Whitepaper

Tackling Big Data: A Data Scientist's Perspective

NVP
NewVantage Partners

So you've got some data. You've read the global reports and the newspaper headlines. 'Big Data' is here to transform the economy and your business. It will make operations more efficient, increase the effectiveness of marketing campaigns, and help create better products for customers. All of these things are true, but too often these headlines miss the crucial question: how do you get there? A recent commissioned study by DataXu found that while 75% of the 350 companies they surveyed agree that more customer data would improve business, 70% said they did not have the skills or tools required to analyze it[1]. Data may be the new oil, but cars don't run on oil. Likewise, businesses don't run on data, but on information to inform their decisions. Raw data must be refined, transformed, and carefully distilled into a coherent story that shares some truth about the process that generated it. This refining process is where the value of data lies, not in the bits locked in servers and warehouses.

This refinement begins with a question: What is 'Big Data'? The definition is fluid. Some data sets are big because their size creates storage problems. Others are big because the variables are connected by a complex set of relationships. However, the precise definition of big is not necessarily central to the question of how to analyze data. Consider two types of businesses. To the first, dealing with the inefficiencies inherent in working with big data is key to their business. Google must find a way to search through all sites on the web or else its service fails. To the second type of business, the size of the data is not the problem, it is the solution. To them, big data offers the opportunity to gather information on parts of their business that were previously inaccessible. The challenge is to apply creative and innovative data driven approaches to aspects of a business that have previously been unconsidered.

As a current PhD student at the Massachusetts Institute of Technology (MIT) and data scientist with New Vantage Partners, my work focuses almost entirely on the 'how'. My work in the MIT Human Mobility and Networks Lab (HuMNet) look for patterns of human behavior in terabytes of location and communication data recorded by mobile phones. I know that this data can make our cities more vibrant, sustainable, and fun. The challenge is figuring out how. Others are often envious of the amount of data we have to work with, but once you start wrangling big data, you quickly realize there is no magic formula to extracting insight and value. However, this does not mean it is useless to have a plan. A gemcutter may have a vision for an uncut diamond, but he must respect the natural structure of the stone. Similarly, the data scientist must balance creative innovation with honesty about what the data is actually saying.

Through my personal struggle to strike this balance, I have developed a few guidelines to keep myself creative, focused, and grounded. I will refer often to the research process I leveraged to complete a recently published study on the role of geography and mass media in the diffusion of innovations[2]. These guidelines are not meant to be a checklist or a rigid procedure. They are best used as a tool for stoking the creative process.

Know your business. Know your data. Even the most sophisticated data mining or business intelligence tools are useless if they are not appropriate for the data they are applied to. Moreover, without an understanding of the data going into a complicated analysis, there is no hope of understanding what comes out. Every data project should start with an intimate introduction to the data itself. Each analysis should begin with the question - Where does the data come from and what process is generating it? At the start of my thesis project, I was given a data set on the first 3.5 million users to sign up for the microblogging platform Twitter. I made sure to find out as much as possible about how the data was generated. I learned that my collaborator had extracted records directly from Twitter's servers with no guesswork or middle men involved. The data included only users who signed up in the US, but there was no way of knowing whether these individuals were ever active on the site. Knowing where your data comes from is the first step in narrowing the list of questions that can be asked about it.

Just as important as the origins of the data is its scope. Is the data structured with columns that include records of time, date, place, or quantities such as dollar amounts? For each of these variables, what does the distribution look like? What is the minimum value, the maximum? A data scientist should be intimately familiar with the mean, median, and variance of his or her data. Are there correlations among these different

[1] http://techcrunch.com/2012/03/26/dataxu-study/

[2] Toole, J. L., Cha, M. Gonzalez, M. C. "Modeling the adoption of innovations in the presenceof geographic and media influences", 2012, 7(1): e29528. doi:10.1371/journal.pone.0029528 2012

variables? To complement the observed statistics of the data, a good data scientist should also form an understanding of what one might expect these values to be. The distribution of incomes within a country should span a much larger range than the distribution of pants size. These expectations help a data scientist spot outliers or biases. Increasingly unstructured data is also collected. Blocks of text from emails or tweets can reveal sentiments and contain patterns of information flow within a process or organization. Though dealing with unstructured data may appear like a daunting task, even simple analysis techniques like counting words can be very valuable. Finally, it is important to understand the link structure of data observations. Is there a unique ID that can link multiple data points to the same customer? Is there a temporal ordering to observations that can be used to track a process from start to finish?

