# Afterword: Data in Transit


**Helen E. Longino**

**Abstract** The naïve fantasy that data have an immediate relation to the phenomena of the world, that they are "objective" in some strong, ontological, sense of that term, that they are the facts of the world directly speaking to us, should be finally laid to rest by the papers collected in this volume. In this afterword, I propose that these papers, investigating data journeys in fields from particle physics to urban planning, show that even the primary, original, state of data is not free from researchers' value- and theory-laden selection and organization.

The naïve fantasy that data have an immediate relation to the phenomena of the world, that they are "objective" in some strong, ontological, sense of that term, that they are the facts of the world directly speaking to us, should be finally laid to rest by the papers collected in this volume. One might think that "data journeys" catalog the way that raw givens are transformed as they move from their original context to other contexts, whether higher levels of abstraction in the same field or other fields of inquiry. These papers, investigating data journeys in fields from particle physics to urban planning, show that even the primary, original, state of data is not free from researchers' value- and theory-laden selection and organization. Once reactivated in a new context, once they have journeyed, the mutability of the data is even more starkly revealed. But it is just this mutability that demands of researchers' creativity and diligence in the preparation and transport of data. Data are the currency of science and, even though not nature announcing itself to us, must be protected as if they were, because they are the closest we have. Where philosophers have in the past treated data as brute, unanalyzed, givens for purposes of inference to hypotheses or theories, the essays in this volume approach data as themselves the outcome of research practices. The practice perspective reveals a multiplicity of data production and manipulation processes. I will underscore the ways activities reported in four of these papers demonstrate this point. I will then engage in some general

H. E. Longino (✉)
Department of Philosophy, Stanford University, Stanford, CA, USA
e-mail: hlongino@stanford.edu



S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_20

reflection on the lessons from the workshop that spawned this volume for thinking about data. The journeys are various, the fields even more so.

The contributions from Ramsden and Müller-Wille illustrate the challenges of obtaining information from data. Of course, one can count and measure any number of things. The trick is to measure the right things and to measure and report them in a way that will serve cognitive and practical purposes – one's own and those of others. Ramsden's paper tracks the progressive improvement in the quality of data on housing in mid-twentieth century United States, setting it alongside changes in the perception of the lives of the poor by middle-class professionals. Mueller-Wille's follows anthropologist Franz Boas's even earlier development of standardization in anthropometrics. This is set alongside changing demographic preoccupations in the US. Both papers concern themselves with the technologies researchers developed to make physical data relevant to social and cultural questions, as well as with the institutions, norms, and practices within which those technologies were deployed.

Ramsden focuses on the efforts of Edith Wood to obtain data that could be used to create national housing standards. When she began her work in the early 1930s she could complain that what data existed were locally variable, so that generalizations across states and cities were not possible. Wood "knew" that housing stock was inadequate, but the available data sets measured different things and used different scales of measurement. A break came in the form of data collected for commercial purposes: the Real Property Index collected for the real estate industry. This was a nationwide survey and classification of real property that made possible meaningful comparisons of residential housing. It included information on availability of utilities (gas, electricity), running water, bathing and toilet facilities). Its utility for Wood was its broad geographical reach and the consistency of items catalogued or measured. Wood was then able to find enough data on various indices of ill health (mortality, morbidity, delinquency) that she could map areas of more intense social ill health onto transparencies and overlay those on maps of housing quality created with the RPI. With correlations established in enough municipalities, the housing data could come to signify the intensity of social ills without need for further study. Wood's purpose was to advocate for minimum building standards as essential to a healthy society. Once her stress on comparability of data was accepted, the data became more fine-grained and the standards more demanding. Ramsden traces the trajectory from increased amounts of data, an increase that eventually made the data unusable, to the selection of key elements that could be taken as informative of a range of qualities. More is not always better.

