CHAPTER

DATA ANALYTICS: FUNDAMENTALS 2

Venkat N. Gudivada
East Carolina University, Greenville, NC, United States

2.1 INTRODUCTION

Data analytics is the science of integrating heterogeneous data from diverse sources, drawing inferences, and making predictions to enable innovation, gain competitive business advantage, and help strategic decision-making. The *data analytics* domain has evolved under various names including online analytical processing (OLAP), data mining, visual analytics, big data analytics, and cognitive analytics. Also the term *analytics* is used to refer to any data-driven decision-making. In fact analytics is a pervasive term and is used in many different problem domains under different names—road traffic analytics, text analytics, spatial analytics, risk analytics, and graph analytics, for example. In the last 3 years, new academic degree programs at the master's level have been introduced under the name *data science*.

The recent emergence of *Big Data* has brought upon the *data analytics* domain a bigger role as well as greater challenges. The bigger role comes from the strategic initiatives across various organizations, small and big, to leverage big data for innovation and competitive advantage. In addition to the predominantly structured data that the data analytics methods used hitherto, there is a need to incorporate both semistructured and unstructured data into the analytic methods. There is greater value in drawing upon heterogeneous but related data from sources such as social media, geospatial data, and natural language texts. This in itself is a very difficult problem. Among the other challenges, both the data volume and the speed of data generation have increased tremendously in the recent years. From 2008 to 2015 the world-wide data has increased from 50 petabytes (PB) to 200 PB [1].

There is a greater expectation that the data analytics methods not only provide insights into the past, but also provide predictions and testable explanations. Moreover, analytics is not limited to predictive models. The IBM Watson's Jeopardy! game championship in 2011 clearly demonstrated the increased role of data analytics. Watson is a question-answering system [2] and exemplifies *cognitive analytics*. It generates multiple hypotheses for answering a question and assigns a degree of confidence to each answer.

Data analytics and Business Intelligence (BI) figure in the top spots for CIO's technology priority list in 2013. They also appear in the top 10 CIO business strategies [3]. Analytics are used for solving a range of problems from improving process efficiency to cost reductions, providing superior customer service and experience, identifying new products and services, and enhancing security capabilities.

Data analytics plays a tremendous role in Intelligent Transportation Systems (ITS). The advent of Internet of Things (IoT) ushers in even a greater role for analytics in ITS. Heterogeneous data

Data Analytics for Intelligent Transportation Systems. DOI: http://dx.doi.org/10.1016/B978-0-12-809715-1.00002-X

originates from diverse sources such as weather sensors embedded in roadways, traffic signal control systems, social media, mobile devices such as smart phones, traffic prediction and forecasting models, car navigation systems, and connected car networks. Several software applications driven by this data are emerging. Such applications include emergency vehicle notification systems, automatic enforcement of speed limits, dynamic traffic light sequencing, vehicle-to-vehicle communication and collaboration, and real-time traffic prediction and rerouting.

The goal of this chapter is to provide a comprehensive and unified view of data analytics fundamentals. This exposition is intended to provide the requisite background for reading the chapters that follow. The intent is not to describe rigorous mathematical and algorithmic details about data analytics methods and practices. Entire books have been dedicated to providing that level of detail for topics such as OLAP, data mining, hypothesis testing, predictive analytics, and machine learning, which have implications for ITS.

The chapter is organized as follows. The four functional facets of data analytics from a workflow perspective—descriptive, diagnostic, predictive, and prescriptive—are described in Section 2.2. Next the evolution of data analytics from the late 1980s is traced in Section 2.3. The progression from SQL analytics, to business analytics, visual analytics, big data analytics, cognitive analytics is described. This evolution should be seen as a gradual increase in data analytics functional sophistication and the range of analytics-enabled applications.

Data science as the foundational discipline for the current generation of data analytics systems is discussed in Section 2.4. Data lifecycle, data quality issues, and approaches to building and evaluating data analytics are discussed in this section. An overview of tools and resources for developing data analytic systems is provided in Section 2.5. Future directions in data analytics are listed in Section 2.6. Section 2.7 summarizes and concludes the chapter. Questions and exercise problems are given in Section 2.8. Machine learning algorithms are a critical component of the state-of-the-art data analytics systems, and are discussed in Chapter 12 in this volume.

2.2 FUNCTIONAL FACETS OF DATA ANALYTICS

Just as a picture is worth a thousand words, data analytics facilitate unraveling several insightful stories from very large datasets. Based on the intended purpose of data analytics, the stories are placed into four broad functional categories—descriptive, diagnostic, predictive, and prescriptive. These four facets are highly interrelated and overlap significantly. From the author's standpoint, this categorization is only for exposition purpose. The facets represent an evolution of the analytics domain rather than a clear demarcation of functions across the categories. It is helpful to think of the facets as representing the sequence of steps in the *analytics workflow*.

The first phase in the workflow is *descriptive analytics*. The focus is on understanding the *current state* of a business unit or an organization. This phase also aims to glean insights into the distribution of data and detection of outliers. Descriptive analytics reveals both desirable and undesirable outcomes. The second phase leads into understanding what is causing that we observed in the first phase—*diagnostic analytics*.

Predictive analytics is the third stage in the analytics workflow. It helps analysts to predict future events using various statistical and mathematical models. There is a great overlap between

predictive and prescriptive analytics and this causes confusion. While predictive analytics forecasts potential future outcomes under various scenarios, *prescriptive analytics* provides intelligent recommendations about how to ensure only a chosen or preferred outcome. In other words predictive analytics forecasts probability of various events, but does not offer concrete steps which need to be executed to realize a chosen outcome. For example, predictive analytics may reveal a strong demand for an automobile model across the entire market space. However, in reality, actionable plans to increase sales across various regions of the marketplace are likely to vary from one region to another. Prescriptive analytics fills this need by providing execution plans for each region by incorporating additional data on weather, culture, and language.

In general as the workflow progresses from the first stage to the last, the diversity of data sources as well as the amount of data required increases. And so do the sophistication of the analytics models and their business impact. According to Kart [3], as of 2013, 70% of the organizations in their survey practice only descriptive analytics. These numbers drastically drop for diagnostic, predictive, and prescriptive analytics to 30%, 16%, and 3%, respectively.

2.2.1 DESCRIPTIVE ANALYTICS

Descriptive analytics provides both quantitative and qualitative information by analyzing historical data. Its goal is to provide insights into the past leading to the present, using descriptive statistics, interactive explorations of the data, and data mining. Descriptive analytics enables learning from the past and assessing how the past might influence future outcomes.

Organizations routinely use descriptive analytics to improve operational efficiencies and to spot *resource drains*. For example, software development organizations have been using descriptive analytics for decades under the name *software metrics and measurements*. The primary goal of these organizations is to produce high-quality and reliable software within specified time and budget. A software metric is a measure of the degree to which a software system possesses some property such as efficiency, maintainability, scalability, usability, reliability, and portability. Data such as total lines of code, number of classes, number of methods per class, and defect density is needed to characterize software metrics. The goal of the Capability Maturity Model (CMM) is to improve existing software development processes of an organization. The CMM model is based on the data collected from numerous software development projects.

2.2.1.1 Descriptive Statistics

Descriptive statistics is one of the approaches for realizing descriptive analytics. It is a collection of tools that quantitatively describes the data in summary and graphical forms. Such tools compute measures of *central tendency* and *dispersion*. Mean, median, and mode are commonly used measures of central tendency. Each measure indicates a different type of typical value in the data. Measures of dispersion (aka variability) include minimum and maximum values, range, quantiles, standard deviation/variance, distribution skewness, and kurtosis.

The *distribution* of a variable in a dataset plays an important role in data analytics. It shows all the possible values of the variable and the frequency of occurrence of each value. The *distribution* of the values of the variable is depicted using a table or function. Though *histograms* are simple to construct and visualize, they are not the best means to determine the shape of a distribution. The shape of a histogram is strongly affected by the number bins chosen. For this reason, a *kernel*

Table 2.1 Anscombe's Quartet Dataset

Dataset 1		Dataset 2		Dataset 3		Dataset 4	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

density plot (discussed later in this section), is a preferred method for determining the shape of a distribution. *Skewness* is a measure of the *asymmetry* of the distribution of a variable and *kurtosis* measures the *tailedness* of the distribution.

Anscombe's Quartet dataset is an elegant demonstration of the dangers involved when central tendency and dispersion measures are used exclusively. The quartet is comprised of four datasets, which appear to be quite similar based on the above measures, but scatter plots reveal how different the datasets are. Each dataset consists of 11 (*x, y*) pairs as shown in Table 2.1.

For all the four datasets, mean of x and y are 9 and 7.5, variance of x and y are 11.0 and 4.12, correlation between x and y is 0.816, and a linear regression to predict the y value for a given x value is given by the equation $y = 0.5x + 3$. However, the dataset differences are clearly revealed in the scatter plots shown in Fig. 2.1. The dataset 1 consists of data points that conform to an approximately linear relationship, though the variance is significant. In contrast there is no linear relationship among the points in dataset 2. In fact, these points seem to conform to a quadratic relationship. The datasets 1 and 3 exhibit some similarity. However, the points in dataset 3 more tightly conform to a linear relationship. Lastly, in dataset 4, x values are the same except for one outlier.

In summary we need multiple methods—measures of central tendency and variance, as well as graphical representations and interactive visualizations—to understand the true distributions of data. Interactive visualizations come under a group of techniques known as *exploratory data analysis* (EDA).

2.2.1.2 Exploratory Data Analysis

EDA techniques are used to interactively discover and visualize trends, behaviors, and relationships in data [4,5]. They also provide clues as to which variables might be good for building data analytic models—*variable selection* (aka *feature selection*). EDA enables three distinct and complementary

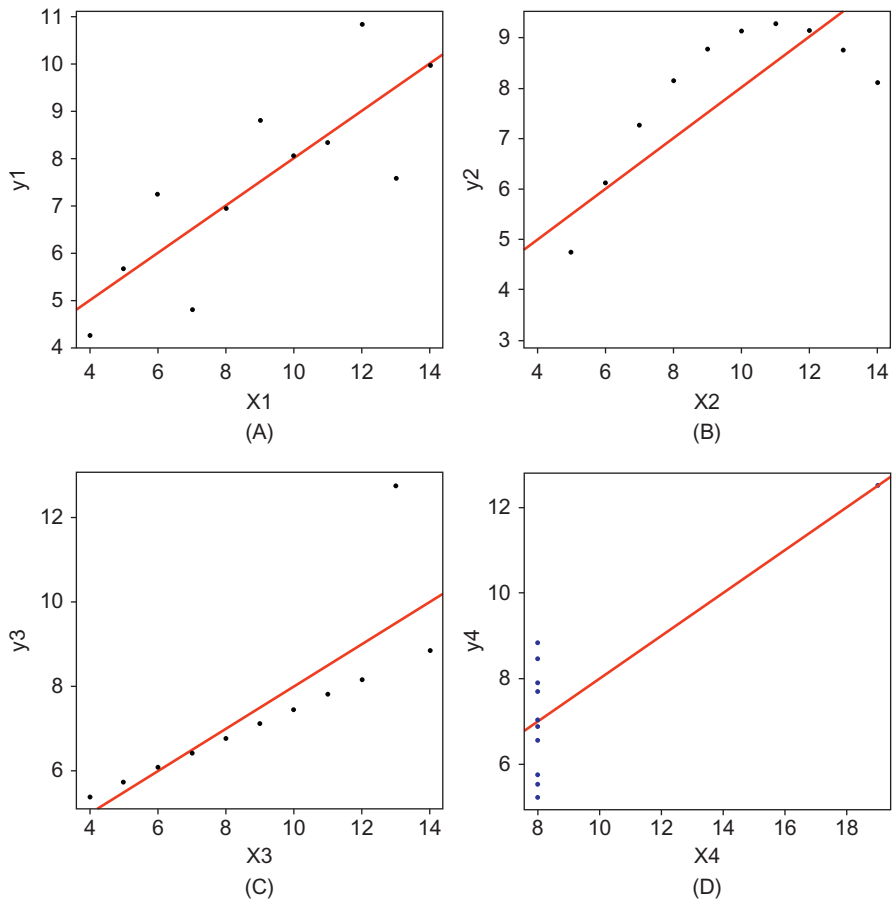

FIGURE 2.1

Scatter plots and linear regression models for Anscombe's dataset. (A) Anscombe's dataset 1; (B) Anscombe's dataset 2; (C) Anscombe's dataset 3; (D) Anscombe's dataset 4.

data analysis processes: presentation, exploration, and discovery. Visualization is an integral aspect of all three processes.

The goal of the *presentation* process is to gain a *quick and cursory familiarity* with the datasets. It involves computing and visualizing various statistics such as mean, median, mode, range, variance, and standard deviation (see Section 2.2.1.1). The type of statistics computed depends on the *data type* of the variable—nominal, ordinal, interval, and ratio. Visualization techniques for the presentation process range a broad spectrum from histograms to scatter plots, matrix plots, box-and-whisker plots, steam-and-leaf diagrams, rootograms, resistant time-series smoothing, and bubble charts.

