# Theoretical Bases of Critical Data Studies

Łukasz Iwasiński
0000-0003-2126-7735
*Department of Information Studies,*
*Faculty of Journalism, Information and Book Studies,*
*University of Warsaw, Poland*

**Abstract**

**Purpose/Thesis:** The paper presents main premises and analyzes the theoretical bases of critical data studies (CDS).
**Approach/Methods:** The article uses critical review of the literature on CDS, social aspects of big data, sociology of knowledge, philosophy of knowledge and science and technology studies.
**Results and conclusions:** Author identifies three main theoretical premises of CDS: (1) A critique of market-oriented instrumental rationality; (2) Rejection of the idea that data is independent from the research process; (3) Rejection of the concept of raw data. Article discusses intellectual roots of CDS. It is argued that CDS derive from constructivist sociology of knowledge, and science and technology studies.
**Originality/Value:** The article brings together theoretical literature and empirical studies from diverse disciplinary fields to examine theoretical bases of CDS and situates it in its intellectual context. It stresses the need of critical view of data and data processing, which is especially important in the big data area. CDS are recognized in cultural studies and media studies (however poorly discussed in related Polish scholarship), but they remain almost absent in Information Studies, which would benefit from it.

**Keywords**

Big data. Critical data studies. Datafication. Instrumental rationality. Social constructivism. Sociology of knowledge.

*Received: 5 June 2020. Reviewed: 16 June 2020. Accepted: 30 September 2020.*

## 1. Introduction

Simply put, critical data studies (CDS) apply critical theory, or, more generally, a critical intellectual attitude, to data. CDS is an attempt to consider data in all its aspects, and to show that their significance extends beyond the technical and the epistemological to the cultural, the social, the economical, the ethical, and the political.

We witness an increasing quantification of reality. Qualitative aspects of the world are reduced to numbers, subject to mathematical processing. Ian Hacking writes that the development of statistics in the nineteenth century made numbers into a fetish. Statistics introduced new styles of reasoning; it imposed new categories on reality, especially on people, transforming the organization of social life and facilitating its surveillance (Hacking, 1990; 1991). This process continues – the possibility of an increasingly precise measurement of further aspects of social life, and of lives of individual persons, changes

our world. It allows the state and the market to monitor their subjects and in the case of self-tracking, it enables auto-surveillance (Iwasiński, 2017) – an operation to a large extent subordinated to the logic and needs of contemporary capitalism (Wróblewski, 2016). Nowadays, we pursue a reduction of all aspects of reality to a sequence of quantitative data (Iwasiński, 2016; Szpunar, 2019). Lev Manovich suggests that we treat contemporary world as a data base (Manovich, 2012, 355). Viktor Mayer-Schonberger and Kenneth Cukier (2013) term this phenomenon "datafication". Other researchers seeking metaphors to describe contemporary reality refer to "metric fixation" (Muller, 2018), "metric culture" (Ajana, 2018), or "data-driven life" (Wolf, 2010). Jose van Dijck uses the term "dataism" in reference to the ideology premised on the assumption that data is the most appropriate means to understanding human behavior (Dijck, 2014, 197–208). It could be said that the ideology of dataism affirms that a mathematical analysis of data is the most effective method of optimizing any and all actions and processes undertaken by people, ensuring the greatest control over reality, the greatest objectivity of its view, and the best decisions. CDS questions the validity of this position.

I would like to highlight three basic premises of CDS, from which all more detailed assertions of the critical data study follow. Firstly, CDS opposes the supremacy of the market-oriented instrumental rationality, driven by datafication. Secondly, it rejects the assumption that data is separate from the process of cognition. Thirdly, it assumes that there are no raw data, i.e., that the concept of "raw data" has no immanent sense but that their sense is always contextual. Before I proceed to a more detailed discussion of these aspects of CDS, I will briefly summarize its history.

At the same time, I would like to make clear that I have no intention of negating the worth and use of research based on data, including big data. It is an important – perhaps, the most important – means to acquiring knowledge. However, the narrative celebrating research based on quantitative data would be incomplete without a critique and a discussion of its bases. In the last few years, the need for such a critical perspective became increasingly obvious[1].

## 2. Critical data studies: a history

The indirect sources of CDS lie in critical theory and the poststructuralist/postmodernist thought; direct – in constructivist sociology of knowledge (as opposed to the classical) and in sociology of scientific knowledge deriving from it, as well as in science and technology studies (STS). These fields constitute CDS's intellectual background. Critical theory developed by the Frankfurt School in the 1930s questioned the positivist view of science; it rejected the division between the subject and object of observation, underlined the normative quality of knowledge – including scientific knowledge – and highlighted the negative aspects of instrumental rationality. In the 1960s, this approach was radicalized by poststructuralism/postmodernism, cultural studies, and more specific sub-disciplines: media studies, feminism, queer theory, postcolonialism and others. In 1970s, it entered

[1] In 2016, the journal *Big Data & Society* published an issue devoted entirely to CDS, see: *Big Data & Society* 3(2), 2016.

research concerned with scientific knowledge and with technology, and enriched by new concepts (the "strong programme" of Edinburgh School and Bath School, actor-network theory [ANT], sociology of statistics, including critical statistics promoted by the Radical Statistics group). It was in this spirit that Siva Vaidyanathan (2005) proposed Critical Information Studies – almost a decade before the term "critical data studies" was introduced. All these disciplines gestured towards the ideological and political complicity of all knowledge; they exposed science and technology as tools maintaining existing power dynamics, and highlighted their own liberating potential. CDS shares these aims.

