Not a Mirror, but an Engine: Digital Methods for Contextual Analysis of "Social Big Data"

Jonas Andersson Schwarz

Introduction

Unlike several of the subsequent chapters in this anthology, I will not focus on audiovisual media, nor new media objects such as Instagram posts (Pennlert et al.), podcasts (Johansson), data visualizations (Uggla) or web art (Wasielewski) for that matter.[1] The medium that I will write about is a much more familiar one for humanists and social scientists: plaintext.

Like Wasielewski, I too would note that the methods for computational text analysis are in many ways less complicated than those for digital image analysis, for example, and that there are several already-established methodological pathways for text-based corpus archiving, distribution, retrieval, and analysis. But the research projects that I will refer to are very novel, however, in that their research objects are what other authors in this volume refer to as "born digital": The initial publication and distribution of the text data in question was made by native internet users in online social media forums. Moreover, as soon as the researcher uses computers to access and select samples from such forums, and then computationally process and analyze his or her findings, we are dealing with methods that are in many ways *very* different from predigital methods for text analysis. In other words, relating

[1] For all of these authors, see this volume.

How to cite this book chapter:
Andersson Schwarz, Jonas. "Not a Mirror, but an Engine: Digital Methods for Contextual Analysis of 'Social Big Data.'" In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 23–47. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.b. License: CC-BY.

back to the editors' introduction,[2] my own interest in "the digital" is equally *as a tool* and *as an object* of research.

As a scholar of media and communications, my own frame of analysis borrows from traditional sociology and public opinion research, where the question of *representativity* is always of high interest, and medium theory, where the question of the *ontological properties* of particular media forms is similarly important—in this case, the properties of online, user-generated and user-distributed text. I am interested in using digital corpora to be able to say something about "the public," but I am equally interested in noting what possibilities and limitations are afforded by online-mediated social discourse—an artifact that is digital from its very inception, as users type their affective expressions in interfaces that come with certain preconditions to begin with.

Initial Definitions

This chapter aims to discuss the epistemology and normativity of data produced and shared on social media platforms, and the attendant challenges for research: problems of access and representativity, and the need for contextualization and, consequentially, for combinatory methods. My work is situated at the intersection of social science and humanities, in the sense that I have experience from engaging in highly empirical endeavors aiming to capture digital and communicational mechanisms of contemporary society, while at the same time agonizing over the issues of social philosophy that arise from such endeavors, including the abovementioned epistemic and normative tangles.[3] Lastly, I will close the chapter by relating back to the initial empirical and epistemological challenges, pointing out some key pitfalls concerning the validity, reliability, and representativity of such data, and some steps toward improving contextualization, especially regarding semantic text data.

[2] This volume.

[3] Due to brevity, I will not focus legal aspects, for example those that address consent or copyright, although these are highly important aspects for the generation and collection of this type of data. It is crucial to maintain an acquaintance with the various administrative and governmental applications of data analysis, since many are both politically and commercially expedient.

What do I mean by *epistemology*? Reflecting upon two recent research projects that I have been part of,[4] this chapter is a meditation on the multifaceted methodological and epistemological challenges that mount when researchers face contemporary social media platform architectures as research objects and data sources. Like other critical theorists of the so-called "big data era," I would not declare that the new data landscape heralds an "end of theory." On the contrary, one needs to contemplate the epistemological implications of the very "data revolution" at hand.[5] While new approaches to data generation, collection, and analyses are enabled that make it possible to ask and answer questions in new ways, it is clear that much of the usefulness of relational, real-time datasets has to be complemented by more established, conventional research methods. Dominique Boullier has shown how the novel methods that contemporary data structures enable—for example, tracing trends and inferring relational patterns in (near) real time—often need to be put in context,[6] for example by being complemented by more established methods of assessing representativity (late 20th-century quantitative social science) and validity (qualitative assessment). Moreover, methods and instruments tend to be suffused with epistemological implications, and so too are the research sites and objects to begin with. Data always has prerequisites for its generation in the first place, even putting to one side the politics of access to it: its alleged validity, reliability, salience, and indeed importance; the various idealistic claims that are made for it; and, lastly, the naturalization throughout key societal sectors (academic, administrative, business, policy) of such viewpoints. Data tends to have a normative function, in that it represents the social relations that it registers as valid in and of themselves. If we are to make a comprehensive overview of

[4] Jonas Andersson Schwarz et al., *Opinioner och offentligheter online* (Stockholm: IIS, 2015); Stefan Dahlberg, *Linguistic Explorations of Society* (Swedish Research Council, 2017).

[5] Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* 1, no. 2 (April–June 2014): 1.

[6] Dominique Boullier, "Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications," in *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*, eds. Franck Cochoy et al. (London, New York: Routledge, 2017).

challenges, we find that they are indeed philosophical, psychological, cybernetic, *and* political-economic. Formalized methods of data analysis tend to inductively probe data that is abundant with internal relations in order to identify latent structures and patterns in such data. I will take natural language processing (NLP) as an example, but there are many more machine learning (ML) approaches, and methods for formalized social network analysis (SNA). Researchers could utilize combinations of these novel methods, in addition to more traditional content analysis (CA) where samples are taken and manually decoded.

What is *data* in this particular context? The research object I have in mind is *user-generated plaintext generated on web-based internet platforms*.[7] As data sources, large volumes of plaintext are suitable for computer-aided digital human sciences tools such as "information retrieval, text analytics, data mining, visualization, and geographic information systems."[8] More specific methods—like the NLP that I will discuss in this chapter—lend themselves to *large quantities* of plaintext, where corpus-based, statistical approaches can be employed. Here, the general item of analysis would be morphemes (minimal meaningful word segments), where NLP can, for example, successfully parse the syntactical position of individual morphemes, given their position relative to other morphemes in the corpus.[9] More specifically, the models hinge upon the so-called *type-token distinction*.[10] There are functional differences between identifying a *class* (type) of objects and naming the individual *instances* (tokens) of that class, and computers are used to statistically identify such types and tokens.

