

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Quantitative Analysis of Culture Using Millions of Digitized Books

Citation

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. Quantitative analysis of cutlure using millions of digitized books. Science 331(6014): 176-182.

Published version

https://doi.org/10.1126/science.1199644

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:8899722

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Quantitative analysis of culture using millions of digitized books

Jean-Baptiste Michel[1,2,3,4*†], Yuan Kui Shen[5], Aviva P. Aiden[6], Adrian Veres[7], Matthew K. Gray[8], The Google Books Team[8], Joseph P. Pickett[9], Dale Hoiberg[10], Dan Clancy[8], Peter Norvig[8], Jon Orwant[8], Steven Pinker[4], Martin A. Nowak[1,11,12], Erez Lieberman Aiden[1,12,13,14,15,16*†]

[1] Program for Evolutionary Dynamics,

[2] Institute for Quantitative Social Sciences,

[3] Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA.

[4] Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[5] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA.

[6] Harvard Medical School, Boston, Massachusetts, 02115, USA.

[7] Harvard College, Cambridge, Massachusetts 02138, USA.

[8] Google, Inc., Mountain View, California, 94043, USA.

[9] Houghton Mifflin Harcourt, Boston, Massachusetts 02116, USA.

[10] Encyclopaedia Britannica, Inc., Chicago, Illinois 60654, USA.

[11] Dept of Organismic and Evolutionary Biology,

[12] Dept of Mathematics,

[13] Broad Institute of Harvard and MIT,

[14] School of Engineering and Applied Sciences,

[15] Harvard Society of Fellows,

[16] Laboratory-at-Large, Harvard University, Cambridge, Massachusetts 02138, USA.

(*) these authors contributed equally

(†) To whom correspondence should be addressed. E-mail: jb.michel@gmail.com (JBM); erez@erez.com (EA).

ABSTRACT

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics', focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. 'Culturomics' extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Introduction

Reading small collections of carefully chosen works enables scholars to make powerful inferences about trends in human thought. However, this approach rarely enables precise measurement of the underlying phenomena. Attempts to introduce quantitative methods into the study of culture (1-6) have been hampered by the lack of suitable data.

We report the creation of a corpus of 5,195,769 digitized books containing ~4% of all books ever published. Computational analysis of this corpus enables us to observe cultural trends and subject them to quantitative investigation. 'Culturomics' extends the boundaries of scientific inquiry to a wide array of new phenomena.

The corpus has emerged from Google's effort to digitize books. Most books were drawn from over 40 university libraries around the world. Each page was scanned with custom equipment (7), and the text digitized using optical character recognition (OCR). Additional volumes – both physical and digital – were contributed by publishers. Metadata describing date and place of publication were provided by the libraries and publishers, and supplemented with bibliographic databases. Over 15 million books have

been digitized (12% of all books ever published [7]). We selected a subset of over 5 million books for analysis on the basis of the quality of their OCR and metadata (Fig. 1A) (7). Periodicals were excluded.

The resulting corpus contains over 500 billion words, in English (361 billion), French (45B), Spanish (45B), German (37B), Chinese (13B), Russian (35B), and Hebrew (2B). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 60 million words per year; by 1900, 1.4 billion; and by 2000, 8 billion.

The corpus cannot be read by a human. If you tried to read only the entries from the year 2000 alone, at the reasonable pace of 200 words/minute, without interruptions for food or sleep, it would take eighty years. The sequence of letters is one thousand times longer than the human genome: if you wrote it out in a straight line, it would reach to the moon and back 10 times over (8).

To make release of the data possible in light of copyright constraints, we restricted our study to the question of how often a given '1-gram' or 'n-gram' was used over time. A 1-gram is a string of characters uninterrupted by a space; this includes words ('banana', 'SCUBA') but also numbers ('3.14159') and typos ('excesss'). An n-gram is sequence of 1-grams, such as the phrases 'stock market' (a 2-gram) and 'the United States of America' (a 5-gram). We restricted n to 5, and limited our study to n-grams occurring at least 40 times in the corpus.

Usage frequency is computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year. For instance, in 1861, the 1-gram 'slavery' appeared in the corpus 21,460 times, on 11,687 pages of 1,208 books. The corpus contains 386,434,758 words from 1861; thus the frequency is 5.5×10^{-5}. 'slavery' peaked during the civil war (early 1860s) and then again during the civil rights movement (1955-1968) (Fig. 1B)

In contrast, we compare the frequency of 'the Great War' to the frequencies of 'World War I' and 'World War II'. 'the Great War' peaks between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as 'World War I' (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as 'slavery'). Linguistic change – which, of course, has cultural roots – affects the words we use for those concepts ('the Great War' vs. 'World War I'). In this paper, we will examine both linguistic changes, such as changes in the lexicon and grammar; and cultural phenomena, such as how we remember people and events.

The full dataset, which comprises over two billion culturomic trajectories, is available for download or exploration at www.culturomics.org.

The size of the English lexicon

How many words are in the English language (9)?

We call a 1-gram 'common' if its frequency is greater than one per billion. (This corresponds to the frequency of the words listed in leading dictionaries (7).) We compiled a list of all common 1-grams in 1900, 1950, and 2000 based on the frequency of each 1-gram in the preceding decade. These lists contained 1,117,997 common 1-grams in 1900, 1,102,920 in 1950, and 1,489,337 in 2000.

Not all common 1-grams are English words. Many fell into three non-word categories: (i) 1-grams with non-alphabetic characters ('l8r', '3.14159'); (ii) misspellings ('becuase, 'abberation'); and (iii) foreign words ('sensitivo').

To estimate the number of English words, we manually annotated random samples from the lists of common 1-grams (7) and determined what fraction were members of the above non-word categories. The result ranged from 51% of all common 1-grams in 1900 to 31% in 2000.

Using this technique, we estimated the number of words in the English lexicon as 544,000 in 1900, 597,000 in 1950, and 1,022,000 in 2000. The lexicon is enjoying a period of enormous growth: the

addition of ~8500 words/year has increased the size of the language by over 70% during the last fifty years (Fig. 2A).

Notably, we found more words than appear in any dictionary. For instance, the 2002 Webster's Third New International Dictionary [W3], which keeps track of the contemporary American lexicon, lists approximately 348,000 single-word wordforms (10); the American Heritage Dictionary of the English Language, Fourth Edition (AHD4) lists 116,161 (11). (Both contain additional multi-word entries.) Part of this gap is because dictionaries often exclude proper nouns and compound words ('whalewatching'). Even accounting for these factors, we found many undocumented words, such as 'aridification' (the process by which a geographic region becomes dry), 'slenthem' (a musical instrument), and, appropriately, the word 'deletable'.

