Process and Purpose, Not Thing and Technique: How to Pose Data Science Research Challenges

Robert Williamson^{1,2}

¹School of Engineering and Computer Science, Australian National University, Canberra, Australia,

²Australian Academy of Science, Acton, Australia

Published on: Sep 30, 2020

DOI: <https://doi.org/10.1162/99608f92.6e525663>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

1. Let's Get Out of Disciplines and Get Into Problems

I welcome the opportunity to respond to the stimulating articles of Jeannette Wing, Xuming He, and Xihong Lin (this issue).

To a first approximation, Wing's challenges are more 'problem-oriented' and He and Lin's more 'technique oriented' (I will explain the distinction below, and argue why I think it matters). I have more to say about the problem-oriented perspective so I organize my response primarily according to Wing's article. I first make three meta-remarks.

First, I consider the issue of whether data science is a discipline, and what one might do about this. While the pro-disciplinary viewpoint is undoubtedly popular nowadays, I have great sympathy for the contrary view of 19th-century scientist Thomas Henry Huxley, who said "Authorities,' 'disciplines,' and 'schools' are the curse of science; and do more to interfere with the work of the scientific spirit than all its enemies" (quoted by Barber, 1961, p. 601). A major function of disciplines is to help identify outsiders and to keep them excluded from the in-group (Krishnan, 2009); notably one of the traditional functions of the 'tribe,' which is perhaps a more precise word to describe the disciplinary enterprise (Becher & Trowler, 2001). Disciplines are about controlling behavior: "[T]here is an important moral dimension to 'discipline' that defines how people should behave or think" (Krishnan, 2009). Rather than engaging in such exclusionary and controlling endeavours, motivated more by the pursuit of power than the pursuit of truth, I prefer the view succinctly articulated by Popper (1993) at the beginning of his *Realism and the Aim of Science*: "Subject matters in general do not exist. There are no subject matters; no branches of learning—or, rather, of inquiry: there are only problems, and the urge to solve them." Indeed. I would be delighted to see no further discussions of creating new disciplines, or arguing where their boundaries lie; I would rather see us collectively getting on with the job of solving problems.

Mention of "problems" brings me to the *second* point that I think is pertinent: the distinction between "problem-oriented" and "method-oriented" articulated by Platt (1964). Machine learning conferences (with which I am familiar) abound in method-oriented contributions: improved algorithms (methods) usually presented in a manner quite divorced from problems other than as a benchmark. Research contributions from a problem-oriented perspective are rarer. Although the boundary is not sharp, as Platt well argued, taking the perspective of the end problem at hand (or the purpose) does empirically seem to be highly valuable. Since Platt made his case so well, I say nothing more in general about this but commend his old article to readers.

My *third* meta-remark, related to the second, and implicit in both the articles' contributions, concerns the pervasive valorization of the 'algorithm.' For example, we see books with stupendous titles such as *The Master Algorithm* (Domingos, 2015). I maintain that algorithms are the wrong concept to organize our thinking around, because we cannot actually say (precisely) what an algorithm is, or when two algorithms are the same.

Not only will you not find a formal definition in the canonical text on the subject (Cormen et al., 2009), but the authors explicitly say that they avoid giving a definition! And well they should, because the task seems quite hopeless as Blass and Gurevich (2004) and Blass et al. (2009) have compellingly argued. Fortunately, we can sidestep this definitional conundrum by focusing our attention on the *problems* of data science to which I now turn. (I recognize that this then begs the question of what a problem is, which I will simply duck for now.)

The authors of the two contributions identify many interesting and pressing problems concerning data science; I make some comments about Wing’s list in particular later in my response. But first I wish to focus on what seems to be a lacuna in their contributions though, a consideration of how we think of data, and why we put so much value on it.

2. The Value of Data

It is instructive to reflect upon how the earliest data scientists thought of data. (I do note that calling them data scientists is especially anachronistic given the very word scientist was only coined much later). A prototypical example is Robert Hooke (1665), who undertook an investigation (observation LVIII) into “a new property of air.” He ended up with his “Table of the Elastick power of the Air”; what might now be called a data set (Hooke, 1665, p. 226). The key point is that his *data was gathered for a purpose*, and it reflected that purpose, and indeed is of much less use for any other purpose (other than pedagogical). Data divorced from purpose is data divorced from context, and is just a string of bits. The power of data resides in its entanglement with a problem or a purpose. (This is not to deny that reuse of old data is valueless; it clearly is not; Pasquetto et al., 2019. Merely that the dangers in reusing are even greater in that a larger number of errors are possible.).

