DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 2

A Genealogy of Distant Reading

Ted Underwood <tunder_at_illinois_dot_edu>, University of Illinois, Urbana-Champaign

Abstract

It has recently become common to describe all empirical approaches to literature as subfields of digital humanities. This essay argues that distant reading has a largely distinct genealogy stretching back many decades before the advent of the internet – a genealogy that is not for the most part centrally concerned with computers. It would be better to understand this field as a conversation between literary studies and social science, inititated by scholars like Raymond Williams and Janice Radway, and moving slowly toward an explicitly experimental method. Candor about the social-scientific dimension of distant reading is needed now, in order to refocus a research agenda that can drift into diffuse exploration of digital tools. Clarity on this topic might also reduce miscommunication between distant readers and digital humanists.

Over the last decade or so, it has become common to describe all empirical approaches to literary history as subfields of digital humanities. At first, I didn't take this conflation seriously; I thought it was journalistic shorthand for a history that scholars understood to be more complex. Writing in *The New York Times*, for instance, Kathryn Schulz described distant reading in 2011 as one of many approaches "currently proliferating under the broad rubric of 'digital humanities'" [Schultz 2011]. A reader who thought it worthwhile to quibble could have replied that neither thing is a subset of the other. Although the projects were certainly in conversation by 2011, the phrases "distant reading" and "digital humanities" had been coined ten years earlier, in different academic communities, to describe different kinds of research. Digital technology hadn't even played a central role in early examples of distant reading. But why quibble? No one expects a short newspaper article to give a full history of academic trends. 1

More recently, however, I have noticed that scholars themselves are beginning to narrate intellectual history in the same way: treating all quantitative or empirical approaches to literary history as aspects of a digital turn in the discipline. In Amy Earhart's genealogy of "digital literary studies," for instance, distant reading is presented as a recent change of course in an intellectual tradition originally centered on editorial theory and the internet. 2

> [T]he digital work I have traced in the first part of this book has been largely representational, with technology primarily used to create idealized or better versions than would be possible in print. Current trends in digital literary studies, and the larger digital humanities, appear to be moving away from representational concerns and toward interpretive functions as contemporary digital scholars, such as Stephen Ramsay, Franco Moretti, Matthew Jockers, Geoffrey Rockwell, and others, are using technology to devolve, manipulate, and reform the literary text. [Earhart 2015]

It may be correct to say that interpretive questions are a relatively late development in "digital literary studies" (a tradition that Earhart traces to the impact of the World Wide Web in the 1990s). But the interpretive questions posed by the scholars she mentions in this passage are much older than the web. Quantitative interpretation of literature is a story that stretches back through book history, sociology, and linguistics to a range of nineteenth-century experiments [Mendenhall 1887], [Sherman 1893]. This tradition is a branch of the digital humanities only in a parochial sense — as we might call pizza a branch of American cuisine. Both things were imported from a different social context, and inherit a longer history of their own. 3

There is nothing wrong with writing a history of food in America, and also nothing wrong with Earhart's decision to focus

on a particular critical tradition initiated by the advent of the web. As long as readers remember that many ingredients of this history have longer backstories elsewhere, no one will be misled. But of course, backstories do get forgotten with the passage of time, and new generations learn to associate pizza mainly with Chicago or New York. Today it is common to see distant reading and the sociology of literature folded into discussions of "Big Data research in digital humanities" [Kaplan 2015]. To my eyes, this puts the emphasis in a strange place, and suggests that we have begun to forget (or at least underplay) important aspects of our own past. *Big data* is a twenty-first century technical buzzword. It is odd to see it used as a filing category for the much older aspiration to survey patterns that organize human culture. 4

This essay will turn the calendar back to the middle of the twentieth century, in order to tease apart intellectual traditions that have begun to be conflated. In particular, I want to emphasize that distant reading is not a new trend, defined by digital technology or by contemporary obsession with the word *data*. The questions posed by distant readers were originally framed by scholars (like Raymond Williams and Janice Radway) who worked on the boundary between literary history and social science. Of course, computer science has also been a crucial influence. But the central practice that distinguishes distant reading from other forms of literary criticism is not at bottom a technology. It is, I will argue, the practice of framing historical inquiry as an experiment, using hypotheses and samples (of texts or other social evidence) that are defined before the writer settles on a conclusion. 5

Integrating experimental inquiry in the humanities poses rhetorical and social challenges that are quite distinct from the challenges of integrating digital media. It seems desirable — even likely — that distant readers and digital humanists will coexist productively. But that compatibility cannot be taken for granted, as if the two projects were self-evidently versions of the same thing. They are not, and the institutional forms of their coexistence still need to be negotiated. 6

"Distant reading"

Large-scale literary history is far from a new idea. Vernacular literary study entered nineteenth-century universities as an already-ambitious project that sought to trace the parallel development of literature, language, and society across a thousand years. It was only in the twentieth century that literary scholarship began to restrict itself paradigmatically to the close reading of single texts. If we take a long view of disciplinary history, recent research on large digital libraries is just one expression of a much broader trend, beginning around the middle of the twentieth century, that has tended to reinstate the original historical ambitions of literary scholarship. 7

But that would be a very long view: it doesn't do much to help us understand current scholarly debate. For that, we need a tighter frame — a frame that can characterize the goals that have energized empirical approaches to literary history over the last half-century or so, without reducing them to an expression of twenty-first-century technology. This essay will provide an account on that intermediate scale. The frame I have chosen to use is the phrase *distant reading*. But I want to make clear from the outset that this phrase is not inevitable; there are other valid options. Andrew Goldstone observes that *distant reading* tends to foreground textual interpretation (reading) at the expense of questions about social structure [Goldstone 2015]. James F. English has shown that a similar account can be organized instead around the phrase "sociology of literature" [English 2010]. "Cultural analytics" could be an equally valid choice, if we wanted to include disciplines other than literary studies. In short, like most historical phenomena, the trend I am describing is composed of multiple overlapping impulses. There is more than one right way to describe it. 8

