# A Technical Writers Guide to Big Data Analysis

My name is Sebastian Barnes and I’m a graduate student in the technical communication program at the University of North Texas. This is the first of a series of blog posts introducing the basics of big data analysis to technical communicators. For this post I will provide a basic overview of what big data is and then explain some of the tools that are available to analyze big data sets.

## http://timoelliott.com/blog/wp-content/uploads/2014/01/big-data-speech-bubble.jpgWhat is Big Data?

Big data is the paradigm shift caused by the huge proliferation of data in the information era. Every two days we create more data than we had created from the beginning of time to 2003 (Marr). This enormous amount of data is hard to manage because of the sheer volume of information. Using traditional sorting methods it would be impossible to wade through all of this information. Luckily we have access to new tools that are specifically designed to handle this data load.

It’s important to note that, while the name “big data” implies that the only shift is in the amount of data there are other factors that make big data special. The amount of data plays a role, but more important is the type of data and how we analyze it. The kinds of data we collect on a daily basis today would be pointless without the amount of data we have. For example, companies are constantly tracking your spending habits and web searches. In isolation the data about the purchases of a single person would be entirely useless, but in the context of the litany of data that companies have about other people’s spending habits and demographic information they can predict what products and services you’ll be most interested in.

One of the classic stories of how powerful big data analysis is revolves around a high school aged girl shopping online at Target in 2012. A few weeks after some shopping she received a coupon in the mail for baby clothes and cribs. Her father went to the local Target and demanded an explanation. The Target manager was understandably taken off guard and apologized profusely. Because the man was so upset the manager decided to call back a few days later. When the father picked up he sounded much more timid than the first time he talked with the manager. You can probably guess where this is going at this point. As it turned out the daughter had been pregnant all along and her buying habits gave her away.

The New York Times explains how Andrew Pole developed the system:

As Pole’s computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Their algorithm was so effective that Target had to start including decoy ads unrelated to people’s interest to avoid making people uncomfortable with the amount of information Target has access to.

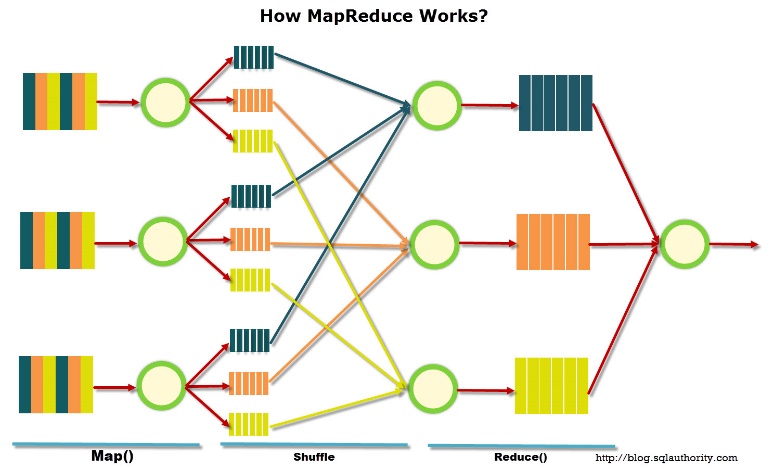
## Big Data Tools

It’s worth noting that if your needs are specific you may be able to use an existing solution, such as Google Analytics. Analytics offers information about web traffic and has a simple and easy to use interface. All you need to do is place the prewritten code in the head of your HTML file and Google takes care of the rest. While Analytics is incredibly useful, it is limited to the parameters Google sets and it offers all of you user’s information directly to Google.

Now that we know just how powerful big data is I’ll outline the two most important tools to begin analyzing your own big data sets.

The first major tool is Apache Hadoop. Hadoop is a distributed file system which enables users to use a large cluster of computers to store a single data set over multiple computers. Not only does it allow for more data to be stored but also gives access to more processors. Hadoop makes it possible to store your data set over a number of computers making the average company which only has access to cheap hardware able to store and analyze data similarly to larger companies.

