Forgetting Practices in the Data Sciences

ANONYMOUS AUTHOR(S)

HCI engages with data science through many topics and themes. Researchers have addressed biased dataset problems, arguing that bad data can cause innocent software to produce bad outcomes. But what if our software is not so innocent? What if the human decisions that shape our data-processing software, inadvertently contribute their own sources of bias? And what if our data-work technology causes us to forget those decisions and operations? Based in feminisms and critical computing, we analyze forgetting practices in data work practices. We describe diverse beneficial and harmful motivations for forgetting. We contribute: (1) a taxonomy of data silences in data work, which we use to analyze how data workers forget, erase, and unknow aspects of data; (2) a detailed analysis of forgetting practices in machine learning; and (3) an analytic vocabulary for future work in remembering, forgetting, and erasing in HCI and the data sciences.

CCS Concepts: • **Human-centered computing** → *Computer supported cooperative work*; **HCI theory, concepts and models**; • **Computing methodologies** → *Cooperation and coordination*; **Supervised learning**.

Additional Key Words and Phrases: forgetting, forgettance, data silence, articulation work, invisible work

ACM Reference Format:
Anonymous Author(s). 2018. Forgetting Practices in the Data Sciences. In *CHI 2022, 30 April - 6 May, 2022, New Orleans, LA, US*. ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

...and our songs about the stories we've forgotten;
and all that we've forgotten we've forgotten...
–Pádraig Ó Tuama [161]

Researchers' work in the data sciences has inspired important and diverse themes in HCI and related research areas, such as crowdmarket labor [100, 101], fairness [13], bias reduction [27, 109], surveillance capitalism [235], human centered data science [121, 232], human centered machine learning [30], labeling [67, 157], explainable AI/XAI [59, 231], co-creativity [44, 137], and the specialized work of AI teams [172, 227, 234]. As we store more and more data about one another, ourselves, and things, we assume that our databases can "remember" what we need to know. As Bowker wrote in *Memory Practices in the Sciences*, human work in many scientific fields requires attention to what we remember, how we remember, how we store or otherwise preserve what we remember, how we re-find what we know (or what we once knew), and whom we remember with [20].

We also forget. Forgetting may initially seem like a "bad" thing in the sciences. And yet, scholars have argued that forgetting can be beneficial to memory [131, 217], and that remembering and forgetting may be seen as facets of a single, unitary phenomenon [148, 152, 224]. Spiel, for example, advocates the gradual removal of less relevant information, in a way that mimics gradual memory degradation in humans [207].

Consistent with this view, in a related field, de Souza and colleagues showed that much of software engineering involves a reversible kind of forgetting through the use of application programming interfaces (APIs) and other strategies of *separation of concerns* [51, 166], which allows developers to focus on their immediate task while encapsulating non-focused complexities outside of their current scope of attention and action [41, 42]. They described several types of tensions in this work, leading to a redefinition of APIs in a more social and infrastructural context (e.g., [43]). Through a series of thoughtful examinations of software practices, they asked to what extent and in what ways *separation* could be beneficial or harmful [42, 193]. We see encapsulation as a type of *reversible* forgetting - i.e., if complexity is forgotten through encapsulation in a particular function call, a computer scientist or engineer can usually access the source code of the function - thus effectively remembering the complexity upon need. In this way, separation of concerns may be seen as a combination of strategized forgetting and strategized remembering.

Data science work seems to involve similar strategies "where data becomes a first-class citizen, on a par with code" [213]. There are similar de facto practices of *forgetting* complexities in favor of pattern-finding in data, and hiding complexities through the addition of layers of sophistication and abstraction during data-cleaning and feature-engineering [97, 150]. Each layer involves its own complexities and challenges, encouraging a data science team to focus on a single problem at-a-time [114, 156, 189]. On this basis, we claim that data science uses both software engineering tools[1] and also software engineering heuristics of work practices through hiding complexity (e.g., [127]). For example, one data science worker may replace certain missing values through a form of missing-values imputation. A second data science worker will then receive that dataset, and will not know which values were initially missing.

We argue that - unlike the software engineering practices of encapsulation and separation of concerns (discussed above) - much of the forgetting practices in data science are, in practical terms, *non-reversible*. Our concern in this paper is to examine how we forget in the data sciences, what we may lose thereby, and how these forms of forgettance [217] (i.e., the inverse of remembrance) may be implicated in the broader politics of data science and "big data." We question the meta-narratives (per Lyotard's influential analysis [141]) that AI technologies are objective and/or infallible (e.g., as critiqued by [23, 36, 74]).

To summarize so far, we propose that forgetting practices can be both beneficial and harmful. The beneficial aspects allow us to focus on particular problems and to build useful higher-level concepts (abstractions). The harmful aspects occur when we forget that we have engaged in those forgetting practices, thereby losing metadata that we may need to understand the surprising, biased, unfair, or injurious outcomes of our work. We will take up additional beneficial aspects of certain socially-motivated strategies of forgetting, when we discuss data silences in Section 2.3. In that section, we will also examine additional harmful aspects of other socially-motivated strategies of forgetting.

In this paper, we consider both extrinsic and intrinsic issues in the work of data science. From an extrinsic perspective, we acknowledge the important discussions of bias in the large-scale selection of entire datasets in data science (e.g., [13, 27, 144, 163, 170, 198]). From an intrinsic perspective, we extend that analysis to show how forgetting occurs within the detailed work practices of data work [151, 175, 197] - i.e., planning, choosing, cleaning, curating, (feature) engineering, and labeling records at the level of the data records themselves. We describe forgetting and forgettance as important human actions that inevitably put a human interpretation into the data in the dataset [156, 192, 198].

We have structured this essay as follows: In Section 2, we begin with a broader consideration of forgetting as social and scientific practices and then briefly review well-known discussions of bias in datasets in Section 3. Section 4 presents our detailed critique of work practices in data work, and the ways in which humans add their knowledges

[1] E.g., libraries, packages, and even Knuth's literate programming [120] in the form of Jupyter notebooks

and interpretations into the detailed data within data records. Following this, we integrate the data work practices of Section 4 with the forgetting practices of Section 2, and we propose changes to those practices that may provide a better balance between strategic forgetting and strategic re-remembering.

With this discussion of strategic forgetting and re-remembering, this paper makes the following contributions: we present a classification of (1) data work practices related to forgetting, omitting, obliviating, and silencing, organized into three higher-level categories of silences; (2) an analysis of forgetting during the detailed steps of data work; and (3) implications of those silences and forgettings in the broader politics of data and algorithms.

1.1 Positionality Statement

The two authors of this paper are actively involved in critical computing. One of us has studied both formal and informal arrangements in civic life and civil society, including online resources that carefully negotiate visibility and invisibility for people who are made vulnerable. One of us has studied data science workers through qualitative and survey methods, including initial investigations into the detailed data work through which data science workers construct the data in their datasets.

2 TYPOLOGIES OF FORGETTING

Remembering - by individuals, groups, and via technological or social mediation - has been a major theme in HCI and in data science [2–4, 6, 20, 35, 82, 154]. In this paper, we attempt an inversion [21, 203], by focusing on the unattended aspects of forgetting as part of memory work. We build on the forgetting aspects of the work of Bowker [20], Engestrom [61], Easterby [58], Connerton [35], Minarova-Banjac [149], and Vinitzky-Seroussi [225], and feminist technoscience work by Harding [85–87], Bardzell [10], Costanza-Chock [36], D'Ignazio and Klein [49], Mulvin [158], Strohmayer et al. [211, 212], and Bellini et al. [14], along with selected political perspectives which turn out to be applicable [28, 129, 165, 184]. We will begin with praise for forgetting, followed by accounts of harms of forgetting. We then focus on an integrated analysis of types of forgetting, which will help to guide the rest of the paper.

2.1 Forgetting Considered Beneficial

On one hand, forgetting can be understood as beneficial. Initially, it seems that forgetting is opposed to remembering. However, recent thinking in the humanities and the social sciences argues for a more complementary and even syncretic view. Lamers et al. suggest that forgetting serves to highlight what we need or want to remember [131]. Mills writes of this phenomenon as ”Forgetting is an important part of memory work” ([148]; see also [224]). Momigliano anticipated this complexity, writing that ”to learn something new or to be reminded of something we had forgotten... is almost the same” [152]. Bowker observed that archives - our large institutional memory repositories - function ”by remembering all and only a certain set of facts/discoveries/observations, consistently and [thereby] actively engage... in the forgetting of other sets” ([20]; see also [58]). Writing in the Conference on Artificial General Intelligence, Thórisson et al. described this memory strategy as *forgettance*, which they defined as ”Removing the least relevant and necessary knowledge, if needed” ([217]; see also [71]).

As we discussed in the previous section, forgetting is also an implicit strategy in data science. If we think of data science as a kind of ”stack” of refinements on data - i.e., from data-acquisition to data cleaning etc. to modeling - then data science workers tend to focus their efforts on the current layer of refinement, and to forget the complexities and uncertainties of the prior layers. As is common in many human activities, we forget the past in order to concentrate on the present. In Section 4, we will consider the potential costs of this implicit strategy.

2.2 Forgetting Considered Harmful

On the other hand however, forgetting in data science can also be harmful or cause violence, not least because our choice of what we deem unimportant enough to forget to improve our memory, impacts on our understanding of histories, data, exploitation, harm, and so on. Similarly forgetting is often considered harmful in political arenas. Forché titled her anthology of human rights poetry *Against Forgetting*, based on her experiences with politically-motivated efforts to erase an inconvenient past so as to valorize an authoritarian present [66]. Orwell famously wrote of *memory holes* into which non-conformant or currently dangerous information could be placed for immediate destruction [165]. For Minarova-Banjac, ”Collective forgetting refers to how states and citizens selectively remember, misremember, and disremember[,] to silence and exclude alternative views and perspectives that counter the official discourse” [149]. In ancient Rome, the current ruler might try to obliterate all memory of a former ruler under the rubric of *damnatio memoriae*. [229]. More recently, Panagopoulou-Koutnatzi proposed the word *oubli* to indicate the information that is to be un-remembered [168].

The research literature on HCI and particularly on infrastructuring also argues against forgetting. Bowker’s *Memory Practices* is a thoughtful, sometimes-ironic, encyclopedic treatment of the nuanced values of remembering in the sciences [20]. Large-scale repositories - in effect, databases of datasets - tend to be carefully constructed and classified for re-use by the original creators of datasets and by other researchers in global communities of scholars in multiple disciplines [126]. Ackerman and colleagues explored technological and work-practice activities to preserve knowledge in organizations [2–4, 82]. Two types of organizational memory - of skills and of facts - were said to be necessary foundations for meeting new challenges through organizational improvisations [154]. Others have emphasized transactive memory systems - i.e., knowing whom to ask - as a third necessary resource, either online [162] or in communities of practice as knowledge-holders [105, 136].

And yet, some researchers are also aware of limitations in how ”welcoming” a data repository may be for information. The reduction of gender identity to a simplified female/male binary has been documented as causing significant harms to people whose identities go beyond that binary [36, 195, 206]. Engestrom discusses ways in which non-conformant information may not be recorded in a structured repository that is designed for only certain categories of data [61]. Bowker concurs, critiquing repositories for including *expected* forms of data while excluding *unexpected* forms of data ([20]; see also [22, 88]). Earlier, De Certeau described how data may be distorted when they are transformed to fit preconceptions or available structures of knowledge ([38]; see also [56, 158, 222]):

> “[T]he operation of walking can be traced on city maps... These thick or thin curves only refer, like words, to the absence of what has [been] passed by... They allow us to grasp only a relic set in the nowhen of a surface of projection. Itself visible, it has the effect of making invisible the operation that made it possible. These fixations constitute procedures for forgetting. The trace left behind is substituted for the practice.”

In summary, despite the widespread view that forgetting may be harmful, there is ample evidence that we deliberately and perhaps necessarily lose data in HCI and data science through diverse forms of what Lamers et al. called ”engines of forgetting” in their study of scholarly forgetting [131]. In the next section, we use Onuoha’s conception of *data silences* ([164]; see also [49]) as a structuring principle for a discussion of multiple analyses of diverse types of forgettings.

