1  
Introduction to Analytics

*Philip T. Keenan, Jonathan H. Owen, and Kathryn Schumacher*

*General Motors, Global Research & Development, Warren, MI, USA*

1.1 Introduction

We all want to make a difference. We all want our work to enrich the world. As analytics professionals, we are fortunate–this is our time! We live in a world of pervasive data and ubiquitous, powerful computation. This convergence has inspired new applications and accelerated the development of novel analytic techniques and tools, while breathing new life into decades-old approaches that were previously too data- or computation-intensive to be of practical value. The potential for analytics to have an impact has been a call to action for organizations of all types and sizes. Companies are creating new C-level positions and departments to grow analytic capability. A torrent of new start-ups have formed to sell analytics products and services. Even governments have created new high-profile offices to leverage analytics. These changes have driven a surge in demand for analytics professionals, and universities are creating departments, curricula, and new program offerings to fill the gap.

But what exactly do we mean when we say “analytics”? The term is widely used, but has vastly different meanings to different people and communities. A number of well-established disciplines, including statistics, operations research, economics, computer science, industrial engineering, and mathematics, have some claim to “analytics” and interpret it to have specialized meaning within their domains. The popular usage of the term is often comingled with other widely used but equally overloaded terms such as “big data,” “data science,” “machine learning,” “artificial intelligence,” and “cognitive computing.” As a result, this seemingly innocuous term has led to much confusion over the last decade as people using the same language often talk right past each other. In the authors' own experience, frustration at all levels of an organization is inevitable when well-intentioned and intelligent people believe they have a shared understanding–on a new project initiative, for example–only to discover weeks or months later that there was a fundamental misunderstanding of what work was to be performed or insights delivered.

In a 2016 article intended to reduce some of this confusion, Robert Rose identified three main usages of the term “analytics” [1]:

1. As a synonym for metrics or summary statistics
2. As a synonym for “data science” (another overloaded term)
3. As a very general term to represent a quantitative approach to organizational decision-making

Our use of the term is closest to the last of these; we consider analytics broadly as a process by which a team of people helps an organization make *better decisions* (the objective) through the *analysis of data* (the activity). This chapter gives a brief, high-level introduction to the subject. We first describe a conceptual framework for analytics, and define three primary categories of analytics (descriptive, predictive, and prescriptive). We then discuss considerations for applying analytics within an organization, and briefly discuss the ethical implications of using analytics. Subsequent chapters dive more deeply into each component of the process of applying analytics, including developing a request for a new project, building a cross-functional team, collecting data, analyzing data with a wide variety of mathematical and statistical methods, and communicating results back to the client.

Interview With Alan Taber

*Alan Taber, System Engineer with Lockheed Martin Missiles and Fire Control, defines analytics in the following way:*

Analytics is both a mindset and a process. The mindset is that instead of simply reacting to what you perceive your environment to be that you gather data understanding the limits and bounds of that data. You feed it into a model. It can be a very detailed model or a simple model about how situations evolve over time if you do take options A or B or C, or some combination thereof, and then you test that hypothesis. You have the continual feedback loop to say if what you're doing makes sense and also keep an eye on your surroundings because what may have made sense a year ago or a month ago may no longer make sense. That's the mindset, to always be paying attention rather than running on autopilot.

The process is to make sure you understand the root problem, figure out if you can frame that as a problem that's amenable to being solved with data, figure out your data sources, and don't limit yourself to the data you have on hand and know how to collect. If you need a different data set, go get it. Once you have your data and can run your test, do that. Over and under and around all that, you're working with your stakeholders so that when you deploy people are familiar enough with what you're doing that they're willing to try it out rather than saying, “I don't understand the model and therefore I'm busy, I don't have time to learn, I'm not interested.” If you are overwhelming people with information but not helping them actually solve the problems that they perceive they have, you simply will not get very far. You will have wasted all your time. So that's the mindset and that's the process.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

1.2 Conceptual Framework

As shown in [Figure 1.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0001), the generic analytics process can be viewed as a continuous cycle where the analysis of data produces insights that inform better decision-making. We use this simple figure to highlight two fundamentally different approaches to analytics: *data-centric* and *decision-centric*.

A diagram of a process

Description automatically generated

[**Figure 1.1**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF1) Simplified visual representation of the analytics process.

1.2.1 Data-Centric Analytics

The philosophy behind *data-centric* analysis is to “let the data speak freely.” Working under this philosophy generally involves pulling together as much relevant data as possible, analyzing that data to identify patterns that lead to insight, and serving up those insights to a decision-maker who (hopefully) will make better informed decisions. As shown in [Figure 1.2](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0002), this follows the natural (clockwise) flow of the analytics process.

A diagram of data analysis

Description automatically generated

[**Figure 1.2**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF2) The *data-centric* approach starts with the data to surface insights.

Not surprisingly, the data-centric approach has gained popularity with the surge in “big data.” Many of the analytic methodologies employed in this arena–including data mining and classification, machine learning, and artificial intelligence–increase in effectiveness with the volume of data available for analysis. Advocates believe that we are in a new “machine age” that is changing the landscape of business and the world [2–4]. Some argue that the data-centric “big data” paradigm is really about eliminating sampling error; they claim that we are no longer reliant on small samples since we have storage capacity to hold and computing power to process vast amounts of data [5]. Others have observed that the promised insights have not always materialized, and that the challenge is “to solve new problems and gain new answers–without making the same old statistical mistakes on a grander scale” [6].

1.2.2 Decision-Centric Analytics

*Decision-centric* analytics begins with an understanding of the *decision* that needs to be made and what *insights* would lead to better expected outcomes. Decision-centric models typically encapsulate subject matter expertise (SME) and codify domain knowledge in order to relate decision variables to the target objective. Data requirements are determined by the chosen analytical model; ideally these data already exist in a convenient form, but often they must be extracted from disparate sources or collected through new instrumentation or market research. As summarized in [Figure 1.3](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0003), this approach starts with the final outcome–the decision–and works backward (counterclockwise) at each step to define and develop needed analysis and data resources.

A diagram of decision making

Description automatically generated

[**Figure 1.3**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF3) The *decision-centric* approach starts with the problem and works backward.

Decisions are often defined as an “irrevocable allocation of resources” [7]. Improving decision-making requires an understanding of the desired outcome (the objective), alternative actions (decision variables), and boundary conditions (constraints), but also the richer context of possible future conditions (scenarios). It also requires that we answer several softer questions: Who is making the decision? What is her or his scope of control and influence? What information is already available to the decision-maker(s) and where are the gaps? In a decision-centric approach, many of these questions are considered as part of upfront framing activities that look ahead toward operational implementation.

1.2.3 Combining Data- and Decision-Centric Approaches

Analytic practitioners and professional communities are often predisposed to either data-centric or decision-centric approaches. In the authors' view, this is attributable to different pedagogical perspectives and experiences. Given the centrality of computing and information technologies for handling large amounts of data, it is not surprising that many organizational IT functions are naturally aligned with a data-centric view. Business operations and the analytic teams that support them often have a natural affinity for decision-centric approaches that leverage their deep understanding of key problems and models that support improvements. [Table 1.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-tbl-0001) summarizes salient features of the two approaches.

[**Table 1.1**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backT1) Comparison of data-centric and decision-centric approaches.

|  | **Data-centric analysis** | **Decision-centric analysis** |
| --- | --- | --- |
|  | **(Data science, computer science)** | **(Decision science, operations research)** |
| Mantra | “Start with the data” | “Start with the decision” |
| Philosophy | Leverage large amounts of data. Let the data “speak freely” by identifying patterns and revealing implicit (hidden) factor relationships | Leverage domain knowledge and subject matter expertise to model explicit variable relationships |
| Data | More is better, especially for “big data” applications (e.g., speech or image recognition) | Custom collection of curated data sets |
| Computing | High-performance computing is often price of entry. Potential need for specialized processors (e.g., GPUs, TPUs) for acceptable execution speeds, especially in contexts requiring real-time analysis | Desktop or server-based computing is typical. Trade-offs between potential benefits of leveraging high-performance computing versus added overhead in development and maintenance |
| Pros | * Increasingly automatable * Potential to extract weak signals from large, unstructured data sets | * Causal focus * Strategic value beyond historical observations |
| Cons | * Risk of conflating correlation with causation * Analysis inferences are limited by history * Noisy data with confounded effects | * Human subject matter expertise required * Cost of data acquisition can be high |
| Key disciplines | * Computer science * Data science * Machine learning and unstructured data mining * Artificial intelligence (AI), deep learning | * Management and decision sciences * Operations research * Mathematics * Classical statistics |
| Example applications | * Image classification * Speech recognition * Autonomous vehicle scene recognition | * Supply chain optimization * Scenario planning * New business model development |

Important opportunity arises from combining elements of the two approaches. There is undeniable potential to leverage increasingly pervasive data and computational power associated with data-centric analysis, but contextual knowledge and subject matter expertise provide needed guardrails so that the resulting insights are meaningful.

Acknowledging the natural tendencies of individuals or analytics organizations toward data- or decision-centric approaches may help practitioners to identify growth opportunities. For example, traditionally decision-centric organizations may benefit by expanding the amount of data used in their analyses, including unstructured data sources. Typically, data-centric groups may improve the fit and predictive power of their models by incorporating domain-specific expertise.

Evidence of the benefit of utilizing a combined approach is seen in recent movements to incorporate “thick data” into marketing analytics (see Refs [8,9], for example). Combining thick data, such as ethnographic studies or focus group responses (see [Figure 1.5](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0005)), with big data, such as transaction data, enables a more complete understanding of customers' preferences and behaviors. Decision-centric framing, domain knowledge, and deep subject matter expertise collectively provide scaffolding that helps big data insights take shape.

1.3 Categories of Analytics

A well-known and useful classification scheme for analytics was proposed by Lustig et al., at IBM [10]. Based on their experience with a variety of companies across a diverse set of industries, they defined three broad categories of analytics: *descriptive*, *predictive*, and *prescriptive*. As summarized in [Figure 1.4](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0004), there is a natural progression in the level of insight provided–and potential value–as an organization moves from descriptive to predictive and ultimately to prescriptive analytics. Typically there is also a progression in the mathematical sophistication of the analysis techniques, as well as the organizational maturity required to absorb and act on resulting insights.

A screenshot of a diagram

Description automatically generated

[**Figure 1.4**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF4) Categories of analytics.

1.3.1 Descriptive Analytics

The purpose of descriptive analytics is to reveal and summarize facts about what has happened in the past or, in the case of real-time analysis, what is happening in the present. This is done by examining and synthesizing data collected from a variety of sources. Raw data are captured and recorded in source systems, eventually to be cleaned, retrieved, and normalized such that entities and relationships can be meaningfully understood. The audience for descriptive analytics is broad, potentially reaching all functions and levels of an organization. Descriptive analytics are at the heart of most business intelligence (BI) systems.

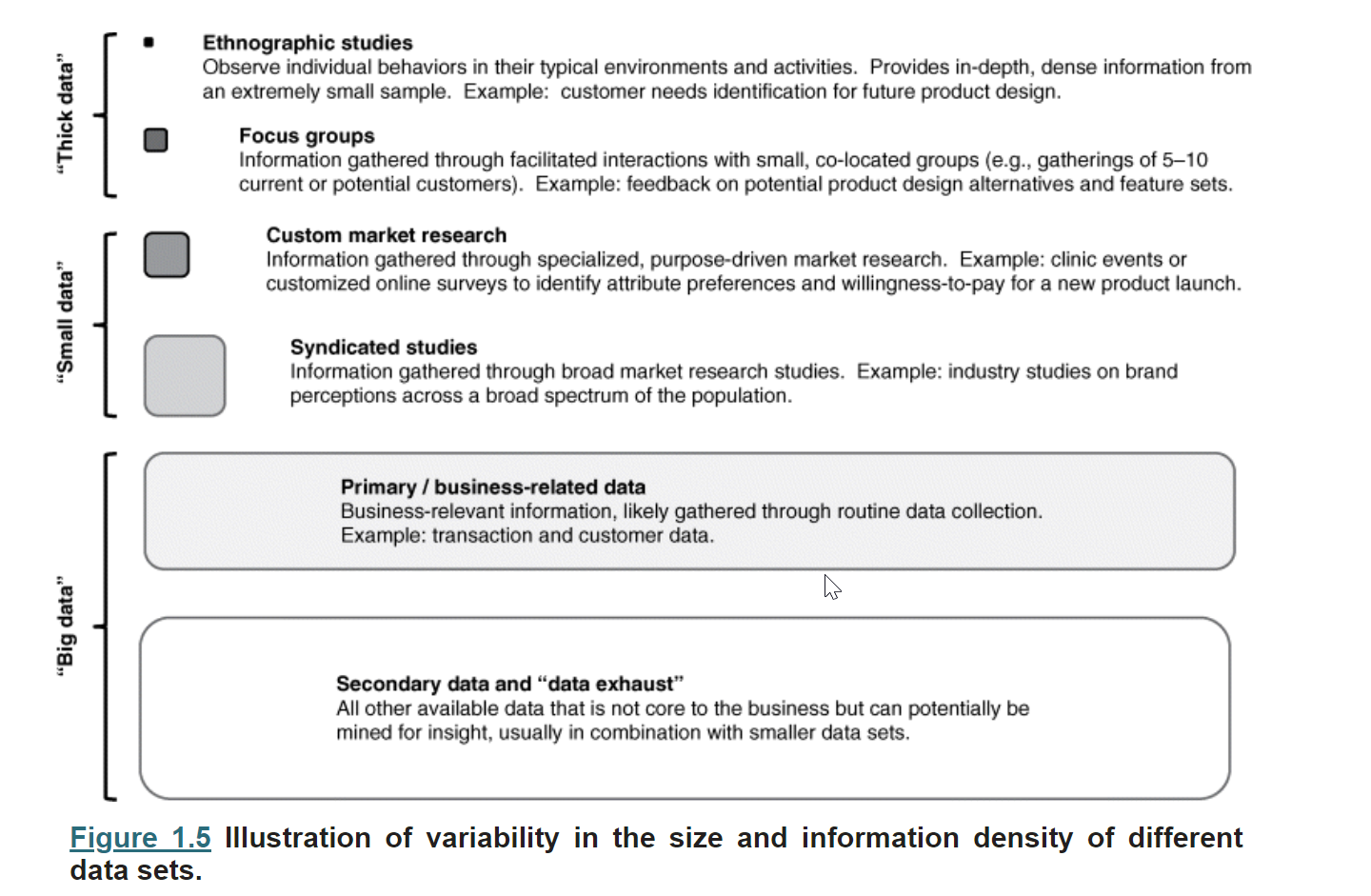
Data Modeling

Many organizations have access to vast quantities of data. Useful descriptive analytics generally involves processing the raw facts into higher level abstractions. Data scientists think in terms of *entities* and *relationships*. For example, a customer database might contain entities like “Household” and “Product,” linked by relationships like “Purchased,” with data elements including the demographics of the households and the price, cost and features of the products.

Sources of data can be highly varied (see [Table 1.2](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-tbl-0002) for examples), as can the size and information density of any given data set (see [Figure 1.5](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0005)). There is also high variability in the expense and effort required to collect different types of data. On one end of the spectrum, ethnographic studies require social scientists to spend many hours shopping with or interviewing individual customers, and thus the data are very carefully curated and very expensive to collect. On the other end of the spectrum, “data exhaust” is logged nearly for free, including data generated from smartphones and online activity [11]. Data exhaust is collected without a specific intended purpose and can be especially messy, so substantial cleanup effort is usually necessary before this type of data are usable.

[**Table 1.2**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backT2) Potential sources of data.

| **Source** | **Examples** |
| --- | --- |
| Transaction data | Data associated with a transactional event. Example: a purchase transaction with details of the specific item purchased, where and when it was purchased, the price paid and any discounts applied, how the customer paid (e.g., cash, credit card, finance), and other contextually relevant data (e.g., inventory of other items for sale at the same time and location) |
| Customer data | Data associated with customers. Examples: detailed demographic or psychographic information on individuals and households, history of interactions (past purchases, Web site visits, customer service requests) |
| Sensor data | Data collected through electronic or mechanical instrumentation. Examples: web browser cookies tracking customer activity, electronic sensors monitoring weather conditions, airplane flight data recorder information |
| Public data | Open-source data from individuals, organizations, and governments. Example: aggregated census data |
| Unstructured | Data without known structure. Examples: text and images from social media, call center recordings, qualitative data from focus groups or ethnographic studies |
| Curated data | Data collected for a specific purpose with downstream analysis in mind. Examples: consumer surveys, designed market research experiments |



[**Figure 1.5**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF5) Illustration of variability in the size and information density of different data sets.

Developing a data model that captures the structure and relationships among the different data elements is a fundamental task. Generic data models are often constructed to efficiently store ingested data, without specific analytic use cases in mind. Although such data models can be useful for general-purpose reporting and data exploration, purpose-built data models are typically needed for efficient analysis. Depending on the size of the organization and the speed with which new data arrives, substantial IT support may be required to run systems that capture and record data, clean it, and store it in a warehouse or lake for eventual retrieval and analysis.

Reporting

The real value of descriptive analytics comes from putting access to this plethora of data into the hands of analysts who can use it to rapidly answer questions. To this end, Lustig et al. proposed a classification of descriptive analytics into three areas [10]:

1. Standard reporting and dashboards
2. Ad-hoc reporting
3. Analysis/query/drill-down

In our experience, standard reporting and dashboards are useful to a point, but users need to be able to “slice and dice” the data on the fly to gain more meaningful insights, computing summary statistics and visualizing comparisons without being limited to predefined reports.

Visualization

Descriptive analytics is often about communication, not math. Authors such as Tufte [12] provide useful guidelines for describing and visualizing data in ways that reduce the cognitive burden on those who must interpret the results. Later chapters will go into more depth on this; however, since the topic is so important, we will elaborate on it later in this chapter (Section 1.4.2) as well when we discuss the communication of project insights.

Software

Software for descriptive analytics is plentiful. At the most basic level are ordinary spreadsheets and databases. At the other end of the spectrum are systems designed specifically to support data visualization, exploration, and reporting–such as Cognos, Tableau, and Spotfire. These systems can greatly increase the accessibility of data and basic analytic insights throughout an organization.

1.3.2 Predictive Analytics

Descriptive analytics describe the world as it is (or as it recently was). In contrast, *predictive analytics* seek to forecast the likely future state of the world through a deeper understanding of the relationships among data inputs and outcomes. This is a much more demanding goal, so there is much more that can go wrong. Inexperienced analysts and leaders often imagine that once you have a good descriptive model, you can use it to make good forecasts. Not true! Statisticians have long understood that correlation does not imply causation. As a result, teams that wish to forecast the future need to use more sophisticated modeling approaches and follow more rigorous validation procedures if they want to have confidence that their forecasts make sense.

As a very simple example of the difference between descriptive and predictive analytics, consider television programs that cover the stock market. Every day, talking heads explain why the stock market behaved the way it did the previous day. But can any of them accurately forecast what the market will do tomorrow? Not a one. If they could, they would be billionaires living on a beach, not reading off a teleprompter in a TV studio. Hindsight may be 20–20, but foresight certainly is not.

Data Mining and Pattern Recognition

The starting point for predictive analytics is often mining data to identify meaningful relationships and patterns. As we work with increasingly large and diverse data sets, there is a growing opportunity to identify hidden relationships that relate disparate data. For example, clustering analysis might be used to segment customer populations into groups that go beyond simple demographic or psychographic characteristics. Or we might apply various machine learning techniques to identify objects and trajectories for autonomous vehicle scene recognition and navigation.

The set of available data mining techniques is highly varied, and practitioners need to be adept at selecting appropriate methods based on an understanding of the pros and cons of each within a given application context. Many methods are based on classical statistical models, often to classify populations into distinct groups (e.g., classification and regression trees) or to estimate the impact of a set of descriptor variables on a metric of interest (regression). Machine learning and artificial intelligence techniques can arguably answer a broader set of questions (e.g., image recognition), but trade the transparent simplicity of classical models for a harder-to-explain “black box” capable of representing more complex relationships. Regardless of the methodology, analysts must be alert to the danger of false positives. Given enough computer time and input data, one can *always* find some sort of “statistically significant” effect that is actually pure noise.[1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-note-0001)

Predictive Modeling, Simulation, and Forecasting

Predicting the future requires a model. Simply collecting and reporting data, or identifying interesting patterns about the past and present is not sufficient.

One of the simplest models assumes that the future will behave like the past; for obvious reasons, this is often referred to as a naive model. For an established company, sales next month will likely be similar to sales last month. However, leaders who request analytics projects generally want deeper insights than that! The next simplest model is trend extrapolation. If sales were 100 units in January, 110 in February, and 120 in March, it seems plausible to predict that they will be 130 in April and 140 in May. Projecting simple trends can be useful, but it is not always appropriate. Suppose you are selling tax preparation software; this forecast would be inaccurate, as sales in May will instead be close to zero, since most customers will have filed their taxes with the IRS by April 15. In this context, a more advanced model that “seasonally adjusts” the data would be appropriate.

More sophisticated models often include other explanatory variables in addition to time. For example, when trying to predict the number of vehicles the US automotive industry will sell next year, it is often helpful to consider macroeconomic data such as the unemployment rate, interest rates, and inflation. The automotive industry is cyclical–sales fall during recessions and rise during periods of economic expansion. Predicting the timing of the next recession can be almost as challenging as predicting the future course of the stock market. As a result, predictive models generally need to report ranges, or uncertainty bounds, rather than simple point forecasts. Unfortunately, many clients have difficulty consuming range estimates and prefer to pretend that point forecasts suffice. This is one of the many challenges the analytics practitioner faces when trying to communicate results in a form accessible to decision-makers.

Deciding what variables to include in a model can also be challenging. Leave out an important causal factor and the model's predictions may be seriously wrong. Including extraneous factors can also cause difficulties. For instance, classical regression models can fail if several input variables are closely correlated, an issue known as multicollinearity.

Analysts often attempt to assess the goodness of fit of their proposed model. For example, when fitting a regression model, most software packages report the “R-squared” metric, a measure of how closely the model matches the data. Analysts often construct a variety of models (perhaps using different subsets of variables in each) and pick the one with the highest R-squared. Unfortunately, this technique of “chasing R-squared” is not, in fact, a good approach–it can easily lead to overfitting, which in turn can lead to poor performance when predicting future values.

To avoid this pitfall, analysts can instead divide the data into a “training sample” used for fitting the model, and a “validation sample” used for assessing and comparing models after they have been fitted. Executed properly, this methodology can dramatically reduce the risk of overfitting, so it should be standard operating policy for all analysts whenever sufficient data are available.

Leveraging Expertise

There are a great many methodologies available for building predictive models. Frequentist statistical models have been used for over a century. Bayesian statistical models became widely used starting around 1995, when faster computers and algorithms made them computationally practical. Machine learning methods have become popular in recent decades, made possible by faster computers and larger data sets. Statistical and machine learning methods work well for analyzing a vast array of situations, but they tend to rely on the computer to *discover patterns* in the historical data and assume these patterns will repeat in the future. However, sometimes the future is different from the past. For example, when launching a new product, historical sales data are not available. How then to predict future sales?