Let us apply this to Twitter's adoption in the US. The data columns included unique user ID, the date and time that user signed up, and the city and state that they had signed up from. Data was collected for a roughly 2.5 year period beginning only a few weeks after Twitter's launch in March, 2006 and extended through August, 2009. Users were recorded in roughly 16,000 unique cities, however, only 408 cities ended up with over 1000 users by the end of the collection period. This small subset of cities accounted for nearly 2.5 million of the 3.5 million total users, which was consistent with the distribution of city population sizes. This type of information focuses questions. The diffusion of innovations is an active line of research, but geography at the inter-city scale had rarely been considered. This presents an opportunity because it is not entirely clear what role geography should play. On the one hand, I can invite a friend who is on the other side of the globe as fast as I can invite my neighbor. On the other hand, I am far more likely to be friends with the person living next door than someone across an ocean. This leads to the question: was Twitter's spread local, relying on word-of-mouth from neighbor to neighbor, or was it completely uncorrelated with space?

There are countless sophisticated statistical tests to determine answers to specific hypotheses. The danger is that these tests are used without providing any intuition. In the exploratory phase, it is often far more useful to visualize than calculate. Plotting and animating the data utilizes the thousands of years evolution has spent optimizing our brains to quickly identify find patterns. Of course, statistical tools should always be used to test these patterns, but for efficiency, few things can beat a visual inspection of the data. To this end, I created a few animations to display how the number of Twitter users in each city grew in time. As I watched the cities swell on a map and the time series of the number of new users steadily rise, patterns began to emerge. Twitter's spread across the country shared features of both local and non-local adoption. Suburbs within the same metropolitan region began adopting around similar times, but there was no obvious flow from one city to its nearest neighbors. The time series held a few unexpected surprises.

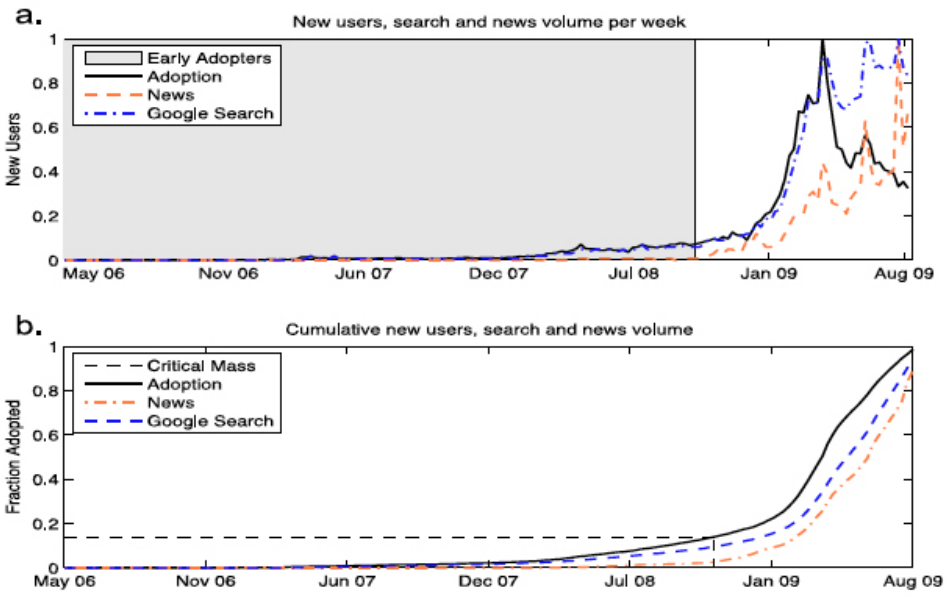

Figure 1: *National Twitter adoption trends. a.) A time series of the number of new users per week as well as news and Google search volume during that time. b.) Cumulative time series of these same values.*

NVP
NewVantage Partners

Twitter's growth in the first few years was steady until a sudden explosion around April, 2009. After this, huge spikes in new users appeared seemingly at random every month (see Figure 1). This was something that needed to be explained.

Point to the fence. The exploratory phase of data analysis can last forever. There are always new ways to combine, multiply, and resample data. At some point, the search must narrow, a hypothesis must be generated, and an effort must be made to test them. Setting a goal for an exploratory data analysis project is a difficult task. It is a balance between creativity and feasibility. Much like the baseball player who points to the location he plans on sending the next pitch over the fence, a data project should have a concrete vision. This vision should be something that finishes a sentence like "Wouldn't it be amazing if..." In the case of the Twitter data, wouldn't it be amazing if the geographic locations of places like Silicon Valley determined the path of Twitter's adoption across the country. Goals like this keep a data scientist motivated and in touch with the overarching goal of the project, to learn something about the process that generated that data.

As important as it is set a focal point, it is equally critical that a data scientist not become attached to it. The inability to give up a hypothesis when the data tells you otherwise is a guaranteed way to get the wrong answer and produce dishonest work. The chance that your initial hypothesis is exactly correct is very slim. Be prepared to revise your goals often, constantly checking them against the data. This process is known as Grounded Theory. It was formalized by two sociologists Glaser and Strauss[3] in 1967. It is a pity that the quantitative domains do not turn to their qualitative colleagues more often. The intersection may be extremely fruitful.

