Effective and ethical use of strategic data analysis for the purposes of election engineering in India.

Karthik Premachandran Geetha

Dissertation written under the supervision of Dr. Jakob Stollberger

Dissertation submitted in partial fulfilment of requirements for the MSc in Strategy & Entrepreneurship, at the Universidade Católica Portuguesa, 06/06/2019.

Acknowledgement

I would like to thank my supervisor Dr. Jakob Stollberger for his guidance and support throughout the completion of my thesis. Also, I would like to thank my cohort and all the lecturers who helped me and supported me through this past year at Catolica. Finally, I would like to thank my family and friends who have encouraged and motivated me throughout my post-graduate studies.

Word Count – 11,166

Table of Content

Effective and ethical use of strategic data analysis for the purposes of election engineering in India.

By
Karthik Premachandran Geetha

Abstract

There has been a rise in the use of data analytics in the election campaigning process, the potential of which can be seen with the 2012 US presidential election. This dissertation aims to do an examination into the effective and ethical use of data analytical models for the purpose of election engineering in an Indian context. It appraises the current analytical models used in India through the collection of primary data such as semi-structured interviews which were conducted with people working with the different political organizations in India. In total, 6 semi-structured interviews were carried out among representatives from 3 political parties, at both the state level and the national level. The study revealed the usage of data analysis by the national level parties with different strategic approaches towards it, whereas, the state party focused more on traditional strategic methods of data analytics. The recommendation for improving the effectiveness of current data analytical models in India, is through providing a structured method to be followed which could be tailored for specific political party requirements. The ethical recommendation involves monitoring and controlling data analytics from breaching the data privacy of individuals through the use of governmental institutions and regulatory bodies.

Keywords : Data analytics, Election engineering, India election campaign, Political campaign.

Uso eficaz e ético da análise de dados estratégicos para fins de engenharia eleitoral na Índia.

Por
Karthik Premachandran Geetha

Abstrato

Houve um aumento no uso de análise de dados no processo de campanha eleitoral, cujo potencial pode ser visto nas eleições presidenciais de 2012 nos EUA. Esta dissertação tem como objetivo fazer um exame sobre o uso efetivo e ético de modelos analíticos de dados para fins de engenharia eleitoral em um contexto indiano. Ele avalia os modelos analíticos atuais usados na Índia através da coleta de dados primários, como entrevistas semi-estruturadas que foram realizadas com pessoas que trabalham com as diferentes organizações políticas na Índia. No total, foram realizadas 6 entrevistas semi-estruturadas entre representantes de 3 partidos políticos, em nível estadual e nacional. O estudo revelou o uso da análise de dados pelas partes de nível nacional com diferentes abordagens estratégicas, enquanto o partido do estado se concentrou mais em métodos estratégicos tradicionais de análise de dados. A recomendação para melhorar a eficácia dos atuais modelos analíticos de dados na Índia é através do fornecimento de um método estruturado a ser seguido, que pode ser adaptado para requisitos específicos de partidos políticos. A recomendação ética envolve monitorar e controlar a análise de dados por violar a privacidade de dados de indivíduos através do uso de instituições governamentais e órgãos reguladores.

Palavras-chave: Análise de dados, Engenharia eleitoral, campanha eleitoral na Índia, Campanha política.

1. Title

Data "of the people, by the people, for the people" – Effective and ethical use of strategic data analysis for the purposes of election engineering in India.

2. Introduction

2.1 Aims

The aim of this project is to do an examination of the effectiveness and ethical implications with respect to the usage of data obtained from the Indian electorate by political organizations in India. This project will also include an appraisal on the current data-driven operational models used by both political and other consultancy organizations. As a part of this project, some key recommendations will be made that improve the effectiveness of the data collection, analysis, and use of such data by political organizations and their ethical implications of it.

2.2 Context

With the emergence and advancement in digital and electronic operations, there has been an explosion in the amount of availability of both personal and private data, at organizational as well as individual levels (Tufekci, 2014; U.S. Federal Trade Commission, 2014). Incidents have been recorded in the use of such vast amounts of data "to make customer-cultivating strategy work" (Rust, Moorman and Bhalla, 2010) by many firms for targeted marketing. Recently this data has also been seen to be used for politically driven election campaigns, given more prominence through the 2012 U.S. presidential election (Shen, 2013). The successful implementation of the 2012 U.S. presidential election opened the doors to a new form of data-driven electoral campaigning, for which the operational and ethical boundaries have not yet been fully drawn out. Due to the presence of ethical and operational ambiguity surrounding this, it has led to situations of illegal infringement of personal rights in many countries. In countries like Germany, political organizations "are eagerly interested in the implementation of individual-based campaigning techniques without thinking about implications for personal privacy, civil liberties and democratic values" (Kruschinski & Haller, 2017).

In the absence of privacy rights and lack of policy framework in the field of data analysis in election campaigning many consultancies have been seen to pay less attention to such rights. The Information Commissioner's Office (ICO) spokesperson in the UK said they "are conducting a wide assessment of the data-protection risks arising from the use of data analytics, including for political purposes, and will be contacting a range of organizations," (Doward, Cadwalladr and Gibbs, 2017). Even after such measures taken by the ICO, the recent Cambridge Analytica scandal has revealed how such consultancies are taking advantage of these lack of privacy right and policy framework. In the past, academic research and study of such election engineering campaigns were predominately focused and limited to the western culture and developed nations such as the United States, United Kingdom and other European Nations. In light of the probability of such infringement on personal privacy, an examination into the operational aspect such as data collection, data analysis and data storage of such data-driven campaigns in an Indian context would be of high academic importance. Also, due to the fact that India has the most diverse electoral population in the world, in term of language, caste, creed, colour. Considering the diversity facts, the accurate sampling of the desired data is always a challenge, as a minor slippage on data selection can result in defeating the whole purpose of the data-driven election campaigns. As per the sources, the use of such data-driven campaign became prominent in India from the 2014 general parliamentary elections onwards.

2.3 Objectives

The objective of this project is to:

- Examine the effectiveness of the use of data analytics in political organizations.
- Discuss the ethical implications of the current use of data analytics within political organizations with appropriate use of case studies.
- Appraise the current data analytical models used by political organizations.
- Recommend how the effectiveness of data analytics can be improved for political organizations in an ethical manner.

3. Literature Review

The literature review describes how data analytics has evolved over time. The review starts with data collection by political organizations followed by a discussion on the analysis of the collected data using predictive models run through algorithmic programs. Thereafter, the literature review considers the ethical breaches and factors that data analytics has on the privacy laws and the response by many governmental institutions to such usage of voters' data and information. It will conclude with the strategies adopted and the ethical implication of these analyses in an Indian context.

3.1 Data collection in political organization

David W. Nickerson and Todd Rogers argue that "the all-encompassing goal of political campaigns is to maximize the probability of victory" (Nickerson and Rogers, 2013). The role of data in election campaigns has increased over the recent years. Kreiss claims that, for a political party or campaign, figuring out what makes the millions of voter's tick has always been a monumental task and that the use of different data help's us "to define what twenty-first-century citizenship looks like" (Kreiss, 2016). Tufekci sheds some light into Kreiss' claim of understanding the voters by describing computational politics, explained as, "applying computational methods to large datasets derived from online and offline data sources for conducting outreach, persuasion and mobilization in the service of electing, furthering or opposing a candidate, a policy or legislation" (Tufekci 2014, p. 2). Nielsen described the use of the database by the political organizations as more of "personalized political communication," where activists directly contact targeted voters (Nielsen 2012. p. 7). These arguments portrayed by all the authors show that a technology-intensive campaigning has reoriented parties and campaigns with the use of data for political campaigns.

Data-driven political campaigns have been growing over the last 10 years, the prominent emergence of which was seen in the 2012 U.S presidential election (Shen, 2013). Sathiaraj, Cassidy and Rohli (2017) attributed two factors to the rise in the use of such data analytics in political campaigns, "first, there has been a substantial decrease in the cost of computing power and maintaining databases. The second factor is a recent increase in political consultants with backgrounds in fields such as computer science, data science, and statistics" (Sathiaraj, Cassidy and Rohli, 2017, p.4). David W. Nickerson and Todd Rogers (2014, p.52-53) described the

availability of "statistical thinking – and the human capital … The supply of quantitatively oriented political operatives and campaign data analysts has increased as predictive analytics has gained footholds in other sectors of the economy" which helped boost the data-driven election campaigns.

