Najah Alshanableh

Agenda

* Important Definitions
* What Data Mining IS and IS NOT
* Steps in the Data Mining Process
* Examples
* Questions

Algorithms

Algorithm - Definition

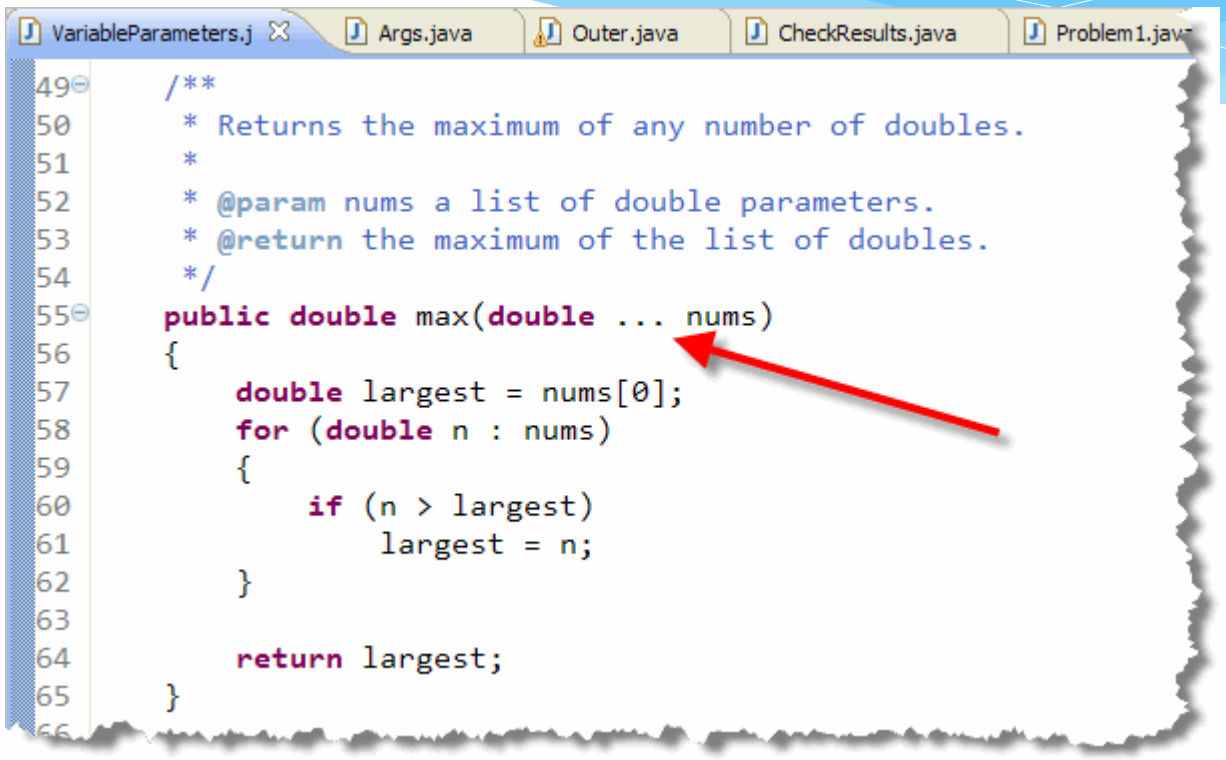
- An algorithm is a set of instructions for solving a problem.
- When the instructions are followed, it must eventually stop with an answer.

