Eckerson Group

Best Practices in Data Science

Ten Keys to Operational Success and Business Value

By Stephen J. Smith

April, 2018

Co-sponsored by

cloudera

About the Author

Stephen J. Smith is the research leader for data science at Eckerson Group. His unique perspective comes from his real-world experience in building the machine learning and data science products Darwin, Discovery Server, and Optas. He is co-author of the highly rated business technology books Data Warehousing, Data Mining, and OLAP and Building Data Mining Applications for CRM (McGraw-Hill Education).

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 20 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.

About this Report

This report is made possible through generous support of our sponsors, Alteryx, Cloudera and SAP. It is based on interviews with leading vendors and practitioners of data science. See the Appendix for the names of individuals who graciously donated their time to this project.

Table of Contents

Executive Summary

Data science is fundamentally different from other analytics technologies because it is prescriptive: it uses enterprise data to tell business users what to do, not what they have already done. It optimizes key business processes, helping companies reduce costs, grow revenues, and manage risk. Given these benefits, many executives are eager to invest in data science.

But many companies fail to achieve success with data science. Many don't recognize the need to manage the data science lifecycle that spans data preparation, model design, deployment, and model management. In addition, some business people refuse to trust models, preferring their own experience over a statistical output. Others have lost confidence in analytical models that contain errors or don't deliver much business value.

However, there is a way to unlock the power of data science. The key is to elevate data science from a craft practiced by solo practitioners to an industrial process that spans the full data science lifecycle and is executed by a team of people with complementary skills. The goal is to create a process that is repeatable, automated, managed, and optimized and consistently delivers quantifiable business value. There are many challenges to implementing data science. It's easy for models to go off the rails and deliver inaccurate, inconsistent, or irrelevant results. The good news is that some companies have created data science programs that systematically turn data into insights and action for business gain. This report presents best practices to overcome ten common data science challenges.

The Discipline of Data Science

The Rise of Data Scientists

Data science has had many names: statistics, data mining, predictive analytics, and now machine learning and artificial intelligence. In 2012, Thomas Davenport wrote a Harvard Business Review article titled, "Data Scientist: The Sexiest Job of the 21st Century." Davenport did not coin the term data scientist, but his article popularized the term "data science" and defined a new role that expanded the skills required for what was previously the domain of statisticians.

Davenport recognized that analytics requires more than academically-trained statisticians. He asserted that a data scientist needs knowledge of business domains, data preparation, data governance, application coding, feature engineering, DevOps, model management, and ROI optimization—basically the full lifecycle of turning data into analytical models that drive business value.

With this expanded list of skills, Davenport created a mythical "super scientist" that doesn't exist. Today, leading companies build data science teams comprised of individuals with complementary skills who work collaboratively to operationalize data science and deliver a positive return on investment.

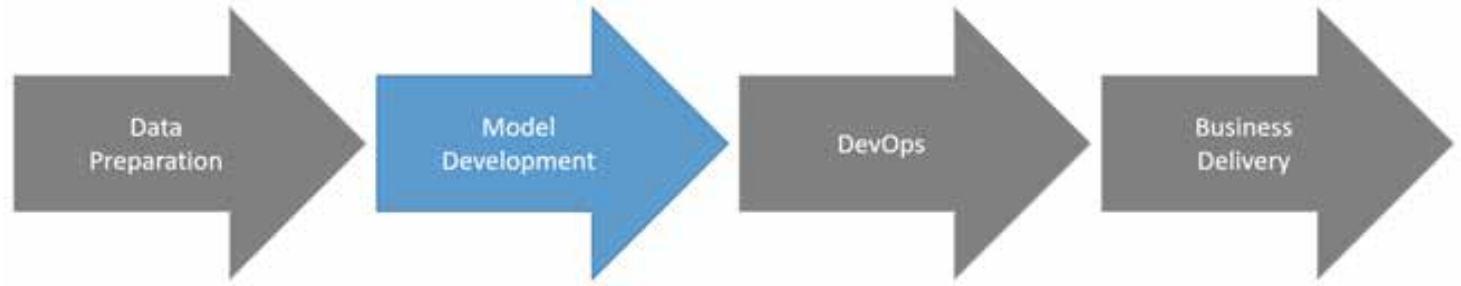
Data Science Lifecycle

Data science can behave a bit like plutonium. It can do great good, but cause big explosions. To keep plutonium safe and make it productive, we manage it using a repeatable, automated, and governed process within a nuclear power plant. To get value from data science, we need to create a similar process to create, manage, and deploy analytical models.