The second, and more important, tool is MapReduce. MapReduce is a Java (although there are work-arounds which allow you to use other programing languages) based system which distributes the sorting process to each individual computer instead of computing all the sorting on a single computer. You could think of it similar to an assembly line. There are many individuals all doing small parts of the process instead of one person doing everything. This overcomes two major issues that data analysis had previously faced. One, it prevents the need to stream data between computers for the data to be analyzed. Two, it allows for access to a larger total processing power because processing power is distributed among all the machine.

MapReduce works in the three steps shown here. First the “Map ()” phase sorts the cluster of data on the specific machine, second the “shuffle” phase transfers the sorted data into different groups based on the parameters of the sort, and finally the “Reduce()” phase double checks the data for consistency and then places it into the appropriate repositories.

## Closing Thoughts

Unfortunately, it would take far too long for me to cover all of the coding required to use Hadoop and MapReduce in a single post. For a more in depth guide on how to use these tools, you would be best served to take an online course on the subject. <https://www.coursera.org/> offers a number of courses at various skill level that can take you from newbie to pro.

All in all, big data only requires a rudimentary knowledge of a handful of Java commands, but opens up a huge amount of knowledge to those who wield it. Next week I’ll be covering a couple of the ways technical communicators can use big data analysis in the workplace.

As new posts are added in the next couple of weeks I’ll add links below.

* Big Data and the Industry of Technical Writing (hyperlinks)
* Big Data and the Study of Technical Writing
* Writing about Big Data

References.

Duhigg, C. (2012, February 18). How Companies Learn Your Secrets. Retrieved December 11, 2015

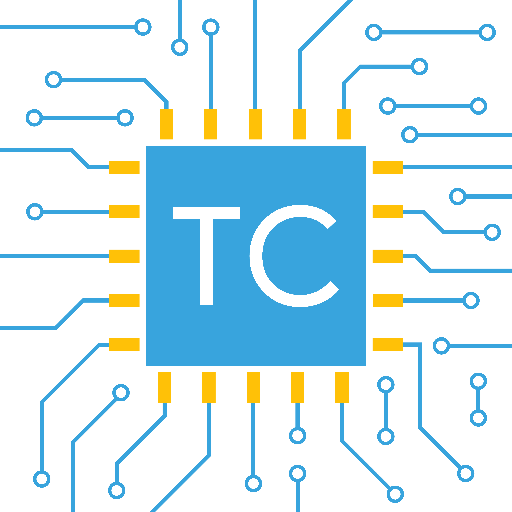
Marr, B. (n.d.). Big data: Using smart big data, analytics and metrics to make better decisions and improve performance.

# Big Data and the Industry of Technical Writing

This is the second post in a series explaining the value big data has to offer technical communication. If you are confused as to what “big data” means go here (hyperlink). I will add hyperlinks to future article here as I post them. Next week’s article will cover how to utilize big data analysis in academia (hyperlink). The final article will cover how to write more effectively about big data (hyperlink).

## Corporate Appreciation

Technical communicators have long been challenged with how they can assert their value in corporations. There are a number of issues which have caused others to struggle to understand the value of technical communicators.

The first issue is that technical communication isn’t just one thing. Technical communication encompasses the development of manuals, content marketing, usability reports, information architecture, science writing, medical writing, software documentation, blogging, instructional videos, white papers, and more. And while we do all that, many people are convinced all we do is sitting in our cubical writing manuals.

Another issue is that most of our contributions lead to cost savings and not profit generation. A successful online help database is unlikely to sell a product, but it will undoubtedly save money on calls to customer service. While that may be obvious to us as the content creators, accountants have a tendency to assume those gains are the result of other departments (Reddish)

