WILEY

ARTICLE

The double bind of validation: distant reading and the digital humanities' "trough of disillusionment"

Adam Hammond 🆔

University of Toronto

Correspondence
Adam Hammond, University of Toronto.
Email: adam.hammond@utoronto.ca

Abstract

The digital humanities (DH) is currently in the phase of the "hype cycle" known as the "trough of disillusionment." Franco Moretti, perhaps the most prominent practitioner of the most prominent discipline of DH—"distant reading," the computational analysis of large quantities of literary texts—recently expressed his exasperation with the state of DH, reflecting "our work could have been better" and asking why, "considering the amount of energy, talent, and tools, going into [DH], that we have such difficulty producing great results." Surveying leading recent work in distant reading by Moretti, Matthew L. Jockers, Laura Mandell, Ryan Heuser, Long Le-Khac, and Joanna Swafford, this paper provides a twofold explanation to the field's failure to produce "great results." Both explanations relate to "validation," the process by which quantitative results are shown to be reliable and trustworthy. Many distant reading projects have produced disappointing results because they have been more interested in validating their tools—showing that their computational methods are able to confirm existing stereotypes—than in pursuing genuine discoveries. Many others, meanwhile, produce provocative results that cannot be meaningfully validated. Although the double bind of validation is real, I propose collaboration and "interdisciplinary adaptation" as promising solutions.

1 | INTRODUCTION

Digital humanities (DH), the most tech-obsessed area of the humanities, is also the most susceptible to the tech-world model of the "hype cycle." Developed by the consulting firm Gartner to describe the vacillating fortunes of new technologies, the hype cycle consists of four phases. First, following an initial "technology trigger," expectations for the new technology begin to grow, investment dollars begin to pour in, and the media takes enthusiastic notice. Next, following a "peak of inflated expectations," negative press begins to accumulate as lofty expectations aren't met and investors begin to panic, causing a descent into the "trough of disillusionment." Only once this phase is completed can the "slope of enlightenment" be climbed, a stage characterized by steady growth and modest expectations built on hard lessons learned from the previous phases. Finally, the dizzying highs and lows give way to the relatively unexciting plane of the "plateau of productivity," where the real work gets done.

Today, DH is sliding down into the "trough of disillusionment." Its "peak of inflated expectations" came in the early years of this decade, on the heels of William Pannapacker's (2009) famous declaration following the 2009 Modern Language Association convention that DH was the "next big thing" in the humanities, promising rapt public attention, limitless funding, and a steady stream of tenure-track lines to a discipline starved of all three. When these impossible expectations weren't met, the downward turn began. Some notable indicators have included Stephen Marche's (2012) "Literature Is Not Data: Against Digital Humanities" in the *LA Review of Books*, the "Dark Side of the Digital Humanities" panel at the 2013 Modern Language Association, Adam Kirsch's (2014) "Technology Is Taking Over English Departments: the False Promise of the Digital Humanities" in *The New Republic*, and most recently and perhaps most forcefully, the *LA Review of Books* piece "Neoliberal Tools (and Archives): A Political History of Digital Humanities" by Daniel Allington, Sarah Brouillette, and David Golumbia (2016). Effectively inverting the picture painted by Pannapacker in his early post, these writers portray DH as engaged in a hostile takeover of English departments that entails abandoning humanistic nuance in favor of the brute positivist force of big data and rejecting such core humanities traditions as theory, critique, and activism. In their view, DH is ethically and intellectually bankrupt, motivated by financial opportunism and serving only to further the neoliberalization of the increasingly dollar-driven university. Indeed, the very susceptibility of DH to the tech-world model of the hype cycle has been read as evidence of its complicity in market-driven logic. As Richard Grusin argued at the "Dark Side of the Digital Humanities" panel, "it is no coincidence that the digital humanities ... emerged as the 'next big thing' at the same moment that the neoliberalization and corporatization of higher education has intensified in the first decades of the 21st century." Rita Raley (2014), quoting these remarks, notes that the DH "gold rush" is "not unlike the start-up frenzy of the 1990s," arguing that "there is more than a hint of speculative mania within the exuberant investment in centers and initiatives" (p. 31).

Although it is only one of countless disciplines in DH, "distant reading" has functioned as the major example of the field's promise and perils. Coined by Franco Moretti (2000) to describe the large-scale computational analysis of massive quantities of literary texts, often in the order of thousands or millions of books in a single experiment, distant reading served as the public face of DH in the upward-trending years of the hype cycle. When the *New York Times* wanted to get a handle on all the excitement, for instance, it was to the work of Moretti and his onetime collaborator Matthew Jockers that they turned, in articles with headlines like "What is Distant Reading?" (Schulz, 2011) and "Literary History, Seen Through Big Data's Lens" (Lohr, 2013). Distant reading has been an equally conspicuous target for DH skeptics. Marche (2012) accused computational studies of relying on dubious definitions of undertheorized key terms like "classic" and "influence," concluding "The problem with 'distant reading' is, naturally, the distance involved." Kirsch cited Moretti's (2009) essay "Style, Inc." as evidence of distant reading's banality. Describing its findings as "not very exciting," he wrote, "It is striking that digital tools, no matter how powerful, are themselves incapable of generating significant new ideas about the subject matter of humanistic study" (Kirsch, 2014). The authors of "Neoliberal Tools," meanwhile, accused practitioners of distant reading of charlatanism, calling their work "a slapdash form of computational linguistics adorned with theoretical claims that would never pass muster within computational linguistics itself" (Allington et al., 2016).

The latter claims—that distant reading has produced no interesting findings and that the findings it does produce are not trustworthy in any case—are carrying the day as we plunge into the "trough of disillusionment." Even Moretti himself, a founder of the discipline, seems to have lost faith. In his appearance in the *LA Review of Books*'s recent interview series, "The Digital in the Humanities" (Dinsman, 2016), he interrupts a line of future-directed questions to turn the tables on his interviewer. He says,

> There wasn't a single question that asked, 'Has the digital humanities done anything?' Leave aside what it can do in the future; has it done anything? And I find this fascinating. Somehow digital humanities has managed to secure for itself this endless infancy, in which, it is always a future promise.

