theories and methodologies

Think Small: On Literary Modeling

ANDREW PIPER

LITERARY STUDIES CONTINUES TO HAVE A PENCHANT FOR GREAT MEN. IN 2015, FOR EXAMPLE, 20% OF AUTHORS LISTED AS SUBJECTS IN the *MLA International Bibliography* accounted for just under 60% of all articles or book chapters published that year.[1] Just the top 1% of authors, or 33 in total, accounted for 1,302 works, or 20.8% of the total. Four of these authors were women, and one was not white (W. E. B. Du Bois). Those numbers are even slightly more concentrated than in 1970, when 1% of authors accounted for 15.9% of all articles and book chapters. In that year, only one of the most frequently mentioned authors was a woman (George Eliot), and all were white.

Male proper names continue to serve, in other words, as convenient metonyms in our field for larger aesthetic or methodological frameworks. They are far-reaching vehicles of particularization and generalization, of how we zoom in to zoom out. One of the most salient contributions of Franco Moretti's work has undoubtedly been its ability to call into question this reliance on nominalization, to introduce the question of scale, and the means through which we might address it, as one of the central theoretical concerns of our time.[2] It is thus ironic, even if not altogether surprising, that a special feature devoted to the study of distant reading would choose to frame itself through the figure of the single author.

In this essay, I use the idea of the literary model to introduce a new way of thinking about traversing scales of critical analysis. Rather than rely on proper names as placeholders or on the visual icons of graphs, maps, or trees, models return us to the process—the tools, techniques, and practices—through which we construct our knowledge of phenomena that exceed our direct observation. Much of the early discourse surrounding the computational understanding of literature has inevitably focused on notions of distance or bigness, on a vocabulary of transcendence or the macrocosm. Like computing culture (Davis) or literary studies (Wellmon) during their beginnings, the nascent field of data-driven literary studies has prioritized a sense of communion with something greater than ourselves.

ANDREW PIPER is professor and William Dawson Scholar in the Department of Languages, Literatures, and Cultures at McGill University. He is director of .txtLAB, a cultural analytics laboratory, and author of the forthcoming book *Enumerations: The Quantities of Literature* (U of Chicago P).

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

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

In thinking about modeling, my aim is to realign our focus on the small ways in which our insights about large numbers of texts or words are mediated. Models represent an important new form of mediation in reading and interpretation, and, like other forms of mediation, their role or agency needs to be better understood.[3] The emphasis on bigness and distance overlooks the minutiae that stand between us and these larger scales. Focusing on models, thinking small to think big, moves us away from a sense of communion and ultimately toward one of craft. It helps draw attention to the constructed nature of knowledge.

A great deal has been written in the history and philosophy of science about the role modeling plays in knowledge creation. This literature should become increasingly integral to our field. Much of the writing has focused on how a model represents, rather than indexically stands for, some real-world phenomenon (Hacking; Hughes; Giere). "The characteristic—perhaps the only characteristic—that all theoretical models have in common," writes Richard Hughes, "is that they provide representations of parts of the world" (S325). By focusing on the representational qualities of models, philosophers of science have encouraged us to move past a form of empiricism that asserts an unproblematic relation between data and the world. Instead, they push us to look at one of the core concerns of our discipline, that of representation, the activities of construction and creativity that are involved in the process of understanding, the way models stand for something but are not to be confused with the thing itself. We could say, using terminology closer to home, that models shift the focus toward the signifiers of research and away from the signifieds.

Such work has drawn explicit connections to the early-twentieth-century philosophy of Hans Vaihinger and his emphasis on the "as if" (*als ob*), the role of fictionalizing in the act of knowing. As Arthur Fine writes in his work resurrecting Vaihinger, "He finds no realm of human activities, even the most serious of them, into which play and imagination fail to enter. These faculties are part of the way we think ('constructively'), approach social and intellectual problems ('imaginatively'), employ metaphor and analogy in our language, and relate to others" (16). Fictionality, for Vaihinger and his intellectual descendants, becomes a core part of epistemology.

