ABSTRACT This chapter defines analytics and traces its evolution from its origin in 1988 to its current stage—cognitive analytics. We discuss types of learning and describe classes of machine learning algorithms. Given this backdrop, we propose a reference architecture for cognitive analytics and indicate ways to implement the architecture. A few cognitive analytics applications are briefly described. The chapter concludes by indicating current trends and future research direction. Keywords: Cognitive analytics, Text analytics, Learning analytics, Educational data mining, Cognitive systems, Cognitive computing, Personalized learning, Data science, Machine learning, Big data analytics, Business analytics 1 INTRODUCTION Cognitive computing is a computational environment which is comprised of (1) a high-performance computing infrastructure powered by special processors such as multicore CPUs, GPUs, TPUs, and neuromorphic chips; (2) a software development environment with intrinsic support for parallel and distributed computing, and powered by the underlying computing infrastructure; (3) software libraries and machine learning algorithms for extracting information and knowledge from unstructured data sources; (4) a data analytics environment whose processes and algorithms mimic human cognitive processes; and (5) query languages and APIs for accessing the services of the cognitive computing environment. We have defined cognitive computing in terms of its functions, since it is not easy to define it precisely and completely by other methods. Cognitive analytics draws upon the cognitive computing Handbook of Statistics, Vol. 35.
http://dx.doi.org/10.1016/bs.host.2016.07.010   environment to generate actionable insights by analyzing diverse heterogeneous data sources using cognitive models that the human brain employs. Classical symbolic and rule-based approaches to problems such as machine translation and speech-to-speech translation are being superseded by statistical learning approaches. For example, consider the problem of recognizing handwritten digits. Rule-based approaches entail developing a number of rules which aim to explicitly capture ways to write digits by different users. This results in too many rules. Furthermore, additional rules are needed to accommodate new users who might write a digit differently from the ways that were reflected in the current rule set. In contrast, artificial neural network (ANN) approaches use several small pieces of evidence in the form of features and combine them to produce higher-level features. ANN approaches are more robust as they perform better with data which is not seen in the training phase. The ubiquity of big data (Gudivada et al., 2015a), abundant computing power, and the resurgence of neural network algorithms are providing scalable solutions to several difficult problems. The performance of newer approaches to problems that have been considered difficult for computers such as finding objects in images and classifying images rival human performance. For example, in the ImageNet Large-Scale Visual Recognition Challenge (Russakovsky et al., 2015), the error rate for some algorithms for detecting objects in video and scene classification is as low as 6%, whereas the error rate for humans is 5%. In another study with deep-learning algorithms (Goodfellow et al., 2014), Google reports 99.8% accuracy in recognizing CAPTCHA images on the hardest category of reCAPTCHA dataset. In another study at Facebook on the classification of images, Taigman et al. (2014) achieved an accuracy of 97.35% on the Labeled Faces in the Wild dataset using a nine-layer deep neural network. Finally, Lake et al. (2015) describe an approach called Bayesian Program Learning, which is used to recognize 1623 handwritten character sets from 50 languages with only a limited training. Though the above problems are diverse, deep neural network algorithms perform exceptionally well in all these domains. The above approaches coupled with advances in information retrieval, natural language understanding (Gudivada et al., 2015b), artificial intelligence (AI), and machine learning are helping to

usher in a new paradigm for strategic decision making. The term data analytics when used in a generic sense refers to any actionable information that results from computational analysis of data using mathematical and statistical methods. Data analytics is an interdisciplinary domain encompassing mathematics, statistics, and computer science. Implicitly, there is a domain associated with data analytics. The domain provides the data for analysis. The primary goal of data analytics is to gain insights into a process or problem so that the latter can be improved or solved. In other words, analytics is a data-driven approach to decision making and problem solving. 170 SECTION B Complex Analytics and Machine Learning Though certain types of analytics are common across various application domains, they tend to vary significantly from domain to another. This has led to the proliferation of names such as business analytics, text analytics, image analytics, video analytics, graph analytics, spatial analytics, visual analytics, and cognitive analytics. However, irrespective of the domain, data analytics is comprised of three components: data acquisition and loading, methods and algorithms, and a computational platform that implicitly embodies workflows and best practices. The data acquisition and loading components enable the preparation of input data and loading it into the computational platform. Various algorithms and approaches for data analysis are provided by the methods and algorithms component. Lastly, the computational platform brings everything together as a system and provides interfaces for users and other applications to interact with it. From a functional perspective, there are three categories of data analytics: descriptive, prescriptive, and predictive. Descriptive analytics provides a dashboard view of the current state of a system or process. It uses descriptive statistics and machine learning algorithms to provide insight into a system. The insight often reveals, in a process, for example, various steps in the process, how the steps are sequenced, what type of resources are consumed, and how much time is spent in each process. As another example, readability of English texts is determined by text analytics such as the Fry readability formula, Automated Readability Index, Flesch-Kincaid, Gunning-Fog, ColemanLiau Index, and SMOG Index. Software metrics and measurements are analytics used to characterize properties of software. Such metrics include the number of classes, number of methods per class, depth of inheritance tree, number of interfaces, and total lines of code. Prescriptive analytics is a natural outcome of descriptive analytics. It suggests ways to improve a current process or system using simulation and optimization algorithms. In the case of software metrics and measurements, prescriptive analytics specifies a range of values for each measurement such as bounds for number of methods in a class. Furthermore, it specifies refactoring techniques if a measurement is not within the specified range. Predictive analytics enables answering “what-if” questions by building predictive models using inferential statistics and forecasting techniques. It enables organizations to make data-driven strategic decisions. Predictive models are built using the operational and historical data. They extract associations and other implicit relationships in the data to build the models. Various regression models such as linear, logistic, Lasso, ridge, Cox proportional hazards, and Bayesian are widely used. Logic regression, for example, is used in clinical trials and fraud detection to associate a probability with a binary outcome. Like cognitive computing, cognitive analytics is pursued from two complementary perspectives. The first is driven by the computer science researchers in both industry and academia. Advances in big data, cloud computing, Cognitive Analytics Chapter 5 171 natural language understanding, and machine learning are enabling extraction of knowledge from vast repositories of unstructured data such as natural language text, images, video, and audio. From this group’s perspective, the knowledge extracted from the unstructured data coupled with statistical inference and reasoning distinguishes cognitive analytics from

