DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 3

Literary Data Mining: A review of Matthew Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013).

Jim Egan <Jim_Egan_at_brown_dot_edu>, Brown University

Abstract

This review finds that Jockers' *Macroanalysis* provides a clear and provocative argument in favor of literary critics' use of data mining in their efforts to understand literary history. The review finds Jockers' case for a blended approach, one that combines data mining with close-reading techniques, compelling, and it finds, in addition, that his claim that such an approach holds the potential to revolutionize literary study to be a fair assessment of the possibities offered by data mining tools and techniques.

A friend recently asked me, with a smile and in the friendliest way possible, “What does a professor of literature actually do? Don’t you just spout off about the real meaning of a great work of literature? Are your books really scholarship or just your own opinions?” He sought me out after his oldest daughter announced that she wanted to get a Ph.D. in English. He had not only heard horror stories about the job market for Humanities PhDs but was also absolutely baffled, even after she explained to him, as to what literature professors published and whether it had any value. 1

When I tell this story to my colleagues, they often refer to the so-called “crisis in the Humanities” as a way to contextualize these questions as part of an attack on the humanities in general. I believe this is, at least to some extent, absolutely correct. Just focusing on English departments (the subsection of the humanities I know best), universities across the country report steep declines in the number of students majoring in English; students who complete a PhD in English face an abysmal job market in which full-time positions are a rare species available to a decreasing number of graduates; and those who get a job at a place where a book is required for tenure face an increasing number of academic presses reluctant to publish their first book because of the plummeting sales of literary studies. My colleagues hear not only the fears of a concerned father but also, and even more powerfully, echoes of what administrators at their universities ask them when they request new faculty positions (if, that is, they are lucky enough to work at a university that is even willing to consider expanding full-time positions in literature departments). 2

The readers of this journal see that the digital humanities, which has a long history dating back at least to the 1940s, has emerged in this crisis as a major topic of conversation among scholars and administrators associated with the humanities. I will forgo offering here my own speculations about why DH has gained the prominence it has amidst the ruins of the institutional humanities. DH occupies, for journalists reporting on the humanities and conference attendees in the humanities, the very similar position of “the next big thing” previously occupied by theory (which, like the term “digital humanities,” I mean not a unified or uniform set of practices, ideas, projects, writings, products, ideologies, methods, etc. but rather as a very imprecise shorthand for a set of practices, etc. that are, for better or worse, grouped together). DH appeals to me for precisely the same reason as theory and, quite frankly, as the questions posed by the friendly parent. DH and theory each ask those interested in the humanities to ask themselves why we do what we do the way we do it. What do we learn when we close-read a literary text? What assumptions about our goals, aims, and values are embedded within the methods of close-reading? What relationship does history bear to literature — what, after all, is the difference between the categories of history and literature, if any, and what ends do these distinctions serve? Indeed, why do we have literature departments at all? 3

Matthew Jockers' *Macroanalysis* asks us — sometimes explicitly, sometimes implicitly — to engage in just this kind of reexamination of how and why literary scholars analyze texts. He makes a clear, engaging, and provocative case for the potential value of data mining for studying literature. Jockers calls the approach that serves as the title of his book "a new methodology" that demands a "new way of thinking about [literary scholars'] object of study" [Jockers 2013, 4]. We are, Jockers tells us, at "a moment of revolution [in] the way we study the literary record" [Jockers 2013, 171], a moment where how — if not why, though this, of course, is always connected to questions of "how" — we study literature is changing in fundamental ways. *Macroanaysis* seeks to buttress Jockers' contention that "the study of literature should be approached not simply as an examination of seminal works but as an examination of an aggregated ecosystem or 'economy' of texts" [Jockers 2013, 32]. Data mining, according to Jockers, will very likely provide "new knowledge" about larger trends in literary history and, in the process, give analysts "a fuller sense of the literary-historical milieu in which a given book exists" [Jockers 2013, 28]. The "data sets" now available to scholars make it possible to understand individual texts in relation to an almost "comprehensive" context of other relevant texts [Jockers 2013, 7]. Far from being a threat to the very soul of literary studies, as some scholars fear, Jockers demonstrates that computational tools can work in tandem with traditional methods of analysis. Jockers argues for what he calls "a blended approach" that puts data mining in conversation with close-reading of individual texts, an approach that, he contends, will allow analysts to "better understand the context in which individual texts exist and thereby better understand those individual texts" [Jockers 2013, 26, 27]. 4

