EMERGING AREAS OF ANALYTICS

- ***Data Science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms.***
- ***Given a problem, the objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that can be used.***
- For example, Target's pregnancy prediction is a classification problem in which customers (or entities) are classified into different groups.
- In the case of pregnancy test, the classes are: 1. Pregnant 2. Not pregnant There are several techniques available for solving classification problems such as logistic regression, classification trees, random forest, adaptive boosting, neural networks, and so on.
- ***The objective of the data science component is to identify the technique that is best based on a measure of accuracy. Usually, several models are developed for solving the problem using different techniques and a few models may be chosen for deployment of the solution.***
- ***Business analytics can be grouped into three types: descriptive analytics, predictive analytics, and prescriptive analytics.***

In God we trust; all others must bring Data —Edwards Deming

- The epigraph captures the importance of analytics and data-driven decision making in one sentence. During the early period of the 20th century, many companies were taking business decisions based on 'opinions' rather than decisions based on proper data analysis (which probably acted as a trigger for Deming's quote).
- Opinion-based decision making can be very risky and often leads to incorrect decisions. ***One of the primary objectives of business analytics is to improve the quality of decision-making using data analysis***.
- Every organization across the world uses performance measures such as market share, profitability, sales growth, return on investment (ROI), customer satisfaction, and so on for quantifying, monitoring, benchmarking, and improving its performance.
- It is important for organizations to understand the association between key performance indicators (KPIs) and factors that have a significant impact on the KPIs for effective management. ***Knowledge of the relationship between KPIs and factors would provide the decision maker with appropriate actionable items.***
- ***Analytics is a body of knowledge consisting of statistical, mathematical, and operations research techniques; artificial intelligence techniques such as machine learning and deep learning algorithms; data collection and storage; data management processes such as data extraction, transformation, and loading (ETL); and computing and big data technologies such as Hadoop, Spark, and Hive that create value by developing actionable items from data.*** Two primary macro-level objectives of analytics are problem solving and decision making.
- Analytics helps organizations to create value by solving problems effectively and assisting in decision making. Today, analytics is used as a competitive strategy by many organizations such as Amazon, Apple, General Electric, Google, Facebook and

Procter and Gamble who use analytics to create products and solutions. Marshall (2016) and MacKenzie et al. (2013) reported that Amazon's recommender systems resulted in a sales increase of 35%.

- Davenport and Harris (2007) and Hopkins et al. (2010) reported that there was a high correlation between use of analytics and business performance. They claimed that the majority of high performers (measured in terms of profit, shareholder return and revenue, etc.) strategically apply analytics in their daily operations, as compared to low performers.

A few of the problems that e-commerce companies such as Amazon and Flipkart try to address are as follows:

1. Forecasting demand for products directly sold by the company; excess inventory and shortage can impact both the top line and the bottom line.
2. Cancellation of orders placed by customers before their delivery. Ability to predict cancellations and intervention can save cost incurred on unnecessary logistics.
3. Fraudulent transactions resulting in financial loss to the company.
4. Predicting delivery time since it is an important service level agreement from the customer perspective.
5. Predicting what a customer is likely to buy in future to create recommender systems.

- Given the scale of operations of modern companies, it is almost impossible to manage them effectively without analytics. Although decisions are occasionally made using the HiPPO algorithm ("highest paid person's opinion" algorithm), especially in a group decision-making scenario, there is a significant change in the form of "data-driven decision making" among several companies.
- Many companies use analytics as a competitive strategy and many more are likely to follow. A typical data-driven decision-making process uses the following steps (Figure 1.1):

1. Identify the problem or opportunity for value creation.
2. Identify sources of data (primary as well secondary data sources).
3. Pre-process the data for issues such as missing and incorrect data. Generate derived variables and transform the data if necessary. Prepare the data for analytics model building.
4. Divide the data sets into subsets training and validation data sets.
5. Build analytical models and identify the best model(s) using model performance in validation data.
6. Implement Solution/Decision/Develop Product.

- ***Analytics is used to solve a wide range of problems starting with simple process improvement such as reducing procurement cycle time to complex decision-making problems such as farm advisory systems that involve accurate weather prediction, forecasting commodity price etc, so that farmers can be advised about crop selection, crop rotation, etc***.

- Figure 1.2 shows the pyramid of analytics applications, at the bottom of the pyramid analytics is used for process improvement and at the top it is used for decision making and as a competitive strategy.

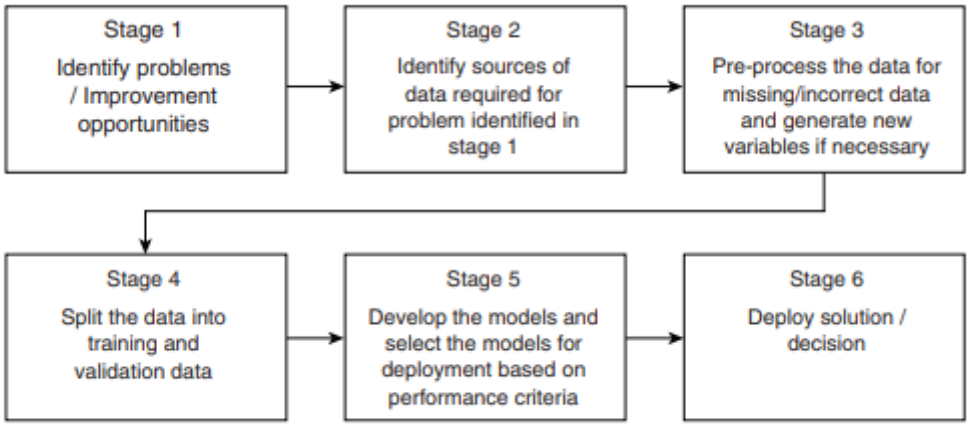

FIGURE 1.1 Business analytics – Data-driven decision-making flow diagram.