business analytics. The second perspective is advanced by cognitive and neuroscience researchers. They employ theories of mind, functional areas of the brain, and cognitive models and processes. For example, an approach in this class might gather analytics about a cognitive process to validate the cognitive model as well as to improve the model (Chakraborty et al., 2014). 1.1 Chapter Organization The overarching goal for this chapter is to present a unified approach to the emerging area of cognitive analytics. More specifically, in Section 2, we trace the evolution of data analytics and discuss central issues. Types of learning used in cognitive analytics are described at a conceptual level in Section 3. In Section 4, we discuss the following classes of machine learning algorithms: logistic regression, decision trees, support vector machines (SVMs), Bayesian networks (BNs), neural networks, and deep learning. This section also includes a discussion on machine learning frameworks and libraries. We propose a reference architecture called Cognalytics for cognitive analytics in Section 5. This section also indicates how this architecture can be implemented using open source tools. Section 6 presents applications of cognitive analytics including learning analytics (LA), personalized learning, cognitive businesses, brain–computer interfaces (BCIs), and assistive technologies. Cognitive analytics trends and future research directions are described in Section 7. Section 8 concludes the chapter. 2 EVOLUTION OF ANALYTICS AND CORE THEMES AI is a subfield of computer science and machine learning is a major area within AI. The recent emergence of big data and cloud computing created AI renaissance. The attendant media coverage of machine learning is making the latter a household name. This is also creating confusion and propagation of misinformation. In blogs and other self-published forums, some authors have declared AI and Computer Science as two distinct disciplines, likewise, AI and machine learning. The scope and the meaning of the term analytics are being reinvented. You cannot manage what you do not measure is an old adage from the management world that is still true today in most organizations and academic disciplines. At the core of analytics are data, mathematical and statistical models built using this data. The types of data needed and the type of processing performed, and the variety of models built varies. The models are used 172 SECTION B Complex Analytics and Machine Learning for a broad range of purposes under the umbrella terms descriptive analytics, prescriptive analytics, and predictive analytics. AI, machine learning, distributed computing, and high-performance computing comprise the computational infrastructure to manage data and enable model building. 2.1 Multiple Perspectives There exists multiple perspective on analytics. The Computer Science perspective is driven by technical considerations related to storing, managing, and querying data. In the early days, there was limited support for analysis. The business perspective views analytics from an organizational level and focuses on actionable insights into data. Visual analytics is a new area of analytics whose goal is analytical reasoning through interactive visual interfaces. Even more recently, other terms such as educational data mining (EDM), LA, and cognitive analytics have emerged. Academia has responded to this unprecedented interest in analytics by creating new interdisciplinary degree programs primarily at the master's level. These programs fall into three categories: (1) programs that have the term analytics somewhere in their name—business analytics, health informatics, health care informatics, and nursing informatics. Other degree programs such as econometrics also fall under this category, though they do not explicitly use the term informatics in the name. These programs are typically administered or led by noncomputer science departments; (2) programs with names such as master of science in analytics and master of science in data science. These programs are typically led by computer science departments; and (3) numerous graduate certificates, tracks, and concentrations in analytics,

data mining, knowledge discovery, machine learning, and big data. 2.2 Analytics Evolution We trace the evolution of analytics from a Computer Science perspective as shown in Fig. 1. Basic analytics functions were part of the relational database management systems (RDBMS) from their early years. RDBMS served as operational databases for conducting day-to-day business transactions— online transaction processing (OLTP). Basic functions for descriptive statistics were provided. In subsequent years, more sophisticated functions were introduced under the name Statistics & SQL Analytics. They included functions for ranking of results, moving and cumulative aggregating values over a range of rows, lag and lead to access data from preceding and following rows, descriptive statistics, correlations, and linear regression. In the early days of RDBMS, analytic functions were implemented outside of the RDBMS system. Each analytic function was implemented by a standalone piece of code which made code optimization across RDBMS and analytic functions difficult. Recently, there have been efforts in implementing analytic functions within the database (Feng et al., 2012).