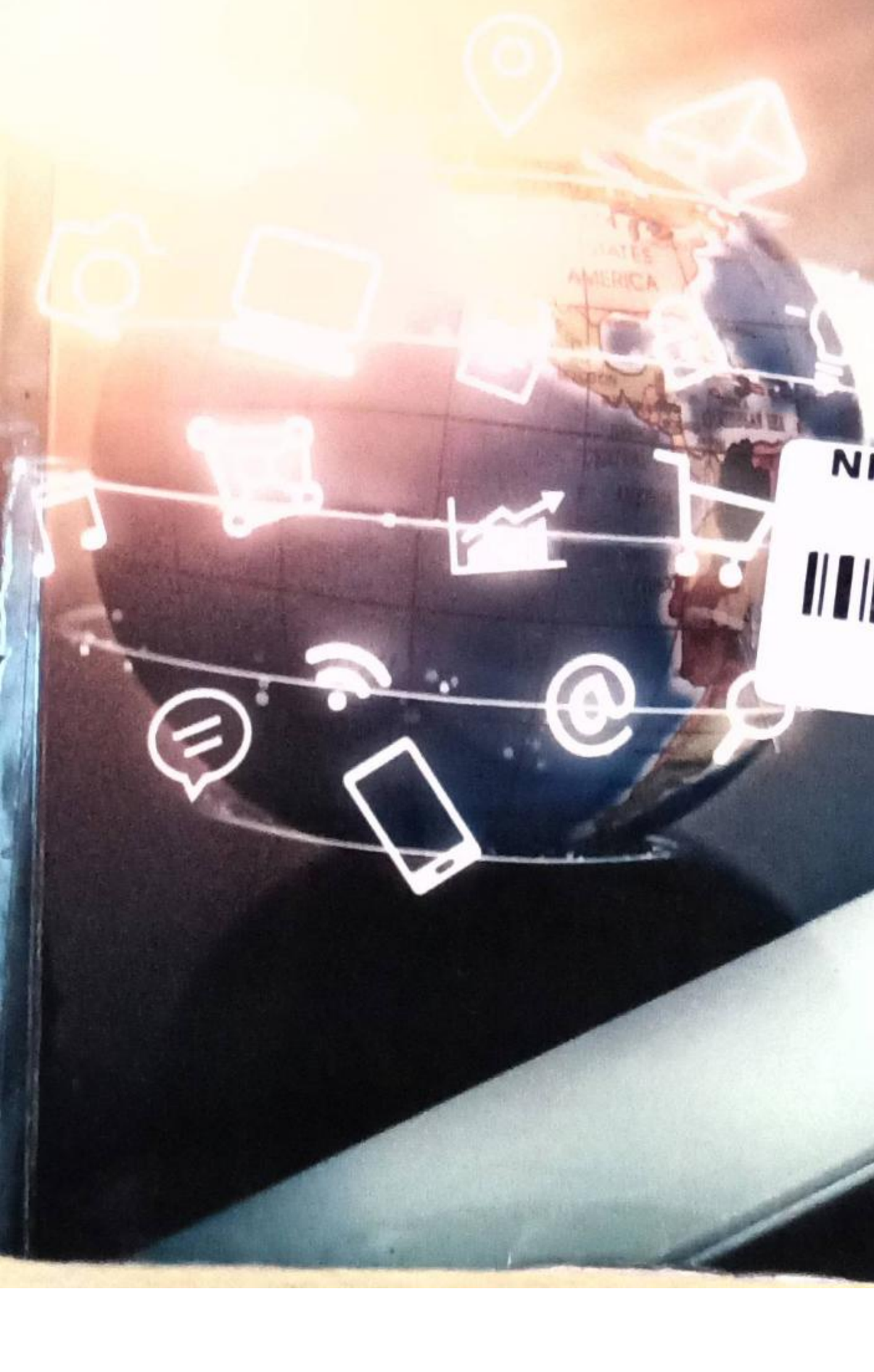
Business Analytics

The Science of Data-Driven Decision Making

U Dinesh Kumar



WILEY

Contents

Introduction to Business Analytics

1

"This Exists, So that Exists This is not there, so that is not there
This Ends, So that Ends This Arises, So that Arises."

— The Buddha

LEARNING OBJECTIVES

- **LO 1-1** Learn foundations of analytics and how it is becoming a competitive strategy for many organizations.
- **LO 1-2** Understand the importance of analytics in decision making and problem solving.
- **LO 1-3** Understand how different organizations are using analytics to gain insights and add value.
- **LO 1-4** Learn how organizations are using analytics to generate solutions and products.
- **LO 1-5** Understand different types of analytical models such as descriptive analytics, predictive analytics, and prescriptive analytics.
- **LO 1-6** Learn framework for analytics model development and deployment.
- **LO 1-7** Understand frequently used tools and techniques in analytics and problems solved using such tools and techniques.

BUSINESS ANALYTICS

Analytics has evolved from a simple number crunching exercise used for solving problems and assisting in decision making to a competitive strategy. In the beginning of the 21st century, analytics became one of the most important verticals within organizations due to its potential benefits including the ability to make better decisions and its impact on profitability of an organization. Today, several products and solutions are driven by analytics; Amazon Go, recommender systems, predictive keyboards used in smart phones and chatbot are few examples of solutions that are driven by analytics. It has become evident that analytics has become an important differentiator between high-performing and low-performing companies. Davenport and Patil (2012) claim that 'data scientist' will be the sexiest job of the 21st century.

IMPORTANT

Analytics is not just about number crunching. It has evolved into a competitive strategy that drives innovation across several organizations.

1.1 | INTRODUCTION TO BUSINESS 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, which is the focus of this book.

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.

Statistical and operations research techniques have been in use for several decades by many companies, but since 2000, companies that use analytics have increased exponentially. One reason for this increase in use of analytics is the *theory of bounded rationality* proposed by *Herbert Simon* (1972). According to Herbert Simon, the increasing complexity of business problems, the existence of several alternative solutions, and the limited time available for decision making demand a highly structured decision-making process using past data for the effective management of organizations. Decision making has become difficult due to reasons such as uncertainty, incomplete information about alternatives, lack of knowledge about cause and effect relationships between parameters of importance, and time available for decision making. For example, fraudulent transactions are a major problem for e-commerce companies; one such fraud is customers returning fake items in place of genuine items that they purchased (for example, buying branded Ray-Ban sunglasses and returning a fake sunglasses). Once the customer returns an item, the e-commerce company, such as Amazon, processes the refund of money within 3 to 5 business days (source: Amazon website). Given such a time constraint, e-commerce companies have to

identify fraudulent transactions in real time or within a very short duration (check whether the returned item is fake or genuine and start the refund process). But it is easier said than done since an expert is required to differentiate a fake product from a genuine one. What makes this problem even more difficult is that the number of stock keeping units (SKUs) sold by e-commerce companies runs into several millions and the number of transactions can run into several millions in a day making it a highly complex problem to deal with. Valentina Palladino (2013) reported that Amazon sold 426 items per second prior to December 2013 Christmas. The scale of operations of 21st century companies is huge and makes it difficult to manage the business without analytics. In 2015, Flipkart sold over 30 million products from more than 50,000 sellers through their platform. The number of visits to their portal was more than 10 million daily and the number of shipments exceeded 8 million per month (Bhansali *et al.*, 2016). 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.

