IS 733 Lesson 3

Data Preprocessing

Some slides based on those by Witten et al., Han et al., Vandana Janeja and James Foulds

Announcements

- Project groups should be formed by next week, 2/16. We will share project ideas in class.

- Project proposal due 2/23!
  - Please send this to me by email, nroy@umbc.edu, and CC your teammates

Announcements

- Reminder: Homework 1 due next week, 2/16
  - Submit on Blackboard, under “Assignments”
  - Make use of Blackboard Discussion Forum for asking questions! Do subscribe to BB IS 733 class discussion forum.
  - Academic integrity, see syllabus on course webpage

    You can discuss homework assignments together, but you **can’t look at each other’s answers or copy each other**

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

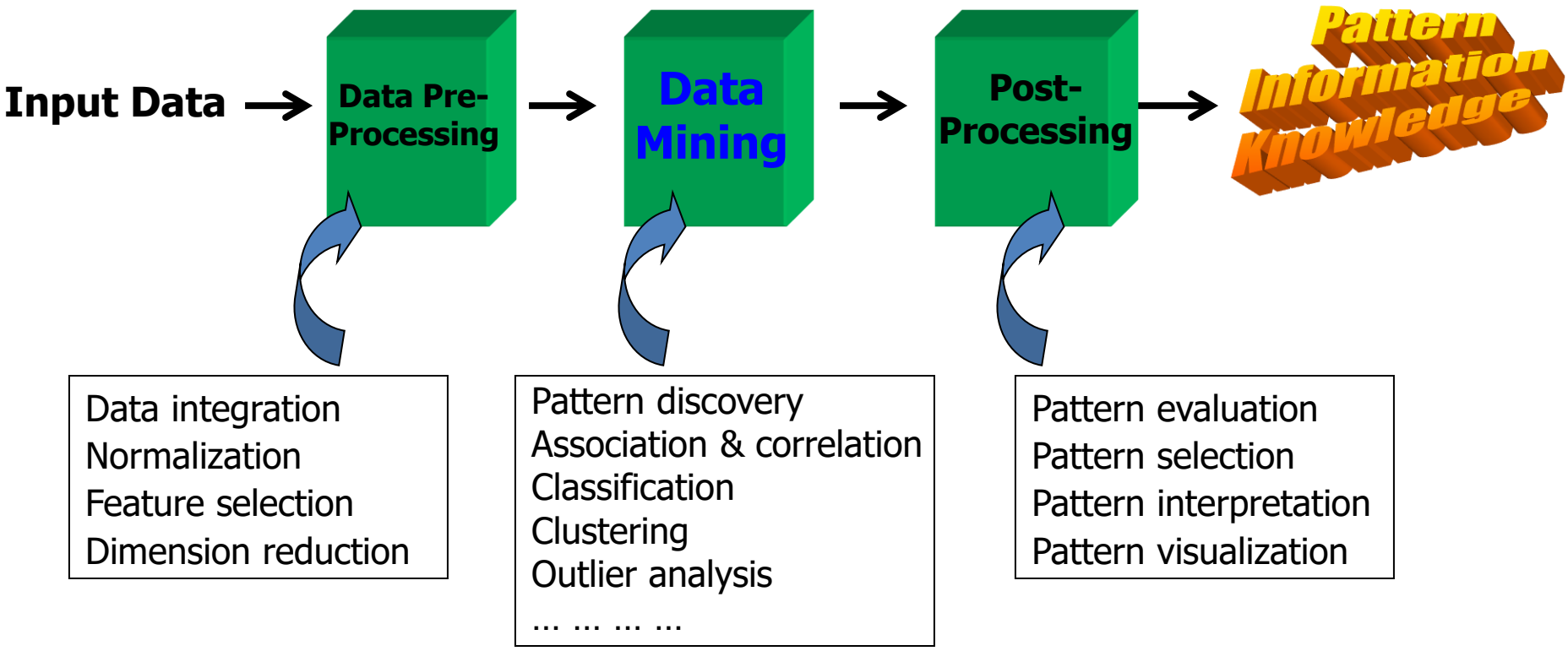
Open poll in your web browser

Learning outcomes

By the end of the lesson, you should be able to:

- **Explain** and **employ** several strategies for "getting to know your data"
- **Discuss** common reasons for real-world data being noisy, and strategies for **data cleaning**
- **Identify** redundancy from **data integration** using the **correlation coefficient** and **covariances**
- **Apply** appropriate **data transformations** and **feature construction** techniques to improve the performance of data mining algorithms, and **justify your choices**

KDD Process: A Typical View from ML and Statistics



How do we know which of these methods we should perform?

Getting to know your data

- It’s pretty hard to know what methods to use, if you don’t know anything about your data

- An initial exploration of your data can help you decide how to proceed

Getting to know your data

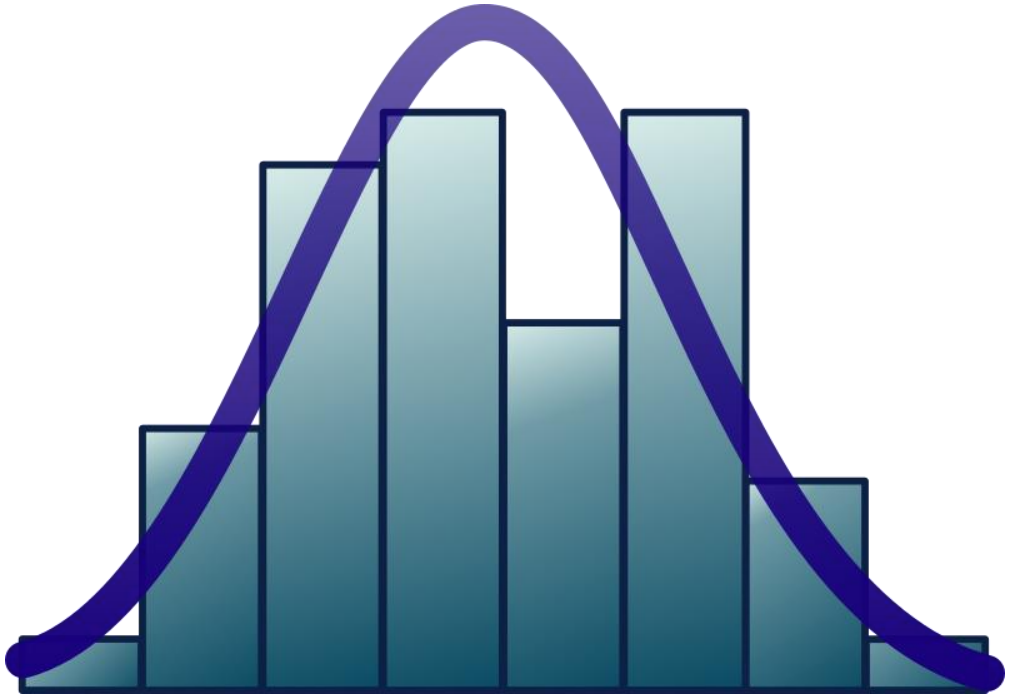
- **Inspecting your data "by hand" is a best practice**

  - Take the time to **open up your data file and have a look**.
    You might be surprised at what you find!
    - You may notice obvious issues with the data, e.g., duplicate records / attributes, nonsensical values, useless attributes,…
    - Too much data to inspect manually?
      Take a sample!

Getting to know your data

- Simple **visualization tools** and **summary statistics** are very useful
  - Make some plots, calculate summary statistics, then think:
  - Is the distribution consistent with background knowledge?
    (You may need to consult domain experts)
  - Any obvious outliers?
  - Are some variables heavily correlated with each other?