I have chosen *distant reading* because the phrase underlines the macroscopic scale of recent literary-historical experiments, without narrowly specifying theoretical presuppositions, methods, or objects of analysis. Although I understand Goldstone's concerns about the word, I think we are free to interpret *reading* as an inquiry about social structures as well as literary forms. *Distant reading* also has the crucial advantage of being vivid, memorable, and less bristly than any of the alternatives that end in "mining" or "analysis." On the other hand, it does have one significant disadvantage: it is often understood to imply a recent origin story that would prevent us from crediting any work done in the previous century. I will need to complicate that story in the pages that follow. It is true that Franco Moretti coined *distant reading* around the year 2000. But although Moretti is an important scholar in the tradition I will be tracing, there is no reason to treat his invention of the phrase as an originating moment for the whole tradition. *Distant reading* was not coined to describe a radically new method. The first occurrence of the phrase, in "Conjectures on World Literature," 9

seems in fact to describe the familiar scholarly activity of aggregating and summarizing previous research [Moretti 2000a]. *Distant reading* has evolved into a name for a more specific approach to literary history, but the approach described significantly predates this particular name for it.

10 Moretti's turn-of-the-century works were important, not because they invented the idea of macroscopic literary inquiry, but because they galvanized an existing project by infusing it with a new sense of possibility and a new polemical rationale. I will have more to say about his contribution, but this essay will mostly take aim at a larger target — a critical tradition, emerging in the later twentieth century, that would include things originally called "book history" or "sociology of literature," as well as more recent, emphatically quantitative experiments. The common denominator that links all these projects is simply that they pose broad historical questions about literature, and answer them by studying samples of social or textual evidence. Those samples may range from a few dozen instances to a million or more. Instead of prescribing a particular mode or scale of representation, I want to highlight the underlying project of experimenting on samples, and the premise that samples of the literary past have to be constructed rather than passively received.

11 This premise is general enough to have cropped up many times, so the tradition I am describing will lack crisp boundaries. Many traditional works of literary history pause at the outset to construct an informal sample of, say, Gothic novels. To the extent that those studies separate the construction of the sample from the process of historical inference, I would say they are approximating distant reading. Since versions of this approach to literature can be traced back to the nineteenth century, it would be pointless to go looking for a moment of origin. The emergence of distant reading was not contained in any eureka moment when a literary scholar decided to try social-scientific methods. It emerged rather through a long sequence of attempts, which gradually transformed casual historiographic practices into an explicitly experimental method.

Mid-twentieth-century developments

12 A longer study might follow this story down many different paths. Marxist literary theory has been one crucial influence; Raymond Williams might deserve a chapter of his own. The books he wrote around 1960 laid a theoretical foundation that still underpins much contemporary research — for instance, by insisting that literary culture is never a unified object, but rather a palimpsest of emergent and residual formations, transformed retrospectively by processes of selection. After reading Williams, it becomes hard to imagine that there could ever be a single definition of literary exemplarity, or a single correct sample of the literary past. In *The Long Revolution,* Williams also intriguingly foreshadows contemporary distant reading by grappling with a *longue durée*, and by emphasizing our ignorance about the past: "nobody really knows the nineteenth-century novel; nobody has read, or could have read, all its examples, over the whole range from printed volumes to penny serials" [Williams 1961, 66].

13 A full account of the emergence of distant reading might also spend a chapter on book history. Book historians have been compelled to explicitly define samples, since libraries don't cover the full range of practices they study. Book historians have also pushed literary history to define its object of study more concretely — separating processes of production, for instance, from circulation and reading practices. But these parts of the story are already well known [Darnton 1982]. In this limited space, I need to skip forward to a later stage of development, when the theoretical premises developed in book history and Marxist literary theory started to combine with an experimental method drawn from the social sciences. One excellent example of this fusion can be found in Janice Radway's *Reading the Romance* (1984).

14 This book became a monument of feminist scholarship by challenging the widespread premise that popular literature simply transmitted ideology. In Radway's view, critics had too quickly extrapolated their own interpretive practices to other readers. A critic may pick up a popular romance, for instance, identify the gender norms that seem implicit in the plot, and conclude that the effect of the book is to reinforce those norms. But how much does this tell us about the actual experience of romance readers? What aspects of the stories do they value? What role do the books play in their lives? Studying a community of women linked by a particular bookstore, Radway concluded that readers have more control over the meaning of stories than critics assume. Romances seemed to function in practice as a "declaration of independence" from the pressure of these readers' responsibilities as wives and mothers, even when the gender roles

represented in the narrative were traditional. Many subsequent arguments about the active agency of reception in fan culture are indebted to Radway's conclusions.

Literary scholars have been much slower to imitate her methods, which depended on questionnaires, interviews, and numbers. 15

TABLE 2.2

Question: What Are the Three Most Important Ingredients in a Romance?

Response	First Most Important Feature	Second Most Important Feature	Third Most Important Feature	Total Who Checked Response In One of Top Three Positions
a. A happy ending	22	4	6	32
b. Lots of scenes with explicit sexual description	0	0	0	0
c. Lots of details about faraway places and times	0	1	2	3
d. A long conflict between hero and heroine	2	1	1	4
e. Punishment of the villain	0	2	3	5
f. A slowly but consistently developing love between hero and heroine	8	9	6	23

Figure 1. Table from Radway 1984, p. 67.

