Data Silences: How to Unsilence the Uncertainties in Data Science

Michael Muller*

Abstract

When we wrangle the data in data science, we design the data to make it fit-for-analysis. Wrangling involves the removal or reduction of uncertainties, such as outliers, missing values, mal-distributions, and the details of feature engineering. Many of the steps of data wrangling go unrecorded or poorly recorded, in terms of both *what was done* and also the rationale for *why it was done.* In this way, we impose multiple types of *data silences* on the data, and often on the sources (people) who are "behind" the data. In this paper, we articulate how we may perform multiple types of silencing. We challenge comfortable conceptions of the nature of data, and we call on the data-science community to devise and adopt methodologies to unsilence data.

1 What is the Blue Sky Idea?

As Bowker observed, " 'Raw data' is both an oxymoron and a bad idea" ([6]; see also [12]). If most data are imperfect [29], then we need to *design* those data to be fit-for-purpose ([10, 23, 43]). This kind of data wrangling [32] - also called *data work* [25, 27] - receives less scholarly attention than work on data science models [33]. By contrast, Tae et al. envision analytic approaches "where data becomes a first-class citizen, on a par with code" ([38]; see also [44]). Inattention to data work can lead to important problems in data quality, potentially including bias and other distortions [3], and also improvisational rather than disciplined approaches to core data practices such as labeling and annotating ground truth [26].

This **Blue Sky Idea** is to understand how both data and rationales are forgotten or even obfuscated, and how to improve our data practices so as to preserve our knowledge of both what was done to data, and why, and to inform recent work in data-centric AI [44].

2 Why is it a Blue Sky Idea? Why should the community ponder over it? Why now?

As we noted above, the design of data and the problems of data work have been under-analyzed in the research literature (e.g., [33]). To make our data fit-for-purpose, we tend to prepare our data through a layered series of analyses and repairs, including steps such as

- discovering and/or capturing the data [24]
- cleaning the data [32]
- transforming data distributions [24]
- defining outliers [19, 35]
- removing some outliers [1]
- imputing values in place of some outliers and of missing values [18, 36]
- engineering features [20, 42]
- creating new data in the form of labels [17]

and then finally the steps of designing, training, and evaluating the model.

Zhang et al. analyzed data science work as a series of activities performed by data science workers with different skills and different job responsibilities [45]. Wang et al. showed that these data science workers tend to focus only on their own steps in a data science pipeline [41], with little attention to how the data came to be in their current state. Thereby, there is a "forgettance stack" of layered data science activities [25], in which the work in each layer hides any problems [3] or imperfections [29] of the lower layers. Within the current work practices of data science, we forget what we have done to and with the data.

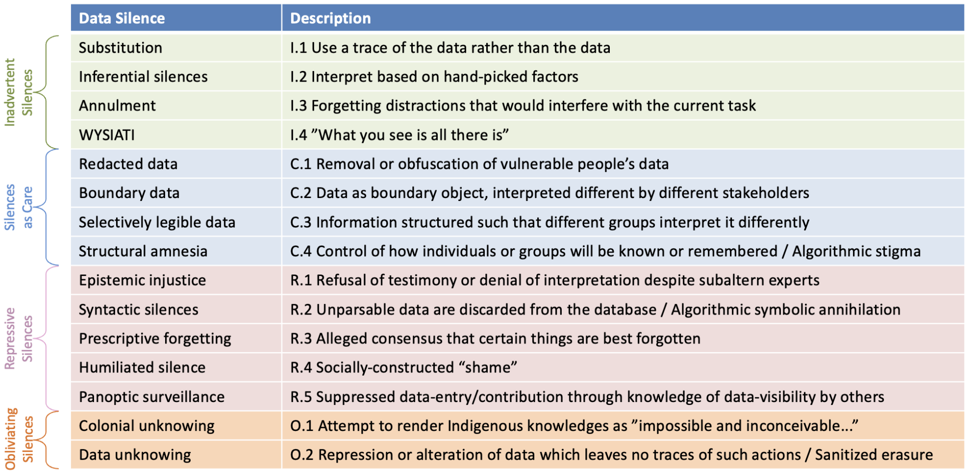
These problems are compounded when we fail to record *how* we design our data [25]. The design of data is complex and errorprone [3, 23], requiring data science knowledge and domain knowledge [10, 22]. Aspects of designing data include knowledge of both *how* to design, and also *why* to design the data in certain ways - i.e., a design rationale of the data in the dataset [26, 40].