Data science is more than model development.

Data science is more than model development. There are important functions both upstream and downstream from the model building process that are critical ingredients to the success of a data science project. The data science lifecycle spans data preparation, model development, DevOps, and business delivery (see figure 1.)

Figure 1. Data Science Lifecycle



Why Now?

Historically, data science solutions have focused on model creation. They have been designed to simplify, automate, and optimize the design, selection, and validation of analytical models.

Today, most vendors recognize that model creation is just one part of the data science lifecycle. On the front-end, they have expanded their workbenches with data preparation features to help data scientists select, profile, clean, format, merge, manipulate, segment, and validate data to feed their models. On the back-end, they have added model management and deployment features to put models into production, monitor their output, and ensure their accuracy.

Data science has moved from a craft practiced by one or two skilled artisans… to an industrial process executed by a data science team that generates dozens, if not hundreds, of high-value models annually.

What's clear is that data science has moved from a craft practiced by one or two skilled artisans who produce a handful of models a year to an industrial process executed by a data science team that generates dozens, if not hundreds, of models annually. Most importantly, each model produces a quantifiable return on investment (ROI) and complies with emerging regulations governing the use of data science output.

The timing is right because many organizations are eager to move beyond reporting to analytics. Rather than analyze events that have already happened, they want to use their vast treasure trove of historical data to make smart predictions, work proactively to meet customer needs, and adapt quickly to changing business conditions. Many organizations that have dabbled with data science and seen its potential now want to expand its footprint and create a world-class data science program that delivers continuous business value.

Data Science Challenges and Strategies

However, the road to data science nirvana is paved with potholes. Organizations face many challenges when implementing data science. Those that succeed develop effective strategies to avoid or counteract the challenges.

Table 1 describes ten common challenges that organizations face when implementing data science and strategies to address them. The challenges and solutions are based on feedback from experts in the field and are organized by stage in the data science lifecycle. The rest of this section describes strategies for addressing each challenge.

Table 1. Ten Data Science Challenges and Strategies by Stage

Process Stage I: Data Preparation		
Challenge	**Description**	**Strategies**
1. Ambiguous Data Semantics	Data scientists can easily misunderstand the meaning of a predictor field which can lead to costly mistakes.	Create a data dictionary; hire a data librarian; create a feature library; embed processes in a data science platform; and use data science as a data governance catalyst.
2. Data Drift, Anomalies, and Errors	Dirty, incomplete, inconsistent, non-valid, or changing data can skew models and must be cleaned and normalized.	Apply data profiling tools to understand the shape of data and then use data preparation tools to clean and normalize data and impute or delete missing values.
Process Stage II: Model Development		
3. Bad Model Validation	Models built without proper training and test data sets or cross-validation processes can generate models that work well in the lab but fail in the field.	Use data science tools that automate model testing with strong cross-validation techniques. Always trial any new model on real world data before launching it.
4. Regulatory Compliance	Models that fail to adhere to privacy and other regulations can lead to costly fines and potential loss of reputation in the market.	Bake in privacy compliance for model building that includes PII detection. Grow cultural awareness of the risks of using certain data.
Process Stage III: DevOps		
5. Model Degradation	Model performance can degrade over time, sometimes because of changes in the data that feeds the model, and sometimes because of changes in the variable being predicted.	Automate model monitoring with tools that provide threshold alerts delivered to operations and only involve data scientists when a new model needs to be created.

