Correction: 11 March 2011

www.sciencemag.org/cgi/content/full/science.1199644/DC1

Supporting Online Material for

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,* Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden*

*To whom correspondence should be addressed. E-mail: jb.michel@gmail.com (J.B.M.); erez@erez.com (E.A.).

Published 16 December 2010 on *Science* Express
DOI: 10.1126/science.1199644

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1199644/DC1)

SOM Data

Correction: SOM references have been added to the SOM PDF, and a set of SOM data has been added in Excel format.

Materials and Methods

"Quantitative analysis of culture using millions of digitized books", Michel et al.

Contents

I. Overview of Google Books Digitization

In 2004, Google began scanning books to make their contents searchable and discoverable online. To date, Google has scanned over fifteen million books: over 11% of all books ever published. The collection contains over five billion pages and two trillion words, with books dating back to as early as 1473 and representing 478 languages. Over two million of these scanned books were given directly to Google by their publishers; the rest were borrowed from large libraries such as the University of Michigan and the New York Public Library. The scanning effort involves significant engineering challenges, some of which are highly relevant to the construction of the historical n-grams corpus. We survey those issues here.

In particular, in this section, we describe the creation of a database of book editions; each entry contains the digital text and associated metadata for a particular book edition. This database is the result of three steps: algorithmic curation of metadata records to create a database of book editions, digitization of the corresponding texts, and decomposition of the digitized text into structural components such as frontmatter, backmatter, headers, footers, page numbers, and body text.

I.1. Metadata

Over 100 sources of metadata information were used by Google to generate a comprehensive catalog of books. Some of these sources are library catalogs (e.g., the list of books in the collections of University of Michigan, or union catalogs such as the collective list of books in Bosnian libraries), some are from retailers (e.g., Decitre, a French bookseller), and some are from commercial aggregators (e.g., Ingram). In addition, Google also receives metadata from its 30,000 partner publishers. Each metadata source consists of a series of digital records, typically in either the MARC format favored by libraries, or the ONIX format used by the publishing industry. Each record refers to either a specific edition of a book or a physical copy of a book on a library shelf, and contains conventional bibliographic data such as title, author(s), publisher, date of publication, and language(s) of publication.

Cataloguing practices vary widely among these sources, and even within a single source over time. Thus two records for the same edition will often differ in multiple fields. This is especially true for serials (e.g., the *Congressional Record*) and multivolume works such as sets (e.g., the three volumes of *The Lord of the Rings*).

The matter is further complicated by ambiguities in the definition of the word ‘book’ itself. Including translations, there are over three thousand editions derived from Mark Twain’s original *Tom Sawyer*.

Google’s process of converting the billions of metadata records into a single nonredundant database of book editions consists of the following principal steps:

Coarsely dividing the billions of metadata records into groups that may refer to the same work (e.g., *Tom Sawyer*).

Identifying and aggregating multivolume works based on the presence of cues from individual records.

Subdividing the group of records corresponding to each work into constituent groups corresponding to the various editions (e.g., the 1909 publication of *De lotgevallen van Tom Sawyer*, translated from English to Dutch by Johan Braakensiek).

Merging the records for each edition into a new "consensus" record.

The result is a set of consensus records, where each record corresponds to a distinct book edition and work, and where the contents of each record are formed out of fields from multiple sources. The number of records in this set -- i.e., the number of known book editions -- increases every year as more books are written.

In August 2010, this evaluation identified 129 million editions, which is the working estimate we use in this paper of all the editions ever published (this includes serials and sets but excludes kits, mixed media, and periodicals such as newspapers). This final database contains bibliographic information for each of these 129 million editions (**Ref. S1**). The country of publication is known for 85.3% of these editions, authors for 87.8%, publication dates for 92.6%, and the language for 91.6%. Of the 15 million books scanned, the country of publication is known for 91.5%, authors for 92.1%, publication dates for 95.1%, and the language for 98.6%.

I.2. Digitization

We describe the way books are scanned and digitized. For publisher-provided books, Google removes the spines and scans the pages with industrial sheet-fed scanners. For library-provided books, Google uses custom-built scanning stations designed to impose only as much wear on the book as would result from someone reading the book. As the pages are turned, stereo cameras overhead photograph each page, as shown in **Figure S15**.

One crucial difference between sheet-fed scanners and the stereo scanning process is the flatness of the page as the image is captured. In sheet-fed scanning, the page is kept flat, similar to conventional flatbed scanners. With stereo scanning, the book is cradled at an angle that minimizes stress on the spine of the book (this angle is not shown in **Figure S15**). Though less damaging to the book, a disadvantage of the latter approach is that it results in a page that is curved relative to the plane of the camera. The curvature changes every time a page is turned, for several reasons: the attachment point of the page in the spine differs, the two stacks of pages change in thickness, and the tension with which the book is held open may vary. Thicker books have more page curvature and more variation in curvature.

This curvature is measured by projecting a fixed infrared pattern onto each page of the book, subsequently captured by cameras. When the image is later processed, this pattern is used to identify the location of the spine and to determine the curvature of the page. Using this curvature information, the scanned image of each page is digitally resampled so that the results correspond as closely as possible

to the results of sheet-fed scanning. The raw images are also digitally cropped, cleaned, and contrast enhanced. Blurred pages are automatically detected and rescanned. Details of this approach can be found in U.S. Patents 7463772 and 7508978; sample results are shown in **Figure S16.**

Finally, blocks of text are identified and optical character recognition (OCR) is used to convert those images into digital characters and words, in an approach described elsewhere (**Ref. S2**). The difficulty of applying conventional OCR techniques to Google's scanning effort is compounded because of variations in language, font, size, paper quality, and the physical condition of the books being scanned. Nevertheless, Google estimates that over 98% of words are correctly digitized for modern English books. After OCR, initial and trailing punctuation is stripped and word fragments split by hyphens are joined, yielding a stream of words suitable for subsequent indexing.

I.3. Structure Extraction

After the book has been scanned and digitized, the components of the scanned material are classified into various types. For instance, individual pages are scanned in order to identify which pages comprise the authored content of the book, as opposed to the pages which comprise frontmatter and backmatter, such as copyright pages, tables of contents, index pages, etc. Within each page, we also identify repeated structural elements, such as headers, footers, and page numbers.

Using OCR results from the frontmatter and backmatter, we automatically extract author names, titles, ISBNs, and other identifying information. This information is used to confirm that the correct consensus record has been associated with the scanned text.

II. Construction of Historical N-grams Corpora

As noted in the paper text, we did not analyze the entire set of 15 million books digitized by Google. Instead, we

Performed further filtering steps to select only a subset of books with highly accurate metadata.

Subdivided the books into 'base corpora' using such metadata fields as language, country of publication, and subject.

For each base corpus, construct a massive numerical table that lists, for each n-gram (often a word or phrase), how often it appears in the given base corpus in every single year between 1550 and 2008.

In this section, we will describe these three steps. These additional steps ensure high data quality, and also make it possible to examine historical trends without violating the 'fair use' principle of copyright law: our object of study is the frequency tables produced in step 3 (which are available as supplemental data), and not the full-text of the books.

II.1. Additional filtering of books

II.1A. Accuracy of Date-of-Publication metadata

Accurate date-of-publication data is crucial component in the production of time-resolved n-grams data. Because our study focused most centrally on the English language corpus, we decided to apply more stringent inclusion criteria in order to make sure the accuracy of the date-of-publication data was as high as possible.

We found that the lion's share of date-of-publication errors were due to so-called 'bound-withs' - single volumes that contain multiple works, such as anthologies or collected works of a given author. Among these bound-withs, the most inaccurately dated subclass were serial publications, such as journals and periodicals. For instance, many journals had publication dates which were erroneously attributed to the year in which the first issue of the journal had been published. These journals and serial publications also represented a different aspect of culture than the books did. For these reasons, we decided to filter out all serial publications to the extent possible. Our 'Serial Killer' algorithm removed serial publications by looking for suggestive metadata entries, containing one or more of the following:

Serial-associated titles, containing such phrases as 'Journal of', 'US Government report', etc.

Serial-associated authors, such as those in which the author field is blank, too numerous, or contains words such as 'committee'.

Note that the match is case-insensitive, and it must be to a complete word in the title; thus the filtering of titles containing the word ‘digest’ does not lead to the removal of works with ‘digestion’ in the title. The entire list of serial-associated title phrases and serial-associated author phrases is included as supplemental data (**Appendix**). For English books, 29.4% of books were filtered using the 'Serial Killer', with the title filter removing 2% and the author filter removing 27.4%. Foreign language corpora were filtered in a similar fashion.

This filtering step markedly increased the accuracy of the metadata dates. We determined metadata accuracy by examining 1000 filtered volumes distributed uniformly over time from 1801-2000 (5 per year). An annotator with no knowledge of our study manually determined the date-of-publication. The annotator was aware of the Google metadata dates during this process. We found that 5.8% of English books had metadata dates that were more than 5 years from the date determined by a human examining the book. Because errors are much more common among older books, and because the actual corpora are strongly biased toward recent works, the likelihood of error in a randomly sampled book from the final corpus is much lower than this value. As a point of comparison, 27 of 100 books (27%) selected at random from an unfiltered corpus contained date-of-publication errors of greater than 5 years. The unfiltered corpus was created using a sampling strategy similar to that of Eng-1M. This selection mechanism favored recent books (which are more frequent) and pre-1800 books, which were excluded in the sampling strategy for filtered books; as such the two numbers (6.2% and 27%) give a sense of the improvement, but are not strictly comparable.

The importance of such filtering is heightened by the observation that many of the original dating errors cluster around particular years. This has multiple causes; for instance, these years might be the first year in which a particular serial commenced publication. Metadata sources also occasionally assign default dates, such as 1899 or 1905, to books whose publication date is unknown. Thus, instead of producing a low, uniform rate of background noise, highly localized, gross errors are apparent in the timelines that derive from unfiltered corpora. The filtering step both reduces overall error rate and, crucially, dramatically reduces the rate of localized error.

Note that since the base corpora were generated (August 2009), many additional improvements have been made to the metadata dates used in Google Book Search itself. As such, these numbers do not reflect the accuracy of the Google Book Search online tool.

II.1B. OCR quality

The challenge of performing accurate OCR on the entire books dataset is compounded by variations in such factors as language, font, size, legibility, and physical condition of the book. OCR quality was assessed using an algorithm developed by Popat et al. (**Ref S3**). This algorithm yields a probability that expresses the confidence that a given sequence of text generated by OCR is correct. Incorrect or *anomalous* text can result from gross imperfections in the scanned images, or as a result of markings or drawings. This algorithm uses sophisticated statistics, a variant of the Partial by Partial Matching (PPM) model, to compute for each glyph (character) the probability that it is anomalous given other nearby glyphs. ('Nearby' refers to 2-dimensional distance on the original scanned image, hence glyphs above, below, to the left, and to the right of the target glyph.) The model parameters are tuned using multi-language subcorpora, one in each of the 32 supported languages. From the per-glyph probability one can compute an aggregate probability for a sequence of glyphs, including the entire text of a volume. In this manner, every volume has associated with it a probabilistic OCR quality score (quantized to an integer between 0-100; note that the OCR quality score should not be confused with character or word accuracy). In addition to error detection, the Popat model is also capable of computing the probability that the text is in a particular language given any sequence of characters. Thus the algorithm serves the dual purpose of detecting anomalous text while simultaneously identifying the language in which the text is written.

To ensure the highest quality data, we excluded volumes with poor OCR quality. For the languages that use a Latin alphabet (English, French, Spanish, and German), the OCR quality is generally higher, and more books are available. As a result, we filtered out all volumes whose quality score was lower than 80%. For Chinese and Russian, fewer books were available, and we did not apply the OCR filter. For Hebrew, a 50% threshold was used, because its OCR quality was relatively better than Chinese or Russian. For geographically specific corpora, English US and English UK, a less stringent 60% threshold was used, in order to maximize the number of books included (note that, as such, these two corpora are not strict subsets of the broader English corpus). **Figure S18** shows the distribution of OCR quality score as a function of the fraction of books in the English corpus. Use of an 80% cut off will remove the books with the worst OCR, while retaining the vast majority of the books in the original corpus.

The OCR quality scores were also used as a *localized* indicator of textual quality in order to remove anomalous sections of otherwise high-quality texts. The end source text was ensured to be of comparable quality to the post-OCR text presented in "text-mode" on the Google Books website.

Of course, OCR errors are not entirely random; certain character combinations can be particularly difficult to distinguish from other character combinations. For instance, the long s character, common in the first few years of the 19th century, is often mistaken for an “f”. Such typographic confounds must be considered as a matter of course in interpreting timelines.

II.1C. Accuracy of language metadata

We applied additional filters to remove books with dubious language-of-composition metadata. This filter removed volumes whose meta-data language tag disagrees with the language determined by the statistical language detection algorithm described in section 2A. For our English corpus, 8.56% (approximately 235,000) of the books were filtered out in this way. **Table S1** lists the fraction removed at this stage for our other non-English corpora. Note that there are special issues that arise with the Hebrew corpus, which contains a significant quantity of Aramaic text written in the Hebrew script.

II.1D. Year Restriction

In order to further ensure publication date accuracy and consistency of dates across all our corpora, we implemented a publication year restriction and only retained books with publication years starting from 1550 and ending in 2008. This additional restriction eliminated a significant fraction of misdated books which have a publication year of 0 or dates prior to the invention of printing. The number of books filtered due to this year range restriction is considerably small, usually under 2% of the original number of books.

The fraction of the corpus removed by all stages of the filtering is summarized in **Table S1**. Note that because the filters are applied in a fixed order, the statistics presented below are influenced by the sequence in which the filters were applied. For example, books that trigger both the OCR quality filter and by the language correction filter are excluded by the OCR quality filter, which is performed first. Of course, the actual subset of books filtered is the same regardless of the order in which the filters are applied.

II.2. Metadata based subdivision of the Google Books Collection

II.2A. Determination of language

To create accurate corpora in particular languages that minimize cross-language contamination, it is important to be able to accurately associate books with the language in which they were written. To determine the language in which a text is written, we rely on metadata derived from our 100 bibliographic sources, as well as statistical language determination using the Popat algorithm (**Ref S3**). The algorithm takes advantage of the fact that certain character sequences, such as 'the', 'of', and 'ion", occur more frequently in English. In contrast, the sequences 'la', 'aux', and 'de' occur more frequently in French. These patterns can be used to distinguish between books written in English and those written in French. More generally, given the entire text of a book, the algorithm can reliably classify the book into one of the 32 supported language types. The final consensus language was determined based on the metadata sources as well as the results of the statistical language determination algorithm, with the statistical algorithm being assigned the higher priority.

II.2B. Determination of book subject assignments

Book subject assignments were determined using a book's Book Industry Standards and Communication (BISAC) subject categories. BISAC subject headings are a system for categorizing books based on content developed by the BISAC subject codes committee overseen by the Book Industry Study Group. They are often used for a variety of purposes, such as to determine how books are shelved in stores. For English, 92.4% of the books had at least one BISAC subject assignment. In cases where there were multiple subject assignments, we took the more commonly used subject heading and discarded the rest.

II.2C. Determination of book country-of-publication

Country of publication was determined on the basis of our 100 bibliographic sources; 97% of the books had a country-of-publication assignment. The country code used is the 2 letter code as defined in the *ISO 3166-1 alpha-2* standard. More specifically, when constructing our US versus British English corpora, we used the codes "us" (United States) and "gb" (Great Britain) to filter our volumes.

II.3. Construction of historical n-grams corpora

II.3A. Creation of a digital sequence of 1-grams and extraction of n-gram counts

All input source texts were first converted into UTF-8 encoding before tokenization. Next, the text of each book was tokenized into a sequence of 1-grams using Google's internal tokenization libraries (more details on this approach can be found in **Ref. S4**). Tokenization is affected by two processes: (i) the reliability of the underlying OCR, especially vis-à-vis the position of blank spaces; (ii) the specific tokenizer rules used to convert the post-OCR text into a sequence of 1-grams.

Ordinarily, the tokenizer separates the character stream into words at the white space characters (\n [newline]; \t [tab]; \r [carriage return]; " " [space]). There are, however, several exceptional cases:

(1) Column-formatting in books often forces the hyphenation of words across lines. Thus the word "digitized", may appear on two lines in a book as "digi-<newline>ized". Prior to tokenization, we look for 1-grams that end with a hyphen ('-') followed by a newline whitespace character. We then concatenate the hyphen-ending 1-gram to the next 1-gram. In this manner, digi-<newline>tized became "digitized". This step takes place prior to any other steps in the tokenization process.

(2) Each of the following characters are always treated as separate words:

! (exclamation-mark)

@ (at)

% (percent)

^ (caret)

* (star)

((open-round-bracket)

) (close-round-bracket)

[(open-square-bracket)

] (close-square-bracket)

- (hyphen)

= (equals)

{ (open-curly-bracket)

} (close-curly-bracket)

| (pipe)

\ (backslash)

: (colon)

: (semi-colon)

< (less-than)

, (comma)

> (greater-than)

? (question-mark)

/ (forward-slash)

~ (tilde)

` (back-tick)

“ (double quote)

(3) The following characters are not tokenized as separate words:

& (ampersand)

_ (underscore)

Examples of the resulting words include AT&T, R&D, and variable names such as HKEY_LOCAL_MACHINE.

(4) . (period) is treated as a separate word, except when it is part of a number or price, such as 99.99 or $999.95. A specific pattern matcher looks for numbers or prices and tokenizes these special strings as separate words.

(5) $ (dollar-sign) is treated as a separate word, except where it is the first character of a word consisting entirely of numbers, possibly containing a decimal point. Examples include $71 and $9.95

(6) # (hash) is treated as a separate word, except when it is preceded by a-g, j or x. This covers musical notes such as A# (A-sharp), and programming languages C#, J#, and X#.

(7) + (plus) is treated as a separate word, except it appears at the end of a sequence of alphanumeric characters or "+" s. Thus the strings C++ and Na2+ would be treated as single words. These cases include many programming language names and chemical compound names.

(8) ' (apostrophe/single-quote) is treated as a separate word, except when it precedes the letter s, as in ALICE'S and Bob's

The tokenization process for Chinese was different. For Chinese, an internal CJK (Chinese/Japanese/Korean) segmenter was used to break characters into word units. The CJK segmenter inserts spaces along common semantic boundaries. Hence, 1-grams that appear in the Chinese simplified corpora will sometimes contain strings with 1 or more Chinese characters.