The essence of visual *exploration* process lies in examining the data from multiple perspectives, identifying interesting patterns, and quantitatively characterizing the patterns to support

decision-making. This process supports both conceptual and insightful understanding of what is already *known* about the data (education and learning perspective) as well as help discover what is *unknown* about the data (research and discovery perspective). In other words the goals of the exploration process are to gain an intuitive understanding of the overall structure of the data and to facilitate analytical reasoning through visual exploration. The latter provides scaffolding for guided inquiry. It enables a deeper understanding of the datasets and helps to formulate research questions for detailed investigation. Recently, this exploration process is popularly referred to as *visual analytics*.

Lastly, the *discovery* process enables a data analyst to perform ad hoc analysis toward answering specific research questions. The discovery involves formulating hypotheses, gathering evidence, and validating hypotheses using the evidence. We illustrate some of the above concepts using R [6], which is a software system for statistical computing and visualization.

2.2.1.3 Exploratory Data Analysis Illustration

First, we define some terminology. A *quantile* is the fraction of data points that fall below a given value. For example, the 0.2 quantile is that data point q at which 20% of the data fall below q and 80% of the data fall above q. Related to quantiles are the four *quartiles* Q_1, Q_2, Q_3, and Q_4. Q_1 is the 0.25 quantile, Q_2 is the 0.50 quantile, Q_3 is 0.75 quantile, and Q_4 is 1.00 quantile. The difference $(Q_3 - Q_1)$ is called the *interquartile* (IQ) *range*. An *outlier* is an observation that is abnormally away from other other observations in a random sample from a population. Observations that are beyond $Q_3 + 1.5 \cdot IQ$ are called *mild outliers* and those even further beyond $Q_3 + 3 \cdot IQ$ are *extreme outliers*. Likewise we define similar outliers with respect to Q_1: values less than $Q_1 - 1.5 \cdot IQ$ are *mild outliers* and those that are even smaller than $Q_1 - 3 \cdot IQ$ are *extreme outliers*.

Several datasets come with the R software distribution, one of which is named *mtcars*. This dataset describes features of thirty-two, 1973−74 automobile models. The features include fuel consumption, and 10 aspects of automobile design and performance. In summary, the dataset has 32 observations, and 11 variables for each observation. This data was extracted from the 1974 Motor Trends US magazine.

The variables are (1) mpg (miles per gallon), (2) cyl (number of cylinders), (3) disp (volume of engine's cylinders) (4) hp (gross horsepower), (5) drat (rear axle ratio), (6) wt (gross weight), (7) qsec (seconds it takes to travel 0.25 miles from resting position), (8) vs (whether the car has a V engine or a straight engine), (9) am (transmission type: automatic or manual), (10) gear (number of forward gears), (11) carb (number of carburetors). Next we perform an EDA of mtcars dataset using boxplots, qqplots, and kernel density plots.

A *boxplot* is a graphical summary of the distribution of a variable. Fig. 2.2 depicts three boxplots. The left plot illustrates how the *mpg* feature varies for the 4-cylinder cars. The horizontal thick line in the box indicates the *median* value (Q_2). The horizontal lines demarcating the box top and bottom denote Q_3 and Q_1. The dotted vertical lines extending above and below the box are called *whiskers*. The top whisker extends from Q_3 to the largest nonextreme outlier. Similarly, the bottom whisker extends from Q_1 to the smallest nonextreme outlier. The center and right boxplots depict the same information for 6 and 8 cylinders cars.

A quantile−quantile (Q−Q) plot is a graphical method for comparing two distributions, by plotting their quantiles against each other. Fig. 2.3 depicts a Q−Q plot of the *mpg* variable against a

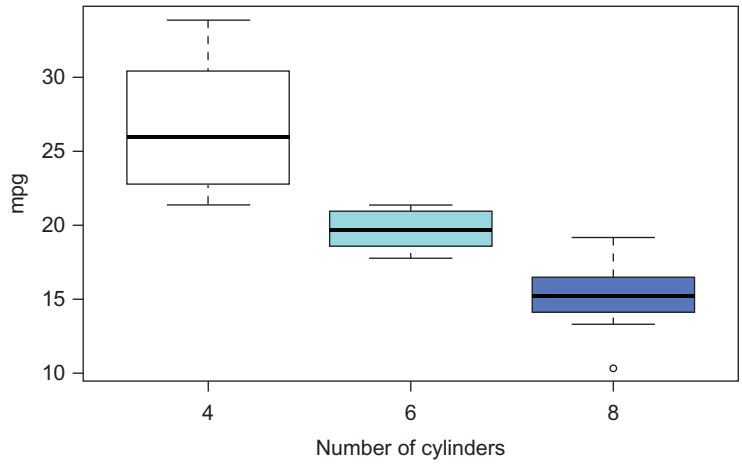

FIGURE 2.2

Boxplots of miles per gallon (mpg) variable for 4, 6, and 8 cylinder cars.

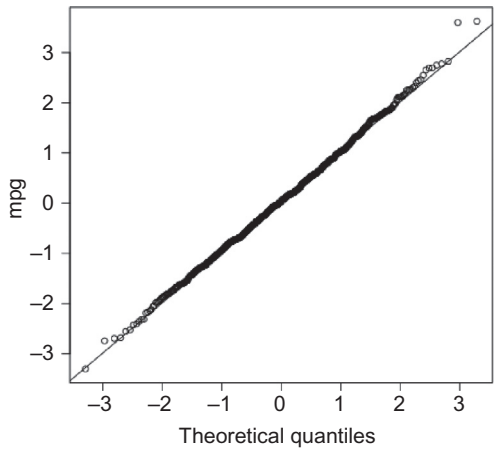

FIGURE 2.3

A Q−Q plot of the miles per gallon (mpg) variable.

theoretical (normal) distribution. A 45 reference line is also plotted. The line passes through the first and third quantiles. If the two datasets come from a population with the same distribution, the points should fall approximately along this reference line. This is the case for the mpg distribution. Therefore we can conclude that the variable mpg is *normally distributed*.

Sometimes it is desirable to look at the relationships between several variables. A *scatter plot matrix* enables such an exploration. Fig. 2.4 shows a scatter plot matrix of four variables—mpg,

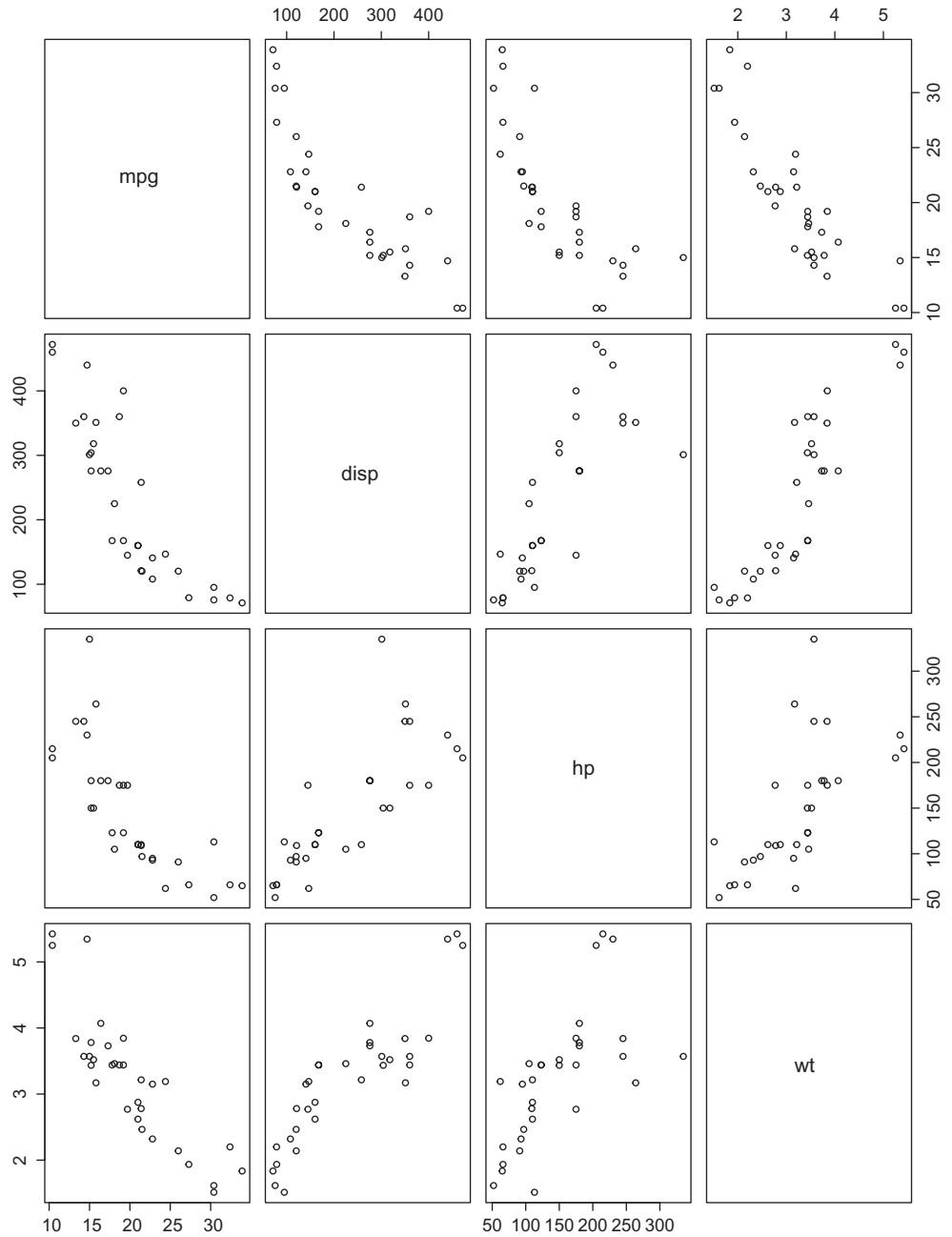

FIGURE 2.4

A scatter plot matrix of the mtcar dataset variables.

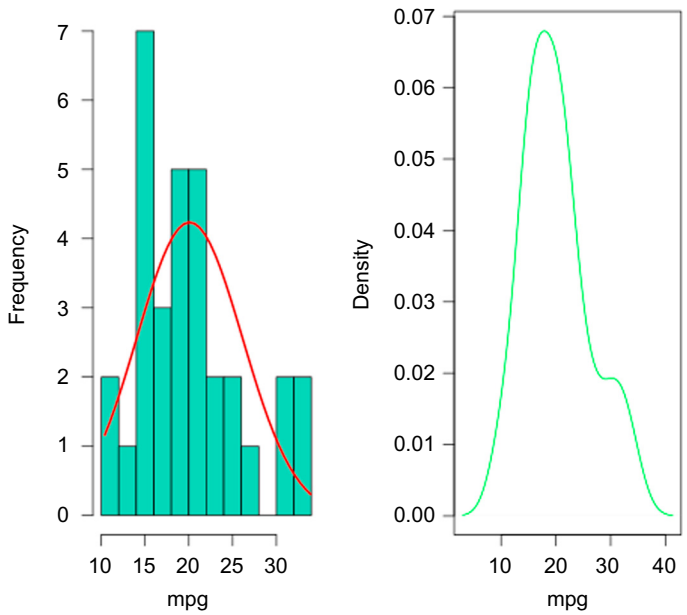

FIGURE 2.5

Histogram and kernel density function of miles per gallon (mpg).

displacement, horsepower, and weight. The number of rows and columns in the matrix is same as the number of variables. We assume that row and column numbers begin with 1. Consider the scatter plot at row 1 and column 2. The *x*-axis is the displacement variable and mpg is the *y*-axis. It appears that there is a good negative correlation between displacement and mpg. As another example, consider the scatter plot at row 4 and column 3. The *x*-axis is the horsepower and the *y*-axis represents the weight variable. There seems to be no correlation between the horsepower and weight variables.

Through a visual exploration of the scatter plot matrix, we can gain insights into correlations between variables. This exploration will also help us identify potential variables that may have greater predictive power.

Shown on the left in Fig. 2.5 is the histogram of the mpg variable superimposed with a *density curve* that fits the histogram. The density curve does not describe the data distribution accurately. A *kernel density plot* is more effective technique than a histogram in illustrating the distribution of a variable. A *kernel* is a probability density function (PDF) with the additional constraint that it must be even. There are several kernel functions and the Gaussian PDF is one of them. *Kernel density estimation* is a nonparametric method of estimating the PDF of a continuous random variable. It is nonparametric since no assumptions are made about the underlying distribution of the variable. Shown on the right in Fig. 2.5 is the kernel density plot of *mpg* constructed using R. Compared to

the density curve, kernel density plot more closely approximates the distribution. The mpg distribution is right-skewed indicating that the number of cars that have high mpg is few and farther.

2.2.1.4 Exploratory Data Analysis Case Studies

Most of the information that we deal with is in the form of text documents. As the number of documents increases, it becomes more difficult to sift through them and glean insights. Keyword-based search, as exemplified in Web search engines, returns too many documents. Furthermore users' information need is often at the *task level*. Text Insight via Automated, Responsive Analysis (TIARA) is a system for locating critical information from large document collections [7,8].

