Multi-level computational methods for interdisciplinary research in the HathiTrust Digital Library

Jaimie Murdock[1,2,+], Colin Allen[1,2,+,*], Katy Börner[1,2,3,4,+], Robert Light[2,+], Simon McAlister[5], Robert Rose[1,6], Doori Rose[1], Jun Otsuka[7], David Bourget[8], John Lawrence[9], Andrew Ravenscroft[5], and Chris Reed[9]

[1]Program in Cognitive Science, Indiana University, Bloomington, IN, USA
[2]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[3]Indiana University Network Science Institute (IUNI), Bloomington, IN, USA
[4]User-Centered Social Media, Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Duisburg, Germany
[5]Cass School of Education & Communities, University of East London, London, United Kingdom
[6]Department of Mathematics, Indiana University, Bloomington, IN, USA
[7]Department of Philosophy, Kyoto University, Kyoto, Japan
[8]Department of Philosophy, University of Western Ontario, London, Ontario, Canada
[9]Centre for Argument Technology, University of Dundee, Dundee, United Kingdom
[+]These authors contributed equally to this manuscript, see "Author's Contributions" for details.
[*]Corresponding author: colallen@indiana.edu

February 6, 2017

Abstract

We show how faceted search using a combination of traditional classification systems and mixed-membership models can move beyond keyword search to inform resource discovery, hypothesis formulation, and argument extraction for interdisciplinary research. Our test domain is the history and philosophy of scientific work on animal mind and cognition. We demonstrate an application of our methods to the problem of identifying and extracting arguments about anthropomorphism during a critical period in the development of comparative psychology. We show how a combination of classification systems and mixed-membership models trained over large digital libraries can inform resource discovery in this domain, using methods that can be generalized to other interdisciplinary research questions. Through a novel approach of drill-down topic modeling, we are able to reduce a collection of 1,315 fulltext volumes to 6 focal volumes that did not appear in the first ten search results in the HathiTrust digital library. This ultimately supports a system for semi-automatic identification of argument structures to augment the kind of "close reading" that leads to novel interpretations at the heart of scholarly work in the humanities, drilling down from massive quantities of text to very specific passages. This multi-level view advances understanding of the intellectual and societal contexts in which writings are interpreted.

Introduction

Just as Britain and America have been described as two nations separated by a common language, different academic disciplines often use the same words with divergent meanings [1]. Interdisciplinary research thus poses unique challenges for information retrieval (IR). Word sense disambiguation [2, 3], differing publication practices across disciplines [4–6] and disjoint authorship networks [7] pose special challenges to information retrieval for interdisciplinary work. When the dimension of time is added, terminological shifts [8, 9], changing citation standards [10–13], and shifting modes of scholarly communication [4, 5, 14, 15] all amplify the challenges for IR to serve the need of interdisciplinary scholars.

Widespread digitization of monographs and journals by HathiTrust [16, 17] and Google Books [18, 19] enable new longitudinal studies of change in language and discourse [8, 9, 12, 20–22], an approach known as "distant reading" [23]. These data-driven distant readings contrast with "close readings", in which short passages and particular details are emphasized for scholarly interpretation. Newly digitized materials, which enable distant reading, differ from born-digital scholarly editions in three key ways: First, the reliance on optical character recognition (OCR) over scanned page images introduces noise into the plain-text representations of the text. Second, the unstructured text does not contain any markup that may differentiate page header and footer information, section headings, or bibliographic information from the main text. Finally, metadata is often automatically extracted and lacks the provenance information important to many humanities scholars. Researchers seeking to marry these "distant readings" to more traditional "close readings" are impacted by these factors [24].

Our goal is to develop computational methods for scholarly analysis of large-scale digital collections that are robust across both the technological inconsistency of the digitized materials and the variations of meaning and practice among fields and across time. A further goal of our approach is that these methods should inform interdisciplinary research by suggesting novel interpretations and hypotheses. These methods should support scholars who wish to drill down from high level overviews of the available materials to specific pages and sentences that are relevant for understanding the various responses of scientists to contentious issues within their fields.

In this paper, we focus on meeting these challenges within the interdisciplinary field of history and philosophy of science (HPS). HPS must not only bridge the humanities and the sciences, but also the temporal divide between historically-significant materials and the present [25–28]. We show how faceted search using a combination of traditional classification systems and mixed-membership models can move beyond keyword search to inform resource

discovery, hypothesis formulation, and argument extraction in our test domain, delivering methods that can be generalized to other domains.

Using a novel approach of drill-down topic-modeling, we demonstrate how a set of 1,315 fulltext volumes obtained by a keyword search from the HathiTrust digital library is reduced to 6 focal volumes that did not appear in the top 10 HathiTrust search results. Topic modeling of these volumes at various levels, from whole book down to individual sentences, provides the contexts for word-sense disambiguation, is relatively robust in the face of OCR errors, and ultimately supports a system for semi-automatic identification of argument structure. We show how visualizations designed for macroanalysis of disciplinary scientific journals can be extended to highlight interdisciplinarity in arguments from book data [29]. This guides researchers to passages important for the kind of "close reading" that lies at the heart of scholarly work in the humanities, supporting and augmenting the interpretative work that helps us understand the intellectual and societal contexts in which scientific writings are produced and received.

While the extension of computational methods to various questions in the humanities may eventually provide ways to test specific hypotheses, the main focus of such research is likely to remain exploratory and interpretative, in keeping with the humanities themselves [24, 30]. This approach nevertheless shares something with the sciences: it is experimental to the extent that it opens up a space of investigation within which quantitatively defined parameters can be systematically varied and results compared. Such exploratory experimentation is common not just in the social sciences, but also in the natural sciences [31, 32].

Our study consisted of six stages. (1) We used a *keyword search* of the HathiTrust collection to generate an initial search space for the faceted search. (2) We constructed *probabilistic topic models* for the volumes in the initial search results. This model is a type of mixed-membership model, which captures the multiple contexts of the selected volumes and allows us to reduce the original search space even further. Topic models are also a type of bag-of-words model, making them well-suited for the unstructured text found in the HT. (3) Third, we used *drill-down topic modeling* to construct page-level models of the reduced set of volumes selected at the previous stage. (4) Using the page-level results to select pages for close-reading analysis, we thus supported semi-automatic *argument extraction* to showcase the interpretive results of our search process. (5) We exploited the close reading of arguments for exploratory investigation of sentence-level topic modeling in a single volume. (6) We used *scientific mapping* to find relevant volumes [33]. As current science maps represent journal data and data overlays are created based on journal names, we used a *classification crosswalk* from the UCSD Map of Science to the Library of Congress Classifications of these journals, allowing us to project books onto the science map.