What Ramsden also shows is that the impact of information depends on the aims of those with the power to use the information. The impact of uniform housing codes depended on the attitudes and aims of those for whom the data were produced. In Wood's time the goal was to improve the quality of life of the poor and indeed rates of disease and crime were (initially) lower in those areas where housing stock conformed to the new standards. But it takes more than square footage and running water to create a good life. In an urban context, it takes transportation, jobs, shops, spaces for social life. In the 1950s and 1960s the social reformers' goal of public health was replaced by urban renewal which came to mean the wholesale destruction of neighborhoods perceived as plagued with substandard housing (and the associated

social ills) and their replacement with high-rise, uniform apartment blocks. In some cases the residents of neighborhoods classified as substandard were moved wholesale into new areas and placed in housing that conformed to the physical requirements validated by the extensive data collection and analysis Ramsden chronicles. But, as the Boston example shows, the uprooting of families and neighbors destroyed social bonds that had come to flourish in less than ideal physical circumstances. Those social bonds, the relations of friendship and acquaintance as well as of commerce, are just as important in overall health as the physical requirements codified by housing advocates. New kinds of data sought and generated in the wake of the West End project in Boston signaled the lack of transportability of quality of life indicators that had to do with the interrelationships of residents with one another. The conclusion seems to be that these indicators are local, and the measurable indices too variable to be applicable across municipalities and states. It might be possible to read the lesson as that the conditions of "healthy" neighborhoods are plural: there is more than one way to make for good quality of life. Such a lesson is not generally welcome in a context looking for uniform and universal standards. And so, the story of data on housing and public health comes in some ways full circle, but through paths that demonstrated the importance of minimal standards, the relation between housing conditions and disease (especially communicable diseases such as tuberculosis) and what from one perspective one could call crime and from another, security.

Mueller-Wille's study of Franz Boas's anthropometric methods is, like Ramsden's, a story of efforts to integrate physical with social information. Boas was a staunch environmentalist, believing that both environment and biology contributed to the expression of specific physical characteristics in individuals. To this end his anthropometric effort was directed at matching continuity and discontinuity of physical characteristics with genealogical relations and tribal identities. Boas selected a small number of physical variables but measured them on a large number of individuals. In particular he conducted a major study of the Chickasaw, a people of the southeastern United States that was relocated together with the Choctaw as Americans of European descent pushed into Indian lands. Mueller-Wille stresses the hybrid character of the data Boas developed. Height and cranial features could be measured with standardized measuring instruments. Genealogical information, including tribal affiliation of parents, was, however, provided by the subjects themselves. Tribal affiliation was, in turn, determined by the informants according to internal standards of kinship and belonging (whether mother or father was of the tribe, etc.). Boas was interested in these data for theoretical purposes. His interest in pursuing this line of research among American Indians flagged when the data came to be seen as relevant to the highly political issues of tribal membership and entitlement. In spite of Boas's loss of enthusiasm for his study, the data have been preserved and have now been used to ground longitudinal studies of the (descendants of) the populations Boas studied and in statistical reanalyses of the data themselves. Mueller-Wille alludes to some of the recent debates about Boas's study of immigrants to the United States, a study he took up after abandoning the research on native Americans. This work, too, was caught up in political controversy, as Boas (and many after him) argued that his data showed the role of environment in the production of physical traits like cranial size. Such findings are not at all to the lik-

ing of those who urge restricting immigration of certain ethnic groups based on the assumed immutability of certain of their physical characteristics.

