# A Strategy for Innovation Analytics

Using analytics to improve innovation processes is, well, innovative in and of itself.

I spoke with [Tom Davenport](http://www.tomdavenport.com/index.html) last week about using analytics to understand innovation. While my own company (EMC) and other corporations use analytics across many different parts of the business,Tom and I agreed that it is unique to use data scientists to gain insight into innovation processes.

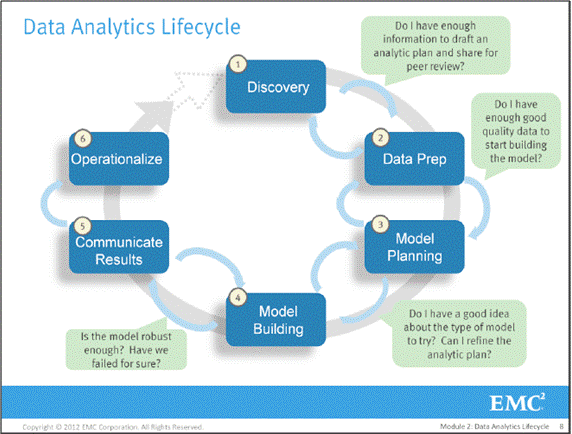
So how exactly is it done?

I have received requests for more detail on the approach that we are using. especially from managers that are unfamiliar with the field of analytics.

In response to these requests I've decided to write a series of posts that describe the evolution of my own experience applying analytics to innovation data. These posts will roughly fall into three categories:

1. A description of the data that my team collects internally.  Most of my posts have focused exclusively on idea submissions from employees, but the larger repository contains innovation and research data beyond just idea submissions.
2. The training that I underwent to manage my own personal team of Big Data Scientists.  This training contains a six-step process that I undertake to gain innovation insight into the data we have collected over the years.
3. My ongoing use of social media to accelerate the collection of innovation and research data world-wide, and the specifics of the approach that I take to motivate teams of "volunteer" data scientists at worldwide locations.

If you are interested in doing some pre-reading along these lines, I recommend taking a look at the posts that I have [already written](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) on the topic. I also recommend taking a [deeper look at the training](https://education.emc.com/guest/campaign/data_science.aspx) that I undertook to increase my own expertise in this area. The diagram below depicts the approach that I use for gaining analytic insight.



Please leave me a note in the comments section if there are specific topics you'd like to see explored along the way.

As Bill Schmarzo put it, be prepared to embark on [a most excellent journey](http://infocus.emc.com/william_schmarzo/bill%E2%80%99s-most-excellent-data-scientist-adventure/).

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 04:50 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/03/a-strategy-for-innovation-analytics.html)

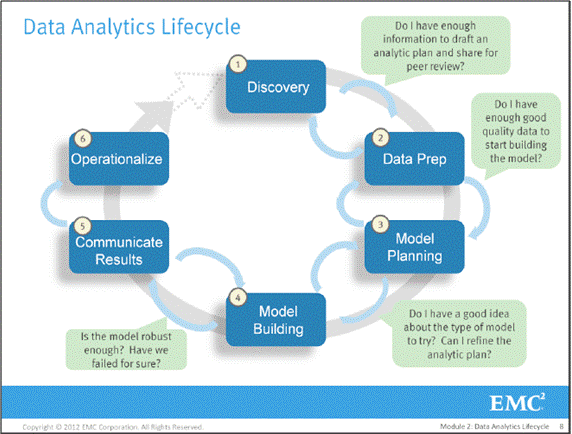
Tags: emc data scientist curriculum big data

March 05, 2012

Phase 1 Innovation Analytics

How do you analyze innovation at a large multi-national corporation?

I've been using the Data Analytics Life Cycle (depicted below):



These steps have helped me to internally construct a strategy for analyzing global innovation processes and methodologies at my own corporation (EMC). I've [published some of the results](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) of this effort in previous posts.

The analytic lifecycle was something I learned while attending the new [Data Science and Big Data Analytics course](http://www.emc.com/about/news/press/2011/20111205-01.htm).  Who came up with these steps? They are essentially an overview of industry best practices and experiences as summarized by EMC Education Services (for a deeper dive into the course content, register and take a look [here)](https://education.emc.com/guest/campaign/data_science.aspx).

The business driver that caused me to attend the course was my belief that analytics could help me discern promising new opportunities in my role as Director of the EMC Innovation Network.

The first thing I learned at the course is that I am not a data scientist. Any successful analytics project has a set of key players. In addition to data scientists, key roles include project sponsors and managers, DBAs, data engineers, and business intelligence analysts. My role is essentially that of a *business user*: someone who consults and advises on how to operationalize the end result of the analytic exercise.

**The Discovery Phase**

Step number 1 in the analytics lifecycle is all about the business domain. Here are some of the key activities that are critical in this phase:

* Frame the business problem as an analytic challenge that can be solved in phases.
* Understand what's been done in the past.
* Assess the resources supporting the project (people, technology, time, and data).
* Form initial hypotheses.
* Determine readiness to move to the next phase.

**Framing the Business Problem**

My company has 50,000+ globally distributed employees, many of whom innovate on a daily basis. The main business problem, from an innovation standpoint, is to ensure that we have an innovation pipeline that continually introduces new revenue sources and cost improvements. Innovation is the lifeblood of the company, especially in the field of high-tech.

The most important word in EMC's innovation lexicon is *knowledge*.

The mission of our EMC Innovation Network is to (a) expand knowledge locally, (b) transfer it globally, and (c) leverage it strategically.  The continual introduction of new revenue sources and cost improvements comes down to leveraging new knowledge that has been transferred and shared across our global employee base.

Our company needs to analyze the expansion, transfer, and leverage of knowledge. The insight gained from this process will improve the innovation pipeline (one of the hypotheses that I will expand upon in future posts).

**Understanding What's Been Done in the Past**

When it comes to a repository for innovation data, EMC has a five-year history of global ideas submitted by employees. The repository amounts to roughly 6000 ideas.

In addition to the idea repository, each business unit has their own repository and collaboration site describing innovation and research activities specific to their business.

As I mentioned in a [previous post](https://stevetodd.typepad.com/my_weblog/2012/03/a-strategy-for-innovation-analytics.html), the idea of running analytics across this type of data is a fairly novel approach. We've already analyzed year-over-year idea submission totals, with an emphasis of the geographic location of the submitter. Other than that, there has been no previous attempt to analyze knowledge expansion and transfer on a global scale. This realization surfaces a clear pain point: any future analysis would require some sort of centralization of global knowledge activities related to innovation. This meant that our project would likely be phased. The team would start with the idea repository and focus on measuring research and innovation activities in a later phase.

**Assessing the Resources**

The resources required to run this analytics initiative is a good news/bad news situation.

The good news is that I work in the CTO Office and my team has an excellent lab with excellent lab managers. Compute/network/storage resources are not an issue.

The bad news is that I need two teams of people to help with this project and none of them report into my organization:

1. I need sponsoring organizations, managers, and high-level DBAs to help with the initial phases.
2. I need data engineers and scientists to execute the low-level work.

I've solved this problem by forming two global, volunteer teams within my own company.  I met with MIT Professor Peter Gloor to ask his advice on how to motivate globally distributed teams outside of my functional organization. [He gave me great advice that worked](https://stevetodd.typepad.com/my_weblog/2011/06/the-flip-side-of-the-coin.html).

Every other Tuesday morning I meet with Team #1: the steering committee for the project.  On alternating Tuesdays I meet with Team #2: data scientists-in-training! These teams are distributed throughout the U.S., China, India, Israel, Egypt, Russia, and Europe.

**Initial Hypotheses and The Analytic Plan**

This next sentence is essentially the jewel of the course:

*The hypotheses and analytic plan form the foundation of everything that comes after it.*

Before taking the course, my hypotheses fell into two buckets:

1. Descriptive analytics of what is currently happening in my organization will spark further creativity, collaboration, and asset generation.
2. Predictive analytics will advise executive management of where it should be investing next.

After taking the course, I realized that I need to spend much more time a more comprehensive set of hypotheses and a formalized analytic plan.

I will be diving into each one of these in future posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 07:39 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics.html)

Tags: emc innovation analytics big data science curriculum

Phase 1 Innovation Analytics: Hypothesis Generation

Diagrama

Descrição gerada automaticamente

In this [series](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) I have introduced the concept of a Data Analytics lifecycle and began to explain how it guides the analysis of innovation at my corporation (EMC).  
As with any lifecycle, the first phase (Discovery) lays the foundation for the rest of the steps. Two of the key activities in the Discovery phase are (a) the creation of hypotheses, and (b) the creation of an analytic plan. In this post I will introduce relevant hypotheses; my next post will dive into with the analytic plan.  
I found it useful to ask myself the following question:  
*Given a global repository representing employee ideas, discussions, minutes, and notes about innovation and research activities, what can I measure?  What measurements can accelerate corporate innovation world-wide?*

I found the exercise below to be the most difficult part of the [curriculum](https://education.emc.com/guest/campaign/data_science.aspx). It is clearly the most relevant, however. Business theories about data must turn into statements that can be proved or disproved.

In this phase of the analytics lifecycle, the curriculum encourages business users to generate as many relevant hypotheses as possible. All of these can help guide data scientists in further phases. Each statement below starts with a "stream of consciousness" on the business problem, and ends with a specific hypothesis that data scientists can either prove or disprove.

***Hypothesis #1: Local Measurement of Innovation Activity***  
  
I believe that innovation can be measured for a given geography. This measurement can take a number of forms, including number of participants, percentage of time dedicated to innovation, local idea-to-implementation timeframes, and geographic reach of innovators (how far outside the workplace does innovation activity extend)?  As the Director of a Global Innovation Network, I'd like to know how these activities map to corporate strategy.