[7] These could be highly publicly known platforms like Twitter, but what is more specifically implied here are message boards and pages on the common web, since these are more easily accessible for scraping than many proprietary platforms, such as those owned by Facebook.

[8] Michael Piotrowski, *Natural Language Processing for Historical Texts* (London, Williston, VT: Morgan & Claypool, 2012), 6.

[9] Lenhart Schubert, "Computational Linguistics," in *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.), ed. Edward N. Zalta (Stanford, CA: Stanford University, 2020).

[10] Linda Wetzel, "Types and Tokens," in *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.), ed. Edward N. Zalta (Stanford, CA: Stanford University, 2018).

Tokenization is the task of chopping up a document into constituent pieces.[11] By using so-called *fuzzy search* heuristics, misspelled and ambiguous tokens can also be typified together with native, unambiguous tokens. In exploratory data analysis, token occurrences in an unknown language can be explored, in order to improve language models by creating particular "watermarks" of the ways in which word embeddings are distributed in different languages. *Word embeddings* is a way of statistically representing the ways in which words have similar meanings, given their contexts in specific languages. Linguist John Firth expressed it in an often-invoked quip: "You shall know a word by the company it keeps!"[12] As Piotrowski has pointed out, computational linguistics should be considered an "auxiliary science" to digital human sciences, which can aid researchers with *formal modeling* of scholarly knowledge and insights in machine-processable form.[13] There are numerous various approaches in this rapidly developing field, and the practical applications are many: machine translation, document retrieval, knowledge extraction (by way, e.g., of recognition of patterns and/or clusters), sentiment analysis, natural language user interfaces, and so on.

More specifically, what do I mean by *social* data? For the purposes of this chapter, I will focus on the societal contingencies of found web-mediated texts, as those texts are produced by social agents, for specific purposes, and mediated by specific technologies with specific affordances. Despite their profusion, the texts collected should not be expected to mirror the social world in a 1:1 fashion. Rather, they are produced under specific circumstances, having specific properties. Hence, I argue that the mathematical rigor and competence of the NLP scientist have to be combined with a sociological sensibility in order not to adjust the inferences

[11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval* (Online ed.) (Cambridge: Cambridge University Press, 2009), 22.

[12] John Firth, "A Synopsis of Linguistic Theory 1930–1955," in *Studies in Linguistic Analysis 1930–1955*, ed. John Firth (Oxford: Basil Blackwell, 1962), 11.

[13] Michael Piotrowski, "Digital Humanities, Computational Linguistics, and Natural Language Processing," presentation, *Lectures on Language Technology and History*, Uppsala University (March 4, 2016), 21.

made and put the texts into societal context. As *research objects*, digital media platforms are "moving targets" in the sense that the sociocultural enactments taking place are ever-changing, making replication very hard. If researchers are to allow future researchers to replicate their analyses, they either have to duplicate the data used and be able to hand this to adjacent researchers in some way or form, or provide clear step-by-step heuristics as regards the data collection,[14] with the proviso that, despite following these very same steps, future data harvests might look entirely different. Moreover, while the platforms in question could be seen as "material substrates" to the sociocultural enactments taking place "on top of them," as it were, it is important to understand them, more aptly, as *enablers of agency* that structure, delimit, and harness social action. The platforms are, in other words, *firm ground* that enable and record social action and, at the same time, *structuring agents* that contain certain affordances and, thus, orchestrate user agency. It is advisable to seek out literature on political-economic and material platform architectures and business models, since this literature can help the analyst identify features that might be critical for understanding what users can and cannot do ("hard" governance; code-as-law), and the various ways in which they are expected to act ("soft" governance; user norms).

What is the *bigness* of "big data"? As I have argued elsewhere,[15] the big data signifier should be understood epistemologically rather than out of mere quantitative assessment. The datasets involved *need* not be terabyte-sized; they can actually be much smaller, yet still fit under the header of big data methodology, if the methods employed address and make available the *multidimensional relationality* of the data involved. In SNA, for example, the set of social relations is always *at least* two-dimensional[16] and is usually represented in the form of a *node table* (listing all the nodes

[14] There are, e.g., open science projects like CommonCrawl that routinely scrape the publicly available web, making available enormous quantities of multilingual corpus text.

[15] Göran Bolin and Jonas Andersson Schwarz, "Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation," *Big Data & Society* 2, no. 2 (December 2015): 1–12.

[16] John Scott, "Networks & Relations," in *Social Network Analysis*, ed. John Scott (London: Sage, 2013), 1–9.

involved) and an *edge table* (listing the relations between them). The operational term in this chapter will therefore be "social big data"—as a term for the forms of data employed *and* its attendant methods and approaches. Arguably, in this case, the two are hard to separate; the data generated becomes generated through analytical operations. Method and data co-constitute each other, as it were. Ultimately, there is an epistemological case to be made for seeing "data" as not only a research object alone but also a *research tool*. I would like to stress the *boundedness* and *ordinariness* of algorithmic infrastructures—that is, their capacity as bureaucratic systems, and co-constitutive of a resurgence of the technocratic "administered society"[17] as a governmental trope. I argue that methodological reductionism should be understood as a prime cause for this resurgent technocracy. When institutional actors understand social reality through a lens of "data idealism," the resultant policies are likely to be mechanistic in nature, as many other scholars have pointed out.[18] In order not to simplify the world views on which important decisions are taken, empirical and theoretical rigor begs a constructionist understanding of social big data. As Evgeny Morozov once pointed out, data-driven companies like Google operate less as mirrors of social reality and more as engines of social change.[19]

Social Big Data: Perspectives

Given the material-semiotic concerns noted above, I will now list a range of particular contexts that are constituent of the generation of social big data. I will begin by noting the important distinction between numerical (quantitative) data, often accrued through binary sensors, and semantic (qualitative) data, comprised of meaning-bearing units.

