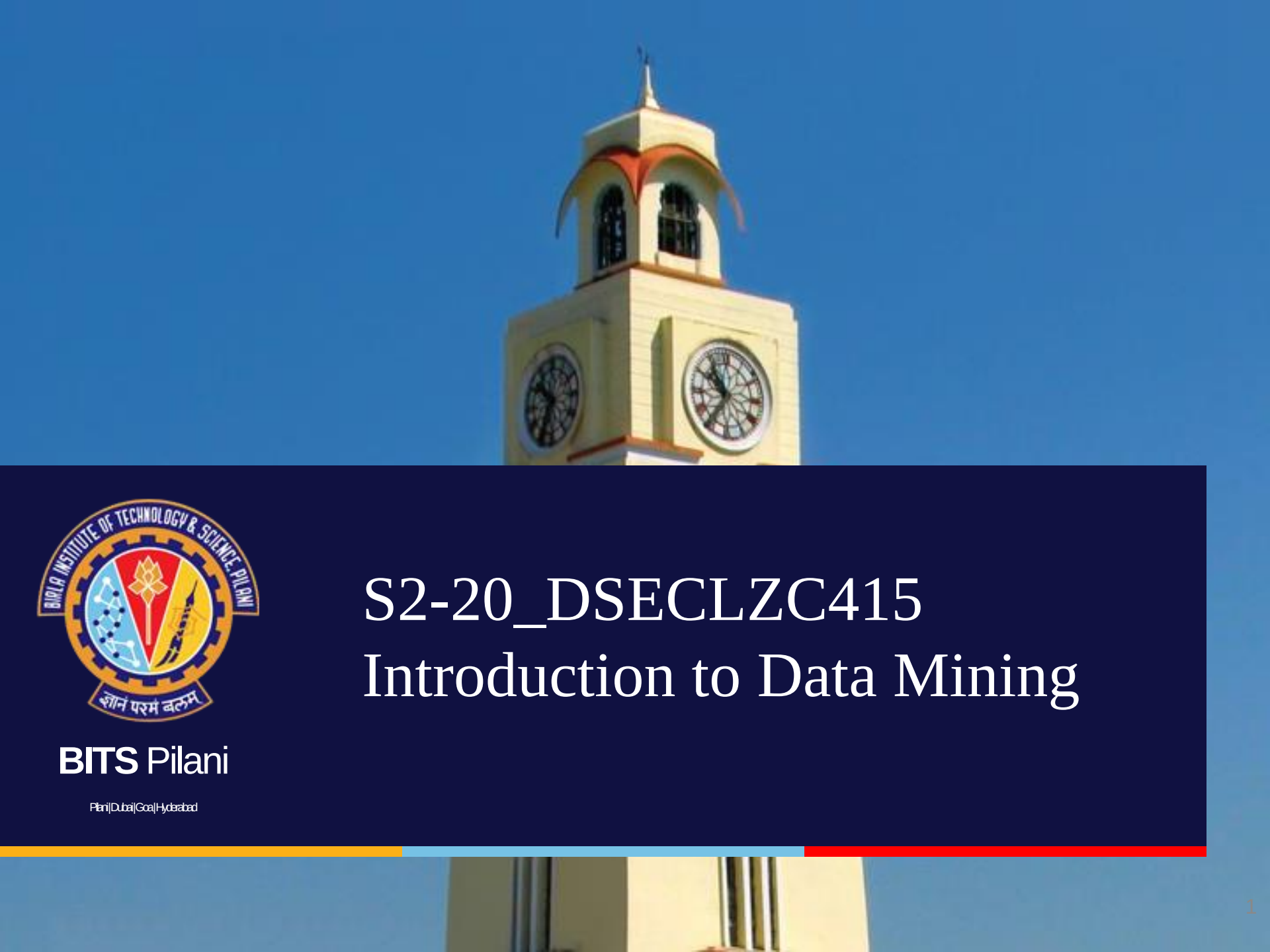


BITS Pilani

Pilani|Dubai|Goa|Hyderabad

# S2-20_DSECLZC415 Introduction to Data Mining


- *The slides presented here are obtained from the authors of the books and from various other contributors. I hereby acknowledge all the contributors for their material and inputs.*
- *I have added and modified a few slides to suit the requirements of the course.*

# Textbooks/Reference Books

## Text Books

| | |
|---|---|
| T1 | Tan P. N., Steinbach M & Kumar V. “Introduction to Data Mining” Pearson Education, 2006 |
| T2 | Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2011 |