6. Deployment Disconnect	One of the most common problems with data science is the need to recode the model to run in a different environment from where it was built. This slows down deployment and can introduce new errors.	Build and score models in the same data ecosystem to minimize recoding and use APIs to execute models in data science platforms. Look to the cloud to help scale support for both model creation and deployment.
Process Stage IV: Business Delivery		
7. Business Disconnect	Businesses often do not build a formal business case for each data science initiative. Skipping this step often results in models that deliver what was requested but not what was needed.	Hire data scientists with business knowledge and strong communications skills; train business managers how to work with data scientists; leverage optimization tools to track ROI.
8. Model Opacity and Repeatability	Some data science algorithms generate models that are not easy to explain, audit, or roll back. This makes them unusable in many environments, especially highly regulated ones.	Use transparent machine learning algorithms and be willing to trade transparency for accuracy to increase business trust in model output and comply with regulations.
9. Finding Data Scientists	Data scientists that have all the skills required to deliver a data science project from start to finish are hard to find, prohibitively expensive, and hard to retain.	Build an agile, multidisciplinary team that combines data engineers, data scientists, data librarians, domain experts, and DevOps specialists.
10. Lack of a Data Science Program	Most companies hire a data scientist or two and conduct a few projects without generating a lot of business value. There is no program, plan, or infrastructure to turn data science into a core competency that delivers high ROI.	Create a formal data science program or competency center; build a plan that specifies mission, vision, values, key stakeholders, and a roadmap; build a technical infrastructure that meets data science data processing needs.

Data Preparation

#1. Ambiguous Data Semantics

Challenge

Data can be deceptive. A column in a table may not be what it appears to be. For example, it might be labeled "Purchase Date" but actually contain a consumer's birth date. This data could be exploited by an advanced analytics algorithm to derive age, which then could be built into the model to discriminate against older or younger customers.

Misinterpreting the data is a common problem for data scientists. Because data preparation is essential to the creation of great models, it is critical that the lineage and metadata for data fields is correct and current. Data scientists often use new data, so they are often the first to expose flaws in data definitions and semantics. The reality is that the productivity of your data science team is limited by the quality of the data and the discipline of your organization's data governance processes.

Best Practices

To avoid problems with data semantics, organizations should consider the following:

1. Create a data dictionary. Implement a data dictionary so data scientists are less likely to misinterpret the meaning of columns in a database. Your data dictionary should be a process, not a thing, as it will be out of date as soon as you have finished it. Also, consider creating a business glossary that contains definitions of commonly used business terms that can convey business understanding to other members of your agile team.

2. Hire a data librarian. A data librarian is responsible for sourcing relevant data assets (purchasing them if necessary), validating their usefulness, pre-scrubbing them when appropriate, and cataloging them so data scientists can easily find what they need. A data librarian can save data scientists considerable time and help increase model accuracy.

3. Build in redundancy and resiliency. The departure of a data librarian or subject matter expert from the company can significantly undermine the productivity of a data science team. It's important to build redundancy in key areas that might leave a data science team vulnerable. It's also critical to implement data science platforms that execute processes in a consistent manner no matter what people are involved.

4. Create a feature library. During feature engineering, data scientists create fields from existing content to optimize the way specific algorithms work and increase model accuracy. Companies should create a library of derived fields so data scientists can reuse them in other projects, if appropriate. It's not a good sign when derived columns and features are only used by the data scientist who created them.

5. Use data science as a catalyst. Data science is a great way to expose flaws in your data. Make sure your data governance program uses investigations by data scientists to highlight and fix data problems. Leverage your data science initiative as a catalyst to drive change in the way your organization governs data.

#2: Data Drift, Anomalies, and Errors

A model's accuracy depends on the algorithm used to build it and the data that feeds it. Most data scientists agree that the data is more important than the algorithm and prefer superior data and simpler techniques over sophisticated algorithms and limited or noisy data.

The most successful predictive models often result from using new data. Data scientists sometimes reach deep into the data lake, source systems, or external and syndicated data to find hidden gems that can yield huge payoffs. But this data is often very dirty, with missing data elements, inconsistent formats, hidden rules and anomalies, and errors.