Answering his own questions, Moretti declares, "the results so far have been below expectations": "I think," he says, "and think others in the field would agree, that our work could have been better." The ride up the "slope of enlightenment" will only begin, he suggests, when we start asking seriously what we have done wrong and how we

can do better. Prophesying a decidedly less exciting future for DH than the one proclaimed around 2009, he concludes, "Perhaps one of the most important things DH should address in the coming phase is the nature of its own results—how to evaluate them—and if necessary, why is it, considering the amount of energy, talent, and tools, going into it, that we have such difficulty producing great results."[1]

Unexciting as the task may seem, this paper takes Moretti up on his challenge. Why has distant reading failed thus far to produce any "great results"? The answer I propose relates directly to Moretti's suggestion that we turn our attention to "the nature of [distant reading's] results"—and especially to the question of "how to evaluate them." What DH faces today, I propose, is the twofold challenge of *validation*. Associated most closely with the discipline of statistics, and seldom employed in the humanities today, the term *validation* refers to the process of determining whether numerical results obtained in quantitative analysis provide a reasonable description of the source data—in the case of distant reading, whether the results produced by quantitative literary analysis make sense given what we know about literature. Those who practice distant reading are engaged in validation in two key moments of the research process: first, when asking whether their tools are working correctly, and, second, when asking whether the conclusions they draw from these tools are sound. As I argue here, the future success of distant reading depends in large part on a more concerted and self-conscious engagement with both types of validation. The widespread disappointment with distant reading, I further propose, stems from what we might call "validation errors": from research that promises "great results" while delivering only validation of its tools or else produces startling results that cannot be validated in any meaningful way. In my concluding section, I draw on my personal experience collaborating with computational linguists to suggest a way out of distant reading's double bind of validation.

2 | VALIDATING TOOLS

Adam Kirsch's (2014) *New Republic* article "Technology Is Taking Over English Departments" is a representative document of the current phase of DH disenchantment. Kirsch hits all the usual notes, calling the rise of DH an effect of the neoliberalization of the university, presenting humanistic work as fundamentally inimical to computation, and calling on humanists to unite in opposition to the digital menace.[2] Kirsch's main objection, however, is the one that digital humanists working on distant reading hear the most often from their colleagues: "the digital analysis of literature tells us what we already know rather than leading us in new directions." His principal example is *Uncharted: Big Data as a Lens of Human Culture* by Erez Aiden and Jean-Baptiste Michel (2013), two of the creators of the Google Ngram Viewer. Introduced in 2010, the Ngram Viewer is perhaps the most accessible and user-friendly tool of distant reading currently available. It does something that is at once simple and profoundly powerful: feed it a phrase of up to five words, along with a date range, and it will visualize changes in relative frequency of that phrase over time, drawing on the massive corpus of the Google Books archive.[3] Kirsch objects to *Uncharted* on grounds of both style and substance. To the authors' hyperbolic claim that the "big data revolution" will "transform how we look at ourselves …, change the humanities, transform the social sciences, and renegotiate the relationship between the world of commerce and the ivory tower," Kirsch responds, "These breathless prophecies are just hype," classing them alongside other examples of DH research in which "language of scholarship" is usurped by "the spirit of salesmanship." Kirsch's disdain is motivated by what he regards as patently uninteresting findings. Most egregious for Kirsch is the authors' claim to have made a discovery in the history of artistic censorship in Nazi Germany. Aiden and Michel devote some 15 pages of their chapter "The Sound of Silence" to the fate of modernist artists in the years 1933–1945, showing that, during the period of the Nazi's campaign against what they called "degenerate art," names like Marc Chagall, Paul Klee, and Wassily Kandinsky nearly disappear from German-language books.[4] For Kirsch, this is an "example of data illustrating a truism rather than discovering a truth." "After all," he writes, "we wouldn't think to search for those names in that time period unless we knew what we were going to find, and why."

Kirsch is right about the Nazi censorship example, and about *Uncharted* more generally. Far from providing a way forward for humanities research, the book performs an effective travesty of hype-ridden, data-laden, cliché-

reinforcing digital scholarship. But Kirsch is wrong to use *Uncharted* as a basis for dismissing DH or distant reading. The first reason for this is that Aiden and Michel are not humanists: They are both mathematicians, and, understandably, they *think* like mathematicians.[5] Despite their claims of providing a "lens on human culture" and stated desire to found a new field of "culturomics," Aiden and Michel are less interested in probing culture than in pursuing the first type of validation—that is, showing that their tool works as expected.[6] As Kirsch notes, we learn nothing new from the Ngram Viewer graph of banned artists: It tells us what we already know. That, for Aiden and Michel, is precisely the point. They are particularly proud of this graph and chose to publish it in their book, because it functions exactly as it should, dipping where we expect it to dip and rising where we expect it to rise. Despite their "breathless prophecies," Aiden and Michel are content to stop once they have validated their tool. Kirsch, for his part, understandably believes that the tool is valuable only insofar as it challenges or unsettles received notions.

Yet the fact that mathematicians can't come up with interesting questions for their tool doesn't mean we should dismiss its use in the hands of humanists. In my own engagement with the Ngram Viewer, I've found it to be a powerful resource for humanistic inquiry, provided one recognizes its limitations[7] and asks the right sorts of questions. While I agree with Kirsch that the examples Aiden and Michel provide leave much to be desired, the fact that they succeeded in the first type of validation—showing their tool works—inspired me to use it to pursue my own ends. When the Ngram Viewer was first introduced in 2010, I was at work on *Modernism: Keywords* with my co-authors Melba Cuddy-Keane and Alexandra Peat (Cuddy-Keane, Hammond, & Peat, 2014). My first reaction on learning about the tool was one of mild dismay. In our entry for the word "Advertising," we had written that the word saw a dramatic increase in usage in the modernist period. We claimed the same for "propaganda" and "international" and argued that modernists rarely used "manifesto" or "avant-garde." What if the Ngram Viewer—validated through examples such as that of Nazi censorship—showed our claims to be false? In practice, we had nothing to fear: The tool showed precisely the trends we had predicted, much to my relief. But even if it had falsified our claims, it would have served a useful purpose. I was confident that the trends we had described were accurate, but the fact of my unease on learning about the Ngram Viewer suggests that I wasn't absolutely certain. Aiden and Michel were mostly concerned with validating the tool itself; once they had validated it for us—once we had learned we could trust it—we could then use it to test our arguments.

The subsequent research of my collaborator Melba Cuddy-Keane suggests that the Ngram Viewer can be used not only for testing hypotheses but also for posing entirely new questions. While investigating the periodization of the Victorian and modernist periods in British literature, Cuddy-Keane discovered a shape she calls a "war hump": a distinctive type of graph for words whose relative frequencies increase markedly during the two World Wars, producing a shape like the silhouette of a Bactrian camel.[8] Predictably, the word "war" exhibits the "war hump" trend; but so do the words "peace," "nation," "democracy," and, less expectedly still, "speed" and "new world." Unlike the charts of *Uncharted*, which confirm what we already know—and the "advertising" example, which confirmed what we thought we knew—Cuddy-Keane's "war humps" open up avenues for exploration. How does the pressure of war alter our conceptions of utopia, our representations of temporality, and our conception of our most cherished self-definitions? The Ngram Viewer can't answer these questions, but by giving us a genuinely new way of looking at the written record, it can raise them.

Humanists practicing distant reading are not exempt from "validation errors" of the kind committed by Aiden and Michel—of mistaking mere tool validation for "great results." The discussion surrounding Matthew Jockers's (2013) widely debated *Macroanalysis: Digital Methods & Literary History* demonstrates this point. Though Jockers is a humanist—a professor of Digital Humanities at the University of Nebraska and a specialist in Irish and Irish American literature—his investigations in *Macroanalysis* fall into a pattern similar to those in *Uncharted*. Jockers's introduction makes a number of bold claims, arguing that his approach ("macroanalysis," effectively a synonym of distant reading) represents a "generational shift away from traditional literary scholarship" (p. 32). Dismissing close reading of individual texts as "totally inappropriate as a method of studying literary history" (p. 7), he claims, "The macroanalysis I describe represents a new approach to the study of the literary record, an approach designed for probing the digital-textual world as it exists today, in digital form and in large quantities" (p. 32). Despite these bold assertions, Jockers seldom

uses his digital methods to ask probing new questions or to seek meaningful new insights into literary history. Instead, his book is largely concerned with the first type of validation—showing that his tools work as expected when applied to a literary corpus—and generally proceeds by demonstrating these tools' ability to reinforce received ideas and confirm stereotypes.

This is particularly clear in Jockers's chapter on "Theme," which investigates the usefulness of topic modeling to the study of literary history. Topic modeling is an application of an algorithm called latent Dirichlet allocation (LDA), which functions, as Jockers puts it, as "the mother of all collocation tools" (p. 123). It employs a mathematically sophisticated method to determine which words tend to co-occur in multiple contexts in a large corpus and groups these words into clusters called "topics."[9] Jockers uses topic modeling to sort a corpus of 3,346 19th-century novels into 500 clusters in order to argue that, in the context of literary analysis, these are better understood as *themes* than as "topics." After explaining his method, Jockers makes the case for the literary merits of LDA with a close reading of topic #271, which he labels "Native Americans" and illustrates through a word cloud visualization. Jockers's discussion of the "Native Americans" theme centers on the presence of the word "stream," which appears in small type in the upper-left region of the word cloud. Jockers praises the ability of the LDA algorithm to overcome the inherent ambiguity of a word like "stream": In this context, he argues, the usage of "stream" is clearly distinguished from usages like "jet stream" or "stream of immigration" and bears no relationship to "media streaming"; among a group of words about "Native Americans," he argues, it simply refers to "a body of flowing water" (pp. 125–127). LDA *works*, in other words: All of the words in the topic, including the most prominent words "indian," "indians," and "chief," are related to 19th-century representations of Native Americans. Yet Jockers goes no further in investigating the literary–historical significance of this theme. As a colleague of mine asked when I showed this image in a talk on leading work in computational literary analysis, "So what? So the nineteenth century novel was racist. We already knew that." Jockers could have chosen any of the 500 topics he produced in order to demonstrate the efficacy of the LDA algorithm in identifying literary themes, yet he deliberately chose the loaded theme #271, "Native Americans." His subsequent choice to offer no discussion of its relevance to the representation of race in the 19th-century American novel, but only to focus on technical questions of word-sense disambiguation, and to conclude that the word cloud offers "a useful and quantifiable representation of a very particular theme" (p. 127) stands as a demonstration of his priorities, which are on the side of showing that his tools work rather than using them to uncover new readings.

Another aspect of *Macroanalysis* that has generated significant debate is its treatment of gender. In the "Theme" chapter, Jockers uses the results of his 500-theme topic model of 19th-century novels to argue that authors of particular genders "have clear thematic 'preferences' or tendencies" (p. 136). Finding that the themes most associated with female authorship are those he labels "Affection and Happiness," "Female Fashion," and "Infants," he writes, "Some of these preferences correspond rather closely to our expectations and our stereotypes." Noting further that female authors show strong preferences for themes he labels "Happiness," "Passion," and "Grief and Sorrow," he continues, "Not surprisingly, women also appear to have the market cornered when it comes to even more specific expressions of strong emotion" (p. 136). All in all, he concludes, his analysis presents "a ringing confirmation of virtually all of our stereotypes about gender" (p. 152). Laura Mandell raises several objections to Jockers's work on gender (in press). First, although Jockers claims to be speaking about gender, he is in fact speaking about sex: His binary division of authors into male and female may correspond to the distinction of sexed bodies into the male and female of the species but does not speak to the nonbinary spectrum of culturally inflected gender identities. In failing to acknowledge the sex–gender distinction, Mandell argues, Jockers flagrantly ignores at least 40 years of scholarship, which she takes as a worrying sign that in carrying out his promised "generation shift" in literary criticism, he will simply ignore the aspects of humanities research he happens to find inconvenient. The rhetoric of objectivity that Jockers employs in describing his findings further serves to naturalize a binary gender distinction that is anything but. In the course of his various attempts to classify texts according to author gender, Jockers produces phrases such as "the accuracy of the classification can be known with absolute certainty" (p. 71) and "the effectiveness of the classification process can be measured with perfect accuracy" (p. 79). Later, he argues that gender is a "strong signal" that can be detected with 86% accuracy from thematic data alone (p. 153). Reflecting on Jockers's language, Mandell reiterates

Johanna Drucker's terminological preference for "capta" rather than "data," a lexical choice that emphasizes the sense that so-called data is not *given* but *taken*. Mandell quotes Drucker's remarks that "The basic categories of supposedly quantitative information, the fundamental parameters of chart production, are already interpreted expressions," and, in particular, "the assumption that gender is a binary category, stable across all cultural and national communities, is an assertion, an argument" Drucker (2014, p. 129). Supplementing this language of scientific objectivity with invocations of "our stereotypes" and phrases like "not surprisingly" serves to further reinforce the supposed givenness of binary gender. Gender is an immensely promising area of investigation for distant reading, Mandell concludes, but only "if we admit that the gender category is being constructed both by the measurer and the measured, who are also agents, not just our patients."

Mandell further argues that Jockers misses the most interesting questions raised by his quantitative work. If female authorship is different from male authorship—if sexed bodies produce texts that can be differentiated computationally—then surely the most interesting question is "Why?" Does the method produce equally reliable results in all cultures at all times? When are the sexes closest to one another stylistically, and when are they furthest apart? What unites the "outliers" of these investigations: Why do some writers elude the grasp of such methods? What social forces produce writers whose sex is easily determined by computational methods, and what social forces muddy such distinctions? Jockers poses such questions neither in his consideration of the "Native Americans" topic nor in his analysis of female themes. Focusing on whether his tools work or not—validation of the first type—he fails to consider what they might reveal. Yet this is not to say that *Macroanalysis* is a failure or that it represents a dead end. In his introduction, Jockers writes that one of his primary hopes for *Macroanalysis* is that it will prove "provocative in the sense of provoking more work, more exploration, and more experimentation" (p. 32). In this respect, it has been an unqualified success. By leaving the most interesting questions unanswered, he opened a door for critics like Mandell to push distant reading from an exercise in tool validation to one directed at broader questions of the relationship between social forces, gender, sex, and authorship—one directed toward what Moretti might call "great results."

3 | VALIDATING RESULTS

When digital humanists engaged in distant reading move from validating tools to producing "great results," another danger manifests itself. Once we have validated our tools and produced results that seem interesting, how do we validate our findings? How do we *know* when distant reading has produced a genuine, as opposed to apparent, discovery? This danger is perhaps most clearly demonstrated by the case of Ryan Heuser and Long Le-Khac's (2012) *Quantitative History of 2,958 Nineteenth-Century British Novels*, published as the fourth number in the Stanford Literary Lab's series of pamphlets. The authors' procedure is relatively straightforward from a statistical standpoint. Using a script they call Correlator, they are able to track words with similar frequency trends across their corpus of nearly 3,000 novels published between 1785 and 1900. (Imagine a version of the Google Ngram Viewer that not only produces a frequency graph for the word you enter but also tells you every other word that follows a similar trend.) Their first finding was that words with similar trends seemed to be related semantically. For instance, a group of words they labeled "abstract values" (including *sensible, envy, prejudice,* and *emotion*) demonstrated very similar downward-trending trajectories; these words collectively represented 1% of all words in the corpus in 1800–1810, but only 0.44% by 1900. Noting that this trend largely confirmed their existing notions of 19th-century literature—a gradual retreat from the abstract emotionality of the Romantic period—they then looked for a group of words exhibiting an exactly opposite trend, with usage increasing steadily through the 19th century. They expected this group to include words closely associated with the Victorian period: a vocabulary of empire, perhaps, or of moral prudishness. Yet they found something else entirely. The group of words with a trend opposite to the "abstract values," which they labeled "hard seed" words, was composed primarily of action verbs, body parts, colors, and numbers. Although this wasn't what they expected, they nonetheless were able to produce a theory that explained the opposing trends of the "abstract values" and "hard seed" groups in terms of a shift in narrative technique from "telling" to "showing": a

movement away from idealism and abstraction toward literary realism predicated on empirical description. The authors presented their findings as proof that computational approaches can yield new discoveries that human-scale reading cannot. "It did not and would not have occurred to us to look toward a group of 'hard seed' type words," they argued: "It took a computational method of finding language trends to discover this other group of words." They concluded that digital humanists must embrace unexpected findings like these if they are to make any significant contribution to literary scholarship:

> If we required all data to make sense—that is, fit our established concepts—quantitative methods would never produce new knowledge. If the digital humanities are to be more than simply an efficient tool for confirming what we already know, if they are to be a method for making new discoveries and exploring the unexplored, then we need to check this tendency to seek validation. (p. 49)

Yet while Heuser and Le-Khac should be commended for moving beyond the mere validation of methods and confirmation of stereotypes,[10] they were too quick to consider their own conclusions validated. In a paper published shortly after that of Heuser and Le-Khac, Ted Underwood and Jordan Sellars (2012) presented an alternate explanation of the shift from "abstract value" to "hard seed" words. Underwood and Sellars focus on trends in literary diction, tracking the relative employment of words with Latinate and Anglo-Saxon roots in a variety of genres. In the 19th century, they find, Anglo-Saxon words increased steadily in usage while Latinate words decreased. Noting that nearly all of Heuser and Le-Khac's "abstract value" words are Latinate and nearly all of their "hard seed" words are Anglo-Saxon, they propose that what Heuser and Le-Khac may be seeing is not a shift from "telling to showing," but simply the manifestation of a larger-scale trend in literary taste away from Latinate words. This hypothesis is bolstered by the trend that Underwood and Sellars perceive for the 18th century, which is opposite, with Latinate words increasing in usage relative to those with Anglo-Saxon roots. Unless Heuser and Le-Khac could argue that literature in the 18th century saw a reverse shift from "showing to telling"—which would indeed represent "new knowledge," for it is quite counterintuitive—their argument seems to fall apart. Yet Underwood and Sellars stop well short of calling "gotcha"; in fact, they describe Heuser and Le-Khac's article as a "work of groundbreaking importance." Their argument is not that Heuser and Le-Khac are wrong, but rather that, given our lack of robust methods for validating our results, we can't be entirely sure. Even when distant readings appear to produce surprising new insights, they suggest, we are not equipped to prove that they are genuine.[11]

The debate in distant reading in which questions of validation have been discussed most vigorously is that surrounding Matthew Jockers's Syuzhet, a tool for exploring plot trajectories through sentiment analysis. Jockers (2015a) announced the release of Syuzhet, a software package for the R statistical programming language, in a blog post published in early February 2015. Using a variety of off-the-shelf sentiment analysis tools—tools for classifying how "positive" or "negative" a given word is, assigning a score of −1 to a negative word like "hate," 0 to a neutral word like "potato," or 1 to a positive word like "love"—Jockers would calculate the sentiment polarity for every sentence of some 40,000 novels, produce a graph of the sentiment values plotted along "narrative time" for each novel, and then smooth these graphs using a technique called a Fourier transform, employed to simplify the shape of the sentiment curve and to render novels of different lengths more readily comparable. Drawing on a lecture by Kurt Vonnegut that reduces all literary plots to a few basic shapes plotted from beginning to end across a y-axis of "fortune,"[12] Jockers speculates that his sentiment-based method can capture basic plot shapes in the manner that Vonnegut suggests. Jockers concludes this post with the cliffhanger suggestion, "In a follow up post, I'll discuss how I measured and compared 40,000+ plot shapes and then clustered the resulting data in order to reveal six common, perhaps archetypal, plot shapes."

This initial post attracted widespread attention from digital humanists and from the mainstream press.[13] In short order, Jockers (2015b) posted his promised follow-up, and another flurry of attention followed, most notably from the *Times* of London, which printed an infographic laying out Jockers's "six or seven" archetypal plot shapes under the dismayed headline, "The only ways to write a novel?" (Whipple, 2015). The most serious response, however, came from digital humanist Joanna Swafford. In a blog post published in early March, Swafford (2015a) provided three compelling

reasons that Jockers's graceful plot graphs were not to be trusted: First, the algorithm did a poor job of sorting the novels into sentences, Jockers's basic unit of analysis, thus muddying results; second, its sentiment analysis methods produced poor results in the testing she conducted (for instance, it found the sentence "I am *happy*" to be less positive than the sentence "*Well*, it's *like* a potato"); and third, the smoothing performed by the Fourier transform produced unreliable and misleading representations of the sentiment data. The latter point became the locus of contention in a series of further blog posts in which Jockers (2015c, 2015d, 2015e, 2015f) defended the Fourier transform and Swafford (2015b, 2015c) elaborated her position. Her main argument, demonstrated with increasing ingenuity and persuasiveness, was that the Fourier transform—which attempts to turn messy data into a repeating, periodic shape like a sound wave—distorts plot trajectories, which unlike sound waves are neither necessarily repeating nor periodic but often jaggedly abrupt. Eventually, Jockers conceded the point; in a blog post entitled "Requiem for a low pass filter" (Jockers, 2015g), he acknowledged that the Fourier transform made assumptions inappropriate to the representation of literary plots.[14] The fact that this acknowledgment came only after two male digital humanists publicly took Swafford's side (Enderle, 2015; Schmidt, 2015) was not lost on those taken aback by his treatment of gender in *Macroanalysis* (cf. Clancy, 2015).

Before Jockers had made his concession, Andrew Piper joined the conversation in order to steer it toward questions of validation. In a post entitled "Validation and Subjective Computing" published on the website of McGill University's .txtLab, Piper (2015b) argued that the major question raised by the Jockers–Swafford debate was how to "validate something that is inherently subjective" such as whether a graph of a plot shape was "correct." Holding that "validation should serve as a form of discovery, not confirmation of previously held beliefs," Piper argues that standard models of validation from computer science aren't appropriate to humanities materials that are "inherently unstable and dissensual."[15] Yet the usual humanities methods ("We don't validate a procedure; we just read until we think we have enough evidence to convince someone of something") won't serve to settle the debate, either. Despite the pressing need for "some sort of process to arrive at interpretive consensus about the validity of our analysis," Piper finds nothing available to him: "right now," he says, "we don't have enough evidence to validate or—and this is a key point—invalidate Jockers' findings."

In her response to Piper, Swafford (2015c) helpfully distinguishes between the two types of validation discussed so far in this paper: The first type, she argues, involves "making sure the tool works as expected" while the second type—"possible only once the first type of validation works"—deals with "analyzing the results of the tool to come up with new theories about history, literary history, or plot." Her contention, bolstered by Jockers's subsequent concession, is that she has contributed to the first type of validation, showing definitively that the tool does not work as expected. Only once the tool is validated can we turn our attention to the question of how to validate the theories generated from its outputs. Yet while Swafford is correct on this point, Piper's larger argument stands. Suppose Jockers's tool *were* valid from a technical perspective: its sentence segmentation faultless, its sentiment analysis impeccable, and its graph smoothing irreproachable. Even then, would we be in a position to say whether Jockers's findings about "six or seven" plot shapes were correct? If we asked 10 human readers to produce a "happiness versus sadness" graph for a given novel, we might reasonably expect to get 10 different graphs. How could we say which of these graphs were most valid? How would we evaluate the performance of *Sjuzhet* relative to that of human readers?

4 | COLLABORATION, OR INTERDISCIPLINARY ADAPTATION

Neither Swafford nor Piper is able to answer these questions definitively, although each points helpfully toward possible means of escaping from the double bind of validation. Swafford, like so many digital humanists before her, argues that collaboration represents the best way forward. Humanists need methods of validating their results, she argues, but are not presently capable of developing them on their own. Only by collaborating throughout the research process with experts outside the humanities—computer scientists, computational linguists, engineers, mathematicians, software developers, and others better trained than humanists in statistical and quantitative methodologies—will we be

able to arrive at a set of procedures that are robust and sensitive to the peculiar challenges presented by literary texts (Swafford, 2015c). Piper, for his part, argues that we must turn our attention from archiving literary *texts* toward archiving literary *interpretations*. "We have the novel data, but not the reader data," he writes; "Right now DH is all texts, but not enough perspectives." If we could build a "public platform" for "subjective validation exercises," he continues, it would shift the burden of validation from the "individual genius" to "new principles of readerly consensus" (Piper, 2015b). Jockers (2015h, 2016) is following precisely this path in his latest research, working to compare the sentiment trajectories produced by Syuzhet with those produced by human readers hired to manually evaluate the sentiment polarity of every sentence in a selection of novels.

My own research suggests the value of both approaches. Since 2011, I have been collaborating with computational linguists Julian Brooke and Graeme Hirst on modeling the phenomenon of dialogism in modernist literature. Rather than following a top-down procedure—developing an algorithm for detecting dialogism based on our intuitions or prevailing critical elaborations of the notion, and then deploying it on a large corpus—we began by tracking the subjective responses of individual readers. We started our investigation by looking at free indirect discourse (FID), an important formal vehicle of dialogism in modernist literature, through which authors deliberately blur the boundaries between the words of narrators and characters. Our initial experiment was simply to test whether readers (some 480 undergraduate students) were able to consistently identify spans of FID in Woolf's *To the Lighthouse* and Joyce's "The Dead" (Brooke, Hammond, & Hirst, 2016). Validating these results proved challenging, given the expectation in computational linguistics that readerly consensus must reach levels where disagreements can be reasonably dismissed as noise—an expectation that was unreasonable in the context of our experiment (Hammond, Brooke, & Hirst, 2013). Faced with the particular challenges posed by humanities materials, then, the existing validation techniques of computer science proved inadequate. Yet the solution was not to abandon collaboration with computer scientists but rather to work closely with them to develop new validation techniques that were both robust and appropriate to the materials—techniques that would have been impossible to develop without pooling their expertise in quantitative methodologies and my domain expertise in literary studies. Our initial aim had been to use the reading data to train a machine learning model that could detect passages of FID in any literary text. Again, the computational linguists used their disciplinary expertise to determine that this was not practicable; Woolf's and Joyce's employments of the device were simply too peculiar to be generalizable (Hammond, Brooke, & Hirst, 2016). As a way forward, I came up with the idea of producing digital editions of *To the Lighthouse* and "The Dead" that would visualize students' vastly differing interpretations of FID in the words—"readers' maps" to show how great an interpretive challenge dialogic modernist works present to their readers.[16] The computational linguists, for their part, found a new use for the reader data, using it to validate a new method for expressing linguistic style in quantitative terms (Brooke, Hammond, & Hirst, in press). Several years on, it is this quantitative approach to linguistic style that we are now using to track dialogism at the scale of distant reading (Hammond & Brooke, 2016). The success of our research program has depended on a continuous process of "interdisciplinary adaptation" by which the disciplinary assumptions of computer science and English literature have been tested, bent, and rethought. This process of adaptation would not have been possible without active, long-term interdisciplinary collaboration.

Following the dizzy heights experienced around 2009, the fun is mostly over for distant reading. As digital humanists attempt to turn the momentum of the precipitous fall into the "trough of disillusionment" into the steady upward progress of the "slope of enlightenment," the real work is only now beginning. This work cannot be done by humanists alone. From its inception, the field known first as humanities computing and then as DH has been defined as interdisciplinary. In an article introducing the term to a general literary audience, Matthew Kirschenbaum (2010) cited approvingly the Wikipedia definition of DH as "a field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities." As Susan Hockey (2004) explains, much of the appeal of distant reading in particular has derived from the promise of applying "the rigor and systematic unambiguous procedural methodologies characteristic of the sciences to address problems within the humanities that [have] hitherto been most often treated in a serendipitous fashion." Taking advantage of this promise will require not only *importing* these methods from the sciences but also working with scientists to *adapt* them to the specific demands of literary

research. It will require humanists and scientists to work together, in real time, at every step of the research process, to apply their expertise toward the validation of results. As it stands, few if any literature experts possess the quantitative or statistical skills to do this work alone, and few if any computer scientists or computational linguists possess the literary knowledge to work independently. Why has distant reading not produced any "great results"? Because too many of its books and articles have single authors, either from the humanities or the sciences, because literature experts spend too little time at computational linguistics conferences, and computational linguists too little time at literature conferences. Collaboration and "interdisciplinary adaptation" are the most promising ways out of the double bind of validation. They will be the defining characteristics of the next phase of distant reading, which may not be as much fun, but which—finally—promises results.

NOTES

[1] A persistent line of argument in literary criticism rejects the enterprise of seeking "great results" as a capitulation to the positivistic values of hard science (see Hammond, 2016, pp. 85–89). In distant reading itself, there is a long tradition of seeking not definite proof but using digital tools to "deform" texts creatively or to identify gaps and silences. Yet even critics working in the latter tradition recognize the pressing need to develop methods of validating computational results in the humanities. Lauren Klein (2013), who uses distant reading techniques to identify archival silences in "The Image of Absence," argues that although "certain key practitioners of digital humanities continue to frame their work as exploratory rather than discursive … the time has now come for digital humanities practitioners to more forcefully theorize the knowledge claims they make" (p. 668).

[2] Kirsch writes, "digital humanities has less to do with ways of thinking than with problems of university administration"; "It makes no sense to accelerate the work of thinking by delegating it to a computer when it is precisely the experience of thought that constitutes the substance of a humanistic education"; "The best thing that the humanities could do at this moment … is not to embrace the momentum of the digital, the tech tsunami, but to resist it and to critique it."

[3] For a more detailed introduction to the Ngram Viewer and Google Books, see Hammond 2016 (pp. 106–113).

[4] The Nazi censorship example also appears in the *Science* paper that introduced the Ngram Viewer (Michel et al., 2011).

[5] None of the 14 entities (13 individuals and one grouping, "The Google Books Team") credited as authors of the original *Science* paper were humanists. The disciplinary affiliations listed for the involved scholars involved are psychology, biology, computer science, medicine, mathematics, and engineering.

[6] Further evidence of the authors' lack of humanities training comes in their opening example, which takes the form of a purported data-driven debunking of the notion that Americans stopped treating "the United States" as a plural noun after the Civil War (pp. 1–5). Citing historian James McPherson's argument that "the war marked a transition of the United States to a singular noun"—and presenting as evidence an 1887 *Washington Post* article noting that "a few years ago … men said 'the United States *are*' " (2; emphasis added)—they then produce an Ngram Viewer chart for "the United States is" versus "the United States are" showing that the singular form eclipsed the plural in 1880, some 15 years after the war. "McPherson, though brilliant, was wrong about the singular form," they conclude; "The [*Washington Post*] eye witness didn't recall events accurately" (p. 5). Setting aside the fact that 1880 was "a few years" before 1887, Aiden and Michel also miss a glaring problem in their approach, which relies exclusively on books, which is to say, *written words.* The author of the *Washington Post* article wasn't referring to what men *wrote* but to what men *said*; McPherson doesn't specify but is no doubt sufficiently conversant with the discipline of history to know that shifts in spoken language can take much longer to appear in written discourse. The authors' claim that the Ngram chart is "true" (p. 5) requires much greater scrutiny, from a humanistic perspective, before it can be validated. For a more methodologically rigorous "distant reading" analysis of the is–are distinction, see Santin et al. (2016).

[7] For a discussion of the main drawbacks of the Ngram Viewer, which relate to metadata and optical character reader errors in the Google Books corpus, see Hammond (2016, pp. 106–113).

[8] Cuddy-Keane's work is illustrative not only of the Ngram Viewer's potential for raising significant humanistic questions but also of the tool's accessibility. Her article (Cuddy-Keane, in press) pushes back against the account of Ted Underwood (2013) in *Why Literary Periods Mattered* that "the organization of English studies into discrete and distinct literary eras arose from a perceived need, around the middle of the 20th century, to claim disciplinary autonomy by advancing a paradigm of rupture and contrast in marked contrast to the 'rhetoric of transition and [gradualist] development' prevailing in social history." The article thus presents a methodological chiasmus: a humanist relatively unfamiliar with computational approaches (Cuddy-Keane) using a user-friendly tool of distant reading (the Ngram Viewer) to challenge the relatively traditional literary–historical work (*Why Literary Period Mattered*) of a digital humanist best known for his innovative work in computational text analysis (Ted Underwood). It is a testament to the usability of the Ngram Viewer that such chiasmic work is possible.

[9] For a more detailed introduction to topic modeling, see Hammond 2016, pp. 116–119.

[10] It is worth noting that even these "surprising" results serve to confirm a well-established critical commonplace: the notion that the movement from the Romantic to Victorian period was accompanied by a rise in realism and a decline in abstract judgment. It would be difficult to argue that this finding represents "new knowledge." It would be more accurate, perhaps, to say that the authors expected their work to confirm a particular literary–historical commonplace and were surprised to find that it confirmed another.

[11] For an insightful reading of Heuser and Le-Khac's work, see Liu (2013).

[12] The fact that Vonnegut's lecture is clearly facetious—a send-up of quantitative thinking about plot shapes, rather than a "ringing endorsement" of the approach (just listen for the audience laughter)—has not been taken sufficiently into account, either by Jockers or those who have engaged with him.

[13] For links to the relevant articles, see Clancy 2015.

[14] In his most recent work, Jockers (2017) again argues for the usefulness of the Fourier transform.

[15] For more on Piper's notions of validation as a process of discovery, see Piper (2015a).

[16] These "readers' maps" are available on the websites of *He Do the Police in Different Voices* (http://hedothepolice.org) and *The Brown Stocking* (http://brownstocking.org/).

WORKS CITED

Aiden, E., & Michel, J. B. (2013). *Uncharted: Big data as a lens on human culture*. New York, NY: Riverhead.

Allington, D., Brouillette, S., & Golumbia, D. (2016, May). Neoliberal tools (and archives): A political history of digital humanities. *LA Review of Books*, Retrieved from https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/

Brooke, J., Hammond, A., & Hirst, G. (2016). Using models of lexical style to quantify free indirect discourse in modernist diction. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqv072

Clancy, E. (2015). A fabula of Syuzhet: A contretemps of digital humanities and sentiment analysis [Blog post]. Retrieved from https://storify.com/clancynewyork/contretemps-a-syuzhet

Cuddy-Keane, M. (in press). Crossing the Victorian/modernist divide: From multiple histories to flexible futures. In A. F. Gillard-Estrada, & A. Levita (Eds.), *Beyond the Victorian/modernist divide: Remapping the turn-of-the-century break in literature, culture and the visual arts*. New York, NY: Routledge.

Cuddy-Keane, M., Hammond, A., & Peat, A. (2014). *Modernism: Keywords*. Chichester, UK: Wiley-Blackwell.

Dinsman, M. (2016, March). The digital in the humanities: An interview with Franco Moretti. *LA Review of Books*, https://lareviewofbooks.org/article/the-digital-in-the-humanities-an-interview-with-franco-moretti/.

Drucker, J. (2014). *Graphesis: Visual forms of knowledge production*. Cambridge, MA: Harvard University Press.

Enderle, S. (2015). What's a sine wave of sentiment? [Blog post]. Retrieved from http://www.lagado.name/blog/sine-of-the-times/

Hammond, A. (2016). *Literature in the digital age: An introduction*. Cambridge, MA: Cambridge University Press.

Hammond A. & Brooke J. (2016). Project dialogism: Toward a computational history of vocal diversity in English-language literature. Paper presented at Digital Humanities 2016, Krakow, Poland, July 2016.

Hammond, A., Brooke J., & Hirst G. (2013). A tale of two cultures: Bringing literary analysis and computational linguistics together. Paper presented at the NAACL 13 Workshop on Computational Linguistics for Literature, Atlanta, GA, June 2013.

Hammond, A., Brooke, J., & Hirst, G. (2016). Modeling modernist dialogism: Close reading with big data. In S. Ross, & J. O'Sullivan (Eds.), *Reading modernism with machines: Digital humanities and modernist literature*. London, UK: Palgrave Macmillan.

Heuser, R., & Le-Khac, L. (2012). *A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method*. Stanford, CA: Stanford Literary Lab.

Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities*. Oxford, UK, Blackwell.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Champaign, IL: University of Illinois Press.

Jockers, M. L. (2015a). Revealing sentiment and plot arcs with the Syuzhet package [Blog post]. Retrieved from http://www.matthewjockers.net/2015/02/02/syuzhet/

Jockers, M. L. (2015b). The rest of the story [Blog post]. Retrieved from http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/

Jockers, M. L. (2015c). Some thoughts on Annie's thoughts ... about Syuzhet [Blog post]. Retrieved from http://www.matthewjockers.net/2015/03/04/some-thoughts-on-annies-thoughts-about-syuzhet/

Jockers, M. L. (2015d). Is that your Syuzhet ringing? [Blog post]. Retrieved from http://www.matthewjockers.net/2015/03/09/is-that-your-syuzhet-ringing/

Jockers, M. L. (2015e). A ringing endorsement of smoothing. [Blog post]. Retrieved from http://www.matthewjockers.net/2015/03/24/ringing_endorsement/

Jockers, M. L. (2015f). My sentiments (exactly?) [Blog post]. Retrieved from http://www.matthewjockers.net/2015/04/01/my-sentiments-exactly/

Jockers, M. L. (2015g). Requiem for a low pass filter [Blog post]. Retrieved from http://www.matthewjockers.net/2015/04/06/epilogue/

Jockers, M. L. (2015h). That sentimental feeling [Blog post]. Retrieved from http://www.matthewjockers.net/2015/12/20/that-sentimental-feeling/

Jockers, M. L. (2016). More Syuzhet validation [Blog post]. Retrieved from http://www.matthewjockers.net/2016/08/11/more-syuzhet-validation/

Jockers, M. L. (2017). Resurrecting a low pass filter (well, kind of) [Blog post]. Retrieved from http://www.matthewjockers.net/2017/01/12/resurrecting/

Kirsch, A. (2014). Technology is taking over English departments: The false promise of the digital humanities. *New Republic*, Retrieved from https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch

Kirschenbaum, M. G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin*, *150*, 55–61.

Klein, L. F. (2013). The image of absence: Archival silence, data visualization, and James Hemings. *American Literature*, *85*(4), 661–688.

Liu, A. (2013). The meaning of the digital humanities. *PMLA*, *128*(2), 409–423.

Lohr, S. (2013, January 26). Literary history, seen through big data's lens. *The New York Times*, Retrieved from http://www.nytimes.com/2013/01/27/technology/literary-history-seen-through-big-datas-lens.html

Mandell, L. (in press). Gender and cultural analytics: Finding or making stereotypes? In M. K. Gold, & L. Klein (Eds.), *Debates in the digital humanities 2017*. Minneapolis, MN: University of Minnesota Press.

Marche, S. (2012, October). Literature is not data: Against digital humanities. *LA Review of Books*, Retrieved from https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Moretti, F. (2009). Style, Inc. reflections on seven thousand titles (British novels, 1740–1850). *Critical Inquiry*, *36*(1), 134–158.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*, 54–68.

Pannapacker, W. (2009). The MLA and the digital humanities [Blog post]. Retrieved from http://chronicle.com/blogPost/The-MLAthe-Digital/19468/

Piper, A. (2015a). Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Literary History*, *46*(1), 63–98.

Piper, A. (2015b). Validation and subjective computing [Blog post]. Retrieved from http://txtlab.org/?p=470

Raley, R. (2014). Digital humanities for the next five minutes. *Differences*, *25*(1), 26–45.

Santin, B., Murphy, D., & Wilkens, M. (2016). Is or are: The "United States" in nineteenth-century print culture. *American Quarterly*, *68*(1), 101–124.

Schmidt, B. M. (2015). Commodius vici of recirculation: The real problem with Syuzhet [Blog post]. Retrieved from http://benschmidt.org/2015/04/03/commodius-vici-of-recirculation-the-real-problem-with-syuzhet/

Schulz, K. (2011, June 24). What is distant reading? *The New York Times*, Retrieved from http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?_r=0

Swafford, J. (2015a). Problems with the Syuzhet package [Blog post]. Retrieved from https://annieswafford.wordpress.com/2015/03/02/syuzhet/

Swafford, J. (2015b). Continuing the Syuzhet discussion [Blog post]. Retrieved from https://annieswafford.wordpress.com/2015/03/07/continuingsyuzhet/

Swafford, J. (2015c). Why Syuzhet doesn't work and how we know [Blog post]. Retrieved from https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-doesnt-work-and-how-we-know/

Underwood, T. (2013). *Why literary periods mattered*. Stanford, UK: Stanford University Press.

Underwood, T., & Sellars, J. (2012). The emergence of literary diction. *Journal of Digital Humanities*, *1*(2), http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

Whipple, T. (2015, February 25). Computer reads between the lines and finds there are only six plots. *The Times*, http://www.thetimes.co.uk/tto/arts/books/fiction/article4365964.ece

Adam Hammond's research focuses on the intersections of technology, politics, and artistic expression in modernist and contemporary literature. He is the author of *Literature in the Digital Age: An Introduction* (Cambridge 2016) and coauthor of *Modernism: Keywords* (Wiley-Blackwell 2014). His research on computational analysis of modernist dialogism, coauthored with Julian Brooke and Graeme Hirst, has appeared in *Digital Scholarship in the Humanities*, *Linguistic Issues in Language Technology*, and the collection *Reading Modernism with Machines* (Palgrave 2016). He is cocreator with Julian Brooke of the websites *He Do the Police in Different Voices* and *The Brown Stocking*, focused respectively on representing voices in T. S. Eliot's *The Waste Land* and Virginia Woolf's *To the Lighthouse*. His work on writers and critics such as Mikhail Bakhtin, Virginia Woolf, Wyndham Lewis, and Marshall McLuhan has appeared in *Modernism/Modernity*, *Style*, and collections such as *Counterblasting Canada* (University of Alberta Press 2016) and *Rereading the New Criticism* (Ohio State University Press 2012).