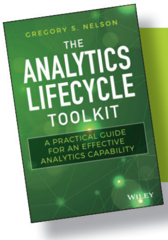
GREGORY S. NELSON

THE ANALYTICS LIFECYCLE TOOLKIT

A PRACTICAL GUIDE FOR AN EFFECTIVE ANALYTICS CAPABILITY

WILEY



Contents

Preface xi

Acknowledgments xv

PART I THE FOUNDATION OF ANALYTICS 1

Chapter 1 Analytics Overview 3

Chapter 2 The People of Analytics 38

Chapter 3 Organizational Context for Analytics 68

Chapter 4 Data Strategy, Platforms, and Architecture 95

PART II ANALYTICS LIFECYCLE BEST PRACTICES 127

Chapter 5 The Analytics Lifecycle Toolkit 129

Chapter 6 Problem Framing 148

Chapter 7 Data Sensemaking 185

Chapter 8 Analytics Model Development 218

Chapter 9 Results Activation 266

Chapter 10 Analytics Product Management 301

PART III SUSTAINING ANALYTICS SUCCESS 349

Chapter 11 Actioning Analytics 351

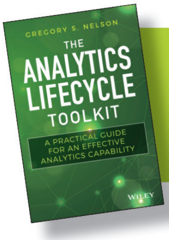
Chapter 12 Core Competencies for Analytics Teams 386

Chapter 13 The Future of Analytics 424

About the Author 433

About the Companion Web Site 435

Index 437



CHAPTER 1

Analytics Overview

> *…what enables the wise commander to strike and conquer, and achieve things beyond the reach of ordinary men, is foreknowledge. Now, this foreknowledge cannot be elicited from spirits…*
>
> *The Art of War*, Sun Tzu (as seen in Giles, 1994)

FUNDAMENTAL CONCEPTS

Peter Drucker first spoke of the "knowledge economy" in his book *The Age of Discontinuity* (Drucker, 1969). The knowledge economy refers to the use of knowledge "to generate tangible and intangible value." Nearly 50 years later, organizations have virtually transformed themselves to meet this challenge, and data and analytics have become central to that transformation.

In this chapter, we highlight the "fundamentals" of analytics by hopefully creating a level playing field for those interested in the moving from the *concept* of analytics to the *practice* of analytics. The fundamentals include defining both *data* and *analytics* using terms that I hope resonate. In addition, I think it is important to consider analytics in the wider context of how it is used and the value derived from these efforts. Finally, in this chapter, I relate analytics to other widely used terms as a way to find both common ground and differentiation with often-confused terminologies.

Data

Data permeates just about every part of our lives, from the digital footprint we leave with our cell phones, to health records, purchase history, and utilization of resources such as energy. While not impossible, it would require dedication and uncanny persistence to live "off-the-grid" in this digital world. Beyond the pure generation of data, we are also voracious consumers of data, reviewing our online spending habits, monitoring our fitness regimes, and reviewing those frequent flyer points for that Caribbean vacation.

But what is data? At its most general form, data is simply information that has been stored for later use. Earliest forms of recording information might have been notches on bones (Sack, 2012). Fast forward to the 1950s, and people recorded digital information on Mylar strips (magnetic tape), then punch cards, and later disks. Modern data processing is relatively young but has set the foundation for how we think about the collection, storage, management, and use of information.

Until recently, we cataloged information that wasn't necessarily computable (e.g., videos, images); but through massive technological change, the class of "unstorable" data is quickly vanishing. Stored information, or data, is simply a model of the real world encoded in a manner that is usable, or for our purposes "computable" (Wolfram, 2010).

The fact that data is a persistent record or "model" of something that happened in the real world is an important distinction in analytics. George Box, a statistician considered by many as "one of the greatest statistical minds of the 20th century" (Champkin, 2013) was often quoted as saying: "All models are wrong, but some are useful." All too often, we find something in the data that doesn't make sense or is just plain wrong. Remember that data has been translated from the real, physical world into something that represents the real world—George's "model." Just as the mechanical speedometer is a standard for measuring speed (and a pretty good proxy for measuring velocity), the model is really measuring tire rotation, not speed. (For those interested in a late-night distraction, I refer you to Woodford's 2016 article "Speedometers" that explains how speedometers work.) In sum, data is stored information and serves as the foundation for all of analytics. In visual analytics, for example, we make sense out of the data using visualization techniques that enable us to perform analytical reasoning through interactive, visual interfaces.

Analytics

Analytics may be one of the most overused yet least understood terms used in business. For some, it relates to the technologies used to "beat data into submission," or it is simply an extension of business intelligence and data warehousing. And yet for others, it relates to the statistical, mathematical, or quantitative methods used in the development of models.

According to Merriam-Webster (Merriam-Webster, 2017), **analytics** is "the method of logical analysis." Dictionary.com (dictionary.com, 2017) defines analytics as "the science of logical analysis." Unfortunately, both definitions use the root word of *analysis* in the definition, which seems a bit like cheating.

The origin of the word *analysis* goes all the way back to the 1580s, where the term is rooted in Medieval Latin (analȳticus) and Greek (analȳtikós), and means to break up or to loosen. Throughout this book, I frame analytics as *a structured approach to data-driven problem solving*—one that helps us break up problems through careful consideration of the facts.

What Is Analytics?

There has been much debate over the definition of analytics (Rose, 2016). While the purpose of this book is not to redefine or challenge anyone's definition, for the current discussion I define analytics as:

> a comprehensive, data-driven strategy for problem solving

I intentionally resist using a definition that views analytics as a "process," a "science," or a "discipline." Instead, I cast analytics as a comprehensive strategy, and as you will see in Part II of this book, it encompasses best practice areas that contain processes, along with roles and deliverables.

Analytics uses logic, **inductive and deductive reasoning**, critical thinking, and quantitative methods—along with data—to examine phenomena and determine its essential features. Analytics is rooted in the scientific method (Shuttleworth, 2009), including problem identification and understanding, theory generation, hypothesis testing, and the communication of results.

Inductive reasoning

Inductive reasoning refers to the idea that accumulated evidence is used to support a conclusion but with some level of uncertainty. That is, there is a chance (probability) that the final conclusions may differ from the underling premises. With inductive reasoning, we make broad generalizations from specific observations or data.

Deductive reasoning

Deductive reasoning on the other hand makes an assertion about some general case and then seeks to prove or disprove it with data (using **statistical inference** or **experimentation**). We propose a theory about the way the world works and then test our hypotheses.

We will explore this in more detail later in this chapter.

Analytics can be used to solve big hairy problems such as those faced by UPS that helped them "save more than 1.5 million gallons of fuel and reduced carbon dioxide emissions by 14,000 metric tons" (Schlangenstein, 2013) as well operational problems like optimizing the scheduling of operating rooms for Cleveland Clinic (Schouten, 2013). With success stories like those, it is no wonder that analytics is an attractive bedfellow with technology vendors (hardware and software) and other various proponents. Of course, the danger in the overuse of analytics can be seen in the pairing of the term with other words such as:

- Big data analytics
- Prescriptive analytics
- Business analytics
- Operational analytics
- Advanced analytics
- Real-time analytics
- Edge or ambient analytics

While these pairings offer distinctive qualifiers on the type and context to which analytics is applied, it often creates confusion, especially in C-suites, where technology vendors offer the latest analytics solutions to solve their every pain. My perspective (and one that is shared with lots of other like-minded, rational beings) is that analytics is not a technology but that technology serves as an enabler.

