Estimating technological breaks in the size and composition of human collective memory from biographical data

C. Jara-Figueroa[1], Amy Z. Yu[1], César A. Hidalgo[1,2]

[1]The MIT Media Lab, Massachusetts Institute of Technology

[2] to whom correspondence should be addressed: hidalgo@mit.edu

Abstract

The ability of humans to accumulate knowledge and information across generations is a defining feature of our species. This ability depends on factors that range from the psychological biases that predispose us to learn from skillful, accomplished, and prestigious people, to the development of technologies for recording and communicating information: from clay tablets to the Internet. In this paper we present empirical evidence documenting how communication technologies have shaped human collective memory. We show that changes in communication technologies, including the introduction of printing and the maturity of shorter forms of printed media—such as newspapers, journals, and pamphlets—were accompanied by sharp changes (or breaks) in the per-capita number of memorable biographies from a time period that are present in current online and offline sources. Moreover, we find that changes in technology, such as the introduction of printing, film, radio, and television, coincide with sharp shifts in the occupations of the individuals present in these biographical records. These two empirical facts provide evidence in support of theories arguing that human collective memory is shaped by the technologies we use to record and communicate information.

Introduction

There is no information without representation [1]. To exist, information needs to be recorded in physical systems, from books to hard drives, and from DNA to the human brain [2,3,4]. The need for information to be physically embodied implies that human *collective memory*—our records of past events, objects, ideas, and individuals—will co-evolve with the ability of humans to physically record information. In other words, human collective memory will change with communication technologies.

The idea that communication technologies shape our collective memory is not new. In fact, it is present in the definition of history. History differentiates itself from other academic efforts to study our past, such as archeology and anthropology, by a breakthrough in communication technologies: the invention of writing. Yet, as the historian Elizabeth Eisenstein argued [5,6], while historians largely agree that writing defines the onset of history, there is no widespread consensus about the relative importance that more recent technological changes had on our species' ability to remember.

But since the invention of writing, humans have developed a plethora of new technologies that have shaped our ability to record and communicate information. These subsequent technological changes have given rise to a fertile set of theoretical studies exploring how humans' ability to record information shapes the content and the volume of information that we record [5,6,7,8,9,10,11,12]. These studies include the work of the economic historian Harold Innis [12], the philosopher of communication Marshall McLuhan [9], and the historian of printing Elizabeth Eisenstein [5,6], among others [7,8,10,11].

Innis is considered by some [13] to be the first scholar who attempted to establish the history of communications as a distinct academic field. Unfortunately, Innis passed away before finishing his "History of Communications," which was poised to become his defining work on this topic [13]. Innis's ideas, nevertheless, were expanded and popularized by one of his colleagues, Marshall McLuhan, who made the study of the social impact of communication technologies globally famous with his 1964 book: *Understanding Media*: *The Extensions of Man.* Understanding Media opened with the now famous phrase: "the medium is the message" [9], which meant that changes in communication technologies are more consequential to society than the messages uttered through them. McLuhan emphasized how changes in media changed the type of messages that can be transmitted through them (i.e. you cannot transmit musical performances using printing). Ultimately, by biasing the production of content, new media changes the type of messages that are transmitted in a society, and thus McLuhan expects changes in media to change a society's collective memory.

Building on Innis and McLuhan's work, the historian of printing Elizabeth Eisenstein expanded the empirical validity of Innis and McLuhan claims and ideas by documenting the impact of printing in early modern Europe [5,6]. Eisenstein showed that prior to the introduction of printing, our species' ability to store information was volatile. Printing, therefore, changed the durability and reliability of the data available to scholars (such as the astronomical tables used by Copernicus and Kepler [6]), and also, helped catalyze new scientific knowledge by dramatically increasing the availability and reliability of the data and literature available to renaissance scholars.

But not all changes in our collective memory have come from changes in the technologies we use to record and transmit information. An alternative communication revolution came with the introduction of the public sphere: a space for public discourse in the eighteenth century described by the sociologist Jurgen Habermas [14]. Habermas argued that the public sphere emerged together with the creation of coffeehouses in London, Paris, Vienna, and Prague, and with the introduction of new publication formats, such as newspapers, journals, and pamphlets. Coffeehouses acted as centers of social interaction that enabled the public use of reason in a rational-critical debate. New formats, such as journals, pamphlets, and newspapers, provided vehicles to share content focused on more recent events than the content described in books, fueling public discussions. According to Habermas, this combination of new formats and new spaces for public discourse gave rise to the public sphere, a revolution that once again shaped our collective memory.

But what exactly do we mean by 'collective memory'? Collective memory is a term characterized by diverging definitions that are particular to each field [15,16,17]. For instance, the sociologist Maurice Halbwachs defined the concept of collective memory as both, the process by which society creates a shared version of the past, and the fact that each individual's memory is shaped by the information that is available to her from society. The French historian Pierre Nora, on the other hand, focused on separating memory from history by exploring what he called *sites of memory*, which are purely material sites invested with a symbolic aura, and studying their role in the construction of a group's identity—in Nora's case, the construction of French identity [18]. More recently, the Egyptologists Aleida Assmann and Jan Assmann [17] described collective memory by focusing on the distinction between *Communicative Memory*—the information we preserve through repeated acts of communication—and *Cultural Memory*—the reusable texts, images, and rituals that are specific to each society. In our work, we subscribe to Aleida and Jan Assmann's definition of Cultural Memory, since our data consists of digitized texts.

In fact, it is the recent widespread availability of digitized texts what allows us to provide novel statistical evidence in support of the ideas of Innis, McLuhan, and Eisenstein, since statistically testing these ideas has been difficult in the past, due to the scarcity of structured datasets on historical events. It is worth noting, however, that digital media has already been used to explore historical patterns [19,20,21,22] that might be otherwise buried in narrative descriptions [23,24]. Examples of these studies include the evolution of language and ideas as recorded in printed books [22], patterns of historical migration [25], the importance of language translations in the global diffusion of information [26], the emotional content of global languages [27], and the dynamics of fame [19,21,28]. These studies are made possible thanks to new datasets that leverage digital sources, such as Wikipedia, Freebase, and digitized books [29].

In this work, we use two large datasets based on biographical records to provide evidence supporting the theories of Innis, McLuhan, and Eisenstein, which state that communication technologies affect the rate and content of human collective memory. These two datasets summarize the occupations and time periods associated with 11,337 and 3,869 globally memorable biographies extracted from multilingual expressions in digital and printed sources. Both of these datasets show evidence of breaks in the rate at which we preserve biographical information that coincide with the transition from scribal to print culture, and with the birth of the Habermas' public sphere which includes the maturity of shorter forms of printed media. Furthermore, the Pantheon 1.0 dataset exhibits sharp changes in the occupations of the biographical records that coincide with changes in communication technologies, such as printing, film and radio, and television, demonstrating that these technological changes do not only shape how much we remember, but also, what we remember.

Data

We use biographical data of memorable historical characters from two sources: The Pantheon 1.0 dataset [19] and the Human Accomplishments dataset [20]. The Pantheon 1.0 dataset contains the 11,337 biographies that had a presence in more than 25 different language editions of Wikipedia as of May 2013, the time when the data was collected. The Pantheon 1.0 dataset uses the number of Wikipedia Language editions of a historical character (*L*) as a rough proxy of its memorability, enabling us to test the robustness of our results for different memorability thresholds. Furthermore, the Pantheon 1.0 dataset associates each biography to a place of birth, a date of birth, and an occupation using a three-level hierarchical classification that disaggregates biographies into 88 distinct occupations—i.e. *Physics* and *Biology* are branches of *Natural Science*, just like *Natural Science* is a branch of *Science*.

The Human Accomplishments dataset contains 3,869 biographies of accomplished individuals from the Arts and Sciences that are recorded in authoritative printed texts in six different languages—for the definition of accomplishment and other details see original source [20]. This dataset classifies people into 5 different inventories—Science, Philosophy, Music, Literature, and Art—instead of the 88 occupations of the Pantheon 1.0 dataset. Hence, the Human Accomplishment dataset provides a limited resolution regarding the composition of our collective memory (see SM for an extended discussion).

The use of historical characters as a proxy of our biographical collective memory can be contextualized by looking at two alternative theories of the role of individuals in history. These are Thomas Carlyle's “great man” theory of history, which sees historical characters as the drivers of change [30], and Herbert Spencer's theory that, in opposition to Carlyle, sees prominent individuals as a reflection of their times [31]. Under Spencer’s view, historical events beget historical characters, and the fame of individuals becomes an artifact of how we record history. Under Carlyle's view historical characters drive historical events. We note that our results hold for either interpretation of the connection between historical characters and history, so we take an agnostic approach with respect to these theories.

Population data comes from the historical world population estimates of the US Census Bureau [32], which reports an aggregated dataset of world population estimates starting from the year 10,000BC. We interpolate the missing years using linear splines.

Both the Pantheon 1.0 and the Human Accomplishments dataset include limitations that need to be taken into account when interpreting our results. We emphasize the need to interpret our results as valid only in the narrow context of the sources used to compile these datasets. The Pantheon 1.0 dataset has all the biases inherent to using Wikipedia as a primary data source [19], therefore our results should be interpreted as statements about the picture of collective memory that is representative only of the people that edit the more than 250 language editions of Wikipedia—a literate and digitally empowered elite of knowledge enthusiasts [33]. For a detailed description of the biases and limitations of the Pantheon 1.0 dataset see [19]. The Human Accomplishments dataset is compiled from literary expressions in six different languages, and is representative only of the elite of people that participated in the creation of printed encyclopedias. An extended discussion of the biases and validation of this dataset can be found in *Human Accomplishments* [20]. We note that the use of biased datasets is not a peculiarity of this study, but rather the norm of historical research, since as both Eisenstein and Innis argued; the study of history is always biased in favor of the groups who have produced written content. That is why it is important to always interpret historical research, including that in this study, as a reflection of the sources used in its creation.

Results

According to the theories advanced by Innis, McLuhan and Eisenstein, the rate at which humans record information changes together with changes in communication technologies. To test this hypothesis we look at the number of memorable people *M* born in a given time window that is remembered prominently today. To make these estimates comparable across time, we normalize *M* by the average population of the world *N* during each time window. This provides us with an estimate of the per-capita number of births occurring in a given year that we still remember today ($m=M/N$). *m* is measured in units of births per year per billion people in the world, or [bpyb].

