Dec 12th, 12:00 AM

Machine Learning for ARUP: Time to Redefine the Ground Truth

Sergey Stroppiana Tabankov
University of Warwick, phd18st@mail.wbs.ac.uk

Machine learning for ARUP: Time to redefine the Ground Truth

Short Paper

Sergey Stroppiana Tabankov
Warwick Business School
University Of Warwick, Scarman Rd, Coventry, UK, CV4 7AL
phd18st@mail.wbs.ac.uk

Abstract

Research accepts that ML-based AI tools' accuracy is a defining characteristic for AI implementation. Yet, the understanding of accuracy in relation to the "ground truth" remains under-researched, especially the understanding of universally recognized practices for the "ground truth" in specific knowledge domains. This short paper investigates how knowledge workers' expertise can be used effectively to redefine the "ground truth" and produce training datasets conducive to more accurate ML predictions. It approaches the question empirically with a case study of ARUP, a global engineering and consultancy firm that uses various AI tools for its advisory services. The paper highlights how executives often overlook data preparation and the role of knowledge workers during this phase, thus questioning the meaning of "ground truth". It provides valuable insights on how a total and constructive collaboration of stakeholders is essential for organizing existing data, contributing to existing literature on ML implementation and data in general.

Keywords: Machine learning, ground truth, accuracy, knowledge workers

Introduction

Today, using existing data to improve the effectiveness of machine learning (ML) models is one of the biggest challenges facing large organizations. As pointed out by Waardenburg et al. (2021:62) *"although Big Data is a big topic, little is written in detail about what datafication actually means for organizations."* Put simply, organizations do not always seem to understand what their data is about and how to use it effectively to build accurate ML models (Agrawal et al 2020; Dwivedi et al., 2019). In fact, although organizations have access to large volumes of data (e.g., Jones, 2019), not many of them understand how to organize existing data to create effective training datasets to train ML models and obtain more accurate predictions (e.g., Aaltonen et al., 2021; Alaimo et al., 2020).

The accuracy of ML models refers to the number of correct predictions in relation to the total number of predictions and, for a number of years, this has been one of the defining characteristics of ML-based artificial intelligence (AI) tools and the main focus of research in this field (e.g., Geiger et al., 2020). As a consequence, managers have traditionally been preoccupied mainly with this criterion to justify the adoption and implementation of new ML models, somewhat partial to the attractiveness of new tools that promise to alleviate workloads and costs while maintaining high-quality services and, at the same time, increasing revenue. This has been the norm in many professional fields, including sales departments (Pachidi et al., 2021), medical (Lebovitz et al., 2021) or criminal justice (Angwin et al., 2016) where the implementation of ML-based AI tools has been increasing exponentially.

However, on the opposite side of the same coin, knowledge workers and ML developers have been working on the process of obtaining accurate output by focusing on the quality of the input, and that is the existing

data. They have been concentrating on the type and quality of data used to train ML models and how the effective labeling of training datasets validated by expert knowledge workers in a specific field (Mitchell et al. 1990; Smyth et al. 1994) is a crucial starting point for accurate predictions and successful ML implementation. More specifically, developers and knowledge workers have been focusing on the importance not only of training datasets labelling, but also of which training datasets are to be considered the "ground truth", that is, the information known to be true in the real world upon which to build ML models (Jones, 2019).

This case study and research in progress build on emerging literature on the "ground truth" (Lebovitz et al., 2021) and domain expertise (van den Broek et al., 2021) which brings into question what is universally recognized as the "ground truth" in some specific domains (i.e., medical and HR). The aim is to shed some light on how organizations can organize for data by emphasizing the importance of reliable and effective data labeling in order to train ML models within specific domains and the role of knowledge workers in sharing their expertise. Thus, the author formulated the following research question:

how can knowledge workers' expertise be used effectively to redefine the ground truth and produce training datasets conducive to more accurate ML predictions?

The author proposes a qualitative study to understand how organizations can organize and classify their existing data in a meaningful way to produce viable training datasets and train their ML models. The author approaches the research question empirically with a case study of ARUP, a global engineering and consultancy firm with its headquarters in London, combining findings from several sources, including interviews and online material. ARUP's work spans from design and engineering to architecture and planning, making the firm an excellent example for the case study because it works on the development of several internal ML-based AI tools, involving different departments and groups of stakeholders. Amongst these tools is Terrain which works with both environmental and satellite imagery data in a setting of supervised ML learning where correct labeling of large-scale data is of vital importance for the effectiveness of the tool itself.