Getting to know your data



Histogram and best-fit normal distribution for sepal lengths of Iris Versicolor flowers

Image obtained from https://commons.wikimedia.org/wiki/File:Fisher_iris_versicolor_sepalwidth.svg author en:User:Qwfp

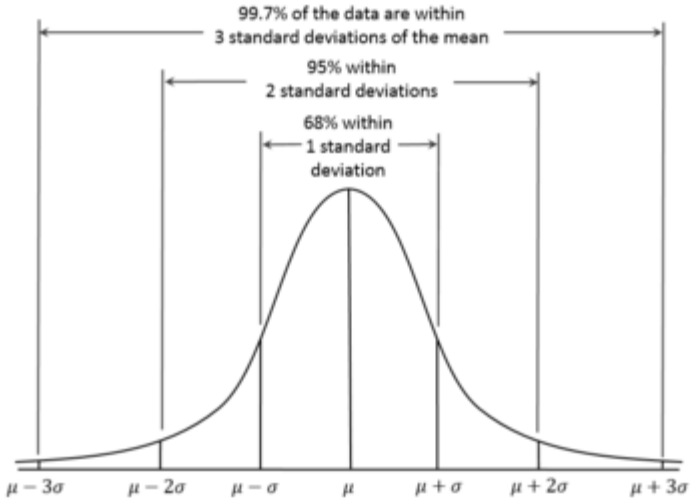
Measuring the Dispersion of Data

- Variance and standard deviation
  - **Variance**:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - **Standard deviation** σ is the square root of variance σ^2

For a normal distribution:



**dividing by N-1 instead of N leads to unbiased estimates. We will ignore this!*

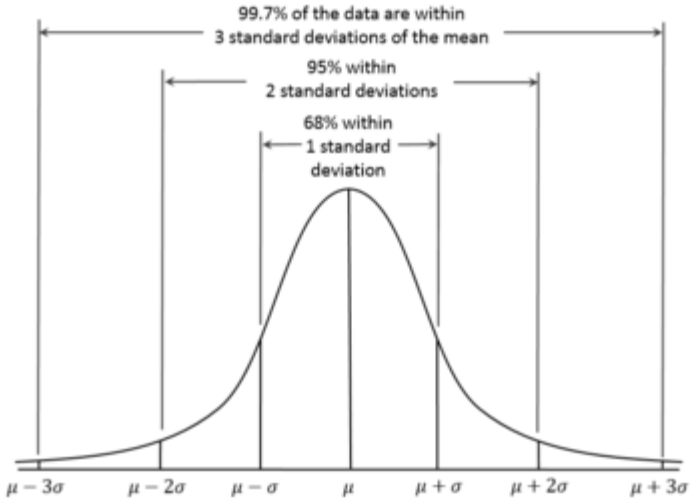
Measuring the Dispersion of Data

- Variance and standard deviation
  - **Variance**:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

  - **Standard deviation** σ is the square root of variance σ^2

For a normal distribution:



**dividing by N-1 instead of N leads to unbiased estimates. We will ignore this!*

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

Your poll will show here

1

Install the app from pollev.com/app

2

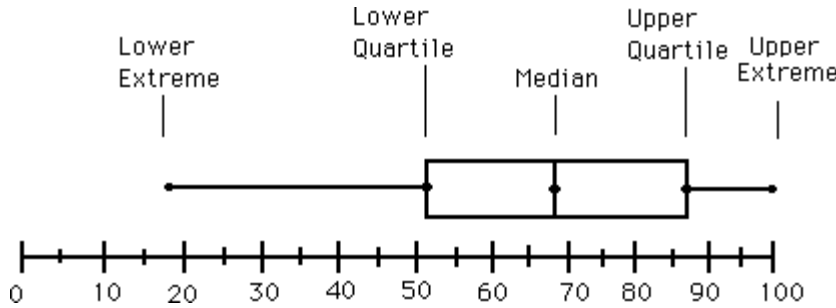
Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

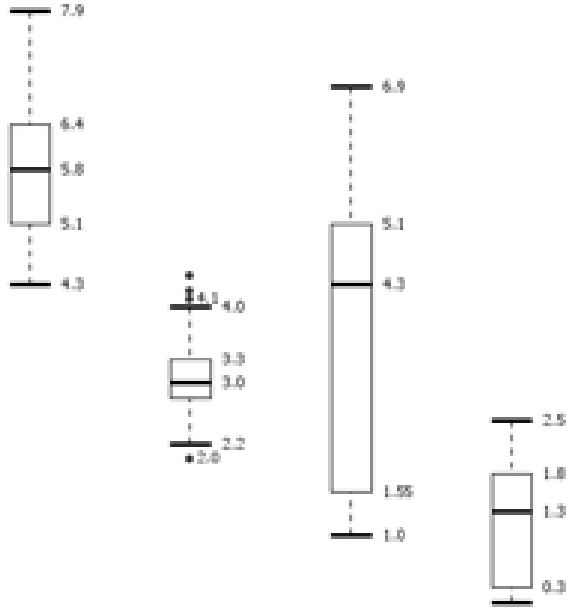
or

Open poll in your web browser

Boxplot Analysis



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

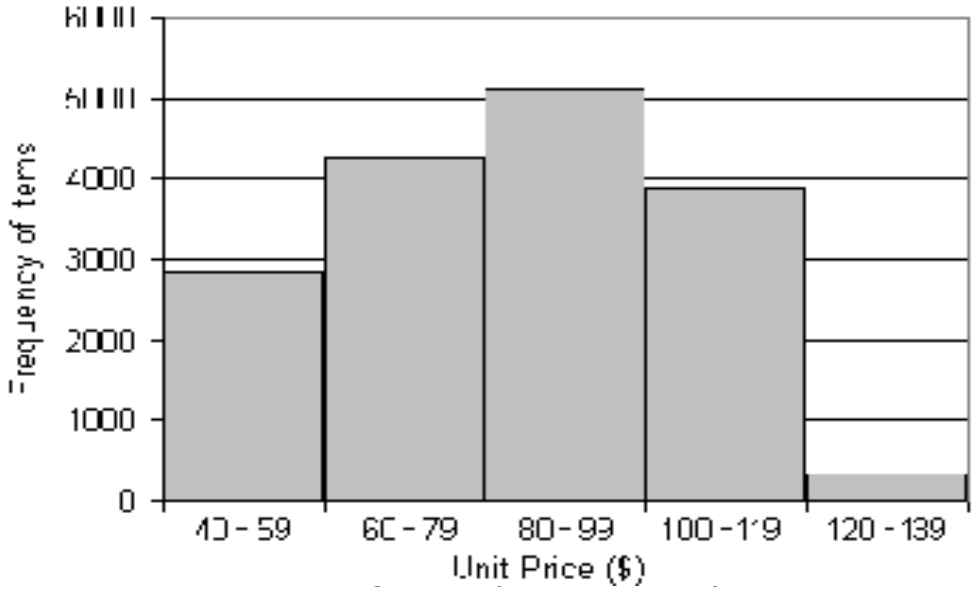
Still not working? Get help at pollev.com/app/help

