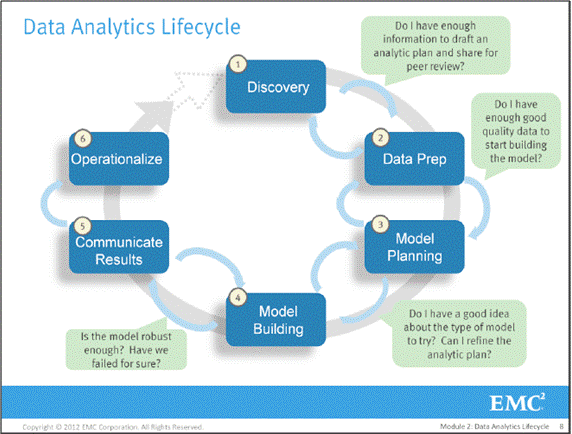
### [Phase 6 Innovation Analytics: Operationalize](https://stevetodd.typepad.com/my_weblog/2012/07/phase-6-innovation-analytics-operationalize.html)

## July 05, 2012

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:

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340177430a19bb970d-pi)

**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)

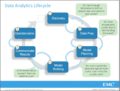
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/07/phase-6-innovation-analytics-operationalize.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/07/phase-6-innovation-analytics-operationalize.html" \l "trackback)

Tags: phase 6 innovation analytics emc

## June 19, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d4900883401630691614f970d-pi)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".

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340163069187f2970d-pi)

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)

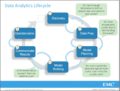
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-5-innovation-analytics-global-knowledge-flight-patterns.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-5-innovation-analytics-global-knowledge-flight-patterns.html" \l "trackback)

Tags: emc data science analytics

## June 13, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340163061d971f970d-pi)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:

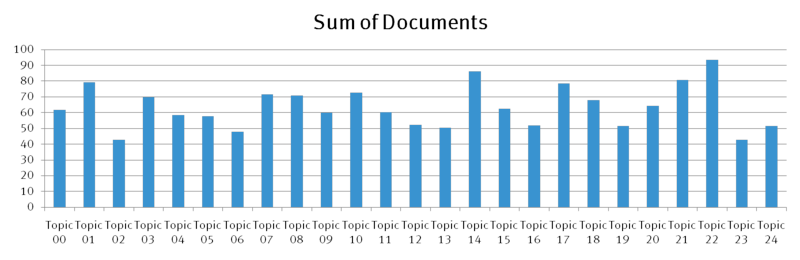
H5: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.

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" \t "_self)) 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.

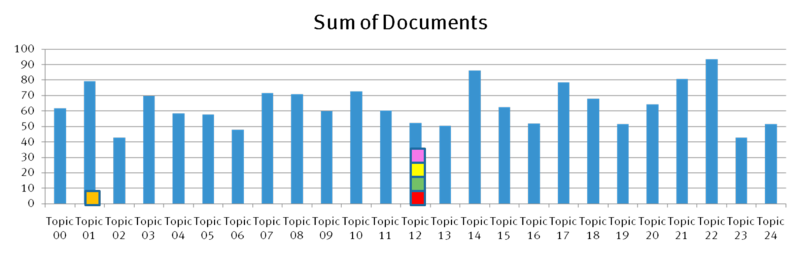
As part of Phase 4 activities, EMC Data Scientist Tao Chen (also of EMC Labs China) ran topic modeling algorithms against the data in the analytic sandbox. In a [previous post](https://stevetodd.typepad.com/my_weblog/2012/02/innovation-and-topic-modeling.html) I highlighted the identification of twenty-five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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340163061da313970d-pi)  
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.

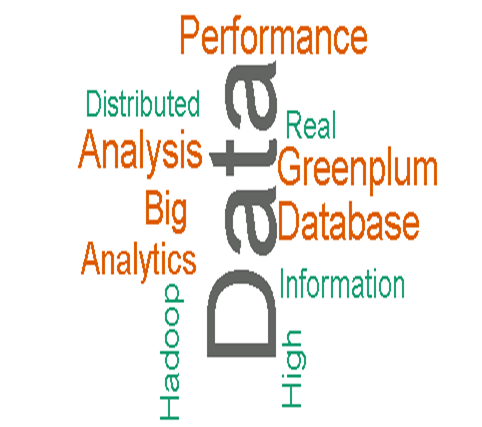
* In 2011 Jidong attended the SIGMOD conference in Greece
* Jidong visited EMC employees in France that are part of the IIG business unit (e.g. Documentum)
* Jidong presented his thoughts on the SIGMOD conference at a Virtual Brownbag session (GREEN) attended by
  + Three employees in Russia
  + One 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)

