Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, 1838-1860

Jaimie Murdock[1,2], Simon DeDeo[1,2,3], and Colin Allen[1,4,5]

[1]Program in Cognitive Science, Indiana University, Bloomington, IN 47405, USA
[2]School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
[4]Department of History and Philosophy of Science and Medicine, Indiana University, Bloomington, IN 47405, USA
[5]School of Humanities and Social Sciences, Xi'an Jiaotong University, Xi'an, China

September 7, 2015

Significance — Charles Darwin's Theory of Evolution by Natural Selection is one of the most revolutionary breakthroughs in the history of science. Between 1838 and 1860, Darwin meticulously documented the books he read. His Reading Notebooks thus enable the study of inputs to his creative process during the 22 years spanning his disembarkment from the Beagle to the publication of *The Origin of Species*. We located 669 of his English nonfiction readings and applied topic modeling to the full-text of these readings. We then used the semantic space of the topic models in a novel way to measure the distances that Darwin traveled between books. These measurements permitted us to investigate the trade-off he made between reading in-depth and leaping to new domains. Our analysis shows that Darwin's behavior shifts from exploitation to exploration on multiple timescales, and that at the longest timescale these shifts correlate with major intellectual epochs of his career. Furthermore, contrasting his reading order with the publication order of the same texts, we find Darwin's consumption of the texts is more exploratory than the culture's production of them.

Abstract

Search in a dynamic environment with an uncertain distribution of resources involves a trade-off between local exploitation and distant exploration. This extends to the problem of information foraging, where a knowledge-seeker shifts between reading in-depth and studying new domains. We examine the choices made by one of the most celebrated scientists of the modern era: Charles Darwin. Darwin built his theory of natural selection in part by synthesizing disparate parts of Victorian science. When we analyze Darwin's extensively self-documented reading, we find shifts, on multiple timescales, between choosing to remain with familiar topics and seeking cognitive surprise in novel fields. On the longest timescales, these shifts correlate with major intellectual epochs of his career, as detected by Bayesian epoch estimation. When we compare Darwin's reading path with publication order of the same texts, we find Darwin more adventurous than the culture as a whole.

Scientific innovation occurs against a cultural background of accumulating ideas. Individual researchers can be viewed as conducting a cognitive search [1] in which they must balance *exploration* of ideas that are novel to them against *exploitation* of existing knowledge in domains in which they are already expert [2]. The general problem of "information foraging" [3] in an environment about which agents have incomplete information has been explored in many fields including cognitive psychology [1, 4], neuroscience [5], economics [6, 7], finance [8], ecology [9, 10], and computer science [11]. In all of these areas, the searcher aims to enhance future performance by surveying enough of existing knowledge to orient themselves in the information space.

Researchers have studied information foraging at timescales of minutes (*e.g.*, laboratory experiments on visual attention [12]) up to, in large populations, the course of years and decades (*e.g.*, in the recombination of patented technologies [13]). New advances in the digitization of historical archives allow us to curate biographically-plausible datasets to study how a single individual, over the course of a lifetime, explores and synthesizes the work of contemporaries and predecessors.

As one of the most successful and celebrated scientists of the modern era, Charles Darwin's scientific creativity has been the subject of numerous narrative and qualitative studies [14–16]. In part, these studies are possible because Darwin left his biographers many resources with which to study his life, including reading diaries maintained for the critical period from 1838 to 1860, culminating in the publication of *The Origin of Species*. See Table 1 for a summary of major events in Darwin's personal and intellectual life.

We present here the first quantitative analysis of these reading diaries, tracking how – over time – he navigated the crucial exploration-exploitation tradeoff of search in structured environments. After linking these records with the full text of the original volumes, we use probabilistic topic models [17, 18] to create a representation of the texts Darwin explored, where the distances between texts are quantified through an information-theoretic analysis of their underlying topics. We use Bayesian epoch estimation over these models, combined with biographical scholarship, to correlate Darwin's long-term behavioral shifts with significant events in his life.

Our focus on the reading patterns of a single individual allows us to see how a single agent explores and arranges available artifacts. This contrasts with previous uses of topic modeling to analyze the large-scale structure of scientific disciplines [19, 20] and the humanities [21–23], which are created through many people's collective behavior. Previous models of historical records have focused on word frequency as an indication of larger shifts in style [24, 25] or content [26, 27] of significant portions of publications in a field. Modeling the collective state of all published works at a particular date may obscure the role of individual foraging behavior. By focusing on a single individual for whom ample records exist, we gain access to what Tria et al. [28] describe as "the interplay between individual and collective phenomena where innovation takes place".

We present three key findings: first, Darwin's reading patterns appear to switch, on multiple timescales, between exploration and exploitation. This is in contrast to an "efficient" surprise-minimization strategy – where 'surprise' is defined information-theoretically in terms of divergence between probability distributions derived from the texts – that exploits content within a local region before moving on. Second, on the longest timescale, mode switching between these strategies falls into three epochs. These correspond to three biographically significant periods: Darwin's post-*Beagle* studies, his extensive work on barnacles, and a final period leading to his synthesis of natural selection in the *Origin of Species*. Third, in comparison to the order in which the texts Darwin read were published, Darwin's reading order shows higher cumulative surprise. This indicates that society-at-large accumulates innovations more gradually than an individual consumes them.

Major Events in Charles Darwin's Life (1809-1882)	
12 Feb 1809	Born in Shrewsbury, England
22 Oct 1825	Matriculates at University of Edinburgh
15 Oct 1827	Admitted to Christ's College, Cambridge
27 Dec 1831	Departs England aboard the *HMS Beagle*
2 Oct 1836	Return to England aboard the *HMS Beagle*
Apr 1838	**Begins reading notebooks**
29 Jan 1839	Marries Emma Wedgwood
Aug 1839	Publication of *The Voyage of the Beagle* (1st edition)
May 1842	Writes the 1st Essay on Species
17 Sep 1842	Moves to Down House in Kent, his home until death
4 July 1844	Writes the 2nd Essay on Species
Aug 1845	Publication of *The Voyage of the Beagle* (2nd edition)
1 Oct 1846	Begins barnacle project
19 Feb 1851	Publishes first volume of barnacle work
23 Mar 1851	9-year-old daughter Annie Darwin dies of scarlet fever
9 Sep 1854	Begins sorting notes on natural selection
14 May 1856	Starts writing "large work" on species
24 Nov 1859	Publication of *The Origin of Species* (1st edition)
13 May 1860	**Last entry in reading notebooks**
24 Feb 1871	Publication of *The Descent of Man*
19 Feb 1872	Publication of *The Origin of Species* (6th and final edition)
21 Apr 1882	Dies at Down House in Kent, England

Table 1: *Timeline* – A brief summary of major events in Charles Darwin's life, including those marked on Fig. 2. This paper focuses on the critical period of his work from 1838-1860 leading to the publication of *The Origin of Species*. See [29] for an expanded, but still condensed, chronology.

1 Methods

Darwin was a meticulous record-keeper—starting in April 1838, he kept a notebook of "books to be read" and "books read". Even more remarkable is that these records span the 22 years from 1838-1860, tracking his reading from just after his return to England aboard the *HMS Beagle* to just beyond the publication of the *The Origin of Species*[1]. In all, the two reading notebooks ([30, 31] – transcribed by [32]) contain 1,248 titles identified by the Darwin Correspondence Project [33], of which 915 were marked "read". We reduced this list of 915 to the 688 English-language non-fiction titles[2], and located 628 of these within the HathiTrust Digital Library[3] and an additional 41 at the Internet Archive[4] for a total of 669 titles[5].

[1] See Table 1 for a condensed timeline of Darwin's life.

[2] We chose to focus on English-only text due to technical complications with multilingual corpora [34], and chose non-fiction text to avoid cross-domain issues.

[3] http://hathitrust.org/

[4] http://archive.org/

[5] See Supplementary A for more details on corpus creations, including the complete corpus, list of excluded volumes, list of stopwords and additional preparation done.

We model these texts using using probabilistic topic models [17, 18], varying the number of topics, k, to test the robustness of our results[6]. This allows us to describe Darwin's reading as taking place in a $(k-1)$-dimensional space, the simplex, where a particular volume is described as a probability distribution, $\vec{p}$, over k topics. Darwin's "semantic voyage" is the track he leaves through this space, as he moves from text to text.

To capture the cognitive structure of this path, we use the Kullback-Leibler (KL) divergence. The KL divergence quantifies the "surprise" of a optimal learner trained on distribution $\vec{q}$, when encountering a new distribution $\vec{p}$. It is defined as

$$D_{\mathrm{KL}}(\vec{p}, \vec{q}) = \sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i}. \tag{1}$$

where $\vec{p}$ is the new distribution, and $\vec{q}$ the baseline. As with many information-theoretic quantities, KL has many conceptually distinct, but consistent, interpretations [35]. Another interpretation is that KL quantifies the inefficency of an optimal code for a distribution $\vec{q}$ when it is used to encode a time-stream drawn from $\vec{p}$.

For these reasons, KL can be considered a measure of cognitive load due to novelty [36], and closely-related measures have been shown in laboratory experiments to attract human attention [37]. Among its many uses in the study of human cognition, it has been used to quantify language-learning [38, 39], selectional preferences in word choice [40, 41], information gathering in syntactic processing [42], and as a measure of resource allocation in syntactic comprehension [43].[7] In general, lower KL in this analysis indicates that the new text is efficiently encoded given knowledge of the previous text: a reader choosing texts nearby to those already encountered.

We use KL divergence in two distinct ways. We measure the text-to-text surprise: given a distribution over topics for the text Darwin just read, $\vec{q}$, how surprised is he upon encountering the distribution $\vec{p}$ associated with the next? We also measure the past-to-text surprise: given all of the volumes that Darwin has encountered so far, how surprised is Darwin by the text that comes next?

Text-to-text surprise and past-to-text surprise provide complementary windows onto Darwin's decision-making. Local decision-making, meaning the choice of the next text to read given the current one, is captured by text-to-text surprise. Global decision-making, the choice of which text to read given the entire history of reading to date, is captured by past-to-text surprise. Low surprise, in either case, is a signal of *exploitation*, while high surprise indicates larger jumps to lesser-known topics, and thus of *exploration*.