This case study contributes to the current literature on ML-based AI tools implementation (e.g., Berente et al., 2021; Gregory et al., 2021) by providing a new insight on how organizations can organize existing data and improve accuracy of ML output by focusing on the data input and specifically on the creation and labeling of training datasets. This case study is important because it will give managers practical tools in order to improve the existing datafication practices or to put in place new ones with the aim of increasing and encouraging the participation of ML developers and knowledge workers. It also contributes to the literature on data in general (e.g., Chen et al., 2012; Gitelman and Jackson, 2013) because the findings point to a possible new perspective for data, focusing on practices for data labeling and what is currently considered the "ground truth" and used as a base to evaluate ML models accuracy. Specifically, it highlights the presence of untapped potential in the existing data accumulated over the years in the fields of planning and construction, currently available in a format that cannot be processed efficiently. This has the potential to help developers and knowledge workers redefine the "ground truth", thus impacting the final accuracy of ML models predictions. Lastly, it emphasizes the importance of effective collaboration among all organizational stakeholders, with a particular focus on executives, ML developers and knowledge workers (i.e., architects and engineers).

Conceptual Background

Machine learning (ML) is a subset of artificial intelligence (AI) that started flourishing in the nineties and is now being developed extensively by organizations in several different professional fields (e.g., law enforcement, Waardenburg et al., 2018, or human resource, van Den Broek et al., 2020). The popularity of ML today is also due to unprecedented access to large amounts of data that organizations can use to train their ML models. For example, some ML models employ image recognition technology and classify data points to identify patterns and "learn". However, it is important to stress the point that the performance and quality of ML models are largely dependent on effective training datasets and their correct initial labeling (Bechmann and Bowker 2019; Dhar 2016; Provost and Fawcett, 2001). Their accuracy is the measure of how close the ML models' predictions are to the "ground truth" (Geiger et al., 2020; Krig, 2016; Milan et al., 2013; Smyth et al., 1994), that is, how often they match the predefined training datasets labeled as "accurate" by the developers (Kohavi and Provost 1998). While recent research has been focusing on the

adoption by knowledge workers of ML outputs in their field of expertise (Knorr Cetina, 2016; Lebovitz, 2019; Pachidi et al., 2021), an underdeveloped research direction is the active role played by knowledge workers in supporting ML developers in the production of high-quality training datasets.

Research Design and Method

ARUP is a multinational professional services firm that provides design, engineering, architecture, planning and advisory services and relies heavily on digital innovation for its work practices. Its operation spreads across 35 countries and 90 offices, and it employs approximately 16,000 staff. The firm worked on a number of famous projects, including the finalization of the Sagrada Familia in Barcelona, the Marina Bay Sands Integrated Resort in Singapore and the National Stadium in Beijing.

Research Settings

The author conducted a 12-month qualitative study that examined to what extent developers use knowledge workers' expertise to produce training datasets conducive to more accurate ML predictions, supporting ARUP's advisory services. As part of these services, for example, architects and engineers used Terrain, an ML-based AI tool, to generate predictions based on environmental data (e.g., weather conditions, ecological data, consequences on animal species and people) and satellite data to provide planning advise, diagnostic and solutions to be used for urban development. With Terrain, developers combined environmental data with data from image recognition technologies to analyze satellite images; they worked on training datasets to be used to train ML models to generate outputs in the form of different land surfaces classifications or satellite image segmentation files. These could then be used for urban planning and building construction.

Data Source and Analysis

The main focus of this case study was to understand how knowledge workers' expertise can be used effectively to produce training datasets conducive to more accurate ML predictions. It used grounded theory's systematic methodology in the process of theory development, that assisted the author during iterative analysis of data, comparing themes and categories emerged during the data collection process (Charmaz, 2014; Glaser and Strauss, 1967). The bulk of the data comes from semi-structured interviews (Spradley, 1979) conducted with ARUP's employees from several offices worldwide from February 2021 to February 2022. The questions of the interviews were designed to gain an understanding of the different stakeholders' involvement in working with and organizing data, in creating and labeling training datasets to train ML models, and in handling issues arising from existing processes of both training and implementing ML-based AI tools.

