LDA Topic Modeling: Contexts for the History & Philosophy of Science

Colin Allen & Jaimie Murdock

Forthcoming in Ramsey, G., De Block, A.(Eds.) *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science.* Pittsburgh University Press; Pittsburgh.

§1 Introduction

In their introduction to a special issue of the *Journal of Digital Humanities* in 2012 about topic modeling for the digital humanities, the editors Elijah Meeks and Scott Weingart began by lampooning the state of the field with its obscure jargon and self-inflicted wounds. We believe in the promise of topic modeling as a research tool for the humanities in general, and for history & philosophy of science in particular. However, as we shall argue in this essay, to realize this potential and minimize additional self-harm requires a shift in the way that topic models have been used and discussed, moving away from a word-centered conception of topics and toward a document- and context-centered conception of the models. The potential of topic modeling makes it worth mastering the jargon and developing an appreciation for the inside jokes in the introduction by Meeks and Weingart:

> Topic modeling could stand in as a synecdoche of digital humanities. It is distant reading in the most pure sense: focused on corpora and not individual texts, treating the works themselves as unceremonious "buckets of words," and providing seductive but obscure results in the forms of easily interpreted (and manipulated) "topics." In its most commonly used tool, it runs in the command line. To achieve its results, it leverages occult statistical methods like "dirichlet priors" and "bayesian models." Were a critic of digital humanities to dream up the worst stereotype of the field, he or she would likely create something very much like this, and then name a popular implementation of it after a hammer.

The "command line" to which Meeks & Weingart refer is that text-only interface to the computer's operating system that is modeled on computer terminals from the 1970s, and whose operation manuals bear some resemblance to a book of spells. The "hammer" to which they refer is the popular implementation of Latent Dirichlet Allocation (LDA) topic modeling named 'MALLET'. To many, LDA topic modeling has indeed seemed like a blunt instrument whose significance is obscure to all but a cult of magicians. How, after all, could a technique that begins by throwing away all the syntactic information in language yield anything about the meanings of text? In this bag-(or bucket)-of-words approach, "man bites dog" is indistinguishable from "dog bites man" and thus the surprising is rendered indistinguishable from the banal. Furthermore, for reasons of computational tractability and model interpretability, the highest frequency words are eliminated before the models are constructed (aka "stoplisting"). Gone are the most common prepositions, conjunctions, and pronouns, as well as important operators such as 'not'. Thus, not only is the surprising reduced to the banal, but exact opposites of meaning collapse to a single representation.

It is hard to see how such a seemingly blunt instrument could be of interest to scholars in the humanities. Philosophers and historians of science may be forgiven for being especially skeptical given their concerns with the subtleties of scientific reasoning and explanation, and the shifts in meaning and understanding that follow theoretical change, but analogous concerns arise for any other humanities discipline. The initial applications of topic modeling rather reinforced the idea that they provide little to sustain the interest of anyone interested in detailed understanding of texts or intellectual history, whether literary or scientific. As Meeks & Weingart put it, given a topic model, "You would marvel at the output, for a moment, before realizing there isn't much immediately apparent you can actually do with it". Nearly a decade later, much has been done with topic models, but they still tend to perplex all but the innermost circle of practitioners. We surmise, however, that the difficulty of seeing what can be done with topic models is partly a product of some common misconceptions about how they are conjured, and partly a product of how they are typically presented.

Our aim in this paper is to illustrate through our own work the application of topic modeling to questions that interest historians & philosophers of science, going beyond simplistic presentations that tend to give scholars the idea that the algorithms produce results that are superficial and perhaps unreliable. The facts about topic models and the ways in which they are often misrepresented and misunderstood frame our attempt in this chapter to convince readers that, despite appearances, that LDA topic modeling provides a lot more of value to HPS research than merely providing for enhanced search and information retrieval from large sets of documents. There is much room for the interplay between human intelligence and sophisticated algorithms to expand the range of questions about science that HPS scholars will ask, and can answer. Although it is impossible to assess the approach without some insight into the workings of the algorithms, we also believe the case for their use can be made and understood without first gaining expert-level understanding. Further below we will provide a brief introduction to LDA topic modeling, but more detailed introductions can be found in various places, including the contributions to the aforementioned special issue of the *Journal of Digital Humanities* edited by Meeks & Weingart (see especially David Blei's contribution to that issue, Blei 2012a, as well as Blei 2012b).

Our take on topic models extends to dissatisfaction with the term "topic modeling" itself and urges a reorientation to documents and the contexts in which they are written. Although not ideal, a better label might have been "context modeling". Such a relabeling would, nevertheless, help to avoid the ordinary connotations of the term 'topic', which suggests something that could be a title for a lecture, a course, or a thesis: 'veterinary medicine in the Andes', for example, or the 'quantum states of electrons'. The implicit but oft-articulated story behind topic models is that they provide a (very partial) theory of the writing process (Boyd-Graber et al. 2017). The texts that authors produced typically combine a few topics. A treatise on veterinary medicine in the Andes is likely, for example, to touch upon some related topics such as physiological adaptations to altitude, or the state of veterinary education in South America, but may also digress into geology or meteorology. From this perspective, topic modeling provides an account of how documents are generated by selecting among words associated with the multiple topics.

We do not challenge the idea that topic models provide a theory about writing. However, by recasting them as context models, we think we get a better account of the relevance of the models to writing, as we shall explain in §3 below.

