Giorcelli, Michela; Lacetera, Nicola; Marinoni, Astrid

Working Paper

Does Scientific Progress Affect Culture? A Digital Text Analysis

CESifo Working Paper, No. 7499

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Giorcelli, Michela; Lacetera, Nicola; Marinoni, Astrid (2019) : Does Scientific Progress Affect Culture? A Digital Text Analysis, CESifo Working Paper, No. 7499, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/198859

CESifo WORKING PAPERS

7499
2019

February 2019

Does Scientific Progress Affect Culture? A Digital Text Analysis

Michela Giorcelli, Nicola Lacetera, Astrid Marinoni

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7499*
Category 13: Behavioural Economics

Does Scientific Progress Affect Culture? A Digital Text Analysis

Abstract

We study the interplay between scientific progress and culture through text analysis on a corpus of about eight million books, with the use of techniques and algorithms from machine learning. We focus on a specific scientific breakthrough, the theory of evolution through natural selection by Charles Darwin, and examine the diffusion of certain key concepts that characterized this theory in the broader cultural discourse and social imaginary. We find that some concepts in Darwin's theory, such as Evolution, Survival, Natural Selection and Competition diffused in the cultural discourse immediately after the publication of *On the Origins of Species.* Other concepts such as Selection and Adaptation were already present in the cultural dialogue. Moreover, we document semantic changes for most of these concepts over time. Our findings thus show a complex relation between two key factors of long-term economic growth – science and culture. Considering the evolution of these two factors jointly can offer new insights to the study of the determinants of economic development, and machine learning is a promising tool to explore these relationships.

JEL-Codes: C190, C890, N000, O000, O390, Z190.

Keywords: science, culture, economic history, text analysis, machine learning.

Michela Giorcelli
University of California
Department of Economics
9262 Bunche Hall, 315 Portola Plaza
USA – Los Angeles, CA 90095
mgiorcelli@econ.ucla.edu

Nicola Lacetera
University of Toronto
Institute for Management and Innovation
3359 Mississauga Road North
Canada – Mississauga, ON
nicola.lacetera@utoronto.ca

Astrid Marinoni
University of Toronto
105 St. George Street
Canada – Toronto, ON
astrid.marinoni@rotman.utoronto.ca

We gratefully acknowledge the financial support of the National Bureau of Economic Research through the Innovation Policy Grant Program. We also thank Ryan Heuser, Graeme Hirst, Xander Manshel, Yang Xu and participants to our presentations at the NBER Productivity Lunch, the REER Conference at Georgia Tech, the Workshop in Memory of Luigi Orsenigo at Bocconi University, and the 2018 Academy of Management Annual Meetings for their helpful feedback.

1. Introduction

> *"It is doubtful if any single book, except the "Principia," ever worked so great and so rapid a revolution in science, or made so deep an impression on the general mind."*
> Obituary for Charles Darwin, Proceedings of the Royal Society of London, 1888.

There is widespread consensus that scientific progress is a major driver of economic growth through its impact on the stock of knowledge and ultimately technological change (Bush 1945, Romer 1990, Stephan 2012). A parallel literature points at other, perhaps less tangible, determinants of economic development. These factors often go under the term "culture", or the beliefs that a population shares and that coordinate activities and transactions. These beliefs lead to cohesion, trust and acceptance of a particular social, economic and political order (Alesina and Giuliano 2015, Bisin and Verdier 2011, Gramsci 1948, Greif 1994, Guiso et al. 2006, Harrison 2002, Landes 2000, Mokyr 2016, Sen 2004). The prevalence of certain beliefs, such as the role of personal responsibility, the value of hard work, and a positive view of progress may have been instrumental to some major economic transformations, such as the Industrial Revolution (Mokyr 2016). Even in current times, we observe how intense public debates characterize scientific and technological development; examples include the development and use of genetically modified organisms, vaccination, and the ethical concerns about the safeguard of privacy as threatened by the development of information and communication technologies.

Despite the qualitative work by historians and humanities scholars, quantitative evidence on the relationship between scientific progress and cultural change is scant. However, quantifying this phenomenon is important to understand the extent to which the interplay between scientific progress and cultural change affects economic development and growth.

This paper proposes the first large-sample empirical study of the relationship between scientific progress and the broader cultural environment in which it occurs. Our underlying hypothesis is that the impact of scientific discoveries on growth does not depend only on the recognition of these advances by the scientific community, but also on their broader public perception, understanding and acceptance.

Performing such an analysis presents several challenges. First, it is difficult to measure social perceptions of science and technology. Second, one would need a long time horizon to analyze the interplay between public discourse and scientific and technological progress. Third, the plausible two-way relationship between science and culture makes it difficult to identify causal

links; if, on the one hand, scientific progress can spur the diffusion and acceptance of certain ideas, on the other hand the presence and diffusion of some ideas can facilitate certain scientific discoveries.

We focus on one major scientific breakthrough: the theory of evolution through natural selection by Charles Darwin. Specifically, we examine whether the key concepts in Darwin's work emerged and diffused in the broader public discourse and social imaginary as a result of the publication of Darwin's theory.

Our methodology takes advantage of the development of machine learning techniques to perform digital text analysis on a corpus of about eight million fiction and non-fiction literary works, between 1820 and 1899. We first measure the cultural discourse by computing the relative frequency of use of certain Darwinian key words and phrases (number of occurrences in a corpus of text every 1,000,000 words). Second, we study the semantic evolution of these words over time.

In Section 2, we outline the historical background of the development of Darwin's theory of evolution, and of the publication of *On the Origin of Species* in 1859. We claim that Darwin did not fully plan the publication date of his treatise; he had to accelerate the publication, and the public reach of his work, to keep scientific priority over it. This specific context and the ensuing natural variation provides us with empirical features that are difficult to find in other cases.

In Section 3, we describe the data and the techniques that we use. We rely on the Google Books corpus, a digitized collection of about eight million books. We use the publication year of *On the Origin of Species* as our reference date, and we focus our analysis on the four decades before and after it. The first goal of our work is to investigate whether the frequency of use of certain words and phrases changed significantly in the years following publication. We mostly focus on a few words and expressions that, according to many accounts, represent the key concepts in Darwin's theory (see, for example, Desmond and Moore 1994, and Mayr 1995): Evolution, Survival, Competition, (Natural) Selection, Survival and Adaptation. These frequencies provide a measure of the adoption and relevance of certain concepts in the public discourse. We compare, both descriptively and in a differences-in-differences econometric framework, the evolution of the frequency of use of Darwinian concepts with the frequency of scientific concepts not related directly to Darwin's theory but extensively used by Darwin in *On the Origins of Species*. We then move to analyzing the semantic evolution of these words. To do

so we apply word-embedding techniques, which are widely used in the Natural Language Processing and Machine Learning literature.

We report our findings in Section 4. We document that some key concepts in Darwin's theory became relevant in the broader cultural discourse in the years immediately following the publication of *On the Origins of Species*: Evolution, Survival and, to a lesser extent, Competition. The patterns of diffusion of these words were similar in the non-fiction and in the fiction literature; this indicates a broad impact on culture as well as the social imaginary as represented, for example, by short stories and novels. Other key concepts such as Selection and Adaptation were already present in the cultural discourse. Although the relative frequency of the term Selection per se did not vary around the publication of *On the Origins of Species*, the expression Natural Selection was virtually nonexistent in the fiction and non-fiction literature before 1859 and diffused rapidly thereafter; this suggests a potential change in the way in which the term Selection was used.

The objective of the second part of our analysis is indeed to explore these potential semantic changes. We first focus on a few interesting pairs of concepts to explore if there was a change in their semantic association. Of particular interest is the increase in semantic association between words such as Competition (as well as Struggle) and Life, as well as between Life and Adaptation, again immediately following the publication of *On the Origins of Species*. This is consistent with Darwin's theories affecting the perception of what existence means and how it unfolds. Second, we "let the data speak" by determining, for each decade in our period of interest, the words with the highest semantic connection to the key words of our interest. Among the most interesting semantic changes, we find, for example, that the term Adaptation became, over the 19th Century, less related to physical terms (such as Mechanism) and increasingly related to concept related to living beings (such as Organism and Reproduction). The term Evolution, which came mostly from chemistry and physics in the first half of the 1800s, later in the century related more to concepts from biology as well as social and human subjects. Finally, Selection became more similar in meaning to other specific "Darwinian" words, such as Survival, Variation, Fittest and Heredity.

Our findings thus show a complex relationship between two key factors of long-term economic growth. Consideration of the joint evolution of science and culture can therefore offer new insights to the study of the determinants of economic development. Empirical approaches

enabled by machine learning techniques provide promising tools to explore these relationships; they allow, on the one hand, for the analysis of a vast amount to textual data, and on the other hand, for a detailed analysis of specific ideas as embedded in key words and phrases.

In Section 5 we further discuss our findings and outline directions for future research.

Related literature and contributions. The stream of literature that is closest to our work is the study of how different cultures are more or less open to scientific and technological change, and how certain scientists may introduce new sets of beliefs in a population. Mokyr (2013, 2016), in particular, defines "cultural entrepreneurs" those scientists who put in motion broader cultural changes. Our paper provides an empirical approach to study this form of cultural entrepreneurship.

We also contribute to the growing use of "text as data", which is developing especially in such fields as finance, marketing and the study of media (Gentzkow et al. 2018). Economists of science, productivity and innovation have also recently begun to rely on these sources of information and related techniques (Balsmeier et al. 2018; Bandiera et al. 2017; Catalini et al. 2015; Kelly et al. 2017). Scholars in linguistics and literary criticism are also increasingly employing computerized text analysis to answer questions about the evolution of literary genres and styles, semantic changes of words and concepts, and the influence of certain issues in the public discourse. These scholars are moving from the direct reading of an inevitably limited set of text from which to offer general insights and interpretations, to the automated or "distant" reading of a much larger set of digitized texts (Heuser and Le-Khac 2011; Heuser 2016; Moretti 2013; Wilkens, 2015). In addition to literary analysis, an area of study known as "cultural analytics" or "culturomics" explores the evolution of culture through text analysis (Manovich 2009; Michel et al. 2011). To our knowledge, there are no applications of these approaches and analyses to studying the public understanding of science and technology.[3]

Finally, our work also relates to the literature on the role of institutions in the diffusion of ideas and innovation (Abramitzky and Sin 2014). Our paper looks at the impact of scientific advancements on the perception of key ideas and concepts in society, and on how these ideas and concepts were already permeating the public discourse.

2. Historical Background and Identification

[3] Current studies on the public understanding of science (Bauer 2009) do not rely on text analysis techniques.

Charles Darwin's interest in the evolution of living organisms largely developed during his voyage on the HMS Beagle (1831-36), a ship of the Royal Navy on its second survey expedition. Over those five years, Darwin collected fossils from the places that he visited and observed their geographical distribution. These investigations led him to elaborate a theory of evolution as early as 1838. Darwin was particularly interested in the geographical distribution of wildlife and fossils that he collected in the voyage. Although his early elaborations built on previous theories (such as Lamarck's) and considered the possibility of the transformation of one species into another (transmutation), he then developed more explicitly his theory of evolution based on the natural selection of the most adaptive (innate) characteristics of a species. Small, gradual variations within a species would emerge randomly, and would eventually lead to branching of new species. Competition for resources and adaptive capacities would determine whether and where a particular species would be more likely to thrive. The developments in genetic research since the mid-20th Century provided corroboration and foundations to Darwin's evolutionary theory.

In addition to being one of the greatest scientific breakthroughs in history, there is broad agreement that the reach of Darwin's theory of evolution had a wider, cultural reach (Desmond and Moore 1994). Mayr (1995) provides a summary of this larger impact. In particular, the ideas of competition for resources, common origins of species, and random variation implied the absence of a teleology or (benevolent) design, that is, a very different conception of Nature and God. The likely common origins of all species also eliminated any special status of humans as compared to other living beings. Mokyr (2013, 2016) includes Darwin among a small set of "cultural entrepreneurs", i.e. scientists whose discoveries questioned deeply held cultural and popular beliefs. Research in literary criticism has analyzed how the production of certain poets and novelists, such as George Eliot and Thomas Hardy, began to reflect ideas of a different role that nature had in its relationship with humans and the environment, of the absence of an "intelligent design", of the presence and importance of innate traits in individuals. Similarly, studies of the literary production prior to the publication of *On the Origin of Species* point out how some of Darwin's ideas connected to images already developed by these writers. A frequently cited example is the work of Alfred Tennyson, and in particular his poem *In Memoriam*, published in 1850; scholars also investigated the connections between broader worldviews, such as Enlightenment and Romanticism, on Darwin's ideas (Cartwright and Baker

2005; Chapple 1986; Gianquitto and Fisher 2014; Lansley 2016; Otis 2009; Richards 2013; Scholnick 2015).