## Reference Book(s) & other resources

| | |
|---|---|
| R1 | Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers © 2015 |
| | Additional references may be given during lectures |

# Modular Structure

| No | Title of the Module |
|---|---|
| M1 | Introduction to Data Mining |
| M2 | Data Preprocessing |
| M3 | Data Exploration |
| M4 | Classification and Prediction |
| M5 | Clustering |
| M6 | Association Analysis |
| M7 | Anomaly Detection |
| M8 | Data mining on unstructured (Big) data |
| M9 | Data Mining Applications |

# Evaluation Scheme

| No | Name | Type | Weight |
|---|---|---|---|
| 1. | Quiz-I | Online | 5% |
| | Quiz-II | Online | 5% |
| | Assignment | Group | 10% |
| 2. | Mid-Semester Test | | 30% |
| 3. | Comprehensive Exam | | 50% |

# Data Mining Defined

# 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

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

# What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- What is 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,)

# 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

# Why Data Mining

A search engine (e.g., Google) receives hundreds of millions of queries every day. Each query can be viewed as a transaction where the user describes her or his information need.

What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time? Some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone.

For example, Google's *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms. A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, *Flu Trends* can estimate flu activity up to two weeks faster than traditional systems can.

This example shows how data mining can turn a large collection of data into knowledge that can help meet a challenge.

# Evolution of Database Technology

1960s:

- Data collection, database creation, IMS and network DBMS

1970s:

- Relational data model, relational DBMS implementation

1980s:

- RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
- Application-oriented DBMS (spatial, scientific, engineering, etc.)

1990s:

- Data mining, data warehousing, multimedia databases, and Web databases

2000s

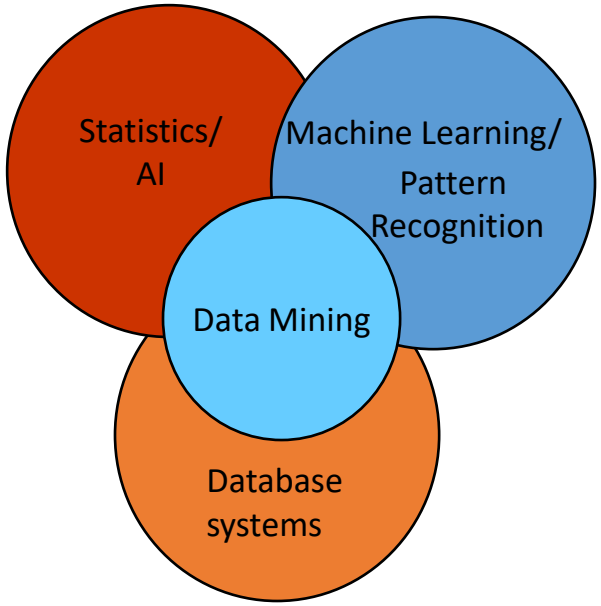
- Stream data management and mining
- Data mining and its applications
- Web technology (XML, data integration) and global information systems