Darwin's decision process is characterized by the combination of text-to-text and past-to-text surprise. These local and global behaviors do not have to align — text-to-text surprise may be high (local exploration) at the same time that past-to-text surprise is low (global exploitation). This pattern can happen, for example, when Darwin repeatedly "sweeps" over a series of topics, interleaving concepts.

Conversely, text-to-text surprise can be low while past-to-text surprise can be high. This local-exploitation/global-exploration pattern can happen when, for example, Darwin has recently begun a novel, but focused, investigation. In this situation, he focuses on a particular subset of topics that

[6] In the main body of the paper we report results for $k = 80$. See Supplemental B for further discussion of model selection and results of models for $k = \{20, 40, 60\}$.

[7] In much of this work, the key concept is "surprise"–the average relative log-likelihood of an event under two different distributions. Kullback-Leibler divergence is simply the average surprisal under the new distribution.

are under-represented in his overall history. Exploration at both scales can happen when Darwin is moving across a space not previously explored. Exploitation at both scales can happen when, Darwin has a sustained focus on material he is already familiar with.

In order to consider the cultural process of creation alongside the individual, cognitive process of consumption, we examine the publication of the texts themselves. To do this, we measure text-to-text and past-to-text surprise in the publication order of the texts relevant to Darwin's research[8].

In all cases, we compare measured values to a null model that allows us to determine when jumps are unusually near (or far) in either rank or distance, while holding Darwin's overall reading list, and dates, fixed. To generate null replicates, we re-sample without replacement from Darwin's original reading list, being sure not to allocate a book to a reading date before its publication time[9]. In contrast to a purely-random permutation, this null captures the dynamics of publication in which a new work can unexpectedly change the information space.

1.1 Bayesian Epoch Estimation

In the foraging literature, individuals are often assumed to persist in either exploration or exploitation for a sustained period. We call this an *epoch*. We are particularly interested in whether or not these epochs align with important events in Darwin's life. Independent of this qualitative biographical scholarship, we can determine whether or not there is quantitative evidence for epochs and the position of epoch switches using Bayesian models.

Bayesian epoch estimation relies upon an underlying model with a variable number of parameters. These parameters describe the epoch start and end points, and the mean and variance of the text-to-text or past-to-text surprise distributions within each epoch. Epoch switches can occur at the local and global level independently. We use a simple model-complexity penalty, Akaike Information Criterion (AIC – [44]), to determine the number of parameters, and thus the number of epochs, that best fit our data. See Supplemental D for more information on our model.

2 Results

2.1 Exploration and Exploitation

Over the 593 records in our corpus, Darwin's reading order led to a below-null cumulative surprise. On average, the KL divergence from text to text in the corpus is 10.65 bits compared to a null expectation of 11.42 bits ($p \ll 10^{-3}$). Darwin's past-to-text average surprise, meanwhile, is 3.03 bits in the data versus to 3.06 bits in the null ($p = 0.039$).

While Darwin's cumulative text-to-text surprise is lower than expected from a null model, it is far larger than many paths that can be found: a greedy shortest-path algorithm, for example, can reduce the average surprise to 0.72 bits. The difference between Darwin's decision process and this shortest-path choice is shown visually in Fig. 1, where his long-range jumps are shown as

[8]We are unable to resolve publication dates to less than a year; this occasionally implies an ambiguity in creation date for texts in the corpus that have the same year of publication. To solve this, we average our results in these cases, in a Bayesian fashion, over all possible within-year orders, with uniform prior.

[9]See Supplemental B for a discussion of why this null is more rigorous than a null model constructed against all the books available to him in Kent and London during these years, rather than books Darwin actually read.

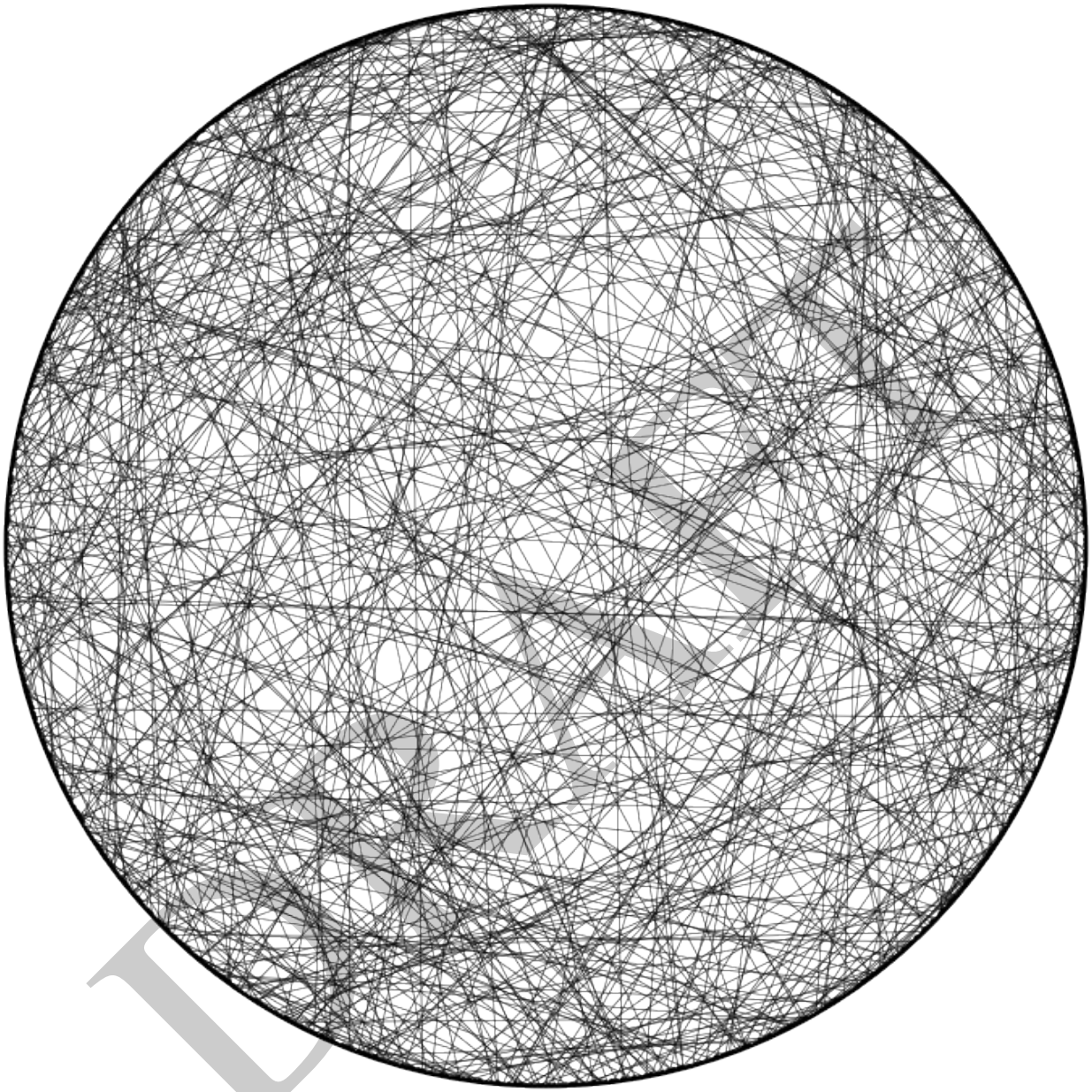

Figure 1: *Darwin's reading network.* The 593 records in the corpus are represented by nodes and arranged around the circumference of the graph. Directed arrows indicate reading order. Nodes themselves are ordered, clockwise from top, according to a greedy shortest-path algorithm. If Darwin choose books by minimizing information-theoretic surprise, his path would be purely on the circumference. While many of Darwin's steps are nearly circumferential (*e.g.*, the red arrow), there are a number of long-range jumps, where Darwin crosses the semantic space (*e.g.*, the blue arrow).

center-crossing directed arrows.

Darwin's reading patterns, in other words, do not show strong effects of pure surprise-minimization constraints. By contrast, we can examine how text-to-text and past-to-text surprise accumulates over time. This is shown in Fig. 2 for the book-to-book case (top panel) and the past-to-book case (bottom panel). A negative (downward) slope in these lines indicates reading decisions by Darwin that produce below-null instantaneous surprise. Tracking the slopes in these charts allows us to see how Darwin moves between low-surprise and high-surprise text-to-text and past-to-text decision rules on a range of timescales. The interaction of these two decision rules characterize Darwin's behavior. Over the entire corpus in both the text-to-text and past-to-text case, Darwin's cumulative surprise is below the null expectation, showing an overall bias towards local and global exploitation.

Bayesian epoch estimation finds three main epochs: (1) from the start of our records in 1837 until Spring 1846, when past-to-text surprise changes from exploitation to exploration, (2) from Spring 1846 until Autumn 1856, when text-to-text surprise changes from exploitation to exploration, and (3) from Autumn 1856 to the end of our data.

In the first epoch, both text-to-text and past-to-text surprise remain low – a regime of simultaneous local and global exploitation. Biographically, the dates of this epoch correspond to the early phase of Darwin's post-*Beagle* writings, where his readings were largely confined in natural history and geology as part of his work on the two editions of *The Voyage of the Beagle*, and the first two drafts of his theory on species transmutation and natural selection in Spring 1842 and Autumn 1844.

The second epoch sees a rise in past-to-text surprise, indicating a shift from global exploitation to global exploration, while maintaining local exploitation. While new readings in this epoch are focused, they now range over a subset of topics previously under-represented – particularly in botany, zoology, and paleontology. Biographically, the dates of this epoch correspond to a significant period in Darwin's life when, in response to concerns about the rigor of his theories on species transmutation, he began an 8-year investigation into the biology of barnacles.

The final epoch is characterized by a shift in text-to-text surprise from low-surprise to high-surprise – a regime of both local and global exploration. Darwin is neither repeatedly returning to well-covered topics (as in epoch one), nor turning his attention to a new, but narrow, range (as in epoch two). Instead, he ranges widely over new, previously understudied topics, such as the ornithological books read in 1856 to build his case study of pigeons. This final epoch begins within days of his journal entry on September 16, 1854 indicating that he has begun sorting his notes for his great synthesis.

2.2 Individual and Collective

One of the central roles of innovation is the recombination of past ideas. While many studies see scientific innovations as following large-scale temporal trends (*e.g.*, [46]), individuals can also be known as "ahead of their time" [15, 47]. By ordering Darwin's readings by the publication date, rather than reading date, we see how the culture gradually accumulates and assimilates knowledge. We can then compare how the culture produced texts to how Darwin, in his reading, consumed them.