Common Challenges

Normalization. Imagine you are building a model to predict the price of a commodity (say, corn). You gather weather data, pricing data from different markets, and purchases of fertilizer. The data comes in from different parts of the world, and some sources may provide updates every minute, while others may provide daily data. Some date fields might be in a format of "year-month-day" and others in "month/day/year." You would need to normalize all of these fields and make them consistent so that the data can be combined before you feed it into your model-building tools.

Missing Data. The fertilizer data might also have a gap where the reporting stores' systems were down. Or an IoT (internet of things) sensor might have malfunctioned and gone silent for several days. Some rationalization would need to be made for the missing data. It could be left blank with the understanding that the machine learning algorithm could navigate around it, or it might be filled in with average values to keep from confusing the algorithm.

Data Drift and Anomaly Detection. Perhaps the historical average for rainfall in July is 2 inches, but data for one month shows 12 inches. This may be an important and rare event, or it may be an anomaly caused by a data entry error. It could also indicate correct data that is changing over time. It can't, however, be automatically corrected. It should be flagged with an automated alerting system and brought to the attention of the data scientists

Best Practices

To clean dirty data, data science teams must take a systematic approach to data cleansing and normalization before data is fed into predictive modeling tools. Here are some tips to ensure that your

data science team uses clean data.

1. Profile data. Data profiling tools show the shape and distribution of your data and make it easy to detect outliers and anomalies, such as data values that are outside historical ranges (e.g., a temperature of 130 degrees Fahrenheit in Boston) or that deviate from a localized norm (e.g., engine vibrations higher than expected for a new engine). Most data preparation tools come with data profiling capabilities.

2. Fix missing values. These data profiling tools can also flag missing values and make it easy for data scientists to impute missing values or delete the data.

3. Treat data as an asset. Build a culture that views the data as a separate asset from the application that generates it. View that data asset as a shared corporate value not owned by the application owner. Hold the application owner accountable for the quality of the data delivered to the enterprise.

4. Create multidisciplinary teams. Build collaborative teams that include data engineers, data scientists, data librarians, domain experts, and program managers that are better prepared to clean, validate, and tag data to foster agility, reuse, and productivity.

Model Development

#3. Poor Model Validation

Common Challenges

It is not uncommon for a predictive model to work spectacularly well in a test environment but fail miserably in the real world. The majority of these failures are caused by mistakes in the construction of training and test data sets. Test data validates results generated by the training data and helps to create a more robust model, so improperly designed training and test data sets can undermine model accuracy.

To understand why model validation is important, consider that models are generally used to predict things. Of course, predicting what you already know is pretty easy. It's harder to predict what will happen in the future or in a new situation. This is why models are trained on one set of historical data and tested on another. The key to model validation is to create distinct training and test sets. Each set draws from the same target database but the data elements are randomly selected. There should be no overlap, duplication, or leakage of records between the two sets.

Real life example. Good model validation practices catch many types of errors. In one case, a data scientist asked his IT department for "random" records to create a "next best offer" model. However, the IT analyst provided records from the first 10% of the database, which was ordered by customer wealth.

As a result, the data scientist created a model based on the behavior of the wealthiest customers, not a random sample. Fortunately, the data scientist caught the error during a small trial prior to launch, saving considerable money from suboptimal recommendations.

Best Practices

1. Check random sampling features. Make sure your data tools support random sampling for both small and large data sets. More importantly, train your data engineers or IT analysts in proper sampling techniques or have them work closely with data scientists when creating training and test data sets.

2. Send alerts. An accomplished data scientist knows when model results are too good to be true, but novice data scientists may not notice. Check to make sure data science workbenches send alerts when a model exhibits the characteristics of being overfit to the training data.

3. Spot temporal leakage. Look for tools that flag variables that are highly correlated with the target variable being predicted. This may indicate that the variable has a portion of the answer encoded into it.

4. Apply cross-validation. Take the time to perform full cross-validation, not just single train-and-test validation. The cross-validation slices the train and test set multiple ways and gives a more robust estimate of accuracy and robustness

5. Conduct a real-world trial. Always test the model in a production environment using live data. Do this by putting the model into production, but don't execute its recommendations. See how the model performs compared to the status quo before moving it into full production. This is also a great way to build trust with senior management.

