Chapter 12

Datagraphy as Historiography of Big Data: Taking Account of Unintentional and Intentional Misrepresentations by Twin Big Brothers and Expert Data Producers

Ulaş Başar Gezgin[1]

Abstract

The purpose of this article is to classify misrepresentations in big data. There are three major sources of misrepresentations:

- Unintentional misrepresentations by surveilling governments and corporations (we call them as 'the twin big brothers') such as recording errors and sampling biases,
- Intentional misrepresentations by the twin big brothers such as distortions and manipulations,
- Intentional self-misrepresentations by data producers (i.e. internet users, consumers and citizens) such as faking data to protect oneself from harms to be inflicted by the health system or to protect right to privacy.

This is a critical thought paper. Thus the methodology consists of philosophical discussions.

The analyses in this article bring forth a new way to classify misrepresentations in big data, the notion of twin big brothers, the proposal to develop a new research area, datagraphy, and the notion of omniresistance.

Against the big data optimists, this article clearly shows that the big data are corrupt from the very beginning due to the conditions set by surveillance capitalism.

This article brings forth new concepts and conceptualizations with regard to big data. It is expected that it will move other researchers to reflect on and develop further the new ideas presented in this work.

1. Introduction

In this article we present and discuss unintentional misrepresentations such as errors and biases, intentional misrepresentations such as distortions

1 Professor, Istanbul Galata University, Faculty of Arts and Social Sciences, Department of Psychology, ulas.gezgin@galata.edu.tr, Orcid: 0000-0002-6075-3501

https://doi.org/10.58830/ozgur.pub135.c628

and manipulations with regard to big data, and big data producers' resistance which misfeeds big data as users', consumers' and citizens' self-misrepresentations. Based on these points, the notion of datagraphy is proposed as the historiography of big data. By datagraphy, we mean the study of the way the data are produced, recorded, interpreted, presented etc. That is because in analogy with history and historiography, the information is not recorded in big data as it is, since data recorders always add their interpretations to their recording through their assumptions about what counts as recordable and not recordable. The article argues that big data can never be true and objective, as they are recorded for governments' control motive and corporations' profit making motive. As an original contribution of this article, in addition to the notion of datagraphy, the notion of twin big brothers is introduced to refer to surveilling governments and corporations whose boundaries are mostly blurred in practice, though they differ in their motive. The article concludes by the notion of omniresistance for truth defenders to resist anywhere, any time and by all means.

In our discussion, we make a distinction between unintentional and intentional misrepresentations in big data. By this distinction, we would like to believe in benevolence of some of the big data optimists, but argue that they are wrong, although they are not aware of it. That is why, their failure is unintentional. But there are other forms of misrepresentations which are clearly intentional. In the next section, we present and discuss unintentional misrepresentations such as errors and biases in big data. But before this discussion, we would like to note that in some cases it is hard to draw this distinction.[2]

2. Unintentional Misrepresentations: Errors and Biases in Big Data

Harford (2014) argues that

"Cheerleaders for big data have made four exciting claims, each one reflected in the success of Google Flu Trends: that data analysis produces uncannily accurate results; that every single data

2 Here is an example in which case whether the misrepresentations are intentional or unintentional is hard to decide: The data collected from and about human beings differ in terms of subjectivities and sensitivities. For example, death counts in American occupation of Iraq or Syrian civil war are associated with emotions which can cause another group of biases. Price & Ball (2014) identify a particular type of bias:

"Event size bias is the variation in the probability that a given event is reported, related to the size of the event: big events are likely to be known, small events are less likely to be known. Event size bias is the variation in the probability that a given event is reported, related to the size of the event: big events are likely to be known, small events are less likely to be known." (p.11).

point can be captured, making old statistical sampling techniques obsolete; that it is passé to fret about what causes what, because statistical correlation tells us what we need to know; and that scientific or statistical models aren't needed because, to quote "The End of Theory", a provocative essay published in Wiredin 2008, "with enough data, the numbers speak for themselves"" (p.14-15).

Harford (2014) rightly proposes that none of them is correct, they are merely oversimplifications. The big data is infected with selection bias, we still need statistical techniques, causality is still the basis of science as correlation can be due to innumerable confounding variables not considered in big data analysis and big data approach itself is a model.

Crawford, Gray & Miltner (2014) state that "[t]he very term big data science is itself a kind of mythological artifact: implying that the precepts and methods of scientific research change as the data sets increase in size" (p.1664). They call the big data enthusiasts big data fundamentalists. In some other works, we see the use of the term 'myth of large n' (e.g. Seely-Gant & Frehill, 2015).