Potential solutions have been developed for such cases, but they are substantially more complicated and time consuming (i.e., expensive) than methods that make use of existing data. For example, when launching a new product, one such approach is to perform primary market research to test how potential customers react to the new product.

In some situations, a practitioner has abundant knowledge of the structure of the real world, and incorporating that knowledge into the model building process can be extremely valuable. Simulation models are particularly useful in such situations. Simulation is based on the understanding of how some entities–individuals, components, or other actors–behave in isolation, and how their interactions lead to consequences under different scenarios. Simulation techniques can be classified based on what interacts and how the interactions occur. [Table 1.3](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-tbl-0003) summarizes key differences between three common types of simulation models: discrete event, agent-based, and system dynamics.

[**Table 1.3**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backT3) Comparative summary of three common simulation models.

|  |  |
| --- | --- |
| Discrete event simulation | Models a system using a central global mechanism, often a network, within which entities interact according to centrally specified rules at discrete points in time (events). Interactions are defined by standardized structures such as queues. Example: call center and discrete manufacturing operations analysis |
| Agent-based simulation | Models a system using autonomous agents (representing both individuals and collective groups), each with their own rules for behavior. Interactions are determined by domain-specific rules potentially based on the state of the agents involved and the overall state of the system. The overall system behavior emerges from the interactions of the agents. Example: flight simulation for a flock of birds |
| System dynamics | Models a system using stocks and flows. Interactions are defined by feedback loops and control policies. System dynamics is to agent-based simulation as thermodynamics is to molecular simulation, in that it aims to reduce the computational and cognitive burden through aggregation. Example: Bass diffusion model of the impact of advertising |

Simulation models require a lot of effort to calibrate to observed history. However, because they model the underlying “physics” (e.g., microeconomics) of the situation, they can incorporate additional data from subject matter experts or market research. Simulation models can be used to evaluate “what-if” scenarios, a capability that is very useful to decision-makers, and is not possible with basic forecasting models.

1.3.3 Prescriptive Analytics

*Prescriptive analytics* seek to go further than forecasting a future state, to make actionable recommendations about what the decision-maker should do to achieve a particular objective, such as maximize profit. With descriptive and predictive analytics, the analytics team shoulders most of the burden of interpreting the results and developing recommendations for action. With prescriptive analytics, the computer helps with that process by evaluating a large number of potential alternative courses of action and reporting the best ones. The team still needs to apply a level of business judgment in interpreting the answers, since all models are incomplete descriptions of reality. Nonetheless, this sort of analytics has the greatest potential to help decision-makers realize tangible benefits through better decision-making.

However, automating the process of generating actionable recommendations requires a higher standard for defining causal relationships. Consider the following hypothetical example. Suppose you develop a time series model that attempts to forecast US automotive sales using imports of cheese from Mexico as the explanatory variable. You may find that the model fits the data well (it is descriptive). You may well also find that the prediction it makes (more cheese imports correlates with more vehicle sales) also turns out to be accurate year after year into the future (it is predictive). Nevertheless, if you were to then make the prescriptive recommendation that auto manufacturers should lobby Congress to reduce tariffs on Mexican cheese in order to stimulate car sales in the United States, you would be making a very foolish error. The relationship is spurious. There is no causal connection, so reducing tariffs would have no actual effect on vehicle sales. Instead, both cheese sales and vehicle sales are correlated with overall gross domestic product (GDP): when people have more money to spend, they use it for cheese and for cars; when they have less, they defer both kinds of purchases.

The lesson of the tale is clear: you need to first understand how the real-world business situation works, and model it appropriately. One huge risk of “big data” is that analysts will simply throw a huge quantity of data at a machine learning system with no thought about what kinds of relationships are plausible. In some settings this is not an issue (think “people who shopped for X also shopped for Y” recommendation engines). But in other settings, recommending nonsensible actions may destroy credibility.

No one knows the future. What we can hope to achieve with prescriptive analytics is simply to help decision-makers make the best decision possible, given the best data available at the time.

Prescriptive analytics typically require a combination of simulation and optimization. You begin by determining what quantity you wish to maximize–for example, the net present value of operating your business. Next, you list the decision levers available to you, such as investments in advertising, new product development, or price cuts for existing products. Next, you build and calibrate a model that is robust under a wide variety of ways of pulling the levers. This may require something like a system dynamics model, since it may need to capture scenarios in which the future does not look like a simple trend extrapolation of the past. Finally, you embed the simulator inside an optimization loop that evaluates a large number of different ways of setting the decision levers and tells you which one maximizes your objective, for example, is most profitable. The optimizer frequently needs to deal with various sorts of constraints, for instance, some decision levers are discrete, others are continuous, and some economic variables, like price and sales volume, cannot be negative.

Prescriptive models must also consider how entities outside of your control (e.g., competitors) will behave or react to your decisions. These may be “random,” as in Monte Carlo simulation, or “strategic,” as in Game Theory. Real life generally includes both.

For a real-life example, consider “Modeling General Motors and the North American Automobile Market” [13]. The client was the then-President of GM North America. The goal was to maximize future profitability. The team developed a system dynamics simulation model combining internal activities such as engineering, manufacturing, and marketing with external factors such as the competition for consumer purchases in the new and used vehicle marketplaces. Eight groups of automotive manufacturers competed for a decade across 18 vehicle segments, making monthly segment-by-segment decisions about price, volume, and investment in future products. The model included Monte Carlo simulation of random effects, such as how attractive future competitor vehicles turned out to be once they entered the marketplace, and when the next recession would occur. This was then embedded inside an optimization loop that evaluated alternative strategies. Instead of point forecasts, it generated probability distributions on future profitability, as illustrated in [Figure 1.6](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0006). Ultimately it was able to show that despite future uncertainty, following a particular proposed strategy (B) would produce a probability density shifted to the right (i.e., toward higher profits) as compared to following an initial strategy (A). This supported a *prescriptive* recommendation to enact strategy B.

A graph of a graph of a strategy

Description automatically generated with medium confidence

[**Figure 1.6**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF6) Output from an example prescriptive analysis of alternative policies [13].

Just as with descriptive and predictive models, prescriptive models require substantial amounts of business judgment and work best when the team iterates between analyzing scenarios and discussing them with subject matter experts. No computer model is perfect. The data may contain valuable information, but inevitably you will get better results if you also incorporate subject matter expertise. At a minimum, this expertise is necessary for qualitatively interpreting the results, and when possible can also be quantitatively incorporated into the model itself.

1.4 Analytics Within Organizations

Suppose you have decided you want to do analytics within your organization. How do you get started?

Until recently, in many large organizations this involved a lot of pushing. Analytically minded employees would see an opportunity, perhaps even build a prototype analysis tool for a particular business challenge, show it to management, and then often watch it die a quiet death at the hands of leaders who did not understand the potential benefits of analytics, or who felt threatened by the thought of being replaced by a computer program.

In the last decade, however, things have changed dramatically. Analytics has become a senior management buzzword and a prominent topic of articles in publications like *Harvard Business Review* and the *McKinsey Quarterly*. These days, it is no longer a question of you, an individual employee, wanting to get more involved. Now the question is: “Your organization has decided it needs to do more analytics. How does it get started?”

The answer is of course unique to each organization, but we will make some general comments, first about the life cycle of an individual analytics project, and then about the alternative ways an organization can implement such projects.

1.4.1 Projects

Analytics projects work best when you have three key ingredients: (1) quantitative analytics professionals who are well-versed in the data and appropriate analytic techniques, collaborating closely with (2) subject matter experts who understand the problem domain, and (3) leadership sponsors in the core business who understand the value of better data-driven decisions and will champion implementation in the organization.

A new analytics project typically begins with a conversation between executives, one with operating responsibility for a difficult business decision and the other with experience doing analytics projects. If they are able to communicate effectively, they will be able to jointly write a framing document: a statement of the problem to be solved that also describes the scope, outputs to be delivered, and a high-level description of the kinds of input data and analytical frameworks that will likely be helpful in creating the desired outputs. The framing document should also include a list of stakeholders whose engagement will be needed to see their project through to implementation.

Next comes a stage we call “invent and pilot.” This is a highly iterative process. The stakeholders assemble a cross-functional team combining analytical experts with business experts. The team gets up to speed on the business problem, obtains samples of available data, tries a variety of methods for analyzing it, discusses the results of each, and eventually settles on an approach that is feasible to execute within the time and resource constraints of the project while also delivering results that make actual business sense to the end clients.

Next comes “productionization.” In a small organization, this could be as simple as providing the client with a spreadsheet. In a large organization, this may be a much longer and more expensive process involving the internal IT organization. Typically IT support is necessary to automate the data feed into the analytical environment, and to provide data security for both the inputs and the results of the analysis. Ideally, IT also provides services such as data cleaning, although often this is beyond their scope and falls to the analytics team instead. This can be a huge undertaking, since a great many real-world data sets have missing values, incorrect values, and are inconsistent with other data sets that are needed for the same project.

IT may also choose to develop some sort of delivery platform, such as a custom app or Web site, in order to simplify the user experience for end client users and to help maintain control of the data for security purposes.

Finally, IT deploys the solution to the client. Typically the analysis team continues to play a major role for the first year or so, conducting ongoing analysis and presenting it to leadership, as well as training people in the client organization to use the system. Often a change management process is required, since the new analytics based method of making decisions may involve a very different process than the one people in the organization are familiar with. It is best if some members of the client organization were participants in the cross-functional analytics team from the beginning, but at a minimum, some members of the client team must be trained as “superusers”–people who can load data, run the model, and present and interpret results, all without requiring much support from the analytics experts who built the system initially.

Additional activities (e.g., training, security, help/support) are often needed to sustain an analytics capability over time and support ongoing business use. As users become more sophisticated with experience and grow in their ability to leverage insights, new questions arise that require model enhancements. The complete life cycle of a typical analytics project is summarized in [Figure 1.7](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0007).

Interview With Eric Stephens

*When asked to identify the key skills needed to obtain the problem definition/problem statement, Eric Stephens, Manager of Population Health Analytics at the Vanderbilt University Medical Center, responded as follows:*

These aren't necessarily going to be in any particular order, but first and foremost I think is communication. This means the ability to listen, as well as to speak and write. In fact, listening is probably even more critical in this context than it may be in others because the ability to listen–and to comprehend and understand the situation–is extremely critical to framing the problem properly.

Although it is typically not something an analytics practitioner can influence, the culture of the organization can have a significant effect on the ability to properly define the problem. In my previous organization, there were many cases where I worked very closely with the president. He would frequently call and ask, “I need this data for this time period” or “I need to see this and this,” and that's all the information he would consider. This is problematic because there may be parameters, circumstances, or other attributes that aren't stated that could significantly impact the output or the result. I would always have to push back on him a little bit to say, “OK, can we step back just a moment and can you give me a little bit more information about the problem you're trying to solve? What is it you're trying to accomplish? What's the overall objective?” Toward the end of my tenure there things got a little better, but I remember when I tried to initiate this type of conversation early on, it was usually met with something like, “it doesn't really matter,” “you don't need to know,” “it's not important right now,” or “I don't have time to go into it.” My effort was to try to communicate with him in order to better understand from his perspective what he was trying to accomplish. In situations like this, it's incumbent upon the analytics professional to convey that he or she is simply trying to provide the executive with the most appropriate solution for their problem.

The communication element is important in terms of being able to really listen and understand what the situation is; this includes the ability to empathize with the other person. From an analytic standpoint, this means being able to understand what the other person's overall situation is. For example, they may be under a lot of pressure from the president of the organization. Let's say that they're a VP or someone who reports to the senior executive team. Their sales may be significantly down, and they're trying to understand why so that they can either reorganize their product selection, or hire new salespeople, or whatever the case may be. That person may be thinking such things as “what could this mean in terms of my employment?” or “what impact would this decision have on the overall organization?” Being able to put yourself in another person's shoes really gives a lot of perspective into what the overall problem is and how it could potentially be addressed with an analytic solution.

Another important skill is the ability to think at the level of the person who is presenting the problem. It goes along with empathy, but it's really more concrete. In other words, if you are dealing with an executive, then the ability to think from the executive's perspective in terms of the business implications of the decision is important. It's not just a problem that you throw some data at and you build some models and that's it. It is important to be able to think at a higher level: to comprehend and understand the business as executives do. Certainly, it doesn't mean that every analytics professional needs to have an MBA in business strategy, but the more accomplished or the more adept the analytics professional is at thinking at that level, the more it opens up or exposes additional potential analytic solutions that may not necessarily have come to mind.

All else being equal, being able to communicate with empathy can make all the difference in how successful an analytics professional is in addressing business problems. Consider a situation in which you've got Analyst A, who is not able to think or converse at an executive level. They're mired in the statistical minutia or spend most of their day thinking in computer language rather than in the language of business. This person may be incredibly skilled at developing technical solutions, but has difficulty communicating with those in the business who are requesting their assistance. Contrast that with Analyst B, who is also very adept at building models and at programming whatever tool necessary to do the work that they need to do, but at the same time can switch perspectives so that they can converse with the business owner or executive at their level. Oftentimes, what I see are analytics professionals who can't bridge that gap, resulting in communication breakdowns at best, and a lack of trust at worst. When this occurs, the executive or businessperson asking the question may feel like the analyst lacks the understanding necessary to be able to deliver effectively. This is definitely not a recipe for analytics success.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

A diagram of a pilot testing

Description automatically generated

[**Figure 1.7**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF7) Life cycle of analytics projects.

1.4.2 Communicating Analytics

The best model in the world is of no value if the team is unable to persuade the decision-maker to act on the recommendation, so clear and transparent communication of recommendations and their rationale is essential. Writing good presentations takes effort. That effort is extremely important, even though it is completely irrelevant to the underlying mathematics. Analysts and executives frequently have very different perspectives and cognitive styles. Analysts are comfortable with mathematical formulae and inherently interested in computation, whereas executives are more focused on people, products, relationships, and results that impact business outcomes.

Junior analysts are prone to presentation pitfalls such as pasting a data table directly into a presentation (complete with six significant digits) and giving the slide a generic topic title like “Future Profit.” Executives look at the mass of numbers and wonder why the analyst is so naive as to believe they can actually distinguish between 10.5678 and 10.5679. Wondering if the analyst is equally naive about other, less obvious issues, the entire analysis is now suspect.

Unfortunately, even experienced analysts can get so caught up in the mathematically interesting details of their work that they neglect to take the time to properly frame their communication. A good presentation uses “sentence titles,” so that a reader who only reads the titles and does not look further into the slide can still follow the gist of the story. Good slides make their point clearly while also looking visually balanced and simple. This requires careful thought. Who is the audience? What is my goal for this meeting? What do I need to tell them to accomplish that goal? Business presentations are not mystery novels: they should lead with the answer and provide supporting details only in backup, for reference in the event they are needed. The analyst has to think about the important themes and illustrate them carefully. This usually means selecting a few key metrics and showing a relevant comparison, such as “benefit if you follow our recommendation versus benefit under the *status quo* plan.”

1.4.3 Organizational Capability

The sketch of the life cycle of an analytics project in [Figure 1.7](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0007) highlights some important issues. One is that the analytics experts who build the initial prototype solution tend to be scarce commodities. Whether the organization maintains its own internal pool of analytics talent or hires external consultants for each project, either way these people are expensive and difficult to recruit and retain. That is why it is essential to train a group of “superusers” who can support and maintain the project after the initial stage, so that the analytics specialists can be reassigned to new projects.

This scarcity leaves organizations with two key questions: how to prioritize analytic initiatives, and whether to use internal or external talent.

Prioritizing opportunities should be based on the impact to the organization as a whole. For businesses, this generally means improving the net present value of future free cash flows, or in simpler words, prioritizing the opportunities with the biggest potential bang for the buck.

There are two main ways this can occur: push and pull. In the “push” version, someone–either a central analytics organization or a central planning function–attempts to model the key drivers of business performance and the available levers for influencing those drivers. Applying a sensitivity analysis to this model results in a prioritized list of opportunities for intervention that have the highest potential to improve profitability. The leader of the organization must then “push” this agenda by socializing it with the leaders of the prioritized functions, who may or may not be receptive to the idea that some outsider thinks they can run the area more efficiently or more profitably. However, depending on the culture of the company, some of these leaders will be intrigued by the possibility of improvement, and will champion the initial projects. If those succeed, other leaders will generally become interested as well.

Over the past decade, many organizations have switched from push to pull, as analytics has become more visible in the C-suite. In the “pull” version, the central analytics organization prioritizes requests as they come in from leaders around the business. This version generally works much better than pushing, because the leaders themselves initiate the project and are pulling for it to happen. Someone still needs to set priorities, however, so it is still valuable to model key performance drivers and have a means for estimating the potential impact of each new project. Generally speaking, a project with a billion dollar potential impact requires only modestly more analytics resources than a project with a million dollar impact, so prioritizing based on the estimated size of the impact can be very helpful.

The prioritization decision is closely linked with the question of using internal or external talent. There are pros and cons to both approaches. External consultants can get up to speed quickly, draw upon a deep experience base within their firm, and already have a base of talented analytics professionals available. However, they are expensive. Moreover, “consulting makes the consultant smarter”–unfortunately, the client rarely gets as much of that benefit. Far too often, consultant-based projects turn out to be difficult to productionize without essentially paying the consulting company forever, because only the consultants really understand the analytics process at a deep level. Moreover, despite internal firewalls within consulting companies that keep specific details of competing clients strategies private, once a consulting firm develops a methodology for solving a particular business problem with one client, they are likely to want to leverage that investment by applying the more “generic” elements of that methodology with other clients. Initially, those new clients may indeed be in different industries, but over time the knowledge often diffuses more broadly, with the risk of eventually benefiting competitors of the original client.

As a result, companies that view analytics as a competitive advantage generally prefer to hire their own permanent analytics staff. This strategy too has downsides however, since it may be difficult to attract and retain sufficiently qualified people. Moreover, sometimes internal groups become insular, cut off from the advances in other industries, whereas consultants in a large firm may benefit from seeing many applications across a variety of industries.

If an organization does hire its own analytics staff, where should they fit in the organizational structure? Some companies centralize them under a Chief Analytics Officer, others spread them among a variety of client organizations, and some use a mix of both approaches. Sometimes analytics is viewed as part of the IT function, other times it is separate. Not surprisingly, it is difficult to make one-size-fits-all recommendations–the right answer depends on the size and shape and culture of the organization. For example, if the IT function's role and culture is primarily to manage infrastructure costs, they will probably not be a good fit for an analytics organization, which by nature is more like a small start-up or internal consulting company. In such cases, a centralized analytics group in conjunction with centers of expertise within client functions may be a good approach.

1.5 Ethical Implications

As analytics become increasingly pervasive, the ethical implications of collecting data and partially or fully automating decision-making become increasingly important. Analytics methods have the potential to provide tremendous value to individual companies and organizations, and to broader society. However, widespread collection of data raises privacy and security concerns. Additionally, broad adoption of algorithms to make decisions may have negative unintended consequences. Analytics professionals should be aware of these potential pitfalls and take actions to ensure that models are deployed in a responsible way.

In many countries, particularly in Europe, laws limit the kind of personal data companies are allowed to collect, store, and share, or provide consumers with the right to have their data erased. Even countries that allow collection of personal information often have laws mandating public notification if the data are inadvertently released or maliciously accessed by hackers. As a result, all organizations that analyze data must now stay informed about the potential legal implications of their data sets and take appropriate security measures to comply with applicable laws.

New technology for collecting data will raise new questions around “who owns data?” For example, who should have access to or be able to sell your Internet search history, your Fitbit health record, or your autonomous vehicle's sensor data? Similarly, as organizations lean more heavily on automated analytics, who will bear responsibility for errors, when for instance a driverless car is involved in a crash? There are many open questions that need to be resolved before the potential advantages of these technologies can be fully realized. In the coming decades, conversations regarding these topics will likely continue and will involve policy makers, lawyers, academics, politicians, and analytics professionals. Analytics professionals have a responsibility to honestly represent the capabilities and limitations of these technologies in these discussions, and to work toward solutions that serve the public good.

Algorithms have the potential to make decisions in ways that are more transparent and objective than a human decision-maker. For example, decades ago loan officers explicitly considered applicants' race when deciding whether to approve their loan applications. Modern credit scoring algorithms explicitly do not consider race as a factor. While not perfect, these algorithms are less discriminatory. However, the predictions or recommendations that come out of a model can be perceived as being completely objective, when in reality they are subject to biases in the data or in the modeling decisions. For example, data collected from smartphone apps are not representative of the whole population, as avid smartphone users skew young, affluent, and urban. Distribution of public services based on smartphone data may potentially exclude individuals who are invisible in the digital data set [14]. Similarly, racial biases in crime data can lead to racial biases in crime predictions, such as those used in predictive policing models [15].

Widespread deployment of certain algorithms can also create self-reinforcing feedback loops. For example, in most states in the United States, auto insurance premiums are substantially higher for people with poor credit [16]. These people then face much higher expenses, increasing the likelihood that their credit remains poor. Cycles like these are not a new phenomenon, but as price discrimination algorithms become more prevalent and more precise, the implications of the cycles become more profound.

Most countries have anti-discrimination laws that forbid discrimination based on factors like race, religion, gender, nationality, disability, and so on. Well-intended modelers who explicitly omit variables representing these categories may still inadvertently discriminate by including variables that correlate with these categories. For example, recidivism models are used to predict the likelihood that criminal defendants will reoffend. While these models do not explicitly use race, they use variables that correlate with race, such as education level and employment history, and thus defendants of different races may receive different risk scores [17].

The potential hazards of using analytics vary widely with the specific application. Sentencing decisions, for example, are fundamentally more morally fraught than decisions regarding which ad to serve on a Web site. Nonetheless, all analytics professionals should be aware of these issues, and should consider the societal consequences of their work. Diakopoulos and Friedler [18] proposed the following five principles that can guide accountability in the application of analytics:

1. *Responsibility:* Someone should have the authority and resources to deal with adverse consequences. Fully automated decision-making does not require a human in the loop, but a human should be involved to monitor the system and be able to make changes if needed.
2. *Explainability:* Decisions should be explainable to people affected by those decisions. Explaining the outcomes of machine learning models is especially difficult, but efforts are underway to develop interpretable machine learning methods, such as the DARPA Explainable Artificial Intelligence program [19]. In some applications, like speech recognition, explainability may be less important. But when used in contexts that have serious consequences for people's lives, such as determining who should receive a loan or be released from prison, clear and accessible explanations are essential.
3. *Accuracy:* Sources of error and uncertainty should be identified, logged, and benchmarked. Any model can make inaccurate predictions or misleading recommendations if it is given flawed data.
4. *Auditability:* Just as third parties are often used to identify security vulnerabilities, auditing could be used to identify potential ethical implications. The third party could exist within the same company, to protect proprietary information, but would have a different perspective from the original algorithm designer and could creatively search for potential unintended consequences.
5. *Fairness:* Biases can be “baked in” to existing data, and automated decisions can amplify structural discrimination. Analysts should be aware of this risk, and evaluate for potential discriminatory effects.

Recognizing the increasing risk of unintended consequences in the growing field of analytics, some organizations and professional societies in the area have taken the step of establishing explicit ethics guidelines to heighten awareness and stress the importance of responsible behavior (see [Figure 1.8](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#c01-fig-0008), for example).

