WQD7005 Data Mining

Course Information

Instructor Information

- Name: Loh Yuen Peng
- Background: PhD (2019) Computer Science
- Research Interest:
  Computer Vision, Image Processing,
  Deep Learning, Machine Learning
- E-mail: lohyuenpeng@365.um.edu.my
- MS Teams Code: 22ol9lm

Course Information

- Coordinator: Prof. Dr. Teh Ying Wah
- Session: 2022/2023
- Semester: 2
- Group: 2
- Location: MM2
- Mode:
  - Physical (F2F)
  - Online (F2F) for selected sessions (TBA)
  - NF2F (TBA)

Software / Tools

- Talend Data Preparation (Open Source)

  https://sourceforge.net/projects/talend-data-preparation/

- SAS Enterprise Miner
  - Lab provided
  - SAS OnDemand for Academics

  https://www.sas.com/en_my/software/on-demand-for-academics.html

Course Information

- Text/Reference Book:

  Jiawei Han, Micheline Kamber, Hanghang Tong. Data Mining Concepts and Techniques, 4th Edition. Morgan Kaufmann Publishers, 2022

- Assessments:
  - Continuous (50%):
    - » Midterm Test (10%)
    - » Group Assignment & Presentation (20%)
    - » Group Project & Presentation (20%)
  - Summative (50%):
    - » Alternative Assessment 1 (25%)
    - » Alternative Assessment 2 (25%)

Assessment Information

- Assignment:
  - 4 members per group
    - Perform exploration and analysis on dataset
- Project:
  - 4 members per group
  - Perform exploration, analysis, and data mining on dataset
- Midterm & Alternative Assessments
  - Discussions and Case Studies

Weeks													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
						Mid-term	Asg.				AA1		Prj. AA2

Topics Information

- Introduction to Data Mining (2 weeks)
- Data Warehouse (2 weeks)
- Pre-mining (3 weeks)
- Classification (1 week)
- Association Rule Mining (2 weeks)
- Clustering (2 weeks)

Note: 2 weeks reserved for assessments.