According to Leonelli (2014), contrary to enthusiasts' claims, big data research still needs to abide with sampling principles and the sheer size does not eliminate biases in collecting and interpreting data. Interestingly enough, he argues that big data is nothing new for natural sciences such as biology, as natural sciences are always engaged in big data analysis anyway. However, he continues, big data proponents' fascination for correlation instead of causation can't be appealing for natural scientists, as correlations can be spurious or due to uncontrolled confounding variables. For Leonelli (2014), big data approach can be only revolutionary for social science areas such as business, economics and politics.

Throughout their discussion based on epidemiological and health services data, Kaplan, Chambers & Glasgow (2014) warn that unlike the enthusiasm about big data, the errors can be much bigger in big data such as "sampling error, measurement error, multiple comparisons errors, aggregation error, and errors associated with the systematic exclusion of information" (p.342). Their pre-big-data example comes from the American presidential election polls conducted with 2.4 million voters. However, the data predicted another candidate. So they would like to remind us that even in a smaller scale, the big data is not infallible. In this case, it suffers from representativeness bias. Likewise, the big data currently collected can be non-representative, considering huge masses of people that are not connected. Accordingly,

Kaplan, Chambers & Glasgow (2014) question "the assumption that large sample sizes yield more meaningful results than small sample sizes" (p.342).

In the case of medical big data, Kaplan, Chambers & Glasgow (2014) aptly warn that not everybody has insurance, and thus not everybody is represented in the relevant databases. Related to this point, big data is not immune to availability bias. Currently, due to the misbelief that bigger is better, all available data is inputted to the system. Even if all the data inputted are true (and we know that that is not the case), it may not necessarily represent the unavailable data. This problem has disastrous consequences in case of a war where limited resources should be allocated across various localities (Price & Ball, 2014).

Selection bias with regard to big data can be defined as

"the likelihood that certain persons or groups are more apt to be picked up by big data collection efforts than others, whether due to their use of social media and open source platforms, the availability of internet connectivity in certain areas, their ability to purchase smart phones and access applications, or any other number of omitted variables" (Seely-Gant & Frehill, 2015, p.31).

Statistically speaking, for samples with the same p level, effect size is negatively associated with the sample size. In other words, larger samples with the same p level have lower effect sizes (Kaplan, Chambers & Glasgow, 2014). How come? There are strong relationships that require a smaller sample size for researcher to detect. Weaker relationships will require a higher sample size to have the same effect. That means big data can unravel unexplored links, but they may be meaningless.

Buelens et al. (2014) warn that

"the measuring mechanisms for Big data sources are unlike those used in survey sampling, where through careful questionnaire design and interviewer training the measurement of well-defined constructs is operationalized" (p.4).

According to Liu et al. (2016), "big data brings lots of "big errors" in data quality and data usage, which cannot be used as a substitute for sound research design and solid theories" such as "inauthentic data collection, information incompleteness and noise of big data, unrepresentativeness, consistency and reliability, and ethical issues" (p.134).

Another point of inconsistency rests in the fact the number of social media users and their identities are not stable, they can change any time.

Thus, a valid conclusion today may not be applicable tomorrow (Liu et al., 2016). Thirdly, smart-phone-based data in many cases reflect class reality in the society: Well-connected areas are usually those populated with middle and upper class. As a result, for example, potholes and mapping features are described in more details for middle and high income areas when they are based on user-generated content (Harford, 2014).

In this context, digital divide also disables some of the claims of big data enthusiasts: According to World Bank statistics, although there is a rising trend, still only half of the world uses internet (World Bank, 2018). It is no surprise that smart phone subscriptions are far less. In that sense, at a world-level analysis, big data is not a matter of today, but tomorrow. To make vast and universal claims, big data enthusiasts should wait longer, as it will take long time for world-level internet penetration.

In fact, because of the potential errors, biases, and especially the sampling problem, big data will always need verification by small data. Data missed and ignored by big data will always point to potential confounding variables which, in the arrogance of big data model, is not even considered. In fact, there may a systematic pattern in missing data. Graham (2008) in his work about missing data reframed as attrition bias recommends "using auxiliary variables, collecting follow-up data on a sample of those initially missing, and collecting data on intent to drop out" (p.549). What big data does not include is even more important than what it covers, as there is no way to properly forecast the former without reference to small data. Of course, there are statistical models for forecasting, but apart from their mathematical beauty, whether they are realistic or not is a moot question. There is another possibility which makes big data use even more problematic: The real factors behind let's say a certain kind of human behavior can be the ones that may not be tractable by surveillance tools and other big data collection devices. Without the help of small data and small data researchers, big data can't identify them. For example, it is common to collect information about basic health indicators, but not personality differences.

