VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Creativity in data work: case of developing training sets for machine learning

Karacic, Tomislav; Günther, Wendy; Sergeeva, Anastasia V.; Huysman, Marleen

2021

Link to publication in VU Research Portal

citation for published version (APA)
Karacic, T., Günther, W., Sergeeva, A. V., & Huysman, M. (2021). *Creativity in data work: case of developing training sets for machine learning*. Paper presented at 37th EGOS Colloquium , Amsterdam, Netherlands.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 12. Dec. 2021

Creativity in data work: case of developing training sets for machine learning

Tomislav Karacic (t.karacic@vu.nl)

Wendy Günther (w.a.gunther@vu.nl)

Anastasia Sergeeva (a.sergeeva@vu.nl)

Marleen Huysman (m.h.huysman@vu.nl)

KIN Center for Digital Innovation; Vrije Universiteit Amsterdam

Work in progress - Do not cite

Abstract Data reuse is seen as an important practice for realizing value from data. But, as scholars have repeatedly shown, the "cooked" character of data can present great challenges for data reuse. Yet, empirical research into how organizations can reuse data despite its "cooked" character is still underresearched. To address this gap, we followed five teams as they developed ML solutions for tackling complex, agricultural challenges. Our research finds that the development teams engage in creative data work which goes beyond mere preparation of data for training a machine learning model. In doing so, the team engaged in three data work practices: *problematization, creative data work,* and *scrutinizing datasets*. Our study shows that, in what seemingly appears as a merely technical and "janitorial" work, developers iteratively learn and interlace their knowledge of available data and a phenomenon in an effort to creatively produce a representation of that phenomenon in a form of a workable training set.

1. Introduction

Scholars have highlighted the importance of organizations opening their data and realizing value from its reuse for knowledge production and innovation (Gunther et al 2017a; Van den Broek and Van Veenstra, 2015; Leonelli 2013; Verhulst 2020). But, reusing data for purposes it was not originally intended for is arguably a major challenge for organizations. Data is embedded in a particular context in which it is produced, collected, stored, and worked with for a particular representational purpose, i.e. data is always "cooked" (Pine 2019; Jones 2019). So, data workers using "cooked" data need to cope with the fact that "cooked" characteristics

This paper was presented at a paper development workshop. As such, it is an early work-in-progress paper. Please do not distribute or cite this work.

of data cannot be removed from some "pure data" which can then be reused freely (Gitelman & Jackson).

Challenge of data reuse is particularly relevant for machine learning (ML). In the discourse on ML, data work, including the one performed for data reuse, is primarily understood as a process of cleaning, augmenting, and assembling data into a particular representation of a target phenomenon (Zhang et al 2020, Kitchin 2014, Jones 2019, Lehr & Ohm 2011). Yet, studies have shown that working with data involves practices of judging, understanding, and contextualizing data (Gitelman & Jackson 2013; Pine 2019; Jones 2019). This suggests that developing training sets for ML from "cooked" data involves more than merely technical work. Moreover, so far, we know little about how actors deal with "cooked" data when aiming to use the data to train ML models, and establish generativity in practice. Thus, there is a need for deeper understanding of work involved in ML development that takes into account all data choices and practices involving data. To address this issue, in this paper we formulate the following research question:

How do data workers cope with the embedded nature of cooked data when creating training sets for machine learning?

To answer the research question, we conducted a qualitative study on five machine learning development teams tackling predefined challenges related to agriculture, thereby relying on open data from a range of different contexts (e.g., satellite data and weather data). We adopt a practice perspective (Feldman & Orlikowski 2011) to analyze these cases and make sense of the work that goes into data reuse. By emphasizing how actors carry out the data work - e.g., how data workers collect, assemble, and transform data - the practice perspective enables us to understand how data from different contexts actually "come to be used" in new contexts (Jones 2019).

Work in progress - Do not cite

Our findings show that data workers enact three practices that emerge as they cope with the "cooked" character of data in order to create ML models and facilitate reuse beyond the original purpose for which the data has been produced. These practices are: problematization, creative data work, and scrutinizing datasets. The three practices show how facilitating data generativity involves creative work of creating workable representations of target phenomena in the form of a training set. Data work performed is creative because developers tackle emerging and unexpected challenges through inventive actions that go beyond mere technical,

linear work. Alongside it, data workers iteratively learn and improve their understanding of data and how particularities of that data enable or constrain them to construct a workable representation of a phenomenon of interest in the form of a training set. By looking at the work involved in production of these data sets from a practice perspective, we are able to explicate practices that often remain hidden or underappreciated, while their enactment is crucial for successful data reuse. This insight is important for organizations that are increasingly implementing machine learning solutions using reused data to cater for their specific organizational needs.

2. Theoretical background

In this section, we discuss scholarly work relevant for our research question. The first one concerns the nature of data. The relevance of this debate is in the centrality of the question of (un)boundedness, i.e. the ability to use data in novel contexts. As such, it is closely related to the literature on data sharing and reuse, with empirical studies engaging with both literatures. Yet, the literature on the properties of data, while bringing new insights, often involves conceptual arguments for or against a certain property with empirical cases being somewhat rare. This motivated us to also include the literature on data work in our research. Besides bringing valuable empirical insights, data work studies also involve uncovering hidden aspects of work involved around data, which makes it a natural setting for our research question. Finally, we also review the literature on digital representations. This research line stresses the enduring problem of representations never being the same as what they represent. As such, it highlighted the key problem data workers face when working with data - data can never fully represent a phenomenon - and enabled us to focus on the way data workers cope with this problem.

Properties of Data Data has increasingly been conceptualized as an "unbounded", "portable" and "open-ended" resource, meaning that it can be used for many different purposes beyond the original purpose, in a wide range of new contexts (Constantiou & Kallinikos 2015; Alaimo, Kallinikos & Aaltonen 2020; Alaimo & Kallinikos 2016; Ekbia 2009; Gerlitz & Helmond, 2013; Kallinikos et al. 2013), often in unexpected ways and with unanticipated consequences (Lycett 2013, Yoo et al 2012; Gunther et al 2017). According to one line of research into properties of data, the unbounded character of data stems from data being continuously editable (Alaimo et al 2020) and non-rival (Shapiro & Varian, 1999; Alaimo et al 2020). This means

Work in progress - Do not cite

that data can always be modified and once used data as a resource is not depleted. Furthermore, data has been characterized as being dynamic (Kallinikos et al., 2013; Yoo et al., 2010), and objective (Davenport & Prusak 2007). Since data represents objective facts of the world, this representational capacity of data stays the same when data travels to new contexts. Hence, the argument goes, data can always be modified for new contexts and in principle all data can be used by everyone.

But, the scholarship on data is becoming increasingly varied and growing in size with critical voices making strong arguments against the supposed unbounded character of data. For example, data has recently been characterized as dissimilar (Jones 2019; Kitchin & McArdle, 2016), contextual (Jones 2019; Strong et al. 1997), constructed (Neff et al. 2017; Kitchin & Lauriault 2014), 'cooked' (Gitelman & Jackson 2013; Jones 2019) and 'dirty' (Muller et al 2019). These critiques of data unboundedness highlight how the way data was produced and what consequences data production has on how data can be (re)used (Jones 2019; Gitelman & Jackson 2013). The critique maintains that data is intrinsically local, situated, and theory-laden, thus having no meaning or truth value outside of the context of use. So, not only are data constructed and "cooked", but these are properties that cannot be disentangled or cleaned from some "pure data" which can be recontextualized or reused freely (Gitelman & Jackson 2013; Kitchin & Lauriault 2014; Jones 2019). This situated and constructed nature of data arguably influences the extent to which the data can be reused for purposes beyond the original purpose. These critiques form a great challenge for the potential value creation through data sharing and reuse. So, the debate emerging around the issue of (un)boundedness is relevant for understanding the potentials and challenges of data reuse, because the ability to create value through data sharing and reuse depends on the ability of organizations to use data in contexts different from the one data was produced in and for.

Data sharing and reuse Organizations are investing increasing amounts of resources into establishing data sharing ecosystems in an effort to create value through interorganizational, collaborative data reuse (Gehlaar & Otto 2020; Lis & Otto 2020). Disparate literature on data sharing suggests that data is not necessarily readily available or useful for reuse and thus cannot be aggregated without considering the context of its production (Pine 2019; Bowker & Star 2000). This line of research supports the critique of data's unbounded character, by illuminating the practices involved in the recontextualization of data needed for data reuse (Birnholtz and Bietz 2003; Gitelman & Jackson 2013; Leonelli 2014; Rolland and Lee 2013).

Work in progress - Do not cite

This work highlights the importance of understanding data, meaning learning how data production is situated and enacted through 'localized work within social, cultural, and political contexts that in turn shape the production and interpretation of data' (Pine 2019). So, even though some data can be transported and recontextualized, this is not an inherent property of data, but an outcome of practices involved in uncovering the historicity of data journeys for this to be successful (Leonelli & Tempini 2020; Gunther et al 2017). As Neff (2017) clearly illustrates, 'the work of making and analyzing data is a journey, not a destination, the product of layers of contributions from multiple people; so data often lead to new questions'. Hence, to uncover what data is and how it can be shared and reused, we need to look into the data work involved in data's "journey".