The emergence of big data and datafication phenomena inspired critical reflection on data. Although big data is the main subject of CDS analysis, CDS is not confined to the issue of big data, even if it is in relation to big data that CDS's central concerns may be viewed with greatest clarity. Critical reflection on big data has been developing for more than a decade[2]. Chris Anderson in his opinion piece from 2008 argued that, in a datified world, where we may use big data technologies to describe reality and predict its future states, theory is no longer necessary, because data speaks for itself:

> Petabytes allow us to say: "Correlation is enough". We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot (...). Correlation supersedes causation, and science can advance even without coherent models, unified theories (...). There's no reason to cling to our old ways (Anderson, 2008).

Anderson's views resonated with the world of business, and opinion journalism responding to developments in science. Marc Prensky wrote that

> [s]cientists no longer have to make educated guesses, construct hypotheses and models, and test them with data-based experiments and examples. Instead, they can mine the complete set of data for patterns that reveal effects, producing scientific conclusions without further experimentation (Prensky, 2009).

Andy Clark stated that big data analysis removed "the human element (...) and, as such, all the human bias that goes with it" (Clark, 2013). However, the academic community questioned these claims. Rob Kitchin's summary of the position shared by Anderson and his supporters is worth citing here in full:

- Big Data can capture a whole domain and provide full resolution;
- there is no need for a priori theory, models or hypotheses;
- through the application of agnostic data analytics the data can speak for itself free of human bias or framing, and any patterns and relationships within Big Data are inherently meaningful and truthful;
- meaning transcends context or domain-specific knowledge, thus can be interpreted by anyone who can decode a statistic or data visualization (Kitchin, 2014).

Kitchin argues that all these claims are false. Briefly, his argument was as follows: firstly, big data analysis always relies only on a part of the potentially available data, and its results constitute only one of the possible images of a given fragment of reality, informed by the quality of the data, the method of its processing, and technology; as such, it is susceptible to bias. Secondly, the supposedly flawless inductive modelling of the world with the help

[2] The most important critics of big data have a background in media studies, sociology, social geography and mathematics; they include Rob Kitchin, Jim Thatcher, Craig M. Dalton, danah boyd (styled lowercase), Kate Crawford, Tracey P. Lauriault, Lisa Gitelman, Stefania Milan, Wendy Hui Kyong Chun, Deborah Lupton, Cathy O'Neil.

of big data does not occur in a theoretical vacuum – data is collected by instruments constructed with reference to theory, according to a methodology based on theory, and processed according to theory or scientific laws (cf. Frické, 2015). Thirdly, as the previous two points show, data does not "speak for itself":

> Data itself does not speak. What is required is a huge amount of background knowledge, or assumptions, or prior research of one kind or another (Fricke, 2015).

We should add that the results of big data analysis are not necessarily intelligible and unambiguous; in fact, they require interpretation (e.g. to prevent apophenia, i.e., the mistaken perception of connections where there are none). Fourthly, it is a specific discourse, rather than a mathematical operation, that invests data with meaning.

Big data processing is useful, and effective when applied instrumentally, e.g. in predictive market analysis. However, it is not sufficient to explain the causes of a given phenomenon, or to illuminate its significance. Kitchin (2014) states that while data may allow us to identify a pattern, it cannot explain it. Interpretation requires theory and knowledge of context. David Sumpter, a professor of mathematics specializing in big data processing, particularly in the analysis of collective behavior, observes that

> (...) when it comes to understanding the world around us, mathematical models don't usually beat humans ... While computers are very good at collecting large numbers of statistical measures, humans are very good at discerning the underlying reasons for these measures (Sumpter 2019, 90–91).

In an article published in a 2012 issue of *Critical Questions for Big Data*, danah boyd and Kate Crawford proposed six "provocations", with the aim of "sparking conversations" about the issues of big data: (1) Big Data Changes the Definition of Knowledge; (2) Claims to Objectivity and Accuracy are Misleading; (3) Bigger Data are Not Always Better Data; (4) Taken Out of Context, Big Data Loses its Meaning; (5) Just Because it is Accessible Doesn't Make it Ethical; (6) Limited Access to Big Data Creates New Digital Divides.

It would be difficult to agree that big data indeed changes the definition of knowledge. It offers new methods of knowledge formation, however, it does not undermine, contrary to what Chris Anderson and others declare, the premises of the previous research methodology. Points (2), (3), and (4) derive from the premises of CDS discussed above, namely its critique of the assumption that data is separate to the process of cognition, and its rejection of the concept of "raw data", whose meaning would be independent from research context. These premises are discussed in more detail below. The last two points, (5) and (6), gesture towards ethical dilemmas of big data analysis. Big data specialists occasionally follow the "capture all" principle, which dictates that they should collect and preserve all available data, as it will allow them to analyze any phenomenon they wish. It is easy to misuse data from the large bases, e.g., intruding on the privacy of a person to whom the data pertains. Even if anonymity is maintained, aggregation of data from many sources allows its deanonymization (Villasenor, 2011; Waszewski, 2015, 245). The increase of data it is possible to capture and the datafication of an expanding part of our lives create new ethical issues, related to the intrusion of privacy, digital surveillance and the possibility of manipulative profiling. The last point, (6), says that data is not equally accessible to every user – it is much easier to access for internet companies. Easy access to data gives these companies a massive advantage, and marginalizes those who lack access to data and to the

tools to analyze it. A different aspect of digital divide may be observed as different people are not equally subject to datafication. Kate Crawford (2013) says that

> [d]ata are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities. (...) With every big data set, we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big datasets?

