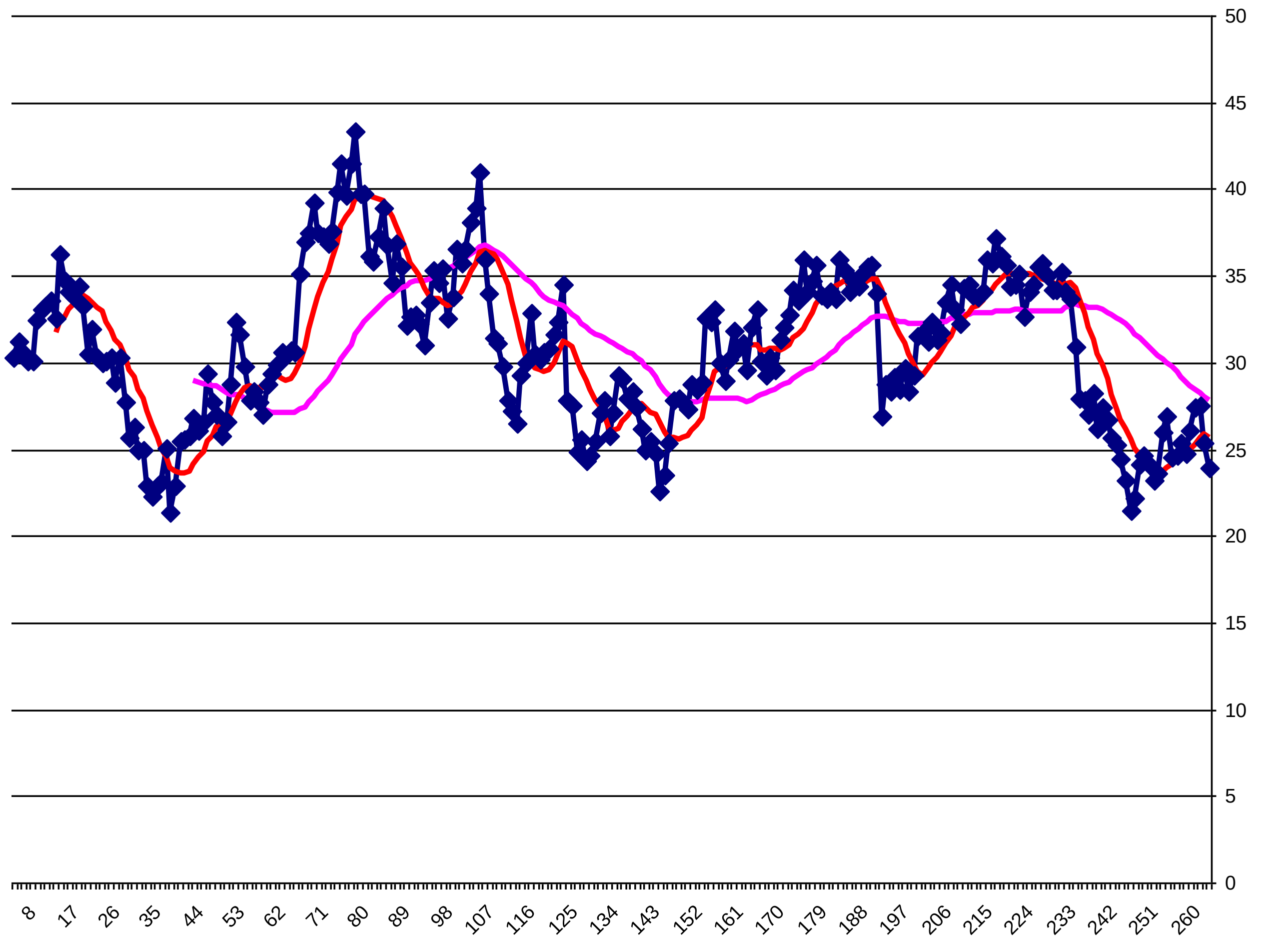
Selecting *k*

- Smoothing effect (large *k*)
- Responsiveness (small *k*)
- Useful to compare results with different *k* values

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt

BCC Stock Price with 10 & 40 week MA

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt



Weighted Moving Average

- Moving average where each value in the window is assigned a unique weight

$$\hat{Y}_{t+1} = w_t Y_t + w_{t-1} Y_{t-1} + \ldots + w_{t-k+1} Y_{t-k+1}$$

$$where: \quad w_t + w_{t-1} + \ldots + w_{t-k+1} = 1$$

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt

Selecting Weights

- Sum is 1.0
- More recent data is often more important
- Other knowledge may skew weights
- Equal weights is the same as single moving average ($w=1/k$)

From http://ie.d.umn.edu/MSEM/Courses/EMGT5230/Lectures/Ch_04a.ppt

Exponential Moving Average

$$F_t = \alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + \alpha(1-\alpha)^2 Y_{t-3} + \cdots$$
$$F_t = \alpha Y_{t-1} + (1-\alpha)\left[\alpha Y_{t-2} + \alpha(1-a)Y_{t-3} + \cdots\right]$$

$$F_t = \alpha Y_{t-1} + (1-\alpha)F_{t-1}$$

Ft :	Forecast for period t
F_{t-1}:	Last period forecast
Y_{t-1}:	Last period actual value

Summary

- Data preparation or preprocessing is a big issue for *data engineers*
- Descriptive data summarization is needed for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature extraction + selection
  - Discretization
- A lot a methods have been developed but data preprocessing is still an active area of research