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.