­­A page of a document

Description automatically generated

[**Figure 1.8**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#_backF8) Guidelines on ethics from analytics professional society INFORMS [20].

1.6 The Changing World of Analytics

The analytics landscape has changed rapidly in recent years, and the pace of change continues to accelerate. Dramatic reductions in the cost to store and transmit data combined with the “Internet of Things” have resulted in much larger and more readily available data sets. Additionally, use of analytics is becoming more widespread, as many influential people and organizations publicize the benefits. Universities are responding to the shortage of trained analysts by developing undergraduate and graduate programs of various levels of rigor, and conferences related to “Big Data and Analytics” abound. At some point the hype will diminish, but because the results are real, analytics will not go away. Indeed, we expect organizations will continue to rely even more on analytics-based decision support in the future, as the benefits become increasingly well understood.

Increased volume of data has motivated the rise of parallel and distributed computing systems and the development of new algorithms for efficiently storing and retrieving data in these systems. Although particular vendors and platforms may rise or fall in popularity, the general theme is clear: problems that involve more data than comfortably fits on a single computer can be distributed over many computers in a way that makes answering certain common types of questions very efficient.

Certain kinds of data, such as real-time transaction data, or web browsing data, can be particularly massive, and the future storage requirements will likely grow astronomically. Distributed information systems that store this type of data are particularly well suited to descriptive analytics. It is straightforward to divide a giant database across multiple machines and let each one report back on the subset of data elements that match a given query. This can make report-generating systems run much faster.

Predictive and prescriptive analytics generally require more sophisticated mathematical models that are more difficult to fit into a distributed computing paradigm. This has led to the development of new algorithms for old methods that are better suited to distributed environments, as well as to entirely new methods. For example, “deep learning” methods are a form of machine learning that rely heavily on access to vast quantities of data. Traditional statistical techniques designed for small data situations rely on structure imposed by the analyst. Deep learning is attractive because (in theory) it allows the computer to find structure in the data without the human analyst having to first teach it a great deal. In practice, this depends on having a sufficiently large and rich data set available with a sufficiently high signal within the noise. Deep learning shows particular promise in situations such as voice and image recognition, where defining a structure is especially challenging, and where vast quantities of data are indeed readily available.

Traditional statistical methods are still the preferred choice in many other settings, where there is known structure and the amount of available data are more limited. For example, market research data are integral in many common business decisions, providing considerable value despite being “small” data.

We expect that both “big” and “small” data situations will remain important. Big data often follows the data-centric framing and bottom-up collection process, whereas small data generally start from the top-down decision-centric view. We predict that the most significant advances in applied analytics will come from combining the best of both worlds–leveraging the deep subject matter expertise required for small data applications to make the most of big data opportunities.

Some people are working on approaches to try to automate the analytics process further. At the moment, it is a very labor-intensive process requiring people with significant levels of education and experience. Will it be possible for computers to automate much of that? Perhaps someday, but at least for the foreseeable future, it seems to us that subject matter experts will continue to play a key role in many analytics projects. The real world is infinitely complex. Explaining the world to a computer is not easy. Cleaning data and interpreting results are complex cognitive tasks not easily replaced by current forms of “artificial intelligence.” Applying existing mathematical methods to a problem once the data are clean can be reasonably straightforward, but that is not the time-consuming, rate-limiting step in the analytics process, so automating it will not really solve the problem of scarce talent. Creating new mathematical methods suited to emerging new decision questions will long remain solely the province of human experts.

1.7 Conclusion

This chapter broadly defined analytics, using conceptual frameworks (data-centric and decision-centric) and high-level classifications (descriptive, predictive, and prescriptive). We introduced considerations for implementing analytics in organizations, and potential ethical implications. The following chapters will describe in more depth how analytics can be successfully implemented, including how to get started, data and organizational requirements, solution methodologies, and management considerations.

Analytics offers exciting and vast possibilities. The analytics landscape is rapidly evolving, and new methods, data sources, and computing resources create new opportunities. Businesses have opportunities to improve profit by growing revenue or reducing cost. Governments and nonprofits have opportunities to use resources more efficiently and deliver better services. For society more broadly, there are opportunities to improve health outcomes, reduce environmental impact, improve quality of life, and increase transparency and fairness. However, capturing these potential gains is not easy. Effectively implementing analytics requires the right data, the right tools, the right people, and the right systems.

Note

[1.](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch1/format/chapter-epub/OPS/c01.xhtml?hmac=1704948814-VvvobKXAqvZ6%2FfQzEoN6eg6J6U2Xc851Uc17j7W57pw%3D#backTNT1)The reader is encouraged to see <https://imgs.xkcd.com/comics/significant.png> for a lighthearted cartoon illustrating the dangers of false significance.

References

1. 1. Rose R (2016) Defining analytics: a conceptual framework. OR/MS Today, 34–38.
2. 2. Rosenbush S, Totty M (2013) How big data is changing the whole equation for business. The Wall Street Journal, March 10.
3. 3. Brynjolfsson E, McAfee A (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W. W. Norton & Company).
4. 4. McAfee A Brynjolfsson E (2017) *Machine, Platform, Crowd: Harnessing Our Digital Future* ( W. W. Norton & Company).
5. 5. Siebel T (2017) IOT: AI at enterprise scale. Comments made at MIT Sloan CIO symposium.
6. 6. Hartford T (2014) Big data: are we making a big mistake? *Financial Times*, March 29.
7. 7. Howard RA (1968) The foundations of decision analysis. *IEEE Trans. Syst. Sci. Cybern.* SCC-4 (3): 211–219.
8. 8. Madsbjerg C, Rasmussen MB (2014) The power of ‘thick’ data: businesses need to know how a product or service fits into the emotional lives of their customers. The Wall Street Journal, March 21.
9. 9. Rasmussen MB, Hansen AW (2015) Big data is only half the data marketers need. *Harvard Business Review*, digital article, November 16.
10. 10. Lustig I, Dietrich B, Johnson C, Dziekan C (2010) The analytics journey, *Analytics Magazine*, 11–18.
11. 11. Noyes K (2016) 5 things you need to know about data exhaust. *PC World*, digital article, May 13. Available at <http://www.pcworld.com/article/3069507/5-things-you-need-to-know-about-data-exhaust.html>.
12. 12. Tufte ER (1986) *The Visual Display of Quantitative Information* (Graphics Press, Cheshire, CT).
13. 13. Keenan P, Paich M (2004) Modeling General Motors and the North American Automobile Market. The 22nd International Conference of the System Dynamics Society, Oxford, England.
14. 14. Crawford K (2013) The hidden biases in big data. *Harvard Business Review Blog*, April 1.
15. 15. Eckhouse L (2017) Big data may be reinforcing racial bias in the criminal justice system. *Washington Post*, February 10.
16. 16.Consumer Reports (2015) The truth about car insurance, July 30. Available at <http://www.consumerreports.org/cro/car-insurance/auto-insurance-special-report/index.htm>.
17. 17. Barry-Jester AM, Casselman B, Goldstein D (2015) Should prison sentences be based on crimes that haven't been committed yet? *FiveThirtyEight*, August 4.
18. 18. Diakopoulos N, Friedler S (2016) How to hold algorithms accountable. *MIT Technology Review*, November 17.
19. 19. Gunning D Defense Advanced Research Projects Agency Explainable Artificial Intelligence (XAI). Available at <https://www.darpa.mil/program/explainable-artificial-intelligence> (accessed July 18, 2017).
20. 20.INFORMS Ethics Guidelines. Available at <https://www-informs-org.proxy3.library.mcgill.ca/content/download/357082/3750804/file/Ethics_Guidelines.pdf> (accessed July 19, 2017).

2  
Getting Started with Analytics

*Karl G. Kempf*

*Decision Engineering, Intel Corporation, Chandler, AZ, USA*

*“The secret of getting ahead is getting started. The secret of getting started is breaking your complex overwhelming task into smaller manageable tasks, and then starting on the first one.”*

Mark Twain [1]

2.1 Introduction

In 1965, Gordon Moore made a prediction that computing would dramatically increase in power and decrease in relative cost at an exponential pace over time [2]. True to this speculation, computing power measured in millions of instructions per second (MIPS) per dollar has grown by a factor of 10 every 4–5 years [3]. Advances in computing have driven or enabled similar results in memory, storage, and networking that have in turn enabled the World Wide Web and a sequence of revolutions in analytics.

Davenport identifies Analytics 1.0, 2.0, and 3.0 [4] and since advances in computing and associated technologies show no signs of slowing, we can expect Analytics 4.0, 5.0, and so on into the future. Prior to 2010, Analytics 1.0 was characterized by small, structured, internally generated data sets. Analytics were confined to reporting and what we would now think of as descriptive analytics. Results took weeks if not months to produce and so organizations could not think of analytics as a competitive advantage. The rise of Analytics 2.0 has dramatically changed each part of that characterization. Data are assembled from a variety of internal and external sources, in a variety of formats, sometimes including real-time streams. Analytics has slowly started to include predictive and prescriptive techniques. Speed in collecting and analyzing data has become paramount. Organizations think of and use analytics as a competitive weapon. Analytics 3.0 is continuing the trend to larger and more varied data sets as well as faster and more powerful analytics including machine learning. Analytics soon will support internal decisions by being embedded into operational processes and will enhance data-based products and services for customers. Predictive and prescriptive analytics are becoming more commonplace and indispensable to organization strategy.

In the construction of this chapter, we assume the reader has no analytics experience or some experience in Analytics 1.0, and in either case is interested in getting started with Analytics 2.0. The goal is utilization of a methodology for examining large data sets to expose as yet undiscovered patterns and correlations, market trends and customer preferences, and other information to help businesses make more informed decisions. Applications enable analytics professionals to analyze growing volumes of data that are often untapped by conventional business intelligence programs. This requires the ability to collect, integrate, manage, and leverage relevant data sources to help identify actionable improvement opportunities. It includes technologies for data manipulation and governance, application of analytics, and communication of results as well as organizational components. Properly executed analytics can point the way to multiple business benefits, including new revenue opportunities, more effective marketing, better customer service, higher impact research and development activities, faster product design-to-market cycles, improved operational efficiency in manufacturing, and faster and more reliable supply chains, all providing competitive advantages in the marketplace [5,6].

2.2 Five Manageable Tasks

In this chapter we introduce and explain the five manageable tasks required to succeed at the seemingly complex overwhelming task of getting started in analytics [7]. These include (i) choosing the business problem on which to focus, (ii) assembling the team, (iii) acquiring and preparing the data, (iv) selecting and applying the analytic tools, and (v) executing to produce an actionable result. Each task is treated in detail in later chapters. The tasks described here for getting started with analytics are ubiquitous. Whether the business problem selected is Descriptive (what events happened and when) or Diagnostic (why events happened in the past) or Predictive (what is likely to happen in the future), the same basic tasks must be completed. Whether the data used are structured, semistructured, or unstructured, internally available or acquired from outside, stored over time or streaming in real time, or a combination of all of these types and sources, the basic tasks must be executed faithfully. One of the tasks is building the team by selecting contributors from across the company and perhaps including personnel from consulting firms or universities. Another task involves selecting from the plethora of analytics tools commercially available today. Armed with a focus on the problem, a strong team, and appropriate data and tools, the last task is to supply an explainable and actionable result that can be communicated across the business as a successful example of applying analytics.

The importance of the five tasks claimed here as foundational and ubiquitous to all analytics is substantiated by a “Big Data” survey conducted by Capgemini Consulting [8]. The survey covered 226 respondents from across Europe, North America, and the Asia-Pacific (APAC) region and spanned multiple industries including retail, manufacturing, financial services, energy and utilities, and pharmaceuticals. When senior executives were asked to identify the major roadblocks to analytics, they pointed specifically to the five tasks described here:

1. *Task 1:* 39% answered “absence of a clear business case for funding and implementation.”
2. *Task 2:* 35% replied “ineffective coordination of big data and analytics teams across the organization” and 27% “lack of sponsorship from top management.”
3. *Task 3:* 46% indicated “scattered data lying in silos across various teams,” 27% “ineffective governance for big data and analytics,” and 15% “data security and privacy concerns.”
4. *Task 4:* 25% mentioned “lack of big data and analytic skills,” 22% “lack of clarity on big data and analytics tools and technologies,” and 18% “cost of tools and infrastructure for big data and analytics.”
5. *Task 5:* 12% pointed to “resistance to change within the organization.”

2.2.1 Task 1: Selecting the Target Problem (see also [Chapter 1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c01.xhtml))

The first task for successfully getting started with analytics is framing the target problem. Carefully determine what you are trying to achieve with your analysis before beginning the initial project. Start with a focus, not a mandate. Otherwise you will try to find out everything and risk finding out nothing. Identify a number of internal business or operational problems to solve or processes to improve. Focus on challenges and clarify key questions and concerns related to the goal. In the early stages of working on Task 1, you may pose a variety of questions expecting to get each considered but not necessarily resolved. This will lead in later stages to the description of the specific problem you need and want to address.

This description must include the value proposition. How will the outcome be measured–more efficient use of a scarce resource, a speed improvement of an important process, higher revenue in a historical market? Will there be a performance impact or an organizational transformation? In considering target problems, scoring and ranking by potential value is a useful exercise. Additional filtering can be accomplished by considering how your result will drive decisions and actions. This will help ensure that there is demand for the answer produced by your initial project. Selecting a good target problem and then supplying insight that does not drive action will not have an impact. Without impact, your project will (and should) be labeled as a failure.

Right sizing is another selection criteria. On the one hand, the problem should be sufficiently large to provide you an opportunity to produce a solution that the organization will recognize as a substantial contribution. On the other hand, the problem should be small enough so that the necessary data can be acquired, managed, and manipulated by a team getting started with analytics. A small but high-impact problem should be the target.

2.2.2 Task 2: Assemble the Team (see also [Chapter 3](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c03.xhtml))

Interview With Greta Roberts

*Greta Roberts, cofounder and CEO of Talent Analytics, shares her thoughts on the optimal analytics team size and the different challenges that teams of different sizes face?*

Optimal size to me is relative. If you have a small team and they have less work and they need to do everything in the analytics workflow, then a team of one could be optimal. What we've seen in the questions we ask our customers and in some of the research work we do is that when we ask people, “How big is your analytics team?” there are still a lot of people who do just have one person out there because they're probably not yet part of this global Analytics Center of Excellence inside of their company.

We also see three to five people working perhaps in retail, and three to five people working on the corporate analytics team, and maybe one person working in HR, and so on. So to my knowledge, you have these little disparate analytics teams. What we've seen in some of our research is that it still is kind of like that–one to five people, except for organizations that are starting to build an analytics Center of Excellence. Is that optimal? I don't know. I think that's what's happening today. I expect that it will start to grow, and maybe there will be more Centers of Excellence that might have a larger presence. I suspect companies will still have analytics professionals embedded in lines of business like marketing and sales and workforce and other lines of business. It is completely correlated to the amount of work needed to be done. There are some people who can do the entire cycle of analysis and programming and algorithms and deploying models and making presentations. There are some that are optimized doing a slice of that workflow.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

The goal of the second task for successfully getting started with analytics is assembling the team. This means establishing a core group that understands the importance of using data to drive business decisions. It involves finding folks with the right mix of skill sets who want to collaborate to develop a new capability for the company.

Collaboration is key. Too often some particular group initiates an effort while claiming that forming a cross-functional team will take too long. Unfortunately, it is seldom if ever the case that a single function in a company has all the skill sets needed for success. When their effort subsequently fails, interest in analytics is damaged. The attempt to speed up adoption in fact slows down the adoption. Getting started with analytics not only affords the opportunity to break down silos, but also demands the sharing of skills and information across departments. As you begin the analytics effort, proceed with collaboration between executives, business experts, operations personnel, information technology specialists, data scientists, engineering, and whomever else can help. Resolving issues and uncovering insights from your data for your target problem requires a variety of close working relationships. There are a number of roles listed, including (but not limited to) the following in no particular order.

Executive Sponsor [9]

This team member can set the vision and tone for the work while ensuring the target problem aligns with organizational goals. The executive sponsor can remove roadblocks and address funding issues as they arise. Serving as the communication channel to other executives, the sponsor ultimately broadcasts the success of the project.

Project Manager

The project needs a committed leader to manage the schedule and the budget while delivering the results. As a “getting started” project in a new endeavor including multi-group collaboration, an experienced project manager will be required to oversee moving the team through uncharted territory. Bonus points if this manager is known and trusted by the executive sponsor.

Domain Expert

A domain expert from the business can help the team better understand the target problem, measure the results of the endeavor, and learn to speak the sponsor's language. It should be the case that this team member was heavily involved in selecting the target problem as well as forecasting the action(s) the solution is intended to drive.

IT Expert

This team member has the necessary systems knowledge to help gather and organize data and information. She/he might know where the appropriate data silos are located and who controls them. Ideally her/his skills (or those of associates of the IT expert) should stretch to such important topics as data quality, security, and governance.

Data Scientist [10,11]

The data scientist understands modeling and algorithms as well as how to explore data sets for new insights. Especially important is the practical knowledge of which of the myriad of tools are available to select from the analytics toolbox for application to the specific target problem and associated data. Of the roles to fill, this might be the most difficult. If that is the case, consultants, advanced degree interns, or other experts from outside the organization can accelerate delivery, reduce risk, and expedite learning.

Stakeholders

You have to be sure to identify all of the affected stakeholders. If you fail to do so, the ones that you neglect are likely to raise objections as the project proceeds. It may be that the domain expert is one of the business stakeholders. Alternatively, the stakeholders may be in operations or engineering or sales and marketing. It may not be necessary to include stakeholders as team members working on the project on a daily basis, but it is certainly necessary to establish consensus on the goals and keep stakeholders updated on status as the project progresses.

To get started with analytics, the team does not necessarily have to be large, but it does need to be creative and fast. There may be many unique obstacles–technical and political–to overcome along the way. There will certainly be many interested parties who will need to be pulled along, so crisp clear reporting will be necessary–especially to help the executive sponsor spread the word once success is achieved. Note that although a successful initial project will have a positive impact on the business, the main benefit may be the team learning captured and the organizational confidence gained.

2.2.3 Task 3: Prepare the Data (see also [Chapter 4](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c04.xhtml))

The data task involves a number of stages and usually these stages proceed in parallel with input from various sources. The initial challenge is to understand the characteristics of the data needed to address the target problem. In the best of cases for getting started in analytics, a search of internal data sources yields most if not all of the data needed. A successful search is followed by acquiring access to data in batch mode or as a real-time stream. In the worst of cases, a major investment in time, personnel, and budget is required to find and access data before it is clear whether the data can be used to successfully solve the target problem. This heavy lift increases the chance of failure and the team might consider rerunning Task 1. Access is followed by validation and cleansing or correction [12] sometimes including transformation (e.g., standardization of units or formats) in preparation for analysis.

Most discussions of data in the context of analytics include the five “V's”: Value, Veracity, Volume, Velocity, and Variety. Value and Veracity are critical to data for finding a good solution to the selected problem. Volume, Velocity, and Variety are crucial in picking the data management approach used. Prior to Analytics 1.0, as well as for many projects during 1.0, databases based on relational algebra were able to manage a few thousand documents or a few thousand transactions per second. For Analytics 2.0 and beyond with higher Volume, higher Velocity, and higher Variety, relational databases are inadequate. Hence, the advent of alternative approaches such as Key-Value Pair databases, Wide-Column stores, Graph databases, and many more. Relational databases work with structured data and scale vertically but not horizontally. Alternative approaches work with unstructured and semistructured data and scale out very well horizontally. Relational databases are over-kill, damaging scalability for data that can be effectively used as key-value pairs, and are under-kill, decreasing performance for data that need more context than just relations like graph structures.

Value means finding the right data to address the target problem. This may not be obvious in the early days of the effort even if the team members have some idea of types of data needed. Obviously, the more potentially relevant data that can be acquired the better since data that are ultimately found to be irrelevant can be ignored. From the opposite perspective, if some relevant data are missing, then some potential insights could be missed in the analysis.

Veracity draws attention to the question of quality. Inaccuracies in the data can quickly compromise all of the effort invested by the team. Insight from the analysis will be lacking and the resulting decisions may be poor. The team must realize that data quality is more important than data quantity. Focus on gathering as much data as possible without considering whether it is accurate will not yield the desired result. On balance, there is the very real question of acceptable accuracy. The team must also guard against setting accuracy standards so high that they are neither relevant nor cost-effective in a business context.

The question of Volume raises an important trade-off. On the one hand, relative to the potential value to the company of the specific project and the general learning to do successful analytics in the future, storage is cheap as is compute time for analysis. Hence, the more data the better. On the other hand, the team resources required to discover, access, cleanse, and transform the data are nontrivial. The objective of the “getting started” project is to solve the target problem, not collect all data that might ever be useful to the company.

How fast data are generated and the speed with which they need to be acquired and used–Velocity–are relative to the speed of business. Companies must be in a position to respond to what is going on around them. The time companies have to make their decisions is decreasing. This means that data have a shelf life and their usefulness decrease with the passage of time. The team must be sensitive to the velocity with which the data must be moved from acquisition to action to satisfy the focus problem [13].

Variety means that useful data will come from many sources, some supplying structured data, some unstructured, and some semistructured. Structured data are comprised of clearly defined data types whose pattern makes it easy to manipulate. Structured data often reside in relational databases in IT data centers. In addition, most companies have a large number of spreadsheets scattered throughout the organization with well-defined data organized in rows and columns. Examples of unstructured data include text, audio and video files, and a majority of social media data. The data can be human or machine generated, but in any case are not structured via predefined data models or schema that make them easy to manipulate. E-mail is an example of semistructured since it is structured in sender, recipient, date, and time, but is unstructured in the contents of the message. Some (or all) of the data required may be acquired from outside the company either publicly accessible or available for purchase. Not all meaningful data will be collected electronically. Thus, the data may be in all different formats and structures, and may be of variable completeness and quality.

In addition to these technical problems, the team may encounter political problems in collecting data. In some circles, data are power and the owners of the data may not be anxious to share. Under some circumstances, even within the same company, Group X might not want Group Y to have access to its data and so might be reluctant to share. Generalizing these specific problems leads to the topic of data governance. Who is responsible for controlling access to the data? Who is responsible for maintaining the data? These and related questions require the team to develop and communicate a set of rules to deal with privacy and security from day 1 of the project [14,15].

Addressing these issues is the job of the IT expert with support from the other team members, especially the data scientist. A step-by-step data process must be strictly followed as part of governance:

* Identify and document each internal and external data source, including location, contents, owner, quality, format, and origin if possible. Negotiate access and conditions of use with special attention to confidentiality.
* Maintain a detailed record of all data captured, including what, when, and from whom. Archive that data in their original format for future reference especially if modifications are required for integration into the analysis set.
* Develop a consistent format suitable for use by analytics. If a change is needed to incorporate new data or to improve the analytics, change the existing data as well and carefully document all format changes. If necessary, it is better to add a new field rather than change the name or meaning of an existing field.
* Granularity is important. Granularity beyond that required to address the target question is more difficult to acquire and manage, but it can be aggregated to suit various needs and purposes. Data of less granularity may be easier to obtain, but will not be as useful since it can't be disaggregated. Granular (disaggregated) data can be aggregated, but aggregated data cannot be disaggregated.