Radway's quantitative methods may at first seem remote from familiar examples of distant reading. She doesn't discuss algorithms. Instead she uses numbers simply to count and compare — in order to ask, for instance, which elements of a romance novel are most valued by readers. Recent examples of distant reading can grow more complex than this. But they can also remain just as simple. Franco Moretti has relied on bibliographies to measure the lifespans of genres; I have quizzed readers about their impressions of elapsed time in ninety novels. 16

Admittedly, contemporary distant reading is usually based on textual evidence, or on social evidence about dead people, rather than questionnaires. Distant readers are certainly concerned with reception [DeWitt 2015]; [Algee-Hewitt and McGurl 2015]. But it is hard to find living witnesses to interview when you're studying the *longue durée,* so few distant readers have characterized reception quite as richly as *Reading the Romance*. These are significant differences. But the central research practice I want to highlight is broad enough to encompass all these different kinds of evidence. It is simply that Radway separates the question she is posing from the evidence she gathers to address it, and from the conclusion she finally draws. Moreover, she organizes these aspects of the research process sequentially. In short, Radway's book is designed as an experiment. It is admittedly an observational experiment: Radway isn't measuring the consequences of an intervention, and she doesn't express her reasoning in strict hypothetico-deductive form. Instead, she proceeds ethnographically, allowing herself to pause and comment when she sees an interesting detail. She is, after all, exploring a new research area, and encountering problems that are not yet formally defined. But *Reading the Romance* is still at bottom "empirical research" which aims to "test the validity of … a hypothesis" [Radway 1984, 11, 13]. Because Radway's voice is candid and engaging, the book may not always sound like social science. But the whole 17

rhetorical performance has been organized around a scrupulous attempt to avoid confirmation bias. That is the point of using a clearly-defined sample of readers and novels, rather than casually adducing quotations and anecdotes that happen to fit a thesis defined in advance.

18 Radway's doctorate was in American Studies; she currently teaches in a department of Communication Studies. But other social-scientific traditions are also hovering in the background of *Reading the Romance*. Its questionnaires and interviews echo the methods of sociology. And when Radway looks at the romance novels themselves, her methods echo both sociology and structural anthropology. Reading systematically through a sample of twenty romance novels, she finds, for instance, a set of "binary oppositions" organizing the heroine, the female foil, the hero, and the male foil into a symmetrical structure [Radway 1984, 122–32]. The plusses and minuses she uses to represent polarity in this structure recall Claude Lévi-Strauss's diagrams in *The Savage Mind* (1965). But her technique of systematically sampling and coding features of each novel also echoes the technique of "content analysis" that sociologists have applied to mass media.

19 Linguistics was not particularly central to Radway's project, and it may be worth pausing for a moment to underline this point. Contemporary distant reading has also been shaped by a different intellectual tradition devoted to quantitative analysis of linguistic detail. That tradition has made vital contributions, which I want to acknowledge. But I think linguistics may be looming a little too large in the foreground of contemporary narratives about distant reading, so much that it blocks our view of other things. Linguistic categories are just as important as the social categories Radway explored; it's not that I want to champion one subject against the other. Rather, I think we need to see both influences at once in order to grasp the generality of the method that organizes this research agenda. Our knowledge about large-scale literary history isn't expanding because there was a special magic in linguistic analysis (or a special moral authority in feminist sociology). The project is succeeding, rather, because scholars have learned how to test broad literary-historical hypotheses in a way that resists confirmation bias. Otherwise it would be very difficult to make progress at this scale. If you're working in a domain where you could potentially cite 100,000 different novels as evidence, confirmation bias will make all generalizations equally true until you invent some procedure to limit your own freedom of selection. As psychologists have expressed this: fields with abundant evidence need some way to limit "researcher degrees of freedom" [Simmons et al. 2011].

Moretti's contribution

20 Although Radway's book was widely celebrated and widely cited in English departments throughout the 1990s, it was not widely imitated there. As James F. English has pointed out, literary scholars are traditionally quick to borrow social scientists' conclusions, but slow to borrow their methods [English 2010, xiii–xiv]. We might justify our hesitation in various ways, but it is rooted, practically, in institutional inertia: literary curricula simply do not teach graduate students how to do content analysis or manipulate numbers. There are, however, a few cases where literary scholarship has developed along the lines suggested in *Reading the Romance* — including, notably, a scholar strongly associated with distant reading. In "The Slaughterhouse of Literature," Franco Moretti developed a coding scheme to describe the role of "clues" in detective fiction [Moretti 2000b]. He then read a sample of about twenty stories, taking notes on the presence or absence of each aspect of clues in order to sort the stories into a tree.

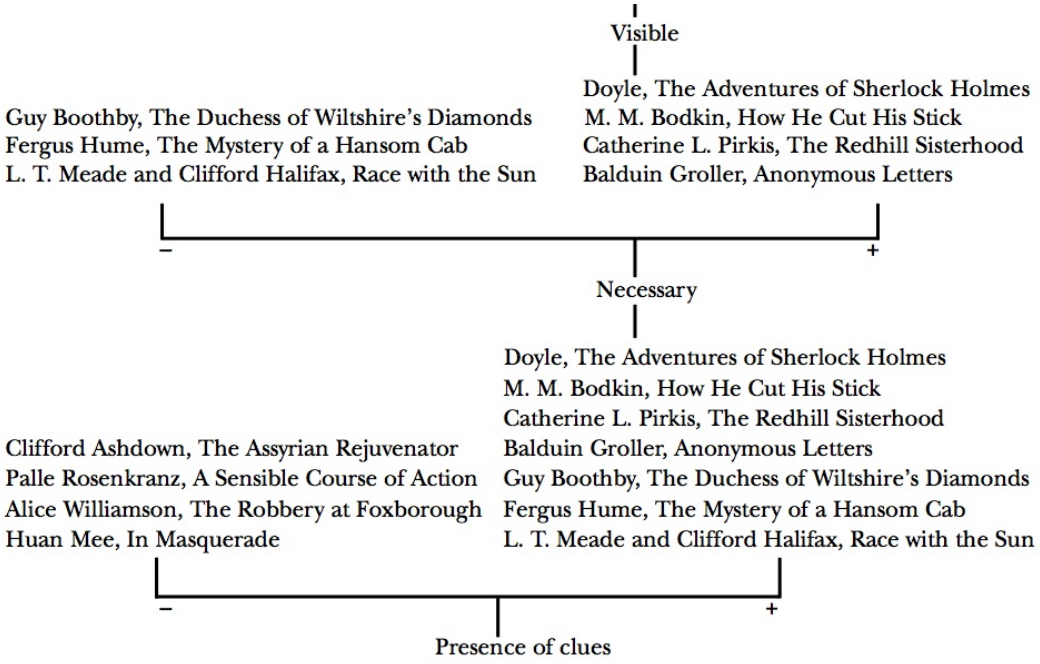

Figure 1 The presence of clues and the genesis of detective fiction

Figure 2. Figure from Moretti 2000b, p. 213.