Do such people choose to escape datafication, or not? Do they lose by that, and if so, how? What do they gain? It seems that some people, aware of the risks posed by datafication, seek to consciously escape it; for others, who did not choose to remain outside its reach, their position outside the realm of datafication may be a source of difficulties, as it may prevent them from benefitting from digital services.

The term "critical data studies" emerged after boyd and Crawford published their 2012 article. Craig M. Dalton and Jim Thatcher are agreed to have coined it. They used in their 2014 article, *What does a critical data studies look like, and why do we care?*, published in an online edition of the *Society & Space* journal. They showed that there were many reasons to identify "critical data studies" as a separate field. Firstly, they point to the increasing role of big data in the contemporary world; furthermore, they observe that data is never raw, and that big data analysis is never neutral, i.e., devoid of cultural, social, and political leanings. They argue that big data techniques have consequences for the society, affect human behavior, shape social dynamics, and inevitably influence various spheres of social life. According to them, the goal of CDS is to expose ideological agendas hidden in data itself, and in the operations conducted on them. To realize it, we must combine big data techniques with research based on "small data", i.e. data which may be analyzed and interpreted by a single individual, allowing an in-depth qualitative description.

## 3. Critical data studies and market-oriented instrumental reason

The increasing prominence of instrumental rationality is a part of the general progress of modernity. Its nature is selection of an optimal means to a given goal (Sztompka, 2003, 57, 65). Actions suggested by instrumental rationality were proven the most effective in controlling reality, manipulating its elements to achieve certain benefits, and predicting its future states. Instrumental rationality is fundamental for technology, but with the progress of modernity, it begins to be applied in other spheres of social life. In other words, an increasing number of areas of life is subject to the rule of technocracy (Habermas, 1977; 1983; Zybertowicz, 2015b, 54). The risks of the domination of instrumental rationality over social life were observed as early as in 1940s, by the founders of critical theory represented by the Frankfurt School. Max Horkheimer, a key member of the School, argues that the development of instrumental rationality leads towards an increasing instrumentalization of the world and of the human, which can only result in the man's enslavement:

> As a result of the development of technical knowledge, the autonomy of the individual subject decreases, and his power to resist the growing apparatus of mass manipulation, his imagination and independent judgement weaken. The development of technical means is accompanied by dehumanization (Horkheimer, 1987, 245),

and

> The more apparatuses to tame nature we invent, the more we must serve them if we wish to survive (Horkheimer, 1987, 245).

Datafication is closely related to instrumental rationality; certainly it facilitates an extension of instrumental reason's dominance. After all, it is easier for instrumental reason to govern what is quantitative. Therefore, the more quantified the reality, the easier it is to follow the rules of the instrumental reason. This is particularly relevant to digitized objects (objects in digital form). In the mid-1990s Nicholas Negroponte (1997, 13–18) wrote that an increasing number of objects was digitized, atoms turned to bites. Material goods, or, more specifically, digital representations of material goods are moved into the virtual world, where they are much easier to modify. Andrzej Kiepas (2017, 39) observes that "digital objects are much easier to manipulate than analog objects." Wiesław Godzic (1998) writes that a digital world is a tamed world. Every element of such a world may be reduced to its basic components (data), subject to analysis, modified, transferred. As far as instrumental reason is concerned, it would be ideal if the entirety of reality were digitized, so that all problems might be solved by strictly mathematical operations[3]. In such a world, even the individual persons would be reduced to data sets and parameters, and their relationships – even their relationships with themselves, i.e., auto-reflection – would be realized by mathematical formulas. Several scholars suggest that this is precisely the direction in which we are headed. Zybertowicz (2015a, 449) writes that

> widely defined postmodernism achieved a conceptual, philosophical, intertextual deconstruction of the subject. Today we face the next step: new technologies make possible not only a conceptual, but also a practical, entirely literal technical transformation, and even a dismantling of the human person.

Jan Waszewski states that

> the assumption that data bases contain the entirety of a human being, his beliefs, personality, motivations, moods, future behaviors and so on, is a key element of the majority of analyses based on the Big Data technology (Waszewski, 2015, 255).

The rapidly developing self-tracking technologies, supported by the idea of Quantified-Self, dating platforms based on parametrization, or Chinese social credit may all be seen as substitutes of, or first steps towards, the fully digitized world.

According to Horkheimer (1987), instrumental rationality discloses the most efficient methods of realizing imposed goals, but it does not account for the reflection on these goals and the values which underlie them. It does not make space for any normative rules. Following the principles of instrumental reason alone, it is impossible to prove that "justice and freedom are by themselves better than injustice and bondage" (Rudziński, 1987, 11). Instrumental rationality deprives people of a part of their reason – the part responsible for value judgements, assessments in moral categories, or questions regarding meaning. It allows one to determine the most economically profitable solution of a given problem, but the decision to prioritize economical profitability, as opposed to preserving health of the

[3] It is an utopian vision for many reasons, but particularly because it would be impossible to mathematize fundamental dilemmas and conflicts of social life resulting from disparities of values and interests.

individuals or of the environment, and so on, goes beyond instrumental rationality. Such awareness belongs to the realm of normative knowledge beyond the reach of instrumental rationality; it occurs in the sphere of values and meanings.