# Origins of Data Mining

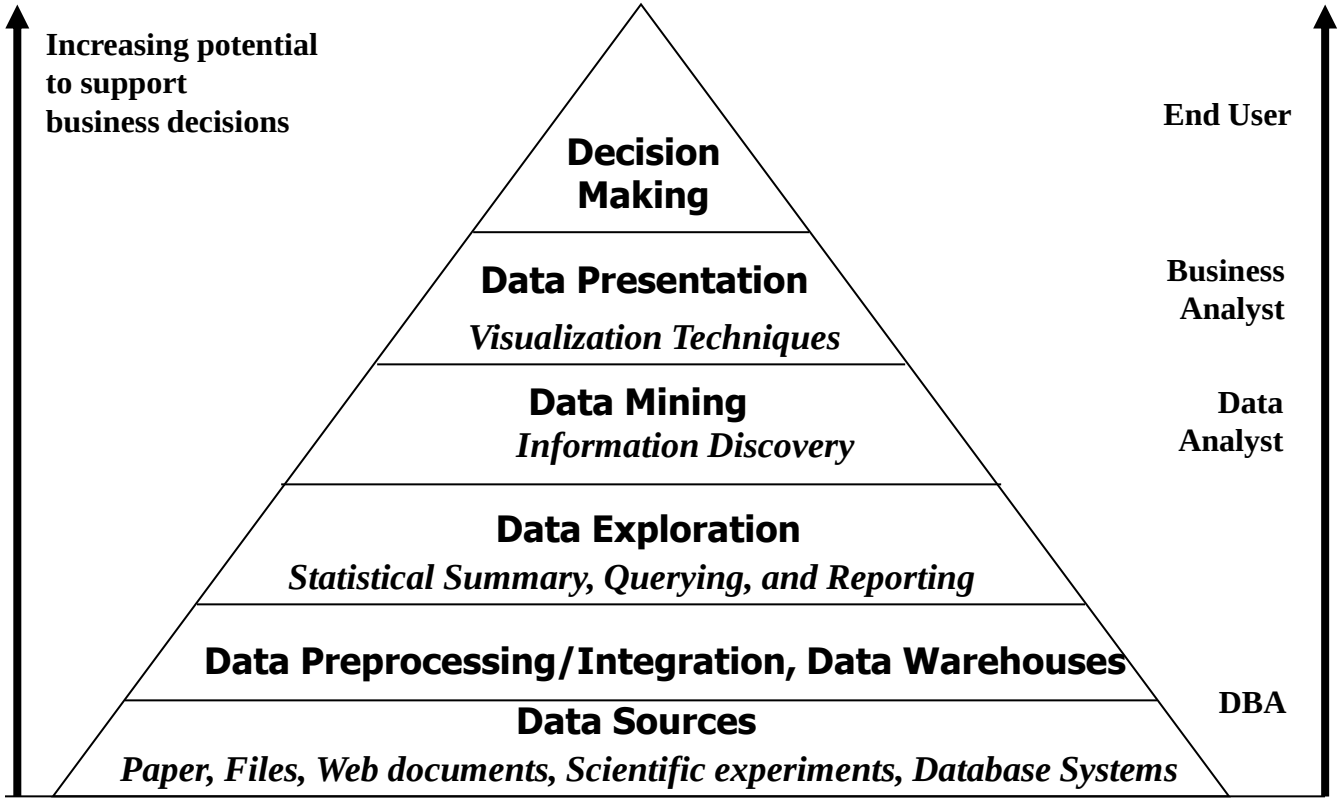
Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

Traditional Techniques may be unsuitable due to

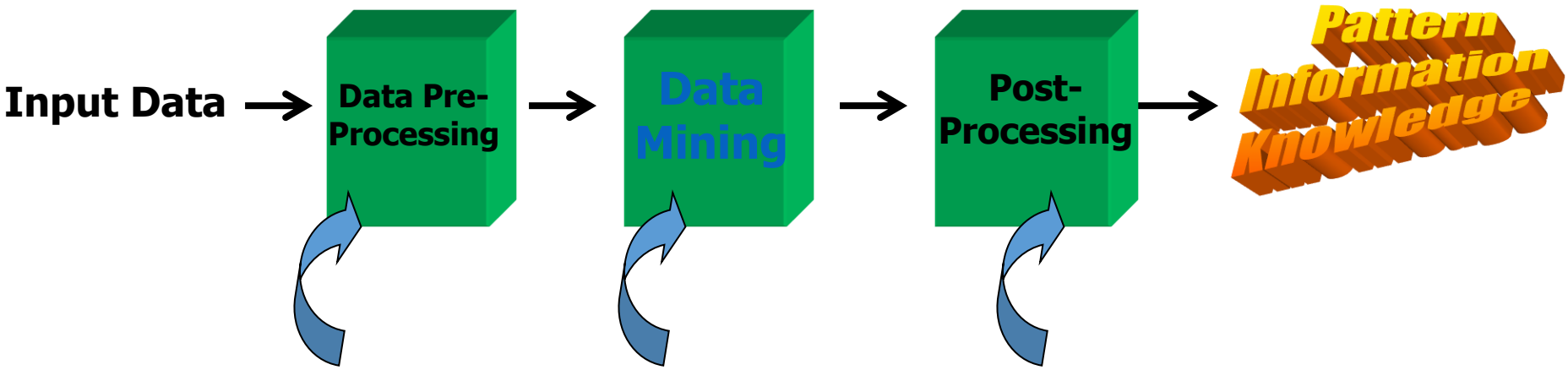
- Enormity of data
- High dimensionality of data
- Heterogeneous, distributed nature of data



