NBER WORKING PAPER SERIES

DOES SCIENTIFIC PROGRESS AFFECT CULTURE? A DIGITAL TEXT ANALYSIS

Michela Giorcelli
Nicola Lacetera
Astrid Marinoni

Working Paper 25429
http://www.nber.org/papers/w25429

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2019

We gratefully acknowledge the financial support of the National Bureau of Economic Research through the Innovation Policy Grants Program. We also thank Dora Costa, Ryan Heuser, Graeme Hirst, Xander Manshel and Yang Xu for their suggestions; and participants to presentations at Brown University, the University of Toronto, the University of Munich, the NBER Productivity Lunch, the 2018 REER Conference at Georgia Tech, the Workshop in Memory of Luigi Orsenigo at Bocconi University, the 2019 NBER Summer Institute and the 2018 Academy of Management Annual Meetings for their helpful feedback. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Scientific Progress Affect Culture? A Digital Text Analysis
Michela Giorcelli, Nicola Lacetera, and Astrid Marinoni
NBER Working Paper No. 25429
January 2019, Revised MAy 2020
JEL No. B55,C55,N00,O30,Z1

ABSTRACT

We focus on a unique episode in the history of science, the elaboration of the theory of evolution by Charles Darwin, to study the interplay between scientific progress and cultural change. We perform text analysis on a corpus of hundreds of thousands of books, with the use of techniques from machine learning. We examine, in particular, the diffusion of certain key ideas of the theory of evolution in the broader cultural discourse and 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, and a more positive sentiment toward these ideas, thus providing further insights about the channels through which Darwin's theory influenced the broader discourse. Our findings provide the first large-sample, systematic quantitative evidence of the relation between two key factors of long-term economic growth (science and culture), and suggest that machine learning and natural language processing offer promising tools to explore this relation.

Michela Giorcelli
Department of Economics
University of California at Los Angeles
Bunche Hall 9262
Los Angeles, CA 90095
and NBER
mgiorcelli@econ.ucla.edu

Nicola Lacetera
University of Toronto
Institute for Management and Innovation
3359 Mississauga Road, Room KN 235
Mississauga, ON L5L 1C6
CANADA
and NBER
nicola.lacetera@utoronto.ca

Astrid Marinoni
Rotman School of Management, University of Toronto
105 St. George St.
Toronto, ON M5S 3E6 CANADA
astrid.marinoni14@rotman.utoronto.ca

1. Introduction

"Economic change in all periods depends, more than most economists think, on what people believe." (Mokyr, The Enlightened Economy, p.1, 2012)

Scientific and technological progress is a fundamental driver of economic growth and human development; as such, it is crucial to understand what its determinants are, and the ways in which it impacts economic outcomes. According to Acemoglu et al. (2001), for example, institutions have a fundamental role in affecting a society's capacity to produce new technologies. Other scholars, such as McCloskey (2016) and Mokyr (2015), emphasize the role that culture played in encouraging scientific discoveries, and, through them, technological advances. Guiso, Sapienza, and Zingales (2003, 2004a) document a strong relationship between culture and macroeconomic outcomes, and Williamson (2000) argues that culture could lead to permanent effects on economic development, given its persistence in the long run.

Despite the evidence on the impact of both scientific progress and cultural change on economic variables, we know little about the relationship between these two broad phenomena. In this paper, we propose the first large-sample empirical study of one aspect of this relationship, namely the impact of scientific progress on cultural change. This exercise presents three main challenges. First, one would need a long-time horizon to analyze the interplay between public discourse and scientific and technological progress. Second, the plausible two-way relationship between science and cultural change 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 the emergence and acceptance of scientific discoveries. Third, the determination of what constitutes culture and cultural change, and how to measure them, is problematic.

We focus on one of the major advancements in the history of science: Charles Darwin's theory of evolution by natural selection, a milestone discovery with implications well beyond biological sciences (Ayala 2009). That Darwin's theory had a broad social and cultural impact is, on the one hand, undisputed; on the other hand, evidence of *how* this influence occurred is scant. This limits our ability to locate the theory of evolution in the broader social and economic history. With our study, we begin to fill this gap.

The publication of *On the Origins of Species*, in 1859, made Darwin's theory known to a vast public; the timing of the publication was largely unplanned, and we rely on this event as our main source of natural variation. To measure cultural change and its diffusion, we perform text analysis using natural language processing methods, based on machine learning, on a large corpus of digitized books. Specifically, we investigate how certain scientific concepts that were central to Darwin's theory spread in the cultural discourse. By operationalizing concepts with specific words and phrases, we are able to document which particular ideas were actually culturally novel in Darwin's contributions, which ones were already present in the public discourse, and which ones were present but influenced different cultural spheres over time.

The main text corpus on which we perform our analysis is Google Books, a digitized collection of about eight million books. We define the publication year of *On the Origin of Species* (1859) as our reference date and concentrate the analysis on the four decades before and after it. We consider words and expressions that, according to many accounts, represent the key concepts in Darwin's theory (Desmond and Moore 1994, Mayr 1982): Evolution, Survival, Competition, (Natural) Selection, and Adaptation. The frequencies of use of these words provide a measure of the adoption and relevance of certain concepts in the public discourse. We compare, both descriptively and in difference-in-differences econometric frameworks, the evolution of the frequency of use of Darwinian concepts with the frequency of a large number of words not related directly to Darwin's theory but extensively present in *On the Origins of Species*.

We complement the frequency analysis on the Google Books corpus with evidence from Congressional Records, which include all the remarks and debates given on the floor of the House and Senate of the United States, again for the period 1820-1899. With this additional corpus, we assess whether certain concepts diffused not only in the cultural discourse, but also in the political arena, thus potentially shaping the policy debate.

We then explore the semantic change of these words as well as the changes in "sentiments" towards them over time. To study semantic and sentiment evolution, we employ word-embedding techniques from the Natural Language Processing and Machine Learning literature.

We present two main sets of results. First, some key concepts in Darwin's theory became relevant in the broader cultural discourse in the years immediately after 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 fiction literature. This indicates that

these concepts had a broad impact on culture as well as on 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 before 1859 and diffused rapidly thereafter; this suggests a potential change in the way the term Selection was used. We also document that some of the key Darwinian ideas entered the policy debate, which we measure with U.S. Congress textual data, after 1859 but with some delay with respect to the entry of these concepts in the broader public discourse. The effects of *On the Origin of the Species* were not specific to the English-speaking world; the Darwinian concepts diffused in non-English speaking countries right after the translation of the book in the corresponding language. Moreover, the translation occurred earlier in countries that industrialized earlier, such as Germany and France, than in "late comers" such as Italy and Spain.

The second set of results concerns changes in the semantics of these words as well as in the types of reactions, or sentiments, that they generated over time. Of interest is the increase in semantic association between certain words, such as Competition (or Struggle) and Life, as well as between Life and Adaptation, 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. Furthermore, 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, indicating a broader reach of this idea in society. Furthermore, Selection became more similar in meaning to other "Darwinian" words, such as Survival, Variation, Fittest and Heredity. The key concepts of Darwin's theory of evolution that we consider also have higher semantic similarity with the word Darwin itself than with the names of other scientists who elaborated theories of evolutions, such as Lamarck, Wallace and Chambers. This suggests that such concepts as Evolution, Selection, Survival, Competition and Adaptation were particularly associated, in the public discourse, with Darwin's work and not just generically with the progress in the biological sciences of the time. Finally, sentiment analysis shows a more

positive attitude toward certain Darwinian concept after the publication of *On the Origins of Species*, in particular Evolution, and a positive attitude toward Darwin himself.

Our results provide insights as to how Darwin's theory of evolution led to the diffusion of some concepts in the broader cultural discourse, and affected the use, meaning and public perception of concepts that were already part of the public conversation. To the extent that a culture that values scientific inquiry and evidence is more likely to promote economic development (Mokyr 2016), a channel through which this appreciation may occur is precisely through scientific progress; enhancing the knowledge of this channel is a meaningful exercise. The relationship between scientific discoveries and the public discourse also helps understanding deep social and political processes, such as the extent to which, to paraphrase Alexander Hamilton's reflections in the *Federalist Papers*, a society is based on a "culture of reason and evidence", i.e. whether scientific inquiry guides shared beliefs and choices.

Related literature. The stream of literature that is closest to our work includes studies 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 with their "macro" discoveries.[3] Mokyr (2013, 2016) calls these scientists "cultural entrepreneurs", and our paper attempts to quantify their impact on societal beliefs.

We also contribute to the growing use of "text as data" in the social sciences. In economics, for example, the use of text corpora is developing especially in such fields as political economy, the study of media, and innovation (Balsmeier et al. 2018; Bandiera et al. 2017; Catalini et al. 2015; Gentzkow et al. 2018, Iaria et al. 2018; Jelveh et al. 2014; Kelly et al. 2017). In business studies, there are several applications to marketing and finance (see Kearney and Liu 2014 for a review). Recent work in sociology addresses questions such as the common understanding of social class using text analysis (Kozlowski et al. 2019). Scholars in linguistics and literary criticism are increasingly employing computerized text analysis to answer questions about the evolution of literary genres and styles, and semantic changes of words and concepts. Digital text analysis or "distant reading" also allows for the inclusion in the analyses of the "great unread", i.e. the large

[3] Mokyr (2013) distinguishes between "macro" and "micro" scientific discoveries, and argues that only the first ones are able to create a discontinuous change in a society development. The latter are important to guarantee a continuous improvement, but do not cause any breakthrough.

quantity of texts that normally scholars do not study, but that, as a whole, represent the broader social and cultural climate at a given time (Cohen 1999, Heuser and Le-Khac 2011, Heuser 2016, Moretti 2013, Wilkens, 2015). An area of study known as "cultural analytics" or "culturomics" explores the evolution of culture through text analysis (Aiden and Michel 2014, Manovich 2009, Michel et al. 2011), and in particular through the study of changes in the frequency and meaning of certain words and expressions over time. To our knowledge, there are no applications to studying the public perception of science.

Finally, our work 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 at how these ideas and concepts were already permeating the public discourse.

Plan of the paper. In Section 2, we provide a brief account of Darwin's elaboration of the theory of evolution by natural selection. We also explain why the publication of *On the Origin of Species* provides natural variation that allows studying the effect of Darwin's theory on the broader public discourse. In Section 3, we describe the text-based data that we use and the techniques and empirical strategies that we adopt to extract information about cultural evolution. Section 4 reports the findings, and, in Section 5, we provide a discussion and propose directions for future research.

2. Historical Background and Identification

> *"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.

2.1. The Development of Darwin's Theory of Evolution

Charles Darwin's interest in the evolution of living organisms largely developed during his voyage on the HMS Beagle, a ship of the Royal Navy, from 1831 to 1836. Over those five years, Darwin collected fossils from the places that he visited and observed their geographical distribution. Although his early conjectures built on previous theories (such as Lamarck's and Chambers') and considered the possibility of the transformation of one species into another (transmutation), he then developed his own 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 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 in 20th century provided corroboration and foundations to Darwin's theory (Desmond and Moore 1994; Mayr 1982).[4]

In addition to being one of the greatest scientific breakthroughs in history, there is a perception that Darwin's theory of evolution had a wider cultural reach (Desmond and Moore 1994; Fuller 2017; Mayr [1982, 2001]). For example, 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 of God.[5] The likely common origins of all species, moreover, eliminated any idea of intrinsic superiority of humans over other living beings, and, within humans, of a race over another. Fuller (2017) argues that Darwin's theory had a major influence on the debate over race, slavery and discrimination in the United States, thus hinting at a major role of this scientific breakthrough in the evolution of the American society. Mokyr (2013, 2016) includes Darwin among a small set of "cultural entrepreneurs", i.e. scientists whose discoveries affected deeply held and broadly shared popular beliefs.

These accounts, however, focus on a narrow set of literary contributions or debates mostly restricted to scientific, political and economic elites; this makes it hard to advance inferences about the broader cultural impact of this scientific advance, and about the cultural climate that preceded that breakthrough. Our approach to answering these questions relies on a massive corpus of fiction and non-fiction literary work, 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.

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

[5] Research in literary criticism analyzed how the production of certain poets and novelists, began to reflect ideas of a different role that nature had in its relationship with humans and the environment. 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).

2.2. Identification Strategy

Some features of how Darwin made his work public enable us to identify the impact of his work on the broader cultural discourse. Although Darwin developed his theory over a long period, there is a precise time at which Darwin's theory reached the broader public, and this is 1859, the year of publication of *On the Origin of Species*.[6] This publication date was largely unplanned. Darwin proceeded slowly initially and had to deal with sickness and deaths in his family that further delayed him. However, eventually he "rushed" in order not to lose priority over Alfred R. Wallace, who was researching on the same topics and had sent Darwin some of his writings that developed similar concepts and reached similar conclusions about natural selection.

The book and Darwin's theory received almost immediate attention and diffusion, thanks to presentations at 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 (see for example Gray 1860; Huxley 1859).

The unplanned publication date of Darwin's theory provides the main source of variation for our empirical study. The rapid diffusion of the theory gives us an opportunity to observe the effect on the diffusion of the main concepts, and to establish which ones were especially novel and had an independent impact on the broader public discourse.

To be sure, *On the Origins of Species* was not the first treatment of evolution. Darwin's theory was novel in several ways and more coherent than previous ones, but earlier in the 19th Century some related ideas were already been elaborated and discussed; examples include the work of Lamarck, the anonymous *Vestiges of the Natural History of Creation* (later attributed to the Scottish journalist and publisher Robert Chambers), and of course the work of Alfred R. Wallace. Our empirical strategy, however, allows assessing whether the publication of Darwin's book represented a discontinuous change in the cultural discourse. In the analyses reported below we attempt to address the issue of whether the theory of evolution as Darwin presented it was already "in the air" with a variety of empirical exercises.

[6] 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 Dickens' *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 presence of those concepts in the public discourse *before* 1859.

3. Data and Methods

> *"The limits of my language mean the limits of my world."* Ludwig Wittgenstein, Tractatus Logico-Philosophicus (1922).

To examine the diffusion and the evolution of the meaning and interpretation of scientific concepts over time, we exploit the increasing availability of digitized text corpora, as well as the tools of natural language analysis. Our first step is to compute relative frequencies of some key words that embody the main concepts in Darwin's theory of evolution, and that 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 as an effective tool for the analysis of semantic and sentiment change (Aiden and Michel 2014; Manovich 2009; Michel et al. 2011; Roth 2014).

Word Frequencies. We first rely on Google N-Grams[7] (Lin et al. 2012) to assess how frequencies of words changed over time in fiction and non-fiction literature. The Google N-Grams data is the result of the Google Book project to build a vast collection of digitized books in partnership with major libraries.[8] First released in 2010, the data consist 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 there is a continuous update. The database includes different languages (besides English: Italian, French, German, Spanish, Russian, Hebrew, and Chinese). The English corpus alone has half a trillion words in it. For the period that we consider (i.e., 1820 to 1899), there are about 380,000 books containing more than 45 billion words in total. The data include both fiction and non-fiction books, but not periodicals, and is aggregated depending on the number of terms considered; for instance, the 1-ngram dataset includes single words and their frequency in a given corpus, and *n*-grams are combinations of *n* words and their frequency. We compute frequencies from 1-ngrams and 2-ngrams data for each year and express them in per-million-words terms.

The ability to separate fiction and non-fiction literature is relevant to us for two main reasons. First, one critique to the N-Grams (and Google Books) corpus is that it may over-represent scientific texts (Pechenick et al. 2015). In our study, increases in the frequency of words related to Darwin's theory may just reflect a disproportionate increase over time of the corpus of scientific

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

books (included in the non-fiction category). Second, separating fiction and non-fiction literature enables the analysis of different types of relationships between Darwinian science and broader culture. The use of Darwin's concepts in the non-fiction literature may better represent higher-educated or more erudite conversations. Conversely, 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 social imaginary (Armstrong 1987; Winans 1975).

Whilst the Google Books data allow us to capture Darwin's influence on the broader cultural discourse, we also aim to assess whether his work had an impact on the *political* discourse. We thus supplement our analyses using the digitized collection of the U.S. Congressional Records, which includes reports of all discussions occurring on the floor of the House and Senate (ProQuest's Congressional Record Permanent Digital Collection)[9]. The Congressional Records offer the opportunity to gauge the importance and diffusion of a given topic in the political arena; they and have been increasingly used by researchers in political science and economics (e.g., Gentzkow and Shapiro 2010).

Semantic Evolution and Sentiment Analysis: Word Embeddings

The analysis of word frequencies is informative, but does not provide insights about how a given word was used and its perception in society. The semantic changes and the evolution of attitudes toward a concept may be a more appropriate measure of cultural change if one interprets the meaning of a word as the association of that word with other concept and ideas, and the attitudes toward a concept as whether that concept had a positive or negative reception.