Critiques believe that even with the exponential rise in the availability of data, the effectiveness of analytical tools to mine and analyze these vast data have not quite caught up. David and Todd argue that "While the adoption of these new analytic methods has not radically transformed how campaigns operate, the improved efficiency gives data-savvy campaigns a competitive advantage" (Nickerson and Rogers, 2013, p.51).

The process of data-driven election campaign starts with the collection of data. At a minimum, campaigns need accurate contact information on citizens, volunteers, and donors. Campaigns would like to record which citizens would be engaging in which specific part of the campaign such as donating money, volunteering, attending rallies, signing petitions, or expressing support for candidates. All of this retrospective data requires tracking citizens over time, which is difficult because people frequently change residences and contact information (Nickerson, 2006). David W. Nickerson and Todd Rogers explained the hardest part of the tech-savvy campaigns as "campaigns struggled to manage and integrate the various sources of their data". The various sources being the digital data collected from the internet, mainly from social platforms, and the data collected from fieldwork such as focus groups and voter list identification. This meant that, in the past, the digital data collected by campaigns hardly matched with the data collected in the field. The "Narwhal" computer program, developed as part of the 2012 campaign to re-elect President Obama, represented a breakthrough in this respect as it merged all three sources of data, that is, digital, field and financial into one database (Gallagher 2012).

The foundation of voter database for all campaign would be the publicly available database, in the case of US would be maintained by the Secretaries of State and in the case of India the Election commission of India (Eci.nic.in, 2018). In India, the official voter's list contains a unique electoral identifier which consists of wide range of information starting not only the name and the age but also, the gender the home address with house number, fathers name and the name of the booth the voter casts his vote as shown below.

2 KASARAGOD Assembly Constituency

Voter's List 2018 Part Number 002

Kasaragod Taluk

Mogralputhur Panchayat

KASARAGOD District

Ward Number : 3, Mogral Puthur , 671124

1	ZII0608125		2	ZII0606608		3	JWQ1751916	
Name	Monappa Poojari		Name	Girija		Name	Kavitha P	
Father's Name	Manyu Poojari		Husband's Name	Monappa Poojari		Husband's Name	Anand M	
House Number	3/30	(2/77)	House Number	3/30	(2/77)	House Number	3/30	(2/77)
House Name	Seetha Nilayam		House Name	Seetha Nilayam		House Name	Seetha Nilaya	
(M/F)/Age	M /68		(M/F)/Age	F /55		(M/F)/Age	F /35	
4	JWQ1232172		5	JWQ1753003		6	ZII0138610	
Name	Udaya P M		Name	Mahesh Kumar M		Name	Jithesh Kumar M	
Father's Name	Monappa Poojari		Father's Name	Monappa Poojari		Father's Name	Monappa Poojari	
House Number	3/30	(2/77)	House Number	3/30	(2/77)	House Number	3/30	(2/159)
House Name	Seetha Nilayam		House Name	Seetha Nilayam		House Name	Seetha Nilaya	
(M/F)/Age	M /34		(M/F)/Age	M /33		(M/F)/Age	M /28	
7	ZII0608349		8	FDT1175348		9	ZII0607408	
Name	Aisha		Name	Hameed		Name	Moosa	
Guardian's Name	Hameed		Mother's Name	Aisha		Father's Name	Abdu Rahman	
House Number	3/30C	(15/271)	House Number	3/30C	(15/271)	House Number	3/31	(2/71)
House Name	Mundakkal House		House Name	Mundakkal House		House Name	Rashid Manzil	
(M/F)/Age	F /51		(M/F)/Age	M /34		(M/F)/Age	M /56	
10	ZII0607432		11	JWQ1788926		12	JWQ1786805	
Name	Jameela		Name	Abdul Kareem		Name	Kairunnisa C H	
Husband's Name	Moosa		Father's Name	Abbassa		Husband's Name	Abdul Kareem P	
House Number	3/31	(2/71)	House Number	3/31	(2/54A)	House Number	3/31	(2/54A)
House Name	Rashid Manzil		House Name	Rashid Manzil		House Name	Rashid Manzil	
(M/F)/Age	F /49		(M/F)/Age	M /42		(M/F)/Age	F /35	

Source: (Chief Election Office, Kerala, 2018)

These provide us with the geographical location including the house number, and other valuable information. The availability of such valuable information can be used for election campaigns as these can help merge relevant census and precinct data to the information on citizens in the voter database (Nickerson and Rogers, 2013). While campaigns can purchase secondary data and information, the vast majority of the useful information the campaigns collect about individuals is provided from individuals directly (Pathak, 2017) for example, the name and details of the volunteers and donors in the past are high-value prospects for volunteer-recruitment and fundraising in the future. In short, David and Todd believe "despite overblown claims about campaigns purchasing individuals data's, very little of this information that is most useful to them is purchased" (Nickerson and Rogers, 2010). But, a recent paper shows that merely using Facebook "likes" is sufficient to model and accurately predict a striking number of personal attributes including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender (Kosinski, et al., 2013) such information have been known to be purchased using third-party agreements as mentioned earlier. However, the sources may vary in an Indian context due to the fact that, Facebook has only 204 million active users which are only 19 % of the Indian population, versus 73% in the U.S (DeFotis, 2017)

3.2 How analysis is done

Many models have come into practice with the use of data analysis in election campaigning around the world. But, Tufekci (2014) in his journal "Engineering the public", talks about modelling programs that allow for acquiring answers about an individual, without directly asking questions to the individual, "thus opening the door to a new wave of techniques reliant on subterfuge and opacity" (Tufekci, 2014). David W. Nickerson and Todd Rogers in their work "Political Campaigns and Big Data" describes the use of "predictive models to make targeting campaign communications more efficient and to support broader campaign strategies" (Nickerson and Rogers, 2013). The 2012 US presidential election had used the same predictive model mentioned by David and Todd with the difference in their analysis interpretation (Sathiaraj, Cassidy and Rohli (2017). These predictive models include three categories of "predictive scores" for each citizen in the voter database (Nickerson and Rogers, 2013):

- behaviour scores
- support scores
- responsiveness scores

Behaviour scores use past behaviour and demographic information to calculate explicit probabilities that citizens will engage in particular forms of political activity (e.g., donate, volunteer or attend a rally for the campaign).

Support scores predict the political preferences of citizens. This is ideally carried out by direct contact with the all the voters, which can be impractical, therefore, the campaign contacts a subset of the voters and use their response to develop a model that could be used for predicting the above-mentioned preference. These support scores typically range from 0 – 100 and generally are interpreted to mean "if you sample 100 citizens with a score of X, X% would prefer the candidate/issue". A support score of "0" means that no one in a sample of 100 citizens would support the candidate/issue, "100" means that everyone in the sample would support the candidate/issue, and "50" means that almost exactly half of the sample would support the candidate/issue. Support scores only predict the preferences at the aggregate-level, not the individual level. Constructing these support scores saves campaigns the time and cost of collecting the political preferences of every citizen in the electorate.

Responsiveness scores predict how citizens will respond to campaign outreach. Gerber and Green explained that the campaign can use a randomized field experiment to calculate the average response of the voters with respect to adopted strategy (Gerber and Green, 2000). Predictive scores in many cases can help identify the voter's behaviour and preference more accurately than direct primary data collected from the voters themselves (Rogers and Aida 2013).

Campaigns are able to predict with greater accuracy which citizens will support their candidates and issues better than which citizens will oppose their candidates or issues (Nickerson and Rogers, 2014). The predictive score campaign can be divided into two types: "The first predicts the behaviour or attitudes of voters (that is, behaviour scores or support scores) …. The second type of score predicts how voters will respond to campaign outreach (that is, responsiveness scores)" (Nickerson and Rogers, 2014 p. 58-59). Similarly, two other types of the same predictive models are explained by Goldstein and Ridout, "the first step, campaigns need to identify people suitable to direct their mobilization efforts at (also known as voter targeting). The second step involves crafting measures that best motivate these people to turn up at the polls, that is, to assure the effectiveness of mobilization" Goldstein and Ridout (2002).

All these predictive models are backed up with a unique statistical human capital and the campaigns rely on simple deterministic rules of statistic. The human capital refers to the people behind the algorithmic designing, typically, a data scientist with expertise in statistics and predictive modelling. For instance, Malchow (2008) promotes a linear probability model as well as tree-like models such as CHAID (Kass, 1980) for political microtargeting. Murray and Scime (2010) suggest decision trees as well. Green and Kern (2012) advocate Bayesian additive trees (BART, Chipman, George and McCulloch, 2010) and Imai & Strauss (2011) propose to use classification trees, which they embed in a decision-theoretic framework for optimal planning of GOTV (Get out the vote) campaigns. Other state-of-the-art approaches that are used include logistic or probit regression. David W. Nickerson and Todd Rogers maintain that most of the analytic techniques employed by campaign data analysts are taught in standard undergraduate econometrics or statistics classes. And that, currently "the vast majority of the predictive scores used by campaigns are created by a campaign data analyst (or a team of them) using simple regression techniques: ordinary least squares for continuous outcomes; logistic regression for

binary outcomes; and, rarely, probit for truncated data like dollars donated or hours volunteered (Nickerson and Rogers, 2014 p. 59).