or

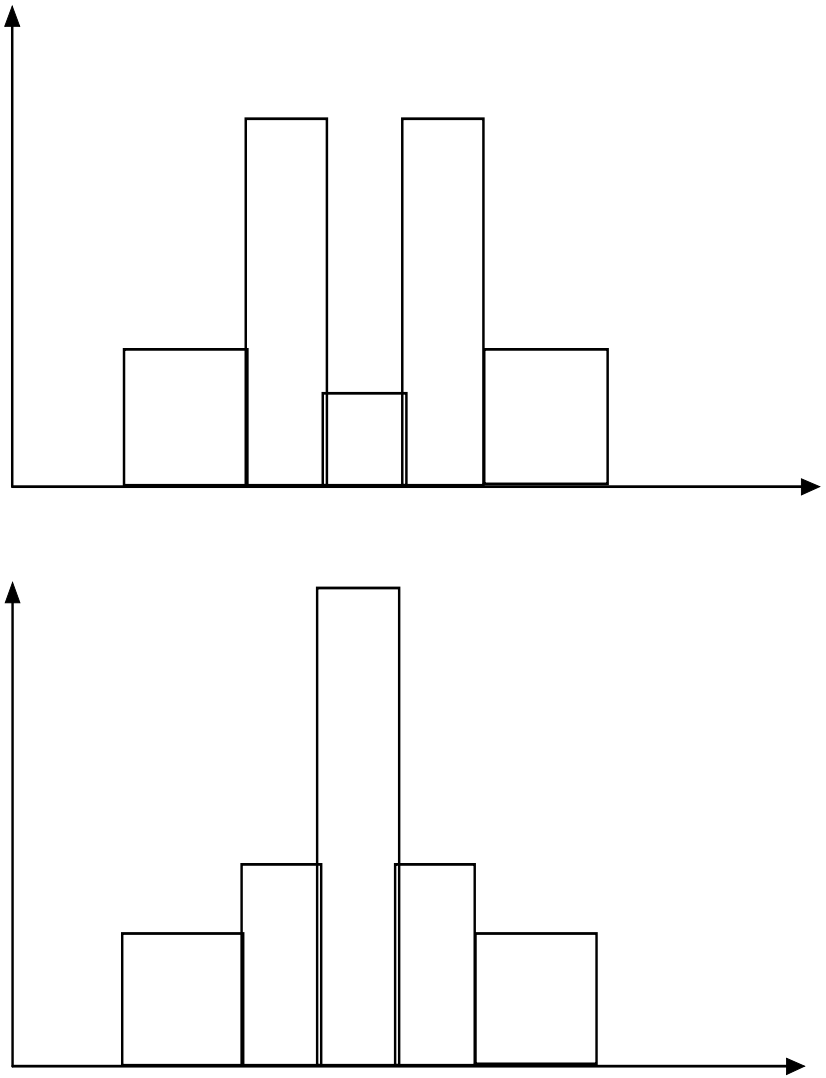
Open poll in your web browser

Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



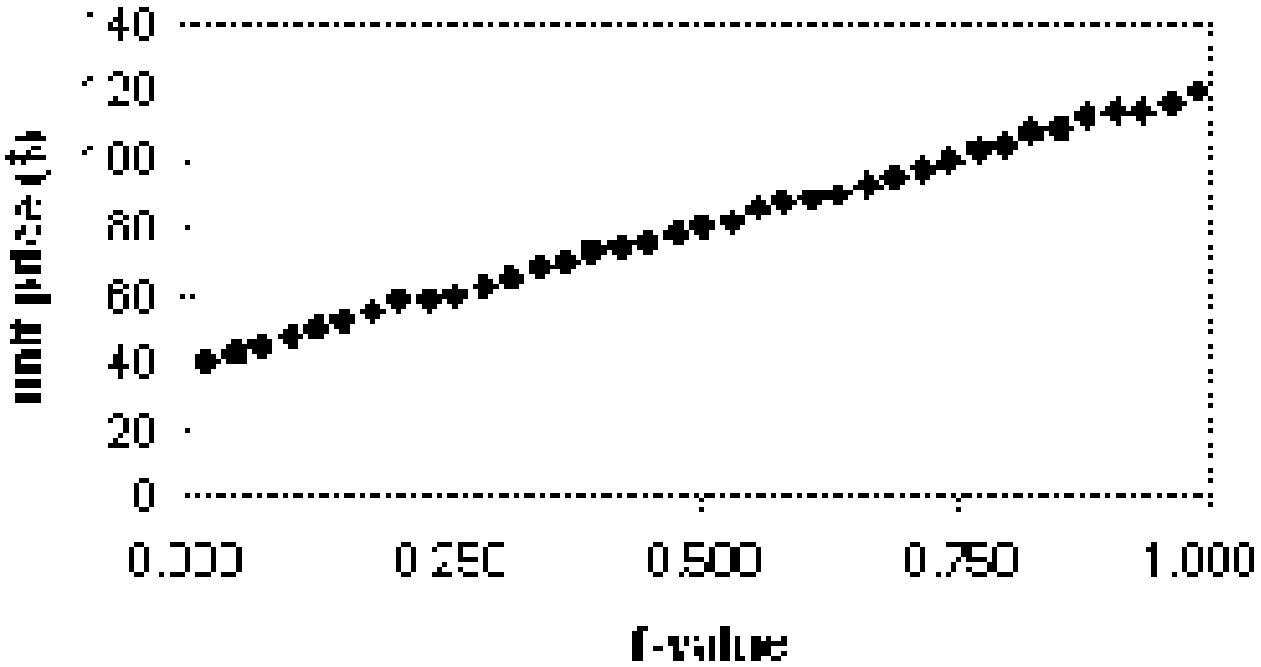
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

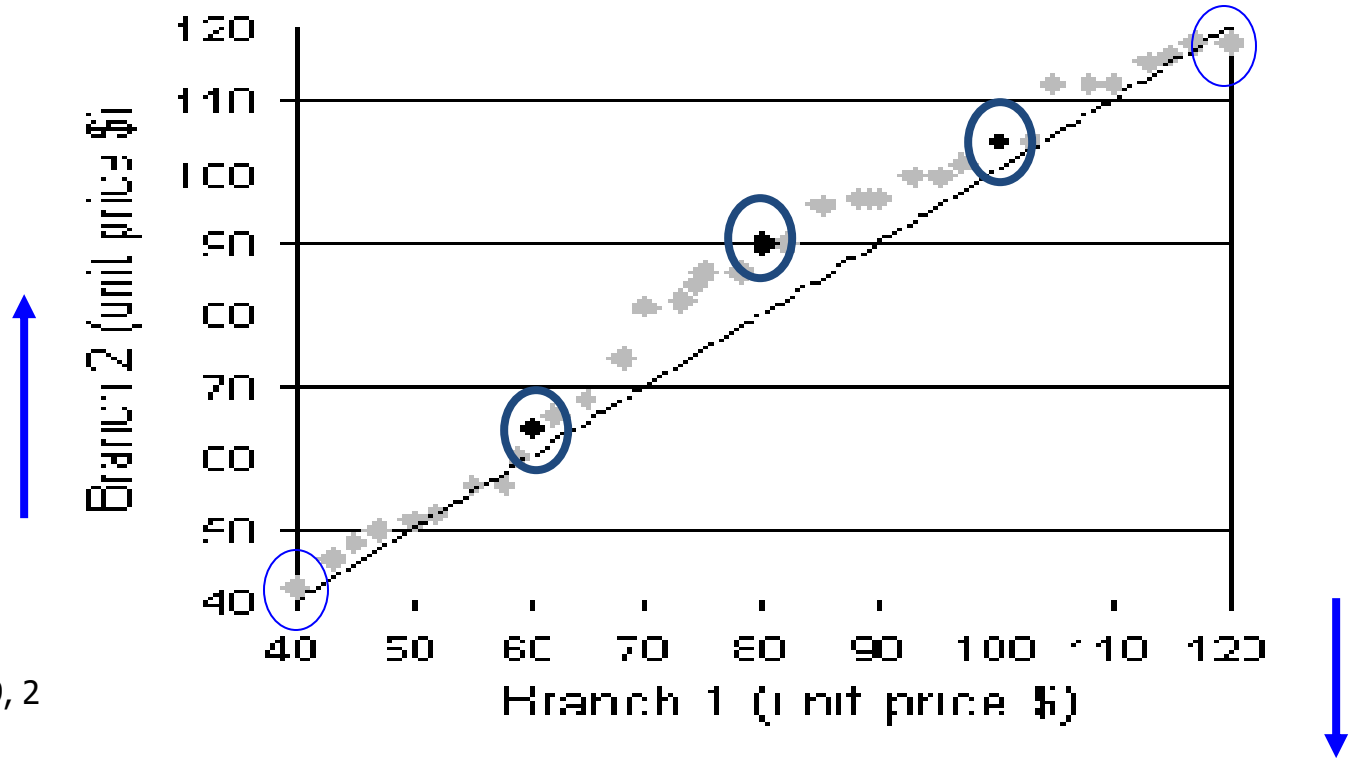
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i % of the data are below or equal to the value x_i