Natural language processing relies on word embedding techniques to determine the meaning of and sentiment toward words from large text corpora, and their evolution over time. The main idea of word embedding is that we can evaluate semantic associations between words by analyzing co-occurrence patterns in a text. Two words of similar meaning are unlikely to appear, say, in the same sentence, but they are likely to be surrounded by similar words. For example, we would not expect that within five words before and after the word "queen" (the "context words") we read the

[9] These includes Congressional Record (1873-1997), the Congressional Globe (1833-1873), the Register of Debates in Congress (1824-1837), and the Annals of Congress (1789-1824). It is worth noticing that before 1873 each House was only required to keep an internal journal of its proceedings. Only from 1874 onwards were external reporters allowed to witness debates and granted full permission to report them (McPherson 1942).

word “monarch”; however, there is plausibly high overlap between the words that appear before and after “queen”, and those that appear close to “monarch”.

The outcome of word embedding algorithms is a set of vectors that includes information about co-occurring patterns among words. Consider for example a text corpus with V words w (w=1, 2,…, V). For each word, one can specify a subset of context words that appear within a window of m words before and after w. The objective is to represent each word w as a Nx1 vector, with $N<V$ determined by the researcher, where each entry is a measure of how frequent is the occurrence of w with each of the context words. Taken together, the N-dimensional V vectors form the VxN embedding matrix.

In order to create word embeddings vectors, we adopt a Word2Vec approach (SkipGram with negative sampling; Mikolov et al. 2013), a technique that has been used extensively in literature interested in measuring semantic change (e.g., Garg et al. 2017, Bolukbasi et al. 2016; Caliskan et al. 2017). The Word2Vec model is based on a neural-network structure that we represented, in simplified form, in **Figure 1**. The starting point is the definition of Vx1 one-hot vectors for each focal word w, i.e. vectors of all 0’s except one value of 1 in correspondence of the word of interest. Two matrices, called the embedding and context matrices, are initially filled with random weights that the training process updates. For each word w, the algorithm multiplies a one-hot vector, or input layer, by the embedding matrix of VxN dimension, to obtain a Nx1 vector, called the hidden layer. This vector simply “copies” the input layer into the embedding matrix that corresponds to the word w. In turn, multiplying the hidden layer by the NxV context matrix produces the Vx1 output layer. The V entries (or scores) in the output layers go through a soft-max activation function, which maps the scores to a probability distribution. The probability vectors have values that range from 0 to 1 and sum up to 1.[10] These vectors can now be compared to the “target” one-hot-encoding vector of a given context word c to obtain a vector of errors by subtracting the probability vector from the “target” vector. Using this information, a backpropagation mechanism (Rumelhart, Hinton and Williams 1986a) updates the weights in the embedded and context matrix. The training process proceeds by considering all combinations of words w and context words c.

[10] The aim of the softmax function is to map scores into probability distributions as follows: $\mathrm{p}(c|w;\theta) = \frac{e^{v_c v_w}}{\sum_{c' \in C} e^{v'_c v_w}}$, where v_c and v_w are vector representations of context word c and focal word w respectively, and C is the set of all possible contexts. The estimation procedure thus consists of maximizing the probability that a given context word occurs within a given window around each focal word of interest.

The final output consists of updated scores in the *V*x*N* embedding matrix. Each row in the matrix is the embedded vector for each of the *V* words *w*, where each of the entries is a coordinate in an *N*-dimensional space, and carries information about the context.

These embedded 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 literature, we expect that, when the word vectors corresponding to *king, kings, queen, queens, man* and *woman*, the following holds: *(king – kings)* ≈ *(queen – queens)* and *(king – man)* ≈ *(queen – woman)*.

The closer two word vectors are in the *N*-dimensional space, the closer the semantic association between the two words.[11] The main metric of the association or proximity between vectors is the cosine between them (Dubossarsky et al. 2015; Gulordava and Baroni 2011; Jatowt and Duh 2014; Kim et al. 2014; Kulkarni et al. 2015). Call γ the angle between two *N*-dimensional vectors $u = (u_1, \ldots u_N)$ and $v = (v_1, \ldots v_N)$. 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 investigate whether the words that defined the main Darwinian concepts shared context words with different terms before and after the publication of *On the Origins of Species*. We rely on previously trained Word2Vec embeddings resulting from the n-grams distributed by Google Books (Hamilton et al. 2016).[12] Figures are available for every decade between 1800 and 1990 and data are designed to enable comparisons across decades. The models use a context window of four and parameters as suggested by Levy et al. (2015) to measure semantic changes in cultural shifts.[13]

Sentiment analysis

The measure of semantic similarity (cosine distance) includes all *N* dimensions. Each dimension explains some of the variance that distinguishes association patterns among all the words in a text corpus. Generally, however, it is hard to interpret each dimension. One might therefore consider projecting these vectors on a limited subset of pre-specified dimensions and evaluate the position

[11] 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).

[12] We set *N*=300 for all decades, whereas V is specific to each 10-year period.

[13] See Hamilton et al. (2016) also for a discussion on the pre-processing methods and parameters used.

of each vector on the new space, to measure the association of a given word within a set of well-defined underlying concepts. Kozlowski et al. (2019), for instance, project vectors related to different musical genres (e.g., jazz, rap, etc.) onto a plan that measures "affluence", to determine whether, say, jazz is more associated with wealthier strata of a population than rap, and how these associations vary over time. Following this approach, we specified some dimensions that might be relevant for gauging the sentiments surrounding key Darwinian concepts over time. We focus in particular on "goodness", "importance", and "morality". In order to build these dimensions, we first identify pairs of antonym words that are closely related to what we aim to capture. We then average the difference of all the vector pairs: $\frac{\sum_{P}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$, where $\overrightarrow{p_1}$ and $\overrightarrow{p_2}$ are the antonym vectors of a given pair belonging to the set of relevant pairs P. Our dimensions are based on the work of Jenkins et al (1958), who specify a list of terms for a variety of cultural dimensions. Once we identify our dimensions, we proceed by calculating the orthogonal projection of the vectors of key Darwinian words on each dimension and track their position over time.

4. Frequency Analysis

In this section, we analyze the evolution of the relative frequency of key words in the Darwin's theory and two-word expressions as measures of the diffusion of key concepts in the public discourse around the time of the publication of *On the Origin of Species* in 1859.

4.1 Darwinian and "Control" Concepts; Fiction and Non-fiction Books

Graphical Analysis. We consider terms (1-grams) that, from many accounts (Desmond and Moore 1994, and Mayr 1995), as well as our own reading, represent the key concepts in Darwin's theory: Evolution, Selection, Adaptation, Competition, Survival, and the expression (2-gram) Natural Selection. **Figure 2** reports their frequency of use, per million words, in each year between 1820 and 1899, separately in fiction and non-fiction books.[14] We scale the y-axes differently for the two categories in each graph.

The expression Natural Selection, perhaps the most defining of Darwin's concepts, was virtually non-existent in both the fiction and non-fiction literature before 1859 and experienced a

[14] We initially included also the word Mutation, but then opted to discard it its occurrence was too low throughout the period of interest to allow for meaningful analyses

significant increase in use since then. On the one hand, this may not be surprising, precisely because of the close association of Darwin with the idea of natural selection. On the other hand, we may consider the significant increase in the diffusion of this concept immediately after the publication of *On the Origin of Species* as a validation of our approach; this initial analysis of frequencies appears to capture what we might have expected.

Moving to other Darwinian concepts, also Evolution and Survival entered the public discourse in the years immediately following the publication of *On the Origin of Species*. The ideas that underlie these words and expressions, therefore, generated interest not only in specialized or more educated circles, but plausibly also in the more general cultural context. Interestingly, the diffusion of these concepts in the fiction literature lagged the diffusion in the non-fiction literature by a few years. Competition was already present in the first part of the 19th Century, especially in the non-fiction literature, and experienced an increase 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*; Adaptation reached a stable relative frequency in the 1840s, whereas the frequency of Selection was constantly increasing since the early 19th Century. Note that Selection was already increasing its presence in the cultural discourse before 1859, whereas Natural Selection appeared after the publication of *On the Origins of Species*. This suggests the possibility that, after 1859, the word Selection might have experienced a change of meaning, perception and use in the public discourse. We will investigate this below.

In **Figure 3**, we display the relative occurrence of some terms that were of frequent use in general and in the sciences, are not specific to Darwin's theory, and appear very frequently 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 in the figure are Number, Life, Animals, Flowers, Plants and Nature. For none of these words is there any discernible change in diffusion in the decades immediately preceding and following the publication of *On the Origin of Species*. These "generic" words are a subsample of the full set of nouns whose frequency we use as "control" in the regression analyses that we describe below.

Regression Results: Word-by-Word Time Series. Table 1 reports estimates from spline regressions of the yearly relative frequency of use (per million words) of each of the Darwinian words and phrases, as well as of the subsample of six generic words that we represented in the

graphs above, on time t=20, 21, …, 99, with a knot at t=59 corresponding to year 1859. The reported values are the estimated slopes of the diffusion curves in the two sub-periods. **Table 2** displays the estimates from spline regressions where the outcome is the natural logarithm of the frequency per million words (+0.01), and the knot is at the natural logarithm of t=59. In these log-transformed models, the estimates thus represent elasticities, and it is therefore easier to make comparisons across words. **Table 3** reports estimates from the same spline regressions in natural log terms, separately for fiction and non-fiction books, and limited to the Darwinian concepts. Finally, **Table 4** reports spline regressions for the Darwinian words, separately for fiction and non-fiction books, with knots at each decade between 1820 and 1899 (20-29, 30-39, …, 90-99), in natural logarithm.

The estimates, both in absolute and relative terms, corroborate the visual evidence from **Figure 2**. For the Darwinian terms discussed above, the increase in slope after 1859 is statistically significant and especially large for Natural Selection, Evolution, Survival and Competition. We do not detect any specific pattern related to the publication of *On the Origins of Species* for the six generic words. We ran the same spline regressions for the full set of the 99 most frequent nouns in *On the Origins of Species* (one of the originally selected 100 words is Selection, which we exclude from the controls; the list of words is in Appendix **Table A1**). Appendix **Figure A1** reports the estimates of the yearly frequency slopes in the 1820-59 and 1860-99 periods. The estimates are split between negative and positive; **Figure A2** shows that the estimates of the slopes in the two sub-periods for a given word are very similar to each other.[15]

Regression Results: Differences in Differences. We use the full set of control words to perform differences-in-differences analyses where we estimate the aggregate diffusion patterns of Darwinian and generic words before and after 1859. We perform these analyses in two ways.

First, in **Table 5** we report the estimates from analyses where, for each year, we sum up the frequencies of the six Darwinian concepts on the one hand and of the ninety-nine control nouns on the other hand, and compare the trends in this aggregate diffusion before and after 1859. Because the aggregate frequency of the generic words is much higher than the frequency of the Darwinian

[15] Appendix Table A3, where we report estimates with specific slopes and steps for each decade between 1820 and 1899 instead of only one "cut" in 1859, also confirms the delay in diffusion in the fiction literature that we observed in the graphical representations above.

concepts pooled together, 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. In this analysis, we also pool together fiction and non-fiction books. The regression model that we estimate is as follows:

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

The data thus include 160 observations, two for each year, with one reporting information about the generic words ($\mathbf{1}(Darwinian) = 0$), and the other about the six Darwinian concepts ($\mathbf{1}(Darwinian) = 1$). Columns 1 and 2 of **Table 3** display estimates of a simplified version of the model, were the left-hand-side variable is the natural logarithm of the sum of frequencies of Darwinian and generic terms separately, regressed on a time trend and the interaction between the indicator for years greater than 1859 and the difference between the current year and 1859. Estimates of the parameters of the full model 3 are in column 3. The estimate on the coefficient on the interaction between the indicator for Darwinian words, the indicator for the post-1859 period and the difference between the current year and 1859 (μ_w) is positive, large and statistically significant, indicating a much larger relative increase in the frequency of Darwinian concepts after 1859. The estimate of θ_w is significantly smaller than the estimate of μ_w, but it is positive and statistically different from zero; this indicates that also before 1859, the frequency of Darwinian concepts was increasing at a higher rate that the combined generic terms. This is likely due to the trend and diffusion that some Darwinian terms, such as Selection and Adaptation, were experiencing also in the first half of the 19th Century.[16] The trend, however, clearly had an additional, fast acceleration after the publication of *On the Origins of Species*.

Second, we consider a model where the outcome variable is the annual frequency (from 1820 to 1899) of each of the six Darwinian concepts and ninety-nine control nouns separately, and we estimate the average difference in frequency for the Darwinian words and the generic words per each decade:

[16] If, for example, we exclude Adaptation and Selection from computing the aggregate frequency of the Darwinian concept, the estimate of θ_w declines from 0.36 to 0.08, whereas the estimate of μ_w increases from 1.74 to 3.08.

$$\ln(y_{wt}) = \alpha_w + \beta_w \mathbf{1}(Darwinian) + \sum_{j=2}^{4} \gamma_i \mathbf{1}(j0 \leq t \leq j9) + \sum_{j=6}^{9} \gamma_j \mathbf{1}(j0 \leq t \leq j9) + \sum_{i=2}^{4} \delta_i \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian) + \sum_{j=6}^{9} \delta_j \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian) + \varepsilon_{wt} \quad (2)$$

This analysis is therefore based on (6+99)*80=8,400 observations. **Figure 4** displays the estimates of the δ_j coefficients and their 95% confidence intervals. The omitted time category is the decade 1850-59 ($50 \leq t \leq 59$). This analysis provides further evidence of the different patterns of diffusion of the Darwinian words immediately following 1859, compared to statistically insignificant differences before the publication of *On the Origins of Species*.

4.2 Translation in Other Languages

Were the effects of *On the Origin of Species* specific to the context in which the book was written and first published? Or did the treatise generate a similar impact in other countries upon its translation? Moreover, did the diffusion of scientific concepts in the cultural environment depend on the status of a country economic development? To answer these questions, we study whether the translations of *On the Origin of Species* generated a similar the diffusion of its key concepts in other languages.

As shown in **Figure 5,** the phrase Natural Selection dramatically increased its diffusion upon the translation of *On the Origin of Species*. The same holds for such words as Evolution, Survival, and Competition (**Figure A3** in the Appendix). Moreover, the diffusion of most words started increasing right after 1859, indicating that, even in the absence of an official translation, Darwin's concepts diffused across borders. These results suggest that the cultural effects of *The Origin of the Species* were not specific to the English-speaking context.

An important caveat in interpreting these results is that we cannot claim the translation year was exogenous. For example, the translation might have occurred first in countries were the interest was higher, and this, in turn, might have affected its diffusion. The likely endogeneity of the publication year, on the other hand, offers the opportunity for additional discussions on the relationship between the broad cultural acceptance of scientific concepts and economic development. For instance, in countries like Italy and Spain, both "late comers" during the Industrial Revolution (Ciccarelli and Nuvolari, 2015), the translation of *On the Origin of the Species* occurred latter than the translation into French and German, i.e., the languages of two

countries where industrialization occurred earlier. Similarly, Russia, which was mostly a feudal country until World War I (Markevich and Zhuravskaya, 2018), had the first translation of Darwin's book even later. In fact, the diffusion of Darwinian words and phrases was extremely limited in Russian books in the 19th Century. These terms, finally, were virtually nonexistent in Chinese books. This is consistent with Mokyr's (2008) idea of Chinese isolation in the scientific debate, and in turn of its delayed industrial development.

4.3 Ideas in the Air and Multiple Attribution

The various findings that we just reported show that some concepts, as measured by the words that embody them, were only marginally part of the public discourse before the publication of *On the Origin of Species*. But even if some words entered the public discourse only after 1859, can we attribute this only to the work of Darwin? In other words, would this diffusion have occurred starting in the 1860s, irrespective of Darwin's contribution? In **Figure 6** we report the frequency of occurrence of the names of four scientists who contributed, in different ways, to our understanding of evolution. In addition to Darwin, we consider Alfred Russell Wallace, Robert Chambers, and Jean-Baptiste Lamarck. Lamarck's theory of the transmission of acquired traits is frequently mentioned as an example of "failed" theory to compare to Darwin's. Chambers' *Vestiges of the Natural History of Creation* introduced, in the 1840s, the idea of an "evolution" of living and non-living beings over time, more as a speculation than as a scientific treatment (note that the author was anonymous until 1884). Alfred Russell Wallace's work was much closer, in time and content, to Darwin's. **Figure 6** shows that both Darwin and Wallace increased their occurrence in the English book corpus in the second half of the 19th Century, but Darwin's frequency increased substantially more. Chambers and Lamarck were already present before then, but their frequency remained stable (and fairly low) after 1860.[17]

In **Figure 7** we report additional tests. Because Lamarck was (and originally wrote in) French, we compare the diffusion of the words Darwin and Lamarck in the French corpus. After 1860, the relative occurrence of the word Darwin in French books surpassed the frequency of Lamarck. We also compare terms that related to the study of the emergence and development of new species:

[17] We add the following combinations of names, middle names and last names: Alfred Russel Wallace, Alfred Wallace, Charles Darwin, Charles Robert Darwin, Robert Chambers, Jean-Baptiste Lamarck, Jean-Baptiste de Lamarck, Jean Baptiste Lamarck, Jean Baptiste de Lamarck.