Subsequent work has increasingly focused on the nature of models' representations, the extent to which models should be understood either isomorphically—that is, as mimetic representations of the world, as in a Newtonian model of planetary behavior that approximates planets by spheres—or more informationally, as in the metaphor of the map (Hunter; Bailer-Jones; Suárez; Chakravartty). Models can more or less faithfully represent the world; the issue at stake is not realism per se but the "surrogative reasoning" that models enable (Contessa). As Gabriele Contessa writes, "Faithful epistemic representation is a matter of degree. A vehicle does not need to be a completely faithful representation of its target in order to be an epistemic representation of it" (55). The important issue is the extent to which a model allows a given user to make "valid inferences" (54), not the extent to which a model looks like the world it claims to represent.

We can understand literary models as part of this process of surrogative reasoning, an assemblage of externalities that allow mental inferences to be made about meanings in the world that are not readily at hand. Following Contessa, we can think of them as the attempt to limit information loss about the world, to arrive at an appropriate approximation of a given territory (here a set of texts). When we cite a passage, for example, there is almost no information lost between the cited passage and the text (or terrain) from which it is drawn. It is reproduced word for word and thus enables a reader to make inferences from

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

it with considerable ease. This ease of inference is close reading's greatest strength. And yet a tremendous amount of information is lost in all the other aspects of the work that are not cited, as well as through the omissions of the material, social, and linguistic contexts from which the passage is drawn. This loss is close reading's greatest weakness, one that computational methods are well-suited to address. Literary modeling is the attempt to minimize such information loss while maximizing the insights that can be gained about the terrain under review.

Such an informational understanding of modeling, however, relies too much on an indexical relation to the world, where difference is seen only as a matter of degree (as in the sense of a cartographic scale of 1:1000). And yet as philosophers of science have pointed out, modeling also involves a certain amount of heterogeneity in representational practices. As Contessa writes, "The same vehicle can be a faithful representation of some aspects of the target and misrepresent other aspects" (55). In most cases, one model, like one map, cannot account for all aspects of a given set of documents. Models may contradict each other: "The focus on modeling a certain aspect of a phenomenon," writes Daniela Bailer-Jones, "sometimes leads to the acceptance of false propositions that address other aspects that are not the focus of a current model" (68). Diversity in modeling, even contradiction, is a core component of what Contessa calls "epistemic representation."

Instead of foregrounding incompatible differences between models, Roman Frigg has emphasized the diversity of representational approaches within models. Here I think we move still closer to the insights of literary studies as well as its relevance to literary modeling. A narrative representation, for example, consists not only of narratological features like point of view, diegetic levels, or temporal frames but also of linguistic aspects that depend on paradigmatic and syntagmatic dimensions of semantic meaning. Both aspects are important for understanding narrative texts, and yet they require different notions of how representation works.

If we wish to unpack and do literary modeling, then, we will need greater clarity about the different kinds of epistemic representation that models encode to facilitate surrogative reasoning. We need to better understand the different representational practices involved in how we construct our knowledge of social and textual fields. In what follows, I offer five such practices, which we can imagine as layers nested inside one another. Identifying these practices is a first step in constructing a model, if you will, of literary modeling.

The Five Layers of Literary Modeling

Theorization. What is the theoretical goal of the model? What hypothesis does one want to test? In a now-classic example, Ian Lancashire and Graeme Hirst set out to understand the relation between mental illness and literary expression in the work of Agatha Christie, who was diagnosed with dementia late in her life (Lancashire and Hirst). At a theoretical level, Lancashire's model encodes a question about the relation between the notions of illness and creativity. It makes explicit a latent connection between different constellations of discourse in the world. Theorization makes this connection manifest and subjects it to interrogation, testing, and critique. The theoretical level operates at the greatest degree of abstraction from the world to draw new lines of connection in the world.

Conceptualization. What are the conceptual proxies for the hypothesis? How will the overarching theoretical terms, terms whose openness is a condition of their efficacy, be particularized? In Lancashire's example, mental illness is understood as age-related, as a physiological process in time, while creativity is translated into the far more concrete notion of "vocabulary richness." Here we see

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

the underlying process of translation at work, as the more general hypothetical structure is translated into more concrete conceptual terms (*age-related*, *vocabulary richness*). This process of translation is also one of specification, since information is lost with each substitution or "carrying over" of an idea into a more concrete realm.[4] Such specification is driven by existing fields of knowledge—Christie's biography is one source but so too is biomedical research into the human experience of dementia. Models rely on the specifications of other models in a larger representational web.

