How communication technologies impact the size and composition of human collective memory

by

Cristian I. Jara-Figueroa

MSc, Universidad de Concepción (2012)
BSc, Universidad de Concepción (2010)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Master in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

© Massachusetts Institute of Technology 2016. All rights reserved.
How communication technologies impact the size and composition of human collective memory

by

Cristian I. Jara-Figueroa

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 6, 2016, in partial fulfillment of the requirements for the degree of Master in Media Arts and Sciences

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 and prestigious people, to the development of technologies for recording and communicating information: from clay tablets to the Internet. How do these communication technologies affect the size and composition of our human collective memory? Here we use two datasets on historical characters to present empirical evidence documenting how communication technologies have shaped human collective memory. We show that changes in communication technologies, including the introduction of movable type printing and 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 given time period found in current online and offline sources. Moreover, changes in technology, such as the introduction of printing, film and radio, and television, coincide with sharp changes in the occupations of the individuals present in these biographical records. These two empirical facts provide evidence in support of theories arguing that communication technologies are more consequential to society than the messages transmitted through them. Finally, this thesis contributes an update to the Pantheon dataset that includes historical geocoded data. We hope this updated version of the Pantheon dataset will enable future work documenting the effect of new communication technologies in ancient and modern civilizations.

Thesis Supervisor: Cesar Hidalgo
Title: Associate Professor, Program in Media Arts and Sciences
How communication technologies impact the size and composition of human collective memory

by

Cristian I. Jara-Figueroa

The following people served as readers for this thesis:

Signature redacted

Iyad Rahwan
Associate Professor
Program in Media Arts and Sciences

Signature redacted

William Powers
Research Scientist
Laboratory for Social Machines
Acknowledgements

Writing is not an easy task. Specially when you have to support your ideas with facts, and connect them to the big pool of knowledge we have accumulated in our collective memory. As in any task, there is a group of people helping by giving direction, support, new ideas, or simply by sharing a cup of coffee or a glass of beer. Because people are the bread and butter of any university, I would like to thank some of them.

First, I wish to thank my thesis advisor Dr. Cesar Hidalgo for treating every one of my interruptions, questions, and complaints as something worthy of his time and attention. I would also like to thank my readers Bill Powers and Iyad Rahwan for their valuable feedback and for taking an interest in my topic. I hope I challenged you as much as you challenged me.

I would like to mention my friends and members of the Macro Connections group, who were the source of a lot of my culturally acquired knowledge. In particular Manuel (Manu) Aristaran, Luisina (Lupa) Pozzo-Ardizzi, Kevin (The Kev) Hu, Guru Mahendran, Mary Kaltenberg, Ambika (the CEO of shout!) Krishnamachar, Mia Petkova, and Jessica Rais, as well as those who used to be part of the team, Dominik (calidad de vida) Hartmann, Flavio (Vauvaro) Pinheiro, Miguel Guevara, Margaret (Maggie) Church, Elisa Castañer, Jermain Kaminski, and Amy Yu. Special mention to the ping pong table; if it is not an honorary group member, it definitely should be.

I wish to acknowledge my long-distance friends Sebastian (Seba) Arpón, Rodolfo Barria, Pía Salazar, Francisco (Pancho) de la Carrera, and Andrés Concha, who found the way to be always there no matter how bad their wi-fi connection was. I also wish to mention my short-distance friends, in particular Xavier Benavides and Juliana Nazare, who I thank for the road-trips, coffee breaks, jokes, beers, home cooked meals, advice, and for all the past and future adventures. I would also like to thank Bianca Datta (listener), Martin (the magician) Saveski, and Ivan Fernandez de Casadevante (that is only one last name, he has another one) for making time for a friend in their busy lives. I should also mention Eddy (Edmond) Awad and Chris (Criistofer) Mark, who are always ready to go grab a burger or a beer.

Finally, I would like to thank my mother Cristina Figueroa and my sister Rocío Jara (and Leeloo). I will never be able to understand how difficult it is to cope with having your son so far away in a country where everyone else stays in their home town.

Special mention to Natasha Jaques, Juliana Cherston, Peter (Pitêrr) Beshai, Andrew Schonhoffer, Maitane Iruretagoyena, Manan Ravan, Sam Beller, and Clark Wood; we should all go out and celebrate!
# Contents

Abstract ................................................. 5

1 Introduction: ........................................... 15

2 Why Collective Memory?: ......................... 19
   2.1 What is collective memory? ...................... 19
   2.2 Understanding the role of media ................ 23
   2.3 Measuring cultural memory ...................... 29

3 Breaks in size and composition of collective memory: .......... 33
   3.1 Breaks in size .................................. 34
   3.2 Changes in composition ......................... 41

4 Pantheon 2.0: ............................................ 53
   4.1 Description of a Wikipedia article ............... 54
   4.2 Collecting articles about a single person ........ 56
   4.3 Birth and death year ............................ 59
   4.4 Birth place ...................................... 61
   4.5 Occupation ..................................... 64
   4.6 Summary ......................................... 66

5 When communicative and cultural memory merge: ............... 69

A Aggregation of Pantheon 1.0: ................................... 73
List of Figures

2-1 Changes in communication technologies. This figure shows the adoption of the communication technologies discussed in the main text. (a) Number of printers per million people in the world [52]. (b) Number of journals, founded between 1600 and 1800, per million people in the world [53]. (c) Fraction of countries from the HCCTA dataset [7] that adopted radio (solid-red) and television (dashed-blue). 28

3-1 The dynamics of collective memory. The grey area indicates the period in which printing was expanding. (a) Per-capita births of globally memorable people (m) measured using a 15-years 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 (Figure 2-1a). (b) Same as (a), but for the Human Accomplishment dataset. 35

3-2 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 (a), 10 (b), 15 (c), 20 (d), and 25 (e) years bins. The vertical dashed lines signal the position of the changepoints, and the horizontal solid lines correspond to the each period’s mean. The vertical axis corresponds to the births per year per billion people (m). 38

3-3 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 (a), 10 (b), 15 (c), 20 (d), and 25 (e) years bins. The vertical dashed lines signal the position of the changepoints, and the horizontal solid lines correspond to the each period’s mean. The vertical axis corresponds to the births per year per billion people (m). 39

3-4 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: 30 (a), 35 (b), and 37 (c). Horizontal lines are the result of the changepoint analysis. Horizontal axis represents the year, and vertical axis the births per year per billion people (m). 41
3-5 Changes in composition. Occupations of the biographies in Pantheon 1.0 for the period between (a) 500 and 1990, and (b) between 1800 and 1990.

3-6 Cross section of Figure 3-5 for (c) 500-1450 (scribal period), (d) 1450-1900 (printing), (e) 1900-1950 (film and radio), and (f) 1950-1990 (television).

3-7 Yearly probability that a memorable character in the Pantheon 1.0 belongs to each category. The diffuse contour corresponds to the 95% confidence interval.

3-8 Chi-squares statistic comparing the composition before and after each 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.

3-9 Probability that a memorable character is a Performing Artist (a), and a Sports Player (b). This figure is a closer look at Figures 3-7g and 3-7h. The diffuse contour corresponds to the 95% confidence interval.

4-1 Example of the Wikipedia page for Albert Einstein [56], one of the characters in Pantheon. Each page has an infobox, and extract, and a talk page.

4-2 Example of the Wikipedia page for New York City [56], an article within the scope of the WikiProject Cities.
List of Tables

3.1 Results of the changepoint analysis for different time windows using the Pantheon 1.0 dataset. The third and fourth columns show the years where each of the two breaks fall, and the last two columns show the mean associated to each period in births per year per billion people in the world.  

3.2 Results of the changepoint analysis for different time windows using the Human Accomplishment dataset. The third and fourth columns show the years where each of the two breaks fall, and the last two columns show the mean associated to each period in births per year per billion people in the world.  

3.3 Results of the changepoint analysis for different values of $L$, where $L$ is the minimum number of Wikipedia language editions a character needs to be featured in the Pantheon 1.0 dataset. $N$ is the number of characters used for the analysis. The second and third columns show the years where each break occurs.  