To make the case for the value of data mining in literary study, *Macroanalysis* is divided into three sections: Foundation, Analysis, and Prospects. Foundation provides an overview in four chapters — titled, respectively, "Revolution," "Evidence," "Tradition," and "Macroanalysis" — of the potential value of data mining for literary scholars. "Revolution" differentiates data mining from close-reading, the method he argues has come to dominate literary studies; "Evidence" contrasts the kind of evidentiary material data mining produces to the kind derived from a close, careful reading of a text by an individual analyst; "Analysis" examines data mining in relation to the history of humanities computing as a way of differentiating it from other forms of computational and/or digital approaches; and "Prospects" briefly discusses some of the issues and questions the methodology of macroanalysis seems particularly well-suited to address, especially in contrast to close-reading. In Part II of the book, the Analysis section, Jockers uses data mining tools on a database composed of nineteenth-century Irish-American fiction to explore questions of Irish-American identity written by writers with Irish roots. If the first section can be said to lay out the theoretical justification for a macroanalytical method, the Analysis section aims to show that, in practice, macroanalysis can, first, yield knowledge about familiar topics that challenge a particular field of literary study's conventional wisdom, a wisdom borne out of earlier models of analysis, and/or, second, generate insights about issues that hadn't occurred to scholars in the field using conventional methods of analysis. Since Jockers focuses here on nineteenth-century American fiction by Irish-American authors about Irish identity, a set of authors and issues about which I have little expertise, I will refrain from judging the success of Jockers' use of data mining in this instance. I do find it significant, though, for reasons I will mention below, that Jockers uses his computational tools to explore four analytical categories, categories which also serve as chapter titles for this section: "Style," "Nationality," "Theme," and "Influence." 5

From my perspective, as a scholar trained in the 1980s, when faculty in the vast majority of literature departments in the United States were entirely ignorant of humanities computing, I find myself excited and energized by the possibilities data mining offers. 6

What about data mining prompts my excitement? First, I think that we don't know what we don't know. While it might end up being true that data mining will provide us exclusively and absolutely only with insights and information that we already knew, I find this possibility to be exceptionally unlikely. Second, as Jockers points out so well in various places in *Macroanalysis*, data mining will allow analysts to read any single work from a large corpus in the context of every work within that corpus in a way that has, prior to this, been virtually impossible. Though Jockers focuses more on the big picture of literary history here, the implications of data mining for our understanding of individual texts will be, among other things, that many, many, many new readings of canonical and non-canonical works will be possible. Who knows what such new readings will produce, but I am willing to wager a great deal that out of the many new readings borne out of data mining at least some — and all one needs is one — will help us see an individual work of literature (and, as a 7 part of the whole of literature, all of literature as well) in a new way. Jockers is certainly right when he says that no one scholar can process all the published writing of the fields in which we tend to divide literary studies. Even if one could read every novel published in the United States between 1800 and 1899, for instance, the human mind just can't process the material — can't, as it were, hold it all in one's head at the same time in the way a computer can. To be sure, the human mind can do many things a computer cannot, but processing enormous bodies of data isn't one of those things.

As I noted above, Jockers speaks of the turn to data mining as a revolution. I think the word "revolution" is well chosen. Trends, relationships, patterns, and meanings in the histories of literatures across the world that had previously been invisible will come into view when we unleash the many digital processing tools on the bits of data that are now readily available. I think it is safe to say that these tools will, taken as a whole, help us see anew the various literary histories we study. Revolutions can occur, at least in some instances, when one sees a familiar object or set of objects in a fundamentally new way. So, it seems to me, if macroanalysis or even other methods that grow out of data mining do, as I think they will, produce new insights about familiar literary histories and new ways of understanding literature — what it is, what it does, how it should be understood, what it's value is, etc. — it will bring about revolutions in the way we conceptualize, read, study, and teach literary works. Perhaps even more exciting, by integrating explicitly computational methods into literary analysis and hiring people who have extensive computational skills in literature departments, we will, I think, change the nature of literary study by broadening the way it can be processed and understood. Works that advocate the use of data mining and other computational methods such as *Macroanalysis* demonstrate the need for at least some literature scholars to receive training as programmers, statisticians, and data visualization specialists. By viewing literature through such new lenses, literary study will remain, I believe, literary study — just as it has while incorporating insights from psychology, history, book production, neuroscience, and other fields — but it will be even more inclusive and extensive in the tools it brings to bear on the literary object. 8