Data work Data work refers to practices of organizing, analyzing, judging, and decision-making concerning data (Foster et al 2018; Bjørnstad & Ellingsen 2019). Research on data work has its roots predominantly in studies on data practices in the healthcare sector (Berg & Bowker 1997; Cabityza et al 2019; Bjørnstad & Ellingsen 2019; Dixon-Woods et al 2012; Elingsen et al 2018; Holten Møller & Bjørn 2011; Pine 2019). These studies emphasized that data is not an independently existing entity and that it can only be understood in its wider sociotechnical context. Furthermore, in agreement with critical voices in the debate on data (Jones 2019; Gitelman & Jackson 2013), data work studies highlight the importance of making visible the practices through which data comes about. These practices often remain hidden due to the social status of people that are usually involved in most of the data work practices such as administrative staff, labelers, and other data workers whose practices are deemed as scutwork (Pine 2019). Studies on data work have expanded to new contexts and repeatedly showed that data work practices require much effort, involve human judgment, reflect political choices (Foster et al 2018; Pine 2019) and sometimes require intensive sensemaking involving multiple stakeholders to put the data in context (Fischer et al 2017). These insights are particularly illuminating in the context of data work for ML. Currently, ML development is often described in a linear and technical way (Lehr & Ohm 2017; Muller et al 2019). Yet, research on data work suggests that many aspects of development are not being highlighted or uncovered with such understanding of ML development.

Representations Opportunities and pitfalls of data reuse both rest on data's representational capacity. Data appear to be in a correspondence relation to a phenomenon it represents (Bailey et al., 2012; Knorr-Cetina, 1999). This supposed representational character of data promises

Work in progress - Do not cite

creation of objective knowledge about the world. Scholars argue that because of the ability to decouple digital representations (or in this case data) from a physical device that produced data, representations promise a radical transformation of work, yet this promise is faced with skepticism due the inherent inability of representations to capture the complexity of the phenomena it stands for (Bailey et al 2012; Monteiro & Parmiggiani 2019), but also critiques because of the detrimental effects reliance on representations can have on expertise (Zuboff 1988). For example, Bailey et al. (2012) shows that in simulations, when representations of vehicles do not match their referents and there is no way to empirically validate them, engineers can neither analyze vehicle performance in an informative way nor find solutions to known problems. Similarly, Monteiro & Parmiggiani (2019) find that sensors detecting sand in oil wells located at the deep-sea levels need to be continuously verified, while representations they produce require expert interpretation accounting for the context of production. Thus, representations cannot reliably stand for their referents, as they are unable to portray the full complexity of the phenomenon they represent. Moreover, in instances where verification is more difficult or even inherently impossible, issues with working with representations are amplified because there is no possibility to inspect the correspondence relationship between a phenomenon and its representation.

Four lines of research we discussed bring many insights for data reuse. The debate on data (un)boundedness is informative for data reuse as it investigates the constraints which can hinder data reuse. But, this is also a key gap we identified as research into how (un)boundedness is actually established in practice is still under developed. Uncovering this issue is important as it would provide us with insights into how data workers can facilitate data's generative potential and realize value from data for organizations.

3. Methods

We conducted a qualitative study on five data science teams tackling predefined challenges related to agriculture through data reuse. The context of agriculture is particularly interesting as it involves global challenges of food security and climate change making this setting as timely as ever. Moreover, agriculture is inherently an interdisciplinary endeavor which requires knowledge spanning several disciplines such as biology, meteorology, and economics. Hence, agriculture is a prime context for studying reuse of diverse data. Furthermore, natural sciences are often spoken of as 'hard' sciences involving objective and exact data. If such data is indeed

Work in progress - Do not cite

'hard', its reuse should be easier since data should preserve its correspondence relationship to the world in novel contexts.

We adopted a practice lens in our study as we aimed to see how data reuse was enacted through ongoing activities, without assuming some inherent properties of data (Feldman & Orlikowski 2011). So, in our interviews, we ask participants to describe in detail their actions during a two-month hackathon from the moment of registration to the end of the final hackathon event. We conducted twenty-one semi-structured interviews with participants (~1 hour each), observed five introductory webinars (~30 minutes each), and three final hackathon events (~2,5 hours each). Overview of the teams is given in Table 1. Studying hackathon teams is highly suitable to address the question of how data scientists reuse data. The datasets that the teams used were not pre-fabricated, and they were produced for different purposes then those of the teams. Furthermore, the short team duration of two months offered us a "pressure-cooker" situation that makes it convenient to observe a full process of development, from finding data sources to evaluating models. Also, the variety of phenomena and methods across challenges provides a rich research context. Of course, its downside is its generalizability, which we will discuss at the end of the paper.

We analyzed our data in several phases. In the first phase, we structured our data in case-based narratives that described in detail the work involved in development of datasets. Through comparison of narratives, several issues came to the fore. Participants struggled with framing the right research questions for their projects, understanding what appropriate data means, and understanding how the output of their tools ought to be represented and understood. This led us to engage with the literature on data work which studied the need for understanding data production in data related projects. So, in the second phase, by continuously going back and forth between the cases and the literature on data and data work, we identified challenges and actions that data workers engage in for facilitating data reuse across the five cases. We created event lists for each case that explicated data related actions and challenges that emerged during the hackathon's two-month period. We noticed a difficulty in describing the relationships between challenges and actions in a stepwise way and, consequently, finding clear developmental phases was difficult. This was surprising given the clarity and linearity with which ML development is usually described. Moreover, challenges and actions we listed were often unexpected and required more than mere technical work and expertise. To address this issue, in the third phase, we temporally bracketed (Langley 1999) our data in three phases:

Work in progress - Do not cite

collecting data, preparing data, creating a dataset, and listed assumptions, realizations, and reactions related to data that occurred in each of the case for each of the three bracketed phases. This enabled us to see more finegrainedly the creative actions of developers in each phase, as they faced specific challenges such as coping with fragmented data, diversity of data, and complexity of ML models. By aggregating the actions and challenges according to the three temporal brackets, we came to three broad themes of digging into domain and data expertise, creativity of data work, and evaluation of newly constructed data. In the fourth phase we structured data from the desert locust team in a case-oriented matrix based on the three themes and searched for a general practice associated with each of the themes (Yin 1984). Then, we compared all other cases to the resulting matrix and saw a great fit with the resulting practices, with all of the activities within a theme having a common goal of producing a specific outcome: phenomenon definition, training set, and evaluation, respectively. From this, the practices of problematization, creative data work, and scrutinizing datasets emerged. But, we also saw a need for more finegrained structuring of data, as activities and challenges within the same general practice involved various motivations, strategies, and impact. So, we grouped the activities in a total of seven subpractices associated with specific data work challenges teams had to address. The practices, subpractices, and outcomes we identified are shown in Table 1.

Work in progress - Do not cite

Table 1 Overview of practices, subpractices, and outcomes

Practice	Problematization	Creative data work	Scrutinizing datasets
Subpractices	Digging into domain expertise; Digging into data expertise; Interlacing domain and data expertise	Integrating data into workable datasets; Creating representational proxies	Scrutinizing ML workability of datasets; Scrutinizing the consequences of use
Practice outcomes	Phenomenon definition	Training set	Evaluation

We opted for using the conceptual composition of reporting our findings, due to a complex nature of problems developers tackled and a larger number of theoretical constructs involved (Berends & Deken 2019). So, we introduce the main concepts beforehand and use them as 'theoretical signposts in narratives that follow and later connected in a theoretical process model'(Berends & Deken 2019). So, we organize our findings according to three main phases, each characterized by a practice that emerges and ends during that phase. Each phase consists of several subpractices enacted for data reuse which gives rise to a general creative practice of coping with the issue of data reuse. Conceptual composition also enables us to present our findings in a space-effective way (Berends & Deken 2019), which is valuable for us due to a larger amount of rich cases. So, the case narratives are not presented in a fully inductive manner and due to considerations of space we use representative data to illustrate concepts we developed. But, this compositional strategy also highlights the theoretical relevance of our findings and shows a strong link between our data and the process model that emerged from it. Also, we make up for the lack of narrative display through continuous use of tables. In Table 2, we provide an overview of the five cases.

Work in progress - Do not cite

Table 2 Case overview

Case	*Land Boundary Detection*	*Desert Locust Outbreak*	*Weather forecast*	*Composite maps*	*Agriclimatic factors*
Aim	Develop a tool for detecting boundaries between crop fields	Develop a tool for estimating the impact and movement of desert locust swarms	Develop a tool for forecasting weather over a small 100x100 meter crop field	Develop a tool for producing agricultural production advice	Develop a tool for calculating agricultural phenomena from climatic data
Participants	Two geomaticians and two machine learning experts	Four remote sensing specialists, one agricultural business owner, one academic agronomist, and one data journalist	A machine learning expert, two software engineers, and a geomatician	An industrial engineer, four academic agronomists, one farmer, business master student, and agriculture master student	Five geomaticians
Data used	Data from a regional Land parcel identification system, Sentinel 2 satellite data (13 bands)	Sentinel 1 radar data, Sentinel 2 satellite data (4 bands), weather data from an international meteorological organization (12 variables), locust GPS locations from UN Food and agriculture organization	Sensor data from a small university owned field (4 weather variables), weather data from a continental meteorological organization (5 weather variables).	Prices of goods (various sources), Yield data (personal and university owned data), Land ownership data (government data), weather data from an international meteorological organization, Sentinel 2 satellite data	Weather data from farmer owned weather stations (various weather variables), weather data from a continental meteorological organization (5 weather variables)

Work in progress - Do not cite

4. Findings

We structure our findings around the three main practices of data reuse for ML projects. We present our data in the form of temporally ordered practices of problematization, creative data work for ML, and scrutinization of data sets. These practices consist of several subpractices data workers enacted to tackle specific changes and produce an outcome of a practice which is then used as input for the subsequent developmental work.

