Aggregation, Public Criticism, and the History of Reading Big Data

BENJAMIN MANGRUM

ONE OF THE STANDARD NARRATIVES REGARDING THE ORIGIN OF digital methods used for humanities research looks to the Jesuit priest and scholar Roberto Busa, who in 1949 published an index of the works of Thomas Aquinas. Busa collaborated with the CEO of IBM, Thomas J. Watson, to use computational methods for organizing and navigating Aquinas's works. Some scholars have argued that the computational humanities were born in July 1945, when, as Eileen Gardiner and Ronald G. Musto explain, "Vannevar Bush, a pioneering engineer in the development of analog computing, published an article in which he introduced the Memex—a hypothetical instrument to control the ever-accumulating body of scientific literature" (67). Bush imagined a kind of interactive desk that would perform the role of an encyclopedic system of storage and retrieval. Gregor Wiedemann adds to the list of "milestones" the development of software called the General Inquirer during the 1960s (334). Basing the prehistory of the digital humanities on accounts of figures such as Busa and Bush or of software designed for content analysis anchors it in seminal intersections between computational technology and conventional instances of humanistic scholarship.

While locating the origins of the digital humanities in such moments of revolution and singular innovation is undoubtedly important, the use of computational methods to analyze cultural and social phenomena has wider intellectual debts. What might we learn about particular digital methods if we consider the history of their conceptual underpinnings? How might examining specific cases in the intellectual genealogies of computational technology inflect our digital practices today? Aspiring to such a task, I consider the intellectual history of one prominent method in the digital humanities: the use of quantitative methods to analyze large corpora or databases

BENJAMIN MANGRUM is a fellow with the Michigan Society of Fellows and an assistant professor of English at the University of Michigan, Ann Arbor. He is the author of *Land of Tomorrow: Postwar Fiction and the Crisis of American Liberalism* (Oxford UP, 2018). His essays have appeared in *New Literary History*, *American Literature*, *Contemporary Literature*, *Genre*, and elsewhere.

© 2018 BENJAMIN MANGRUM

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

of texts and thereby to explain the contours of literary history. Large corporations and research institutions alike curate these so-called text corpora, which include the Internet Archive, HathiTrust and Google's database of texts from scanned books, and Harvard University's closely related Culturomics project. In 2011, a partnership between Google and researchers at Harvard reported "the creation of a corpus of 5,195,769 digitized books containing ~4% of all books ever published." The researchers complain that, before this corpus existed, "[a]ttempts to introduce quantitative methods into the study of culture ha[d] been hampered by the lack of suitable data" (Michel et al. 176). The creation of this corpus facilitated Google's *Ngram Viewer*, which allows any user to chart the frequency of search strings across history through the text corpus.

But what does this intersection of literary criticism, technology, and statistical aggregation presuppose about the public sphere? How did statistical graphing and the types of knowledge yielded by large corpora of literary data gain legitimacy? There are a variety of ways to evaluate the uses, material conditions, and historical assumptions of such large corpora. Instead of, say, identifying the first humanists to use large data stores and computational and quantitative methodologies, this essay traces a historical arc through the structures of thought, sentiment, and plausibility underlying a particular set of quantitative methods that use large corpora of texts. What I am describing as a "set of quantitative methods" is admittedly quite diverse. It includes the "quantitative patterns" that James M. Hughes and his coauthors use to formulate "a simple evolutionary model for stylistic influence" through Project Gutenberg's digital library corpus (Hughes et al. 7682); the n-gram tools of Google and the Culturomics team; the quantitative analysis published in many of the Stanford Literary Lab's pamphlets; and recent, related work by Matthew Jockers, Matt Erlin and Lynne Tatlock, and Franco Moretti (*Distant Reading*).

This diverse body of scholarship is representative of what Heather Love groups under the rubric "new sociologies of literature," which "distance themselves from texts and from practices of close reading altogether" (373). Despite the differences in their emphases and tools, and in the text corpora that they consider, these quantitative methodologies have become some of the most prominent means for analyzing large textual data in the digital humanities. In 2000, Moretti laid out the ethos of these methods in terms of "distant reading," defining "distance" as "*a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text disappears, well, it is one of those cases when one can justifiably say, Less is more" ("Conjectures" 57). Individual texts drop out as the objects of inquiry, and large corpora analyzed for data become the relevant standards of knowledge (Schöch 2–4; Love 374).[1]

Quantitative analysis of large text corpora has generated new ways of thinking about literary history. However, fundamental to such analysis is the notion of an integrated public that is made accessible through aggregated statistical data. Such a view of the public has recent conceptual origins. Attention to the intellectual and technological precursors to distant reading shows that the assumptions and methodologies of statistically aggregating literary publics (e.g., the readers of British novels during a particular decade) have as much of a constitutive function as a quantitative one. Instead of unearthing knowledge sub specie aeternitatis, the precursors of distant reading in twentieth-century science, technology, and media helped give legitimacy to the very idea of an integrated public made accessible through technologies of aggregation. The history of this idea sug-

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

gests that there are important (although not necessarily disqualifying) limits to the application of aggregate quantitative methods to literary history. These methods are limited specifically because they flatten the diverse publics of literary history by applying a recent notional configuration of statistically aggregate populations to cultural fields that were not constituted or "imagined" according to such aggregate terms and technologies.[2]

Before moving to the genealogy of ideas and technologies underwriting these methods, I explain certain relevant aspects and assumptions of macroanalysis, distant reading, and other related big data projects. I refer to these as big data research methods following a somewhat informal usage of the term *big data* common in literary criticism and the digital humanities.[3] Following this scholarly trend, I consider big data research in the humanities to involve quantitative methodologies for investigating aggregate-scale trends in large corpora of digitized texts. The interconnected histories of ideas, technology, and media show several ways in which such big data analysis not only is incomplete as a measure of knowledge for literary history but also has important disjunctions in its theoretical assumptions about the public sphere.

Literary History and Reading Aggregate Data

In *Graphs, Maps, Trees*, Moretti conceives of three abstract models, named in the study's title. These models account for what he describes as "distinct 'sections' of the literary field." Moretti says graphs, in particular, signify "the system of novelistic genres as a whole" (91). Within this "whole" system, Moretti quantifies the rise and fall of particular subgenres of the novel, revealing cycles in this literary form's broader history. In the British novel, for example, the epistolary genre was largely replaced by the Gothic genre in the first decade of the nineteenth century, and then the Gothic was replaced by the historical novel beginning around 1820, and so on. Moretti argues that generational cycles of political conditions explain the rise and fall of these genres and that such cycles allow us to map patterns onto the novel's literary history. Moretti thus assumes that genre trends can be explained through rubrics of knowledge that are political and economic. The subgenres of the novel "subordinated narrative logic to the tempo of the short span, . . . and thus they also disappeared with the short span" (24). Temporary political and social conditions, according to Moretti, explain the contours revealed by what he elsewhere describes as "quantitative formalism" (*Distant Reading* 180). What is more, the graphing model moves from a data set to observations about cycles, and this methodology consequently requires the critic to view "genres as [a cycle's] morphological embodiment" (*Graphs* 17). Genre and cycle are thus intimately related. According to Moretti, their relation hinges on the explanation of causal mechanisms external to the genres themselves, which he says in turn relies on "some kind of generational mechanism . . . to account for the regularity of the novelistic cycle" (*Graphs* 22). Reading the cycles of the novel's literary history distantly, in other words, demands an account of the averaged arc of generational shifts.