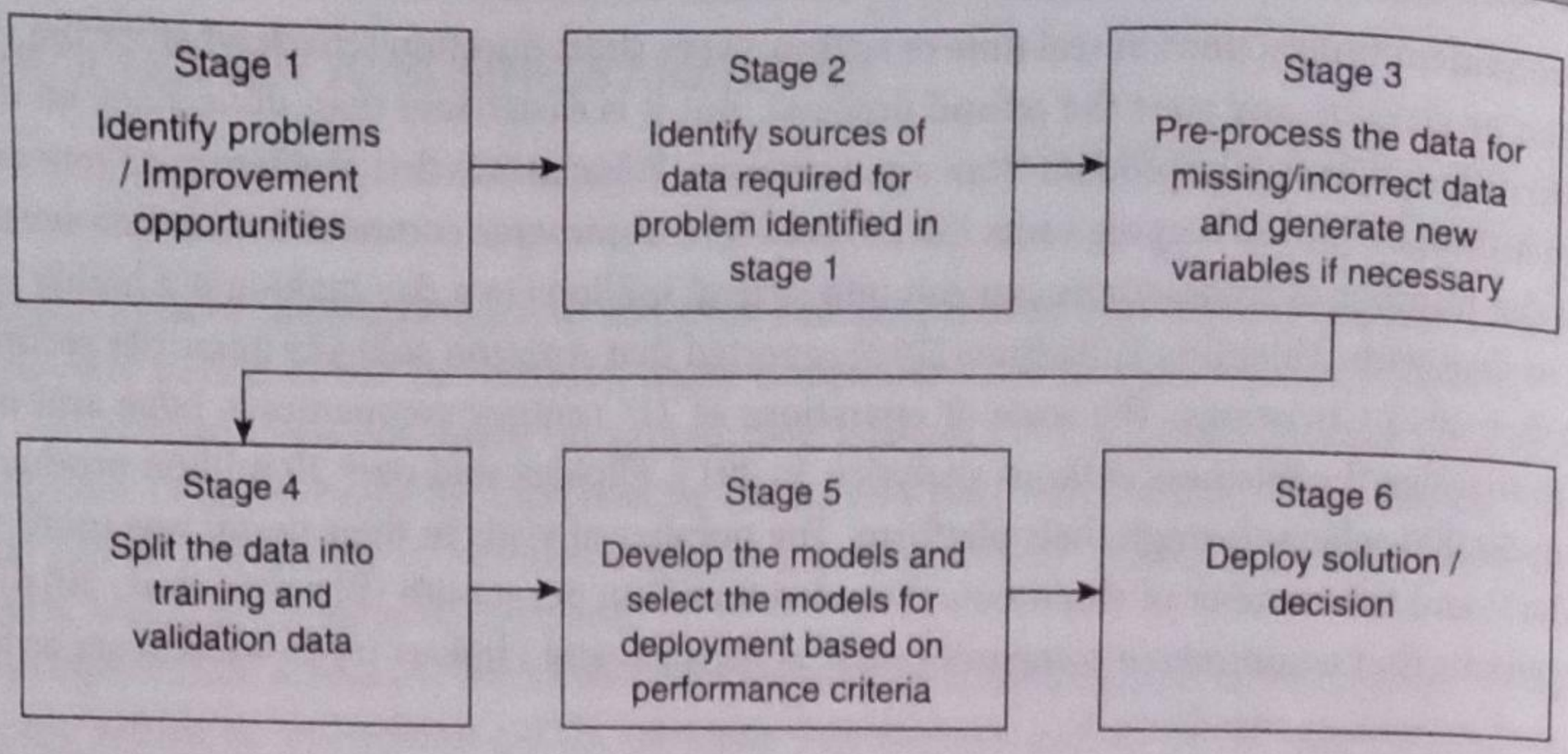


FIGURE 1.1 Business analytics – Data-driven decision-making flow diagram.

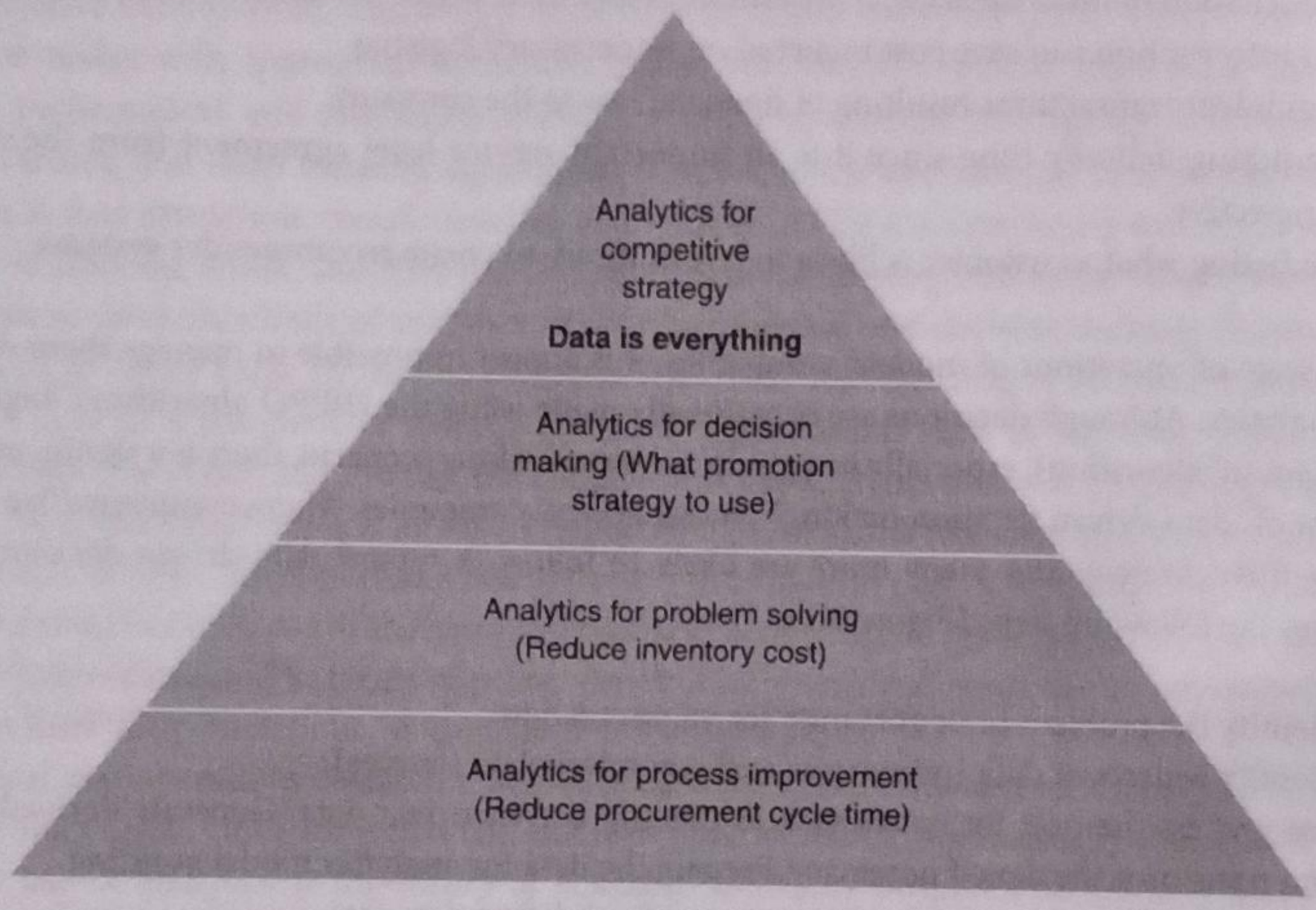


FIGURE 1.2 Pyramid of analytics.

Ransbotham and Kiron (2017) reported that they observed an increasing trend in companies using analytics to drive innovation and several companies reported competitive advantage from use of analytics.

1.2 | 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.

The profit earned by the firm would depend on how well they manage to minimize the transaction costs. Profit maximization or transaction cost minimization would mean making right decisions about the market, product/service, processes, supply chain, and so on. 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

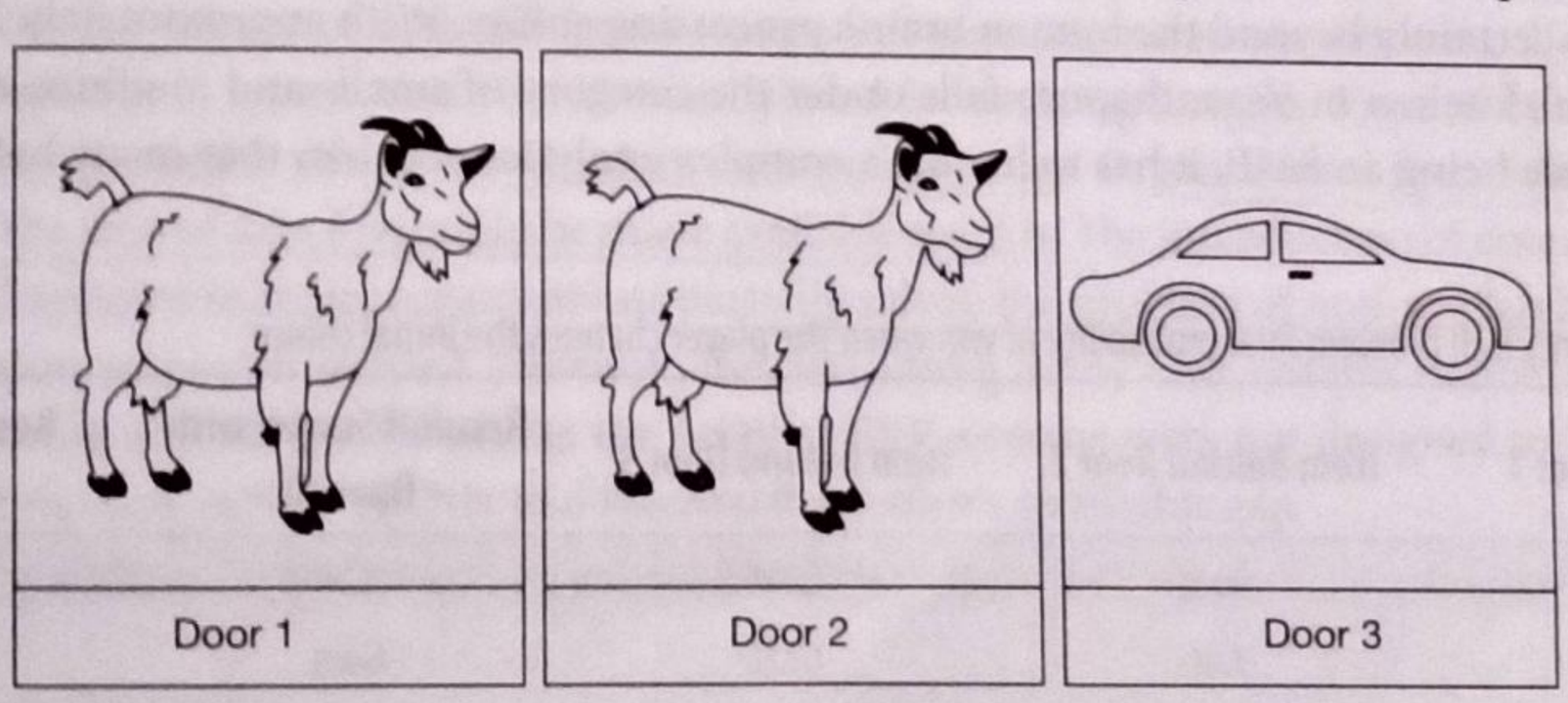


FIGURE 1.3 Monty Hall problem.