Table 1. Definitions and Types of Forgetting

Number	Definition	Source
1	**Data Silences.** ”blank spots that exist in spaces that are otherwise data-saturated.”	[164]
2	**Syntactic Silences | Exclusionary Principles.** Small instances of unparseable data that can form patterns of un-inclusion for unnoticed sub-populations.	[20, 53, 61, 88]
3	**Inferential Silences.** Developing an interpretation based on isolated or hand-picked factors	[88]
4	**Substitution of Trace for Actual Experience or Data.** Use of traces or other proxies in place of actual events themselves or persons.	[38, 158]
5	**WYSIATI** (”What You See Is All There Is”) Assumption that *easily available data* are all that are needed.	[88, 106]
6	**Annulment.** Forgetting what is unimportant, or what would interfere with remembering what is important.	[35]
7	**Prescriptive Forgetting.** Alleged consensus that certain things are best forgotten.	[35]
8	**Repressive Erasure.** Use of [political] power to destroy records so as to benefit the powerful	[35]
9	**Humiliated Silence.** Pressure to forget (or not to mention) what is socially-constructed as ”shameful.”	[35]
10	**Colonial Unknowing.** Attempt to render Indigenous knowledges as ”impossible and inconceivable... . normative acts of ignoring, disavowal, and epistemicide” of national identities that pre-date the ”colonial present.”	[60, 79, 223]
11	**Structural Amnesia.** A person [206], institution, or state [35] wants to control how it/they will be remembered - related to impression management [76]	[6, 35, 190]
12	**Redacted Data.** Deliberate obfuscation or removal of data to protect vulnerable persons or groups.	[47, 185, 209]
13	**Covert Silences | Sanitized Erasures | Historical Amnesia.** Removal or alteration of selected data - typically about others - such that the alteration cannot be easily detected	[20, 129, 225]
14	**Selectively Legible Data.** Data are available but serve as boundary objects, interpreted differently by different persons or groups.	[25, 187]

2.3 Data Silences

Data silences are physical or conceptual sites of forgetting - i.e., in the language of Thórisson et al. [217], sites where forgettance is practiced. The concept of data silences may help to bridge between domains of analysis, such as HCI, data science, critical computing, and contemporary social concerns. Onuoha defined data silence as follows:

> ”‘Missing data sets’ are the blank spots that exist in spaces that are otherwise data-saturated. Wherever large amounts of data are collected, there are often empty spaces where no data live... Spots that we’ve left blank reveal our hidden social biases and indifferences.” [164]

Table 1 is an inevitably incomplete synthesis of positions and descriptions of, or related to, data silences. For breadth of coverage, we include descriptions from HCI/CSCW, data science, and more diverse fields of study. We will focus in this paper on the silences that are related to human practices in data science.

We divided the rows in Table 1 into three groups. The silences in the first group of rows (1-6), ”Modest Silences,” are often relatively innocuous actions that are likely to happen, but without negative intentions. The silences in the second group of rows (7-10), ”Silence as Force,” are more deliberate, and may represent intentions to erase or obscure information to the disadvantage of others. The silences in the last group of rows (11-14), ”Ambivalent Silences,” are complex actions that may be done for mixed or uncertain motivations. Context is important to interpret any of these silences, but is particularly important for the silences in rows 11-14.

2.3.1 Modest Silences (rows 1-6). Above, we stated that forgetting can be understood to be beneficial as well as harmful; though of course some of these ”practices” of forgetting may be more complicated. Having said this, these practices contribute to the selective silences that Onuoha wrote about. When Bowker [22] and Engestrom [61], describe data repositories that resist non-conformant data, these are examples of Syntactic Silences and Exclusionary Principles (Table 1 row 2) - e.g.,

> ”One data silence is syntactic gaps, which is a proportionately small amount of data in a very large data set that will not parse (be converted from raw data into meaningful observations with semantics or meaning) in the standard way. A common response is to ignore them under the assumption there are too few to really matter. The problem is that oftentimes these items fail to parse for similar reasons and therefore bear relationships to each other. So, even though it may only be .1% of the overall population, it is a coherent sub-population that could be telling us something if we took the time to fix the syntactic problems.” (Bradley S. Fordham, quoted in [88])

Syntactic Silences have the effect of denying aspects of some people’s experiences, identities, or realities. They are thus aspects of epistemic injustice [69, 128]. Examples include databases that code ”gender” as either female or male, which deny the existence of LGBTQIA2S+ people [36, 206], or restrictions of ascii characters that can be used in ”name” fields, which render some Indigenous names as non-recordable in official records [153].

As Seager observed, ”what gets counted counts” [196]. Systematic patterns of Syntactic Silences may cause certain populations to be undercounted or entirely uncounted. The result is a selective silence. As analysts, we may not be aware that our data processing has caused us to forget a systematic part of our data, and therefore we forget as well a part of our understanding of the people or phenomena that we are studying. As a civic society, we may not properly fund, care for, or otherwise support the people whom we have under-counted or uncounted - i.e., whom we have forgotten. In some cases, it may be necessary to write data from or about certain sources or people, back into the dataset - or to record these data in a separate dataset. As an example, through generations of activism and struggle [55, 103, 222], the Indigenous Nations in North America have begun to make their own tally of murdered and missing Indigenous women and girls (#mmiwg and #mmiwg2s)[2] [204, 221], because most non-Indigenous police departments do not keep such statistics [83, 205]. Syntactic Silences are summarized in row 2 of the Table 1.

We now move to Substitution (Table 1 row 4). Earlier in this section, we discussed de Certeau’s example of how a trace of activity may take the place of original data [38]. When we make this kind of substitution, we create a silence in the data that obscures (forgets) the original data for which we have chosen a substitute or proxy value [158].

The principle of selective-forgetting-to-remember-what-matters [20, 131, 148, 152, 224] is an aspect of Annulment in row 6 of Table 1. We render certain phenomena silent, so as not to be distracted by them: We annul them. In the Introduction, we discussed separation of concerns and encapsulation, which may also be understood as *reversible* forms of Annulment.

[2]The abbreviation ”2s” refers to Two-Spirit people as a generic reference to well-established non-female, non-male gender identities in some North American Indigenous cultures [54]). Two-Spirit people may share some experiences with non-binary people in non-Indigenous cultures, but they may also have a distinct roles and positions within Native cultures [186, 215].

2.3.2 *Silence as Force.* The second set of silences (rows 7-10) are more active, and therefore more likely to have been strategized. Prescriptive Forgetting (Table 1 row 7) is based on a consensus that certain things are best forgotten [35]. But who is included in that consensus? Value Sensitive Design (VSD) suggests that we consider the interests of multiple stakeholders in a design, practice, or policy [70, 90]. Feminist standpoint theories also encourage us to consider the perspectives of multiple other persons, roles, and interested parties - as well as our own perspectives [85, 86, 139] - often starting from the margins [10, 14, 133]. We may thereby ask: *Who* is included in the group, nation, class, or workplace-constituency that forms the consensus in Prescriptive Forgetting? If the claimed consensus is incomplete or illusory, then Prescriptive Forgetting may devolve into one of the more abusive forms of silence in rows 8-11 - either through intention or inadvertence. When a majoritarian position of binary gender is presented as a kind of consensus view, then people with non-binary identities may suffer. These harmful silences can be repaired. For example, several governments recently took steps towards reducing harms, by adding non-binary options for gender identities on passports, thus relieving some trans* people of the burden of being misgendered [17].[3]

Repressive Erasure and Humiliated Silence (rows 8-9 of Table 1) are more related to the political realms that we mentioned in our earlier discussion of the harmful aspects of forgetting - i.e., the imposition of silence on people who wish to be known, seen, heard. We briefly note here that Syntactic Silences may, in the extreme, become an implementation of a kind of Repressive Erasure.

Colonial Unknowing (Table 1 row 10) may provide distinct lessons for data science. In the classic form of Colonial Unknowing, a powerful group attempts to suppress knowledge of certain subordinate persons or peoples, or to hide knowledge of crimes done against those groups [79, 223]. There is a related concept of Colonial Amnesia [60] which may seem less deliberate - i.e., ”lost” knowledge rather than ”suppressed” knowledge.

The strong case of *unknowing* may help us to think about certain politics of data and knowledge. Earlier in this Section, we mentioned the concept of damnatio memoriae, in which information about a prior ruler is suppressed by the current ruler. In Whitling’s account [229], this practice often led to ironic outcomes, causing greater interest in the deposed ruler. Damnatio memoriae thus involves information that is simultaneously remembered and forgotten - but by different interested parties. The non-reversible forgetting practices of data science, which we described in the Introduction, present a similar case: Data science workers at each step are aware of the complexities of data-processing, but data science workers at the next step prefer not to know about these complexities (see also Section 4). When the data reach the model, the claims of modeling excellence are dependent on no longer remembering any potential weak-points in how the dataset was processed.

A contemporary example of Colonial Unknowing is the on-going crisis of the so-called ”residential schools” in former British colonies [146].[4] Tens of thousands or hundreds of thousands of Indigenous children (the Stolen Generation [28]) were legally abducted from their parents and sent to boarding schools, where they were physically punished for speaking their birth languages, and were minimally educated for menial occupations in the colonizers’ economies [77]. At the time of writing, Indigenous-led use of ground-penetrating radar [202] has revealed the unmarked and/or hidden graves of nearly 10,000 of children at the locations of the North American ”residential schools.” Survivor testimony makes it clear that these thousands of children died through malnourishment, physical and sexual abuse, additional forms of torture, and preventable diseases [1, 34, 146, 218].

[3] We note that this approach - while an improvement - continues to treat gender-identity as a single, fixed attribute, and thus does not reflect the realities of people who are gender-fluid and/or intersex.

[4] Where possible, we have cited Indigenous scholars’ works [146], or works that were written by mixed groups of Indigenous and non-Indigenous authors [34, 218]. In the remaining citations, we have consulted non-Indigenous scholars who call for ”unsettling the settler within” [184] or whose collections of papers contain contributions from Native and non-Native scholars in dialog [56, 221, 222].

Many non-Indigenous people in these former colonies are learning about this genocide for the first time in 2021, despite the existence of multiple authoritative books [68, 77, 184], the Truth and Reconciliation reports in Canada [218] and in the US State of Maine [34], and the Abouresk hearings in the US Senate in 1978 [1]. Clearly, the Indigenous Nations know the bitter truth of these institutions [146]. The religious organizations that operated most of these places kept records (currently sealed or sent overseas [102]), and thus are also in a position to know what they have done. In some cases, the religious institutions remembered enough to remove grave markers [160], and in other cases local governments remembered enough to pave over the gravesites [96]. Colonial Unknowing is a way of constructing a selective silence - a selective forgetting - among a public who might condemn the genocide. In this case, it is not that the information has simply ”become unknown.” Similarly to Whitling's description of damnatio memoriae [229], Colonial Unknowing becomes a form of *motivated forgetting*, in which a knowledgeable party tries to perform an act of forgettance - of silence - upon the knowledge of others. Some people might argue that Colonial Unknowing is a form of Prescriptive Forgetting (Table 1 row 7) - i.e., the alleged consensus that some things are best forgetten. However, this claim would require that the prescriptive ”consensus” deliberately excludes the Indigenous Nations, who very much want *their* view of history to be told. We will return to the topic of motivated forgetting in Section 4.

2.3.3 Ambivalent Silences. The last four rows of Table 1 are more multi-valent. Structural Amnesia (Table 1 row 11) involves an attempt to control one's impression to others - what Goffman called the ”frontstage” or public view of self, which could be managed through ”backstage” work [76]. Certain aspects of the discussion (above) about Colonial Unknowing may be relevant here (e.g., distortions in historical records), but so are the practices of asserting a new identity following e.g. a gender-identity transition (e.g., [36, 206]). In the latter case, the to-be-forgotten information (the oubli, in the language of Panagopoulou-Koutnatzi [168]) may remain known to others (e.g., as a deadname), but it is clearly not the preferred self-presentation. Institutions (publishers, universities) may sometimes resist this kind of individually-based Structural Amnesia, if those institutions fail or refuse to propagate new identities from one record-keeping system to other such systems. They pit one *individual* form of Structural Amnesia against a second *institutional* form of Structural Amnesia. Like other forms of motivated forgetting, it is important to consider Structural Amnesia in personal, institutional, and political contexts.

Sometimes data silences can also be seen as mechanisms of safety. Unlike Structural Amnesia, Redacted Data (Table 1 line 12) is a deliberate effort to obscure one's own data - usually for reasons of safety. A benign example is the *right to be forgotten* under the European Union's General Data Protection Regulation [32]. In other cases, this is not an easy or clear-cut task and something that can mean losing parts of oneself to be safer. A particularly current and prescient example of this is currently taking place in Afghanistan. At the time of writing this article, the Taliban have taken over leadership of the country after the US's and NATO's removal of troops. This take-over resulted in a scramble to try to bring out of the country many Afghan citizens and others who had worked for the West, because their prior work would make them a target for the Taliban. Stokel-Walker spoke to a former translator, known as Muhibullar, who burned the documents that showed that he had worked for the US [209]. As Stokel-Walker writes, he did this ”knowing that such paperwork is vital to gain a visa and a potential route out of Afghanistan. But it remains a horrific quandary: Taliban militia are already reportedly going door-to-door to find those who have worked with foreign governments and non-governmental organisations.”