About electronic health data, Moore & Furberg (2015) conclude that

"The high levels of variability in almost every parameter render findings difficult to replicate and vulnerable to substantial bias, either as an accident of data and method selection or through intentional manipulation of study criteria.

At present, few studies have been conducted to assess the likelihood that risk assessments based on electronic health data systematically underestimate

the adverse effects of drugs. Unless great caution is used in interpreting studies that do not detect a drug effect, society is at substantial risk that evidence of important drug harms may be masked, potentially blinding us to safety concerns that could affect millions of patients" (p. 608).

As a solution they recommend additional studies to maintain scientific principles such as validity, reproducibility, reliability, consistency, accuracy etc. (Moore & Furberg, 2015) which require small data. Once again, we see that big data alone (without the support of small data) can't be helpful in scientific advancement as it may mislead the research with spurious correlations which can have deadly implications for public health. Fortunately enough, big data is not in a position to guide public health policies despite of the keen proposals to that direction.

A number of studies offer statistical explanations for the biases in big data (e.g. Kaur & Arora, 2015; Lu & Li, 2013), which can be defined as "a measurable difference between the observed sample and the underlying population of interest" (Price & Ball, 2014, p.16), however they assume that the data is properly collected and recorded, and no corporate, government or expert distortion and manipulation were applicable, which is not a tenable assumption. Of course, these statistical big data bias studies are useful, but we claim that there are more fundamental problems associated with big data. So far we have seen unintentional misrepresentations such as errors and biases. In the next section, we study intentional misrepresentations such as distortions and manipulations.

3. Intentional Misrepresentations: Distortions and Manipulations in Big Data

When it comes to big data, as in many other cases, it is hard to see the boundary between government surveillance and corporate surveillance. We call these surveilling government and corporations as twin big brothers, a new term first introduced in this article. Breur (2016) points out that CEO of Trump's presidential campaign (2016) was a board member of Cambridge Analytica, a data firm later on associated with deliberate manipulation and fabrication of Facebook data (cf. Laterza, 2018; Tarran, 2018; Tuttle, 2018). Secondly, it is a known fact that social media data which is a part of big data can easily be faked and distorted (Breur, 2016).

As explained by Liu et al. (2016), Google data based on search terms are inherently distorted, as the system provides auto-complete options which distracts users' attention and hampers independent decision making. Converging with this point, social media platforms are manipulating the

users by suggesting them accounts to connect. The users may have the illusion that they are free on social media, but that is not the case. Secondly, the primacy of similar pages on search terms leads to self-confirmation, with the misperception that the virtual world is mostly in favor of a particular view rather than another. Another example for distortion of big data provided by Liu et al. (2016) is about geo-tagged tweets. Regardless of where s/he is, the user can tag his/her tweets with any place. S/he can post from Hong Kong and New York consequently, which shows that Twitter data are not verified at all.

Corporate world is well-known in manipulating and distorting data for profit maximization (cf. Van Dijk, 2014). Pharmaceutical companies exert pressure over data recorders such as doctors to prescribe rather than let the illness recover by itself in a couple of days, and also to prescribe the drugs produced by particular companies (cf. Civaner, 2012; Hsu, Fang & Lee, 2009; Ijoma et al., 2010). Unethical 'scientific' practices are often associated with those companies (cf. Gøtzsche et al., 2009; Moffatt & Elliott, 2007; Smith, 2005). Similar cases can be reported about insurance companies and some other relevant industries. Surveillance capitalism has inherent reasons to distort public health big data. Compared to small sample research, big data is again more inclined towards big errors, since these distortions, manipulations and sampling problems go unnoticed and the errors get multiplied without any measure taken against them.

Lukoianova & Rubin (2014) introduces veracity as the 4th V, as big data comprise various degrees of error. Veracity is defined on the basis of the degree of objectivity (vs. subjectivity), truthfulness (vs. deceptiveness) and credibility (vs. implausibility) of the data (Lukoianova & Rubin, 2014). Without a veracity score for each data, it is impossible to evaluate whether the findings proposed based on big data is valid or not (Lukoianova & Rubin, 2014).

In article which asks whether big data is big brother, Khanna (2015) reminds us another trouble with big data, although it is not a bias: Anyone can buy them and use them for their own interests. The interest in medical data is not benevolent per se, there are also financial motives to identify who would likely to be costly for health institutions. Insurance companies denying services on the basis of the link they discovered in big data can be a science fiction theme right now, but as for many other cases, it can become the reality. The data set also includes sensitive personal data about who has depression, diabetes, bed-wetting, erectile dysfunction etc., as Khanna (2015) warns us. That my private health data can be sold to anyone who

pays for it is incredible in that sense. If that is the case, in the near future, it is no surprise to see that people will try their best to avoid being recorded or fake their personal data.

