Modern Analytics
Methodologies
This page intentionally left blank
Modern Analytics Methodologies

Driving Business Value with Analytics

Michele Chambers
Thomas W. Dinsmore
To my son, Cole, may you help make the world a better place with your math, science, and technology talent. To my mother, who taught me how to be graceful and loving. To my father, who passed his math gene on to me and taught me that there are no limits in life other than those you impose on yourself. To my adopted family, Lisa, Pei Yee, Patrick, Jenny, and Angel, thank you for your love and support.

To the heroes on the front line and those behind the scenes who are working toward eradicating slavery from the face of the earth—may analytic insights help in some small way to achieve this quest in your lifetime.

—Michele

To my wife, Ann; my two sons, Thomas and Michael; my late nephew Jeffrey Thomas Dinsmore; my father, Ralph Boone Dinsmore; and to my grandfather E.W. Egee Jr., who loved new technology.

—Thomas
This page intentionally left blank
## Contents

### Section I  Why You Need a Unique Analytics Roadmap  
**Chapter 1** Principles of Modern Analytics  
- Deliver Business Value and Impact  
- Focus on the Last Mile  
- Leverage Kaizen  
- Accelerate Learning and Execution  
- Differentiate Your Analytics  
- Embed Analytics  
- Establish Modern Analytics Architecture  
- Build on Human Factors  
- Capitalize on Consumerization  
Summary  
**Chapter 2** Business 3.0 Is Here Now  
**Chapter 3** Why You Need a Unique Analytics Roadmap  
- Overview  
- Business Area  
- Data  
- Approach  
- Precision  
- Algorithms  
- Embedding  
- Speed  
Summary  
### Section II  The Analytics Roadmap  
**Chapter 4** Analytics Can Supercharge Your Business Strategy  
- Overview  
- Case Studies  
Summary
### Chapter 5 Building Your Analytics Roadmap

- **Overview** ........................................... 57
- **Step 1: Identify Key Business Objectives** ........ 57
- **Step 2: Define Your Value Chain** .................. 58
- **Step 3: Brainstorm Analytic Solution Opportunities** 60
- **Step 4: Describe Analytic Solution Opportunities** 66
- **Step 5: Create Decision Model** .................... 70
- **Step 6: Evaluate Analytic Solution Opportunities** 73
- **Step 7: Establish Analytics Roadmap** ............. 81
- **Step 8: Evolve Your Analytics Roadmap** .......... 84
- **Summary** ........................................... 84

### Chapter 6 Analytic Applications

- **Overview** .......................................... 87
- **Strategic Analytics** ................................ 88
- **Managerial Analytics** ............................. 94
- **Operational Analytics** ............................. 96
- **Scientific Analytics** ............................... 99
- **Customer-Facing Analytics** ....................... 100
- **Summary** .......................................... 103

### Chapter 7 Analytic Use Cases

- **Overview** .......................................... 105
- **Prediction** ........................................ 108
- **Explanation** ....................................... 112
- **Forecasting** ....................................... 113
- **Discovery** .......................................... 114
- **Simulation** ......................................... 120
- **Optimization** ...................................... 121
- **Summary** .......................................... 121

### Chapter 8 Predictive Analytics Methodology

- **Overview: The Modern Analytics Approach** ...... 123
- **Define Business Needs** ............................. 126
- **Build the Analysis Data Set** ....................... 133
- **Build the Predictive Model** ....................... 139
- **Deploy the Predictive Model** ...................... 147
- **Summary** .......................................... 152
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>End-User Analytics</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>User Personas</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>Analytic Programming Languages</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Business User Tools</td>
<td>167</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>171</td>
</tr>
<tr>
<td>10</td>
<td>Analytic Platforms</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Predictive Analytics Architecture</td>
<td>174</td>
</tr>
<tr>
<td></td>
<td>Modern SQL Platforms</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>202</td>
</tr>
<tr>
<td>III</td>
<td>Implement Your Analytics Roadmap</td>
<td>203</td>
</tr>
<tr>
<td>11</td>
<td>Attracting and Retaining Analytics Talent</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>Culture</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>Data Scientist Role</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>232</td>
</tr>
<tr>
<td>12</td>
<td>Organizing Analytics Teams</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Centralized versus Decentralized Analytics Team</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Center of Excellence</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>Chief Data Officer versus Chief Analytics Officer</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>Lab Team</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Analytic Program Office</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>242</td>
</tr>
<tr>
<td>13</td>
<td>What Are You Waiting For? Go Get Started!</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>247</td>
</tr>
</tbody>
</table>
Foreword

In the Information Age, those who control the data control the future. I’ve invested my career in helping develop and bring to market the technologies that help make sense of data. While serving alongside Michele Chambers and Thomas Dinsmore as an executive at Netezza, the firm that invented and launched the world’s first data warehouse appliance, I met thousands of business people around the world and heard stories of the challenges they faced turning information into insight. Most of all, I saw an overwhelming demand for practical guidance on how to best take advantage of the wealth of new analytic technologies that were emerging to help organizations make better sense of data. The pace of innovation in analytic technologies makes best practices a moving target, and keeping up with them is a huge challenge. As the executives leading initiatives involving our firm’s most advanced analytical capabilities, Michele and Thomas developed deep insights into what it took for firms to fully utilize the most sophisticated analytic technologies in the market. Because of this experience, there are no better people than Michele and Thomas to offer this timely roadmap on Modern Analytic Methodologies.

The world around us has become increasingly digital. An ever-expanding collection of digital devices—computers, mobile phones, IPTVs, smart homes, connected appliances, smart hospitals, smart utilities, and more—are creating an explosion of data as we interact with them. This “digital exhaust” produces so much data that yesterday’s analytic technologies often fall short. Fortunately, innovation in analytic technologies has accelerated to keep pace with the data deluge. Hadoop, NoSQL, MPP (massively parallel processing) databases, in-memory databases, streaming/CEP (complex event processing) engines, and more—these modern platforms are fully capable of capturing relevant information about nearly every digital interaction occurring anywhere in the world.
Importantly, extracting insights from this growing data volume requires not only new technologies, but also new methodologies. There is no one-size-fits-all approach for analytic architecture, and it follows that business processes and organizational structures must be tailored to support each firm’s unique technical approach to data analysis. Because analytics must be linked to business strategy in order to deliver value, best practices are unique to each problem, and each firm’s path to success will be unique. What Michele and Thomas offer here is a compelling review of the patterns of success drawn from a diverse set of experiences helping firms address a variety of different business challenges with analytics. This makes Modern Analytic Methodologies an invaluable tool in helping you craft your own unique path to success.