Figure 1A shows the per-capita number of memorable biographies from a given year (*m*) in both the Pantheon 1.0 dataset and the Human Accomplishment dataset. For both datasets *m* is constant for the 2000 years preceding the introduction of the removable type press, at rates of 3.5 [bpyb] for Pantheon 1.0 and 1.3 [bpyb] for Human Accomplishments, and increased respectively to 5.3 [bpyb] and 6.4 [bpyb] for the 300 years following the introduction of printing. We note that the larger increase in *m* observed for the Human Accomplishment dataset is consistent with its selection criteria which focuses only on biographies from artists and scientists who become more memorable after the introduction of printing. Nevertheless, the discontinuous change observed in both datasets is evidence in favor of the hypothesis that the introduction of printing introduced a discontinuity in the per-capita number of people from a time period that we remember today.

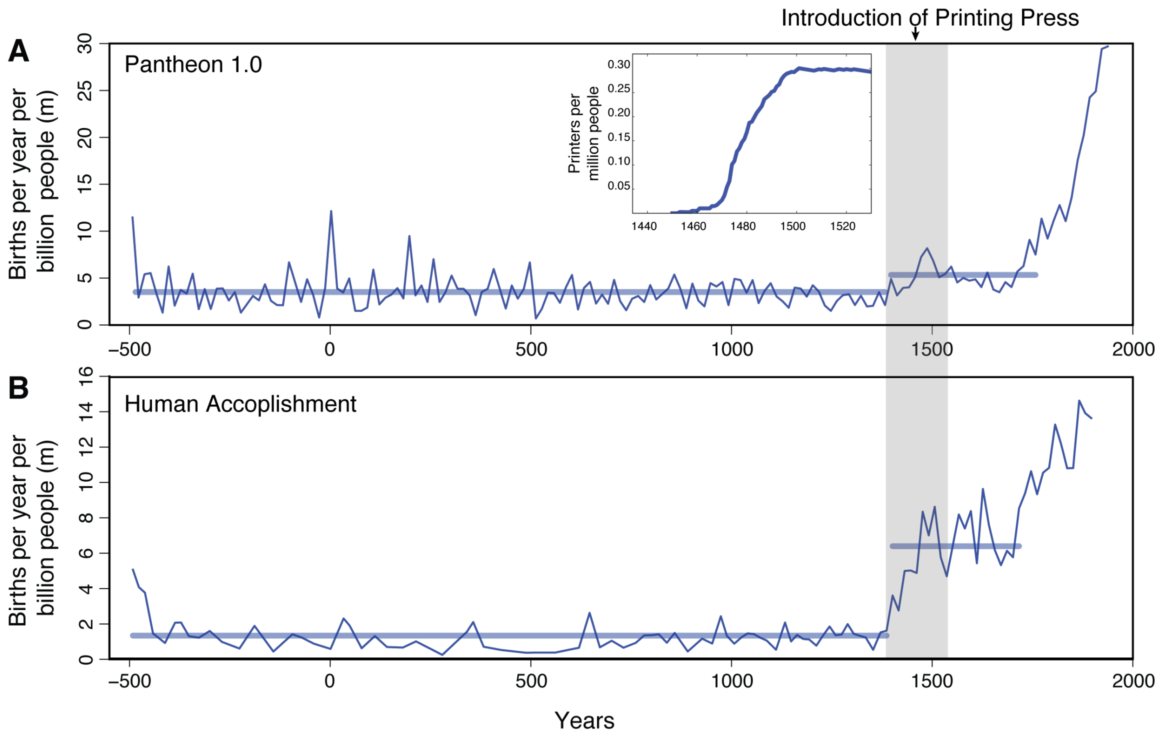

Figure 1: The dynamics of collective memory. **A** Per-capita births of globally memorable people (*m*) measured using a fifteen-tear time window for the Pantheon 1.0 dataset. Inset shows the number of printers per million people in the World between 1440 and 1530, highlighting the rapid adoption of printing. **B** Same as **A** but for the Human Accomplishment dataset. The grey are indicate the period in which printing was expanding.

To statistically test that these abrupt changes are not the result of fluctuations we use the changepoint estimation technique for time series analysis described in [34]. The changepoint analysis estimates the position and number of changepoints in a time series by assuming that the time series can be modeled by a distribution with a fixed mean. The changepoints in a time series are the points that require updating the mean of the distribution used to model the data. To find the changepoints, the technique minimizes a test statistics that depends on the number and position of the changepoints. For both time series the analysis detects two major changepoints. For Pantheon the first break occurs at 1375, and the second at 1750. For Human Accomplishment, the first break occurs at 1379 and the second at 1709.

The two changepoints found by the statistical analysis can be directly mapped to two major breaks in our ability to record and spread information (note that dates are birthdates, meaning that a person born in 1390 was of 60 years of age by the time the printing press was being introduced). The first changepoint occurs right before the introduction of the removable type press in 1450, and it seems to lead to a new stationary state that lasts roughly three centuries. Printing spread and plateaued quickly [35], making the introduction of printing a relatively discrete event in history (Figure 1 A inset), a fact that is consistent with the observed discrete jump. The second break coincides with the birth of the public sphere in the eighteenth century described by Habermas and the maturation of shorter forms of printed media that we described in the introduction (see SM for more details). The time series reveals that starting in the middle of the eighteenth century the per-capita number of globally memorable births begins an era of continuous growth. What this means is that the number of people from the last three centuries that we remember prominently today has increased at a rate that is faster than the global rate of population growth (roughly as the square of the population).

Beyond the rate at which we record information, the theories of Innis, McLuhan, and Eisenstein also suggest that changes in technology should change the type of information we record. We test this hypothesis by looking at changes in the occupations associated with the biographies in the Pantheon 1.0 dataset.

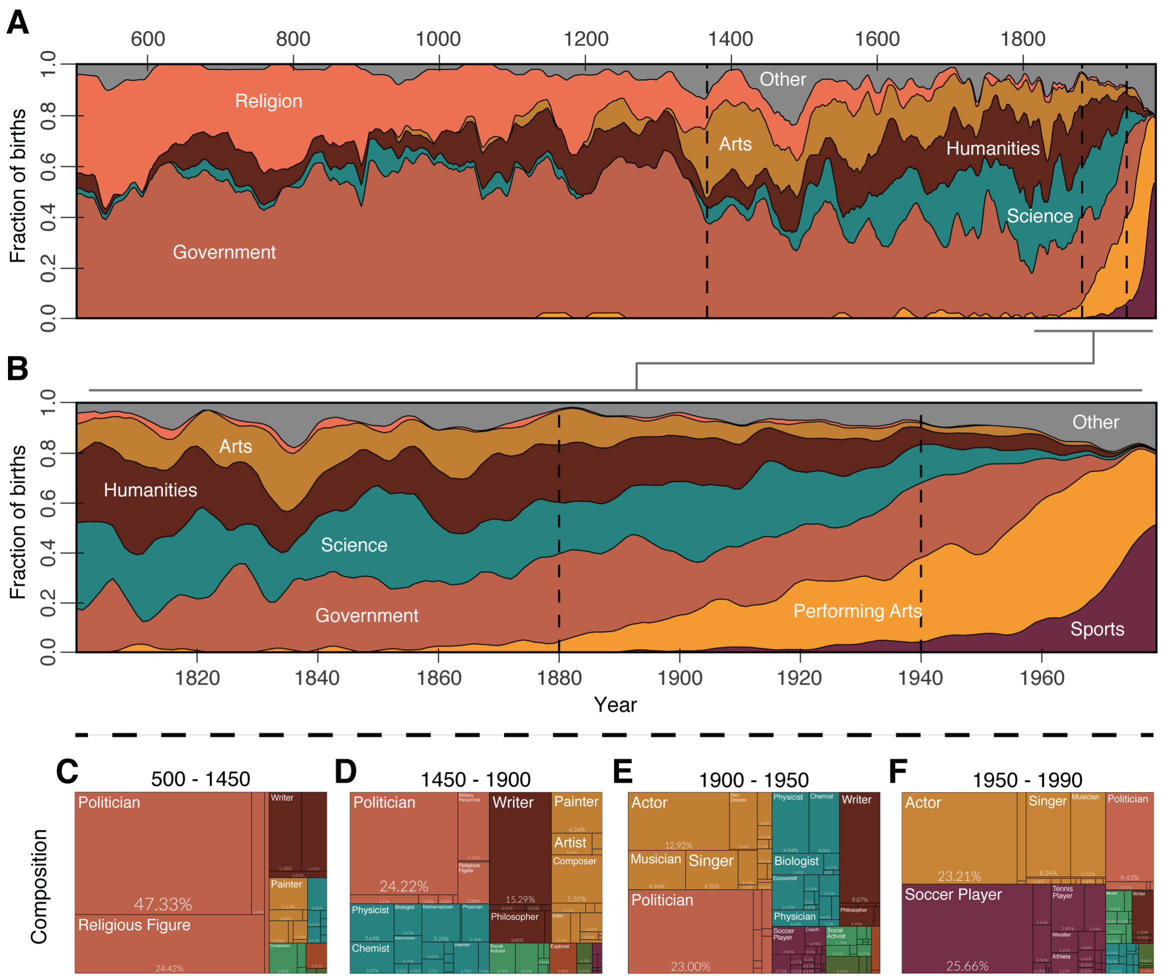

Figure 2: Changes in composition. Occupations of the biographies in the Pantheon 1.0 dataset for (**A**) the period between 500 and 1990 and (**B**) between 1800 and 1990. **C**-**F** Cross-sections for each technological period: (**C**) 500-1450 (scribal period), (**D**) 1450-1900 (printing), (**E**) 1900-1950 (film and radio), and (**F**) 1950-1990 (television).

Figures 2A and B show the fraction of biographies corresponding to individuals associated with each cultural domain. Figure 2C to F shows cross-sections for different technological periods: scribal culture (<1450), printing (1450-1900), film and radio (1900-1950), and television (1950-2000).

Figure 2 provides evidence in support of the hypothesis that changes in communication technologies are accompanied by changes in the information from each time period that we remember today. The transition from scribal culture to printing is associated with a sharp increase in the number of painters, composers, and scientists that we remember today, but also, with a large decrease in the fraction of biographies associated with religious figures (Figures 2C-2D).