These accounts, however, focus on a narrow set of literary contributions, thus making it hard to advance inferences about the broader cultural impact of this scientific advance, and about the cultural climate that preceded that advance. Our approach to answering these questions relies on a massive corpus of literary work (fiction and non-fiction), and therefore offers a methodological contribution that allows going beyond the analysis of a small set of texts and authors as a way to extrapolate general cultural views and trajectories.

Some features of how Darwin made his work public and of how his theories diffused crucially enhance our ability to identify the impact of Darwin's work. Although Darwin developed his theory over a long period, there is a precise time at which one could identify Darwin's theory of evolution as reaching a broader public, and this is 1859, the year of publication of *On the Origin of Species*.[4] This publication date was largely unplanned. Darwin proceeded slowly initially and had to deal with sickness and deaths in his family that delayed him. However, eventually he had to "rush" in order not to lose priority over Alfred Russel Wallace, who was researching on the same topics and had sent Darwin on some of his writings that used similar concepts and reached similar conclusions about natural selection as the theory that Darwin was elaborating.[5]

The book and Darwin's overall theory received almost immediate attention and fast diffusion, also thanks to presentations at prestigious scientific meetings such as the Linnaean Society (of a joint paper with Wallace in 1858) and the British Association for the Advancement of Science (in 1860) as well as reviews in the popular press. Note, finally that a few historians attribute this long preparation period also to Darwin's concern about society not being "ready" to receive his radical ideas that would contrast prevailing views about the origins of life from a divine or intelligent design (Gopnik 2010).

[4] The year 1859 saw also the publication of other important works, John Stuart Mill's *On Liberty*, Tennyson's *Idylls of the King*, Eliot's *Adam Bede* and Dicken's *A Tale of Two Cities*. These publications make it harder to identify a connection between the publication of *The Origins of Species* and changes in the public discourse. However, in our study, we focus on rather specific concepts that are central in Darwin's work but not in the other works mentioned above; we also consider the role of those concepts in the public discourse *before* 1859.

[5] See in particular Desmond and Moore (1994) for details on the personal and intellectual biography of Darwin.

The unplanned publication date and the rapid diffusion that Darwin's theory had also with the broader public provide the main sources of natural variation that we employ in our analysis below.

3. Data and Methods

To examine the diffusion and semantic evolution of scientific concepts over time, we exploit the increasing availability of digitized historical text corpora, as well as new tools of natural language analysis. Our first step is to compute relative frequencies of some key words that embody the main concepts advanced in Darwin's theory of evolution, and that in fact Darwin used extensively in his own work. These frequencies represent a basic measure of the adoption of certain ideas in the broader cultural and social discourse. The second step of our investigation focuses on word embeddings, which are widely used in the Natural Language Processing and Machine Learning literature as an effective tool for the analysis of semantic change.

Word Frequencies. We rely on Google N-Grams[6] (Lin et al. 2012) to assess how frequencies of words have changed over time. The Google N-Grams data is the result of the Google Book digitization project whose aim is to build a vast collection of digitized books in partnership with major research libraries[7]. First released in 2010, the data consists of a set of corpora of roughly eight million books, an estimated 6% of all books ever published (Lin et al. 2012). The texts cover roughly a 500-year span and they are continuously updated. The Google Books data includes different corpora and different languages (besides English: Italian, French, German, Spanish, Russian, Hebrew, and Chinese). The English corpus alone has half a trillion words in it. This currently represents the largest available collection of digitized books in existence. The data focuses on of both fiction and non-fiction work, but not periodicals. The data is aggregated depending on the number of terms considered; for instance, the 1-ngram dataset includes single, unique words and their frequency in a given corpus, and 2-grams and higher take into account combinations of multiple words and their frequency.

We compute frequencies from 1-ngram and 2-ngrams data for each year and express them in per-million-words terms. Although this is a crude measure of the diffusion of words and the

[6] Available at: http://books.google.com/ngrams.
[7] http://books.google.com/googlebooks/library/partners.html

underlying ideas, the literature in digital humanities and computational linguistics widely employs this approach as a tool for measuring cultural trends (Michel et al.2011; Roth 2014).

The ability to separate fiction and non-fiction literature is particularly relevant for addressing our questions of interest. First, one critique to the N-gram (and Google Books) corpus is that it may over-represent scientific texts (Pechenick et al. 2015). In our study, for example, an uptake in the (absolute and relative) frequency of words related to Darwin's theory may reflect just a disproportionate increase over time of the corpus of scientific books (included in the non-fiction corpus). Second, separating fiction and non-fiction literature enables the analysis of different types of relationships between Darwinian science and broader culture. The use of Darwinian concepts in the non-fictional literature may better represent higher-educated or more erudite cultural conversations. Conversely, and given the diffusion of the novel and the relatively high literacy rates especially in England and the United States in the 19th Century, fictional literature may better measure the prevailing topics in the broader social imaginary (Armstrong 1987, Winans 1975).

Semantic evolution and word embeddings. The analysis of word frequencies is informative, but it does not provide insights about the evolution and, in particular, the associations between words over time. To this aim, we employ a distributional natural-language processing technique, known as word embeddings, which is able to capture semantic and contextual change of words in a given period. This technique represents words as embedded in high-dimensional vector spaces according to their co-occurrence statistics.

Consider a vocabulary with V distinct words in it. One way to represent a word w in a V-dimensional vector space is $w_{1xV} = (0,0,\ldots,0,1,0\ldots,0)$, where all values are equal to zero except the entry that corresponds to the word of interest (suppose the words are in alphabetical order). This characterization is simple, but it does not allow to compare words, or to study the evolution of the meaning of a word over time. For example, any pair of word vectors would be orthogonal by construction. An alternative approach to represent words as vectors is to define them in relation to the other words with which they occur in texts, within a certain "window" or number of words before and after the focal one. We employ machine-learning algorithms, and in particular a Word2Vec (SkipGram with negative sampling) approach (Mikolov et al. 2013), to

predict the surrounding words given a target word. Specifically, the model computes estimates of parameters θ that solve:

$$\arg\max_{\theta} \prod_{w \in T} \prod_{c \in C(w)} p(c|w;\theta), \quad (1)$$

where *w* is a focal word in *T*, the text of the corpus. The term *c* represents a context word included in *C(w)*, the set of possible context words of *w*. These can be defined as all the words that appear within a context window *m*. Smaller windows around a focal word tend to capture functionally similar words (e.g., 'respect' and 'deference'), whereas larger windows capture context relatedness or topic similarity (e.g., 'respect' and 'love') (Levy and Goldberg, 2014). *c* can also be expressed in terms of the focal word as: $c = w_{t+j}$ where *t* is a given position in the corpus and *m* is the number of context words to consider before and after *w*; therefore, $-m \leq j \leq m$ and $j \neq 0$. The parameters of the models need to be set such that the probability of context words appearing near the target words are as high as possible. After expressing equation (1) as a negative log-likelihood, we parametrize the model following the neural-network literature, using a soft-max function:

$$p(c|w;\theta) = \frac{e^{v_c v_w}}{\sum_{c' \in C} e^{v'_c v_w}}, \quad (2)$$

where v_c and v_w are vector representation of *c* and *w* respectively. *C* is the set of all possible contexts. The derivation of the final word embeddings is the result of a training process in which randomly initialized vectors are "pulled closer or apart" depending on the actual word co-occurrence. This will result in word embeddings where similar words will have similar vectors. Moreover, the final vectors satisfy some "linearity" features in the relationship between, for example, the singular and plural form of a word, or feminine and masculine versions. Using a frequent example in the technical literature, we expect, when the words *king*, *kings*, *queen*, *queens*, *man* and *woman* are in distributed vector form, that the following holds: *(king – kings)* ≈ *(queen – queens)* and *(king – man)* ≈ *(queen – woman)*.

Because the geometry of the vectors captures semantic relations between words, one can compare vectors among each other to find a word's nearest neighbors and across periods to examine the degree of change a word underwent from one-time interval to the next.[8] The main

[8] Embeddings can measure close semantic relationships between words as well as more global ones. For instance, beyond successfully measuring shifts in word meanings over time (Hamilton et al. 2016), embedded vectors have also been used to track demographic and occupational social shifts (e.g., Garg et al. 2017) and gender stereotypes (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017).

metric to compare the word's vector representations between two points is the cosine distance (Dubossarsky et al. 2015; Gulordava and Baroni 2011; Jatowt and Duh 2014; Kim et al. 2014; Kulkarni et al. 2015). Call γ the angle (generalized to *N*-dimensions) between two *N*-dimensional vectors $u = (u_1, ... u_N)$ and $v = (v_1, ... v_N)$. Then, $u'v = \sqrt{\sum_{i=1}^{N} u_i^2} * \sqrt{\sum_{i=1}^{N} v_i^2} * \cos(\gamma) = \|u\|\|v\| \cos(\gamma)$, or: $\cos(\gamma) = \frac{u'v}{\|u\|\|v\|} \in [-1,1]$. The more similar the two vectors, the closer to one the cosine. We use previously trained Word2Vec embeddings resulting from the n-grams distributed by Google Books (Hamilton et al., 2016). Figures are available for every decade between 1800 and 1990 and data are specifically designed to enable comparisons across decades. The training parameters of the embeddings employed in the analysis allow a sufficiently large window so that we can explore topic similarity. A simplification of how the resultant word vectors might appear is offered in Figure SM1, where the vector for the word 'Darwin' is projected on a three-dimensional space with its closest 20 vectors. In order to obtain a feasible visualization of this word embeddings, we reduced the dimensionality by using Principal Component Analysis.

4. Findings

We first describe the evolution of the relative frequency of certain selected words and two-word expressions as measures of the diffusion of certain concepts in the public discourse around the time of the publication of *On the Origin of Species* in 1859. The second part of the analysis will focus on semantic evolution.

4.1 Word Frequency and Diffusion of Concepts

4.1.1 Key Concepts

Graphical Analysis. We consider terms that, from many accounts (as well as our own reading), represent the key concepts in Darwin's theory: Evolution, Selection, Adaptation, Competition, Survival, Mutation and the 2-gram Natural Selection. Figure 1 reports their frequency of use, per million words, in each year between 1820 and 1899, separately in fiction and non-fiction books (we excluded the word Mutation from the graphs because its occurrence was too low throughout the period of interest to allow for meaningful analyses). We scale the y-axes differently for the two groups in each graph. Evolution and Survival entered the public discourse in the years

immediately following the publication of *On the Origin of Species*. Similarly, the expression Natural Selection was virtually non-existent in both fiction and non-fiction literature before 1859 and experienced a significant increase in use since then. The concepts that underlie these words and expressions, therefore, generated interest in not only specialized or more educated circles, but plausibly also in the more popular cultural context. Competition was already present in the first part of the 19th Century, but especially in the non-fiction literature, experiences and uptick in frequency after about 1860. Selection and Adaptation, in contrast, did not see a further increase in relative frequency around the publication of *On the Origins of Species*; the relative frequency of Selection was constantly increasing since the early 19th Century, whereas Adaptation reached a stable relative frequency in the 1840s.

In Figure 2, we display terms of frequent use in general and in the sciences, not specific to Darwin's theory, which we found with high occurrence in *On the Origins of Species*. In looking at these terms, our objective is to assess whether there were general trends in the use or diffusion of scientific concepts. The words that we consider are Number, Life, Animals, Flowers, Plants and Nature. For none of these words was there any discernible change in diffusion in the decades immediately preceding and following the publication of *On the Origin of Species*.[9,10]

Regression Results: Word-by-Word Time Series. Table 1 reports estimates from regressions of the yearly relative frequency of use of each of the Darwinian words and phrases that we represented in the graphs above, and Table 2 reports the estimates for the general scientific terms. We first regress the annual frequency y_{tw} for each word *w* on just a linear time (year) trend *t*:

$$y_{wt} = \alpha_w + \beta_w t + \varepsilon_{wt} \qquad (3)$$

We consider also an alternative specification in which we compare the average relative frequency of each word before and after 1859:

$$y_{wt} = \alpha_w + \gamma_w \mathbf{1}(t > 1859) + \varepsilon_{wt} \qquad (4)$$

[9] Interestingly, the data show a secular decline in the relative use of Nature in the literature, and an uptake of Number at the end of the 19th Century. The decline in the relative frequency of Nature may depend on the evolution of literature away from Romanticism, a literary movement and style that exalted nature.

[10] In material available upon request, we generated similar graphs with additional science-related words, including some that pertains to other established theories (e.g. Gravity); again, we find no significant changes in the trend of their diffusion over the 19th Century.