Every individual climbs a learning curve over time with respect to his or her ability to leverage data analysis to create success. The uniform truth that I’ve seen is that individuals that climb this curve the fastest are the ones that win. They are the most highly sought after—and highly compensated—professionals in the world. The payoff is clearly worth the effort. What Michele and Thomas offer in Modern Analytic Methodologies is a roadmap for accelerating your journey to take full advantage of the state of the art in analytics.

They have rigorously tested the techniques and best practices they share across a breadth of diverse firms with different business challenges. As a result, they are able to reveal what works and what doesn’t. Their real-world experiences have been a crucible for discovering the challenges that analytics professionals face and for evaluating which solutions really work. Modern Analytic Methodologies is a battle-tested blueprint for practitioners who want to increase their odds of success.

**Brad Terrell**  
Former VP & General Manager,  
Netezza and Big Data Platforms, IBM  
Boston, MA
Acknowledgments

Imagine how hard it is to write a book, then quadruple it, and you’ll start to feel how much work it takes to write a book. We undertook this project as a labor of love for our field and to give back to others the value of our insights and knowledge. Although a book on technology is never complete because the industry is constantly evolving and morphing, we have finally approached the end for now.

Along the way, we have had the distinct pleasure of collaborating with many thought leaders and who are experts in their own rights. We’d like to thank them for their time, support, and contributions.

Thank you for your contributions:
Sujha Balaji—Philadelphia University
George Matthew—Alteryx
Greta Roberts—Talent Analytics
Les Sztandera—Philadelphia University

Thank you for sharing your experiences:
Dean Abbott—Smarter Remarketer & Abbott Analytics
Thomas Baeck, Ph.D.—Divis Intelligent Solutions
Brandy Baxter—Alteryx
Michael Forhez—CSC
Bob Gabruk—Cognizant
Rayid Ghani—EdgeFlip & University of Chicago
Kevin Kostuik—Charlotte Software Systems
Doug Laney—Gartner
Bob Muenchen—r4stats.org
Tess Nesbitt, Ph.D.—DataSong
Karl Rexer—Rexer Analytics
Greta Roberts—Talent Analytics
George Roumeliotis—Intuit
Thank you for your support:

Thank you, Jeffrey Brown with Accenture, for being a sounding board.

Thank you, Bill Jacobs, Lee Edlefsen, Neera Talbert, Rich Kittler, and Derek McCrae Norton, for your valuable review and feedback.
About the Authors

**Michele Chambers** is the Vice President of Marketing for MemSQL. Prior to this, she served as Chief Strategy Officer and Vice President of Product Management & Marketing for Revolution Analytics, General Manager and Vice President for IBM Big Data Analytics, and General Manager and Vice President for Netezza Analytics. In these roles, Michele has worked with hundreds of customers to help them understand how to use analytics and technology to achieve high-impact business value.

**Thomas W. Dinsmore** is the Director of Product Management for Revolution Analytics. Previously, he served as an Analytics Solution Architect for IBM Big Data, SAS Consulting, and PricewaterhouseCoopers. Thomas has helped more than 500 enterprises around the world use analytics more effectively. He uniquely combines hands-on skill in predictive analytics with business, organization, and technology experience.
This page intentionally left blank
Principles of Modern Analytics

There was a time, not long ago, when enterprise analytics was simple: You bought software from the leading vendor and installed it on a box. If your needs changed, you bought more software from the same vendor, and installed it on a bigger box. Analytics was a niche field populated by specialists, all of whom used the same software they learned in graduate school. People still believed that a single data warehouse could hold everything worth knowing.

The business cadence was, in retrospect, leisurely: If it took two years to implement a predictive model, well, that was just how things worked. Not that long ago, a big bank ran four campaigns per year to promote its credit card; at the time, executives thought that was an accomplishment.

Well, good-bye to all of that. Digital media is here; so are Web 2.0, mobile, cloud, and Big Data. The volume, velocity, and variety of data are exploding; enterprises are abandoning the ideal of the single data warehouse because it is impossible to stay on top of the tsunami. Diversity rules—we have a plethora of sources, an alphabet soup of platforms, and data everywhere: on premises, hosted by third parties, and in the cloud.

The changing landscape of data brings with it sweeping changes to the field of analytics: new business questions, applications, use cases, techniques, tools, and platforms. Techniques now considered mainstream were exotic five years ago. A single vendor once dominated analytic software; today, there are 851 analytic startups listed in Crunchbase, the leading source of information about startups. Open
source software continues to eat the software world: two of the four Leaders in The Gartner Group’s most recent Advanced Analytics Magic Quadrant are open source projects, and surveyed analysts prefer open source analytics to the most popular commercial software by more than two to one.

Above all, the cadence of business accelerates exponentially. Yesterday, we ran four campaigns a year; now we can run four campaigns an hour. Nobody can afford to take two years to implement a predictive model; we will be out of business by then.

We can no longer afford the luxury of the blue chip, single-vendor proprietary analytics architecture. In its place, we see enterprises building an open analytics platform based on diverse commercial and open source tools, tied together through open standards. In this new world, each organization must define a unique analytics architecture and roadmap, one that recognizes the complexity of the modern organization and business strategy. This architecture will include many vendors and open source projects because no single vendor can meet all needs.

In this book, we propose an approach based on nine core principles:

• **Deliver Business Value and Impact**—Building and continuously evolving analytics for high-value business impact

• **Focus on the Last Mile**—Deploying analytics into production to attain repeatable, ongoing business value

• **Leverage Kaizen**—Starting small and building on success

• **Accelerate Learning and Execution**—Doing, learning, adapting, and repeating

• **Differentiate Your Analytics**—Exploiting analytics to produce new results

• **Embed Analytics**—Building analytics into business processes
• **Establish Modern Analytics Architecture**—Leveraging commodity hardware and next generation technology to drive out costs

• **Build on Human Factors**—Maximizing and grooming talent

• **Capitalize on Consumerization**—Leveraging choices to innovate

Next, we fully explore each of these principles because they are the foundation upon which Modern Analytics are built.

