BIG DATA: MORE THAN JUST A BUZZWORD?

Matthew Glatt
Interdisciplinary Studies 2420: Data Driven Societies
Professors Gaze and Gieseking
May 12, 2014

Introduction

Early History and Definitions

While the current understanding of big data has existed in some form since the late 1990's, the origins can be traced to the 1880 United States Census and the resulting data "panic" that ensued. Without any mechanical devices to aid in data-tabulation, the total collection and organization processes took a record seven years to complete,[1] and with that, the United States had finally entered the era of "Information Overload." Since that pivotal Census year, humans have been plagued with the task of managing huge amounts of data with finite resources. The 1880 census yielded another important piece of information: the first definition of big data. NASA researchers Cox and Ellsworth explain, "[Data] visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*."[2] This definition remained insulated in the engineering community until other entities began experimenting in big data projects.

Other, less technical, definitions soon appeared. One such classification emerged from the National Institute of Standards and Technology, who defines big data as data that "exceeds the capacity or capability of current or conventional methods and systems."[3,4] However, as technology development has skyrocketed in the past two decades, the

---

[1] Gil Press, "A Very Short History Of Big Data," *Forbes*, May 9, 2013, http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.

[2] Michael Cox and David Ellsworth, "Application-controlled demand paging for out-of-core visualization," *IEEE Computer Society Press*, 235-ff., 1997.

[3] "NIST Big Data," The National Institute of Standards and Technology, last modified June 13, 2013, http://bigdatawg.nist.gov/home.php.

[4] Jonathan Ward, and Adam Barker, "Undefined By Data: A Survey of Big Data Definitions," *arXiv*. (2013), http://arxiv.org/pdf/1309.5821v1.pdf.

traditional definition of big data has become somewhat obsolete. Manovich claims that using the industry's definition of big data all but nullifies the existence of big data in today's market because the computing can be done on commercial computers, rather than supercomputers.[5] Although, one day, projects involving huge amounts of data may require advanced computing power, the majority of current big data projects only demand nominal resources. The elusive big data definition needed to be updated.

Boyd and Crawford attempt to redefine this term by including three clauses.

> We define big data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:
> (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
> (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
> (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.[6]

Boyd and Crawford's take on big data has gained traction in the humanities space because of their social stance. However, Gartner, an information technology research and advisory firm, has developed a more popular definition that abstracts away any classifications from specific disciplines. They employ the "three V's" to define big data: Volume, Velocity, and Variety. Specifically, " Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision

---

[5] Manovich, Lev, "Trending: the Promises and the Challenges of Big Social Data," last modified April 28, 2011, http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

[6] Danah Boyd and Kate Crawford, " Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* 15, no. 5 (2012): 662-679.

making, insight discovery and process optimization."[7] This definition has been employed across all industries, and is the most prevalent in the technology industry world.

Big Data Today

Gartner's annual *Hype Cycle for Emerging Technologies* puts big data at "The Peak of Inflated Expectations." This stage of an emerging technology "produces a number of success stories—often accompanied by scores of failures. Some companies take action; many do not."[8] However, big data will soon devolve into the "Trough of Disillusionment," as expectations fall and existing implementations of big data fail to fully deliver on initial promises (see Figure 1). It is hard to pinpoint an exact time frame for this shift, as accurately measuring hype is a difficult task.