We then enhance the model to allow for a structural break (a change in the slope of the linear trend) to occur after 1859:

$$y_{wt} = \alpha_w + \beta_w t + \gamma_w (t - 1859) * \mathbf{1}(t > 1859) + \varepsilon_{wt} \quad (5)$$

Table 3 reports regression estimates Model 5 above, limited to the Darwinian concepts, separately for fiction and non-fiction books, and also estimates of the following interaction model:

$$y_{wt} = \alpha_w + \beta_w t + \gamma_w (t - 1859) * \mathbf{1}(t > 1859) + \delta_w \mathbf{1}(fiction)$$
$$+\theta_w t * \mathbf{1}(fiction) + \lambda_w (t - 1859) * \mathbf{1}(t > 1859) * \mathbf{1}(fiction) + \varepsilon_{wt} \quad (6)$$

For the "Darwinian" concepts (Table 1), the estimated annual increase in frequency is significantly higher after 1859 for most words with the exception of Adaptation that saw its frequency of use stabilizing after the publication of *On the Origins of Species*. Tests for structural breaks in the time trend estimate the breaks to be in the neighborhood of 1859 for Evolution (1864), Survival (1864) and Natural Selection (1859), to have occurred significantly before 1860 for Selection (1840) and Adaptation (1833), and after for Competition (1978). We can statistically reject the null hypothesis that no break occurred in 1859 for any of the terms, however. Regarding the more generic scientific words that we considered, the estimates in Table 2 do not show any systematic patterns in their frequencies around 1859. Estimates of a single structural break are close to 1859 only for Number (1860). When we test directly for a break in 1859, the estimated Wald χ^2 statistics are much smaller than for the Darwinian concepts.

Table 3 reports the regression estimates separately for fiction and non-fiction books (limited to the Darwinian concepts), following Models 5 and 6 above. In the analyses based on Model 6, the data for each single word reports two observations per year, one with the relative frequency of the word with respect to non-fiction books, and one with the relative frequency in fiction texts. The regression estimates confirm the similar diffusion patterns in fiction and non-fiction books especially for Evolution, Survival and Natural Selection.

Regression Results: Differences in Differences. After having studied the diffusion over time of each word separately, we proceed with some differences-in-differences analyses where we estimate the aggregate diffusion patterns of Darwinian and generic scientific concepts before and after 1859. We perform these analyses in two ways.

First, in Table 4 we report the estimates from regressions where, for each year, we sum the frequency of occurrence of the six Darwinian concepts on the one hand, and of the six generic scientific words on the other hand, and compare the trend in the aggregate diffusion before and after 1859. Because the aggregate frequency of the generic words is much higher than the frequency of the Darwinian concepts pooled together, in order to make more immediate comparisons we transform these frequencies into their natural logarithms and include the logarithm of the time trend in the regression analyses. Therefore, we compare scale-free elasticities.[11] In this analysis, we also pool together fiction and non-fiction books. The full regression model that we estimate is as follows:

$$\ln(y_{wt}) = \alpha_w + \beta_w \ln(t) + \gamma_w(\ln(\mathrm{t}) - \ln(59)) * \mathbf{1}(t > 59) + \delta_w \mathbf{1}(Darwinian\ word) + \theta_w \ln(t) * \mathbf{1}(Darwinian\ word) + \lambda_w(\ln(\mathrm{t}) - \ln(59)) * \mathbf{1}(t > 59) + \mu_w(\ln(\mathrm{t}) - \ln(1859)) * \mathbf{1}(Darwinian\ word) * \mathbf{1}(t > 59) + \varepsilon_{wt} \quad (7)$$

Estimates of the parameters of Model 7 are in Column 3 of Table 4. Columns 1 and 2 report estimates of a simplified version of the model, were the left hand side variable is the difference in the natural logarithm of the frequencies of Darwinian and generic terms, regressed on either a simple time trend, or a time trend and the interaction of an indicator for years greater than 1859 and the difference between the current year and 1859 (similar to Model 5 above). A Wald test for a single structural break in the constant and time trend identifies 1855 as the most likely break, and the value of the χ^2 statistics is very similar when we test for a break in 1859. The estimate on the coefficient on the interaction between and indicator for the post-1859 period and the difference between the current year and 1859 is positive, large and statistically significant, indicating a much larger relative uptick in the frequency of Darwinian concepts in fiction and non-fiction books after 1859. The estimates of the full model in column 3 confirm these findings; the R-squared for this model is 99.8%. Figure 3 provides a graphical representation of the results.[12]

[11] We also estimated Models 3, 4, and 5 with the variables expressed as natural logarithms. The results provide the same insights as the analyses in levels.

[12] We also explored the evolution of the concept of gradualism as applied to the type of change and evolution that Darwin considered. The idea of small, continual changes at the basis of the evolution for species and more generally of biology is key in Darwin's work; it is also a philosophical contribution or worldview. We also considered expressions such as Gradual Change, Gradual Adaption, Gradual Divergence and Gradual Mutation. The very low

Second, we estimate a model where we consider the average frequency of the six Darwinian concepts and the six generic scientific words per each decade from 1820-29 to 1890-99:

$$\ln(y_{wt}) = \alpha_w + \beta_w \mathbf{1}(Darwinian\ word) + \sum_{i=3}^{5} \gamma_i \mathbf{1}(i^{th}\ decade\ of\ 1800) + \sum_{i=7}^{10} \gamma_i \mathbf{1}(i^{th}\ decade\ of\ 1800) + \sum_{i=3}^{5} \delta_i \mathbf{1}(i^{th}\ decade\ of\ 1800) * \mathbf{1}(Darwinian\ word) + \sum_{i=7}^{10} \delta_i \mathbf{1}(i^{th}\ decade\ of\ 1800) * \mathbf{1}(Darwinian\ word) + \varepsilon_{wt} \qquad (8)$$

Figure 4 displays the estimates of the δ_i's (and 95% confidence intervals). The omitted time category is the 6th decade of 1800, i.e. 1850-59.

relative frequencies of these expressions in our corpus, however, do not allow making any clear inference. Data are available upon request. There was no strong trend throughout the 19th Century or, especially for the most frequent of the di-grams (Gradual Change), nor any specific change in adoption rates around 1860. This lack of a clear effect may be consistent with the idea that this concept was already part of a "Victorian" view of society and this contributed to the acceptance of several aspects of Darwin's theories. But, again, given the very low overall frequencies, we need to be cautious in drawing conclusions.

4.1.2 Lamarck and Darwin; Transmutation and Evolution

If a word frequency analysis is a valid way to measure the diffusion and acceptance of an underlying scientific theory in the broader cultural discourse, then this analysis should also be able to identify the decline of certain theories. A natural counterfactual for Darwin's elaboration is Lamarck's theory of the transmission of acquired traits. We plot the relative frequency of the use of the words "Darwin" and Lamarck" in books. Because Lamarck was French, we do the same exercise also on the corpus of French books. For English texts, we also isolate the frequency in the fiction literature; this is not possible for French texts in Google books. Figure 5 reports the frequency graphs. The frequency of the word Darwin became increasingly greater than the frequency of Lamarck both in the English and French literature; in the latter case, the frequency of Darwin surpassed that of Lamarck soon after 1859. Darwin seems to have had a larger presence in English fiction than Lamarck, too. We also compare in Figure 4 two terms that related to the study of the emergence and development of species: Evolution and Transmutation. Although Evolution, which we already analyzed above, is typically associated with Darwin's work, earlier works in biology (including some of Darwin's) used the term Transmutation to characterize (gradual or discrete) transformations of plants and animals. By comparing these two words, we want to assess whether the broader literature and cultural discourse also picked up the "newer" word to express these changes. For books in French, we consider the word Transformism (Transformisme in French), which was used, for example, by Lamarck. The general pattern is that Evolution became progressively more frequent than Transmutation, with a significant change in frequency after 1859.

4.2 Semantic Changes

Looking at frequency of use as a measure of the interplay between a major scientific discovery and the broader cultural climate is a natural first step in the analysis. However, the role of a particular construct does not only depend on how often that construct occurs. Words can change their meaning over time. These changes, even keeping frequency constant, provide further evidence of cultural evolution potentially linked to certain scientific events.

Figure 6 introduces the second part of our study, where we move from the analysis of the frequency of use of certain words, expressions and the concepts underlying them, to the analysis

of whether the semantic evolution of certain words and concepts, to see whether this evolution occurred in ways that we can relate to the elaboration of Darwin's theory.

One aspect of Darwin's theory, for example, is that life (or existence) includes adaptation, as well as competition, among its defining aspects. We do see an increase in the semantic association between Life on the one hand, and Adaptation, Struggle and Competition on the other hand, especially after 1859. For Life and Struggle we see a trend since the early 19th Century. Several of the studies mentioned above that relate Darwin's work to the Romantic literary climate of the first half of the 19th Century, with a more tumultuous view of nature in particular, seem therefore to have captured a more general trend. Meaningful cosine similarities between survival and competition started in the 1860s and increased since then.

Finally, a controversial implication of Darwin's theory is that evolution applies to humans in the same way as it applies to other animals; although Darwin did not explicitly treat the human species in his 1859 book, this was the topic of his 1871 *The Descent of Man and Selection in Relation to Sex*. The semantic evolution of the word Human shows an increase in its similarity with Animal especially in the late 1800s.

A second analysis of semantic changes focuses on some of the key words and concepts that we considered so far. Instead of investigating the similarity of these words with a select sample of other concepts, we "let the data speak" by determining, for each decade, the words with the highest semantic connection to these key words, again in terms of cosine similarity among word vectors. Figures 7 through 12 report the results for the words Adaptation, Competition, Evolution, Nature, Selection and Survival (in this order). We excluded from the rankings of semantic similarity the words that had the same root as the focal key word as well as the most obvious synonyms (e.g. Compete or Competitor for Competition); we also defined a lower bound to the relevant cosine similarity to be equal to 0.05. The closer a word is to the time axis in the figures, the closer to one the cosine similarity. Finally, we use a "color system" to classify words according to some broad category; in addition to being interest in changes in the type of most similar words, we also want to assess whether, for example, concepts more distant from Darwin disciplines related to Darwin's major concepts and whether these similarities evolved over time.

The figures identify a few interesting facts. First, the term Adaptation became, over the 19th Century, less related to physical or "mechanical" terms (such as Mechanism) and increasingly similar to concepts that represented living beings (such as Organism, Reproduction).

Second, perhaps the biggest changes in meaning and association concern the word Evolution. In the first half of the 19th Century, the terms that were closest to Evolution came mostly from chemistry and physics. Later in the 1800s concepts from biology as well as related to human society were semantically more similar to Evolution. Examples include Social and Progress. Note also how the word Darwinian itself became closely associated with Evolution; this is consistent with a direct role of Darwin's theory in changing the meaning of this concept over time.

Third, Selection appeared more closely related to the concept of Choice (and qualification for the choice such as "careful" or judicious") in the first half of 1800; the similarity in meaning with Choice remained also later, but in the overall literature, Selection became more similar in meaning to other specific "Darwinian" words, such as Survival, Variation, Fittest and Heredity.

Fourth, very few words had a similarity in meaning with Survival, likely because the word itself was only rarely used in the first half of the 19th Century. Later in the century, the word was increasingly associated in the overall literature to other concepts related to evolutionary theory, notably Fittest, Evolution, Struggle and Selection. The increasing relatedness with Fittest toward the end of the 1880s is likely due also to the publication of the *Principles of Biology* by Herbert Spencer in 1864, where this concept applies also to society and ethics and not only to the natural sphere. Competition, in contrast, maintained an association with a stable set of words, mostly related to production and markets, throughout the century.

Finally, Nature is perhaps too generic (and more widely used) of a term to expect a close relation with specific concepts. Interestingly, however, words such as Divine and Perfection disappear from the concepts most closely related to Nature in the second half of 1880.

5. Conclusions

Do scientific discoveries influence broader cultural and social domains? Because both scientific progress and culture are long-term factors of growth, analyzing the interplay between these two spheres is a question of economic relevance. Historians and literary critics have long claimed the existence of such an interplay. It is challenging, however, to assess the relevance of these

phenomena quantitatively on a large scale. The analysis of large corpora of text through machine learning techniques promises to provide information about the nature and evolution of cultural beliefs and the public discourse in history. In this paper, we adopted this approach to study how the social and cultural environment of the 19th Century received one of the major scientific breakthroughs in history, the theory of evolution by Charles Darwin.

Our evidence shows that some key concepts in Darwin's theory, as expressed by single specific words or phrases such as Evolution, Survival, and Natural Selection, diffused in the cultural discourse immediately after the publication of *On the Origins of Species*. Other key concepts such as Selection and Adaptation were already present in the cultural discourse as represented by a corpus of about eight million fiction and non-fiction books. The adoption of some of these words and phrases in the broader cultural conversation also led to a change in the meaning of the concepts. Overall, these findings are consistent both with a view that Darwin's theory had broad cultural implications. Our methodology allows us to identify more specifically which specific concepts were new to society and which ones were already part of the public conversation and the social imaginary.