Increasingly, data science applications are used by institutions and governments to determine and govern the lives of people in life situations such as access to finances, [16], healthcare [14], carceral systems [7], and the removal of children from their families [34]. There is thereby an urgent need to understand how data come to be, how they are constituted, and why, so as to reduce harms from the operation of data science applications.

3 Does the Blue Sky Idea challenge our current set of assumptions or does it take a bold approach to solve a wicked problem?

We addresss this urgent need through Onuoha's concept of *data silences*, which she described as "blank spots

*IBM Research, michael_muller@us.ibm.com



	Data Silence	Description
Inadvertent Silences	Substitution	I.1 Use a trace of the data rather than the data
	Inferential silences	I.2 Interpret based on hand-picked factors
	Annulment	I.3 Forgetting distractions that would interfere with the current task
	WYSIATI	I.4 "What you see is all there is"
Silences as Care	Redacted data	C.1 Removal or obfuscation of vulnerable people's data
	Boundary data	C.2 Data as boundary object, interpreted different by different stakeholders
	Selectively legible data	C.3 Information structured such that different groups interpret it differently
	Structural amnesia	C.4 Control of how individuals or groups will be known or remembered / Algorithmic stigma
Repressive Silences	Epistemic injustice	R.1 Refusal of testimony or denial of interpretation despite subaltern experts
	Syntactic silences	R.2 Unparsable data are discarded from the database / Algorithmic symbolic annihilation
	Prescriptive forgetting	R.3 Alleged consensus that certain things are best forgotten
	Humiliated silence	R.4 Socially-constructed "shame"
	Panoptic surveillance	R.5 Suppressed data-entry/contribution through knowledge of data-visibility by others
Obliviating Silences	Colonial unknowing	O.1 Attempt to render Indigenous knowledges as "impossible and inconceivable..."
	Data unknowing	O.2 Repression or alteration of data which leaves no traces of such actions / Sanitized erasure

Figure 1: Data Silences: "[B]lank spots that exist in spaces that are otherwise data-saturated" [28]. *Inadvertent Silences* (I.1-1.4) may occur without specific motivation. *Silences as Care* (C.1-C.4) are intended to protect people or data. *Repressive Silences* (R.1-R.5) are used to suppress persons, Nations and their information. *Obliviating Silences* (O.1-O.2) are used to deny the existence of data, and sometimes of persons, groups, or Nations.

that exist in spaces that are otherwise data-saturated" [28]. As we noted above, the day-to-day activities of data science are layered, one above another, in a forgettance stack [25]. Each activity assumes perfect data from the preceding activity that is "below" it in the stack. This assumption creates a series of silences, beginning with the selection of the dataset [30] and culminating in the dataset that is ready for analysis [33]. We silence our own data.

Muller and Strohmayer described multiple types of data silences [25]. Figure 1 is a further development of that analysis. We describe them briefly here.

3.1 Inadvertent Silences. The first category of *Inadvertent Silences* occur without apparent motivation, in the course of preparing data for analysis, such as *(I.1) Substituting* a trace or summary of data for the phenomena themselves. Most real-world phenomena are complex, and data science workers use *(I.2) Inferential Silences* to highlight the attributes that they think are most relevant to the problem they are solving. The highlighted data can be made more salient by *(I.3) Annulment*, in which less relevant attributes are removed or forgotten. Together, these human actions shape or design the data, leading to the phenomenon of *(I.4) WYSIATI*, "What you see is all there is" [13].