This gap between dictionaries and the lexicon results from a balance that every dictionary must strike: it must be comprehensive enough to be a useful reference, but concise enough to be printed, shipped, and used. As such, many infrequent words are omitted. To gauge how well dictionaries reflect the lexicon, we ordered our year 2000 lexicon by frequency, divided it into eight deciles (ranging from $10^{-9} - 10^{-8}$ to $10^{-2} - 10^{-1}$), and sampled each decile (7). We manually checked how many sample words were listed in the OED (12) and in the Merriam-Webster Unabridged Dictionary [MWD]. (We excluded proper nouns, since neither OED nor MWD lists them.) Both dictionaries had excellent coverage of high frequency words, but less coverage for frequencies below 10^{-6}: 67% of words in the $10^{-9} - 10^{-8}$ range were listed in neither dictionary (Fig. 2B). Consistent with Zipf's famous law, a large fraction of the words in our lexicon (63%) were in this lowest frequency bin. As a result, we estimated that 52% of the English lexicon – the majority of the words used in English books – consists of lexical 'dark matter' undocumented in standard references (12).

To keep up with the lexicon, dictionaries are updated regularly (13). We examined how well these changes corresponded with changes in actual usage by studying the 2077 1-gram headwords added to AHD4 in 2000. The overall frequency of these words, such as 'buckyball' and 'netiquette', has soared since 1950: two-thirds exhibited recent, sharp increases in frequency (>2X from 1950-2000) (Fig. 2C). Nevertheless, there was a lag between lexicographers and the lexicon. Over half the words added to AHD4 were part of the English lexicon a century ago (frequency $>10^{-9}$ from 1890-1900). In fact, some newly-added words, such as 'gypseous' and 'amplidyne', have already undergone a steep decline in frequency (Fig. 2D).

Not only must lexicographers avoid adding words that have fallen out of fashion, they must also weed obsolete words from earlier editions. This is an imperfect process. We found 2220 obsolete 1-gram headwords ('diestock', 'alkalescent') in AHD4. Their mean frequency declined throughout the 20th century, and dipped below 10^{-9} decades ago (Fig. 2D, *Inset*).

Our results suggest that culturomic tools will aid lexicographers in at least two ways: (i) finding low-frequency words that they do not list; and (ii) providing accurate estimates of current frequency trends to reduce the lag between changes in the lexicon and changes in the dictionary.

The evolution of grammar

Next, we examined grammatical trends. We studied the English irregular verbs, a classic model of grammatical change (14-17). Unlike regular verbs, whose past tense is generated by adding –ed (jump/jumped), irregulars are conjugated idiosyncratically (stick/stuck, come/came, get/got) (15).

All irregular verbs coexist with regular competitors (e.g., 'strived' and 'strove') that threaten to supplant them (Fig. 2E). High-frequency irregulars, which are more readily remembered, hold their ground better. For instance, we found 'found' (frequency: $5x10^{-4}$) 200,000 times more often than we finded 'finded'. In contrast, 'dwelt' (frequency: $1x10^{-5}$) dwelt in our data only 60 times as often as 'dwelled' dwelled. We defined a verb's 'regularity' as the percentage of instances in the past tense (i.e., the sum of 'drived', 'drove', and 'driven') in which the regular form is used. Most irregulars have been stable for the last 200 years, but 16% underwent a change in regularity of 10% or more (Fig. 2F).

These changes occurred slowly: it took 200 years for our fastest moving verb, 'chide', to go from 10% to 90%. Otherwise, each trajectory was *sui generis*; we observed no characteristic shape. For instance, a few verbs, like 'spill', regularized at a constant speed, but others, such as 'thrive' and 'dig', transitioned in fits and starts (7). In some cases, the trajectory suggested a reason for the trend. For example, with 'sped/speeded' the shift in meaning from 'to move rapidly' and towards 'to exceed the legal limit' appears to have been the driving cause (Fig. 2G).

Six verbs (burn, chide, smell, spell, spill, thrive) regularized between 1800 and 2000 (Fig. 2F). Four are remnants of a now-defunct phonological process that used –t instead of –ed; they are members of a pack of irregulars that survived by virtue of similarity (bend/bent, build/built, burn/burnt, learn/learnt, lend/lent, rend/rent, send/sent, smell/smelt, spell/spelt, spill/spilt, and spoil/spoilt). Verbs have been defecting from this coalition for centuries (wend/went, pen/pent, gird/girt, geld/gelt, and gild/gilt all blend/blent into the dominant –ed rule). Culturomic analysis reveals that the collapse of this alliance has been the most significant driver of regularization in the past 200 years. The regularization of burnt, smelt, spelt, and spilt originated in the US; the forms still cling to life in British English (Fig. 2E,F). But the –t irregulars may be doomed in England too: each year, a population the size of Cambridge adopts 'burned' in lieu of 'burnt'.

Though irregulars generally yield to regulars, two verbs did the opposite: light/lit and wake/woke. Both were irregular in Middle English, were mostly regular by 1800, and subsequently backtracked and are irregular again today. The fact that these verbs have been going back and forth for nearly 500 years highlights the gradual nature of the underlying process.

Still, there was at least one instance of rapid progress by an irregular form. Presently, 1% of the English speaking population switches from 'sneaked' to 'snuck' every year: someone will have snuck off while you read this sentence. As before, this trend is more prominent in the United States, but recently sneaked across the Atlantic: America is the world's leading exporter of both regular and irregular verbs.

Out with the Old

Just as individuals forget the past (18, 19), so do societies (20). To quantify this effect, we reasoned that the frequency of 1-grams such as '1951' could be used to measure interest in the events of the corresponding year, and created plots for each year between 1875 and 1975.

The plots had a characteristic shape. For example, '1951' was rarely discussed until the years immediately preceding 1951. Its frequency soared in 1951, remained high for three years, and then underwent a rapid decay, dropping by half over the next fifteen years. Finally, the plots enter a regime marked by slower forgetting: collective memory has both a short-term and a long-term component.

But there have been changes. The amplitude of the plots is rising every year: precise dates are increasingly common. There is also a greater focus on the present. For instance, '1880' declined to half its peak value in 1912, a lag of 32 years. In contrast, '1973' declined to half its peak by 1983, a lag of only 10 years. We are forgetting our past faster with each passing year (Fig. 3A).