The issue of averages is of course a function of statistical aggregation itself. Technologies of aggregation and methods of quantification in the digital humanities construe ideas such as "form" and "genre" in terms of "signals" of potential data (Heuser and Le-Khac 81). This approach tries to establish formal and generic differences based on equally registered information; that is to say, it places the "signals" of form and genre on a flattened plane of information to find patterns in the data. As Moretti puts it in a recent pamphlet for the Stanford Literary Lab, "Images come first, in our pamphlets, because—by visualizing empirical findings—they constitute *the specific object of study of*

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

computational criticism; they are our 'text'; the counterpart to what a well-defined excerpt is to close reading" (*Literature* 3). The data set is the object of reading or analysis. The authors of the Stanford Literary Lab's first pamphlet, published in 2011, acknowledge that the pursuit of empirical data lends itself to a troubling level of generality: "If all men in an audience wore pink, and all women blue, the colours would differentiate them *perfectly*, and tell us *nothing* about them" (Allison et al., *Quantitative Formalism* 18). The pamphlets are laudably experimental and often involve methodological self-criticism. In the first pamphlet, the authors note that the quantifiable measurement of genre only charts "differential features," not a given form's "inner structure" (18). Subsequent pamphlets attempt to move beyond "traits that classify so well, and explain so little," and indeed the Stanford Literary Lab's research often offers innovative interpretive work (24). Nonetheless, its computational methods of quantification present the literary field as a plot of nonspecific and leveled aggregate data.

This methodological assumption is not unique to the Stanford Literary Lab. It also appears in a recent analysis of the "evolution of literature" published in *Proceedings of the National Academy of Sciences of the United States of America*. The authors of the study quantify the notion of a literary "style of time" by referring to "content-free words" as signals of the stylistic relation between words and authors (Hughes et al. 7683). The authors recognize that this research makes interpretive or nonquantitative claims about the results of their data analysis (7684–85). And this move toward interpretation after quantitative research is consistent with the softer empiricism in many of the Stanford Literary Lab's pamphlets. For example, in *Style at the Scale of the Sentence*, the authors acknowledge that their attempts to measure style prove that "the 'digital' clearly needed the 'humanities' to make sense of its findings" and that the analysis of a large corpus "is a feedback loop wherein concepts inform measurements, and further measurements bring into play further concepts" (Allison et al., *Style* 28). While it is important for digital humanists to keep in mind the problem of putatively pure description, as well as the related "feedback loop" phenomenon (Liu 415), the problem I am considering here instead focuses on how quantification conceives of texts within large corpora as information in an integrated system, which is to say how it configures such data in an aggregate relation.[4]

Katie Trumpener complicates one aspect of this configuration by raising questions about the taxonomizing labels in Moretti's essay "Style, Inc." Trumpener shows that Moretti's method neglects the fact that the "same generic designation . . . could mean really rather different, if distantly related, things, and describe texts written on completely different scales" (164). The variety and multiplicity of a genre's meaning are foreclosed in distant reading. The measurement of an aggregated field of data about book titles or subgenres makes it especially difficult to account for compound, fragmented, and fluctuating styles, genres, and identities. Yet it is intrinsic to methodologies of statistical aggregation that such measurements require fixed and unitary structuring categories. Matthew Wickman similarly raises questions about Moretti's use of big data by pointing to a transformative crisis among eighteenth-century Scottish philosophers and mathematicians regarding the very concept of "number." For these philosophers of mathematical concepts, number becomes a contested ontological category, one that is "multiple" and "complex" (5). According to Wickman, Moretti's visualizations of big data "tacitly express a kind of historical wish to circumvent the crises of number and hence, to a degree, of history itself as the frustrated compulsion toward succession" (6). In other words, by applying the eighteenth-century

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

philosophical crisis in the concept of number to twenty-first-century methods of counting and analyzing texts, Wickman shows, first, how historical uncertainties about "wholeness" are lost within notions of integrated text corpora and, second, how the stability and integrity of mathematical conceptions underlying measurement cannot be taken for granted. The methods of aggregating and quantifying texts threaten to elide these historical and philosophical problems.

While the ontological presuppositions of the measurement of literature are problematic in the sense that they flatten texts into uniform data in large corpora, methods of aggregation are inadequate in other ways. In particular, big data methods explicitly depend on a statistically driven conception of the population of texts in a database. This underlying conception of a statistically aggregated population surfaces in Jockers's method, which he calls macroanalysis. Instead of positioning his method as "a strictly scientific practice," Jockers explains that "through the study and processing of large amounts of literary data, [macroanalysis] calls our attention to general trends and missed patterns that we must explore in detail and account for with new theories" (29). Jockers claims that aggregate data sets allow for "investigations at a scale that reaches or approaches a point of being comprehensive. The once inaccessible 'population' has become accessible and is fast replacing the random and representative sample" (7). This statement expresses two key assumptions: first, that aggregate data open up a "once inaccessible" public for analysis; second, that aggregation configures the "large amounts of literary data" as an integrated "population" that is intelligible as a coherent data set. Restating both assumptions, Jockers explains that the "object of analysis" in macroanalysis "shifts from looking at the individual occurrences of a feature in context" (i.e., close reading) "to looking at the trends and patterns of that feature aggregated over an entire corpus" (24). It is, as he also says, a "study of word 'data' or derivative 'information' about word behavior at the scale of an entire corpus" (25). Crucially, Jockers configures aggregate data as more than an ostensibly comprehensive archive; it also becomes an integrated corpus, a unified network of interconnected social information encased in numbers.

Moretti expresses related presuppositions when he says in reference to thousands of nineteenth-century British novels that "a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases: it's a collective system, that should be grasped as a whole" (*Graphs* 4). Jockers again says that the "study of literature should be approached not simply as an examination of seminal works but as an examination of an aggregated ecosystem or 'economy' of texts" (32). The ecosystem metaphor does not characterize the actual procedures and assumptions of microanalysis. More important, the relation of texts to one another is not equivalent to the configuration of data in macroeconomics, which Jockers takes as an analog for his methodology (24–26). This is because the economic value attributed to a single, small unit in macroeconomics (e.g., one dollar) is not equivalent to one sentence, one book, or one title in the cultural field. In the case of genres, for instance, generic categories such as the Gothic were crafted across interpretative histories and often in relation to certain exemplary texts that operated as centers of gravity in the constellation of books in any given marketplace. The category of the Gothic, which is one of the genres most commonly analyzed in distant reading, is a case in point. In the address to the reader prefacing Clara Reeve's *The Champion of Virtue* (1777), for example, the writer describes her book as the "literary offspring" of Horace Walpole's *The Castle of Otranto* (1764), a novel whose "plan" she follows in her own work (vi). Big data methods by definition would flatten out

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

Walpole's book, reducing it to a statistically negligible data point in aggregate literary history. In ways that the distant flattening of information within a system cannot adequately address, close reading can often provide a textured account of not only the particular novel but also the formation of its genre.