During the execution of these steps, the overriding concern should be quality. New data need to be as complete as possible as well as correct and consistent. It should be time stamped for temporal alignment with existing data. Bad data such as outliers and missing values should be detected and either repaired or eliminated as soon as possible (note that these may be informative anomalies!). A substantial part of data quality control can be accomplished with a new data set before it is introduced into the existing data set. Quality control continues as new data are checked as much as possible for consistency and alignment with the existing data. Documentation of issues detected and remedies employed is a good practice.

2.2.4 Task 4: Selecting Analytics Tools (see also [Chapters 5](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c05.xhtml) and [6](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c06.xhtml))

Any attempt at listing, describing, or comparing available open-source or commercial analytic tools faces at least two challenges. The first challenge is that this list would be very long indeed. Even a basic online search will identify scores of possibilities. The second challenge is that this list would very quickly become obsolete. Research advances the state of the art, new tools are introduced, current tools are updated, and old tools are discontinued in short order. These challenges are not addressed here but can be managed by regularly consulting the appropriate periodicals [16,17]. In this section we can only describe the characteristics of tools that should be taken into consideration when making a selection [18].

The technical factors for selecting the analytic tools to be used are (a) the target problem selected and the type of solution desired, (b) the data available with which to solve the problem, (c) the computational infrastructure available to the team, and (d) visualization requirements.

Analytical Specificity or Breadth

Examples of analyses requiring very specific tools include video analytics for extracting information from video footage and voice or speech analytics for examining audio recordings. Sentiment analysis (also known as opinion mining) might require combined video and audio understanding. Image analytics for working with photographs or medical images and text analytics for use with large quantities of unstructured text data are other examples that require tools tailored to the specific task. For a “getting started” project, it is more likely that the data being examined is numeric and the analytics is some form of statistics, simulation, or optimization. In this case, breadth of techniques is important. For example, a single software package supporting clustering, segmentation, decision trees, time series, classification, and regression is more useful than a tool with only a single technique since many approaches will need to be considered when identifying which is most appropriate. Of course, the broader the functionality, the higher the price and the more the sophistication presumed of the user.

Access to Data

The factors for selecting among the various tools must be based on your team's specific requirements for accessing and processing data volumes and varieties. Whether your data are in a columnar or in-memory or nonrelational database on a private or public cloud, the analysis tool must be able to efficiently access the data. The IT expert and the data scientist need to collaborate in selecting storage technology and analysis technology that are compatible.

Execution Performance

The need for performance is determined by size and complexity of the data set and the velocity with which the analysis needs to run to drive the actions desired. Overall execution performance is determined by the analysis technique employed as well as the computational, storage and networking infrastructure available. Assuming your initial project is a great success, assessing how the tool scales to larger problems and more capable infrastructure will save money and resources in the future.

Visualization Capability

Clear communication of the results of the initial project will be crucial to success. The visualization capabilities of the analysis tool or some other specific visualization adjunct tool are therefore of great interest to the team. Equally important is the utilization of the extraordinary pattern recognition skills of human beings. Visualization of the raw data and the intermediate results of analysis for inspection by team members and stakeholders is often important–and sometimes crucial–to solving the target problem.

Nontechnical factors are also involved in the tool selection process. These include (a) the skill set of the data scientist(s) on the team, (b) the pricing from the vendor, (c) the budget available to the team, and (d) collaboration requirements.

Data Scientist Skillset

Available analysis tools range from those that target novice users to those engineered for expert users. One perspective is that the selected tool should match the skill level of the data scientist(s) on the team. Unfortunately, if the analytics the tool supplies are not appropriate to solve the target problem, the team will fail to deliver the desired solution. Another perspective (the correct one) is that the selected tool should match the needs of the project. If that exceeds the skill level of the data scientist(s), some skill enhancement is warranted. While this might delay the project, effective and relatively inexpensive training is available from multiple sources, including classes at conferences, universities, and tool vendors. Alternatively, the team may add a scientist who holds the required skill set.

Vendor Pricing

There are almost as many pricing models for analytics tools as there are vendors of analytics tools. At one extreme is a free or open-source version of a tool with charges for support. At the other extreme are large vendors who offer a massive portfolio of analytics tools and ideally (for them) are interested in site-wide or enterprise-wide licenses for the whole portfolio. Somewhere between these extremes are small vendors who license one or a few tools. Not surprisingly, each of these arrangements can include consulting (at an additional cost). Another common pricing determinant is number of simultaneous users, or number of processor cores in the infrastructure, or size of memory serviced. The latter two relate to the infrastructure available.

Team Budget

There is a range of strategies here. If budget is a constraint, then the least expensive path to providing a high-quality solution to the test problem may be the only feasible option. Even if budget is not a constraint, this may be the wise choice. But if budget is available and the team is confident of success, speculation concerning future projects can enter into the purchasing decision. In addition, perceived risk can be a consideration. Some teams feel safer dealing with a large vendor that offers well-tested software of proven reliability and has an extensive user community. Others prefer a small vendor that offers the potential of a closer working relationship.

Sharing and Collaboration

Although perhaps not critical for a small team working on a “getting started” project, a criterion looking to the future of larger teams is tool capability for sharing. Imagine a large team spread over multiple sites and multiple time zones (perhaps including multiple continents). Advantage could be gained in terms of velocity and brain power by sharing models and collaborating regarding interpretation of the results across multiple data scientists and potentially including business users. Some scientists have actually found that software and programming languages provide a common lingua franca that helps bridge language gaps; they can often communicate more succinctly and clearly through the code they write and the models they build than they can through their spoken and written languages.

Addressing each of these issues is the job of the data scientist supported by the other team members, especially the IT expert. During the decision process, it should be kept in mind that vendors often supply “evaluation” versions of their offerings for potential customers to test. Access is strictly time limited, and capacity is frequently limited; sometimes these vendors provide “typical” data sets for demonstration to help speed a purchasing decision.

2.2.5 Task 5: Execute (see also [Chapter 7](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch2/format/chapter-epub/OPS/c07.xhtml))

With the focus problem defined and at least some initial data and an analytic tool in hand, the team can start iterating. Although “getting started” in analytics should be equivalent to building the minimum viable system to provide actionable results, if the first round of data collection and application of analytics supplies a solution, there are only two possible conclusions: either the team is incredibly lucky or the focus problem is far too easy. A more realistic execution trajectory involves iterating around data, analytics, and learning. The initial data set will likely be inadequate and the learning will point to gathering more data, restructuring the existing data, eliminating some data, and improving the data in some way. Probably the initial analytic approach will be lacking in some way and the learning will lead to modifying the approach or trying a different technique or combining techniques–refining the analytics in some way to move closer to solution. This is the art of the team. This is the moving through uncharted territory that was mentioned earlier. It requires persistence, creativity, and confidence. This is where the learning how to succeed at analytics takes place. The more quickly the iterations can take place the better.

This process can be aided by establishing a schedule for reporting to interested stakeholders. Showing visualizations of the data as well as the intermediate analytics results to experts in the domain could produce insights and suggestions that will be of great value to the team. In any case, this will serve to keep the stakeholders engaged. The team must remember, however, that getting started in analytics is, with some stakeholders, an “old dogs, new tricks” exercise and there will be skeptics and naysayers. Useful guidance can come from these folks too, if the team has the patience to listen to and fully comprehend what these stakeholders are missing or not understanding. The stakeholder with a negative perspective may be correct (even insightful) and have especially useful feedback! In any case, these stakeholders have self-identified as folks to invite to the celebration once the focus problem has been cracked.

Once the project is successful and an actionable result is generated, make sure it is applied and implemented correctly. As mentioned repeatedly, validating the exercise with a real measurable business result is the real goal. Once that result is demonstrated in practice, project success needs to be socialized. The executive sponsor should drive this top-down, while the domain expert can broadcast bottom-up. This will convince the executive team that the organization can succeed and benefit with analytics, and educate employees on the value of creating a data culture. Documentation of lessons learned makes the success easier to duplicate by other teams with other team members. In any case, do not rest on your laurels! Next problem please!

Interview With Harrison Schramm

*Harrison Schramm, who recently retired after a 20-year career as a helicopter pilot and operations research analyst in the U.S. Navy, suggests ways an analytics professional can overcome the challenges of data politics to educate clients and get stakeholder buy-in.*

You have to present yourself as a human being. Before another human being will trust you, you have to present yourself as someone with whom they have at least some (maybe even superficial) commonality. Finding a way to have a relationship on which the work you are doing together can build is key. Let me tell you a parallel story about that (I'm going back to storytelling). I recently had a great deal of dental work done. I thought my dentist was fantastic. Why did I think my dentist was fantastic? I did not go to another dentist and have that dentist check to see if the drilling and filling and whatever my dentist did was right, but she sure seemed like she knew what she was doing. She had a great bedside manner, and she was someone I could relate to. It is interesting to consider that the criteria on which I judged the quality of my dentist had absolutely nothing to do with the quality of the way she repaired my teeth–I actually don't know how well she did it! Of course, the quality of the work you can do is very important, but you might not get an opportunity to demonstrate your capabilities if your potential client isn't comfortable with you.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

2.3 Real Examples

Unfortunately, few groups publish information about the projects they successfully executed, especially while getting started with analytics. Small but high-impact projects for a particular company are not often deemed worthy of broadcast. Reports do sometimes come out in trade magazines [19–21] that are then collected and used as data to support broader abstractions and theories [22]. A few are briefly described here to provide examples of successful small but impactful projects using data from various sources and analytics of many types all delivering beneficial actionable results.

Case 1: Sensor Data and High-Velocity Analytics to Save Operating Costs [23]

U.S. Xpress is the third largest privately owned trucking company in the United States. Fortunately, it had already partially addressed Task 3 before the BDA case described here. It had embarked on a general overhaul of its information management system since business questions took at least weeks and sometimes months to answer. The task was selected out of exasperation during an industry downturn. The topic was how to cut costs, and the request from one executive was to focus on truck idling times as a potential key to saving on fuel. Since every truck had some sort of communications device, the IT team was able to respond in less than 6 weeks with an application to supply data on how much time trucks were stationary with their engines running and using up fuel but not going anywhere. Simple analysis across data from 8000 tractors and 22,000 trailers showed that implementing policy changes for the drivers could indeed result in substantial results. This insight gained through the collection, management, and analysis of operations data saved U.S. Xpress roughly $6 million per year.

Case 2: Social Media and High-Velocity Analytics for Quick Response to Customers [24,25]

At the time of this example activity, Starbucks had already become proficient at mining social media like Facebook and Twitter as well as niche coffee forum discussion groups. An obvious task was to assess the reaction of customers to newly introduced products and the sooner the better. (Here is a paradigmatic example of the value of acquiring, analyzing, and acting on data in near-real time as opposed to stockpiling data over time for analysis and response later.) The specific question was whether customers would think a particular brew tasted too strong. Real-time efforts began as soon as the first short/tall/grande/venti/trenta was poured. By mid-morning, it was clear from social media that the taste was agreeable but the price was not. It was perceived to be too high. The price was reduced across the Starbucks network by early afternoon, and at the end of the first day there were no further negative comments. Note the punch line: not end of quarter or end of month, not even end of week, but rather *end of the first day*.

Case 3: Sensor Data and High-Velocity Analytics to Save Maintenance Costs [26]

Petróleos Mexicanos (also known as Pemex) is a producer, refiner, and distributor of crude oil, natural gas, and petroleum products. It is one of the largest petroleum companies in the world, and it relies on analytics to solve basic but important operational question. Oil refineries use water to heat fluids and cool equipment during the refining process. A major component of the water system is the cooling tower, and a typical refinery has many of them. Each of these towers has a number of large cooling fans that regulate the temperature of the water contained in the tower. Due to mechanical wear, axis misalignment, oil leaks, and other problems, the fan motors and gear boxes sometimes begin to vibrate. This shortens their productive life and risks unexpected shutdowns that cost time and money and have a negative impact on refinery operations. Maintenance crews sometime waste effort by addressing a fan that is not yet vibrating, and in other instances do not arrive at a fan until it is too late to address the issue. Pemex has found a way to avoid the vast majority of fan-related problems by mounting wireless vibration sensors and collecting and analyzing high volumes of data in real time, thereby reducing parts and labor expenses as well as dramatically reducing shutdown risk.

Case 4: Using Old Data and Analytics to Detect New Fraudulent Claims [27]

Infinity Property & Casualty Corporation, headquartered in Birmingham, Alabama, is a national provider of nonstandard car insurance for individuals who are unable to secure coverage through standard insurance companies due to a driving record with accidents, tickets, prior DUI, or vehicle type. Considering its business model, it is not surprising that fraud management is a particularly important part of Infinity's operations. Thinking through this task, Infinity speculated that automobile insurance claims could be scored in the same way as consumer credit applications. Furthermore, Infinity realized that it had archived years of adjusters' reports that could be analyzed and correlated to instances of fraud. It built an analytics-based system around that data to assign fraud probability “scores” when initial accident reports are filed. Based on the score, suspicious claims are sent to fraud investigators within a day or two for deeper analysis. This has resulted in a roughly 50× decrease in time taken to identify attempted fraudulent claims and increased their success rate in catching fraudulent claims, which has resulted in $12 M in recoveries.

Case 5: Using Old and New Data Plus Analytics to Decrease Crime [28]

PREDPOL Inc. is located in Santa Cruz, California, and was incorporated in 2012. Its business is predictive policing: using data and predictive analytics to supply law enforcement agencies with predictions for the places and times where and when crimes are most likely to occur. Depending on the granularity of the input data, PREDPOL can provide predictions with a resolution of 500 ft × 500 ft segments on a map of the patrol area. It uses no personal information, eliminating any concerns about personal liberties or profiling. The predictions are based on the observation that certain crime types tend to cluster in time and space and are based solely on crime type, crime location, and crime date/time. Historical data are obtained from the target police department's records to build the initial model. Fresh data are collected and used daily to create updated predictions for each patrol area and shift. In areas of Los Angeles where the predictions have been used to focus policing efforts, there has been a 20+% reduction in violent crimes and a 30+% decrease in burglaries.

Case 6: Collecting the Data and Applying the Analytics Is the Business [29]

Chicago start-up Food Genius (FG) is a foodservice data provider. They scrape open content from the Internet, specifically menus with prices posted online by independent and chain restaurants around the country. When possible, they break down menu items into ingredients. With this data and analysis, FG can tell, for example, whether restaurants are still luring their burger customers with added bacon or whether customers' tastes have switched over to avocado. Furthermore, FG can determine the asking price delta of such a switch. This gives FG the ability to determine what combinations of ingredients, flavors, and buzzwords are being offered in attempts to make dishes more appealing and perhaps worth an increased price to diners. Customers of the analysis supplied by FG fall into two categories. One is the restaurants: independent restaurants that want to understand what the local competition is doing, and chains that want to see regional trends across the country. The other is the companies that supply raw ingredients to the restaurants that are being analyzed.

References

1. 1 Mark Twain. Available at <https://www.brainyquote.com/quotes/quotes/m/marktwain118964.html> (accessed May 20, 2017).
2. 2 50 years of Moore's law. Available at <http://www.intel.com/content/www/us/en/silicon-innovations/moores-law-technology.html> (accessed May 27, 2017).
3. 3 Trends in the cost of computing (2015) Available at <http://aiimpacts.org/trends-in-the-cost-of-computing/> (accessed April 30, 2017).
4. 4 Davenport TH (2013) Analytics 3.0. *Harv. Bus. Rev.* 91 (12): 64–72.
5. 5 Davenport TH (2006) Competing on analytics. *Harv. Bus. Rev.* 84 (1): 98–107.
6. 6 Barton D, Court D (2012) Making advanced analytics work for you: a practical guide to capitalizing on big data. *Harv. Bus. Rev.* 90 (10): 78–83.
7. 7 Davenport TH (2015) 5 Essential principles for understanding analytics. *Harv. Bus. Rev.*, October 21, 2015.
8. 8 Cracking the data conundrum: how successful companies make big data operational. Available at <https://www.capgemini-consulting.com/cracking-the-data-conundrum> (accessed April 28, 2017).
9. 9 McShea C, Oakley D, Mazzei C (2016) The reason so many analytics efforts fall short. *Harv. Bus. Rev.*, August 29, 2016.
10. 10 Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. *Harv. Bus. Rev.* 90 (10): 70–76.
11. 11 Data Scientist Report. Available at <https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html> (accessed May 29, 2017).
12. 12 Xiong H, Pandey G, Steinbach M, Kumar V (2006) Enhancing data analysis with noise removal. *IEEE Trans. Knowl. Data Eng.* 18 (3): 304–319.
13. 13 Davenport TH, Barth P, Bean R (2012) How big data is different. *MIT Sloan Manag. Rev.* 54 (1): 43.
14. 14 Richards NM, King JH (2013) Three paradoxes of big data. *Stanf. Law Rev. Online* 66, 41–46.
15. 15 Richards NM, King JH (2014) Big data ethics. *Wake Forest Law Rev*. 49, 393–432.
16. 16 Analytics Magazine Available at <http://analytics-magazine.org/> (accessed May 14, 2017).
17. 17 OR/MS Today. Available at <https://www-informs-org.proxy3.library.mcgill.ca/ORMS-Today> (accessed May 14, 2017).
18. 18 Redlon M (2015) A guide to selecting an analytics vendor. *Harv. Bus. Rev.* October 23, 2015.
19. 19 Laskowski N (2013) Ten big data case studies in a nutshell. Available at <http://searchcio.techtarget.com.proxy3.library.mcgill.ca/opinion/Ten-big-data-case-studies-in-a-nutshell> (accessed May 10, 2017).
20. 20 Laskowski N (2015) Ten analytics success stories in a nutshell. Available at <http://searchcio.techtarget.com.proxy3.library.mcgill.ca/opinion/Ten-analytics-success-stories-in-a-nutshell> (accessed May 10, 2017).
21. 21 Petersen R (2016) 37 big data case studies with big results. Available at <https://www.businessesgrow.com/2016/12/06/big-data-case-studies/> (accessed May 10, 2017).
22. 22 He W, Wang F, Akula V (2017) Managing extracted knowledge from big social media data for business decision making. *J. Knowl. Manag.* 21 (2): 275–294.
23. 23 Lemos R (2011) Big data: how a trucking firm drove out big errors. Available at <http://www.cio.com/article/2410714/data-center/big-data--how-a-trucking-firm-drove-out-big-errors.html> (accessed April 26, 2017).
24. 24 Gallaugher J, Ransbotham S (2010) Social media and customer dialog management at Starbucks. *MIS Q. Exec.* 9 (4): 197–212.
25. 25 Chua A, Banerjee S (2013) Customer knowledge management via social media: the case of Starbucks. *J. Knowl. Manag.* 17 (2): 237–249.
26. 26 Espinosa R, Spinoso R (2012) Wireless monitoring asset protection. Available at <http://www.controleng.com/channels/process-control/process-control-news/single-article/wireless-monitoring-asset-protection/b60ecfe04dcf887c656d14522d4fe728.html> (accessed April 28, 2017).
27. 27 Dibble W (2011) Infinity property and casualty builds a smarter system for Fraud. InformationWeek Insurance & Technology. Available at <http://www.insurancetech.com/infinity-property-and-casualty-builds-a-smarter-system-for-fraud/a/d-id/1313275> (accessed May 17, 2017).
28. 28 PREDPOL: The Predictive Policing Company (2017) How PredPol works: we provide guidance on where and when to patrol. Available at <http://www.predpol.com/> (accessed May 24, 2017).
29. 29 Rekdal A (2016) Table tech: these 22 Chicago companies are changing the food industry. Available at <http://www.builtinchicago.org/2016/10/07/chicago-food-tech-companies> (accessed May 14, 2017).

Further Reading: Papers

1. Ahlawat and Rambola, 2016 Ahlawat T, Rambola R (2016) Literature review on big data. *Int. J. Adv. Eng. Technol. Manag. Appl. Sci.* 3 (5): 21–30.
2. Chen et al., 2016 Chen Y, Chen H, Gorkhali A, Lu Y, Ma Y, Li L (2016) Big data analytics and big data science: a survey. *J. Manag. Anal.* 3 (1): 1–42.
3. Chong and Shi, 2015 Chong D, Shi H (2015) Big data analytics: a literature review. *J. Manag. Anal.* 2 (3): 175–201.
4. Elgendy and Elragal, 2014 Elgendy N, Elragal A (2014) Big data analytics: a literature review paper, Industrial Conference on Data Mining. Springer International Publishing, pp. 214–227.
5. Gandomi and Haider, 2015 Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inform. Manag.* 35 (2): 137–144.
6. Power, 2014 Power D (2014) Using ‘big data’ for analytics and decision support. *J. Decis. Syst.* 23 (2): 222–228.
7. Soon et al., 2016 Soon K, Lee C, Boursier P (2016) A chronology of big data adoption. *Int. J. Appl. Bus. Eco. Res.* 14 (1): 521–544.
8. Tsai et al., 2015 Tsai C, Lai C, Chao H, Vasilakos A (2015) Big data analytics: a survey. *J. Big Data* 2 (1): 21–53.

Further Reading: Books

1. Ankam, 2016 Ankam V (2016) *Big Data Analytics* ( Packt Publishing, Birmingham, UK). ISBN 978–1785884696 (Hadoop).
2. Bunnik et al., 2016Bunnik A, Cawley A, Mulqueen M, Zwitter A, eds. (2016) *Big Data Challenges: Society, Security, Innovation and Ethics* ( Palgrave Macmillan, London, UK). ISBN: 978–1349948840 (Security).
3. Davis and Patterson, 2012 Davis K, Patterson D (2012) *Ethics of Big Data: Balancing Risk and Innovation* ( O'Reilly Media, Sebastopol, CA). ISBN-13: 978–1449311797 (Ethics).
4. Foreman, 2013 Foreman J (2013) *Data Smart: Using Data Science to Transform Information into Insight* ( John Wiley & Sons, Inc., Indianapolis, IN,). ISBN 978–1118661468 (Techniques).
5. Maheshwari, 2014 Maheshwari A (2014) *Data Analytics Made Accessible* ( Amazon Digital Services) (Case Studies).
6. Provost and Fawcett, 2013 Provost F, Fawcett T (2013) *Data Science for Business* ( O'Reilly Media, Sebastopol, CA). ISBN 978–1449361327 (Concepts).

3  
The Analytics Team

*Thomas H. Davenport*

*Technology, Operations, and Information Management, Babson College, Wellesley, MA, USA*

3.1 Introduction

Analytics are created or initiated by people. People frame the question to be answered by analytics, select the data to be analyzed, propose and test hypotheses, and determine how well the hypotheses are supported in the data. Even in relatively automated machine learning environments, analysts or data scientists select the data and the tools, and kick off the process of finding a model that fits the data. The capabilities of human analysts are among the most important factors in determining the success of an analytics initiative.

In organizations of any size, it is impossible for one analyst to do all the necessary analytics work. Therefore, the topic of human analytical resources quickly becomes one of assembling and managing an analytical team. In addition, there may often be too many different skills required for high-quality analytical work for one person to possess them all. It is usually the case that some sort of division of labor and skills across a team is necessary.

This chapter, then, will focus on assembling and managing teams of analytical people to analyze data and assist the organization in making analytical and data-driven decisions. It addresses not only an organization's requirements for analytical capabilities but also the individual skills required to make analytics successful. It will also address some of the ways in which analytical teams can be organized within a company.