Analytics is also often referred to as "any solution that supports the identification of meaningful patterns and relationships among data." Analytics is used to parse large or small, simple or complex, structured and unstructured, quantitative or qualitative data for the express purposes of understanding, predicting, and optimizing. Advanced analytics is a subset of analytics that uses highly developed and computationally sophisticated techniques with the intent of supporting the fact-based decision process—usually in an automated or semi-automated manner.

Advanced analytics typically incorporates techniques such as data mining, econometrics, forecasting, optimization, predictive modeling, simulations, statistics, and text mining.

How Analytics Differs from Other Concepts

Vincent Granville, who operates Data Science Central, a social network for data scientists, compared 16 analytics disciplines to data science (Granville, 2014). Without repeating those here (but definitely worth the read!), it is useful to highlight the differences between analytics and similar concepts as a way to clarify the meaning of analytics. Here, analytics will be described as it relates to concepts and methods:

- **Concepts**
  - Business intelligence and reporting
  - Big data
  - Data science
  - Edge (and ambient) analytics
  - Informatics
  - Artificial intelligence and cognitive computing
- **Methods**
  - Applied statistics and mathematics
  - Forecasting and time series
  - NLP, text mining, and text analytics
  - Machine learning and data mining

To start with, it is important to distinguish between **concepts** and **methods**.

Concepts

Concepts are generalized constructs that help us understand what something is or how it works.

Methods

Methods, in this context, are the specific techniques or approaches that are used to implement an analytic solution.

Another way to think about this is that methods describe approaches to different types of problems. For example, we might consider something as an optimization problem or a forecasting problem, whereas big data is a mental model that helps us understand the complexity of modern data challenges. Similarly, as we will see later in this chapter, machine learning can be thought of as simply the current state of artificial intelligence—the latter being the concept and the former being the method.

ANALYTICS CONCEPTS

An analytics concept can be thought of as an abstract idea or a general notion. We differentiate concepts from implementation to highlight the fact that the idea necessarily can take on multiple forms when implemented. For example, the concept of artificial intelligence can be seen in self-driving cars, chatbots, or recommendation engines. The specific implementations are essentially the current state implementation of the concept.

In the following section, I outline my interpretation of what I see as the fundamental definition of **business intelligence**, **reporting**, **big data**, **data science**, **edge analytics**, **informatics**, and the world of **artificial intelligence** and **cognitive computing**.

Business Intelligence and Reporting

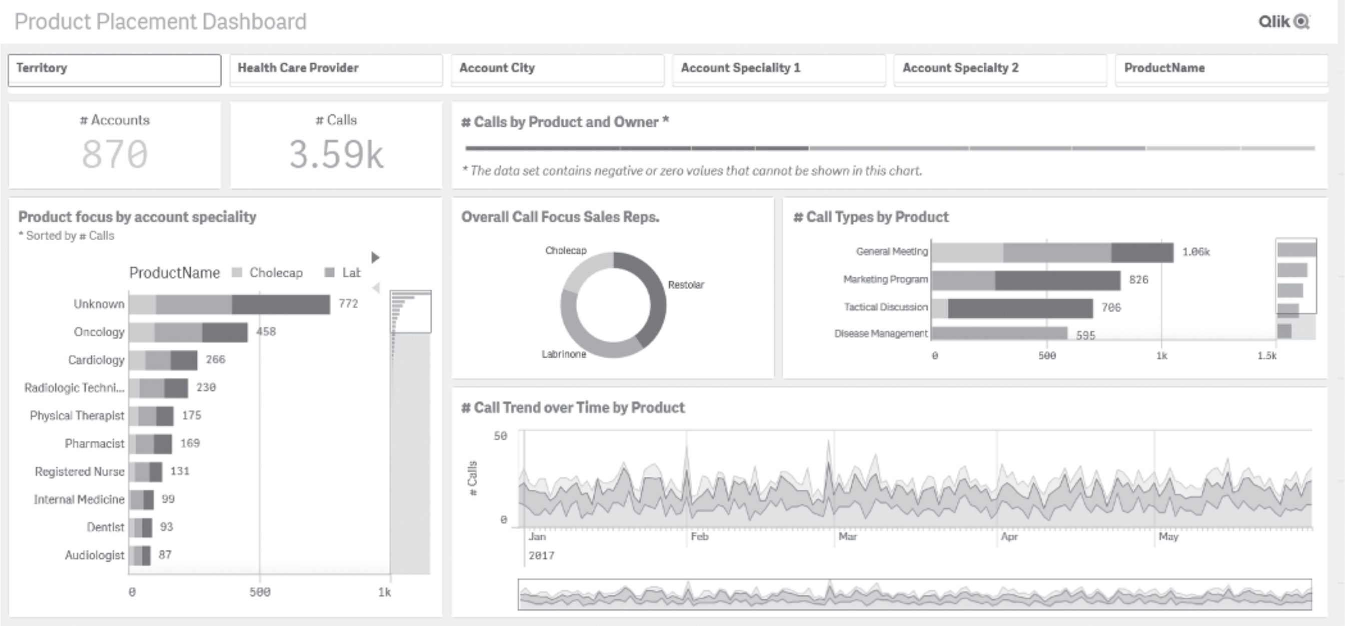
There is little consensus as to how analytics and business intelligence differ. Some categorize analytics as a subset of business intelligence, while others position analytics in an entirely different box. In a paper I wrote in 2010 (Nelson, 2010), I defined business intelligence (BI) as "a management strategy used to create a more structured and effective approach to decision making ... BI includes those common elements of reporting, querying, Online Analytical Processing (OLAP), dashboards, scorecards and even analytics. The umbrella term 'BI' also can refer to the processes of acquiring, cleansing, integrating, and storing data."

There are those who would classify the difference between analytics and business intelligence as differences between (1) the complexity of the quantitative methods used (i.e., algorithms, mathematics, statistics) and (2) whether the focus of the results is historical or future-oriented. That is to say, business intelligence is focused on the presentation of historical results using relatively simple math, while analytics is thought of as much more computationally sophisticated and capable of predicting outcomes of interest, determining causal relationships, identifying optimal solutions, and sometimes also used to prescribe actions to take.

The limit of most business intelligence applications lies not in the constraints of technology but, rather, in the depth of analysis and the true insights created that inform action. Telling me, for example, that something happened doesn't help me understand what to do to change the future—often, that is left for *offline analysis*. The promise of analytics is that it creates actionable insights about what happened (and where, why, and under what conditions), what is likely to occur in the future, and then what can be done to influence and optimize that future.

Note that the BI dashboard depicted in Figure 1.1 relays facts about the past such as sales, call volumes, products, and accounts, making it easy to get a quick snapshot of the current state of the organization's sales position or activities.