Materials

HathiTrust Digital Library

The HathiTrust Digital Library is a collaboration between over ninety institutions to provide common access and copyright management to books digitized through a combination of Google, Internet Archive, and local initiatives. As of October 24, 2016, it consisted of over 14.7 million volumes represented both as raw page images and OCR-processed text[1].

Due to copyright concerns, access is given only to pre-1928 materials, which are assumed to be in the public domain in the United States.[2] When the work described in this paper was initiated in 2012, the public domain portion of the HathiTrust consisted of approximately 300,000 volumes. At the end of the funding period in 2014, the public domain consisted of 2.1 million volumes. That number is now 5.7 million volumes, as of October 24, 2016.

While the corpus size has increased 20-fold, the methods presented in this paper are aimed to reduce the portion of the corpus for analysis. For example, the first step described below is *keyword search*, with our initial results returning 1,315 volumes (referred to as the HT1315 corpus). Using the same query on October 24, 2016, we returned 3,497 volumes. Both of these datasets are computationally-tractable on modern workstations, in contrast to (for example) the 1.2 terabyte HTRC Extracted Features Dataset, derived from 4.8 million volumes [35].

From the HT1315 corpus, we selected 86 volumes to model at the page-level (the HT86 corpus). This corpus was then further reduced to a 6-volume collection for argument mapping (HT6).

Stop Lists

Before analyzing the texts, it is common to apply a 'stop list' to the results, which omits words that are poor index terms [36]. Frequently, these are high-frequency words such as articles ('a', 'an', 'the'), prepositions ('by', 'of', 'on'), and pronouns ('he', 'she', 'him'), which contain little predictive power for statistical analysis of semantic content [37]. We use the English language stop list in the Natural Language Toolkit, which contains 153 words [38]. Additionally, we filtered words occurring 5 or fewer times, which both excludes uncommon words and infrequent non-words generated by OCR errors.

[1] https://www.hathitrust.org/statistics_info

[2] During the funding period, the fulltext of post-1928 materials were impossible to access for computational analysis from the HathiTrust. Recently, the HathiTrust Research Center (HTRC) Data Capsule has been developed to enable tightly restricted access to in-copyright materials [34].

UCSD Map of Science

For our macroanalysis, we want to see how our selected texts divide among the different academic disciplines. As a base map for the disciplinary space (analogous to a world map for geospatial space), we use the UCSD Map of Science [29] which was created by mining scientific and humanities journals indexed by Thomson Reuters' Web of Science and Elsevier's Scopus and laying them out as a map of 554 sub-disciplines – e.g., Contemporary Philosophy, Zoology, Earthquake Engineering – that are further aggregated into 13 core disciplines – e.g., Biology, Earth Sciences, Humanities. Each of the 554 sub-disciplines has a set of journals and keywords associated with it.

Library of Congress Classification Outline (LCCO)

The Library of Congress Classification Outline (LCCO) is a system for classifying books, journals, and other media in physical and digital libraries. It is different from the Library of Congress Control Number (LCCN), which provides an authority record for each volume. The HathiTrust stores the LCCN, which we then use to query the Library of Congress database for the call number, which contains the LCCO, providing us with a disciplinary classification for each volume in the HT1315, HT86, and HT6 datasets.

Target Domain: History and Philosophy of Scientific Work on Animal Cognition

Our specific test domain is the history and philosophy of scientific work on animal cognition [39–41]. We aimed to identify and extract arguments about anthropomorphism from a relevant subset of the scientific works published in the late 19th and early 20th century. This period represents a critical time for the development of comparative psychology, framed at one end by the work of Charles Darwin and at the other end by the rise of the behaviorist school of psychology (see [42] for a full historical review). Using the methods described in this paper, we progressively narrowed the 300,000 volumes to a subset of 1,315 selected for topic modeling at the full-volume level, then 86 of these selected for page-level topic modeling, and then 6 specific volumes selected for manual analysis of the arguments.

The term "anthropomorphism" itself illustrates the problem of word sense disambiguation. In the theological context, anthropomorphism refers to the attribution of human-like qualities to gods. In the animal cognition context, it refers to the projection of human psychological properties to animals. Given the theological controversy evoked by Darwin, our inquiry demands our system be robust in partitioning these separate discourses.

Methods and Results

Keyword Search: From Library to Reading List

Methods

We began by conducting a keyword search in the HathiTrust collection using the HathiTrust's Solr index. We searched using terms intended to reduce the hundreds of thousands of public domain works to a set of potentially relevant texts that could be efficiently modeled with the available computing resources. Specifically, we searched for "Darwin", "comparative psychology", "anthropomorphism", and "parsimony". While the specificity of our query may be seen as too restrictive, we emphasize that we are following an exploratory research paradigm - we are not narrowing in on a particular fact, but rather surveying the available literature at the intersection of our interest in the history and philosophy of animal mind and cognition.

Results

The search yielded a set of 1,315 books published between 1800 and 1962. We refer to this set of results as HT1315.[3] The same query conducted in August 2015 yielded 3,027 full-text results. Notably, it took Charles Darwin 23 years to read a number of books comparable in size to HT1315, as documented in his Reading Notebooks [43]. Even at the unlikely rate of one book a day, it would take nearly four years to read this set of books in its entirety. About one fifth of the volumes retrieved were course catalogs, but even eliminating those would leave a daunting, if not quite Olympian, reading task. As the majority of the volumes selected by keyword search were not directly relevant to the research project, the potential payoff made possible by more sophisticated computational analysis of the full texts is critical for information retrieval tasks.

[3] A list of titles and HT handles is provided in the supplemental materials. Because the HT collection has changed over time, this exact set of results cannot be recreated by doing the same keyword search at hathitrust.org (see http://bit.ly/1LBbqnS). At the time our project started, there were over 300,000 public domain works in the HathiTrust digital library. Currently there are over 5.5 million public domain works in the collection (see https://www.hathitrust.org/visualizations_dates_pd).