Who then, or what, decides what we should aim for? In contemporary capitalism, a key mechanism for determining goals to be realized by the society is the market[4]. Therefore actions geared towards optimizing the accumulation of profit are prioritized. Datafication of reality facilitates the subordination of further spheres of life to market's rule, as quantification promotes valuation. Furthermore, the development of information and communication technologies, and related economic networkization fosters the global expansion of capitalism – even if it does not determine it (Szumlewicz, 2005, 175). Therefore, it should not come as a surprise that big data analysis primary object is to establish economic value derived from data. IDC, a firm offering big data analysis for markets declares:

> Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis (Villars et al., 2011).

Mathematics and the market have long been connected:

> Scholarly work on early modern mathematics in its cultural context has ably demonstrated the relationship between evolving protocapitalist market economies and increasingly ubiquitous mathematical discourses related to mercantile activity (Williams, 2013, 41; cf. Brine & Poovey, 2013).

Georg Simmel (2012) observed it as early as in 1900 in his famous study *The Philosophy of Money*. Under the late datacentric capitalism, which emphasizes individualism and flexible management, hegemony of instrumental reason does not pose a risk of the progressing standardization of the world and people, prophesized by the classical critical theorists of the Frankfurt School. Instead, it threatens to subordinate further spheres of social life to market logic, and a resulting commodification of the datified man. After all, the currency with which we pay for various services is data we generate; therefore, we do not pay with money, but with our privacy, sharing information regarding our interactions, relationships, views, interests, thoughts and feelings, with market subjects. In the light of CDS, our life – our very identity – becomes subject to market's exploitation. Furthermore, reduced to data, "mapified" man or social group becomes an easy target for manipulation – as evidenced by

[4] Obviously, it is not the only mechanism. However, as Kazimierz Krzysztofek observes, the aims suggested by social institutions, such as the family, school, or Church, erode "insofar as the patterns of control they establish by promoting social roles are incompatible with the individual's performance of his role as a producer and a consumer". Furthermore: "Why not remake Acropolis into a hotel with a McDonald's, Colosseum into a stadium and a gym; why not clean up the ruins of Forum Romanum and erect there a Hiperforum mall, re-make Troy into an antique theme park, and Cheops's pyramid into a Cheops Beer pub. It would certainly be rational. Fortunately, nobody thinks in these categories, even the most ardent supporters of the market; fortunately culture with its ethical and religious systems provides us with some restraints". However, the market is capable of accommodating the needs generated by other, cultural systems. The author goes on to ask if the cultural imperative to protect heritage is rational from the market's point of view: "Yes and no. It is irrational, as it stops its [the market's – ŁI] expansion; it is rational as without preserving heritage, there would be no tourism, which is nowadays perhaps the largest business in the world; the ruins of Forum Romanum generate more profit that Hiperforum mall ever would" (Krzysztofek, 2000, 125).

the results of political microtargeting and neuromarketing (Hegazy, 2019). Such practices interest the CDS researchers.

## 4. Critical data studies and the assumption of the data's separation from the research process

The proponents of CDS reject the assumption that data is independent from the research process. It comes from their opposition to the objectivist model of knowledge, premised on the assumptions that reality consists of objects independent from the mind, and that its full, unambiguous and true description is possible (Zybertowicz, 1995, 72–77; cf. Szahaj, 2014, 213). Objectivist model of knowledge seems intuitive, and is implicitly accepted by researchers who believe that by conducting appropriate operations on data, they uncover facts and adequately reconstruct the properties and qualities of the world. CDS, however, has its basis the constructivist model of knowledge. It questions the separation of two spheres: the sphere of reality, defined by qualities pre-existing and independent from the process of knowing, and the sphere of knowledge, which accurately and objectively reflects such reality (Zybertowicz, 1995, 101). According to this model, knowing of the object changes it. This idea is the basis of postpositivism (understood as an epistemological position), characteristic of poststructuralism/postmodernism (understood as a general approach to cultural analysis). It is a foundation of constructivist sociology of knowledge and its derivatives: sociology of scientific knowledge, as well as science and technology studies (STS). Postpositivism and poststructuralism/postmodernism problematize (or reject) the idea of representationalism, which claims that it is possible to create unambiguous models of reality, or its fragment, basing on the methods of cognition available to us. Tomasz Szkudlarek and Zbyszko Melosik introduce,

> the concept of (re)presentation – I write it thus because no expression of the world is not, and cannot be a mirror reflection of it; it is always a presentation – shaped by the dynamics of knowledge/power, interpretation, biography ... In its attempt to present reality, representation – itself an integral part of reality – creates it. It constitutes an indispensable and substantial element of the dynamic of reality's existence (Szkudlarek & Melosik, 1998, 42).

According to the objectivist model, knowledge is uncovered, and according to the constructivist model, it is always constructed. The former model assumes that, correctly conducted, with the use of rigorous procedures, research uncovers the truth (as defined by the classical correspondence theory); the latter – that research presents only one of the few possible interpretations of a given fragment of reality (with the problematized understanding of truth, the term used only in the non-classical sense – constructivist, pragmatic, or as defined by the coherence theory of truth).