Given a sequence of *n* 1-grams, we denote the corresponding n-gram by concatenating the 1-grams with a plain space character in between. A few examples of the tokenization and 1-gram construction method are provided in **Table S2**.

Each book edition was broken down into a series of 1-grams on a page-by-page basis. For each page of each book, we counted the number of times each 1-gram appeared. We further counted the number of times each n-gram appeared (e.g., a sequence of n 1-grams) for all n less than or equal to 5. Because this was done on a page-by-page basis, n-grams that span two consecutive pages were not counted.

II.3B. Generation of historical n-grams corpora

To generate a particular historical n-grams corpus, a subset of book editions is chosen to serve as the base corpus. The chosen editions are divided by publication year. For each publication year, total counts for each n-gram are obtained by summing n-gram counts for each book edition that was published in that year. In particular, three counts are generated: (1) the total number of times the n-gram appears; (2) the number of pages on which the n-gram appears; and (3) the number of books in which the n-gram appears.

We then generate tables showing all three counts for each n-gram, resolved by year. In order to ensure that n-grams could not be easily used to identify individual text sources, we did not report counts for any n-grams that appeared fewer than 40 times in the corpus. (As a point of reference, the total number of 1-grams that appear in the 3.2 million books written in English with highest date accuracy ('eng-all', see below) is 360 billion: a 1-gram that would appear fewer than 40 times occurs at a frequency of the order of 10^{-11}.) As a result, rare spelling and OCR errors were also omitted. Since most n-grams are infrequent, this also served to dramatically reduce the size of the n-gram tables. Of course, the most robust historical trends are associated with frequent n-grams, so our ability to discern these trends was not compromised by this approach.

By dividing the reported counts in a given year by the corpus size in that year (measured either in words, pages, or books), it is possible to determine the normalized frequency with which an n-gram appears in the base corpus in a given year.

Eleven corpora were generated, based on eleven different subsets of books. Five of these are English language corpora, and six are foreign language corpora.

Eng-all

This is derived from a base corpus containing all English language books which pass the filters described in section 1.

Eng-1M

This is derived from a base corpus containing 1 million English language books which passed the filters described in section 1. The base corpus is a subset of the Eng-all base corpus.

The sampling was constrained in two ways.

First, the texts were re-sampled so as to exhibit a representative subject distribution. Because digitization depends on the availability of the physical books (from libraries or publishers), we reasoned that digitized books may be a biased subset of books as a whole. We therefore re-sampled books so as to ensure that the diversity of book editions included in the corpus for a given year, as reflected by BISAC subject codes, reflected the diversity of book editions actually published in that year. We estimated the latter using our metadata database, which reflects the aggregate of our 100 bibliographic sources and includes 10-fold more book editions than the scanned collection.

Second, the total number of books drawn from any given year was capped at 6174. This has the net effect of ensuring that the total number of books in the corpus is uniform starting around the year 1883. This was done to ensure that all books passing the quality filters were included in earlier years. This capping strategy also minimizes bias towards modern books that might otherwise result because the number of books being published has soared in recent decades.

Thus, the Eng-1M corpus more closely resembles a traditional ‘balanced’ corpus. Of course, the balancing algorithm employed is very crude. Further, it must be noted that the types of books that are published has changed quite radically over time; thus one must be wary that crude attempts to ensure ‘balance’ might in fact skew the resulting corpus in ways that made it less reflective of texts actually published at the time. Ultimately, the development of high-throughput methods for balancing large corpora consistent with a particular corpus-builder’s desiderata is an important avenue for future research.

Eng-Modern-1M

This corpus was generated exactly as Eng-1M above, except that it contains no books from before 1800.

Eng-US

This is derived from a base corpus containing all English language books which pass the filters described in section 1; however the OCR quality threshold is set to 60% instead of 80%. In addition we require that the book be published in the United States, as reflected by the 2-letter country code "us" in the country-of-publication metadata.

Eng-UK

This is derived from a base corpus containing all English language books which pass the filters described in section 1; however the OCR quality threshold is set to 60% instead of 80%. In addition we require that the book be published in the United Kingdom, as reflected by the 2-letter country code "gb" in the country-of-publication metadata.

Eng-Fiction

The base corpus used was the subset of the Eng-all base corpus that had a BISAC top level category attribution of "Fiction". Note that this includes many works which are not themselves works of fiction, such as scholarly analyses of fictional works. Because the subject filtering is overly crude, we do not use this corpus for any of the analyses appearing in the paper, but subject-specific subcorpora of this type are an important area for future exploration.

Fre-all

This is derived from a base corpus containing all French language books which pass the series of filters described in section 1.

Ger-all

This is derived from a base corpus containing all German language books which pass the series of filters described in section 1.

Spa-all

This is derived from a base corpus containing all Spanish language books which pass the series of filters described in section 1.

Rus-all

This is derived from a base corpus containing all Russian language books which pass the series of filters described in section 1C-D.

Chi-sim-all

This is derived from a base corpus containing all books written using the simplified Chinese character set which pass the series of filters described in section 1C-D.

Heb-all

This is derived from a base corpus containing all Hebrew language books which pass the series of filter described in section 1.

The computations required to generate these corpora were performed at Google using the MapReduce framework for distributed computing (**Ref S5**). Many computers were used; these computations would take many years on a single ordinary computer.

The size of these base corpora is described in Tables S3-S6.

III. How the analyses were performed

In this section we describe the computational techniques we use to analyze the historical n-grams corpora.

III.0. General Remarks

III.0.1 Sources of bias and error

There is significant variation in the quality of the various corpora during various time periods and their suitability for culturomic research. All the corpora are adequate for the uses to which they are put in the paper. In particular, the primary object of study in this paper is the English language from 1800-2000; this corpus during this period is therefore the most carefully curated of the datasets. However, to encourage further research, we are releasing all available datasets - far more data than was used in the paper. We therefore take a moment to describe the factors a culturomic researcher ought to consider before relying on results of new queries not highlighted in the paper.

1) Volume of data sampled. Where the number of books used to count n-gram frequencies is too small, the signal to noise ratio declines to the point where reliable trends cannot be discerned. For instance, if an n-gram's actual frequency is 1 part in n, the number of words required to create a single reliable timepoint must be some multiple of n. In the English language, for instance, we restrict our study to the years after 1800, where at least 40 million words are found each year. Thus an n-gram whose frequency is 1 part per million can be reliably quantified with single-year resolution. In Chinese, there are fewer than 10 million words per year prior to the year 1956. Thus the Chinese corpus in 1956 is not in general as suitable for reliable quantification as the English corpus in 1800. (In some cases, reducing the resolution by binning in larger windows can be used to sample lower frequency n-grams in a corpus that is too small for single-year resolution.) In sum: for any corpus and any n-gram in any year, one must consider whether the size of the corpus is sufficient to enable reliable quantitation of that n-gram in that year.

2) Composition of the corpus. The full dataset contains about 4% of all books ever published, which limits the extent to which it may be biased relative to the ensemble of all surviving books. The corpus contains mostly books borrowed from participating libraries, and thus its composition reflects library acquisition practices. Still, marked shifts in composition from one year to another are a potential source of error. For instance, book sampling patterns differ markedly for the period before the creation of Google Books (2004) as compared to the period afterward. As a result, we caution users that results from after 2000 are not generally comparable with results from before 2000 and often reflect changes in corpus composition. This was an important reason for our choice of the period between 1800 and 2000 as the target period.

3) Quality of OCR. This varies from corpus to corpus as described above. For English, we spent a great deal of time examining the data by hand as an additional check on its reliability. The other corpora may not be as reliable.

Known issues with OCR include the old spelling for 's' which resembles a modern 'f'. The s/f issue sharply decreases in prevalence in the first years of the nineteenth century.

For simple statistical reasons, short words are much more likely to arise due to OCR mistakes (for instance, the word 'hat' can be the result of an OCR error recognizing an 'a' instead of an 'o' in 'hot') than longer words (such as 'outstanding') or phrases. Thus results for short 1-gram must also be taken with caution.

4) Quality of Metadata. Note that the ability to study the frequency of words or phrases in English over time during the period between 1800 and 2000 was our primary focus in this study. As such, we went to significant lengths to ensure the quality of the general English corpora and their date metadata (i.e., Eng-all, Eng-1M, and Eng-Modern-1M) during this period. Still, the resulting dataset is not perfect. For instance, one metadata provider uses the default publication date '1899' when in doubt. This can cause spurious peaks in this year. Nonetheless, the filtering steps we have described above have managed to significantly reduce the number of systematic errors of this kind.

In contrast, the accuracy of place-of-publication data in English is not as reliable as the accuracy of date metadata. In addition, the foreign language corpora are affected by issues that were improved and largely eliminated in the English data. For instance, their date metadata is not as accurate.

The Hebrew corpus metadata deserves special mention for a number of reasons. First, a significant fraction of the earliest texts annotated as Hebrew are in fact hybrid texts fusing Hebrew and Aramaic, the latter written in Hebrew script (indeed, this Aramaic text is often collinear with the Hebrew text). Second, the Hebrew corpus during the 19th century is composed largely of reprinted works, whose original publication dates far predate the metadata date for the publication of the particular edition in question. These issues arise in part because Hebrew was revived as a spoken modern language during the late 19^{th} century; for many centuries prior to its revival, it was used mostly for scholarly purposes and not for ordinary speech. Thus this corpus can be especially difficult to work with. That said, the corpus spans a period of time in which the language was revived, a rare phenomenon not reflected in the other corpora.

All the above issues will likely improve in the years to come. In the meanwhile, users must use extra caution in interpreting the results of culturomic analyses, especially those based on the various non-English corpora, or based on time-periods where few books are available. Nevertheless, as illustrated in the main text, these corpora already contain a great treasury of useful material, and we have therefore made them available to the scholarly community without delay. We have no doubt that they will enable many more fascinating discoveries.

III.0.2 Measures of n-gram frequency

For each n-gram in the corpus, we provide three measures as a function of year of publication:

the number of times it appeared

the number of pages on which it appeared

the number of books in which it appeared

It is often helpful to normalize these measures by taking the value of a measure in a given year and dividing by the total number of words/pages/books in the corpus in that year.

For the sake of consistency, we only make use of the first measure in the body of the paper. But in fact, the different counts can be used for different purposes. The usage frequency of an n-gram, normalized by the total number of words, reflects both the number of authors using an n-gram, and how frequently they use it. An increase may reflect a broad effect spanning many authors and books, but a marked effect can also be discerned when a rare n-gram is used very frequently by a single author; for instance, a biography of 'Gottlieb Daimler' might mention his name many times. This latter effect is sometimes undesirable. In such cases, it may be preferable to examine the fraction of books containing a particular n-gram: texts in different books, which are usually written by different authors, tend to be more independent. Book counts can be particularly suitable for studying grammatical change, since we might be especially interested in which fraction of the population uses one form or another; because authors rarely publish multiple books in a single year, book frequencies can potentially serve as a proxy for surveying the authors that were alive in a given year.

For example, the **Appendix** includes a complete set of measures for the frequency of the word 'evolution'. In the first three columns, we give the raw counts, the normalized number of times it appeared (relative to the number of words in the corpus in that year), the normalized number of pages it appeared in, and the normalized number of books it appeared in, all as a function of the date.

III.0.3 On the number of books published

In the text, we report that our corpus contains about 4% of all books ever published. Obtaining this estimate relies on knowing how many books are in the corpus (5,195,769) and estimating the total number of books ever published. The latter quantity is extremely difficult to estimate, because the record of published books is fragmentary and incomplete, and because the definition of book is itself ambiguous.

One way of estimating the number of books ever published is to calculate the number of editions in the comprehensive catalog of books which was described in Section I of the supplemental materials. This produces an estimate of 129 million book editions. However, this estimate must be regarded with great caution: it is conservative, and the choice of parameters for the clustering algorithm can lead to significant variation in the results. More details are provided in **Ref S1**.

Another independent estimate we obtained in the study "How Much Information? (2003)" conducted at Berkeley (Ref S6). That study also produced a very rough estimate of the number of books ever published and concluded that it was between 74 million and 175 million.

The results of both estimates are in general agreement. If the actual number is closer to the low end of the Berkeley range, then our 5 million book corpus encompasses a little more than 5% of all books ever published; if it is at the high end, then our corpus would constitute a little less than 3%. We report an approximate value (about 4%) in the text; it is clear that, in the coming years, more precise estimates of the denominator will become available.

III.1. Generation of timeline plots

III.1A. Single Query

The timeline plots shown in the paper are created by taking the number of appearances of an n-gram in a given year in the specified corpus and dividing by the total number of words in the corpus in that year. This yields a raw frequency value. Results are smoothed using a three year window; i.e., the frequency of a particular n-gram in year X as shown in the plots is the mean of the raw frequency value for the n-gram in the year X, the year X-1, and the year X+1.

III.1B. Multiple Query/Cohort Timelines

Where indicated, timeline plots may reflect the aggregates of multiple query results, such as a cohort of individuals or inventions. In these cases, the raw data for each query we used to associate each year with

a set of frequencies. The plot was generated by choosing a measure of central tendency to characterize the set of frequencies (either mean or median) and associating the resulting value with the corresponding year.

Such methods can be confounded by the vast frequency differences among the various constituent queries. For instance, the mean will tend to be dominated by the most frequent queries, which might be several orders of magnitude more frequent than the least frequent queries. If the absolute frequency of the N query results is not of interest, but only their relative change over time, then individual query results may be normalized so that they yield a total of 1. This results in a probability mass function for each query describing the likelihood that a random instance of a query derives from a particular year. The mean value of this collection of probability mass functions in each year may then be computed. (This is equivalent to summing all of the functions and dividing by N.) The resulting function, which we call a 'Mean PMF', may be used to characterize the dynamics of the original set of N queries. This approach eliminates bias due to inter-query differences in frequency, making the change over time in the cohort easier to track.

III.2. Note on collection of historical and cultural data

In performing the analyses described in this paper, we frequently required additional curated datasets of various cultural facts, such as dates of rule of various monarchs, lists of notable people and inventions, and many others. We often used Wikipedia in the process of obtaining these lists. Where Wikipedia is merely digitizing the content available in another source (for instance, the blacklists of Wolfgang Hermann), we corrected the data using the original sources. In other cases this was not possible, but we felt that the use of Wikipedia was justifiable given that (i) the data – including all prior versions - is publicly available; (ii) it was created by third parties with no knowledge of our intended analyses; and (iii) the specific statistical analyses performed using the data were robust to errors; i.e., they would be valid as long as most of the information was accurate, even if some fraction of the underlying information was wrong. (For instance, the aggregate analysis of treaty dates as compared to the timeline of the corresponding treaty, shown in **Figure S1**, will work as long as most of the treaty names and dates are accurate, even if some fraction of the records is erroneous.

We also used several datasets from the Encyclopedia Britannica, to confirm that our results were unchanged when carefully curated, high-quality data was used. For the lexicographic analyses, we relied primarily on existing data from the American Heritage Dictionary.

We avoided doing manual annotation ourselves wherever possible, in an effort to avoid biasing the results. When manual annotation had to be performed, such as in the classification of samples from our language lexica, we tried whenever possible to have the annotation performed by a third party with no knowledge of the analyses we were undertaking

III.3. Controls

To confirm the quality of our data in the English language, we sought controls in the form of n-grams that should exhibit very strong peaks around a date of interest. We used three categories of such n-grams: heads of state ('President Truman'), names of treaties ('Treaty of Versailles'). In addition, we studied geographical name changes ('Byelorussia' to 'Belarus'). In this latter case, the new name should exhibit a strong peak around the year of interest, and the old name should begin to decay. We used Wikipedia to obtain appropriate n-grams, coupled with manual curation.

The list of heads of state included all US presidents and British monarchs who gained power in the 19th or 20th centuries (we removed ambiguous entries, such as 'President Roosevelt'). The list of treaties is taken from the list of 198 treaties signed in the 19th or 20th centuries (**S7**); but we kept only the 121 names that referred to only one known treaty, and that have non-zero timeseries. The list of geographic name changes is taken from **Ref S8**. The lists are given in the **Appendix.**

We recentered each timeline so that year 0 was when the event in question (rise to power of a head of state, signing of a treaty, or changing of a country's name) occurred. We then computed the mean PMFs for each group of queries and plotted the results in **Figure S1.**

The correspondence between the expected and observed presence of peaks was excellent. 42 out of 44 heads of state had a frequency increase of over 10-fold in the decade after they took office (# expected if the year of interest was random: 1 out of 44). Similarly, 85 out of 92 treaties had a frequency increase of over 10-fold in the decade after they were signed (# expected: 2 out of 92). Finally, 23 out of 28 new country names became more frequent than the country name they replaced within 3 years of the name change; exceptions include Kampuchea/Cambodia (the name Cambodia was later reinstated), Iran/Persia (Iran is still today referred to as Persia in many contexts) and Sri Lanka/Ceylon (Ceylon is also a popular tea, potentially confounding this analysis).

We performed controls in the French, German and Spanish languages over the same time period, using the translated list of country name changes described above. Here, too, the correspondence between the expected and observed presence of peaks was excellent.

III.4. Lexicon Analysis

III.4A. Estimation of the number of 1-grams defined in leading dictionaries of the English language.

American Heritage Dictionary of the English Language, 4th Edition (2000)

We are indebted to the editorial staff of AHD4 for providing us the list of the 153,459 headwords that make up the entries of AHD4. However, many headwords are not single words (“preferential voting” or “men’s room”), and others are listed as many times as there are grammatical categories (“to console”, the verb; “console”, the piece of furniture).

Among those entries, we find 116,156 unique 1-grams (such as “materialism” or “extravagate”).

Webster’s Third New International Dictionary (2002)

We obtained the number of “boldface entries” in the dictionary via personal communication with the editorial staff: 476,330. This is the number of n-grams defined in the dictionary. However, we needed to instead estimate the number of 1-grams defined in the dictionary.

The editorial staff also communicated the number of multi-word headwords (74,000) and the total number of headwords (275,000); thus ~73% of headwords were 1-grams.

We used this fraction to estimate the number of boldface entries which were 1-grams: 0.73*476,330, or approximately 348,000.

Oxford English Dictionary, 2nd Edition (1989)

The OED website notes that the “number of word forms defined and/or illustrated” is 615,100 and that of these, 169,000 are “italicized-bold phrases and combinations”.

