Before starting to expand on the theoretical, methodological, and analytical work of this project, I would first like to briefly contextualize the election. The 2015 Canadian federal election is an important moment in Canadian history, the results of this decision impact many on a global and national scale. Relative to the rest of the world's political processes the Canadian election might seem a little subdued, and to some it might seem that this was an important election strictly because in the next year there is set to be another fairly important moment in history in the United States. Although some issues before the Canadian elections are often coloured with prospect of what the relationship of the potential Canadian government will be with the potential American government, the Americans are not the only reason why this election is an important moment.

Incumbents in this election, Stephen Harper and the Progressive Conservative party, have been in office for a decade preceding the election. Due to the Conservative's apparent successes, they were followed by a significant portion of the Canadian populous for that decade (Parliment of Canada, 2015). An entirely undesirable government does not pass through votes of non-confidence and reach a majority government as the Harper Conservatives did in 2011. Such successes include the *relatively weak* effects of the 2008 financial crisis, although many argue this was unrelated to the Harper Conservative's policies. As Canadians watched the rest of the world struggle through economic turmoil and uncertainty, their own lives seemed easier. Many Canadians gave credit for that ease to the Conservatives, although some Canadians do not give this same praise. While the incumbents typically do well in elections, many Canadians leading up the election regarded their vote as 'ABC', or 'Anything But Conservative'[[1]](#footnote-2).

Parties

There are four major national parties typically considered as those with a shot at winning the election or receiving major media coverage. One of which is the Progressive Conservative Party, led by Stephen Harper. As noted previously, the PC's are the incumbents to the election after a decade of office. Ideologically, they are slightly right of center, being more conservative with social and economic policy than most of the other parties. Their campaign platform in this election played off their political experience and reputation for responsible economic policy (Conservative Platform, 2015). Some party campaign ambitions were tax reductions, balancing the federal budget, job growth, and reduction of drug-related crime. Most of their platform was centred around the economic stability of Canada that they could provide and the other parties could not, even notable was one of the party's slogans “Protect our Economy”.

As a member of the 'ABC' parties, the Liberal Party of Canada is often seen as the traditional opposition party when there is a Conservative government. Most Federal governments in Canada swing between the Liberals and the Conservatives with few exceptions. This was not the case in the 2015 election however, the NDP was the opposition leading up to this election. Ideologically the Liberal Party is a little left of centre in Canadian politics (Liberal Party, 2015). For many Canadian's the Liberals represented a change from the Conservative party, but in the form of change they were familiar with due to their long history in Canada. This familiarity with the Liberals in part comes from their many federal governments, but may also be because the leader of the party is the son of one of Canada's most well known prime ministers, Pierre Trudeau. Justin Trudeau is also the youngest candidate in the election, and many of the Liberal Party's more progressive policies which reflect an appeal to a younger demographic may well have come from their young leader. The party slogan 'Real Change' can be seen as a reflection of this attitude in the Liberal Party.

While the Liberals may traditionally be the opposition party in Canada, in this election they were not they only party with potential to win. Another ABC party, the NDP led by Thomas Mulcair, was thought to have strong candidacy. Part of the reason for this perception was Mr. Mulcair's strong presence in the House of Commons (Global News, 2015 & Canada.ca, 2014 & Ottawa Citizen, 2014 & CBC, 2014). As leader of the NDP, from 2011 up to the election, Mr. Mulcair was the leader of the official opposition in the house of commons, and by no means could it be said that he quietly held that position. Mulcair was frequently seen in headlines, depicted as a strong voice opposing the policies and plans of Prime Minister Harper, leading a lot of voters to see him and the NDP as very strong possible candidates for the 'anything but conservative' vote. Ideologically the NDP is left of centre, quite a distance away from the Conservatives on a number of key issues (NDP, 2015). This distance from the Conservatives was reinforced with the recent win of Albertan provincial election for the NDP. Being one of the more conservative leaning provinces in Canada, the win in Alberta was thought by some to confirm the NDP's apitude for replacing the Conservatives federally (Global News, 2015 & Toronto Sun, 2015 & Ipsos, 2015 & Winnepeg Free Press, 2015).

The Green party, led by Elizabeth May, has traditionally gotten very low percentage of votes in Canada (Wikipedia, 2015). This party's main focus is traditonally acting as an advocate for environmental issues, raising such problems before the house so as to give them a proper hearing in government as a response to the lack in the top parties (Green Party, 2015). Never holding a federal government, and never being in the official opposition, meant that the Green party had a strong reputation for being a 'wasted vote'. Although this reputation carried some doubts leading into the 2015 election, climate change and the global impact that some of Canada's industries has on the environment became an issue more present on voter's minds. Concerns about climate change, along with May's repeated assertions that they are not a single issue party, citing Canada's problems of inequality, student debt, sustainability, meant that it is not unreasonable that the Green party's policies might not be very far out of line from what many voters would like to see in government.

Issues

There were a host of issues and political talking points that each of the parties had a stance on. While it may be unnecessary to discuss how each party felt about every issue, it is very important to discuss what kinds of things Canada was dealing with leading up to the election. One talking point of the federal election that almost always comes up is the national economy. There were some major issues, one of which was the coming recession which Canada was expects to endure (CBC, 2015 & Time, 2015 & Macleans, 2015). A low dollar meant rising cost of imports, a problem leading to Canadian's expecting to see their their day to day living getting harder.

A major economic talking point for the election was the national budget, and each party felt that they had to define their position on this problem. Liberal's felt that a balanced budget was not that important, stating they would run a deficit for years after the election (Liberal Party Platform, 2015). The Conservatives and the NDP felt that the budget was important enough to voters that they explicitly stated would try to balance it during their time in office (Conservative Party Platform, 2015 & NDP Party Platform, 2015). Growth was expected to slow, and each party more or less towed the ideological line of their parties in order to achieve said growth. A related issue to this was proposed pipelines, exporting oil from Canada to the United States and another across Canada (Global News, 2015 & The Globe and Mail, 2015 & Financial Post, 2015). This was an issue of economy vs. the environment, and while it represented some subtle distinctions between the three major parties, it effectively positioned the Greens against the rest. Also of note, nearing the end of the campaign the Trans-Pacific-Partnership agreement was approaching its conclusion. While this was not a major issue in the news cycles, it was an important issue which did divide the parties to some degree.

Healthcare is another theme for the election, often being a source of major division between the parties. One issue was euthanasia or assisted suicide. This is a topic that has been relatively well developed in other countries and on a provincial level, but this was one of the first elections in which the issue became something that some parties felt they needed to include in their platforms (Conservative Party Platform, 2015 & Liberal Party Platform, 2015 & NDP Party Platform, 2015). There was also the issue of pharmacare, an alleged hole in the Canadian national healthcare plan (The College of Family Physicans In Canada, 2015). This was framed by each party differently, some focusing on the economic impact of such a policy and others as a health and fundamental rights issue. Another issue that bleeds the lines between main themes is marijuana legalization, which also represented a very strong division between the parties. Conservatives framed this issue as one of public safety and crime (Conservative Party Platform, 2015), while as Liberals tended to argue that this could not only be an health issue but one of economics as well (Liberal Party Platform, 2015).

There was also the theme of climate and the environment. This issue is not strictly Canadian, much of the policy that results of these discussions will have a global impact. In the months following the election, the world is set to meet to at COP21, a global conference on climate change[[2]](#footnote-3). The government elected in this campaign would need to go to represent Canada in these conferences, a nation which has a global reputation for being one of the worst emitters (E.D.G.A.R Joint Research Center, 2015). Part of this comes from our strong economic dependence on oil and agricultural industries, an issue which divides the parties sharply. Again, one party may feel that the oil dependence is a strictly economic issues and ignore the climate consequences, while another party may feel just the opposite.

~~Another main theme for this election is foreign policy. The decisions made in this area have global impacts, and each party made it clear where their intention on this problem lay. It was in these intentions that many of the parties differ. An especially important issue surrounding foreign policy was the global threat of terrorism represented by ISIS. While no party takes this issue lightly, their responses to this issue vary greatly and because this issue is so emotionally charged it represented a major voting factor for Canadians. Some felt that the response was a military invasion to isolate the problem to the affected areas, and others felt that a more hands off approach was best for Canada, sometimes citing the problems that the American's felt during their involvement in Iraq and Afghanistan.~~

~~This problem leads into two domestic problems as-well, which also were approached differently by the parties. First was the issue of the Bill C-51, which many believed to be an infringment on Canadian's privacy while others believed to be a nessecary course of action to protect Canadian's security. The policies presented by this bill were fairly broadly scoped and so its impacts were aswell, but one of the most concerning of issues surrounding this bill was the implications that it had for digital communications. In the wake of the Snowden reveliations about the NSA and the patriot act, Canada was argued to be getting its own version of digital spying. Another domestic issue that greatly imapcted the election was the global refugee crisis, a direct consequence of the Syrian civil war. Few of the issues of the election were as divisive or as emotionally charged as the refugee crisis, probably because it engaged with such vigor so many issues - economics, humantarianism, foreign policy were all shades of this discussion.~~

~~Distinctly Canadian were the First Nations issues, some of which were brought to the Canadian awareness after the 'Idle No More' protests which occured years prior to the election. Despite a number of urgent issues, for a long time they were unfortunately disregarded by politicians. They include the missing and murdered aboriginal women, who greatly outnumber any other demographic in Canada, and yet there has been little investigation to the cause of this difference. Another issue is the failing reserve system, where many live without safe drinking water or stable housing. The prevelance of drug and alcohol abuse in this community is thought to be a direct consequence of the Canadian tendency to ignore these problem of integration of the communities. The Truth and Reconciliation commision had its hearings in years prior, and was set to release its final report in the months after the election. Many first nation peoples found this report important as whoever gets elected will have to deal with the Comission and it's report and likely act on those findings. This report could represent a significant change moving forward in how Canada works towards resolving these issues, and also represents an issue which is highly divisive and important to the election.~~

Twitter and Politics

These examples of issues facing Canada are not meant to be exhaustive of every single election topic, only to be illustrative of the some of political, social, and economic contexts in which the election occurred. These issues are all relevant to the election and Canadian political process, but what is of particular importance to this work is how people discuss these issues, and how that discussion can be mapped along a narrative. One way people discuss political matters is via the internet, and on channel through which they do that is Twitter. Twitter is a fairly commonly used micro-blogging social network, where in 140 characters or less users can share their thoughts, feelings, and opinions with the world. This makes Twitter a fairly unique tool for politics, since discussion is limited in some ways and cast to a large enough audience. Many have thought that Twitter has a significant impact on the political process and in this case political analysis (Small, 2011 & Tumasjen, 2010 & Choy, 2011 & Wang, 2012 & Eltantawny and Wiest, 2011).

