What good is distant reading for our reading?

Introduction

My question for today's talk is, "What good is distant reading for our reading?" By "our reading" I simply mean the various kinds of reading that form the basis of professional literary study today.

I frame this talk around the idea of "our reading" because I think that, if it is to be successful, the digital humanities must eventually seem more like the New Historicism than a new discipline. Just as the New Historicism broadened the range of texts that literary critics studied, so too should computational literary study offer the discipline *new kinds of evidence* that support its goal of reading well.[1]

[OUTLINE SLIDE]

I'll begin by explaining two interdependent answers to today's question that emerge from my first book project. The first is that computational approaches *occasion attention*. The second is that they require us to think about the *representativeness* of our reading in a usefully different way.

Then, I will show how the research for my book supports those answers. Specifically, I'll discuss what we can learn from experimentally treating 1865 as a hinge point in the history of US prose fiction.

In his widely circulated *New Yorker* essay from a few weeks ago, Nathan Heller asks a student what the dominant vernacular of the contemporary university is. The student responds, "the language of statistics."[2] I will argue that prose fiction in Gale shows that this process began earlier, in different places, and in different ways than we usually imagine.

After sharing how I draw these conclusions using computational methods, I will show how these findings inform a reading of fiction by W.E.B. Du Bois, US literature's most serious statistical thinker.

Although these answers to today's question are a part of my book project, they also exceed it.

At the end of the talk, I'll briefly show how they inform my larger program of research by turning to a pilot study that is described in my article forthcoming in *PMLA*. That pilot study will

---

[1] Broadened by emphasizing the circulation and mutual influence of literary and non-literary texts. H. Aram Veeser, ed., *The New Historicism* (New York: Routledge, 1989), xi.
[2] Nathan Heller, "The End of the English Major," *The New Yorker*, February 27, 2023, https://www.newyorker.com/magazine/2023/03/06/the-end-of-the-english-major.

expand into a larger project over the coming years, hopefully with the support of a pending NEH grant. These answers will inform that project as it continues to grow.

First proposition

So, what good is distant reading for our reading? Counterintuitively, distant reading *defends* close reading.

Twenty-three years ago, Franco Moretti coined the term distant reading.[3] Moretti was not the first to argue for using quantitative methods to study literature, efforts that go back to the work of scholars like Josephine Miles, Janice Radway, and J.F. Burrows, among others.[4]

[*CONCORDANCE* SLIDE]

As my co-authors and I argued in *The Cambridge Companion to the Novel*, we can find evidence of scholars extracting tabular data from novels at least as early as 1907, with the publication of Mary Williams's *Dickens Concordance*, which collects data about Dickens's character systems.[5]

If his was not the first expression of this idea, Moretti's coinage of "distant reading" has nevertheless been influential. However, it has also created a certain mistrust of computational literary studies in some corners of the academy. That's because Moretti presented distant reading *in opposition* to reading.

[Moretti SLIDE]

"We know how to read texts, now let's learn how *not* to read them."[6] One fear this framing created is the idea that people like me would teach students of literature how to read spreadsheets *instead of* poems or write code *instead of* paragraphs.[7]

I *categorically* reject these ideas. Distant reading does not obviate our reading. On the contrary, computational findings *defend* our reading by *occasioning attention*.[8]

---

[3] Franco Moretti, "Conjectures on World Literature," *New Left Review*, no. 1 (February 1, 2000): 54–68.
[4] Brad Pasanek, "Extreme Reading: Josephine Miles and the Scale of the Pre-Digital Digital Humanities," *ELH* 86, no. 2 (June 4, 2019): 355–85, https://doi.org/10.1353/elh.2019.0018; John Frederick Burrows, *Computation Into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987); Janice A. Radway, *Reading the Romance: Women, Patriarchy, and Popular Literature* (Chapel Hill: University of North Carolina Press, 1984).
[5] Alex Woloch, *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel* (Princeton University Press, 2009).
[6] Moretti, "Conjectures on World Literature," 57.
[7] This stance contributed to an argument made in *The Los Angeles Review of Books* that the digital humanities are a Trojan horse for the so-called "neoliberalization" of literary studies. Daniel Allington, Sarah Brouillette, and David Golumbia, "Neoliberal Tools (and Archives): A Political History of Digital Humanities," Los Angeles Review of Books, May 1, 2016, https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/.
[8] Contra Moretti, this can also motivate our reading *beyond* the literary canon.

Alan Liu writes about the success of this approach in an analysis of a computational literary study by Ryan Heuser and Long Le-Khac.

[LIU SLIDE]

Liu describes how Heuser and Le-Khac move from analyzing data to analyzing texts, using "lines of interpretation generated by machine observation."[9]

For Liu, computational literary study provides new motivations for reading *these* texts in *this* way. It also provides a new justification for any generalizations made from those texts.

Liu's analysis suggests that "machine observation" creates a *demand for interpretation*, especially in cases where we observe under-explained patterns.[10]

If distant reading occasions attention and creates a demand for interpretation, then, as Liu suggests, that work *pays off* at the level of the text.

Second proposition

My second proposition builds on the first.

[slide with propositions]

Distant reading requires us to think about the representativeness of our reading in a *usefully different* way.

There is much that we do not know about what Margaret Cohen has called "the great unread," mostly because we haven't read it.[11] Computational approaches can help with this fact. But a corollary to this is also true: No one *needs* to read all 41 of Horatio Alger's novels for young people to know something about the rags-to-riches story.

My argument is not that statistical conceptions of representativeness ought to supplant longstanding critical practices. Rather, computational literary study makes *usefully different* claims of representativeness because they are *bounded* and *falsifiable*. Andrew Piper asks how literary scholars can know when we are wrong without occasionally making such claims.[12] By making our criteria of representativeness explicit, we can open a methodological conversation that has for too long been implicit.