The essays of both Ramsden and Mueller-Wille, like those of Cambrosio and colleagues, Bechtel, Boumans and Leonelli, and Ankeny, make evident the struggle to identify and provide usable data. In both cases, the data gathered in one context needed to be comparable with data taken in another context. This required identifying what data could be found in all the contexts that needed to be compared, and selecting what among those variables would be most informative. Uniform measuring tools were necessary as well as universally available targets of measurement. Observers needed to be trained to use the tools, such as calipers, and to perform the measurements of, for example, the right cranial dimensions. And, in both cases, techniques of visualization are also required to make the import of the data evident. Overlaying transparencies marked by frequency of socially undesirable phenomena over municipal maps marked according to quality of housing stock enables even the non-statistically literate to see the connection between quality of housing and health. Drawing up kinship trees and putting information in tabular form again enables a "reading" of the anthropometric data not facilitated by mere reporting without attention to presentation. The first journey is the journey to comparability and association, facilitated by the selection of categories of data and of tools and the training required to use them; the second to visibility, facilitated by techniques of presentation. Furthermore, in both cases, the data are enlisted for further purposes. These purposes also leave their mark on the character of the data. So, the conceptions of public health current in mid-century support a focus on internal features of a home – square footage, running water, availability of a toilet – but not the elements of social glue that bind the inhabitants of those homes into a community. The need to have all members of a population represented for a comprehensive kinship mapping of the physical traits requires that obtaining the data involve no violation of modesty (disrobing) thus limiting what measurements can be performed. Finally, there is also a sense in which the data break free (or are broken free) from the contexts of their production and get deployed towards aims that the original research may not have envisioned and might not even endorse. The emphasis on the internal, physical requirements for healthy living determined the kinds of data that were available about neighborhoods and when urban planning shifted its emphasis slightly from public health, the welfare of the inhabitants, to urban renewal, a more comprehensive and impersonal value, the role of the data expanded from supporting building standards and encouraging or enforcing renovation to supporting destruction of neighborhoods whose housing stock seemed unamenable to updating and displacing their former inhabitants. Boas could see how his data could be used not just for the theoretical environmentalism he advocated but also in debates about pure-bloodedness, tribal membership and access to the benefits associated with both. When data are comprehensive (that is, include information on selected variables for an entire or very large segment of a population) they lose some of the apparent mooredness to their context and take on an independent life that makes them available for reuse in other contexts. As Mueller-Wille reminds us however, in spite of appearances, a crucial part of the data Boas gathered on the Chickasaw was indissolubly rooted to its context, as tribal identity was recorded based on the testimony of the subjects.

Koray Karaca takes us inside the black box of a particle detector. This is a research world characterized by very different challenges than the social worlds explored by Boas and by Edith Wood and her followers. His detailed description of the data preparation process in the ATLAS experiment at the CERN LHC nevertheless reveals some similar patterns as characterize the housing and anthropometric data generation, but at a vastly greater scale. The similarity is in the need to find the telling data among quantities of events, just as in the previous two cases the challenge was to identify the relevant variables. But whereas the issue in the previous two cases was to identify the particular variables that were both universal and variable enough to enable data to exit the contexts of their production for comparison with similarly produced data, in the case of the High Energy Particle Physics world, the challenge is to winnow down the unmanageable amounts of data produced in any given run of the collider. A succession of triggers selects events that will be informative given the aims of any given experiment. Collision events produce masses of different kinds of particle at various energy levels. Many, perhaps most, of these are already well understood. The point of the experiments is to find the rare events that are evidence of predicted but not yet detected particles, like the Higgs boson was for ATLAS, or particles or events that indicate physics processes not foreseen in current physical theory, the Standard Model. At the first selection level, the point is to thin or prune the data to a more manageable size, so that events of interest, that is, potentially informative events, will be more salient. What remains as data are the products of (a very small proportion of) the collision events and at the next stage these are further reduced while the remaining products are amplified by adding information about the trajectories known to produce those particular products or signatures. So, the collision data is reintroduced based on theoretical understandings of the particle and energy properties. Finally, at the third level selection triggers reflecting just what it is the researchers hope to find are applied to the products of the second level of selection. The products of this third selection are the data that will be subject to analysis.

Clearly a great deal of theory is required to design the triggers. So, while the data are not theory-laden in the old Kuhnian sense they are certainly theory-mediated (to adopt a phrase of George Smith's). Karaka describes a process of transforming the "blooming, buzzing confusion"[1] of collision events into data suitable for analysis and for use as the basis of inferences about particles. While on a first reading it may seem that the sequence of triggers renders the data hopelessly theory-dependent, it is important to remember that their "provenance" remains available. Unless one wants to call into question the entire enterprise of High Energy Physics, the relation of the final data to the wild unmanageable dance in the collider is readable through the technical specifications of the series of triggers. Seen this way, on a second reading, what Karaka has described at the LHC is in some ways a better documented production of data than most. James's phrase refers first to the human infant's experience of the sensible world, but can also apply to the indefinite number of ways we can individuate the contents of our sense perception. That we humans perceive as we do (three-dimensional medium sized objects perceived via wave lengths within

[1] To repurpose William James' famous description of infant perception (James 1890).

a small part of the full spectrum) is a product of our evolutionary history, but even within what we could perceive, we attend only to a small portion. While some of the selections we make can be reconstructed by reference to the interests we have in extracting information from a given perceptual experience, many of the criteria by which selections are made are buried in pre-conscious neural processes. Unlike the processes of the detector, our processes are opaque to us, revealed only by researchers studying mammalian perceptual systems. Whatever the world out there is like, its signals must be processed in order that we can begin to make some sense of them. In both the constructed detector and the human detector, what is important is that the selection processes, however much they may transform the inputs, retain the relationships (of relative magnitude, of time, of extension) as the data travel from input to cognitively accessible.