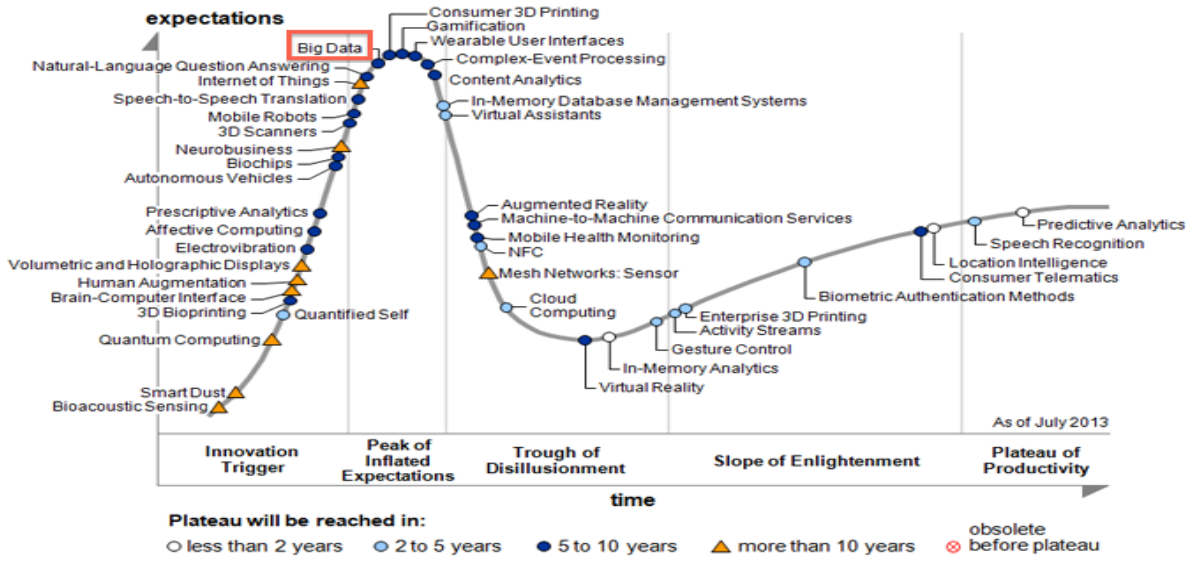

Figure 1. "Big Data" in relation to other emerging technologies in Gartner's 2013 Hype Cycle. (Source: Gartner, 2013.)

---

[7] Laney, Doug, "3D data management: Controlling data volume, velocity and variety," *META Group Research Note* 6 (2001).

[8] Fenn, Jackie, and Mark Raskino, "Understanding Gartner's Hype Cycles," *Gartner Group Monthly Research Review* (2013).

Big data's current position at "The Peak of Inflated Expectations" leads to some unexpected consequences. One of them is the "buzzwordification" of big data, or the transition of a term (in this case, big data) from a worthwhile technical term to a voguish phrase that holds little meaning. A common way to extract a buzzword's position in the network of overused corporate slang is to measure its activity (i.e. number of mentions) in corporate conference calls. In 2013, "Big Data" was mentioned 841 times in conference calls, up 43% from 589 in 2012.[9] On the other hand, "Cloud Computing" (another corporate buzzword) saw a 37% decrease in usage during conference calls from 2012 to 2013.[10]

There is validity to this type of analysis, but a buzzword cannot only be judged by its prevalence in the corporate world, rather it should put through a series of pertinent methods of measuring its popularity in social spheres, geolocation details, and investment history. Tweets containing the hashtag "#BigData" are mostly representative of the public's reaction to company announcements or general industry news related to the big data "Revolution". Twitter users are always happy to make a tweet buzzword compliant[11], and including a breaking story about big data will certainly satisfy the users' collective desires for maximizing retweets and favorites. While this causes some repetition and sub-par content among tweets dealing with big data news, best practices (of tweeting) almost

---

[9] McDuling, John, "Proof that 'big data' is one of the most overused corporate buzzwords of 2013," *Quartz*, last modified November 27, 2013, http://qz.com/151109/proof-that-big-data-is-one-of-the-most-overused-corporate-buzzwords-of-2013/.

[10] Ibid.

[11] Buzzword compliance usually pertains to products that have incorporated current jargon, however, it is also used to describe content that generates hype for a new buzzword.

always prevail. Consequently, it is always interesting to see the evolution of tweets of a topic while it is undergoing " buzzwordification ".

Hypotheses and Research Outline

Using a Twitter dataset, we can generate multiple analyses. The first of which is a word count frequency of tweet content. After visualizing the frequency in a word cloud, it is expected to yield commonly associated words with big data, such as "business," "analytics," "MapReduce," and "NoSQL." Word clouds are useful to gain perspective into what the big data community is talking about during a certain time period, but more in depth analysis is required to explore all aspects of this term.

In addition to examining the content of the tweets, it also is worthwhile to see when the tweets are published, specifically the day of week, as there is a lot to be learned about a Twitter community from a time-based analysis. I expect that most of the tweets will be submitted during business days, particularly on Monday and Tuesday, when most companies make product announcements.

The next logical step is to see where the tweets are coming from. For this section, I will focus on the United States, rather than the entire world in order to distinguish the discrete big data communities in a smaller geographic area. The map should show hotspots in all of the tech hubs in the US, such as New York City, Palo Alto (and the rest of Silicon Valley), Boston, and Chicago. I will go further and investigate areas with the highest Twitter activity to reveal any trends that may arise. After seeing *where* the tweets originate, it makes sense to see *who* in the big data world is tweeting and to determine their role in the community. This is a hard prediction to make before running any analyses, however a safe

bet to make is that the larger Twitter accounts will play a central role in the big data Twitter community.

So far, all of my research propositions and hypothesis have been qualitative in nature. It is imperative to also concentrate on the larger objective of the big data sector: driving investment. My analysis of big data investment will be twofold; first I will look at the different stage investments for big data startups over the past 20 decades, then I will be broaden the analysis to total investment in companies located internationally. Why international firms? Well, it is evident that funding is pouring into United States-based big data firms, but less obvious is the international investment market for the same types of companies. I expect that international investment will have increased over the past 20 years. Similarly, I predict that overall venture capital investment will have increased over the past two decades, although it is hard to tell exactly which specific stages will see the biggest increases.

Data Collection

Amusingly enough, amassing a dataset of tweets bypasses a limitation of big data suggested by Boyd and Crawford (2012). They claim that the "specialized tools of big data also have their own limitations and restrictions. For example, Twitter and Facebook are examples of big data sources that offer very poor archiving and search functions." In this way, Twitter may not be the best way to fully flesh out a developing topic; however, we are able to overcome the pains of superficiality of short-lived content so present in Web 2.0 websites by manually archiving pertinent information.

