What is Literary Mathematics?

Michael Gavin

University of South Carolina

mgavin@mailbox.sc.edu

What is literary mathematics?

The nascent field of cultural analytics has exposed a gap in literary and historical theory. Scholarship written in this mode goes under various names — distant reading, digital history, humanities computing — but shares a common methodological orientation toward quantification and abstraction, usually for the purpose of exposing large trends or otherwise latent patterns in the historical record. To date, this research has proceeded without any guiding methodological framework nor even much of a shared field of theoretical debate.[1] In the information sciences, the social sciences, psychology, mathematics, and even in physics, scholars have developed mathematical models for the study of cultural phenomena and subjected those models to intense scrutiny.[2] But there has been no similar effort in the humanities, where discussions have been limited almost entirely to reflections about quantification in general and have included very little sustained analysis of any specific approach. We need to develop a theory of literary computation.[3] Such a discourse would ask: What kinds of things are the objects of literary computation? What mathematical forms are appropriate for their description?

---

[1] Underwood (2017) explains that the phrase distant reading "underlines the macroscopic scale of recent literary-historical experiments, without narrowly specifying theoretical presuppositions, methods, or objects of analysis."

[2] For overviews and examples across a variety of fields and methodologies, chosen more or less arbitrarily, see Watts and Strogatz (1998), Barabási and Albert (1999), Epstein (2006), Pachucki and Breiger (2010), Krioukov et al. (2010), Mohr and Bogdanov (2013) and, most recently, Barron et al. (2018).

[3] I am not alone in identifying this need. See So (2017) and Algee-Hewitt (2017). Algee-Hewitt expresses a view close to my own: "Establishing metrics, finding patterns, and linking these metrics and patterns with meaningful concepts of literary criticism: these are tasks that digital humanities now faces" (751).

My goal in this essay is to sketch an outline for future research, not to review the history of humanities computing nor to add another think-piece on DH, but a few words of framing context may be helpful. Although digital humanities has been around now for more than a decade and humanities computing for much longer, the need for a theory of literary computation has appeared only very recently.[4] In English studies, this urgency can be precisely identified with the public release of large-scale textual archives — *Early English Books Online*, HathiTrust, *Chronicling America*, etc. — that make it not just possible but convenient for scholars to do relevant computational research.[5] Although Franco Moretti contrasted distant reading from close reading, for most scholars close reading goes hand-in-hand with historical contextualization.[6] Texts are intuitively chosen for close reading because they exemplify general historical phenomena. Moretti's provocateurship aside, the problem with distant reading wasn't that it violated some sacrosanct interpretive ritual but that, when it came to large trends, scholars lacked the data and the know-how to produce analyses that rivaled their intuition in quality.[7] Only with the advent of large textual collections has it become possible to perform bread-and-butter humanities research computationally because only now do scholars have corpora from which to make adequately trustworthy generalizations about the past, and so only now does computing offer sufficient promise for literary scholarship such that investing time

---

[4] Earlier research in humanities computing was usually focused on the study of relatively small corpora and individual texts or authors. In addition to the overview provided in Hoover (2008), see for example Burrows (1987), McCarty (2005), and Ramsay (2011).

[5] For recent studies that draw from these corpora, see Witmore and Hope (2016), Smith, Cordell, and Mullen (2015) and Underwood, Bamman, and Lee (2018).

[6] See Moretti (2005) and Moretti (2013). Jay (2017) usefully situates distant reading in the context of close reading's vexed relationship with literary history.

[7] The complaint that computational studies merely reinforce intuition is a common theme among critics. Such accusations are not always grounded in careful readings of the work. For example, Timothy Brennan (2017) pretends to summarize what he calls the "elaborate article" by Hoyt Long and Richard Jean So (2016): "After 30 pages of highly technical discussion, the payoff is to tell us that haikus have formal features different from other short poems. We already knew that."

and effort in learning quantitative methods can be justified for people outside the narrow field of humanities computing itself.

The data we now have.[8] The know-how continues to lag behind. What to do with all this information remains a large and complicated question. Several aspects of our critical discourse are at least partly to blame for this lag. Polemical debates over institutional politics surround this work like a social-media hubbub of bad feeling. Whatever the merits of some scholars' critique of "digital humanities" as an institutional construct, those critiques do nothing to clarify matters for any Americanist wondering how to learn from text files available on *Chronicling America*.[9] Even sympathetic discussion frequently points in the wrong directions. Too often, "quantitative methods" are invoked as a monolithic paradigm that marks a division between "the humanities" and "the sciences."[10] But quantification is a broad and diverse category of thought that encompasses many very different theories, from information science and Bayesian statistics, to graph theory, to geospatial modeling. These bodies of research have divergent goals and sometimes conflicting assumptions. What quantitative methods are appropriate in what research situations? The work of sorting through these ideas would be arduous under the best of scenarios, and without a clear-eyed (and shared) understanding of this issue, the difficulty is amplified.

Matters are made worse when digital humanists rely too heavily on black-box software to perform their analyses and on visualization to advance their arguments. Data-analysis software

---

[8] I do not mean to suggest here that the work of data curation is complete. Quite the contrary.
[9] For the most high-profile example of this critique, see Allington, Brouillette, and Golumbia (2016). See also Golumbia (2014).
[10] See Berry (2011) and Liu (2013). A somewhat clearer sense of variety can be found in Manovich (2016). Commentary on the "two cultures" of the humanities and the sciences is too copious and too tedious to cite. Snow (1959) is the common touchstone for such debates.

insulates scholars from the underlying mathematical procedures.[11] Often, this is a good thing. Writing algorithms to parse data can be complicated work, and most critics usually have more important things to do with their time, even most digital humanists. But if the entire field relies wholly on software produced by engineers and scientists for other purposes, we will be forever limited to the interpretive structures they provide. Even if a few eager DHers win grants to build new software tools, they can do little to remedy this problem.[12] Why? Because there are lots of ways to count things, and mathematicians, physicists, and information and social scientists are devising new ways every day. No point-and-click software could ever offer more than the thinnest of slices from this very big pie. The work of building humanities software, though valuable on its own terms, cannot substitute for sustained engagement with quantitative theory.

So we're left in a perverse situation where self-appointed defenders of critique lob sweeping attacks on scholars whose use of quantitative reasoning is actually quite limited and who feel so much pressure to demonstrate findings — "See, you really can learn something with digital methods!" — that few have time to just sit down with the mathematical propositions and think them through. Critics attack the whole project as insidious, tout court. Practitioners defend themselves by writing essays that use digital tools to explore privileged topics. Meanwhile, the underlying theories remain unexplored and unexplained.

I ask again: What quantitative theories are appropriate for studying what literary topics? Stop for a moment to notice how utterly unprepared we are as a profession to answer this question. Adopting it as an informing rubric accomplishes several things simultaneously. First,

---

[11] Ben Schmidt (2016) states the problem this way: "This instrumental approach to software, however, promises us little in the way of understanding; in hoping that algorithms will approximate existing meanings, it in many ways precludes them from creating new ones."

[12] Perhaps the best-known software dedicated to humanities data is *Voyant*, developed by project leads Stéfan Sinclair and Geoffrey Rockwell. See https://voyant-tools.org/. The challenges involved in translating point-and-click software to meaningful insight can be inferred from the paper-thin analyses in their co-authored 2016 study, *Hermeneutica*.

this question de-fetishizes computer technology, visualization, and rhetorical provocation. Second, it opens to literary scholars the entire field of mathematics, which is now available for our application and critique. Third, it exposes how little we know, because few among us have given the question more than passing consideration. Fourth, it deflates the ideological stakes, because confident declarations about the state of the discipline or the nature of interpretation, whether meant to encourage or resist quantitative study, can hardly be compelling in the context of near-total and near-universal ignorance, once that context is openly acknowledged. Lastly, it shifts attention to theoretical problems that invite experiment and critique, suggesting an agenda for research that is wide open with many paths of inquiry already clearly visible.

Such is the rationale for a new practice of criticism I call "literary mathematics." In what follows, I'll highlight what I take to be the key issues this practice raises, and, for scholars interested in pursuing research in this area, I'll sketch out several directions for future work that I believe are most promising. For scholars who don't do digital work, much of what I'll say may seem far afield of their research and pedagogical interests. This, too, is a good thing. It would be neither practical nor desirable for literary scholars *en masse* to put down their novels and pick up algebra textbooks. But, that doesn't let them off the hook. They have a responsibility — as reviewers, as colleagues, and as credentialed experts expressing opinions about research in their fields — to keep abreast of current methods. Like it or not, statistical techniques are now included in that mix. For this reason, I'll conclude with a rubric for evaluating research in cultural analytics, which I hope will prove useful both for my fellow digital humanists and for others tasked with judging our work.

The presumption of quantifiability

Literary mathematics is math applied to literary problems; it is the practice of representing critical concepts using formal expressions that describe relations among

literature's countable features. In this regard, literary mathematics shares a great deal with research called distant reading, algorithmic criticism, or cultural analytics.[13] With distant reading, the goal is to use computation in service of literary history, and mathematical models become instruments for evaluating and describing historical trends.[14] Similarly but in contrast, literary mathematics uses computation in service of literary theory, focusing on the models themselves as newly invented theoretical constructs.

Mostly this a difference of emphasis and mode of presentation. Research in cultural analytics usually does literary math while focusing on other things. Consider, for example, Ted Underwood's and Jordan Sellers's 2012 study of "literary diction" and their visualization of changing word-use patterns over time. (Figure 1.) Working from a corpus of poetry, drama, fiction, and nonfiction, they ask what percentage of words used in each type entered the English language before 1150. Over the course of the eighteenth and nineteenth centuries, they show, literary diction separated from nonliterary diction. To Underwood and Sellers, this analysis exposes the gradual, centuries-long emergence of literature as a social category. Their emphasis is on the historical process of transformation and change, and so it makes sense for them to represent their idea in the form of a time-series graph reflecting familiar large-scale changes.

[13] For the terms "distant reading," "algorithmic criticism," and "cultural analytics" as I'm using them here, see, respectively, Underwood (2017), Ramsay (2011) and Manovich (2016).
[14] By identifying distant reading's primary method with mathematical modeling rather than with visualization, I adopt a view more closely aligned with So (2017) than Moretti (2005) or Moretti (2013).