Donald MacKenzie (1978), using as an example the research of Karl Pearson and George Udny Yule, and the discussion between the two, argues that the development of seemingly abstract, rigid field detached from a socio-cultural background, such as the measurement of statistical relations, was informed by cognitive interests resulting from various interests of the groups from which these scholars came. If we follow MacKenzie's argument, we must admit that even the research of statistical formulas is not ideologically neutral, and its result depends on sociocultural factors. CDS states that facts and data is socially

constructed, rather than objective – which I discuss in more detail below. According to CDS, creating knowledge is a social process, in which values, convictions, and interests all play a role (to use the term of Florian Znaniecki, it is characterized by the humanistic coefficient)[5]. It could be said that a critical, constructivist understanding of knowledge – which is a part of CDS – always involves reflecting on the society (Zybertowicz, 1995, 94). It is not surprising then that a major inspiration and one of the main intellectual sources of CRS is non-classical (constructivist) sociology of knowledge and related disciplines (science and technology studies, actor-network theory).

## 5. CDS and the problem of raw data

The objectivist model of knowledge considers data as objective particular pieces of information, observed and registered. This model distinguishes between data (empirical) and interpretations (speculative). Data must be accepted, while interpretations may be discussed. It follows from the commonsensical assumption that one does not argue with facts. It is a common belief that there are no facts harder than raw data, as it would seem that data precedes facts. However, the proponents of CDS write

> [a]t first glance data are apparently before the fact: they are the starting point of what we know (...). This shared sense of starting point with data often leads to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself. If we're not careful, in other words, our zeal for more and more data can become a faith in their neutrality and autonomy, their objectivity (Gitelman & Jackson, 2013, 2–3).

CDS assumes that there is no raw data; that it is impossible to separate raw data from interpretation "Data need to be imagined as data to exit and function as such, and the imagination of data entails an interpretive base" (Gitelman & Jackson, 2013, 3). How are such claims justified? Zybertowicz (1995, 86–97) identifies key arguments. Firstly, knowledge is always conditioned by conceptual resources and values of a given culture. Secondly, he refers to the Duhem-Quine thesis, which claims that data in and of itself does not determine theory; there are no meaningful claims based on data alone; any claim must refer to knowledge beyond data. In other words, to make any use of data, we must "pollute" it with outside knowledge. Therefore, basing on one set of data, we may propose empirically equal, but mutually exclusive interpretations (Zybertowicz, 1995, 90–91). Thirdly, data is always selectively dissociated from a fragment of reality.

Let us consider specific examples. Firstly, construction of data involves assigning a value to select objects or their properties. We may consider the number of ill persons in a given society. We will see that it is not unambiguous, as it depends on a given definition of illness (a set of concepts and values). For example, it encompassed homosexual persons until 1972 when American Psychiatric Association ruled that homosexuality was not an illness. Therefore, a given object is assigned different values depending on cultural context. Often, these values are assigned basing on indicators or indices, which are obviously socially

[5] Znaniecki related it only to humanistic knowledge, however, as Jerzy Kmita argues (1985, 47), we should reject the claim that "natural sciences study exclusively the phenomena observed without the humanistic coefficient (objectivized)".

constructed (there are different methods of accounting for unemployment or inflation, based on differently constructed indicators). Secondly, a practical example of the Duhem-Quaine thesis, claiming that data by itself does not determine conclusions and highlighting the social aspect of knowledge formation, is the research of scientific controversies conducted by the Bath School. It showed that there is a degree of flexibility to the interpretation of laboratory data. When a dispute arises, it is often social factors, rather than the nature of the studied object, that determine which interpretation will be accepted (Afeltowicz, 2012, 76–77). Thirdly, any selection of data to be considered in a given study is arbitrary: certain data is shared, while other is neglected, which has impact on results, or sometimes may be a result of a conscious manipulation. If we say that the proportion of drivers punished for traffic offenses rose from 5% to 15% of the entire population, it will seem that the drivers have been less cautious. However, if we add that, at the same time, the number of highway patrols tripled, or that the regulations became more strict, our conclusions will be different. The constructivist critique of statistics refers to "statistical wars", i.e., the use of statistics with the intention of forcing a specific view of reality and justifying specific claims (Miś, 2017, 82).

Alongside big data, CDS is interested in "thick data", i.e., qualitative data aiming to capture as many relevant contexts as possible. The same set of data may have a different sense in different contexts:

> Three different "likes" on a Facebook status may reflect three disparate emotional responses: from intense agreement to sardonic recognition to sympathetic pity. However, when it is analyzed simply as a "like" (...), the thickness of the data and its variety of meanings is lost. In practice, data are not simple evidence of phenomena, they are phenomena in and of themselves (Dalton, Thatcher, 2014).

Hypothetically, we may assume that any context might be datified, and therefore that thick data might be integrated into big data (Der, 2017). From this point of view, the thickness of data is simply a function of its amount and density. However, the number of contexts is potentially infinite, and therefore its selection and assigned importance will always be arbitrary. Thus, Tom Boellstorff (2013) writes that "[w]hat makes data thick is recognizing its irreducible contextuality".

