theories and methodologies

“All Models Are Wrong”

RICHARD JEAN SO

SEVERAL YEARS AGO, THE FIRST THING I LEARNED IN MY INTRODUCTORY STATISTICS CLASS WAS THE FOLLOWING DECLARATION, WHICH THE instructor had written in capital letters on the blackboard: “all models are wrong.” Models are statistical, graphic, or physical objects, and their primary quality is that they can be manipulated. Scientists and social scientists use them to think about the social or natural worlds and to represent those worlds in a simplified manner. Statistical models, which dominate the social sciences, particularly in economics, are typically equations with response and predictor variables. Specifically, a researcher seeks to understand some social phenomenon, such as the relation between students’ scores on a math test and how many hours the students spent preparing for the exam. To predict or describe this relation, the researcher constructs a quantitative model with quantitative inputs (the number of hours each student spent studying) and outputs (each student’s test score). The researcher hopes that the number of hours a student spent preparing for the exam will correlate with the student’s score. If it does, this quantified relation can help describe the overall dynamics of test taking.

Literary scholars have long cast a suspicious and critical gaze toward modeling, which strikes them as offensively simpleminded and naive: models run counter to the deep and intensive reading that literary critics take pride in, the exposing of nuance and singularity in texts, writers, and human beings. What about gender? What about race? Don’t they influence how well a student does on a test? And even if you could quantify gender and race and add them to the model, there are always additional dimensions of identity and experience to include. In the end, an individual exceeds socially constructed identity categories, so what does a model do besides reify such categories? Historians of finance, such as Mary Poovey and Donald MacKenzie, have provided many examples of how modeling imperils the social world; for instance, when economists fixate on the model as a tool for reasoning, they conflate the model’s internal logic with the logic of the social world, assuming—in a frightful inversion

RICHARD JEAN SO is assistant professor of English and cultural analytics at McGill University. His recent articles, both written with Hoyt Long, are “Turbulent Flow: A Computational Model of World Literature,” in *MLQ*, and “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning,” in *Critical Inquiry*.

PMLA 132.3 (2017), published by the Modern Language Association of America

of what is real and what is represented—that what holds true for the model must hold true for the world. And today we know the costs of such fallacies: the 2008 economic crisis was in part caused by unrealistic market assumptions built into financial models.

Scholars in literature departments often regard statistics and economics as hubristic endeavors. And so the radical modesty of the aphorism "All models are wrong" might be disarming, alien to those scholars' prior views of the fields. Despite the frequent misuse of models in applied fields like finance, this modesty is a strong feature of statistics. George E. P. Box, a prominent statistician, famously coined the aphorism in 1976 to sketch out a methodology for statistics. He assumes that models must always be wrong; they are just numbers, and numbers, which are simplifying and coarse, cannot represent the complexity of the social and natural world. Yet a model allows the researcher to isolate aspects of an interesting phenomenon, and in discovering certain properties of such aspects, the researcher can continue revising the model to identify additional properties. In this "iterative process," the "truth" of that phenomenon resides at some asymptomatic point that can never be reached. But along the way, the modeling process yields productive insights. Box qualifies his slogan: "All models are wrong, but some are useful" (201). For him, models are recursive mechanisms for generative exploration.

The consensus is that in *Di* -*i* Franco Moretti has made a major methodological intervention into literary studies. But there is disagreement—or, rather, confusion—about what he has done. Accounts run the gamut: he has quantified novels or introduced computers into the analysis of literature or brought science to bear on aesthetic forms. The diverse rhetoric that now surrounds his work underscores this confusion: "algorithms" (Galloway), "quantification" (English), "data" (English; Fitzpatrick; Galloway), or simply "bean counting" (Galloway). In responses to *Di* for or against—keywords proliferate. To assess his work, we need a more precise account of what he has done, and I suggest that *Di* is innovative because it has introduced not merely "data" or "algorithms" into literary studies but also, and more significant, quantitative modeling as a form of reasoning and analysis. This is not just a pedantic point. Once we clarify his intervention, we can better assess its limits, particularly its relation to error, or being wrong, as well as its achievements.