Although our arguments are addressed explicitly to scholars in the history & philosophy of science (HPS), we believe they generalize to topic modeling in other humanities disciplines, including history and literary studies. It is worth acknowledging, however, that some of our interpretive concerns concerning topic models will seem more relevant to those with historical interests than those with primarily literary concerns.[1] We also will make the case that philosophers of science are particularly well placed to make a contribution to the interpretive questions because of their attention to models in science

§2 On the Use and Abuse of Topic Models

A frequent target of topic modelers — the low-hanging fruit as it were — has been the back issues of scholarly journals, identifying the changing distribution of topics over time (e.g., for *Science* by Blei & Lafferty 2006, 2007; for *Cognition* by Cohen Priva & Austerweil 2014; for the *Journal of the History of Biology* by Peirson et al. 2017; for the Proceedings of the Cognitive Science Society by Rothe, Rich, and Li 2018; and for *Philosophy of Science* by Malaterre et al. 2020). Similar projects have been pursued with other temporally-sequenced datasets (see Brauer & Fridlund 2013 for an early review), from parliamentary debates in France (Barron et al. 2018) and Britain (Guldi & Williams 2018; Guldi 2019a,b;) to the 18th C. *Encylopédie* (Roe et al. 2016) and 19th Century novels (Jockers 2013). As tantalizing as the prospect might be that this technique would reveal something novel about the history of ideas, the presentations of the distributions and fluctuations of topics uncovered in this way rarely seem to go far enough. Part of the problem concerns the highly variable intelligibility of the word distributions identified as "topics" by the LDA algorithm. The flexible way in which people understand these so-called topics has been analogized to "reading tea leaves" (Chang et al. 2009), and the tendency of topic modelers to use relatively short lists of 10-20 words to represent each topic exacerbates the difficulty of coming to a good understanding of the models.

Our intermittent use of scare quotes around "topics" is intended to flag places where over-interpretation looms. Technically, what LDA outputs are not *topics* as English speakers would primarily understand that term. Rather, each "topic" is a total probability distribution over the vocabulary in the corpus — that is, the sum of all the word probabilities within one topic is equal to 1 — and every word is assigned a non-zero probability in every topic (albeit that most words are assigned a vanishingly small probability in most topics). Simultaneously, each document is represented as a total probability distribution over the topics, and every topic is assigned a non-zero probability in every document, albeit skewed to relatively few topics. The sum of the topic probabilities within one document is likewise equal to 1. The number of topics is chosen by the modeler, but their content is not. The model is initialized with random probabilities assigned to

[1] Thanks to Jo Guldi for this observation.

the word-topic and topic-document distributions. It is only through an iterative training process that updates these distributions that anything interpretable emerges. Specifically, the models are trained by a Bayesian process which tests document-word distributions generated from the model against the observed distributions sampled from the documents, and concurrently adjusts the word-topic and topic-document probability distributions so as to be capable of better matching the word distributions found in the actual documents. The probability assignments in the models become stable with repeated training passes through the full corpus, making it reasonable to terminate the training after a few hundred iterations of this process.

The shapes of the word-topic and topic-document distributions are controlled by two parameters — technically “hyperparameters” or “priors” on the Dirichlet distribution (named for the 19th C. mathematician Gustav Dirichlet) — that are also chosen by the modeler. These hyperparameters skew the algorithm towards producing word-topic and topic-document distributions that have most of their probability mass (aka “weight”) concentrated in relatively few of the words and topics assigned to topics and documents respectively. As Blei (2012a) explains, the choice of the hyperparameters represents a trade-off: “On both topics and document weights, the model tries to make the probability mass as concentrated as possible. Thus, when the model assigns higher probability to few terms in a topic, it must spread the mass over more topics in the document weights; when the model assigns higher probability to few topics in a document, it must spread the mass over more terms in the topics.” The practical upshot here is that when topics are heavily loaded on a few words, they will be less successful at accounting to the words in any given document, so more topics will need to be assigned to that document to account for its word distribution, but this runs counter to the imperative to load documents with relatively few topics. In the extreme, imagine a topic that puts nearly all of its probability mass on one word – e.g. ‘lion’ -- another that skews equally heavily on ‘tiger’, a third on ‘elephant’, etc. A document containing normal mixture of words – “Lions and tigers mostly avoid elephants in Africa and India respectively. …” -- would need low weightings on lots of such skewed topics to represent its actual word distribution well. But distributing the probability mass over lots of topics is not compatible with the hyperparameter setting that favors distributing the probability mass over fewer topics per document, so the training process must compromise by assigning some of the probability mass to more words in each topic.

Given reasonable and typical selections for these hyperparameters, even if the model is trained with as many topics as there are documents, LDA assigns a mixture of topics to each document, although many of these topics will be relatively hard to interpret. With too many topics, some of the topics are specialized on just a few documents, also making them less useful for discovering relationships in the wider corpus. When training models with considerably fewer topics than documents the LDA process achieves generalizable and interpretable results via a form of data compression. However, with too few topics, each topic becomes very general and less useful for identifying informative relationships among the documents. While statistical methods exist for computing the number of topics which give the best statistical fit for a given corpus, scholars and other human users may prefer a coarse-grained scheme (fewer topics) for some purposes while preferring a more fine-grained scheme (more topics) for other purposes. Furthermore, the value of more specialized topics may be more apparent to domain experts

than to other users. The fact that statistical fit of the topic model to the documents in the corpus does not correlate with user judgments about topic quality was the point of the 'tea leaves' paper by Chang et al. (2009).