# Data Mining in Business Intelligence



# Data Mining/KDD Process




Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

KDD – Knowledge Discovery in Databases

# Data Mining & Machine Learning

According to Tom M. Mitchell, Chair of Machine Learning at Carnegie Mellon University and author of the book *Machine Learning* (McGraw-Hill),

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E.
>
> We now have a set of objects to define machine learning:
>
> Task (T), Experience (E), and Performance (P)
>
> With a computer running a set of tasks, the experience should be leading to performance increases (to satisfy the definition)

Many data mining tasks are executed successfully with help of machine learning

# Multi-Dimensional View of Data Mining

**Data to be mined**

- Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

**Knowledge to be mined (or: Data mining functions)**

- Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
- Descriptive vs. predictive data mining
- Multiple/integrated functions and mining at multiple levels

**Techniques utilized**

- Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

**Applications adapted**

- Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

# Data Mining on Diverse kinds of Data

Besides relational database data (from operational or analytical systems), there are many other kinds of data that have diverse forms and structures and different semantic meanings.

Examples of data can be :

- time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data),
- data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
- spatial data (e.g., maps),
- engineering design data (e.g., the design of buildings, system components, or integrated circuits),
- hypertext and multimedia data (including text, image, video, and audio data),
- graph and networked data (e.g., social and information networks), and
- the Web (a widely distributed information repository).

Diversity of data brings in new challenges such as handling special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity)

# Data Mining Activities

# Data Mining Tasks

## Prediction Methods

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

## Description Methods

- Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

# Data Mining Tasks...

Classification [Predictive]

Clustering [Descriptive]

Association Rule Discovery [Descriptive]

Sequential Pattern Discovery [Descriptive]

Regression [Predictive]

Deviation Detection [Predictive]

# Classification: Definition

Given a collection of records (*training set* )

- Each record contains a set of *attributes*, one of the attributes is the *class*.

Find a *model* for class attribute as a function of the values of other attributes.

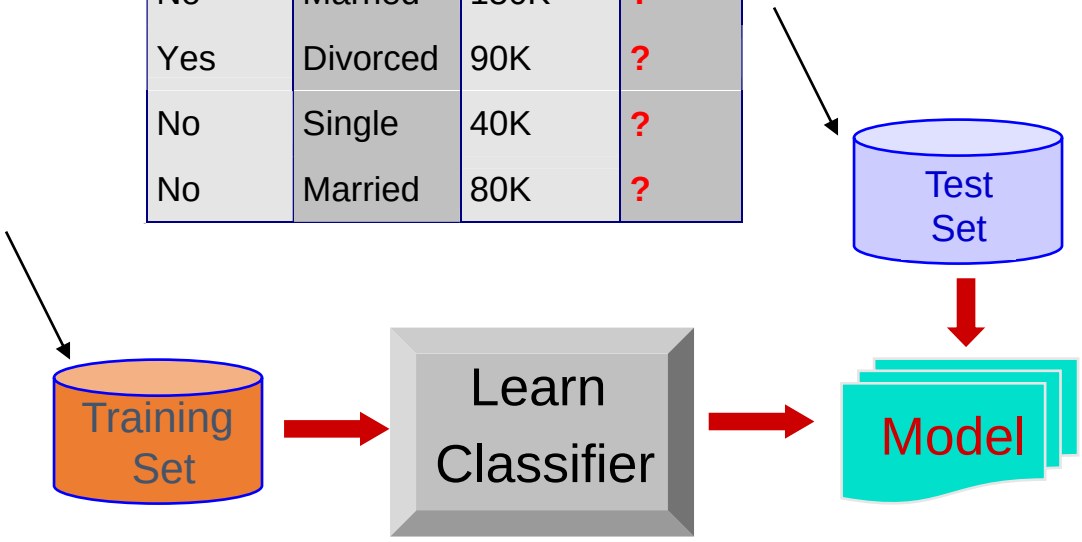
Goal: previously unseen records should be assigned a class as accurately as possible.

- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

# Classification Example

| Tid | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 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 |

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



# Classification: Application 1

Direct Marketing

- Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
- Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

# Classification: Application 2

Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

# Classification: Application 3

Customer Attrition/Churn:

- Goal: To predict whether a customer is likely to be lost to a competitor.
- Approach:
  - Use detailed record of transactions with each of the past and present customers, to find attributes.
    - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal.
  - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

# Clustering Definition

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Similarity Measures:

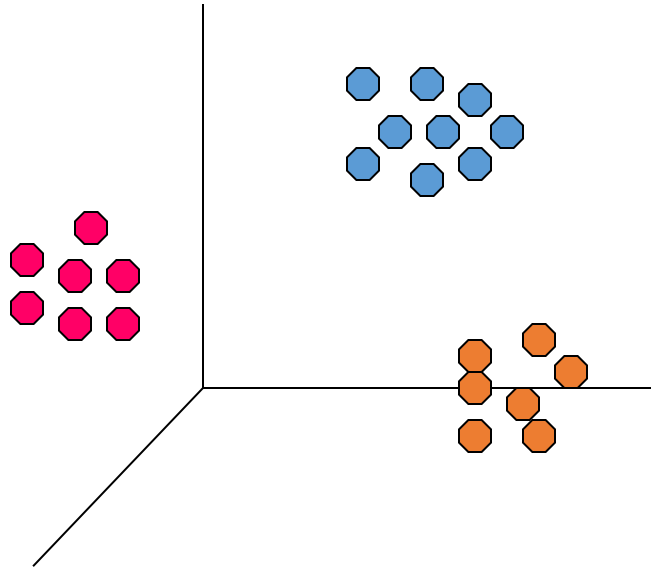
- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures.

# Illustrating Clustering

Intracluster distances are minimized

Intercluster distances are maximized

Euclidean Distance Based Clustering in 3-D space



# Clustering: Application 1

Market Segmentation:

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

# Clustering: Application 2

Document Clustering:

- Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

# Association Rule Discovery: Definition

Given a set of records each of which contain some number of items from a given collection;

- Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

| TID | Items |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Butter, Beans |
| 3 | Milk, Diaper, Butter, Coke |
| 4 | Bread, Milk, Diaper, Butter |
| 5 | Bread, Milk, Diaper, Coke |

**Example of Association Rules**

{Diaper} $\rightarrow$ {Butter},
{Milk, Bread} $\rightarrow$ {Beans, Coke},
{Butter, Bread} $\rightarrow$ {Milk},

# Association Rule Discovery: Application 1

Marketing and Sales Promotion:

- Let the rule discovered be

  *{Bagels, … } --> {Potato Chips}*

- Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
- Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
- Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

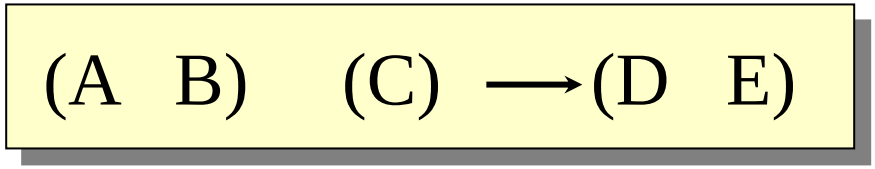
# Association Rule Discovery: Application

Inventory Management:

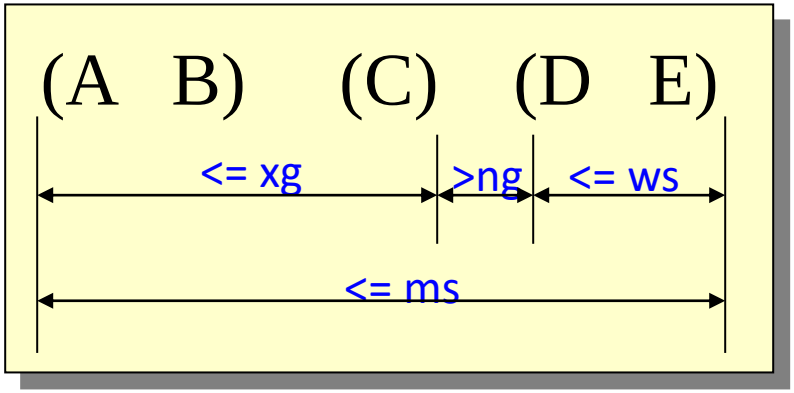
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

# Sequential Pattern Discovery: Definition

Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.




Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.




Timing constraints include maxgap (xg), mingap (ng), windowsize (ws), maxspan (ms)

# Sequential Pattern Discovery: Examples

In telecommunications alarm logs,

- (Inverter_Problem  Excessive_Line_Current)
  (Rectifier_Alarm) --> (Fire_Alarm)

In point-of-sale transaction sequences,

- Computer Bookstore:
  (Intro_To_Visual_C)  (C++_Primer) -->
  (Perl_for_dummies,Tcl_Tk)
- Athletic Apparel Store:
  (Shoes) (Racket, Racketball) --> (Sports_Jacket)

# Prediction/Regression

Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

Greatly studied in statistics, neural network fields.