Implementation. How will these concepts be made actionable—that is, how will they be measured? Here we arrive at the crucial dimension, and also point of contention, in the emerging field of computational criticism (Piper, "There"). Why must we measure? Because it's reductive! We shouldn't shy away from measurement's reductiveness or even reductiveness itself but acknowledge that such reductiveness is a necessary component of all generalization. (We have, after all, already engaged in two levels of reduction before this moment of measurement.) Measurement is no more or less reductive than selecting a passage from a single author and having it stand for all European literature. The antipode to measurement is not subtlety or complexity but personal authority. Measurement replaces charisma as the guiding vehicle of generalization. It diffuses power, away from the persona (the proper name) and into a more dispersed array of technologies, techniques, and practices among which the individual is enmeshed (Latour). Unlike the power of personal authority, of the proper name, the power of measurement can be made explicit (though of course one can also try to hide its explicitness). Measurement, in this sense, is a form of explication. In Lancashire and Hirst's work, the concept of vocabulary richness is explained by the measurements of "type-token ratio," "repeating phrases," and "indefinite words" (somehow, anyone, etc.).[5] These measurements can be demonstrated—and disputed or complemented. It would be naive to suggest that measurements are devoid of power relations, but neither are the acts through which proper names account for knowledge. Models make us aware of the externalities through which we arrive at truth claims.

Selection. At this point the model needs to be applied to some terrain (data), a process that involves still further reduction. The approximation of the world through the model is applied to another approximation of the world through data. Each of these approximations represents some underlying phenomenon, and each involves distortions, like laying two maps on top of each other. What data are appropriate for answering the question that my model poses? And what model is the most appropriate given the data I have selected? Lancashire and Hirst use the first fifty thousand words of fourteen novels by Christie out of a total of eighty-five. Here the writer, a selection from the pool of all writers, has her work approximated by a selection of all works, which are approximated by a selection of a portion of each work. Sometimes selection can lose too much information to be useful. Other times, through curation, it can ensure a better representation of the world we are trying to model. More is not always better.

Validation. How do we know that the inferences we draw from our model are valid? This question is far more challenging than it might seem at first. As philosophers of science have pointed out, a large amount of implicit knowledge is at work when scholars interpret models (Contessa; Polanyi). Sometimes the interpretation of models can take the form of a statistical test, when we attempt to measure the extent to which what we are observing exceeds the bounds of chance. Lancashire and Hirst use regression analysis to assess the relation between age and vocabulary richness. The better the fit—that is, the more the assumptions of the test approximate the observed

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

data (that vocabulary richness declines in a linear fashion)—the more valid the inferences that can be drawn from the test. The selection of an appropriate test is a key component of modeling. What this selection cannot answer is whether the model is an appropriate approximation of the phenomenon that one is claiming to observe. Are type-token ratio or indefinite words a good representation of vocabulary richness? Is vocabulary richness a good representation of mental illness? We can quickly see how problematic the notion of validation becomes when we think about modeling as representation. As I have argued elsewhere (Piper, "Novel Devotions"), reading remains a core tool of validating whether a model captures the theoretical and conceptual frameworks it is meant to approximate. Close reading reenters distant reading through the layer of validation.

We can see how at every layer in the modeling process we are engaged in some act of representation, whether theorization, conceptualization, measurement, data collection, or even validation. All these representations are different in nature and interact with one another as they allow surrogative reasoning to take place. The point is not to invalidate the exercise as hopelessly biased or distorted—to indulge in the fantasy of the text *an sich* ("by itself")—but instead to acknowledge that these representational dimensions are a core new area of research in computational criticism. We must take seriously the fictionalism that Vaihinger saw as the foundation of reason.

Vulnerability

I conclude with an example drawn from a new project that I have been working on to show this process of modeling in action. The project studies vulnerability in poets' careers, which I define as the moments or periods in poets' lives when their stylistic changes exceed expectations, when they open themselves up to new modes of expression. It is part of a larger effort to think about the author's work as a corpus, as a body of writing over time, one that is not only distinctive in shape but also marked by openings and fissures. Inspired by Edward W. Said's work on "late style," in which Said writes about musicians' and writers' late works, I ask a larger theoretical question concerning the corporality of writing: What is the relation between embodiment, time, and poetic creativity?