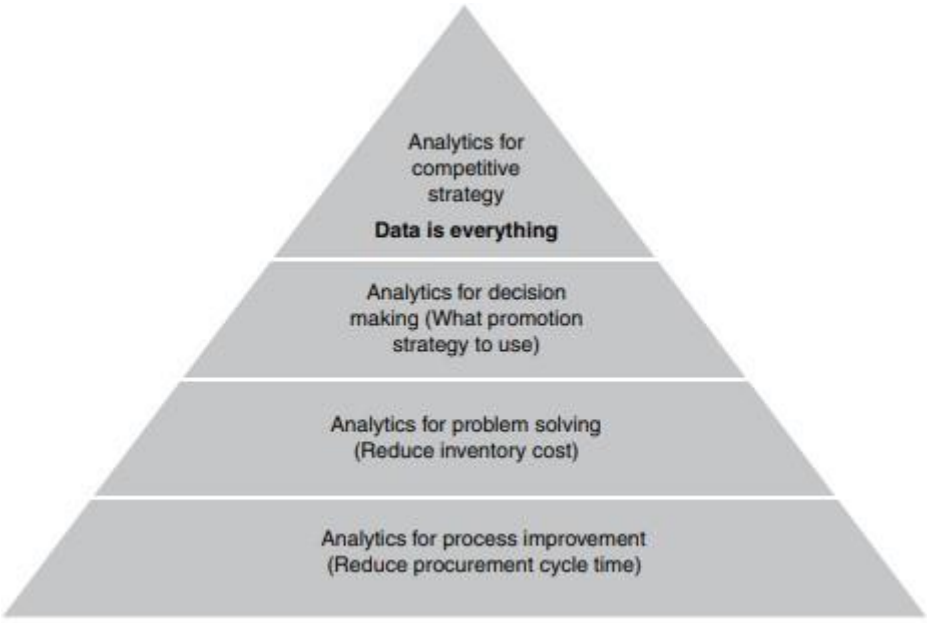

FIGURE 1.2 Pyramid of analytics.

WHY ANALYTICS

- According to the theory of firm (Coase, 1937 and Fame, 1980) as proposed by several economists, firms exist to minimize the transaction cost. Transactions take place when goods or services are transferred to customers from the supplier.
- The cost of decision making is an important element of transaction cost. Michalos (1970) groups the costs of decision making into three categories:
  1. Cost of reaching a decision with the help of a decision maker or procedure; this is also known as production cost, that is, cost of producing a decision.

2. Cost of actions based on decisions produced; also known as implementation cost.
3. Failure costs that account for failure of an organization's efforts on production and implementation.

- For example, consider a firm that would like to sell product such as a ready made shirt. The firm has to take several decisions such as fabric, colour, size, fit, price, promotion, and so on.
- Each of these attributes has several options. The real problem starts with decision-making ability of firms, especially the techniques and processes used in decision making; unfortunately human beings are inherently not good at decision making.

A great example for human's inability to take decisions is the famous Monty Hall problem (Savant, 1990) in which the contestants of a game show are shown three doors (Figure 1.3).

- Behind one of the doors is an expensive item (such as a car or gold); while there are inexpensive items behind the remaining two doors (such as a goat).
- The contestant is asked to choose one of the doors. Assume that the contestant chooses door 1; the game host would then open one of the remaining two doors.
- Assume that the game host opens door 2, which has a goat behind it. Now the contestant is given a chance to change his initial choice (from door 1 to door 3).
- The problem is whether or not the contestant should change his/her initial choice. Note that the contestant is given an option to switch door irrespective of the item behind his/her original choice of door.
- The problem is based on a famous television show "Let's make a deal" hosted by Monty Hall in 1960s and 1970s (Selvin, 1975).
- In this problem, the contestant — the decision maker — has two choices: he/she can either change his/her initial choice or stick with his/her initial choice.
- When Marilyn vos Savant, a columnist at the Parade Magazine, posted that the contestant should change the initial choice (Savant, 1990), 92% of the general public and 65% of the university graduates (many of them with PhDs) who responded to her column were against her answer.1 Although Marilyn vos Savant provided a simple decision tree

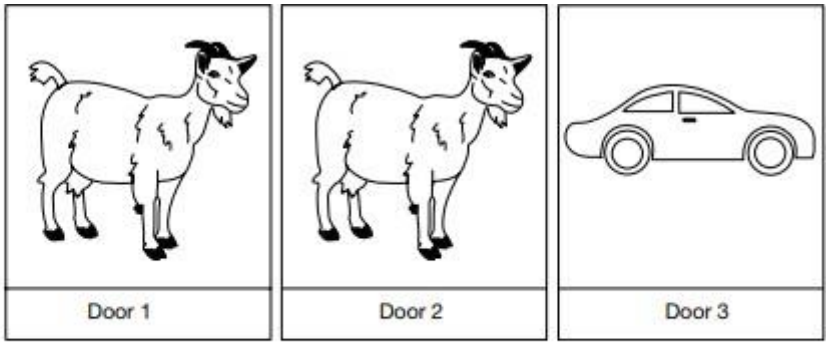

FIGURE 1.3 Monty Hall problem.

