Member since 2019
JM14475

Are They Doing Artificial Intelligence? (Re)Constructing the Primary Activity in Data Science

Remigiusz Żulicki
University of Lodz, Poland

https://doi.org/10.18778/1733-8069.20.4.09

Keywords:
social worlds, primary activity, artificial intelligence, data science

Abstract: Data science (DS) is concerned with building so-called artificial intelligence, i.e., computer systems that automate tasks based on historical data. This article is the first attempt to examine DS using Adele E. Clarke's framework of social worlds. The main goal of this paper is to show the (re)construction of primary activity based on the example of the social world of DS in Poland. Methodological reflection on this (re)construction is an underdeveloped element in the study of social worlds; therefore, this paper strives to make this process explicit. The empirical background is a three-year ethnographic study, following Clarke's situational analysis approach. The methodological results demonstrate the indispensability of collaborative ethnography in (re)constructing primary activity and the importance of finding palpable elements as those being crucial to understanding primary activity. The substantive results focus on the idea that data scientists do not refer to their activity as doing artificial intelligence.

Remigiusz Żulicki
Ph.D., works at the Department of Social Research Methods and Techniques, Institute of Sociology, Faculty of Economics and Sociology, University of Lodz. His research interests include social science research methodology, social worlds/arenas, critical data studies, and digital sociology. He harmonizes panel survey data regarding precarity – *CNB-Young project* – and investigates illegal trash dumping. He is an open-source and open-science enthusiast. Member of interdisciplinary Generative Artificial Intelligence Team at the University of Lodz.
e-mail: remigiusz.zulicki@uni.lodz.pl

Introduction

This paper focuses on (re)constructing the primary activity (Strauss, 1978). It employs the theoretical framework of social worlds (Strauss, 1978; Clarke, 1997) for an exploratory and descriptive study of data science (DS) that I conducted for my doctoral dissertation. It was a three-year ethnographic study in the Polish DS community, following a situational analysis approach (Clarke, 2003; 2005). This paper argues that collaborative ethnography is indispensable in reconstructing primary activity, and finding "palpable"[1] (Strauss, 1978: 121) elements of action is crucial to its understanding. Therefore, this article presents a methodological case study that examines the process of reconstructing the primary activity within the social world of DS, as outlined by Adele E. Clarke (1997).

There are three reasons why data science (DS) is a valuable training ground for a methodological analysis of reconstructing the primary activity. First, the framing of primary activity – initially perceived as "data analysis" and later as "problem-solving involving data, code, and analytical methods" – was found to be inaccurate during the fieldwork. Second, the hype around DS hinders and distorts an outsider's understanding of the primary activity. It is said that data scientists (DSTs) have the sexiest job in the 21st century, that they work with big data, and that they create artificial intelligence (Loukides, 2010; Davenport, Patil, 2012; Cao, 2017). Such statements sound impressive, but that is what creates the hype. Third, the DS example demonstrates the significance of identifying the palpable element of an insider's activities in understanding the primary activity.

Methodological reflection on this process of reconstruction is an underdeveloped element in the study of social worlds. Sociologists are content with Anselm Strauss' (1978: 122) classic framing – the primary activity is just "strikingly evident." However, it may not be sufficient for studying a social world such as DS and other social worlds centered around digital technologies or heavily hyped.

The paper also provides substantive knowledge about DS, but only as much as is necessary to understand the methodological contribution of the article. DS is involved in building systems called artificial intelligence (AI) or, until recently, big data (Elish, Boyd, 2018). These systems are used by billions of people globally and shape our world. Based on historical data, they automate tasks using statistical models; they may show us ads, play us another video, recommend similar clothes, recognize objects in photos, or assign tasks to gig workers. Such systems are deployed across various applications, from Internet platforms owned by Meta, Google, Amazon, and Tencent to telecommunications, banking, health care, or justice. There are also new generative AI systems, such as ChatGPT, Dall•E, or Stable Diffusion, which make it possible to produce graphical or text content from prompts. Still, terms such as big data and AI need to be better understood from the technical perspective, as they are burdened with pop-culture associations and are subject to hype.

1 Palpable in terms of something that can be physically touched, tangible.

I agree with Krzysztof T. Konecki (2020) that sociology does not pay enough attention to modern information technology and its impact on human life. Sociologists can help explain emerging technological issues in an accessible way that goes beyond the hype using the framework of social worlds. However, sociologists who use this framework should problematize the evidentness of the primary activity and the process of reconstructing that activity. Otherwise, they are likely to reproduce clichés and hype, especially when researching emerging technologies such as AI.

The paper is structured as follows. The "Methods" section presents a report on field research and analysis conducted on the DS social world in Poland. The upcoming "Data science literature review" section familiarizes readers with substantive concepts such as DS, machine learning (ML), big data, and AI. The following section shows the connection between these substantive concepts and the sociological theory that frames this work. The "Findings" section details the reconstruction of the primary activity. The subsequent "Discussion" section compares the final formulation of the DS primary activity with an alternative formulation that was initially considered but ultimately deemed a business cliché. The concluding segment summarizes the case study by highlighting insights derived from the primary activity formulation. The paper also addresses the limitations of this formulation.

Methods

I studied the social world of DS in Poland for almost three years following Clarke's situational analysis approach as a theory/methods package (Clarke, 2003; 2015). The research used heterogeneous data sources inspired by a multi-site ethnography (Marcus, 1995; Clarke, 2005; Clarke, Friese, Washburn, 2015) and took place between October 2016 and May 2019. It included the following:

1) individual in-depth "intense" interviews (Charmaz, 2006: 25–27) – almost 31 hours recorded with 26 interviewees;
2) participant observation (Charmaz, 2006; Angrosino, 2010) at 47 sites (including meetup.com gatherings, coding workshops, DS courses, conferences, lectures, and other meetings, both online and offline) – ~295 hours in the field;
3) elements of collaborative ethnography (Lassiter, 2005);
4) parts of analytic autoethnography (Anderson, 2006);
5) and netnography (Kozinets, 2003).

I was the sole researcher, conducting 26 interviews (IDIs). I obtained informed consent to digitally record the IDIs, with only the transcription of these interviews being outsourced. The interviews were unstructured conversations, starting with open questions based on a short list of topics to be covered. This list changed iteratively throughout the research in line with the theoretical sampling strategy.

Participant observation, a research technique equated with ethnography (Angrosino, 2010), enables researchers to engage with parts of the lives of the group they are studying, directly observing

phenomena (Charmaz, 2006). During the initial stages of the investigation, covert participant observation was carried out. As the research progressed, I adopted an increasingly overt approach, revealing my identity to establish friendships and recruit interview participants.

Most observations were completed between October 2016 and June 2018 (176 out of 295 hours). More than half of the observation time was conducted online, covering 16 sites out of 46. The observation sites included local and international meetings in several major Polish cities. During the 2016/2017 academic year, I completed a postgraduate university program in data analysis, programing, and ML.

Collaborative ethnography was used, albeit in a limited form, by having informal discussions with the interviewees days or weeks after the formal interviews. These discussions proved groundbreaking for reconstructing DS core activity, which is discussed extensively in the "Findings" section. Lassiter (2005) defined collaborative ethnography as an approach to ethnographic research that deliberately and openly acknowledges and accentuates collaborative elements throughout the entire research process. On occasion, an initial excerpt of the doctoral dissertation was shared with interviewees for their input. However, consulting theoretical ideas with them before text composition was rare, collaboration was limited, and I retained final decision-making authority. This approach differed from Lassiter's (2005) proposal, where the consultants are regarded as equal co-authors.

Analytical autoethnography was used, which involved questioning and exposing the research process as well as my relationship to the social world under study. Evocative autoethnography was used marginally (Anderson, 2006). I integrated elements of self-reflection into my notes and memos, and created twelve autoethnographic notes. Analytical autoethnography proved more suitable than the evocative one for addressing my positions within the full situation of inquiry (Clarke, 2003).

