MSc GFIS

Information Insights I

Objectives:

- Business analytics
- Data mining
- Goal, design, techniques & implementation of data mining (CRISP-DM)
- Data mining application
- Understand data

Technologies to Create Insights: Using Data Mining to Create New Explicit Knowledge

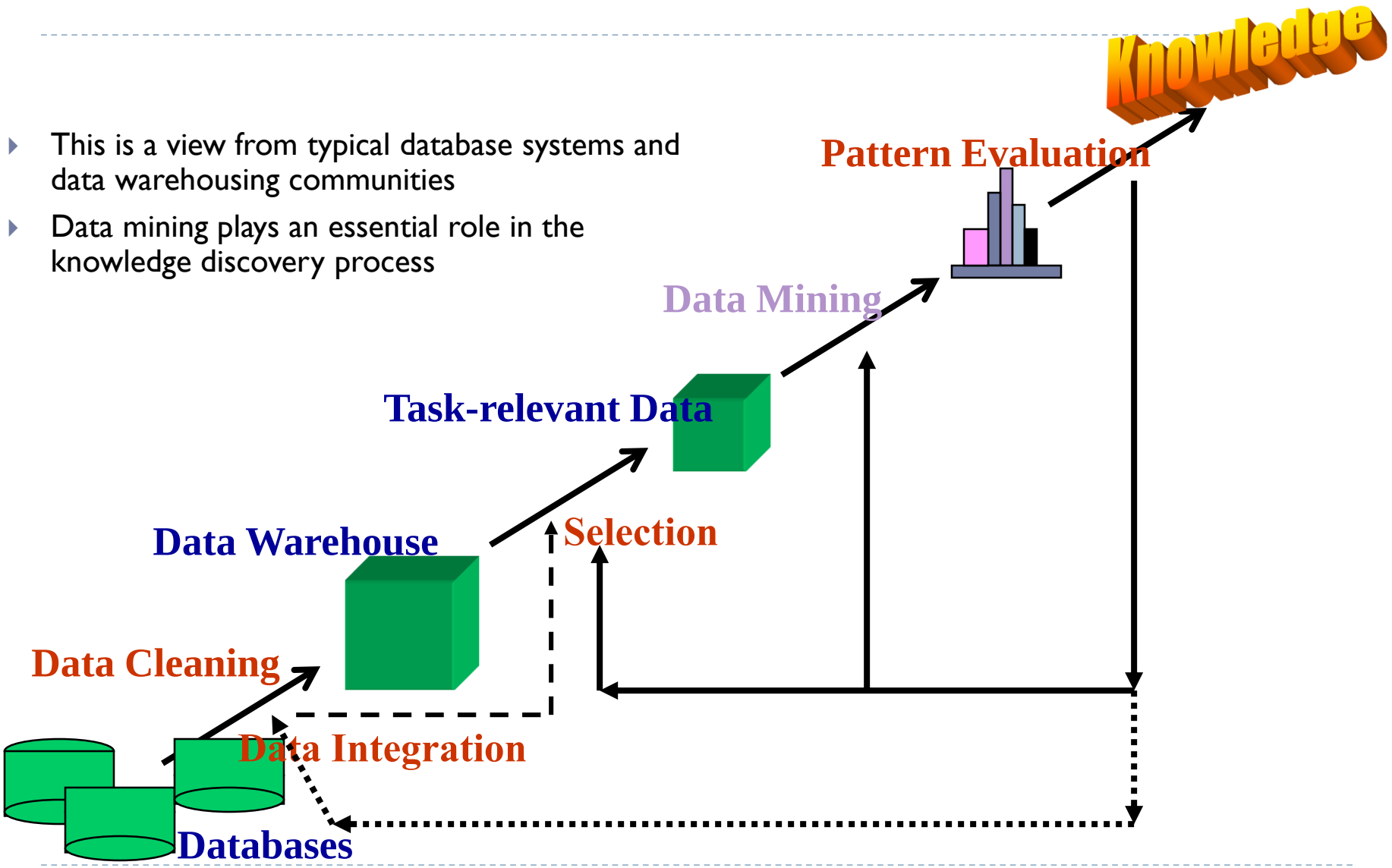
- Business analytics
- *Knowledge discovery in databases* (KDD), or more commonly, *data mining* (DM)
- Analytics comes with hypothesis testing
- Data mining is more the act of discovery that lacks a hypothesis

Business Analytics

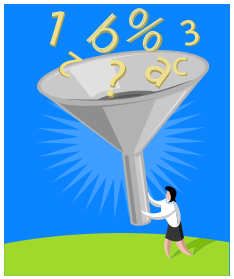
- The essence of business analytics is to embed analytical decisions into business processes on an ongoing, repeatable basis instead of treating analytics as an ad hoc activity (Davenport and Harris, 2007).
- Relevant technologies for business analytics, including different data mining technologies that have their origin in statistics and artificial intelligence are covered in this section of the module.
- Firms need to follow a business analytics strategy to embed analytics in their most important business processes.

