**IBM Analytics**
White Paper

# Foundational Methodology for Data Science

In the domain of data science, solving problems and answering questions through data analysis is standard practice. Often, data scientists construct a model to predict outcomes or discover underlying patterns, with the goal of gaining insights. Organizations can then use these insights to take actions that ideally improve future outcomes.

There are numerous rapidly evolving technologies for analyzing data and building models. In a remarkably short time, they have progressed from desktops to massively parallel warehouses with huge data volumes and in-database analytic functionality in relational databases and Apache Hadoop. Text analytics on unstructured or semi-structured data is becoming increasingly important as a way to incorporate sentiment and other useful information from text into predictive models, often leading to significant improvements in model quality and accuracy.

Emerging analytics approaches seek to automate many of the steps in model building and application, making machine-learning technology more accessible to those who lack deep quantitative skills. Also, in contrast to the "top-down" approach of first defining the business problem and then analyzing the data to find a solution, some data scientists may use a "bottom-up" approach. With the latter, the data scientist looks into large volumes of data to see what business goal might be suggested by the data and then tackles that problem. Since most problems are addressed in a top-down manner, the methodology in this paper reflects that view.

## A 10-stage data science methodology that spans technologies and approaches

As data analytics capabilities become more accessible and prevalent, data scientists need a foundational methodology capable of providing a guiding strategy, regardless of the technologies, data volumes or approaches involved (see Figure 1). This methodology bears some similarities to recognized methodologies[1-5] for data mining, but it emphasizes several of the new practices in data science such as the use of very large data volumes, the incorporation of text analytics into predictive modeling and the automation of some processes.

The methodology consists of 10 stages that form an iterative process for using data to uncover insights. Each stage plays a vital role in the context of the overall methodology.

### What is a methodology?

**A methodology is a general strategy that guides the processes and activities within a given domain. Methodology does not depend on particular technologies or tools, nor is it a set of techniques or recipes. Rather, a methodology provides the data scientist with a framework for how to proceed with whatever methods, processes and heuristics will be used to obtain answers or results.**

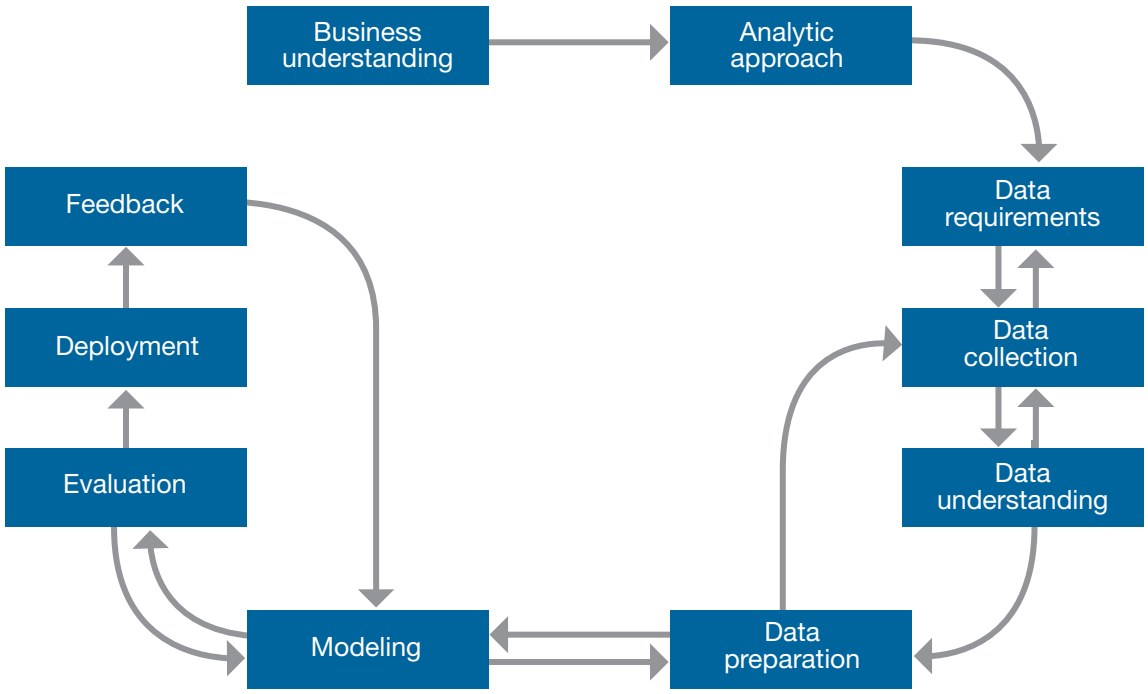


*Figure 1*. Foundational Methodology for Data Science.

### Stage 1: Business understanding

Every project starts with business understanding. The business sponsors who need the analytic solution play the most critical role in this stage by defining the problem, project objectives and solution requirements from a business perspective. This first stage lays the foundation for a successful resolution of the business problem. To help guarantee the project's success, the sponsors should be involved throughout the project to provide domain expertise, review intermediate findings and ensure the work remains on track to generate the intended solution.

### Stage 2: Analytic approach

Once the business problem has been clearly stated, the data scientist can define the analytic approach to solving the problem. This stage entails expressing the problem in the context of statistical and machine-learning techniques, so the organization can identify the most suitable ones for the desired outcome. For example, if the goal is to predict a response such as "yes" or "no," then the analytic approach could be defined as building, testing and implementing a classification model.