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 by Lamarck. The general pattern is that Evolution became progressively more frequent than Transmutation, with a significant change in frequency after the 1950s.

Overall, this evidence suggests that Darwin, with his own work and especially his 1859 book, caused a discontinuous change in the cultural discourse.

4.4 The Diffusion of Darwinian Concepts in the Political Debate

We perform frequency analyses on U.S. Congress data to assess whether Darwin's theory spilled over not only to the cultural discourse, but also to the political debate; arguably, this is a condition for science-driven cultural change to affect outcome of social and economic relevance. The corpus of Congressional Records includes the transcripts of all legislative debates occurring on the floor of Congress. It also contains additional materials, such as communications from the president and the executive branch agencies memorials, petitions, and supplementary information on the current legislation. As such, it represents the official and most comprehensive daily account of the political discussion happening in the House and the Senate.

The findings, displayed in **Figure 8**, indicate an increase of the frequencies of such words and concept as Evolution, Survival and Natural Selection in the Congress data after 1859. Spline regression analyses confirm that the words more related to Darwin's theory were more likely to enter the political debate, compared to generic words. The increase becomes significant about two decades after Darwin's publication (estimates are in Appendix **Table A4**).

Overall, these results suggest that, after diffusing in the cultural environment, key Darwin's concepts also reached the political debate. The lag with which this happened, however, seems to indicate that the cultural diffusion was faster than, and perhaps a pre-condition for political diffusion. A limit of this analysis is that we cannot directly estimate the extent to which Darwin's theory affected policy decision making or economic impacts. However, that fact that the theory reached the political is a necessary condition for it affecting policies as well.

5. Semantic and Sentiment Analysis

Word embedding techniques require very large sample sizes to produce reliable results and insights. For this reason, in this section we limit the analysis to the Google Book database, and aggregate the data at the decade level.

5.1 Semantic Analysis

Figure 9 introduces the second part of our study, where we move from the analysis of the frequency of use of certain words and the concepts underlying them, to the analysis of the semantic evolution of certain words and concepts, to see whether this evolution occurred in ways that we can relate to Darwin's theory. In the graphs, the horizontal axis reports decades (the time unit of reference), and the vertical axis indicates the cosine between the two word vectors of interest.

One aspect of Darwin's theory is that life (or existence) includes adaptation, as well as competition, among its defining aspects. We 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, characterized by a more tumultuous view of nature, seem therefore to have captured a more general trend. Greater cosine similarity 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.

In **Figure 10** we assess the semantic association between the five key words in On the Origins of Species that we took as expressing Darwin' contribution (Evolution, Selection, Survival, Competition and Adaptation), and the names of the four scientists (including Darwin) we considered in Section 4.3 above. With this exercise, we aim to explore whether these key terms that defined the theory of evolution by natural selection were, in fact, specifically associated with Darwin or were part of a discourse that included also the contribution of other scientists. In general, the similarity of these words with Darwin is systematically positive and greater than the similarity with the other names. Lamarck generally shows higher similarity with the five key words than

Chambers and Wallace. This suggest that Darwin and Lamarck remained the two most prominent figures, among students of evolution, in the cultural discourse.

A second analysis of semantic changes focuses, again, on the key words and concepts that we considered so far. However, 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 (cosine similarity) to these key words. **Figure 11** (A through E) reports the findings. 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 horizontal (time) axis in the figures, the closer to one the cosine similarity.

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 and Reproduction).

Second, substantial 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.

Third, Selection was 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 broader 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.

5.2 Sentiment analysis

Figure 12 (A through C) displays the evolution over time of the perceptions or sentiments about the key Darwinian concepts in English books, as well as about Darwin himself. We focus on the proximity to three categories of antonyms: Unimportant vs. Important, Bad vs. Good, and Immoral vs. Moral. These dichotomies help assessing whether Darwin's concepts gained relevance and had a positive or negative connotation in the public discourse.

Although the evidence is not clear-cut, the term Evolution is, especially after 1859, perceived as more important, moral, and good, and so is Competition. Therefore, these two key concepts in Darwin's theory not only experienced an increase in use and evolution of their meaning (especially Evolution, as described in Section 5.1), but also were received positively. The term Darwin also shows positive reception, with spikes around the publication of *On the Origins of Species*.

6. Conclusions

To the extent that both cultural and scientific change are major drivers of long-term economic outcomes, the investigation of how these two phenomena interact with each other promises to offer a deeper understanding of their role in enhancing growth.

We focused on one of the major scientific breakthroughs, the theory of evolution via natural selection of Charles Darwin, and explored its impact on the public discourse. Given the undoubted importance of Darwin's theory, there is a diffused perception that it affected culture in many different ways, from changing the interpretation of nature to influencing ideas about race and equality among humans. Existing accounts, however, largely rest on qualitative or narrative evidence limited to scientists or cultural elites in society, whereas little is known about the wide diffusion of Darwin's ideas into society. Arguably, to affect cultural change (to be, in the terminology of Mokyr [2013, 2016], a cultural entrepreneur), a scientist should have an impact on the imaginary of a broader population. Moreover, it is difficult to identify, from existing accounts, which Darwinian concepts were actually novel in the cultural discourse, and which ones were already part of it. We address these challenges by analyzing the diffusion and the semantic evolution of the key words and phrases that embody Darwin's main concept in hundreds of thousands of books, with the use of techniques from machine learning. We rely on the largely unplanned publication date of *On the Origin of Species* as source of natural variation, and compare the use of these words and phrases with more generic terms that Darwin used.

Our analysis shows that the key concepts expressed by Evolution, Survival, Competition and Natural Selection diffused in fiction and non-fiction literature immediately after the publication of *On the Origins of Species*. Competition, a theme already present in the broader literature, diffused significantly more rapidly after 1859. Other key concepts such as Selection and Adaptation were already gaining relevance in the cultural discourse before 1859. The adoption of some of these words and phrases in the broader cultural conversation led also to a change in the meaning of the concepts, providing further evidence of the impact of Darwin's theory in society at large; overall, the attitude toward these concepts was positive rather than adversarial.

Our approach has several inductive and descriptive aspects. The choice of the concepts on which to focus may seem somewhat arbitrary; however, we based our selection on the main topics that Darwin developed, 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. 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 reception are difficult to make. Empirical approaches enabled by machine learning techniques provide promising tools to explore this relationship beyond the specific historical episode on which we focus. Examples of relevant scientific breakthroughs 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.

Aiden, E. and Michel, J.B. (2014). *Uncharted: Big data as a lens on human culture*. Penguin.

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.

Ayala, F. J. (2009). Darwin and the scientific method. *Proceedings of the National Academy of Sciences*, *106*, 10033-10039.

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., and 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., and 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.

Cohen, M. (2002). *The sentimental education of the novel*. Princeton University Press.

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., and 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.

Fuller, R. (2017). *The Book that Changed America: How Darwin's Theory of Evolution Ignited a Nation*. Penguin.

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.

Gentzkow, M and Shapiro, J. "What drives media slant? Evidence from US daily newspapers." Econometrica 78.1 (2010): 35-71.

Gianquitto, T., and 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*.

Gray, A. (1860). Darwin on the Origin of Species. *The Atlantic*, July issue.

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., and 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.

Huxley, T. (1859). Darwin on the origins of species. *The Times*, 26 December: 8-9.

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

Jelveh, Z., Kogut, B., and Naidu, S. (2014). Detecting latent ideology in expert text: Evidence from academic papers in economics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1804-1809.

Jenkins, J., Russell, W, and Suci, G. (1958). An Atlas of Semantic Profiles for 360 Words. American Journal of Psychology 71(4):688–99

Kearney, C., and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis,* 33, 171-185.

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.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5): 905-949.

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., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Trans. ACL, 3.*

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*: 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. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.

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

Mayr, E. (2001). The philosophical foundations of Darwinism. Proceedings of the American Philosophical Society, 145(4), 488-495.

McPherson, E. G. (1942). Reporting the debates of congress.

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*: 3111-3119.

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.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

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.

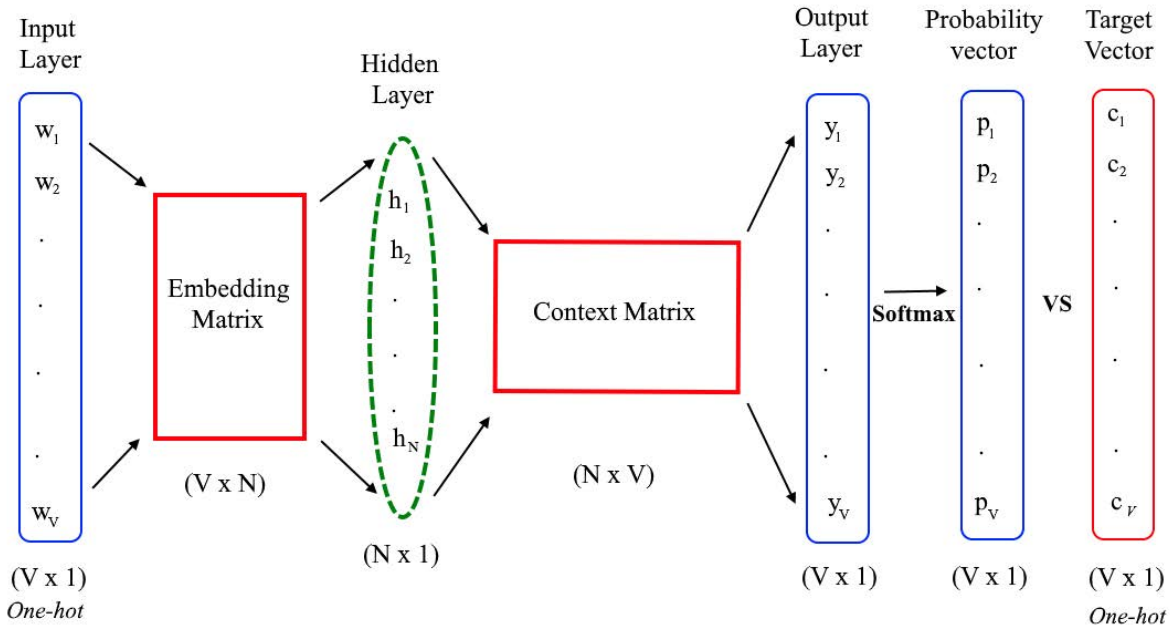
Stephan, P. E. (2012). *How economics shapes science* (Vol. 1). Cambridge, MA: Harvard University Press.

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.

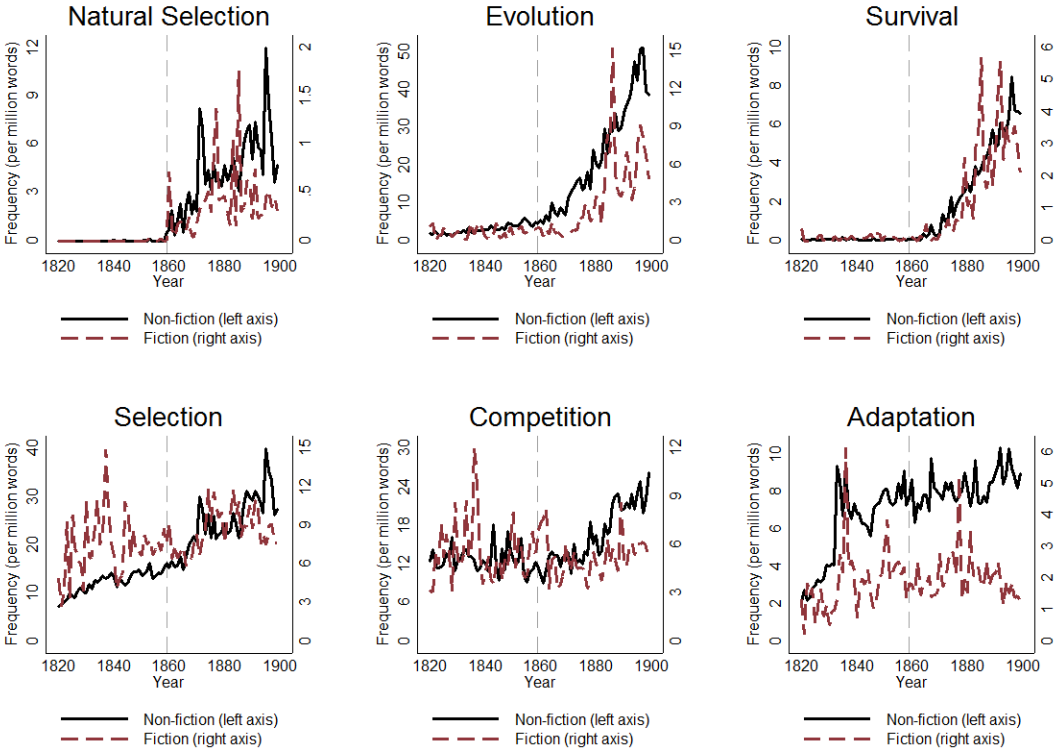
Winans, R. B. (1975). The Growth of a Novel-Reading Public in Late-Eighteenth-Century America. Early *American Literature*, 9(3), 267-275.

Figure 1: Word2Vec model



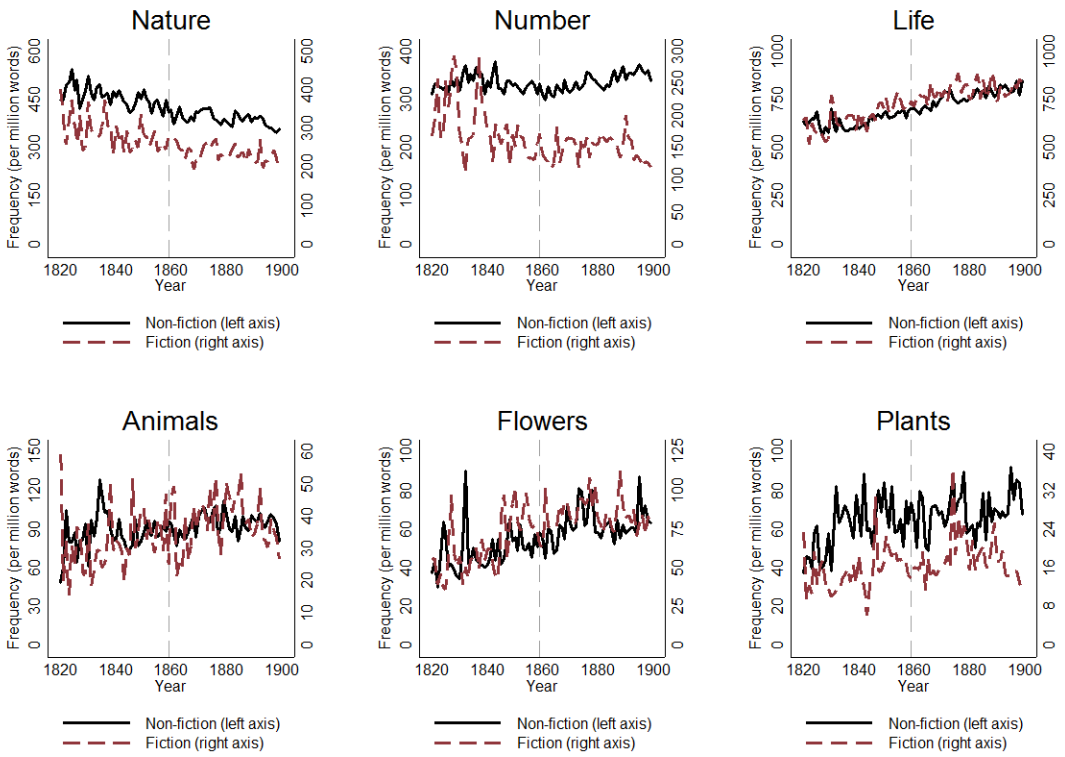
Notes: The diagram illustrates the structure of a Word2Vec model. Each word is encoded into binary vectors (one-hot) of dimension Vx1. The embedding matrix (VxN) and the context matrix (NxV) are initialized with random weights (note that $N<V$). The multiplication of the initial one-hot vector and the embedding matrix gives us the embedding vector of the input word we are currently considering. This embedding vector forms a hidden layer of dimension Nx1. The multiplication of the hidden layer and the context matrix forms the output vector, which becomes a probability vector after a soft-max transformation. This vector can be readily compared to the one-hot vector that identifies the considered context word (i.e., target vector). The difference between the probability and the target vector modifies the scores of the embedding and context matrix through a backpropagation mechanism so that the weight can be adjusted accordingly to real words co-occurrence.