Knowledge Discovery from Data (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining



- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases - Fayyad et al., (1996)
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable
- Data mining: a misnomer?
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging

Watch out: Is everything "data mining"?

The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases

Data Mining:

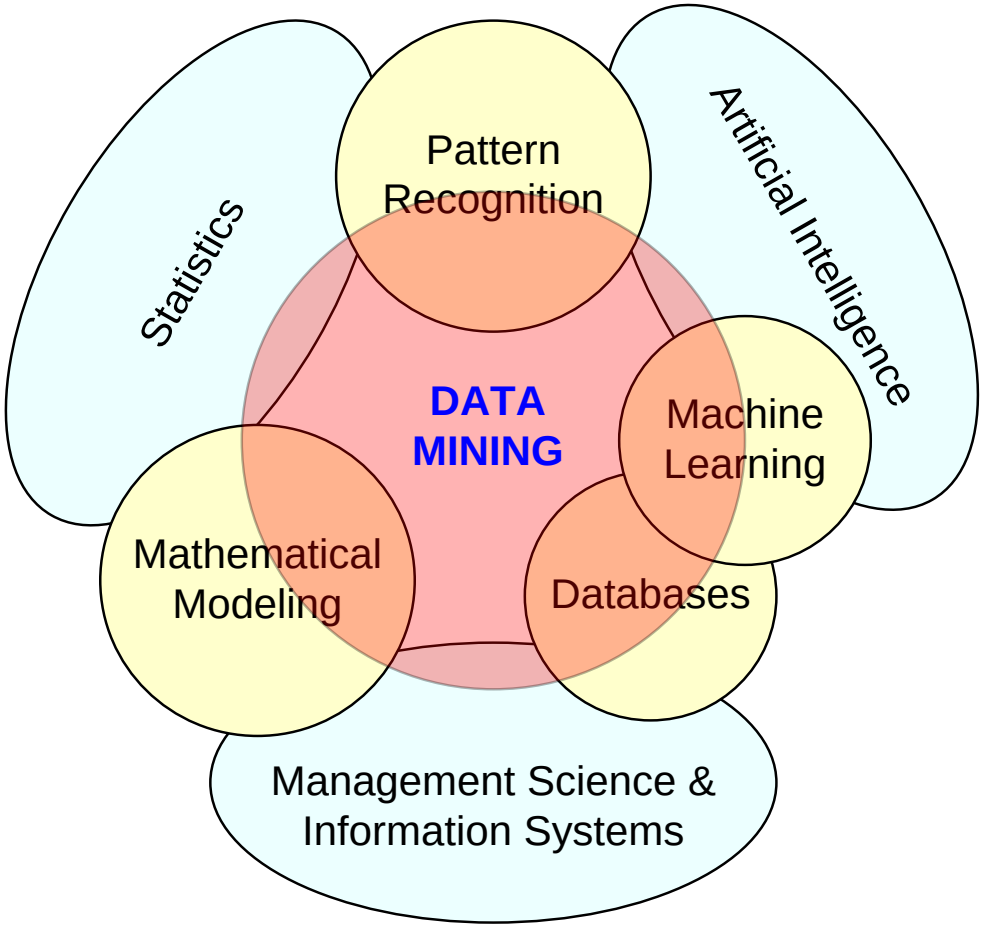
- Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly…in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, etc.)