We used this to estimate an upper bound on the number of unique 1-grams defined by this dictionary: 615,100-169,000 = 446,000.

III.4B. Estimation of Lexicon Size

How frequent does a 1-gram have to be in order to be considered a word? We chose a minimum frequency threshold for ‘common’ 1-grams by attempting to identify the largest frequency decile that remains lower than the frequency of most dictionary words.

We plotted a histogram showing the frequency of the 1-grams defined in AHD4, as measured in our year 2000 lexicon. We found that 90% of 1-gram headwords had a frequency greater than 10^{-9}, but only 70% were more frequent than 10^{-8}. Therefore, the frequency 10^{-9} is a reasonable threshold for inclusion in the lexicon. (See **Fig. S3.**)

To estimate the number of words, we began by generating the list of common (i.e., frequency > 10^{-9}) 1-grams at 11 different time points spanning the period from 1900 until 2000 (1900, 1910, 1920, ... 2000). We excluded all 1-grams with non-alphabetical characters.

For three of the time points (1900, 1950, 2000), we took a random sample of 1000 forms from the resulting lists. We then asked a native English speaker to classify the candidate words into one of 8 categories:

M if the word is a misspelling or a typo or seems like gibberish*

N if the word derives primarily from a personal or a company name

P for any other kind of proper nouns

H if the word has lost its original hyphen

F if the word is a foreign word not generally used in English sentences

B if it is a 'borrowed' foreign word that is often used in English sentences

R for anything that does not fall into the above categories

U unclassifiable for some reason

Beyond this description of the task, the speaker had no knowledge of the analysis that was being performed. The results of the classification task are found in **Appendix**.

[Note that a *typo* is a one-time typing error by someone who presumably knows the correct spelling (as in *improtant*); a *misspelling*, which generally has the same pronunciation as the correct spelling, arises when a person is ignorant of the correct spelling (as in *abberation*).]

We computed the fraction of these 1000 words at each time point that were classified as **P, N**, **B**, or **R**, which we call the 'word fraction for year X', or WF_X. To compute the estimated lexicon size for 1900, 1950, and 2000, we multiplied the word fraction by the number of common alphabetical forms in those years.

For the other 8 time points, we did not perform a separate sampling step. Instead, we estimated the word fraction by linearly interpolating the word fraction of the nearest sampled time points; i.e., the word fraction in 1920 satisfied WF_{1920}=.WF_{1900}+.4*(WF_{1950}.- WF_{1900}). We then multiplied the word fraction by the number of alphabetical forms in the corresponding year, as above.

We repeated the sampling and annotation process using a second native speaker for the year 2000 lexicon. The results were similar, which confirmed that our findings were independent of the person doing the annotation.

Figure S4 shows the estimates of the lexicon excluding the category 'N', which contains most proper nouns. The presence or absence of these proper nouns has no bearing on the trends we observe.

III.4C. Dictionary Coverage

To determine the coverage of the OED and Merriam-Webster's Unabridged Dictionary (MWD), we performed the above analysis on randomly generated subsets of the lexicon in eight frequency deciles (ranging from 10^{-9} – 10^{-8} to 10^{-3} – 10^{-2}). The samples contained 500 candidate words each for all but the

top 3 deciles; the samples corresponding to the top 3 deciles (10^{-5} – 10^{-4}, 10^{-4} – 10^{-3}, 10^{-3} – 10^{-2}) contained 100 candidate words each.

A native speaker with no knowledge of the experiment being performed determined which words from our random samples fell into the P, B, or R categories (to enable a fair comparison, we excluded the N category from our analysis as both OED an MWD exclude them). The annotator then attempted to find a definition for the word in both the online edition of the Merriam-Webster Unabridged Dictionary or in the online version of the Oxford English Dictionary's 2nd edition. Notably, the performance of the former was boosted appreciably by its inclusion of Merriam-Webster's Medical Dictionary. Results of this analysis are shown in **Appendix**. For each decile, we show the fraction of words for which a definition was found.

To estimate the fraction of 'dark matter' in the English language, we summed P_{word}*$P_{OED/MWD}$*N_{1gram} over all deciles, where:

N_{1gram} is the number of 1grams in the decile

P_{word} is the fraction of 1-gram in this decile which a words (R, B, or P)

$P_{OED/MWD}$ is the fraction of words in the decile that are defined in OED or MWD.

This results in an estimate of 297,000 'dark matter' words out of a total of 572,000 words in the R, B, and P categories. Note that the size of either OED or MWD, when added to the 'dark matter', exceeds 572,000; this is mostly because these sources define many words which have a frequency lower than 10^{-9}; were we to drop our frequency threshold for inclusion in our lexicon, the 'dark matter' would become even more abundant.

III.4D. Analysis of New and Obsolete words in the American Heritage Dictionary

We obtained a list of the 4804 vocabulary items that were added to the AHD4 in 2000 from the dictionary's editorial staff. These 4804 words were not in AHD3 (1992) – although, on rare occasions a word could have featured in earlier editions of the dictionary (this is the case for "gypseous", which was included in AHD1 and AHD2).

Similar to our study of the dictionary's lexicon, we restrict ourselves to 1-grams. We find 2077 1-grams newly added to the AHD4. Median frequency (**Fig 2D**) is computed by obtaining all frequencies of this set of words and computing its median.

Next, we ask which 1-grams appear in AHD4 but are not part of the year 2000 lexicon any more (frequency lower than one part per billion between 1990 and 2000). We compute the lexical frequency of the 1-gram headwords in AHD, and find a small number (2,220) that are not part of the lexicon today. We show the mean frequency of these 2,220 words (**Fig 2F**).

III.5. The Evolution of Grammar

III.5A. Ensemble of verbs studied

Our list of irregular verbs was derived from the supplemental materials of **Ref 18** (main text). The full list of 281 verbs is given in the **Appendix.**

One of our objectives in this section is to study the way word frequency affects the dynamics of irregular and regular forms. As such we filter out two classes of verbs:

Verbs whose regular or irregular forms are ambiguous, such as “dive/dove”: “dove” is a common noun that can refer to a bird. Similarly, in the case of “bet/bet”, the present tense and past tense forms cannot be distinguished. Because the current version of the corpus does not include features such as part-of-speech tagging, we are unable to accurately estimate the frequency of the relevant form using our data.

Irregular verbs that are compounds, such as “overpay” or “unbind”, often parasitize the memory effect of the more frequent verb to which they are associated. As such, it the usage frequency of the compound form does not reflect the frequency of the pattern as a whole, or its hold on a person’s memory.

What remains is a list of 106 verbs that we use in our study (marked by the denomination ‘True’ in the column “Use in the study?” in the **Appendix**)

III.5B. Computing verb frequency and regularity

For each verb, we computed the frequency of the regular past tense form (i.e., the form generated by appending the –ed suffix), and the frequency of the irregular past tense (summing preterit and past participle). Examples are shown in **Fig 3A** and **Fig S5**.

The regularity of a verb is the fraction of times that the regular conjugation is used out of all occasions in which the verb is conjugated into the past tense. Thus the regularity of a verb in a given year is r=R/(R+I) where R is the number of times the regular past tense form was used, and I the number of times the irregular past tense form was used. Regularity is a continuous variable that ranges between 0 and 1 (or 0% - 100%).

In **Figure 3B**, we display each verb, determining its position along the x-axis based on its mean regularity between 1800 and 1825, and determining its position along the y-axis based on its mean regularity between 1975 and 2000.

III.5C. Translation of verb frequencies into population statistics

We occasionally use the verb frequency results to estimate the fraction of the population using one form or the other. These rough estimates are based on two simplifying assumptions: (i) all speakers are equally likely to use every verb, and (ii) each speaker uses only one of the two forms.

Using this technique, we can further estimate the number of speakers who are adopting one form or another. Here we make the additional assumption that the changes we observe are driven largely by speakers adopting one form over another, rather than by the entry of new speakers using a particular form. Note that we do not mean to suggest that changes in population size, and in particular differences between the forms preferred by speakers entering and exiting the population, do not play a role.

For instance, the regularity of “sneak/snuck” has decreased from 100% to 50% over the past 50 years, or approximately 1% per year. Since the population of US English speakers is roughly 300 million, this corresponds to 3 million speakers per year, or about six speakers per minute.

III.5D. Classification of Verbs

The verbs were assigned to different classes using the method of **Ref 18** (main text). **Fig 3C** shows the median regularity for the verbs ‘burn’, ‘spoil’, ‘dwell’, ‘learn’, ‘smell’, and ‘spill’ in each year.

III.6. Collective Memory

One hundred timelines were generated for every year between 1875 and 1975. Amplitude for each plot was measured by either computing ‘peak height’ – i.e., the maximum of all the plotted values, or ‘area-under-the curve’ – i.e., the sum of all the plotted values. The peak for year X always occurred within several years of year X itself. The lag between a year and its peak is partly due to the length of the authorship and publication process. For instance, a book about the events of 1950 may be written over the period from 1950-1952 and only published in 1953.

For each year, we estimated the slope of the exponential decay from this peak value. The exponent was estimated by plotting the frequency over time on a logarithmic axis and estimating the slope between the year Y+5 and the year Y+25. This estimate is not dependent on the specific choice of interval, as long as the interval begins after the peak and ends sometime in the 50 years that follow. See inset of **Figure 4A**.

Another method of estimating half-life is to ask how many years must elapse after the peak before the frequency drops below half of the peak value. These values exhibit the same trend as in **Figure 4A, Inset** (not shown), but are noisier.

III.7. Analysis of Fame

We study the fame of individuals with biographical records in the databases of Encyclopaedia Britannica and in Wikipedia. Given the encyclopedic objective of these sources, they furnish reasonable sources from which to obtain comprehensive lists of famous individuals. We perform the analyses independently using each source. In both cases, we use the data obtained to produce a database of all individuals born between 1800 and 1980, in particular obtaining records for their full name and year of birth. Because a person might often be referred to by a name other than their full name (for instance, Oliver Joseph Lodge tends to be referred to as Oliver Lodge), we consider many possible names for each person in order to determine which one is most frequently used to refer to the person in question. In particular, we exclude names that might refer to multiple people from consideration. Ultimately, we are able to associate many of these biographical records with a timeseries reflecting mentions of the person in question in books. Finally, we compare the fame of individuals as a function of birth year, and as a function of occupation.

Note that a very significant motivation behind the study of fame is that, unlike other n-grams, such as words, the referent of a personal name tends to remain constant over time and often, with relatively minor modifications to the n-gram, across languages. Thus the problem of ensuring consistency of meaning, which can be very significant when tracking words in general, is more tractable in the case of fame.

III.7A. Curation of Biographical Records and associated fame trajectories

III.7A.1. Create a list of individuals with biographies in Wikipedia.

Overview:

Creating a database of records referring to people born between 1800 and 1980 in Wikipedia.

a. Using the DBPedia framework, find all articles which are members of the categories ‘1700_births’ through ‘1980_births’. Only people born in 1800-1980 are used for the purposes of fame analysis. People born in 1700-1799 are used to identify naming ambiguities as described in section III.7A.7 of this Supplementary Material.
b. For all these articles, create a record identified by the article URL, and append the birth year.

c. For every record, use the URL to navigate to the online Wikipedia page. Within the main article body text, remove all HTML markup tags and perform a word count. Append this word count to the record.
d. For every record, use the URL to determine the page's traffic statistics for the month of March 2010. Append the number of views to the record.

Detailed Description: Wikipedia contains many biographies of people. In this step, we generate a database of biographical Wikipedia articles, associate each record with the birth year of the profiled individual, and further associate each record with additional statistics enabling us to roughly assess the importance of the person profiled.

We identified biographical records by using the DBPedia engine (**Ref S9**), a relational database created by parsing Wikipedia. For our purposes, the most relevant component of DBPedia is the "Categories" relational database. These categories are derived from Wikipedia category headings, each of which encompasses many articles related to a specific topic. The DBPedia "Categories" database includes, for all articles within Wikipedia, a complete listing of the categories of which this article is a member. As an example, the article for Albert Einstein (http://en.wikipedia.org/wiki/Albert_Einstein) is a member of 73 categories, including "German physicists", "American physicists", "Violinists", "People from Ulm" and "1879_births". Likewise, the article for Joseph Heller (http://en.wikipedia.org/wiki/Joseph_Heller) is a member of 23 categories, including "Russian-American Jews", "American novelists", "Catch-22" and "1923_births".

Some of the biographical records refer to fictional people. We recognize articles as referring to non-fictional people by their membership in a "year_births" category. The category "1879_births" includes Albert Einstein, Wallace Stevens and Leon Trotsky; "1923_births" includes Henry Kissinger, Maria Callas and Joseph Heller; while "1931_births" includes Michael Gorbachev, Raul Castro and Rupert Murdoch. If only the approximate birth year of a person is known, their article will be a member of a "decade_births" category such as "1890s_births" and "1930s_births". We treat these individuals as if born at the beginning of the decade.

For every parsed article, we append metadata which enables us to assess the importance of the article within Wikipedia, namely the size in words of the article and the number of page views which it obtains. The article word count is created by accessing the article using its URL and parsing the result. The traffic statistics for Wikipedia articles are obtained from http://stats.grok.se/.

Table S7 displays specific examples of records from the resulting database, including name, year of birth, year of death, approximate word count of the main article, and traffic statistics for March 2010.

III.7A.2 – Determine occupation of Wikipedia biographees.

Overview: Associate Wikipedia records of individuals with occupations using relevant Wikipedia "Categories" and "Lists" pages. For every occupation to be investigated:

a. Manually create a list of Wikipedia categories and lists associated with this defined occupation.
b. Using the DBPedia framework, find all the Wikipedia articles which are members of the chosen Wikipedia categories.
c. Using the online Wikipedia website, find all Wikipedia articles which are listed in the body of the chosen Wikipedia lists.
d. Intersect the set of all articles belonging to the relevant Lists and Categories with the set of people born in 1800-1980. For people in both sets, append the occupation information to the person's record.

Detailed Description:

When it is available, we append information about a person's occupation to their record in our database.

Two types of structural elements within Wikipedia enable us to identify the occupations of many of the people listed therein. Wikipedia Categories were described earlier, and contain information pertaining to occupation. The categories "Physicists", "Physicists by Nationality", and "Physicist stubs", along with their subcategories, pinpoint articles that relate to people who's occupation is being a physicist. The second are Wikipedia Lists, special pages dedicated to listing Wikipedia articles which fit a precise subject. For physicists, relevant examples are "List of physicists", "List of plasma physicists", and "List of theoretical physicists". Because they are redundant, these two structural elements may be used in combination to reliably identify the occupation of an individual.

As will be described later, we ultimately selected the top 50 most famous individuals in a variety of occupations. For these individuals, we manually confirmed the occupation assignment by reading the associated Wikipedia article. This eliminated a number of misleading assignments. For instance, "Che Guevara" was listed under Biologists. Although he was a medical doctor by training, this is not the field in which he made his primary historical contribution, and so we manually removed his name from our list of biologists. The most famous individuals of each category born between 1800 and 1920 are given in **Appendix.**

III.7A.3 – Create a list of individuals with biographical information in the Encyclopedia Britannica database.

Overview: Create a database of records referring to people born 1800-1980 in Encyclopedia Britannica.

a. Using the internal database records provided by Encyclopedia Britannica Inc., find all entries referring to individuals born 1700-1980. Only people both in 1800-1980 are used for the purposes of fame analysis. People born in 1700-1799 are used to identify naming ambiguities as described in section III.7A.7 of this Supplementary Material.

b. For these entries, create a record identified by a unique integer containing the individual's full name, as listed in the encyclopedia, and the individual's birth year.
c. For every record, count the number of encyclopedic informational snippets present in the Encyclopedia Britannica dataset, and append this count to the record.

Detailed Description: Encyclopedia Britannica is a hand-curated, high-quality encyclopedic dataset with many detailed biographical entries. Encyclopedia Britannica Inc. provided use with a structured biographical database through a private communication. These datasets include a complete record of all entries relating to individuals in the Encyclopedia Britannica. Each record contains the birth and death year of the person described, as well as set of informational snippets summarizing the most critical biographical information about the person.

We extract all records of individuals born in between 1800 and 1980, retaining the number of biographical snippets as a measure of their notability. **Table S8** displays examples of records resulting from this step of the analysis procedure.

III.7A.4 – Generate orthographic variants associated with the full names of individuals.

Overview: In both databases, we create a set of orthographic name variants associated with each record. To create the set:

a. Include the original full name of the person.
b. If the name includes apostrophes or quotation marks, include a variant where these elements are removed.
c. If the first word in the name contains a hyphen, include a name where this hyphen is replaced with a whitespace.
d. If the last word of the name is a numeral, include a name where this numeral has been removed.
e. For every element in the set which contains non-Latin characters, include a variant where this characters have been replaced using the closest Latin equivalent.

Detailed Description: We ultimately wish to identify the name which is most frequently used to refer to a particular individual in our dataset. Because of the current state of OCR technology and the manner in which our n-gram dataset was created, certain typographic elements such as accents, hyphens or quotation marks may disappear from references to an individual. To compensate, for every full name present in our database of individuals, we append variants of the full name where these typographic elements have been removed or replaced. **Table S9** presents examples of spelling variants for multiple names.

III.7A.5 – Generate possible names used to refer to individuals.

