GROUP C

TITLE:

Write a Case study on global innovation network and analysis (GINA). Components of analytic plan are 1.discovery business problem framed, 2.Data, 3.Model planning analytic technique and 4. Result and key findings.

CASE STUDY: GLOBAL INNOVATION NETWORK AND ANALYSIS (GINA)

EMC's Global Innovation Network and Analytics (GINA) team is a group of senior technologists located in centers of excellence (COEs) around the world. This team'scharter is to engage employees across global COEs to drive innovation, research,and university partnerships. In 2012, a newly hired director wanted to improve these activities and provide a mechanism to track and analyze the related information. In addition, this team wanted to create more robust mechanisms for capturing the results of its informal conversations with other thought leaders within EMC, in academia, or in other organizations, which could later be mined for insights.

The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separatedgeographically. It planned to create a data repository containing both structured and unstructured data to accomplish three main goals.

The GINA case study provides an example of how a team applied the Data Analytics Lifecycle to analyze innovation data at EMC.

Phase 1: Discovery

In the GINA project's discovery phase, the team began identifying data sources .Although GINA was a group of technologists skilled in many different aspects of engineering, it had some data and ideas about what it wanted to explore but lacked a formal team that could perform these analytics. After consulting with various experts including Tom Davenport, a noted expert in analytics at Babson College, and Peter Gloor, an expert in collective intelligence and creator of CoIN (Collaborative Innovation Networks) at MIT, the team decided to crowdsource the work by seeking volunteers within EMC.

The data for the project fell into two main categories. The first category represented five years of idea submissions from EMC's internal innovation contests, known as the Innovation Roadmap (formerly called the Innovation Showcase). The Innovation Roadmap is a formal, organic innovation process whereby employees from around the globe submit ideas that are then vetted and judged. The best ideas are selected for further incubation. As a result, the data is a mix of structured data, such as idea counts, submission dates, inventor names, and unstructured content, such as the textual descriptions of the ideas themselves. The second category of data encompassed minutes and notes representing innovation and research activity from around the world. This also represented a mix of structured and unstructured data. The structured data included attributes such as dates, names, and geographic locations. The unstructured documents contained the “who, what, when, and where” information that represents rich data about knowledge growth and transfer within the company. This type of information is often stored in business silos that have little to no visibility across disparate .

Phase 2: Data Preparation

The team partnered with its IT department to set up a new analytics sandbox to store and experiment on the data. During the data exploration exercise, the data scientists and data engineers began to notice that certain data needed conditioning and normalization. In addition, the team realized that several missing datasets were critical to testing some of the analytic hypotheses. As the team explored the data, it quickly realized that if it did not have data of sufficient quality or could not get good quality data, it would not be able to perform the subsequent steps in the lifecycle process. As a result, it was important to determine what level of data quality and cleanliness was sufficient for the project being undertaken. In the case of the GINA, the team discovered that many of the names of the researchers and people interacting with the universities were misspelled or had leading and trailing spaces in the datastore. Seemingly small problems such as these in the data had to be addressed in this phase to enable better analysis and data aggregation in subsequent phases.

Phase 3: Model Planning

In the GINA project, for much of the dataset, it seemed feasible to use social network analysis techniques to look at the networks of innovators within EMC. In other cases, it was difficult to come up with appropriate ways to test hypotheses due to the lack of data. In one case (IH9), the team made a decision to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property. This data collection would enable the team to test the following two ideas in the future.

Phase 4: Model Building

In Phase 4, the GINA team employed several analytical methods. This included work by the data scientist using Natural Language Processing (NLP) techniqueson the textual descriptions of the Innovation Roadmap ideas. In addition, he conducted social network analysis using R and RStudio, and then he developed social graphs and visualizations of the network of communications related to innovation using R's ggplot2 package

Phase 5: Communicate Results

The team found several ways to cull results of the analysis and identifythe most impactful and relevant findings. This project was considered successful inidentifying boundary spanners and hidden innovators. As a result, the CTO officelaunched longitudinal studies to begin data collection efforts and track innovationresults over longer periods of time. The GINA project promoted knowledge sharing related to innovation and researchers spanning multiple areas within the company and outside of it. GINA also enabled EMC to cultivate additional intellectual property that led to additional research topics and provided opportunities to forge relationships with universities for joint academic research in the fields of Data Science and Big Data. In addition, the project was accomplished with a limited budget, leveraging a volunteer force of highly skilled and distinguished engineers and data scientists.

One of the key findings from the project is that there was a disproportionately high density of innovators in Cork, Ireland. Each year, EMC hosts an innovation contest, open to employees to submit innovation ideas that would drive new value for the company. When looking at the data in 2011, 15% of the finalists and 15% of the winners were from Ireland. These are unusually high numbers, given the relative size of the Cork COE compared to other larger centers in other parts of the world. After further research, it was learned that the COE in Cork, Ireland had received focused training in innovation from an external consultant, which was proving effective. The Cork COE came up with more innovation ideas, and better ones, than it had in the past, and it was making larger contributions to innovation at EMC.

It would have been difficult, if not impossible, to identify this cluster of innovators through traditional methods or even anecdotal, word-of-mouth feedback. Applying social network analysis enabled the team to find a pocket of people within EMC who were making disproportionately strong contributions. These findings were shared internally through presentations and conferences and promoted through social media and blogs.

Phase 6: Operationalize

Running analytics against a sandbox filled with notes, minutes, and presentations from innovation activities yielded great insights into EMC's innovation culture.

Key findings from the project include these:

• The CTO office and GINA need more data in the future, including a marketing initiative to convince people to inform the global community on their innovation/research activities.

• Some of the data is sensitive, and the team needs to consider security and privacy related to the data, such as who can run the models and see theresults.

In addition to running models, a parallel initiative needs to be created to improve basic Business Intelligence activities, such as dashboards, reporting, and queries on research activities worldwide.

• A mechanism is needed to continually reevaluate the model after deployment. Assessing the benefits is one of the main goals of this stage, as is defining a process to retrain the model as needed. In addition to the actions and findings listed, the team demonstrated how analytics can drive new insights in projects that are traditionally difficult to measure and quantify. This project informed investment decisions in university research projects by the CTO office and identified hidden, high-value innovators. In addition, the CTO office developed tools to help submitters improve ideas using topic modeling as part of new recommender systems to help idea

submitters find similar ideas and refine their proposals for new intellectual property.

Innovation is an idea that every company wants to promote, but it can be difficult to measure innovation or identify ways to increase innovation. This project explored this issue from the standpoint of evaluating informal social networks to identify boundary spanners and influential people within innovation subnetworks. In essence, this project took a seemingly nebulous problem and applied advanced analytical methods to tease out answers using an objective, fact-based approach. Another outcome from the project included the need to supplement analytics with a separate datastore for Business Intelligence reporting, accessible to search innovation/research initiatives. Aside from supporting decision making, this will provide a mechanism to be informed on discussions and research happening worldwide among team members in disparate locations

Summary

This chapter described the Data Analytics Lifecycle, which is an approach to managing and executing analytical projects. This approach describes the process in six phases.

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Through these steps, data science teams can identify problems and perform rigorous investigation of the datasets needed for in-depth analysis. As stated in the chapter, although much is written about the analytical methods, the bulk of the time spent on these kinds of projects is spent in preparation—namely, in Phases 1 and 2 (discovery and data preparation). In addition, this chapter discussed the seven roles needed for a data science team. It is critical that organizations recognize that Data Science is a team effort, and a balance of skills is needed to be successful in tackling Big Data projects and other complex projects involving data analytics