When reading critiques of *Di* -*i* (and, more generally, research in literary studies that uses quantitative and computational methods), we often encounter a desire to find error in the work and to demand correction. For example, if Moretti has used such and such novels to represent a literary tradition and found such and such historical trends, why didn't he also include other novels (Freedman)? This type of critique comes from a zero-sum mentality in how literary scholars often evaluate arguments: either you buy a reading or you don't, and if you don't, that interpretation needs to be displaced or ignored. This tendency is heightened regarding scholarship in digital humanities. Some scholars feel that digital humanists, with their computers and numbers, seek to assert ineluctable facts about literature that one must yield to or resist—for example, that by 1800 "long titles" for novels disappear (Moretti 184). The outcome is a self-consuming cycle. On the one hand, critics of digital humanities imagine computational researchers as hubristically bringing forth empirical knowledge that one must accede to because it has been produced by calculation. On the other hand, computational researchers imagine critics of digital humanities as ungenerous readers who are obsessed with finding error and who, when they do, demand a public confession of sorts, behavior that echoes moralizing eighteenth-century scholarly practices of correction (Lerer 19).

In this back-and-forth, much of the frustration, shared by advocates and opponents of digital humanities, rests on a misunderstanding of how best to use numbers for the analysis of literature. For example, in “Style, Inc.: Reflections on 7,000 Titles,” Moretti analyzes the titles of a large corpus of British novels published between 1740 and 1850. In his first analysis of these data, he computes the mean and median of the length of titles for novels by year. So, for instance, the median length of such titles in 1740 is approximately nine words, and the mean length is approximately twenty-seven (183). He computes these values for every year between 1740 and 1850 and produces a scatterplot in which they are placed in a two-dimensional space and in which the x-axis represents the year and the y-axis represents the number of words. He then eyeballs this scatterplot and discerns some obvious trends. He finds that both the mean length and the median length of British book titles decrease as we move from 1740 to 1850. Book titles are getting shorter, and from this empirical observation, he makes a number of compelling historical inferences that facilitate further analysis.

What he has done is produce a statistical *description* of his corpus; what he has not done is produce a statistical *model* of this corpus. The distinction is that medians and means summarize the novels in his large corpus, giving us a descriptive overview of what is in it. And once we have such quantitative summaries, we can visualize the results, as Moretti does, and see that they point to apparent trends. These results don’t answer some questions, though: How strong is this upward or downward trend? How confident are we that this trend is significant and not just noise? Even though the points in his scatterplot appear to move downward over time, isn’t it possible that this trend is random or trivial? How can we be sure?

A statistical model allows a more nuanced analysis of the data. As Mary Morgan has argued, statistical models are not just summaries of or reports about data, they are mechanisms with which individuals reason and think (31). Here, for example, we could use a common statistical model called regression to better understand Moretti’s corpus. We start where his scatterplot leaves off and try to discern whether there is a true trend in this data—upward or downward. If a perfect linear trend exists, we could put all the points on a straight line. Each point would fit on, say, a flat line, in which case the trend is neither upward nor downward; or we could put each point on a diagonal line that increases by one on both the x- and y-axes, in which case the trend is perfectly increasing by a value of one. Each time we move forward by one year, the number of words that appear in book titles increases, on average, by a value of one. Few data sets have such clean trends. Nearly always, each point will be above or below that line. The more distant the points are from that imagined line, the less confident we are that there is a real trend. And vice versa. We can think of that line as the platonic ideal of a trend, and the distance between the points and the line signals the strength or weakness of that trend.