**Deliver Business Value and Impact**

Later in the book, we describe how to go about creating a unique analytics roadmap and how to prioritize projects. For now, suffice it to say that one of the principles of Modern Analytics is a focus on analytic projects with potential for game-changing value to your organization. To hold the organization accountable for delivering value, measure your current state to establish a baseline and set initial quantifiable target business objectives and ongoing business objectives. For example, current revenue is $100 million with CAGR 4%. The initial target is to identify 15% net new revenue with an ongoing net new revenue contribution of 10% annually.

Although such a metric can be easy to identify and measure, other metrics can be harder to identify and measure. To discover these potential metrics, identify points where business decisions are typically made. Start by measuring impact at these points. Then work toward establishing metrics that have a direct impact on the business. Whereas in the past, companies typically aspired for either a revenue metric or an operational cost metric but not both, today mature analytic organizations often establish metrics on both sides of the balance sheet. This sends a clear signal to the team that revenue growth has to be accomplished cost efficiently.
Savvy organizations identify potential analytic opportunities by thinking outside the box. Typically, the hardest, most entrenched problems in an industry or company have been around so long that people start to think about them as hard-and-fast constraints for their business. However, often, the barriers that existed in the past that made them impossible to solve no longer exist. Unleashing the bottleneck typically results in massive business value creation. Analytic-driven organizations dare to think outside the box and identify some of the most challenging problems facing their industry or business. When that is done, they work toward identifying how they solve or reduce the problem through innovative data- and technology-driven approaches. This is usually accomplished with a clean sheet brainstorming approach and imagining that all the resources needed to solve the problem exist. After ideas are vetted, the team typically has another brainstorming round to determine how to get everything they need to solve the problem without settling. Instead of using samples or backward validation\(^1\) to estimate a solution, the team will identify potential new resources—data, symbiotic partnerships, or technology—that will help them achieve their business objectives.

To realize the business value both initially and over an extended period of time, you need to deploy the analytics into production. Before any analytics can be deployed, the results of the analytic model need to be validated for accuracy. Today, that typically occurs in a “sandbox” with a limited subset of the data and in an artificial, non-production environment. It is all too common for an analytic model to meet or exceed business criteria in a sandbox but significantly underperform in a production environment. Be sure to evaluate your analytic models based on the environment that they will be deployed into, not any idealistic environment. Deploy the analytic models into a replicated production environment to fully test the model prior to going live to get a realistic assessment against the target business.

\(^1\) Also commonly referred to as backtesting.
objectives. Where deploying used to be a “post” process after the model was built, deployment is now part of the full life-cycle analytic process. Once all potential technical deployment barriers are identified, obtain legal and/or procedural process validation before the “go-live” launch into production. After an analytic model is deployed, measure the initial business impact and identify quick ways to continuously improve on the results.

Focus on the Last Mile

Today, very few organizations get to the promised land of deploying analytics into production environments to drive game-changing business value for their organization. To get to this end goal, start with the end in mind and work backward. Understand day-to-day issues at every level in the organization by speaking with frontline workers—from strategy through to execution. These domain experts are acutely aware of the issues, problems, and constraints impeding their success. Clearly understand what it will take to achieve success—not how success can be attained. With this understanding, set quantifiable and ambitious goals for your analytics. For example:

- What is the target business value to be obtained?
  A 3% lift in revenue?
  Inventory saving of $10 million annually?
  Total cost savings of $100 million in the first year of deployment?

- What is the expected service level agreement (SLA) for the business?
  Reevaluated credit scores nightly?
  Portfolio evaluation within 5 minutes?

- What is the operational model?
  How does the model get moved into production?
Does this analytic model need to integrate with other business systems? If so, how do the operational processes and decisions change?

Is this analytic model triggered from another business system?

Is this analytic model deployed in one location or multiple locations?

Are there multinational or localized requirements?

What is the frequency of updates to the model?

- What are the key success factors that measure the business impact?
  - How is success measured?
  - What constitutes failure?
  - How long does the team have to achieve success?

- What is the model accuracy?
  - Is the accuracy “good enough” to realize immediate business value?
  - How much should the model be improved in what period of time?

Traditionally, one team—quants, statisticians, or data miners—has been responsible for the model creation while a second team—typically IT—has been responsible for the production deployment. Because this often crosses organizational boundaries, there can be long lags and disconnects between the model creation and the model deployment or scoring. The teams must function as if they are one team even if organizational boundaries exist and will persist. A full life-cycle methodology can serve to bring these two teams into alignment if the analytics methodology goes beyond just creating and assessing the initial analytic model to encompass the actual production deployment and ongoing reassessment of the analytic model to achieve the business objectives.
With Modern Analytics, teams focus on delivering results quickly rather than waiting to build the “perfect” analytic model. They do so by starting with Proof-of-Concepts (POCs) or prototypes that may be limited in scope, but help the organization increment toward realizing business value. They quickly mature and harden the POCs or prototypes into a production deployment where the rewards can be systematically reaped.

**Leverage Kaizen**

Kaizen, the manufacturing movement for continuous improvement, is being adapted into many different disciplines, including analytics. The core tenements of Kaizen are to

- Start small.
- Remove overly complicated work.
- Perform experiments to identify and eliminate waste.

There is an emphasis on delivering value quickly rather than completeness. Testing and learning can make small improvements along the way while working toward the end goals.

This is a marked contrast to the current tendency to spend long development cycles building the “perfect” models. Today, building and deploying analytics are complex, custom projects that cross over multiple functional areas. In this new era, modern analytic teams are removing the ivory tower academic shackles from traditional analytic methodologies to eliminate unnecessary, time-consuming steps from the project cycle. This helps increase agility and responsiveness while incorporating business feedback into the process so results can be improved.