To begin, I translate the question of corporality through the concept of vulnerability: the way a body is fundamentally defined not by a sense of distinctiveness (me or not me) but by implication in or entanglement with an environment, by an openness to the world. Said imagines lateness as a state in which a writer or musician challenges or resists, whereas I am interested in the creative state of openness—not a hardening into something but an accessibility born of letting go. I further specify vulnerability as a sense of stylistic excess, as unreasonable change. Vulnerability is marked by a temporal window in which the distinctiveness of one's writing breaks down, where the reasons one may have constructed for one's art no longer appear to operate.

In the third layer (implementation), I choose two different measurements to represent my model. The first constructs the poet's career as a network in which poems are connected to those they are most similar to. Drawing on the classic notion of percolation in network science (Newman), I then gradually remove poems (nodes) from the network until the network breaks into two similarly sized components, meaning there are now two distinct smaller networks that are no longer connected to each other.[6] I am asking how vulnerable these global relations are to being fractured into two distinct groups, where the unity of the author begins to look double. The second measure I use looks at a more local level of vulnerability to identify those moments when a given poem's similarity to the

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

rest of the corpus exceeds expectations up to that point (fig. 1).[7] These moments are what I identify as the stylistic openings that capture this second sense of poetic vulnerability.

But what makes two poems more or less similar? This question is not straightforward and requires its own model. In other words, what we have here are nested models, each requiring its own process of representation and validation. I represent similarity between poems using a vector-space model that includes lexical, semantic, phonetic, and syntactic features. Whereas I validate the other measures using tests of statistical significance, here I validate the model by reading outputs of different specifications (giving different weights to these different dimensions). The valid model is the one that produces the most appropriate connections between poems. However much I may try to share this process with readers, appropriateness will always inevitably be based on what Polanyi calls a "tacit dimension" to knowledge. Researchers are always implicated in their models. Taking these models together, I then select my data by observing the vulnerability of seventy-eight poets' careers in a collection of over thirty thousand poems spanning three languages and three centuries.

As can be seen by even this cursory description, the modeling process entails a tremendous amount of subjectivity. But the process is not arbitrary. It allows for much creativity and intellectual inventiveness, far more than I think the discourse surrounding distant reading has initially led us to believe. Researchers' entanglement in modeling has largely been overlooked, and it makes room for many of the values we typically associate with literary interpretation. What Susan Stewart said of miniature things, we can say of models: they are "limited in physical scope yet fantastic in content" (44). Models provide new ground for debate and interpretation but