3.4 Sample bounds to determine the composition of each technological era.  

3.5 Results of the Pearson's chi-squared test comparing the occupations of characters from each technological period.  

3.6 Results of the Pearson's chi-squared test comparing the occupations of characters from each technological period for $L = 37$.  

3.7 Table showing the location of the break in composition associated with the invention of the movable type press using different thresholds for the minimum number of languages.  

A.1 Description of Pantheon 1.0 categories and the aggregation used in our analysis. $N$ is the number of people in each occupation.
Chapter 1

Introduction

"The spoken word was the first technology by which men was able to let go of his environment in order to grasp it in a new way."

Marshall McLuhan

In 1824, when Ludwig van Beethoven created his Symphony No. 9 in D minor he distributed it in the only way he could at the time; he wrote it down. Later on, when The Beatles created their "Yellow Submarine" they distributed it in the obvious way at the time; a vinyl disc. We rarely think about the constraints that the medium poses to the production of content perhaps because those constraints are so strong that they are difficult to see. For example, Yellow Submarine lasts 2 minutes and 48 seconds because the media standard for music recording in the 1960s was the 45 rpm 7 inch vinyl, which could record a maximum of three minutes [29]. The constraints imposed by the medium imply that whenever we experience a change in media (e.g. from written music to recorded music) we also experience a change in the content we can produce. Some changes are small, such as changing the length of songs, but others are more drastic, such as changing from compositions to performances. New media stimulates the production of new kinds of content because it changes the
constraints imposed over such content. Nowadays songs can last anywhere from 1.3 seconds (You Suffer, by Napalm Death) to 69 minutes (The Devil Glitch, by Chris Buttler). Moreover, the shift in the content we produce also implies that what we record will also change. We remember Beethoven as a great composer and Elvis Presley as a great singer because of the type of media that they used to record their creations. Therefore, whenever there is a change in communication media, we should see a change in the types of content we record.

But not only content is affected by media. Communication technologies also condition the amount of content we produce by making it easier or harder to encode information [10]. For example, due to the low cost of producing digital content, the new digital technologies have stimulated an unprecedented explosion in the amount of available information. Whenever it becomes easier to store or to spread information, whenever there is a change in media, we expect to see an increase in the amount of content we are able to remember. Hence, our collective memory—our records of past events, objects, ideas, and individuals—should co-evolve with our communication technologies.

Of course, this idea is not new. Scholars such as Harold Innis, Elizabeth Eisenstein, and Marshall McLuhan have already discussed the different societal effects of new communication technologies [11, 12, 24, 30, 36, 38]. In Innis’ view, communication media shapes empires and their power dynamics, by enabling new monopolies of knowledge [24]. For McLuhan, the medium is the message [30] since technology shapes our understanding of the world in ways more profound than the messages it carries. Finally, for Eisenstein the printing press acted as an agent of change in early modern Europe [11] by increasing the quantity and type of available information, as well as its durability and reliability—a revolution that helped catalyze new scientific knowledge [12].

The precise connection between theories of media and our collective memory, however, has often been overlooked. In this work we start by analyzing what the theories
of Innis, McLuhan, and Eisenstein tell us about how communication technologies affect our collective memory. According to their theories, changes in communication technologies should be accompanied by sharp changes in the size and composition of human collective memory—what we remember and how much we remember.

We empirically test these hypotheses by using two large datasets on biographical records—the Human Accomplishment dataset [34] and the Pantheon 1.0 dataset [58]—summarizing more than three thousand years of human history, and more than eleven thousand historical characters. We find that the per-capita birth rate of globally memorable characters from both datasets changes following the introduction of the movable type printing press and the maturity of shorter forms of printed media, such as newspapers and journals. We also find that the nature and mix of occupations of the memorable characters in the Pantheon 1.0 dataset change abruptly following the introduction of the printing press, the radio and cinema, and the television. These facts serve as empirical evidence in support of the theories of media introduced by Innis, McLuhan, and Eisenstein.

Our results, however, are limited to recent changes in communication media—the printing press is a recent technology when viewed in the context of human history. Earlier changes in media such as the introduction of papyrus or the alphabet spread much more slowly and locally compared to technologies like printing or television. Therefore, ancient changes in communication media should be accompanied by civilization-specific breaks in collective memory. For example, even though the alphabet existed before the rise of Ancient Greece, it only affected Greek's collective memory during the time of Socrates [38]. These “civilization-specific” breaks cannot be captured by any of our datasets due to both the lack of information about each character's civilization of birth, and the scarcity of our data for earlier years. To start tackling these issues we use the Pantheon 1.0 dataset as a starting point to build a more comprehensive dataset including more than forty thousand characters, connect-
ing each of them to their place of birth using historically accurate political borders instead of modern day borders. We hope this new revision will enable researchers to study a more fine-grained picture of human collective memory across all time periods.
Chapter 2

Why Collective Memory?

In this chapter we will discuss the concept of collective memory and build a rationale about its connection with communication technologies. We will start by reviewing the origin of human collective memory and how communication technologies rose as an evolutionary advantage. Finally, we will discuss why records of historical characters provide a good way to measure human collective memory.

2.1 What is collective memory?

"We shape our tools and thereafter our tools shape us."

Marshall McLuhan

We live in a world full of complex technologies that we take for granted. For example, even though I do not know how to synthesize sodium chloride, when I brush my teeth every morning I access the knowledge about the effect that sodium chloride has on human beings, how to synthesize it, and how to distribute it [22]. Toothpaste is just one of many technologies that make up the human toolkit. Most of these
technologies have been passed from generation to generation by a social system that figured out ways to remember collectively. What are the mechanisms by which our society creates its collective memory? To identify the features that enable modern humans to remember collectively, let us start by looking at two examples of early civilizations.

Almost ten thousand years ago, the Tasmanians’ toolkit included many complex bone tools. Archeological records show that after the island of Tasmania was disconnected from Australia, the tribes living there lost the ability to make many of the more complex technologies invented centuries earlier, bone tools being among them [20]. The anthropologist Joseph Henrich argues that this technology loss is related to how early Tasmanian society recorded information. The knowledge about the process required to make a particular tool could only be recorded by embedding it in a person, which made this particular kind of memory very fragile. When a tool-making process is simple, the death of those who know it is not a big problem since everyone can learn the process and pass it on to successive generations. But when the process is complex, the number of people who will learn it is inevitably smaller. Therefore, when there is a considerable decrease in the population of a society—as when Tasmania was separated from Australia—the knowledge and knowhow needed to manufacture complex technologies can be lost. Put in simple terms, everyone who knew how to effectively make bone tools died before they could teach the craft to enough people to guarantee preserving that knowledge.

The example of Tasmania teaches us that our ability to accumulate information is a centerpiece of our knowledge-generating enterprise. Forgetting is expensive, and because of this, our species has evolved many cultural adaptations that enable us to remember as what might be called a collective brain. The origins of some adaptations are relatively evident, such as the invention of writing, but others are more obscure, such as the creation of public gathering spaces or the practice of yearly rituals such as
the ones practiced by the hunter-gatherers in western Australia, our second example.

In 1943 a group of hunter-gatherers in the western desert of Australia faced a severe and enduring drought. An old man named Paralji saved the tribe by leading them through 350 kilometers of desert to the Mandora Station, a place that Paralji did not even know existed. How did he do it? Paralji was able to access the knowledge accumulated in the form of song cycles that his people sang during rituals. By telling a story about a group of his ancestors wandering through a sequence of places, the song worked very much like a map pointing to the location of the Mandora Station. The group was saved because they evolved a cultural adaptation to collectively remember a map. The songs of Paralji’s tribe are a good example of a cultural adaptation that enables information to be passed from one generation to the other. They abstract the particular knowledge a single person might have about the location of a place into a shared body of knowledge.