First, from a *semiotic* perspective, the theory of Charles Sanders Peirce would imply that what most digital sensor signals do is

[17] Herbert Marcuse, *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society* (Boston, MA: Beacon Press, 1964), 9–20.

[18] Notable names are Mireille Hildebrandt, Antoinette Rouvroy, Erich Hörl, Nick Couldry, Evgeny Morozov, John Cheney-Lippold, Ted Striphas.

[19] Evgeny Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism* (New York: PublicAffairs, 2013), 145.

to *index* physical reality—as smoke to a fire, a pedometer tick would index a walking step. Similarly, a ticker would count binary "likes" in social media.[20] While this might appear straightforward, indications are in actual real life rarely that straightforward or clear. In practice, a single sensor can indicate many things. This suggests that one should consider data epistemology from the ground up: *signals* are transformed into *data*, which in turn can signify *information*, which in turn requires *interpretation* so as to be generative of *knowledge*. There are many links in this chain, where degrees of arbitrariness might seep in; it is rarely a foolproof chain of evidence. In other words, as even binary sensor data might contain significant ambiguity, it is easy to see how semantic (in the Peircian parlance, we would call it *symbolic*) data is likely to raise even more complex concerns regarding context complexity.

Second, from a *material* perspective, the way interfaces are constituted is also of vital importance. Different interfaces and design choices make for different affordances,[21] including so-called "dark patterns" deliberately designed to steer the user in various directions. The generation of data, as an ontological and epistemological object, is also directly resultant from such interface design choices. As Lev Manovich has argued, data (archives) and interfaces (algorithms) are co-constitutive of each other: "The two goals of *information access* and *psychological engagement* compete within the same new media object."[22] The means of interacting with a digital archive always takes place through an algorithmic interface, and is thus determined by algorithms. Vice versa: what ends up in a database is an outcome of the interface's ability to record its surrounding social world.

Third, *situational/phenomenological* perspectives are deeply entangled in interpretation of data. Pennlert et al.[23] use

[20] Dawn Nafus, *Quantified: Biosensing Technologies in Everyday Life* (Cambridge, MA: MIT Press, 2016), xx.

[21] James J. Gibson, *The Ecological Approach to Visual Perception* (New York, London: Psychology Press, 1979).

[22] Lev Manovich, *The Language of New Media* (Cambridge, MA: MIT Press, 2001), 216, emphasis added.

[23] This volume.

optical metaphors in order to account for methodological operations—"cuts" or "folds" through the corpus, as it were—and similar optical metaphors are also indispensable for understanding relational arrangements of users of interfaces versus databases being examined through interfaces. As Lorraine Daston and Peter Galison have shown, the very notion of objectivity is deeply intertwined with vision. Interestingly, the old meaning of "'objective' referred to things as they are presented to consciousness, whereas 'subjective' referred to things in themselves."[24] That is, things would have "'objective reality' […] by virtue of their clarity and distinctness, regardless of whether they [would] exist in material form."[25] Premoderns thought of things as objective merely by being "objects of the mind and not of the world," so to speak. This, however, is not too far removed from the modern notion of what Walter Lippmann coined "phantom publics." "Public opinion" is something that can never be physically experienced; it only exists as a figment of the imagination. But, to a modern human, the *objectivity* of a poll would be a matter not of its abstractness but of its veracity, trustworthiness, or probability, and the decisive factor would be how statistically representative the underlying samples or polls would be. However, there is a blind spot here: the tendency to think of population samples in terms of representativity means that modern subjects constantly risk making the mistake of seeing the observed sample as somehow representative of the whole. This tendency can of course be exploited; I leave it to the reader to think of such examples.

Consequently, the complex production of data on social platforms cannot be examined without methodologically taking into account several converging factors. If we continue to employ optical metaphors, we find that the "ways of seeing" into digital infrastructure is premised on some rather different mechanics compared to the ways in which physical social reality allows itself to be observed.

For example, digital archives do not lend the observer any vantage point from which (s)he can attain a full overview of it as a

[24] Lorraine Daston and Peter Galison, *Objectivity* (New York: Zone Books, 2007), 29.

[25] Daston and Galison, *Objectivity*, 29.

"social totality." Observing a town square or, say, a prison yard at a distance, one can (at least theoretically) gain a *panoptic* overview.[26] With digital infrastructures, the mode of vision is instead near-sighted, *oligoptic*,[27] more akin to traversing a sequestered space, say a house or a labyrinth, where one only sees the particular room or corridor that is at hand, never the whole structure. To continue this optical metaphor, the means of access can conveniently be understood through computer science parlance of *back-end* versus *front-end* architectures,[28] the front end usually being premised on psychological engagement (personalized feeds, little or no overview) and the back end more primarily on information access (access to database, queries with aspirations to be objectivizing). The former precipitates manual local reading, continuously or in (near) real time, while the latter enables automated distant reading and systematic, retroactive overview. The former entails a risk of context collapse or misattribution, while the latter entails a risk of context loss.[29]

Moreover, in online milieus, especially those characterized by advertising and various personalization algorithms,[30] users tend to be enticed to make interactions that will in turn be farmed into interesting content for other users to interact with. Crucially, users will be shown content and advertisements; the platform operators expect to make these so that the user finds them interesting. Arguably, users adapt their behaviors in order to suit the algorithm: in order to maintain peer visibility, users update their own posts according to what the algorithm deems

[26] Michel Foucault, "Panopticism," in *Discipline & Punish: The Birth of the Prison*, trans. Alan Sheridan (London: Allen Lane, 1977).