Business intelligence and its little sister, "reporting," are the techniques used to display information about a phenomenon, usually at the tail end of the data pipeline where visual access to data and results



* The data set contains negative or zero values that cannot be shown in this chart.

Figure 1.1 BI dashboard

Source: © QlikTech International AB. Reprinted with permission.

are surfaced. Analytics, on the other hand, goes beyond description to actually understand the phenomenon to predict, optimize, and prescribe appropriate actions.

Business intelligence has traditionally suffered from two shortcomings. These shortcomings are related to the fact that BI typically (1) focuses on creating awareness of the facts of the past in that it measures and monitors rather predicting and optimizing; and (2) it is often not quantitatively sophisticated enough to build accurate insights that could be used to influence meaningful change (although the right report or visualization can influence change).

In cases where business intelligence is properly coupled with in-depth "analysis" rather than the mere awareness of facts, it gets closer to analytics. But it often lacks the advanced statistical and mathematical sophistication or "learning" seen in advanced analytics solutions.

To that end, I view analytics as a natural evolution of the concepts contained within the general framework of business intelligence. It places more emphasis on the full pipeline of activities necessary to create insights that fuel action. Analytics is more than just the predefined visual elements used in self-service dashboards or reporting interfaces.

Big Data

Big data is a way to describe the cacophony of information that organizations must deal with in their efforts to turn data into insights. The phrase *big data* was first used by Michael Cox and David Ellsworth in 1997 (Cox, 1997) who referred to the "problem" as follows:

> Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.

Think of big data as a concept that highlights the challenge of utilizing traditional methods of data analysis because of the size and complexity of that data. We contrast big data with traditional "small" data by its volume (how much data we have), velocity (how fast the data is coming at us), and variety (numbers, text, images, video).[1]

If big data is a concept used to describe the complexity of today's information, analytics is used to help us analyze that complexity in proactive (predictive and prescriptive) ways versus reactive ways (i.e., the realm of business intelligence).

Data Science

It would seem that defining big data was a cakewalk as compared with data science as such little consistency can be found in the term. There is a lot of debate about what it means and whether it is different at all from analytics. Even those who would attempt a definition might do so by discussing the people (data scientists), the skills they need to have, the roles they play, the tools and technologies used, where they work, and their educational backgrounds. But this doesn't give one a meaningful definition.

Rather than describing data science by the people or the types of problems they address, it might be more accurate to define it as follows:

> Data Science is the scientific discipline of using quantitative methods from fields like statistics and mathematics along with modern technologies to develop algorithms designed to discover patterns, predict outcomes, and find optimal solutions to complex problems.

The difference between data science and analytics is that data science can help support or even automate the analysis of data, but analytics is a human-centered strategy that takes advantage of a variety of tools, including those found in data science, to understand the true nature of the phenomenon.

Data science is perhaps the broadest of these concepts in that it relates to the entirety of the science and practice of dealing with "data." I think data science is analytics engineered by computer scientists. In practice, however, data science tends to focus on macro,

[1] Note: The three Vs of Big data have evolved into five Vs that also include veracity (trustworthiness) and value.

generalized problems, whereas analytics tends to address particular challenges within an industry or problem space. In Chapter 10, I extend this by defining the relationship between data science and analytics by referring to data science as an enabler of analytics.

Edge (and Ambient) Analytics

Analytics is a predominant activity for most modern organizations that see it as their directive to **democratize data** through data-driven, human-centered processes. Edge analytics refers to distributed analytics where the analytics are built into the machinery and systems where information is generated or collected as part of the "unconscious" activities of an organization.

Edge analytics is often associated with smart devices where the analytical computation is performed on data at the point of collection (e.g., equipment, sensor, network switch, or other device). Rather than relying on traditional data-pipeline methods where data is collected, transmitted, cleansed, integrated, and warehoused, analytics are embedded within the device or nearby.

DEMOCRATIZE DATA

The democratization of data refers to "freeing" data so that everyone that can and should have access to data is given the tools and the rights to explore the data and these are not limited to the privileged few.

As an example, consider the fact that traditional credit card fraud detection relies on a machine (e.g., card reader) and a connection to an authorization "broker" to validate the transaction by sending a request and very quickly (hundredths of a millisecond) applying an algorithm to authorize or flag the operation and the device receives the authorization. In edge analytics, the algorithm would run on the instrument itself (think smart chip reader with embedded analytics).

Edge analytics is often linked with the Internet of Things (IoT), and a recent IDC FutureScape for IoT report found that "by 2018, 40% of IoT data will be stored, processed, analyzed and acted upon at the edge of the network where it is created" (Marr, 2016). As IoT evolves, we will likely see more attention paid to the Analytics of Things (AoT), which refers to the opportunity of analytics to bring unique value to IoT data.

Ambient analytics is a related term whose name implies that "analytics are everywhere." Just as the lighting or acoustics of a room often go unnoticed but set the stage for mood, ambient analytics will support and influence the context of how we work and play. We are seeing ambient intelligence play out in everyday scenarios, such as detecting glucose levels and administering insulin. Similarly, home automation devices can detect when you are nearing your home and adjust the temperature and turn on lighting. Ambient analytics goes beyond simple decision rules and utilizes algorithms to decide on the appropriate course of action.

There is little doubt that edge and ambient analytics will continue to challenge the traditional human-centered processes for operationalizing (e.g., understanding, deciding, and acting) analytics.

Informatics

Informatics is a discipline that lies at the intersection of information technology (IT) and information management. In practice, informatics relates to the technologies used to process data for storage and retrieval. In essence, informatics deals with the realm of how information is managed and refers to the ecosystem of data and systems that support process workflows rather than the analysis of the data found therein.

Often used in information sciences and used heavily in healthcare and research, health informatics is a specialization that sits between health IT and health information management and links information technology, communications, and health care to improve the quality and safety of patient care. It lies at the heart of where people, information, and technology intersect.

Health policy refers to decisions, plans, and actions that are undertaken to achieve specific health-care goals within a society. Because health policy makers want to see health care become more affordable, safer, and of higher quality, information technology and health informatics are often the means prescribed to do this. In fact, one of the

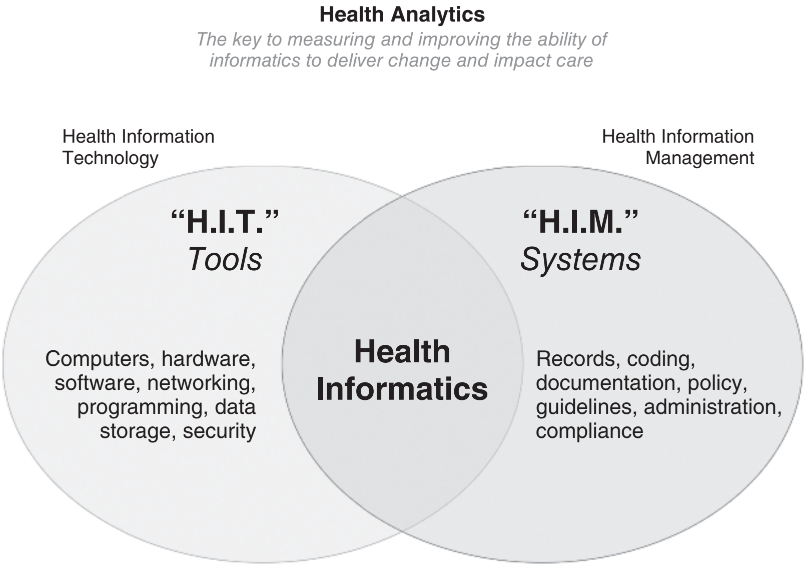

Health Analytics
The key to measuring and improving the ability of informatics to deliver change and impact care
Health Information Technology
Health Information Management
"H.I.T."
Tools
"H.I.M."
Systems
Computers, hardware, software, networking, programming, data storage, security
Health Informatics
Records, coding, documentation, policy, guidelines, administration, compliance