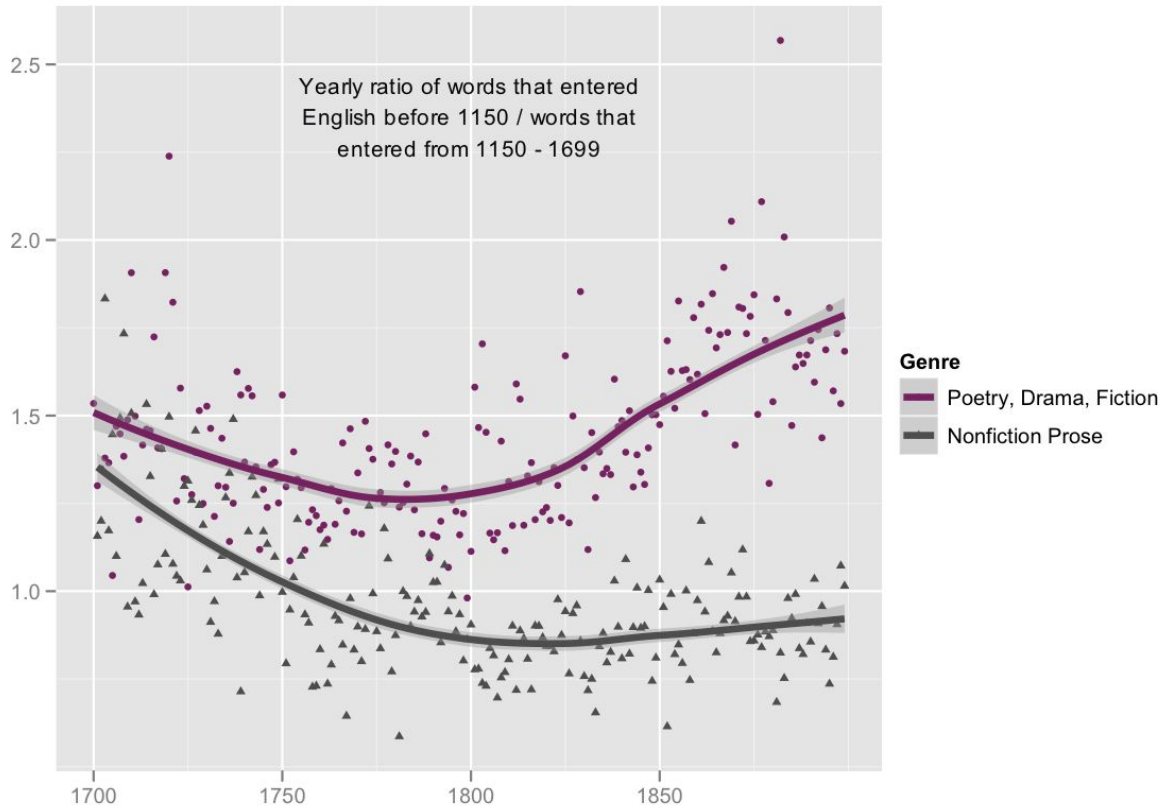

Figure 1. "The Emergence of Literary Diction," by Ted Underwood and Jordan Sellers, *Journal of Digital Humanities* (2012).

Another way to describe Underwood's and Sellers's theory of poetic simplicity would be to set chronology aside and to focus on the critical concept directly. Rather than as a graph, the idea could be expressed like this:

$$\text{Simplicity of diction} \rightarrow F_{pre} / F_{post} \quad ,$$

where F_{pre} denotes the frequency of pre-1150 words, F_{post} refers to words that entered English after 1150, and the symbol → reads "is analogous to" or "can be approximated by." Rendered into prose, this would be stated as a proposition: "Simplicity of diction can be approximated by dividing the frequency of pre-1150 words by the frequency of post-1150 words." By disconnecting the critical proposition from the empirical findings, such phrasing highlights the underlying

claim and opens that claim to further manipulation. Frequencies might be compiled into yearly totals, as in Figure 1, but they could just as easily be grouped by author, place of publication, or any other category of interest. (Below, I'll refer to such organizational choices as "topologies.") More importantly, such phrasing makes explicit the analogy that maps a concept from the qualitative domain of literary criticism onto the quantitative domain of corpus analytics.[15] The purpose of literary mathematics is to discover such analogies.

Analogies are never perfect, and maps like these are never bijective. A "bijective" map is a function that achieves perfect one-to-one correspondence between two sets, where every possible x can be transformed into a unique value of y, and every y can be transformed inversely to its corresponding x.[16] This never works out so neatly across qualitative and quantitative domains, because the qualitative realm is by definition undefined. The phrase "simplicity of diction" means different things to different scholars, and indeed even to the same person it may be taken in different senses as the situation demands. So there's a residue left behind — poems that use simple language but aren't discovered in the model — and there's excess created in the output — texts that score highly on the metric but aren't actually simple. Fuzzy definitions on the qualitative side correspond to fuzzy results on the other. Fuzziness is everywhere. This, again, is a good thing. Any critical concept worthy of serious consideration should be impossible to measure perfectly. However, skeptics should stop to ask: How could so reductive a model sort a corpus so effectively? How could it trace in outline so familiar a historical trajectory? Critics often claim that distant reading merely reflects scholarly intuition, but there's nothing intuitive about the basic argument that simplicity of diction can be measured in this way, no matter how fuzzily.

---

[15] Such analogies can be called "models." In addition to So (2017), see McCarty (2004), who cites Gentner (2002) on analogy in the sciences. For the classic statement on scientific models and the "logic of analogy," see Hesse (1966).

[16] Mendelson (1962, p. 13).

Literary mathematics is thus motivated by an entirely counterintuitive presumption: for every qualitative literary property, there exists, or might exist, an analogous quantitative approximation. This presumption is just that — a premise, a postulate, not a hypothesis or theory — because it is neither verifiable nor falsifiable. The presumption of quantifiability is a tentatively held, guiding belief that motivates research in cultural analytics.

Three misunderstandings often cause people to reject this belief out of hand. The first misunderstanding wrongly equates quantification with objectivity and scientism. Mathematics is regularized structuralism.[17] It has no special epistemology and provides no privileged access to reality, and any scholar who suggests otherwise is selling something.[18] The second misunderstanding equates mathematics with the culture of mathematics, which is tainted by

---

[17] Of course, there are many ways to define mathematics, none of which are satisfactory to all. In their classic 1941 study *What is Mathematics?*, Richard Courant and Herbert Robbins pointedly failed to answer their titular question. A sociological, "humanist" definition is offered in Hersh (1986). Although critics still sometimes gloat over disagreements about the phrase "digital humanities," terms of art like "mathematics" and "literary studies" are necessarily polyvalent.

[18] Unfortunately, hucksters abound, though more in the popular press than in digital humanities. Bode (2017) trenchantly, if ungenerously, critiques Franco Moretti and Matthew Jockers on this score. In other disciplines, the adoption of quantitative methods was sometimes accompanied with inflated and ill-considered claims regarding "scientific" objectivity. See Burton (1963) on the "quantitative revolution" in geography for an argument very similar to mine in general but differing very sharply on this score. Fish (1975) very correctly points out that interpreting evidence involves many common epistemological and rhetorical problems whether that evidence is qualitative or quantitative. Within mathematics as a discipline, the question of its relation to the external world is long settled. Courant and Robbins (1941) write,

> Throughout the ages mathematicians have considered their objects, such as numbers, points, etc., as substantial things in themselves. Since these entities had always defined attempts at an adequate description, it slowly dawned on the mathematicians of the nineteenth century that the question of the meaning of these objects as substantial things does not make sense within mathematics, if at all. The only relevant assertions concerning them do not refer to substantial reality; they state only the interrelations between mathematically "undefined objects" and the rules governing operations with them. What points, lines, numbers "actually" *are* cannot and need not be discussed in mathematical science. What matters and what corresponds to "verifiable" fact is structure and relationship, that two points determine a line, that numbers combine according to certain rules to form other numbers, etc. A clear insight into the necessity of a dissubstantiation of elementary mathematical concepts has been one of the most important and fruitful results of the modern postulational development. ("Preface")

hypercompetitive geek masculinity, test anxiety, and essentialist theories of intelligence, not to mention a long history of complicity with racism, sexism, empire, and capital.[19] On this point it's worth remembering that many disciplines, perhaps most, have similar negative associations, but in those fields scholars resist oppressive culture through ethical practice, not by rejecting whole categories of knowing. Lastly, many people reject quantification because they have an impoverished view of math's expressive capabilities, either dismissing it as mere bean counting or elevating it as some opaque wizardry.

The first two misunderstandings are not likely to be assailed by argument, bound up as they respectively are in institutional and identity politics, but I do hope to make some headway against the third. The interpretive paradigm I propose can be thought of as formalism for digital archives. Mathematics provide a rich vocabulary for describing structure. Given any collection of literary objects, it's possible to identify a formal system of relations among them. Once such structures have been identified, it's possible to describe the distribution of difference across those forms. Further, statistics can be used to evaluate the significance of those differences. How are things structured? How are some things different from others? Which differences are most significant or surprising? These three broad categories of critical inquiry — form, difference, and significance — correspond loosely to three kinds of quantitative reasoning — discrete mathematics, linear algebra, and probability theory.

In what follows, I'll describe these three areas and highlight what I believe to be their most important concepts and their most direct applications to literary inquiry. Before beginning, let me reiterate what I am *not* arguing. I do not believe that literary mathematics should displace other theoretical concerns nor that literary scholars as a community should feel compelled to

[19] On the topic of gender and discourses of mathematical competence, see Mendick (2006). For a discussion of ethical approaches to mathematics in the context of global justice, see Ernest, Greer, and Sriraman (2009).

master these ideas. However, I do argue that at least some familiarity with such principles is necessary for responsible peer review or public critique of distant reading scholarship. I further argue that increased attention to mathematical principles would make cultural analytics more intellectually rigorous and therefore more interesting and, frankly, more fun. Whether or not I can persuade my colleagues on that last point, everybody reading this essay should feel welcomed in a common project of knowing.

I invite you to join me on a happy tour through a land of strange thoughts.

Form: sets, topologies, and graphs

Discrete structures are the simplest and most fundamental maths in humanities computing.[20] By "simplest" I do not mean "easiest." Key concepts for their study can be found in set theory, graph theory, and general topology.[21] Often, scholars doing digital work use these concepts more or less unknowingly while designing data structures and research plans, describing them as "choices" made while preparing data for analysis.[22] As a practical matter, it's possible (easier, probably) to make such choices while ignoring general principles. However, attending to those principles makes them visible in a new way. Choices are evanescent happenings. Topologies are concrete things that can be evaluated, replicated, and treated as objects of study in their own right.