As well as the issue of *purpose* there is the question of the *fact-likeness* or *indisputability* of data. Not everyone is like Hooke, able to design the experiment, collect the data, and analyze it themselves. Far more common nowadays (especially for people known as data scientists) is the situation exemplified by Hooke’s contemporary (and *bête noire*) Newton, who had to beg, cajole, and persuade *others* for their data, for instance, from the Astronomer Royal John Flamsteed:

For all the world knows that I make no observations my self & therefore I must of necessity acknowledge their Author: *And if I do not make a handsome acknowledgement, they will reckon me an ungratefull clown.* Isaac Newton to John Flamsteed 16 February 1694/5 (Scott, 1967, p. 87)

Newton’s exasperation in writing letters to get hold of another’s data remains today (Noy & Noy, 2020). He was well aware of the need for the data to be worthy of his trust, and he consequently wanted his data as raw (unprocessed) as possible: “I want not your calculations but your Observations only” (Newton to Flamsteed, 29 June 1695; Scott, 1967, p. 134). At issue was that Flamsteed provided Newton with ‘corrected’ data, taking account of parallax and refraction. Kollerstrom and Yallop (1995, p. 238) asked “were the data unduly theory laden?” signaling that this is a matter of degree rather than a categorical distinction. They concluded (ironically, given Newton’s evident frustrations) that “there is little doubt that the data sent by Flamsteed to

Newton comprise the most accurate solar-lunar positional observations then made.” But, of course, Newton was not to know this at the time: Newton trusted Flamsteed’s raw observations, but not his data processing skills!

Now it might be argued that such antique examples are irrelevant to the modern enterprise of data science. Indeed, through the 20th century there was a gradual abstraction of the notion of data. Rather than experiments being done for a purpose, it was argued that they were done solely to gather ‘information’ (which is vaguely conceived of as refined version of data); for a stimulating critique of the traditional view, see Tuomi (1999). Nowadays, data is commonly gathered for its own sake (with no particular purpose in mind, or at least declared), in such large quantities that it is stored in “lakes” (Fang, 2015). While there are sometimes reasons given for the widespread collection of data ‘just-in-case,’ it seems some such claims have been overstated (“NSA Surveillance Exposed,” 2020).

I claim that data only has value (for inference) when considered relative to a *purpose* and that purpose needs to be kept in mind from the outset. There is a view that much data analysis is not solving a decision problem, but is rather merely about ‘gathering information.’ Criticism of this perspective is hardly new: confer De Groot’s (1961) understated conclusion: “the distinction sometimes made ... between decision problems and problems in which the experimenter simply wants to gain knowledge may not be very sharp.” The point is that one does not analyze data solely to gather information. Ultimately, one acts, and in so acting there must be some criteria judging the quality of the act, in other words, a loss function. In fact, widely used mathematical definitions of ‘information’ can be shown to be simple reparameterizations of Bayes’ risks of experiments; that is, the smallest average loss attainable for the given statistical experiment (Reid & Williamson, 2011). Thus, the moment one chooses a sensible measure of information, one is implicitly choosing a loss function, and hence any appearance that one is merely gathering information (and not implicitly optimizing a loss) is indeed mere appearance.

The idea that data can have a purpose-independent value is closely related to how data is conceived of in the first place. The very language that is commonly used to describe our data belies our mental mode: we conceive of data as *set*, arguably the mathematical object with the least possible structure: *a set is merely a container*. The ‘set’ view makes some sense when data is viewed as a ‘fact,’ described by Poovey (1998) as “deracinated particulars”—elementary statements about the world that are divorced from context. We tend to think of data as *given*—confer the nice contrast between ‘data’ (which is *given*) and ‘capta’ (which is *taken*) (Kitchin, 2014). The set view construes data as a *thing*, a presumption made even in subtle analyses of data’s role, for example, as capital (Sadowski, 2019).

A more significant distinction arises from the degree to which data can be *disputed*. It is natural to seek some solid ground on which to build, and that is true of our inductive exercises. Mary Poovey looks at the history of this idea in her wonderful *A History of the Modern Fact*, which she opens by asking

What are facts? Are they incontrovertible data that simply demonstrate what is true? Or are they bits of evidence marshalled to persuade others of the theory one sets out with? Do facts somehow exist in the world like pebbles, waiting to be picked up? Or are they manufactured and thus informed by all the social and personal factors that go into every act of human creation? (Poovey, 1998, p.1).