#4: Regulatory Compliance

Common Challenges

Like young children, machine learning algorithms are ruthless observers: they quickly notice and expose things that are potentially embarrassing or uncivil. For example, if ethnicity is a helpful predictor for loan acceptance or gender correlates with credit scores, then algorithms will find and use that data. Algorithms have no social conscience or understanding of higher societal goals.

Compliance is a huge issue for data scientists. Regulatory agencies penalize organizations whose models exhibit bias towards certain classes of people. Companies that use models to make credit, loan, and other decisions need to demonstrate the rationale used to justify those decisions, putting a premium on model transparency. In addition, privacy laws (e.g., GDPR, HIPAA, FERPA, COPPA) are getting stricter, making it challenging to use personally identifiable information in analytic models.

Best Practices

There are many things organizations can do to ensure compliance with regulations:

1. Detect PII exposure. Look for tools that automatically recognize many different types of personally identifiable information (PII). These tools should alert data scientists to obvious things, such as age, gender, and IP addresses. The tools should flag these fields if a column is mislabeled or the metadata is out of date.

2. Set flags. Your metadata should indicate the privacy level of all fields and flag fields that may not be appropriate to use in a model, such as ethnicity, gender, or age. Tools should also be able to detect if a variable is strongly correlated to a sensitive field, such as a person's neighborhood and ethnicity.

3. Appoint a privacy advocate. Assign someone on the data science team to serve as a privacy advocate. Train this person about industry privacy regulations, the company's privacy policies, and how to spot potential violations.

4. Establish policies. Successful companies have clear enterprise policies governing the proper usage of sensitive data. These companies typically have a chief privacy officer or an active security group that addresses how to handle sensitive data.

5. Abolish private sandboxes. Strictly enforce policies that deter data scientists from downloading data into a private sandbox or home computer. At the same time, make it easy for data scientists to access and analyze new data sources on a corporate server so they aren't tempted to circumvent data and privacy policies.

DevOps

#5: Model Degradation

Common Challenges

No model lasts forever. Most models are effective for a short time. In many cases, the world changes and the models no longer accurately reflect current reality. In other cases, the value of a model declines as competitors exploit the same data sets, variables, algorithms, and techniques.

Models are also cursed by success. If a business acts on a predictive model, the lift from applying the model declines over time. For instance, an organization that proactively reaches out to customers likely to churn with discounts and other incentives may find that the model loses its effectiveness over time because the company has succeeded in reducing churn.

Automated Detection. When a model degrades, not all is lost. Operations may be able to automatically

retrain the existing model or data scientists may need to produce a new one, if events have changed significantly. Either way, the model will need to run through the full testing, and release process again.

The challenge is detecting when a model has degraded enough that it needs to be retrained or rebuilt. This can be difficult at a large company that might have thousands of active models. An automated alerting system is required, as data scientists cannot track each model.

Best Practices

There are many things an organization can do to maintain the accuracy and effectiveness of its predictive models:

1. **Offload data scientists.** As much as possible, push model monitoring to the DevOps team so data scientists can focus on building new models. Organizations with hundreds or thousands of active models need a separate team to monitor the models, otherwise data scientists would never have time to build new models.
2. **Create threshold alerts.** Look for model management tools that automatically alert data scientists when the accuracy of a deployed model falls below a certain threshold. Similarly, look for tools that make it easy to set and store that threshold within metadata associated with the model.
3. **Monitor lift curves and ROI.** Train DevOps personnel to check whether the lift curve and ROI of a model are degrading. Ideally, this is accomplished via automated alerts.
4. **Consider self-modifying models.** There are emerging techniques that enable models to retrain themselves when they go out of date. This practice can save time for data scientists and be highly effective in specific situations such as retraining a forecast model each quarter with new rolling data.
5. **Conduct sensitivity analysis.** Before delivering a model, data scientists should conduct sensitivity analysis by varying input data values and even removing input variables completely. A model that maintains its accuracy when an input field is missing should be considered robust. A model that is overly dependent on small changes in data values or on a particular field may indicate poor performance when in production on new data.