Many articles report the results of topic modeling in ways that drive misunderstanding of the models. First, it is typical to display only a subset of the topics found by LDA, and only the most readily interpretable, thus making the overall model seem more interpretable than it would otherwise seem. The embarrassment of so-called "junk" and "jargon" topics — i.e., topics that are hard to interpret — is thereby swept under the rug. For example, Malaterre et al. 2020 write: "47 [of the 200 topics] appeared to be either too generic or polysemic to be precisely related to any meaningful issue in philosophy of science. We therefore grouped these 47 topics under the label 'Jargon' and set them aside." Similarly, albeit with different goals, Lambert et al. (2020) report that they "studied the *British Medical Journal* between 1960 and 2008, identifying 100 topics using latent Dirichlet allocation, which we filtered for those directly concerned with clinical practice or medical research using the words most highly associated with each topic, leaving us with 73 topics." To be fair, these researchers understand that their decision to omit certain topics from their analyses is driven by their particular explanatory interests, and that different interests might entail making the distinctions between "junk" and "jargon" more precise, with the former perhaps providing a guide to better corpus preparation prior to modeling and the latter proving useful for the genre or style analyses. Our point is only that the practice of ignoring such topics serves to reinforce a fraught strategy that centers on directly assigning meaning to topics.

Second, by showing only the ten or so highest-weight words for each topic, such presentations neglect most of the words that contribute to the topics' roles in representing the corpus documents. For example, in the 200-topic model that we constructed from 665 non-fiction English-language books read by Charles Darwin between 1837 and 1860 (Murdock, Allen & Dedeo 2017), typically 500-600 words are required to account for 50% of the probability mass for any given topic. Looking only at the first ten or twenty words may provide little understanding of why that topic has been assigned a high weight for a given document.

Third, this limited way of presenting the topics also leads readers who don't fully understand LDA to the incorrect assumption that documents are assigned a high proportion of a given topic because they contain all the words listed as "in" the topic. While it is indeed somewhat likely that a high probability word from a given topic appears in any document for which the topic is highly weighted, it is not guaranteed. One of the strengths of topic modeling is that thematically similar documents may be assigned similar topic profiles despite considerable differences in the vocabulary they contain. Another source of misunderstanding is that phrase "words in the topic" is easily taken to encompass all and only the words that the authors have presented for each topic.

Analyses of topic models that focus on interpretation of the topics often distort a fundamental feature of topic modeling: it is not a discriminative model (i.e., one which sorts entities into distinct categories), but rather a mixed membership model (Airoldi et al., 2014): i.e., a document does not "belong" to a single topic, nor does a "word". The distributions themselves,

independent of any interpretation of their components, can illuminate a collection of documents. For example, in our analysis of Darwin's readings we did not select a subset of meaningful topics and analyze only those. Rather, we analyzed the time sequence of Darwin's reading choices in respect to fluctuations in an information-theoretic distance measure of "surprisal" (i.e., Kullback-Leibler divergence, see below) applied to the entire topic distribution, junk, jargon, or otherwise. Indeed, it is worth emphasizing that the machine itself has no understanding, and thus to it, all the topics are on a par, whether or not they make sense to a human reader who comes to them with the bias that "topics" must make sense. Those who work with topic models should, as Binder (2016) says, "resist attempts to present computational results in forms that readily appeal to our assumptions and intuitions about language."

It seems important to emphasize here that no currently available form of artificial intelligence or machine learning can supply the meanings that matter for genuine scholarship and understanding (many have made this point, but see Ravenscroft & Allen 2019 for specific discussion of the strengths and weaknesses of LDA for argument-based analysis in HPS; for general discussion of the limits of current AI/ML, see Mitchell 2019; Smith 2019; Marcus & Ernest 2019). For the foreseeable future, computers will remain mere tools for humans to use creatively. But the stupidity of AI/ML does not render it useless.

§3 Topics as Contexts

The shift to thinking of the "topics" in a topic model as representing contexts helps deal with the problems outlined above in various ways. First, it helps us to reconceptualize the issue of junk topics. The assumption that LDA finds topics in the ordinary English sense practically forces one into interpretive mode, so that for topics where there is no easily available interpretation, it is natural to label them as "junk". For instance, here are the ten highest probability words from two "topics" in an 80-topic LDA model of the 881 letters selected by Randolph (1829) from among the thousands written by Thomas Jefferson, author of the U.S Declaration of Independence in 1776, the United States' first secretary of state from 1790-1793, and its third president from 1801-1809:

> vessels, war, British, vessel, port, Britain, sea, peace, enemy ships.
>
> honor, respect, obedient, humble, servant, also, sentiments, esteem, take, think.

Setting aside for the moment the concern about guessing the content of a topic from just ten words, the first of these suggests an obvious interpretation, one that is borne out by it being assigned the most weight in a letter concerning British seizure of a ship inbound to the U.S. from the French West Indies, written in 1792 by Jefferson, while he was secretary of state, to Edmond-Charles Genêt, who was then the French envoy to the United States. The second "topic" consists of words one might expect to find in letters, but is not obviously topical in the ordinary sense. It turns out that this particular "topic" is most represented in the letters Jefferson wrote to various diplomats during his term as Secretary of State in the early 1790s, especially to the British envoy George Hammond, but also in many of his letters to Genêt. These are not just

any old letter words, but words more likely to be used in the specific context of writing letters to diplomats.

Writing takes place in historical contexts which influence the words selected. The contexts of writing may include topics in the ordinary sense (i.e., the subjects addressed in the writing) as well as the situation of the writer in the historical moment, as influenced by the particular networks of family, friends, colleagues, and culture at large, and the author's roles in institutions and society more broadly. Different writing contexts entail different audiences: letters to friends and family vs. business associates or diplomats, philosophical treatises, public speeches, scientific publications, etc. Each of these contexts changes the likelihood of the author selecting certain words even when the topic of discussion (in the ordinary sense) is nominally the same. Conversely, the appearance of the same word in different contexts may produce minor or major variations in meaning. The meaning of 'realism' in the context of philosophy of science is fairly similar to this use of this word in the context of general metaphysics but quite different from the meaning of 'realism' in the context of political philosophy. Likewise, 'topic' in a discussion of LDA topic modeling does not mean the same as 'topic' in the context of a library or in a public debate. And lest it goes without saying, the contexts in which words are used change over time, so the context of 'mass' in Einstein's usage is not the same as in Newton's, yet both these contexts are more similar to each other than the post-1940 context in which the phrase 'probability mass' emerges.

