Module 3

Business Intelligence Concepts and Application

1. Explain BIDM cycle with diagram.

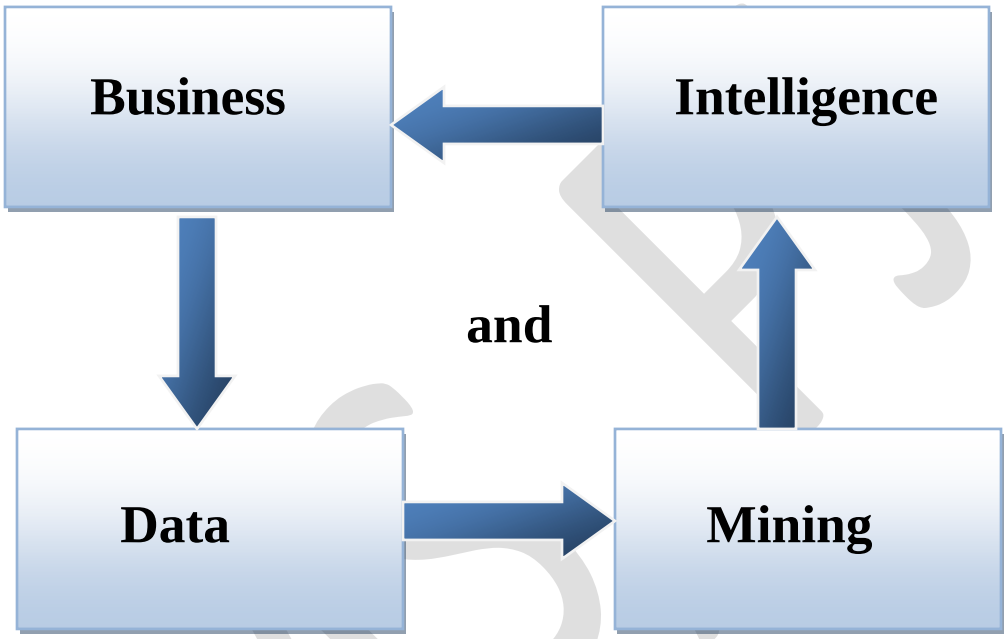


Figure 2.1: BIDM cycle

- Business intelligence (BI) is an umbrella term that includes a variety of **IT applications** that are used to **analyze an organization's data** and c**ommunicate** the information to **relevant users.**
- Information is the life-blood of business. Businesses use many techniques for **understanding their environment and predicting the future** for their own **benefit and growth**.
- **Decisions** are made from **facts and feelings**. Data-based decisions are more 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, analyze 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 hyper connected 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.

2. Explain BI decision types and tools.

- **BI 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**.

    Eg: 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**.

    Eg: 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.
  - ❖ **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 **manager**s 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 futur**e.
  - ❖ 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 **dash boarding 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.**

3.Explain the applications of BI in

i) CRM ii) Healthcare and wellness. iii)Education

iv) Retail v) Telecom

- **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**.

i) Customer Relationship Management (CRM)

- 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.

ii) 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. 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

iii) Education

As higher education becomes more expensive and competitive, it becomes a great user of data based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

1. *Student Enrollment (Recruitment and Retention)*:

Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend. Schools can develop models of what kinds of students are attracted to the school, and then reach out to those students. The students at risk of not returning can be flagged, and corrective measures can be taken in time.

2. *Course offerings*:

Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students. This can help increase class size, reduce costs, and improve student satisfaction.

3. *Fund-raising from Alumni and other donors*:

Schools can develop predictive models of which alumni are most likely to pledge financial support to the school. Schools can create a profile for alumni more likely to pledge donations to the school. This could lead to a reduction in the cost of mailings and other forms of outreach to alumni.

iv) Retail

Retail organizations grow by meeting customer needs with quality products, in a convenient, timely, and cost-effective manner. Understanding emerging customer shopping patterns can help retailers organize their products, inventory, store layout, and web presence in order to delight their customers, which in turn would help increase revenue and profits. Retailers generate a lot of transaction and logistics data that can be used to diagnose and solve problems.

1. *Optimize inventory levels at different locations*:

Retailers need to manage their inventories carefully. Carrying too much inventory imposes carrying costs, while carrying too little inventory can cause stock-outs and lost sales opportunities. Predicting sales trends dynamically can help retailers move inventory to where it is most in demand. Retail organizations can provide their suppliers with real time information about sales of their items, so the suppliers can deliver their product to the right locations and minimize stock-outs.

2. *Improve store layout and sales promotions*:

A market basket analysis can develop predictive models of which products sell together often. This knowledge of affinities between products can help retailers co-locate those products. Alternatively, those affinity products could be located farther apart to make the customer walk the length and breadth of the store, and thus be exposed to other products. Promotional discounted product bundles can be created to push a nonselling item along with a set of products that sell well together.

3. *Optimize logistics for seasonal effects*:

Seasonal products offer tremendously profitable short-term sales opportunities, yet they also offer the risk of unsold inventories at the end of the season. Understanding which products are in season in which market can help retailers dynamically manage prices to ensure their inventory is sold during the season. If it is raining in a certain area, then the inventory of umbrella and ponchos could be rapidly moved there from nonrainy areas to help increase sales.

4. *Minimize losses due to limited shelf life*:

Perishable goods offer challenges in terms of disposing off the inventory in time. By tracking sales trends, the perishable products at risk of not selling before the sellby date, can be suitably discounted and promoted.

v) Telecom

BI in telecom can help with the customer side as well as network side of the operations. Key BI applications include churn management, marketing/customer profiling, network failure, and fraud detection.