Liu et al. (2016) warn that big data is not collected by scientific research institutions based on scientific methods, but by businesses motivated by profit. Thus, data collection is problematic from the very beginning as businesses don't feel any need to abide with scientific research principles. Kitchin & Lauriault (2015), in this context, reminds us that big data collected by the businesses are rarely accessible by the public. Without transparency of big data and declaration of underlying algorithms and any changes to these, big data can't meet the scientific standards required for empirical research. Furthermore, even if we have access to data, no question unsettling the data provider could be formulated in a research program, as they can cut data access as a response (boyd & Crawford, 2012).

Couper (2014) discusses file drawer bias with regard to predictions based on big data. Although it is a bias, at a meta level, it becomes a distortion and manipulation to legitimate big data: When the prediction is successful it extensively appears on media, while in most of the cases the predictions fail, but they are not publicly presented and discussed. Thus, people only hear about the success stories of big data. The rest are kept in file drawers. We can add to this point the following: The sensational success stories also hide eventual failure of big data analytics in some of the cases. For instance, as mentioned before, Google Flu Trends was hailed as a successful example outcompeting even health authorities in predicting flu epidemics based on Google search items. In public discussions, this is usually the most typical example to legitimate Big Data. However, eventually it was found that Google's predictions were wrong either because people search news about flu, rather than their symptoms and Google as a search engine is set to lead to a particular kind of results through its algorithm and auto-complete function (see Harford, 2014). In a similar case, Couper (2014) reminds us why election predictions based on big data fail: Some parties such as Pirate Party has strong internet presence which unrealistically skews data. In other words, from the very beginning, political parties are not equally represented on internet. In fact, carefully designed small sample surveys are still more predictive than big data in predicting election results as mentioned before. However, when big data succeed in predicting election results in a single case among thousands of elections, it is considered as an indicator for its success. Another example is about the company predicting that a customer is pregnant based on her purchases, before even her father notices

it (Harford, 2014). The serious privacy violation aside, again they never tell us how many times they failed in their predictions.

In the next section, we briefly see the third major source of misrepresentation in big data.

4. Big Data Producers' Resistance: Users', Consumers' and Citizens' Self-Misrepresentations

Big data enthusiasts are in fact characterized by an old fallacy which is called 'naturalism':[3] This is the fallacy to treat human data as natural data. They can't be same, as humans unlike nature have agency. They can change the way they live, think, eat etc. They are not as predictable as natural events. That is why we can't talk about universal laws about social aspects of human beings. For instance, people can lie to the health care provider in order not to be harmed by the health insurance system (cf. Werner et al., 2004). They can state that they were never hospitalized, if there is no health record about their past hospitalizations. They can lie about illnesses of their parents, in order not to be blacklisted by insurance companies. They can lie to their employer about their health status to get the job etc. We can claim that this big-datafication will create an undatafied black market for all services recorded in favor of the surveillance capitalism. For instance, in some countries, teen abortion is illegal or the teen is scared of informing parents about her pregnancy (cf. Klick & Stratmann, 2007; McKay & Barrett, 2010), but there are informal (illegal) ways to serve pregnant teens who have unwanted babies. Just like this example, due to datafication of doctors and hospitals, people may prefer informal health service providers for the needed services and pharmacies to receive health information.

Many other misrepresentations in big data are yet to be mentioned and discussed in details: The knowledge that what I am saying will be recorded can change structure and form of my input to big data. I will be more formal, and look like more educated. We can call this as 'the verbal Hawthorne effect'.[4] Secondly, my imaginations about the audience of my data will bias my data. Thirdly, fictionalized and non-fictionalized accounts of the same event (e.g. 9/11) will make a big difference. Fourthly, each text has a style. For example, I can exaggerate what happened to me today to draw interests of my audiences. Also in biographical vs. autobiographical statements we

3 Not to be confused with naturalistic fallacy which "refers to "is-ought" confusions in which empirical descriptions of nature are seen as dictating moral conclusions" (Friedrich, 2005, p.59).

4 For Hawthorne effect, see Barnes, 2010; Benedetti, Carlino & Piedimonte, 2016; McCarney et al., 2007.

can observe differences questioning the epistemic status of big data. In an autobiography, usually the author is in a self-confirmation mode. That is why autobiographical accounts of the same events by different authors have the potential to clash with each other.

