University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Data Mining Methodology and Preprocessing

Dr. Surbhi Khan

School of Science, Engineering & Environment
University of Salford

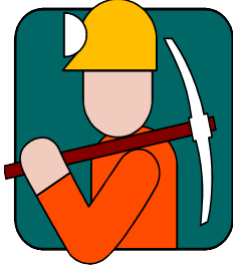
Outline

- Introduction
- Data Mining Methodology
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD
- Getting to Know Your Data
- Data Preprocessing

Recap: Why Data Mining?

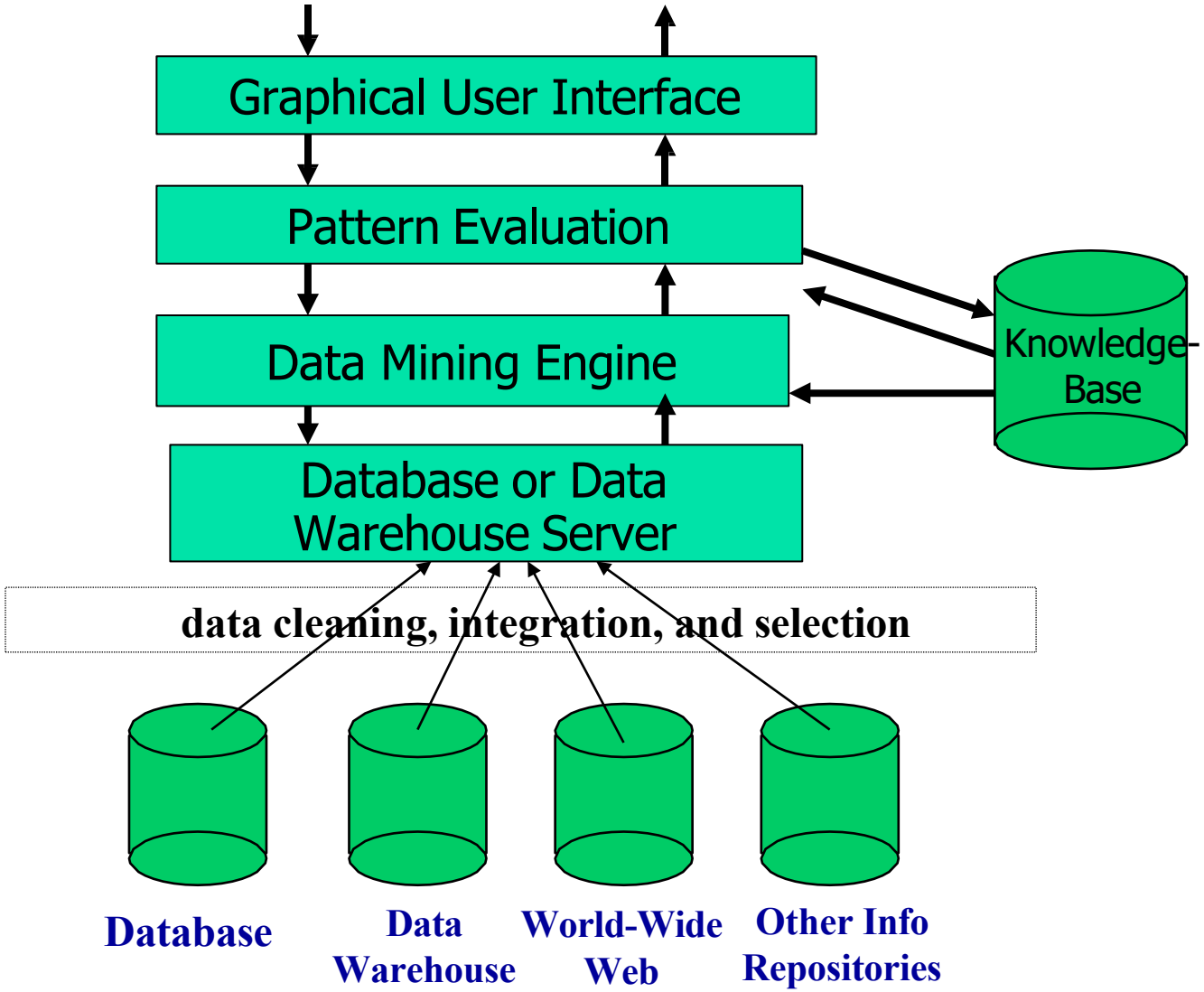
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

What Is Data Mining?



- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Architecture: Typical Data Mining System



Data Mining Methodology (Process)

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard methodologies:
  - **CRISP-DM**
    (Cross-Industry Standard Process for Data Mining)
  - **SEMMA**
    (Sample, Explore, Modify, Model, and Assess)
  - **KDD**
    (Knowledge Discovery in Databases)

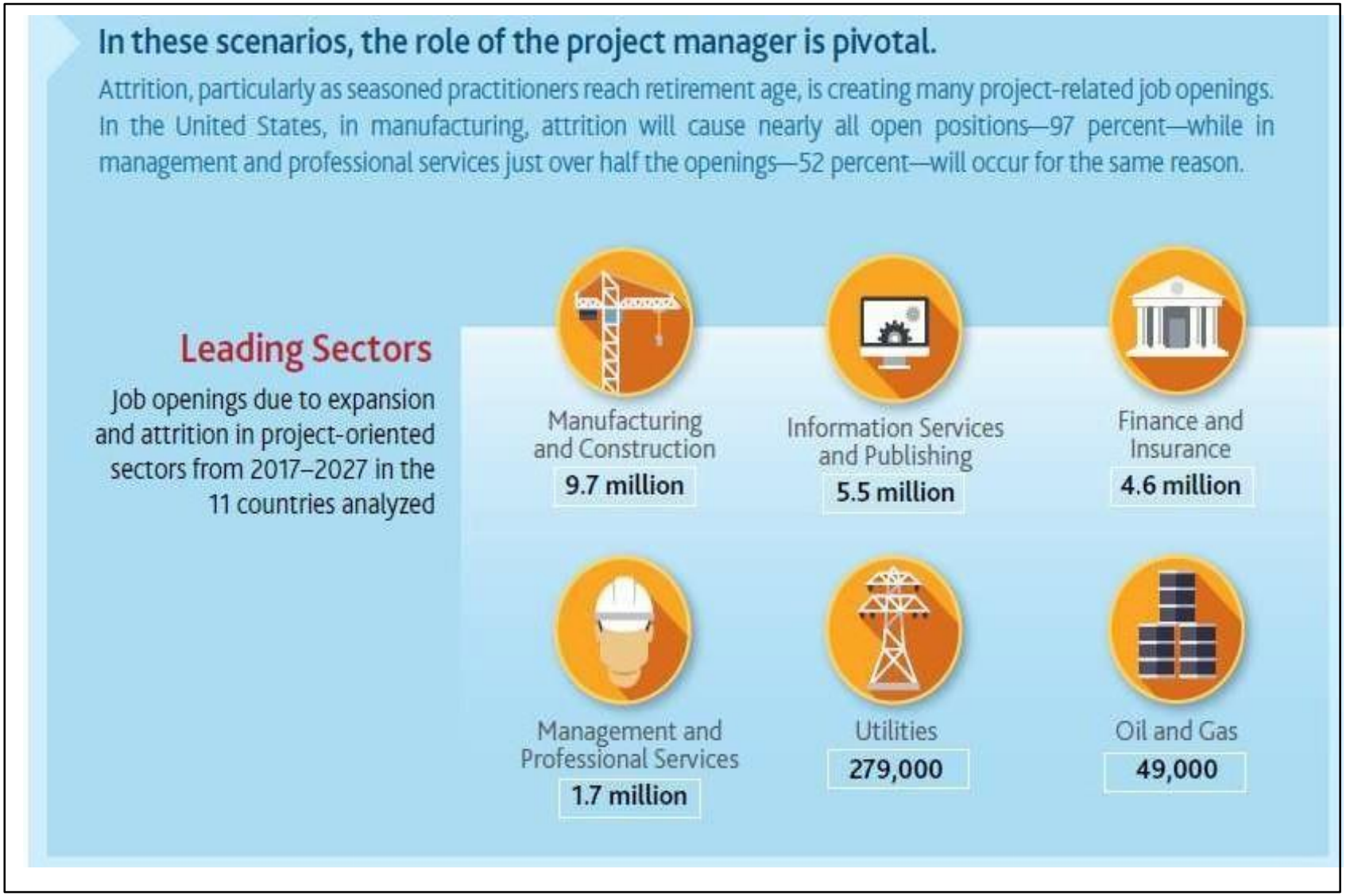
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Need of methodology in projects