[1] Source: http://en.wikipedia.org/wiki/Monty_Hall_problem

argument to prove that the probability of winning increases to 2/3 when the contestant changes his/ her initial choice, many scholars did not accept her argument that changing the initial option is the right decision.

- Table 1.1 shows why changing the initial option increases the probability of winning. The expensive item can be behind any one of the three doors as shown in Table 1.1 (rows 2−4).
- Assume that the contestant has chosen door 1 initially, columns 4 and 5 (last row) give the probability of winning the car if contestant stays with door 1 (column 4) and the door 1 is changed (column 5), respectively.
- The above argument can be extended to any number of doors without loss of generality. In the case of Monty Hall problem, the number of alternatives available to the player is just two. Even when the number of options is only 2, many find it difficult to comprehend that changing the initial choice will increase the probability of winning

TABLE 1.1 Monty Hall problem final probability of win when the player changes the initial choice

Item Behind Door 1	Item Behind Door 2	Item Behind Door 3	Result if Stayed with Door #1	Result if the Door is Changed
Car	Goat	Goat	Car	Goat
Goat	Car	Goat	Goat	Car
Goat	Goat	Car	Goat	Car
Probability of Winning			1/3	2/3

Business analytics is a set of statistical and operations research techniques, artificial intelligence, information technology and management strategies used for framing a business problem, collecting data, and analysing the data to create value to organizations.

Business Analytics can be broken into 3 components:

1. Business Context

2. Technology

3. Data Science

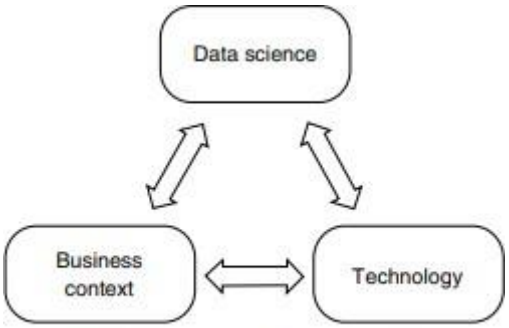

FIGURE 1.4 Components of business analytics.

DESCRIPTIVE ANALYTICS "If the statistics are boring, then you've got the wrong numbers". —Edward R. Tufte

- ***Descriptive analytics is the simplest form of analytics that mainly uses simple descriptive statistics, data visualization techniques, and business related queries to understand past data.***
- ***One of the primary objectives of descriptive analytics is innovative ways of data summarization. Descriptive analytics is used for understanding the trends in past data which can be useful for generating insights***.
- Figure 1.5 shows visualization of relationship break-ups reported in Facebook. It is clear from Figure 1.5 that spike in breakups occurred during spring break and in December before Christmas. There could be many reasons for increase in breakups during December (we hope it is

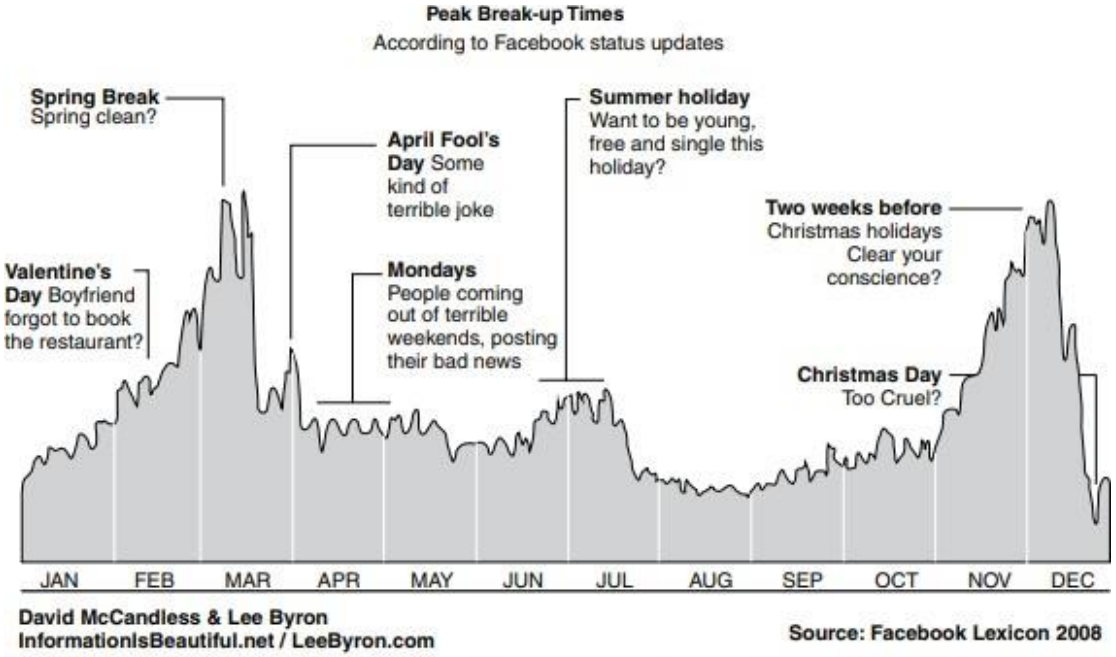

FIGURE 1.5 Peak breakup times according to Facebook status update. *Source*: David McCandless and Lee Bryon.

not a New Year resolution that they would like to change the partner). Many believe that since December is a holiday season, couples get a lot of time to talk to each other, probably that is where the problem starts.

- However, descriptive analytics is not about why a pattern exists, but about what the pattern means for a business.
- The fact that there is a significant increase in breakups during December we can deduce the following insights (or possibilities):
  1. There will be more traffic to online dating sites during December/January.
  2. There will be greater demand for relationship counsellors and lawyers.
  3. There will be greater demand for housing and the housing prices are likely to increase in December/January.
  4. There will be greater demand for household items.

5. People would like to forget the past, so they might change the brand of beer they drink.

Descriptive analytics using visualization identifies trends in the data and connects the dots to gain insights about associated businesses.

- In addition to visualization, descriptive analytics uses descriptive statistics and queries to gain insights from the data.
- The following are a few examples of insights obtained using descriptive analytics reported in literature:
  1. Most shoppers turn towards the right side when they enter a retail store (Underhill, 2009, pages 77−79). Retailers keep products with higher profit on the right side of the store since most people turn right.
  2. Married men who kiss their wife before going to work live longer, earn more and get into less number of accidents as compared to those who do not (Foer, 2006).
  3. Correlated with Facebook relationship breakups, divorces spike in January. According to Caroline Kent (2015), January 3 is nicknamed 'divorce day'.
  4. Men are more reluctant to use coupons as compared to women (Hu and Jasper, 2004). While sending coupons, retailers should target female shoppers as they are more likely to use coupons. Trends obtained through descriptive analytics can be used to derive actionable items

PREDICTIVE ANALYTICS

If you torture the data long enough, it will confess. —Ronald Coase

- ***In the analytics capability maturity model (ACMM), predictive analytics comes after descriptive analytics and is the most important analytics capability.***
- ***It aims to predict the probability of occurrence of a future event such as forecasting demand for products/services, customer churn, employee attrition, loan defaults, fraudulent transactions, insurance claim, and stock market fluctuations.***
- ***While descriptive analytics is used for finding what has happened in the past, predictive analytics is used for predicting what is likely to happen in the future.***
- The ability to predict a future event such as an economic slowdown, a sudden surge or decline in a commodity's price, which customer is likely to churn, what will be the total claim from auto insurance customer, how long a patient is likely to stay in the hospital, and so on will help organizations plan their future course of action.
- ***Anecdotal evidence suggests that predictive analytics is the most frequently used type of analytics across several industries.***
- The reason for this is that almost every organization would like to forecast the demand for the products that they sell, prices of the materials used by them, and so on.
- Irrespective of the type of business, organizations would like to forecast the demand for their products or services and understand the causes of demand fluctuations.
- ***The use of predictive analytics can reveal relationships that were previously unknown and are not intuitive. The most popular example of the application of predictive analytics is Target's pregnancy prediction model***.

- In 2002, Target hired statistician Andrew Pole; one of his assignments was to predict whether a customer is pregnant (Duhigg, 2012). At the outset, the question posed by the marketing department to Pole may look bizarre, but it made great business sense.
- Any marketer would like to identify the price-insensitive customers among the shoppers, and who can beat soon-to-be parents? A list of interesting applications of predictive analytics is presented in Table 1.2

TABLE 1.2 List of predictive analytics applications

Organization	Predictive Analytics Model
Polyphonic HMI	Predicts whether a song will be a hit using machine learning algorithms. Their product 'Hit Song Science' uses mathematical and statistical techniques to predict the success of a song on a scale of 1 to 10 (Anon, 2003).
Okcupid	Predicts which online dating message is likely to get a response from the opposite sex (Siegel, 2013).
Amazon.com	Uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie *et al.*, 2013).
Hewlett Packard (HP)	Developed a flight risk score for its employees to predict who is likely to leave the company (Siegel, 2013).
University of Maryland	Claimed that dreams can predict whether one's spouse will cheat (Whitelocks, 2013).
Flight Caster	Predicts flight delays 6 hours before the airline's alerts.
Netflix	Predicts which movie their customer is likely to watch next (Greene, 2006). 75% of what customer watch at Netflix is from product recommendations (MacKinzie *et al.*, 2013).
Capital One Bank	Predicts the most profitable customer (Davenport, 2007).
Google	Predicted the spread of H1N1 flu using the query terms (Carneiro and Mylonakis, 2010).
Farecast	Developed a model to predict airfare, whether it is likely to increase or decrease, and the amount of increase/decrease.[4]

[4] Source: http://www.crunchbase.com/company/farecast

PRESCRIPTIVE ANALYTICS

Every decision has a consequence. —Damon Darrel

- ***Prescriptive analytics is the highest level of analytics capability which is used for choosing optimal actions once an organization gains insights through descriptive and predictive analytics.***
- In many cases, prescriptive analytics is solved as a separate optimization problem. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives.
- ***Operations Research (OR) techniques form the core of prescriptive analytics. Apart from operations research techniques, machine learning algorithms, metaheuristics, and advanced statistical models are used in prescriptive analytics.***
- Note that actionable items can be derived directly after descriptive and predictive analytics model development; however, they may not be the optimal action.