[11] Darwin did not track the month and day of reading until October 2, 1838.

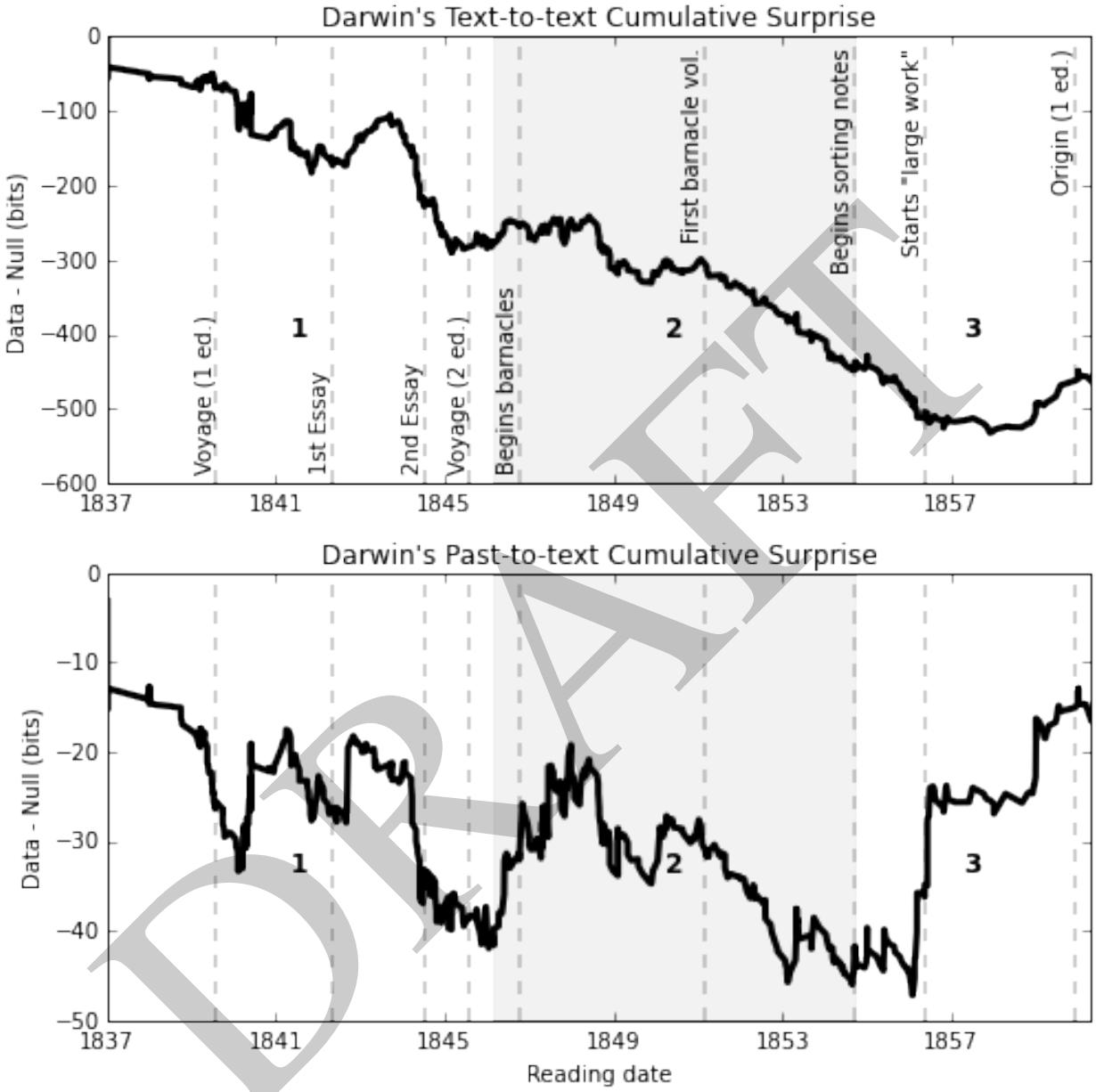

Figure 2: *Cumulative Surprise*. Average book-to-book (top) and past-to-book (bottom) cumulative surprise over the reading path measured as the cumulative KL divergence (bits). The three epochs are marked as alternating shaded regions with key biographical events marked as dashed lines and labeled in the top graph. The first epoch shows global and local exploitation (decreasing surprise). The second epoch shows local exploitation and global exploration (increasing surprise, bottom only). The third epoch shows local and global exploration (increasing surprise)

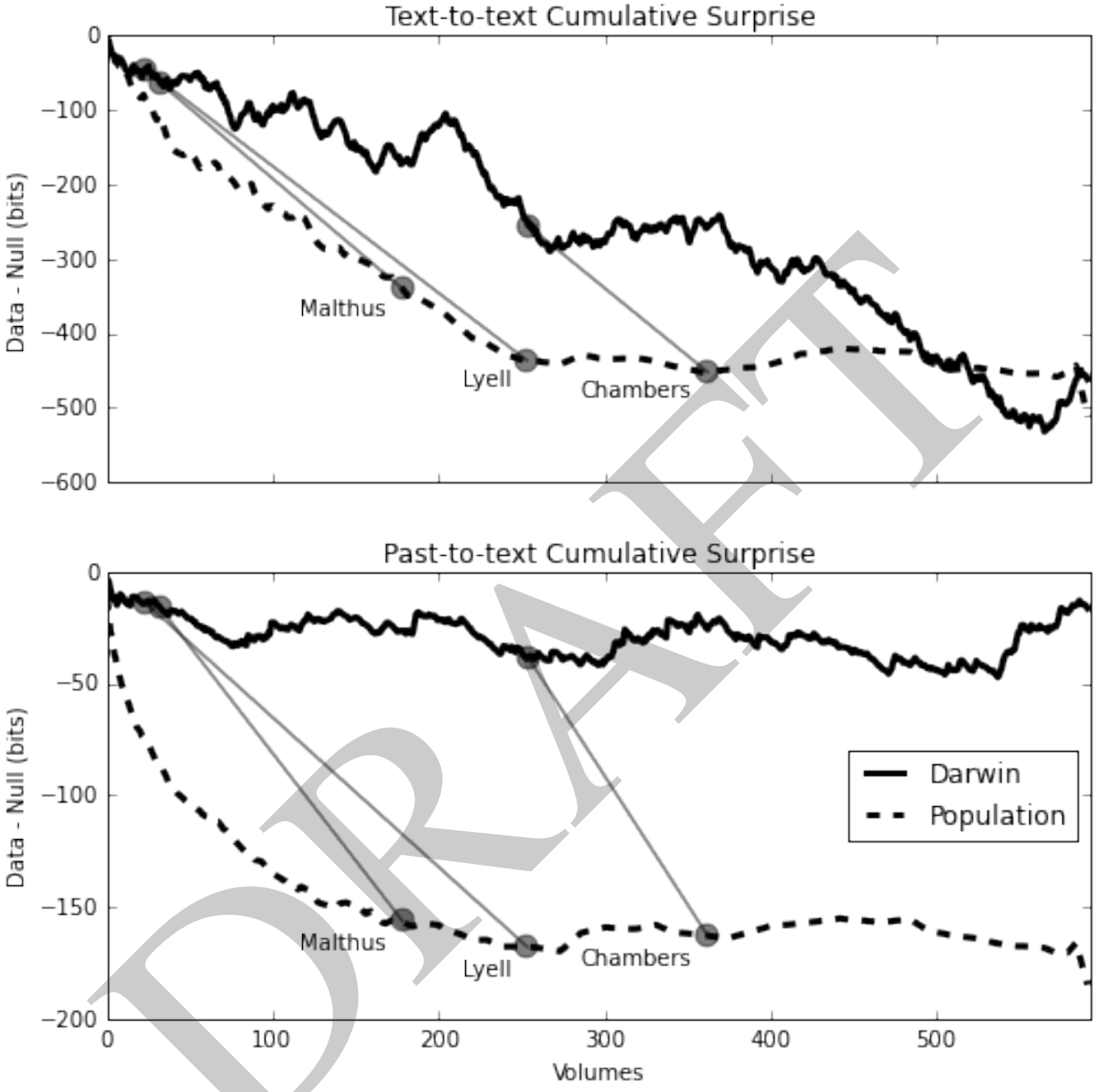

Figure 3: *Cumulative Surprise, Individual and Collective* — Average book-to-book (top) and past-to-book (bottom) cumulative surprise over the reading order (solid) and over the publication order (dashed), measured in bits. In both book-to-book and past-to-book, Darwin's cumulative surprise remains lower than the time-dependent null, and the publication order remains even lower. Three key volumes in Darwin's intellectual development are marked: Charles Lyell's *Principles of Geology* (3rd ed., 1837; read in 1837[11]), Thomas Malthus's *An Essay on the Principle of Population* (1803; read on October 3, 1838), and Robert Chambers's *Vestiges of the Natural History of Creation* (1844; read on November 20, 1844).

Figure 3 shows the text-to-text and past-to-text surprise for Darwin's reading order (solid line) compared to the publication date order (dashed line). Since volumes are published and read at different times, the x-axis is now ordinal (*i.e.*, by position in the reading or publication sequence), rather than temporal (*i.e.*, by date read or published). This allows us to compare his reading order to the publication order independent of time.

Compared to Darwin's reading practices, cultural production has far lower rates of surprise. While cumulative text-to-text surprise for Darwin often shows either flat or positive (above-null text-to-text surprise) slope, the publication order of Darwin's readings is far less divergent in both text-to-text and past-to-text cumulative surprise. This suggests that among the texts Darwin read, society accumulates facts gradually and largely exists in a regime of exploitation.

Epochs like those corresponding to Darwin's shifting reading strategies are non-existent in the population data. **TODO:** add evidence.

3 Discussion

Darwin's seminal work rendered a large body of diverse evidence intelligible, transforming biology in the process. He synthesized ideas from all branches of biology, and from a broad range of other areas represented in his reading list, such as geology, political economy, and philosophy. Darwin's willingness to explore widely beyond biology had serendipitous consequences for the development of his theory of evolution. He wrote in his autobiography:

> In October 1838, that is, fifteen months after I had begun my systematic inquiry, I happened to read for amusement Malthus on Population, and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favorable variations would tend to be preserved, and unfavorable ones to be destroyed. The results of this would be the formation of a new species. Here, then I had at last got a theory by which to work.

