Originally published at *Ethnography Matters*
https://ethnographymatters.net/blog/2017/05/24/persuasion-and-the-other-thing-a-critique-of-big-data-methodologies-in-politics/
May 24, 2017

Persuasion and the other thing: A critique of big data methodologies in politics

Earlier this year, a company called Cambridge Analytica shot to the forefront of the debate over big data and elections when it claimed responsibility for the upset victories of both Donald Trump and the Brexit Campaign. Reports have cast the firm as a puppet master "propaganda machine" able to mint voters through a proprietary blend of psychometric data, primarily Facebook "likes" and targeted nudges. In this story, repeated by Mother Jones and The Guardian among others, Cambridge Analytica [working in conjunction with an "election management" firm called SCL Group] is both a king maker and a Pied Piper: voters are unable to resist attempts at political manipulation, as they are seamlessly integrated with voters' online environment and pulled by strings too deeply anchored in voters' psyches to be ignored.

I'm uninterested in the actual snake content of Cambridge Analytica's snake oil. As noted by the MIT Technology Review and BuzzFeed, the company has made some big claims and has been happy to take credit for several of 2016's startling electoral results. But Cambridge Analytica relies heavily on the techno-magic of under-described big data psychographics and algorithmic nudging. Both the Tech Review and BuzzFeed point out that the amount and types of data that the company appears to use are not much different than types of data acquisition and analysis already commonly in use.

Instead I'm interested in the ways that Cambridge Analytica's sales pitch reflects how the subjects of these big data analytics projects are viewed by those conducting the research, and the entitlements held by advertisers, tech firms, and researchers who deploy big data analytics in support of political campaigns or other political projects. This sense of entitlement matters. I'd like to posit that the use of "big data" in politics strips its targets of subjectivity, turning individuals into ready-to-read "data objects," and making it easier for those in positions of power to justify aggressive manipulation and invasive inference. I would like to further suggest that when big data methodology is used in the public sphere, it is reasonable for these "data objects" to, in turn, use tactics like obfuscation, up to the point of actively sabotaging the efficacy of the methodology in general, to resist attempts to be read, known, and manipulated.

Cambridge Analytica's willingness to throw its brand behind causes some might consider to be cacklingly evil, along with well-publicized incidents like the Facebook "emotional contagion" experiment in 2014, have dramatized these issues for the general public, but many researchers, particularly Zeynep Tufekci, have been sounding the alarm for big data's specific methodological implications in the political arena for quite some time. In 2014, Tufekci described the information asymmetry problem of big data methodology, that it is not merely that subjects do not know as much about the researchers as the researchers know about them, but that as a core aspect of the methodology, subjects often do not even know they are being studied. While previous models of data collection allowed for the modeling of rough populations, Tufekci notes that big data analytics allow for the modeling of individuals without the researcher ever having to encounter that individual, or the individual being aware their actions are being taken into the political sphere.

It is reasonable for individuals to use tactics like obfuscation to resist attempts to be read, known, and manipulated as data objects.

By avoiding any responsibility to encounter research subjects in their own contexts, researchers are free to imagine that the individuals in their datasets, their data objects, neatly align with the researcher's pre-made analytic categories, and further to imagine that these categories describe the whole of that individual. This is a familiar problem with large-N study methodologies, but in the political sphere, let us substitute "constituents" for research subjects, and "elected officials" for researchers. While, as Tufekci notes, polling and rough inferential population modeling have long been part of the political sphere, the appeal of big data modeling is its purported ability to specifically model individuals with high degrees of reflective and predictive accuracy. The rhetoric of big data methodologies as deployed by Cambridge Analytica and others provides the mathematical, methodological justification for political campaigns and governments to ignore constituents in favor of data models of those constituents.

The rhetoric of big data is the echo overtaking the voice, the map overtaking the territory.

This is the "data doppelganger" (a term coined by critic Sara Marie Watson) overtaking the individual who is ostensibly its source, the echo overtaking the voice, the map overtaking the territory. In so much as these data doppels are used to directly impact, direct, and influence the lives of those individuals from whose actions they are derived, the sense in which their knowability is both assumed and constructed solely from building blocks provided by (powerful) others and rendered machine readable, they have a potential diminishing effect on these individuals' subjectivity and agency. The researcher (or the advertiser or campaign manager) is no longer dealing with a person possessed of their own self-determining agency and unmeasurable subjectivity, but rather manipulating a fully comprehensible data object.

Cathy O'Neil's book *Weapons of Math Destruction* starkly describes the ways in which big data methodologies are used to model and then influence individual lives, either through slight nudges like the slow drip of targeted advertisements, or the more forceful shove of not getting a job or failing to qualify for a loan. O'Neil has ably highlighted the dangers of substituting suites of machine-readable behaviors or characteristics for actual encounters with people as they are, calling these types of algorithmic modeling "opinions embedded in mathematics" (p. 21). Mathematizing subjective knowledge can make it appear objective, in this context creating the impression that algorithmically modeling a person is a useful or beneficial or even superior way to know them.