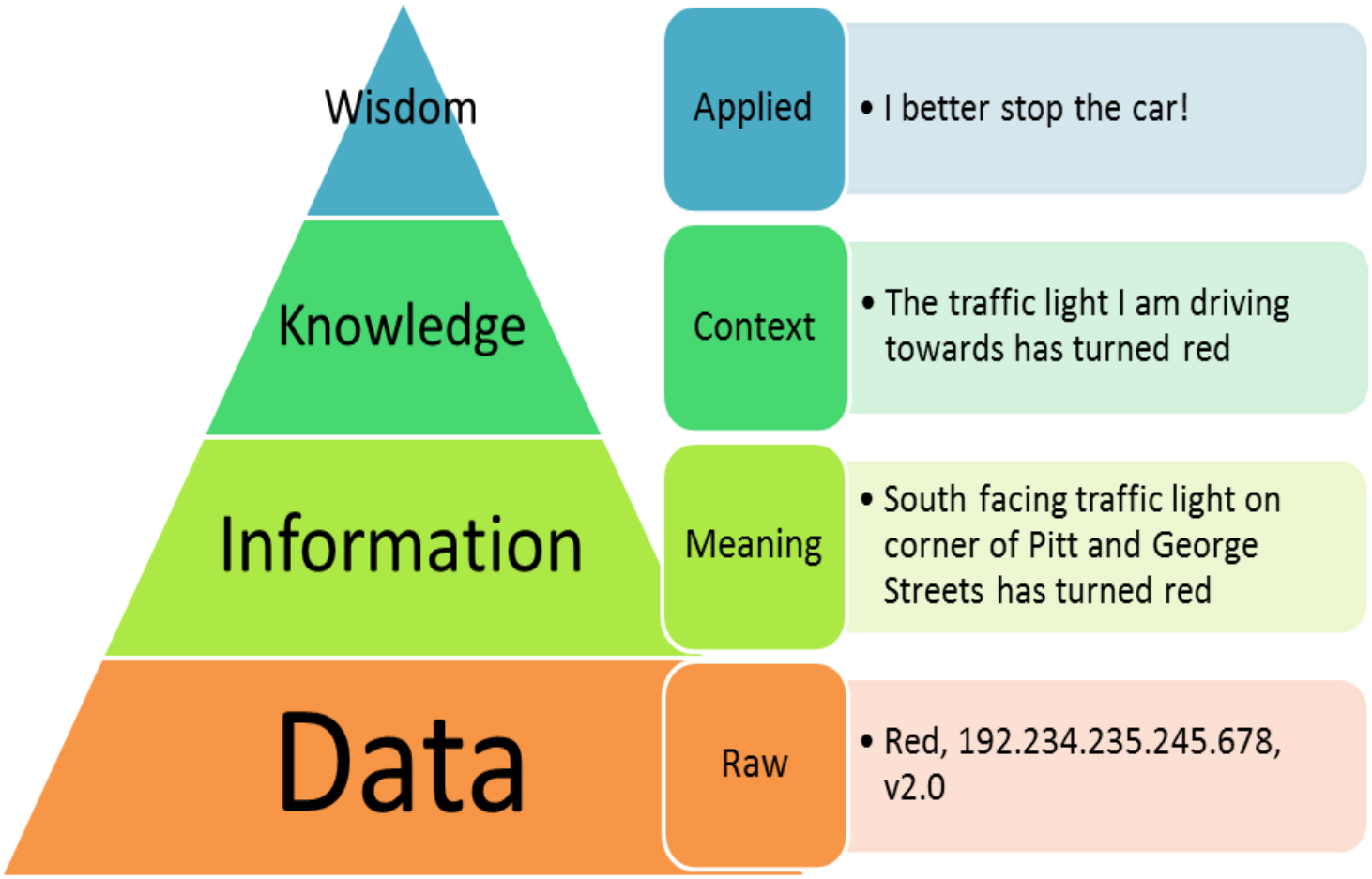
Example

Algorithm 1 Basic K-means Algorithm.