*IH1: Innovation activity in different geographic regions can be mapped to corporate strategic directions.*  
  
***Hypothesis #2: Geographic Innovators***  
  
In every locale world-wide there are typically a set of people who pursue innovation with passion and consistency.  Their contributions may be hidden from the corporate eye. They also may be focused on particular activities, such as idea contests, visits to customers/partners, frequent visitors to a university, or the generation of intellectual property.  If they have not yet taken the initiative to extend their visibility to the global stage, I believe that they can be found via analytics and connected to relevant knowledge sources.  I believe their skill in the delivery of ideas would be improved. For this topic there are two hypotheses to prove:

*IH2a: The length of time it takes to deliver ideas decreases when global knowledge transfer occurs as part of the idea delivery process.*

*IH2b: Innovators that participate in global knowledge transfer deliver ideas more quickly than those that do not.*

***Hypothesis #3: Effective Ideators***  
  
When it comes to idea generation, some employees have an advanced ability to suggest ideas that resonate with the decision makers. Rarely are their ideas dismissed outright. They often (but not always) have a track record of idea delivery as well.  I believe that analytics can help me find these people. I also believe that the form of their ideas can be analyzed to reveal clues as to why their ideas are likely to be funded. In turn, the format of any idea submission can be evaluated for its value.

*IH3: An idea submission can be analyzed and evaluated for the likelihood of receiving funding.*  
  
***Hypothesis #4: Geographic Knowledge***  
  
Very often certain geographies have a reputation for excellence in a certain area of knowledge.  I believe that analytics will reveal that this knowledge can be found in other locales as well (or conversely it may not be found where it was assumed to be).  In general, different geographies will likely reveal themselves to be hubs of expertise in any number of areas. Knowing this fact would not only facilitate the matching of problems to local innovators in a certain region, it also may provide the opportunity to join different locales together for problem-solving exercises.

*IH4: Knowledge discovery and growth for a particular topic can be measured and compared across geographic regions.*  
  
***Hypothesis #5: Knowledge transfer facilitation via boundary spanners***  
  
There are certain employees that have arisen within a geography and made connections with other geographies for the purpose of collaboration. They may not have high visibility within a corporation aside from the direct connections that they have made on their own.  I believe that not only can analytics identify these people, but it can also classify the type of knowledge that these individuals are transferring. These "boundary spanners" can be targeted and trained as "innovation facilitators" and united at a corporate level.

*IH5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.*  
  
***Hypothesis #6 Corporate research gaps and assignments***  
  
A corporate research roadmap needs a portfolio of initiatives to go along with it.  I believe analytics can enable a dashboard view of particular strategic initiatives (e.g. cloud computing) and determine how much research activity (if any) is occurring across the corporation. This view can also be extended to profile funding activity on particular themes. I also believe that analytics can recommend the best place to perform research as new items are added to corporate research roadmaps.

*IH6: Strategic corporate themes can be mapped to geographic regions.*

***Hypothesis #7 Incubation Lineage and Asset Generation***

I believe that the path that knowledge takes, from a local innovator, to a corporate boundary spanner, to an implementation team, to a delivered asset, can be traced and measured. I also believe that this measurement, once studied, can reveal ways to accelerate innovation and point out areas of knowledge that are yet to be converted. I've long been a fan of provenance, and I love the concept of "idea lineage". The lineage can be studied to reduce asset delivery time.

*IH7a: Frequent knowledge expansion and transfer events reduce the amount of time it takes to generate a corporate asset from an idea.*

*IH7b: Lineage maps can reveal when knowledge expansion and transfer did not (or has not) result(ed) in a corporate asset.*

***Hypothesis #8 New areas of innovation and research***  
  
Finally, I believe that predictive analytics can reveal areas of focus for future innovation, research, and investment. What knowledge should be expanded? Who should collaborate on that theme? What kind of assets could result?

*IH8: Emerging research topics can be classified and mapped to specific ideators, innovators, boundary spanners and assets.*  
  
If I were to sum up the list above into one hypothesis, it would look something like this:

***An increase in geographic knowledge transfer improves the speed of idea delivery.***  
  
In my next post I will describe how this hypothesis can be integrated into an analytic plan.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 04:57 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html)

Tags: emc data science and big data curriculum analytics lifecycle hypothesis hypotheses

Phase 1 Innovation Analytics: Creating the Analytics Plan

Diagrama

Descrição gerada automaticamente

The finishing touch for Phase 1 of the Data Analytics lifecycle is the creation of an Analytic Plan. In the same way that requirements drive all phases of a software project, the analytic plan lays the foundation for all of the work in an analytics project.

I've mentioned in previous posts that this part is not easy. Analytic Plans are new to me. Before starting I need to give credit where credit is due. [David Dietrich](https://www.twitter.com/imdaviddietrich) has been a driving force behind our[Data Science and Big Data Analytics curriculum](https://education.emc.com/guest/campaign/data_science.aspx), and a regular contributor to this [series of articles](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/).

There are four initial components of an Analytic Plan:

1. Framing of the Business Problem

I am trying to accelerate innovation within my corporation (EMC). Three problems faced by the corporation are (a) the tracking of knowledge growth throughout our global employee base, (b) ensuring that this knowledge is effectively transferred within the corporation, and (c) that this knowledge is most effectively converted into corporate assets. Executing on these three elements more effectively should accelerate innovation, which is the lifeblood of our company.

2. Initial Hypothesis

In my [last post](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) I described eight different initial hypotheses theorizing how analytics can assist in solving the business problem. These eight hypotheses were boiled down to one high-level hypothesis statement:

***An increase in geographic knowledge transfer improves the speed of idea delivery.***

This hypothesis paves the way for what data we will need and what type of analytic methods we will likely use.

3. Data

The data that the project will rely on fall into two categories.

1. The first category represents five years' worth of idea submissions into EMC's Innovation Showcase process. The Showcase process is a formal, organic innovation process whereby employee ideas from around the globe are submitted, vetted, judged, and incubated. The data is a mix of structured (idea counts, submission dates, inventor names) and unstructured (the ideas themselves) content.
2. The second category encompasses minutes and notes representing innovation and research activity from around the world. This data is also a mix of structured and unstructured. The structured data, once again, includes items such as dates, names, and geographic location. The unstructured documents contain the "who, what, when, and where" information that represents rich data about knowledge growth and transfer within the company. This type of information, however, is often stored in business silos that have little to no visibility across disparate research teams.

The first repository (the idea submissions) is centralized. The second data set (centralized research and innovation minutes/notes) will be gathered from throughout the corporation and contain 6 months worth of global data.

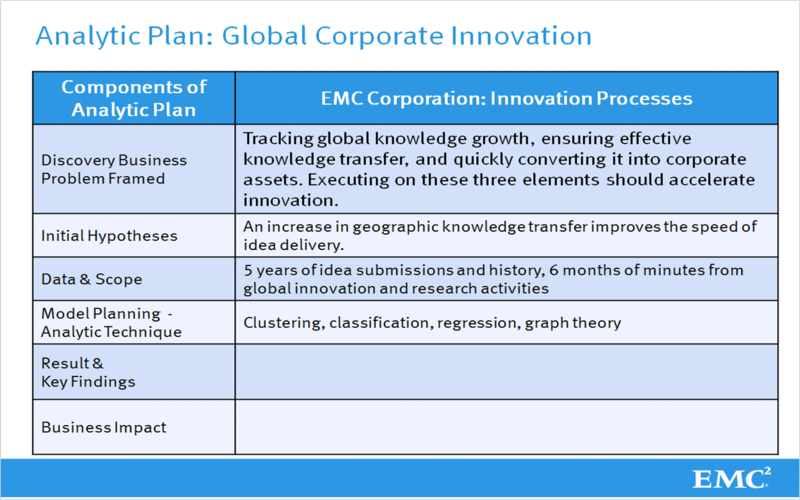
4. Model Planning - Analytic Technique

Model Planning represents the conversion of the business problem into a data definition and a potential analytic approach. In other words the rubber is beginning to hit the road in terms of creating algorithms. A model contains the initial ideas on how to frame the business problem as an analytic challenge that can be solved quantitatively. There is a strong link between the hypotheses and the analytic techniques that will eventually be chosen. Described below are a few algorithms and approaches that make sense given the hypotheses. They do not represent a complete list but give the reader a sense for this activity within the analytic plan.

Keep in mind that model selection is an "art form". Some people are better at it than others. It requires iteration and overlap with phase 2 (Data Prep). Multiple types of models are applicable to the same business problem.  Selection of methods can vary depending on the experience and comfort zone of the Data Scientist. In other cases model selection is more strongly dictated by the problem set.

* Use Map/Reduce for extracting knowledge from unstructured documents. At the highest level, Map/Reduce imposes a structure on unstructured information by transforming the content into a series of key/value pairs. Map/Reduce can also be used to establish relationships between innovators/researchers discussing the knowledge.
* Natural language processing (NLP) can extract "features" from documents, such as strategic research themes, and can store them into vectors.
* After vectorization, several other techniques would be appropriate:
  + Clustering (e.g. k-means clustering) can find "clouds" within the data (e.g. create 'k' types of themes from a set of documents).
  + Classification can be used to place documents into different categories (e.g. university visits, idea submission, internal design meeting).
  + Regression analysis can focus on the relationship between an outcome and its input variables. What happens when an independent variable changes?  It can help in predicting outcomes. This could suggest where to apply resources for a given set of ideas.
  + Graph theory (e.g. Social Network Analysis) will be an important way to establish relationships between employees who are submitting ideas and/or collaborating on research.

At this point I have generated some hypotheses, described potential data sets, and chosen some potential models for proving or dis-proving the hypotheses.  During this process I have been sharing my thoughts in bits and pieces with my peers, and I feel confident that I have enough data to draft a high-level analytic plan and submit it for formal review. I've attached a template slide below.



The last two rows in the Analytic Plan overview (Results & Key Findings, Business Impact) are a reminder to me that I am working toward Step 5 of the Analytic Lifecycle: Communicate the Results. As the business user I participate most heavily in the beginning and the end of the Lifecycle.

