TDWI RESEARCH

TDWI CHECKLIST REPORT

Seven Steps for Executing a Successful Data Science Strategy

By David Stodder

Sponsored by:

sas

tdwi.org

tdwi
Advancing all things data.

JANUARY 2015

TDWI CHECKLIST REPORT

Seven Steps for Executing a Successful Data Science Strategy

By David Stodder

TABLE OF CONTENTS

tdwi
Advancing all things data.

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

Data science is a hot topic among business and IT leaders. Excitement about the potential benefits of data science is tempered, however, by anxiety about how hard it is to find, hire, and train data science personnel, not to mention the difficulty of defining the term within the context of an organization's goals and objectives.

There is no single definition of *data science*, nor one solution or technology. It is a term that joins together contributions from several fields, including statistics, mathematics, operations research, computer science, data mining, machine learning (algorithms that can learn from data), software programming, and data visualization. It can cover the entire process of acquiring and cleaning data, methods for exploring the data and extracting value from it, and techniques for making insights actionable for humans and automated processes.[1] Most often, the focus of data science is to optimize decisions and realize higher value from data through advanced analysis.

One factor that makes data science distinct, however, is the word *science*. Data science is about applying scientific methods to explore and test hypotheses about the data. Indeed, many data scientists come from hard science fields such as chemistry and physics or professions such as neurobiology and nuclear physics. Data science pioneers have contributed mightily to the growth of social media and e-commerce; now, firms in other industries are keen to apply data science to their decision-making processes.

Continuous experimentation through examination of data to test hypotheses is at the heart of most data science projects. At the same time, the availability of technologies that can work with enormous data volumes and variety enables professionals to complement scientific methods with hypothesis-free approaches that employ machine learning to examine data and discover unforeseen patterns before articulating a hypothesis. This enables organizations to use data science to find previously hidden risks and opportunities and apply analytics to improve outcomes.

To solve business problems, develop new products and services, and optimize processes, organizations increasingly need analytics insights produced by data science teams with a diverse set of technical skills and business knowledge who are also good communicators. This TDWI Checklist Report describes seven steps to achieve a successful data science strategy.

[1] For a good discussion of how to define data science, see: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

NUMBER ONE
IDENTIFY YOUR ORGANIZATION'S KEY BUSINESS DRIVERS FOR DATA SCIENCE

Data science may not be for everyone. Before embarking on a data science project, the first question to ask is a simple one: Do we need data science? Users may appear content with spreadsheets, business intelligence (BI) applications, and the selection of structured data available through data warehouses or other IT-managed repositories. Existing reports and dashboards may seem sufficient. From this perspective, investing in data science and technology to expand the reach of analytics into more data types, including semi-structured and unstructured, may appear unjustified.

To evaluate whether it is worth engaging in data science, organizations need to look at the value it could bring beyond what it already realizes from traditional BI, analytics, and data warehousing. The place to begin such an evaluation is the potential business drivers: What business value could be gained by developing a data science strategy? What are the questions the organization needs to solve to be more competitive, effective, and proactive? How well does the organization understand—and know how to respond to—the interplay of factors that affect customer behavior, the success of its website, or the impact of key trends? Often, such analysis will reveal knowledge gaps the organization has been unable to fill with its current BI and data warehousing systems.

For this reason, one of the most important qualities to seek when selecting personnel for a data science effort is knowledge of and curiosity about the business. Data scientists often come into an enterprise possessing exceptional technical and scientific skills. However, it is critical that they also develop the business domain expertise to uncover questions the organization needs to ask using analytics, and how to make the resulting data insights actionable.

At this stage, organizations should identify where data science could contribute most to realizing business objectives. Some classic areas include achieving greater personalization and computational efficiency in marketing and advertising; monitoring social media; modeling attribution to determine what drives purchasing; establishing a dynamic pricing strategy across multiple channels; uncovering fraudulent activity; and autonomic analysis of important documents or images such as call center logs or checks.