What Jockers and I find exciting and potentially revolutionary about data mining often do not coincide, though. I found that Jockers' decision to focus on Style, Nationality, Theme, and Influence in his database drained the revolutionary potential out of data mining. From my perspective as a scholar of American literature trained in the 1980s, the categories Jockers choose seemed more counter-revolutionary than revolutionary. Each of these categories is quite old and familiar in the study of American literature. They constitute, in fact, foundational categories. The professional study of American literature in the academy can be dated to the 1920s and 30s. The flagship journal in the field, *American Literature*, lists March 1929 as the date of its first issue, and the journal's first few years include many essays devoted to questions of theme, style, nationality, and influence. My quick search of essay titles from the first five years of *American Literature* just focused on influence, and I found plenty, including "The Influence of European Ideas in Nineteenth-Century America"; "The Influence of Edgar Allan Poe into Ambrose Bierce"; "The Influence of Milton on Colonial American Poetry"; and "The Influence of Persian Poetry upon Emerson's Work." 9

On the one hand, the categories Jockers chose are foundational for a reason. They are, well, foundational. Sticking again with influence, I found it disappointing that Jockers chose not to engage with some of the most recent works that offer rather complex analyses, theories, meditations, etc. on the very notion of influence and what its various forms might look like. He cites Harold Bloom's *Anxiety of Influence*, which was published more than forty years ago (in 1973). Jockers, of course, gets to write his own book, but I think an engagement with the way "influence" — and style, theme, and nationality — has been discussed by writers subsequent to Bloom, writers who were/are part of the so-called theory revolution in literary studies, would have added greatly to Macroanalysis' contribution to the current dialogue about the digital humanities. This revolution happened immediately after Bloom's book and, to some extent, managed to remain the dominant way to read literature in the top journals in the discipline. At Brown, and many other peer universities across the United States, the theory revolution so transformed the way literature was studied and taught in English departments that few of my colleagues mention theme, style, or influence in their published writings (and, if they do, their use of these terms would be almost unrecognizable to the authors of the essays from *American Literature* I list above). For better or for worse, style, theme, and influence are rarely discussed at any length in courses taught by faculty here, and I include myself in this group. Literature is often not taught in terms of themes, influence, or style (for the most part) but rather in terms of problems, issues, questions, enabling contradictions, and ideologies. 10

I have no idea whether Jockers wants his focus on such rather old-school categories understood in ways that predate the theory revolution to be part of a counter-revolution that will take back literary studies from the theorists. I do not find enough evidence in *Macroanalysis* to support the conclusion that Jockers argues for the more “objective” method of data mining in order to displace the admittedly less “objective” analyses one finds in works steeped in literary theory — and, quite frankly, it means little to me whether he intends the book to be part of a move to wrest literary studies away from the theorists. Influence, theme, nationality, and style remain important categories in literary study, in spite of the changes wrought in the field over the last fifty years, and they deserve to be studied in a variety of different ways. My point here is simply that, from my perspective, the provocative nature of data mining lies precisely in its potential to help us think about literature in “traditional” and “nontraditional” ways. Jockers uses macroanalysis to learn about individual authors in relation to the economy of nineteenth-century American literature. I want to use data mining to learn about the economy of nineteenth-century American literature without recourse to individual authors. Jockers focuses on the potential of data mining to reveal things about individual authors and large movements in literary history. I am interested in how data mining might reveal trends that exceed the control of individual authors. In many respects, Jockers and I are not so far apart. He speculates that after much research we might find that literary trends obey an “evolutionary” model — a speculation, whatever its plausibility, and I admit I am skeptical — that puts the question of individual human agency at the forefront [Jockers 2013, 155–156]. Who is driving the evolutionary steps: individual authors or forces that are beyond the control of individuals? 11

In spite of these differences in emphases, I found *Macroanalysis* to offer a clear and engaging argument for the possibilities of data mining for literary scholars. I hope many scholars take up his call. Literary scholarship will benefit immensely from the use of big data in its analyses of individual works and massive histories. We need books like *Macroanalysis* to offer ways of processing and incorporating and using the new forms of data scholars will be using. While I might differ from Jockers on just how to parse the data, these differences are precisely the kind that have animated literary scholarship since its beginnings in the nineteenth century. 12

Bring on the revolution. 13

Works Cited

Jockers 2013 Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History* Urbana: University of Illinois Press, 2013.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.