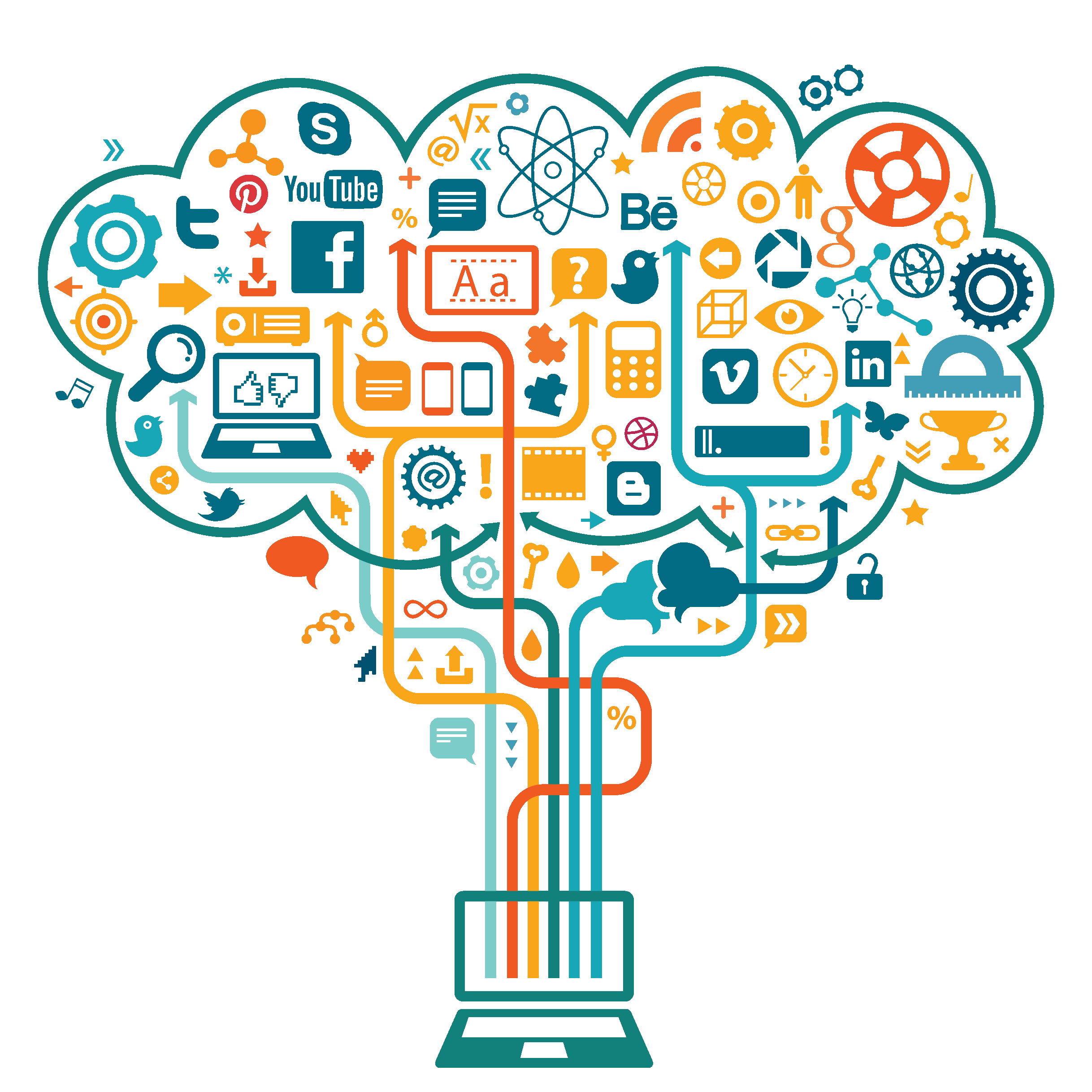
Most technical communicators today will agree that technical knowledge directly correlates with salary. Writing today is thought of as a commodity, something that can be outsources and edited by a small team of technical communicators. As a result to remain relevant in the workplace, we would be best suited to invest ourselves in the hottest new technologies and there is no technology hotter right now than big data.

Luckily our skills as technical communicators have prepared us for much of the interpretive process of data analysis. As Jordan Frith explains the “practices of technical communication will be necessary for the success of big data projects... data never speak for themselves. Someone must always speak for them.” Big data analysis is one way technical communicators can remain relevant in the face of the shrinking need for traditional writers.

## Our Future as Data Analyst

It may surprise you but most technical communicators already interact with big data sets. Tools such as Google Analytics (website stats) and Sysmosos (social media listening tool) are widely used already in the industry, but these services don’t allow you to analyze data sets outside of their specialization. By learning how to use tools like Hadoop and MapReduce we can find much more relevant information in our own datasets.