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



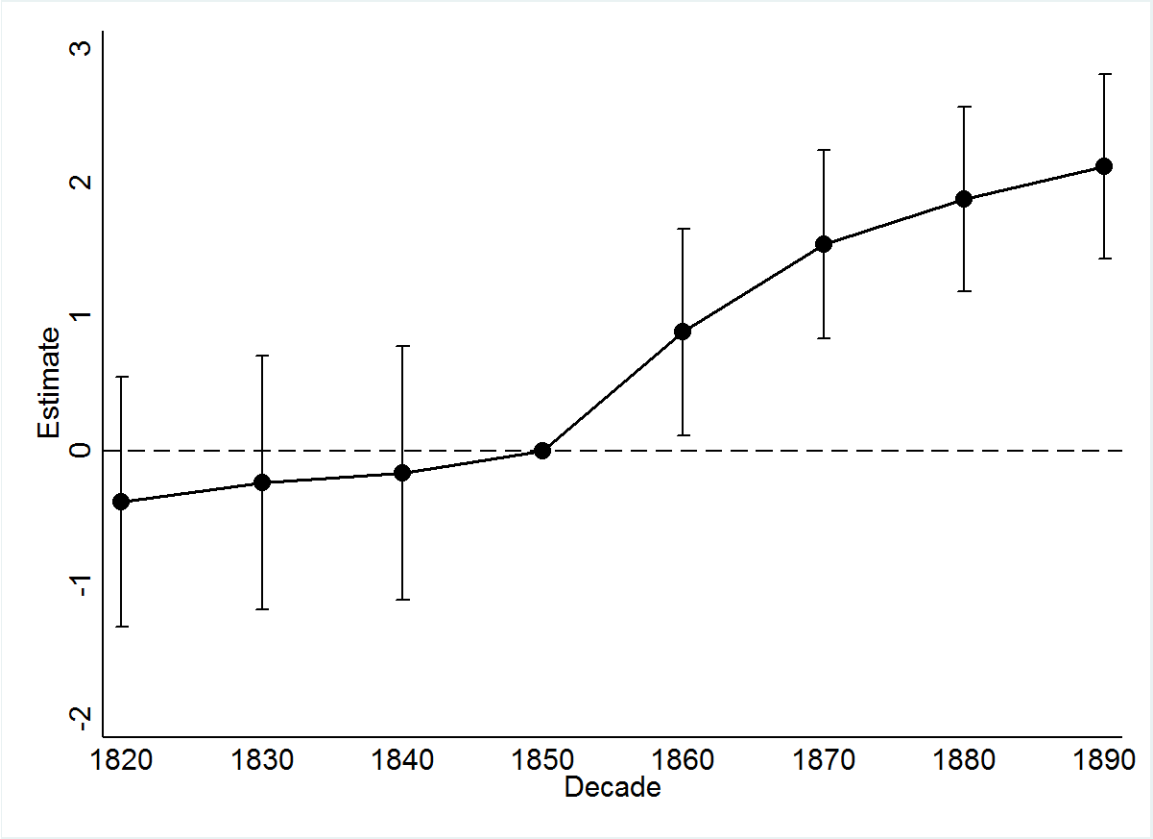
Notes: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right shows the scale for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

Figure 3: Frequencies (per 1 Million Words) of Select "Generic" Words in the Google Books Corpora



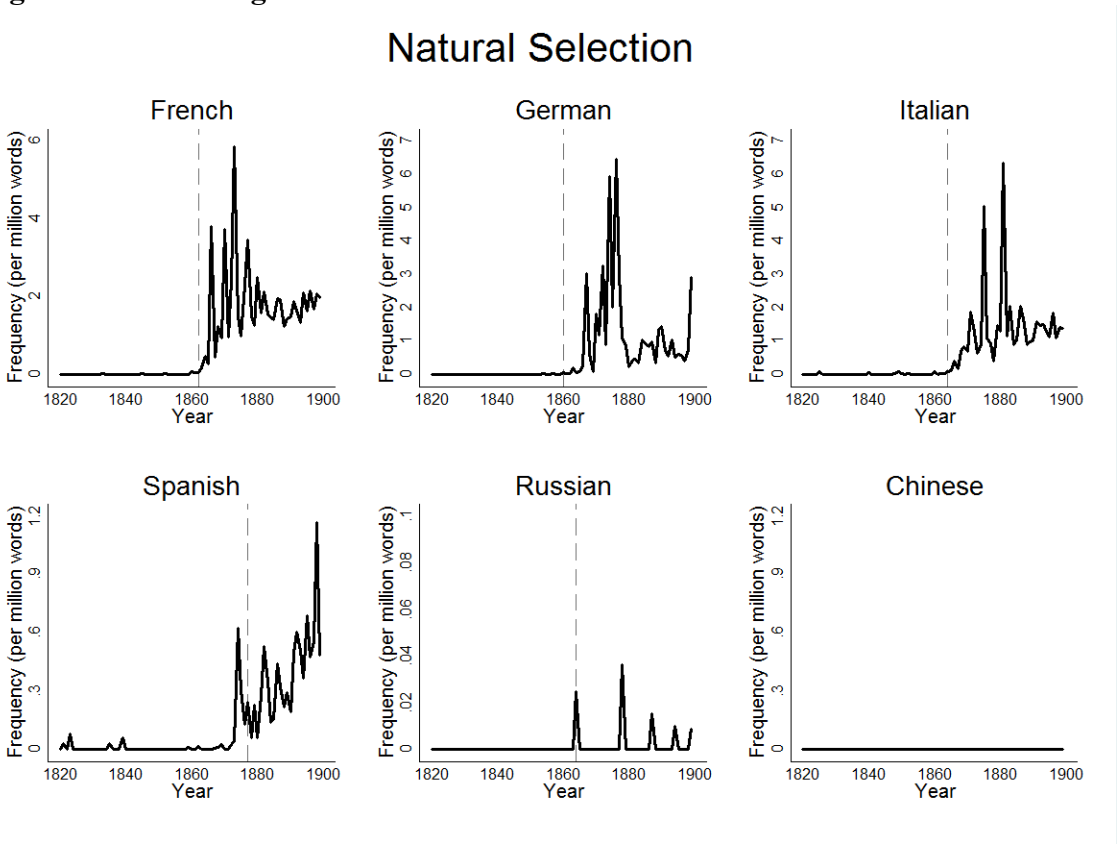
Notes: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words, separately for fiction and nonfiction texts. The y-axis on the left of each graph reports the reference scale for nonfiction, whereas the y-axis on the right shows the scale for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

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



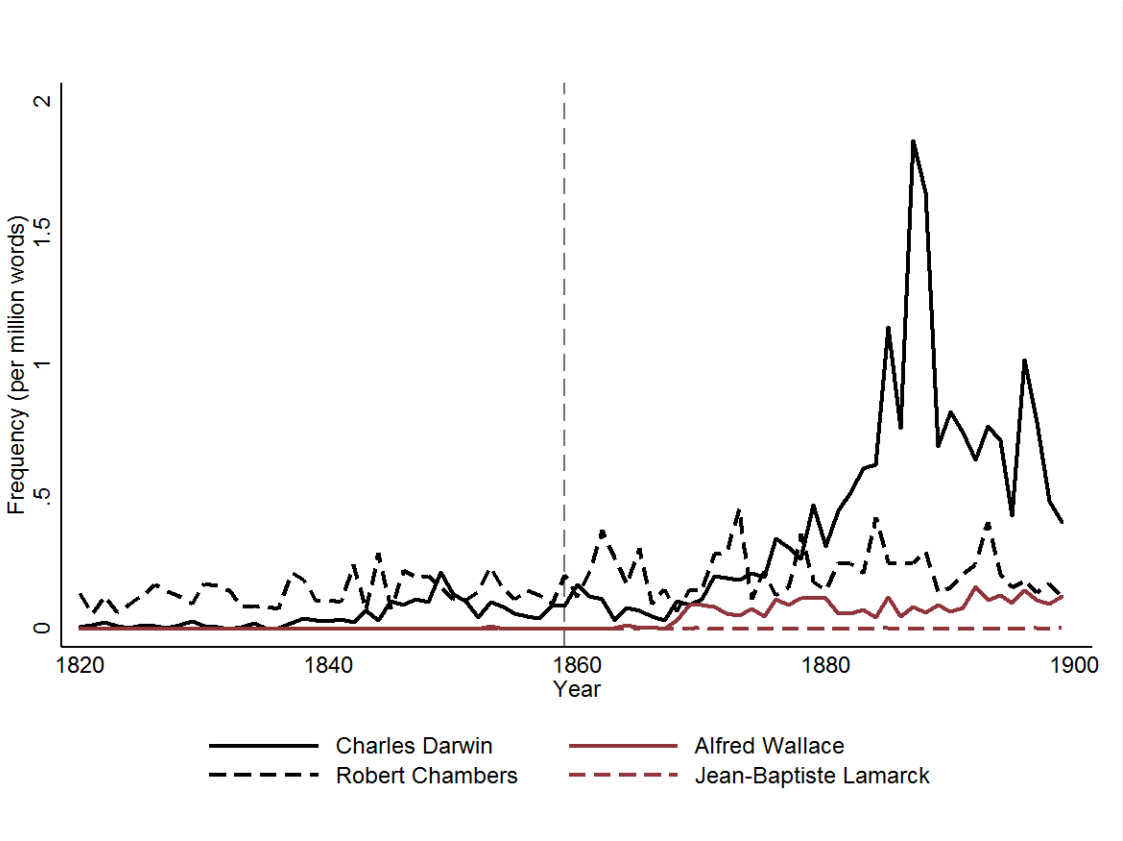
Notes: Each dot in the graph represents the estimate of the parameters δ_j from the following regression model: $\ln(y_{wt}) = \alpha_w + \beta_w \mathbf{1}(Darwinian) + \sum_{j=2}^{4} \gamma_i \mathbf{1}(j0 \leq t \leq j9) + \sum_{j=6}^{9} \gamma_j \mathbf{1}(j0 \leq t \leq j9) + \sum_{i=2}^{4} \delta_i \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian) + \sum_{j=6}^{9} \delta_j \mathbf{1}(j0 \leq t \leq j9) * \mathbf{1}(Darwinian) + \varepsilon_{wt}$, where y_{wt} is the frequency of use of a word per million words used (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, 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

Figure 5: Frequencies (per 1 Million Words) of the phrase "Natural Selection" in six languages other than English



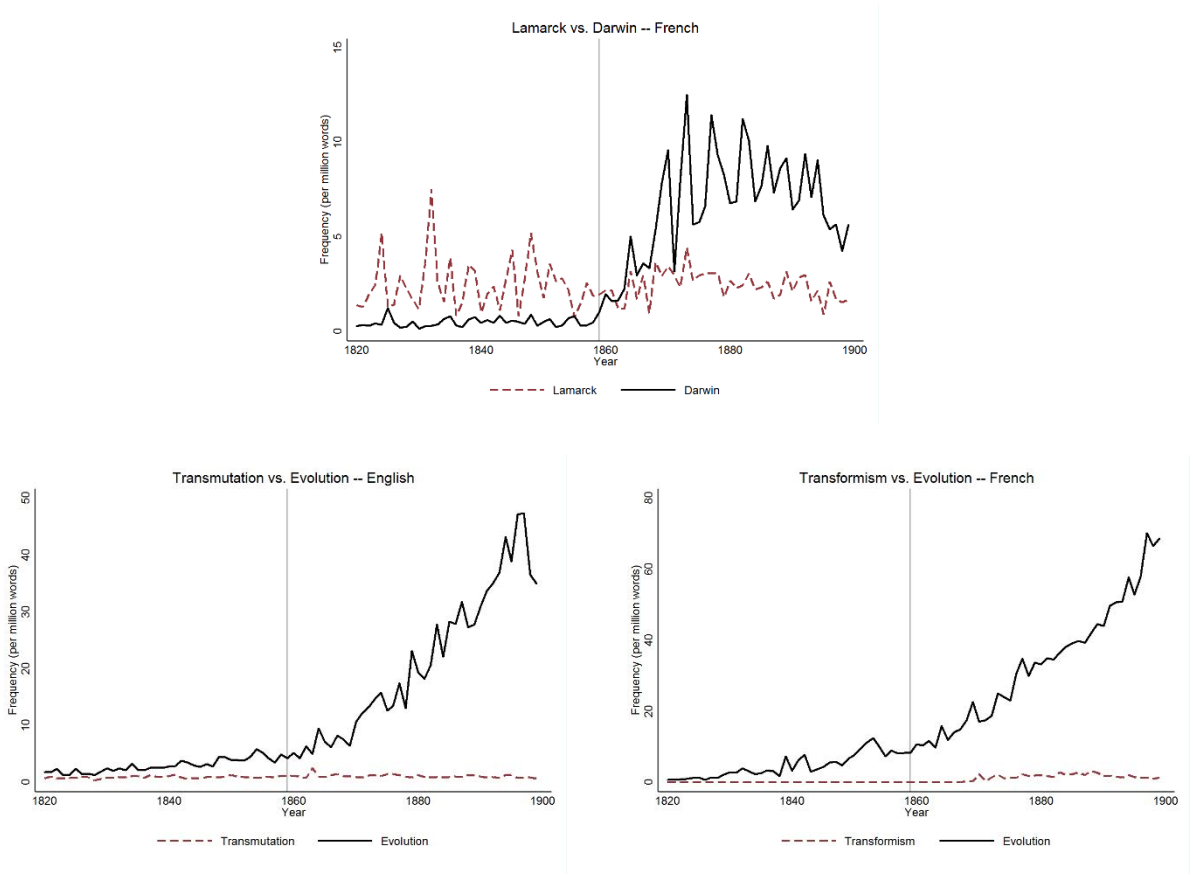
Notes: For each year, the figures report the number of occurrences (per million words) of the phrase "Natural Selection" in the language indicated on top of a graph; the vertical dashed line are in correspondence of the year of the first published translation of *On the Origin of Species* in a given language.

Figure 6: Frequencies (per 1 Million Words) of Occurrences of the names Charles Darwin, Alfred Wallace, Robert Chambers and Jean-Baptiste Lamarck in the English Google Books Corpus



Notes: For each year, the figures report the number of occurrences (per million words) of the name indicated in the legend. We include different combinations of the full names of the four scientists: Alfred Russel Wallace, Alfred Wallace, Charles Darwin, Charles Robert Darwin, Robert Chambers, Jean-Baptiste Lamarck, Jean-Baptiste de Lamarck, Jean Baptiste Lamarck, Jean Baptiste de Lamarck

Figure 7: Frequency of Occurrence (per million words) of the Words Darwin, Lamarck, Transmutation, Transformism and Evolution in the English and French Google Book Corpora



Notes: For each year, the figures report the number of occurrences (per million words) of the name indicated in the legend.