In order to create the most representative Twitter dataset for big data, it was necessary to choose the best hashtag to base the collection. Luckily, most of the tweets involving big data simply used the hashtag #bigdata, so it was relatively painless to locate. Using ScraperWiki, I amassed a sample of N=116,961 tweets with a collection period of a month (from 1/16/14 to 2/16/14). Due to Twitter's strict location-based privacy restrictions, only about 0.2% (N=218) of the data were georeferenced. The set did contain, however, tweet content, number of retweets, mentions, external URL's, and linked images.

ScraperWiki functioned relatively well as a tweet collection method, but there were a few shortcomings. First, data collection was only semi-automated. Every week, ScaperWiki's Twitter API handler stopped pushing new tweets from Twitter to the dataset. Additionally, ScraperWiki could not guarantee collection of all Tweets, as we do not have "firehouse" access to Twitter's API. Although not ideal, this implementation of software-sorting, or "separating privileged and marginalized groups and places across a wide range of sectors and domains"[12], ensures the privacy of Twitter users by limiting access to unverified queries.

A second data set was necessary to compile because Twitter is not a useful resource for measuring investment. Since big data is a relatively new field, it proved useful to acquire investment information from a database that tracks startups. Crunchbase, a subsection of Techcrunch, is an all-but-exhaustive database of startups, investors, and prominent people. It regularly releases a dump of its entire company database in either API or excel format in order to "make it even easier for people to dive into the history of

---

[12] Graham, Stephen DN, "Software-sorted geographies," *Progress in Human Geography* 29, no. 5 (2005): 562-580.

startups and formulate their own insights into how industries are evolving."[13] For this paper, I used Crunchbase's April 1, 2014 export, the most recent edition of their database, which includes over 700 companies and 7500 unique investments.

Findings and Analysis

Big Data Mentions in Twitter: Who, What, When and Where

Word clouds can provide an interesting look into a dataset. For this dataset, I removed common English words (a, it, the), common Twitter phrases (RT, QT, @, #), and all permutations of #BigData (big, data, bigdata, etc.). Figure 2 is the result. Prominent words in the visualization all relate to the larger buzzword "Big Data". For example,

Figure 2. Word Cloud of Tweets Containing #BigData. Common English words have been removed.

"Hadoop", is a programming framework purpose-built to handle large datasets. It is interesting and unsurprising that this phrase makes its way into so many tweets. Some of

[13] "CrunchBase Data Exports," *CrunchBase*, accessed April 1, 2014, http://info.crunchbase.com/about/crunchbase-data-exports/.

the more familiar words include "analytics," "marketing," "cloud," "IoT," and "DevOps," all of which play different roles in the big data world. However, one popular term in the big data world, MapReduce, did not make it into the word cloud. Perhaps it is too technical of a term to be included in tweets aimed towards the general consumer, or it may not often be included in product announcement tweets. IBM also has a large presence in the #BigData realm, as indicated by the word cloud. For one, they have the most followed big data twitter handle: @IBMbigdata. Additionally, they have been trying to catch up to other startups and companies in terms of big data product offerings. Their big data brand is certainly being supplemented by its Twitter presence, which will be discussed in further detail later in this section.

The number of tweets per day was, for the most part, completely cyclical, as evident in Figure 3. On average, the number of tweets during the weekend ($\bar{x}$ = 2,754) was lower than during the week ($\bar{x}$ = 5,889). This is likely due to many factors. Big data is mostly a