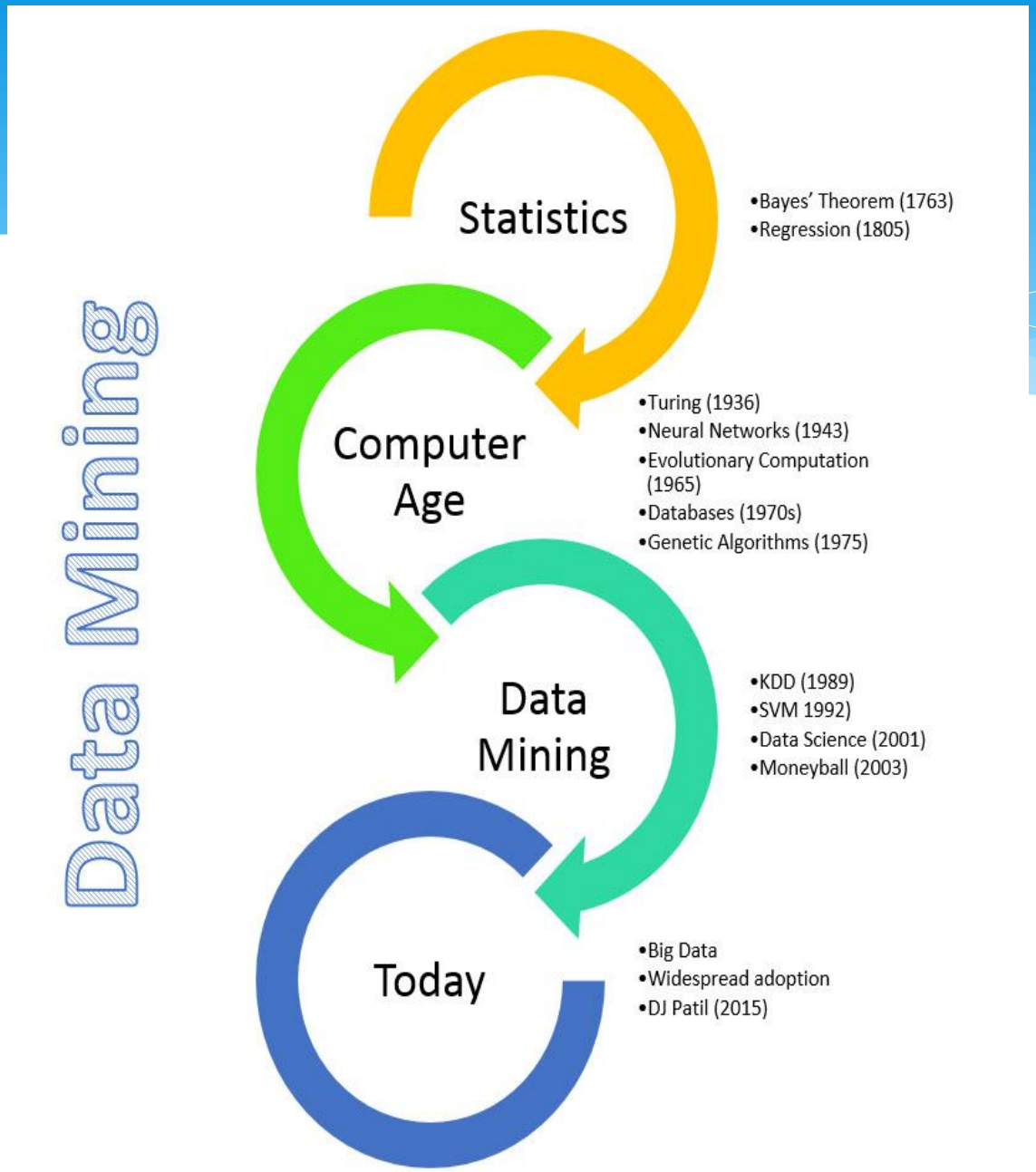
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Translate the algorithm to a working program






data
mining


Data Mining
Statistics
•Bayes' Theorem (1763)
•Regression (1805)
Computer Age
•Turing (1936)
•Neural Networks (1943)
•Evolutionary Computation (1965)
•Databases (1970s)
•Genetic Algorithms (1975)
Data Mining
•KDD (1989)
•SVM 1992)
•Data Science (2001)
•Moneyball (2003)
Today
•Big Data
•Widespread adoption
•DJ Patil (2015)