A closely related problem with the patterns in literary history identified by big data methods is that they transform an interpretive concept (the Gothic novel) into an aggregate measurable category. This method commits a type of equivocation fallacy, or at least performs a sleight of hand, a veiled transformation of the granular and responsive idea of genre into a quantifiable category in a data set. Distant reading and macroanalysis are shifting not only the methods of inquiry but also its objects, for they redefine the senses of terms such as *genre* in ways that borrow the names of historical disciplinary terms but not the underlying assumptions about textual meaning. More generally, many big data methods obscure historical conceptions of literary texts as objects forged in cultural fields—fields that existed before technologies of aggregation. These methods further transform already altered categories (e.g., the Gothic novel as viewed through the conceptual prism of data) by configuring their principal relations according to an anachronistic (and seemingly ahistorical) account of a statistically aggregate public. The analysis of large corpora thus employs but transforms the concepts established by the historicist, theoretical, archival, and close reading methods that aggregate quantification purports to supplant. What these quantitative methods mean by "the novel," for instance, is an object transfigured by computational technologies and unrecognizable to its various granular births and readerly conceptions in literary history.

Big data methods rely on large corpora composed of texts generally no longer subject to copyright, and most texts analyzed by these methods were produced long before there were technologies of public aggregation and computational measurement. The gap between the texts' composition and their analysis attests to the disjunction between method and object, big data and the local features of literary production. Methods such as distant reading and macroanalysis often eschew interpretive and conventional historicist work in order to apply big data analysis to epochs of literary history whose social and cultural fields were constituted through contingent and often irregular practices of reading, not aggregation. As an example of this disjunction, Jockers's method proceeds in terms of "analysis" rather than "reading," because the former term "places the emphasis on the systematic examination of data, on the quantifiable methodology. It deemphasizes the more interpretive act of 'reading'" (25). However, what Jockers affirms is not so much a quantifiable methodology as a certain configuration of the aggregated objects under scrutiny. The methodology entails an epistemic structure—a way of constituting knowledge. "The result of such macroscopic investigation," as Jockers puts it, "is contextualization on an unprecedented scale. The underlying assumption is that by exploring the literary record writ large, we will better understand the context in which individual texts exist . . ." (27).

The Epistemic History of Aggregate Publics

The literary record writ large. The macroeconomy of texts. A now-accessible aggregate population. Such conceptual scaffolding was built across a series of historical developments, which provided the conditions of possibility and the intellectual plausibility for the quantitative analysis of large corpora of texts. Other scholars have rightly considered digital methods of analysis and pedagogical practices from the perspective of changes to higher education and the nature of academic disciplines. Richard Grusin, for example, has argued that the advent of the digital humani-

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

ties dovetails with the corporatization of the university and also manifests a neoliberal political economy. Rachel Sagner Buurma and Laura Heffernan argue that recent rejections of "inherited or rote interpretive practices" by critics such as Moretti are in fact forms of historicist self-critique rather than the "displacement" of previous forms of academic criticism (618, 617). David Golumbia discusses the origins of computational technologies in terms of an emerging biopolitical regime. While these accounts consider digital methodologies as phenomena of broader shifts in American higher education and the political economy of a globalized technological system, what interests me is that big data analysis patterns literary history after an integrated mass public—a network of potentially analyzable social information encased in numbers.

To evaluate this configuration of literary history's populations as data sets, we have to ask questions about the conceptual framework that supports and legitimizes the type of knowledge yielded by big data analysis. The history of ideas and big data analysis ought not to be taken as "immaculately separate human and machinic orders," something that Alan Liu argues many digital humanities projects tend to do (416), nor should the methods of aggregation be used and viewed as if inoculated from the contingencies of an epistemic history. It is true that Jockers and Moretti acknowledge the *Annales* school of historiography as a precursor in the pursuit of, in Jockers's words, "applying quantitative and social-scientific methods in order to study the history of the 'long-term'" (19). Both also acknowledge debts to the Russian formalists, particularly in the scholarship of quantitative formalism that characterizes much of the research of the Stanford Literary Lab. But the terms of knowledge production's legitimacy have an intellectual history extending beyond any one practitioner's theoretical debts. The conditions for computational modes of reading and the types of knowledge they yield are strands woven from a multifarious historical fabric.

What are the epistemic assumptions and intellectual genealogies of reading or analyzing aggregate-scale data? And how is the aggregate analysis of social information bound up with other methods of quantification, as well as with nonscientific ways of imagining order? One of the genealogical strands derives from changes to the wider media ecology of the United States during the twentieth century. In particular, a unique brand of social information encased in numbers gained new intellectual traction in the United States beginning in the 1920s and 1930s, penetrating into the most private domains of personal existence by the middle of the century. Mary Poovey has shown that late-nineteenth-century scholars in the social and natural sciences earlier "sought not to generate knowledge that was simultaneously true to nature and systematic but to *model* the *range of the normal* or sometimes simply to create the most sophisticated models from available data" (3). The integrated system of knowledge subsumed granular particularity within "the *range of the normal*."

By the first decades of the twentieth century, methods of quantification such as surveys and the graphing of social data allowed social scientists to make claims about the contours of American life. As Sarah E. Igo demonstrates, the effects of these new methods created a feedback loop about the nature of the public under scrutiny. These new social-scientific technologies served a constitutive function in the public sphere: they began not only to imagine the public but also to inform the public imagination. As Igo puts it, "[A]ggregate data gave shape and substance to a 'mass public.' Midcentury social scientists were covert nation-builders, conjuring up a collective that could be visualized only because it was radically simplified." Scholars relying on aggregate-scale social data "offered more than simple summaries of data:

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

they encouraged new ways of seeing, perceiving, and imagining." Through their research methods, the quantifiers of social data "subtly transformed the entities under investigation." Igo concludes, "Ultimately, it would become nearly impossible to know the nation apart from their charts and curves" (18). In other words, this understanding of information encased in aggregate data had unique effects on how the collectives being studied understood themselves; it helped the notion of an aggregate public emerge as a legitimate account of the population. The idea of "public opinion," for instance, transformed from what G. W. F. Hegel described in the nineteenth century as "the unorganized way in which a people's opinions and wishes are made known," primarily through "transactions" in print media (353), to what George Horace Gallup and other twentieth-century pollsters conceived of as a "distinctly precise and systematic" measurement (Igo 106). Increasingly, individuals came to conceive of themselves in relation to those measurements, which also shaped university disciplines, marketing practices, and consumer behavior (Igo 107–10). Such aggregate knowledge helped create and reinforce the idea of a measurable aggregate public.

Other converging historical developments allowed for the legitimacy of the idea of an intelligible and accessible aggregate public to take shape. For one, the social sciences were undergoing significant transformations during the 1940s and 1950s, particularly as many influential academics repositioned themselves on what John Guillory describes as "the epistemic hierarchy of the disciplines." Social scientists, according to Guillory, forged a new professional status by "discard[ing] interpretation as much as possible from their methodological repertoire" (498). For example, Geoffrey Hawthorn notes that the Harvard sociologist Talcott Parsons formulated a method that he called "structural functionalism" in order to articulate social facts on the basis of avowedly scientific observations of macrolevel phenomena. As Hawthorn puts it, "[I]n exactly the way in which the instruments of survey analysis served to constitute a professional technique, functionalism served to constitute a professional value" (214). The method and its theories were part of disciplinary struggles over what counted as legitimate academic knowledge.

