MACROANALYSIS
TOPICS IN THE DIGITAL HUMANITIES

Humanities computing is redefining basic principles about research and publication. An influx of new, vibrant, and diverse communities of practitioners recognizes that computer applications are subject to continual innovation and reappraisal. This series publishes books that demonstrate the new questions, methods, and results arising in the digital humanities.

SERIES EDITORS
Susan Schreibman
Raymond C. Siemens
MACROANALYSIS
DIGITAL METHODS AND LITERARY HISTORY
MATTHEW L. JOCKERS

UNIVERSITY OF ILLINOIS PRESS
Urbana, Chicago, and Springfield
Additional material referred to in this book, including confusion matrices, an expanded stop-words list, and additional graphs and color images, can be found at http://www.matthewjockers.net/macroanalysisbook/.
For my father,

Maestro Artigiano
CONTENTS

Acknowledgments ix

PART I. FOUNDATION
1 REVOLUTION 3
2 EVIDENCE 5
3 TRADITION 11
4 MACROANALYSIS 24

PART II. ANALYSIS
5 METADATA 35
6 STYLE 63
7 NATIONALITY 105
8 THEME 118
9 INFLUENCE 154
PART III. PROSPECTS

10 ORPHANS 171

References 177
Index 187
ACKNOWLEDGMENTS

Without initial support from John Bender, Lois Brooks, Mike Keller, Makoto Tsuchitani, and Ramón Saldívar, this work would not have been started. Without ongoing support from the Stanford University Department of English and the Stanford University Library, this work would not have continued. For years Glen Worthey has been my text pusher, my sounding board, my trusted digital humanities colleague, and my friend. His contributions to the completion of this project are innumerable. From Franco Moretti I have learned and been given much. Our partnership over the past seven years has been the most rewarding collaboration of my career. A special thank-you goes to Susan Schreibman, who convinced me to put these ideas into a book and then read drafts and offered kind and honest feedback. I thank Alan Liu for being such a generous and thoughtful reader of an early draft of the first few chapters; we should all be such gracious scholars. Stéfan Sinclair provided a careful and expert review of the entire manuscript, and I could not have asked for a better reader of the work. For chapter 8, David Mimno was essential; he indulged my layman’s questions about latent Dirichlet allocation with kindness and good humor. For chapter 9, Elijah Meeks was similarly tolerant and gave his time, energy, and expertise to help me understand and leverage the power of Gephi. For patient advice with R in particular and statistics more generally, I acknowledge indirect contributions from Vijoy Abraham, Claudia Engel, Ken Romeo, and Daniela Witten. For their direct and essential contributions to the final manuscript, I thank my copy editor, Annette Wenda, along with the top-notch University of Illinois Press team, especially Bill Regier and Tad Ringo, who fostered this project through to completion. A special and personal thank-you to my wife, Angela, who has been a patient reader, thoughtful editor, and tolerant friend for more than twenty years.
Importantly, enthusiastically, and with deep appreciation, I acknowledge my debt to the students who enrolled in my first course in “macroanalysis” at Stanford. I doubt that a future group of students could ever capture the enthusiasm and excitement of that year. Their energy and earnestness were the catalyst behind the founding of the Stanford Literary Lab, and they have been in my mind most often in the writing of this book: Richard Alvarez, Cameron Blevins, Ryan Heuser, Nadeen Kharputly, Rachel Kraus, Alison Law, Long Le-Khac, Rhiannon Lewis, Madeline Paymer, Moritz Sudhof, Amir Tevel, Ellen Truxaw, Kathryn VanArendonk, and Connie Zhu, you are all, quite simply, the best a teacher could hope for.
PART I FOUNDATION
1 REVOLUTION

The digital revolution is far more significant than the invention of writing or even of printing.
—Douglas Carl Engelbart

An article in the June 23, 2008, issue of Wired declared in its headline “Data Deluge Makes the Scientific Method Obsolete” (Anderson 2008). By 2008 computers, with their capacity for number crunching and processing large-scale data sets, had revolutionized the way that scientific research gets done, so much so that the same article declared an end to theorizing in science. With so much data, we could just run the numbers and reach a conclusion. Now slowly and surely, the same elements that have had such an impact on the sciences are revolutionizing the way that research in the humanities gets done. This emerging field we have come to call “digital humanities”—which was for a good many decades not emerging at all but known as “humanities computing”—has a rich history dating back at least to Father Roberto Busa’s concordance work in the 1940s, if not before.* Only recently, however, has this “discipline,” or “community of practice,” or “field of study/theory/methodology,” and so on, entered into the mainstream discourse of the humanities, and it is even more recently that those who “practice” digital humanities (DH) have begun to grapple with the challenges of big data.† Technology has certainly changed some things about the way literary scholars go about their work, but until recently change has been

* Roberto Busa, a Jesuit priest and scholar, is considered by many to be the founding father of humanities computing. He is the author of the Index Thomisticus, a lemmatized index of the works of Thomas Aquinas.