While certainly not the intent of this work, to illustrate its political salience there is some existing work which suggests the discussions on Twitter may be good predictors of the outcome of elections. Tumasjen (2010) from Germany has some work on sentiment analysis of tweets, closely related to the type of methodology which is used in this project, which suggests that sentiment of Tweets in an issue is actually a fairly good predictor of the preferences of voters. While this is certainly not to say that polls can be fully replaced by tweets, it is interesting to see it as a microcosm of the social world reflected in tweets. This isn't the only work, Wang (2012) has done a very similar process for American elections, finding much of the same results. One interesting note that will be taken up later is that this analysis would be impossible maintaining the same types of limitations imposed on this work. Prediction is not totally impossible too, Choy (2011) found that by using Twitter they could fairly accurate predict the top contenders for elections in Singapore. One thing to note about the existing work that is done, is that it varies on degrees of openness of technique and software. As will be explained later, this openness is of fundamental importance to digital methodology.

One of the most well-known examples of the influence that Twitter has on the political process is the 2011 Egyptian Revolution. Unlike this project, the impact that Twitter was said to have on the political process was not only to discussion, but also to political organization (Schwarz, 2011). The political process itself in Egypt is different than the one studied in this project. While this project studies an official federal election, the case in Egypt was a full blown revolution. This meant there are factors which make it a very different case. None the less, it’s important to note some of the existing findings from these studies because they can show how important Twitter can be to politics. During a revolution, one of the most significant factors in it's failure or success is resource mobilization(Jenkins, 1983 & Walsh, 1981 & Anduiza, 2014 & Theocharis, 2013, & Edwards, 2013), which is something that Eltantawny and Wiest (2011) found to be a fairly significant point of salience for Twitter. The platform allowed for mass organization and execution of social movements like boycotts protests and demonstrations. Not just Egypt, but around the world also because Twitter is a global platform there were countless 'solidarity' movements in countries around the globe (Wikipedia, 2015). Part of this global nature of these movements is thought by some to have sparked the Arab spring, contributing to many of the issues which face the Canadian election in more recent years.

Another fairly significant factor to political processes is dissemination of information, and even more so to something like a revolution in Egypt. We are All Kalhed Said (Eltantawny, 2011) was a Facebook group dedicated to one activist who was killed, this was thought to be a precipitating factor for the Egyptian revolution(Jaradaliyya, 2012). This group quickly became a digital vector of information about the protests to pass through, often giving activists near real-time information about what is happening in many parts of the country and at protests. Although this group was on a different platform, the capability is there for Twitter just readily as it is for Facebook. The biggest influence that a platform like Twitter has on the political process is this, the speed and scope of information and discussion.

In Egypt there were a number of dangers associated with the digital discussions that are worth discussing briefly. First, the revolutions may have grown too attached to the digital platforms. This was a fairly big problem during the revolution, the government had the power to cut internet access across the country and decided to excersize that(The Gaurdian, 2011, BBC, 2011). What this meant is that much of the protester’s lines of communication were cut and without the infrastructure to make more resliant versions of these communication lines the protest would not last a particularly long time. This was a problem for a few days in Egypt, but eventually the internet was back on, but it does warrant activist's introspection about depending on something state owned when protesting the state. There is also the problems of spying and tracking, which can be prevelant in all forms of public discussion. Spying is really only possible when the identities of people are reveled in their discussion, an issue which can be mitigated much better online with digital anonymity. Protestors discussing on Twitter are not anonymous however, often underestimating their digital footprint. This means a government can theoretically track a protestor despite their efforts, and while not much of an issue in Canada, during an Egyptian revolution opposing a dictator it can be critical. Finally, there is the problem of astroturfing, which is the act of faking an opinion in a public forum to give the illusion of the number of opposing views(Cho, 2011 & Zhang, 2013 & Lee, 2010$). What this really means for digital discussion is that information passed so quickly through these channels is incredibly difficult to verify, and from that it makes trust a fairly difficult thing to have. Again, much of these problems in Egypt are not particularly important for an official election, but they do shed some light on the types of ways in which Twitter is salient to the political process.

Twitter still plays a role in the Western world, where revolutions are significantly less common and more regulated political processes happen. When Barack Obama was elected President of the United States in 2008, and then again in 2012 it was thought that much of his success came from his contributions to online discussion (Cogburn, 2011 & Shirky, 2011). This is of course not to say that he won because he tweeted and updates a Facebook status, since he probably doesn't do that himself anyway. But much like Kennedy's success might be in part due to his television appeal, it is important in the contemporary political landscape to have a digital prescence (Piasecki, 2009 & Time, 2010). While uncertain at this time, its still possible to see similar trends in the 2016 election, with most candidates spending a good amount of time contouring their digital personas. Much of the work being done in this project has already been done for the 2012 presidential elections, though it is important to note that this work was done with a far smaller budget than those projects and much of the findings cannot be as confidently stated as theirs.

And now Canada is starting their own election, which no doubt will be discussed on Twitter and is the subject of this project. A lot of discussion and political influence has sprung up in the online domain regarding the election, and one of the topics which centers this discussion is the “ABC vote. Having had the same prime minister for a decade, and being thought to have a good shot at winning another election, many non-conservative voters feel that it is time for a new prime minister leading to the popular stop harper catch phrase prevelant online[[3]](#footnote-4). While these kinds of political discussion were not only on Facebook, some of the properties of Facebook loaned themselves to fuller discussion. The limit on how long a post can be on Facebook is considerably larger than the 140 characters that one finds on Twitter, which means people can express fuller opinions. Also people can reply in line with comments to posts that people make, including news storys, and it is much easier to isolate whole discussions in one place. Having lived in Alberta, and most of my own social network existing there it was especailly relevant during the large amounts of oilfield layoffs, many of the people laid off and their families were ready to share their experiences and in long form($).

It is very easy to think that Twitter and digital social media represent a major change to how the political process occurs, but it is important to temper that with realism. Yaroslav Baran feels this way, that these types of communication mediums represent an evolution, not a revolution, of the political process. Much of the online communication that happens online can be thought of for politicians as a potential hazard more than it is a site of good PR(Baran, 2011). This comes from the decreased ability of polticians and campaigns to control the amount of information that gets released online, meaning it is much harder to control an politicans image. There is also the problem that digital media doesn't reach all parts of a nation equally. It can be thought that digital media in this way really only reaches the young, since they are most often on the more bleeding edge of these types of technologies(Statistia, 2014 & Pew 2014 & Statistics Canada, 2010). This means that many of the messages that can happen online are likely to be targeted towards a younger audience, traditionally this demographic voters closer to the left and so the discussion online should have somewhat of a left leaning bias($). While this isn't nessecarily a problem, it does mean that the discussion will be somewhat biased, and therefore much of the findings of this project too will be biased. Again, this is not a fundamental problem, but an external force that should be noted before claiming some major objective findings.

Narrative

The real theoretical framework through which this election is narrative. Paul Ricouer (2010) is likely the point of first contact for many scholars of narrative, and his concept of narrativity is a relatively simple qualitative metric by which to measure discussions on Twitter. Narrative is simply the connection of two or more events in time, an example is the classic 'story' of a man going to the store. A man puts on his shoes, and then walks out his door and onto the street, then walks the route to the store and finally arrives. There are a number of events isolated here, and it is only when adding the variable of time which connects these events that a story starts to emerge. This example is a massive oversimplification. Narrative study has many nuances which lead to even more developments of this theory, but the theoretical tool used in this project is narrativity, or the degree of connections in time, or how *strong* is the narrative. This need not be a connection of many things, it can be a connection of just two but the strength of these connections is narrativity.

Narrative is an important area of study for the reason that it enables researchers to connect ideas and events over time in a way unlike a lot of other metrics. Many researchers seek to tell a story with their work, and there is no better tool than searching for story than looking through a narrative lens. Despite the form of the analysis in this project, it is important to not try and make quantative arguments based on the narrative data that can be present. Instead, narrative is a means to interpret data. For this reason, it is important to note that all interpretations of data which use the narrative lens are nessecarily qualitative. What this qualitative nature of narrative interpretation means is that researchers have control to 'spin' the message, and while efforts can be taken to minimize this risk it is impossible to be totaly objective while viewing data through a narrative lens.