Probabilistic Topic Modeling of Volumes: Narrowing the Reading Lists

Methods

Probabilistic topic models [44] are a family of mixed-membership models that describe documents as a distribution of topics, where each topic is itself a distribution over all words in a corpus. Topic models are *generative* models, that we interpret as providing a theory about context blending during the writing process [43].

To construct the topic models used in this study, we use *Latent Dirichlet Allocation* (LDA – [45]) with priors estimated via Gibbs sampling [46] as implemented in the InPhO Topic Explorer [47].

We initially modeled the HTRC1315 set at four different values for the number of topics, $k = \{20, 40, 60, 80\}$. We applied cosine-similarity measures to the topic mixtures attributed to each volume by the model.

Results

Manual inspection of the topics generated for the different values of k showed that while all four of the models produced interpretable results, we judged that $k = 60$ provided the best balance between specificity and generality for our HPS goals.

Table 1 shows the top ten topics related to the word 'anthropomorphism' in the k=60-topic model. Inspection of this list indicates that 'anthropomorphism' relates to a theological topic (38), a biological topic (16), a philosophical topic (51), an anthropological topic (58), etc. The topic model checking problem [44] – i.e., how to assess the quality of the model's topics – remains an important open problem in topic modeling. Nevertheless, most of the topics in the model can be quickly summarized, with the second topic (16) being the most obvious attractor for researchers interested in comparative psychology. The second-to-last topic (1) is targeted on bibliographic citations, and is dominated by common German words that were not in the English language stop list used during initial corpus preparation.

We use the topic model to narrow the search by querying topics with a combination of words. We do this by finding the topic or topics with the highest sum of the probabilities for each word. For example, Table 2 shows the top ten topics returned using 'anthropomorphism', 'animal', and 'psychology' as input. This new query reveals two relevant topics (numbers 26 and 10) that were not returned using 'anthropomorphism' alone.

Subsequently, we used all three topics (10, 16, and 26) to filter relevant books from the original set of 1,315 books. We took the cosine distance between each of the three topics to each book in HT1315. We took the sum of these three distances and filtered them at the

Topic	10 most probable words from topic
38	god, religion, life, man, religious, spirit, world, nature, spiritual, divine
16	**animals, evolution, life, animal, development, man, species, cells, living, theory**
51	philosophy, nature, knowledge, world, thought, idea, things, reason, truth, science
58	man, among, tribes, primitive, men, people, also, races, women, race
12	child, children, first, development, movements, play, life, little, mental, mother
21	social, life, new, mind, upon, individual, human, mental, world, subfield
11	motion, force, must, forces, matter, changes, us, parts, like, evolution
1	pp, der, vol, die, de, des, und, ibid, university, la
31	gods, religion, p, name, see, god, india, ancient, one, worship

Table 1: Topics ranked by similarity to 'anthropomorphism' in the HT1315 corpus. Topic 16 (highlighted with bold text) was used to derive the HT86 corpus, as it was most relevant to the inquiry.

Topic	10 most probable words from topic
26	**consciousness, experience, p, psychology, process, individual, object, activity, relation, feeling**
16	**animals, evolution, life, animal, development, man, species, cells, living, theory**
10	**animals, water, animal, food, birds, one, leaves, insects, species, many**
47	college, university, professor, school, law, work, students, degree, education, new
49	subfield, code, datafield, tag, ind2, ind1, b, d, c, controlfield
1	pp, der, vol, die, de, des, und, ibid, university, la
12	child, children, first, development, movements, play, life, little, mental, mother
58	man, among, tribes, primitive, men, people, also, races, women, race
21	social, life, new, mind, upon, individual, human, mental, world, subfield
2	test, tests, age, group, children, mental, table, per, cent, number

Table 2: Topics ranked by similarity to 'anthropomorphism', 'animal', and 'psychology' in the HT1315 corpus. Topics 26, 16, and 10 (highlighted with bold text) were used to derive the HT86 corpus, as they were most relevant to the inquiry.

Document	Distance
Comparative studies in the psychology of ants and of higher ...	0.64083
Secrets of animal life	0.66966
Ants and some other insects; an inquiry into the psychic ...	0.67043
The colours of animals, their meaning and use, especially ...	0.69174
The colour-sense: its origin and development	0.70837
The foundations of normal and abnormal psychology	0.70985
The animal mind; a text-book of comparative psychology	0.71216
The nature of life; a study in metaphysical analysis	0.72213
General zoology	0.72864
The animal mind; a text-book of comparative psychology	0.72944

Table 3: Documents ranked by similarity to topics 10, 16, and 26 in the HT86 corpus.

threshold of 1.25, yielding a smaller corpus of 86 volumes which we refer to as the HT86 collection. The top ten volumes identified in this way are shown in Table 3.

Drill-down Topic Modeling: From Books to Pages

Methods

We re-modeled the HT86 set at the level of individual pages, moving towards our goal of identifying arguments in text by "zooming in " to select books which had a high number of apparently relevant pages. These reduced sets of pages become appropriate targets for manual argument identification by a human reader.

The notion of a "document" in LDA topic modeling is flexible. One can consider a full volume as a single document with a particular topic distribution. However, finer-grained models can also be made, in which each page, paragraph, or sentence receives its own topic distribution. Since OCR document scans in the HathiTrust have very little structural information — there is no encoding for section headings or paragraph breaks, let alone chapter breaks — page-level was the next level below the full volume that we could reliably recover.

Results

For the sake of direct comparison to results reported above with the HT1315 model, we probed the $k = 60$ page-level model with 'anthropomorphism' as the query term. Results are shown in Table 4. Note that topic numbers do not correlate across the HT86 and HT1315 models. Although a theological topic (18) is at the top of the list, it is clear that biological and psychological topics have become more prevalent. Even within topic 18, 'evolution' and 'science' are now among the ten highest probability words indicating that the topic is closer

Topic	Top Ten Most Probable Words from Topic
18	god, religion, evolution, religious, man, human, science, world, christian, belief
3	mind, man, facts, life, evolution, instinct, subjective, instincts, organic, development
1	animal, animals, may, stimulus, experience, would, instinct, reaction, one, stimuli
51	sense, sensation, qualities, touch, perception, sensations, extension, sight, senses, us

Table 4: Topics ranked by similarity to 'anthropomorphism' in the HT86 corpus, as modeled at the page level.

