IBM Big Data & Analytics Hub

Blogs

Why we need a methodology for data science

AUGUST 24, 2015

by John Rollins
Data Scientist, IBM Analytics, IBM

Those who work in the domain of data science solve problems and answer questions through data analysis every day. They build models to predict outcomes or discover underlying patterns, all to gain insights leading to actions that will improve future outcomes. And the tools and technologies used in data analysis are evolving rapidly, enhancing data scientists' abilities to reach their goal.

But a chronic inhibitor to success somehow remains even in such rapid growth. Although many business analysts are moving into data scientist roles, they and the businesspeople whose problems need solving lack sufficient understanding of how to go about solving problems using data science techniques. As a result, they sometimes arrive at solutions that fail to adequately address the problem at hand. For them, the gap is often a failure to understand and then follow a proper methodology for problem solving.

Like traditional scientists, data scientists need a foundational methodology that will serve as a guiding strategy for solving problems. This methodology, which is independent of particular technologies or tools, should provide a framework for proceeding with the methods and processes that will be used to obtain answers and results. I have described such a methodology: the Foundational Methodology for Data Science, depicted in the following diagram. Its 10 stages represent an iterative process leading from solution conception to solution deployment, feedback and refinement.

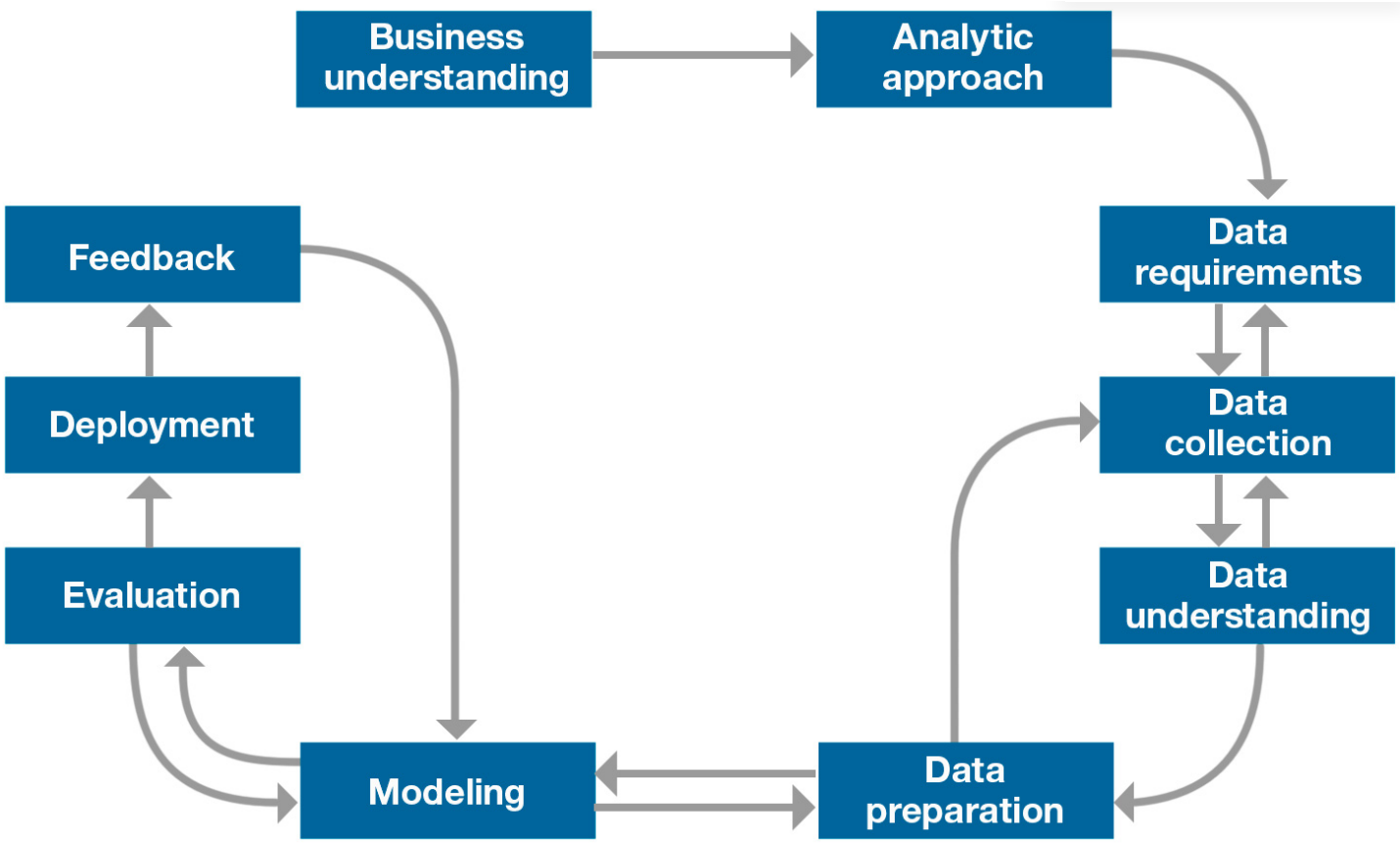

Figure 1. Foundational methodology for data science.

1. Business understanding
Every project, regardless of its size, starts with business understanding, which lays the foundation for successful resolution of the business problem. The business sponsors needing the analytic solution play the critical role in this stage by defining the problem, project objectives and solution requirements from a business perspective. And, believe it or not—even with nine stages still to go—this first stage is the hardest.

2. Analytic approach
After clearly stating a business problem, the data scientist can define the analytic approach to solving it. Doing so involves expressing the problem in the context of statistical and machine learning techniques so that the data scientist can identify techniques suitable for achieving the desired outcome.

3. Data requirements
Choice of analytic approach determines the data requirements, for the analytic methods to be used require particular data content, formats and representations, guided by domain knowledge.

4. Data collection
The data scientist identifies and gathers data resources—structured, unstructured and semi-structured—that are relevant to the problem domain. On encountering gaps in data collection, the data scientist might need to revise the data requirements and collect more data.

5. Data understanding
Descriptive statistics and visualization techniques can help a data scientist understand data content, assess data quality and discover initial insights into the data. A revisiting of the previous step, data collection, might be necessary to close gaps in understanding.