---

[20] For a general overview of these topics with special emphasis on their application in computing, see Gersting (2002).

[21] The classic introduction to set theory is Halmos (1960). For graph theory, see Ore (1962), Flament (1963), and Harary, Norman, and Cartwright (1965). Recent introductions to graph theory oriented towards the study of networks can be found in Newman (2010, chapter 6) and Barabási (2015, chapter 2). For general topology, see Mendelson (1962), Lee (2000), and Shick (2007). Munkres (1984) is the classic graduate-level textbook for algebraic topology but presupposes a mastery of basic point-set and geometric concepts.

[22] Buurma (2017) briefly summarizes commentary on this point.

A *set* is any collection of objects. In literary history, sets usually involve language, one way or another. A bibliography is a set of titles and other metadata. A text is a set of words. Sets can contain other sets. A digital archive is a set of words organized by documents, which in turn are subsets of the larger whole. In this very typical situation, the documents represent a *topology* over the corpus; that is to say, they divide the corpus into parts and describe its overall structure.[23] A *topological space* is a total collection — an "underlying set" — combined with a set of subsets that organize it. To be a topology, these subsets must obey certain rules guaranteeing their coherence. All subsets must reside completely inside the underlying set, for example, and any areas of overlap among them must also be included as distinct elements. These rules give a topology the properties of space, of hereness and thereness, such that it's possible to talk meaningfully of some words appearing over here in these books but not over there in those. Textuality is the prevailing topology of historical discourse: texts divide archives into segments distinguished from each other and from unstructured discourse.[24] At its simplest, textuality is an indexing function that associates word tokens with word types and document titles.[25] However, textuality is not the only available topology. Authorship associates words with names.

---

[23] I do not know of any linguists who associate topology with core concepts in corpus analysis. Corpus linguists seem curiously indifferent to the theoretical foundations of their work and seem to prefer instead to adopt an ethos of no-nonsense practicality. See Baayen (2008) and McEnery and Hardie (2012). Jenset and McGillivray (2017) describe "quantitative historical linguistics" in terms that closely parallel Underwood's genealogy of distant reading. In both fields, the usual research imperative seems to be to demonstrate findings commensurable with existing theory, rather than to advance theory in new directions.

[24] Outside the context of corpus-based study, a topological theory of textuality is most directly implied in Silverstein and Urban (1996).

[25] Scholars trained in digital editing will be familiar with the "ordered hierarchy of content objects" (OHCO) model of textuality. See Goldfarb (1981), DeRose et al. (1990), Sperberg-McQueen (1991), and Renear (2004). Editing with XML similarly involves devising a topological model for a corpus, but the ordinary topology of distant reading is in practice more flexible than any viable markup system, because it presumes each character to be a discrete element and relegates all categories and other structures to the status of an attribute.

Chronology lumps them together by date, as in Figure 1 above, when Underwood and Sellers distributed their corpus over the topological space of time.

The phrase "topological space of time" has a funny ring to it. I mean to call attention to a theoretical challenge at the center of cultural analytics. When we — all scholars — go about our reading and research, we bring to that work a wide range of concepts that organize our attention and guide our arguments. What is literature? What is a nation-state, a historical period, a sonnet? These categories are refined through training but continue to dwell within our holistic sense of reality. We know what ideologies are in part because we feel their power. We know what meanings are because we mean them (or don't, as the case may be). Not so in distant reading. As soon as the corpus is introduced as an intervening object of analysis, all theories of textuality, historicity, and subjectivity need to be analogized to corresponding data structures. In cultural analytics, the "author function" is just that — a function, written into an algorithm, that reorganizes a corpus by grouping documents according to the value of the "author" field in the metadata. The same holds true for time, geography, syntax, genre, and whatever category we might want to impose on the corpus.[26] Each critical concept must be represented as a distinct topology that places the underlying set under the description of its subsets. Every space must be accompanied by a corresponding function that maps elements of the underlying set onto itself. The challenge is to devise such topologies and to invent analytical procedures that navigate among them.[27]

---

[26] Whereas Michel Foucault (1984) denaturalized authorship, corpus-based historical research denaturalizes all literary-critical concepts.

[27] Usually this challenge is described imprecisely as "interpretation" that goes into designing computer models. E.g. Alvarado (2012, p. 52) and Brown (2015, p. 30). The word "interpretation" is too vague to do more good than harm to a theory of distant reading, as is, for that matter, "reading."

At bottom, this is a problem of counting.[28] General principles get very abstract, but please bear with me for just a moment. If literary mathematics describes "literature's countable features," we might ask: What kinds of things are features, and how can they be counted? To answer this question, let us consider the typical situation. A literary historian is armed with a collection of text files and a bibliography of metadata about the original sources. (A more general and more rigorous account can be found in Table 1.) Every instance of every word in the corpus can be treated as an atomic element with a variety of attributes. One attribute is the word type.[29] Other attributes often apply to whole documents. Although in shorthand we might say Shakespeare's *Much Ado about Nothing* uses the word "love" 93 times, a more precise phrasing would say instead that 93 tokens in the corpus have the following attributes: they are contained in a document of the title "Much Ado about Nothing," they were written by "Shakespeare, William. 1564-1616," and they are instances of the type "love." Other attributes might include part of speech, or the fact that they appear in certain sentences or certain grammatical patterns.[30] By rule, tokens can be counted together if they share at least one attribute. We might count all instances of a type that occur throughout the corpus, no matter in what document, or we might count how many tokens are in a document, no matter of what type. Once counted, frequencies can be added together, but only if their characteristic attributes belong to a common class. A vocabulary of word types is a class; in a bibliography, classes are categories of metadata, like title, author, or date. Scholars might think of some reason why it's interesting to add all the

[28] Many of the issues I raise here can be approached as problems in combinatorics. See Gersting (2002, p. 188-222).

[29] On the distinction between word types and word tokens, see McEnery and Hardie (2012, p. 50).

[30] This way of thinking about literary data presumes an implied Resource Description Format (RDF) tuple for every observable and describable fact in a corpus, only some of which are explicit in what we conventionally describe as the metadata to a collection, and many of which emerge only from the analysis itself — as do, for example, syntactic forms like those discussed in Shore (2018). For an introduction to RDF "Linked Data" conceptual models, see van Hooland and Verborgh (2014).

tokens written by Shakespeare to those written by Jonson, but they would never add all of Shakespeare's words to all the words published in, say, 1601. The two subsets might intersect — Shakespeare might have published some words in 1601 — but they aren't conceptually commensurable categories, so there's no qualitative historical concept their union could ever approximate.

Table 1. Principles of Literary Topology

The ordinary topological space for distant reading includes a set of elements **X**, a collection of subsets **Θ**, and an index of tuples **I** joining elements *x* to attributes *a*. An element *x* with an attribute *a* is written x^a. An attribute *a* belonging to an element *x* is written a^x. Attributes belong to attribute classes *c*.

Θ is defined by **I** according to the following rules:

1. **X** and ∅ are in **X**.	Borrowed directly from the general definition of a topology, these rules ensure the integrity of the system. The complete set and the empty set are both included, and no subsets can overlap unless their overlap is specified.
2. For any subsets O_1, O_2 in **Θ**, the intersection $O_1 \cap O_2$ is also in **Θ**.	
3. For any elements x_1, x_2 in **X**, if there exists an *a* in **I** such that $a^{x1} = a^{x2}$, then there exists an O in **Θ** such that the union $x_1 \cup x_2$ is in O.	These additional constraints define the relationship between the data, **X**, and the metadata, **I**. Two elements can be joined into a subset (that is, they can be counted together) only if they share some attribute. Each subset is defined by a characteristic attribute, and subsets can be joined only if their characteristic attributes belong to a common class. The number of elements in a subset is the *frequency* of that subset's characteristic attribute. Taken together, these constraints define the phrase, "literature's countable features."
4. For all O in **Θ**, there exists a unique *a* in **I** such that, for every *x* in O, $a^x = a$.	
5. For any subsets O_1, O_2 in **Θ**, the intersection $O_1 \cap O_2$ contains all elements *x* such that $x^{a1} = x^{a2}$. The value *a* for such subsets is written $\{ a_1 \cap a_2 \}$.	
6. For any attributes a_1, a_2 in **I**, the union $a_1 \cup a_2$ is in **I** if and only if there exists an attribute class *c* such that $c^{a1} = c^{a2}$. The value *a* in such cases is written $\{ a_1 \cup a_2 \}$.	
7. For every *x* in **X** such that $x^a = x^{\{ a1 \cap a2 \}}$, there exists a link between a_1 and a_2.	This final rule defines a graph over **I**. Two features are linked if their characteristic attributes have elements in common.

Taken together, these rules define what it means to count words in a corpus. A literary topology is historically valid if it conforms to these rules. Analogies drawn from valid topologies

can be trusted as veridical, but only insofar as the corpus sufficiently represents historical discourse and as the metadata sufficiently describes its formal properties.[31] Such, at least, are several principles that seem true to me. The study of literary topology, which I am outlining here, would debate such principles by showing how different data structures express different literary-historical concepts. The goal is to preserve the integrity of the analogy between the qualitative and quantitative domains as well as to discover new homeomorphisms, new possible correlations between the corpus and the past. Possibly there is a theoretical paradigm under which { Shakespeare ∪ 1601 } makes sense as a thing that actually existed. More likely, scholars would decide that concepts like authorship and temporality require additional constraints or otherwise more sophisticated forms of expression.

Topologies enter the study of literary form at the point of data curation and design. How does the corpus look differently when distributed over, say, the spaces of sentences, paragraphs, and full texts, and how does this vary among authors or across historical periods? The basic method begins by identifying multiple candidate topologies, then analyzing their structure and evaluating their most significant points of overlap and divergence. (I will discuss analytical and statistical concepts appropriate for this work below.)

However, topological considerations are not restricted to questions of literary form. Social and geographical forms can be studied in much the same way. Two research areas with the most direct applications to literary history include network science and geospatial topology.[32] Many aspiring digital humanists first encounter literary mathematics when they draw their

---

[31] What counts as "sufficiently" will vary from case to case. The question of what cultural data are sufficient to serve as proxies for what cultural phenomena is among the most difficult and important questions in literary mathematics and cultural analytics. McEnery and Hardie (2002, p. 15-19) address this issue, briefly, in terms of the "representativeness" and "balance" of a corpus.