I want to just make a disclaimer here about the power of narrative, and try to curb some expectations for what it might be able to do for this research. In the end, this project will not be able to claim that it has uncovered some true narrative of the election or even the discussion about the election, that is not the purpose here. Nor is it the purpose to measure the 'narrativity' for that matter, these are just an analytical framework I will be working from. I'm not entirely sure if anything really can be called a true narrative, where the full story is captured using these techniques and offered as fact. Instead, I'm just using it as a way to tell a story about discussion, and furthermore use this analysis as a means to explain the methodology which surrounds digital research. The aim in the end is to attempt to look for some kinds of patterns, make some connections between events and feelings, and try to get an overview of the discussion from above. But as will be discussed in the coming section on methodology, this overview is an interpretation and because of this it will lend importance to the following section on the nature of open source research techniques.

It might be helpful to bridge the gap between narrative and the story I'm trying to uncover by working within a slightly more limited framework of hypertext fiction and the methodological implications therein. Hypertext literature or texts are pieces of writing in a digital context where certain phrases or words can be hyperlinked to other texts which are tangential to the main storyline(Ryan, 2001). A writer may include as many hyperlinks as they want. Recall that Ricouer defines narrative as the connection of two events in time. In traditional literature, the narrative of the story may be either defined as plot narrative which can be separate from the order in which it is presented to the reader, or one could also defined narrative as the order of presentation itself. One benefit of hypertext fiction is that it allows for the expansion of each point, wherein the narrative can diverge from one point to many others. The nature of narrative in traditional literature can change then from a linear fashion into a type of network of narrative. It is this network of narrative that connects both hypertext fiction and the type of narrative found on Twitter.

There are two separate binaries that Ryan describes as part of her exploration of hypertext. This exploration is more within the discipline of game studies, but Ryan does mention quite a lot about narrative and this paper is specifically about the intricacies of dealing with narrative in hypertext. One is the binary of external and internal, being the position of the 'reader' in the context of narrative. In the internal, the reader is a part of the world in which they are consuming a narrative, and in the external they are as Ryan says a kind of 'god' like figure. The other binary that Ryan explores is the exploratory and ontological binary. In the exploratory experience, the reader moves about the narrative without changing it's path. For hypertext novels this might mean changing perspectives of characters or locations, but the overall path that the narrative takes is the same. In the ontological side of the binary, the readers decisions change the narrative itself. This doesn't necessarily mean they can't change perspective, but their interactions with the narrative can change it.

The narrative of the 2015 Canadian Federal Election as seen through Twitter is much like that of a hypertext fiction – before any researcher looks at it the narrative represents the potential to be any permutation of the classification scheme that Ryan sets out. It is through the interaction or absence of interaction that the user classifies their own experience of narrative on Twitter. It starts with the bounds between internal and external not being made explicitly clear. In one sense, the user who explores Twitter can either choose to embed themselves directly into the narrative by tweeting about the election, who they might vote for and why, and in that the narrative would be internal. This user who chooses to participate in the narrative in such a way also is transforming that potential narrative into real – in this case moving from the potential of either binary towards the ontological. By tweeting, the user not only embeds themselves into the narrative but they also change that narrative. If the user chooses not to tweet and limit themselves only to the observation of tweets, then their experience of the narrative changes into something which is more external and exploratory. This is the classic 'lurking' experience, wherein the user is much more like a ghost invisible to the real interactions.

Another way that the Twitter experience is a lot like the reading of a hypertext fiction is in the actual links themselves, of which there are a few varieties. There are classic hyper-links, not at all unlike those found all over the web. These links can represent tangents to the narrative, or by their very inclusion they can be thought of as artifacts of the narrative itself. The experience is that a user clicks on these hyper-links and can explore this new line of thought, often this artifact contains links to other areas of the web, sometimes tangential to the main story of the narrative. Other times these artifacts are left as symbols for larger thoughts, where the hyperlink may just be a placeholder.

Narrative analysis isn't nessecarily important for elections in the sense that it is a very good predictor of winners or losers of the election, but rather it can be thought of as an important thing because it contributes to the chroniciling of history. Again, this work isn't going to replace the polls and it isn't going to predict a winner, but it will attempt to do is tell a story about how people were feeling about general election topics during the campaign. While this is a broadly scoped work, the election is not the only concern of this project itself. Another major concern of this project is to document the challenges and charm of digital methodology when doing this kind of work from a very limited time and funding window.

Methodology

Digital methodology has many uses, and in many ways is not unlike traditional technqiues used in the social scientists. Rather than seeing these techniques as a subsets of either the qualitative or quantative stream, it is more an application of new technology to both. In some cases, digital methodologies are used as augmentations of traditional techniques. An example of this could be as simple using a digital forum or the internet to gather a larger number of repsonses to a survey than would be possible in a traditional research model. Digital methodology can be as simply defined as anything that uses a computer in part of the research process, small or large. Researchers might not rely on comptuers for data collection, but might use a script to process that data in many different ways quickly. This type of augmentation of existing techniques isn't excluded in what I would like to define as digital methodology, but I feel as though this isn't what people are refering to when they talk about digital methodology. Instead, they are likely refering to the types of methods which can realistically only be practised computationally. One major source of data for collection in the contemporary world is social media websites, and the user generated content which is contained therein.

The methods for collecting user generated content from the internet is something unique to digital methodology, it cannot be replicated offline and not using a computer. An example of one such method, and one used in the research branch of this project, is using an API (application programming interface) to collect data (Twitter, 2015). These APIs provided by social media platform corporations are tools through which researchers can directly access the user generated content on the platform. There is an abundance of issues surrounding APIs and their use in research which will be explained in later sections, but for now suffice to say that the use of these is something only possible with digital methodology.

The API method of data collection is not the only method unique to digital methodology, there is also the bulk collection of smartphone and computer use data which is possible for some researchers with certain resources and access privledges. Sometimes this data is platform specific, an example being the Twitter Search API providing the location data along with Tweets if the user had their location settings configured to provide that information(Twitter, 2015). Of course this isn't true of only Twitter, this type of data is present for most social media platforms as well and it is up to the user if they want to submit it or not. While location data is readily accessible by the public, it is not the only data which gets transmitted by devices back to a database. Quite often our smart phones and laptops are sending back all types of information, location data most commonly being checked in minute by minute intervals and logged. As one can expect, these logged transmissions mostly end up into databases which are then used by companies in other products (location history in maps for example makes it a better product, but could also be sold to advertisers).

These types of massive databases are what most commonly get refered to as Big Data. A common definition of Big Data that I enjoy is anything that cannot be processed using traditional techniques and the hardware you have in your office or computer labs (Wickam, 2015). The databases which get used for Big data applications are orders of magnitudes larger than what one collects through traditional techniques, and the computing power or time needed to go through it all with traditional hardware and techniques would make it nearly impossible. Instead, the computational social sciences form to merge the ideas of database management and computer sciences with the social sciences as a way of managing both the unweidlyness of big data and the broad implications of how to use it. It is the methodology of these computational social sciences that I find particularly interesting and will spend much of this section discussing, though I will try not exclude relevant information elsewhere.

The question that many social scientists, comuptational or traditional, have on their minds is how can we effectively use this type of big data to explain our world. There are two streams of thought, akin to those in statistics, of infertential or description. Can we only explore the social world through this lens of digital methodology or is it possible that we can explain and predict things in the social world with this type big data. It is easy to think that the science of it gets much easier with the additional of digital methods, with gigantic sample sizes and a huge number of different variables across the globe. It is tempting to think that its possible to really 'know' things for sure using digital methods. Another stream of thought is that it the world does not get easier to understand in any appreciable way($), in part this comes from the wall of issues that digital methodology presents that have no solution. These issues will be discussed shortly, but first it is important to clarify something about this project.

This project does not use 'big data'. The data collected from Twitter for this project amounts to a little under 50mb in total, and while that represents over a million tweets, it is miniscule in the shadow of real big data. For comparison, the GDELT is over 100gb (2015), so my work is 0.0005% of that and it's not immedately clear that the GDELT should even be called 'big data'. That doesn't mean a lot of the digital methodology questions and concerns don't apply to this project, just that much of the techniques and issues surrounding big data might not be described as detailed as they derserve to be. This section is more addressing the big brothers of the methods used in this project. Many of the methods used in this project are relatively naive when compared to their big data counterparts, many of them being very impractical if not impossible to deal with if the dataset used was much larger. Never the less, there are many issues and contruibutions a section on big data methodology that can be contributed to that of this project, and so it is worth while to explore some of the ontological and epistempological implications of this methodology, as well as the limitations, strengths, and ethics of such methodology.

