Why you can’t model away bias

Preprint: *Modern Language Quarterly* 80.3

Katherine Bode

Quantitative literary studies[1] is often understood as homogenous in its methods; many literary scholars also perceive the field as incapable of contributing literary critical or historical insight.[2] This article contests both perceptions – but not by arguing that quantitative literary research is inevitably sound, justified, or beneficial. Rather, I elaborate and extend a central, ongoing disagreement in the field, with considerable bearing on its current state as well as its future. That disagreement is between what I’ll call its scholarly and statistical approaches: between those who maintain that literary insight depends on critically analysing and historicizing datasets prior to statistical analysis of patterns and trends and those who argue that statistical analysis, in the absence of or with minimal investigation and contextualization of datasets, is sufficient for critical and historical understanding.

One side of this disagreement has been recently and forcefully expressed in Ted Underwood’s *Distant Horizons: Digital Evidence and Literary Change* (2019). Underwood is one of the most esteemed researchers in quantitative literary studies and this is his first book-length contribution to the field. *Distant Horizons* articulates key principles of the statistical approach, and its concept of “perspectival modeling” is more comprehensive than previous frameworks – of distance and laws (Moretti 2005), systems and flows (Long and So 2016), patterns and interpretation (Moretti 2017), or repetition and translation (Piper 2018) – in justifying the application of quantitative and machine learning methods to literary phenomena. Arguing that statistical analyses can stand on their own, Underwood presents the concerns that I and others have expressed about gaps and biases in literary datasets as debates about data “representativeness” that are, in essence, “well-intentioned red herrings” (176). These issues are resolvable, he argues, by statistical means, particularly data sampling and comparison.

I see three main problems in Underwood’s account: it misinterprets the scholarly approach to quantitative literary studies; it misconstrues key statistical principles; and its theoretical framework of perspectival modeling neglects critical, political, and ethical issues implicated in using data to understand literature. Ultimately, Underwood maintains the longstanding assumption that those investigating dominant cultures and cultural forms make generalizable statements, while scholarship from or about the cultural margins is not relevant to everyone and so distracts from “important questions” (184). Yet the challenges of literary data construction and curation apply not only to quantitative literary studies but to how the discipline, in general, conceptualises fundamental issues including the relationship between critical and scholarly practices; the epistemological standing and implications of computer models and algorithms; and the relative importance of aesthetic and political arguments. More broadly, still, the stakes of this disagreement pertain to the growing influence – in most if not all academic disciplines – of a discourse of “big data,” the core heuristic claim of which is that scale liberates us from ethical decisions.

While Underwood’s questions draw on literary and humanistic paradigms, his answers privilege statistical methods and aesthetic sensibilities. In his own words, “the real challenge to large-scale literary analysis is not epistemic or ethical but aesthetic: it is simply hard to write with sweep and verve about thousands of books” (156). To the contrary, interpreting large sets of literary data brings real epistemic, ethical, and political challenges; to be defensible and meaningful, quantitative literary studies needs to keep them to the fore when formulating questions, constructing datasets to ask them with, and building a field in which humanities and statistical ways of knowing coexist and enhance each other.

I *A narrow debate that's had its day?*

Underwood attributes concerns about data "representativeness" to a misreading of an out-of-date justification for "distant reading" (174). Around the turn of the twenty-first century, to make the method interesting and relevant to literary scholars, Franco Moretti cast distant reading as an extension of the literary canon wars. Rather than just adding women, non-white, working-class, and non-Western authors to the canon, distant reading would incorporate *all* literature into its analyses. But Underwood argues that Moretti's actual agenda was comparison rather than completeness: investigating multiple different samples of the literary past to understand it at a new scale. Some quantitative literary scholars miss this nuance in Moretti's argument, Underwood maintains, because they don't understand statistics:

> I suspect critics of distant reading have spent most of their energy on otiose debates about corpus construction because it's the part of this project literary scholars are best equipped to critique. Debates about representativeness have long been central to literary study. But statistical inferences are almost unknown, and we're not yet comfortable testing them. (180)

For Underwood, then, debates about data construction and curation import ideas from the canon wars into quantitative literary studies and wrongly apply a political logic to questions that require statistical methods. Based on this misreading of Moretti, "critics of distant reading" supposedly insist that quantitative literary studies should not proceed unless and until we have "a single master list of representative works." If this list cannot represent all literary works, it must at least be "correctly balanced" in terms of the literary genres, social groups, historical periods, and so on, that it incorporates (178).

But the arguments that Underwood invokes are not calling for such a list. Instead, I have proposed, as has Ryan Cordell (and others not cited by Underwood), that quantitative literary studies should begin by trying, as much as possible, to consider the nature of ontological gaps and epistemological biases in its evidence.[3] These gaps and biases can be random or systematic, and they can arise from multiple sources, including historical conditions, cultural and institutional practices, economic factors, and/or technological processes. Investigating them requires attention to the histories of transmission and "infrastructures of knowledge-making" through which literary data are constituted: a process in which meaning is inevitably transformed, if not lost entirely.[4] One way to describe and manage the losses and changes is to adapt, for the digital context, the bibliographical and editorial theories, practices, and technologies that were developed by the earlier field of textual studies to investigate the analog documentary record.[5] Incorporating editorial principles into quantitative literary studies will often involve statistical investigations (for instance, of the proportions of particular documents that are translated into literary data; or the reliability of Optical Character Recognition when applied to a particular collection). But these inquiries would be couched in a bibliographical and editorial understanding of literary works as distributed across (though not reducible to) a vast network of specific material objects; and they would necessarily precede, rather than being folded into, statistical investigations of historical trends and patterns.

In this vein, Cordell (2016: 201) argues "that scholars must understand mass digitized texts as assemblages of new editions, subsidiary editions, and impressions of their historical sources, and that these various parts require sustained bibliographic analysis and description." By claiming these "complex, palimpsestic, editional" assemblages as evidence for literary studies, he challenges the "myths of surrogacy – or worse – replacement" that surround mass-digitized collections. In *A World of Fiction: Digital Collections and the Future of Literary History*, I propose that quantitative literary researchers ground statistical analyses in "scholarly editions of literary systems." A literary system is some collection of documents (and/or people, positions, institutions, and so on) that related to, and generated literary effects in relation to, each other in the past; a scholarly edition models an ideational concept – typically a literary work, in this case, a literary system – by demonstrating and justifying its interpretation of documentary evidence.[6] Where a scholarly edition of a literary work fulfils its purpose with a reading text and critical apparatus, a scholarly edition of a literary system might accompany a curated dataset with a critical essay exploring the selections, assumptions, and uncertainties about evidence underpinning the model.[7]

In contrast to what he sees as unrealistic demands for agreement on a single, "correctly balanced sample," Underwood (2019: 11) argues for openness to diverse forms of balance responsive to publishing patterns and library holdings:

> For instance, one could argue that it is actually appropriate to study a sample that represents 60% of nineteenth-century literary production but only about 20% of titles published in the twentieth century, because the long tail of publishing becomes vastly longer over time, without becoming more important. If we take that view, the existing balance in libraries might be roughly the balance we want. (178–79)

But the idea that adapting bibliographical and editorial principles to quantitative literary studies intends to create a dataset (or even datasets) with equal numbers or proportions of different types of literature across all times and places overlooks a fundamental premise of those arguments: that transformations (including very radical, as well as minor, ones) occur at every stage in the history of transmission from literature in the world to literary data. Accordingly, it would be highly unlikely – if not impossible – ever to construct a dataset in which all types of literature from all periods are equally represented or "balanced." The aim, instead, is to characterize, as much as possible, the transformations that have produced the available evidence. The resulting dataset might encompass substantially more nineteenth- than twentieth-century literature. But a 60/20 dataset based on the principles of textual scholarship would be conceived differently from the 60/20 dataset that Underwood advances. The latter describes the literary past, or some proportion of it: it is a mirror, even if it reflects only part of the whole. By contrast, adapting bibliographical and editorial approaches to mass-digitized collections is intended to position any resulting dataset as an argument about the relationship between literary phenomena and data, where the available evidence is, by definition, contested.