[32] For a succinct review of the history of network analysis within the social sciences, see Scott (2012). Barabási (2017) attends more closely to developments in mathematics and physics.

first network graph or their first map of historical data. However, in graph theory, a graph isn't something you can look at.[33] A *graph* is a particular kind of topological structure that organizes objects, called *nodes* or *vertices*, into pairs connected by *links* or *edges*. Across these simple connections, nodes join together into large, complex networks.[34] The goal of network science is to correlate the local formation of individual links with emergent patterns reflected in a network's overall structure.[35] In geospatial topology, features like cities and countries may or may not have fixed boundaries, but they often have discrete structures that are well defined nonetheless. Regardless of its precise area, London is in England, and England sits adjacent to Scotland and Wales. Geospatial topology provides a mathematical framework for such putatively qualitative concepts, providing the basis for much work in geographical information science.[36] When combined, network science and GIS enable the study of discrete structures distributed over geographic space, whether those structures are physical networks like roads, socially stipulated entities like political territories, or lived practices like kinship and communication.[37]

The study of networks is well underway in literary history: studies have appeared that use graphs to represent connections among people in publication or epistolary networks, or connections among characters in novels and plays.[38] So too, mapping.[39] The most sophisticated

---

Geospatial topology has a much shorter history: its conceptual foundations are traced back to Egenhofer and Herring (1990) and Egenhofer and Franzosa (1991).

[33] Graphs in this technical sense are very different objects from those described in Moretti (2005).

[34] For this reason, network science shares much overlap with the study of complex systems. See Strogatz (2001) and Barabási (2011).

[35] Brandes et al. (2013) define network science more broadly as "the study of the collection, management, analysis, interpretation, and presentation of relational data" (2).

[36] The phrase geographical information science (rather than geographical information *systems*) can be traced to Goodchild (1992).

[37] For a detailed overview of research into spatial networks, see Barthélemy (2011).

[38] For drama networks, see Moretti (2011) and Algee-Hewitt (2017). For publishing networks see Smith, Cordell, and Mullen (2015), Greteman (2015), and Gavin (2016). Epistolary networks are described in Mandell (2013), Ahnert and Ahnert (2015), Ahnert (2016), and Edelstein et al. (2017).

and promising areas of research cross multiple topological domains to learn how the distributions of things and people affect the distribution of ideas, and vice versa.[40]

The first task of distant reading is to identify countable proxies for qualitative concepts. Rhetorically effective arguments often suppress this aspect of the research, preferring to emphasize intuitive connections across the qualitative and quantitative domains. That's a shame, because here's where the heavy intellectual lifting usually occurs. General topology is the abstract theory of such structures, applied in fields where scientists and mathematicians are tasked with reconciling measurements across different coordinate systems, like differential geometry, geodesy, cartography, and physics.[41] In software engineering, these topics arise when designing conceptual models that establish data categories and relationships.[42] When applied to corpora by linguists and computer scientists, these questions are central to experiment design; much research in computation and language hinges on the question of how corpora can be divided and how resulting measurements can be correlated.[43] Analogized to literary studies, such research would identify new topologies that expose different aspects of the cultural record. Such questions strike to the heart of literary theory: How do texts exist in time? How are persons connected to words? Across what apparent discontinuities are real continuities imaginable, even necessary? Comparative analyses of literary topologies would help to clarify

---

[39] The "spatial humanities" represent one of the more vibrant subfields of digital humanities. For applications to literary study in particular, see Cooper, Donaldson, and Murrieta-Flores (2017).
[40] Research in digital humanities that bridges spatial and textual analysis includes Gregory and Hardie (2011), Wilkens (2013), Broadwell and Tangherlini (2016), and Gavin and Gidal (2017). In the field of geospatial computing, similar lines of inquiry are pursued in Kuhn (2005) and Schwering (2008). Among the more imaginative extensions of GIS within literary studies is Murrieta-Flores and Howell (2017).
[41] Carlsson (2009) emphasizes the suitability of topology for reasoning across the qualitative and quantitative domains in a range of disciplines, from image processing to neuroscience.
[42] Gersting (2002) defines these problems in terms of set theory and combinatorics.
[43] Most visibly, in the long-standing debate over the "bag-of-words" hypothesis. David Lewis describes the issue as old-hat as early as 1998: "An ongoing surprise and disappointment is that structurally simple representations produced without linguistic or domain knowledge have been as effective as any others."

these issues, both as a practical matter for digital humanists designing research projects and more generally for scholars interested in the questions themselves.

Difference: matrices and metric spaces

A *matrix* is a rectangular array of numbers with fixed rows and columns, like a table or a spreadsheet.[44] Matrices are the most common and most important structure for analyzing data. Indeed, the topological issues discussed in the previous section can all be defined in terms of matrices; that is to say, any topology over a corpus can be defined by the matrices it makes available for analysis. A topology of documents implies a *term-document matrix*, where words are taken as the rows, documents as the columns, and the value of each cell records the frequency of each word in each document.[45] Social-network data can be described similarly; in a typical *bipartite network* represented by an *incidence matrix*, the rows represent people and the columns stand for events that connect them, like social meetings they attended or, in the case of citation-network studies, academic papers in which they are mutually cited.[46] Geographical data is stored in matrices where the rows represent places and the columns contain statistical measurements, like census or climate records.[47] Sometimes rows are referred to as *observations* and columns are called *variables* or *attributes*. Tabular data can be organized in any number of

[44] The basics of matrix algebra are briefly reviewed in Gersting (2002). More comprehensive treatments are widely available. My discussion below depends most heavily on Lay (1999) and Axler (2004).

[45] Often credited with introducing the term-document matrix format is Salton, Wong, and Yang (1975). For a more detailed discussion see Salton and McGill (1983). Turney and Pantel survey research applications in the information sciences; see Clark (2015) for linguistics.

[46] For the incidence matrix format and its projection onto univariate network models, including co-citation networks, see Newman (2010). Its explanatory potential was first explored in Homans (1951).

[47] Berry (1968) first described the use of matrices, stacked over a third dimension of time, to represent the changing distribution of variables over places.

variations depending on the adopted topology, and complex topologies often imply systems of related matrices.

Virtually all forms of quantitative analysis involve computing over matrices at some point or another, and so if there's one area of general mathematics that digital humanists should review as part of their training, it's matrix algebra.[48] The central idea of matrix algebra is to represent numbers in the form of a fixed sequence, sometimes called a *vector*. Vectors can be multiplied together by adding the products of their respective elements, reducing them to a single value. The *dot product* (or inner product) takes this form:

$$a \cdot b = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n \quad ,$$

where a and b represent two sequences of numbers of the same length, n.[49] Each a is multiplied against its corresponding b, and the sum is taken over the whole. The inner product is useful analytically because it represents the overlap shared between any two vectors. If two rows of a matrix have a lot in common, the inner product between them will be high. If not, it'll be low. For example, if you were analyzing a term-document matrix representing a corpus of genre fiction, and you compared the rows for "detective," "police," and "dragon," you'd likely find that the inner product between "detective" and "police" is higher because they're used frequently in many of the same novels; the big *a*s get multiplied by the big *b*s. By contrast, "detective" and "dragon" would be lower, because they tend to appear in different kinds of books; the big *a*s are lost when multiplied against low *b*s. Most of the fantasy words get cancelled out by zero values

[48] Matrix algebra receives strangely cursory treatment in Juola and Ramsay (2017), which, despite its title, does not appear to have been written with any humanities applications in mind. Unfortunately, there exists no minimally adequate general book on mathematical structures for literary study. Scholars hoping to get their feet wet with matrices should start instead with Widdows (2004), which offers an elegant and compact introduction to matrices in the context of semantic analysis, or with Newman (2010), which introduces many of the same concepts in network science.

[49] The inner product is somewhat more general than the dot product, because it can handle complex numbers. See Axler (2004, p. 98-99). I use the terms interchangeably.

in detective fiction, and vice versa.[50] Thus, the dot product measures the degree of alignment between any two vectors of numbers, showing how much any two observations have in common when defined over the same variables.

By reducing two vectors to a single value that describes their overlap, the dot product becomes a distance function, and the matrix over which it's computed represents a continuous metric space. A *metric space* is a topological space where all elements exist across describable intervals.[51] Chronology is a one dimensional metric space. Along a timeline, the years 1616 and 1623 sit seven years apart. But not all topologies are metric in this way. Authorship is not metrizable: you can't subtract Shakespeare from Jonson. However, if Shakespeare and Jonson are observed over fixed variables — over the words used in their plays, for example — then it is possible to take their inner product and so to characterize how much overlap they share. Shakespeare and Jonson probably share more words than, say, Chaucer shares with Dickens. They'll be closer in semantic space.[52] As a distance metric, the inner product suggests that the rows and columns of a table exist along some implied continuum of possibility, and so matrices are sometimes said to exist over *inner product spaces* or, more generally, *latent spaces*.[53] (See Table 2.) A latent space is a metric space that organizes a topology without being explicit in that topology's definition. Most scholars use some version of the dot product for this purpose. Cosine distance, a popular metric, is just normalized to the unit vectors. Euclidean distance is based even more closely on the Pythagorean theorem and so works, not by multiplying each element, but by taking the sum of the squares of their difference. Some distance metrics can get quite

---

[50] For more extended examples like this, see Widdows (2004) and Gavin (2018).
[51] For the distinction between metric spaces and topological spaces, see Mendelson (1962, ch. 3) and Shick (2007, ch. 9).
[52] The concept of "semantic space" is fundamental to computational studies of meaning. See Lowe (2001), Sahlgren (2006), and Gavin, Jennings, Kersey, and Pasanek (2018).
[53] For inner-product spaces, see Axler (2004, ch. 6). Though the term is often used in various applications, I know of no general definition of the phrase "latent space."

complicated and exotic, but they usually share a basic structure: combine the respective elements, then combine the combinations into one value.[54] This final value represents the proximity between any two objects in latent space. Such proximities are analogous to qualitative notions of similarity and difference.

Table 2. Continuous Latent Spaces
A metric space is a set of elements and a distance function that describes intervals among them. A vector **v** represents frequencies over a fixed list of attributes belonging to a common class (Rule 6 above), such as a vocabulary of word forms or a list of book titles. Vectors placed in a matrix **A** are compared using a distance measurement, *d*. Proximity in latent space is analogous to qualitative similarity.