As the twin big brothers' surveillance will be more biting, we will see more people faking their data or avoiding the twins by the assistance of counter-surveillance measures. In mass revolts, it will be no surprise to see that the first casualty would be the CCTVs. Anonymous use of internet will become more common to defend the right to privacy. All these efforts will feed big data as self-misrepresentations that will go mostly unnoticed. Thus, big data has also been including internet users'-citizens'-consumers' faked data.

5. Discussion: A Proposal for Datagraphy as the Historiography of Big Data

Just like the discussions about the notion of historiography, in such cases, who produces and records the data becomes important. This brings up the possibility that, contrary to the champions of big data, the data can be subjective. Human life is more complicated than computational models of mind: North Korean, South Korean, Chinese and Japanese historiographies narrating the same events in a completely different way (much like the Rashomon effect[5]) will be parts of big data based on the official ideology of each country (cf. Beal, Nozaki, & Yang, 2001; Bukh, 2007; Schmid, 2000; Schneider, 2008; Seo, 2008). Which account is correct is a moot issue. In that sense, we can propose a new research area, datagraphy which would study how data is produced, recorded, processed, presented etc. by whom, with what purpose, with what kind of a methodology to avoid biases, with what biases or errors.[6]

Historical studies started with the conceptualization of history, then proceeded to historiography, and that is the current site of historical discussion. Originally, the idea of the history was to narrate what happened in history as it is. So it was a sort of a story, but a realistic one (Iggers, 1997). But historiographical accounts emerged from the fact that history is not such an area: Who writes the history and for whom determines what is written in

5 For Rashomon effect, see Heider, 1988; Roth & Mehta, 2002; Sanahuja, 2013.

6 The term 'datagraph' is not a new term in academic research, however it is used in a different way, by referring to data and graph (e.g. Batarfi et al., 2015; Estabrook & McMorris, 1977; Gottron, Knauf & Scherp, 2015; Neumann & Weikum, 2010; Qiu & Hancock, 2006). Contrary to this usage, we develop this idea on the basis of our analogy of history and historiography. In this new sense, the first time datagraphy is presented and discussed is in this current article.

the name of history and in what forms, distortions and manipulations. For example, Maoist Chinese history was narrating the story of an assassin (Jing Ke, 荊軻, ?- 227 B.C.) who tried to kill the Chinese emperor, as people's hero, but post-Mao China narrates the same figure as a traitor against the 'unity of Great Middle Kingdom' (i.e. China) (cf. Pines, 2008; Rawnsley, 2007). Likewise, conflicts have various sides: Liberation of a city can be considered as the fall of the city (e.g. Saigon). Conquest of a city can be considered as a loss of the city (e.g. Constantinople). A freedom fighter can be considered as a terrorist from another account. So the question is by whom and for whom will the data be recorded, for what purpose. The answer is clear: The data will be recorded by governments and corporations from their perspective, to control and to make profit respectively. So unlike the statistical models trying to mitigate the error rate in big data, the big data is corrupt from the very beginning.

From epistemological and ontological points of view, one of the major assumptions behind big data is false. That assumption rests on the outdated notion of correspondence theory of truth.[7] When we human beings represent something in real life, we can never record and represent as it is / they are. That is because we more or less add our interpretation. We select and unselect certain phenomena. For example, when we observe a group of people, and aim to record let's say the number of young people, we can miss other age groups. Our observation also depends on how we define age groups. For some, 40 is the last year of youth, for other countries where longevity is limited, this is extreme. They would prefer younger ages as the maximum value. There is no historical standard either. Nowadays people live longer, but the average longevity was quite low in the past centuries. That may also imply that, if this trend continues, the last age of youth may be extended.[8] Furthermore, let's suppose that we agreed on 40 as the top limit. But some cultures (especially Asian cultures) consider the newborn as one year old, counting the period from conception to birth as an age, while according to other cultures (especially globally Western cultures) newborn's life starts by Age 0. Thus, 40 year-olds in the Western system will be 41 year-olds in the Asian system, and will accordingly be excluded. So there is not even an agreement about age groupings at a universal and ahistorical level.