Figure 8: Frequencies (per 1 Million Words) of Selected Darwinian Words and Phrases in the U.S. Congressional Records

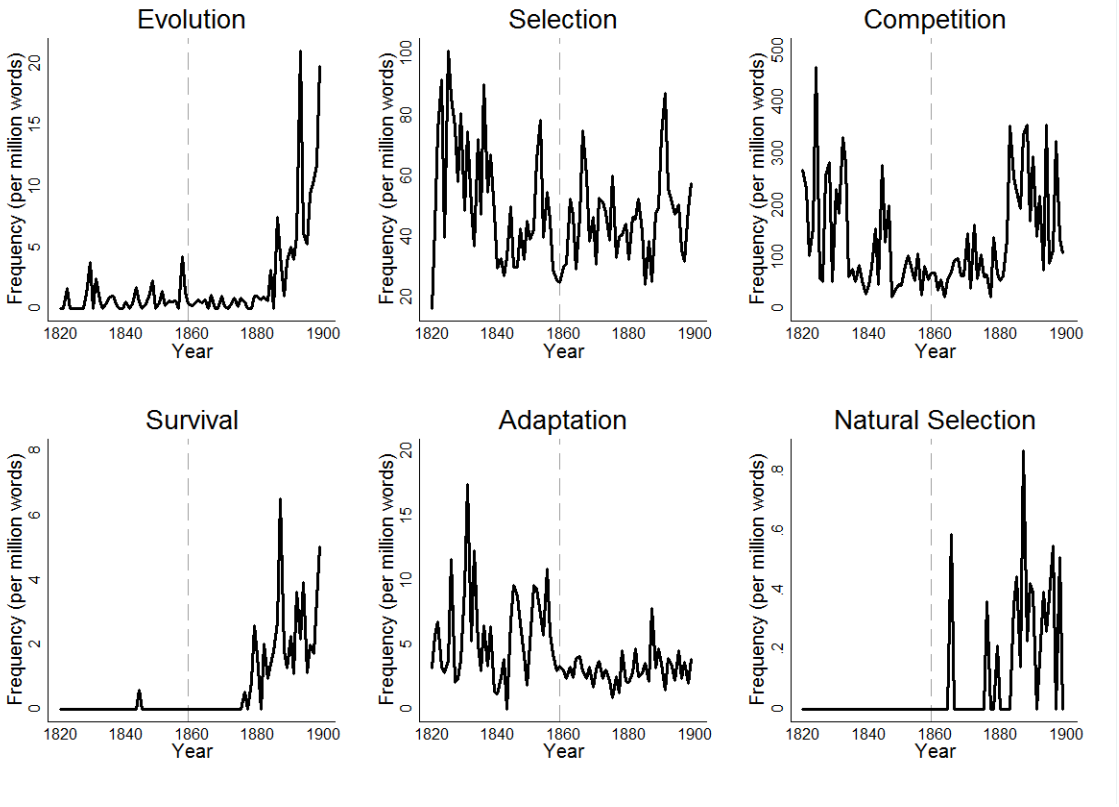
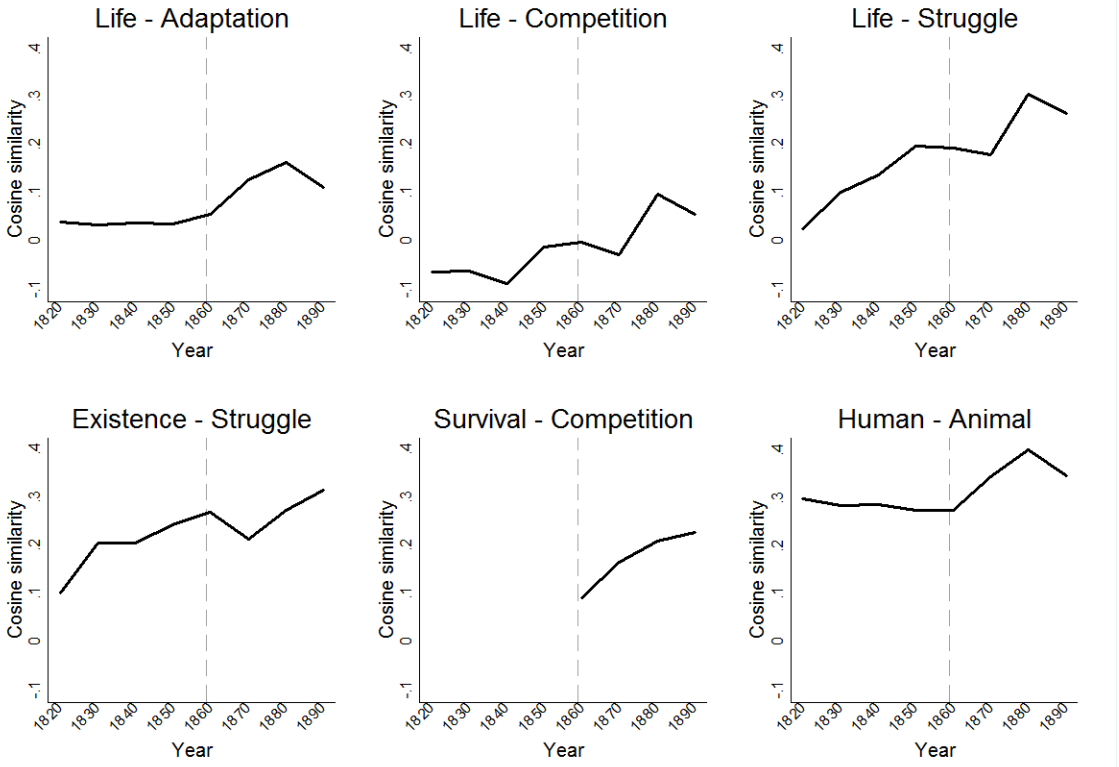
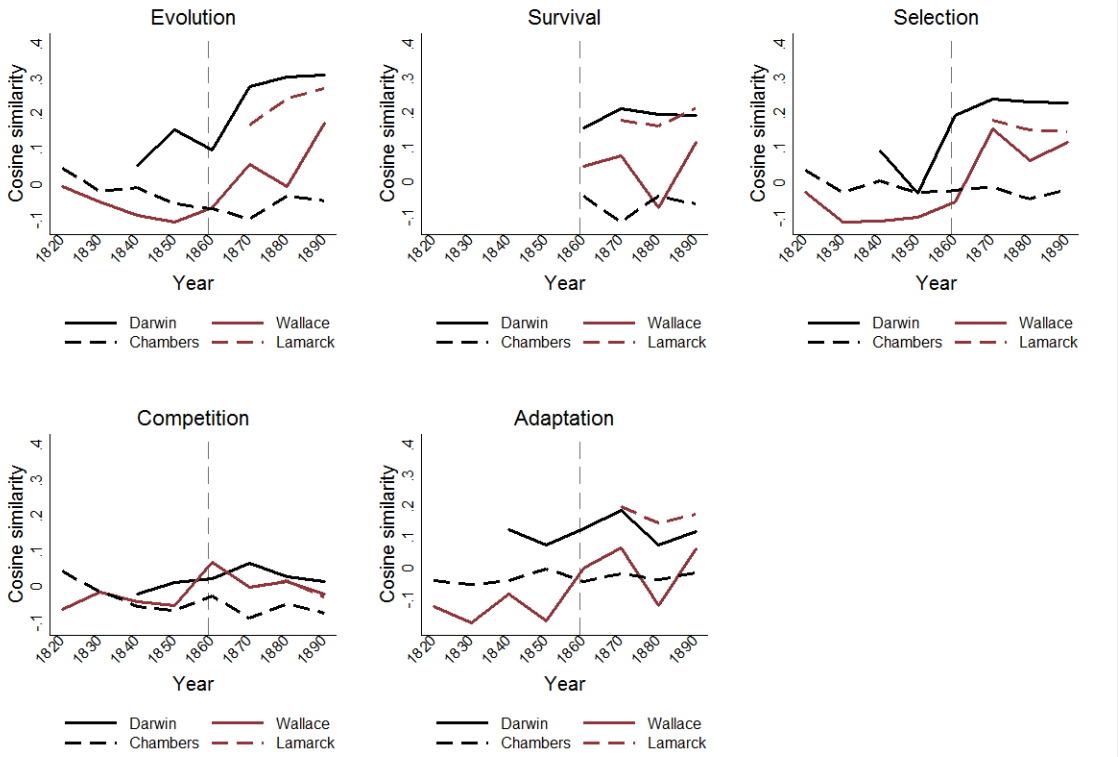


Figure 9: Semantic Associations between Selected Pairs of Words



Notes: The graphs 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. On the x-axis, 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

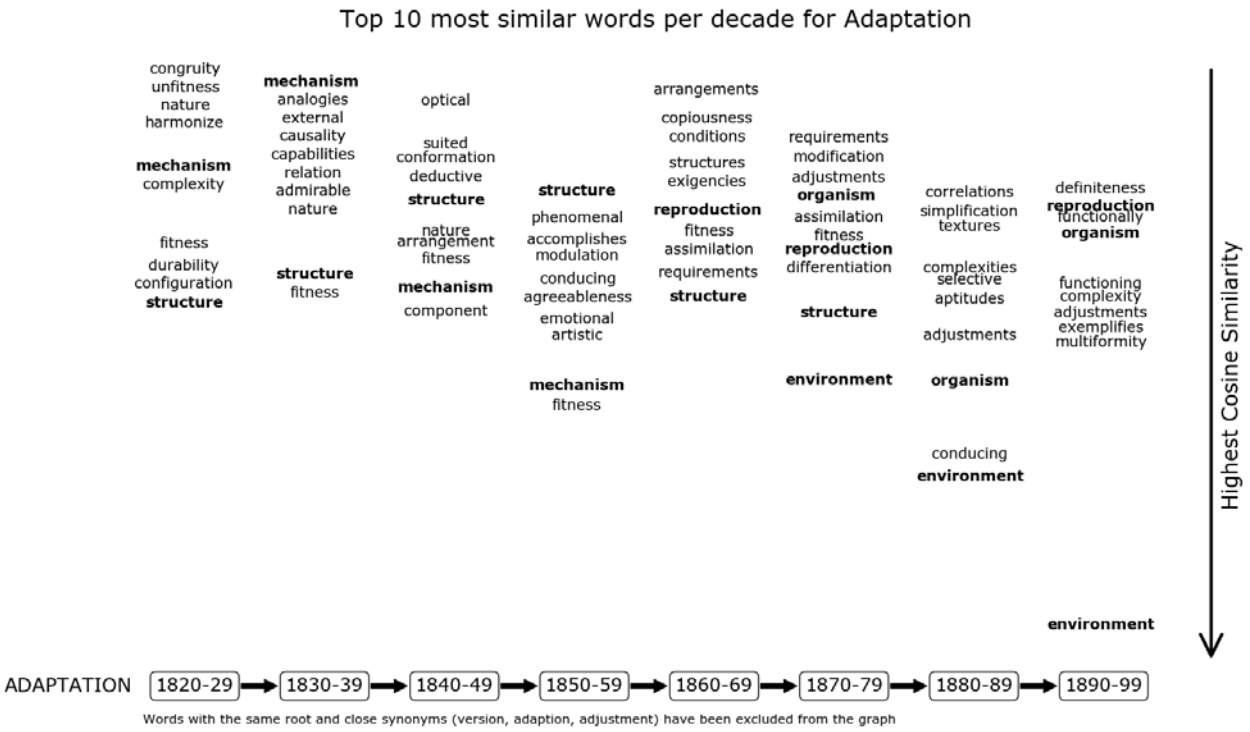
Figure 10: Semantic Associations between the Key Words in *On the Origins of Species* and the names Darwin, Wallace, Chambers and Lamarck



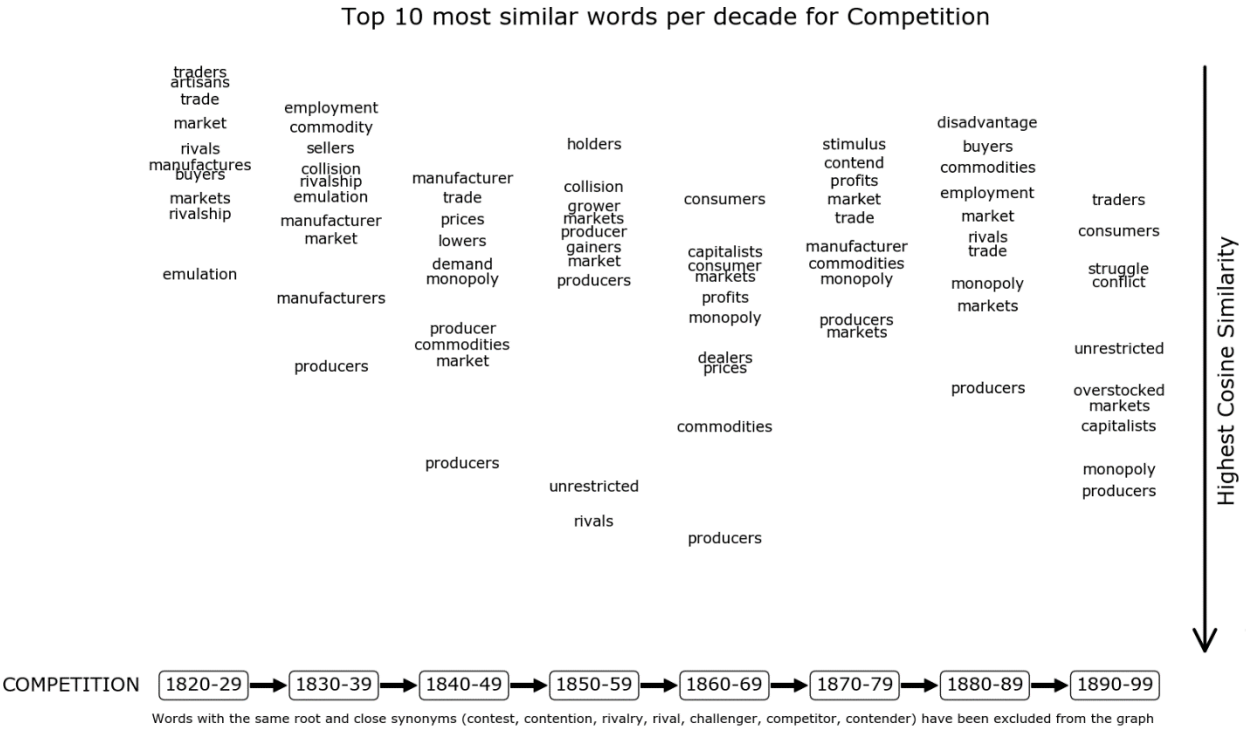
Notes: The graphs report the similarity between word on top of each chart and each of the four names in the legend. The weights in the word vectors were calculated with a Word2Vec algorithm. On the x-axis, 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