The scholarly approach to quantitative literary studies thus differs from the statistical one in terms of where computational modeling begins. Underwood (2019: xii, fn 3) aligns *Distant Horizons* "with several recent arguments" by Andrew Piper and Richard Jean So, who maintain that computational models are not "mimetic representations of the world" (Piper 2017a: 652; see also Piper 2018: 11, 19) or "summaries or reports about data, [but] … mechanisms with which individuals reason and think" (So 2017: 670). But where Underwood and So (though not Piper, as I'll discuss) describe modeling as a means of reasoning and thinking that is done with literary data, the scholarly approach takes modeling to encompass reasoning and thinking about and with literary data. It thus aligns with a tradition in digital humanities of theorizing modeling as occurring "at two different stages, in the creation of the model and in its application and successive manipulation" (Ciula et al. 2018: 345; see also Buzzetti 2002; Flanders and Jannidis 2016; McCarty 2005).[8] A dataset of digital texts is not simply found – or extracted, harvested, downloaded, mined, and so on – prior to being modeled using computational methods; rather, modeling is the means by which literary concepts and artefacts are both made computable and computed.

Bibliography and scholarly editing resonate with the view of modeling as a non-positivist framework for reasoning. Where many people perceive bibliographies as facts about literature and scholarly editions as especially accurate versions of literary works, in contemporary textual studies, they are conceived as "embodied arguments about textual transmission" (Eggert 2009: 177) or "hypothetical platform[s]" for historical enquiry (McGann 2005: 203, 230). They offer mechanisms with which not only individuals but the discipline can reason and think about literary phenomena. Rather than asserting a one-to-one correspondence between literary works and literary data, in adapting these non-positivist approaches to the digital documentary context, the scholarly approach resists Underwood's conception of mass-digitized collections and literary datasets as even partial mirrors of the past. In the process, it aims to connect modeling in quantitative literary studies to both scientific and humanities traditions of inquiry.

Underwood mistakes the effects as well as the aims of focusing on data modeling. Like Piper previously (2017b), he argues that demands for data representativeness stifle computational innovation and weigh quantitative literary studies down with the requirement, if not for a single, authoritative list of literary works, then certainly for stable and permanent datasets. But this claim is a "red herring," from both a pragmatic and a conceptual standpoint. Pragmatically, no one can impose a stable model – whether a text or a dataset – on a field. The continuing use of any dataset (as with any bibliography or edition) would, and should, be a function of its usefulness. If it meets a need for literary scholars it will be used and if it doesn't, it won't (and shouldn't) be. With digital publication, models are also updatable: for instance, *To be continued: The Australian Newspaper Fiction Database* maintains a version of the curated dataset that underpins *A World of Fiction*. But it also publishes thousands of other stories, discovered in the initial analysis of *Trove*'s digitized newspapers or added, subsequently, by members of its crowdsourcing community.[9]

The more abstract concern might be expressed as the idea that a stable or falsely permanent whole will be privileged over multiple exploratory samples. Yet a sample requires a "whole" from which to be drawn. The absence of a framework for defining wholes renders Underwood's conception of samples unstable. For instance, he writes that "quantitative inquiry about literary history has never, in reality, limited itself to a single representation of the past. Sometimes distant readers take the whole library as it stands, but often we select subsets, in order to contrast them" (Underwood 2019: 175). Here, "a single representation of the past" becomes "the whole library as it stands," and "the whole library" becomes literary history. Subsets or samples are, thus, both of the literary past and of the whole library, and perhaps because of this ambiguity, are selected primarily with reference not to either of those wholes, but to each other: "we select subsets, in order to contrast them."

A number of quantitative literary researchers explore the field's difficulty with defining the relationship between sample/s and whole/s. *Stanford Literary Lab Pamphlet 11*, "Canon/Archive," differentiates an ideational whole ("the published," or all literary works made public in history) from a material one ("the archive," or the portion of what was published that has been preserved and is now increasingly digitized), and both from "the corpus" (the segment of the archive selected for a research question). Although their terminology ("the archive" – like Underwood's "the whole library as it stands") conflates a diverse and distributed multitude of collections and documents into a single, stable entity, the distinction the authors draw between whole/s usefully recognizes that we never investigate the literary past directly, only some forms of some of its preserved records. However, the authors immediately undercut their distinction by claiming to identify "bias" in their corpus through statistical comparison of it with analog bibliographies: as if bibliographies materialize an ideational whole (Algee-Hewitt et al. 2016: 2).

Piper criticizes the assumptions underlying this approach in his introduction to *Enumerations: Data and Literary Study*. Discussing "representativeness," or the "circular nature [of] computational reading," he notes that quantitative literary studies lacks recourse to "some stable, knowable whole against which to limit one's bias" because "the ideas, meanings, and feelings of literature are never capturable once and for all, but are themselves always forms of … represented reality" (Piper 2018: 9). Elsewhere he observes that while statistical tests can measure the functioning of the model ("the extent to which what we are observing exceeds the boundaries of chance"), they cannot confirm "whether the model is an appropriate approximation of the phenomenon that one is claiming to observe" (Piper 2017a: 655). Piper's terminology is somewhat confusing, given that he uses a statistical term – representativeness – to argue against a statistical approach to defining literary samples. Yet his argument aligns with the central premise of the scholarly approach in quantitative literary studies: that modeling of as well as modeling with literary data "foreground[s] the constructedness of knowledge and the observer's place within it" (Piper 2018: 9). While Piper acknowledges that all "observations are dependent on the collections … assembled for" the purposes of inquiry, he does not explore those dependencies, differences, or interactions. Instead, his *PMLA* article on modeling argues that the only way to ground models is in the judgment of the modeler as to whether the operations and outputs make sense or are "appropriate" in relation to what's already known about the phenomena being investigated (Piper 2017a: 656). And in his book, he leaves it to "future work … to explore the extent to which these and other observations depend on different kinds of collections [and] to better understand the interactions between methods and archives" (Piper 2018: 10).

In both, Piper shifts the focus away from the constructedness of the relationship of samples to wholes. And in his book, he compounds this decision by arguing that computational models keep "[p]art and whole always … simultaneously in view." By transforming "part" into words and "whole" into an amalgam of collected corpus, literary formation, and computational model, Piper positions computational models as well-calibrated mechanisms for generalization that "allow us to understand how an individual word (or feature) is being used relative to a larger context (the twentieth century, the nineteenth-century novel), just as they allow us to understand how an individual word (or feature) is being used in the context of other words and features" (Piper 2018: 16). Ultimately, although conscious of what he calls the "positivism" of models – or their status as "represented reality" – in imagining them as self-contained mechanisms Piper privileges a non-referential expression of literary meaning. He quotes Gilles Deleuze's description of the diagram to explain literary models, arguing that both produce "a new kind of reality, a new model of truth. … mak[ing] history by unmaking preceding realities and significations" (Gilles Deleuze, *Foucault*, cited in Piper 2018: 12).