When utilizing netnography (Kozinets, 2003), contents from social media, blogs, forums, and pertinent websites were analyzed. Exposure to these sources provided insights into the culture, communication channels, and practice aspects of the DS social world, such as Kaggle and StackOverflow. Netnography deepened my understanding of the discourse, which encompassed diverse definitions of DS (Conway, 2010; Taylor, 2016; Kaggle, 2017), varying perspectives on DS functions (Biecek, 2015; Wickham, Grolemund, 2017), and the media hype surrounding DS (Loukides, 2010; Davenport, Patil, 2012).

Theoretical sampling was employed, which entails the purposive selection of successive subjects as sources of qualitative data[2] to address emerging theoretical interests (Clarke, 2005; Charmaz, 2006). It is not intended to provide a representative sample (Clarke, 2005; Charmaz, 2006). Nevertheless, I aimed to achieve a heterogeneous sample of data sources to avoid unintentionally overemphasizing the most visible or easily accessible aspects. The interviewees included individuals of different ages

2 By qualitative data sources, I refer to interviewees, participant observation sites, and various online "places." Notably, in the case of autoethnography, I did not employ any specific sampling or data-eliciting strategy.

with varying professional experience and educational backgrounds. They held various job positions (not exclusively "data scientists"), and some were not professionally involved in DS.

Unintended biases were identified in the sample. Regarding the interviews, the interviewees highlighted the dominance of individuals who attended DS social world events, such as meet-ups and conferences. My participant observation was notably limited to workshops/courses that utilized GNU R rather than Python[3] due to familiarity with R as the initial language of choice.

Two procedures of situational analysis were employed: inductive coding of qualitative data, following grounded theory principles (Charmaz, 2006), and mapping (Clarke, 2003; 2005). The coding process involved iterative "initial" and "focused" cycles, and concurrently generated memos and theoretical notes. QDA Miner CAQDAS facilitated qualitative data work.

Clarke's framing of social worlds serves as a "potent theoretical code" (Martin, 2006: 122), and in this article, the primary activity is such a potent theoretical code. Undoubtedly, from the moment I decided to frame my research as a situational analysis of the social world of DS, this theoretical code shaped further study and analysis.

Three types of Clarke's maps (2005) were generated recurrently. Some initial maps were hand-drawn while awaiting an interviewee in a café, and several final versions were published (Żulicki, 2022). However, these maps were not designed for primary activity reconstruction but, rather, they provided a "bird's eye" view (Uri, 2015: 146). This article employed a different processual diagram (see Fig. 3).

Since a "good ethnographic study" combines data from observations, interviews, and an analysis of the existing data (Angrosino, 2010: 102), in June and July 2019, a quantitative analysis of two types of existing data was included. Secondary analysis of the data from the internal DS surveys[4], treated

3 GNU R is a domain-specific programing language for working with quantitative data, particularly statistics and data visualization. Python is a general-purpose programing language, and has numerous libraries for working with data, especially ML. Both languages are considered simple and suitable for non-programers (Nunns, 2017; Thieme, 2018).

4 The data was from six sources: [1] the registration form of the first Polish R programing language user conference "WhyR?" from 2017 (https://raw.githubusercontent.com/WhyR2017/konkursy/master/dane_z_formularza_rejestracyjnego.csv); [2] the registration form of the second Polish Python user conference "PyData 2018" (https://raw.githubusercontent.com/stared/random_data_explorations/master/201811_pydatawaw2018/pdwc2018_anonym.csv); the first two Kaggle surveys from 2017; [3] https:// www.kaggle.com/kaggle/kaggle-survey-2017 and 2018; [4] https:// www.kaggle.com/kaggle/kaggle-survey-2018. Kaggle is an online platform where people participate in contests for the best models; Stack Overflow Annual Developer Survey 2018; [5] https://drive.google.com/uc?export=download&id=1_9On2-nsBQIw3JiY43sWbrF8EjrqrR4U and 2019; [6] https://drive.google.com/file/d/1QOmVDpd8hcVYqqUXDXf68UMDWQZP0wQV/. Stack Overflow is an online forum for help in coding, not only DS-related. To summarize the results: the DS social world is concentrated in Poland's largest urban centers, with Warsaw dominating. Approximately 80% are men, and 50% are aged 25–34. Polish data scientists mostly have Master's degrees in computer science (30–50%) or mathematics/statistics (15%), but also in medical or life sciences (~10%). In 2019, there were over 260 DS companies operating throughout Poland; about ⅓ of them earned income mainly from foreign sources. At the time, novice data scientists in private companies received salaries above the national average (over 5,000 PLN gross monthly); those with experience earned several times higher. My analysis in R code is available on Github (https://github.com/zremek/survey-polish-data-science).

as acts of self-knowledge, was performed. Additionally, data from meetup.com[5], a popular platform for organizing DS in-person meetings, was web-scraped and analyzed. Most of this work was done using the GNU R programing language. This part of the research aimed to capture "the full situation of inquiry" (Clarke, 2005: xxviii) and to embody and enculturate myself in an activity similar to the primary activity in DS.

Data science literature review

The first known use of the term "data science" dates to the 1970s (Naur, 1974), although the first use in a contemporary sense was in the early 2000s (Cleveland, 2001). As works on DS history indicate (Andrus, Cook, Sood, 2017; Cao, 2017; Donoho, 2017), two media publications contributed to hyping the term DS (Loukides, 2010; Davenport, Patil, 2012). Both articles cite Hal Varian, Google's chief economist, who said that the next decade's "sexy job" will be a statistician (Lohr, 2009). Both articles discussed the opportunities that big data and DS presented as well as the labor market demand for DS professionals. Although Varian used the word "statistician" (Lohr, 2009), Loukides (2010), Davenport and Patij (2012) had already used the term "data scientist."

Attempts have been made to define DS (Big Data Borat, 2013; Azam, 2014; Baiju, 2014; Jarvis, 2014; O'Neil, Schutt, 2015; Taylor, 2016; Delapenha, 2017; Trzpiot, 2017; Doran, 2018), and some have debated whether DS is a rebranding of statistics/data mining or a new field (Hyndman, 2014; Taylor, 2016; Donoho, 2017; Jesionek, 2017). During the netnographic research, I encountered Kaggle's definition (2017), which was close to a palpable element and the proper way of performing the activity. The first Kaggle survey defined the DST as "someone who uses code to analyze data" (Kaggle, 2017). Focusing on this definition was one of the crucial steps in reconstructing the primary activity of DS.

DS has been referred to as working with big data (Loukides, 2010; Davenport, Patil, 2012) and as creating AI (Cao, 2017). However, big data and AI are different concepts, and identifying DS with either is an oversimplification. Big data is about storing and processing digital data, characterized by what Laney (2001) describes as the three Vs: volume, velocity, and variety. Big data is, therefore, not terabytes of structured data in tables (like survey or census data) but a mix of structured and unstructured data (text, graphics, audio, or video) that is a record from social media or e-commerce (Kitchin, 2014). AI, on the other hand, is concerned with the computer automation of tasks, especially those "hard for people to describe formally […] like recognizing spoken words or faces in images" (Goodfellow, Bengio, Courville, 2016: 1). What links big data and AI is ML.

5 The Data Science Warsaw meetup.com group, the biggest in Poland, had over 3,500 enrolled members in mid-2019. It may be an approximation of the total size of the DS world, as people from all over Poland tended to sign up for both the local meet-up group and this Warsaw group. Data access via API and my analysis in R code are available on Github (https://github.com/zremek/meetup-harvesting).

ML is one of many approaches to creating AI, which involves automating tasks using statistical models based on historical data. ML algorithms – including logistic regression, decision trees, or neural networks (Goodfellow, Bengio, Courville, 2016) – were known as far back as the 1960s. However, as the Internet did not yet exist, preparing the digital data necessary to build models was laborious. Access to large amounts of digital data, known as big data, and high-powered computing machines has made it possible to build better-fitting ML models (Goodfellow, Bengio, Courville, 2016). Therefore, in Conway's (2010; see Fig. 1) diagram, ML is at the intersection of "hacking skills" – the ability to use big data and powerful computers – and "math and statistics knowledge" – the use of algorithms to extract patterns from digital data. As this article shows, ML modeling is part of the core activity in DS.