[27] Bruno Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford: Clarendon, 2005).

[28] Metaphors that could, in turn, be related to Erving Goffman's spatial metaphors—"front stage" and "back stage"—which were applied to broadcast and telecommunications media by Joshua Meyrowitz.

[29] Jonas Andersson Schwarz and Johan Hammarlund, "Kontextförlust och kontextkollaps: Metodproblem vid innehållsanalys av sociala medier," *Nordicom-Information* 38, no. 3 (2016): 41–55.

[30] James Webster, *The Marketplace of Attention: How Audiences Take Shape in a Digital Age* (Cambridge, MA: MIT Press, 2014), 88–89.

popular or recognizable to a large audience.[31] There is a kind of built-in conformism to this, a popularity bias.[32] Individuals seem to follow trends based on machine-based predictions and thereby coproduce popular culture by delegating what content should be distributed and when.

Lastly, online accumulations rarely have *Gaussian distributions* (normal curve); they tend to have *Pareto distributions* (skewed or "long tail" curve). Early observations by Albert-Lászlό Barabási[33] showed that, although anyone can publish content on an internet website, there is no guarantee that the content will be read by any major audience or at all. This insight was expanded upon by Yochai Benkler[34] and, later, Matthew Hindman.[35] Ten years later, Hindman expanded his critique of the systemic obstacles to growing digital audiences; he attributed it to a combination of minuscule design factors and sheer mathematical odds.[36] He showed how the so-called "long-tail" distributions online[37] are too steep for small sites to add up to substantial collective market share. Such distributions, found everywhere in networked spaces online, result in "starkly inegalitarian outcomes."[38] This means, as we will see below, that conventional tests of sample representativity—which generally require that the total distributions, from which random samples are made, are following Gaussian

[31] Tarleton Gillespie, "The Relevance of Algorithms," in *Media Technologies: Essays on Communication, Materiality, and Society*, ed. Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Cambridge, MA: MIT Press, 2014), 183–188.

[32] Webster, *The Marketplace*, 89–91.

[33] Albert-László Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means* (New York: Perseus, 2002).

[34] Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (New Haven, CT: Yale University Press, 2006).

[35] Matthew Hindman, *The Myth of Digital Democracy* (Princeton, NJ: Princeton University Press, 2009).

[36] Matthew Hindman, *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy* (Princeton, NJ: Princeton University Press, 2018).

[37] More correctly, these should be labelled "lognormal," "power-law," or Pareto distributions.

[38] Hindman, *Internet Trap*, 41.

distributions—are very hard, or even impossible for these types of online corpora.

For all of the above factors, actual understanding of the media and constitutive milieus in question is premised on tacit experiential knowledge of these media and milieus in question. Researchers operate with situational knowledge—much like archivists, librarians, and historians operate with implicit assumptions about the calligraphic- or type-based universes that they traverse.[39] In other words, the particular types of visibility, homosociality, and accumulation engendered online are hard to fully appreciate without actual experience of the applications and sites in question. Ultimately, for researchers of corpuses of socially mediated text, it is indispensable to have an awareness of the situational contexts for the generation of the data in the first place, before one even begins to ask formal questions of provenance, representability, and so on.[40]

Social Big Data: Contexts

Generating observational knowledge about social media platforms requires taking these optical and material-semiotic concerns into account. But it does not stop here; to conclude this section, we will have to consider a range of further structural and global contexts in addition to the concerns listed above.

Microstructural context is crucial, as local network topologies and communities make for a range of sociological factors to consider. Primarily, one should heed the analytically useful differentiation between the "subjective" surrounding that the user finds herself immersed in (phenomenological *Umwelt*, or *constitutive milieu*),[41] and the more "objective" notion of a surrounding, general ecology (*Umgebung*), as it would constitute itself for an external

[39] Marshall McLuhan, *The Gutenberg Galaxy: The Making of Typographic Man* (Toronto: University of Toronto Press, 1962).

[40] Derek Ruths and Jürgen Pfeffer, "Social Media for Large Studies of Behavior," *Science* 346, no. 6213 (2014): 1063–1064.

[41] Jonas Andersson Schwarz, "Umwelt and Individuation: Digital Signals and Technical Being," in *Digital Existence*, ed. Amanda Lagerkvist (London, New York: Routledge, 2018).

observer.[42] However, since any act of observation is premised on the observer's own local faculties and schemata, social reality is never "purely" observed. In other words, those claiming access to a "more objective" overview would in fact be doing so through *their own* constitutive milieu. I have argued that this is particularly decisive when considering social media platforms as interactive design objects, where intelligence-based design decisions intend to help analysts tracing user preferences through various dashboards (i.e., intelligence-compiling interfaces). Such dashboards capture discrete signals given off by users—while the users, on the other hand, are entangled in a mode of usage that is always and forever based on interpretations of the digital interfaces made by the designers. Here, too, we see a co-constitutive "dance" of interpretive agency: everyday users try to draw conclusions from signals provided through interfaces, while the designers of these interfaces try to draw conclusions from intelligence on the signals generated.[43] To really grasp the complexity at hand, it is instructive to peruse the literature on first- and second-order cybernetics and its debates on "observer problems."[44]

Macrostructural context would address the ways in which internet-based social media platforms are shaped by their various biases pertaining to ownership and the attendant economic incentives motivating their owners. This has to be understood in combination with fluctuations of world affairs, as these form a political-historical context. In conventional mass media system theory, macrostructural concepts—that is, the notion of (national) "public opinion"—as well as functional entities—that is, various notions of mass media systems—have been established epistemological tools for the last hundred years.[45] In recent years, this functional theory has been complemented by notions of "social media logic," which act to explain the workings of social media systems by reference to structural incentives. Importantly,

[42] Andersson Schwarz, "Umwelt."