---

[9] Alan Liu, "The Meaning of the Digital Humanities," *PMLA/Publications of the Modern Language Association of America* 128, no. 2 (March 2013): 409–23, https://doi.org/10.1632/pmla.2013.128.2.409.
[10] Creating a demand for interpretation is of no small value to the profession at the present time, something that John Guillory's new book touches on throughout, including with respect to DH and literary history. John Guillory, *Professing Criticism: Essays on the Organization of Literary Study* (Chicago: University of Chicago Press, 2022), 98.
[11] Margaret Cohen, *The Sentimental Education of the Novel* (Princeton, N.J: Princeton University Press, 1999), 23.
[12] Andrew Piper, *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*, Elements (Cambridge University Press, 2020), https://doi.org/10.1017/9781108922036.

[SLIDE robust is obvious]

A recent controversy reveals the importance of this claim to a different kind of representativeness. Nan Z. Da argues that "…the problem with computational literary analysis as it stands is that what is robust is obvious (in the empirical sense) and what is not obvious is not robust."[13] Da suggests that literary studies can therefore ignore computational approaches because what is robust is always already obvious to readers.

Da's argument echoes an occasional criticism of computational literary study, which says that some findings merely confirm preexisting theories. We need only think about what the sciences call the replication crisis to understand one reason why such confirmation might be valuable.

[SLIDE Underwood]

Ted Underwood offers a different response, arguing that, "the fact that something [e.g., a pattern in literary history] is *retrospectively plausible* doesn't mean we already knew it."[14] If we can explain new data with old theories, that strengthens those theories; it doesn't devalue the data.

Da concludes with an arithmetical case for why literary scholars should read rather than compute. In making her case, she concedes the value of computational literary studies' distinct claim to representativeness. However, her arithmetic inadvertently makes a case *for* computation:

[SLIDE one thousand people]

Da suggests that it would only take 1,000 people one year to read 15,000 novels, a number close to some of the larger corpora used in computational literary studies, such as the Gale corpus of US fiction, which we will be discussing shortly.

But this presupposes that one thousand professional readers could *organize* and *distribute* their reading in a way that would serve any specific research goal.

The question is not whether scholars could read any number of texts. The question is whether our efforts could be *coordinated* such that we could ever say "what is obvious" about thousands of different texts read by a thousand different people.

It's also important that this theory of representativeness—using evidence gathered from numerous examples to identify their shared characteristics—is the *inverse* of how students of

---

[13] Nan Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45, no. 3 (March 1, 2019): 601, https://doi.org/10.1086/702594.

[14] Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: The University of Chicago Press, 2019), 14.

literature often learn about major concepts like literary movements. Instead, as Heather Brink-Roby has suggested, literary studies' idea of representativeness often treats exemplary texts as "suitable bases of inference" to the categories of which they are taken to be representative.[15]

For example, you would be hard-pressed to find a theory of English-language modernism that did not make room for *Ulysses*. Yet conversely, few in the field would look askance if you suggested that features of *Ulysses* might be characteristic of modernism.

This logic of representativeness is tacitly justified not on *breadth* of observation but on *depth* of influence. The salience of *Ulysses* to the idea of modernism means that literary criticism's theory of representativeness often becomes one of *resemblance* between the exemplar and the example.[16] Computational approaches can do something usefully different.

Summary thus far

So far, I've suggested two ways that distant reading is good for our reading. By observing patterns in larger volumes of texts than scholars can gather evidence from, computational findings *occasion attention* and create a *demand for interpretation*. Second, computational approaches require us to think about the *representativeness* of both our evidence and our claims in a *usefully different way* than we ordinarily do.[17]

From theory to practice

Now, I will show how the research for my first book supports these propositions.

My book began by asking an old question of US literary history in a new way:

[slide why 1865]

Why 1865? It's the boundary line for most survey courses in the field, and between the halves of *The Norton Anthology of American Literature*. 1865 reliably recurs as a start or end date on graduate reading lists, at conferences, and in book titles. There are very good historical reasons: the Thirteenth Amendment passes, Lee surrenders to Grant, Booth shoots Lincoln, Reconstruction begins. But what does all this tell us about the literature?

Scholars from Granville Hicks to Amy Kaplan to Henry Louis Gates, Jr. have used 1865 to think about the ways in which the end of the Civil War upended social relations, and how that, in turn,

---

[15] Heather Brink-Roby, *Something Larger than Themselves: Representativeness and Modernity*, Forthcoming.
[16] Hanna Fenichel Pitkin, *The Concept of Representation* (Berkeley: University of California Press, 1967); Suzanne Dovi, "Hanna Pitkin, The Concept of Representation," in *The Oxford Handbook of Classics in Contemporary Political Theory*, ed. Jacob T. Levy (Oxford University Press, 2015), https://doi.org/10.1093/oxfordhb/9780198717133.013.24.
[17] Neither of these claims imply that computational literary study can only engage with highly frequent or centrally distributed features of large numbers of texts. These methods do not entail a turn toward the center or away from the periphery. Indeed, as I show in my article that discusses Hawthorne's distinctive use of the word "likewise," computational methods can show us how *unusual* a given pattern or text is within its broader context.

remade literature.[18] But others say that 1865 tells us relatively little. Christopher Hager and Cody Marrs have argued "Against 1865," pointing out that it misrepresents the careers of authors like Walt Whitman whose work crosses that line.[19]

I realized that it could be informative to take the opposite approach of Hager and Marrs. Where they begin from the premise that 1865 *must* be criticized because it so often goes without saying, I wanted to know what we would learn if, instead, we were to take 1865 a little *too* seriously.

If we experimentally treat that year as a boundary in US literary history, what does that reveal about the literature on either side of the line?

