



## Engineering the public: Big data, surveillance and computational politics

by Zeynep Tufekci

### Abstract

Digital technologies have given rise to a new combination of big data and computational practices which allow for massive, latent data collection and sophisticated computational modeling, increasing the capacity of those with resources and access to use these tools to carry out highly effective, opaque and unaccountable campaigns of persuasion and social engineering in political, civic and commercial spheres. I examine six intertwined dynamics that pertain to the rise of computational politics: the rise of big data, the shift away from demographics to individualized targeting, the opacity and power of computational modeling, the use of persuasive behavioral science, digital media enabling dynamic real-time experimentation, and the growth of new power brokers who own the data or social media environments. I then examine the consequences of these new mechanisms on the public sphere and political campaigns.

### Contents

#### [Introduction](#)

#### [Engineering the public: From broadcast to the Internet](#)

#### [New dynamics of persuasion, surveillance, campaigning and social engineering](#)

#### [Consequences and power of big data analytics](#)

#### [Discussion and conclusion](#)

### Introduction

Emergence of networked technologies instilled hopes that interactivity in the public sphere could help limit, or even cure, some of the ailments of late modern democracies. In contrast to broadcast technologies, the Internet offers expansive possibilities for horizontal communication among citizens, while drastically lowering the costs of organizing and access to information (Shirky, 2008). Indeed, the Internet has been a critical tool for many social movements (Tufekci and Freelon, 2013).

However, Internet's propensity for citizen empowerment is neither unidirectional, nor straightforward. The same digital technologies have also given rise to a data-analytic environment that favors the powerful, data-rich incumbents, and the technologically adept, especially in the context of political campaigns. These counter-trends arise specifically from an increased exploitation on *big data*, that is, very large datasets of information gleaned from online footprints and other sources, along with analytic and computational tools.

Big data is often hailed for its ability to add to our knowledge in novel ways and to enrich our understanding (Lazer, *et al.*, 2009; Lohr, 2012). However, big data also needs to be examined as a political process involving questions of power, transparency and surveillance. In this paper, I argue that big data and associated new analytic tools foster more effective — and less transparent — “engineering of consent” (Bernays, 1947) *in the public sphere*. As a normative (but contested) ideal, the public sphere is envisioned by Habermas (1989) as the location and place in which rational arguments about matters concerning the public, especially regarding issues of governance and the civics can take place, freed from constraints of status and identity. The public sphere should be considered at once a “normative ideal” as well as an institutional analysis of historical practice (Calhoun, 1993). As actual practice, the public sphere pertains to “places” — intersections and commons — where these civic interactions take place, and which are increasingly online. This shift to a partially online public sphere, which has brought about the ability to observe, surveil and collect these interactions in large datasets, has given rise to *computational politics*, the focus of this paper.

*Computational politics* refers applying computational methods to large datasets derived from online and off-line data sources for conducting outreach, persuasion and mobilization in the service of electing, furthering or opposing a candidate, a policy or legislation. Computational politics is informed by behavioral sciences and refined using experimental approaches, including online experiments, and is often used to profile people, sometimes in the aggregate but especially at the individual level, and to develop methods of persuasion and mobilization which, too, can be individualized. Thus, computational politics is a set of practices the rise of which depends on, but is not solely defined by, the existence of big data and accompanying analytic tools and is defined by the significant *information asymmetry* — those holding the

data know a lot about individuals while people don't know what the data practitioners know about them (U.S. Federal Trade Commission, 2014).

While computational politics in its current form includes novel applications, the historical trends discussed in this paper predate the spread of the Internet. In fact, there was already a significant effort underway to use big data for purposes of marketing, and the progression of using marketing techniques for politics — and “selling of the President” — clearly reflects longer-term trends (McGinniss, 1988). However, computational politics introduces significant qualitative differences to that long march of historical trends. Unlike previous data collection efforts (for example, collating magazine subscriptions or car type purchases) which required complicated, roundabout inferences about their meaning (does a magazine subscription truly signal a voter preference?) and allowed only broad profiling in the aggregate, this data provides significantly more individualized profiling and modeling, much greater data depth, and can be collected in an invisible, latent manner and delivered individually.

Computational politics turns political communication into an increasingly personalized, private transaction and thus fundamentally reshapes the public sphere, first and foremost by making it less and less *public* as these approaches can be used to both profile and interact *individually* with voters outside the public sphere (such a Facebook ad aimed at that particular voter, seen only by her). Overall, the impact is not so much like increasing the power of a magnifying glass as it is like re-purposing the glass by putting two or more together to make fundamentally new tools, like the microscope or the telescope, turning unseen objects into objects of scientific inquiry and manipulation.

*Big data's* impact on the public sphere through *computational politics* operates through multiple intertwined dynamics, and the purpose of this paper is to define, explain and explore them both individually, and also within the context of this intertwining. **First**, the rise of digital mediation of social, political and financial interactions has resulted in an exponential increase in the amount and type of data available, especially to large organizations that can afford the access, *i.e.*, big data. **Second**, emergent computational methods allow political targeting to move beyond aggregated group-based analysis and profiling to modeling of *specific individuals*. **Third**, such modeling allows 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. **Fourth**, advances in behavioral sciences have resulted in a move away from models of the “rational human” towards more refined and realistic models of human behavior. In combination with the other dynamics outlined here, these models allow for enhanced, network-based *social engineering*. **Fifth**, digital networks enable these methods to be experimentally tested in real time and for immediate deployment, adding a level of previously unfeasible dynamism and speed to molding the public sphere. **Sixth**, the data, tools and techniques that comprise these methods require access to proprietary, expensive data, and are driven by *opaque algorithms* — opaque algorithms refers to “black box” algorithms the operations of which are proprietary and undisclosed, and most of which are controlled by a few Internet platforms. In other words, an ordinary user sees an opaque algorithm as a black box.

Though the field is growing — and significant contributions have been made by boyd and Crawford (2012), Kreiss (2012b), Bryant and Raja (2014) and Lazer, *et al.* (2009), among others — there has been fairly little conceptual theory-building especially about the political and civic consequences of big data. Popular media, on the other hand, rarely goes beyond exploring big data as a hot, new topic and an exciting new tool, and rarely consider issues of power.

Previous scholarly exploration of new media technologies and politics includes, most notably, the prescient analysis by Howard (2006), which anticipates some of the key aspects examined here, especially with regard to new media, as well as Kreiss (2012b), Howard and Kreiss (2010), and Kreiss and Howard (2010) which focus more on privacy aspects. However, social media platforms that are increasingly integral to the practice of computational politics have fully blossomed only recently. These new practices build upon the growing ability of campaigns to use technology to “manage” the electorate (Howard, 2006), a dynamic which has so far been examined in case studies of Barack Obama's campaigns (Bimber, 2014; Carty, 2011; Issenberg, 2012; Kreiss, 2012a; Tufekci, 2012), as well as an ethnographic account of a Congressional effort (Nielsen, 2012). While I, too, pick multiple examples from the Obama campaign, as it is the most recent, best studied and most relevant one, this paper is not meant as a study or indictment of any particular campaign, nor to imply that all the developments outlined here fully practiced by any single campaign [1]. Rather, this is an empirically based conceptual, theory-building paper that grapples with the consequences of newly emergent computational politics.

Consequently, this paper focuses on the intertwined dynamics of computational politics and big data with an emphasis on their implications for power, politics and the public sphere and engages in empirically based, conceptual theory-building required for both more conceptual and empirical research. While many of the aspects explored here apply in commercial, corporate and other spheres, albeit with different emphases, questions of political computation deserve independent analysis as the considerations are not identical, and since the practice of politics is central to questions of civics (Hillygus and Shields, 2008; Kreiss, 2012a; 2012b).