Even though this insight came near the beginning of the first epoch described in this paper, Darwin's continuing exploration and exploitation of the content of the books available to him was critically important to the theory of evolution he eventually developed. The process of natural selection acting on populations of organisms, linked via descent with modification, provided an innovative solution to the problem of explaining the great diversity and distribution of life on Earth.

Darwinian theory has in turn provided insights into processes of cultural change, including the study of innovation itself as a multi-level combinatoric process, in which assemblages of ideas are subject to forms of cultural selection analogous to natural selection [48, 49]. Such applications of an evolutionary framework typically consider cultural change at the population level, as new ideas are created, spread, and modified by the crowd. More generally, a variety of recent studies covering conceptual formation in science, technology, and the humanities have also pursued a population-level perspective. These include work on the recombination of patents [13], novelties [28], and citations [50]. Sociological studies of scientific practice have investigated how disciplines [46] or "communities of practice" [51] are formed.

There is room for both population-level and individual-level processes in the study of cultural innovation, just as there is in evolutionary biology itself [52]. The mechanisms driving population-

level processes leading to cultural innovation cannot, however, be fully understood without taking into account the cognitive processes operating at the level of individual scientists. We have taken a step towards modeling such individual-level processes by studying the information foraging behavior of one preeminent scientist, using an information-theoretic framework applied to probabilistic topic models of his reading behavior. Information-theoretic quantities such as KL divergence connect both analytically and empirically to cognition [36–43]. Because of its generality, an information-theoretic approach to foraging may also provide a new means to tie individual-level studies into more general frameworks, such as that proposed by Berger-Tal et al.[2]

4 Conclusion

An analysis of the full-text of Darwin's readings shows him switching between exploration and exploitation modes at multiple timescales. While his overall path shows lower surprise than a null-model permutation of his reading list, he does not follow a strategy of pure surprise-minimization, a form of exploitation that might be expected from simple coding efficiency arguments. On the longest timescales, our epoch estimation technique uncovers three epochs that align with important historical periods of his intellectual life: his post-*Beagle* studies, his extensive work on barnacles, and a final period leading to his synthesis of natural selection with modification by descent. In contrast to his actual reading order, the publication order appears to follow a lower-surprise path characterized by exploitation.

In this paper, we laid the groundwork for information-theoretic analysis of an individual scientist's reading habits, represented by topic models of full-text readings, to measure exploitation and exploration of culturally available resources. By focusing on a single individual for whom ample records exist, we gain access to what Tria et al. [28] describe as "the interplay between individual and collective phenomena where innovation takes place". In doing so, we have made possible a more detailed, multi-leveled understanding of scientific innovation. Future investigations will move beyond reading strategies and explore the relationship between reading and writing, or consumption and production, of scientific output.

Maintaining a reading diary was not unique to Darwin, as evidenced by the more than 30,000 records in the UK Reading Experience Database, 1450-1945[12]. Furthermore, many figures have maintained extensive "commonplace books" recording quotes, readings, and interactions that may become useful in their later intellectual life, from Marcus Aurelius [53] to Francis Bacon [54], John Locke [55], and Thomas Jefferson [56].[13] While Darwin's notebooks are particularly well-organized for our purposes, we hope these techniques can be applied to other great thinkers, giving insight into the individual cognitive processes from which our collective culture springs.

Independently, Darwin's sustained engagement with the products of his culture is remarkable. Including works of fiction and foreign-language texts not included here, he averaged one book

[12] http://www.open.ac.uk/Arts/reading/

[13] As Virginia Woolf [57] notes, these commonplace books are all too common: "[L]et us take down one of those old notebooks which we have all, at one time or another, had a passion for beginning. Most of the pages are blank, it is true; but at the beginning we shall find a certain number very beautifully covered with a strikingly legible hand-writing. Here we have written down the names of great writers in their order of merit; here we have copied out fine passages from the classics; here are lists of books to be read; and here, most interesting of all, lists of books that have actually been read, as the reader testifies with some youthful vanity by a dash of red ink."

every ten days for twenty-two years. For some months in our data, Darwin appears to be reading one book every other day, a fact even he was astonished by:

> When I see the list of books of all kinds which I read and abstracted, including whole series of Journals and Transactions, I am surprised at my industry.
> — *Autobiography of Charles Darwin*, p. 119 .

Darwin not only consumed information, it consumed him. In the words of Herbert Simon, "what information consumes is rather obvious: it consumes the attention of its recipients" [59]. Even the most ambitious individuals must confront and manage the limits of their own biology in allocating attention. They leave traces of that management in the records they leave behind. Now we can exploit those traces through innovative new methods of digital scholarship.

Acknowledgments

We thank Peter M. Todd for extensive comments on a draft of this manuscript. We also thank numerous members of the Indiana University Cognitive Science Program for their presentation feedback. Jaimie Murdock and Simon DeDeo thank the Santa Fe Institute for hospitality while this work was completed. We thank Tom Murphy for assistance with corpus curation and Robert Rose for programming assistance. Tools for corpus preparation and modeling were produced by Robert Rose and Jaimie Murdock while supported by the National Endowment for the Humanities Digging Into Data Challenge (NEH HJ-50092-12, Colin Allen, co-PI). Jaimie Murdock and Colin Allen were supported by an Indiana University (IU) Office of the Vice Provost for Research (OVPR) Faculty Research Support Program (FRSP) Seed Funding Grant and Bridge Funding Grant. Jaimie Murdock was also supported by an IU Cognitive Science Program Supplemental Research Fellowship.

References

[1] Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins, editors. *Cognitive Search: Evolution, Algorithms, and the Brain*. MIT Press, Cambrdige, MA, 2012.

[2] Oded Berger-Tal, Jonathan Nathan, Ehud Meron, and David Saltz. The exploration-exploitation dilemma: A multidisciplinary framework. *PLoS ONE*, 9(4), 2014.

[3] Peter Pirolli and Stuart Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

[4] Thomas T Hills, Peter M Todd, David Lazer, A David Redish, and Iain D Couzin. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1):46–54, March 2015.

[5] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, May 2007.

[6] James G March. Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1):71–87, February 1991.

[7] Rina Azoulay-Schwartz, Sarit Kraus, and Jonathan Wilkenfeld. Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision Support Systems*, 38(1):1–18, October 2004.

[8] Juha Uotila, Markku Maula, Thomas Keil, and Shaker A Zahra. Exploration, exploitation, and financial performance: analysis of S&P 500 corporations. *Strategic Management Journal*, 30(2):221–231, 2009.

[9] David W Stephens and John R Krebs. *Foraging theory*. Princeton University Press, 1986.

[10] Sigrunn Eliassen, Christian Jø rgensen, Marc Mangel, and Jarl Giske. Exploration or exploitation: life expectancy changes the value of learning in foraging strategies. *Oikos*, 116(3):513–523, March 2007.

[11] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT Press Cambridge, 1998.

[12] Marvin M Chun and Jeremy M Wolfe. Just say no: How are visual searches terminated when there is no target present? *Cognitive psychology*, 30(1):39–78, 1996.

[13] Hyejin Youn, Deborah Strumsky, Luis M.A. Bettencourt, and José Lobo. Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society*, 12(106):1–8, 2015.

[14] Howard E Gruber and Paul H Barrett. *Darwin on man: A psychological study of scientific creativity.* EP Dutton, 1974.

[15] Steven Johnson. *Where good ideas come from: The natural history of innovation*. Penguin UK, 2010.

[16] D. Van Hulle. *Modern Manuscripts: The Extended Mind and Creative Undoing from Darwin to Beckett and Beyond*. Historicizing Modernism. Bloomsbury Academic, 2014.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[18] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April 2012.

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

[20] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[21] John W. Mohr and Petko Bogdanov. Introduction — topic models: What they are and why they matter. *Poetics*, 41(6):545 – 569, 2013. Topic Models and the Cultural Sciences.

[22] David Blei. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1):8–11, 2012.

[23] M.L. Jockers. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press, 2013.

[24] James M Hughes, Nicholas J Foti, David C Krakauer, and Daniel N Rockmore. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686, 2012.

[25] Ted Underwood and Jordan Sellers. The Emergence of Literary Diction. *Journal of Digital Humanities*, 1(2), 2012.

[26] 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, 2011.

[27] Andrew Goldstone and Ted Underwood. The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, 45(3):359–384, 2014.

[28] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H Strogatz. The dynamics of correlated novelties. *Scientific reports*, 4, 2014.

[29] Tim M Berra. *Charles Darwin: the concise story of an extraordinary man*. John Hopkins University Press, Baltimore, MD, 2009.

[30] Charles Darwin. *'Books to be read' and 'Books Read' notebook*. 1838-1851. CUL-DAR119.-Transcribed by Kees Rookmaaker.

[31] Charles Darwin. *'Books to be read' and 'Books Read' notebook*. 1852-1860. CUL-DAR128.-Transcribed by Kees Rookmaaker.

[32] Peter J. Vorzimmer. The Darwin reading notebooks (18381860). *Journal of the History of Biology*, 10(1):107–153, 1977.

[33] Frederick Burkhardt and Sydney Smith, editors. *Darwin's Reading Notebooks*. 1989. Darwin Correspondence Project. http://www.darwinproject.ac.uk/darwins-reading-notebooks.

[34] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82, Arlington, Virginia, United States, 2009. AUAI Press.

[35] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[36] John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[37] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[38] Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124, 2013.

[39] Shira Calamaro and Gaja Jarosz. Learning general phonological rules from distributional information: A computational model. *Cognitive Science*, 39(3):647–666, 2015.

[40] Philip Stuart Resnik. Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series*, page 200, 1993.

[41] Marc Light and Warren Greiff. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281, 2002.

[42] Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193 – 210, 2008.

[43] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177, 2008.

[44] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.

[45] Charles Darwin. *Journal of Researches into the Geology and Natural History of the various countries visited by H.M.S. Beagle*. Henry Colburn, 1838.

[46] Xiaoling Sun, Jasleen Kaur, Stasa Milojevic, Alessandro Flammini, and Filippo Menczer. Social Dynamics of Science. *Sci. Rep.*, 3, January 2013.