[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. In many real-life decision-making scenarios, the number of options available to the decision maker can be several millions or billions. The travelling salesman problem (TSP) is one such decision-making problem which many companies encounter in their business. In a TSP, given a list of cities and the distances between each pair of cities, the objective is to find the shortest possible route a salesperson should take to visit each city exactly once and return to the origin city. Many organizations that need to deliver products to many locations on regular basis encounter TSP. For example, in 2015 the Akshaya Patra Foundation (TAPF), which provides mid-day meals to approximately 1.4 million under privileged school children across India, faced this decision-making challenge (Srujana *et al.*, 2015). In 2015, through their Vasanthapura kitchen in Bangalore, approximately 84000 school children from 650 schools in South Bangalore were provided mid-day meals. Providing high quality food at an affordable price is one of the challenges faced by Akshaya Patra. The Vasanthapura kitchen used 35 vehicles to distribute the cooked food (Srujana *et al.*, 2015). To minimize the cost of distribution, they need to solve a complex vehicle routing problem (VRP). To simplify this problem, assume that they divide the number of schools equally among the vehicles; each vehicle would then have to deliver food to approximately 20 schools (few vehicles are kept as standby). For each vehicle, we need to find the best route. This problem can be formulated as a TSP with a solution space of 20 factorial ($20! = 2.4329 \times 10^{18}$). If a computer can evaluate one million routes per second, it would take more than 77146 years to evaluate all possible routes! For Akshaya Patra, every rupee saved would enable them to add more children to their mid-day meal programme. Given that the human brain lacks the ability to take the right decision in the Monty Hall problem that has just two alternatives, a problem with 20 factorial alternatives is certainly beyond the human brain's processing ability. With approximately 270 employees, Akshaya Patra's kitchen in Vasanthapura falls under the category of small- and medium-size enterprises (SMEs). Despite being an SME, it has to handle a complex analytics problem that many believe is relevant

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

only for big organizations. One of the misconceptions about analytics and big data technologies is that it is appropriate only for large organizations; however, the truth is that any organization, small or big, can benefit from the use of analytics. TSP is a problem that is encountered by several e-commerce companies for delivery of items placed by the customers and logistics service providers.

In today's world, data-driven decision making through business analytics is just not an option, but an essential capability that every organization should acquire irrespective of its size. As the competition increases, organizations cannot afford to shield inefficiencies. Analytics provides the capability for the organizations to be efficient and effective. Based on a survey of 3000 executives, Hopkins *et al.* (2010) claimed that there is a striking correlation between an organization's analytics sophistication and its competitive performance. The biggest obstacle to adopting analytics is the lack of knowledge about the tools and techniques that are required.

1.3 | BUSINESS ANALYTICS: THE SCIENCE OF DATA-DRIVEN DECISION MAKING

"Go down deep enough into anything and you will find Mathematics".

— Dean Schlicter

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.

Increasing complexities associated with businesses in the form of scale of operations and competition demand deeper understanding of the market and customers to serve better and succeed in the market. One of the main reasons for analytics is the scale of operations. If Walmart was a country, its GDP would be 28th in the world, its revenue in 2014 was 485.7 billion US dollars (Snyder, 2015). Merchandizing, shelf space allocation, promotions, brand monitoring, managing talent at the scale of operations of Walmart, Target, and Amazon requires solving complex problems in real time. The human mind lacks the ability to choose the right decisions due to the complexity of the problems that the organizations are facing and the limited time available for decision making (Simon, 1972).

In the 1980s, the culture of data collection was poor. Many organizations did not collect data or the data collected was not in a form that could be easily used for deriving insights. Even in 2017, many companies collect data manually which may result in data quality issues. Organizations found decision making difficult due to the lack of data that could be made available quickly. The introduction of enterprise resource planning (ERP) systems in many organizations partially solved the problem of non-availability of data that can be called upon whenever needed. However, the data sitting in the ERP systems needed to be analysed for problem solving and decision making; the original ERP systems were not designed to build analytics models. Platforms such as SAP HANA and Microsoft Azure try to fill this gap.

Business Analytics can be broken into 3 components:

1. Business Context
2. Technology
3. Data Science

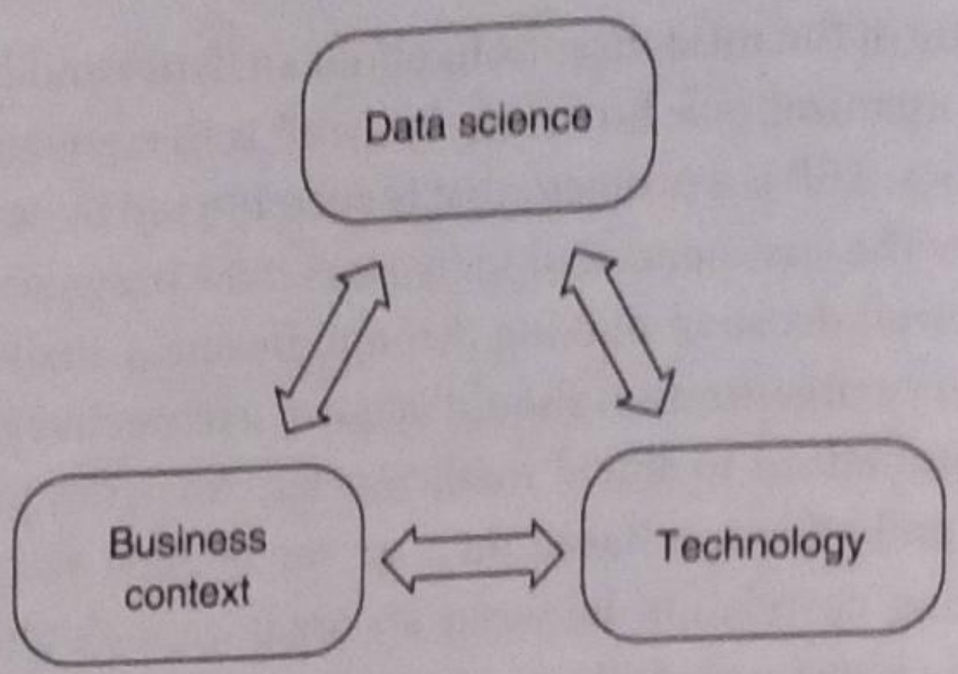


FIGURE 1.4 Components of business analytics.

Interaction between the three components is shown in Figure 1.4.

1.3.1 | Business Context

Business analytics projects start with the business context and ability of the organization to ask the right questions. Consider Target's Pregnancy prediction (Duhigg, 2012), which is a great example for organizations' ability to ask the right questions. Target is one of the largest retail chains in the world and in 2015, the revenue of Target Corporation was approximately US $ 73 billion. According to Duhigg (2012), Target developed a model to assign a pregnancy score to each female customer among their shoppers which could be further used for target marketing. However, what is so special about this prediction, and why pregnant women? This is where the knowledge of business context plays an important role. The following business knowledge manifests the importance of pregnancy prediction from a retailer's perspective:

1. Pregnant women are likely to be price-insensitive, so they become the Holy Grail for retailers such as Target. Expectant women are usually willing to spend more for their comfort as well as the babies'.
2. US Department of Agriculture reported that the expenses on children in 2015 ranged between US Dollars (USD) 9,600 and USD 19,700 (Lino *et al.*, 2017). According to National Vital Statistics report (2017), close to 4 million children were born in 2015, that is, the market size of baby-related products was at least USD 38 billion (Martin *et al.*, 2017). The market size was probably similar during the early 2000s when Target developed the pregnancy prediction model.
3. For many customers shopping is a habit, and most do not respond to promotions since shopping is a routine for them. The shopping behaviour changes during special events such as marriage and pregnancy and it becomes easy to target them during these special events.

The 'pregnancy prediction' is based on the insights about price-insensitive customers and the market size of baby products. In analytics, knowledge of business context is important for the ability to ask the right questions to start the analytics project.

Another good example of business context driving analytics is the 'did you forget feature' used by the Indian online grocery store bigbasket.com (Abraham *et al.*, 2016). Many customers have the tendency to forget items they intended to buy. Fernandes *et al.* (2013) reported that on average, customers forget

30% of the items they intend to buy. Forgetfulness can have significant cost impact for the online grocery stores. The customers may buy the forgotten items from a nearby store where they live, but since she/he is already in the store she/he may buy more items resulting in reduction in basket size in the future for online grocery stores such as bigbasket.com. Alternatively, the customer may place another order for forgotten items, but this time, the size of the basket is likely to be small and results in unnecessary logistics cost. Thus, the ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com.

Another problem that online grocery customers face while ordering the items is the time taken to place an order. Unlike customers of Amazon or Flipkart, online grocery customers order several items each time; the number of items in an order may cross 100. Searching for all the items that a customer would like to order is a time-consuming exercise, especially when they order using smart phones. Thus, bigbasket created a 'smart basket' which is a basket consisting of items that a customer is likely to buy (recommended basket) reducing the time required to place the order (Abraham *et al.*, 2016).

The above two examples (Target's pregnancy test and 'did you forget' and smart basket feature at bigbasket.com) manifest the importance of business context in business analytics, that is, the ability to ask the right questions is an important success criteria for analytics projects.