Ontology of Digital Methodology

There needs to be some discussion of what types of questions digital methodology can ask and which it can answer. This is a window into many of the fundamental assumptions which digital methodology takes for granted, and so it sheds light on the project as a whole. Although it might be obvious, a digital approach to research can really only take a look at what is found online and on computers. This project assumes that there is some type of knowledge and truth present online which can be studied, that it isn't just chaotic noise. And while this is most likely true, if the goal of research is to study 'society', then it isn't really clear that what you are looking at online can really tell you any truths about society. Especially in the case of this project, which ostensibly looks at the 2015 Canadian Federal Election, it is not clear that what is being found online and in digital culture really is a refelection of anything that exists offline too. Without distracting from the main point, it's not clear if one is able to study an election by studying Twitter. One is able to study the discussion of that election on Twitter, but at this time it is not entirely clear that this is a good proxy for the real election. A reason for this is that there is a significant number of people who aren't on Twitter. In 2013, Macleans reported that less than one in five internet users in Canada use Twitter, with over four fifths of Canadians using the Internet in their lives (2013). StatsCanada's data suggests that most of these internet users are English, educated, young people with a reasonably good income (2010). From a strictly sampling perspective this is not a very representative sample of Canadians. More to the point, if the object of study is the election, then voters and their opinions is going to be an important thing to study. The starkest difference is in the respective ages, where Twitter users are mostly young(Pew, 2015), and voters in Canada are mostly old according to Elections Canada (2014). When employing digital methodology to research things offline, it is an important thing to question if the methods used can really answer the question asked. In the case of this project, the question is what does the narrative of the election look like online and so does not run against this problem at first, but in the application of these findings and their implications, it is important to note the methodological limitations set upon them.

A related topic is the types the data which can be collected using digital methods. Often times, the data that can actually be collected is limited in so many ways that it often limits the questions that can be asked. For example, it would be exceedingly difficult to make a research question which depended on the private interactions on Facebook using a digital method such as an API request. Not even private messaging, just the types of things which show up on user's timelines strictly because of the structure of the platform and privacy settings. It is simply not possible to answer certain types of questions because of this structure in a real sense. In a practical sense too, for example, if say somebody bulk collected tweets it would be exceptionally difficult to do an analysis based on the age or gender of those who are tweeting. The type of questions which sociologists often concern themselves with would be exceptionally difficult to answer using a lot of digital methodologies, questions such as age, race, gender, income and so forth are all shrouded by the anonymoity which digital spaces often provide to users (Mislove, 2011). It is possible with some clever computational analysis techniques to make educated guesses at some of these traits but it is not always something inspires much confidence ($). Instead, researchers need to tailor their questions to which data they can actually use online, something which is not uncommon for researchers in general but the types of limitations digital methods present are often unique.

Digital methods do have some strengths however, researchers often use them for good reason. The most obvious strength that a lot of digital methods have is that the data which can be analysed can be huge. Without an API, it would have been impossible for this project to get over 1 million tweets for analysis, it just simply wouldn't have been possible. And there is no real limit set for how big data can be, the only practical one is set by a researchers ability to weild such data. This means that with careful sampling, the quality and size of data that researchers can get can be huge. That is not to say that sample size is not confidence however, there is existing work which cases some doubt onto the attempts to argue that big data is a silver bullet to research ($). Sample size is not always going to give definate answers for a multitude of reasons and limits, but it is something that is afforded by digital methods.

Researchers also have at their disposal with digital methods the entire toolkit of computer and data scientists when they do their research. This is especially relevant for research which takes on a more qualitative approach, in that the qualitative gaze can be taken from a much higher vantage point than in traditional research. In this project for example, it would take an unreasonable amount of time to read each tweet and score it based on certain criteria and move on to the next. One could sit and read tweets and form a general opinion, but I can't help but question if a researchers bias and fatigue would tarnish the results more than if one allowed the computer to do the work for them. I'm not going to make estimates, all I'm arguing is that a computer can do it's analysis on a million tweets faster than a person can look at 100. I'm not saying the digital method is better at all, if the research questioned how do people feel the discussion went after reading Twitter and tried to research that experience than I would think that the traditional method might be better. But what the digital method allows for is just a different approach to answer a different question, and the way that it does that is through amazing computational speed. This computational speed is not only limited to qualitative work, obviously quanatative analysis is a strong suit for computers. If the data is strictly numeric, then the types of processes by which it can undergo are only accelerated by digital methods.

There is also the strength of being able to access the data directly. One downside to traditional research is that the direct data needed to be analyzed is often through some secondary source. Consumer Price Index for example can change somewhat depending on the source and their calculation of the CPI. Not a significant problem for most research, and I don't want to try to downplay the difference between these methodologies, but an example of the difference here is to try to imagine if you got the sentiment from Twitter from some source who calculated it, or if you could calculate it yourself. For CPI, this would be like getting each item to report its price to some database and being able to access that directly. The benefit of being able to use most digital methodologies is that you get a direct connection to the users and content which you are studying. This obviously has some issues associated with it. The data and content in the online world is unverifiable and often not perfectly in the format you might want. This means that while you have some direct access to that information you still need the processing capability to go through it.

This leads to one of the most poingant limitations of using digital methodologies, the priviledge of who can actually answer these questions. For this project, most of the techniques have been set out before hand by other researchers and I mostly applied them to this specific election data. Not only that, but because of the relatively small dataset that I worked with, it didn't require much of an understanding of how to weild large data sets. In studies which analyze big data however, the methods used to analyze some of that data require a researcher to be very competent in many different areas in order to achieve their goals. The impact that this has is that even if the tools are available too all, it takes a skillful hand to use these tools approproately. Not only that, but if these works are ever to be used and published, they would have to undergo some type of peer-review process, requiring a large enough body of peers who are also able to understand not only the theoretical premise of the research but to also understand the intricacies of the methods themselves. A counter-argument to this limitation though is that this is true of much of science, I'm sure not every researcher knows all the intricacies of some quantum physics experiments and could peer review them. This doesn't mean that computational social science and digital methods are inpenetrable, only that they raise the barrier of entry to other researchers to a point where it might not be entirely possible to accomplish some research tasks on a small enough scale. So while digital methods give researchers a lot of power, these researchers also are then required to know how to use it.

Another limitation alluded to before is the lack of control that researchers have over the data they use. This was alluded to somewhat in the section on APIs, but I will just briefly overview what the problems are here with digital methodology in general. Often this problem presents itself in the collection of data, as noted in the API section, the format of this data is not often something immediately useful for research. The problem is that any change one makes in this data, no matter how small it might immediately be, is echoed throughout the entire research process. Even small changes can have dramatic effects. Worse still, there are some cases where the actual collection of data is somewhat masked to researchers. This can happen through the API black boxing effect mentioned in it's section, but it is also prevelant in other cases aswell. Researchers have little control over this aspect of their research, and as such it makes being totally transparent difficult when publishing or communicating work. Imagine a researcher being questioned on the data collection process, and them being effectively left in the dark about a fundamental part of their research.

The other problem that is associated with this lack of control is that the quality of the data which researchers use is more or less unverifiable. This means that researchers often have little in the way of a mechanism which can confirm the quality of the data that they are collecting, and by quality I mean to say the relevance and appripriateness. This is in part a product of digital anonymity, with much of the same types of limitations on identity being placed upon it here. This difficultly verifying is a problem in this project that I would be remiss not to mention. There is no such mechanism in this project verifying the real identity of the users who make these tweets. It does nothing to ensure that the users are real people and not bots, and that the posts made are at least somewhat sincere. This is also a problem with the actual scoring and measurement aswell, in that the algorithim has no way to detect sarcasm and other more nuanced approaches. Of course more intricate and developed research designs can attempt to account for such things, but I doubt if they will be able to do it as well as the human brain anytime soon ($ dec 3). These problems are not death knells to the project though, I cavet all the findings by saying that in digital methodology one can only really approach content and not users. It is fair enough to say that I have somewhat accurately described the discussion that occured on Twitter, but I would not be able to say anything about the users therein.

One major academic issue with this type of digital methodology becomes apparent when discussing the open platform nature of much of the work, including this project, and the issues surrounding instituonalizing and publicizing the work. The end product of this work is a web based data explorer, which users can visit and explore the election data for themselves. All of the work, the project and analysis code, data files, and the web site code will be made available and given to the public as free domain. One hopes that in larger projects the same type of behaviour may invite other users to contribute to making the project better, more accurate. This is an exciting area to discuss, but will have to wait for the next section. What is relevant now is the relative lack of space in academia for publishing and spreading these types of work. While in an strictly pedagogical sense the Universities are fairly well equipped digitally, they are less so in the way of publishing, and more importantly contributing to research. There is no channel by which the university can publish the link to the data explorer, and this project will likely only be printed a few times and stored in some cabinet. Much the same as with hypertext fictions, digital mediums can fundamentally change the way in which this content is published. In digital contexts there are so many other options that are available, and it will take a long time for this shift to occur. For now, this remains a limtation on those who wish to use digital methodologies and be published, they eventually they will need to convert to a more analog method of publishing. In that the researchers lose much of the reason for choosing a digital method in the first place. I know that for this project one thing that makes it unique is the ability to interact with the data, and while my own findings from it might not be important, it could certainly be seen as a resource for those willing to pick up the torch. As a short tag on point to the one prior, it is also important to note that this difficulty publiushing and instituonalizing means that reproducibility is a problem. I've done what I could in my own project to try to be as transpartent about my methods, mostly for the sake of review. But it should be noted that by not being able to publish and communicate every aspect of the research, then future research would suffer from an increased difficutly in reproduction. While it would be difficult to collect data on the election in the future, one could always run the code on a different system and audit it for faults this way as long as it is included in some form of publishing, though somehow I doubt whoever finds this thesis in said cabinet will also find the codebook attached to it.

Ethics of Methodology

Finally what can be said about the ethics of digital methodology? The biggest issue that I would like to discuss is privacy online, and how that might affect the approach that some researchers take. If say one major tenant of digital methods is transparency, and if part of that transparency means that researchers much open their data sets, then I can forsee a problem with privacy. Each of these data sets would have the tweets and usernames associated with them, but if say one person were to have deleted their tweet for whatever reason after it had been collected, then it would have not been deleted fully. While unlikely, incredbily unlikely, it means that the user has effectively lost some degree of control over the tweets. The counter to this is that the person should simply accept that what they post online is going to stay up there forever, and that they should account for that when they post. I am more inclined to stay wth the latter, that I don't believe anybody should post aything online they aren't prepared to have saved and potentially published through fair use in some way.