We were curious whether our increasing tendency to forget the old was accompanied by more rapid assimilation of the new (21). We divided a list of 154 inventions into time-resolved cohorts based on the forty-year interval in which they were first invented (1800-1840, 1840-1880, and 1880- 1920) (7). We tracked the frequency of each invention in the n^{th} after it was invented as compared to its maximum value, and plotted the median of these rescaled trajectories for each cohort.

The inventions from the earliest cohort (1800-1840) took over 66 years from invention to widespread impact (frequency >25% of peak). Since then, the cultural adoption of technology has become more rapid: the 1840-1880 invention cohort was widely adopted within 50 years; the 1880-1920 cohort within 27 (Fig. 3B).

"In the future, everyone will be world famous for 7.5 minutes"

-Whatsisname

People, too, rise to prominence, only to be forgotten (22). Fame can be tracked by measuring the frequency of a person's name (Fig. 3C). We compared the rise to fame of the most famous people of different eras. We took all 740,000 people with entries in Wikipedia, removed cases where several famous individuals share a name, and sorted the rest by birthdate and frequency (23). For every year from 1800-1950, we constructed a cohort consisting of the fifty most famous people born in that year. For example, the 1882 cohort includes 'Virginia Woolf' and 'Felix Frankfurter'; the 1946 cohort includes 'Bill Clinton' and 'Steven Spielberg'. We plotted the median frequency for the names in each cohort over time (Fig. 3D-E). The resulting trajectories were all similar. Each cohort had a pre-celebrity period (median frequency $<10^{-9}$), followed by a rapid rise to prominence, a peak, and a slow decline. We therefore characterized each cohort using four parameters: (i) the age of initial celebrity; (ii) the doubling time of the initial rise; (iii) the age of peak celebrity; (iv) the half-life of the decline (Fig. 3E). The age of peak celebrity has been consistent over time: about 75 years after birth. But the other parameters have been changing. Fame comes sooner and rises faster: between the early 19th century and the mid-20th century, the age of initial celebrity declined from 43 to 29 years, and the doubling time fell from 8.1 to 3.3 years. As a result, the most famous people alive today are more famous – in books – than their predecessors. Yet this fame is increasingly short-lived: the post-peak half-life dropped from 120 to 71 years during the nineteenth century.

We repeated this analysis with all 42,358 people in the databases of Encyclopaedia Britannica (24), which reflect a process of expert curation that began in 1768. The results were similar (7). Thus, people are getting more famous than ever before, but are being forgotten more rapidly than ever.

Occupational choices affect the rise to fame. We focused on the 25 most famous individuals born between 1800 and 1920 in seven occupations (actors, artists, writers, politicians, biologists, physicists, and mathematicians), examining how their fame grew as a function of age (Fig. 3F).

Actors tend to become famous earliest, at around 30. But the fame of the actors we studied – whose ascent preceded the spread of television – rises slowly thereafter. (Their fame peaked at a frequency of $2x10^{-7}$.) The writers became famous about a decade after the actors, but rose for longer and to a much higher peak ($8x10^{-7}$). Politicians did not become famous until their 50s, when, upon being elected President of the United States (in 11 of 25 cases; 9 more were heads of other states) they rapidly rose to become the most famous of the groups ($1x10^{-6}$).

Science is a poor route to fame. Physicists and biologists eventually reached a similar level of fame as actors ($1x10^{-7}$), but it took them far longer. Alas, even at their peak, mathematicians tend not to be appreciated by the public ($2x10^{-8}$).

Detecting Censorship and Suppression

Suppression – of a person, or an idea – leaves quantifiable fingerprints (25). For instance, Nazi censorship of the Jewish artist Marc Chagall is evident by comparing the frequency of 'Marc Chagall' in English and in German books (Fig.4A). In both languages, there is a rapid ascent starting in the late 1910s (when Chagall was in his early 30s). In English, the ascent continues. But in German, the artist's popularity decreases, reaching a nadir from 1936-1944, when his full name appears only once. (In contrast, from 1946-1954, 'Marc Chagall' appears nearly 100 times in the German corpus.) Such examples are found in many countries, including Russia (e.g. Trotsky), China (Tiananmen Square) and the US (the Hollywood Ten, blacklisted in 1947) (Fig.4B-D).

We probed the impact of censorship on a person's cultural influence in Nazi Germany. Led by such figures as the librarian Wolfgang Hermann, the Nazis created lists of authors and artists whose 'undesirable', 'degenerate' work was banned from libraries and museums and publicly burned (26-28). We plotted median usage in German for five such lists: artists (100 names), as well as writers of

Literature (147), Politics (117), History (53), and Philosophy (35) (Fig 4E). We also included a collection of Nazi party members (547 names, ref 7). The five suppressed groups exhibited a decline. This decline was modest for writers of history (9%) and literature (27%), but pronounced in politics (60%), philosophy (76%), and art (56%). The only group whose signal increased during the Third Reich was the Nazi party members (a 500% increase; ref 7).

Given such strong signals, we tested whether one could identify victims of Nazi repression *de novo*. We computed a 'suppression index' *s* for each person by dividing their frequency from 1933 – 1945 by the mean frequency in 1925-1933 and in 1955-1965 (Fig.4F, *Inset*). In English, the distribution of suppression indices is tightly centered around unity. Fewer than 1% of individuals lie at the extremes ($s<1/5$ or $s>5$).

In German, the distribution in much wider, and skewed leftward: suppression in Nazi Germany was not the exception, but the rule (Fig. 4F). At the far left, 9.8% of individuals showed strong suppression ($s<1/5$). This population is highly enriched for documented victims of repression, such as Pablo Picasso ($s=0.12$), the Bauhaus architect Walter Gropius ($s=0.16$), and Hermann Maas ($s<.01$), an influential Protestant Minister who helped many Jews flee (7). (Maas was later recognized by Israel's *Yad Vashem* as a 'Righteous Among the Nations'.) At the other extreme, 1.5% of the population exhibited a dramatic rise ($s>5$). This subpopulation is highly enriched for Nazis and Nazi-supporters, who benefited immensely from government propaganda (7).

These results provide a strategy for rapidly identifying likely victims of censorship from a large pool of possibilities, and highlights how culturomic methods might complement existing historical approaches.

Culturomics

Culturomics is the application of high-throughput data collection and analysis to the study of human culture. Books are a beginning, but we must also incorporate newspapers (29), manuscripts (30), maps (31), artwork (32), and a myriad of other human creations (33, 34). Of course, many voices – already lost to time – lie forever beyond our reach.

Culturomic results are a new type of evidence in the humanities. As with fossils of ancient creatures, the challenge of culturomics lies in the interpretation of this evidence. Considerations of space restrict us to the briefest of surveys: a handful of trajectories and our initial interpretations. Many more fossils, with shapes no less intriguing, beckon:

(i) Peaks in 'influenza' correspond with dates of known pandemics, suggesting the value of culturomic methods for historical epidemiology (35) (Fig. 5A).