These considerations challenge our idea of data as "objective", "immutable", and "given" by the nature of reality (e.g, as critiqued in [5, 9]). The data are, by now, quite removed from their original condition. When we prepare data for analysis, we shape and design the data. However, there are more types of silences to consider.

3.2 Silence as Care. In some cases, researchers or data "subjects" act to *(C.1) Redact Data* as a form of care, to protect vulnerable persons, populations, or Nations. One strategy is to present data in a way that is interpreted differently by different parties (e.g., as a *(C.2) Boundary Object* [37]). A classic example is the "Drinking Gourd" song, which was simultaneously perceived by enslavers as a simple work song, and by enslaved peoples as a set of travel directions about how to escape to freedom [15] - a form of *(C.3) Selective Legibility*. Bellini et al. described updated forms of *Redacted* and *Boundary Silences* in their work with survivors of domestic abuse [4]. Other researchers have described *(C.4) Structural Amnesia* - i.e., the selective deletion of data to protect people from harmful or dangerous labels, thus reducing the potential for *Algorithmic Stigma* [2].

These considerations further challenge our idea of data as a set of dispassionate objects. We can, instead, understand data as an opportunity for care, and as a form of social relation.

3.3 Repressive Silences. Repressive Silences present more disturbing dynamics. *(R.1) Epistemic Injustice* may become a refusal to acknowledge or undderstand the perspectives of others, and thus to deny their reality. This form of oppression may be enacted through *(R.2) Syntactic Silences*, in which the dataset is constructed to disallow certain forms of expression (e.g., a validation rule that enforces binary gender data, rather than accommodating diverse gender expressions [9]). The imposition of these constraints may be argued in terms of *(R.3) Prescriptive Forgetting*, which can be a (false) consensus on what is to be remembered, and what should be forgotten, leading to *(R.4) Humiliated Silencing* of people who do not conform. Finally, people's expression of their identity and/or their perspectives may made hazardous, so as to impose a *(R.5) Panoptic Silence* in which constant surveillance suppresses speech, action, and thought [11].

These considerations further challenge our idea of data as neutral. We can, instead, understand data as a means to exercise of power.

3.4 Obliviating Silences. Finally, *Obliviating Silences* may impose ignorance [31] of data that describe harms, crimes, or inconvenient knowledge. *(O.1) Colonial Unknowing* is a broad category of acts to hide evidence of acts of colonial powers upon lands and peoples whom they have conquered. A contemporary example of *Colonial Unknowing* is the on-going crisis of the so-called ”residential schools” in former British colonies, in which hundreds of thousands of Indigenous children (the Stolen Generation [8]) were legally abducted from their parents and sent to boarding schools, where they were subjected to harsh conditions intended to subjegate them into servants of colonial powers [21, 39]. Evidence of these crimes was then suppressed - i.e., the data became “unknown” because certain parties wanted to prevent others from knowing about them. The unintended but systematic hiding of data uncertainties in the “forgettance stack” described above, may constitute a similar form of *(O.2) Data Unknowing*.

These considerations further challenge our idea of data as neutral. We can, instead, understand data as a means to erase persons, groups, or Indigenous Peoples.

4 What are the challenges?

Simply put, the challenges are to remember not only the finished form of our data, but also the steps to design the data, and the uncertainties and rationales of the data science workers who modify and create the data.

5 What will success look like?

In *Memory Practices in the Sciences*, Bowker described human work in many scientific fields in terms of what we remember, how we remember, how we re-find what we once knew, and whom we remember with [6]. Based on this analysis, success will involve a series of sociotechnical reforms of both data practices and the technologies that support those practices, extending data-centric AI [44] with more human-informed practices. The current state of the art preserves only the *current* form of a dataset, often without provenance or change-record, and with no *social* record of who modified the dataset. We anticipate that dataframes and other representations and repositories will be enhanced with

- versioning features to recover prior configurations of the data;
- annotation features, through which data science workers can record metadata tuplets to describe how they modified (wrangled, designed) the data in terms of *who* performed each action, *when* they performed it, *what* they did, and *why* they did it, and what uncertainties remain; and
- social features to enable searching and pooling of knowledge across the steps of data wrangling.