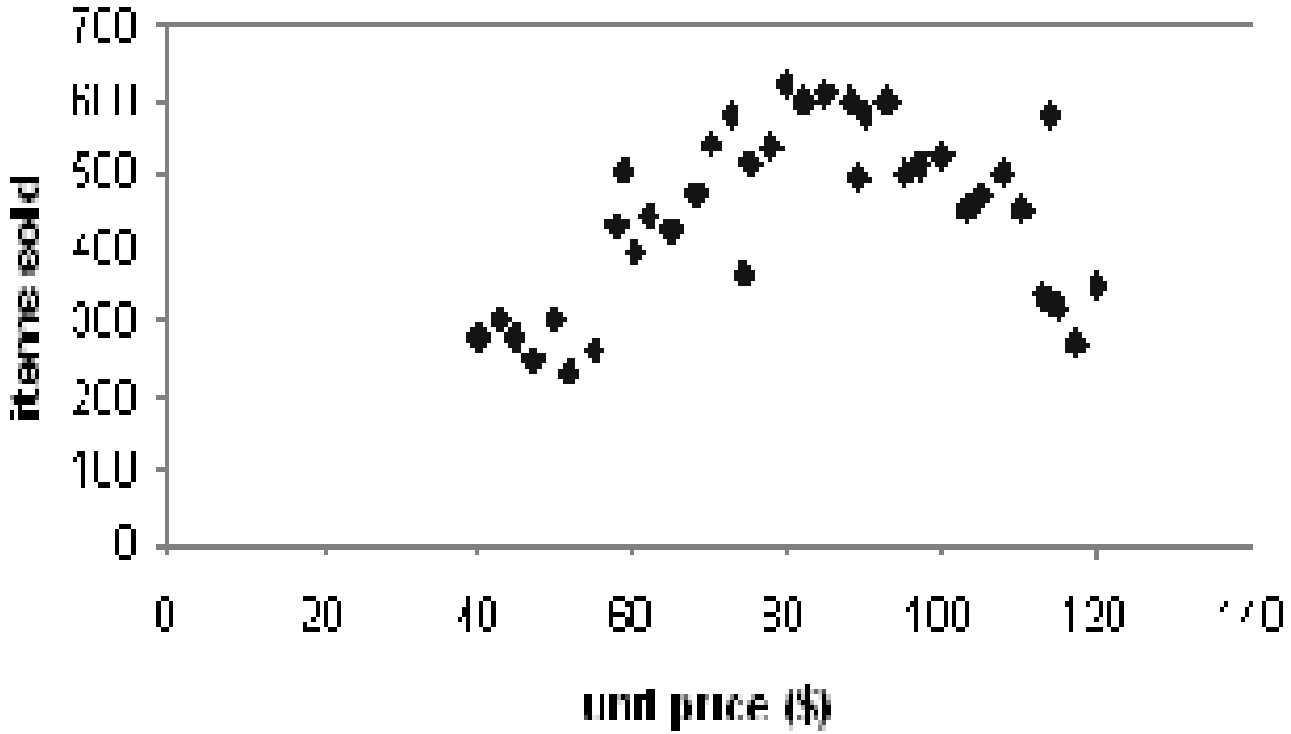
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Summary: Simple statistical plots

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Your poll will show here

1

Install the app from pollev.com/app

2

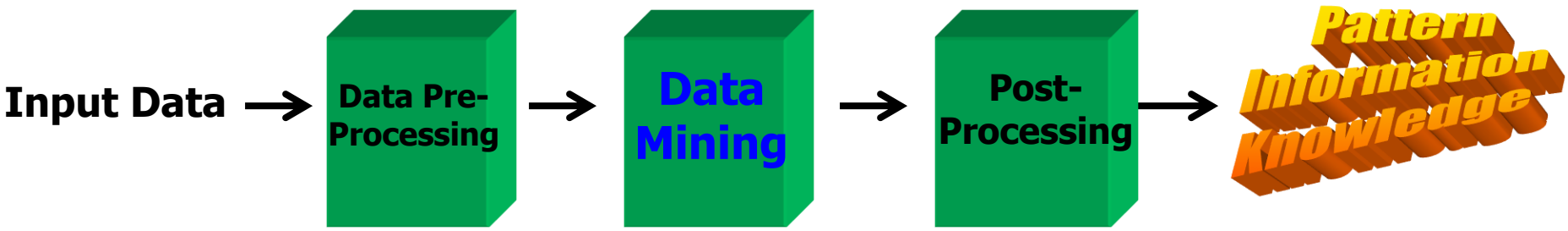
Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