To explain these type of processes, the Egyptologists Aleida and Jan Assman [1] introduced the distinction between the short-term memory preserved by a small group of people through repeated acts of communication—communicative memory—and the long-term memory that forms part of a group’s culture—cultural memory. Communicative memory is shared within a social group, is defined by personal memories, and has a lifespan of 80 to 100 years—the lifespan of a generation. Family stories, jokes, and personal stories are all part of our communicative memory. Communicative memory is unstructured due to the fact that everyone has his or her own version of the recorded idea. Only a small number of ideas are able to transcend the boundaries of one generation and become part of our cultural memory.

Cultural memory is the large body of reusable texts, images, rituals, songs, ideas, social norms, etc. that societies accumulate across generations. It has a more exclusive character than communicative memory, since not everyone is endowed with the legitimacy to influence the content of cultural memory. According to Aleida and
Jan Assman cultural memory is intrinsically related to power and tradition, hence it
covers a much longer period of time compared to communicative memory. Rituals,
songs, folk stories, and cave paintings enable people to transfer the short-lived pieces
of information stored in our communicative memory into our cultural memory. In
short, cultural memory begins where communicative memory ends, and also some of
the latter is captured in the former.

The distinction between communicative and cultural memory was originally intro-
duced to organize the divergent definitions of the term collective memory [1, 19, 49].
The sociologist Maurice Halbwachs first used the term collective memory to refer
to 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. Later on, the French historian Pierre Nora focused on separating
memory from history by exploring what he called sites of memory—which are purely
material “sites”, such as events, places, concepts, or objects, invested with a sym-
bulic aura, such as the Eiffel Tower, the Statue of Liberty, or the Boston Marathon
bombing [35]—and studying their role in the construction of a group’s identity—in
Nora’s case, the construction of French identity [16]. Nora triggered a whole body
of literature that focused on the relation between how a group of people creates a
shared version of the past and how it perceives itself [43, 49]. This line of literature
diverges from our definition of collective memory in that it seeks to distinguish the
subjective reconstruction of the past (memory) from the objective reconstruction of
the past (history).

The following work focuses on the cultural memory we have accumulated as a human
species. In particular, we focus on the role that communication technologies had in
the creation of this vast corpus of preserved information.
2.2 Understanding the role of media

"Written text stands there as it were alive, but if anyone asks them anything, they remain most solemnly silent."

Socrates

The idea that communication technologies shape our collective memory is not new. In fact, the Egyptians were already concerned about the effect that writing could have on their ability to remember. The story goes that Thoth, the Egyptian god credited for inventing writing, presented his creation to the king as a tool for making Egyptians wiser and improving their memory. The king, however, thought that his invention would have the opposite effect because once something is recorded in this external way, there was no need to remember it. Today, historians largely agree about the importance of the invention of writing, so much so that history differentiates itself from other academic disciplines, such as archeology and anthropology, by the invention of writing. Yet, as the historian Elizabeth Eisenstein argued [11, 12], while historians agree that writing defines the emergence of history, there is no widespread consensus about the relative importance that more recent technological changes had on our species’ ability to remember.

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 [2, 11, 12, 24, 28, 30, 33, 37]. These studies include the work of the economic historian Harold Innis [24], the philosopher of communication Marshall McLuhan [30], and the historian of printing Elizabeth Eisenstein [11, 12], among others [2, 28, 33, 37].
Harold Innis is considered by some [36] to be the first scholar who attempted to establish the history of communications as a distinct academic field. Innis was more interested in the effect of communication technologies in the rise and fall of civilizations, rather than its influence on our collective memory. For example, he saw the invention of writing as a way of building a centralized control system such as the one used by the Romans [24]. Even though Innis’ focus was not memory, his powerful theories point towards a clear relation between media and memory.

Innis points out that the nature and uses of a medium will determine how knowledge can be communicated and recorded, and thereby shape how each civilization monopolized knowledge. If the medium is light and easily transported, like hieroglyphs drawn on papyrus, it is better suited to spread information across space. On the other hand, if the medium is heavy and durable, like monuments, tombs, and stone calendars, it is better suited to disseminate knowledge over time. Unfortunately, Innis passed away before finishing his “History of Communications,” which was poised to become his defining work on this topic [36].

Innis’s ideas were expanded and popularized by one of his younger colleagues, Marshall McLuhan, who made famous the study of the social impact of communication technologies with his 1964 book: “Understanding Media: The Extensions of Man.” Understanding Media opened with the now famous phrase: “the medium is the message” [30], which meant that changes in communication technologies are more consequential to society than the messages transmitted through them. McLuhan emphasized how changes in media changed the type of messages that can be communicated (e.g. 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, according to McLuhan, we expect 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's claims and ideas by documenting the impact of printing in early modern Europe [11, 12]. Eisenstein showed that prior to the introduction of printing, our species' ability to store information was volatile. Printing changed the durability and reliability of the data available to scholars (such as the astronomical tables used by Copernicus and Kepler [12]), and also, helped catalyze new scientific knowledge by dramatically increasing the availability and reliability of the data and literature available to renaissance scholars.

Although these ideas have been discussed by many scholars, statistically testing them 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 [32, 34, 44, 58] that might be otherwise buried in narrative descriptions [46, 47]. Examples of these studies include the evolution of language and ideas as recorded in printed books [32], patterns of historical migration [42], the importance of language translations in the global diffusion of information [41], the emotional content of global languages [9], and the dynamics of fame [44, 48, 58]. These studies are made possible thanks to new datasets that leverage digital sources, such as Wikipedia, Freebase, and digitized books [25].

Here, 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, respectively, 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 changes in communication technologies. Furthermore, the Pantheon 1.0 dataset exhibits sharp changes in the occupations represented in biographical records that coincide with changes in communication technologies.
In this work, I focus on four major communication revolutions: the printing press, the public sphere, radio and cinema, and television.

The printing press was a technology that spread very quickly over Europe, making it an invention with relatively discrete historical effects. Before the invention of printing, the number of manuscript books in Europe could be counted in the thousands. By 1500, after only 50 years of printing, there were more than 9,000,000 books [3]. Figure 2-1a shows the number of printers per million people in the world, until 1530 [52].

Elizabeth Eisenstein discussed extensively the effect of the printing press in her 1979 book “The Printing Press as an Agent of Change,” and her later book “The Printing Revolution in Early Modern Europe.” In her books, Eisenstein showed that printing not only changed the number of books printed during the Renaissance, but also who the authors of these books were and what these books were about. The renewed availability of printed texts had the effect of turning reading, which had been an oral and social practice, into a solitary practice that would allow people to reflect and rethink the content they were consuming. Probably the strongest impact of printing, however, would come later with the development of new forms of printed media that co-evolved with what sociologists call the public sphere.

The public sphere was a space for public discourse described by the sociologist Jürgen Habermas [18]. Habermas argued that the public sphere emerged during the early 1700s, after the arrival of coffee in Western Europe due to the establishment of coffeehouses in London, Paris, Vienna, and Prague, and with the introduction of new publication formats, such as newspapers, journals, and pamphlets. These periodical publications afforded easy sharing of news, opinion, and other content about recent current events, fueling conversations among coffeehouse patrons that books (with their much longer production cycles) could not have enabled. Coffeehouses thus acted as centers of social interaction that afforded for the first time the widespread public use of reason in rational-critical debate. By the dawn of the eighteenth century, contem-
poraries counted over 3,000 coffeehouses in London alone [17], each one a node of periodical-inspired topical conversations. According to Habermas, this combination of new publication formats and new spaces for public discourse gave rise to the public sphere.

The public sphere was not itself a new communication technology, but rather a socio-cultural innovation that relied on new technological innovations. It had such a big impact, however, on the way people communicate, that it seems unfair not to consider it as one of the greatest communication revolutions of our recent history. Its impact was the reinvention of public discourse [18]. The public sphere did not evolve due to a single piece of technology, but rather due to a combination of many technologies. By the early 1700s the cost of gathering and distributing information within a city was high, preventing the newspaper and other forms of daily printed formats from spreading widely [10]—in 1700 neither London nor Paris had a daily newspaper. Over the following decades new technologies, such as turnpikes built using new construction techniques, changed the transmission costs by cutting the time of intercity travel [14]. By 1760, London had four daily and six tri-weekly newspapers [3]. The coffeehouses, on the other hand, played an important role by increasing accessibility to these new publication formats, and the relative freedom of the press at the time enabled the spread of opinions, rather than just facts. During this period arose an intense competition between journals, which peaked around 1720, and eventually settled down, suggesting a maturation of the technology. Figure 2-1b shows the number of functioning journals, established between 1600 and 1800, per million people in the world.

