Module-3

Business Intelligence Concepts and Applications

Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization‘s data and communicate the information to relevant users.

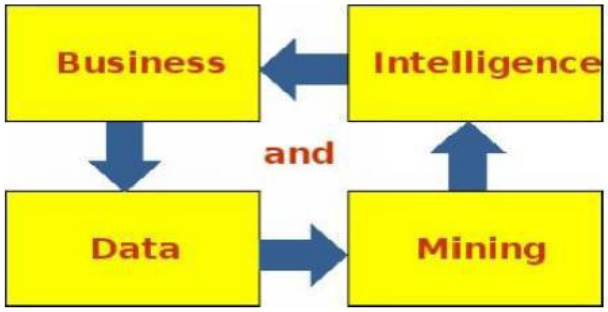

Fig 3.1: BIDM cycle

The nature of life and businesses is to grow. Information is the life-blood of business. Businesses effective than those based on feelings alone. Actions based on accurate data, information, knowledge, experimentation, and testing, using fresh insights, can more likely succeed and lead to sustained growth. One‘s own data can be the most effective teacher. Therefore, organizations should gather data, sift through it, analyse and mine it, find insights, and then embed those insights into their operating procedures.

There is a new sense of importance and urgency around data as it is being viewed as a new natural resource. It can be mined for value, insights, and competitive advantage. In a hyperconnected world, where everything is potentially connected to everything else, with potentially infinite correlations, data represents the impulses of nature in the form of certain events and attributes. A skilled business person is motivated to use this cache of data to harness nature, and to find new niches of unserved opportunities that could become profitable ventures.

BI for better decisions

The future is inherently uncertain. Risk is the result of a probabilistic world where there are no certainties and complexities abound. People use crystal balls, astrology, palmistry, ground hogs, and also mathematics and numbers to mitigate risk in decision-making. The goal is to make effective decisions, while reducing risk. Businesses calculate risks and make decisions based on a broad set of facts and insights. Reliable knowledge about the future can help managers make the right decisions with lower levels of risk.

The speed of action has risen exponentially with the growth of the Internet. In a hypercompetitive world, the speed of a decision and the consequent action can be a key advantage. The Internet and mobile technologies allow decisions to be made anytime, anywhere. Ignoring fast-moving changes can threaten the organization‘s future. Research has shown that an unfavorable comment about the company and its products on social media should not go

unaddressed for long. Banks have had to pay huge penalties to Consumer Financial Protection Bureau (CFPB) in United States in 2013 for complaints made on CFPB‘s websites. On the other hand, a positive sentiment expressed on social media should also be utilized as a potential sales and promotion opportunity, while the opportunity lasts.

Decision types

There are two main kinds of decisions: strategic decisions and operational decisions. BI can help make both better. Strategic decisions are those that impact the direction of the company. The decision to reach out to a new customer set would be a strategic decision. Operational decisions are more routine and tactical decisions, focused on developing greater efficiency. Updating an old website with new features will be an operational decision.

In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal. The consequences of the decision would be apparent some time later. Thus, one is constantly scanning for new possibilities and new paths to achieve the goals. BI can help with what-if analysis of many possible scenarios. BI can also help create new ideas based on new patterns found from data mining.

Operational decisions can be made more efficient using an analysis of past data. A classification system can be created and modeled using the data of past instances to develop a good model of the domain. This model can help improve operational decisions in the future. BI can help automate operations level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way. For example, a bank might want to make decisions about making financial loans in a more scientific way using data-based models. A decision-tree-based model could provide a consistently accurate loan decisions. Developing such decision tree models is one of the main applications of data mining techniques.

Effective BI has an evolutionary component, as business models evolve. When people and organizations act, new facts (data) are generated. Current business models can be tested against the new data, and it is possible that those models will not hold up well. In that case, decision models should be revised and new insights should be incorporated. An unending process of generating fresh new insights in real time can help make better decisions, and thus can be a significant competitive advantage.

BI Tools

BI includes a variety of software tools and techniques to provide the managers with the information and insights needed to run the business. Information can be provided about the current state of affairs with the capability to drill down into details, and also insights about emerging patterns which lead to projections into the future. BI tools include data warehousing, online analytical processing, social media analytics, reporting, dashboards, querying, and data mining.

BI tools can range from very simple tools that could be considered end-user tools, to very sophisticated tools that offer a very broad and complex set of functionality. Thus, even executives can be their own BI experts, or they can rely on BI specialists to set up the BI mechanisms for them. Thus, large organizations invest in expensive sophisticated BI solutions that provide good information in real time.

A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated statistical analysis.

A dashboarding system, such as IBM Cognos or Tableau, can offer a sophisticated set of tools for gathering, analyzing, and presenting data. At the user end, modular dashboards can be designed and redesigned easily with a graphical user interface. The back-end data analytical capabilities include many statistical functions. The dashboards are linked to data warehouses at the back end to ensure that the tables and graphs and other elements of the dashboard are updated in real time

Data mining systems, such as IBM SPSS Modeler, are industrial strength systems that provide capabilities to apply a wide range of analytical models on large data sets. Open source systems, such as Weka, are popular platforms designed to help mine large amounts of data to discover patterns.

BI Skills

As data grows and exceeds our capacity to make sense of it, the tools need to evolve, and so should the imagination of the BI specialist. ―Data Scientist‖ has been called as the hottest job of this decade.

A skilled and experienced BI specialist should be open enough to go outside the box, open the aperture and see a wider perspective that includes more dimensions and variables, in order to find important patterns and insights. The problem needs to be looked at from a wider perspective to consider many more angles that may not be immediately obvious. An imaginative solution should be proposed for the problem so that interesting and useful results can emerge.

