بسمه تعالی

Data mining: Standard Methodology

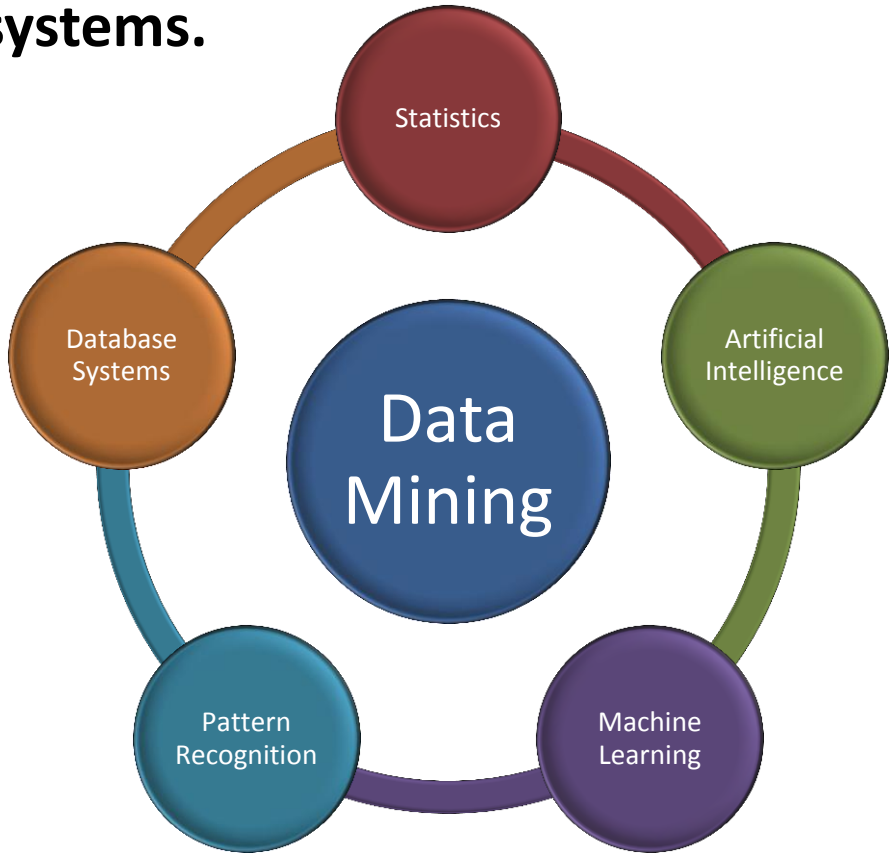
Dr O. Pournik ***MD, MPH, MSc, PhD*** **pournik@gmail.com**

Origins of Data Mining

Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.

Traditional Techniques may be unsuitable due to::

- Enormity of data
- High dimensionality of data
- Heterogeneous, distributed nature of data



Data Mining Methodology

CRISP-DM

CRoss-Industrv Standard Process for Data Mining

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
  - Reduces dependency on "stars"

Process Standardization

- **Initiative launched in late 1996 by three "veterans" of data mining market.**

  Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR

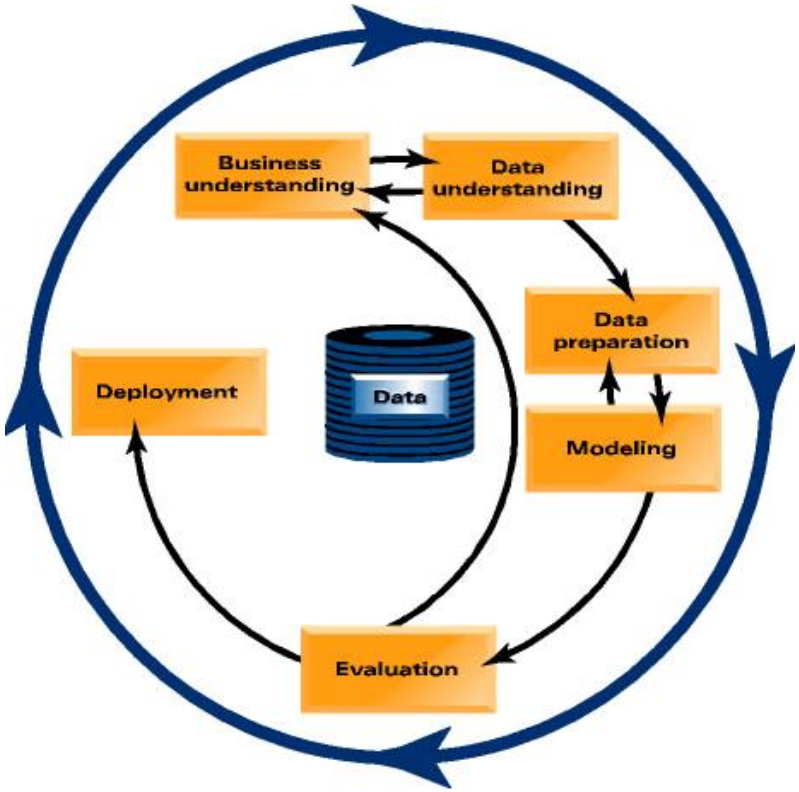
- **DevelInitiative launched in late 1996 by three "veterans" of data mining market.**

  Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR

- **Developed and refined through series of workshops** (from 1997-1999)
- **Over 300 organization contributed to the process model**
- **Published CRISP-DM 1.0** (1999)
- **Over 200 members of the CRISP-DM SIG worldwide**

  - **DM Vendors** - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.
  - **System Suppliers / consultants** - Cap Gemini, ICL Retail, Deloitte & Touche, etc.
  - **End Users** - BT, ABB, Lloyds Bank, AirTouch, Experian, etc.