1. *Churn management*:

Telecom customers have shown a tendency to switch their providers in search for better deals. Telecom companies tend to respond with many incentives and discounts to hold on to customers. However, they need to determine which customers are at a real risk of switching and which others are just negotiating for a better deal. The level of risk should to be factored into the kind of deals and discounts that should be given. Millions of such customer calls happen every month. The telecom companies need to provide a consistent and databased way to predict the risk of the customer switching, and then make an operational decision in real time while the customer call is taking place. A decision-tree- or a neural network-based system can be used to guide the customer-service call operator to make the right decisions for the company, in a consistent manner.

2. *Marketing and product creation*:

In addition to customer data, telecom companies also store call detail records (CDRs), which can be analyzed to precisely describe the calling behavior of each customer. This unique data can be used to profile customers and then can be used for creating new products/services bundles for marketing purposes. An American telecom company, MCI, created a program called Friends & Family that allowed free calls with one's friends and family on that network, and thus, effectively locked many people into their network.

3. *Network failure management*:

Failure of telecom networks for technical failures or malicious attacks can have devastating impacts on people, businesses, and society. In telecom infrastructure, some equipment will likely fail with certain mean time between failures. Modeling the failure pattern of various components of the network can help with preventive maintenance and capacity planning.

4. *Fraud Management*:

There are many kinds of fraud in consumer transactions. Subscription fraud occurs when a customer opens an account with the intention of never paying for the services. Superimposition fraud involves illegitimate activity by a person other than the legitimate account holder. Decision

rules can be developed to analyze each CDR in real time to identify chances of fraud and take effective action.

4.Explain the applications of BI in

i) Banking

ii) Financial Services

iii)Insurance

iv)Manufacturing

v)Public Sector

i) Banking

Banks make loans and offer credit cards to millions of customers. They are most interested in improving the quality of loans and reducing bad debts. They also want to retain more good customers, and sell more services to them.

1. *Automate the loan application process*:

Decision models can be generated from past data that predict the likelihood of a loan proving successful. These can be inserted in business processes to automate the financial loan approval process.

2. *Detect fraudulent transactions*:

Billions of financial transactions happen around the world every day. Exception-seeking models can identify patterns of fraudulent transactions. For example, if money is being transferred to an unrelated account for the first time, it could be a fraudulent transaction.

3. *Maximize customer value (cross-, up-selling).*

Selling more products and services to existing customers is often the easiest way to increase revenue. A checking account customer in good standing could be offered home, auto, or educational loans on more favorable terms than other customers, and thus, the value generated from that customer could be increased.

4. *Optimize cash reserves with forecasting*.

Banks have to maintain certain liquidity to meet the needs of depositors who may like to withdraw money. Using past data and trend analysis, banks can forecast how much to keep and invest the rest to earn interest.

ii) Financial Services

Stock brokerages are an intensive user of BI systems. Fortunes can be made or lost based on access to accurate and timely information.

1. *Predict changes in bond and stock prices*:

Forecasting the price of stocks and bonds is a favorite pastime of financial experts as well as lay people. Stock transaction data from the past, along with other variables, can be used to predict future price patterns. This can help traders develop long term trading strategies.

2. *Assess the effect of events on market movements*.

Decision models using decision trees can be created to assess the impact of events on changes in market volume and prices. Monetary policy changes (such as Federal Reserve interest rate change) or geopolitical changes (such as war in a part of the world) can be factored into the predictive model to help take action with greater confidence and less risk.

3. *Identify and prevent fraudulent activities in trading*:

There have unfortunately been many cases of insider trading, leading to many prominent financial industry stalwarts going to jail. Fraud detection models seek out-of-the-ordinary activities, and help identify and flag fraudulent activity patterns.

iii) Insurance

This industry is a prolific user of prediction models in pricing insurance proposals and managing losses from claims against insured assets.

1. *Forecast claim costs for better business planning*:

When natural disasters, such as hurricanes and earthquakes strike, loss of life and property occurs. By using the best available data to model the likelihood (or risk) of such events happening, the insurer can plan for losses and manage resources and profits effectively.

2. *Determine optimal rate plans*:

Pricing an insurance rate plan requires covering the potential losses and making a profit. Insurers use actuary tables to project life spans and disease tables to project mortality rates, and thus price themselves competitively yet profitably.

3. *Optimize marketing to specific customers*:

By micro-segmenting potential customers, a data-savvy insurer can cherry pick the best customers and leave the less profitable customers to its competitors. Progressive Insurance is a US-based company that is known to actively use data mining to cherry pick customers and increase its profitability.

4. *Identify and prevent fraudulent claim activities*.

Patterns can be identified as to where and what kinds of fraud are more likely to occur. Decision-tree-based models can be used to identify and flag fraudulent claims.

iv) Manufacturing

Manufacturing operations are complex systems with inter-related subsystems. From machines working right, to workers having the right skills, to the right components arriving with the right quality at the right time, to money to source the components, many things have to go right. Toyota's famous lean manufacturing company works on just-in-time inventory systems to optimize investments in inventory and to improve flexibility in their product-mix.

1. *Discover novel patterns to improve product quality*:

Quality of a product can also be tracked, and this data can be used to create a predictive model of product quality deteriorating. Many companies, such as automobile companies, have to recall their products if they have found defects that have a public safety implication. Data mining can help with root cause analysis that can be used to identify sources of errors and help improve product quality in the future.