A good data mining project begins with an interesting problem to solve. Selecting the right data mining problem is an important skill. The problem should be valuable enough that solving it would be worth the time and expense. It takes a lot of time and energy to gather, organize, cleanse, and prepare the data for mining and other analysis. The data miner needs to persist with the exploration of patterns in the data. The skill level has to be deep enough to engage with the data and make it yield new useful insights.

BI Applications

BI tools are required in almost all industries and functions. The nature of the information and the speed of action may be different across businesses, but every manager today needs access to BI tools to have up-to-date metrics about business performance. Businesses need to embed new insights into their operating processes to ensure that their activities continue to evolve with more efficient practices. The following are some areas of applications of BI and data mining.

Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer, sell more of its offerings to the existing customers, and also, expand the pool of customers it serves. BI applications can impact many aspects of marketing.

1. *Maximize the return on marketing campaigns*: Understanding the customer‘s pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.

2. *Improve customer retention (churn analysis)*: It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit, can help the business design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.

3.Maximize customer value (cross-, up-selling): Every contact with the customer should be seen as an opportunity to gauge their current needs. Offering a customer new products and solutions based on those imputed needs can help increase revenue per customer. Even a customer complaint can be seen as an opportunity to wow the customer. Using the knowledge of the customer‘s history and value, the business can choose to sell a premium service to the customer.

4. *Identify and delight highly-valued customers*. By segmenting the customers, the best customers can be identified. They can be proactively contacted, and delighted, with greater attention and better service. Loyalty programs can be managed more effectively.

5. *Manage brand image*. A business can create a listening post to listen to social media chatter about itself. It can then do sentiment analysis of the text to understand the nature of comments, and respond appropriately to the prospects and customers.

Healthcare and Wellness

Health care is one of the biggest sectors in advanced economies. Evidence based medicine is the newest trend in data-based health care management. BI applications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.

1. *Diagnose disease in patients*: Diagnosing the cause of a medical condition is the critical first step in a medical engagement. Accurately diagnosing cases of cancer or diabetes can be a matter of life and death for the patient. In addition to the patient‘s own current situation, many other factors can be considered, including the patient‘s health history, medication history, family‘s history, and other environmental factors. This makes diagnosis as much of an art form as it is science. Systems, such as IBM Watson, absorb all the medical research to date and make probabilistic diagnoses in the form of a decision tree, along with a full explanation for their recommendations. These systems take away most of the guess work done by doctors in diagnosing ailments.

2. *Treatment effectiveness*: The prescription of medication and treatment is also a difficult choice out of so many possibilities. For example, there are more than 100 medications for hypertension (high blood pressure) alone. There are also interactions in terms of which drugs work well with others and which drugs do not. Decision trees can help doctors learn about and prescribe more effective treatments. Thus, the patients could recover their health faster with a lower risk of complications and cost.

3. *Wellness management*: This includes keeping track of patient health records, analyzing customer health trends and proactively advising them to take any needed precautions.

4. *Manage fraud and abuse*: Some medical practitioners have unfortunately been found to conduct unnecessary tests, and/or overbill the government and health insurance companies. Exception reporting systems can identify such providers and action can be taken against them.

5. *Public health management*: The management of public health is one of the important responsibilities of any government. By using effective forecasting tools and techniques, governments can better predict the onset of disease in certain areas in real time. They can thus be better prepared to fight the diseases. Google has been known to predict the movement of certain diseases by tracking the search terms (like flu, vaccine) used in different parts of the world.

EDUCATION:
RETAIL:
FINANCIAL SERVICES
INSURANCE
MANUFACTURING
TELECOM
PUBLIC SECTOR (Refer textbook for more information on these topics)

Data Warehousing

A data warehouse (DW) is an organized collection of integrated, subject oriented databases designed to support decision support functions. DW is organized at the right level of granularity to provide clean enterprise-wide data in a standardized format for reports, queries, and analysis. DW is physically and functionally separate from an operational and transactional database.

Creating a DW for analysis and queries represents significant investment in time and effort. It has to be constantly kept up-to-date for it to be useful. DW offers many business and technical benefits.

DW supports business reporting and data mining activities. It can facilitate distributed access to up-to-date business knowledge for departments and functions, thus improving business efficiency and customer service. DW can present a competitive advantage by facilitating decision making and helping reform business processes.

DW enables a consolidated view of corporate data, all cleaned and organized. Thus, the entire organization can see an integrated view of itself. DW thus provides better and timely information. It simplifies data access and allows end users to perform extensive analysis. It enhances overall IT performance by not burdening the operational databases used by Enterprise Resource Planning (ERP) and other systems.

Design Considerations for DW

The objective of DW is to provide business knowledge to support decision making. For DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up-to-date. Here are some requirements for a good DW:

1. *Subject oriented*: To be effective, a DW should be designed around a subject domain, i.e. to help solve a certain category of problems.

2. *Integrated*: The DW should include data from many functions that can shed light on a particular subject area. Thus the organization can benefit from a comprehensive view of the subject area.

3. *Time-variant (time series):* The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.

4. *Nonvolatile*: DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.

5. *Summarized*: DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reduces the number of variables or dimensions of the data to make them more meaningful for
the decision makers.

6. *Not normalized*: DW often uses a star schema, which is a rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries.

7. *Metadata*: Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.