## Engineering the public: From broadcast to the Internet

This debate on meaningful participation in governance for a society that is too large for frequent and direct face-to-face interaction — any social organization bigger than a small village or a hunter gatherer tribe — goes back at least to Plato and Aristotle in written records. At its heart, this debate asks in large societies where centralization of power and delegation appears inevitable, whether citizenry — within its gradually expanding, historically variable definition — can ever be fully equipped to undertake or understand all the complex decisions that are required for governance, and further, what, if anything can keep those with power in check so that they do not assure perpetuation of their own rule.

Plato, famously, called for kings to be philosophers so that they would rule justly for the good of society but not necessarily by being truthful or accountable (Plato, 2012). A modern incarnation of Plato's call for powerful but benevolent "philosopher kings" emerged in the early twentieth century Lippmann–Dewey debates (Dewey, 1927; Lippmann, 1925). Walter Lippmann expressed pessimism at the possibility of a public actually in charge of governance, and argued that the powerful would always be able to manipulate the opinions, beliefs and ultimately, voting behavior of ordinary citizens. They would be "social engineers", in Karl Popper's terms, who manipulated the public to achieve their own goals. John Dewey, however, believed that it was possible to build social and political institutions — ranging from a free press to a genuinely enriching education — that would expose and counter the manipulations of the powerful and allow for meaningful self-governance by an educated, empowered citizenry. Though both Dewey and Lippmann worried about the powerful controlling the public, neither had experienced the full force of broadcast media, yet to come.

The rise of broadcast media altered dynamics of politics in fundamental ways. Public relations pioneer Edward Bernays explained the root of the problem in his famous "Engineering of consent" article where, discussing the impact of broadcast on politics, he argued that the cliché "the world has grown smaller" was actually false (Bernays, 1947). The world is actually much bigger and today's leaders, he pointed out, are farther removed from the public compared to the past. The world feels smaller partly because modern communication allows these leaders, potent as ever, to communicate and persuade vast numbers of people, and to "engineer their consent" more effectively.

Bernays saw this as an unavoidable part of any democracy. He believed, like Dewey, Plato and Lippmann had, that the powerful had a structural advantage over the masses. However, Bernays argued that the techniques of "engineering of consent" were value-neutral with regard to message. He urged well-meaning, technologically and empirically enabled politicians to become "philosopher-kings" through techniques of manipulation and consent engineering.

The techniques can be subverted; demagogues can utilize the techniques for antidemocratic purposes with as much success as can those who employ them for socially desirable ends. The responsible leader, to accomplish social objectives, ... must apply his energies to mastering the operational know-how of consent engineering, and to out-maneuvering his opponents in the public interest. [2]

Bernays recommended study of the public through opinion research, and controlling it through managing of communication and media. The techniques of opinion control espoused by Bernays became bread-and-butter of political campaigns in the post-war West. At its heart, this has been driven by "public opinion research" which seeks to understand, categorize and label the public and views differentiating as key to effective "persuasion" — whether marketing a politician or a soft drink. Soon after public opinion research started seeping into politics, cultural critic Adorno called the forms of "classifying, organizing and labeling" as a form of propaganda in which "something is provided for all so that none may escape." [3] In other words, Adorno feared a public sphere in which politicians correctly identified all subcategories of voters and served each of them with a palatable message.

However, messaging and mobilization based on such sub-categorization has limits intrinsic to the method, as all categorization hides variation. The match between demographics or political profiles and a specific person is, at best, broadly probabilistic and often very muddled. During the broadcast era, most targeting was necessarily course-grained, because TV audiences were measured in broad demographics. The best that aspiring micro-targeters could do was to define potential segmented audiences, like "soccer moms", by gender and age, and target programs to that gender and age group. Because audiences could not be tightly defined, messaging had to be broader as well. "Soccer moms" surely include a variety of political views and personalities. Many exposed to the ads would not fit the target group, and many members of the target group would be excluded. Research showed that such political advertisements on broadcast TV remained largely ineffective in tipping the scale between existing candidates, at least when compared with more structural factors such as the unemployment rate or economic growth [4].

Similarly, almost all voter canvassing and turnout campaigns have traditionally been based at the precinct level simply because demographic data has been available at that level. However, precinct data are probabilistic in that no precinct uniformly votes for a single party, so campaigns tend to pour resources into a specific precinct in the hopes that they will mobilize more supporters than opponents, and that their canvassing efforts will not aggravate too many supporters of the other party (Issenberg, 2012). Political campaigns, in turn, ignore precincts which contain many of their own voters, but less than those of their opponent.

Unsurprisingly, targeting individuals as individuals rather as members of broadly defined aggregates has long been the holy grail of political campaigns. Such efforts have been underway for decades. Culling information from credit cards, magazine subscriptions, voter registration files, direct canvassing efforts and other sources, political parties, as well as private databases, have compiled as much information as they can on all individual voters. However, until recently, the collection of individual level data was messy and fragmented, and targeting was still on done by aggregate groups, which were simply based on richer individualized data than before. Much of this has changed with the rise of the Internet, which greatly increases the type and amount of individual data, and computational analytics, altering what information can be gleaned from these sources.



## New dynamics of persuasion, surveillance, campaigning and social engineering

The recent rise of big data and computational techniques are changing political communication with citizens once again. If the twentieth century engineers of consent had magnifying glasses and baseball bats, those of the twenty-first century have acquired telescopes, microscopes and scalpels in the shape of algorithms and analytics. In this section, I examine six, intertwined dynamics which create a new environment of surveillance and social engineering.

**1. Big data:** The advent of digital and networked technologies has caused an explosion in the amount and variety of data available on each individual, as well as the velocity with which such data become available (Bryant and Raja, 2014; U.S. Federal Trade Commission, 2014). These large collections of data, referred to as big data, are not just more of old kind of data. Rather, in some ways, its effects are like the invention of the microscope (Brynjolfsson and McAfee, 2011) which makes visible the previously existed unseen, and in other ways, or like a telescope that allows the observer to “zoom out” and observe at a different scale, often at loss of subtlety and individuality of the data points but with powerful aggregate effects. While no single metaphor fully captures its novel impacts, big data, like the microscope and the telescope, threatens to upend our understanding of multiple fields and to transform the practice of politics.

What has changed is not just the depth and the scope of available data: the fundamental nature of data available for aggregation has undergone a significant shift. In the past, data collection was primarily “pull” (questions answered voluntarily as in surveys), supplemented by a layer of “latent data,” which are data which exist as imprints of actions we carry out as we go about our lives. In the pre-digital era, such *latent*, imprint data was limited — financial transactions, magazine subscriptions, credit card purchases. Political campaigns were faced with the task of *inferring* what such a transaction meant. Does a subscription to *Better Homes & Gardens* imply a party affiliation? Does it correspond to a position on progressive taxation? The answer often was, maybe, but only weakly. Such data provided some correlational guidance at the group level but did not allow precise individual targeting.

The rise of the Internet itself as a social, civic and political space engendered a tremendous growth in a different category of data often called “user-generated” data. Some of this growth is of *latent* data; transactions which are carried out online for a wide variety of purposes now leave behind harvestable imprints. In *latent* data, the user is going about her day, say, purchasing products and participating on social media. The imprints she leaves behind, however, carry important information and include her actual conversations. Hence, unlike the explicit process in which a respondent is asked by a pollster about her choices, and the answer recorded, campaigns can now capture actual utterances of people as they talk about a wide variety of topics of interest to them. Data brokers increasingly scrape and examine user behavior in these environments and collate the responses, which they match with vast amounts of other online and offline data about the person (U.S. Federal Trade Commission, 2014). This type of user-generated content is directly semantic, and rather than convoluted inferences, such data lends itself to deeper and direct insight into a person’s opinions, dispositions and behaviors (through computational methods discussed below).