4.1 Initiating a data reuse project in the context of ML - Problematization

The initiation of ML development is often not a straightforward issue for developers as it requires development teams to face the complex nature of phenomena they aim to capture with their training sets. This means that developers have to formulate their developmental trajectory with respect to a particular goal they wish to attain, e.g. milestones to develop a tool for improvement of agricultural practices. As a result, developers need to understand well what a particular need the tool they are developing has to solve which requires involvement with, and understanding of, domain specific requirements of their tool. This issue can be even harder to solve in case of data reuse, as developers are constrained with respect to available resources they can use to model a phenomenon.

So, early on each team aimed at agreeing what exactly the target phenomenon was, what knowledge about the phenomenon they wanted to produce with ML, i.e. what type of outcomes should the resulting model provide, and how that tool can help agricultural practitioners to perform their work. Also, developers looked closely how their data relates to the target phenomenon. So, each team aimed at inspecting how well the available data can represent a phenomenon and how can data be refined into a better representation of that phenomenon. To achieve this, the teams also explored what similarities and differences there were between what data represented and what the phenomenon was to see if there are opportunities or threats for constructing workable representations from available data. In the following subsections, we provide examples and explanations of the two subpractices developers engage in. We show how these subpractices result in developers constructing a *phenomenon definition* as the product of the general problematizations practice. We present the practices of problematization across cases in Table 3.

Work in progress - Do not cite

Table 3 Problematization

Case	*Land Boundary Detection*	*Desert Locust Outbreak*	*Weather forecast*	*Composite maps*	*Agriclimatic factors*
Subpractice – Digging into domain expertise	The team needs to understand the need that land use agencies currently have, while thinking of other potential uses of reused data. Hard to define a goal because of the ambiguity of the word 'boundary'. Explore different ways boundary can be understood and how it aligns to potential data sources.	It is not possible to observe the locust, while it is dangerous for livelihoods of people, so there is a great need for a tracking system. The team also know little about the locust so they search for academic articles on the topic to be able to think of potential proxy phenomena.	There is a need for more local forecasts that farmers can rely on. There is a need for more long-term, yet reliable, forecasts to manage agricultural practices in a better way.	The team needs to know which exact questions would be relevant for the biggest amount of farmers and if those questions can be translated in computational terms. They consult agricultural experts on the matter.	The team did not know a priori which factors are of special interest for farmers. The team interviewed farmers to see what kind of phenomena they would deem interesting and relate them to the team's agricultural expertise.
Subpractice – Digging into data expertise	The team needs to define what kind of outputs a resulting model should provide and define what kind of datasets the team should develop. The team also investigates how available data is structured to identify potential uses, but also deficiencies of sources.	It is hard to define a goal because there is no particular data referencing the actual locust. The team understands that even the locust GPS points are actually just sighting reports. They realize a need for establishing a good data proxy for the locust.	The team investigates closely what data they have available and look for relationship between them. They perform statistical analysis on the data to evaluate the potential of 'uncovering hidden patterns'.	The team needed guidance on which exact data to search for as they were not sure what exactly the phenomenon they should represent is. They consider what data might be available for them to inform them of potential phenomena to represent.	The team did not know what data they can use for local predictions of events, while they realized global data has issues with the accuracy and internal uncertainty due to microclimatic differences.
Subpractice – Interlacing domain and data expertise	The domain expertise informs the way satellite images need to be constrained to be able to represent the land boundaries in a workable way. In the same	Digging into domain enables the team to consider which phenomena can serve as proxies for the locust – e.g. vegetation change. On the	By learning about the data, the team sees potential opportunities of combining local and global weather data, due to correlations between the	The team creates a framework explicating three questions they want to answer – what and where to plant, and where to sell – each associated with a	The team iteratively compares available weather variables and phenomena farmers highlighted during the interviews. By comparing the

Work in progress - Do not cite

	time, LPIS data is used to define what a land boundary is.	other hand, understanding the opportunities and deficiencies of satellite and radar images enable the team to work out a way to represent vegetation change specifically for tracking the locust.	two. Yet, the team lacks the domain expertise to interpret what underlines those correlations, so they search for alternative strategies to find important parameters to consider	list of data sources that can potentially be used to answer the questions	two, they map one onto another to find which factors to make calculations on.
Outcome – Phenomenon definition	Land boundary is defined so that it fits the way data is structured in the LPIS.	Read blogs and papers on the problem, while relating their findings to potential data sources. They realize vegetation can serve as a proxy and they can use satellite data for it.	They combine similar, yet different weather phenomena from two data sources in case the pair seems to correlate enough to be treated as a single phenomenon.	The team infers from the general framework the notion of best practices as a guide to where and when to plant a particular plant and where to sell it.	The team formulates a list of 7 factors that farmers have highlighted and the team believes they can be calculated from available data.

Work in progress - Do not cite

Digging into domain expertise Development teams we studied were faced with an issue of understanding how exactly their tools can improve actual agricultural practices. This was crucial for these teams as they wanted to ensure the relevance of their solutions for potential end users. Also, they believed that understanding the need can help them realize what kind of data they would need to search for to address that need. Hence, overcoming the issue of finding potential data for reuse and defining data requirements for ML involves developers learning about the practices they want to improve and relate them to the potential inputs that some algorithms can work with.

All five development teams engaged in this subpractice in their first meetings where they discussed aims of their respective projects. To illustrate, the composite maps team was early on faced with a confusing situation in which different participants had very different ideas on what need they wanted to satisfy for farmers. As an agronomist that participated in this team explained:

> *First, the challenge was to ourselves what really do we want to do and why we want to do something. So in the course of the discussion we came to, I was prompted by the fact that we've not asked ourselves these questions. What we needed to get from this question is to realize what data needs to be available to answer the question. (Agronomist, Composite maps)*

This required the team to discuss which problems exactly they wanted to solve and then if those problems can be framed in terms of a ML problem. Moreover, as the team mentor explained, their solution needed to have enough information for the advice to be understandable by farmers that would use it:

> *'[Our aim] is to detect the good and the wrong places on the field and you need to convince farmers that this is reality and for example if there is a discussion with them why this part of the field is bad for this and this reason. So it is necessary to then have such dialogue in this stage not only to offer farmers a "black box". They want to have some evidence and to explain to them what you are doing with this data.' (Mentor, Composite maps)*

As these examples illustrate, the issue of understanding the need presses developers to understand the domain and think about the way that their tools will be embedded in the practices they are developing the tool for. This is important for the developers, as understanding

Work in progress - Do not cite

the need can help them in agreeing what exact data they can search for, as well as in what way they can define the kind of outputs their ML tool can have.

Digging into data expertise Besides understanding the needs that agricultural practitioners have, the teams also had to figure out what kind of a training set they need to develop and how to evaluate the usefulness of data sources available for reuse. In doing so, the developers worked on agreeing what kind of training data, and ultimately model outcomes, they wanted to produce. The challenge to constructing a dataset is particularly salient in case of data reuse, as developers are dependent on the existing data, which was not produced, formatted, or even stored for the purpose that they want to use it for. Hence, developers can face the problem of potentially useful data being fragmented or inaccessible. This was especially evident in the case of the composite maps team which struggled with the fact that land ownership data is both fragmented across national databases and often inaccessible due to privacy concerns. This is why other teams relied on open data, with the exception of the agricultural factors team which reached out directly to farmers to collect their data.

Furthermore, since the data that teams searched for was produced for a different purpose, developers faced large discrepancies between what the available data represents and what aspects of the phenomenon developers need to represent. As the case of the desert locust team nicely illustrates, when there is no data available on the phenomenon, the path to defining a phenomenon can be hard. The team was considering what data they can use to represent the locust and they found that the only available data directly referencing the locust were GPS locations of reported locust occurrences. But, as one participant explained, when looking into the data, the team soon faced an issue:

> '*there's not that much information in the GPS points. Basically, you have like the name of the city or the village, and then you have the GPS location, and the date. So we were missing some information to identify how big it can be.*'*(Geomatician 1, Desert locust)*

In this example, we can clearly see how the differences between what the data represent and what the developers want to represent can pose problems. This issue can be further complicated when there is a lack of domain knowledge. The team working on the desert locust didn't know much about the insect prior to the challenge, besides hearing about infestations. As one participant pointed out:

> *'I didn't know what to do, because I'm not like a biologist guy. So I didn't know at all how we can measure the size of the swarm' (Geomatician 2, Desert locust)*

Work in progress - Do not cite

Interlacing domain and data expertise After digging into the available domain and data expertise, the team looked for ways to use this information to construct a phenomenon definition that will be used to define what training sets and model outcomes they need to reach. This work involves interlacing domain and data expertise the developers dug into. Interlacing knowledge involves examining '*what, how, and why of the various [design] options*' that enables developers '*to recontextualize and transform that knowledge to improve or even radically alter their own designs*' (Tuertscher et al 2014). An interesting illustration of interlacing comes from the agriclimatic factors team that reached out directly to farmers and interviewed them about the kinds of phenomena that would be interesting for them, to ensure they understand the need of actual agricultural practitioners. As one participant said bluntly: *'we know that what we call agroclimatic factors is something of interest to farmers because we discussed with farmers what they want'(Geomatician 1, Agriclimatic factors).*

As they considered which phenomena they can represent, the team investigated how various data sources relate to specific phenomena highlighted by farmers. The team came to the idea to use satellite-based climate data, but as one participant explained, there can be quite a discrepancy between what the data states and what is the actual state of affairs:

> *'If we talk about the temperature, we can get influences like if you are by the river or by water, you can expect different micro climate than if you're somewhere else. This is why it is hard to use the global data for this. Global data can give us some, let's say some overview, but then you usually need some meteorological station to work with.' (Geomatician 2, Agriclimatic factors)*

Since the team wanted to produce a visualization of factors across a region, e.g. probability that frost will be present on a specific date in a specific location, they needed data that is more fine-grained than global meteorological data. Yet, when they turned to local, farmer owned meteorological stations, they found issues too. While global meteorological data is uniform across the globe, not all-weather stations are the same, because they don't necessarily measure the same kind of phenomena – some measure multiple temperature related phenomena, but not wind, while others measure precipitation and wind, but only one temperature related phenomena, for example. Moreover, even if they measure the same kind of phenomena, they are not necessarily the same kind of instrument, meaning that their accuracy and fine-grainedness might diverge too. To solve these issues, the team decided to define a set of agricultural factors that match the insights from interviews they conducted, but

Work in progress - Do not cite

that are also least prone to complications due to the specific issues that the two identified data sources can have.