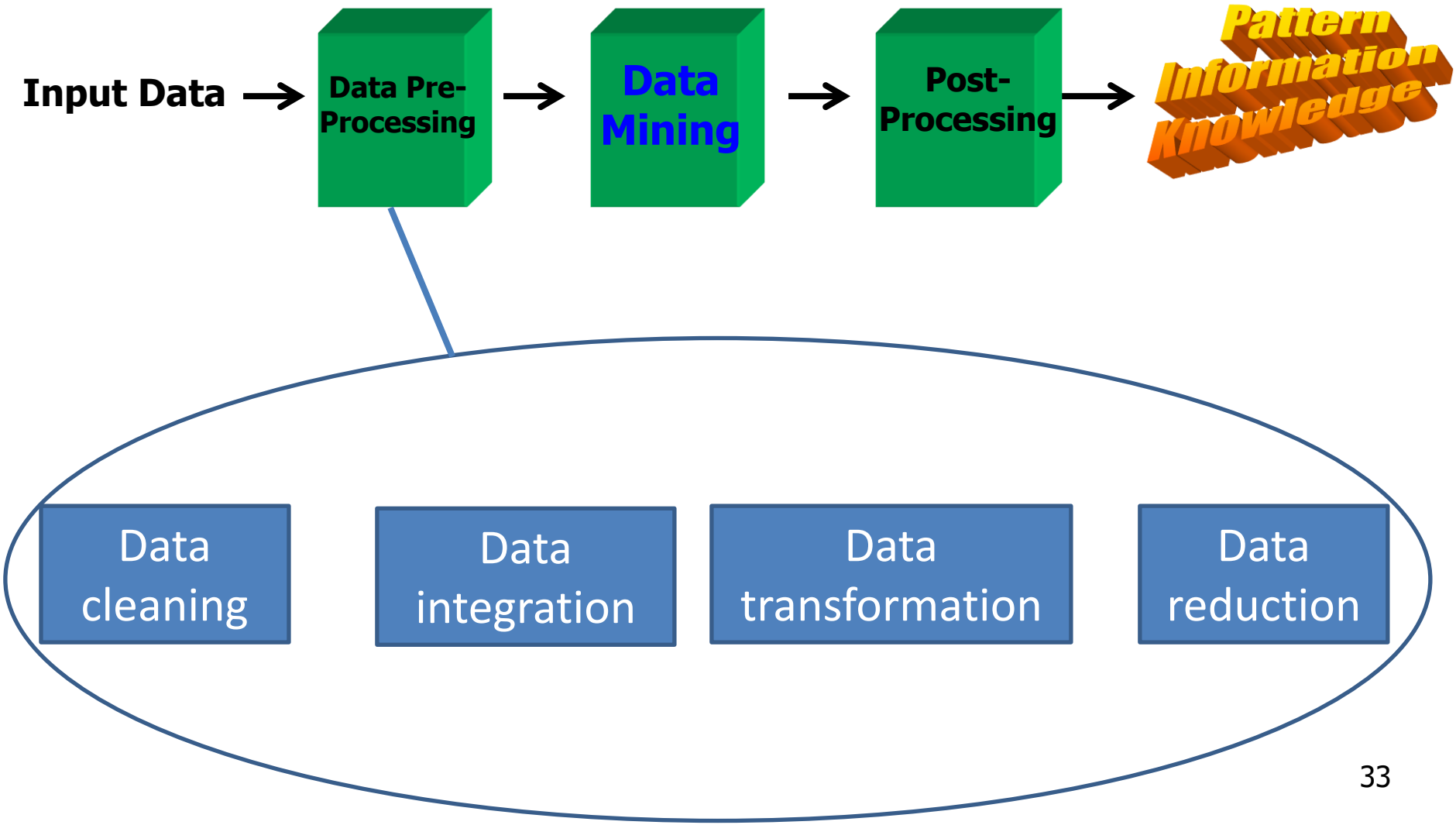
Open poll in your web browser

Key takeaway: take the time to get to know your data



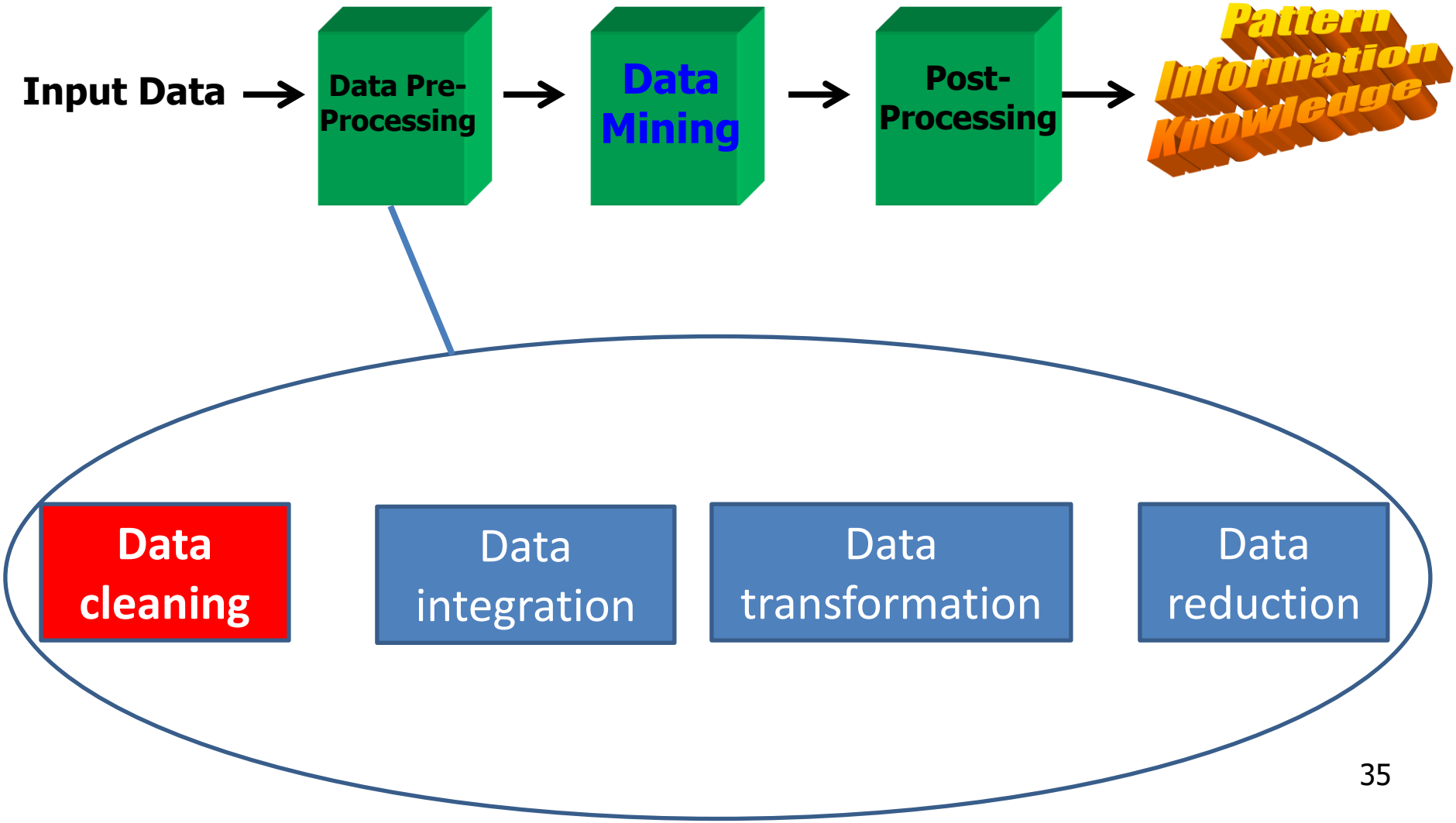
How do we know which of these methods we should perform?

Manual inspection, visualization, statistical plots, summary statistics are useful tools which help to answer this question!


Input Data
Data Pre-Processing
Data Mining
Post-Processing
Pattern
Information
Knowledge
Data cleaning
Data integration
Data transformation
Data reduction

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data transformation**
  - Normalization and aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: part of data reduction, of particular importance for numerical data


Input Data
Data Pre-Processing
Data Mining
Post-Processing
Pattern
Information
Knowledge
Data cleaning
Data integration
Data transformation
Data reduction

Data Cleaning

- No quality data, no quality mining results!

  "Garbage in, garbage out"

  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics

  - "Data cleaning is the number one problem in data warehousing"—DCI survey

  - Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers
  - smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- *Ignore the tuple:* usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- *Fill in the missing value manually:* tedious + infeasible?
- *Fill in it automatically* with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

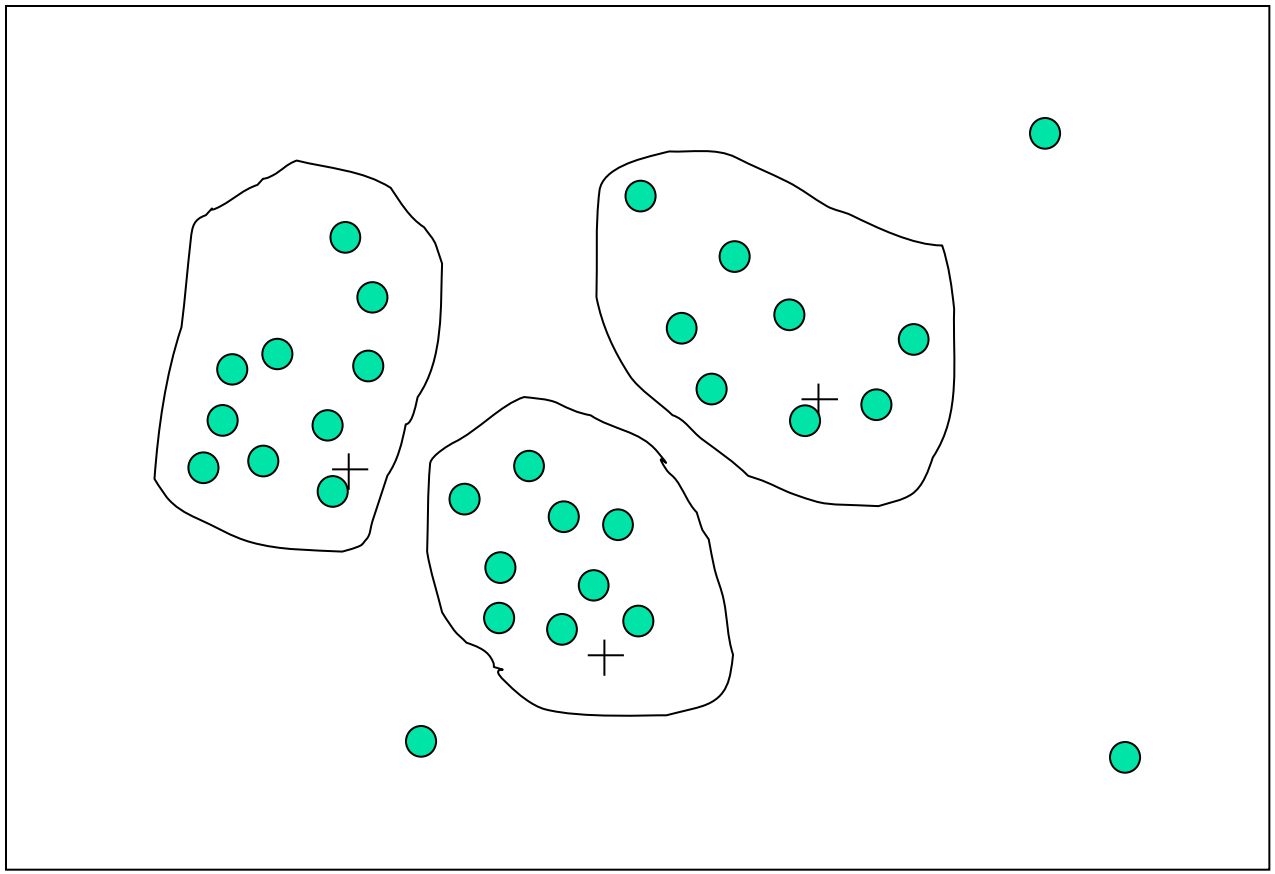
Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

