GINA Case Study

- The GINA case study provides an example of how a team applied the Data Analytics Lifecycle to analyse innovation data at EMC.
- Innovation is typically a difficult concept to measure, and this team wanted to look for ways to use advanced analytical methods to identify key innovators within the company.
- GINA is a group of senior technologists located in centres of excellence (COES) around the world.
- The GINA Team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically

It planned to create data repository containing both structured and unstructured data to accomplish three main goals

1) Store formal data and informal data
2) Track research from global technologists.
3) Mine the data for patterns and insights to improve the team's operation and strategy

Phases

1) Phase 1: Discovery

In the GINA Project's discovery phase, the team began identifying data sources

Following person are involved in this phase

1) Business user, project sponsor, project manager – Vice president from office of CTO
2) BI Analyst – person from IT
3) Data engineer and DBA – people from IT

4) Data scientist – distinguished engineer

The data for the project fell into two main categories.

1) Innovation Roadmap
2) data encompassed minutes and notes representing innovation and research activity from around the world

Hypothesis

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

2) Phase 2: Data Preparation

- IT department to set up a new analytics sandbox to store and experiment on the data. The data scientists and data engineers noticed 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.
- Important to determine what level of data quality and cleanliness was sufficient for the project being undertaken.

3) Phase 3: Model Planning

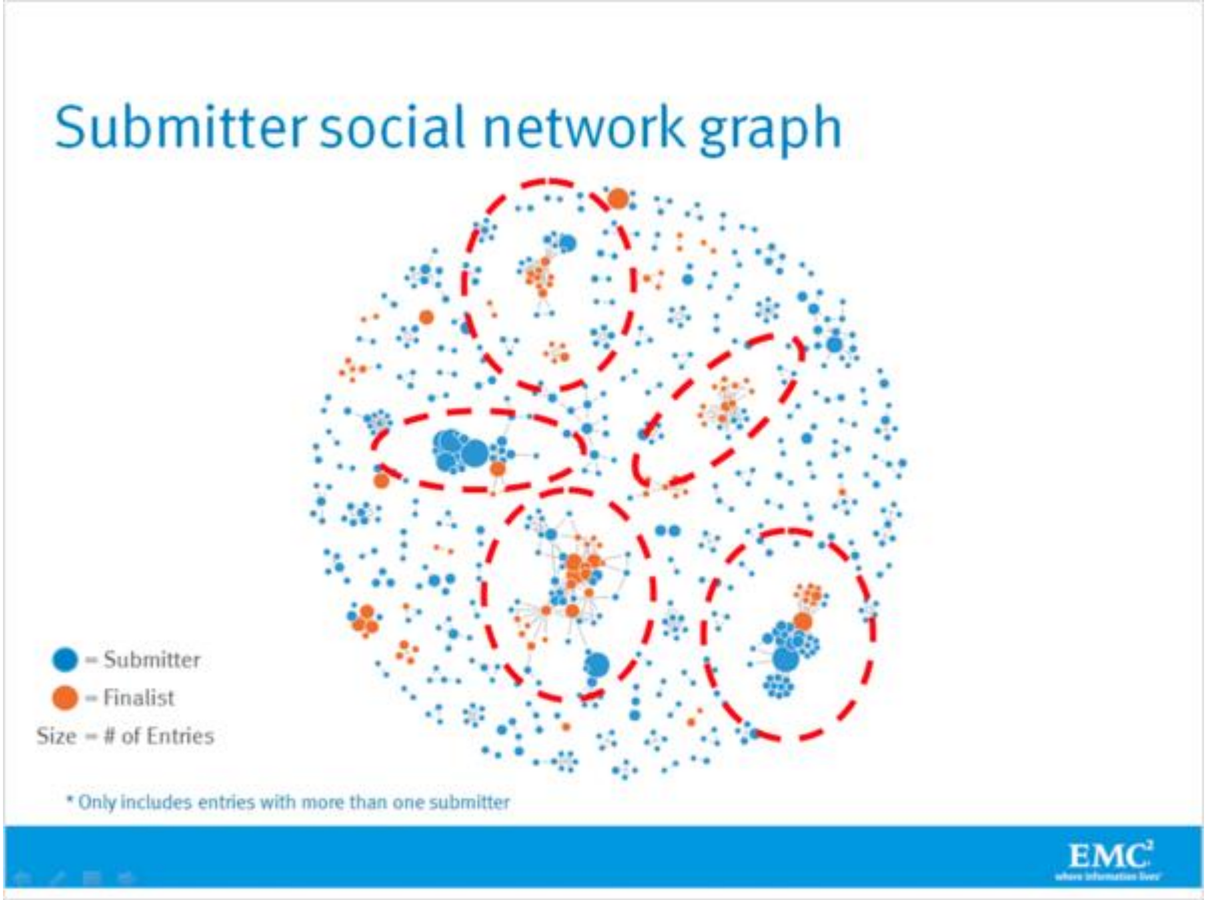
The team decided to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property.