Our approach has several inductive and descriptive aspects. Although the choice of the concepts on which to focus may seem somewhat arbitrary, we based our selection on the main topics that Darwin developed in his treatise, as well as on the analysis of several interpretations of Darwin's theory of evolution. Moreover, it is generally hard to provide causal identification with this type of analysis. However, the unplanned publication date of *On the Origins of Species*, the reliance on very large amount of data, and the consistency in the patterns of different words, phrases and concepts, give us some confidence about the nature of the patterns that we established.

Finally, this is a single case study, and generalizations about the relationship between major scientific discoveries and their cultural and reception are difficult to make. The methodology that we employed in our study is applicable to other relevant cases, where there is a perception that progress in sciences related to broader cultural change but, thus far there has not been large sample empirical evidence to substantiate these claims. Example include the theory of relativity or the indeterminacy principle in physics, the discovery of the DNA and the emergence of biotechnology and genetic engineering. In fact, one could go beyond scientific discoveries and

employ a similar approach to explore the cultural antecedents and effects of new technologies as well as of new industries, such as computers and the Internet (see for example Turner 2010).

References

Abramitzky, R. and Sin, I. (2014). Book translations as idea flows: The effects of the collapse of communism on the diffusion of knowledge. *Journal of the European Economic Association*, 12(6):1453-1520.

Alesina, A., and Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, *53*(4), 898-944.

Armstrong, N. (1987). *Desire and domestic fiction: A political history of the novel*. Oxford University Press.

Balsmeier, B., Li, G.C., Assaf, M., Chesebro, T., Zang, G., Fierro, G., Johnson, K., Lück, S., O'Reagan, D., Yeh, B. and Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economics & Management Strategy*, forthcoming.

Bandiera, O., Hansen, S., Prat, A., & Sadun, R. (2017). CEO Behavior and Firm Performance (No. w23248). National Bureau of Economic Research.

Bauer, M. W. (2009). The evolution of public understanding of science-discourse and comparative evidence. *Science, technology and society*, 14(2):221-240.

Bisin, A., & Verdier, T. (2011). The economics of cultural transmission and socialization. In *Handbook of social economics* (Vol. 1, pp. 339-416). North-Holland.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, pages 4349-4357.

Bush, V. (1945). *Science, the endless frontier: A report to the President*. US Govt. print.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183-186.

Cartwright, J. H. and Baker, B. (2005). Literature and science: Social impact and interaction. *Abc-Clio*.

Catalini, C., Lacetera, N., and Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences*, 112(45):13823-13826.

Chapple, J. (1986). *Science and Literature in the 19th Century*. London: Macmillan.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493-2537.

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4):447-464.

Desmond, A. J., & Moore, J. (1994). *Darwin*. WW Norton & Company.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. *NetWordS*, pages 66-70.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2017). Word embeddings quantify 100 years of gender and ethnic stereotypes. *arXiv preprint* arXiv:1711.08412.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2018). Text as data. *Journal of Economic Literature*, forthcoming.

Gianquitto, T., & Fisher, L. (Eds.). (2014). *America's Darwin: Darwinian Theory and US Literary Culture*. University of Georgia Press.

Gopnik, A. (2010). *Angels and ages: A short book about Darwin, Lincoln, and modern life*. Vintage.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of political economy*, *102*(5), 912-950.

Gramsci, A. (1948). 2003. Selections from the prison notebooks. *The civil society reader. Hanover and London: University Press of New England*.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books ngram corpus. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67-71. Association for Computational Linguistics.

Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic perspectives*, *20*(2), 23-48.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint* arXiv:1605.09096.

Harrison, L. E. (2002). *Culture matters: How values shape human progress*. Basic books.

Heuser, R. (2016). Word vectors in the eighteenth century. *IPAM workshop: Cultural Analytics*.

Heuser, R. and Le-Khac, L. (2011). Learning to read data: Bringing out the humanistic in the digital humanities. *Victorian Studies*, 54(1):79-86.

Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference*, pages 229-238. IEEE.

Kelly, B., P. D. S. A. and Taddy, M. (2017). Measuring technological innovation over the long run. *Working paper*.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint* arXiv:1405.3515.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, pages 625-635. International World Wide Web Conferences Steering Committee.

Landes, D. (2000). Culture makes almost all the difference. *Culture matters: how values shape human progress*, 2-13.

Lansley, C. M. (2016). Charles Darwin-s debt to the Romantics. *PhD thesis, University of Winchester*.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (2):302-308.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google books ngram corpus. Pro*ceedings of the ACL 2012 system demonstrations*, pages 169-174. Association for Computational Linguistics.

Manovich, L. (2009). *Cultural analytics: visualising cultural patterns in the era of -more media*. Domus March.

Marshall, A. (1890). Principles of political economy. Maxmillan, New York.

Mayr, E. (1995). Darwin's impact on modern thought. Proceedings of the American Philosophical Society, 139(4):317-325.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176-182.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 3111-3119.

Mokyr, J. (2010). The Enlightened economy an economic history of Britain. Yale University Press.

Mokyr, J. (2013). Cultural entrepreneurs and the origins of modern economic growth. *Scandinavian Economic History Review*, 61(1), 1-33.

Mokyr, J. (2016). A culture of growth: the origins of the modern economy. Princeton University Press.

Moretti, F. (2013). *Distant reading*. Verso Books.

Otis, L. (2009). Literature and science in the nineteenth century: an anthology. Oxford University Press.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10).

Richards, R. J. (2013). The impact of German romanticism on biology in the nineteenth century. The impact of Idealism: The legacy in philosophy and science, Cambridge University Press.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, *98*(5, Part 2), S71-S102.

Roth, S. (2014). Fashionable functions: A Google ngram view of trends in functional differentiation (1800-2000). *International Journal of Technology and Human Interaction*, 10(2):35-58.

Scholnick, R. (2015). American literature and science. University Press of Kentucky.

Sen, A. (2004). How does culture matter? In Rao, V. (2004). *Culture and public action*. Orient Blackswan.

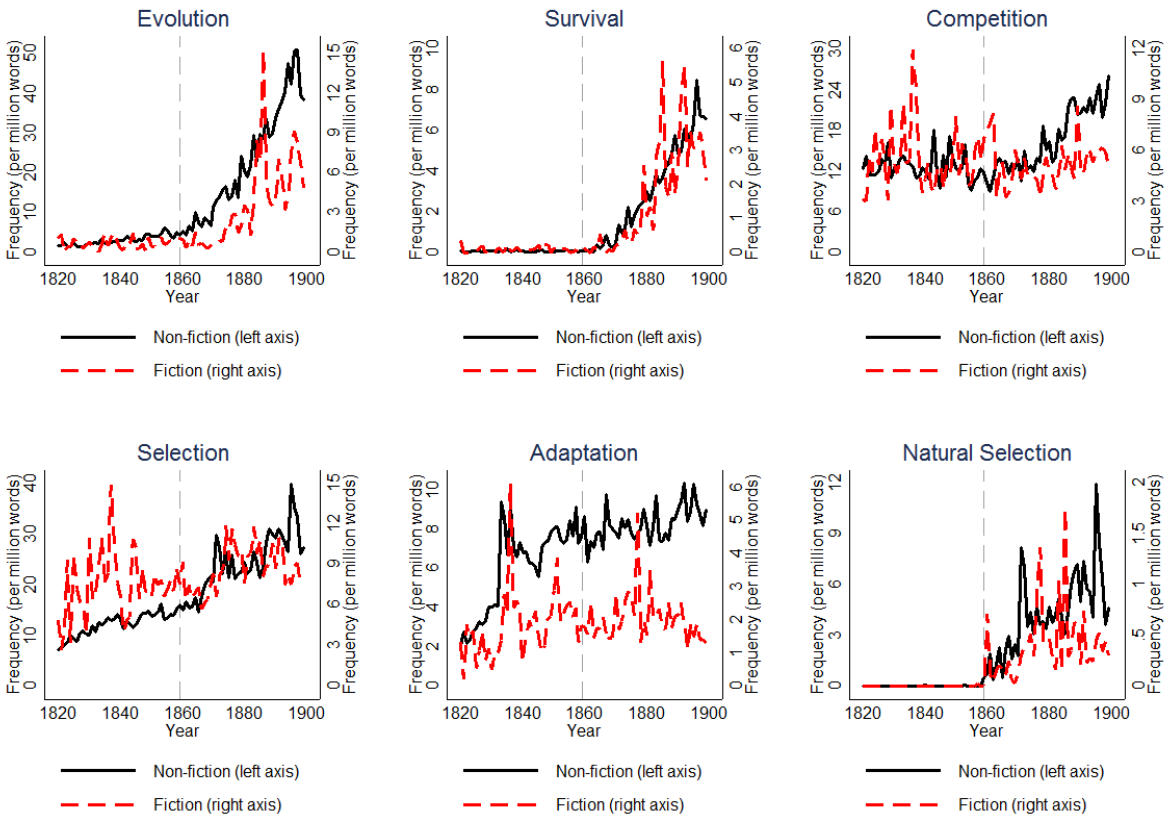
Stephan, P. E. (2012). *How economics shapes science* (Vol. 1). Cambridge, MA: Harvard University Press.

Turner, F. (2010). *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. University of Chicago Press.

Wilkens, M. (2015). Digital humanities and its application in the study of literature and culture. *Comparative Literature*, 67(1):11-20.

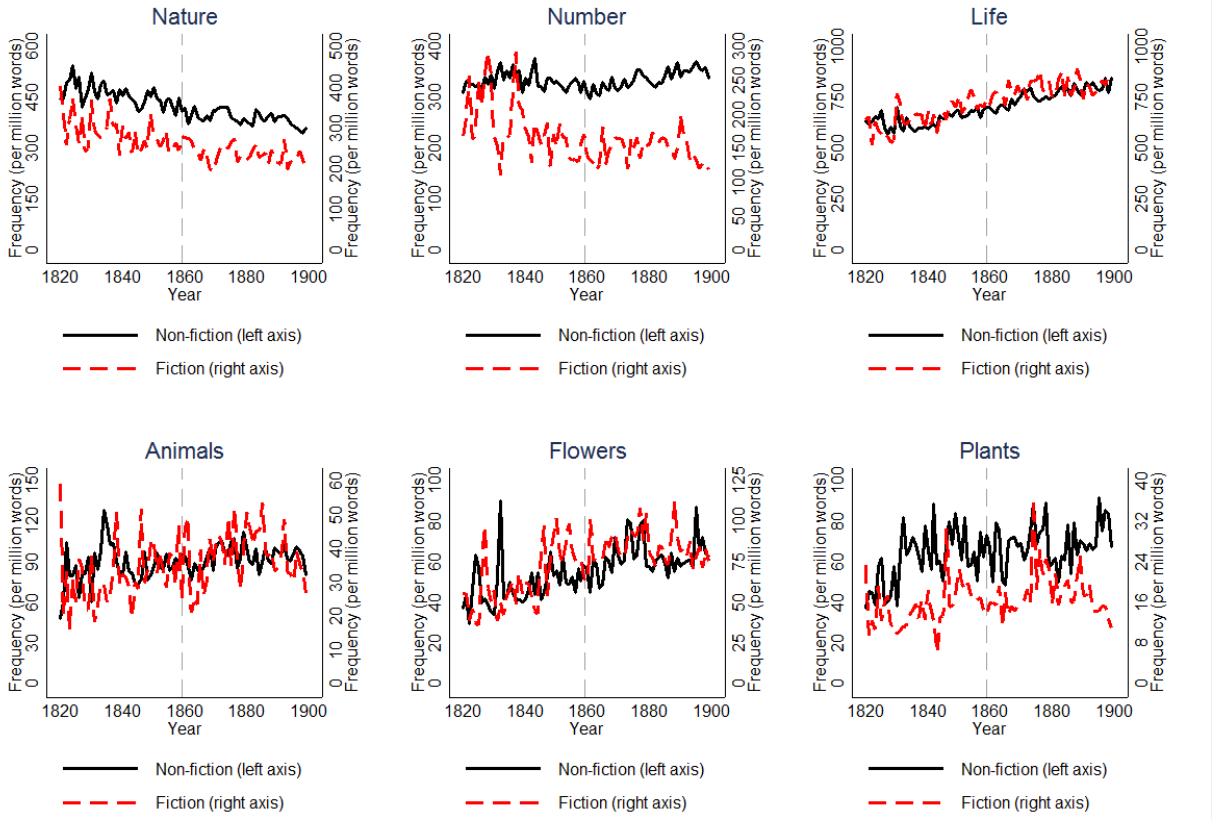
Winans, R. B. (1975). The Growth of a Novel-Reading Public in Late-Eighteenth-Century America. Early *American Literature*, 9(3), 267-275.