As the *Stanford Literary Lab Pamphlet 11* recognizes, the whole to which we refer is both ideational (the literary phenomenon we aim to understand) and material (the inevitably imperfect documentary evidence accessible to researchers, because collected by particular institutions and remediated in specific digital collections). We create corpora (or curated datasets or bibliographies of mass-digitized collections or scholarly models) in order to understand conceptual wholes using evidence derived from material ones. And as Piper recognizes, because we have access only to incomplete materializations of the ideal or conceptual whole, there is no direct or absolute way to determine how much, or what parts, of the past we construct for analysis. While these researchers ultimately back away from the implications of their arguments, the scholarly approach is an attempt to foreground – and devise theoretical and technical frameworks for investigating, embodying, and managing – the complex distinctions and entanglements of ideational and materialized wholes.

Cordell's bibliographical approach proposes to embody an argument about the relationship of a mass-digitized collection (a materialized whole) to the documentary context it was drawn from (a conceptual one); my scholarly edition investigates this relationship so as to create a dataset (a materialized whole) that models a literary system (a conceptual one). In neither case is the materialized whole a sampling from the conceptual one, because that would mean that the conceptual whole was an actual rather than an ideational phenomenon. But nor is the conceptual whole a construction of the material one, because that would mean that the literary phenomenon investigated had no historical meanings or effects. Instead of one representing the other, the materialized whole responds, and is a response, to the ideational one. As Karen Barad (2007: 152) puts it, in a principle as essential to textual studies and digital humanities theories of modeling as to her agential realism, "materiality is discursive … just as discursive practices are always already material." Because the scholarly approach then offers its materialized wholes for investigation – in total or in parts – it seeks to enable the two forms of quantitative literary studies that Underwood defends: exploring "the whole" or "select[ing] subsets."

Ultimately, the debate about data representativeness (or data modeling) in quantitative literary studies goes much deeper than the contrast Underwood draws between those who misunderstand Moretti's rhetoric (and replay the canon wars with digital libraries) and those who subscribe to what is supposedly Moretti's actual agenda (and practice sampling and comparison). In terms of disciplinary traditions, Underwood's insistence that modeling with literary data is non-positivist inquiry while modeling of literary data is naïve and reductionist perpetuates the longstanding disciplinary division of literary critical and scholarly practices (and marginalization of the latter as "narrowly technical") (McGann 2014: 2). This was the same division I identified, in an earlier article in this journal, as the reason for Moretti's focus on analysing literary data, to the extent of minimizing or ignoring its construction (Bode 2017). But really, the situation is more complex than a single disciplinary inheritance, because the division of statistical and scholarly approaches also has a geographical dimension, with the former prevalent in North America and the latter in Europe and Australasia. Where many North American researchers and labs focus on exploring machine learning and algorithmic methods for analyzing textual corpora (that are often digitized texts extracted from one or more proprietary collections), in Europe multiple digital projects devote collaborative effort to constructing datasets that model the production, circulation, and transmission of literature within and across particular historical, regional, and linguistic contexts.[10] Certainly, there are important crossovers; but the debate about quantitative literary studies in North America is perhaps insufficiently aware of its difference from approaches elsewhere.

Beyond these disciplinary and geographical dynamics, there is another – arguably more interesting, certainly more challenging – reason why "representativeness," and questions of relationships between world, data, and model, have become a sticking point in quantitative literary studies. It relates to a "crisis of representation" (Noth 2003) experienced across the humanities and social sciences in this era of radical technological change. In *Friending the Past: The Sense of History in the Digital Age*, Alan Liu (2018: 4) proposes that cultural criticism is transitioning from a regime of "*rhetoric-representation-interpretation*" to one of "*communication-information-media*." Previously, "representations" of culture had rhetorical power (or not) and could be interpreted. But the new fundamentals of contemporary knowledge production, including data and models, do not align well with this framework. Data reports on reality, but in "some offset way (inferred, derived, sampled, or reduced)." Rather than describing reality "data is a *parameter* of how reality can be addressed, represented or interpreted – that is, a measure *to the side* ('para') of rhetorical measures, representational frames, and interpretive paradigms." Equally, models are "not exactly like any of the[se] concepts …. Models, uncannily, are all, and none, of the above." In Liu's words, as our old regime is "hollowed out … from the inside," we are left only with "abstract terms, like *communication*, *information*, and *media* [that] increasingly mask the fact that we don't really know what we are doing, let alone signifying, when speaking and listening, writing and reading, and so on" (5). Another way to express this challenge is in N. Katherine Hayles's (2004: 68) terms of an awakening. While the "unrecognized assumptions" of "five hundred years of print" led us to take the materiality of literature for granted, the digital age awakens us to the "importance of media-specific analysis" and the entanglement of instantiation and signification.

This shift, however we understand it, underpins the debate about "representativeness" in quantitative literary studies. Our "crisis of representation" is perhaps heightened because we engage with data and models from the standpoint of a discipline that has long been unsettled by the uncanniness of literary works: their status as never just interpretive (rhetoric, representation) or informational (communication, media). Having long struggled (and failed) to connect text to book, we now aim to explore tens of thousands, or millions, of texts/books, while connecting them to new, strange fellow travellers: data and models. Neither the scholarly nor the statistical approach solves this challenge. And scholarly projects also have clear practical and rhetorical limitations. Most obviously, they are more time-consuming than harvesting and using existing records from mass-digitized collections. The insights they propose to offer are more constrained than those claimed by statistical enactments: oriented to particular collections, communities, and/or cultures rather than to broad swathes of the literary past. Perhaps they will ultimately prove too "heavy" with the "associations" of "print culture" (Price 2009: np) to usefully serve in distributed, dynamic, and increasingly monopolistic digital environments. Doubtless, scholarly interventions are less uplifting than statistical ones, foregrounding the uncertainties of literary history rather than claiming to uncover its true patterns and trends. Yet with all these problems, the scholarly approach attempts to reflect on and respond to what Liu (2018) calls a regime shift and Hayles (2004) an awakening to emergent mediality. Its slowness, specificity, and emphasis on uncertainty are, I argue in what follows, part of the ethical framework that the scholarly approach brings to inquiry into the literary past, in the uncanny conditions of the (post)digital present.

II *A statistical solution?*

If there is a problem, in quantitative literary studies, with "single master list[s] of representative works" (Underwood 2019: 178), it's the dominance of large digital libraries, such as *HathiTrust* and *Google Books*, as proxy models for past literature.[11] Of course, even the largest digital collections are inevitably selections from a much broader documentary record, with *Google Books* and *HathiTrust* based on the holdings of certain elite American and British university and major public libraries. As he has done in the past,[12] in *Distant Horizons* Underwood acknowledges that the nature of these collections can influence the results of quantitative analysis.

Yet Underwood maintains that statistical methods, particularly data sampling and comparison, can identify and resolve distortions in literary data derived from those collections. He finds support for this argument in data practices in the social sciences:

> Social scientists don't, after all, begin by stabilizing a single list of people who will count as a representative sample for all future inquiry. They assume that there are far too many questions about human behaviour for any one sample to suffice. Some questions will need detailed evidence about a small group of people who are all the same age; other questions will need a broad sample ranging across ages, nations, and socioeconomic backgrounds. Although framed differently, the results of these two studies may eventually interlock and become mutually illuminating. Similarly, in literary study, researchers who focus on reception may want a sample that emphasizes popular works and includes every edition of them. Researchers interested in literary production may need a broader sample limited to first editions. There is no need for scholars to decide which of those samples more correctly represents the past. Instead, we can compare them and explore the differences. (177–8)

Although Underwood seems to align statistical methods in social sciences and quantitative literary studies, his account differentiates them. In the social sciences, comparison is a possible outcome, but it's a question for the future, not a basis for ascertaining the reliability of a sample. Datasets for literary study, by contrast, are assumed to refer to an actual system. Different samples might emphasize different parts of this collective whole – for instance, its popular works or its first editions – but drawn from that whole, they are inevitably and, in this passage at least, immediately comparable.[13]

Underwood reiterates the comparability of literary datasets throughout *Distant Horizons*. I have already quoted him arguing that "we select subsets, in order to contrast them" (175); he also affirms Benjamin Schmidt's claim that "the goal is not to construct an unbiased sample but to understand each 'source *through* its biases'" (178); notes that the solution to imbalance in digital libraries is "not to stabilize a single sample … but to try different possibilities and compare them" (179); and proposes that, "Once researchers have created digital collections of several thousand titles, it becomes unnecessary to decide on a single representation of the past, since digital collections are easy to subdivide and rebalance" (177). Underwood objects to the technological determinism of phrases such as "big data" (164). Yet the assertion of ease in this final recommendation reflects his conception of quantitative literary studies as an agile and dynamic movement across multiple, interoperable datasets; the kind of approach that has been critiqued for imagining that big data could be a "shortcut to ground truth" (Crawford, Miltner, and Gray 2014: 1670).