To illustrate a different strategy, the land boundary team aimed at identifying boundaries between crop fields and one issue they encountered was that a "field" can be understood in relation to plants as an area where they grow, in administrative terms as an area that is registered with the municipality, or as an area that is physically surrounded with a fence. The team had access to a regional land parcel identification system (LPIS, which contains images of fields with marked boundaries based on farmers' reports. Those farmers' reports defined a field as a continuous land covered in a single crop and owned by a single farmer. As they found out about the definition of a field from the LPIS, the team decided to define a boundary as the edge of a LPIS documented field. This also enabled them to use LPIS data as labels. Interestingly, as the team wanted to use LPIS data to label satellite images, one participant soon pointed out a problem:

> *'the crops are growing between May and July. Then it's cut. [We] were all the time discussing the best season for the choice of [satellite] data. If you, for example, choose autumn you cannot see those boundaries properly because there is no crop.' (Geomatician, Land Boundary)*

So, as the team learned more about the domain and the data, they were able to anticipate challenges that they will be facing in the course of constructing workable training sets for ML. These examples show how the interlacing of the interpretation of the phenomenon and the interpretation of data can be used to guide definition of the developmental goals. Moreover, this also shows that, besides it being important to identify information needs to see what kind of data might be relevant for representing the phenomena, it is also very important to learn about the data so that developers can see what reuse opportunities they can leverage. So, data reuse crucially depends on interlacing knowledge about data with the knowledge about the domain.

Work in progress - Do not cite

4.2 Creating a training set from reused data – Creative data work

The main objective that the development teams had was to overcome the challenges stemming from the tension between a phenomenon definition and the ability to represent that phenomenon with available data coming from diverse sources. Each team aimed at constructing a workable and representative training set, meaning that the training sets the teams made had to be adequate for the purposes of ML, in terms of their format and size, but also

representative of the phenomenon in a sense that the resulting model should provide informative outcomes for end-users. To achieve this, the team had to face and overcome the heterogeneity of data coming from different sources, as well as the complexities of working with diverse and large data sets. In our cases, we found that these challenges triggered the development teams to engage in the practice of creative data work. Hence, the creative work consists of two subpractices: integrating data into workable datasets and creating representational proxies. We present the practices of creative data work across cases in Table 4.

Work in progress - Do not cite

Table 4 Creative data work

Case	*Land Boundary Detection*	*Desert Locust Outbreak*	*Weather forecast*	*Composite maps*	*Agriclimatic factors*
Subpractice – Integrating data into workable datasets	The team is faced with a need to do multiple transformations to be able to use such diverse data. The team relies on domain and data expertise to brute force the data into a unique format.	To produce a proxy representation of the locust, the team needs to integrate satellite images, radar images, weather data, and GPS locations. The team seeks outside expert help for combining all data sources.	There is a great discrepancy in the frequency, resolution, and measures coming from two available data sources. The team relies on statistical analysis and time warping algorithm to integrate the two datasets.	The team is faced with an increasing amount of identified data sources needed. The team performs statistical analysis and consultations with experts to find which data can be made obsolete.	Available data sources exhibit great diversity which is a challenge. The team searches for the biggest common set of variables present in all sources and builds their training set from that data.
Subpractice – Creating representational proxies	Representing fields in a rich way is in tension with representing fields outside of the embedded environment fields are in. The team considers temporal constraints they can place on satellite data and the richness of resulting representations. Moreover, there is a tradeoff between the number of types of satellite imagery and complexity of their model. They enforce strict temporal constraints, but then also use all types of images.	The team faces a tradeoff between the amount and finegrainedness of weather data and the complexity of resulting data. The tradeoffs are directly related to the type of proxy representation the team can make. They use statistical analysis to make data choices. They also face representational tradeoffs between satellite and radar images and opt for overcoming deficiencies of both through integration.	The team faces the issue of raising complexity of their dataset, so they decide to partition their data into one-to-one pairs of weather variables from the two datasets. To further reduce complexity, they use autoencoders to extract relevant features of datasets through machine learning.	The team sees great differences in the level of finegrainedness of data and choosing one level over the other enables the use of some data while it makes other data unusable. They decide to use data coming from a single organization to ensure it can be integrated due to same data governance requirements that applied to all data types.	The team sees a need for making data choices regarding the amount of weather variables that is ultimately included in the training set, as they anticipate the issue of complexity that might arise. They rely on their domain expertise to make computationally most efficient calculations.

Work in progress - Do not cite

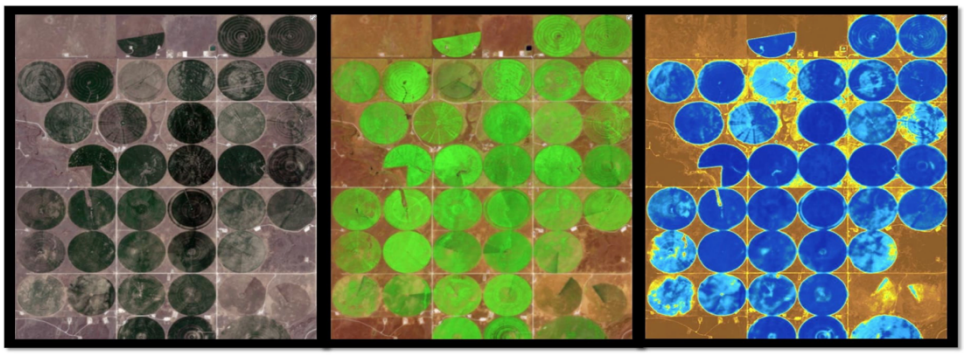
Outcome – Training set	14-layered images – 13 layers of transformed satellite images with the 14th layer being the labels created form LPIS data	Timeseries dataset correlating vegetation and weather change with reported locust occurances.	Three datasets integrating specific pairs of variables produced by autoencoders.	The team does not produce a training set, but a collection of diverse datasets. They lack a framework to deal with data reuse challenges.	The team produces 7 training sets for each factor respectively by integrating the two data sources.

Work in progress - Do not cite

Integrating data into workable datasets Developing workable ML tools always depends on usable ways to integrate large and diverse datasets coming from different sources and the same is the case in data reuse. Yet, standardized and reliable strategies to overcome these issues do not exist, as the specific issues that can emerge between two data sources depends on how the data was created, curated, and what is the context of its use. So, data diversity was a challenge faced by all development teams and this issue was particularly salient in these cases of data reuse because data that is being used is not produced for the purpose of being combined together. This makes the work of transforming and integrating data even more challenging.

To illustrate, for the land boundary team this meant a great deal of inspecting images they collected. Based on the way the team interpreted what a field boundary is and how satellite images and the LPIS data represent the details of crops, the team discussed how to construct the training set from the data they had available. The team was faced with the decision on which out of 13 different types of satellite images they can use. Sentinel 2 satellites capture images that differ in terms of the light-wave they capture, as well as in their resolutions (having a 10, 20, or 60-meter resolutions). So, the team had to judge on the ability to use these images to accurately represent fields. The tradeoff was between choosing a larger amount of data which they can use for training by using all types of images or selecting only more fine-grained types of images. Another potential issue was that not all images could show clearly the boundaries between fields as they capture diverse aspects of nature. Figure 1 illustrates how three different combinations of several satellite bands can represent the fields differently.

In progress - Do not cite



This example shows agricultural fields in various Sentinel-2 band combinations: Natural Color (left) displays the optical wavelengths our eyes naturally detect, Short Wave Infrared Vegetation (Middle) showing the most vigorous vegetation in bright green, and a Water Moisture Index (Right) with the highest moisture levels shown in blue.