8. *Near Real-time and/or right-time (active*): DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

DW Development Approaches

There are two fundamentally different approaches to developing DW: top down and bottom up. The top-down approach is to make a comprehensive DW that covers all the reporting needs of the enterprise. The bottom-up approach is to produce small data marts, for the reporting needs of different departments or functions, as needed.
The smaller data marts will eventually align to deliver comprehensive EDW capabilities. The top-down approach provides consistency but takes more time and resources. The bottom-up approach leads to healthy local ownership and maintainability of data.

	Functional Data Mart	Enterprise Data Warehouse
Scope	One subject or functional area	Complete enterprise data needs
Value	Functional area reporting and insights	Deeper insights connecting multiple functional areas
Target organization	Decentralized management	Centralized management
Time	Low to medium	High
Cost	Low	High
Size	Small to medium	Medium to large
Approach	Bottom up	Top down
Complexity	Low (fewer data transformations)	High (data standardization)
Technology	Smaller scale servers and databases	Industrial strength

Table 3.1: Comparing Data Mart and Data Warehouse

DW Architecture

9.

DW has four key elements .The first element is the data sources that provide the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the methods of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

Data Sources

Data Warehouses are created from structured data sources. Unstructured data such as text data would need to be structured before inserted into the DW.

1. *Operations data:* This includes data from all business applications, including from ERPs systems that form the backbone of an organization‘s IT systems. The data to be extracted will

depend upon the subject matter of the data warehouse. For example, for a sales/marketing data mart, only the data about customers, orders, customer service, and so on would be extracted.

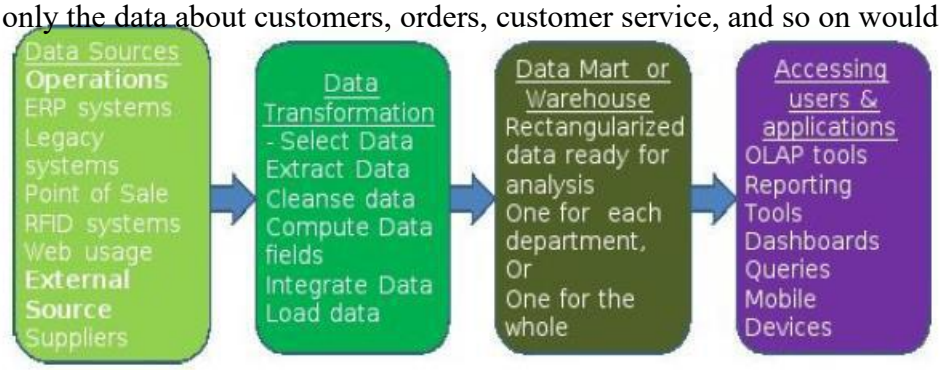

Fig 3.2: Data Warehousing Architecture

2. *Specialized applications*: This includes applications such as Point of Sale (POS) terminals, and e-commerce applications, that also provide customer-facing data. Supplier data could come from Supply Chain Management systems. Planning and budget data should also be added as needed for making comparisons against targets.
3. *External syndicated data*: This includes publicly available data such as weather or economic activity data. It could also be added to the DW, as needed, to provide good contextual information to decision makers.

Data Loading Processes

The heart of a useful DW is the processes to populate the DW with good quality data. This is called the Extract-Transform-Load (ETL) cycle.

1. Data should be extracted from the operational (transactional) database sources, as well as from other applications, on a regular basis.

2. The extracted data should be aligned together by key fields and integrated into a single data set. It should be cleansed of any irregularities or missing values. It should be rolled-up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.

3. This transformed data should then be uploaded into the DW.This ETL process should be run at a regular frequency. Daily transaction data can be extracted from ERPs, transformed, and uploaded to the database the same night. Thus, the DW is up to date every morning. If a DW is needed for near-real-time information access, then the ETL processes would need to be executed more frequently. ETL work is usually done using automated using programming scripts that are written, tested, and then deployed for periodically updating the DW.

Data Warehouse Design

Star schema is the preferred data architecture for most DWs. There is a central fact table that provides most of the information of interest. There are lookup tables that provide detailed values for codes used in the central table. For example, the central table may use digits to represent a sales person. The lookup table will help provide the name for that sales person code.

Here is an example of a star schema for a data mart for monitoring sales performance Other schemas include the snowflake architecture. The difference between a star and snowflake is that in the latter, the look-up tables can have their own further look up tables.
There are many technology choices for developing DW. This includes selecting the right database management system and the right set of data management tools. There are a few big and reliable providers of DW systems. The provider of the operational DBMS may be chosen for DW also.
Alternatively, a best-of-breed DW vendor could be used. There are also a variety of tools out there for data migration, data upload, data retrieval, and data analysis.