Topic	Top Ten Most Probable Words from Topic
1	animal, animals, may, stimulus, experience, would, instinct, reaction, one, stimuli
51	sense, sensation, qualities, touch, perception, sensations, extension, sight, senses, us
18	god, religion, evolution, religious, man, human, science, world, christian, belief
3	mind, man, facts, life, evolution, instinct, subjective, instincts, organic, development

Table 5: Topics ranked by similarity to 'anthropomorphism', 'animal', and 'psychology' in the HT86 corpus.

to a "religion and science" topic than the more general religion topic 38 from the HT1315 model (Table 1), and reflecting the tighter range of books in the HT86 subset.

Using 'anthropomorphism', 'animal' and 'psychology' in combination as the query, topic 1 is the highest ranked topic (Table 5). In comparison to the earlier topics 10 and 16 from the HT1315 results in Table 2, this topic has more terms relevant to psychology (i.e., stimulus, experience, instinct, reaction), suggesting that for the purposes of locating specific pages in HT86 collection relevant to our initial interests, topic 1 provides the best starting point. Table 6 shows the first rows of a list of 800 highest ranked pages from HT86 using topic 1 as the query.

Document	Distance
The animal mind, 1st ed., p. 43	1.00000
The animal mind, 2nd ed., p. 47	1.00000
The animal mind, 2nd ed., p. 16	0.99999
The animal mind, 2nd ed., p. 263	0.99993
Mind in the lower animals, p. 179	0.99988
The animal mind, 1st ed., p. 219	0.99937
The animal mind, 2nd ed., p. 71	0.99893
The animal mind, 1st ed., p. 232	0.99887
The animal mind, 1st ed., p. 57	0.99778
The animal mind, 2nd ed., p. 48	0.99700

Table 6: Pages ranked by similarity to Topic 1.

We selected six volumes from the HT86 collection which had the most pages in the top 800 highest ranked pages. None of these volumes were in the top 10 keyword search results. These volumes formed the HT6 collection:

1. *The Animal Mind: A Textbook of Comparative Psychology*, 1908 (first edition), by Margaret Floy Washburn, psychologist. Washburn's textbook was foundational for comparative psychology and she is notable as the second woman to be president of the American Psychological Association.
2. *Comparative studies in the psychology of ants and of higher animals*, 1905, a monograph by Erich Wasmann, an entomologist who only partly accepted evolution within species, rejecting common descent, speciation via natural selection, and human evolution.
3. *The Principles of Heredity*, 1906, a scientific monograph by G. Archdall Reid, a physician who argued against the Lamarckian idea of inheritance of acquired characteristics.
4. *General Biology*, 1910, a text book by James G. Needham, entomologist and limnologist.
5. *The Nature and Development of Animal Intelligence*, 1888, a compilation of articles by Wesley Mills, physiologist, physician and veterinarian.
6. *Progress of Science in the Century*, 1908, a book on the history of science for general readers by J. Arthur Thomson, naturalist.

These books provide a broad array of perspectives on animal intelligence and psychology, from specialist monographs to textbooks to general-audience nonfiction. The texts were written by two Americans (Washburn and Needham), two Scots (Reid and Thomson), a Canadian (Mills), and an Austrian (Wasmann).

Argument Extraction: From Pages to Arguments

Methods

From the HT6 collection, we selected 108 pages for further analysis (Table 7). These pages were annotated using the Argument Interchange Format ontology (AIF – [48]), which defines a vocabulary for describing arguments and argument networks. We generated 43 argument maps using AIF annotated documents, providing a visual representation of the structure of each argument (e.g., Figure 1).

The argument content was marked up with OVA+[4], an application which links blocks of text using argument nodes. OVA+ provides a drag-and-drop interface for analyzing textual arguments. It also natively handles AIF structures. Each argument, as selected in the previous section, was divided into propositions and marked up as a set of text blocks.

[4]http://ova.arg-tech.org/, see also [49]