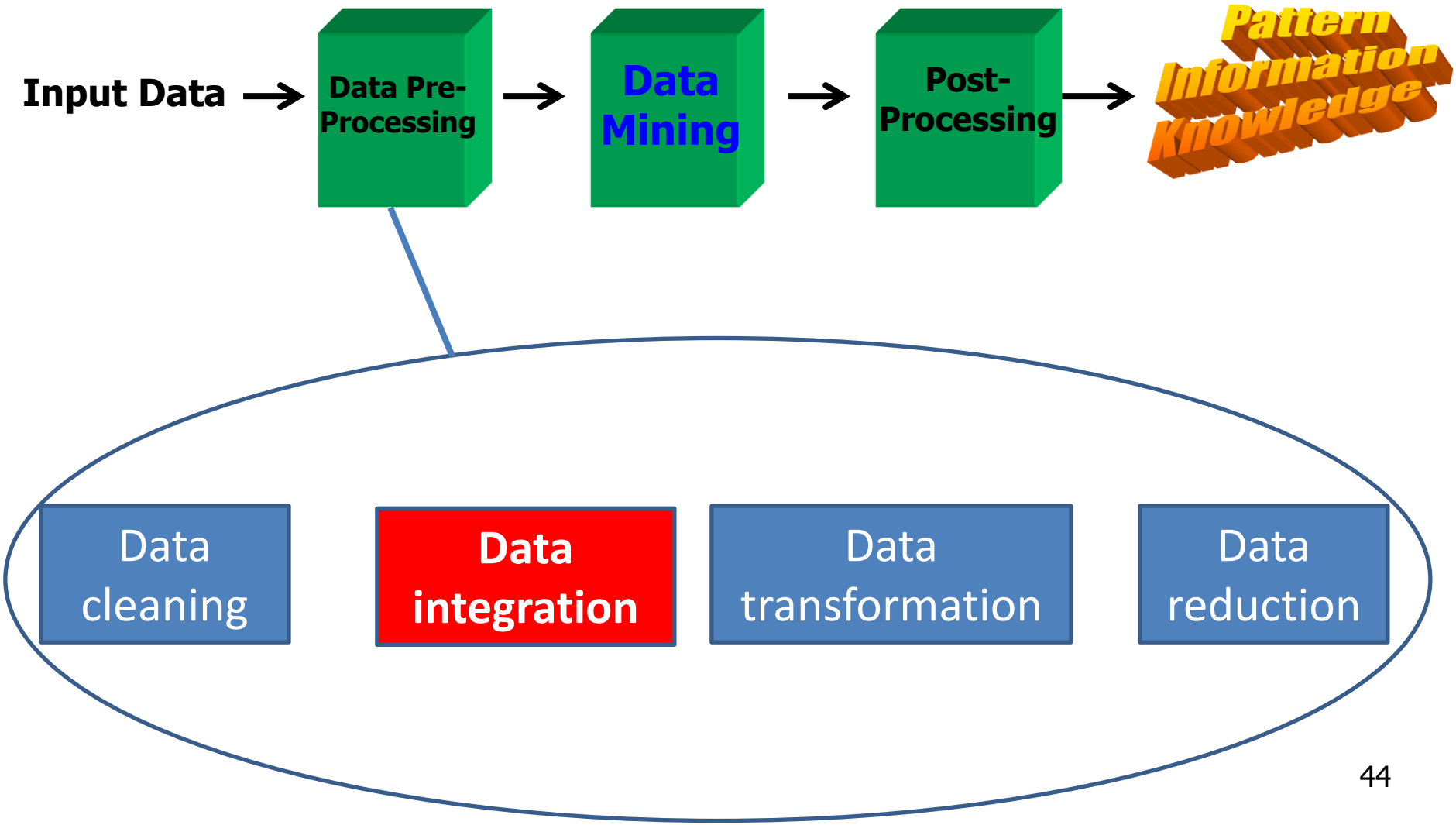
How to Handle Noisy Data?

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

Cluster Analysis



Can be applied on specific attribute(s) for smoothing


Input Data
Data Pre-Processing
Data Mining
Post-Processing
Pattern
Information
Knowledge
Data cleaning
Data integration
Data transformation
Data reduction

Data Integration

- **Data integration:**
  - Combines data from multiple sources into a coherent store

  - **Schema integration problem:**

    e.g., A.cust-id ≡ B.cust-#
    - Integrate metadata from different sources

Data Integration

- **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., annual revenue

Handling Redundancy in Data Integration

- Redundant attributes may be able to be detected by *correlation analysis*

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment correlation coefficient)

$$r_{p,q} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(q_i - \bar{q})}{\sigma_p \sigma_q}$$

where n is the number of tuples, $\bar{p}$ and $\bar{q}$ are the respective means of p and q, σ_p and σ_q are the respective standard deviation of p and q.

- If $r_{p,q} > 0$, p and q are **positively correlated** (p's values increase as q's). The higher, the stronger correlation.
- $r_{p,q} = 0$: **independent**; $r_{pq} < 0$: **negatively correlated**

Correlation (viewed as linear relationship)

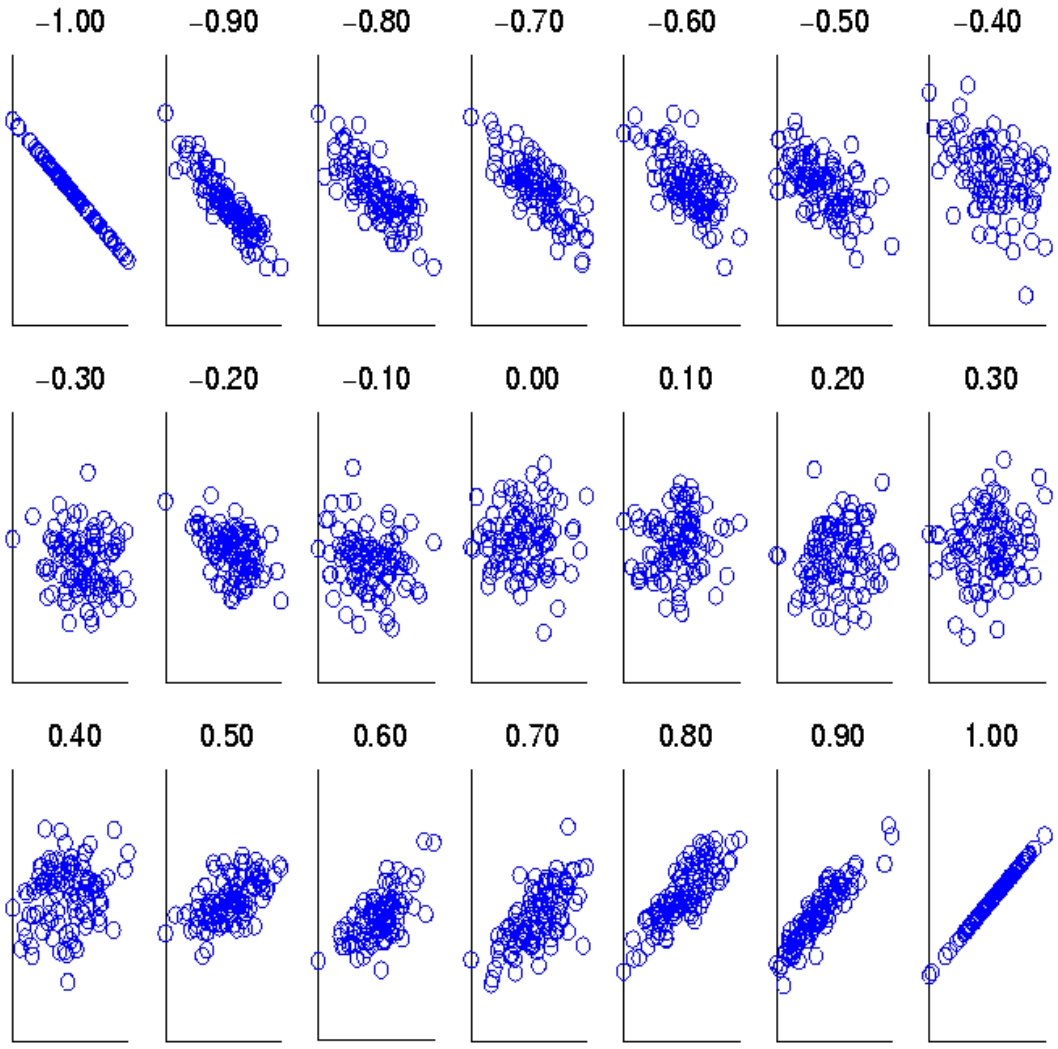
- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values at the same time, or smaller at the same time.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ does not imply independence:
  - Some pairs of random variables may have a covariance of 0 but are not independent. If data are multivariate normal, if Cov(A,B)=0 then independent