After printing we have the introduction of film and radio, which was accompanied by a shift in the arts and a sharp increase in the number of memorable performers—such as actors, singers and musicians (Figure 2D-2E). Finally, Figure 2E and 2F show that athletes, such as soccer players, basketball players, and racecar drivers, became memorable with the adoption of television. We note that these observations suggest a direction of causality, since actors, singers, and musicians have been part of society for centuries (both Ancient Greeks and Elizabethan England playwrights like Shakespeare employed actors), but performers and their performances were not memorable in the absence of media capable of recording performances—such as film and radio. A similar story holds for the rise of memorable athletes, since athletes already existed at the time of Ancient Greece. The other direction of causality—that the rise of performers caused the invention of film and radio, or that soccer players invented television—is unlikely.

We test the statistical significance of the changes in composition by means of a Pearson's chi-squared test, used to compare categorical variables. We compare four non-overlapping time periods: 500-1300, 1450-1850, 1900-1930, and 1950-1990. The changes are significant with p-values of less than 10^{-190} (see SM). To check the robustness of our results, we perform the same analysis using different thresholds for our proxy of memorability *L*, finding that these results are robust to these changes (see SM).

We note that observations for most recent years need to be interpreted carefully because of two reasons. First, the more recent biographies in our dataset contain a mix of characters that are memorable (i.e. Barack Obama, as the first African American president of the United States), with characters whose presence in today's collective memory may not necessarily be long lasting (like teen pop icons and reality show celebrities). So the picture obtained for recent decades is not the one we expect to be representative of those decades in the future. Nevertheless, we can safely assume that this issue does not affect our historical data prior to the twentieth century, since these transient effects should be minimal centuries after a person's death. Second, we also note that data for the most recent years is affected by differences in the life cycle of individual's memorability, since individuals with different careers peak at different ages. Soccer players, for example, peak around their late twenties or early thirties [36], so our dataset should contain all soccer players born in the 1950s that became memorable as players. Politicians and scientists, on the other hand, often become globally memorable much later in life [37], and hence, we may be missing some influential individuals that are yet to reach global recognition. Both of these effects imply that fifty years from now the fraction of our collective memory allocated to sports players will be smaller than what we observe in our data today. In other words, we expect the focus of history to adjust as time continues to elapse.

6. Discussion

Here we used two large biographical datasets to test the ideas of Innis, McLuhan, and Eisenstein regarding the effect of communication technologies in human collective memory. First, we studied the size of our collective memory by looking at changes in the per-capita number of biographies from a time period that we remember today. Here, both datasets revealed two breaks in our collective memory that coincided with the introduction of printing, and the birth of the public sphere. The second break gave rise to a period where the number of biographies that we remember today begun to grow faster than global population.

Second, we documented a strong connection between the predominant communication technology of a time period and the occupations of the biographies recorded in both datasets. We showed that as communication technologies shifted from writing to printing, and from printing, to film, radio, and television, new elites became memorable. Prior to printing our historical record was biased towards institutional elites, including mostly political and religious leaders (which represent >70% of all biographies in the Pantheon 1.0 dataset). The introduction of printing, however, enabled the emergence of a new cultural elite, populated by scientists and artist. Similar shifts happened with the introduction of radio and cinema, and television, which gave rise to an elite of performers, including actors, musician, and athletes. We note, however, that new communication technologies do not always replace older forms of recording and diffusing information. While printing did replace the process of manual transcriptions that characterized scribal culture, film and radio did not replace printing, but grew together with it (the twentieth century was for the most part a good century for printing). So the patterns observed for most recent centuries should be interpreted as the patterns that emerge from a combination of communication technologies.

But while our study helps provide evidence in favor of the questions explored by Innis, McLuhan, and Eisenstein, it also motivates some new questions to explore. One of those questions is how recent changes in communication technologies are shaping the volume and types of information that we record. While the biographical nature of our data limits our ability to empirically answer this question, the connections between media and messages that are prevalent in our study provide us with a few hints. Prior to printing, history was limited to the most powerful institutional elites in the world. Now, we live in a world in which history is almost personalized, since billions of individuals now leave traces that could be used to reconstruct biographical data through personal acts of communication (emails, text messages, and social media posts). Of course, this does not mean that everyone will become memorable, but maybe, that memorability will now have a chance to spread over a wider number of people who may now enjoy intermediate levels of memorability and fame. This is an effect that has already been observed in the context of creative industries [38]. Going forward, however, hypothesis like this one will not need to remain as mere speculations, since the rise of digitized historical records is providing us with an increasing ability to statistically study historical events.

7. Acknowledgments

All authors acknowledge support from the MIT Media Lab Consortia and from the Metaknowledge Network at the University of Chicago. We thank comments from Manuel Aristaran, Domink Hartmann, Iyad Rahwan, Flavio Pinheiro, William Powers, and Ethan Zuckerman.

References

[1] W. H. Zurek, "Decoherence and the Existential Interpretation of Quantum Theory, or 'No Information Without Representation'," in *From Statistical Physics to Statistical Inference and Back*., 1994, pp. 341-350.

[2] Rolf Landauer, "The Physical Nature of Information," *Physics Letters A*, vol. 217, pp. 188-193, 1996.

[3] Seth Lloyd, "Computational Capacity of the Universe," *Physical Review Letters*, vol. 88, no. 23, p. 237901, June 2002.

[4] César A. Hidalgo, *Why Information Grows*. New York: Basic Books, 2015.

[5] Elizabeth L. Eisenstein, *The printing press as an agent of change*. Cambridge, United Kingdom: Cambridge University Press, 1979.

[6] Elizabeth L. Eisenstein, *The Printing Revolution in Early Modern Europe*. Cambridge: Cambridge University Press, 1983.

[7] Shaheed Nick Mohammed, *Communication and the Globalization of Culture: Beyond Tradition and Borders*. Plymouth, UK: Lexington Books, 2011.

[8] Asa Briggs and Peter Burke, *A Social History of the Media: From Gutenberg to the Internet*. Cambridge, UK: Polity Press, 2009.

[9] Marshall McLuhan, *Understanding Media: The Extensions of Man*. London: Routledge & K.Paul, 1964.

[10] Marshall Poe, *A History of Communications: Media and Society from the Evolution of Speech to the Internet*. New York: Cambridge University Press, 2011.

[11] Bill Kovarik, *Revolutions in Communication: Media History from Gutenberg to the Digital Age*. London: Continuum, 2011.

[12] Harold Innis, *The Bias of Communication*. Toronto: Toronto University Press, 1951.

[13] Harold Innis, *Harold Innis's History of Communications: Paper and Printing-Antiquity to Early Modernity*, Michael R. Cheney, Paul Heyer, John Durham Peters William J. Buxton, Ed. London: Rowman & Littlefield, 2015.

[14] Jurgen Habermas, *The Structural Transformation of the Public Sphere*. Cambridge, United States: MIT Press, 1991.

[15] James V. Wertsch, *Voices of Collective Remembering*. United Kingdom: Cambridge University Press, 2002.

[16] Maurice Halbwachs, *La Memoire collective*. Paris: Albin Michel, 1997[1950].

[17] Jan Assmann, *Das Kulturelle Gedächtnis: Schrift, Erinnerung und Politische Identität in frühen Hochkulturen*. Munich: Beck, 1997.

[18] Pierre Nora, "Between Memory and History: Les Lieux de Memoire," *Representations*, vol. 26, no. Special Issue: Memory and Counter-Memory, pp. 7-24, Spring 1989.

[19] Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo, "Pantheon: A Dataset for the Study of Global Cultural Production," *arXiv:1502.07310v1 [physics.soc-ph]* , February 2015.

[20] Charles Murray, *Human Accomplishment*. New York, NY, USA: Harper Collins, 2003.

[21] Steven Skiena and Charles Ward, *Who's Bigger? Where Historical Figures Really Rank*. Cambridge, UK: Cambridge University Press, 2013.

[22] 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 Jean-Baptiste Michel, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, vol. 331, no. 6014, pp. 176-182, January 2011.

[23] Peter Turchin, "Arise 'Cliodynamics'," *Nature*, vol. 454, no. 3, July 2008.

[24] Laura Spinney, "History as science," *Nature*, vol. 488, pp. 24-26, 2012.

[25] Maximilian Schich et al., "A network framework of cultural history," *Science*, vol. 345, no. 6196, pp. 558-562, August 2014.

[26] Shahar Ronen et al., "Links that speak: The global language network and its association with global fame," *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, December 2014.

[27] Eric M. Clarka, Suma Desuc, Morgan R. Frankc, Andrew J. Reagana, Jake Ryland

Williamsa, Lewis Mitchelld, Kameron Decker Harrise, Isabel M. Kloumannf, James P. Bagrowa, Karine Megerdoomiang, Matthew T. McMahong, Brian F. Tivnanb, Christopher M. Danfortha Peter Sheridan Dodds, "Human language reveals a universal positivity bias," *Procedings of the Natural Academy of Sciences of the United States of America*, vol. 112, no. 8, pp. 2389-2394, 2015.

[28] Arnout van de Rijt, Eran Shor, Charles Ward, and Steven Skiena, "Only 15 Minutes? The Social Stratification of Fame in Printed Media," *American Sociological Review*, vol. 78, no. 2, pp. 266-289, 2013.

[29] Ludvig Bohlin, Ann Samoilenko, Martin Rosvall, Andrea Lancichinetti Fariba Karimi, "Quantifying national information interests using the activity of Wikipedia editors," *arXiv:1503.05522*, 2015.

[30] Thomas Carlyle, *On Heroes, Hero-Worship, and the Heroic in History*. London: James Fraser, 1841.

[31] Herbert Spencer, *The Study of Sociology*. London: Henry S. King, 1873.

[32] United States Census Bureau. World Population. [Online]. http://www.census.gov/population/international/data/worldpop/table_population.php

[33] Wikimedia Foundation, "Second 2011 Wikipedia Editor Survey," http://en.wikipedia.org/wiki/Wikipedia:Systemic_bias, 2011.

[34] Idris A. Eckley Rebecca Killick, "changepoint: An R Package for Changepoint Analysis," *Journal of Statistical Software*, vol. 58, no. 3, 2014.

[35] Encyclopaedia Britannica. History of publishing. [Online]. http://www.britannica.com/topic/publishing/Newspaper-publishing#toc28665

[36] Ben Carter. When do footballers reach their peak? [Online]. http://www.bbc.com/news/magazine-28254123

[37] Wikimedia Foundation. List of US Presidents by age. [Online]. http://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_age

[38] Chris Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion, 2006.