In another article [185] an unnamed Afghan woman writes about how she has hidden or burned all of her school certificates - achievements she has been proud of and worked towards for her whole life. She writes ”Why should we hide the things that we should be proud of? In Afghanistan now we are not allowed to be known as the people we are." Later in the article, she writes: "Having any ID card or awards from the American University is risky now."

Both of these examples show how data, paperwork, and other pieces of information about us can cause us harm - and how we can silence these data for our safety. However, this safety is complex - in Muhibullar's case the documents he and others like him have burned are also perhaps their only way of proving that they worked with the West, meaning it may be their only way of leaving the country. In the case of the women who have had to hide their educational certificates, they must do this as being affiliated with an American university can be dangerous for them. In doing this though, they must hide important parts of their selves, identities, jobs, experiences - they are a little more safe than before, but they are no longer whole.

To explain the concept of Selectively Legible Data (Table 1 row 14), we begin with a song:

> *When the sun comes back and the first quail calls,*
> *Follow the Drinking Gourd.*
> *For the old man is a-waiting for to carry you to freedom*
> *Follow the Drinking Gourd*
> –Traditional Freedom Spiritual, US, "Follow the Drinking Gourd"

A particular form of selective legibility takes advantage of specialized knowledge among marginalized or at-risk people. The song "Follow the Drinking Gourd" provides an historical example from the US. The spiritual is a *song map*, i.e., a map that uses words - usually in an oral culture - to communication geographic knowledge [25, 187]. Using only words, it told enslaved people in the US South how to reach a particular point along the Ohio River where someone could ferry them across to a non-slave-holding region. Beyond that safer free state was an assisted path north to greater safety in a non-slave-holding country. Martin Luther King Jr. wrote,

> *"Our spirituals... were often codes... One of our spirituals, 'Follow the Drinking Gourd,' in its disguised lyrics contained directions for escape. The gourd was the big dipper, and the north star to which its handle pointed gave the celestial map that directed the flight to the Canadian border."* [115]

Later verses of the song provided more navigational details, such as two smaller rivers, a pass between two hills, and dead trees to *"show you the way."* Brunson reminds us that the verse about the dead trees "refers to the fact that in the northern hemisphere, moss grows on the north side of the trees and can thus be used to point travelers in the right direction in the absence of the North Star." [25].

Because of the differential legibility of the song, enslaved people could sing it and teach it without punishment - sometimes even within the hearing of the enslavers (for whom the content was not legible). Among people who were not allowed to own property, the song was a fully portable map that could be carried and used anywhere, because it persisted solely in human memory and human voices. The selective legibility of the song made it memorable for enslaved people, and forgettable for the enslavers.

A contemporary example makes a similar point in an inverse way. The US conducts a decennial count of the population (a census). Two aspects of the census are crucial for this paper: (a) The census is a count of *people*, not limited to *citizens*; (b) There has historically been a clear, protective data-boundary (a localized silence) between census data and law-enforcement agencies. That is, the census data-collection was explicitly defined with Selective Legibility, to encourage a full count which would safely include people who needed to remain unknown to legal authorities. The outcome of the census is used to compute governmental aid to localities, and to revise the number of elected representatives, as well as who can vote for which reprsentatives. However, under a reactionary President, there was a threatened *breach* of that Selective Legibility (between census and law enforcement) in the 2020 census, which would allow immigration police to discover and deport people who did not have citizenship or immigration papers. In this case, the preceding guarantees of selective legibility (through de-identified Census data) were placed into doubt, apparently with an intention to reduce Census counts from urban and Latinx areas [117].

We also find issues of Selective Legibility in contemporary HCI research. For example, Bellini et al. describe the tensions between the safety of silence and the importances of connected conversations among people who are survivors of domestic violence and those who support them [14]. Similarly, Strohmayer et al. describe how sex workers need to share life-saving information about potentially dangerous clients at the community level, to keep one another safe. However, they must do this in ways that are only selectively legible to ensure that this information remains illegible to non-sex working communities and some legal authorities [210, 212]. These communications are often shared in various media and formats, both digitally and non-digitally (such as on flyers or in online fora), in non-public venues with varying degrees of privacy. Looking towards a different community, Yarosh and colleagues explored tensions between privacy concerns and participation in both face-to-face and online twelve-step programs [93, 188], also indicating the need for selectively legible data that can help support those within the program, while ensuring their privacy remains intact outside of the program.

3 SOURCES OF BIAS

As we have shown in section 2, forgetting is complicated and may further the safety of individuals and communities, but may also cause additional harm. But why is it that we forget, intentionally or not? Here, we want to distinguish between two major approaches to bias in data science: extrinsic bias and intrinsic bias.

Extrinsic bias is concerned with a view of a biased dataset ”from the outside.” The argument is that an already-biased dataset can cause even innocent software to produce a biased outcome - and may look like people saying things such as ”the data made me do it.” This has already been well-documented as a domain of active study in the data science literature, particularly when looking towards discourses on ”fairness” - such as [11, 13, 27, 49, 91, 140, 144, 163, 182, 198]. Recent important projects are developing ways to detect, analyze, and mitigate bias in datasets [13, 182], and there are now so many definitions of fairness that entire papers are written to compare those conceptual and computational models and whom to include in evaluating those models [11, 144, 181, 198]. If we fail to remember that a dataset is biased, then we may treat it as ”fair” or ”representative,” harming people who have been excluded from it.

But what if our software is not so innocent? Through practices of data wrangling, curation, and feature-engineering, humans make a series of decisions about how to treat their data, and those decisions may inadvertently introduce bias into the data (see detailed examples in Aragon et al. [7]). Researchers have paid less attention to *intrinsic bias* - i.e., the ways in which we change the data ”from the inside” of data science work-processes while we are preparing the data for modeling. Some of the current research in this area was summarized in [156, 192]. We extend those arguments in the next section of this paper, and we propose sociotechnical improvements in Section 5.2. We claim that *forgetting* currently occurs in many of the activities related to data-preparation. This kind of bias is concerned with a view of potentially biased *data work practices* - a view ”from the inside” of the ways that we add distortions to particular records and fields through methods like cleaning, curating, wrangling, etc. We understand that these are necessary steps in data work, and we emphasize that people with goodwill, will try to do these steps as responsibly as they can [157, 192]. The classes of problems that we want to highlight are paired:

- Much of our work to make these necessary changes is not governed by concerns for bias, fairness, or even a strong awareness of the consequences of our actions. We do the work that needs to be done, and we make changes that appear to be obvious and common-sense (e.g., [189]). Sadly, unexamined common-sense decisions can introduce bias beyond the intentions of the practitioner [33, 181].
- For each change that we make, there is little infrastructure (of practices or of technologies) to record those changes, and even less infrastructure to record the rationale for those changes.

Having had a look at both extrinsic and intrinsic bias in our understanding of how we *forget* in the data sciences, we claim that *forgetting* currently occurs in many of the activities related to data-preparation. But it is also our understanding, that there is the relative lack of tooling to detect, analyze, and mitigate bias *within* the processing steps of record-by-record or variable-by-variable data work [29]. We present a description of how this happens in practice, in section 4, by building a *forgettance stack* as it occurs in machine learning projects.

4 INTRINSIC BIAS: BUILDING A FORGETTANCE STACK IN MACHINE LEARNING

"Why don't we know what we don't know any longer?"
–Proctor and Schiebinger [180]

In their provocative definition of *agnotology* (a science of forgetting - see also *amnesiology* [177]), Proctor and Schiebinger ask a series of questions about how forgetting happens in organizations and societies, and what the positive and negative consequences may be [180]. In this section, we attempt to answer their plaintive question (above) as it applies to data sciences, and specifically to machine learning projects.[5] This is important, because "[c]urrent practices of data cleaning and data readiness assessment for machine learning tasks are mostly conducted in an arbitrary manner" [5], and machine learning practices tend not to preserve disciplined histories of what was done to data, or how it was done, or by whom [112, 114]. Later in this section, we will consider the broader issues that may motivate the forgettance in data science.

Data science work in machine learning typically goes through a series of stages. It has sometimes been convenient to think of machine learning as a sequential process [81, 125, 138, 156, 227]. However, more recently, researchers and practitioners have described a more iterative process [94, 226, 230]. Nonetheless, as with many scientific endeavors [132], data science workers tend to focus on the current step, and to move *forward* to the next sequential challenge after they have solved that problem. Often, the current step is demanding, and data science workers may concentrate all of their energy on informal problem-solving activities [114] rather than on documenting their work - i.e., on exploration rather than explanation [189].

In this paper, we are concerned with what we forget at each step in this process - and so far, we have described what some of the reasons for this forgettance may be. Now, we present a specific example of how this forgettance is put into practice - intentionally or not - through data science work. We describe machine learning as a process in which data science workers gradually create layers of knowledge [156, 171], with each layer built "on top of" the previous layers, as shown diagrammatically in Figure 1. The layers become a kind of "stack" in which the data are processed from bottom to top, and in which the knowledge extracted from the data becomes more and more sophisticated and productively abstracted as the data move "up" the stack [57, 123, 236]. Our concern in this paper is for the knowledge that we *lose* while building this stack. We will show in this section that, while we are building more sophisticated knowledge, we

[5]We have focused on supervised machine learning for convenience. Nearly all of our concerns about the *human construction of data* apply equally to unsupervised machine learning, reinforcement learning, genrative AI, etc.

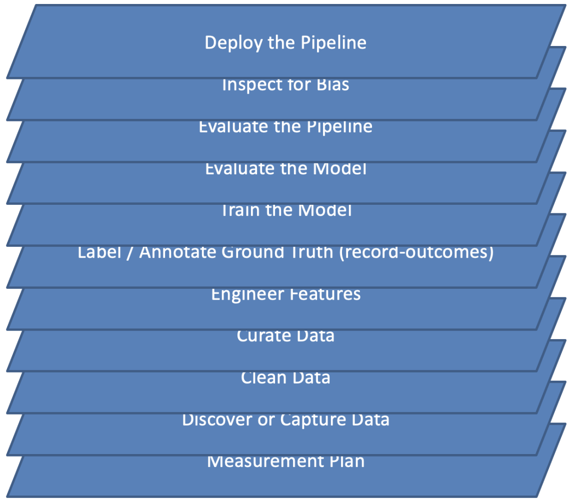

Fig. 1. Forgettance stack of data work on the records and variables of data science. Each action tends to push previous actions into the infrastructure, where the action itself and its consequence are easily forgotten. We indicate this reduction in legibility and remembrance by partially overlapping the layers, such that lower layers are made less legible by upper layers.

are also forgetting earlier knowledge. Later, we will consider the nature of those forgetting processes, and their possible motivations.

Multiple machine learning lifecycle models have been published (e.g., [72, 179, 226]). For this section, we built on an earlier sequential description of specifically *human* actions during the machine learning lifecycle [156]. We believe that the points we make in this section apply to other published models. Using this description, we will build one of many possible *forgettance stacks* of data science, and we will describe the forgettance that occurs at each level of the stack.

4.1 Measurement Plan / Syntactic Silences; WYSIATI

There are many diverse accounts of the data science cycle or process. As Pine and Liboiron have shown [176], most accounts begin with a ”measurement plan” that describes data sources, analytic intentions, expected outcomes, and sometimes clients or customers. As Pine and Liboiron describe explicitly, there is often a politics to these measurement plans [176].

> Though often described as ‘raw,’ this data is produced by techniques of measurement that are imbued with judgments and values that dictate what is counted and what is not, what is considered the best unit of measurement, and how different things are grouped together and “made” into a measurable entity... It is usually assumed that the human element has been scrubbed from the database and that significant political and subjective interventions come from the analysis or use of data after the fact. Instead, we argue that human-computer interactions start before the data reaches the computer because various measurement interfaces are the invisible premise of data and databases, and these measurements are political.

Aspects of these problems may have their roots in a data science team’s understanding of what problem they are trying to solve - which can be a complex and difficult process to solve [143, 170]. Working to reduce bias from an extrinsic perspective (see above), Selbst et al. describe five types of errors (”traps”) that can lead to biased outcomes through mismatches of human needs with existing or prior systems [198]. Martin et al. propose that data science teams should include a larger and more diverse group of stakeholders, including the people and organizations that may be affected by a data science system or deployment. They note that the language of data science analysis may present an obstacle to community involvement, and they hope that a more participatory approach might solve that problem [144].

Crucially, a measurement plan defines not only timelines and project activities, but also *the data themselves* - i.e., what measurements are considered to be ”data” [48, 176, 196]? What are the quantitative or qualitative attributes of the data? What data attributes qualify as ”valid”? These are human decisions [156] requiring human discernment [64, 171] that are often the reflection of human social negotiations [95, 176], especially in inter-disciplinary projects and in bespoke projects that have to meet both intrinsic definitions of rigor and extrinsic client-originated definitions of relevance [157].