There has been significant interest in the social implications of developing technologies such as big data or AI, called critical data/algorithm studies (van Dijck, 2014; Krzysztofek, 2015; Tufekci, 2015; Dalton, Taylor, Thatcher, 2016; Batorski, Grzywińska, 2018; Zuboff, 2019; Iwasiński, 2020; Crawford, 2021), as well as a methodological and epistemological critique (Boyd, Crawford, 2011; Kitchin, 2014; Elish, Boyd, 2018; Desai et al., 2022). Despite this, few studies have considered DS itself as a research subject. In particular, qualitative research on DS is rare (Lowrie, 2016; 2017; 2018; Grommé, Ruppert, Cakici, 2018; Thomas, Nafus, Sherman, 2018). Thus, this article is the first attempt to use the framework of Clarke's social worlds for studying DS.

The theoretical framework – social worlds

The social worlds theory originates from and is rooted in the Chicago symbolic interactionist tradition. As Clarke (1997) points out, it is based on George Herbert Mead's (1972) notion of a universe of discourse as the organized medium within which the universality of meaning obtains and prevails. The theory of social worlds was developed mainly by Tamotsu Shibutani (1955) and further expanded upon by Anselm Strauss (1978). Clarke made significant contributions regarding the situational analysis approach as a theory/methods package for studying social worlds and arenas (Clarke, 2003; 2015).

The social world is a social whole with shared commitments to certain activities and a universe of discourse. Such a social whole is not clearly distinguishable by geographical, membership, or other formal boundaries. The boundaries are fuzzy and blurry, determined by the interaction and effective communication of the participants (Unruh, 1980; Clarke, 1997). A wide range of studies have employed Strauss' concept to investigate social worlds such as computers (Kling, Gerson, 1978), reproductive science (Clarke, 1997), tattooing and opera (Vail, 1999), and climbing (Kacperczyk, 2016).

Situational analysis is Clarke's constructivist approach to collecting and analyzing qualitative materials within the grounded theory framework, designed as a theory/methods package to study social worlds/arenas (Clarke, 2003; 2005; 2015). Clarke pushed grounded theory into the postmodern turn (Clarke, 2005)

and later the interpretive turn (Clarke, Friese, Washburn, 2017). Her approach focuses on the empirical construction of the situation of inquiry, with the situation becoming the unit of analysis (Clarke, 2005).

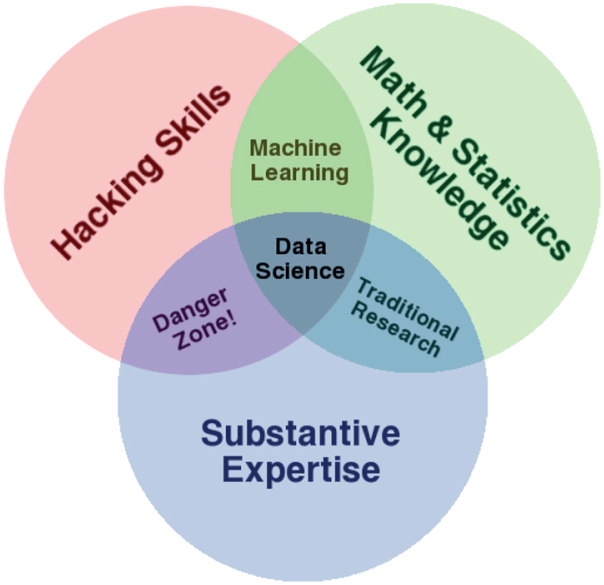
Researchers and their positions are within the full situation of inquiry (Clarke, 2003). Through this approach, Clarke calls for "de/repositioning the researcher from 'all-knowing analyst' to 'acknowledged participant'" (Clarke, 2003: 555–556). This call to reposition the researcher served as a guiding principle in the primary activity reconstruction presented in this paper. I confronted my interpretations and formulations with DSTs during collaborative ethnography, and throughout the autoethnography, I revealed my "intellectual wallpaper" (Clarke, 2005: 85), which is presented further below. Furthermore, challenging the evidentness of Strauss's primary activity follows situational analysis, which aims to help "silences" speak and avoid oversimplification (Clarke, 2003; 2005; 2015; Clarke, Star, 2008: 119, 124, 129).

According to Strauss (1978), the most important defining feature of the social world is the existence of one strikingly obvious primary activity, which becomes the criterion for distinguishing the social world (Kacperczyk, 2016). Perhaps this unquestionable evidentness of a primary activity (or "core activity") has precluded the development of any method to reconstruct the essence of such an activity within a particular social world. To the best of my knowledge, there has been no investigation into how to proceed when the researcher's perception of a seemingly evident core activity proves wrong when confronted with the interpretation of those immersed in this particular social world.

There is a universe of discourse on DS. The primary activity of DSTs is evident; namely, it involves the manipulation and analysis of digital data using a computer. It is difficult to identify alternative criteria to distinguish DS as a social whole. Therefore, during my fieldwork, I decided to treat DS as a social world. I take the practices of self-defining DS as markers for applying the framework of a social world. DS is understood as a "child" born of "parental fields," such as statistics and mathematics, computer science, and business (Andrus, Cook, Sood, 2017). It can also be seen as a social whole located at the intersection of these fields – shown on various Venn diagrams (Conway, 2010; Taylor, 2016). The first diagram, well-known in the DS community, was proposed by Drew Conway (Fig. 1).

In Conway's diagram (Fig. 1), DS is shown at the intersection of three fields. In red, "Hacking Skills" represent technical virtuosity with computers, programing languages, and digital data. Drawn in green are skills in mathematics and statistics, especially familiarity with algorithms for data modeling. In purple, "Substantive expertise" is the most discussed part of the diagram in the world of DS. Some believe that DSTs should know the domain in which they are modeling. Others believe this concerns business knowledge or acumen (Conway, 2010; Taylor, 2016; Andrus, Cook, Sood, 2017).

Figure 1. Data science as a social whole at the intersection of more mature fields of science and practice – Venn diagram



Source: Conway, 2010.

Findings

One cannot formulate the primary activity in the social world of DS as "data analysis". The term is too narrow, because "data analysis" is an activity *within* the primary activity of DS, yet it is also too broad because more than a dozen different communities could be described as performing data analysis (Azam, 2014; Granville, 2014). Thus, the term does not exhaust the complexity of the primary activity. Moreover, there is the job of a "data analyst," which can be seen as being on the fringes of DS if the analyst uses coding to perform their job. It can also be seen as something beyond and much older than DS if the job uses graphical user interface (GUI) tools such as MS Excel or SQL[6] code to query databases. The term "data analysis" also does not convey the proper way to perform the primary activity. In a technology-dependent social world such as DS, using the proper tools separates those who can be considered participants in this social world from those who cannot.

I formulated **the primary activity in the social world of DS as writing codes for data processing, analysis, and modeling**.

6 Structured Query Language (SQL) is a domain-specific language used for work with tabular data held in a relational database.

Writing codes refers to a human typing instructions with a keyboard (called lines of code) in a programing language that a computer can understand (Nowosad, 2019). The word "programing" is not used here to draw attention to the physical act of writing code, but to point out that DS is a separate world from that of IT programing, i.e., software engineering and development. The distinction between coding and programing is also found in the jargon of those who can program computers, which is discussed further below.

Participants of the DS social world can program. They perform the mental process of translating a human idea into a programing language, but when they program, they write a code; they do not create software. Therefore, DSTs are neither programers nor software developers.