During this period, social scientists also disavowed the subjectivity of reading and turned instead to neutral observation—a methodological shift that helped reconstitute disciplinary boundaries during the 1940s and 1950s (Schryer 34–40). What is more, new types of social data became relevant and scrutinized. Data culled by emerging modern political polling provide a case in point. Before the 1936 presidential election in the United States, straw polls were conducted through imprecise methodologies by major magazines. During the 1936 election campaign, *The Literary Digest* predicted a win for the Republican, Alf Landon; however, Franklin D. Roosevelt won the election by the largest margin in history. The only researchers to predict the result accurately were Gallup and his colleagues, Archibald Crossley and Elmo Roper. By the 1940 presidential election, Gallup's team had nearly cornered the polling market. "The political wisdom of the common people can now be settled," Gallup pronounced, "on the basis of a mountain of factual data" (qtd. in Igo 122). Gallup and his fellow researchers claimed to have found the statistical methodology for arriving at "the average American," a coupling of method and concept that shaped not only the science for gathering and explaining social data but also how Americans understood the democratic process itself.

As Gallup, Crossley, and Roper acknowledged, the science of polling led them to weigh more heavily the answers of those who were more likely to vote, and so their polls undercounted women, racial minorities, and the impoverished. Political conditions shaped the epistemic contours of what counted as

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

relevant social data. In addition to bestowing predictability on election coverage, the polling concept of the "average American" deployed what Gallup insisted on calling "scientific evidence" and empirically verifiable scales to produce the concept of a measurable population. The science of averages and social quantification therefore legitimized a particular depiction of collective identity—a veneer that obfuscated the inequalities and political disparities of mid-century public life. Indeed, the irony underlying the rise of this version of the mass public is that both the quantitative methodology and its attendant view of an accessible population are coterminous with the postwar resurgence in voter-suppression tactics. If some voices or data points count less in measuring an imagined community because empirically those voices or data contribute less to the institutional shape of the population of that community, such a fact attests as much to what quantification elided as to what it revealed or demonstrated. The methods for accessing a newly emerging aggregate body entailed political and cultural structures of thought, which enabled the verification of the methods themselves while also relying on problematic notions of statistical averages. Studies such as Gallup and Roper's culling of public data aimed "to sketch the collective whole of society" (Igo 14), but such an idea of the collective was predicated on an aggregate body rife with politicized conditions of possibility.

While the emerging disciplinary boundaries of twentieth-century sociology and Gallup's new technologies of polling both signal a shift in the ways American communities imagined themselves as an aggregated collective, the notion of a mass public or even of statistical measurement is not unique to the twentieth century. Harald Westergaard, for example, explains that the "centralisation" of statistics (i.e., "a single institution having charge of the chief statistical subjects" of a nation or government [243]) began in most European countries in the 1870s and became complete after the turn of the century (242–45). Still, earlier kinds of statistics differ from the new types of data collected in the twentieth century. An especially influential example is Alfred C. Kinsey, Wardell Baxter Pomeroy, and Clyde E. Martin's *Sexual Behavior in the Human Male* (1948), which collects and interprets data on adolescent orgasms (182–92), the relation between masturbation and age (238–42), the frequency of coitus among various social groups (335–62), and even "animal contacts" of a sexual nature (667; see 667–79). The Kinsey report was not so much a break with the older idea of a mass public as it was a further permutation of it. For instance, Stacey Margolis has analyzed how the notion of a mass public inflected American conceptions of democracy before the existence of public opinion polls. According to Margolis, American writers and intellectuals tended to conceive of mass democratic life in the nineteenth century in informal and often ad hoc ways. Yet the Kinsey report reveals an emerging conception of publics as measurable through data about behavior.

The method of statistical aggregation and the idea of an aggregate public had other important precursors. For example, Herman Hollerith, who founded the company that would become IBM, created a machine to help tabulate the 1890 census. As Bill Kovarik explains, this early computer "[made] it easier to search through massive amounts of data for specific facts, such as how fast manufacturing was growing, or how many Irish now lived in New York, or the birth rate in Chicago" (352). The scale, scope, and type of social data collected, and the emerging technological means for analyzing them, increasingly gave public weight to aggregate social data in the United States. By the first decades of the twentieth century, as Alain Badiou explains, professional statisticians and sociologists formalized sweeping attempts "to submit the figure of communitarian bonds to number" (2)—that

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

is, to reconceive the public sphere as an aggregation of quantifiable behavior. From data sets of census information to scientific surveys, from Robert Lynd and Helen Merrell Lynd's two-part *Middletown* (1929, 1937), a sociological study of Muncie, Indiana, to *Sexual Behavior in the Human Male*, a spectrum of new modes of information gathering and analysis was, to borrow again from Igo, "trumpeted as both a sign of, and a route toward, a modern culture that prized empirical investigation over faith, tradition, approximation, common sense, and guesswork" (5). The unprecedented influx of facts and figures helped shape public understandings of a certain type of collective body. It had widespread political and social effects on a newly aggregated America.[5]

At the same time that such aggregate data garnered new legitimacy and helped create a distinct conception of public life during the middle decades of the twentieth century, the rise of an aggregate public benefited from the emergence of new technologies in American government and the corporate world. For example, as Paul E. Ceruzzi explains, the UNIVAC line of mainframe computers "inaugurated the era of large computers for what is now called 'data processing' applications" (30). General Electric initially purchased the UNIVAC line in 1954 for such tasks as "long-range planning, market forecasting based on demographic data, revamping production processes to reduce inventories and shipping delays, and similar jobs requiring a more ambitious use of corporate information" (33). As Kovarik says, data tabulators had also become the principal technology for census gathering: "In the 1940 census, for example, IBM tabulators . . . could process cards containing census information at the rate of 400 a minute, and from these, twelve separate bits of statistical information could be extracted" (353). Here, an aggregate public met computational technologies of scale. Yet the first public test of a computer was broadcast on election night in 1952 on CBS, when a UNIVAC (unlike most other means of expert forecasting) accurately predicted the number of electoral votes Dwight D. Eisenhower won in his victory (Kovarik 358). The public met its aggregator on prime-time television.

Such moments in the public history and intellectual genealogy of the science and technology behind big data analysis helped create a sense of an aggregate mass public and simultaneously shored up the legitimacy of the technology itself. The idea of accessing literary history through massive text corpora is a child of this intellectual and technological history. Consequently, this vein of the digital humanities is freighted with that history's assumptions and limitations, even if it is not necessarily bound to repeat its errors. Indeed, the ability of computational technologies to make demonstrable claims about certain objects of inquiry required that the underlying theoretical assumptions entailed in such empirical claims were first legitimized. As Richard Powers puts it in his novel *Gain* (1998), "the greatest merchandising prize" of postwar corporate marketing technologies "was the idea of market research itself." "By the time Sputnik left the earth," Powers's narrator says, "the industry of needs creation had learned to see the blind taste test as its own product." In other words, the science of marketing research "had to sell science, scientifically. And the resulting combination serviced huge sectors of the psychic economy" (326). Even as various postwar aggregation technologies helped sell the idea of an aggregatable consumer society, they also created new forms of consumable desire and opened up new markets for production and consumption. With the birth of an aggregate public comes the notion of aggregate desires. This mode of quantification and analysis was thus configured in terms of market preferences and the projection of incentives to an entire population (Kovarik 217–48). The data used in this quantification and analysis not only tell us about ourselves; they also constitute

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

our selves. Computing technology, the possibilities of analyzing an aggregate public to predict marketing outcomes, and the earlier advent of mass-communications technologies all reinforced the value of big data and disseminated a new public conception for the now-aggregated masses.