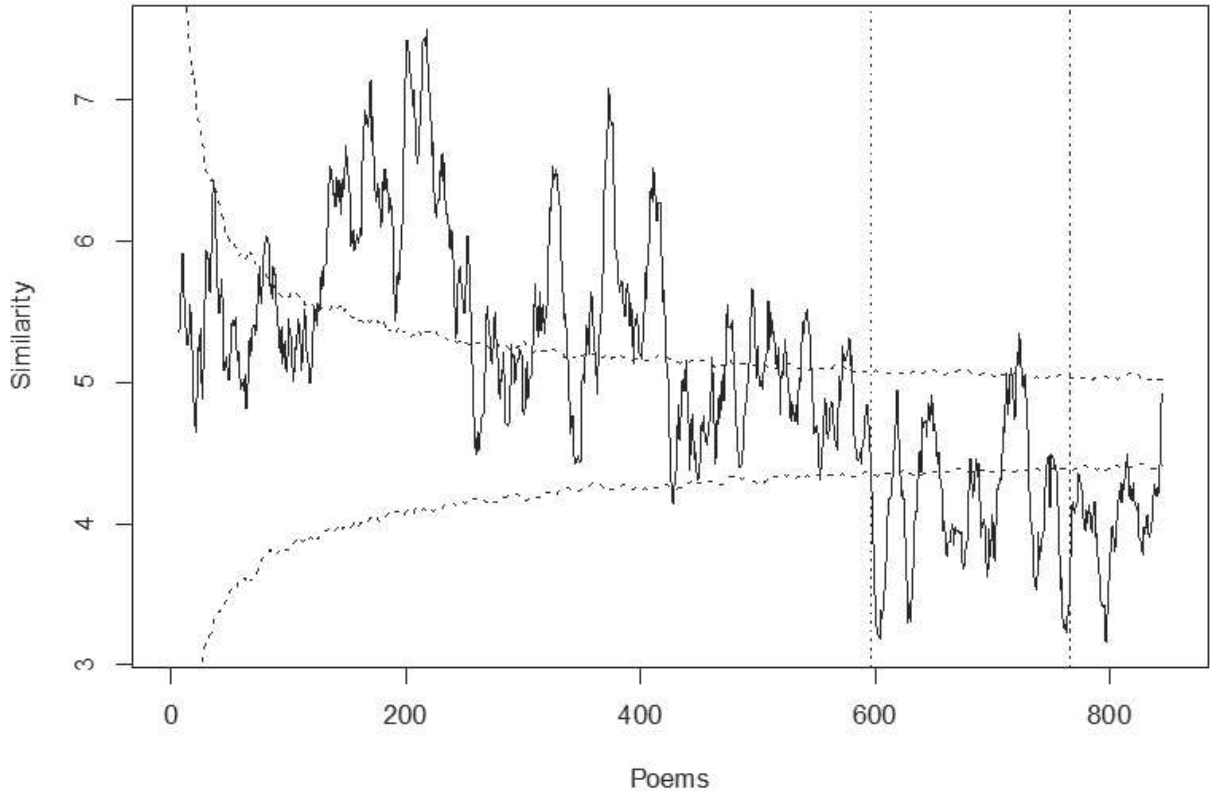


FIG. 1
A representation of the career of Wanda Coleman.

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

also for imagination. In this way, they are similar to existing critical methods.

Models, however, also provide a new basis for consensus and for the architectonic construction of knowledge. Former critical methods depended on an affective allegiance to one's own evidence and a tendency to dispute that of others. Modeling presupposes a more mediated, craft-like relation to evidence. You can enter into my model and I yours. A model allows us to be more vulnerable with our ideas, making knowledge cumulative but also conglomerative. It emphasizes collectivity, departing from our field's historical focus on singularities. Models open the door to new kinds of critical sociability, away from the intimacies of books (Lynch) and toward the more festive nature of the commons. Instead of just book love I hope we can also find model enjoyment.

Not

1. For a description of the collection process and all the code and data, see Piper, Data.

2. See, e.g., the essays collected in English and Underwood.

3. For some recent work that has begun to address modeling in literary studies, see Moretti; Long and So; Underwood; Piper, "Novel Devotions."

4. As B. Schmidt has argued, rather than simply understand algorithms we need to understand the "transformations" that they enact.

5. Type-token ratio compares the number of individual word types to the overall number of words (or tokens) in a given passage. The values range between 0 and 1, where higher scores equal more "richness"—that is, fewer words are repeated less frequently. A score of 1 would mean that no word was ever repeated in a passage.

6. I define the breakpoint of the network (its vulnerability) as the moment when a second component is at least half as large as the largest component during the process of removing nodes (percolation).

7. The horizontal axis represents in chronological order the poems Wanda Coleman wrote over her lifetime, and the vertical axis represents the similarity of a given poem to all other poems in the collection prior to that point. The higher the value, the more similar a poem is to the rest of the previous poems, while the lower the value, the more dissimilar. The horizontal dotted lines represent bands of significance, meaning that the values above or below them occur in less than 1% of all random permutations of her corpus. Poems that fall below the lower line are thus significantly different from the rest of her career up to that point. We can see how with the publication of Coleman's collection *Bathwater Wine* (1998), here marked with the vertical dotted lines, Coleman radically departs from the stylistic orientation that had governed her writing until then. As Coleman writes, "By the end of 1996 everything was in shambles and 32-years of sacrifice to become a writer was tantamount to nothing" (T. Schmidt 134).

Works Cited

Bailer-Jones, Daniela M. "When Scientific Models Represent." *International Studies in the Philosophy of Science*, vol. 17, no.1, 2003, pp. 59–74.

Chakravartty, Anjan. "Informational versus Functional Theories of Scientific Representation." *Synthese*, vol. 172, no. 2, 2010, pp. 197–213.