TIARA provides two major functions. The first function is the topic generation. A topic represents thematic information that is common to a set of text documents. A topic is characterized by a distribution over a set of keywords. The set of keywords associated with a topic are called *topic keywords*. Each topic keyword is assigned a probability, which measures the likelihood of the keyword appearing in the associated topic. TIARA uses the Latent Dirichlet Allocation (LDA) model to automatically extract a set of topics from a document collection. The LDA output includes a set of topics, keywords associated with each topic including the keyword probability distributions. To enable easy comprehension of the LDA output, TIARA provides rich user interaction tools. This second function help users interpret and examine the LDA output and summarized text from multiple perspectives.

TIARA also enables visualization of how topics have evolved over a period of time. Furthermore users can view and inspect the text analytic results at different levels of granularity using drill-down and roll-up functions. For example, using the drill-down function, users can navigate from a topic to the source documents that manifest the topic. The effectiveness of TIARA has been evaluated using two datasets. The first one is a collection of email messages. The second is the patient records from the National Hospital Ambulatory Medical Care Survey (NHAMCS) data.

Queensland Hospital Admitted Patient Data Collection (QHAPDC) is a public health dataset, which is independently generated by multiple hospitals [9]. The dataset features both clinical and demographic data. The clinical data is coded using the International Classification of Diseases taxonomy. Various visualization techniques, which are explanatory in nature, are used to expand the audience for the QHAPDC data. Furthermore visualization techniques are used to assess data quality, detect anomalies, identify temporal trends, spatial variations, and potential research value of QHAPDC. The goal for data quality assessment is to identify potential improvements to QHAPDC. Both positive and negative anomaly detection is used to promote improvements in clinical practice. Temporal trends and spatial variations are used to balance allocation of healthcare resources.

The visualization techniques used for the QHAPDC data include histograms, fluctuation plots, mosaic plots, time plots, heatmaps, and disease maps. These techniques provide insights into patient admissions, transfers, in-hospital mortality, morbidity coding, execution of diagnosis and treatment guidelines, and the temporal and spatial variations of diseases. This study discusses relative effectiveness of visualization techniques and associated challenges.

Modeling user interactions for exploratory analysis of spatiotemporal trend information using a visualization cube is discussed in [10]. The cube is comprised of four axes: spatial, temporal, statistics-value, and type-of-views. The model is implemented and the resulting prototype is used in elementary schools. It is demonstrated that the system features sufficient usability for fifth grade students to perform EDA.

Different levels of granularity in space, time, and data are intrinsic to spatiotemporal data. Coordinating these levels for an integrated visual exploration poses several challenges. Decomposing the data across various dimensions and displaying it has been proposed as a solution in Ref. [11]. The approach is demonstrated on election and poll data from Germany's various administrative levels and different time periods.

Space-Filling Multidimensional Data Visualization (SFMDVis) is a technique for viewing, manipulating, and analyzing multidimensional data [12]. It uses horizontal lines to represent multidimensional data items, which reduces visual clutter and overplotting. Every data point is directly selectable and this feature makes SFMDVis unique. SFMDVis is conceptually simple and supports a variety of interactive tasks for EDA. It enables direct data selection with AND and OR operators, zooming, 1D sorting, and K-means clustering.

Association rule mining typically generates a large number of rules. A visualization mechanism is needed to organize these rules to promote easy comprehension. *AssocExplorer* is a system for EDA [13] of association rules. AssocExplorer design is based on a three-stage workflow. In the first stage, scatter plots are used to provide a global view of the association rules. In the second stage, users can filter rules using various criteria. The users can drill-down for details on selected rules in the third stage. Color is used to delineate a collection of related rules. This enables users to compare similar rules and discover insights, which is not easy when the rules are explored in isolation.

2.2.2 DIAGNOSTIC ANALYTICS

While descriptive analytics reveals insights into the past, it does not necessarily answer the question "why did it happen?" This is exactly what diagnostic analytics aims to achieve. It answers the *why did it happen* question by employing several techniques including data mining and data warehousing techniques. Diagnostic analytics is considered an advanced technique and uses OLAP's roll-up and drill-down techniques (discussed in Section 2.3). Diagnostic analytics is both exploratory in nature and labor-intensive. Diagnostic analytics has been practiced in the education and learning domain for quite some time under the name *diagnostic assessment*. We motivate diagnostic analytics using a few use cases.

2.2.2.1 Diagnostic Analytics Case Studies

Our case studies are drawn from the teaching and learning domain in educational institutions. A range of datasets are used in learning analytics research for improving teaching and learning. The datasets fall into two broad categories—data that is tracked within the learning environments such as learning management systems (LMS), and linked data from the Web. The latter complements learning content and enhances learning experience by drawing upon various connected data sources.

The goal of LinkedUp project [14] is to catalog educationally relevant, freely accessible, linked datasets to promote student learning. In 2014 the LinkedUp project organized the second LAK Data Challenge based on a LAK dataset [15]. The LAK dataset is a structured corpus of full-text of the proceedings of the LAK and educational data mining conferences, and some open access journals. In addition to the full-text, the corpus includes references, and metadata such as authors, titles, affiliations, keywords, and abstracts. The overarching goal of the structured corpus is to advance data-driven, analytics-based research in education and learning.

2.2.2.1.1 Student Success System

Student Success System (S3) is a diagnostic analytics system for identifying and helping at-risk students, developed by Essa and Ayad [16]. Its comprehensive functional capability encompasses descriptive, diagnostic, predictive, and prescriptive analytics. S3 uses both risk analytics and data visualization to achieve its goals. An ensemble of predictive models are used to identify at-risk students. Essa and Ayad's approach is both flexible and scalable for generating predictive models to address significant variability in learning contexts across courses and institutions.

S3 defines a generic measure called *success index*, which is characterized using five subindices—preparation, attendance, participation, completion, and social learning. Each subindex is a composite of a number of activity-tracking variables, which are measured on different scales. These subindices are the basis for applying an ensemble method for predictive modeling.

S3 provides a course instructor with a color-coded lists of students—red for at-risk, yellow for possibly at-risk, and green for not at-risk. The instructor can drill-down to get more details about a student including projected risk at both the course and institution level. Visualizations for diagnostic purposes include risk quadrant, interactive scatter plot, win-loss chart, and sociogram. For example, interactive scatter plot is used to visualize changes in learners' behaviors and performance over time. The *win-loss chart* enables visualizing the performance of a student relative to the entire class based on success indicator measures.

S3 builds a separate predictive model for each aspect of the learning process. Initial domains for the predictive models include attendance, completion, participation, and social learning. Consider the attendance domain. The data collected for this domain encompasses the number of course visits, total time spent, average time spent per session, among others. A simple logistic regression or generalized additive model is appropriate for the attendance domain. In contrast, for the social learning domain, text analytics and social network analysis is required to extract suitable risk factors and success indicators. Therefore a simple logistic regression is inappropriate. Next a stacked generalization strategy is used to combine the individual prediction models using a second-level predictive modeling algorithm.

2.2.2.1.2 COPA

Enhancing learning analytics with students' level of cognitive processing presents new opportunities to gain insights into learning. Gibson, Kitto, and Willis [17] propose COPA, a framework for mapping levels of cognitive engagement into a learning analytics system. More specifically, COPA provides an approach for mapping course objectives, learning activities, and assessment instruments to the six levels represented by the Bloom's category verbs—remember, understand, apply, analyze, evaluate, and create. This entails a flexible structure for linking course objectives to the cognitive demand expected of the learner. The authors demonstrate the utility of COPA to identify key missing elements in the structure of an undergraduate degree program.

2.2.2.1.3 Diagnostic Analytics in Teaching and Learning

Vatrapu et al. [18] proposed a visual analytics-based method for supporting teachers' dynamic diagnostic pedagogical decision-making. They introduce a method called *teaching analytics*, which

draws upon three components named *teaching expert*, *visual analytics expert*, and a *design-based research expert*.

2.2.3 PREDICTIVE ANALYTICS

Based on the past events, a predictive analytics model forecasts what is likely to happen in future. Predictive analytics is critical to knowing about future events well in advance and implementing corrective actions. For example, if predictive analytics reveal no demand for a product line after 5 years, the product can be terminated and perhaps replaced by another one with strong projected market demand. Predictive models are probabilistic in nature. Decision trees and neural networks are popular predictive models, among others. Predictive models are developed using training data.

One aspect of predictive analytics is *feature selection*—determining which variables have maximal predictive value. We describe three techniques for feature selection: correlation coefficient, scatter plots, and linear regression.

Correlation coefficient (r) quantifies the degree to which two variables are related. It is a real number in the range -1 to 1. When $r = 0$, there is no correlation between the variables. There is a strong association between the variables, when r is positive. Lastly, when r is negative, there is an inverse relationship between the variables.

The correlation coefficient for the variables *cyl* and *mpg* is $-.852162$. Though this value of r suggests good negative correlation, the scatter plot (the top plot in Fig. 2.6) indicates that the Cyl is not a good predictor of mpg. For example, the mpg value for a four-cylinder car varies from 22 to 34, instead of being one value.

Shown in the scatter plot is also the superimposed *linear regression line*. The purpose of this line is to predict mpg given the cyl. The slope of the regression line is *negative*, therefore, the correlation between the variables is also *negative*. In other words, when the value of one variable increases, the value of the other decreases. The slope of the regression line can also be *positive*. In that case the association between the variables is *positive*—when the value of one variable increases, the value of the other also increases.

The middle scatter plot in Fig. 2.6 depicts the relationship between the car wt and mpg. Unlike the top scatter plot the points in this plot are generally well aligned along the regression line. The line has negative slope, therefore, correlation between the variables is negative. The r value for the variables is $-.8676594$, which corroborates the negative slope of the regression line.

The bottom scatter plot in Fig. 2.6 depicts the relationship between the cyl and hp. Like the top scatter plot, all the data points are vertically stacked at three places and do not generally align well along the positively-sloped regression line. The r value for the variables *cyl* and *hp* is .8324475. For the same reasons as in the case of the top scatter plot, the cyl is not a good predictor of hp.

In summary, *scatter plots* are considered as part of the standard toolset for both descriptive and predictive analytics. They are different from the linear regression and do not fit lines through the data points.

Simple linear regression is just one technique used for predictive analytics. Other regression models include discrete choice, multinomial logistic, probit, logit, time series, survival analysis, classification and regression trees (CART), and multivariate adaptive regression splines. In addition

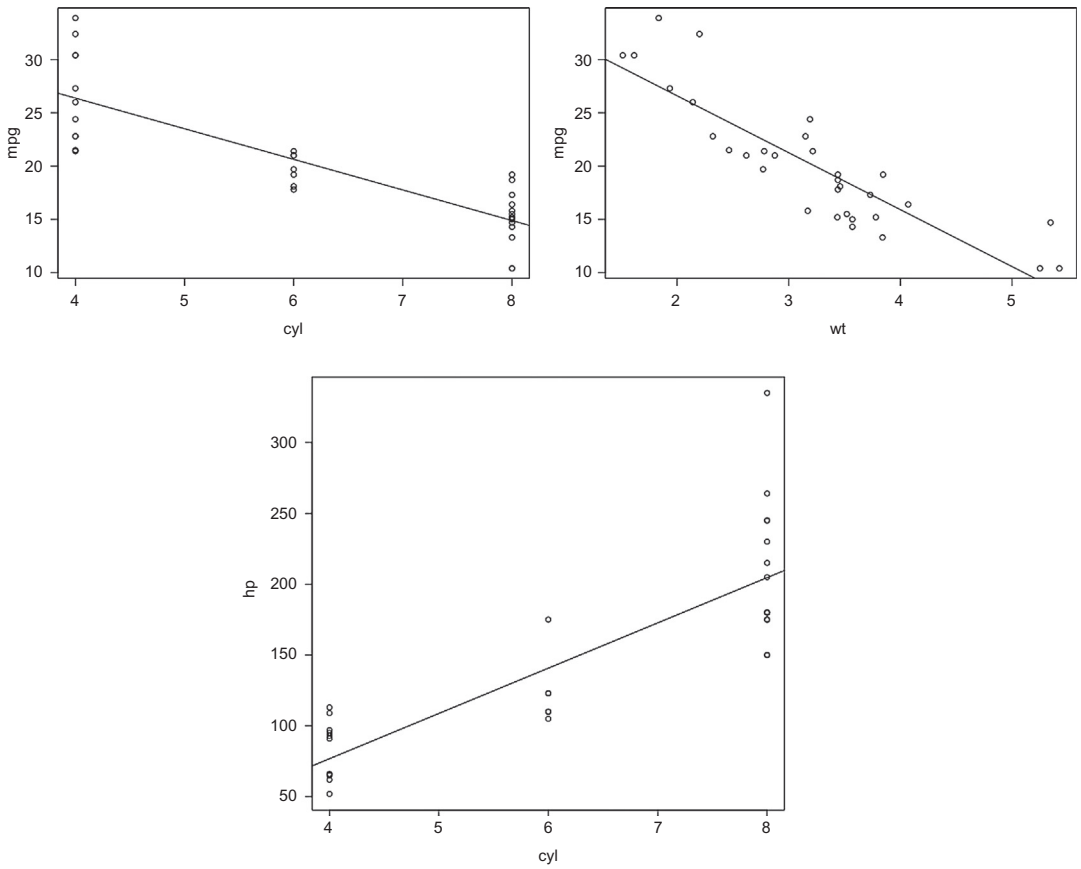

FIGURE 2.6

Scatter plots with linear regression lines superimposed.

to regression, several machine learning algorithms such as naive Bayes, multilayer perceptron, neural networks, radial basis functions, support vector machines, and k-nearest neighbors are also used in predictive analytics.