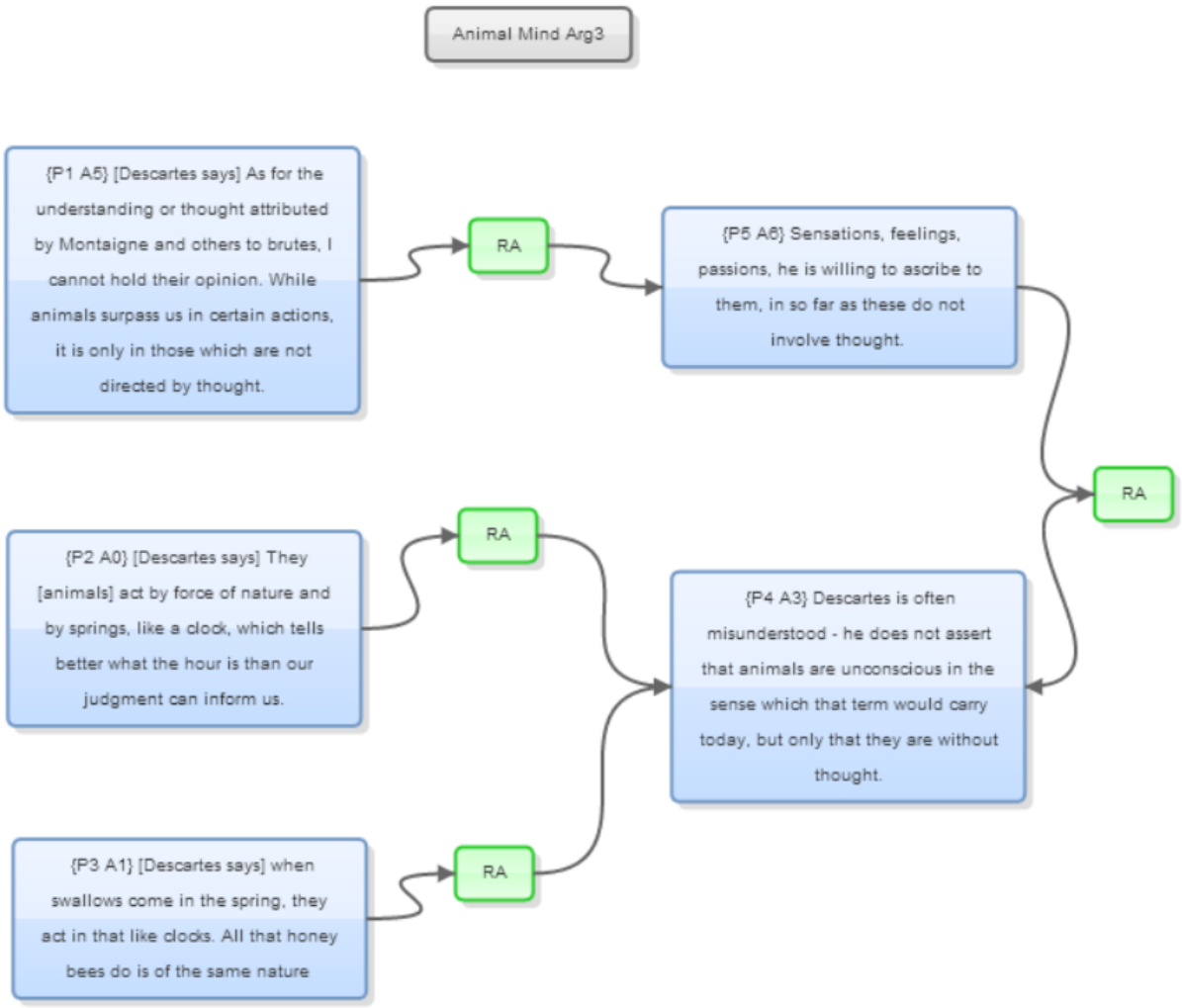

Figure 1: An argument map derived from *The Animal Mind*, represented in OVA+.

Volume	Pass 1	Pass A	Total	Pages
The Animal Mind	9	15	40	13-16, 16-21, 24-27, 28-31, 31-34, 58-64, 204-207, 288-294
The Psychology of Ants	8	10	37	Preface, 15-19, 31-34, 48-53, 99-103, 108-112, 206-209, 209-214
The Principles of Heredity		8	10	374, 381, 382, 385, 386, 390, 394, 395
General Biology		2	3	434-435, 436
The Nature and Development of Animal Intelligence		5	12	16-18, 21-26, 30-32
Progress of Science		3	6	479-484
Total	17	43	108	

Table 7: Pages for which OVA+ argument maps were created.

These text blocks containing propositions were linked to propositions that they support, or undercut, to create an argument map.

Results

We performed two types of argument analysis: Pass 1 aimed to *summarize* the arguments presented in each volume. Pass A aimed to *sequence* the arguments presented in each volume. All argument maps can be found at http://bit.ly/1bwJwF9. A full description of the study, including analysis of the arguments can be found in [50].

As a proof of concept, these arguments show the utility of new techniques for faceted search enabling access from a library of over 300,000 books to volume-level analysis of a subset of 1,315 books all the way down to page-level analyses of 108 pages for the purpose of identifying, encoding, modeling, and visualizing arguments. These argument diagrams function as a type of close reading, common in the humanities, and drawing on a rich tradition of philosophical literature (reviewed in [51]).

Drilling Down Again: From Arguments to Sentences

Methods

To further investigate the utility of combining distant reading methods with close reading, we applied topic modeling to the relatively small units of text comprising the sentences within a single volume. We selected Washburn's *The Animal Mind* text book because of its historical significance, and modeled its 17,544 sentences as a collection of documents. To explore the power of topic modeling to identify latent but meaningful relationships at the micro-level,

we arbitrarily chose a sentence from Argument 15 in the Washburn set and used it to query the sentence-level model of *The Animal Mind* for the most similar sentences using the cosine of the sentence-topic vectors.

Results

The query sentence and the first half dozen results (with their similarity scores) are shown below.[5]

Query: Every statement that another being possesses psychic qualities is a conclusion from analogy, not a certainty; it is a matter of faith. (1.0000)

1. If any consciousness accompanies it, then the nearest human analogy to such consciousness is to be found in organic sensations, and these, as has just been said, must necessarily be in the human mind wholly different in quality from anything to be found in an animal whose structure is as simple as the Amoeba's. (0.8413)
2. Fancy, for example, one of us entering a room in the dark and groping about among the furniture. (0.8239)
3. This, of course, does not refer to the power to judge distance. (0.8235)
4. Again, a bodily structure entirely unlike our own must create a background of organic sensation which renders the whole mental life of an animal foreign and unfamiliar to us. (0.8224)
5. She disposes of the psychic learning by experience theory of Nagel by saying that the only experience upon which the animal could reject the filter paper must be experience that it is not good for food. (0.8198)
6. We speak, for example, of an "angry" wasp (0.7924)

Sentence 1 is obviously related in meaning to the query sentence: they overlap in some words, and directly express related ideas. But the relevance of the other examples is less direct. Sentence 6 provides a nice illustration of anthropomorphic attribution with no word overlap whatsoever. The inclusion of sentences 2 and 3 is, more puzzling. However, in the context of where these sentences appear in Washburn's book, the relationship become plainer. Sentence 2 comes in the context of the discussion of what it might be like to be an amoeba. It is thus related to sentence 1, and it is used by Washburn to make the point that our experience in the dark, which still involves visual imagination and memories of what we touch, must be "wholly different in quality" (per sentence 1) from what an amoeba might

[5]It is important to note that LDA topic modeling is a "bag of words" approach; i.e., it uses only an unordered list of words in each document. It has no information about word order, punctuation, or other formatting in the text, and some of the most common words are not included. The full sentences are shown here only to aid the reader.

experience. Sentence 3 occurs in a footnote on page 238, and it is worth quoting the footnote in full:

> Porter observed that the distance at which spiders of the genera Argiope and Epeira could apparently see objects was increased six or eight times if the spider was previously disturbed by shaking her web (612). This, of course, does not refer to the power to *judge* distance. [Italics in original.]

Here, then, we see the author cautioning the reader not to jump to a high-level interpretation of the spider behavior. The spiders may perceive objects at various distances but they don't judge it. The term 'judge' here is philosophically interesting, as it suggests an influence of Immanuel Kant on framing the debate. While Kant's name does not appear in Washburn's book, the term 'judgment' is important to Kant's theory of cognition, and fundamental to the cognitive divide he posits between humans and animals. We emphasize that this is just a speculative suggestion about Washburn's influences, but it does show how the topic modeling process can bring certain interpretive possibilities to the fore, moving the digital humanities another step closer to the goal of generating new insight into human intellectual activity.

Zooming Out Again: Macroanalysis by Science Mapping

Methods

We created visualization of the retrieved books overlaid on a map of science [33] to help understand the distribution of the retrieved books with respect to scientific disciplines.