Since the first half of the twentieth century, radio and television have also spread relatively fast, at least in the industrialized world. Figure 2-1c shows the fraction of the countries in the Historical Cross-Country Technology Adoption (HCCTA) Dataset [7] that have radio (solid-red) and television (dashed-blue).
Figure 2-1: Changes in communication technologies. This figure shows the adoption of the communication technologies discussed in the main text. (a) Number of printers per million people in the world [52]. (b) Number of journals, founded between 1600 and 1800, per million people in the world [53]. (c) Fraction of countries from the HCCTA dataset [7] that adopted radio (solid-red) and television (dashed-blue).
2.3 Measuring cultural memory

The ideal dataset for quantifying our cultural memory should include data summarizing information produced by people from all origins and including all forms of historical information—biographical data, fictional works and characters, other artifacts and constructions that people created, social norms, etc. Since no such dataset exists, we use a simpler picture of our cultural memory that arises when we focus only on biographical information. Historical characters are both easy to structure—they can be counted, classified according to their occupation, and each can be assigned to a particular birthplace and time—and can provide tracking points to study social change [45].

We use biographical data on memorable characters from two sources: The Pantheon 1.0 dataset [58] and the Human Accomplishments dataset [34]. The Pantheon 1.0 dataset contains the 11,337 biographies that had a presence in more than 251 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 with a place of birth, a date of birth, and an occupation using a three-level hierarchical classification that disaggregates biographies into 88 distinct occupations—e.g. Physics and Biology are branches of Natural Science, just as Natural Science is a branch of Science.

The Human Accomplishments dataset [34] contains 3,869 biographies of accomplished\(^2\) 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—

---

\(^1\)The choice of the \(L > 25\) threshold is guided by a combination of criteria. For more details see [58].

\(^2\)For the definition of accomplishment and other details see original source [34].
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.

The use of historical characters as a proxy for 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 [5], and Herbert Spencer’s theory that, in opposition to Carlyle, sees prominent individuals as a reflection of their times [45]. In Spencer’s view, historical events beget historical characters, and the fame of individuals becomes an artifact of how we record history. In 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.

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 in using Wikipedia as a primary data source [58], therefore our results should be interpreted as statements about the picture of collective memory that is representative only of the people who edit the more than 250 language editions of Wikipedia—a literate and digitally empowered elite of knowledge enthusiasts [51]. For a detailed description of the biases and limitations of the Pantheon 1.0 dataset see [58]. The Human Accomplishments dataset is compiled from literary expressions in six different languages, and is representative of the elite cadres of people who participated in the creation of printed encyclopedias. An extended discussion of the biases and validation of this dataset can be found in Human Accomplishments [34]. 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. All historical research is ultimately a reflection of the sources used in its creation.

Because some occupations in the Pantheon 1.0 dataset include very few characters, we use the second level of aggregation provided by the dataset. Pantheon classifies characters into 8 domains, 27 industries, and 88 occupations. For our analysis, we aggregate occupations into categories distinguishing mainly between politicians and religious figures, and between arts and performing arts.

The arts domain is split into performing arts (including dance, and film and theater industries, plus all occupations from the music industry except for composer) and arts (including 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. The science and technology, and humanities industries remain unchanged. Finally, individual sports, along with the domains business and law, exploration, and public figure are grouped together as other. Table A.1, from the appendix, shows a summary of the aforementioned aggregation.

The three largest occupations aggregated under other—tennis player, social activist, and racecar driver—are very small—with 161, 114, and 104 individuals respectively. Therefore, any change in the category other will also captured by the 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, a change already captured by the sports category.
Chapter 3

Breaks in size and composition of collective memory

This chapter constitutes the main empirical result of this work. First, we show that the number of yearly births of globally memorable individuals recorded in each dataset, normalized by the world population, is constant for the two thousand years preceding the invention of the movable type press and jumps to a new rate soon after the adoption of this new technology. The per-capita birth rate remains constant until the birth of the public sphere, where it transitions to a period of unbounded growth. Second we show that the distribution of occupations of the characters recorded in the Pantheon 1.0 dataset exhibits a sharp change coinciding with the introduction of the movable type press, and that the introduction of radio and cinema, and television, were accompanied by the rise of occupations that were not present before. Together, these two facts point towards a picture of our collective memory that is deeply affected by the advent of new communication technologies.
3.1 Breaks in size

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, 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. We use population data from the historical world population estimates of the US Census Bureau [4], which reports an aggregated dataset of world population estimates starting from the year 10,000 BC. The missing years are interpolated using linear splines. This quantity provides us with an estimate of the per-capita number of births occurring in a given year that we still remember today \( m = \frac{M}{N} \). \( m \) is measured in units of births per year per billion people in the world, or [bpyb].

Figure 3-1 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, using a 15-years time window. For both datasets \( m \) was constant for the 2000 years preceding the introduction of the movable 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 remembered today.
Figure 3-1: The dynamics of collective memory. The grey area indicates the period in which printing was expanding. (a) Per-capita births of globally memorable people (m) measured using a 15-years 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 (Figure 2-1a). (b) Same as (a), but for the Human Accomplishment dataset.

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 [27]. 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 statistic that depends on the number and position of the changepoints. The analysis detects two major changepoints in both time series. 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.

All changepoint analyses were performed using the *changepoint* package available for R, developed by Rebecca Killick and Idris Eckley from Lancaster University [27]. In particular, we used the `cpt.mean` function to determine changes in mean. We used “CUSUM” as a test statistic, 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 this method see reference [27].

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 60 years of age by the time the printing press was being introduced). The first changepoint occurs right before the introduction of the movable type press in 1450, and it seems to lead to a new stationary state that lasts roughly three centuries. As discussed before, printing spread and plateaued quickly [3], making the introduction of printing a relatively discrete event in history (Figure 3-1a), 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 shorter forms of printed media that we described in the previous sections. The time series reveals that starting in the middle of the eighteenth century, the per-capita number of births of globally memorable individuals begins an era of continuous growth. This means 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).

In what follows we perform a series of robustness checks of our result to prove that it is independent of the time window used and of the particular definition of what a memorable character is. We start with repeating the changepoint analysis for
Table 3.1: Results of the changepoint analysis for different time windows using the Pantheon 1.0 dataset. The third and fourth columns show the years where each of the two breaks fall, and the last two columns show the mean associated to each period in births per year per billion people in the world.

<table>
<thead>
<tr>
<th>Window size [years]</th>
<th>Number of points</th>
<th>1st break</th>
<th>2nd break</th>
<th>1st mean [bpyb]</th>
<th>2nd mean [bpyb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>490</td>
<td>1445-1450</td>
<td>-</td>
<td>3.530</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>245</td>
<td>1420-1430</td>
<td>1760-1770</td>
<td>3.515</td>
<td>5.625</td>
</tr>
<tr>
<td>15</td>
<td>163</td>
<td>1375-1390</td>
<td>1750-1765</td>
<td>3.514</td>
<td>5.336</td>
</tr>
<tr>
<td>20</td>
<td>122</td>
<td>1420-1440</td>
<td>1720-1740</td>
<td>3.520</td>
<td>5.311</td>
</tr>
<tr>
<td>25</td>
<td>98</td>
<td>1400-1425</td>
<td>1700-1725</td>
<td>3.515</td>
<td>5.147</td>
</tr>
</tbody>
</table>

Table 3.2: Results of the changepoint analysis for different time windows using the Human Accomplishment dataset. The third and fourth columns show the years where each of the two breaks fall, and the last two columns show the mean associated to each period in births per year per billion people in the world.