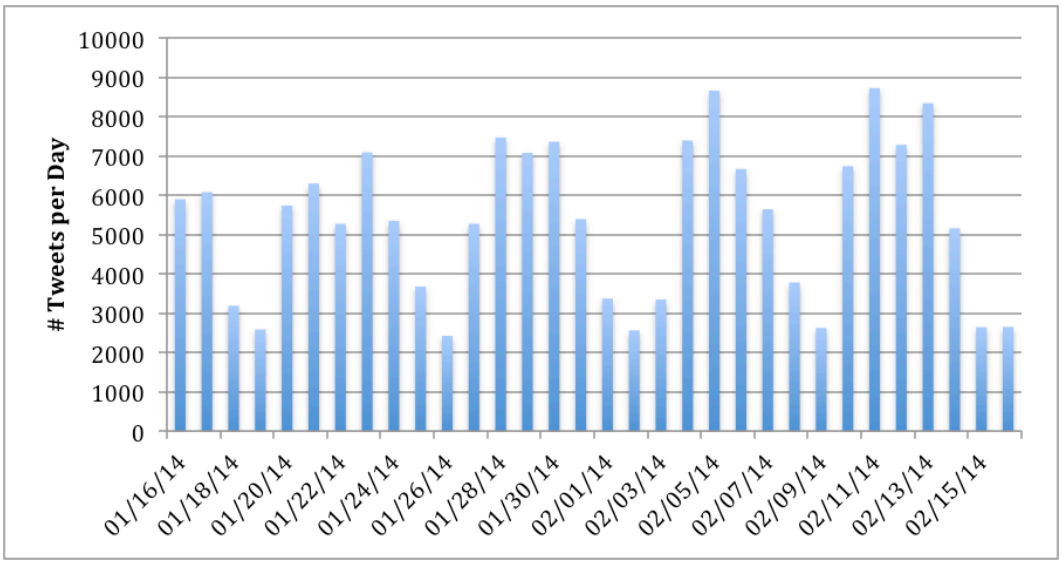

Figure 3. Number of Tweets Containing #BigData per day from 1/16/14 to 2/16/14

business-related topic, so it is appropriate that the trough section of the tweet cycle takes

place during the weekend period. Furthermore, most, if not all, product announcements and press releases take place during the week. The peak number of tweets took place on Tuesday through Thursday, which confirms this suspicion. If we were to draw a trend-line, we would see a slight increase in the number of tweets per day through this period of one month. This is not due to any activity relating to big data; rather, the uptick can be attributed to how data was collected on ScraperWiki. There was a simple week-over-week increase of number of tweets pulled from the Twitter API. We now know *what* the tweets contain and *when* they are submitted, but what about *where* they originated? Two sets of map provide a clear picture.

The first map showing the location of big data tweets (Figure 4) is a distribution of the georeferenced tweets by density (number of tweets in a certain location) in the United States. Notice that highly populated areas, such as New York, San Francisco, Los Angeles, and Chicago all have higher tweet densities. Additionally, Silicon Valley in California has a high tweet per capita, as it is a hotspot for all things big data.

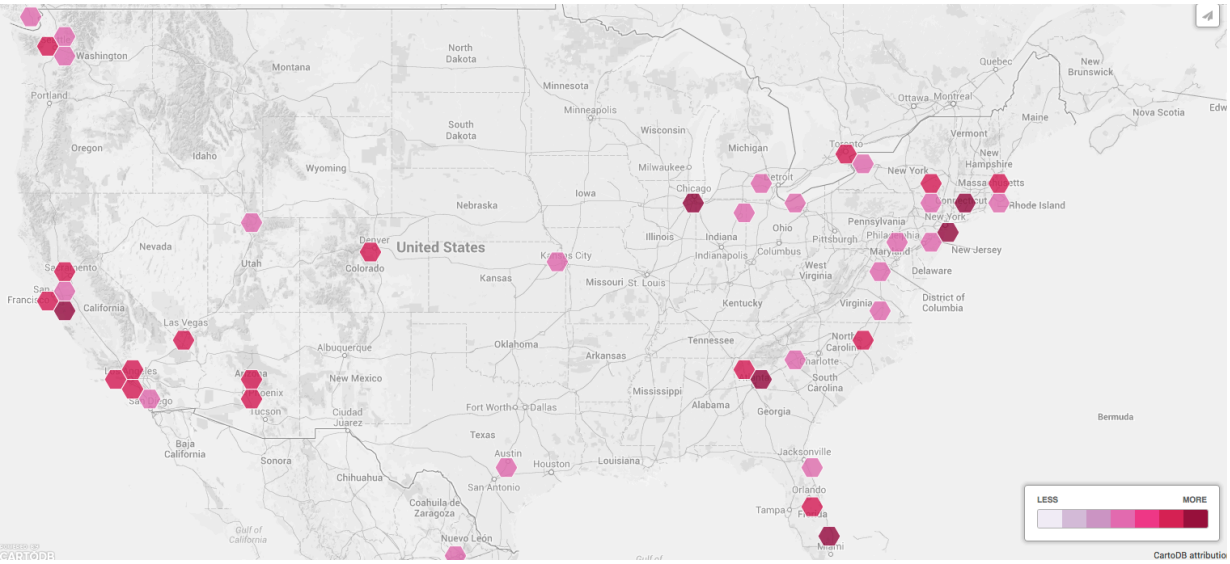

Figure 4. Density Map of Georeferenced #BigData Tweets in the United States.

So far we've covered where tweets are originating, when they are being tweeted, and why they are tweeted when they are. But what can be said about the "who" aspect? Who is tweeting about big data? Well, this question is easy to answer if it is only looked at with location in mind. It is harder, however, to see what kinds of people tweet about this hashtag. Using census data, we can visualize demographics that are usually associated with a certain interest. For the case of big data, we will look at family income changes over time. Specifically, the following maps (Figure 5) show the increase of the percentage of families