Estimating technological breaks in the size and composition of human collective memory from biographical data; Supplementary Material

C. Jara-Figueroa[1], Amy Zhao Yu[1], César A. Hidalgo[1,2]

[1]Macro Connections, The MIT Media Lab

[2] to whom correspondence should be addressed: hidalgo@mit.edu

Table of Contents

Summary of the data sources

The Pantheon 1.0 dataset [1] contains the 11,337 biographies that had a presence in more than 25 different language editions of Wikipedia as of May 2013. The Pantheon dataset associates each biography to a place of birth, a date of birth, and an occupation using a three-level hierarchical classification that disaggregates into 88 distinct occupations—i.e. *Physics* and *Biology* are branches of *Natural Science*, just like *Natural Science* is a branch of *Science*.

The Human Accomplishments dataset [2] contains 3,869 biographies of accomplished individuals from the Arts and Sciences that are recorded in authoritative printed texts in six different languages. This dataset classifies people into 5 different inventories—Science, Philosophy, Music, Literature, and Art.

Population data comes from the historical world population estimates of the US Census Bureau [3], which reports an aggregated dataset of world population estimates starting from the year 10,000BC. We interpolate the missing years using linear splines.

The data on technology adoption used on this SM comes from two sources. We calculate the yearly number of printing presses per-capita, and the number of journals per-capita using data from Wikipedia [4,5]. The data for radio and cinema, and television comes from the Historical Cross-Country Technology Adoption (HCCTA) Dataset [6], in particular we use the variables “Radios”, and “Televisions”.

Changes in communication technologies

In this section we will review data on the changes in communication technologies discussed in the main text: printing, mass media, radio and cinema, and television.

The printing press was a technology that spread very quickly over Europe, making its invention a relatively discrete event in history. Before the invention of printing, the number of manuscript books in Europe could be counted in thousands. By 1500, after only 50 years of printing, there were more than 9,000,000 books [7]. Figure SM1-A shows the number of printers per million people in the world between 1420 and 1530.

The early eighteenth century was a period of adjustment to the new forms of printed media. Newspapers were created in the seventeenth century, but had not yet matured as a technology. The coffeehouses increased the accessibility to these new formats, and the freedom of press enabled the spread of opinions, rather than just facts. It became a period of intense competition between journals that peaked around 1720, and eventually settled down. Figure SM1-B shows the number of functioning journals, established between 1600 and 1800, per million people in the world between.

The radio and the television were technologies that also spread relatively fast, at least in the industrialized world. Figure SM3-C shows the fraction of the countries in the HCCTA dataset that have radios (ref) and television (blue).

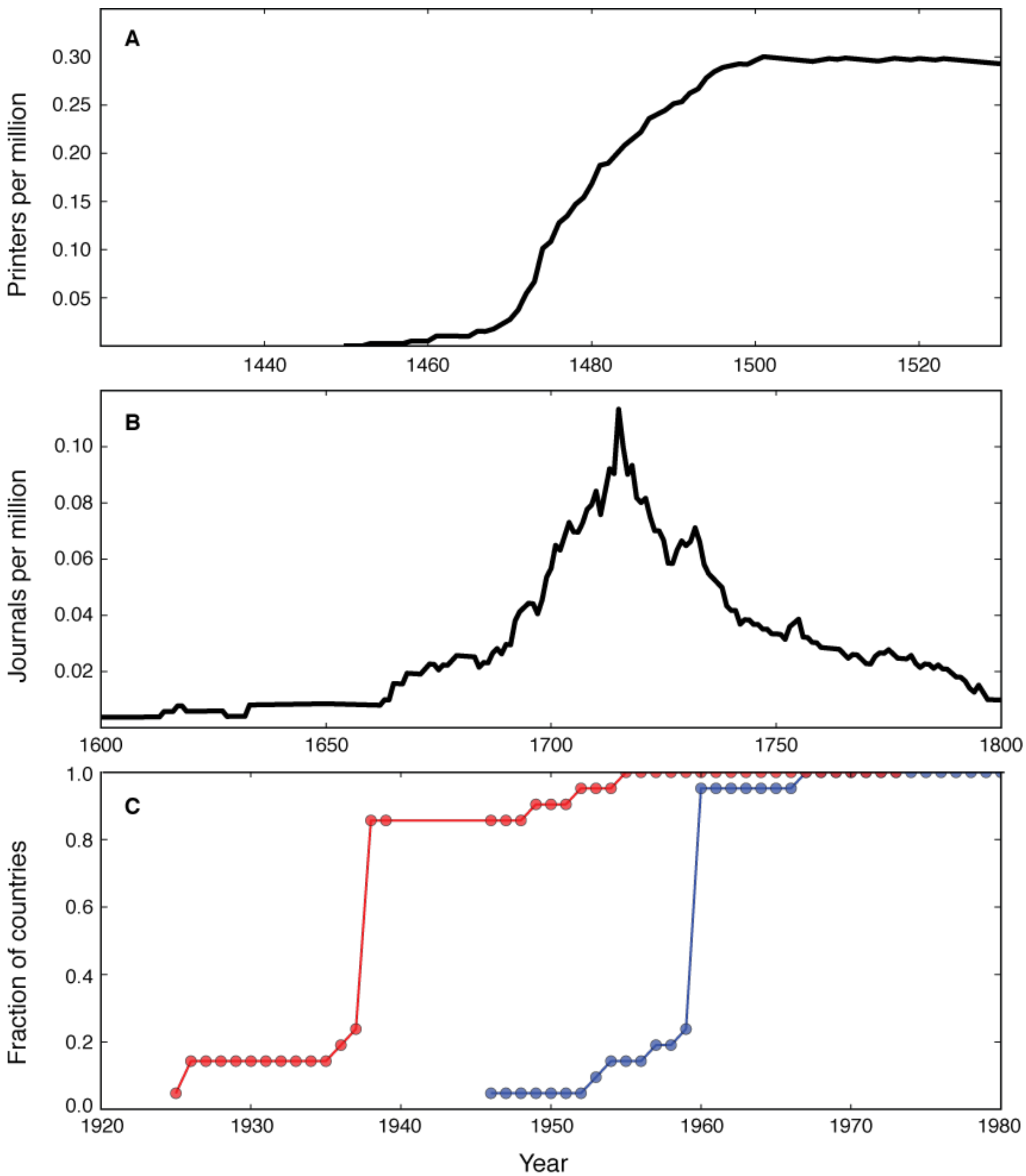

Figure SM1: Changes in communication technologies. This figure shows the adoption of the communication technologies discussed in the main text. **A** shows to the number of printers per million people in the world [4], **B** shows the number of journals—founded between 1600 and 1800—per million people in the world [5], and **C** shows the fraction of countries from the HCCTA dataset [6] that adopted radio (red) and television (blue).

Changes in size

In this section we explain the changepoint analysis for the per-capita birthrate of memorable characters and perform some robustness checks of our result. Note that we restrict our analysis between 500 BC and 1960 because of the recency effects discussed in the main text.

A 15-years window was used for the analysis presented on the main text. We divide the number of people born into each window by the average world population and the size of the time window. The world population for all years was calculated starting from the US Census Bureau data on population [3] and interpolating it using linear splines.

All changepoint analyses were performed using the 'changepoint' package available for R, developed by Rebecca Killick and Idris Eckley from Lancaster University [8]. In particular we used the *cpt.mean* function to determine changes in mean. We used 'CUSUM' as a test statistics, the 'SegNeigh' method to minimize the cost function, and a scaling of the "SIC" (Schwarz Information Criterion) as penalty function—the scaling factor is 10^{-6} for Pantheon 1.0 and 10^{-4} for Human Accomplishment, when using births per million. For more details on the working of this method see reference [8].

Varying time window

To check the robustness of our results, we repeat the changepoint analysis for different time windows—5, 10, 15, 20, and 25-year windows.

Due to the scarcity of the data in the Human Accomplishment dataset, especially in the first millennia, we use variable size time windows. We set a lower threshold of 5 characters per time window. The 5-years bins, for example, are built such that each time bin is the smallest time window bigger than 5 years, having at least 5 characters born inside it. If we do not include this lower bound, only 68% of all time windows have non-zero people when using 5-years bins, and this noise makes the changepoints harder to detect. Pantheon 1.0 does not have this problem.