Another particularly dangerous factor of this publushing data is the publishing of location data also. Part of the Twitter API pulls out location data from tweets, and the same can be said about many of the other APIs and platforms used in digital methods. Should this data be published aswell? Again much of the same arguments can be made, I don't think it puts anybody in sigificant harm considering the time difference between publishing the tweet and publushing the data.

APIs

API”s are simply an interface for the platform, and as noted by ($) they are not designed for researchers. This issue is one that has some implications for all research, including the work undertaken here. One question that ($) brings up about API's that researchers need to ask themselves when they use API's as a part of their research, are they using them as a part of their methodology and as a tool to study the content within the platform, or are they studying the platform itself. This is a totally unnessecary dichotomy, as ($) mentions the work can begin in one area and move along to the other, research can do well to embrace this tension instead of ignoring it. In the case of this research, this tension is not explicitly explored but is present. It is not possible to separate the impact of the Twitter platform on how people express their feelings and opinions from the constraints that Twitter provides for them. While this isn't entirely an issue with the Twitter Search API, it is certainly an issue with the data that outputs from the API. This data is predetermined by the platform from the point at which users interact with the website. It would be impossible for users to interact with eachother outside of the constraints imposed by the platform, and as such the data itself represents these constraints. This is why it is so critical to question if the object of study is the content or the platform itself, since it is impossible to separate the two from eachother, at least in this case.   
 Bowker($) wonders if raw data is an oxymoron. I think what they mean by this is that it is impossible to have data in a totally raw sense, that everything is influenced by the methodology. This is not only the case for data collected through api's, but a condition true of all kinds of methodologies. In the case of APIs this is particularly salient issue, researchers must address the kind of inherant implications for their data when they use these tools. In the case of this project, there are a number of platform specific influences which come up and influence the data, both while it is being collected and before.

While the API has many fields, this project used only the date and strings of text for the tweets. The first influence that the Platform has on my data is the 140 character limit that Twitter has. This means that all users who wish to express themselves must either do it within the constraints of one string <140 characters, or over the span of many tweets. The impact of this limitation could be seen as twofold; either people are more expressive than they normally would be trying to fit their feelings into a tweet, or there is some 'skew' in the score because that expression is spread over many tweets. Without a much more extensive algorithim to detect this spreading of expression, it is impossible to tell to what extent this is happening in such a large data set. This is partly why it is so important to question if the analysis is trying to look at people's opinions, or people's opinions on the platform. It is clear by this limitation that the project steps away from looking strictly at narrative, and must include the tension between platform and content, already we have moved away from 'raw' data.

A more practical limitation imposed by the API is rate limiting. This is a constraint from Twitter which varies depending on the API used, in the case of the Search API used on this project the rate limit was 180 calls per 15 minute window. Twitter is justified in creating these limitations, this API was designed for developers and not researchers. This API was to be used in conjuntion with the other's for those developing apps to build up the development of Twitter through third party applications. This API is free to use, which makes it great for doing research but to avoid it being abused Twitter needs to implement a rate limit. To gain access to this API, developers need to register a Twitter developer account, and with that account comes tokens for the API. These tokens are used by developers to 'log in' to the API and begin making calls. On this project, the calls made to the API had to be few enough not to be 'rate limited' too often, lest those tokens be blacklisted for abuse of the API. This means lower sample sizes and a less representative sample of the conversation happening online. Not a significant constraint, but one that influences the data as it is collected.

Another API constraint is that the REST APIs (of which the Search function is part of) do not access historical data from anywhere between 1-3 weeks before. This also true for the Streaming API's that Twitter also provides. This constraint has profound impacts for research on Twitter on the whole, because researchers cannot delve into the past in order to collect data. Instead, data needs to be collected during events, or very shortly thereafter. The problem is that it is exceptionally difficult to predict what data to collect (a limitation imposed by rate limits) until its relevance is known and that relevance is not always apparent immediately or soon after the event. For this project, the Election was known to occur on October the 19th, and some conversation was expected to occur in the weeks leading up to the election. This project was fortunately in the right place at the right time, but research is not always this fortuneate.

The collection process itself is also almost entirely blackboxed in terms of what and when it collects. The calls for data collection in this project were made in batches, typically not everyday but every few days. Most calls were made for ~4000 tweets, but the API obscures which 4000 tweets of that day were selected to retriveal. It is unclear if these were biased towards the time, location, language, network, retweet count of the calls being made. The implications of this influence are profound. Researchers set out to take a representative sample, and yet are unable to truly know where the data they are using comes from. There is one noteable way around this last problem, although unfortuantely it was realized too late for data to be collected using this technique. Developers and researches need to register their tokens, and all calls made to the API are returned through the perspective of these tokens user. It is well known that Twitter caters its user experience based on location language, time, networks and so on, and so this single perspective problem is unavoidable using one token. In order to work around this constraint, researchers would need to register many tokens and appear to access the API from a representative sample. This would 'trick' the API into returning different results for the same calls, based on apparent different users. Researchers would likely not need to travel the country however, its possible that they may be able to do the same with using simple VPN software.

Another question researchers need to ask themselves is if they really want corporations to have this much control over their research. Clearly, the digital world of the internet and social networks is an area of social research that should not go ignored, but the level of control which private enterprise has over data means that it is exceptionally difficult to have any claims of impartiality or objectivity in the findings. Everything must be taken through the lens of which platform the data comes from, which API (if any) was used to gather the data, the manipulation of that data before it is analysed, and finally the process of analysis itself. One major issue is that the data which comes through these API's is often formatted by the corporations, either by the platform itself as discussed earlier or to insure faster transfer rates. One can imagine this to a study being conducted with surveys, except the only surveys which made it possible for the research to be undertaken were those already written by another institution or private company. This isn't to suggest that companies nessecarily are involved in any wrong doing when they provide public API's, only that researchers need to question how involved they would like to be with this type of arrangement, and if their chosen methodology truly does what they think it does.

One issue that arises from corporate control in the research process is that of reproducibility. It is common for API's to change from time to time, adding or removing new features. If part of the research process depended on a feature that was removed, it would then be impossible for future research to reproduce these tpyes of findings or methods. Sometimes these changes are documented, other times they are not. These kinds of changes often come not from a desire to inhibit research, but for a company to retain control over the development surrounding their product, or as a response to changing corporate values. While API's are a powerful tool for online research, researchers need to question if they feel comfortable undertaking research using tools which are often obscure in design, platform specific, and under the whim of corporate entities.

Open Source

A major source of inspiration for this work was the communities surrounding open source data and software. I also firmly believe that digital methods who depend on closed software are severely lacking in their capability. In the following section I will describe some of the benefits to using open source software in research, but I will also try to expand on these ideas to relate them to research itself. The vision is to do for research what open source does for software.

Since digital methods are computational in nature, the computer used in some analysis is often an important aspect of the process. There is the tendency to think that there might be some difference based on the operating system that gets used. I'm not entirely convinced that there remains much of a difference here, at least between Linux, Windows, or OSX. If there is any difference, it would be in how the hardware is utilized for computations, and in this way it might make a difference in terms of researcher work flow to use something like Linux which is highly customizable and controllable. There is also the benefit in Linux of being able to make some commands on the system level, but again this is probably something that relates more to workflow than it does to research. I'm also not entirely convinced that its realistic for researchers to be modifying their kernels, or auditing their OS source to gain some benefit from a Linux.

There is one instance where having an open source OS would make some difference, and that is purely ideological. Some Linux users feel very strongly that all computation should be something that a user has total access to in their systems. While on an enterprise level this is probably fairly impractical, on a research level this might engage some interesting ideas. In the section on API's, I mentioned there being a risk to users who do not question the effect that closing source might have on a research project. If it is important to researchers to understand and control every single step in their research, then using a closed source OS is simply not an option for them. For example, using an operating system like Windows in some digital research projects would involve the researcher to pass their computations through a layer which obscures exactly what is happening to their data. This is almost certainly a negligible risk, I really don't want to over sell this point, but it is a risk none the same. Practically speaking this is a non-issue, but this is why I say that it is so much more an ideological issue. This argument is also kind of silly to make, because at the hardware level there is almost nothing that is open. This means that literally, practically all computations go through a black box and researchers will never be able to claim total transparency. Again, not a siginificant enough problem to ring the alarm on digital methods, but something that researchers and consumers of research need to be aware of. The control that researchers give up isnt given to entropy or chaos, it is given to corporations. Much of the same issues brought up in the section on APIs can come back here, and again researchers need to ask themselves how comfortable they are giving control away to major corporate entities.

