Theorizing Research Practices We Forgot to Theorize Twenty Years Ago

Ted Underwood, University of Illinois

Humanists are gearing up to have a conversation about digital research methods that will be interesting for many reasons — not least, because it's oddly belated. Algorithmic mining of large electronic databases has been quietly central to the humanities for two decades. We call this practice "search," but "search" is a deceptively modest name for a complex technology that has come to play an evidentiary role in scholarship. Many of the features that seem new to us about data mining (its "bigness" or quantitative character, for instance) have been invisibly naturalized in our disciplines since humanists started using full-text search in the 1990s. Although data mining is widely framed as a novel technology now being imported to the humanities, I'll argue that it is better understood as a philosophical discourse that can help humanists think more rigorously and deliberately about existing practices of algorithmic research.

First, what does it mean to say that search plays an "evidentiary role in scholarship"? The appearance of paradox here is partly produced by the word *search* itself, which blurs boundaries between distinct technologies. Bibliographic search can be little more than an aid to memory — for instance, if a scholar is recovering the call number for a known title. Full-text search looks similar: we may even enter search terms in the same box where we would have entered a title. But the underlying technology, and its scholarly applications, are different.

In practice, a full-text search is often a Boolean fishing expedition for a set of documents that may or may not exist. For instance, if I suspect that blushes are symbols of moral consciousness in nineteenth-century poetry, I can go to a database of primary sources and search for poems that contain both *blush* and *conscious*. If I find enough examples, I flesh out an article. If not, I usually keep trying until I succeed. Perhaps *blush* and *shame* would work better? Here search is not just a finding aid; it's analogous to experiment — although, to be sure, there's something a bit dubious about experiments that get repeated until they produce a desired result. The search terms I have chosen encode a tacit

Accepted for publication in a *Representations* special forum to appear later in 2014.

hypothesis about the literary significance of a symbol, and I feel my hypothesis is confirmed when I get enough hits. It's possible that the article I finally write will discuss only a few of these sources, because I may not believe that the problem requires "big data." But in fact I've used algorithms to explore a big dataset, and the search process may well have shaped my way of framing the subject, or my intuitions about the representativeness of sources.

The internal mathematics of full-text search also have more in common with data mining than with bibliographic retrieval. If I do a title search for *Moby-Dick*, the results are easy to scan. But in full-text search, there are often too many matches for the user to see them all. Instead, the algorithm has to sort them according to some measure of relevance. Relevance metrics are often mathematically complex; researchers don't generally know which metric they're using; in the case of web search, the metric may be proprietary.

In short, full-text search is not a finding aid analogous to a card catalog. It's a name for a large family of algorithms that humanists have been using for several decades to test hypotheses and sort documents by relevance to their hypothesis. Simple forms of full-text search were already available in the 1970s (LEXIS was an early example), but CD-ROM databases of historical sources weren't widely distributed until the 1990s. Even today, the technology may not have permeated the discipline of history as deeply as it has literary studies, since historians rely more heavily on unpublished sources. One recent study suggests, however, that humanists across a range of disciplines rely heavily on search engines, and use them for research in ways that are not very different from the general public. (Like everyone else, we begin with Google.)[1] The scholarly consequences of search practices are difficult to assess, since scholars tend to suppress description of their own discovery process in published work.[2] But as someone who began

1 Max Kemman, Martijn Kleppe, and Stef Scagliola. "Just Google It — Digital Research Practices of Humanities Scholars." CoRR 2013. arXiv:1309.2434 (2013).

2 The invisibility of search practices is related to a pattern Lisa Gitelman has observed: "media become authoritative as the social processes of their definition and dissemination are ... forgotten." *Always Already New: Media, History, and the Data of Culture* (Cambridge, MA: MIT Press, 2006), 7.

a dissertation just before full-text databases became available, I remember that I seemed to be finishing it in a different world.

The most obvious effect of the new technology was that, like many other literary scholars in the 90s, I found myself writing about a wider range of primary sources. But I suspect that the questions scholars posed also changed to exploit the affordances of full-text search. Before 1990, narrowly-defined themes were difficult to mine: there was no Library of Congress subject heading for "descriptions of work as 'energy' in British Romantic-era writing." Full-text search made that kind of topic ridiculously easy to explore. If you could associate a theme with a set of verbal tics, you could suddenly turn up dozens of citations not mentioned in existing scholarship, and discover something that was easy to call "a discourse." I remember feeling uneasy about this. The rules of the research game seemed to have changed in a way that made it impossible to lose. After all, how many sources do you need to establish the importance of a theme? Twenty? When searches were limited by networks of previous citations, that was a meaningfully high bar. But in a database containing millions of sentences, full-text search can turn up twenty examples of anything. Even at the time, it was clear that this might strengthen confirmation bias.