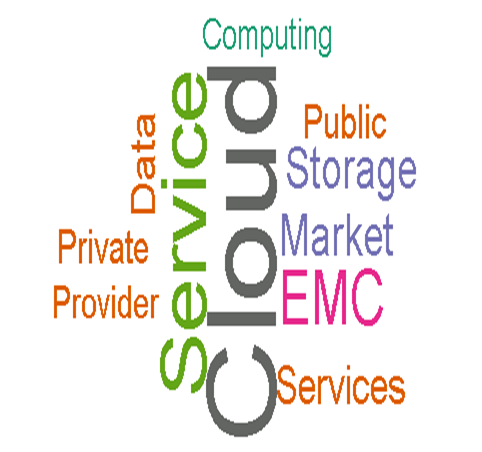
The Stanford Topic Modeling toolkit can take the minutes, notes, and/or presentations from these meetings and map them against all twenty-five topics. For each color listed above, here is a visualization of the mapping of Jidong's activities:

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).

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340163061dcd23970d-pi)

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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340163061dc5cf970d-pi)  
  
Interestingly, Jidong's VMware meeting mapped to the following word cloud:

[](https://stevetodd.typepad.com/.a/6a00e5500d49008834016767116220970b-pi)  
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

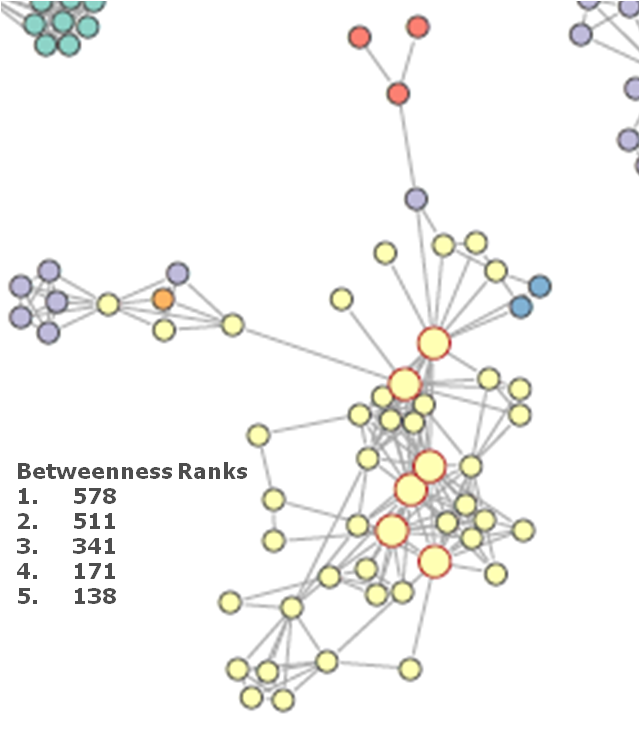
Posted at 07:15 AM in [Data Science and Big Data Curriculum](https://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/), [Innovation Analytics](https://stevetodd.typepad.com/my_weblog/innovation-analytics/) | [Permalink](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-voice-measurement.html) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-voice-measurement.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-voice-measurement.html" \l "trackback)

Tags: emc innovation analytics topic modelling

## June 07, 2012

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

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.

[](https://stevetodd.typepad.com/.a/6a00e5500d49008834016766c79b4f970b-pi)

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.

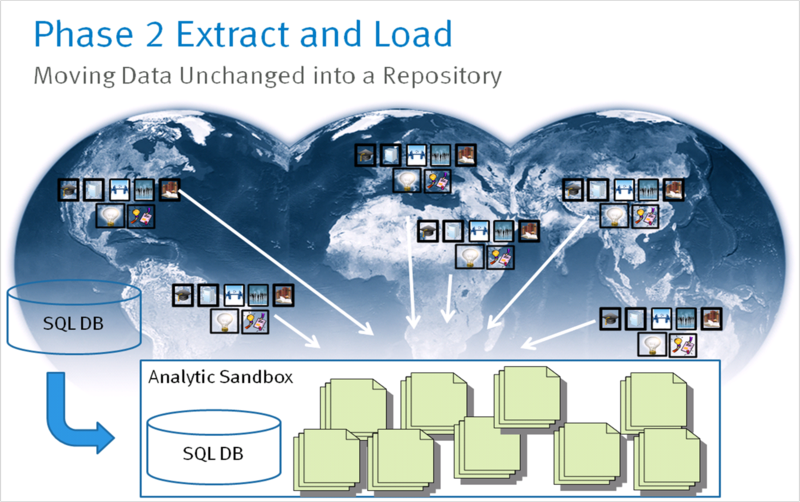
Jidong has been a researcher in EMC's Office of the CTO for five years. In fact, I wrote a blog post about Jidong's research several years ago, in which he and his team proposed a [vastly different approach to desktop search.](https://stevetodd.typepad.com/my_weblog/2009/08/search-your-memory.html)

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.