Although this chapter appears in a book published by INFORMS, it is not a commercial for that organization. Nevertheless, there should be little doubt that certification of analytical skills is a useful exercise to ensure that the necessary skills are present in an individual's repertoire. INFORMS has created one of the more effective certification programs in its CAP, or Certified Analytics Professional. The CAP certification requires work experience in analytics, but the Associate CAP (aCAP) does not. I won't discuss these further–there are many materials available on the program's website (<https://www.certifiedanalytics.org/>)–but as a member of the Analytics Certification Board, I will testify to its quality and urge individuals and organizations to pursue this certification.

3.2 Skills Necessary for Analytics

The skills necessary to work with analytics have evolved considerably over the several decades that companies have been pursuing business analytics. I'll describe the evolution in this chapter, beginning with the basic skills that have been necessary since the 1960s or 1970s, when the use of analytics in businesses began to take off.

Quantitative skills–broadly speaking, the ability to extract, meaning from numbers–are the core requirement for any type of quantitative analyst. But tuning a regression equation or manipulating a spreadsheet is only the beginning. Effective analysts need to be proficient not only with data but also with people.

* *Quantitative and technical skills* are the foundation. All analytical people must be proficient in both general statistical analysis and the quantitative disciplines specific to their industry or business function: lift analysis in marketing, stochastic volatility analysis in finance, biometrics in pharmaceutical, and informatics in health care firms, for example. Some types of analysts–those involved in “business intelligence” or reporting work–may get by without substantial statistical knowledge, but this lack would probably limit their careers today. Analytical people must also know how to use the specific software associated with their type of analytical work, whether it be to build statistical models, generate visual analytics, define decision-making rules, conduct “what-if” analyses, or present a business dashboard.
* *Business knowledge and design skills* enable analysts to be more than simple backroom statisticians. They must be familiar with the business functions and processes to which analytics are being applied–marketing, finance, HR, new product development, and the like. They need enough general business background to work at the interfaces of business processes and problems. They also must have insight into the key opportunities and challenges facing the company, and know how analytics can be used to drive business value. One study of quantitative analysts suggested that they have more business acumen than their nonanalytical counterparts [1].
* *Data management skills* are perhaps even more important to analytical professionals than statistical and mathematical expertise. It is often commented that such professionals spend the majority of their time manipulating data–finding, integrating, cleaning, matching, and so on. And the most commonly sought software skill by employers of data scientists is not a statistical program, but rather SQL–a query language for data management [2]. There is little doubt that analytical professionals need skills in managing and manipulating data, and for some this activity will constitute a major component of their jobs.
* *Relationship and consulting skills* enable analysts to work effectively with their business counterparts to conceive, specify, pilot, and implement analytical applications. Relationship skills–advising, negotiating, and managing expectations–are vital to the success of all analytical projects. Furthermore, an analyst needs to communicate the results of analytical work: either within the business to share best practices and to emphasize the value of analytical projects or outside the business to shape working relationships with customers and suppliers, or to explain the role of analytics in meeting regulatory requirements (e.g., utility company rate cases). This skill has been described as “telling a story with data [3].”
* *Coaching and staff development skills* are essential to an analytical organization, particularly when a company has a large or fast-growing pool of analysts, or when its analytical talent is spread across business units and geographies. All analytical professionals may not need them, but they are certainly required for supervisors and managers of large teams. When analytical talent is not centralized, coaching can ensure that best practices are shared across the company. Good coaching not only builds quantitative skills but also helps people understand how data-driven insights can drive business value.

One survey of quantitative analysts' activities suggested that there are really several categories of the role [4]. Based on their self-reported time allocation across 11 different analytical activities, the analytical professionals surveyed were clustered into four groups: generalists, data preparation specialists, programmers, and managers. Every participant indicated they did a little of each activity; however, managers mostly managed, programmers mostly programmed, and data prep folks mostly worked on data acquisition and preparation. The generalists do all these activities, of course, but focus more on analysis, interpretation, and presentation than other activities. Across all four categories, the least amount of time was spent on data mining and visualization.

Of course, few individuals come equipped with the full array of skills I've described; this is where teaming comes in. To constitute effective teams, a company needs the right mix of analytical talent in its teams of analysts. For example, it is often a good idea to balance hard-core quantitative experts–who focus on more advanced analytical techniques–with business-oriented “translators”–who have a broader skill set, combining strong analytics with business design and management skills to link professionals to their customers.

3.2.1 More Advanced or Recent Analytical and Data Science Skills

The practice of analytics has changed substantially over several different “eras [5].” However, the skills I've described for basic analytical work don't go away over time. That's in part because companies still have a need for descriptive analytics and the other activities performed in early analytics periods, and also because the skills required for that era still apply in later eras of analytics. However, as analytical practice has evolved, new skills are added. That is, the skills for doing analytics across the different eras of analytical practice, unfortunately, are cumulative. To be more specific, none of the statistical, business acumen, relationship, data management, and coaching capabilities required for traditional quantitative analysis go away when organizations move into the era of “big data.” This occurred around the turn of the twenty-first century in Silicon Valley, when organizations needed new data management and analytical approaches for the rise of online business.

But there are new skills required in the big data era. Data scientists–the new term for people doing high-skill analytical and data management work in this environment–typically have advanced degrees in technical and scientific fields [6]. Because they are testing many different approaches to online operations and commerce, they need experimentation skills, as well as the ability to transform unstructured data into structures suitable for analysis. In Silicon Valley, performing these tasks also requires a familiarity with open-source development tools. If the data scientists are going to help develop “data products”–products and services based on data and analytics–they need to know something about product development and engineering. Perhaps because visual displays are a good way to comprehend a very large data set, the time that big data took off was also the time that visual analytics became widely practiced in large organizations, so a familiarity with visual display of data and analytics also became important during this period.

The next era, which I would argue began around 2012 or 2013 in the most sophisticated companies, involved the combination of both big and small data for analytics within large organizations. What skills got added at this point? In addition to mastering the new technologies used in combining big and small data, there's a lot of organizational and process change to be undertaken. If operational analytics means that data and analytics will be embedded into key business processes, there's going to be a great need for change management skills. At UPS, for example, which initiated a large real-time driver routing initiative called ORION during this period, the most expensive and time-consuming factor by far in the project was change management–teaching about and getting drivers to accept the new way of routing. This period was also marked by the early use of statistical machine learning approaches, which were necessary to handle the large and fast-changing data environment of the period.

The current era, which started perhaps 5 years ago in online businesses and 2 years ago in other industries, involves extensive use of artificial intelligence or cognitive technologies. This means that analysts and data scientists need a heavy dose of new technical skills–machine and deep learning, natural language processing, and so forth. There is also a need for work design skills to determine what tasks can be done by smart machines, and which ones can be performed by (hopefully) smart humans.

The cumulative nature of these additional skills over time means that it is even more important to take a team-based approach to analytical and data science projects. It is impossible to imagine, for example, that someone who possesses the rare skill of deep learning analytics would also have all the other skills I've mentioned thus far in this chapter. The only way to have all the necessary skills on a team is to staff projects with people who hold different–and hopefully complementary–skill sets.

3.2.2 The Larger Team

Analytics were initially created to improve human decision-making. But there are many circumstances in organizations in which analytics aren't enough to ensure an effective decision, even when orchestrated by a human analyst. In order for analytics to be of any use, a decision-maker has to assess the analytical outcomes, make a decision on the basis of them, and take action. Since decision-makers may not have the time or ability to perform analyses themselves, such interpersonal attributes as trust and credibility between analysts and decision-makers come into play. If the decision-maker doesn't trust the analyst or simply doesn't pay attention to the results of the analysis, nothing will result from the analytical work, and the statistics might as well never have been computed.

I cited one such example in my first book on analytics [7]. In the course of research for that book, I talked to analysts at a large New York bank who were studying the profitability of the bank's branch network. The analysts went through a detailed and highly analytical study in the New York area–identifying and collecting activity-based costs, allocating overheads, and even projecting current cost and revenue trends for each branch in the near future. The outcome of the analysis was an ordered list of all branches and their current and future profitability, with a clear red line drawn to separate the branches that should be left open from those that should be closed.

The actual outcome, however, was that not a single branch was shut down. The retail banking executive who sponsored the study was mostly just curious about the profitability issue, and he hardly knew the analysts. He probably wasn't aware of all the work that would go into the analysis process. He knew–but the analysts didn't–that there were many political considerations involved in, say, closing the branch in Brooklyn near where the borough president had grown up, no matter where it ranked on the ordered list of branches. Basing actions on analytics often require a close, trusting relationship between analyst and decision-maker, and that was missing at this bank. Because of the missing relationship, the analysts didn't ask the right questions about the analysis, and the executive didn't frame the question for them correctly.

Instead of just analysts and data scientists, there are really three groups whose analytical skills and orientations are at issue within organizations. One is the senior management team–including the CEO–that sets the tone for the organization's analytical culture and makes the most important decisions. Then there are the professional analysts and data scientists, who gather and analyze the data, interpret the results, and report them to decision-makers. The third group is a diverse collection I have referred to as analytical amateurs. They comprise a large category of “everybody else,” whose daily use of the outputs of analytical processes is critical to their job performance. These could range from frontline manufacturing workers, who have to make multiple small decisions on quality and speed, to middle managers, who also have to make decisions with respect to their functions and units–which products to continue or discontinue, for example, or what price to charge for them. IT employees who put in place the software and hardware for analytics also need some familiarity with analytical topics, and also qualify as analytical amateurs.

To really succeed with analytics, a company will need to acquaint a wide variety of employees with at least some aspects of analytics. Managers and business analysts are increasingly being called on to conduct data-driven experiments, interpret data, and create innovative data-based products and services [8]. Many companies have concluded that their employees require additional skills to thrive in a more analytical environment. One survey found that more than 63% of respondents said their employees need to develop new skills to translate big data analytics into insights and business value [9]. Bob McDonald, at one point CEO of Procter & Gamble and then head of the U.S. Veterans Administration, said about the topic of analytics (and business intelligence more broadly) within P&G:

We see business intelligence as a key way to drive innovation, fueled by productivity, in everything we do. To do this, we must move business intelligence from the periphery of operations to the center of how business gets done.

With regard to the people who would do the analysis, McDonald stated:

I gather there are still some MBAs who believe that all the data work will be done for them by subordinates. That won't fly at P&G. It's every manager's job here to understand the nature of statistical forecasting and Monte Carlo simulation. You have to train them in the technology and techniques, but you also have to train them in the transformation of their behavior [10].

Of course, all senior executives are not as aggressive as McDonald in their goals for well-trained analytical amateurs. But in even moderately sophisticated companies with analytics, there will be some expectations for analytical skills among amateurs of various types. As Jeanne Harris and I wrote in the new edition of our book *Competing on Analytics*,

To succeed at an analytical competitor, information workers and decision-makers need to become adept at three core skills [11]:

1. *Experimental:* Managers and business analysts must be able to apply the principles of scientific experimentation to their business. They must know how to construct intelligent hypotheses. They also need to understand the principles of experimental testing and design, including population selection and sampling, in order to evaluate the validity of data analyses. As randomized testing and experimentation become more commonplace in the financial services, retail, and telecommunications industries, a background in scientific experimental design will be particularly valued. Google's recruiters know that experimentation and testing are integral parts of their culture and business processes. So job applicants are asked questions such as “How many tennis balls would fit in a school bus?” or “How many sewer covers are there in Brooklyn?” The point isn't to find the right answer but to test the applicant's skills in experimental design, logic, and quantitative analysis.
2. *Numerate:* Analytical leaders tell us that an increasingly critical for their workforce is to become more adept in the interpretation and use of numeric data. VELUX's [Anders] Reinhardt [until recently global head of business intelligence at the Danish window company] explains that “Business users don't need to be statisticians, but they need to understand the proper usage of statistical methods. We want our business users to understand how to interpret data, metrics, and the results of statistical models.” Some companies, out of necessity, make sure that their employees are already highly adept at mathematical reasoning when they are hired. Capital One's hiring practices are geared toward hiring highly analytical and numerate employees into every aspect of the business. Prospective employees, including senior executives, go through a rigorous interview process, including tests of their mathematical reasoning, logic, and problem-solving abilities.
3. *Data literate:* Managers increasingly need to be adept at finding, manipulating, managing, and interpreting data, including not just numbers but also text and images. Data literacy is rapidly becoming an integral aspect of every business function and activity. Procter & Gamble's former chairman and CEO Bob McDonald is convinced that “data modeling, simulation, and other digital tools are reshaping how we innovate.” And that changed the skills needed by his employees. To meet this challenge, P&G created “a baseline digital-skills inventory that's tailored to every level of advancement in the organization.” The current CEO, David Taylor, also supports and has continued this policy. At VELUX, data literacy training for business users is a priority. Managers need to understand what data are available, and to use data visualization techniques to process and interpret them. “Perhaps most importantly, we need to help them to imagine how new types of data can lead to new insights,” notes Reinhardt [12].

As with analytical professionals, additional function unit- or business unit-specific expertise in analytics may be needed by amateurs. In the case of IT professionals, for example, those who provision and support data warehouses and lakes should have some sense of what analyses are being performed on data, so that they can ensure that stored data are in the right formats for analysis. HR workers need to understand something about analytics so that they can hire people with the right kinds of analytical skills–and how analytics can be employed to identify promising employees, or those likely to leave the company soon. With the rise of artificial intelligence, even the corporate legal staff may need to understand the implications of a firm's approach to automated decision-making in case something goes awry in the process. There are also an increasing number of AI applications in corporate litigation as well.

Interview With Greta Roberts

*When asked for her thoughts on the essentials of assembling an analytics team, cofounder and CEO of Talent Analytics Greta Roberts responded:*

One thing that has been a curiosity for us has been the question of whether there a quintessential data scientist. If you had a list of all the necessary attributes, could you find it all in one person? I never believed that for an instant, but it was anecdotal. We wanted to study it quantitatively. It was interesting because, since it was on the analytics side, you would think the first thing that analytics people would do is to use a quantitative approach to understand the people who are actually doing the analysis. Harlan Harris, Marck Vaisman, and Sean Murphy wrote a book–*Analyzing the Analyzers*–that makes me ask, “Why not just apply analytics to understand the analyzers?” That's when we really said that, instead of just saying anecdotally, “We think this is what a data scientist is,” we would actually see if there is a way to really understand them.

Just as in IT there is not just one kind of IT person, there is not one kind of analytics person. I think because analytics is still relatively forming in its new iteration–you do have people that do the entire analytics workflow. There are people that need to do everything from forming the question to data acquisition and collection to visualization, programming, interpretation, presentation, communication…you name it. I think we're starting to see that some of that is breaking into specialization. We've seen anecdotally that sometimes you have data scientists who now just do the data acquisition and collection (and maybe preparation) side, and then turn the work over to people to do the programming and the design of the algorithms. The programmers, in turn, then turn it over to people who actually do the presentation. We've been very interested because of the work that we do here, always really understanding the people that do the analytics work. We've been interested in moving beyond anecdotes around analytics professionals.

There is not a single kind of marketer or one kind of sales rep or one kind of IT person…or one kind of anything, really. For any role, there is always a lot of granularity inside the role. We are really only interested in the analyzers and whether there is a way to analyze the analyzers. Is there a way to categorize a little better to make it easier for people to identify the people who are doing the work? I think the passion comes from the interest in being involved. It is interesting that there is a quantitative approach, and it's particularly interesting to use the same approach–analysis–to understand the analyzers. That is fun on all kinds of different levels!

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

3.3 Managing Analytical Talent

In addition to hiring people with the right kinds and levels of skills, there are a number of activities that are involved in ongoing management of analytical talent [13]. One such activity is to conduct an assessment of the analytical skills within your organization and a “gap analysis” of the differences between the current state and the desired future state. The level of details in the assessment may vary by the purposes of the organization, but may include a roles inventory of analytical jobs and their locations within the organization, a skills inventory, and an analytics talent map. The skills inventory might include a listing of the analytical skills required, and a comparison to the desired skill levels and numbers of people possessing them. A talent map is a high-level mapping of current roles and skills, comparisons to desired future objectives, and elements of plans to close the gap–all ideally in some visual format that is easily comprehensible by busy executives.

The factors measured in the assessment will also differ by the strategies and priorities of the organization conducting the assessment. Some typical examples of factors include the following:

* How many people are there in each of the major analytics functions?
* What percentage of analytical professionals are capable of predictive and prescriptive analytics, as opposed to descriptive analytics?
* How many data scientists are able to use machine learning to build models?
* What percentage of data management-oriented employees have any experience with Hadoop and other big data tools?
* How many employees are familiar with each of the software tools in our approved portfolio for analytics?
* How many analytics staff have close and trusting relationships with the business leaders in the units and functions they serve?
* What analytics/technical skills exist within the current staff by type and number of years of experience?
* What percentage of analytics team members have more than 3 years, or less than 1 year, of experience in analytics?
* Which software/tools have the most and least number of skilled resources available for development and support activities?

Answering these types of questions can allow analytics leaders to build the initial foundation of their organization's talent strategy and roadmap. In order to ensure that the information is relevant to the entire enterprise, it is important to involve all analytics leaders within the company and to include questions or decision points that address the unique nature of the organization and industry. In addition to providing important information, for highly decentralized analytics groups such an inventory can also be a first step toward building a greater level of cohesion.

After doing the assessment, a company will normally want to formulate some objectives and plans for what to do about the results, with a time frame for planned changes. One company, for example, determined that only 5% of its analytics staff were comfortable with predictive analytics, and it wanted to shift to 95% with that skill over 5 years. Another organization determined that its staff lacked close relationships with business leaders, so it developed clearer assignments of analysts to business units, and asked business leaders to participate in annual performance assessments for analytics staff.

A one-time talent assessment is of limited value. People, their skills, and objectives for new capabilities change all the time. Organizations should reassess their analytical roles and skills every year or two. Once an assessment process is in place, it can be repeated relatively easily.

3.3.1 Developing Talent

Many analytics organizations primarily think about hiring people with needed skills. But it is often less expensive and more effective to develop skills through education and training programs, either in-house or in partnership with universities. If there is a critical analytical skill that an organization identifies that is particularly important, it is not difficult to arrange a training program for it. There are, for example, training programs available for organizations that want their analysts to achieve CAP certification from INFORMS.

For another example within a specific firm, Cisco Systems has been expanding for several years into advanced services that analyze the data thrown off by devices like routers and switches. In addition, Cisco has been using analytics extensively for internal purposes, such as sales propensity modeling and demand/production forecasting.

However, managers within Cisco felt that they lacked the data science skills to effectively perform all these activities. Desmond Murray, a Senior Director for IT at Cisco, was running Enterprise Data and Data Security for the company in 2012. His team was adopting new big data technologies (Hadoop, SAP HANA, etc.) for the company to use, but demand within the business was limited. He concluded that a set of educational offerings around data science would build awareness and stoke demand for these new technologies.

Murray designed a distance education (an obvious approach, given Cisco's distance conferencing business offerings) program on data science with two different universities. The program would last for 9 months and results in a certificate in data science from the university. Students attend virtual classes a couple of nights a week, and of course had homework as well. Cisco is now on its sixth student cohort with 40 students in each. About 300 data scientists have been trained and certified, and are now based in a variety of different functions and business units at Cisco.

But Murray, by now head of the Enterprise Data Science Office within the IT organization, didn't stop there. He realized that the newly trained data scientists needed some support from their managers if they were going to be satisfied in their new roles. So Cisco also created a 2-day executive program led by business school professors on what analytics and data science are and how they are typically applied to business problems. The program also covers how to manage analysts and data scientists, and how to know whether their work is effective. Cisco's initiatives to develop its employees' analytics and data science skills are relatively unusual, but they don't have to be. Any company that is serious about analytics and data science could undertake similar steps.

3.3.2 Working with the HR Organization

Analytics and data science organizations in companies can do a lot to identify and inculcate needed skills. At some point, however, it will probably be wise to collaborate with a company's human resources (HR) function to institutionalize talent management processes. HR groups can help to establish formal job titles, create linkages between skill and seniority levels and compensation, and provide internal and external resources for training. If analytics and data science skills are considered strategic, HR groups can help to source, nurture, and manage them. Many HR organizations are themselves interested in doing more with analytics in their own functions, so a partnership with analytics groups can be mutually beneficial.

HR organizations can provide guidance about the type of future skills that the organization will need. Additionally, HR leadership can describe the types of nontechnical skills that they are planning to develop or already have available to support the analytics function (e.g., business acumen, relationship, or communication skills).

At Cisco, the creation of data science skill development programs revealed that there was no standard at Cisco–or at many other firms, for that matter–for who is a serious data scientist and who isn't. So they created a “Pyramid of Analytic Knowledge” to classify different levels of expertise and establish a career track. Murray and his successor worked with Cisco's HR organization to incorporate these into official job classifications and compensation bands.

Interview With Russell Walker

*When asked about soft skills and the analytics professional, Kellogg School of Management Clinical Associate Professor Russell Walker responded:*

Many of us have personality traits and interests that we cannot divorce ourselves from very easily. However, I suspect that with appropriate awareness and training, there are probably effective tools for doing so. I also suspect that this that would be an enormous undertaking and perhaps even an exercise in great frustration for many people.

I would not necessarily expect employees to be someone other than who they are. But in a business setting, employees should be aware of the impact of their work on others and the contribution of others to their work. This is best achieved by creating some sort of collaborative environment, and this is an approach that more analytical professionals should embrace. Perhaps you have some tasks at work that you can–or even should–do alone. However, many tasks cannot be accomplished alone. Your manager is involved, your coworkers are involved, your customers are involved–someone pulls the data, someone performs the analyses, someone creates the PowerPoint deck, etc. In an analytics project, there is generally an enormous amount of work to do across the team, and the team is often rather disparate. So being mindful and respectful of this aspect of analytics is critical.

In a seemingly simple exercise, I ask students I have assigned to work on a project in teams to go to dinner as a team and interview each other. The objective is to gain a better understanding of the specific strengths and interests of their teammates. I think this goes a very long way in simply helping everybody communicate. There are people who are much better at so-called soft skills, just as there are people who are better at drawing than others, and again this is probably driven by personality. Can we teach everybody to draw? With sufficient effort, we probably could make everybody good–not necessarily an artist, but reasonably accomplished at drawing. But could we make everybody a grand master craftsman or an artist? I do not think so!

So we should all recognize limitations and strengths in others and in ourselves, and by doing so see that a team can be better if someone takes the lead role in the presentation, someone takes the lead role in model building, etc. Someone should not go to work and be expected to dwell on and toil in their weaknesses. Corporations are actually pretty good at addressing this; at the end of the year, your manager will perform an annual review and say, “Here are all the things you didn't do well, and this is why we're not going to give you a full bonus.” But if we look at sports as an example, no coach or general manager goes to the quarterback at the end of a season and says, “You didn't kick any field goals this year, so we're going to take money away from you.” You ask the quarterback to do his job, and you let the kicker do his job.