2.2.3.1 Predictive Analytics Use Cases

Use cases for predictive analytics are numerous. Retail businesses such as Walmart, Amazon, and Netflix critically depend on predictive analytics for a range of activities. For example, predictive analytics helps identify trends in sales based on customer purchase patterns. Predictive analytics is also used to forecast customer behavior and inventory levels. These retailers offer personalized product recommendations by predicting what products the customers are likely to purchase together. Real-time fraud detection and credit scoring applications are driven by predictive analytics, which are central to banking and finance businesses.

2.2.4 PRESCRIPTIVE ANALYTICS

As the name implies, prescriptive analytics helps to address problems revealed by diagnostic analytics. Also prescriptive analytics is used to increase the chance of events forecast by predictive models actually happen. Prescriptive analytics involves modeling and evaluating various what-if scenarios through simulation techniques to answer what should be done to maximize the occurrence of good outcomes while preventing the occurrence of potentially bad outcomes. Stochastic optimization techniques are used to determine how to achieve better outcomes, among others. Also prescriptive analytics draws upon descriptive, diagnostic, and predictive analytics.

Business rules are one important source of data for prescriptive analytics. They encompass best practices, constraints, preferences, and business unit boundaries. Furthermore prescriptive analytics requires software systems that are autonomous, continually aware of their environment, and learn and evolve over time. *Cognitive computing* in general [19], and *cognitive analytics* in particular [20], are needed to implement prescriptive analytics.

Cognitive computing is an emerging, interdisciplinary field. It draws upon *cognitive science*, data science, and an array of computing technologies. There are multiple perspectives on cognitive computing, which are shaped by diverse domain-specific applications and fast evolution of enabling technologies.

Cognitive Science theories provide frameworks to describe models of human cognition. *Cognition* is the process by which an autonomous computing system acquires its knowledge and improves its behavior through senses, thoughts, and experiences. *Cognitive processes* are critical to autonomous systems for their realization and existence.

Data science provides processes and systems to extract and manage knowledge from both structured and unstructured data sources. The data sources are diverse and the data types are heterogeneous. The computing enablers of data science include high-performance distributed computing, big data, information retrieval, machine learning, and natural language understanding.

Cognitive analytics is driven by cognitive computing. Cognitive analytics systems compute multiple answers to a question, and associates a degree of confidence for each answer using probabilistic algorithms. We will revisit cognitive analytics in Section 2.3.5. Because of the inherent complexity and nascency of the field, very few organizations have implemented cognitive analytics.

2.3 EVOLUTION OF DATA ANALYTICS

The genesis of data analytics is in Database Management Systems (DBMS), which are the foundations of today's software applications. The initial version of the Integrated Database Management System (IDMS) was released in 1964, which is considered the first DBMS. IDMS is based on the network (aka CODASYL) data model and runs on mainframe computers. IBM Information Management System (IMS) is another mainframe DBMS, which was released in 1968. IMS is based on the hierarchical data model. Both IDMS and IMS withstood the test of time and continue to be used even today, especially in mission-critical online transaction processing (OLTP) applications.

The mid-1970s ushered in dramatic changes to the DBMS landscape. In 1974 IBM began developing *System R*, a DBMS prototype based on the relational data model [21]. IBM commercialized System R and introduced it under the name SQL/DS in 1981. Oracle Corporation also released its

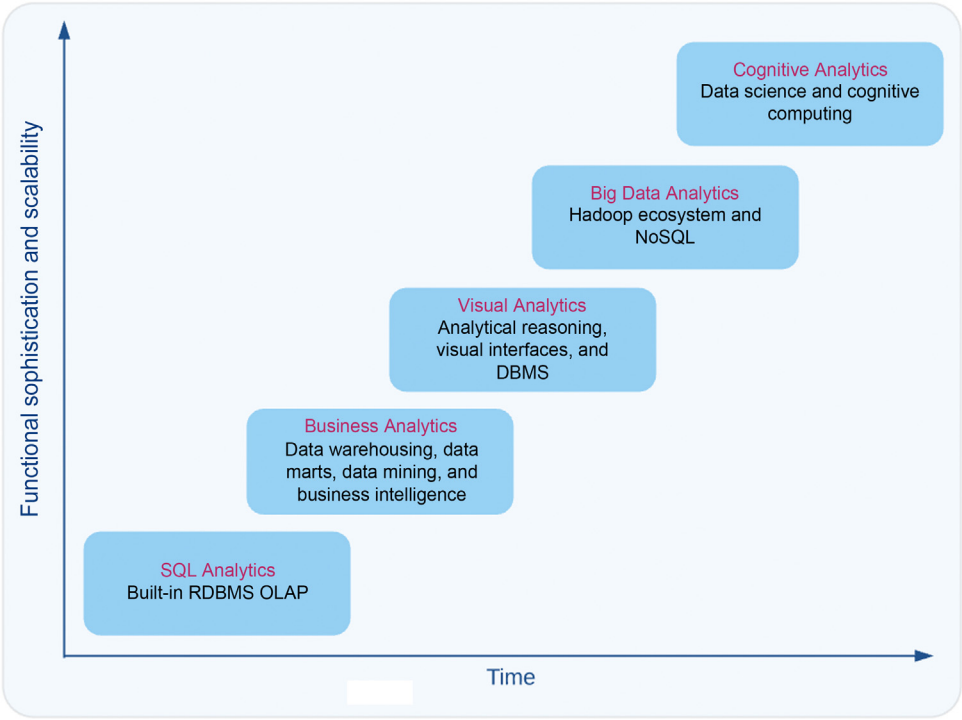

FIGURE 2.7

Evolution of data analytics.

DBMS based on the relational data model in 1979 under the product name *Oracle*. In subsequent years tens of DBMS based on the relational data model followed. These systems are called Relational DBMS (RDBMS) and have become the de facto standard for managing all types of data until recently.

RDBMS have been maintaining their market dominance for over three decades now. The recent emergence of Big Data [22] and NoSQL systems [23] are posing challenges to the long-held RDBMS domination. Fig. 2.7 depicts the evolution of data analytics over the last 35 years.

2.3.1 SQL ANALYTICS: RDBMS, OLTP, AND OLAP

Though the primary focus of RDBMS have been on real-time OLTP, more and more functions for analyzing data have been incrementally introduced under the name *online analytic processing* (OLAP). This is the true beginning of *data analytics* in the era of computers. One might argue that the concept of electronic spreadsheets originated in 1961. However, the first generation electronic spreadsheets were limited to small datasets, and the data was manually entered through keyboards.

The OLAP functions in RDBMS have evolved to a high degree of sophistication over the years [24]. They include an array of mathematical, statistical, and financial libraries to enable

the production of sophisticated reports, answering ad hoc queries, experimenting with what-if queries, and building limited predictive models.

SQL analytics is the set of OLAP functions that are accessible through the SQL database query language. One great advantage of SQL analytics is its performance—computations take place where the data is. However, the types of analytics that can be performed using just the RDBMS data are limited. More useful analytics need data from outside the RDBMS. For example, the analytics needed to develop an intelligent transportation application requires data from connected car networks, traffic signal control systems, weather sensors embedded in roadways, weather prediction models, and traffic prediction and forecasting models. Other issues such as *data cleaning* and *data integration* come to the fore with such external data.

From a SQL query processing standpoint, data organization in the RDBMS for OLTP and OLAP workloads is quite different. The OLTP requires a *row-wise* organization to fetch entire rows efficiently. On the other hand, *column-wise* organization is required for the OLAP workloads. For example, SQL analytics often compute aggregates using mathematical and statistical functions on entire columns. Another issue is the query latency requirements. The OLTP workloads need real-time response, whereas batch processing is tolerated for the SQL-based OLAP tasks.

Given the competing data organization requirements of the OLTP and OLAP tasks, it is difficult to optimize database design to meet the performance and scalability requirements of both. RDBMS practitioners and researchers recognized the need to address the OLAP requirements separately through *data warehousing* and *data marts* technologies. Also the term *business intelligence* was born to differentiate between the RDBMS built-in SQL analytics from the more encompassing *business analytics*. The latter goes beyond the RDBMS data and leverages data mining and machine learning algorithms.

2.3.2 BUSINESS ANALYTICS: BUSINESS INTELLIGENCE, DATA WAREHOUSING, AND DATA MINING

The overarching goal of business analytics is to enable organizations become nimble by responding to changes in the marketplace in a timely manner. Business analytics employs an iterative approach to (1) understanding past business performance, (2) gaining insights into operational efficiency and business processes, (3) forecasting market demand for existing products and services, (4) identifying market opportunities for new products and services, and (5) providing actionable information for strategic decision-making. Business analytics is a set of tools, technologies, and best practices.

2.3.2.1 Business Intelligence

BI is an umbrella term which refers to an assortment of data analysis tasks to help improve business functions and achieve organizational goals. Before the emergence of the term business analytics, the role of BI was similar to that of the descriptive analytics—understanding the past. BI encompasses a range of data sources, technologies, and best practices such as operational databases, data warehouses, data marts, OLAP servers and cubes, data mining, data quality, and data governance. The usage of the term BI is on the decline since 2004, while the usage of the term business analytics has been on a sharp increase beginning 2009. The term BI is being superseded by the term business analytics.

BI's focus has been on the enterprise and provided support for strategic decision-making. It is requirements-based and follows a traditional top-down design approach. Data is primarily structured, and tremendous effort is required for extraction, cleaning, transformation, and loading of data. BI projects are typically reusable. In contrast business analytics focuses on innovation and new opportunities. There are no specific project requirements, and throwaway prototypes are the norm. Bottom-up experimentation and exploration take the center stage. In summary BI was primarily concerned about *what has happened* aspect of the business. On the other hand, business analytics encompasses a broader spectrum by addressing the three questions: what has happened (descriptive analytics), why it has happened (diagnostic analytics), what is likely to happen (predictive analytics), and what should be done to increase the chance of what is likely to happen (prescriptive analytics).

2.3.2.2 Data Warehouses, Star Schema, and OLAP Cubes

Data warehouses are large databases that are specifically designed for OLAP and business analytics workloads. Their data organization is optimized for column-oriented processing. Data is gathered from multiple sources including RDBMS, and is cleaned, transformed, and loaded into the warehouse. The data model used for data warehouses is called the *star schema* or *dimensional model*.

The star schema is characterized by a large *fact* table, to which several smaller *dimensional tables* are attached. Each row in the fact table models an event in the organization. Typically the fact table rows include temporal information about the events such as the order date. Consider a retailer such as Walmart and its data warehousing requirements. The dimensions for the star schema may include a geographic region (e.g., northeast, midwest, southeast, and northwest), time (calendar year quarters 1, 2, 3, and 4), and item category (electronics, garden and outdoor, and home fashion).

Fig. 2.8 shows a star schema for a healthcare data warehouse. Shown at the center is the fact table that has a large number of attributes. There are 8 dimension tables—gender code, race code, year, admission sources, pay sources, and others.

The star schema enables the generation of multidimensional OLAP cubes, which can be *sliced* and *diced* to examine the data at various levels of detail across the dimensions. The term cube is synonymous with *hypercube* and *multicube*. We limit our discussion to three dimensions for the ease of exposition.

OLAP summarizes information into multidimensional views and hierarchies to enable users quick access to information. OLAP queries are generally compute-intensive and place greater demands on computing resources. To guarantee good performance, OLAP queries are run and summaries are generated a priori. Precomputed summaries are called *aggregations*.

Consider a cube whose dimensions are *geographic region*, *time*, and *item category*. Data and business analysts use such cubes to explore sales trends. For example, the cell at the intersection of a specified value for each dimension represents the corresponding sales amount. For example, the cell at the intersection of *midwest* for the geographic region dimension, *quarter 4* for the time dimension, and *electronics* for the item category dimension denotes electronic products sales revenues in the fourth quarter. It is also possible to have finer granularity for the dimensions. For instance, quarter 4 can be subdivided into the constituent months—October, November, and December. Likewise the more granular dimensions for geographic region comprise the individual states within that region.