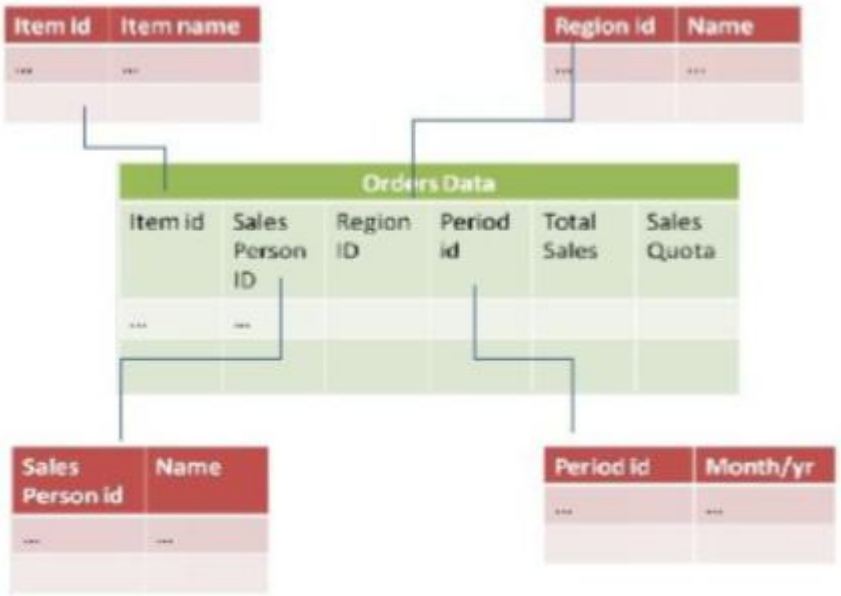

Figure 3.2: Star Schema Architecture for DW

DW Access

Data from the DW could be accessed for many purposes, by many users, through many devices.

1. A primary use of DW is to produce routine management and monitoring reports. For example, a sales performance report would show sales by many dimensions, and compared with plan. A dashboarding system will use data from the warehouse and present analysis to users. The data from DW can be used to populate customized performance dashboards for executives. The dashboard could include drill-down capabilities to analyze the performance data for root cause analysis.

2. The data from the DW could be used for ad-hoc queries and any other applications that make use of the internal data.

3. Data from DW is used to provide data for mining purposes. Parts of the data would be extracted, and then combined with other relevant data, for data mining.

DW Best Practices

A data warehousing project reflects a significant investment into information technology (IT). All of the best practices in implementing any IT project should be followed.

1. The DW project should *align with the corporate strategy*. Top management should be consulted for setting objectives. Financial viability (ROI) should be established. The project must be managed by both IT and business professionals. The DW design should be carefully tested before beginning development work. It is often much more expensive to redesign after development work has begun.
2. It is important to *manage user expectations*. The data warehouse should be built incrementally. Users should be trained in using the system so they can absorb the many features of the system.

3. *Quality and adaptability* should be built in from the start. Only relevant, cleansed, and high-quality data should be loaded. The system should be able to adapt to new tools for access. As business needs change, new data marts may need to be created for new needs.

Data Mining

Data mining is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data. Patterns must be valid, novel, potentially useful, and understandable. The implicit assumption is that data about the past can reveal patterns of activity that can be projected into the future.

Data mining is a multidisciplinary field that borrows techniques from a variety of fields. It utilizes the knowledge of data quality and data organizing from the databases area. It draws modeling and analytical techniques from statistics and computer science (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of business management.

The field of data mining emerged in the context of pattern recognition in defense, such as identifying a friend-or-foe on a battlefield. Like many other defense-inspired technologies, it has evolved to help gain a competitive advantage in business.

For example, ―customers who buy cheese and milk also buy bread 90 percent of the time‖ would be a useful pattern for a grocery store, which can then stock the products appropriately. Similarly, ―people with blood pressure greater than 160 and an age greater than 65 were at a high risk of dying from a heart stroke‖ is of great diagnostic value for doctors, who can then focus on treating such patients with urgent care and great sensitivity.

Past data can be of predictive value in many complex situations, especially where the pattern may not be so easily visible without the modelling technique. Here is a dramatic case of a data-driven decision-making system that beats the best of human experts. Using past data, a decision tree model was developed to predict votes for Justice Sandra Day O‘Connor, who had a swing vote in a 5–4 divided US Supreme Court. All her previous decisions were coded on a few

variables. What emerged from data mining was a simple four-step decision tree that was able to accurately predict her votes 71 percent of the time. In contrast, the legal analysts could at best predict correctly 59 percent of the time. (Source: Martin et al. 2004)

Gathering and selecting data

The total amount of data in the world is doubling every 18 months. There is an ever-growing avalanche of data coming with higher velocity, volume, and variety. One has to quickly use it or lose it. Smart data mining requires choosing where to play. One has to make judicious decisions about what to gather and what to ignore, based on the purpose of the data mining exercises.It is like deciding where to fish; as not all streams of data will be equally rich in potential insights.

To learn from data, quality data needs to be effectively gathered, cleaned and organized, and then efficiently mined. One requires the skills and technologies for consolidation and integration of data elements from many sources. Most organizations develop an enterprise data model (EDM) to organize their data. An EDM is a unified, high-level model of all the data stored in an organization‘s databases. The EDM is usually inclusive of the data generated from all internal systems. The EDM provides the basic menu of data to create a data warehouse for a particular decision-making purpose. DWs help organize all this data in an easy and usable manner so that it can be selected and deployed for mining. The EDM can also help imagine what relevant external data should be gathered to provide context and develop good predictive relationships with the internal data. In the United States, the various federal and local governments and their regulatory agencies make a vast variety and quantity of data available at data.gov.

Gathering and curating data takes time and effort, particularly when it is unstructured or semistructured. Unstructured data can come in many forms like databases, blogs, images, videos, audio, and chats. There are streams of unstructured social media data from blogs, chats, and tweets. There are streams of machine-generated data from connected machines, RFID tags, the internet of things, and so on. Eventually the data should be rectangularized,that is, put in rectangular data shapes with clear columns and rows, before submitting it to data mining.

Knowledge of the business domain helps select the right streams of data for pursuing new insights. Only the data that suits the nature of the problem being solved should be gathered. The data elements should be relevant, and suitably address the problem being solved. They could directly impact the problem, or they could be a suitable proxy for the effect being measured. Select data could also be gathered from the data warehouse. Every industry and function will have its own requirements and constraints. The health care industry will provide a different type of data with different data names. The HR function would provide different kinds of data. There would be different issues of quality and privacy for these data.