NOT Data Mining:

- Look up phone number in phone director
- Query a Web search engine for information about "Amazon"

Data Mining at the Intersection of Many Disciplines

A process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases



Data mining commercial view

- Lots of data is being collected:
  - E-commerce, web data
  - Any time you go to a grocery store
  - Any time you use your bank or credit card
- Computers are cheaper and more powerful
- Competitive pressure is strong

Data Mining Applications

- Travel industry
- Healthcare
- Medicine
- Entertainment industry
- Homeland security, crime
- Sports

Data Mining Applications

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading

- Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

Data Mining Applications

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross- or up-selling)
  - Identify and treat most valued customers

- Banking & Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Maximize customer value (cross- and up-selling)
  - Optimizing cash reserves with forecasting

Data Mining Applications

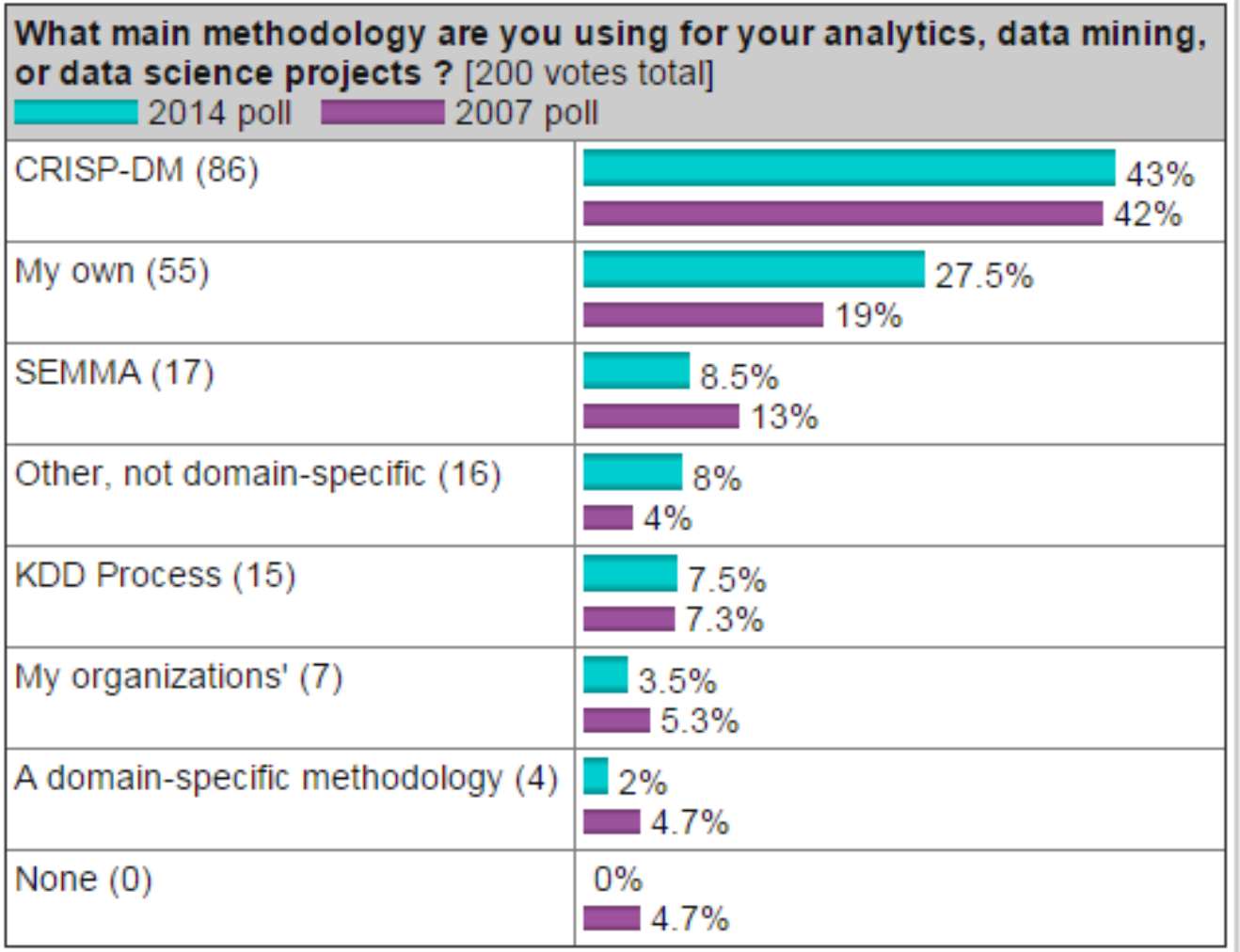
- Retailing and Logistics
  - Optimize inventory levels at different locations
  - Improve the store layout and sales promotions
  - Optimize logistics by predicting seasonal effects
  - Minimize losses due to limited shelf life