We already know quite a few literary historical arguments about what happens across this divide. Literacy rises. Book production and distribution costs decline. Education becomes compulsory in most states. Public libraries expand as robber barons like Andrew Carnegie spread the "gospel of wealth."[20] Professional authorship becomes increasingly possible for Americans not named James Fenimore Cooper. Works by Frederick Douglass, Frances E. W. Harper, and Charles Chesnutt suggest some of the innovations happening in a range of African American literary traditions. Building on the sensations of Susan Warner and Harriet Beecher Stowe in the 1850s, publishers increasingly focus on middle-class women as a key literary market.[21] Writing accessible to young people also grows in importance and profitability, as exemplified by *Little Women* (1868-9). The lyric increasingly becomes the dominant poetic form.[22] The literary status of prose fiction increases, as we see in Henry James's 1884 essay "The Art of Fiction." F.O. Matthiessen's *American Renaissance* and US romanticism are said to die with Hawthorne in 1864.[23] Then, at some point, realism, naturalism, and modernism emerge.

My question was which of these literary histories, if any, would be evident in data comparing large volumes of US antebellum and postbellum writing.

Corpus

[18] Granville Hicks, *The Great Tradition: An Interpretation of American Literature since the Civil War*, Rev. ed (New York: Biblo and Tannen, 1967); Amy Kaplan, *The Social Construction of American Realism* (Chicago: University of Chicago Press, 1988); Henry Louis Gates Jr., *The Signifying Monkey: A Theory of African-American Literary Criticism*, Twenty-Fifth Anniversary ed. (New York: Oxford University Press, 2014).
[19] Christopher Hager and Cody Marrs, "Against 1865: Reperiodizing the Nineteenth Century," *J19: The Journal of Nineteenth-Century Americanists* 1, no. 2 (September 30, 2013): 259–84, https://doi.org/10.1353/jnc.2013.0026; Cody Marrs and Christopher Hager, eds., *Timelines of American Literature* (Baltimore: Johns Hopkins University Press, 2019).
[20] Andrew Carnegie, "Wealth," in *The North American Review*, ed. Allen Thorndike Rice, vol. CXLVIII (New York: No. 3 East Fourteenth Street, 1889), 653–64.
[21] Richard H. Brodhead, "The American Literary Field, 1860-1890," in *The Cambridge History of American Literature: Volume 3: Prose Writing, 1860–1920*, ed. Sacvan Bercovitch, vol. 3, The Cambridge History of American Literature (Cambridge: Cambridge University Press, 2005), 22, https://doi.org/10.1017/CHOL9780521301077.
[22] Virginia Jackson, *Before Modernism: Inventing American Lyric* (Princeton: Princeton University Press, 2023).
[23] F. O. Matthiessen, *American Renaissance: Art and Expression in the Age of Emerson and Whitman* (New York: Oxford University Press, 1954).

In order to make these comparisons, I first needed a corpus of US texts with a large number of examples before and after 1865. Beyond mere size, the corpus would also need to make a specific claim to representativeness, as Katherine Bode insists by her axiom that "you can't model away bias."[24] Lauren Klein and Catherine D'Ignazio also argue that attending to the conditions of data's production is one of the key tenets of data feminism.[25]

The Stanford Literary Lab holds a full-text copy of the Gale *American Fiction* corpus, which I have chosen to use and modify for my research, and which I will refer to as "Gale" for short. Like any commercial corpus, Gale has an institutional history. It builds on prior authorities and represents a compromise between what is desirable, what is available, and, for Gale and the Cengage Group, what is profitable.

Its key advantages are its comprehensive digitization of its bibliographic sources; its coverage of place, time, and literary form suitable for the research question; its hand-made metadata; and its large scale.

[SLIDE with corpus stats]

It is hard to imagine more than a billion words, so I provide a few comparators here, of which the eighty or so IKEA bookcases seems to me the most imaginable.

The texts were digitized using optical character recognition, or OCR, which is standard in the field and has a variable error rate depending on both image quality and the age of a text.

Unlike the OCR, which is machine generated, Gale's metadata was created by editors using information from bibliographies. Other large corpora, like Project Gutenberg, have more accurate texts than Gale does, but they have atrocious metadata.

[bibliography citations slide]

The works included in Gale come from four bibliographies of US fiction: three by Lyle Wright covering the period from 1774 to 1900, and one by Geoffrey Smith covering 1901 to 1925.

[slide describing principles of selection]

Wright's bibliographies included and excluded works of fiction based on the criteria he identifies here [gesture to slide].[26] Smith states in his preface that it was his goal in making his bibliography to "remain true to Wright's" criteria.[27]

---

[24] Katherine Bode, "The Equivalence of 'Close' and 'Distant' Reading; or, toward a New Object for Data-Rich Literary History," *Modern Language Quarterly* 78, no. 1 (2017): 77–106; Katherine Bode, "Why You Can't Model Away Bias," *Modern Language Quarterly* 81, no. 1 (2020): 95–124.
[25] Catherine D'Ignazio and Lauren F. Klein, *Data Feminism*, Strong Ideas Series (Cambridge, Massachusetts: The MIT Press, 2020).
[26] Lyle Henry Wright, *American Fiction, 1774-1850: A Contribution toward a Bibliography*, 2d rev. ed, Huntington Library Publications (San Marino, Calif: Huntington Library, 1969), viii.
[27] Geoffrey D. Smith, *American Fiction, 1901-1925: A Bibliography* (Cambridge, U.K: Cambridge University Press, 1997), ix.