The user-generated content environment has undergone such a dramatic change that even a mere eight years ago, when the Internet was already widespread, political campaigns had to resort to some degree of trickery to compel users to provide content. Howard (2006) documents how a political company stealthily operated a discussion forum that provided the participants with the aggregated voting record data of politicians *primarily so they could have access to the participants’ discussions*. People had to be coaxed into user-generated data. These days, such data is voluntarily and widely generated by individuals themselves as a by-product of civic participation that is *digitally mediated* — in other words, people are commenting and discussing about politics on general-purpose digital platforms and this *digital mediation* of their activities leaves behind a trove of data that is harvested by companies and data brokers.

Further, the quantitative depth of big data composed of online imprints is exponentially richer than pre-digital data. A large commercial database may easily contain from thousands data points on each individual — a recent report found that some data brokers had 3,000 individual data points per person, and were adding to it at a rapid pace (U.S. Federal Trade Commission, 2014). The volume and variety of this kind of big data is qualitatively different. If anything, the problem of data analysis today is data that is too much, too deep, and too varied. However, rise of computational methods and modeling is quickly catching up to the challenge of turning this deluge of data into usable information at the hands of political campaigns and others.

**2. Emergent computational methods:** All this data is burdensome without techniques to acquire usable information from the dataset. Computational methods used by political campaigns depend on multiple recent developments. First, technical developments in storage and database systems mean that large amounts of data can be stored and manipulated. Second, new methodologies allow processing of *semantic*, unstructured information contained in user-generated natural language outputs such as conversations — as opposed to already structured data such as a financial transaction which come in the shape of already neatly packaged fields. Third, new tools allow human interactional data can to be examined through a structural lens using such methods as social network analysis. Fourth, the scale of the data allows for new kinds of correlational analyses that would have been previously hard to imagine.

*First*, given the amount of data that is being generated, even mere storage has been a challenge, and has required developing new methods. YouTube has 72 hours of video uploaded every minute. As of last year, Facebook was processing about 2.5 billion pieces of content, 2.7 billion “like” actions, 300 million photos and, overall, 500 terabytes of data everyday [5]. Handling such large datasets has recently become easier through the development of techniques like “Hadoop clusters” which provide a system of shared storage along with “Map Reduce” which provides the analytic layer allowing for reliable and quick access to such large datasets. Facebook reportedly holds its data in a 100-petabyte Hadoop cluster.

*Second*, new computational processing techniques allow for extracting semantic information from data without using an army of human coders and analysts, as would have been required under old techniques. Techniques that automatically “score” words to generate estimates of ideological content of texts and

sentiment analysis, or group sentences into topics and themes ("Latent Dirichlet Allocation" or LDA), allows for a probabilistic but fairly powerful method of categorizing a person's approach to an issue as represented in her textual statements, but without the costly step of a human reading the actual content. Without these computational techniques, the texts would have to be read and summarized by a large number of people; even then, just aggregating the results would pose a challenge. While algorithms come with pitfalls and limitations — for example "Google Flu Trends" which was once hailed as a great innovation has turned out to produce misleading data (Lazer, *et al.*, 2009), they can be useful in providing information that would otherwise be prohibitively costly or impossible.

*Third*, social network analysis, the roots of which go to sociology in the 1950s, has seen greatly broadened utility and technical expansion in allowing analysts not just map people's views, but also to situate them within social networks. The broadened utility has occurred partly because data that is in the form a network has increased significantly due to online social network platforms that are used for a variety of ends, including politics (Howard and Parks, 2012). Previously, gathering social network information from people was a difficult and costly endeavor and various biases and difficulties in recalling social network information led to great many difficulties as even small social networks required hundreds of interviews where people were expected to name dozens, if not hundreds, of social ties. Understandably, such research has always been very difficult and carried out only on small samples.

With the advent of networks that were encoded by the software, network analysis became possible without the difficult step of collecting information directly from individuals. Researchers also started applying network analysis to broader topics where the "connections" could be interpreted as "links" in a network — such as blogosphere (with each link constituting a network connection) or citation networks of scholars. Network analysis allows identifying various structural features of the network such as "centrality", clustering (whether there are dense, distinct groupings), bridges that connect clusters and much more, which provide very valuable political information in deciding how to target or spread political material. For example, people with high centrality are useful propagators of information and opinions, and once identified, can be targeted by political campaigns as entry points to larger social networks.

*Fourth* and finally, researchers can now look for "correlations" in these huge datasets in ways that would have been difficult to impossible before. This has, of course, led to many positive applications. For example, researchers have started identifying drug interactions by looking at Google searches of multiple drugs matched by symptoms — a feat that simply cannot practically be done any other way which would mean surveying all users of all drugs about all side effects. However, for political campaigns, that also opens doors to better individualized identifying of target individuals by looking or correlations between their political choices and other attributes. Considering that data brokers have thousands of data points on almost every individual in the United States, these new computational methods available to campaigns with which to analyze, categorize and act upon the electorate.

**3. Modeling:** In this context, modeling is the act of inferring *new* information through analysis of data based on creating a computational relationship between the underlying data and target information. Modeling can be vastly more powerful than aggregate profiling. Aggregate profiling attempts to categorize a user, by putting her in a category with many others, by combining available data. Someone who answered a survey question on environmental issues as "very important" to her, for example, is likely an environmentally conscious voter. Combined with purchase data ("shops at Whole Foods"), a campaign may profile her as an environmental voter.

However, the advent of big datasets that contain imprints of actual behavior and social network information — social interactions, conversations, friendship networks, history of reading and commenting on a variety of platforms — along with advances in computational techniques means that political campaigns (and indeed, advertisers, corporations and others with the access to these databases as well as technical resources) can model *individual* voter preferences and attributes at a high level of precision, and crucially, often without asking the voter a single direct question. Strikingly, the results of such models may match the quality of the answers that were only extractable via direct questions, and far exceed the scope of information that could be gathered about a voter via traditional methods.

For example, 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) Researchers' models which solely used Facebook "likes" — a fraction of the data available to any data broker — correctly discriminated whether the Facebook user is heterosexual or not in about 88 percent of the cases; and predicted race with about 95 percent of the time and political party affiliation about 85 percent of the time (Kosinski, *et al.*, 2013). In other words, just access to a fraction of Facebook data, processed through a computational model, allows for largely correctly delineating Republicans and Democrats without looking into any other database, voter registration file, financial transactions or membership in organizations.

While parts of this example may seem trivial since some of these, such as age and gender, are traditional demographics and are usually included in traditional databases, it is important to note that these are being estimated through modeling, and are not asked or observed from the user. This means that that these attributes can also be modeled in platforms where anonymous or pseudonymous postings are the norm. This type of modeling also furthers information asymmetry between campaigns and citizens; campaigns learn about a given voter with the voter having no idea about this modeling in the background.