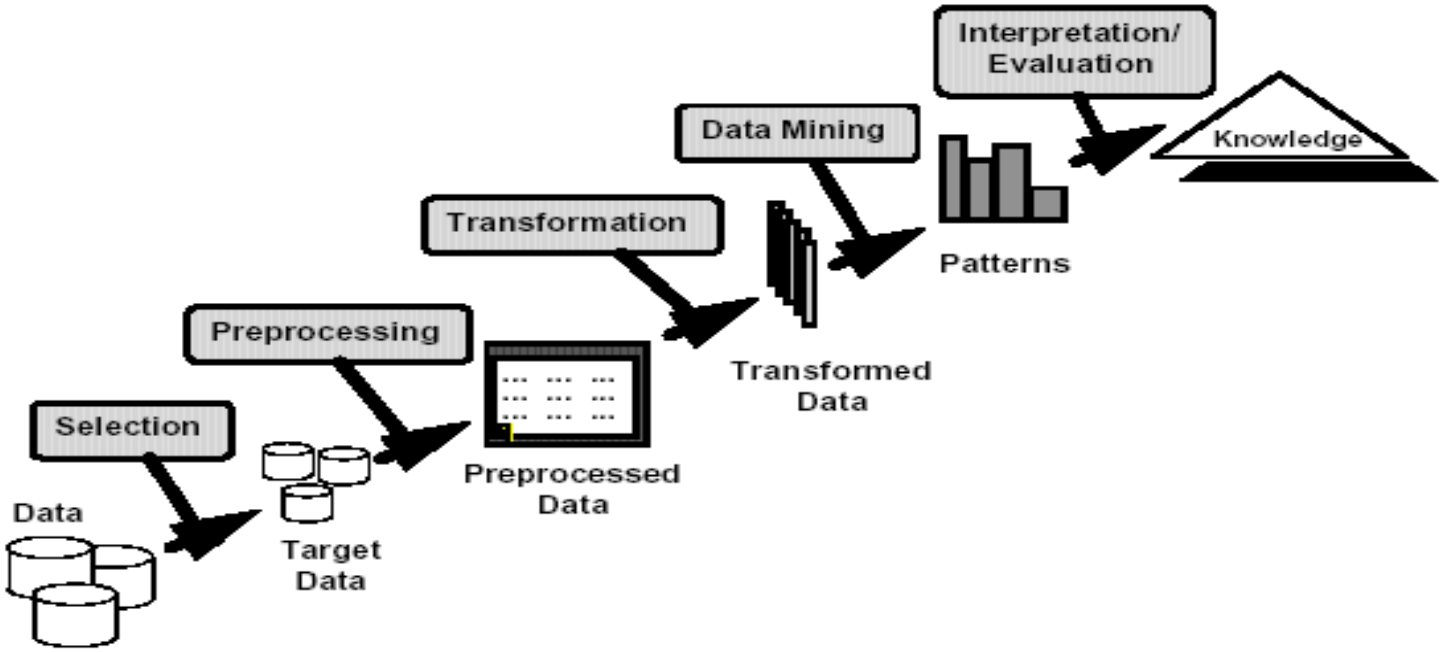
Data mining definition

Data mining is part of a group of concepts or techniques related to business intelligence, or e-business intelligence. Data mining involves **obtaining information** from a variety of sources that is stored in a data warehouse.

Data mining definition

What is Data Mining?

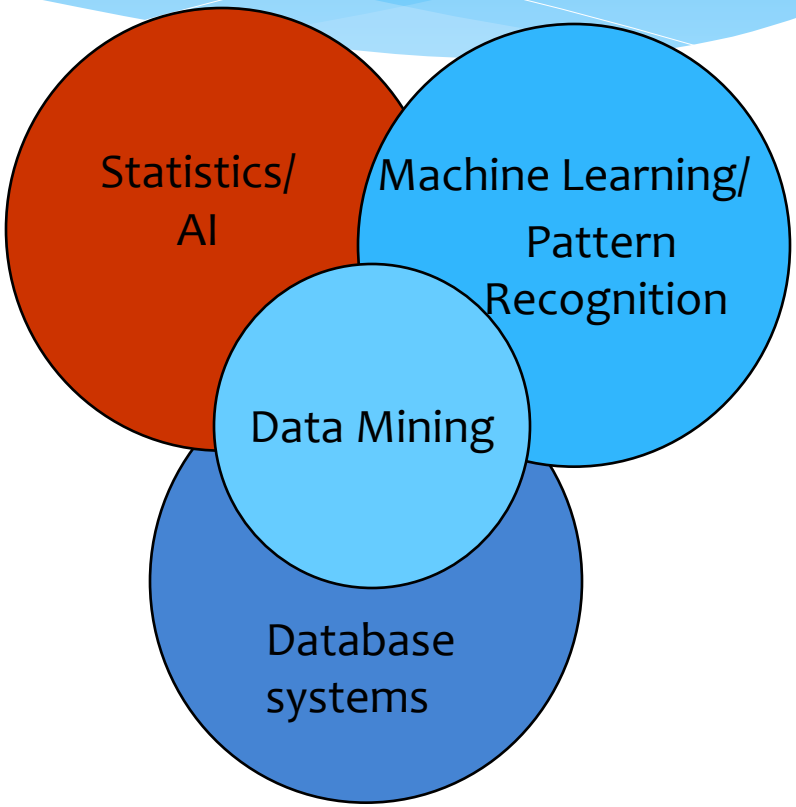
Data mining is the process of automatically discovering useful information in large data repositories.



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



Why Mine Data? Scientific Viewpoint

Traditional techniques infeasible for large data sets

Data mining may help scientists in classifying and segmenting data in hypothesis formation

What is wrong with conventional statistical methods ?