21 This method is very close to Radway's approach to romance novels: from the sample of twenty texts, to the plan of reading systematically for particular features, to the little plusses and minuses that represent polarity in the diagram. I don't mean to suggest that Moretti was specifically influenced by *Reading the Romance*. It is more likely that both scholars drew their methods directly from structural anthropology and sociology. But whether the lines of influence run through literary criticism, or through social science, there is a coherent tradition to be traced here. Moretti adds an evolutionary hypothesis that is missing in Radway, and this may have been the aspect of his argument that most strongly shocked and provoked readers in the year 2000. But from the perspective of the present day, I think we can see that Moretti's evolutionary hypothesis was no more decisive than Radway's reliance on questionnaires. The crucial underlying similarity between these works, which has made both of them durably productive models for other scholars, is simply the decision to organize critical inquiry as an experiment.

22 To experiment on the past admittedly stretches the definition of *experiment* beyond its ordinary association with beakers and prisms. We cannot intervene in the past and then ask whether it changed as our hypothesis predicted. But this is a problem shared by geologists, astronomers, and computer scientists who run "experiments" on fixed datasets [Ullman 2013]. Distant reading is a historical science, and it will need to draw on something like Carol Cleland's definition of scientific method, which embraces not only future-oriented interventions, but any systematic test that seeks "to protect the hypothesis from misleading confirmations" [Cleland 2001, 988]. Literary historians can minimize misleading confirmations, for instance, by framing testable hypotheses about a sample of texts that are selected before the researcher settles on a conclusion. In calling this approach minimally "scientific," I don't mean to imply that we must suddenly adopt all the mores of chemists, or even psychologists. Imaginative literature matters because readers enjoy it; criticism would gain nothing if we let meticulous hypothesis-testing drain all the warmth and flexibility from our writing. Literary historians who use numbers will have to somehow combine rigor with simplicity, and prune back a thicket of fiddly details that would be fatal to our reason for caring about the subject. But within those rhetorical limits, distant reading can, let us say, *aspire* to the methods of social science: it is defined not only by a commitment to historical breadth, but by a version of the scientific method appropriate for a historical discipline.

23 Of course, not everyone will agree with this definition. For many scholars, the term *distant reading* is still shaped by the polemical context that surrounded it circa 2000, when it seemed to be the culmination of a long argument over the

canon. The process of canon-revision that began by addressing imbalances of race and gender had evolved, by the late 1990s, into a more systematic expansion that sought to recover a larger "great unread" [Cohen 1999, 23]. Although the political implications of this project had become increasingly diffuse, it still retained some of the moral fervor of the canon wars. Readers of Moretti's early experiments on large collections were accordingly tempted to interpret them as a normative argument that the only valid sample of literature is the largest possible one. This is not a position that his articles affirm systematically, but they sometimes leave themselves open to that interpretation. The decision to characterize the archive of forgotten books as a "slaughterhouse" of literature, for instance, echoes the moral pathos associated with a mission of recovery. I don't think the normative force of that pathos has turned out to be the most durable and influential part of the project. But it's the part that readers were primed to pay attention to, and thus the part they often remember.

24 Moretti's insistence on reconstructing a maximally complete archive is also the part of distant reading that scholars have spent most time debating. Many critics have pointed out that it is impossible to recover everything [Bode 2014, 20–21]. From that indisputable premise, they sometimes infer (more disputably) that comprehensiveness is not even an appropriate goal [Rosen 2011]. I won't rehearse the debate here; it seems to me a waste of energy, since there are many valid ways to represent the past. Scholars interested in literary production may want to approximate completeness, while scholars interested in reception prefer to focus on a subset of influential works. Some social questions hinge on the demographic identity of writers, others on readers — while for other modes of moral engagement with human history, the whole issue of synchronic social breadth is less urgent than diachronic scope. All of these sampling strategies have their uses, and there is no reason to make a final choice between them. The legacy of the canon wars has perhaps made literary scholars a little too eager to force such a choice. Having seen many premature arguments on the topic, I try not to join any debate about the representativeness of different samples until I have seen some evidence that the debate makes a difference *to the historical question under discussion.* Considering more than one sample can be worthwhile, but samples are provisional, purpose-built things. They are not canons. It makes no sense to argue about their representativeness in the abstract, before a question is defined. Moreover, it often turns out that the same patterns are visible, whether you look at ten thousand obscure texts, or at two hundred lovingly curated editions. So it would be a mistake to stall out at the initial stage of research, in "an argument about what constitutes an historically relevant and justifiable sample for analysis" [Bode 2017, 17]. That question has no right answer. We will get much further if we defer the argument, and start instead by comparing different samples.