Aggregation and the Digital Humanities

The epistemic history of technologies of aggregation qualifies the scope and applicability of big data methods of analysis in the humanities. This history reveals the contingency of the methods' assumptions. These methods often lack textured accounts of the public spheres that they aggregate through large corpora. The notion of mapping the evolutions of the subgenres of late-nineteenth-century British novels, for instance, is a fascinating project that, as Paul A. Youngman and Ted Carmichael say, "can enhance the conclusions researchers draw or have drawn" (287). But these methodologies first need to be more thoroughly attuned to the particular conditions of the publics and of the texts contained in the corpus, as well as to the granular features of the texts themselves. The methodological, historical, and theoretical limitations of quantitative analyses of large corpora often go unheeded among digital humanists who reject nonquantitative modes of reading. After all, they argue, why spend twenty pages of one's scholarly writing on a British novel from 1854 that sold only a few hundred copies and that, perhaps, only a handful of people have since read?

This dismissive view of the limited scale of historicist, philological, archival, and close reading methodologies presents aggregation as a fulfillment of an empirical and scholarly mandate. As Jockers puts it, "Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history" (7). For scholars "concerned with incorporating larger numbers of texts and viewing works in a broader social, industrial, and even transnational context," Erlin and Tatlock similarly contend that close reading is insufficient because it "does not adequately answer questions about the production and circulation of books, taste formation, or even necessarily about the relative position of texts in the literary field." Close reading might yield certain limited claims about individual texts, these scholars argue, but it is unsuitable as a literary-historical method by virtue of what "evidence" it yields in comparison with the "review and interpretation of data" (9).

It is true that scholarly analyses of the grammatical features of style, such as certain claims in Hoyt Long and Richard Jean So's essay on the diffusion of stream of consciousness techniques, count features of form instead of relying directly on models of literary publics. Such work often uncovers unexpected formal trends, such as Long and So's discovery that there is a greater concentration of stream of consciousness writing in the romance novels of the British writer Jeffery Farnol than in modernist texts like Djuna Barnes's *Nightwood* ("Turbulent Flow" 354). However, moving from identifying the concentration of formal features in individual texts to wider claims about a literary field would require historically particularized theories of the public spheres behind text corpora. For instance, literary publishing is embedded within granular material conditions that are often governed by inequality—who reads, who buys, who writes and when, who edits and sells. Archival work and the close analysis of work by women of color begin to correct this structural inequality. However, taking text corpora as the measure of literary history privileges the influence of publishers and editors, with their prejudices and appeals to market preferences. Big data methods take the structural inequalities of the literary marketplace and describe them as sources of a newly found objectivity.

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

Distant reading and the notion of a statistically analyzable population also privilege an averaged aggregate public above textual or cultural outliers. For example, during the 1930s the modernist writer Mina Loy wrote but was not able to publish her autobiographical novel "Goy Israels," seemingly because of her uncertain position among the avant-garde modernists, her Jewish heritage, her mystical commitments to Christian Science, the gender politics of the moment, and her endorsement of "crossbreeding between races as a means of asserting control over evolution" (Vetter 55; see also Churchill et al.). As a result, the quantification by Hughes and his colleagues measures certain signals in the institutionally sanctioned objects of published modernism but does not account for Loy's text (7685). "Goy Israels" falls outside the measurement of modernist style. Yet, does being unpublished make Loy's novel less representative of the literary-historical objects of analysis called "the novel" or "modernist style"? Given the repeated claims that large corpora give scholars access to the great unread of published but very-little-read books—and that such texts make up the neglected data that is essential for literary history—it seems hard to imagine that unpublished works would be any less significant to the methodological aims of big data analysis than a little-read "industrial novel" (Moretti, "Conjectures" 55).

Ontological problems thus plague the large corpora that produce the objects of inquiry for computational and quantitative methodologies. The trouble is not only with how the methodologies evaluate their "*specific object of study*" but also the source of the objects themselves (Moretti, *Literature* 3). Unpublished texts are either irrelevant, perhaps because they would be averaged out in the other data of the large corpora, or available in theory to be incorporated into ever-more-comprehensive corpora. The impossibility of incorporating the vast body of unpublished texts attests to the incompleteness of the large corpora, even though such comprehensiveness would be required to meet the demands of scale that digital humanists often invoke to justify big data methods. In either case, though, the more troubling problem is that these massive text corpora subordinate—and perhaps even elide—the political and social inequality inherent in the process of publication in order to establish a coherent object of study. Big data methods' neglect of the structural inequalities of publication is all the more ironic given that Moretti links genre development and political cycles in *Graphs, Maps, Trees* (24–26).

It is easy to imagine another sense in which text corpora are limited and misleading—a problem that computer scientists call dirty data. If thousands of novels are written but not published every year, works are also regularly published disjunctively, which is to say fraudulently, posthumously, in multiple editions, and so on. For example, in 1916 an important and troubling novel by Mark Twain titled *The Mysterious Stranger* was published. The problem, however, was that Twain never wrote the book. As John S. Tuckey shows in his monograph on Twain's manuscripts, the book was a fraud pieced together by Twain's literary executor, Albert Bigelow Paine, and an editor at Harper and Brothers, Frederick A. Duneka. As Robert H. Hirst explains, the 1916 book was based on "the earliest rather than the latest of the manuscripts," and Paine and Duneka "deleted fully one-fourth of the author's words; they wrote into the story the character of an astrologer, who did not even appear in the manuscript . . . [and] they appropriated the concluding chapter Mark Twain had written for his latest and longest version," altering the names of many characters to make it all fit (202). None of this literary fabrication would be known had Tuckey not conducted archival work on Twain's late unpublished manuscripts, which had generated virtually no interest among scholars (Csicsila). Given the extensive archival work

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

required to expose this false Twain novel, one wonders how regularly publishers produced such texts and how this practice might affect the quantification of style and genre.

Texts such as "Goy Israels" and *The Mysterious Stranger* can be described as uncertain cases of counting. Does the vast trove of unpublished texts in literary history count in quantitative measures? What about, say, James Welch's untitled and unpublished first book (Orton)?

How can we count texts with other types of temporal disjunctions, such as fictions serialized across several decades and multiple generations? Indeed, both early modern printing and the history of the book attest to the variability of the objects we associate with writing (J. Fleming). Counting is a problem not only when one quantifies genre but also when one measures what Hughes and his colleagues describe as the "evolution" of literary style. Technologies of aggregation and measurement obscure the temporal fluidity of publication and the disjunctions in "the regularity of the novelistic cycle" (to use one of Moretti's explanatory phrases [*Graphs* 22]) in order to shore up the conceptual coherence of large corpora. The ambiguities of counting get lost in these particular technologies of aggregation and in many digital humanists' methods of quantification, despite the rigorous tradition of philosophical and sociological debate about the nature of enumeration and statistical aggregation (Wickman 5; Stigler 229–33). In short, what counts is not an easy question to answer, and referring to the aggregate scale of the database to dismiss the question merely elides the potential for ambiguity in large corpora. Indeed, to legitimize big data by reference to the bigness of the data is only a very powerful tautology.