Overview: For every record, use the set of orthographic variants to create a set of possible names that might be used to refer to an individual. Possible names are 2-grams or 3-grams that will be used in order to measure the fame of the individual. The following procedure is iterated on every orthographic variant associated with a given record. Steps for which the record source (Wikipedia or Britannica) is not specified are carried out for both.

a. For Encyclopedia Britannica records, truncate the raw name at the second comma; reorder so that the part of the name preceding the first comma follows the part after the first comma.
b. For Wikipedia records, replace the underscores with whitespaces.
c. If any parenthesis or commas are present, truncate the name at the first parenthesis or comma.
d. Truncate the name at the first instance of one of the words ‘in’, ’In’, ’the’, ’The’, ’of’ or ‘Of’. (Of course, in many cases none of these words appear.)
e. Determine the last name. Moving from word to word in the name, choose the first word with all of the following properties:
    i. It begins with a capitalized letter.
    ii. It is longer than 1 character.
    iii. It does not end in a period.
f. If the word preceding this last name is identified as a common last name prefix (i.e., is a member of the set: 'von', 'de', 'van', 'der', 'de' , ‘d’’, 'al-', 'la', 'da', 'the', 'le', 'du', 'bin', 'y', 'ibn', in either capitalized or non-capitalized versions), then prepend the prefix to the last name; i.e., the last name is a 2-gram consisting of the prefix followed by a space, followed by the word determined in step *e*.
g. If the last name contains capitalized character which is not the first character in the last name, create a variant last name where all capital letters except for the first are replaced by lower-case forms. At the end of this step, we have a list of one or more possible last names associated with a particular orthographic variant of a person’s full name.
h. Identify ‘possible first names’. Moving from word to word through all the words that precede the last name determined in *f*, identify all ‘candidate first names’, i.e., names that satisfy the following properties :
    i. They begin with a capital letter.
    ii. They are longer than 1 character.
    iii. They do not end in a period.

iv. They are not a title (i.e., they are not a member of the set: ‘Archduke’, 'Saint', 'Emperor', 'Empress', 'Mademoiselle', 'Mother', 'Brother', 'Sister', 'Father', 'Mr', 'Mrs', 'Marshall', 'Justice', 'Cardinal', 'Archbishop', 'Senator', 'President', 'Colonel', 'General', 'Admiral', 'Sir', 'Lady', 'Prince', 'Princess', 'King', 'Queen', 'de', 'Baron', 'Baroness', 'Grand', 'Duchess', 'Duke', 'Lord', 'Count', 'Countess', 'Dr')

i. Generate a list of possible names associated with a given orthographic variant by generating all possible pairs of the form {possible first name} {possible last name}

j. Repeat this procedure for every orthographic variant, adding to the master list of possible names each time.

III.7A.6 – Associate each possible name in each record with the trajectory of the corresponding n-gram.

Overview: In the previous step, we generated a list of possible names which may refer to a particular individual. In this step we retrieve the trajectory of the corresponding n-gram.

For each possible name of each record, append the trajectory of the corresponding n-gram from the Eng_all corpus.

III.7A.7 – Identify possible names that are ambiguous and may refer to multiple individuals.

Overview: Identify homonymity conflicts. Homonymity conflicts arise when the possible names of two or more individuals share a 2-gram substring. These conflicts are identified using the following algorithm:

a. For every possible name of every record, generate the set of all 2-gram substrings it contains.

b. For every possible name of every record, intersect the substring set associated with the possible name with the substring set of every possible name of every other record.

c. If the intersection is non-empty, then one of two types of conflicts has occurred.

i. A *bidirectional homonymity conflict* occurs when two possible names associated with different records are identical. This name could potentially be used to refer to both individuals.

ii. A *unidirectional homonymity conflict* occurs when one of the substrings associated with a possible name is itself a possible name for a different record. In this case, the conflicted name can refer to one of the individuals, but is part of a name that can be used to refer to a different person. In such a case, the fame of the latter may be erroneously assigned to the former.

Detailed Description: Certain names are particularly popular and are shared by multiple people. This results in ambiguity, as the same possible name may refer to multiple people. 'Homonymity conflicts' occur between a group of individuals when they share all or part of a name. When these homonymity conflicts arise, the word frequency of a specific name may not reflect references to a unique person, but instead to any one of the members of the conflict group. As such, the word frequency for the possible name cannot be used to tracking the fame of the concerned individuals. We identify homonymity conflicts by finding instances of individuals whose possible names overlap one another. Typical homonymity conflicts are shown in Table S11.

Sometimes, one member of the group is obviously much more famous than the other members. In this case, the homonymity conflict can be resolved. Otherwise, possible names associated with homonymity conflicts must be set aside and cannot be used to analyze fame over time. This procedure is described in the next step.

III.7A.8 – Resolve homonymity conflicts when possible.

Resolve homonymity conflicts.

a. When a possible name is associated with multiple biographical records, conflict resolution is employed to determine whether the conflict can be resolved by determining whether a single biographical record is likely to be responsible for the vast majority of mentions in the corpus, and if so, which one. Conflict resolution is performed differently for Wikipedia and for Britannica.
b. *Wikipedia.* Conflict resolution for Wikipedia records is carried out on the basis of the word count and traffic statistics associated with a given Wikipedia biography. The conflict is resolved as follows:
    i. Compute the total word count of all articles corresponding to records involved in the conflict.
    ii. Determine the total number number of pageviews of all articles corresponding to records involved in the conflict.
    iii. For every record in the conflict, determine the fraction of total words and total pageviews that result from the corresponding article.
    iv. For each record, check whether the corresponding article satisfies all of the following conditions:
        1. Has more words than any other article in the conflict
        2. Has more pageviews than any other article in the conflict
        3. Has over 66% of the total words or 66% of the total pageviews of the entire group.

v. Any record which satisfies these conditions is the 'winner' of the conflict and the possible name is assigned to that record and removed from all other records. (It is impossible for multiple winners to arise.)
vi. If there is no winner, conflict cannot be resolved and the possible name is removed from all of the records.

c. *Encyclopedia Britannica*. Conflict resolution for Encyclopedia Britannica records is carried out on the basis of the number of informational snippets present in the Britannica dataset for a given biographical record.
    i. Find the total number of informational snippets related to all records in the conflict.
    ii. If one record is responsible for more than 66% of all informational snippets, it is the 'winner'. The possible name is assigned to that record and removed from all other records.
    iii. If there is no winner, conflict cannot be resolved and the possible name is removed from all of the records.

Detailed Description: The problem of homonymity limits our ability to associate n-grams with specific people. As such it is a special case of a far more challenging general problem: the plasticity of meaning that words exhibit over time and as a function of context. Because the class of things to which a name can refer tends to be quite restricted, it is sometimes possible to determine the likely referent, even without the extensive context that is often necessary for such disambiguation tasks. Our strategy rests on the fact that, in a small group of individuals, it is sometimes the case that the prior probability of a reference to one of these individuals is far higher than to the rest, i.e., there is reason to believe that, although there may be many people named X, one of them might be much more famous than the rest. For the Britannica database, we use the quantity of information available about an individual as a proxy for fame. For records in Wikipedia, we use both the size of the article about them and the quantity of traffic that it generates. Examples of conflict resolution are shown in **Table S12** and **S13**.

III.7A.9 Determine which of the possible names is most frequently used to refer to the individual described in a given record.

Overview: Determine the 'tracking name' which best tracks the fame of the person described in each biographical record.

a. Rank all the possible names associated with a record in descending order on the basis of the total number of mentions of the possible name from the year of birth until the year 2000.

b. Beginning with the top-ranked possible name and moving downward one at a time, select the first possible name which satisfies all three of the following requirements:
    i. Unambiguously refers to the record (i.e., was not involved in a conflict or was the winner of the conflict)
    ii. The average frequency of the possible name in the window [year of birth – 10 years : year of birth + 10 years] is either less than 10^{-9}, or is an order of magnitude smaller than the average frequency of the possible name from the year of birth till the year 2000.
    iii. (*Wikipedia Only*). The possible name, when converted to a Wikipedia URL by replacing whitespaces with underscores, either directs the browser to the article associate with the record or does not direct the browser to any article at all. If the possible name directs the browser to another article or to a disambiguation page, the query name is rejected.
c. Exception: If step *b* results in the selection of a 2-gram, and another possible name with a lower-rank is a 3-gram which ends with the selected 2-gram (i.e., if the 2-gram is A B then the 3-gram is of the form X A B), and furthermore the total number of mentions of the 3-gram from birth till 2000 is >80% of the total number of mentions of the 2-gram during that period, then the 3-gram is chosen as the tracking name in order to maximize specificity. In case multiple 3-grams satisfy this description, the top-ranked 3-gram is the one chosen.

Detailed Description: Until this point, we have identified a large set of possible names that are associated with each biographical record in our database. We want to identify which of these possible names is most frequently used to refer to the person in question. We will then be able to use the frequency of the corresponding n-gram to track the person's fame over time. This 'tracking name' is identified primarily on the basis of total usage frequency of a possible name. There is one exception: if one of the possible names is a 2-gram and another is a 3-gram which ends with that 2-gram, and in addition the 3-gram is almost (>80%) as frequent as the 2-gram, then we use the full 3-gram as the 'tracking name' in order to increase the specificity of the results. Examples are shown in **Fig S20** and **S21.**

III.7A.10 – Rank the records in descending order of fame

Overview: Assemble cohorts on the basis of a shared record property.

a. Fetch all records which match a specific record property, such as year of birth or occupation.
b. For year-of-birth cohorts, rank individuals in descending order of fame, computed as the average usage frequency of an individual from birth until the year 2000.

c. For occupation cohorts, rank individuals in descending order of fame, computed as usage frequency in 20th best year.

Detailed Description: We use the frequency of the tracking name over time as a metric for the fame of the corresponding individual. In this analysis, we group people according to characteristics such as occupation or birth year.

Figure S19 displays the number of records parsed from Wikipedia and from Encyclopaedia Britannica, as well as the number retained after the ten processing steps described above and which were the basis for the final cohort analyses.

III.7B. Cohorts of fame

For each year, we defined a cohort of the top 50 most famous individuals born that year. Individual fame was determined using the average frequency of the tracking name over all years after one's birth. We compute cohorts separately for the Wikipedia and Encyclopaedia Britannica databases. In **Figure 5**, we used cohorts computed with names from Wikipedia, but the results are similar in either case.

As part of the analysis, we aggregate all trajectories in the cohort; for each year, we defined the frequency of the cohort as a whole using the median value of the frequencies of all individuals in the cohort during that year.

For each cohort, we define:

(1) Age of initial celebrity. This is the first age when the cohort's frequency is greater than 10^{-9}. This corresponds to the point at which the median value of the cohort exceeds the threshold for entry into the "English lexicon" described earlier.

(2) Age of peak celebrity. This is the first age when the cohort's frequency is greater than 95% of its peak value. The threshold is defined in this way in order to diminish the effects of noise.

(3) Doubling time of fame. This is the rate at which fame increases between the 'age of initial celebrity' and the 'age of peak celebrity.' It is measured by fitting an exponential to the timeseries during that interval using the methods of least squares.

(4) Half-life of fame. This is the rate at which fame decreases after passing its peak. For the results shown, the interval examined begins 5 years after the peak and ends 25 years after the peak; the results are robust to modifications of this interval. Again, use the methods of least squares to fit an exponential to the curve.

The results for all four parameters are shown as a function of the birth year of the cohort in **Figure S8**.

As noted above, the results are similar for both Britannica and Wikipedia. However, because Britannica contains fewer individuals, the cohorts from the early 19th century are much noisier. As such we show in **Figure S9** the fame analysis conducted with the cohorts from Britannica, but restrict our analysis to the years 1840-1950.

In **Figure 5E**, we show display the characteristics of fame as a function of occupation using a carotgram. For each occupation, we select the top 25 most famous individuals born between 1800 and 1920. We define the region associated with an occupation as the locus of points within a threshold distance of at least 2 members of the occupational cohort. Representative individuals are shown, marked by last name.

Of course, people leave more behind them than a name. Like her fictional protagonist Victor Frankenstein, Mary Shelley is survived by her creation: Frankenstein took on a life of his own within our collective imagination (**Figure S22**). Such legacies, and all the many other ways in which people achieve cultural immortality, fall beyond the scope of this initial examination.

III.8. History of Technology

A list of inventions invented between 1800 and 1960 was taken from Wikipedia (**Ref S10**).

The year listed is used in our analysis. Where multiple listings of a particular invention appear, the year retained in the list is the one reported in the main Wikipedia article for the invention. (e.g. "Microwave Oven" is listed in 1945 and 1946; the main article lists 1945 as the year of invention, and this is the year we use in our analyses).

Each entry's main Wikipedia page was checked for alternate terms for the invention. Where alternate names were listed in the main article (e.g. **thiamine** or **thiamin** or **vitamin B_1**), the frequency of all terms was examined, and the dominant term was chosen. Where there was no single dominant term (e.g., MSG or monosodium glutamate) the invention was eliminated from the list. If a name other than the originally listed one appears to be dominant, the dominant name was used in the analysis (e.g. between electroencephalograph and EEG, EEG is used).

The trajectory of each invention was normalized using its peak value. Inventions were then grouped into three cohorts (1800-1840, 1840-1880, 1880-1920, and 1920-1960), and the median percentage of the peak frequency was calculated for each cohort for each year following invention. These were plotted in **Fig 4B**.

It is worth noting certain biases that affect this initial examination.

One source of bias is that recent inventions may not have reached their peak yet, and as such we underestimate their adoption time. To rule this out, we show that the vast majority of the inventions we study are decidedly past their peaks in all three cohorts (**Fig S7**). We also repeat the analysis using two different cohorts of inventions (early vs. late 19th century), each of which is tracked for exactly one hundred years after the initial invention. Thus all inventions would be equally susceptible to the problem of ‘insufficient time to peak’ and no bias between the cohorts is introduced. Again, the result is unambiguous and consistent with the findings reported in the paper (**Fig S7**).

It is possible that some older inventions peaked rapidly after they were first developed, and are by now forgotten and not listed in the Wikipedia article at all. This sampling bias would be more extreme for the earlier cohorts, and would therefore tend to exaggerate the lag between invention date and cultural impact in the older invention cohorts. Future analyses would benefit from the use of historical invention lists to control for this effect, but such lists would require exquisite curation, and our approach in this initial paper was to avoid overly manipulating our underlying dataset for fear of introducing bias and reducing the simplicity and transparency of our method.

III.9. Censorship

III.9A. Comparing the influence of censorship and propaganda on various groups

To create panel E of **Fig 6,** we analyzed a series of cohorts; for each cohort, we display the mean of the normalized probability mass functions of the cohort, as described in section III.1B. We multiplied the result by 100 in order to represent the probability mass functions more intuitively, as a percentage of lifetime fame. People whose names did not appear in the cohorts for the time periods in question (1925-1933, 1933-1945, and 1955-1965) were eliminated from the analysis.

The cohorts we gen=erated were based on four major sources, and their content is given in **Appendix.**

The Blacklists of Wolfgang Hermann

The lists of the infamous librarian Wolfgang Hermann were originally published in a librarianship journal and later in *Boersenblatt,* a publishing industry magazine in Germany. They are reproduced in **Ref S11**. A digital version is available on the German-language version of Wikipedia (**Ref S12**). We considered digitizing **Ref S10** by hand to ensure accuracy, but felt that both OCR and manual entry would be time-consuming and error prone. Consequently, we began with the list available on Wikipedia and hired a manual annotator to compare this list with the version appearing in **Ref S11** to ensure the accuracy of the resulting list. The annotator had no knowledge of the nature of our study, did not have access to our data, and made these decisions purely on the basis of the text of **Ref S11**. The following changes were made:

Literature:
1) “Fjodor Panfjorow” was changed to “Fjodor Panferov”.
2) “Nelly Sachs” was deleted.

History:
1) “Hegemann W. Ellwald, Fr. v.” was changed to “W. Hegemann” and “Fr. Von Hellwald”

Art:
4) “Paul Stefan” was deleted.

Philosophy/Religion:
1) “Max Nitsche” was deleted.

The results of this manual correction process were used as our lists for Politics, Literature, Literary History, History, Art-related Writers, and Philosophy/Religion. The lists for Art and Literary History were very short and are therefore not shown in the primary figure, but instead appear in **Fig S12**.

The Berlin list

The lists of Hermann formed the basis for a continually expanding blacklist supported by the Nazi regime. We also analyzed a version from 1938 (**Ref S13)**. This version was digitized by the City of Berlin to mark the 75th year after the book burnings in 2008 (**Ref S14)**. The City of Berlin website, berlin.de, was also generous enough to make the underlying database available to us. The list of authors appearing on the website occasionally included multiple authors on a single line, or errors in which the author field did not actually contain the name of a person who wrote the text. These were corrected by hand to create an initial list.

We noted that many authors were listed only using a last name and a first initial. Our manual annotator attempted to determine the full name of any such author. The results were far from comprehensive, but did lead us to expand the dataset somewhat; names with only first initials were replaced by the full name wherever possible.

Some authors were listed using a pseudonym, and on several occasions our manual annotator was able to determine the real name of the author who used a given pseudonym. In this case, the real name was added to the list.

In addition, we occasionally included multiple spelling variants for a single author. Because of this, and because an author’s real name and pseudonym may both be included on the list, the number of author names on the list very slightly exceeds the number of individuals being examined. The numbers reported in the figure are the number of names on the list.

It is worth pointing out that Adolf Hitler appears as an author of one of the banned books from 1938. This is due to a French version of *Mein Kampf*, together with commentary, which was banned by the Nazi authorities. Although it is extremely peculiar to find Hitler on a list of banned authors, we did not remove Hitler’s name, as we had no basis for doing so from the standpoint of the technical authorship and name criteria described above: Adolf Hitler is indeed listed as the author of a book that was banned by the Nazi regime. This is consistent with our stance throughout the paper, which is that we avoided making judgments ourselves that could bias the outcome of our results. Instead, we relied strictly upon our secondary sources. Because Adolf Hitler is only one of many names, the list as a whole nevertheless exhibits strong evidence of suppression, especially because the measure we retained (median usage) is robust to such outliers.

Degenerate artists

The list of degenerate artists was taken directly from the catalog of a recent exhibition at the Los Angeles County Museum of Art which endeavored to reconstruct the original 'Degenerate Art' exhibition (**Ref S15**).

People with recorded ties to the Nazis

The list of Nazi party members was generated in a manner consistent with the occupation categories in section 7. We included the following Wikipedia categories: *Nazi_leaders, SS_officers, Holocaust_perpetrators, Officials_of_Nazi_Germany,* and *Nazis_convicted_of_war_crimes*, together with all of their subcategories. In addition, the four categories *Nazis_from_outside_Germany, German_Nazi_politicians, Nazi_physicians,* and *Nazis* were included without their respective subcategories.

III.9B. *De Novo* Identification of Censored and Suppressed Individuals

We began with the list of 56,500 people, comprising the 500 most famous individuals born in each year from 1800 – 1913. This list was derived from the analysis of all biographies in Wikipedia described in section 7. We removed all individuals whose mean frequency in the German language corpus was less than 5×10^{-9} during the period from 1925 – 1933; because their frequency is low, a statistical assessment of the effect of censorship and suppression on these individuals is more susceptible to noise.