Figure 1 Differences and similarities of satellite data representations

Because of the number of images they had at their disposal, the team looked at the small sample of them and compared them. Using a naked eye, the team agreed that they share enough similarity that they can all be used – e.g. when they selected a particular field and checked all 13 images, they could have seen the boundaries themselves. As Figure X (above) illustrates, there is a correlation between moisture, vegetation, and color, among other phenomena different satellite images capture. Nevertheless, the geomatician raised concerns about using all of the types of satellite images, because in remote sensing *'the combination of satellite bands is a cornerstone for a good result' (Geomatician, Land Boundary Detection)* and using all of them can lead to many problems pertaining to the transformation that need to be done to integrate all of them. Yet, as one of the machine learning experts argued, the team already lost a lot of data due to limiting images to spring time:

> *'for machine learning you need a lot of data and we didn't have many anymore. We probably didn't have enough data as the net needed to train it'. (Machine learning expert, Land Boundary Detection)*

Note, the team decided to use only the images taken between May and July, so they leverage the fact that fields have crops on them, making it easier for them to see where the boundaries between different fields are. The mentor agreed with the concerns over the amount of data they had on their disposal, but integration of all of the images was a great challenge. The team had to solve the problem of different resolutions (10, 20, or 60 meter resolution), as well as the fact that the LPIS data was expressed in a different coordinate system than the satellite images. This required the team to perform multiple transformations of images so that they can all be mapped one onto another. While the machine learning experts solved the resolution problem and patching satellite images together, the geomaticians performed geomatic transformation of LPIS expressing its coordinates in the same system satellite images were formatted in. After combining all of the images, the team produced 14-layered images of crops that should have correlated various aspects of crops with the existence of reported boundaries. Yet, the team was concerned with how well their training set can be used due to, what they regarded as, the lack of training examples. Moreover, they were concerned with how well the outputs of their ML model will look like and if they will be useful for detecting the land boundaries.

Work in progress - Do not cite

These issues were encountered by all teams, since the data they were acquiring came from multiple organizations, with different data governance systems, formats, frequencies, etc. The example above illustrates an issue that any interorganizational data sharing that does not have common standardization will face.

Creating representational proxies The second issue the teams faced was that of evaluating tradeoffs when creating proxy representations of their respective phenomena. Multiple problems can emerge that developers face when creating proxies. Different data sources can have both complementary and competing ways of representing a phenomenon. So choosing one over the other or integrating both can be complicated and with far reaching consequences. Also, developers need to consider what computational resources are available and what developers can do with it. Here, the tradeoff is between how many different data sources are being combined and how complex the model resulting from that training set will be. This is an important tradeoff to consider, as the use of multiple resources was seen as needed to create a representation of defined phenomena from reused data, yet by including lots of different sources or types of data,, they might not have the computational power needed to actually train or run their model. This issue was particularly evident for the weather forecast, agriclimatic factors, and desert locust teams as they tried to estimate how many different weather variables they can use without rendering their data too complex. Hence, developers need to strike a balance between representations being created and model complexity.

To illustrate this challenge and how it was addressed, consider the desert locust team which discovered that, since the danger with the locust is precisely in it eating vast amounts of vegetation, sudden and widespread changes in vegetation were known to be linked to the locust infestation. Furthermore, they found that the desert locust numbers and movement are linked to rain since locusts tend to lay eggs in moist areas. During a team meeting, participants brainstormed ideas on how they could use this information for solving the problem of representing the locust. A remote specialist from France had experience with proxy representations, as he worked on a project where he tracked immigration by representing change of urban environments through satellite imagery. He argued that natural phenomena such as vegetation and weather change might serve as reliable proxies for the locust in the same way. As he explained:

Work in progress - Do not cite

[we were] noticing that there is like specific temperature or specific precipitation [connected] to it. So soil moisture was [also included in these factors]. And then maybe you can add other data, but I mean, with three types of data you already have, a lot of data to collect, and you have a lot of information to analyze to get something interesting. (Geomatician 2, Desert locust)

So, the remote specialists suggested that they can use NDVI – a vegetation index derived from satellite imagery whose sudden changes can be indicative of desert locust location. But, as another remote specialist pointed out, the satellite imagery has issues of so-called "cloud cover" (clouds stand in a way of the light being reflected from Earth), and therefore creates poor representations of vegetation during rain. As he explained, this was a great disadvantage:

the desert locust crisis is happening after some meteorological events, like rain, a specific temperature on the ground, soil moisture, I don't know, so I knew that maybe the clouds will be a problem as [the crisis is] happening after rains. And so, if we are missing the time when we have the locust we're losing some information. (Geomatician 1, Desert locust)

As an alternative the team considered using radar images. Unlike the satellite which works by collecting light reflected from the Earth, radars send their signals which can penetrate the clouds and are collected once they contact Earth's surface. But, the way radars work also means that they can provide information on the texture of land cover, but not also on the amount of vegetation. As the French remote sensing specialist explained:

Radar measurement is based on the signal that's touching the ground and let's say the shape or the texture of the ground, but then when you are like… sometimes when you have a signal on the mountain and signal on the forests, you will get the same signal. (Geomatician 1, desert locust)

The team also had to decide if and how to incorporate weather data in their training set. They believed that if they could combine images with weather data, they could construct a reliable representation of the desert locust. But, the team worried that including too many weather variables would render their model too complex. So, they had to make a judgement on the number of different variables that they would use. The French remote sensing expert had

Work in progress - Do not cite

some experience with ML and he suggested that keeping more than three variables would greatly increase the complexity of their model, while this would not greatly improve the predictive power of their model. The team first decided to list several variables and correlate them with the time stamped GPS data on locust presence by doing statistical analysis. Finally, the team was faced with choosing between the numerous potential data sources which ones to use and in what amount, making evident the representational tradeoffs they had to make.

4.3 Identifying emerging issues – Scrutinizing datasets

After developing the training sets, the main objective that the teams set out to achieve is evaluating their training sets. The teams were interested in how well they managed to construct a training set that represents a phenomenon and that can be used as input for ML. Moreover, as they have also faced potential challenges related to consequences of use and workability of their training sets for ML, the teams engaged in the practice of *scrutinizing datasets*. In doing so, development teams inspected potential issues that can emerge from the use of tools based on their training sets and realized how consequential the tasks they performed in the creative work were on their training sets - which often surprised them. While investigating what have happened during the creative data work and what the consequences for their project are, the teams realized that there is a discrepancy between what their training sets represents and can achieve, and what they initially set out to do with them. Moreover, they realize that the complexities of data have increased and that tradeoffs they made at start have had a large effect on what they can do with their training sets now.

The main outcome of scrutinization is the *evaluation* of their training set. In this evaluation, a development team pin points specific issues that can emerge due to the way a dataset is constructed. As a result, development teams realize that they need to work again on improving and reinterpreting their understanding of the need they are trying to solve, as well as to improve their understanding of the newly constructed data. This practice shows how anticipation of issues that can emerge from future use impacts the developmental practices while the project is still running. The concern that the developers had comes directly from the fact that data is being reused and, after seeing how unexpected issues can emerge while producing datasets, developers operate under great uncertainty over how that data will impact the risks connected with the future tool use. We present the practices of scrutinizing datasets across cases in Table 4.

Work in progress - Do not cite

Table 5 Scrutinizing datasets

Case	*Land Boundary Detection*	*Desert Locust Outbreak*	*Weather forecast*	*Composite maps*	*Agriclimatic factors*
Subpractice – Scrutinizing ML workability of datasets	Numerous transformations led to loss of large amounts of data, while data that was left does not represent the boundaries in the same way as they defined them in phenomenon definition.	The team considers how good their data is for the problem definition. They find specific issues for areas near water or inability to differentiate between phenomena that can have similar data trace (e.g. forest on a mountain and in a plaine).	Some datasets perform well, while others have a bias in predictions. The team considered how well their datasets represent phenomena, but due to lack of domain expertise cannot evaluate them properly.	The data the team collected is, in general, not good enough for ML as it is in unusable formats, often outdated, and lacks crucial metadata. The team needs to improve on the quality and format of their data/	The team realizes that they cannot capture everything they planned due to specific discrepancies between some data and phenomena they wanted to represent. They modify the factors based on specific representational problems data has.
Subpractice – Scrutinizing the consequences of use	The team deliberates on the use of their tool and realizes the need for perfect accuracy as their tool should be used for determining amounts of subsidies and taxation for government agencies. So, mistakes can bring large negative consequences both on farmers and agencies.	The team is concerned about the possibility of their tool having false negative predictions, this not alarming the population about an incoming infestation which can have grave consequences on the livelihoods of people living in those areas.	The team notices that there can be issues with explainability of model outcomes due to the stacked architecture they created. They notice a bias in outcomes, but due to lack of expertise cannot identify how to fix the bias. They see this as problematic for end users.	The team agrees that their initial phenomenon definition was too broad and use of such diverse data can lead to lack of explainability with respect to the causes of certain recommendations the tool would make to farmers. On the other hand, lowering the amount of data needed for the same phenomenon can lead to a lack of justification for recommendations.	There is a great risk attached to the uncertainty of data being used. While small uncertainties can sometimes be nonconsequential on the end-users, for some factors that same data can have detrimental consequences on agricultural practices.
Outcome – Evaluation	The team sees a need to redefine the phenomenon based on identified need for more understanding of constructed data and need that exist.	The requirement to faithfully represent the locust was dropped and the ability to estimate which communities to warn of danger became the focus for redefinition.	The team discusses other parameters that they had to use for training the model, and seek domain expertise needed for redefinition of phenomena.	The lack of data made them unable to define the phenomenon and due to the lack of time, the team did not manage to fix the issue. This case stresses the need to substantially connect the phenomenon with available data to be able to solve the challenge.	The team sees a need for redefining some factors to cope with uncertainty of data. They seek consultations with farmers and analyze data to come up with new definitions of phenomena.