Figure 1: Frequencies (per 1 Million Words) of Selected Darwinian Words in the Google Books Corpora



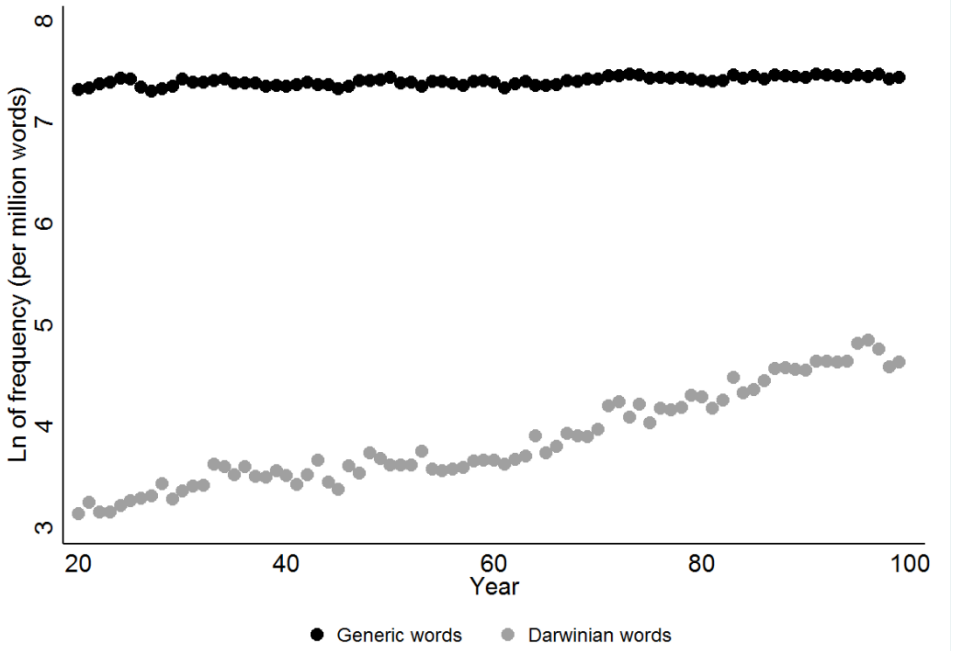
Notes: For each year, the graphs report the number of occurrences of the word or phrase indicated on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right reports the scale for fiction. Note that also the denominators for the calculation of the relative frequencies are separate for fiction and non-fiction.

Figure 2: Frequencies (per 1 Million Words) of Selected "Generic" Words in the Google Books Corpora



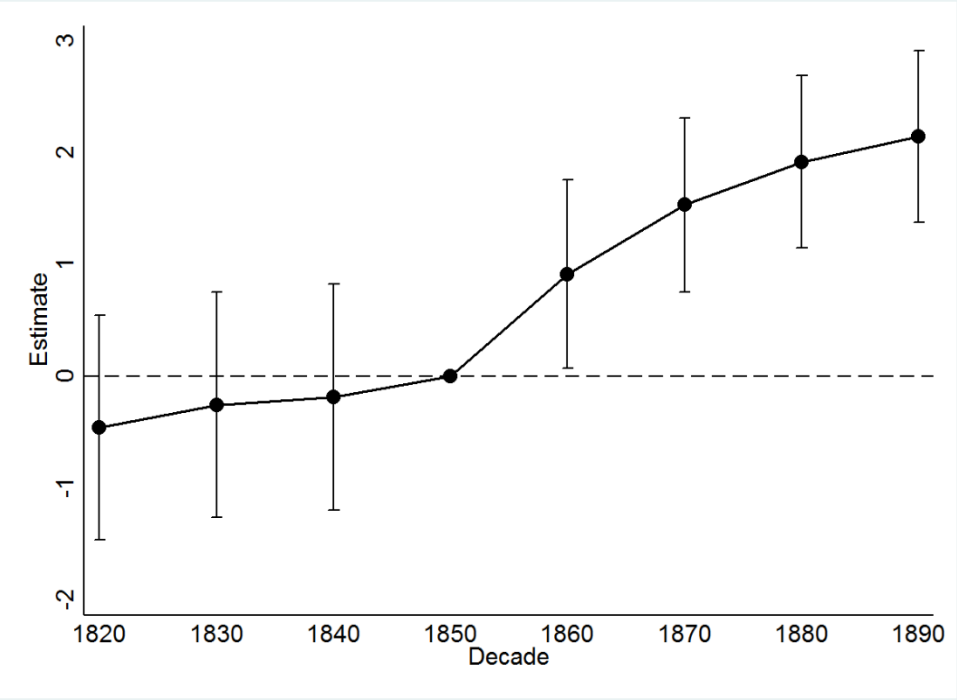
Notes: For each year, the graphs report the number of occurrences of the word or phrase indicated on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right reports the scale for fiction. Note that also the denominators for the calculation of the relative frequencies are separate for fiction and non-fiction.

Figure 3: Aggregate Yearly Frequency of Darwinian and Generic Scientific Terms, in Natural Logarithm



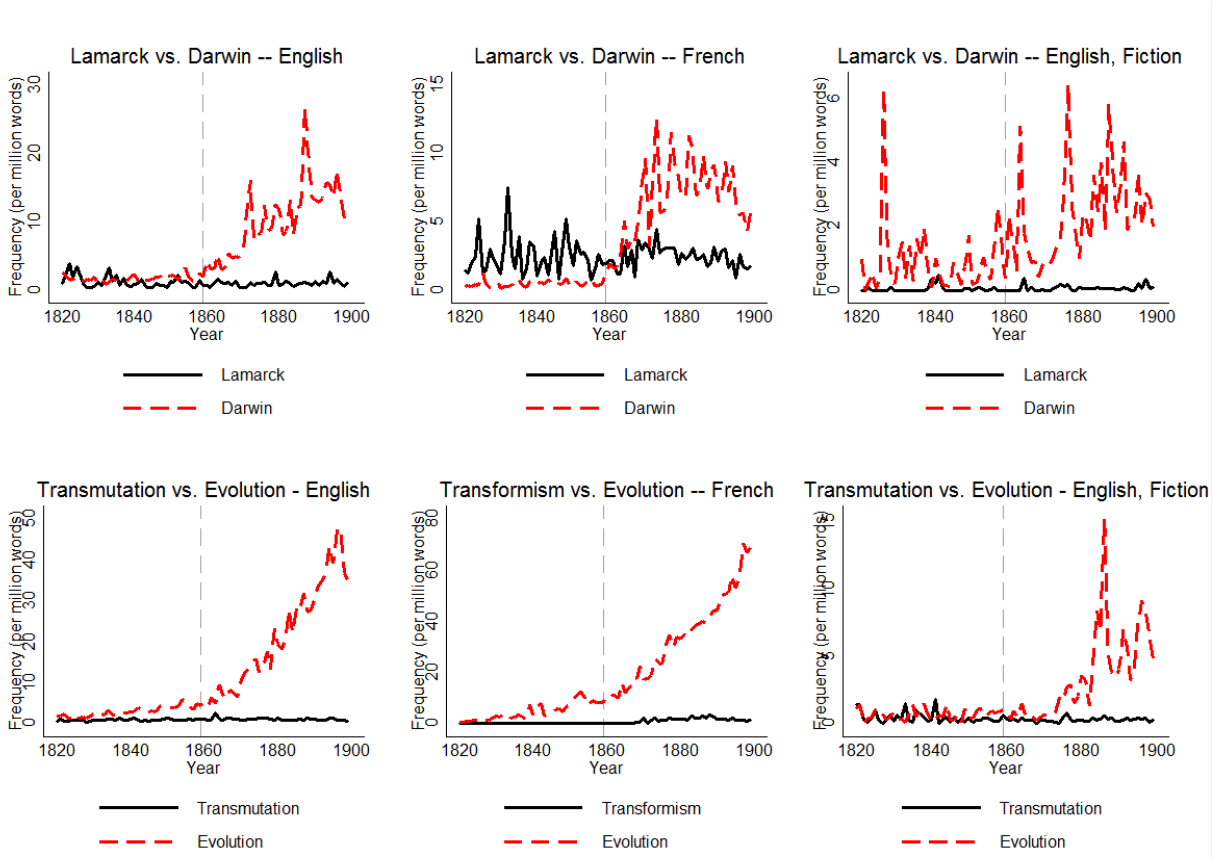
Notes: The graph displays the annual aggregate relative frequency (per million words) of the six generic scientific terms considered in the study, and the six Darwinian terms, expressed in natural logarithm.

Figure 4: Differences-in-Differences estimates of the average frequency of Darwinian and generic concepts in each decade between 1820 and 1899



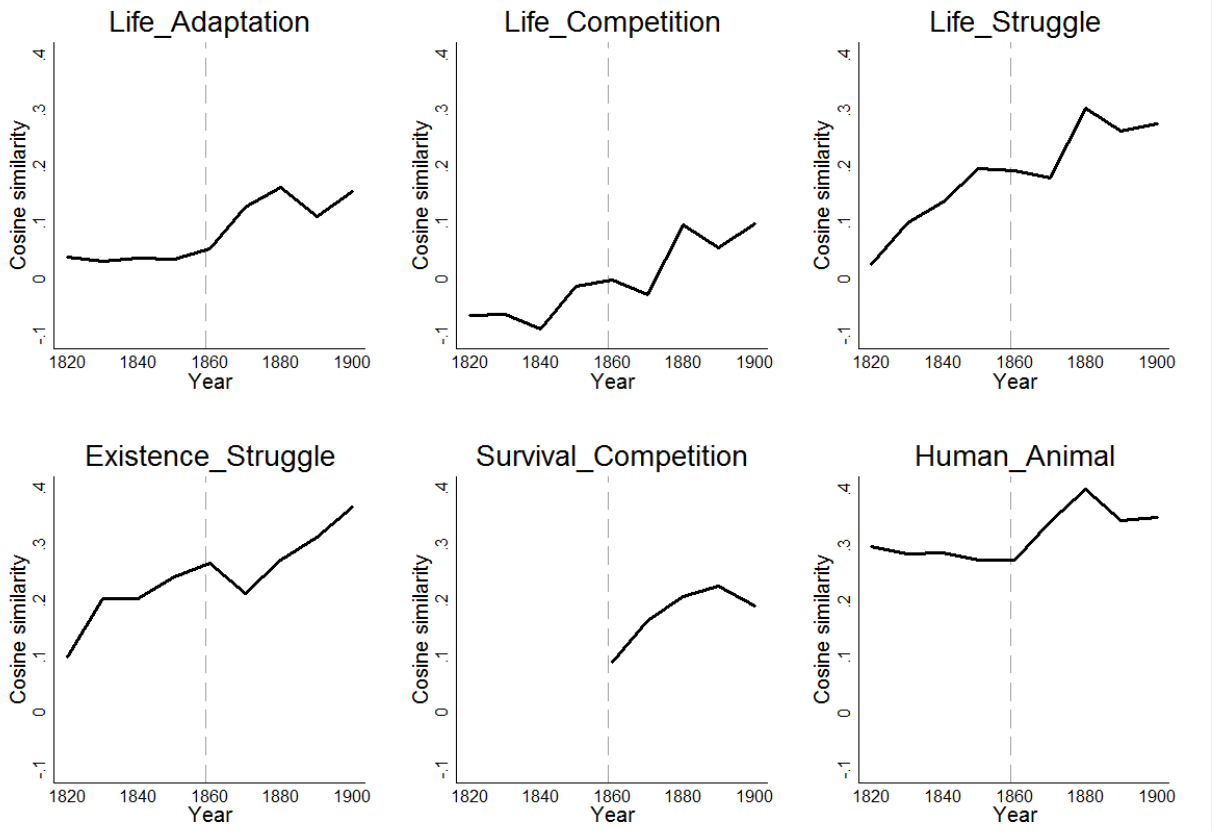
***Notes*:** Each dot in the graph represents the estimate of the parameters δ_i from the following regression model: $\ln(y_{wt}) = \alpha_w + \beta_w 1(Darwinian\ word) + \sum_{i=3}^{5} \gamma_i 1(i^{th}\ decade\ of\ 1800) + \sum_{i=7}^{10} \gamma_i 1(i^{th}\ decade\ of\ 1800) + \sum_{i=3}^{5} \delta_i 1(i^{th}\ decade\ of\ 1800) * 1(Darwinian\ word) + \sum_{i=7}^{10} \delta_i 1(i^{th}\ decade\ of\ 1800) * 1(Darwinian\ word) + \varepsilon_{wt}$, where $y_{wt}$is the frequency of use of a word per million words (plus 0.01) and the omitted (or baseline) decade is 1850-59. The vertical bars report 95% confidence intervals (from robust standard errors). On the x-axis, the 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