The suppression index is computed for the remaining individuals using an observed/expected measure. The expected fame for a given year is computed by taking the mean frequency of the individual in the German language from 1925-1933, and the mean frequency of the individual from 1955-1965. These two values are assigned to 1929 and 1960, respectively; linear interpolation is then performed in order to compute an expected fame value in 1939. This expected value is compared to the observed mean frequency in the German language during the period from 1933-1945. The ratio of these two numbers is the suppression index *s*. (When the observed frequency from 1933-1945 was zero, the suppression index was given a finite value of 200)

The complete list of names and suppression indices is included as supplemental data. The distribution of *s* was plotted for using a logarithmic binning strategy, with 100 bins between 10^{-2} and 10^{2}. Three examples of individuals who received scores indicating suppression in German are indicated on the plot by arrows (Walter Gropius, Pablo Picasso, and Hermann Maas).

As a point of comparison, the entire analysis was repeated for English; these results are shown on the plot.

III.9C. Validation by an expert annotator

We wanted to see whether the findings of this high-throughput, quantitative approach were consistent with the conclusions of an expert annotator using traditional, qualitative methods. We created a list of 100

individuals at the extremes of our distribution, including the names of the fifty people with the largest s value and of the fifty people with the smallest s value. We hired a guide at *Yad Vashem* with advanced degrees in German and Jewish literature to manually annotate these 100 names based on her assessment of which people were suppressed by the Nazis (S), which people would have benefited from the Nazi regime (B), and lastly, which people would not obviously be affected in either direction (N). All 100 names were presented to the annotator in a single, alphabetized list; the annotator did not have access to any of our methods, data, or conclusions. Thus the annotator's assessment is wholly independent of our own.

The annotator assigned 36 names to the S category and 27 names to the B category; the remaining 37 were given the ambiguous N classification. Of the names assigned to the S category by the human annotator, 29 had been annotated as suppressed by our algorithm, and 7 as elevated, so the correspondence between the annotator and our algorithm was 81%. Of the names assigned to the B category, 25 were annotated as elevated by our algorithm, and only 2 as suppressed, so the correspondence was 93%.

Taken together, the conclusions of a scholarly annotator assessing the names qualitatively by hand closely matched those of our automated approach. These findings confirm that our computational method provides an effective strategy for rapidly identifying likely victims of censorship given a large pool of possibilities.

III.10. Epidemics

Disease epidemics have a significant impact on the surrounding culture (**Fig. S14 A-C**). It was recently shown that during seasonal influenza epidemics, users of Google are more likely to engage in influenza-related searches, and that this signature of influenza epidemics corresponds well with the results of CDC surveillance (**Ref S16**). We therefore reasoned that culturomic approaches might be used to track historical epidemics. These could help complement historical medical records, which are often extremely hard to obtain.

We examined timelines for 4 diseases: influenza (main text), cholera, HIV, and poliomyelitis. In the case of influenza, peaks in cultural interest showed excellent correspondence with known historical epidemics (the Russian Flu of 1890, leading to 1M deaths, the Spanish Flu of 1918, leading to 20-100M deaths; and the Asian Flu of 1957, leading to 1.5M deaths). Similar results were observed for cholera and HIV. However, results for polio were mixed. The US epidemic of 1916 is clearly observed, but the 1951-55 epidemic is harder to pinpoint: the observed peak is much broader, starting in the 30s and ending in the 60s. This is likely due to increased interest in polio following the election of Franklin Delano Roosevelt in 1932, as well as the development and deployment of Salk's polio vaccine in 1952 and Sabin's oral version in 1962. These confounding factors highlight the challenge of interpreting timelines of cultural interest: interest may increase in response to an epidemic, but it may also respond to a stricken celebrity or a famous cure.

The dates of important historical epidemics were derived from the Cambridge World History of Human Diseases (1993) 3rd Edition.

For cholera, we highlighted only the epidemics which most affected the Western world:

1. 1830-35 (Second Cholera Epidemic)
2. 1848-52, and 1854 (Third Cholera Epidemic)
3. 1866-74 (Fourth Cholera Epidemic)
4. 1883-1887 (Fifth Cholera Epidemic)

Supplementary References

"Quantitative analysis of culture using millions of digitized books", Michel et al.

S1. L. Taycher, "Books of the world stand up and be counted", 2010. http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

S2. Ray Smith, Daria Antonova, and Dar-Shyang Lee, **Adapting the Tesseract open source OCR engine for multilingual OCR**, *Proceedings of the International Conference on Multilingual OCR*, Barcelona Spain, 2009, http://doi.acm.org/10.1145/1577802.1577804

S3. Popat, Ashok. "A panlingual anomalous text detector." DocEng '09: Proceedings of the 9th ACM symposium on Document Engineering, 2009, pp. 201-204.

S4. Brants, Thorsten and Franz, Alex. "Web 1T 5-gram Version 1." LDC2006T13 http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

S5. Dean, Jeffrey and Ghemawat, Sanjay. "MapReduce: Simplified Data Processing on Large Clusters." OSDI '04 p137--150

S6. Lyman, Peter and Hal R. Varian, *"How Much Information",* 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/print.htm#books

S7. http://en.wikipedia.org/wiki/List_of_treaties.

S8. http://en.wikipedia.org/wiki/Geographical_renaming

S9. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann." DBpedia – A Crystallization Point for the Web of Data." Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 2009, pp. 154–165.

S10. http://en.wikipedia.org/wiki/Timeline_of_historic_inventions

S11. Gerhard Sauder: *Die Bücherverbrennung.10. Mai 1933.* Ullstein Verlag, Berlin, Wien 1985.

S12. http://de.wikipedia.org/wiki/Liste_der_verbrannten_Bücher_1933.

S13. Liste Des Schädlichen Und Unerwünschten Schrifttums: Stand Vom 31. Dez. 1938. Leipzig: Hedrich, 1938. Print.

S14. http://www.berlin.de/rubrik/hauptstadt/verbannte_buecher/az-autor.php

S15. Barron, Stephanie, and Peter W. Guenther. *Degenerate Art: the Fate of the Avant-garde in Nazi Germany.* Los Angeles, CA: Los Angeles County Museum of Art, 1991. Print.

S16. Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (2008): 1012-014.

Supplementary Tables

"Quantitative analysis of culture using millions of digitized books", Michel et al.

Table S1. The fraction of the corpus removed during various stages of filtering. Filters are applied in sequence; a book is filtered by the first matching filter it encounters.

Language	Fraction of volumes filtered due to all filters	Fraction of Books filtered due to Serial Killer (1A)	Fraction of Books filtered due to OCR quality (1B)	Fraction of Books filtered due to Language Correction (1C)	Fraction of Books filtered due to year restriction (1D)
English	45.53%	29.45%	10.61%	3.90%	1.56%
French	56.97%	8.94%	44.29%	3.57%	0.17%
Spanish	30.54%	9.50%	16.41%	4.57%	0.06%
German	63.88%	2.98%	49.36%	10.88%	0.66%
Russian	1.88%	0%	0%	1.88%	0%
Chinese	37.08%	0%	0%	37.08%	0%
Hebrew	37.45%	7.13%	15.97%	14.34%	0.01%

Table S2. Examples of tokenization algorithm

Sample Text	Representation as n-gram
I'm seeing the man with the telescope.	[Shown as a series of 2-grams] I ' ‘ m m seeing seeing the the man man with with the the telescope telescope .
[Apostrophes] parents' children's won't it’s St. John’s	 parents ' (2 tokens) children's (1 token) won ' t (3 tokens) it’s (1 token) St . John’s (3 tokens)
[URLs] http://www.google.com	http : / / www . google . com (9 tokens)
[Hyphenation] Mother-in-law	Mother - in - law (5 tokens)
[Prices] $0.99	$0.99 (1 token)
[Numbers] 0.02	0.02 (1 token)
[Double Quotes] “Hello, world”	“ Hello , world ” (5 tokens)

Table S3. Size of the base corpora for the various historical n-grams corpora.

	#Words	#Pages	#Books
eng-all	360,717,742,667	928,281,052	3,288,288
eng-1M	111,452,435,771	299,741,012	1,000,000
eng-modern-1M	112,022,840,823	300,213,207	1,000,000
eng-us	157,659,851,011	402,032,807	1,358,921
eng-uk	51,121,081,518	128,656,551	422,797
spa-all	45,360,124,978	128,560,210	521,768
fre-all	45,197,623,833	123,589,493	389,857
ger-all	37,439,210,527	104,891,090	406,666
rus-all	35,781,205,650	113,205,486	355,852
chi-sim-all	13,439,617,812	75,251,324	196,839
heb-all	2,846,870,172	9,193,246	36,499

Table S4. Number of words in the base corpora as a function of century. Columns denote the first year of the century; for example 1600 represents 1600-1699. "2000" represents the period from 2000-2008.

#Words	1600	1700	1800	1900	2000
eng-all	18,199,198	766,893,675	42,248,037,352	218,937,513,003	98,745,664,789
eng-1M	18,199,198	766,893,675	39,755,701,505	64,774,212,947	6,135,993,796
eng-modern-1M	0	0	39,919,506,888	65,850,721,190	6,252,612,745
eng-us	4,830,321	39,884,018	22,926,851,420	107,683,819,692	27,003,574,219
eng-uk	42,842,141	1,153,687,131	17,494,590,210	24,520,469,907	7,908,366,259
spa-all	3,809,332	323,589,560	3,858,853,932	32,179,073,535	8,993,990,284
fre-all	8,307,032	378,906,074	16,164,449,627	22,061,581,837	6,583,857,322
ger-all	881,396	55,864,117	5,812,359,978	24,084,313,167	7,485,772,895
rus-all	0	6,818,799	2,045,477,289	30,743,688,651	2,985,220,911
chi-sim-all	9,624	58,844	1,270,867	9,818,064,426	3,620,208,047
heb-all	498,844	682,924	122,328,838	2,101,953,851	621,048,116

Table S5. Number of pages in the base corpora as a function of century.

#Pages	1600	1700	1800	1900	2000
eng-all	51,386	2,842,023	115,160,541	563,854,908	246,367,994
eng-1M	51,386	2,842,023	107,984,598	173,659,520	15,199,285
eng-modern-1M	0	0	108,393,169	176,361,396	15,458,642
eng-us	11,055	108,789	57,636,933	278,821,551	65,452,122
eng-uk	113,420	3,700,029	45,402,047	61,539,541	17,897,185
spa-all	13,193	1,216,656	11,461,394	91,798,644	24,067,785
fre-all	32,252	1,502,250	49,664,237	56,559,447	15,829,391
ger-all	2,474	260,596	17,584,619	66,296,308	20,746,939
rus-all	0	69,145	9,723,997	94,465,658	8,946,686
chi-sim-all	2,049	10,040	204,759	56,850,527	18,182,986
heb-all	1,291	2,600	367,103	6,743,198	2,077,881

Table S6. Number of books in the base corpora as a function of century.

#Books	1600	1700	1800	1900	2000
eng-all	216	10,928	338,361	2,035,005	903,759
eng-1M	216	10,928	315,873	617,399	55,564
eng-modern-1M	0	0	317,224	626,399	56,376
eng-us	54	364	173,499	960,774	224,226
eng-uk	423	12,902	122,256	223,001	64,182
spa-all	44	3,317	33,380	386,780	98,235
fre-all	90	3,921	126,595	197,582	61,659
ger-all	4	877	56,838	249,522	99,424
rus-all	0	190	21,262	307,783	26,617
chi-sim-all	4	22	449	149,496	46,861
heb-all	5	11	1,905	27,355	7,217

Table S7: Examples of records parsed from Wikipedia.

Full Name	Views in March 2010	Approximate Word Count	Year of Birth	Year of Death
Alan Turing	54085	12037	1912	1954
Cecil Rhodes	30357	9330	1853	1902
Charles Weissmann	310	334	1931	-
Jean-Baptise Lemire	96	1852	1867	1945
Margaret Thatcher	207904	34466	1925	-
Mark Twain	210905	15696	1835	1910
Niels Bohr	44496	6634	1885	1962
Steve Jobs	369948	15106	1955	-

Table S8 : Examples of records parsed from Encyclopedia Britannica

Full Name	# of Metadata Elements	Date of Birth	Date of Death
Eliot, Charles William	6	20 Mar 1834	22 Aug 1926
Legendre, Adrien-Marie	18	18 Sep 1752	10 Jan 1833
Maxwell, James Clerk	19	13 Jun 1831	5 Nov 1879
Piaf, Edith	8	19 Dec 1915	11 Oct 1963
Schelling, Thomas C.	6	14 Apr 1921	-
Strauss, Johann, the Elder	11	14 Mar 1804	24 Sep 1849
Strauss, Johann, the Younger	10	25 Oct 1825	3 Jun 1899
Vanderbilt, William Henry	15	8 May 1821	8 Dec 1885

Table S9 : Examples of raw name variant creation

Raw Name	Variants
Encyclopedia Britannica	
Aba Novák, Vilmos	Aba Novak, Vilmos
Salih, 'Ali 'Abd Allah	Salih, Ali Abd Allah
Poincaré, Henri	Poincare, Henri
Wikipedia	
Søren Kierkegaard	Soren Kierkegaard
Nicolae Ceaușescu	Nicolae Ceausescu
Henry Ford II	Henry Ford
Jean-Baptiste Lemire	Jean Baptiste Lemire

Table S10: Examples of query names sets

Raw Name	Name after 4f	Query names set
Encyclopedia Britannica		
Thoreau, Henry David	Henry David Thoreau	Henry Thoreau; David Thoreau; Henry David Thoreau
Salih, Ali Abd Allah	Ali Abd Allah Salih	Ali Salih; Abd Salih; Allah Salih
Baker, Houston A., Jr.	Houston A. Baker	Houston Baker
Lloyd Webber, Andrew, Baron Lloyd-Webber of Sydmonton	Andrew Lloyd Webber	Andrew Webber; Lloyd Webber; Andrew Lloyd Webber
MacMillan, Sir Kenneth	Sir Kenneth MacMillan	Kenneth MacMillan; Kenneth Macmillan
Wikipedia		
Thomas Babington Macaulay, 1st Baron Macaulay	Thomas Babington Macaulay	Thomas Macaulay; Babington Macaulay; Thomas Babington Macaulay
Alexandre Dumas, père	Alexandre Dumas	Alexandre Dumas
Sir Charles Wheatstone	Sir Charles Wheatstone	Charles Wheatstne
Mary Johnston (novelist)	Mary Johnston	Mary Johnston
Franklin D. Roosevelt	Franklin D. Roosevelt	Franklin Roosevelet
Douglas MacArthur	Douglas MacArthur	Douglas MacArthur; Douglas Macarthur
Princess Marie Gabriele of Luxembourg	Princess Marie Gabriele	Marie Gabriele
Duke Maximilian Joseph in Bavaria	Duke Maximilian Joseph	Maximilian Joseph

Table S11: Examples of identified conflicts

Conflicted Name	Individuals Concerned	Comments
Encyclopedia Britannica		
Greenleaf Whittier	John Greenleaf Whittier, 1807	
	Greenleaf Whittier Pickard, 1877	
Hermann Muller	Hermann Muller, 1876	
	Hermann Muller, 1890	
	Paul Hermann Muller, 1899	
Cornelius Vanderbilt	Cornelius Vanderbilt, 1794	
	Cornelius Vanderbilt Whitney, 1899	
Wikipedia		
Winston Churchill	Winston Churchill, 1874	Bidirectional all individuals.
	Winston Churchill, 1940 (grandson)	
	Winston Churchill, 1871 (novelist)	
John Maynard	John Maynard, 1969 (cricketer)	The italicized names unambiguously refer to a specific individual (unidirectional conflicts), but matches for John Maynard may come from references to any of these.
	John Maynard, 1959 (actor)	
	John Maynard Keynes, 1883	
	John Maynard Smith, 1920	
	John Maynard Woodworth, 1837	
Harvey Oswald	Harvey Emerson Oswald, 1918	Rare form of conflict, where first and last names match a middle and last name.
	Lee Harvey Oswald**,** 1939	
Abraham Lincoln	Abraham Lincoln, 1809	
	Abraham Lincoln II, 1873	
	Abraham Lincoln, 1744 (captain)	

	Abraham Lincoln Marovitz, 1905	
Emil Hansen	Robert Emil Hansen, 1860	The conflict name is a query name for all three conflicted individuals.
	Emil Christian Hansen, 1842	
	Georg Emil Hansen, 1833	

Table S12 : Resolution of Encyclopedia Britannica conflicts from S11

Conflict	Most likely origin of conflict name matches	Cumulative Information Fraction	Information >66% of total ?	Winner of conflict ?
Greenleaf Whittier	John Greenleaf Whittier, 1807	0.5	No	No
Hermann Muller	Hermann Muller, 1876	0.54	No	No
Cornelius Vanderbilt	Cornelius Vanderbilt Whitney	0.8	Yes	Yes

Table S13 : Resolution of Wikipedia conflicts from S11

Conflict	Most likely origin of conflict name matches	Word count fraction	Traffic fraction	Most words and traffic ?	>66% of words or traffic?	Winner of conflict
Winston Churchill	Winston Churchill, 1874	0.937	0.992	Yes	Yes	Yes
Harvey Oswald	Lee Harvey Oswald,1939	0.980	0.999	Yes	Yes	Yes
John Maynard	John Maynard Keynes, 1883	0.786	0.923	Yes	Yes	Yes
Abraham Lincoln	Abraham Lincoln, 1809	0.887	0.982	Yes	Yes	Yes
Emil Hansen	Emil Christian Hansen	0.335	0.455	No	No	No

Supplementary Figures

“Quantitative analysis of culture using millions of digitized books”, Michel et al.