Those who support big data methods might still ask, does the scale of distant reading not average out the uncertainty in large corpora? After all, what are a few unpublished texts or mendaciously published novels among a sea of data? Yet, instead of being merely anecdotal, the cases of *The Mysterious Stranger* and "Goy Israels" demonstrate the possibility of uncounted and uncategorized cases, which substantially increase the uncertainty about the information in a system. The great unknown, in contrast to the great unread, haunts text corpora. These uncounted and disjunctive data attest to the informational shortcomings in methodologies that often style themselves as supplanting other methods (which undoubtedly have their own limitations). As Youngman and Carmichael note, the problem of deficiencies in data is well-known in computer science: "there is no such thing as a clean data set, even when the data set is small" (291). But my point is not that there are "many technical compromises and approximations one must typically accept in order to get on with digital humanities projects," as one recent pamphlet from the Stanford Literary Lab puts it (Algee-Hewitt and McGurl 3). The technical problems with curating large corpora illustrate the uncertainty about the relation between the data of the corpora and the historical realities to which they refer. The data point representing a Gothic novel in a large corpus is ontologically not the same object as the material book published in a given literary-historical moment.

Because proponents of computational quantitative methods often assert the innovative importance of those methods by pointing to the limitations of close reading, they regularly gloss over the philosophical and theoretical problems intrinsic to the objects of quantification. The problem of counting that underlies aggregate-scale analysis troubles the unitary character of the data—or, more specifically, the very notion of aggregate integration that legitimizes large corpora. What counts and when and why? The historicist, philological, archival, and close reading methodologies that have been the mainstays of literary criticism negotiate these questions in undeniably

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

inefficient, contingent, and limited ways; there has never been some innocent or flawless moment in the history of interpretation. However, aggregate-scale analysis cannot achieve greater certainty by producing notional configurations of fields of data that elide the problems of counting or the conditions of book publication. There are ontological questions that quantification often skirts and that many of its advocates even disregard by deferring to the logic of scale. But of course this logic has a conceptual history. Without more-textured theorizations of text corpora themselves, the granular uncertainty and multiplicity of texts become perilously irrelevant.

The intellectual history sketched in the previous section shows not only that there are problems with the data of large corpora but also that exhaustive, unambiguously integrated, and theoretically "clean" corpora of texts (even if they were achievable) would still bring the fraught presupposition of an aggregate public to the task of quantitative analysis. As I have argued, many big data methods overlay patterns onto populations not imagined according to such terms, making them into statistically aggregate mass publics. Census technologies have existed since antiquity, but the types of counting relevant to the style of modernism or to a particular novelistic subgenre cannot be flattened to the same form as census data, because, as big data methodologists acknowledge, such data derive from cultural and social fields. Yet the social fields of the historical publics themselves—e.g., the competing accounts of collective identity in Victorian England, or in China during the early development of its novel forms—were not imagined in terms of statistical aggregation. Therefore, distant reading's application of twentieth- and twenty-first-century models of aggregation often relies on an anachronistic and nonspecific theory of the public sphere, as if all publics were quantifiable in the same way. Such an approach can even dehistoricize itself methodologically at the same time that it dehistoricizes its data, ignoring the heterogeneity of the scores of publics across literary history. Insofar as the distant reading of big data theorizes social information about publics sub specie aeternitatis, such a method positions its claims using ahistorical epistemic assumptions.

If big data quantitative methods tend to understand history and public spheres as an unchanging repository of potentially aggregated social data, then surely there would be disjunctions in the results those methods produced. And in fact there seem to be. In one essay in *Distant Reading*, for example, Moretti contrasts social data about literacy in eighteenth-century Europe, the growing market for European novels, and contemporaneous novels in China to explain both the rise of the European novel in terms of distraction and the creation of a marketplace for novels predicated on a lack of concentrated reading (176). However, in a different essay in the same book his analysis of stylistic data about the development of the European novel suggests that literary devices such as the "difficulty" of metaphors hold "the secret" to market success (203). In the first account, the rapid expansion of the eighteenth-century novel is explicable in terms of "reading a lot more than in the past, avidly, at times passionately, but probably more often than not also superficially, quickly, even a little erratically" (174). In the second account, however, Moretti's distant reading of the relation between the novel's style and commercial success (from 1740 to 1850) suggests that "by puzzling and challenging readers, metaphors induced them to take *an active interest in the novel* from the very first word" (204). Perhaps Moretti's claims about distracted consumption and the novel's encouragement of active interest through difficult metaphor are not contradictory, but the tension between the two accounts at least attests to a public sphere that is more heterogeneous than the integrated network of text corpora would suggest.

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

This tension between Moretti's two accounts also shows that literary publics are not more accessible through statistical aggregation; rather, these publics are shown to be less than coherent and thus require analyses of their granular features and individual texts.

Other recent scholarship offers additional examples of disjunctions within aggregated corpora. Specifically, in Long and So's article on the "turbulent flow" of stream of consciousness techniques in world literature, their analysis of grammatical structures in their early-twentieth-century corpus suggests how language itself "threw up obstacles in some places and not others and that these patterns of interference internally varied" ("Turbulent Flow" 364). The corpus did not always yield access to intelligible patterns or trends at the aggregate level. Long and So convincingly identify local patterns that afford fascinating new readings of particular texts (as I discuss above). However, in a way that confirms the importance of theorizing the variegated texture of literary cultures and publics, Long and So's analysis attests to the disaggregation and disruptions that exist throughout a global literary field. The diffusion of a certain technique throughout the text corpus exposes, as Long and So put it, "how this diffusion is marked by constant, heterogeneous variance" (365).[6] The intellectual history I have offered in this essay suggests why variance and heterogeneity characterize the publics of literary history. As a result of such heterogeneity, these publics are often incommensurable with the aggregations created by large corpora.

The blanket application of aggregate methodologies would simply ignore the diversity of the literary-historical terrain to which large corpora putatively give us access. Trumpener's response to Moretti's "Style, Inc." raises yet another example of the problems with aggregation. In his analysis of seven thousand British novels published from 1740 to 1850, Moretti argues that longer titles disappeared "because between the size of the market, and the length of titles, a strong negative correlation emerged: as the one expanded, the other contracted" ("Style" 139). Moretti shows a correlation between the commodification of titles in a competitive book market and changes to certain formal aspects of the novel. However, Trumpener says such an explanation would be ill-suited to account for phenomena during the same period in Germany, where "the absence of a single literary center and a massified book trade meant literary life remained more diffuse, less commercialized." There was no mass integration of the German cultural arena at this moment, according to Trumpener, and so the underlying conditions of the public sphere were multifarious and diffuse. The changes in the forms of titles in Germany, which "at first glance appear consonant with contemporaneous British publishing practices" (167), are not in fact explicable as equivalent effects in an integrated system. Thus, even as there are significant ambiguities underlying counting and data, there is also historical incommensurability among literary publics—a lack of shared standards that are obscured by the information system of text corpora. Indeed, Trumpener's point further disarticulates the fields of global literary history, which demand textured theories of its various publics and literary cultures.