With Kaizen serving as a guiding principle, modern analytic teams build and deploy models immediately and then improve on the
models in short burst cycles dovetailing the work of analysts and IT to create a frictionless environment that continuously delivers business value. On the ground, the teams often use hybrid agile or rapid application development methodologies to improve cycle time and reduction in barriers with cross-functional teams.

**Accelerate Learning and Execution**

Today, modern analytic teams are trying new things—experimenting with new and combined approaches, tools, visualizations, and algorithms to uncover patterns in the growing mass of data. By trying new things, experimenting and transferring lessons from one industry and problem to a completely different industry and problem, modern analytic teams have significantly accelerated their learning and are driving new business value. However, to foster this level of innovation through experimentation, there has to be a culture that tolerates and expects failure as a path to learning and improving.

As an example, as data sizes have increased, modern analytic teams have started shifting away from constrained, statistical-only approaches to predictive and machine-learning approaches that can leverage the power of all of the data. One of the key lessons learned as data has increased is that the underlying tools and infrastructure need to minimize data movement in order to meet business objectives, especially for service-level agreements. Modern analytic teams have quickly realized the transferability of this lesson to various types of analysis and have incorporated these lessons into requirements at the onset.

Today, by industry standards, 60–80% of an analyst’s development time is spent doing data preparation or data munging. Another valuable lesson that modern analytic teams have discovered is that the upfront manual data munging should be minimized, and instead, data
prep tasks should be automated and/or handled as part of the analytics processing activity. This dovetails with the need for businesses to move at a faster pace to be ahead of the competition. The ability of an organization to learn in as close to real-time as possible is a trend whose momentum will continue to build. The ability to uncover patterns in real-time, act on them almost instantaneously, and continue to discover deeper insights to improve the next cycle is a requirement in the modern business world.

**Differentiate Your Analytics**

Businesses strive to create competitive differentiation through a combination of products, customer service, and operational processes that are delivered to the market. Analytics can simply support each independent activity by delivering comparable insights as realized by competitors. Or they can be used to highly differentiate your competitive strategy. That could mean being a first mover—the first or on the leading edge of using analytics in your industry—or it could mean distinguishing your analytic approach or the speed at which you deploy your analytics into a production environment.

Many firms look around the marketplace and try to learn from their competitive landscape. However, that copycat approach typically means settling for a secondary position in the market rather than a leading position. Instead, analytic leaders look at other industries and how they’re using analytics. They draw analogies between problems from other industries and the problems in their business. They discover how other companies are using analytics to solve their problems. They start to look for new data and approaches outside the four walls of their organization. They forge new symbiotic alliances to obtain data and methods that benefit their organization. They apply the new knowledge to problems in their industry or business. To do
this, they look beyond the silo of their own team, department, or location. They look for opportunities to integrate with other data and processes to build an analytic solution with a broader impact for the organization. They remove the constraints they’ve lived with and find new ways to inspire innovative analytics. They use the full breadth and combination of available analytics—not just what they’ve always used—to drive game-changing value. They don’t just create predictions; they use their predictive model to systematically perform at their best by optimizing the predictive model to prescribe the best course of action. This relentlessly drives the best course of action—day in and day out—to help them achieve their competitive differentiation.

**Embed Analytics**

On-demand or ad hoc analytics are analytic models that are executed occasionally and provide a “point in time” insight that can be used by a human to inform decisions and take a course of action. While this approach is useful and provides value, it is slowed down by the human interaction. Think about financial traders of years gone by. Traders would run desktop tools on the trading floor to understand complex financial market interdependencies. The tools would produce a “spot”—or instance in time—view of the market, and the trader would use that to inform them as to what to buy or sell. Today, the capital markets are dominated by “algorithmic trading,” where a sophisticated program, which embodies a newer generation of the algorithms that were in the desktop tools, automatically makes trades. Eliminating human interaction and embedding analytics into the complex financial market process eliminates friction in the overall system. When the analytic models are built into the process, repeatability and scalability are achieved. This relentless execution drives incalculable business value in the marketplace.
Establish Modern Analytics Architecture

As analytics has matured over the past 20 years, analytic architectures have gone through a substantial transition from standalone desktops to enterprise data warehouses to Big Data platforms such as Hadoop. High-performance computing environments, such as clusters and grids, which were once specialty environments, are becoming mainstream environments for analytics. This has created a hodgepodge hardware and software legacy in data centers across the globe. All of this has occurred while the cost of computing power has dramatically decreased and open source software has gone mainstream.

The paradigm shift underway is a movement toward building lean analytic architectures, as illustrated in Exhibit 1.1, based on simplicity and open standards that leverage commodity hardware and open source software to drive the costs out of the architecture, provide platform scalability, and leverage the latest innovation. This innovation supports the execution of thousands of computationally and data-intensive predictive models in production deployments by large user bases with differing analytic and service-level requirements. Building, managing, and supporting the ecosystems required to deliver on these requirements means incorporating many different hardware and software products—both open source and proprietary. Even within a single vendor, products oftentimes don’t integrate seamlessly due to generational changes in software and acquisitions. Lean analytic architectures use a proprietary hardware and software solution when the solution provides a unique value but insist that the solution has open interfaces that make it readily integrated to other solutions.

The streamlining reduces the complex administration and maintenance costs while creating efficiencies for analyzing and deriving insights from the data.
Build on Human Factors

Flashy news headlines and hype have organizations believing that there are elusive individuals, known as data scientists, who embody the consummate triumvirate of deep expertise in computer science (software engineering, programming languages, and database skills), analytics (statistics, data mining, predictive analytics, simulation, optimization, and visualization skills), and domain expertise (industry, functional, or process expertise). Although there are some (actually precious few) individuals with this combination of skills, there is a growing realization that the elusive data scientist is actually a team of individuals who work closely together to fulfill the objective embodied in the data scientist role. These teams usually include a handful of
multidisciplinary data scientists who are part of the senior leadership in the team. To analytic-mature organizations, this is readily apparent and mirrors their experience in growing into an analytic-mature organization.