But what of data? Is data more fundamental than facts? And is it really ‘incontrovertible’? Rosenberg (2013) observes that Joseph Priestley (in 1788) referred to the facts of history as “data” and goes on to observe that “if facts can be deconstructed—if they can be shown to be theory-laden—surely data can be too.” He concludes by asking “what makes the concept of data a good candidate for something we would *not* want to deconstruct” (Rosenberg, 2013, p. 18).

Schaffer (2009, p. 246) privileges “information” over “knowledge”: “‘Information’ here is a term designed to describe matters more broadly shared and less explicitly challenged than formalized knowledges. Information is the commonly taken-for-granted, rather less disputed and less disputable; knowledge looks more mutable, its status certainly more debatable.”

Whether it is information, fact, or data that sits at the bottom, the crucial thing seems to be the commonly accepted point that there needs to be *something* one rules out of bounds for dispute, questioning, and deconstructing; something that we trust absolutely. (Popper, 1960, pointed out the intrinsically authoritarian nature of this presumption of a source of absolute truth.) Deeming data to be a ‘fact’ means we are henceforth unable to question it *by definition*; it can thus no longer be challenged in argument, a device not unknown in rhetoric:

A fact loses its status as soon as it is no longer used as a possible starting point, but as the conclusion of an argumentation. *It can recover this status only by being detached from the context of the argument*; in other words, there must once again be an agreement that does not depend on the terms of the argument for its proof (Perelman & Olbrechts-Tyteca, 1969, p. 68, italics added).

But there is an alternative view, which I think, if adopted, could assist in solving some of challenges in the two articles under consideration. Rather than viewing data as a thing, view *data as a process* (Borgman, 2019; Jones et al., 2019; Leonelli, 2019; Wing, 2019). In the words of Heuer (1999):

There is no such thing as “the facts of the case.” There is only a very selective subset of the overall mass of data to which one has been subjected that one takes as facts and judges to be relevant to the question at issue.

I do not have space here to develop this notion further, other than to observe that construing data as a process means one needs to consider the *entire* process, including the *choices made in gathering the data*, and the *uses to which the data will be put*. Thus, rather than simply *trusting* the bottom level (our nondisputable data), we

need ensure it is *worthy* of our trust (i.e., *trustworthy*) precisely by *not* excluding it from critical scrutiny, and by documenting the chain of evidence that makes it worthy of our trust.

3. Addressing Challenges From the Perspective of ‘Data as Process’

I now offer some telegraphic reactions to Wing’s 10 challenges, largely from this nascent ‘data as process’ viewpoint.

1. *Deep learning algorithms.* These are worthy questions, but are method oriented rather than underlying problem oriented. It is striking that so much work is being done on these methods for modeling, relative to the paucity of work performed on understanding the goals of the data analysis, and the uses to which it is put.
2. *Causal reasoning.* Turning the ‘why’ question (Pearl & Mackenzie, 2018) behind causal reasoning upon itself, I suggest it is worth asking *why do we want to determine causes?* I think a general answer is ‘so we can use knowledge of the cause to (partially) control the world.’ This begs the question of why not just pose the problem as a control problem in the first place? Such a viewpoint seems rare in the causal literature: like other aspects of data analysis, there is a (false, I claim) dichotomy between the ‘why’ of matters, and the ‘how’ (indeed answering ‘how’ questions with explanations serves to satisfy our need to know ‘why’). This suggests an extension of Wing’s question: how to properly incorporate the use to which ‘causes’ will be put into the causal inference process; and if one does this, does one actually have any need for causes at all? Control engineers, who design complex but reliable modern technologies eschew the use of the notion of ‘cause’ (if for no other reason than it becomes very problematic in systems with feedback) (Aström & Murray, 2008, p. 1).
3. *Precious data* is arguably a less clear question. But behind this is the deeper (and important) question of how to formalize and model the ‘preciousness’ of the data. I maintain (per my previous points) that this is only answerable extrinsically, relative to the context: data that is precious for one purpose can be useless for another. So, my more refined version is to ascertain how to quantify the value of data *relative to the purpose to which it will be put*. Some results on a simple case are presented in van Rooyen and Williamson (2018).
4. *Multiple heterogenous data sources.* Again, my refinement of this question is to pose it contextually. The related question of the marginal value of an additional data source also (as I see it) needs to be posed, and answered, *in context*.
5. *Flawed data* (not just noisy). This is indeed an important and still understudied problem. If one interprets ‘noise’ broadly enough, it is better still: I am thinking of the extraordinary challenges of selection bias when one does not have a full provenance of the data (so it is logically impossible to fully understand selection bias mechanisms). The question is closely related to He and Lin’s “post selection inference.” Regarding this, I make two points: 1) ‘Data mining,’ which is sometimes used as synonym for data analysis, used to be a pejorative appellation given to the practice of throwing every analytical technique in one’s arsenal at the data in the hope one finds something valuable (or publishable, not even the same thing), and 2) I worry that the