A second benefit of the shift towards thinking in terms of contexts rather than topics is that it reorients us back towards a document-centered view of LDA. In fact, we disagree with the statement by Meeks & Weingart that topic modeling is "focused on corpora and not individual texts". It would be more accurate to say that topic modeling is typically deployed in ways that lead scholars to focus on corpora and not individual texts. Studies focusing primarily on changes in topic distributions through the life of a journal exemplify exactly this. Thinking that the job of LDA is to find topics that are latent in a corpus, sends one tripping towards the problems that Meeks & Weingart identify. But contexts matter to the understanding of particular documents. It is indeed banal to be told that Darwin's *On the Origin of Species* contains topics that are related to botany; one hardly needs topic modeling for that. But it is far less banal to use a topic model to help identify the way in which, say, one document provides part of the context for the production of another.

To illustrate, consider how one might compare Darwin's *Origin of Species* to the books that he read. At any given moment in his reading sequence, the books read so far contain some mixture of topics. Each new book read changes the aggregate mixture slightly. In the process of writing the *Origin*, Darwin assembles a new mixture. Using Kullback-Leibler divergence, we can assess how much different the mixture in the *Origin* is from the mixture of topics aggregated across the subset of books Darwin has read up to any point in the sequence. This is illustrated in Figure 1, but with a twist. Because there is a random element in the way in which the topics derived from the reading list are assigned to the written book, it is necessary to check that we obtain the same or similar answers from a repeated sampling process (for details see Murdock et al. 2018). When we run the process multiple times, we find that the samples fall into distinct

clusters, and one such cluster reveals a large signature for one particular book: Hugh Falconer's 1852 *Report on the Teak Forests of the Tenasserim Provinces* provides part of the context in which Darwin produced the *Origin* (the topmost cluster of lines in Figure 1).

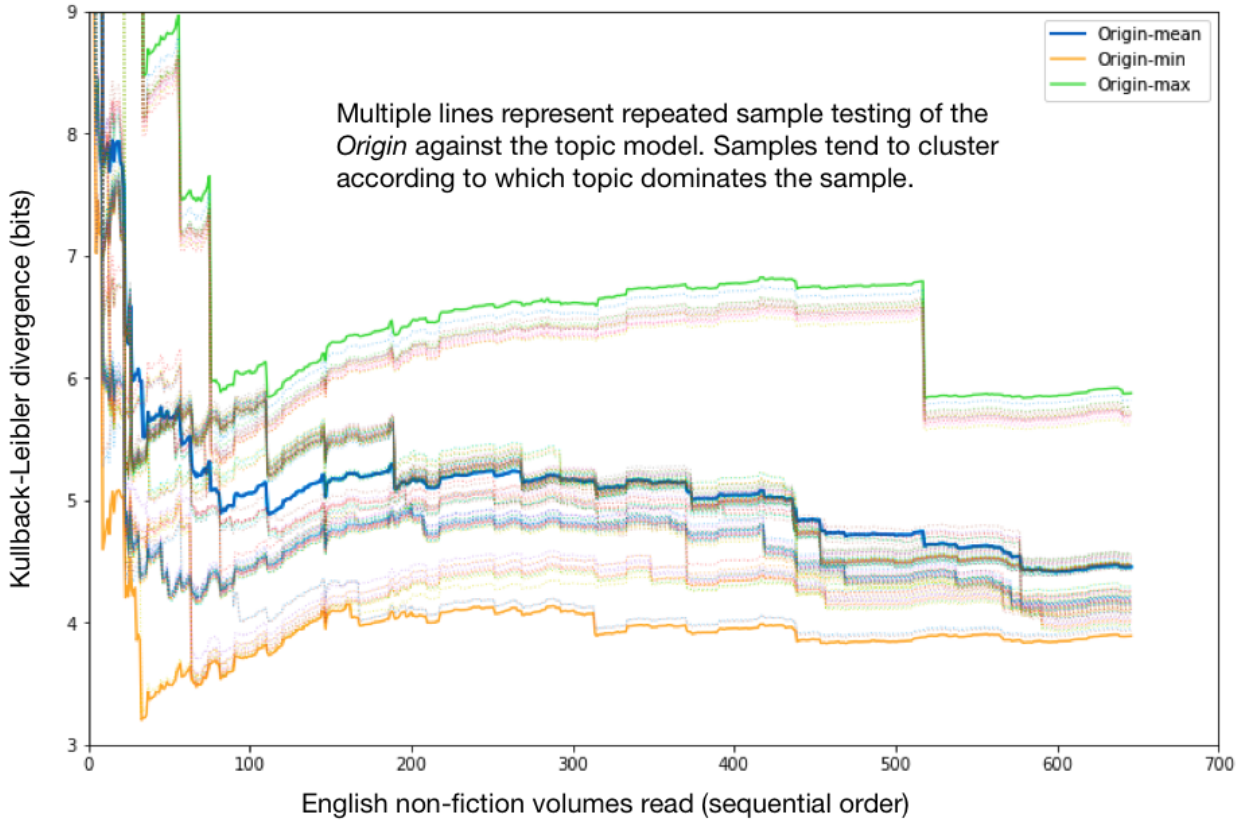

Figure 1: Cumulative divergence of *Origin* from English non-fiction books read by Darwin 1837-1859 generated by repeatedly fitting the *Origin* to a 200-topic model of the reading corpus ("sampling"). The X-axis represents the number of books read in the order that Darwin read them. The Y-axis represents KL-divergence between the books read and samples of the *Origin* Solid (bolder) lines represent the mean (blue line) and the upper (green line) and lower (orange line) bounds of divergence at each point in the reading sequence. Lighter (dashed) lines each represent the results of one sample, and they are colored according to which topic is most dominant for each sample. Samples evidently form clusters according to dominant topic. The cluster just below the max (red dashed lines) has the greatest divergence from the reading, always close to the max. The samples in this cluster are dominated by a topic whose highest probability words are 'forests', 'timber', 'teak', 'forests', 'Ceylon'. The big drop in divergence just after the 500th item indicated on the X-axis, corresponds to Falconer's report on the teak forests in Burma, which Darwin recorded finishing on August 11, 1853. (For details of the sampling process and the clusters that emerge see Murdock et al. 2018.)