- Manual hypothesis testing:
  Not practical with large numbers of variables

- User-driven… User specifies variables, functional form and type of interaction:
  User intervention may influence resulting models

- Assumptions on linearity, probability distribution, etc.
  May not be valid

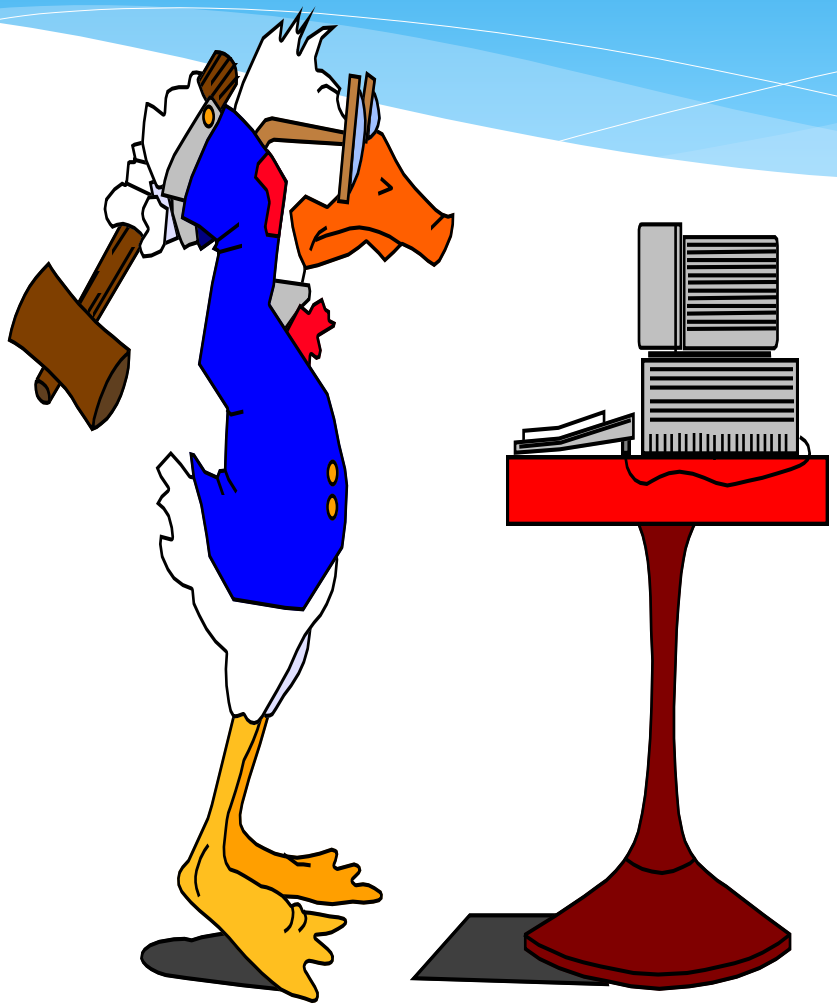
- Datasets collected with statistical analysis in mind
  Not always the case in practice

Statistics vs. Data Mining: Concepts

Feature	Statistics	Data Mining
Type of Problem	Well structured	Unstructured / Semi-structured
Inference Role	Explicit inference plays great role in any analysis	No explicit inference
Objective of the Analysis and Data Collection	First – objective formulation, and then - data collection	Data rarely collected for objective of the analysis/modeling
Size of data set	Data set is small and hopefully homogeneous	Data set is large and data set is heterogeneous
Paradigm/Approach	Theory-based (deductive)	Synergy of theory-based and heuristic-based approaches (inductive)
Signal-to-Noise Ratio	STNR > 3	0 < STNR <= 3
Type of Analysis	Confirmative	Explorative
Number of variables	Small	Large