The value of this method is that it significantly expands the terrain by which the person who has made the model and the person who wishes to evaluate the model can judge the results of the model. Specifically, both parties can focus on the question of error. For example, when a researcher constructs a regression model, a crucial feature of that model (which, again, is a mathematical equation) is its error term, often written as ε. This term (pronounced “epsilon”) measures how well one’s model describes the data. In our example, we can think of it as the sum of the distances of all the points in our scatterplot from our ideal line. This value quantifies the strength of the trend in our data, if one exists.

A large error term indicates that most of the points in our scatterplot fall far from our platonic ideal of the trend, a small one that most of the points are on or near that ideal line. For this reason, ε is also called a disturbance term. It captures the degree to which the reality of our data disturbs the model's ideal imagining and visualization of a trend—a straight line.

Much of the pushback against *Distant Reading* comes from a resistance to its perceived crude empiricism. Moretti has collected a lot of data, run many statistical tests on this data, and, finally, produced a series of scatterplots and bar charts. And the reader must accept such statistical representations as documenting some empirical underlying truth of literary history or, at least, the version of literary history instantiated by his corpus. What's left to debate is the significance of the results: what does the density of gothic-novel titles in the form "the x of y" mean for the evolution of the gothic novel in the nineteenth century compared to that of other genres (Moretti 207)?

Statistics don't quite work this way, though. The advantage of statistical modeling is that it does not present cut-and-dried results that one accepts or rejects. Built into the modeling process is a self-reflexive account of what the model has sought to measure and the limitations of its ability to produce such a measurement. Again, as Box reminds us, "all models are wrong." What's important is not to insist on how the model is right or nearly right but rather to understand how it is wrong. This approach invites the reader to think through with the analyst the mediating figure of the model. Rather than see the model as a potential antagonist, an entity that has brought forth intractable empirical truths that one must accept or reject, the reader is encouraged to reason with the model and its maker to better grasp the data.

We could thus imagine a different version of "Style, Inc." The virtue of applying a regression model to Moretti's data is that we can determine if there is a downward trend in the length of titles over time and, if so, how strong the trend is. But more important, the model allows us to discern the amount of disturbance in that trend. And in locating such error, we can identify the outliers: novels that account for the greatest amount of disturbance. Spotting and then reading outliers help us understand how the model is not working; then by identifying the distinctive features of such texts, we can rebuild our model to account for those features. To his credit, Moretti continuously explores his data and formulates new questions from each set of results, a method that leads to new analyses. But the process, largely animated by intuition, could be more systematic: it could consist of understanding the errors inherent in one's modeling of data, then altering the model on the basis of what one has learned from the errors, a procedure that leads to a model that better fits the data.

Perhaps the greatest benefit of using an iterative process is that it pivots between distant and close reading. One can only understand error in a model by analyzing closely the specific texts that induce error; close reading here is inseparable from recursively improving one's model. Critics who fault Moretti for a lack of close reading are unconvincing when they insist that one must close-read simply because that is what literary scholars have always done. There is a natural, necessary link between distant reading as a modeling project and close reading. This approach helps to attenuate the weakest argumentative aspects of "Style, Inc." Quite often, in assessing his statistical analysis, Moretti confirms his results with a statement such as "it makes sense" (196). Intuition here stands in for what could be a more rigorous method of first understanding how one's analysis doesn't make sense and then refining that analysis until it better fits the data at both distant and close scales. His prose is often charming. But in this instance, it asserts his quantitative analysis as a kind of common sense rather than as a reasoning process that achieves precision and meaning through critical inspection and constant revision.

I conclude by presenting an example from my research to help clarify what this method might look like in practice. In a recent essay, Hoyt Long and I describe our attempt to build a language model to identify English-language haiku in a large corpus of modernist poems published in English-language little magazines between 1913 and 1928. We gave a computer a corpus of several hundred poems that we had manually identified as haiku, as well as several hundred that were identified as not haiku. The computer then studied the features of these poems and learned what qualities most distinguish haiku from poems that are not haiku: it might recognize words that haiku are likely to employ or notice that haiku do not use rhyme. Then the computer quantified this information—by calculating, for example, that haiku are three times more likely to use *house* than are poems that are not haiku—and these rules represent our model (257). Finally, on the basis of the information encoded in our language model, the computer could determine whether new, unseen poems are examples of haiku.