NUMBER TWO

CREATE AN EFFECTIVE TEAM FOR ACHIEVING DATA SCIENCE GOALS

"It's like chasing unicorns" is a phrase often used to describe the difficult task of finding and keeping those rare individuals with the experience and ability to perform all that is required of a data scientist. In this exclusive group, many have a Ph.D. and a good number come from diverse, non-computer-science backgrounds.

The pioneers of data science—"half hackers and half scientists," as one person put it—often took a do-it-yourself approach through hands-on implementation of Hadoop and other open source technologies to store, access, and analyze massive and varied sources of big data. Although firms have benefited from their innovation, the artisan approach has left them vulnerable if and when their data scientists are lured away by competitors.

Rather than focus on finding one or a few individuals who seem to be able to do it all, a wiser course is to develop a stable team that brings together the talents of multiple experts. As discussed in the previous step, the team's members must understand business drivers and not lose sight of the goal of delivering actionable business value. Each member of the team should also have enthusiasm, curiosity, and creative energy for working with business leadership on data and analytics projects.

Depending on the project, the team will need personnel with a combination of skills that include expertise in the business domain (for example, customer engagement or marketing), business analytics, statistics, data mining, machine learning, data and information retrieval, programming, prototyping, and visualization. Organizations should assemble a team that includes individuals with communication skills, not just technical acumen.

Although it is valuable to look externally for data scientists and leadership such as chief data officers, taking a team approach allows organizations to look internally. Many organizations already have personnel who could join a data science team. Indeed, TDWI Research finds that the majority of organizations plan to train internal personnel to handle data science projects. Personnel could include business analysts, statisticians, software developers, data analysts, and other data professionals.

In this step, organizations should bring business and IT leadership together to develop a strategy for creating effective and sustainable data science teams. Their plan should include training and incentives to attract internal personnel.

NUMBER THREE

EMPHASIZE COMMUNICATION SKILLS TO REALIZE DATA SCIENCE'S VALUE

Organizations that use data science successfully almost universally point to communication as a key ingredient to their success. Insights provided by analytics are of little value unless the data science team articulates what the findings say and why they are significant to business goals. Often this is not easy, especially if the presentation of the findings calls into question executives' "gut feel" assumptions about business strategy, strays from tightly controlled modes of BI reporting and analysis, or suggests that established processes are ineffective or outdated. Data science often points to the need for change—and change can be difficult.

Communication is also vital to improving collaboration in a data science project. Often, along with data scientists, key players (such as statisticians, business analysts, data analysts, and developers) are scattered in silos across the organization, or business and data analysts may work in a separate department than the business stakeholders, who should also be part of the data science effort. Important new perspectives can come if data science teams are able to work across divisions or silos to gain a more global view.

For example, to identify which actions are most influential to the buying behavior of an important cluster of customers, it is valuable if data science teams can examine data from a number of sources that might be managed in different divisional silos such as e-commerce, brick-and-mortar stores, contact centers, and field service offices. This "big data" has never fit easily into a data warehouse, much less a spreadsheet. The data science team could make a great contribution just by pulling together a global, holistic view of this scattered data.

Working across the organization is also important when the goal of data science is to optimize a process by developing algorithms that will automate decisions. Communication is essential; the team must be aware of how optimization will impact dependent processes, including how data is collected and analyzed. Without good communication, optimization could have unintended consequences.

Communication by and among data science teams is essential to building a data-driven analytics culture. In this step, organizations should emphasize the value of communication and make it a priority as they evaluate candidates for data science teams.

NUMBER FOUR
EXPAND THE IMPACT OF DATA SCIENCE THROUGH VISUALIZATION AND STORYTELLING

Data science fits into a larger objective of creating a data-driven "analytics culture" that is energized by a shared desire to improve decision making at all levels, from executives to frontline personnel. The key goal is to supplant uninformed, emotional decision making based on inaccurate theories with decision processes that are supported by empirical evidence, testing of hypotheses, and impactful data analysis. Although inspiration will always be vital, companies with healthy analytics cultures accept the notion that assumptions should be questioned by looking closely at the data.