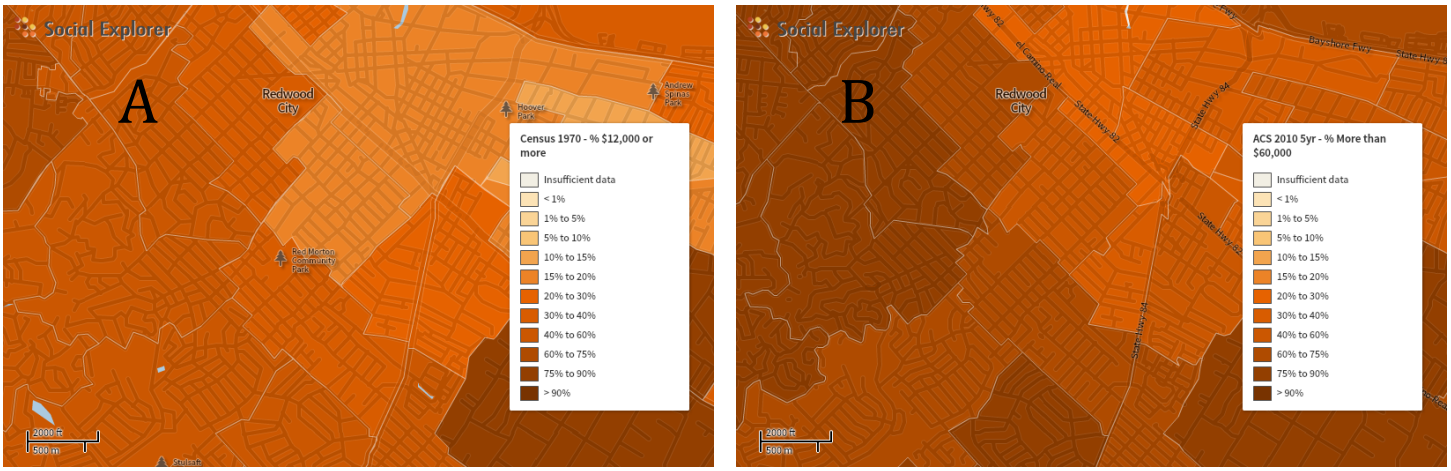

Figure 5. Change in percentage of Palo Alto families with incomes over $75,000 (adjusted for 2014 values). Figure 5a depicts the 1970 landscape, while Figure 5b shows 2010.

earning over $75,000 (adjusted for inflation) from 1970 to 2010 in the Palo Alto region of Silicon Valley.

The results confirmed my suspicions; income has increased over time, most likely due to Palo Alto's location in Silicon Valley. In 1970, only about four out of the ten census tracts had over 60% of its residents earned over $75,000. In 2010, in comparison, seven out of the ten tracts had over 60% of its residents earning over $75,000. As mentioned by

Marwick, most of the residents of Silicon Valley were able to regain their fortunes after the 2001 dot-com crash by reaping the benefits of Web 2.0 startups and communities[14].

The motivation behind the next chart (Figure 6), a network of 3,000 tweets, their submitters, and their retweeters, was to reveal patterns within the big data community on Twitter. I wanted to see whose tweets gained the most traction, what consisted of the "retweet culture", and how many users' tweets get retweeted. My initial predictions about the data proved to be wrong for this representation.

Figure 6. Network of 3,000 tweets containing the #BigData hashtag. Users are represented as nodes, and edges are retweets of the parent node.

[14] Alice Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age* (New Haven: Yale University Press, 2013), 24.

I thought that the large big data accounts (@IBMBigData, @Hadoop, etc.) would be more central in this network. However, smaller accounts, which most likely retweet tweets from the big players, saw more success in terms of degree and betweenness centrality magnitude. The big cluster in the middle of the network is actually a component consisting of smaller users (with fewer followers than the big players). This component is a giant hub, connecting almost the entire network. Upon further inspection, the nodes in this component are made up of big data reporters, engineers and evangelists with a substantial following on Twitter. This discovery produces a very interesting result and somewhat

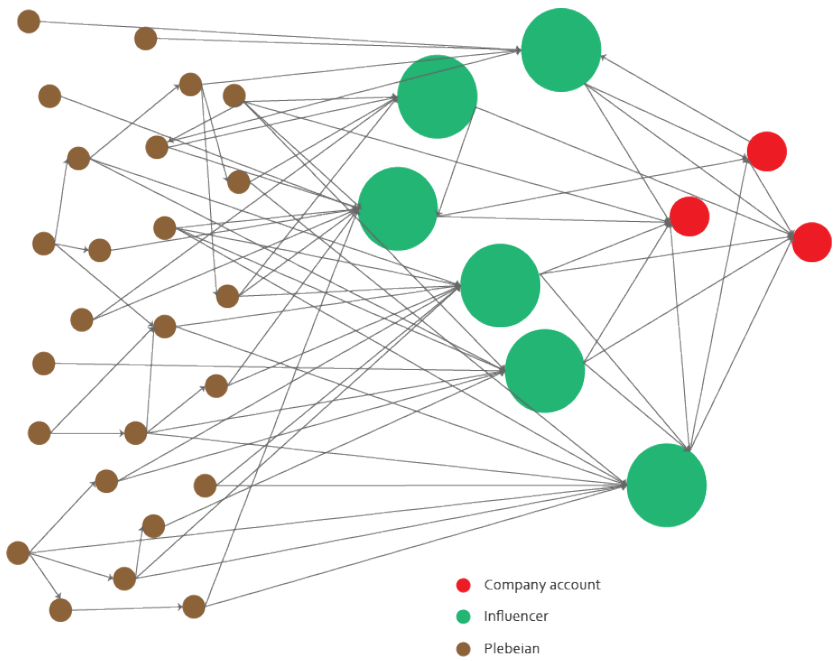

Figure 7. An example network depicting Company Accounts, Influencers, and Plebeians