(ii) Trajectories for 'the North', 'the South', and finally, 'the enemy' reflect how polarization of the states preceded the descent into war (Fig. 5B).

(iii) In the battle of the sexes, the 'women' are gaining ground on the 'men' (Fig. 5C).

(iv) 'féminisme' made early inroads in France, but the US proved to be a more fertile environment in the long run (Fig. 5D).

(v) 'Galileo', 'Darwin', and 'Einstein' may be well-known scientists, but 'Freud' is more deeply engrained in our collective subconscious (Fig. 5E).

(vi) Interest in 'evolution' was waning when 'DNA' came along (Fig. 5F).

(vii) The history of the American diet offers many appetizing opportunities for future research; the menu includes 'steak', 'sausage', 'ice cream', 'hamburger', 'pizza', 'pasta', and 'sushi' (Fig. 5G).

(viii) 'God' is not dead; but needs a new publicist (Fig. 5H).

These, together with the billions of other trajectories that accompany them, will furnish a great cache of bones from which to reconstruct the skeleton of a new science.

References and Notes

1. Wilson, Edward O. Consilience. New York: Knopf, 1998.
2. Sperber, Dan. "Anthropology and psychology: Towards an epidemiology of representations." Man 20 (1985): 73-89.
3. Lieberson, Stanley and Joel Horwich. "Implication analysis: a pragmatic proposal for linking theory and data in the social sciences." Sociological Methodology 38 (December 2008): 1-50.
4. Cavalli-Sforza, L. L., and Marcus W. Feldman. Cultural Transmission and Evolution. Princeton, NJ: Princeton UP, 1981.
5. Niyogi, Partha. The Computational Nature of Language Learning and Evolution. Cambridge, MA: MIT, 2006.
6. Zipf, George Kingsley. The Psycho-biology of Language. Boston: Houghton Mifflin, 1935.
7. Materials and methods are available as supporting material on Science Online.
8. Lander, E. S. et al. "Initial sequencing and analysis of the human genome." Nature 409 (February 2001): 860-921.
9. Read, Allen W. “The Scope of the American Dictionary.” American Speech 8 (1933): 10–20.
10. Gove, Philip Babcock, ed. Webster's Third New International Dictionary of the English Language, Unabridged. Springfield, MA: Merriam-Webster, 1993.
11. Pickett, Joseph, P. ed. The American Heritage Dictionary of the English Language, Fourth Edition. Boston / New York, NY: Houghton Mifflin Pub., 2000.
12. Simpson, J. A., E. S. C. Weiner, and Michael Proffitt, eds. Oxford English Dictionary. Oxford [England]: Clarendon, 1993.
13. Algeo, John, and Adele S. Algeo. Fifty Years among the New Words: a Dictionary of Neologisms, 1941-1991. Cambridge UK, 1991.
14. Pinker, Steven. Words and Rules. New York: Basic, 1999.
15. Kroch, Anthony S. "Reflexes of Grammar in Patterns of Language Change." Language Variation and Change 1.03 (1989): 199.
16. Bybee, Joan L. "From Usage to Grammar: The Mind's Response to Repetition." Language 82.4 (2006): 711-33.
17. Lieberman*, Erez, Jean-Baptiste Michel*, Joe Jackson, Tina Tang, and Martin A. Nowak. "Quantifying the Evolutionary Dynamics of Language." Nature 449 (2007): 713-16.
18. Milner, Brenda, Larry R. Squire, and Eric R. Kandel. "Cognitive Neuroscience and the Study of Memory."Neuron 20.3 (1998): 445-68.
19. Ebbinghaus, Hermann. Memory: a Contribution to Experimental Psychology. New York: Dover, 1987.
20. Halbwachs, Maurice. On Collective Memory. Trans. Lewis A. Coser. Chicago: University of Chicago, 1992.
21. Ulam, S. "John Von Neumann 1903-1957." Bulletin of the American Mathematical Society 64.3 (1958): 1-50.
22. Braudy, Leo. The Frenzy of Renown: Fame & Its History. New York: Vintage, 1997.
23. Wikipedia. Web. 23 Aug. 2010. <http://www.wikipedia.org/>.
24. Hoiberg, Dale, ed. Encyclopaedia Britannica. Chicago: Encyclopaedia Britannica, 2002.
25. Gregorian, Vartan, ed. Censorship: 500 Years of Conflict. New York: New York Public Library, 1984.
26. Treß, Werner. *Wider Den Undeutschen Geist: Bücherverbrennung 1933*. Berlin: Parthas, 2003.
27. Sauder, Gerhard. *Die Bücherverbrennung: 10. Mai 1933*. Frankfurt/Main: Ullstein, 1985.

28. Barron, Stephanie, and Peter W. Guenther. Degenerate Art: the Fate of the Avant-garde in Nazi Germany. Los Angeles: Los Angeles County Museum of Art, 1991.
29. Google News Archive Search. Web. <http://news.google.com/archivesearch>.
30. Digital Scriptorium. Web. <http://www.scriptorium.columbia.edu>.
31. Visual Eyes. Web. <http://www.viseyes.org>.
32. ARTstor. Web. <http://www.artstor.org>.
33. Europeana. Web. <http://www.europeana.eu>.
34. Hathi Trust Digital Library. Web. <http://www.hathitrust.org>.
35. Barry, John M. The Great Influenza: the Epic Story of the Deadliest Plague in History. New York: Viking, 2004.
36. J-B.M. was supported by the Foundational Questions in Evolutionary Biology Prize Fellowship and the Systems Biology Program (Harvard Medical School). Y.K.S. was supported by internships at Google. S.P. acknowledges support from NIH grant HD 18381. E.A. was supported by the Harvard Society of Fellows, the Fannie and John Hertz Foundation Graduate Fellowship, the National Defense Science and Engineering Graduate Fellowship, the NSF Graduate Fellowship, the National Space Biomedical Research Institute, and NHGRI Grant T32 HG002295 . This work was supported by a Google Research Award. The Program for Evolutionary Dynamics acknowledges support from the Templeton Foundation, NIH grant R01GM078986, and the Bill and Melinda Gates Foundation. Some of the methods described in this paper are covered by US patents 7463772 and 7508978. We are grateful to D. Bloomberg, A. Popat, M. McCormick, T. Mitchison, U. Alon, S. Shieber, E. Lander, R. Nagpal, J. Fruchter, J. Guldi, J. Cauz, C. Cole, P. Bordalo, N. Christakis, C. Rosenberg, M. Liberman, J. Scheidlower, B. Zimmer, R. Darnton, and A. Spector for discussions; to C-M. Hetrea and K. Sen for assistance with Encyclopaedia Britannica's database, to S. Eismann, W. Treß, and the City of Berlin website (berlin.de) for assistance documenting victims of Nazi censorship, to C. Lazell and G.T. Fournier for assistance with annotation, to M. Lopez for assistance with Figure 1, to G. Elbaz and W. Gilbert for reviewing an early draft, and to Google's library partners and every author who has ever picked up a pen, for books.