If you are acutely aware of your strengths and your limitations, you can avoid being assigned to a role in which you are set up to fail by having a frank discussion with your employer. And employers can avoid assigning employees–their human capital and valuable assets–roles for which they might not have the requisite traits or interests.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

3.4 Organizing Analytics [14]

One of the key questions to address in managing analytical teams is “How should we best organize our analysts and data scientists?” It is a common question arising from a common situation: Analysts and analytics/big data projects are often scattered across the organization. That is how companies get started with analytics–here and there as pockets of interest arise. However, when an organization starts to get serious about analytics and data science, it often adopts an enterprise perspective in order to develop analysts effectively and deploy them where they create the greatest business value. In order to achieve these objectives, pockets of analytics and data science usually need to be coordinated, consolidated, or centralized.

The trend over the past decade has clearly been toward centralization of analysts, and that makes sense for several good reasons. If a company wants to differentiate itself in the marketplace through its analytical capabilities, it doesn't make sense to manage analytics locally. Skilled and experienced analysts and data scientists are a scarce and high-demand resource. A central function can deploy them on the most important projects, including cross-functional and enterprise-wide projects that may be otherwise difficult to staff. Centralization also facilitates analyst development because people have more opportunity to connect with and learn from one another. In addition, a central group with a critical mass of people helps with recruiting analysts by demonstrating the organization's commitment to analytics and providing new hires with a community. Finally, research led by my frequent coauthor Jeanne Harris [15] suggests that analysts in centralized and well-coordinated structures are more engaged and less likely to leave their employer than their decentralized counterparts.

However, recent trends suggest that analytics and data science teams are not immune from the normal pressures that move centralized functions in a more decentralized direction. Previously centralized analytics groups have been decentralized and dispersed in several different companies over the past year or two. The leaders of these groups cite several reasons for the decentralization, including the visibility of centralized budgets, complaints of lack of responsiveness by business unit and function leaders, and perceptions of excessive bureaucracy in large analytics groups. It seems likely, then, that despite the efficiency and effectiveness benefits of centralization, there will be the usual oscillation between centralization and decentralization in analytics and data science groups.

Another common situation among organizations I encounter is a significant analytics presence in one or two business functions, plus small pockets of analytics across the rest of the organization. The lead functions vary by industry–risk management and trading in financial services, engineering and supply chain in manufacturing, and marketing in consumer businesses. The challenge here is simultaneously to connect the pockets of analytics and spread the wealth of expertise resident in the advanced units. In these cases, full centralization could be unnecessarily disruptive, so the organization needs other mechanisms to coordinate analyst talent supply.

In the book *Analytics at Work*, Jeanne Harris, Bob Morison, and I discuss five common organizational models [16]. They are a useful place to start, but organizing your analysts isn't as simple as just picking one. There are different organizational circumstances, with many variables and mitigating factors in play, and many variations on these five options. This section attempts to decompose the organizational models for analysts and data scientists, and provides tools for developing and tuning your own model.

3.4.1 Goals of a Particular Analytics Organization

When debating alternative organizational structures for analytical and data science groups, it is important to keep the overriding goals for the organization in mind. Typically, the following are some of the goals of analytical groups and their leadership within companies:

* Supporting business decision-makers with analytical capabilities
* Helping to develop new customer-oriented products and features involving data and analytics
* Providing leadership and a “critical mass” home for analytical and data science-oriented people, and the ability to easily share ideas and collaborate on projects across analysts
* Fostering visibility for analytics and big data throughout the organization, and ease in finding help with analytical problems and decisions
* Creating standardized methodological approaches, tools, and processes
* Researching and adopting new analytical and data science practices
* Reducing the cost to deliver analytical outcomes
* Building and monitoring analytical capabilities and expertise

Different priorities for these goals may lead to different organizational models. For example, the goal of supporting business decision-makers with analytics may be best served by locating analysts directly in business units and functions that those decision-makers lead. That decentralized approach may also be the most effective one for development of products and services based on analytics and data. However, such decentralization may work against the goal of giving analysts and data scientists the ability to easily share ideas and collaborate.

Note that throughout this section (and the chapter in general) I have generally mentioned analysts and data scientists in the same breath. This usage is intentional; I believe that it was always difficult to clearly differentiate between “traditional” quantitative analysts and data scientists, and it is becoming increasingly difficult over time. At one point, data scientists tended to be more experimentally focused than traditional analysts, and also were likely to write code to transform unstructured data into structured formats for analysis. But now the tasks that these two groups perform certainly overlap, and the cachet of the data scientist title means that it is being applied to more jobs. My assumption is whatever organizational structure makes sense for one group also makes sense for both; that is, analysts and data scientists should be part of the same larger group. Of course, there are always exceptions to any organizational structure rule.

As I suggested above, no set of organizational structures and processes is perfect or permanent, so organizations must decide what particular goals are most important at any point in their analytical life cycles. For example, if an organization has had a centralized group of analysts and data scientists for a while and it has become unresponsive to business unit needs, it may be time to establish stronger ties between analysts and specified business units and leaders. A company with highly localized analytics may need to switch, at least for a while, to a more centralized structure. If possible, however, organizations should avoid rapid swings from highly centralized structures to highly decentralized structures, and back again. There are usually less disruptive ways to achieve the desired goals.

3.4.2 Basic Models for Organizing Analytics

[Figure 3.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#c03-fig-0001) shows the common organizational models described in *Analytics at Work*.

[**Figure 3.1**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#_backF1) Common organizational models described in *Analytics at Work*. (Adapted from Davenport, Harris, and Morison, 2010.)

In a *centralized* model, all analyst groups are part of one corporate organization. Even if located in or primarily assigned to business units or functions, all analysts report to the corporate unit. This obviously makes it easier to deploy analysts on projects with strategic priority, as well as to develop skills and build community. However, especially if the analysts and data scientists are all housed in the corporate location, it can create distance between them and the business (although this can be mitigated by other factors, as I describe below). Implementing a centralized model for analytics is easiest where there is successful precedent for operating other functions or managing scarce resources as centralized shared services.

In a *consulting* model, all analysts/data scientists are part of one central organization, but instead of being deployed from corporate to business unit projects, the business units “hire” analysts for their analytical projects. This model is more market driven, and especially important here is the analyst/consultants' ability to educate and advise their customers on how to utilize analytical services–in other words, to make the market demand smart. This model can be troublesome if enterprise focus and targeting mechanisms are weak, because analysts may end up working on whatever business units choose to pay for (or whatever wheel is squeakiest) rather than what delivers the most business value.

In a *functional* or “best home” model, there is one major analyst/data scientist unit that reports to the business unit or function that is the primary consumer of analyst services. This analyst unit typically also provides services in a consulting fashion (or even better, strategic prioritization) to the rest of the corporation. As already mentioned, many financial services and manufacturing firms have, in effect, a functional model today, with one or two well-established analyst groups in functions like marketing or risk management. The best home may migrate as analytical applications are completed and the analytical orientation of the corporation changes, typically from operations to marketing.

A *center of excellence* model is a somewhat less centralized approach that still incorporates some enterprise-level coordination. In this structure, analysts are based primarily in business functions and units, but their activities are coordinated by a small central group. The CoEs are typically responsible for issues such as training, adoption of analytical tools, and facilitating communication among analysts. The CoE builds a community of analysts/data scientists and can organize or influence their development and their sharing across units. It is most appropriate for large, diverse businesses with a variety of analytical needs and issues, but that still would benefit from central coordination. This is perhaps the most popular of the five models. In the era of business intelligence, this model was sometimes called a “business intelligence competency center.”

There are many variations on this model, depending on the powers of the CoE. Do analysts report to it dotted line? Does it control the staff development agenda and resources? Does it double as a Program Management Office (PMO), with powers to coordinate priorities and resources across business units? Or are the business units solidly in charge of their analysts?

In a *decentralized* model, analyst groups are associated with business units and functions, and there is likely an analytics group or groups for corporate functions, but there is no corporate reporting or consolidating structure. This model makes it difficult to set enterprise priorities and difficult to develop and deploy staff effectively through borrowing and rotation of staff. It is most appropriate in a diversified multibusiness corporation where the businesses have little in common. But even then it makes sense to build a cross-business community of analysts so that they can share experience. As a result, this is the model I (and my *Analytics at Work* coauthors) am least likely to endorse.

Beneath the surface, each of these models is essentially either centralized or decentralized. The consulting and functional models are variations on centralization–the consulting model has different funding and deployment methods, and the functional model is centralized, just not at corporate. The CoE model is an overlay on a decentralized structure. So are other hybrid models, most commonly a combination of decentralized analyst groups in business units plus a central group at corporate that focuses on cross-functional, cross-organizational, and enterprise-wide initiatives.

These five models have pros and cons and trade-offs in terms of deployment and development and other objectives. [Figure 3.2](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#c03-fig-0002) indicates the strengths of each in terms of four specific goals.

[**Figure 3.2**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#_backF2) The strength of the five models. (Adapted from Davenport, Harris, and Morison, 2010.)

3.4.3 Coordination Approaches

One basic structure may be the best general fit, but no model will be best in terms of meeting all goals. Whatever the basic model, there will be a need to coordinate across analyst groups or across different parts of the business that are consuming analyst services. In a sense, all models are hybrids. Even if all analysts and data scientists work in one centralized corporate unit, the customers for their services are spread across the enterprise. You need coordination mechanisms to manage and meet demand for analytics.

There are a variety of common coordination mechanisms, some of which we've already mentioned. The mechanisms can supplement the formal reporting structure for purposes of enabling groups to plan and work together, and developing an enterprise view of priorities and resources. Think of them as ways of supplementing and fine-tuning a basic centralized or decentralized model, or of compensating for its inherent weaknesses. And note that all present challenges.

Program Management Office

This is a formal corporate unit for setting enterprise priorities, coordinating analytics and big data initiatives, influencing resource deployment on key initiatives, and facilitating the borrowing of staff across analytics groups. As mentioned above, it may be a function within a center of excellence. PMOs are especially useful where potential business value from analytics is high and resources are scarce and distributed. Under a PMO, the deployment process must be sophisticated to meet the dual needs of project staffing and analyst development.

Federation

Analyst groups and their associated business units work together on priorities, coordination of initiatives, resource deployment, and analyst development under a set of “guidelines of federation.” The most basic form of federation is a clearly chartered enterprise governance or steering committee. These committees add an immediate enterprise view, but they sometimes lack clout and even commitment. Some firms have considered federation as a sixth type of organization model.

Community

Decentralized analysts can be encouraged to share ideas and analytical approaches in a community. Such a community would typically involve occasional meetings, seminars, written communications, or electronic discussions or portals. It may be facilitated by a community organizer, and typically benefits from some budget. In most cases, this is a relatively weak coordination mechanism.

Matrix

Analyst groups report both to their associated business units and to a corporate analytics unit, with one line solid and the other dotted. Establishing dotted-line reporting to a central organization injects an imperative to get coordinated, but dotted-lines can lose their force over time if they're not regularly exercised.

Rotation

Some of the analysts in a centralized model are physically located in and dedicated to business units on a rotational basis. Or there is an enterprise-wide program facilitating the lending and migration of analysts across decentralized units. The strength and success of rotation programs are easy to gauge–analysts really do have mobility across the enterprise.

Assigned Customers

Some centralized analytics groups, such as the one at Procter & Gamble, have assigned or “embedded” analysts to work exclusively with particular business units and the leaders of those units. The assignments fall short of a matrixed tie in the organizational structure, but they help to ensure that the analytical needs of the units and their leaders are met. Recently, however, some of the embedded analysts at P&G have been put into a matrix structure; business units and functions were more comfortable having their analysts report to them.

For purposes of deploying analysts on the most important business initiatives, the PMO is the strongest mechanism. For purposes of developing analysts, all of the mechanisms can help the cause, but rotation programs may have the most profound effect. The coordination mechanisms can be used in combination–for example, a PMO focused on deployment and a community focused on development, or a federation focused on coordination and a matrix focused on ensuring alignment with business needs.

What Model Fits Your Business?

Any basic organizational design for analyst may look good on paper, but it's got to work in the context of how the business already operates. To evaluate, design, implement, and refine organizational structures, you've got to look behind the organization chart and consider some basic variables that have to be working together for any organizational model to succeed. These factors can either mitigate or strengthen the effects of any particular organizational structure. [Figure 3.3](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#c03-fig-0003) lists six key variables [17].

A diagram of a work location

Description automatically generated

[**Figure 3.3**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#_backF3) Six key variables for tuning organizational designs.

*Home location* is the geographical location where analyst groups officially reside for administrative purposes. Home base and formal reporting lines have been the dominant variables in organizational design, especially in companies where more headcount has indicated more power. However, in today's more fluid and collaborative organizations, home location means less and less (especially if coordination mechanisms are effective). Home location is a matter of convenience, with the goals of limiting travel to work locations, accommodating employees' preferences, and getting enough people in one place regularly to sustain a community. In many firms today, analysts are based offshore, either as employees or contractors.

*Work location* is where the work of business analytics is performed, typically a mix of in the field (wherever the business customers of analytical models and services may be) and in regional or corporate analytics centers (where colleagues and support services are readily available). It is generally best to locate analytics work, insofar as possible, where the corresponding business work is. This greatly facilitates communication with business leaders and those who perform the work process under analysis. Make sure that home location and reporting structure don't erect barriers to analysts' working close to the business.

*Reporting structure* is the formal lines of connection, direction, and administration. Analysts and their groups typically report to local business units, to corporate, or to an intermediate unit (e.g., business sector or region) if the corporation is so structured. Some reporting structures are matrixed, with analysts reporting solid-line to business units and dotted-line to the corporate analytics organization, or vice versa. Reporting structure may be predetermined if analytics is part of another organization, such as marketing or IT. Reporting lines should not be so rigid as to impede the flexible staffing and development of analysts. Given the advantages of enterprise coordination of analytics, a least a dotted line to a central group or CoE makes sense in most organizations.

*Business structure* is the shape of the enterprise. Are its business units highly autonomous? Or are they closely coordinated? To what extent do business units already share functions, services, and important-but-scarce resources? Is power concentrated at the regional level? Centralizing analysts and data scientists may seem the logical thing to do, but then prove very impractical if that flies in the face of a locally autonomous or regionalized business structure.

Centralized analytics groups are a natural match for an integrated “one business” business. If business units are intertwined and must work with and rely on one another regularly, you need a centralized or consulting model, or else a strong federation. If business units are autonomous with little interconnection, analysts may stay decentralized, but a center of excellence helps in sharing experience and building the analyst community. And if the enterprise relies extensively on business partners to perform major processes, you may need a centralized structure, especially if there's need or opportunity to coordinate analytics with partners.

*Funding sources* are seldom considered in the context of organizational design, even though paralysis is guaranteed if organizational structure and funding sources are at odds. Friction is minimal if funding follows the lines of formal reporting, but matters are seldom that simple because business services like analytics often have multiple funding sources. These may include funds from corporate, business unit assessments, direct funds from business units, chargeback to business units for analyst time, and project-based funding from the sponsoring business unit or units. The organizational questions are as follows: To what extent does the basic model under consideration align with funding sources? How does funding need to be revised or influenced by coordination mechanisms to support the analytics organization and its work?

Project-based funding is the most market and demand driven, but it requires a certain level of maturity among business customers in setting analytics ambitions and priorities, and among analyst groups in advising customers and marketing their services. Project-based funding (or other funding for services performed) should in most cases be supplemented by seed funding (to foster innovation) and infrastructure funding (to build capability), usually from corporate.

*Infrastructure* includes the configuration and ownership of other essential resources, especially technology and data. This variable is similar to funding sources–alignment is essential to the success, but the variable is seldom considered in organizational design. Analysts cannot work across business processes and units if local systems and databases, inconsistent tools, and fragmented infrastructure prevent it. And business units cannot incorporate new technologies and techniques for analytical applications of corporate standards prevent it. To capitalize on analytics, the infrastructure must be local-but-interoperable or corporate-but-flexible.

As a practical matter, those six variables are never perfectly aligned, and organizations will have to experiment with and adjust the coordination mechanisms over time. As a common example, if data and technical infrastructure are fragmented, a company might phase an organizational consolidation alongside (or slightly in advance of) the rationalization and consolidation of those resources.

3.4.4 Organizational Structures for Specific Analytics Strategies and Scenarios

There are at least seven scenarios [18] for how enterprises approach and employ analytics ([Table 3.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#c03-tbl-0001)). These different emphases suggest different basic organizational models.

[**Table 3.1**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch3/format/chapter-epub/OPS/c03.xhtml?hmac=1704948814-ikIWNa7%2BHwtKo6aqx2mKzmtQEqtTfuCQsk505HAWXDY%3D#_backT1) Other factors driving effectiveness of analytics organizational structures.

| **Scenario** | **Definition** | **Basic model** |
| --- | --- | --- |
| Traditional analytics and BI | Make analytics tools and resources available to meet a broad variety of business needs | Centralized |
| Analytics for the masses | “Democratize” analytics and spread their use broadly across the organization | Centralized, with considerable effort to create self-service approaches |
| Big data | Tap the analytical potential of unstructured and nonquantitative data | Functional if one unit is in the lead leveraging these data; otherwise, consulting or centralized |
| Decision-centered | Enable the rapid and accurate execution of business decisions–both frequent/structured and infrequent/new | Model relatively unimportant if analysts can work closely with decision-makers, with a means of sharing methods and experience |
| Embedded analytics | Make real-time, automated analytical decisions part of core business processes and systems | Centralized or consulting, and close relationship with IT |
| Function- or process-specific analytics | Use specialized analytical technologies and applications to excel at a differentiating business process | Functional if there's an organization focused on the process; otherwise, consulting or centralized |
| Industry-specific analytics | Use specialized analytical technologies and applications to excel at processes common to an industry | Centralized or consulting, or functional if focus is on very specific applications |

3.4.5 Analytical Leadership and the Chief Analytics Officer

Another key organizational question is the leadership role for analytics within organizations. There are already a substantial number of “Chief Analytics Officers (CAO)”, and I expect that more will emerge. The role may not always have that title (it may, for example, be combined with Chief Data Officer–particularly in financial services), but there is a need–at least for each of the three centrally coordinated models described above–for someone to lead the analytics organization. The CAO could be either a permanent role, or a transitional role for an organization wanting to improve its analytical capabilities. There are a few Chief Data Scientists in organizations, but often these roles are combined with Chief Analytics Officer titles.

The roles of a Chief Analytics Officer could include any or all of the following:

* Mobilizing the needed data, people, and systems to make analytics succeed within an organization.
* Working closely with executives to inject analytics into company strategies and important decisions.
* Supervising the activities and careers of analytical people.
* Consulting with business functions and units on how to take advantage of analytics in their business processes.
* Surveying and contracting with external providers of analytical capabilities.

One key issue for the CAO role is whether analytical people across the organization should report to it. While an indirect reporting relationship (as one dimension of a matrixed organization) may be feasible, a CAO without any direct or indirect reports seems unlikely to be effective.

In one insurance firm, for example, the CEO was passionate about the role of analytics, and named a CAO as a direct report. But the CAO had only a couple of staff; all other analytics people in the organization did not report to him. The CEO did not want to “rock the organizational boat” by having such traditional analytical functions in insurance as actuaries and underwriters report to the CAO. As a result, the CAO felt he had no ability to carry out his objectives; he resigned from the role, and the CEO did not replace him.

3.5 To Where Should Analytical Functions Report?

There are a variety of different places in the organization to which centralized analytical/data science groups and their CAO leaders can report. While there is no ideal reporting relationship, each one has its strengths and weaknesses. In the following section each alternative is discussed.

Information Technology

Some organizations, such as a leading consumer products firm, have built analytical capabilities within the IT organization, or transferred them there. There are several reasons why this reporting relationship makes sense:

* Analytics are heavily dependent upon both data and software, and expertise on both of these is mostly likely to reside in an IT function.
* The IT function is used to serving a wide variety of organizational functions and business units.
* Analytics are closely aligned with some other typical IT functions, for example, business intelligence and data warehousing.

Of course, there are some disadvantages as well. IT organizations are sometimes slow to deliver analytical capabilities, and may have a poor reputation as a result. They may also overemphasize the technical components of analytics, and not focus sufficiently on business, organization, behavior, skill, and culture-related issues. Finally, IT organizations typically want to produce standardized and common solutions, and this may inhibit one-off analytical projects. In principle, however, there is no reason why IT organizations cannot overcome these problems.

Strategy

A few analytical groups, including those at a large retailer, report to a corporate strategy organization. This relationship allows analysts to become privy to the key strategic initiatives and objectives of the organization. Another virtue is that strategy groups are often staffed by analytically focused MBAs who may understand and appreciate analytical work, even if they cannot perform it themselves. The possible downsides to this reporting relationship are that strategy groups may not be able to marshal the technical and data resources to make analytical projects succeed, and strategy groups are usually relatively small.

Shared Services

In organizations with a shared administrative services organization, an analytics group can simply be part of that capability. The primary benefit of such a reporting structure is that analysts can serve anyone in the company–and often there are charging and resource allocation mechanisms in place for doing so. The downside is that analytics may be viewed as a low-value, nonstrategic resource like some other shared service functions. With the appropriate mechanisms in place, this problem can surely be avoided.

Finance

Being a numbers-focused function, finance organizations have the potential to be a home for business analytics groups. The obvious virtue of this arrangement would be the ability to focus analytics on the issues that matter most to business performance, including enterprise performance management itself. For some unknown reason, however, most CFOs have not embraced analytics, and the finance function remains a logical, if uncommon, home for analytical groups. At some firms, however, including Deloitte (for internal analytics) and Ford, the finance function is beginning to play a much stronger role in championing analytical projects and perspectives.

Marketing or Other Specific Function

As noted above, if an organization's primary analytical activities are concentrated on marketing or some other specific function, then it makes sense to incorporate the analytical group within it. The resulting structure would allow a close focus on the analytics applications and issues in the functional area. Caesars Entertainment, for example, has put analytics in a reporting relationship to marketing. Obviously, it would also make it more difficult for analytical initiatives outside those functional areas to be pursued.

Product Development

The most likely industries for having analytics (and data science) reporting to product development are those–like online businesses–where there are a substantial number of “data products,” or products and services based on analytics and data. There are, for example, analytics groups at LinkedIn, Facebook, and Google who report into product development organizations.

3.5.1 Building an Analytical Ecosystem

Most of the foregoing discussion about analytical capabilities has been focused on organizing and developing internal analytical capabilities. But there is a broad set of analytical offerings that are available from a wide variety of external providers as well. The providers include consultants, IT (primarily software) vendors, offshore analytical outsourcers, data providers, and other categories of assistance. Some provide general analytical help across industries, but in almost every industry there are also specialized analytics and data providers. Many firms can benefit from working with such “analytical ecosystems” to improve their capabilities.

The key in constructing an effective analytical ecosystem is not to let it grow at random, but to identify the analytical capabilities the organization needs overall. Then a decision should be made as to whether internal or external capabilities are most appropriate to fill a specific need. In general, external capabilities make sense when the need is highly specialized, not likely to be needed frequently, and not critical to the organization's ongoing analytical capabilities.

A major pharmaceutical firm's Commercial Analytics group, for example, has a well-developed ecosystem. There is a large group–more than 30–of internal analysts, but their capabilities are supplemented by outside help when necessary. The group has worked with specialized consultants on analysis of physician targeting, for example. The company's primary prescription data provider also works with it on analytics issues. Software vendors have consulted on analytical methods and techniques. Finally, the group supplements its work with help from an offshore analytics vendor in India.