I've spent a lot of time on this first step. Any analytic project lead should do the same. With the Analytic Plan as the foundation, it's time to move on to Step 2: Data Prep.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 07:31 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html)

Tags: emc big data analytics analytic plan lifecycle

Phase 2 Innovation Analytics: Data Preparation

Diagrama

Descrição gerada automaticamente

There is a massive amount of innovation and research data globally distributed amongst the 50,000+ employees at my corporation (EMC). In this current [series of blog posts](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) I've been theorizing that analytics can allow my Innovation Team to unlock key insights from the data and accelerate innovation world-wide. The problem is large but I had the good fortune to attend the first offering of my company's [Data Science and Big Data Analytics course](http://www.emc.com/about/news/press/2011/20111205-01.htm). The course described a logical set of steps for testing business theories as part of a Data Analytics Lifecycle. I've been following these steps and after a number of posts describing the critical first step (the generation of an analytic plan), it is a good time to move on to Step 2: Data Prep.

As the arrows on the diagram indicate, these steps are iterative in the process. Proceeding to Phase 2 is often a matter of whether or not you are comfortable sharing the analytic plan with your peers. If so, then the data preparation phase can begin.

The analytic plan assists the data scientist in identifying the business problem, a set of hypotheses, the data set, and a preliminary plan for the creation of algorithms that can prove or disprove the hypotheses.  Once the analytic plan has been delivered and socialized, the next step is all about the data. In particular, the next step is all about *conditioning* the data.

The data must be in the right shape, structure, and quality to enable the subsequent analysis.

**Building an Analytic Sandbox**

In my [last post](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html) I mentioned that the data set in question falls into two categories: (a) a production "idea submission" server (essentially a large-scale database containing structured data), and (b) a globally-distributed set of unstructured documents representing knowledge expansion within the corporation in the form of minutes and notes about innovation/research activities.

These data sets cannot be analyzed in their current production formats. In addition, it is possible that the data is not of sufficient quality. Furthermore, the data is likely inconsistent. All of these possibilities add up to the fact that a separate analytic sandbox must be created to run experiments. Industry practice states that on average the size of this sandbox should be roughly ten times the size of the data in question (e.g. the current size of your enterprise data warehouse). Keep these things in mind when creating the sandbox:

* You are going to need strong bandwidth and network connections to your sandbox.
* Collect as much data as you can, including summary data, structured/unstructured, raw data feeds, call logs, web logs, etc. This is why the sandbox needs to be large.
* Determine the type of transformations you will need to assess data quality and derive statistically useful measures.
* Transform the data *after* it is in the sandbox (ELT: Extract, Load, Transform, as opposed to ETL). This allows analysts to choose to (a) transform the data or (b) use the data in its raw form. It's worth pointing out that this method is the opposite of best practice for some Data Warehousing use cases. While ETL is a widely accepted practice, the sandbox approach prefers ELT.
* Acquire the right set of tools for the transformation. Good examples would be [Hadoop](http://www.greenplum.com/industry-buzz/hadoop) for analysis, [Alpine Miner](http://www.greenplum.com/community/forums/forumdisplay.php?13-Alpine-Miner) for creating analytic workflows, and [R](https://www.r-project.org/)for many general purpose transformations.

Sandbox creation typically requires assistance from IT, a DBA, or the person that controls the enterprise data warehouse.

Once the sandbox is created, there are three key activities that allow a data scientist to conclude whether or not the data is "good enough".

1. Familiarize yourself with the data thoroughly. List out all the data sources and determine whether key data is available or more information is needed. This can be done by referring back to the analytic plan to determine if you have what's needed, or if more data must be loaded into the sandbox.
2. Perform data conditioning. Clean and normalize the data. During this process discern what to keep versus what to discard.
3. Survey & Visualize the data. Create overviews, zoom and filter, get details, and begin to create descriptive statistics and evaluate data quality.

I learned in the course that this part of the process is expected to take at least 50% of the time spent on the entire data analytics lifecycle (and 80% is not uncommon)!  Indeed, as our team went through this process for innovation analytics we had to work through quite a few issues before being able to work on implementing a model.

I will describe these issues in the next post.

Thanks again to [David Dietrich](https://twitter.com/#!/imdaviddietrich) for his research on the data lifecycle and ongoing support throughout this series of blog posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 05:09 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html)

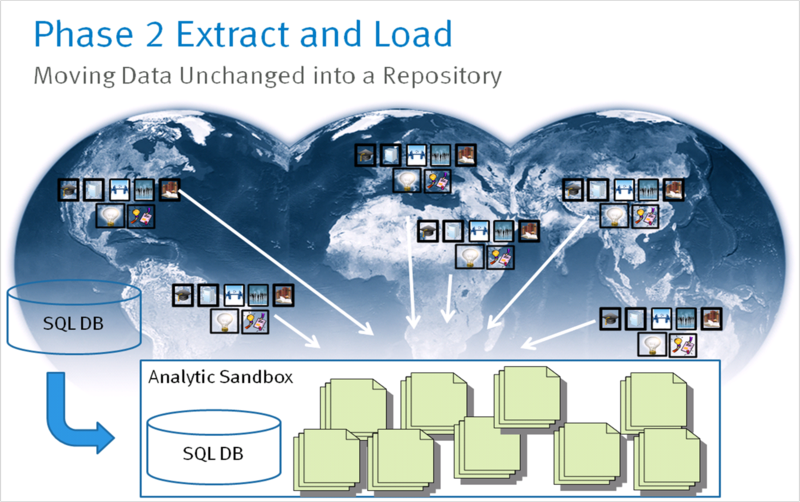
Tags: emc data analytics data science curriculum lifecycle big data innovation

Phase 2 Innovation Analytics: ELT

Diagrama

Descrição gerada automaticamente

This [series of articles](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) describes an analytic lifecycle being used to gain insight into the innovation and research practices of a multi-national corporation (EMC). After creating an [analytic plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html) in Phase 1, a previous post [described the Data Preparation phase](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html) of the lifecycle. This phase involves the creation of an enormous sandbox (e.g. ten times the size of a data warehouse configuration).  Data scientists and engineers are encouraged to extract data from many sources and load it into the sandbox unchanged. This approach may seem a bit revolutionary (most processes transform the data first). This lifecycle, however, is geared towards the data scientist. The possession of the raw data allows for more robust analysis. The diagram below provides an overview of this approach.



As I mentioned in the last post, there are two types of data that will allow a data scientist to analyze innovation. The first type, depicted on the left, is a structured SQL database containing thousands of innovation ideas submitted by employees over a five year period. The second type of data consists of minutes and notes from global innovation and research activities. This content is highly unstructured.

It’s worth taking a moment to discuss how the global team of users and data scientists came up with relevant innovation activities (depicted below). EMC’s product line consists of high-tech products and services that have been introduced into the marketplace. Tracing the lineage of these products and services usually results from an idea that happened long ago during a specific activity. The team came up with a candidate list of activities that is often associated with innovation and research:



Visiting universities, creating publications, attending conferences, visiting customers and partners, holding internal knowledge sessions, holding idea contests, and creating intellectual property are all activities commonly associated with innovation, and therefore an effort was made to gather six months worth of these activities from data sources worldwide.

After resisting the urge to transform these documents before loading them, a next logical step in Phase 2 is to explore the data. I will step through an example of this process in my next post.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 06:47 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html)

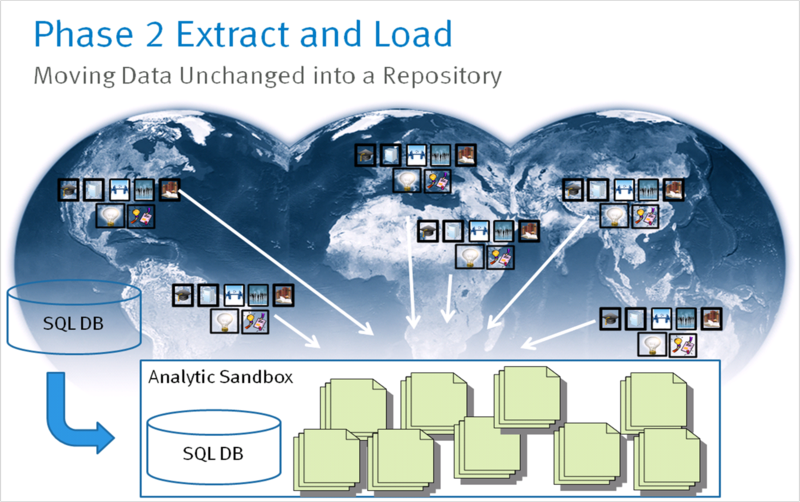
Tags: emc data science and big data curriculum innovation analytics

### Phase 2 Innovation Analytics: Exploring the Data

Diagrama

Descrição gerada automaticamente

At this point in the [series of articles](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) on the Data Analytic Lifecycle, raw data has been identified and imported into a Data Analytics sandbox. The data (a mix of structured and unstructured data) is depicted below. Contained within the sandbox now lies a large amount of data representing research and innovation activities occurring globally throughout my corporation (EMC).



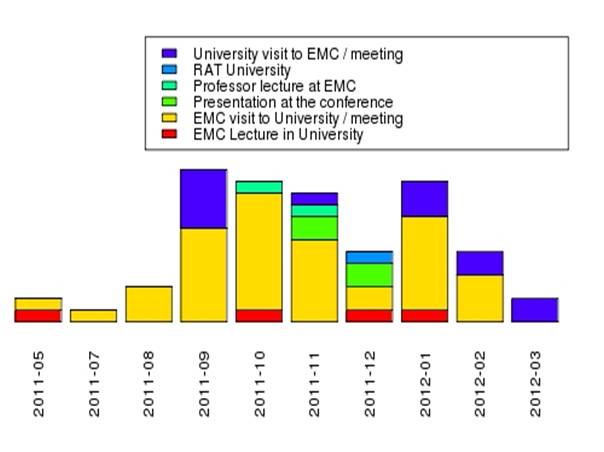
At this point in the lifecycle it is recommended that the data scientists and engineers begin to “get used to the data”. This means that they can use any number of tools to inspect the format, structure, and quality of the data. They have already immersed themselves into the [analytic plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html) generated in Phase 1, and they are likely looking for data sources to validate the hypotheses in the analytic plan.