1.3.2 | Technology

To find out whether a customer is pregnant or to find out whether a customer has forgotten to place an order for an item, we need data. In both the cases, the point of sale data has to be captured consisting of past purchases made by the customer. Information Technology (IT) is used for data capture, data storage, data preparation, data analysis, and data share. Today most data are unstructured data; data that is not in the form of a matrix (rows and columns) is called unstructured data. Images, texts, voice, video, click stream are few examples of unstructured data. To analyse data, one may need to use software such as R, Python, SAS, SPSS, Tableau, etc. Technology is also required to deploy the solution; for example, in the case of Target, technology can be used to personalize coupons that can be sent to individual customers. An important output of analytics is automation of actionable items derived from analytical models; automation of actionable items is usually achieved using IT.

1.3.3 | Data Science

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 the following sections, we shall discuss the three types of analytics in detail.

1.4 | 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

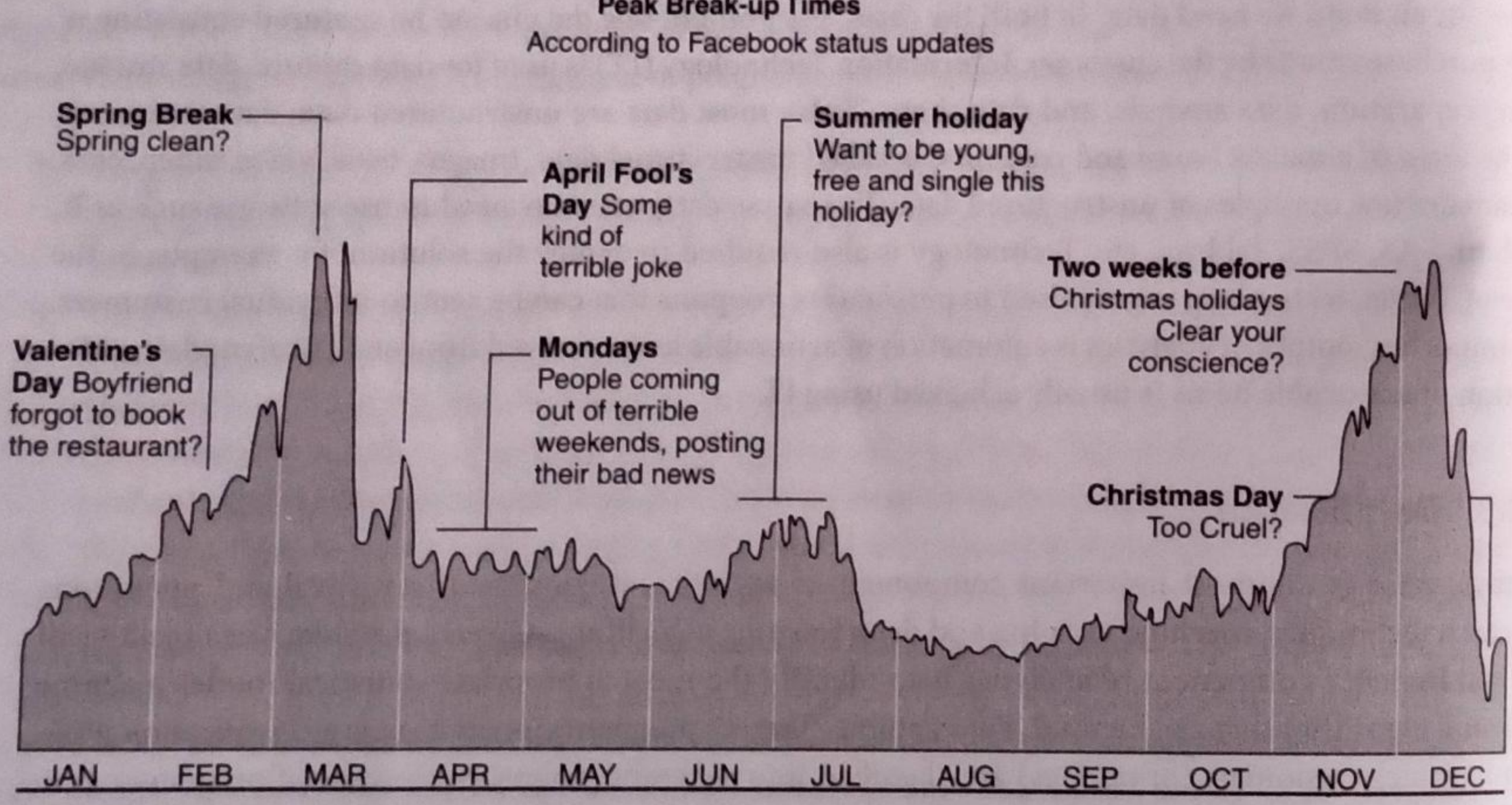


FIGURE 1.5 Peak breakup times according to Facebook status update. *Source*: David McCandless and Lee Bryon.

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. For example, when Hurricane Charley struck the U.S. in 2004, Linda M. Dillman, Walmart's Chief Information Officer, wanted to understand the purchasing behaviour of their customers (Hays, 2004). Using data mining techniques, Walmart found that the demand for strawberry pop-tarts went up over 7 times during the hurricane compared to their normal sales rate; the pre-hurricane top-selling item was found to be beer. These insights were used by Walmart when the next hurricane — Hurricane Frances — hit the U.S. in August–September 2004; most of the items predicted by Walmart sold quickly. Although the high pre-hurricane demand for beer can be intuitively predicted, the demand for strawberry pop-tarts was a complete surprise.

Data visualization to understand hidden facts/trends has been in use for several centuries. Dr. John Snow's cholera spot map is an interesting application of data visualization. Cholera claimed millions of lives across the world during the 19th century. Medical practitioners did not have a clear understanding of the causes of the disease (Cameron and Jones, 1983). Between 1845 and 1856, over 700 articles were

published in London on the causes of cholera and how the epidemic could be prevented (Snow, 2002). However, none of them offered any real solution. The breakthrough in cholera epidemiology was made by Dr. John Snow based on the data of cholera outbreak in central London in 1854. Between 31 August and 10 September 1854, over 500 people died of cholera in London. John Snow marked the locations of the homes of those who had died during the epidemic and prepared a spot map (Figure 1.6)[2]. The spot map revealed that the highest number of deaths occurred in the Golden Square area (Snow, 1999). The most striking difference between this area and the other districts of London was the source of water (Brody *et al.*, 2000); thus, Snow established that water contamination was the main source of cholera.

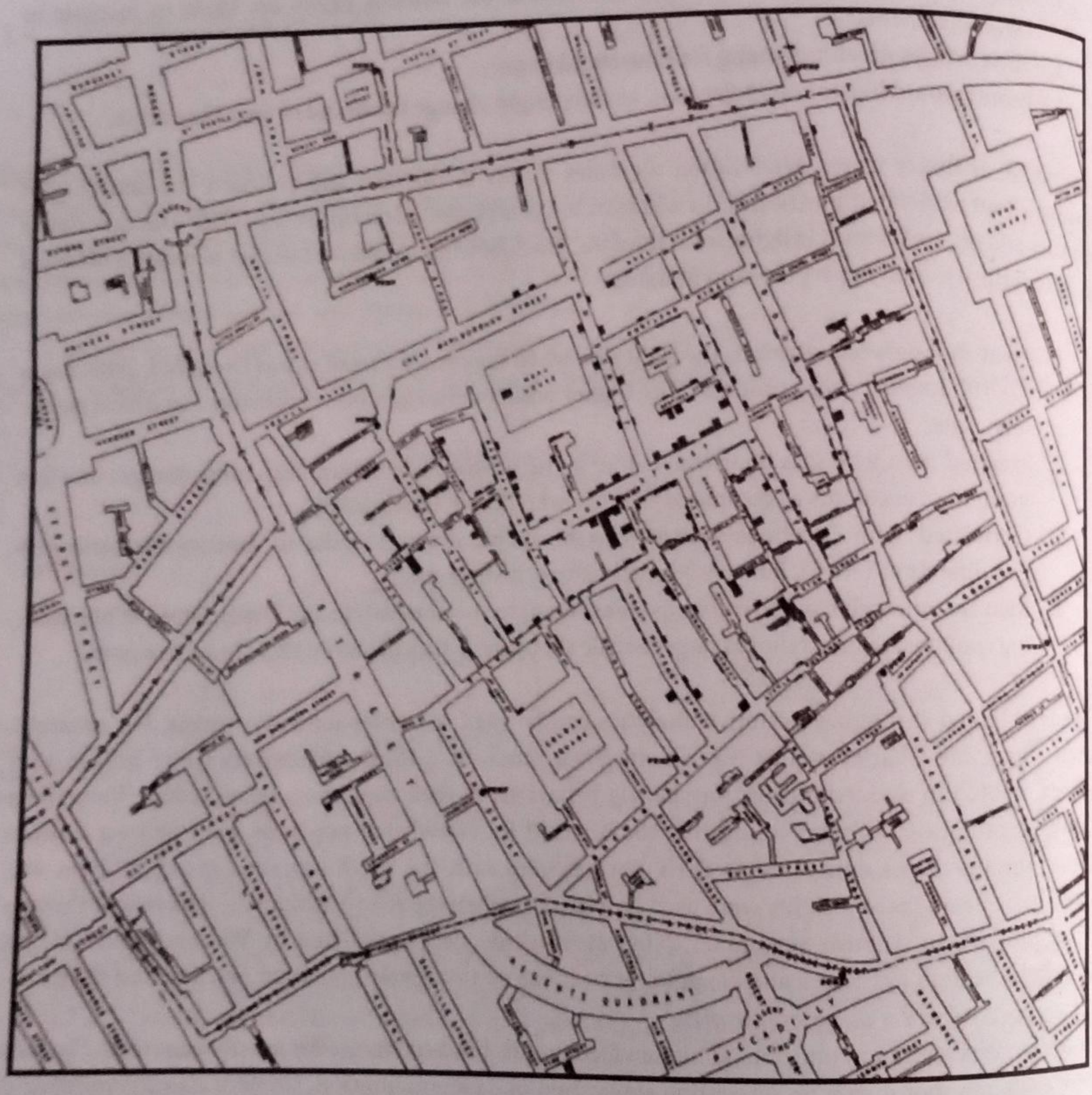


FIGURE 1.6 John Snow's spot map of cholera outbreak in London, 1854.