This statistical implication poses two limitations. First, the practitioners of these new data-driven statistical tools invoke a phrase coined by behavioural economist Dan Ariely (2010), "You are what you measure" that means correlations is highly dependent on the talent of the campaign data analyst employing them. Second, "modelling by hand" approach where using regression techniques in campaign models typically needs to be constructed uniquely for different regions as the issues and area vary with a different region thus they provide few economies of scale (Nickerson and Rogers, 2014). These two problems of dependence on the skill of data analyst and the economies of scale can be addressed using machine learning.

Machine learning in simple terms is an attempt to compress a huge database of possibilities and their outcome into a data structure of reasonable size (Witten et al., 2016). Consequently, data mining algorithms have become popular among campaign analyst, as this is the 1st step towards data analysis and machine learning process, the clustering and classification algorithms such as k-means clustering or k-nearest neighbour classifiers (Gan, Ma, and Wu 2007). Even though they are part of the unsupervised machine learning which to an extent are "less useful for campaign data analysts because campaign planning often requires having individual-specific probabilities for particular outcomes on which to make strategic cost-benefit decisions. For this reason, supervised learning algorithms are typically more appropriate for the task of modelling political data (Nickerson and Rogers, 2013).

Supervised machine learning include some of the tools like regression tree and classification (Breiman et al. 1984) "in a regression tree approach, the algorithm grows a "forest" by drawing a series of samples from existing data; it divides the sample based on where the parameters best discriminate on the outcome of interest; it then looks at how regressions based on those divisions would predict the rest of the sample and iterates to a preferred fit" (Breiman 2001). The researcher chooses the number of "trees"—that is, how many times the data will be divided the most used is the "random forest" algorithm, where the program pulls randomly drawn subset of the variable in each tree to find the best fit rather than going through the whole number of variables. In this algorithm, the program identifies what parameters add the most predictive power when other parameters are unchanged.

Supervised learning has 3 main advantages for campaign analysis (Nickerson and Rogers, 2014 p. 61) such as:

1. The classes of estimators are mainly non-linear in nature for e.g. the relationship between age and turnout of voters where the older cohorts vote at higher rates than younger cohorts but this relationship peaks among group 60–70 years old and then reverses can be accommodated in these algorithms.
2. Less involvement of individual campaign data analysts as the programs are pre-written and thus program runs test identically for every citizen in the voter database.
3. The data-mining algorithms are highly scalable and cost of constructing additional models is lower for these algorithms rather than building newer regression models from scratch with the same subset of the database.

Predictive scores help us improve the efficiency of communicating with the voter in the campaign and help classify voter, which helps to target expenditures for voter mobilization and persuasive communications (Endres and Kelly, 2017). The use of data-generated propensity scores to identify likely voters and potential supporters is commonly known as microtargeting (Hillygus and Shields 2009). Selective targeting of individuals is beneficial in situations where the other party has the most support (Kramer 1966). When properly executed, microtargeted predictive scores allow campaigns to identify supporters and swing voters who might otherwise not be contacted under traditional-based geographic strategy because they might live in areas largely populated by the opposition.

3.3 Ethical implications

Political campaigns have increasingly combined data-driven voter research with personalized political advertising, which is known as online political microtargeting which "may be both a blessing and a curse to democracies" (Zuiderveen Borgesius et al., 2018). It would improve the participation, and lead to more knowledge among voters about certain topics (Bodó, Helberger and Vreese, 2017). But microtargeting also brings risks such as profiling which entails a loss of user privacy, targeting opens the door for selective information exposure, and potential manipulation (Bennett, 2015). Also, political parties could, misleadingly, present itself as a one-issue party to different individuals. And data collection for microtargeting raises privacy concerns" (Zuiderveen Borgesius et al., 2018).

The revelations into how Cambridge Analytica have accessed voter data and its usage have brought new scrutiny to how campaigns target individuals (Mak, 2018). "On March 17, 2018, The New York Times, along with The Guardian and The Observer, reported that Cambridge Analytica and its related company, SCL, pilfered data on 50 million Facebook users and secretly kept it" (Valdez, 2018). Cambridge Analytica has claimed to be targeting voters based on their psychological profile as reported by Issie Lapowsky senior writer in wierd. "psychographic targeting approach supposedly builds on traditional ad targeting metrics like demographics (age, race, income) and behaviour (voting, spending, online habits) by adding a person's psychological profile" (Lapowsky, 2018). In order to generate such models, Cambridge Analytica had employed data scientists and psychologists to draws on personality surveys which they had conducted by telephone, email, and social media since 2013.

The implications of such targeted marketing carry ethical challenges, as many believe "psychological targeting and persuasion might be used to exploit weaknesses in a person's character"(Matz et al., 2017). The use of psychological targeting was recently reported in media, suggesting that one of the 2016 US presidential campaigns had used psychological profiles of millions of US citizens to suppress their votes and keep them away from the ballots on election day (Grassegger and Krogerus, 2018). This calls to question the basic values and meaning of democracy where, the will of the people, in the mentioned scenario, is not the true representation of the people rather an altered one.

In light of such privacy data breaches forced newer policy framework to be formulated around the world of which, the General Data Protection Regulation (GDPR) was passed and implemented in the United Kingdom and the European Union from 25 May 2018(Information Commissioner's Office, 2018). GDPR replaces the previous 1995 data protection directive, this focuses on the rights for people to access the information companies hold about them, obligations for better data management for businesses, and a new regime of fines (Burgess, 2018). GDPR explicitly focus on the below mentioned 12 criteria (Information Commissioner's Office, 2018);

- Individuals' rights
- The right to be informed
- The right of access
- The right to rectification

- The right to erasure
- The right to restrict processing
- The right to data portability
- The right to object Rights related to automated decision making and profiling
- Accountability and governance
- Breach notification
- Transfer of data
- National derogations

GDPR defines two types of data, personal data and sensitive personal data. Personal data, a complex category of information, broadly means a piece of information that can be used to identify a person... Sensitive personal data encompasses genetic data, information about religious and political views, sexual orientation, and more (Galdies, 2018). Similarly, other noticeable changes are Accountability and compliance where the companies are accountable for their handling of people's personal information, Access to our data by which we are given more power to access the information that is collected by the companies about us and finally "one of the biggest, and most talked about, elements of the GDPR has been the ability for regulators to fine businesses that don't comply with it" (Burgess, 2018).

3.4 Indian context: strategy and ethical implications

India being the largest democracy in the world, had its fair share of data-driven electoral campaigning from the last 2014 general parliamentary election. CNBC reported the use of data analysis was carried out by the current ruling party Bhartiya Janata Party (BJP) which swept a major victory over the two-time consecutive winner, the Congress party. It was reported, "Through data analysis, they have helped raise funds, rework advertisements and create detailed models for voter engagement in swing states as well as gender and minority voter clusters to increase the power of their micro-targeted strategy" (Jetley, 2014). Many had called this a data-driven election campaign not very different from that of 2012 U.S. President Obama's (Shen, 2013), albeit somewhat smaller in size, scale and perhaps style" (Jetley, 2014). Amit Sheth, a professor at Wright State University's Knowledge Computing Center in Ohio even went to the extent in calling the then leader of BJP and the current prime minister of India Mr Narendra Modi as "one of the

most tech-savvy politicians in the world and certainly the most active in India" (Pawha Jetley, 2014).

A report by the election commission of India estimates that around 26 million new young voters are added to the voter's age bracket of 18 years. It is assessed that approximately 100 million first time voters will be added by the start of 2019 parliamentary election, who are often better informed, more educated and tech-savvy than the rest of their family, and who can take a stand that goes against the family's established political leanings (Sharma, 2018). India has seen the influence of many major players in the data-driven electoral campaigns, it was reported that Cambridge Analytica had already worked in India in the 2010 Bihar election where it was reported that they "had achieved a landslide victory, with over 90% of total seats targeted by CA being won (Srivas, 2018). The largest democracy in the world coupled with the second largest population in the world supported with the spread of 4G connectivity to rural India and cheaper smartphones mean that the 2019 general parliamentary election is likely to be fought as much on phones as in the streets (Sharma, 2018).