Instead of focusing on the high probability words associated with the topics, we use the topics more holistically to address the interaction between the documents (books) in Darwin's reading list and the book that he wrote. The context in which Falconer's report is relevant to Darwin's

writing is his discussion of the similar adaptations of different species of trees in similar mountain climates found at different altitudes and latitudes worldwide, constituting part of his argument for the power of natural selection to shape the characteristics of species.

The turn to a document-centered view of topic models as context models applies not just to a theory of writing, but to an account of reading as well. In our previous work we showed that the sequence of Darwin's readings between 1837 and 1860 was not haphazard, but neither did he show the same pattern of choices through time; shifts in the pattern correspond to changes in his work context (Murdock, Allen & Dedeo, 2017). More specifically, using the topic weights assigned to the immediately prior book that he had read, as well as the aggregate of the topic weights assigned to the totality of books he had read to that point, we were able to apply an information-theoretic measure of distance to the next book he read. By focusing on Darwin's reading trajectory through the documents, rather than on trying to interpret individual topics, we were able to detect shifting patterns in how close he stayed to the last-read book, and whether he returned to books similar to those he had read previously, or whether he went fishing for information in areas where he had not previously done any reading. These shifts correlated with three main phases of his work life: first, organizing and publishing his notes from the voyage of the *Beagle*, next taking up the intensive study of barnacles from 1846 to 1854, and then turning to the organization of his notes for the book that would eventually be published in 1859, *On the Origin of Species*. Thus we showed how the reading contexts extracted via LDA are related to Darwin's intellectual endeavors.

Subsequently, we have begun to extend the reading model to Darwin's writings, with preliminary results published by Murdock et al. (2018). Thinking about how LDA helps us to identify contexts led us to become less concerned with finding labels for topics (as a way of interpreting them), and more interested in how the models help us find significant relationships between the documents that he read and wrote. For instance, we have results (again preliminary) indicating that the KL-divergence to Darwin's *Origin* is as actually lowered by Whewell's (1837) *History of the Inductive Sciences*, which he read soon after coming off the *Beagle*, whereas it is increased by the works of Francis Bacon (ed. Montagu 1825-34), which he read a couple of years after reading Whewell. This provides some additional evidence in support of Ayala's (2009) claim that Darwin's methodology owes more to Whewell than to Bacon, despite Darwin's overt claim to be following Bacon's inductivist method.

The examples of the preceding two paragraphs describe our ongoing attempts to apply topic modeling to questions within HPS using a document-centered approach. Although these methods occasionally look at the signature contributions of specific "topics" to the documents, they work without regard to whether those topics are directly interpretable as standalone artifacts or in respect to the corpus as a whole. Similarly, Murdock et al. 2018 described methods for comparing Darwin's writings on evolution to Alfred Russel Wallace's 1858 essay using the model of Darwin's readings. The preliminary results reported there indicate greater similarity, in information-theoretic terms, of Wallace's essay to Darwin's earlier essays than to the *Origin*. Darwin himself recognized this similarity to his earlier essays immediately, remarking

to Lyell in a letter dated 18 June 1858: “If Wallace had my MS. sketch written out in 1842, he could not have made a better short abstract!”

§4 Towards More Robust Modeling Practices

In repeatedly emphasizing the preliminary nature of the results, we may seem again to be tripping towards problems that Meeks & Weingart (2012) identified in their continuation of a sentence already quoted above: “You would marvel at the output, for a moment, before realizing there isn’t much immediately apparent you can actually do with it, ***and the article would list a few potential applications along with a slew of caveats and dangers***” (bold italics added to emphasize the continuation). Reasons for caution are many. In addition to the tendencies towards over-interpretation of topics mentioned above, there are multiple technical issues. A far from exhaustive list includes: sensitivity of the models to which volumes were obtained for the corpus and how the documents were fed into the model (whether as whole books, chapters, journal articles, pages, 400-word chunks, etc.); sensitivity to the digitization process and cleanup of the text (treatment of hyphenation, inclusion or removal of headers and footers, etc.); unitization of terms as single words or using multi-word phrases; removal of terms (by stoplist, by word frequency, as “foreign”, etc.); parameterization of the models (number of topics, choice of Dirichlet hyperparameters, number of cycles of training) and stochasticity in the training process (stemming from choice of seed for initial randomization, the nature of the chosen sampling process, etc.).

Philosophers of science are well positioned to bring their expertise to bear on assessing topic models, because of their attention to modeling practices in science, including issues such as model robustness and the representational status of models. We agree with the sentiment expressed in the (different) context of philosophy of cognitive science by Paul Smaldino (2017), who titled his article, “Models are stupid, and we need more of them”, itself a twist on the much earlier proclamation attributed to statistician George Box, “Essentially, all models are wrong, but some are useful” (Box & Draper 1987, p.424). Despite the necessarily partial view provided by any particular modeling approach, and the stupidity of AI/ML, computational models in general and topic models in particular provide useful contexts for discoveries about documents important to the history and philosophy of science. In our own work we have tried to establish that results are robust across models with different numbers of topics (Murdock, Allen & Dedeo. 2017, appendix D.1), and we have also attempted to investigate systematically the behavior of models with different numbers of topics across different sample sizes (Murdock, Zeng, & Allen 2016). The space of possible investigations is huge, however, and much more work of this kind needs to be done.