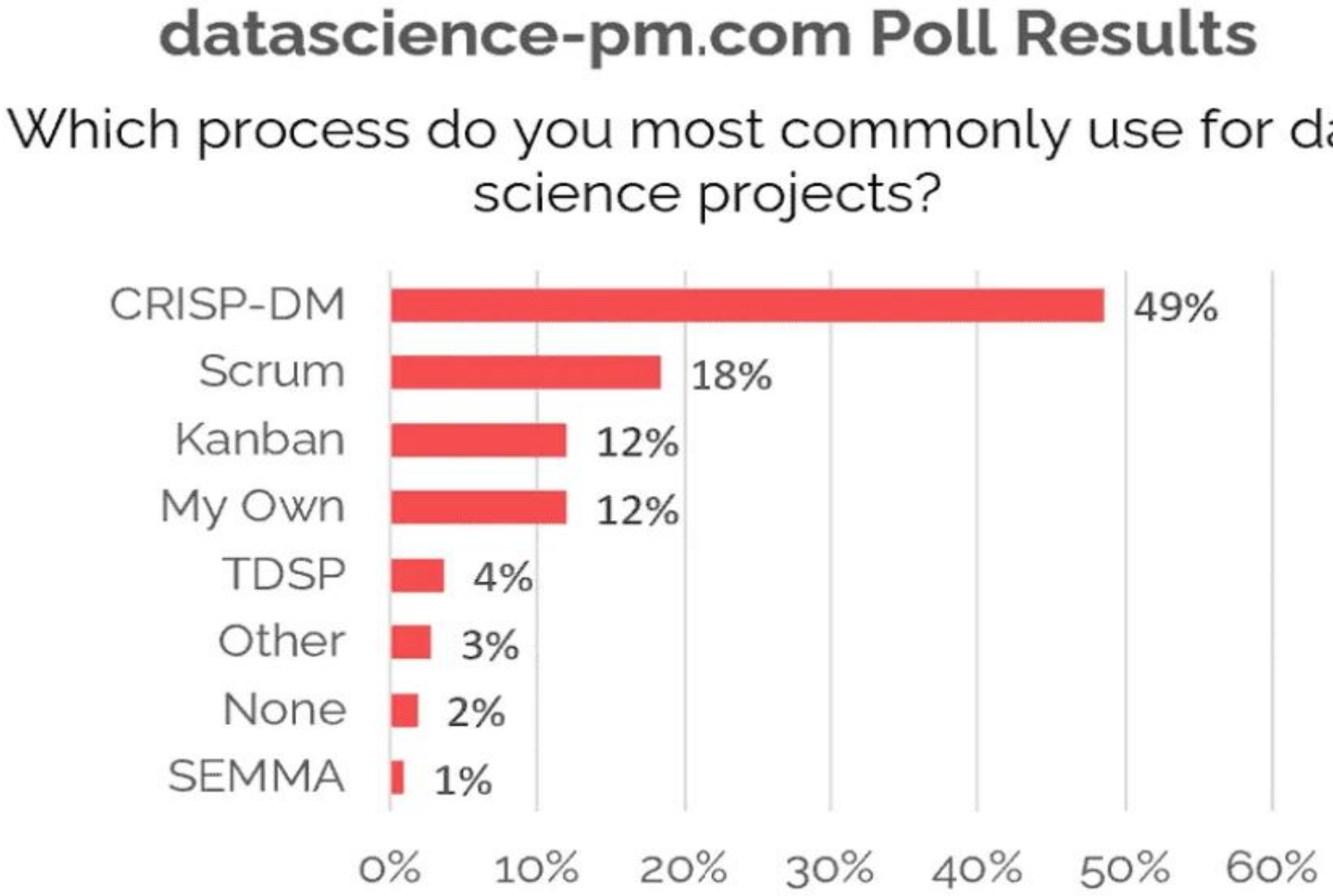
Employment Outlook in Selected Countries

Anderson Economic Group (AEG) and PMI analyzed project-oriented employment opportunity in 11 countries on five continents that represent developed and/or growing economic powers.

Leading Sectors



What main methodology are you using for your analytics, data mining, or data science projects ?



TDSP, Scrum & Kanban

TDSP (Team Data Science Process): Launched by Microsoft in 2016, TDSP defines 5 stages of the data science life cycle (Business understanding, Data Acquisition & Understanding, Modeling, Deployment, and Customer Acceptance)

- Scrum: Scrum is most commonly used framework for software development projects and the de facto agile project management framework. It divides a project into a series of mini-projects, each of a consistent and fixed-length, called a sprint. Scrum also defines meetings and roles to help guide a team in executing a project.

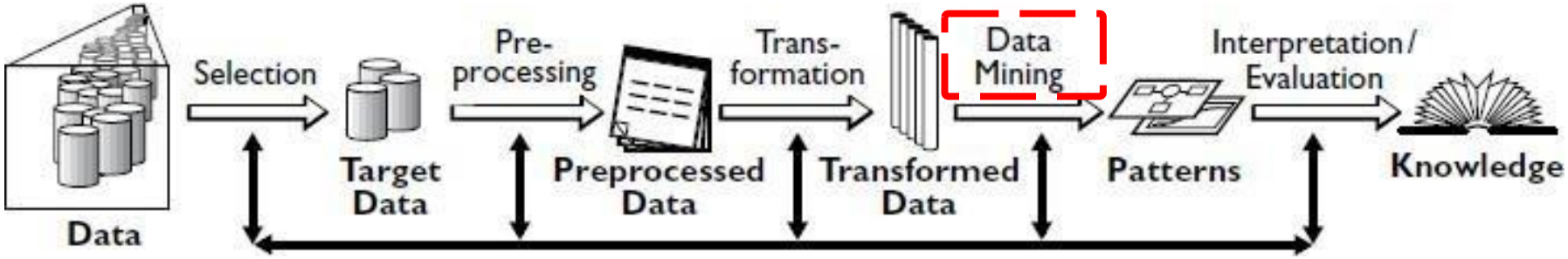
- Kanban: Kanban's two key Principles are to (1) visualize the flow and (2) minimize work-in-progress. In short, by limiting tasks that are being completed simultaneously, Kanban enables agility.

Why Should There be a Standard Process?

The data mining process must be reliable and repeatable by people with little data mining background.

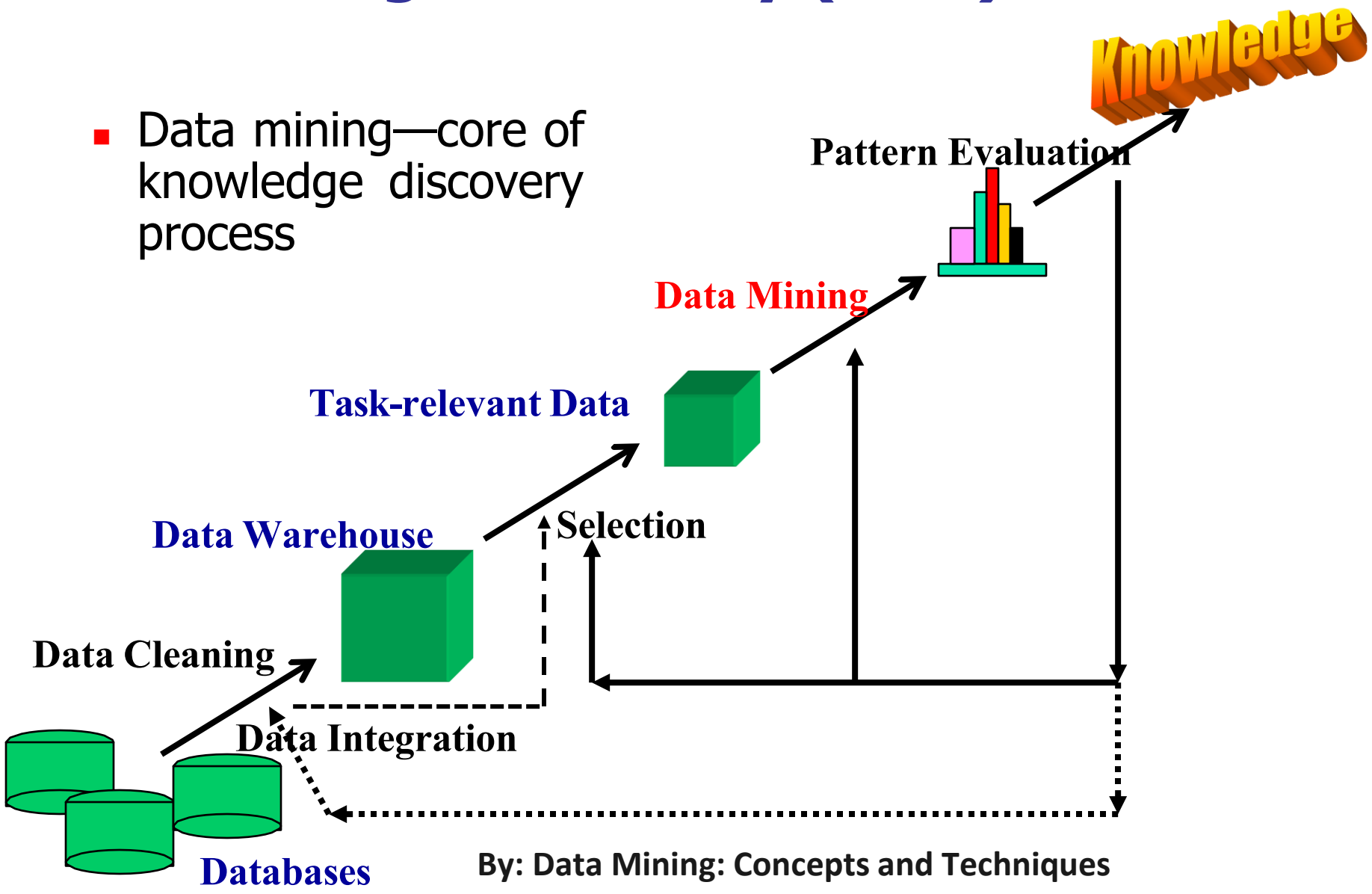
- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining.