Figure S1

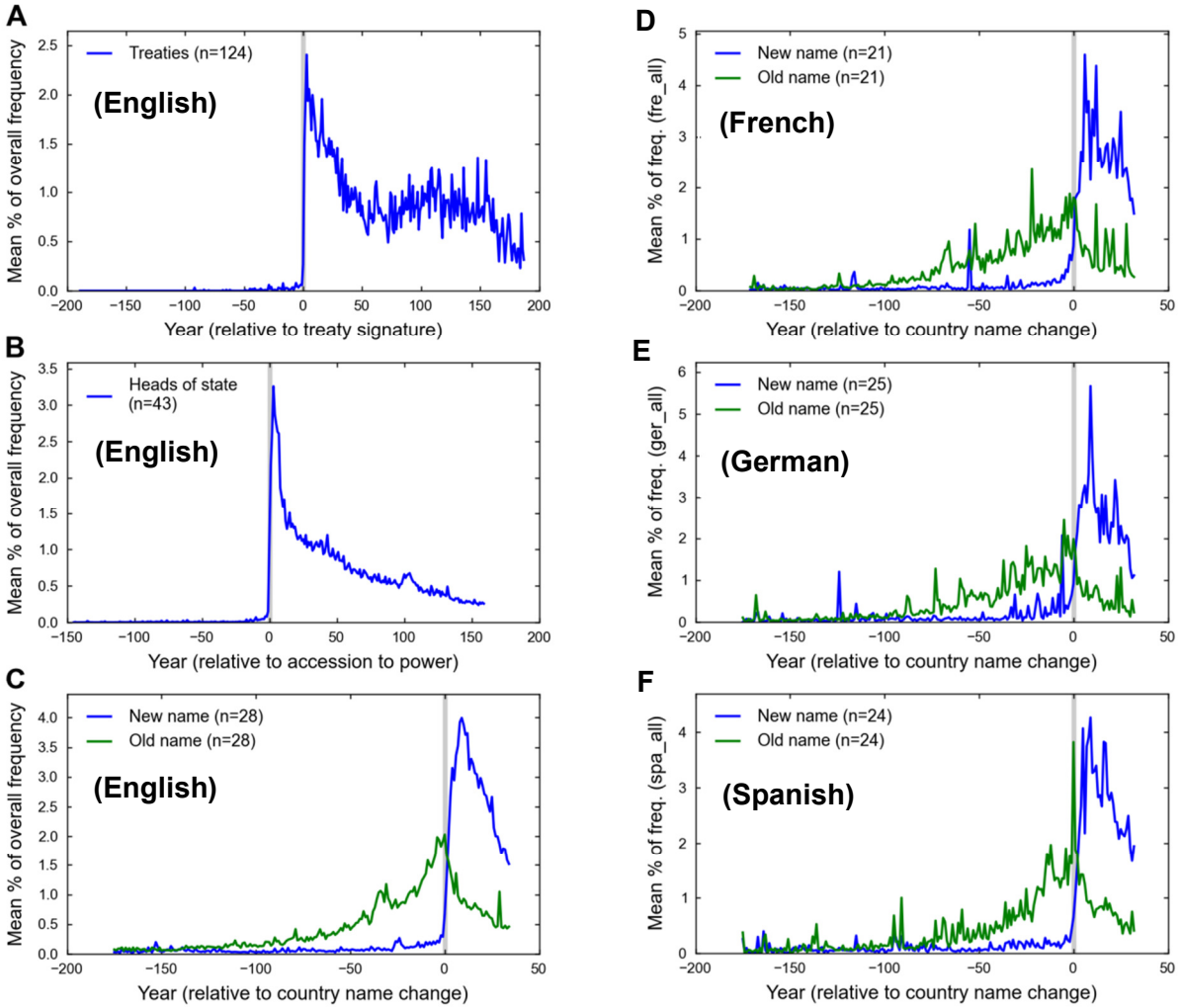

Fig. S1. Controls: known events exhibit sharp peaks at date of occurrence. We can validate the quality of our trajectories in two ways: (i) by examining the data and metadata used in their construction, and (ii) by checking that known events produce reasonable trajectories. Here, we examine events that occurred between 1800 and 2000 produce peaks and transitions at the expected dates. We consider three different types of events: (i) the signing of treaties; (ii) the rise to power of heads of state, and (iii) changes to the names of countries, such as when "Upper Volta" was renamed "Burkina Faso". In the first two cases, we expect little or no signal before the relevant date (serving as a negative control), and a large peak immediately thereafter (serving as a positive control); for the last case, we expect the new name to peak, and the old name to plummet, when the renaming takes place. In each case we show the normalized sum of all the trajectories in a particular category. **(A-C) Controls for books written in English between 1800 and 2000**: **(A)** 124 treaty names; the year of signing is shown in grey. As expected, treaties are rarely discussed before signing, and surge afterward. **(B)** 43 US presidents and UK monarchs; the year in which they rose to power is shown in gray. As expected, people become much more famous upon becoming heads of state. **(C)** 28 instances in which a country's name changed. The

year of the name change is shown in gray; the old name is shown in green, and the new name in blue. The new names surge immediately after the name change; usage of the old name persists, but loses ground. **(D-F) Controls for books written in French, in German, in Spanish, between 1800 and 2000.** We used the same list of country name changes as above, translating where required. In a few cases, one of the translations led to a null result in the database and was consequently removed. The new names surge immediately after the name change; usage of the old name persists, but loses ground.

Figure S2

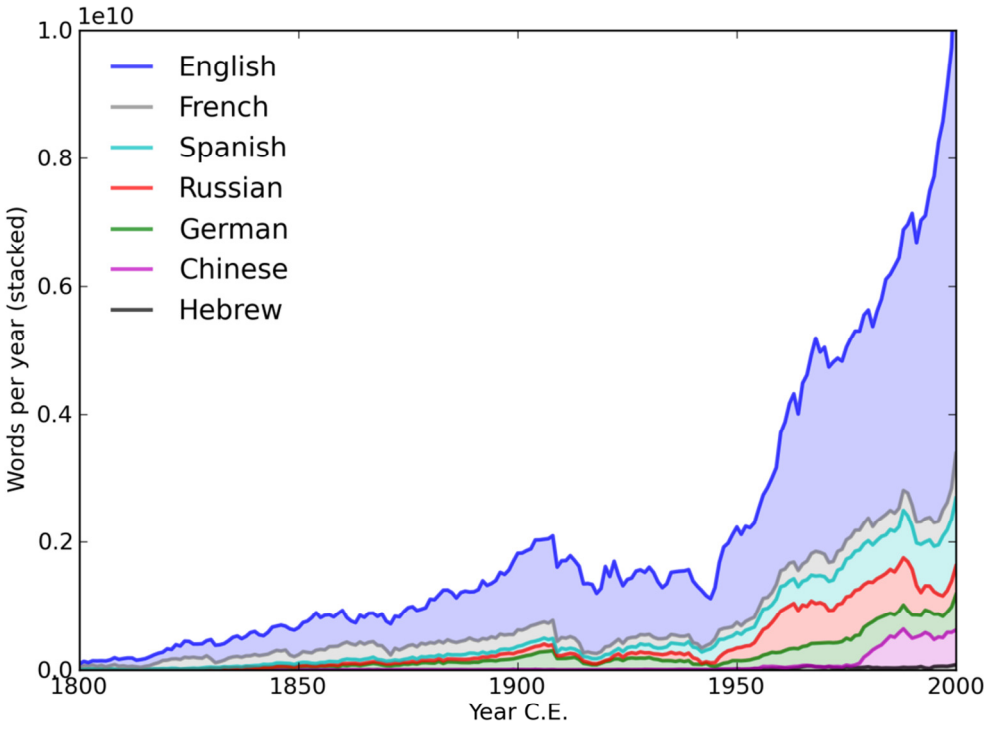

Fig. S2. The size of the corpus as a function of time and of language, 1800-2000. The distributions are stacked on top of one another; thus the total size of all the corpora is reflected by the absolute height of the blue line. The size of the corpus grows rapidly after 1950. The corpus contains 98 million words in 1800, 1.8 billion in 1900, and 11.5 billion in the year 2000. (The corresponding numbers for English alone are 60 million, 1.4 billion, and 8 billion, respectively. There are additional words in the corpus from the period before 1800 and after 2000 (not shown). More words are available in books written before 1800 or after 2000.

Figure S3

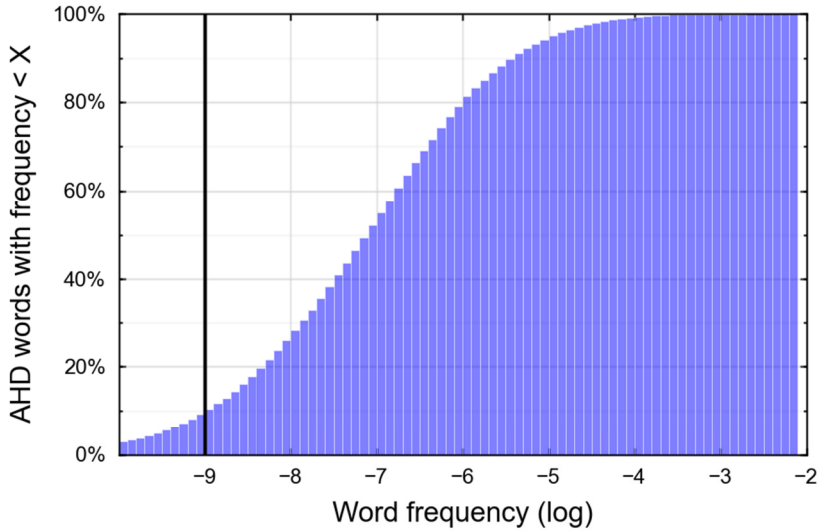

Fig. S3. Frequency distribution of words in the dictionary. We compute the frequency in our year 2000 lexicon for all 116,156 single word (i.e., 1-gram) headwords in the American Heritage Dictionary, 4th Edition (AHD, published in 2000). We show the cumulative distribution: for each frequency value (x-axis), we show the percentage of the words in the AHD whose frequency is smaller than that value. The x-axis is shown on a logarithmic scale using base 10. The cumulative distribution is relatively flat up till 1 part per billion (10^{-9}); 90% of all words in AHD are more frequent than 1 part per billion. Afterwards, it rises much more rapidly; for instance, only 75% of the words are more frequent than 1 part per 100 million (10^{-8}). Thus we use 1 part per billion as a frequency threshold for inclusion in our lexicon.

Figure S4

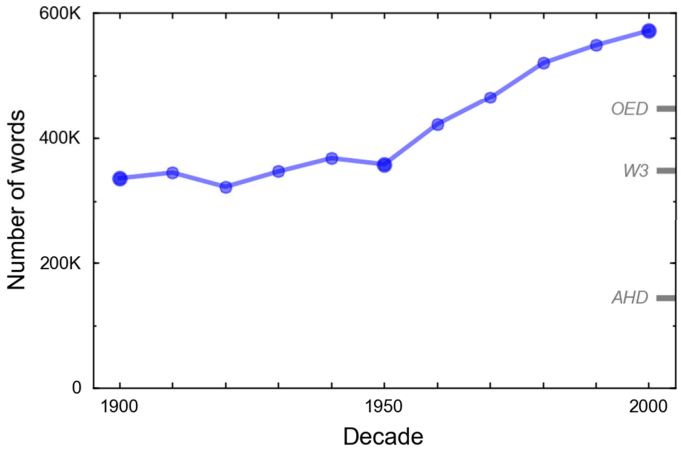

Fig. S4. The size of the English lexicon over time with stricter inclusion criteria. We estimate the number of words exactly as in Figure 1A, but instead exclude proper nouns derived from a person or a company. Our English lexicon still exceed the number of single words (i.e., words which are also 1-grams) listed in OED, W3 or AHD. Even with proper nouns excluded, we observe the same upward trend in the latter half of the century, with the lexicon nearly doubling in size over fifty years.

Figure S5

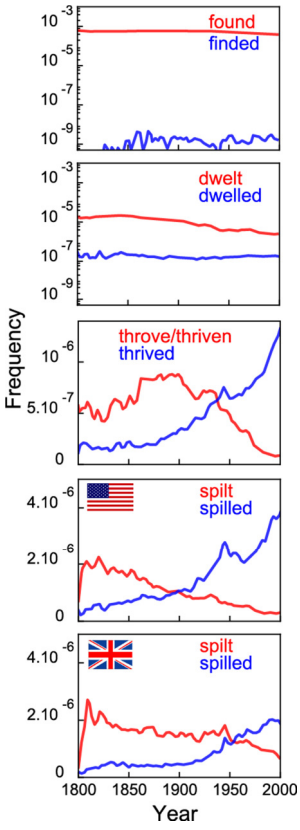

Fig. S5. Quantitative analysis of grammatical trends for past-tense conjugation in English. We track the usage frequency of several irregular verbs (red) and their regular counterparts (blue). Frequent irregulars (find/found) are rarely used in the regularized form (finded); more heterogeneity is seen for infrequent irregulars (dwelt vs. dwelled). Some verbs (thrive/thrived) have regularized during the last two centuries; the trajectory of each regularization event has a unique shape. Verbs do not regularize in all places simultaneously. The verb 'spill' first regularized in the US (US flag) and later in the UK (UK flag).

Figure S6

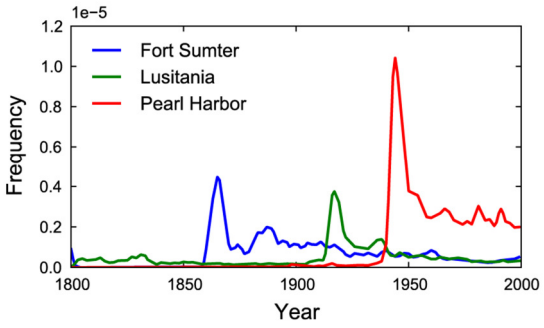

Fig. S6. The persistence of collective memory. Events that precipitated US involvement in three major wars: the Civil War (attack on 'Fort Sumter', 1861, blue), World War I (the sinking of the 'Lusitania', 1915, green), and World War II (the attack on 'Pearl Harbor', 1941, red). Interest in the events is initially very high, but declines sharply as the years go by.

Figure S7

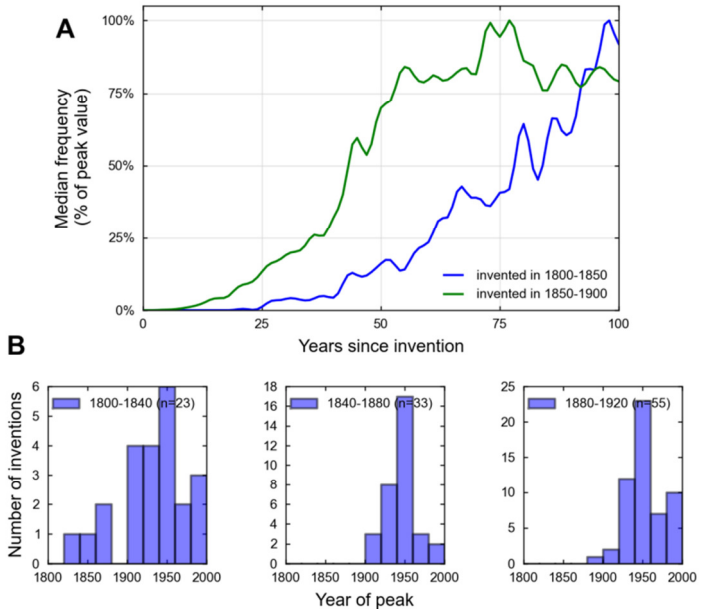

Fig. S7. New technologies spread faster than ever. (A) Two cohorts of inventions that were invented during different periods (blue: 1800-1850; green: 1850-1900) observed over the century following their invention. The more recent cohort rises much faster. **(B)** Most of the inventions studied have already reached their peak. Here we show three histograms corresponding to the three cohorts studied in the main text. The histograms show how many invention from the cohort peaked during various time-periods. The resulting distributions suggest that the vast majority of invention peaked during the period studied (i.e., before 2000).

Figure S8

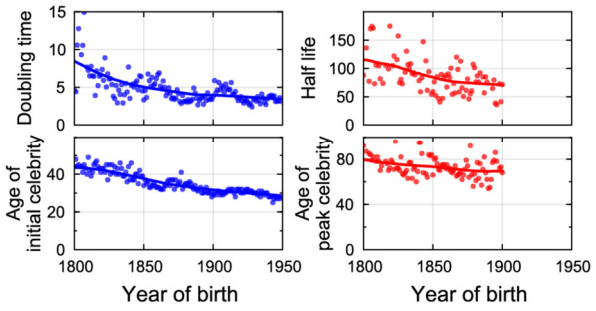

Fig. S8. How to become famous. Values of the four parameters over time. 'Age of peak celebrity' (75 years old) has been fairly consistent. Celebrities are noticed earlier, and become more famous than ever before: 'Age of initial celebrity' has dropped from 43 to 29 years, and 'Doubling time' has dropped from 8.1 to 3.3 years. But they are forgotten sooner as well: the post-peak half-life has declined from 120 years to 71.

Figure S9

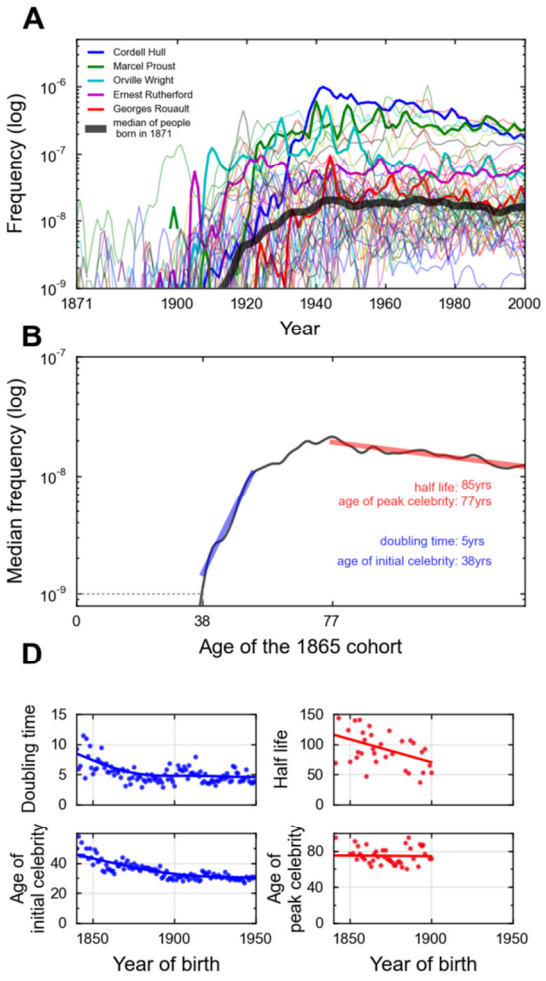

Fig. S9. Analysis of fame using biographical data derived from Encyclopaedia Britannica yields similar results. Here, we use a curated database of biographical information provided by Encyclopaedia Britannica, instead of the Wikipedia-derived database shown in the main text. Data before 1850 is excluded due to small sample size. Results for the four parameters are in agreement with those shown in the main text.