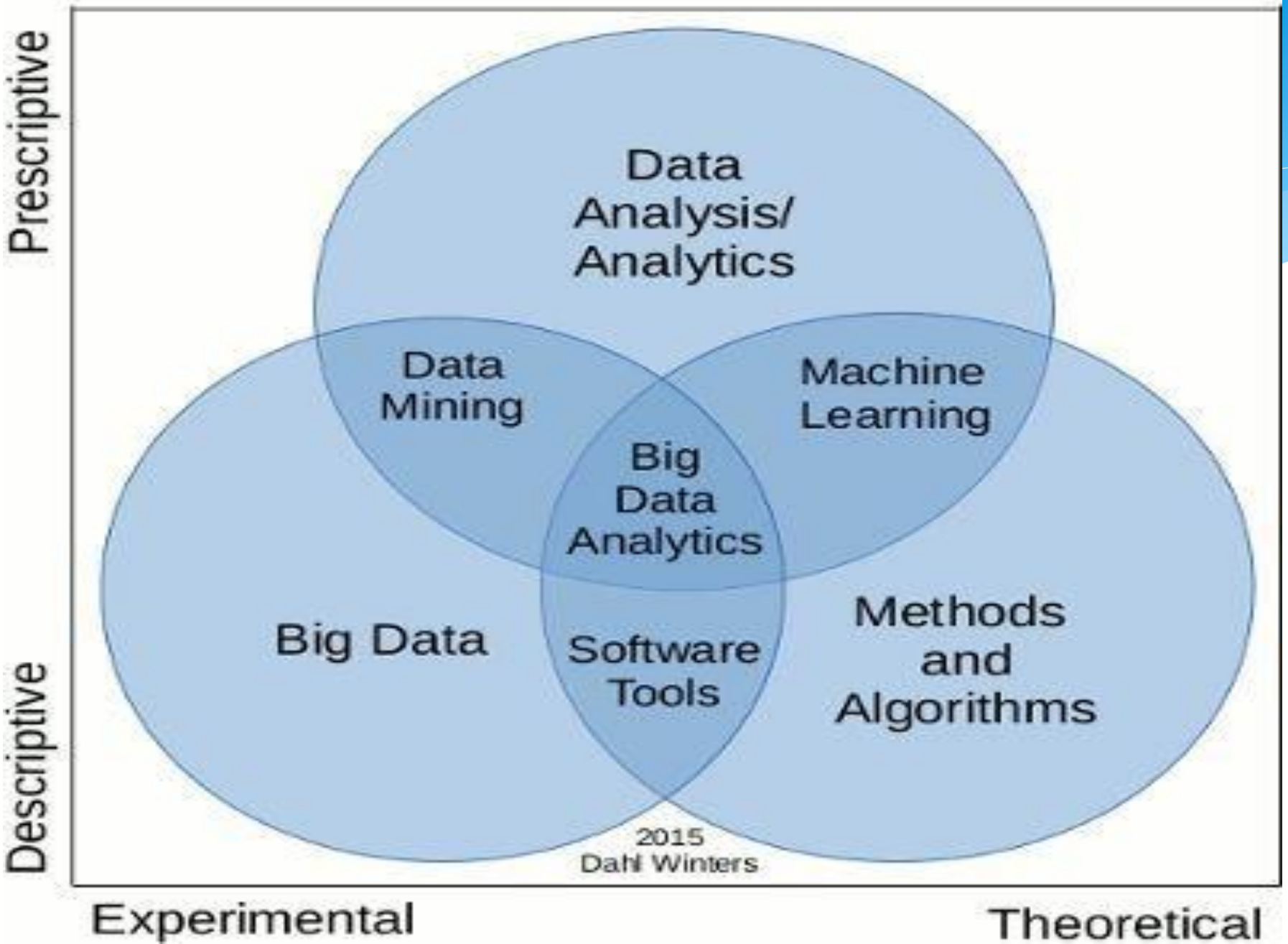
Introduction to Data Mining

Chapter 1

What is Data Mining?

- Statistics?
- Programming?
- Business?
- All about data?
- What about data science? What is the difference?

The Fields of Data Science



Data Mining Definition I

- The nontrivial extraction of **hidden**, previously **unidentified**, and potentially **valuable knowledge** from data
- A variety of techniques such as neural networks, decision trees or standard statistical techniques to **identify nuggets of information** or **decision-making knowledge** in bodies of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting, and estimation.

Data Mining Definition II

- Finding hidden information in a database

Consider these scenarios

- A restaurant owner looking to improve business
- A farmer intend to find out farm output
- An insurance company aiming to improve efficiency