This model of grasping another person solely through pre-set categorizations and machine-readable actions means never being forced to encounter difference. Pre-established and machine-readable categories and actions are fundamentally aspects which are already familiar: they are recognized as important by the person collecting the data, hence, almost tautologically, their inclusion. But difference that breaches the bounds of the dataset becomes invisible. Kelly Oliver discusses the limitations of this "recognition" based model:

> *Any real contact with difference or otherness becomes impossible because recognition requires the assimilation of difference into something familiar… Only when we begin to think of the recognition of what is beyond recognition can we begin to think of the recognition of difference. (p. 9)*

The assumption of knowability, that a person can be grasped with mathematical completeness through their digital shadow-selves, is coupled with a paradoxical problem, a certain entitlement of inference.

Tufekci has provided a basic description of this problem in the quote above, "such modeling allows for acquiring answers about an individual without directly asking questions to the individual…" Tufekci states she is concerned primarily with the opaque deployment of influence or nudge techniques in the political sphere, and obviously these effects concern me as well. I am additionally concerned with the sense of entitlement required to infer personal data that may have been intentionally withheld by users for whom that ability, the ability to not disclose, may be one of few true privacy protections available.

Tufekci highlights a 2013 paper published by Michal Kosinski, David Stillwell, and Thore Graepel, entitled, "Private traits and attributes are predictable from digital records of human behavior." Kosinski and his collaborators use Facebook "like" data to model individuals, predicting "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.*

Tufekci further notes that these traits are being inferred though available data and modeling algorithms, not "asked or observed from the user," pointing out that this type of modeling could be deployed in spaces where anonymous or pseudonymous behavior is common. The combination of separation from the user-subject, coupled with the assumption that that same user-subject is fully graspable from the standpoint of the researcher, results in the collapse of privacy rights for that user-subject in the face of the right to know on the part of the researcher, the entitlement of inference on display in the Kosinski project and in so many others, commercial, political, and academic.

It is a short hop from thinking you know someone to thinking you know what they want or what is good for them, without any need to persuade or even to ask. And removing persuasion as a necessary step from the political sphere removes consent from the political sphere as well.

There are two risks to the deployment of big data methodologies in the political sphere: the first has been repeatedly articulated by Tufekci and others, that big data methodologies will allow secret or opaque influence techniques to be unleashed upon the electorate, creating a storm of personally tailored propaganda that blends seamlessly into a user's media feed. Tufekci notes that this type of "privatized" targeting allows political campaigns to play directly on the fears and reactionary impulses of a certain set of voters, or to make promises without revealing them to others with whom such promises might backfire. This creates a national political fabric not of broad communities of multiple points of address and compromise to be governed holistically, but of schism-ed individuals and groups, each believing that they are the whole of the community that needs to be addressed, and anyone else is an interloper. Democracy shifts from a form of governance at least theoretically concerned with public debate and persuasion to one focused on private, opaque manipulation and emotional coercion.

The second risk occurs when politicians and governments, stuffed with psychographic data and algorithmic models, no longer feel the need to encounter the governed at all.

If we believe that widespread social-media-based big data modeling poses genuine risks to democracy, what is the best way to mitigate these risks?

Both of these situations remove the consent of the governed from the political sphere. The invisible observation methods intrinsic to social-media-based big data preclude meaningful consent, as does inferential modeling intended to collect non-disclosed information. And while elections still, in an imperfect fashion, allow individuals a voice in their government, most of the business of modern representative democracies takes place in the times in between elections. Already alienated from their vote through gerrymandering, corporate lobbying, and the failures of campaign finance reform, could voters be pushed further from their elected representatives by their data doppelgangers? Might noisy town halls, which require elected representatives to travel, which are vulnerable to disruptive in-person demonstrations, and which many Republicans have taken to actively (and comically) ducking since January, be replaced with Cambridge Analytica/Kosinski-style silent, acquiescent constituent models?

If we acknowledge that the risks to democracy posed by widespread social-media-based big data modeling are genuine, what is the best way to mitigate these risks? Short of establishing that, as a matter of political ethics, this type of constituent modeling is unethical and anti-democratic, or convincing social media firms to not sell "likes" or other psychographic data, what forms of resistance might be deployed at the individual or local level?