Figure 5: Frequencies (per 1 Million Words) of the Words "Lamarck" and "Darwin", and "Transmutation" and "Evolution" in the English and French Google Books Corpora



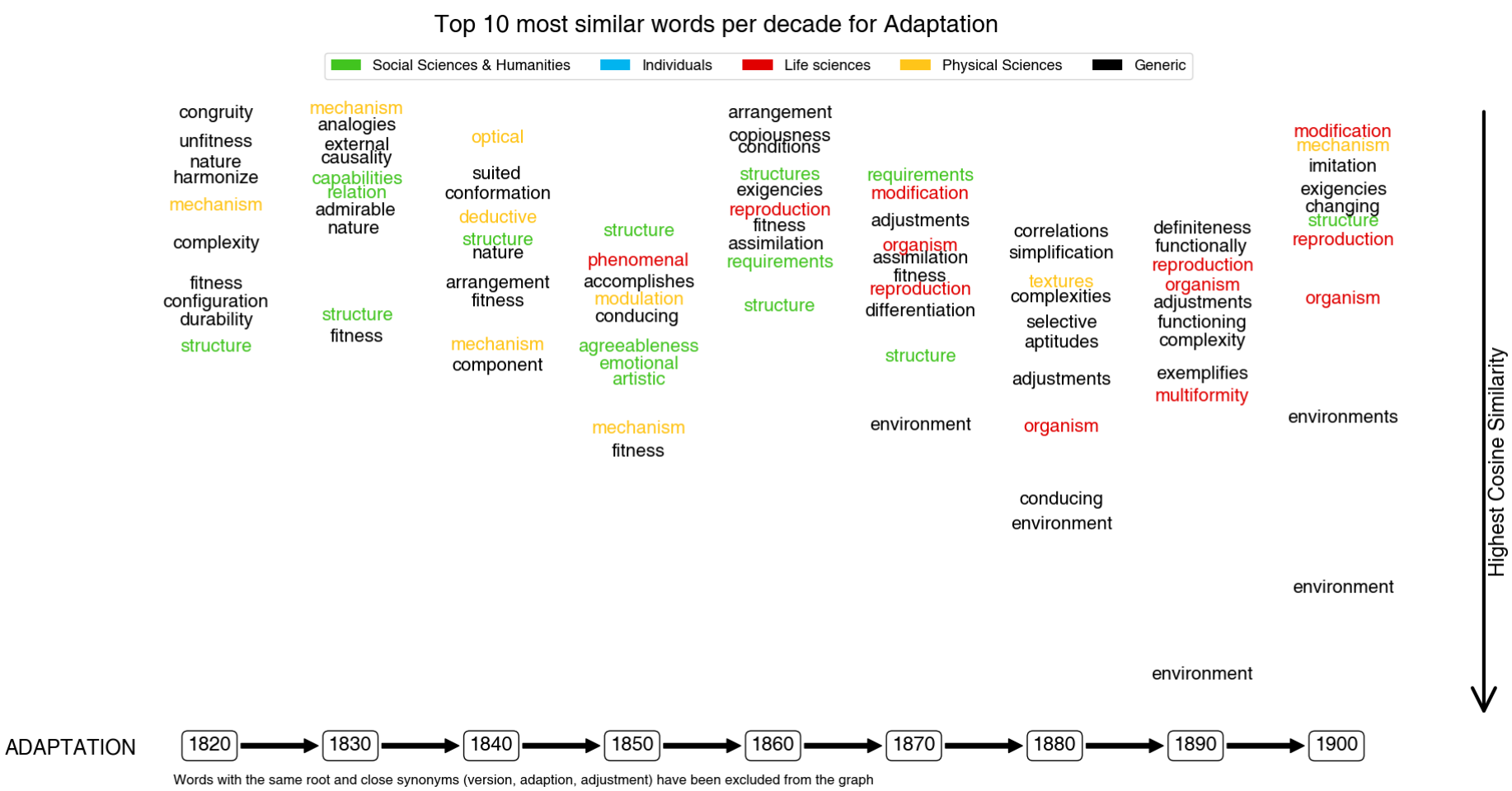
Notes: For each year, the figures report the number of occurrences (per million words) of the word or phrase indicated on top of a graph.

Figure 6: Semantic Associations between Selected Pairs of Words

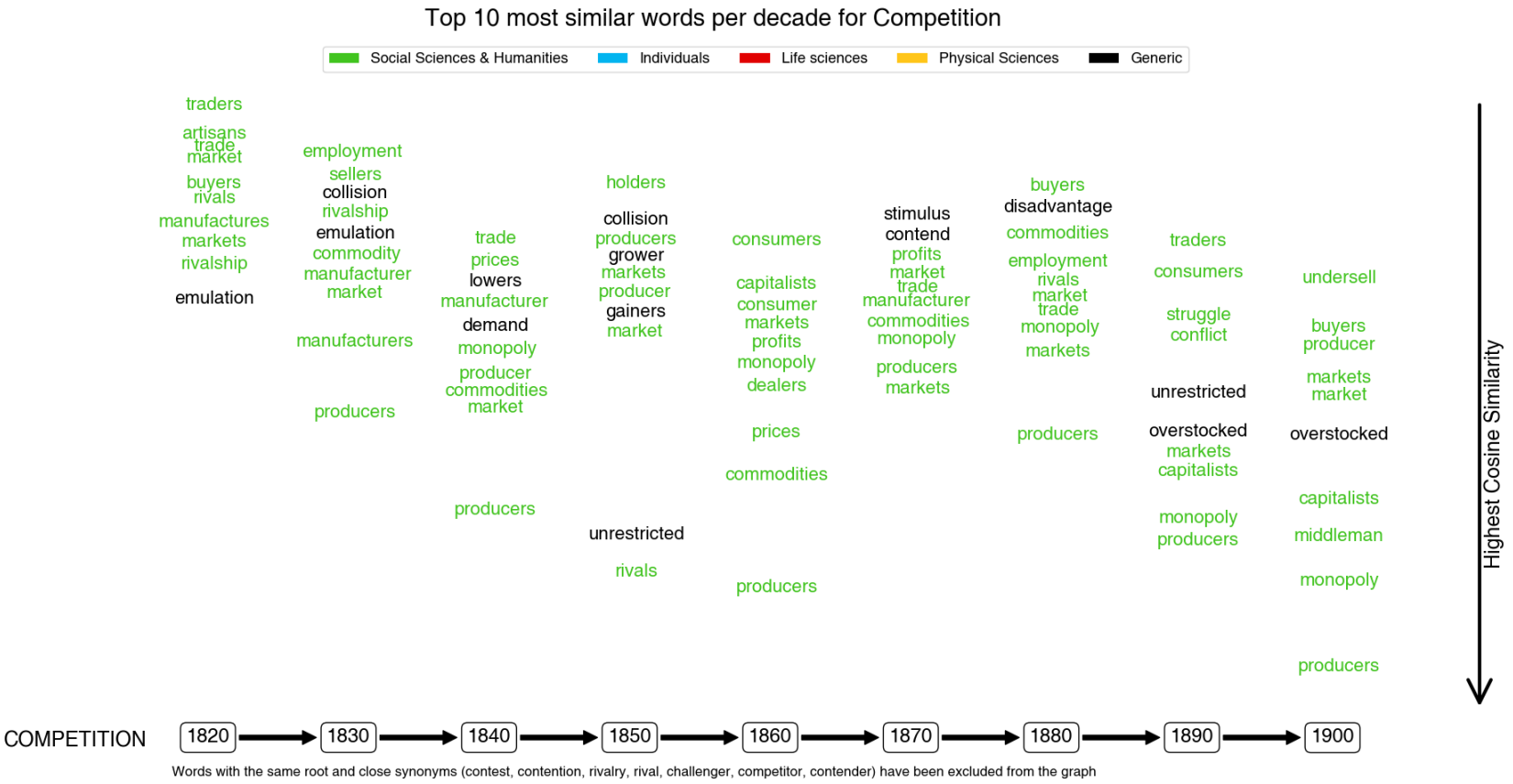


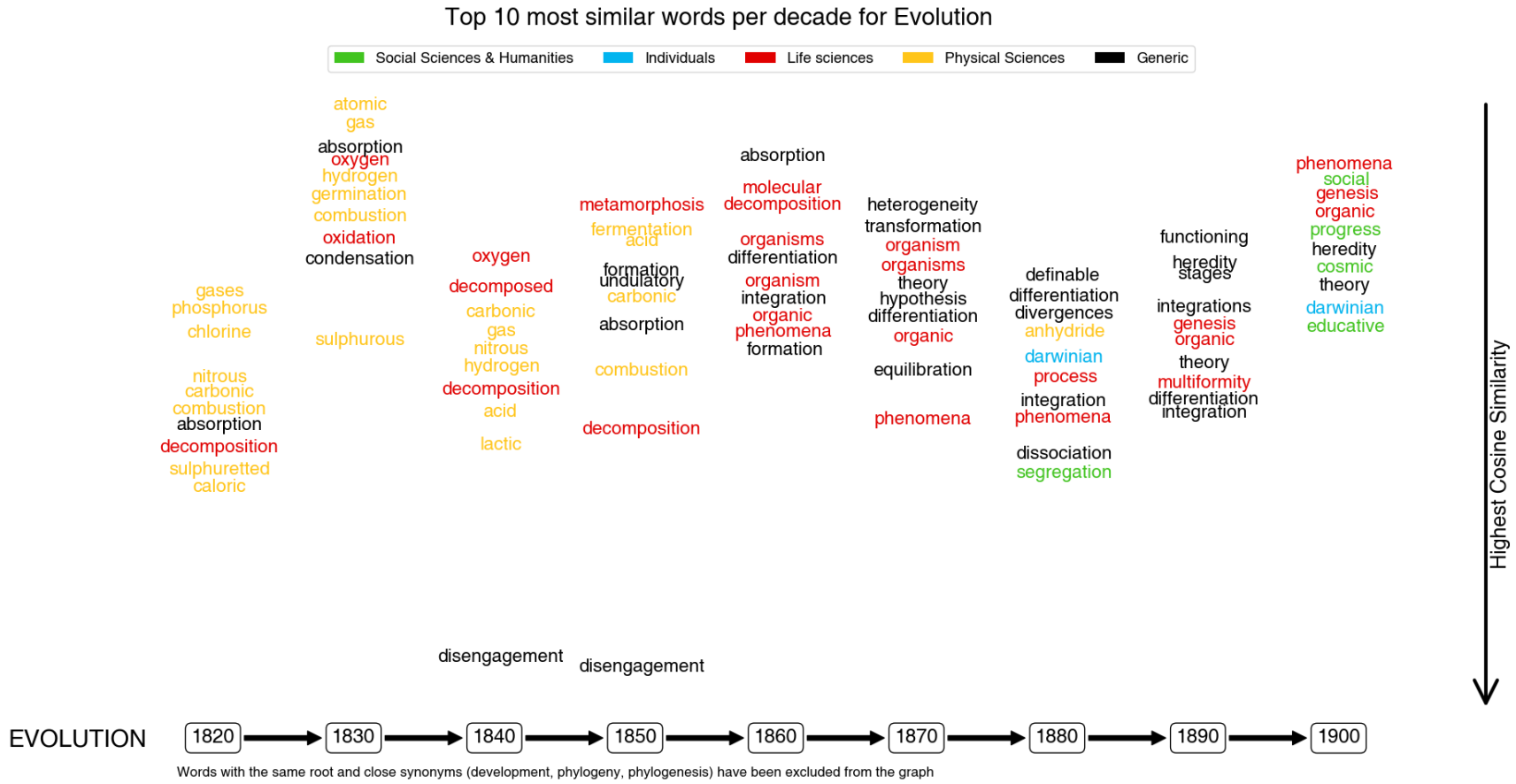
Notes: The graphs below report the similarity between each pair of words, as measured by the cosine of the angle between each pair of word vectors. The weights in the word vectors were calculated with a Word2Vec algorithm.