Working with multiple models simultaneously, fosters the kind of “interpretive pluralism” that characterizes humanities computing (Rockwell & Sinclair 2016). It is also consonant with the kind of ensemble modeling used for weather forecasting (Gneiting & Raftery 2005). Emphasizing the interaction between models and human needs reflects the origins of LDA topic modeling in the field of information retrieval (Blei, Ng & Jordan 2003) -- a corner of information science that aims to support people to find what they need when confronted with large amounts

of text. As topic modeling has been taken up within the digital humanities, and in the time since Weingart & Meeks wrote their *quo vadis* in 2012, the level of understanding and analysis of what can be done with topic models has continued to develop among those who work with them intensively. This understanding has proceeded on two fronts: one is the recognition of the role of human interpretation, in rebuttal of the oft-expressed worry that computational methods seek to replace or reduce human understanding with (mere) mechanically derived statistical summaries of the text; the second is based on successful application of topic modeling to questions that scholars in the humanities should and do care about. To the first point, Geoffrey Rockwell & Stéfan Sinclair (2016), as indicated in the title of their book, *Hermeneutica*, stress the way in which computational models become themselves objects of interpretation by scholars who have different interests and interpretive strategies.

Andrew Piper (2018) continues the theme of critical engagement with the models in his book *Enumerations*. He adds necessary nuance to earlier arguments by Franco Moretti (2013), who coined the phrase that became the title of his book, *Distant Reading*, and Matthew Jockers (2013), whose book *Macronanalysis* pioneered the use of network visualizations based on topic models of very large corpora. Piper is aware that computational analyses do not intrinsically gain credence due to the quantity of data, but rather by their representativeness of texts that may have been overlooked or otherwise marginalized by traditional means of analysis. The field of literary studies has traditionally held to the view that generalizations about literature can only be justified on the basis of close readings of particular texts, but Piper's book shows how the same kinds of generalizations can be supported independent of these close readings. Ted Underwood (2019) argues that topic modeling in literary studies supports a hypothesis testing approach that can reduce hindsight bias that arises when computational methods are applied unguided by specific hypotheses, addressing the problem of how we prove these methods illuminate more than we already know.

This growing literature theorizing and justifying the practice of topic modeling for the humanities has been dominated so far by scholars of history and literature. Philosophers have barely engaged, predominantly sharing the prejudice that such techniques have little to say on issues of philosophical interest. The applications of topic modeling that are acclaimed by authors such as Piper or Underwood -- for example, to determine the historical emergence of genre in literature -- do not strike philosophers as central to their concerns. By having a foot in history, the integrative study of history & philosophy of science is in an interesting position of being forced to close the gap (or better, following Schickore 2011, to eliminate the false opposition) between concern for the particular (in the details of specific episodes of science) and the quest for abstraction (via generalizations that contribute to understanding the significance of those episodes).

Although our own work on Darwin focused on the very particular reading and writing behavior of a single scientist, we pursued this project with the rather general goal of developing methods for measuring and tracking such things as conceptual similarity, conceptual change, the sensitivity of meanings to context, and pathways of intellectual influence. Topic models are both wrong and oblivious to these higher-order goals of understanding. They only partially capture aspects

of language that are relevant to the sorts of meaning that are extracted by competent close readers of the texts. However, the evidence they provide speaks in new ways to existing questions, and leads to new questions that could have been reached by traditional close reading or by other computational means, but are perhaps less likely to have been reached in without the assistance of LDA.

Our focus on LDA is not rooted in any pretence that it is the only tool that matters for computational text analysis. Other approaches have their uses too. Nor is it automatically the right tool for any given purpose. Historian Jo Guldi (pers. comm., ms. in prep.) argues that older, simpler algorithms for analyzing and comparing documents—such as *ngram* counts (simple quantification of occurrences of n-word phrases) and *tf-idf* (term frequency - inverse document frequency; originally described by Jones 1972)—provide "white box" methods that are preferable because of their greater comprehensibility to non-experts than "black box" methods such as LDA. Researchers who are among the LDA cognoscenti may gravitate towards more sophisticated tools for the wrong reasons, and without checking to see whether something else would work just as well for the purpose at hand. One worries about swatting the proverbial fly with the proverbial elephant gun. Direct comparison of methods may however help justify the use of both. Malaterre et al. (2020), for instance, footnote their decision to use LDA because of "its proven reliability for identifying topics in large corpora" and they also mention benchmarking LDA against a simpler approach, *k*-means clustering, preferring LDA because it produced just as good quantitative results that were also more interpretable. (Exercise for the reader: Insert suitable warning about interpretability here!) To their credit, Malaterre and coauthors do provide their readers with titles of documents related to the topics in their model, thus taking a step towards the document-centered view we are advocating.

Similar points apply not just to the choice of modeling approach, but also to the methods used to analyze the models. In our own work on Darwin's reading behavior, we have preferred (the relatively complex) Kullback-Leibler (KL) divergence measure to assess similarity between document-topic distributions rather than the conceptually simpler cosine similarity sometimes used on these vectors of topic weights. We also found that while both a simpler (white-box) rank-ordering method and the fully quantitative KL measure (relatively black-box to some, due to its complexity, although perhaps less opaque than neural network models) could capture Darwin's overall pattern of reading selections, only KL allowed us to adequately quantify shifts in his behavior between exploitation and exploration and correlate these shifts with the major epochs in his research career already described above. Where the methods converge, one has more confidence in both, and where they diverge, we are confident that the extra complexity is worthwhile. A legitimate concern (due here to an anonymous reviewer) is that our choice to use KL was entirely post hoc, based on finding a method the yields the results we wish. In fact, however, we chose KL not for those reasons, but (as explained at the beginning of Murdock, Allen & Dedeo 2017) because it is a widely used measure within cognitive science.