A set of instructions written in a programing language that performs specific tasks on the data is called a script. Jacek[7] explained the jargon of the DS world as follows:

> **Jacek:** So, for sure, data scientists program […] if someone writes a script that does something, let's say, takes some data, processes it, and spits out, I don't know, a table at the end, then it can be said that this is not, although it is also a bit of jargon, that this is not programing but coding or writing scripts […] because data scientists are more concerned with data, right? {Interview 9}

Here is an example of a simple script[8] (adopted from Biecek, 2015). The script consists of three commands saved in the file "plot_blood_ds_level.R".

```
sequence <- seq(0, 10, 0.1)
level <- exp(sequence)
plot(x = sequence,
     y = level,
     xlab = "time spent with R",
     ylab = "blood Data Science level")
```

The first two commands create data as two variables – a sequence of numbers from 0 to 10 with a step of 0.1 (named "sequence") and the exponent of this sequence ("level"). The third command plots those variables with axis labels.

7 All names changed, quotes translated from verbatim transcripts.

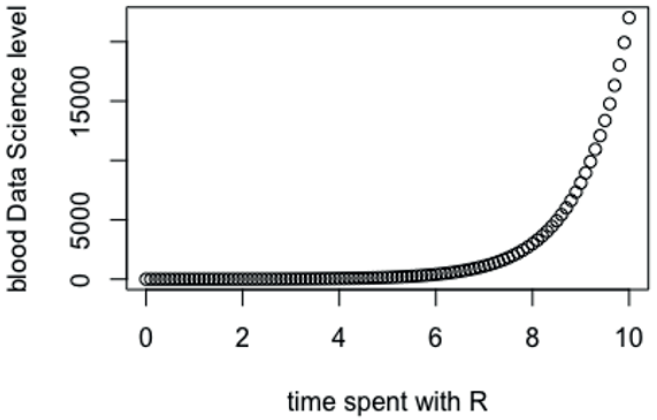
8 Code in GNU R 4.1.2 (R Core Team, 2021) exported to Libre Office Writer with R Markdown (Xie, Allaire, Grolemund 2018).

The script runs the following command:

```
source("plot _ blood _ ds _ level.R")
```

which gives the plot (Fig. 2):

Figure 2. Blood Data Science level vs. time spent with R programing language



Source: own adaptation from R beginners' course (Biecek, 2015).

The plot is the output of the script. The position of each drawn point shows values stored in the variables "sequence" and "level". This script is not only a technical example, but it also provides information on the primary activity of DS. R beginners are told that as they spend time writing codes, they will improve exponentially in DS. Writing a code is a physical and mental activity comparable to playing a musical instrument (Shaw, 2014). DS skills are both embodied and rational. Embodiment is represented by the message: it is the blood data science level (Fig. 2). The rational elements are the medium – data, code, and plot.

The DS code is written for data processing, analysis, and modeling, i.e., activities that make up the process that DS projects follow. Data processing, also called "wrangling" (Wickham, Grolemund, 2017), is the first element of the process. It consists of three stages: importing, cleaning, and transforming data. Data can be imported into the coding environment from a variety of sources. A typical scenario in commercial settings is to import it from a database. There may also be a need to acquire data from internal sources (such as a company email or repository of MS Office files) or external sources (such as social media or web pages).

The next step is to clean/tidy up the data. It may involve combining data from different sources unifying its storage, and defining observation units and variables. Although this process refers to tabular data, in principle, it could also apply to other types of data, such as photographs, audio, or texts.

Data transformations create subsets of data using filtering, variable selection, or sampling. The goal is to reduce the number of variables in favor of quality in preparation for modeling. Identifying and addressing missing data, spurious values, and outliers is also considered as transformations or cleaning.

Data processing is perceived in DS as "dirty work" in a double sense. First, it is tedious, uninteresting, easy, and routine work. But it is also something a little beneath the dignity (Hughes, 1958) of a DST – it is just an ordinary software developer's job that does not contribute much to the understanding of the data. However, DSTs are not ashamed of such work (see Hughes, 1958: 50). Although they may be bored of it, they are also proud of it in a way. Data processing legitimizes a data scientist. DSTs often refer to real-world, unprocessed data as dirty or untidy. They cannot analyze or model such data immediately. Thus, they must be able to clean and transform this data; they cannot avoid getting their hands dirty. Dirty work "gives the occupation its charisma" (Hughes, 1958: 52).

The second step in the process – data analysis and modeling – comprises an iterative process of data transformation and visualization (these two parts are called "analysis"), as well as modeling. Although transformations may initiate the process, and modeling may conclude it; these activities do not have a defined order. The model measures can lead to ideas for further transformations to improve selected measures.

Models are a tool for extracting patterns from data (Wickham, Grolemund, 2017). The model learns from historical data by experience (Goodfellow, Bengio, Courville, 2016) to perform classification or prediction not based on pre-set rules but on patterns extracted from the dataset.

Anna describes modeling as follows:

> **Me:** […] what is modeling anyway?
> **Anna:** Kind of making plasticine puzzles [laughs]. […] So modeling is about finding that balance, so that the errors in one [data]set and in the other set are as close to each other and as small as possible, of course.
> **Me:** Is modeling the most important part of DS?
> **Anna:** The most important part is preparing the data because if you have a good model, but your data is fucked up, you won't really get anything out of it. So, the key is to clean and prepare the data well. {Interview 8}

The saying "garbage in – garbage out" underscores the fundamental importance of data preparation in the context of modeling and analysis. It implies that regardless of the complexity or sophistication of a model, its outputs are only as reliable as the quality of the input data.

Modeling is often seen as DS projects' most engaging and challenging aspect. These models are the supposedly intelligent part of AI/big data systems. While there is a common belief that "robots" may

replace people due to these models, DSTs know that modeling takes only 10–40% of the time in a DS project (CrowdFlower, 2017).

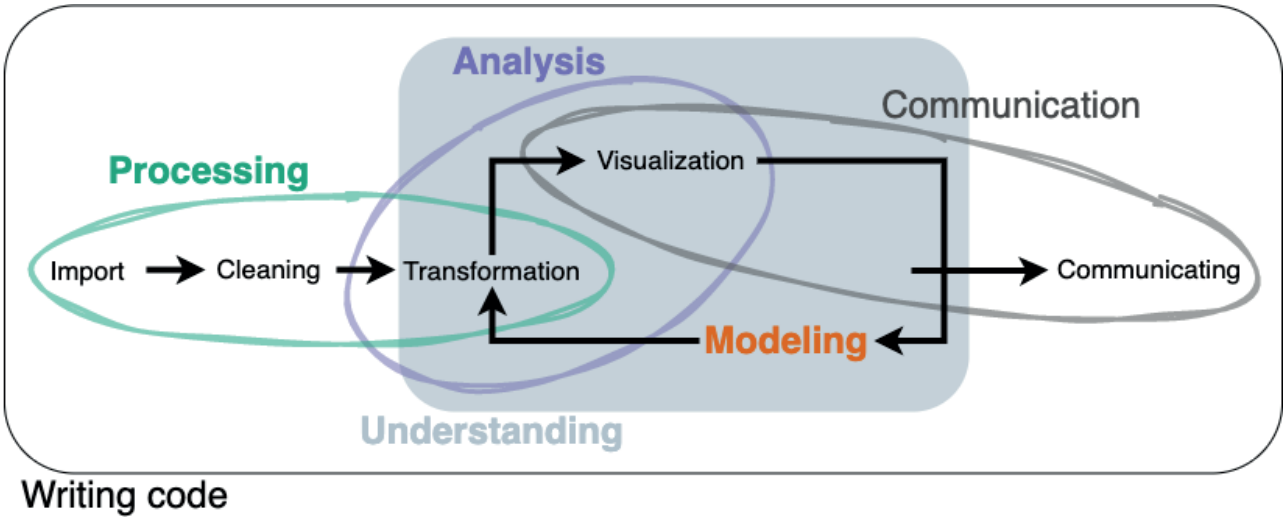
Communicating modeling results to stakeholders is the final element of the DS process (Wickham, Grolemund, 2017). The elements of the communication stage done in code are considered part of the primary activity of the DS world; other elements are considered auxiliary activities. In the code, data visualizations are realized – from graphs for DSTs to those for stakeholders.