2. *Predict/prevent machinery failures*:

Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failures could be constructed using past data. Preventive maintenance can be planned, and manufacturing capacity can be adjusted, to account for such maintenance activities.

v) Public Sector

Government gathers a large amount of data by virtue of their regulatory function. That data could be analyzed for developing models of effective functioning. There are innumerable applications that can benefit from mining that data. A couple of sample applications are shown here.

1. *Law enforcement*:

Social behavior is a lot more patterned and predictable than one would imagine. For example, Los Angeles Police Department (LAPD) mined the data from its 13 million crime records over 80 years and developed models of what kind of crime going to happen when and where. By increasing patrolling in those particular areas, LAPD was able to reduce property crime by 27 percent. Internet chatter can be analyzed to learn of and prevent any evil designs.

2. *Scientific research*:

Any large collection of research data is amenable to being mined for patterns and insights. Protein folding (microbiology), nuclear reaction analysis (sub-atomic physics), disease control (public health) are some examples where data mining can yield powerful new insights.

5. Define data warehouse and write the design considerations for DW.

- 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.DW supports business reporting and data mining activities.
- 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.

6. Explain DW architecture with a neat diagram.

DW has four key elements (Figure 3.1).

- 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.

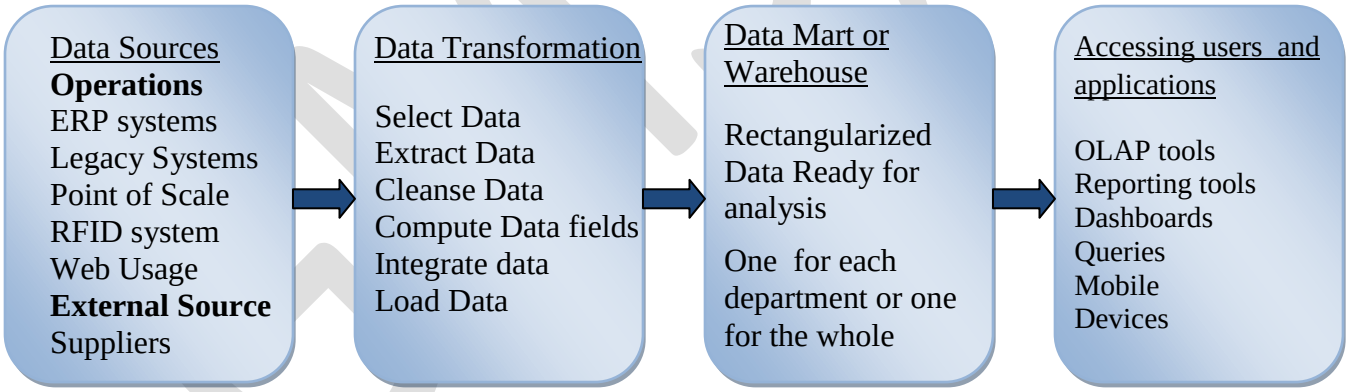


Figure 3.1: Data Warehousing Architecture

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.

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 (Figure 3.2).

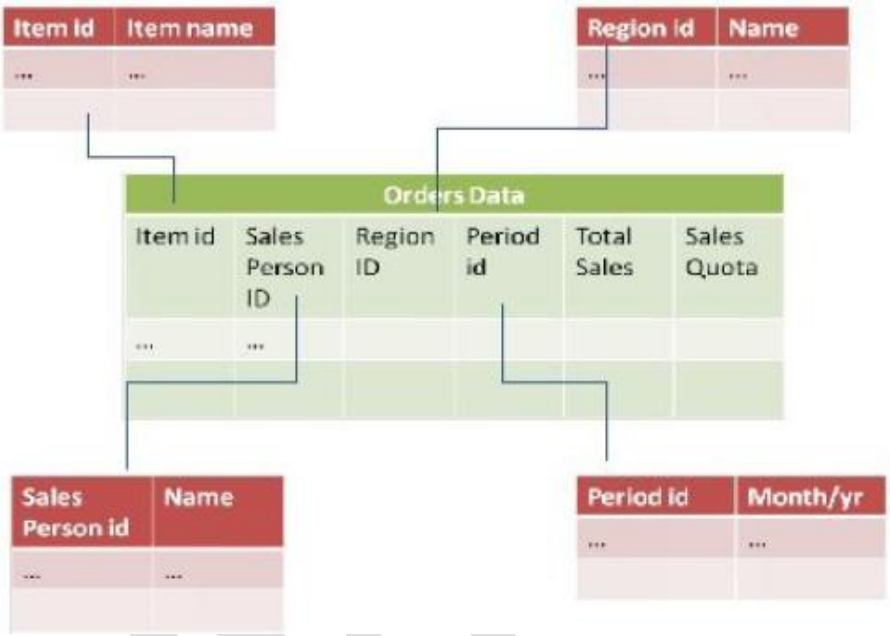


Figure 3.2: Star Schema Architecture for DW

- 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.

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.

7. Explain DW development approaches

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 (Table 3.1).

	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

8. Define Data mining and explain the phases in 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.
- 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.

Phases in Data mining are:

i) Gathering and selecting data

ii) Data cleansing and preparation

iii) Outputs of Data Mining

iv) Evaluating Data Mining Results

i)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 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.
- Gathering and selecting 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.

- 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.

ii) 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 de-duped.