Wright and Smith's criteria usually produce results that literary historians would approve of, but not always. To take what may seem like the most egregious example today, Wright's choice to exclude "juveniles" leads to the inclusion of Louisa May Alcott's other novels and short story collections, but the exclusion of the *Little Women* series.[28]

Even more than such exclusions based on intended readership, however, my major concern in working with a set of texts identified by bibliographers of the mid to late twentieth century is that the corpus could be unusably racist and sexist. Although its editors created useful metadata, Gale only contains information identified in the bibliographies, which provide no data about author race or sex. The only information available about either of these comes from the 3.8% of the more than 8,500 unique authors whose name on the title page includes "Mrs," "Miss," or a description such as "written by a Lady." These markers are by no means universal among women authors, nor consistent over time, much like categories of race and gender themselves.[29]

To find out the extent to which Gale excluded authors who are women and/or people of color, I compared its authors to an eligible subset of the 464 authors who have ever been selected for any edition of *The Norton Anthology of American Literature*.

[SLIDE Gale diversity]

That subset contained all Norton authors who are women and/or people of color whose publishing lives fell within Gale's dates. But this initial subset included poets, dramatists, and writers of other forms whose exclusion from Wright and Smith's bibliographies is best explained by reason of literary form. For example, David Walker's *Appeal* belongs in the Norton but not in Gale, given its focus on prose fiction.

43 of those 115 candidates were eligible for inclusion because they published works of prose fiction during Gale's dates.

36 of 43 (or 84%) are included in the Gale corpus.

Each of these seven authors [gesture to slide] had one or more works that fit Wright and Smith's criteria but were excluded. However, some of these eligible works have been discovered or recuperated since these bibliographies were published, such as Mary Boykin Miller Chesnut's novels.

These seven represent a principled beginning of an addendum to the Gale corpus, a project that I would be happy to discuss further in the Q&A.

---

[28] Nigel Austin Lepianka, "'Yet of Books There Are A Plenty': Bibliography, Data, and the Construction of American Fiction" (Thesis, 2019), https://oaktrust.library.tamu.edu/handle/1969.1/186183.

[29] Sally Haslanger, "Gender and Race: (What) Are They? (What) Do We Want Them to Be?," *Noûs* 34, no. 1 (2000): 33–55.

Although these results from Gale are better than many would have predicted, it is still safe to assume that Gale overrepresents white male authors relative to the *demography* of the US in the period. If women authors are correspondingly underrepresented, they are far from absent in Gale.

[SLIDE top authors by word count]

Eight of the top twenty most prolific authors, measured in words, are women, including the top author (by a lot), EDEN Southworth, who the equivalent of seven or eight *Clarissa*s.

This fact broaches my second provocation about representativeness and its relation to our reading. It remains unclear how literary criticism would want to adjudicate such questions of overrepresentation. Ideally, would we want to sample authors in a corpus such that the balance of texts parallels the population at the approximate time of their circulation? Should we instead compare corpora against the historical demography of known authors? Of literacy rates? Of our own students today? Whatever answer we choose will be consequential, and it is a question on which I am aware of no consensus about what we would expect or how we would measure it.

Having glimpsed the scope of this corpus, we return to Da's earlier provocation. Gale's eighty IKEA bookshelves are *not* readable, at least not without a year and a team of one thousand. Absent that, *what* can we take data derived from this corpus to represent?

I would argue that Gale helps us think about *that which has been repeatedly recognized as US prose fiction*, a process mediated through more than two centuries of publication, preservation in research libraries, and digitization. Roopika Risam identifies this characteristic—repeated recognition—as *the* challenge posed by the digital cultural record: Without active efforts to push back against it, digitization today can reproduce yesterday's biases.[30]

Findings

So, what does experimentally treating 1865 as a firm boundary line in this corpus reveal? The first thing is the absolute growth of US fiction publication:

[Scatterplot slide]

This scatterplot shows the total number of works published per year in the corpus. As you can see, the trend is not continuously upward but involves periods of growth and contraction, as in the decline leading up to the beginning of the Civil War.[31]

---

[30] Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Evanston, Illinois: Northwestern University Press, 2019).

[31] Regardless of how you measure, there is a lot more fiction after 1865. We need to control for that variability in making comparisons between the corpora. There are a variety of ways to handle these discrepancies. For some applications, you might want to randomly sample an equivalent number of tokens from each group. Generally, however, I have taken data from the entire corpus and controlled for the variation in size. The argument for sampling would be that you would want to have an equivalent number of opportunities for observation for certain tests. I would counter this by arguing that some of our work in the digital humanities can be less predictive than descriptive. I'm less interested in the characteristics of works that would have been *likely* to exist than the characteristics of those that exist.

My hypothesis was that quantitatively distinctive differences between antebellum and postbellum US fiction would *not* pertain to the literary historical arguments I reviewed a moment ago. Instead, I expected the most significant differences to reflect changes in the ordinary usage of moderately frequent words. This would include banal words like *lunch*, which is increasingly used to refer to the midday meal during this period.[32]

I also hypothesized that industrialization would introduce a distinctive vocabulary in the postbellum period, with words like *telegraph, telephone, train, bicycle,* and *automobile*.[33] As Thoreau said of living through the beginnings of these changes, "we do not ride upon the railroad; it rides upon us."[34]

I will now describe several experiments testing these hypotheses, moving from simpler to more complex metrics. Following Andrew Goldstone and Johanna Drucker, I prefer whenever possible to use tables rather than data visualizations.[35]

The renowned linguist Adam Kilgarriff argued that what he called "simple math" provides one of the best ways to identify distinctive words across corpora.[36]

[Kilgarriff simple math explanation]

First, we need an important distinction: types versus tokens. A type refers to the abstract form of a word, whereas a token refers to the concrete instantiation of a word type. In the case of this table, we're counting the *tokens*—or instances—of the *type*—or word—*the*.

For Kilgarriff's method, we divide the rate of a type's use in a given corpus by the rate of its use in a reference corpus, plus a smoothing factor or frequency cutoff to deal with this method's tendency to overrate infrequent words. If the two corpora use the type at near-equivalent rates, the ratio would be close to 1. If more prevalent in the target corpus, which goes in the numerator, the value is greater than one.