- whole premise of this postselection inference is a euphemism for the more direct (and rather to be deplored) “HARKing”—Hypothesising after the Results are Known (Kerr, 1998).
6. *Trustworthy AI*. I do not like the AI moniker, but I very much like ‘trustworthy’—that is *worthy of our trust*. My single point here is to compare the notion with the original goals of the scientific enterprise—the production of trustworthy knowledge, which evolved out of the social practices of 17th-century English gentlemen, especially the great importance placed on a gentleman’s word (Shapin, 1994). Trustworthiness is paramount in the inductive virtues, and it serves as a fine touchstone to judge all of our practices and innovations. I think a fine question to ask of all potential research programs is ‘to what degree will this enhance the trustworthiness of our data science?’
 7. *Computer systems*. Unquestionably, energy matters. Although not directly related to my process and problems theme, I cannot resist drawing attention to a nice reflexive point: not only is it necessary (as well argued by Wing) to manage the energy usage of data analytic computations, but the very foundations of data analysis underpin the fundamental energy requirements of *any* computation. As first identified by Rolf Landauer, there is no lower limit to the energy necessary to compute something in itself; rather it is the *erasure of information* that has an unavoidable energy cost, which can be precisely quantified: $kT \ln(2)$ joules per bit erased, where k is Boltzmann’s constant, and T is the temperature (Parrondo et al., 2015). Intriguingly, this “Landauer barrier” can be elegantly interpreted in terms of prediction of information (Still et al., 2012): “any system constructed to keep memory about its environment and to operate with maximal energetic efficiency has to be predictive.” Further elucidating the implications of this fundamental limit (which turns out to be a more refined version of the second law of thermodynamics) makes a very fine research challenge.
 8. *Automating the front end of the data life cycle*. From the ‘data as process’ perspective, I think that the development of systems that support the tracking and management of data provenance particularly important. We should not underestimate the complexity of this task: witness the complexity of Newton’s “information order,” the social network through which he gathered the data necessary for his scientific endeavours (Schaffer, 2009). While this is hardly a novel direction of enquiry (Shamdasani et al., 2015; Yang et al., 2013; Rundle et al., 2015), there is currently a paucity of work integrating such concerns into traditional data science practice as typified by most of the articles in *HDSR*, for example. Two recent examples are “datasheets for datasets” (Gebru et al. 2018), which still takes the thing-perspective of data, but is an attempt to capture some elementary provenance information, and the explication of a measurement model behind data (Jacobs & Wallach 2019). Building systems for provenance over the *entire* data lifecycle directly aligns with the ‘data as process’ perspective even though it is not named as such (Bechhofer et al., 2013).
 9. *Privacy*. Most of the work on data privacy, like indeed much of the work on data analytics, is couched in terms of information, not a task or problem being solved. When one considers the context of these problems more closely, it turns out that there are quite different concerns that arguably swamp the traditional narrow view of privacy as secrecy (Nissenbaum, 2010). Considering privacy in context demands that we ask

questions about *the uses to which the data is put*, not merely whether secret information has been seen by others (Oberski & Kreuter, 2020). This dichotomy directly mirrors the two views of the goal of data science earlier mentioned: to gather information or to make consequential decisions and actions.

10. *Ethics*. Wing, like many commentators, leans on an ill-defined notion of ‘bias’ that data analytic systems are said to sometimes have. This too presupposes an intrinsic perspective, rather than contextual. In the same way that the notion of a biased estimator (from classical statistics) presumes a true value of the parameter, so too does the use of the word in an ethical context. The alternative contextual and end-oriented viewpoint has no use for the concept: indeed, from a statistical decision theory perspective ‘bias’ is irrelevant, only the regret (the difference in expected loss between your estimator and the best possible) matters.