#6: Deployment Disconnect

Common Challenges

Unused Models. More than half of all models that are formally requested by the business (and carefully built and validated by data scientists) are never deployed by the business user who requested them. The primary reason is that it often takes too long to build and deploy the model. By the time the model is ready, the business problem changes or the opportunity passes.

In many organizations, models can take from three to nine months between initial request and delivery

to the business. Only a fraction of this time is spent actually building the model. Often, the business changes its mind as it discovers new patterns in the data, forcing data scientists to revise their model. This can be time consuming, especially once the model is deployed into production, which may require recoding the model in a different language so it can run within the target application. Coding and recoding models makes rapid iteration impossible.

Best Practices

There are many techniques an organization can use to ensure that business gains value from models that are built.

1. **Just say no.** Implement a data science solution where models don't have to be recoded before they are deployed. There are many options where the same system can be used for both model development and model execution. Emerging data science platforms offer APIs that organizations can use to execute models inside a host application. These APIs score incoming records and return results in milliseconds.
2. **Automate recoding.** If models must be built in one system and deployed in another, look for model-building tools that can compile the models and automatically recode them into Java or another language that can be executed on the production database. Rewriting the code by hand is extremely time-consuming and can introduce new errors into the model.
3. **Deployment representative.** Make sure a representative from the DevOps team is part of the agile team delivering the model.
4. **Rationalize options.** Try to bound the number of different tools and environments that are used in your model delivery process. Each new data science tool, language, or algorithm should be carefully considered before it is embraced. As with any decision, the value gained should be balanced against the cost of introducing additional complexity.
5. **Leverage the cloud.** Create a plan for a data science platform that scales to slightly exceed current data sizes. Elastic cloud technologies are a future-proofing boon to your solution.

Business Delivery

#7: Business Disconnect

Data science algorithms are ruthless at exploiting the smallest predictors, but they lack business sense. A model might optimize a local process, but sub-optimize a larger process.

For example, a mobile phone company built an excellent model to predict customer churn. But when it delivered special discounts and offers to its high-risk customers, the churn rate increased rather than

decreased. What happened? The company's outreach campaign to reduce churn inadvertently reminded its dissatisfied customers about their upcoming contract renewal, prompting many of them to cancel their subscription. The model was fine, but needed "parental oversight" to be implemented effectively.

Best Practices

There are many things an organization can do to ensure analytic models deliver business value:

1. **Hire business savvy data scientists.** A data scientist without business domain knowledge can't deliver value. The data scientist won't choose the right questions to answer, select the right variables to model, or know how to interpret the model so the business can understand its implications and how to best apply it.
2. **Hire data science savvy business managers.** Data scientists work best with business people who understand the value and limits of data science and how best to define, manage, and implement data science projects. A good business manager can help guide a data science project to ensure it focuses on business value rather than statistical purity.
3. **Pair novices with experts.** It's best to pair a veteran data scientist with a newbie business manager who doesn't have prior experience working with data scientists. The converse is true too: pair a newbie data scientist with a veteran business manager who has worked successfully with data scientists in the past.
4. **Spend time defining projects.** The tendency among data scientists and business managers is to jump into a project without adequately defining its goals, objectives, and success measures. Without adequate project definition, the project will go astray and begin predicting things of tangential value.
5. **Integrate business rules.** Model scores should be just one part of an overall business decision process and interpreted within the constraints of business rules. For instance, a model may propose a different sales discount for each item in a physical store. While this may be optimal, it can't be implemented and should be overridden by a more intuitive business rule. Look to integrate business rules with model scores into your process and have a preference for tools that support them.
6. **Calculate ROI.** Look for tools that calculate the return on investment (ROI) of your models. This feature helps focus a data science project on what counts and provides ammunition to executives when selling the value of data science projects.