In order to extract knowledge from large dispersed bases, to conduct research in interdisciplinary, dispersed teams, and for the platforms reliant on dispersed online data to realize their tasks, data need to be "communicated and reshaped" (Nafus, 2017). Data movement and reshaping processes do not occur spontaneously; rather, they are initiated by specific subjects and subordinated to specific rules. These processes are not always smooth, as come up against legal, economic, cultural and physical (technical, infrastructural) barriers; they are sometimes entangled with ideological and political issues. All these barriers and entanglements may be referred to as data friction (Bates, 2017; Edwards, 2010). We may take GDPR as an example: there is no doubt that the regulation has a significant impact on data movements, shaping the relations of the subjects using personal data.

I have argued above that any form of critical, constructivist understanding of knowledge, including CDS, involves a measure of reflection on the society. I may add that the researchers associated with CDS are interested not only in the question of the relation between data and results of the analyses conducted on it to reality or truth, but also in its relation to the society and culture; they ask if these analyses are not biased, if they do not serve specific interests, on what systems of value they rely. It is precisely to answer such questions that the research of algorithmic bias has emerged (Iwasiński, in press).

## 6. Conclusion

The discussion above should not suggest that data has no use in research. On the contrary, it is the basic material of knowledge formation and science. Also, I am in no way arguing that the meaning of data is completely relative. But we should be aware that an observation acquires the status of data, and the knowledge derived from is considered objective (if never absolutely so), only in relation to specific assumptions – strictly methodological and technical, and social – referring to values and interests involved in the society's process of knowledge formation. CDS studies precisely these assumptions. In Gitelman's phrase, it is concerned with "looking into data or, better, looking under data to consider their root assumptions" (Gitelman & Jackson, 2013, 4). These assumptions comprise the context of data and of knowledge generated on their basis. Often this context becomes transparent, because such assumptions are invisible, accepted as obvious and unproblematic; often it is simply unconscious. But sometimes it is hidden as a result of an intentional tactic.

I do not mean to negate the value of knowledge formed with the use of operations conducted on data. However, we should not forget that such knowledge might be – and, according to some, always is – tendentious, at least to a degree. It is particularly the case with the algorithms processing big data to predict behaviors or future states. David Sumpter, cited above, argues that no algorithms are free from ideological leanings. Every algorithm is, from some point of view, unfair; it always discriminates against some group subject to the analysis. The group is discriminated not on the basis of mathematics, but of axiology – beliefs and sense of fairness of the algorithm's author:

> Unfairness is like those whack-a-mole games at the fairground where the mole keeps popping up in different places. You hammer it down in one place and another one comes out somewhere else ... There isn't an equation for fairness. Fairness is something human. It is something we feel (Sumpter, 2019, 83–84).

Finally, I should observe that while datafication definitely facilitates market-oriented instrumental rationality, it may also drive developments of a different character – bottom up, social, not prioritizing economical profit. However, they may emerge and succeed only if the relevant data is made available to the groups interested in promoting such initiatives, rather than monopolized by market subjects, and especially not by large internet firms (Morozov, 2016, 22–26).

## References

Afeltowicz, Ł. (2012). *Modele, artefakty, kolektywy. Praktyka badawcza w perspektywie współczesnych studiów nad nauką*. Toruń: Wydaw. Nauk. UMK.

Ajana, B, ed. (2018). *Metric Culture: Ontologies of Self-tracking Practices*. Bingley: Emerald Group Publishing.

Anderson, C. (2008). The End of Theory: The data deluge makes the scientific method obsolete. *Wired* [online], 16(7), [04.06.2020], https://www.wired.com/2008/06/pb-theory/

Bates, J. (2018). The Politics of Data Friction. *Journal of Documentation,* 74 (2), 412–429. https://doi.org/10.1108/JD-05–2017-0080

Boellstorff, T. (2013). Making Big Data, in Theory. *First Monday* [online], 18 (10), 7 Oct, https://doi.org/ 10.5210/fm.v18i10.4869

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.

Brine, K. R., Poovey, M. (2013). From Measuring Desire to Quantifying Expectations: A Late Nineteenth-century Effort to Marry Economic Theory and Data. In: L. Gitelman (ed.). *Raw Data is an Oxymoron* (61–76). Cambridge, Mass.: MIT Press.

Clark, L. (2013). No Questions Asked: big data firm maps solutions without human input. *Wired* [online], 16, [01.02.2016], http://www.wired.co.uk/news/archive/2013–01/16/ayasdibig-data-launch

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review* [online], 4, [04.06.2020], https://hbr.org/2013/04/the-hidden-biases-in-big-data

Dalton, C., Thatcher, J. (2014). What Does a Critical Data Studies Look Like, And Why Do We Care? Seven points for a critical approach to "big data". *Society and Space* [online], 29, [04.06.2020] https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Dijck, J. van (2014). Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. *Surveillance & Society*, 12, 197–208.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming.* Cambridge, Mass.: MIT Press.

Frické, M. (2015). Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology*, 66(4), 651–661, https://doi.org/ 10.1002/asi.23212

Gitelman, L., Jackson, V. (2013). Introduction. Raw Data is an Oxymoron. In: L. Gitelman (ed.). *Raw Data is an Oxymoron* (1–15). Cambridge, Mass.: MIT Press.

Godzic, W. (1998). Cyfrowy film i analogowy widz. *Kino*, 12, 42 – 45.