- For example, in a Business to Business (B to B) sales, the proportion of sales conversions to sales leads could be very low. The sales conversion period could be very long, as high as 6 months to one year.
- ***Predictive analytics such as logistics regression can be used for predicting the propensity to buy a product and actionable items (such as which customer to target) can be derived directly based on predicted probability to buy or using lift chart.***
- However, the values of the sale are likely to be different, as are the profits earned from different customers. Thus, targeting customers purely based on probability to buy may not result in an optimal solution.
- Use of techniques such as binary integer programming will result in optimal targeting of customers that maximize total expected profit. That is, while actionable items can be derived from descriptive and predictive analytics, use of prescriptive analytics ensures optimal actions (choices or alternatives).
- The link between different analytics capability is shown in Figure 1.7. Ever since their introduction during World War II, OR models have been used in every sector of every industry.
- The list of prescriptive analytics applications is long and several companies across the world have benefitted from the use of prescriptive analytics tools. Coca-Cola Enterprises (CCE) is the largest distributor of Coca-Cola products. In 2005, CCE distributed 2 billion physical cases

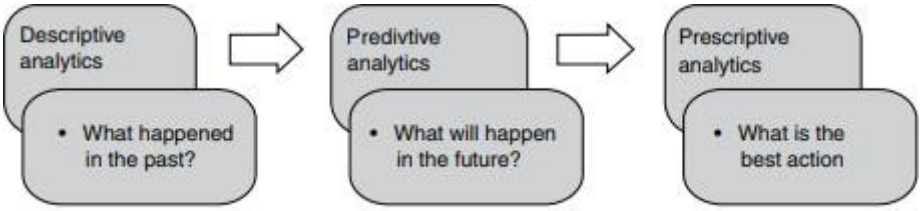

FIGURE 1.7 Link between different analytics capabilities.

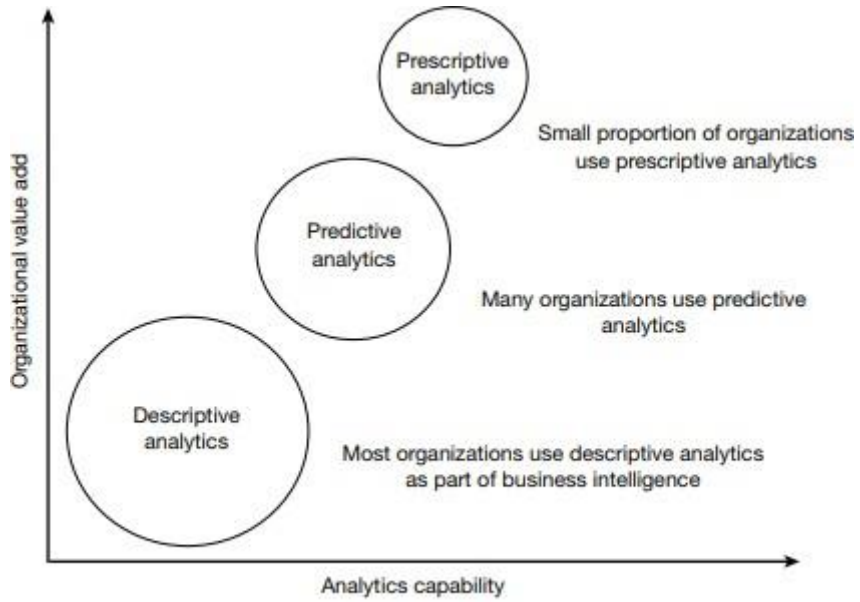

FIGURE 1.8 Analytics capability versus organizational value add.

DESCRIPTIVE, PREDICTIVE, AND PRESCRIPTIVE ANALYTICS TECHNIQUES

- ***The most frequently used predictive analytics techniques are regression, logistic regression, classification trees, forecasting, K-nearest neighbours, Markov chains, random forest, boosting, and neural networks.***
- ***The frequently used tools in prescriptive analytics are linear programming, integer programming, multi-criteria decision-making models such as goal programming and analytic hierarchy process, combinatorial optimization, non-linear programming, and meta-heuristics.***
- In Table 1.3, we provide a brief description of some of these tools and the problems that are solved using these tools. We have highlighted a few tools that are frequently used by analytics companies.