Crucially, this type of modeling allows access to psychological characteristics that were beyond the reach of traditional databases, as invasive as those might have been considered. Personality traits such as "openness" or "introversion" or "neuroticism" are traditionally measured by surveys, which have been developed and validated by psychologists and used on a large number of people for decades. While such traits themselves may be generalizations, they are significantly more detailed than the crude demographics employed by political campaigners ("soccer moms"). Kosinski, *et al.* (2013) demonstrated that models based on Facebook likes were as good as scientific scales. In other words, without asking a



single question, researchers were able to model psychological traits as accurately as a psychologist administering a standardized, validated instrument. Given that social media data have been used to accurately model attributes, ranging from suicide rates to depression to other emotional and psychological variables (De Choudhury, *et al.*, 2013; Culotta, 2014; Quercia, 2013) and given that social media is just one facet of information available for big data modeling, it is clear that political campaigns can have a much richer, more accurate categorization of voters, and without and before necessarily having knocked on their door a single time to ask a single question.

To understand why this is a major shift, consider how different it is compared with cruder, more basic profiling that has been used in traditional survey research to identify “likely” voters — a key political concern to campaigns [6]. For decades, traditional polling organizations and campaigns have been trying to model “likely” voters in their surveys with varying degrees of success. Campaigns do not want to spend resources on people who are unlikely to vote and pollsters need this data to weigh their data correctly. Asking the question itself (“are you likely to vote?”) has not proven that useful, due to the well-known socially desirable answer bias: many “unlikely” voters declare their intention to vote. Previous voting records are also tenuous predictors — besides, there are many young voters entering the rolls. Gallup, whose likely voter model had long been considered the gold standard, asks a series of seven questions that include intent, knowledge (where is your voting booth?), past behavior, and implicit measures (how often do you think about the election). However, even with decades of expertise, Gallup has been missing election predictions, due to its inability to correctly predict likely voters through survey data. The gravity of the situation such that Gallup became a punch line: at the White House Correspondents’ Association dinner, President Obama made a predictable joke followed by a joke about who didn’t see the [predictable] joke coming: “Show of hands? Only Gallup?” he quipped (Yagoda, 2013).

In contrast, during the 2012 election, the Barack Obama campaign developed a fairly sophisticated model of “likelihood of turnout” based on its datasets, which do not just rely on surveys but incorporate the kinds of data discussed in this paper, and generated an index from 0 (not going to vote) to 100 (will certainly vote) for each potential voter. This resulted in a targeted, highly efficient persuasion and turnout effort which focused mostly on turning out voters that were already Obama supporters rather than spending a lot of effort persuading voters who would not end up voting. This left the Romney campaigns, reliant on more traditional efforts, so far behind that after their loss, Romney staffers were left exclaiming that the Obama campaign turned out voters that the Romney campaign “never even knew existed” (Rutenberg and Zeleny, 2012). In 2014, Obama campaign staffers told a gathering at the Personal Democracy Forum that in key states, they were able to go deep into Republican territory, to individually pick voters that they had modeled as likely Democrats within otherwise Republican suburbs, breaking the lock of the precinct at voter targeting. The advantages of stronger, better modeling, an expensive undertaking that depends on being able to purchase and manipulate large amounts of data, can hardly be overstated.

Finally, big data modeling can predict behaviors in subtle ways and more effectively oriented toward altering behavior. For example, for years, the holy grail of targeting for commercial marketers has been pregnancy and childbirth, as that is a time of great change for families, resulting in new consumption habits which can last for decades. Previously, retailers could look at obvious steps like creation of a baby registry; however, by then, the information is often already public to other marketers as well, and the target is well into the pregnancy and already establishing new consumption patterns. By combining rich data with predictive modeling, however, the retail giant Target is not only able to identify potential pregnancy very early on, in the first or second trimester, it can also estimate the due date “within a small window” so that it can send coupons and ads timed to stages of pregnancy and parenting. In a striking example, Duhigg (2012) recounts the tale of an angry father walking into Target, demanding to see the manager to ask why his teenage daughter was being sent advertisements for maternity clothing, nursery furniture and baby paraphenelia. The manager, it was reported, apologized profusely, only to receive an apology himself when the father went back home to talk with his daughter, who was, indeed, pregnant. Data modeling ferreted out facts that a parent did not know about his own child living under his own roof.

**4. Behavioral science:** These predictive analytics would not be as valuable without a corresponding rise in sophistication of behavioral science models of how to persuade, influence and move people to particular actions. Developing deeper models of human behavior is crucial to turning the ability to look, model and test big data into means of altering political behavior.

The founder of public relations, Edward Bernays, himself had posited that people were fundamentally irrational. However, the rational and “utility-maximizing” aspects of human behavior have long been emphasized in dominant academic literature, especially in fields such as economics and political science. While Habermas’ (1989) ideal of the public sphere imagined status-free actors carrying out rational conversations based on merit, political practitioners have long recognized that the “rational voter” model did not correspond to their experience in the world. However, until recently, there was fairly little systematic analysis of “hooks” for steering “irrationality” into desired outcomes — tools required for the instrumental and rational manipulation of human irrationality had not yet been developed.

All this changed thanks to research which emphasized the non-rational aspects of human behavior, and with attempts to measure and test such behavior modification within political contexts. Just as behavior analysis became more sophisticated, for the first time in modern political history, an influx of scholars from the behavioral sciences moved into practical politics, starting with the 2008 Obama campaign. Hence, as recounted by Issenberg (2012), there was a significant shift from “grant narratives” and tidy conclusions produced by pundits towards an operational fight focusing on altering the behavior of individual voters, aided, crucially, by insights from in-house behavioral scientists.

It wasn’t that behavioral science could overcome the structural conditions of a campaign like a bad economy or a profoundly unattractive candidate. Increasingly, however, elections are fought at the margins in part because of pre-existing polarization, a winner-takes-all system in the case of United States, and low turnout. Under these conditions, the operational capacity to find and convince just the right number of individual voters becomes increasingly important [7]. In such an environment, small

differences in analytic capacity can be the push that propels the winning candidate. For example, behavioral political scientists working for political campaigns found “plain white envelopes” work better than glossy mailings in signaling credibility; hence these were increasingly used by the Obama campaign. Consequently, merging empirically based behavioral science with “psychographic” profiles, obtained from computationally modeling big data as discussed earlier, can create a significant advantage.

Combining psychographics with individual profiles in a privatized (*i.e.*, non-public) communication environment can be transformative for political campaigns. For example, campaigns have often resorted to fear (or other tactics that appeal to the irrational) such as the infamous “daisy/nuclear war” ad during the 1964 U.S. Presidential campaign. However, research shows that when afraid, only some people tend to become more conservative and vote for more conservative candidates. Campaigns, though, until now had to target the whole population, or at least a substantial segment, all at once, with the same message. In contrast, by modeling individual psychologies through computational methods applied to big data, a political campaign hoping to garner votes of a conservative candidate can plausibly (and relatively easily) identify voters that were more likely to react to fear by voting conservative and target them individually through “fear-mongering” tactics designed for their personal weaknesses and vulnerabilities [8] while bypassing individuals on whom fear-mongering would not have the desired effect, perhaps even the opposite of the desired one, all the while communicating with them in a manner invisible to broader publics, say via Facebook ads.

**5. Experimental science in real-time environments:** The online world has opened up the doors to real-time, inexpensive and large-scale testing of the effectiveness of persuasion and political communication, a significant novelty to political campaigns. Much campaigning in the past was directed by “tacit knowledge,” “gut feel,” and deference to traditional expertise and punditry (Issenberg, 2012). Empirical discussions about politics would, at most, focus on surveys and there has been surprisingly little testing or experimentation in political campaigns (Gerber and Green, 2000).