#8: Model Opacity and Repeatability

Common Challenges

In her book Weapons of Math Destruction, Cathy O'Neil gives many examples where algorithms and models impact individuals' lives, but no one understands how they work. Teachers are effectively fired

by algorithms for poor test results, and students are denied admission to college by models that no one understands. Most business people don't trust models unless someone can explain in plain English what the model has discovered and how, and the model results make intuitive sense.

Models vary significantly in their transparency. Neural networks and deep learning are very opaque, while decision-tree algorithms are much easier to understand. Data scientists need to evaluate the need for model opacity when selecting algorithms for building models. Sometimes, they will need to choose models that favor transparency over accuracy.

Audits. In highly regulated industries, such as finance and healthcare, organizations need to audit models. This means that models need to be recreated long after they were built, using the same data, fields, features, parameters, and libraries that were used to create the original. In addition, auditors need to know:

- Who built the model?
- Who modified the model?
- How was the model validated?
- Who else has used the model?
- How old is the model?
- What were the release versions of the libraries and software used?

Best Practices

There are many things organization can do to improve model transparency and auditability.

1. Be ready to reproduce models and scores. In industries where audits are commonplace you will need to be able to explain why and how a model was created and how a particular score was calculated. To do this you'll need to be able to recreate data, metadata and software at the time the model was built or used for scoring. There are emerging tools on the market that automate this process.

2. Choose tools that capture model and data lineage. Look for tools that manage your models like software versions, where you can branch off a model and view its lineage. These systems should also track the data used to build the model and may need to store the data used that is scored by the model. This will imply the need for good master data management, data dictionaries, and other metadata.

3. Trade perfection for transparency. Choose your model-building algorithms with transparency in mind. You may find that you choose a less powerful model in exchange for an explainable model (e.g., a decision tree or nearest neighbor predictive model rather than deep learning or another neural network).

4. Try a hybrid approach. Consider a hybrid, where one model is not transparent but has high

accuracy and another model that does the reverse. For example, some financial institutions use a decision tree to cast a wide net of potential money laundering transactions, and they use a neural network to rank the events within each decision tree node. The decision tree provides explanations, and the neural network minimizes false positives.

5. Support collaboration. Use data science tools that provide for collaboration among data scientists so that each can understand and utilize models created by the others.

#9: Finding Data Scientists

Common Challenges

As mentioned in the introduction, data scientists that can manage every step in the data science lifecycle are hard to find. These so-called unicorns have a PhD in statistics or machine learning; know how to write SQL, Java, and Python; can query, clean, and manipulate data; understand the business domain in which they work; have strong communications skills; and know how to deploy, monitor, and scale projects on servers or in the cloud. It is rare to find one person with two of these skills, let alone all of them. For instance, a statistician who loves math and statistics often finds it difficult to communicate effectively with business users.

Best Practices

As mentioned in the outset, successful companies create a complete data scientist by forming a multidisciplinary team of people with complementary skills. Here are few recommendations for building a data science team.

1. Give up on the unicorn. Recognize that you can't find just one person possessing all of the skills required to be a "data scientist."

2. Build an agile team. Construct an agile team with these team members:

- Data engineer—to structure the data and build the data models
- Data librarian—to understand the data
- Statistician or machine learning expert—to build the models
- DevOps—to deploy and monitor the models
- Privacy expert—to own processes to preserve privacy and stay compliant
- Business analyst—to translate the business needs
- Business ROI owner—to own the overall business results

3. Centralize data scientists. Centralize data scientists on a corporate team so they can collaborate and learn from each other, benefit from training and certification programs, and rotate through various assignments.

4. Align and co-locate data scientists. At the same time, align each data scientist with a functional area

so they learn the domain inside and out. Have them sit with the business several times a week—especially during the design phase of a project—so they become well versed in the business and are part of the planning process, not just an order taker.

5. Create citizen data scientists. Smart analysts who have a proclivity for statistics, but may not have a PhD in statistics or machine learning can be turned into serviceable data scientists through training and mentorship programs. The key is not how much they know, but that they know enough to know their own limitations.