† Some have already begun thinking big. In 2008 I served on the inaugural panel reviewing applications for the jointly sponsored National Endowment for the Humanities and National Science Foundation “Digging into Data” grants. The expressed goals of the grant are to promote the development and deployment of innovative research techniques in large-scale data analysis; to foster interdisciplinary collaboration among scholars in
mostly at the level of simple, even anecdotal, search. The humanities computing/digital humanities revolution has now begun, and big data have been a major catalyst. The questions we may now ask were previously inconceivable, and to answer these questions requires a new methodology, a new way of thinking about our object of study.

For whatever reasons, be they practical or theoretical, humanists have tended to resist or avoid computational approaches to the study of literature.* And who could blame them? Until recently, the amount of knowledge that might be gained from a computer-based analysis of a text was generally overwhelmed by the dizzying amount of work involved in preparing (digitizing) and then processing that digital text. Even as digital texts became more readily available, the computational methods for analyzing them remained quite primitive. Word-frequency lists, concordances, and keyword-in-context (KWIC) lists are useful for certain types of analysis, but these staples of the digital humanist’s diet hardly satiate the appetite for more. These tools only scratch the surface in terms of the infinite ways we might read, access, and make meaning of text. Revolutions take time; this one is only just beginning, and it is the existence of digital libraries, of large electronic text collections, that is fomenting the revolution. This was a moment that Rosanne Potter predicted back in the digital dark ages of 1988. In an article titled “Literary Criticism and Literary Computing,” Potter wrote that “until everything has been encoded, or until encoding is a trivial part of the work, the everyday critic will probably not consider computer treatments of texts” (93). Though not “everything” has been digitized, we have reached a tipping point, an event horizon where enough text and literature have been encoded to both allow and, indeed, force us to ask an entirely new set of questions about literature and the literary record.

* I suspect that at least a few humanists have been turned off by one or more of the very public failures of computing in the humanities: for example, the Donald Foster Shakespeare kerfuffle.
While still graduate students in the early 1990s, my wife and I invited some friends to share Thanksgiving dinner. One of the friends was, like my wife and me, a graduate student in English. The other, however, was an outsider, a graduate student from geology. The conversation that night ranged over a wine-fueled spectrum of topics, but as three of the four of us were English majors, things eventually came around to literature. There was controversy when we came to discuss the “critical enterprise” and what it means to engage in literary research. The very term *research* was discussed and debated, with the lone scientist in the group suggesting, asserting, that the “methodology” employed by literary scholars was a rather subjective and highly anecdotal one, one that produced little in terms of “verifiable results” if much in the way of unsupportable speculation.

I recall rising to this challenge, asserting that the literary methodology was in essence no different from the scientific one: I argued that scholars of literature (at least scholars of the idealistic kind that I then saw myself becoming), like their counterparts in the sciences, should and do seek to uncover evidence and discover meaning, perhaps even truth. I dug deeper, arguing that literary scholars employ the same methods of investigation as scientists: we form a hypothesis about a literary work and then engage in a process of gathering evidence to test that hypothesis.

After so many years it is only a slightly embarrassing story. Although I am no longer convinced that the methods employed in literary studies are exactly the same as those employed in the sciences, I remain convinced that there are a good many methods worth sharing and that the similarities of methods exist in concrete ways, not simply as analogous practices.

The goal of science, we hope, is to develop the best possible explanation for some phenomenon. This is done via a careful and exhaustive gathering of evi-
We understand that the conclusions drawn are only as good as the evidence gathered, and we hope that the gathering of evidence is done both ethically and completely. If and when new evidence is discovered, prior conclusions may need to be revised or abandoned—such was the case with the Ptolemaic model of a geocentric universe. Science is flexible in this matter of new evidence and is open to the possibility that new methods of investigation will unearth new, and sometimes contradictory, evidence.