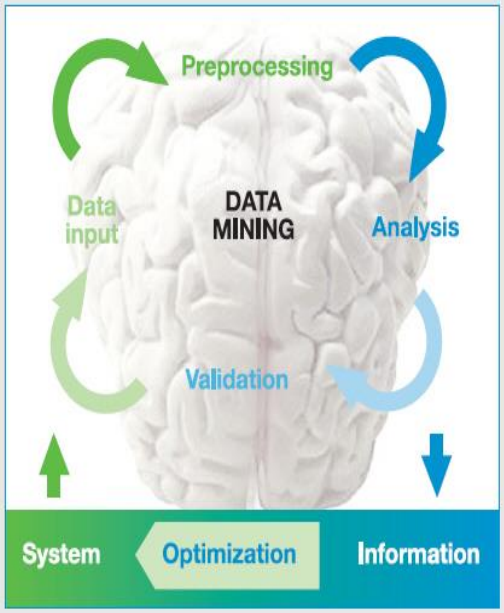
14

Data mining is not

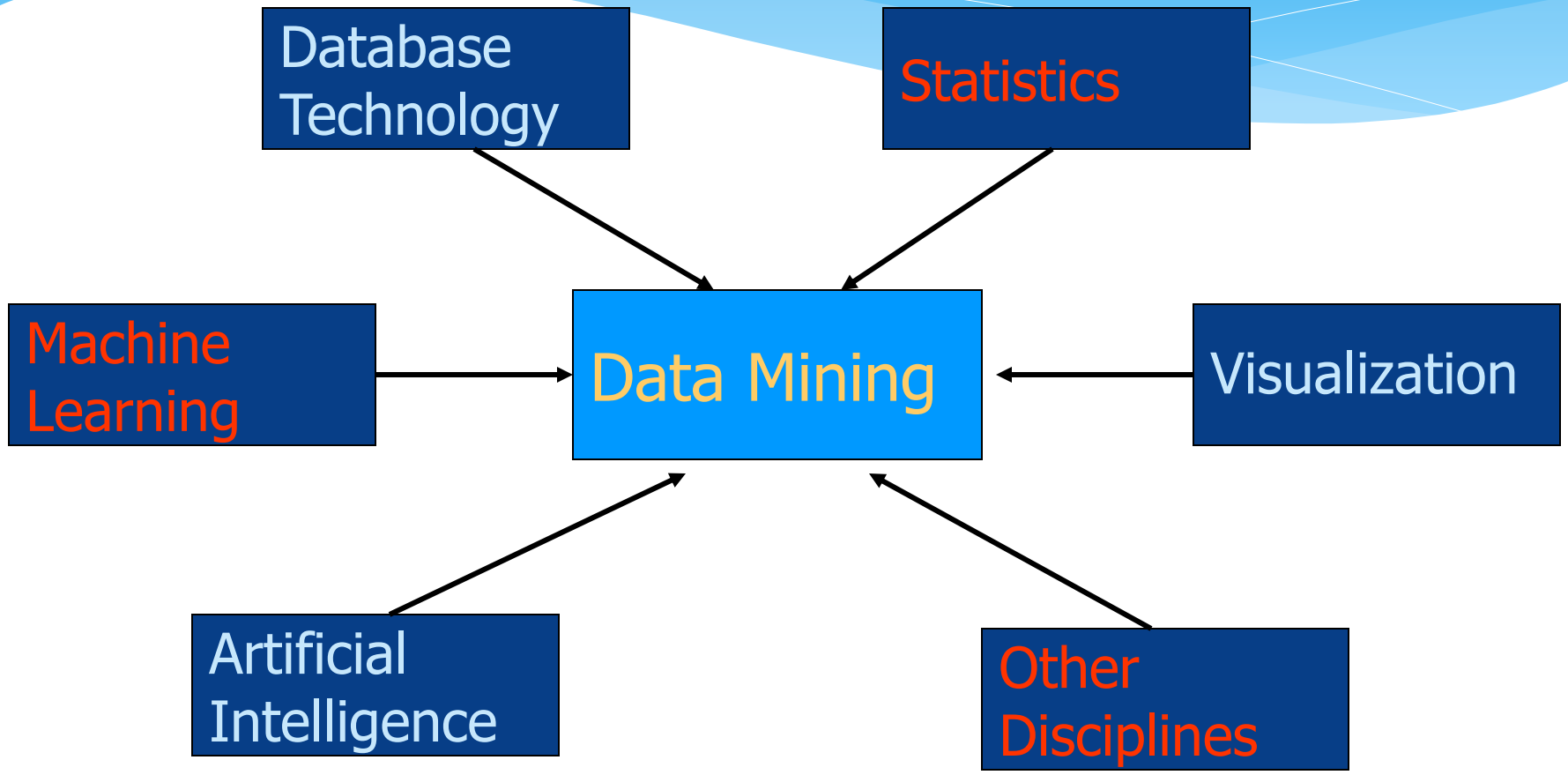


Data Mining is NOT

* Data Warehousing
* (Deductive) query processing
  * SQL/ Reporting
* Software Agents
* Expert Systems
* Online Analytical Processing (OLAP)
* Statistical Analysis Tool
* Data visualization