Consider these scenarios

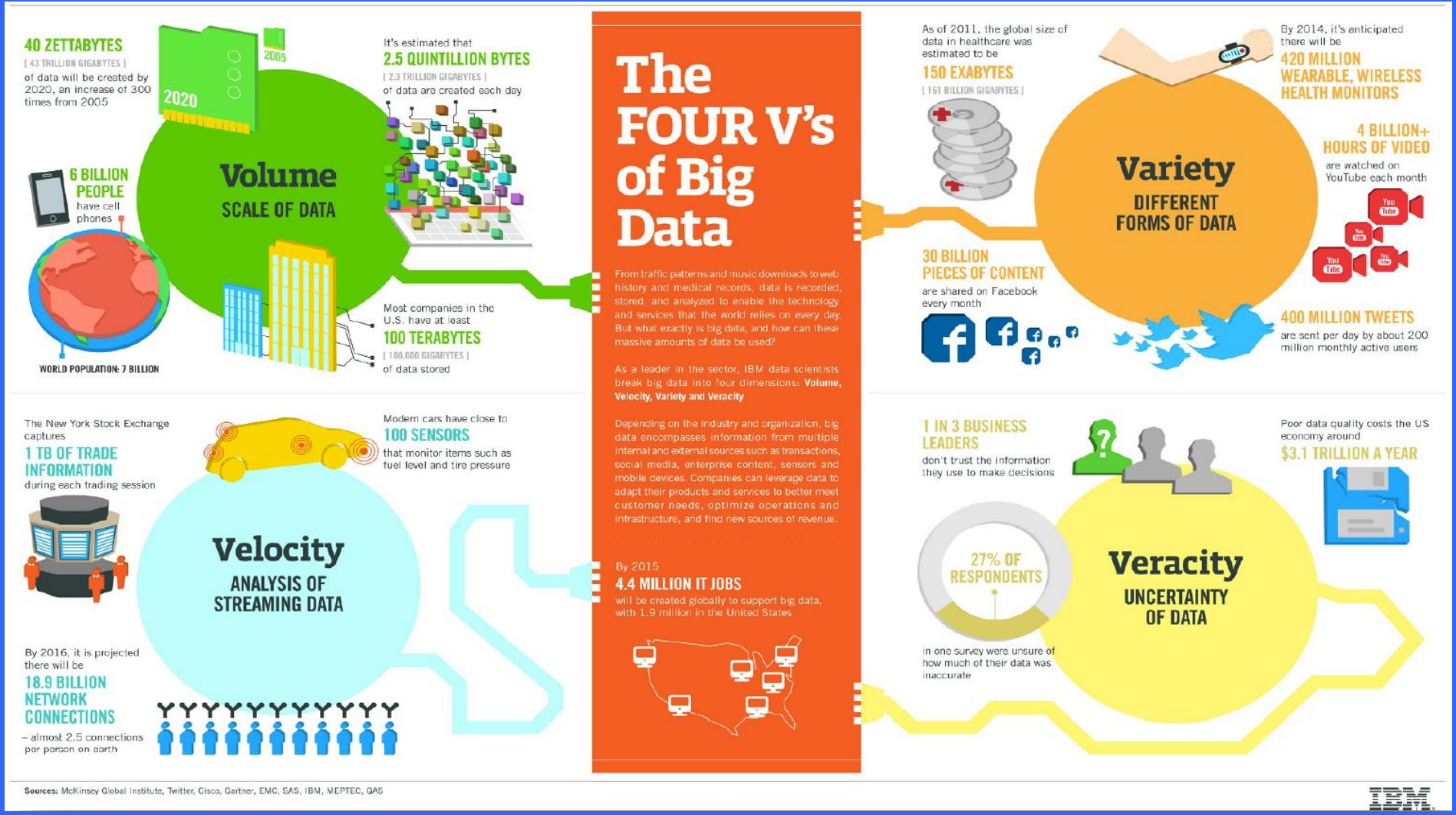
What data would be useful?

- A restaurant owner wants to identify common set of preferences among customers.
- A farmer wants to predict yield of rice in the next harvest.
- An insurance company wants to speed up claims approval process.

Hidden Information

- Number of years of experiences
- Great secret recipes
- Success Factors

Motivation


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Database Processing vs. Data Mining Processing