In this chapter, following our previous work, we have focused on the most basic, unembellished form of LDA topic modeling. More sophisticated applications of LDA exist, such as 'Dynamic Topic Modeling' (Blei and Lafferty 2006), as well as other approaches to modeling text, such as

neural network based word embeddings (i.e., Word2vec by Mikolov et al. 2013), and combined methods (such as lda2vec by Moody 2016). While such methods provide incremental advances on standard information retrieval benchmarks, they have yet to be shown to have specific benefits for computational humanities in general, or history & philosophy of science in particular. Furthermore, dynamic LDA, and related approaches that make topics variant rather than invariant entities over time, may suggest relationships among documents that are spurious. For example, Robert Rose in our research group at Indiana University ran some simple experiments with dynamical topic models showing that if topics were allowed to evolve in response to new documents being added to the corpus, a topic that was previously prominent in early documents concerning (e.g.) theology could morph to become dominant in later documents about (e.g.) symbolic logic, due to simple word ambiguities (e.g., 'church' as referring initially to the Catholic Church and later appearing as the name of a seminally important logician). Of course, this particular issue could have been solved by tokenizing the name and the noun differently instead of treating both as instances of the same 'word', but it is indicative of a much broader problem of colexification of different concepts that cannot easily be solved algorithmically. It also might be thought that the problem can be avoided because typical corpus for HPS research would not contain such a strange mixture of theology and mathematical logic. This optimism is undermined by actual experience, however. Thus, for instance, the corpus of 1,315 books we used for the study reported by Murdock, Allen, Börner et al. (2017) contained examples from logic, theology, comparative psychology, and many other areas, and also coincided with a shift in the use of the term 'anthropomorphism' from theological and anthropological contexts to the context of discussions about the nature of animal minds. A dynamic topic modeling approach with a fixed number of topics might have been forced to repurpose a topic assigned to early texts about theology to fit later texts about comparative psychology, whereas the simpler, static approach we took differentiated these topics within the corpus taken as a whole.

So, although such a simple, unpublished experiment as the one we conducted with "church" and "Church" is not definitive, it suggests that more work needs to be done to determine whether more sophisticated versions of topic modeling are appropriate for the needs of HPS scholars. In our view, the document-centered view helps to keep the correct goal in sight. Fluctuations or other changes in topics are not the important outputs of LDA topic modeling. Rather, the power of the topic models lies in their ability to reveal relationships among appropriately contextualized documents.

§5 Conclusion

In the paragraph from Meeks and Weingart (2012) with which we opened this essay, they refer to the "seductive but obscure results in the forms of easily interpreted (and manipulated) 'topics.'" By reorienting the consumers of topic models away from the "topics" and back to the documents we believe that the results of LDA topic modeling may be rendered less seductive and less obscure. While the need for good information retrieval makes it worthwhile to model large corpora, often comprising documents that have been aggregated for institutional reasons such as library collections or the contents of professional journals, we have chosen a different path in our HPS work: to model the reading and writing behavior of specific individuals. This

focus on people and the documents they read has led us to ask questions about their exploration and exploitation of the cultural contexts in which they found themselves, and to view topic models as tools for identifying influence and measuring creativity within those contexts. The documents are, after all, the ultimate repositories of authors' meanings, and can only be read and understood by human beings given the current limitations of all forms of AI and machine learning. But we hope to have convinced our readers that judiciously used, LDA topic modeling is among the algorithms that are worthy of exploration and exploitation by historians & philosophers of science, who will continue to supply the meanings that inform our understanding of science and its philosophical significance.

Acknowledgements

We are grateful Jo Guldi and two anonymous referees for helpful comments on an earlier draft. CA also would like to thank the audiences for his talks at the LEAHPS II conference at the University of Hannover in August 2019, and in the University of Pittsburgh Mellon-Sawyer Information Ecosystems lecture series in February 2020. Finally, we are grateful for the invitation to contribute to this volume and the patience of the editors during the writing process.

References

Airoldi, Edoardo M., Blei, David M., Erosheva, Elena A. & Fienberg, Stephen E. (2014). *Handbook of Mixed Membership Models and Their Applications*. London: Chapman & Hall.

Ayala, Francisco J. 2009. Darwin and the scientific method. *Proceedings of the National Academy of Sciences* 106 (Supplement 1): 10033-10039; doi: 10.1073/pnas.0901404106.

Barron, Alexander T. J., Huang, Jenny, Spang, Rebecca L. & Dedeo, Simon (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences* 115(18): 4607–4612; doi: 10.1073/pnas.1717729115.

Binder Jeffrey M. (2016) Alien Reading: Text Mining, Language Standardization, and the Humanities. In Debates in the Digital Humanities 2016. Matthew K. Gold and Lauren F. Klein (Eds.). Minneapolis: U. of Minnesota Press: 201-217.

Blei, D. (2012a). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, 2, 8–11.

Blei, D. (2012b). Probabilistic Topic Models. *Communications of the ACM,* 55, 77-84; doi: 10.1145/2133806.2133826.

Blei D.M & Lafferty, J.D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML'06), Pittsburgh, PA.

Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of *Science*. *The Annals of Applied Statistics*, 17–35; doi: 10.1214/07-AOAS114.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

Box, G. E. P. & Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*. Hoboken, NJ: John Wiley & Sons.

Boyd-Graber, J., Hu, Y. & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval,* 11: 143-296; doi: 10.1561/1500000030.