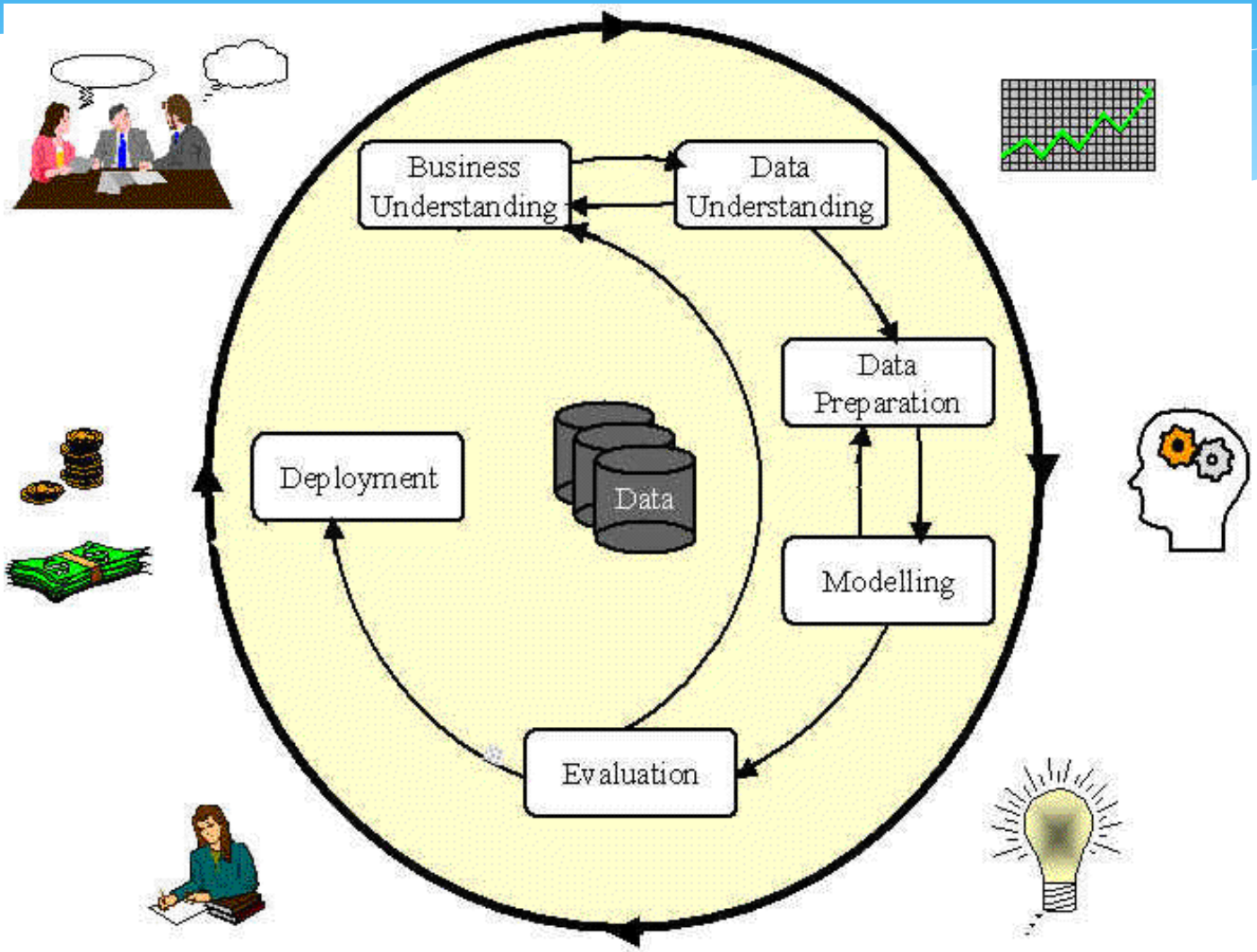
Multidisciplinary Field

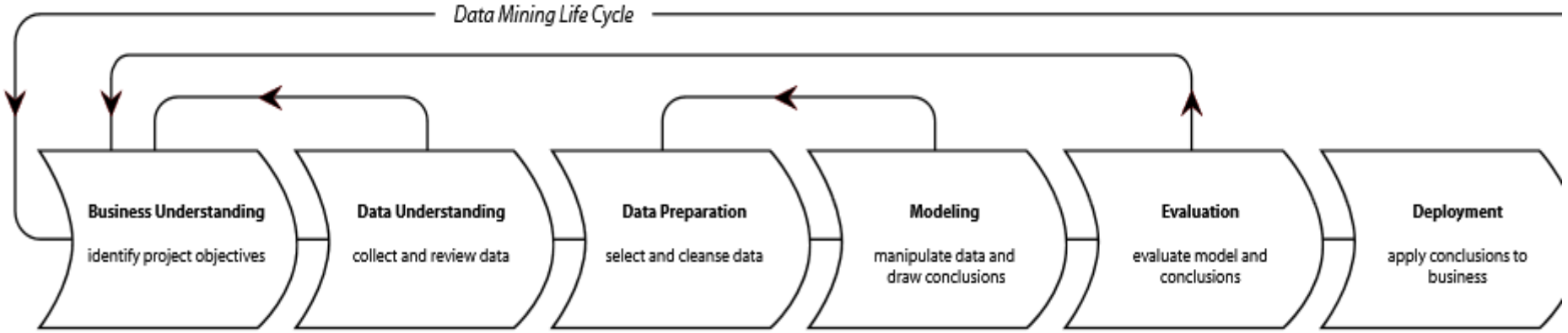


Results of Data Mining Include:

* Forecasting what may happen in the future
* Classifying people or things into groups by recognizing patterns
* Clustering people or things into groups based on their attributes
* Associating what events are likely to occur together
* Sequencing what events are likely to lead to later events

Phases in the DM Process: CRISP-DM





Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

a visual guide to CRISP-DM methodology

Generic Tasks
Specialized Tasks
(Process Instances)

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

Data Mining Applications

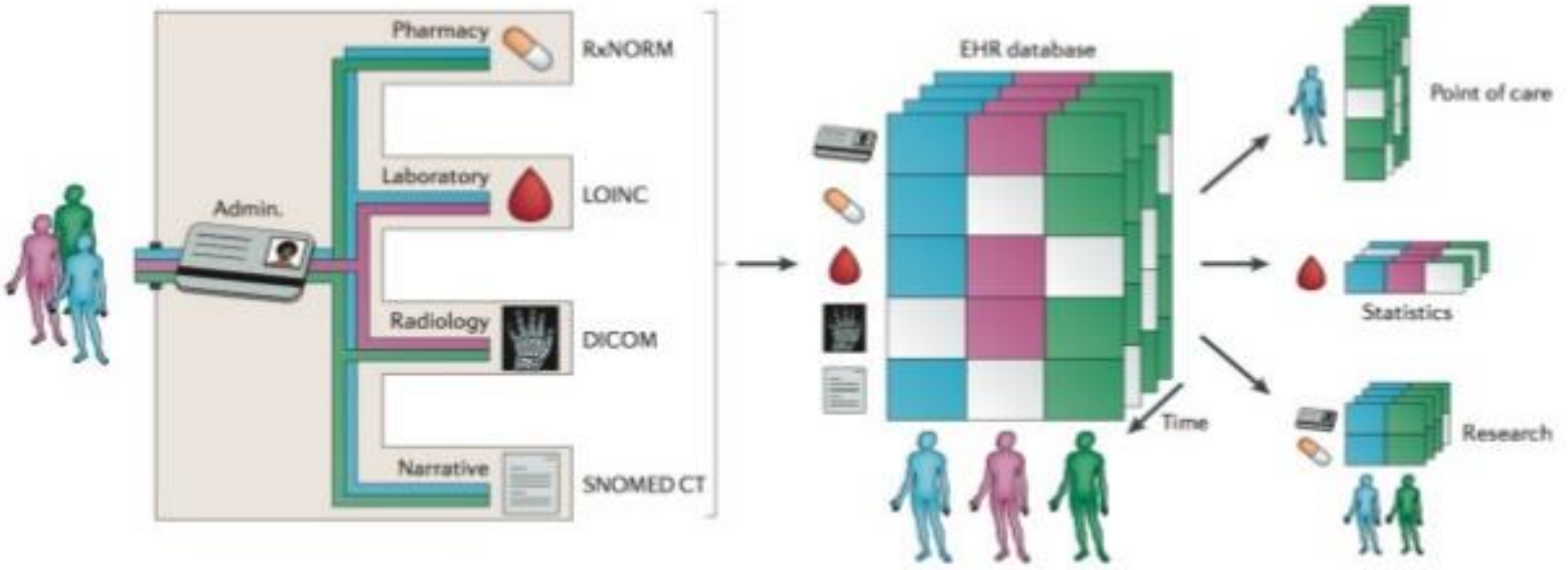
* **Pharmaceutical companies, Insurance and Health care, Medicine**
  * Drug development
  * Identify successful medical therapies
  * Claims analysis, fraudulent behavior
  * Medical diagnostic tools
  * Predict office visits

Examples


Scrn
Scroll Lock
Pause
rs112054285
example
Insert
Home
MediaFocus

Medical data mining

Linking diseases, drugs, and adverse reactions



Lars Juhl Jensen

Customers Who Bought This Item Also Bought

Pattern Recognition and Machine Learning (Infor... by Christopher M. Bishop

★★★★☆ (41) $67.96

The Elements of Statistical Learning by T. Hastie

★★★★☆ (27) $75.17

Introduction to Data Mining by Pang-Ning Tan

★★★★★ (10) $77.60

Questions ???

Thank You…