The Genesis of EMC’s Data Analytics Lifecycle

When I developed a new Data Analytics Lifecycle for EMC’s Data Science & Big Data Analytics course in 2011, I had no idea the attention it would receive. Although I have been doing analytical work for most of my career, I needed to do considerable research to create a solid process for others to follow. After some preliminary research, I realized that there were surprisingly few existing frameworks for conducting data analytics.

The best sources that I came across were these:

- CRISP-DM**,** which provides useful inputs on ways to frame analytics problems and is probably the most popular approach for data mining that I found.
- Tom Davenport’s **DELTA** framework from his text “Analytics at Work.”
- “MAD Skills: New Analysis Practices for Big Data” provided inputs for several of the techniques mentioned in Phases three to five of my Data Analytics Lifecycle which focus on model planning, execution, and key findings.
- Doug Hubbard’s Applied Information Economics (AIE) approach from his work “How to Measure Anything.” The focus of this work differs a bit from a classic data mining approach. Hubbard’s approach emphasizes estimating and measuring for the purpose of making better decisions. It has some very useful ideas, and helps one understand how to approach analytics challenges from a unique angle and treat them more like decision science problems.
- The Scientific Method**.** Although it has been in use for centuries, it still provides a solid framework for thinking about and deconstructing problems into their principal parts. One of the most valuable ideas of the scientific method relates to forming hypotheses and finding ways to test ideas.

After reading these other approaches to problem solving, I read additional industry articles, and also interviewed multiple data scientists, including several now at Pivotal Data Science Labs, as well as Nina Zumel, a Data Scientist at an independent company, Win-Vector.

This research fueled the creation of a new model for approaching and solving data science or Big Data problems, which is portrayed in this diagram:

This diagram was designed to convey several key points:

1) Data science projects are iterative. Each phase does not represent static stage gates, but reflects the cyclical nature of real-world projects.

2) The best gauge of advancing to the next phase is to ask key questions to test whether the team has accomplished enough to move forward.

3) Ensure teams do the appropriate work both up front, and at the end of the projects, in order to succeed. Too often teams focus on Phases two through four, and want to jump into doing modeling work before they are ready.

I’ve seen people get excited about this approach when taking our data science classes, and have talked about it online, in blogs and even in books. Last year, I co-authored a blog series with EMC Fellow Steve Todd describing how to apply the Data Analytics Lifecycle approach to measure innovation at EMC. This work has been cited many times, both in terms of the project itself (which was mentioned in Business Week) and the methodology, which was highlighted in CRN magazine. In addition, it was also recently featured in Bill Schmarzo’s new book on Big Data.

This Data Analytic Lifecycle was originally developed for EMC’s Data Science & Big Data Analytics course, which was released in early 2012. Since then, I’ve had people tell me they keep a copy of the course book on their desks as reference to ensure they are approaching data science projects in a holistic way.

I’m glad that practitioners, theorists, and readers have found this methodology useful. If you would like to learn more about frameworks for approaching Big Data projects, I’d suggest you check out our EMC Education course materials on Data Science and also review some of the resources mentioned above.