The more interesting side of the open-source debate is in the software side of things. Open source software, and the movements which surround it, are one of the more interesting things happening with computers in the contemporary world. There exists an ideology which argues that it should not be possible to 'own' software. Instead, software should be free to share and alter and share and fork and do whatever you want with because all it is is just some programming information. I am not going to argue for or against this ideology, but I am going to outline how open source software can benefit the research process. In this project, I write all of the scripts and code in R, an open source programming language and computing environment. I also intend to publicly share all of the code I have written. One of the things R is used for is it's data manipulation, analysis and visualization packages. The function of these packages are comparable to those which are found with SPSS or Stata. In the methods section I describe the suite of packages I use in R, and the IDE which I also used for this project called Rstudio. All of the packages and programs I used in this project are also open source, meaning I was able to open the source code and check for any mistakes or errors or other kinds of code that would not work with my project. I did not actually do this, but the point is that I was able to do it, and if I was able to so was anybody else, and since R is quite a commonly used language I am am satisfied that it does not contain something which might be malicious to my computer or my project. In fact, through Github and Arch Linux, I was able to compile a copy of Rstudio from source and install directly on my machine, and since the compiler itself is open source I can be fairly confident that the program I've built this project with is safe. More to the point, I would be able to do this on any computer, and because this software is open I would be able to do it without paying anything to anyone for it. The same things cannot be said for programs like Stata or SPSS. I don't want to go too far and suggest that all research should meet this level of scruinty for its source code, as I states before it is unreasonable and impractical to think that all research needs to be undertaken with strictly open source software. But it does say something about the repeatability and levels of access of this research. It wouldn't make sense to open a project like this up, if anybody wanting to contribute or audit it would not have access to the same materials. This also engages with the point that it costs nothing to use this software. If research is to be engaging on this level for others, it is nessecary that each aspect of the project be done in an environment which is available to all. Had I done the same things in SPSS or Stata for example, those who wish to contribute can expect to pay a fairly large licensing fee for access to the same software, not to mention that both are propriatary software and cannot be audited.

Another big reason that research benefits from using open source software is that it can breed innovation, in part due to the low barriers to entry. R-base for example, is a set of packages which are bundled together in R, these include basic statistical and analytical tools on a comparable level to what one may find in SPSS. The difference is that with R, third party developers can create a function and package it into a distributable format, upload it to a server and make it available for all R users to download and install on their own machines for use. This means that it is not only a modular program, but that there is no limit on what these modules can be. This is impossible on a program like SPSS, which has non-free modules and developers cannot write third party functions for the program. Of course, these third party modules are open-source, and can be audited at any time. The effect of open modules is two fold, one that a researcher can make sure the function does what it is supposed to, but I have also found that one of the best ways to understand a function is to review the source. The open-source nature of R also means that these packages can be expanded and updated at any time, and that the nature of what an R package can grow as far as developers can write, and then can be taken up at any time. 'ggplot2' is the star-child of R third party developer Hadley Wickam, a name I mention in the methods section, but is a great example of this. R already contained basic visualization tools in its R-base, but Wickam created a tool which not only helps him in his own research, but was able to make that available to all researchers using the same tools. While it is without a doubt that many of the same functionalities of R are available for other proprietary softwares, it is often that not one can do all and R might be akin to like a Swiss Army Knife. But that is not to say that R is the only programming environment/language which can boast a large user base and strong development. Python is another great example of many of the same features of open source data software. There are just as many wonderful packages for Python as there are for R, and in fact because both are open source, it's possible to translate many of the packages of one language into another. Ggplot2 for example has both a package for R and for Python, and so much of the same syntax can be used that visualizations written in R that use ggplot2 can be done in Python just as easily.

OpenData is another area of the open-source movement which I find interesting, and has many research implications I would like to discuss. Many times, data is the product of some private enterprise. If say a mining company takes Sonar readings of a certain tract of land using equipment they own, then the data that they collect is by all rights their own to do with whatever they please. If they chose however, they could release that data to the public and it would be open. If they submitted it to a much larger database which includes data about tracts of land from the entire region, and this database was accessible to all it would be said that they have contributed to an open database. This type of open-data movement is present in most forms of science, which benefits enormously from sharing resources like this, and sociology is no exception.

One question that Gurdstein brings up however is if researchers are able to make effective use of this data. A part of this question mirrors the concerns I brought up in the digital methodology section, while the data might be available to all, it might not be accessible to all. Much like digital methods, accessing datasets such as these might not be possible for those without sophisticated knowledge of data management and computer science. While it is often touted as a hallmark of the internet and digital culture which empowers all, Gurdstein argues that it might actually only empower those with the infastructure and skills ot make use of the data. I personally sway towards this argument too, because as noted before, many of the bottlenecks to access are not just at a knowledge level but also at a resource level.

The GEDLT is a great example I will use to illustrate this. From their website, “GDELT monitors print, broadcast, and web news media in over 100 languages from across every country in the world to keep continually updated on breaking developments anywhere on the planet”. The dataset that makes up the GDELT event database is quoted at over 100GB, and even the website states that users will require 'deep technical knowledge' and 'extensive experience' to make use of this data. GDELT.org fortunately makes a visualization and analysis service available, but the code for this is obscured and therefore a little dangerous. The data itself has been argued to be not of great quality as well, often getting representation of events incorrect through overrepresentation and double-counting. As I mentioned before, if I wanted to I wouldn't be able to process the data in the GDELT at all, while I have the memory I simply don't have the hardware nessecary to make many of the computations. I also don't have the required 'deep technical skills' that make this kind of analysis possible. Effectively, the barriers to entry for analysis on this data is so high strictly because it is so big means that it is a different animal that regular open data. This is a really important distinction to make, that while data might be open it doesn't mean that is is able to be used effectively. One must question if this is a supply side problem, or if it is a demand side problem. Whose responsibility is to to make data usable? GDELT does provide a reduced data set, which reduces all data down to a single day metric, but this is effectively a different data set than the actual GDELT. Is this problem a technological one? Can hardware capability, and more sophisticated software can bridge the gap between these massive data sets and casual researchers?

One of the most common questions about open source software, and one that will no doubt be thrown at open source research, is why would anybody contribute? If the whole project relies on contribution from people who seemingly are unrelated to the project, why would it ever survive. The answer to this question for software is likely the same for research, and that it is because people are much more connected to these projects than they seem. There is plenty of research which seeks to answer this very question, all seem to come up with siliar answer ($). One of the reasons is that it is a great way to learn. Software coding is a very difficult skill to develop, and one that requires a lot of hands on experience. Sometimes, those who are just starting out benefit most from being able to see the full code of working software to learn how it is done, much like how just watching somebody cook a meal teaches somebody a lot about cooking. Once a beginner starts to make contributions to the project, they learn from mistakes they make and how to effectively implement their ideas into software. If they are good enough, they start to build a reputation. This reputation is another reason that researchers have found people make contributions to open source projects. Often these contributors use their contributions to display their skills to potential employers, or simply to build up credit in the softwre development industtry in general.

A really important reason that people contributions to open source projects is because it is a piece of software that they use often enough that working on it benefits them directly, and the project indirectly. This is a case that I described earlier when arguing for the benefits of R and third party developers who create packages for their own sake, and simply commit changes altruistically. This is without a doubt probably one of the more salient reasons that people contribute, at least from the perspective of many active projects with a large user base. These contributions are not just made by solo coders either, large companies also make massive contributions. This is less obvious with programs like R, but with Linux this is one of the more common ways in which things get developed. Red Hat Enterprise Linux (RHEL hereafter) is a great example of this. RHEL is a distribution of Linux which specifically targets industry instead of personal uses, and is used quite widely. The RHEL website claims that many corporations, universities, and non-profits have relied on RHEL for any number of computational uses. These institutions use RHEL for their day to day operations, and often times make contributions to the distribution for their own benefit, which is also contributing to the Linux Kernel itself, a shared resource among all Linux distributions. These types of contributions don't just help these institutions who use RHEL, but all who use the Linux Kernal. What this means for the larger world is that developers using Linux benefit from those who contribute, and these developers are very wide spread. Google's Android operating system is based on the Linux kernel for example, which powers a great number of people's smart phones who all benefit from controbutions to open source.

Another reason people do these types of contributions is just for fun. Sometime the project themselves are not made out of some serious nature or use, and sometimes the people who many contributions are doing it strictly as a hobby. This is the same thing that powers much of digital culture, that being a part of a community and a shared project is often a rewarding experience for those in it beyond any type of reward that they might get from being a part of some larger project. It isn't really the case that this is what is going on for all of the open source projects that exist in the world, but it is certainly a factor for some. Some coders approach it as just a hobby, and an excerise in their skills.

One question that often comes up from those wishing to harness to power of open source software on their own project is how to facilitate this. There is no secret sauce to facilitating contribution that is unique to open source software, it is just the same as what one may find in any other form of project. It might be helpful to think of open source software development as a kind of conversation between developers, and asking how to facilitate contributions is like asking how to faciliate conversation. I personally think that the biggest factor that can lead to contributions is the level of interest that a project generates. The Linux kernal for example is used by so many users that there is an undeniable interest in the Linux kernal, and so there is a great number of contributions to it on all fronts. Some of these come from building up the weakest points in the Kernal, some of them are contrbitions which allow spport for newer hardware, some are just to fix mistakes and bugs in previous code. It is this constant use of Linux however which generates so much interest, and so the faciliation of contribution to open source comes from this.

One way that projects can contribute facilitation is by becoming a part of a network of contributors, on a platform which makes it easy for them to make contributions. GitHub is a perfect example of how platforms can facilitate contributions to projects. Github is a collaboration platform, created to make the use of Git easier. Git being the source code management tool created by Linus Tarvolds, creator of Linux. By using Github to manage their code, developers are sending an invitaation to other developers to take a look at their code and to use it as they see fit, or sometimes contribute to it. If they want, these code files can be copied to their own computers filesystem, and when alterations are made these changes can be comitted to the main branch of code. Many IDE's, including R studio, have Git extensions, allowing for easy sharing and collaboration on code projects. It is not nessecary for an open source project to be on github, a really big part of facilitating contribution is lowering barreirs to contribution, and being in an easily discoverable place.