The epistemic history of technologies of aggregation suggests that big data methodologies tend to imagine publics that did not exist under—or imagine themselves through—such statistically aggregate terms. Quantitative analyses of the Victorian novel make presuppositions about populations and public spheres that differed significantly from those in the United States in the 1980s or 2010s, not just in terms of how these communities imagined themselves but also in the technologies that bound, separated, and constituted them. An unqualified distant reading of literary data sets implies that the cultural and social spheres throughout history are

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

uniformly measurable, as if what are being measured were merely the internal changes of data within a monolithic social totality. The interpretive alternatives have their own complexities. Michael Warner notes that the "kind of public that comes into being only in relation to texts and their circulation" differs significantly from the concrete publics of an audience, city, or nation (66). Yet quantitative methodologies for investigating aggregate-scale trends tend to level their objects of inquiry, as if there were no difference between vital statistics and statistics about the rise and fall of the number of certain types of texts. Such an approach posits undifferentiated, statistically imagined populations. Tracing the historical development of an integrated, aggregate public and its relevant technologies suggests that the conditions of possibility for their modes of quantification are in large part imprints of the notional arrangements that big data projects bring to the critical table.

Notes

This essay has benefited from the insightful feedback of J. D. Connor, Suzanna Geiser, and Daniel Powell, as well as Molly Des Jardin and the other members of the Word Lab at the University of Pennsylvania.

1. In his account of Moretti's computational analysis of *Antigone*, P. Fleming argues that Moretti's method falters in its attempt to sustain a dialectic between measurement and other modes of interpretation.

2. Anderson's classic account similarly argues that print media create publics: "the newspaper reader, observing exact replicas of his own paper being consumed by his subway, barbershop, or residential neighbours, is continually reassured that the imagined world is visibly rooted in everyday life. . . . [F]iction seeps quietly and continuously into reality, creating the remarkable confidence of community in anonymity which is the hallmark of modern nations" (35–36). According to Anderson, public life was imagined through the media of print capitalism. The statistical aggregations of distant reading and microanalysis presuppose social imaginaries that differ from the print media that created confidence in earlier conceptions of public identity. Nineteenth-century newspaper readers did not conceive of themselves as part of a statistical aggregation, nor did readers of novels view the objects in their hands as quantifiable units that tied them to a specifically statistical body or numerical network. The discrepancy between computational methods and their objects of inquiry ought to give us pause when considering what sort of public those big data methods disclose. It is often like using oranges to measure apples.

3. Jockers 7–8; Tangherlini and Leonard 726; Drouin 110–12; Anderson and Blanke 150–52; "Digging into Data"; Wickman 4n4; Birkerts; Looser; Youngman and Carmichael; Buurma and Heffernan 616. The meaning of the term *big data* ranges widely across humanistic academic disciplines, computer science, and the tech industry. For example, as Manovich explains the term's common meaning in the tech industry, "big data" refers to "data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time" (460). According to industry standards, supercomputers are required for processing these data sets. In contrast, the largest data sets in the digital humanities, as Manovich notes, "are much smaller than big data used by scientists; in fact, if we use the industry's definition, almost none of them qualify as big data" (461). However, as Ward and Barker show, the term was fashioned simultaneously in industry, media, and various disciplines in the academy; as they put it, there are "various stakeholders" in the meaning of "big data," and so the phrase's uses in industry are by no means original or final.

This essay evaluates the intellectual history that enables aggregate data to have plausibility or legitimacy as "research assets" for a type of scholarship "independent of interpretations." Therefore, despite the fact that computer scientists might not view, say, seven thousand British novels as big data, this essay follows the relatively informal usage of the term as it is common in the disciplinary terrain of literary criticism and the digital humanities. Indeed, Schöch makes the helpful observation that "the distinctive mark of big data in the humanities seems to be a methodological shift rather than a primarily technological one" (6–7). By using the term *big data*, then, this essay aims to investigate a methodology of statistical aggregation and quantification, not the volume of the large corpora themselves.

4. In terms of the history of literary criticism, there is good reason to welcome this turn toward wider fields of analysis. As So and Long observe, a desire for quantitative and sociological methods stretches back to writings from the 1930s and 1940s, including work by Kenneth Burke and Caroline Spurgeon ("Network Analysis" 151–54). Work by Ian Watt and others in the 1950s similarly tried to situate literary texts within broader social histories, thus working against the theoretical grain of the New Criticism that was then dominant in American academic departments. The move away from the New Critical view of a literary text as a self-enclosed work of art was supported by such studies as Radway's and by

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

American interpretations of Pierre Bourdieu's work beginning in the 1980s (e.g., Smith 132). The interpretive modes of close reading and philological research before the rise of quantification and sociological methods were by no means flawless or innocently ahistorical.

5. For parallel developments in Great Britain, see Adams; Thompson; Schwarzkopf.

6. In an earlier article, So and Long appear to be more sanguine about the continuities and patterns that may be discovered through computational approaches: "quantitative techniques like network analysis and network visualization can be a useful aid for rendering aspects of social structure visible at a large enough scale to observe the 'strict, nonrandom regularity' that small-scale random phenomena tend to create in their collective action" ("Network Analysis" 155).

Works Cited

Adams, Maeve E. "Numbers and Narratives: Epistemologies of Aggregation in British Statistics and Social Realism, c. 1790–1880." Crook and O'Hara, pp. 101–20.

Algee-Hewitt, Mark, and Mark McGurl. *Between Canon and Corpus: Six Perspectives on Twentieth-Century Novels*. *Stanford Literary Lab*, Jan. 2015, litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Allison, Sarah, et al. *Quantitative Formalism: An Experiment*. *Stanford Literary Lab*, 15 Jan. 2011, litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Allison, Sarah, et al. *Style at the Scale of the Sentence*. *Stanford Literary Lab*, June 2013, litlab.stanford.edu/LiteraryLabPamphlet5.pdf.

Anderson, Benedict. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Revised ed., Verso, 2006.

Anderson, Sheila, and Tobias Blanke. "Taking the Long View: From E-science Humanities to Humanities Digital Ecosystems." *Historical Social Research / Historische Sozialforschung*, vol. 37, no. 3, 2012, pp. 147–64.

Badiou, Alain. *Number and Numbers*. Translated by Robin Mackay, Polity, 2008.

Birkerts, Sven. "The Little Magazine in the World of Big Data." *Sewanee Review*, vol. 123, no. 2, Spring 2015, pp. 224–31.

Buurma, Rachel Sagner, and Laura Heffernan. "Interpretation, 1980 and 1880." *Victorian Studies*, vol. 55, no. 4, Summer 2013, pp. 615–28.

Ceruzzi, Paul E. *A History of Modern Computing*. 2nd ed., MIT P, 2003.

Churchill, Suzanne W., et al. *Mina Loy: Navigating the Avant-Garde*. July 2016, mina-loy.com/.

Crook, Tom, and Glen O'Hara, editors. *Statistics and the Public Sphere: Numbers and the People in Modern Britain, c. 1800–2000*. Routledge, 2011.

Csicsila, Joseph. "John S. Tuckey's *Mark Twain and Little Satan*." *The Mark Twain Annual*, vol. 8, 2010, pp. 14–18.

"Digging into Data Challenge." *Digging into Data*, diggingintodata.org/about.

Drouin, Jeffrey. "Close- and Distant-Reading Modernism: Network Analysis, Text Mining, and Teaching *The Little Review*." *The Journal of Modern Periodical Studies*, vol. 5, no. 1, 2014, pp. 110–35.

Erlin, Matt, and Lynne Tatlock. "'Distant Reading' and the Historiography of Nineteenth-Century German Literature." Introduction. *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Erlin and Tatlock, Camden House, 2014, pp. 1–25.