<table>
<thead>
<tr>
<th>Window size [years]</th>
<th>Number of points</th>
<th>1st break</th>
<th>2nd break</th>
<th>1st mean [bpyb]</th>
<th>2nd mean [bpyb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>183</td>
<td>1421-1428</td>
<td>1723-1728</td>
<td>1.596</td>
<td>6.775</td>
</tr>
<tr>
<td>10</td>
<td>132</td>
<td>1386-1396</td>
<td>1716-1726</td>
<td>1.436</td>
<td>6.424</td>
</tr>
<tr>
<td>15</td>
<td>106</td>
<td>1379-1394</td>
<td>1709-1724</td>
<td>1.341</td>
<td>6.395</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>1375-1395</td>
<td>1695-1715</td>
<td>1.266</td>
<td>6.380</td>
</tr>
<tr>
<td>25</td>
<td>78</td>
<td>1370-1395</td>
<td>1720-1745</td>
<td>1.219</td>
<td>6.573</td>
</tr>
</tbody>
</table>

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 millennium, we use variable size time windows built by setting 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, with at least 5 births. If we do not include this lower bound, only 68% of all time windows have non-zero people when using 5-years bins, making the changepoints harder to detect. We choose this method over a smoothing—using a moving average for example—because our goal is to detect breaks, and the smoothing might smooth out the breaks. We note that most of the time windows after year 1000 do not have this problem. Furthermore, Pantheon 1.0 does not have this problem.
Figure 3-2: 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 (a), 10 (b), 15 (c), 20 (d), and 25 (e) years bins. The vertical dashed lines signal the position of the changepoints, and the horizontal solid lines correspond to the each period’s mean. The vertical axis corresponds to the births per year per billion people (m).
Figure 3-3: 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 (a), 10 (b), 15 (c), 20 (d), and 25 (e) years bins. The vertical dashed lines signal the position of the changepoints, and the horizontal solid lines correspond to the each period’s mean. The vertical axis corresponds to the births per year per billion people (m).
Table 3.3: Results of the changepoint analysis for different values of $L$, where $L$ is the minimum number of Wikipedia language editions a character needs to be featured in the Pantheon 1.0 dataset. $N$ is the number of characters used for the analysis. The second and third columns show the years where each break occurs.

<table>
<thead>
<tr>
<th>L</th>
<th>N</th>
<th>1st break</th>
<th>2nd break</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>8238</td>
<td>1433</td>
<td>1763</td>
</tr>
<tr>
<td>35</td>
<td>5841</td>
<td>1427</td>
<td>1727</td>
</tr>
<tr>
<td>37</td>
<td>5112</td>
<td>1432</td>
<td>1792</td>
</tr>
</tbody>
</table>

Figures 3-2 and 3-3 show that the position of the first and second breaks are independent of the time window. The second break could not be detected when using the 5-years window in the Pantheon 1.0 dataset because increase is too smooth at this resolution. Tables 3.1 and 3.2 summarize the results of the changepoint analysis for the different binning size for both the Pantheon 1.0 and the Human Accomplishment datasets.

Now we move to check the robustness of our results in relation to the different definition of memorability. We first note that our results are independently validated by two datasets, collected with two different definitions of “memorability”—the Pantheon 1.0 dataset focuses on prominent characters and the Human Accomplishment dataset focuses on accomplished characters. The agreement of both datasets implies that the results are independent of the particular definition of memorability. In what follows we will check our results to varying the threshold $L$ of characters in Pantheon 1.0. $L$ is the minimum number of Wikipedia language editions a character needs to appear in within the Pantheon 1.0 dataset [58]. We use $L = 30, 35, \text{and } 37$. Higher values of $L$ are difficult to test, because the sample size decreases significantly and is biased towards the latter years—less than 30% of the characters have $L \geq 45$ and are mostly concentrated after 1500. Lower values of $L$ cannot be tested due to a limitation in the dataset [58]. The results presented in Table 3.3 and in Figure 3-4 show that the two previously discussed breaks are still present for $L = 30, 35, \text{and } 37$, suggesting that our results are independent of the chosen value of $L$.  

40
Figure 3-4: 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: 30 (a), 35 (b), and 37 (c). Horizontal lines are the result of the changepoint analysis. Horizontal axis represents the year, and vertical axis the births per year per billion people (m).

3.2 Changes in composition

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 Pantheon 1.0.

Figures 3-5a and 3-5b show the fraction of biographies corresponding to individuals associated with each cultural domain. Figures 3-6a to 3-6d show the cross-sections for
different technological periods: scribal culture (< 1450), printing (1450-1900), film and radio (1900-1950), and television (1950-2000).

The stacked area chart from Figures 3-5a and 3-5b 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 characters in each category born in each time window.

3. Apply a moving average to smooth the noise. Frequencies are normalized after the smoothing. Figure 3-5a and Figure 3-5b use an 11-year and a 5-year moving average, respectively.

Figure 3-5: Changes in composition. Occupations of the biographies in Pantheon 1.0 for the period between (a) 500 and 1990, and (b) between 1800 and 1990.
Figure 3-6: Cross section of Figure 3-5 for (c) 500-1450 (scribal period), (d) 1450-1900 (printing), (e) 1900-1950 (film and radio), and (f) 1950-1990 (television).

Figure 3-5 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 3-6a and 3-6b).

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 (Figures 3-6b and 3-6c). Finally, Figures 3-6c and 3-6d show that sports players—such as soccer players, basketball players, and race-car drivers—became memorable with the adoption of television. We note that these obser-
Observations suggest a direction of causality, since actors, singers, and musicians have been part of society for centuries (both Ancient Greek and Elizabethan English playwrights, such as 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 note that observations for most recent years need to be interpreted carefully for two reasons. First, the more recent biographies in our dataset contain a mix of characters that are memorable (e.g. 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 not last for centuries after a person’s death. Second, we note that data for the most recent years is also affected by differences in the life cycle of an individual’s memorability, since individuals with different careers peak at different ages. Soccer players, for example, peak around their late twenties or early thirties [6], 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 [54], and hence, we may be missing some influential individuals who 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.
The fraction of memorable characters belonging to each category can be thought of as the probability that a memorable character born in a given year belongs to a each category. Figure 3-7 shows the estimator of this probability for each year and for each category—calculated the same was as in Figures 3-5a and 3-5b.

In what follows, we will test the statistical significance and robustness of our claims. Due to the scarcity of the data in older eras, we restrict our analysis of the occupations of memorable characters to the time window between 500 and 1990. We test the statistical significance of the changes in composition by means of a Pearson’s chi-squared test [55]. 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 statistic has a chi-squared distribution.

To sample each technological period we use the bounds shown in Table 3.4. These bounds were chosen to reduce the transition effects from one technology to the other. For example, since the movable type press was invented in 1450, its effects should start appearing in our dataset during the early 1400s because our dataset is based on birth years. Therefore, we choose the upper bound for the “writing era” well before 1400, and the lower bound for the “printing era” at 1450. The p-values resulting from comparing each technological period are shown on Table 3.5.

As before, we check the robustness of our claims relative to the change in the definition of memorability by changing the minimum number of Wikipedia language editions each character needs in order to be featured in Pantheon 1.0. We use $L = 30, 35,$ and $37$, and find that the results are consistent (p-values for $L = 37$ are shown in Table 3.6).
Figure 3-7: Yearly probability that a memorable character in the Pantheon 1.0 belongs to each category. The diffuse contour corresponds to the 95% confidence interval.
Table 3.4: Sample bounds to determine the composition of each technological era.

<table>
<thead>
<tr>
<th></th>
<th>Sample limits</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>500-1300</td>
<td>540</td>
</tr>
<tr>
<td>Printing</td>
<td>1450-1850</td>
<td>1350</td>
</tr>
<tr>
<td>Radio and cinema</td>
<td>1900-1930</td>
<td>1123</td>
</tr>
<tr>
<td>Television</td>
<td>1950-1990</td>
<td>3041</td>
</tr>
</tbody>
</table>

Table 3.5: Results of the Pearson's chi-squared test comparing the occupations of characters from each technological period.