In these ways, we can begin to recognize data as a responsible entity that can be held accountable for meeting human and cultural needs. We call on the data science community to develop the details of such methodologies, and also invent new approaches that can unsilence our silenced data.

References

[1] C.C. Aggarwal and P.S. Yu, *Outlier Detection with Uncertain Data*, Proc. 2008 SDM.

[2] N. Andalibi, C. Pyle, K. Barta, L. Xuan, A.Z. Jacobs, and M.S. Ackerman, *Conceptualizing algorithmic stigma.* Proc. CHI 2023, Art 373 (2023).

[3] C. Aragon, S. Guha, S. Kogan, M. Muller, and G. Neff, *Human centered data science: An introduction.* MIT Press (2022).

[4] R. Bellini, A. Strohmayer, Patrick, O., and C. Crivellaro, *Mapping the margins: Navigating the ecologies of domestic violence service provision*, Proc. CHI 2019, 1-13.

[5] E.M. Bender, T. Gebru, A. McMillan-Major, and M. Mitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proc. FAT* 2021, 610-623 (2021).

[6] G.C. Bowker, *Memory practices in the sciences.* MIT Press (2008).

[7] M. Brennan, J.S. Gersen, M. Haley, M. Lin, A. Merchant, R.J. Millett, S.K. Sarkar, and D. Wegner, *Constitutional Dimensions of Predictive Algorithms in Criminal Justice*, Harv. CR-CLL Rev. 55, 267 (2020).

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

[9] C. D'Ignazio and L.F. Klein, *Data Feminism*, MIT Press (2020).

[10] M. Feinberg, *Everyday Adventures with Unruly Data* MIT Press (2022).

[11] M. Galič, T. Timan, and B.-J. Koops, *Bentham, Deleuze and Beyond: An Overview of Surveillance Theories from the Panopticon to Participation*, Phil. & Tech. 30, 9-37 (2017).

[12] L. Gitelman, *"Raw Data" Is an Oxymoron*, MIT Press (2013).

[13] J. Harris, *"Data silence"*, OCDQ Blog (2021).

[14] K.N. Keya, R. Islam, S. Pan, I. Stockwell, and J. Foulds, *Equitable Allocation of Healthcare Resources with Fair Survival Models*, Proc. 2021 SDM.

[15] M.L. King, *Conscience for Change: Massey Lectures, Seventh Series*, Canadian Broadcasting Corporation I (1967).

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

[17] X. Kong, Z. Wu, L.-J. Li, R. Zhang, P.S. Hu, H. Wu, and W. Fan, *Large-Scale Multi-Label Learning with Incomplete Label Assignments*, Proc. 2014 SDM (2014).

[18] R. Leibrandt and S. Günnemann, *Making Kernel Density Estimation Robust towards Missing Values in Highly Incomplete Multivariate Data without Imputation*, Proc. 2018 SDM.

[19] S. Lin and D.E. Brown, *An Outlier-based Data Association Method For Linking Criminal Incidents*, Proc. 2003 SDM.

[20] K. Liu, H. Huang, W. Zhang, A. Hariri, Y. Fu, and K. Hua, *Multi-Armed Bandit Based Feature Selection*, Proc. 2021 SDM.

[21] Maine Wabanaki-State Child Welfare Truth & Reconciliation Commission, *Beyond the Mandate: Continuing the Conversation: Report of the Maine Wabanaki-State Child Welfare Truth & Reconciliation Commission*, 2015.

[22] H.M. Mentis, A. Rahim, and P. Theodore, *Crafting the Image in Surgical Telemedicine*, Proc. CSCW 2016.

[23] M. Muller, L. Aroyo, M. Feinberg, H. Mentis, S. Passi, S. Guha, and H. Candello, *Designing data in data science.* FAccT CRAFT 2022, https://facctconference.org/2022/acceptedcraft.html.

[24] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q.V. Liao, C. Dugan, and T. Erickson, *How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation*, Proc. CHI 2019 Paper 126.