Work in progress - Do not cite

Scrutinizing the consequences of use Scrutinizing datasets involves considerations of explainability or utility of tools for end users, as well as risks that might emerge through use due to the type of data that was reused. These considerations proved to be very consequential on how the teams approached the further development of their solutions. In case of weather forecast and composite maps teams, they were most concerned with the explainability of their products to end-users. The issue that the former had was that their stacked deep learning architecture disabled themselves to explain the outputs of their model. This raised concerns since they saw some kind of bias in the outputs which made their predictions imprecise, yet without understanding why. For the composite maps on the other hand, the issue of explainability was tied closely with the inability to collect all desired data, which led them to believe they cannot produce a tool good enough for farmers.

The other three teams were concerned with the risks attached to the use of their tools. To illustrate this challenge, the team calculating the agriclimatic factors realized that how much of a gap between data and phenomena is acceptable depends on each individual case. As one participant explained:

> *'this is kind of dangerous for talking about some freezing periods. Because if you like have the temperature plus one or minus one, it's a big difference for the crop in the area. So, for the crop related issues it's not a good data source. But if you calculate the accumulated, for example, soil solar radiation or the accumulated temperature, you say, like okay, for this hour, it was like 10 degrees Celsius. And even if it was 11 or 9, it doesn't matter that much.' (Geomatician 3, Agriclimatic factors)*

As the GIS expert explained, inaccuracy of measurements is something that is normal and expected, but the degree of inaccuracy constrains the number of phenomena that can be calculated from it without potentially bringing harm to farmer's crops. A one-degree Celsius difference can amount to a difference between a field that is covered in frost and a field that is not, which is a difference between a healthy plant and a frozen plant. So, at least when it comes to sharing predictions with farmers, the team was very cautious about which phenomena could be reliably calculated with the available data. Similar considerations were deliberated by the locust team as they thought of effects their early warning system can have on infested areas or the land boundary team when it comes to the distribution of subsidies or charging of taxes based on the boundaries their tool identified.

Work in progress - Do not cite

Scrutinizing ML Workability of datasets Developing ML solutions requires data not only to represent a phenomenon, but that it is also constructed in a way that it can be used as training examples for ML. This requires the data to be of a certain volume and variety, but also quality and format. Yet, these properties can be hard to achieve as work invested in one, can be detrimental for the other. For example, as we have shown, for three teams inclusion of the large amount of weather variables can be understood as both positive, as well as detrimental for ML development. Hence, workability is something that can be competingly interpreted by the developers.

When evaluating the effects of transformations done on images, the land boundary team was unpleasantly surprised because '*when you clean up those data, it causes a loss of big portions of area' (Machine learning expert, Land boundary)*. Due to numerous transformations to deal with the differences in resolution and format, many resulting images turned out to be simply black, missing land boundary data, or the team simply wasn't able to recognize what the images showed. Also, there was a discrepancy between the way the team defined land boundaries and what the 'good images' showed, namely the algorithm segmented all of the pixels in two classes, those that fell within the boundaries of a field and those that did not seen on Figure 2. So, there weren't any pixels that actually referred to a boundary between a field and a non-field at all.

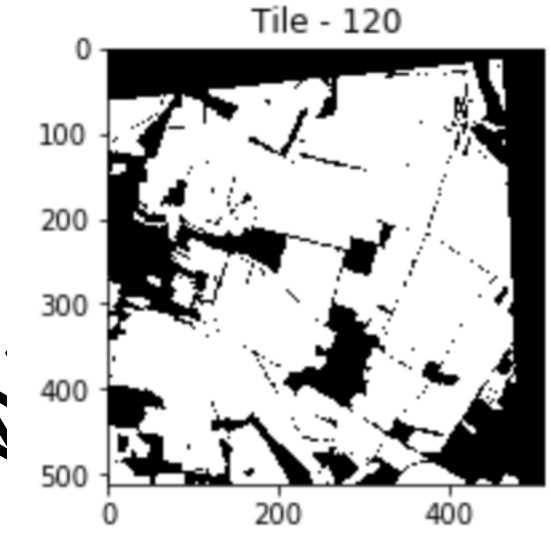


Figure 2 Example of a land boundary model output

The team decided to check how all images looked like and were still faced with some discrepancies due to the interpretative nature of land boundaries. As Figure 3 illustrates, boundaries between crops are far from fixed due to plants being living beings embedded in the natural world, making the observed boundary fundamentally different from the one that is

In Press - Do not cite

registered in the LPIS. Moreover, as unused land can exhibit physical boundaries due to over growing or a single crop field can have some visible boundary due to a disease, but these aspects are not reported in the LPIS, there is an inherent imperfectness of the representation with respect to capturing the actual boundaries.

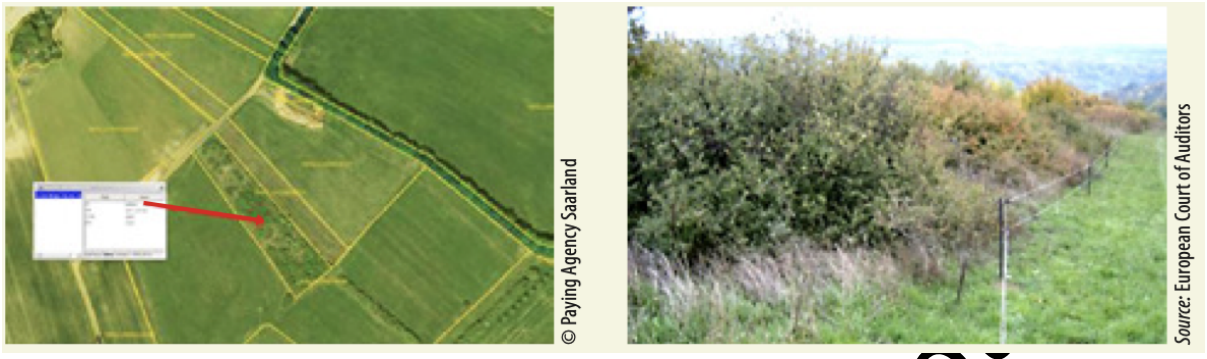


Figure 3 Differences between representations and phenomena

This example shows how considerations of workability, such as considerations of what an algorithm will learn, is important to understand for developers, so that they can address these issues before they move on to the next development stage.

Work in progress - Do no

5. Discussion

5.1 Model of creative data work for machine learning

Our findings illustrate how actors facilitate data reuse in the context of ML by engaging in creative work and actively co-constructing data and phenomena. Such creative work involves learning about how the data "came to be" (Jones, 2019) and what the data represent, as well as learning more about the phenomenon of interest for which actors aim to reuse these data. By continuously inspecting the existing and emerging discrepancies between what data represents and how the phenomenon is defined, data workers search for ways to reinterpret the meaning of data and phenomena, so they can repurpose the data in new, innovative ways and use them as input for ML models. We identified three practices that emerged as teams coped with challenges of repurposing data coming from different contexts: problematization, creative data work, and scrutinizing datasets. Furthermore, these practices seem to form a closed loop, as the last practice of scrutinizing datasets reveals challenges that require redefinition of a phenomenon initiating the practice of problematization again. We present the process model we developed in Figure 1.

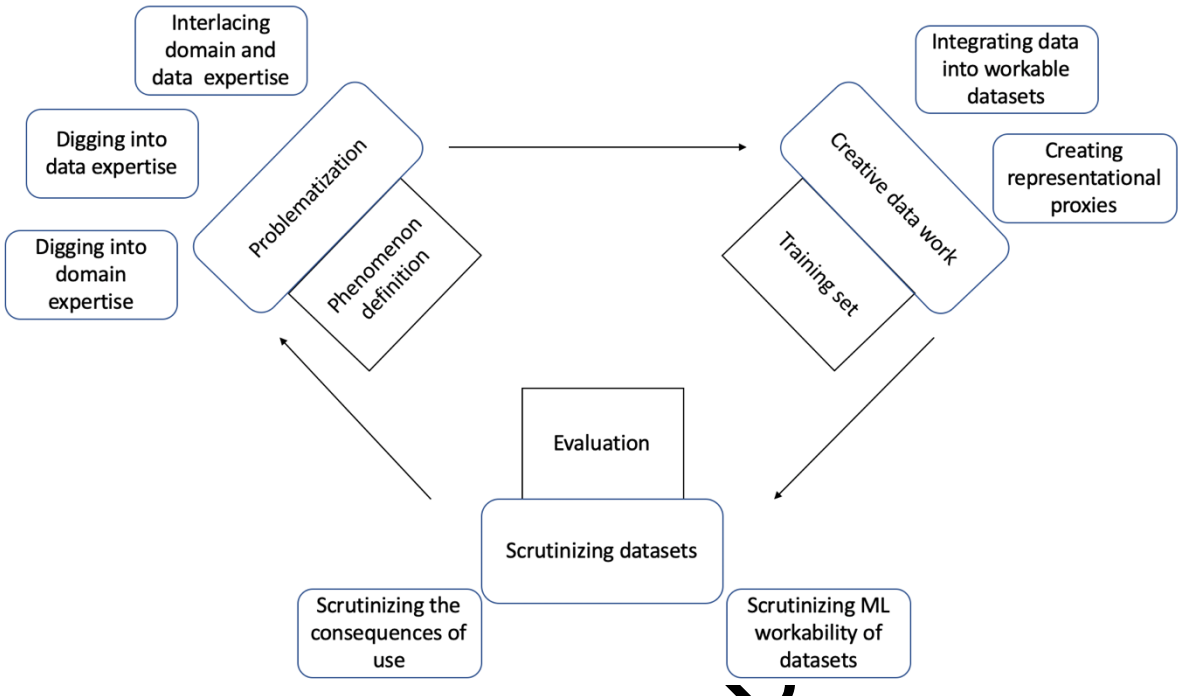


Figure 4 Model of creative data work

Work in progress

Initiating data reuse projects can be more challenging, compared to producing data for machine learning, since the data from which the value needs to be realized is already produced with all of the limitations which that process placed on data. As a result, a target phenomenon has to be defined in terms of that data while accounting for all of its "cooked" characteristics. Since reuse involves data that was not produced for the purpose it is being used for, identifying similarities and differences between what data represents and what aspects of a phenomenon developers want to represent is very important. To successfully identify those, the developers need to both have in mind the internal discrepancies of available data, as well as how each of them, or combination thereof, can be used to capture a phenomenon through a proxy representation. Hence, the practice of problematization clearly shows the importance of tight knowledge interlacing (Tuertscher et al 2014) between domain and data expertise for successful data reuse. This interlacing is beneficial because if developers are not involved in data production and they might not be informed about the opportunities, but also threats of reusing some data