25 I have been at pains to downplay several aspects of Moretti's contribution to distant reading that are often seen as definitional: his coinage of the phrase itself, and his emphasis on comprehensive samples that include many non-canonical works. However, I do think Moretti is rightly credited with sparking the twenty-first-century expansion of this research project. To illustrate why, I can't do better than quote the last paragraph of "Slaughterhouse":

> Fantastic opportunity, this uncharted expanse of literature, with room for the most varied approaches, and for a truly *collective* effort, like literary history has never seen. Great chance, great challenge … which calls for a maximum of methodological boldness: since no one knows what knowledge will mean in literary studies ten years from now, our best chance lies in the radical diversity of intellectual positions, and in their completely candid, outspoken competition. Anarchy. Not diplomacy, not compromises, not winks at every powerful academic lobby, not taboos. Anarchy. [Moretti 2000b, 227]

26 Two contributions are vital here. First, the recognition that literary history is not an exhausted, well-mapped field, but an "uncharted expanse," because we actually know little about its macroscopic shape. When I say that Moretti galvanized distant reading by infusing it with a new sense of possibility, this is the primary thing I mean. But I would also emphasize, secondly, his inference that the diplomatic reconciliation of conflicting normative claims is less urgent than many literary scholars assume.

27 Here we reach a zone of persistent miscommunication between distant readers and their colleagues. The discipline of literary studies has long organized itself around prescriptive debates that seek to define the proper concern of a literary critic. We inherit this polemical emphasis from nineteenth-century criticism, and it survives today in vigorous arguments

that pit history against form, surface against depth, and critique against appreciation. Scholars rooted in this tradition understandably want to interpret distant reading as a normative stance of the same kind. Perhaps distant readers are expressing a principled opposition to, say, close reading? In that case, the natural next move would be to dialectically sublate the tension between close and distant. Observers are often quite willing to offer this sort of compromise solution [Freedman 2015]. For literary critics, it is an obvious response. But from within the project of distant reading, it feels like a non sequitur. At bottom, distant readers are not arguing against close reading. They're just pointing to a blank space on our map of the past — where questions about large samples or long timelines might be located — in order to say "none of us really know what's in there yet." A confession of ignorance isn't something one can meaningfully strike compromises about; it calls for a different genre of response. Instead of interpreting distant reading as a normative argument about the discipline, it would be better to judge it simply by asking whether the blind spot it identified is turning out to contain anything interesting.

28 I am of course a biased observer. But personally, I became confident that new scales of inquiry were paying off in 2012, when Ryan Heuser and Long Le-Khac published evidence of a massive, steady shift from abstraction to concrete description in nineteenth-century novels [Heuser and Le-Khac 2012]. In subsequent years, distant readers have grappled with social questions about money, gender, race, geography, and literary circulation, as well as formal questions about genre, plot, emotion, and time [Wilkens 2013]; [Klein 2013]; [Cordell 2015]; [Jockers and Kirilloff 2016]. Some of these publications are still working their way through the press; in many cases, scholars are still struggling to reach consensus about the meaning of the evidence they uncovered. For instance, the shift toward concreteness discovered in fiction by Heuser and Le-Khac has alternatively been described as a broader parting of the ways between literary and nonliterary language, affecting poetry and nonfiction as well as the novel [Underwood and Sellers 2012]. If all of these discoveries are things we already knew in a tacit or unconscious way — as skeptics sometimes suggest — then our unconscious must have known so many conflicting things that the verb *know* seems oddly generous. Consensus about new evidence emerges very slowly: inventing an air-pump doesn't immediately convince readers that vacuums exist. So wariness about particular conclusions is definitely still warranted. But at this point, there is no doubt in my mind that literary scholarship turned out to have a blind spot. Many important patterns in literary history are still poorly understood, because they weren't easily grasped at the scale of individual reading.

Distant reading and computational methods

29 Up to this point, I have said relatively little about numbers, and nothing at all about computers. I have characterized distant reading as a tradition continuous with earlier forms of macroscopic literary history, distinguished only by an increasingly experimental method, organized by samples and hypotheses that get defined before conclusions are drawn. The interdisciplinary connections that mattered most for this tradition were, until recently, located in the social rather than computational sciences.

30 However, it is true that this sociological approach to literature has, over the last twenty-five years, fused with a computational tradition. The history of that fusion is complex, and I won't try to detail it fully here; one could point to Mark Olsen and the ARTFL project at Chicago, or to Matthew Jockers and the Stanford Literary Lab, or to John Unsworth and an archipelago of people involved with the MONK Project. In any case, it is clear that large-scale literary history is now suffused with ideas drawn from corpus linguistics, information retrieval, and machine learning. I don't intend to downplay the significance of this fusion; it has been the most exciting part of my career, and I'm indebted to everyone I just mentioned.

31 Nor do I want to suggest that computation was merely a means to achieve an end that Radway and Moretti had already fully defined. Critics of digital humanities often assume that computer science ought to remain merely instrumental for humanists; it should never "challenge" our "fundamental standards or procedures" [Columbia 2014, 164]. This misunderstands the place of computational disciplines in intellectual history. The value of computation is not merely to accelerate literary research or expand its scale; on the contrary, ideas drawn from computer science have given literary scholars new questions [Jones 2014, 31–32], and have encouraged us to frame existing questions in a more explicitly theorized way [Piper 2016]. Machine learning, for instance, represents a new way of thinking about literary concepts, like genre, that may be organized around loose family resemblances rather than crisp definitions [Long and So 2016].

In short, I am not at all motivated to shore up disciplinary boundaries or insist on a strictly internalist history of literary studies. And yet I have to admit that, for me, distant reading remains the name of an approach to literary history rather than a computational method. To be sure, it has multiple genealogies, and roots in many disciplines. But in tracing connections to the past I would still, on the whole, emphasize the thread that runs back through Moretti, Radway, and Williams. My rationale is simple. An approach to literature informed by social science can produce significant historical results by itself — with or without computers. But the converse has not generally turned out to be true. Computational methods, by themselves and without a social scale of inquiry, have not been enough to transform literary history. 32