In hindsight, I underestimated the scope of the problem. It's true that full-text search can confirm almost any thesis you bring to it, but that may not be its most dangerous feature. The deeper problem is that by sorting sources in order of relevance to your query, it also tends to filter out all the alternative theses you didn't bring. Search is a form of data mining, but a strangely focused form that only shows you what you already know to expect. This limitation would be a problem in any domain, but it's particularly acute in historical research, since other periods don't always organize their knowledge in ways we find intuitive. Our guesses about search terms may well project contemporary associations and occlude unfamiliar patterns of thought.

Humanists didn't spend a lot of time debating this problem in the 1990s, because search engines were usually our only mode of access to large digital collections. But in recent years, research practices have diversified, and the hermeneu-

tic limitations of search are becoming obvious. In computer science, the subfields of data mining and machine learning have specialized in the problem of extracting knowledge from large collections.[3] They've come up with a range of alternatives to search based on a more self-conscious, philosophically rigorous account of interpretation.

I realize that last sentence may be an eye-opener. Humanists tend to think of computer science as an instrumental rather than philosophical discourse. The term "data mining" makes it easy to envision the field as a collection of mining "tools." But that's not an accurate picture. The underlying language of much data mining — Bayesian statistics — is a way of reasoning about interpretation that can help us approach large collections in a more principled way.[4] In particular, it emphasizes a hermeneutic spiral that will be familiar to humanists, acknowledging that we approach every question with some previous assumptions (called "prior probabilities"), as well as particular kinds of uncertainty. When we encounter new evidence, our interpretation is at once shaped by existing assumptions, and (possibly) capable of reshaping them. This hermeneutic cycle is intuitive enough when we're talking about a single text; the task of data mining is to explain how it can work at the level of a collection too large to be surveyed by a single reader. All mapping strategies are going to make some assumptions about the patterns we expect to find. But some strategies are also able to reveal evidence that challenges prior assumptions.

For instance, literary scholars' habit of using keyword search to probe for intersections of themes (like *blush/shame* or *work/energy*) is tacitly based on a assumption that the co-occurrence of words will reveal a connection between their meanings. This assumption is related to a model of meaning that linguists call the "distributional hypothesis," which postulates that the meaning of a word

3 For a brief history of data mining see Frans Coenen, "Data Mining: Past, Present, and Future," *The Knowledge Engineering Review* 26 (2011): 25-29.

4 For a philosophical approach to this topic see Luc Bovens and Stephan Hartmann, *Bayesian Epistemology* (Oxford, 2003). A practical introduction can be found in John K. Kruschke, *Doing Bayesian Data Analysis* (Burlington, MA, 2011).

is related to its distribution across contexts.[5] This may not be a perfect model, but it has proven to be a useful one in computer science as well as literary study, and if we want to continue using it as a heuristic, there are more flexible ways to use it than iteratively guessing particular pairs of words. Algorithms based on distributional assumptions can map the language that was in practice associated with any term in a given period.[6] For instance, the word most commonly associated with *blush* in a collection of 4,820 eighteenth- and nineteenth-century volumes turns out to be not *shame* but *artless* — a detail that might interestingly complicate a scholar's assumptions about moral consciousness, if they use an exploratory strategy flexible enough to reveal it.[7] Mapping strategies like this won't replace keyword search for all purposes. When you already know what you're looking for, a search engine is the appropriate tool. But in historical scholarship, there are times when we don't know what we're looking for as well as we think.

In fact, perhaps it's already hasty to assume that the topic I'm exploring can be associated with a single word like *blush*. Maybe a different term, that I can't begin to guess, was more important in this period, or maybe the social phenomena relevant to my question take shape at the intersection of many different terms. If we want a more open-ended strategy, we can map the print record by allowing an algorithm to organize the language of a collection into clusters of terms that tend to occur in the same contexts. This strategy (known as "topic modeling") is capable of revealing discursive patterns that the researcher didn't necessarily go looking for.[8]

---

5 Magnus Sahlgren, "The Distributional Hypothesis," *Rivista di Linguistica* 20 (2008): 33-53.

6 Peter D. Turney and Patrick Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (2010): 141–188.

7 I produced this result by measuring the "cosine similarity" of word distributions over a collection of 4,820 volumes that Jordan Sellers and I assembled, with assistance from TCP-ECCO and the Brown Women Writers Project. Measuring cosine similarity is a simple approach; there are more sophisticated ways of using a distributional model to assess similarity between terms. (See Turney and Pantel, cited above.) Of course no single collection of volumes is perfectly representative of print culture; in practice, the best way to address questions of representativeness is often to pose the same question in multiple collections that have been selected in different ways.