This process will almost certainly identify “gaps” in the data that will prevent the data scientists from being able to prove or disprove the hypotheses.

At my corporation we have created a volunteer team of data scientists that ensure that our approach scales globally.  One of these data scientists is Vladimir Suvorov of Saint Petersburg, Russia. Vladimir described the tools and approach he used to explore the data:

R-studio provides the easiest way for initially examining the data placed within a sandbox. R-studio provides a simple connection string and SQL query on one side and a powerful statistics and graphing package on the other. Besides that, R is open-source software, so it can present additional value for the community.  I recommend R for fast prototyping and quick summaries of the data.

Vladimir created a chart that focuses solely on university activities. This chart uses color coding techniques to categorize the types of university activities that have occurred in previous months. What is actually happening here is broadly considered "data exploration".  In this phase, people like Vladimir explore the data, assess data quality, examine basic information about the data itself (relationships in the data, trends, etc.), and they try to understand what kind of data they actually have.  This lets them form ideas of how they can test in future phases, and what kinds of insights they may be able to drive towards.



Vladimir’s simple visual chart shows the pace of visits to universities (yellow), as well the emergence of a dark blue color, which represents meetings of the Research Advisory Team (RAT) for the purpose of discussing university research funding for 2012. These meetings continue throughout the first quarter of the current year (2012-03).

The relatively small amount of employee lectures at universities (red) and professor visits to EMC (aqua) is a potential indicator that programs could be put in place to accelerate these types of exchanges.

This chart falls under the category of “descriptive statistics”.

It’s an important activity in Phase 2 which, as just explained, can already lead to actionable conclusions.

During the Data Science and Big Data course, I learned that bar charts like this, while helpful for the data exploration phase, may not be the best choice in future phases. They can actually be improved via a few simple tips. Stacked bar charts are good for showing a few aggregated data points and exploring the data points.  When looking at data over time (as above), it is generally better to display line charts, to make it easy to see trends (people generally have a more difficult time examining trends in stacked bar charts).  Likewise, best practice is to use very soft colors and strong, emphasis colors as a way of highlighting key points; red is also a signal for things like “danger” or problem areas, especially in certain world countries.  In summary, a graphic like this gives us good initial information at this stage in the process. For more polished visualizations (used later in the lifecycle), Data Scientists need to address these kinds of aesthetic considerations.

During the data exploration exercise the data scientists and engineers begin to notice that (a) certain data needs conditioning and/or normalization, and (b) there are data sets that cannot be found anywhere which are critical to proving analytic hypotheses.

These activities will be further profiled in the next post. Once again, thanks to [David Dietrich](https://twitter.com/#!/imdaviddietrich), who not only taught the course I attended, but continues to oversee this series of posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 07:16 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-exploring-the-data.html)

Tags: emc innovation analytics big data science curriculum

### Phase 2 Innovation Analytics: Data Quality

Diagrama

Descrição gerada automaticamente

I am blogging my way through the Data Analytics Lifecycle as taught in EMC's Data Science and Big Data Analytics course.  I am running a Data Analytics project that employs a team of volunteer data scientists from around the world, and I have communicated an [analytic plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html) to them (along with a set of [hypotheses](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html)). The entry into Phase 2 of the process (typically the longest phase) has resulted in [preparing the data](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html), [loading it](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html) (without transform), and [exploring](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-exploring-the-data.html)it.

At this point it is worth mentioning a quote that I heard during the course:

If you do not have data of sufficient quality or cannot get good data, you will not be able to perform the subsequent steps in the lifecycle process.

**How Clean is the Data?**

One of the data scientists on my team was evaluating a tool called Tableau, which can be used for data exploration (among other things). They began to use the tool and explore the data that had been previously loaded into the analytics sandbox. They sent me the following screenshot (I zoomed in and circled my name):

[Tableau2](https://stevetodd.typepad.com/.a/6a00e5500d49008834016765a58b30970b-pi)

I am showing up twice in the database because some entries have a space before my first name. This is a classic problem (and not always easy to fix).  Addressing this problem within the sandbox is clearly a much easier proposition than doing so in the production database. However, it could take a long time to get it right (another reason why phase 2 takes so long).

Who typically does this work?  Is it a database admin (DBA)? A data engineer? Both typically play a role in Phase 2. The "Database Administrator" provisions and configures the database environment to support the analytical needs of the working team. The “Data Engineer” tends to have deep technical skills to assist with tuning SQL queries for data management and extraction. They also support data ingest to the analytic sandbox.  These people can be one in the same, but many times the data engineer is an expert onqueries and data manipulation (and not necessarily analytics as such).  The DBA may be good at this too, but many times they may simply be someone who is primarily skilled at setting up and deploying a large database schema, or product, or stack.

In addition to mis-spelled names, the data scientists exploring the data are starting to uncover missing data that will help them prove the hypotheses. For example, consider one of the hypotheses generated in phase 1:

H5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.

The association of boundary spanners to the geographic location where they work requires that any names found in the sandbox (e.g. "Todd,Steve") have an associated location (e.g. Hopkinton, MA).

Our data scientists found that this data was nowhere to be found within the sandbox.

In addition to DBAs and Data Engineers, IT often plays a large role in Phase 2.  For our project, once the names were "cleansed", we had to bring in IT resources to help generate geographic associations via our employee database. In our particular case we were fortunate: not only did IT grant access to our request, but the IT resource had Data Engineering skills and cleansed the data for us! In general, bringing in additional data from the IT realm is no easy task. Access to these types of assets is typically a very tough, time-consuming part of Phase 2.

I could write paragraph upon paragraph describing issues that we've come across (and solved) for Phase 2. It may be more useful, however, to summarize some of the lecture material that describes common problems:

* Consistency of data types (e.g. confirm that all numeric types contain numeric fields)
* Data feeds can often change over time (e.g. someone removes a column without telling anyone)
* Fields that contain calculations (e.g. interest charges) may change over time (if interest rates change over time)
* What are the legal ranges of data and are there any values that are out of bounds?
* Is the data standardized/normalized? If so, what is the scale?
* Are geospatial data sets consistent (e.g. metric versus english units, two-letter state abbreviations versus full-names)?

During this phase the data scientist may discern what to keep and what to discard. They had probably formed an opinion of what model they will use. Data exploration and cleansing has either validated their assumptions or caused them to select a different model. Data cleansing is a big job, so the objective should be to determine "what is enough?". What is clean enough data? What is sufficient quality for the operating context? What will properly enable the analysis?  These questions give people boundaries for the data cleaning, which is quite intensive.

Phase 3 is about model planning. How does one know when they are ready to leave phase 2 and move on to phase 3 (keep in mind that a return to Phase 2 is highly likely!)?

In general, Phase 3 begins when the data quality is "good enough" to start building the model. In my case, once we had cleaned up erroneous names and associated the names with geographies, the team had sufficient reason to enter Phase 3.

I will relate our team experience with Phase 3 in future posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 08:13 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-data-quality.html)

Tags: emc data science big data curriculum analytics innovation

Phase 3 Innovation Analytics: Model Planning

Diagrama

Descrição gerada automaticamente

This [series of posts](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) describes the efforts of a team of global data scientists. These data scientists are attempting to measure innovation at a large multi-national corporation. The approach they are following has been taken from the [Data Science and Big Data Analytics course](http://www.emc.com/about/news/press/2011/20111205-01.htm) created by their corporation (EMC).

After spending a good amount of time in Phase 1 (Discovery) and Phase 2 (Data Prep) of the Data Analytics Lifecycle, Phase 3 (Model Planning) is entered once the data scientists conclude that the data in their analytic sandbox is of sufficient quality. In Phase 2 the quality of the data was improved through various data cleaning and conditioning techniques.

As I learned in the course (via [David Dietrich](https://twitter.com/imdaviddietrich)):

*“Phase 3 represents the last step of preparations before executing the analytical models and, as such, requires you to be thorough in planning the analytical work and experiments in the next phase.”*

In Phase 3 the data scientists move closer to the algorithms that they will use to prove or disprove the [hypotheses](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) generated as part of the [Analytic Plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html).  The hypotheses frame the analytics that will be executed in Phase 4. Choosing the right methods to validate the hypotheses means that the team needs to consider some of the following conditions:

* The structure of the data will dictate what tools and analytic techniques can be used in Phase 4. Is textual data being analyzed?  If so, then maybe [Sentiment Analysis](https://en.wikipedia.org/wiki/Sentiment_analysis) using [Hadoop](https://hadoop.apache.org/) is the right approach. Does the sandbox contain structured financial data? Perhaps [regression](https://en.wikipedia.org/wiki/Linear_regression) via the [R analytics platform](https://www.r-project.org/) is the right method to use.
* The analytical technique that is chosen must map back to the business objectives. The objectives are met when the working hypotheses are proved or disproved. This condition clearly highlights why the generation of an Analytic Plan is so important.
* Determine whether or not the situation warrants a series of tests, or only one test. If a series of techniques must be used as part of a larger analytic workflow, then the team may benefit from an analytic workflow tool such as [Alpine Miner](http://www.alpinedatalabs.com/product/).

Some people may be tempted to jump directly to Phase 4 after loading, exploring, and conditioning the data in Phase 2. However, there is more exploring that needs to be done, and this phase of exploration is subtly different.

In Phase 2, the data exploration was mainly about data hygiene and quality.

In Phase 3, additional data exploration should focus on relationships between variables. These relationships will help to further understand the problem domain. The unbiased view of the data scientist is extremely valuable in this phase. Stakeholders (e.g. business users) bring their gut feelings and pre-defined hunches to the problem. Data scientists can translate these hunches into actual correlations between inputs and outcomes. They identify candidate predictors and outcomes, all within the framework of the business problem.

Our experience in Phase 3 has been valuable. As part of the analytic plan, we had theorized that the following analytic techniques would be valuable (described more fully in a [previous post](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html)):

* Use Map/Reduce …
* Natural language processing (NLP) …
* Several other techniques would be appropriate:
  + Clustering (e.g. k-means clustering) …
  + Classification …
  + Regression analysis …
  + Graph theory (e.g. Social Network Analysis) …

In Phase 3, the data scientist team began applying some of these models to the sandbox, and the results were mixed.