Figures 7 through 12: Top 10 most similar words for selected Darwinian words

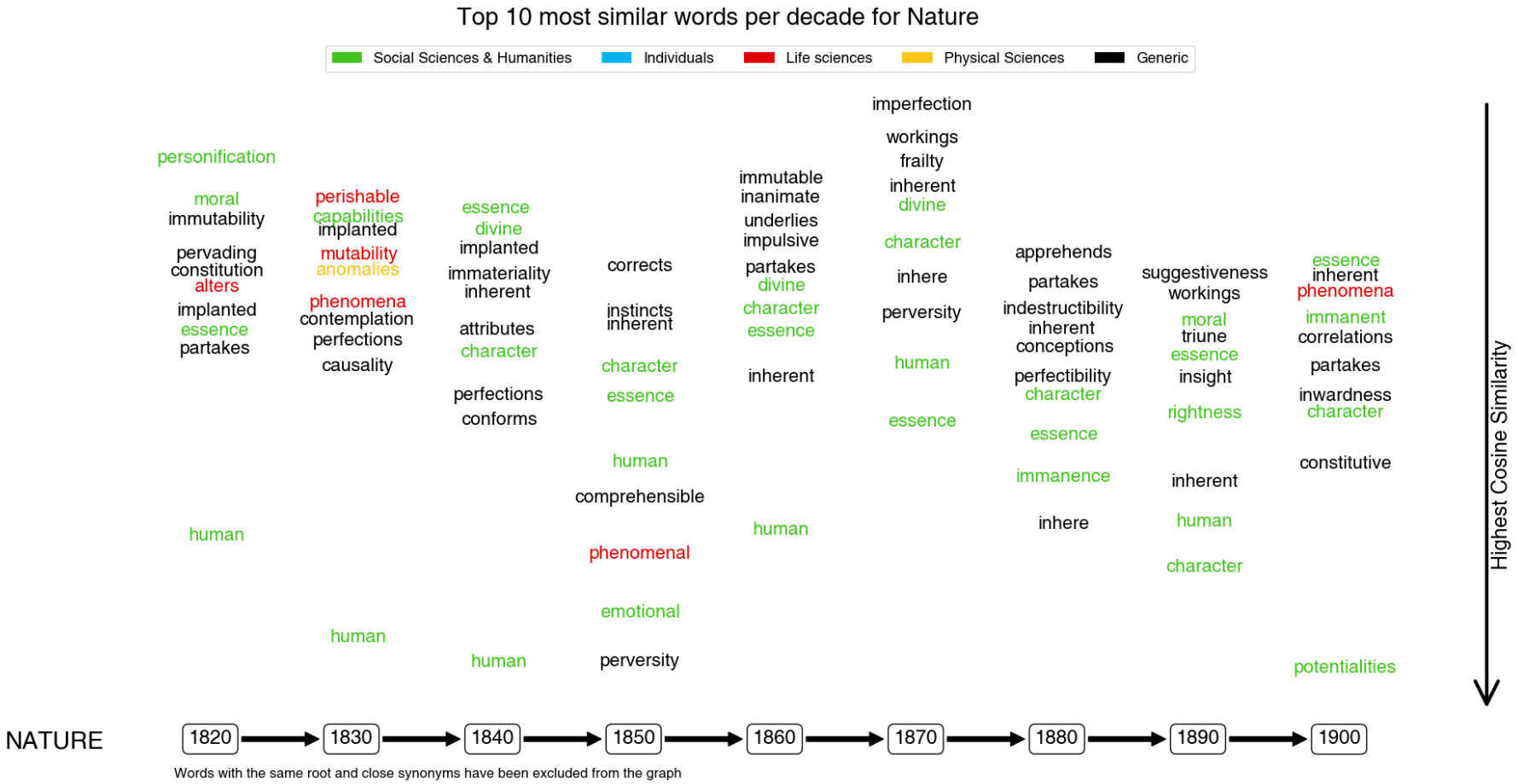


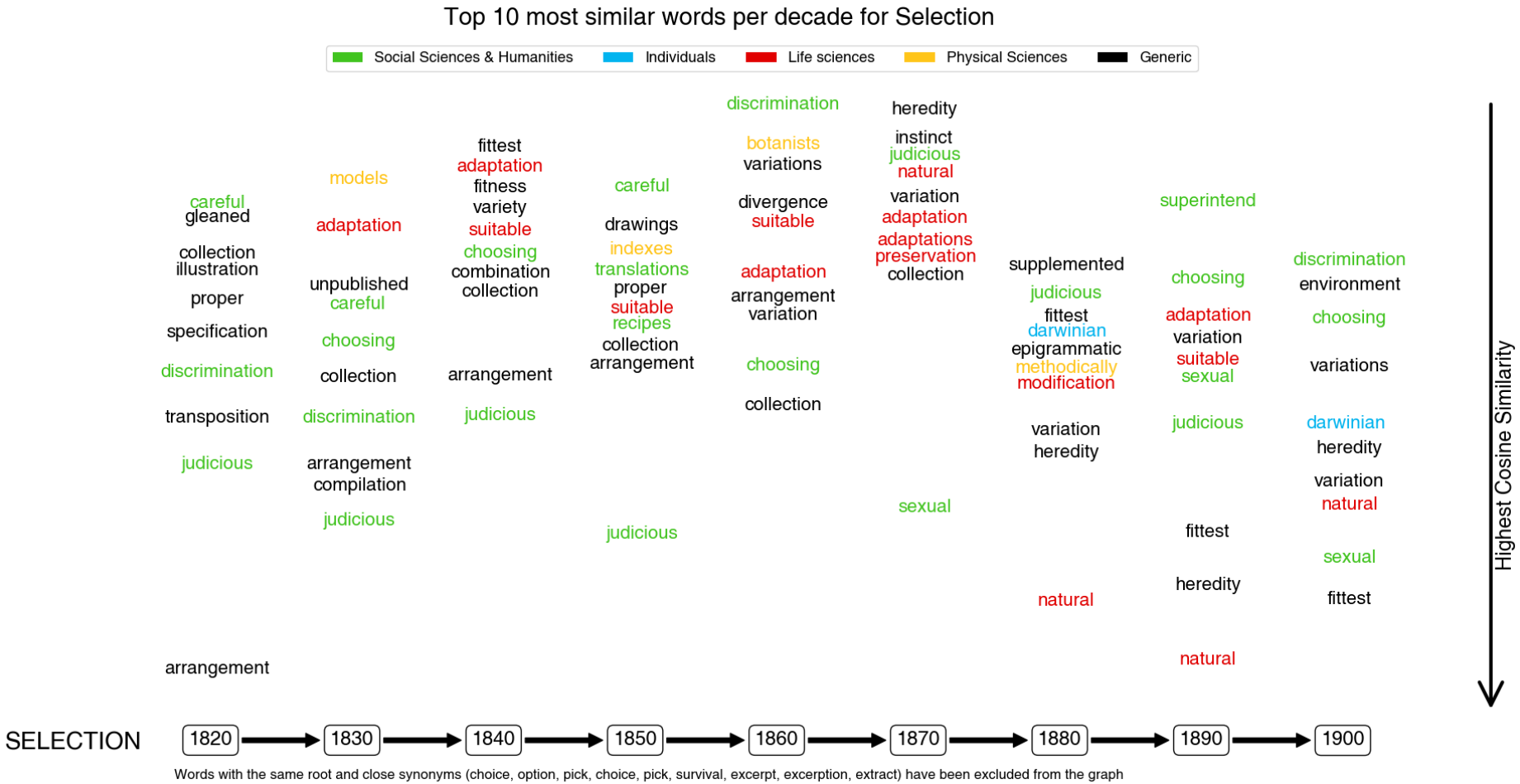
Words with the same root and close synonyms (version, adaption, adjustment) have been excluded from the graph


Top 10 most similar words per decade for Competition
Social Sciences & Humanities
Individuals
Life sciences
Physical Sciences
Generic
traders
artisans
trade
market
buyers
rivals
manufactures
markets
rivalship
emulation
employment
sellers
collision
rivalship
emulation
commodity
manufacturer
market
manufacturers
producers
trade
prices
lowers
manufacturer
demand
monopoly
producer
commodities
market
producers
holders
collision
producers
grower
markets
producer
gainers
market
unrestricted
rivals
consumers
capitalists
consumer
markets
profits
monopoly
dealers
prices
commodities
producers
stimulus
contend
profits
market
trade
manufacturer
commodities
monopoly
producers
markets
buyers
disadvantage
commodities
employment
rivals
market
trade
monopoly
markets
producers
traders
consumers
struggle
conflict
unrestricted
overstocked
markets
capitalists
monopoly
producers
undersell
buyers
producer
markets
market
overstocked
capitalists
middleman
monopoly
producers
Highest Cosine Similarity
COMPETITION
1820
1830
1840
1850
1860
1870
1880
1890
1900
Words with the same root and close synonyms (contest, contention, rivalry, rival, challenger, competitor, contender) have been excluded from the graph



Words with the same root and close synonyms (development, phylogeny, phylogenesis) have been excluded from the graph


Top 10 most similar words per decade for Nature
Social Sciences & Humanities
Individuals
Life sciences
Physical Sciences
Generic
personification
moral
immutability
pervading
constitution
alters
implanted
essence
partakes
human
perishable
capabilities
implanted
mutability
anomalies
phenomena
contemplation
perfections
causality
human
essence
divine
implanted
immateriality
inherent
attributes
character
perfections
conforms
human
corrects
instincts
inherent
character
essence
human
comprehensible
phenomenal
emotional
perversity
immutable
inanimate
underlies
impulsive
partakes
divine
character
essence
inherent
human
imperfection
workings
frailty
inherent
divine
character
inhere
perversity
human
essence
apprehends
partakes
indestructibility
inherent
conceptions
perfectibility
character
essence
immanence
inhere
suggestiveness
workings
moral
triune
essence
insight
rightness
inherent
human
character
essence
inherent
phenomena
immanent
correlations
partakes
inwardness
character
constitutive
potentialities
Highest Cosine Similarity
NATURE
1820
1830
1840
1850
1860
1870
1880
1890
1900
Words with the same root and close synonyms have been excluded from the graph



Words with the same root and close synonyms (choice, option, pick, choice, pick, survival, excerpt, excerption, extract) have been excluded from the graph

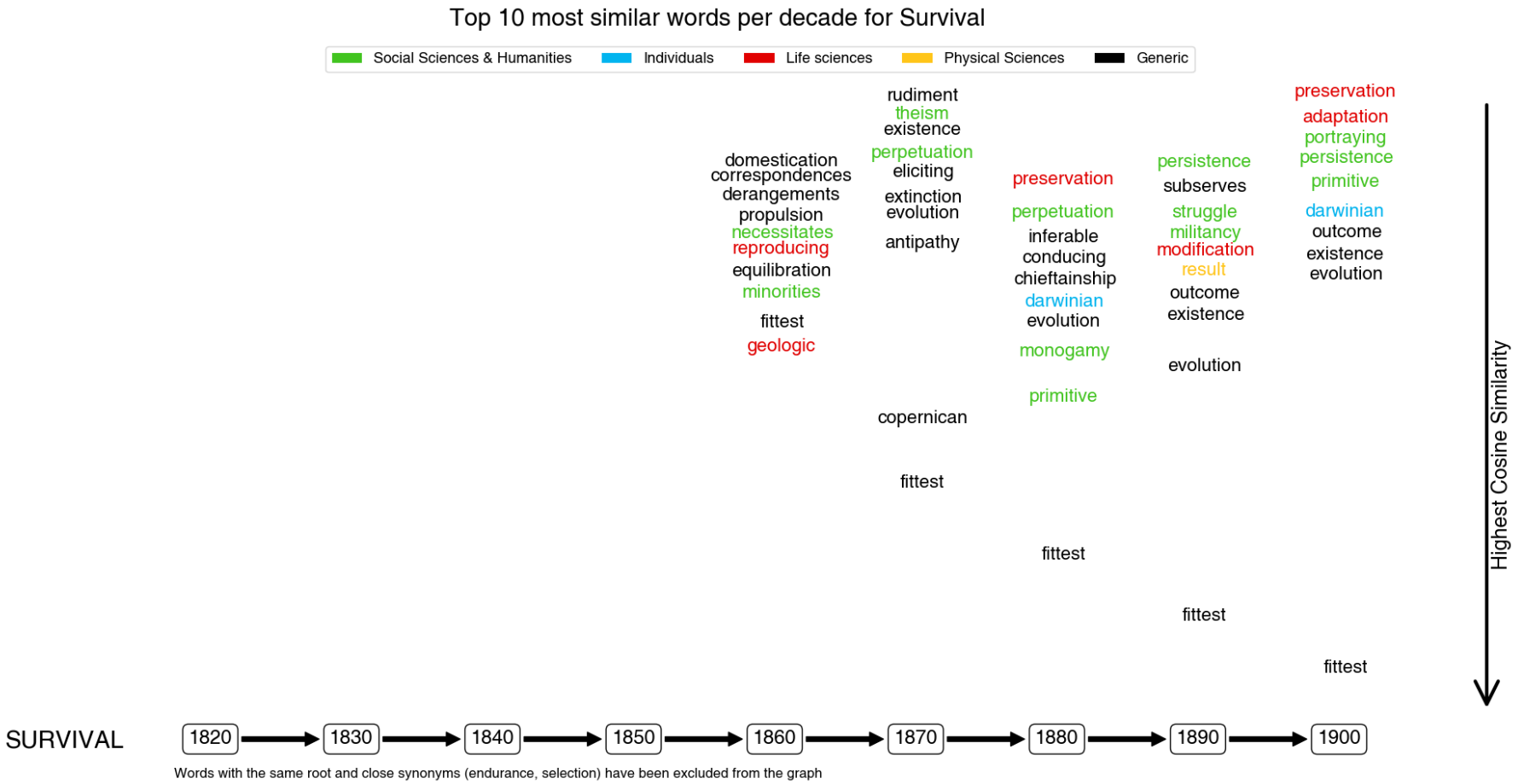

Top 10 most similar words per decade for Survival
Social Sciences & Humanities
Individuals
Life sciences
Physical Sciences
Generic
domestication
correspondences
derangements
propulsion
necessitates
reproducing
equilibration
minorities
fittest
geologic
rudiment
theism
existence
perpetuation
eliciting
extinction
evolution
antipathy
copernican
fittest
preservation
perpetuation
inferable
conducing
chieftainship
darwinian
evolution
monogamy
primitive
fittest
persistence
subserves
struggle
militancy
modification
result
outcome
existence
evolution
fittest
preservation
adaptation
portraying
persistence
primitive
darwinian
outcome
existence
evolution
fittest
Highest Cosine Similarity
SURVIVAL
1820
1830
1840
1850
1860
1870
1880
1890
1900
Words with the same root and close synonyms (endurance, selection) have been excluded from the graph