One of the problems with measurement plans is the changing understanding of the people who are doing the planning. Mao et al. described the often-lengthy process through which teams initially try to determine how to find an *answer* to a question, only to discover that they need to revise or redefine the *question* itself - or to find a different and more powerful question [143]. Passi and Barocas [170] criticize simple applications of known or ”normative” problem assessments (similar to the ”traps” of Selbst et al. [198]. They observe that ”the specification and operationalization of the problem are always negotiated and elastic.” They emphasize that the data science team has to perform a translation task from a problem in-the-world, into a problem in-the-business, and then into a data science formulation. Their work, along with that of Mao et al. [143], adds an extended temporal dimension to the analysis of Pine and Liboiron [176]. Each translation step requires additional interpretation into data sources and data formulations, imposing further decisions upon the humans who carry out the work.

Measurement plans tend to record conclusions, not rationales [176]. Other people then work with those conclusions, and have no way to access those unrecorded rationales. The intentional or unintentional omissions may lead to the unintentional creation of Syntactic Silences (Table 1 row 2). If the data are incomplete (perhaps through Syntactic Silences), then there is the further risk of *assuming* that the data are nonetheless sufficient - e.g., WYSIATI (”what you see is all there is,” Table 1 row 5). Are these social-process criteria recorded? Do we forget the initial criteria, and their intentions, as we revise the questions and rewrite the plan? And what happens to the measurement plan in the next stages of data science?

4.2 Choosing the Data / Substitutions; Annulments

The measurement plan is intended to guide the selection of data for analysis. Within the data sciences, ”the data” are usually considered as a concrete, unquestionable set of ”facts” that describe a similarly unquestioned ”real world” [92, 220]. However, according to D'Ignazio and Klein [48] and boyd and Crawford [23], the selection of data is also a human process, requiring human discernment. Bilis goes a step further, distinguishing between data that are ”discovered” vs. data that are ”captured” ([16]; see also [89]). While the action of capture implies active human intervention, even the action of discovery requires a human to perform or *make* that discovery. Further, data science teams often replace one data source with another to respond to project needs - e.g., from surveys to videos to mobile phone records [156]. As we switch from one data source to another - often for reasons of efficiency or economy - then we may also be moving from relatively direct data and into indirect traces of the data (Substitution, Table 1 row 4) [38, 56]. While we may use processes of Annulment (Table 1 row 6) to focus attention on a subset of data of particular problem of interest, there is again the risk of WYSIATI if we forget how we focused our attention through Annulment.

In this process of human *recognition* and *selection* of data, there is a subtle shift in the status of the data itself. The perspective of the data sciences might initially treat data in the abstract as having an "objective" existence that is independent of human action. However, by the time we have discovered or captured the data, we have engaged in multiple human and collective *interpretive* actions (see again [170] for a discussion of interpretation and translation in data science). The origin of the data may remain in a realistic world, but the data as *taken for use* in data science now also reflect the views, assumptions, and biases (conscious or unconscious) of the humans who engaged in the speech act of saying "These are the data in our project." The contents of the measurement plan set up these speech acts by defining data in certain ways, and implicitly refusing to define data in other ways. The supposed realism of the data is constructed (reified) in the measurement plan.

However, the relevance of the measurement plan seems to fade as data science workers improvise their data sources when faced with issues of effort, scale, and cost. As the measurement plan becomes less relevant, people are less likely to record how and why they deviated from that original plan. The changes in practice, which could also be changes to the measurement plan, are rarely recorded, and tend to be lost.

4.3 Cleaning the Data / Syntactic Silences; Substitutions; Sanitized Erasures

The effort of choosing data is small compared with the effort of cleaning (or "wrangling") the data [81, 108, 183]. While descriptions of the cleaning of data are often phrased in terms of statistical transformations [183] or the replacement of missing values ("imputation"), it is clear that these are often human decisions that require human skill and discernment [52, 156]. In a recent paper about *reforming* the practices of data cleaning through the MLCLEAN toolset, Tae et al. provide examples of common-sense reduction of duplicated records and replacement of an outlier data field with "a reasonable value" [213]. However, even in this reform effort, the authors adopt the conventions of computer science, and do not tell us *who* decides whether two similar (but not identical) records are actually duplicates, and who decides what a reasonable value may be. Substitution (Table 1 row 4) and Syntactic Silences (Table 1 row 2) appear to be quite likely - and undetectable afterwards, because we have no way to remember what we did (i.e., Sanitized Erasures, Table 1 row 13).

We know from the standpoint literature ([87]; see also [86, 139]) and the literature on boundary objects [122, 135, 174] that people with different backgrounds may live and work in different social worlds, where "reasonable" is a local, situational, and/or social construction. There are many similar accounts of disembodied reasonableness in the data cleaning literature [81, 108, 183] that become another form of data science forgetting. When we forget *who* did the cleaning, then we correspondingly forget *whose* definitions of reasonableness were involved. If we do not preserve the lineage or provenance of these detailed changes to the data, then we cannot inspect, interrogate, and reverse those changes upon need. We implicitly engage in a form of Prescriptive Erasure (Table 1 fow 8). When we forget *who* acted, we also forget *how* and *why* they acted, and *what* they did, and we forget how to reverse those actions. Without self-documentation of what we have done [189], we may forget our own data-cleaning actions.

In the event of *quantitative* imputation there are many choices of mathematical methods [18, 130], while in *qualitative* imputation (e.g., for classifications or categories) there may be simple statistical approximations [104]. In the cases of statistical methods, it is often possible to apply the imputation to an entire variable or factor in a single conditional operation based on the most-frequent of the non-missing classes or cateogory-labels (e.g., "if missing, compute..."). In other cases, there may be important dependences on domain knowledge, during a manual process of replacing missing values on a record-by-record basis [9, 213], especially when human familiarity with the data and its domain suggests that "something doesn't look right" in the data ([46]; see also [72]).

Knuth proposed *literate programming* with a goal of rethinking software as a means of communication among humans, as well as between humans and machines [119]. Thirty years later, a contemporary environment for literate programming, inspired by Knuth's ideas, is the Jupyter notebook, in which "code cells" of software are intermixed with "markdown cells" of formatted documentation. Jupyter notebooks are commonly used in data science, and they seem to offer an opportunity to serve as memory aids [113] in which we write code for processing data (in a code cell) and simultaneously document the rationale for that code for others or for our future selves (in a markdown cell). At first glance, a Jupyter notebook appears to be a superb tool for remembering the purpose, rationale, and strategy of data-processing code.

However, Rule et al. analyzed a million Jupyter notebooks from Github, observing the relative scarcity of self-documentation [189]. It seems clear that many of these human decisions about imputation strategies and operations go unrecorded, despite the ease of using Jupyter notebooks in what might be called a "memorious way" - i.e., a way to support the writing and sharing of knowledge.

We might think that an analyst could examine a colleague's code to find out how that colleague wrangled the data. That strategy could work well if people wrote a single, unified set of code while cleaning their data. However, Rule et al. also reported that data science workers often pursue multiple, contradictory, parallel or sequential experiments in finding the best data treatment. To coin a phrase, data science workers are "coding out loud" (similar to "thinking out loud") as they try different alternatives. Without documentation, it may be too difficult and too uncertain to determine *which* transformation was made among many trial transformations, and *which* imputation scheme was applied among diverse imputation strategies.

Kery et al. provide examples of this kind of forgetting within data science code [114]. They reported a series of questions that programmers wished to answer when inspecting their own code and data, such as: "Find me how I cleaned the data from start to finish"; "What questions did I ask that didn't pan out?"; and "[P]revious test result for this particular dataset". In practice, the details of wrangling are often lost, and so is the ability to ask the kinds of questions that participants suggested in Kery et al.'s study [114]. Because we can no longer answer questions of this type, we tend to pass the dataset along to the next step, as if there were no uncertainties and nothing that we might need to revise later. A strategy of Annulment to focus on the current problem (Table 1 row 6), tends to become an unintentional strategy for Prescriptive Forgetting in which there seems to be a consensus that certain things are best forgotten (Table 1 row 7). Kery et al. recently created the Verdant system [113] which shows promise for making past coding decisions more legible and understandable. A more data-centric version of Verdant could provide a memory aid to address some of the issues we have raised here.

4.4 Curating the Data / Syntactic Silences; Prescriptive Forgettings; Repressive Erasures

Definitions of data curation vary. Some scholars even write about a complex process that includes aspects of wrangling as part of "purging of dirty data" [8]. In this section, we are concerned with a narrower interpretation of curation as data-selection *within a dataset* [12, 45]. This can become a strategy of Syntactic Silence (Table 1 row 2) that tends toward Prescriptive Forgetting (Table 1 row 7) and can lead, for certain deliberately-rejected classes of data, to a form of Repressive Erasure (Table 1 row 8).

In the HCI tradition, curation can refer to how a data science worker prepares data for use by another entity - either a human [145, 216] or an algorithm [8, 183]. A typical activity is the removal of outliers [8, 65, 107, 108], based on the values in one or more fields of each data record. We note here that the person who is removing outliers may not be the person who performed the operations described above in data cleaning. They may not know which values are "original" from the dataset, and which values were altered through data-cleaning, or imputed to replace missing values. In a manner of speaking, the dataset has "forgotten" about those prior operations, because there is no record of them (see Syntactic Silences, Table 1 row 2). All data appear with the *same degree of confidence or certainty.* The experiential knowledge of which fields have been modified, is often lost.

The stakes of these outlier decisions may be high, especially if each record corresponds to a person or a family [199]. When faced with the risk of (e.g.) removing most or all BIPOC or disabled people on the basis of income-level or home-ownership status, then it would be important to know how trustworthy each outlier data value is. We may need to know which values were altered, but we may not be able to access records of how the data were modified through human or algorithmic actions. All we have are the data in their current form. While there are multiple proposals to record the source of individual or combined *datasets* [26, 110, 201], the corresponding concept of provenance of individual data *records* has received less attention. Even the proposed records of data transformations by Glavic et al. deal with an entire factor or variable at-a-time, without recording individual decisions at the level of the data record [75]. And so, we do our best to remove outliers, but we forget both the outlier records themselves (i.e., they are no longer in the dataset) and also the reason why we decided that those records were outliers.

As we showed in Section 4.3, we tend to forget (in mind and in data-records) the metadata that could help with questions of who, what, why, and how. The same lesson applies to the curation of outliers in this section. We may also forget any steps that we took (or did not take) to see if our outlier criteria might be erasing categories or classes of records (e.g., of people).

4.5 Feature Engineering / Prescriptive Forgettings, Structural Amnesias

Many machine learning models make good use of existing values (factors) in the dataset. Often, however, there is additional information being constructed through non-linear combinations of data fields [167, 219]. As we noted earlier (Section 4.1), these are human decisions. Data science workers apply their general knowledge or (in some cases) their domain knowledge to translate [170] those ideas into features that "make sense" in the context of other data and their background knowledge of the field [173]. In this way, they *design* the data that the model will subsequently consume [63, 64, 197], "handcrafting" aspects of their data [116, 156]. Common examples of engineered features are ratios (i.e., non-linear combinations of more basic predictors), such as weeks of employment divided by total lived weeks to compute a common sense "percent of weeks of full employment" during an adult's employment history.

Even with simple ratios, there can be important decisions. The computation of work history could be constructed as percent-of-weeks-worked divided by percent-of-weeks-lived. The denominator can make a big difference, especially for younger people - e.g., was there a correction factor for the number of weeks-in-school? Each form of computation carries human social knowledge or assumptions, such as an upper-class assumption that people in school do not also work, or that people below a certain age do not also work while in school. Unless the feature-engineering is carefully documented, we forget how we designed that part of our data, and we may unintentionally encode our own standpoint (i.e., the assumptions of our social position, based in class, race, gender...) in this buried step.

We may think of the data in the original dataset as first-order predictors. In that framework, the engineered features become second-order predictors. The quality of the second-order predictors depends on both the human's knowledge and also the quality of its components - i.e., the first-order predictors. If the person who is translating concepts into features did not also clean and curate the data, then they may not know about uncertainties or reasons to be skeptical of certain first-order data. The result is engineered features that appear to be reliable. Their earlier history of human decisions is lost - another possible instance of Annulment (focus on the data of interest) and Prescriptive Forgetting

(Table 1 rows 6-7). This forgetting is of course convenient for people who performed the earlier data-wrangling, because their well-intentioned decisions are less subject to scrutiny or question (see Structural Amnesia, Table 1 row 11).

4.6 Labeling and Annotating (Ground Truth Practices) / Prescriptive Forgettings; Colonial Unknowings

Often in machine learning, there is a need to *create* data. For supervised machine learning, there is usually a need for a predicted (or "dependent") variable that the machine-learning model is supposed to predict, especially if the prediction is about classes or categories of data [73, 124, 178]. These predicted values are often called *ground truth*, and may be produced through anonymous crowdsourcing [78, 142] or through the applied knowledge of domain experts [67, 194]. The contents of the ground truth data-field on a particular record has been called a *label* or an *annotation*, and the role of the people who assign these values has been referred to as both *labeler* and *annotator*.

