Distant Reading and Recent Intellectual History

Ted Underwood

I love the phrase "distant reading." It's vivid, it doesn't overemphasize technology, and it candidly admits that new methods are mainly useful at larger scales of analysis. It's how I describe what I do. But the phrase does have two disadvantages.

First, since "distant reading" was coined by Franco Moretti on or around the year 2000, the phrase may seem to name a completely new project.[1] In fact, as Katherine Bode has noted, the questions posed by distant readers are often continuous with the older tradition of book history (*Reading*); as Jim English has noted, they are also continuous with the sociology of literature ("Everywhere").

The second disadvantage of the phrase "distant reading" is more serious. By defining a new mode of "reading," the phrase suggests to some that this project is still contained in literary studies — just another stage of our debate about the right way to interpret literature. That assumption has made conversation on the topic needlessly parochial and polemical. We have spent too much time on inward-looking debates that pit distant against close reading, and not enough time understanding connections to other disciplines.

Distant reading is better understood as part of a broad intellectual shift that has also been transforming the social sciences. The best-publicized part of this shared story is an increase in the sheer availability of data, mediated by the Internet and digital libraries. Because changes of scale are easy to describe, journalists often stop here — reducing recent intellectual history to the buzzword "big data." The more interesting part of the story is philosophical rather than technical, and involves what Leo Breiman, fifteen years ago, called a new "culture"

1 See Moretti's reflection on the origin of the term in *Distant Reading,* 43–44.

of statistical modeling (Breiman). The conceptual premises informing models may at first seem arcane, but they're playing a crucial role behind the scenes: this is the fundamental reason why disciplines that used to seem remote from humanists are now working with us on shared problems.

In the twentieth century, the difficulty of representing unstructured text divided the quantitative social sciences from the humanities. Sociologists could use numbers to understand social mobility or inequality, but they had a hard time connecting those equations to the larger and richer domain of human discourse. Over the last twenty years, that barrier has fallen. A theory of learning that emphasizes generalization has shown researchers how to train models that have thousands of variables without creating the false precision called "overfitting."[2] That conceptual advance would be interesting in itself. But it also allows researchers to include qualitative evidence like text in a quantitative model by the simple expedient of using lots of variables (say, one for each word). Social scientists can now connect structured social evidence to loosely structured texts or images or sounds, and they're discovering that this connection opens up fascinating questions.[3]

Humanists are discovering the same thing. Distant reading may have begun with familiar forms of counting akin to book history. (How many novels were published in 1850?) But much of the momentum it acquired over the last decade came from the same representational strategies that are transforming social science. Instead of simply counting words or volumes, distant readers increasingly treat writing as a field of relations to be modeled, using equations that connect

2 The field of machine learning is actually founded on a theory of learning. Specific new algorithms have mattered less than the general implications of this theory—for instance, that there is a tradeoff between bias and variance, and that models should ideally be tested on out-of-sample evidence (Breiman).

3 A brief survey of computational social science can be found in O'Connor, Bamman, and Smith; see also Wallach.

linguistic variables to social ones.[4] Once we grasp how this story fits into the larger intellectual history of our time, it no longer makes much sense to frame it as a debate *within* literary studies. The change we are experiencing is precisely that quantitative and qualitative evidence are becoming easier to combine, blurring disciplinary boundaries. We're working on a methodological continuum now that extends from history and literature through linguistics and sociology. Scholars are still free to specialize in parts of the continuum, of course, and specialization is still valuable. But nothing prevents us from ranging more widely. Since human affairs are also a continuum, we should feel free to use whatever mixture of methods gives us leverage on a particular problem.

Although distant readers are still a tiny minority in literary studies, they receive admonitions from all corners of the field (Spivak 107-9; Marche). Much of this boils down to gatekeeping, and it is rarely informed by a clear understanding of the thing that is to be kept out. We are often warned about "big data," for instance, because the term is new, terrifying, and so poorly defined that it can signify a wide range of threats. But the substantive methodological changes that have actually created new disciplinary connections are rarely mentioned. Conversation of this kind amounts to an empty contest of slogans between the humanities and social sciences, and I think Thomas Piketty spends the right amount of time on those contests: "Disciplinary disputes and turf wars are of little or no importance" (*Capital,* 33).

Recent debates may also tend to overstate the technical challenges of interdisciplinarity. Distant readers admittedly enjoy discussing new unsupervised algorithms that are hard to interpret.[5] But many useful methods are supervised, comparatively straightforward, and have been in social-science courses for

---

4 Supervised models often use linguistic evidence to predict a social variable. For differences of literary prestige, see Underwood and Sellers. For genre, gender, and nationality, see Jockers.

5 Although topic modeling is slippery in a way humanists find fun to argue about, I don't believe it's actually paradigmatic of new methods. If you like fun arguments, however, compare Liu ("Meaning of the Digital Humanities") to Goldstone and Underwood ("Quiet Transformations").

decades. A grad student could do a lot of damage to received ideas with a thousand novels, manually gathered metadata, and logistic regression.

What really matter, I think, are not new tools but three general principles. First, a negative principle: there's simply a lot we don't know about literary history above the scale of (say) a hundred volumes. We've become so used to ignorance at this scale, and so good at bluffing our way around it, that we tend to overestimate our actual knowledge.[6] Second, the theoretical foundation for macroscopic research isn't something we have to invent from scratch; we can learn a lot from computational social science. (The notion of a statistical model, for instance, is a good place to start.) The third thing that matters, of course, is getting at the texts themselves, on a scale that can generate new perspectives. This is probably where our collaborative energies could most fruitfully be focused. The tools we're going to need are not usually specific to the humanities. But the corpora often are.

6 For an illuminating parable about this problem, see Lincoln, "Confabulation in the Humanities."

Bibliography

Bode, Katherine. *Reading by Numbers.* London: Anthem, 2014.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199–231.

English, James F. "Everywhere and Nowhere: The Sociology of Literature after 'The Sociology of Literature.'" *New Literary History* 41, no. 2 (2010): v–xxiii.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Study: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45, no. 3 (2014): 359–84.

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

Lincoln, Matthew. "Confabulation in the Humanities." Author's blog, March 21 2015. http://matthewlincoln.net/2015/03/21/confabulation-in-the-humanities.html.

Liu, Alan. "The Meaning of the Digital Humanities." *PMLA* 128 (2013): 409–23.

Marche, Stephen. "Literature Is Not Data: Against Digital Humanities." *Los Angeles Review of Books.* Oct 28, 2012. https://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities

Moretti, Franco. *Distant Reading.* London: Verso, 2013.

O'Connor, Brendan, David Bamman, and Noah Smith. "Computational Text Analysis for Social Science: Model Assumptions and Complexity." *Second Workshop on Computational Social Science and Wisdom of the Crowds (NIPS 2011),* December 2011. http://brenocon.com/oconnor+bamman+smith.nips2011css.text_analysis.pdf.

Piketty, Thomas. *Capital in the Twenty-First Century.* Translated by Arthur Goldhammer. Cambridge, Mass: Harvard University Press, 2014.

Spivak, Gayatri C. *Death of a Discipline.* New York: Columbia University Press, 2003.

Underwood, Ted, and Jordan Sellers. "How Quickly Do Literary Standards Change?" Figshare. http://dx.doi.org/10.6084/m9.figshare.1418394

Wallach, Hanna. "Computational Social Science: Toward a Collaborative Future." *Data Science for Politics, Policy, and Government.* Cambridge: Cambridge University Press, 2015. http://dirichlet.net/pdf/wallach15computational.pdf