One issue that comes up with this is the Free Rider problem. This project itself is a fine example of this. I have benefitted from the contributions of others to open source projects, while making no contributions of my own. This isn't just a problem of software, it is a problem in a lot of digital products and culture. Another good example of this besides Linux is Wikipedia, where many go as their first source for knowledge and yet not most do not make contributions. Some might be quick to say that Wikipedia could be so much better if everybody contributed something, but this isn't necessary. Running Wikipedia is not free, it costs money to run the website and serve content, and so Wikipedia makes up for this as a non-profit and asks for donations during some campaigns during the year. This is the area where the Free-Rider problem really comes into account for me. As Wikipedia grows and grows, how can they sustain themselves without the financial contributions of the people who use it. Free rider problems don't just exist online, they are everywhere in the world. I feel that the best solution is the same as the reason for faciliating contributions, that if there is enough interest the project will likely be sustained on contributors. I believe that because Wikipedia is used by so many, it is this interest which will likely sustain it.

Finally, a question that comes to mind when thinking about what all of this might mean for research. Who actually owns and controls the research? This is a really important question for research, which may have findings that could change the world, but it is also important for software which can impact millions of users. It is important that some level of control be kept over the projects, but luckily being open-source doesn't mean that projects are totally anarchic. With Wikipedia, absolutely anybody can edit any article, but that doesn't mean that these changes will be always accepted by administrators. Often times, these edits are simply reversed and things go back to how it was before a malicious or errenous contributon was made. For research, one has to question if anybody wants to truly 'own' research. I know that some researchers treat the projects like their children, but at the cost of the expansion or colloboration on that project I'm not totally convinced that it is worth it to retain total control simply for the sake of it. The potential that comes from transparency, and opening of digital methods to all who are interested not only helps communicate the findings and methods of research, but also helps to enable those who are interested to contribute or fix mistakes. This is so in line with traditions of the scientific method, where peer-review, repeatability, and transparency are so critical. I'm not arguing that all research pursuits consider open source methods, of course this is unthinkable in some circumstances, only that for those who practice digital methods this may be one thing to think about.

Methods

The research process is done in an open-source programming language called 'R', and most of the code is written in RStudio. This project is indebted to the hard work of R programmers who designed the development environment that this project is designed in, and also those who developed the packages that the analysis uses as tools. R, while relatively new at this time, proves to be one of the most useful tools available to the contemporary social science researcher – especially those employing digital methods.

One reason that R is so useful is it's being open-source on all levels. This has effects across the entire research process. For example, analysis in R often involves imputing data into functions and getting some output from it. In some closed-sourced programs, the actual processing of the data that occurs by these functions is masked to the user. In R the user can inspect the source code of that function, and determine exactly what is happening to their data as it passes through said function. The benefits in this case are many. The user has a very clear idea of what is happening with their research, but it also empowers the user to have total control over that function. Another benefit to using open source languages is the ability to freely share packages between users, users who often create packages for their own use but decide to share them gratis to anybody else who may need them. This benefit brushes against the benefits of using open source software in general, an interesting topic but one that is well beyond the scope of this work.

The language itself when used in a good IDE might seem challenging to some users who primarily use graphic analysis tools, but once the initial hump of discomfort is overcome R is a monstrously powerful tool for researchers. The most important reason that R is so powerful is because the baseline packages can be augmented with anything developers can come up with, meaning that most research methods can be done in R, and often times it is much easier. The following methods that are used in this analysis could have been done with a suite of research tools, but because R has such a strong base and a very active community of developers the project never needed to move away from it.

Data Collection

The first stage of research was to gather 'data'. The term data will be used hereafter, but must be qualified first to clarify that it does not represent numerical data in the strictest sense at this point. The form of the data at this stage was a 'tweet' and the corresponding meta-data from Twitter. This data had 16 different variables when collected from Twitter, the most important of which are: who wrote the tweet, the text of the tweet, if the tweet was favourited, screen name who it was a reply to if it was a reply, the date and time it was created down to the second, the device used to publish the tweet, if it was a re-tweet, who originally tweeted it if it was a re-tweet, how many times it had been re-tweeted, and the latitude and longitude of where that tweet was published. This level of data collection was made possible by the public facing API (application programming interface) that Twitter provides for researchers and developers. This API allows a user to develop an application which uses the data directly from Twitter, the level of which depends on which API the user has access to. Due to the limited resources on this project, the public facing Search API was used – although there are far more powerful API's available to developers and researchers willing to pay a fee. The problem with this level of access is that a users access to Twitter data is limited in time going back a week and 180 calls on data per 15 minute window, a limitation that is explored later in this work.  
 This research project did not involve access to this API directly however, instead this project depends on a package for R named “Twitter” (Gentry, 2015). This package was created for this accessing the Twitter Search API in R. It should be made clear that this package does not have any tools for analysis, it is just a tool made to collect tweets. As mentioned before, the API was limited in the number of calls it can make in a window, and this 180 calls per 15 minutes might sound like plenty, but each command in the data collection stage takes more than 1 call, meaning the early iterations of this project brushed up against this rate limit often. One method to solve this problem is to write all the commands into a script, and run that whole script in R, enabling the program to keep making commands if it was rate limited until the program entered a new window.

An example of part of this script:

*write.csv(twListToDF(searchTwitter("Harper",n=2000, since="2015-09-28", until="2015-09-29")),file="Harper0928.csv")*

Starting from the inside brackets and moving out there are 4 arguments: the search term, the number of tweets to collect, the dates of the earliest tweets to pull, and the dates from the late tweets to pull. This innermost bracket is what gets passed of first function of the Twitter package, the 'searchTwitter' function which searches Twitter based on user specified information. That 'searchTwitter' function is argument for the second function, the 'twListToDF' function which transforms the output of the 'searchTwitter' function into a data-frame, which is the main way that data is handled in R. This data-frame transforms what is essentially a series of vectors into the familiar column and row format that many are familiar with. Finally, that data-frame is then passed through the 'write.csv' function from the base packages in R, which saves the data frame as a. .csv file with the name specified in the second argument of the function.

This same type of function was used for each search term, on each day of the data collection period. There were 14 search terms in total: Harper, Trudeau, Mulcair, Elizabeth May, Conservative, Liberal, NDP, Green Party, #cdnpoli, #elxn42, #elxn2015, Canadian Politics, Canada, #oct19. Some search terms were added partway through the data collection period, as a response to the trending topics on Twitter regarding the election. The earliest data is from September 28th, and the latest data is from November 1st. The average N for each of these functions was roughly 4000 tweets to be collected for that search term on that day, though that is sometimes is high as 6000 and as low as 100. Since there are 14 terms it often came to be that each day there was anything from 50-60 thousand tweets being collected each day over 30 days, the total number of tweets analyzed in this project comes in at a little over 1.5 million. As mentioned earlier there is a limit on how far back the Twitter API lets a user collect data from, for this reason the event that was observed has to be something occurring near the time of data collection. Using this API it is impossible to go back and research 'historical' tweets, a topic for later on in this project.

It is important to discuss the decision process involved with defining the search parameters. The two arguments of the searchTwitter function that define what gets collected that enable the researcher any real control over were the arguments for the search term, and for the number of tweets to be collected. This number could be much higher than 4000 tweets, but higher than 4000 meant being rate limited far more often. This means that the data collection process would be very slow process, if the data got collected at all without being the program crashing. This problem is also coupled with the potential consequence of Twitter revoking access to the API or the application being blacklisted when the rate limit gets abused. For these reasons, a reasonable N is between 4000-5000 tweets per search term per day. That N does not guarantee that there is always going to be that many tweets on that day for that search term, in many cases the actual number of tweets fulfilling the date argument fell short of the N argument, resulting in data-frames that were short in cases relative to other days. For this reason, not every day can be directly analyzed and needed to be normalized in the processing portion of the project.   
 Since the subject in question was the 2015 Canadian Federal Election, it is wise to include as many search terms related to that as resources allow for. The major parties: Liberal, Conservative, NDP, and Green Party. The leaders of those parties: Harper, Trudeau, Mulcair, and Elizabeth May. These search terms had to be short enough to not exclude important tweets, but specific enough to include only those which pertained to the election. “Liberal” for example would search for any tweet including the word “liberal”, in upper or lower case. 'Liberal' was chosen rather than “liberal party” because that would exclude tweets that might have left out the word 'party'. The balance is trying to avoid being too general, this search term would all pull tweets out that have the word “liberalized” which might have nothing to do with the election for example. The same logic has to be applied, but in the case of the Green Party and Elizabeth May it was not possible to just have the last name of the leader and 'Green' since those would be too vague of a search term. For this reason, it might be said that there is a fundamental difference in Twitter data starting at this point, but the line has to be drawn someplace. This practice will be discussed in later sections of the project.

The dates that selected were based on the 2015 Election campaign, and the date of the election. The data actually begins after the start of the campaign, but the start does serve as a good point to lead towards the election date. This is a project on the narrative of the election, and so the data does not necessarily need to be begin on the same day as the start of this historically long campaign. It would have been more beneficial to start on that day, but due to project limitations this was not possible. It is also because of those limitations that it is not possible to confidently say much about the start of the campaign narrative, there is only confidence in the early days and the wind up period before the election. The data also extends beyond the date of the election, not strictly at the end of the campaign. This is to try to capture the ending sentiments and the wind down period after the election. The range of time that this data has effectively captures the narrative of the 2015 election, but of course this is one area that could be seen as a limitation on the end results.