**dividing by n-1 instead of n leads to unbiased estimate of Cov. We will ignore this!*

Covariance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 – 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

Covariance: An Example

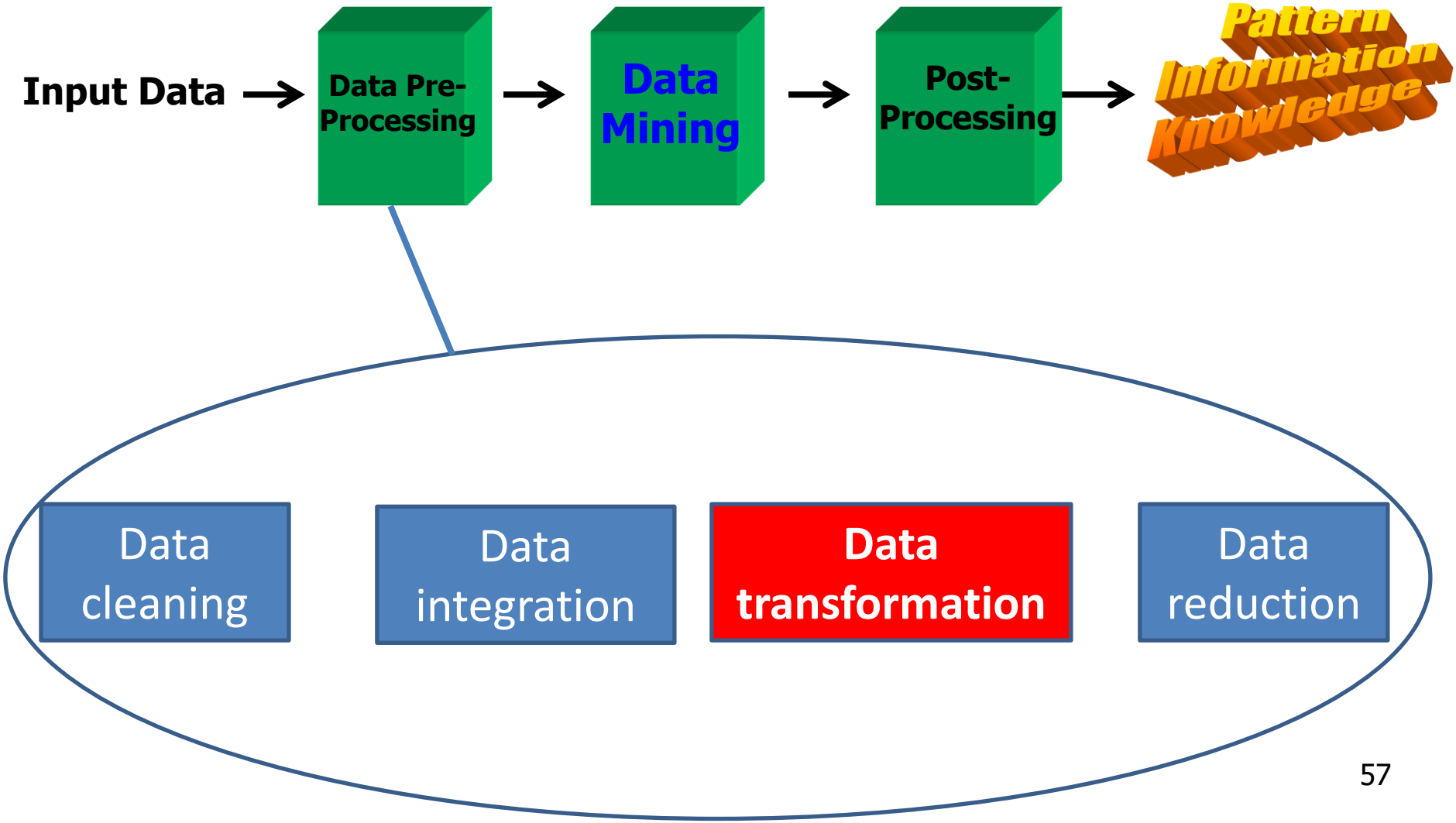
$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- The data are (5.1,1.4) (4.3,1.1) (7.0,4.7) (5.0,3.5) (7.1,5.9) (5.1,3.0).
  - E(A) = (5.1 + 4.3 +7.0 + 5.0 + 7.1 + 5.1)/ 6 = 5.6
  - E(B) = (1.4 + 1.1 + 4.7 + 3.5 + 5.9 + 3.0) /6 = 48/5 = 3.27
  - Cov(A,B) = (5.1*1.4 + 4.3*1.1 + 7.0*4.7 + 5.0*3.5 + 7.1*5.9 + 5.1*3.0)/6 – 5.6*3.27 = 19.91- 18.31 = 1.6
- Thus, A and B rise together since Cov(A, B) > 0.

(These data points are actually sepal length and petal length values from the Iris dataset!)


Input Data
Data Pre-Processing
Data Mining
Post-Processing
Pattern
Information
Knowledge
Data cleaning
Data integration
Data transformation
Data reduction

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values

- Methods:
  - **Smoothing:** Remove noise from data
  - **Aggregation:** Summarization, data cube construction
  - **Generalization:** Concept hierarchy climbing
  - **Normalization:** Scaled to fall within a small, specified range
  - **Attribute/feature construction**
    - New attributes constructed from the given ones

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 \$73,600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

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 \$73,600 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 $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