#10: Lack of a Data Science Program

Common Challenges

Most data science teams dive into projects without creating the business or technical infrastructure required to succeed. Without a business plan, the data scientists are journeyman bouncing from one project to the next without much commonality or consistency among projects. Without a technical infrastructure, the data scientists are continually hamstrung by lack of compute, data, and storage services to support the data science lifecycle. Their output is small and delivers minimal business value.

Best Practices

There are many ways organizations can build a data science program that scales and delivers continuous business value. Here are a few.

1. Create a data science program. Make sure you create a program or competency center, staffed by trained experts in the field of data science and supported by a technical infrastructure designed to maximize the productivity of data science teams.

2. Create a business plan. Every data science program starts with a business plan that defines the mission, vision and values of the program. The plan also identifies key stakeholders, success metrics, an organizational reporting structure, optimal methods for engaging with the business (embedded, agile, etc.), and a technical infrastructure. Most importantly, it outlines a roadmap of projects that will be delivered over a period of years.

3. Create a technical infrastructure. Nothing hamstrings a data science team more than lack of an adequate data architecture and technical infrastructure. Fortunately, the cloud provides a perfect arena for data scientists, with unlimited (and cheap) storage, processing, and elasticity.

Summary

Promise. Data science offers tremendous value to organizations that treat it as a core competency and create a formal program with a business plan, technical infrastructure, and adequate sponsorship and funding. Data science promises to help companies move from working reactively with customers and suppliers to proactively. Models can help optimize processes, reducing costs and growing revenues. What executive doesn't want that?

Challenges. But as this report points out, there are many challenges. It's hard to find data scientists that can support the full lifecycle and the quality of data can have an outsized impact on model accuracy and business value. Privacy and compliance regulations may restrict how data scientists interact with the data, while lack of model transparency can make it impossible for business people to trust model output.

As a result, most companies have not yet moved beyond the age of artisans. If your company is still executing data sciences projects with one or a handful of superstar data scientists, you are putting your company at risk. You are also leaving money on the table. It's critical that you begin to automate and operationalize data science to achieve its full benefits. That is what your competitors are doing.

Strategies. The good news is that if you create a bonafide data science program with ample sponsorship, funding, and infrastructure you can achieve sky high ROI. The best programs build teams of people with complementary skills that support the entire data science lifecycle from data preparation to model building to DevOps and model management.

These programs embrace the credo "garbage in is garbage out" so they pay close attention to the quality of the raw material (data) that feeds data science models. They maximize impact by creating agile teams with people who have skills in different functional areas. They centralize their data science teams to foster training, collaboration, and mentorship, and focus on building a platform that supports a long term plan for success.

About the Research Sponsor

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. We deliver the modern platform for machine learning and analytics optimized for the cloud. The world's largest enterprises trust Cloudera to help solve their most challenging business problems. Learn more at www.cloudera.com.

Appendix: Industry Experts

We would like to thank those who contributed their time and expert opinions to this research:

- Abhiram Gujjewar, Informatica
- Baran Durukan, SAP
- Benjamin Baer, FICO
- Bertrand Cariou, Trifacta
- Chad Meley, Teradata
- Dave Duling, SAS
- Fernando Jorge, FICO
- Ingo Mierswa, RapidMiner
- James Serra, Microsoft
- Jonathan Von Rueden, SAP
- James Swanson, Monsanto
- Katherin Rincon, STRIIM
- Kristin Rahn, Pitney Bowes
- Kurt Thearling, Wex
- Mac Steele, Domino Data Lab
- Murthy Mathiprakasam, Informatica
- Naveen Singla, Monsanto
- Pauline Brown, Dataiku
- Richard Mooney, SAP Saurabh Gupta, SAS
- Seth Dobrin, IBM
- Southard Jones, Domino Data Lab
- Sri Raghavan, Teradata
- Steve Belcher, Microsoft
- Steven Hillion, TIBCO
- Steve Sarsfield, MicroFocus
- Steve Sparano, SAS
- Steve Wilkes, STRIIM
- Susara Van Den Heever, IBM
- Ted Fischer, IBM
- Thomas Dinsmore, Cloudera
- Tom Wentworth, RapidMiner
- Veronique Venditti, SAP