Data cleansing and preparation

The quality of data is critical to the success and value of the data mining project. Otherwise, the situation will be of the kind of garbage in and garbage out (GIGO). The quality of incoming

data varies by the source and nature of data. Data from internal operations is likely to be of higher quality, as it will be accurate and consistent. Data from social media and other public sources is less under the control of business, and is less likely to be reliable.

Data almost certainly needs to be cleansed and transformed before it can be used for data mining. There are many ways in what data may need to be cleansed – filling missing values, reigning in the effects of outliers, transforming fields, binning continuous variables, and much more – before it can be ready for analysis. Data cleansing and preparation is a labor-intensive or semi-automated activity that can take up to 60-70% of the time needed for a data mining project.

1. Duplicate data needs to be removed. The same data may be received from multiple sources. When merging the data sets, data must be deduped.
2. Missing values need to be filled in, or those rows should be removed from analysis. Missing values can be filled in with average or modal or default values.

3. Data elements should be comparable. They may need to be (a) transformed from one unit to another. For example, total costs of health care and the total number of patients may need to be reduced to cost/patient to allow comparability of that value. Data elements may need to be adjusted to make them (b) comparable over time also. For example, currency values may need to be adjusted for inflation; they would need to be converted to the same base year for comparability.
They may need to be converted to a common currency. Data should be (c) stored at the same granularity to ensure comparability. For example, sales data may be available daily, but the sales person compensation data may only be available monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.

4. Continuous values may need to be binned into a few buckets to help with some analyses. For instance, work experience could be binned as low, medium, and high.

5. Outlier data elements need to be removed after careful review, to avoid the skewing of results. For example, one big donor could skew the analysis of alumni donors in an educational setting.

6. Ensure that the data is representative of the phenomena under analysis by correcting for any biases in the selection of data. For example, if the data includes many more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.

7. Data may need to be selected to increase information density. Some data may not show much variability, because it was not properly recorded or for other reasons. This data may dull the effects of other differences in the data and should be removed to improve the information density of the data.

Outputs of Data Mining

Data mining techniques can serve different types of objectives. The outputs of data mining will reflect the objective being served. There are many ways of representing the outputs of data mining.
One popular form of data mining output is a decision tree. It is a hierarchically branched structure that helps visually follow the steps to make a model-based decision. The tree may have certain attributes, such as probabilities assigned to each branch. A related format is a set of businessrules, which are if-then statements that show causality. A decision tree can be mapped to business rules. If the objective function is prediction, then a decision tree or business rules are the most appropriate mode of representing the output.

The output can be in the form of a regression equation or mathematical function that represents the best fitting curve to represent the data. This equation may include linear and nonlinear terms. Regression equations are a good way of representing the output of classification exercises. These are also a good representation of forecasting formulae.

Population ―centroid‖ is a statistical measure for describing central tendencies of a collection of data points. These might be defined in a multidimensional space. For example, a centroid could be ―middle-aged, highly educated, highnet worth professionals, married with two children, living in the coastal areas‖. Or a population of ―20-something, ivy-league-educated, tech entrepreneurs based in Silicon Valley‖. Or it could be a collection of ―vehicles more than 20 years old, giving low mileage per gallon, which failed environmental inspection‖. These are typical representations of the output of a cluster analysis exercise.

Business rules are an appropriate representation of the output of a market basket analysis exercise. These rules are if-then statements with some probability parameters associated with each rule. For example, those that buy milk and bread will also buy butter (with 80 percent probability).

Evaluating Data Mining Results

There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many techniques for classification,decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.

Predictive Accuracy = (Correct Predictions) / Total Predictions

Suppose a data mining project has been initiated to develop a predictive model for cancer patients using a decision tree. Using a relevant set of variables and data instances, a decision tree model has been created. The model is then used to predict other data instances. When a true positive data point is positive, that is a correct prediction, called a true positive (TP). Similarly, when a true negative data point is classified as negative, that is a true negative (TN). On the other hand, when a true-positive data point is classified by the model as negative, that is an incorrect prediction, called a false negative (FN). Similarly, when a true-negative data point is classified as positive, that is classified as a false positive (FP). This is represented using the confusion matrix (Figure 4.1).

ConfusionMatrix		True Class	
		Positive	*Negative*
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
Predicted class	Negative	False Negative (FN)	True Negative (TN)

Figure 4.1: Confusion Matrix

Thus the predictive accuracy can be specified by the following formula.
Predictive Accuracy =(TP+TN)/(TP+TN+FP+FN).

Data Mining Techniques

Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved .The most important class of problems solved using data mining are classification problems. Classification techniques are called supervised learning as there is a way to supervise whether the model is providing the right or wrong answers. These are problems where data from past decisions is mined to extract the few rules and patterns that would improve the accuracy of the decision making process in the future. The data of past decisions is organized and mined for decision rules or equations, that are then codified to produce more accurate decisions.

Decision trees are the most popular data mining technique, for many reasons.

1. Decision trees are easy to understand and easy to use, by analysts as well as executives. They also show a high predictive accuracy.
2. Decision trees select the most relevant variables automatically out of all the available variables for decision making.
3. Decision trees are tolerant of data quality issues and do not require much data preparation from the users.
4. Even non-linear relationships can be handled well by decision trees.

There are many algorithms to implement decision trees. Some of the popular ones are C5, CART and CHAID.