<table>
<thead>
<tr>
<th></th>
<th>Printing</th>
<th>Radio and cinema</th>
<th>Television</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>1.3e-93</td>
<td>4.6e-151</td>
<td>&lt;1.3e-311</td>
</tr>
<tr>
<td>Printing</td>
<td>-</td>
<td>3.1e-92</td>
<td>&lt;1.3e-311</td>
</tr>
<tr>
<td>Radio and cinema</td>
<td>-</td>
<td>&lt;1.7e-287</td>
<td>&lt;1.7e-287</td>
</tr>
</tbody>
</table>

Table 3.6: Results of the Pearson's chi-squared test comparing the occupations of characters from each technological period for \( L = 37 \).

<table>
<thead>
<tr>
<th></th>
<th>Printing</th>
<th>Radio and cinema</th>
<th>Television</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing</td>
<td>5.0e-71</td>
<td>1.7e-95</td>
<td>6.6e-215</td>
</tr>
<tr>
<td>Printing</td>
<td>-</td>
<td>3.6e-53</td>
<td>5.9e-285</td>
</tr>
<tr>
<td>Radio and cinema</td>
<td>-</td>
<td>-</td>
<td>9.9e-153</td>
</tr>
</tbody>
</table>

From Figure 3-7 we see that the category *performing arts* and the category *sports* were hardly present before 1900 and 1960, respectively. The printing break, however, is not as clear-cut. In what follows we show that between 500 and 1850 there is no other significant discrete break, besides the one related to the invention of the movable 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.
Table 3.7: Table showing the location of the break in composition associated with the invention of the movable type press using different thresholds for the minimum number of languages.

<table>
<thead>
<tr>
<th>$L$</th>
<th>Mean year</th>
<th>Min year</th>
<th>Max year</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>1363</td>
<td>1348</td>
<td>1377</td>
</tr>
<tr>
<td>35</td>
<td>1364</td>
<td>1344</td>
<td>1394</td>
</tr>
<tr>
<td>37</td>
<td>1376</td>
<td>1354</td>
<td>1395</td>
</tr>
</tbody>
</table>

3. Calculate the chi-squared statistic, and reject with a p-value smaller than 0.005.

4. For every group of years for which the hypothesis has been rejected, report the average year.

Figure 3-8a shows the value of the test statistic as a function of the year. This analysis yields only one group of years for which we reject the hypothesis. Such group of years has an average of 1368 and a range between 1359 and 1377. This means 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.

To show that this break is not dependent on the particular definition of memorability, we repeat the analysis with a different threshold in number of languages $L$. Figure 3-8b shows the evolution of the chi-squared statistic for $L = 30, 35, $ and 37. Table 3.7 summarizes the position of the breaks. The curve for lower years is missing some values because of the scarcity of the data; it was not possible to build the time windows for those years according to the described procedure.

The printing break is different from the breaks for radio and cinema, as well as television, because it is a more generalized change. The breaks for radio and cinema on one hand, and television break, on the other hand, had the effect of bringing a new category to collective human memory. Since the time between these two inventions is very short for historical time scales—roughly 50 years—neither of them can be
Figure 3-8: Chi-squares statistic comparing the composition before and after each 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.

modeled as a discrete break because the system was not given enough time to stabilize. This also implies that detecting a change in the size of our collective memory due to the introduction of these new technologies is very difficult using this method.

Figure 3-9 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.

Next, we briefly discuss the limitations of the Human Accomplishment dataset to capture changes in the composition of our collective memory. The Human Accomplishment dataset classifies people into 5 different categories—science, art, philosophy,
Figure 3-9: Probability that a memorable character is a Performing Artist (a), and a Sports Player (b). This figure is a closer look at Figures 3-7g and 3-7h. The diffuse contour corresponds to the 95% confidence interval.

The Human Accomplishment dataset does not include people related to performing arts nor sports; therefore it is not possible to use it to track changes in the composition of our collective memory due to the introduction of film, radio, and television. Since the classification is not as rich as in Pantheon 1.0, the changes in composition are not as clear.

The Human Accomplishment dataset is an effort to track people who had an effect in society, even though they are not globally remembered. The large number of artists present in this dataset, born between 1250 and 1350—41% of all births—is 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, that promoted the arts and trade between east and west. Europe, on the other hand, was still struggling through the Middle Ages, and its 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 do not appear in the Pantheon dataset—but their contribution had an effect on modern society—hence they appear in 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 of very limited utility for tracking changes in the composition of our collective memory.
Chapter 4

Pantheon 2.0

“We drive into the future using only our rearview mirror.”
Marshall McLuhan

The advantage of the Pantheon 1.0 dataset, over other similar datasets, lies on its multilingual sources, and its classification of characters into occupations based on their major field of contribution [58]. The historical precision of this dataset, however, is limited due to its use of present day political borders rather than historical borders. Hence, we introduce the Pantheon 2.0 dataset, an update to Pantheon 1.0 that assigns historical characters to a place of birth based on historical political borders and on the most fined grained birth place available in the data sources used to compile the dataset—i.e. if data about the character’s birth city is present, the character is assigned to a birth city, if only data about the character’s birth country is present, the character is assigned to a birth country. Furthermore, by decreasing the minimum number of Wikipedia language editions that a character needs in order to be featured in the dataset, Pantheon 2.0 includes nearly four times as many characters as Pantheon 1.0. We hope this new dataset will enable future research on the effect on our collective memory of communication technologies that spread very
slowly across ancient societies, such as the alphabet, the papyrus, or road building.

The following chapter outlines the data collection process for the Pantheon 2.0 dataset. Such process follows an approach that mixes text parsing, natural language processing, and manual verification. We use Wikipedia [56] as our primary data source, and complement it with Freebase, Wikidata, and Pantheon 1.0 [15,50,58].

4.1 Description of a Wikipedia article

Since Wikipedia is our primary data source, we start by describing the anatomy of a Wikipedia article (hereafter “wikipage”). Figure 4-1 shows the wikipage for Albert Einstein as an example. The most relevant parts of a wikipage for our purpose are the infobox, the extract, and the talk page.

Figure 4-1: Example of the Wikipedia page for Albert Einstein [56], one of the characters in Pantheon. Each page has an infobox, and extract, and a talk page.
The infobox\(^1\) is a summary box for the character that includes fields such as birth date, birth place, death date, name, etc. Each infobox is built using a different template\(^2\) written in WikiCode\(^3\). The following is an example of the WikiCode for the infobox scientist\(^4\):

```wiki
{{ Infobox scientist
    | honorific_prefix = 
    | name = 
    | birth_date = 
    | birth_place = 
    | death_date = 
    | death_place = 
}}
```

There are a total of 147 different infobox templates that can be used for biographies, including scientist, writer, football biography, person, etc. Not all infoboxes contain the same data, and the type of infobox used for each wikipage is a decision of the wikipedians\(^5\), who can choose more than one infobox per wikipage. For example, the article about James FitzJames has the infoboxes nobility and military person.

The infobox data is not necessarily structured. For example, the field corresponding to a date of birth can be filled using some of the multiple WikiCode templates for dates—such as Birth date and age\(^6\)—or using any other date format—such as January, 1204.

---

or sometime during 1204.

The extract is the first part of the wikipage text, and consists on a brief description of the character. It typically includes the character’s origin, occupation, and main accomplishments. In some cases the extract contains more data than the infobox. For example, a character’s date of birth can be missing from the infobox, but be present in the extract.

There is no explicit format for the extract, but wikipedians have converged to writing extracts in a way that the first sentence is a statement about who the character is. For example, the first sentence of Albert Einstein’s wikipage reads: *Albert Einstein was a German-born theoretical physicist.* The first sentence also contains a parenthesis with other names of the character, the correct pronunciation for his or her name, and dates and places of birth and death.

The talk page, also known as discussion page, is a page which editors can use to discuss improvements to the wikipage. It also includes information about which WikiProject the page belongs to, and its relative quality level. This last part is important for us, since we will start our data collection process with all the Wikipedia pages that belong to the WikiProject Biography [57].