Figure 1.2 The difference between health information management, health IT, and informatics

biggest mandates is to position data resources as to enable a 360-degree view of every patient, and only data sharing can accomplish this (see Figure 1.2).

Analytics integrates with all of these concepts and relies on the underlying data, supporting technologies, and information management processes.

Artificial Intelligence and Cognitive Computing

Artificial intelligence (AI) is "the science of making computers do things that require intelligence when done by humans" (Copeland, 2000).

The difference between artificial intelligence (AI) and **machine learning** is that AI refers to the broad concept of using computers to perform the "intelligence" work of discovering patterns, whereas machine learning is a part of AI that relates to the notion that computers can learn from data.

Machine learning is a subset of artificial intelligence that can learn from and make predictions based on data. Rather than following a particular set of rules or instructions, an algorithm is trained to spot patterns in large amounts of data.

Artificial intelligence (and machine learning) can be used in the Analytics Lifecycle to support discovery (e.g., how data is structured, what patterns exist). The application of artificial intelligence found in analytics usually comes in the form of machine learning (as seen above) or cognitive computing.

Cognitive computing is a unique case that combines artificial intelligence and machine-learning algorithms in an approach that attempts to reproduce (or mimic) the behavior of the human brain (Feldman, 2017).

Cognitive computing systems are designed to solve problems the way people solve problems, by thinking, reasoning, and remembering. This approach gives cognitive computing systems an advantage that allows them to "learn and adapt as new data arrives" and to "explore and uncover things you would never know to ask about." (Saffron Technologies, 2017). The advantage of cognitive computing is that once it learns—unlike humans—it never forgets.

> *In the battle of man vs. algorithm, unfortunately, man often loses. The promise of Artificial Intelligence is just that. So if we're going to be smart humans, we must learn to be humble in situations where our intuitive judgment simply is not as good as a set of simple rules.*
>
> Farnham Street Blog (Parrish, 2017, Do Algorithms Beat Us at Complex Decision Making?)

In slightly pejorative terms, artificial intelligence acts on behalf of a human, whereas cognitive computing provides information to help people decide.

Learn More

To learn more about the difference between AI and cognitive computing, please review the referenced article by Steve Hoffenberg (Hoffenberg, 2016).

THE METHODS OF ANALYTICS

In the prior section, we discussed analytics and some of the related concepts such as big data and data science. We now turn our attention to the practical methods used in analytics, including the tools in the analytics toolbox.

Specifically, in this section, I will outline the methods found in statistics, time series analysis, natural language processing, machine learning, and operations research.

Applied Statistics and Mathematics

Like many of the concepts that have already been discussed, there is a wide disparity about how people define **statistics** and how it differs from mathematics in general. Some would argue that statistics is a branch of mathematics (Merriam-Webster, 2017b), and others (like John Tukey (Brillinger, 2002)) suggest that it is a science. Most would agree that like physics, statistics uses mathematics but is not math (Milley, 2012).

For present purpose, statistics deals with the collection, organization, analysis, interpretation, and presentation of data. Using that broad definition, it sounds an awful lot like analytics. However, analytics and data science both use the quantitative underpinnings of statistics but their focus is wider than that of traditional statistics. While there are dozens of perspectives about the conceptual relationship between statistics and other disciplines (Taylor, 2016), I have represented what I see as the relationships among those concepts discussed here in Figure 1.3.

Mathematics has a certain absolute and determinable quality about it, and the way that math is taught (at least in US schools) imbues a deterministic way of viewing the quantitative world around them. That is, we are taught to believe that all facts and events can be explained. Statistics, on the other hand, views quantitative data as probabilistic or stochastic. That is, facts may lead to conclusions that may be generally true (beyond simple randomness), but it must be acknowledged that there is some random probability distribution or pattern that cannot be predicted precisely.

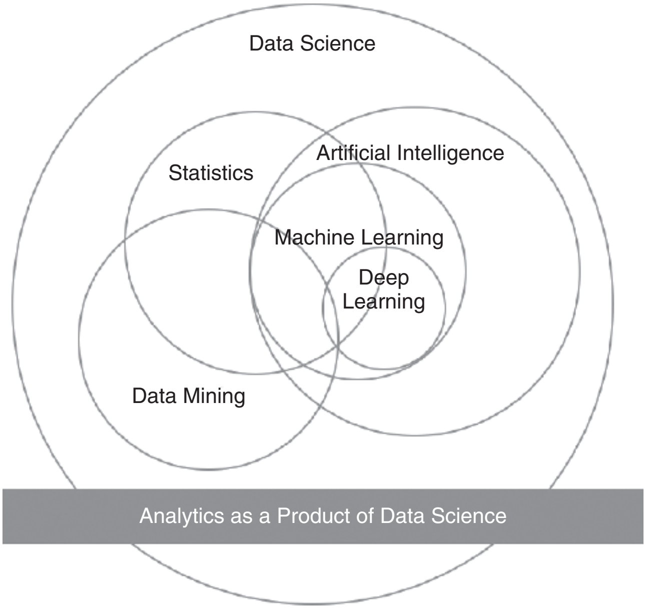


Figure 1.3 The relationship between statistics and other quantitative disciplines

Learn More

To learn more about the history of statistics and how it transformed science, please see David Salsburg's book *The Lady Tasting Tea* (Salsburg, 2002).

As shown in Figure 1.4, mathematical thinking is deductive (i.e., it infers a particular instance by applying a general law or principle) whereas statistical reasoning is inductive (i.e., it infers general laws from specific instances).

This difference is important in the context of analytics in that we apply both inductive and deductive reasoning to analytics problem solving. Thus, the application of both mathematics and statistics to analytics is appropriate and necessary. If analytics is a comprehensive strategy, then statistics and mathematics are tools in our proverbial analytics toolbox that help deliver on that strategy.