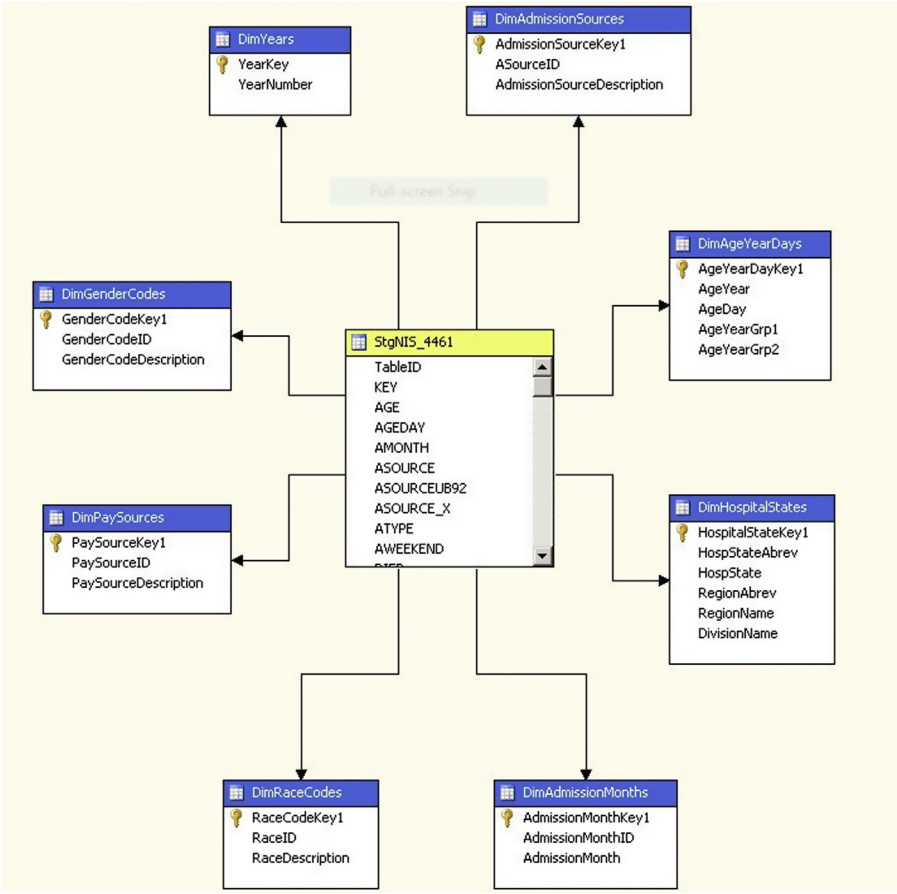

FIGURE 2.8

A star schema.

The structure of the OLAP cube lends itself to interactive exploration through the *drill-down* and *roll-up* operations. Fig. 2.9 depicts an OLAP cube showing an aggregation. The OLAP view in Fig. 2.10 shows a drill-down detail.

2.3.2.3 ETL Tools

Extract, Transform, and Load (ETL) are a set of tools for designing, developing, and operating data warehouses. A data warehouse development is a resource-intensive activity in terms of both people and computing infrastructure. Identifying, cleaning, extracting, and integrating relevant data from multiple sources is a tedious and manual process even with ETL tools. Some organizations build just one comprehensive data warehouse, which is called the *enterprise data warehouse*. In contrast others take the *data mart* approach. Data marts are mini data warehouses with limited

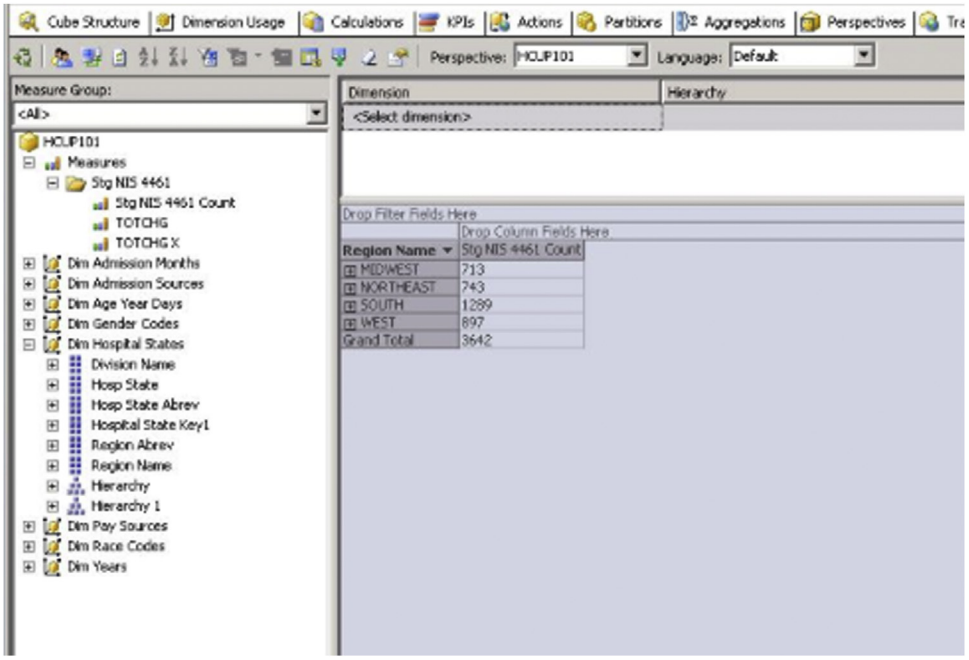

FIGURE 2.9

A OLAP cube view showing a roll-up aggregation.

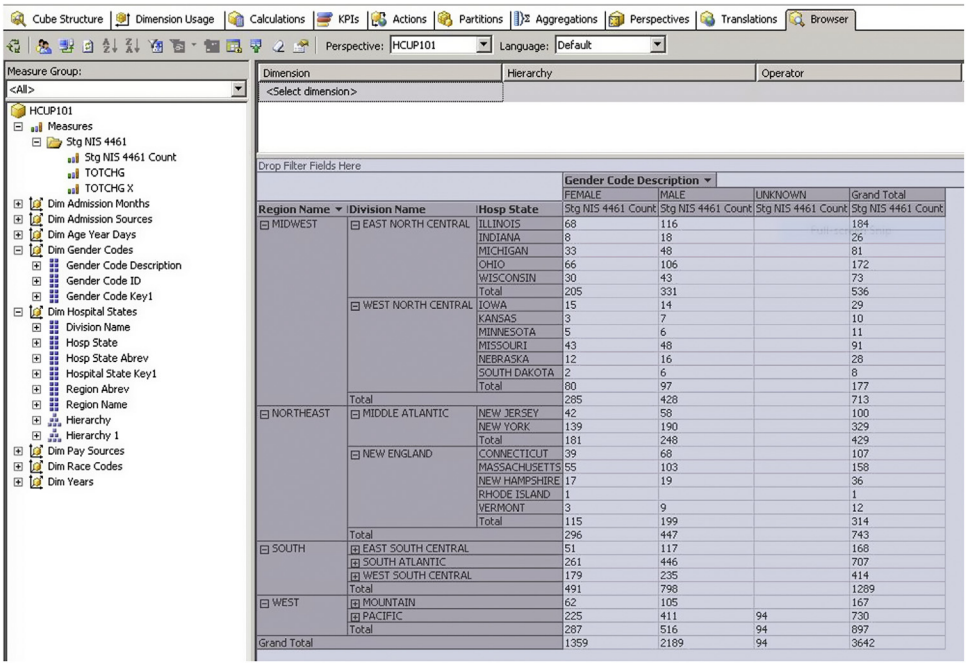

FIGURE 2.10

A OLAP cube view showing a drill-down detail.

scope, often serving the needs of a single department. Data marts are also constructed from an existing enterprise data warehouse. Typically both data warehouses and data marts are implemented using an RDBMS.

2.3.2.4 OLAP Servers

Once the data warehouses and data marts are constructed, they are accessed by various clients including query and reporting, analysis, and data mining tools. From the client tools' perspective, SQL is a low-level language to access data warehouses and data marts. *OLAP servers* remove this shortcoming by providing a higher-level data access abstraction in the form of an OLAP multidimensional cube with roll-up and drill-down operations. OLAP servers act as intermediaries between the data warehouses and the client tools. As noted earlier, OLAP cubes also provide a performance advantage.

OLAP servers are implemented using one of four approaches—relational OLAP (ROLAP), multidimensional OLAP (MOLAP), Hybrid OLAP (HOLAP), and specialized SQL DBMS. A ROLAP server is an extended RDBMS, which maps operations on the multidimensional cubes to standard SQL operations. The latter are executed by the underlying RDBMS. This contributes to scalability of ROLAP servers and avoids data redundancy.

A MOLAP is a special-purpose server, which directly implements multidimensional data through array-based multidimensional storage engines. Two copies of the data exists—one in the RDBMS and another in the multidimensional storage engines. Array-based storage engines enable precomputing summarized data using fast indexing. Relative to ROLAP, MOLAP approaches do not scale well.

The hybrid OLAP combines ROLAP and MOLAP by taking advantage of the scalability of ROLAP and the faster computations of MOLAP. Finally, specialized SQL servers provide query languages that are specifically designed for the star schema. They natively support roll-up and drill-down operations.

The data analytics functions provided by the OLAP servers are useful for meeting reporting requirements, enabling EDA, identifying opportunities for improving business processes, and assessing the performance of business units. Typically business analytics requires significant human involvement. The advent of Big Data and attendant NoSQL systems [25] coupled with near real-time applications overshadowing batch systems, the critical role of data warehouses and data marts is diminishing.

2.3.2.5 Data Mining

Data mining goes beyond OLAP's drill-down and roll-up features. It enables automatic extraction of actionable insights from data warehouses and data marts by discovering correlations and patterns hidden in the data. Such patterns may be used for purposes such as improving road traffic by reducing congestion, providing superior customer support, reducing the number of defects in the shipped products, increasing revenues and cutting costs. Data mining is typically performed on the data which resides in the warehouses. However, it can also be performed on data kept in flat files and other storage structures.

As noted earlier, data goes through cleaning, transformation, and integration steps before mining can be performed. Preparing data for mining is an important step, which is both tedious and

difficult. This aspect will be discussed in Section 2.4. Essentially data mining involves finding frequent patterns, associations, and correlations among data elements using machine learning algorithms [26].

Frequent patterns include itemsets, subsequences, and substructures. A frequent *itemset* refers to a set of items that frequently appear together in a grocery store sales receipt, for example. Such insight is useful for understanding customers' purchase patterns and their variation over the year. This in turn helps to plan inventory and to promote customer loyalty by issuing relevant coupons. In the case of ITS, if two types of undesirable traffic events seem to occur concurrently and frequently, such information can be used to design effective controls to reduce their occurrence.

A frequent *subsequence* refers to frequently observing in the dataset scenarios such as buying a house first, home insurance next, and finally furniture. Unlike the itemset, the purchases in the subsequence are temporally spaced. Knowing the frequent subsequences of customers will help to execute a targeted marketing campaign. An ITS example in this case is temporally spaced traffic jams caused by a single accident in a road network. Information about such frequent subsequences is used to implement just-in-time traffic rerouting.

A *substructure* refers to structural forms such as trees and graphs. Mining frequent subgraph patterns has applications in biology, chemistry, and web search. For example, chemical compounds structures and Web browsing history can be naturally modeled and analyzed as graphs. Finding recurring substructures in graphs is referred to as *graph mining*. Graph mining applications include discovering frequent molecular structures, finding strongly connected groups in social networks, and web document classification.

Mining frequent patterns helps to reveal interesting relationships and correlations among the data items. As an example, consider the following association rule: *act-math-score*(X, "29−34") $\wedge$ *ap-courses-in-high-school* (X, "4−8")) $\Rightarrow$ *success*(X, cs-major) [support = 15%, confidence = 75%]. This rule states that a student X whose ACT math score is in the range (29−34) and completed 4−8 AP courses in high school will succeed as a computer science major in college. And support = 15% means that 15% of the records in the dataset met the conditions in the rule. Lastly confidence = 75% states that the probability of future students who meet the rule conditions will succeed as computer science majors is .75.

The other data mining tasks include classification, cluster analysis, outlier analysis, and evolution analysis.

The classification problem involves assigning a new object instance to one of the predefined classes. The system that does this job is known as the classifier, which typically evolves through learning from a set of training data examples. The classifier is represented using formalisms such as *if-then* rules, *decision trees*, and *neural networks*.

Consider the task of recognizing handwritten zip codes as a classification problem. Each handwritten digit is represented by a two-dimensional array of pixels and *features* such as the following are extracted for each digit: the number of strokes, average distance from the image center, aspect ratio, percent of pixels above horizontal half point, and percent of pixels to right of vertical half point. These features of a digit are assembled into a structure called the *feature vector*.

The *if-then* rules, *decision trees*, and *neural networks* are developed using the feature vectors. The features are manually identified. In some problem domains, a large number of features are available and a *feature selection* task determines a subset of the features, which have significance for the classification task. Though the features vary from one domain to another, the process of training and validating the classifier using feature vectors is domain independent.

Clustering (aka *cluster analysis*) is the problem of nonoverlapping partitioning of a set of n objects into m classes. The number of classes, m, is not known a priori. The objects are grouped into clusters based on the principle of *maximizing the interclass similarity* while *minimizing the intraclass similarity*. In other words objects within a class are cohesive and similar to each other, whereas the objects in one cluster are quite dissimilar to objects in another cluster. Like the classification algorithms, clustering algorithms also use feature vectors.