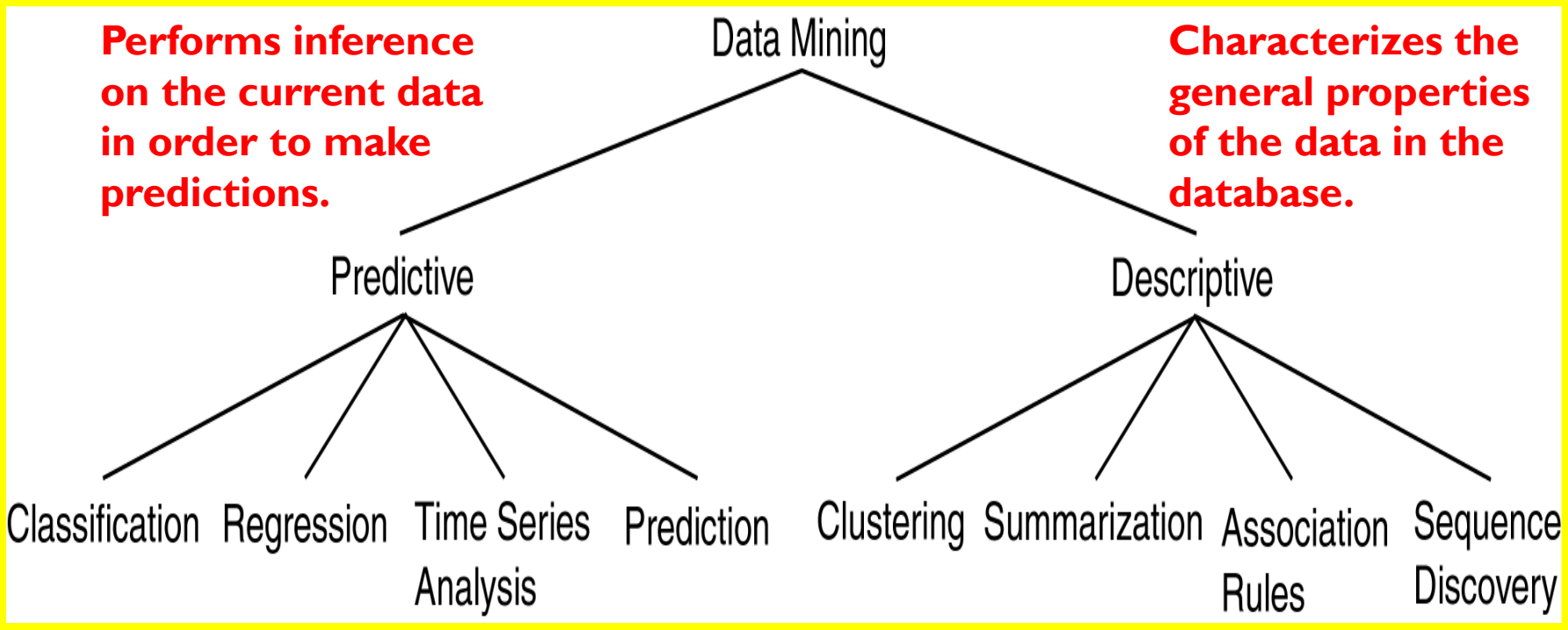
- Query
  - Well defined
  - SQL
- Data
  - Operational data
- Output
  - Precise
  - Subset of database

- Query
  - Poorly defined
  - No precise query language
- Data
  - Not operational data
- Output
  - Fuzzy
  - Not a subset of database

Query Examples

- Database
  - Find all credit applicants with surname name of Lee.
  - Identify customers who have purchased more than $100,000 in the last year.
  - Find all customers who have purchased bread
- Data Mining
  - Find all credit applicants who are good credit risks. (classification)
  - Identify customers with similar eating habits. (Clustering)
  - Find all items which are frequently purchased with bread. (association rules)

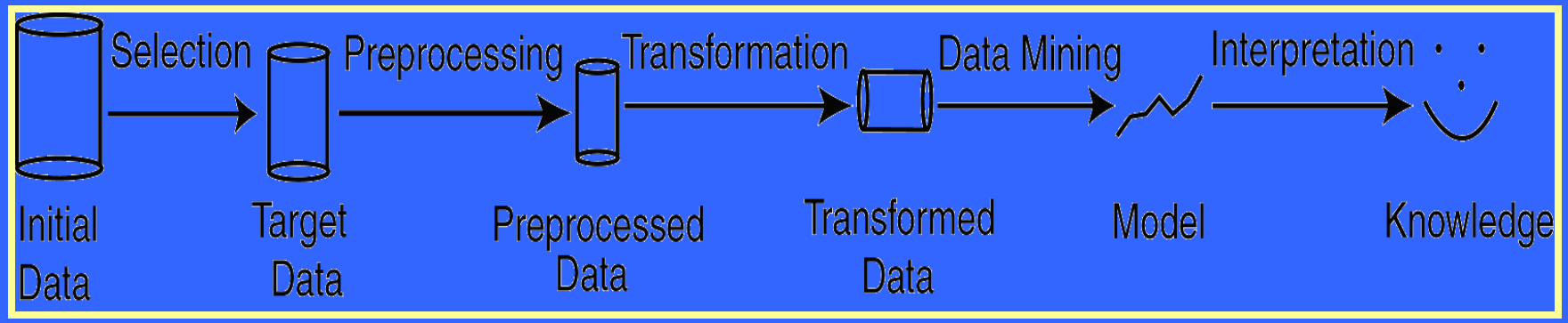
Data Mining Models and Tasks



Data Mining vs. KDD

- ***Knowledge Discovery in Databases (KDD):*** process of finding useful information and patterns in data.
- ***Data Mining:*** Use of algorithms to extract the information and patterns derived by the KDD process.