There are a multitude of reasons for limited political experiments including the fact that campaigns are also businesses — and the ecology of consultants who make a living by running campaigns by “gut feel” were never warm to experimentation which might devalue their own strategies and hurt their bottom line (Issenberg, 2012). But most importantly, field experiments are costly and time-consuming, and money and time are the resources on which political campaigns already place the highest premium.

In spite of these obstacles, some experiments were conducted; however their results were often published too late for the election in question. The field experiments conducted in 2001, demonstrating that face-to-face canvassing was most effective for turnout, were published three years later (Green, *et al.*, 2003). These experiments increased awareness that many methods that campaigns traditionally spent money on (for example, slick mailers or phone calls) were not very effective. The cultural shift in emphasizing metrics came fully of age with the 2008 and 2012 Obama campaigns which were notable for their “data-driven culture”. Campaign manager Jim Messina declared as early as 2011 that the 2012 Obama campaign was going to be “metric driven” (Scherer, 2012). A culture of experimentation was encouraged and embraced.

This shift, however, wasn’t just a change in outlook but a change in technical infrastructure as well. The rise of digital platforms allowed incorporating real-time experimentation into the very act delivery of the political message. Sometimes called “A/B” testing, this method involves creating multiple versions of a screen or a message to be delivered separately, and delivering them to randomly selected control groups. The results are measured in real time and quickly integrated into the delivery as the winning message becomes *the* message. Methodologically, of course, this is traditional experimental science but it has become possible because campaigns now partially take place over a medium that allows for these experimental affordances: cheap delivery of messages, immediate measurement, ability to randomize recipients, and quick turnaround of results which can then be applied to the next round. Campaign operatives involved in such “metric-driven” campaigns report that their “gut feel” was often shown to be wrong by these experiments (author interviews).

The Obama campaign had incorporated experiments into its methods as early as 2007. For example, in December 2007, when the Obama campaign was still in its early stages, the campaign created 24 different button and media combinations for its splash page (the first page that visitors land on). Each variation were seen by 13,000 people — an incredibly large number for running a field experiment by old standards, but a relatively easy and cheap effort in the digital age (Siroker, 2010). In the end, the winning combination (showing a picture of his family and his children) had a 40 percent higher “sign-up rate” — translating it to the total number of people who signed up, the impact may have been as high as an extra 2,888,000 people who signed up (though, of course, that is a maximum effect). Considering that the average contribution was US\$21 and that 10 percent of people who signed up volunteered, the difference would be an additional US\$60,000,000 and 288,000 more volunteers that came through a cheap, massive and immediate experiment. Through such experimentation, the Obama campaign was led to predominantly feature his family in much campaign material.

Such an on-the-fly, complicated, “large N” true randomized experimentation has been traditionally rare in social sciences, let alone political campaigns, due to costs, efforts and ethical considerations. To consider the complications of such massive online experiments note the recent furor over Facebook and Cornell’s emotional contagion experiment (Kramer, *et al.*, 2014) in which almost 700,000 Facebook user’s newsfeeds were manipulated to see whether those seeing more “sad” posts would also post more “sad” posts — and similarly for “happy” posts — as measured by semantic analysis software. The increasing digitization of political campaigns as well as political acts by ordinary people provides a means through which political campaigns can now carry out such experiments with ease and effectiveness.

**6. Power of platforms and algorithmic governance:** Much political and civic speech occurs in the “fifth estate”, composed of blogs, micro-blogs and online social media and social networking platforms, a great many of which are privately owned corporations (Dutton, 2009; Gillespie, 2010). These platforms operate via algorithms the specifics of which are mostly opaque to people outside the small cadre of

technical professionals within the company with regards to content visibility, data sharing and many other features of political consequence.

These proprietary algorithms determine the visibility of content and can be changed at will, with enormous consequences for political speech. For example, Twitter, an emergent platform that plays a significant role in information sharing among politicians, journalists and citizens, selects and highlights 10 “trending topics” per region, which then gain visibility as they are advertised on the platform itself. The algorithm that selects topics, however, is proprietary, which has led to political actors wondering if they were censored (Lotan, 2011) while others try to “game it” by reverse engineering it (Tufekci, 2013). Similarly, non-profits that relied on Facebook to reach their audiences faced a surprise in 2013–2014. Facebook changed its algorithms, meaning that fewer and fewer people who had *chosen* to like their page were seeing status updates, going down from 20 percent in 2012 to one to two percent, with a new option that allowed for pay-to-play to promote one’s own postings (Traven, 2014).

The implications of opaque algorithms and pay-to-play are multiple: first, groups without funds to promote their content will become hidden from public view, or will experience changes to their reach that are beyond their ability to control. Second, since digital platforms can deliver messages individually — each Facebook user could see a different message tailored to her as opposed to a TV ad that necessarily goes to large audiences — the opacity of algorithms and private control of platforms alters the ability of the *public* to understand what is ostensibly a part of the public sphere, but now in a privatized manner.

These platforms also own the most valuable troves of “big data” that campaigns most desire. Campaigns can access this data either through favorable platform policies which grant them access to user information. For example, political “apps” such as those of the Romney and Obama campaigns can acquire user information as well as information on users who are friends of the original individual who accepts a campaign app, a potential privacy issue campaigns bypass. These private platforms can make it easier or harder for political campaigns to reach such user information, or may decide to package and sell data to campaigns in ways that differentially empower the campaigns, thus benefiting some over others.

Further, a biased platform could decide to use its own store of big data to model voters and to target voters of a candidate favorable to the economic or other interests of the platform owners. For example, a study published in *Nature* found that civic “go vote” messages that were targeted in Facebook through users’ social networks (thanks to a voting encouragement app deployed by Facebook) resulted in a statistically significant increase in voter turnout among those targeted, compared with a similar “go vote” message that came without such embedding in social ties (Bond, *et al.*, 2012). A platform that wanted to manipulate election results could, for example, model voters who were more likely to support a candidate it preferred and then target a preponderance of such voters with a “civic” message narrowcast so that most of the targets were in the desired target group, with just enough thrown in from other groups to make the targeting less obvious. Such a platform could help tilt an election without ever asking the voters whom they preferred (gleaning that information instead through modeling, which research shows is quite feasible) and without openly supporting any candidate. Such a program would be easy to implement, practically undetectable to observers (since each individual only sees a portion of the social media stream directed and nobody sees the totality of messages in the whole platform except the platform owners), easily deniable (since the algorithms that go into things like Facebook’s news feed are proprietary and closely guarded secrets), and practically unconfirmable [9].

A similar technique could be possible for search results. Ordinary users often never visit pages that are not highlighted on the first page of Google results and researchers already found that a slight alteration of rankings could affect an election, without voter awareness (Epstein and Robertson, 2013). Indeed, based on randomized experiments, Epstein and Robertson (2013) concluded that “with sufficient study, optimal ranking strategies could be developed that would alter voter preferences while making the ranking manipulations undetectable.” By holding on to the valuable troves of big data, and by controlling of algorithms which determine visibility, sharing and flow of political information, the Internet’s key sites and social platforms have emerged as inscrutable, but important, power brokers of networked politics.



## Consequences and power of big data analytics

Big-data driven computational politics engenders many potential consequences for politics in the networked era. In this section, I examine three aspects: deep and individualized profiling and targeting; opacity of surveillance; and, assault on (idea) of a Habermasian public sphere.