[47] N.T. Bliss, B.R.E. Peirson, D. Painter, and Manfred D. Laubichler. Anomalous subgraph detection in publication networks: Leveraging truth. In *48th Asilomar Conference on Signals, Systems and Computers*, pages 2005–2009, Nov 2014.

[48] F Jacob. Evolution and tinkering. *Science*, 196(4295):1161–1166, June 1977.

[49] Andreas Wagner and William Rosen. Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. *Journal of the Royal Society, Interface*, 11(97):20131190–, 2014.

[50] Eugene Garfield. *Citation Indexing – Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, Inc., 1979.

[51] Luis M.A. Bettencourt and David I Kaiser. Formation of Scientific Fields as a Universal Topological Transition. *arXiv eprint*, 1504.00319v1, 2015. SFI Working Paper 2015-03-009.

[52] Elisabeth Lloyd. Units and levels of selection. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[53] Marcus Aurelius. *Meditations*. Sheba Blake Publishing, 2015.

[54] Francis Bacon. *The Promus of Formularies and Elegancies*. Longman, Greens and Company, 1883.

[55] John Locke. *A New Method of Making Common-Place-Books*. Printed for J. Greenwood, bookseller, at the end of Cornhil, next Stocks-Market, 1706.

[56] D.L. Wilson. *Jefferson's Literary Commonplace Book*. Papers of Thomas Jefferson, Second. Princeton University Press, 2014.

[57] Virginia Woolf. *Hours in a Library*. Harcourt, Brace and Co., 1958.

[58] Charles Darwin. *The life and letters of Charles Darwin, including an autobiographical chapter*. John Murray, 1887.

[59] Herbert A Simon. Designing organizations for an information-rich world. *Computers, communication, and the public interest*, 37:40–41, 1971.

[60] Jaimie Murdock and Colin Allen. Visualization techniques for topic model checking. *Proceedings of 29th Association for the Advancement of Artificial Intelligence (AAAI-15)*, 2015.

[61] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. OReilly Media Inc., 2009.

[62] Margaret Roberts, Brandon Stewart, and Dustin Tingley. Navigating the Local Modes of Big Data: The Case of Topic Models. In *Data Analytics in Social Science, Government, and Industry*. Cambridge University Press, New York, 2015.

A Corpus Characterization

Despite our use of a digital library, it is important to remember that books are originally physical artifacts (see Fig. 4), and Victorian publishing practices often spread a single title over multiple volumes for portability and ease of use. In this paper, we use *volume* to refer to each physical artifact. Each individual entry of Darwin's notebooks is refered to as a *title*. In the case of books, a *title* gathers together one or more volumes. In the case of journal articles, a *title* is merely a subpart of a particular volume. We model at the level of a *catalog record*, which corresponds to a *title* in almost all cases, except for journals, where it corresponds to the aggregate of all issues listed as read across entries in the notebooks.

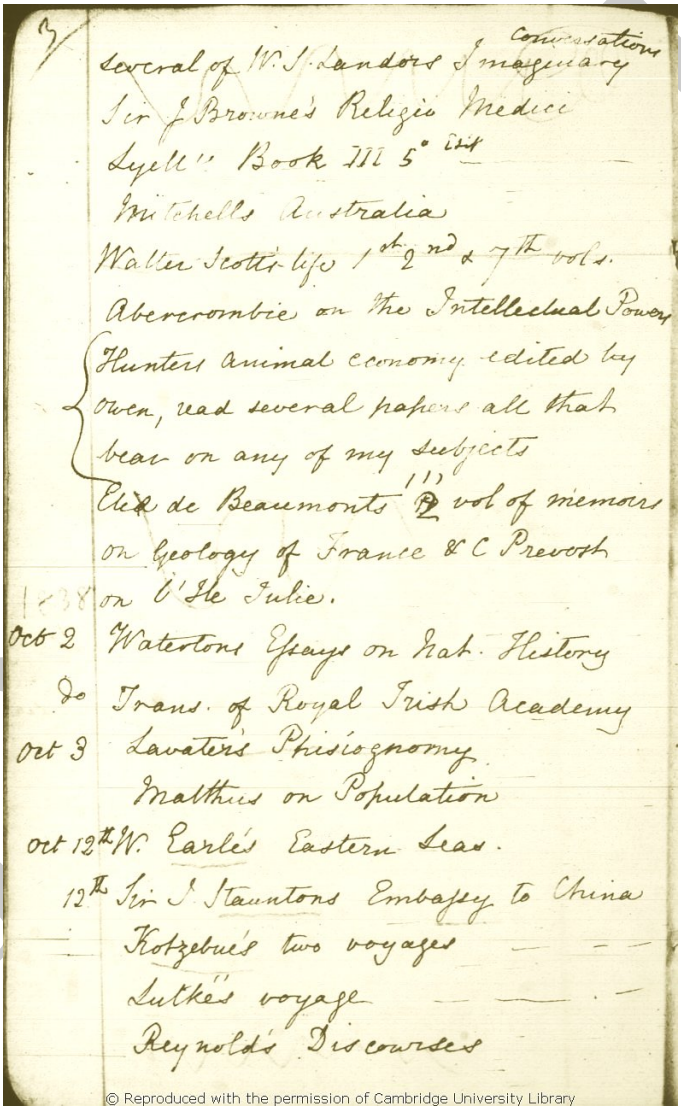

Figure 4: *Darwin's Reading Notebooks*. Page 3a of Darwin's first notebook (DAR 119), during which he began to track the exact dates. Note the reading of Malthus's *On Population* on October 3, 1838. Photo courtesy of Cambridge University Libraries and the Darwin Online Project.

Darwin also read French, German, and Latin texts. We reduced the corpus to English-only to reduce cross-linguistic effects in the model [34]. Additionally, we focused only on non-fiction texts. An examination of the influence of fiction on Darwin is a topic for further exploration.

There are 593 catalog records and 1132 volumes corresponding to the 669 titles modeled in this study. Some volumes in the corpus alignment were unable to be matched to the exact edition listed by the DCP, and thus there is occasionally a difference between the volume Darwin read and the volume whose text we use for topic modeling. Table 2 shows the summary of the items which were located and remain missing. All texts are available in the source file listed in Supplemental E.

Our publication dates are those listed by the Darwin Correspondence Project (DCP); the DCP uses the publication date of the volume, if found in Darwin's library, otherwise the date of first publication. The reading order is determined by dates listed in the reading notebook. When multiple titles are listed at a particular date, we use their natural ordering in the notebooks — titles written at the top of the page are assumed to be read before those at the bottom.

	Located	Non-located	Total
Total	**811**	**104**	**915**
- Fiction	- 79	- 1	- 80
- Non-English	- 63	- 84	- 147
English Non-fiction	**669**	**19**	**688**

Table 2: *Corpus Composition:* Composition of the Reading List in terms of fiction, non-fiction, English, and non-English texts. Located titles refers to the number identified in the HathiTrust (`http://hathitrust.org/`), Internet Archive (`http://archive.org/`), and Project Gutenberg (`http://gutenberg.org/`). Non-located texts were unavailable in the HathiTrust, Internet Archive, or Project Gutenberg as of August 5, 2015.

We use the InPhO Topic Explorer [60] for tokenization and modeling of texts. First, plain-text editions downloaded from the HathiTrust, the Internet Archive, and Project Gutenberg are normalized by merging cross-line hyphens into single words, normalizing into ASCII using Unidecode[14], removing all words containing punctuation and numerals (often due to OCR errors), and lower-casing all words. Then, words appearing in the English stopwords corpus from the Natural Language Toolkit (NLTK – [61]) are removed. Finally, words occurring less than 5 and more than 12,000 times were excluded from the corpus. After pre-processing, the corpus consisted of 38,719,172 tokens drawn from 204,562 unique tokens. We made no attempt to apply stemming or clean up OCR errors, other than the filtering of words occurring fewer than 5 times.

Fig. 5 shows the density of Darwin's readings modeled. Notice the large jump in 1840 corresponds to a period when he was reading entire series of journals, each article of which was a separate title in his notebook. Also, note that Fig. 5 shows both the density of the selection modeled and the entire reading notebook list.

Fig. 6 indicates that as Darwin's readings progress he begins reading more recently published, contemporary sources. We also show the regression for the un-modeled texts, showing that his total reading also progressed toward contemporary sources, although at a slower rate.