New datasets are overlaid on this map by matching records via journal names or keywords to the 554 sub-disciplines. However, the present work is the first instance of using book data on a science map. We constructed a *classification crosswalk* to align the journal-based sub-disciplines with a book classification system. The Library of Congress Classification Outline (LCCO) provides a disciplinary taxonomy similar to that of the UCSD Map of Science. By using the Library of Congress Control Numbers (LCCN) assigned to each of the 25,258 journal sources in the UCSD Map of Science, we were able to assign likelihoods of each LCCN belonging to each subdiscipline.

We assigned each book in our HathiTrust collection a UCSD sub-discipline based on its LCCN. A number of items in the HathiTrust collection never receive LCCNs. For example, university library collections frequently contain course bulletins that are not catalogued by the Library of Congress. We removed the uncatalogued items and projected the remaining volumes onto the UCSD map of science.

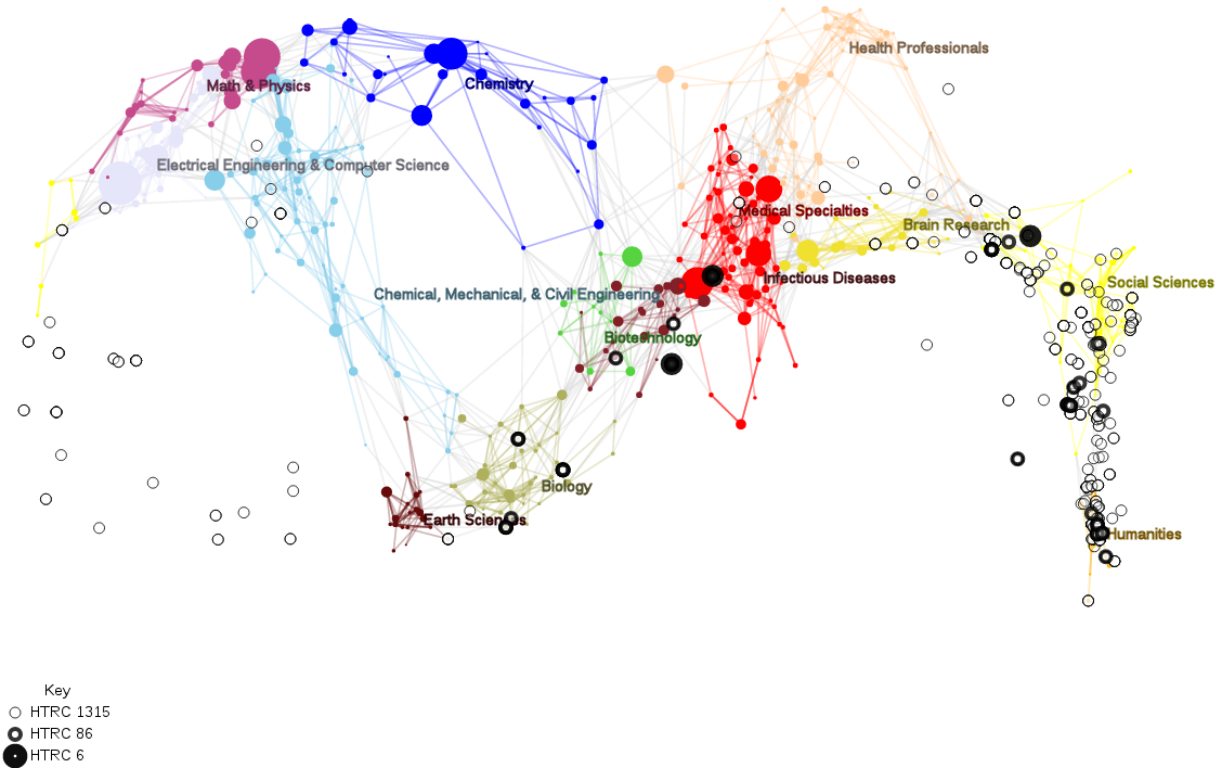

Figure 2: **UCSD Map of Science with overlay of HathiTrust search results shows topical coverage of humanities and life science data.** The basemap of science shows each sub-discipline denoted by a circle colored according to the 13 core disciplines. Links indicate journal co-citations from the basemap. The 776 volumes of HT1315 with LCCN metadata are shown on the map as circles. Volumes also in HT86 are shown with thicker circles, and those in HT6 are shown in the thickest circles. An online, interactive version can be explored at http://inpho.cogs.indiana.edu/scimap/scits.

Results

Using the LCCO classification crosswalk, we located 776 out of 1,315 books on the UCSD Map of Science, as shown in Figure 2.

In general, the map confirms that the initial keyword-based selection from the HathiTrust retrieved books that are topically positioned below the "equator" of the map, with particular concentrations in the life sciences and humanities, as was to be expected. The map provides additional visual confirmation that the further selections via topic modeling to a subset of 86 and then six of the original collection of 1,315 managed to target books in appropriate areas of interest. In the interactive online version, nodes can be selected, showing which volumes are mapped and providing the title and links to various external sources of metadata.

Ultimately, the map overlay provides a grand overview and a potential guide to specific books that were topic modeled, although without further guidance from the topic models,