I’ve already explained my favorite story about corporate use of big data in the first post. If you’re interested in how Target figured out one of its customers was pregnant before her father found out you can read my first article here (hyperlink)

Similar to how you might already analyze your web presence with Google Analytics, other big data sets provide “a deeper level of insight into what customers are thinking and how the business operates. The potential payoffs are improving customer retention, selling individual customers more products, and producing items with higher quality and lower rates of return.” (Oracle). Much of the time you’ll be trying to find similar data to what Google Analytics looks for, but within your own data sets. For example, you might analyze how many customers utilize your company’s app in store. With this information you can create targeted ads for those users based off of their previous purchases, their demographic info, or even where they tend to spend the most time in your store.

It’s hard to define exactly what information you might look for, as it differs from business to business, but having access to multiple streams of data from your customers only adds to the information your company can make decisions based upon.

Content marketing is a huge avenue for technical communicators to use big data analysis. While you may be able to do some rudimentary analysis with tools like Sysomos having the ability to create your own data sets let you search more specifically and access to more information than Sysomos.

An interesting example comes from Arby’s social media interaction. Through data analysis, Arby’s determined many of their customers would watch the 2014 Grammy awards while interacting with Twitter. With this knowledge Arby’s had its social media department watch the Grammy awards to look for opportunities for social media interaction. When Pharrell entered the stage wearing a ten gallon hat, Arby’s seized the opportunity and posted “Hey @Pharrell, can we have our hat back? #GRAMMYs.” This tweet was retweeted over seventy thousand times and was discussed on many social media blogs the next day. Without the analytic work on the back-end this marketing opportunity would have been totally missed.

## Conclusions

This is only a couple of examples of where data analysis can be applied to the industry of technical communication. I believe that if technical communicators developed advanced technical skills, such as big data analysis, then there would be much more respect for the profession in corporate culture. As more technical communicators develop the skills to do big data analysis I expect we will find new connections between other parts of our discipline. I’m personally interested in how big data can affect usability testing, but who know what important information we might find about how to best relate to customers as big data analysis become more commonplace.

References

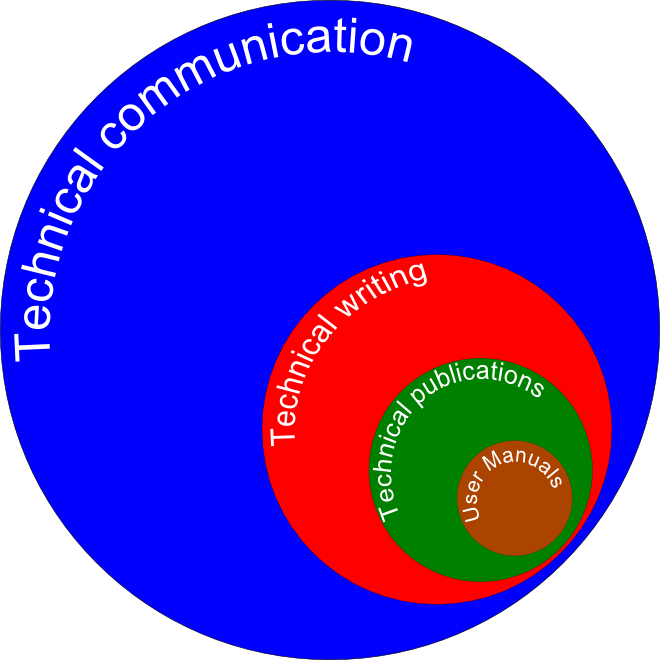
Frith, J. (in press). Big data, technical communication, and the smart city. *Journal of Business and Technical Communication.*

Reddish, J. (2003). Adding Value as a Professional Technical Communicator. Technical Communication, 50(4), 505-518

# Big Data and the Study of Technical Writing

This is the third post in a series explaining the value big data has to offer technical communication. If you are confused as to what “big data” means go here (hyperlink). If you’re looking for information about how to utilize big data analysis at your job check out this article (hyper link). Next week’s article will cover how to write more effectively about big data and I’ll link to it here when it is posted (hyperlink).

## Technical Communication and Universities

Technical communication has been in a tenuous situation with universities ever since it was created. As far back as the 1979 when Carolyn Miller was defending the rhetorical value of technical communication to her university, there have been disputes about the value of technical communication. Even today there are only five standalone technical communication departments in the United States. Because we’ve long been in the shadow of English we must take every opportunity we can to show that while we study similar rhetorical principles to traditional English departments, we point those skills solely at utilitarian applications. We’ve made huge strides in the past with concepts like usability and today we have another opportunity to do so again with big data analysis.