---

[32] While we're on the topic of food, I also expected to see evidence of commodities and consumer goods that became more widely available and more widely advertised like coffee, which had been rationed during the Civil War. Mark Pendergrast, *Uncommon Grounds: The History of Coffee and How It Transformed Our World*, New edition (New York: Basic Books, 2019), 46.

[33] Alexander Manshel, "The Lag: Technology and Fiction in the Twentieth Century," *PMLA* 135, no. 1 (January 2020): 40–58, https://doi.org/10.1632/pmla.2020.135.1.40.

[34] Henry David Thoreau, *Walden*, 150th Anniversary ed. (Princeton University Press, 2016), 92.

[35] Johanna Drucker, *Visualization and Interpretation: Humanistic Approaches to Display* (Cambridge, Massachusetts: The MIT Press, 2020).

[36] Adam Kilgarriff, "Comparing Corpora," *International Journal of Corpus Linguistics* 6, no. 1 (January 1, 2001): 97–133, https://doi.org/10.1075/ijcl.6.1.05kil.

Comparing US fiction before and after 1865, we find that the corpora use just 15% of their types at nearly identical rates.[37] Types most distinctive of antebellum fiction include words rarely used in the later period, such as *privateer* and *brandywine*. This seems fine, if unsurprising.

However, some of the distinctive postbellum words began to catch my attention:

[postbellum Kilgarriff]

Where we already have explanations from the history of prose genres to account for the predominance of a word like *detective* on this list,

[postbellum Kilgarriff highlighted]

I was surprised to see terms associated with probabilistic thinking and people in the abstract appear among the most twenty most distinctive in a set of some 39,000 candidates.

Digging further down through the top 200 types, I found many of the words I had hypothesized I would see—*railway*, *telegraphed*, *revolver*, *receiver*.

[postbellum Kilgarriff top 200 selected 1]

Yet there were still more words seemingly about probability and representativeness: *normal*, *typical*, and *average* among them.

[postbellum Kilgarriff top 200 selected 2]

Two things struck me about those words: First, by this measure, they appear to be *more distinctive* than many of the words I had expected to see near the top. Second, unlike with *detective*, I had weak intuitions as to *why* they would be so distinctive of the postbellum period.

Though easy to use, Kilgarriff's method is hardly enough to confirm this pattern. Other techniques allow us to quantify the degree of our *surprise* at a result like this. Dunning's log-likelihood is frequently used for this purpose.

[Dunning explanation]

Dunning's log-likelihood builds on the same insights from Kilgarriff's simple math: By comparing the rate of a type's observed usage to the rate at which we would expect it to be used in the corpus, log-likelihood tells us how large and how statistically significant any discrepancies between expectation and reality are, with higher values indicating greater degrees of surprise and statistical significance.[38]

---

[37] Surprisingly, only about 15% of the words in common across the corpora are used at rates within 10% of each other, i.e., Kilgarriff ratios of 0.95 to 1.05. Put another way, 85% of words are used at rather different rates across the corpora.

[38] Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidince," in *Using Large Corpora*, ed. Susan Armstrong, 1st MIT Press ed (Cambridge, Mass: MIT Press, 1994), 61–74. Rayson and Garside further

[Dunning slide 1]

Here are the most distinctive postbellum words per log-likelihood. Log-likelihood is sensitive to the amount of available evidence, which is one reason why we see highly frequent words like *she* at the very top of the distribution.

[Dunning slide 2]

However, just a little further down this list, we see those same probabilistic words, plus new compelling information about related words that are as or more distinctive of the postbellum period: *guess*, *chance*, *queer*, and *type* among others.

To be clear, by drawing your attention to these words, I am *not* arguing that they are the only story in the data; they aren't and they couldn't be.

But, to return to my first proposition, their consistent and persistent co-presence among the most distinctive features of US prose after 1865 *occasioned* my attention to them.

At this point, we have a set of words that appear to have some semantic relationship. But any computational literary scholar quickly learns that even good readers usually do a bad job guessing which words are associated with which ideas.

Collocates—words that *co-locate* with one another in a corpus—can help us correct for our biases by revealing the contexts in which the words that interest us usually appear.

[collocation examples]

Linguist J. R. Firth famously argued in 1957 that "you shall know a word by the company it keeps."[39] Collocates help us move from word frequencies toward word meanings by measuring the company words keep. As you can see in these examples, collocation measures words based on their *proximity*.

[Collocation PMI]

We sort significant collocates by a measure of association called pointwise mutual information, which is abbreviated PMI. As with log-likelihood, PMI measures how much more often two words collocate than we would expect given their distributions.

New and *York* collocate a lot; *York* and *death* rarely if ever. *New* and *York* have a high PMI; *York* and *death* have a lower one.

---

described how to apply Dunning's log-likelihood to the comparison of corpora: Paul Rayson and Roger Garside, "Comparing Corpora Using Frequency Profiling," in *The Workshop on Comparing Corpora* (Hong Kong, China: Association for Computational Linguistics, 2000), 1–6, https://doi.org/10.3115/1117729.1117730.

[39] J. R. Firth, "A Synopsis of Linguistic Theory, 1930-1955," in *Studies in Linguistic Analysis* (Oxford: Blackwell, 1962), 11.

[Collocation results]

These are significant collocations for the word *average* split by decade. As you can see, *average* collocates with some numerical words like *height*. But it also collocates as or more strongly with words with which it describes the representative member of groups: the *average American, woman, man, citizen, reader,* etc.

[*normal* collocates]

We see similar results in the collocates of *normal*. Early on, we mostly see *normal condition* and *normal school*. But later, we also see the increasing association of normalcy with health, the self, the mind, gender, and humanity. These are all features that both disability studies and queer theory have drawn to our attention.[40]

[*typical* collocates]

Typical is used in a more limited domain than *average* or *normal* to characterize social classes and geographic differences rather than how individuals compare to central tendencies.