Even if literature were a single system, comparison of samples wouldn't indicate (at least) two types of selection bias. The first is exclusion bias: where features of the population affect what is sampled. Say you have a barrel containing red and green balls. You sample from the barrel by taking ten balls, tallying up the number of red and green, and throwing them back in. After multiple samples where you discover an approximately equal number of red and green balls, you surmise that there is a high probability the barrel contains an equal number of both. But what if the red and green balls had different qualities that you didn't know about: for instance, the green ones were heavier or more slippery, so were less likely to be captured by your sampling process?[14] Comparison, in this case, hasn't illuminated features of your underlying population, and it won't help to sample from another barrel containing red and green balls when one type of ball is more resistant to sampling than the other. Rather than an unfortunate and rare circumstance, it's highly likely that different literary collections will manifest the same systemic biases – for instance, the relative absence of women authors, or of particular regional or national literatures.

Comparison, alone, also won't indicate the nature, or extent, of bias in our sampling methods. Suppose you were interested in investigating characteristics of realist fiction in different national contexts so you asked a research assistant to identify prominent titles. Simply comparing the resulting samples won't determine which differences between them, if any, relate to national trends in realist fiction and which, if any, to variation in the knowledge or assumptions of your research assistant. Yet the basis of these differences is essential to your conclusions, because in one case they provide interesting grounds for critical discussion, whereas in the other they precluded effective comparison.[15] One response to this hypothetical example might be that using computational methods, for instance, classification algorithms, to define samples means that identical parameters are applied in all instances. But this understanding treats data and algorithms as divisible and definable in terms of discrete effects, when they are, rather, "moving targets" in constantly changing networked environments (Ananny 2016: 108), "intertwined in the production of social and cultural meaning" (Hoffman 2019: 909). More basically, the assumptions embedded in algorithms used for sampling in quantitative literary studies are often not discussed, and certainly haven't been adequately theorized, even though different methods produce "utterly different sample[s]" (Guldi 2018: np; see also Shore 2018).

Underwood argues that critics of his approach must demonstrate the consequence of any problem they claim to identify:

> Having gone to great lengths to make it possible for skeptics to prove me wrong, I think it is fair to expect that they actually do so. … it is no longer enough to draw up a list of scattered errors and omissions that might – who knows? – have altered the book's conclusions. Armed with the author's code, a critic who finds a genuinely consequently oversight should be able to take the next step and demonstrate that it was in fact consequential. (183)

This insistence has its own political implications, which I'll return to. But first, in the spirit of the request, I'll use some Australian literary data to demonstrate why sampling and comparison do not, as Underwood argues, preclude gender bias from his conclusions about literary authorship.

In chapter 4 of *Distant Horizons*, Underwood employs a dataset derived from *HathiTrust* to identify a decline in the proportion of English-language fiction by women from around 50% of titles in the late nineteenth-century to roughly 20% by 1970, before a return to women writing a bit under half of all titles at the end of the twentieth century. Noting that elite university and public library collections might simply have collected more books by men than women, Underwood seeks to test whether this bias influenced his results by comparing proportions of male and female authors in *HathiTrust* to those in manual samples from four years of *Publishers' Weekly* listings. Because the *Publishers' Weekly* samples indicate an even more dramatic fall in women's writing, Underwood claims that the comparison "addresses … doubts" about "how well … those collections represent the wider world of fiction" (135).[16] While *Publishers' Weekly* incorporates a great deal of popular fiction that does not figure in academic collections, it indexes almost no titles by even the most prominent and prolific popular romance fiction publisher of the twentieth century, Mills and Boon. Women authors predominate in this genre and the period of decline in the proportional representation of women authors and characters in Underwood's results – the 1950s to the 1970s – also corresponds with the heyday of popular romance fiction publishing.

To explore how much the exclusion of romance fiction might have influenced his results, Figure 1 amends Underwood's Figure 4.9, using data on Australian women's novels from 1945 to 2000.[17] If American and British women wrote romance fiction at similar levels to that recorded in the Australian context, then rates of fiction by women would remain relatively flat through the 1940s, 1950s, and 1960s, at equivalent levels to the turn of the twentieth century.[18] There would be a decline in women's fiction in the 1970s, but a less dramatic one than Underwood reports, and the general trend across the twentieth century would be fairly stable or growing. The Australian data thus undermine Underwood's conjecture that the decline in female characterization was due to a decline in women authors of fiction and exposes the fragility of inferences made on the basis of literary datasets that have not been adequately historicized. I am not saying that my results show what actually happened; I am using them here as another sample. My point is that comparing two – or three or four or five or however many – samples cannot exclude the possibility that they manifest similar biases; nor can it define the degree or limits of bias introduced by sampling methods.

Working with large datasets tends to increase confidence in the capacity of statistical methods to answer questions while decreasing recognition of gaps or other biases, especially in datasets created for purposes other than the specific research question addressed. These impressions are endemic in a range of fields that face the challenge of reconciling established research practices with new, much larger, datasets. As Daniel McFarland and Richard McFarland (2015: 1) comment with regards to the social sciences,

> since these large data sets easily meet the sample size requirements of most statistical procedures, they give analysts a false sense of security … [with the result that] most analyses performed on Big Data today lead to 'precisely inaccurate' results that hide biases in the data.

While Underwood sees data modeling and curation as the opposite of statistical inference, ultimately, the principle known as GIGO (garbage in, garbage out) applies to both. We need to be wary of claiming precision for results that are artefacts of distortions in digital libraries, rather than effects of the literary phenomena we wish to understand.