## The Big Data Revolution

Big data analysis offers technical communicators powerful and objective tools to analyze language. Much writing in technical communication has called for such objective tools (Blakeslee). The reason this level of objectivity is important is two-fold.

First it is opens up room for new theories about how to apply rhetoric. For example, in “The Naked Truth about the Naked this: Investigating Grammatical Prescriptivism in Technical Communication” the authors discuss the use of the unattended this. An ‘unattended this’ is the use of ‘this’ as a definite article without referencing what the ‘this’ is referring (e.g. Bring this with you.) While traditionally an unattended this would be considered a stylistic error, the research indicated that it was often used when giving bad news to avoid reminding the reader of the bad news. While this research relied on a smaller data set than big data analysis it illustrates the principle that rhetoric is often not as simple as the rules we’ve created for it and with research we could develop better principles for writing.

The second major reason is outlined by Patrick Moore in his article “Cruel Theory.” His argument is essentially that universities base the value of departments off of “prestige.” He explains this issue of prestige is a huge part of the reason there is tension between technical communication and traditional English programs. Some English programs have painted the study of technical communication as skills courses which teach principles without encouraging individual thought. If we intend to separate ourselves from the caricature some English departments have portrayed us as our best tool is technology.

I personally believe anything we can do to construct ourselves as “STEM writing” will only benefit the discipline. Research methods like big data analysis will help to construct technical communication as an objective field of study. Even if practitioners collectively recognize that all truths are constructions of rhetoric, we will still benefit from the grants and funding that are offered to so call “objective studies”.

## What Can We Do?

Developing big data in the university setting will actually not be that challenging. All that is necessary to get started is a singular repository with a large number of student papers. Once we have a singular place with a large data set the analysis opportunities are endless. Looking further into the future additional repositories containing content released by corporations or data about social media interactions with companies would also benefit the field (although privacy issues may prevent us from gaining access to such data sets).

With access to a huge amount of writing researchers could determine a multitude of things. Some of the ideas for research I’ve personally come up with include:

* What stylistic/grammatical errors affect grades/perception of the document the most
* What English as a second language students struggle with most and how to improve methods for teaching those concepts
* How document design changes how we perceive documents
* How English has changed throughout different periods of history

It probably goes without saying, but I also think technical communications programs should include big data analysis in their research methods classes. Not only would this offer all technical communicators new skills, but also help to build the prestige Moore discusses in both academia and the field. If more technical communicators entered the workforce with skills that could fundamentally change how companies interact with customers then it would reflect positively on the discipline and potentially lead to more grants and funding.

Next week I’ll be concluding the series discussing how technical writers can write about big data more effectively and how their skills uniquely facilitate the best possible big data analysis.

References.

Blakeslee, A., & Spilka, R. (2004). The State of Research in Technical Communication. Technical Communication Quarterly, 73-92.

Boettger, R., & Wulff, S. (2014). The Naked Truth about the Naked This : Investigating Grammatical Prescriptivism in Technical Communication. Technical Communication Quarterly, 115-140.

Miller, C. (1979). A Humanistic Rationale for Technical Writing. College English, 610-610.

Moore, P. (2008). Cruel Theory? The Struggle for Prestige and Its Consequences in Academic Technical Communication. Journal of Technical Writing and Communication, 207-240.

# Writing about Big Data

This is the fourth post in a series explaining the value big data has to offer technical communication. If you are confused as to what “big data” means go here (hyperlink). If you’re looking for information about how to utilize big data analysis at your job check out this article (hyperlink). If you’re interested in how technical communicators in academia can use be data go here (hyperlink).

To round off the series, this week I’ve decide to discuss two topics. First, how we as technical writers can write about big data without being intimidating, and second how our skills of rhetorical analysis apply themselves to big data analysis.

## Explaining Big Data

The phrase big data is intimidating to most people. I’ll admit, when I began this project I was worried I wouldn’t be able to communicate the concept clearly, but as I did more research I quickly realized that big data is nothing more than regular data analysis that requires the aid of a computer because of the amount of data. I think many people don’t realize the tools like Google Analytics are big data analysis tool with sleek packaging.