As Bowker and Gitelman have observed, "'raw data' is an oxymoron" [20, 74]. Ground truth is often constructed ("cooked" to remove its rawness, as it were) by humans, and then predicted through training a model. It is worthwhile to consider how this raw-to-cooked construction takes place. Traditional accounts of machine learning seem to treat the crowdworkers and domain experts as types of sensors, as if humans could provide an objective and infallible reflection of the nature of the world. However, studies of the construction of ground truth show that these values are *made* by humans (e.g., [63, 64, 197, 214]), and reflect not objective reality, but rather human sensibilities and also the specific contextualized demands of labeling as situated practices [49, 84, 147, 156, 157]. D'Ignazio and Klein write that "data are not neutral or objective," but are "products of unequal social relations" [49], and they argue that data begin to lose their meaning when they are abstracted away from their context. Borgman [19] and Bowker [20] note the importance of context in the human activity of *making sense* of data - including both formal data structures and informal social relations (e.g., [37, 191]). In these terms, "'Ground truth' begins to look less like a formal or 'objective' truth, and more like a worthwhile social accomplishment" [157].

In many projects, data science workers collect more than one label for each record. This can be a kind of quality control [67, 157] or even a way to estimate the reliability of citizen science labelers [98]. Miceli et al. showed that people who create ground truth labels may disagree about the most appropriate label for a particular record, with diverse protocols used to resolve those conflicting labels ([147]; see also [37, 98, 157]), such as choosing the label that was most popular among the labelers. In contrast to records on which all labelers agreed on the label, the existence of these disagreements could signal lower confidence in the contested labels - *if* we had a way to record that lower confidence, and *if* we had a way to use that confidence metadata while computing the model.

Disagreements based on different standpoints or worldviews among the labelers, may be particularly important for data that have social implications. However, in the data sciences, our practices are designed to forget those disagreements. In common practice, each record in the dataset is supposed to have a single, unitary ground truth value. Thus, when data science workers take the dataset to the next stage of the process, all ground truth labels are treated as being equally and *uniformly* authoritative. The assumptions of uniformity reflect points that we made earlier in this subsection, about positivist assumptions of humans as noisy "sensors" of a single, unified reality. Because labeling is a relatively expensive part of the data science cycle [67, 183, 194], there may be incentives to forget that contested labels might be less reliable than unanimous labels, and might require further labeling with a larger number of labelers. Because the epistemology of data assumes uniformity among labelers, the existence of different perspectives and situated perceptions are also forgotten. Syntactic Silences again tend to become Prescriptive Forgetting (Table 1 rows 2 and 7). If there are "minority" or "disfavored" perspectives among disagreeing labelers - perhaps reflecting different experiences of gender, race, or class - or different interests of developers vs. clients - then we may also see genteel forms of Humiliated Silence and/or

Colonial Unknowing (Table 1 rows 9 and 10). The inconvenient information is silenced. The metadata about potentially lower-confidence labels is lost.

4.7 Training the Model and Deploying the Pipeline / Prescriptive Forgettings; Repressive Erasures; Colonial Unknowings

All of these activities become forgotten antecedents when it is time to train a model [183]. As Sambasivan et al. have observed, ”Everyone wants to do the model work, not the data work” [192]. The antecedent ”data work” [151] tends to fade into the background, becoming layers of invisible human infrastructural work (e.g., [208]). Hutchinson et al. observe that the ”Datasets that power machine learning are often used, shared, and reused with little visibility into the processes of deliberation that led to their creation” [99], because of the devaluing of data work as contrasted with model work that Sambasivan et al. described [192].

The dataset now becomes ”the data” and becomes infrastructural to the modeling work. There is, within data science workers, a constituency to support this form of Prescriptive Forgetting (Table 1 row 7) - if only for matters of convenience - e.g., working with a single dataset is much easier and also *less questionable* than working with multiple, partially-contradictory versions of the dataset. People cannot easily perceive or make use of the forgotten knowledges of choices, improvisations, and uncertainties. After the model has been perfected, it is typically wrapped inside of a monolithic deployable pipeline [40, 200], which both contains and obscures the ways in which the data have been captured or discovered, cleaned, wrangled, curated, and labeled.[6] Indeed, some important machine-learning products are deliberately rendered entirely opaque, with the stated motivation of protecting intellectual property. However, the products that contain these opaque pipelines influence or control important human decisions in areas such as criminal justice [24, 140, 228], bank loans [91, 118], and who is stopped and searched by legal authorities [36].

Opaque pipelines are more difficult to challenge or interrogate. We cannot analyze how they operate on data [228]. We can only analyze the outcomes - e.g., through methods for detection, analysis, and mitigation of bias [13, 27, 163]. We forget the complex and tension-filled work that creates ”the data,” which ceases to be construed as a ”dataset” as it becomes part of an opaque ”system” or ”algorithm.” These deliberately unknowable (or ”pre-forgotten”) algorithms may provide examples of Repressive Erasure (Table 1 row 8), if there are possible problems with the predictive model. Because the creators of the opaque algorithms presumably know about potential weaknesses, while the rest of us do not, these situations may also be analyzed in terms of Colonial Unknowing (Table 1 row 10), in which one interested party wants to make certain data unknown (i.e., selectively silent, hence forgotten) to other interested parties.

In this case, the verb *forget* takes on a peculiar, transformative function. In English-language grammar, we might say that it takes a ”direct object” - i.e., the point of the action is to remove or erase a *target* kind of memory, to transform it into an oubli (something that is forgotten) [168] or a damnatio memoriae (something that one person wants other people to forget) [229]. As we *forget the dataset* by tranforming it (cognitively) into ”the data”, the data entity (and the human decisions that have shaped it) fade into the infrastructure [99, 159, 214]. When this happens, the data acquire a sense of inevitability and objectivity [23, 80], as if they reflected the nature of the world, rather than the constructions of a particular group of humans [63, 64, 156, 157, 169, 171, 197]. Through that transformative forgetting, we remove the knowledges of both the uncertainties that people experienced during data-preparation, and the potential weaknesses or other reasons to re-examine the processes leading to the creation of the data.

[6] We note that there are research projects that are experimenting with more transparent pipelines, such as: [31, 99, 134, 233].

Table 2. Vocabularies of Remembrance and Forgettance

Perspective:	**Remembrance**	**Forgettance**	**Erasure**
Actor:	Rememberer	Forgetter	Obliviator/Unknower
Object:	Memory	Oubli	Damnatio memoriae
Assistance:	Aids to memory	Engines of forgetting	Forces of unknowing
Capability:	Memory	Forgettery	Doublethink/Unknowing
Community:	Remembering community	Forgetting community	Unknowing community
Stakeholders:	Beneficiaries	Beneficiaries	Beneficiaries and Maleficiaries

We are left with a seemingly perfect thing that we call "data." That seeming perfection aligns with the meta-narrative (e.g., [141]) of powerful, objective, and inevitable outcomes that seem to be based *in data*, rather than *in the human processes of the construction of a dataset*, which becomes reified as the data [15, 36, 49, 149]. When we accept those silences, we contribute to a kind of god-trick [84], in which (inevitably fallible) human actions are made to appear to be authoritative, naturally-given, "true," and consequently difficult to interrogate or challenge.

4.8 Summary: The Forgettance Stack

In Section 4, we provided a linearized sequence of activities in the data science cycle [81, 125, 138, 156, 227], while acknowledging that the lived work of data science is even more complex than our simplified version [94, 226, 230]. We hope that we have shown how much information is forgotten in the simplified sequence, in which the original measurement plan may be overridden without being overwritten (so to speak). The decisions about the definitions of data are quickly forgotten beneath a series of additional decisions, opportunities, improvisations, assumptions, and enactments - each of which renders previous human actions less and less known. Humans add value to their data, and they build their value-additions into their processing software. With the best of intentions, humans forget - or never know - what other humans have done while making the human decisions that result in the data-processing steps [46, 95, 111, 156, 192]. They may even forget what they themselves have done [114, 189, 234].

5 DISCUSSION

By bringing together the typology of forgetting and the notion of a forgettance stack, we have presented a variety of ways of 'forgetting' that take place in data sciences. Throughout the paper, we have presented various contemporary and historic examples, often focusing on experiences of those who have historically been marginalised or excluded.

5.1 Implications for Conceptualizations

To summarize the detailed arguments of the paper, we propose a simpler vocabulary that can integrate the traditional HCI concerns of actor and object/artifact, clarified by concepts from studies of remembrance, forgettance, and erasure (see Table 2).

The **Remembrance** column presents a conventional understanding of memory practices, based on Bowker's *Memory Practices* [20]. In this rendering, HCI and data science workers collectively strive to record, curate, and categorize matters of shared concern for use by selves or by a future community of scholars and engineers, using well-known and powerful aids to memory (e.g., databases) [2–4, 82, 125, 136, 154, 162]. The view of stakeholders is similarly straightforward - i.e., as *benficiaries* of the sociotechnical work of remembrance, workers and scholars gain knowledge and computational power from from these records. This view reflects the assumption of *innocent software* that correctly and completely models the data for non-injurious use by others. Any bias in the outcome is assumed to be due to problems with the data, and *not* with the human decisions that shape the software [13, 27, 49, 163, 198]. We combined concepts from human centered data science [121, 155, 156, 232], human centered machine learning [30], ground-truth labeling/annotation studies [67, 157] and feminist technoscience [14, 36, 47, 84, 86, 158, 211] to trouble this simple view.

The **Forgettance** column summarizes our perspective in this paper , which we propose as a necessary and complementary view to that of Remembrance. Workers in HCI and datascience use well-recognized tools for forgetting (principally curation practices for datasets) to help self and others to focus on the data of current concern. As discussed in Section 2, Forgettance has been considered as both the opposite of Remembrance [66, 129, 149, 165, 229], and also as a *component* of Remembrance (i.e., of successful memory practices) [20, 39, 58, 131, 148, 152, 224]. In these terms, the stakeholders for Forgettance are as uncomplicated as those for Remembrance - i.e., workers in HCI and data science are primarily *beneficiaries* of Forgettance in the service of Remembrance. Bowker [20] and Lamers [131] wrote of the heuristic need to remember what matters by forgetting what doesn't matter (e.g., [148, 152, 224]). We noted in Section 1, de Souza and colleagues described a reversible kind of forgetting in their study of API-related work-practices in programming [41, 42, 42, 43, 193], and we showed in Section 4 that much of the work-practices of data science do not provide such reversibility in our data science forgetting practices [48, 95, 145, 157, 176, 216]. Nonetheless, the Forgettance column is also a predominantly "innocent" view, which reflects some of the less worrisome silences from Table 1, namely: Syntactic Silences (row 2), Inferential Silences (row 3), Substitution (row 4), WISIATI (row 5), Annulment (row 6), and often Prescriptive Forgetting (row 7).

We therefore summarize a third perspective, the **Erasure** column, which could also be called the **Unknowing** column. What distinguishes this column from the Forgettance column is primarily matters of *intention*. The social forces that practice erasure or unknowing are generally intended to hide or erase data that others may wish to know. There may be helpful reasons for erasure or for selective legibility (e.g., [14, 206, 209, 211, 212], but there may also be harmful reasons for such actions [28, 34, 146, 146, 218]. This third perspective provides an re-entry-point to the more critical perspectives of the paper. As we discussed in Sections 2.3 and 4, forgetting can become a form of obscuring, of hiding what we wish to forget, or what we wish someone else will forget, or of what we want to prevent someone else from ever knowing. When motivated, this kind of erasure can become a form of deliberately silencing or obliviating. Here is where we might apply the more worrisome concepts from Table 1, including Repressive Erasure (row 8), Humiliated Silence (row 9), Colonial Unknowing (row 10), Structural Amnesia (row 11). In the two previous propositions, we could assume benevolent intent. However, for Erasure, the characterization of stakeholders becomes more complicated as we consider who benefits (beneficiaries) and who may be harmed (maleficiaries) through these strategies and actions.

Scholars of value sensitive design [70, 90] and feminist technoscience [10, 36, 85–87, 158] have argued that we need to look not only at the data - we must also consider the people involved, as well as their intentions and contexts. We propose that these lessons apply as well to our readings and formalisms for remembrance, forgettance, and erasure. As we have presented in Table 2, the notions of remembrance, forgettance, and erasure relate to the actor (person or persons doing the remembering, etc.), as well as the object and any assistance they have with the practices (e.g., [145, 156, 170, 216]). This then of course also relates to the capabilities (e.g., memory, forgettery, and doublethink/unknowing). Of course, all of this relates to the communities in which these practices sit [21, 131], as well as the various stakeholders who are involved in these processes. In keeping with these thoughts, we recognize that four of the complex silences from Table 1 are more difficult to describe as being *either* simply "innocent" *or* "harmful":

- Structural Amnesia - which could be beneficial for someone in gender transition, or harmful if enacted as propaganda;
- Redacted Data - which could be beneficial for people who redact *their own data* because do not want to be found by authorities, or harmful if someone wants to remove knowledge of *other* people;
- Covert Silences - which could be beneficial to secure the effects of protective Redacted Data, or harmful if it amounts to removing/erasing evidence;
- Selective Legibility - which could be live-saving, as in the example of "The Drinking Gourd", or harmful, as in the example of covert political messaging sometimes known as "dog-whistles."