8 "Topic modeling" is a name for a large family of algorithms, but the algorithm most commonly used by humanists is Latent Dirichlet Allocation. For an accessible humanistic introduction, see

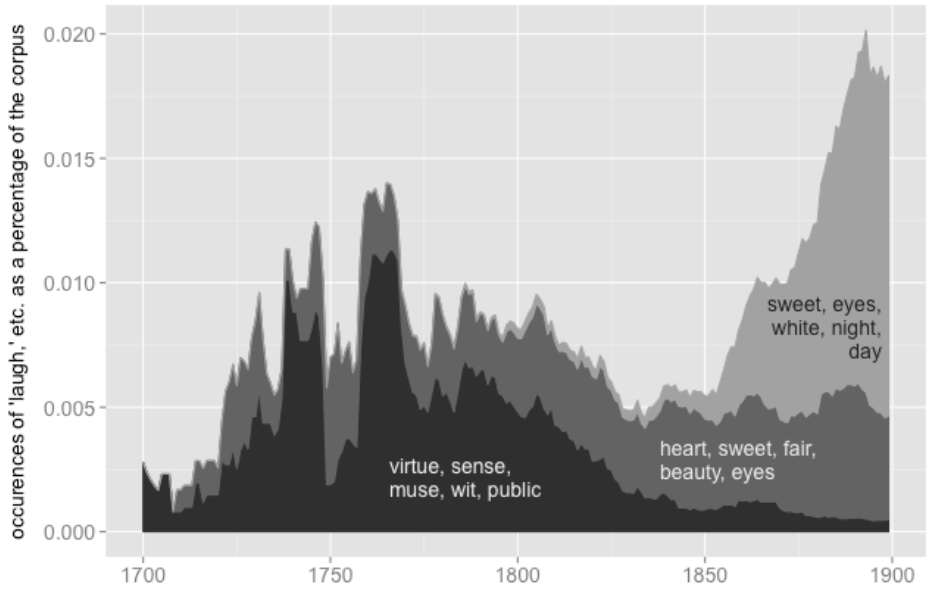

Figure 1: Occurrences of *laugh, laughter,* etc. in a collection of 13,789 volumes of poetry, divided by the topic each occurrence was assigned to. Among 120 topics, I have plotted the three where *laugh-* occurs most often; each topic is labeled with its most frequent words.

Because topic modeling allows a word to belong to more than one "topic," it can reveal patterns of association that shift across time. Figure 1, for instance, plots occurrences of *laugh* (and words derived from that root) in eighteenth- and nineteenth-century poetry, dividing the occurrences by their association with three different algorithmically-created topics.[9] It would be possible to consider each topic separately — in fact, that's how topic modeling is commonly used —

Ted Underwood, "Topic Modeling Made Just Simple Enough," *The Stone and the Shell,* April 7, 2012. http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/. For a more technical account, see David Blei, "Probabilistic topic models," *Communications of the ACM*, 55(2012): 77–84.

9 These 13,798 volumes of poetry were selected from a larger set of 469,000 eighteenth- and nineteenth-century volumes in HathiTrust Digital Library. To identify the poetry in this large collection, I used tools for genre mapping described in Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu, "Mapping Mutable Genres in Structurally Complex Volumes," 2013 IEEE International Conference on Big Data, pp. 95-103.

but here I've added an additional twist by showing how references to laughter are so to speak passed from one topic to another over time. (The algorithm also created 117 other topics; these only are the three where *laugh-* occurred most often.) Each topic is labeled with its most common words, giving us a sense of the changing contexts where poets mention laughter. A contrast is visible between the public, satirical function of laughter in much eighteenth-century poetry, and a different pair of contexts where laughter is associated with personal description of a sentimental or amatory kind ("sweet," "fair," "eyes"). This use of laughter for characterization is already present in the eighteenth century, but becomes more prominent as the association of laughter with public wit fades.

I don't mean to imply a causal connection between these changes (for one thing, there are many other topics in the model; these three don't constitute a closed system). The illustration is only meant to show how topic modeling can generate suggestive leads. But pursued in more depth, leads become results. Matthew Jockers has used topic modeling to map nineteenth-century novels; Robert K. Nelson has used it to correlate thematic emphases in a Civil-War-era newspaper with the changing fortunes of the war. Andrew Goldstone and I have used the technique to chart the rise and fall of different critical vocabularies in twentieth-century literary study.[10]

Instead of dwelling specifically on topic modeling, I want to consider the way innovations of this kind are prompting a belated conversation about algorithmic exploration in general. Topic modeling will be and should be controversial — as full-text search, actually, should have been controversial twenty years ago. Researchers can never afford to treat algorithms as black boxes that generate mysterious authority. If we're going to use algorithms in our research, we have to crack them open and find out how they work. Topic modeling, fortunately,

10 Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana, 2013). Robert K. Nelson, "Mining the Dispatch," http://dsl.richmond.edu/dispatch/. Andrew Goldstone and Ted Underwood, "What can Topic Models of PMLA Teach Us About the History of Literary Scholarship?" *Journal of Digital Humanities* 2.1 (2012). http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/ (A much-expanded version of this project is forthcoming in *New Literary History.*)

is not proprietary, like many algorithms behind web search. Topic modeling algorithms are public, and humanists have proven to be quite capable of understanding them and changing them to fit humanistic goals.