Figure S10

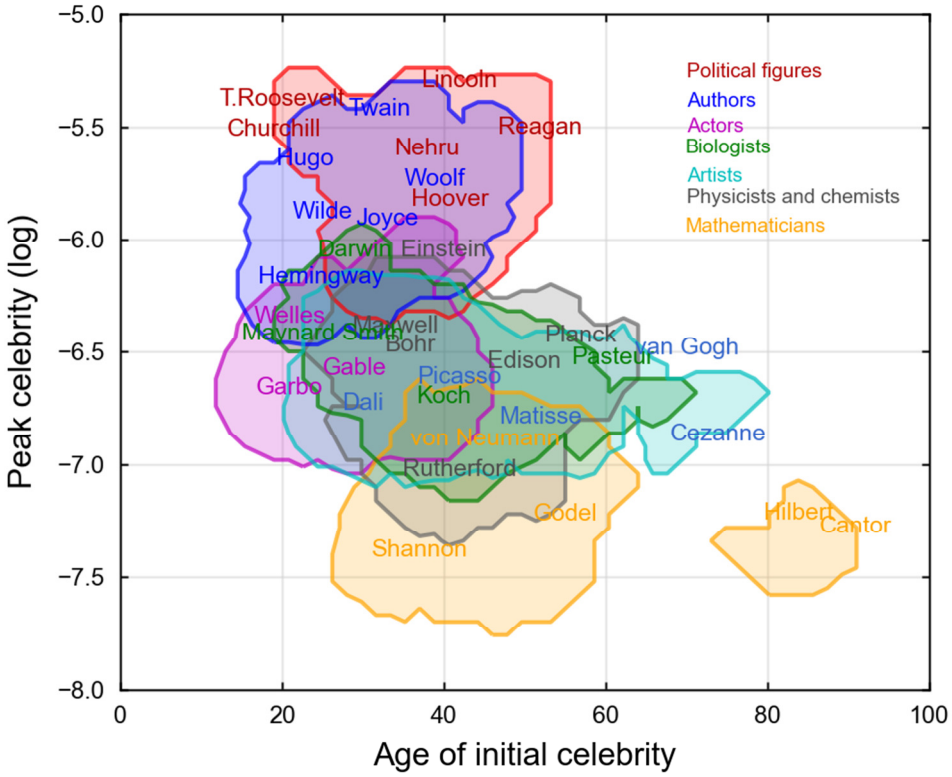

Fig. S10. Fame trade-offs for different occupations. We study the 25 most famous individuals born in 1800-1920, in each of seven occupations. For each, we calculate the age at which they become famous (x-axis), and their fame at the age of peak celebrity (y-axis). A cartogram highlights regions of the plane that are enriched for particular professions, revealing characteristic profiles. We also plot the last name of representative individuals in each occupation.

Figure S11

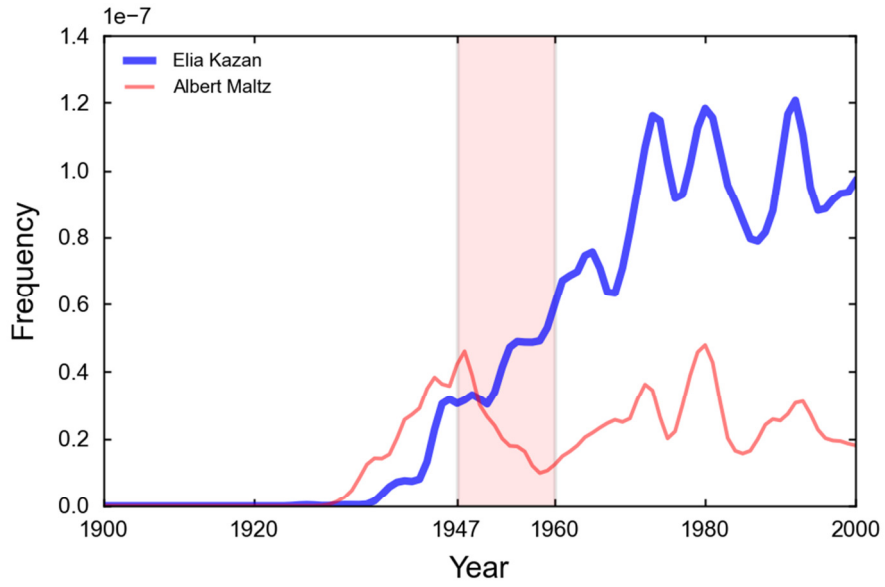

Fig. S11. Moral dilemmas, choices, and their consequences. Up to the time when they were asked to testify before the House Un-American Activities Committee, the fame trajectories of screenwriter Albert Maltz and Director Elia Kazan were fairly similar. But Maltz, after refusing to testify, was blacklisted by American movie studios as one of the 'Hollywood Ten', and his fame declined. Kazan eventually agreed to name names; though his reputation suffered, usage of his own name did not decline, instead rising throughout the period in which the blacklist was in force (red highlight).

Figure S12

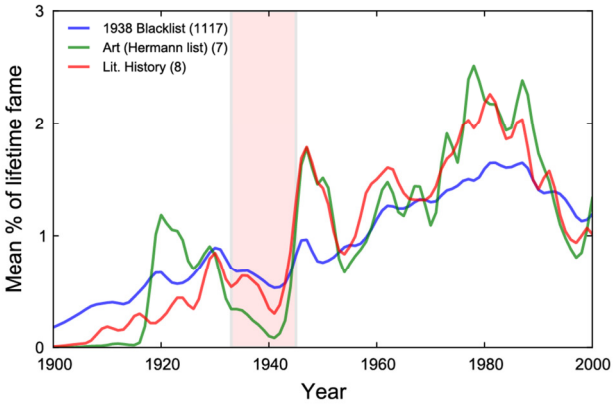

Fig. S12. Effect of censorship in Nazi Germany. We plot the median trajectories of authors found on Herman's lists for Art (green) and Literary History (red). We also plot the median trajectories for people found in the more extensive blacklist of 1938 (blue). The Nazi regime (1933-1945) is highlighted, and corresponds to significant drops in the trajectories of these authors.

Figure S13

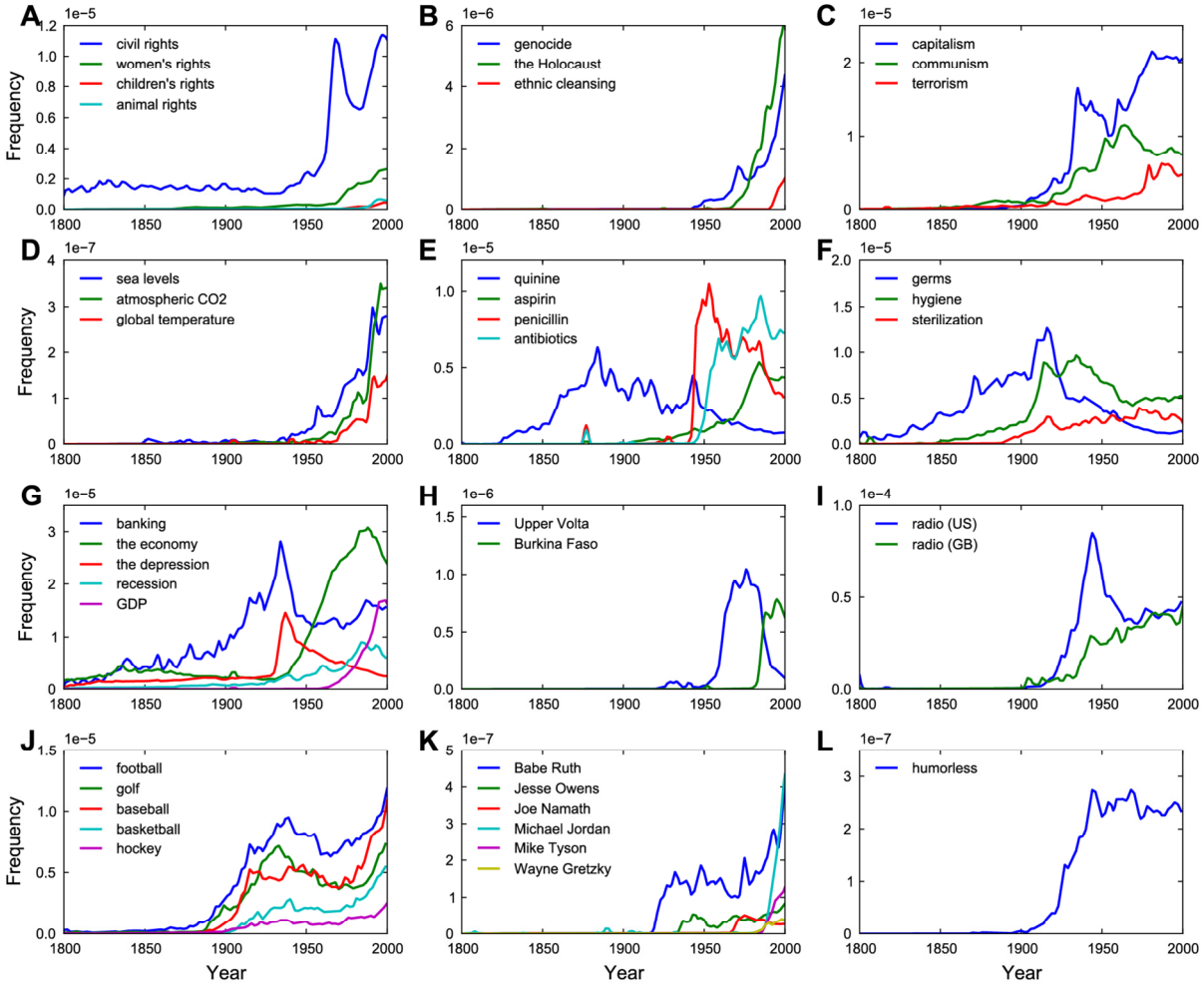

Fig. S13. Culturomic 'timelines' reveal how often a word or phrase appears in books over time. (A) 'civil rights', 'women's rights', 'children's rights' and 'animals rights' are shown. (**B**) 'genocide' (blue), 'the Holocaust' (green), and 'ethnic cleansing' (red) (**C**) Ideology: ideas about 'capitalism' (blue) and 'communism' (green) became extremely important during the 20th century. The latter peaked during the 1950s and 1960s, but is now decreasing. Sadly, 'terrorism' (red) has been on the rise. (**D**) Climate change: Awareness of 'global temperature', 'atmospheric CO2', and 'sea levels' is increasing. (**E**) 'aspirin' (blue), 'penicillin' (green), 'antibiotics' (red), and 'quinine' (cyan). (**F**) 'germs' (blue), 'hygiene' (green) and 'sterilization' (red). (**G**) The history of economics: 'banking' (blue) is an old concept which was of central concern during 'the depression' (red). Afterwards, a new economic vocabulary arose to supplement the older ideas. New concepts such as 'recession' (cyan), 'GDP' (purple), and 'the economy' (green) entered everyday discourse. (**H**) Geographical name changes: 'Upper Volta' (blue) and 'Burkina

Faso' (green). (**I**) 'radio' in the US (blue) and in the UK (red) have distinct trajectories. (**J**) 'football' (blue), 'golf' (green), 'baseball' (red), 'basketball' (cyan) and 'hockey' (purple) (**K**) In the 1980s, the fame of 'Michael Jordan' (cyan) leaped over other that of other great athletes, including 'Jesse Owens' (green), 'Joe Namath' (red), 'Mike Tyson' (purple), and 'Wayne Gretsky' (yellow). Presently, only 'Babe Ruth' (blue) can compete. One can only speculate as to whether Jordan's hang time will match that of the Bambino. (**L**) 'humorless' is a word that became popular during the first half of the 20th century. Such words can serve as a marker that a text was written during a specific period of time.

Figure S14

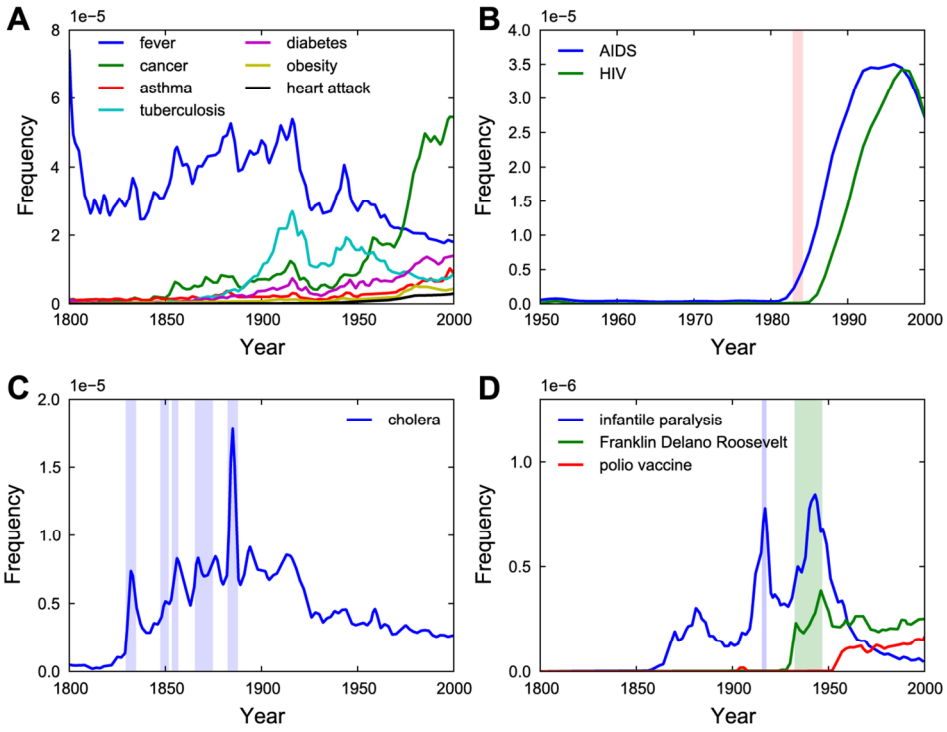

Fig. S14. Tracking historical epidemics using their influence on the surrounding culture. (**A**) Usage frequency of various diseases: 'fever' (blue), 'cancer' (green), 'asthma' (red), 'tuberculosis' (cyan), 'diabetes' (purple), 'obesity' (yellow) and 'heart attack' (black). (**B**) Cultural prevalence of AIDS and HIV. In 1983 (green highlight) it was shown that AIDS was caused by a viral agent, dubbed HIV. (**C**) Usage of the term 'cholera' peaks during the cholera epidemics that affected Europe and the US (blue shading). (**D**) Usage of the term 'infantile paralysis' (blue) exhibits one peak during the 1916 polio epidemic (blue shading), and a second around the time of a series of polio epidemics that took place during the early 1950s. But the second peak is anomalously broad. Discussion of polio during that time may have been fueled by the election of 'Franklin Delano Roosevelt' (green; the period of his presidency is shown in green highlight), as well as by the development of the 'polio vaccine' (red) in 1952. The vaccine ultimately eradicated 'infantile paralysis' in the United States.

Figure S15

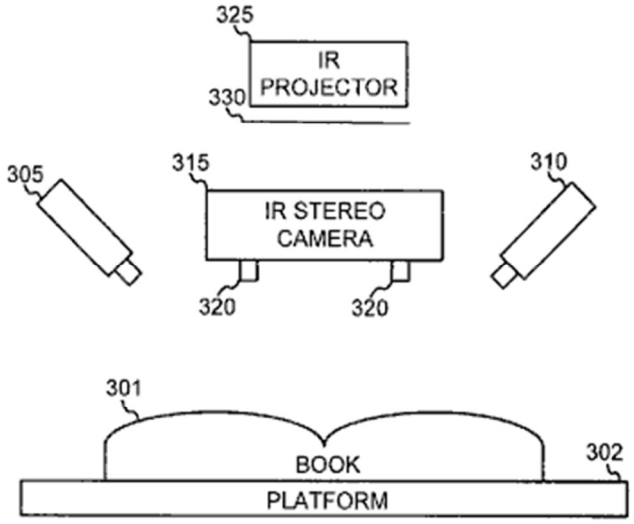

Fig. S15. Schematic of stereo scanning for Google Books.

Figure S16

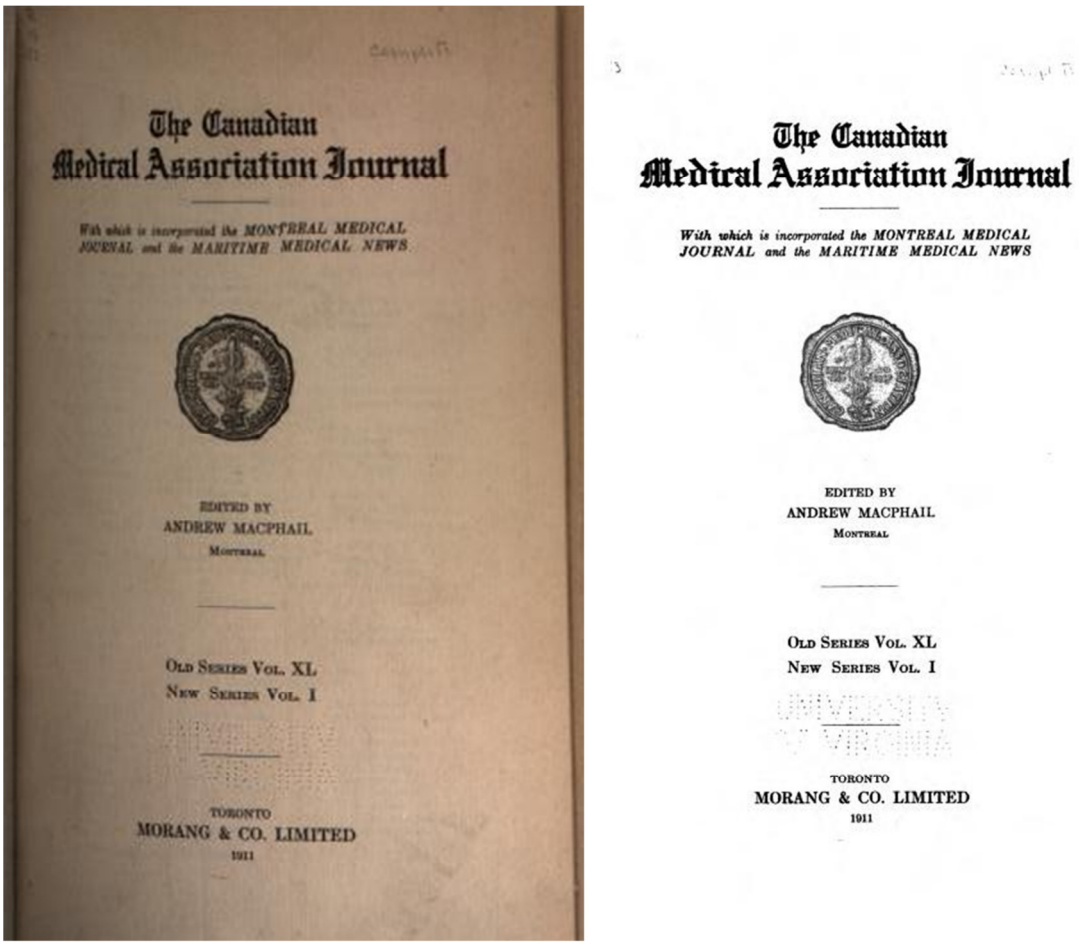

Fig. S16. Example of a page scanned before (left) and after image processing (right).