Habermas, J. (1977). Technika i nauka jako „ideologia". In: J. Szacki (ed.). *Czy kryzys socjologii?* (342–396). Warszawa: Czytelnik.

Habermas, J. (1983). Postęp techniczny i społeczny świat życia. In: Z. Krasnodębski (ed.). *Teoria i praktyka: wybór pism* (357 – 369). Warszawa: PIW.

Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.

Hacking, I. (1991). How Should We Do the History of Statistics?. In: G. Burchel et al. (eds.). *The Foucault Effect: Studies in Governmentality* (181–196). Chicago: The University of Chicago Press.

Hegazy, I. M. (2019). The Effect of Political Neuromarketing 2.0 on Election Outcomes: The Case of Trump's Presidential Campaign 2016. *Review of Economics and Political Science* [online], ahead--of-print, https://doi.org/10.1108/REPS-06–2019-0090

Horkheimer, M. (1987). Krytyka instrumentalnego rozumu. In: M. Horkheimer (ed.). *Społeczna funkcja filozofii: wybór pism* (244–413). Warszawa: PIW.

Iwasiński, Ł. (2016), Społeczne zagrożenia danetyzacji rzeczywistości. In: B. Sosińska-Kalata et al. (eds.). *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa* (135–146). Warszawa: Wydaw. SBP.

Iwasiński, Ł. (2017). Quantified Self. Self-tracking a problem tożsamości. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 55(2), 126–136, https://doi.org/10.36702/zin.369

Iwasiński, Ł. (in press). Social Implications of Algorithmic Bias. In: B. Sosińska-Kalata et al. (eds.). *Nauka o informacji w okresie zmian. Rewolucja cyfrowa – dziś i jutro: Infrastruktura, usługi, użytkownicy*. Warszawa: Wydaw. SBP

Kmita, J. (1985). *Kultura i poznanie.* Warszawa: PWN.

Kiepas, A. (2017). *Filozofia techniki w dobie nowych mediów*. Katowice: Wydaw. UŚ.

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, 1(1), https://doi.org/10.1177/2053951714528481

Krzysztofek, K. (2000). Racjonalność, tabu kulturowe i kontrola przez wolność. In. L. Zacher (ed.). *Racjonalność myślenia, decydowanie i działanie* (121–132). Warszawa: Wyższa Szkoła Przedsiębiorczości i Zarządzania im. Leona Koźmińskiego.

MacKenzie, D. (1978). Statistical Theory and Social Interests: A Case-study. *Social Studies of Science*, 8(1), 35–83.

Manovich, L. (2012). *Język nowych mediów*. Warszawa: Oficyna Wydaw. Łośgraf.
Mayer-Schönberger, V., Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston; New York: Houghton Mifflin Harcourt.
Miś, L. (2007). *Problemy społeczne: teoria, metodologia, badania*. Wydawnictwo UJ.
Morozov, E. (2016). Neoliberalizm na google'owskich sterydach. *Krytyka Polityczna* [online], 44, [04.06.2020], https://krytykapolityczna.pl/gospodarka/morozov-neoliberalizm-na-google-owskich-sterydach/
Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton: Princeton University Press.
Nafus, D., (2017). Data Friction. *Playtus* [online] [04.06.2020] http://blog.castac.org/2017/02/data--friction/
Negroponte, N. (1997). *Cyfrowe życie: jak się odnaleźć w świecie komputerów*. Warszawa: Książka i Wiedza.
Prensky, M. (2009). H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom. *Innovate: Journal of Online Education* [online], 5(3), [04.06.2020] https://nsuworks.nova.edu/cgi/viewcontent.cgi?article=1020&context=innovate
Rudzinski, R. (1987). Wstęp. Filozofia Maxa Horkheimera. In: M. Horkheimer (ed.). *Społeczna funkcja filozofii: wybór pism* (5–33). Warszawa: PIW.
Sikora, M. (2007). *Problem reprezentacji poznawczej w nowożytnej i współczesnej refleksji filozoficznej*. Poznań: Wydaw. Nauk. Instytutu Filozofii UAM.
Simmel, G. (2012). *Filozofia pieniądza*. Warszawa: Wydawnictwo Aletheia.
Sumpter, D. (2019). *Osaczeni przez liczby: o algorytmach, które kontrolują nasze życie: od Facebooka i Google'a po fake newsy i bańki filtrujące.* Kraków: Copernicus Center Press.
Szahaj, A. (2004). *Zniewalająca moc kultury: artykuły i szkice z filozofii kultury, poznania i polityki*. Toruń: Wydaw. UMK.
Szkudlarek, T., Melosik, Z. (1998). *Kultura, tożsamość i edukacja – migotanie znaczeń*. Kraków: Impuls.
Szpunar, M. (2019). Kwantyfikacja rzeczywistości. O nieznośnym imperatywie policzalności wszystkiego. *Zeszyty Prasoznawcze*, 3(239), 95–104.
Sztompka, P. (2002). *Socjologia. Analiza społeczeństwa.* Kraków: Znak.
Szumlewicz P. (2005), Technika jako polityka prowadzona innymi środkami. In: P. Żuk (ed.). *Dogmatyzm, rozum, emancypacja: tradycje Oświecenia we współczesnym społeczeństwie polskim* (169–185). Warszawa: Scholar.
Vaidhyanathan, S. (2005). Critical Information Studies: A Bibliographic Manifesto [online]. SSRN [04.06.2020], https://papers.ssrn.com/sol3/papers.cfm?abstract_id=788984
Villasenor, J. (2011). *Recording Everything: Digital Storage as an Enabler of Authoritarian Governments* [online]. Brookings. Center for Technology Innovation [04.06.2020], https://www.brookings.edu/research/recording-everything-digital-storage-as-an-enabler-of-authoritarian-governments/
Villars, R.L., Eastwood, M., Olofson, C.W. (2011). IDC White Paper: Big Data:What It Is and Why You Should Care [online]. The Data Analytics Report [04.06.2020], https://dataanalytics.report/whitepapers/big-data-what-it-is-and-why-you-should-care/4519
Waszewski, J. (2015). Ewolucja systemów nadzoru. In: A. Zybertowicz (ed.). *Samobójstwo Oświecenia? Jak neuronauka i nowe technologie pustoszą ludzki świat* (233–283). Kraków: Wydaw. Kasper
Williams, T. D. (2013). Procrustean Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers. Raw Data Is an Oxymoron. In: L. Gitelman (ed.). *Raw Data is an Oxymoron* (41–59). Cambridge, Mass.: MIT Press.
Wolf, G. (2010). The Data-driven Life. *The New York Times* [online], April 28, [04.06.2020], https://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html
Wróblewski, M. (2016). Nowe szaty healthismu. Self-tracking, neoliberalizm i kapitalizm kognitywny. *Acta Universitatis Lodziensis. Folia Sociologica*, (58), 5–23.