CRISP-DM: Overview



- Data Mining methodology
- Process Model
- For anyone
- Provides a complete blueprint
- Life cycle: 6 phases

CRISP-DM: Phases

- Business Understanding

  Project objectives and requirements understanding

- Data Understanding

  Initial data collection and familiarization, Data quality problems identification

- Data Preparation

  Table, record and attribute selection, Data transformation and cleaning

- Modeling

  Modeling techniques selection and application, Parameters calibration

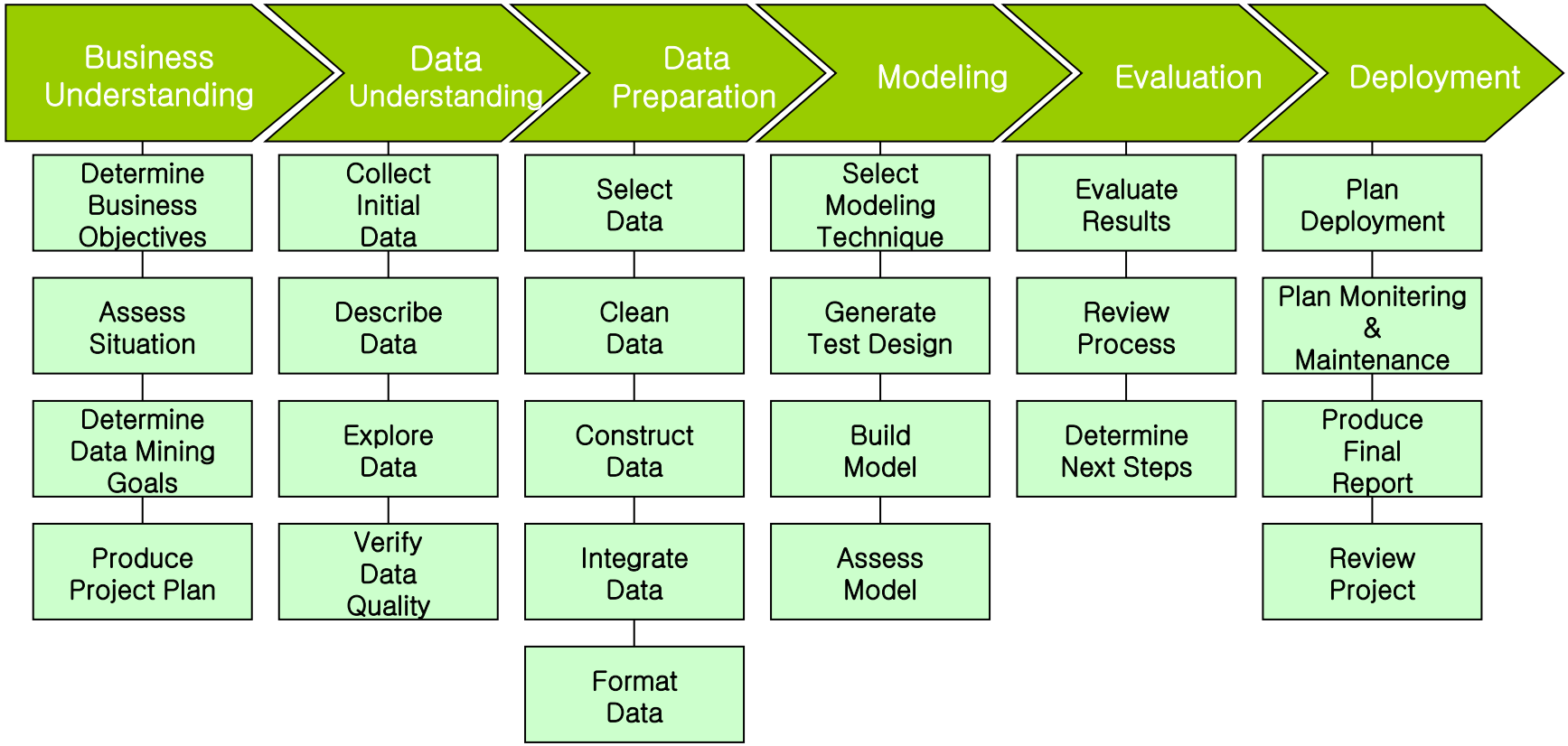
- Evaluation

  Business objectives & issues achievement evaluation

- Deployment

  Result model deployment, Repeatable data mining process implementation

Phases and Tasks



Data Mining Concept Methods

- ***Prediction Methods***
  - *Use some variables to predict unknown or future values of other variables.*

- ***Description Methods***
  - *Find human-interpretable patterns that describe the data.*

Data Mining Tasks

- ***Classification [Predictive]***
- ***Clustering [Descriptive]***
- ***Association Rule Discovery [Descriptive]***
- ***Sequential Pattern Discovery [Descriptive]***
- ***Regression [Predictive]***
- ***Deviation Detection [Predictive]***

Any Questions?