5.2 Implications for Sociotechnical Practices

Improvements to work-practices and to infrastructures could (a) clarify the intentions of remembrance and forgettance, and (b) reduce the extent of subtle erasures.

Data wrangling, feature engineering, and labeling are actions taken through technologies that make a dataset fit-for-purpose - i.e., well-formed for modeling [151, 192]. As we noted above, these actions inenvitably assert human interpretations into the data [64, 145, 156, 170, 197, 216]. We propose that data science and HCI workers should engage in memory-practices while using these technologies, recording the changes that they make to the data. Correspondingly, we propose that the technologies should be enhanced with straightforward tools that support these remembrance actions. In effect, we are recommending that sociotechnical software engineering concepts, such as separation of concerns and encapsulation [50, 119], should be applied to the sociotechnical practices and infrastructures of data science tools as well. One way to think about this is to add a change-history to conventional dataframes - preferably in ways that support transparency but not surveillance. The change-history could include both simple data-transformations and (where this can be done safely) an automated signature of the data science worker who made that change, as suggested by Passi and Barocas [170]. In appropriate circumstances, a rationale could also be attached.

We note also that some aspects of bias occur subtly, over a range of data records. For example, during curation of records [145, 216], a data science worker might inadvertently exclude members of marginalized or minoritized groups. The exclusion would be difficult to detect *while doing the work*. Using today's tools, the exclusion might go unnoticed, or might have to await a post-wrangling or post-modeling bias analysis such as described by Bellamy et al. [13]. We propose a second sociotechnical approach in which a diligent data science worker could pre-designate a set of sensitive or protected attributes, such as race, class (e.g., ownership-status in housing), or gender-identity in a new form of work-tracking tool in the wrangling software. The software would keep a running tally of inclusions, exclusions, and potentially other outcomes, summarized across records, while the wrangling work proceeded. The data science worker could then check their outcomes on a periodic basis; they also might set some threshold exclusionary values, and request to be notified by the wrangling software if they exceeded the limits that they themself had set.

The effort to configure this kind of tool would be minimal - simply designate a small number of factors to be watched, and then use the automated tallies of the values of those factors. If the social signature (from the preceding paragraph) were included, then the data science worker could revisit the records that they had modified, to understand the patterns of their work, and to make changes where needed. Such a tool would enable people to prevent harm by becoming aware of the intended and unintended consequences of their wrangling work while there was still time to make changes. Principles of social translucence [62] could be applied, so that individual workers could revisit their own changes, but other workers and managers would only be able to perceive that the data had been anonymously changed.

6 CONCLUSION

To conclude this paper, we have complicated and unpicked our understanding of "forgetting" in data science practices, with the intention of advocating for increased understanding of and attention paid to forgetting and forgettance in HCI, CSCW, and data science communities. To begin the paper, we summarized prior work on the benefits and harms of forgetting. With this, we presented our first contribution: a classification of data practices related to forgetting, omitting, obliviating, and silencing by presenting a typology of forgettance (as outlined in table 1). In this typology, we analyse three classes of silences that can cause or invoke forgetting: modest silences, silence as force, and ambivalent silences.

Following this typology, we look towards our second contribution: a detailed description of where these kinds of forgetting take place in the data science process, by building a *forgettance stack*. In doing this, we provide a detailed analysis of forgetting the data work in data science, with an emphasis on silences that lead to different dynamics of forgetting throughout the data work cycle.

Data silences and forgettance within data work are complex and multi-valenced processes. We hope to have inspired data scientsits to consider how their data work relates to forgettance, and hope to see other scholars expand our typology, forgettance stack, and thinking by writing about their own categories of silences, and their own interpretations.

REFERENCES

[1] James Abourezk. 1978. Indian Child Welfare Act. 25 U.S.C. §§ 1901–1963, November 8 1978.

[2] Mark S Ackerman and Christine Halverson. 2004. Organizational memory as objects, processes, and trajectories: An examination of organizational memory in use. *Computer Supported Cooperative Work (CSCW)* 13, 2 (2004), 155–189.

[3] Mark S Ackerman and Eric Mandel. 1999. Memory in the small: Combining collective memory and task support for a scientific community. *Journal of Organizational Computing and Electronic Commerce* 9, 2-3 (1999), 105–127.

[4] Mark S Ackerman and David W McDonald. 1996. Answer Garden 2: merging organizational memory with collaborative help. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. 97–105.

[5] Shazia Afzal, Manish Kesarwani, Sameep Mehta, Hima Patel, et al. 2020. Data Readiness Report. *arXiv preprint arXiv:2010.07213* (2020).

[6] Michel Anteby and Virag Molnar. 2012. Collective memory meets organizational identity: Remembering to forget in a firm's rhetorical history. *Academy of Management Journal* 55, 3 (2012), 515–540.

[7] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human centered data science: An introduction*. MIT Press.

[8] Otmane Azeroual. 2020. Data wrangling in database systems: purging of dirty data. *Data* 5, 2 (2020), 50.

[9] B Mathura Bai, Nimmala Mangathayaru, and B Padmaja Rani. 2015. An approach to find missing values in medical datasets. In *Proceedings of the The International Conference on Engineering & MIS 2015*. 1–7.

[10] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips tutorial* 1 (2017), 2017.

[12] Amin Beheshti, Boualem Benatallah, Alireza Tabebordbar, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh, and Reza Nouri. 2019. Datasynapse: A social data curation foundry. *Distributed and Parallel Databases* 37, 3 (2019), 351–384.

[13] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[14] Rosanna Bellini, Angelika Strohmayer, Patrick Olivier, and Clara Crivellaro. 2019. Mapping the margins: Navigating the ecologies of domestic violence service provision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[16] Hélène Bilis. 20188. "Mapping fiction: Social networks and the novel", Wellesley College conference Shifting (the) Boundaries.

[17] Anthony Blinken. 2021. *Proposing Changes to the Department's Policies on Gender on U.S. Passports and Consular Reports of Birth Abroad*. https://www.state.gov/proposing-changes-to-the-departments-policies-on-gender-on-u-s-passports-and-consular-reports-of-birth-abroad/ Accessed Sep 4, 2021.

[18] Shahin Boluki, Siamak Zamani Dadaneh, Xiaoning Qian, and Edward R Dougherty. 2018. Optimal Clustering with Missing Values. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 593–594.

[19] Christine L Borgman. 2020. Big Data, Little Data, or No Data? Why Human Interaction with Data is a Hard Problem. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 1–1.

[20] Geoff Bowker. 2005. *Memory Practices in the Sciences.* MIT.

[21] Geoffrey C Bowker, C Geoffrey, W Bernard Carlson, et al. 1994. *Science on the run: Information management and industrial geophysics at Schlumberger, 1920-1940.* MIT press.

[22] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences.* MIT press.

[23] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[24] Michael Brenner, Jeannie Suk Gersen, Michael Haley, Matthew Lin, Amil Merchant, Richard Jagdishwar Millett, Suproteem K Sarkar, and Drew Wegner. 2020. Constitutional Dimensions of Predictive Algorithms in Criminal Justice. *Harv. CR-CLL Rev.* 55 (2020), 267.

[25] Brianna Brunson. 2020. "Jes Go Back to de Fiel a Singin": The Spiritual as a Vehicle of Resistance in the Antebellum South. *Global Africana* (2020), 3.

[26] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory.* Springer, 316–330.

[27] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency.* PMLR, 77–91.

[28] Julie Cassidy. 2009. The Canadian response to Aboriginal residential schools: Lessons for Australia and the United States. *eLaw J.* 16 (2009), 38.

[29] Aleksandar Chakarov, Aditya Nori, Sriram Rajamani, Shayak Sen, and Deepak Vijaykeerthy. 2016. Debugging machine learning tasks. *arXiv preprint arXiv:1603.07292* (2016).

[30] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[31] Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible ai algorithms: Issues, purposes, and challenges. *arXiv preprint arXiv:2101.02032* (2021).

[32] Jean-Marie Chenou and Roxana Radu. 2019. The "right to be forgotten": Negotiating public and private ordering in the European Union. *Business & Society* 58, 1 (2019), 74–102.

[33] D Christozov, K Rasheva-Yordanova, and S Toleva-Stoimenova. 2019. Designing data science curriculum in a way to address expected students entry competences. In *Proceedings of INTED2019 Conference.* 2635–2640.

[34] Maine Wabanaki-State Child Welfare Truth & Reconciliation Commission. 2015. *Beyond the Mandate: Continuing the Conversation: Report of the Maine Wabanaki-State Child Welfare Truth & Reconciliation Commission.* https://d3n8a8pro7vhmx.cloudfront.net/mainewabanakireach/pages/17/attachments/original/1468974047/TRC-Report-Expanded_July2015.pdf?1468974047 Accessed Aug 26, 2021.

[35] Paul Connerton. 2008. Seven types of forgetting. *Memory studies* 1, 1 (2008), 59–71.

[36] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

[37] Luca De Alfaro, Ashutosh Kulshreshtha, Ian Pye, and B Thomas Adler. 2011. Reputation systems for open collaboration. *Commun. ACM* 54, 8 (2011), 81–87.

[38] Michel de Certeau. 1984. *The Practice of Everyday Life.* University of California Press.

[39] Pablo Martin de Holan. 2004. Managing Organizational Forgetting. *MIT Sloan management review* 45, 2 (2004), 45–51.

[40] Andrea De Lucia and Evi Xhelo. 2019. Data Science Pipeline Containerization. *17th SC@ RUG* 2020 (2019), 39.

[41] Cleidson RB de Souza, Fernando Figueira Filho, Müller Miranda, Renato Pina Ferreira, Christoph Treude, and Leif Singer. 2016. The social side of software platform ecosystems. In *Proceedings of the 2016 CHI conference on human factors in computing systems.* 3204–3214.

[42] Cleidson RB de Souza, David Redmiles, Li-Te Cheng, David Millen, and John Patterson. 2004. How a good software practice thwarts collaboration: the multiple roles of APIs in software development. In *Proceedings of the 12th ACM SIGSOFT twelfth international symposium on Foundations of software engineering.* 221–230.

[43] Cleidson RB de Souza, David Redmiles, Li-Te Cheng, David Millen, and John Patterson. 2004. Sometimes you need to see through walls: a field study of application programming interfaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work.* 63–71.

[44] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 628–635.

[45] José Dias, Jácome Cunha, and Rui Pereira. 2020. Data curation: towards a tool for all. In *International Conference on Human-Computer Interaction.* Springer, 176–183.

[46] Florian Diekmann. 2012. Data practices of agricultural scientists: Results from an exploratory study. *Journal of Agricultural & Food Information* 13, 1 (2012), 14–34.

[47] Catherine D'Ignazio. 2021. Making data work for social justice. Presentation at NotEqual Social Justice Summer School.

[48] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.

[49] Shalini D'Ignazio. 2020. Making Data Work for Social Justice, Social Justice Summer School, UKRI.

[50] E. Dijkstra. 1979. Go to statement considered harmful. In *Classics in software engineering (incoll).* Yourdon Press, Upper Saddle River, NJ, USA, 27–33. http://portal.acm.org/citation.cfm?id=1241515.1241518

[51] Edsger W Dijkstra. 1982. On the role of scientific thought. In *Selected writings on computing: a personal perspective.* Springer, 60–66.

[52] AnHai Doan. 2018. Human-in-the-loop data analysis: a personal perspective. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–6.

[53] Mary Douglas. 1986. *How institutions think*. Syracuse University Press.

[54] Qwo-Li Driskill. 2010. Doubleweaving two-spirit critiques: Building alliances between native and queer studies. *GLQ: A Journal of Lesbian and Gay Studies* 16, 1-2 (2010), 69–92.

[55] MARISA ELENA DUARTE and MORGAN VIGIL-HAYES. 2021. How We Connect. *Indigenous Peoples Rise Up: The Global Ascendency of Social Media Activism* (2021), 64.

[56] Marisa Elena Duarte, Morgan Vigil-Hayes, Sandra Littletree, and Miranda Belarde-Lewis. 2019. Of Course, Data Can Never Fully Represent Reality": Assessing the Relationship between "Indigenous Data" and "Indigenous Knowledge,""Traditional Ecological Knowledge," and "Traditional Knowledge. *Human biology* 91, 3 (2019), 163–178.