invalidates my hypothesis. The large companies were not, in fact, immediately important to the rest of entities.

Figure 7 shows a hypothetical example of my network's behavior, built in Adobe Illustrator. This example network provides a possible explanation of why the company accounts do not play a more central role in the community. Here, the "influencers" have a

bigger role; they connect the "plebeians" to the company accounts. While the company accounts produce content to be retweeted, most of the "long-tail" of this community prefers to retweet the "influencers."

Measuring Investment

Figure 8 shows venture capital rounds of analytics startups over the past 16 years. It further confirms the hype surrounding big data in the recent years. All series investments have increased drastically, especially series A. Investors are eager to see what startups are capable of in this space. In the next few years, we will most likely see continued increase of series A investment into big data, a stagnation of series B investment (as companies fail to meet expectations), and a slight decrease of series C investment (as companies fail, get acquired, or become profitable). Comparing the correlation values for series A and series C investment versus series A and series B investment confirm this notion. While both sets are variables are strongly correlated, the series A/series B relationship is $R^2 = 0.04$ stronger than the series A/series C relationship ($R^2 = 0.88569$ vs $R^2 = 0.84522$).

So far, we have focused on big data within the United States. In order to provide a more complete picture of the big data landscape, I needed to broaden my analysis to international activity. Using the Crunchbase dataset, I extracted information about international analytics companies. Specifically, I looked at total investment in international analytics companies from 1999 to (Q1) 2014. What I found aligned pretty well with my expectations.

First, I believed that there would be a surge in international investment toward the end of my selected time period, and I was right. Average investment per country grew from

about six million dollars year in 2007-2010 to about 32 million dollars per year in 2011-2014 (about a 433% increase in investment per year). I also predicted that a few countries would receive most of the money. The five countries represented in Figure 9 (see Appendix), Canada, Germany, United Kingdom, India, and Israel, received 90% of the foreign investment over this time period. Also evident from the chart is the disproportionate growth of investment in the UK an India compared to the other three countries. This is likely due to the startup culture in India and the UK; the “Silicon Roundabout” and Bangalore (the Silicon Valley of India) are hotspots for big data innovation and discussion.

Discussion and Conclusions

All of these analyses shed a different light on the current standing of big data in today's market. The world cloud is a successful primary visual representation of tweet content, but there is more to be derived from the 140 characters that make up a submission.

Interestingly, most tweets containing an external URL share a news story about a new development in big data, and not an article covering ways to incorporate big data into existing business. Buzzwords typically fit this mold; for example, the term "The Cloud" has moved from the buzzword definition stage to the buzzword utilization stage. Tweets including #Cloud or #CloudComputing now often share stories explaining how to decrease storage costs or increase web application performance, rather than revealing new developments in the larger "Cloud" industry that changes how the buzzword is used. Technology industry analysts are fully aware of this process, as evident from the first page of Google results for the term "buzzword stage"; an article about big data is the first result.

Even more surprising is *who* in the community gained the most traction from tweeting about big data. A select set of individuals is able to consistently garner higher engagement from their followers than the companies tweeting out substantive content. Think about other twitter communities and the influencers that run them. Celebrities are both content generators and influencers in the world of fame, and news outlets serve as a destination for real-time content. So why does the big data community have a separate set of influencers and generators? Further research should explore this anomaly.

Big data will continue to be a hot topic for the next few years. While not in the scope of this project, it is important to remember that a representation of the maturation of this technology only using Twitter or investment data does not paint a complete picture. Other mediums should be incorporated to fully cover the evolution of big data as a topic (e.g. research, real-world implementations, financial trends, etc.). Future research about big data should lead to a society "that we can use big data to *design* organizations, cities and government that work better than the ones we have today."[15] Furthermore, big data creates the opportunity to revolutionize personal and social analytics. As long as corporates remain (relatively) in the realm of information-as-freedom, or unimpeded end-to-end communication and access in order to facilitate the growth of a vibrant, broadly participatory culture,[16] the general public will have the ability to apply big data strategies to solve otherwise unsolvable problems. Big data may exist as a buzzword in corporate settings, but in the realm of research it is very much a practical term.

---

[15] Alex Pentland, "The Data-Driven Society," *Scientific American*, October 2013, 82.

[16] Julie E. Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*, Yale University Press, 2012, 3-31.