India is not yet a party to any convention on the protection of personal data which is equivalent to the GDPR or the Data Protection Directive. However, India has adopted or is a party to other international declarations and conventions such as the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, which recognize the right to privacy (Talwar Thakore & Associates, 2018). As such, while the next 2019 Indian general parliamentary election is right around the corner, a study into how the top national political parties are approaching the campaign with; the level of dependency on the data-driven electoral system and the extent in which the ethical constraints are considered.

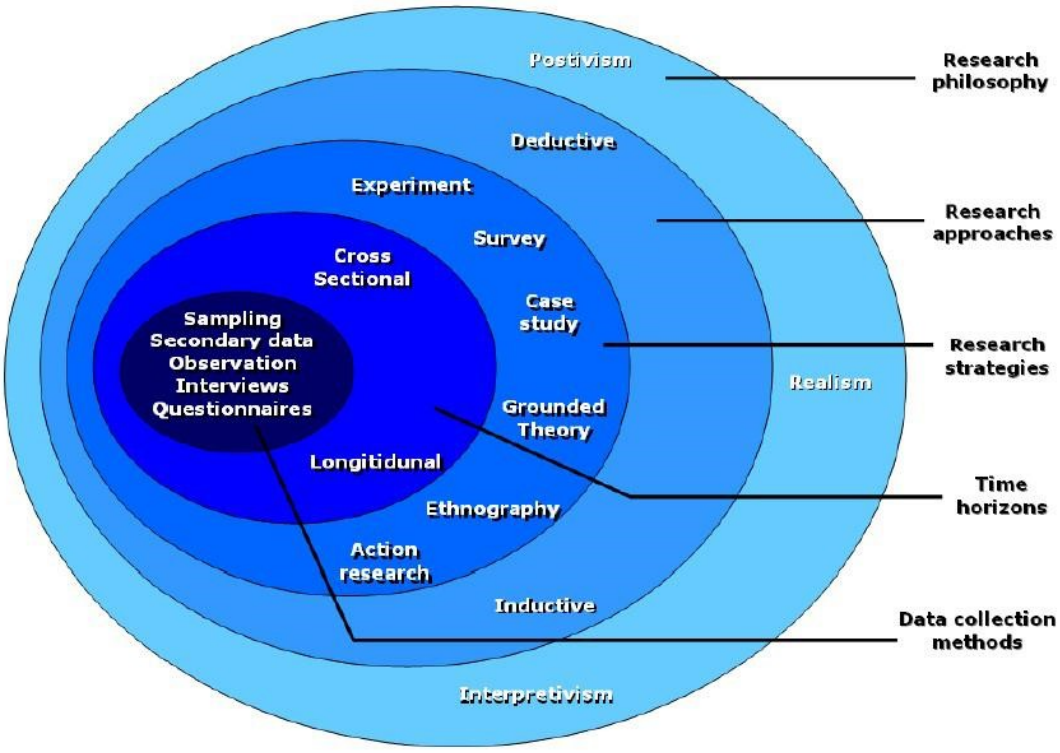
4. Methodology

This study considers in-depth interviews, which will be consisting of people working in both national level political parties and small regional political parties. The people will be remaining anonymous as they have opted out of being on record, abiding by the ethical consent signed by them. These interviews are semi-structured interviews and involved the interviewer working through a list of discussion point which covers these main points;

1. Has there been any primary or secondary data collected for the election campaigning?
2. Do you know the purpose for collecting the electoral data?
3. Please confirm if the data collected is in accordance with the data privacy law?
4. Have you revealed the purpose of data collection to the concerned participants?
5. Do you have any mechanism to protect the collected data?
6. Will the collected data be used in-house or will it be shared with an external party?

This gave both the parties involved in the interview some common ground to build upon (Bryman 2012). The interview provides some valuable primary data but can be criticized for being conducted in an artificial environment (Brinkmann, 2013). Semi-structured interviews are valuable elite interviews, where "elite" are the group of people who have exclusive information not available in the public domain (Dexter, 2006).

The research onion model (Saunders et al., 2007) developed by Saunders is used for explaining the research methodology adopted for this study. An effective progression with more detailed stages in the research process is described as each of the 6 layers are unwrapped and explained (Sahay, 2016) as shown below.



Source: (Institut Numerique, 2012)

4.1 Research Philosophy

This study relates to the philosophy of positivism in the outer or 1st layer of the onion model of philosophical instances. Since, Positivism assumes that the reality exists independently of the thing being studied, in this case, the thing studied being, the data analytic campaign of the common electorate (Newman, 1998). Thus, the study is independent of the reality and the environment of study.

4.2 Research Approach

In the second stage, the study identifies itself to inductive approach and qualitative approach, since observations in the reality mentioned in the research policy of positivism of the 1st stage are the starting point of the research. Furthermore, observation of how data is being collected by political parties and finding a pattern of the data being collected led to the creation of the conjecture. The conjecture examines the manner in which the processes are carried out by these political organizations. Interviews are carried out concerning specific phenomena and then the data may be

examined for patterns between the respondents (Flick, 2011). The qualitative approach was undertaken as the study revolves around what and how are the processes carried out in the political organizations in India. A qualitative approach in this study made sure that the researcher avoided imposing their own perception of the meaning of social phenomena upon the respondent (Banister, Bunn and Burman, 2011) The aim was to investigate how the respondent interprets their own reality (Bryman & Allen, 2011).

4.3 Research Strategy

For the research strategy adopted in the 3rd layer, grounded theory was used. Tim May has described the grounded theory as a qualitative methodology which draws on an inductive approach by identifying the patterns which are derived from the data as a precondition for the study (May, 2011). This means that the results obtained in the research are fundamentally from the research that has been done, rather than the data being examined to establish whether the results fit with pre-existing frameworks we assumed (Flick, 2011). In this study, this research is done based on the semi-structured interviews conducted.

4.4 Choices

The 4th layer of the onion model, the choice, for this study we are using a mono-method by using the qualitative analysis. The mono-method involves using one research approach for the study and in this research the approach taken in just one qualitative method. Since the philosophical choices and the strategies adopted in the beginning are in line with that of qualitative analysis it reinforces this decision. This analysis helps us to explore personal accounts, descriptions and option using the semi-structured interviews.

4.5 Time Horizons

The Time Horizon, the 5th layer of the research onion, is the time framework within which the project is intended for completion (Saunders et al., 2007). Since the data collection is qualitative in nature, the opinions and description of the people will be focusing on the current working process of the political organization. Similarly, the discussion and result obtained are in respect to this current data collected. Thus, in this case, we are using the cross-sectional time horizon as the cross-sectional time horizon is one which is already established, whereby the data must be

collected. This is dubbed the snapshot time collection, where the data is collected at a certain point (Flick, 2011).

4.6 Information Collection and Analysis

Information collection and analysis, the 6th and the final layer of the research onion, the process used at this stage of the research underwrites the overall reliability and validity (Saunders et al., 2007) to the study. There are two types of data that can be collected regardless of the type of approach we use, they are

- Primary data
- Secondary data

We depend mainly on primary data which are the semi-structured interview conducted with people working with the different political organization in India. Uwe Flick (2011) argues that data that is derived from other researches can also be used as primary data in that case the voters list publicly released by the election commission of India will also be considered as primary data in this report. Newman (1998) describes secondary data as that which is derived from the work or opinion of other researches. Secondary data used In this report consist of newspaper analysis and results and the conclusion drawn by other researchers in the field of the data-driven election campaign.

The questioners for the collection of primary data will cover the area such as

- Is data collected
- How data is collected
- How is it analyzed
- Is there any data protection law/policy violation?
- What measures are taken to ensure the safeguarding of the stored data

5. Data Collection & Analysis

As prescribed in the methodology above, the primary data is collected from three political organization, the data collection method for this research is performed through a defined structural interview. The interviewing candidates are selected from two national level parties which include Indian National Congress (INC), Bhartiya Janata Party (BJP) and one state-level party Revolutionary Socialist Party (RSP). One state level and one national level representative have been selected for the qualitative data research from all 3 political organizations, thus in total 6 semi-structured interviews were carried out. The interview was carried out through Skype and telephonic media due to the limited time constraints involved in the research. The naming convention for the selected candidates is in the form of acronyms of political party_1 for national representative and acronyms of political party_2 for state representative. In the current context, the role names are INC_1, BJP_1, RSP_1 for national representatives from the three-political organization Indian national congress (INC), Bhartiya Janata Party (BJP), and Revolutionary Socialist Party (RSP) respectively. Similarly, INC_2, BJP_2, RSP_2 will be the state representatives from the three-political organization Indian national congress (INC), Bhartiya Janata Party (BJP), and Revolutionary Socialist Party (RSP) respectively.