The research began from meeting the executive director who led and coordinated AI-related projects for 6 informal interviews approximately 2-hour long to discuss how ARUP organizes its data for current and future ML initiatives. The director was interviewed at different times during the research period to account for any differences in how ARUP organized its practices for data and how this data was used in ML implementation. Next, the author spoke to ARUP's operational directors, associate directors and executive board members who also managed AI-related projects and participated in negotiations with AI developers; this was in the form of 12 semi-structured interviews approximately 1-hour long to deepen the understanding of the issues faced during ML implementation. After obtaining the management's perspective, the author decided to talk to the HR department, conducting 6 semi-structured interviews approximately 30-minute long to understand how they bridged ARUP's need to implement AI and the employees' clear understanding of this need. In addition, the author talked to the digital and technology leads, and the ML developers involved in creating training datasets, designing and implementing AI tools, and conducted 6 semi-structured interviews approximately 1-hour long.

In order to triangulate the findings and gain a 360-degree perspective, this was followed by 20 semi-structured interviews with architects and engineers who actually use ARUP's AI tools in order to compare their perspectives with the ones previously collected from stakeholders and decision makers. This helped gain a better understanding of the nature of AI-mediated practices in ARUP, to identify issues faced by developers when creating training datasets and by employees when adopting these tools (Table 1).

Informants	Data Type	Topics
ARUP decision makers	Informal interviews (n=6; 2 hr.)	Data; AI; ML; org. practices
ARUP directors; associate directors & executive board members	Semi-structured and recorded interviews (n=12; 1 hr.)	Data; AI; ML; org. practices
ARUP HR personnel & personal assistants	Semi-structured and recorded interviews (n=6; 30 min.)	Bridging practices between the firm's needs and employees' expectations; executives' unawareness; employee's resistance
ARUP digital and technology leads & ML developers	Semi-structured and recorded interviews (n=6; 1 hr.)	Data; AI; ML; org. practices
ARUP architects & engineers	Semi-structured and recorded interviews (n=20; 1hr.)	Data; AI; ML; design & org. practices
ARUP supplementary materials	Analysis of ARUP's official YouTube channel (n=378). Online material and documentation (i.e., websites & reports) related to the unique projects that relay on big data	Opportunities & challenges while implementing AI and ML

Table 1. Data Collection Breakdown 2021-2022

Preliminary Findings

ARUP's executive are continuously trying to refine an operational model that would include the use of technologies such as a machine learning-based (ML) artificial intelligence (AI) tools for design practices and advisory services. One of the recurrent themes to emerge from the interviews is in fact the firm's intention to "*reuse our existing data in order to be able to help our developers and enable even more sophisticated technologies*".

The Need for global Data Standards

For the last two decades, a team of ARUP's executives has been pushing for global data standards with the aim to organize and distribute existing data in a single and shared format…

> *…to move away from traditional data analysis towards advanced data analysis…and have the right data with the right quality, so we can use artificial intelligence and machine learning effectively…*

This focus assisted the firm *"in addressing a part of the already estimated data-misuse loss of $15.8 billion alone in the US"*. However, this has not come without its challenges due to the heterogeneous nature of data and the various data formats traditionally used in the engineering and construction field, as pointed out by one of the engineers at ARUP:

> *…we need a lot of data to assess the state of the majority of the bridges in the Netherlands that were built after WW2…to make a digital twin of a particular bridge…We are now working on having a database with its analysis, resulting in the right datasets etc., so we can search through it much quicker without going through the bulk of information to get where we want to…BUT it is a very difficult task to clean and organize all data that is available because nobody can find anything since, traditionally, this was done with Word or, PDF documents, JPEG*

> *scans and filing cabinets, with less than ideal resolution, insufficient quality, inconsistent naming practices…*

The Improvement of AI Tools

An important theme to emerge from this research is ARUP's strong drive for the continuous improvement of its own ML-based AI tools. Throughout the years ARUP has in fact developed several internal tools optimized for specific tasks or tailored to find specific solutions for complex projects. Some of these tools have been quite successful in setting ARUP apart as one of the leader firms in the construction and engineering field. Terrain is perhaps one of ARUP's most well-developed ML-based tools currently employed in land use analysis to help architects and engineers in urban planning. The tool combines satellite imagery and environmental data with ML's ability to process large quantity of visual data and recognizes patterns to produce accurate land use maps. These are used to illustrate risks, implications and opportunities of new projects within the context of land settings or even to produce new types of land classifications. The unique selling point of these technologies is arguably their high level of accuracy which depends firstly and foremost on the data used to train ML models.