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340168ebc90c6d970c-pi)  
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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/06/phase-4-innovation-analytics-boundary-spanner-validation.html" \l "trackback)

Tags: emc innovation analytics lifecycle data science scientist

## June 04, 2012

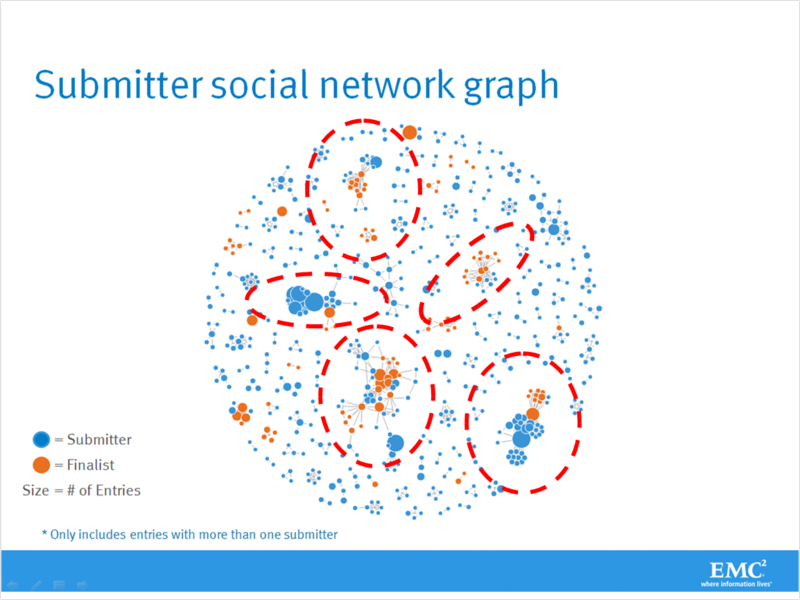
### [Phase 4 Innovation Analytics: Finding Boundary Spanners](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html)

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:

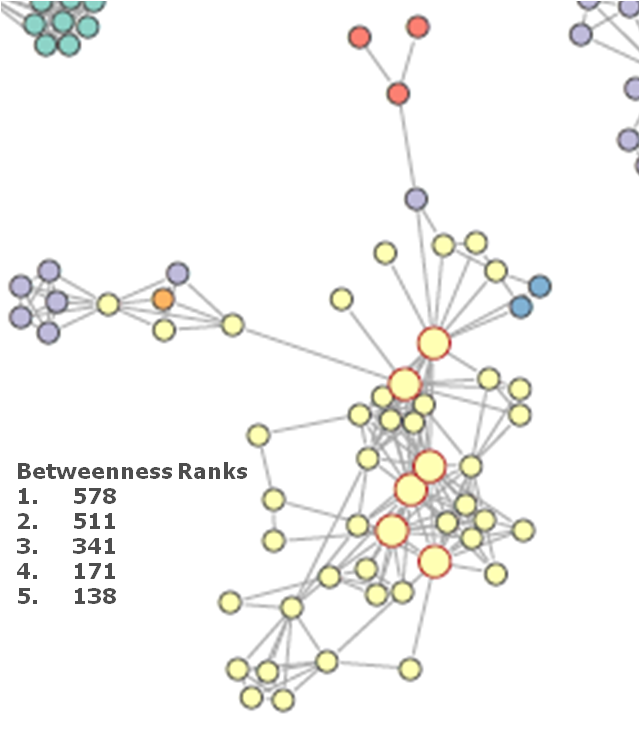
The input for this graph is from EMC's Innovation Showcase (an idea submission contest). Each circle in the graph represents an idea submitter that was part of a team (i.e. more than one submitter on an entry). Gray lines between circles represent team relationships - two circles connected by a line 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.

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340162ff81dc30970d-pi)

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:

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340162ffaf00fd970d-pi)  
  
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)

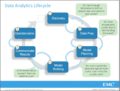
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html" \l "trackback)

Tags: emc innovation analytics betweenness

## May 29, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340168eba87a1f970c-pi)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 simulatedModel-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:

[](https://stevetodd.typepad.com/.a/6a00e5500d49008834016766a6b7e7970b-pi)

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

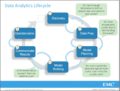
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-longitudinal-studies.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-longitudinal-studies.html" \l "trackback)

Tags: emc innovation data science scientist

## May 15, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340168eb7dfe4e970c-pi)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

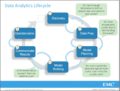
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-hypothesis-exploration.html" \l "trackback)

Tags: emc data science curriculum innovation modeling sna

## May 03, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d490088340168eb143281970c-pi)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)

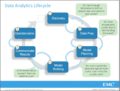
Director, EMC Innovation Network

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) |[Comments (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html#comments) | [TrackBack (0)](https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html" \l "trackback)

Tags: emc data science scientist innovation curriculum

## April 26, 2012

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

[](https://stevetodd.typepad.com/.a/6a00e5500d49008834016765bffa3e970b-pi)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