Figure S17

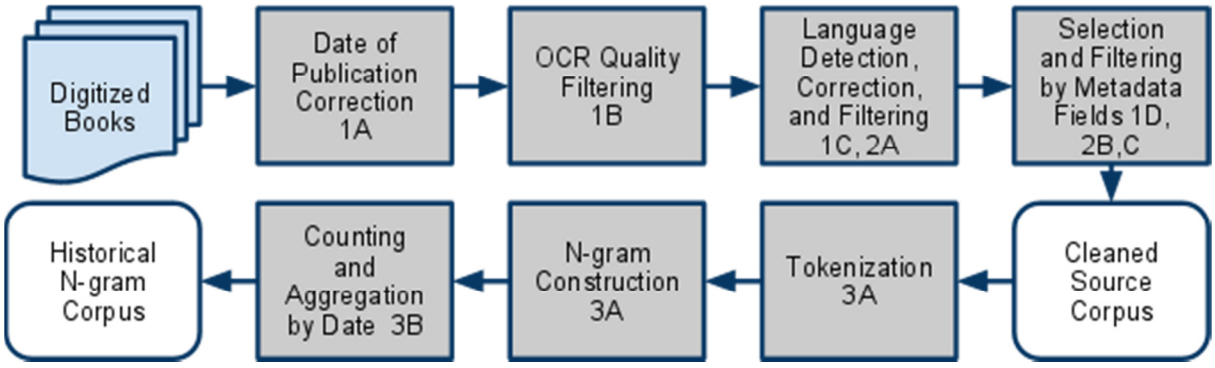

Fig. S17. Outline of n-gram corpus construction. The numbering corresponds to sections of the supplemental text which describe the procedure in detail.

Figure S18

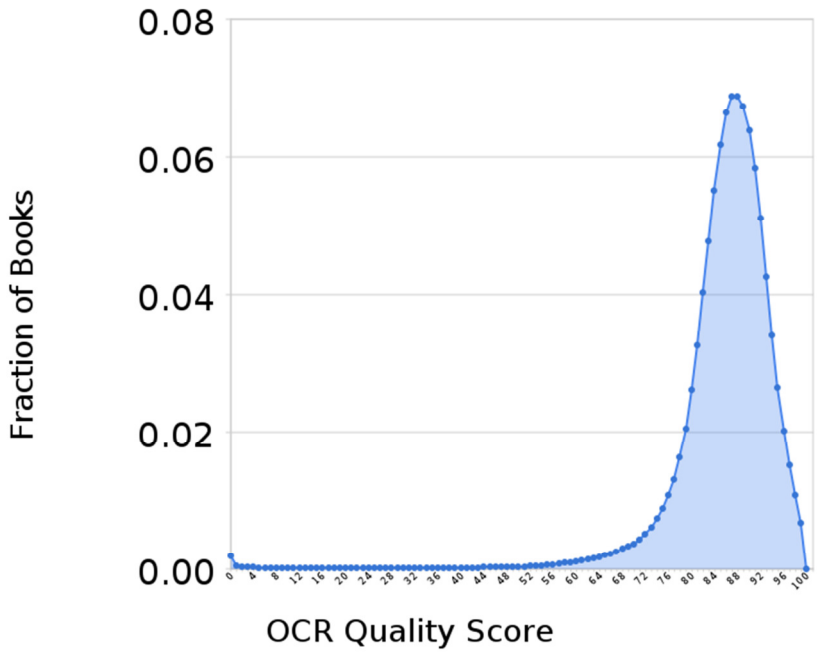

Fig. S18. Distribution of OCR scores in English books. The Google OCR quality score is a measure of the quality of the optical character recognition results for a book. We show here the distribution of these scores for all English books that have been digitized by Google. We use a minimum of OCR cut off of 80 for the Eng_all corpus, and of 60 for Eng-US and Eng-UK.

Figure S19

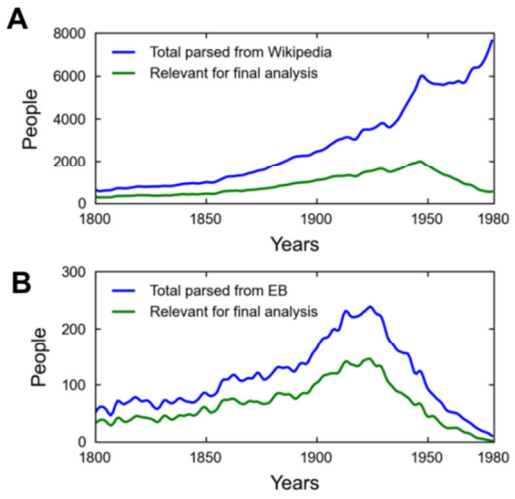

Fig. S19. Generating lists of individuals with unambiguous fame trajectories. The fame analysis was performed independently using two different sources: Wikipedia and Encyclopaedia Britannica. Each source provided a list of individuals; but only a subset of those people turned out to be usable for our analysis. There were many other cases where a person shared a name with a much more famous colleague, and these names were excluded. (Since we studied the most famous people of their era, excluding far less famous individuals with the same name had little impact on our results. Here we show the total number of people listed in each source(blue), and the number that could be unambiguously associated with a specific trajectory (green), as a function of time.

Figure S20

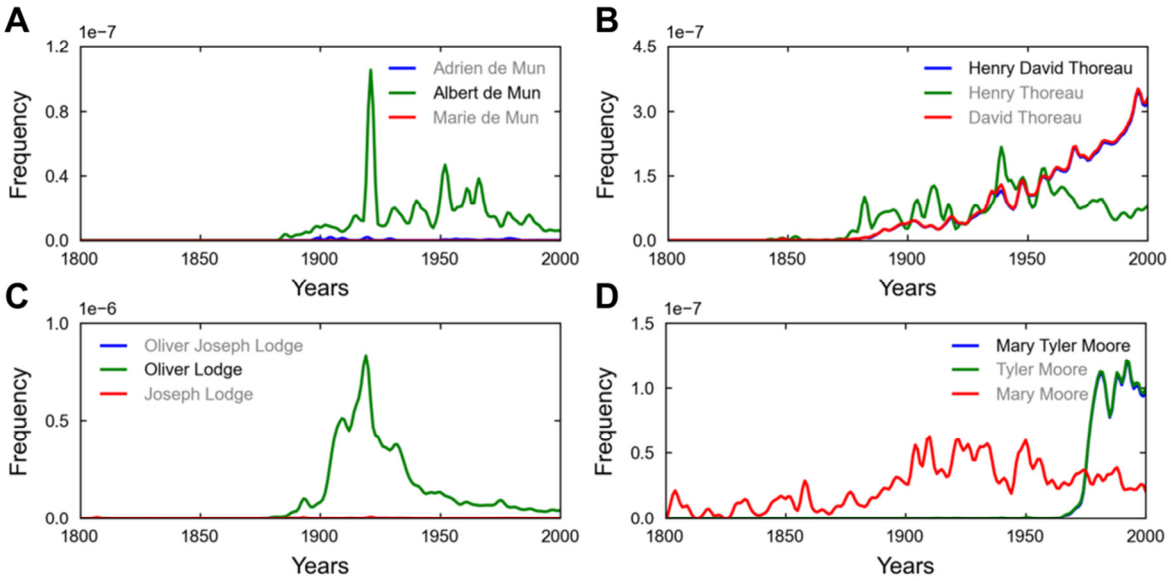

Fig. S20. Selection of query name. The databases from Wikipedia and Encyclopaedia Britannica contain biographical information such as birth date associated with a person's full name, possibly including a title. We must then attempt to determine the name most commonly used to refer to the person in books. This name is called the 'query name'. This step is part of our vetting process, used to associate a biographical record to a trajectory. Here we show four examples; a few of the possible names are shown in grey, the final query name is shown in black. (A) Adrien Albert Marie de Mun. The most frequent possible name is "Albert de Mun", and this is what we use as the query name; (B) Oliver Joseph Lodge. The most frequent possible name is "Oliver Lodge", and this is what we use as the query name; (C) Henry David Thoreau. The most frequent possible name is "David Thoreau", but this name is a substring of "Henry David Thoreau". The latter accounts for more than 80% of the fame of "David Thoreau" fame. To ensure maximal specificity, in such a case we use the full name, "Henry David Thoreau", as the query name. (D) Mary Tyler Moore. The most frequent possible name is "Mary Moore", but this name is rejected as inconsistent with her birthdate. (It has too strong a signal before 1936.) The next most frequent name is "Tyler Moore", but this is a substring of "Mary Tyler Moore", which accounts for over 80% of the fame of "Tyler Moore". Thus "Mary Tyler Moore" is chosen as the query name.

Figure S21

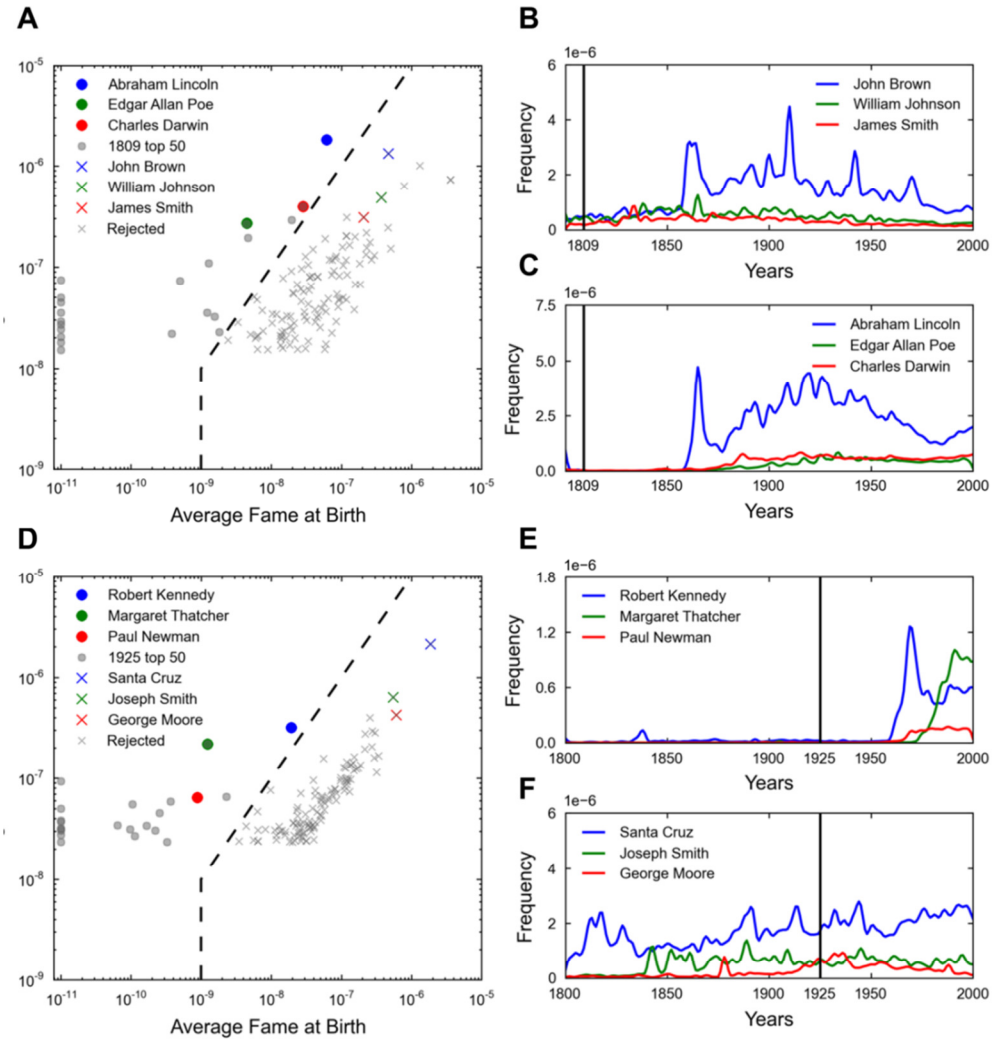

Fig. S21. Filtering out names whose trajectories are inconsistent with the individual's known birthdate. (A,D) The databases from Wikipedia and Encyclopaedia Britannica contain biographical information such as birth date associated with a person's full name. We must determine whether various possible names that could refer to the person in question may in fact be referring to someone else. One helpful technique leverages the fact that a person is very rarely famous at birth, and never before; thus if there are many hits for a name before the person's birth, then it is very likely that there is someone else who is famous and who went by that name. We use this criterion to weed names out as 'ambiguous'; in particular, for most individuals, if the average fame of one of their possible names around the time of birth exceeds one part per billion, that possible name is excluded as ambiguous. However, very famous individuals will appear so often that the small fraction of misdated books in the corpus will cause the frequency of spurious mentions of their names before their birth to exceed one part per billion. To account for this, we do not use the same cutoff rule, and instead rule out possible names on the basis of the ratio of the frequency of the name during their lifetime and around the time

of their birth. Taken together, this strategy results in a separatrix, indicated by the dashed line. The 50 most frequently used possible names derived from people born in 1809 are plotted in **(A)**. Filled circles indicate possible names which are consistent with the known birthdate; crosses indicate possible names which are not consistent with the known birthdate. Examples of the former are shown in **(C)**, and of the latter in **(B)**. The vertical line shows the date of birth appearing in the records. The 50 most frequently used possible names derived from people born in 1925 are plotted in **(D)**, examples of usable names are indicated in **(E),** and of unusable names in **(F)**.

Figure S22

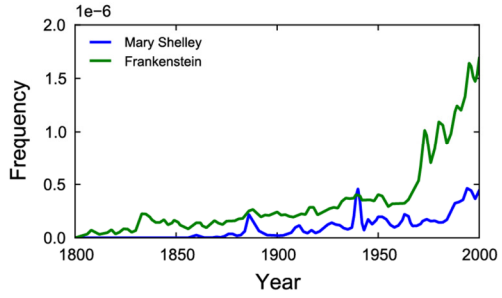

Fig. S22. Many routes to Immortality. People leave more behind them than their name: 'Mary Shelley' (blue) created the monstrously famous 'Frankenstein' (green).

Figure S23

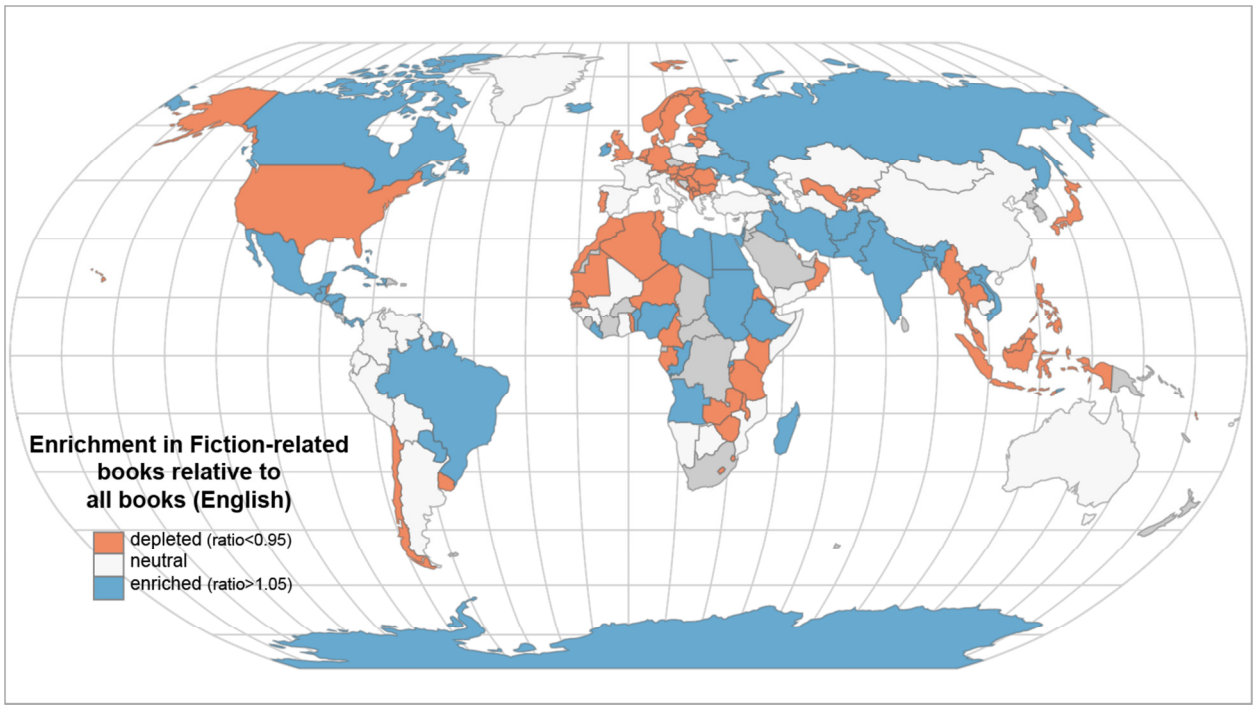

Figure S23. Mapping the human imagination. We measured the average frequency in 1990-2000 of all country names that are single words ('France', but not 'South Africa'; we used USA and UK for proxies in the case of the United States of America and the United Kingdom), comparing the corpus of all English books and in our corpus of Fiction-associated books. We compute the ratio of these two numbers, plotting in blue the countries whose name is enriched in the Fiction-associated corpus (ratio >1.05); and in red those whose name is depleted (ratio <.95). White indicates neutral values; countries in gray were not included in this analysis. Brazil, Antarctica, Egypt and India are enriched in Fiction-associated books; the US, most of Europe and the South-East Asia are underrepresented.