[2] Source: http://en.wikipedia.org/wiki/John_Snow_(physician)

Edward Tufte (2001), in his book titled *The Visual Display of Quantitative Information*, demonstrated how innovative visuals can be used to effectively communicate data. Google search keywords are used to predict demand for different apparel styles, jewellery, footwear, and so on to understand demand trends for many products. These trends help retailers take better decisions regarding procurement and inventory planning. Dashboards are created using innovative visuals form the core of business intelligence and are an important element of analytics. Tableau and Qlik Sense are popular visualization tools that are used by several organizations to create dashboards to monitor several key performance indicators relevant for the organization in real time. Indian companies such as Gramener[3] have used innovative data visualization tools to communicate hidden facts in the data. Descriptive analytics could be the initial stage of creating analytics capability.

Simple analysis of data can lead to business practices that result in financial rewards. For instance, companies such as RadioShack and Best Buy found a high correlation between the success of individual stores and the number of female employees in the sales team (Underhill, 2009). Underhill (2009) also reported that the conversion rate (percentage of people who purchased something) in consumer durable shops was higher among female shoppers than among male shoppers. Many organizations across the globe have to deal with fraudulent transactions. Sometimes, a simple query can lead to fraud detection. In 2014, China Eastern Airline found that a man had booked a first class ticket more than 300 times in a year and cancelled it before its expiry for full refund so that he could eat free food at the airport's VIP lounge (David K Li, 2014). In India, insurance frauds accounted for 2500–3500 crore in 2010 (Anon, 2013). It is always a good practice to start analytics projects with descriptive analytics.

1.5 | 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.

---

[3] Source: https://gramener.com/

The most popular example of the application of predictive analytics is Target's pregnancy prediction model discussed earlier in the chapter. 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.

The examples shown in Table 1.2 represent a tiny fraction of the predictive analytics applications used in the industry. Companies such as Procter & Gamble use analytics as a competitive strategy — every critical management decision is made using analytics (Davenport, 2013). If one were to search for the reasons behind highly successful companies, one would usually find analytics being deployed as the competitive strategy. Google — without which many people think the world would end — uses Markov chains for page ranking (Hayes, 2013). Google also developed accurate prediction models that could predict events such as the outcome of political elections, the launch date of a product, or action(s) taken by competitors (Coles *et al.*, 2007). Davenport and Harris (2007) reported how companies such as Amazon, Capital One, Harrah's, and the Boston Red Sox have dominated their business by using analytics. The application of analytics is not restricted to big corporates only; many sports clubs have successfully used analytics to manage their clubs. The most famous application of analytics in sports is by Oakland Athletics, which used analytics to put together a team with the limited resources available

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

for purchasing players (Lewis, 2003). Oakland Athletics had the third lowest payroll among the major league baseball teams in 2002. The manager of Oakland Athletics, Billy Beane, used statistical techniques to identify player qualities that made an impact on the match outcome and to also identify relatively cheaper skill. Oakland Athletics revised their team management strategy and with a payroll of USD 41 million, they were able to compete successfully in the league. In 2002, they won 20 games in a row.

1.6 | 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

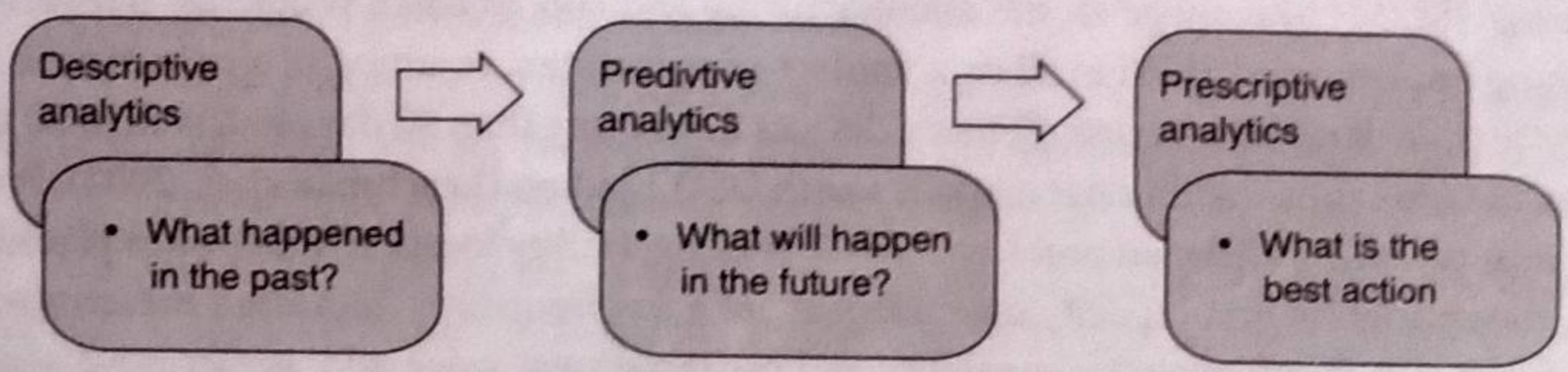


FIGURE 1.7 Link between different analytics capabilities.

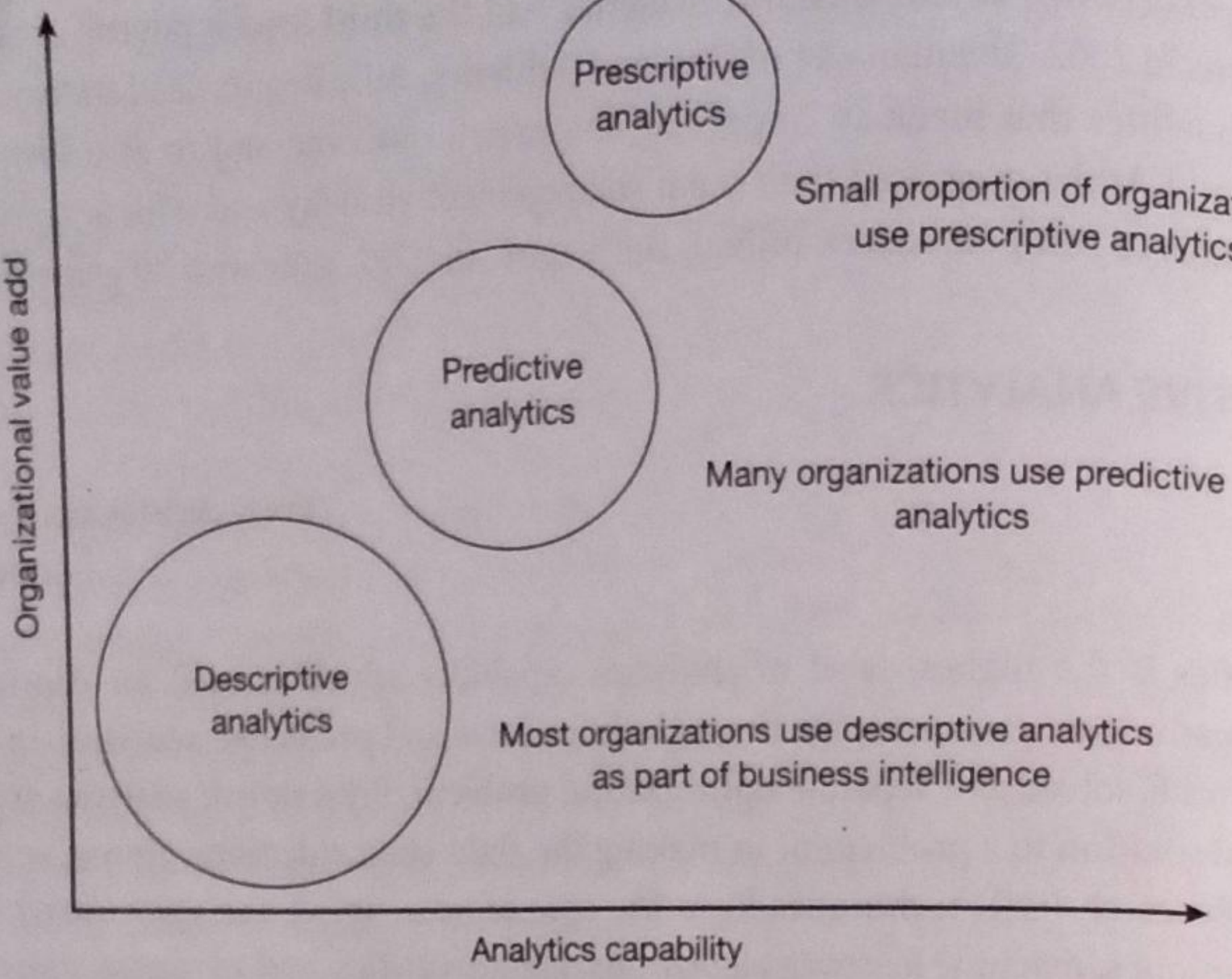


FIGURE 1.8 Analytics capability versus organizational value add.

containing 42 billion bottles and cans of Coca-Cola (Kant *et al.*, 2008). CCE developed an OR model that would meet several objectives such as improved customer satisfaction and optimal asset utilization for their distribution network of Coca-Cola products from 430 distribution centres to 2.4 million retail outlets. The optimization model resulted in cost savings of USD 54 million and improved customer satisfaction. The Akshaya Patra Midday Meal Routing and Transportation Algorithm (AMRUTA) was developed to solve the vehicle routing problem (discussed in Section 1.1); this was implemented at Akshaya Patra's Vasanthapura campus, resulting in savings of USD 75000 per annum (Mahadevan *et al.*, 2013). A major challenge for any e-commerce company is to improve the conversion of visits to transactions and order sizes. Hewlett Packard (HP) established HPDirect.com in 2005 to build online sales. HP Global Analytics developed predictive and prescriptive analytics techniques to improve sales. The analytical solutions helped HP to increase conversion rates and order sizes (Rohit *et al.*, 2013).