Regression is a most popular statistical data mining technique. The goal of regression is to derive a smooth well-defined curve to best the data. Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature. Simply plotting the data may show a non-linear curve. Applying a non-linear regression equation will fit the data very well with high accuracy. Once such a regression model has been developed, the energy consumption on any future day can be predicted using this equation. The accuracy of the regression model depends entirely upon the dataset used and not at all on the algorithm or tools used.

Artificial Neural Networks (ANN) is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure: Neurons receive stimuli, process them,and communicate their results to other neurons successively, and eventually a neuron outputs a decision. A decision task may be processed by just one neuron and the result may be communicated soon. Alternatively, there could be many layers of neurons involved in a decision task, depending upon the complexity of the domain. The neural network can be trained by making a decision over and over again with many data points. It will continue to learn by adjusting its internal computation and communication parameters based on feedback received on its previous decisions. The intermediate values passed within the layers of neurons may not make any intuitive sense to an observer. Thus, the neural networks are considered a black-box system.

Cluster Analysis is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural groupings of things. Data instances that are similar to (or near) each other are categorized into one cluster, while data instances that are very different (or far away) from each other are categorized into separate clusters. There can be any number of clusters that could be produced by the data. The K-means technique is a popular technique and allows the user guidance in selecting the right number (K) of clusters from the data. Clustering is also known as the segmentation technique. It helps divide and conquer large data sets. The technique shows the clusters of things from past data. The output is the centroids for each cluster and the allocation of data points to their cluster. The centroid definition is used to assign new data instances can be assigned to their cluster homes. Clustering is also a part of the artificial intelligence family of techniques.

Association rules are a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities. This is the heart of the personalization engine used by ecommerce sites like Amazon.com and streaming movie sites like Netflix.com. The technique helps find interesting relationships (affinities) between variables (items or events). These are represented as rules of the form X ® Y, where X and Y are sets of data items. A form of unsupervised learning, it has no dependent variable; and there are no right or wrong answers. There are just stronger and weaker affinities. Thus, each rule has a confidence level assigned to it. A part of the machine learning family, this technique achieved legendary status when a fascinating relationship was found in the sales of diapers and beers.

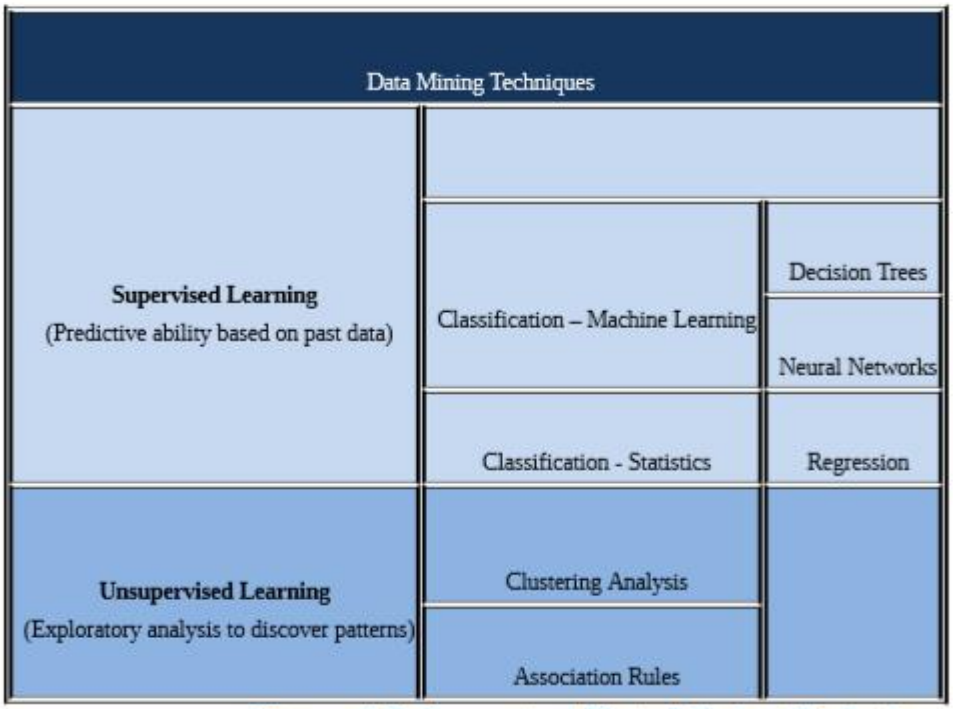

Figure 4.2: Important Data Mining Techniques

Tools and Platforms for Data Mining

Data Mining tools have existed for many decades. However, they have recently become more important as the values of data have grown and the field of big data analytics has come into prominence. There are a wide range of data mining platforms available in the market today.

1. Simple or sophisticated: There are simple end-user data mining tools such as MS Excel, and there are more sophisticated tools such as IBM SPSS Modeler.

2. Stand-alone or Embedded: There are stand alone tools and there are tools embedded in an existing transaction processing or data warehousing or ERP system.

3. Open source or Commercial: There are open source and freely available tools such as Weka, and there are commercial products.

4. User interface: There are text-based tools that require some programming skills, and there are GUI-based drag-and-drop format tools.

5. Data formats: There are tools that work only on proprietary data formats and there are those directly accept data from a host of popular data management tools formats.

Table 4.1: Comparison of Popular Data Mining Platforms

Feature	Excel	IBM SPSS Modeler	Weka
Ownership	Commercial	Commercial, expensive	Open-source, free
Data Mining Features	Limited; available with add-on modules	Extensive features, unlimited data sizes	Extensive, performance issues with large data
Stand-alone	Stand-alone	Embedded in BI software suites	Stand-alone
User skills needed	End-users	For skilled BI analysts	Skilled BI analysts

User interface	Text and click, Easy	Drag & Drop use, colorful, beautiful GUI	GUI, mostly follows menu design
Data formats	Industry-standard	Variety of data sources accepted	Proprietary

Data Mining Best Practices

Effective and successful use of data mining activity requires both business and technology skills. The business aspects help understand the domain and the key questions. It also helps one imagine possible relationships in the data, and create hypotheses to test it. The IT aspects help fetch the data from many sources, clean up the data, assemble it to meet the needs of the business problem, and then run the data mining techniques on the platform. An important element is to go after the problem iteratively. It is better to divide and conquer the problem with smaller amounts of data, and get closer to the heart of the solution in an iterative sequence of steps. There are several best practices learned from the use of data mining techniques over a long period of time. The Data Mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM). It has six essential steps :

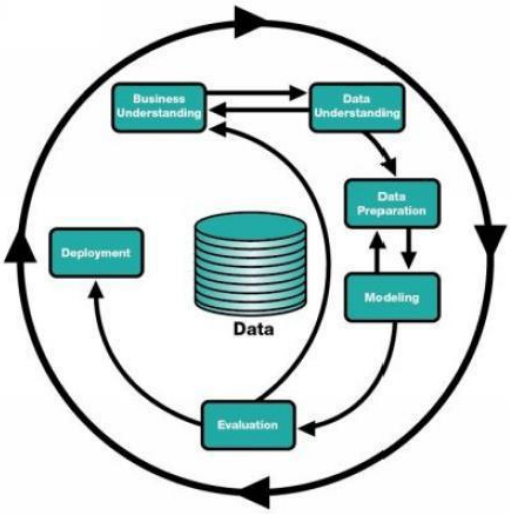

Fig 3.3: CRISP-DM Data Mining cycle

1. *Business Understanding*: The first and most important step in data mining is asking the right business questions. A question is a good one if answering it would lead to large payoffs for the organization, financially and otherwise. In other words, selecting a data mining project is like any other project, in that it should show strong payoffs if the project is successful. There should be strong executive support for the data mining project, which means that the project aligns well with the business strategy. A related important step is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in terms of a proposed model as well in the data sets available and required.

2. *Data Understanding*: A related important step is to understand the data available for mining. One needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem. Without relevant data, the hypotheses cannot be tested.

3. *Data Preparation*: The data should be relevant, clean and of high quality. It‘s important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. Data cleaning can take 60-70% of the time in a data mining project. It may be desirable to continue to experiment and add new data elements from external sources of data that could help improve predictive accuracy.

4. *Modeling*: This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported. Patience is required in continuously engaging with the data until the data yields some good insights. A host of modeling tools and algorithms should be used. A tool could be tried with different options, such as running different decision tree algorithms.

5. *Model Evaluation*: One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques, and conducting many what-if scenarios, to build confidence in the solution. One should evaluate and improve the model's predictive accuracy with more test data. When the accuracy has reached some satisfactory level, then the model should be deployed.

Dissemination and rollout: It is important that the data mining solution is presented to the key stakeholders, and is deployed in the organization. Otherwise the project will be a waste of time and will be a setback for establishing and supporting a data-based decision-process culture in the organization. The model should be eventually embedded in the organization's business processes.

MYTHS ABOUT DATA MINING

There are many myths about this area, scaring away many business executives from using Data mining. Data Mining is a mindset that presupposes a faith in the ability to reveal insights. By itself, data mining is not too hard, nor is it too easy. It does require a disciplined approach and some cross disciplinary skills.

Myth #1: Data Mining is about algorithms. Data mining is used by business to answer important and practical business questions. Formulating the problem statement correctly and identifying imaginative solutions for testing are far more important before The data mining algorithms gets called in. Understanding the relative strengths of various algorithms is helpful but not mandatory.

Myth #2:Data Mining is about predictive accuracy. While important, predictive accuracy is a feature of the algorithm.As in myth#1,the quality of output is a strong function of the right problem,right hypothesis, and the right data.

Myth #3:Data Mining requires a data warehouse.While the presence of a data warehouse assists in the gathering of information, sometimes the creation of the data warehouse itself can benefit from some exploratory data mining. Some data mining problems may benefit from clean data available directly from the DW, but aDW is not mandatory.

Myth #4:Data Mining requires large quantities of data. Many interesting data mining exercises are done using small or medium sized data sets, at low costs, using end-user tools.

Myth #5:Data Mining requires a technology expert. Many interesting data mining exercises are done by end-users andexecutives using simple everyday tools like spreadsheets.

DATA MINING MISTAKES

Refer text book for this section.

DATA VISUALIZATION

- DV is the art and science of making data easy to understand and consume for the end user.DV is the last step in the data life cycle. This is where the data is processed for presentation in an easy-to-consume manner to the right audience for the right purpose.

Excellence in visualization

- Data can be represented in the form of rectangular tables or it can be presented in colorful graphs of various types.

Objectives for graphical excellence

- Show , and even reveal, the data.
- Induce the viewer to think of the substance of the data.
- Avoid distorting what the data have to say.
- Make large datasets coherent.
- Encourage the eyes to compare different pieces of data.
- Reveal the data at several levels of detail.
- Serve a reason ably clear purpose
- Closely integrate with the statistical and verbal descriptions of the dataset.

Types of charts

- Line graph
- Scatter plot
- Bar graph
- Stacked bar graphs
- Histograms
- Pie charts
- Box charts
- Bubble graph
- Dial
- Geographical data maps
- Pictographs

Diagram of charts

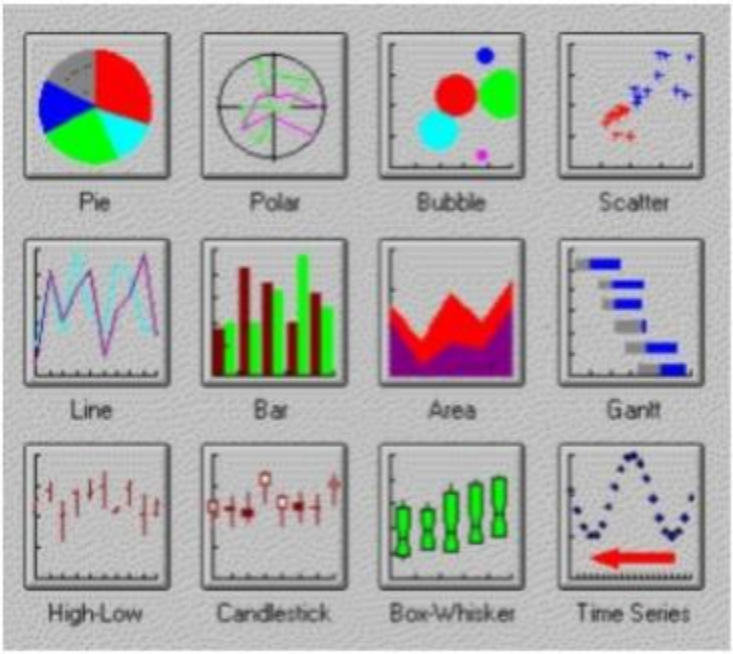

Figure 5.1: Many types of graphs

Figure 5.3: US tweet map (Source: Slate.com)

Ex for Pictographs

11. *Pictographs*: One can use pictures to represent data. E.g. Figure 5.2 shows the number of liters of water needed to produce one pound of each of the products, where images are used to show the product for easy reference. Each droplet of water also represents 50 liters of water.

Figure 5.4: Pictograph of Water footprint (source : waterfootprint.org)

Visualization example

To demonstrate how each of the visualization tools could be used, imagine an executive of a company who wants to analyze the sales performance of his division.

Product	Revenue	Orders	SalesPers
AA	9731	131	23
BB	355	43	8
CC	992	32	6
DD	125	31	4
EE	933	30	7
FF	676	35	6
GG	1411	128	13
HH	5116	132	38
JJ	215	7	2
KK	3833	122	50
LL	1348	15	7
MM	1201	28	13

Table 5.1: Raw Performance Data

important ratios to the right of the table (Table 5.2).

Product	Revenue	Orders	SalesPers	Rev/Order	Rev/SalesP	Orders/SalesP
AA	9781	131	23	74.3	423.1	5.7
HH	5116	132	38	38.8	134.6	3.5
KK	3683	122	30	31.4	76.7	2.4
GG	1411	128	13	11.0	108.5	9.8
LL	1345	15	7	89.9	192.6	2.1
MM	1201	28	13	42.9	92.4	2.2
CC	992	32	6	31.0	165.3	5.3
EE	933	30	7	31.1	133.3	4.3
FF	676	35	6	19.3	112.7	5.8
BB	355	42	8	8.5	44.4	5.4
JJ	215	7	2	30.7	107.5	3.5
DD	125	31	4	4.0	31.3	7.8
Total	25936	734	177	35.3	146.5	4.1

Table 5.2: Sorted data, with additional ratios

proportion drops significantly from the first product to the next. (Figure 5.5). It is interesting to note that the top 3 products produce almost 75% of the revenue.

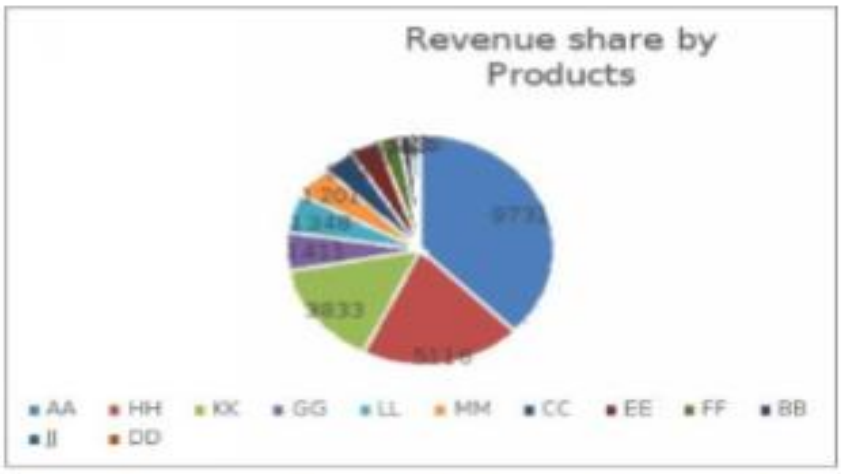

Figure 5.5: Revenue Share by Product

Figure 5.6: Orders by Products

Tips for DV

- Fetch appropriate and correct data for analysis.
- Sort the data in the most appropriate manner.
- Choose appropriate method for present the data.
- The dataset could be pruned.
- The visualization could show additional dimension for references.
- The numerical data may need to be binned into few categories.

- High level visualization could be backed by more detailed analysis.
- Need to present additional textual information.