Linear programming, for example, can be used to support special types of problems in analytics that are loosely defined as an

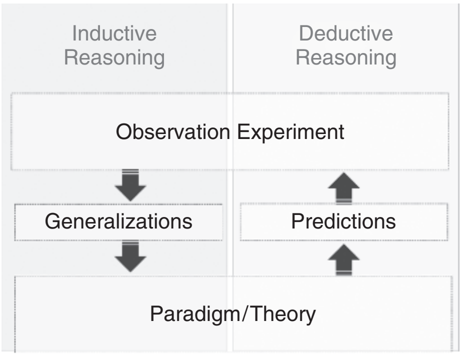


Figure 1.4 Inductive reasoning compared to deductive reasoning

optimization problem. For example, The Walt Disney Company uses linear, nonlinear, mixed integer, and dynamic programming in its data science work to support the optimization of restaurant seating, reduce wait times for park rides, and schedule staff (i.e., Cast Members). Note that I do not call out operations research, **mathematical optimization**, **decision sciences**, or **actuarial sciences** separately for the purposes of this discussion, as my perspective is that they serve as tools in our proverbial analytics toolkit—just as critical thinking and problem solving.

LINEAR PROGRAMMING

Linear programming is a mathematical method for problem solving where the output is a function of a linear model. For example, we might want to optimize emergency department throughput by looking at several factors, including surgical complexity, number of staff required, and potential complications, for example.

Forecasting and Time Series

In discussing the methods that support analytics, forecasting and time series methods are grouped together, not because they are the same thing but, rather, because they both fall into the same class of problem—the process of characterizing and predicting time-ordered data based on historical information.

Forecasting and time series refer to methods for analyzing time-sequenced data to extract meaningful characteristics from the data. Most often, forecasts are seen as trends represented as a visual display of historical data values, with some providing future predictions. Time series analysis is different than forecasting. You need time series data to make forecasts, but not all time series analysis is done to make a forecast. For example, time series analysis can be used to find patterns or similar features in multiple time series, or to perform statistical process control. Similarly, seasonality can be used to identify patterns.

Time series analysis utilizes a variety of approaches, including both quantitative and qualitative methods. The objective of time series analysis is to discover a pattern in the historical data (or time series) and then extrapolate the trend into the future. In Figure 1.5, note that there are generally four types of time series approaches.

Quantitative methods are the most common type of forecasting, but both qualitative and decision analysis approaches are in widespread

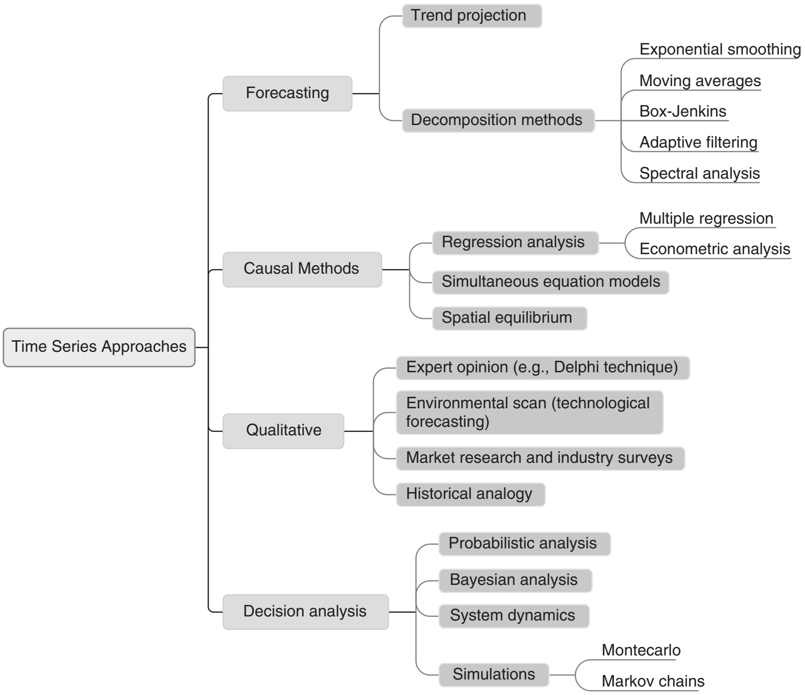


Figure 1.5 Approaches to forecasting and time series analysis

use, where historical, quantitative data may not be available or in cases where the "cone of uncertainty" is at its broadest (Saffo, 2007).

Natural Language Process

Natural language processing (NLP) refers to approaches used to understand and generate "natural language" through the use of computers.

NLP is a field of study that focuses on the interactions between human language and computers and sits at the intersection of computer science, artificial intelligence, and computational linguistics. Text mining and text analytics are often used interchangeably and refer to precursor activities to NLP as well as the application of NLP itself.

The goal of NLP is the understanding of natural language in computerized text. NLP is used for classification, extraction, and summarization, but the advances both in our understanding and in technology are quickly driving NLP to the forefront of a number of applications in analytics and beyond. For example, in analytics, we have historically taken information found in narratives (text, documents, tweets, speech) and processed them to categorize (taxonomy development) or to understand the sentiment. Sentiment analysis is particularly useful for understanding how people perceive products or services. In health care, sentiment analysis is used to measure "joy" in patients (Freed, 2017) as well as those at risk of heart failure (Eichstaedt, Schwartz, & Kern, 2015). These abstractions of text are then used as input to analytics processes such as predictive modeling, decision analysis, search, or question-answer applications.

Figure 1.6 highlights this as a generalized process.

A practical application of NLP can be seen in marketing; text is used to understand the overall "sentiment" of something (usually a brand or a product). Sentiment refers to the concept of understanding how emotions can be characterized in a body of work. Beyond sentiment analysis, NLP can be used in a variety of applications including:

- Grammar checking
- Entity extraction
- Translation
- Search
- Standardization
- Question answering

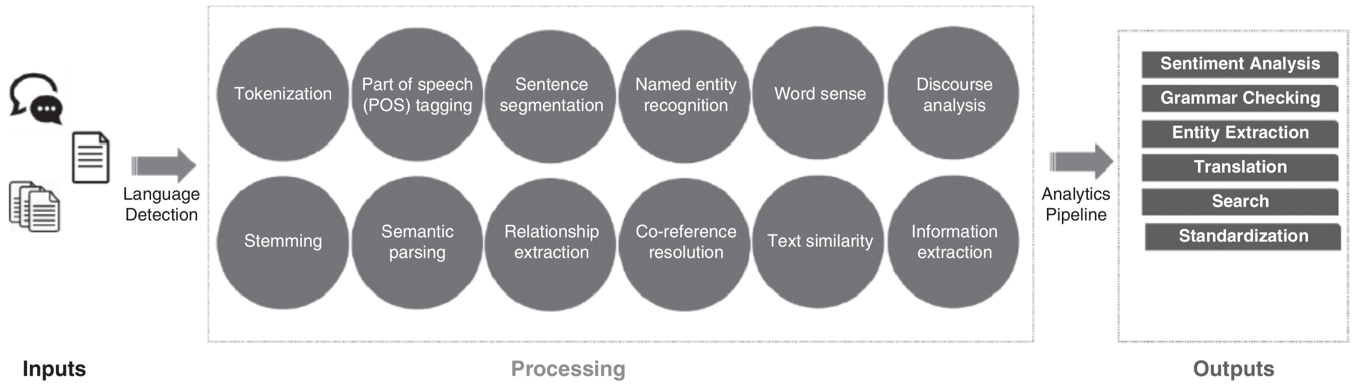


Figure 1.6 Natural language processing conceptual workflow

Learn More

To learn more about the terminologies used in natural language processing, please see the article by Matthew Mayo at https://www.kdnuggets.com/2017/02/natural-language-processing-key-terms-explained.html.