Examples:

- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.

# Deviation/Anomaly Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection
- Network Intrusion Detection

# DM Process & Challenges

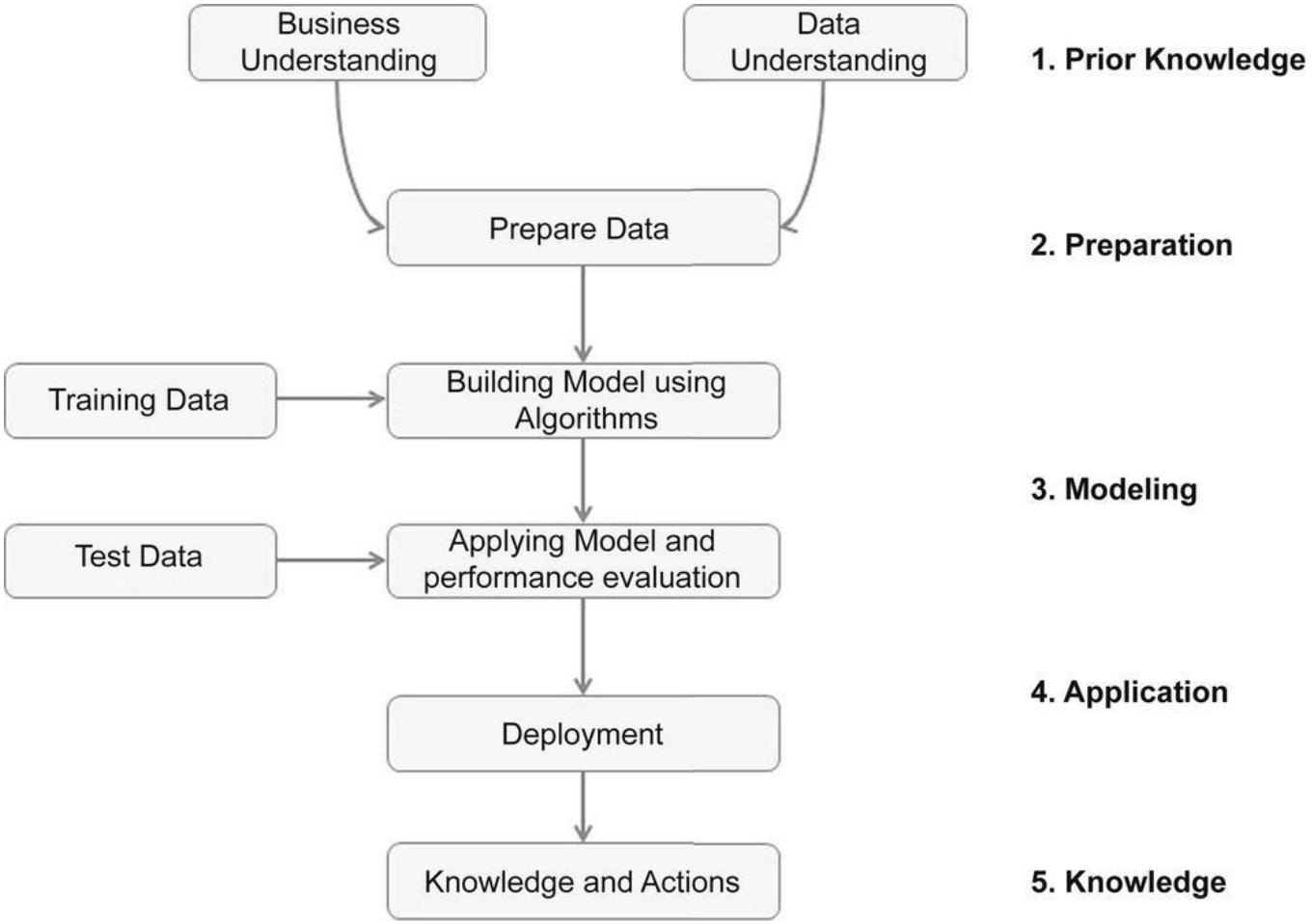
# DM Process

The standard data mining process involves

1. understanding the problem,
2. preparing the data (samples),
3. developing the model,
4. applying the model on a data set to see how the model may work in real world, and
5. production deployment.