Data science thrives in an analytics culture. However, not all personnel in an organization are going to be part of data science teams, nor should they be. To bring more users into the analytics culture, organizations should explore technologies that can support the "democratization" of BI, analytics, and data discovery. These products are increasingly able to address users' self-service demands for data access and interaction without IT hand-holding. The tools go beyond simple spreadsheets and canned reporting to deliver different perspectives on metrics, help users uncover trends, and enable them to personalize dashboards.

Data visualization is an essential technology for data science and most self-service BI, analytics, and data discovery use cases. Across organizations, users' visualization requirements can be diverse; some need simple interfaces that emphasize how to respond to a situation while others demand more varied types of visualizations. Leading tools have libraries of visualization types, and more are available through open source libraries. Organizations should take advantage of maturing data visualization technologies for both advanced data science and data interaction by nontechnical users.

Visualization enables "data storytelling." This hot trend fuses visualization, data analysis, and usually verbal or written discussion, often in an infographic, to provide interpretation of data science results and why they are significant. Storytelling can be an effective way for data science teams to communicate accurately what they have found rather than just present numbers that could be misinterpreted. Organizations should encourage data storytelling and provide training so data science teams and other users can do it well.

NUMBER FIVE
GIVE DATA SCIENCE TEAMS ACCESS TO ALL THE DATA

Data is the raw material of data science. Like chefs looking for new taste sensations, data scientists need to work closely with data at every step so they know what they have and can extract fresh insights to deliver business value. Although valuable for reporting and proscribed forms of analysis, most traditional BI and data warehousing systems offer users only selected data samples, subsets, and pre-aggregated reports that have been carefully scrubbed and manicured by data professionals. Instead of raw data, most BI users work with reports or dashboards. What they leave behind are unincorporated structured sources and a vast universe of semi- and unstructured data and content that has never easily fit into BI systems and data warehouses.

Structured data can, of course, be voluminous and varied, especially when brought in from diverse applications. However, data science is often more closely associated with the desire to analyze semi- and unstructured data because these sources are growing rapidly and have been analyzed little, if at all. Preparing this breadth of data, assessing its quality, looking for gaps and errors, and performing exploratory analysis to determine relevant extracts are essential data science activities. They can take up the lion's share of a data science team's time. Although tools can automate steps, data science teams need to get close to the data to properly move forward with analytics and algorithm development.

Computer logs, social media, sensor data, and other new sources can be messy and chaotic; organizations should be realistic about the effort it will take to investigate and prepare the data. Organizations should ensure that data science teams include personnel who are comfortable working with raw data. In most cases, the team will need personnel who are knowledgeable about Hadoop and related technologies and are familiar with data lake and data hub concepts for gathering, storing, and accessing raw data.

Data science teams should always be on the lookout for interesting and potentially relevant data sources. Often, more than one application will be recording diverse (or sometimes the same or similar) data about customers, transactions, or other objects. Data scientists can play a valuable role by uncovering discrepancies and data quality problems.

NUMBER SIX

PREPARE DATA SCIENCE PROCESSES FOR OPERATIONALIZING ANALYTICS

Businesses can execute at a higher level if they can strengthen the connection between analytics and business processes. The first step is to move beyond purely "descriptive" analytics, which only answers *what* and *why* questions about historical trends and events, to predictive analytics, which can help discern what is likely to happen next. By streamlining how they develop and deploy predictive models, organizations can expand their use into more operational processes.

However, getting business value from this expansion requires more than just producing more analytic models faster. Firms must move to the next stage: to "prescriptive" analytics, which is about producing not just predictive insights but also suggested actions. Prescriptive analytics can be useful to both humans responsible for business processes and for guiding emerging automated decision systems.