On the other hand, some tools developed for addressing buildings' malfunctioning (such as the one used for the City Hall project in the Hague) proved more difficult to scale due to the nature of the field itself, where the specificity of each building makes it impossible to create default training datasets applicable to all buildings. These and other similar tools are the continuous focus of developers' work and at top of management's agenda. In fact, according to one of the directors…

> *…what we do in terms of developing and continuously pushing forward and setting the bar higher and higher is to really work in identifying areas of our organization and the services we provide where we could really push the computational power that AI offers to the max.*

This reinforces the executives' view that the human cost of time is probably one of the biggest barriers to achieving progress at their desired speed and that, effectively, software cost is fairly insignificant compared to human costs. One thing that everyone at ARUP agreed on is that the very nature of AI means that you cannot limit human intervention, nor eradicate it completely, especially when it comes to knowledge workers' expertise, although, based on this research, there seemed to be two different perspectives on this topic.

Executives' Perspective versus Knowledge Workers' Perspective

Although ARUP agreed that the current approaches to data are not always as efficient as they should or could be, the author noted a difference between the executives' perspective on one side and the developers' and knowledge workers' perspective on the other side with regard to training data. The executives' perspective seemed to suggest that accuracy depends on the amount of existing data to train ML models despite admitting that they had access to large volumes of data already:

> *I used 10 years of various projects data and financial data to train our machine learning model to find out certain patterns within that data. We were looking for hidden correlations, so we could learn about which client we worked for, how we actually managed our projects, what we charged for each project, and so on and so forth. However, although I had data related to 10,000 projects, I do not think I had enough data.*

Interestingly, the same executives also touched on the possibility that the existing data might not be of a good-enough quality or of the correct type, calling some of the data "incoherent" or "inconsistently collected", but at the same time stating that *"this was not necessarily part of the problem"*.

On the opposite side, developers and knowledge workers recognized not only the role of large volumes of data as a starting point to train ML models, but also the importance of a systematic way of data collection, highlighting issues with ARUP's datafication practices:

> *AI and ML are very data-hungry, but we must know how to harvest data in an efficient way to be able to train ML models properly. With the right data collected in the right way, we can then look at creating the right training datasets.*

Domain Expertise's Role in defining the "Ground Truth"

A recurrent theme in the conversation with architects and engineers was the recognition that ML-based AI tools will not replace completely human intervention and, specifically, their expertise in the near future; such tools *"can only really be an enabler of human decision making"* and are meant to explore the computational power of ML to augment human performance and experts' knowledge. For this reason, architects and engineers strongly expressed the need for them to be involved in the initial phases of ML training, that is, the creation and labeling of the training datasets from existing data, where their expert knowledge would be a crucial help:

> *although developers are there…at present us architects and engineers are still trying to understand how we can speak to them before we push something into mass motion to see how we could create a layering system to remove the data that we do not need…this is needed because there is a lot of data that does not add any value, slowing down our processes and impacting accuracy…we need to be involved from the beginning, rather than being asked to contribute when the ML model is not performing as expected…*

Some of ARUP's ML-based AI tools operating in a supervised learning environment and use image recognition and classification to identify and learn recurring patterns between input data and output data using previously labelled datasets. The quality of the tools depends on how often the tools' predictions match the labels predefined as "accurate". However, this presupposes that the initial training datasets have been labelled correctly and, most importantly, it depends on what has been initially defined as the "ground truth". In more than one occasion, architects and engineers have highlighted their lack of involvement in the data labeling phase, questioning the initial parameters used for labeling at the bases of ARUP's ML models. This shifts the focus on what is initially defined as the "ground truth" and whether this represents faithfully the knowledge in ARUP's expert domains. In Figure 1 the current approach at ARUP - with the knowledge workers' feedback being provided once the ML model has already been trained – is compared to a hypothetically more effective approach suggested by knowledge workers that illustrates their active role in reviewing and therefore redefining the "ground truth" at the very beginning of the process.