In summary, the primary activity in the social world of DS follows a process that comprises the following phases:

1) In the first phase, the operations of importing, cleaning, and transforming data are performed sequentially, called data processing.
2) In the second phase, iterative and not necessarily consecutive data transformation, visualization, and modeling are carried out. This phase should give an understanding of the data.
3) In the third phase, communicating the results obtained in phases 1 and 2 may be considered partly a primary activity and partly an auxiliary activity.

This process is shown in Figure 3.

Figure 3. The primary activity in the social world of DS as a process that comprises data processing, analysis, and modeling



Source: own elaboration, adapted from the R handbook (Wickham, Grolemund 2017).

The whole process is written in a code. Thus, my formulation of the primary activity in the social world of DS is writing code for data processing, analysis, and modeling. In the context of this formulation, the DS community joke becomes clear: "If it is written in Python, it's probably machine learning; if it is written in PowerPoint, it's probably AI" (Alekseichenko, 2019b). Computer code, represented

by Python, is a technical tool that DSTs use for real ML work, while PowerPoint slide deck is a non-technical tool used by business spokespeople to communicate the "AI" packaging.

My formulation of the primary activity of DS refers to people skills and the indispensability of digital data, ignoring the utilitarian purpose of projects. It took me over two years to reconstruct the primary activity, which comprises all my DS research. However, I would like to focus on three aspects.

First, IDIs alone were an insufficient data collection technique for this reconstruction. As something obvious to the participants of the social world, the primary activity was not directly verbalized. In the coded qualitative interviews, it was apparent how much attention the interviewees paid to DS-specific technologies, especially Python and R. However, no codes/categories indicated palpable elements of "doing something" that, for me, are essential to capture the primary activity. The insights gained from the interviews were somehow too far from the primary activity.

Second, my observant participation in writing the DS code was much more important for the reconstruction. A slow embodiment of programing skills into my mind, hands, and eyes was an irreplaceable experience, and indeed, the usefulness of such an embodiment is well established in symbolic interactionism (Becker, 1953; Wacquant, 2004). However, I emphasize it, because I am writing about the social world of DS. With such a digitally-oriented research subject, sociologists might easily overlook the possibility of embodying the researcher in the action. It is probably easier than when studying communities that are focused on bodily action and bodily experience, such as marijuana users (Becker, 1953), boxers (Wacquant, 2004), or liver transplant patients (Plummer, 2012).

Nevertheless, it was also helpful to participate in a community of concentration with other coders during live workshops. Sitting shoulder to shoulder, laptops on our laps, I could see others typing their code, how fast they were using the keyboard, and their focused stares. I could see their sticker-covered computers (see Lowrie, 2016), operating systems, and integrated development environment[9] settings, and, finally, see and ask them about specific lines of code and compare them to what I had written. Therefore, I find that adopting the role of "observing participant" and not just "participating observer" (Gold, 1958; Junker, 1960) is useful for reconstructing a core activity.

However, the insights gained by participating in the activity were too close to the activity to be synthesized. As Josh Seim (2021: 21) points out, "observant participation favors incarnation over inscription" – researchers doing observant participation are more focused on embodying experiences than taking notes. Such experiences contributed a lot to getting an "interactional" level of expertise – "enough expertise to interact interestingly with participants" (Collins, Evans, 2002: 254) – and enabled collaborative ethnography.

9 Integrated development environment (IDE) is software for writing code in programing languages.

Third, the collaborative ethnography led to the breakthrough in my research and revealed the social world. I started the consultation with the interviewees with the following proposal to capture the DS primary activity: "problem-solving with data, code, and analytical methods." I asked whether this framing answered the question "What do people do in DS?" and whether they thought it captured the "essence of DS." In hindsight, the term "essence" was misleading, because the interviewees saw it as a request to define DS. It was also pointed out that the term "problem-solving" is a business cliché that I repeat.

During the consultation, I compared DSTs and their activities with cooks and cooking. The cook can be an amateur home cook, a chef in a five-star hotel, or work in a roadside diner, but all cooks cook. In this context, the interviewees accepted the definition from Kaggle (2017), sometimes saying it was too broad or did not address what the practical purpose of DS was:

> **Me:** I want to answer the question, "What do people do in DS?" This includes commercial, academic, and hobbyist DS. I suggest: "They solve problems using data, code, and analytical methods." […] For you, does that capture the essence of DS? […]
> **Aleksander:** So, what do people do in DS in terms of how you can summarize their work? […]
> **Me:** To summarize. Just like cooks cook […] How do you rate the definition of DS from Kaggle surveys: "writes code to analyze data"?
> **Aleksander:** Well then, it's like one-to-one with the cook cooking. But I miss the element that the cook cooks to make it taste good {LinkedIn chat consultation – Interviewee 15}

From the above conversation, I have learned that writing a code (for data processing, analysis, and modeling) is a primary activity for DS, just as cooking is for cooks. One has to undertake that activity to be considered a participant in the social world. The utilitarian purposes, the desired results the good or true participant is obliged to provide – be that cooking to taste good or providing business problem solutions with data and code – show the values of the social world and its legitimation strategies, but not the primary activity. Besides, this excerpt presents my focus on summarizing DS, but it was groundbreaking for my research. The hype around DS hindered and distorted my understanding of its primary activity. I was influenced by the business legitimacy of its existence – unlike academia, it solves real problems and delivers practical value. However, that is a cliché.

Discussion

Various authors consider problem-solving to be what mathematics and computer science do. DS has developed at the intersection of these fields (Schoenfeld, 1992; Gallopoulos, Houstis, Rice, 1994; de St. Germain, 2008; Lanthier, 2011). The Polish DSTs indicated that DS is a different way of solving problems than software development (Kuncewicz, 2019), and that defining the problem to be solved is the initial and most crucial stage of a DS project (Alekseichenko, 2019a). The famous Google Brain DS team praised the capabilities of ML to "solve challenging problems" (Dean, 2019). True DSTs

are considered to have a good approach to problem-solving, and DS is a field that deals with real-world problems, unlike mathematics and computer science (Lowrie, 2018). Studies using Strauss's framework of social worlds also used the concept of problem-solving. The primary activity of the social world of computers was defined as "generalized problem-solving" (Kling, Gerson, 1978: 27). It was also claimed that scientific research was "problem-solving work" (Gerson, 1983: 358–359) and that "scientific work, then, centers on problem-solving" (Clarke, 1997: 71).

A primary activity may be strikingly obvious, but not necessarily for the researcher. In this study, conventional qualitative data collection and analysis techniques, excluding participant consultations, did not reveal the primary activity. Discursive techniques such as IDIs, netnography, and participant observations of talks or meet-ups provided valuable insights but were too far from the primary activity. This challenge is especially notable in the initial stages of research when the researcher grapples with formulating questions and confronting their interpretations of the social world they are studying. Notably, the less discursive technique of observing participation, particularly in R coding, yielded insights so close to the primary activity that articulating and integrating them into the scientific discourse proved challenging without an external perspective.

Formulating the primary activity of the DS as "solving problems" with data, code, and statistics would have reproduced a business cliché and produced a research artifact. The non-technical, business, and media hype around DS/AI/big data hindered and distorted my understanding of the core activity. As researchers, outsiders trying to get a glimpse of a social world, and, at best, novices, we only have access to the hype beforehand. We have the easiest access to this epidermal layer of discourse, to the colorful and attractive packaging of the social world given to us by its spokespeople. And it is given to us to further the interests of this world, to help it survive and develop, to legitimize its existence. Long-term ethnographic research leads to a deeper understanding of the jargon, the technologies, and the concepts beyond the hype. Nevertheless, confronting the researcher's interpretations with that of the participants is crucial (see Lassiter, 2005). For a field as poorly recognized as DS, this is the only empirical tool with which to verify the accuracy of our formulations.

In this study, the primary activity of DS was not discovered but (re)constructed. (Re)construction, in this context, means formulating the primary activity based on fieldwork and consultations, striving to comprehend it from the viewpoint of participants in that social world. Secondly, (re)construction manifests a constructivist approach (Clarke, Star, 2008), driven by data and theory.