TABLE 1.3 Predictive and prescriptive analytics techniques

Analytics Techniques	Applications
Regression	Regression is the most frequently used predictive analytics tool. It is a supervised learning algorithm. In management and social sciences, almost all hypotheses are validated using regression models. In business, irrespective of the sector, the decision maker would like to know how the key performance indicators (KPIs) of the business are related to macro-economic parameters and other internal process parameters. Regression is an excellent tool for establishing the existence of an association relationship between a response variable (KPI) and other explanatory variables. Unfortunately, regression is one of the most highly misused techniques in analytics.
Logistic and Multinomial Regression	Logistic and multinomial logistic regression techniques are used to find the probability of occurrence of an event. Logistic regression is a supervised learning algorithm. Logistic regression is used for solving classification and discrete choice problems. Classification problems are common in many businesses. For example, banks and financial institutions would like to classify their customers into several risk categories. Companies would like to predict which customer is highly likely to churn in the next quarter. Marketers would like to know which brand a customer is likely to buy and whether promotions can make a customer change his/her brand loyalty. Credit scoring and fraud detection are other popular applications of logistic regression.
Decision Trees	Decision trees or classification trees are usually used for solving classification problems. There are several types of classification tree models. Chi-Squared Automatic Interaction Detection (CHAID) and Classification Trees (CART) are frequently used for solving classification problems. Although the decision trees are usually used for solving classification problems (in which the outcome variable is discrete), they can also be used when the outcome variable is continuous.
Markov Chains	Olle Haggstrom (2007) wrote an article stating that problem solving is often a matter of cooking up an appropriate Markov chain. One of the initial applications of Markov chains was implemented by the American retail giant Sears. They used a Markov Decision Process to decide the optimal mailing policy for their catalogues (Howard, 2002). Today, Markov chains are one of the key analytics tools in marketing, finance, operations, and supply chain management.
Random Forest	Random forest is one of the popular machine learning algorithms that uses ensemble approach to solve the problem by generating a large number of models.
Linear Programming	Since its origins during World War II, linear programming is one of the most frequently used techniques in prescriptive analytics. Problems such as resource allocation, product mix, cutting-stock problem, revenue management, and logistics optimisation are frequently solved using linear programming.
Integer Programming	Many optimization problems in real life may have variables that can take only integer values. When one or more variables in the problem can take only an integer solution, the model is called an integer programming model. Capital budgeting, scheduling, and set covering are a few problems that are solved using integer programming.

Analytics Techniques	Applications
Multi-Criteria Decision-Making Model	In many cases, the decision makers may have more than one objective (or KPIs). For example, a company may like to increase the profit as well as the market share. It is possible that the various objectives identified by the organization may conflict with one another. In such cases, techniques such as Analytic Hierarchy Process and Goal Programming are used to arrive at the optimal decisions.
Combinatorial Optimisation	Combinatorial optimization involves choosing the optimal solution from a large number of finite solutions. The travelling salesman problem (TSP), the vehicle routing problem (VRP), and the minimum spanning tree problem (MST) belongs to this category. Many industry problems are analogical to TSP, VRP, and MST.
Non-Linear Programming (NLP)	Large classes of problems faced by the industry have non-linear objective functions and/or non-linear constraints. Many engineering design optimization problems belongs to this category. NLP are also difficult set of problems to solve due to limitations of the existing algorithms. NLP is an integral part of several machine learning algorithms such as neural networks. The loss function which is used for finding weights for input variables is a non-linear function.
Six Sigma	Six Sigma and its problem-solving methodology DMAIC (Define, Measure, Analyse, Improve, and Control) are frequently used in process improvement problems.
Social Media Analytics Tools	Social media analytics is a collection of tools and techniques used for analysing unstructured data such as texts, videos, photos, and so on. With the exponential growth in the use of social media by the general public, tools designed for analysing unstructured data will be frequently used by organizations.

BIG DATA ANALYTICS

The world is one big data problem. —Andrew McAfee

- Big data is a class of problems that challenge existing IT and computing technology and existing algorithms. Traditionally, big data is defined as a big volume of data (in excess of 1 terabyte) generated at high velocity with high variety and veracity.
- That is, ***big data is identified using 4 Vs, namely, volume, velocity, variety, and veracity*** which are defined as follows:
    1. ***Volume is the size of the data that an organization holds. Typically, this can run into several petabytes (1015 bytes) or exabytes (1016 bytes).*** Organizations such as telecom and banking collect and store a large quantity of customer data. ***Data collected using satellite and other machine generated data such as data generated by health and usage monitoring systems fitted in aircrafts, weather and rain monitoring systems can run into several exabytes since the data is captured minute by minute.***
    2. 2. ***Velocity is the rate at which the data is generated. For example, AT&T customers generated more than 82 petabytes of data traffic on a daily basis (Anon, 2016).***
    3. ***3. Variety refers to the different types of data collected.***
    In the case of telecom, the different data types are voice calls, messages in different languages, video calls, use of Apps, etc. TABLE 1.3 (Continued) Analytics Techniques Applications Multi-Criteria Decision-Making Model In many cases, the decision makers may have more than one objective (or KPIs).
        - For example, a company may like to increase the profit as well as the market share. It is possible that the various objectives identified by the organization may conflict with one another. In such cases, techniques

such as Analytic Hierarchy Process and Goal Programming are used to arrive at the optimal decisions. Combinatorial Optimisation Combinatorial optimization involves choosing the optimal solution from a large number of finite solutions. The travelling salesman problem (TSP), the vehicle routing problem (VRP), and the minimum spanning tree problem (MST) belongs to this category. Many industry problems are analogical to TSP, VRP, and MST. Non-Linear Programming (NLP) Large classes of problems faced by the industry have non-linear objective functions and/or non-linear constraints. Many engineering design optimization problems belongs to this category. NLP are also difficult set of problems to solve due to limitations of the existing algorithms. NLP is an integral part of several machine learning algorithms such as neural networks. The loss function which is used for finding weights for input variables is a non-linear function. Six Sigma Six Sigma and its problem-solving methodology DMAIC (Define, Measure, Analyse, Improve, and Control) are frequently used in process improvement problems. Social Media Analytics Tools Social media analytics is a collection of tools and techniques used for analysing unstructured data such as texts, videos, photos, and so on. With the exponential growth

in the use of social media by the general public, tools designed for analysing unstructured data will be frequently used by organizations.

4. Veracity of the data refers to the quality issues. Especially in social media there could be biased or incorrect data, which can result in wrong inferences.