Once all the Twitter data was collected, it is then concatenated each. .csv file which represented an individual search term on a single day into one main .csv file for each search term spanning the whole data collection process. This was one of the only processes not done inside R, but rather using a simple Linux command. The end result of this data collection stage was 14 different main files each containing over 100 thousand tweets worth of Twitter data, spanning over the course of the month of October.

Data Processing

There are a number of tutorials[[4]](#footnote-5) using R to analyze Twitter data based on sentiment, this so called sentiment analysis was the basis for this analysis. Not everything from all of these sources is used on this project, but part of the research process was discovering uses of R for this and these tutorials are indispensable.

One function that used from another source is Jeffery Breen's sentiment scoring function as seen here:

*score.sentiment <- function(sentences, pos.words, neg.words,.progress='none')*

*{*

*require(plyr)*

*require(stringr)*

*scores <- laply(sentences, function(sentence, pos.words, neg.words){*

*sentence <- gsub('[[:punct:]]', "", sentence)*

*sentence <- gsub('[[:cntrl:]]', "", sentence)*

*sentence <- gsub('\\d+', "", sentence)*

*sentence <- tolower(sentence)*

*word.list <- str\_split(sentence, '\\s+')*

*words <- unlist(word.list)*

*pos.matches <- match(words, pos.words)*

*neg.matches <- match(words, neg.words)*

*pos.matches <- !is.na(pos.matches)*

*neg.matches <- !is.na(neg.matches)*

*score <- sum(pos.matches) - sum(neg.matches)*

*return(score)*

*}, pos.words, neg.words, .progress=.progress)*

*scores.df <- data.frame(score=scores, text=sentences)*

*return(scores.df)*

*}*

There are 3 main parts to this function. The first section dealing with formatting the text so as to remove punctuation and change everything to lower case. The next part is splitting each tweet apart into word while still keeping them inside the same tweet wrapper. The last part is comparing each word to a list of words called a lexicon, in this case 'pos.words'. The score that each tweet receives is the sum of the positive matches minus the sum of the positive matches. This is a fairly simple algorithm, whose function is to just describe how many words from a list of words are found in any given tweet. This sentiment score represents the first dimension that these tweets are analyzed on.

The list of words for sentiment is actually called the 'Subjectivity Lexicon' from Theresa Wilson , Jaynce Weibe, and Paul Hoffman all from the University of Pittsburgh who used it as a part of the paper describing this type of phrase level analysis described earlier. The actual word list itself comes off of Jeffery Breen's github, the same lexicon he uses for his sentiment analysis. The process that Wilson,Wiebe and Hoffman describe is much more involved that a Bayesian matching algorithm, but due to a lack of resource it could not be implemented in this research.

Sentiment is not the only metric that these tweets are scored on, there are many other dimensions of analysis that can be explored. These other dimensions include: economics, healthcare, climate, and foreign policy. Each dimension required its own lexicon, and without having the benefit of existing lexicons it is a requirement to write them. The lexicons are more or less just words pertaining to that topic, and the creation of these lexicons was little more than modifying existing glossaries and indexes of these words to fit the algorithm. This is another area where it is important to find a proper balance and having to draw an arbitrary line. Like the search term definitions, the researcher runs the risk of being either too inclusive or too exclusive in this definition of what might be an 'economic' word. Being too inclusive means a type I error is more common and being too exclusive means its a higher chance of a type II error. There is no perfect lexicon, and it is a relatively small problem since all tweets regardless of search term are passed through the same lexicon, and so any bias present in one will be present in the other. This means on the whole, things might seem more 'economic' than another for example, but there will be no bias between search terms.

There is also the problem of some words having the potential to belong to many different lexicons and being double counted. Rather than picking and choosing which words should go where and imbuing the analysis with my own bias, this analysis includes such words in both lexicons and accepts the risk of double counting a few tweets. This is a problem that is wholly unavoidable, and while it does present a limitation on this work, that limitation is the product of the methodology itself. To do this type of analysis at all means that there will be some new types of unavoidable limitations, and so rather than scrap the whole work it should just simply be noted that there is these types of issues.

The sentiment scoring function proved useful not just for analyzing sentiment, but also for scoring the tweets on all the above dimensions with some slight modifications. The measurement of sentiment can either be positive or negative, which is reflected in the score function, but how focused on 'healthcare' a tweet is cannot be negatively focused, and so that score must only be positive. As such the scoring function is just equal to how many matches each tweet has on the lexicon of whichever dimension is being tested for.

The parsing code is available for review outside of this section as it is too long to document here, but is to be described in short detail. The first part of the code imports the data set to be analyzed, and does some preliminary formatting as found in this tutorial[[5]](#footnote-6). One major function of the preliminary formatting is that it deletes the duplicate tweets, but not the re-tweets which should serve to reinforce the narrative present in the tweets. This is a conscious decision to leave re-tweets in the analysis as they are part of the narrative as well, and while they do not necessarily represent separate ideas from the original tweet they are none the less important.

Next is the definition of the scoring function for each of the election themes, and the loading of each themes corresponding lexicon. After all has been defined, the data is passed through each scoring function, each tweet is given a score for each dimension as it passes through each function and those scores are tallied up into one final data-frame. This data frame has the columns for each possible permutation of the dimension scores and the date, and a number of tweets that fit those parameters. This number of tweets varies from day to day, sometimes due to the specified N and sometimes due to the variance of tweets for that day, and so at this stage these numbers were normalized to represent a proportion of all tweets for that day. That data-frame with aggregate data is saved, one for each of the search terms tested for. It is this aggregate data that goes on to be visualized and then analyzed in the next stage.

Data Visualization

The visualization stage of the project is the stage in which the most creative liberties can be taken. Before this stage, the data is still in a fairly cumbersome format, while it is not as hard to manage as 1.5 million tweets it is still a lot of numbers to keep track of. The visualization stage is an attempt to format this aggregate data into something more manageable, to understand the story that the data tells us in a single glance. One of R's biggest strengths that it provides to users, and an area with some of the most active development, is in visualization.

Part of digital research is going to have to be the publication and communication of the work being done. This means that there has to be tools available to researchers to accomplish this, and there is no better tool for sharing information than the internet. There are tools available across the board for self-publishing figures and interactive charts, but none as integrated directly into the work as Shiny is to R. Shiny is a platform for making digital interactive web-apps, and is available for the most commonly used R environment, RStudio. What Shiny allows for is the free publication of a researcher's data, directly from the researcher and directly from the workspace that it was developed in. This means that for very large projects, the early stages can be published and researchers can provide daily updates with a few clicks.

This section on visualization depends heavily on the work of Hadley Wickham, an R developer who has made much of the power of R accessible to users. The package that Wickham developed that was most useful for this project was his 'ggplot2' package, a data visualization tool for R. Like Shiny, there are many tools available to researchers which enable similar tasks to be accomplished, but none are as fast or as integrated as the packages are to R. The visualization technique that I used for this project is a simply bubble chart, wherein the size of the bubble corresponds to the proportion of tweets that day fitting the co-ordinates of the plot. The Y-axis in these plots was always the sentiment score, and the X-axis was user-specified using a pull-down menu which selected any of the other election dimensions. All of these plots rendered in 'ggplot2' were then passed through another package called 'plotly', which enables interactivity such as zoom and pan with the graphs.

The final product of this visualization is a website in which a user can visit and explore the data for themselves. While this project has its own conclusions, that does not mean that curious others can't visit the website and explore the Twitter data for themselves. It is this that represents a fundamental shift between traditional methodology and digital methodology, wherein the entire process can be built upon by others (provided the tools stay available). This entire project is intended to be public, and so at any stage in the future anybody with the desire can audit the project to find and fix flaws, or to build on it themselves, or to simply explore it for themselves and make their own conclusions.

1. <http://www.anyonebutharper.ca/>,<http://www.votetogether.ca/>,<http://anythingbutconservative.ca/>,<http://atlantic.ctvnews.ca/angry-veterans-launch-anything-but-conservative-campaign-1.2522054>, http://news.nationalpost.com/news/canada/canadian-politics/veterans-group-starts-anyone-but-conservative-campaign-plans-to-picket-pro-harper-events-in-uniform [↑](#footnote-ref-2)
2. http://www.cop21paris.org/ [↑](#footnote-ref-3)
3. http://stopharper.ca/ [↑](#footnote-ref-4)
4. [https://sites.google.com/site/miningTwitter/questions/sentiment/sentiment](https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment), [http://www.r-bloggers.com/Twitter-sentiment-analysis-with-r/](http://www.r-bloggers.com/twitter-sentiment-analysis-with-r/), [http://thinktostart.com/create-a-wordcloud-with-your-Twitter-data/](http://thinktostart.com/create-a-wordcloud-with-your-twitter-data/), https://jeffreybreen.wordpress.com/2011/07/04/Twitter-text-mining-r-slides/ [↑](#footnote-ref-5)
5. [http://analyzecore.com/2014/04/28/Twitter-sentiment-analysis/](http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/) - This tutorial is one of the more fundamental works for the parsing segment of this project, and so this along with the prior tutorials and Jeffery Breen deserve credit for their influence on this project. Much of the code I used is just slightly modified versions of what I found there and so these authors/developers/educators deserve the credit for the parsing section of the analysis. None of the work used is proprietary, but it is still important to give credit where it is due. [↑](#footnote-ref-6)