Figure 11: Top 10 most similar words for selected Darwinian words

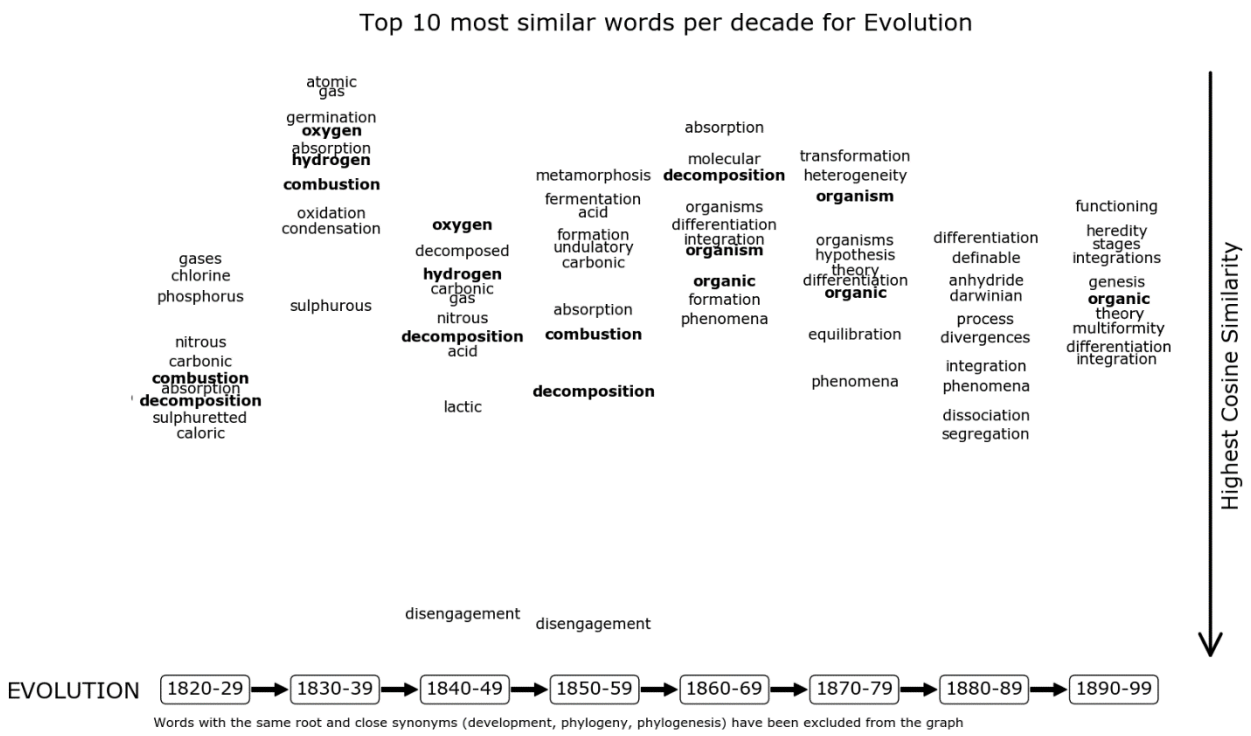
Panel A



Panel B

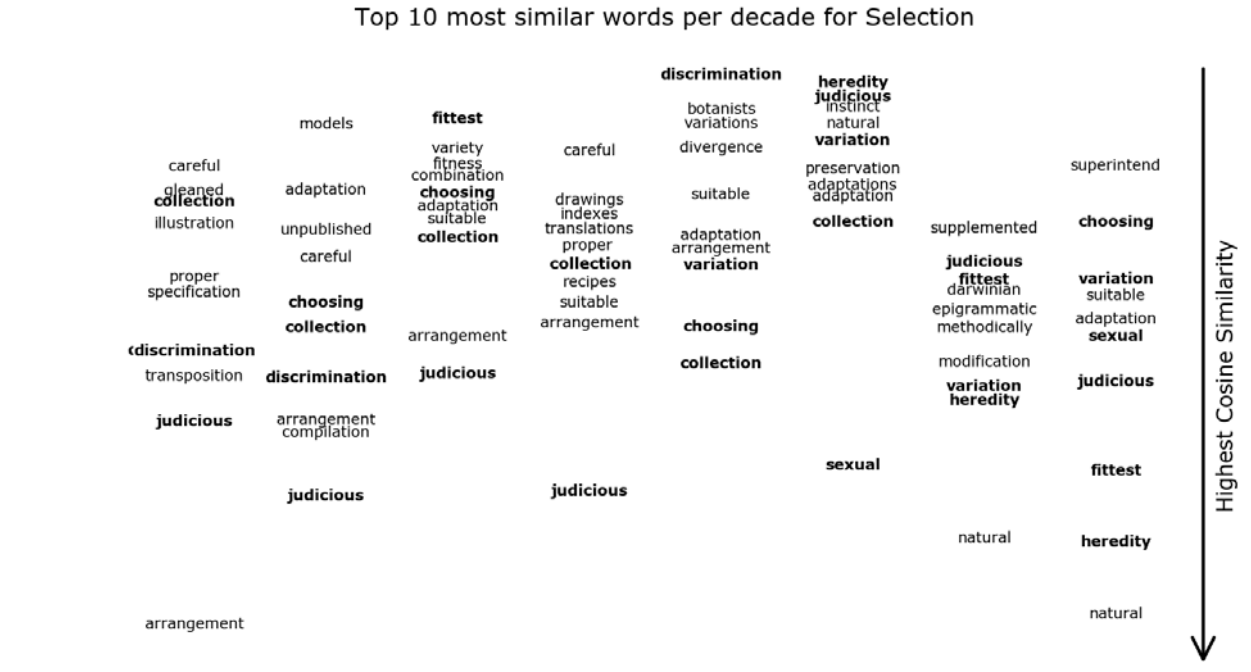


Panel C



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

Panel D



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

Panel E

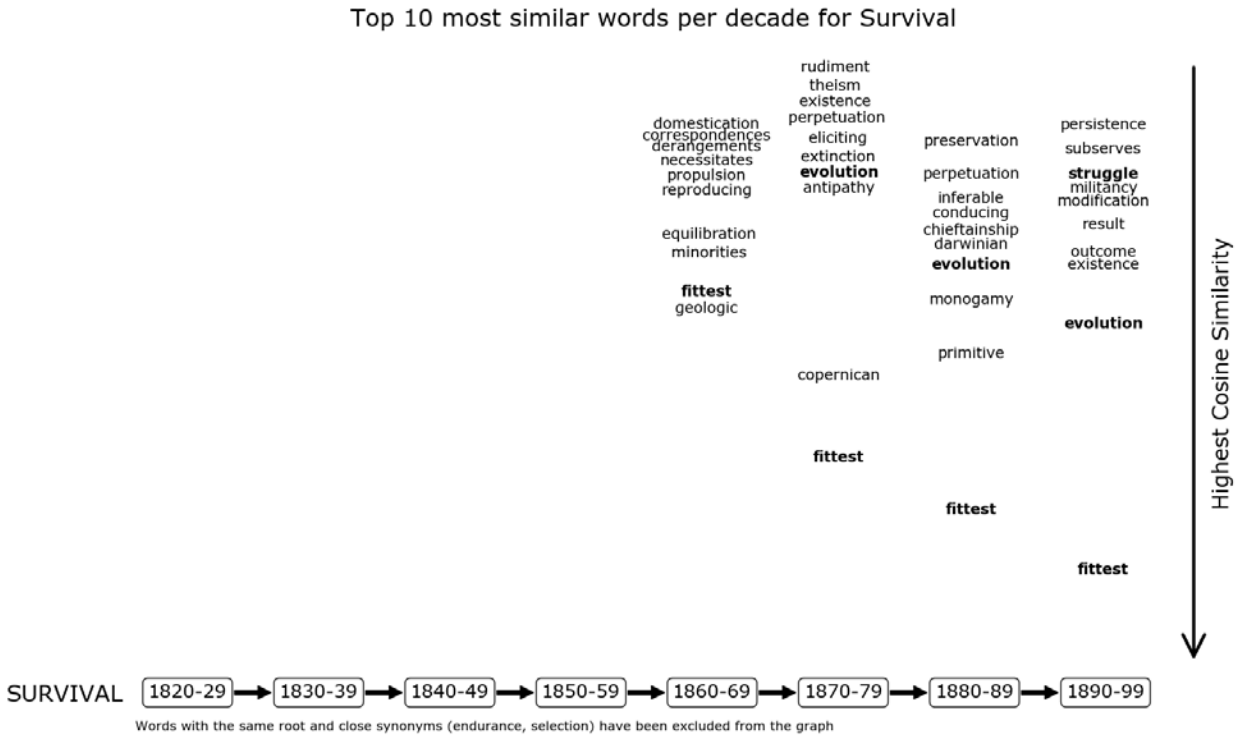
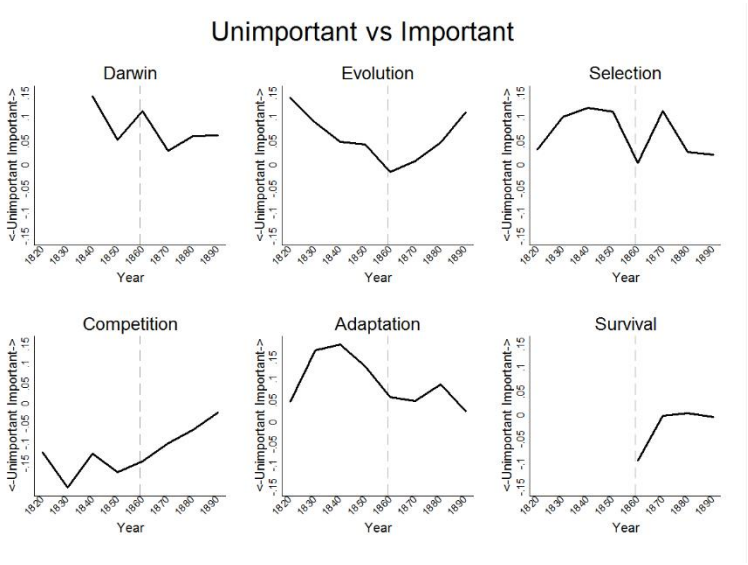
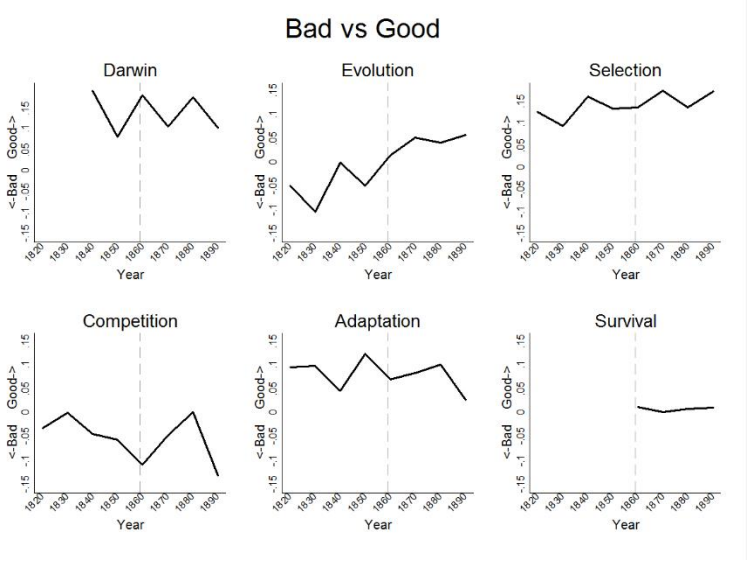

Top 10 most similar words per decade for Survival
domestication
correspondences
derangements
necessitates
propulsion
reproducing
equilibration
minorities
fittest
geologic
rudiment
theism
existence
perpetuation
eliciting
extinction
evolution
antipathy
copernican
fittest
preservation
perpetuation
inferable
conducing
chieftainship
darwinian
evolution
monogamy
primitive
fittest
persistence
subserves
struggle
militancy
modification
result
outcome
existence
evolution
fittest
Highest Cosine Similarity
SURVIVAL
1820-29
1830-39
1840-49
1850-59
1860-69
1870-79
1880-89
1890-99
Words with the same root and close synonyms (endurance, selection) have been excluded from the graph

Figure 12: Sentiment Analysis of Selected Darwinian Words in the Google Books Corpus

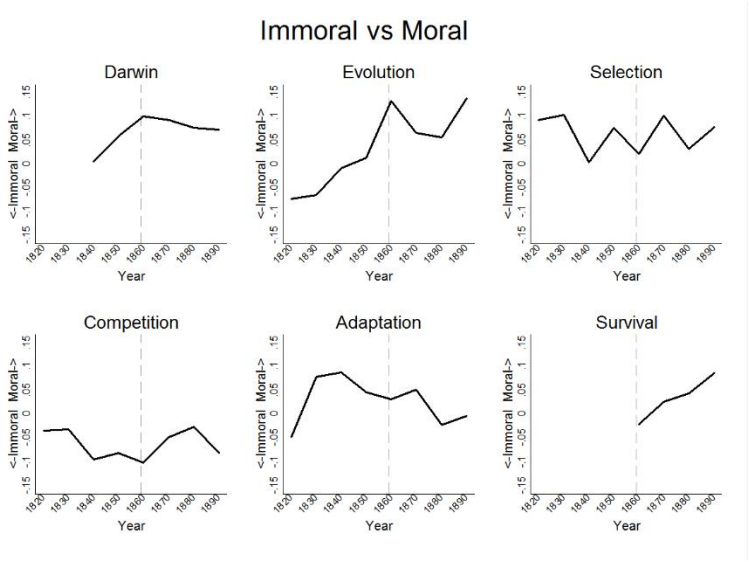
Panel A



Panel B



Panel C



Notes: The graphs report the similarity between word on top of each chart, and set of antonyms within a certain category. On the y-axis positive values of the cosine indicate higher similarity with the "positive" end of a category (Important, Good, Moral), whereas negative values indicate closer association with the negative end (Bad, Unimportant, Immoral). On the x-axis, 1820 represents the decade 1820-29, 1830 represents the decade 1830-39, and so on.

Table 1: Spline Regression Analyses – Frequency of Darwinian Concepts and Select Generic Words (Levels)

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-59	-0.003	0.177***	-0.071***	-0.020***	0.117***	0.027***
	(0.022)	(0.019)	(0.017)	(0.004)	(0.013)	(0.010)
1860-99	0.950***	0.372***	0.269***	0.171***	0.002	0.165***
	(0.044)	(0.033)	(0.020)	(0.007)	(0.008)	(0.020)
Observations	80	80	80	80	80	80
R-squared	0.952	0.888	0.720	0.952	0.658	0.779

Word/phrase:	Nature	Number	Animals	Flowers	Plants	Life
	(1)	(2)	(3)	(4)	(5)	(6)
1820-59	-1.992***	-0.593***	0.237	0.399***	0.427***	2.415***
	(0.313)	(0.166)	(0.156)	(0.106)	(0.117)	(0.376)
1860-99	-1.352***	0.640***	0.065	0.279***	0.003	3.536***
	(0.196)	(0.125)	(0.083)	(0.072)	(0.104)	(0.249)
Observations	80	80	80	80	80	80
R-squared	0.767	0.206	0.094	0.438	0.216	0.851

Notes: For each word and phrase, the two estimates refer to the slope coefficient of the nest linear fit from a spline regression of frequency (per million words) on years from 1820 to 1899, expressed as t=20, 21, …, 99, with one knot at 59. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Spline Regression Analyses – Frequency of Darwinian Concepts and Select Generic Words (Natural log)

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-59	1.103***	0.583***	-0.187***	0.518**	1.048***	3.217***
	(0.106)	(0.042)	(0.055)	(0.197)	(0.085)	(0.572)
1860-99	4.585***	1.332***	1.253***	9.432***	-0.133	10.442***
	(0.131)	(0.089)	(0.102)	(0.330)	(0.095)	(1.021)
Observations	80	80	80	80	80	80
R-squared	0.972	0.925	0.663	0.947	0.764	0.823

Word/phrase:	Nature	Number	Animals	Flowers	Plants	Life
	(1)	(2)	(3)	(4)	(5)	(6)
1820-59	-0.162***	-0.061***	0.154*	0.309***	0.352***	0.124***
	(0.030)	(0.021)	(0.085)	(0.079)	(0.078)	(0.024)
1860-99	-0.292***	0.127***	0.048	0.390***	-0.030	0.392***
	(0.037)	(0.029)	(0.077)	(0.084)	(0.118)	(0.026)
Observations	80	80	80	80	80	80
R-squared	0.761	0.163	0.147	0.490	0.292	0.828

Notes: For each word and phrase, the two estimates refer to the slope coefficient of the nest linear fit from a spline regression of natural logarithm of frequency (per million words, plus 0.01) on years from 1820 to 1899, expressed as ln(t) =ln(20), ln(21), …, ln(99), with one knot at ln(59). Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Spline Regression Analyses (one knot) – Frequency of Darwinian Concepts (natural log, separate for fiction and non-fiction books)

Word/phrase:	Evolution		Selection		Competition	
Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(1)	(2)	(3)	(4)	(5)	(6)
1820-59	1.125***	0.232	0.595***	0.310*	-0.184***	0.017
	(0.107)	(0.536)	(0.042)	(0.161)	(0.056)	(0.142)
1860-99	4.639***	5.003***	1.404***	0.321**	1.332***	-0.021
	(0.130)	(0.471)	(0.090)	(0.149)	(0.105)	(0.181)
Observations	80	80	80	80	80	80
R-squared	0.973	0.607	0.928	0.217	0.683	0.000
Word/phrase:	Survival		Adaptation		Natural Selection	
Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(7)	(8)	(9)	(10)	(11)	(12)
1820-59	0.526***	0.568	1.064***	0.832***	3.220***	1.800***
	(0.199)	(0.610)	(0.086)	(0.260)	(0.572)	(0.375)
1860-99	9.547***	8.750***	-0.075	-0.470*	10.453***	6.661***
	(0.336)	(0.714)	(0.099)	(0.248)	(1.022)	(0.758)
Observations	80	80	80	80	80	80
R-squared	0.944	0.711	0.771	0.224	0.823	0.769

Notes: For each word and phrase, the two estimates refer to the slope coefficient of the nest linear fit from a spline regression of natural logarithm of frequency (per million words, plus 0.01) on years from 1820 to 1899, expressed as ln(t) =ln(20), ln(21), …, ln(99), with one knot at ln(59), separately for fiction and non-fiction books. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Spline Regression Analyses (eight knots) – Frequency of Darwinian Concepts (natural log, separate for fiction and non-fiction books)

Word/phrase:	Evolution		Selection		Competition	
Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.191	0.385	1.020***	1.616*	0.256	1.477**
	(0.420)	(3.323)	(0.116)	(0.884)	(0.271)	(0.677)
1830-39	1.410***	1.094	0.581***	0.324	-0.357	-0.659
	(0.376)	(1.501)	(0.172)	(0.912)	(0.301)	(0.842)
1840-49	1.571***	-1.520	0.283	-0.588	0.490	-0.399
	(0.387)	(2.096)	(0.196)	(0.778)	(0.349)	(0.999)
1850-59	0.862	4.485**	0.176	0.373	-1.184**	1.410
	(0.569)	(2.174)	(0.271)	(0.548)	(0.515)	(0.880)
1860-69	4.388***	-4.938***	2.863***	-0.577	1.121**	-2.346**
	(0.778)	(1.354)	(0.459)	(0.544)	(0.503)	(0.908)
1870-79	5.321***	13.882***	0.129	3.085***	0.857	0.286
	(0.953)	(2.067)	(0.471)	(0.696)	(0.601)	(1.099)
1880-89	4.517***	4.982**	2.006***	-0.895	3.331***	1.846
	(0.856)	(2.461)	(0.423)	(0.818)	(0.656)	(1.267)
1890-99	3.167***	1.223	0.666	-1.690**	0.721	0.267
	(1.135)	(3.048)	(0.952)	(0.702)	(0.764)	(1.398)
Observations	80	80	80	80	80	80
R-squared	0.976	0.667	0.944	0.374	0.765	0.182