[43] Andersson Schwarz, "Umwelt."

[44] See, e.g., Heinz von Foerster, Niklas Luhmann, Siegfried Schmidt, and Gregory Bateson.

[45] See, e.g., Walter Lippmann, Wilbur Schramm, and, later, Jürgen Habermas, David Altheide, Daniel Hallin, and Paolo Mancini.

the hybrid entanglement of mass media and social media has become something that is increasingly emphasized.[46] Mass media and social media interact, primarily in their mutual interplay of references. "Mass media logic and social media logic get incrementally entangled in defining the popularity of issues and the influence of people."[47]

Lastly, there is a larger cultural *global context*: behavioral, linguistic, narratological, and rhetorical modes of habit. These are patterns more stable over time than the geopolitical and events-based fluctuations in the above category. Habit and cultural norms rarely shift rapidly, and tend to follow generational patterns. It would be spurious to attribute agency to macro factors such as "globalization," "imperialism" *et cetera* without considering the intermediaries or mediators of such forces.[48] The study of so-called "bots" or "trolls" in social media settings would be an example of this; they can be regarded as concrete specimens whose very occurrence can be (albeit speculatively) traced to particular politico-historical conditions. The explanatory power of abductive theory can, if we are to apply this Latourian perspective, never be as certain as the original descriptive observations of these actual mediators. More on such abduction below.

Epistemological Challenges of Social Big Data for Scholarly Research

I will now turn to two research projects that I have had experience from, illustrating the challenges at hand. The first one addresses indexical, numerical data, and the second one symbolic, semantic data.

The first one is a study on *The Transformation of a Swedish Twittersphere* (#svpol, 2014 to 2018), targeting the logics of sharing and retweeting in the Swedish Twittersphere, and, in particular, the interactive logics between social and editorial media.[49]

[46] See e.g., Andrew Chadwick, Tarleton Gillespie, David Beer.

[47] José van Dijck and Thomas Poell, "Understanding Social Media Logic," *Media & Communication* 1, no. 1 (2013): 8.

[48] Latour, *Reassembling*.

[49] Andersson Schwarz et al., *Opinioner*.

Through a combination of methods, mainly SNA and quantitative content analysis, we were able to find patterns and could categorize the shape of affinities among people using the hashtag #svpol on Swedish-language Twitter, during the election year 2014. We could also trace the recycling of tweets in political debate and opinion formation, as well as illustrating some important aspects of the interaction between Twitter and editorial media.

Our study was based on an original data harvest of approximately 109,000 tweets covering the distribution of Sweden's most popular hashtag, #svpol, on 25 different days during the election year of 2014.[50] We combined SNA (14,412 nodes, 37,959 edges[51]) that showed network topology (cluster identification) with a rather conventional content analysis (n=500), in which a systematic selection of tweets was interpreted manually in order to assess semantic meaning. Moreover, our study was a typical example of collaboration between industry and academia. We reflected on the nature of such collaboration, and chose to emphasize the challenges—especially for established organizations within competitive intelligence and editorial media—inherent to the interpretation of social media flows.[52]

Four years later, I was able to make a repeat sample, representative of the same 25 days (i.e., election year 2018) from the same data provider (approx. 125,000 tweets; 42,645 nodes; 77,810 edges). By letting the procedural collection remain the same, the potential for comparability over time increases. While the analysis of this more recent data is still under way, I would like to stress the point that such comparability provides an important aspect of contextualization: regardless of how exogenously representative the data samples in question would be (and I maintain to exhaustively list the various ways in which they would be lacking in that respect), there would be endogenous consistency, making it revealing so as to see how the shape of this particular social space changes over time.

[50] Data was provided by intelligence company Retriever, which in turn fetched Twitter data from data provider Sysomos.

[51] Each edge is one instantiation of a tweet being retweeted.

[52] Andersson Schwarz and Hammarlund, "Kontextförlust."

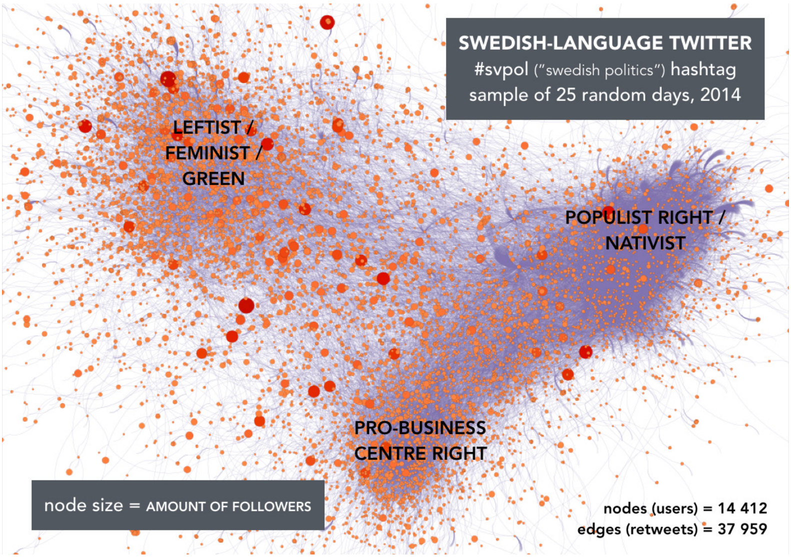

Figure 1: Political clusters in the Swedish Twittersphere, election year 2014. Copyright: Author. License: CC BY-SA.