In statistical analysis outliers are data points that are drastically different from the rest. We have seen in Section 2.2.1.3 how a boxplot is used to detect outliers manually. Usually many statistical analyses methods discard outliers. However, in the data mining context, outliers are the data points of interest as they reveal, e.g., fraudulent credit card transactions for a bank, and security breaches for a surveillance system. Outliers can be detected using statistical tests, which assume a distribution for the data. They can also be detected using distance measures. If a data point is above a certain threshold distance from all the clusters, the point is considered an outlier.

Evolution analysis models show how object behaviors change over time. For example, such an analysis will help to spot regularities and trends in time-series data such as the stock market prices. Knowledge of such patterns can be leveraged to predict stock market trends and to make stock market investment decisions. In the context of ITS, knowledge of how traffic patterns and volumes have been evolving over time will help in capacity planning of roadways and devising traffic decongestion measures.

We conclude this section with a graph data mining application. Consider the Web document classification problem. We outline an approach to solve this problem using graph data mining. First we extract all the unique words in a set of Web documents, remove commonly occurring grammatical function words known as *stop words* (e.g., words such as a, an, the, or, and), reduce the alternative forms of words into their most frequently occurring form using lemmatization techniques, and keep only the words whose frequency of occurrence is greater than a specified threshold. Second, designate each word as a vertex in the graph. If a word w_2 follows the word w_1, create an edge between the vertices corresponding to w_1 and w_2. Edges are labeled based on the sections of the document in which the words w_1 and w_2 occur. Third, mine the graph for determining frequent subgraphs, and call the subgraphs *terms*. Fourth, use a *term frequency* (tf) and *inverse document frequency* (idf) based measure to assign characteristic terms to documents. Lastly, use a *clustering* algorithm and a k-nearest neighbors classification algorithm to classify the Web documents.

2.3.3 VISUAL ANALYTICS

Visual analytics is a new interdisciplinary field [27]. It draws upon the following areas: analytical reasoning; planning and decision-making; visual representations; interaction techniques; data representations and transformations; production, presentation, and dissemination of analytical results. Analytical reasoning methods enable users to gain insights that directly support assessment. The next three techniques—planning and decision-making, visual representations, and interaction techniques—leverage the high bandwidth connection between the human eye and mind to explore and understand large amounts of information in parallel mode. Data representations and transformations methods transform conflicting and dynamic data into forms that support visualization and analysis. Lastly, techniques for production, presentation, and dissemination of analytical results are used to effectively communicate information to a range of audience.

Visual analytics combines the computational prowess of computers with the robust decision-making capability of humans. It enables efficient analysis of large data by combining data analytics with visual devices and interaction techniques. Visual analytics applications that are based on large, time-evolving graphs include real-time information-based evacuation decision support system for emergency management, and forecasting influenza, among others.

The data used in visual analytics may come from disparate data sources including RDBMS, NoSQL systems, data warehouses, data marts, and flat files. Many visual analytics software vendors provide data connectors to these sources.

Visual analytics software products are available from several vendors including SAS, Tableau, IBM, Microsoft, and Qlik. They vary widely in terms of functional features and scalability. Some products seamlessly integrate query, exploration, and visualization processes and also automatically recommend effective visualizations based on context. Others encourage *visual thinking* by augmenting human perception and enable *visual perspective shifting and linking*—easily switch among alternative visualizations and multiple related visualizations are semantically linked. Other considerations include the range of visualizations supported and multidimensional expressiveness. Lastly, collaborative visualization, which enables multiple, geographically dispersed users to collaboratively develop and share visualizations.

2.3.4 BIG DATA ANALYTICS

The goals for big data analytics are similar to the other data analytics techniques we discussed earlier. However, several issues make big data analytics quite challenging [28,29]. First, is the issue of data heterogeneity and attendant complex data types. The unprecedented data volumes and the speed at which the data is generated are the other issues. This calls for specialized computing hardware and software. Data quality issues are also more pronounced in the big data context. From a practical standpoint, scalability and performance are defining success factors [30].

Analytics prior to big data primarily dealt with structured data nestled in the relational data model. In contrast big data analytics encompasses unstructured data in the form of natural language text, short twitter tweets, and multimedia data such as audio clips, images, and video. Feature extraction is another challenge. In many cases, the *feature selection* task itself is extremely difficult. Furthermore the knowledge and skills needed to analyze and interpret terabyte- or petabyte-sized datasets are quite different from those for small-scale datasets.

On the positive side, big data is interesting because it has the potential to reveal emergent phenomena that do not manifest in small- and medium-scale data. This is where the EDA and visualization serve as preliminary tools to gain insight into big data. A number of high-performance computational infrastructures are available for big data analytics. The Hadoop ecosystem and NoSQL databases are the popular choices. We will revisit big data analytics related issues in Section 2.4.

2.3.5 COGNITIVE ANALYTICS

Cognitive analytics is the natural evolution and merger of visual analytics and big data analytics. Cognitive analytics removes humans from the loop. It replaces human involvement with cognitive agents

whose goal is to mimic human cognitive functions. Cognitive analytics draws upon advances in several areas but the primary contributions are from the computing and the cognitive science disciplines.

Cognitive analytics systems extract features from structured, semistructured, and unstructured data. They also employ taxonomies and ontologies to enable reasoning and inference. Cognitive analytics systems extract both low-level features and high-level information from the data. These systems also rely on an assortment of machine learning algorithms and inference engines to realize human-like cognitive capabilities. Learning and adaptation are integral to cognitive analytics.

Cognition endows autonomous systems such as the self-driving cars the ability to self-regulate themselves, and acquire knowledge from their environments including situational awareness. Learning, development, and evolution are integral to the existence and functioning of autonomous systems, and are realized through *cognitive processes*. Cognitive science provides theories, methods, and tools to model cognitive processes [31].

A *cognitive architecture* is a blueprint for developing cognitive systems. It specifies fixed structures as well as interactions among them toward the goal of achieving intelligent behavior like humans. A *cognitive model* focuses on a single cognitive process such as language understanding for spoken language interfaces. A cognitive architecture may support multiple cognitive models. Lastly, a *cognitive computing system* provides the necessary computing infrastructure for developing and deploying cognitive systems.

In contrast with all other analytics we have seen so far, cognitive analytics generates multiple answers for a question and associates a degree of confidence measure to each answer. Cognitive analytics systems rely on probabilistic algorithms to come up with multiple answers with varying degrees of relevance. Each answer corresponds to a hypothesis. Evidence is gathered in support of each hypothesis and is used to score the relevance of a hypothesis.

Some aspects of big data analytics and cognitive analytics overlap with data science, which is discussed in the next section. *Data science* provides the framework for solving problems using cognitive analytics.

2.4 DATA SCIENCE

Data Science is a new interdisciplinary domain whose goal is to provide data-driven solutions to difficult problems. The latter are ill-posed for precise algorithmic solutions. Such problems abound in natural language understanding and autonomous vehicle navigation. Data science provides solutions to problems by using probabilistic and machine learning algorithms. Often multiple solutions to a problem are provided and a degree of confidence is associated with each solution. The higher the degree of confidence, the greater is the relevance of the solution to the problem. This approach is a natural consequence of working with data, which may be incomplete, inconsistent, ambiguous, and uncertain. The data science approach closely reflects the way humans solve problems, which are difficult to characterize algorithmically—obstacle detection and avoidance in self-driving cars.

The emergence of big data is the genesis of the data science discipline. Big data enables scientists to overcome problems associated with small data samples. With big enough data, (1) certain assumptions of theoretical models can be relaxed, (2) over-fitting of predictive models to *training data* can be avoided, (3) noisy data can be effectively dealt with, and (4) models can be validated with ample *test data*.

Halevy, Norvig, and Pereira [32] argue that the accurate selection of a statistical model is not critical if it is compensated by *big enough* data for the model training and validation. In other words ill-posed problems are amenable to solution by managing the complexity of the problem domain through building simple but high-quality models by harnessing the power of big data. For example, a true random and additive noise can be eliminated through *image averaging*. The quality of an image can be dramatically improved by averaging a sequence of successive images of the same object. The more the number of images used in averaging, the greater is the visual quality of the resulting image. This technique is routinely used with satellite imagery.

Like computer science, data science is a scientific discipline and also an engineering discipline. It is a scientific discipline because it uses an experiment-oriented scientific approach. Based on empirical evidence, a hypothesis is formulated, and evidence is gathered to perform the hypothesis testing. It is an engineering discipline since it is used to develop and deploy practical systems such as autonomous navigation vehicles and cognitive assistants. This is where cognitive science plays a natural role in complementing data science. Data science deals with data issues, experimental design, and hypothesis testing [33], whereas cognitive science provides theories, methods, and tools to model cognitive tasks [31]. More specifically, cognitive science provides frameworks to describe various models of human cognition including how information is represented and processed by the brain.

There are multiple perspectives on the data science. The computing perspective deals with the computational issues of storing and retrieving big data, authentication and authorization, security, privacy, and provenance issues. It provides high performance, scalable computing infrastructure to perform dataflows. The mathematical perspective is concerned with experimental design, hypothesis testing, inference, and prediction. The scientific perspective encompasses empirical observations, posing interesting questions, visualizing, and interpreting the results. In summary data science is about EDA, discovering patterns, building and validating models, and making predictions [1]. In the rest of this section, we discuss various aspects of data science.

2.4.1 DATA LIFECYCLE

The data lifecycle describes the various processes that data goes through from its inception to end of life. The data lifecycle stages are (1) generation, (2) acquisition, (3) cleaning, (4) sampling, (5) feature extraction, and (6) model development. Many data analytics projects enhance in-house data with data acquired from third-party vendors. Before selecting data vendors, issues that arise in integrating in-house data with the vendor data should be investigated. Acquisition is concerned with identifying data vendors, examining data formats, resolving licensing issues, and selecting vendors. Vendors may have special restrictions on how the data should be used and royalties on embedding data in downstream applications.

The effort involved in the cleaning stage can differ widely. Cleaning refers to a broad range of activities including duplicate detection, data imputation, outlier detection, resolving conflicting and ambiguous data, and establishing procedures for record linking (i.e., associating related data for an object from multiple sources).

Data sampling refers to subsetting the data for the next stage—feature extraction. Since feature extraction is a computationally intensive task, an important decision is determining how much data be used for feature extraction. Various sampling methods are available including simple random

sampling, stratified sampling, systematic sampling, cluster sampling, and probability-proportional-to-size sampling, among others.

Each observation in the data is characterized by multiple attributes (aka features). For example, a historical airline flights database may record the following information about each flight: scheduled departure time, actual departure time, scheduled arrival time, actual arrival time, departure airport code, arrival airport code, flight duration, and day of the week. Some of these attributes may not be independent of each other. For model building, such attributes do not add value but make the model building task more difficult—*dimensionality curse*. As the number of dimensions increases, so does the dramatic increase in the volume of space. This creates data sparsity and requires additional data to obtain a statistically significant result. Unfortunately the amount of data needed to support statistical significance often grows exponentially with the dimensionality. *Feature selection* refers to determining a subset of the attributes that are maximally effective for model building.

Once feature selection is made, the next step is to extract features from each observation. If the number of features selected is n, the feature vector will also be an n-dimensional vector of numeric values. Assuming that the data is comprised of m observations, the result of feature extraction is a list of m feature vectors, each of which is an n-dimensional vector. Often features are extracted in batch mode using a computational framework such as Hadoop (see Section 2.5). Model building is discussed in Section 2.4.3.

2.4.2 DATA QUALITY

The critical role of data quality in decision-making and strategic planning precedes the computing era. Data quality research is characterized by multiple facets including data constraints, data integration and warehousing issues, data cleaning, frameworks and standards, data quality metrics, among others [34]. Facets may have subfacets. For example, data cleaning encompass error detection, outlier detection, record linking, and data imputation.

Some of the data quality problems are easy to solve, whereas solutions to problems such as record linking remains elusive. The record linking problem involves associating different pieces of data about an entity in the absence of a unique identifying attribute. Another source of data quality problems stems from data transformation. As data goes through various transformations, data quality errors propagate and accumulate.

RDBMS use integrity constraints (ICs) as a primary means for enforcing data quality. ICs are specified declaratively. If ICs are not expressive enough, *triggers* is used to specifies constraints procedurally.

Quantitative assessments of data quality measure the severity and extent of data defects. An approach to assessing data quality in existing databases as discussed in [35] requires three major tasks: static analysis of the database schema, measurement of the complexity of database structure, and validation of the database content. Static *database schema analysis* performs checks such as the use of vendor-specific features, and violation of database design normal forms. Three types of metrics are proposed for measuring the *complexity of database structure*: size, complexity, and quality. Lastly, *validating the database content* requires verifying that the existing data passes assertions stated using predicate logic. Assertions capture data quality semantics that are not expressible using the ICs.

2.4.3 BUILDING AND EVALUATING MODELS

We illustrate model building and validation for the classification problem using the airline on-time performance dataset [36,37]. It consists of flight arrival and departure details for all commercial flights within the United States, from October 1987 to April 2008. There are nearly 120 million records in the dataset. The dataset in compressed form is 1.6 gigabytes, and the same in uncompressed form is 12 gigabytes. For the year 2007 alone, the dataset has information about 7,453,216 flights. Each flight record is comprised of 29 variables. Some of these variables are year, month, day of the month, day of week, actual departure time, scheduled departure time, actual arrival time, scheduled arrival time, unique carrier code, flight number, origin airport, and destination airport. The dataset will be enhanced with supplemental data about airports, carrier codes, planes, and weather.