[57] Richard Dybowski. 2020. Interpretable machine learning as a tool for scientific discovery in chemistry. *New Journal of Chemistry* 44, 48 (2020), 20914–20920.

[58] Mark Easterby-Smith and Marjorie A Lyles. 2011. In praise of organizational forgetting. *Journal of Management Inquiry* 20, 3 (2011), 311–316.

[59] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[60] Fatima El-Tayeb. 2020. The Universal Museum: How the New Germany Built its Future on Colonial Amnesia. *Nka: Journal of Contemporary African Art* 2020, 46 (2020), 72–82.

[61] Yrjö Engeström, Katherine Brown, Ritva Engeström, and Kirsi Koistinen. 1990. Organizational forgetting: An activity-theoretical perspective. (1990).

[62] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.

[63] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2952–2963.

[64] Melanie Feinberg. 2017. Material Vision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 604–617.

[65] Melanie Feinberg, Daniel Carter, and Julia Bullard. 2014. A story without end: writing the residual into descriptive infrastructure. In *Proceedings of the 2014 conference on Designing interactive systems*. 385–394.

[66] Carolyn Forché et al. 1993. *Against forgetting: Twentieth-century poetry of witness*. WW Norton New York.

[67] Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

[68] Suzanne Fournier and Ernie Crey. 1997. *Stolen From Our Embrace: The Abduction of First Nations Children and the Restoration of Aboriginal Communities*. ERIC.

[69] Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

[70] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

[71] Mikhaila Friske, Jordan Wirfs-Brock, and Laura Devendorf. 2020. Entangling the Roles of Maker and Interpreter in Interpersonal Data Narratives: Explorations in Yarn and Sound. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 297–310.

[72] Rolando Garcia, Vikram Sreekanti, Neeraja Yadwadkar, Daniel Crankshaw, Joseph E Gonzalez, and Joseph M Hellerstein. 2018. Context: The missing piece in the machine learning lifecycle. In *KDD CMI Workshop*, Vol. 114.

[73] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[74] Lisa Gitelman. 2013. *Raw Data is an Oxymoron*. MIT Press.

[75] Boris Glavic, Klaus R Dittrich, A Kemper, H Schöning, T Rose, M Jarke, T Seidl, C Quix, and C Brochhaus. 2007. Data provenance: A Cctegorization of existing approaches. *BTW'07: Datenbanksysteme in Buisness, Technologie und Web* 103 (2007), 227–241.

[76] Erving Goffman et al. 1978. *The presentation of self in everyday life*. Vol. 21. Harmondsworth London.

[77] Agnes Grant. 1996. *No End of Grief: Indian Residential Schools in Canada*. ERIC.

[78] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.

[79] Derek Gregory. 2004. *"The colonial present"*. Oxford.

[80] Joel Grus. 2019. *Data science from scratch: first principles with python*. O'Reilly Media.

[81] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. 2011. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 65–74.

[82] Christine A Halverson and Mark S Ackerman. 2003. Yeah, the Rush ain't here yet-Take a break: Creation and use of an artifact as organizational memory. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 10–pp.

[83] John G Hansen and Emeka E Dim. 2019. Canada's missing and murdered indigenous people and the imperative for a more inclusive perspective. *International Indigenous Policy Journal* 10, 1 (2019).

[84] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[85] Sandra Harding. 1992. Rethinking standpoint epistemology: What is" strong objectivity?". *The Centennial Review* 36, 3 (1992), 437–470.

[86] Sandra Harding. 2007. Feminist standpoints. *Handbook of feminist research: Theory and praxis* (2007), 45–69.

[87] Sandra G Harding. 2004. *The feminist standpoint theory reader: Intellectual and political controversies.* Psychology Press.

[88] Jim Harris. 2020. [Online]. "Data Silence". OCDQ Blog. http://www.ocdqblog.com/home/data-silence.html

[89] Elliott Hauser. 2020. [Online]. "Logical force in the giving and taking of data". Interrogating Data Science workshop at CHI 2020 (title only). https://sites.google.com/view/interrogating-data-science-wk/home

[90] David G Hendry, Batya Friedman, and Stephanie Ballard. 2021. Value sensitive design as a formative framework. *Ethics and Information Technology* 23, 1 (2021), 39–44.

[91] Andrew Hertzberg, Jose Maria Liberti, and Daniel Paravisini. 2010. Information and incentives inside the firm: Evidence from loan officer rotation. *The Journal of Finance* 65, 3 (2010), 795–828.

[92] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery.* Vol. 1. Microsoft research Redmond, WA.

[93] Jeremy Heyer, Zachary Schmitt, Lynn Dombrowski, and Svetlana Yarosh. 2020. Opportunities for enhancing access and efficacy of peer sponsorship in substance use disorder recovery. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

[94] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.

[95] Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. 2020. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society.* 1–12.

[96] Tristan Hopper. 2021. The graves were never a secret: Why so many residential school cemeteries remain unmarked. *National Post* (2 June 2021).

[97] Jessica Hullman and Andrew Gelman. 2021. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review* (2021).

[98] Jane Hunter, Abdulmonem Alabri, and Catharine van Ingen. 2013. Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience* 25, 4 (2013), 454–466.

[99] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 560–575.

[100] Lilly Irani. 2015. The cultural work of microwork. *New Media & Society* 17, 5 (2015), 720–739.

[101] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 611–620.

[102] Julie Ireton. 2021. *Residential school records once held in Canada now in Rome, researchers say (CBC 15 Nov 2021).* https://www.cbc.ca/news/canada/ottawa/residential-school-records-now-in-rome-researchers-survivors-concerned-1.6241449 Accessed 5 Jan 2022.

[103] Devon S Isaacs and Amanda R Young. 2019. Missing and murdered Indigenous women (MMIW): Bringing awareness through the power of student activism. *Journal of Indigenous Research* 7, 1 (2019), 2.

[104] Tsunenori Ishioka. 2014. Investigations into Missing Values Imputation Using Random Forests for Semi-supervised Data. In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services.* 296–301.

[105] A Javadpour and S Samiei. 2017. Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Management Science Letters* 7, 2 (2017), 81–86.

[106] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[107] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank Van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (2011), 271–288.

[108] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 3363–3372.

[109] Shalini Kantayya. 2020. *Coded Bias.*

[110] Grigoris Karvounarakis, Zachary G Ives, and Val Tannen. 2010. Querying data provenance. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.* 951–962.

[111] Gunay Kazimzade and Milagros Miceli. 2020. Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-oriented Data Annotation Practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 71–71.

[112] Mary Beth Kery, Amber Horvath, and Brad A Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists.. In *CHI*, Vol. 10. 3025453–3025626.

[113] Mary Beth Kery, Bonnie E John, Patrick O'Flaherty, Amber Horvath, and Brad A Myers. 2019. Towards effective foraging by data scientists to find past analysis choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–13.

[114] Mary Beth Kery and Brad A Myers. 2017. Exploring exploratory programming. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC).* IEEE, 25–29.

[115] Martin Luther King Jr. 1967. Conscience for Change: Massey Lectures, Seventh Series. *Toronto: Canadian Broadcasting Corporation I* 967 (1967).

[116] Ákos Kiss and Tamás Szirányi. 2013. Evaluation of manually created ground truth for multi-view people localization. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*. 1–6.

[117] Edward Kissam. 2019. How low response among Latino immigrants will lead to differential undercount if the United States' 2020 census includes a question on sensitive citizenship. *Statistical Journal of the IAOS* 35, 2 (2019), 221–243.

[118] Atay Kizilaslan and Aziz A Lookman. 2017. Can Economically Intuitive Factors Improve Ability of Proprietary Algorithms to Predict Defaults of Peer-to-Peer Loans? *Available at SSRN 2987613* (2017).

[119] Donald Ervin Knuth. 1984. Literate programming. *The computer journal* 27, 2 (1984), 97–111.

[120] Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.)*. Addison Wesley Longman Publishing Co., Inc.

[121] Marina Kogan, Aaron Halfaker, Shion Guha, Cecilia Aragon, Michael Muller, and Stuart Geiger. 2020. Mapping Out Human-Centered Data Science: Methods, Approaches, and Best Practices. In *Companion of the 2020 ACM International Conference on Supporting Group Work*. 151–156.

[122] Christoph Kollwitz, Mximilian Perez Mengual, and Barbara Dinter. 2018. Cross-Disciplinary Collaboration for Designing Data-Driven Products and Services. In *2018 Pre-ICIS SIGDSA Symposium on Decision Analytics Connecting People, Data & Things*. 1–15.

[123] Robert Kozma, Roman Ilin, and Hava T Siegelmann. 2018. Evolution of abstraction across layers in deep learning neural networks. *Procedia computer science* 144 (2018), 203–213.

[124] Scott Krig. 2016. Ground truth data, content, metrics, and analysis. In *Computer Vision Metrics*. Springer, 247–271.

[125] Sean Kross and Philip J Guo. 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[126] Sean Kross and Philip J Guo. 2021. Orienting, Framing, Bridging, Magic, and Counseling: How Data Scientists Navigate the Outer Loop of Client Collaborations in Industry and Academia. *arXiv preprint arXiv:2105.05849* (2021).

[127] Kateryna Kuksenok, Cecilia Aragon, James Fogarty, Charlotte P Lee, and Gina Neff. 2017. Deliberate Individual Change Framework for Understanding Programming Practices in four Oceanography Groups. *Computer Supported Cooperative Work (CSCW)* 26, 4 (2017), 663–691.

[128] Neha Kumar and Naveena Karusala. 2021. Braving citational justice in human-computer interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–9.

[129] Milan Kundera and Michael Henry (trans.) Heim. 1981. The Book of Laughter and Forgetting. 1979. (1981).

[130] Xiaochen Lai, Liwen Mao, Zheng Zhang, and Genglin Zhang. 2021. Multi-task learning modeling of attribute mutual association based on confidence and imputation of missing values. In *2021 2nd International Conference on Computing, Networks and Internet of Things (CNIOT 2021)*. 1–6.

[131] Han Lamers, Toon Van Hal, and Sebastiaan G Clercx. 2020. How to Deal with Scholarly Forgetting in the History of the Humanities: Starting points for discussion. *History of Humanities* 5, 1 (2020), 5–29.

[132] Larry Laudan. 1978. *Progress and its problems: Towards a theory of scientific growth*. Vol. 282. Univ of California Press.

[133] Debora de Castro Leal, Angelika Strohmayer, and Max Krüger. 2021. On Activism and Academia: Reflecting Together and Sharing Experiences Among Critical Friends. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[134] Dokyun Lee, Emaad Manzoor, and Zhaoqi Cheng. 2018. Focused concept miner (FCM): Interpretable deep learning for text exploration. *Available at SSRN 3304756* (2018).

[135] Susan Leigh Star. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, Technology, & Human Values* 35, 5 (2010), 601–617.

[136] Eric Lesser, Michael Fontaine, and Jason Slusher. 2009. *Knowledge and communities*. Routledge.

[137] Antonios Liapis, Gillian Smith, and Noor Shaker. 2016. Mixed-initiative content creation. In *Procedural content generation in games*. Springer, 195–214.

[138] Sijia Liu, Parikshit Ram, Deepak Vijaykeerthy, Djallel Bouneffouf, Gregory Bramble, Horst Samulowitz, Dakuo Wang, Andrew Conn, and Alexander Gray. 2020. An ADMM based framework for automl pipeline configuration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4892–4899.

[139] Helen E Longino. 1993. Feminist standpoint theory and the problems of knowledge.

[140] Alexandra Lyn. 2020. Risky Business: Artificial Intelligence and Risk Assessments in Sentencing and Bail Procedures in the United States. *Available at SSRN 3831441* (2020).

[141] Jean-François Lyotard. 1984. *The postmodern condition: A report on knowledge*. U of Minnesota Press.

[142] Diana Lynn MacLean and Jeffrey Heer. 2013. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the american medical informatics association* 20, 6 (2013), 1120–1127.

[143] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[144] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572* (2020).

[145] Helena M Mentis, Ahmed Rahim, and Pierre Theodore. 2016. Crafting the image in surgical telemedicine. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 744–755.

[146] Edmund Metatawabin and Alexandra Shimo. 2015. *Up Ghost River: A Chief's Journey Through the Turbulent Waters of Native History*. Vintage Canada.

[147] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[148] Barbara J Mills. 2008. Remembering while forgetting: depositional practices and social memory at Chaco. *Memory work: Archaeologies of material practices* (2008), 81–108.

[149] Cindy Minarova-Banjac. 2018. Collective Memory and Forgetting: A Theoretical. (2018).

[150] Hrushikesha Mohanty. 2015. Big data: An introduction. In *Big Data*. Springer, 1–28.

[151] Naja Holten Møller, Claus Bossen, Kathleen H Pine, Trine Rask Nielsen, and Gina Neff. 2020. Who does the work of data? *Interactions* 27, 3 (2020), 52–55.