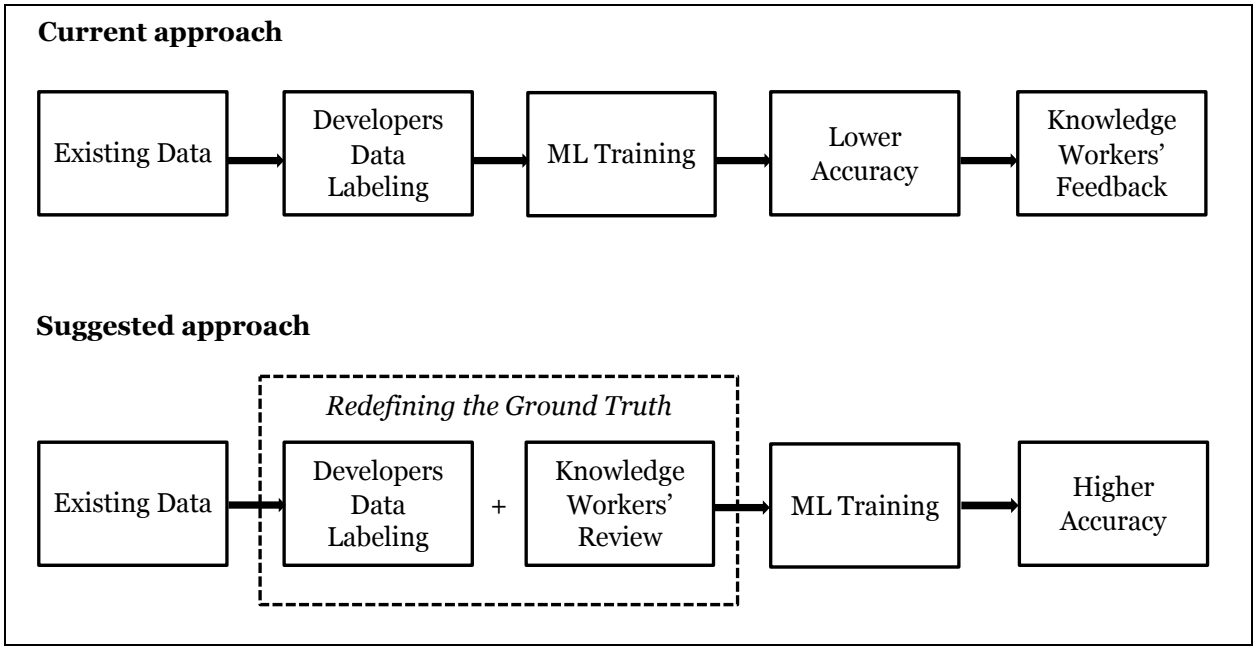

Figure 1. Current Versus Suggested Approach

The suggested approach emerging from this case study appears to be aligned with recent research directions in other knowledge domains (e.g., medical field in Lebovitz et al., 2021); however, in ARUP's current settings, the defining characteristic appears to be the presence of gaps in the process of collaboration among stakeholders, and specifically ML developers, knowledge workers and executives.

Theoretical Implications and Future Outlook

Machine learning (ML) based artificial intelligence (AI) tools have become increasingly popular in many fields of work and many organizations have adopted tools or invested in developing and improving their own internal tools (e.g., Smith et al., 1994). This has been possible due to the large amount of existing data to produce training datasets to train ML models. Despite the fact that considerable amount of research has been conducted on ML models' accuracy and its measurement (e.g., Mitchell et al., 1990), not much research has been done on what is recognized as the "ground truth" and the role of knowledge workers' expertise in defining it, which is at the very basis of ML models.