Knowledge Discovery in Databases (KDD)



Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



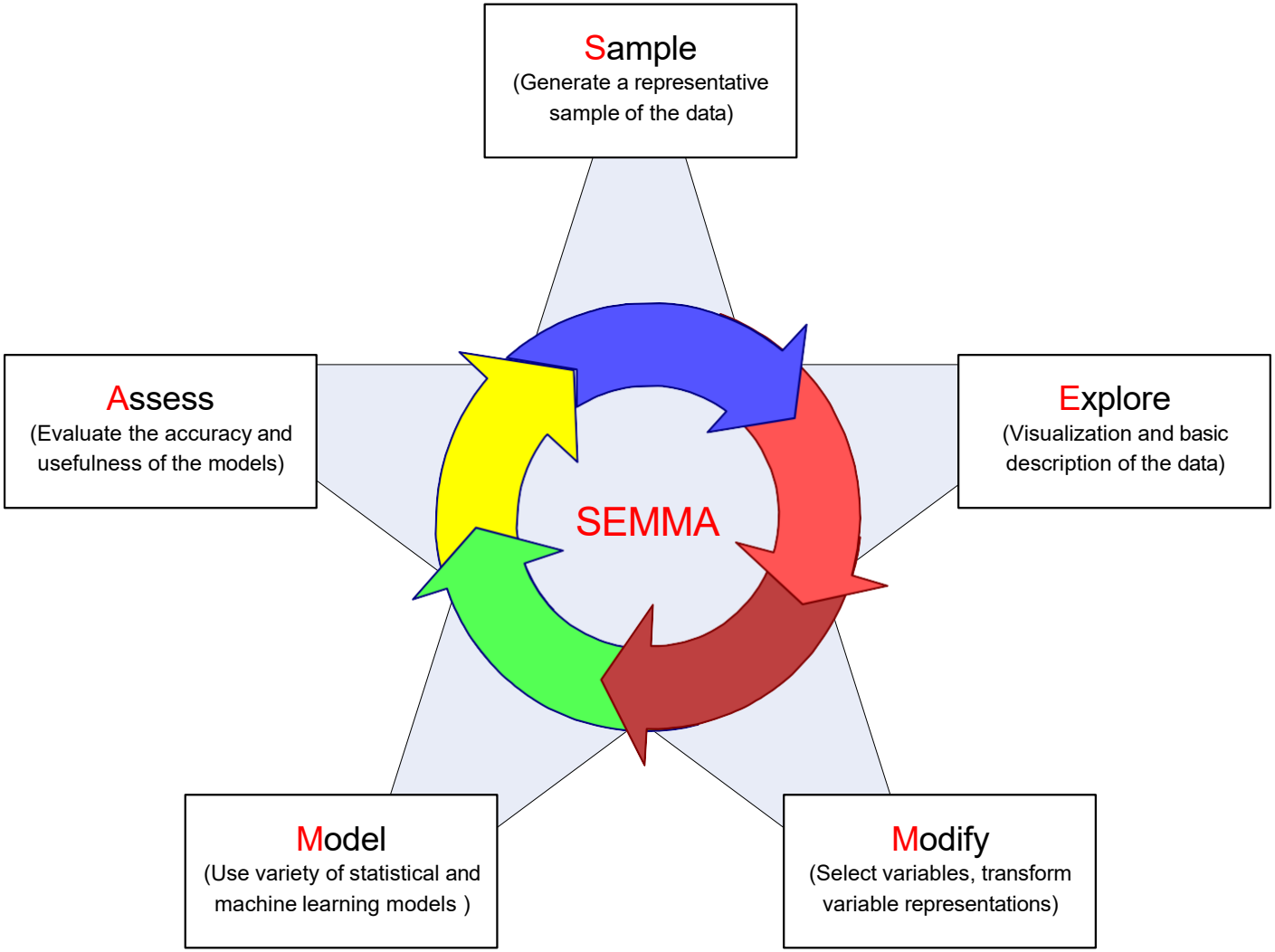
By: Data Mining: Concepts and Techniques
by Jian Pei, Micheline Kamber

SEMMA methodology (SAS Enterprise Miner)

- The core process of conducting data mining study includes the following steps (SEMMA):
  - Sample
  - Explore
  - Modify
  - Model
  - Assess

- SEMMA is a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining.
- SEMMA is focused on the model development aspects of data mining

SEMMA


Sample
(Generate a representative sample of the data)
Explore
(Visualization and basic description of the data)
Modify
(Select variables, transform variable representations)
Model
(Use variety of statistical and machine learning models)
Assess
(Evaluate the accuracy and usefulness of the models)
SEMMA

CRISP-DM

- Cross-Industry Standard Process for Data Mining (CRISP-DM) Launched in 1996
- SPSS/ISL,NCR,Daimler-DEMZ, Ohra
- European Community funded effort to develop framework for data mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process
  - Take the mystery/high-priced expertise out of simple data mining tasks

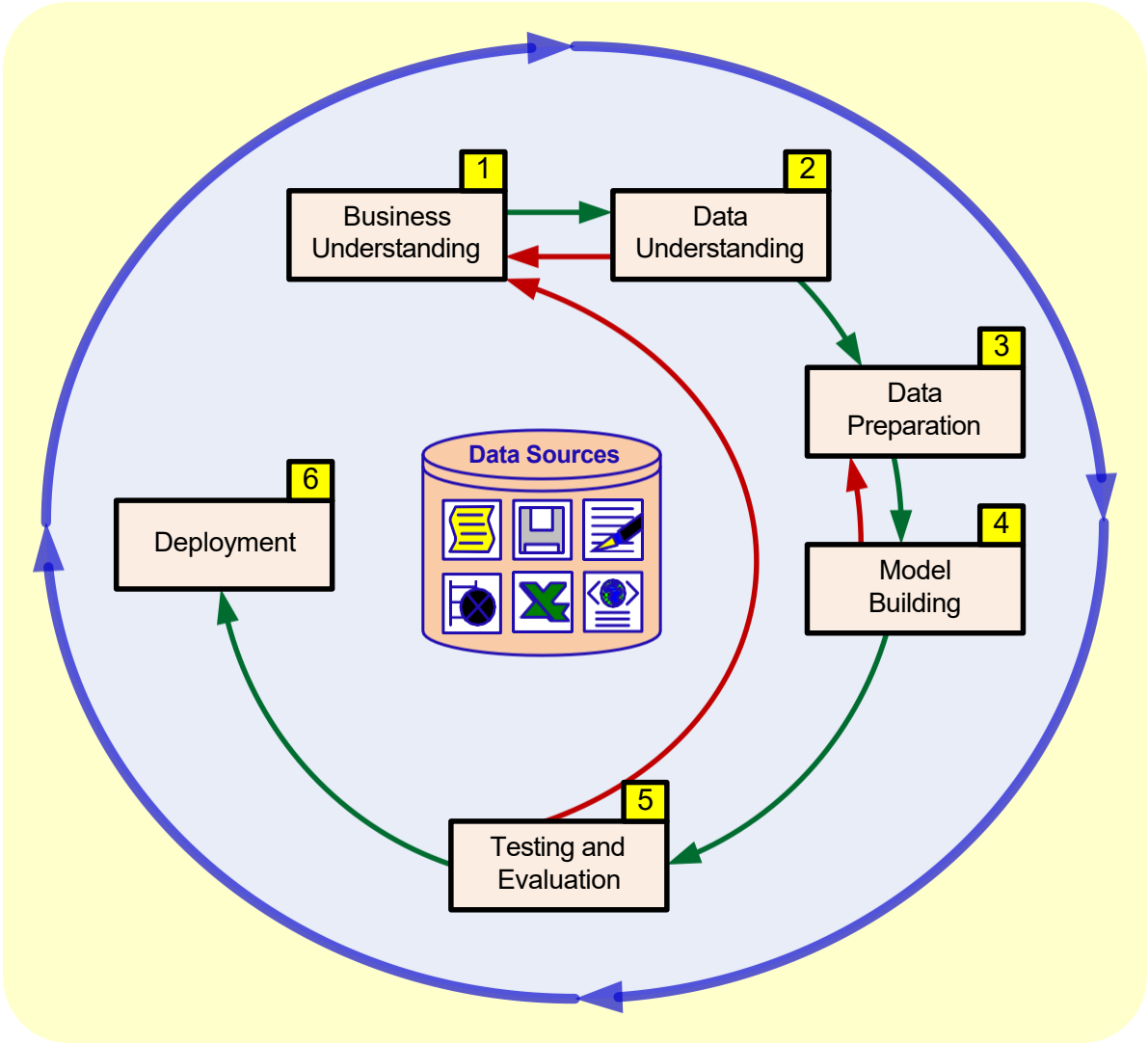
Process Standardization

- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996
- SPSS/ISL, NCR, Daimler-Benz, OHRA
- Funding from European commission
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

Data Mining Methodology: CRISP-DM



CRISP-DM: Phases

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data
Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evalution