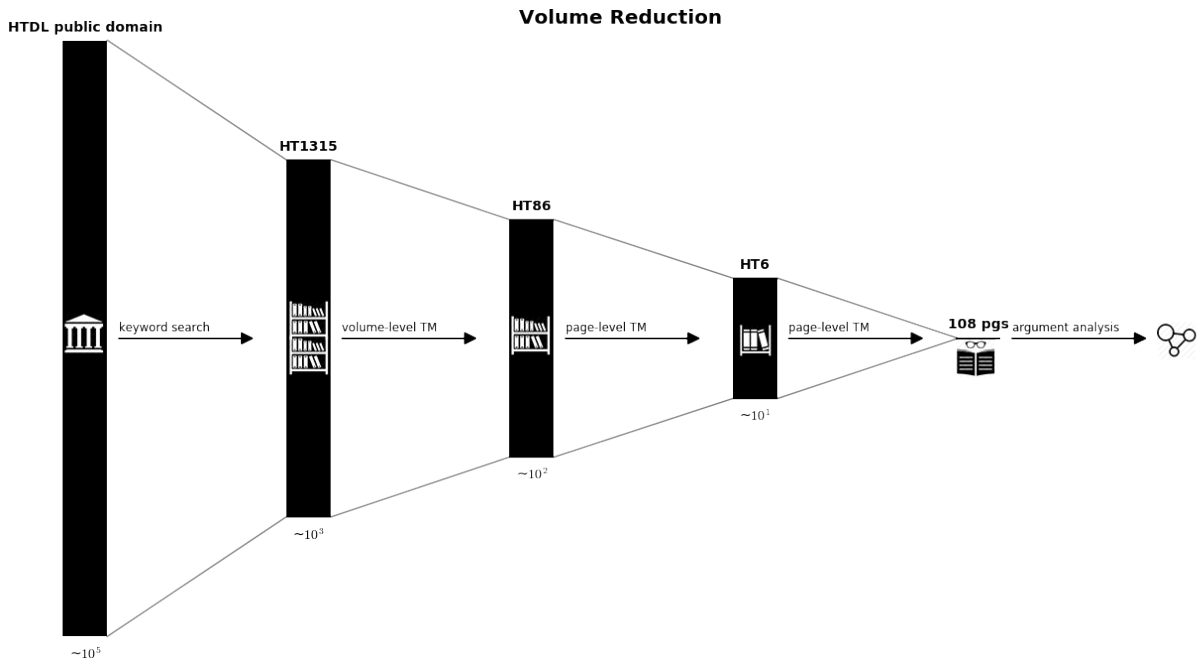

Figure 3: **Schematic rendering of the drill-down modeling process. The approximate order of magnitude is listed below each bar, which is scaled logarithmically.**

the map does not fully meet the desired objective of linking a high-level overview to more detailed textual analysis.

General Discussion

The notion of "distant reading" [23] has captured the imagination of many in the digital humanities. But the proper interpretation of large-scale quantitative models itself depends on having a feel for the texts, similar to Barbara McClintock's stress on having a "feeling for the organism" [52] or Richard Feynman on the importance for nascent physicists of developing "a 'feel' for the subject" beyond rote knowledge of the basic laws [53]. The interpretation of data and models, whether in science or the humanities, is itself (as yet, and despite a few small successes in fields such as medical diagnosis) a task at which humans vastly outperform machines. For this reason, the digital humanities remain a fundamentally hermeneutic enterprise [30], and one in which distant readings and close readings must be tightly linked if anything is to make sense.

In this paper we have motivated, introduced, and exemplified a multi-level computational process for connecting macro-analyses of massive amounts of documents to micro-level close reading and careful interpretation of specific passages within those documents. Thus we

have demonstrated how existing computational methods can be combined in novel ways to go from a high-level representation of many documents to the discovery and analysis of specific arguments contained within documents.

We have also shown how to zoom out to a macro-level overview of the search results. We presented a novel classification crosswalk between the Library of Congress Classification Outline (LCCO) and the UCSD Map of Science, which was constructed using only journal data, to extend the data to books. Because of the mismatch between the book data and the journal metadata, the crosswalk is not perfect, and the method of averaging locations places many books in uninterpretable regions of the map. Nevertheless, the visualization provides some useful information about the effectiveness of a simple keyword search in locating items of interest within a collection of hundreds of thousands of books.

That our method succeeded in discovering texts relevant to a highly specific interdisciplinary inquiry shows its robustness to inconsistent and incomplete data. The HathiTrust Digital Library had OCR errors in 2.4% of volumes as of May 2010 [54]. While the quality of the HathiTrust has increased in the intervening years, it is still a pervasive issue in digital archives [55].

Multi-level topic modeling combined with an information-theoretic measure of distance can efficiently locate materials that are germane to a specific research project, going from more than a thousand books, to fewer than a hundred using book-level topic models, and further narrowing this set down to a small number of pages within a handful of books using page-level topic models. The similarity measure we used is mediated by the topics in the model, and because every topic assigns a probability to every word in the corpus, this approach is highly adept at finding implicit relationships among the documents. Typical applications of topic modeling, such as graphing the rise and fall of topics through time, may show large-scale trends, but do not mediate the interplay between distant reading and close reading that leads to deeper understanding. By connecting abstract, machine-discovered topics to specific arguments within the text, we have shown how topic modeling can bridge this gap.

Conclusion

The process and results of our iterative drill-down method are summarized in Figure 3, showing the reduction of the 300,000 public domain volumes in the HathiTrust in August 2012 to the HT1315 collection, to the roughly 32,000 pages in the HT86 collection, to the over 17,000 sentences of the HT6 collection, to smaller set of the 108 pages selected for close reading and argument markup. This reduction allowed us to identify key elements of late 19th

and early 20th Century arguments about anthropomorphizing of nonhuman organisms, and to uncover the surprising taxonomic range of these arguments to include consideration even of consciousness in amoebae. The alternative approach of simply counting the occurrence of species names within these books would only have hinted at the presence of such discussions whereas, by putting words into context, topic modeling enabled researchers to zero in on passages worthy of detailed analysis and humanistic interpretation.

Acknowledgments

This work was funded by the National Endowment for Humanities (NEH) Office of Digital Humanities (ODH) Digging Into Data Challenge ("Digging by Debating"; PIs Allen, Börner, Ravenscroft, McAllister, Reed, and Bourget; award no. HJ-50092-12). The authors thank the Indiana University Cognitive Science Program for continued supplemental research funding, and especially for research fellowships for Jaimie Murdock and Robert Rose. We also thank the HathiTrust Research Center (HTRC) for their support of research activities and generous access to materials.

Author's contributions

CA motivated the research for the history and philosophy of comparative psychology. RL and JM constructed the LoC-UCSD crosswalk. DR, JO, and RR programmed the `vsm` topic model implementation. JM constructed the interactive UCSD map of science, with feedback from RL and KB. SM carried out the argument analysis using the OVA+ tool built by CR and JL, and a protocol devised by CA, DB, and AR. CA, RL, JM and KB wrote the paper.

References

[1] Bradford Demarest and Cassidy R Sugimoto. Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, 66(7):1374–1387, 2015.

[2] Rada Mihalcea. *Word Sense Disambiguation*, pages 1027–1030. Springer US, Boston, MA, 2010.

[3] Eneko Agirre and Philip Edmonds. Word Sense Disambiguation: Algorithms and Applications. *Text Speech and Language Technology*, 33:384, 2006.

[4] Christine L Borgman and Jonathan Furner. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36(1):2–72, jan 2002.