Fleming, Juliet. *Graffiti and the Writing Arts of Early Modern England*. U of Pennsylvania P, 2001.

Fleming, Paul. "Tragedy, for Example: Distant Reading and Exemplary Reading (Moretti)." *New Literary History*, vol. 48, no. 3, Summer 2017, pp. 437–55.

Gardiner, Eileen, and Ronald G. Musto. *The Digital Humanities: A Primer for Students and Scholars*. Cambridge UP, 2015.

Golumbia, David. *The Cultural Logic of Computation*. Harvard UP, 2009.

Grusin, Richard. "The Dark Side of the Digital Humanities: Dispatches from Two Recent MLA Conventions." *Differences: A Journal of Feminist Cultural Studies*, vol. 25, no. 1, 2014, pp. 79–92.

Guillory, John. "The Sokal Affair and the History of Criticism." *Critical Inquiry*, vol. 28, no. 2, Winter 2002, pp. 470–508.

Hawthorn, Geoffrey. *Enlightenment and Despair: A History of Sociology*. Cambridge UP, 1976.

Hegel, G. W. F. *Elements of the Philosophy of Right*. Translated by H. B. Nisbet, Cambridge UP, 2003.

Heuser, Ryan, and Long Le-Khac. "Learning to Read Data: Bringing Out the Humanistic in the Digital Humanities." *Victorian Novels*, vol. 54, no. 1, 2011, pp. 79–86.

Hirst, Robert H. "Note on the Text." *No. 44, the Mysterious Stranger*, by Mark Twain, edited by William M. Gibson, U of California P, 2003, pp. 201–02.

Hughes, James M., et al. "Quantitative Patterns of Stylistic Influence in the Evolution of Literature." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 20, 15 May 2012, pp. 7682–86.

Igo, Sarah E. *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Harvard UP, 2007.

Jockers, Matthew J. *Macroanalysis: Digital Methods and Literary History*. U of Illinois P, 2013.

Kinsey, Alfred C., et al. *Sexual Behavior in the Human Male*. W. B. Saunders, 1948.

Kovarik, Bill. *Revolutions in Communication: Media History from Gutenberg to the Digital Age*. 2nd ed., Bloomsbury, 2016.

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press

Liu, Alan. "The Meaning of the Digital Humanities." *PMLA*, vol. 128, no. 2, Mar. 2013, pp. 409–23.

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

Looser, Devoney. "British Women Writers, Big Data and Big Biography, 1780–1830." *Women's Writing*, vol. 22, no. 2, 2015, pp. 165–71.

Love, Heather. "Close but Not Deep: Literary Ethics and the Descriptive Turn." *New Literary History*, vol. 41, no. 2, Spring 2010, pp. 371–91.

Lynd, Robert, and Helen Merrell Lynd. *Middletown: A Study in Contemporary American Culture*. Harcourt, Brace, 1929.

———. *Middletown in Transition: A Study in Cultural Conflicts*. Harcourt, Brace, 1937.

Manovich, Lev. "Trending: The Promises and the Challenges of Big Social Data." *Debates in the Digital Humanities*, edited by Matthew K. Gold, U of Minnesota P, 2012, pp. 460–75.

Margolis, Stacey. *Fictions of Mass Democracy in Nineteenth-Century America*. Cambridge UP, 2015.

Michel, Jean-Baptiste, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, vol. 331, no. 6014, 14 Jan. 2011, pp. 176–82.

Moretti, Franco. "Conjectures on World Literature." *New Left Review*, no. 1, Jan.–Feb. 2000, pp. 54–68.

———. *Distant Reading*. Verso, 2013.

———. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2007.

———. *Literature, Measured*. *Stanford Literary Lab*, Apr. 2016, litlab.stanford.edu/LiteraryLabPamphlet12.pdf.

———. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)." *Critical Inquiry*, vol. 36, no. 1, Autumn 2009, pp. 134–58.

Orton, Thomas. "Keening Woman and Today: James Welch's Early Unpublished Novel." *Studies in American Indian Literatures*, vol. 18, no. 3, Fall 2006, pp. 52–57.

Poovey, Mary. *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. U of Chicago P, 2009.

Powers, Richard. *Gain*. Picador, 1998.

Radway, Janice. *Reading the Romance: Women, Patriarchy, and Popular Literature*. U of North Carolina P, 1984.

Reeve, Clara. *The Champion of Virtue: A Gothic Story: By the Editor of the Phoenix: A Translation of Barclay's Argenis*. London, 1795. *Eighteenth Century Collections Online*, find.galegroup.com.i.ezproxy.nypl.org/ecco/infomark.do?&source=gale&docLevel=FASCIMILE&prodId=ECCO&userGroupName=nypl&tabID=T001&docId=CW3312449510&type=multipage&contentSet=ECCOArticles&version=1.0.

Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities*, vol. 2, no. 3, Summer 2013, pp. 1–13.

Schryer, Stephen. *Fantasies of the New Class: Ideologies of Professionalism in Post–World War II American Fiction*. Columbia UP, 2012.

Schwarzkopf, Stefan. "The Statisticalization of the Consumer in British Market Research, c. 1920–1960: Profiling a Good Society." Crook and O'Hara, pp. 144–64.

Smith, Barbara Herrnstein. *Contingencies of Value*. Harvard UP, 1988.

So, Richard Jean, and Hoyt Long. "Network Analysis and the Sociology of Modernism." *Boundary 2*, vol. 40, no. 2, 2013, pp. 147–82.

Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard UP, 1986.

Tangherlini, Timothy R., and Peter Leonard. "Trawling the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research." *Poetics*, no. 41, 2013, pp. 725–49.

Thompson, James. "Printed Statistics and the Public Sphere: Numeracy, Electoral Politics and the Visual Culture of Numbers, 1880–1914." Crook and O'Hara, pp. 121–43.

Trumpener, Katie. "Paratext and Genre System: A Response to Franco Moretti." *Critical Inquiry*, no. 36, Autumn 2009, pp. 159–71.

Tuckey, John S. *Mark Twain and Little Satan: The Writing of* The Mysterious Stranger. Purdue UP, 1963.

Vetter, Lara. "Theories of Spiritual Evolution, Christian Science, and the 'Cosmopolitan Jew': Mina Loy and American Identity." *Journal of Modern Literature*, vol. 31, no. 1, Fall 2007, pp. 47–63.

Ward, Jonathan Stuart, and Adam Barker. "Undefined by Data: Survey of Big Data Definitions." *ArXiv*, 20 Sept. 2013, arxiv.org/abs/1309.5821.

Warner, Michael. *Publics and Counterpublics*. Zone Books, 2002.

Westergaard, Harald. *Contributions to the History of Statistics*. Agathon Press, 1968.

Wickman, Matthew. "Robert Burns and Big Data; or, Pests of Quantity and Visualization." *Modern Language Quarterly*, vol. 75, no. 1, Mar. 2014, pp. 1–28.

Wiedemann, Gregor. "Opening Up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences." *Historical Social Research / Historische Sozialforschung*, vol. 38, no. 4, 2013, pp. 332–57.

Youngman, Paul A., and Ted Carmichael. "Big Data, Pattern Recognition, and Literary Studies: N-Gramming the Railway in Nineteenth-Century German Fiction." *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, Camden House, 2014, pp. 285–99.

https://doi.org/10.1632/pmla.2018.133.5.1207 Published online by Cambridge University Press