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

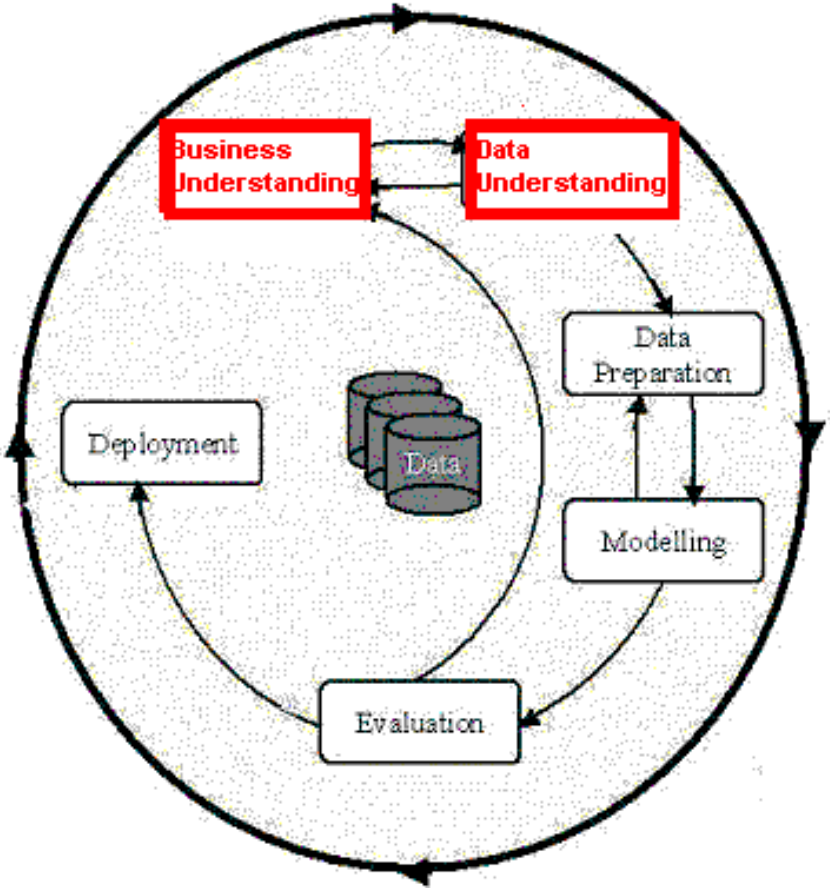
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Phases in the DM Process (1 & 2)

- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria

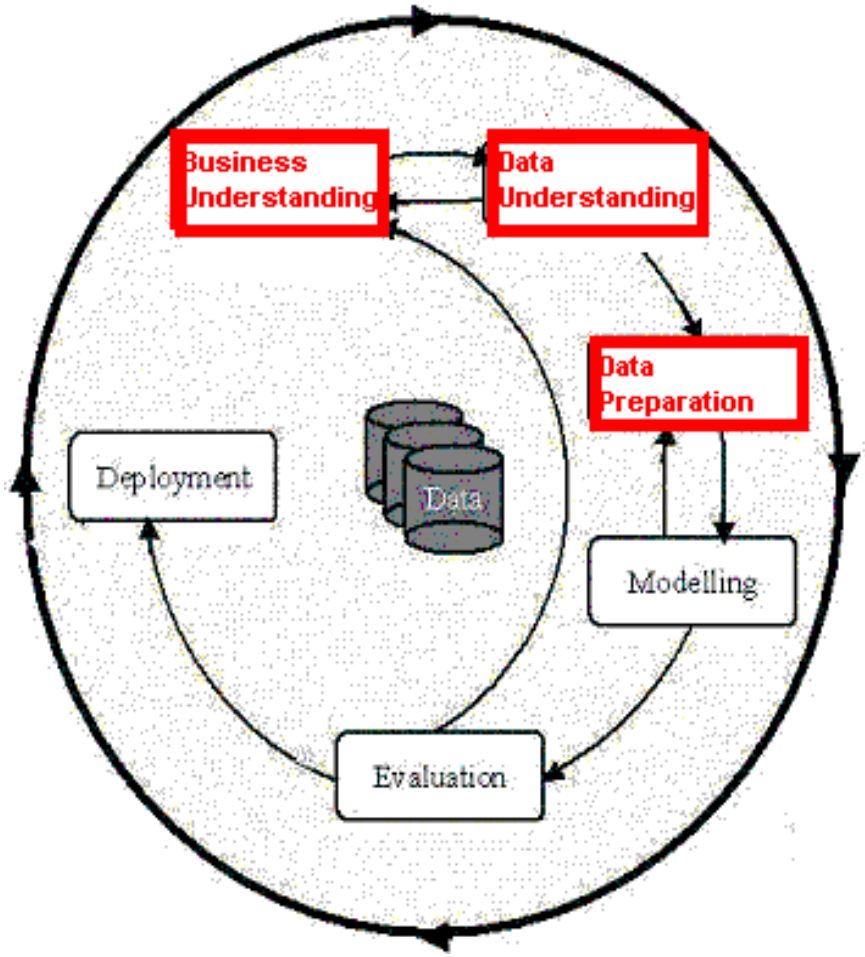


- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

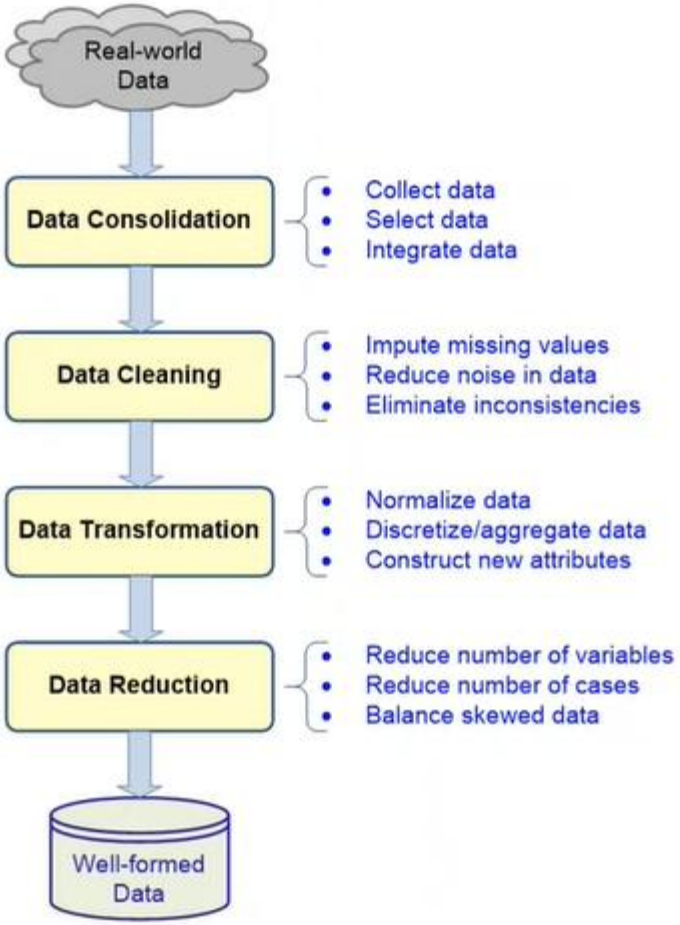
Phases in the DM Process (3)

Data preparation:

- Takes usually over 80% of the time
  - Collection
  - Assessment
  - Consolidation and Cleaning
    - table links, aggregation level, missing values, etc
  - Data selection
    - active role in ignoring non-contributory data?
    - outliers?
    - Use of samples
    - visualization tools
  - Transformations - create new variables