[14] https://pypi.python.org/pypi/Unidecode

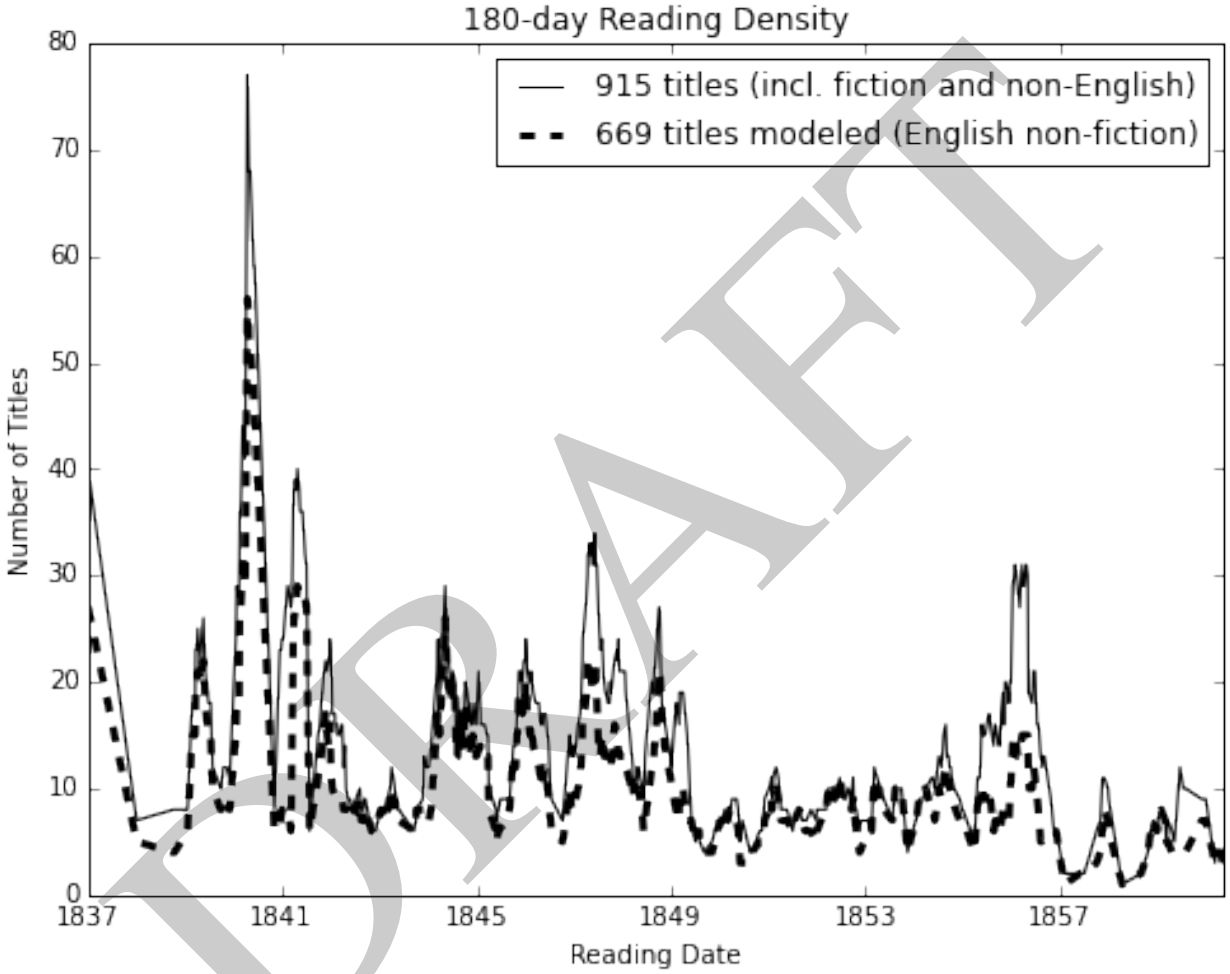

Figure 5: *Reading Density* – Reading density, smoothed over a 6-month window. The dashed line shows the 669 titles here modeled, while the thin solid line represents all 915 titles in the reading notebooks.

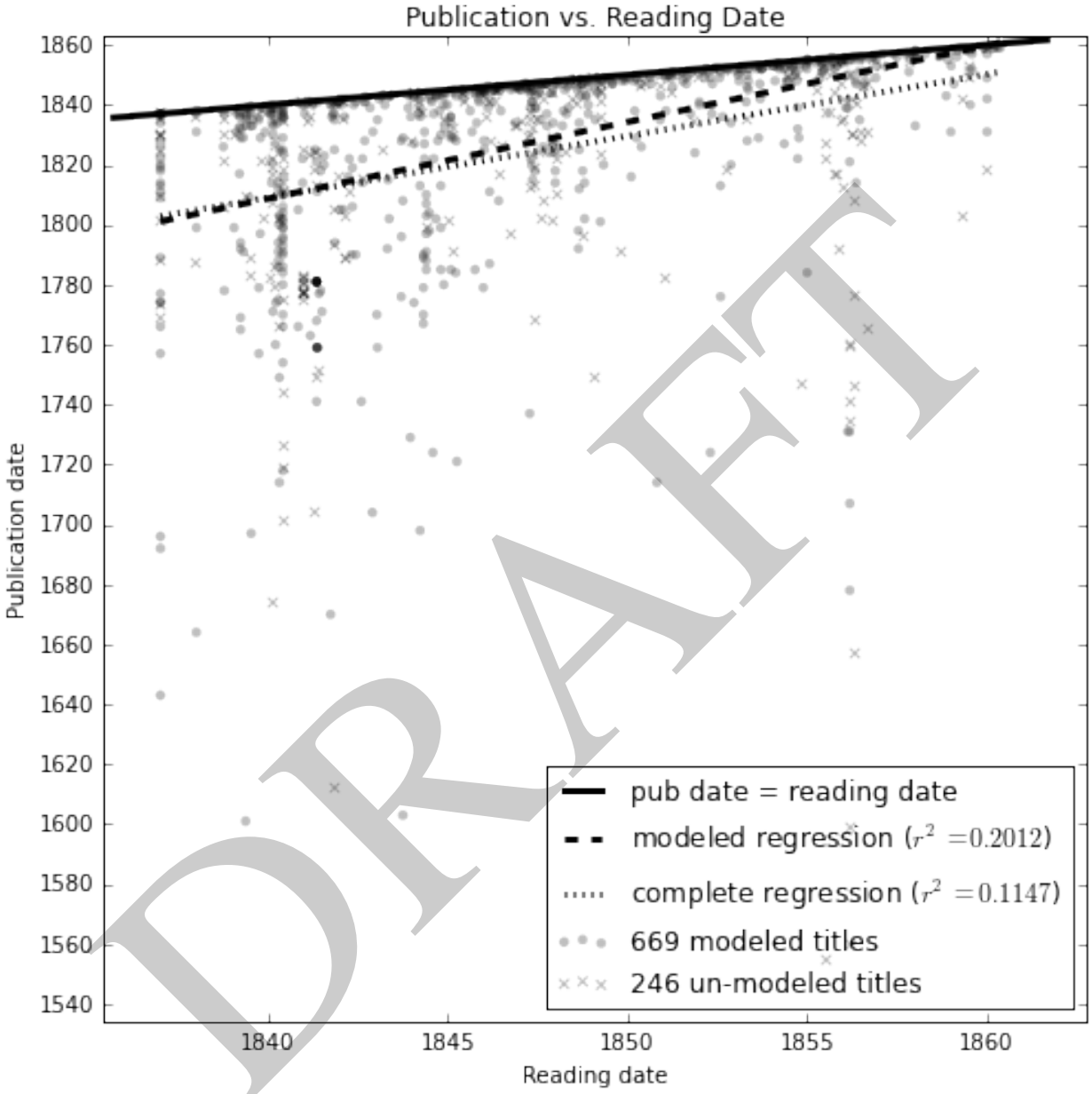

Figure 6: *Publication vs. Reading Dates* – Scatter plot of the publication and reading dates of the titles in Darwin's reading list. The 669 modeled titles are shown with dots, while the remaining 246 titles are shown as xs. The solid line indicates when the reading date and publication date are equal. The dashed line indicates a linear regression over the dots ($r^2 = 0.2012$), and the dotted line indicates a linear regression over the dots and xs combined ($r^2 = 0.1147$).

B Null Model Justification

A more complete representation of the state of Victorian science (i.e. Darwin's entire search space) would require the null model to be constructed against all the books available to him in Kent and London during these years, rather than books Darwin actually read. To construct and model such a corpus would be a monumental task, and would be circumscribed anyway by the subsequent curatorial decisions that have shaped present access to digitized Victorian era texts. Fortunately, the null model based on Darwins own reading list provides a more rigorous test of our results. This is because text-to-text surprise in the larger set is expected to be greater, thus accentuating the difference between data and null for Darwin's lower-surprise trajectory. Similarly, past-to-text surprise should be greater in a null model constructed against a broader set of books. This is because whether the prior state of the "null" reader based on one text or many, the model of the larger corpus provides more opportunities for long range jumps.

C Local and Cumulative KL

Table 3 shows the raw local text-to-text and cumulative past-to-text KL divergence data, along with the greedy shortest path and greedy longest path single-visit traversals of the KL distance matrix.

	Local (bits/step)	Cumulative (bits/step)
Measured	9.81	2.84
Null	10.3 ± 0.3	$2.85^{+0.03}_{-0.02}$
(p-value)	$\ll 10^{-3}$	n.d.
Greedy Shortest Path	2.82	2.75
Greedy Longest Path	13.71	3.01

Table 3: *Exploration habits*. Average book-to-book and past-to-book KL Divergence (bits/step) over the reading path. Past-to-book KL is much lower, as Darwin's reading spreads out to cover topic space and lowers the information-theoretic surprise of subsequent books. Book-to-book, Darwin achieves lower step-size than the null—but far larger than (retrospective) shortest paths.

C.1 Model Robustness

The "model checking problem" is an enduring problem for applied topic modeling [18], but recent work on selection of a "reference model" in the social sciences provided guidance to selecting a value of $k = 80$ for the number of topics to use in our analysis [62]. More specifically, setting $k = 80$ produced a set of topics subjectively deemed more interpretable than the lower values of k suggested by more "objective" measures of model fit to data.

In addition to the $k = 80$ topics results shown in the main paper, the same analyses are also shown below for $k = 20, 40, 60$ in Figs. 7, 8, 9, respectively.

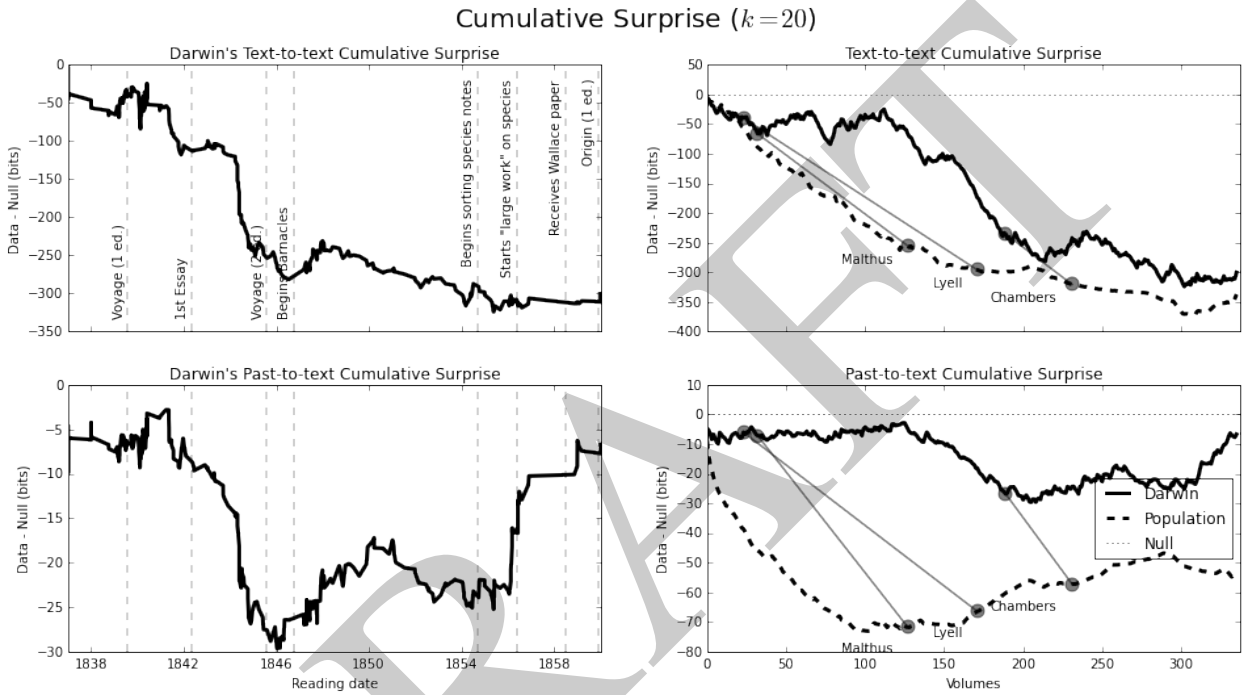

Figure 7: *Cumulative and Cultural Surprise* – Analysis of Figures 2 (left) and 3 (right) repeated for $k = 20$. *Left:* Average book-to-book (top left) and past-to-book (bottom left) cumulative surprise over the reading path and over the publication history, measured as the cumulative KL divergence (bits). As Darwin drops below zero in these plots, his choices are producing surprises lower than expected in the null. *Right:* Average book-to-book (top) and past-to-book (bottom) cumulative surprise over the reading order (solid) and over the publication order (dashed), measured in bits. In both book-to-book and past-to-book, Darwin's cumulative surprise remains lower than the time-dependent null, and the publication order remains even lower.