WEB AND SOCIAL MEDIA ANALYTICS

- Social media and mobile devices such as smart phones are becoming an important source of data for all organizations, small and big.
- ***Business Analytics helps to create a buzz or electronic word-of-mouth (WoM) effectively. Stelzner (2013) claimed that 86% of the marketers indicated that social media is important for their business. Stelzner (2013) identified the following questions as the most relevant for any marketers when dealing with social media engagement (also valid for mobile devices):***
- ***1. What is the most effective social media tactic?***
- ***2. What are the best ways to engage the customers with social media?***
- ***3. How to calculate the return on investment on social media engagement?***
- ***4. What are the best social media management tools?***
- ***5. How do we create a social media strategy for the organization?***

"Machine learns with respect to a particular task T, performance metric P following experience E, if the system reliably improves its performance P at task T, following experience E". Let the task T be a classification problem. To be more specific, consider customer's propensity to buy a product. The performance P can be measured through several

metrics such as overall accuracy, sensitivity, specificity, and area under the receive operating characteristic curve (AUC).

- The experience E is analogous to different classifiers generated in machine learning algorithms such as random forest (in random forest several trees are generated and each tree is used for classification of new case).
- ***Carbonell et al. (1983) list the following three dimensions of machine learning algorithms:***
- ***1. Learning strategies used by the system.***
- ***2. Knowledge or skill acquired by the system.***
- ***3. Application domain for which the knowledge is obtained.*** Business Analytics Carbonell et al. (1983) classifies learning into two groups: knowledge acquisition and skill refinement.

They give an example of knowledge acquisition as learning concepts in physics whereas skill refinement is similar to learning to play the piano or ride a bicycle. Machine learning algorithms imitate both knowledge acquisition as well as skill refinement process.

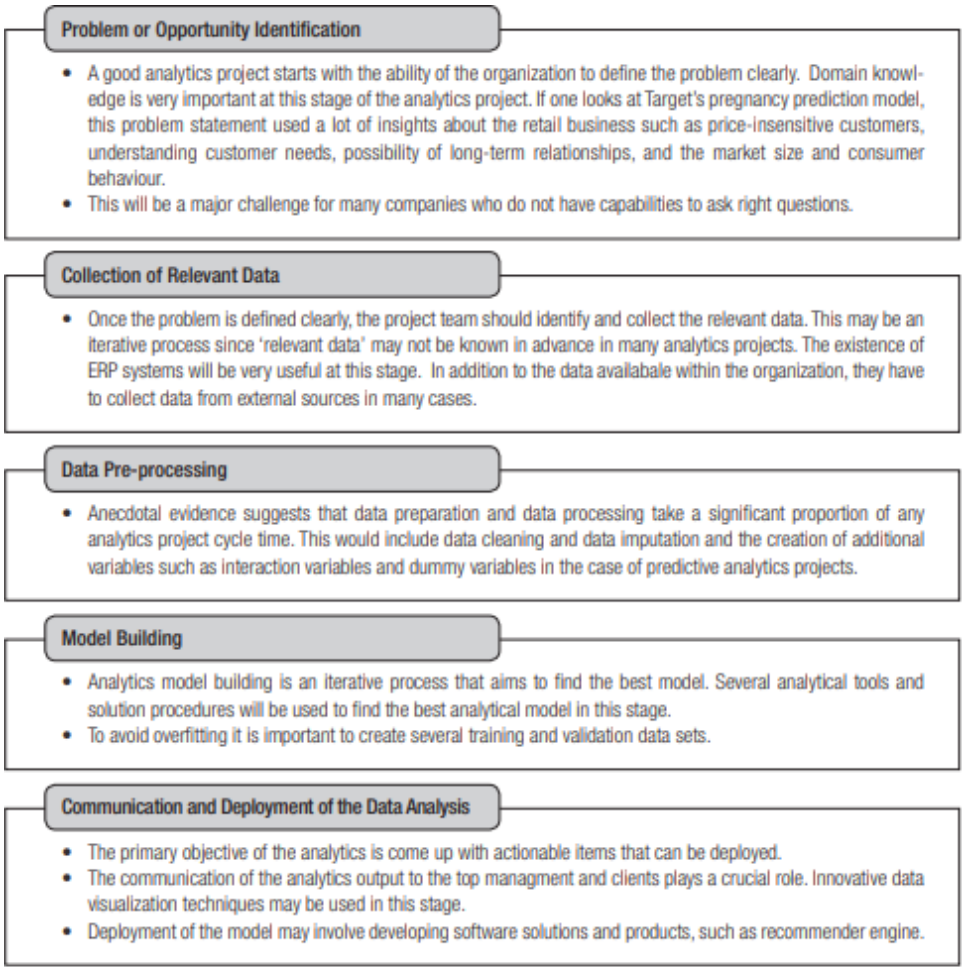

FIGURE 1.9 Framework for data-driven decision making.

- Descriptive analytics is the starting point of analytics based solution to problems. It helps to understand the data and provide directions for predictive and prescriptive analytics. Business Intelligence (BI), which largely involves creating reports and business dashboard that lead to actionable insights, is essentially a descriptive analyticsexercise.

Pattern Discovery

Revise Association Rule Mining notes and the understand the following presented algorithms

1. ***Suppose you have the set C of all frequent closed itemsets on a data set D, as well as the support count for each frequent closed itemset. Describe an algorithm to determine whether a given itemset X is frequent or not, and the support of X if it is frequent.***

Answer:

Algorithm: Itemset Freq Tester. Determine if an itemset is frequent.