5.1 Interview observations

The transcripts from the interviews and the meeting minutes were then analyzed. Summary of the data was derived from each participant so as to draw the similarities and dissimilarities among the participants in the same party and different parties.

INC_1

A clear and in-depth knowledge into the data analysis processes and strategies adopted by his party was portrayed by the participant. The answers provided by the INC_1 gave a clear picture of the nature of data collected and the methods used to analyze, the below summary describes further details:

- "We had started off with the collection and categorizing of both primary and secondary data after the backlash we had obtained in the last 2014 election. The primary data consists of 1) voters list published by the election commission and 2) direct data collected by the worker in state and local level, about the volunteers and donors during party meeting and demonstrations. Once the data is collected it would be stored in a single database for the

whole of Indian sub-continent and a copy would be retained in the respective state respectively. A two back up storage option is provided one at the central database and one at their respective state database…. The collected data would be run through a specific customized data analysis model designed in-house by our data scientist. The primary segmentation of this data is in a region, which are north India, south India, west India and east India. The segmented data is further merged with the secondary data for attaining data variety. Secondary data for the analysis are mainly obtained from the open source governmental publication such as the economic survey of India and the census report of every state published regularly. Proper ethical guidelines were given to the state level for the direct data collection from the participants. After the clustering of these data, voters are grouped based on categories for effective communication and knowledge transfer between the party and the voters, so that party ideologies can be translated and tailored according to the likes of the public in each category. The implementation of this communication strategy is carried out using social media, telephonic call and other direct means. The main interactive social media platform is Facebook."

INC_2

The interview started off with clear answers to the question provided initially, but as the interview progressed a level of ambiguity regarding the strategic usage of the collected data was seen. They had followed the instruction by the national level authority in collecting the required data but the ethical aspect regarding the consent and clear explanation to what analysis would be carried on these data were vague as can be seen by the summary.

- "We were directed to collect primary data from the participants like 1) volunteer, 2) donors and 3) random survey report from public places. Once the data is collected they are manually entered into a common template which would later be stored in our local database and copy would be sent to the central database. We usually focus less on the ethical consent of the people, but we make sure to explain the reason and make their survey anonymous if directed by the participants, but it's rarely followed when it comes to the data gathering from donors and volunteers…. At the end of the month specifically designed posters and media content would be provided by the national committee with specific instruction to what needs to be published and when. We believe these contents are derived from the survey and data collected from our side."

BJP_1

The interview followed a more relaxed and open-ended discussion with respect to other, as the ideas and the analysis carried out in BJP were more focused towards the current trend and less on the structure of working and collection and analysis of data. Priority was given to social media analysis and electoral campaigning method rather than primary data collecting from their field volunteers. They believe this helps them address the most recent and trending social incident as

they get picked by more voters and user which would improve the attitude and help in marketing the voters toward the party as shown in the summary.

- "We focus on collecting primary data from the social media platform directly by using data mining tools such as application program interface (API) in twitter of both historic tweet and current tweet. We focus less on buying existing twitter database as they are considered outdated and we spend money on training our already established group of data miner and data scientist. Once the data is collected we keep track of these users to study their change in political affinity with their response to the newer situation and find the trending social issue in twitter. Once the programming algorithms identify the trending social causes, we focus our resources on addressing that issue for which we mainly use WhatsApp and then Facebook in that order. Facebook is the implementing social media platform where we address and tackle already identified issues from Twitter. The content for these social issues and other matter implemented in Facebook would be regularly sent down the traditionally available hierarchal chain which is from central, state, regional, district and local committee respectively. Also, we are not just focusing on the social media data collection alone, rather traditional means of strategic implementation like field work, demonstration and other works related responsibility have been given to the state level committee so that they can incorporate their strategy locally"

BJP_2

In terms of data collection and analysis, the responses were very much limited to yes and no as their involvement in these areas were minimal in nature. Their focus lies on field work such as cross-checking the voter's list and targeting people in the list based on location by using traditional means of targeting like the door to door and telephonic intervention. The usage of data analysis is limited to the categorization of different voters using basic k-cluster algorithms based on the local survey which is sectile randomly from different areas in the state. But they feel they are directed by the national level committee to improve their involvement in social media platform without compromising on the local fieldwork already existing as shown below:

- "The primary data collection is based on the voter list published by the election commission of India which we analyse locally. We split the voter's list based on booth level, which is the lowest level of a committee in the hierarchical chain, consisting of roughly 1000 voters in each booth to the local volunteer and party worker. They will then assign a score to the people whom they know based on their interactions with these voters and then conduct a random sample survey on these voters in that booth to figure out the efficiency of the given score. Once, the scores are analyzed with the given score we start targeting the people with direct face to face engagement and telephonic communication to convert the swing votes to confirmed votes. The ethical concerns for us are minimal as the only data collection point we need to worry about is, the survey for which we try to get the ethical consent, but

the people are less worried about it and usually skip the consent and directly be involved in the survey…. We've been instructed by the central committee to increase our involvement in the social media platform like Facebook for which the content and material to be posted in social media are send from the higher-level committee. But, we fell that, these issues are of national prominence and fail to address the local issues."

RSP_1

The interview went through a more insightful talk about the more traditional approach of election campaigning used by them involving the use of primary data from the voter list. The analysis part of the data is limited to software such as Excel and basic coding programs for small categorization of the electoral masses. Once, the small segments of data is ready, the packets of data will be shared with the state level committee. The localized survey would be conducted to identify the need of the people by using focus group study and random informal face to face interview among the public. No program or algorithm is used for the later analysis of the survey data rather, they use the experience of the party leader and volunteer to address the need of the voters directly with door to door campaign and telephonic communication. The limitation found to be was the lack of financial resources available to the party, as currently the party focus only 2 to 3 states, due to which upgradation of their data analytic tools is not feasible. But, the advantage is that they need to focus only on 2 states as they are a state-level party, which result in need of fewer resources as shown below:

- "Our priority is addressing the problem arising in the state politics rather than focusing much in national politics. The election commission data is the primary data in the form of voters list which help to classify different segments like sex, caste, occupation using simple excel and basic program. The program identifies the caste by comparing the name in the voter's list to 100's of already pre-loaded surnames commonly seen in each caste in these states. These compressed and categorized data is provided to the lower level party workers who filter out the report using their knowledge and experience to identify the anonymity in this report. Later a specific focus group is created by the local committees to study the overall need and problems, on an aggregate level which will be summarized for all the voters in that category group. The report later will be submitted to the leader in the state level who decided on what the appropriate responses should be and how to address it. Ethical aspects are limited since the data collection is through a defined focus group where the ethical aspects are clearly driven hence the ethical breaches are highly unlikely. The data collection aspects are transparently communicated to the voters with an assurance that the collected data will not be utilized for any other purpose other than electoral data processing purposes.”

RSP_2

This interview focused more on the collection of primary data from field work such as focus groups study, random casual interviews with voters and record of local volunteers and supporters are categorized and stored in an area-based criterion. An initial dataset for all the voters will be received from the national committee which will be used to determine the focus group criteria. The unique strategic approach they maintain is comparing the focus group needs and their opinion with that of random face to face interviews and interaction in the public domain. If both the opinion of the focus group and interview are in line with each other the result obtained is expected to be accurate as summarized below:

- "The data collected are purely based on fieldwork activates. The base data containing the categorized voter's list are available through the internal committee of the party structure. Based on the revised electoral data we figure out how to sample these data to get an optimum selection of candidates for the focus group study, as this is the primary data we collect on which analysis is carried out. The sample size criteria vary from place to place based on the most dominant factor and the least dominant factor for e.g., in the constituency of Kollam in the state of Kerala criteria commonly used are occupation, area of living, age and caste as the first three are the dominating factor and caste the least dominating factor. Proper ethical consent of the people is taken, and standardized guidelines are explained as the ethical privacy is of high priority in our case. Once the data from the focus group is available we cross-check these by conducting an informal random interview in the public places from where we have selected candidates for the focus group. This is more of an opinion poll survey to calculate the accuracy of the focus group results. Once the results are seen to have optimum accuracy we submit the report before the state leadership, where a further decision regarding the decisions to be taken and factor to be considered by the state leadership to determine the future course of action."