Choosing the right tool. At this point, a data scientist should have a good understanding of what he or she is working with as well as a temporary target to aim for. Now is the time to choose a tool. Weighing the pros and cons of any particular technique is beyond the scope of this article. Instead, I aim to offer a process that increases the likelihood that the correct one is chosen. For example, because I had visualized the time series of Twitter's weekly user growth, I knew to choose a model that generated an S-shaped growth curve, rather than one designed to capture periodic seasonal trends.

Moreover, the tools you choose should be compatible with the fence you have pointed to. I wanted to test how the inclusion of geography affected forecasts of Twitter's adoption. For this, I needed an analysis technique that would allow me to explore the dynamics of innovation spread in space and time. There was already a large amount of literature studying the structure of social networks (who is friends with who) empirically. In addition, a related body of work had simulated information exchange on these social networks, even going as far as to include a special type of individual in their simulations who had an extremely high propensity to adopt a technology. Relating this back to my initial vision, I wondered if these early adopters might be over represented in some regions of the country like Silicon Valley. Moreover, I knew that I could measure this by comparing the number of users who signed up in each city during the very early months of its existence. Were high concentrations of early adopters the key to predicting technology adoption from city to city?

Realign your vision. As well as it would have fit with my initial vision, simply accounting for more early adopters in places like Silicon Valley did not lead to results that supported my hypothesis. To stay true to the data, another hypothesis was needed. Eventually, I came across research suggesting that people are much more likely to be friends with others living near them, than those far away. In my model, I had put more early adopters in certain cities, but never specified that they choose friends who lived near them. My hypothesis needed to pivot.

At this point, it is helpful to put yourself in the data's shoes. I am a member of Twitter, why did I sign up? Where do most of my friends live? In answering these questions, the solution seems obvious. Friends with influence over the things you buy or sign up for tend to be those who you interact with most often. Even in the world of email and instant online communication, these people still tend to be those who live close to you. My model needed to reflect this. The creative combination of geography and social networks was the key to making good forecasts of adoption.

[3] Glaser, Barney G. and Strauss, Anselm L. (1967) *The discovery of grounded theory: strategies for qualitative research*. Chicago.: Aldine.

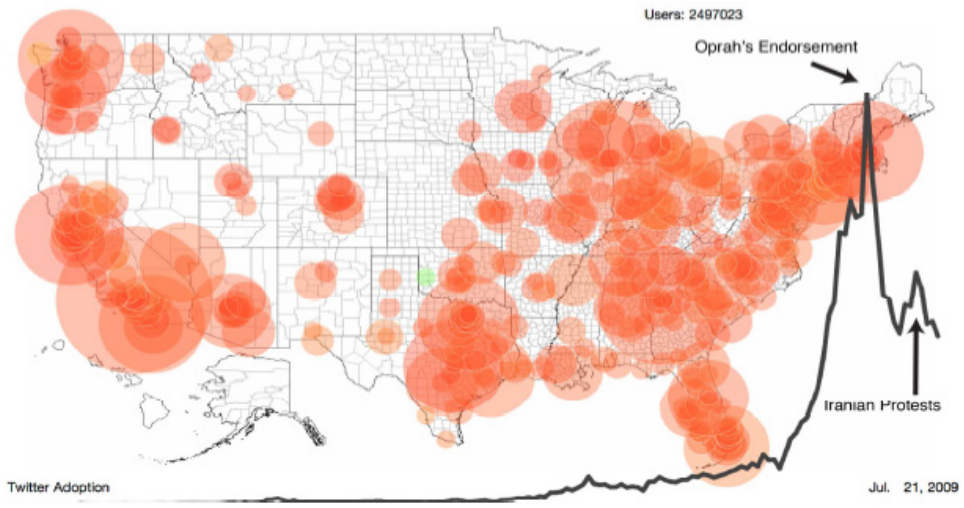

Figure 2: *Each city is represented by a circle whose size is proportional to the number of users. Large spikes in new users correspond to celebrity endorsements and major news events.*

While the insight of geography helped predictions greatly, it did not explain everything. In fact, our forecasts were very good for early stages of Twitter's adoption, but greatly underestimated Twitter's explosive growth in later stages. The key to correcting this was to put the data back in context. Remembering the results from the initial data visualization, the time series showed a huge spike in new users around April 2009. A quick Google search revealed that during this month, Ashton Kutcher appeared on the Oprah Winfrey show where both endorsed the service. This event resulted in the single largest increase in Twitter's user base over the nearly 2.5 years of data collected (see Figure 2). Again, the vision had to be expanded to include this. As important and efficient as word-of-mouth viral marketing campaigns were supposed to be, the data suggested they were at most only half of the story. Mass media and celebrity endorsements were responsible for at least doubling Twitter's users during the time frame studied. This example highlights the importance of seeking out additional sources of data relevant to your project. No matter how big your data, they are incomplete. Data are not generated in a vacuum. The world continues to spin.