- Manufacturing and Maintenance
  - Predict/prevent machinery failures
  - Identify anomalies in production systems to optimize manufacturing capacity
  - Discover novel patterns to improve product quality

Data Mining Process

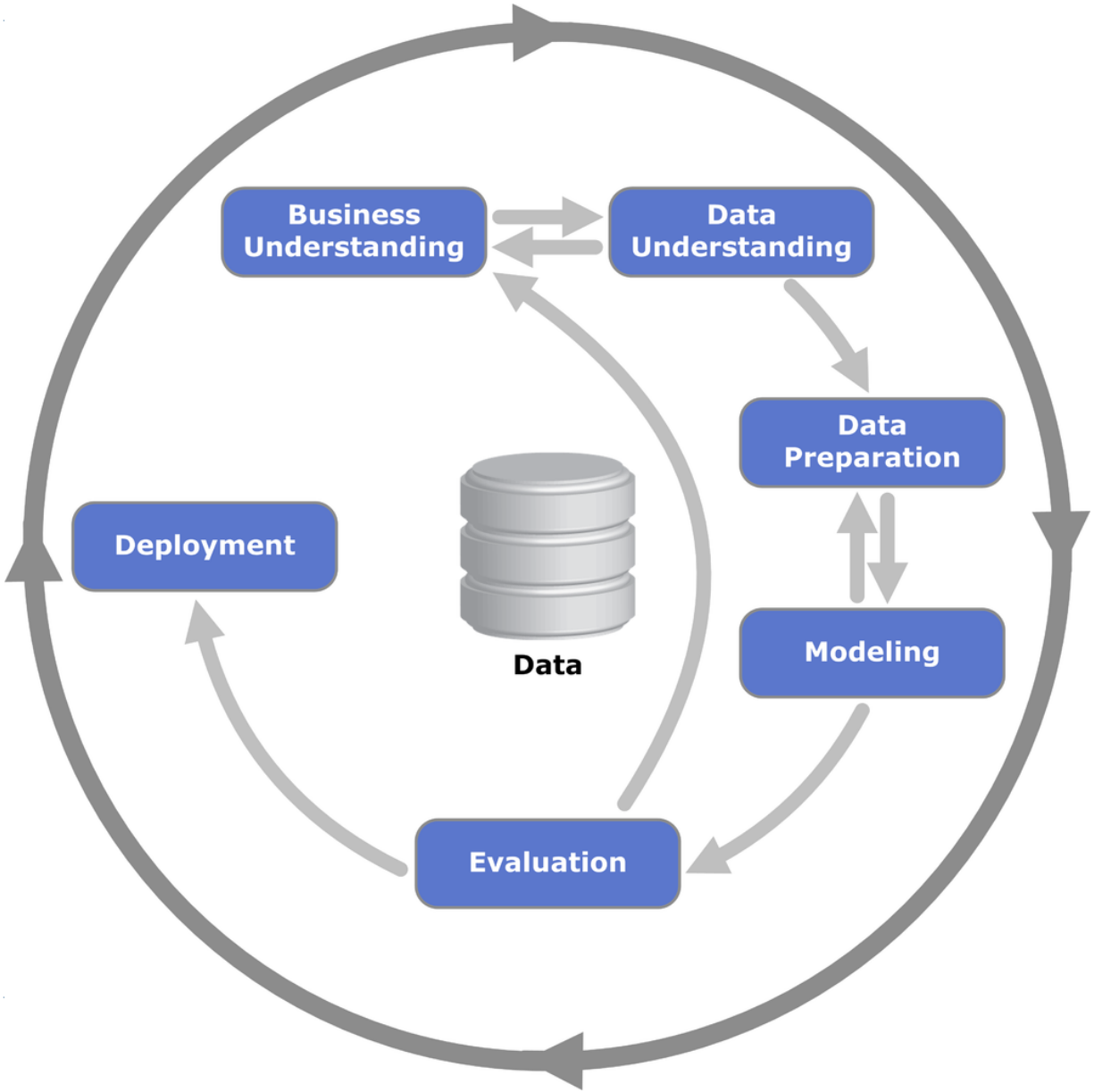
- A manifestation of best practices
- A systematic way to conduct DM projects
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Data Mining Process



Source: KDNuggets.com, Oct 2014

Data Mining Process: CRISP-DM



Data Mining Process: CRISP-DM

Step 1: Business Understanding

Step 2: Data Understanding

Step 3: Data Preparation (!)