As the field of analytics has matured, the breadth and scope of analytics in organizations have increased. There is no longer just one type of role that builds, uses, and understands analytics in an enterprise. Instead, there are multiple roles or personas, each with different skills and responsibilities. Analytic-savvy organizations build on the human capital in the organization and understand what personas they have and what personas they need to achieve their business objectives. Various roles and personas contribute to the business differently, and all the personas are typically critical to achieving the business objectives. When there is a gap in skills, these organizations invest to bring the individuals or team up to speed. Analytic personas value new knowledge, and that’s a key component to keeping the scarce resources committed to your organization. By broadening and elevating the interests, awareness, and skills of your analytic professionals, you’ll keep the team engaged and innovating for your organization.

**Capitalize on Consumerization**

Consumerization of information technology continues to gain momentum in the marketplace. Consumerization today takes several forms, including “app stores,” crowdsourcing, and BYO (bring your own).

B2B app stores and markets have analytic apps emerging. Some of these apps are very narrow, discrete use cases, such as a credit scoring formula, whereas other apps are more comprehensive end-to-end use cases, such as multichannel marketing attribution. Although none of these may be a complete 100% fit, they can provide a starting point to speed up time to insight and drive down costs.
Crowdsourcing is a type of outsourcing through which an enterprise solicits contributions from an online community to perform a specific task. Crowdsourcing analytic models or algorithms provide access and leverage of outside expertise that would be difficult or impossible to access at an economical rate.

The “bring your own” era of self-service is well underway with analytic professionals who are demanding to use their favorite tools, data sources, and models rather than the standard information technology or mandated equivalents. Although information technology typically seeks to standardize and consolidate vendors and tools for cost and ease of support, analytic professionals typically value other considerations, such as user interface ease of use, flexibility to tailor via programming interfaces, and breadth of analytic models. Out of this tension has arisen the self-service approach to bringing your own:

- Data (BYOD)
- Tools (BYOT)
- Models (BYOM)

BYOD—bring your own data—is a way for organizations to combine their noncompetitive data to discover patterns and derive insights with new rich data sources. BYOT—bring your own tools—is a way to mix and match open source and proprietary technology tools to address specific service-level agreement requirements. BYOM—bring your own models—is a way to leverage app stores and crowdsourcing to derive value.

**Summary**

The principles of Modern Analytics chart the path toward analytic transformation and maturity in an organization. On the ground, these newly emerging principles of Modern Analytics are reshaping analytics from the bottom up in organizations. Today’s business leaders are
reshaping the next era of business—Business 3.0—from the top down driving automated, fact-based decisions, execution, and results.

To power your competitive differentiation, you must have a unique analytics roadmap to supercharge your business strategy to make the transformation from the Information Age into the next evolving age of faster changing business competition where epiphanies and agility to use those epiphanies to shift your business to the next level are key to thriving.

This book starts with the revolutionary stories to inspire you and which you can learn from and apply to your industry and business. We then transition into a framework that will allow you to identify opportunities throughout your organization for applying Modern Analytics. Some of these have never been attempted before; therefore, you may be breaking new ground. Don’t let this discourage you; the risk takers reap the highest rewards. Even if your initial attempts fail, learn from them so that when you use the technology and approaches for other applications, you learn from your mistakes. Some of your ideas for applications may not be completely new ideas but may be innovative in their approach, which may produce better results for your organization. This framework will allow you to systematically identify a wide variety of opportunities that align with your specific business strategies and objectives to uncover hidden value in your business that is lying dormant waiting to be discovered.

In the following chapters, you learn more about organizations that are transforming their businesses with analytics in this new age and how you can create a unique analytics roadmap for your organization.
This page intentionally left blank
Index

A
Abbott, Dean, 26
case study, 45-50, 52
accuracy of classification, 143
ACID (Atomic, Consistent, Isolation, Durability), 189
ad hoc analysis, 12, 90-91
Adams, John, 210
agent-based modeling, 51
aggregate functions, 166
Air Liquide case study, 43-45, 52
Akaike information criterion (AIC), 146
algorithms in analytics roadmap, 27
Alpine, 168, 178
Alteryx, 167-168
ambition factors of data scientists, 215-216
analysis data set
building, 133-139
assembling data, 134
evaluating data, 134-135
investigating outliers, 135-136
missing data, 138-139
table operations, 137-138
transforming data, 136-137
partitioning, 140-141
analytic applications in
customer-facing analytics, 102-103
analytic libraries (SQL), 192
analytic personas, 14-15
analytic solutions
brainstorming, 60-65
describing, 66-70
establishing roadmap, 82
prioritizing, 70-72
scoring, 73-81
analytic talent. See data scientists
analytics
business culture and, 206-207
change, 210
curiosity, 207-208
evidence, 210-211
experimentation, 209
problem solving, 208-209
business user tools
Alpine, 168
Alteryx, 167-168
IBM SPSS Modeler, 168-169
RapidMiner, 169
SAS Enterprise Guide, 170
SAS Enterprise Miner, 170-171
Statistica, 171
usage by user personas, 171
complexities of deployment, 205
customer-facing analytics, 62, 100-103
analytic applications, 102-103
consumer analytics, 103
prediction services, 101-102
descriptive analytics, 62-63
managerial analytics, 61, 94-96
operational analytics, 61-62, 96-99
possible uses, 31-32
predictive analytics, 63
   analysis data set, building, 133-139
   architectures, 174-187
   business needs, defining, 126-133
deployability, 99
methods, 124-126
predictive model, building, 139-147
predictive model, deploying, 147-152
use case, 108-111
prescriptive analytics, 63
principles of
   accelerate learning and execution, 10-11
   building human factors, 14-15
   consumerization, 15-16
   deliver business value and impact, 5-7
   differentiation, 11-12
   embedded analytics, 12
   establish lean architecture, 13-14
   focus on last mile, 7-9
   leverage Kaizen, 9-10
   list of, 4-5
programming languages, 160
   R Project, 160-163
   SAS programming language, 163-165
   SQL (Structured Query Language), 165-167
scientific analytics, 62, 99-100
simulation analytics, 63
strategic analytics, 61, 88-93
   ad hoc analysis, 90-91
   business simulation, 92-93
   econometric forecasting, 92
   market segmentation, 91-92
success factors, 154-155
teams
   analytics program office, 242
   Center of Excellence (COE), 238-240
centralized versus decentralized, 233-238
   chief data officer versus chief analytics officer, 240-242
   lab teams, 242
   organizing, 233
types of, 61-62, 88, 245
unique analytics roadmap, 23-24
   algorithms, 27
   analytic solutions, brainstorming, 60-65
   analytic solutions, describing, 66-70
   analytic solutions, establishing roadmap, 82
   analytic solutions, prioritizing, 70-72
   analytic solutions, scoring, 73-81
   approach, 26
   building, 245
   business area, 25
   case study, 52
data sources, 25-26
deploying, 243-244
   embedded analytics, 27-28
evaluating, 84
   importance of, 57
   key business objectives identification, 57-58
   precision, 26-27
   rough-cut project estimates, creating, 81
   speed, 28
   updating, 84
   value chain definition, 58-60
use cases
   discovery, 114-120
   explanation, 112
   forecasting, 113-114
   optimization, 121
overview, 105-108
prediction, 108-111
simulation, 120