3.5.2 Developing the Analytical Organization over Time

A final point is that analytical organization structures should develop and evolve over time. An internal structure and ecosystem that makes sense at the beginning of developing analytical capabilities will become obsolete later on. For example, it may be very reasonable to have a highly decentralized organizational model early on, but most firms create mechanisms for coordination and collaboration around analytics as they mature in their analytical orientations. It may also make sense to “borrow” a number of external resources in a firm's early stages of analytical maturity before making the commitment to build internal capabilities. In addition, companies may want to add data science capabilities to existing analytics groups to take advantage of the potential of big data.

The best way to adapt organizational capabilities to current needs is with a strategy or plan. Admittedly, in the early stages, there may not be anyone with the formal authority to even create a plan. However, if it appears that analytics are going to be key to an organization's future, it may make sense for a small group of analysts or data scientists to get together and create a bottom-up one.

At a large U.S. bank, for example, the head of the distribution organization (including physical branches, call centers, ATMs, and online channels) realized that she had a large number of analysts in her organization, but they weren't providing the value of which they were capable. She met with the managers of the diverse analytics and reporting groups in her business unit, and asked one of them to take the lead in assessing the problem. His work determined that the vast majority of the groups worked on reports rather than more predictive analytics, and that there were virtually no resources devoted to cross-channel analytics. With this start, the group began to develop a plan to remedy the situation and shift the balance toward predictive analytics and a cross-channel perspective. There was also a major focus on reducing the amount and frequency of reports. Later, this same sponsor moved a different business unit toward heavier use of machine learning technologies.

Plans should probably be revised every year or so, or with major changes in the demand or supply around analytics. There are usually clear signs–if anyone is looking–that the current model has become dysfunctional. It is a key step in an organization's analytical development that someone takes responsibility–either informally or formally–for assessing the organization of analytical resources, and for creating a better model.

No set of skills, plans, or organizational structures is perfect–even for a given time and situation–and every structure or skillset, if taken beyond its limits, will become a limitation. The leaders of contemporary organizations will need to become conversant with their analytical capabilities and how they are organized. Most importantly, they will need to realize when their current organizational approach and team no longer functions effectively, and needs to be restructured and/or reskilled.

References

1. 1 Harris J, Craig E, Egan H (2010) *Counting on Analytical Talent* (Accenture Institute for High Performance), March.
2. 2 Muenchen RA (2017) *The Popularity of Data Science Software*, R4Stats website. Available at <http://r4stats.com/articles/popularity/>. June 19 version.
3. 3 Davenport TH (2014) *Ten Kinds of Stories to Tell with Data*. Harvard Business Review blog post, May 5. Available at <https://hbr.org/2014/05/10-kinds-of-stories-to-tell-with-data>.
4. 4 Roberts P, Roberts G (2013) *Research Brief: Four Functional Clusters of Analytics Professionals*, Talent Analytics Corp, July. Available at <http://www.talentanalytics.com/wp-content/uploads/2012/05/ResearchBriefFunctionalClusters.pdf>.
5. 5 Davenport TH (2013) *Analytics 3.0*. Harvard Business Review, December. Available at <https://hbr.org/2013/12/analytics-30>.
6. 6 Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century. *Harvard Bus. Rev.* 90 (10): 70–76.
7. 7 Davenport TH, Harris JG (2007) *Competing on Analytics* ( Harvard Business Review Press).
8. 8 Davenport TH, Kudyba S (2016) Designing and developing analytics-based data products. *MIT Sloan Management Review*, Fall 2016. Available at <http://sloanreview.mit.edu.proxy3.library.mcgill.ca/article/designing-and-developing-analytics-based-data-products/>.
9. 9 Hartman T (2012) *Is Big Data Producing Big Returns?* Avanade Insights blog post, June 5. Available at <http://blog.avanade.com/avanade-insights/data-analytics/is-big-data-producing-big-returns/>.
10. 10 McDonald quotations from Davenport TH, Iansiti M, Serels A (2013) *Managing with Analytics at Procter & Gamble*, Harvard Business School case study, April.
11. 11 The three core skills needed for analytical managers is adapted research originally published in Harris J (2012) Data is useless without the skills to analyze it. *Harvard Business Review*, September 13. Available at <https://hbr.org/2012/09/data-is-useless-without-the-skills>.
12. 12 Davenport TH, Harris JG (2017) *Competing on Analytics* ( Harvard Business Review Press, revised edition).
13. 13 This section draws on content from a research brief by Harrington E (2014) *Building an Analytics Team for Your Organization*, International Institute for Analytics, September. Available at <http://iianalytics.com/research/building-an-analytics-team-for-your-organization-part-i>.
14. 14 This section is a revised and updated version of a chapter by Morison R, Davenport TH (2012) Organizing analysts, in *Enterprise Analytics*, Davenport TH, ed. ( Prentice Hall).
15. 15 Accenture Institute for High Performance (2010) *Counting on Analytical Talent*.
16. 16 Davenport TH, Harris JG, and Morison R (2010) *Analytics at Work: Smarter Decisions, Better Results* ( Harvard Business Press), pp. 104–109.
17. 17 Framework is based on *Building an Analytical Organization*, Business Analytics Concours and nGenera Corporation, 2008.
18. 18 Framework is based on Morison R, Davenport TH (2008) *Mastering the Technologies of Business Analytics*, Business Analytics Concours and nGenera Corporation.

4  
The Data

*Brian T. Downs*

*Accenture Digital, Data Science Center of Excellence, Dallas, TX, USA*

4.1 Introduction

Regardless of one's area of specialization or interest, it is true that most analytics students and professionals devote most of the effort and energy that goes into training to learning analytics methods and algorithms. A review of a typical curriculum in business analytics will reveal a sharp focus on the tools and techniques required, often in a specific context such as marketing or operations, to be a successful analytics practitioner. Therefore, it is often a surprise for people starting out in the field to discover that on most analytics projects, most of one's time is not spent on using the algorithms recently mastered with such great effort and determination. Rather, it is the lot of an analytics professional to spend most of their time messing with data. This chapter provides a practitioner's view of the different types of data, and some of the challenges in identifying, collecting, and preparing data for analysis.

4.2 Data Collection

4.2.1 Data Types

Before exploring data collection, a review of the various types of data will be useful. [Figure 4.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0001) shows a useful hierarchy for describing these.

A diagram of a data preparation process

Description automatically generated

[**Figure 4.1**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#_backF1) The analytics process.

*Qualitative* data result from classifying something or labeling its attributes. There are three main types of qualitative data. *Nominal* data results when we identify things with named categories that do not have any natural or intrinsic value associated with them. For example, the wooden poles a utility company uses to transmit power to its customers can be classified by the species of tree from which they are made. Pine, fir, and cedar are meaningful categories in that each has intrinsic properties that affect their performance in this application, but there is no obvious way to rank them based on the nominal classification alone. One could use this classification to perform an analysis to see if there are statistically significant differences in the lifespan of wood poles made from each species.

An important special case of nominal data is *binary* data. This type of data places something into one of two mutually exclusive and collectively exhaustive categories, often implying opposite states. A quality inspection of an item on a production line can result in a pass/fail. A production process can be in control/out of control. A magazine subscriber can renew/not renew. This type of data has become increasingly important as methods for predicting how likely an event of interest is to occur have seen widespread use in a variety of contexts. A manufacturing company may wish to predict how likely a machine is to fail given current operating conditions using data that can be collected from the production process. A cable television provider may wish to predict how likely a customer is to drop their cable service given demographic information and their history of problems and complaints. There are many number of classification methods that can predict events with binary outcomes effectively. The challenge in many cases is that historical data that contain multiple instances of each outcome may not be available, or will require some time to collect.

*Ordinal* data are created when one classifies things into categories where there is an implicit relationship between categories. The use of small, medium, and large to describe the size of things has an implicit meaning in many contexts. We expect a medium drink to contain more than a small drink, and a large drink to contain more than a medium drink. In the context of completing a survey, one might be asked to rank something from worst to best on a scale of 1–10, with the expectation that 5 is better than 2, 10 is better than 6, and so on. The problem with both of these examples is that the rank ordering does not tell us the magnitude of the difference between each category. There is no way to know from the classification that a medium drink is 33% bigger than a small drink, or how much better in absolute terms a rank of 10 is than a rank of 5.

*Quantitative* data are created when things are counted or measured. *Discrete* data result from counting things, and therefore is typically expressed as an integer value. The number of nights one has stayed with their preferred hotel chain is an example of discrete data. The number of warranty claims received on a model of smart phone is another. Neither of these things are recorded as fractional values as they refer to discrete events that have occurred an integer number of times.

*Continuous* data are generally anything that can be measured, and as such may have fractional values depending on how fine of a measurement is made. The flow rate of crude oil through a pipeline, the exhaust temperature of a diesel engine, and the daily output of a chemical process are all things that can be measured and the result will generally be a real number. One thing to be cautious about when using continuous data is that the quality and reliability if the data can be affected by the method of collection. Devices such as electronic sensors can be unreliable or influenced by the surrounding environment. Data recorded by human interaction are naturally prone to errors.

Time is also a potential consideration. Data collected from several subjects at approximately the same point in time are referred to as *cross-sectional* data. Examples of cross-sectional data are candidate preferences of voters immediately prior to an election, the high temperature on a given date in the 100 most populated U.S. cities, or the sizes of donations given to a charity during a fund raising drive. The most common purpose for collecting cross-sectional data is to develop an understanding of characteristics of a population at a particular point in time. Data collected from a single subject at several approximately equally spaced points in time are referred to as *time series* data. Examples of time series data are weekly sales for an item, the daily number of visitors to a museum, or the monthly rainfall measured at a weather station. The most common purpose for collecting time series data is to predict future values of the time series, such as when a sales history is used to predict future sales for planning purposes. In many cases, a time series will have measurable components that can be estimated using appropriate analytical methods. A trend indicates a long-term shift in the overall level of the time series, while seasonality is a cyclic pattern that repeats within a specific time interval such as a day or a year. If a single subject is observed over several points in space, the data are referred to as *spatial data*. Spatial data are similar to time series data and are often analyzed with methods designed for time series data. Finally, data are increasingly collected from several subjects at several approximately equally spaced points. These data, which have characteristics of both cross-sectional and time series data, are referred to as *panel data* (or *longitudinal data* or *cross-sectional time series data*). Time series and cross-sectional data are each a special case of panel data in which either the number of periods of time or the number of subjects is one.

Another type of data that has proliferated in recent years is unstructured *text data*. New technologies have been developed to collect and store this type of data, which can be collected from Web sites, social media, and discussion groups in the form of comments, reviews, and opinions. This type of data is stored as documents and may be used for text mining and sentiment analysis. As will be discussed later, the lack of structure in this type of data makes it difficult to store in a traditional database. Therefore, it has been a driving force in the evolution of nonrelational database technologies in recent years.

Interview With Harrison Schramm

*Sometimes clients have data that could be useful to an analytics project. Harrison Schramm, who recently retired from a 20-year career as a helicopter pilot and an operations research analyst in the U.S. Navy, shares his thoughts on obtaining data for an analytics project from a client:*

There are two ways to approach this, and the choice depends on the stakeholder. One way to go about it is to get stakeholders excited about what you are doing and make them want to help you by giving you their data. This is the preferred route.

The other route is to make stakeholders utterly terrified of what you are going to do if they don't give you data. This is a horrible route to take, but sometimes you have to go down this path. If you are working with a large organization, you cannot expect every segment of that organization to be excited about what you're going to do. So if one department is recalcitrant, you just end up having to say, “If you don't give us the data, then we're going to assume this, this, and this…” and you pathologically craft those assumptions so all of a sudden that giving you their data looks a lot better to them than those assumptions you're threatening to make. It's a varsity move–it's not for freshman.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

4.2.2 Data Discovery

There are two types of analytics projects that are often encountered in practice. *Management Consulting*-type projects involve the use of analytics to solve a problem or answer a particular set of questions. These types of projects deal with one-time decisions and the “leave behind” from the effort is a report that contains analysis and recommendations. The questions addressed can be relatively simple, such as “Should I add storage capacity at a facility?”, or they can be complex such as “How can I reduce my variable conversion cost while pursuing a high variety, highly customized make-to-order production strategy?” The analytics practitioner may use a single technique, or a combination of predictive analytics, simulation, and optimization. The data may come from a variety of internal and external sources, but are generally discarded after the project is complete. As such, there may not be a need to collect and merge the data into a permanent and sustainable environment. Data often will be collected in spreadsheet format, from a variety of sources, and will require considerable manual effort to prepare. Detective work may be required to locate some data elements as there may not be a system of record that contains what is needed, or even worse the data that are in the system of record may not be accurate. These types of projects can often be completed by people with analytics as their primary skill set as such people usually have some basic data management skills as well. Larger projects with high volumes of data may require data integration specialists to assist with data preparation.

The other common type of analytics project is *Application Development*. In these projects, analytics tools and algorithms are imbedded in an information technology system to support a set of business processes. Examples of such processes include forecasting and demand planning, sales and operations planning, and production process monitoring and control. In these applications, the analytics component is executed on a periodic basis. This can be anywhere from fractions of a second to monthly, depending on the type of process. Supporting a recurring process requires that the data needed by the analytics models be current and complete. This generally requires the development of a data warehouse, or at least an operational data store (ODS), which may involve combining data from a variety of sources in a single environment, and developing extract, transform, and load (ETL) procedures that capture data from source systems and move it into the analytics environment. ETL processes will also clean and transform data, creating tables and views that can be loaded directly into analytics applications. For these types of projects, the “leave behind” will be a functioning application, as well as the data infrastructure necessary to support it. In addition, there will be documented procedures in place to maintain the integrity of both the analytics models and the data. Application Development projects usually require skill sets beyond analytics, including data integration, system architecture, and data visualization.

At a high level, most analytics projects follow a similar process flow, although the amount of effort and complexity can vary widely depending on the scale of the initiative. [Figure 4.2](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0002) shows a high-level view of typical activities that an analytics practitioner will undertake to complete a project. Note that a sizable proportion of the activities are data related. Data discovery is a critical first step as it is necessary to define the objectives of the work, as well as to determine the likelihood of a successful outcome.

A diagram of data

Description automatically generated

[**Figure 4.2**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#_backF2) Types of data.

Data discovery must begin with discussions with subject matter experts (SME) to understand the research question to be addressed with analytics. In a business context, these are typically people who work in a client business and who are familiar with the attendant operational and business processes. It is the job of the analytics practitioner to understand the business issues at stake, and to frame those issues as testable hypotheses or to propose an approach for addressing a business need. An example of the former is “The results of our supplier audits can be used to predict which of them are most likely to be noncompliant in the next audit cycle.” An example of the latter could be “We can use simulation modeling to understand the effect of different lot size policies on our manufacturing conversion costs.”

Once the problem and analytics approach have been identified, it is necessary to determine whether the proposed approach is feasible. Critical elements for answering this question are the availability and quality of the necessary data. This requires carefully listing all of the required data elements for the analytics work and identifying possible sources for each. A data source will have an owner, whether it resides in a corporate information system or in a spreadsheet on a personal computer. Enlisting the cooperation of a data source owner is a key to success in analytics projects, as access to data and help understanding its format and structure are essential.

[Figure 4.3](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0003) provides a way to categorize potential data sources. Along one dimension, one can think of data that are collected manually versus data that are collected by an automated process. Any process that involves a human being recording data on paper or through a form, electronic or paper, is manual data entry. Automated data collection does not require human intervention. Along the other dimension, there are data that are collected specifically in support of the analytics project at hand versus data that have been collected to support another business process but which can be used for the analytics project at hand. This last can be problematic since the specifications of the data being collected were designed to support a different objective, and there is a good chance that it will not be an exact fit for the needs of the current effort. It will likely require additional effort to augment and transform such data into a form fitting the current objective.

A diagram of a company

Description automatically generated with medium confidence

[**Figure 4.3**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#_backF3) Data sources.

Surveys, audits, and inspections are all examples of methods requiring manual data collection that are designed to investigate specific questions using analytics. Surveys use statistical methods to identify a representative sample of a target population to measure their response to a set of research questions. Audits measure compliance to a set of standards, usually along several dimensions. Inspections entail a point-by-point examination of specific operating criteria, usually with a binary (pass/fail) outcome. For each of these, people are directly involved in the gathering and recording of the data, and thus there are opportunities for errors to occur in the process. These can take the form of simple keystroke errors, known as “fat fingering,” or the failure to correctly record a response or observation. Other concerns relate to the effect of human judgment on the data collection process. Survey respondents may not answer truthfully because of a reluctance to express an unpopular viewpoint. Different auditors may evaluate the same situation as having different levels of compliance. Different inspectors may employ different thresholds as the standard for a pass/fail recommendation, or even fail to complete the entire inspection process. It is essential that an analytics practitioner be mindful of these potential sources of trouble when using such data to analyze the populations on which these tools are used.

There are transaction-oriented systems that rely upon manual data entry as the means by which data are digitized. Manual processes are often used to create and process invoices, creating records of customer, product, pricing, and shipment information. Such records are initially created for accounting purposes as a financial record of the transaction, but the same data can be used for other analysis such as sales forecasting and production planning. An industrial equipment manufacturer had a system that relied on data entered manually by technicians at their dealer sites to create warranty claims. The system had an electronic form that needed to be completed with data such as the product serial number, time of failure, the parts replaced, and a failure code to classify the nature of the claim. While initially used as a way to receive and process warranty claims so that dealers can be reimbursed for warranty service, over time this system creates a history of warranty claims that can be used by other analyses such as predictive maintenance, root cause analysis, and fraud detection.

An anecdote from the last example highlights the importance of a thoughtful design when creating tools for manual data entry. The field in the form that requested a code to classify the specific failure mode was free text, rather than a pull-down list that provided specific choices. Often the technicians would be in a hurry, or would not have the right code handy, and would enter an invalid code “99” just to complete the form and get the claim submitted. They would write short details in a free text field to describe the work performed. While in most cases this was enough information to get the claim reimbursement, it created a history of warranty claims that did not have the correct failure mode associated with many of the records. This made the data almost unusable for deeper analysis without someone trying to manually review the free text fields to recode the problem records, a task that proved impractical from both time and accuracy perspectives.

Some basic guidelines for design of forms for manual data collection can alleviate some of the data quality risks inherent in this method of data collection:

* Automate the workflow as much as possible, eliminating intermediate steps that use paper or spreadsheets. The use of tablets or other mobile devices for data collection in the field will improve both the accuracy and completeness of data.
* Limit the use of text fields.
* Use pull-down lists, radio boxes, and check boxes wherever possible instead of text fields to limit the potential for errors.
* Make clear which fields are required for the form to be submitted.
* Be sure that required data formats (e.g., dates, currency) are clearly indicated on the form.
* Validate the data and correct errors before allowing the form to be submitted.

This list is not exhaustive and there are many resources available on the Internet to assist in designing forms for data collection. It is important that an analytics practitioner be mindful of these issues when designing tools or applications for these types of applications.

A common lament from many companies is that they are drowning in data, but do not know how to extract value from all the data they possess. This can be attributed in part to factors such as the low cost of data storage, and the proliferation of inexpensive sensors that can be used to collect data from equipment at intervals as small as a fraction of a second. In many industries, the collection of process data from sensors and other systems such as SCADA (Supervisory Control and Data Acquisition) is a prevalent form of automated data collection that is used to monitor and control processes, as well for other analytics-driven applications such as predictive maintenance. This type of data is the life blood of the Internet of Things (IoT), as the automation of the data collection process enables the automation of monitoring and control processes. This type of data consists of high-frequency time series, typically measurements of process parameters such as pressure, temperature, or velocity. It is usually captured and stored in a *data historian*, such as AspenTech's InfoPlus.21 or OSISoft's PI, that is designed for the efficient storage and retrieval of time series data. Examples of this type of automated data collection can be found in a variety of industries. According to *Aviation Week* (<http://aviationweek.com/connected-aerospace/internet-aircraft-things-industry-set-be-transformed>), there are now jet engines that have over 5000 sensors, and produce over10 GB of data a second. Industrial equipment manufacturers that serve the mining industry have developed sensors and software that collect operational data from shovels and haul trucks. Utilities collect process data from power generation equipment such as turbines. And processes throughout the oil and gas industry are closely monitored using automated data collection, from upstream drilling and extraction through refining and downstream chemical production. Value can be extracted from this enormous volume of data, but not without considerable effort in the data preparation step of the analytics process.

Other types of automated data collection occur in systems that are used for transaction management purposes such as point of sale (POS) systems. Such systems are used to process transactions in retail operations and perform critical tasks such as invoice preparation, payment and membership discount processing, inventory management, and promotion processing. Since tools such as bar code and credit card readers are a part of the system, the need for manual data entry is nearly eliminated and errors are minimized. The resulting sales records, which contain information about both customers and products, are captured in a data base that can be used for a variety of analytics applications. These include customer segmentation to allow targeted promotions, supply chain segmentation to enable segment-specific strategies such as make to order (MTO) and make to stock (MTS), and forecasting and demand sensing to allow in-season adjustments to production quantities and inventory placement.

Another important data source to be considered is called third party data. These are data provided by an external source that will collect, cleanse, and transform data into usable form, typically on a subscription basis. This includes industry-focused companies such as S&P Global Platts, which provides energy and commodities information, including pricing. Experian is known as a credit reporting corporation with a global footprint. The U.S. Bureau of Labor Statistics compiles a variety of price and production indices, which range from the aggregate level to the industry or commodity specific such as construction, natural gas, and electric utilities. These indices are time series, and are based on good and services specific to the sector that they measure. They are often available in a seasonally adjusted form, with seasonal variation removed, making it easier for analytics models to identify correlations between different time series. This is especially useful when an index is a leading indicator, giving it predictive power for other related time series.

4.3 Data Preparation

Referring to [Figure 4.1](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0001), one can see that the next step in the analytics process is data preparation. This step can be divided into two parts: data cleansing and data transformation. The objective of the data preparation step is to collect the data that have been gathered from various sources into a single location, and transform it into a form that can be consumed by analytical tools and software. [Figure 4.4](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0004) shows the flow of data through the process, and the important activities conducted at each step.

A diagram of data processing

Description automatically generated

[**Figure 4.4**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#_backF4) Data preparation.

Data profiling involves a univariate analysis of each of the variables in a data source, as well as a record-by-record evaluation of the completeness of the data. This is to allow the analyst to evaluate the suitability of the data for the project at hand. For quantitative data, this analysis will involve plotting the distribution of the variable, and identifying measures of central tendency such as the mean and median, as well as measure of dispersion including maximum, minimum, range, variance, and skewness. In addition to providing a sense of the overall shape of the data, this analysis provides insight about the possible probability distributions that may apply to the data, including whether the assumption of normality is warranted. In addition, profiling can help with the identification of missing values, extreme values, or problems with scaling.

Data profiling of qualitative data will involve the creation of frequency histograms to confirm that the data values are valid and complete. This aids in the identification of common data errors, such as missing or inconsistent values, or problems such as high cardinality of values in a categorical variable. Most commercially available analytics software has standard routines that are available for data profiling. These can greatly reduce the amount of effort involved in the process. However, this task can be completed using the tools that are available in a typical spreadsheet program, although the time and effort required to do so will be much greater than using a tool designed specifically for that purpose. No matter what the tool employed, a key output of data profiling is a list of data issues that need to be remediated to proceed with the analysis. What follows is a discussion of some common data problems that need to be addressed at this stage of the process.