Inventory management is one of the problems that is most frequently addressed using prescriptive analytics. Samsung implemented a set of methodologies under the title '*Short Life and Low Inventory in Manufacturing*' (SLIM) to manage all the manufacturing and supply chain problems. Between 1996 and 1999, Samsung implemented SLIM in all its manufacturing facilities, resulting in a reduction in the manufacturing cycle time of random access memory devices from more than 80 days to less than 30 days. SLIM enabled Samsung to capture additional markets worth USD 1 billion (Leachman *et al.*, 2002). Product mix, marketing mix, travelling salesman problem, vehicle routing, facility location, manpower planning, capital budgeting, transportation and capacity management are a few frequently addressed prescriptive analytics problems. Figure 1.8 shows analytics capability and organizational value add; prescriptive analytics provides maximum value add to the organization since the benefits are realized during every period.

1.7 | 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.

(*Continued*)

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.

1.8 | 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 (10^{15} bytes) or exabytes (10^{16} 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. 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. 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.

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.

Big data is mostly misused terminology since a large proportion of analytics projects are not big data projects, in the sense that they do not challenge the existing computing technology and algorithms. Many companies see big data as a technology problem. Although it is true that technology is a constraint while addressing big data problems, a true big data problem challenges the algorithms that we have today, that is the algorithms are inadequate to solve these problems within a reasonable time. However, many problems in predictive and prescriptive analytics belongs to the big data category. When the problem associated with the data can challenge the existing computing technology due to its volume, velocity, and variety, then we have a big data problem. For example, Google processes 24 petabytes of data every day (Mayer-Schonberger and Cukier, 2013). Google was the first to exploit big data for targeted advertising using clickstream data. Google also predicted the spread of H1N1 flu based on the search terms entered by Google users (Ginsberg *et al.*, 2009). According to Richard Kellet, Director of Marketing at SAS, we (human) created 500 times more data in the last 10 years than what we had done prior to that, since the beginning of humanity (Scott, 2013). Every B787 flight creates half a terabyte of machine-generated data. For large banks, the automatic teller machine (ATM) transactions themselves will run into several billion transactions per month. Telecom call data, social media data, banking and financial transactions, machine-generated data, and healthcare data are a few examples of the sources of big data. Alternatively, any problem that can challenge the existing computing power, IT systems, and/or algorithms constitutes a big data problem. Big data problems need innovative ideas in order to handle the 4 Vs for deriving insights. The volume of the data generated is increasing every day, and most of this data is user generated, mainly from social media platforms, or machine generated. With the increase in Internet penetration and autonomous data capturing, the velocity of data is also increasing at a faster rate. It is estimated that 2.5 exabyte of data is created every day; this figure is likely to double in the near future (Mayer-Schonberger and Cukier, 2013). As the velocity of the data increases, traditional models such as regression and classification techniques may become unstable and invalid for analysis. The variety of data is another challenge within big data. Even in the case of text mining, users may use different languages within the same sentence, especially in a country such as India, which is home to hundreds of languages. Natural language processing (NLP), which is an essential component of big data, is challenging when multiple languages are used in the text data.

Innovative parallel processing capabilities such as Apache Hadoop (that comes with the ability to process large-scale data sets in multiple clusters using the Hadoop Distributed File System) and MapReduce (which enables parallel processing of large data sets) are used by organizations to handle big data. Big data technologies are still in a nascent stage and are far from maturity. However, these big data technologies do provide better computing power compared to existing technologies.

1.9 | 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. Social media is also an important marketing channel for marketers since it

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?

From the literature on social media and interviews conducted by **IIMB** (Dinesh Kumar *et al*, 2014) with industry experts, it is evident that social media is important for marketing products and services. However, the effectiveness of the social media marketing is still an emerging subject. Suhruta *et al* (2013) claimed that there is a relationship between social media engagement and the box-office collection of movies based on the data obtained from the Bollywood movie '1920 evil returns'. Social media has several advantages over conventional media as discussed below. Social media is measurable in terms of impressions, visits, views, clicks, comments, shares, likes, followers, fans, subscribers, etc. Impact of conventional media cannot be measured, for example, views of a hoarding or newspaper ad cannot be measured.

Social media is less expensive than conventional media and has the potential to reach a wider audience. Social media can create viral impact in a short duration and can reach a larger number of people. A key challenge in social media strategy will be assessing the return on investment. Return on Investment (ROI) should be calculated by the formula

$$\text{ROI} = (\text{Gain from Social Media Marketing} - \text{Cost of Social Media Marketing})/\text{Cost of Social Media Marketing}$$

However, it is difficult to quantify the actual gain from social media marketing. Hence, several variations are used to calculate ROI as given below (Hemann and Burbary, 2013, pages 276-284):

1. **Return on Engagement (ROE):** This measures the impact of social media marketing on users' engagement on the premise that higher engagement leads to higher awareness and thus greater likeliness to make a purchase decision (Hemann and Burbary, 2013). ROE calculation for some of the social media platforms is provided below:
   - Facebook – (Number of likes, comments, and shares on a post)/(Total number of Facebook page likes)
   - Twitter – (Number of replies, re-tweets)/(Number of followers)
   - YouTube – (Number of comments, ratings, and likes)/(Number of video views) OR (Number of comments, ratings, and likes)/(Number of subscribers)
2. **Return on Influence:** This tries to measure how social media activity changes the behaviour of users.
3. **Anecdotes:** This measures verbal sharing of sales activity or intent of purchase on the social media platforms.

4. **Correlation:** This measures the relationship between any social media engagement activity and actual sales.
5. **Multivariate Testing:** This measures the relationship between multiple social media engagement activities and actual sales and enables providing the right kind of offers and promotions to different users.
6. **Linking and Tagging:** This approach provides links on the social media to the buyers to make their purchase and thus it is possible to relate sales and social media engagement. Another way is to embed 'Cookies' (a piece of software), which track consumers' online activity, thus providing the connect between social media engagement and actual sales. However, this approach is more effective when the sales are conducted online.
7. **Social Commerce Approach:** In this method, sales are directly conducted through social media; for example, a store front is set up on Facebook page.
8. **Share of Conversation:** (Volume of conversation for a particular brand)/(Volume of conversation for entire industry)
9. **Sentiment Analysis:** Tracks overall brand perception by crawling through all the data available on the net. Kumar and Mirchandani (2012) have proposed new measures such as customer influence effect (CIE), stickiness index, and customer influence value (CIV).

1.10 | MACHINE LEARNING ALGORITHMS

Machine learning algorithms are part of artificial intelligence (AI) that imitates the human learning process. Humans learn through multiple experiences to perform a task. Similarly, machine learning algorithms usually develop multiple models and each model is equivalent to an experience. For example, consider someone trying to learn tennis. Getting the service right requires much practice especially, to learn to serve an ace (serve such that the opponent player is unable to reach the ball). To master the service in tennis (especially ace), a player probably has to practice several thousand times; each practice session is a learning. AI is still a developing field and nowhere near the human cognitive process. In machine learning algorithms, we develop several models which can run into several hundreds and each model is treated as a learning opportunity. Mitchell (2006) defined machine learning as follows:

> "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.

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. Machine learning algorithms are classified into the following four categories:

1. **Supervised Learning Algorithms:** When the training data set has both predictors (input) and outcome (output) variables, we use supervised learning algorithms. That is, the learning is supervised by the fact that predictors (X) and the outcome (Y) are available for the model to use. Techniques such as regression, logistic regression, decision tree learning, random forest, and so on are supervised learning algorithms.
2. **Unsupervised Learning Algorithms:** When the training data has only predictor (input) variables (X), but not the outcome variable (Y), then we use unsupervised learning algorithms. Techniques such as K-means clustering and Hierarchical clustering are examples of unsupervised learning algorithms.
3. **Reinforcement Learning Algorithms:** In many cases, the input variable X and the output variable Y are uncertain (predictive keyboards and/or spell check). The algorithms are also used in sequential decision-making scenarios; techniques such as dynamic programming and Markov decision process are examples of reinforcement learning algorithms.
4. **Evolutionary Learning Algorithms:** These are algorithms that imitate human/animal learning process. They are most frequently used to solve prescriptive analytics problems. Techniques such as genetic algorithms and ant colony optimization belongs to this category.

In this book, we will be discussing techniques from supervised, unsupervised, and reinforcement learning algorithms.