Natural language generation (NLG) is a subset of both artificial intelligence and NLP and is the process of automatically producing text from structured data in a readable format with meaningful phrases and sentences. Unlike NLP, the goal of NLG is to go the other way. That is, NLG takes data or some other form of information as input and produces text as output.

NLG has been popularized by chatbots which range in applications from customer service (Pathania & Guzma, Chatbots in Customer Service) to diagnosing symptoms (Facebook, 2017). Chatbots are only one application of NLG, and others include the automation of such things as:

- Summarizing business intelligence reports into complete narratives (Qlik, Tableau, TIBCO, Microstrategy, Sisense, Information Builders)
- Automatically creating financial reports complete with analysis (Nanalyze)
- Producing daily sports recaps (StatsMonkey)
- Providing automated performance reviews on customer service representatives (Quill by Narrative Science)
- Suggesting opportunities for customer relationship management systems by automatically creating CRM scripts (Yseop's Savvy)
- Helping a small business with a "financial analyst in a box" (Recount by Arria)

The field of natural language processing historically has involved the direct hand coding of rules—ontologies—that defined the structure, content, and context of words and how they are used in everyday language. Modern advances in statistical computing, computational linguistics, and machine learning are transforming the world of NLP at unprecedented rates.

Text Mining and Text Analytics

Perhaps one of the most confusing aspects of text analytics in general is the distinction between NLP and text mining. Think of this like **data mining**, where we are trying to extract useful information from data. The data, in this case, happens to be text and the extracted information includes the discovery of patterns and trends found in the textual data.

Text mining deals with the text data itself, where we attempt to answer questions such as the frequency of words, sentence length, and presence or absence of certain text strings. We can solve problems such as those outlined in Chapter 8 (e.g., classification using techniques found in NLP). In essence, text mining is often a precursor to NLP.

Text analytics often refers to advanced methods than span statistical analysis, machine learning, and other techniques but is generally recognized as equivalent to text mining. I think this is a gray area. Note the phrase *text analytics* is often used by the BI folks and represents more simple actions that can be done automatically and visualized via typical reporting means (e.g., word clouds, frequencies, etc.). Text mining would be the moniker used by the data scientists who have a wide swath of more advanced methods, but they would still do all of the counting of things as part of their effort. I think this fits my perspective that analytics is a natural evolution of BI and makes the important point of how different communities use different words, which can cause confusion. See, for example, www.linguamatics.com/blog/are-terms-text-mining-and-text-analytics-largely-interchangeable.

Machine Learning

SAS, the largest privately held US software firm and analytics giant, defines machine learning as (SAS, 2017):

> … a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

At its core, machine learning is a class of quantitative methods that uses algorithms to build analytical models, helping computer models "learn" from data. It differs from human-centered processes in that the computer learns to find patterns in the data rather than the person building the model directly. The concept of model building and model management, in general, is that it brings repeatability to ongoing decision making rather than the high touch analysis that often accompanies statistical analysis.

With recent advances in computing power, machine learning can be used to automatically apply complex mathematical calculations to big data that heretofore would have been impossible.

> *Humans can typically create one or two good models a week; machine learning can create thousands of models a week.*
>
> Thomas H. Davenport, Analytics thought leader (Davenport, 2013)

Figure 1.7 highlights common methods used in machine learning.

Learn More

To learn more about these and other terms used in machine learning, please visit the Google Developers Machine Learning Glossary found here: developers.google.com/machine-learning/glossary/.

Machine learning algorithms are most often categorized by the "learning style" (remember, machine learning is all about having computers learn what matters by looking at patterns in the data). That is, there are different ways an algorithm can model the real world (problem) based on the data it sees.

There are four learning styles, or learning models, that an algorithm can use. Each differs in the roles the input variable can take on and how the data must be prepared for the model.

Table 1.1 highlights the differences in machine learning algorithms.

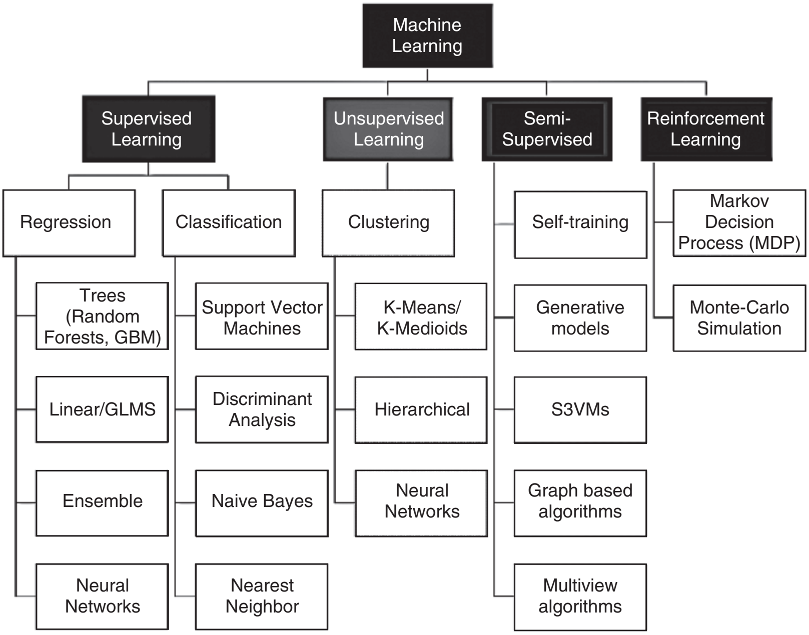


Figure 1.7 Techniques found in machines learning

Data Mining

DATA MINING

Data mining is a process of discovering and interpreting patterns in (often large) data sets in order to solve business problems.

Data mining was popularized in the late 1990s and early 2000s as a way to analyze large databases in order to generate new or novel information. The holy grail of data mining is finding the "needle in the

Table 1.1 Styles of machine learning

Learning Style	Model Building	Example
Supervised	The model is trained through a human-centered process of identifying which outcomes are examples (labels) of true cases versus those that are not.	Patients with and without a given disease are labeled in a historical dataset. The supervised model is intended to use that historical information to predict patients with the disease in a new (unseen) database.
Unsupervised	The model attempts to self-describe or organize data in an attempt to discover novel interactions.	In an exploration of consumer behavior, we might want to understand what distinguishes people who visit our website. Without an a-priori hypothesis, unsupervised methods can help us classify types of visitors.
Semi-supervised	In cases where labeling is impractical, the model is built by labeling "some" of the cases and letting the algorithm learn.	Perhaps the most common example of semi-supervised models is image categorization. The model can identify "events" in a picture—say, of a child with a basket on a grassy lawn at an Easter Egg hunt.
Reinforcement	The algorithm "decides" on a course of action in response to seeing a new data point, and the model is "rewarded" based on how good that decision was.	NLG algorithms can be taught to improve how it structures content and syntax, uses punctuation, or expresses colloquialisms in spoken dialogue systems (SDS) based on reward systems. This essentially maximizes rewards through reinforcement mechanisms.

haystack″ and differs from statistics in that there is not necessarily an **a-priori** theoretical-driven hypothesis before discovery can begin.