By looking at how organizations reuse existing data, this case study examines how knowledge workers' expertise can be used by developers to produce better quality training datasets to improve ML models accuracy. Specifically, it highlights the importance of correct labeling for the datasets and which training datasets are to be considered the "ground truth". This case study has opened the door for substantial contribution to existing research (e.g., Aaltonen et al., 2021; Alaimo, 2020; Gitleman and Jackson, 2013; Jones, 2019) because it highlights the importance of domain expertise in the production of training datasets. It questions the meaning of "ground truth" actively looking at the role of knowledge workers and what they provide as labels for the "ground truth", thus shifting the focus on the preparation of data, a phase very often overlooked in the development of ML models. Specifically, the knowledge workers' input should ensure that the "ground truth" datasets used are a reliable representation of the knowledge in their domain, possibly establishing benchmarks to improve "ground truth" acquisition (e.g., Krig, 2016; Milan et al., 2013). This might require revisiting the responsibilities of the stakeholders involved, and particularly of the knowledge workers, to ensure the new organizing for data practices become part and parcel of all stakeholders' roles and of the organizational fiber at all levels.

This case study also provides valuable insights on how organizing existing data and effective datafication in large organizations such as ARUP are not achievable without a total and constructive collaboration of all stakeholders involved. Effective data reuse and development of ML models require open and clear communication among developers, knowledge workers and executives to bring together the different perspectives needed in order to have a more complete view of the bigger picture (Berente et al., 2021; Gregory et al., 2021). The extended version of this research paper includes a section with the interview questions and the summaries of all interviews to further corroborate the importance of collaboration and communication among all stakeholders as a base to the effective implementation of ML-based AI tools. It also includes a detailed analysis of the data gathered and its visual representation for ease of reference and understanding to corroborate the importance of the investigated topic. The data is grouped by stakeholders and gives an overview of the different priorities and all elements emerging from the interviews (including type of issues, expectations of each stakeholder, proposals for solutions, outcomes of pilot initiatives and actions taken). In addition, the full paper provides a section discussing the current process for the creation and labeling of training datasets and the conditions of its specific settings at ARUP with limitations and boundaries related to the features of the field and the type of ML-based AI tools.

This case study is also characterized by some limitations with its research design, generalizability and data sources.

Firstly, it is a *single* case study exploring datafication practices and their link to the "ground truth" in a *specific organization* and *specific knowledge domain* where knowledge workers' expertise is under-researched. The author is aware that the design of this single case study presupposes a population sample that might not be representative of the general population (i.e., all stakeholders involved in ML work and research) and for this very reason the findings might not be applicable to other situations or settings such as different industries, knowledge domains or stakeholders. The author is currently working in securing access to two more firms operating in the same field as ARUP in order to gather further data as a starting

point to assess to what extent the findings can be transferrable or useful in different settings. This will also shed some light on whether other firms operating in the same industry might be experiencing similar stakeholders' collaboration issues with regard to practices for labeling and defining the "ground truth". Based on this, it will be possible to evaluate the efficiency of the suggested approach in the specific knowledge domain of architecture and construction by means of a comparative study (e.g., current vs suggested approach) and analysis of key performance indicators.

Secondly, due to the Covid-19 pandemic breakout overlapping with part of the study, the entire research had to be conducted remotely (with the exception of the meetings in London, Milan and Amsterdam with one of ARUP's top executives) and the author was, therefore, unable to conduct any field observations. To this regard, the author is looking into gaining access to ARUP's headquarter in London to conduct field observations of executives, ML developers and knowledge workers in their own working and cultural environment. The author intends to integrate with the existing findings further insights on the extent to which knowledge workers' expertise is used and a better understanding of the more technical aspects of data preparation and labeling. The author is looking at examining closely Terrain, one of the ML-based AI tools developed internally by the firm, and the cooperation between developers, architects and engineers in creating datasets and defining the "ground truth".

References

Aaltonen, A., Alaimo, C., and Kallinikos, J. 2021. "The making of data commodities: data analytics as an embedded process," *Journal of Management Information Systems*.

Agrawal, A., Gans, J., and Goldfarb, A. 2020. "How to win with machine learning," *Harvard Business Review*.

Alaimo, C., Kallinikos, J. and Aaltonen, A. 2020. "Data and Value," in The Handbook of Digital Innovation Nambisan, S. Lyytinen, K. and Yoo, Y. (eds.), Cheltenham: Edward Elgar Publishing, 162-178.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. "Machine Bias," ProPublica (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Bechmann, A., and Bowker, G. C. 2019. "Unsupervised by Any Other Name: Hidden Layers of Knowledge Production in Artificial Intelligence on Social Media," *Big Data & Society* (6:1).