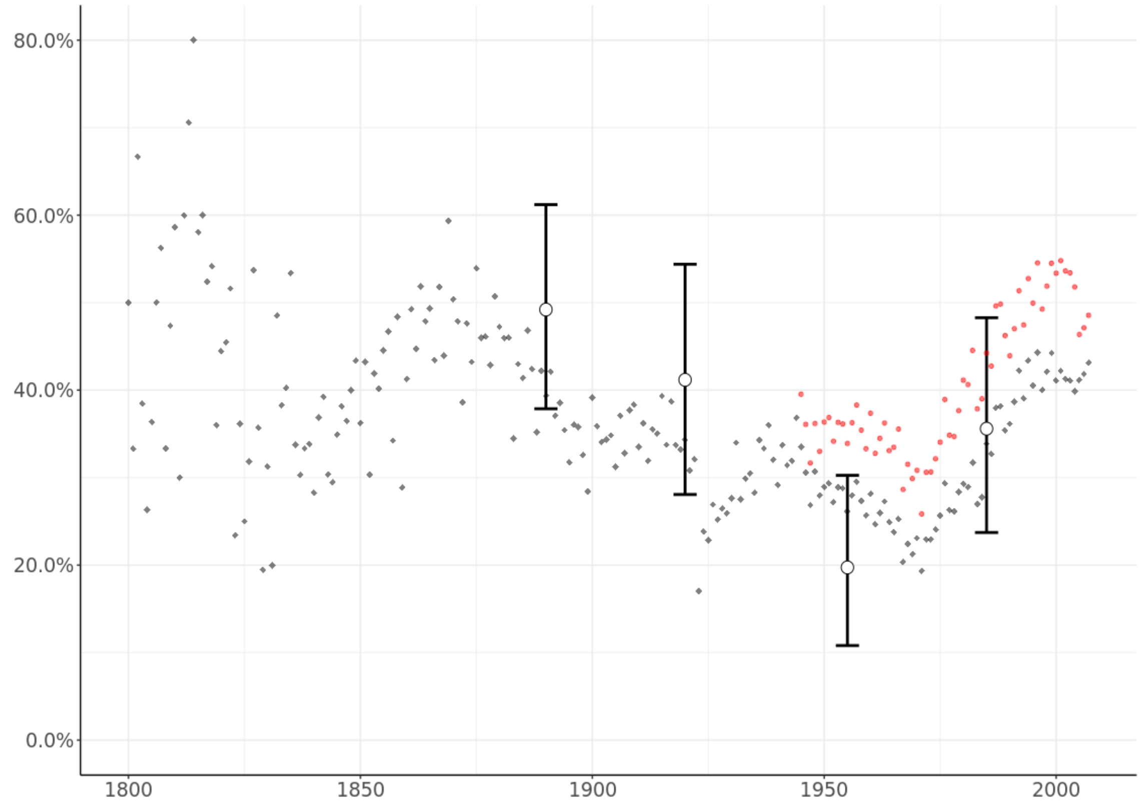

Figure 1 Percentage of English-language fiction titles written by women in *HathiTrust* (diamonds) and *Publishers Weekly* (error bars) (Underwood 134, Fig. 4.9). Circles estimate percentages for *HathiTrust* based on rates of romance fiction recorded for Australian novels. (Results exclude reprints and juvenile titles.)

III ***Perspectives on bias***

Underwood's sampling and comparisons are often more intentional and "explicitly limited" than is the case with analyses of overall trends (2019: xx). Indeed, he has been instrumental in demonstrating the critical potential of using machine learning models to categorize and compare different literary samples, and his example has been widely emulated. These analyses typically identify a literary phenomenon, propose a way of measuring it, create a sample to represent that phenomenon, use the sample as a training set for a machine learning model, then test the model's classification against a random sample. For instance, in chapter 3 of *Distant Horizons*, Underwood seeks to understand perceptions of literary value, which he argues can be measured in terms of literary works reviewed in particular British and American literary journals. He selects a small number of reviewed prose and poetry titles from *HathiTrust* and uses that sample to train his model. Underwood then tests its classifications against what he presents as a random sample (also extracted from *HathiTrust*) to demonstrate his model's capacity to distinguish reviewed from non-reviewed works. This method is not uncontroversial,[19] nor is it immune to the selection biases discussed above. But Underwood also investigates possible confounding features of his dataset (such as the inclusion, in *HathiTrust*, of much more obscure American than British writing). With attention to the histories and features of particular datasets, and where there is adequate evidence for the questions asked, such use of machine learning has promise as a means of reducing distortions arising from the history of transmission of literary datasets.

But as with his approach to overall trends, with machine learning Underwood wants to claim statistical methods as substitutable for scholarly ones. To this end, he proposes using perspectival modeling to identify biases in literary datasets. The machine learning algorithms that he employs for this purpose present "a huge problem for institutions," like banks and courts, "that are expected to be neutral arbiters" because they "pick up the assumptions of prejudices latent in a particular selection of evidence" and "absorb the biases of people" whose decisions that model trained on." In contrast:

> When we're reasoning about the past, … our aim is usually to acknowledge and explore biases, not to efface them. Understanding the subjective preferences implicit in a particular selection of literary works, for instance, may be exactly the goal of our research. For this kind of project, it is not a problem but a positive advantage that machine learning tends to absorb assumptions latent in the evidence it is trained on. By training models on evidence selected by different people, we can crystallize different social perspectives and compare them rigorously to each other. (xv)

These people might include modern researchers, who explicitly embed their decisions about literature in training supervised models, or historical readers whose literary judgements – for instance, whether a novel belongs to a particular genre, or a preference for concrete physical description – are learned by the model as it trains on data manifesting those judgments.

While Underwood notes, at one point, that "*bias* might not be exactly the right word to apply in this context" (94), his account of perspectival modeling folds a range of social and cultural attitudes – "assumptions," "prejudices," "biases," and "preferences" (elsewhere, "subjective beliefs" [xv], "historical vantage points" [xviii], "different perspectives" [35], "subjective evidence" [141], "implicit associations" [147], "historical vantage points" [192–3]) – into a single notion of "perspective." Machine learning then allows these "perspectives" to be precisely measured (crystallized) and compared, such that predictive models enable researchers "to represent perspectives of other eras, in a form solid enough to allow one perspective to be compared rigorously to another" (37). Underwood also equates these models to "sensors that can measure and compare implicit associations from other periods" (193); evokes the geometrical notion of "parallax" to describe the form of measurement enabled (xv); and aligns perspectival modeling with Pierre Bourdieu's empirical research, arguing that the method "allows us to pose the same questions about a century where people are no longer able to reply to surveys" (98). Biases that in textual studies are intrinsic to any document or dataset (as residues of acts of witness or testimony) become discrete and classifiable perspectives that can be extracted and/or implemented, thereby making a machinic, as opposed to an ethical, response feasible and justifiable. Opposing the agility and dynamism of his approach to the unrealistic demands of data construction and curation, Underwood argues that perspectival modeling "make[s] it easy to look at the same evidence from several different angles and measure the interactions between social categories." In particular, it "measure[s] the effects of imbalances" – he cites historical biases against women writers – "without forcing our sample of the past to be unrealistically balanced" (94).

The problems with this framework are multiple. Conceptually, it renders equivalent the range of factors that can shape a literary dataset, so that aesthetic and formal judgments operate on the same plane as the latent assumptions of the modeler and the multiple factors – historical, technical, economic, cultural, political, geosocial – affecting what literary documents are collected and how they are documented and remediated. In practice, because the only "perspectives" Underwood investigates are judgments about the nature of the literary, he ignores other influences while claiming to manage them by modeling. The idea that literary scholars can "acknowledge and explore biases" because those "perspectives" are embedded in literary datasets accords machine learning a transparent, heuristic function that creates a logical tautology. On this account, perspectives are present in the results because (and only if) they are found, and they are absorbed by the machine learning model because (and only if) they are discernible.[20] In arguing that it is "easy" to measure inequality in this way, Underwood claims for "big data" methods an outcome that has eluded generations of humanities researchers, and accords quantitative literary researchers a level of insight into machine learning models that he believes governments and corporations cannot achieve.[21]

In proposing perspectival modeling and in foregrounding subjective aspects of quantitative analysis and machine learning, Underwood seeks an alternative to the positivism that has often hovered over, or been explicitly asserted for, quantitative literary studies: "there is no single authoritative account of the past," so "instead of seeking an objective metric to solve the problem, the best course of action is often to consider different perspectives" (27). This judgment echoes Piper's (2018: 12) notion that, because data and models are constructions, computational models can be put forward as "a new kind of reality" that "make[s] history by unmaking preceding realities and significations." Because there are no unbiased – or unconstructed – collections, these statements imply that it is acceptable to proceed as if all biases and constructions are equally valid and deserving of expression and inquiry. But as Donna Haraway (1988: 584) shows, totalism and relativism are not opposites but "perfect mirror twin[s]." Both are unsituated views, even if one claims to see the world from a single, authoritative perspective, and the other from "everywhere equally." This twinning perhaps explains why Underwood's relativism regarding datasets is combined with a totalizing belief in the ability of machine learning to transparently and rigorously absorb and measure perspectives.