2.11. Use the two methods below to *normalize* the following group of data:

$$200, 300, 400, 600, 1000$$

(a) min-max normalization by setting $min = 0$ and $max = 1$

(b) z-score normalization

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

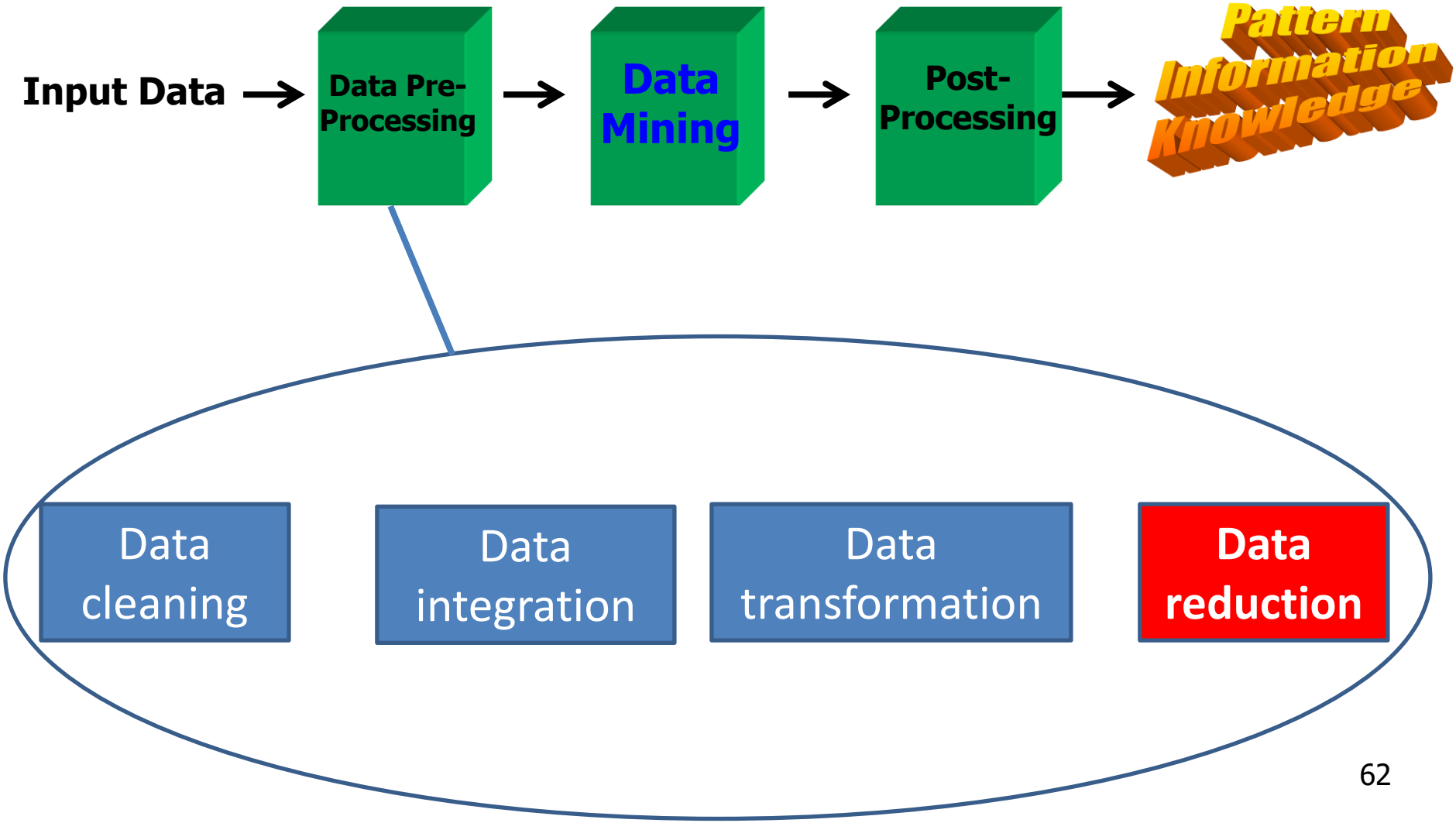
(a) min-max normalization by setting $min = 0$ and $max = 1$

original data	200	300	400	600	1000
[0,1] normalized	0	0.125	0.25	0.5	1

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

b) Z-score normalization

average	500				
stdev	316.2278				
original	200	300	400	600	1000
Zscore	-0.94868	-0.63246	-0.31623	0.316228	1.581139

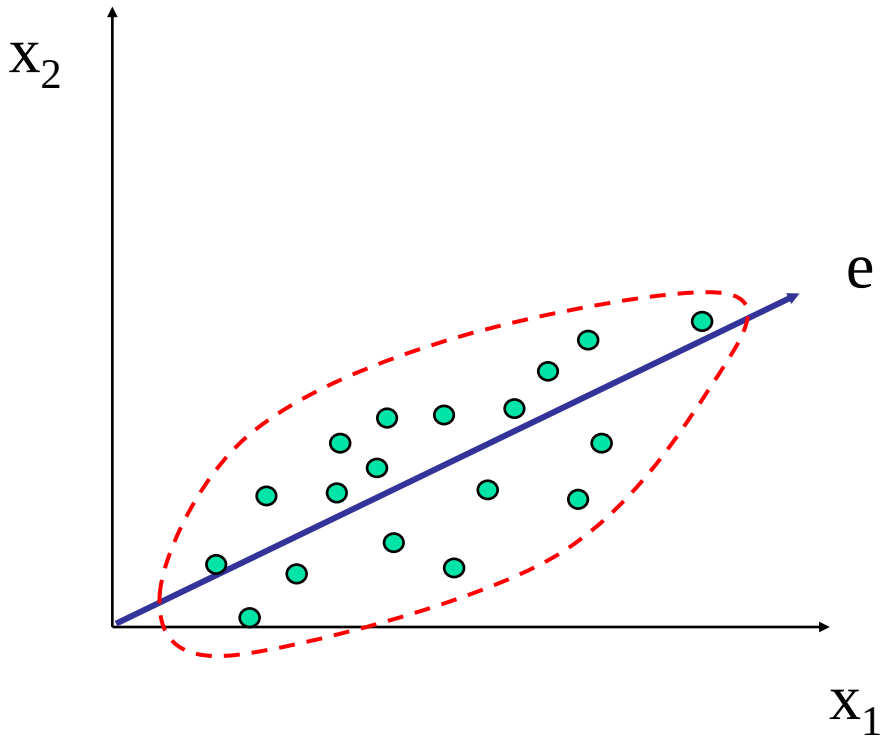

Input Data
Data Pre-Processing
Data Mining
Post-Processing
Pattern
Information
Knowledge
Data cleaning
Data integration
Data transformation
Data reduction

Principal component analysis (PCA)

- PCA is a method for *dimensionality reduction*
- Unsupervised method for identifying the important directions in a dataset
- We can then rotate the data into the (reduced) coordinate system that is given by those directions

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- Throw away the projected dimensions that capture less variations
- The original data are thus projected onto a lower dimensional space



Attribute 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

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Greedy algorithms are often used
  - Best single attribute chosen by significance tests, or classification performance
  - **Forward selection**:
    - The best single-attribute is picked first
    - Then next best attribute conditioned on the first, ...
  - **Backward elimination:**
    - Repeatedly eliminate the worst attribute
  - More sophisticated search options possible, e.g. branch and bound

Attribute Creation (Feature Generation)

- Adding transformed attributes can increase the flexibility of a model, e.g. letting a linear model capture non-linear relationships
  - Log, exp, or powers of attributes
  - Products of pairs of attributes (e.g. length * width = area)
  - Indicator functions for pairs of nominal attributes' values
  - Can sometimes be done implicitly by the learning algorithm (the kernel trick for SVMs)
  - Flexibility of concept representation should be considered

Attribute Creation (Feature Generation)

- Feature selection vs feature generation
  - **Feature selection**:
    - simplify data/models
    - improve interpretability
    - avoid overfitting
    - reduce running time (train and test)
  - **Feature generation**:
    - increase discriminative power, traded against the above
    - Most beneficial for simple methods, e.g. linear models
      - May be able to get away with a simpler model to achieve good performance
      - Therefore indirectly helps model simplicity, interpretability, overfitting, running time …

Considerations for Preprocessing

- Preprocessing changes the data, introduces bias.
- **Be careful that it won't hurt performance on new data!**
  - E.g. balancing the classes makes classification easier, but **biases the classifier to expect balanced classes**

Considerations for Preprocessing

- Think about likely **causes of noise** and errors when correcting them
  - E.g. is this "outlier" really an outlier, or is there a reasonable explanation for it?

Considerations for Preprocessing

- Consider the data mining **methods to be used**
  - Simple methods may benefit from feature generation.
  - Accurate distances are important for some classifiers, for which normalization may be important

Think-pair-share:
Predicting High School Student Performance

- Consider the Student Performance Data Set, at

  http://archive.ics.uci.edu/ml/datasets/Student+Performance
- The task is to predict student achievement (grades) in math and Portuguese, based on demographic, social, and school related features.
- Identify as many **preprocessing steps** as you can that might be useful for this dataset (and **explain why**).

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships
(numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school
(numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends
(numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption
(numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption
(numeric: from 1 - very low to 5 - very high)
29 health - current health status
(numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric)

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999
- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995