In the book, I discuss this period distinction between the average and the typical by analyzing a 1901 essay in *Everybody's Magazine* written by the Geographer of the US Census, Henry Gannett.

[Gannett slide]

Gannett's essay uses the 1900 census data to quantitatively describe "The Average American."[41] Gannett begins his essay by arguing that, properly understood, the average American is Black. This claim is so provocative that Gannett's editor prefaces his essay with a paragraph arguing that the average must not be mistaken for the typical, which he claims is bounded by geography and ideals rather than numbers. Of course, the editor's nervousness also reveals how these terms are used to *confirm* rather than *contradict* expectations.

With these collocates, we see evidence of one of the key features I study in the book: the use of statistical and probabilistic thinking to *authorize generalizations* about individuals and groups. The use of statistical language—even or perhaps especially in cases *without* supporting evidence—provides a veneer of authority behind which speakers mask subjective judgments as objective facts.

---

[40] Lennard J. Davis, *Enforcing Normalcy: Disability, Deafness, and the Body* (London ; New York: Verso, 1995); Michael Warner, *The Trouble with Normal: Sex, Politics, and the Ethics of Queer Life* (New York: Free Press, 1999).

[41] Henry Gannett, "The Average American," in *Everybody's Magazine*, vol. V (Ridgeway Company, 1901), 318–20.

The philosopher Ian Hacking suggests that this is part of a larger process over the long nineteenth century whereby the concept of "human nature" is supplanted by the idea of "normal humanity."[42]

We see evidence of Hacking's claim in the data, especially in the changing meaning of the average person. For its inventor Adolphe Quetelet in the 1830s, the average person was an *ideal*.[43] Quetelet regarded real people as erroneous approximations of the idealized average.

But by the latter half of the nineteenth century, the notion of the average person had moved from Quetelet's ideal to the modern conception of the statistically representative individual as a *mediocrity*.

Francis Galton criticized Quetelet's idealization of the average on both statistical and aesthetic grounds. Galton argued that Quetelet's view was analogous to someone who, upon seeing Switzerland's dramatic landscape, concluded that, "if its mountains could be thrown into its lakes, two nuisances would be got rid of at once."[44]

As his aesthetic preference for inequality might suggest, Galton's interpretation of normal humanity as mediocrity is inextricably linked to his support for eugenics.

At this point, we know that these words are among the most distinctive of postbellum US fiction across multiple metrics, and that they collocate with words that suggest they are used to characterize individuals and groups. But how widely dispersed within the corpus is this discourse?

I answer that question using a two-step process: First, I use the terms already identified to create what Heuser and Le-Khac call a semantic field.[45] I use those words to identify other words that significantly correlate with any of our key terms across the whole corpus.

[top correlates of "normal"]

As this chart shows, correlation tells us to what extent words *move together* in a linear relationship, even when they move at different rates, as we see in the distinction between *normal* and *worry*. Put another way, mutually correlated words help us think about what *else* we say when we say a certain word.

---

[42] Ian Hacking, *The Taming of Chance*, Ideas in Context (Cambridge: Cambridge University Press, 1990), 160, https://doi.org/10.1017/CBO9780511819766. See also Elizabeth B. Bearden, "Before Normal, There Was Natural: John Bulwer, Disability, and Natural Signing in Early Modern England and Beyond," *PMLA* 132, no. 1 (January 2017): 33–50, https://doi.org/10.1632/pmla.2017.132.1.33.

[43] Lambert Adolphe Jacques Quetelet, *A Treatise on Man and the Development of His Faculties*, ed. T. Smibert, trans. R. Knox, Cambridge Library Collection - Philosophy (Cambridge: Cambridge University Press, 2013), https://doi.org/10.1017/CBO9781139864909.

[44] Francis Galton, *Natural Inheritance* (London ; New York: Macmillan and Co, 1889), 62.

[45] Ryan Heuser and Long Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," *Pamphlets of the Stanford Literary Lab*, no. 4 (May 2012), https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

But not all correlates are semantically related. So, following Heuser and Le-Khac, I filtered the correlates based on their inclusion or exclusion in a common category of the *OED*'s *Historical Thesaurus of English*.[46]

[keywords correlated with probabilistic seeds]

In this case, terms associated with probabilistic thinking fall under the thesaurus category Relative Properties (01.16), which includes words that relate to relationship, kind, order, number, measurement, quantity, and wholeness.

We can then measure the relative frequency of this semantically unified and correlated group of words over time.

[boxplot showing proportion of frequencies per year in probability discourse]

This boxplot shows the number of correlated probabilistic words *in each text* in each five-year group. The midpoint of the box represents the rate at which this discourse is used in the median text of that period.

As the trend suggests, the use of probabilistic discourse nearly triples over this period.

Even though this research began with the artifice of treating 1865 as a rupture, the growth displayed here suggests that this process is more continuous than abrupt.[47]

One final way that we can evaluate the significance of statistical thinking computationally is by measuring how well this semantic field can *classify* antebellum from postbellum works.

This is a clear case to apply a widely used form of machine learning called logistic regression.

[logreg slide]

We can "train" a logistic regression model to classify texts based on any of their features. Through the training process, the model "learns" how to weigh the significance of the evidence it receives about each object that it tries to classify.

In this example, the model predicts the probability that a student will pass an exam given the number of hours they spend studying. In our case, given data about the distribution of specific words in a work, the model will predict how likely it is that work was written in the antebellum or postbellum period.

[comparison of model performance: random vs. probability correlates vs. top dunning]

---

[46] Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon (eds.). 2023. *The Historical Thesaurus of English* (2nd edn., version 5.0). University of Glasgow. https://ht.ac.uk/.