A subset of the 2007 data is used for EDA using RStudio and R software. This analysis aims to answer questions such as (1) Which airport has the maximum number of delays? (2) Which airline experienced the maximum number of delays? (3) Do older planes experience more delays than the newer planes? (4) Does the weather play a role in the delays? (5) Do delays in one airport contribute to delays in other airports?

The next step is to build a predictive model. But first we need to create the feature matrix. We are not going to use all 29 variables and use *feature selection* to select only those variables that have more predictive power. For example, it is natural to suspect that the winter months may contribute to more delays than summer months. Therefore we keep the *month* variable. Some airports like Chicago and Atlanta experience more delays due to heavy air traffic. Following this line of reasoning, we also select day of the month, day of the week, hour, flight distance, days from holiday. The last variable is the number of days from the nearest holiday. For example, if the departure date is July 2, days from the holiday is 2 (since July 4th is the nearest holiday). All the six variables/features are numeric.

All the flights departing from Chicago in 2007 will be used to train the model. The same data for the year 2008 will be used to test and validate the model. To make the following exposition more concrete, we quote the analysis results reported in [37].

The training data consists of 359,169 flights, and 6 features characterize each flight. Next Python's *Scikit-learn* machine learning package is used to develop two predictive models: logistic regression with *L2* regularization, and random forest. To compare the performance of predictive models built in successive iterations, we need to define some metrics. First, we define the *confusion matrix*, which is shown in Table 2.2.

Table 2.2 Structure of a Confusion Matrix

		Predicted Value		
		Positive	**Negative**	**Row Total**
Actual value	Positive	True Positive (TP)	False Negative (FN)	P'
	Negative	False Positive (FP)	True Negative (TN)	N'
	Column Total	P	N	

Consider the cell in the table labeled "True Positive (TP)." Given feature vectors comprised of six variables as input, the value in this cell indicates the number of times the model predicted delays (i.e., predicted values are positive) for the flights and the flights are actually delayed (i.e., actual values are positive). A bigger value for this cell is better. The value in the cell labeled "False Positive (FP)" indicates the number times the model predicted delays when actually there were no delays. Smaller values are better for this cell. Next the value in cell labeled "False Negative (FN)" indicates the number of times the model did not predict delays when there were delays. A smaller number for this cell is desirable. Lastly, the value in the cell labeled "True Negative (TN)" indicates the number of times the model did not predict delays when there were no delays. Here again a small number is better.

Let $P = TP + FP$ and $N = FN + TN$. We also define four metrics—PPV, TPR, ACC, and F1—as follows:

$$\text{Precision or positive predictive value (PPV)} = \frac{TP}{TP + FP} \tag{2.1}$$

$$\text{Recall (sensitivity or true positive rate (TPR))} = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{Accuracy (ACC)} = \frac{TP + TN}{P + N} \tag{2.3}$$

$$\text{F1 score (harmonic mean of precision and sensitivity)} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.4}$$

The confusion matrix for logistic regression model with $L2$ regularization is shown in Table 2.3. Using these values, the following metrics evaluate to: precision (PPV) = 0.38, recall (TPR) = 0.61, accuracy (ACC) = 0.60, and F1 = 0.47.

Next we try a different model for flight prediction delays. A random forest classifier with 50 trees is developed. The metrics for the random forest model evaluate to: precision (PPV) = 0.41, recall (TPR) = 0.31, accuracy (ACC) = 0.68, and F1 = 0.35. Compared to the logistic regression model, accuracy is increased for the random forest model, but the F1 score decreased. Since the model is designed for predicting flight delays, accuracy is a more relevant measure.

Sometimes it is desirable to turn *categorical variables* into binary ones to indicate simply whether or not a feature is present. Even other variables such as *departure delay* can be mapped to binary values. Assume that the values for the variable *departure delay* are in the range 0−20.

Table 2.3 Confusion Matrix of the Logistic Regression Model With L2 Regularization

		Predicted Value		
		Positive	**Negative**	**Row Total**
Actual value	Positive	58,449 (TP)	36,987 (FN)	95,436 (P')
	Negative	96,036 (FP)	143,858 (TN)	239,894 (N')
	Column Total	154,485 (P)	180,858 (N)	

Select a threshold, for instance, 10. If a value is less than 10, assign 0 as the new value of the variable, otherwise, assign 1.

In the next iteration a random forest model is developed after converting both categorical variables as well as those which are categorical in nature to binary variables. This increases the dimension of feature vectors from 6 to 409. The metrics for this new random forest model are: precision (PPV) = 0.46, recall (TPR) = 0.21, accuracy (ACC) = 0.71, and F1 = 0.29. The accuracy measure has improved from 0.68 from the previous iteration to 0.71.

In the last iteration another random forest model is developed after enhancing the dataset with weather data. This increases the feature vector dimensionality from 409 to 414. Also, the number of trees in the random forest is increased to 100. The metrics for this new random forest model evaluate to: precision (PPV) = 0.63, recall (TPR) = 0.23, accuracy (ACC) = 0.74, and F1 = 0.34. The accuracy measure has improved from the previous iteration value of 0.71 to 0.74.

In summary *feature selection* is both a science and an art. EDA on a carefully sampled dataset will help to gain insights into which variables may hold greater predictive value. Also there are many other machine learning algorithms that we can use for the airline delay prediction problem. Note that some algorithms may perform very well in one domain, but the same algorithms may perform poorly in other domains.

2.5 TOOLS AND RESOURCES FOR DATA ANALYTICS

Data analytics needs lots of data, and high-performance, highly-parallel computing infrastructure. In the era of big data, there is data everywhere. However, high-quality data commands a premium price. On the other hand cloud-hosted computing infrastructures such as the Amazon Web Services (AWS) is relatively inexpensive.

1. *Data resources for written and spoken language analytics*: The *World Atlas of Language Structures* (WALS) is a large database of phonological, grammatical, lexical properties of languages gathered from descriptive materials such as reference grammars [38].

   WordNet is a lexical database for English [39].

   The *one-billion word benchmark* for measuring progress in statistical language modeling is available at Ref. [40].
2. *General data sources*: The UC Irvine Machine Learning Repository maintains 350 datasets for promoting machine learning research [41].

   Data.gov is a US government open data initiative [42]. Currently over 185,675 datasets are available for public use.

   Over 50,000 free ebooks are available at Project Gutenberg [43].

   Other resources such as data.stackexchange.com, DBpedia, LinkedIn, and ResearchGate provide programmatic access to their data. For example, one can submit parameterized SQL queries to data.stackexchange.com and retrieve highly targeted data. Data usage restrictions, cost, and license types vary.
3. *Software libraries and tools*: Organizations are competing to make their software and tools open-sourced, but not the data. For example, Google recently open-sourced TensorFlow, an

open source software library for machine intelligence [44]. Google also open-sourced SyntaxNet, an open-source neural network framework for developing Natural Language Understanding (NLU) systems [45]. SyntaxNet is implemented in TensorFlow.

Weka 3 is a collection of Java machine learning algorithms for data mining tasks [46].

The Apache UIMA project provides open source frameworks, tools, and annotators for facilitating the analysis of unstructured content such as text, audio, and video [47].

R is a celebrated open source language environment for statistical computing, graphics, and visualization [6]. R provides a wide range of statistical tools for linear and nonlinear modeling, classification, clustering, time-series analysis, classical statistical tests, among others. R is highly extensible and thousands of packages are available to meet domain-specific needs.

Due to the extreme popularity of R, data access connectors are available to access RDBMS and other data sources from R. One such package is dplyr, which is more than a database connector. It provides data manipulation syntax, translates queries written in dplyr to SQL, executes SQL against the database, and fetches query execution results into the R environment. dplyr uses lazy evaluation—computation of the results is delayed until the results are needed—and streams results. The databases that dplyr connects to include SQlite, MySQL, MariadB, PostgreSQL, and Google BigQuery. Other tools that are similar in functionality to R are SparkR, RHadoop, RJDBC, RODBC, and DBI.

Hadoop is the most widely used open source computing platform for big data processing and cognitive analytics. Hadoop's primary components are the Hadoop Distributed File System (HDFS), Hadoop MapReduce—a high-performance parallel data processing engine, and various tools for specific tasks. The HDFS, MapReduce, and the tools are collectively referred to as the *Hadoop ecosystem*. Tools include Pig, Pig Latin, Hive, Cascading, Scalding, and Cascalog.

Pig is a declarative language to specify dataflows to ETL data and compute analytics. Pig generates MapReduce jobs which execute the dataflows. Pig provides a higher-level abstraction to MapReduce. The *Pig Latin* enhances Pig through a programming language extension by providing common data manipulation operations such as grouping, joining, and filtering. *Hive* provides SQL-based data warehouse functions for large datasets stored in HDFS-compatible file systems.

Cascading is a popular, high-level Java API for MapReduce programming. It effectively hides many of the complexities of MapReduce programming. *Scalding* and Cascalog are even higher-level and concise APIs compared to Cascading. Scalding enhances Cascading with matrix algebra libraries, whereas Cascalog adds logic programming constructs. Scalding and Cascalog are accessible from the Scala and Clojure programming languages, respectively.

Storm and Spark are the Hadoop ecosystem tools for event stream processing, and real-time stream data processing, respectively. Storm features an array of spouts specialized for receiving streaming data from disparate data sources. Also it enables incremental computation, and computing metrics on rolling data windows in real-time. Spark features several additional libraries for database access, graph algorithms, and machine learning.

Apache Tez is a new distributed execution framework for executing analytics jobs on Hadoop. Tez enables expressing computations as a dataflow graph, which is a higher-level abstraction compared to MapReduce. More importantly Tez eliminates storing intermediate results to HDFS by directly feeding the output of one process as input to the next process. This gives a tremendous performance advantage to Tez over MapReduce.

Python is a popular programming language for scientific computing and cognitive analytics. Pydoop is a Python interface to Hadoop, which enables writing MapReduce applications in pure Python. Python and Matplotlib are ideal tools for EDA.

AWS Elastic MapReduce (EMR) is a cloud-hosted commercial Hadoop Ecosystem from Amazon. Microsoft's StreamInsight is another commercial product for stream data processing with special focus on complex event processing.

bigml.com provides an assortment of machine learning algorithms and datasets (https://bigml.com/) for a fee. The algorithms are available through a REST API.

Cloudera and Hortonworks provide Hadoop and its ecosystem components as open source Platform-as-a-Service (PaaS) distribution. An advantage of Hadoop PaaS is a diminished learning curve and product support can be purchased for a fee. Product support is critical for mission-critical, enterprise deployments. Cloudera and Hortonworks are two companies that provide support and consulting services for Hadoop-based big data analytics.

Cloudera Data Hub (CDH) is Cloudera's open source Apache Hadoop platform distribution [48]. It includes Hadoop and related components in its ecosystem. End-to-end big data workflows can be executed using the CDH. The *QuickStarts* is a single-node CDH which runs in a virtual machine (VM) environment. It eliminates the complexities of installing and configuring multiple components in the Hadoop ecosystem. QuickStarts is ideal for personal learning environments.

Hortonworks Data Platform (HDP) is an open source Hadoop platform for developing and operating Big Data analytics and data-intensive applications [49]. Like Cloudera, Hortonworks also provides a single-node HDP which runs in a VM environment.

2.6 FUTURE DIRECTIONS

Though the majority of current big data analytics projects are executed in batch mode, the trend will be toward near real-time to real-time processing. For some data analytics applications such as fraud and anomaly detection, spotting cybersecurity attacks, and identifying impending terror attacks need real-time processing even today.

Advances in big data processing, cognitive computing, and IoT are poised to bring about dramatic changes to the data analytics domain. Just as the DBMS is the foundation of current software applications, data analytics will be an integral component of all future software applications.

With the increasing Internet connectivity in the developing countries and the ubiquity of the handheld computing devices such as smart phones and tablets, increasingly more data will be created through social media and Web 3.0 applications. Furthermore the medium for this data will transition from the written form to the spoken for the reasons that follow.

There are more than 6000 spoken languages in the world [50]. Some of these languages are spoken widely, while others are spoken by small communities whose populations are in the order of a few hundreds. Almost all humans communicate through the speech medium, though only a small fraction of these people can read and write, let alone fluently. Spoken language interfaces will be the natural and dominant mode of communication with computing systems in the next decade.

Real-time, speech-to-speech translation systems for the dominant spoken languages already exist as pilot implementations. This trend will continue to encompass speech translation between more and more languages. This has far-reaching implications for both the data analytics domain and ITS. From a traveler's perspective, spoken language interfaces will accelerate the adoption of new transportation paradigms such as self-driving cars. Spoken language interfaces will also improve accessibility to services such as travel kiosks and broaden the user base significantly.

2.7 CHAPTER SUMMARY AND CONCLUSIONS

Data is everywhere. However, for consumers, identifying, cleaning, transforming, integrating, and curating the data is both an intellectual and financial challenge. Given the far-reaching implications of data analytics, it is imperative that data quality be assessed before embarking on data analytics.