Digging into domain knowledge and relating the findings to technical knowledge of data, enabled teams we studied to infer how to define a phenomenon in terms of available data. Learning about data production can be hindered by the lack of recorded information on how

data was produced or even subsequently altered. Yet, knowledge of this can prove to be important for development of best strategies for value generation later on or even reconsideration of which data can best serve the aims defined in the previous practice. This is so because particularities of data inform developers of possibilities to use certain data as a representational proxy. Lack of knowledge about data can lead to unexpected consequences of its use, or to nonuse of potentially valuable data.

Production of a phenomenon definition enables developers to engage in data work needed to produce a training set that matches the definition. Data work in ML development is often regarded as merely technical, described in a straightforward and stepwise way, and in general taken to be dull, janitorial work of cleaning and preparing data (Lehr & Ohm 2011). Yet, data work we observed seemed to go beyond mere technical considerations, as it required innovative thinking, and was filled with judgement calls based on domain expertise and/or data science experience. Having a phenomenon definition, developers engaged in creating a training set that can workably represent the defined phenomenon. Data work that developers perform in this practice is creative for several reasons. First, as we show in our cases, challenges that developers face are emerging and unanticipated, which leads to developers finding novel ways to represent phenomena that are very complex. Furthermore, developers cannot simply follow some predefined steps for constructing workable representations, because tradeoffs and integrations they have to make depend on their judgement calls and understanding of the similarities and differences between what data represents and a target phenomenon. Developers need to work with what they have and depending on the computational resources, domain and data expertise, data accessibility, and many other factors they try to find the best way to ensure workability and utility of their main outcome - a training sets.

As they cope with issues pertaining to the lack of standardization and representational tradeoffs that they have to make, development teams also think and try to anticipate how consequential their responses to emerging challenges will be on the final product of their development. So, engaging in a practice of scrutinizing datasets enabled developers to realize that the data they created from repurposed data sources still bears marks of the initial way data was created. Hence, the teams became aware of the enduring consequences of data reuse that their data work was not able to fully eliminate and use data in new contexts without considering its journey. Furthermore, this practice emerged as a key moment in which developers, faced with the unexpected consequence of their work, addressed potential issues of future use of their tools and impact they can have on end users. This anticipation fed back into their work and

Work in progress - Do not cite

enabled them to put data they constructed in a new context. For developers to be able to do this though, they need to have a good understanding of the practices in which their tools can be potentially embedded.

As teams were learning about the specific issues that might emerge, we identified that the practice of problematization reemerged, as a phenomenon is again being redefined based on the newly constructed training set. So, we offer insights that suggest that these practices are cyclical in nature. This suggests that knowing the reasons and actions through which data was refined is crucial for successful data work throughout the process. Since the decoupling of data and phenomena is always present to a certain degree and there is "no such thing as a perfect tuning of machines dictated by material agency as a thing-in-itself" (Pickering 2010), we have a reason to believe that these practices are cyclical and emerge in data work more generally.

Work in progress - Do not cite

5.2 Theoretical implications

Practices of creating proxy representations for ML Our study informs the scholarship on digital representations by explaining practices of dealing with their imperfections. While representations are known to be socially, politically and materially constructed, less is known about the work that goes into coping with, and overcoming, their limitations. We find that the rich literature on data work enables us to uncover the critical role of interlacing domain and data expertise in ML development. Tight interlacing of domain and data expertise plays a role in coping with imperfect representations by guiding data related choices – such as making representational tradeoffs and integration in the face of data diversity. So, beyond showing that domain expertise matters for making ML meaningful or understandable, our case emphasizes how it has an active role in continuous coping with imperfectness of representation. Moreover, evaluating relevant pieces of domain knowledge is performed with references to technical specifications of how to capture this knowledge by available data. Hence, it is through interlacing of expertise that developers make their data choices and discover imperfections of the data in the first place. This interlacing can be instantiated through collaboration of experts with different backgrounds and through consultation with external researchers and academic articles, but it can also involve hybrid expertise instantiated in one person or distributed across teams. This insight calls for more research in different arrangements of epistemic dependencies developers can find themselves in as they work on reusing data.

Data reuse for ML requires creative data work Our findings also show the relevance of the research on data work for the literature on data sharing and reuse (Gehlaar & Otto 2020; Lis &

Otto 2020; Leonelli 2014). Production of data sets involves a messy process filled with judgement calls aimed at coping with the imperfect nature of data. Facilitating data reuse does not solely rely on standardization of datasets or cleaning noise, but data workers also have to actively engage in reinterpreting and realigning data and a phenomenon. By leveraging the focus on the broader data work practices, we uncovered how exactly developers create training sets and how they cope with "cooked" data. Our findings are in accordance with the general belief that preparatory work takes up most of the time and effort in ML development. Interestingly, this work is often regarded as menial and boring in the discourse on ML (Lehr & Ohm 2011). Yet, our insights suggest that this is precisely the work that requires most creativity and innovativeness, and instead of it being merely 'janitorial', this data work seems to be most consequential on the successfulness of data reuse projects and as such should be regarded as data work that requires most attention from scholars and organizations alike. So, our findings have an implication for future research on data reuse for ML by highlighting the practices that often remain hidden in the discourse on ML, but are very consequential on the success of ML projects.

(Un)boundedness of data We also bring valuable insights to the debate on the properties of data. We complement existing conceptualizations of data as potentially unbounded (Ekbia 2009; Alaimo et al 2020), yet situated resources (Jones 2019; Strong et al. 1997) by illustrating that unboundedness is not something inherent to the data, but an outcome of practices that involve a creative and messy process aimed at coping with different representations of the world. In doing so, we agree with the critical voices arguing for need to give attention to the historicity of data, but also show how data can come to be reused despite their "cooked" nature (Gitelman & Jackson 2013; Jones 2019). Through iterative development of representational proxies, data workers can bridge the deficiencies of data and produce workable ML datasets.

By taking a practice perspective, we haven't looked at data work as involving arrangements of disparate entities (e.g. people, technology, data) (Feldman & Orlikowski 2011), but at concrete practices that emerge in data work, thus uncovering how (un)boundedness comes about in practice. Our findings show that the same data can pose different issues, as well as opportunities, for reuse depending on the work that is invested in it. So, instead of assuming data unboundedness or uncovering the biases and constraints that limit unboundedness of data, our study shows how data (un)boundedness is being established in the first place thus enabling or limiting data reuse.

Work in progress - Do not cite

5.3 Practical implications

Organizing for data reuse Our findings have important implications for organizations that aim at leveraging data reuse both intra- and interorganizationally (Gelhaar & Otto 2020; Lis & Otto 2020). Intraorganizationaly, organizations constructing data lakes which enable data use and reuse need to be aware that, besides ensuring data accessibility and lack of noise in form of data standardisation, data reuse requires creative ways of engaging with data, understanding how it can be interpreted and how it relates to the domain which it stands for. This suggests a need for a greater interdepartmental collaboration on ML development where data and domain experts can collaborate on identifying and overcoming challenges related to data reuse. This insight builds on top of the need for appropriate data curation and thorough documentation that brings clarity to the differences and similarities available data has to phenomena that can potentially be represented with it (Leonelli & Tempini 2020). Interorganizationally, our findings add to the already known challenges of establishing data ecosystems (Lis & Otto 2020). Besides the collaborative and competing challenges, organizations can face challenges related to the need for domain expertise to understand data. This challenge is particularly salient when data is shared across contexts and the expertise instantiated in organizations can differ greatly. Similarly to intraorganizational insights we provided, besides sharing data, organizations also need to share their expertise to be able to realize the maximal value of reused data.

Machine learning development Our findings have implications for how we understand machine learning development, as well as how to train data and domain experts, and organize collaboration between them. Instead of perceiving and treating preparatory work as undesirable, automatable, boring, and merely technical, our findings suggest that data work involved in construction of training sets is the most creative and important part of ML development, that it involves deep expertise in data and the domain, and which is crucial to invest in to realize value from data reuse. Hence, both in education and during employment, preparatory work that data workers do needs to be given its due credit and organizations. Also, both domain and data experts, and those that are developing to become one, need to be aware of the need for interlacing the two expertise in practice. So, they should be trained to have hybrid expertise to tackle data reuse challenges in a particular domain or be trained on how to collaborate with other experts on such projects.