I argue that human actors use their bodies with some physical infrastructure in any social world. Therefore, the palpable element is findable and could be crucial to understanding the primary activity. The assertion here is not that the final formulation of the core activity must, for any social world, include an element of tangible action. Instead, it is suggested that researchers probing the core activity should pose materially focused questions. For instance: What physical elements do participants interact with (or not)? In what manner do they engage their bodies? What objects do they handle, manipulate, or utilize? Is the physical aspect distinguishing them as participants in the social world being studied?

Conclusion

The final formulation of the primary activity in DS made it possible to understand how technologies, which are essential for executing this activity, impact various processes within this social world. These processes encompass setting the boundaries of the social world, legitimization, claiming authenticity, and segmentation/professionalization (Kling, Gerson, 1978; Strauss, 1982; 1984). For example, it clarified the rationale behind DSTs joking about Excel spreadsheets for data analysis: Excel is a GUI business tool capable only of small, tabular datasets. This joking is a boundary-setting practice that shows that Excel is not an appropriate tool for DS. DS needs more powerful coding environments, such as Python or R; it cannot be performed in GUI software (Wickham, 2018).

Based on the final formulation of the primary activity, three major arenas in which DS is involved become apparent. Firstly, a debate revolves around the choice of programing language, i.e., Python or R. This debate illustrates the proper way of performing the core activity in DS – emphasizing the use of code rather than GUI software. The second arena concerns gender as a boundary object within DS, intertwined with the broader discussion on underrepresentation in STEM fields for non-male and non-White/Asian minorities. Lastly, the multiple-world arena encompasses the ML model as a boundary object. There are ongoing debates concerning power dynamics, resources, values, and development across multiple social worlds, including DS, business, academia, IT, media, politics, and law. DS, IT, and parts of academia treat the boundary object as ML. The legal world uses this category, as well as AI. Other worlds mainly use only the non-technical concept of AI, which is the colorful packaging of ML models (and many other technologies). DSTs do not say that they are "doing artificial intelligence." However, the HR department may hire for the "AI Products & Solutions" team, while the sales department may give presentations on "AI-powered software."

This paper has also demonstrated the usefulness of Clarke's theory/methods package in helping "silences" speak (Clarke, Star, 2008: 119, 124, 129). The DS community remained silent in critical data/algorithm studies, which is surprising since DSTs have direct agency over ML models via code in programing languages.

One limitation of the study presented here is my position in the social world of DS. As a novice user of the R language, fascinated by the programing possibilities, I may have overestimated the importance of coding in DS. This is my known "intellectual wallpaper" (Clarke, 2005: 85). Another limitation is the time and place of my research. A primary activity may change over time as DS tools evolve. DSTs tend to believe that DS is still in its infancy, especially in Poland. On the other hand, there is an ongoing professionalization process (Kling, Gerson, 1978; Strauss, 1984) in DS. Modeling data is now being performed by more specialized roles, such as ML engineer and ML researcher, and each role forms its distinct core activity. The term DS may become obsolete. It was already being avoided during my research as too broad and vague.

More research should investigate the presence of the palpable element (Strauss, 1978) in primary activity. It is conceivable that there is, or will be, a social world and a corresponding construct of primary activity devoid of any tangible element yet acceptable to the participants in that world. Methodological contemplation regarding the (re)construction of the core activity also holds considerable potential. Questioning the obviousness of a primary activity and making the process of reconstructing it explicit will be of great use in studying social worlds, especially those that revolve around digital technologies or are experiencing hype. Such efforts could lead to the development of a research method for reconstructing primary activity in any social world. This method might resemble situational analysis methods such as mapping (Clarke, 2005), but with a focus on the core of a social world rather than taking a bird's eye view.

Acknowledgements

The research was conducted for my PhD dissertation, *Data Science in Poland. Ethnography of The Social World* (https://repozytorium.uni.lodz.pl/handle/11089/35209), and further analysis was conducted for the submitted paper. Part of the research was funded by the Department of the Sociology of Culture, Faculty of Economics and Sociology, University of Lodz, for publication: Remigiusz Żulicki (2019), *Pułapki Myślowe Data-Driven. Krytyka (Nie Tylko) Metodologiczna*, "Marketing i Rynek", vol. 8(XXVI), pp. 3–14, https://doi.org/10.33226/1231-7853.2019.8.1

I want to thank Anna Kacperczyk for helping shape the main argument of this paper.

References

Alekseichenko Vladimir (2019a), *10 właściwych pytań przy wdrażaniu uczenia maszynowego*, https://biznesmysli.pl/10-wlasciwych-pytan-przy-wdrazaniu-uczenia-maszynowego/ [accessed: 30.04.2019].

Alekseichenko Vladimir (2019b), *The difference between AI vs ML*, https://www.linkedin.com/feed/update/urn:li:activity:6501030890754314240/ [accessed: 4.05.2019].

Anderson Leon (2006), *Analytic Autoethnography*, "Journal of Contemporary Ethnography", vol. 35(4), pp. 373–395, https://doi.org/10.1177/0891241605280449

Andrus Calvin, Cook Jon, Sood Suresh (2017), *Data Science: An Introduction*, https://en.wikibooks.org/wiki/Data_Science:_An_Introduction [accessed: 21.03.2018].

Angrosino Michael (2010), *Badania etnograficzne i obserwacje*, Warszawa: Wydawnictwo Naukowe PWN.

Azam Anum (2014), *The First Rule of Data Science*, "Berkeley Science Review", 27.04.2014, https://web.archive.org/web/20170922061629/https://berkeleysciencereview.com/article/first-rule-data-science/ [accessed: 26.01.2018].

Baiju Nt (2014), *What is a data scientist? 14 definitions of a data scientist!*, https://web.archive.org/web/20171207002047/https://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/ [accessed: 11.01.2018].

Batorski Dominik, Grzywińska Ilona (2018) *Three dimensions of the public sphere on Facebook*, "Information Communication and Society", vol. 21(3), pp. 356–374, https://doi.org/10.1080/1369118X.2017.1281329

Becker Howard S. (1953), *Becoming a Marihuana User*, "The American Journal of Sociology", vol. 59(3), pp. 235–242.

Biecek Przemysław (2015), *Pogromcy Danych. Przetwarzanie danych w programie R*, https://web.archive.org/web/20161205211608/http://pogromcydanych.icm.edu.pl/ [accessed: 28.12.2016].

Big Data Borat (2013), *@BigDataBorat: Data Science Is Statistics on Mac*, https://twitter.com/bigdataborat/status/372350993255518208 [accessed: 19.02.2018].

Boyd Danah, Crawford Kate (2011), *Six Provocations for Big Data*, "SSRN Electronic Journal", s. 1–17, https://doi.org/10.2139/ssrn.1926431

Cao Longbing (2017), *Data Science: A Comprehensive Overview*, "ACM Computing Surveys", vol. 50(3), pp. 1–42, https://doi.org/10.1145/3076253

Charmaz Kathy (2006), *Constructing grounded theory*, London: Sage Publications.

Clarke Adele E. (1997), *A Social Worlds Research Adventure: The Case of Reproductive Science*, [in:] Anselm L. Strauss, Juliet Corbin (eds.), *Grounded Theory in Practice*, Thousand Oaks: Sage Publications, pp. 63–94.

Clarke Adele E. (2003), *Situational Analyses: Grounded Theory Mapping After the Postmodern Turn*, "Symbolic Interaction", vol. 26(4), pp. 553–576.

Clarke Adele E. (2005), *Situational Analysis. Grounded Theory After the Postmodern Turn*, London: Sage Publications.

Clarke Adele E. (2015), *From Grounded Theory to Situational Analysis. What's New? Why? How?*, [in:] Adele E. Clarke, Carrie Friese, Rachel S. Washburn (eds.), *Situational Analysis in Practice. Mapping Research with Grounded Theory*, Walnut Creek: Left Coast Press Inc., pp. 84–118.

Clarke Adele E., Star Susan Leigh (2008), *The Social Worlds Framework: A Theory/Method Package*, [in:] Edward J. Hackett, Olga Amsterdamska, Michael Lynch, Judy Wajcman (eds.), *The Handbook of Science and Technology Studies*, Cambridge–London: The MIT Press, pp. 113–158.