---

7 For correspondence theory of truth, see Kuukkanen, 2007; Lewis, 2001; Patterson, 2003.

8 For discussions of historical and demographic accounts of human longevity see Gurven & Kaplan, 2007; Wilmoth, 1998, 2000

We don't just observe as it is, as discussed in philosophy of science. It may be claimed that the machines are not as vulnerable as human beings in recording real life. That is not true. Because machines are also bounded by the assumptions and principles set by their designers and relevant industries. For example, a big data collection device, in fact, does not record everything. Some data are worth recording, some others are not. For instance, when medical data are collected by automated machines, a particular group of indicators that are not considered to be associated with diseases are ignored. But those uncollected data can unravel the factors behind the illnesses. Ditto for other transactions such as shopping. People can shop at traditional markets if they don't want to be identified during or after purchase. Thus, big-datafication rather than covering everything, can end up covering formal economy only and enormously contribute to the rise of informal economy. The internet of things will expand the attacks against privacy even worse by recording everything we do at home as well, including how many times we take a bath, clean the house, cook etc. (cf. Caron et al., 2016; Winter, 2014, Zuboff, 2015).

In fact, the rise of big data is completely against public health and public interest in general. For instance, insurance companies and governments can collaborate to blacklist those who are expected to live shorter. For example, through the big data collected by various sources, the insurance companies can identify who drinks and smokes, who has a less regular life etc. This will have ideological implications at the hands of corrupt, conservative or health-obsessed governments. Hand in hand with insurance companies, by identifying who does not vote for them, they can deny certain citizenship services such as healthcare, public transports, and even passports for those blacklisted. The Chinese social credit system which involves classifying all citizens on the basis of all big data collected through various means shows that this is not science fiction (see Creemers, 2018; Kotska, 2018). Welcome to the future! Even worse than these, since the files compiled for each citizen are not transparent, there is no way to object to or appeal what is written. Legal checks and balances are eliminated. Which means this double surveillance system is more error-prone than the more traditional ways of government control and corporate manipulation.

In this article, we have showed multiple ways through which big data can be misrepresented and thus wrong. Based on misrepresented big data, the citizen files will be misrepresented as well, which will have enormously negative implications for citizens and society in general. Accidentally or with a valid explanation such as crime, those blacklisted will be unfairly treated

anywhere they go and any time, thanks to a system all the time reminding the individual and society that Citizen A is on the black list.

6. Conclusion: Omniresistance

In summary, in order to identify the intentional and unintentional misrepresentations of the twin big brothers and intentional self-misrepresentations of internet users/consumers/citizens, we propose the new area of datagraphy which studies intentional and unintentional misrepresentations and intentional self-misrepresentations while producing, recording, storing, deciphering, mining, presenting, processing and interpreting big data. Such a new research area is definitely needed against the unrealistic enthusiasm of big data optimists which would likely have negative consequences for citizens.

Against the omniscience, omnipotence and omnipresence of the twin big brothers, their big data and surveillance based on intentional and unintentional misrepresentations, the truth defenders will need to be omniresistant. That is, as they will be surveilled everywhere and all the times by all means, they have to resist everywhere, all the times and by all means. Our prediction is the following: As a way of resistance against big-datafication and surveillance, the informal economy and informal life in general will be expanded. Historically speaking, when feudalism and later on capitalism had full control over the cities, the unlawful bandits moved to mountains where they could enjoy freedom. Some of them were bloody, ruthless criminals, but some others emerged as people's heroes (cf. Hobsbawn, 1969). We expect that there will be metaphorical mountains in our times, or let's say pockets of unsurveilled freedom or free islands. This time, the fight will be against surveillance for equiveillance, whereby the citizens will have the mechanism to watch the watchers which will remind them that their powers are not unlimited.

Hopefully, one day big data will be for people, for democratization, for public interest, for social welfare, for truly scientific advancement abiding with firm scientific methods and principles rather than for twin big brothers, control and profit making. But we need to realize that under the current conditions of the surveillance capitalism, this is just a pipe dream. Nevertheless, we need to "be realistic and demand the impossible".[9]

Declaration of Conflicting Interests

The author has no conflicting interests. No funding was received for this work.

9 A quote from Che Guevara (1928-1967), popularized by the 1968 movements.

References

Barnes, B. R. (2010). The Hawthorne Effect in community trials in developing countries. *International Journal of Social Research Methodology*, *13*(4), 357-370.

Batarfi, O., El Shawi, R., Fayoumi, A. G., Nouri, R., Barnawi, A., & Sakr, S. (2015). Large scale graph processing systems: survey and an experimental evaluation. *Cluster Computing*, *18*(3), 1189-1213.

Beal, T., Nozaki, Y., & Yang, J. (2001). Ghosts of the past: the Japanese History Textbook controversy. *New Zealand Journal of Asian Studies*, *3*, 177-188.

Benedetti, F., Carlino, E., & Piedimonte, A. (2016). Increasing uncertainty in CNS clinical trials: the role of placebo, nocebo, and Hawthorne effects. *The Lancet Neurology*, *15*(7), 736-747.