[47] This graph suggests that Rita Felski may have been right to argue that "Context Stinks!" Rita Felski, "Context Stinks!," *New Literary History* 42, no. 4 (2011): 573–91, https://doi.org/10.1353/nlh.2011.0045.

After you train a logistic regression model on some of the data, you *test* it on data it has never seen before to measure how well that model *generalizes*.

After generating 100 different models with 100 equal random samples of training and test data from Gale for three different lists of words, I find that the words associated with probabilistic thinking classify unseen texts *nearly as well as* an equivalent number of top-ranked results from the measure of distinctiveness we considered earlier, Dunning's log-likelihood.

This is true despite the fact that the Dunning words were more frequent, and our probabilistic keywords have been filtered to ensure their *semantic* coherence rather than optimizing for distinctiveness alone.

For comparison, both the probabilistic words and the Dunning's log-likelihood words perform more than three standard deviations better than an equivalent number of random words.

So, this discourse surrounding statistical and probabilistic thinking is a distinctive, widespread phenomenon in US prose fiction; it is often used to characterize individuals and their relations to social groups; and its prevalence can be used to discern whether any given work was written before or after 1865 with about 87% accuracy.

After explaining these findings, my book turns to several case studies to understand the role played by this semantic field in US literature of the period.

I make two major arguments about it: First, the rise of statistical and probabilistic thinking is a more significant feature of US *literary* history than is generally acknowledged.[48] As a corollary to this point, this rising discourses of the "average person" and "normal humanity" emerge in US prose fiction *earlier* and more *pervasively* than other accounts acknowledge.[49]

Catherine Gallagher's essay on "The Rise of Fictionality" may make the best-known case for the relationship between English language prose fiction and probability. Gallagher argues that the idea of fictionality emerges as texts cease to explicitly solicit readers' belief in their truthfulness, as in well-known examples like *Robinson Crusoe*, which claimed not to have been authored by Defoe but by Crusoe.[50]

---

[48] When English literary criticism has touched on these themes in the eighteenth and nineteenth centuries, it has most often done so in the British rather than the American context, as in work by Mary Poovey, Audrey Jaffe, and Tina Young Choi, among others. In the US context, Maurice Lee has given the most thorough attention to these topics. Mary Poovey, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society* (Chicago: University of Chicago Press, 1998); Audrey Jaffe, *The Affective Life of the Average Man: The Victorian Novel and the Stock-Market Graph*, Victorian Critical Interventions (Columbus: Ohio State University Press, 2010); Tina Young Choi, *Victorian Contingencies: Experiments in Literature, Science, and Play* (Stanford, California: Stanford University Press, 2022); Maurice S. Lee, *Uncertain Chances: Science, Skepticism, and Belief in Nineteenth-Century American Literature* (New York: Oxford University Press, 2012); Maurice S. Lee, *Overwhelmed: Literature, Aesthetics, and the Nineteenth-Century Information Revolution* (Princeton University Press, 2019).

[49] Anna Creadick's recent book is good example, arguing that we should read this discourse around the desire to be average as a post-World War II phenomenon. We know now that it was happening much earlier.

[50] Catherine Gallagher, "The Rise of Fictionality," in *The Novel*, 2006, 336–63.

I suggest that, in light of this data, we can nuance Gallagher's point: US prose fiction solicits readers' belief not only through the probability of its events, but also through the discourse of probability itself.

My second argument is that one of the major ways in which statistical rhetoric functions in works of literature is to *authorize generalizations* made about individuals and groups. As you would expect, this is often used to make racist, sexist, and nationalist assertions. Yet at the same time, the *increasing* need to make such assertions betrays their *increasing* contingency.

And, as we will see next, some authors used statistical thinking toward emancipatory ends.

Du Bois

W.E.B. Du Bois begins his long writing life in the last twenty-five years of the Gale corpus. Among all major American authors, Du Bois is perhaps the most insightful statistical thinker, not least because he studied statistical theories and methods, and put them into practice for his book *The Philadelphia Negro*, which describes the lives of Black residents of Philadelphia's Seventh Ward based on data gathered using surveys Du Bois designed.

While Du Bois engages with statistical thinking throughout his writing life, his relation to it changes over time. Early in his career, he expresses hope that statistics will disabuse whites of the idea of the uniformity of Black experience. Du Bois suggests that, instead, statistics may be the best way to reckon with the true diversity of a population. Moreover, he suggests that statistics could characterize social problems, and were therefore a key step to reform.

[illiteracy rate]

Here, for example, is one of the sixty data visualizations Du Bois prepared for the 1900 exhibition on Black progress since emancipation. This chart measures and projects improving Black literacy rates, represented as the decline of illiteracy.

We also see Du Bois use statistics for reform in his famous idea of the Talented Tenth.

[talented tenth slide]

Du Bois initially framed the Talented Tenth as a counter to "the blind worshippers of the Average," those who have substituted human nature for normal humanity.[51]

But later in life, I argue that Du Bois becomes more pessimistic about the persuasive power of statistical thinking because it is so often used cynically.

In a satirical story published in 1923, Du Bois imagines a conversation between a fictionalized version of himself and a white man "of that famous average" on the question of racial

[51] W. E. B. Du Bois, "The Talented Tenth," in *The Negro Problem: A Series of Articles by Representative American Negroes of Today*, ed. Booker T. Washington (James Pott & Co., 1903), 43.

superiority.[52] Du Bois's interlocutor takes the white supremacist position and argues that, if measured, whites would be found to have greater "native gifts" than Blacks.

[measure incommensurable slide]

Exasperated, Du Bois's persona replies, "Are we not seeking to measure incommensurable things; trying to lump things like sunlight and music and love?"[53] Here, Du Bois takes issue not only with the method but also the *attempt* at measurement, balking at the use of statistical thinking for questions it cannot seriously address.