Word/phrase:	Survival		Adaptation		Natural Selection	
Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(7)	(8)	(9)	(10)	(11)	(12)
1820-29	1.211	0.377	1.849***	0.395	-0.863	-0.062
	(0.934)	(4.063)	(0.357)	(1.757)	(0.647)	(0.059)
1830-39	0.345	2.550	2.117***	2.342*	1.724	0.240
	(0.828)	(3.197)	(0.283)	(1.208)	(1.347)	(0.216)
1840-49	-0.795	1.339	-0.339	-0.311	-2.520	-1.167
	(0.900)	(3.666)	(0.412)	(1.474)	(1.975)	(1.001)
1850-59	0.370	-3.735	0.511	-0.028	14.204***	8.890**
	(1.322)	(4.307)	(0.415)	(1.148)	(3.906)	(3.830)
1860-69	12.473***	5.301	0.168	-0.496	22.948***	10.206*
	(2.021)	(4.333)	(0.406)	(1.048)	(5.088)	(5.162)
1870-79	12.218***	18.346***	-0.057	2.220	-1.772	7.757*
	(2.125)	(4.345)	(0.454)	(1.739)	(2.702)	(4.190)
1880-89	4.682***	7.920***	0.625	-1.651	4.471***	-3.444
	(0.921)	(2.927)	(0.441)	(2.010)	(1.180)	(3.231)
1890-99	3.666***	-0.543	0.716	-3.573***	-0.836	-0.728
	(0.820)	(2.886)	(0.563)	(1.095)	(2.175)	(3.339)
Observations	80	80	80	80	80	80
R-squared	0.964	0.747	0.875	0.302	0.947	0.885

Notes: For each word and phrase, the two estimates refer to the slope coefficient of the nest linear fit from a spline regression of natural logarithm of frequency (per million words, plus 0.01) on years from 1820 to 1899, expressed as ln(t) =ln(20), ln(21), …, ln(99), with eight knots at ln(19), ln(29), …,(ln(99), separately for fiction and non-fiction books. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Differences-in-Differences regressions – Darwinian and Generic Scientific Concepts

Outcome variable:	ln(aggregate frequency)	ln(aggregate frequency)	ln(aggregate frequency)
Sample:	Generic words	Darwinian words	Darwinian and generic words
	(1)	(2)	(3)
Regressors:			
ln(Year)	0.042***	0.405***	0.042***
	(0.010)	(0.035)	(0.010)
1(Year>1859) x ((ln(Year)-ln(59))	-0.021	1.723***	-0.021
	(0.020)	(0.092)	(0.020)
1(Darwinian word			-7.471***
			(0.134)
1(Darwinian word) x ln(Year)			0.363***
			(0.036)
1(Darwinian word) x 1(Year>1859)			1.744***
x ((ln(Year)-ln(59))			(0.095)
Constant	9.482***	2.011***	9.482***
	(0.038)	(0.129)	(0.038)
Observations	80	80	160
R-squared	0.407	0.966	1.000

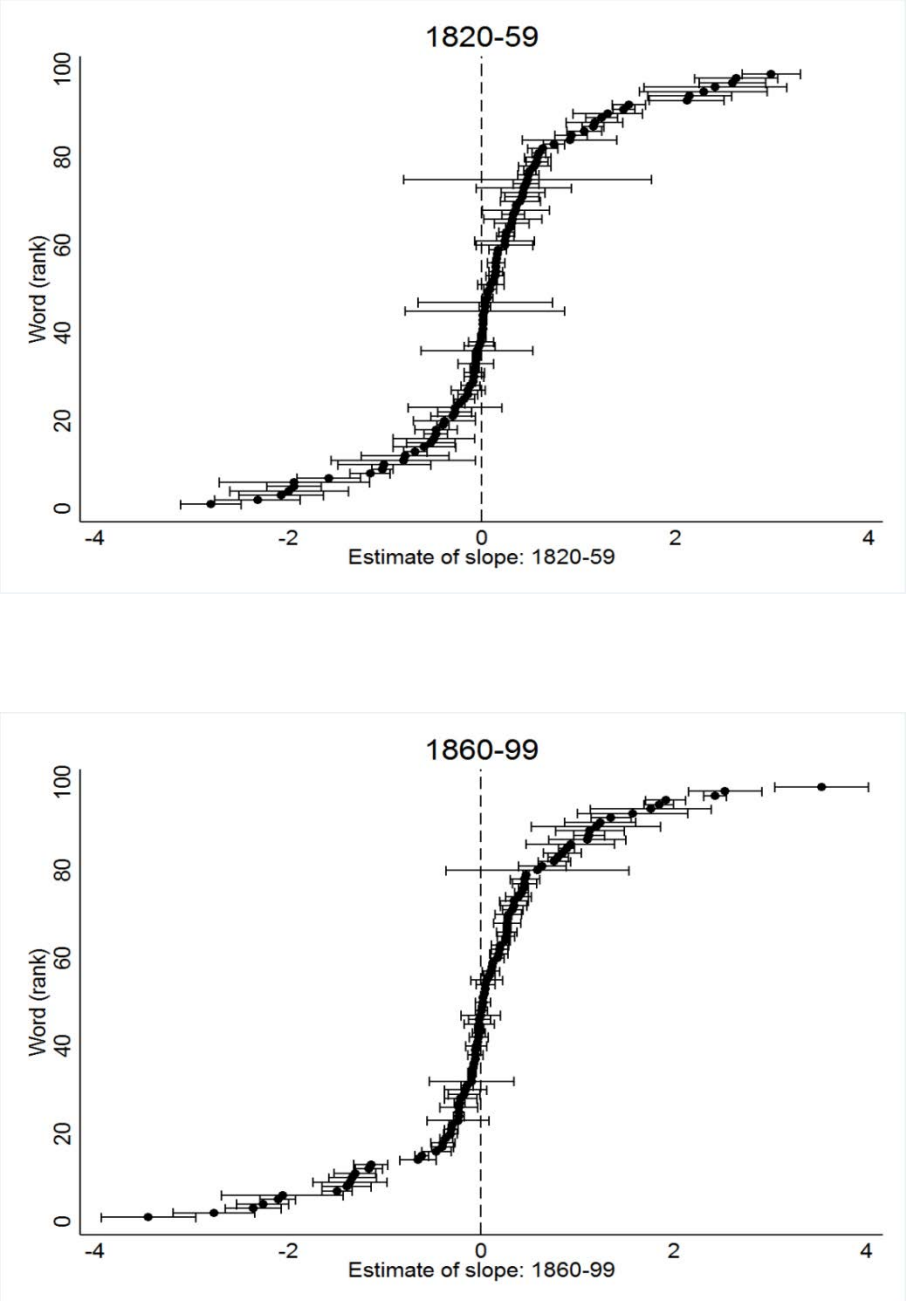
Notes: Columns 1 and 2 report estimates from regressions where the outcome variable is the natural logarithm of the aggregate frequency of the 99 most frequent nouns in *On the Origins of* Species (column 1) and of the aggregate yearly frequencies of the six Darwinian word and concepts (column 2). The regression estimates in column 3 come from combining the data used for the regressions in columns 1 and 2; therefore there are two observations per year (N=160). Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

How Does Scientific Progress Affect Cultural Changes? A Digital Text Analysis

Michela Giorcelli, University of California – Los Angeles and NBER
Nicola Lacetera, University of Toronto and NBER
Astrid Marinoni, University of Toronto

ONLINE APPENDIX – NOT FOR PUBLICATION

Figure A1: Estimates of slopes in the frequency of 99 high-frequency nouns in the 1820-59 and 1860-99 period



Notes: Each dot represents the estimates of parameters $\gamma_w$and δ_w (top and bottom graph respectively) from regression model 4, for the 99 high frequency generic nouns. Each word is represented by a number between 1 and 99 on the vertical axis (Table A1 reports the list of these words). The horizontal lines and bars are the 95% confidence intervals of the estimates.

Figure A2: Correlation in the estimates of slopes in the frequency of 99 high-frequency nouns in the 1820-59 and 1860-99 period

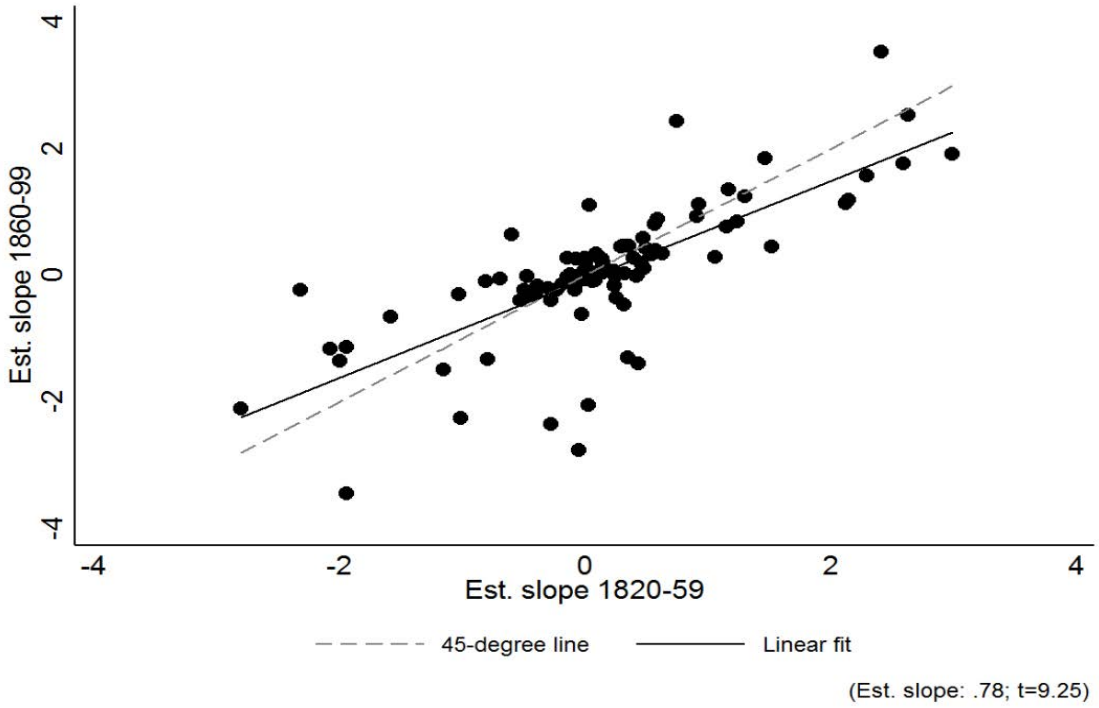
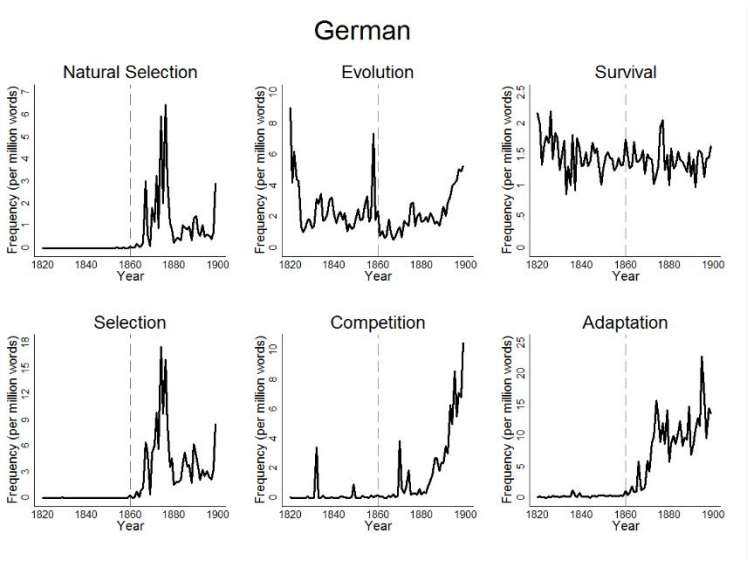
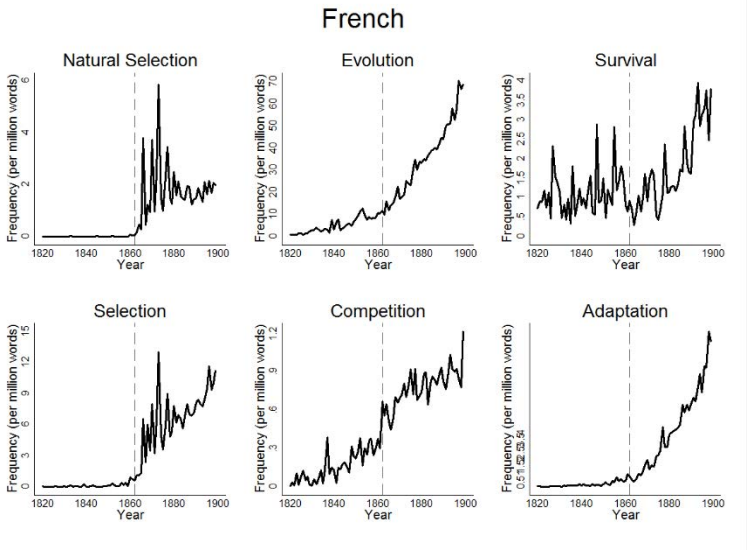


Figure A3: Frequencies (per 1 Million Words) of the phrase "Evolution", "Survival", "Competition" in six languages other than English

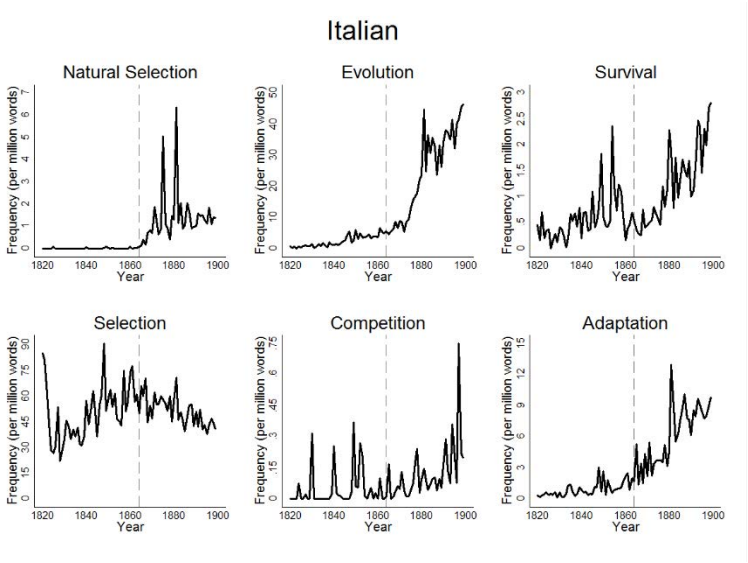
Panel A



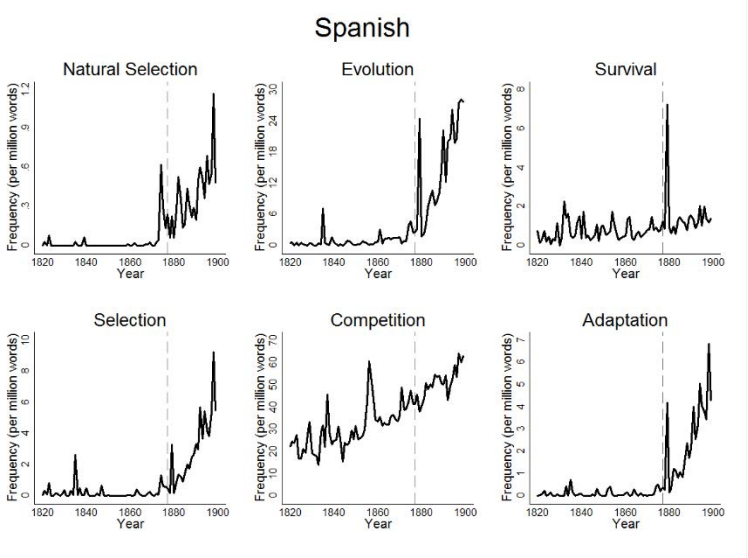
Panel B



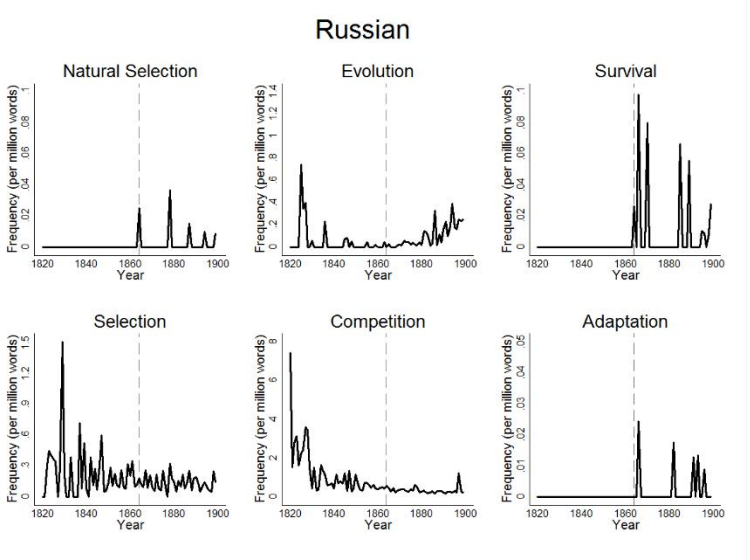
Panel C



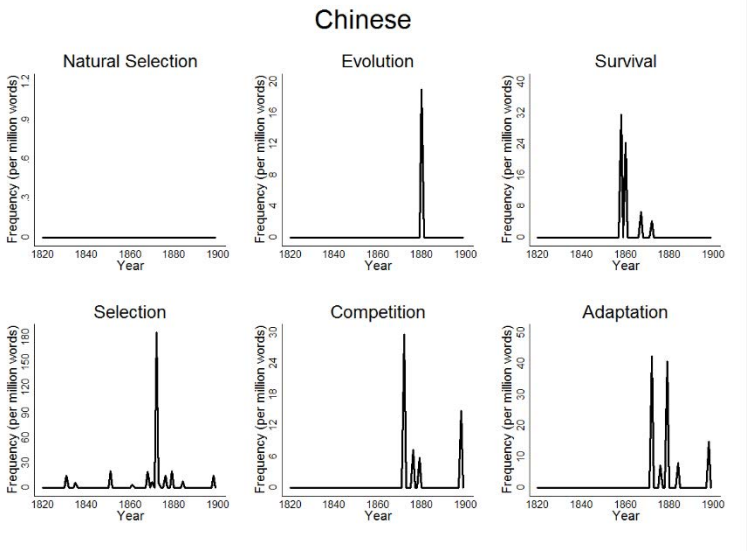
Panel D



Panel E



Panel F



Notes: For each year, the figures report the number of occurrences (per million words) of the phrase "Evolution", "Survival", "Competition" in the language indicated on top of a graph; the vertical dashed line are in correspondence of the year of the first published translation of *On the Origin of Species* in a given language.