One interim conclusion from these different analyses is that, while the first dataset (Figure 1) clearly showed three political clusters (leftist/feminist/green; pro-business/center-right; and populist-right/nativist), the later dataset (Figure 2) indicates that the "middle" center-right cluster has all but disappeared, while the populist-right/nativist cluster appears to have intensified in activity. This should not be understood to be showing a popularization of populist-right/nativist sentiment, as the dataset only captures tweets being deliberately tagged #svpol, and it remains unclear how much of the change is due to changing propensities among Swedish-language Twitter users to use that hashtag. However, it should be seen as indication that, for those who choose to use this hashtag, there seems to be an intensification and radicalization of use.

By making a strategic sample, interesting questions about representativity can be answered. The relatively small sample was big enough to generate a reliable visualization of the ways in which this Twitter hashtag propagated, simply by tracing the dissemination of a simple parameter: occurrences of retweets (disregarding the content of each tweet). While the study only observes

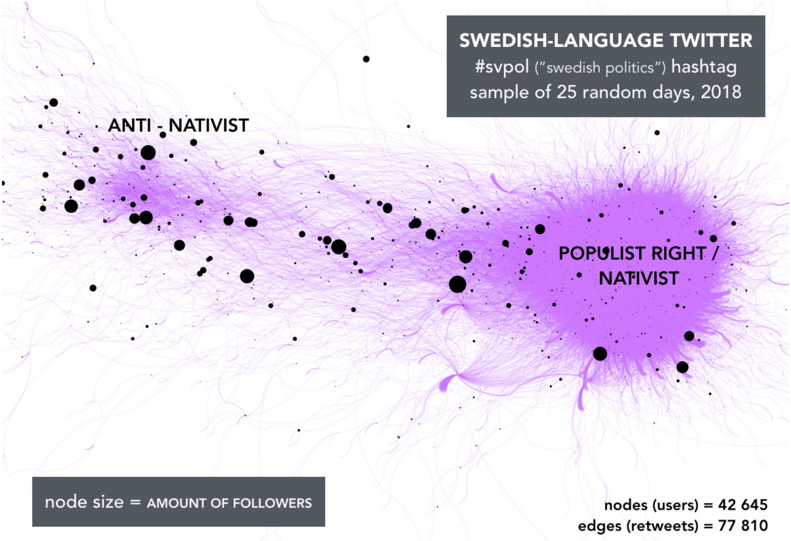
SWEDISH-LANGUAGE TWITTER
#svpol ("swedish politics") hashtag
sample of 25 random days, 2018
ANTI - NATIVIST
POPULIST RIGHT /
NATIVIST
node size = AMOUNT OF FOLLOWERS
nodes (users) = 42 645
edges (retweets) = 77 810

Figure 2: Political clusters in the Swedish Twittersphere, election year 2018. Copyright: Author. License: CC BY-SA.

the formations of a particular hashtag during particular time periods, some of the more principal logics observed are likely to be apparent also in social media in general. The repetition of the exact same sampling over time strengthens internal reliability, recording different historical states of this propagation.

The second project, *Linguistic Explorations of Society*, is transdisciplinary in nature and aims to investigate the very notion of representative text data for populations in various countries, primarily in order to lay the groundwork for better transnational comparability between open-text answers from official national surveys and found text data from online (editorial and/or social) sources.[53] The purpose is to extract large-scale language collections, to make possible purely statistical, computer-driven methods like NLP in order to find intralinguistic dimensions (e.g., entity recognition and linking, relationship extraction, sentiment analysis, topic detection, and author identification). In a preceding project,[54] the same researchers had employed similar methods in

[53] Dahlberg, *Linguistic*.

[54] Stefan Dahlberg, *Language Effects in Surveys* (Swedish Research Council, 2014).

order to find patterns in corpuses of accumulated responses to open-ended questions in formalized surveys on political opinion, around the globe. The purpose was to be able to more accurately estimate the inherent biases that would skew the comparability due to, for example, Spanish-language speakers, Russian ones, and Swedish ones meaning different things when they respond to words like "democracy."

Currently, I am in the process of surveying available online text data sources across countries. In addition to surveying things like internet penetration, media use, available social media platforms, modes of censorship, freedom of speech indices and so on, one of the key tasks is to unravel and critically describe the complex ecosystem of commercial vendors/providers of user-generated text data, for different languages and in different countries. Access to user-generated online text data is determined by the largely commercial nature of not only the sources, but also the redistributors of such data; complex arrangements of interlinking commercial vendors, each providing different modes of access and collections of sources (generally by way of different APIs[55] and/or visual dashboards). Each vendor offers a particular selection of sources, and moreover remains largely opaque as regards the provenance of these sources. The data in question is not Facebook data (which is, nowadays, practically unavailable, especially for commercial purposes) and rarely even Twitter data (due to arrangements to do with the terms of use of the Twitter API). Rather, the data in question is often scraped from wikis and other online corpora, blogs, and message boards on the open web, alongside various web-based editorial news sources. Moreover, vendors normally only make available the most recent 30 days of data, shelving older data on magnetic tape (thus making it significantly less accessible). Each vendor also presents its own, institutional terms of use, by default often making reuse of the data in question impossible. For these reasons, and several more, it really hard to employ traditional standards of statistical representativity and replicability. Thus, certain types of data are generally very hard to get hold of:

[55] Application programming interfaces.