[25] M. Muller and A. Strohmayer, *Forgetting practices in the data sciences.* Proc. CHI 2022, Art. 323.

[26] M. Muller, C.T. Wolf, J. Andres, M. Desmond, N.N. Joshi, Z. Ashktorab, A. Sharma, K. Brimijoin, Q. Pan, E. Duesterwald, and C. Dugan, *Designing Ground Truth and the Social Life of Labels.* Proc. CHI 2021, Art. 94.

[27] N.H. Møller, C. Bossen, K.H. Pine, T.R. Nielsen, G. Neff, *Who does the work of data?*, Interactions 27(3), 52-55 (2020).

[28] M. Onuoha, *The Library of Missing Datasets*, https://mimionuoha.com/the-library-of-missing-datasets (2016).

[29] R.K. Pearson, *Mining imperfect data: With examples in R and Python*, SIAM (2020).

[30] K.H. Pine and M. Liboiron, *The Politics of Measurement and Action*, Proc. CHI 2015, 3147–3156 (2015).

[31] R. Proctor and L. Schiebinger, *Agnotology: The making and unmaking of ignorance.* Stanford University Press (2008).

[32] T. Rattenbury, J.M. Hellerstein, J. Heer, S. Kandel, and C. Carreras *Principles of Data Wrangling: Practical Techniques for Data Preparation*, O'Reilly Media (2017).

[33] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L.M. Aroyo, *"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI*, Proc. CHI 2021, Art. 39 (2021).

[34] D. Saxena, K. Badillo-Urquiola, P.J. Wisniewski, and S. Guha *A Human-Centered Review of Algorithms used within the U.S. Child Welfare System*, Proc. CHI 2020, 1-15.

[35] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, *On Evaluation of Outlier Rankings and Outlier Scores*, Proc. 2012 SDM.

[36] S.M. Shortreed, E.B. Laber, J. Pineau, and S.A. Murphy, *Imputing missing data from sequential multiple assignment randomized trials*, in M.R. Kosorok and E.E.M. Moodie (eds), Adaptive treatment strategies in practice. Proc. 2015 SDM.

[37] S.L. Star, *This is not a boundary object: Reflections on the origin of a concept*, Science, Technology, & Human Values 35(5), 601-617 (2010).

[38] K.H. Tae, Y. Roh, Y.H. Oh, H. Kim, and S.E. Whang, *Data cleaning for accurate, fair, and robust models: A big data-AI integration approach*, Proc. MOD 2019, 1-4.

[39] Truth and Reconciliation Commission, *Truth & Reconciliation Commission of Canada – Findings & Reports*, (2015).

[40] A. Vande Moere, A. and S. Patel, *The physical visualization of information: designing data sculptures in an educational context.* In M.L. Huang, Q.V. Nguyen Q.V., and K. Zhang (eds.), Visual Information Communication, 2010, pp. 1–23. Springer (2010).

[41] D. Wang, Q.V. Liao, Y. Zhang, U Khurana, H. Samuelowitz, S. Park, M. Muller, and L. Amini, *How Much Automation Does a Data Scientist Want?*, arXiv:2101.03970 (2021).

[42] M.Xiao, D. Wang, M. Wu, Z. Qiao, P. Wang, K. Liu, Y. Zhou, and Y. Fu, *Traceable Automatic Feature Transformation via Cascading Actor-Critic Agents*, Proc. 2023 SDM.

[43] H.D. Zajac, N.R. Avlona, F. Kensing, T.O. Anderson, and I. Shklovski, *Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI*, Proc. AIES 2023, 351-362.

[44] D. Zha, Z.P. Bhat, K.-H. Lai, F. Yang, and Xia Hu, *Data-centric AI: Perspectives and Challenges*, Proc. 2023 SDM.

[45] A.X. Zhang, M. Muller, and D. Wang, *How do Data Science Workers Collaborate? Roles, Workflows, and Tools*, Proc. CSCW 2020, Art. 22.