Table A1: Generic Words

action	forms	parent
advantage	genera	parts
animal	generations	period
animals	genus	periods
beings	group	plant
birds	groups	plants
breeds	habits	points
case	hand	pollen
cases	hybrids	power
change	importance	principle
changes	individuals	process
character	inhabitants	productions
characters	insects	reason
class	instance	respect
climate	instincts	sea
conditions	islands	seeds
country	kind	size
degree	kinds	species
descendants	land	state
descent	life	sterility
development	man	structure
difference	manner	subject
differences	means	tendency
difficulty	modification	theory
eggs	naturalists	time
fact	nature	variation
facts	number	variations
fertility	numbers	varieties
flower	offspring	variety
flowers	older	view
form	organ	water
formation	organization	world
formations	organs	years

Notes: The table lists the 99 most frequent nouns in *On the Origins of Species*, which we used as controls for the Darwinian concepts in some of the analyses.

Table A2: Spline Regression Analyses (eight knots) – Frequency of Darwinian Concepts and Select Generic Words

Panel A. Levels

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.024	0.390***	0.149	0.003	0.289***	-0.001
	(0.033)	(0.052)	(0.135)	(0.003)	(0.085)	(0.001)
1830-39	0.076***	0.183***	-0.145	0.001	0.284***	0.001
	(0.024)	(0.063)	(0.117)	(0.002)	(0.057)	(0.003)
1840-49	0.135***	0.101	0.147	-0.001	-0.023	-0.001
	(0.031)	(0.061)	(0.102)	(0.002)	(0.061)	(0.010)
1850-59	0.053	-0.011	-0.264**	-0.003	0.051	0.005
	(0.052)	(0.085)	(0.113)	(0.005)	(0.051)	(0.039)
1860-69	0.438***	0.828***	0.174*	0.045***	0.017	0.369***
	(0.093)	(0.159)	(0.090)	(0.014)	(0.046)	(0.109)
1870-79	0.941***	0.008	0.168	0.190***	0.001	0.007
	(0.179)	(0.144)	(0.114)	(0.017)	(0.047)	(0.096)
1880-89	1.336***	0.579***	0.619***	0.253***	0.046	0.240***
	(0.226)	(0.123)	(0.125)	(0.024)	(0.042)	(0.064)
1890-99	1.151**	0.129	0.140	0.192***	0.033	-0.008
	(0.460)	(0.280)	(0.159)	(0.049)	(0.046)	(0.145)
Observations	80	80	80	80	80	80
R-squared	0.969	0.906	0.792	0.982	0.773	0.814

Word/phrase:	Nature	Number	Animals	Flowers	Plants	Life
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.760	2.655**	2.664*	0.803	1.523*	-2.180
	(2.677)	(1.095)	(1.557)	(1.111)	(0.774)	(3.208)
1830-39	-3.178*	0.054	0.260	-0.456	1.210	-1.842
	(1.894)	(1.186)	(0.846)	(0.687)	(0.733)	(2.505)
1840-49	-0.509	-1.644*	-1.190*	1.216***	0.162	7.510***
	(1.269)	(0.978)	(0.695)	(0.338)	(0.617)	(1.028)
1850-59	-4.066***	-1.231	0.660	-0.480	-0.703	0.723
	(1.421)	(0.777)	(0.508)	(0.397)	(0.657)	(1.140)
1860-69	-0.656	0.351	0.463	1.429***	0.631	7.011***
	(1.255)	(0.722)	(0.455)	(0.432)	(0.527)	(1.475)
1870-79	-1.611	1.531**	0.469	0.187	-0.052	0.978
	(1.023)	(0.768)	(0.572)	(0.514)	(0.638)	(1.321)
1880-89	0.712	1.003	-0.553	-0.837*	-0.380	3.458**
	(0.771)	(0.811)	(0.532)	(0.500)	(0.657)	(1.357)
1890-99	-4.699***	-0.366	-0.258	1.182**	1.215*	2.131
	(0.680)	(0.919)	(0.613)	(0.472)	(0.725)	(2.232)
Observations	80	80	80	80	80	80
R-squared	0.792	0.400	0.264	0.515	0.349	0.901

Notes: The table reports estimates from regressions of the relative annual frequency of use (per million words) of a given word or phrase on a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the current year and 1829, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, 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.

Panel B. Natural logarithms

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.183	1.034***	0.282	1.199	1.825***	-0.859
	(0.399)	(0.225)	(0.288)	(0.798)	(0.314)	(1.564)
1830-39	1.371***	0.557**	-0.389	0.321	2.080***	1.717
	(0.402)	(0.227)	(0.291)	(0.805)	(0.316)	(1.578)
1840-49	1.546***	0.257	0.467	-0.570	-0.338	-2.514
	(0.506)	(0.286)	(0.366)	(1.014)	(0.398)	(1.986)
1850-59	0.868	0.174	-1.127**	0.033	0.491	14.195***
	(0.622)	(0.351)	(0.449)	(1.246)	(0.489)	(2.441)
1860-69	4.244***	2.727***	0.976*	12.327***	0.094	22.920***
	(0.739)	(0.417)	(0.534)	(1.480)	(0.581)	(2.898)
1870-79	5.473***	0.255	0.912	12.222***	0.060	-1.754
	(0.856)	(0.483)	(0.618)	(1.714)	(0.673)	(3.358)
1880-89	4.351***	1.808***	3.159***	4.770**	0.418	4.450
	(0.986)	(0.556)	(0.712)	(1.974)	(0.775)	(3.866)
1890-99	2.918**	0.434	0.525	3.297	0.448	-0.850
	(1.297)	(0.732)	(0.937)	(2.597)	(1.020)	(5.087)
Observations	80	80	80	80	80	80
R-squared	0.976	0.942	0.749	0.967	0.870	0.947

Word/phrase:	Nature	Number	Animals	Flowers	Plants	Life
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.050	0.191**	0.889***	0.352	0.673*	-0.091
	(0.111)	(0.089)	(0.280)	(0.340)	(0.369)	(0.083)
1830-39	-0.222*	0.017	0.144	-0.171	0.832**	-0.094
	(0.112)	(0.090)	(0.282)	(0.343)	(0.373)	(0.083)
1840-49	-0.064	-0.218*	-0.614*	1.018**	0.033	0.508***
	(0.141)	(0.113)	(0.355)	(0.432)	(0.469)	(0.105)
1850-59	-0.504***	-0.209	0.465	-0.411	-0.566	0.067
	(0.173)	(0.139)	(0.437)	(0.531)	(0.576)	(0.129)
1860-69	-0.126	0.058	0.265	1.445**	0.630	0.606***
	(0.205)	(0.166)	(0.518)	(0.631)	(0.684)	(0.153)
1870-79	-0.296	0.369*	0.419	0.262	-0.084	0.109
	(0.238)	(0.192)	(0.601)	(0.731)	(0.793)	(0.177)
1880-89	0.162	0.240	-0.490	-1.060	-0.420	0.370*
	(0.274)	(0.221)	(0.692)	(0.841)	(0.913)	(0.204)
1890-99	-1.198***	-0.085	-0.292	1.661	1.672	0.243
	(0.360)	(0.291)	(0.910)	(1.107)	(1.201)	(0.269)
Observations	80	80	80	80	80	80
R-squared	0.806	0.404	0.312	0.553	0.403	0.897

Notes: The table reports regressions of the natural logarithm relative annual frequency of use (per million words +0.01) of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the natural logarithm of the current year and the natural logarithm of (18)29, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, 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 A3: Spline Regression Analyses (eight knots) – Frequency of Darwinian Concepts: Fiction and Non-fiction

	Word/phrase:	Evolution		Selection		Competition	
	Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
		(1)	(2)	(3)	(4)	(5)	(6)
1820-29		0.024	-0.045	0.389***	0.359*	0.140	0.312**
		(0.035)	(0.039)	(0.053)	(0.210)	(0.142)	(0.144)
1830-39		0.081***	0.034	0.194***	0.091	-0.137	-0.095
		(0.025)	(0.026)	(0.064)	(0.196)	(0.123)	(0.143)
1840-49		0.144***	-0.031	0.113*	-0.148	0.158	-0.085
		(0.033)	(0.029)	(0.062)	(0.153)	(0.108)	(0.147)
1850-59		0.052	0.053**	-0.015	0.044	-0.283**	0.161*
		(0.055)	(0.025)	(0.090)	(0.079)	(0.120)	(0.096)
1860-69		0.482***	-0.078***	0.899***	-0.053	0.206**	-0.206***
		(0.101)	(0.024)	(0.169)	(0.068)	(0.095)	(0.074)
1870-79		0.957***	0.298***	-0.032	0.339***	0.164	0.011
		(0.189)	(0.077)	(0.154)	(0.080)	(0.121)	(0.071)
1880-89		1.467***	0.313**	0.668***	-0.092	0.684***	0.128
		(0.240)	(0.152)	(0.130)	(0.091)	(0.133)	(0.088)
1890-99		1.335***	0.008	0.210	-0.166**	0.199	-0.005
		(0.493)	(0.225)	(0.306)	(0.069)	(0.179)	(0.094)
Observations		80	80	80	80	80	80
R-squared		0.970	0.677	0.908	0.331	0.810	0.168

	Word/phrase:	Survival		Adaptation		Natural Selection	
	Sample:	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
		(7)	(8)	(9)	(10)	(11)	(12)
1820-29		0.004	-0.007	0.298***	0.005	-0.001	-0.000
		(0.002)	(0.014)	(0.087)	(0.059)	(0.001)	(0.000)
1830-39		0.000	0.005	0.300***	0.115*	0.001	0.001
		(0.002)	(0.006)	(0.058)	(0.059)	(0.003)	(0.001)
1840-49		-0.001	0.001	-0.023	-0.026	-0.001	-0.004
		(0.002)	(0.006)	(0.060)	(0.079)	(0.010)	(0.003)
1850-59		-0.003	-0.001	0.056	-0.013	0.005	0.017
		(0.005)	(0.007)	(0.054)	(0.046)	(0.039)	(0.012)
1860-69		0.048***	-0.010	0.027	-0.009	0.371***	0.006
		(0.016)	(0.014)	(0.050)	(0.033)	(0.110)	(0.013)
1870-79		0.193***	0.124**	-0.012	0.077	0.007	0.049**
		(0.018)	(0.047)	(0.052)	(0.060)	(0.097)	(0.020)
1880-89		0.258***	0.222***	0.070	-0.068	0.243***	-0.018
		(0.027)	(0.077)	(0.045)	(0.065)	(0.064)	(0.022)
1890-99		0.221***	-0.044	0.059	-0.060**	-0.007	-0.018
		(0.052)	(0.088)	(0.052)	(0.027)	(0.146)	(0.021)
Observations		80	80	80	80	80	80
R-squared		0.981	0.789	0.785	0.196	0.813	0.522

Notes: The table reports regressions of the natural logarithm of the relative annual frequency (per million words + 0.01) of use of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, interactions of these indicators with the difference between the natural logarithm of the current year and the natural logarithm of 29, 39, 49, 59, 69, 79 and 89, respectively, and interactions of all these previous terms with an indicator for whether an observation pertains to fiction books as opposed to non-fiction books. There are two observations per year, one based on the corpus of non-fiction books, and the other on the corpus of non-fiction books (N=160). Each regression is limited to one word or phrase as indicated in the corresponding columns. 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<0.10.

Table A4: Spline Regression Analyses (eight knots) – Frequency of Darwinian Concepts in US Congressional Records

Panel A. Levels

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.169	2.908	-1.334	-0.002	0.548	-0.000
	(0.105)	(2.146)	(8.500)	(0.002)	(0.352)	(0.000)
1830-39	-0.111	-3.557***	-12.213*	0.006	-0.501*	0.000
	(0.077)	(1.130)	(6.148)	(0.005)	(0.281)	(0.000)
1840-49	0.065	0.326	1.669	-0.002	0.412**	-0.001
	(0.045)	(1.035)	(3.567)	(0.003)	(0.199)	(0.001)
1850-59	0.007	-0.760	-4.716	0.003	-0.302*	0.003
	(0.065)	(0.722)	(3.451)	(0.007)	(0.177)	(0.003)
1860-69	-0.066	1.450*	3.532	-0.030*	-0.177*	0.003
	(0.067)	(0.797)	(2.500)	(0.016)	(0.098)	(0.004)
1870-79	0.011	-1.458*	0.764	0.133***	0.006	-0.000
	(0.059)	(0.753)	(3.671)	(0.048)	(0.067)	(0.010)
1880-89	0.358**	1.602*	19.141***	0.140	0.146	0.035**
	(0.169)	(0.937)	(5.115)	(0.091)	(0.096)	(0.013)
1890-99	1.067***	-0.606	-15.709**	0.061	-0.126	-0.019
	(0.255)	(1.117)	(6.805)	(0.133)	(0.114)	(0.020)
Observations	80	80	80	80	80	80
R-squared	0.725	0.271	0.393	0.674	0.285	0.447

Notes: The table reports estimates from regressions of the relative annual frequency of use (per million words) of a given word or phrase on a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the current year and 1829, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, 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.

Panel B. Natural logarithms

Word/phrase:	Evolution	Selection	Competition	Survival	Adaptation	Natural Selection
	(1)	(2)	(3)	(4)	(5)	(6)
1820-29	0.169	2.908	-1.334	-0.002	0.548	-0.000
	(0.105)	(2.146)	(8.500)	(0.002)	(0.352)	(0.000)
1830-39	-0.111	-3.557***	-12.213*	0.006	-0.501*	0.000
	(0.077)	(1.130)	(6.148)	(0.005)	(0.281)	(0.000)
1840-49	0.065	0.326	1.669	-0.002	0.412**	-0.001
	(0.045)	(1.035)	(3.567)	(0.003)	(0.199)	(0.001)
1850-59	0.007	-0.760	-4.716	0.003	-0.302*	0.003
	(0.065)	(0.722)	(3.451)	(0.007)	(0.177)	(0.003)
1860-69	-0.066	1.450*	3.532	-0.030*	-0.177*	0.003
	(0.067)	(0.797)	(2.500)	(0.016)	(0.098)	(0.004)
1870-79	0.011	-1.458*	0.764	0.133***	0.006	-0.000
	(0.059)	(0.753)	(3.671)	(0.048)	(0.067)	(0.010)
1880-89	0.358**	1.602*	19.141***	0.140	0.146	0.035**
	(0.169)	(0.937)	(5.115)	(0.091)	(0.096)	(0.013)
1890-99	1.067***	-0.606	-15.709**	0.061	-0.126	-0.019
	(0.255)	(1.117)	(6.805)	(0.133)	(0.114)	(0.020)
Observations	80	80	80	80	80	80
R-squared	0.725	0.271	0.393	0.674	0.285	0.447

Notes: The table reports regressions of the natural logarithm relative annual frequency of use (per million words +0.01) of a given word or phrase on the natural logarithm of a linear time trend, indicators for the years after 1829, 39, 49, 59, 69, 79 and 89, and interactions of these indicators with the difference between the natural logarithm of the current year and the natural logarithm of (18)29, 39, 49, 59, 69, 79 and 89, respectively. Each regression is limited to one word or phrase as indicated in the corresponding column, 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.