Underwood's proposal to replace data modeling with sampling and comparison, alone or in concert with perspectival modeling, is ultimately inseparable from his view of the literary past as a single system of interlocking parts. But the literary past, and (quantitative) literary studies, are spaces of highly differentiated knowledges, opportunities, constraints, resources, and possibilities, and these must be considered in any arguments about, or propositions for the future of, the discipline.[22] At the very least, many areas of literature are not well served by existing mass-digitized collections like *HathiTrust*. In my work, for instance, I have found that most novels, novellas, and short stories by Australian authors were never published as books and were not recorded in the national bibliography. Those that were published as books appear to have made this transition, at least partly, in the context of decisions relating to racial bias.[23] Even the Australian works that were published as books are much less likely to appear in large digital libraries than the works of American and British authors; and when they are digitized, are less likely to be usefully indexed or to receive the quality assurance accorded to dominant literary traditions (Barnett 2015).

Australia is an Anglophone, Western, wealthy nation with a relatively short written documentary record, where there has been extensive public investment in digitization. The situation must be worse for many other literary cultures, for which the documentary record is often hardly or poorly digitized, or not at all. Underwood's argument that quantitative literary studies should rely on sampling and comparison, perspectival modeling, and existing, large digital collections is only conceivable from the standpoint of well-resourced and culturally dominant areas of literary research. The debate about data, in other words, is not "a residual controversy, hanging on beyond the point where it makes sense" (174) since we are far from having a single literary system that is fully (or even adequately) modeled by existing digital libraries.

The mistaken view of quantitative literary studies as an equal playing field connects to other aspects of Underwood's argument. For instance, identifying "long timelines," particularly "century-spanning" ones, as the proper focus of quantitative literary studies overlooks the political implications of this orientation: namely, that the existence of digitized records encompassing centuries is an effect of historical and contemporary advantages that pertain only to certain literary contexts (Underwood 2019: 179).[24] Given the unequal grounds on which quantitative literary studies operates, we might also reconsider Underwood's proposal that researchers should cease "otiose" (180) discussions of data representativeness and, if they want to contribute, either engage with his dataset or provide their own to rebut his claims. His invitation is no doubt intended to encourage dialogue, but it makes access to extensive, century-spanning datasets a precondition. This requirement focuses debate in quantitative literary studies on dominant literary traditions and overlooks the more limited resources and opportunities available to many participants and communities. Even the political arguments that Underwood then makes – for instance, that women are underrepresented in books by men and women – are limited because inattentive to inequalities as they relate to the documentary and disciplinary contexts in which he works.

Underwood ultimately advocates a "compromise" between scholarly and statistical approaches. Yet that compromise privileges statistical analyses as necessary "to motivate that work" of improving digital collections by showing that "our collections, as they stand, are already capable of addressing important questions" (184). Given that the construction and curation of literary data is now often a precondition for literary study, quantitative or not, scholarly work should not be reduced to supporting quantitative analyses; nor should computational methods, such as text mining and machine learning, be conceived only as tools for statistical arguments. To the contrary, both scholarly and statistical arguments are necessary – and should be deployed – to make the (digital) literary record better: less reliant on multinational corporations that have a vested interest in maintaining the status quo;[25] more responsive to the historical and contemporary inequalities that marginalize certain authors, cultures, and groups; more aware and respectful of the labor that goes into creating digital collections; and more conscious of the colleagues and communities that have less privileged access to the literary past.

I am by no means the first to suggest that quantitative literary studies struggles with political issues. The field's problems with gender inequality have particularly been emphasized. Speaking about "Distant Reading after Moretti," Lauren Klein (2018: np) points to systemic problems relating to the use of "computational methods that derive from statistical modeling and computational linguistics … applied to analyze texts at scale," arguing that their "unduly masculinized rhetorical positioning" creates difficulties in dealing with "conceptual issues that relate to women" and other marginalized groups.[26] Shawna Ross (2018: 212) describes how collaborative bibliographical and editorial projects, especially feminist ones, are excluded from accounts of quantitative literary studies that focus on "the coolness of one's tool, the bigness of one's data, or the goodness of one's intentions." More broadly, Natalia Cecire (2011) writes of digital humanities that its epistemological and ethical claims are often normative. In Underwood's book, both the masculinized rhetoric he applies to machine learning and his marginalization and feminization of scholarly approaches reflect an inattention to epistemology and ethics that undermines his genuine concern for gender inequality in literary history.

The broader, normative claim of *Distant Horizons* is common to all (quantitative) literary research that proposes to survey literary history using datasets derived from one or more dominant literary cultures. It's more than disquieting, decades after the discipline stopped (or at least agreed it was necessary to try to stop) propagating narratives about British and/or American literature as generalizable to all literary cultures and communities, to find this rhetoric re-emerging in statistical approaches to quantitative literary studies. And this rhetoric is reanimated by its resonance with contemporary epistemological claims, particularly that datafication is a "new gold standard of knowledge" (van Dijck 2014: 201) and that data and digital methods offer freedom from local constraints. This latter idea, which is called "digital universalism" in critical data studies (Milan and Treré 2019), has been characterized as "a new kind of colonialism in which practitioners at the 'periphery' are made to conform to the expectations of a dominant technological culture" (Loukissas 2019: 10; see also Risam 2019).[27] The recent history of the discipline has (or should have) taught us that relativism (everyone has their own perspective) is ethically moribund and politically disastrous (Latour 2004), and that it's insufficient – as was attempted during the canon wars – to respond to systemic failures with discrete additions (more women, more working-class, more non-Western writers). We need instead, as Anna Lauren Hoffman (2019: 910) states, to pay "sustained and iterative critical attention. … to the kinds of worlds being built – both explicitly and implicitly – by and through the design, development, and implementation of data-intensive, algorithmically mediated-systems."

There are many ways to include diverse literary researchers, cultures, and communities in quantitative literary studies; all begin with recognizing and reassessing the political and ethical implications of our approaches and assumptions. Literary datasets are not ontologically obvious wholes, but events: complexly contingent sequences and networks of action and reaction, arising from and altering how the human record is put together. Constructing and curating, or modeling, literary datasets – including any interventions we make in existing ones – involves political, social, and ethical decisions, the outcomes of which are neither inevitable nor inevitably justified. Creating (or adapting) frameworks to explain and demonstrate the effects of those decisions certainly does not mean that all of them will be justified – or ethical. But it does create the conditions for recognizing that our engagements with data have consequences for knowledge, for the future of the discipline, and for the others with whom we stand in embodied, situated, and inevitably partial relations. It means taking responsibility and being accountable for those decisions and assumptions and their capacity either to sustain or inhibit "the possibility of webs of connections called solidarity in politics and shared conversations in epistemology" (Haraway 1988: 584).

References

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. "Canon/Archive. Large-scale Dynamics in the Literary Field." *Stanford Literary Lab Pamphlet* 11, January: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Ananny, Mike. 2016. "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness." *Science, Technology & Human Values* 41 no. 1: 93–117.

Barad, Karen. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Barnett, Tully. "The Fate of Australian Texts in Mass Digitization Projects." *Literary Studies Convention* Wollongong, Australia. July 7–11.