Clarke Adele E., Friese Carrie, Washburn Rachel S. (2015), *Introducing Situational Analysis*, [in:] Adele E. Clarke, Carrie Friese, Rachel S. Washburn (eds.), *Situational Analysis in Practice. Mapping Research with Grounded Theory*, Walnut Creek: Left Coast Press Inc., pp. 11–75.

Clarke Adele E., Friese Carrie, Washburn Rachel S. (2017), *Situational Analysis: Grounded Theory After the Interpretive Turn*, Los Angeles: Sage Publications.

Cleveland William S. (2001), *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*, "International Statistical Review", vol. 69(1), pp. 21–26, https://doi.org/10.1111/j.1751-5823.2001.tb00477.x

Collins Harry M., Evans Robert (2002), *The Third Wave of Science Studies: Studies of Expertise and Experience*, "Social Studies of Science", vol. 32(2), pp. 235–296, https://doi.org/10.1177/0306312702032002003

Conway Drew (2010), *The Data Science Venn Diagram*, https://web.archive.org/web/20110225163125/http://www.dataists.com/2010/09/the-data-science-venn-diagram/ [accessed: 18.02.2018].

Crawford Kate (2021), *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven: Yale University Press.

CrowdFlower (2017), *2017 Data Scientist Report*, https://visit.crowdflower.com/rs/416-ZBE-142/images/data-scientist-report-dec.pdf [accessed: 11.02.2018].

Dalton Craig M., Taylor Linnet, Thatcher Jim (2016), *Critical Data Studies: A dialog on data and space*, "Big Data & Society", vol. 3(1), https://doi.org/10.1177/2053951716648346

Davenport Thomas H., Patil D.J. (2012), *Data Scientist: The Sexiest Job of the 21st Century*, "Harvard Business Review", https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century [accessed: 3.10.2016].

Dean Jeff (2019), *Deep Learning to Solve Challenging Problems (Google I/O'19)*, https://www.youtube.com/watch?v=rP8CGyDbxBY [accessed: 11.06.2019].

Delapenha Lauren (2017), *42 Essential Quotes by Data Science Thought Leaders*, https://www.kdnuggets.com/2017/05/42-essential-quotes-data-science-thought-leaders.html [accessed: 6.02.2018].

Desai Jules, Watson David, Wang Vincent, Taddeo Mariarosaria, Floridi Luciano (2022), *The epistemological foundations of data science: a critical analysis*, "SSRN Electronic Journal", pp. 1–26, https://doi.org/10.2139/ssrn.4008316

Dijck José van (2014), *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology*, "Surveillance and Society", vol. 12(2), pp. 197–208, https://doi.org/10.24908/ss.v12i2.4776

Donoho David (2017), *50 Years of Data Science*, "Journal of Computational and Graphical Statistics", vol. 26(4), pp. 745–766, https://doi.org/10.1080/10618600.2017.1384734

Doran Derek (2018), *Data Scientist*, [in:] Laurie A. Schintler, Connie L. McNeely (eds.), *Encyclopedia of Big Data*, Cham: Springer International Publishing, pp. 1–4, https://doi.org/10.1007/978-3-319-32001-4_61-1

Elish M.C., Boyd Danah (2018), *Situating methods in the magic of Big Data and AI*, "Communication Monographs", vol. 85(1), pp. 57–80, https://doi.org/10.1080/03637751.2017.1375130

Gallopoulos Efstratios, Houstis Elias, Rice J.R. (1994), *Computer as thinker/doer: problem-solving environments for computational science*, "IEEE Computational Science and Engineering", vol. 1(2), pp. 11–23, https://doi.org/10.1109/99.326669

Gerson Elihu M. (1983), *Scientific Work and Social Worlds*, "Knowledge: Creation, Diffusion, Utilization", vol. 4(3), pp. 357–377.

Gold Raymond L. (1958), *Roles in Sociological Field Observations*, "Social Forces", vol. 36(3), pp. 217–223, https://doi.org/10.2307/2573808

Goodfellow Ian, Bengio Yoshua, Courville Aaron (2016), *Deep Learning*, http://www.deeplearningbook.org/ [accessed: 5.11.2017].

Granville Vincent (2014), *16 analytic disciplines compared to data science*, https://web.archive.org/web/20140808055923/http://www.datasciencecentral.com/group/resources/forum/topics/16-analytic-disciplines-compared-to-data-science [accessed: 2.01.2017].

Grommé Francisca, Ruppert Evelyn, Cakici Baki (2018), *Data scientists: a new faction of the transnational field of statistics*, [in:] Hannah Knox, Dawn Nafus (eds.), *Ethnography for a data-saturated world*, Manchester: Manchester University Press, pp. 33–61.

Hughes Everet C. (1958), *Men and Their Work*, London: The Free Press.

Hyndman Rob (2014), *Am I a data scientist?*, https://robjhyndman.com/hyndsight/am-i-a-data-scientist/ [accessed: 5.01.2018].

Iwasiński Łukasz (2020), *Theoretical Bases of Critical Data Studies*, "Zagadnienia Informacji Naukowej – Studia Informacyjne", vol. 58(1A(115A)), pp. 96–109, https://doi.org/10.36702/zin.726

Jarvis Jeremy (2014), *@jeremyjarvis: A Data Scientist Is a Statistician Who Lives in San Francisco*, https://twitter.com/jeremyjarvis/status/428848527226437632 [accessed: 7.12.2017].

Jesionek Robert (2017), *Uczenie maszynowe i sztuczna inteligencja w opiniach polskich CIO*, https://digitalandmore.pl/uczenie-maszynowe-i-sztuczna-inteligencja-w-opiniach-polskich-cio/ [accessed: 24.04.2018].

Junker Buford H. (1960), *Field Work: An Introduction to the Social Sciences*, Chicago: University of Chicago Press.

Kacperczyk Anna (2016), *Społeczne światy. Teoria – empiria – metody badań: na przykładzie społecznego świata wspinaczki*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Kaggle (2017), *2017: The State of Data Science & Machine Learning*, https://web.archive.org/web/20180222175627/https://www.kaggle.com/surveys/2017 [accessed: 27.03.2018].

Kitchin Rob (2014), *Big Data, new epistemologies and paradigm shifts*, "Big Data & Society", vol. 1(1), pp. 1–12, https://doi.org/10.1177/2053951714528481

Kling Rob, Gerson Elihu M. (1978), *Patterns of Segmentation and Intersection in the Computing World*, "Symbolic Interaction", vol. 1(2), pp. 24–43, https://doi.org/10.1525/si.1978.1.2.24

Konecki Krzysztof (2020), *Uwagi na temat tego, co jest postrzegane jako ważne i nieważne w socjologii*, "Przegląd Socjologii Jakościowej", vol. XVI, no. 2, pp. 188–207, https://doi.org/10.18778/1733-8069.16.2.11

Kozinets Robert V. (2003), *The Field behind the Screen: Using Netnography for Marketing Research in Online Communities*, "Journal of Marketing Research", vol. 39(1), pp. 61–72, https://doi.org/10.1509/jmkr.39.1.61.18935

Krzysztofek Kazimierz (2015), *Technologie cyfrowe w dyskursach o przyszłości pracy*, "Studia Socjologiczne", vol. 4(219), pp. 5–31, https://journals.pan.pl/Content/91277/mainfile.pdf [accessed: 2.11.2018].

Kuncewicz Łukasz (2019), *Lukasz Kuncewicz on LinkedIn: „Data Science Job Interview – How the Questions Will Change in 5 Years?*, https://www.linkedin.com/feed/update/urn:li:activity:6556607403457155074 [accessed: 31.07.2019].

Laney Douglas (2001), *3-D Data Management: Controlling Data Volume, Velocity and Variety*, https://web.archive.org/web/20120813181324/https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [accessed: 14.03.2016].

Lanthier Mark (2011), *An Introduction to Computer Science and Problem Solving*, [in:] Mark Lanthier (ed.), *COMP 1405*, Ottawa: Carleton University, pp. 1–38.

