Recensioner

Katherine Harrison, *Behind the Science: The Invisible Work of Data Management in Big Science*, Bristol University Press 2025.

Behind the Science: The Invisible Work of Data Management in Big Science is a book about a neutron-scattering facility in Lund, and in particular their data management practices (as told by data scientists); or in Harrison's own words, a book "on the technical and organisational challenges involved in selecting, connecting, and aligning the software and hardware that must deliver the brand new 'valid' data" (p. 8). If data is so essential to the scientific outcomes of these facilities, why has data management remained largely invisible – both in scholarly discussions and within the facilities themselves? And why does this invisibility matter for the way knowledge is produced at Big Science institutions like the European Spallation Source (ESS)? These are the questions Harrison's study seeks to shed light on with the help of the two concepts: *invisible work* and the *black box*.

Invisible work is a term that encompasses the intricate processes of adaptation, translation, and alignment that users (at some level) undertake to make technological systems functional. Invisible work thereby represents the additional cognitive and practical efforts required to bridge potential gaps between technological design and actual usefulness, revealing how professionals and users reshape their contexts and activities to accommodate these new technologies. The concept is particularly significant because it illuminates the hidden labour that creates organisational value while remaining systematically unacknowledged. The concept of invisible work serves as a bridge between critical data studies and science and technology studies (STS). But why not include critical studies of algorithms as well? We find ourselves wondering: What is the difference between algorithm studies and the study of data processed by machines? Why focus solely on data management? What about other instantiations of data, such as visualisations, models, algorithms, or information architecture?

The black box (or "blackboxing") is a somewhat related metaphor, originally introduced by Bruno Latour in STS. Derived from cybernetic theory in the 1940s, the concept deals with processes where input and output are central, but where intermediate processes are less visible. To "open a black box" means to examine the hidden aspects that appear neutral and self-evident, but which, upon closer examination, turn out to contain a complex history of claims, policies, and positions. The more user-friendly an artefact appears to be, the more difficult it is to examine its development history,

https://doi.org/10.37062/sf.62.27833
Sociologisk Forskning, årgång 62, nr 1–2, s. 187–190.

which obscures much of the work required to create it. Latour advocated "opening" the black box through historical reconstruction, which means studying artefacts before they become closed and stable. Here, Harrison had a unique opportunity in her study to examine the black box before it closed as the studied facility was literally under construction. This potential openness made it easier to recognise the invisible work being done, as it had not yet become hidden as part of the infrastructure, which otherwise often happens once everything appears to run smoothly and the black box is closed.

As such, the book is built on a study that Harrison did in conjunction with the establishment of ESS and in particular the Data Management and Software Centre (DMSC). DMSC is used as the primary source of information for understanding how infrastructural decisions reflect and shape knowledge-producing practices in Big Science. Harrison defines (New) Big Science as follows: "The term is often now associated with any kind of scientific endeavour involving large, geographically distributed research teams producing or curating large data sets, and not necessarily tied to a single facility or institution" (p. 7). Big Science is thus a global spectacle with an international audience in a fiercely competitive and costly arena, as Harrison says: "Money, expertise, people and equipment flow into the ESS. Data flows out" (p. 8)

So, what are the *practical* consequences of these observations of (mis)managed data? What does it *matter*? What are the (hidden) *purposes*? Harrison makes a point out of the fact that the study sits on two chairs, telling two parallel stories; one that regards raw data as an oxymoron and one where it is a precondition for conducting proper science (p. 32).

The empirical data most readily visible to the reader comes from interviews, mainly with individuals in leadership roles within the previously mentioned part of the organisation known as DMSC. A total of seven people were interviewed once a year over three years (2015–2017). The interviews focused on three themes: organisational and production patterns; practices; and artefacts, which also form the basis for three empirical chapters in the book. Harrison carefully highlights the limitations of the study, noting that this particular methodological approach and selection can only provide a snapshot of the complexity and diversity involved in data management.

Harrison goes on to examine the "data flows" managed by DMSC, tracing the journey of data (Leonelli 2016) from the experiments conducted to the end users who engage with the generated data. The data takes its journey across various software and hardware instances, and a key example of this is Kafka – an open-source software designed to handle and store streams of data. This software must be aligned and integrated with other systems to ensure that data flows as smoothly as possible. This synchronisation process (or alignment work), which involves both humans and machines, also requires data to be adapted for different "relays" and recipients within the network. At its core, this process concerns the precision and categorisation with which a given instrument measures something – a fundamentally Baradian perspective (2003, 2007). The informants navigate between presenting data as "raw" or unaltered (primarily when it has only been processed by machines) and acknowledging that some data may be lost or renegotiated (especially when handled by humans). While

this is neither new nor unique to the ESS, Harrison makes it clear that "tidying up" is not merely an innocent act of organisation; it actively shapes how data is interpreted, understood, and legitimised within scientific practices.