Bode, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press.

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; Or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78 no. 1: 77–106.

Bode, Katherine. 2012. *Reading by Numbers: Recalibrating the Literary Field*. London: Anthem Press.

Bode, Katherine and Carol Hetherington, eds. 2018–. *To be continued: The Australian Newspaper Fiction Database*. http://cdhrdatasys.anu.edu.au/tobecontinued/.

Buzzetti, Dino. 2002. "Digital Representation and the Text Model." *New Literary History* 33 no. 1: 61–88.

Cecire, Natalia. 2011. "Introduction: Theory and the Virtues of Digital Humanities." *Journal of Digital Humanities* 1 no. 1: http://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire/.

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle. 2018. "Models and Modelling between Digital and Humanities. Remarks from a Multidisciplinary Perspective." *Historical Social Research / Historische Sozialforschung* 43 no. 4: 343–61.

Cordell, Ryan. 2017. "'Q i-jtb the Raven': Taking Dirty OCR Seriously." *Book History* 20: 188–225.

Crawford, Kate, Kate Miltner, and Mary L. Gray. 2014. "Critiquing Big Data: Politics, Ethics, Epistemology." *International Journal of Communication* 8: 10 pages.

Da, Nan Z. 2019. "The Computational Case against Computational Literary Studies." *Critical Inquiry* 45 no. 3: 601–39.

Eggert, Paul. 2009. *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge University Press.

Elliott, Jack. 2014. "Patterns and Trends in Harlequin Category Romances." In *Advancing Digital Humanities: Research, Methods, Theories*, edited by Paul Longley Arthur and Katherine Bode, 54–67. London: Palgrave Macmillan.

Flanders, Julia and Fotis Jannidis. 2016. "Data Modeling." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 229–37. John Wiley & Sons.

Fyfe, Paul. 2016. "An Archaeology of Victorian Newspapers." *Victorian Periodicals Review* 49 no. 4: 546–77.

Gavin, Michael. 2019. "How to Think about EEBO." *Textual Cultures* 11 no. 1-2: 70–105.

Guldi, Jo. 2018. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora." *Journal of Cultural Analytics* December 20. http://culturalanalytics.org/2018/12/critical-search-a-procedure-for-guided-reading-in-large-scale-textual-corpora.

Haraway, Donna. 1988. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14 no. 3 (1988): 575–99.

Hayles, N. Katherine. 2004. "Print is Flat, Code is Deep: The Importance of Media-Specific Analysis." *Poetics Today* 25 no. 1: 67–90.

Hoffmann, Anna Lauren. 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22 no. 7: 900–15.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Champaign, Illinois: University of Illinois Press.

Klein, Lauren. 2018. "Distant Reading after Moretti." *ARCADE. Literature, the Humanities, & the World*. https://arcade.stanford.edu/blogs/distant-reading-after-moretti.

Lahti, Leo, Jani Marjanen, Hege Roivaninen, and Mikko Tolonen. 2019. "Bibliographic Data Science and the History of the Book (c.1500–1800)." *Cataloguing & Classification Quarterly* 57: 5–23.

Latour, Bruno. 2004. "Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30 no. 2: 225–48.

Liu, Alan. 2018. *Friending the Past: The Sense of History in the Digital Age*. Chicago: Chicago University Press.

Long, Hoyt and So, Richard Jean. 2016. "Turbulent Flow: A Computational Model of World Literature." *Modern Language Quarterly* 77 no. 3: 345–67.

Loukissas, Yanni Alexander. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society*. Cambridge, Mass.: The MIT Press.

Luengo, Óscar G. and Jaime Peláez-Berbell. 2019. "Exploring the accuracy of electoral polls during campaigns in 2016: only bad press?" *Contemporary Social Science* 14 no. 1: 43–53.

Mak, Bonnie. 2014. "Archaeology of a Digitization." *Journal of the Association for Information Science and Technology* 65.8 (2014): 1515–26.

Manovich, Lev. 2016. "The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics." *Journal of Cultural Analytics*, May 23. http://culturalanalytics.org/2016/05/the-science-of-culture-social-computing-digital-humanities-and-cultural-analytics/.

McCarty, Willard. 2005. *Humanities Computing*. London: Palgrave Macmillan.

McFarland, Daniel A. and H. Richard McFarland. 2015. "Big data and the danger of being precisely inaccurate." *Big Data & Society* July–December: 4 pages.

McGann, Jerome. 2014. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, Massachusetts: Harvard University Press, 2014.

McGann, Jerome. 2005. "From Text to Work: Digital Tools and the Emergence of the Social Text." In *The Book as Artefact: Text and Border*, edited by Anne Mette Hansen, Roger Lüdeke, Wolfgang Streit, Cristina Urchueguía, and Peter Shillingsburg, 225–40. Amsterdam: Rodopi.

Milan, Stefania and Emiliano Treré. 2019. "Big Data from the South(s): Beyond Data Universalism." *Television & New Media* 20 no. 4: 319–35.

Moretti, Franco. 2017. "Patterns and Interpretation." *Stanford Literary Lab Pamphlet* 15, September: https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

NINES. nd. *Nineteenth-century Scholarship Online*. http://www.nines.org/about/.

Noth, Winfried. 2003. "Crisis of Representation?" *Semiotica* 143 no. 1: 9–15.

Ortega, Élika. 2018. "Archives, Libraries, Collections, and Databases: A First Look at Digital Literary Studies in Mexico." *Hispanic Review* 86 no. 2: 229–47.

Piper, Andrew. 2018. *Enumerations: Data and Literary Study*. Chicago: Chicago University Press.

Piper, Andrew. 2017a. "Think Small: On Literary Modeling." *PMLA* 132.3 (2017): 651–58.

Piper, Andrew. 2017b. "Data, data, data. Why Katherine Bode's new piece is so important and why it gets so much wrong about the field." *txtLAB* (blog), June 23. https://txtlab.org/2017/06/data-data-data-why-catherine-bodes-new-piece-is-so-important-and-why-it-gets-so-much-wrong-about-the-field/.

Prescott, Andrew. 2014. "I'd Rather be a Librarian." *Cultural and Social History* 11 no. 3: 335–41.

Price, Kenneth M. 2009. "Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?" *Digital Humanities Quarterly* 3: http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html.

Radcliffe, David Hill. ed. nd. *Spenser and the Tradition: English Poetry 1579–1830*. http://spenserians.cath.vt.edu/index.php.

Risam, Roopika. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evaston, Illinois: Northwestern University Press.

Ross, Shawna. 2018. "Toward a Feminist Modernist Digital Humanities." *Feminist Modernist Studies* 1 no. 3: 211–29.

Shore, Daniel. 2018. *Cyberformalism: Histories of Linguistic forms in the Digital Archive*. Baltimore: Johns Hopkins University Press.

So, Richard Jean. 2017. "All Models Are Wrong." *PMLA* 132 no. 3: 668–73.

Tenen, Dennis Yi. 2018. "Toward a Computational Archaeology of Fictional Space." *New Literary History* 49 no. 1: 119–47.

Thylstrup, Nanna Bonde. 2019. *The Politics of Mass Digitization*. Harvard, Mass.: The MIT Press.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: Chicago University Press.

Underwood, Ted. 2015. "A dataset for distant-reading literature in English, 1700–1922." *The Stone and the Shell: Using Large Digital Libraries to Advance Literary History* (blog), August 7, https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/.

Underwood, Ted and Jordan Sellers. 2016. "The *Longue Durée* of Literary Prestige." *Modern Language Quarterly* 77 no. 3: 321–44.

van Dijck, José. 2014. "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." *Surveillance & Society* 12 no. 2: 197–208.

van Dijck, José, Thomas Poell, and Martijn de Waal. 2018. *The Platform Society: Public Values in a Connective World.* Oxford: Oxford University Press.