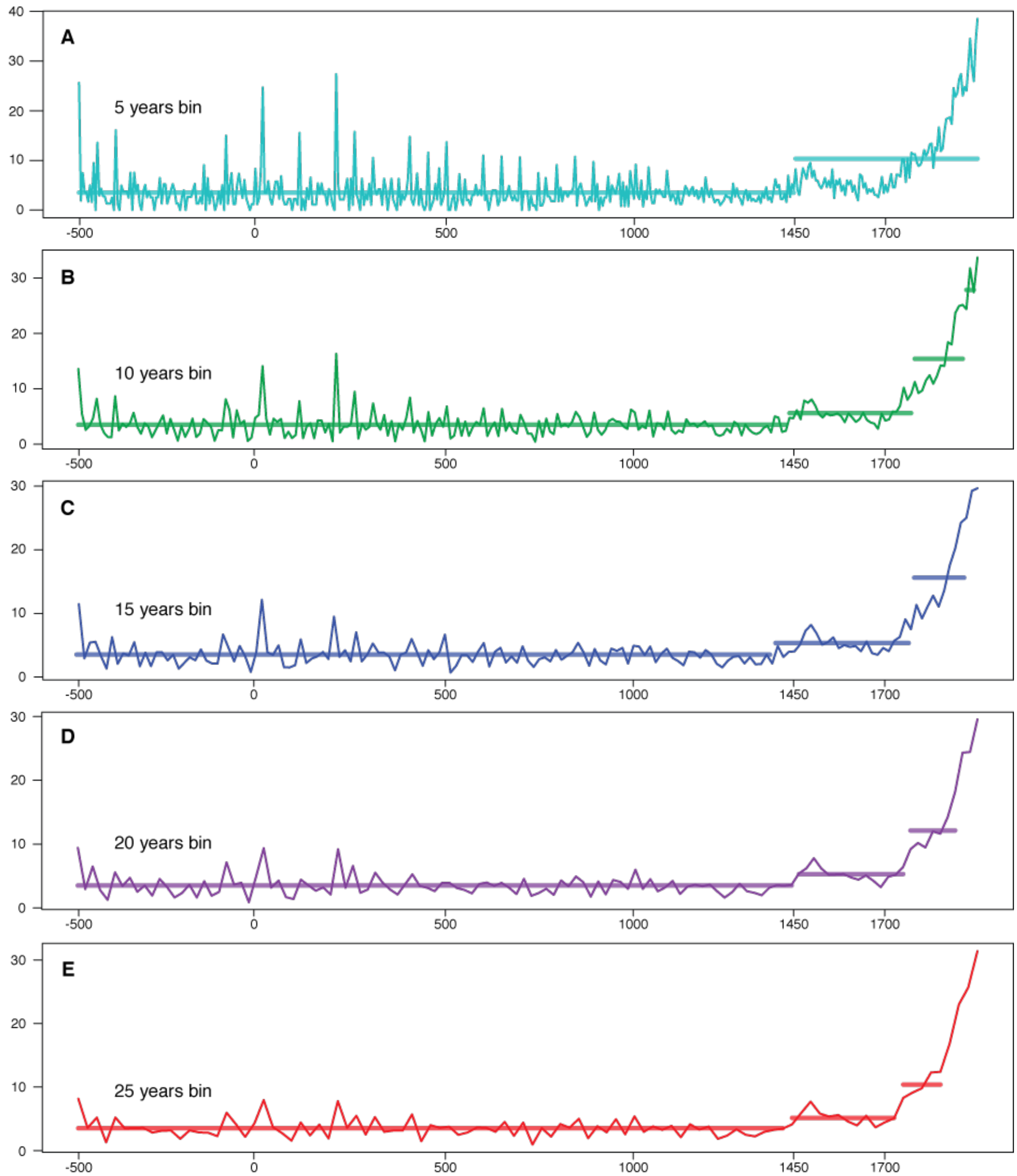

Figure SM2: Per capita yearly birth rate, in births per year per billion, of people in the Pantheon 1.0 dataset, and results of the changepoint analysis, using 5, 10, 15, 20, and 25 years bins.

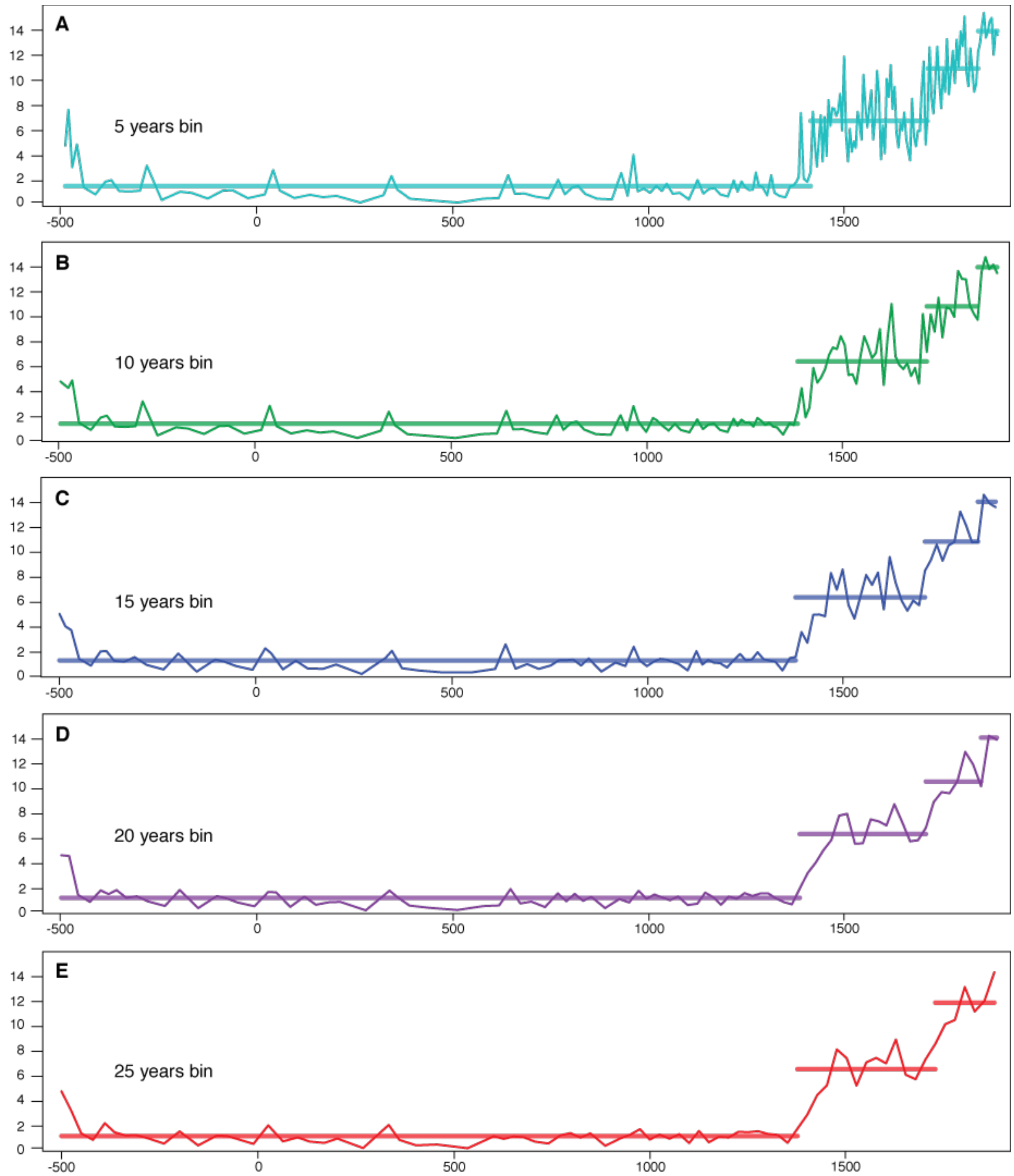

Figure SM3: Per capita yearly birth rate, in births per year per billion, of people in the Human Accomplishment dataset, and results of the changepoint analysis, using 5, 10, 15, 20, and 25 years bins.

From Figure SM2 and SM3 we see that the position of the first break is independent of the time window. The second break also remains the same, with the only exception being the 5-years window in Pantheon 1.0 dataset.

Table SM1 and SM2 summarize the results of the changepoint analysis for the different binning size for both the Pantheon 1.0 and the Human Accomplishment datasets.

Window size	N	1st break	2nd break	1st mean [bpyb]	2nd mean [bpyb]
5	490	1445-1450	-	3.530	-
10	245	1420-1430	1760-1770	3.515	5.625
15	163	1375-1390	1750-1765	3.514	5.336
20	122	1420-1440	1720-1740	3.520	5.311
25	98	1400-1425	1700-1725	3.515	5.147

Table SM1: Results of the changepoint analysis for different time windows using the Pantheon 1.0 dataset.

Window size	N	1st break	2nd break	1st mean [bpyb]	2nd mean [bpyb]
5	183	1421-1428	1723-1728	1.596	6.775
10	132	1386-1396	1716-1726	1.436	6.424
15	106	1379-1394	1709-1724	1.341	6.395
20	90	1375-1395	1695-1715	1.266	6.380
25	78	1370-1395	1720-1745	1.219	6.573

Table SM2: Results of the changepoint analysis for different time windows using the Human Accomplishment dataset.

Higher number of languages

The fact that our results are independently validated by two datasets, collected with two different definitions of "fame"—the Pantheon 1.0 dataset focuses on memorable characters and the Human Accomplishment dataset focuses on accomplished characters—imply that they are not dependent of the particular definition of fame.

As yet another robustness check, we repeat the changepoint analysis with a different threshold in number of languages L from the Pantheon 1.0 dataset. L is the minimum number of Wikipedia language editions a character needs to appear in the Pantheon 1.0 dataset [1]. We use $L = 30$, 35, and 37. The results are shown in Table SM3 and in Figure SM4.

L	N	1st break	2nd break
30	8238	1433	1763
35	5841	1427	1727
37	5112	1432	1792

Table SM3: Results of the changepoint analysis for different values of L. L is the minimum number of Wikipedia language editions a character needs to appear in the Pantheon 1.0 dataset [1].

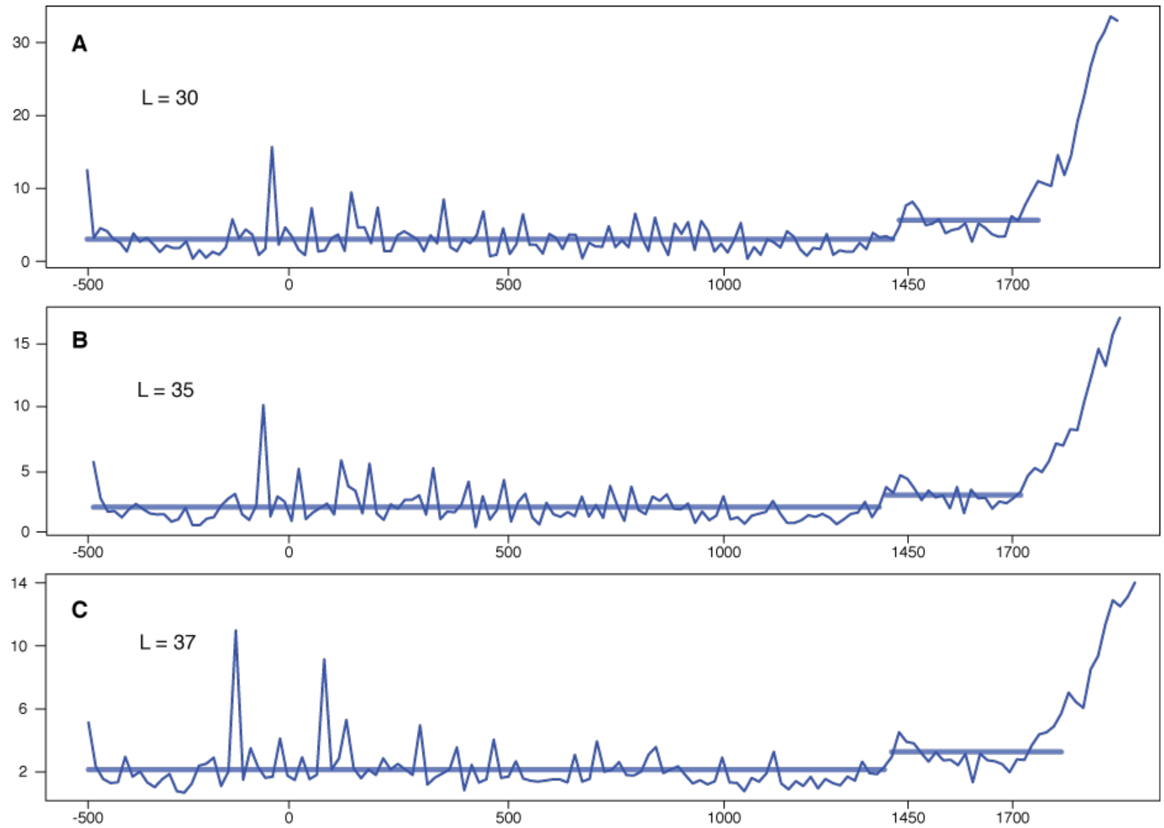

Figure SM4: Yearly per-capita birthrate of globally memorable characters from the Pantheon 1.0 dataset, in births per year per billion people in the world, for different values of *L*. The horizontal lines are the result of the changepoint analysis.