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 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.

iii) 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 business rules, 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 "vehicle 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).

iv)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).
- Thus the predictive accuracy can be specified by the following formula.

$$Predictive\ Accuracy\ = \frac{(TP\ + TN)}{(TP\ +\ TN\ +\ FP\ +\ FN)}$$

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

- All classification techniques have a predictive accuracy associated with a predictive model. The highest value can be 100%.
- In practice, predictive models with more than 70% accuracy can be considered usable in business domains, depending upon the nature of the business.
- There are no good objective measures to judge the accuracy of unsupervised learning techniques such as Cluster Analysis. There is no single right answer for the results of these techniques. For example, the value of the segmentation model depends upon the value the decision-maker sees in those results.

9. Explain 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 (Figure 4.2).
- 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.

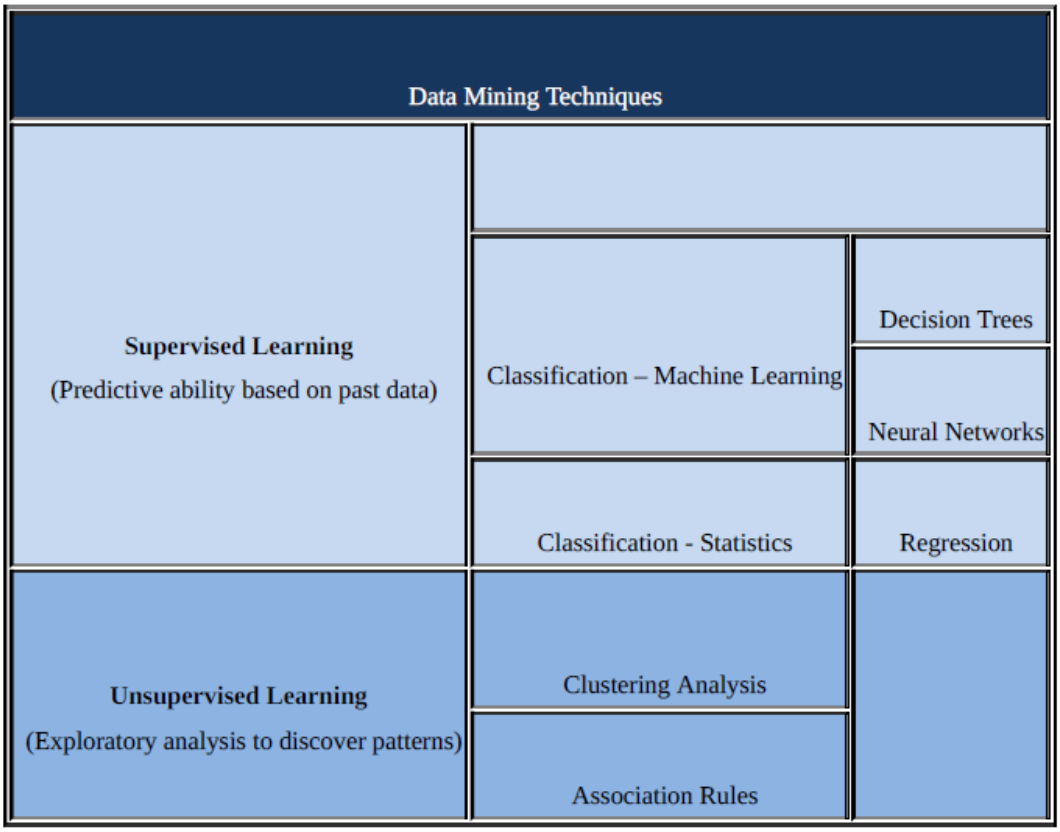


Figure 4.2: Important Data Mining Techniques

Decision trees :

These 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

- This 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.
- At some point, the neural network will have learned enough and begin to match the predictive accuracy of a human expert or alternative classification techniques. The predictions of some ANNs that have been trained over a long period of time with a large amount of data have become decisively more accurate than human experts.
- At that point, the ANNs can begin to be seriously considered for deployment, in real situations in real time. ANNs are popular because they are eventually able to reach a high predictive accuracy. ANNs are also relatively simple to implement and do not have any issues with data quality. However, ANNs require a lot of data to train it to develop good predictive ability.

Cluster Analysis

- It 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

- These 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 $X \rightarrow Y$, 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.

10. Explain CRISP-DM Data Mining cycle with a neat diagram

The Data Mining industry has proposed a Cross-Industry Standard Process for Data Mining (CRISP-DM). It has six essential steps (Figure 4.3):

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.