Bibliography

Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15, no. 5 (2012): 662-679.

Cohen, Julie E. *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*. Yale University Press, 2012.

Cox, Michael, and David Ellsworth. "Application-controlled demand paging for out-of-core visualization." *IEEE Computer Society Press*. 235-ff., 1997.

"CrunchBase Data Exports." *CrunchBase*. Accessed April 1, 2014. http://info.crunchbase.com/about/crunchbase-data-exports/.

Fenn, Jackie, and Mark Raskino. "Understanding Gartner's Hype Cycles." *Gartner Group Monthly Research Review* (2013).

Graham, Stephen DN. "Software-sorted geographies." *Progress in Human Geography* 29, no. 5 (2005): 562-580.

Laney, Doug. "3D data management: Controlling data volume, velocity and variety." *META Group Research Note* 6 (2001).

Manovich, Lev. "Trending: the Promises and the Challenges of Big Social Data." Last modified April 28, 2011. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

Marwick, Alice. *Status Update Celebrity, Publicity, and Branding in the Social Media Age*. New Haven: Yale University Press, 2013.

McDuling, John. "Proof that 'big data' is one of the most overused corporate buzzwords of 2013." *Quartz*. Last modified November 27, 2013. http://qz.com/151109/proof-that-big-data-is-one-of-the-most-overused-corporate-buzzwords-of-2013/.

"NIST Big Data." *The National Institute of Standards and Technology*. Last modified June 13, 2013. http://bigdatawg.nist.gov/home.php.

Pentland, Alex. "The Data-Driven Society." *Scientific American*, October 2013.

Press, Gil. "A Very Short History Of Big Data." *Forbes*. May 9, 2013. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.

Ward, Jonathan, and Adam Barker. "Undefined By Data: A Survey of Big Data Definitions." *arXiv*. (2013). http://arxiv.org/pdf/1309.5821v1.pdf.