Serial autocorrelation

We take a closer look at the yearly per-capita birth rate during the time before the first break for the Pantheon 1.0 dataset. In order to model a signal as white noise, there should be no statistical evidence of serial autocorrelation within the time series. In this section we perform the Ljung-box test on the first segment of the time series, to test for autocorrelations up to some lag. The null hypothesis of this test is that there are no correlations between different observations. We only use Pantheon 1.0, since the Human Accomplishment dataset is too sparse in this time period.

We use $h = ln(N)$ as the time lag, where N is the number of points in the time series.

Time window	h	Time lag [years]	N	p-value
5 years	5	25	380	2.008e-03
5 years	4	20	380	1.383e-02
10 years	5	50	192	2.797e-06
10 years	4	40	192	0.267

15 years	4	60	125	0.680
20 years	4	80	96	0.385
25 years	4	100	76	3.796e-06
25 years	3	75	76	1.256e-03

Table SM4: Result of the Ljung-box test for different time binnings. There is no statistically significant evidence of serial autocorrelations, therefore the data can be modeled as white noise.

According to the p-values displayed in Table SM4, there is no statistical evidence of serial correlation, so the data could be modeled as white noise.

Changes in composition

Due to the scarcity of the data in older years, we restrict our analysis of the occupations of memorable characters to the time window between 500 and 1990.

Grouping description for stacked area chart

Because some occupations of the Pantheon 1.0 dataset include very few characters, we use the second level of aggregation provided by the dataset. Pantheon classified characters into 8 "domains" and 27 "industries". For our analysis, we disaggregate domains, based on industries, into categories separating mainly between politicians and religious figures, and between arts and performing arts.

The chart on Figure 2A-B of the main text and Figure SM5 was built using 8 categories (see Table SM5): *Arts*, *Performing arts*, *Humanities*, *Government*, *Religion*, *Science*, *Sports*, and *Other*.

The *arts* domain is split into *performing arts*—including the *dance*, and *film and theatre* industries, plus all occupations from the *music* industry, excluding *composer*–and *arts*—including the *design* and *fine arts* industries and the *composer* occupation. The *religion* industry is grouped by itself, and all the other industries under the *institutions* domain are grouped together under *government*. The *team sports* industry is considered under *sports*. Industries *science and technology* and *humanities* remain unchanged. Finally, *individual sports*, along with the domains *business and law*, *exploration*, and *public figure* are grouped as *other*.

The three largest occupations aggregated under *other* are very small, corresponding to *tennis player*, *social activist*, and *racecar driver*, with 161, 114, and 104 individuals respectively. Let us note that changes in the category *other*, are also captured by other categories. For example, there is an observed increase in the number of *tennis players* in the second half of the 1900s due to the adoption of television, but that change is already captured by the change in the *sports* category.

Category	Occupation	Industry	Domain	Number of people
Arts	Fashion designer	Design	Arts	10
Arts	Game designer	Design	Arts	4
Arts	Designer	Design	Arts	16
Arts	Comic artist	Design	Arts	24
Arts	Architect	Design	Arts	73
Arts	Photographer	Fine arts	Arts	12
Arts	Sculptor	Fine arts	Arts	21
Arts	Artist	Fine arts	Arts	88
Arts	Painter	Fine arts	Arts	178
Arts	Composer	Music	Arts	225
Performing arts	Dancer	Dance	Arts	12
Performing arts	Comedian	Film and theatre	Arts	4
Performing arts	Film director	Film and theatre	Arts	177
Performing arts	Actor	Film and theatre	Arts	1193
Performing arts	Musician	Music	Arts	381
Performing arts	Singer	Music	Arts	437
Performing arts	Conductor	Music	Arts	11
Humanities	Historian	History	Humanities	48
Humanities	Critic	Language	Humanities	5
Humanities	Journalist	Language	Humanities	19
Humanities	Linguist	Language	Humanities	21
Humanities	Writer	Language	Humanities	955
Humanities	Philosopher	Philosophy	Humanities	281
Government	Judge	Government	Institutions	9
Government	Public worker	Government	Institutions	14
Government	Diplomat	Government	Institutions	36
Government	Nobleman	Government	Institutions	116
Government	Politician	Government	Institutions	2528
Government	Military personnel	Military	Institutions	223
Government	Pilot	Military	Institutions	9
Religion	Religious figure	Religion	Institutions	517
Science	Computer scientist	Computer science	Science and technology	34
Science	Engineer	Engineering	Science and technology	41
Science	Inventor	Invention	Science and technology	67
Science	Statistician	Math	Science and technology	4
Science	Mathematician	Math	Science and technology	157

Science	Physician	Medicine	Science and technology	143
Science	Geologist	Natural sciences	Science and technology	10
Science	Archeologist	Natural sciences	Science and technology	13
Science	Astronomer	Natural sciences	Science and technology	83
Science	Biologist	Natural sciences	Science and technology	141
Science	Chemist	Natural sciences	Science and technology	220
Science	Physicist	Natural sciences	Science and technology	268
Science	Political scientist	Social sciences	Science and technology	7
Science	Anthropologist	Social sciences	Science and technology	11
Science	Geographer	Social sciences	Science and technology	14
Science	Sociologist	Social sciences	Science and technology	15
Science	Psychologist	Social sciences	Science and technology	38
Science	Economist	Social sciences	Science and technology	102
Sports	American football player	Team sports	Sports	1
Sports	Hockey player	Team sports	Sports	2
Sports	Cricketer	Team sports	Sports	2
Sports	Baseball player	Team sports	Sports	5
Sports	Referee	Team sports	Sports	10
Sports	Basketball player	Team sports	Sports	71
Sports	Coach	Team sports	Sports	75
Sports	Soccer player	Team sports	Sports	1064
Other	Producer	Business	Business and law	12
Other	Businessperson	Business	Business and law	79
Other	Lawyer	Law	Business and law	17
Other	Astronaut	Explorers	Exploration	32
Other	Explorer	Explorers	Exploration	70
Other	Social activist	Activism	Public figure	114
Other	Companion	Companions	Public figure	101
Other	Pornographic actor	Media personality	Public figure	11
Other	Celebrity	Media personality	Public figure	21
Other	Chef	Media personality	Public figure	2
Other	Magician	Media personality	Public figure	4
Other	Presenter	Media personality	Public figure	19
Other	Model	Media personality	Public figure	30
Other	Pirate	Outlaws	Public figure	9
Other	Mafioso	Outlaws	Public figure	13
Other	Extremist	Outlaws	Public figure	34
Other	Golfer	Individual sports	Sports	2

Other	Snooker	Individual sports	Sports	3
Other	Mountaineer	Individual sports	Sports	5
Other	Gymnastic	Individual sports	Sports	7
Other	Martial arts	Individual sports	Sports	7
Other	Skater	Individual sports	Sports	9
Other	Boxer	Individual sports	Sports	14
Other	Skier	Individual sports	Sports	17
Other	Swimmer	Individual sports	Sports	20
Other	Cyclist	Individual sports	Sports	29
Other	Chessmaster	Individual sports	Sports	30
Other	Wrestler	Individual sports	Sports	44
Other	Athlete	Individual sports	Sports	74
Other	Racecar driver	Individual sports	Sports	104
Other	Tennis player	Individual sports	Sports	161

Table SM5: Description of categories used to analyze the changes in composition.

Area chart

The stacked area chart, Figure 2A-B from the main text, was constructed in the following way:

1. Build the smallest time window that starts in the given year, that includes at least 50 characters, and that is smaller than 100 years.
2. Calculate the fraction of the characters in each category born into each time window.
3. Apply an 11-year moving average to Figure 2A, and a 5-year moving average to Figure 2B to smooth the noise. Normalize the frequencies after the smoothing.

The non-smoothed stacked chart is shown in Figure SM5.

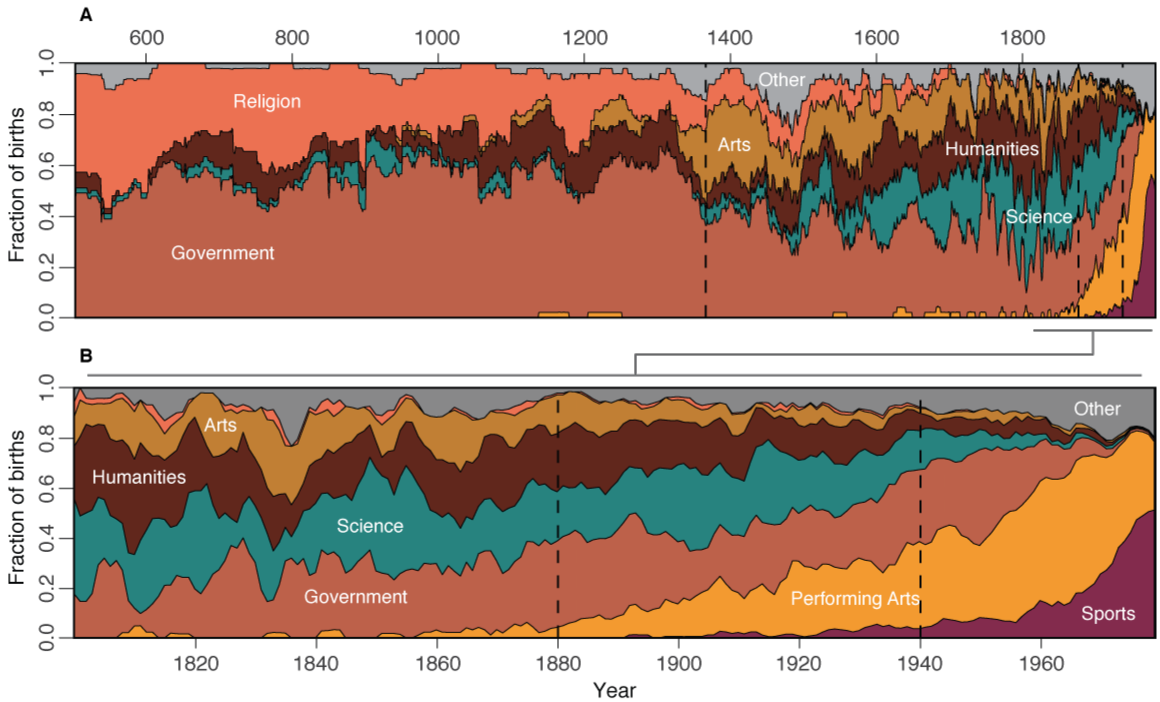

Figure SM5: Non-smoothed stacked area chart showing the composition of human collective memory between 500 and 1990 (**A**), and between 1800 and 1990 (**B**).