Bukh, A. (2007). Japan's history textbooks debate: National identity in narratives of victimhood and victimization. *Asian Survey*, *47*(5), 683-704.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Breur, T. (2016). US elections: How could predictions be so wrong?. *Journal of Marketing Analytics, 4*(4), 125-134.

Buelens, B., Daas, P., Burger, J., Puts, M., & van den Brakel, J. (2014). Selectivity of Big data. Accessed http://www. pietdaas. nl/beta/pubs/pubs/Selectivity_Buelens. pdf.

Caron, X., Bosua, R., Maynard, S. B., & Ahmad, A. (2016). The Internet of Things (IoT) and its impact on individual privacy: An Australian perspective. *Computer Law & Security Review*, *32*(1), 4-15.

Civaner, M. (2012). Sale strategies of pharmaceutical companies in a "pharmerging" country: The problems will not improve if the gaps remain. *Health Policy*, *106*(3), 225-232.

Couper, M. P. (2014). What big data may mean for surveys. In Proceedings of Statistics Canada's 2014 International Symposium on Methodological Issues, Gatineau, Quebec. Accessed

http://www.statcan.gc.ca/eng/conferences/symposium2014/program/14272-eng.pdf

Crawford, K., Gray, M. L., & Miltner, K. (2014). Critiquing Big Data: Politics, ethics, epistemology| special section introduction. *International Journal of Communication*, *8*, 1663-1672.

Creemers, R. (2018). China's Social Credit System: An evolving practice of control. Accessed http://www.iberchina.org/files/2018/social_credit_china.pdf

Estabrook, G. F., & McMorris, F. R. (1977). When are two qualitative taxonomic characters compatible?. *Journal of Mathematical Biology*, *4*(2), 195-200.

Friedrich, J. (2005). Naturalistic fallacy errors in lay interpretations of psychological science: Data and reflections on the Rind, Tromovitch, and Bauserman (1998) controversy. *Basic and Applied Social Psychology*, *27*(1), 59-70.

Gottron, T., Knauf, M., & Scherp, A. (2015). Analysis of schema structures in the linked open data graph based on unique subject uris, pay-level domains, and vocabulary usage. *Distributed and Parallel Databases*, *33*(4), 515-553.

Gøtzsche, P. C., Kassirer, J. P., Woolley, K. L., Wager, E., Jacobs, A., Gertel, A., & Hamilton, C. (2009). What should be done to tackle ghostwriting in the medical literature?. *PLoS Medicine*, *6*(2), e1000023.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Gurven, M., & Kaplan, H. (2007). Longevity among hunter-gatherers: a cross cultural examination. *Population and Development review*, *33*(2), 321-365.

Harford, T. (2014). Big data: A big mistake?. *Significance*, *11*(5), 14-19.

Heider, K. G. (1988). The Rashomon effect: When ethnographers disagree. *American Anthropologist*, *90*(1), 73-81.

Hobsbawn, E. (1969). *Bandits*. London: Weidenfeld & Nicolson.

Hsu, Y. H., Fang, W., & Lee, Y. (2009). Ethically questionable behavior in sales representatives–An example from the Taiwanese pharmaceutical industry. *Journal of Business Ethics*, *88*(1), 155.

Iggers, G.G. (1997). *Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge*. London: Wesleyan University Press.

Ijoma, U., Onwuekwe, I., Onodugo, O., Aguwa, E., Ejim, E., Onyedum, C., Onah, L., Okwudire, E. & Ugwuonah, G. (2010). Effect of Promotional Strategies of Pharmaceutical Companies on Doctors' Prescription Pattern in South East Nigeria. *TAF Preventive Medicine Bulletin, 9*(1), 1-6.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and Translational Science*, *7*(4), 342-346.

Kaur, P., & Arora, S. (2015). Regression and Endogeneity Bias in Big Marketing Data. *Procedia Computer Science*, *70*, 41-47.

Khanna, R. (2015). Big Data= Big Brother?. Accessed

http://law.emory.edu/ecgar/_documents/volumes/2/1/khanna.pdf

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, *80*(4), 463-475.

Klick, J., & Stratmann, T. (2007). Abortion access and risky sex among teens: parental involvement laws and sexually transmitted diseases. *The Journal of Law, Economics, & Organization*, *24*(1), 2-21.

Kotska, G. (2018). China's social credit systems and public opinion: Explaining high levels of approval. Accessed https://www.researchgate.net/profile/Genia_Kostka/publication

Kuukkanen, J. M. (2007). Kuhn, the correspondence theory of truth and coherentist epistemology. *Studies in History and Philosophy of Science Part A*, *38*(3), 555-566.