user personas
analytics consumers, 159
business analysts, 158-159
data scientists, 157-158
list of, 155
power analysts, 155-157
tools used by, 171

analytics consumers, 159, 171
analytics generalists, 224-225
analytics managers, 222-224
analytics program office, 242
analytics programmers, 220-222
“Analytics: The Widening Divide”
(MIT Sloan School of Management), 33
anomaly detection in Discovery use case, 117-119
ANSI SQL standard, 166-167
ant systems, 44
Apache Giraph, 183
Apache Hive, 183, 196
Apache Mahout, 183
Apache Spark, 184
app stores, 15
appliances, 193
applied decision systems, 98
approach in analytics roadmap, 26
approving predictive model, 148
architectures
lean analytics architectures, 13-14
predictive analytics architectures, 174-187
  in cloud, 185-187
freestanding analytics, 175-177
in-database analytics, 179-182
in-Hadoop analytics, 182-185
partially integrated analytics, 177-179

asset management in predictive analytics, 151-152
association in Discovery use case, 116-117

At Home in the Universe: The Search for Laws of Self-Organization and Complexity (Kauffman), 50
attracting data scientists, 213-214, 230-231
attribution analysis, 96
automation, 20-22

B
backtesting, 6, 37
backward validation, 6
batch deployment in predictive analytics, 132, 150
batch SQL, 191
Bayesian information criterion (BIC), 146
Beiersdorf case study, 42-43, 52
BiosGroup, 50
brainstorming analytic solutions, 60-65
bring your own data (BYOD), 16
bring your own models (BYOM), 16
bring your own tools (BYOT), 16
budget in analytics roadmap, 75
built-in functions (SQL), 192
business analysts, 158-159, 171
business and technology history, 20-22
business area in analytics roadmap, 25
business culture, 206-207
  change, 210
curiosity, 207-208
evidence, 210-211
experimentation, 209
problem solving, 208-209
business needs, defining, 126-133
business objectives, identifying, 57-58
business simulation, 92-93
business user tools
  Alpine, 168
  Alteryx, 167-168
  IBM SPSS Modeler, 168-169
  RapidMiner, 169
  SAS Enterprise Guide, 170
SAS Enterprise Miner, 170-171
Statistica, 171
usage by user personas, 171
business value, measuring, 5-7
BYOD (bring your own data), 16
BYOM (bring your own models), 16
BYOT (bring your own tools), 16

C

call center example (workforce performance), 39-42
case studies
  Air Liquide, 43-45, 52
  Beiersdorf, 42-43, 52
  DataSong, 34-38, 52
  Dean Abbott, 45-50, 52
  P&G (Procter and Gamble), 50-52
  Talent Analytics, 38-42, 52
Center of Excellence (COE), 238-240
centralized teams, decentralized versus, 233-238
Champy, James, 21
change in business culture, 210
Charlotte Software Systems, 45
chief analytics officer, chief data officer versus, 240-242
chief data officer, chief analytics officer versus, 240-242
classification
  accuracy, 143
  regression versus, 129
  sensitivity, 143-144
cloud-based analytics, 185-187
Cloudera Impala, 196
clustering techniques, 116
Clustrix, 198
Codd, Edgar, 188
COE (Center of Excellence), 238-240
Competitive Advantage: Creating and Sustaining Superior Performance (Porter), 58
competitive differentiation, history of, 20-22

Competitive Strategy: Techniques for Analyzing Industries and Competitors (Porter), 23
connectors (Hadoop), 195
customer analytics in customer-facing analytics, 103
customer-facing analytics, 62, 100-103
analytic applications, 102-103
consumer analytics, 103
prediction services, 101-102
use cases, 107

data integration. See architectures
data munging, 46, 137
data prep professionals, 219-220
data science in analytics roadmap, 26
data scientists, 157-158
  ambition factors, 215-216
  characteristics of, 211-218
  finding, 229-230
  recruiting, 213-214, 230-231
  roles of, 215
    analytics generalists, 224-225
    analytics managers, 222-224
    analytics programmers, 220-222
    data prep professionals, 219-220
    sample job structure, 225-229
tools used by, 171
INDEX  251

data sets
  building, 133-139
  assembling data, 134
  evaluating data, 134-135
  investigating outliers, 135-136
  missing data, 138-139
  table operations, 137-138
  transforming data, 136-137
partitioning, 140-141

data sources in analytics roadmap, 25-26

data warehouses, 189
databases. See also SQL (Structured Query Language)
  ACID (Atomic, Consistent, Isolation, Durability), 189
in-database analytics, 175, 179-182, 188
modern SQL platforms, 188-192
  future of, 200-202
  MPP (massively parallel processing) databases,
    192-195
  NewSQL databases, 197-200
  SQL-on-Hadoop, 195-197
NewSQL, 190
NoSQL, 189-190
DataSong case study, 34-38, 52
Day-in-the-Life-of-Scenarios, 76-81
DB Lytix, 181
DB2 Intelligent Miner, 180
de Bono, Edward, 209
decentralized teams, centralized versus, 233-238
decision models
  creating, 70-72
  scoring, 73-81
decision trees, 46-50
deployability in predictive analytics, 99
deploying
  analytics, complexities of, 205
  predictive model, 147-152
    asset management, 151-152
    measuring performance, 150-151
  reviewing and approving, 148
  scoring, 148-150
unique analytics roadmap, 243-244
deployment environment in predictive analytics, 132-133
descriptions of analytic solutions, 66-70
descriptive analytics, 62-63
differentiation of analytic strategy, 11-12
The Discipline of Market Leaders (Treacy and Wiersema), 21
discovery
  anomaly detection, 117-119
  association, 116-117
  graph and network analysis, 119-120
  segmentation, 116
  text and document processing, 114-115
  use case, 114-120
Disney, Walt, 207
Divis, 43
document analytics, 114-115
DS2 programming language, 165