Interpreting your results. At this point, you have selected the proper tool for the data. Though much of the heavy lifting is done, the interpretation stage should not be overlooked. I could accurately forecast when a city would reach a critical mass of users and how important mass media was to Twitter's growth, but these numbers and forecasts were not the value of the project. The value is understanding why the forecasts are accurate. Better predictions of future growth are useful, but understanding that this growth is the result of the very specific way that people choose their friends is transformative. You can build a business around the latter.

Everyone likes the perceived accuracy of numbers. Forecasts of profits or sales are often reported down to the penny, completely ignoring the uncertainty in these numbers. Business intelligence should not be about adding needless decimal places to estimates. Analyses should be honest about forecast errors (see Figure 3). It should be about learning something fundamental about your organization or your customer and using this to improve your products and services.

Big data is here to stay. There are still important issues to be worked out from data privacy to management and curation, but potential benefits far outweigh the costs. The 2011 MGI Report on Big Data projects shortages of more than 100,000 individuals with the analytic skills to tap data potential and suggests over a million managers will need to gain skills interpreting data and making decisions based on it[4]. The major challenge moving forward will not be how to collect and store more data. Instead, it will be finding creative ways to use this data and finding the individuals capable of realizing the data's potential. Those organizations that are best at internalizing what the data tells them stand to gain the most.

[4] http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_front ier_for_innovation).

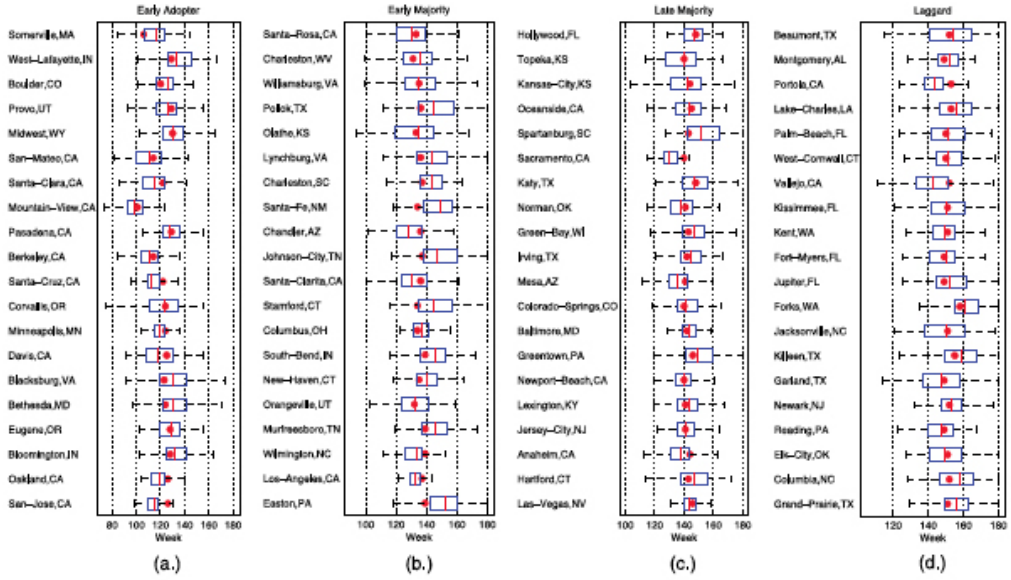

Figure 3: *Plots showing the accuracy of final adoption forecasts. A heavy emphasis is placed on displaying uncertainty in predictions.*

About NewVantage Partners

Decades before the advent of "big data", a body of business leaders, technologists, and academic thinkers were engaged by the prospect of helping companies leverage their information assets to gain insights that would lead to more informed business decisions. These individuals were pioneering new information-driven initiatives through database marketing and advanced analytics, online strategy and digital capabilities, and business intelligence.

NewVantage Partners (NVP) is a boutique management consulting firm established in 2001, comprising many of the early business and technology pioneers of information management and analytics.

Today, NVP are trusted advisors and senior consultants to a roster of Fortune 1000 clients, operating as a core team of experienced c-level business and technology executives, and industry advisors, augmented by subject-matter experts, working in small teams with executive management.

NewVantage recently announced the establishment a "Big Data" Data Science Group comprised of data scientists and experts in aspects of big data, advanced analytics, unstructured data, and high performance computing. This group will support Big Data and data science initiatives, and undertake independent research in these areas.

For more information, contact Randy Bean at rbean@newvantage.com.

NVP
NewVantage Partners

Boston I San Francisco I New York
Tel: 857-991-1404 I info@newvantage.com I www.newvantage.com

NVP
NewVantage Partners