It is said that the biggest challenge for big data is ensuring data quality. For example, InfoUSA is a case in point. InfoUSA sells mailing lists data about consumers and businesses, which can be used for customer profiling and targeted marketing. Data about businesses is gathered from over 4000 phone directories and 350 business sources such as new business filings, utility connections, county court houses, and public record notices. Likewise consumer information is collected from over 1000 sources including real estate records, voter registration files, and tax assessments. Missing data, incomplete and inconsistent data, duplicate and ambiguous data are the norms. InfoUSA data is not even remotely big data and they employ over 500 full-time employees for data collection and curation.

The big data and cognitive analytics pose additional problems. The data is typically procured from multiple data vendors, who produce data without any specific context attached to the data. In other words the data is produced for a generic context. The general context of the procured data must be evaluated for context congruence with the proposed data analytics application. Added to this are the concerns related to personal privacy and data provenance.

Data vendors typically use multiple approaches to data collection and curation. Crowdsourcing is a relatively new process for obtaining ideas, services, and data from a large group of people from online communities. The work is divided among the participants and they collectively accomplish the task. For example, assigning keywords to digital images is accomplished through crowdsourcing. Some crowdsourcing service providers such as the Amazon Mechanical Turk, provide participants a fee. Wikipedia and DBpedia are great examples of crowd-sourced projects. However, not all crowd-sourced data collection and curation projects may be open for public comments and scrutiny.

Though some data for ITSs is machine generated from sources such as weather sensors embedded in roadways and connected car networks, this data is also subject to data quality problems. However, integration of the above data with third-party data is critical to realize the true potential of ITSs.

The U.S. Department of Transportation (USDOT) Connected Vehicle Real-Time Data Capture and Management (DCM) Program is a testimony to the critical role that Big Data and Data Analytics will play in the transportation domain [51]. Data Analytics will enable a wide range of strategies aimed at improving safety and mobility, and reducing environmental damage.

Furthermore it will reduce the need for traditional data collection mechanisms such as traffic detectors by replacing them with connected vehicle probes.

IoT technologies enable real-time monitoring of motor vehicles through data collection, integration, and analytics. This entails improved situational awareness both for the vehicle and the driver, which in turn can be used to predict problems and address them before they manifest. Furthermore the IoT data integrated with geospatial and traveler models will enable delivering personalized services to the traveler.

2.8 QUESTIONS AND EXERCISE PROBLEMS

1. What are the sources of big data in the context of ITS?
2. Describe the various steps in the data analytics process?
3. List ways in which big data analytics is different from data analytics. In other words, how does big data impact the role of data analytics?
4. What is the relationship between data science and data analytics?
5. What properties of data is revealed by descriptive analytics?
6. What is the relationship between descriptive analytics and descriptive statistics?
7. Consider the Anscombe's quartet dataset shown in Table 2.1 (5). What is the significance of Anscombe's quartet? What lessons can you learn from it from data analytics perspective?
8. How difficult is it to create another dataset that serves the same purpose as the Anscombe's quartet?
9. In many data analytics projects, outliers are detected and eliminated from analysis. Why?
10. In some applications such as fraud detection, outliers are the observations are prime importance. Why?
11. Several datasets are available at https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html. Seatbelts dataset provides monthly totals of car drivers in Great Britain, killed or seriously injured in the time period January 1969 to December 1984. Construct a scatter plot matrix for this dataset using R or some other statistical software. What insights can you glean from the scatter plot matrix?
12. In what ways are descriptive and diagnostic analytics similar? In what aspects are they different?
13. In what ways are predictive and prescriptive analytics similar? In what aspects are they different?
14. Simple linear regression is just one technique used for predictive analytics. Name and describe three other regression models. Include in your discussion how each of the three models are different from the simple linear regression model.
15. Discuss the technical and business drivers for the data analytics evolution.
16. What is a start scheme? In the context of ITS, design a star schema for connected vehicles of the future.
17. Discuss the reasons for diminishing interest in data warehouses and data marts.
18. What are the computing technology limitations for implementing cognitive analytics?

19. Consider the predictive model building for the airline on-time performance dataset discussed in Section 2.4.3. For this model, we selected 6 variables—month, day of the month, day of the week, hour, flight distance, and days from holiday. Next we developed logistic regression and random forest predictive models.

Select a different set of variables and build logistic regression and random forest predictive models. Do you get different results? Are these results better?

Also experiment with other predictive models such as decision trees and support vector machines.

20. Current research in ITS seems to focus primarily on road traffic in urban areas. What transportation problems can be solved in rural areas using data analytics?

REFERENCES

[1] V. Dhar, Data science and prediction, Commun. ACM 56 (12) (2013) 64–73.

[2] W. Zadrozny, V. de Paiva, L.S. Moss, Explaining watson: Polymath style, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2015, pp. 4078–4082.

[3] L. Kart, Advancing analytics. <http://meetings2.informs.org/analytics2013/Advancing%20Analytics_LKart_INFORMS%20Exec%20Forum_April%202013_final.pdf>.

[4] J.T. Behrens, Principles and procedures of exploratory data analysis, Psychological Methods 2 (1997) 131–160.

[5] NIST, Exploratory data analysis (2002). URL <http://www.itl.nist.gov/div898/handbook/eda/eda.htm>.

[6] The R Foundation, The r project for statistical computing. URL <https://www.r-project.org/>.

[7] S. Liu, M.X. Zhou, S. Pan, W. Qian, W. Cai, X. Lian, Interactive, topic-based visual text summarization and analysis, Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, NY,, 2009, pp. 543–552.

[8] F. Wei, S. Liu, Y. Song, S. Pan, M.X. Zhou, W. Qian, et al., Tiara: A visual exploratory text analytic system, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, NY, 2010, pp. 153–162.

[9] W. Luo, M. Gallagher, D. O'Kane, J. Connor, M. Dooris, C. Roberts, et al., Visualizing a state-wide patient data collection: a case study to expand the audience for healthcare data, Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management – vol 108, HIKM '10, Australian Computer Society, Inc., Darlinghurst, Australia, 2010, pp. 45–52.

[10] Y. Takama, T. Yamada, Visualization cube: modeling interaction for exploratory data analysis of spatio-temporal trend information, Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology – vol 03, WI-IAT '09, IEEE Computer Society, Washington, DC, 2009, pp. 1–4.

[11] H.-J. Schulz, S. Hadlak, H. Schumann, A visualization approach for cross-level exploration of spatiotemporal data, Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, i-Know '13, ACM, New York, NY, 2013, pp. 2:1–2:8.

[12] T.-H. Huang, M.L. Huang, Q.V. Nguyen, L. Zhao, A space-filling multidimensional visualization (sfmdvis for exploratory data analysis), Proceedings of the 7th International Symposium on Visual Information Communication and Interaction, VINCI '14, ACM, New York, NY, 2014, pp. 19:19–19:28.

[13] G. Liu, A. Suchitra, H. Zhang, M. Feng, S.-K. Ng, L. Wong, Assocexplorer: an association rule visualization system for exploratory data analysis, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, ACM, New York, NY, 2012, pp. 1536–1539.
[14] M. d'Aquin, A. Adamou, S. Dietze, Assessing the educational linked data landscape, Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, ACM, New York, NY, 2013, pp. 43–46.
[15] H. Drachsler, S. Dietze, E. Herder, M. d'Aquin, D. Taibi, M. Scheffel, The 3rd lak data competition, in: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, LAK '15, ACM, New York, NY, pp. 396–397.
[16] A. Essa, H. Ayad, Student success system: risk analytics and data visualization using ensembles of predictive models, Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, LAK '12, ACM, New York, NY, 2012, pp. 158–161.
[17] A. Gibson, K. Kitto, J. Willis, A cognitive processing framework for learning analytics, Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, LAK '14, ACM, New York, NY, 2014, pp. 212–216.
[18] R. Vatrapu, C. Teplovs, N. Fujita, S. Bull, Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making, Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK '11, ACM, New York, NY, 2011, pp. 93–98.
[19] V. Gudivada, V. Raghavan, V. Govindaraju, C. Rao (Eds.), Cognitive Computing: Theory and Applications, vol. 35 of Handbook of Statistics, Elsevier, New York, NY, 2016 ISBN: 978-0444637444.
[20] V. Gudivada, M. Irfan, E. Fathi, D. Rao, Cognitive analytics: Going beyond big data analytics and machine learning, in: V. Gudivada, V. Raghavan, V. Govindaraju, C.R. Rao (Eds.), Cognitive Computing: Theory and Applications, vol. 35 of Handbook of Statistics, Elsevier, New York, NY, 2016, pp. 169–205.
[21] E.F. Codd, A relational model of data for large shared data banks, Commun. ACM 13 (6) (1970) 377–387.
[22] V. Gudivada, R. Baeza-Yates, V. Raghavan, Big data: promises and problems, IEEE Computer 48 (3) (2015) 20–23.
[23] V. Gudivada, D. Rao, V. Raghavan, Renaissance in database management: navigating the landscape of candidate systems, IEEE Computer 49 (4) (2016) 31–42.
[24] J. Celko, Joe Celko's analytics and OLAP in SQL, Morgan Kaufmann, San Francisco, California, 2006.
[25] V.N. Gudivada, D. Rao, V.V. Raghavan, NoSQL systems for big data management, 2014 IEEE World Congress on Services, IEEE Computer Society, Los Alamitos, CA, 2014, pp. 190–197.
[26] J. Han, M. Kamber, J. Pei, Data mining: concepts and techniques, third ed., The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Burlington, MA, 2011.
[27] F. Greitzer, C. Noonan, L. Franklin, Cognitive Foundations for Visual Analytics, Pacific Northwest National Laboratory, Richland, WA, 2011.
[28] V. Govindaraju, V. Raghavan, C. Rao (Eds.), Big Data Analytics, vol. 33 of Handbook of Statistics, Elsevier, New York, NY, 2015 ISBN: 978-0444634924.
[29] J. Hurwitz, M. Kaufman, A. Bowles, Cognitive Computing and Big Data Analytics, Wiley, New York, NY, 2015.
[30] K.H. Pries, R. Dunnigan, Big Data Analytics: A Practical Guide for Managers, CRC Press, Boca Raton, FL, 2015.
[31] A. Abrahamsen, W. Bechtel, History and core themes, in: K. Frankish, W.M. Ramsey (Eds.), The Cambridge Handbook of Cognitive Science, Cambridge University Press, 2012.
[32] A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, IEEE Intelligent Systems 24 (2) (2009) 8–12.

[33] B. Baesens, Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications, Wiley, New York, NY, 2014.
[34] S. Sadiq, N.K. Yeganeh, M. Indulska, 20 years of data quality research: themes, trends and synergies, Proceedings of the Twenty-Second Australasian Database Conference – vol. 115, Australian Computer Society, Inc., Darlinghurst, Australia, 2011, pp. 153–162.
[35] H.M. Sneed, R. Majnar, A process for assessing data quality, Proceedings of the 8th International Workshop on Software Quality, ACM, New York, NY, 2011, pp. 50–57.
[36] American Statistical Association, Airline on-time performance. URL <http://stat-computing.org/dataexpo/2009/>.
[37] O. Mendelevitch, Data science with Apache Hadoop: Predicting airline delays. URL <http://hortonworks.com/blog/data-science-apacheh-hadoop-predicting-airline-delays/>.
[38] Max Planck Institute for Evolutionary Anthropology, World atlas of language structures (wals). URL <http://wals.info/>.
[39] G.A. Miller, Wordnet: a lexical database for english, Communications ACM 38 (11) (1995) 39–41. Available from: http://dx.doi.org/10.1145/219717.219748.
[40] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, One billion word benchmark for measuring progress in statistical language modeling, Computing Research Repository (CoRR) abs/1312.3005.
[41] University of California, Irvine, Machine Learning Repository. URL <http://archive.ics.uci.edu/ml/>.
[42] DATA.GOV, Data catalog. URL <http://catalog.data.gov/dataset>.
[43] Project Gutenberg. Free ebooks by project gutenberg. URL <https://www.gutenberg.org/>.
[44] TensorFlow. An open source software library for numerical computation using data flow graphs. URL <https://www.tensorflow.org/>.
[45] SyntaxNet. An open source neural network framework for tensorflow for developing natural language understanding (nlu) systems. URL <https://github.com/tensorflow/models/tree/master/syntaxnet>.
[46] The University of Waikato. Weka 3: Data mining software in java. URL <http://www.cs.waikato.ac.nz/ml/weka/>.
[47] Apache, The UIMA Project. URL <http://uima.apache.org/>.
[48] Cloudera. QuickStarts. URL <http://www.cloudera.com/downloads.html>.
[49] HORTONWORKS. Hortonworks sandbox. URL <http://hortonworks.com/products/sandbox/>.
[50] Ethnologue: Languages of the world (Oct. 2016). URL Online version: <http://www.ethnologue.com>.
[51] Big Data's Implications for Transportation Operations: An Exploration (December 2014). URL Online version: <http://ntl.bts.gov/lib/55000/55000/55002/Big_Data_Implications_FHWA-JPO-14-157.pdf>.