A-PRIORI

A-priori is defined as "from the earlier," or simply, beforehand. An a-priori hypothesis is one that is stated prior to an experiment being conducted or the data collected.

Data mining uses traditional statistical methods as well as artificial intelligence and machine learning techniques. It is ultimately focused on identifying previously unknown patterns in our data and in making predictions.

Just like other techniques in analytics, data mining follows a lifecycle that usually begins with framing the problem, then making sense of the data, doing model construction, and acting on the results. In typical fashion, the data miner identifies the outcome variable of interest and then uses a variety of techniques to pre-process the data (such as clustering, principal components analysis, and association rule learning), then applies those outputs as inputs to data mining algorithms such as regression, neural networks, decision trees, or support vector machines. A critical part of the data mining process is in model evaluation and ensuring that we don't overfit the model. I will discuss this in greater detail in Chapter 8.

THE GOAL OF ANALYTICS

> *Things get done only if the data we gather can inform and inspire those in a position to make a difference.*
>
> Mike Schmoker, PhD, author, former administrator,
> English teacher, and football coach

Analytics is a comprehensive strategy to support change. It informs interventions or change strategies. The goal of analytics is to support a data-driven, fact-based process of discovery. It is all about building confidence in our knowing and using that knowledge to understand, explain, predict, and optimize.

Analytics Is about Improving Outcomes

We analyze to understand, frame, and solve problems, make decisions, and create insights that can be used to drive change. We use what we know to make sense out of our worlds—that is, we "describe, discover, predict, and advise" (Blackburn & Sullivan, 2015). But advice falls short when analytics neither creates change nor produces outcomes. Results are interesting at best. The litmus test we should use is whether analytics has real-world impact. Fortunately, there are plentiful examples of how analytical thinking and its resultant products create change throughout organizations across various industries.

Analytics has the power to transform businesses and has proven its utility in hundreds if not thousands of examples. So why all the attention? To answer that, here are some of the outcomes that can be achieved with analytics.

Case Studies @ www.analyticslifecycletoolkit.com

See case studies organized by industry, method, and outcome.

In sum, the biggest opportunities for analytics may include those areas where the need involves:

1. An integrated, unified view of data
2. Going beyond description and discovery of the unknown to prediction, prescription, and optimization
3. A problem significant enough to be considered urgent and solvable

All three are required and not merely a "nice to have" to ensure that analytics has its permanent place in businesses.

Analytics Is about Creating Value

> *We are already overwhelmed with data; what we need is information, knowledge, and wisdom.*
>
> Dr. John Halamka, CIO Beth Israel Deaconess Medical Center

As indicated, outcomes are a critical component of analytics in that we must create something that is worthwhile. In discussing analytics, it's hard not to talk about value creation. After all, many view analytics in the same light as the countless information technology (IT) projects that they see fail. There are lots of reasons why projects fail (Bartels, 2017). As Jeremy Petranka, PhD notes, failed projects often relate to the fact that the linkage between the undertaking and organizational strategy is absent.

While the reasons may vary, I find in my advisory work that projects often fail when the fundamental value proposition was never fully realized. That is, the promise of the value to be "delivered, communicated, and acknowledged" was not accomplished (Value Proposition, n.d.).

Value has many definitions including the net result of [benefits − costs]. I prefer to look at value with a quality component, depicted as:

$$\text{Value} = \frac{(\text{Quality} + \text{Outcomes})}{\text{Cost}}$$

The quality component is critical in that analytics without quality presents risk, uncertainty, and unrealized potential. Quality comes in the form of robustness, repeatability, reliability, and validity. When the ratio is less than one—that is, when costs outweigh the quality plus outcomes—then we have failed to meet our fundamental value proposition (the reason for doing analytics in the first place). However, the full value isn't enough. Instead, we expect analytics to be a "force multiplier" (Kaufman, 2010) for organizations, and as such, it should be far greater. Of course, return on investment (ROI) for analytics isn't the only measure of value, as other things must be considered, such as mitigating risk, avoiding missed opportunities, or committing to improving the lives of customers, patients, or other stakeholders.

Rather, analytics is about creating value in that we achieve outcomes through the scientific approach we refer to as the **Analytics Lifecycle**. Analytics requires a multidisciplinary approach to achieving value.

ANALYTICS LIFECYCLE

Analytics Lifecycle refers to the series of changes that occur during the life of an analytics product. Throughout this book, we consider the evolution of a business question to what changes must be affected in order to improve the organization and its processes.

We will discuss tools and best practices for measuring value in Chapter 10 when we outline the "Value Management" as a critical component of Analytics Product Management.

Analytics Is about Discovery

Even more than what you think, how you think matters.

Dr. Atul Gawande, author and surgeon

If business intelligence (BI) is about knowing the knowable, then analytics helps us with knowing the unknown. As the adage goes, "We can never know something until we discover it." The power of analytics is that it supports discovery. We use our skills of reasoning and sensemaking to unearth patterns in data and are, in fact, often pivoting between deductive and inductive reasoning when we "problem solve" with data.

Discovery runs throughout this book and is specifically described in a number of best practice areas as a way to frame problems (Chapter 6), unearth patterns and discover relationships (Chapter 7), and activate analytics results (Chapter 9).

Analytics Is about Change

Don't rely solely on data to drive decisions; use it to help drive better leadership behaviors.

John W. Boudreau, PhD, professor and research director, University of Southern California's Marshall School of Business and Center for Effective Organizations

I know of few people who like and embrace change. Yet change is inevitable and perhaps the only constant organizations should count on. The impetus for change can come in many forms but for some organization it can come in the form of a crisis such as disaster, deregulation, declining profits, government mandates, failed systems, or public health scares.

Change has afforded entire industries the opportunity to transform how they operate. Take, for example, the case of the Oakland A's as depicted in the book *Moneyball* (Morris, 2014) and their use of analytics to drive competition. Major League Baseball has been transformed by analytics, and its decisions around players will never be the same.

I began this section with words like value, outcome, and impact. For me, these are key to analytics—driving change to improve outcomes and creating value. If we look at some of the most celebrated cases of analytics we see evidence of organizational change (i.e., the impact of how decisions are made or work is performed) as a result of analytics:

- Disney uses analytics and linear programming to optimize party size at restaurants throughout their resorts to optimize capacity and resource utilization.
- Cleveland Clinic uses advanced forecasting models to schedule operating room staff.
- Boston Public Schools improves how they make bus stop assignments to support over 25,000 school bus riders.
- University of Utah predicts outbreaks of Respiratory Syncytial Virus (RSV) three weeks before it happens for high-risk patients.

In each of these cases, whether we use analytics to improve customer experience, spur innovation, or redesign and optimize service delivery processes, change is inevitable, as the impact to the organization necessarily involves altering how work gets done.