First, the shift to tailored, individualized messaging based on profiling obtained through modeling brings potential for potential significant harms to civic discourse. Howard (2006) and Hillygus and Shields (2008) had already presciently warned of the dangers of data-rich campaigns. As Howard (2006) crucially argued, the ability to eliminate unlikely or unpersuadable voters via modeling means that a strategy of focusing Presidential politics on “swing states” can be implemented at an individual level. A home judged as “non-voter” can be skipped while the next one will be flooded with campaign material, thus introducing a new form of categorical inequality into the public sphere. Previously inefficient data collection and modeling made such “redlining” difficult and confined it to precinct and state levels. Computational politics removes a “beneficial inefficiency” (Karpf, 2012) that aided the public sphere.

Messaging will also likely fracture further, encouraging further deployment of potent “wedge” issues at the expense of broadly engaged topics. Campaigns have long tried to use “wedge” issues — issues that are highly salient and important to specific segments of a voting population, such as abortion or gun rights. However, these can be double-edged for campaigns in that they elicit significant passion on all sides. Hence, campaigns aim to put wedge issues in front of sympathetic audiences while hiding them from those who might be motivated in other directions (Hillygus and Shields, 2008; Howard, 2006). Until now,



the ability to do just that has so far been limited by availability of data (finding the exact wedge voter) and means to target individuals (Barocas, 2012).

The use of “wedge issues” in direct mail is a telling example and a taste of the power of targeting capacities. Hillygus and Shields (2008) demonstrated that even by 2004, the use of “wedge issues” by campaigns was significantly more prevalent in direct mail, which is only seen by the recipient, compared with broadcast. Further, the opacity of individualized targeting through digital networks creates a new type of “dog whistle” politics, whereby the campaign emphasizes a provocative position only to sympathetic audiences, while remaining invisible to others. Prevalence of wedge issues is further damaging in that it allows campaigns to remain ambiguous on important but broadly relevant topics (economy, education) while campaigning furiously (but now also secretly) on issues that can mobilize small, but crucial, segments.

Further, the construction of “wedges” need no longer pertain merely to issues. It can also incorporate psychographic profiles modeled from online social data — data collected without directly interfacing with an individual. Hence, fear-mongering messages can be targeted only to those motivated by fear. Unlike broadcast, such messages are not visible to broad publics and thus cannot be countered, fact-checked or otherwise engaged in the shared public sphere the way a provocative or false political advertisement might have been. This form of big data-enabled computational politics is a private one. At its core, it is opposed to the idea of a civic space functioning as a public, shared commons. It continues a trend started by direct mail and profiling, but with exponentially more data, new tools and more precision.

The second negative effect derives from information asymmetry and secrecy built into this mode of computational politics. The current surveillance environment has been compared to the “panopticon”, Jeremy Bentham’s model of a prison, later used as a metaphor for modern surveillance by Foucault (1977). While the observational aspect is similar, computational politics is currently exercised in a manner opposite of the panopticon. The panopticon operates by making very visible the act and possibility of observation, while hiding actual instances of observation, so that a prisoner never knows if she is being watched but is always aware that she could be. Modern social engineering operates by making surveillance as implicit, hidden and invisible as possible, without an observed person being aware of it [10].

While browsers, cell phone companies, corporate and software companies, and, as recently revealed, the U.S. government, accumulate extensive information about individuals, the depth and the scale of the accumulated data remains opaque and inaccessible to the ordinary person. That the “guard” in such cases, *contra* Bentham, hides the fact of his observation from a prisoner flows from the fact that we are not actual prisoners but rather citizens who may be upset about surveillance and loss of privacy — and take action against it. As we are not prisoners, the model of control sought by these systems is not one of pure fear, as in George Orwell’s *1984*, but rather an infrastructure of surveillance (and targeted fear aimed at “underclass” subgroups) along with direct overtures toward obtaining assent and legitimacy through tailored, fine-tuned messaging. This model of hegemony is more in line with that proposed by Gramsci (2000) which emphasizes manufacturing consent, and obtaining legitimacy, albeit uses state and other resources in an unequal setting, rather than using force or naked coercion. Combined with voter information obtained without asking, political campaigns are moving to a paradigm of modern marketing, which aims to appear as a “pull” system, in which the user seeks and views the ad because it is enjoyable and cool, whereas the actual effort is actually a highly expensive “push” to design the most “pullable” product, without hiding the effort as much as possible.

This “hiding” of effort is well founded in human behavior. Research shows that people respond more positively to messages that they do not perceive as intentionally tailored to them, and that overt attempts are less persuasive than indirect or implicit messages. Political campaigns are acutely aware of this fact. As advisor and consultant to the Democratic party, Hal Malchow puts explicitly: “People want information, they don’t want advertising. When they see our fingerprints on this stuff, they believe it less.” (Issenberg, 2012).

A third way in which big data driven computational politics can undermine the civic experience is the destruction of “status-free” deliberation of ideas on their own merit, as idealized by Habermas (1989). Habermas’ ideal public sphere envisioned a public composed of interactions between status-free individuals where ideas were debated on their merits, regardless of who uttered them. Regardless of whether one thought this ideal existed at all, and taking into account its critics (for example, Fraser, 1990), new developments constitute an anti-Habermasian public sphere in which every interaction happens between people who are “known quantities”. The public is constituted unequally; the campaign knows a great deal about every individual while ordinary members of the public lack access to this information. In this information environment, notions of “status-free” and equal deliberation are removed from the equation.

Even when identity information is not embedded into a platform (such as Twitter where people can and do use pseudonyms), identity often cannot be escaped. Modeling can ferret out many characteristics in a probabilistic but highly reliable manner (Kossinko, 2013). Commercial databases which match computer IP to actual voter names for an overwhelming majority of voters in the United States (Campaign Grid, 2012; U.S. Federal Trade Commission 2014) are now available. Thus, political campaigns with resources can now link individual computers to actual users and their computers without the consent. Big data makes anonymity difficult to maintain, as computer scientists have shown repeatedly (Narayanan and Shmatikov, 2008). Given enough data, most profiles end up reducing to specific individuals; date of birth, gender and zip code positively correlate to nearly 90 percent of individuals in the United States.




## Discussion and conclusion

On the surface, this century has ushered in new digital technologies that brought about new opportunities for participation and collective action by citizens. Social movements around the world, ranging from the Arab uprisings to the Occupy movement in the United States (Gitlin, 2012), have made use of these new technologies to organize dissent against existing local, national and global power [11].

Such effects are real and surely they are part of the story of the rise of the Internet. However, history of most technologies shows that those with power find ways to harness the power of new technologies and turn it into a means to further their own power (Spar, 2001). From the telegraph to the radio, the initial period of disruption was followed by a period of consolidation in which challengers were incorporated into transformed power structures, and disruption gave rise to entrenchment. There are reasons to think that the Internet's trajectory may have some differences though there is little reason to think that it will escape all historical norms.

The dynamics outlined in this paper for computational politics require access to expensive proprietary databases, often controlled by private platforms, and the equipment and expertise required to effectively use this data. At a minimum, this environment favors incumbents who already have troves of data, and favors entrenched and moneyed candidates within parties, as well as the data-rich among existing parties. The trends are clear. The selling of politicians — as if they were “products” — will become more expansive and improved, if more expensive. In this light, it is not a complete coincidence that the “chief data scientist” for the Obama 2012 campaign was previously employed by a supermarket to “maximize the efficiency of sales promotions.” And while the data advantage is held, for the moment, by the Democratic party in the United States, it will likely be available to the highest bidder in future campaigns.

A recent peek into public's unease with algorithmic manipulation was afforded by the massive negative reaction to a study conducted by Facebook and Cornell which experimentally manipulated the emotional tenor of the hundreds of thousands of people's newsfeed in an effort to see if emotional contagion could occur online (Kramer, *et al.*, 2014). The authors stated their results “indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks.” While both the level of stimulus and the corresponding effect size were on the small side, the broad and negative reaction suggests that algorithmic manipulation generates discomfort exactly because it is opaque, powerful and possibly non-consensual (study authors pointed to the Facebook's terms-of-service as indication of consent) in an environment of information asymmetry.