We know this, to be quite blunt, because computational methods were applied to literature for thirty years without making a great impact on the discipline. The journal *Computers and the Humanities* was founded in 1966. It became the center of an ambitious intellectual community, making important contributions to phonology and concordance-building, database design and the teaching of language. But the whole project made very little difference for literary history. Stanley Fish observed as much in the 1970s [Fish 1973], and Mark Olsen couldn't really disagree, writing in the pages of the journal itself in 1993: "Computer-aided literature studies have failed to have a significant impact on the field as a whole" [Olsen 1993]. According to Olsen, the mistake lay in trying to explain "how a text achieves its literary effect" by examining "subtle semantic or grammatical structures in single texts or the works of individual authors." Computers had turned out to be "very poorly suited" to those New Critical questions, and concentration on them had "tended to discourage researchers from using the tool to ask questions to which it is better adapted, the examination of large amounts of simple linguistic features" [Olsen 1993, 309]. The irony, Olsen goes on to say, is that this broader and simpler kind of text processing is exactly what recent developments in literary theory and semiotics would seem to demand. (He cites Roland Barthes, Michel Foucault, and M. A. K. Halliday.) If only these two branches of research could be connected, computational analysis might finally assume a central role in literary studies. 33

This was the article that originally pulled me toward distant reading in the mid-1990s [Underwood 1995, 124]. I still find it a prescient argument. One of Olsen's strengths is that he ignores the false opposition between allowing our research to be shaped by properly literary questions, and allowing it to be guided by the capacities of digital tools. Instead, he considers both aspects of the landscape at once, and highlights a zone of intersection, where new literary questions happen to overlap with new technical opportunities. That zone of intersection turned out to be extremely productive, and Olsen's prophecies have almost all come true. With the important (but isolated) exception of authorship attribution, computers still contribute relatively little to our understanding of individual texts and authors. But computational methods now matter deeply for literary history, because they can be applied to large digital libraries, guided by a theoretical framework that tells us how to pose meaningful questions on a social scale. Olsen's article may overlook some scholars who were already moving in the direction he recommended [Radway 1984]; [Brunet 1989]. And the framework we use today may be more sociological, and less semiotic, than Olsen predicted. But as crystal balls go, his 1993 article isn't bad. It explains at once how the tradition embodied in *Computers and the Humanities* could eventually become significant for literary history, and why that significance wasn't for the most part achieved until the twenty-first century. 34

Moreover, Olsen's remarks are still a useful warning for scholars working in the area of overlap between digital humanities and distant reading. Algorithms are genuinely important; they aren't merely instrumental. But they also aren't sufficient for this project. So far, computation has only made a difference for literary history in combination with reasonably broad samples aimed at historical questions. A broad sample does not have to be an exhaustive collection; it might only amount to a few dozen books. But framing questions about dozens of books still tends to require a complete rethinking of received research questions. So I understand why scholars are often tempted to start with the algorithms instead, hoping that they will produce something interesting when applied to familiar author-sized questions. Unfortunately, in my experience, this is false economy. Olsen's warning has not been superseded by any technical advance: computers still can't teach us much about New Criticism. (Maybe someday, but not quite yet.) Within the sprawling ecumenical community called "digital humanities," it can be impolitic to insist on this barrier to easy assimilation of digital methods. But I make a point of distinguishing distant reading from digital humanities partly in order to signpost the problem: using computation, and reframing the scale of literary inquiry, are two distinct things. The first will not give you the results of the second. 35

The elision of social science

Writing in *Computers and the Humanities,* Olsen was naturally inclined to tell a story whose central characters were humanists and computers. He acknowledged the relevance of social science, but didn't foreground its methods. The same thing can be said about much contemporary work in distant reading. The best distant readers do in practice approach their projects as experiments. (We don't wander around aimlessly counting things.) But the experimental structure of our research is not always foregrounded when we write it up for publication. An article organized in social-scientific fashion (methods, then results, then conclusions) might not be warmly received by an audience of literary critics who are used to rhetorical panache. It can be more effective to pretend that your work grew in a casually discursive, thesis-driven way, and then happened to be illustrated with some scatterplots you had lying around. 36

I'm as guilty as anyone of striking this casual pose. It is often unavoidable. I have suggested that distant readers aspire to a version of the scientific method appropriate for a historical discipline. But we are also literary critics, and critics have an obligation to be interesting. This means that we sometimes have to tuck methods in an appendix, or make the analytical task look a bit easier than it truly was.[1] On the whole, I accept this rhetorical double-bind as a consequence of our location on a tricky disciplinary boundary. But it does have the side-effect of obscuring an engine that powers the project. Readers can see why broad historical questions matter, and they can see the role of computers. But the value of an explicitly experimental approach can be hard to discern: distant readers have an incentive to play down that part of our work. And yet, the experimental framing of research questions is really the key to this field. Great work can still be done with Janice Radway's quantitative methods, which require little more than paper and pencil. On the other hand, it's very hard to do social research at scale without imitating Radway's explicitness about hypotheses, samples, and results. 37

Unfortunately, social-scientific methodology has not been a central subject of conversation in digital humanities, or in the forms of distant reading that cluster under the DH rubric [Clement 2016]. Andrew Goldstone is right that the word "reading" itself contributes to the elision of social science [Goldstone 2015]. But the recent tendency to treat distant reading as a subfield of digital humanities may also play a role. The term *digital humanities* stages intellectual life as a dialogue between humanists and machines. Instead of explicitly foregrounding experimental methods, it underlines a boundary between the humanities and social science. 38

That's why I have written this article — to tease out the elided social-scientific genealogy behind distant reading. There are other threads one could trace. For instance, as I have acknowledged, machine learning is exerting a powerful influence on the contemporary scene. I don't want to disparage any subfield, but I do want to insist that the genealogy of distant reading should be traced by disentangling its central intellectual impulses, not just by following the zone of overlap between computers and textual study as far back as possible. Roberto Busa's concordance of Aquinas was a valuable thing, but a concordance of a single author does not constitute an important origin moment for distant reading. If we wanted to trace this tradition back to the middle of the twentieth century, we would need to follow different threads in several different directions. We might end up asking what Raymond Williams was doing with literature in the late 1950s, what Claude Lévi-Strauss was doing at the same time with social anthropology, and what Frank Rosenblatt was doing with the perceptron. 39