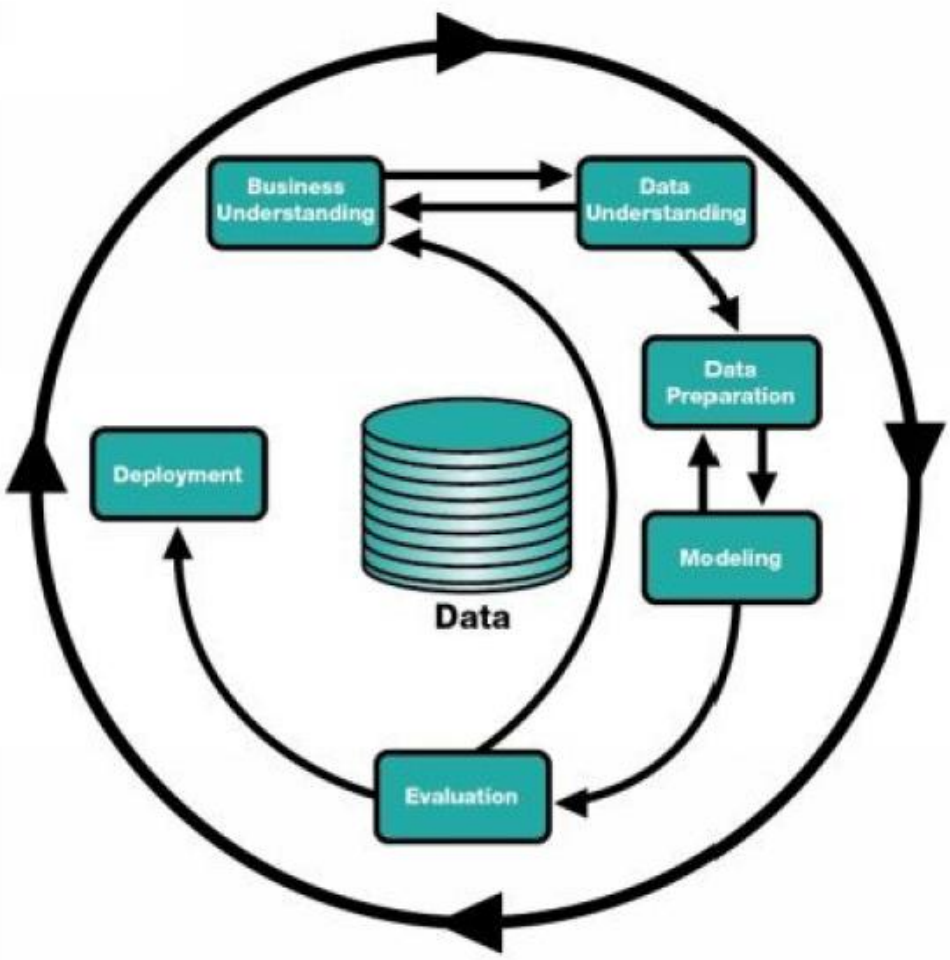


Figure 4.3: CRISP-DM Data Mining cycle

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.

6. *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.

11. What are the major mistakes to be avoided when doing data mining?

- Data mining is an exercise in extracting non-trivial useful patterns in the data. It requires a lot of preparation and patience to pursue the many leads that data may provide.
- Much domain knowledge, tools and skill is required to find such patterns. Here are some of the more common mistakes in doing data mining, and should be avoided.

  Mistake #1*: *Selecting the wrong problem for data mining

  Without the right goals or having no goals, data mining leads to a waste of time. Getting the right answer to an irrelevant question could be interesting, but it would be pointless from a

business perspective. A good goal would be one that would deliver a good ROI to the organization.

***Mistake #2: Buried under mountains of data without clear metadata*:**

It is more important to be engaged with the data, than to have lots of data. The relevant data required may be much less than initially thought. There may be insufficient knowledge about the data, or metadata. Examine the data with a critical eye and do not naively believe everything you are told about the data.

Mistake #3: Disorganized data mining:

Without clear goals, much time is wasted. Doing the same tests using the same mining algorithms repeatedly and blindly, without thinking about the next stage, without a plan, would lead to wasted time and energy. This can come from being sloppy about keeping track of the data mining procedure and results. Not leaving sufficient time for data acquisition, selection and preparation can lead to data quality issues, and GIGO. Similarly not providing enough time for testing the model, training the users and deploying the system can make the project a failure.

Mistake #4: Insufficient business knowledge:

Without a deep understanding of the business domain, the results would be gibberish and meaningless. Don't make erroneous assumptions, courtesy of experts. Don't rule out anything when observing data analysis results. Don't ignore suspicious (good or bad) findings and quickly move on. Be open to surprises. Even when insights emerge at one level, it is important to slice and dice the data at other levels to see if more powerful insights can be extracted.

Mistake #5: Incompatibility of data mining tools and datasets.

All the tools from data gathering, preparation, mining, and visualization, should work together. Use tools that can work with data from multiple sources in multiple industry standard formats.

Mistake #6: Looking only at aggregated results and not at individual records/predictions.

It is possible that the right results at the aggregate level provide absurd conclusions at an individual record level. Diving into the data at the right angle can yield insights at many levels of data.

Mistake #7: Not measuring your results differently from the way your sponsor measures them.

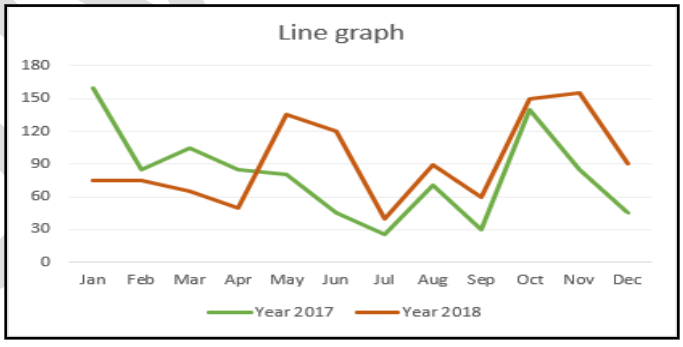
If the data mining team loses its sense of business objectives, and beginning to mine data for its own sake, it will lose respect and executive support very quickly

12. Define data visualization and explain the types of charts.

- Data Visualization is the art and science of making data easy to understand and consume, for the end user. Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information.
- The right visualization requires an understanding of the consumer's needs, nature of the data, and the many tools and techniques available to present data. The right visualization arises from a complete understanding of the totality of the situation. One should use visuals to tell a true, complete and fast-paced story.
- Data visualization 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.
- Some of the popular chart types and their usage:

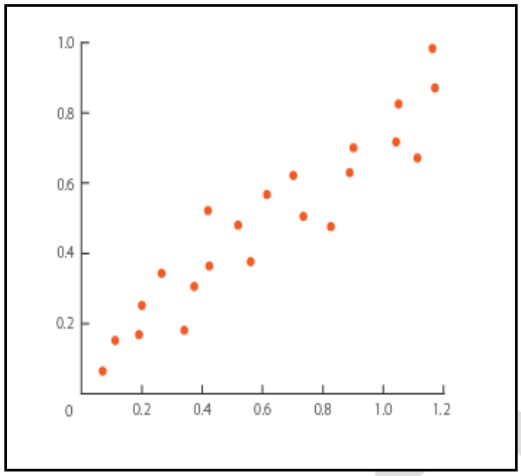
1. *Line graph.*

This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.



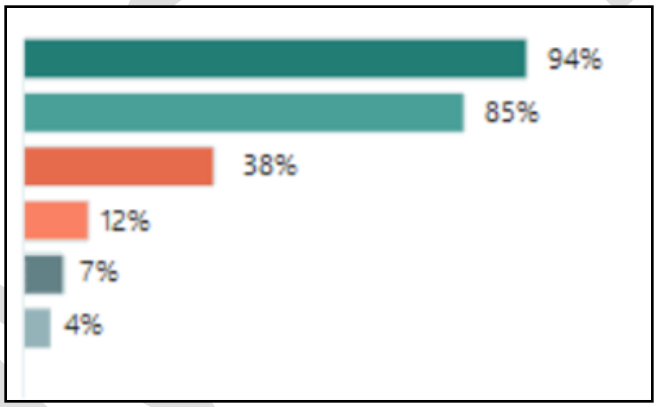
2. *Scatter plot:*

This is another very basic and useful graphic form. It helps reveal the relationship between two variables. In the above caselet, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.



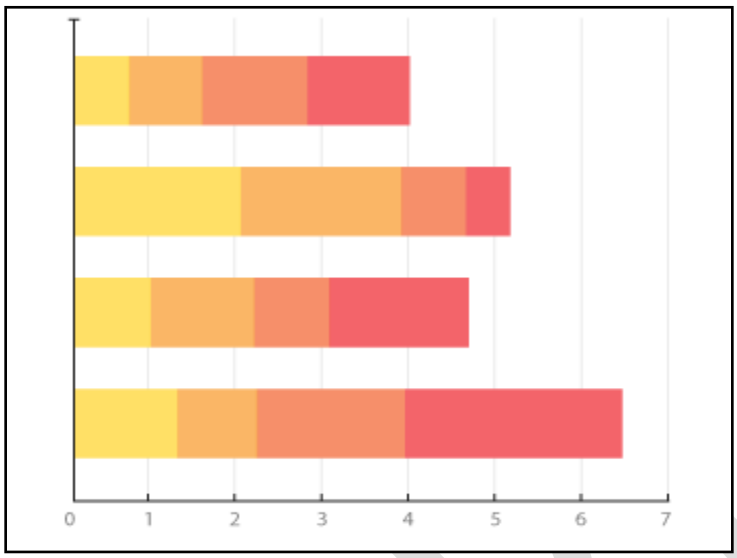
3. *Bar graph:*

A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.



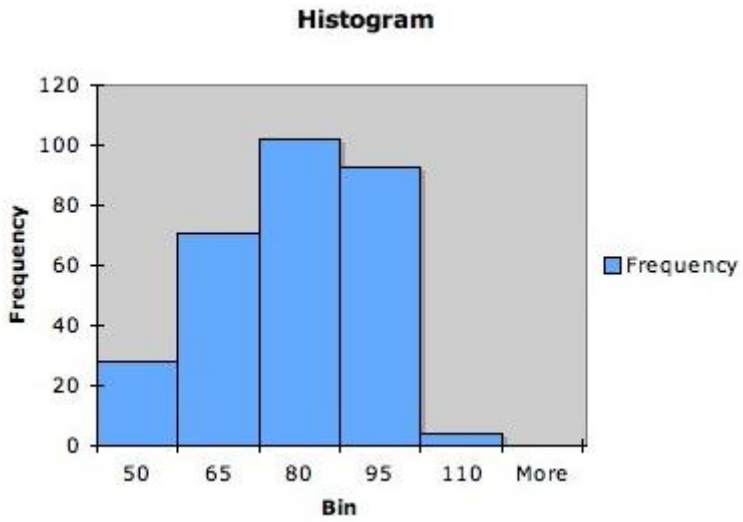
4. Stacked Bar graphs:

These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.



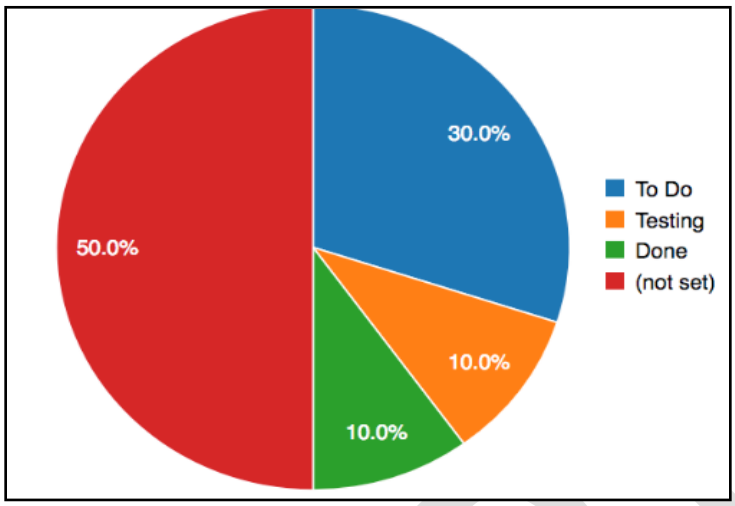
5. *Histograms*:

These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.