The methods of computational politics will, and already are, also used in other spheres such as marketing, corporate campaigns, lobbying and more. The six dynamics outlined in this paper — availability of big data, shift to individual targeting, the potential and opacity of modeling, the rise of behavioral science in the service of persuasion, dynamic experimentation, and the growth of new power brokers on the Internet who control the data and algorithms — will affect many aspects of life in this century. More direct research, as well as critical and conceptual analysis, is crucial to increase both our understanding and awareness of this information environment, as well as to consider policy implications and responses. Similar to campaign finance laws, it may be that data use in elections needs regulatory oversight thanks to its effects on campaigning, governance and privacy. Starting an empirically informed, critical discussion of data politics now may be the first important step in asserting our agency with respect to big data that is generated *by* us and *about* us, but is increasingly being used *at* us. 

## About the author

Zeynep Tufekci is an assistant professor at the University of North Carolina, Chapel Hill at the School of Information and Library Science with an affiliate appointment at the Department of Sociology, a fellow at Center for Information Technology Policy at Princeton University and a faculty associate with the Berkman Center for Internet and Society at Harvard University.  
E-mail: zeynep [at] unc [dot] edu

## Notes

1. Further, 2012 also saw computational methods besides polling spread to outside of campaigns, such as that of Nate Silver's simulation models; that, however, is beyond the conceptual scope of this paper which focuses on campaigns and political actors.

2. Bernays, 1947, p. 115.

3. Horkheimer and Adorno, 2002, p. 123.

4. The political advertisement climate, and the need to advertise on broadcast, arguably, has a stronger effect in determining who can be a candidate in the first place, and not so much in selecting a winner among those who make it to that level.

5. Constine, 2012. In contrast, an online depository of books by leading large research libraries in the world contain a mere 78 terabytes of information in total (Anderson, 2008).

6. As Bryant and Raja (2014) astutely point out, this kind of analysis can be double-edged sword.

7. The number of votes that needed to flip to change outcome in the 2012 Presidential election was about a mere 400,000 distributed in the right states.

8. For example, scares about children for parents; about safety to people who've suffered from accidents; and terrorism, health, or petty crime.

9. An piece published just as this paper was about to go to press suggested a similar scenario, and called it digital gerrymandering (Zittrain, 2014).

10. While the latest NSA revelations due to leaks by Edward Snowden may change that, the level of surprise and outrage they generated speaks to both lack of awareness of surveillance as well as efforts to keep it hidden.

11. Some proponents have claimed that “big data” could predict some of these events (Leetaru, 2011) while others correctly emphasize the continuing “necessity for insight, modelling, and theorizing” (Bryant and Raja, 2014).

## References

Nate Anderson, 2008. “Universities launch elephantine 78 terabyte digital library,” *Ars Technica* (13 October), at <http://arstechnica.com/uncategorized/2008/10/universities-launch-elephantine-78-terabyte-digital-library/>, accessed 14 September 2013.

Solon Barocas, 2012. “The price of precision: Voter microtargeting and its potential harms to the democratic process,” *PLEAD '12: Proceedings of the First Edition Workshop on Politics, Elections and Data*, pp. 31–36.  
doi: <http://dx.doi.org/10.1145/2389661.2389671>, accessed 25 June 2014.

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler, 2012. “A 61-million-person experiment in social influence and political mobilization,” *Nature*, volume 489, number 7415 (13 September), pp. 295–298.  
doi: <http://dx.doi.org/10.1038/nature11421>, accessed 25 June 2014.

danah boyd and Kate Crawford, 2012. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Information, Communication & Society*, volume 15, number 5, pp. 662–679.  
doi: <http://dx.doi.org/10.1080/1369118X.2012.678878>, accessed 25 June 2014.

Edward L. Bernays, 1947. “The engineering of consent,” *ANNALS of the American Academy of Political and Social Science*, volume 250, number 1, pp. 113–120.  
doi: <http://dx.doi.org/10.1177/000271624725000116>, accessed 25 June 2014.

Bruce Bimber, 2014. “Digital media in the Obama Campaigns of 2008 and 2012: Adaptation to the personalized political communication environment,” *Journal of Information Technology & Politics*, volume 11, number 2, pp. 130–150.  
doi: <http://dx.doi.org/10.1080/19331681.2014.895691>, accessed 25 June 2014.

Anthony Bryant and Uzma Raja, 2014. “In the realm of Big Data ...,” *First Monday*, volume 19, number 2, at <http://firstmonday.org/article/view/4991/3822>, accessed 25 June 2014.  
doi: <http://dx.doi.org/10.5210/fm.v19i2.4991>, accessed 25 June 2014.

Erik Brynjolfsson and Andrew McAfee, 2011. “The big data boom is the innovation story of our time,” *Atlantic* (21 November), at <http://www.theatlantic.com/business/archive/2011/11/the-big-data-boom-is-the-innovation-story-of-our-time/248215/>, accessed 25 June 2014.

Campaign Grid, 2012, at <http://www.campaigngrid.com/>, accessed 25 June 2014.

Victoria Carty, 2011. *Wired and mobilizing: Social movements, new technology, and electoral politics*. New York: Routledge.

Josh Constine, 2012. “How big is Facebook’s data? 2.5 billion pieces of content and 500+ terabytes ingested every day,” *TechCrunch* (22 August), at <http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/>, accessed 25 June 2014.

Aron Culotta, 2014. “Estimating county health statistics with Twitter,” *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1,335–1,344.  
doi: <http://dx.doi.org/10.1145/2556288.2557139>, accessed 25 June 2014.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz, 2013. “Predicting depression via social media,” *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, at <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124>, accessed 25 June 2014.

John Dewey, 1927. *The public and its problems*. New York: H. Holt and Co.

Charles Duhigg, 2012. “How companies learn your secrets,” *New York Times* (16 February), at <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>, accessed 25 June 2014.

William H. Dutton, 2009. “The fifth estate emerging through the network of networks,” *Prometheus: Critical Studies in Innovation*, volume 27, number 1, pp. 1–15.  
doi: <http://dx.doi.org/10.1080/08109020802657453>, accessed 25 June 2014.

Robert Epstein and Ronald E. Robertson, 2013. “Democracy at risk: Manipulating search rankings can shift voters’ preferences substantially without their awareness,” paper presented at the 25th annual meeting of the Association for Psychological Science; version at [http://www.fraw.org.uk/files/politics/epstein\\_robertson\\_2013.pdf](http://www.fraw.org.uk/files/politics/epstein_robertson_2013.pdf), accessed 25 June 2014.

Michel Foucault, 1977. *Discipline and punish: The birth of the prison*. Translated from the French by Alan Sheridan. New York: Pantheon Books.

Nancy Fraser, 1990. "Rethinking the public sphere: A contribution to the critique of actually existing democracy," *Social Text*, numbers 25/26, pp. 56–80.  
doi: <http://dx.doi.org/10.2307/466240>, accessed 25 June 2014.

Alan S. Gerber and Donald P. Green, 2000. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment," *American Political Science Review*, volume 94, number 3, pp. 653–663.

Tarleton Gillespie, 2010. "The politics of 'platforms'," *New Media & Society*, volume 12, number 3, pp. 347–364.  
doi: <http://dx.doi.org/10.1177/1461444809342738>, accessed 25 June 2014.