Lassiter Luke Eric (2005), *The Chicago Guide to Collaborative Ethnography*, Chicago–London: The University of Chicago Press.

Lohr Steve (2009), *For Today's Graduate, Just One Word: Statistics*, "The New York Times", http://www.nytimes.com/2009/08/06/technology/06stats.html [accessed: 15.02.2018].

Loukides Mike (2010), *What is data science?*, https://www.oreilly.com/ideas/what-is-data-science [accessed: 8.09.2016].

Lowrie Ian (2016), *Caring for Computers: How Russian Data Scientists Refashion Their Laptops*, "Anthropology Now", vol. 8(2), pp. 25–33, https://doi.org/10.1080/19428200.2016.1202578

Lowrie Ian (2017), *Algorithmic rationality: Epistemology and efficiency in the data sciences*, "Big Data & Society", vol. 4(1), pp. 1–13, https://doi.org/10.1177/2053951717700925

Lowrie Ian (2018), *Becoming a Real Data Scientist. Expertise, Flexibility and Lifelong Learning*, [in:] Hannah Knox, Dawn Nafus (eds.), *Ethnography for a data-saturated world*, Manchester: Manchester University Press, pp. 62–81.

Marcus George E. (1995), *Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography*, "Annual Review of Anthropology", vol. 24, pp. 95–117.

Martin Vivian B. (2006), *The Postmodern Turn: Shall Classic Grounded Theory Take That Detour? A Review Essay*, "The Grounded Theory Review", vol. 5(2/3), pp. 119–129.

Mead George H. (1972), *The Philosophy of the Act*, Chicago: University of Chicago Press.

Naur Peter (1974), *Concise Survey of Computer Methods*, Lund: Studentlitteratur.

Nowosad Jakub (2019), *Elementarz programisty. Wstęp do programowania używając R*, Poznań: Space A., https://jakubnowosad.com/elp/ [accessed: 16.03.2020].

Nunns James (2017), *How Python rose to the top of the data science world*, "Computer Business Review", https://www.techmonitor.ai/technology/data/python-rose-top-data-science-world [accessed: 2.10.2018]

O'Neil Cathy, Schutt Rachel (2015), *Badanie danych: raport z pierwszej linii działań*, Gliwice: Wydawnictwo Helion.

Plummer Ken (2012), *My Multiple Sick Bodies: Symbolic Interactionism, Autoethnography and Embodiment*, [in:] Bryan S. Turner (ed.), *Routledge Handbook of Body Studies*, New York: Routledge, pp. 75–93.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, https://www.r-project.org/ [accessed: 3.12.2021].

Schoenfeld Alan H. (1992), *Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics*, [in:] Douglas Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning*, New York: Macmillan Publishers Limited, pp. 334–370.

Seim Josh (2021), *Participant Observation, Observant Participation, and Hybrid Ethnography*, "Sociological Methods & Research", vol. 53(1), pp. 1–32, https://doi.org/10.1177/0049124120986209

Shaw Zed A. (2014), *Learn Python the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code*, Donnelley: Addison-Wesley.

Shibutani Tamotsu (1955), *Reference Groups as Perspectives*, "American Journal of Sociology", vol. 60(6), pp. 562–569, https://doi.org/10.1086/221630

St. Germain James H. de (2008), *Problem Solving*, https://www.cs.utah.edu/~germain/PPS/Topics/problem_solving.html [accessed: 27.01.2019].

Strauss Anselm L. (1978), *A Social World Perspective*, [in:] Norman Denzin (ed.), *Studies in Symbolic Interaction*, vol. 1, Greenwitch: JAI Press, pp. 119–128.

Strauss Anselm L. (1982), *Social Worlds and Legitimation Processes*, [in:] Norman Denzin (ed.), *Studies in Symbolic Interaction*, vol. 4, Greenwitch: JAI Press, pp. 171–190.

Strauss Anselm L. (1984), *Social Worlds and Their Segmentation Processes*, [in:] Norman Denzin (ed.), *Studies in Symbolic Interaction*, vol. 5, Greenwitch: JAI Press, pp. 123–139.

Taylor David (2016), *Battle of the Data Science Venn Diagrams*, https://web.archive.org/web/20170428061035/https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html [accessed: 14.12.2017].

Thieme Nick (2018), *R generation*, "Significance", vol. 15(4), pp. 14–19, https://doi.org/10.1111/j.1740-9713.2018.01169.x

Thomas Suzanne L., Nafus Dawn, Sherman Jamie (2018), *Algorithms as fetish: Faith and possibility in algorithmic work*, "Big Data & Society", vol. 5(1), pp. 1–11, https://doi.org/10.1177/2053951717751552

Trzpiot Grażyna (2017), *Rozumienie Data Science*, [in:] Grażyna Trzpiot (ed.), *Statystyka a Data Science*, Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, pp. 6–30.

Tufekci Zeynep (2015), *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, "Telecomm & High Tech", vol. 203, pp. 203–218.

Unruh David R. (1980), *The Nature of Social Worlds*, "The Pacific Sociological Review", vol. 23(3), pp. 271–296, https://doi.org/10.2307/1388823

Uri Therese (2015), *The Strengths and Limitations of Using Situational Analysis Grounded Theory as Research Methodology*, "Journal of Ethnographic & Qualitative Research", vol. 10(1), pp. 135–151.

Vail D. Angus (1999), *The Commodification of Time in Two Art Worlds*, "Symbolic Interaction", vol. 22(4), pp. 325–344.

Wacquant Loïc (2004), *Body and Soul: Notebooks of an Apprentice Boxer*, New York: Oxford University Press.

Wickham Hadley (2018), *You Can't Do Data Science in a GUI*, https://www.youtube.com/watch?v=cpbtcsGE0OA [accessed: 27.11.2018].

Wickham Hadley, Grolemund Garrett (2017), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, Beijing–Boston–Farnham–Sebastopol–Tokyo: O'Reilly.

Xie Yihui, Allaire Joseph J., Grolemund Garrett (2018), *R Markdown: The Definitive Guide*, Boca Raton: Chapman and Hall/CRC.

Zuboff Shoshana (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, New York: PublicAffairs.

Żulicki Remigiusz (2022), *Data science: najseksowniejszy zawód XXI wieku w Polsce. Big data, sztuczna inteligencja i PowerPoint*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Cytowanie

Remigiusz Żulicki (2024), *Are They Doing Artificial Intelligence? (Re)Constructing the Primary Activity in Data Science*, „Przegląd Socjologii Jakościowej", t. XX, nr 4, s. 190–213, https://doi.org/10.18778/1733-8069.20.4.09

Czy oni tworzą sztuczną inteligencję? (Re)konstrukcja działania podstawowego w *data science*

Abstrakt: *Data science* (DS) zajmuje się budowaniem tzw. sztucznej inteligencji, czyli systemów komputerowych automatyzujących zadania na podstawie danych historycznych. Niniejszy artykuł jest pierwszą próbą zbadania DS z zastosowaniem ramy teoretycznej światów społecznych Adele E. Clarke. Głównym celem opracowania jest przedstawienie (re)konstrukcji działania podstawowego na przykładzie świata społecznego DS w Polsce. Refleksja metodologiczna nad tą (re)konstrukcją jest słabo rozwiniętym elementem badań nad światami społecznymi; niniejszy artykuł stara się ten proces wyeksplikować. Podstawą empiryczną jest trzyletnie badanie etnograficzne, przeprowadzone zgodnie z podejściem analizy sytuacyjnej Clarke. Wyniki metodologiczne prezentują niezbędność etnografii opartej na współpracy w (re)konstruowaniu działania podstawowego oraz znaczenie namacalnych elementów jako kluczowych dla zrozumienia tego działania. Substancjalne wyniki koncentrują się na spostrzeżeniu, że osoby zajmujące się *data science* nie określają swego działania z użyciem pojęcia sztucznej inteligencji.

Słowa kluczowe: społeczne światy, działanie podstawowe, sztuczna inteligencja, *data science*