Contessa, Gabriele. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science*, vol. 74, no. 1, 2007, pp. 48–68.

Davis, Erik. *TechGnosis: Myth, Magic, and Mysticism in the Age of Information*. Harmony Books, 1998.

English, James F., and Ted Underwood, editors. *Scale and Value: New and Digital Approaches to Literary History.* Special issue of *Modern Language Quarterly*, vol. 77, no. 3, 2016.

Fine, Arthur. "Fictionalism." *Midwest Studies in Philosophy*, vol. 18, no. 1, 1993, pp. 1–18.

Frigg, Roman. "Fiction and Scientific Representation." *Beyond Mimesis and Convention: Representation in Art and Science*, edited by Frigg and Matthew Hunter, Springer, 2010, pp. 97–138.

Giere, R. N. "Using Models to Represent Reality." *Model-Based Reasoning in Scientific Discovery*, edited by Lorenzo Magnani et al., Kluwer Academic / Plenum Publishers, 1999, pp. 41–57.

Hacking, Ian. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge UP, 1983.

Hughes, R. I. G. "Models and Representation." *Philosophy of Science*, vol. 64, no. 4, 1997, pp. S325–36.

Hunter, Matthew C. "Experiment, Theory, Representation: Robert Hooke's Material Models." *Beyond Mimesis and Convention: Representation in Art and Science*, edited by Roman Frigg and Hunter, Springer, 2010, pp. 193–219.

Lancashire, Ian, and Graeme Hirst. "Vocabulary Changes in Agatha Christie's Mysteries as an Indication of De-

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press

mentia: A Case Study." Nineteenth Annual Rotman Research Institute Conference, 8–10 Mar. 2009, Intercontinental Centre Hotel, Toronto.

Latour, Bruno. *On the Modern Cult of the Factish Gods*. Translated by Catherine Porter, Duke UP, 2010.

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

Lynch, Deidre Shauna. *Loving Literature: A Cultural History*. U of Chicago P, 2015.

Moretti, Franco. "Operationalizing; or, The Function of Measurement in Literary Theory." *New Left Review*, no. 84, Dec. 2013, pp. 103–20.

Newman, M. E. J. *Networks: An Introduction*. Oxford UP, 2010. *Oxford Scholarship Online*, doi:10.1093/acprof:oso/9780199206650.001.0001.

Piper, Andrew. Data for "Think Small: On Literary Modeling." *Figshare*, 15 Feb. 2017, doi.org/10.6084/m9.figshare.4653913.v1.

———. "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel." *New Literary History*, vol. 46, no. 1, Winter 2015, pp. 63–98.

———. "There Will Be Numbers." *Cultural Analytics*, May 2016, doi: 10.22148/16.006.

Polanyi, Michael. *The Tacit Dimension*. 1966. U of Chicago P, 2009.

Said, Edward W. *On Late Style: Music and Literature against the Grain*. E-book, Pantheon Books, 2006.

Schmidt, Ben. "Do Digital Humanists Need to Understand Algorithms?" *Debates in Digital Humanities*, 2016, dhdebates.gc.cuny.edu/debates/text/99.

Schmidt, Tyler T. "'Womanish' and 'Wily': The Poetry of Wanda Coleman." *Obsidian III: Literature in the African Diaspora*, vol. 6, no. 1, 2005, pp. 120–41.

Stewart, Susan. *On Longing: Narratives of the Miniature, the Gigantic, the Souvenir, and the Collection*. Duke UP, 1993.

Suárez, Mauricio. "Scientific Representation: Against Similarity and Isomorphism." *International Studies in the Philosophy of Science*, vol. 17, no. 3, 2003, pp. 225–44.

Underwood, Ted. "The Life Cycles of Genres." *CA: Journal of Cultural Analytics,* May 2016, culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

Vaihinger, Hans. *The Philosophy of "As If": A System of the Theoretical, Practical and Religious Fictions of Mankind*. Routledge and Kegan Paul, 1965.

Wellmon, Chad. "Sacred Reading from Augustine to the Digital Humanists." *Hedgehog Review*, vol. 17, no. 2, Fall 2015, pp. 70–84.

https://doi.org/10.1632/pmla.2017.132.3.651 Published online by Cambridge University Press