### Stage 3: Data requirements

The chosen analytic approach determines the data requirements. Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

### Stage 4: Data collection

In the initial data collection stage, data scientists identify and gather the available data resources—structured, unstructured and semi-structured—relevant to the problem domain. Typically, they must choose whether to make additional investments to obtain less-accessible data elements. It may be best to defer the investment decision until more is known about the data and the model. If there are gaps in data collection, the data scientist may have to revise the data requirements accordingly and collect new and/or more data.

While data sampling and subsetting are still important, today's high-performance platforms and in-database analytic functionality let data scientists use much larger data sets containing much or even all of the available data. By incorporating more data, predictive models may be better able to represent rare events such as disease incidence or system failure.

### Stage 5: Data understanding

After the original data collection, data scientists typically use descriptive statistics and visualization techniques to understand the data content, assess data quality and discover initial insights about the data. Additional data collection may be necessary to fill gaps.

### Stage 6: Data preparation

This stage encompasses all activities to construct the data set that will be used in the subsequent modeling stage. Data preparation activities include data cleaning (dealing with missing or invalid values, eliminating duplicates, formatting properly), combining data from multiple sources (files, tables, platforms) and transforming data into more useful variables.

In a process called *feature engineering*, data scientists can create additional explanatory variables, also referred to as *predictors* or *features*, through a combination of domain knowledge and existing structured variables. When text data is available, such as customer call center logs or physicians' notes in unstructured or semi-structured form, text analytics is useful in deriving new structured variables to enrich the set of predictors and improve model accuracy.

Data preparation is usually the most time-consuming step in a data science project. In many domains, some data preparation steps are common across different problems. Automating certain data preparation steps in advance may accelerate the process by minimizing ad hoc preparation time. With today's high-performance, massively parallel systems and analytic functionality residing where the data is stored, data scientists can more easily and rapidly prepare data using very large data sets.

### Stage 7: Modeling

Starting with the first version of the prepared data set, the modeling stage focuses on developing predictive or descriptive models according to the previously defined analytic approach. With predictive models, data scientists use a *training* set (historical data in which the outcome of interest is known) to build the model. The modeling process is typically highly

iterative as organizations gain intermediate insights, leading to refinements in data preparation and model specification. For a given technique, data scientists may try multiple algorithms with their respective parameters to find the best model for the available variables.

### Stage 8: Evaluation

During model development and before deployment, the data scientist evaluates the model to understand its quality and ensure that it properly and fully addresses the business problem. Model evaluation entails computing various diagnostic measures and other outputs such as tables and graphs, enabling the data scientist to interpret the model's quality and its efficacy in solving the problem. For a predictive model, data scientists use a testing set, which is independent of the training set but follows the same probability distribution and has a known outcome. The testing set is used to evaluate the model so it can be refined as needed. Sometimes the final model is applied also to a validation set for a final assessment.

In addition, data scientists may assign statistical significance tests to the model as further proof of its quality. This additional proof may be instrumental in justifying model implementation or taking actions when the stakes are high—such as an expensive supplemental medical protocol or a critical airplane flight system.

### Stage 9: Deployment

Once a satisfactory model has been developed and is approved by the business sponsors, it is deployed into the production environment or a comparable test environment. Usually it is deployed in a limited way until its performance has been fully evaluated. Deployment may be as simple as generating a report with recommendations, or as involved as embedding the model in a complex workflow and scoring process managed by a custom application. Deploying a model into an operational business process usually involves additional groups, skills and technologies from within the enterprise. For example, a sales group may deploy a response propensity model through a campaign management process created by a development team and administered by a marketing group.

### Stage 10: Feedback

By collecting results from the implemented model, the organization gets feedback on the model's performance and its impact on the environment in which it was deployed. For example, feedback could take the form of response rates to a promotional campaign targeting a group of customers identified by the model as high-potential responders. Analyzing this feedback enables data scientists to refine the model to improve its accuracy and usefulness. They can automate some or all of the feedback-gathering and model assessment, refinement and redeployment steps to speed up the process of model refreshing for better outcomes.

## Providing ongoing value to the organization

The flow of the methodology illustrates the iterative nature of the problem-solving process. As data scientists learn more about the data and the modeling, they frequently return to a previous stage to make adjustments. Models are not created once, deployed and left in place as is; instead, through feedback, refinement and redeployment, models are continually improved and adapted to evolving conditions. In this way, both the model and the work behind it can provide continuous value to the organization for as long as the solution is needed.

## For more information

A new course on the Foundational Data Science Methodology is available through Big Data University. The free online course is available at: http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

For working examples of how this methodology has been implemented in actual use cases, visit:

- http://ibm.co/1SUhxFm
- http://ibm.co/1IazTvG


## Acknowledgements

Thanks to Michael Haide, Michael Wurst, Ph.D., Brandon MacKenzie and Gregory Rodd for their helpful comments and to Jo A. Ramos for his role in the development of this methodology over our years of collaboration.


## About the author


John B. Rollins, Ph.D., is a data scientist in the IBM Analytics organization. His background is in engineering, data mining and econometrics across many industries. He holds seven patents and has authored a best-selling engineering textbook and many technical papers. He holds doctoral degrees in petroleum engineering and economics from Texas A&M University.





IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
June 2015




[1] Brachman, R. & Anand, T., "The process of knowledge discovery in databases," in Fayyad, U. et al., eds., Advances in knowledge discovery and data mining, AAAI Press, 1996 (pp. 37-57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, "Cross Industry Standard Process for Data Mining," http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard, C., Rollins, J., Ramos, J., Perkins, A., Hale, R., Dorneich, A., Milner, E., and Chodagam, J.: Dynamic Warehousing: Data Mining Made Easy, IBM Redbook SG24-7418-00 (Sep. 2007), pp. 9-26.

[5] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, Oct. 28, 2014, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Please Recycle


IMW14828-USEN-00