Finally, Coopmans and Rappert take us into the exotic (or, per their coda, perhaps not so exotic) world of art and art forgery. Here the stakes are complicated. In the three previous contemplated papers, the stakes are faithfulness and accessibility in service to pragmatic or theoretical values. This paper makes clear that, in the art context, faithfulness and accessibility have more than an epistemic interest. Conceptually speaking, the appraisal of a (putative) work of art has (at least) two dimensions, not always distinguished, and often conflated. One is the identification, perception, and communication of the aesthetic values exhibited by a particular object: the interactions of color, form, figure, and meaning. The other is the attribution to a particular hand, the hand of Leonardo, Rembrandt, or skipping ahead a few centuries, Pollock. In the former case the aesthetic properties of a work and one's abilities as a connoisseur (whether professional or amateur) who can detect them are at stake. In the latter, millions of dollars. The art world has always been a world of mixed motives and mixed values. Artists need to live, after all. But the twentieth century, in particular, has seen a boon in the secondary art market of dealers and auction houses, where works attributable to the hand of a Picasso, a van Gogh, a Monet, a Rembrandt, fetch sums the makers themselves never dreamed of. In such a world, the temptation to produce look-alikes, forgeries, is overwhelming. What is relevant to determining the real from the fake?

Coopmans and Rappert draw on the memoirs of various players in the art world to bring out the contested nature of evidence in this world. Hoving, the connoisseur, "knows" at a glance both whether a work possesses aesthetic value *and* whether it is a genuine work by the artist to whom it is attributed. Indeed, for him, there is no difference between these judgments. Like Boyle summoning credibility for his air-pump by inviting gentlemen onlookers, Hoving tries to enlist us as witnesses by pointing to the telltale details of the work that give away a forgery as fake. We, the non-connoisseurs, ought to be able to see as well as he, once tutored (and given enough study and tutelage might even come to be able to make such judgments on our own). Is this a real X? Here, let me show you. But there is another kind of detail: the physical trace, especially the fingerprint, but chemical analysis, too. Chemical analysis can help identify a fake as a fake, but the fingerprint links the object to a body in a way that does not depend on our being able to appreciate what the connoisseur asks us to "see.". Here the presumption is a causal chain from the print on the object to the hand of the painter and here we can see the orthogonal system of

value at work. A not very good Pollock or Picasso can still fetch higher sums at auction than a (judged by aesthetic criteria) better painting by an unknown. Does the worth of the work lie in its intrinsic aesthetic qualities or in its origin? If someone has been clever enough to produce a painting that looks every inch a Matisse or a Renoir (but is not), should we care? Would we put it up in our living room or entrance way? Would we enjoy it less? This depends on the nature of the pleasure it gives us, the pleasure of responding to aesthetic qualities or the pleasure of pride of ownership. Of course, there are complicated issues of originality, as well as of quality, that could be pursued in a fuller discussion of the relative value of originals versus forgeries, but this paper draws our attention to the wealth of data extractable from a work of art, as from "nature." What data are relevant, what data will travel well and what data will not depend on the purposes for which data are sought and the contexts in which they are to be deployed.

These papers reveal a variety of forms of travel. Data can be made visible through juxtaposition with other data, can be repurposed (from commercial valuation purposes to public health purposes, from theoretical to political), can be rearranged, can be reduced and recreated, can replace and be displaced. Other papers in the volume reveal how data from one source must be manipulated and adjusted in order to be compared or integrated with other data from different sources. Much discussion centers on the importance of preserving metadata in order to preserve the integrity and meaning of the data.