5.4 Limitations

Work in progress - Do not cite

Limitations of the research design can be built upon as several research questions arise from the boundary conditions. The setting of a hackathon enabled us to observe the full process of development, from finding data sources to evaluating models, but the pressure-cooker setting has several boundary conditions. First, as the development process finished after two months of hacking, we were not able to observe if the development process is indeed cyclical and in what manner. Future research can address this issue by performing a longitudinal study. Also, this limitation raises an interesting question for future research that can study how business and research organizations cope with the perpetual issues of imperfectness of representations. Second, participants in a hackathon often did not know each other, their expertise were not necessarily compatible, there were not any organizational pressures, and, due to the global pandemic, the participants collaborated exclusively virtually. So, it remains an open question if the dynamics that we observed will also emerge in organizations and in what way. Nevertheless, by bringing data to the fore of our paper, we aimed at showing how practices emerged due to the particularities of data being reused. Hence, we expect that the general practices replicate also in an organizational setting. Third, we have focused our attention on the case of ML development for agriculture, so our findings might not generalize to all cases of data sharing and reuse or even ML development in some other context. So, it remains for future research to investigate data reuse in other contexts and using other technologies.

Work in progress - Do not cite

6. References

Alaimo, C., & Kallinikos, J. 2016. "Encoding the everyday: The infrastructural apparatus of social data," In Big data is not a monolith: Policies, practices, and problems, C. Sugimoto, H. Ekbia, & M. Mattioli (Eds.), Cambridge, MA: MIT Press, pp. 77–90.

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.

Bailey, D. E., Leonardi, P. M., & Barley, S. R. 2012. The lure of the virtual. Organization Science, 23(5), 1485-1504

Berends, H., & Deken, F. 2021. Composing qualitative process research. Strategic Organization, 19(1), 134-146.

Berg M and Bowker G. 1997. The multiple bodies of the medical record. Sociol Quart; 38(3): 513–537.

Birnholtz JP and Bietz MJ. 2003. Data at work: supporting sharing in science and engineering. In: Proceedings of the 2003 international ACM SIGGROUP conference on supporting group work, Sanibel Island, FL, 9–12 November 2003, pp. 339–348. New York: ACM Press.

Bjørnstad, C., & Ellingsen, G. 2019. Data work: A condition for integrations in health care. Health informatics journal, 25(3), 526-535'

Bowker, G., & Star, S. L. 1999. Sorting things out: Classification and its consequences. Cambridge, MA: MIT Press.

Cabitza, F, Locoro, A, Alderighi, C, Rasoini, R, Campagnone, D, Berjano, P. 2019. The elephant in the record: On the multiplicity of data recording work. Health Inform J; 25: 475–490.

Constantiou, I. D., & Kallinikos, J. 2015. New games, new rules: big data and the changing context of strategy. Journal of Information Technology, 30(1), 44-57.

Davenport, T. H. 2006. Competing on analytics. Harvard business review, 84(1), 98

Davenport, T.H., Prusak, L., 1997. Information Ecology. Oxford University Press, Oxford.

Dixon-Woods M, Leslie M, Bion J, et al. 2012. What counts? An ethnographic study of infection data reported to a patient safety program. Milbank Q; 90(3): 548–591.

Ekbia, H. R. 2009. Digital artifacts as quasi-objects: Qualification, mediation, and materiality. Journal of the American Society for Information Science and Technology, 60(12), 2554–2566.

Ellingsen, G, Monteiro, E, Roed, K. Integration as interdependent workaround. 2013 Int J Med Inform; 82(5): e161–e169.

Feldman, M. S., and W. J. Orlikowski. 2011. “Theorizing Practice and Practicing Theory,” Organization Science (22) pp. 1240-1253.

Fischer, J. E., Crabtree, A., Colley, J. A., Rodden, T., & Costanza, E. 2017. Data work: how energy advisors and clients make IoT data accountable. Computer Supported Cooperative Work (CSCW), 26(4-6), 597-626.

Foster, J., McLeod, J., Nolin, J., & Greifeneder, E. 2018. Data work in context: Value, risks, and governance. Journal of the Association for Information Science and Technology, 69(12), 1414-1427.

Gelhaar, J., & Otto, B. 2020. “Challenges in the Emergence of Data Ecosystems,” In PACIS 2020 Proceedings. 175. https://aisel.aisnet.org/pacis2020/175

Gerlitz, C., & Helmond, A. 2013. The like economy: Social buttons and the data-intensive web. New Media & Society, 15(8), 1348–1365.

Gitelman, L. and Jackson, V. 2013. ’Introduction’, in L. Gitelman (ed.), ‘Raw Data’ is an Oxymoron. MIT Press, Cambridge, MA, 1–14

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. 2017. Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems, 26(3), 191-209.

Günther, W., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, J. F. M. 2017. Rushing for gold: tensions in creating and appropriating value from big data. In International Conference on Information Systems 2017.

Holten Møller N and Bjørn P. 2011. Layers in sorting practices: sorting out patients with potential cancer. Comp Support Comp W; 20(3): 123–153.

Jones, M. 2019 What We Talk About When We Talk About (Big) Data, The Journal of Strategic Information Systems (28:1), 3-16.

Kallinikos, J., Aaltonen, A., & Marton, A. 2013. The ambivalent ontology of digital artifacts. MIS Quarterly, 37(2), 357–370.

Kitchin, R. 2014. "Big Data, New Epistemologies and Paradigm Shifts," *Big Data & Society* (1:1), pp. 1-12.

Kitchin, R., & McArdle, G. 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130.

Knorr-Cetina, K. 1999. Epistemic cultures: How the sciences make knowledge. Cambridge, MA: Harvard University Press.

Langley, A. 1999. Strategies for theorizing from process data. Academy of Management review, 24(4), 691-710.

Lehr, D., & Ohm, P. 2017. Playing with the data: What legal scholars should learn about machine learning. U.C. Davis Law Review, 51(2), 653-718.

Leonardi, P. M. 2012. *Car crashes without cars: Lessons about simulation technology and organizational change from automotive design*. MIT Press.

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

Leonelli, S. 2013. Integrating data to acquire new knowledge: Three modes of integration in plant science. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 44(4), 503-514.

Leonelli, S., & Tempini, N. 2020. Data journeys in the sciences (p. 412). Springer Nature.

Lycett, M. 2013. 'Datafication': Making sense of (big) data in a complex world. European Journal of Information Systems, 22:4, 381-386, DOI: 10.1057/ejis.2013.10

Work in progress

Lis, D., & Otto, B. 2020. "Data Governance in Data Ecosystems–Insights from Organizations," AMCIS 2020 Proceedings. 12. https://aisel.aisnet.org/amcis2020/strategic_uses_it/strategic_uses_it/12

Monteiro, E., & Parmiggiani, E. 2019 Synthetic knowing: The politics of the internet of things. MIS Quarterly, 43(1), 167-184.

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., ... & Erickson, I. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. Big data, 5(2), 85-97.

Newell, S. 2015. "Managing Knowledge and Managing Knowledge Work: What We Know and What the Future Holds," *Journal of Information Technology* (30:1), pp. 1-17.

Østerlie, T., & Monteiro, E. 2020 Digital sand: The becoming of digital representations. Information and Organization, 30(1), 100275.

Parmiggiani, E., & Monteiro, E. 2015. The nested materiality of environmental monitoring. Scandinavian Journal of Information Systems, 27(1).

Pickering, A. 2010. The mangle of practice: Time, agency, and science. University of Chicago Press.

Pine, K. H. 2019. The qualculative dimension of healthcare data interoperability. Health informatics journal, 25(3), 536-548.

Rolland B and Lee CP. 2013. Beyond trust and reliability: reusing data in collaborative cancer epidemiol- ogy research. In: Proceedings of the 2013 conference on Computer supported

Do not cite

cooperative work, San Antonio, TX, 23–27 February 2013, pp. 435–444. New York: ACM Press.

Shapiro, C., & Varian, H. R. 1999. Information rules: A strategic guide to the network economy. Boston, MA: Harvard Business School Press

Strong, D. M., Lee, Y. W., & Wang, R. Y. 1997. Data quality in context. Communications of the ACM, 40(5), 103-110.

Susha, I., Janssen, M., & Verhulst, S. 2017. Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences. Retrieved https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1347&context=hicss-50

Tuertscher, P., Garud, R., & Kumaraswamy, A. 2014. Justification and interlaced knowledge at ATLAS, CERN. Organization Science, 25(6), 1579-1608.

Van den Broek, T., Van Veenstra, A.F., 2015. Modes of governance in inter-organizational data collaborations. In: Proceedings of the Twenty-Third European Conference on Information Systems, Münster, Germany, May 26–29.

Verhulst, S., 2021. *Unlock the Hidden Value of Your Data*. [online] Harvard Business Review. Available at: <https://hbr.org/2020/05/unlock-the-hidden-value-of-your-data> [Accessed 8 June 2021].

Yin, R.K., 1984. Case Study Research: Design and Methods. Beverly Hills, Calif: Sage Publications.

Yoo, Y. 2010. Computing in everyday life: A call for research on experiential computing. MIS Quarterly, 34(2), 213–231.

Yoo, Y., Boland Jr., R. J., Lyytinen, K., & Majchrzak, A. 2012. Organizing for innovation in the digi- tized world. Organization Science, 23(5), 1398–1408.

Work in progress - Do not cite

Zhang, Z., Nandhakumar, J., Hummel, J., & Waardenburg, L. 2020. Addressing the key challenges of developing machine learning AI systems for knowledge-intensive work. *MIS Quarterly Executive*, *19*(4).

Zuboff, S. 1988. In the age of the smart machine: The future of work and power. New York, NY: Basic Books.

Work in progress - Do not cite