Berente, N., Gu, B., Recker, J. and Santhanam, R. 2021. "Managing Artificial Intelligence," *MIS Quarterly* (45:3) pp. 1433-1450.

Charmaz, K. 2014. "Constructing Grounded Theory," Thousand Oaks, CA: SAGE Publications.

Chen, H., Chiang, R.H.L., and Storey, V.C. 2012. "Business intelligence and analytics: From big data to big impact," *MIS Quarterly* (35:4), pp. 1165–1188.

Dhar, V. 2016. "When to Trust Robots with Decisions, and When Not To," *Harvard Business Review*.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... and Williams, M. D. 2019. "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, 101994.

Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J. Tang, R., and Huang. J. 2020. "Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

Gitelman, L., and Jackson, V. 2013. "'Raw Data' is an Oxymoron," L. Gitelman (ed.), MIT Press, Cambridge, MA, pp. 1-14.

Glaser, B., and Strauss, A. 1967. "Discovering Grounded Theory," Chicago: Aldine Publishing Company.

Gregory, R. W., Henfridsson, O., Kaganer, E., and Kyriakou, H. 2021. "The Role of Artificial Intelligence and Data Network Effects for Creating User Value," *Academy of Management Review* (46:3), pp. 1-18.

Jones, M. 2019. "What we talk about when we talk about (big) data," Journal of Strategic Information Systems, (28:1), pp. 3–16.

Knorr Cetina, K. 2016. "What If the Screens Went Black? The Coming of Software Agents," in Beyond Interpretivism? New Encounters with Technology and Organizations, L. Introna, D. Kavanah, S. Kelly, W. Orlikowski, and S. Scott (eds.), Berlin: Springer.

Kohavi, R., and Provost, F. 1998. "Glossary of Terms," *Machine Learning* (30), pp. 271-274.

Krig, S. 2016. "Ground Truth Data, Content, Metrics, and Analysis," Chapter 7 in Computer Vision Metrics: Survey, Taxonomy, and Analysis, Cham: Springer International Publishing, pp. 247-271.

Lebovitz, S. 2019. "Diagnostic Doubt and Artificial Intelligence: An Inductive Field Study of Radiology Work," in *Proceedings of the 40th International Conference on Information Systems*, Munich.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. 2021. "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *MIS Quarterly* (45:3), pp. 1501-1525.

Milan, A., Schindler, K., and Roth, S. 2013. "Challenges of Ground Truth Evaluation of Multi-Target Tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 735-742.

Mitchell, T. M., Mabadevan, S., and Steinberg, L. I. 1990. "LEAP: A Learning Apprentice for VLSI Design," in Machine Learning, Y. Kodratoff and R. S. Michalski (eds.), San Francisco: Morgan Kaufmann, pp. 271-289.

Pachidi, S., Berends, H., Faraj, S., and Huysman, M. 2021. "Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing," *Organization Science* (32:1), pp. 18-41.

Provost, F., and Fawcett, T. 2001. "Robust Classification for Imprecise Environments," *Machine Learning* (42:3), pp. 203-231.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. 1994. "Inferring Ground Truth from Subjective Labelling of Venus Images," Advances in Neural Information Processing Systems, pp. 1085-1092.

Spradley, J. 1979. "The Ethnographic Interview", New York: Holt, Rinehart and Winston.

van den Broek, E., Sergeeva, A., and Huysman, M. 2020. "Managing Data-Driven Development: An Ethnography of Developing Machine Learning for Recruitment," *Academy of Management Proceedings* (2020:1), p. 17689.

van den Broek, E., Sergeeva, A., and Huysman, M.H. 2021. "When the machine meets the expert: An ethnography of developing AI for hiring," *MIS Quarterly*.

Waardenburg, L., Huysman, M., and Agterberg, M. 2021. "Managing AI Wisely," UK: Cheltenham.

Waardenburg, L., Sergeeva, A., and Huysman, M. 2018. "Hotspots and Blind Spots," in Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology, U. Schultze, M. Aanestad, M. Mähring, C. Østerlund, and K. Riemer (eds.), New York: Springer, pp. 96-109.