6. Outcome and Recommendation

The research had identified how data analysis has helped add value to the traditional election campaign model to attain a competitive advantage. The similarities in the interview responses among the participant portray that the political organization are utilization such data analysis. The ethical factors for Indian National Congress are taken through a constructed ethical committee which reports to the national president. Bhartiya Janata Party does not have an ethics committee however the ethical factors are managed through the national president. The ethical factors for Revolutionary Socialist Party are managed through a central committee governed through a national secretary. As there is no generally applicable rule in the ethical aspects this would be considered on a case by case manner and the decision would be taken with that specific issue. The outcome and recommendation for the effectiveness of the electoral engineering process in India, based on the conducted interviews, are explained here on its similarities, uniqueness, strength & weakness among different political parties. This helps us identify what each party is trying to achieve with their uniqueness in approach towards election campaigning, their similarities which helps us identify on what common ground do they intersect each other and finally their weakness and strength with respect to their accuracy in the campaigning process

6.1 Similarities

In an Indian context, the electoral data analysis and electoral engineering started in the 2014 general election. All three parties adopted the possible mechanism of electoral data analysis from then on. However, the degree of utilization was different with BJP being the highest utilized, INC in moderate and RSP at low utilization. The first mover advantage was secured by the BJP hence they could attain a competitive advantage whereas in the case of INC is on a medium level as they started a bit late. Even though INC had a late start, the interview showed INC to have a more structured and larger voter coverage with their current data analytical model. RSP in this respect has just scratched the surface of this new field of electoral campaigning which would be a hurdle to overcome in the long run.

All three-political organization has been using the electoral data analysis however the level of utilization varies based on the model which they choose. In the case of both INC and BJP, they had opted for higher level data analysis and usage, as they are a national level-based party where they use data analysis in all the 29 states and 7 union territories for a total of 36 entities in India.

Since RSP being a state party focuses only on 2 to 3 states which limit their financial and human resources thus, their dependence on data analytical tools is less.

The source data is similar in these three cases but each of three political organizations has their own different data blending with other primary and secondary data to obtain a successful result. Inc uses a structured model but using the voter's list and blending it with secondary data like the census report for each state and field data collection, whereas BJP focuses itself on using social media data as their primary information to be blended with their voter list. In the case of RSP, their alone primary data is the voter's list published by the election of India.

The top to bottom approach is used by all three political organizations in data sourcing, decision making and implementation. Even though, these political parties have different organizational structure and hierarchies they use top to bottom approach

6.2 Uniqueness

INC focuses on their in-house strength of data miners, data scientist and unique programming algorithms. The reliance's on primary data and complementary secondary data helps improve their categorization and segmentation process and is expected to increase their efficiency in tailoring solutions to the identified need. Since the dependence is purely based on data and numbers any error in the assumption or entry of these data could lead to a cascading effect.

BJP, on the other hand, depends on social media platform like Twitter for identifying the trendy issues to be addressed and using Facebook to address these identified issues. This gives them an upper hand in identifying the issues earlier and addressing them on priority than the other two parties, which in turn will help BJP to target the youth segment that results in gaining dynamic support from youth. But, as reported Facebook has only 204 million active users which are only 19 % of the Indian population (DeFotis, 2017) thus using Facebook as an implementation tool could address only a maximum of 19% of the total population directly. These numbers are expected to increase in future but for the time being, the influence such a social media platform is limited to above mentioned 19%.

RSP has a different take on this whereby they still follow a mix of both traditional and a new electoral data analysis model with more prominence to the former. As RSP is a state party their concentration is much lesser the other two in term of the geographical area, this allows them better

manageability and good bondage among the party workers along with fast pace communication and implementation. The interview outcome narrates that the use of electoral data analysis and electoral engineering will increase in the coming years, considering this fact, lacking on such a decisive factor may backfire in the electoral field until they are adapted to the changing electoral strategies.

Each party in its own way tries to attain a competitive advantage over the other, from the conducted interviews it's clear that the deciding factor in the next and future election would be based on the accuracy of each electoral engineering model that the parties adopt. With the increase in the 4G connectivity and the number of the online users, the amount of data that would be available for each party would also be increasing. This is expected to increase the number of new variables and categories to be considered for the analytical models in future which would improve the accuracy of the prediction as well.

6.3 Strength & Weakness

The three parties in this study have shown their strength and weakness with respect to their analysis accuracy. Since the study was between two national parties and one state party the criteria for defining accuracy was not seen to be just data analytical capability rather, in the case of RSP their strength depended on a closed communication channel from top to bottom this enhanced the working efficiency. In the case of BJP and INC, it is hard to communicate from top to bottom due to their long hierarchical chain and the large volume of workers who cover a larger geographical segment. In the case of INC and BJP, their strength lies in the accuracy of their adopted programming models. BJP's strength lies in its ability to respond to the social issues faster and address it with a bigger reach as compared to INC, this is possible as BJP takes primary data directly from the social responses of the people from social media platform like twitter directly. INC draws its strength with the amount of data they have collected and the use of both primary and secondary in a proper bend which increases the accuracy for tailored targeting of the bigger masses compared to the less tailored mass marketing of social issues by BJP.

INC depends on numerical data that is provided by the voter's list and the secondary data sets as mentioned this fails to address the emotional sensitivity of the people. The structure of INC provides for a data segmentation, data categorization and data analysis without having any direct contact or feedback from the voter. Once the analysis part is completed these data is tallied with

the voter opinion and feedback in the fieldwork conducted using the lower level committee. Whereas, BJP bases its primary data based on the emotional sensitivity and outlook of the people from social media and addresses the needful. But, the access to data is limited as only to 22.2 million people (HuffPost India, 2018) of the around 814 million who were registered to vote in the 2014 parliamentary election (JAZEERA, 2014). This accounts for only 2.7% of the total votes that were cast in the 2014 parliamentary election, this could be a challenge in identifying the overall need and emotion of the total electorate mass. RSP still depends highly on its hand-on traditional campaigning method, even though this method has its advantages in the smaller geographical area where the party has its influence. It has not immersed itself in the electoral data analysis process which in the long run when competition with these national places would lose their competitive advantage. This lack of competitive advantage would reduce their chances to grow from a state party to a national party which are dominated by data analytically drive political organizations

6.4 Recommendation

As this study has been executed through a practical method, an ideal election engineering methodology can be proposed with the support of proper data gathering, and data analysis. The construction and sequences of this process would be

1. Identifying the sources of data
    a. Identify the current data source
    b. Identify the historic data source
2. Evaluating the relatability of the sources
3. Gathering the desired data
4. Data analysis
    a. Current data analysis
    b. Historic data analysis
    c. Blending the current and historical data analysis
5. Deriving the key pillar categories
6. Refining the pillar points with impact assessment
7. Prepare a proposed plan

1. Identifying the sources of data

The source of data is the prime most important aspect in this framework, as the result is depended on the quality of this data. There are two segments of data required and the first one is the current data source and historical data, for which an identification of source which can deliver both desired quantity and quality. The source for current data will not be limited as a single source, rather it can go up to the desired number will be factored depending on the requirement. An ideal primary current data contains information with identification aspect like name age, occupation, address and social characteristics example voters list as used by INC and social media inputs which the BJP utilizes.

The historic data consist of the previous electoral information such as overall voting percentage, male-female voting proportion, age-wise segmentation, occupational segments, religious outlook social views and the key socio-economic factors that had influenced the previous elections. This source could be from any private or public archives such as newspaper, television census report published by the government and any previously processed electoral consultancy data. Inc was seen to be utilizing the historic data in its collection process which included the use of census report and the economic survey of India as explained by INC_1.

2. Evaluating the relatability of the sources

The reliability of the utilized data needs to be cross-checked as input data determine the accuracy of the final output or result. If the data collected is not from reliable sources the error that would be included in the study would multiply over each process and the accuracy of the result could be compromised. In the interviews conducted they had failed to portray much cross-checking of their input data. INC has shown a form of blending of the historic data and the current data which could act as a filter and flag and high variants, but this too is minimal in the current state from the analysis of the conducted interviews.

3. Gathering the desired data

The desired data to be collected depend on the strength and the overall influence of the political organization across the country as the collected data depends on the availability of workers. The accuracy of the data analysis depends on the size of the data collected and it could be either random sample collection or categorized and planned collection of the population using the census report

published by the Election Commission of India. A collection of data based on the area would be recommend to the volunteers with each volunteer in the party being assigned the task of conducting an individual survey in the area of his influence using a common centralized survey system database. The survey could be conducted using an electronic device such as iPad or smartphone and even using paper which later can be entered into the centralized survey database. The database can be located in each state which can be accessed and collected by the central analysis team.