Literary studies should strive for a similar goal, even if we persist in a belief that literary interpretation is a matter of opinion. Frankly, some opinions are better than others: better informed, better derived, or just simply better for being more reasonable, more believable. Science has sought to derive conclusions based on evidence, and in the ideal, science is open to new methodologies. Moreover, to the extent possible, science attempts to be exhaustive in the gathering of the evidence and must therefore welcome new modes of exploration, discovery, and analysis. The same might be said of literary scholars, excepting, of course, that the methods employed for the evidence gathering, for the discovery, are rather different. Literary criticism relies heavily on associations as evidence. Even though the notions of evidence are different, it is reasonable to insist that some associations are better than others.

The study of literature relies upon careful observation, the sustained, concentrated reading of text. This, our primary methodology, is “close reading.” Science has a methodological advantage in the use of experimentation. Experimentation offers a method through which competing observations and conclusions may be tested and ruled out. With a few exceptions, there is no obvious corollary to scientific experimentation in literary studies. The conclusions we reach as literary scholars are rarely “testable” in the way that scientific conclusions are testable. And the conclusions we reach as literary scholars are rarely “repeatable” in the way that scientific experiments are repeatable. We are highly invested in interpretations, and it is very difficult to “rule out” an interpretation. That said, as a way of enriching a reader’s experience of a given text, close reading is obviously fruitful; a scholar’s interpretation of a text may help another reader to “see” or observe in the text elements that might have otherwise remained latent. Even a layman’s interpretations may lead another reader to a more profound, more pleasurable understanding of a text. It would be wasteful and futile to debate the value of interpretation, but interpretation is fueled by observation, and as a method of evidence gathering, observation—both in the sciences and in the humanities—is flawed. Despite all their efforts to repress them, researchers will have irrepressible biases. Even scientists will “interpret” their evidence through a lens of subjectivity. Observation is flawed in the same way that generalization from the specific is flawed: the generalization may be good, it may even explain a total population, but the selection of the sample is always something less than
Evidence

perfect, and so the observed results are likewise imperfect. In the sciences, a
great deal of time and energy goes into the proper construction of “representa-
tive samples,” but even with good sampling techniques and careful statistical
calculations, there remain problems: outliers, exceptions, and so on. Perfection
in sampling is just not possible.

Today, however, the ubiquity of data, so-called big data, is changing the sam-
ping game. Indeed, big data are fundamentally altering the way that much
science and social science get done. The existence of huge data sets means that
many areas of research are no longer dependent upon controlled, artificial ex-
periments or upon observations derived from data sampling. Instead of con-
ducting controlled experiments on samples and then extrapolating from the
specific to the general or from the close to the distant, these massive data sets
are allowing for investigations at a scale that reaches or approaches a point of
being comprehensive. The once inaccessible “population” has become accessible
and is fast replacing the random and representative sample.

In literary studies, we have the equivalent of this big data in the form of big li-
braries. These massive digital-text collections—from vendors such as Chadwyck-
Healey, from grassroots organizations such as Project Gutenberg, from nonprofit
groups such as the Internet Archive and HathiTrust, and from the elephants in
Mountain View, California, and Seattle, Washington*—are changing how literary
studies get done. Science has welcomed big data and scaled its methods accord-
ingly. With a huge amount of digital-textual data, we must do the same. Close
reading is not only impractical as a means of evidence gathering in the digital
library, but big data render it totally inappropriate as a method of studying liter-
ary history. This is not to imply that scholars have been wholly unsuccessful in
employing close reading to the study of literary history. A careful reader, such as
Ian Watt, argues that elements leading to the rise of the novel could be detected
and teased out of the writings of Defoe, Richardson, and Fielding. Watt’s study is
magnificent; his many observations are reasonable, and there is soundness about
them.† He appears correct on a number of points, but he has observed only a
small space. What are we to do with the other three to five thousand works of