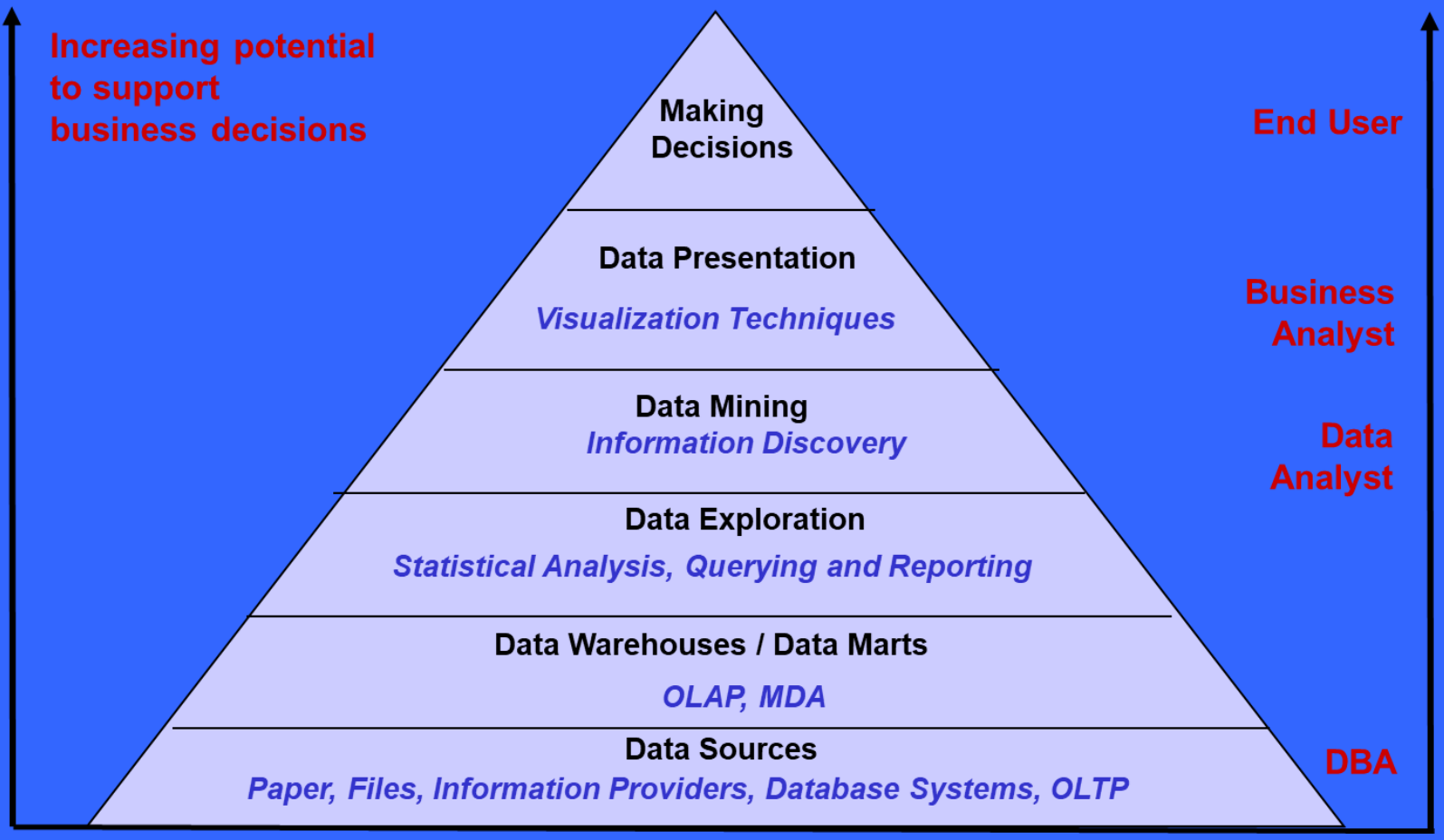
KDD Process



Modified from [FPSS96C]