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.

4) Phase 4: Model Building

- The GINA Team employed several analytical methods. This included work by the data scientist using natural language



Submitter social network graph

* Only includes entries with more than one submitter

processing (NLP) techniques on the textual descriptions of the innovation roadmap ideas.

- Fig shows social graphs that portray the relationships between idea submitters within GINA.
- 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.

5) Phase 5: Communicate Results

- 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.
- The study was successful in identifying hidden innovators. Found high density of innovators in Cork, Ireland
- The CTO office launched longitudinal studies

6) Phase 6: Operationalize

Deployment was not really discussed

Key findings:

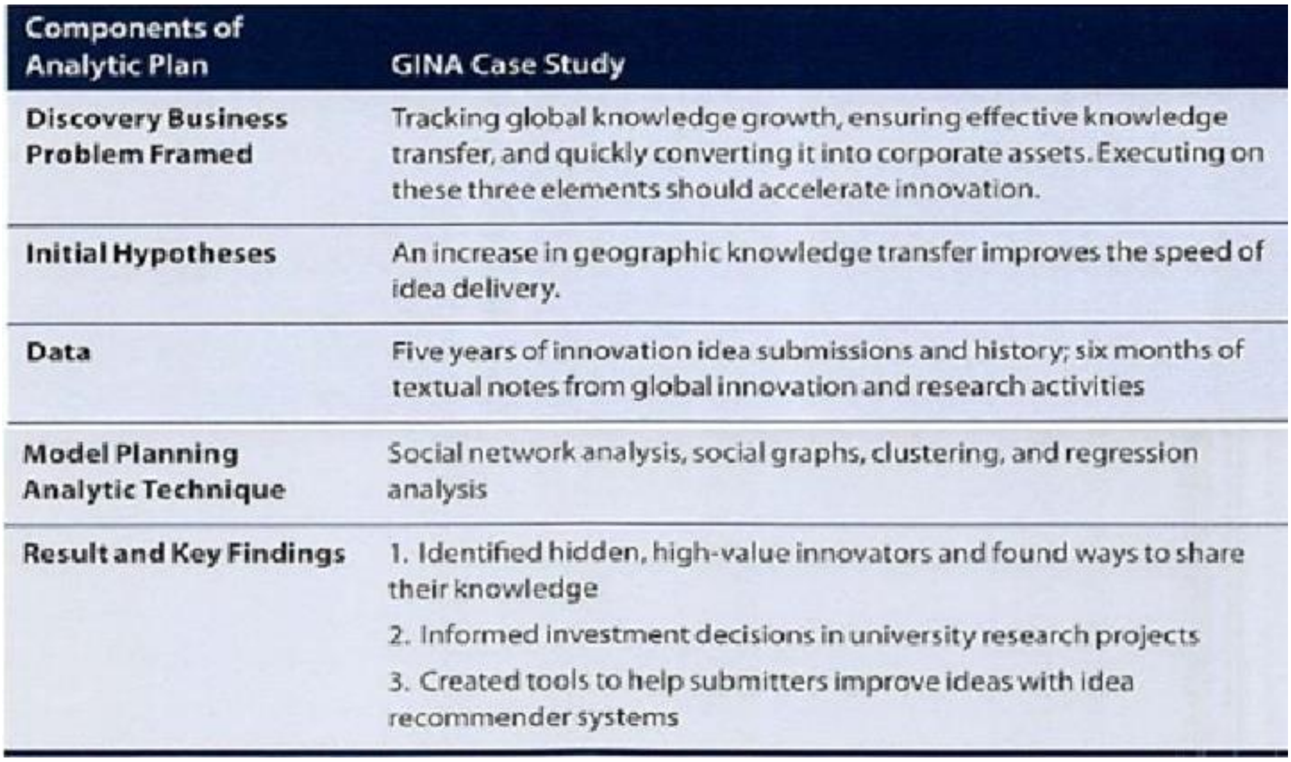
1) Need more data in future

2) Some data were sensitive

3) A parallel initiative needs to be created to improve basic Bl activities

4) A mechanism is needed to continually revaluate the model after deployment

Analytic Plan from EMC GINA Project



Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems