REASONS TO ENGAGE COMPOSITION THROUGH BIG DATA

BENJAMIN MILLER AND AMANDA LICASTRO

Especially during hard times, hard data are a godsend.

—RICHARD H. HASWELL

Recognizing that quantitative assessments of writing aren't likely to vanish, Ed White has repeatedly urged writing program administrators (WPAs) to *assess or be assessed*—that is, to take an active role in designing assessments, so as to infuse them with the values and questions important to WPAs, rather than ceding control to outside forces. We propose that a similar dynamic is now at work with regard to datafication and large-scale computational analysis, which is to say, big data.

Big data has changed the way information is processed, and thus the environment in which writing happens. Corporations analyze patterns in what people buy, how far they run, where they spend their time; they quantify habits to create more effective advertisements and cross-promotions. This development has been lauded by commerce, but it also raises the specter of deep intrusions into private spaces, such that big data can seem one step from Big Brother in its constant surveillance. Within academia, criticisms of such surveillance states coexist with enthusiasm for the potential new insight offered by studying texts at greater scale than individuals can read without algorithmic approaches. As the media scholars danah boyd and Kate Crawford have noted, "Data sets that were once obscure and difficult to manage—and, thus, only of interest to so-

cial scientists—are now being aggregated and made easily accessible to anyone who is curious, regardless of their training" (664). What's "big" about big data, then, is not the information itself, but the number of people able to access and interrogate that data. Given the developments in plug-and-play platforms for data analysis, more researchers, administrators, and students have the tools they need to find evidence for wide trends or correlations, and they can easily make slick infographics to represent and communicate those findings.

In recent years, a small but growing number of scholars in Rhetoric, Composition, and Writing Studies (RCWS)[1] have begun to apply the tools of big data—computational analysis, data visualization, natural language processing, etc.—to their own research questions. These studies aim both to better understand the contours of the field at large (Mueller, "Grasping"; Almjeld et al.; Miller; Gatta; Lauer) and to test the claims of earlier scholarship against larger bodies of writing than individuals could reasonably parse (Lancaster; Moxley; Licastro; Jamieson et al.; Aull). They demonstrate that it is increasingly possible to examine thousands of documents and peer-review comments, labor hours and citation networks, both in Composition courses and beyond. What questions will we ask of this data? What further data do we want to collect or examine, given the field's longstanding and emerging questions? And what protections or special considerations need to be considered for RCWS contexts? We have not, as a field, had a collective reckoning with the role that algorithmic and computational approaches can or should play in our research and teaching.

After all, Composition has in many ways long been concerned with matters of scale: how to teach thousands of students, how to prepare hundreds of teaching assistants, how to manage the paper load. Some savvy WPAs have always gathered their own data, so as to set the terms in which they frame conversations with upper administration. But recent developments in computational power and access change *the way* we work at scale, and we haven't yet dealt with the implications and possibilities of these changes. Perhaps some of us have been having conversations, but locally. This book brings these conversations together and into the open.

It is important that we do so, because—to paraphrase White—if we do not use data, we may well be used *by* data. For instance, RCWS research often entails working with potentially vulnerable populations, or populations in vulnerable moments. Because our work focuses on learners at varying stages of literacy or rhetorical savvy, and involves sharing

drafts and revisions in which ideas are still being developed, we must be cautious in how we represent data gleaned from these sources. This is especially true given what we know about statistical representation in large datasets: those who are already overrepresented can be amplified if we are not careful, while those underrepresented can be flattened or elided.[2] The risks multiply now that many institutions and writing programs are adopting learning management systems, institutional repositories, and commercial databases that may digitally archive hundreds—if not thousands or tens of thousands—of student compositions from across levels and disciplines. These repositories can often increase access for researchers, but at the same time can decrease transparency in the process of opting into participation in these systems, complicating the ethical issues surrounding research on the texts contained within.

Parallel to efforts in writing studies, digital humanists have investigated the specifically digital affordances of archives, and opened the possibility of algorithmic criticism. Digital Humanities (DH) scholars are using computational analysis to identify patterns in literary texts, historical documents, image archives, and sound, all of which has added to the body of knowledge in humanities theory and methodology.[3] As the editors of *Rhetoric and the Digital Humanities* point out, there is much to be gained by considering these two titular fields in tandem—and especially the subfields of Computers & Writing and "computational rhetorics" (Ridolfo and Hart-Davidson 10). As Jerome McGann argues in "On Creating a Usable Future," "We are clearly beginning to see new interpretive possibilities emerge from computerized information processing on one hand and graphical interface design on the other" (185).[4] This is echoed in RCWS by Derek Mueller's advocacy for what he calls—combining Franco Moretti's "distant reading" and Heather Love's "thin description"—a *distant-thin methodology* for disciplinary inquiry (*Network* 3; our emphasis), whereby databases and abstracting practices, such as visual models, make possible the detection and representation of patterns across a variety of scales. Mueller is especially interested in the capacity for such an approach to study disciplinarity itself, enabling both newcomers and old hands to maintain a "network sense" of how their work (teaching, research, administration) is situated within a larger activity system than one can ordinarily encounter without digital tools. And, as Annie Swafford and others have argued, appeals to databases and algorithms also come with the benefit of facilitating replication—or, for that matter, falsification.

What makes all this big data work possible is code, of two kinds: (1) the schemas that transform life into data, and (2) the scripts that interact with that data. Together, these two kinds of code allow reconfigurations, re-representations, and reexaminations of the same phenomena. Rather than "necessitating the abandonment of social, psychological, and historical preconceptions," as Berlin feared empiricism would require (*Rhetoric and Reality* 8), computer programs allow researchers to deliberately vary and so investigate the "preconceptions" by which we filter or compare the texts and processes we study. Like writing, both forms of coding are epistemic: they generate new meaning and understanding in the process of composing. In the course of marking up some real-life phenomenon into a format amenable to data processing, we're making choices among possible constructions—and those choices teach us more about our coding schema and what it helps us see. And in the course of programming some function to be run on our data, we're learning more about the program and how different pieces of code work together. We're leveling up, both on this program and on others like it.

Coding in the first sense, as quantitative categorization of empirically observed behaviors, dates back within Composition at least to Janet Emig's pioneering work in *The Composing Processes of Twelfth Graders* (1971), and was picked up by Sondra Perl, Linda Flower, and John Hayes, and others influenced by the process movement; even post-process, a number of presentations and workshops in recent years at the Conference on College Composition and Communication (CCCC) (e.g., Howard et al.; Lunsford et al.; Smith et al.) show the endurance of this kind of quantification and analysis.

In the second sense, similarly, *coding* as the composing of computer algorithms has a long history in the context of writing studies. As far back as Hugh Burns's TOPOI (1980) and Helen Schwartz's SEEN (1984)—open-ended programs for invention, revision, and peer review—compositionists have used programming to aid in writing instruction and research (see DeWitt 45–47).[5] From the design of MOOs[6] through Flash ActionScript on up through Karl Stolley's suggestion that everyone learn JavaScript and Ruby, the archives of *Computers & Composition* (founded in 1983) and *Kairos* (founded in 1996) are replete with examples of Writing and Rhetoric scholars composing in programming languages. This kind of code "has not only made its way into our research," as Annette Vee and Mark Sample noted in 2012, "it has also found its way into our classrooms."

In recent years, we have seen a newfound comfort in talking about such coding work as work *with data*. RCWS research is in the midst of a flowering of new possibilities for integrative and comparative studies—for networks of significance that will help us recognize and value not only the big patterns but also the interesting departures from those patterns. We see this as a swing of the pendulum back to large-scale, aggregative research after a long time away.

A major feature of the oft-referenced social turn in Composition was resistance to anything approaching the "positivistic position of modern science" (Berlin, "Contemporary Composition" 777). Like James Berlin, Ann Berthoff claimed not only that "empirical research requires that meaning be left out of account" but also that it would therefore be incompatible with claims of "relevan[ce] to pedagogy" (746). Part of the problem, as David Foster claimed, was the sense that supporting empirical methods would seem to assert those methods' primacy over humanistic, dialectical methods (37–38). Yet, as Carol Berkenkotter argued in response, the net result was frequently not balance, but rather a rejection of data-driven study; concluding that, "as an empirically oriented researcher, it's important to me to be accountable to what I call 'the data,'" she pointed out that "that very phrase, 'the data,' implies a model of knowing that's different from yours" [i.e., Foster's] (80). But once we move away from data, as even Foster noted, we risk throwing away any possibility of cumulative understanding of writing-related phenomena, for "rhetoric's understandings are not cumulative but dialectical" (36).

Indeed, as Richard Haswell has documented, publication of replicable/aggregable/data-supported (RAD) studies declined across all of the Composition journals published by the National Council of Teachers of English (*College Composition and Communication*, *College English*, *Research in the Teaching of English*) from the 1980s to the early 2000s ("NCTE").[7] The dominant form of evidence became, instead, anecdote (Johanek 9–11). Cindy Johanek notes the irony: "To argue . . . that narratives, anecdotes, and stories are *always* more true than numbers, that numbers are *always* for some reason out of context and narratives are not, that it is *always* appropriate to share a researcher's personal voice ignores the very thing to which we claim to be rhetorically most sensitive: context" (88). We agree with Johanek that if we take seriously the value of individual, contextualized experience, we should also value the contextualizing power of large-scale, aggregate experience.

In the last decade, RCWS has increasingly returned to this shared project of data-driven research. Since 2009, CompPile and the Council of Writing Program Administrators have solicited and published twenty-six (and counting) state-of-the-research bibliographies (*WPA / CompPile Research Bibliographies*), the prior lack of which Haswell had included in documenting the headwinds for RAD research in 2005 (206, 213–16). Since 2011, Dartmouth College has sponsored a summer institute specifically to support scholars in developing data-driven research projects (Donahue). The International Conference on Writing Analytics, dedicated to "Actionable Data for Teaching and Learning Writing," has met annually since 2014, and recently launched a journal publishing long-form studies, the *Journal of Writing Analytics*. Even generalist journals, including *College Composition and Communication*, have now given pride of place to data mining as an equally valued method of research (Lang and Baehr).

With *Composition and Big Data*, we take stock of this *re*turn, and consider where it might be leading. From ethical reflections to database design, from corpus linguistics to quantitative autoethnography, the authors in this book interpret and implement the drive toward data in diverse ways. Their work takes place in various contexts, including programmatic assessment, first-year pedagogy, stylistics, and learning transfer across the curriculum. What we have valued in assembling the research in this volume, and what we have striven for in our own research, is work that combines qualitative and quantitative methods, recognizing that data doesn't speak for itself, but must be spoken into and from, based on deep disciplinary knowledge.[8]

We are aware of skepticism that any data-driven arguments can comprehend something as variable, context-specific, and interpersonal as writing. David Smit, for example, has argued that Composition should "capitalize on the fact that it is now localized, historicized, and contingent, both theoretically and pedagogically" (230) by openly declaring that we do not—and cannot—know anything cumulative or transferable about writing. Metaphorically speaking, says Smit, "There is no such thing as 'tree-ness'; there are only particular trees" (230).[9] Even digital humanists see reason to be skeptical. "Too often," boyd and Crawford warn, "Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions" (668). This is surely possible,

and we have no need to multiply examples of spurious correlations, especially when there are websites dedicated to preserving some of the most amusing ones (Vigen).

Yet we believe it remains important to ask: if we don't look for patterns in the data at all, what real effects, developmental patterns, and influences on writers might we be missing? In themselves, algorithmic analyses are not guaranteed to be conclusive; but neither is any other kind of evidence, in itself. The primary principle is one of corroboration: multiple indicators, from multiple vantages, pointing in the same direction. Big data can both provide further support—or challenges—to our existing hypotheses, or it can generate new ones that future researchers and teachers can put to the test in their own local contexts. In saying so, we put data analysis in the same camp as case studies, ethnographies, and philosophical inquiry, which is to say, right alongside other widely accepted means of studying Composition, Rhetoric, and Writing.

This volume is organized to highlight the range of disciplinary questions that can be addressed through algorithmic analysis of large datasets. The studies in section one have direct application to writing classrooms; here you can find evidence-based claims about how students compose, as well as exercises to teach students to perform their own data analysis. Section two broadens the scope of inquiry to consider programmatic perspectives and questions of placement and genre. As section three demonstrates, big data is particularly useful for tracking complex disciplinary shifts over time, or for facilitating conversations about how we choose to focus our collective and individual time. And while the authors throughout this collection address the ethical questions raised by their projects, section four squarely centers these questions with a series of arguments about responsible design and interpretation of big data research.

SECTION ONE: DATA IN STUDENTS' HANDS

We open with a chapter that shows students pursuing their own big-data research projects. In chapter 1, "Learning to Read Again: Introducing Undergraduates to Critical Distant Reading, Machine Analysis, and Data in Humanities Writing," Trevor Hoag and Nicole Emmelhainz share student reflections from an Introduction to Digital Humanities course run by Writing/Rhetoric faculty. Contrasting students' engaged uses of digital text-visualization tools such as Voyant and Textalyzer to

published fears about such tools' potentially dulling effects, Hoag and Emmelhainz highlight the metacognitive skills students evince in writing about their distant readings, especially in adjusting their predictions and hypotheses about the texts they were studying. They note the continuities with existing theories of learning: because "distant reading . . . reveals how no one thinks alone," it demonstrates that "thinking, writing, and communication are fundamentally networked" (25). Chapter 1 articulates something we want to highlight: computer-assisted data analysis is compatible with the work that we do in writing studies, both in our research and teaching.

The next two chapters bring the evidence of big datasets to long-standing questions of practical stylistics, and so give writing teachers and students a clearer articulation of the trajectory from beginning to expert academic prose. In chapter 2, "A Corpus of First-Year Composition: Exploring Stylistic Complexity in Student Writing," Chris Holcomb and Duncan A. Buell consider a writerly trait in which some say academics often go too far, obscuring rather than enhancing comprehension. Rather than treat expert writing as purely worth emulating, then, Holcomb and Buell propose that we teach the range of stylistic options available, helping students to choose a rhetorically appropriate style. Data-driven studies like theirs will help teachers "to be informed about the stylistic conventions in question," so as to "gauge where student writers fall in relation to them" (36). Taking advantage of an important affordance of data-supported work, this is a replication-extension study, building on—and complicating—the findings of Biber et al. on the ways in which student writing adopts features of spoken versus academic discourses. The data tables presented in chapter 2 allow for direct comparisons among corpora, and set the stage for future aggregation and comparison.

In chapter 3, "Expanding Our Repertoire: Corpus Analysis and the Moves of Synthesis," Alexis Teagarden begins to fill a gap between expected outcomes and instructional materials: though colleges often want students to make arguments synthesizing multiple sources, very few textbooks or scholarly articles explicitly show students how to do so. Comparing corpora of published scholarship and student essays, Teagarden identifies clusters of phrases commonly used to build shared perspectives from multiple texts. In a hopeful finding for the teachability of writing, she finds that when students demonstrate successful synthesis, the writing takes essentially the same shape as that in published articles:

"The important divide thus appears not between experts and students but rather between successful and unsuccessful student writing" (60).

SECTION TWO: DATA ACROSS CONTEXTS

While section one develops applications for single classrooms, section two expands the scope to programmatic and cross-curricular concerns. The chapters in this section employ machine-learning approaches such as topic modeling and keyword extraction to pursue signals of knowledge transfer across several divides: from high school into college, from one writing course to another, and from First-Year Composition (FYC) into other coursework. This section will be particularly helpful for those seeking to make arguments about the efficacy of Writing Across the Curriculum (WAC) initiatives or the importance of writing programs in higher education.

In chapter 4, "Localizing Big Data: Using Computational Methodologies to Support Programmatic Assessment," David Reamer and Kyle McIntosh begin with two ubiquitous, but often criticized, sources of data about course outcomes: student grades and course evaluations. Working with their university registrar, they were able to obtain anonymized records that preserved links between two different versions of an introductory writing course and subsequent writing-intensive courses; this allowed them to investigate the impact of recent revisions in the course sequencing, while preserving individual students' privacy. Their chapter provides a clear example of how quantitative measures can provide useful feedback to WPAs on curricular changes, building on questions and hypotheses that WPAs are likely already asking.

One such question, regarding student placement within a sequence of writing courses, is taken up by Laura Aull in chapter 5, "Big Data as Mirror: Writing Analytics and Assessing Assignment Genres." Working with two comparable versions of Directed Self-Placement (DSP) prompts, Aull asks how students' treatment of source material changes when the writing assignment primes them to think of either argument or explanation. She finds changes at the level of word and syntax that reflect different understandings of the relationship between author and source cued by these different genre markers. These results suggest both that WPAs should consider the language they use in DSP assignments, and how well the signaled relationships to source material will align with the common tasks and learning outcomes of the writing courses students will soon enter.

Moving outside the writing program proper, in chapter 6, "Peer Review in First-Year Composition and STEM Courses: A Large-Scale Corpus Analysis of Key Writing Terms," Chris M. Anson, Ian G. Anson, and Kendra Andrews offer corroborating evidence that specific wording matters in writing prompts, and in the process they demonstrate a key feature of large-scale data-driven research: its interoperability. Their chapter combines data from a survey of writing instructors and administrators with a large body of writing from peer-review assignments in chemistry, gathered through the My Reviewers platform. Through a series of analyses, they find that transfer of metacognitive writing knowledge out of FYC appears to be uneven—but that transfer may be encouraged by a "shared vocabulary that faculty can use systematically" across disciplines, which would then "[help] students to apply rhetorical and discourse-related concepts as they move across the landscape of higher education" (121).

Somewhat complicating that goal, in chapter 7, "Moving from Categories to Continuums: How Corpus Analysis Tools Reveal Disciplinary Tension in Context," Kathryn Lambrecht argues that efforts to forge cross-disciplinary, interdisciplinary, or transdisciplinary collaborations can falter when shared vocabulary masks unshared meanings. Lambrecht triangulates an analysis of academic subsets of the Corpus of Contemporary American English (COCA) against evidence from ethnographic observations and interviews with both students and professors across four disciplines. Her work demonstrates the power of what we might call *medium data*: a collection large enough to support modeling and trend detection, but small enough to sustain reentry into individual transcripts for verification and nuanced understandings of the patterns detected.

SECTION THREE: DATA AND THE DISCIPLINE

The chapters in this section ask and respond to a few essential questions about the relationship between big data and RCWS. For example: Why do we need to collect, archive, and study data in the field of RCWS? Who should steward this work and what considerations should be taken into account during the process? These chapters alternate between systematic planning strategies to support finding and preparing data for use, and examples that illustrate the research opportunities those strategies make possible.

Demonstrating big data's capacity to help manage complexity, in chapter 8, "From 1993 to 2017: Exploring 'A Giant Cache of (Disci-

plinary) Lore' on WPA-L," Chen Chen conducts a distant reading of the Council of Writing Program Administrators' listserv. She notes that WPA-L, though often informal, is an active space for knowledge production, debate, and responses to moments of crisis, making it an important space for disciplinary culture. Yet that same high activity—and contention—also makes it a fraught space for newcomers, who can sometimes feel either lost or excluded from the list's history. (Between the writing and publication of chapter 8, such feelings led to the public and acrimonious departure of many subscribers, some of whom formed a new listserv, NEXTGEN.) Chen distills the subject lines from a quarter century of posts, culled from before the split, into a series of visualizations and tables, identifying the most frequently engaged topics of discussion. So doing, she enables comparisons with other datasets to highlight WPA-L's role in advancing disciplinary conversations, sometimes ahead of publications and conference presentations.

Drawing on her experience working with the National Archive of Composition and Rhetoric, in chapter 9, "Composing the Archives with Big Data: A Case Study in Building a Collaboratively Authored Metadata Information Infrastructure," Jenna Morton-Aiken argues for the use of folksonomic tagging practices. Morton-Aiken demonstrates how such a system, which she calls a relational architecture, enables us to "habitually interrogate who can enter systems, what rhetorical forces inform—and possibly constrain—organization, and how users can/not engage with resources" (168). By easing the process of adding digital metadata, we can facilitate collaboration not only in real time but also across time by preserving prepublication connections made by researchers in the archive.

In chapter 10, "Big-Time Disciplinarity: Measuring Professional Consequences in Candles and Clocks," Kate Pantelides and Derek Mueller use data visualization as an opportunity for both personal and professional reflection, to "usefully complicate our understanding of how we *do* disciplinary time" (189). Pantelides and Mueller model the process of developing a "(small) big dataset," working week by week to fill in the pieces of a recurring puzzle: how disciplinary commitments like conferences come to fill far more time across a year than their meeting dates alone suggest. Ultimately, Pantelides and Mueller's efforts both challenge RCWS to plan more intentionally as it expands, and offer a way to assemble the information we will need to make those plans.

A common thread across the chapters in section three is the degree to which our data—like our lives—are more interconnected than is often clear at first glance. Aiming to solve the problem of bringing these data together, in chapter 11, "The Boutique Is Open: Data for Writing Studies," Cheryl E. Ball, Tarez Samra Graban, and Michelle Sidler outline four principles by which to make writing studies data more open—to shift from a repository mindset to a collaboratory mindset. Key to their proposal is a renewed understanding of data aggregation: not as settling into some fixed body of knowledge, but rather as participating in an ongoing process of emergent reimagining. For example, rather than imagine that data mining produces "scientistic representations of what is there," they urge us to use data mining to generate "topoi indicating what could be there" (202). Such goals are not without their challenges: two projects they had initially seen as promising, REx and rhetoric.io, have gone dormant in the past few years. Yet we share their optimism that practitioners in the field of writing studies will work together to build and maintain a platform for sharing data in the future. Such a platform could address sustainability concerns by allowing researchers outside of local incentive structures to pick up investigations where the initial data collectors are forced to leave off.

SECTION FOUR: DEALING WITH DATA'S COMPLICATIONS

How do we ensure that our research practices are responsive to those affected by them, and who is responsible for overseeing these assurances? Institutional Review Boards (IRBs) are often tasked with investigating the ethical implementation of research, but may not yet have considered the ways in which humanities engage with data. This is especially true for digital data generated by protected classes of subjects, including students—federal guidance on internet research and data mining notwithstanding. Along with minimizing harm, the chapters in this section take up the question of the researcher's own perspective, including how to keep your head up and your eyes open in the face of inevitable challenges. As we continue to update our norms in response to a data-rich environment, this section begins, but does not end, the conversation about how to research responsibly.

Drawing on long experience working with multiple IRBs and using the My Reviewers platform, in chapter 12, "Ethics, the IRBs, and Big Data Research: Toward Disciplinary Datasets in Composition," Johanna Phelps describes the challenges of getting new data sets approved for

analysis. Instead, she argues for "widely accessible, readily shared, de-identified big data sets" (220) to facilitate replication studies, minimize conflicts of interest, and smooth the review process. This chapter is essential reading for anyone preparing to apply for IRB approval for a big data project, especially graduate students working on their dissertations. Phelps manages to simultaneously interrogate the ethical quandaries that make applying for approval for big data work difficult, while ultimately demystifying the IRB process for RCWS researchers.

In chapter 13, "Ethics in Big Data Composition Research: Cybersecurity and Algorithmic Accountability as Best Practices," Andrew Kulak reminds us that the algorithms researchers rely on to process large datasets are "mindless" but "not theoryless," a combination that means our starting assumptions about who and what is represented can compound and intensify as we scale up. Composition's long-standing commitment to expanding access to the academy, and interrogating the power dynamics of students and teachers—especially across lines of race, class, and gender—should therefore not disappear behind numbers and code; rather, Kulak argues, we should begin from principles of transparency and security, considering (so as to counter) the potentially pernicious effects and reuses of the data we gather and the stories the data can tell.

Similarly cautioning that "data do not speak for themselves," in chapter 14, "Data Do Not Speak for Themselves: Interpretation and Model Selection in Unsupervised Automated Text Analysis," Juho Pääkkönen turns a careful eye to the subjectivity at the heart of a common big data algorithm: topic modeling with Latent Dirichlet Allocation (LDA). Though proponents of LDA sometimes imply that it enables researchers to "find" inherently interpretable "topics" within a corpus, in the form of subsets of documents with co-occurring words, Pääkkönen points out that the reported topics are not only subject to interpretation but constructed by means of their interpretability. That is, the same corpus can provide an infinite number of models, with varying numbers of topics; researchers tend to decide what version to report based not on something inherent to the text or the model, but based on whether they have something to say about it. As a result, Pääkkönen argues, we may be overstating the possible conclusions from such an analysis.

In chapter 15, "'Unsupervised Learning': Reflections on a First Foray into Data-Driven Argument," Romeo García looks back at his own initial aims and anxieties when he was a graduate student. Reflecting on

his use of keyword frequencies and correlations in a study of racism and antiracism in writing centers, García explores a number of issues that can complicate researchers' attempts to work with big data: questions of genre and audience, of self-teaching versus directed study, of linguistic difference and the potential elisions or flattening effects of distant reading. Chapter 15 both illustrates one learner's path and raises important points about ethics and responsibility, exclusion and access.

Picking up on a cautionary thread also sounded by Reamer and McIntosh, in chapter 16, "Making Do: Working with Missing and Broken Data," Jill Dahlman describes a number of ways that the data on hand may be incomplete, messy, or even "broken." From vague or broad questions in a survey instrument, to mismatched genres of student writing in the same files, to numbers generated unevenly by applying subjective processes, there's a lot that could lead a potential researcher to throw up their hands. Rather than despair, Dahlman offers heuristic ways of mitigating or even taking advantage of the complications thrown up by data-driven research.

From ethical reflections to database design, from corpus linguistics to quantitative autoethnography, the authors in *Composition and Big Data* interpret and implement the drive toward data in diverse ways. Their work takes place in various contexts, including programmatic assessment, first-year pedagogy, stylistics, and learning transfer across the curriculum. In assembling this collection, our aim was to bring together a range of scholars, teachers, and administrators in RCWS working with big data methods and datasets in order to kick-start a needed conversation about the role that algorithmic and computational approaches can, or should, play in our research and teaching.

In other words, this collection is far from the final word on these matters, nor do we intend it to be. In addition to direct responses to the questions raised here, we know more work is needed on whether and how to apply machine learning to assessment of individual texts, to take just one example. We would welcome further large-scale analysis of RCWS networks of citation and participation, especially to investigate representation along lines of gender, race, nationality, and institutional type, parallel to work in DH by Roopika Risam and Amy Earhart, Tara McPherson, and Lauren Klein. And we look forward to the new efforts that can be inspired by such studies, in the form of collaborative bibliographies, databases, and other resources.

Boyd and Crawford's note about training, cited at the beginning of this introduction, is doubly important: where we have not yet, as a field, centered computational and algorithmic thinking, we have limited our ability both to perceive patterns beyond direct experiences and also to communicate across fields to colleagues and evaluators who read our data—and the term *data* itself—through different assumptions and concerns. In other words, learning to speak the language of data will allow us to communicate across local contexts within the field, and across fields within a local context. But this language, like others, will continue to move and grow. We offer this collection as a starting point on the never-ending path of training and learning.

NOTES

1. Naming the field is always difficult, and no choice fully satisfactory; see the 2003 special issue of *enculturation* for arguments about alternatives. Rhetoric, Composition, and Writing Studies (RCWS)—which was recently adopted by the Modern Language Association as the name of a high-level forum of scholarly and professional interest—has the advantages of an umbrella term, signaling a breadth of related research, teaching, and administrative concerns. Other choices throughout the book, and indeed in the book's title, are used less to convey substantive differences and more in service of euphony and economy.

2. We believe that some types of data visualization actually offer ways to discern and amplify quieter signals that might go unnoticed when only some parts stand in for wholes, but that is a discussion for another essay.

3. See, for example, the ongoing Debates in Digital Humanities series (eds. Gold; Gold and Klein), which traces the roots of DH and identifies trends in the ever-evolving field.

4. McGann's call for open-access repositories is also taken up and complicated in this collection (see chapter 11).

5. For a thorough history of the early days of Computers and Composition, see Gail E. Hawisher, Paul LeBlanc, Charles Moran, and Cynthia L. Selfe's *Computers and the Teaching of Writing in American Higher Education, 1979–1994: A History* (Praeger, 1995), or the summary timeline posted by Joseph Wilferth and Paul Cesarini.

6. MOO stands for "MUD, Object Oriented," and MUD stands for Multi-User Domain. For a history of MOOs in Writing Studies, see Haynes and Holmevik.

7. Haswell graphed RAD studies in the CompPile index of publications in Composition/Rhetoric to demonstrate their ongoing production outside of

NCTE's mainline journals, even as they faded from view. Note that this also expands our understanding of what counts as *data* to include a bibliography, when treated as a queryable database.

8. That said, we are persuaded by Derek Mueller's argument in *Network Sense* that big data, and the distant-thin reading it affords, are increasingly valuable ways of building and maintaining such disciplinary knowledge.

9. See also his *End of Composition Studies*.

WORKS CITED

Almjeld, Jen, Allison Michelli, Chelsea Weatherhead, Kortney Frederick, Samantha Perez, Julia Germain, Mallory O'Shea, Meghan Lavin, Tyler Haas, Rebekah Pitts, Troy Fultz, Rachel Fisher, Peggy Michel, Kelly Vingelis, Sierra McAliney, Megan O'Neill, Kinzie Stanley, Judy Hong, Hillary Chester, Morgan Shaughnessy, and Brooklyn Steele. "The F-Word: A Decade of Hidden Feminism in Kairos." *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*, vol. 20, no. 2, 2016, http://technorhetoric.net/20.2/reviews/almjeld-et-al/.

Aull, Laura. *First-Year University Writing: A Corpus-Based Study with Implications for Pedagogy*. Palgrave Macmillan, 2015.

Berkenkotter, Carol. "The Legacy of Positivism in Empirical Composition Research." *JAC: Journal of Advanced Composition*, vol. 9, no. 1/2, 1989, pp. 69–82.

Berlin, James A. "Contemporary Composition: The Major Pedagogical Theories." *College English*, vol. 44, no. 8, 1982, pp. 765–77. *JSTOR*, doi:10.2307/377329.

Berlin, James A. *Rhetoric and Reality: Writing Instruction in American Colleges, 1900–1985*. Southern Illinois UP, 1987.

Berthoff, Ann E. "Is Teaching Still Possible? Writing, Meaning, and Higher Order Reasoning." *College English*, vol. 46, no. 8, 1984, pp. 743–55. *JSTOR*, doi:10.2307/377206.

boyd, danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society*, vol. 15, no. 5, June 2012, pp. 662–79. *CrossRef*, doi:10.1080/1369118X.2012.678878.

Donahue, Christiane. "Dartmouth Summer Seminar for Composition Research: 'Got Data—Now What?'" *Council of Writing Program Administrators*, http://wpacouncil.org/node/2968. Accessed 28 Sept. 2018.

DeWitt, Scott Lloyd. *Writing Inventions: Identities, Technologies, Pedagogies*. SUNY P, 2001.

Earhart, Amy. "Can We Trust the University? Digital Humanities Collabo-

rations with Vulnerable Populations." *Bodies of Information: Feminist Debates in Digital Humanities*, edited by Jacqueline Wernimont and Elizabeth Losh, vol. 3, U of Minnesota P, 2018.

Emig, Janet. *The Composing Processes of Twelfth Graders*. National Council of Teachers of English, 1971.

Flower, Linda, and John R. Hayes. "The Cognition of Discovery: Defining a Rhetorical Problem." *College Composition and Communication*, vol. 31, no. 1, Feb. 1980, pp. 21–32.

Foster, David. "What Are We Talking About When We Talk About Composition?" *JAC: Journal of Advanced Composition*, vol. 8, no. 1/2, 1988, pp. 30–40.

Gatta, Oriana. "Connecting Logics: Data Mining and Keyword Visualization as Archival Method/ology." *Peitho*, vol. 17, no. 1, Fall/Winter 2014, p. 15.

Gold, Matthew K., editor. *Debates in the Digital Humanities*. U of Minnesota P, 2012.

Gold, Matthew K., and Lauren F. Klein, editors. *Debates in the Digital Humanities 2016*. U of Minnesota P, 2016.

Haswell, Richard H. "NCTE/CCCC's Recent War on Scholarship." *Written Communication*, vol. 22, no. 2, Apr. 2005, pp. 198–223. *Sage Journals Online*, doi:10.1177/0741088305275367.

Haswell, Richard H. "Quantitative Methods in Composition Studies: An Introduction to Their Functionality." *Writing Studies Research in Practice: Methods and Methodologies*, edited by Lee Nickoson and Mary P. Sheridan, Southern Illinois UP, 2012.

Haswell, Richard, and Glenn Blalock. *CompPile*. Site currently maintained (2021) by Glenn Blalock and Susan Wolff Murphy. https://comppile.org.

Haynes, Cynthia, and Jan Rune Holmevik, eds. *High Wired: On the Design, Use and Theory of Educational MOOs*. U of Michigan P, 1998.

Hawisher, Gail E., Paul LeBlanc, Charles Moran, and Cynthia L. Selfe. *Computers and the Teaching of Writing in American Higher Education, 1979–1994: A History*. Praeger, 1995.

Howard, Rebecca Moore, Rebecca Rickly, Jo Mackiewicz, and Karen Lunsford. "What Coding Means and Why We Should Do It." Concurrent session, Conference on College Composition and Communication, Las Vegas, TX, 2013.

Jamieson, Sandra, Rebecca Moore Howard, and Tricia Serviss. "What Is the Citation Project?" *The Citation Project: Reframing the Conversation about Plagiarism*, http://www.citationproject.net/about/. Accessed 25 Oct. 2018.

Johanek, Cindy. *Composing Research: A Contextualist Paradigm for Rhetoric and Composition*. Utah State UP, 2000. *HathiTrust Digital Library*, https://babel.hathitrust.org/cgi/pt?id=usu.39060010190998;view=1up;seq=4.

Journal of Writing Analytics. https://journals.colostate.edu/analytics/. Accessed 26 Oct. 2018.

Klein, Lauren F. "The Image of Absence: Archival Silence, Data Visualization, and James Hemings." *American Literature*, vol. 85, no. 4, Dec. 2013, pp. 661–88. *Duke University Press*, doi:10.1215/00029831–2367310.

Lancaster, Zak. "Do Academics Really Write This Way? A Corpus Investigation of Moves and Templates in 'They Say / I Say.'" *College Composition and Communication*, vol. 67, Feb. 2016, pp. 437–64.

Lang, Susan, and Craig Baehr. "Data Mining: A Hybrid Methodology for Complex and Dynamic Research." *College Composition and Communication*, vol. 64, no. 1, 2012, pp. 172–94.

Lauer, Claire. "Expertise with New/Multi/Modal/Visual/Digital/Media Technologies Desired: Tracing Composition's Evolving Relationship with Technology through the MLA *JIL*." *Computers and Composition*, vol. 34, Dec. 2014, pp. 60–75. *ScienceDirect*, doi:10.1016/j.compcom.2014.09.006.

Licastro, Amanda. "The Problem of Multimodality: What Data-Driven Research Can Tell Us About Online Writing Practices." *Communication Design Quarterly Review*, vol. 4, no. 4, Dec. 2016, pp. 55–74.

Lunsford, Karen, Jason Swarts, Jo Mackiewicz, and Rebecca Rickly. "MW.08 Coding for Data Analysis." Workshop, Conference on College Composition and Communication, Indianapolis, IN, 2014.

McGann, Jerome. "On Creating a Usable Future." *Profession*, vol. 2011, no. 1, Nov. 2011, pp. 182–95. *Cambridge Core*, doi:10.1632/prof.2011.2011.1.182.

McPherson, Tara. "Why Are the Digital Humanities So White? Or Thinking the Histories of Race and Computation." *Debates in the Digital Humanities*, edited by Matthew K. Gold, U of Minnesota P, 2012.

Miller, Benjamin. "Mapping the Methods of Composition/Rhetoric Dissertations: A 'Landscape Plotted and Pieced.'" *College Composition and Communication*, vol. 66, no. 1, 2014, pp. 145–76.

Moxley, Joe. "Big Data, Learning Analytics, and Social Assessment." *The Journal of Writing Assessment*, vol. 6, no. 1, 2013, http://www.journalofwritingassessment.org/article.php?article=68.

Mueller, Derek. "Grasping Rhetoric and Composition by Its Long Tail: What Graphs Can Tell Us about the Field's Changing Shape." *College Composition and Communication*, vol. 64, no. 1, Sept. 2012, pp. 195–223.

Mueller, Derek. *Network Sense: Methods for Visualizing a Discipline*. WAC Clearinghouse and UP of Colorado, 2017, https://wac.colostate.edu/books/network/sense.pdf.

Perl, Sondra. "The Composing Processes of Unskilled College Writers." *Research in the Teaching of English*, vol. 13, no. 4, 1979, pp. 317–36.

Ridolfo, Jim, and William Hart-Davidson, editors. *Rhetoric and the Digital Humanities*. U of Chicago P, 2015.

Risam, Roopika. "Navigating the Global Digital Humanities: Insights from Black Feminism." *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, U of Minnesota P, 2016.

Smit, David. *The End of Composition Studies*. Southern Illinois UP, 2004.

Smit, David. "Stephen North's The Making of Knowledge in Composition and the Future of Composition Studies 'Without Paradigm Hope.'" *The Changing of Knowledge in Composition: Contemporary Perspectives*, edited by Lance Massey and Richard C. Gebhardt, Utah State UP, 2011, pp. 213–35.

Smith, Jordan, Karen Lunsford, and Jo Mackiewicz. "MW.10 Basics of Coding: Analyzing Data and Reporting Findings." Workshop, Conference on College Composition and Communication, Houston, TX, 2016.

Stolley, Karl. "Source Literacy: A Vision of Craft." *Enculturation*, no. 14, Oct. 2012, http://enculturation.net/node/5271.

Swafford, Annie. "Problems with the Syuzhet Package." *Anglophile in Academia: Annie Swafford's Blog*, 2 Mar. 2015, https://annieswafford.wordpress.com/2015/03/02/syuzhet/.

Vee, Annette, and Mark Sample. "Introduction to 'The Role of Computational Literacy in Computers and Writing.'" *Enculturation*, no. 14, Oct. 2012, http://enculturation.net/computational-literacy.

Vigen, Tyler. "15 Insane Things That Correlate With Each Other." *Spurious Correlations*, http://tylervigen.com/spurious-correlations. Accessed 3 Aug. 2018.

Wilferth, Joseph, and Paul Cesarini. "A Timeline for Computers and the Teaching of Writing in American Higher Education, 1979–1994: A History." 10 June 1997, https://web.archive.org/web/20200410042253/http://personal.bgsu.edu/~pcesari/week4.html.

WPA / CompPile Research Bibliographies. https://wac.colostate.edu/comppile/wpa/. Accessed 28 Sept. 2018.