1.11 | FRAMEWORK FOR DATA-DRIVEN DECISION MAKING

The framework for data-driven decision making and problem solving can be divided into five integrated stages: *problem and opportunity identification; collection of relevant data; data pre-processing; analytics model building; and model deployment*. The various activities carried out during these different stages are described in Figure 1.9. The success of analytics projects will depend on how innovatively the data is used by the organization as compared to the mechanical use of analytical tools. Although there are several routine analytics projects such as customer segmentation, clustering, forecasting, and so on, highly successful companies blend innovation with analytics.

1.12 | ANALYTICS CAPABILITY BUILDING

Although many companies have successful analytics verticals within their organization, many are still in the process of creating one. In this section, we will discuss the pillars of building a centre of analytics excellence.

1. **Top Management Support:** Like many other initiatives, creating a data-driven decision-making process requires a change in the organizational culture, and without the support of the top-level

Problem or Opportunity Identification

- A good analytics project starts with the ability of the organization to define the problem clearly. Domain knowledge is very important at this stage of the analytics project. If one looks at Target's pregnancy prediction model, this problem statement used a lot of insights about the retail business such as price-insensitive customers, understanding customer needs, possibility of long-term relationships, and the market size and consumer behaviour.
- This will be a major challenge for many companies who do not have capabilities to ask right questions.

Collection of Relevant Data

- Once the problem is defined clearly, the project team should identify and collect the relevant data. This may be an iterative process since 'relevant data' may not be known in advance in many analytics projects. The existence of ERP systems will be very useful at this stage. In addition to the data availabale within the organization, they have to collect data from external sources in many cases.

Data Pre-processing

- Anecdotal evidence suggests that data preparation and data processing take a significant proportion of any analytics project cycle time. This would include data cleaning and data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

Model Building

- Analytics model building is an iterative process that aims to find the best model. Several analytical tools and solution procedures will be used to find the best analytical model in this stage.
- To avoid overfitting it is important to create several training and validation data sets.

Communication and Deployment of the Data Analysis

- The primary objective of the analytics is come up with actionable items that can be deployed.
- The communication of the analytics output to the top managment and clients plays a crucial role. Innovative data visualization techniques may be used in this stage.
- Deployment of the model may involve developing software solutions and products, such as recommender engine.

FIGURE 1.9 Framework for data-driven decision making.

management, it may be difficult to create a strong analytics culture within an organization. Data-driven decision making may not result in immediate benefits, especially when the process is new to an organization.

2. **Analytics Talent:** The second important factor in creating a successful analytics vertical is the talent. It is important that organizations identify the right talent and nurture them within the organization to avoid attrition. The organization should have the ability to differentiate the true analytics talent from the mediocre analytics professionals. Davenport and Patil (2012) listed 10 ideas for finding the right data scientists such as recruiting from top universities, using social media such as LinkedIn, looking for evidence, and so on.
3. **Information Technology (IT):** IT plays a crucial role in implementing analytics. Data capturing, data storage, data transfer, data analysis through analytical models, and finally, communication of the model output cannot be achieved without proper data architecture supported by other IT infrastructure. In addition to data handling capabilities, software tools such as R, Python, SAS,

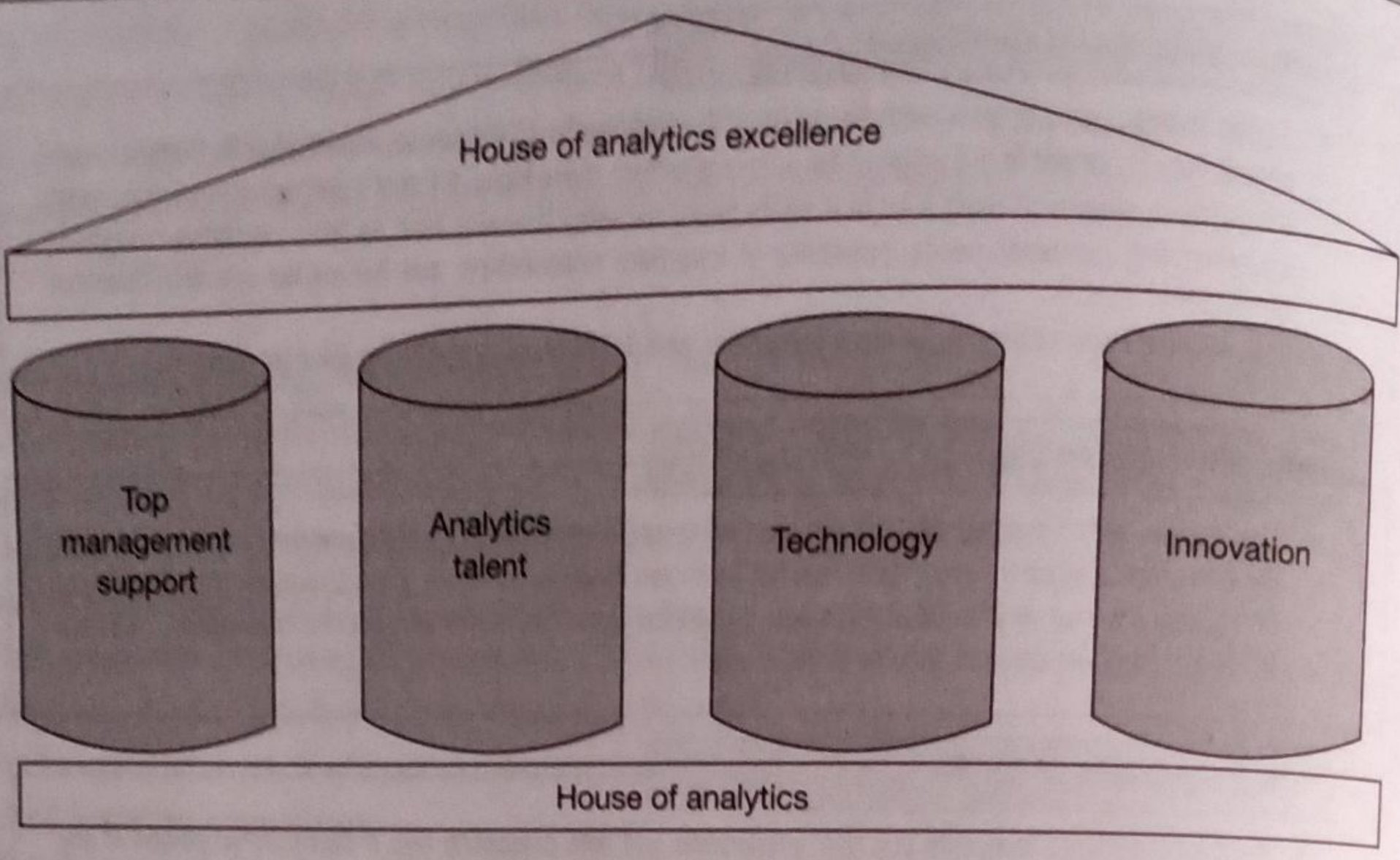


FIGURE 1.10 House of analytics excellence.

SPSS, STATA, Tableau, and so on are an essential part of IT support. A large number of organizations, including multinational companies, prefer open-source software such as R instead of proprietary software.[5] Organizations expect real-time decisions; thus, in-memory computing is becoming popular among analytics companies.

4. **Innovation:** The fourth pillar of analytics excellence is innovation. Without innovation, the analytics function may not achieve its full potential. All these pillars need to be integrated with the domain knowledge of the business; otherwise, the analytics may end up solving non-value-adding problems. The house of analytics is shown in Figure 1.10.

1.13 | ROADMAP FOR ANALYTICS CAPABILITY BUILDING

Many organizations, whether small or big, have a large number of low-hanging fruits that can be targeted with simple analytics tools such as descriptive statistics, data visualization, pivot tables, correlation analysis, basic quality tools, lean and Six Sigma. Data summarization tools such as pivot tables of Microsoft Excel can be used for targeting several small improvement opportunities. Lean and Six Sigma concepts are usually a good way to start analytics practices in an organization if they do not have strong analytics expertise and would like to initiate analytics practice within their organization. Companies who are planning to start analytics divisions to support decision making may use the framework shown in Figure 1.11. Sample industry-wide applications of analytics are captured in Table 1.4.

In addition to the primary data, the companies need to use data from secondary sources such as social media and other data sources such as Centre for Monitoring Indian Economy (CMIE), Capitaline, Bloomberg, Thomson Reuters, A C Nielson, Indiastat.com, etc.

[5] Many universities across the world teach analytics using R due to its capability, and not simply because it is open-source software.

1.14 | CHALLENGES IN DATA-DRIVEN DECISION MAKING AND FUTURE

Analytics requires a cultural change in an organization and managing this change will be the most significant challenge for many companies. This is true for any new initiative and is not specific to analytics. However, unlike many other initiatives, developing analytics skills can be a major hurdle, if these skills are not already present in the organization. Employees who are not skilled in analytical tools would need to be trained. Unlike other training programmes, analytics training can be long and expensive. Sirkin *et al.* (2005) identified duration, performance integrity, commitment, and effort as important factors that would determine the outcome of the change initiative.

If analytics talent is not available internally, the company should establish a system to identify the right talent. Building the right talent pool and retaining the talent would be a key challenge. Another big challenge would be the investment — the IT infrastructure required for advanced analytical techniques can be expensive. However, small and medium enterprises can achieve significant improvements by using simple tools such as MS Excel and open source software such as R and Python.