These results will be described more fully in future posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 08:21 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html)

Tags: emc data science scientist innovation curriculum

Phase 3 Innovation Analytics: Hypothesis Exploration

Diagrama

Descrição gerada automaticamente

In a previous post I described [a set of hypotheses](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) about innovation at my corporation (EMC). One of the hypotheses focused on the role of a boundary spanner in the transfer of geographic knowledge:

*H5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.*

In Phase 3 of the Data Analytics lifecycle (described in my [last post](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html)), I mentioned that:

*“Phase 3 represents the last step of preparations before executing the analytical models and, as such, requires you to be thorough in planning the analytical work and experiments in the next phase.”*

The third phase also gives the data scientist another opportunity to explore the data in ways that are specific to the set of hypotheses. In the case of identifying boundary spanners, EMC Distinguished Engineer and Data Scientist John Cardente already had a gut feel for the model that he wanted to use:

*"I believe that the boundary spanner hypothesis can be explored via social network analysis. I increased my knowledge of SNA by reading Albert-Laszlo Barabasi's excellent book, "Linked: How Everything is Connected to Everything Else and What it Means". While exploring some of the data in the analytic sandbox (e.g. EMC’s 2011 Innovation Showcase data), I confirmed that I had enough information to re-construct the social network associated with the contest entries. The social dynamics of EMC’s Innovation Showcase have always fascinated me. Was it the result of lone geniuses or large teams of collaborators? How connected was EMC's innovation network? Were there key innovators that acted like "hubs" tying ideas together?*

*I decided to use R for the social network analysis. I chose to use the igraph R package based on my knowledge of a great talk by Drew Conway entitled "Social Network Analysis in R". Thanks to the power of R and packages like plyr, it only took a small amount of code to transform the names associated with each contest submission into a form suitable for use with igraph. I produced the social network graph that was published as part of the “*[*Irish Butterfly*](https://stevetodd.typepad.com/my_weblog/2011/11/irish-butterfly.html)*” post. From there, I explored the capabilities of the igraph package and experimented with using cliques, components, degrees, and betweenness metrics to identify individual and groups of highly effective innovators."*

John’s use of R to explore his model choice (Social Network Analysis) took only a few lines of R code, but the code yielded immediate impact within EMC. A highly active network of Irish innovators was identified. As a result of his work I contacted the Irish team and invited them to share their effective approach with other countries.

The igraph package also allowed John to explore the “betweenness” metrics of individual innovators, and we began to consider the possibility that betweenness may help us to prove that research-specific boundary spanners are alive, well, and active at EMC (which is one of our hypotheses).

John’s success in this regard allows us to ask a pivotal question for moving to the next phase:

*“Do we have a good idea about the type of model to try”?*

In this case, the answer is yes. For other hypotheses, we have found the answer to be no. I will dive further into these Phase 3 scenarios in future posts.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: @SteveTodd

Director, EMC Innovation Network

Posted at 05:18 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html)

Tags: emc data science curriculum innovation modeling sna

Phase 3 Innovation Analytics: Longitudinal Studies

Diagrama

Descrição gerada automaticamente

This [series of posts](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/)is describing a cradle-to-grave Data Analytics project using the lifecycle taught at the [Data Science and Big Data Analytics course](http://www.emc.com/about/news/press/2011/20111205-01.htm) created by EMC. The steps of the lifecycle are being observed by a business user (myself) who is trying to gain insight into the innovation culture at EMC via a large amount of innovation and research data from around the world. The insight gained as part of this lifecycle should allow us to operationalize new plans and increase the pace of innovation.

As discussed in a [previous post](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html), Phase 3 is all about trying out analytical models and continuing to explore collected data. In many cases, Phase 3 can be exited because all of the required data is present and of high quality, and the selected analytical model appears to be promising. My team of global data scientists has had the good fortune of [experiencing this situation](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html).

What happens, however, when the data is incomplete, or the selected analytic model does not look promising?  This has also happened to our team and it is well worth telling the story.

As we considered the [list of hypotheses for our project](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) we focused on #7:

***Hypothesis #7 Incubation Lineage and Asset Generation***

I believe that the path that knowledge takes, from a local innovator, to a corporate boundary spanner, to an implementation team, to a delivered asset, can be traced and measured. I also believe that this measurement, once studied, can reveal ways to accelerate innovation and point out areas of knowledge that are yet to be converted. I've long been a fan of provenance, and I love the concept of "idea lineage". The lineage can be studied to reduce asset delivery time.

*IH7a: Frequent knowledge expansion and transfer events reduce the amount of time it takes to generate a corporate asset from an idea.*

*IH7b: Lineage maps can reveal when knowledge expansion and transfer did not (or has not) result(ed) in a corporate asset.*

When data scientists look at a hypothesis, some potential analytic models come to mind. My colleague [Dave Dietrich](https://twitter.com/imdaviddietrich) proposed an approach for Hypothesis #7:

We could in theory apply text mining techniques to address the concept of idea lineage.  That is, perhaps you could parse the ideas and descriptions, and then classify them (e.g. using a [Topic Modeling](https://stevetodd.typepad.com/my_weblog/2012/02/innovation-and-topic-modeling.html)approach).  Run an automated classification algorithm, such as naïve bayes, to parse and classify certain kinds of ideas.  Then create an outcome, such as patent or no-patent, publication or no-publication, new product or no product.  That is, you could identify the right outcomes and see if there is a relationship between clusters of certain types of text with discrete outcomes that represent innovation.

Dave's suggestion would use a naïve bayes model, and would appear to go a long ways towards proving the second hypothesis.

The first hypothesis, however, has a strong focus on elapsed time. During our discussion on analytic models and potential visualizations, Data Scientist Dong Xiang from EMC Labs China decided to do some *simulated*Model-3 data exploration using an impressive javascript visualization tool called [d3.js](https://mbostock.github.com/d3/%20). Using this tool he presented me with a time-lapse view of different phases of an idea:



I liked this approach so much that I commissioned Dong to try and use this data against the Phase3 data in the Analytic Sandbox. Tracking the progress of an idea and visualizing when it crosses thresholds would bring a time dimension into our study that would be useful for proving our hypothesis.

The sandbox contains a set of unstructured ideas, reports, minutes, and notes about global innovation and research activities. Unfortunately, Dong and the team found out the hard way that the data did not provide us with a good way to visualize the transition of an idea to new phases. EMC internally uses a variant of the [Technology Readiness Level](https://en.wikipedia.org/wiki/Technology_readiness_level) (TRL) approach for tracking phases, but the data found in the sandbox did not contain TRL levels. Further searching throughout EMC confirmed that this type of data was nowhere to be found.

Our ability to prove hypothesis #7 was in jeopardy. This realization was not the end of the world. In data scientist terms, it was time to begin a [longitudinal study](https://en.wikipedia.org/wiki/Longitudinal_study) (making a series of observations over a long period of time).The team began to design a method whereby TRL levels would be gathered and recorded as a regular part of the reporting and gathering of global innovation activities. Over time, we would eventually have enough data to take a good, hard look at our hypothesis.

Our longitudinal study would involve the following:

* Establish a goal criteria.  For our case, what would be the end goal of a successful idea that has traversed the entire journey?
* Identify the right milestones to achieve this goal
* Trace how people move ideas from each milestone towards the goal.
* Once this is done, trace ideas that die, and trace others that reach the goal.  Compare the journeys of ideas that make it and ideas that don't.
* 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](https://en.wikipedia.org/wiki/Student%27s_t-test), or perhaps different types of classification algorithms.

A longitudinal study has a similar motto to the Data Analytic Lifecycle: plan everything thoroughly up front!

This post described a hypothesis that fell flat in Phase 3. My previous post described a hypothesis that moved forward into Phase 4 because the model seemed right.

With our analytic plan refined, the team moved to Phase 4. I will introduce this phase in my next post.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: @SteveTodd

Director, EMC Innovation Network

Posted at 10:03 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-longitudinal-studies.html)

Tags: emc innovation data science scientist

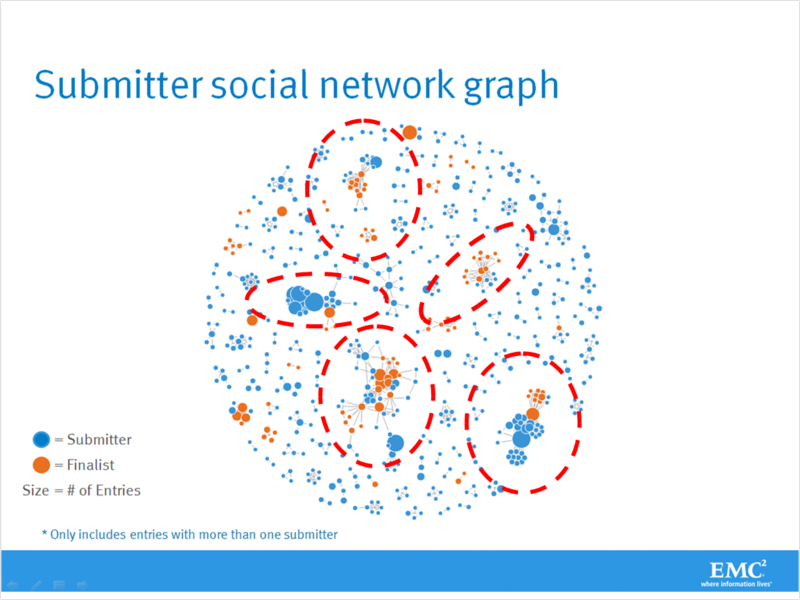
Phase 4 Innovation Analytics: Finding Boundary Spanners

In [previous posts](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) I described the first three phases of the Data Analytics Lifecycle. The fourth phase is where the [rubber meets the road](https://en.wiktionary.org/wiki/the_rubber_meets_the_road). Data Scientists begin running their models and trying to prove the [hypotheses](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html)established in Phase 1.  In Phase 3 I [mentioned that the use of Social Network Analysis (SNA)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html) could help prove the following hypothesis:

*H5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.*

The visualization below is described as follows:

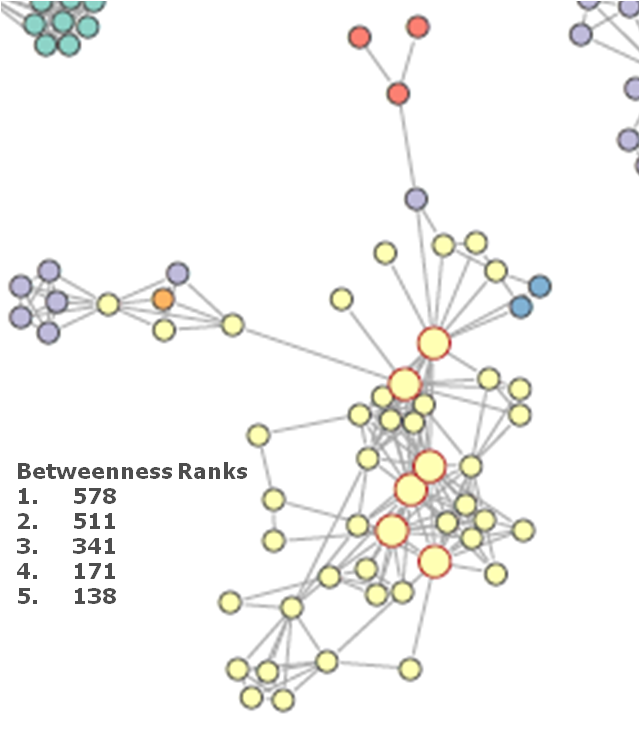
*The input for this graph is from EMC's Innovation Showcase (an idea submission contest). Each circle in the graph represents an idea submitter that was part of a team (i.e. more than one submitter on an entry). Gray lines between circles represent team relationships - two circles connected by a line indicate that those participants submitted an entry together. The size of each circle represents the associated participant's number of contest entries. Orange circles represent contest participants with an entry selected as a finalist.*



I've highlighted five "clusters" of idea submitters/inventors. In a previous post, I drilled down into one of these clusters and discovered that [each submitter was Irish](https://stevetodd.typepad.com/my_weblog/2011/11/irish-butterfly.html).  I followed up with several of these individuals and discovered that the cluster had formed as a direct result of [targeted innovation training](https://stevetodd.typepad.com/my_weblog/2011/12/irish-cocoon.html) that had occured at EMC's facility in Ireland.

The analytics run by EMC Data Scientist John Cardente turns our employee idea database into numerical representations, and these values are then displayed visually to help prove or disprove the hypotheses. Does the knowledge transfer activity that occurs as part of our Innovation Showcase contest identify knowledge-specific geographic boundary spanners?

I asked John to drill down into the lower central cluster. He put together this color-coded visualization:



Red dots represent EMC employees from Israel. Purple dots represent employees from the United States. The two blue dots located towards the upper right are French EMC employees. The lone orange dot on the left represents an inventor from Australia.  The yellow dots, representing one of the largest clusters in the entire experiment, are Chinese innovators.

Any large dot with a red outline represents a "hub". A hub has a large number of connections and high betweenness (as discussed in a [previous post](https://stevetodd.typepad.com/my_weblog/2011/11/irish-butterfly.html)).

When looking at my co-workers from China I saw two values that jumped out at me. I've listed the top five Chinese "betweenness" rankings within the graph above. Two out of the five Chinese employees had betweenness scores that were much, much higher than the rest. I asked John to elaborate on what betweenness means in this context:

*The social network analysis metric of*[*betweenness*](https://en.wikipedia.org/wiki/Centrality#Betweenness_centrality)*is a measure of a node’s importance to the connectivity of a graph. In the case of our Innovation contest, if a person has a high betweenness score, this means that they have a high degree of influence on the other inventors that are submitting ideas.*

Has betweenness identified boundary spanners in China? In my next post I will explore who some of these people are, and whether or not they truly serve as boundary spanners for geographic knowledge.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

[Twitter: @SteveTodd](https://www.twitter.com/SteveTodd)

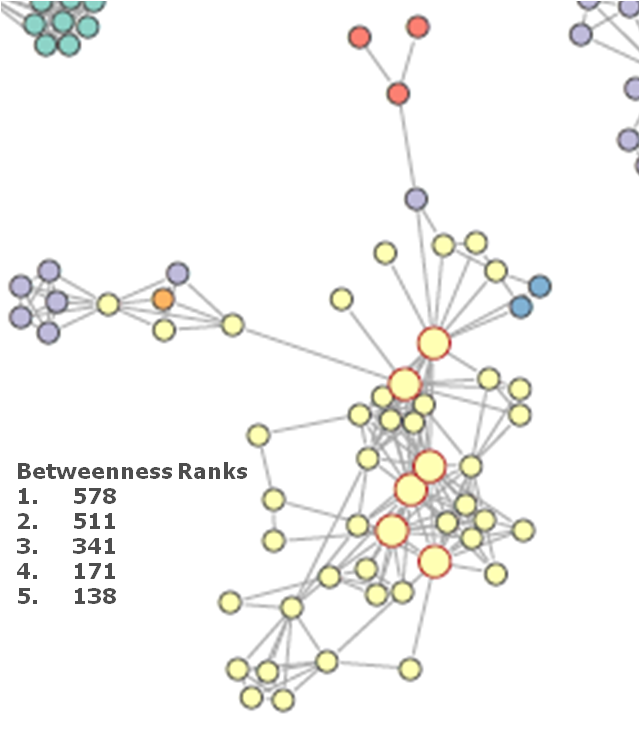
Director, EMC Innovation Network

Posted at 06:55 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html)

Tags: emc innovation analytics betweenness

### Phase 4 Innovation Analytics: Boundary Spanner Validation

In my [last post](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html) I described the entry into Phase 4 of the [Data Analytics Lifecycle](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/).  This phase is all about running analytic models against the high-quality data in an [Analytic Sandbox](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html). EMC Data Scientists are attempting to prove that knowledge transfer activity (e.g. idea submissions that are part of our global Innovation Showcase) can identify research-specific boundary spanners. The first clue that our hypothesis is correct can be found on the chart below.



As I stated in my last post, each color 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 high betweeness. I chose this cluster because it contains geographic variety, which is critical to prove our hypothesis about geographic boundary spanners.

One person in this graph has a betweenness score of 578, which is "sky-high" when compared to the rest of the nodes in the graph.

The name of this person is Jidong Chen.

I was not surprised to hear his name.

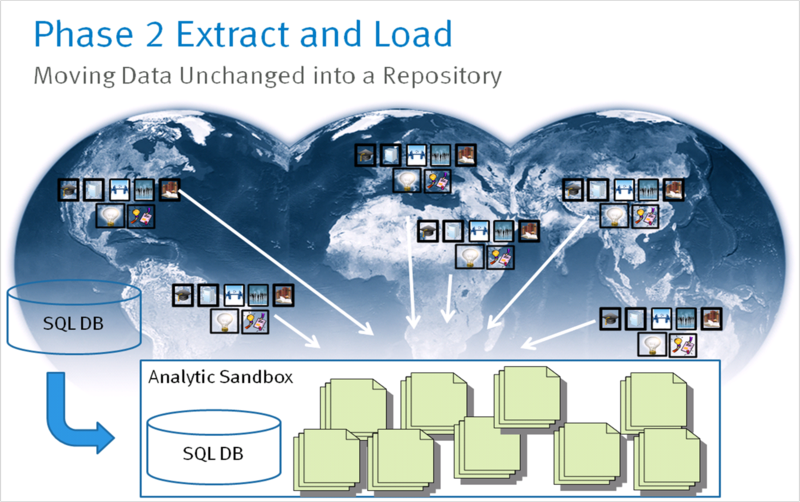
Jidong has been a researcher in EMC's Office of the CTO for five years. In fact, I wrote a blog post about Jidong's research several years ago, in which he and his team proposed a [vastly different approach to desktop search.](https://stevetodd.typepad.com/my_weblog/2009/08/search-your-memory.html)

With Jidong's name in hand, I return to my [original hypothesis #5](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) from Phase 1:

***Hypothesis #5: Knowledge transfer facilitation via boundary spanners***  
  
There are certain employees that have arisen within a geography and made connections with other geographies for the purpose of collaboration. They may not have high visibility within a corporation aside from the direct connections that they have made on their own.  I believe that not only can analytics identify these people, but analytics can also classify the type of knowledge that these individuals are transferring. These "boundary spanners" can be targeted and trained as "innovation facilitators" and united at a corporate level.

IH5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.

Does betweenness indicate that Jidong is a boundary spanner? In order to further prove this hypothesis, I can simply run a query against Jidong's name within our Analytic Sandbox. This sandbox was described in a [Phase 2 post](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html), and is depicted below as a reminder.



What did the collection of entries in the sandbox tell me about Jidong's innovation and research activities?

* In 2011 Jidong attended the SIGMOD conference in Greece
* Jidong visited EMC employees in France that are part of the IIG business unit (e.g. Documentum)
* Jidong presented his thoughts on the SIGMOD conference at a Virtual Brownbag session attended by
  + Three employees in Russia
  + One employee in Cairo
  + One employee in Ireland
  + One employee in India
  + Three employees in the U.S.
  + One employee in Israel
* In 2012 Jidong attended the SDM 2012 Conference in California
* On the same trip he visited innovators and researchers at Greenplum and VMware
* Later on that trip he stood before the monthly CTO Council and introduced two of his researchers to dozens of EMC innovators and researchers

The bottom line is that (part of) our hypothesis is indeed correct. The data (and the model) have identified a boundary spanner. What is the nature of the knowledge that Jidong transfers as part of his boundary spanner activity? In other words, how can the model identify the "research-specific" aspects of the boundary spanner?

In order to answer this question, we will turn to a different Phase 4 modeling exercise, which willl be described in the next post.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 07:48 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html)

Tags: emc innovation analytics lifecycle data science scientist

### Phase 4 Innovation Analytics: Voice Measurement

Diagrama

Descrição gerada automaticamente

This [series of blog posts](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) has been describing a methodology (EMC's Data Analytics LifeCycle) for using analytics to measure innovation at a multi-national corporation.