We set our model loose on thousands of poems in our modernist corpus to find haiku. We were in search of large-scale patterns for the emergence, growth, and diffusion of the haiku in this period. In terms of chronology, we discovered an ebb and flow to the use of the haiku starting in the early 1910s and ending in the late 1920s, confirming modernist literary history. But, surprisingly, we found that the haiku spread far beyond the elite groups of poets who initially absorbed and promoted it—poets in Chicago and New York often associated with the imagist movement in the late 1910s—to penetrate unexpected milieus, such as coteries of writers on the West Coast.

The value of our language model is that it allows us to sketch out a new history of the English-language haiku in modernist poetry as practiced in the United States, but the question of error hangs over that history. The computer will incorrectly guess whether a poem is a haiku about 12% of the time, a nontrivial percentage. This problem troubled us, so we closely examined (close-read) the poems that the machine had misclassified as haiku. These texts had certain consistent textual features, which gave us ideas for how to alter and improve our model, leading to better results.

Moretti is not afraid of being wrong. He has welcomed criticism and been generous to his interlocutors. Recently, he has revealed a keen interest in experimentation and trial and error in a series of Stanford Literary Lab pamphlets (e.g., Allison et al.). This generosity, curiosity, and openness to intellectual play are the distinguishing characteristics of *Distant Reading* and have become an intellectual model for the next generation of digital humanists, me included. The purpose of my response is not to expose error and demand correction in his work; rather, it is to argue that error is a constitutive part of science and that quantitative literary criticism would benefit from viewing error as less something to be tolerated or avoided and more something to be integrated formally into our research (Wimsatt). Accepting that all models are wrong might prove liberating.

Works Cited

Allison, Sarah, et al. "Style at the Scale of the Sentence." *Stanford Literary Lab*, pamphlet 5, June 2013, litlab .stanford.edu/pamphlets/. PDF download.

Box, George E. P. "Robustness in the Strategy of Scientific Model Building." *Robustness in Statistics*, edited by R. L. Launer and G. N. Wilkinson, Academic Press, 1976, pp. 201–36.

English, James F. "Morettian Picaresque." *Franco Moretti's* Distant Reading*: A Symposium*, 27 June 2013. *Los Angeles Review of Books*, lareviewofbooks.org/article/franco-morettis-distant-reading-a-symposium.

Fitzpatrick, Kathleen. "The Ends of Big Data." *Franco Moretti's* Distant Reading*: A Symposium*, 27 June 2013.

Los Angeles Review of Books, lareviewofbooks.org/article/franco-morettis-distant-reading-a-symposium.

Freedman, Jonathan. "After Close Reading." *The New Rambler*, 13 Apr. 2015, newramblerreview.com/book-reviews/literary-studies/after-close-reading.

Galloway, Alexander R. "Everything Is Computation." *Franco Moretti's* Distant Reading*: A Symposium*, 27 June 2013. *Los Angeles Review of Books*, lareviewofbooks.org/article/franco-morettis-distant-reading-a-symposium.

Lerer, Seth. *Error and the Academic Self*. Columbia UP, 2002.

Long, Hoyt, and Richard Jean So. "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry*, vol. 42, no. 2, Winter 2016, pp. 235–67.

MacKenzie, Donald. *An Engine, Not a Camera: How Financial Models Shape Markets*. MIT P, 2008.

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

Morgan, Mary. *The World in the Model: How Economists Work and Think*. Cambridge UP, 2012.

Poovey, Mary. "On 'the Limits to Financialization.'" *Dialogues in Human Geography*, 10 July 2015, journals.sagepub.com/toc/dhga/5/2.

Wimsatt, W. C. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard UP, 2007.