One of my favorite anecdotes I came across when researching this subject is that “Within decades of Gutenberg’s printing press, people were already bemoaning that there were more books than one could possibly read” (Frith). It seems the moment we’re confronted with more data than we’re used to we tend to think that it is simply overwhelming. That’s not to say I know everything there is to know about big data. For example, I have a very limited understanding of machine learning which is likely to have a huge impact on big data. That said, much like our relationship with the fields we write about, I don’t think it’s crucial that technical communicators be at forefront of big data analysis, but simply literate in its tools and the skills required to use them.

## How to Avoid the Hype

First, we should emphasize that the only technically challenging part of analysis is the creation of the parameters for the sort (and even that is just a simple Java program). While that may be enough to scare off some from the back-end part of the process, once we have the data in hand it’s nearly identical to data set we would normally have access to. Graphs don’t fundamentally change just because the data included was found with Hadoop and MapReduce.

Second is that big data isn’t simply about the amount of data. As I explained in the first post of this series (hyperlink), big data is just as much about the kinds of data, and the tools we analyze it with, as the actual amount of data. It’s also important to note, that to be useful most of this data will be very similar (e.g. the collection of customer spending habits) which means while there may be a lot of data, it’s not all that much to wrap your head around.

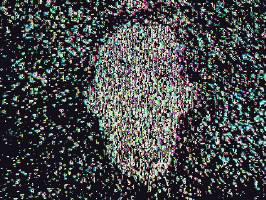
Oracle emphasize in their white paper Integrate for Insight:

“What’s different in big data is the new data sources; new types of data not previously captured, stored, or analyzed. Very large volumes of data are acquired very rapidly, and may not be neatly structured, which can make storing and analyzing it a challenge.”

So while big data is an accurate name it doesn’t quite paint the entire picture. It’s just as important to emphasize what kinds of data is collected and how quickly it is collected.

The final thing we must emphasize leads me into the second half of this article: the data analysts.

## Technical Communicators Role in Big Data Analysis

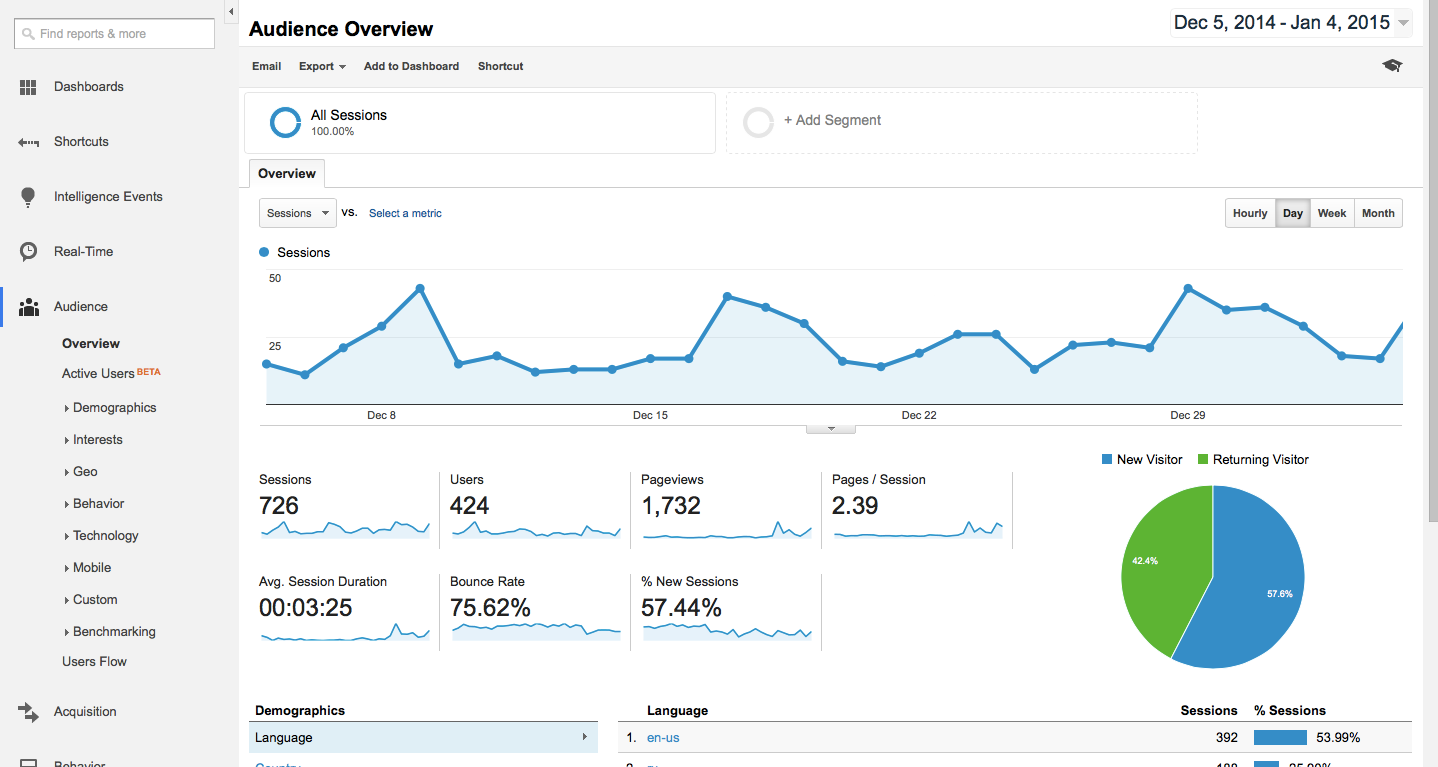
The number one thing people fail to consider when discussing big data is that even with data sets this massive, interpretation biases largely affect how the data is presented. Micheal Salvo explains big data analysis as “finding patterns in the proverbial noise of vast, unstructured data sets,” and the most important word in that description is noise. Much like white noise, big data sets are often so large that meaningless connection arise out of coincidence.