Harrison further demonstrates how data management in Big Science is becoming increasingly professionalised. In the past, software development and data handling were managed informally, often by individual researchers within a research group. However, as datasets grow larger and more complex, specialised technical expertise is now required. As such, gatekeeping and protecting through boundary work (Gieryn 1983) would have been interesting as an analytical lens. Harrison examines both the standardisation and commercialisation of data management software, as well as the professionalisation of data handling as an occupation. While the book argues that traditional boundaries are dissolving, the divide between scientific and technical work has never been as clear-cut as much research seems to suggest – especially if you ask the technicians. The boundaries that do exist are likely already well known to those involved and are maintained through both invisible, routine labour and deliberate strategic demarcations.

Still, Harrison presents a valuable research overview highlighting how the role of technicians in scientific and knowledge production has often been overlooked. However, it remains unclear how the debate she engages in – between technicians and scientists – differs from similar discussions in other fields, or how it fundamentally impacts knowledge production itself. Instead, the primary effect seems to be about access to funding and visibility in academic publications. Data management professionals perform skilled craftsmanship when they "cook" data, yet they place great importance on keeping the data "raw". In doing so, they themselves contribute to the invisibility of this work. This makes us wonder about alternative ways to describe this professionalisation. To us, it sounds more like a commercialisation and commodification of science, where research unfolds within a market at every stage of the data's journey.

By the end of the book Harrison has returned to the intriguing concept of "rawness" in the data produced by ESS. The central debate revolves around whether data is inherently mediated and influenced or can exist in an unaltered, "raw" state. Harrison argues that behind-the-scenes processes at DMSC impact data production, raising questions about transparency. While transparency (at least in terms of data) might benefit experts, it could arguably also overwhelm most audiences. Harrison resolves this tension by focusing on how group leaders at DMSC perceive, modify, and negotiate data "rawness". To us, their discussions echoe clearly of Karen Barad's theoretical framework, which includes concepts such as phenomena, apparatus, and cuts, as the chapter highlights that "rawness" is context-dependent (situated), shaped by tools, processes, purposes, and intended audiences.

Data itself is something that remains undefined in the book – what exactly is data? Computation consists of both data and algorithms, so drawing sharp distinctions between different fields (e.g., critical data studies (CDS), STS on science, algorithmic regimes) may be misleading. After all, computers are fundamentally algorithmic data processors. As such, we would love to see a follow-up book by the author exploring concepts, their trajectories, and the invisible work behind these.

The title of this book might lead readers to expect Harrison to unveil the hidden mechanisms of Big Science with a dramatic: tada! – revealing the machinery that sustains its magic as truth. And as such to burst out: "I told you so: 'raw data is an oxymoron and a bad idea' (Bowker & Star, 2000; Gitleman, 2013), and it is always shaped by management procedures, practices, and technologies – and thus always already cooked." And yes, Harrison does indeed do this (even though this is not her primary goal), but she also does something more – something subtle and nuanced. She does reveal the invisible labour of data management in Big Science, which may not come as a surprise to CDS or STS researchers. However, what makes her approach distinctive is that, as we trace the journey of data through the Big Science facility, we also follow Harrison's own journey. We particularly appreciated this aspect of the book. The book is written with a clear well-organised structure that makes the project journey easy to follow. The writing is both enjoyable and exemplary in its simplicity, relying on just two key concepts – black box and invisible work – without introducing unnecessary neologisms. Harrison's study not only brings much-needed attention to the hidden labour of data management but also offers a thoughtful examination of how knowledge infrastructures take shape through daily practices. With its clear structure, engaging writing, and well-grounded theoretical approach, the book is a valuable resource for anyone interested in the intersections of data, technology, and scientific work.

Lina Rahm and Jörgen Behrendtz

Lina Rahm is an associate professor in the History of Media and Environment with specialization in Artificial Intelligence and Autonomous Systems, at KTH.

Jörgen Behrendtz is professor in media studies at Stockholm University.

References

Barad, K. (2003)"Posthumanist performativity: Toward an understanding of how matter comes to matter", *Signs: Journal of Women in Culture and Society* 28 (3):801–831. https://doi.org/10.1086/345321

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meanin*g. Durham: Duke University Press.

Bowker, G.C. and S.L. Star (2000) *Sorting things out: Classification and its consequences.* Cambridge: MIT Press.

Gieryn, T.F. (1983) "Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists", *American Sociological Review* 48 (6):781–795.

Gitelman, L. (2013) *Raw data is an oxymoron*. Cambridge: MIT Press.

Harrison, K. (2025) *Behind the science: The invisible work of data management in Big Science*. Bristol University Press.

Leonelli, S. (2016) *Data-centric biology: A philosophical study*. Chicago University Press.