4. Data Analysis

This stage takes into account in finding any visible underlying aspects and category segments, among the two-collected data (current data & historic data) and blending these two data to find any visible aspects or category segments between them. This is can be achieved by using customized algorithm programs like k-clusters for finding patterns among the data and the use of "modelling by hand" technique for creating unique programs to understand the specific local condition and specific results.

Each of the political entities has their own outlook into the data analysis part wherein BJP they focus more on social media platforms in identifying the patterns and doing category segmentation of the electoral masses. INC focuses more on the above mentioned "modelling by hand" where they use very specialized programs to blend the current and historic data to identify the pattern and category within the electoral masses.

5. Deriving the key pillar category

Once the analysis is done the output will consist of many segments and categories from which we identify and define who the high impact potential factor is, which could range from 4 to 5 segments. These main segments are referred to as the key pillar categories as their changes in outlook will influence high electoral masses. This process of identifying these pillar categories can be based on the reactions of these categories on previous situation or circumstances and how they reacted to such a situation and who made the most influence in such a situation. Once they are identified our target segment will be ready for study and experiment and the results of these sample experiment give a small picture of how the strategies adopted by the team are seen by the electoral mass.

6. Refining the pillar points with impact assessment

The identified pillar points need to be refined more by experimenting them with sample method which would fourths help us identify the levels of impact each category have over the masses. Refining the pillars would also help segment the electoral masses more accurately with respect to the dynamic changing environment during the election time.

7. Prepare a proposed plan

The final stage of this would be the preparation of the strategies to be adopted by the political organizations, the implementation of which would be carried out with the currently available party structure by tailoring it. The efficiency of implementation will be high for a cadre-based political organization like the BJP and RSP but less for democratic parties like INC. The feedback from the party workers on the execution part is the prominent resources which are likely to be ignored. The feedback would be regarding the implementation method or even the strategies adopted for implementation. This response from the worker will help tailor the strategies and the way its conveyed which implemented properly can make the final change.

6.5 Ethical Considerations

Ethical infringement starts from the moment political parties collect information from the people be it primary or secondary to the moment they implement their proposed plan such as micro-targeting. Two ways this can be detected and controlled are

1. Having an informed society - The people are aware of their data privacy law and refrain from providing any personal data without knowing why
2. An informed Governmental Institution – Where the government is aware of such possible data breach and regulate strict measure for the collection and analysis of any such data by a third party.

An ethically ideal situation would be a mix of both the two conditions, but in an India context with 1.3 billion people and a developing economy the best option would be the 2nd where the governmental institution steps in to regulate the data collection and analysis of not only electoral engineering but rather the whole data privacy. As the inclusion of technology is increasing in an exponential rate the general population of India will also attain its ethically ideal situation of

having both an informed society and informed government. Till then the best solution outcome would be the setting of regulatory bodies supported by strict supporting rules and laws laid out by the government.

7. Limitation

The research was conducted under some assumptions and restriction, but measures were taken to minimize the error possible and reduce the number of the assumption made. Some of the limitation faced while this research was done was;

1. Dependence on third-party source
2. Vastness of the topic
3. Diversity in the Indian context
4. Time

1. Dependence on the Third-party source

The primary data collected for this research is through a semi-formal interview which has its own limitation, the six-question asked has helped us identify the way and means. But, the interview was conducted with people who are working for different political organizations, thus are very much prone to bias. Similarly, the flaws in their own model will not be exposed by themselves thus the negatives and critics for each model were based on comparison with other models adopted by other political entities in India and around the world. Finally, the specifics of each programming algorithm was not exposed by any of the interviewees thus we have to take it for face value and cannot cross-check their model or its functionality. But the interviewee was conducted through personal references thus increasing the chances for what they had said in the interview to be true. The personal reference here related to the people who had referred us to the interviewee which include close friends and family.

2. Vastness of the topic

Data analysis in electoral campaign though developed in the last 10 years, has developed a high degree of vastness and extend in terms of data availability, data analysis using uniquely designed programming algorithms and implementation. In such a case limiting to study to its effectiveness and ethical implications were a challenge in terms of what to select and what to omit from the study such as detailed discussion of the "Narwhal" program used in 2012 presidential election (Gallagher 2012) which revolutionized the data-driven campaigns. Similarly, a detailed discussion of many programming algorithms had to be avoided in order to stay within the confines of the

current study's context. But, relevant information's and research materials are provided which gives the idea of how effective and how ethically this currently carried out and how it can be improved

3. Diversity in an Indian context

The numbers of diversified variables are high with respect to other countries where such data-driven campaigns are successfully conducted like the United States and the United Kingdom. To give the diversity in understandable terms there are 27 officially recognized official languages and more than 50 local languages widely used similarly there are 29 states and 7 union territories summing up to 36 different entities with each state having its own variance in culture, caste and social norms. With such high range of variables in one country, it would hard to identify specific categories or segments that could address the whole need or want of the people thus, identifying the pillar segments and sample category would be a challenge in its own. The solution for which could be a localized data analysis which would bring up the cost of implementation high.

4. Time

The time frame was a factor which changed the research method from a detailed sample sized survey of the people involved in the analysis and the people analyzed to an interview-based research. The time required for sample questionnaire-based research was not pragmatic even though the research would have been more detailed.

8. Conclusion

With the advancement in technology and increased access to information with the help of these advancements, the election campaigning methods to needs to be adaptable to these dynamic environments. The election campaigning model has changed from area-based targeting to individual level micro-targeting with the help of newly available voter database. Data-driven election campaigning has occupied an important position in the electoral process mainly due to the predictive models that produce individual-level scores that predict citizens' likelihoods of performing certain political behaviours, supporting candidates and issues, and responding to targeted intervention.

These predictive model and micro-targeted interventions have made communication between the political parties and the electoral masses more effective and efficient. This has always been the core strategic requirement of all political parties, which was to communicate their ideology and what they have to say to the public. The in-depth data analytical methods such psychographic analysis of individual voter helps in tailoring the means and content of the message that the political organizations hope to communicate with different segments of the population with never before seen accuracy. The ethical implications of such microtargeting methods need to be addressed and controlled by the law of the land as right and wrong are highly subjective from the eye of the viewer. But the law is always objective and help draw and strict boundary on which such data analysis could be carried out.

In this era of technological advancement depriving the political organization of such tools is not practical but it is possible to strictly monitor and control them and data privacy breach. Many western countries have taken initial measures to address such infringement of privacy issues like the General Data Protection Regulation (GDPR) passed by countries like the US, UK and the European Union. The way forward for India in this ethical dilemma would be to learn and adapt from the already existing new data privacy law such as the General Data Protection Regulation (GDPR) and tailoring them to an Indian context. As explained in the recommendation the population of India are not fully aware of their data right as such the governmental institutions are expected to step in prevent any exploitation of such right of its citizens.

9. Reference

Ariely, D. (2010). Column: You Are What You Measure. *Harvard Business Review*. [online] Available at: https://hbr.org/2010/06/column-you-are-what-you-measure [Accessed 16 Apr. 2018].

Banister, P., Bunn, G. and Burman, E. (2011). *Qualitative methods in psychology*. Maidenhead: McGraw-Hill.

Bodó, B., Helberger, N. and Vreese, C. (2017). Political micro-targeting: a Manchurian candidate or just a dark horse?. *INTERNET POLICY REVIEW*, [online] 6(4). Available at: https://policyreview.info/node/776/pdf [Accessed 5 Jan. 2018].

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5-32.

Breiman, L., Friedman, J., Stone, C. and Olshen, R. (1984). *Classification and Regression Trees*. Boca Raton, FL: Taylor & Francis.

Brinkmann, S. (2013). *Qualitative interviewing*. New York: Oxford University Press.

Bryman, A. (2012). *Social research methods*. 4th ed. Oxford: Oxford University Press.

Bryman, A. and Bell, E. (2011). *Business research methods*. Oxford: Oxford University Press.

Burgess, M. (2018). *What is GDPR? The summary guide to GDPR compliance in the UK*. [online] Wired.co.uk. Available at: https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018 [Accessed 4 Jul. 2018].

Chief Election Office, Kerala (2018). *Voter's List 2018*. 2 Kasaragod Assembly Constituency. [online] Kasaragod: Government of Kerala. Available at: http://www.ceo.kerala.gov.in/electoralrolls.html [Accessed 5 Jul. 2018].

Chipman, H., George, E. and McCulloch, R. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), pp.266-298.

DeFotis, D. (2017). *India Facebook Users Surpass U.S.: Is It Demonetization, Apple?*. [online] Barrons.com. Available at: https://www.barrons.com/articles/india-facebook-users-surpass-u-s-is-it-apple-demonetization-1499982716 [Accessed 6 Apr. 2018].