- Chat app data from, for example, WhatsApp, WeChat, Snapchat, Kik, and Facebook Messenger.
- Individual posts (private and public) from Facebook, Instagram, and LinkedIn.
- Geo-tagged data.
- Demographic variables (e.g., gender, age, and income) for the sources in question.
- Historical data (older than 30 days).

Some interim conclusions can also be noted here. To begin with, few vendors provide adequate service-level agreements concerning data coverage and/or latency. There are generally no formal guarantees whatsoever concerning text quality. Vendors tend to guarantee only certain quantities of text and/or frequencies of specimens.

Moreover, legal conditions—such as copyright, data protection (e.g., the EU's GDPR), and terms of service—dictate a lot of archival uses of web-mediated data. These conditions are stacked on top of each other: the original platform (e.g., Twitter) would stipulate certain terms of use, but also the attendant commercial retailers stipulate their own, extraneous terms of use, often explicitly prohibiting any sharing of the data, due to its business value. All of these factors are inimical not only to the provenance, reliability, and representativity of the data—but also to the reproducibility of scientific results. Despite searching for the same time period and language, it is not hard to imagine one vendor providing a slightly different set of data than another vendor, depending on how they handle things like geo-tagging and/or language detection, original API access, search query designs, and so on. Even the point in time when the query is made would have effect, since the original platform might have removed old content at any time. There are many potential points of failure and/or bias, and providers are unlikely to give any formal guarantees regarding data completeness.

In order to enable at least formal reproducibility/replicability of scientific results, one can employ large-scale web scraping services, for example CommonCrawl, or commercial tools for scraping (so-called "crawler-as-a-service" tools), as alternative means to gather data for the purpose of large-scale processing and pattern recognition. However, manual web scraping is nontrivial: Data is always noisy, and requires considerable processing before use. For any scientific research project, this suggests that considerable resources should be considered for the purpose of scraping, cleaning, and preparing data alone.

Conclusions

Digitization has precipitated an ever-growing, immensely huge glut of behavioral data that I propose to label social big data, consisting not only of skeletal *indices* of human activity but also of *semantic* data user-initiated circulations of user-created bulletins in real time, rapidly shifting in terms both of volume and speed. This data can now be fetched in enormous volume, in near real time, and combinatory approaches can reveal patterns in this data glut. It is becoming increasingly necessary to critically interpret these data streams, while at the same time *not* seeing automatically generated tallies and anecdotal examples as automatically self-explanatory and representative.

Classic criteria for validation remain. *Validity:* does it measure the right thing? *Reliability:* does it measure it reliably, carefully, and comprehensively? *Representativity:* does it accurately correspond with that which is signified? *Replicability:* does the empirical endeavor allow for others to repeat it? *Salience:* does it have features which skew the observer's attention? Out of these important criteria, the three last ones appear to be the most troubling ones, when it comes to found specimens of online discourse. Overstating the certainty, veracity, and quality of social big data leads to four distinct problems.

First, in terms of claims to *objective realism* or representativity, what imaginaries of "the public" are generated? Social media are sometimes alleged to be "closer" to the public, a more "direct" mediation of public sentiment. The ontological observation that social media would "mirror" populations inevitably leads to the epistemological position that social media would be *better positioned* than, for example, editorial media to represent reality. Not only are the arguments presented in this chapter a rebuff to such claims; recent events such as the Cambridge Analytica debacle seem to have led many people also outside of academia to understand that such a position is no longer tenable.

Second, the *persuasive capacity* of numbers becomes urgent. This applies to discourses appearing in social media, as well as to references to social media appearing in mass media. Both tend to have heavily metrological characteristics, in that they divulge numerical tallies of different kinds. Numbers can in themselves have normative effects. The invocations to alleged quantitative impacts that onlookers would be able to make can often be rash and uncorrected misinformation can spread like wildfire. In order to be better at identifying possible misaligned claims to representativity, not only do we need better literacy on behalf of citizens; we also need to see improved transparency on behalf of platform providers.

Third, there are numerous *power dimensions* at hand. End users (not only ordinary citizens but researchers, planners, and journalists as well) are only allowed a filtered front-end access to the data glut to which platform providers like Google and Facebook have unfiltered back-end access. Hence, as end users, we are demoted to a secondary subject position in terms of knowledge generation. In that capacity, we are often forced to merely second-guess actual distributions. Lack of access thus enforces a form of *abductive* reasoning (qualified "guesstimation"). The inherent possibilities that this reinforces tendencies pertaining to the shaping of opinion and knowledge in society is a highly urgent sociological question.

Lastly, *contextualization* is unavoidable. Human beings cannot *not* interpret, and any data that is scrutinized by humans will (unwittingly or not) be categorized, valued, and even perhaps judged—even before formalized explications for such endeavors are formulated. However, it would be valuable to make continuous assessments also of the limits for *overcontextualization*. For the purpose of generating new, repurposed archives, the question of how one should relay archival contexts without overcontextualizing the archive is central.[56] It is also important for singular analytical projects, when abductive conclusions are drawn from data, so as to avoid qualitative overinterpretation of one's findings.

[56] There are many important differences between creating archives for posterity, and creating temporary datasets intended to be used only as interim working tools.

Ultimately, I advocate a pragmatic approach, combining data science with more conventionally interpretative methods. By carefully combining methods,[57] it is possible to maintain the benefits of conventional content analysis (systematic stringency, contextual sensitivity, hermeneutic depth) while providing effective capabilities for overview and/or enumeration through computational methods based on automation and/or algorithms (e.g., visualizing relationships and covariations, as well as detecting the relative sizes of different aggregations).