In all the tracings of data *journeys* we go back to the origin of the data, the start of the journey, and here the papers reveal the variety of means by which data are gathered, stored, and deployed. So, no, there is no such thing as "raw data."[2] There are the phenomena of the world, some perceptible to us, some not, always filtered by our means of perception. The resulting data, however closely linked to the phenomena, are always symbolic representations in some medium. Even the samples collected by dragging a receptacle through the sea become symbolic of what is left behind. To be informative about those phenomena, a single datum must be set in the context of other data: as discussed by several chapters, and particularly Mary Morgan's and James Griesemer's, we must have a data set or a singularity set against a background of other data. The data are always selected and produced, a function of the techniques of observation, of measurement, of recording. The expression "raw" is, however, trying to get at some aspect of the process. As we study the journeys we need to go back to some originary point where the data have been subjected to the least processing, a point closest to their context of generation. There are different ways to convey this notion. Niccolò Tempini suggested "source" versus "derivative." In another context we might think of less versus more defeasible. If data travel, there is a place or a status from which they travel. So we might think in terms of base level measurements that can be used to plot against other base level or against higher level measurements to engage in comparative analysis or to tease out the significance of the measurements. What counts as base level will depend on the context and what counts as a higher-level set of correlations in one context may be base level in another.

[2] A point also emphasized by the title and contents of Gitelman, ed. (2013).

But even if there is no such thing as self-announcing data, it doesn't follow that all data are "just interpretations." A naturalistic versus an interpretive approach may be too coarse a distinction for data, although does reflect a difference in how science studies has approached data. The naturalistic approach (found among science practitioners and some philosophers) is accused of naiveté, not understanding the work that goes into generating data that can be used to support scientific inference. The interpretive approach (found among constructivist science studies scholars) is accused of undermining the trust we rightly place in science. While data need other data and sometimes also need theory in order to "speak", to focus on the interpretative dimensions does not mean that the data are unreliable or fake, but that data must be selected, must be classified, must be set in relation to other data. What requires attention are the methods for obtaining, recording, and storing data and how well those methods serve the purposes for which data was sought in the first place. This doesn't place data practices above criticism, but helps us see the multiple places where criticism can be directed and therefore the multiple places where data practices may require defense. The perspective of science practice reveals a wealth of epistemologically relevant moves in the research context. Our understanding both of the trustworthiness of science and of its limits will be enhanced by the attention to data practices manifested in these essays.

## References

Ankeny, Rachel A. this volume. Tracing Data Journeys Through Medical Case Reports: Conceptualizing Case Reports Not as "Anecdotes" but Productive Epistemic Constructs, or Why Zebras Can Be Useful. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Bechtel, William. this volume. Data Journeys Beyond Databases in Systems Biology: Cytoscape and NDEx. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Boumans, Marcel, and Sabina Leonelli. this volume. From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycle Analysis. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Cambrosio, Alberto, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R. Jordan, and Pascale Bourret. this volume. 'Overcoming the Bottleneck': Knowledge Architectures for Genomic Data Interpretation in Oncology. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Coopmans, Catelijne, and Brian Rappert. this volume. Data Journeys in Art? Warranting and Witnessing the 'Fake' and the 'Real' in Art Authentication. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press.

Griesemer, James. this volume. A Data Journey Through Dataset-Centric Population Genomics. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

James, William. 1890. *Principles of Psychology*, 488. New York: Henry Holt.

Karaca, Koray. this volume. What Data Get to Travel in High Energy Physics? The Construction of Data at the Large Hadron Collider. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Müller-Wille, Staffan. this volume. Data, Meta Data and Pattern Data: How Franz Boas Mobilized Anthropometric Data, 1890 and Beyond. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Ramsden, Edmund. this volume. Realizing Healthful Housing: Devices for Data Travel in Public Health and Urban Redevelopment in the Twentieth Century United States. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Tempini, Niccolò. this volume. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

**Helen E. Longino** received her PhD in Philosophy from the Johns Hopkins University in 1973. Her teaching and research interests are in philosophy of science, social epistemology and feminist philosophy. She is particularly interested in the relations between scientific inquiry and its social, cultural and economic contexts. In addition to *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality* (University of Chicago Press, 2013*),* she is the Author of *Science as Social Knowledge* (Princeton University Press, 1990), *The Fate of Knowledge* (Princeton University Press, 2001) and many articles in the philosophy of science, feminist philosophy and epistemology and Coeditor of *Scientific Pluralism* (University of Minnesota Press, 2007). She has taught and lectured at universities in Europe, Asia and South America, as well as in the United States. She is currently C.I. Lewis Professor in Philosophy at Stanford University.