There are [eight different hypotheses](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html) that the project is attempting to prove. My last few posts have focused on one specific hypothesis:

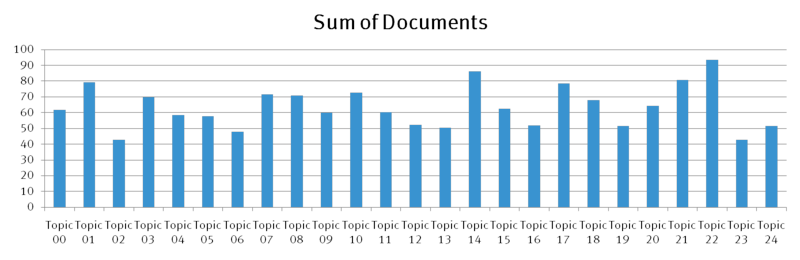
H5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.

Phase 4 of the lifecycle focuses on running analytic models against high-quality data contained in an [analytic sandbox](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html). The chosen analytic method (Social Network Analysis) seemed to indicate that the measurement of knowledge transfer activity did indeed identify [geographic boundary spanners](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html). This was confirmed by focusing on a specific innovator (EMC Labs China innovator [Jidong Chen](Permalink:%20http://stevetodd.typepad.com/my_weblog/2012/05/phase-4-innovation-analytics-boundary-spanner-validation.html%20Edit)) and observing that he indeed attended a large number of meetings with geographically-dispersed innovators.

Our hypothesis, however, insists on identifying research-specific boundary spanners.

Can the "voice" of an innovator be mapped to one or more specific research themes? If the answer is 'yes', then our global team of data scientists have come up with a model that proves the hypothesis.

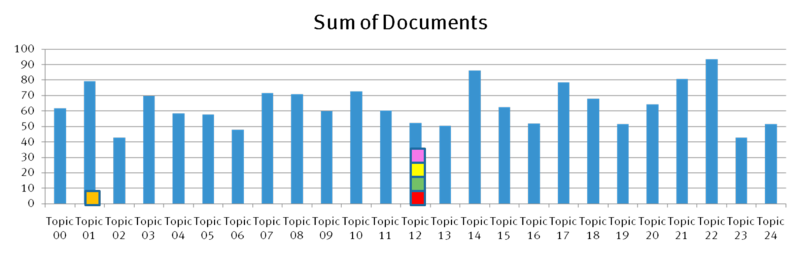
As part of Phase 4 activities, EMC Data Scientist Tao Chen (also of EMC Labs China) ran topic modeling algorithms against the data in the analytic sandbox. In a [previous post](https://stevetodd.typepad.com/my_weblog/2012/02/innovation-and-topic-modeling.html) I highlighted the identification of twenty-five topics that emerge from this analysis (number 00-24 in the graphic below):

  
If Jidong's minutes, notes and presentations from all of his meetings were mapped against this topic model, what would it look like?  As a reminder, the following events that Jidong participated in were queried from the analytic sandbox. I have color coded several of them in order to map them to the model.

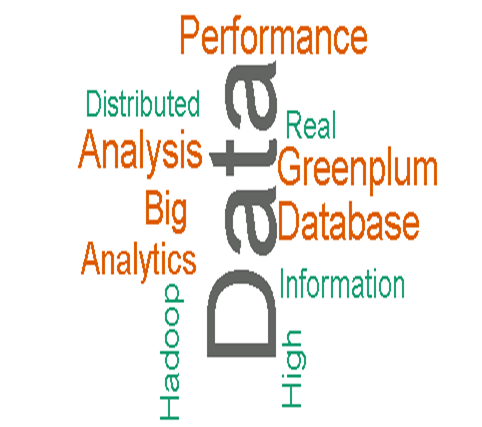
* In 2011 Jidong attended the SIGMOD conference in Greece
* Jidong visited EMC employees in France that are part of the IIG business unit (e.g. Documentum)
* Jidong presented his thoughts on the SIGMOD conference at a Virtual Brownbag session (GREEN) attended by
  + Three employees in Russia
  + One employee in Cairo
  + One employee in Ireland
  + One employee in India
  + Three employees in the U.S.
  + One employee in Israel
* In 2012 Jidong attended the SDM 2012 Conference in California (RED)
* On the same trip he visited innovators and researchers at Greenplum (PURPLE) and VMware (ORANGE)
* Later on that trip he stood before the monthly CTO Council and introduced two of his researchers (and his research) to dozens of EMC innovators and researchers (YELLOW)

The Stanford Topic Modeling toolkit can take the minutes, notes, and/or presentations from these meetings and map them against all twenty-five topics. For each color listed above, here is a visualization of the mapping of Jidong's activities:

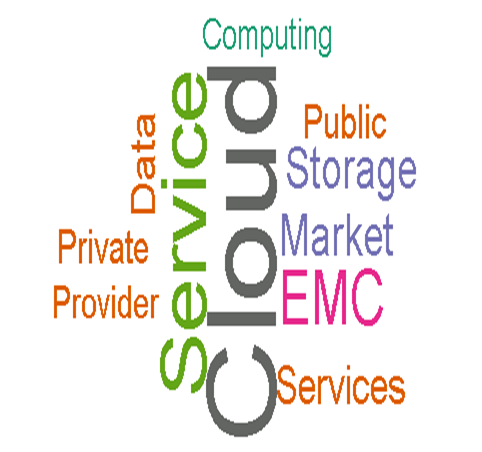
Four out of five of Jidong's activities map to topic #12. The fifth activity (his meeting with VMware) mapped to topic #1 (just barely edging out topic #12).



What is topic #12? The graphic below identifies the word cloud, and we see the theme of "Big Data" begin to emerge:



Interestingly, Jidong's VMware meeting mapped to the following word cloud:



This post has demonstrated that Phase 4 of the Data Analytics Lifecycle allowed us to prove one of our eight hypotheses: analyzing a global repository of research and innovation activity identifies research-specific boundary spanners.

There are seven more hypotheses to prove (or disprove). As our team of global data scientists begins to run their models, I will continue to publish the results.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

Posted at 07:15 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-voice-measurement.html)

Tags: emc innovation analytics topic modelling

### Phase 5 Innovation Analytics: Global Knowledge Flight Patterns

Diagrama

Descrição gerada automaticamente

Our team of data scientists have not reached Phase 5 of the Data Analytics LifeCycle for measuring innovation at EMC. We will likely be in Phase 4 (running analytic models) for several months. In the interest of finishing up this series, however, I'd like to share my thoughts about preparing for Phase 5.

Phase 5 is labeled as "Communicate Results". Presumably a business user received permission (and resources) to create and execute an [Analytic Plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-creating-the-analytics-plan-.html). Before the creation of this plan there was a vague notion, an idea, or a hunch that needed to be proven. Phase 1 represented the beginning stages of scoping the problem, assessing the risks, and defining success. The business value was fuzzy and hard to quantify.

In Phase 5 the results will be communicated.The vague, fuzzy notions of earlier phases should be replaced by quantifiable conclusions. This may seem like a straightforward step (and it can be if the steps are followed diligently). The main takeaway that I learned about Phase 5 can be summarized as follows:

Hypotheses about unlocking value from corporate data were either proven or not. Did we succeed? Did we fail? In Phase 5, the business value imagined before Phase 1 is quantified and presented back to the corporation.

One way to prepare for Phase 5 is to start thinking about this question:

What are the three most significant findings in the observation of the data?

Here are three significant experiences our team has had during this project:

1. Early visualizations yield enormous (and actionable) insight. In Phase 3 our global data scientists began exploring the data. Social network analysis visualizations identified [Irish Butterflies](https://stevetodd.typepad.com/my_weblog/2011/11/irish-butterfly.html) and [Chinese Boundary Spanners](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html). The at-a-glance insight that these graphs revealed led to the immediate operationalization of new processes (which I will describe in Phase 6). Visualizations help immensely (as opposed to staring at the raw data).
2. Measuring innovation delivery is difficult. Correlating ideas with successful (or unsuccessful) delivery could not be accomplished with our data sets. As a result we have launched a [Longitudinal Study](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-longitudinal-studies.html).
3. "Knowledge Flight Patterns" can identify global gaps in communication.

I have direct responsibility for the output of dozens of top-notch researchers at our EMC Labs China location. In 2011 some of my Chinese co-workers came up with a graphical visualization of "Knowledge Flight Patterns".



The capture of meeting minutes within my corporation [identified boundary spanner Jidong Chen](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html). His lecture on SIGMOD findings (June 2011) was shared with six different countries. The EMC Labs China team wrote software that placed red dots on a map to represent participating innovators (hovering over the red dots reveals the names). The animated yellow dots represent the knowledge being transferred across geographic boundaries (from Jidong to the team). In the above example, the yellow dots represents "Big Data" knowledge.

Jidong's talk is just one entry in our database (which contains records of hundreds of such conversations). If each entry is "animated" in the same fashion, eventually a trans-global grid of knowledge transfer "flights" will be displayed. Different colors can be used for different types of knowledge. When looking at this type of grid, a few insights can emerge:

* Some global locations are the source of a particular kind or class of knowledge
* Some global locations are on the receiving end for a particular kind of knowledge, while some are left out in certain conversations
* Some global locations are more active than others when it comes to knowledge transfer participation.
* Some flight patterns regularly leave (and land) at defined intervals, while some flights are more sporadic.

Communicating this learning in Phase 5 is important.  The [course that taught me these steps](http://www.emc.com/about/news/press/2011/20111205-01.htm) recommends a template to use when entering Phase 5. The template is an excellent stimulus for communication, and ties back directly to the "plea for resources" made before Phase 1.