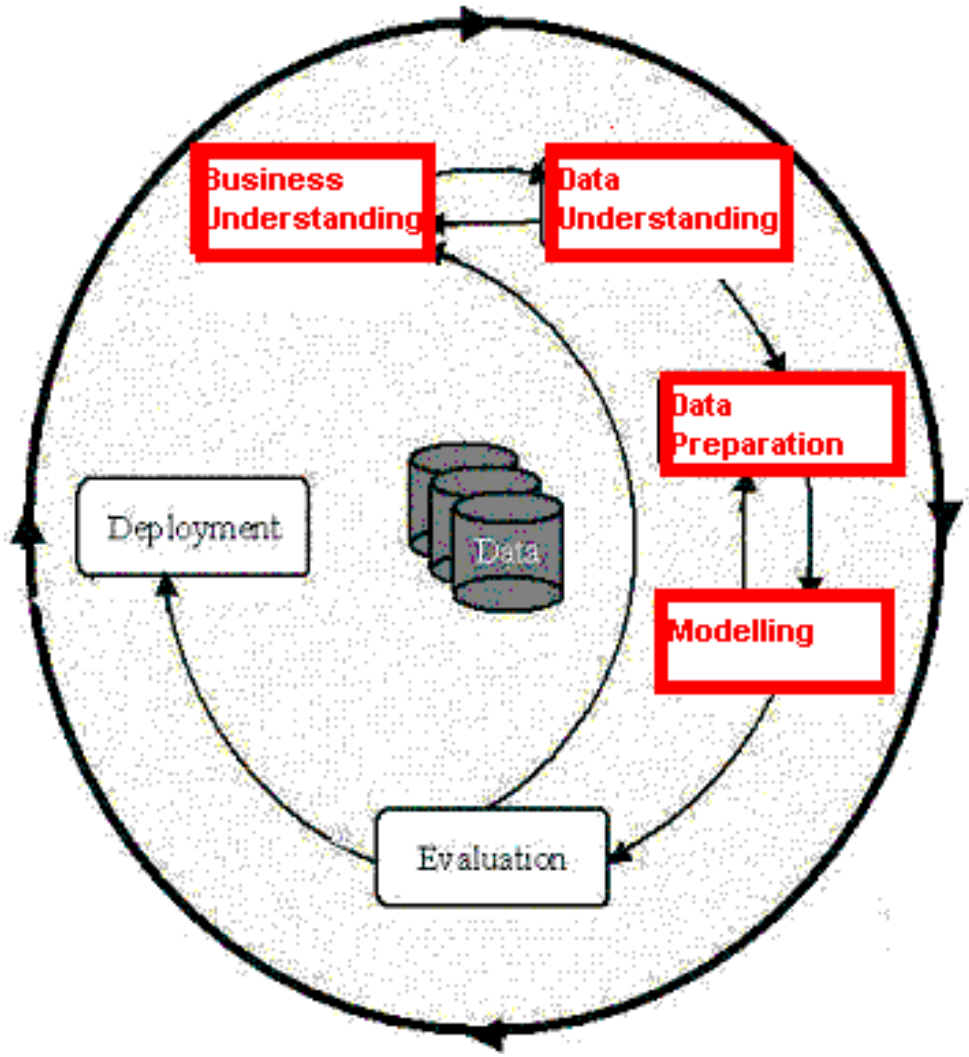


Data Preparation Task



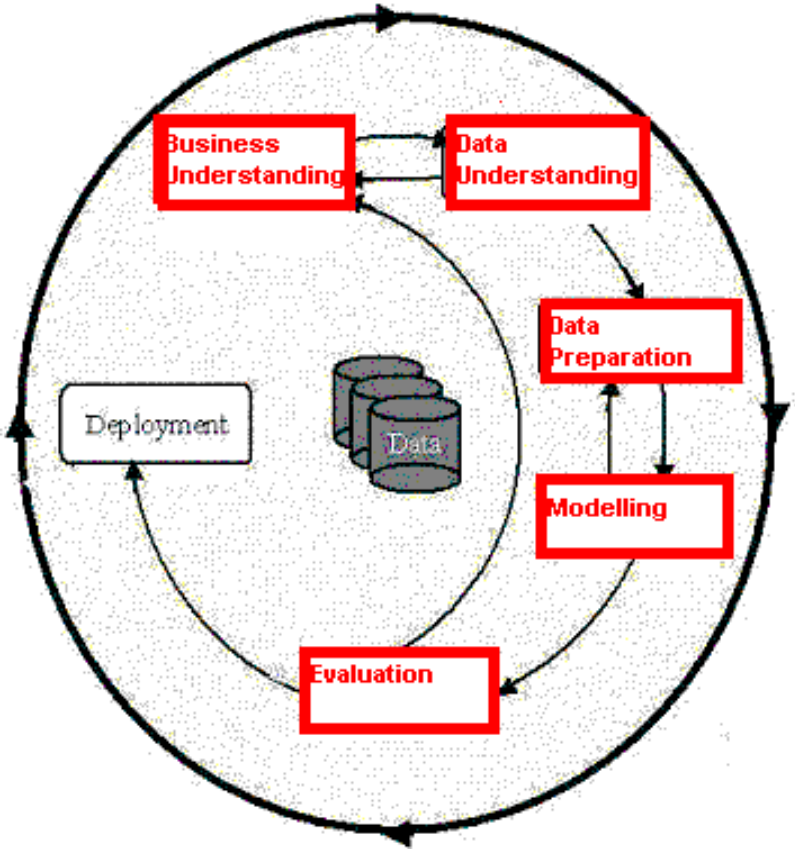
Phases in the DM Process (4)

- Model building
  - Selection of the modeling techniques is based upon the data mining objective
  - Modeling is an iterative process - different for supervised and unsupervised learning
    - May model for either description or prediction



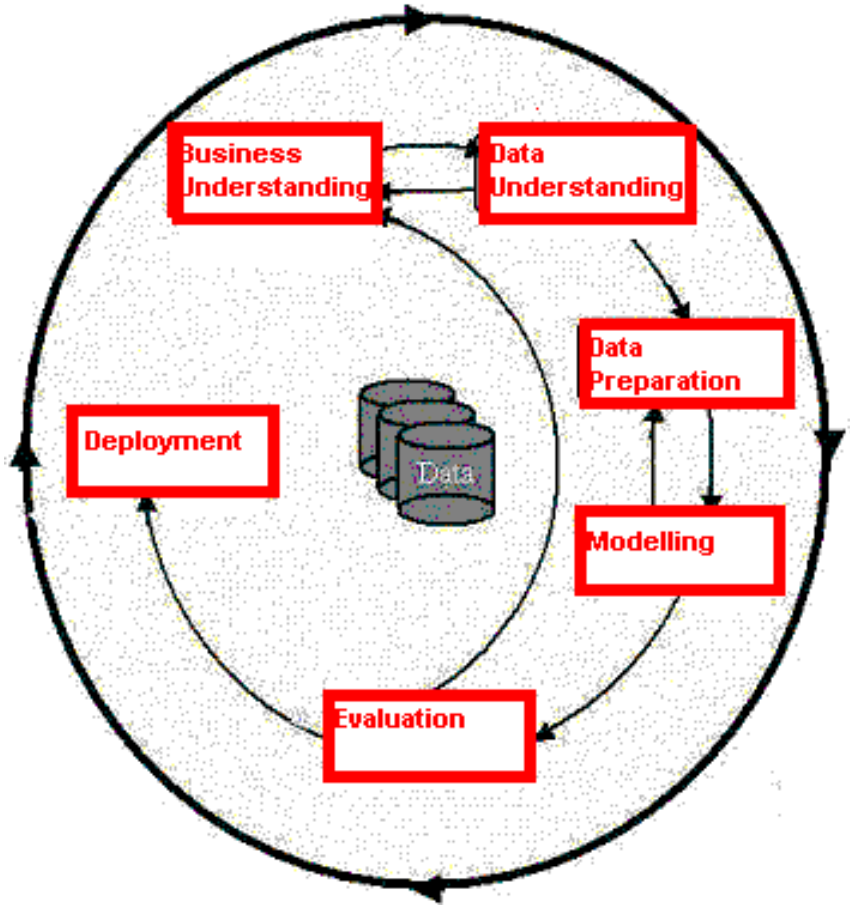
Phases in the DM Process (5)

- Model Evaluation
  - Evaluation of model: how well it performed on test data
  - Methods and criteria depend on model type:
    - e.g. matrix with classification models, mean error rate with regression models
  - Interpretation of model: important or not, easy or hard depends on algorithm

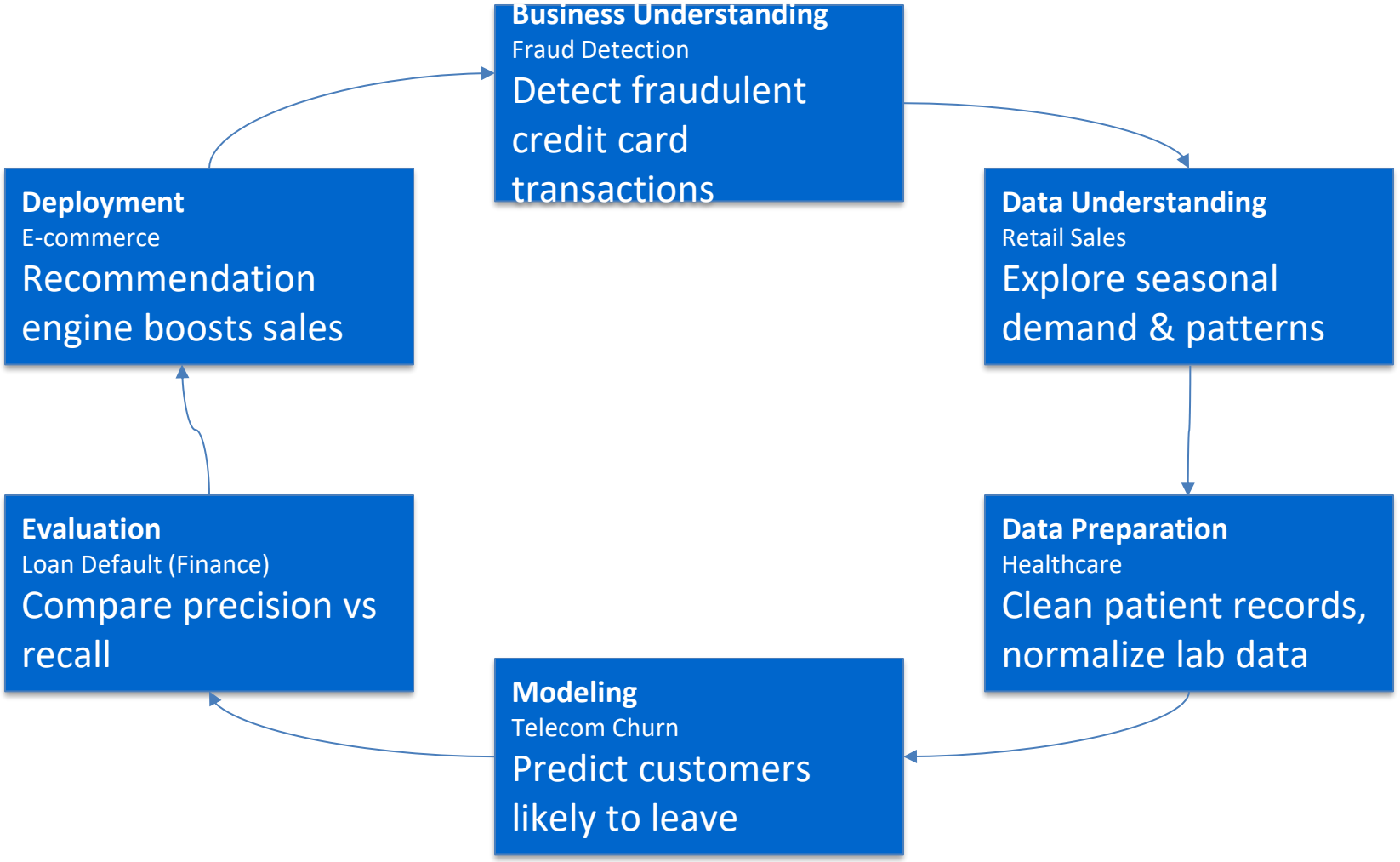


Phases in the DM Process (6)

- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Utilizing results as business rules
  - interactive scoring on-line



Case Study – Data Mining Methodology in Action



CRISP-DM: Details

- Available on-line: www.crisp-dm.org
  - 20 pages model (overview)
  - 30 page user guide (step-by-step process, hints)
  - 10 page “output” (suggested outline for a report on a data mining project)
- Has SPSS written all over it
  - But not a plug for a product (or even customized toward that product)

Summary of Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

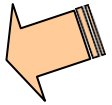
Sample Question

You have been given a project to construct a small dam. Currently, you are visiting your project stakeholders and asking about their needs and expectations. You also have had many workshops and brainstorming sessions with them. Which process is this?

(a) Identify stakeholders
(b) Collect requirements
(c) Define scope
(d) Control scope

Getting to Know Your Data

- Data Objects and Attribute Types



- Basic Statistical Descriptions of Data
- Data Visualization

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
  - Ordinal

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings
- **Numeric**
  - Quantity (integer or real-valued)

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Graphical analysis
  - Boxplot
  - Histograms
  - Scatter Plots

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

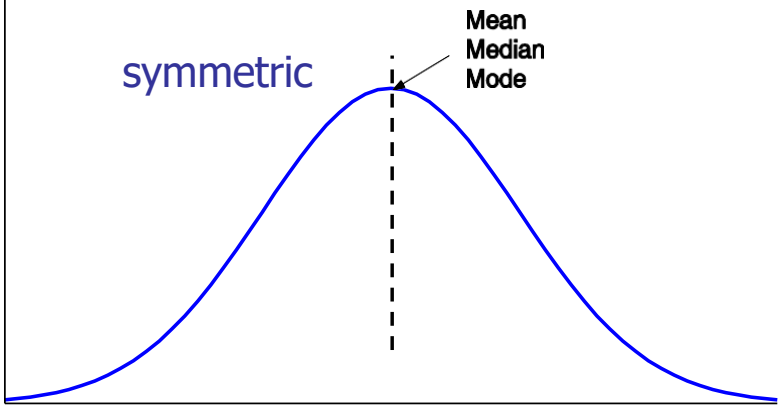
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

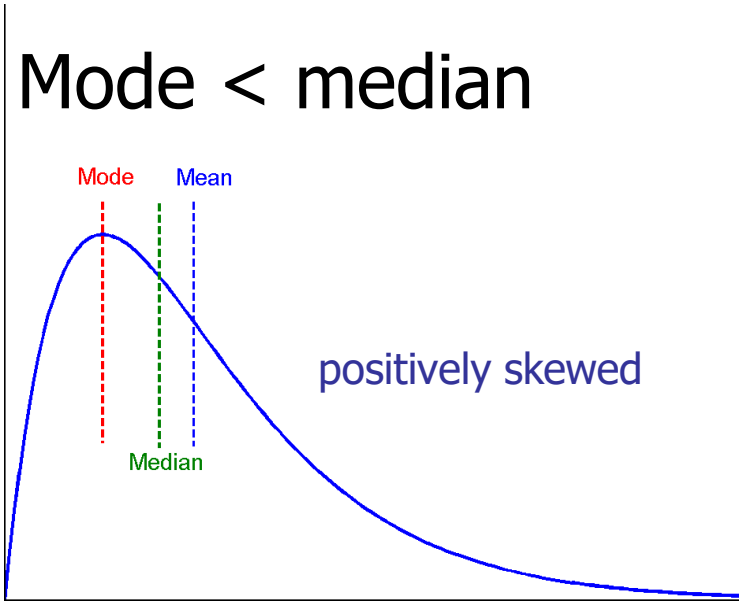
  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

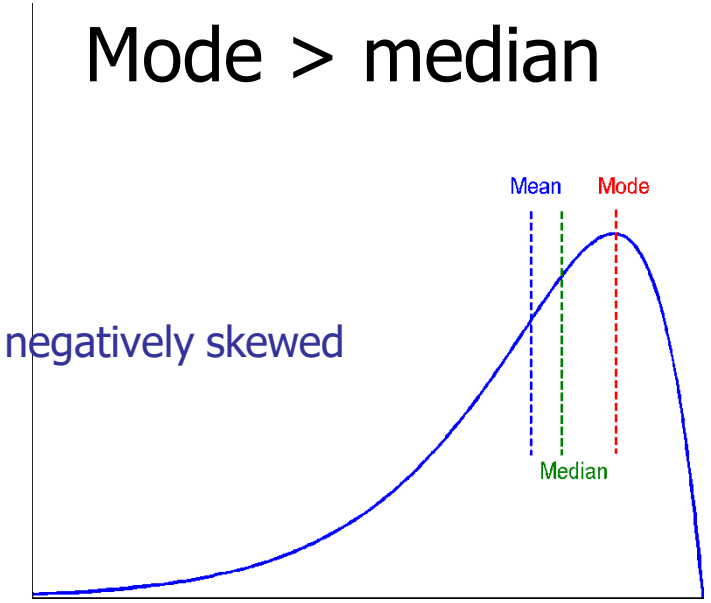
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Note: sort the given data before computing median
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Note: sort the given data before computing model

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th 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
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

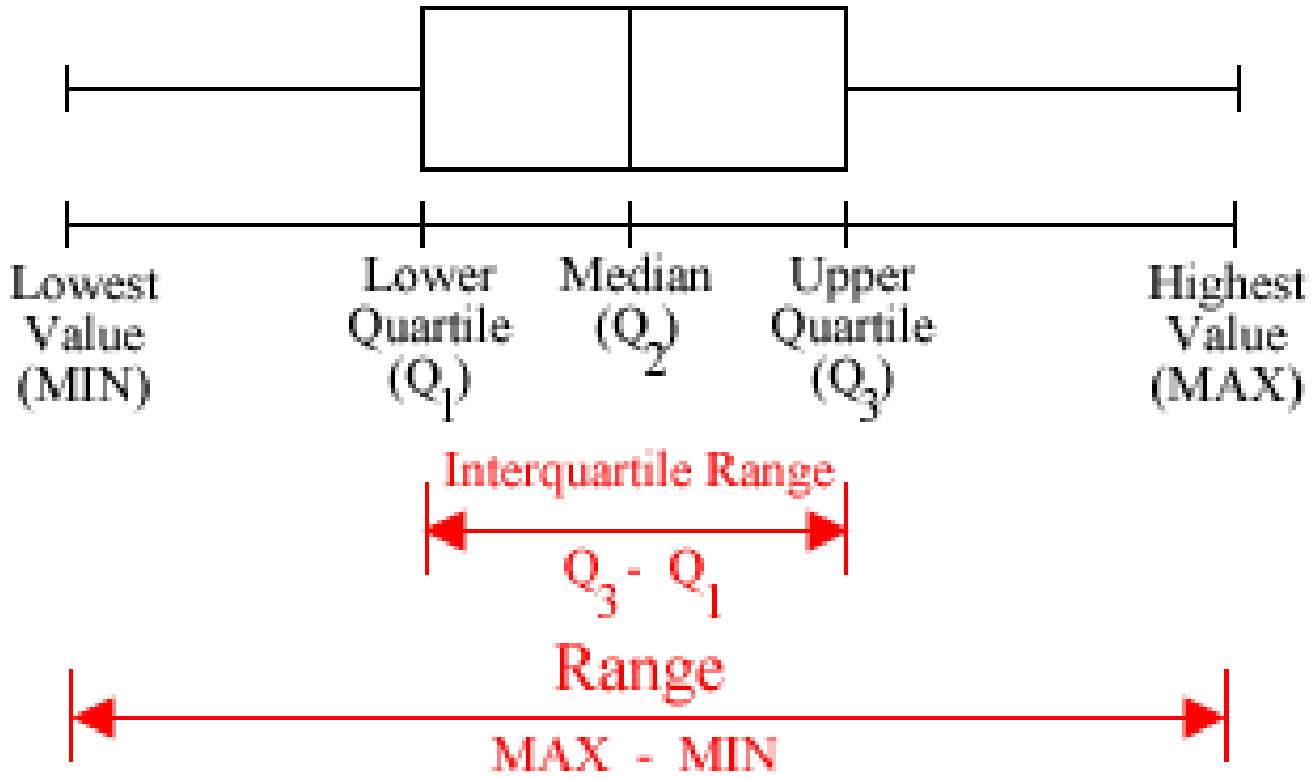
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Example

- Data: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110
- Solution
  - Data already in sorted order
  - Total data = 12
  - Q1 = 25th percentile = 3rd number here = 47
  - Q3 = 9th number = 63
  - Q2 = 6th number = median = 52
  - IQR = 63-47 = 16

Box plot

- Drawing

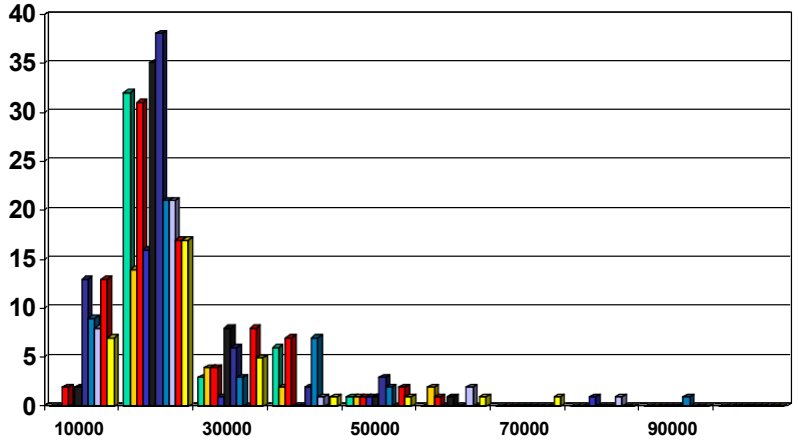


Graphic Displays of Basic Statistical Descriptions

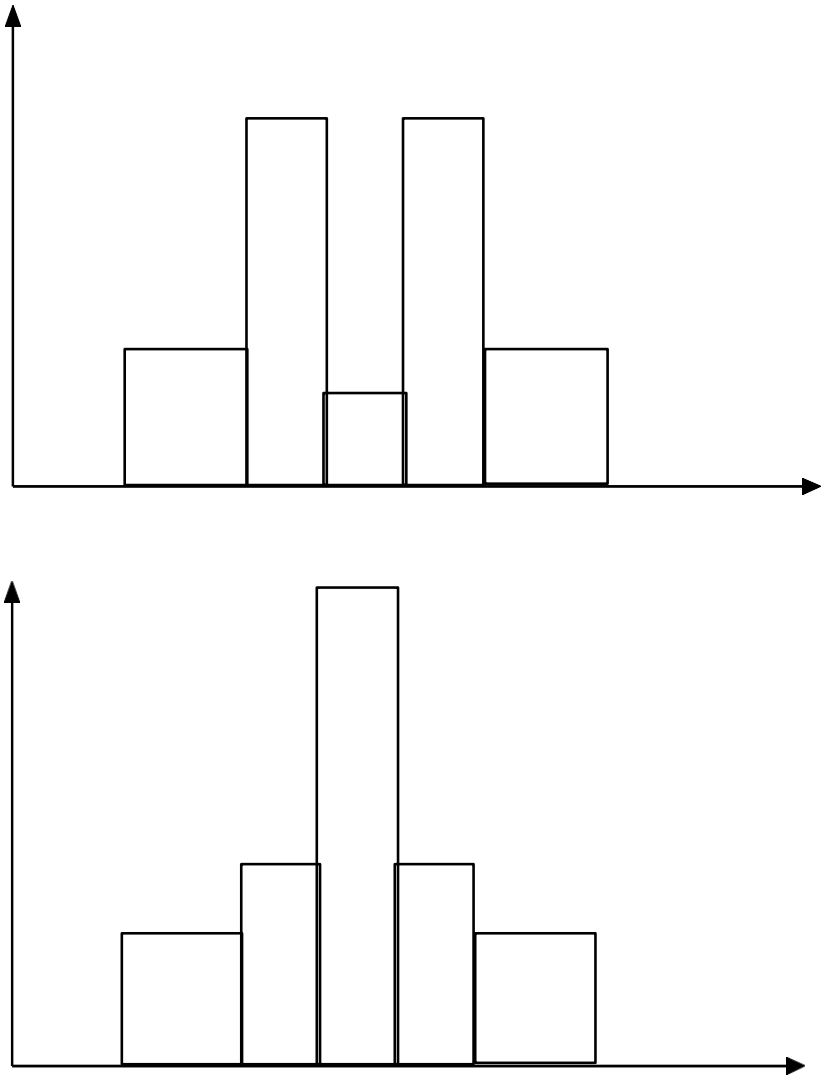
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



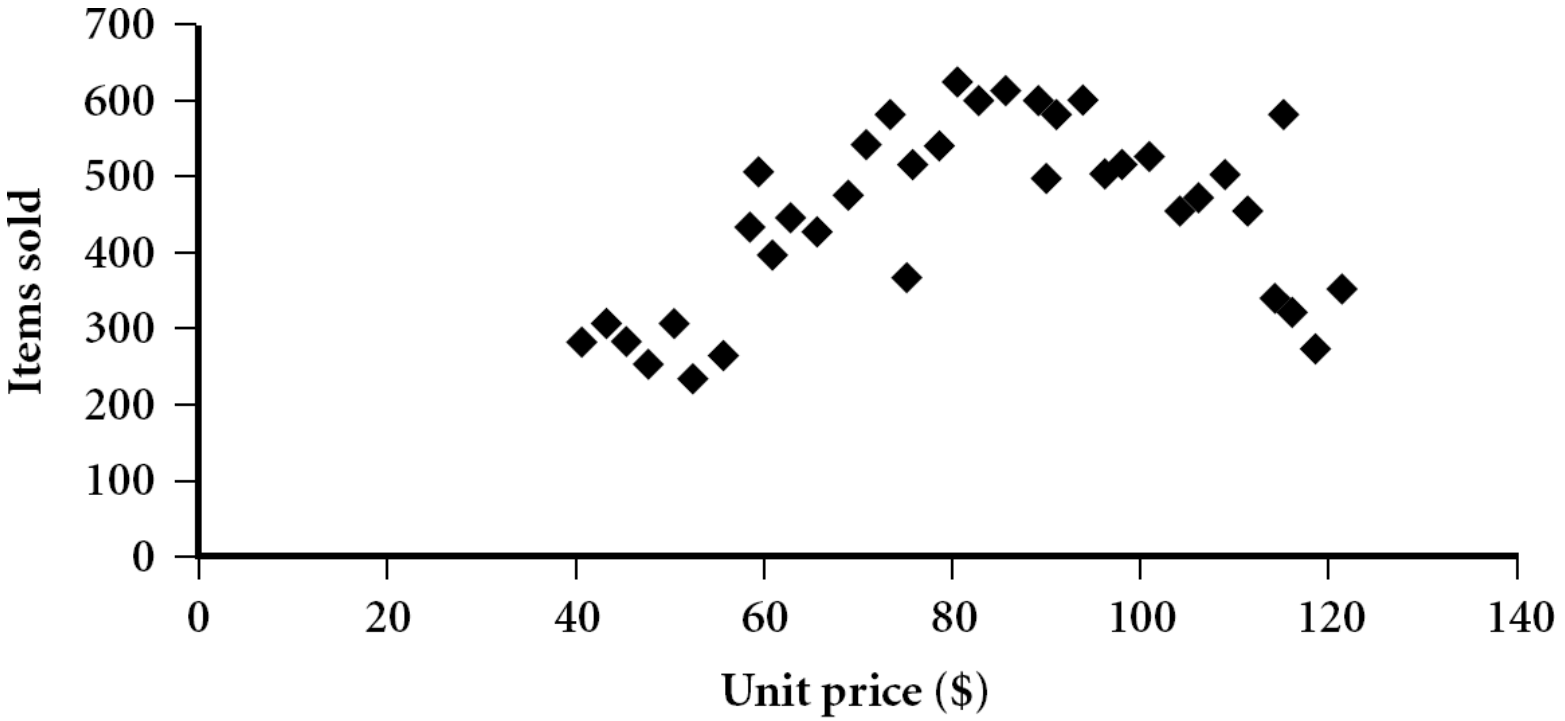
Histograms Often Tell More than Boxplots



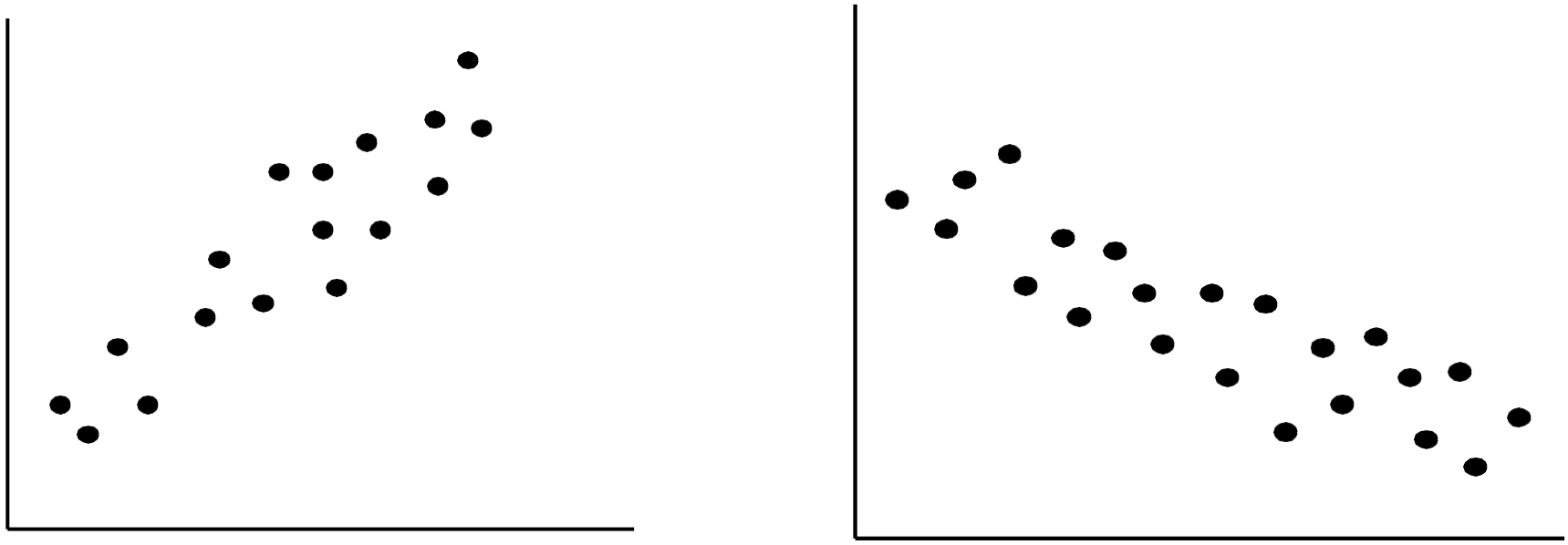
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