Supporting Online Material
www.sciencemag.org/cgi/content/full/???
Materials and Methods
Figs. S1 to S19

Figure Legends

Fig. 1. 'Culturomic' analyses study millions of books at once. (**A**) Top row: authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: a culturomic 'timeline' shows the frequency of 'apple' in English books over time (1800-2000). (**B**) Usage frequency of 'slavery'. The Civil War (1861-1865) and the civil rights movement (1955-1968) are highlighted in red. The number in the upper left (1e-4) is the unit of frequency. (**C**) Usage frequency over time for 'the Great War' (blue), 'World War I' (green), and 'World War II' (red).

Fig. 2. Culturomics has profound consequences for the study of language, lexicography, and grammar. (**A**) The size of the English lexicon over time. Tick marks show the number of single words in three dictionaries (see text). (**B**) Fraction of words in the lexicon that appear in two different dictionaries as a function of usage frequency. (**C**) Five words added by the AHD in its 2000 update. Inset: Median frequency of new words added to AHD4 in 2000. The frequency of half of these words exceeded 10^{-9} as

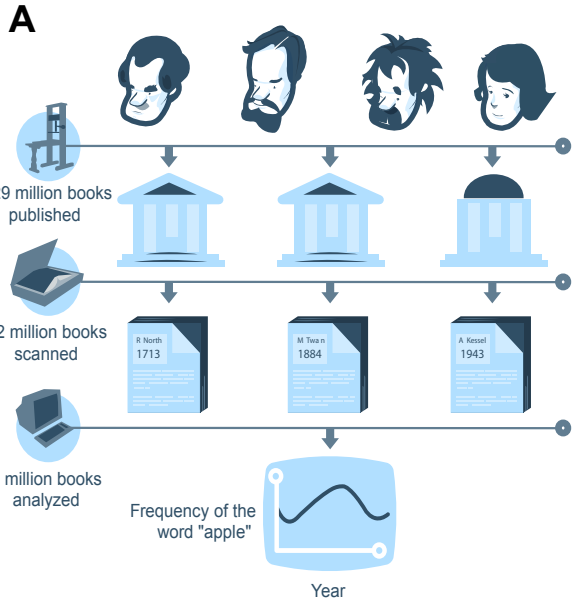
far back as 1890 (white dot). (**D**) Obsolete words added to AHD4 in 2000. Inset: Mean frequency of the 2220 AHD headwords whose current usage frequency is less than 10^{-9}. (**E**) Usage frequency of irregular verbs (red) and their regular counterparts (blue). Some verbs (chide/chided) have regularized during the last two centuries. The trajectories for 'speeded' and 'speed up' (green) are similar, reflecting the role of semantic factors in this instance of regularization. The verb 'burn' first regularized in the US (US flag) and later in the UK (UK flag). The irregular 'snuck' is rapidly gaining on 'sneaked'. (**F**) Scatter plot of the irregular verbs; each verb's position depends on its regularity (see text) in the early 19th century (x-coordinate) and in the late 20th century (y-coordinate). For 16% of the verbs, the change in regularity was greater than 10% (large font). Dashed lines separate irregular verbs (regularity<50%) from regular verbs (regularity>50%). Six verbs became regular (upper left quadrant, blue), while two became irregular (lower right quadrant, red). Inset: the regularity of 'chide' over time. (**G**) Median regularity of verbs whose past tense is often signified with a –t suffix instead of –ed (burn, smell, spell, spill, dwell, learn, and spoil) in US (black) and UK (grey) books.

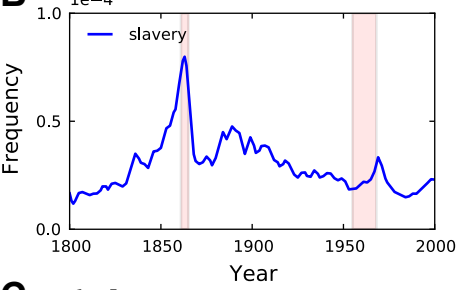
Fig. 3. Cultural turnover is accelerating. (**A**) We forget: frequency of 1883 (blue), 1910 (green) and 1950 (red). *Inset:* We forget faster. The half-life of the curves (grey dots) is getting shorter (grey line: moving average). (**B**) Cultural adoption occurs faster. Median trajectory for three cohorts of inventions from three different time periods (1800-1840: blue, 1840-1880: green, 1880-1920: red). *Inset:* The telephone (green, date of invention: green arrow) and radio (blue, date of invention: blue arrow). (**C**) Fame of various personalities born between 1920 and 1930. (**D**) Frequency of the 50 most famous people born in 1871 (grey lines; median: dark gray). Five examples are highlighted. (**E**) The median trajectory of the 1865 cohort is characterized by four parameters: (i) initial 'age of celebrity' (34 years old, tick mark); (ii) doubling time of the subsequent rise to fame (4 years, blue line); (iii) 'age of peak celebrity' (70 years after birth, tick mark), and (iv) half-life of the post-peak 'forgetting' phase (73 years, red line). *Inset*: The doubling time and half-life over time. (**F**) The median trajectory of the 25 most famous personalities born between 1800 and 1920 in various careers.