Brauer, R., & Fridlund, M. (2013). Historizing topic models: a distant reading of topic modeling texts within historical studies. In Nikiforova, L.V. & Nikiforova, N.V. (Eds.) *Cultural Research in the Context of "Digital Humanities": Proceedings of International Conference 2-5 October 2013*, St Petersburg, pp. 152–163.

Chang, Jonathan, Boyd-Graber, Jordan, Wang, Chong, Gerrish, Sean, & Blei, David M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Proceedings of *Neural Information Processing Systems* 2009: 1–9, Vancouver, BC.

Cohen Priva, U. & Austerweil, J.L. (2015). Analyzing the history of *Cognition* using Topic Models. *Cognition*. 135:4-9. doi: 10.1016/j.cognition.2014.11.006.

Darwin, Charles (1859). *On the Origin of Species*. London: Murray & Co.

Falconer, Hugh (1852). Report on the teak forests of the Tenasserim provinces. Calcutta.

Gneiting, Tilmann & Raftery, Adrian E. (2005). Weather Forecasting with Ensemble Methods. *Science*, 310(5746): 248-249; doi: 10.1126/science.1115255.

Guldi, J. (2019a) Parliament's Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change. *Technology and Culture* 60: 1–33.

Guldi, J. (2019b) The Measures of Modernity: Word Counts, Text Mining and the Promise and Limits of Present Tools as Indices of Historical Change. *International Journal for History, Culture and Modernity* 7; doi: 10.18352/hcm.589.

Guldi, Jo (ms. in prep.) *The Dangerous Art of Text Mining* (working title).

Guldi, J. & Williams, B. Synthesis and Large-Scale Textual Corpora: a Nested Topic Model of Britain's Debates over Landed Property in the Nineteenth Century. *Current Research in Digital History* 1; doi: 10.31835/crdh.2018.01.

Jockers, M. (2013) *Macroanalysis: Digital Methods and Literary History.* Champaign, IL: Univ. Illinois Press.

Jones, Karen Spärck (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* 28: 11-21.

Lambert, Ben, Kontonatsios, Georgios, Mauch, Matthias, Kokkoris, Theodore, Jockers, Matthew, Ananiadou, Sophia, Leroi, Armand M. (2020). The pace of modern culture. *Nature Human Behaviour*, 2397-3374; doi: 10.1038/s41562-019-0802-4.

Marcus, Gary & Davis, Ernest (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.

Meeks, Elijah & Weingart, Scott B. (2012). The Digital Humanities Contribution to Topic Modeling. *Digital Humanities* 2(1): http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/ (accessed 2020-02-03).

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., and Dean, Jeff (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* 26 (NIPS 2013), (Eds.) C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger.

Mitchell, Melanie (2019). *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.

Moody, Christopher E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. Online at https://arxiv.org/abs/1605.02019 (accessed 2020:02:02).

Montagu, Basil, Ed. (1825-1834) The works of Francis Bacon. London: W. Pickering.

Moretti, Franco (2013). *Distant Reading*. New York: Verson.

Murdock, Jaimie, Zeng, Jiann, & Allen, Colin (2015). Towards Cultural-Scale Models of Full Text. In *Proceedings of the 2016 International Conference on Computational Social Science*, Evanston, Illinois.

Murdock, Jaimie, Allen, Colin & DeDeo, Simon (2017a). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. *Cognition* 159: 117-126; doi:10.1016/j.cognition.2016.11.012.

Murdock, Jaimie, Allen, Colin, Börner, Katy, Light, Robert, McAlister, Simon, Ravenscroft, Andrew, Rose, Robert, Rose, Doori, Otsuka, Jun, Bourget, David, Lawrence, John, Reed, Christopher (2017b). Multi-level Computational Methods for Interdisciplinary Research in the HathiTrust Digital Library. *PLoS ONE* 12(9): e0184188; doi: 10.1371/journal.pone.0184188.

Murdock, Jaimie, Allen, Colin & DeDeo, Simon (2018). Quantitative and Qualitative Approaches to the Development of Darwin's Origin of Species. *Current Research in Digital History* 1; doi: 10.31835/crdh.2018.14.

Peirson, B. R. Erick, Bottino, Erin, Damerow, Julia; Laubichler, Manfred D. (2017). Quantitative Perspectives on Fifty Years of the Journal of the History of Biology. *Journal of the History of Biology* 50:695–751; doi: 10.1007/s10739-017-9499-2.

Piper, Andrew (2018). *Enumerations: Data and Literary Study*. Chicago: University of Chicago Press.

Randolph, Thomas Jefferson (Ed.) (1829). *Memoir, Correspondence and Miscellanies: from the Papers of Thomas Jefferson*. Charlottesville, VA: F. Carr, and Co.

Ravenscroft, Andrew & Allen, Colin (2019). Finding and Interpreting Arguments: An Important Challenge for Humanities Computing and Scholarly Practice. Digital Humanities Quarterly 13(4): http://www.digitalhumanities.org/dhq/vol/13/4/000436/000436.html (accessed 2020-02-08).

Rockwell, Geoffrey & Sinclair, Stéfan (2016). Hermeneutica. Cambridge, MA: MIT Press.

Rothe, A., Rich, A.S. and Li, Z. (2018). Topics and Trends in *Cognitive Science*. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Schickore, Jutta (2011). More Thoughts on HPS: Another 20 Years Later. *Perspectives on Science* 19(4): 453-481.

Smaldino, Paul E. 2017. Models Are Stupid, and We Need More of Them. In *Computational Models in Social Psychology*, R. R. Vallacher, A. Nowak, & S. J. Read (Eds.). New York: Psychology Press.

Smith, Brian Cantwell (2019). The Promise of Artificial Intelligence: Reckoning and Judgment. Cambridge, MA: MIT Press.

Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Whewell, William (1837). *History of the Inductive Sciences*. London: John W. Parker.