Chi2 test for changes in composition

The pearson's chi-squared test is applied to sets of categorical data. In particular, a test of independence compares two samples of a categorical variable (in our case the occupation of each character) and tests whether they come from the same distribution. The null hypothesis is that both samples are drawn from the same distribution, and the test statistics has a chi-squared distribution [9].

	Sample limits	**Sample size**
Writing	500-1300	540
Printing	1450-1850	1350
Radio and cinema	1900-1930	1123
Television	1950-1990	3041

Table SM6: Sample bounds to determine the composition of each technological era.

To sample each technological period we use the limits shown on Table SM6. The p values resulting from comparing each technological period are shown on Table SM7.

	Printing	Radio and cinema	Television
Writing	1.3e-93	4.6e-151	<1.3e-311
Printing	-	3.1e-92	<1.3e-311

Radio and cinema		-	<1.7e-287

Table SM7: Results of the pearson's chi-squared test comparing the occupations of characters from each technological period.

We repeat the same analysis, but changing the minimum number of Wikipedia language editions the biography of each character must have to appear on the dataset. We use *L = 30, 35, and 37*, and find that the results are consistent. The p-values of *L = 37* are shown in Table SM8.

	Printing	Radio and cinema	Television
Writing	5.0e-71	1.7e-95	6.6e-215
Printing	-	3.6e-53	5.9e-285
Radio and cinema		-	9.9e-153

Table SM8: Results of the pearson's chi-squared test comparing the occupations of characters from each technological period, for *L = 37*.

Fraction of characters in each category

The fraction of memorable characters belonging to each category can be thought as the probability that a memorable character, born into each year belongs, to a given category. Figure SM6 shows the estimation of this probability, for each year and for each category, calculated following the sane procedure used to create Figure SM5 and Figure 2 (from the main text). Figure SM6 also shows the 95% confidence interval for this estimator.

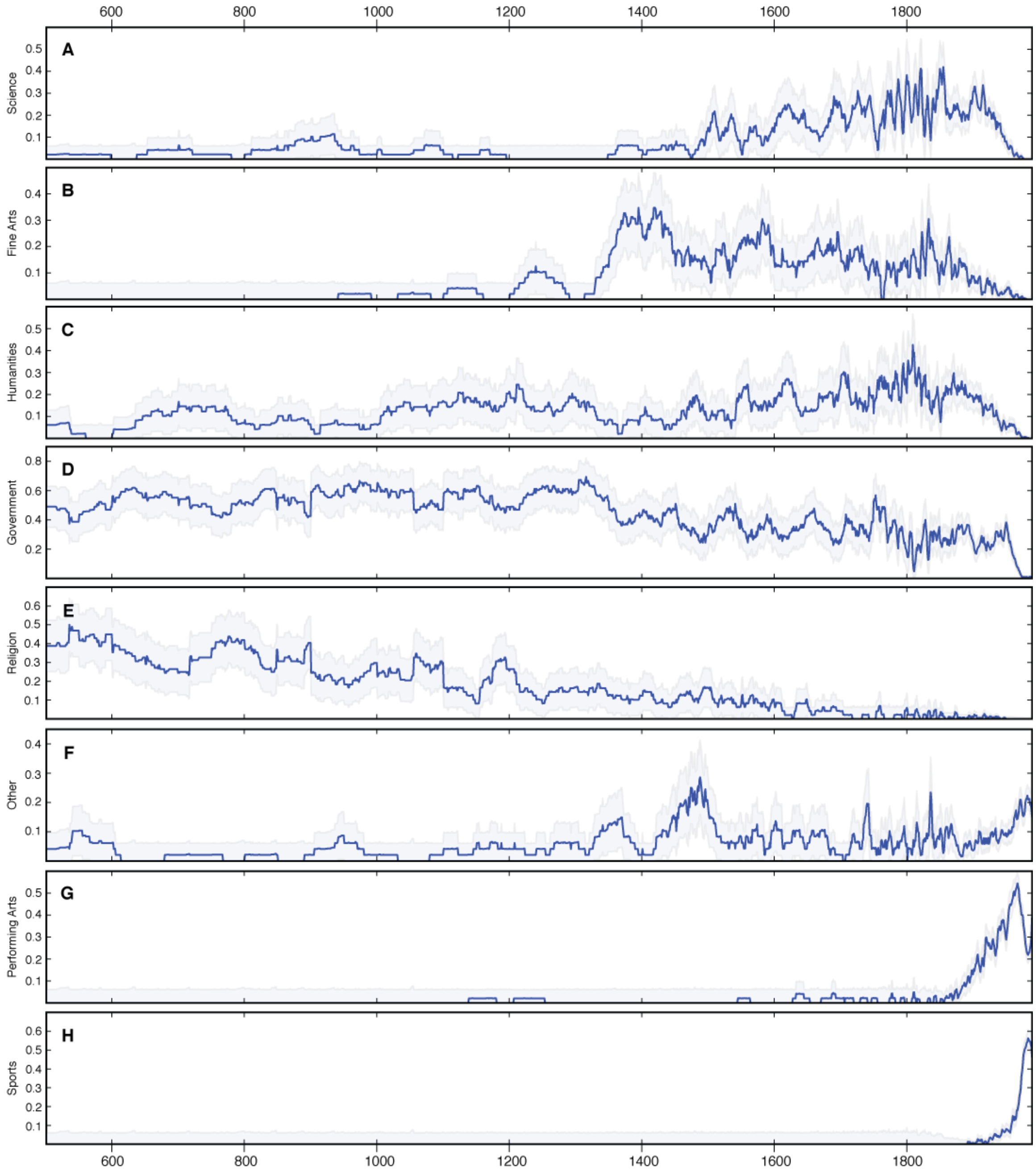

Figure SM6: Yearly probability that a memorable characters in the Pantheon 1.0 belongs to each category. The diffuse blue contour corresponds to the 95% confidence interval.

Printing break

In the main article we state that the major changes in the composition of human collective occur previous to the introduction of printing, radio and cinema, and television. Here we show that between 500 and 1850 there aren't any other significant discrete breaks, besides the one related to the invention of the removable type press.

We analyze each year in the following way:

1. Build the smallest time window that ends in the given year, that includes at least 50 characters, and that is smaller than 100 years.
2. Build the smallest time window that starts in the given year, that includes at least 50 characters, and that is smaller than 100 years.
3. Calculate the chi-squared statistics, and reject with a p-value smaller than 0.005.
4. For every group of years for which the hypothesis has been rejected, calculate the mean

Figure SM7-A shows the value of the test statistics as a function of the year. This analysis yields only one group of years for which we reject the hypothesis, with an average of 1368, with a range between 1359 and 1377. Meaning that the composition of our collective memory after these years was significantly different from the composition before these years. We should not forget that all dates correspond to birthdates of historical characters.

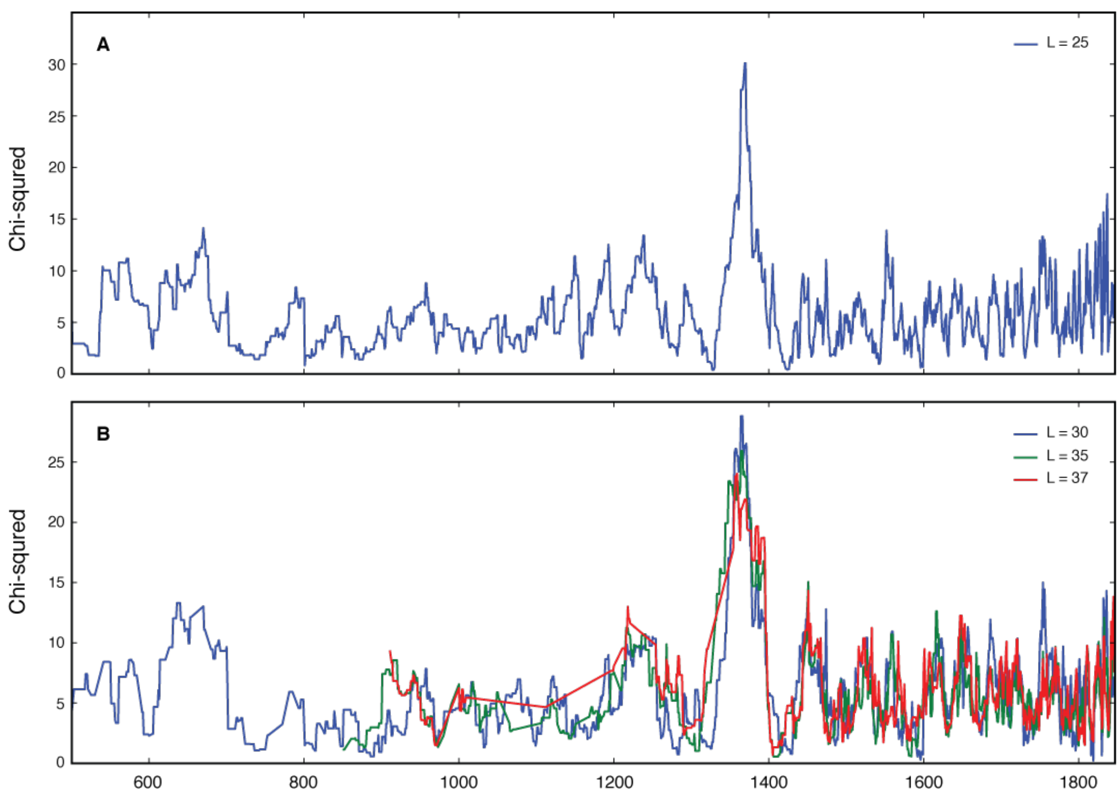

Figure SM7: Chi-squares statistics comparing the composition before and after a given year for the full Pantheon 1.0 dataset (**A**), and for different values of *L* (**B**). Most of the big breaks occur around year 1368.