[5] Blaise Cronin. Scholarly communication and epistemic cultures. *New Review of Academic Librarianship*, 9(1):1–24, dec 2003.

[6] Kim Holmberg and Mike Thelwall. Disciplinary differences in Twitter scholarly communication. *Scientometrics*, 101(2):1027–1042, 2014.

[7] Riitta Kärki. Searching for bridges between disciplines: an author co-citation analysis on the research into scholarly communication. *Journal of Information Science*, 22(5):323–334, oct 1996.

[8] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Association for Computational Linguistics (ACL)*, abs/1605.0, 2016.

[9] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *CoRR*, abs/1606.0, 2016.

[10] Norman Kaplan. The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, 16(3):179–184, jul 1965.

[11] Mengxiong Liu. Progress in Documentation - The Complexities of Citation Practice: A Review of Citation Studies. *Journal of Documentation*, 49(4):370–408, 1993.

[12] Vincent Larivière, Éric Archambault, and Yves Gingras. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, 59(2):288–296, jan 2008.

[13] Sarah de Rijcke, Paul F Wouters, Alex D Rushforth, Thomas P Franssen, and Björn Hammarfelt. Evaluation practices and effects of indicator use–a literature review. *Research Evaluation*, 25(2):161–169, apr 2016.

[14] Andrew Odlyzko. The rapid evolution of scholarly communication. *Learned Publishing*, 15(1):7–19, jan 2002.

[15] James A Evans. Electronic Publication and the Narrowing of Science and Scholarship. *Science*, 321(5887):395 LP – 399, jul 2008.

[16] Jeremy York. This library never forgets: Preservation, cooperation, and the making of HathiTrust Digital Library. *Archiving 2009: Final Program & Proceedings*, 2009(1):5–10, 2009.

[17] Heather Christenson. HathiTrust: A research library at web scale. *Library Resources and Technical Services*, 55(2):93–102, 2011.

[18] Karen Coyle. Mass Digitization of Books. *The Journal of Academic Librarianship*, 32(6):641–645, nov 2006.

[19] Luc Vincent. Google Book Search: Document Understanding on a Massive Scale. In *International Conference on Document Analysis and Recognition, ICDAR'2007*, 2007.

[20] Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLoS ONE*, 4(11):e7678, nov 2009.

[21] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, The Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, jan 2011.

[22] Germinal Cocho, Jorge Flores, Carlos Gershenson, Carlos Pineda, and Sergio Sánchez. Rank Diversity of Languages: Generic Behavior in Computational Linguistics. *PLoS ONE*, 10(4):e0121898, apr 2015.

[23] Franco Moretti. *Distant Reading*. Verso Books, London, illustrate edition, 2013.

[24] Ted Underwood. Theorizing Research Practices We Forgot to Theorize Twenty Years Ago. *Representations*, 127(1):64–72, 2014.

[25] Thomas Kuhn. The Relations Between the History and the Philosophy of Science. In *The Essential Tension*, pages 3–20. 1979.

[26] Larry Laudan, Arthur Donovan, Rachel Laudan, Peter Barker, Harold Brown, Jarrett Leplin, Paul Thagard, and Steve Wykstra. Scientific change: Philosophical models and historical research. *Synthese*, 69(2):141–223, 1986.

[27] Ian Hacking. Two Kinds of "New Historicism" for Philosophers. *New Literary History*, 21(2):343–364, 1990.

[28] Scott B. Weingart. Finding the History and Philosophy of Science. *Erkenntnis*, 80(1):201–213, 2015.

[29] Katy Börner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Larivière, and Kevin W. Boyack. Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE*, 7(7):1–10, 2012.

[30] Geoffrey Rockwell and Stéfan Sinclair. *Hermeneutica*. MIT Press, Cambrdige, MA, 2016.

[31] Friedrich Steinle. Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science*, 64(S1):S65–S74, dec 1997.

[32] C. Kenneth Waters. The Nature and Context of Exploratory Experimentation : An Introduction to Three Case Studies of Exploratory Research. *History and Philosophy of the Life Sciences*, 29(3):275–284, 2007.

[33] Katy Börner. *Atlas of Science: Visualizing What We Know*. The MIT Press, 2010.

[34] Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. Cloud Computing Data Capsules for Non-Consumptive Use of Texts. In *ScienceCloud '14: Proceedings of the 5th ACM Workshop on Scientific Cloud Computing*, pages 9–16, Vancouver, BC, Canada, 2014.

[35] Boris Capitanu, Ted Underwood, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, and J. Stephen Downie. Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes (0.2)[Dataset], 2015.

[36] Christopher Fox. A stop list for general text. *SIGIR Forum*, 24(1-2):19–21, September 1989.

[37] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.

[38] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[39] Colin Allen and Marc Bekoff. *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. A Bradford book. MIT Press, Cambridge, 1999.

[40] Kristin Andrews. Animal cognition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2016 edition, 2016.

[41] Colin Allen and Michael Trestman. Animal consciousness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition, 2015.

[42] Robert Boakes. *From Darwin to Behaviourism: Psychology and the Minds of Animals*. Cambridge University Press, 1984.

[43] Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. page 11, sep 2015.

[44] David M Blei. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84, apr 2012.

[45] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[46] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[47] Jaimie Murdock and Colin Allen. Visualization Techniques for Topic Model Checking. In *Proceedings of the 29th AAAI Conference (AAAI-15)*, Austin, TX, 2015. AAAI Press.

[48] Carlos Chesñevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316, dec 2006.

[49] John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. AIFdb: Infrastructure for the Argument Web. In *Frontiers in Artificial Intelligence and Applications*, volume 245, pages 515–516. Vienna, 2012.

[50] John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, jun 2014. Association for Computational Linguistics.

[51] Chris Reed, Douglas Walton, and Fabrizio Macagno. Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review*, 22(1):87–109, mar 2007.

[52] Evelyn Fox Keller. *A Feeling for the Organism: The Life and Work of Barbara McClintock*. W.H. Freeman and Co., San Francisco, 1983.

[53] Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *Lectures on Physics, Volume 1*. Addison-Wesley, Redwood City, CA, 1964.

[54] Paul Conway. Measuring Content Quality in a Preservation Repository: HathiTrust and Large-Scale Book Digitization. In *Proceedings of 7th International Conference on Preservation of Digital Objects, iPres 2010*, pages 95–102, Vienna, 2010.

[55] Diana Kichuk. Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books , 2015.