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

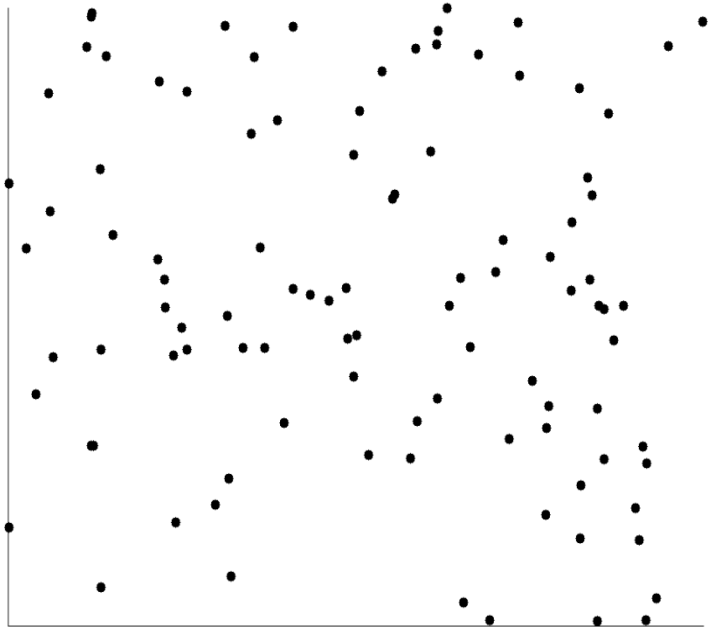


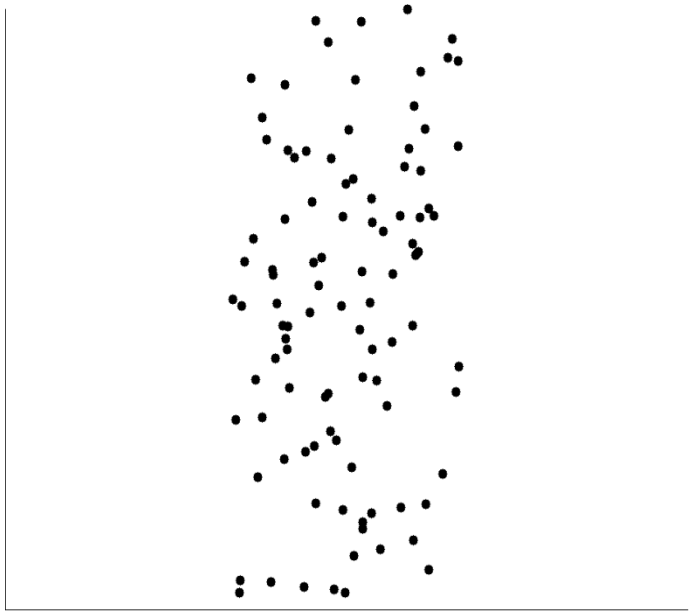
Positively and Negatively Correlated Data



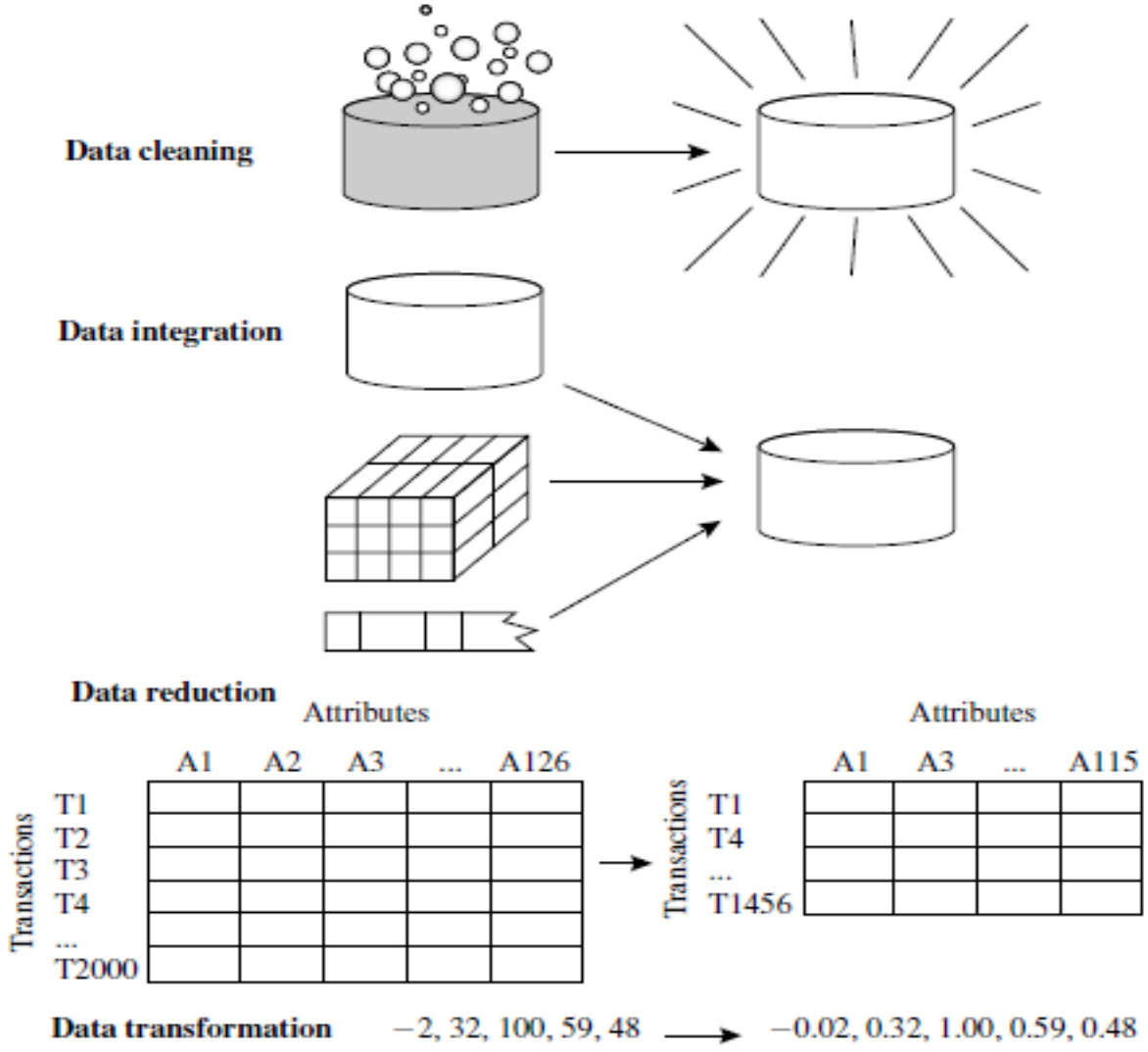
- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data





Major Tasks in Data Preprocessing



Details of Tasks

- **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 reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation = " " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., Salary = "−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - Age = "42", Birthday = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?

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 be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data? - Data smoothing methods

- 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. (Example: next slide)
- Regression
  - smooth by fitting the data into regression functions. Find the best line to fit 2 attributes so that one attribute can be used to predict other
- Clustering
  - detect and remove outliers. So values that fall outside clusters are outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Example data: 4,8,15,21,21,24,25,28,34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Integration

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

- Redundancy is a problem here. 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
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality
- Redundant attributes can be detected by correlation analysis and covariance analysis

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. So mining on the reduced dataset is more efficient
- 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, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Data Transformation

- Data are transformed or consolidated into forms appropriate for mining
- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction

Cond..

- Normalization: Scaled to fall within a smaller, specified range such as [−1, 1] or [0.0, 1.0].
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- Data mining systems and architectures
- CRISP-DM (Cross-Industry Standard Process for Data Mining)
- SEMMA (Sample, Explore, Modify, Model, and Assess)
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

Breakout room

- Suppose you are hired as Data Miner for an Electronic company. The company needs to discover interesting patterns related to its customers. Your advisor has asked you to prepare sample of the customer data as shown below.

No	ID	First Name	Last Name	Gender	Age	City	Number of Orders in 2023	Total Revenue from the customer	Total Tax charged (GBP)
1	10001	Mason	Smith	Male	19	Belfast	10	100	5
2	10002	Jackson	Jones	Male	53	Birmingham	3	1000	50
3	10003	Alex	Taylor	Female	28	Bristol	-30	6000	300
4	10006	Jack	Williams	Male	23	Manchester	10	3000	150
5	10005	Ella	Johnson	Female	300	Cardiff	35	7000	350
6	10004	Wyatt	Wilson	Male	30	Leeds	4	2000	100
7	10007	Rose	Campbell	Male	32	Liverpl	29	1000	000008
8	10010	Grayson	Evans		27	Newcastle	14	1550	77.5
9	10003	Asma	Alkhaled	Female	28	Nottingham	30	6000	300
10	10008	◣	Wright	Male	24	Sheffield	1	399	19.95
11	10009	Ibrahem	Almansour	Male	18	London	12	980	49

Question to Solve

A) **What are the data issues that you have found in the dataset?**

B) **How would you resolve these issues (from the previous question – A).**