My mentor and teacher on this topic, EMC's [David Dietrich](https://twitter.com/#!/imdaviddietrich), points out that Data Scientists often don't enjoy going to the effort of messaging their results to different audiences in different ways. To quote Dave:

"Many people who are great at the analytics do not enjoy this story telling or evangelization portion of the project.  As a result, they may give it short shrift.  Instead, I have come to view these opportunities as a way to (a) fine tune my message and (b) drive change.   Sharing a strong message means you have a chance to reach multiple groups and influence behavior if  what you are embarking on is worth driving change."

I'm nearing the end of this journey through the six steps. My last post will describe Step 6 (Operationalize). Our team is far from arriving at Step 6, but I will share some changes that we have already introduced as a result of executing the first four steps.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/stevetodd)

Director, EMC Innovation Network

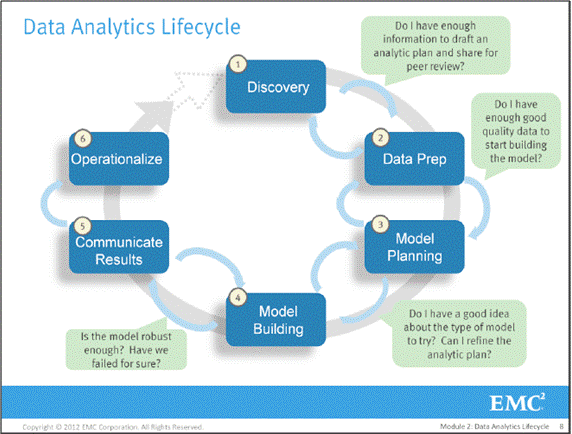
Posted at 06:53 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/06/phase-5-innovation-analytics-global-knowledge-flight-patterns.html)

Tags: emc data science analytics

Phase 6 Innovation Analytics: Operationalize

This is the last in a series of posts describing a methodology (EMC's Data Analytics LifeCycle) for using analytics to measure innovation at a multi-national corporation. This lifecycle is taught at the [Data Science and Big Data Analytics course](http://www.emc.com/about/news/press/2011/20111205-01.htm) created by EMC, and I've blogged my way through each phase of the lifecycle and have arrived at the end (Phase 6).

As a review, here is a graphical view of the lifecycle, followed by a summary of all the posts written thus far:



**Phase 5**

[Global Knowledge Flight Patterns](https://stevetodd.typepad.com/my_weblog/2012/06/phase-5-innovation-analytics-global-knowledge-flight-patterns.html)

**Phase 4**

[Voice Measurement](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-voice-measurement.html)

[Boundary Spanner Validation](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html)

[Finding Boundary Spanners](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html)

**Phase 3**

[Longitudinal Studies](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-longitudinal-studies.html)

[Hypothesis Exploration](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html)

[Model Planning](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html)

**Phase 2**

[Data Quality](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-data-quality.html)

[Exploring the Data](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-exploring-the-data.html)

[ELT](https://stevetodd.typepad.com/my_weblog/2012/04/phase-2-innovation-analytics-elt.html)

[Data Preparation](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html)

**Phase 1**

[Creating The Analytics Plan](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html)

[Hypothesis Generation](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html)

[Introduction to Phase 1](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics.html)

[Introduction to Innovation Analytics](https://stevetodd.typepad.com/my_weblog/2012/03/a-strategy-for-innovation-analytics.html)

Given the foundation of the first five phases, let's finish with the final phase.

Phase 6 is called "Operationalize". My team and I have not yet reached this phase. My understanding of Phase 6, however, is influencing our journey through the steps. The journey that my team has undergone so far can be summarized as follows:

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

Phase 6 moves the analytic models out of the sandbox and into production. The course advises a production "pilot" be run first (as opposed to deploying the model on a wide-scale). This approach minimizes risk. Smaller-scale deployment allows the team to learn about the performance and make adjustments before a full deployment.

Phase 6 may require a new team of people to join the initiative (the people that are responsible for running the production environment). These people will help feed data sets into the production model. During the execution of the model in the production environment, it is important to detect anomalies on inputs *before* they are fed into the model. This may not be 100% possible. Consider doing a logistic regression on a training set of the data if possible.

What does this specifically mean for the project I've been running? The points mentioned below are key aspects to remember for any company wishing to run innovation analytics:

* We need more data, which means we need a marketing initiative to convince people to submit (or inform) the global community on their innovation/research activities.
* This data is sensitive and some thought needs to go into "who" can run the model and "who" see the results
* In addition to running models, a parallel initiative will likely be to access the repository for search (people want to search for innovation/research initiatives). This may impact the performance of the analytics.
* We need a mechanism to continually re-evaluate the model after deployment. Assessing the benefits is one of the main goals of this stage, as well as defining a process to retrain the model as needed.

This last point represents a challenging and often overlooked aspect of Phase 6. The team needs to assess whether the model is meeting goals and expectations, and if desired changes are actually occurring. The data may change over time, or live data may morph to the point where the model needs to be updated or retrained.

As I reach the end of this series of blog posts, I'd like to thank[Dave Dietrich](https://twitter.com/imdaviddietrich), who has proofread nearly all of my posts for accuracy!

As the efforts of the data scientists come to a close, Dave has one final piece of advice:

*Hold a post-mortem with the analytic team to discuss what would change in the process or project if you had the chance to do it over again.*

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

[Twitter: @SteveTodd](https://www.twitter.com/SteveTodd)

Director, EMC Innovation Network

Posted at 05:37 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/07/phase-6-innovation-analytics-operationalize.html)

Tags: phase 6 innovation analytics emc

## June 17, 2013

### [Researching Different Minds Thinking Greatly](https://stevetodd.typepad.com/my_weblog/2013/06/researching-different-minds-thinking-greatly.html)

Several weeks ago I published a blog post ([Different Minds Think Greatly)](http://reflectionsblog.emc.com/2013/04/different-minds-think-greatly.html) that explored the topic of cognitive diversity and innovation. At the time, I had read a [Techonomy article](http://techonomy.com/2013/04/too-much-like-mindedness-hurts-companies-and-the-country/) by John Hagel and John Seely Brown, who basically asserted that “too much like-mindedness hurts companies”, and I quoted the following:

Organizations that host a diverse and broad range of members have a resilience that results from cross pollination.

As part of the article I echoed my agreement with this assertion and referred to some Social Network Analysis [from my own company](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/) (EMC). The data that we modeled highlighted that a good degree of geographic diversity can result in higher quality ideas.  “Higher quality ideas” are typically defined as ideas that receive a high score from judges in our yearly [Innovation Roadmap](https://stevetodd.typepad.com/my_weblog/2013/04/innovation-roadmap-2013-push-the-improve-button.html) program (especially ideas that reach finalist or funded status).

Our data shows that when diverse minds from different cultures collaborate on new approaches, good ideas result.  The conclusions we drew from this analysis have resulted in behavioral change at EMC. Most notably we’ve formed a global “Innovation Best Practices Community” in order to intentionally stimulate this behavior.

After publishing some of our findings, we were approached by two universities on an interesting joint research project. They asked if we wouldn’t mind sharing a filtered view of our employee  idea activity in order to correlate it against the public Twitter profiles of these same people. Professors [Eoin Whelan](http://www.nuigalway.ie/business-public-policy-law/cairnes/subjectareas/bis/bisstafflisting/eoinwhelan/) (NUI Galway) and [Salvatore Parise](http://www.babson.edu/faculty/profiles/Pages/parise-salvatore.aspx) (Babson) invited us to focus on an offshoot of cognitive diversity known as “structural holes”.  Eoin explains structural holes in the following manner.

Ron Burt’s[*theory of structural holes*](https://www.amazon.com/Structural-Holes-Social-Structure-Competition/dp/0674843711)has proven to be influential in explaining how innovation transpires.  Burt proposes that gaps in a social network, structural holes, create brokerage opportunities.  A structural hole indicates that the people on either side of the hole circulate in different flows of information and advantages accrue to those individuals whose relationships span the structural hole.  In his best selling book The Tipping Point, Malcolm Gladwell argues that the success of Paul Revere’s midnight ride was due to his quite diverse social networks – ranging from hunting and card playing to theatre and business. Therefore, he knew which doors to knock on when arriving in a town.  Network brokers like Revere not only disseminate information more broadly, they also benefit by receiving a greater novelty of information from their diverse social contacts.  Indeed, studies within organizations have shown that employees, teams, and even entire companies with more diverse network connections tend to be more innovative.

As a result of our conversation, we polled our global Innovation Best Practices community and asked idea submitters to voluntarily share their Twitter handles with Eoin and Sal. We packaged up the Twitter handles with the employees’ corresponding level of innovation activity over a period of several years.

This research has been ongoing for several months, and in an upcoming post I plan to share some of the results and what it might mean for our organization.  Before I do, however, I’d like to discuss this Data Science project in the context of Phase 1 of the[Data Analytics Life Cycle](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/): [Hypothesis Generation](https://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics-hypothesis-generation.html).

Our hypothesis could be stated as follows:

EMC employees with diverse external Twitter networks submit higher quality ideas.

By “diverse” we mean “disconnected” or “fragmented”. In other words, employees that follow people that are not “like-minded” tend to submit better ideas due to the diversity of their network.

If this hypothesis proves to be true, we can brainstorm ways of stimulating additional behavioral change via encouraging our employees to fragment their Twitter networks. Proving the hypothesis involves iterating through the additional phases of the Life Cycle (e.g. [Data Prep](https://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html), [Data Modeling](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html), etc).

In future posts I will share the results of the modeling exercise conducted by Eoin and Sal.

Steve

[http://stevetodd.typepad.com](https://stevetodd.typepad.com/)

Twitter: [@SteveTodd](https://twitter.com/SteveTodd)

EMC Fellow

Posted at 06:42 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Global Innovation](https://stevetodd.typepad.com/my_weblog/global-innovation/), [Storage Research](https://stevetodd.typepad.com/my_weblog/storage_research/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2013/06/researching-different-minds-thinking-greatly.html) | [Comments (0)](https://stevetodd.typepad.com/my_weblog/2013/06/researching-different-minds-thinking-greatly.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2013/06/researching-different-minds-thinking-greatly.html#trackback)

Tags: emc structural holes research twitter innovation innovation roadmap NUI galway babson