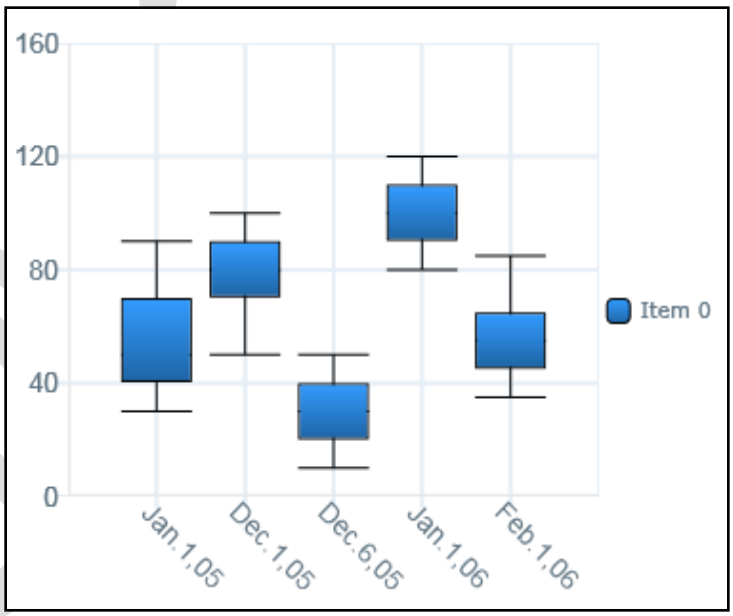
***6.Pie charts*:**

These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.



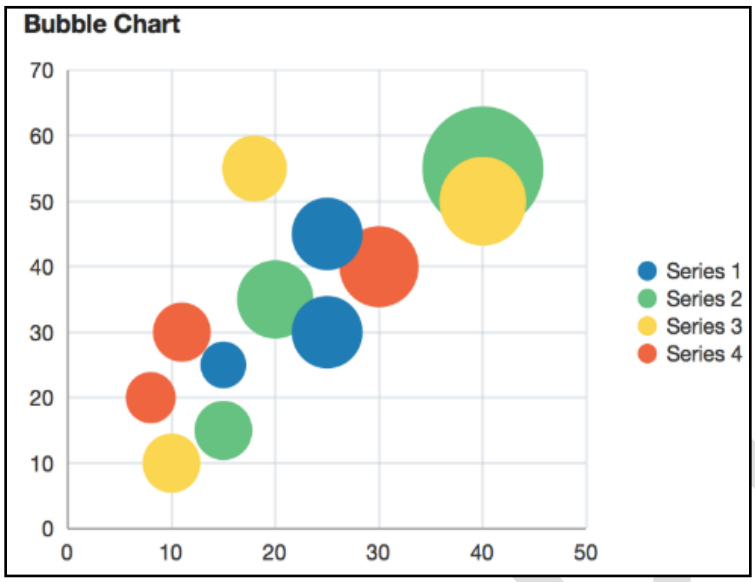
7. *Box charts:*

These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.



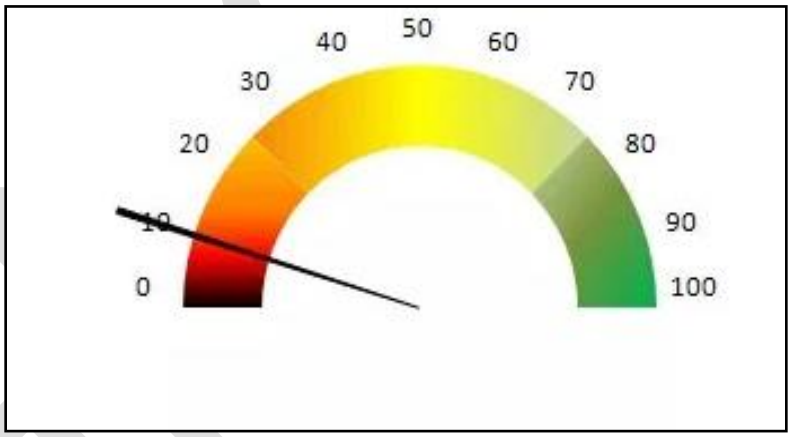
8. *Bubble Graph*:

This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) … the size of the circle and the color fill in the circle could represent two additional dimensions.



9. *Dials*:

These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and green to give an instant view of the data.



10. *Geographical*

Data maps are particularly useful maps to denote statistics. Figure 5.3 shows a tweet density map of the US. It shows where the tweets emerge from in the US.