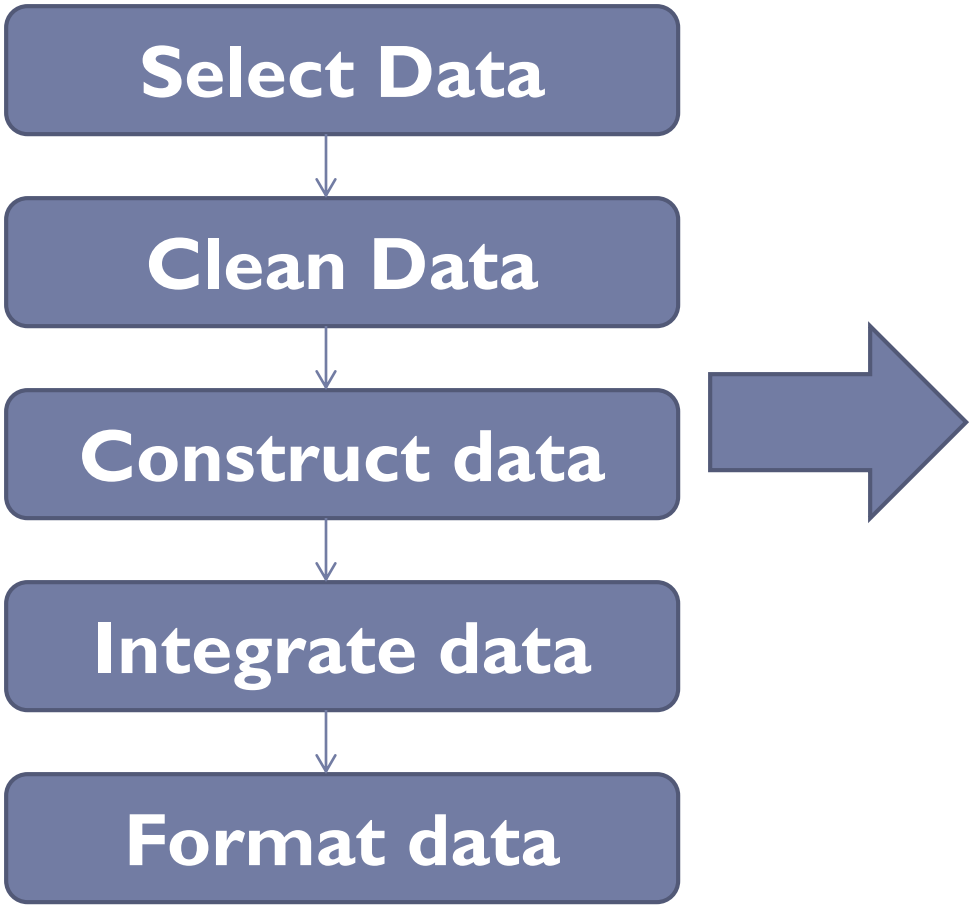
Step 4: Model Building

Step 5: Testing and Evaluation

Step 6: Deployment

- The process is highly repetitive and experimental

Data Preparation – A Critical DM Task



Output:

Dataset and a dataset description, both used for modeling or for the major analysis work of the project.

Data Preparation – A critical task

Select Data

- Activities:
  - Source data collected.
  - Significance and correlation tests to determine inclusion
  - Data selection criteria revisited (step2)
  - Sampling techniques
  - Document rationale for inclusion/exclusion
- Output:
  - List of data to be used/excluded and reasons

Data Preparation – A critical task

Clean Data

- Activities:
  - Noise
  - Special values.
  - Data selection criteria
- Output:
  - Data cleaning report

Data Preparation – A critical task

Construct Data

- Activities:
  - Construction mechanisms and tools
  - Production of derived attributes, new records, transformed values
  - Selection criteria
- Output:
  - Derived Attributes

Data Preparation – A critical task

Integrate Data

- Activities:
  - Integrate sources
  - Data selection criteria
- Output:
  - Merged data

Data Preparation – A critical task

Format Data

- Activities:
  - Rearranging attributes
  - Reordering records
  - Reformatting values
- Output:
  - Reformatted data

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is the column and is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is the row and is also known as record, point, case, sample, entity, observation or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data in Data Mining