Note that, when discussing the principles of literary topology, *a* was used to denote attributes, but here *a* designates their frequency, as defined in Rules 3, 4, 5, & 6 above.

$\mathbf{v} = \{ a_1, a_2, \dots , a_n \}$ $\mathbf{A} = \{ \mathbf{v}_1, \mathbf{v}_2, \dots , \mathbf{v}_m \}$	A vector **v** represents the distribution of frequencies over a fixed list of joinable attributes. When represented as a matrix **A**, vectors are rows or columns of numbers, and the matrix represents a linear map that distributes one class over the space of another.
$\mathbf{u} \cdot \mathbf{v} = a_{u1}a_{v1} + a_{u2}a_{v2} + \dots + a_{un}a_{vn}$ $\cos(\mathbf{u},\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} \ / \ \|\mathbf{u}\| \cdot \|\mathbf{v}\|$	The dot product combines two vectors **u**, **v** into a single value, which can be normalized to a distance metric *d* that varies along the continuous real interval [0, 1].
$\mathbf{B} = \mathbf{A}\mathbf{A}^T$	A matrix multiplied by the transpose of itself results in a square matrix that records how often attributes of one class collocate with each other, when distributed over the space of a second class.
$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$	Singular value decomposition exposes a spectrum **S** of variation that subtends vector space. *Dimensionality reduction* discards values of **S** that display low variance, removing "noise" from the model. In many situations, distance measurements over reduced forms of **US** or $\mathbf{S}\mathbf{V}^T$ are more closely analogous to human notions of similarity than when calculated naively over **A**.

Matrix multiplication extends this operation, merging two matrices into a third by systematically taking the inner products of their respective rows and columns. This process can

[54] For example, Mahalanobis distance, described in Mahalanobis (1936), works much like Euclidean distance, but transforms each distance metric over a covariance matrix, thus adjusting the weight given to each element's difference based on its typical patterns of variation.

be reversed, much like in regular arithmetic. Just as any two matrices of similar shape can be multiplied together to create a third, so too any matrix can be decomposed into its component parts. "Matrix division" exists in slightly different forms depending on context, but you'll see it referred to as *principal component analysis*, *matrix factorization*, or *singular value decomposition*.[55] Multiplying two matrices together creates a detailed picture of their points of overlap. Decomposing a matrix exposes its underlying structure.

When applied to a term-document matrix, singular-value decomposition is called *latent semantic analysis*.[56] Beginning with a table of word frequencies in books, decomposition identifies three component matrices: one of words, showing their relative axes of difference, another of books, showing theirs, and a third that represents the size of each latent dimension. This latent matrix is called the *spectrum* because it organizes the system along a scale of gradually decreasing intensity. Each level in this spectrum is marked by a special number called an *eigenvalue* that sets the scale; the large eigenvalues point to areas where the system exhibits widest variation. Partitioning a matrix along the axes of its spectrum identifies groups of objects that appear together in meaningful patterns. When computed over words, the spectrum exposes hidden axes of meaning that structure *semantic spaces*. The phrase *semantic similarity* refers to the proximities among words and documents in such spaces.

Matrix decomposition supports many other kinds of analysis as well. In the study of networks, spectral partitioning uses the second-smallest eigenvalue to separate a graph into

---

[55] For a detailed discussion of principal component analysis, see Jolliffe (1986), especially chapter 7, which distinguishes PCA from factor analysis.

[56] LSA represents the clearest point of contact between literary theory and the information sciences. For compact introductions to its mathematical properties, see Bellegarda (2007) and Martin and Berry (2007). More thorough discussion is provided in Landauer (2007). In that collection, see also Dumais (2007), which describes the origins of the theory in the context of information retrieval.

modules.[57] In image processing, it's a crucial step in tasks like zooming, compression and decompression, and pattern recognition (including face detection).[58] Similarly, in geospatial modeling, principal component analysis finds regions in the data; that is, it finds points that share similar statistical profiles and so sit near each other in latent space, whether or not they sit together in physical space.[59] All of these applications share a common theoretical base. All are computed over matrices that represent data in prescribed rectangular structures. Once placed in such an array, the numbers of any given row or column are never quite identical to themselves, because the matrix itself is an elaborate proposition about their mutual interrelation. Every vector of numbers carries the latent possibility of its comparison and recombination along countless possible axes. In this way, matrix decomposition formalizes dialectical reason by systematically toggling among parts to create varying snapshots of the wholes they constitute.

New machine-learning algorithms like topic models and word-embeddings are designed to efficiently mimic this procedure for semantic analysis.[60] For large datasets, computing all the necessary linear combinations can be too much for a desktop machine, so software engineers have devised various shortcuts using randomized sampling. These methods are much more complicated than latent semantic analysis, but through an unfortunate twist in the history of literary criticism, this area of research was introduced to scholars through such software packages. Not only were literary historians insulated from the mathematics behind the software, as mentioned above, but they were also misled to believe that these very complicated operations are the most immediately appropriate maths for their critical questions. If you're trying to

---

[57] For a compact but thorough explanation of this and other applications of spectral graph theory to network data, see van Mieghem (2011).
[58] E.g., Kienzle et al. (2005).
[59] Demšar et al. (2013) surveys of applications of PCA geographical problems.
[60] The relationship between LSA and topic modeling is described in Steyvers and Griffiths (2007). Levy and Goldberg (2014) argue that word-embedding models have mathematical properties very similar to factorized matrices.

understand topic modeling, the worst place to start is with topic modeling.[61] Simple matrix factorization was always the motivating inspiration for those algorithms, and matrix algebra has a long intellectual tradition across many disciplines.

Exploring applications of this paradigm to the cultural record should be an explicit priority for the digital humanities. Possible lines of general inquiry might ask: What matrices are appropriate for studying semantic, social, geographical, and temporal networks? How can they be combined and decomposed, and with what effect? To what qualitative differences are latent distances analogous? What spaces organize what kinds of meaning? What is the shape of history?

Significance: information and probability

Discrete structures describe the forms of literary data. Metric spaces describe the distribution of differences across those forms. Statistical models describe the significance of those differences. What does it mean for something to be significant, interesting, or surprising? These questions, too, can be asked quantitatively.

A *probability space* shares much in common with a topological space, but one key difference is worth noting.[62] Valid topologies ensure continuity by requiring that all areas of overlap are included. In a probability space, continuity is not required. Instead, every subset must be paired with its complementary opposite. If you want to say how likely a coin toss will come up heads, you also have to say how likely it will come up tails. What are the odds you'll roll snake eyes? The precise opposite of the odds you'll roll anything else. Every countable subset in

[61] Of course, if it wasn't for the hard way, who'd get anything done? Ted Underwood (2012) followed precisely this path.

[62] For the formal definition of a probability space, see Laha and Rohatgi (1979). The concerns of probability theory also overlap significantly with measure theory, a field of mathematics that Shannon (1949) off-handedly refers to as providing the conceptual foundations for information theory. For measure theory and its relation to probability, see Tao (2011).

a probability space exists in relation to its countable opposite, its *complement*.[63] With classes of events like coin tosses and rolls of the dice, this complementarity seems obvious — too obvious to be mentioned, really, which is why introductory books and lectures on statistics usually suppress this central idea, even though the general foundations of probability theory begin with it.[64] It's confusingly abstract. Any statement declaring how likely something is requires a taken-for-granted categorical background of other somethings, distinct from the first but similar enough to belong to a common class of events. Thinking probabilistically always means thinking categorically. Evaluating significance means contrasting instances of one kind from instances of a class of kinds.

In this, probability spaces also share a lot in common with vector spaces, which also organize elements into interlocking categories. Despite the very different notation and very different jargon, probabilities are very similar to vectors. A *discrete probability distribution* is just like a vector, except, instead of showing the raw frequency for each variable, it shows the relative or proportional frequency. All this means is that you divide each value by the total and represent it as a percentage. Rather than say Shakespeare used "love" 93 times in *Much Ado about Nothing*, you say that he uses "love" 0.12% of the time. A *joint probability distribution* is just like a matrix, except that every value is divided by the sum of the matrix as a whole. It's like asking: If you selected any token at random from the corpus, what are the odds you'd get one that is both from *Much Ado about Nothing* and of the type "love"? As we'll see, these extra steps of processing do nothing to interfere with operations over the matrix — normalizations of this kind adjust the scale of each vector without destroying its basic shape — but the subjunctive

[63] Halmos (1960) describes complementation in the context of set theory.
[64] Probability theory, like topology, is concerned primarily with the underlying premises and structures derived from set theory, while statistics are applied to specific research cases. Digital humanists are for this reason likely to find statistics, such as those introduced in Dowdy and Wearden (1991) or Bolstad (2007), will be more directly applicable to their day-to-day work.

mindset implied by probabilistic thinking invites creative comparisons across various categorical baselines, any of which might reveal significant structures in the observed data.

But what does "significant" mean in this context? Before looking at such a baseline in detail, let's pause for a moment to think about two calculations commonly performed over probability distributions, to develop a slightly deeper sense of how this field of mathematics differs from matrix algebra. Remember, the inner product represents two sequences of numbers by multiplying their corresponding elements and taking the sum of the products. *Entropy* and *relative entropy* work in a similar way.[65] Given a length-n sequence of proportional frequencies, entropy is calculated by taking the sum of each element multiplied by the logarithm of itself:

$$H = -(p_1 \log(p_1) + p_2 \log(p_2) + \ldots + p_n \log(p_n)) = -\sum p_i \log(p_i) \quad ,$$

and relative entropy, like the dot product, weaves two sequences of values together:

$$D(p||q) = -(p_1 \log(p_1 / q_1) + p_2 \log(p_2 / q_2) + \ldots + p_n \log(p_n / q_n)) = -\sum p_i \log(p_i / q_i) \quad .$$

These calculations differ from those of matrix algebra in one important respect: the logarithm is inserted at each point. By taking each value and representing it as the exponent of some base, the logarithm has the effect of smoothing out variation along the sequence by measuring it against an underlying scale. It also responds to large values of p: as p approaches 1, its logarithm approaches 0, so distributions with high values concentrated in just a few elements will exhibit lower entropy, while more evenly distributed values exhibit higher entropy. This metric is sometimes called the quantity of *information* in a system.[66] Relative entropy measures how much variation one vector contributes to another, so is sometimes called *information gain* or, because of its similarity to a more conventional distance metric, *divergence*.[67] The key point to

---

[65] These concepts are succinctly defined in Cover and Thomas (1991, chap. 2).
[66] The best account of "information" as that term is used here probably remains Shannon's discussion of discrete noiseless systems in Shannon and Weaver (1949).
[67] More precisely, *Kullback-Leibler* divergence. Like the inner product, divergence is sensitive to large differences between individual elements: the further apart any value of p is from its corresponding q, the further the ratio between them will deviate from 1, and therefore the larger

take away here is to notice how similar these metrics are to calculations common in linear algebra, but also to see that they further describe the level of variation in a system while identifying areas where that variation is most densely concentrated.