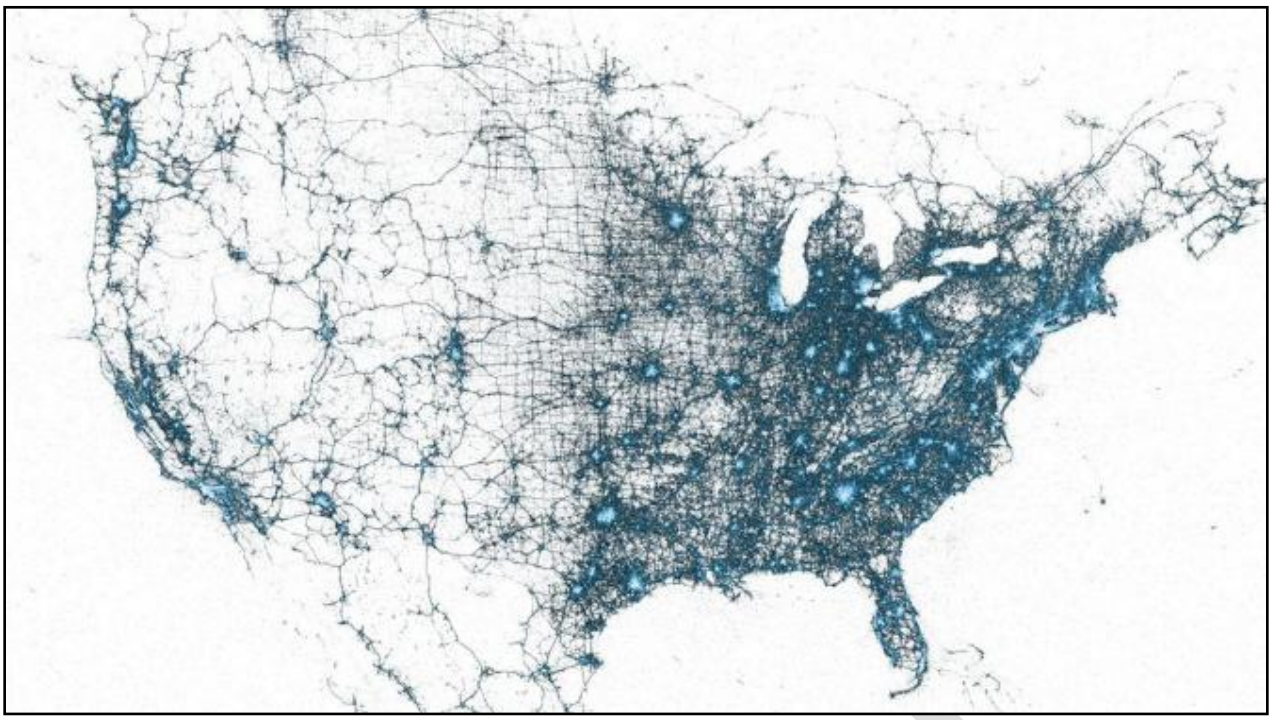


Figure 5.3: US tweet map

11. *Pictographs*:

One can use pictures to represent data. Fig 5.4 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.

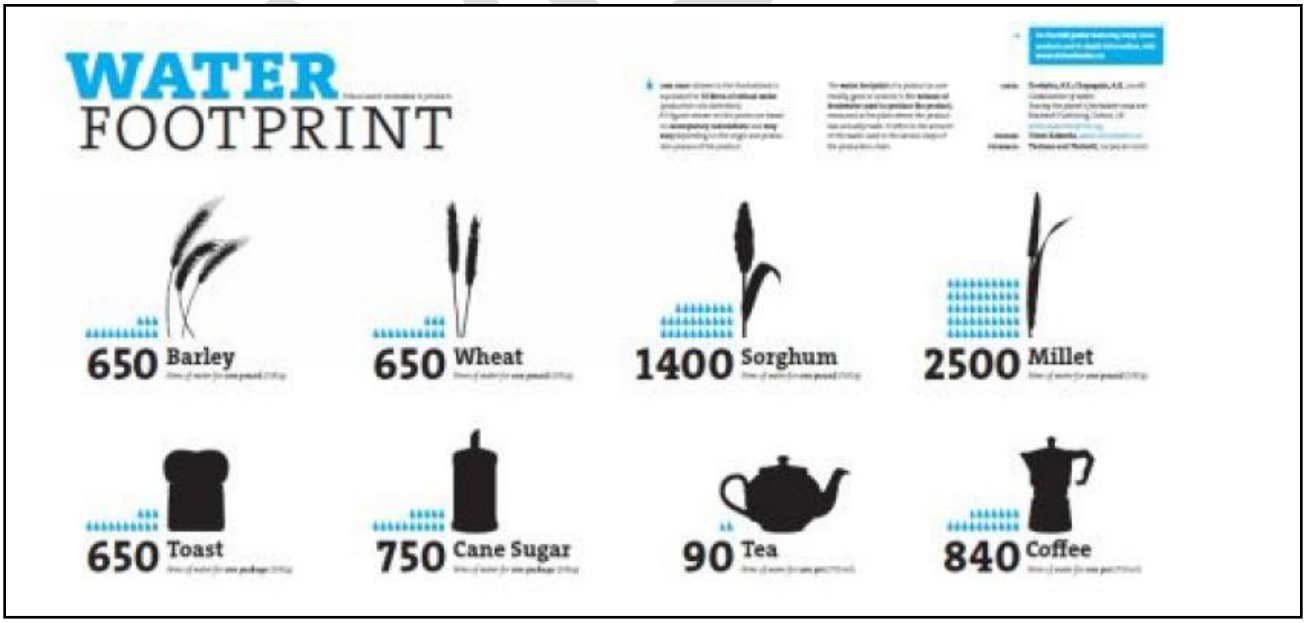


Figure 5.4: Pictograph of Water footprint