Todd Gitlin, 2012. *Occupy nation: The roots, the spirit, and the promise of Occupy Wall Street*. New York: itbooks.

Donald P. Green, Alan S. Gerber, and David W. Nickerson, 2003. "Getting out the vote in local elections: Results from six door-to-door canvassing experiments," *Journal of Politics*, volume 65, number 4, pp. 1,083–1,096.  
doi: <http://dx.doi.org/10.1111/1468-2508.t01-1-00126>, accessed 25 June 2014.

Jürgen Habermas, 1989. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Translated by Thomas Burger with the assistance of Frederick Lawrence. Cambridge, Mass.: MIT Press.

D. Sunshine Hillygus and Todd G. Shields, 2008. *The persuadable voter: Wedge issues in Presidential campaigns*. Princeton, N.J.: Princeton University Press.

Max Horkheimer and Theodor W. Adorno, 2002. "Enlightenment as mass deception," In: Max Horkheimer and Theodor W. Adorno. *Dialectic of enlightenment: Philosophical fragments*. Edited by Gunzelin Schmid Noerr; translated by Edmund Jephcott. Stanford, Calif.: Stanford University Press, pp. 94–136.

Philip N. Howard, 2006. *New media campaigns and the managed citizen*. New York: Cambridge University Press.

Philip N. Howard and Malcolm R. Parks, 2012. "Social media and political change: Capacity, constraint, and consequence," *Journal of Communication*, volume 62, number 2, pp. 359–362.  
doi: <http://dx.doi.org/10.1111/j.1460-2466.2012.01626.x>, accessed 25 June 2014.

Philip N. Howard and Daniel Kreiss, 2010. "Political parties and voter privacy: Australia, Canada, the United Kingdom, and United States in comparative perspective," *First Monday*, volume 15, number 12, at <http://firstmonday.org/article/view/2975/2627>, accessed 14 September 2013.  
doi: <http://dx.doi.org/10.5210/fm.v15i12.2975>, accessed 25 June 2014.

Sasha Issenberg, 2012. *The victory lab: The secret science of winning campaigns*. New York: Crown.

David Karpf, 2012. *The MoveOn effect: The unexpected transformation of American political advocacy*. New York: Oxford University Press.

Michał 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.

Adam D.I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, 2014. "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, volume 111, number 24 (17 June), pp. 8,788–8,790.  
doi: <http://dx.doi.org/10.1073/pnas.1320040111>, accessed 25 June 2014.

Daniel Kreiss, 2012a. *Taking our country back: The crafting of networked politics from Howard Dean to Barack Obama*. New York: Oxford University Press.

Daniel Kreiss, 2012b. "Yes we can (profile you): A brief primer on campaigns and political data," *Stanford Law Review*, volume 64, at <http://www.stanfordlawreview.org/online/privacy-paradox/political-data>, accessed 25 June 2014.

Daniel Kreiss and Philip N. Howard, 2010. "New challenges to political privacy: Lessons from the first U.S. Presidential race in the Web 2.0 era," *International Journal of Communication*, volume 4, pp. 1,032–1,050, and at <http://ijoc.org/index.php/ijoc/article/view/870/473>, accessed 25 June 2014.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne, 2009. "Computational social science," *Science*, volume 323, number 5915, pp. 721–723.  
doi: <http://dx.doi.org/10.1126/science.1167742>, accessed 25 June 2014.

Kalev Leetaru, 2011. "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space," *First Monday*, volume 16, number 9, at <http://firstmonday.org/article/view/3663/3040>, accessed 25 June 2014.



Walter Lippmann, 1925. *The phantom public*. New York: Harcourt, Brace.

Steve Lohr, 2012. "The age of big data," *New York Times* (11 February), at <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>, accessed 25 June 2014.

Joe McGinniss, 1988. *The selling of the President*. New York: Penguin.

Rasmus Kleis Nielsen, 2012. *Ground wars: Personalized communication in political campaigns*. Princeton, N.J.: Princeton University Press.

Arvind Narayanan and Vitaly Shmatikov, 2008. "Robust de-anonymization of large sparse datasets," *SP 2008: IEEE Symposium on Security and Privacy*, pp. 111–125. doi: <http://dx.doi.org/10.1109/SP.2008.33>, accessed 25 June 2014.

Plato, 2012. *Plato's Republic: A dialogue in 16 chapters*. Translated by Susan Spitzer. New York: Columbia University Press.

Daniele Quercia, 2013. "Don't worry, be happy: The geography of happiness on Facebook," *WebSci '13: Proceedings of the Fifth Annual ACM Web Science Conference*, pp. 316–325. doi: <http://dx.doi.org/10.1145/2464464.2464484>, accessed 25 June 2014.

Michael Scherer, 2012. "Inside the secret world of the data crunchers," *Time* (7 November), at <http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/>, accessed 25 June 2014.

Debora L. Spar, 2001. *Ruling the waves: Cycles of discovery, chaos, and wealth from the compass to the Internet*. New York: Harcourt.

Clay Shirky, 2008. *Here comes everybody: The power of organizing without organizations*. Penguin Press.

Dan Siroker, 2010. "How Obama raised \$60 million by running a simple experiment," *Optimizely* (29 November), at <http://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/>, accessed 25 June 2014.

B. Traven, 2014. "Facebook Is throttling nonprofits and activists," *Valleywag* (30 April), at <http://valleywag.gawker.com/facebook-is-throttling-nonprofits-and-activists-1569877170>, accessed 25 June 2014.

Zeynep Tufekci, 2013. "'Not this one' Social movements, the attention economy, and microcelebrity networked activism," *American Behavioral Scientist*, volume 57, number 7, pp. 848–870. doi: <http://dx.doi.org/10.1177/0002764213479369>, accessed 25 June 2014.

Zeynep Tufekci, 2012. "Beware the smart campaign," *New York Times* (17 November) p. A23, and at <http://www.nytimes.com/2012/11/17/opinion/beware-the-big-data-campaign.html>, accessed 25 June 2014.

Zeynep Tufekci and Deen Freelon, 2013. "Introduction to the special issue on new media and social unrest," *American Behavioral Scientist*, volume 57, number 7, pp. 843–847. doi: <http://dx.doi.org/10.1177/0002764213479376>, accessed 25 June 2014.

U.S. Federal Trade Commission, 2014. *Data brokers: A call for transparency and accountability*. Washington, D.C.: U.S. Federal Trade Commission, at <http://permanent.access.gpo.gov/gpo49352/140527databrokerreport.pdf>, accessed 25 June 2014.

Ben Yagoda, 2013. "The comic stylings of POTUS," *Chronicle of Higher Education* (6 May), at <http://chronicle.com/blogs/linguafranca/2013/05/06/the-comic-stylings-of-potus/>, accessed 25 June 2014.

Jonathan Zittrain, 2014. "Facebook could decide an election without anyone ever finding out: The scary future of digital gerrymandering — and how to prevent it," *New Republic* (1 June), at <http://www.newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering>, accessed 25 June 2014.

---

## Editorial history

Received 23 October 2013; revised 11 June 2014; revised June 29 2014; accepted 30 June 2014.

---

Copyright © 2014, *First Monday*.

Copyright © 2014, Zeynep Tufekci.

Engineering the public: Big data, surveillance and computational politics  
by Zeynep Tufekci

*First Monday*, Volume 19, Number 7 - 7 July 2014

<http://firstmonday.org/ojs/index.php/fm/rt/prINTERfriendly/4901/4097>

doi: <http://dx.doi.org/10.5210/fm.v19i7.4901>