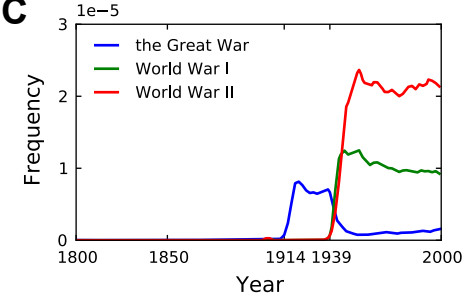
Fig. 4. Culturomics can be used to detect censorship. (**A**) Usage frequency of 'Marc Chagall' in German (red) as compared to English (blue). (**B**) Suppression of Leon Trotsky (blue), Grigory Zinoviev (green), and Lev Kamenev (red) in Russian texts, with noteworthy events indicated: Trotsky's assassination (blue arrow), Zinoviev and Kamenev executed (red arrow), the 'Great Purge' (red highlight), perestroika (grey arrow). (**C**) The 1976 and 1989 Tiananmen Square incidents both lead to elevated discussion in English texts. Response to the 1989 incident is largely absent in Chinese texts (blue), suggesting government censorship. (**D**) After the 'Hollywood Ten' were blacklisted (red highlight) from American movie studios, their fame declined (median: wide grey). None of them were credited in a film until 1960's (aptly named) 'Exodus'. (**E**) Writers in various disciplines were suppressed by the Nazi regime (red highlight). In contrast, the Nazis themselves (thick red) exhibited a strong fame peak during the war years. (**F**) Distribution of suppression indices for both English (blue) and German (red) for the period from 1933-1945. Three victims of Nazi suppression are highlighted at left (red arrows). *Inset*: Calculation of the suppression index for 'Henri Matisse'.

Fig. 5. Culturomics provides quantitative evidence for scholars in many fields. (**A**) Historical Epidemiology: 'influenza' is shown in blue; the Russian, Spanish, and Asian flu epidemics are highlighted. (**B**) History of the Civil War. (**C**) Comparative History. (**D**) Gender studies. (**E,F**) History of Science. (**G**) Historical Gastronomy. (**H**) History of Religion: 'God'.

Fig. 1


A
129 million books published
12 million books scanned
5 million books analyzed
R North 1713
M Twain 1884
A Kessel 1943
Frequency of the word "apple"
Year


B
1e−4
slavery
Frequency
1.0
0.5
0.0
1800
1850
1900
1950
2000
Year


C
1e−5
the Great War
World War I
World War II
Frequency
3
2
1
0
1800
1850
1914 1939
2000
Year

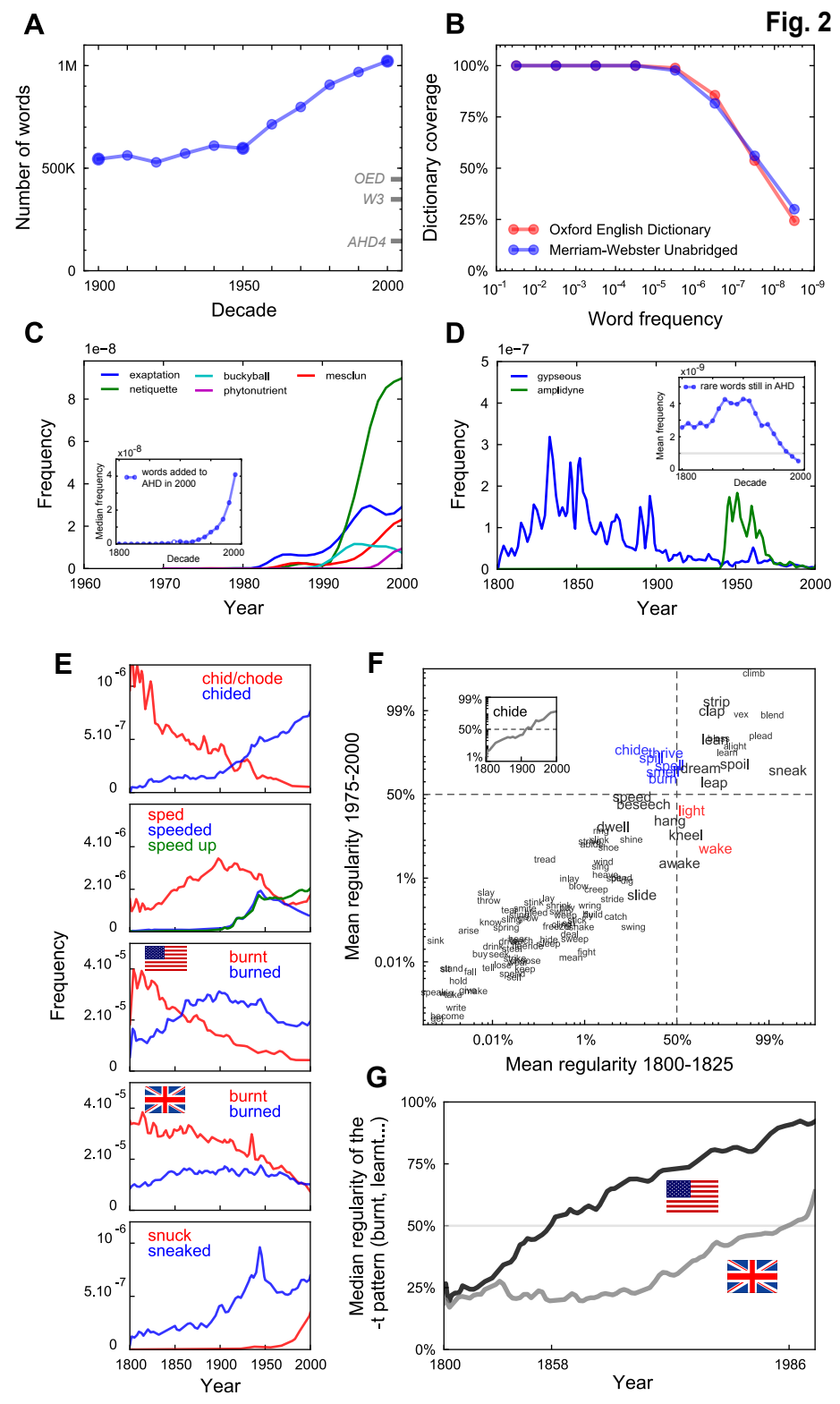

Fig. 2
A
Number of words
1M
500K
0
1900
1950
2000
Decade
OED
W3
AHD4
B
Dictionary coverage
100%
75%
50%
25%
0%
Oxford English Dictionary
Merriam-Webster Unabridged
Word frequency
C
1e−8
Frequency
exaptation
buckyball
mesclun
netiquette
phytonutrient
words added to AHD in 2000
Median frequency
Decade
Year
D
1e−7
Frequency
gypseous
amplidyne
rare words still in AHD
Mean frequency
Decade
Year
E
Frequency
chid/chode
chided
sped
speeded
speed up
burnt
burned
burnt
burned
snuck
sneaked
Year
F
Mean regularity 1975-2000
Mean regularity 1800-1825
chide
climb
strip
clap
vex
blend
plead
slight
learn
chide
thrive
spill
spell
smell
burn
dream
spoil
sneak
leap
speed
beseech
light
hang
dwell
kneel
wake
awake
slide
G
Median regularity of the -t pattern (burnt, learnt...)
Year