Warren, Christopher N. 2018. "Historiography's Two Voices: Data Infrastructure and History at Scale in the *Oxford Dictionary of National Biography (ODNB)*." *Journal of Cultural Analytics* November 22: http://culturalanalytics.org/2018/11/historiographys-two-voices-data-infrastructure-and-history-at-scale-in-the-oxford-dictionary-of-national-biography-odnb/.

[1] This essay considers research that uses quantitative (counting, statistical, and algorithmic) methods to investigate literary datasets. Many names are applied to this research, including distant reading (Moretti 2005), macroanalysis (Jockers 2013), cultural analytics (Manovich 2016), data-rich literary studies (Bode 2018), and computational literary studies (Da 2019). I use quantitative literary studies partly because it's the term Underwood uses (interchangeably with distant reading) and partly because it's not as explicitly aligned, as many of these other terms are, with a particular author.

[2] For a prominent, recent example of both views see Da 2019.

[3] For example, Bode 2018; Cordell 2017. For related research not cited by Underwood see Fyfe 2016; Gavin 2019; Warren 2018.

[4] "Infrastructures of knowledge-making" is Bonnie Mak's (2014: 1519) phrase.

[5] Bibliography and scholarly editing are broad churches; but the scholarly trajectory in quantitative literary studies is particularly influenced by post-war sociological, reader-oriented approaches, often aligned with D. F. McKenzie and Jerome McGann.

[6] Of course, even if both are ideational, the literary work is an idea shored up by substantial historical, cultural, social, legal, political, institutional, and other practices, conventions, and discourses, whereas the literary system has simply been scholarly shorthand for the operations of these practices, conventions, and discourses in particular times and places. Now that digital resources and methods suggest possibilities for investigating rather than simply invoking this broader phenomenon, we face the challenge of determining what vast and distributed network of specific material objects invoked in such investigations: that's the motivation and intention of the scholarly edition of a literary system.

[7] Other digital literary projects adapt bibliographical and editorial principles to literary phenomena ranging from the influence of an individual author (Radcliffe nd) to the development of book formats in modern Europe (Lahti et al. 2019) to the nineteenth-century material archive (NINES nd).

[8] In this vein, in surveying the creation of EEBO in order to present the collection as an object of study in its own right, Michael Gavin argues that, "The transparency and comprehensiveness provided by this surrogate is, of course, not real, but neither is it an illusion. The EEBO-TCP is more like a simulation or model of extant print" (2019: 71).

[9] It includes over 13,000 publications not analysed in *A World of Fiction*, approximately 1,800 of which (as of June 2019) were added by members of the public (Bode and Hetherington, 2018–).

[10] The Helsinki Computational History Group employs algorithmic methods to harmonize and integrate large bibliographical datasets maintained by different research libraries to establish their capacity to sustain quantitative research into literary phenomena such as "the rise of octavo printing in Europe" (Lahti et al. 2019:

7); "Distant Reading for European Literary History" is building a "multilingual European Literary Text Collection" with attention to the literary works published and read in Europe (https://www.cost.eu/actions/CA16204/#tabs|Name:overview); "Travelling Texts 1790–1914: The Transnational Reception of Women's Writing at the Fringes of Europe" uses "systematic scrutiny of reception data from large-scale sources … [as] the basis for the study of women's participation" in nineteenth-century literary culture, from the perspective of five countries (Finland, the Netherlands, Norway, Slovenia, and Spain) (http://travellingtexts.huygens.knaw.nl/); and *Living with Machines* bases its exploration of industrialization on a corpus constructed by a "Sources Lab" with the aim of creating "a more representative account of the past" by identifying and balancing "different viewpoints and voices … to explore the plurality of histories, including minority voices too easily lost in the mix" (http://livingwithmachines.ac.uk/introducing-the-sources-lab/).

[11] On the reciprocal influence of digital archives on digital literary studies see Ortega 2018; on the silences created by these "digital canonicities," including of working-class voices, see Prescott 2014.

[12] See, for example, Underwood and Sellers 2016: 338.

[13] This logic resonates with Moretti's (2005: 4) claim that literary history "isn't a sum of individual cases: it's a collective system, that should be grasped as such, as a whole."

[14] Thanks to Galen Cuthbertson for this illustration.

[15] The difference between the two forms of selection bias I am discussing is quite subtle so a non-literary example might assist. Both are discussed as contributing to the failure of political polling in recent elections. Exclusion bias might occur because a poll is conducted online (or using landlines) and certain sectors of the population are more or less likely to use either media; sampling bias might occur when the wording of a survey makes a certain type of answer more likely (Luengo and Peláez-Berbell 2019).

[16] He is more circumspect on the following page, saying of the results of this process of sampling and comparison: "it is safe to say that we have found no evidence that the broad trends in HathiTrust are produced merely by library purchasing patterns" (136).

[17] This publication data was derived from *AustLit*, the Australian literary bibliography, and formed the basis for my analysis of post-war Australian novels in Bode 2012. The full dataset is available as "Australian Novels, 1945 to 2006" (https://katherinebode.wordpress.com/data/); for the proportions of Australian women's novels recorded as romance fiction see "Figure 1: Data" (https://katherinebode.wordpress.com/articles-chapters/).

[18] Actually, my study of gender trends in Australian novels has been criticized for underrepresenting romance fiction (Elliott 2014: 57–8).

[19] For an important argument against predictive and for explanatory models see Tenen 2018.

[20] Although Underwood notes in an appendix that some machine learning models are "difficult to interpret," and his own are only "relatively transparent," he generally insists that "scandalized discussion" of the opacity of these methods "has expanded far out of proportion to its practical significance" (191). To the contrary, "there are many ways to open the so-called black box of machine learning and find answers" (62); and "models produced by machine learning do not have to be black boxes. They can be complex but transparent moves in an interpretive argument" (160).

[21] If "perspectives" really were easy to measure, then banks and courts could do the same and exclude biases, subjective decisions, implicit associations, and so on, from their decisions.

[22] See Thylstrup 2019 on mass-digitization as a political (or infrapolitical) project that implicates cultural memory institutions in diverse legal, cultural, and ethical investments and uncertainties.

[23] It's been assumed that nineteenth-century Australian fiction didn't represent Aboriginal characters, but they're widely evident in newspaper fiction of the period, though not in (mostly British-published) books.

[24] Underwood argues that quantitative literary studies "helps mainly with questions where the evidence is simply too big to fit in a single reader's memory" (xxi). But it is not clear why this would lead to century-long spans, given that no reader can remember all the novels published in a single year, let alone a longer timeframe. Similarly, why should "the list of volumes we choose … matter a great deal" (179) when considering a decade but not three centuries?

[25] Underwood has done much to make literary data available to others, particularly by curating literary datasets from *HathiTrust*. See Underwood 2015.

[26] Klein cites multiple scholars who have made similar arguments, including Moya Bailey, Tanya Clement, Jessica Marie Johnson, Laura Mandell, Bethany Nowviskie, and Miriam Posner.

[27] Loukissas also ties digital universalism to "the assumptions of an encompassing and rarely questioned free market ideology" (10). Along with the greater historical, regional, and linguistic frictions implicated in Europe's literary cultures, it's interesting to consider whether that region's (changing but still) more collectivist orientation contributes to the previous-noted distinction in attitudes to data modeling in that context as opposed to in North America, where free market ideology is more accepted. For more on how this political contrast plays out in perceptions and regulations of digital platforms see van Dijck, Poell, and de Waal 2018.