In the twenty-first century, admittedly, these disciplinary stories are tending to converge and fuse. That creates an exciting challenge, but also a problem for graduate training. Scholars preparing to work as distant readers probably need some exposure to programming, social theory, and statistics, as well as fairly deep knowledge of a literary-historical tradition. Right now, the flexible interdisciplinary community called "digital humanities" may be the best home for students trying to combine these modes of preparation. 40

But if these two projects are to coexist under one roof, the differences between them need candid discussion. Digital humanists don't necessarily share distant readers' admiration for social science. On the contrary, they are often concerned to defend a boundary between quantitative social science and humane reflection (see, e.g., [Burke 2016]). If DH is unified at all, it is unified by reflection on digital technology, in a mood that ranges from playful exploration to monitory critique. Distant reading, on the other hand, is not primarily concerned with technology at all: it centers on a social-scientific approach to the literary past. This tension sets up a predictable conflict, which has already begun to unfold. Introductory courses and workshops in DH rarely teach students what they need to know in order to practice distant reading. So distant readers will have to agitate for a different kind of curriculum, with more emphasis on 41

quantitative methods. The agitation is already underway (e.g. in [Tenen 2016] and [Goldstone 2017]), but in a conversation framed by the adjective *digital,* it can easily be misinterpreted as an attempt to push digital humanists in a more technical direction. For instance, distant readers' interest in broad historical questions — which runs back at least to Raymond Williams — is widely conflated with the recent technical buzzword, *big data.* Conflations of that kind could begin to create an unproductive debate, where parties to the debate fail to grasp the reason for disagreement, because they misunderstand each other's real positions and commitments.

42 This article has tried to clarify the commitments that define distant reading. I have not aimed to produce consensus: I know that many scholars cited here will disagree with my definition of the field. In particular, I know that many scholars maintain strong ties to both digital humanities and distant reading, and I expect people who do both things will resist the conclusion that these are intellectually distinct projects. Certainly, the projects are at present fused, in ways that matter deeply to academics as human beings. For instance, job advertisements usually call for a "digital humanist" — almost never for a "distant reader." So it is pragmatically unwise for junior scholars to separate the two terms, and a purely descriptive account of the contemporary social scene might well fold them together. This essay has separated digital tools from experimental methods for reasons that are not purely pragmatic or descriptive. I have tried to ground the separation in a genealogical narrative, but I would also admit that it has a forward-looking prescriptive purpose.

43 Over the last fifteen years, as distant readers have seized technological opportunities, the goals of the project have become diffuse. Often our immediate goal has really been exploratory: "let's see what can be done with these tools." The exploration has been fruitful, but I think the field is ready to move past exploration. Large-scale literary history could now reorganize itself around clear research questions and rigorously advance our knowledge of the past. But in order to do that, I believe we need to set fascination with technology to one side and rediscover the guiding principle of experiment. I have defended that opinion by pointing to the history of the field, and especially to the importance of social science in Williams, Radway, and Moretti. But it is also, in the end, an opinion. This essay promises only *a* genealogy of distant reading. There will be other versions of the story, and I look forward to reading them.

Acknowledgments.

44 This essay benefited greatly from conversation at the Instant History symposium at Loyola University Chicago, organized by Paul Eggert and Steven Jones in the fall of 2016. Respondents at the symposium included Ian Cornelius, Lydia Craig, Casey Jergenson, and Justin Hastings. The argument was also influenced by conversation with Andrew Goldstone and Eleanor Courtemanche, and it was improved by the editors and reviewers of *DHQ*.

Notes

[1] Even here, I am simplifying for rhetorical effect. In reality, many writers have developed informal ways of stressing the experimental character of their research. Scholars who have worked in the Stanford Literary Lab tend to be particularly good at highlighting moments when the evidence they gather fails to support their original assumptions. If I am right that resistance to confirmation bias is the main point of experimental method in the humanities, this narrative device may be just as effective as the more formal templates (methods/results/conclusions) that have developed in the social sciences. Freely sharing code and data is another way to reveal the experimental infrastructure that the disciplinary mores of literary study would tend to conceal.

Works Cited

Algee-Hewitt and McGurl 2015 Algee-Hewitt, Mark, and Mark McGurl. *Between Canon and Corpus: Six Perspectives on Twentieth-Century Novels,* 2015, retrieved from http://litlab.stanford.edu/LiteraryLabPamphlet8.

Bode 2014 Bode, Katherine. *Reading by Numbers: Calibrating the Literary Field* London: Anthem, 2014.

Bode 2017 Bode, Katherine. "The Equivalence of 'Close' and 'Distant' Reading; Or, Towards a New Object for Data-Rich Literary History." Draft, retrieved from https://katherinebode.files.wordpress.com/2014/07/equivalence1.pdf

Brunet 1989 Brunet, Etienne. "L'exploitation des grands corpus: Le bestiaire de la littérature française," in *Literary and Linguistic Computing* 4(1989), no. 2, pp. 121-134.

Burke 2016 Burke, Timothy. "The Humane Digital," in *Debates in the Digital Humanities 2016,* edited by Matthew K. Gold

and Lauren F. Klein. Minneapolis: University of Minnesota Press, 2016, pp. 514-518.

Cleland 2001 Cleland, Carol E. “Historical Science, Experimental Science, and the Scientific Method,” in *Geology* 29(2001), no. 11, pp. 987-990.

Clement 2016 Clement, Tanya. “Where is Methodology in Digital Humanities,” in *Debates in the Digital Humanities 2016,* edited by Matthew K. Gold and Lauren F. Klein. Minneapolis: University of Minnesota Press, 2016, pp. 153-175.

Cohen 1999 Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton: Princeton University Press, 1999.

Cordell 2015 Cordell, Ryan. “Reprinting, Circulation, and the Network Author in Antebellum Newspapers.” *American Literary History* 27(2015), no. 3: pp. 417-445.