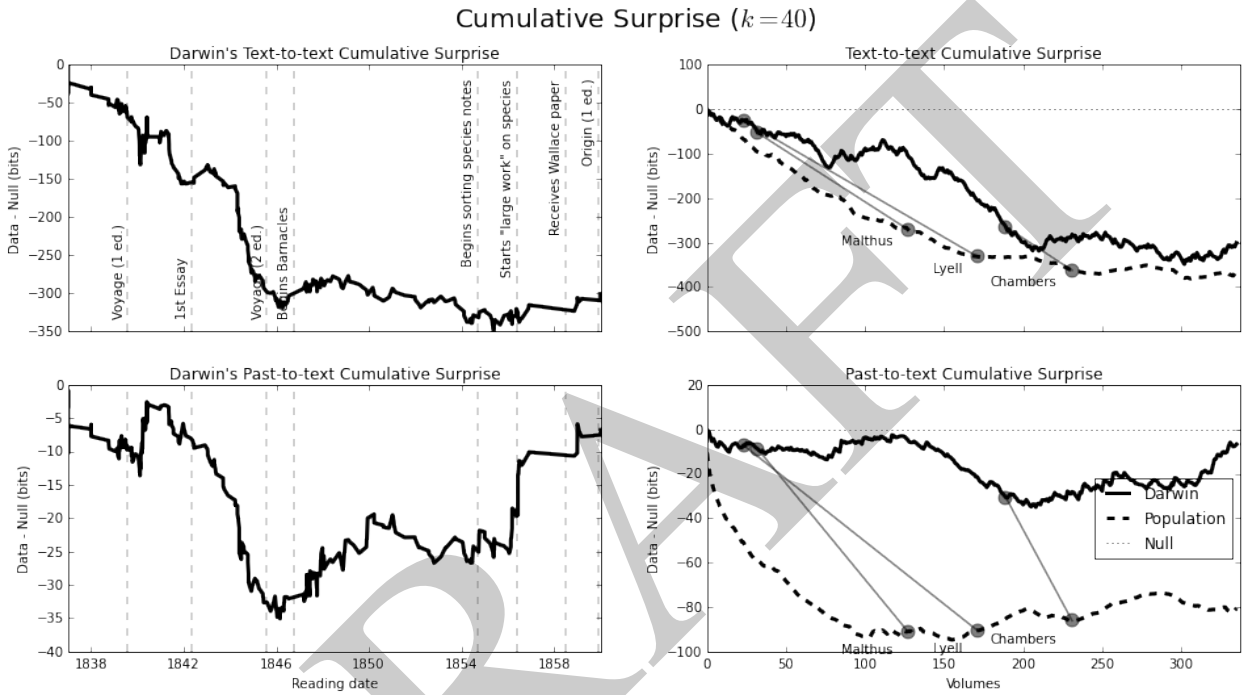

Figure 8: *Cumulative and Cultural Surprise* – Analysis of Figures 2 (left) and 3 (right) repeated for $k = 40$. *Left:* Average book-to-book (top left) and past-to-book (bottom left) cumulative surprise over the reading path and over the publication history, measured as the cumulative KL divergence (bits). As Darwin drops below zero in these plots, his choices are producing surprises lower than expected in the null. *Right:* Average book-to-book (top) and past-to-book (bottom) cumulative surprise over the reading order (solid) and over the publication order (dashed), measured in bits. In both book-to-book and past-to-book, Darwin's cumulative surprise remains lower than the time-dependent null, and the publication order remains even lower.

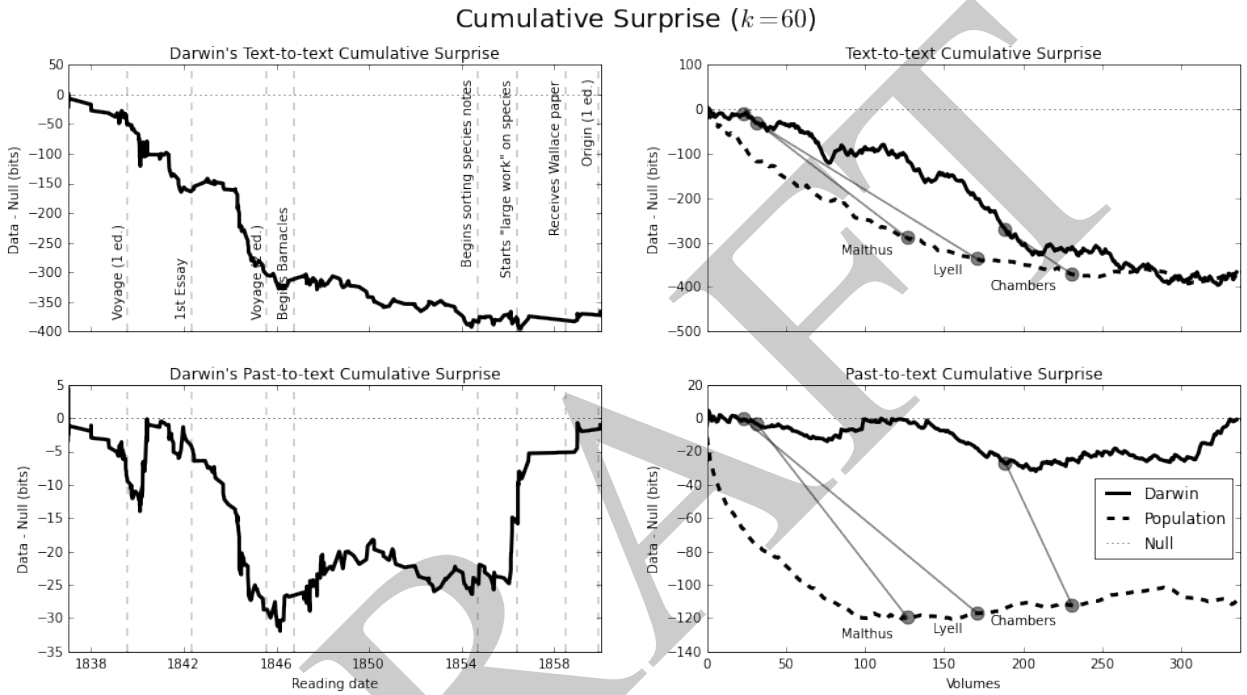

Figure 9: *Cumulative and Cultural Surprise* – Analysis of Figures 2 (left) and 3 (right) repeated for $k = 60$. *Left:* Average book-to-book (top left) and past-to-book (bottom left) cumulative surprise over the reading path and over the publication history, measured as the cumulative KL divergence (bits). As Darwin drops below zero in these plots, his choices are producing surprises lower than expected in the null. *Right:* Average book-to-book (top) and past-to-book (bottom) cumulative surprise over the reading order (solid) and over the publication order (dashed), measured in bits. In both book-to-book and past-to-book, Darwin's cumulative surprise remains lower than the time-dependent null, and the publication order remains even lower.

C.2 Rank Distribution

In addition to the information-theoretic measures described in the paper, descriptive statistics also capture Darwin's explore-exploit behavior. For each volume, we look at the rank of the KL divergence to the next volume by reading order compared to all other volumes in the corpus, as shown in Fig. 10. We can compare this to a null model, as described in the Methods.

Interestingly, Darwin is 11 times more likely than the null model to pick the nearest neighbor as to pick a volume farther away, indicating that explorations are overall rarer than exploitations, and emphasizing that exploitations do indeed occur on a text-to-text basis.

D Bayesian Epoch Estimation

Our generative model for Bayesian epoch estimation has $(n-1)$ parameters to describe the end points of the first $n-1$ epochs, and $2n$ parameters to describe the mean and variance of the text-to-text (or past-to-text) surprise within each epoch. We estimate these parameters using an approximate maximum-likelihood procedure. Within each epoch i, we assume the surprise is constant and Gaussian distributed with a particular mean μ_i and variance σ_i^2. We write the $3n-1$ parameters as a vector $\vec{v}$; then the distribution over $\vec{v}$ given the data s, equal to a list of surprises, $\{s_i\}$, is

$$\log P(\vec{v}|s) = \log P(s|\vec{v}) + C = \sum_{i=1}^{n} \frac{(e_{i+1} - e_i - 1}{2} \left(1 + \ln(2\pi \hat{\sigma^2}_i\right) + C, \tag{2}$$

where e_i is the start point of epoch i and C depends on the prior. The start point of the first epoch, e_1, is fixed to be volume zero; given our conventions, e_{n+1} is fixed to be the final volume plus one. The sigma estimator, $\hat{\sigma^2}_i$, is the standard maximum likelihood estimator of the variance,

$$\hat{\sigma^2}_i = \frac{1}{e_{i+1} - e_i - 1} \sum_{k=0}^{e_{i+1}-e_i-1} \left(s_k - \hat{\mu}_i\right)^2, \tag{3}$$

and $\hat{\mu}_i$ is defined as

$$\frac{1}{e_{i+1} - e_i - 1} \sum_{k=0}^{e_{i+1}-e_i-1} s_k. \tag{4}$$

To do Fisher maximum-likelihood estimation, we ignore the effect of the prior $P(\vec{v})$ on the maximum; equivalently, we do maximum a posteriori estimation and assume that $P(\vec{v})$ is flat over the region of interest. To choose the number of parameters, we use the Akaike Information Criterion [44]; we increase the number of epochs until the increase in the log-likelihood is less than the complexity penalty, equal to the number of parameters.