Opting out, or social media abstinence immediately jumps to mind. If you do not wish to be modeled or tracked, simply do not participate in those systems which expose you to these tracking and modeling algorithms. However, this strategy is ineffective on a number of levels. First, it only protects those who are able to opt out from these tracking and modeling systems; given the central role social media plays in many people's social and professional lives, opting out is simply not a viable option for everyone. Second, regarding the specific issue of constituent modeling, opting out at an individual level would remove even the shadow-representation offered by the data doppelganger. As long as enough people participate in the systems that permit this type of modeling, those who opt out will be simply not represented, and in all likelihood, not missed.

Obfuscation, as described by Finn Brunton and Helen Nissenbaum in their 2015 book of the same name, may be the best mode of resistance to pervasive surveillance and modeling systems that are unlikely to be rejected by those in power (or those who seek power) due to perceptions of their efficacy and profitability. By utilizing obfuscatory methods, Cambridge Analytica-style systems of constituent profiling and manipulation can be render ineffective for the targeted population as a whole, discouraging their use.

Brunton and Nissenbaum highlight several different examples of obfuscations that could be deployed to render this type of constituent modeling less effective. Some of these obfuscatory methods also show ways constituent modeling could be easily gamed by those who wish to influence polling or a politician's perception of their constituents, which further underlines the democratic dangers posed by encouraging a separation between the people and their representatives.

Eventually, attempting to find your "true" data stream among your mob of data clones would be like trying to find a needle in a haystack of other needles.

Several of these methods create noise, either at the level of the platform or the individual profile.

- Platform level noise generation might look like stacking a big data channel, like Twitter or Facebook, with noisy bots that share just enough characteristics with the targeted dataset to be included.
- Another method, "like-farming," involves paying individuals to "like" products or brands on Facebook, often thousands at a time. This behavior could devalue "likes" as psychographic data.
- At the individual obfuscation level, Brunton and Nissenbaum note several add-ons or experiments that operate on the logic of their TrackMeNot browser extension, which obfuscates an individual's genuine search history by generating a background hum of "fake" search requests for every "real" one.
- AdNauseam works in the background of your web browser, invisibly clicking every ad on every page you visit. This activity floods ad tracking networks with useless and inaccurate data, and also allows those websites you visit to collect revenue from the pay-per-click ads they feature.
- FaceCloak creates a network within the Facebook network, allowing users to store personal data with FaceCloak instead of Facebook. Users of the FaceCloak add-on can see your personal data as integrated with your Facebook page, but Facebook never possesses it.
- Brunton and Nissenbaum also discuss "Bayesian flooding," which involves individuals actively feeding false information into their Facebook profiles: "The trick is to populate your Facebook with just enough lies as to destroy the value and compromise Facebook's ability to sell you" (Cho quoted in Brunton + Nissenbaum, p. 39).

- One tactic that entails both individual and platform obfuscation is a patent held, interestingly enough, by Apple, entitled "Techniques to Pollute Electronic Profiling." Brunton and Nissenbaum describe it as a "cloning service," intended to "automate and augment the process of producing misleading personal information, targeting online data collectors." This cloning service would mimic a user's personal rhythms and behaviors, but "may begin to diverge from those interests in a gradual, incremental way," (p. 36) automatically browsing, clicking, signing up for websites and newsletters, chatting with other clones, maybe ordering small physical items from time to time. Eventually, in theory, attempting to find your "true" data stream among your mob of data clones would be like trying to find a needle in a haystack of other needles.

For Brunton and Nissenbaum, obfuscation is of particular utility in cases of information and power asymmetries, which, as noted above, are core issues with big data analytics and inferential modeling. Regarding inevitable accusations of "data pollution" or damage, Brunton and Nissenbaum conclude

[I]n order for a charge of data pollution to stick, a data assemblage must be shown to hold greater value than whatever the obfuscator aims to protect….Data pollution is unethical only when the integrity of the data flow or data set in question is ethically required. Moreover whether the integrity of the data outweighs other values and interests at stake must be explicitly settled. (p. 69)

In political deployments of big data analytics and inferential modeling, what is at stake is the ability of the powerful to see and meaningfully engage with a consenting electorate. A claim could be made that if current trends continue, and big data psychographic methodologies become a primary means of electioneering and governance, actively attempting to reduce the effectiveness of that dataset would be an unethical move. But there are other ways for governments and campaigns to encounter their electorates, to figure out how best to represent their constituents. A town hall may have the potential to be messy, loud, and unpredictable, but it allows an encounter between the people and their elected representatives. Referendums on specific issues, given to the public for a direct vote, can't reveal private information: your vote on medical marijuana legalization won't reveal if you're gay.

As a modern democracy, the US excels in developing new mechanisms to distance the individual from the power of their vote. Big data methodologies and the inferential analytics they power as deployed in elections present yet another move to push people, in all their loud, messy, demanding changeability, out of politics. But unlike gerrymandering or the electoral college, this move can be actively resisted on the individual level.