### 4.2 Collecting articles about a single person

We begin the data collection for Pantheon 2.0 with full list of 8,764,792 articles in the English Wikipedia taken from the Wikipedia dump from December 3rd, 2015. We select the articles that belong to the WikiProject Biography. The Biography

---


9Includes redirect and disambiguation pages.

10For all the English Wikipedia dumps visit [https://dumps.wikimedia.org/enwiki/](https://dumps.wikimedia.org/enwiki/).
WikiProject concerns the creation, development, and organization of Wikipedia's articles about real humans [57]. Articles that belong to this project are identified through the WikiProject Biography Template\textsuperscript{11} in their talk page. In its basic form, the template includes the fields \textit{living}, \textit{class}, and \textit{listas}:

\begin{verbatim}
{{ WikiProject Biography
 | living  = yes
 | class   =
 | listas  =
 }}
\end{verbatim}

The \textit{living} parameter indicates whether the character is alive or dead, \textit{class} is a statement about the relative quality of the article, and \textit{listas} is the key that Wikipedia uses to sort the article. This and all other WikiCode templates—infoboxes, date templates, WikiProject templates, etc.—are parsed using the package \texttt{mwparserfromhell}\textsuperscript{12}.

At the time of data collection we successfully identified 1,297,150 articles within the scope of the WikiProject Biography, with the Pantheon 1.0 dataset fully contained within this group\textsuperscript{13}.

We will restrict ourselves to wikipages about individual persons, therefore articles such as the \textit{Wright Brothers}\textsuperscript{14} will be excluded from our dataset. The articles within the scope of the WikiProject Biography, however, are not guaranteed to be about individual persons. When a substantial section of an article is a biography of a person, then an article about a organization or a group will be part if this WikiProject. Therefore, there is a considerable number of articles with biographical information that do not correspond to a single individual. For example, articles about music

\textsuperscript{12}\url{https://github.com/earwig/mwparserfromhell}.
\textsuperscript{13}There are two articles from Pantheon 1.0 that were deleted from the English Wikipedia. For more information see \url{https://en.wikipedia.org/wiki/Wikipedia:Articles_for_deletion}.
\textsuperscript{14}\url{https://en.wikipedia.org/wiki/Wright_Brothers}.
bands or duos, lists of monarchs, terrorist groups, etc. Since these articles lack some of the fields that a biography about a single person typically has (music bands do not have a gender, articles about siblings do not have a single birth year, etc.) they tend to be singled out by our data collection process. We will apply a set of filters to select our the bulk to these articles.

Next, we use the Wikipedia API to collect the number of language editions for each of the 1,297,150 articles related to a biography\textsuperscript{15}, for the full documentation of the Wikipedia API see [31]. For Pantheon 1.0 the authors chose a threshold of 25 language (for more information see [58]), here we choose $L = 15$ leaving us with 51,676 articles and nearly quadrupling the dataset's coverage.

So far, our data collection process has been centered on the English Wikipedia. To expand the reach of our dataset to multiple language editions, we complement the WikiProject Biography with Wikidata. We use the Wikidata RDF dump from December 2015\textsuperscript{16} to collect all the instances of\textsuperscript{17} the class human\textsuperscript{18}. We find a total of 520 characters that did not appear within the scop of the WikiProject Biography either because they do not have a english Wikipedia page\textsuperscript{19}, they do not belong to the WikiProject Biography\textsuperscript{20}, or they were in the process of being moved\textsuperscript{21}. This gives us a total of 52,196 articles.

To filter out the articles that do not correspond to a biography of a single person we use three methods:

\textsuperscript{17}Property P31 https://www.wikidata.org/wiki/Property:P31.
\textsuperscript{18}Element Q5 https://www.wikidata.org/wiki/Q5.
\textsuperscript{19}We only found 86 characters with more than 15 language editions and without an english Wikipedia edition. The French journalist Beate Klarsfeld is an example https://de.wikipedia.org/wiki/Beate_Klarsfeld.
\textsuperscript{21}The pharaoh Ramesses V for example https://en.wikipedia.org/wiki/Ramesses_V.
1. Heuristics regarding the title: We review all the articles with a title that starts with “list”, or that includes either “and” or “&”.

2. Exploit the structure of Wikipedia articles: We review all the articles that contain either the word “band”, “duo”, or “group” after the verb “to be” in the first sentence, as well as all the articles where the verb “to be” from the first sentence appears in its plural form. For example, an article about a music band will start by saying: *Midnight Oil were an Australian rock band originally performing as Farm from 1972 with drummer Rob Hirst, bass guitarist Andrew James and keyboard player/lead guitarist Jim Moginie.*

3. Collect the gender from Wikidata: We use the Wikidata id\(^2\) (hereafter wdid) to collect the gender field from Wikidata\(^3\), and review all the articles that do not have a gender\(^4\).

The aforementioned process leaves us with 49,898 articles about individual persons.

### 4.3 Birth and death year

The birth and death years are collected mainly from the Wikipedia infobox, complemented mostly with the first parenthesis from the extract, and cross validated with Wikidata. Based on the *living* parameter from the WikiProject Biography template we can identify the dead characters and collect a death year only for them.

---


\(^4\)We must note that there were 9 characters classified as transgender females that we classified as female.
We collect the infoboxes using the Wikipedia API\textsuperscript{25}. The fields corresponding to a birth year are “born”, “birth”, and “birth\_year”, and the fields corresponding to a death year are “died”, “death”, and “death\_year”. When a death year could not be parsed from any of these fields, we used the field “reign” to assign the last year of the character’s reign to his or her death. When a date was in a format not within the scope of our parser we manually revised it.

When a year was not found in the infobox we parsed the first parenthesis in the first sentence of the character’s wikipage. Our parser searches for key the strings \textit{died} or \textit{d.} to identify the death year. If no year could be identified and the parenthesis contains a number, it is manually verified.

We use the Freebase dump collected on June 2015, and the Pantheon 1.0 dataset to fill the missing birth years.

Finally, we collect the birth and death years from Wikidata and use them as a validation of our previous results and as a new source of data. We revise every biography where there was a discrepancy between Wikipedia and Wikidata greater than: 5 years for biographies after 2000, 10 years for biographies after 1900, 15 years for biographies after 1700, 25 years for biographies after 1500, 50 years for biographies after 1000, and 75 years for the rest. We then fill the missing birth years with the data from Wikidata.

Each year data point is tagged according to its source:

- \textit{ibox}: Wikipedia infobox.
- \textit{pt}: Wikipedia parenthesis.

• *est*: Manually estimated. Corresponds to characters for which only the birth century is known, or characters for which the birth (death) year was estimated based on their death (birth) year.

• *estwd*: Automatically estimated from Wikidata. Corresponds to characters for which only the birth century is known.

• fdt: Freebase dump.

• pan: Pantheon 1.0.

• wd: Wikidata.

• text: Manually collected from the Wikipedia text.

• ext: Manually collected from external sources.

### 4.4 Birth place

In order to connect each character in the dataset to a civilization of birth rather than a country of birth, we need to assign each character to the most fine-grained location of birth possible. Some characters will be assigned to a country (such as George Huntington born in the United States), others to a city (such as John Jay born in New York City), and others to a neighborhood (such as Joan Rivers born in Brooklyn, New York). We will then use these coordinates together with their birth year to obtain their place of birth based on The GeaCron Project [40]. The GeaCron Project provides a geo-temporal database to make historic information universally accessible, and intends to be a facilitator for those websites that have historical content.
New York City
From Wikipedia, the free encyclopedia

"NYC" and "New York, New York" redirect here. For other uses, see New York City (disambiguation), NYC (disambiguation), and New York, New York (disambiguation).

The City of New York, often called New York City or simply New York, is the most populous city in the United States. Located at the southern tip of the State of New York, the city is the center of the New York metropolitan area, one of the most populous urban agglomerations in the world. A global power city, New York City exerts a significant impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment, its fast pace defining the term "New York pace." Home to the headquarters of the United Nations, New York is an important center for international diplomacy and has been described as the cultural and financial capital of the world.