Darnton 1982 Darnton, Robert. “What is the History of Books?” in *Daedalus* 111(1982): pp. 65-83.

DeWitt 2015 DeWitt, Anne. “Advances in the Visualization of Data: The Network of Genre in the Victorian Periodical Press,” in *Victorian Periodicals Review* 48(2015), no. 2: pp. 161-182.

Earhart 2015 Earhart, Amy. *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*. Ann Arbor: University of Michigan Press, 2015, retrieved from http://dx.doi.org/10.3998/etlc.13455322.0001.001

English 2010 English, James F. “Everywhere and Nowhere: The Sociology of Literature after ‘the Sociology of Literature’,” in *New Literary History* 41 (2010): pp. v-xxiii.

Fish 1973 Fish, Stanley. “What Is Stylistics and Why Are They Saying Such Terrible Things About It,” in *What is Aesthetics,* ed. Seymour Chatman, New York: Columbia University Press, 1973, pp. 109-152.

Freedman 2015 Freedman, Jonathan. “After Close Reading,” in *The New Rambler,* April 13, 2015, retrieved from http://newramblerreview.com/book-reviews/literary-studies/after-close-reading.

Goldstone 2015 Goldstone, Andrew. “Distant Reading: More Work to be Done,” August 8, 2015, retrieved from https://andrewgoldstone.com/blog/2015/08/08/distant/

Goldstone 2017 Goldstone, Andrew. “Teaching Quantitative Methods: What Makes It Hard.” Forthcoming in *Debates in the Digital Humanities 2018*, edited by Matthew K. Gold and Lauren F. Klein. Retrieved from https://andrewgoldstone.com/teaching-litdata.pdf.

Golumbia 2014 Golumbia, David. “Death of a Discipline,” in *differences* 25(2014), 1, pp. 156-176.

Heuser and Le-Khac 2012 Heuser, Ryan and Long Le-Khac. *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*. Stanford Literary Lab, 2012, retrieved from http://litlab.stanford.edu/LiteraryLabPamphlet4.

Jockers and Kirilloff 2016 Jockers, Matthew, and Gabi Kirilloff. “Gender and Character Agency in the 19th Century Novel,” in *Cultural Analytics* 2016, retrieved from http://culturalanalytics.org/2016/12/understanding-gender-and-character-agency-in-the-19th-century-novel/.

Jones 2014 Jones, Steven E. *The Emergence of the Digital Humanities*. New York: Routledge, 2014.

Kaplan 2015 Kaplan, Frédéric. “A Map for Big Data Research in Digital Humanities,” in *Frontiers in Digital Humanities* 2 (2015), 2, pp. 1-7.

Klein 2013 Klein, Lauren. “The Image of Absence: Archival Silence, Data Visualization, and James Hemings,” in *American Literature* 85(2013), no. 4: pp. 661-688.

Long and So 2016 Long, Hoyt, and So, Richard Jean. “Literary Pattern Recognition,” in *Critical Inquiry* 42 (2016), 2, pp. 235-267.

Mendenhall 1887 Mendenhall, T. C. “The Characteristic Curves of Composition,” in *Science,* vol 9 (1887), 214, pp. 237-246.

Moretti 2000a Moretti, Franco. “Conjectures on World Literature,” in *New Left review* 1(2000), retrieved from https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature.

Moretti 2000b Moretti, Franco. “The Slaughterhouse of Literature,” in *Modern Language Quarterly* 61(2000), 1, pp. 207-227.

Olsen 1993 Olsen, Mark. “Signs, Symbols, and Discourses: A New Direction for Computer-Aided Literary Studies,” in

Computers and the Humanities 27 (1993/1994), 5/6, pp. 309-314.

Piper 2016 Piper, Andrew. “There Will Be Numbers.” *Cultural Analytics* 2016, retrieved from http://culturalanalytics.org/2016/05/there-will-be-numbers/

Radway 1984 Radway, Janice. *Reading the Romance: Women, Patriarchy, and Popular Literature*. Chapel Hill: University of North Carolina Press, 1984.

Rosen 2011 Rosen, Jeremy. “Combining Close and Distant, or the Utility of Genre Analysis: A Response to Matthew Wilkens’s ‘Contemporary Fiction by the Numbers’,” in *Post45*, December 3, 2011, retrieved from http://post45.research.yale.edu/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers/.

Schultz 2011 Schultz, Kathryn. “What Is Distant Reading?” in *The New York Times,* June 24, 2011, retrieved from http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

Sherman 1893 Sherman, L. A. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn, 1893.

Simmons et al. 2011 Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,” in *Psychological Science* 22 (2011), no. 11, pp. 1359-1366.

Tenen 2016 Tenen, Dennis. “Blunt Instrumentalism: On Tools and Methods.” In *Debates in the Digital Humanities 2016,* edited by Matthew K. Gold and Lauren F. Klein. Minneapolis: University of Minnesota Press, 2016, pp. 83-91.

Ullman 2013 Ullman, Jeffrey D. “Experiments as Research Validation — Have We Gone Too Far?” July 9, 2013. Retrieved from http://infolab.stanford.edu/~ullman/pub/experiments.pdf

Underwood 1995 Underwood, Ted. “Productivism and the Vogue for ‘Energy’ in Late Eighteenth-Century Britain,” in *Studies in Romanticism* 34(1995), no. 1, pp. 103-125.

Underwood and Sellers 2012 Underwood, Ted, and Jordan Sellers. “The Emergence of Literary Diction,” in *Journal of Digital Humanities* 1(2012), no. 2, retireved from http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/.

Wilkens 2013 “The Geographic Imagination of Civil War-Era American Fiction,” in *American Literary History* 25(2013), no. 4, pp. 803-840.

Williams 1961 Williams, Raymond. *The Long Revolution*. New York: Penguin, 1985.