Dexter L. A. 2006. *Elite and Specialized Interviewing*. Colchester: ECPR Press.

Eci.nic.in. (2018). *Election Commission of India*. [online] Available at: http://eci.nic.in/eci_main1/Linkto_erollpdf.aspx [Accessed 1 Aug. 2018].

Endres, K. and Kelly, K. (2017). Does microtargeting matter? Campaign contact strategies and young voters. *Journal of Elections, Public Opinion and Parties*, 28(1), pp.1-18.

Flick, U. (2011). *Introducing Research Methodology: A Beginner's Guide to Doing a Research Project*. Los Angeles, CA: Sage.

Galdies, P. (2018). *A Summary of the EU General Data Protection Regulation | www.dataiq.co.uk.* [online] Dataiq.co.uk. Available at: https://www.dataiq.co.uk/blog/summary-eu-general-data-protection-regulation [Accessed 5 Aug. 2018].

Gallagher, S. (2012). *Built to win: Deep inside Obama's campaign tech.* [online] Ars Technica. Available at: https://arstechnica.com/information-technology/2012/11/built-to-win-deep-inside-obamas-campaign-tech/ [Accessed 30 Mar. 2018].

Gan, G., Ma, C. and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, Pa.: SIAM, Society for Industrial and Applied Mathematics.

Gerber, A. and Green, D. (2000). The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment. *American Political Science Review*, 94(03), pp.653-663.

Goldstein, K. and Ridout, T. (2002). *Political Behavior*, 24(1), pp.3-29.

Grassegger, H. and Krogerus, M. (2018). *The Data That Turned the World Upside Down.* [online] Motherboard. Available at: https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win [Accessed 4 Jun. 2018].

Green, D. and Kern, H. (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, 76(3), pp.491-511.

Hillygus., D. Sunshine., Shields., Shields, T. and Todd G. (2009). *Persuadable Voter*. Princeton: Princeton University Press.

HuffPost India. (2015). *India Has 22.2 Million Twitter Users: Report*. [online] Available at: https://www.huffingtonpost.in/2015/01/28/twitter-india-userbase-

report_n_6562950.html?guccounter=1&guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvLnVrLw&guce_referrer_cs=xP3U_6JGk2l1EziZ3Y-aMg [Accessed 10 May 2018].

Imai, K. and Strauss, A. (2011) "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign," *Political Analysis*. Cambridge University Press, 19(1), pp. 1–19. doi: 10.1093/pan/mpq035.

Information Commissioner's Office (2018). *Overview of the General Data Protection Regulation (GDPR)*. [online] UK: Information Commissioner's Office. Available at: https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-13.pdf [Accessed 4 Aug. 2018].

Institut Numerique, (2012). *Research Methodology*, http://www.institut-numerique.org/chapter-3-research-methodology-4ffbd6e5e3391 [retrieved 3rd October, 2014].

JAZEERA, A. (2014). *India announces election dates*. [online] Aljazeera.com. Available at: https://www.aljazeera.com/news/asia/2014/03/indian-announces-election-dates-2014355402213428.html [Accessed 10 Apr. 2018].

Jetley, N. (2014). *How big data has changed India elections.* [online] CNBC. Available at: https://www.cnbc.com/2014/04/10/how-big-data-have-changed-india-elections.html [Accessed 5 Aug. 2018].

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2), p.119.

Kramer, Gerald H. 1966. "A Decision Theoretic Analysis of a Problem in Political Campaigning." *In Mathematical Implications in Political Science. Vol. 2*. edited by Joseph L.Bernd, 137–160. Dallas: Southern Methodist University Press, Dallas.

Kreiss, D. (2016). *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. Oxford: Oxford University Press.

Lapowsky, I. (2018). *A Lot of People Are Saying Trump's New Data Team Is Shady*. [online] WIRED. Available at: https://www.wired.com/2016/08/trump-cambridge-analytica/ [Accessed 4 Mar. 2018].

Mak, T. (2018). *Cambridge Analytica Scandal Raises New Ethical Questions About Microtargeting*. [online] Npr.org. Available at: https://www.npr.org/2018/03/22/596180048/cambridge-analytica-scandal-raises-new-ethical-questions-about-microtargeting?t=1533299048407 [Accessed 3 Mar. 2018].

Malchow, H. (2008). *The new political targeting*. Washington, DC: Predicted Lists.

Matz, S., Kosinski, M., Nave, G. and Stillwell, D. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), pp.12714-12719.

May, T. (2011). *Social research: Issues, Methods and Research*. Berkshire: McGraw Hill.

Michal Kosinski, David Stillwell, and Thore Graepel, 2013. "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, volume 110, number 15 (9 April), pp. 5,802–5,805, at http://www.pnas.org/content/early/2013/03/06/1218772110, accessed 30 March 2013. doi: http://dx.doi.org/10.1073/pnas.1218772110, accessed 25 June 2014.

MURRAY, G. R. and SCIME, A. (2010). Microtargeting and electorate segmentation: Data mining the American national election studies. *Journal of Political Marketing* 9 143–166.

Newman, I. and Benz, C. (2001). *Qualitative-quantitative research methodology*. Carbondale, Ill: Southern Illinois Univ. Press.

Nickerson, D. and Rogers, T. (2010). Do You Have a Voting Plan?. *Psychological Science*, 21(2), pp.194-199.

Nickerson, D. and Rogers, T. (2013). Political Campaigns and Big Data. *SSRN Electronic Journal.*

Nickerson, D. and Rogers, T. (2014). Political Campaigns and Big Data. *Journal of Economic Perspectives*, [online] 28(2), pp.51-74. Available at: https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.51 [Accessed 6 Feb. 2018].

Nickerson, David W. 2006a. "Hunting the Elusive Young Voter," *Journal of Political Marketing* 5(3):47-69.

Pathak, S. (2017). *How To Use Voter Data In Your Campaigns - CallHub.* [online] CallHub. Available at: https://callhub.io/use-voter-data-campaigns/ [Accessed 1 Aug. 2018].

Pawha Jetley, N. (2014). *How big data has changed India elections.* [online] CNBC. Available at: https://www.cnbc.com/2014/04/10/how-big-data-have-changed-india-elections.html [Accessed 4 Mar. 2018].

Rogers, Todd, and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." *American Politics Research*, published online before print, September 5

Sahay, Arunaditya. (2016). *Peeling Saunder's Research Onion*. Shodh Gyan. ISSN 2395-0617

Sathiaraj, D., Cassidy, W. and Rohli, E. (2017). *Improving Predictive Accuracy in Elections. Big Data*, [online] 5(4), pp.325-336. Available at: https://www.liebertpub.com/doi/full/10.1089/big.2017.0047 [Accessed 26 May 2018].

Sharma, S. (2018). *How the new age voters will be crucial in deciding the political fortunes in 2019*. [online] The Economic Times. Available at: https://economictimes.indiatimes.com/news/politics-and-nation/why-political-parties-cant-ignore-these-10-cr-voters-in-their-quest-for-2019/articleshow/63347913.cms [Accessed 5 Jul. 2018].

Shen, G. (2013). *Big data, analytics and elections - Analytics Magazine.* [online] Analytics Magazine. Available at: http://analytics-magazine.org/big-data-analytics-and-elections/ [Accessed 24 May 2018].

Shirky, C. (2008). *Here comes everybody*. New York: Penguin.

Srivas, A. (2018). *Facebook-to-Votes Scandal Turns Spotlight on Cambridge Analytica's India Inroads*. [online] The Wire. Available at: https://thewire.in/politics/facebook-to-votes-scandal-turns-spotlight-on-cambridge-analyticas-india-inroads [Accessed 15 Apr. 2018].

Talwar Thakore & Associates (2018). *General | Data Protection Laws*. Data Protected - India. [online] london: linklaters. Available at: https://www.linklaters.com/en/insights/data-protected/data-protected---india [Accessed 5 May 2018].

Tufekci, Z. (2014). *Engineering the public: Big data, surveillance and computational politics.* [online] First Monday. Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4901/4097 [Accessed 24 May 2018].

Valdez, A. (2018). *Everything You Need to Know About Facebook and Cambridge Analytica.* [online] WIRED. Available at: https://www.wired.com/story/wired-facebook-cambridge-analytica-coverage/ [Accessed 3 Mar. 2018].

Witten, I., Frank, E., Hall, M. and Pal, C. (2016). *Data Mining*. 4th ed. Saint Louis: Elsevier Science.

Zuiderveen Borgesius, F., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B. and De Vreese, C. (2018). Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review*, 14(1), p.82.