We will explore the impact of analytics change throughout this book but special attention will be paid to this topic in Part III where we explore making analytics actionable.

CHAPTER SUMMARY

Analytics is resilient, in large part, because of its ability to impact the way that we work, the decisions we make, and the outcomes we achieve. Analytics is often seen hanging in the same circles as big data, data science, informatics, and even business intelligence.

However, analytics should be considered as an organizational strategy, and the **Analytics Lifecycle** is a set of best practices, each having complementary processes. In fact, I operationally defined analytics for the purposes of this book as *a comprehensive, data-driven strategy for problem solving*. This definition does not diminish those who would tie it to statistics, computational algorithms, data visualization, or massively large databases. But in doing so, it does acknowledge that (1) they aren't the same thing and (2) while useful, many of these things are tools and not disciplines required to engage in a data-driven, fact-based discovery, and problem-solving process. That is, I know plenty of capable, analytical thinkers who do not have a PhD in statistics.

Analytics should be framed as a process, as supported by the following observations. Analytics:

- … is not a destination, but rather the process of gaining insights to effect change. Analytics is the art and science of turning data into actionable interventions.
- … allows us to discover meaningful patterns in data and supports the examination of data with the intent of drawing a conclusion (taking action).
- … is not a technology, although technology is used to support the process.
- … is more than simply counting things or using basic math, but takes advantage of what we know about the past to predict and optimize the future.
- … can include but does not require computationally intense algorithms that can only be driven by the "data scientists" of Silicon Valley but rather by the curious—anticipators we call Data Champions.
- … necessarily creates an artifact that is used as input to a "decisioning" process; that is, the process of analytics creates a data product—large or small, reusable or not—that feeds another process.

Like many organizational strategies that have preceded modern analytics, we continue to evolve our thinking and raise our proficiency in how we make sense of the world. We benefit from the accumulated knowledge of our prior work including the scientific process, statistics, exploratory data analysis, data mining, artificial intelligence, data visualization, computational sciences, psychology and behavioral economics, lean thinking, Six Sigma, and so on. The world of analytics has benefited from each of their unique contributions as applied to thinking, learning, problem solving, decision making, and behavior change.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (or EDA) is the process of understanding data and determining what questions to ask before undertaking analytics.

In Part III of this book, we offer a practical perspective on *actioning* analytics. While not officially a word, actioning is used because it implies movement; that is, actioning is an organized activity to accomplish an objective or outcome.

REFERENCES

Bartels, E. (2017). Jeremy Petranka on IT strategy. Retrieved from events.fuqua.duke.edu/facultyconversations/2017/06/20/jeremy-petranka-on-it-strategy/.

Blackburn, F., & Sullivan, J. (2015). Field guide to data science. Retrieved from www.boozallen.com/s/insight/publication/field-guide-to-data-science.html.

Brillinger, D. R. (2002). John Wilder Tukey (1915–2000). *Notices of the AMS, February 2002*. Retrieved from Notices of the AMS website: www.ams.org/notices/200202/fea-tukey.pdf.

Champkin, J. (2013). George Box (1919–2013): a wit, a kind man and a statistician. *Significance*. Retrieved from www.statslife.org.uk/history-of-stats-science/448-george-box-1919-2013-a-wit-a-kind-man-and-a-statistician.

Copeland, J. (2000). What is artificial intelligence? Retrieved from www.alanturing.net/turing_archive/pages/Reference%20Articles/What%20is%20AI.html.

Cox, M., & David, E. (1997). *Application-Controlled Demand Paging for Out-of-Core Visualization*. Paper presented at the Proceedings of the 8th Conference on Visualization ’97, Phoenix, Arizona, USA.

Davenport, T. H. (2013). Industrial strength analytics with machine learning. Retrieved from blogs.wsj.com/cio/2013/09/11/industrial-strength-analytics-with-machine-learning/.

dictionary.com. (2017). Analytics. Retrieved from www.dictionary.com/browse/analytics.

Drucker, P. (1969). *The age of discontinuity: guidelines to our changing society* (1st ed.). New York: Harper & Row.

Eichstaedt, J., Schwartz, H. A., & Kern, M. L. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*.

Facebook. (2017). Florence chat. Retrieved from www.messenger.com/t/florence.chatbot.

Feldman, S. E. (2017). Cognitive computing. Retrieved from en.wikipedia.org/wiki/Cognitive_computing.

Freed, D. (2017). Joy for Facebook Messenger. facebook.com/hellojoyai/.

Granville, V. (2014). 16 Analytic disciplines compared to data science. Retrieved from www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared.

Hoffenberg, S. (2016). IBM's Watson answers the question, "What's the Difference Between Artificial Intelligence and Cognitive Computing?" Retrieved from www.vdcresearch.com/News-events/iot-blog/IBM-Watson-Answers-Question-Artificial-Intelligence.html.

Kaufman, J. (2010). *The Personal MBA*.

Marr, B. (2016). Will 'analytics on the edge' be the future of big data? Retrieved from www.ibm.com/think/marketing/will-analytics-on-the-edge-be-the-future-of-big-data/.

Merriam-Webster. (Ed.) (2017a) Merriam-Webster.

Merriam-Webster. (Ed.) (2017b).

Nelson, G. S. (2010). BI 2.0: Are we there yet? Paper presented at the SAS Users Group International.

Pathania, A., & Guzma, I. (Chatbots in Customer Service). Retrieved from www.accenture.com/t00010101T000000__w__/br-pt/_acnmedia/PDF-45/Accenture-Chatbots-Customer-Service.pdf.

Rose, R. (2016, June). Defining analytics: a conceptual framework. *ORMS Today, 43*.

Sack, J. (2012). Early human counting tools. Retrieved from mathtimeline.weebly.com/early-human-counting-tools.html.

Saffo, P. (2007). Six rules for effective forecasting. *Harvard Business Review*.

Saffron Technologies. (2017). Retrieved from saffrontech.com/saffronresources/.

SAS. (2017). Machine learning—what it is and why it matters. Retrieved from www.sas.com/en_us/insights/analytics/machine-learning.html.

Schlangenstein, M. (2013). UPS crunches data to make routes more efficient, save gas. Retrieved from www.bloomberg.com/news/articles/2013-10-30/ups-uses-big-data-to-make-routes-more-efficient-save-gas.

Schouten, P. (2013). Better patient forecasts and schedule optimization improve patient care and curb staffing costs. Retrieved from www.beckershospitalreview.com/hospital-management-administration/better-patient-forecasts-and-schedule-optimization-improve-patient-care-and-curb-staffing-costs.html.

Shuttleworth, M. (2009). What is the scientific method? Retrieved from explorable.com/what-is-the-scientific-method.

Taylor (2016). Battle of the data science Venn diagrams. Retrieved from http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html/2.

Value Proposition (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Value_proposition.

Wolfram, S. (2010). Making the world's data computable. *Stephen Wolfram Blog.* Retrieved from blog.stephenwolfram.com/2010/09/making-the-worlds-data-computable/.

Woodford, C. (2016). Speedometers. Retrieved from www.explainthatstuff.com/how-speedometer-works.html.