For example you may have data that shows a correlation between customers who wear red and time spent in store bathrooms. While this data point may be interesting, it is highly unlikely that this information could be applied to anything relevant. It is up to data analyst to determine which correlation are meaningful and which aren’t.

So what makes technical communicators ideal data analyst?

Well the most obvious argument is that we are experts in analyzing technical information. Most technical communicators have backgrounds in research methods and are likely to dismiss nonsensical understandings of big data that suggest that all correlations are relevant and self-evident. Furthermore, technical communicators are adept at understanding things from a broad perspective. Even the most reductive understanding of technical communication which paints us as nothing more than manual writers requires the ability to see the forest through the trees. Not only that but we excel in explain complicated technical concepts as well. Data is only useful if the right people understand it and technical communicators can ensure everyone will be able to understand data points.

The next reason would be our experience with information architecture. Tools like XML data basing are crucial to the work of technical communicators and that understanding of how to effectively structure data will play into the hands of technical communicators when creating parameter for sorting big data sets.

The final criteria that sets technical communicators apart as exemplary data analysts is our experience with data visualization. Google Analytics is a perfect illustration of the importance of data visualization in big data analysis. While this information may be useful without visualization most users would not spend the time to understand it if it was simply a table with all the data points listed. As big data interacts more with the public, technical communicators will be in an advantageous position to explain it succinctly.

## Conclusions

So over the course of these posts we’ve discovered that big data isn’t that scary after all. In fact, the only real difference between big data and the research methods we use now are the tools used to draw out those conclusions. We’ve also learned that big data can apply itself to both the corporate and academic sides of technical communication.

I think the most important conclusion we can draw from the information I’ve collected (aside from needing to learn Hadoop and MapReduce) is that the way we communicate about big data will have an impact on how it’s perceived and how widespread its use becomes. If we continue to pretend that big data just magically constructs truths out of white noise then big data will remain scary and not be applied as widely as it could be. If we properly emphasize the role analyst play in big data analysis and explain that big data is not dramatically different from how we analyze data now, then we can help to secure our place in this rapidly growing field of study.

References

Frith, J. (in press). Big data, technical communication, and the smart city. *Journal of Business and Technical Communication.*

Oracle. (n.d.). Integrate for Insight. Redwood, CA: Oracle

Salvo, M. (2012). Visual rhetoric and big data. Commun. Des. Q. Rev Communication Design Quarterly Review, 37-40.