Fig. 3

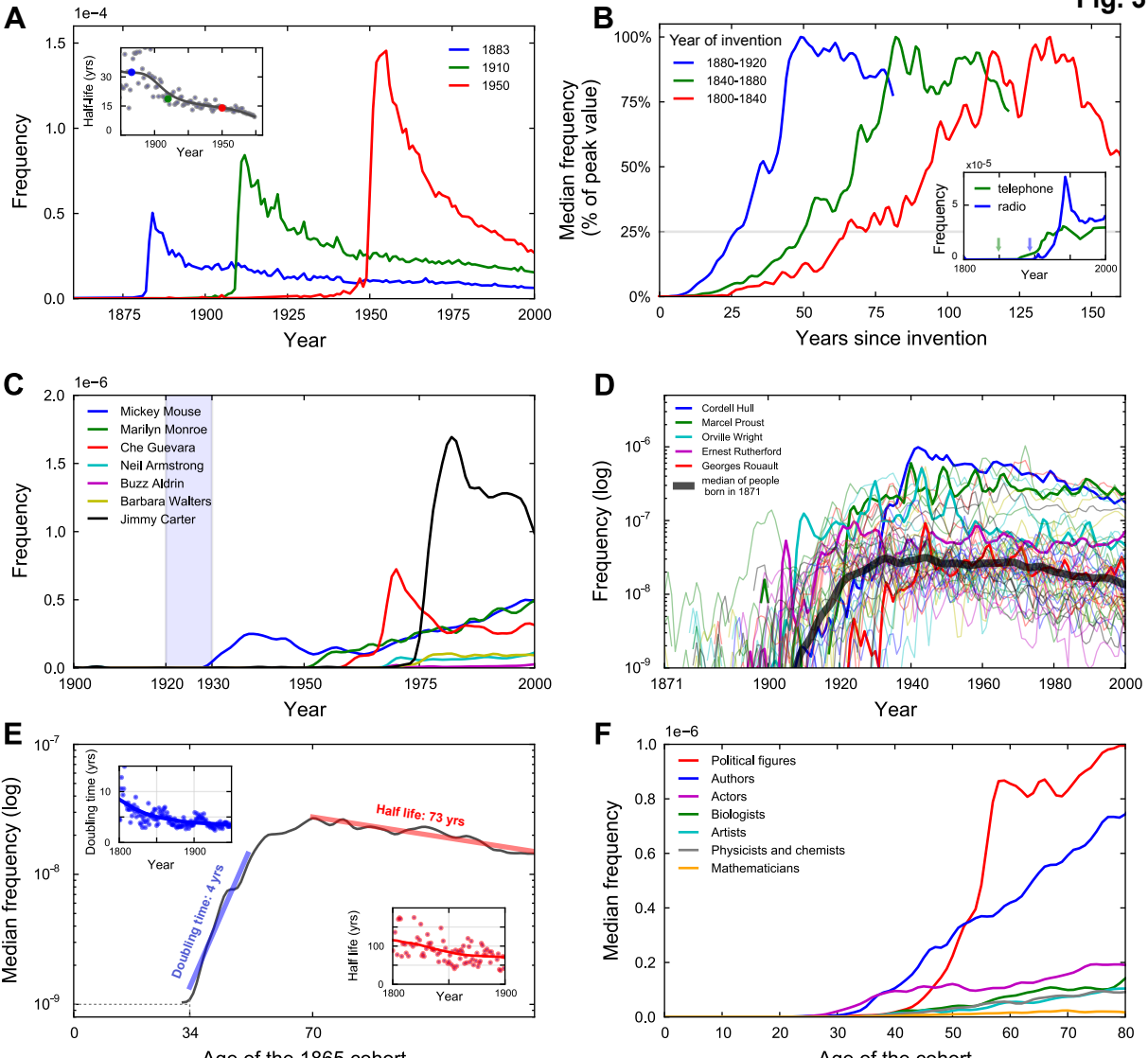

A
1e−4
Frequency
Year
1883
1910
1950
Half-life (yrs)
B
Year of invention
1880-1920
1840-1880
1800-1840
Median frequency (% of peak value)
Years since invention
x10-5
telephone
radio
C
1e−6
Mickey Mouse
Marilyn Monroe
Che Guevara
Neil Armstrong
Buzz Aldrin
Barbara Walters
Jimmy Carter
D
Cordell Hull
Marcel Proust
Orville Wright
Ernest Rutherford
Georges Rouault
median of people born in 1871
Frequency (log)
E
Median frequency (log)
Doubling time (yrs)
Half life: 73 yrs
Doubling time: 4 yrs
Half life (yrs)
Age of the 1865 cohort
F
1e−6
Political figures
Authors
Actors
Biologists
Artists
Physicists and chemists
Mathematicians
Median frequency
Age of the cohort