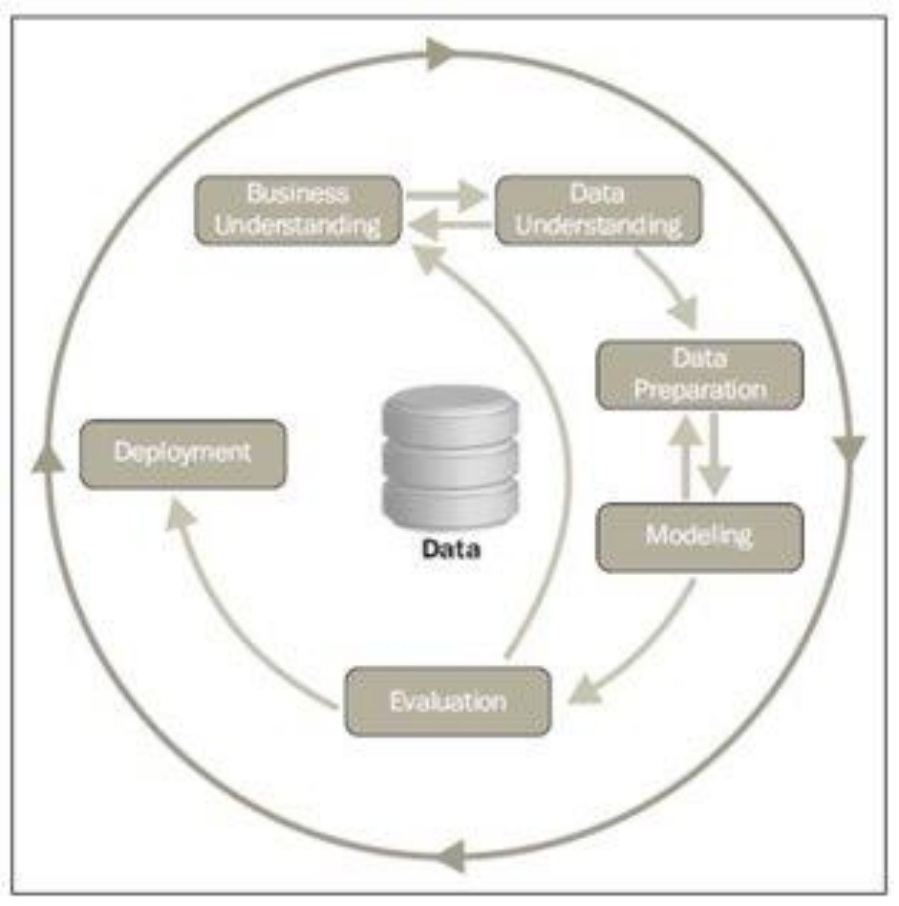
- ***Selection (Pre-Mining 1):*** Obtain data from various sources.
- ***Preprocessing (Pre-Mining 2) :*** Cleanse data.
- ***Transformation (Pre-Mining 3):*** Convert to common format. Transform to new format.
- ***Data Mining:*** Obtain desired results.
- ***Interpretation/Evaluation (Post-Mining):*** Present results to user in meaningful manner.

