CHAPTER 1: INTRODUCTION

GIN (Global Innovation Network) is a globally organized network of interconnected and integrated functions and operations by firms and non-firm organizations engaged in the development or diffusion of innovations. The network focusses to engage senior technologists as well as firms across all global COEs (Centres Of Excellence) to drive innovation, research and university partnerships.

Innovation is typically a difficult concept to measure, and this team wanted to look for ways to use advances analytical methods to identify key innovators within the company. The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among the geographically separated GINA members.

CHAPTER 2: OBJECTIVE

GINA planned to create a data repository containing both structured and unstructured data to accomplish three major goals:

1. Store formal and informal data
2. Track research form global technologists
3. Mine the data for patterns and insights to improve the GINA's operations and strategy

CHAPTER 3: IMPLEMENTATION

3.1. Phase 1: Discovery

3.1.1. Phase description:

In this phase, the data science team performs following actions:

1. Learn about problem.
2. Investigate the problem.
3. Develop context and understanding.
4. Examine the available sources of data.
5. Frame the initial hypothesis.
6. Perform a feasibility analysis for the project problem statement.

At the end of this phase, the objectives of the project are precisely established and success criteria is clearly defined.

3.1.2. Stakeholders involved in this phase:

1. Business user, project sponsor, representative of CTO office.
2. BI Analyst
3. Data Engineer and DBA Administrator
4. Data Scientist

3.1.3. GINA Implementation:

After the complete data collection and discovery process, there were 2 major categories of project that were defined which were:

1. Innovation Roadmap to be followed.
2. Data that has been gathered which represents research and innovation activities from around the world.

Hypothesis of Phase 1:

1. Descriptive Analytics of what is currently happening to spark further creativity, collaboration and asset generation.
2. Predictive Analysis to advertise executive management of where it should be investing in future.

3.2. Phase 2: Data Preparation

3.2.1. Phase description:

In this phase, the data science team performs following actions:

1. Setting up an isolate workspace.
2. Performing ETL processes (Extract, Transform and Load).
3. Learning basics about the data.
4. Data conditioning is performed since the data from multiple sources can be joined together which would result in data redundancy.
5. Data Visualization to get an idea of basic characteristics of data.

3.2.2. GINA Implementation:

It was observed by the data scientists and data engineers that certain data needed conditioning and normalization.

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. Hence the data preparation techniques were performed to avoid any misleading results in upcoming phases of the project.

3.3. Phase 3: Model Planning

3.3.1. Phase description:

In this step, the team explores the data and evaluates the possible data models that can be implemented for the data. The team can also try multiple models before finalizing one of them.

The 2 main activities performed in this phase are:

1. Data Exploration: Gaining in depth knowledge of data by understanding all the attributes and complex relationships between them.
2. Model Selection.

3.3.2. GINA implementation:

The team made a decision to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property.

The parameters related to the scope of the study included the following considerations:

1. Identify the right milestones to achieve this goal.
2. Trace how people move ideas from each milestone toward the goal.
3. Once this is done, trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.
4. Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled). These could be as simple as t-tests or perhaps involve different types of classification algorithms.

3.4. Phase 4: Model Building

3.4.1. Phase description:

In this phase, the team actually starts to build the analytical model. The available datasets are divided into: training dataset, testing dataset and production dataset.

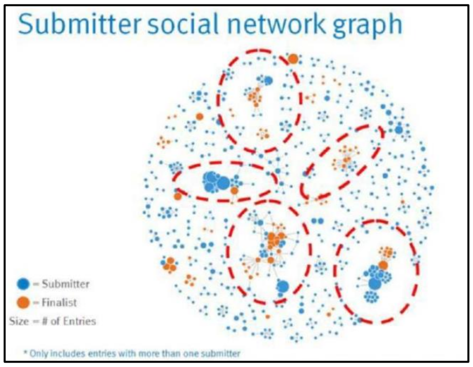
1. The training dataset is used to train the model for different values.
2. The testing dataset is used to test, whether the model is able to predict the values accurately or not.
3. And finally, the production dataset is used to in the production environment which is as good as the deployment environment.

The performance of the model for all the datasets is catalogued and documented.

3.4.2. GINA implementation:

The GINA team employed several analytical methods. This included work by the data scientist using Natural Language Processing (NLP) techniques on the textual descriptions of the Innovation Roadmap ideas.

The figure below, shows social graphs that portray the relationships between idea submitters within GINA.


Submitter social network graph
● = Submitter
● = Finalist
Size = # of Entries
* Only includes entries with more than one submitter

Each colour represents an innovator from a different country. The large dots with red circles around them represent hubs. A hub represents a person with high connectivity and a high "betweenness'" score.

The team used Tableau software for data visualization and exploration and used the Pivotal Greenplum database as the main data repository and analytics engine.

3.5. Phase 5: Communicate Results

3.5.1. Phase description:

1. After building, testing and executing the model, the team compares the model performance with the pre-establishes success criteria.
2. The team then articulates the findings and documents the results.
3. These findings are communicated to the project stakeholders.
4. It is also possible that the team faces failure in the Model Building phase, but the results are always documented and conveyed to the stakeholders.

3.5.2. GINA implementation:

This project was considered successful in identifying boundary spanners and hidden innovators.

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.

Study was successful in identifying hidden innovators such as that in Cork, Ireland.

3.6. Phase 6: Operationalize

3.6.1. Phase description:

1. In this phase, the model is deployed in a staging environment similar to the production environment. This is done to perform a last check before actually deploying it in production. If any changes are required, they they are made and the model is tested again.
2. The project outcome is shared with: Business User, Project Sponsor, Project Manager, BI (Business Intelligence) Analyst, Database Administrator, Data Engineers and Data Scientists.

3.6.2. GINA implementation:

Key findings:

1. Need more data in future.
2. Some data were sensitive.
3. A parallel initiative needs to be created to improve basic Business Intelligence activities.
4. A mechanism is needed to continually re-evaluate the model after deployment.