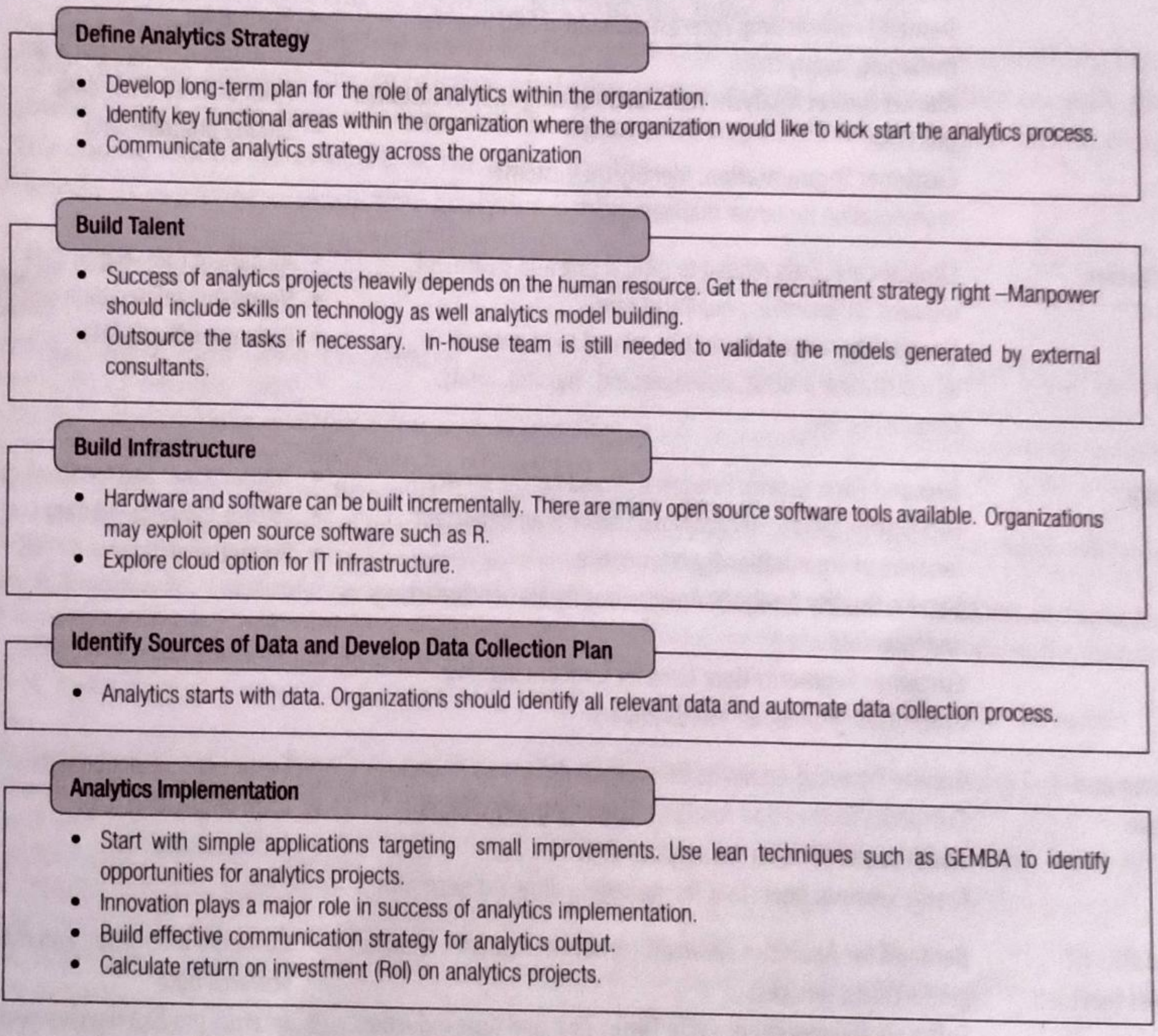


FIGURE 1.11 Roadmap for analytics capability building.

TABLE 1.4 Examples of industry-wise analytical problems and data resources

Industry Sector	Sample Analytical Problems	Data Sources
Manufacturing	**Supply Chain Analytics:** Inventory management, procurement, vendor selection, distribution management **Quality and Process Improvement:** Product quality, manufacturing quality, process improvement **Revenue and Cost Management:** Revenue maximization and cost minimization. **Warranty Analytics:** Manage end customer warranty and after sales support.	▪ Procurement, sales and production data. ▪ Warranty and after sales service data ▪ Commodity price data ▪ Manufacturing data. ▪ Macroeconomic data.
Retail	**Assortment Planning:** Category and SKU (stock keeping unit) management that will maximize the revenue and improve loyalty. **Promotion Planning:** Decide promotion strategy such as temporary price cuts, markdowns, bundling, etc. **Demand Forecasting:** Forecast demand at SKU level for managing supply chain. **Market Basket Analysis:** Association among SKUs in customer purchase. **Customer Segmentation:** Identify the customer segmentation for target marketing.	▪ Price data. ▪ Demand data at SKU and at category level. ▪ SKU level sales data with and without promotions. ▪ Planogram ▪ Customer demographics data. ▪ Point of Sales (PoS) data. ▪ Loyalty program data.
Healthcare	**Clinical Care:** Data related to clinical care and treatment required for improving quality of care. **Hospitality related data:** Data related to issues such as registration process, housekeeping, nursing, utility, diagnostics, etc.	▪ All patient care related data ▪ Hospitality related data ▪ Patient feedback data
Service	**Demand Forecasting:** Forecast demand for the service. **NPS Optimization:** Net Promoters Score is an important measure of organizational performance. **Service Quality Analysis:** Analyse quality for benchmarking and improvement. **Customer Segmentation:** Used for target marketing. **Promotion:** Promotion and its impact.	▪ Transactional and feedback data ▪ Pricing and demand data ▪ Promotional data
Banking and Finance	**Service Demand Analysis:** Demand for different services. **Customer Transaction Analysis:** Used for many different analytics and decision-making insights. **Credit Scoring:** Important for managing different portfolios.	▪ Customer transactional data ▪ Loan originating data ▪ Credit scoring data
IT and ITES (IT enables services)	**Demand for Analytics Services:** Identify demand for analytics products and services, **Software Development Cycle Time:** Cost and time reduction.	▪ Customer interaction and market research data ▪ Internal product development data

Analytics will become an integral part of organizations and majority of the decisions will be made using data in the future. Innovation will be the key success factor for analytics deployment.

1.15 | ORGANIZATION OF THE BOOK

The focus of the book is on data science. The book starts with basic concepts in statistics such as descriptive statistics, axioms of probability, concept of random variables, discrete and continuous probability distributions, hypothesis testing, and correlations. After the introduction to the basic concepts in probability and statistics, advanced concepts such as regression, logistic regression, decision trees, forecasting, and clustering are introduced with practical examples and case studies. Final few chapters are dedicated to prescriptive analytics, stochastic models, and Six Sigma.

We will be discussing several examples and case studies throughout the book for better understanding of the concepts discussed. Exercise questions from Chapter 9 onwards require deeper understanding of the concepts. Case studies provided in this chapter are distributed through Harvard Business Publishing case portal and are used by many institutions across the world.

The data sets used in the book can be downloaded from the following website:

https://www.wileyindia.com/business-analytics-the-science-of-data-driven-decision-making.html

The readers may go through the predictive analytics course offered by the author on edX platform. The course is part of the Indian Institute of Management Bangalore's massive open online course (MOOC). The course videos are available at the following link:

https://www.edx.org/course/predictive-analytics-iimbx-qm901x

REFERENCES

1. Anon (2003), "*Major Music Labels Use Artificial Intelligence to help determine Hitability of Music*", Music Industry News Network, 25 February 2003.
2. Anon (2013), "*Fraud in Motor and Health Insurance Global Perspective: Indian Approach*", Bimabazar.com Insurance Knowledge Portal, available at http://www.bimabazaar.com/index.php/2013-04-05-07-10-11/86-fraud-in-motor-and-health-insurance-global-perspective-indian-approach, accessed on 20 March 2017.
3. Anon (2016), "Harnessing the power of telecom data", *Hewlett Packard Enterprise Business White Paper,* available at https://h20195.www2.hpe.com/V2/getpdf.aspx/4AA6-4370ENW.pdf?ver=1.0.
4. Abraham P, Pradhan M, Lakshminarayanan, Iyer G and Kumar U D (2016), "*Customer Analytics at Bigbasket – Product Recommendations*", Indian Institute of Management Bangalore Case Study, IMB 573, available for download at: https://cb.hbsp.harvard.edu/cbmp/product/IMB573-PDF-ENG
5. Bhansali N, Rudravaram J, Grover S and Kumar U D (2016), "*Customer Analytics at Flipkart*", IIM Bangalore Case IMB 555.
6. Brody H, Rip M R, Johansen P V, Paneth N, and Rachman S (2000), "Map Making and Myth Making in Broad Street: The London Cholera Epidemic, *THE LANCET*, **356**(1), 64–68, 2000.
7. Cameron D and Jones I G (1982), "John Snow, the Broad Street Pump and the Model Epidemiology", *International Journal of Epidemiology,* **12** (4), 393–396, 1983.
8. Duhigg C (2012), "*The Power of Habit*", William Heinemann, London, 2012.
9. Coase R H (1937), "The Nature of the Firm", *Economica,* **4,** 386–405, 1937.
10. Coles P A, Lakhani K R and Mcafee A P (2007), "Prediction Markets at Google", *Harvard Business School Case* (Case Number 9-613-045).
11. Carneiro H A and Mylonakis E (2009), "Google Trends: A Web Based Tool for Real Time Surveillance of Disease Outbreaks", **49**(10), 1557–1564.