Table 1.1 Comparison of data-centric and decision-centric approaches.

| | Data-centric analysis (Data science, computer science) | Decision-centric analysis (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<br>• Potential to extract weak signals from large, unstructured data sets | • Causal focus<br>• Strategic value beyond historical observations |
| Cons | • Risk of conflating correlation with causation<br>• Analysis inferences are limited by history<br>• Noisy data with confounded effects | • Human subject matter expertise required<br>• Cost of data acquisition can be high |
| Key disciplines | • Computer science<br>• Data science<br>• Machine learning and unstructured data mining<br>• Artificial intelligence (AI), deep learning | • Management and decision sciences<br>• Operations research<br>• Mathematics<br>• Classical statistics |
| Example applications | • Image classification<br>• Speech recognition<br>• Autonomous vehicle scene recognition | • Supply chain optimization<br>• Scenario planning<br>• New business model development |

## 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:



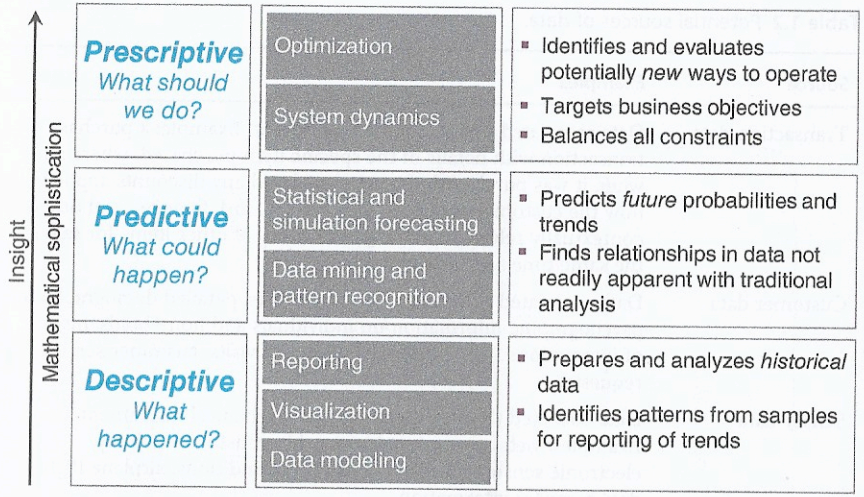


Figure 1.4 Categories of analytics.

*descriptive*, *predictive*, and *prescriptive*. As summarized in Figure 1.4, 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.

### 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

Table 1.2 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 |

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 for examples), as can the size and information density of any given data set (see Figure 1.5). 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.