A popular data mining process frameworks is CRISP-DM (Cross Industry Standard Process for Data Mining). This framework was developed by a consortium of companies involved in data mining

# Generic Data Mining Process



# Prior Knowledge

Data Mining tools/solutions identify hidden patterns.

- Generally we get many patterns
- Out of them many could be false or trivial.
- Filtering false patterns requires domain understanding.

Understanding how the data is collected, stored, transformed, reported, and used is essential.

# Data Preparation

Data needs to be understood. It requires descriptive statistics such as mean, median, mode, standard deviation, and range for each attribute

Data quality is an ongoing concern wherever data is collected, processed, and stored.

- The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.
- it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models to ensure a certain degree of data quality

Missing Values

- Need to track the data lineage of the data source to find right solution

Data Types and Conversion

- The attributes in a data set can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical
- data mining algorithms impose different restrictions on what data types they accept as inputs

Transformation

- Can go beyond type conversion, may include dimensionality reduction or numerosity reduction

Outliers are anomalies in the data set

- May occur legitimately or erroneously.

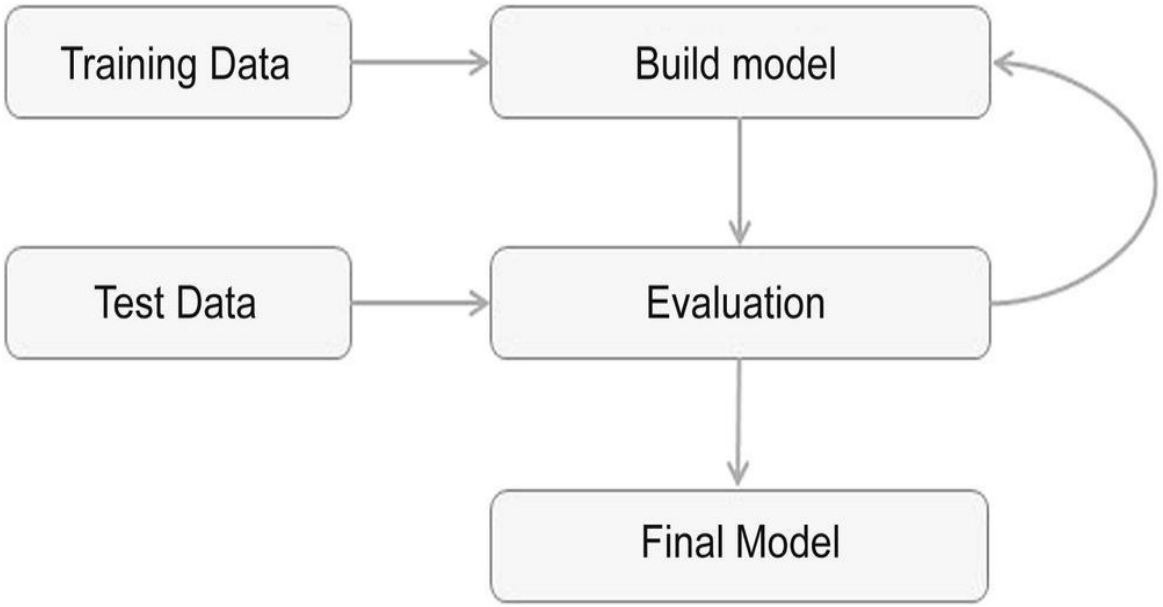
Feature Selection

- Many data mining problems involve a data set with hundreds to thousands of attributes, most of which may not be helpful. Some attributes may be correlated, e.g. sales amount and tax.

Data Sampling may be adequate in many cases

# Modeling & Evaluation

A model is the abstract representation of the data and its relationships in a given data set.
Data mining models can be classified into the following categories: classification, regression, association analysis, clustering, and outlier or anomaly detection.
Each category has a few dozen different algorithms; each takes a slightly different approach to solve the problem at hand



# Application

The model deployment stage considerations:

- assessing model readiness, technical integration, response time, model maintenance, and assimilation

Production Readiness

- Real-time response capabilities, and other business requirements

Technical Integration

- Use of modeling tools (e.g. RapidMiner), Use of PMML for portable and consistent format of model description, integration with other tools

Timeliness

- The trade-offs between production responsiveness and build time need to be considered