To understand the interpretive limitations of an algorithm, you need to understand its mathematical basis. For instance, in the most common form of topic modeling, the number of topics to be produced is one of the initial assumptions you bring to the modeling process. As a consequence, the algorithm can't provide authoritative answers about the unity of any discourse, or about its boundaries. It's always possible to model the same collection with a larger or smaller number of topics, which would lump or split results differently. On the other hand, the algorithm is quite good at revealing patterns of association we might otherwise overlook.

Using algorithms for discovery raises an interesting but unfamiliar set of philosophical questions. Humanists are still more comfortable with quantitative methods when they can be presented in their familiar role as instruments of verification in the late stages of research. Using an algorithm as a source of initial leads seems perilously close to pulling a rabbit out of a hat (in spite of the fact that we've been doing this with search engines for several decades). In a recent issue of *PMLA*, for instance, Alan Liu wonders whether topic modelers aspire to the goal of "tabula rasa interpretation — the initiation of interpretation through the hypothesis-free discovery of phenomena."[11]

If this were true, it would create a real philosophical impasse. And one can certainly find technophilic rhapsodies in *Wired* magazine suggesting that we have reached that impasse: an endgame where "data" finally displaces all "theory."[12] But those rhapsodies are not well informed about the statistical models involved in data mining. It isn't the case that topic modeling (or any other data mining algorithm) pretends to be truly "hypothesis-free." A model is an abstraction created by human beings, and computer scientists have long acknowledged

11 Alan Liu, "The Meaning of the Digital Humanities," *PMLA* 128 (2013): 414.

12 Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired,* June 23, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

this.[13] The Bayesian probabilistic models now common in the discipline are especially meticulous about specifying initial interpretive assumptions.

A researcher who wants to fit a topic model to a collection of documents has to start by specifying, for instance, the number of topics she expects to find, and the degree of blurriness she expects those topics to possess. In the modeling process, the computer doesn't generate insights from nothing; its calculations are rather a way of harmonizing these initial human assumptions with the complex evidence presented by the documents themselves. (To use a term of art, the computer helps us "fit" a model to the evidence.) This mode of exploration can be more open-ended than keyword search, since assumptions about degrees of blurriness are more flexible than a specific assumption that, say, blushes will symbolize shame. But the interpretive process is still shaped and initiated by human assumptions.

I haven't had room here to make a detailed argument about the humanistic value of quantitative methods.[14] But doing that would be almost beside the point, since humanists are already mining large datasets quantitatively every time we use a web browser. The problem is that we are using search algorithms we have never theorized, and arguably using them in a strongly projective way at odds with historicism. Although the statistical language of computer science may seem alien to our disciplinary tradition, I think the paradoxical truth is that humanists will need to understand that language in order to design research practices that allow us to work in large collections while remaining true to our own hermeneutic principles.

This is admittedly a new kind of interdisciplinary conversation for humanists, and we may initially have a lot to learn. But we also have a lot to contribute. I've suggested that quantitative disciplines have their own useful ver-

13 "[F]undamentally, computer science is a science of abstraction — creating the right model for thinking about a problem ..." Alfred Aho and Jeff Ullman, *Foundations of Computer Science* (New York, 1994), 1.

14 For more examples of topic modeling, see a special issue of *Poetics* on "Topic Models and the Cultural Sciences" 41.6 (2013). See also Lisa M. Rhody, "Topic Modeling and Figurative Language," *Journal of Digital Humanities* 2.1 (2012): http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

sion of hermeneutic theory, but they aren't without blind spots. The difficulty of modeling historical change, for instance, is not well understood outside the humanities. Scientists who try to model the print record over a significant timespan often make assumptions about continuity that humanists would recognize as confining.[15] On this topic, and many others, a rare opportunity is emerging for a genuinely productive exchange between scientific methodology and humanistic theory.

15 Attempts to frame explicitly diachronic versions of topic modeling (like Dynamic Topic Modeling and Topics Over Time) have tended to invoke dubious assumptions about historical continuity. Historians are probably better advised to rely on a simpler algorithm like Latent Dirichlet Allocation, which remains blissfully ignorant of dates and yet in practice tends to produce coherent diachronic patterns. See the appendix to Benjamin M. Schmidt, "Words Alone: Dismantling Topic Models in the Humanities," *Journal of Digital Humanities* 2.1 (2012). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/