In this way, descriptive statistics analogize significance to deviance. Something is significant insofar as it differs observably from the categories constructed to define it. To evaluate significance is a two step process: first you generate an expected value based on averages over whole classes of data, then you compare those expected values to the actual, observed frequencies.[68] For example, the most significant values in a matrix can be identified by taking the *pointwise mutual information*, sometimes abbreviated as PMI or PPMI.[69] (See Table 3.) Given a matrix of Shakespeare's plays, you begin by calculating two probability distributions, one of plays and another of words. Of all tokens in the corpus, what percentage appear in *Hamlet*? What percentage are of the type "ghost"? Each categorical value is multiplied together to return the *expected value* for each cell in the table, $p(x)p(y)$. You might find that the expected value for "love" in *Much Ado about Nothing* is merely 0.03%. By taking the logarithm of the ratio between the *observed value*, $p(x, y)$, and this expected value, you get the PPMI score, in this case as log(.0012 / .0003) = log(4) = 0.6. When the actual value is lower than the observed value, the logarithm turns negative, and in the final step of processing all negative values are disregarded and set to zero. The algorithm is complete. You started with a matrix of simple word counts, and PPMI returned a matrix showing which terms are most significantly

---

its logarithm will be. Because at each step those logarithms are multiplied against p, the relative entropy is higher when large deviations correspond with high values of p. Essentially, relative entropy asks how much one distribution's highest, most important values deviate from another's. See Kullback and Leibler (1951).

[68] The "expected value" E is typically the average or, when comparing two variables, the line of best fit. The "error" or "residual" is the difference between that expected value and the actual, observed value. See Dowdy and Wearden (1991, p. 229-38).

[69] Cover and Thomas (1991) define mutual information as an extension of relative entropy. As a normalizing procedure for semantic analysis, PPMI is described in Turney and Pantel (2010).

overrepresented where. Every value in the new matrix reflects the outcome of a complex series of interconnected reasoning about how Shakespeare's works are structured. Rather than mere data, each value is an explicit claim.

Table 3. Descriptive Statistics for Literary Study
The elaborate formulae of statistical reasoning are built like nested dolls with layers of calculations that, taken individually, are little more than simple arithmetic. Given any sequence of numbers, a, most calculations begin by comparing that sequence against its sum, $\Sigma(a_i)$, its length, n, or some combination of them.

$a' = \Sigma(a_i) / n$ $Var(a) = \Sigma(a_i - a') / n$ $Dev(a) = Var(a)^{½}$ $Cov(a,b) = \Sigma(a_i - a')(b_i - b') / n$ $Corr(a,b) = Cov(a,b) / [Dev(a) \times Dev(b)]^{½}$	The most common measure for summarizing a sequence of numbers, a, is to take its arithmetic mean, a' (the average). Averages provide a baseline against which individual values a_i can be evaluated. The *variance*, measures the average distance that separates each value from the mean, and the *standard deviation* is simply the square root of the variance. Much like the dot product, the *covariance* weaves two values together by multiplying their respective elements, and the *correlation* scales the covariance by dividing the square root of the product of the deviations.
$E(a) = a'$ $E(a) = \delta + \mu b$ $a = E(a) + \varepsilon_i$	When working with a single variable, a, the most common expected value is simply the mean. When comparing two variables, a and b, a linear regression model can be used to identify a line of best fit, which assigns a different expected value based on the slope, μ, and the intercept, δ, of the regression line. The residual or error for each value, ε_i, describes how far each actual value deviates from the model.
$p(a) = a / \Sigma(a_i)$; hereafter p $H = -\Sigma\, p_i \log(\, p_i\,)$ $D(p\|\|q) = -\Sigma\, p_i \log(\, p_i / q_i\,)$	Entropy, H, describes the total variation that exists in a sequence of probabilities, $p(x)$, by multiplying each value against the logarithm of itself. Relative entropy, sometimes called Kullback-Leibler divergence, D, identifies where high values in one distribution correspond with big differences from another.
$PMI(a,b) = \log[p(a,b) / p(a) \times p(b)]$ $PPMI(a,b) = PMI(a,b)$, if $PMI(a,b) > 0$, otherwise 0	Pairwise mutual information (PMI) describes how a joint probability distribution, $p(a,b)$ relates to the distributions of its underlying categories, $p(a)$ and $p(b)$. When the observed values are lower than expected, the logarithm turns negative, and positive PMI (PPMI) sets such values to zero.

This calculation and others like it, such as TF-IDF, are useful in many interpretive contexts.[70] They're often performed prior to matrix decomposition in semantic analyses and have been shown to dramatically improve results. In information science, to improve results means to perform better at complex tasks like guessing the correct answers on vocabulary tests or returning the right books in a library catalog search. When scholars use these methods for distant reading, to improve results means to delineate more clearly categories of words, groups of authors, and genres of books.[71]

However, insofar as attention turns away from evaluating texts within genres and towards evaluating the genres themselves, the question of statistical significance shifts in subtle but important ways. How do we know whether the categories we use to describe literary texts are adequate to explain them? Statisticians describe this problem as *uncertainty*.[72] I remarked above that numerical expression does not in and of itself entail any special epistemology, but if there's one group of mathematicians most preoccupied with questions involving belief and inference, it's statisticians. Their goal is to evaluate the trustworthiness of data-driven claims, given that, in most cases, such claims are extrapolated from samples but generalized to whole populations.[73] In voter opinion polls, clinical trials, and other experimental contexts, researchers try to explain what's true about everybody while constrained to observing just a few people. Under a classical, frequentist paradigm, statistical tests imagine an ideal universe where

---

[70] Term-frequency inverse-document frequency (TF-IDF) weighting places greater emphasis on lower frequency words and is used in many information-retrieval systems. Karen Spärck Jones is conventionally credited with its invention. (See Robertson and Spärck Jones, [1976]). Term weighting techniques are reviewed in Salton and Buckley (1988).

[71] Ramsay (2011) uses TF-IDF weighting to creative effect in a reading of Virginia Woolf's *The Waves*. Hoover (2016) critiques the method as inappropriate to Ramsay's critical goals.

[72] This is the central theme of Lindley (2006), a highly accessible overview of statistics for non-mathematicians.

[73] The difficulty of reasoning from samples often causes problems for statisticians. See Lindley (2006, p. 51-54), Dowdy and Wearden (1991, p. 21-26), and Bolstad (2007, p. 13-21), for discussions of sampling errors and other kinds of interpretive dangers.

experiments can be performed infinitely many times, then estimate the likely shape of data in that imaginary world, based on actual data collected in this one. Under a Bayesian framework, scholars begin with a subjective, *prior* expectation about the likelihood of events, then update those expectations by comparing them against new observations.[74] In either case, statistical findings are evaluated in the end based on how well they predict future events.

Because so much of the intellectual scaffolding of statistics deals with questions of uncertainty and prediction, it may seem far outside the bounds of literary scholarship. If there's uncertainty in literary studies, it's usually about what texts mean, not about what words they contain. How do you predict the literary past? If you could, why bother?

The reason comes back to questions of explanatory sufficiency. Some literary features might appear most significant under statistical analysis, but are those features actually sufficient to differentiate among literary kinds? If, say, words like "detective" and "suspect" are overrepresented in detective fiction, do those words provide, in and of themselves, a trustworthy indication of what we want to know? If not, what statistical properties are sufficient? To answer questions like these, scholars begin by curating a special subset of the corpus, called a *training set*, with all relevant metadata carefully noted. Statistical models of the training set are compiled, then used as a baseline for comparing other documents in the corpus. If the statistical model of a training set accurately predicts the metadata of books not included in it, there's good reason to believe that the model provides a trustworthy and accurate representation of the literary phenomenon at issue. For this reason, the goal of literary prediction is not to predict the future, but to evaluate whether quantitative models adequately represent the qualitative properties they claim to analogize.

---

[74] Dowdy and Wearden (1991) reflect a classical, frequentist perspective, while Bolstad (2007) discusses the Bayesian framework.

In computer and information science, the line of inquiry I've been describing is called *machine learning*. Closely related to Bayesian statistics, machine learning offers a complex theory of how beliefs are tested and how information is incorporated into knowledge.[75] To better integrate these theories into our understanding of humanities computing, we need intellectual histories of statistical theory that are oriented directly to problems of literary-critical and historical explanation. We also need more and better case studies in machine learning to compare how different conceptions of probability and information produce differently interpretable results when tested against cultural data.[76] Most broadly, this line of inquiry tackles a question nestled among the thorniest problems exposed by literary computation: What theories of literature and history are implied by the statistical concepts like bias, probability, and uncertainty, and how might distant-reading projects shed new light on those concepts' fundamental premises? Given what we've learned and will continue to learn about corpora, how should we revise our account of the relation between textuality and actuality?

Conclusion

Literary mathematics names the point of contact between cultural analytics and literary theory, where scholars connect the measurable with the meaningful. Whereas cultural analytics is instrumentalist and results-oriented, literary mathematics is theoretical and concept-oriented. This distinction is neither hard nor fast and would break down if taken too literally — people

[75] Discussions of machine learning's general implications tend to be dominated by futuristic, political, and ethical considerations that are largely outside the concerns of this essay, but questions of interpretation and belief are treated well in the standard textbook on machine learning and artificial intelligence, Russell and Norvig (2010).
[76] This work has begun in several excellent studies: Bamman, Underwood, and Smith (2014), Underwood, Bamman, and Lee (2017), Long and So (2016), and Lee, Greteman, Lee, and Eichmann (2018).

doing distant reading are doing literary math, and vice versa — but it names a difference in emphasis that feels tangible enough.