E Source Code

E.1 Corpus

The complete list of citations, together with the reading dates, HathiTrust Catalog Number, and individual volume IDs are available at `http://darwinsmind.org/reading/arXiv.csv`.

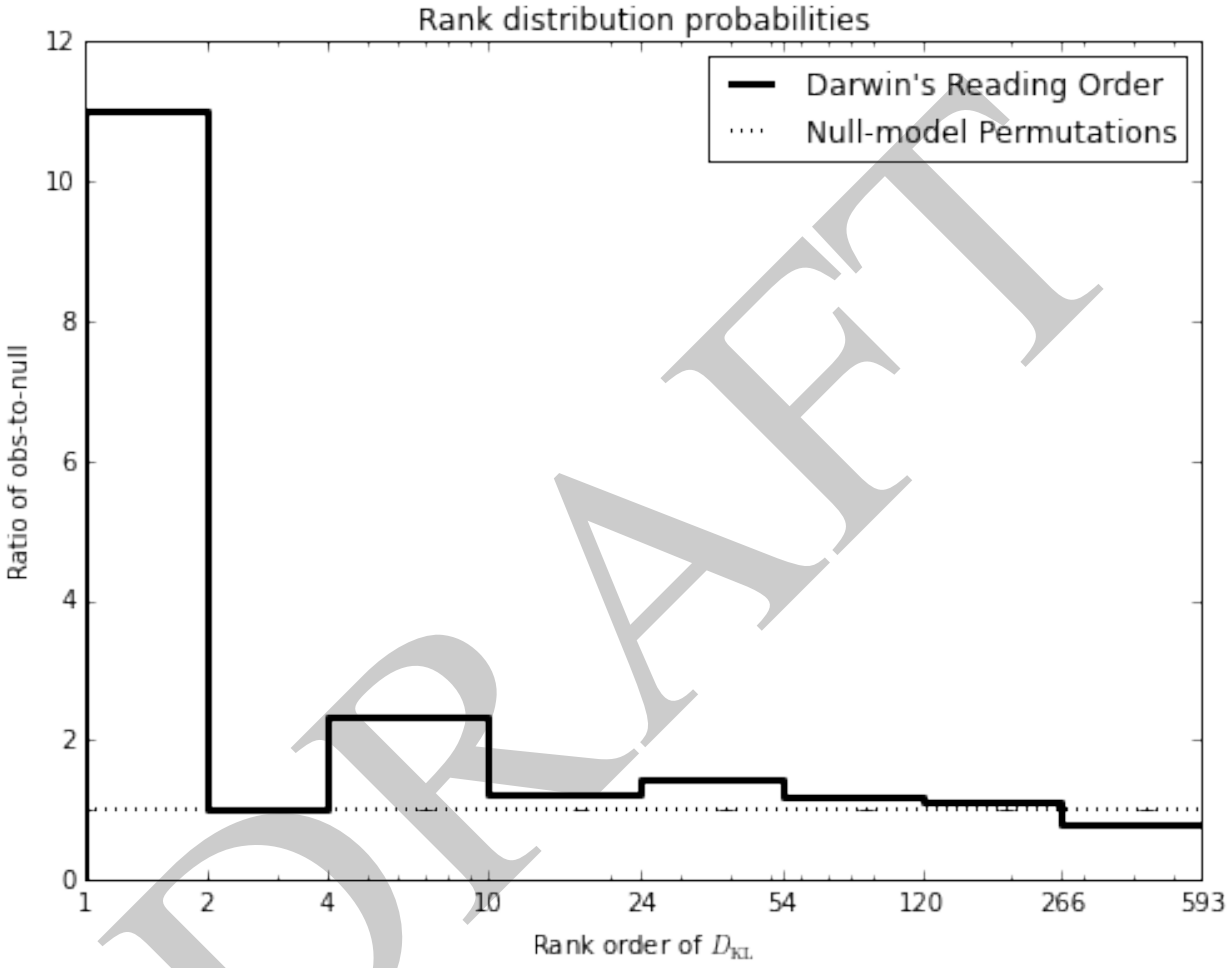

Figure 10: *Rank Distribution*. Rank distribution of $D_{\mathrm{KL}}(\theta_i, \theta_{i+1})$ for Darwin's reading notebooks relative to a null-model permutation of his reading order, as indicated by the dashed line, with 95% confidence intervals shown. The lines are logarithmically binned, showing clearly that Darwin is 11 times more likely to select the nearest KL neighbor, as opposed to volumes further away, which are selected 0.80 times as likely than the null.

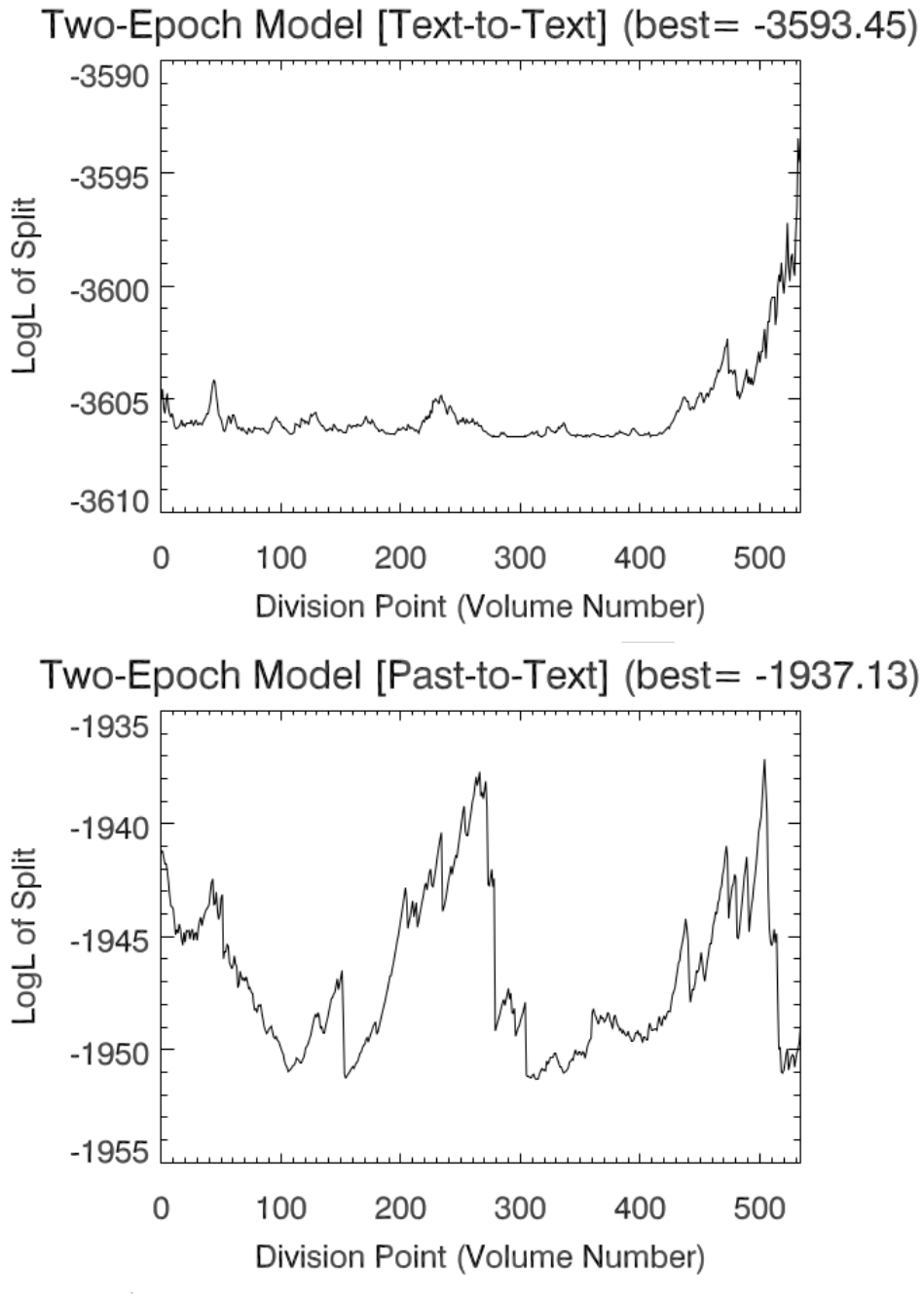

Figure 11: *Bayesian Epoch Estimation* – Fisher maximum-likelihood estimation for a 2 epoch Bayesian Epoch Estimation model over the text-to-text and past-to-text $k = 80$ models of 593 of Darwin's readings. Note the phase transition at the 307th volume in the past-to-text case and the 500th volume in the text-to-text case. Note also that the past-to-text case comes close to transition at the 500th volume as well, indicating the strength of the transition to exploration in the third epoch on both local and global scales.

The list of stopwords is available at http://darwinsmind.org/reading/arXiv/v1/stopwords.txt. The pre-processed corpus is available at http://darwinsmind.org/reading/arXiv/v1/XXX.zip.

E.2 Models

All models were generated with the InPhO Topic Explorer, available at http://github.com/inpho/topic-explorer/. Model files for $k = \{20, 40, 60, 80\}$ are available at http://darwinsmind.org/reading/arXiv/v1/models.zip.

The raw topic-document matrix and word-topic matrix are available as http://darwinsmind.org/reading/arXiv/v1/theta-top-dox.csv and http://darwinsmind.org/reading/arXiv/v1/phi-word-top.csv, respectively. The list of Darwin's reading order, ranked by KL divergence, is available at http://darwinsmind.org/reading/arXiv/v1/ranked-kl.csv and may be of particular use to Historians of Science interested in what leaps were most "surprising" to Darwin.

E.3 Figures

All Jupyter[15] Notebooks used to render these graphs are available at http://darwinsmind.org/reading/arXiv/v1/figures.zip.

[15] née IPython