Table 1: Regression Analyses – Darwinian Concepts

Word/phrase:	Evolution			Selection			Competition		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Year (1820-99)	0.483***		-0.003	0.277***		0.177***	0.102***		-0.071***
	(0.036)		(0.022)	(0.013)		(0.019)	(0.014)		(0.017)
1(Year>1859)		18.207***			11.257***			3.488***	
		(2.019)			(0.896)			(0.720)	
(Year-1859)x1(Year>1859)			0.953***			0.194***			0.340***
			(0.062)			(0.049)			(0.033)
Constant	-16.618***	3.000***	2.498***	1.235*	12.058***	5.136***	7.934***	12.287***	14.762***
	(1.980)	(0.183)	(0.740)	(0.660)	(0.346)	(0.719)	(0.792)	(0.291)	(0.750)
Estimated single structural break (Wald chi2)	1864 (557.6)			1840 (70)			1878 (154.5)		
Wald chi2 test for structural break in 1859	474.6			36.7			128.8		
Observations	80	80	80	80	80	80	80	80	80
R-squared	0.765	0.510	0.952	0.861	0.669	0.888	0.426	0.231	0.720

Word/phrase:	Survival			Adaptation			Natural Selection		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Year (1820-99)	0.077***		-0.020***	0.059***		0.117***	0.098***		0.027***
	(0.007)		(0.004)	(0.007)		(0.013)	(0.008)		(0.010)
1(Year>1859)		2.824***			1.975***			4.266***	
		(0.365)			(0.356)			(0.378)	
(Year-1859)x1(Year>1859)			0.191***			-0.115***			0.138***
			(0.010)			(0.019)			(0.028)
Constant	-3.098***	0.071***	0.725***	3.378***	5.873***	1.071	-3.656***	0.020	-0.890**
	(0.376)	(0.005)	(0.148)	(0.543)	(0.335)	(0.685)	(0.431)	(0.013)	(0.345)
Estimated single structural break (Wald chi2)	1864 (1414.2)			1833 (169.9)			1859 (229.5)		
Wald chi2 test for structural break in 1859	1201.4			60.4			229.5		
Observations	80	80	80	80	80	80	80	80	80
R-squared	0.688	0.434	0.952	0.530	0.283	0.658	0.693	0.620	0.779

Notes: The table reports regressions of the relative annual frequency of use (per million words) of a given word or phrase on a linear time trend, an indicator for the years after 1859, and an interaction of this indicator with the difference between the current year and 1859. Each regression is limited to one word or phrase as indicated in the corresponding columns, and includes 80 observations, one for each year from 1820 to 1899. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Regression Analyses – Generic Scientific Words

Word:	Nature			Number			Life		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Year (1820-99)	-1.666***		-1.992***	0.035		-0.593***	2.986***		2.415***
	(0.115)		(0.313)	(0.070)		(0.166)	(0.170)		(0.376)
1(Year>1859)		-70.365***			-2.560			126.595***	
		(6.033)			(3.562)			(9.133)	
(Year-1859)x1(Year>1859)			0.640			1.232***			1.121**
			(0.464)			(0.256)			(0.545)
Constant	515.198***	451.252***	528.041***	323.587***	326.965***	348.308***	528.062***	642.419***	550.545***
	(8.294)	(5.092)	(15.258)	(4.815)	(2.544)	(8.139)	(12.055)	(6.237)	(19.107)
Estimated single structural break (Wald chi2)	1890 (31)			1860 (40.9)			1844 (76.7)		
Wald chi2 test for structural break in 1859	7.8			32			11.5		
Observations	80	80	80	80	80	80	80	80	80
R-squared	0.760	0.636	0.767	0.003	0.007	0.206	0.844	0.711	0.851

Word:	Animals			Flowers			Plants		
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Year (1820-99)	0.149**		0.237	0.337***		0.399***	0.211***		0.427***
	(0.065)		(0.156)	(0.047)		(0.106)	(0.053)		(0.117)
1(Year>1859)		6.097**			14.494***			6.299**	
		(2.554)			(2.116)			(2.557)	
(Year-1859)x1(Year>1859)			-0.173			-0.120			-0.424**
			(0.214)			(0.154)			(0.197)
Constant	78.756***	84.586***	75.294***	36.665***	49.499***	34.257***	50.137***	59.520***	41.631***
	(4.716)	(2.221)	(8.057)	(3.200)	(1.617)	(5.311)	(3.645)	(2.018)	(5.421)
Estimated single structural break (Wald chi2)	1870 (16.5)			1879 (13)			1837 (39)		
Wald chi2 test for structural break in 1859	0.9			0.3			6.7		
Observations	80	80	80	80	80	80	80	80	80
R-squared	0.087	0.068	0.094	0.434	0.376	0.438	0.172	0.072	0.216

Notes: The table reports regressions of the relative annual frequency (per million words) of use of a given word or phrase on a linear time trend, an indicator for the years after 1859, and an interaction of this indicator with the difference between the current year and 1859. Each regression is limited to one word or phrase as indicated in the corresponding columns, and includes 80 observations, one for each year from 1820 to 1899. The time trend is expressed as the last two digits of the corresponding year (e.g. 28 indicates 1828). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Regression Analyses – Darwinian concepts: Fiction and Non-fiction

Word/phrase:	Evolution					Selection					Competition				
Sample:	Non-fiction		Fiction		All	Non-fiction		Fiction		All	Non-fiction		Fiction		All
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Year (1820-99)	0.519***	-0.009	0.077***	-0.022***	-0.009	0.299***	0.183***	0.035***	0.031	0.183***	0.117***	-0.074***	-0.004	-0.008	-0.074***
	(0.040)	(0.024)	(0.011)	(0.008)	(0.024)	(0.015)	(0.020)	(0.011)	(0.025)	(0.020)	(0.015)	(0.018)	(0.008)	(0.019)	(0.018)
1(Fiction)					-1.388**					0.984					-9.378***
					(0.682)					(1.329)					(1.321)
Year x Fiction					-0.013					-0.152***					0.066**
					(0.019)					(0.029)					(0.028)
(Year-1859)x1(Year>1859)		1.037***		0.195***	1.037***		0.229***		0.009	0.229***		0.374***		0.007	0.374***
		(0.068)		(0.025)	(0.068)		(0.052)		(0.035)	(0.052)		(0.036)		(0.029)	(0.036)
(Year-1859)x1(Year>1859)x1(Fiction)					-0.842***					-0.220***					-0.367***
					(0.061)					(0.058)					(0.048)
Constant	-18.021***	2.781***	-2.523***	1.393***	2.781***	0.550	5.138***	5.937***	6.122***	5.138***	7.644***	15.150***	5.629***	5.771***	15.150***
	(2.168)	(0.807)	(0.517)	(0.305)	(0.810)	(0.736)	(0.746)	(0.771)	(1.269)	(0.749)	(0.865)	(0.790)	(0.575)	(0.943)	(0.792)
Estimated single structural break (Wald chi2)	1864 (547.8)		1874 (74.5)			1867 (70.6)		1874 (30.7)			1878 (158.3)		1873 (23.7)		
Wald chi2 test for structural break in 1859	469.5		64.9			42.7		1.1			129.9		0.4		
Observations	80	80	80	80	160	80	80	80	80	160	80	80	80	80	160
R-squared	0.760	0.949	0.450	0.629	0.952	0.859	0.890	0.145	0.145	0.912	0.451	0.740	0.003	0.004	0.890

Word/phrase:	Survival					Adaptation					Natural Selection				
Sample:	Non-fiction		Fiction		All	Non-fiction		Fiction		All	Non-fiction		Fiction		All
	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Year (1820-99)	0.080***	-0.021***	0.042***	-0.013***	-0.021***	0.065***	0.122***	0.007	0.025***	0.122***	0.098***	0.027**	0.009***	0.005***	0.027**
	(0.007)	(0.004)	(0.005)	(0.004)	(0.004)	(0.008)	(0.014)	(0.004)	(0.008)	(0.014)	(0.008)	(0.010)	(0.001)	(0.002)	(0.010)
1(Fiction)					-0.255*					-0.244					0.721**
					(0.133)					(0.655)					(0.335)
Year x Fiction					0.008**					-0.097***					-0.022**
					(0.004)					(0.013)					(0.010)
(Year-1859)x1(Year>1859)		0.199***		0.109***	0.199***		-0.112***		-0.037***	-0.112***		0.139***		0.007*	0.139***
		(0.011)		(0.012)	(0.011)		(0.020)		(0.012)	(0.020)		(0.028)		(0.004)	(0.028)
(Year-1859)x1(Year>1859)x1(Fiction)					-0.090***					0.075***					-0.132***
					(0.014)					(0.020)					(0.027)
Constant	-3.228***	0.762***	-1.671***	0.507***	0.762***	3.302***	1.047	1.549***	0.804*	1.047	-3.681***	-0.891**	-0.304***	-0.170***	-0.891**
	(0.396)	(0.156)	(0.252)	(0.132)	(0.156)	(0.553)	(0.710)	(0.300)	(0.411)	(0.712)	(0.434)	(0.347)	(0.051)	(0.055)	(0.348)
Estimated single structural break (Wald chi2)	1864 (1331.2)		1863 (101.3)			1833 (158.7)		1889 (39.8)			1859 (229.0)		1828 (55.3)		
Wald chi2 test for structural break in 1859	1116.9		99.8			56.6		11.4			229.0		55.9		
Observations	80	80	80	80	160	80	80	80	80	160	80	80	80	80	160
R-squared	0.685	0.950	0.514	0.725	0.892	0.567	0.673	0.029	0.088	0.891	0.693	0.779	0.382	0.397	0.820

Notes: For each word or phrase, the first two columns report estimates from specifications as in Models 3 and 5 above, limited to non-fiction books, whereas the third and fourth columns display estimates from the corpus of fiction books. The fifth column of each block reports parameter estimates from Model 6 above, which also includes interactions for whether the corpus is that of fiction or non-fiction books but separately for fiction and non-fiction books. The estimates in the fifth column are based on regressions on 160 observations, two per each year between 1820 and 1899. Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Differences-in-Differences regressions – Darwinian and Generic Scientific Concepts

Outcome variable:	Difference in ln of frequence (per million words) between Darwinian terms and generic scientific terms		ln of Frequence (per million words)
	(1)	(2)	(3)
ln Year (1820-99)	0.908***	0.385***	0.020
	(0.049)	(0.037)	(0.016)
(ln (Year)-ln(1859))x1(Year>1859)		1.597***	0.126***
		(0.090)	(0.031)
1(Darwin)			-3.731***
			(0.019)
(ln (Year)-ln(1859)) x 1(darwin)			0.385***
			(0.038)
(ln (Year)-ln(1859))x1(Year>1859)x1(Darwin)			1.597***
			(0.097)
Constant	-7.162***	-5.303***	7.313***
	(0.191)	(0.140)	(0.060)
Estimated single structural break (Wald chi2)	1855 (390.9)		
Wald chi2 test for structural break in 1859	336.1		
Observations	80	80	160
R-squared	0.825	0.965	0.998

Notes: Columns 1 and 2 report estimates from regressions where the outcome variable is the difference in the natural logarithm of the aggregate yearly frequencies of Darwinian and generic terms. The outcome variable in the regression whose parameter estimates are in column 3 is the natural logarithm of the yearly aggregated frequency of either the Darwinian concepts or the generic scientific words that we considered. Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.