The central argument of this essay is that digital humanities would benefit from a deeper engagement with quantitative theory. Why would this be valuable? When scholars simply contrast qualitative from quantitative methods — when, for example, they offer a new take on the relation between close and distant reading — they exhibit a strong tendency toward totalizing statements that could not withstand scrutiny when applied to either field individually. Only the most conservative versions of "reading" and "counting" exist in dichotomous opposition. As the survey above makes clear, quantitative methods are quite varied, and so if we hope to make credible interpretations of the corpora now available, we need to develop a shared field of disagreement about which methods are most appropriate for studying which questions.

Critics often ask: What can you *really learn* from computers? Though it seems reasonable enough, I believe this question is based on a false premise. There is no fact, claim, interpretation, or explanation that exists independently of theory and which can therefore be provided by technology. Because of this, anything you really learn from a distant-reading project will depend on a good-faith understanding of its informing conceptual framework. Interpreting quantitative evidence with a purely qualitative theory will rarely produce good thinking. This epistemic problem manifests in digital humanities in two ways: in the form of scholars who naively throw data into software packages, hoping to interpret the resulting visualizations without needing to know what the numbers mean, and in the form of critics who discount everything they don't immediately understand and then complain when all that's left feels too familiar. I call these the Digital Methods Are Just Tools and We Already Knew That fallacies. That both fallacies are in fact fallacies and represent very poor models for critical thinking will, I trust, be obvious to all.

Another goal of this essay has been to provide a theoretical primer that makes explicit some of the intellectual tasks that go into distant reading projects. Often, the best thinking is left implicit or hazily expressed. We need to do a better job of recognizing a wider variety of tangible contributions to scholarly discourse. To this end, I propose the following rubric. (Table 4.) The first two criteria emphasize that designing new topologies for literary history is in and of itself a difficult task that needn't always be relegated to a data and methods section, an appendix, or a footnote. This practice is a form of literary theory and deserves recognition as such. The next two criteria look at how numbers are used. In cultural analytics, scholars can sometimes borrow metrics directly from the sciences, but usually they can't. Humanities computing asks questions that are subtle and strange; often scholars are forced to design new measures or to interpret existing metrics in novel ways. This is fun but challenging work. Lastly, the rubric asks whether the analysis revises our previous, qualitative understanding of the literary topic. Critics too often skip to this last question, but its answer depends on a full understanding of the others, and in any case it shouldn't be taken as a standard unto itself. The most innovative and exciting scholarship will exhibit originality across several criteria, but much valuable research will tackle just one issue at a time.

Table 4. A rubric for evaluating quantitative research in literary history

Questions to ask:	*You know you've really learned something if:*
What literary-critical topic is being addressed?	the topic has not been studied quantitatively before;
How is that topic analogized to a data-curation task?	a new data model has been proposed;
What metrics are used to describe the resulting data?	new metrics have been conceived;
What are those metrics taken to mean?	new general interpretations of existing metrics have been offered; or
How does that new meaning contribute to the original understanding?	our understanding of the literary-critical topic has been revised.

There's another advantage of this rubric. You don't need to know the math. You just need to notice it when it's there, ask for it when it's missing, and insist that it be credibly explained and properly cited. As a rule of thumb, any essay with more graphs in the body than technical secondary sources in the bibliography should be returned for another round of revision. We would all benefit if distant readers did a better job of showing their work.

The other purpose of this essay has been to sketch an outline of mathematical concepts most directly applicable to literary history. I identify three broad categories that correspond, not to particular methods or fields of inquiry, but to general concepts in literary studies: form, difference, and significance. To study *form* under the paradigm of literary mathematics is to discover or invent new ways of describing literary objects. As a critical practice, it involves designing data models that support quantitative analysis while ensuring historical validity. To analyze data is to describe the distribution of *difference* across a model. This usually means converting a discrete representation of a corpus, network, or map into a metric structure over which fine-grained comparisons can be made. In latent space, persons of social networks can be shown to join something like communities, towns can be shown to occupy something like regions, books can be shown to coalesce into something like genres, and words can be shown to indicate something like topics. All such analytical procedures depend on the mathematical terms chosen, and the precise analogies between qualitative cultural phenomena and their quantitative representation will vary, sometimes radically, based on the methods used. Lastly, the *significance* of any literary feature is determined by constructing a detailed picture of its defining categories, then contrasting the feature against those categories. Global metrics like averages, deviations, and entropies show how much variance to expect in a system and expose where that variance is most surprising or, perhaps, most interesting. Taken together, these methods constitute a vast, heterogeneous, and highly sophisticated body of theory that remains

almost wholly unknown to literary scholars but will prove crucial to studying the large-scale digital collections now available.

Scholars hoping to do this kind of work face the same task students face in our courses. We need to imagine anew what it means to be interesting, to relax our minds to new ideas and new pleasures, and to extend a wider scope for literary knowledge.

Works Cited

Ahnert, Ruth, and Sebastian E. Ahnert. 2015. "Protestant Letter Networks in the Reign of Mary I: A Quantitative Approach." *ELH* 82, no. 1: 1-33.

Algee-Hewitt, Mark. 2017. "Distributed Character: Quantitative Models of the English Stage, 1550–1900." *New Literary History* 48.

Allington, Daniel, Sarah Brouillette, David Golumbia. 2016. "Neoliberal Tools (and Archives): A Political History of Digital Humanities." *Los Angeles Review of Books*. May 1.

Alvarado, Rafael C. 2012. "The Digital Humanities Situation." In *Debates in Digital Humanities*. Ed. Matthew Gold. Minnesota.

Axler, Sheldon. 2004. *Linear Algebra Done Right*. Springer.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge.

Bamman, David, Ted Underwood, and Noah Smith. 2014. "A Bayesian Mixed Effects Model of Literary Character." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 370–79.

Barabási, Albert-László. 2016. *Network Science*. Cambridge.

_____. 2011. "The Network Takeover." *Nature Physics* 8: 14-16.

Barabási, Albert-László and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286: 509-12.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *PNAS* 115, no. 18: 4607-12.

Barthélemy, Marc. 2011. "Spatial networks." *Physics Reports* 499, no. 1-3: 1-101.

Bellegarda, Jerome. 2007. *Latent Semantic Mapping: Principles & Applications*. Morgan Claypool.

Berry, Brian J. L. 1968. "Approaches to Regional Analysis: A Synthesis." In *Spatial Analysis: A Reader in Statistical Geography*. Ed. Brian J. K. Berry and Duane F. Marble. Prentice-Hall.

Berry, David. 2011. "The Computational Turn: Thinking about the Digital Humanities." *Culture Machine*, 12: 1-22.

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78, no. 1: 77-106.

Brandes, Ulrik, Garry Robins, Ann McCranie, and Stanley Wasserman. 2013. "What is Network Science?" *Network Science* 1, no. 1: 1-15.

Brennan, Timothy. 2017. "The Digital-Humanities Bust." *Chronicle of Higher Education*. Oct 15.

Broadwell, Peter M. and Timothy R. Tangherlini. 2016. "WitchHunter: Tools for the Geo-Semantic Exploration of a Danish Folklore Corpus." *The Journal of American Folklore* 129, no. 511: 14-42.

Brown, James J., Jr. 2015. "Crossing State Lines: Rhetoric and Software Studies." In *Rhetoric and the Digital Humanities*. Ed. Jim Ridolfo and William Hart-Davidson. Chicago.

Burrows, John. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press.

Burton, Ian. 1968. "The Quantitative Revolution and Theoretical Geography." In *Spatial Analysis: A Reader in Statistical Geography*. Ed. Brian J. K. Berry and Duane F. Marble. Prentice-Hall.

Buurma, Rachel Sagner. 2017. "The Preparation of the Topic Model." *Digital Humanities 2017*. https://dh2017.adho.org/abstracts/332/332.pdf

Carlsson, Gunnar. 2009. “Topology and Data.” *Bulletin (New Series) of the AMS* 46, no. 2: 255–308.

Clark, Stephen. 2015. “Vector Space Models of Lexical Meaning.” In. *The Handbook of Contemporary Semantic Theory*. Ed. Shalom Lappin and Chris Fox. Wiley.

Cooper, David, Christopher Donaldson, and Patricia Murrieta-Flores, eds. 2017. *Literary Mapping in the Digital Age*. Routledge.

Courant, Richard and Herbert Robbins. 1941, repr. 1996. *What is Mathematics? An Elementary Approach to Ideas and Methods. Second Edition*. Revised by Ian Stewart. Oxford.

Cover, Thomas A. and Joy Thomas. 1991. *Elements of Information Theory*. Wiley.

Demšar, Urška, Paul Harris, Chris Brunsdon, A. Stewart Fotheringham, and Sean McLoone. 2013. “Principal Component Analysis on Spatial Data: An Overview.” *Annals of the Association of American Geographers* 103, no. 1: 106–128

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen Renear. 1990. “What is a Text, Really?” *Journal of Computing in Higher Education* 1, no. 2: 3-26.

Dowdy, Shirley and Stanley Wearden. 1991. *Statistics for Research, 2nd Edition*. Wiley.

Dumais, Susan T. 2007. “LSA and Information Retrieval: Getting Back to Basics.” In *Handbook of Latent Semantic Analysis*. Ed. Thomas Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. 293-322.

Edelstein, Dan, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. “Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project.” *The American Historical Review* 122, no.2: 400–24.

Egenhofer, Max J. and Robert D. Franzosa. 1991. “Point-set Topological Spatial Relations.” *International Journal of Geographical Information Systems* 5, no. 2: 161-74.

Egenhofer, Max J. and John R. Herring. 1990. “A Mathematical Framework for the Definition of Topological Relationships.” *Fourth International Symposium on Spatial Data Handling*. Zurich, Switzerland.

Epstein, Joshua M. 2006. *Generative Social Science: Studies in Agent-based Computational Modeling*. Princeton.

Ernest, Paul, Brian Greer, and Bharath Sriraman, ed. 2009. *Critical Issues in Mathematics Education*. Information Age Publishing, 2009.

Fish, Stanley. 1980. “What is Stylistics and Why are They Saying Such Terrible Things About It?” In *Is There a Text in this Class?: The Authority of Interpretive Communities*. Harvard.

Flament, Claude. 1963. *Applications of Graph Theory to Group Structure*. Prentice-Hall.

Foucault, Michel. 1984. “What is an Author?” In *The Foucault Reader*. Ed. Paul Rabinow. Pantheon.

Gavin, Michael. 2016. “Historical Text Networks: The Sociology of Early Criticism.” *Eighteenth-Century Studies* 51, no. 3.