6. Data preparation
The data preparation stage comprises all activities used to construct the data set that will be used in the modeling stage. These include data cleaning, combining data from multiple sources and transform
variables. Moreover, feature engineering and text analytics may be used to derive new s

enriching the set of predictors and enhancing the model's accuracy. The data preparation stage is the most time-consuming. Although I have seen it account for 90 percent of overall project time, that figure is usually more on the order of 70 percent. However, it can drop as low as 50 percent if data resources are well managed, well integrated and clean from an analytical—not merely a warehousing—perspective. And automating some steps of data preparation may reduce the percentage even farther: Members of a telecommunications marketing team once told me that the team had reduced the average time required to create and deploy promotions from three months to three weeks in just this way.

7. Modeling
Starting with the first version of the prepared data set, data scientists use a training set—historical data in which the outcome of interest is known—to develop predictive or descriptive models using the analytic approach already described. The modeling process is highly iterative.

8. Evaluation
The data scientist evaluates the model's quality and checks whether it addresses the business problem fully and appropriately. Doing so requires the computing of various diagnostic measures—as well as other outputs, such as tables and graphs—using a testing set for a predictive model.

9. Deployment
After a satisfactory model has been developed that has been approved by the business sponsors, it is deployed into the production environment or a comparable test environment. Such a deployment is often limited initially to allow evaluation of its performance. Deploying a model into an operational business process usually involves multiple groups, skills and technologies.

10. Feedback
By collecting results from the implemented model, the organization gets feedback on the model's performance and observes how it affects its deployment environment. Analyzing this feedback enables the data scientist to refine the model, increasing its accuracy and thus its usefulness. This often overlooked stage can yield substantial additional benefits if undertaken as part of the overall process.

The flow of this methodology illustrates the iterative nature of the problem-solving process. Models should not be created once, then deployed and left in place unchanged. Instead, through feedback, refinement and redeployment, a model should continually adapt to conditions, allowing both the model and the work behind it to provide value to the organization for as long as the solution is needed.

I will present this methodology as part of the three-day course "Practical Data Science on Spark & Hadoop" at the Strata + Hadoop World conference, which will be held from September 29 to October 1 in New York City. I hope you'll have a chance to attend.

Explore the power of data science and advanced analytics at the IBM Analytics resource page.