Laterza, V. (2018). Cambridge Analytica, independent research and the national interest. *Anthropology Today*, *34*(3), 1-2.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 2053951714534395.

Lewis, D. (2001). Forget about the 'correspondence theory of truth'. *Analysis*, *61*(272), 275-280.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 134-142.

Lu, J., & Li, D. (2013). Bias correction in a small sample from big data. *IEEE Transactions on Knowledge and Data Engineering*, *25*(11), 2658-2663.

Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible?. Advances In Classification Research Online, 24(1), 4-15.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology*, *7*(1), 30.

McKay, A., & Barrett, M. (2010). Trends in teen pregnancy rates from 1996–2006: A comparison of Canada, Sweden, USA and England/Wales. *Canadian Journal of Human Sexuality*, *19*(1-2), 43-52.

Moffatt, B., & Elliott, C. (2007). Ghost marketing: pharmaceutical companies and ghostwritten journal articles. *Perspectives in Biology and Medicine*, *50*(1), 18-31.

Moore, T. J., & Furberg, C. D. (2015). Electronic health data for postmarket surveillance: a vision not realized. *Drug Safety*, *38*(7), 601-610.

Neumann, T., & Weikum, G. (2010). The RDF-3X engine for scalable management of RDF data. *The VLDB Journal—The International Journal on Very Large Data Bases*, *19*(1), 91-113.

Patterson, D. (2003). What is a correspondence theory of truth?. *Synthese*, *137*(3), 421-444.

Pines, Y. (2008). A Hero Terrorist: Adoration of Jing Ke Revisited. *Asia Major*, 21(2), 1-34.

Price, M., & Ball, P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Review of International Affairs*, *34*(1), 9-20.

Qiu, H., & Hancock, E. R. (2006). Graph matching and clustering using spectral partitions. *Pattern Recognition*, *39*(1), 22-34.

Rawnsley, G. D. (2007). The Political Narrative (s) of Hero. *Media Asia*, *34*(1), 20-26.

Roth, W. D., & Mehta, J. D. (2002). The Rashomon effect: Combining positivist and interpretivist approaches in the analysis of contested events. *Sociological Methods & Research*, *31*(2), 131-173.

Sanahuja, J. A. (2013). Narrativas del multilateralismo:«efecto Rashomon» y cambio de poder/Narratives of multilateralism:" Rashomon effect" and change of power. *Revista Cidob d'afers Internacionals, 101*, 27-54.

Schmid, A. (2000). Colonialism and the 'Korea problem'in the historiography of modern Japan: A review article. *The Journal of Asian Studies*, *59*(4), 951-976.

Schneider, C. (2008). The Japanese history textbook controversy in East Asian perspective. *The Annals of the American Academy of Political and Social Science*, *617*(1), 107-122.

Seely-Gant, K., & Frehill, L. M. (2015). Exploring Bias and Error in Big Data Research. *Washington Academy of Sciences. Journal of the Washington Academy of Sciences*, *101*(3), 29-37.

Seo, J. (2008). Politics of memory in Korea and China: Remembering the comfort women and the Nanjing massacre. *New Political Science*, *30*(3), 369-392.

Smith, R. (2005). Medical journals are an extension of the marketing arm of pharmaceutical companies. *PLoS Medicine*, *2*(5), e138.

Tarran, B. (2018). What can we learn from the Facebook—Cambridge Analytica scandal?. *Significance*, *15*(3), 4-5.

Tuttle, H. (2018). Facebook Scandal Raises Data Privacy Concerns. *Risk Management*, *65*(5), 6-9.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197-208.

Werner, R. M., Alexander, G. C., Fagerlin, A., & Ubel, P. A. (2004). Lying to insurance companies: The desire to deceive among physicians and the public. *The American Journal of Bioethics*, *4*(4), 53-59.

Wilmoth, J. R. (2000). Demography of longevity: past, present, and future trends. *Experimental gerontology*, *35*(9-10), 1111-1129.

Wilmoth, J. R. (1998). The future of human longevity: A demographer's perspective. *Science*, *280*(5362), 395-397.

Winter, J. S. (2014). Surveillance in ubiquitous network societies: normative conflicts related to the consumer in-store supermarket experience in the context of the Internet of Things. *Ethics and Information Technology*, *16*(1), 27-41.

World Bank (2018). Individuals using the Internet (% of population). Accessed

https://data.worldbank.org/indicator/IT.NET.USER.ZS

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75-89.