Furthermore, to show that this break is not dependent of the particular definition of memorability, we repeat the analysis with a different threshold in number of languages *L*. We use *L = 30*, *35*, and *37*. The dataset doesn't provide lower values of *L*, so is not possible to try them. Higher values of *L* are difficult to test, because the sample size already decreases to less than 30% with *L=45* and concentrated after 1500. Figure SM7-B shows the evolution of the chi-squared statistics and Table SM9 summarizes the position of the breaks. The curve for lower years is missing some values because of the scarcity of the data since it was not possible to build the time windows for those years according to the described procedure.

L	Mean year	Min year	Max year
30	1363	1348	1377
35	1364	1344	1394
37	1376	1354	1395

Table SM9: Table showing the location of the break in composition associated with the invention of the removable type press with different thresholds of the minimum number of languages.

Radio and cinema, and television breaks

The radio and cinema, and television breaks had the effect of bringing a new category to human collective memory. Since the time between these two inventions is very short for historical time scales—roughly 50 years—neither of them can be modeled as a discrete break because the system didn't have enough time to stabilize. Figure SM8 shows that a little before the invention of radio and cinema—around 1900— the number of Performing Artists began to increase, and before the massification of television—around 1950—the number of Sports Players began to increase.

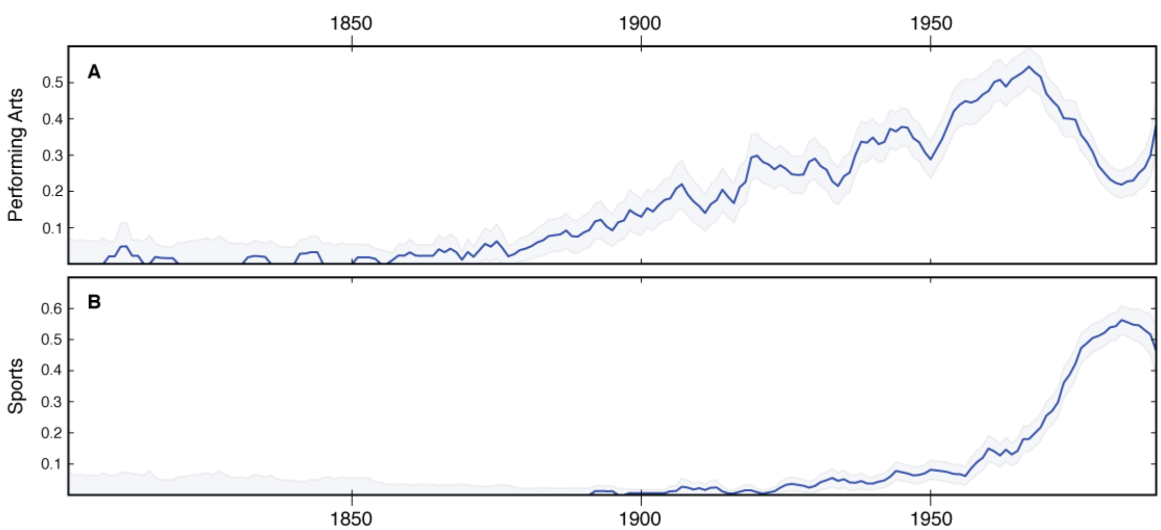

Figure SM8: Probability that a memorable character is a Performing Artist (**A**), and a Sports Player (**B**). This figure is a closer look at Figure SM8G-H.

Changes in composition using HA

The Human Accomplishment dataset has 5 different categories—*science*, *art*, *philosophy*, *music*, and *literature*—of which only science disaggregates into subcategories..

The Human Accomplishment dataset does not include people related to performing arts, therefore is not possible to use it to track changes in composition of our collective memory due to the introduction of film and radio. Since the classification is not as rich as Pantheon 1.0, the changes in composition are not as clear. Figure SM9 shows the changes in composition according to the Human Accomplishment dataset.

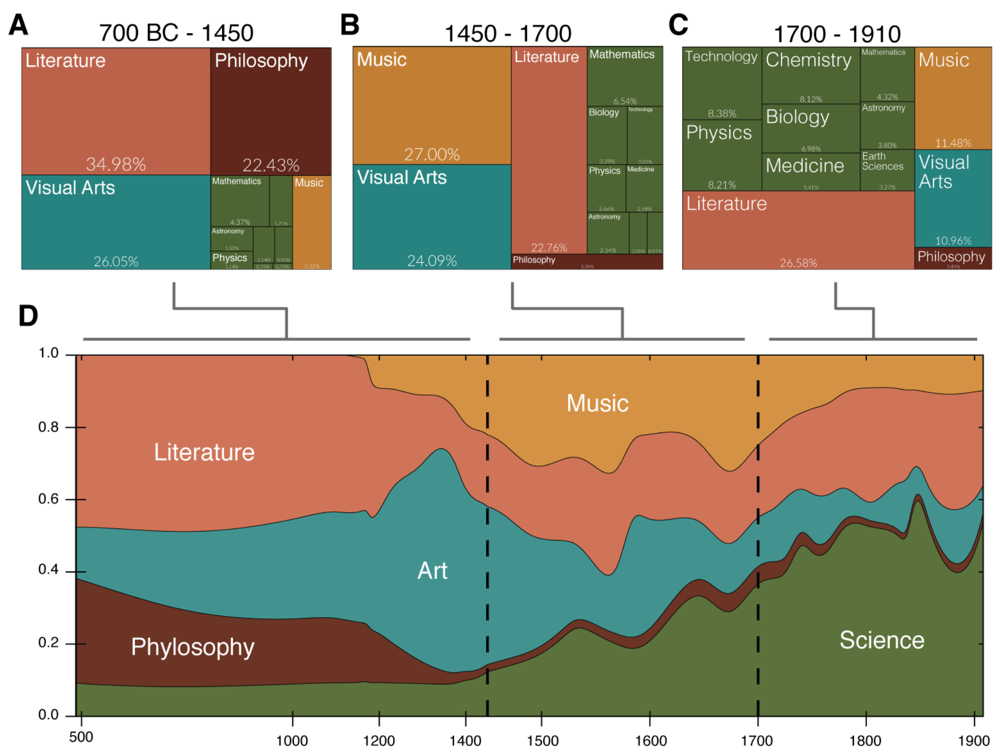

Figure SM9: Changes in composition according to the Human Accomplishment dataset. The periods 1450 -1700 and 1700-1910 are stretched.

When looking at the first time period—from 500 BC to 1450—we have to consider that the HA classification is fundamentally different from Pantheon 1.0. From Figure SM9A we see that around 1200s there is a decrease in the number of writers and an increase in the number of musicians. By 1300 there is a huge peak in the number of artists. We argue that these two events are an artifact of the dataset rather than actual changes in our collective memory.

Human Accomplishment classifies most trovers as musicians rather than writers. For example Alfonso X of Castile, king of Castille, and writer of *Cantigas de*

Santa Maria is classified in Pantheon 1.0 as a writer, and as a musician in HA. Furthermore HA does not distinguish between *composers*, and *singers*, hence performing arts blend with other forms of art.

The Human Accomplishment dataset is an effort to track people who had an effect in society, even though they are not globally remembered. The big number of artists present in this dataset, born between 1250 and 1350—41% of all births—is an evidence of this. By using the tools developed in pantheon.media we find that 34.8% of all artists born between 1250 and 1350 come from China. In this era China was ruled by the Yuan dynasty, successors to the Mongol Kublai Kahn. This dynasty promoted the arts and trade between east and west. Europe, on the other hand, was still struggling with the middle Ages, and their production of artists was low compared to China's. Most of the people born in China in this period are not globally remembered—hence they don't appear in the Pantheon dataset—but their contribution had an effect on modern society—hence they appear on the Human Accomplishment dataset.

The Human Accomplishment dataset does not include a category for religious figures. Some of the accomplished religious figures are grouped as philosophers—like Confucius—or are not included at all—like Jesus Christ. Therefore, the decrease in the memorable religious figures documented using the Pantheon 1.0 dataset, cannot be captured by the Human Accomplishment dataset.

In conclusion, the Human Accomplishment dataset is very limited to track changes in the composition of our collective memory.

Bibliography

[1] Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo, "Pantheon: A Dataset for the Study of Global Cultural Production," *arXiv:1502.07310v1 [physics.soc-ph]* , February 2015.

[2] Charles Murray, *Human Accomplishment*. New York, NY, USA: Harper Collins, 2003.

[3] United States Census Bureau. World Population. [Online]. http://www.census.gov/population/international/data/worldpop/table_population.php

[4] Wikimedia foundation. Global spread of the printing press. [Online]. https://en.wikipedia.org/wiki/Global_spread_of_the_printing_press

[5] Wikimedia foundation. List of early-modern journals. [Online]. https://en.wikipedia.org/wiki/List_of_early-modern_journals

[6] B. Hohijn D. Comin, "Cross-Country Technological Adoption: Making the

Theories Face the Facts," *Journal of Monetary Economics*, pp. 39-83, January 2004.

[7] Encyclopaedia Britannica. britannica. [Online]. http://www.britannica.com/topic/publishing/Newspaper-publishing#toc28665

[8] Idris A. Eckley Rebecca Killick, "changepoint: An R Package for Changepoint Analysis," *Journal of Statistical Software*, vol. 58, no. 3, 2014.

[9] Wikimedia Foundation. Wikipedia. [Online]. https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

[10] Ed Bullmore, Dietmar Plenz Jeff Alstott, "powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions," *PLoS ONE*, vol. 9, no. e85777, 2014, 10.1371/journal.pone.0085777.

[11] Thomas Carlyle, *On Heroes, Hero-Worship, and the Heroic in History*. London: James Fraser, 1841.

[12] Herbert Spencer, *The Study of Sociology*. London: Henry S. King, 1873.