Input: C, set of all frequent closed itemsets along with their support counts; test itemset, X.

Output: Support of X if it is frequent, otherwise -1.

Method:

(1) s = ∅;

(2) for each itemset, l ∈ C

(3) if X ⊂ l and (length(l) < length(s) or s = ∅) then {

(4) s = l;

(5) }

(6) if s≠ ∅ then {

(7) return support(s);

(8) }

(9) return -1;

2. ***An itemset X is called a generator on a data set D if there does not exist a proper sub-itemset Y ⊂ X such that support(X) = support(Y). A generator X is a frequent generator if support(X) passes the minimum support threshold. Let G be the set of all frequent generators on a data set D. (a) Can you determine whether an itemset A is frequent and the support of A, if it is frequent, using only G and the support counts of all frequent generators? If yes, present your algorithm. Otherwise, what other information is needed? Can you give an algorithm assuming the information needed is available? (b) What is the relationship between closed itemsets and generators?***

Algorithm: InferSupport. Determine if an itemset is frequent.

Input:

• l is an itemset;

• F G is the set of frequent generators;

• P Bd(F G) is the positive border of F G;

Output: Support of l if it is frequent, otherwise -1.

Method:

(1) if l ∈ F G or l ∈ P Bd(F G) then {
(2) return support
(3) } else {
(4) for all l′ ⊂ l and l′ ∈ P Bd(F G)
(5) Let a be the item such that l″ = l′ − {a} and l″ ∈ F Gand support(l″)=support(l′);
(6) l = l − {a};
(7) if l ∈ F G or l ∈ P Bd(F G) then {
(8) return support(l);
(9) } else {
(10) return -1;
(11) }
(12) }
Reference: LIU, G., LI, J., and WONG, L. 2008. A new concise representation of frequent itemsets
using generators and a positive border. Knowledge and Information Systems 17, 35-56.
(b) Very generally, they can be considered as opposites. This is because a closed itemset has no proper
super-itemset with the same support, while a generator has no proper sub-itemset with the same
support.

3. A database has four transactions. Let min sup = 60% and min conf = 80%.

cust_ID	*TID*	*items_bought* (in the form of *brand-item_category*)
01	T100	{King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
02	T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
01	T300	{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}
03	T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

(a) At the granularity of item category (e.g., itemi could be "Milk"), for the following rule template,
∀X ∈ transaction, buys(X, item1) ∧ buys(X, item2)⇒buys(X, item3) [s, c]
list the frequent k-itemset for the largest k and all of the strong association rules (with their
support s and confidence c) containing the frequent k-itemset for the largest k.
k = 3 and the frequent 3-itemset is {Bread, Milk, Cheese}. The rules are
Bread ∧ Cheese ⇒ M ilk, [75%, 100%]
Chees ∧ M ilk ⇒ Bread, [75%, 100%]
Chees⇒M ilk ∧ Bread, [75%, 100%]
(b) At the granularity of brand-item category (e.g., itemi could be "Sunset-Milk"), for the following
rule template,
∀X ∈ customer, buys(X, item1) ∧ buys(X, item2)⇒buys(X, item3)
list the frequent k-itemset for the largest k. Note: do not print any rules.
k = 3 and the frequent 3-itemset is {(Wheat-Bread, Dairyland-Milk, Tasty-Pie), (Wheat-Bread,

Sunset-Milk, Dairyland-Cheese)}.

4. ***Give a short example to show that items in a strong association rule may actually be negatively correlated.***
   Consider the following table:
   Let the minimum support be 40%. Let the minimum confidence be 60%. A⇒B is a strong rule because it satisfies minimum support and minimum confidence with a support of 65/150 = 43.3% and a confidence of 65/100 = 61.9%. However, the correlation between A and B is corrA,B = 0.433 0.700×0.667 = 0.928, which is less than 1, meaning that the occurrence of A is negatively correlated with the occurrence of B.

	A	$\overline{A}$	Σ_{row}
B	65	35	100
$\overline{B}$	40	10	50
Σ_{col}	105	35	150

5. The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, hotdogs refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and hamburgers refers to the transactions that do not contain hamburgers

	hot dogs	$\overline{hotdogs}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

   (a) Suppose that the association rule "hot dogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?
   (b) (b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?
   (c) (c) Compare the use of the all confidence, max confidence, Kulczynski, and cosine measures with lift and correlation on the given data.

(a) Suppose that the association rule "hotdogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?
For the rule, support = 2000/5000 = 40%, and confidence = 2000/3000 = 66.7%. Therefore, the association rule is strong.
(b) Based on the given data, is the purchase of hotdogs independent of the purchase of hamburgers?
If not, what kind of correlation relationship exists between the two?
corr{hotdog,hamburger} = P({hot dog, hamburger})/(P({hot dog})
P({hamburger}))=0.4/(0.5 × 0.6) = 1.33 > 1. So, the purchase of hotdogs is NOT independent of the purchase of hamburgers.
There exists a POSITIVE correlation between the two.

Refer Class Notes

Linear Regression and Logistic Regression: Logit transform, ML estimation, Tests of hypotheses, Wald test, LR test, score test, test for overall regression, multiple logistic regression, forward, backward method, interpretation of parameters, relation with categorical data analysis. Interpreting Regression Models, Implementing Predictive Models.

ALL THE BEST