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



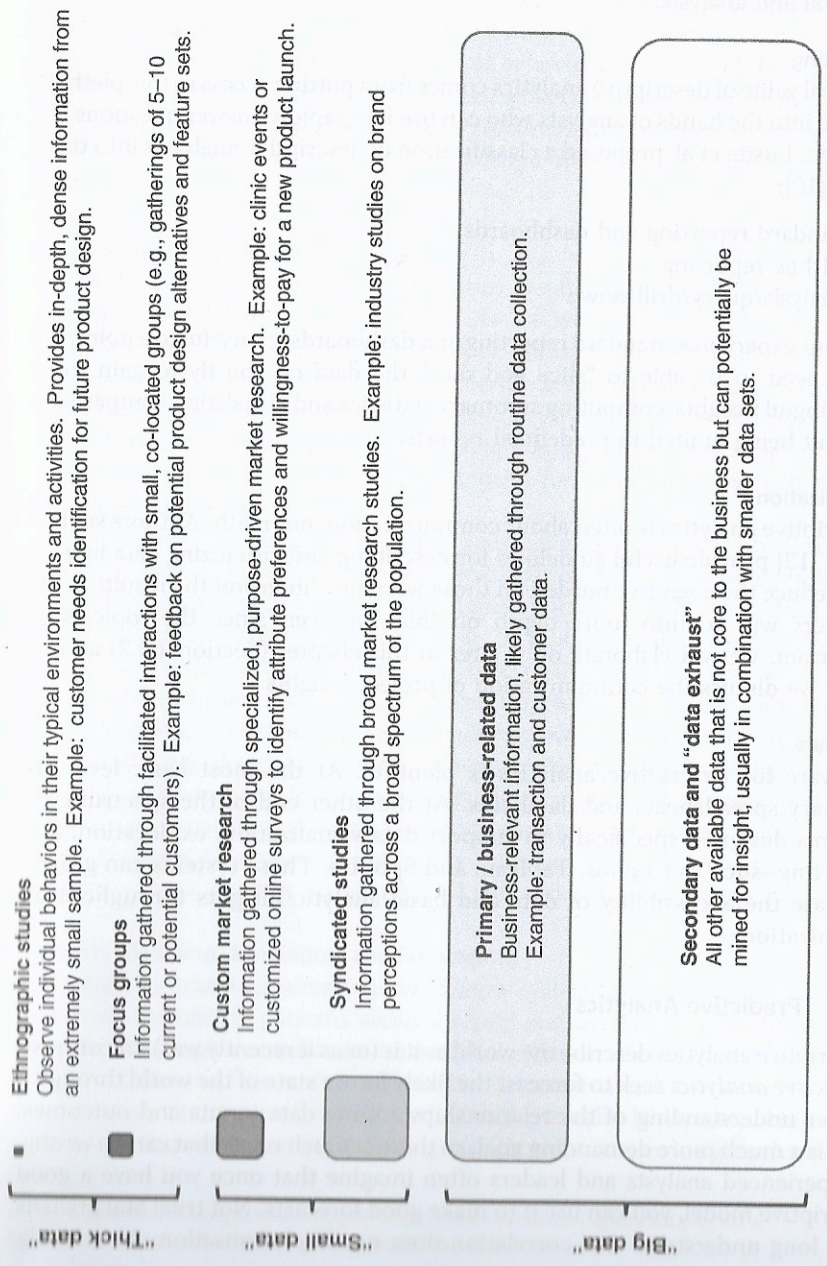


Figure 1.5 Illustration of variability in the size and information density of different data sets.

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]

#### 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!

1 The reader is encouraged to see https://imgs.xkcd.com/comics/significant.png for a lighthearted cartoon illustrating the dangers of false significance.

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 summarizes key differences between three common types of simulation models: discrete event, agent-based, and system dynamics.

**Table 1.3** 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. 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.

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

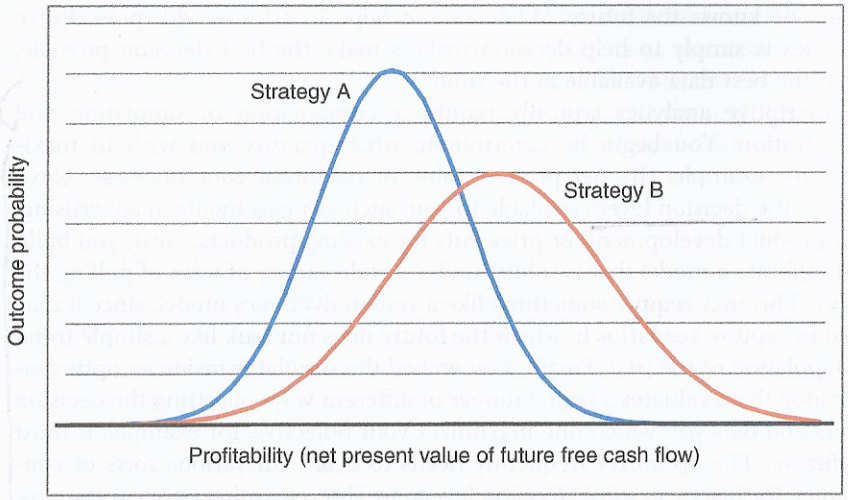


**Figure 1.6** Output from an example prescriptive analysis of alternative policies [13].

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