Insofar as data science and technology supports the enterprise of science (broadly construed), it inherits the underpinning values of science, paramount of which is trustworthiness. There can be no *intrinsic* trustworthiness. It needs a context in which it is built, and a reliable chain of provenance regarding the conclusions drawn. Much of the great progress made in the last century in the field of statistics has been to develop better and more reliable methods and theories that can provide some guarantee of trustworthiness (mitigating certain classes of problems, such as sampling error). This is all admirable, and needs further extension.

But what has received startlingly little attention is the trustworthiness of the data itself for the purposes at hand. I posit that part of the problem lies with our fundamental conception of data. If we equate data with uninterpreted fact, then we literally ‘can’t argue with the facts’ because we have ruled out such argument a priori!

The alternative is to eschew this fact-like thing-like view of data. Don’t think of data as indisputable facts that are just given. Don’t think of it as the indisputable starting point of a chain of inference, but merely a constituent part of that chain, one that can be challenged along with the rest of the chain. Such a processual reconceptualization of data then readily suggests a range of other research challenges, which I think should be considered alongside those of the two articles under discussion: how to represent the chain of provenance of data in a manner that plays nicely with our data analytic practices and how to build sociotechnical certification and validation mechanisms for data; and how to ensure all this improves the trustworthiness of data for our purposes at hand.

These questions are inescapably contextual. They are logically impossible to solve in terms of a deracinated free-floating ‘data set.’ We need to embrace that context, represent it, put it up in lights, and, to use Leonelli’s apt phrase (2016, 2019), to improve our “data journeys.”

Acknowledgment

My thinking has been improved and sharpened by enjoyable discussions with Atoosa Kasirzadeh with whom I am preparing a more detailed presentation of some of these views.

Disclosure Statement

Robert Williamson has no financial or non-financial disclosures to share for this article.

References

- Aström, K. J., & Murray, R. M. (2008). *Feedback systems: An introduction for scientists and engineers*. Princeton University Press. https://www.cds.caltech.edu/~murray/books/AM05/pdf/am08-complete_22Feb09.pdf
- Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, 134(3479), 596–602. <https://doi.org/10.1126/science.134.3479.596>
- Becher T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines* (2nd ed.). The Society for Research into Higher Education and Open University Press.
- Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., & Gamble, M. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2), 599–611. <https://doi.org/10.1016/j.future.2011.08.004>
- Blass, A., Dershowitz, N., & Gurevich, Y. (2009). When are two algorithms the same? *The Bulletin of Symbolic Logic*, 15(2), 145–168. <https://doi.org/10.2178/bsl/1243948484>
- Blass, A., & Gurevich, Y. (2004). Algorithms: A quest for absolute definitions. In G. Plun, G. Rozenberg, & A. Salomaa (Eds.), *Current trends in theoretical computer science* (pp. 283–311). World Scientific. https://doi.org/10.1142/9789812562494_0051
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.9a36bdb6>
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. MIT Press.
- De Groot, M. H. (1961). Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2), 404–419. <https://doi.org/10.1214/aoms/1177704567>
- Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it become popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 820–824). IEEE. <https://doi.org/10.1109/CYBER.2015.7288049>

Gebri, T., Morgenstern, J., Vecchione, B., Wortmann Vaughan, J., Wallach, H., Daume III, H., & Crawford, K. (2018). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>

Heuer Jr., R. J. (1999). *Psychology of intelligence analysis*. Central Intelligence Agency. <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/PsychofIntelNew.pdf>

Hooke, R. (1665). *Micrographia: Or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon*. Jo. Martyn and Ja. Allestry for the Royal Society. <https://doi.org/10.5962/bhl.title.904>

Jacobs, A. Z., & Wallach, H. (2019). Measurement and fairness. In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). <https://doi.org/10.1145/3442188.3445901>

Jones, M., Blackwell, A. F., Prince, K., Meakins, S., Simpson, A., & Vuylsteke, A. (2019). Data as process: From objective resource to contingent performance. In T. Reay, T. B. Zilber, A. Langley, & H. Tsoukas (Eds.), *Institutions and organizations: A process view*. Oxford University Press. <https://doi.org/10.1093/oso/9780198843818.003.0013>

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE Publications. <https://doi.org/10.4135/9781473909472>