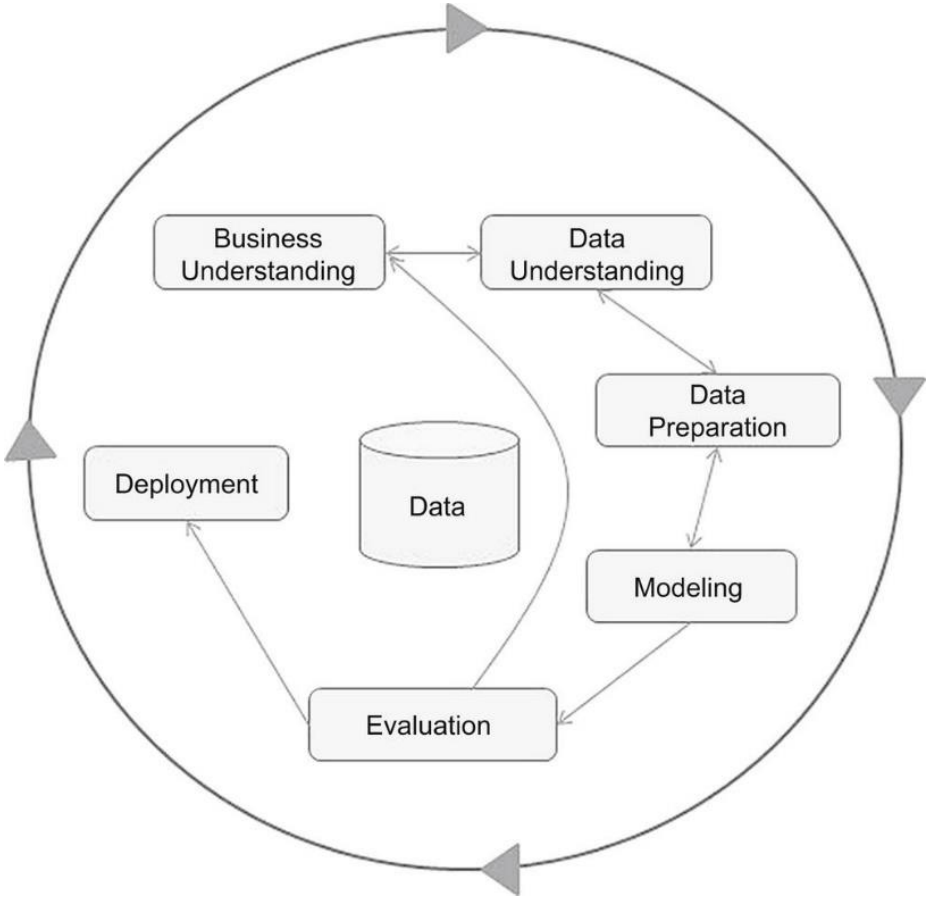
Remodeling

- The conditions in which the model is built may change after deployment

Assimilation

- The challenge is to assimilate the knowledge gained from data mining in the organization. For example, the objective may be finding logical clusters in the customer database so that separate treatment can be provided to each customer cluster.

# CRISP data mining framework




CRISP is the most popular methodology for analytics, data mining, and data science projects, with 43% share as per 2014 KDnuggets Poll.

CRISP-DM was conceived in 1996. In 1997 it got underway as a European Union project, led by SPSS, Teradata, Daimler AG, NCR Corporation and OHRA.

# DM Issues/Challenges

## DM Issues/Challenges – Mining Methodology

Mining various and new kinds of knowledge

Mining knowledge in multidimensional space

Data mining—an interdisciplinary effort

Boosting the power of discovery in a networked environment

Handling uncertainty, noise, or incompleteness of data

Pattern evaluation and pattern- or constraint-guided mining

# DM Issues/Challenges

## DM Issues/Challenges – User Interaction

Interactive mining

Incorporation of background knowledge

Ad hoc data mining and data mining query languages

Presentation and visualization of data mining results

## DM Issues/Challenges - Efficiency and Scalability

Efficiency and scalability of data mining algorithms

Parallel, distributed, and incremental mining algorithms

Cloud computing and cluster computing

## Text Books

| | Author(s), Title, Edition, Publishing House |
|---|---|
| T1 | Tan P. N., Steinbach M & Kumar V. "Introduction to Data Mining" Pearson Education |
| T2 | Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers |
| R1 | Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers |

# Thank You