However, the best expression of Du Bois's complex relationship to statistical thinking appears in his fiction. It is not contained in one novel, but rather in one *character* who he rewrites across two novels. In *The Quest of the Silver Fleece*, her name is Caroline Wynn. In *Dark Princess*, her name is Sara Andrews.

Caroline and Sara are almost the same: They are both young Black women of the Talented Tenth who can pass. They both work in politics in major US cities and exert a great degree of political power through their ability to use statistics to read and sway voters. They are both drawn to their novels' male protagonists—Bles and Matthew, respectively—because they believe they will help them advance socially and politically. They both prepare the protagonists to become politicians by checking their idealism. They both try to marry the men to secure their power. And they both appear in the middle of their respective novels as interludes before Bles and Matthew return to the more radical political commitments they had begun developing earlier.

What makes Caroline and Sara worthy of our attention is that the source of their extraordinary power—their ability to accurately predict how populations will think and act—is also the source of their undoing. They can read people, but they misread individuals.

In the end, this costs both of them everything that they had worked for. Caroline and Sara are sometimes described as the novels' villains or anti-heroes, but I argue that Du Bois paints a more sympathetic portrait of these two women, who can be read as cautionary tales about statistical thinking: What is true of a population may be false of any individual. Du Bois foreshadows this theme near the beginning of *The Quest of the Silver Fleece*.

[SLIDE drat statistics]

Miss Smith states the problem that Caroline and Sara face plainly: "'Drat statistics! … These are folks.'"[54]

Caroline and Sara's fates bring us back to the provocations that began this talk. Computational approaches *occasioned attention* to the theme of statistical thinking in Du Bois's fiction. Those

---

[52] William Edward Burghardt Du Bois, *Dusk of Dawn: An Essay Toward an Autobiography of a Race Concept*, The Oxford W. E. B. Du Bois (Oxford University Press, 2007).
[53] Du Bois, 74.
[54] W. E. B. Du Bois, *The Quest of the Silver Fleece*, The Oxford W. E. B. Du Bois (Oxford ; New York: Oxford University Press, 2007), 7.

findings told me something about where and what to look for, but their demand for interpretation compels a return to reading.

My second provocation argued that computational approaches require literary scholars to think about the representativeness of our reading in a usefully different way. This suggests that Caroline and Sara are not merely a compelling echo of each other within Du Bois's fiction, but also echo significant discursive changes in US fiction over the long nineteenth century.

Anthologies

[slide coda anthologies]

I want to end by briefly showing how both of these arguments about computational literary study, while central to my book, also apply to my larger program of research.

My co-author J.D. Porter and I created a relational database describing every work and every author selected for each of the ten editions of *The Norton Anthology of American Literature* published between 1979 and 2022.

I only have time to discuss the key finding here, but I would be very happy to tell you about many of the other discoveries during Q&A.

According to Leah Price, "the canon wars of the 1980s were fought over anthologies' tables of contents."[55] One of the results of the canon wars thus far has been widespread agreement that US literary scholars should study *more* authors than we used to.

The Norton Anthology of American Literature has substantially diversified its roster of authors with respect to race and ethnicity, and more modestly with respect to gender.

But we also show that editors have achieved this goal through a strategy of *growth*, which we argue is the most significant trend in the anthology over the past 50 years.

[selection slide]

The number of authors anthologized per edition has increased from 131 in the first edition to 288 in the most recent, far more than doubling. At the same time, the total number of anthologized works has decreased by about 2%.[56]

The editors have diversified the anthology, but they have done so by giving every author selected for it *less*: fewer works, fewer pages, and less attention than authors selected for earlier editions received. This strategy devalues anthologization itself.

---

[55] Leah Price, *The Anthology and the Rise of the Novel: From Richardson to George Eliot* (Cambridge: Cambridge University Press, 2000), 2.

[56] Page counts explain some but not all of this. The number of pages per edition has only increased by about 17%.

Being one of 288 authors today simply can't mean as much as being one of the 131 authors selected for the first edition. And it certainly doesn't mean as much as being one of the 103 authors from the first edition who we reveal have never been cut from any edition.

This growth strategy and its attendant devaluation of *all* anthologized authors is all the more troubling since it is the mechanism by which 89% of the women and/or people of color ever anthologized were added.

Having described the main way the anthology has changed, we can also use this data to predict how this strategy will play out.

[prediction slide]

Slightly more than half of the authors who have ever been selected get reselected for *every* revision following their initial selection. Put another way, the majority of authors in any edition of the anthology are as likely as not to *never* be cut going forward. This high indefinite retention rate is one of the main causes of growth.

At the average pace set by the first ten editions, women authors will not reach demographic parity with men until the 19th edition, which will be published around 2065 and will feature 445 authors.

Editors could *increase* the significance of their interventions if they were to shift from a model premised on growth to one premised on *redistribution*.

[slide: network]

This network shows our database of *The Norton Anthology of American Literature*. It is a force-directed bimodal network that visualizes the selection of individual works, which are the small black nodes, in individual editions of the anthology, which are the large orange nodes.

It reveals the relationship between the core and the periphery of the anthology's selection history: Just 9% of works ever selected have been reselected for every edition; everything else around them has changed.

The larger point is that we could not have known these aggregate effects without first aggregating.

In my book project, computational methods *occasioned attention* to a discursive needle in a haystack: the rising prevalence of statistical thinking in US prose fiction of the long nineteenth century. In the case of my work on anthologies, computational methods do the opposite: They draw our attention *away* from individual needles and *toward* [gesture to slide] this haystack.

So, what good is distant reading for our reading? Computational approaches occasion attention and create a demand for interpretation. They also require us to think about the representativeness of our reading in a usefully different way. These advantages cut in *both directions*: They help us

move from the unreadably large down to the readably small, and from individual choices to collective patterns that are invisible until they are aggregated. In both of these ways, computational literary study defends our reading.

Thank you.