Fig. 4

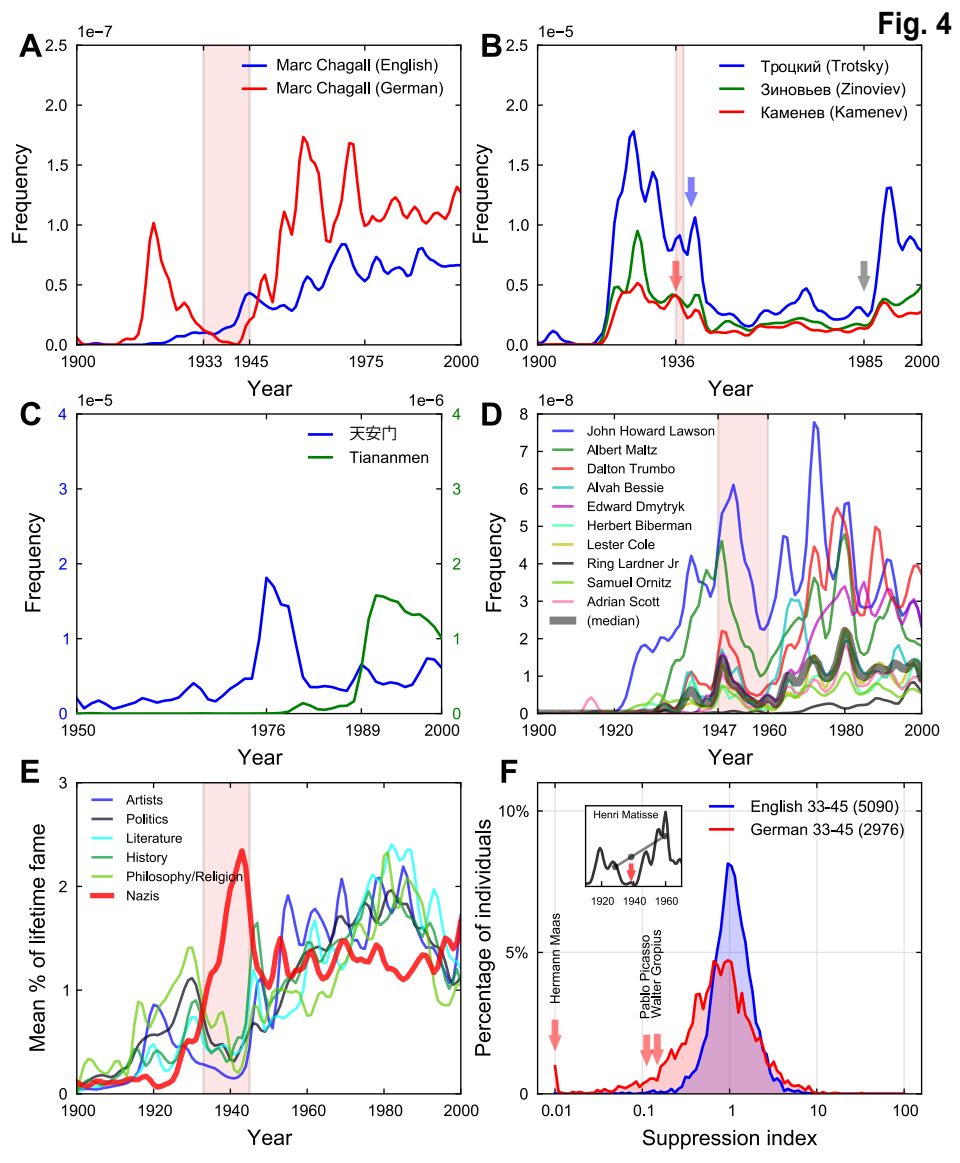
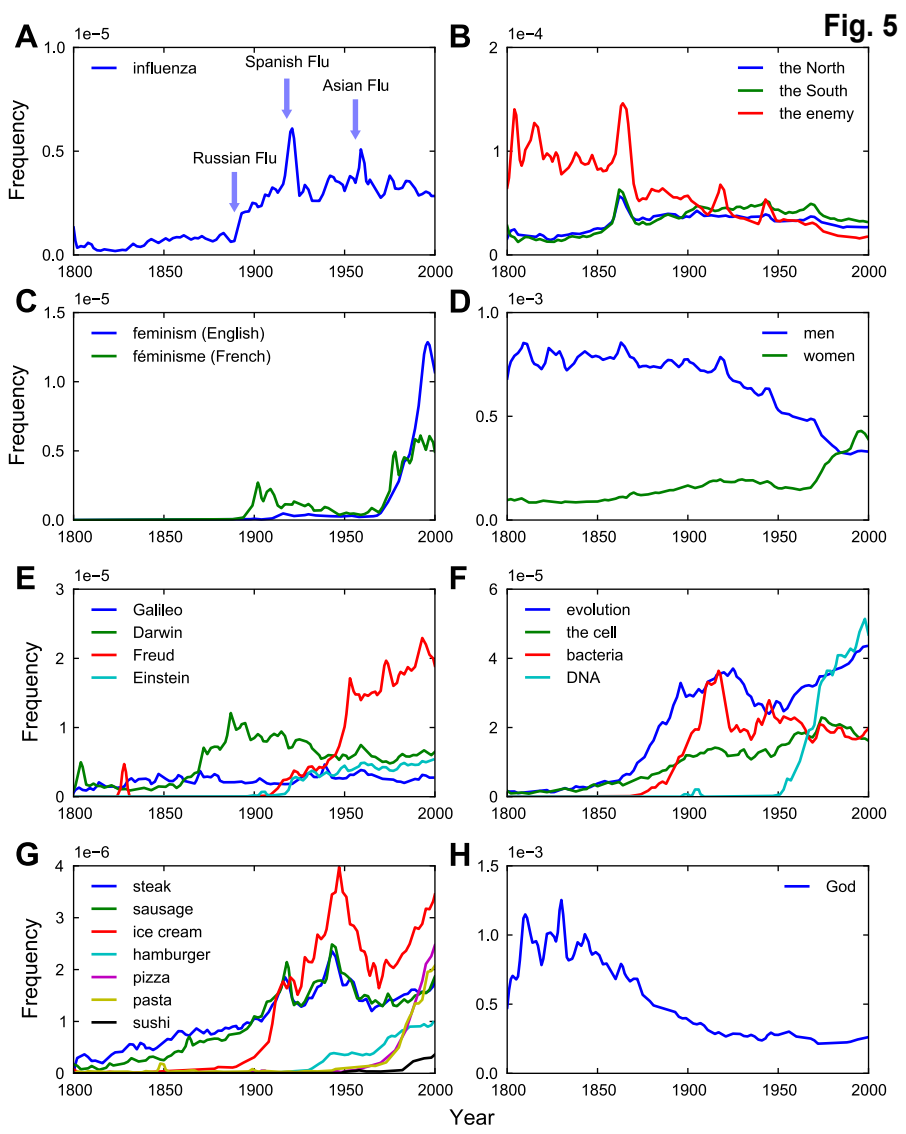

A
1e−7
Marc Chagall (English)
Marc Chagall (German)
Frequency
Year
B
1e−5
Троцкий (Trotsky)
Зиновьев (Zinoviev)
Каменев (Kamenev)
Frequency
Year
C
1e−5
1e−6
天安门
Tiananmen
Frequency
Year
D
1e−8
John Howard Lawson
Albert Maltz
Dalton Trumbo
Alvah Bessie
Edward Dmytryk
Herbert Biberman
Lester Cole
Ring Lardner Jr
Samuel Ornitz
Adrian Scott
(median)
Frequency
Year
E
Artists
Politics
Literature
History
Philosophy/Religion
Nazis
Mean % of lifetime fame
Year
F
Henri Matisse
English 33-45 (5090)
German 33-45 (2976)
Hermann Maas
Pablo Picasso
Walter Gropius
Percentage of individuals
Suppression index

Fig. 5


A
1e−5
Frequency
influenza
Spanish Flu
Asian Flu
Russian Flu
B
1e−4
the North
the South
the enemy
C
1e−5
feminism (English)
féminisme (French)
D
1e−3
men
women
E
1e−5
Galileo
Darwin
Freud
Einstein
F
1e−5
evolution
the cell
bacteria
DNA
G
1e−6
steak
sausage
ice cream
hamburger
pizza
pasta
sushi
H
1e−3
God
Year