[152] Arnaldo Momigliano, Glen Warren Bowersock, Tim Cornell, Glen Warren Bowersock, and Tim Cornell. 1994. *Studies on modern scholarship*. University of California Press Berkeley.

[153] Lenard Monkman. 2022. *'There's got to be ways' Indigenous language names can be registered in Manitoba, says SCO grand chief (CBC 18 Feb 2022)*. https://www.cbc.ca/news/indigenous/manitoba-indigenous-language-baby-names-1.6357478 Accessed 19 Feb 2022.

[154] Christine Moorman and Anne S Miner. 1998. Organizational improvisation and organizational memory. *Academy of management Review* 23, 4 (1998), 698–723.

[155] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating Data Science. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 467–473.

[156] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 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*. 1–15.

[157] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[158] Dylan Mulvin. 2021. *Proxies: The cultural work of standing in*. MIT Press.

[159] Gina Neff, Anissa Tanweer, Brittany Fiore-Gartland, and Laura Osburn. 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data* 5, 2 (2017), 85–97.

[160] Katherine L Nichols. 2020. 3 The Brandon Indian Residential School Cemetery Project. *Working with and for Ancestors: Collaboration in the Care and Study of Ancestral Remains* (2020), 43.

[161] Pádraig Ó Tuama. 2019. Makebelieve. Retrieved August 12, 2021 from https://poets.org/poem/makebelieve

[162] Anne Oeldorf-Hirsch and Darren Gergle. 2020. 'Who Knows What' Audience Targeting for Question Asking on Facebook. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–20.

[163] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[164] Mimi Onuoha. 2016. "The Library of Missing Datasets [Online]". https://mimionuoha.com/the-library-of-missing-datasets

[165] George Orwell. 2021. *Nineteen eighty-four*. Oxford University Press.

[166] Harold Ossher and Peri Tarr. 2001. Using multidimensional separation of concerns to (re) shape evolving software. *Commun. ACM* 44, 10 (2001), 43–50.

[167] Sinan Ozdemir and Divya Susarla. 2018. *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing Ltd.

[168] Fereniki Panagopoulou-Koutnatzi. 2012. The right to be forgotten in the digital era. In *Proceedings of Interntional Conference on Information Law ICIL 2012*. 305–322.

[169] Samir Passi. 2021. *Making Data Work: The Human and Organizational Lifeworlds of Data Science*. Cornell University.

[170] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[171] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2436–2447.

[172] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[173] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 667–676.

[174] Deana Pennington, Imme Ebert-Uphoff, Natalie Freed, Jo Martin, and Suzanne A Pierce. 2020. Bridging sustainability science, earth science, and data science through interdisciplinary education. *Sustainability Science* 15, 2 (2020), 647–661.

[175] Kathleen Pine and Claus Bossen. 2020. [Online]. "Multiple Virtues in Data Work: CDIS' Practices of Generating Healthcare Data". Interrogating Data Science workshop at CHI 2020 (title only). https://sites.google.com/view/interrogating-data-science-wk/home

[176] Kathleen H Pine and Max Liboiron. 2015. The politics of measurement and action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3147–3156.

[177] Liedeke Plate. 2016. Amnesiology: Towards the study of cultural oblivion. *Memory Studies* 9, 2 (2016), 143–155.

[178] Ivens Portugal, Paulo Alencar, and Donald Cowan. 2018. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications* 97 (2018), 205–227.

[179] Line Pouchard. 2015. Revisiting the data lifecycle with big data curation. (2015).

[180] Robert N Proctor and Londa Schiebinger. 2008. Agnotology: The making and unmaking of ignorance. (2008).

[181] Paloma Rabaey, Matis De Schutter, Toto De Brant, and Matthias Derudder. [n.d.]. Fairness in Data Science. ([n. d.]).

[182] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[183] Tye Rattenbury, Joseph M Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. 2017. *Principles of data wrangling: Practical techniques for data preparation*. " O'Reilly Media, Inc.".

[184] Paulette Regan. 2010. *Unsettling the settler within: Indian residential schools, truth telling, and reconciliation in Canada*. ubc Press.

[185] Kabul Resident. [n.d.]. *An Afghan woman in Kabul: 'Now I have to burn everything I achieved', The Guardian 15 Aug, 2021*. Retrieved Aug 25, 2021 from https://www.theguardian.com/world/2021/aug/15/an-afghan-woman-in-kabul-now-i-have-to-burn-everything-i-achieved

[186] Margaret Robinson. 2017. Two-spirit and bisexual people: Different umbrella, same rain. *Journal of Bisexuality* 17, 1 (2017), 7–29.

[187] Marina Roseman. 2003. *Singers of the Landscape: Song, History, and PropertyR ights in the Malaysian Rainforest*. Duke University Press.

[188] Sabirat Rubya and Svetlana Yarosh. 2017. Interpretations of online anonymity in Alcoholics Anonymous and Narcotics Anonymous. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[189] Adam Rule, Aurélien Tabard, and James D Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[190] Mihai Stelian Rusu et al. 2011. The Colonization of the Past and the Construction of Mnemonic Order. *Studia Universitatis Babes-Bolyai-Sociologia* 56, 2 (2011), 39–57.

[191] James Sadd, Rachel Morello-Frosch, Manuel Pastor, Martha Matsuoka, Michele Prichard, and Vanessa Carter. 2014. The truth, the whole truth, and nothing but the ground-truth: Methods to advance environmental justice and researcher–community partnerships. *Health education & behavior* 41, 3 (2014), 281–290.

[192] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[193] Viviane Santos, Alfredo Goldman, Eduardo Guerra, Cleidson De Souza, and Helen Sharp. 2013. A pattern language for inter-team knowledge sharing in agile software development. In *20th Conference on Pattern Languages of Programs, Monticello, IL*.

[194] Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. 2020. Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[195] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[196] Joni Seager. 2016. "Missing women, blank maps, and data fields: What gets counted counts", Boston Public Library Lectures in Cartography.

[197] Cathrine Seidelin, Yvonne Dittrich, and Erik Grönvall. 2018. Data Work in a Knowledge-Broker Organisation: How Cross-Organisational Data Maintenance Shapes Human Data Interactions. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference* (Belfast, United Kingdom) *(HCI '18)*. BCS Learning & Development Ltd., Swindon, GBR, Article 14, 12 pages. https://doi.org/10.14236/ewic/HCI2018.14

[198] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[199] Dilruba Showcat and Eric P.S. Baumer. 2020. [Online]. "Outliers: More Than Numbers?". Interrogating Data Science workshop at CHI 2020. https://www.researchgate.net/profile/Dilruba-Showkat/publication/345774805_Outliers_More_Than_Numbers/links/5fad80eda6fdcc9389b15a15/Outliers-More-Than-Numbers.pdf

[200] Miguel-Ángel Sicilia, Elena García-Barriocanal, Salvador Sánchez-Alonso, Marçal Mora-Cantallops, and Juan-José Cuadrado. 2018. Ontologies for data science: On its application to data pipelines. In *Research Conference on Metadata and Semantics Research*. Springer, 169–180.

[201] Yogesh L Simmhan, Beth Plale, and Dennis Gannon. 2005. A survey of data provenance in e-science. *ACM Sigmod Record* 34, 3 (2005), 31–36.

[202] Eric Simons, Andrew Martindale, and Alison Wylie. 2020. Bearing witness: What can archaeology contribute in an Indian Residential School context? In *Working with and for Ancestors*. Routledge, 21–31.

[203] Jesper Simonsen, Helena Karasti, and Morten Hertzum. 2020. Infrastructuring and participatory design: Exploring infrastructural inversion as analytic, empirical and generative. *Computer Supported Cooperative Work (CSCW)* 29, 1 (2020), 115–151.

[204] Lucy Simpson and Cherrah Giles (Eds.). 2022. Special edition on Missing and Murdered Indigenous Women. *Restoration* 18, 4 (2022).

[205] Stolen Sisters. 2004. A Human Rights Response to Discrimination and Violence against Indigenous Women in Canada. *Amnesty International* (2004).

[206] Katta Spiel. 2021. " Why are they all obsessed with Gender?"—(Non) binary Navigations through Technological Infrastructures. In *Designing Interactive Systems Conference 2021*. 478–494.

[207] Katta Spiel, Julia Makhaeva, and Christopher Frauenberger. 2016. Embodied companion technologies for autistic children. In *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. 245–252.

[208] Susan Leigh Star and Anselm Strauss. 1999. Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work (CSCW)* 8, 1-2 (1999), 9–30.

[209] Chris Stokel-Walker. [n.d.]. *Afghans are racing to erase their online lives: Every photo and every data point is a link to the old way of life in Afghanistan – and a reason for Taliban retribution, Wired 17 Aug, 2021.* Retrieved Aug 25, 2021 from https://www.theguardian.com/world/2021/aug/15/an-afghan-woman-in-kabul-now-i-have-to-burn-everything-i-achieved,Aug17,2021

[210] Angelika Strohmayer, Jenn Clamen, and Mary Laing. 2019. Technologies for social justice: Lessons from sex workers on the front lines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[211] Angelika Strohmayer, Mary Laing, and Rob Comber. 2017. Technologies and social justice outcomes in sex work charities: Fighting stigma, saving lives. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3352–3364.

[212] Angelika Strohmayer, Janis Lena Meissner, Alexander Wilson, Sarah Charlton, and Laura McIntyre. 2020. " We come together as one... and hope for solidarity to live on" On Designing Technologies for Activism and the Commemoration of Lost Lives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 87–100.

[213] Ki Hyun Tae, Yuji Roh, Young Hun Oh, Hyunsu Kim, and Steven Euijong Whang. 2019. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*. 1–4.

[214] Anissa Tanweer, Brittany Fiore-Gartland, and Cecilia Aragon. 2016. Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society* 19, 6 (2016), 736–752.

[215] Lisa Tatonetti. 2021. *Written by the Body: Gender Expansiveness and Indigenous Non-Cis Masculinities.* U of Minnesota Press.

[216] Alex S Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-place: Thinking through the relations between data and community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2863–2872.

[217] Kristinn R Thórisson, Jordi Bieger, Xiang Li, and Pei Wang. 2019. Cumulative learning. In *International Conference on Artificial General Intelligence*. Springer, 198–208.

[218] Truth and Reconciliation Commission. 2015. Truth & Reconciliation Commission of Canada – Findings & Reports. https://archive.org/stream/TruthReconciliationCanada/Volume_4_Missing_Children_English_Web_djvu.txt.

[219] Muhammad Fahim Uddin, Jeongkyu Lee, Syed Rizvi, and Samir Hamada. 2018. Proposing enhanced feature engineering and a selection model for machine learning processes. *Applied Sciences* 8, 4 (2018), 646.

[220] Wil MP Van der Aalst. 2014. Data scientist: The engineer of the future. In *Enterprise interoperability VI*. Springer, 13–26.

[221] Morgan Vigil-Hayes, Marisa Duarte, Nicholet Deschine Parkhurst, and Elizabeth Belding. 2017. # indigenous: Tracking the connective actions of native american advocates on twitter. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1387–1399.

[222] Morgan Vigil-Hayes, Nicholet Deschine Parkhurst, and Marisa Duarte. 2019. Complex, contemporary, and unconventional: Characterizing the tweets of the# NativeVote movement and Native American candidates through the 2018 US midterm elections. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[223] Manu Vimalassery, Juliana Hu Pegues, and Alyosha Goldstein. 2017. Colonial unknowing and relations of study. *Theory & Event* 20, 4 (2017), 1042–1054.

[224] João Vinagre and Alípio Mário Jorge. 2012. Forgetting mechanisms for scalable collaborative filtering. *Journal of the Brazilian Computer Society* 18, 4 (2012), 271–282.

[225] Vered Vinitzky-Seroussi and Chana Teeger. 2010. Unpacking the unspoken: Silence in collective memory and forgetting. *Social forces* 88, 3 (2010), 1103–1122.

[226] Dakuo Wang, Josh Andres, Justin D Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[227] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[228] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.

[229] Frederick Whitling. 2010. Damnatio memoriae and the power of remembrance. *A European memory* (2010).

[230] Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. 2021. Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities. In *Proceedings of the 2021 International Conference on Management of Data*. 2639–2652.

[231] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions* 26, 4 (2019), 42–46.

[232] Matthew Yapchain. 2018. Human-Centered Data Science: A New Paradigm for Industrial IoT. In *Ethnographic Praxis in Industry Conference Proceedings*, Vol. 2018. Wiley Online Library, 53–61.

[233] Carlos Vladimiro González Zelaya. 2019. Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE, 2086–2090.

[234] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[235] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs.

[236] Jean-Daniel Zucker. 2003. A grounded theory of abstraction in artificial intelligence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358, 1435 (2003), 1293–1309.