Zybertowicz, A. (1995). *Przemoc i poznanie: studium z nie-klasycznej socjologii wiedzy*. Wydaw. UMK.
Zybertowicz, A. (2015a). Bilans. In: A. Zybertowicz (ed.). *Samobójstwo Oświecenia? Jak neuronauka i nowe technologie pustoszą ludzki świat* (429–452). Kraków: Wydaw. Kasper.
Zybertowicz, A. (2015b). Oświecenie – utopia, która działa. In: A. Zybertowicz (ed.). *Samobójstwo Oświecenia? Jak neuronauka i nowe technologie pustoszą ludzki świat* (29–57). Kraków: Wydaw. Kasper.

# Teoretyczne podstawy critical data studies

**Abstrakt**

**Cel/Teza:** Celem artykułu jest przedstawienie głównych założeń oraz analiza podstaw teoretycznych nurtu critical data studies (CDS).
**Koncepcja/Metody badań:** Analiza opiera się na krytycznym przeglądzie literatury z zakresu CDS, społecznych aspektów Big Data, a także socjologii wiedzy, filozofii wiedzy oraz studiów nad nauką i techniką.
**Wyniki i wnioski:** Autor wskazuje trzy główne teoretyczne postulaty CDS: (1) krytyka rynkowo zorientowanej racjonalności instrumentalnej; (2) Odrzucenie założenia o niezależności danych od procesu badawczego; (3) Odrzucenie koncepcji surowych danych. W artykule omówiono intelektualne źródła CDS. Autor argumentuje, że nurt CDS wyrasta z konstruktywistycznej socjologii wiedzy oraz studiów nad nauką i technologią.
**Oryginalność/Wartość poznawcza:** Artykuł czerpie z literatury teoretycznej i studiów empirycznych z różnych dziedzin w celu zbadania teoretycznych podstaw CDS i ulokowania tego nurtu na historycznej mapie idei. Podkreśla potrzebę krytycznego patrzenia na dane i ich przetwarzanie, co jest szczególnie istotne w obszarze big data. Nurt CDS jest rozpoznany na gruncie kulturoznawstwa i nauk o mediach (choć słabo dyskutowany w polskiej literaturze naukowej z tych dziedzin), ale nieobecny w informatologii, której dorobek mógłby istotnie wzbogacić.

**Słowa kluczowe**

Big Data. Critical Data Studies. Danetyzacja. Konstruktywizm społeczny. Racjonalność instrumentalna. Socjologia wiedzy.

*Dr ŁUKASZ IWASIŃSKI received the M.E. degree in Organization and Management from Lodz University of Technology in 2006, and the M.A. and Ph.D. degrees in Sociology from the University of Lodz in 2007 and 2013, respectively. At present, he is Associate Professor at the Faculty of Journalism, Information and Book Studies at the University of Warsaw. His major publications include: Iwasiński, Ł. (2016),* Socjologiczne dyskursy o konsumpcji. *Gdańsk: Wydawnictwo Naukowe Katedra* [Sociological discourses of consumption]*; Iwasiński, Ł. (2016),* Społeczne zagrożenia danetyzacji rzeczywistości [Social risks of datafication of reality]. *In: B. Sosińska-Kalata (eds.). Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa [Information science in the age of change. Informatology and digital humanities] (135–146). Warszawa: SBP; Iwasiński, Ł. (2017).* Quantified Self. Self-tracking a problem tożsamości [Quantified Self. Self-tracking and the question of identity]. *Zagadnienia Informacji Naukowej – Studia Informacyjne, 55 (2), 126–136.*

*Contact to the Author*
*l.iwasinski@uw.edu.pl*
*Katedra Informatologii*
*Wydział Dziennikarstwa, Informacji i Bibliologii*
*Uniwersytet Warszawski*
*Krakowskie Przedmieście 69*
*00-927 Warszawa, Poland*