* That is, Google.com and Amazon.com.
† A similar statement could be made of Erich Auerbach’s Mimesis. It is a magnificent
bit of close reading. At the same time, Auerbach was acutely aware of the limitations of
his methodology. In the epilogue to Mimesis, he notes the difficulties of dealing with
“texts ranging over three thousand years” and how the limitations of his library in Istan-
bul made it “probable that [he] overlooked things which [he] ought to have considered.”
Interestingly, however, he says at the same time that “it is quite possible that the book
owes its existence to just this lack of a rich . . . library.” If it had been possible to access
the greater archive, he “might never have reached the point of writing” (1953).
fiction published in the eighteenth century? What of the works that Watt did not observe and account for with his methodology, and how are we to now account for the works not penned by Defoe, by Richardson, or by Fielding? Might other novelists tell a different story? Can we, in good conscience, even believe that Defoe, Richardson, and Fielding are representative writers? Watt's sampling was not random; it was quite the opposite. But perhaps we only need to believe that these three (male) authors are representative of the trend toward "realism" that flourished in the nineteenth century. Accepting this premise makes Watt's magnificent synthesis into no more than a self-fulfilling project, a project in which the books are stacked in advance. No matter what we think of the sample, we must question whether in fact realism really did flourish. Even before that, we really ought to define what it means "to flourish" in the first place. Flourishing certainly seems to be the sort of thing that could, and ought, to be measured. Watt had no such yardstick against which to make a measurement. He had only a few hundred texts that he had read. Today, things are different. The larger literary record can no longer be ignored: it is here, and much of it is now accessible.

At the time of my Thanksgiving dinner back in the 1990s, gathering literary evidence meant reading books, noting "things" (a phallic symbol here, a biblical reference there, a stylistic flourish, an allusion, and so on) and then interpreting: making sense and arguments out of those observations. * Today, in the age of digital libraries and large-scale book-digitization projects, the nature of the "evidence" available to us has changed, radically. Which is not to say that we should no longer read books looking for, or noting, random "things," but rather to emphasize that massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making. The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random "things" gathered from a few, even "representative," texts. † We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly "uninteresting" texts.

* Yes, a simplification, but close enough to serve as a heady foil in this introductory polemic. Along similar lines, Susan Hockey writes of the "somewhat serendipitous noting of interesting features" (2000, 66).
† When writing of "anecdotal" here, I am not thinking of the use made of anecdote in the new historical tradition that we find expressed in, for example, Greenblatt's "cultural poetics." Rather, I use the word in the sense of "anecdotal evidence": that is, evidence that is atypical, informally gathered, speculative, or purely interpretive, which is to say not empirical. On this point, the type of literary data I am exploring allows me to adopt a fundamentally empirical position. Having said that, there is a place for anecdotal evidence in literary study, and I do not intend here a critique of anecdotalism per se, but rather to simply make a distinction and separation between two types of evidence.
“Strictly speaking,” wrote Russian formalist Juri Tynjanov in 1927, “one cannot study literary phenomena outside of their interrelationships” (1978, 71). Unfortunately for Tynjanov, the multitude of interrelationships far exceeded his ability to study them, especially with close and careful reading as his primary tools. Like it or not, today’s literary-historical scholar can no longer risk being just a close reader: the sheer quantity of available data makes the traditional practice of close reading untenable as an exhaustive or definitive method of evidence gathering. Something important will inevitably be missed. The same argument, however, may be leveled against the macroscale; from thirty thousand feet, something important will inevitably be missed. The two scales of analysis, therefore, should and need to coexist. For this to happen, the literary researcher must embrace new, and largely computational, ways of gathering evidence. Just as we would not expect an economist to generate sound theories about the economy by studying a few consumers or a few businesses, literary scholars cannot be content to read literary history from a canon of a few authors or even several hundred texts. Today’s student of literature must be adept at reading and gathering evidence from individual texts and equally adept at accessing and mining digital-text repositories. And mining here really is the key word in context. Literary scholars must learn to go beyond search. In search we go after a single nugget, carefully panning in the river of prose. At the risk of giving offense to the environmentalists, what is needed now is the literary equivalent of open-pit mining or hydraulicking. We are proficient at electronic search and comfortable searching digital collections for some piece of evidence to support an argument, but the sheer amount of data now available makes search ineffectual as a means of evidence gathering. Close reading, digital searching, will continue to reveal nuggets, while the deeper veins lie buried beneath the mass of gravel layered above. What are required are methods for aggregating and making sense out of both the nuggets and the tailings. Take the case of a scholar conducting research for a hypothetical paper about Melville’s metaphysics. A query for whale in the Google Books library produces 33,338 hits—way too broad. Narrowing the search by entering whale and god results in a more manageable 3,715 hits, including such promising titles as American Literature in Context and Melville’s Quarrel with God. Even if the scholar could further narrow the list to 1,000 books, this is still far too many to read in any practical way. Unless one knows what to look for—say, a quotation only partially remembered—searching for research purposes, as a means of evidence gathering, is not terribly practical.* More interesting, more exciting, than panning for nuggets in digital archives

* In revising this section before publication, I went back to Google Books and discovered that the number of hits for this particular search had grown significantly since I first tested. No doubt readers will find even higher numbers today.