*Missing values* are endemic to many data sources, and they can occur in a variety of ways often as the result of human involvement in the data collection process. Operational data collected in the field are particularly prone to missing values. Technicians may neglect to enter key information in a form such as identification codes for assets that they are inspecting simply because they cannot see through obstructions such as vegetation. Survey data may have missing values. It is typical that high-income respondents are reluctant to answer questions about their income level and may not respond to them. Whatever the source of the missing values, the critical question to answer is whether the missing values affect the representativeness of the data relative to the population from which it comes. If the sample size is large and the number of missing values is few, the missing values can be discarded without altering the results of the analysis. However, if the number of missing values is large, or if the incidence of the missing values is due to some systemic cause as described in the second example above, it will be necessary to attempt the estimation of the missing values.

In cases where data are gathered from operational systems, it may be necessary to pull data together from multiple sources to create records for analysis. For example, in the case described above, suppose there are technicians performing inspections in the field, and some of the data elements in the inspection form are missing. If there is a master list of assets, it may be possible to fill gaps such as missing identification codes by comparing timestamps from inspection records, and ascertaining the location of the crew at the time the inspection was performed. Comparing this type of data with the geographic coordinate data contained in the master list of assets may make it possible to identify the assets for which the identification codes are missing. In practice, this type of forensic approach to missing value correction is quite common, although it is labor-intensive and usually requires the assistance of someone who has a profound understanding of the data.

When there is a sensitivity to responding to certain questions, perhaps about topics such as income or politics, survey data may contain missing values that indicate response bias, called not missing at random (NMAR). One way to identify this is to create a new binary variable coded as response/no response, and to compare mean values of response variables between the two groups. If there are significant differences, this indicates a nonresponse bias and one should be careful about making inferences using such data.

In other situations, there will not be significant differences between the response and no response group for variables of interest. In this case, the data are missing at random (MAR). One can proceed with the analysis without fear of nonresponse bias, although there will be smaller sample sizes for the questions where there are no responses. If we repeat the comparison of means for all response variables and find no significant differences between the response and no response groups, then we have the best outcome and the data are said to be missing completely at random (MCAR).

There are two approaches often employed in situations where there are missing data due to no response on survey instruments. Pairwise deletion occurs when the responses to each question are summarized individually and the missing value is just excluded from the analysis. Similarly, one can perform correlation analysis on such data, but with a smaller sample size due to the missing values. List-wise deletion is used when using tools such as multiple regression or classification models. Since these methods seek to determine the relative influence of each of a group of predictor variables, any missing value requires that the entire record be excluded from the analysis.

Other techniques for handing missing values fall into the category of imputation. This is the substitution of some value for the missing values using mathematical methods of estimation. The simplest is to use the mean value of all observations for a missing value. This has the desirable property of not changing the mean of the variable, although it will dilute the correlation between that variable and any other. There are many other methods available for imputation of missing values. The reader is advised that many of them are quite advanced and will require some experience and skill to properly implement.

Another common problem with data comes in the form of *nonstandard values*. This happens when the same categorical data value is represented in the data with more than one set of characters. For example, the category “not applicable” may be represented as NA, N/A, n/a, N\_A, and so on. This often results where there is manual data entry and nothing in the data entry process enforces the standardization of the response. This is common with abbreviation as well, including those for state and country names. This can also happen when sharing data between countries. A recent example involved data being shared between different groups within a multinational company. Certain special characters such as the @ and % that appear to be the same character have different values in Chinese and English character sets, creating instances of variables that looked the same but were different. While nonstandard values appearing in data sets are very common, it is a straightforward issue to fix. A frequency histogram of all observed values makes it easy to identify such cases, and a best practice is to automate the replacement of nonstandard values as part of an ongoing ETL process.

A related issue with categorical data can occur when a variable has a high cardinality. This occurs with data such as zip codes where the number of unique values is very high. These could also be variables such as e-mail addresses, user names, and social security numbers. The high number of unique values makes these variables impossible to use in tools such as linear models as such variables will not have enough observations per level of the variable to create a model. In practice, this type of data will either be discarded or transformed into a new variable using a technique called binning. For example, a long list of phone numbers may be mapped to a new variable made of just the area code. Zip codes could be mapped to a new variable called region, made up of a small number of geographical areas in the country. Binning is one example of data transformations that will be discussed in a later section of the chapter.

When dealing with quantitative data, the problem of outliers will often occur. An outlier is said to occur when a value is observed for a quantitative variable that is more than three standard deviations away from its mean. Outliers happen for many reasons, and understanding the reason for their occurrence is essential to knowing the proper remedy. Often an outlier is due to a mistake or malfunction. For example, heavy mining trucks have many sensors that monitor critical systems. These track important operating parameters such as the temperature and pressure of oil, fuel, and cooling systems, as well as tire pressure and payload, at intervals of a fraction of a second. However, the normal operating condition of a mining truck is to be carrying hundreds of tons of payload over rough terrain, or to have 40 ton loads dropped into the bed of the truck while loading, often in extremes of altitude and temperature. Under such harsh conditions, sensors can malfunction, as can the software used to collect the data. Data collected from operating assets will often have extreme values that are known to be erroneous. In such environments, observations with extreme values are discarded.

If the outliers in a data set are recurring and predictable, discarding those observations can mean a loss of valuable information that can cause bias in statistical models. However, many statistical modeling techniques are sensitive to extreme values. In such cases, an appropriate variable translation may be necessary to keep the predictive power of the variable reducing the influence of the outliers. An example of such a transformation is the creation of a new variable that is the base 10 logarithm of the observed variable, often after adding a constant to eliminate zeros and negative values.

Another problematic characteristic of some data is that it can be very *noisy*. This occurs when there is a high level of random variability in the data. The data collected from mining trucks described above is an example of data that are noisy. The shocks and vibration endured by the equipment result in data that have a high variance, obscuring the directional changes in the key operating parameters that may be occurring. Another example of data that is noisy is the intra-day price for a stock. Over the course of the day, there may be substantial variation in the price of individual transactions. This obscures directional changes that are of interest. High variability in time series data can be handled by transformations such as smoothing. One approach is to create a new variable that is based on a rolling moving average of a fixed number of the observed values. Since the high and low extreme values are effectively negated by the averaging, the result will be a new time series that has lower variability that will more transparently reveal any trends in the overall level of the time series. Another approach is to reduce the frequency of the time series by taking the minimum, maximum, and average of the time series over fixed intervals such as an hour. This creates three new time series that can be used to monitor not only the average value but also the range and variability.

Another issue with data that may require transformation is *skewness*. This happens when the distribution of continuous data has a long tail on one side, often because the values of the variable are bounded by zero on one side, leading to a long upper tail in the distribution. Such distributions are not consistent with the assumption of normality that is required for many parametric methods, and a transformation is used to mitigate this difficulty. Such transformations involve using a function that will impact the long upper tail the most. Examples of these transformations include logarithmic (ln(*x*), log10(*x*)), inverse (1/*x*), and square root (sqr(*x*)).

Data sets with multiple observations taken on the same population may experience high correlation between variables. While this correlation can be informative and a useful output of the data profiling process, one is advised to be cautious about using correlated variables in predictive models. These collinear variables contain redundant information, and can make it difficult to estimate the parameters of the models. Often this is due to a hierarchical relationship among variables, such as supplier name and source country. The analyst is encouraged to limit the inclusion of all but the most descriptive variables when modeling.

In the previous discussion, the methods for data transformation that have been discussed included *binning*, *smoothing*, and *fitting.* Binning divides the values of a continuous variable into intervals. Binnig discretizes the data, turning quantitative data into categorical data. Most statistical software will have a capability to create new categorical variable from bins using any number of methods such as assigning an equal number of observations to each bin, or creating bins of equal width and assigning a record to a bin if its value falls within the defined interval. Binnig can also be done based upon prior knowledge of the data.

There are mixed views about the use of binning. It does involve the loss of information, and the use of too few bins can hide information such as a multiple modality in the continuous data. However, it does have advantages. Binning reduces the influence of outliers on the model by converting them to a level of a categorical variable. It can also help with the interpretation of the coefficients of the final predictive models as it has the effect of scaling variables that are of different magnitudes. It can also increase the number of degrees of freedom of a model.

Smoothing is a technique most often used to reduce the volatility of a time series. As already mentioned, simple multi-period moving averages can be effective for reducing volatility. In this approach, a new time series is created from the old one by taking successive averages of a fixed number of periods. For each new point in the new time series, the oldest observation in the previous average is dropped and the next one in the series is added to the calculation. Another popular method for smoothing a time series is called Loess Regression. This is a nonparametric method that performs least-squares regression on a local neighborhood of the time series. The new time series is predicted within a specified range, or span, and may include other predictor variables. The result is a new time series with a smoothness that increases with the width of the span, although this does not minimize the sum of squared errors of the Loess Regression. This functionality is available in commercial and open-source tools.

Several approaches for transforming variables by fitting functions are mentioned above. Another technique of interest for transforming data is *normalization*. Not to be confused with normalization in a database context, this refers to scaling data to eliminate differences of magnitude between continuous variables that can create numerical issues solving for model coefficients, as well as difficulties in comparing and interpreting the estimated coefficients. Typical methods include the following:

* Min–max: The value is scaled by subtracting the minimum from the value, and dividing by the range (max–min) of the observed values.
* *Z*-score: The value is scaled by subtracting the mean from the value, and dividing by the standard deviation of the observed values.
* Decimal scaling: The value is divided by some power of 10, to adjust the range of the observed values.

Another transformation that can be used to discretize time series data is to count the number of events that occur within a specific time interval. For example, suppose data are collected from a diesel engine. A sensor collects data for the temperature of the engine coolant at regular intervals. The engine has a protection system that causes the engine to be derated (the power reduced) when the coolant temperature exceeds 225°F. The raw time series will be noisy and difficult to use. One approach is to consider an event of interest to occur whenever the temperature exceeds this threshold. A new variable can be created that will count the number of these events that occur within a specified interval. The transformed variable can now be used to examine the relationship between these events, which may be transitory in nature, and the occurrence of other events such as unplanned maintenance. This is a valuable transformation as the collection of events and alarms is quite common in asset monitoring systems.

A final consideration on the topic of data transformation is *data reduction*. With the advent of inexpensive data storage and inexpensive devices that can collect data at high frequencies, it is not uncommon for data warehouses to become quite large. Analyses that run using data sets with terabytes of data can become impractical due to the processing time required. Data reduction seeks to reduce the size of the data warehouse while preserving the information contained in the data. There are many techniques used to perform data reduction. This discussion will focus on two examples.

The first is called *principal components analysis* (PCA). PCA finds new variables, called components, that represent the data in a lower dimensional space. PCA reduces the dimensions by an orthogonal transformation of the data that is achieved through the following process:

* Start with a data matrix of *m* observations of *n* variables.
* Subtract the mean of each variable from each observation.
* Calculate the *n* × *n* covariance matrix.
* Calculate the eigenvalues and eigenvectors of the covariance matrix.
* The principal component is the eigenvector with the largest eigenvalue.
* Select some subset of *p* eigenvectors with the *p* largest eigenvalues.
* Derive the new data by creating a matrix of *p* eigenvectors and transposing it. Multiply this by the mean adjusted to complete the transformation.

If the correlation between the original *n* variables is high, the difference between *n* and *p* will be significant and there will be substantial reduction in the size of the data. These new data can be used for model development in a fashion like the original data. However, much of the redundancy and unimportant information is removed by the projection into the lower dimensional space.

Another commonly used technique for data reduction is data sampling. There are a variety of sampling methods that can be used to reduce the number of instances submitted to an algorithm while retaining the original characteristics of the data. Simple random sampling without replacement (SRSWOR) is used to select *n* records from a set of *m* records, where *n* < *m* and every record has an equal probability of being selected. Simple random sampling with replacement (SRSWR) is similar, except that each record that is selected is replaced and may be selected again on the next draw. If the population from which the sample is not homogeneous, then a stratified sample may be taken. Suppose that a sample of individuals consists of three groups or strata: youth, adults, and seniors. A simple random sample (SRS) may be taken from each stratum to accurately reflect the data of the entire population. One challenge with using SRS methods for data reduction is that while they do reduce the size of the data, which will improve computational performance and memory usage, they also increase the sample variance. This will make it more difficult to detect small differences between groups, and will generally reduce the effectiveness of statistical algorithms. More complex algorithms are available for data reduction, also sometimes called *data squashing*. These methods select *n* records from a set of *m* records, where *n* is much smaller than *m*, and add an additional column that contains a weight that is representative of the frequency of occurrence of that record in the original population. Numerous references to data squashing methods and their application and effectiveness can be found in the statistical literature.

4.4 Data Modeling

4.4.1 Relational Databases

After the data have been cleaned and transformed, the ETL process will deposit them into a data warehouse. The most common type of data warehouse is built using a *relational database*. The software underlying the structure of relational databases is called a relational database management system (RDBMS). Originally proposed by an IBM researcher named E.F. Codd in 1970, a relational database stores data in tables. Each table consists of rows called records that usually represent one entry of the content of the table. For example, each record can contain information about a customer, an asset, or a purchase order. Each record consists of columns or fields that contain data related to that instance. So, a customer record might contain account number, first name, last name, phone number, and e-mail address.

[Figure 4.5](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#c04-fig-0005) illustrates the critical concept in a relational database. Each table will have a *primary key* that serves as a unique identifier for that record in the table. In this example, Asset Type, Asset ID, Work Order Number, and Task Code all serve as a primary key in a table. When they appear in other tables, they are called *foreign keys*. The relationships between the tables are highlighted by the connections. For example, in the Assets table, Asset ID is the primary key as each asset will have a unique Asset ID. In the Work Orders table, Asset ID is a foreign key and the relationship between the two keys is said to be *one to many*. The Asset ID may appear many times in the Work Orders table as the asset may have been serviced many times. However, the information uniquely related to the asset appears just once in the Asset table. This allows the relational database to store the data in a more compact form, and master data such as we see here regarding Assets, Asset Types, and Tasks needs to be retrieved only if we have a need to associate it with transactional data such as we see in the Work Orders table.

A diagram of a work order

Description automatically generated

[**Figure 4.5**](https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/reader/content/17facba3b5e/10.1002/9781119505914.ch4/format/chapter-epub/OPS/c04.xhtml?hmac=1704948814-ix%2FnweEac2e9TDoBw9eVK%2F0GuleLL3gU1wkE36QK678%3D#_backF5) Relational database.

Essentially all relational databases use structured query language (SQL) to write queries and maintain the database. A query allows the user to extract data from several different tables to create a new record format specific to a required purpose, such as being processed by an algorithm. Suppose that one wanted to examine the work order history and compare the maintenance costs by manufacturer and by work crew foreman. The following example of SQL code is called a *query.* It will create a new *view* of the data that will be useful for the analysis:

SELECT

[Work Orders].[Work Order Number],

[Work Orders].Foreman,

[Work Orders].[Asset ID],

[Asset Types].[Asset Type],

[Asset Types].Manufacturer,

[Task].[Labor Cost]

FROM

[Task],

[Assets],

[Asset Types],

[Work Orders]

WHERE

[Assets].[Asset ID] = [Work Orders].[Asset ID]

AND

[Task].[Task Code] = [Work Orders].[Task Code]

AND

[Asset Types].[Asset Type] = Assets.[Asset Type]

ORDER BY

[Work Orders].[Service Date];

The query will select individual fields from records within the Work Orders, Asset Types, and Task tables, and join them together using the relationships defined between the primary and foreign keys. The SELECT portion of the query lists the fields that are to go into the new view. The FROM portion of the query lists the target tables or views from which records are to be selected. The WHERE portion of the query lists conditions that must be satisfied for selected records to be displayed in the view. The ORDER BY portion of the query allows the user to sort the new records in the view based upon one or more specified fields. Note that the fields specified in either the WHERE or ORDER BY sections need not appear in the view itself. The resulting data record will contain the following fields:

* Work Order Number
* Foreman
* Asset ID
* Asset Type
* Manufacturer
* Labor Cost

In addition, the records in the view will be sorted by the Service Date contained in the Work Orders table. A variety of operators are available in SQL that can be used to create complex views from many tables or views, giving it tremendous power for query development and data maintenance.

Another powerful feature of a relational database is the ability to enforce specific data types for each field, as well as adding constraints on the values allowed for each field. For example, it is possible to restrict a field to only integer values within an allowed range, or to require that a date/time value be in a specified format. This facilitates the enforcing of necessary business rules and prevents introduction of incorrect or erroneous data into the database. This structured approach to maintaining *data integrity* is one of the primary advantages of a relational database.

For efficiency and security, *stored procedures* are sometimes used to perform complex or frequently repeated tasks. A stored procedure is a block of code, either SQL or some other language such as Java or C++, that can be used to implement business logic. Because they are stored in the database and run on the database server, they typically run with less overhead and better security than when applications send dynamic queries to the database from outside the database server.

4.4.2 Nonrelational Databases

A nonrelational database, also sometimes called a NoSQL database, is any database that does not rely on the tabular structures and primary and foreign key relationships supported by a traditional RDBMS. These databases have become popular in recent years as part of the so-called *big data* explosion that has been driven in part by the sheer volumes of data that an Internet-connected world can create. But it has also been driven by the fact that these data are also more unstructured as social media-driven content has proliferated.

Interview With Robert Clark

*When asked his work with* big data, *RTI International Senior Research Biologist Robert Clark provided the following example:*

On the LungMAP project, we are considering the normal development of the human lung. On that project, we are looking at imaging and we are looking at genomics, transcriptomics, proteomics, metabolomics, lipidomics–and all have very large data sets. The human cell has 3.3 billion base pairs and 20,000 genes. Then you probably have 100,000 different proteins in each of your cells, and then the RNA is probably, oh, 300,000 RNAs per cell, 12,000 different RNAs, things like that. Trying to align all that information and then analyze it separately and then together is a huge feat. We use all kinds of data analysis tools as well as imaging and machine learning to draw on and annotate images so that people understand what's on these images rather than having people go in and manually draw them. And that is just one mapping project.

On the LungMAP project, we are currently storing everything in the cloud using tools developed for use in the cloud. But, for example, we had a great deal of 3D image data that came from one of our research centers–It was 150 Terabytes of data, and it took a whole day to download it on a special drive within the Amazon cloud that we're using. Then it took another entire day to retrieve it from the cloud so that we could examine it. And that's just a first few steps in a complex analytics project.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

Nonrelational databases encompass a variety of different technologies, but tend to share some characteristics. Since they do not rely upon a relational model, no predefined schema must be constructed before data can be loaded into them. In addition to having a flexible schema, they can handle unstructured data that do not fit into the tabular structures of the relational model. Most of them *scale horizontally*, meaning they can be increased in size by adding additional clusters of inexpensive, commodity servers. And with a few exceptions, they follow an open-source model and do not require expensive license and maintenance fees to get started.

*Columnar* databases such as Redshift and Cassandra organize data by columns, not rows as in the traditional model. Queries are still processed using traditional SQL, but for many applications the efficiency is greater as the input/output process is more efficient for many common types of queries used in analytics applications. This is because such queries often touch on many rows but only a few columns in the data; and in the row-oriented structure of a traditional relational database, this means scanning across each row to retrieve the required columns. In a columnar database, only the relevant columns need to be scanned for increasing performance.

A *key-value store* is a database without a schema that stores all the data in a single blob. Each value of data can have a different form, and will have a unique key identified with it. The fundamental structure of the key-value store is called an associative array that contains what are called key-value pairs. This approach does not allow the processing of complex queries using SQL. The only way to retrieve data is using the unique key associated with it. Since this means a direct request to the data object in memory or on disk, it will be very fast. However, since the operations typically performed with SQL such as joins are not available in the database, they will need to be done in the code calling the database. A key reason to use a key-value store is scalability as the simple architecture makes this easy to do. Aerospike and Cassandra are examples of products that offer key-value stores.

A *document-oriented* database is designed to store semistructured data, typically using Java Script Object Notation (JSON) or XML. MongoDB is an example of this type of database, which can be thought of as a subset of a key-value store. This is because document-oriented databases use a key to document look-up like the key-value relationship. The difference is that while the characteristics of the data in a key-value store are not visible, a document-oriented database will typically have an API for developing queries based upon the structure of the documents.

A *graph* database is used to map complex relationships between objects such as people, things, and locations. In a graph database, objects are stored as vertices and directed edges. For example, the vertices may represent people and activities, and the directed edges may represent relationships such as “friend” and “likes.” Graph databases are useful for analyzing data where there are complex relationships between entities. Examples of this can be found in customer relationship management applications, such as identifying product bundles to suggest to a customer. Another example is market segmentation, where one might seek to identify interests of customers that are strongly related to preferences for certain products. Titan is an example of a graph database.

4.5 Data Management

Any company or organization that has physical assets or human resources is likely to have processes in place to ensure that those are maintained and protected, recognizing the value they create. An organization's data are another asset that should be actively managed and maintained just like any other asset. In recent years, most organizations have acquired data at rates far exceeding those at which they acquire new assets or employees. To ensure that their data are reliable, usable, and available to create value for the consumers of data within the organization, it needs to have a thoughtful and systematic approach to data management. The literature about data management and data governance is extensive, and has been the subject of entire volumes. It is useful to highlight some concerns that a successful data management strategy will address.

As described earlier in this chapter, a great deal of effort can be expended to clean data when gathered from its original source. Therefore, it is important that methods for *data capture* are implemented to prevent the entering of incomplete or incorrect data into the system. This should also incorporate the necessary business rules to ensure that the data collected matches the requirements for the business processes it is intended to support.

Effective data management programs will create roles for *data stewards* who are responsible for monitoring the quality of specific elements of data on an ongoing basis. Data stewards may use dashboards and reports combined with their unique understanding what constitutes data quality for their data, and will not allow changes to be made to the data structure without approval from the appropriate governance body within the organization. Data stewards make sure that usable data stay usable.

*Metadata* is often called “data about data.” It serves as the documentation for the data and can contain information such as when and by whom the data were created, as well as the structural elements of the data such as tables, fields, and relationships. It can also contain information about who has access to the data and at what security level. Metadata can also describe the various elements of the data in relation to the business processes that generated it so that the user can have a practical understanding to complement the technical description. Collection and maintenance of metadata is an important data management function, especially if the data are to be archived, or retained and published and made available to other users within the organization. This is especially the case as organizations seek to build large data platforms to support self-service analytics on demand across the entire enterprise.

*Master data* consists of the key objects necessary to describe and run a business. These are typically lists of people, places, or things core to the business such as customers, regions, and products. They may be scattered across multiple systems in the organization. *Master data management* (MDM) is a process by which a single view of these core objects can be presented to users across the entire enterprise in a consistent and current format. Each of these objects has a unique life cycle, sometimes called the CRUD cycle (create, read, update, and delete), that must be managed according to defined procedures for each step in the objects life cycle. Effective MDM will require frequent updates to the master data store, and will often rely upon specialized applications such as customer data integration (CDI) or product information management (PIM) tools.