Potential use cases abound. The most common is to improve customer marketing to offer targeted cross-sell, up-sell, and next-best-action offers at the moment of engagement. Another example occurs in complex, high-volume supply chains. Leading firms today apply predictive modeling to forecast what might happen given the probability of factors that could affect product manufacturing, packaging, and shipping. To get maximum value from their analysis, these firms are moving toward prescriptive analytics to develop recommended options for automated rules and complex event processing systems. This evolution could also be important for organizations seeking to operationalize analytics to fight fraud, assess risks, position mobile assets, and more, in real time.

To operationalize analytics, data science teams must focus on reducing the time it takes to develop and deploy analytic models. With cleaner workflows and processes, data science teams can move *away from* uncoordinated, artisanal model development and *toward* practices that include quality feedback sessions to correct flaws. Along with process improvements, organizations can take advantage of new technology practices such as in-database scoring, which can help eliminate time-consuming data movement to specialized data stores, improve the performance of analytic models, and make models available for multiple applications as stored procedures.

Teams must continue to improve communication with business stakeholders. Delays in model development and deployment are often due as much to communication difficulties as they are to other factors.

NUMBER SEVEN

IMPROVE GOVERNANCE TO AVOID DATA SCIENCE "CREEPINESS"

Data science teams must keep in mind that the outside world contains another set of stakeholders: the general public, including current and prospective consumers of the firm's products and services. Fear and concern are at a high level with the continued unfolding of news about data thefts, hacking, surveillance, online and geolocation tracking, and marketing retargeting. Leading retail firms have had their reputations sullied by security breaches. Commentators rail about the "creepiness" factor: that is, the extent of knowledge firms are amassing about customers' purchasing and other observed behavior that through powerful, real-time analytics can be (and often is) turned into highly personalized marketing. "Creepiness" is the label given to what some call the "dark side" of data science.

Data science teams, along with business leadership, must be cognizant of the right balance between what they can achieve through advanced analysis of consumer data and what is tolerable—and ethical—from the public's perspective. Often there is no single standard; companies report that younger "millennial" demographics groups are more tolerant of personalized targeting than are older groups. Some consumers appreciate having the flow of advertising and marketing be more relevant to their buying patterns and shopping interests, while others are surprised and upset by it. Some will voice their concerns through social media, proving the observation that marketing is always a conversation, not one-way communication.

Enterprises should ensure that ethics and consumer tolerance are part of data science planning discussions, along with adherence to standard data governance policies. Data science teams must make sure they are not cloistered from the outside world and that they hear about how consumers and the public in general are responding to actions taken based on their data insights. The teams should consult with business leaders to gain their feedback about how certain programs could affect the conversation between the company and the public—and consider the possible ramifications on the company's reputation.

Governance policies should address how to protect sensitive data during data science processes, particularly personally identifiable information. Anonymizing data may not be sufficient. Organizations should examine how they can protect data used in algorithms so that consumers' behavior patterns cannot be hacked by those looking to identify specific people.

ABOUT OUR SPONSOR

sas.com

SAS is the leader in advanced analytics software and services and is the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 70,000 sites worldwide benefit from advanced analytics — whether for revealing relationships between pricing and demographics, understanding how social media influences your customers, or predicting risk and fraud. With an unwavering focus on analytics since 1976, SAS offers the breadth and depth of advanced algorithms and techniques such as machine learning, text analytics, and a broad set of predictive techniques.

Learn more about SAS advanced analytics at http://www.sas.com/en_us/software/analytics.html

ABOUT THE AUTHOR

David Stodder is director of TDWI Research for business intelligence. He focuses on providing research-based insight and best practices for organizations implementing BI, analytics, performance management, data discovery, data visualization, and related technologies and methods. He is the author of TDWI Best Practices Reports on mobile BI and customer analytics in the age of social media, as well as TDWI Checklist Reports on data discovery and information management. He has chaired TDWI conferences on BI agility and big data analytics. Stodder has provided thought leadership on BI, information management, and IT management for over two decades. He has served as vice president and research director with Ventana Research, and he was the founding chief editor of *Intelligent Enterprise*, where he served as editorial director for nine years. You can reach him at dstodder@tdwi.org.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.