Gavin, Michael and Eric Gidal. 2017. “Scotland’s Poetics of Space: An Experiment in Geospatial Semantics.” *Cultural Analytics*. doi:10.22148/16.017

Gavin, Michael, Collin Jennings, Lauren Kersey, and Brad Pasanek. 2018. “Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading.” In *Debates in Digital Humanities*. Ed. Matthew Gold and Lauren Klein. Minnesota.

Gentner, Dedre. 2002. “Analogy in Scientific Discovery: The Case of Johannes Kepler.” In *Model-Eased Reasoning: Science, Technology, Values*. Ed. Magnani, Lorenzo and Nancy J. Nersessian. Kluwer Academic/Plenum.

Gersting, Judith L. 2002. *Mathematical Structures for Computer Science. Fifth Edition*. W. H. Freeman and Company.

Goldfarb, Charles. 1981. “A Generalized Approach to Document Markup,” *ACM SIGPLAN Notices* 16, no. 6.

Golumbia, David. 2014. “Death of a Discipline.” *Differences: A Journal of Feminist Cultural Studies* 25 no. 1: 156-76.

Good, Phillip I. and James W. Hardin. 2009. *Common Errors in Statistics (and How to Avoid Them) 3rd Edition*. Wiley.

Goodchild, Michael. 1992. “Geographical Information Science,” *International Journal of Geographical Information Systems* 6, no. 1: 31-45.

Gregory, Ian N. and Andrew Hardie. 2011. “Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems.” *Literary & Linguistic Computing* 26, no. 3: 297-314

Greteman, Blaine. 2015. “Milton and the Early Modern Social Network: The Case of the *Epitaphium Damonis*.” *Milton Quarterly* 49, no. 2: 79-95.

Halmos, Paul R. 1960. *Naive Set Theory*. Princeton.

Hersh, Reuben. 1986. *What is Mathematics, Really?* Oxford.

Hesse, Mary. 1966. *Models and Analogies in Science*. Notre Dame.

Homans, George. 1951. *The Human Group*. Routledge and Kegan Paul. CITED IN SCOTT NEED TO REVIEW.

Hoover, David L. 2008. “Quantitative Analysis and Literary Studies.” In *A Companion to Digital Literary Studies*. Ed. Susan Schreibman and Ray Siemens. Blackwell.

_____. 2016. “Argument, Evidence, and the Limits of Digital Literary Studies.” In *Debates in Digital Humanities*. Ed. Matthew Gold and Lauren Klein. Minnesota.

Jay, Jin. 2017. “Problems of Scale in ‘Close’ and ‘Distant’ Reading.” *Philological Quarterly* 96, no. 1.

Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag.

Juola, Patrick and Stephen Ramsay. 2017. *Six Septembers: Mathematics for the Humanist*. Zea Books.

Kienzle, Wolf, Matthias O. Franz, Bernhard Schölkopf, and Gökhan H. Bakir. 2005. "Face Detection — Efficient and Rank Deficient." *Advances in Neural Information Processing Systems* 18: 673-80.

Krioukov, Dmitri, et al. 2010. "Hyperbolic geometry of complex networks." *Physical Review E* 82, no. 3.

Kuhn, Werner. 2005. "Geospatial Semantics." *Journal on Data Semantics III*: 1-24.

Kullback, Solomon, and Richard A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22, no. 1: 79-86.

Laha, R.G. and V.K. Rohatgi. 1979. *Probability Theory*. Wiley.

Landauer, Thomas, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, eds. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Lay, David C. 1999. *Linear Algebra and its Applications. 2nd Edition*. Addison Wesley.

Lee, James Jaehoon, Blaine Greteman, Jason Lee, and David Eichmann. 2018. "Linked Reading: Digital Historicism and Early Modern Discourses of Race around Shakespeare's *Othello*." *Cultural Analytics*. doi:10.22148/16.018

Lee, John M. 2000. *Introduction to Topological Manifolds*. Springer.

Lewis, David D. 1998. "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval." In *Machine Learning: ECML-98*. Ed. C. Nédellec and C. Rouveirol. Springer.

Lindley, Dennis V. 2006. *Understanding Uncertainty*. Wiley.

Liu, Alan. 2013. “The Meaning of Digital Humanities.” *PMLA* 128, no. 2: 409-23.

Long, Hoyt and Richard Jean So. 2016. “Literary Pattern Recognition: Modernism between Close Reading and Machine Learning.” *Critical Inquiry*, 42.

Lowe, Will. 2001. “Towards a Theory of Semantic Space.” *Proceedings of the Annual Meeting of the Cognitive Science Society* 23, no. 23.

Mahalanobis, Prasanta Chandra. 1936. “On the Generalized Distance in Statistics.” *Proceedings of the National Institute of Sciences of India* 2, no. 1: 49–55.

Mandell, Laura. 2013. “How to Read a Literary Visualisation: Network Effects in the Lake School of Romantic Poetry.” *Digital Studies/Le champ numérique* 3, no. 2.

Manovich, Lev. 2016. “The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics.” *Cultural Analytics*. doi: 10.22148/16.004

Martin, Dian I. and Michael W. Berry. 2007. “Mathematical Foundations Behind Latent Sematic Analysis.” In *Handbook of Latent Semantic Analysis*. Ed. Thomas Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. 35-55.

McCarty, Willard. 2005. *Humanities Computing*. Palgrave Macmillan.

_____. 2004. “Modeling: A Study in Words and Meanings.” In *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens, John Unsworth. Blackwell.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice.* Cambridge.

Mendelson, Bert. 1962, repr. 2016. *Introduction to Topology, Third Edition*. Dover.

Mendick, Heather. 2006. *Masculinities in Mathematics*. Open University Press.

Mohr, John and Peter Bogdanov. 2013. “Topic Models: What They Are and Why They Matter?” *Poetics* 41, no. 6.

Moretti, Franco. 2005. *Graphs, Maps, and Trees*.

_____. 2013. *Distant Reading*. Verso.

Munkres, James R. 1984. *Elements of Algebraic Topology*. Addison-Wesley.

Murrieta-Flores, Patricia and Naomi Howell. 2017. “Towards the Spatial Analysis of Vague and Imaginary Place and Space: Evolving the Spatial Humanities through Medieval Romance.” *Journal of Map & Geography Libraries* 13, no. 1: 29-57.

Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford.

Ore, Øystein. 1962. *Theory of Graphs*. American Mathematical Society.

Pachucki, Mark A. and Ronald L. Breiger. 2010. “Cultural Holes: Beyond Relationality in Social Networks and Culture.” *Annual Review of Sociology* 36: 205–24. http://dx.doi.org/10.1146/annurev.soc.012809.102615.

Piper, Andrew. 2016. “There Will Be Numbers.” *Cultural Analytics*. DOI: 10.7910/DVN/MOFE2N

Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Illinois.

Renear, Allen. 2004. “Text Encoding.” In *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens, and John Unsworth. Blackwell.

Robertson, S.E. and Karen Spärck Jones. 1976. “Relevance Weighting of Search Terms.” *Journal of the American Society for Information Science* 27, no. 3: 129-46.

Rockwell, Geoffrey and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces*. PhD Thesis. Stockholm University.

Salton, Gerard and Michael McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw Hill.

Salton, Gerard, A. Wong and C.S. Yang. 1975. "A Vector Space Model for Automatic Indexing," *Communications of the ACM* 18, no. 11: 613-20.

Salton, Gerard and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24, no. 5: 513-23.

Schmidt, Ben. 2016. "Do Digital Humanists Need to Understand Algorithms?" In *Debates in Digital Humanities*. Ed. Matthew Gold and Lauren Klein. Minnesota.

Schwering, Angela. 2008. "Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey." *Transactions in GIS* 12, no. 1: 5-29.

Scott, John. 2012. *Social Network Analysis*. Sage.

Shannon, Claude and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Illinois.

Shick, Paul L. 2007. *Topology: Point-Set and Geometric*. Wiley.

Shore, Daniel. 2018. *Cyberformalism: Histories of Linguistic Forms in the Digital Archive*. Johns Hopkins.

Silverstein, Michael and Greg Urban. 1996. "The Natural History of Discourse." In *The Natural History of Discourse*. Ed. Michael Silverstein and Greg Urban. Chicago.

Smith, David, Ryan Cordell, and Abby Mullen. 2015. "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers." *American Literary History* 27, no. 3.

Snow, C. P. 1959, repr. 1998. *The Two Cultures*. Cambridge.

So, Richard Jean. 2017. “All Models Are Wrong.” *PMLA*.

Sperberg-McQueen, C. M. 1991. “Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts,” *Literary & Linguistic Computing* 6, no. 1: 34-46.

Steyvers, Mark and Tom Griffiths. “Probabilistic Topic Models.” In *Handbook of Latent Semantic Analysis*. Ed. Thomas Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. 427-48.

Strogatz, Steven H. 2001. “Exploring Complex Networks.” *Nature* 410, no. 6825: 268-76.

Tao, Terence. 2011. *An Introduction to Measure Theory*. American Mathematical Society.

Turney, Peter D. and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37: 141-188.

Underwood, Ted. 2017. “A Genealogy of Distant Reading.” *DHQ: Digital Humanities Quarterly*.

_____. 2012. “Topic Modeling Made Just Simple Enough.” *The Stone and the Shell*. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Underwood, Ted, David Bamman, and Sabrina Lee. 2018. “The Transformation of Gender in English-Language Fiction.” *Cultural Analytics*. doi: 10.7910/DVN/TEGMGI

Underwood, Ted and Jordan Sellers. 2012. “The Emergence of Literary Diction.” *Journal of Digital Humanities* 1, no. 2.

van Hooland, Seth and Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums*. Facet Publishing.

van Mieghem, Piet. 2011. *Graph Spectra for Complex Networks*. Cambridge.

Watts, Duncan J. and Steven H. Strogatz . “Collective Dynamics of ‘Small-World’ Networks.” *Nature* 393: 440-42.

Widdows, Dominic. 2004. *Geometry and Meaning*. CSLI Publications.

Wilkens, Matthew. 2013. “The Geographic Imagination of Civil War-Era American Fiction.” *American Literary History* 25, no. 4: 803-40

Witmore, Michael and Jonathan Hope. 2016. “Books in Space: Adjacency, EEBO-TCP, and Early Modern Dramatists.” *New Technologies in Medieval and Renaissance Studies*, 6.