E
Earhart, Amelia, 19
econometric forecasting, 92
Edison, Thomas, 209, 242
Einstein, Albert, 208
embedded analytics, 12, 27-28
end-user analytics
  business user tools
    Alpine, 168
    Alteryx, 167-168
    IBM SPSS Modeler, 168-169
    RapidMiner, 169
    SAS Enterprise Guide, 170
    SAS Enterprise Miner, 170-171
    Statistica, 171
  usage by user personas, 171
success factors, 154-155
user personas
  analytics consumers, 159
  business analysts, 158-159
  data scientists, 157-158
  list of, 155
  power analysts, 155-157
  tools used by, 171
ensemble models, 48
errors, cost of, 129-131
estimates, creating rough-cut project estimates, 81
evaluating
  data sets, 134-135
  unique analytics roadmap, 84
evaluation criteria for analytic solutions, 70-71
evaluation rubric for analytic solutions, 71-72
evidence in business culture, 210-211
evolutionary strategy, 43
experimentation
  in business culture, 209
  innovation via, 10-11
explanation use case, 112

F
false negatives, 131
false positives, 131
fault tolerance, 189
finding data scientists, 229-230
forecasting use case, 113-114
freestanding analytics, 175-177
future of modern SQL platforms, 200-201
fuzzy logic, 43
Fuzzy Logix, 181

G
goals
  of predictive analytics, 123-124
  setting, 7-9
Google Spanner, 199

graph and network analysis in
  Discovery use case, 119-120
GraphLab, 184

H
H2O, 184
Hadapt, 197
Hadoop, 189
  in-Hadoop analytics, 175, 182-185
  SQL-on-Hadoop, 195-197
Hammer, Michael, 21
heroes, 19-20, 22
history
  of business and technology, 20-22
  of SQL (Structured Query Language), 187-188
Hortonworks Stinger, 196
HP Vertica, 193-194
HTAP (Hybrid Transaction and Analytical Processing), 201
human role in analytics, 14-15
hybrid teams, 235
Hybrid Transaction and Analytical Processing (HTAP), 201

I
IBM Netezza Analytics, 181
IBM PureData, 180, 194
IBM SPSS Modeler, 168-169, 178
in-database analytics, 175, 179-182, 188
in-Hadoop analytics, 175, 182-185
innovation via experimentation, 10-11
integration of data. See architectures interactive SQL, 191
inventory supply chain, P&G (Procter and Gamble) case study, 50-52

J-L
Kaizen, 9-10
Kauffman, Stuart, 50
Keller, Helen, 19
key business objectives, identifying, 57-58

lab teams, 242
lean analytics architectures, 13-14

M
managerial analytics, 61, 94-96, 107
MapReduce, 183
market segmentation in strategic analytics, 91-92
marketing mix modeling, 34
massively parallel processing (MPP) databases, 192-195, 200-202
measuring
  business value, 5-7
  performance of predictive model, 142-146
MemSQL, 199
methodologies for predictive analytics, 124-126
  analysis data set, building, 133-139
  business needs, defining, 126-133
  predictive model
    building, 139-147
    deploying, 147-152
missing data, 138-139
Model Building subcase (Prediction use case), 108
model repository management in predictive analytics, 151-152
Model Scoring subcase (Prediction use case), 108
model training plan, executing, 141
modeling plan, developing, 139-140
modern SQL platforms, 188-192
  future of, 200-202
  MPP (massively parallel processing) databases, 192-195
  NewSQL databases, 197-200
SQL-on-Hadoop, 195-197
MPP (massively parallel processing) databases, 192-195, 200-202
munging, 46

N
Nesbitt, Tess, 35
“New Path to Value” (MIT Sloan School of Management), 33
NewSQL databases, 190, 197-202
N-fold cross-validation, 147
NoSQL databases, 189-190
NuoDB, 199

O
on-demand analytics, 12
open source R, 181
operational analytics, 61-62, 96-99, 107
operational costs (Air Liquide case study), 43-45
operational forecasting systems, 98-99
operational SQL, 191
optimization use case, 121
Oracle, 180
  Advanced Analytics Option, 181
  Exadata, 194
  R distributions, 181
order size policy, P&G (Procter and Gamble) case study, 51-52
organizational culture. See business culture
organizing teams, 233
  analytics program office, 242
  Center of Excellence (COE), 238-240
  centralized versus decentralized, 233-238
  chief data officer versus chief analytics officer, 240-242
  lab teams, 242
outliers
  detecting, 117-119
  investigating, 135-136
out-of-time sample validation, 147
P