Data Mining: The Pyramid



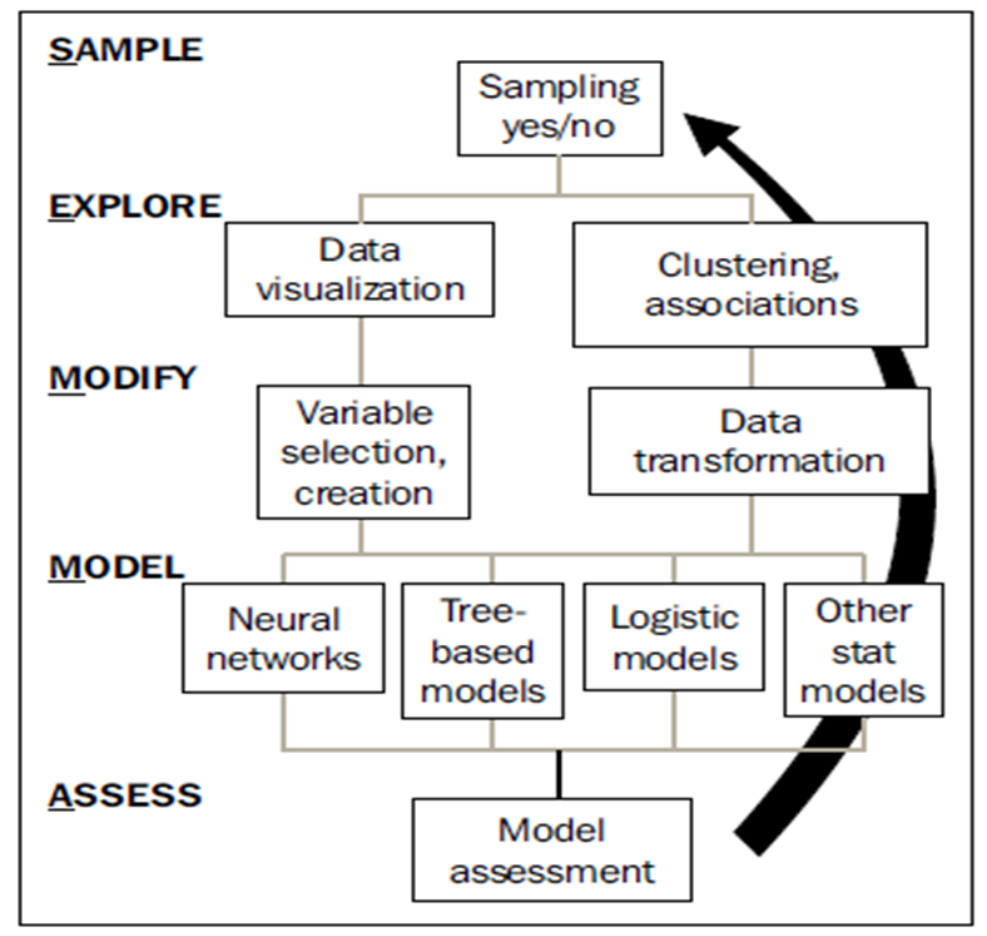
CRISP-DM

- Cross-Industry Standard Process for Data Mining
- Open standard process model



SEMMA

- Sample, Explore, Modify, Model, Assess
- By SAS



Data Mining Development

- Similarity Measures
- Hierarchical Clustering
- IR Systems
- Imprecise Queries
- Textual Data
- Web Search Engines

- Relational Data Model
- SQL
- Association Rule Algorithms
- Data Warehousing
- Scalability Techniques

- Bayes Theorem
- Regression Analysis
- EM Algorithm
- K-Means Clustering
- Time Series Analysis

- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Neural Networks
- Decision Tree Algorithms

What kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - The WWW

Purpose and Ability

- Descriptive and Prescriptive
- Association and cluster analysis
- Classification and regression
- Models and algorithms

Interestingness

- A pattern is interesting if:
  - It is easily understood by humans.
  - It is valid on new or test data with some degree of certainty.
  - It is potentially useful.
  - It is novel.
  - It validates a hypothesis that the user sought to confirm.
- An interesting pattern represents **knowledge**.

Major Issues

- Data Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem

Major Issues

- Performance and scalability
  - Efficiency and scalability of data mining algorithms
    - Parallel, distributed and incremental mining methods
- Issues relating to the diversity of data types
  - Handling relational and complex types of data.
  - Mining information from heterogeneous databases and global information systems (WWW).
- Issues related to applications and social impacts
  - Application of discovered knowledge
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem.
  - Protection of data security, integrity, and privacy.

Form Groups

WQD7005 Groups (230318)