Kollerstrom, N., & Yallop, B. D. (1995). Flamsteed's lunar data, 1692–95, sent to Newton. *Journal for the History of Astronomy*, 26(3), 237–246. <https://doi.org/10.1177/002182869502600303>

Krishnan, A. (2009). *What are academic disciplines? Some observations on the disciplinary vs interdisciplinarity debate*. Economic and Social Research Council (ESRC) National Centre for Research Methods (NCRM) Working Paper Series, 03/09. <http://eprints.ncrm.ac.uk/783/>

Leonelli S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A*, 374, Article 20160122. <https://doi.org/10.1098/rsta.2016.0122>

- Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.17405bb6>
- Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books. <https://doi.org/10.1515/9780804772891>
- Noy, N., & Noy, A. (2020). Let go of your data. *Nature Materials*, 19, 128. <https://doi.org/10.1038/s41563-019-0539-5>
- NSA surveillance exposed by Snowden was illegal, court rules seven years on. (2020, September 3). *The Guardian*. <https://www.theguardian.com/us-news/2020/sep/03/edward-snowden-nsa-surveillance-guardian-court-rules>
- Oberski, D. L., & Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1), <https://doi.org/10.1162/99608f92.63a22079>
- Parrondo, J. M. R., Horowitz, J. M., & Sagawa, T. (2015). Thermodynamics of information. *Nature Physics*, 11(2), 131–139. <https://doi.org/10.1038/nphys3230>
- Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creator's advantage. *Harvard Data Science Review*, 1(2), <https://doi.org/10.1162/99608f92.fc14bf2d>
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Perelman C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. University of Notre Dame Press. <https://doi.org/10.2307/j.ctvpj74xx>
- Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, 146(3642), 347–353. <https://doi.org/10.1126/science.146.3642.347>
- Popper, K. R. (1960). On the sources of knowledge and of ignorance. *Proceedings of the British Academy*, 46. Reprinted in *Conjectures and refutations: The growth of scientific knowledge* (4th ed.) (pp. 3–30). Routledge and Kegan Paul, 1972.
- Popper, K. R. (1993). *Realism and the aim of science*. Routledge.
- Poovey, M. (1998). *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. University of Chicago Press. <https://press.uchicago.edu/ucp/books/book/chicago/H/bo3614698.html>
- Reid, M. D., & Williamson, R.C. (2011). Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22), 731–817. <https://jmlr.org/papers/v12/reid11a.html>

- Rosenberg, D. (2013). Data before the fact. In L. Gitelman (Ed.), *“Raw data” is an oxymoron* (pp. 15–40). MIT Press. <https://doi.org/10.7551/mitpress/9302.003.0003>
- Rundle, R. T., Vuurboom, R., & Duroy, Y. (2015). End-to-end data provenance. In *SPE Annual Technical Conference and Exhibition* (SPE-174803-MS). Society for Petroleum Engineers. <https://doi.org/10.2118/174803-MS>
- Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data and Society*, 6(1). <https://doi.org/10.1177/2053951718820549>
- Schaffer, S. (2009). Newton on the beach: The information order of *Principia Mathematica*. *History of Science*, 47(3), 243–276. <https://doi.org/10.1177/007327530904700301>
- Scott, J. F. (Ed.). (1967). *The correspondence of Isaac Newton, Volume IV, 1694–1709*. Cambridge University Press (for the Royal Society).
- Shamdasani, J., McClatchey, R., Branson, A., & Kovacs, Z. (2015). Analysis traceability and provenance for HEP. *Journal of Physics: Conference Series*, 664, Article 0032028. <https://doi.org/10.1088/1742-6596/664/3/032028>
- Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226148847.001.0001>
- Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of prediction. *Physical Review Letters*, 109(12), Article 120604. <https://doi.org/10.1103/PhysRevLett.109.120604>
- Tuomi, I. (1999). Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems*, 16(3), 103–117. <https://doi.org/10.1080/07421222.1999.11518258>
- van Rooyen, B., & Williamson, R. C. (2018). A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(288), 8501–8550. <https://jmlr.org/papers/v18/16-315.html>
- Wing, J. (2019). The data life cycle. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.e26845b4>
- Yang, E., Matthews, B., & Wilson, M. (2013). Enhancing the core scientific metadata model to incorporate derived data. *Future Generation Computing Systems*, 29(2), 612–623. <https://doi.org/10.1016/j.future.2011.08.003>
-

©2020 Robert Williamson. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.