P&G (Procter and Gamble) case study, 50-52
partially integrated analytics, 175, 177-179
partitioning data sets, 140-141
“Pass Through” integration, 178
performance of predictive model, measuring, 142-146, 150-151
personalas. See user personas
Pivotal Database, 181
Pivotal Greenplum Database, 195
Pivotal SQLFire, 199
PMML (Predictive Model Markup Language), 111
POCs (Proof-of-Concepts), 9
Porter, Michael, 23, 58
power analysts, 155-157, 171
precision in analytics roadmap, 26-27
prediction services in customer-facing analytics, 101-102
prediction windows, determining, 131-132
predictive analytics, 63
architectures, 174-187
in cloud, 185-187
freestanding analytics, 175-177
in-database analytics, 179-182
in-Hadoop analytics, 182-185
partially integrated analytics, 177-179
deployability, 99
goals, 123-124
methodologies, 124-126
analysis data set, building, 133-139
business needs, defining, 126-133
predictive model, building, 139-147
predictive model, deploying, 147-152
use case, 108-111
predictive model
building, 139-147
measuring performance, 142-146
model training plan, 141
modeling plan, developing, 139-140
partitioning data set, 140-141
deploying, 147-152
asset management, 151-152
measuring performance, 150-151
reviewing and approving, 148
scoring, 148-150
validating, 146-147
Predictive Model Markup Language (PMML), 111
predictor variables in modeling plan development, 139-140
prescriptive analytics, 63
principles of analytics
accelerate learning and execution, 10-11
building human factors, 14-15
customerization, 15-16
deliver business value and impact, 5-7
differentiation, 11-12
embedded analytics, 12
establish lean architecture, 13-14
focus on last mile, 7-9
leverage Kaizen, 9-10
list of, 4-5
prioritizing analytic solutions, 70-72
private knowledge, 99
probing questions in value chain brainstorming, 63-65
problem solving in business culture, 208-209
Procter and Gamble (P&G) case study, 50-52
product innovation (Beiersdorf case study), 42-43
Index

programming languages for analytics, 160
  R Project, 160-163
  SAS programming language, 163-165
  SQL (Structured Query Language), 165-167

Proof-of-Concepts (POCs), 9
public knowledge, 99
“Push Down” integration, 178

Q-R

R Project, 160-163
  open source R, 181
  Oracle R distributions, 181
RapidMiner, 169
real-time SQL, 191
Receiver Operating Characteristic (ROC) curve, 145
recruiting data scientists, 213-214, 230-231

Reengineering the Corporation (Hammer and Champy), 21
regression, classification versus, 129
response measures, defining, 127-129
response-attribution analysis, 112
reviewing predictive model, 148
risk assessment, 93
RMSE (root mean square error), 145
roadmap. See unique analytics roadmap
Roberts, Greta, 39, 212
ROC (Receiver Operating Characteristic) curve, 145
root mean square error (RMSE), 145
rough-cut project estimates, creating, 81
Roumeliotis, George, 236, 241

S

SAP HANA, 200
SAS Enterprise Guide, 170
SAS Enterprise Miner, 170-171, 179
SAS programming language, 163-165, 178
SAS Scoring Accelerator, 179
scalar functions, 166
scientific analytics, 62, 99-100, 107
scoring
  analytic solutions, 73-81
  partially integrated analytics, 177-179
  prediction versus, 108-111
  predictive model, 148-150
segmentation
  in Discovery use case, 116
  market segmentation in strategic analytics, 91-92
sensitivity of classification, 143-144
sentiment analysis, 115
service calls (Dean Abbott case study), 45-50
simulation analytics, 63, 120
Six Thinking Hats (de Bono), 209
Smarter Remarketer case study, 45-50
Socrates, 210
software. See architectures; business user tools; modern SQL platforms
speed in analytics roadmap, 28
Splice Machine, 197
split-sample validation, 147
SQL (Structured Query Language), 165-167
  history of, 187-188
  modern SQL platforms, 188-192
    future of, 200-201
    MPP (massively parallel processing) databases, 192-195
    NewSQL databases, 197-200
    SQL-on-Hadoop, 195-197
SQL-on-Hadoop, 195-197, 201
State University of New York (SUNY) at Buffalo, 100
Statistica, 171
strategic analytics, 61, 88-93
  ad hoc analysis, 90-91
  business simulation, 92-93
INDEX

econometric forecasting, 92
market segmentation, 91-92
use cases, 107
streaming SQL, 191
Structured Query Language (SQL).
See SQL (Structured Query Language)
SUNY (State University of New York) at Buffalo, 100
supply chain, P&G (Procter and Gamble) case study, 50-52
support activities, 58
survival analysis, 35

table operations, building data sets, 137-138
talent. See data scientists; teams, organizing
Talent Analytics case study, 38-42, 52
targeting and routing systems, 98
teams, organizing, 233
analytics program office, 242
Center of Excellence (COE), 238-240
centralized versus decentralized, 233-238
chief data officer versus chief analytics officer, 240-242
lab teams, 242
technology and business history, 20-22
Teradata, 180-181, 195
Teradata SQL-H, 197
text analytics, 46, 114-115
text mining, 115
time series analysis, 113-114
timeline in analytics roadmap, 75
tools. See architects; business user tools; modern SQL platforms
transactional deployment in predictive analytics, 132, 150
transforming data, 136-137
TransLattice, 200
Treacy, Michael, 21

U
UDFs (user-defined functions), 192
unique analytics roadmap, 23-24
building
analytic solutions, brainstorming, 60-65
analytic solutions, describing, 66-70
analytic solutions, establishing roadmap, 82
analytic solutions, prioritizing, 70-72
analytic solutions, scoring, 73-81
key business objectives
identification, 57-58
rough-cut project estimates, creating, 81
steps in, 245
value chain definition, 58-60
capabilities
algorithms, 27
approach, 26
business area, 25
data sources, 25-26
embedded analytics, 27-28
list of, 24
precision, 26-27
speed, 28
case study, 52
deploying, 243-244
evaluating, 84
importance of, 57
updating, 84
updating unique analytics roadmap, 84
Upstream, 34-38, 52
use cases
discovery, 114-120
anomaly detection, 117-119
association, 116-117
graph and network analysis, 119-120
segmentation, 116
text and document processing, 114-115
explanation, 112
forecasting, 113-114
optimization, 121
overview, 105-108
prediction, 108-111
simulation, 120

user personas
analytics consumers, 159
business analysts, 158-159
data scientists, 157-158
list of, 155
power analysts, 171

user-defined functions (UDFs), 192

V
validating predictive model, 146-147
value at risk, 129-131
value chain, defining, 58-60
VoltDB, 200
volumetric modeling, 34

W-Z
Wanamaker, John, 34
Wayne, John, 19
weighting analytic solutions, 73-81
Wiersema, Fred, 21
window functions, 166
workforce performance (Talent Analytics case study), 38-42
World Programming System (WPS), 164

[x+1] platform, 102