References

Andersson Schwarz, Jonas, and Johan Hammarlund. "Kontextförlust och kontextkollaps: Metodproblem vid innehållsanalys av sociala medier." *Nordicom-Information* 38, no. 3 (2016): 41–55.

Andersson Schwarz, Jonas, Johan Hammarlund, Stefan di Grado, and Magnus Kjellberg. *Opinioner och offentligheter online: Slutrapport för projektet Vad gör en politisk utsaga framgångsrik? Den användardrivna kommunikationens villkor*. Research report. Stockholm: IIS, 2015.

Andersson Schwarz, Jonas. "Umwelt and Individuation: Digital Signals and Technical Being." In *Digital Existence*, edited by Amanda Lagerkvist, 61–80. London, New York: Routledge, 2018.

Andreotta, Matthew, Robertus Nugroho, Mark J. Hurlstone, Fabio Boschetti, Simon Farrell, Iain Walker, and Cecile Paris. "Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis." *Behavior Research Methods*, no. 51 (2019): 1766–1781.

Barabási, Albert-László. *Linked: How Everything Is Connected to Everything Else and What It Means*. New York: Perseus, 2002.

[57] Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida, "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods," *Journal of Broadcasting & Electronic Media* 57, no. 1 (2013); Matthew Andreotta et al., "Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis," *Behavior Research Methods*, no. 51 (2019).

Benkler, Yochai. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press, 2006.

Bolin, Göran, and Jonas Andersson Schwarz. "Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation." *Big Data & Society* 2, no. 2 (December 2015): 1–12. DOI: https://doi.org/10.1177/2053951715608406

Boullier, Dominique. "Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications." In *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*, edited by Franck Cochoy, Johan Hagberg, Magdalena Petersson McIntyre, and Niklas Sörum, translated by Jim O'Hagan, 20–40. London, New York: Routledge, 2017.

Dahlberg, Stefan. *Language Effects in Surveys*. Research project "The Advantage of Country Comparisons: Towards a New Method for Estimating Language Effects in Cross-Cultural Surveys." Swedish Research Council/Vetenskapsrådet, 2014.

Dahlberg, Stefan. *Linguistic Explorations of Society*. Research project "Studying Opinions and Societies Through Communicative Behavior Online: Assessing the Validity, Reliability, and Representativity of Online Text Data" (ongoing). Swedish Research Council/Vetenskapsrådet, 2017.

Daston, Lorraine, and Peter Galison. *Objectivity*. New York: Zone Books, 2007.

Firth, John. "A Synopsis of Linguistic Theory 1930–1955." In *Studies in Linguistic Analysis 1930–1955*, edited by John Firth, 1–32. Oxford: Basil Blackwell, 1962.

Foucault, Michel. "Panopticism." In *Discipline & Punish: The Birth of the Prison*, translated by Alan Sheridan, 195–228. London: Allen Lane, 1977.

Gibson, James J. *The Ecological Approach to Visual Perception*. New York, London: Psychology Press, 1979.

Gillespie, Tarleton. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 167–194. Cambridge, MA: MIT Press, 2014.

Hindman, Matthew. *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton, NJ: Princeton University Press, 2018.

Hindman, Matthew. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press, 2009.

Kitchin, Rob. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1, no. 2 (April–June 2014): 1–12. DOI: https://doi.org/10.1177/2053951714528481

Latour, Bruno. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Clarendon, 2005.

Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting & Electronic Media*, 57 no. 1 (2013): 34–52. DOI: https://doi.org/10.1080/08838151.2012.761702

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Online ed. Cambridge: Cambridge University Press, 2009. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Manovich, Lev. *The Language of New Media*. Cambridge, MA: MIT Press, 2001.

Marcuse, Herbert. *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society*. Boston, MA: Beacon Press, 1964.

McLuhan, Marshall. *The Gutenberg Galaxy: The Making of Typographic Man*. Toronto: University of Toronto Press, 1962.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs, 2013.

Nafus, Dawn. *Quantified: Biosensing Technologies in Everyday Life*. Cambridge, MA: MIT Press, 2016.

Piotrowski, Michael. "Digital Humanities, Computational Linguistics, and Natural Language Processing." Presentation, *Lectures on Language Technology and History*, Uppsala University (March 4, 2016).

Piotrowski, Michael. *Natural Language Processing for Historical Texts*. London, Williston, VT: Morgan & Claypool, 2012.

Ruths, Derek, and Jürgen Pfeffer. "Social Media for Large Studies of Behavior." *Science* 346, no. 6213 (2014): 1063–1064.

Schubert, Lenhart. "Computational Linguistics." In *The Stanford Encyclopedia of Philosophy*. Spring 2020 ed., edited by Edward N. Zalta. Stanford, CA: Stanford University, 2020. https://plato.stanford.edu/archives/spr2020/entries/computational-linguistics.

Scott, John. "Networks & Relations." In *Social Network Analysis*, edited by John Scott, 1–9. London: Sage, 2013.

van Dijck, José, and Thomas Poell. "Understanding Social Media Logic." *Media & Communication* 1, no. 1 (2013): 2–14.

Webster, James. *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. Cambridge, MA: MIT Press, 2014.

Wetzel, Linda. "Types and Tokens." In *The Stanford Encyclopedia of Philosophy*. Fall 2018 ed., edited by Edward N. Zalta. Stanford, CA: Stanford University, 2018. https://plato.stanford.edu/archives/fall2018/entries/types-tokens.