Appendix

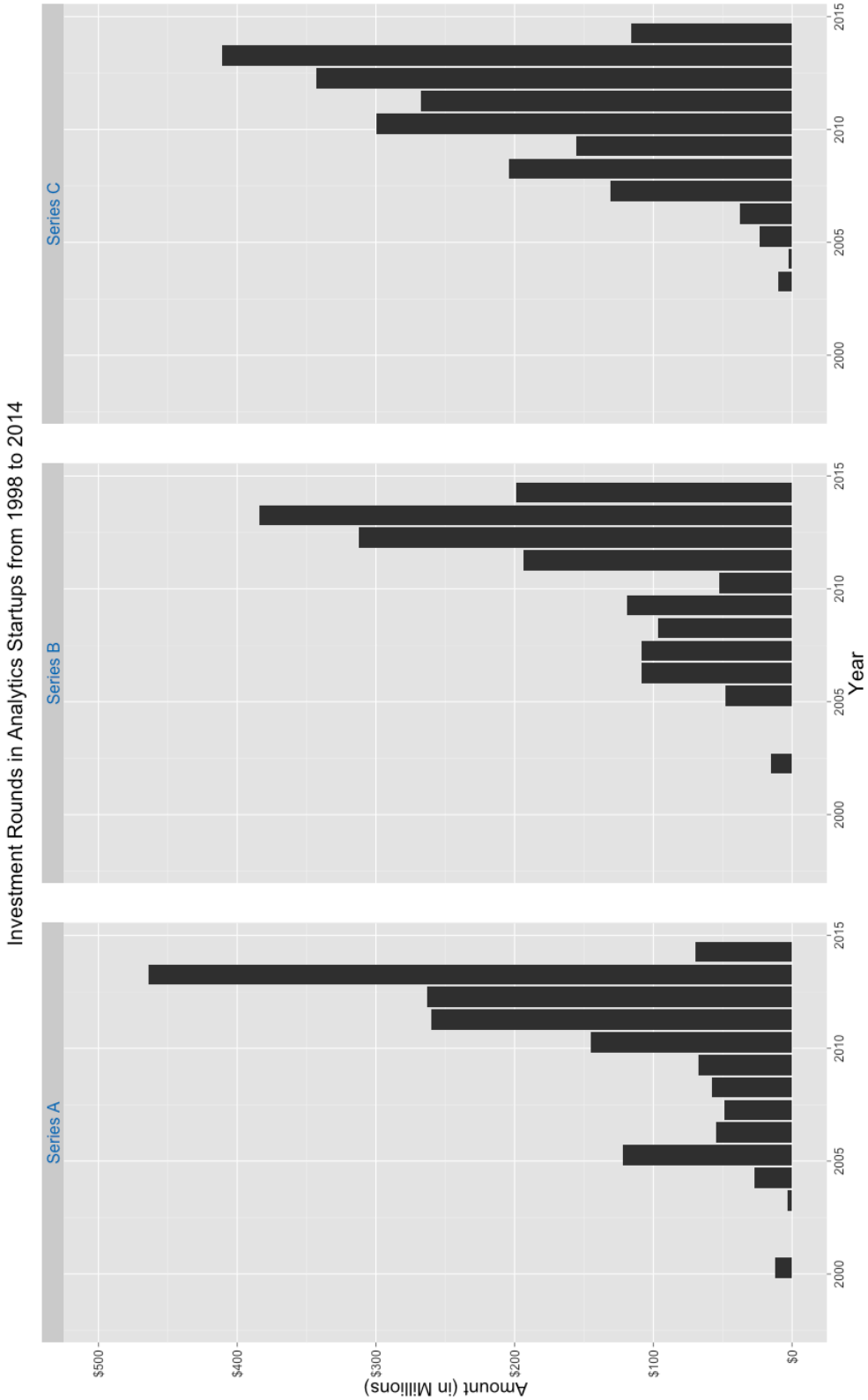

Figure 8. This chart shows the overall increase of funding for analytics companies over the past two decades

Appendix

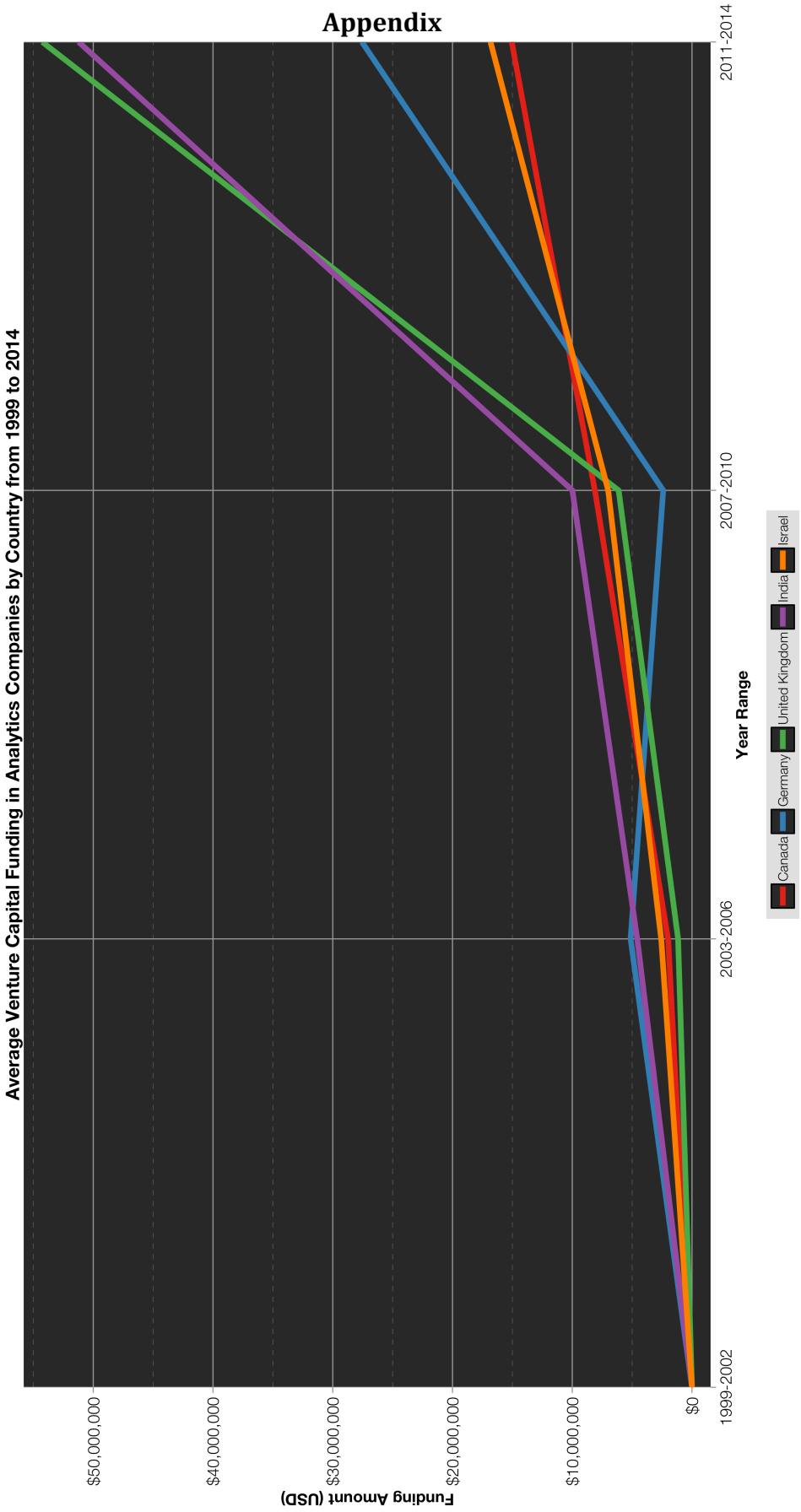

Figure 9. The five countries with the largest average venture capital investment in big data startups