Situated on one of the world's largest natural harbors, New York City consists of five boroughs, each of which is a separate county of New York State. The five boroughs — Brooklyn, Queens, Manhattan, The Bronx, and Staten Island — were consolidated into a single city in 1898. With a census-estimated 2015 population of 8,550,405 distributed over a land area of just 305 square miles (790 km²), New York is the most densely populated major city in the United States. The city and its metropolitan area constitute the premier gateway for legal immigration to the United States and as many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. By 2015, the New York City metropolitan region remains by a significant margin the most populous in the United States, as defined by both the Metropolitan Statistical Area (20.2 million residents) and the Combined Statistical Area (23.7 million residents). In 2013, the metro area's gross metropolitan product (GMP) was over US$1.71 trillion, while in

Figure 4-2: Example of the Wikipedia page for New York City [56], an article within the scope of the WikiProject Cities.

We will restrict ourselves to locations with a wikipage, with special emphasis on those pages that belong to the WikiProject Cities. Figure 4-2 is an example of the wikipage of a location within the scope of the WikiProject Cities. Wikipages about locations can be identified because they have coordinates associated to them. Some places, however, have their coordinates hidden in the text, posing a problem with their identification. To solve this issue is that we put a special emphasis to the WikiProject Cities. We must note, however, that the WikiProject Cities has a very low coverage for cities in some European countries, such as France and Italy, hence we cannot restrict ourselves to places within the scope of this project.

The data collection process for Wikipedia is done according to steps from 1 to 4, outlined below. We implicitly assume that the data from step 1 is more accurate.

---

than step 2, and step 2 is more accurate than step 3, etc. Once the Wikipedia data is collected in this form, we will cross-verify it with data from Wikidata [50].

1. Attempt to collect the birth place from the infobox by looking at the fields "birth_place", "birth", and "born". Look for both the first page corresponding to a city and the first page with geographical coordinates. Ex: *Born: 23 February 1976 (age 40) Glasgow, Scotland.*

2. Attempt to collect the birth place and birth city from the first parenthesis in the extract. Ex: *Pavel Kohout (born July 20, 1928 in Prague) is a Czech and Austrian novelist, playwright, and poet.*

3. Search the wikipage for the sentence with the string "born." Identify in this sentence, the first link to a wikipage corresponding or to a city Ex: *Rufinus was born in 344 or 345 in the Roman city of Julia Concordia (now Concordia Sagittaria), near Aquileia (in modern-day Italy) at the head of the Adriatic Sea.*

4. Finally, get the first link in the page that corresponds to a page with coordinates or to a city. Despite not being very reliable, this method yields acceptable results since it tends to identify the biggest city the character lived in, or the city that operated as capital of the empire that gave them their nationality. Ex: *Sophonisba (also Sophonisbe, Sophoniba; in Punic) (fl. 203 BC) was a Carthaginian noblewoman who lived during the Second Punic War, and the daughter of Hasdrubal Gisco Gisgonis (son of Gisco).*

More than 30,000 characters were successfully assigned to a birth city, from which more than 80% came from the infobox, ~ 2% from the first parenthesis, ~ 7% were found after the word "born", ~ 3% from the text, and less than 40 characters came from external sources. More than 47,000 characters were successfully assigned to
a birth place, from which more than 89% came from the infobox, ~ 2% from the parenthesis, ~ 6% from the text, only 66 characters came from external sources, and 159 came from Pantheon.

Same as before, we use Wikidata to verify our results and to fill the missing characters.

Finally, we assign every instance of lat, lon, year to a civilization based on a GeaCron polygon [40]. GeaCron polygons capture the historical political boundaries of empires, kingdoms, countries, and city states. For more details see [40].

4.5 Occupation

The occupations associated with characters in Pantheon are meant to capture the way the character is recorded in our collective memory. Many characters hold an occupation that is not what they are remembered for. For example, Margaret Thatcher is a chemist and a lawyer, but she is recorded in our collective memory as the longest-serving British Prime Minister of the 20th century, and the only woman to have held that office so far\textsuperscript{28}; she is remembered as a politician. The available databases that contain information about historical characters, such as Freebase, Wikidata, and DBpedia, fail to associate a character to a single, most relevant, occupation.

We use machine learning to classify each character according to their major field of contribution. The approach we follow is to train a SVM classifier based on the Pantheon 1.0 dataset. Since 6 new occupations were added in Pantheon 2.0\textsuperscript{29}, we complement Pantheon 1.0 with manually classified characters for each new occupation in order to build the training set. We use the package “Natural Language Toolkit” [39] to implement the classifier.

\textsuperscript{28}https://en.wikipedia.org/wiki/Margaret_Thatcher.
\textsuperscript{29}Youtuber, gamer, go player, poker player, table tennis player, badminton player, and handball player.
The features we select for the classifier are drawn from Wikidata and Wikipedia, and are meant to characterize the character’s main field of contribution. We select the following features:

- **Infobox type**: The type of the infobox templates used in the wikipage associated with the character. We filter out all the infoboxes that do not correspond to a biography.

- **Wikidata occupations**: For each character we collect all the values from the property “occupation”\(^{30}\) from Wikidata. We delete the occupations that appear only once and build a controlled vocabulary of 350 occupations—e.g. design engineer is mapped to engineer, marine biologist to biologist, etc.

- **Extract words**: We get the top 5 most frequent words from the character’s extract, selected from a manually curated list of 750 words that are meant to capture the character’s work. For example: backstroke, summit, reign, contribution, youtube, football, testament, etc.

We train and implement three classifiers, one for each of the three levels of the Pantheon hierarchical classification (see [58] for more information about Pantheon classification). We finally compare the results of the three classifiers and manually solve the inconsistencies—e.g. if a character is classified as a writer, language, and public figure, there is an inconsistency since writer belongs to humanities, not to public figure.

The challenge of building a classifier that is able to distinguish different occupations lies on the fact that even for human beings, is very difficult to assign a single field of contribution to each character [58]. Therefore, we expect our classification not to be perfect for all characters, but to yield good aggregate results. When testing the

\(^{30}\text{https://www.wikidata.org/wiki/Property:P106.}\)
classifier by setting aside 10% of the data, we get an in-sample error of 15%, before the cross-validation between the three datasets.

4.6 Summary

We have outlined the data collection process for the Pantheon 2.0 dataset, an extension that incorporates four times as many characters and expands the historical coverage by assigning each character to a birth place based on historical borders rather than modern day borders. In principle, the methods used to generate Pantheon 2.0 can be scaled to further reduce the threshold in the number of Wikipedia language editions.

In Pantheon 2.0 each character instance contains the following fields:

- *en_curid*: Page ID for the character's english wikipage.
- *wd_id*: Wikidata ID.
- *birth_year*: Character's birth year.
- *birth_year_source*: Source of the data used for the character's birth year.
- *death_year*: Character's death year.
- *death_year_source*: Source of the data used for the character's death year.
- *birth_city_name*: City of birth, based on the WikiProject Cities.
- *birth_civilization*: Civilization of birth based on the GeaCron Project [40].
• *birth_place_curid*: English Wikipedia page ID for the character’s most fined grained place of birth.

• *birth_place_lat*: Latitude for the character’s most fined grained place of birth.

• *birth_place_lon*: Longitude for the character’s most fined grained place of birth.

• *occupation*: Character’s occupation.

The Pantheon 2.0 dataset will be made available in an online platform in the forthcoming months.
Chapter 5

When communicative and cultural memory merge

“In this electronic age we see ourselves being translated more and more into the form of information, moving toward the technological extension of consciousness.”
Marshall McLuhan

My grandmother’s most valuable possession is a shoebox containing newspaper scraps and pictures from her younger days. This way of remembering might seem very foreign to us in a world where we only need to save a couple of key words from an article of our interest, in order to access it whenever we need it. The stories that live inside my grandmother’s shoebox belong to the communicative memory of my family and have not yet found a place in the cultural memory of our society. I might tell future generations about my great-grandfather, the man who beat up the devil¹, or my grandfather and his radio show, but that will be the end of it. Or would it?

¹My great grandfather lived in a mining town where the mining company hired a man to dress up as the devil and scare the miners at night so that they would not stay up late drinking. He liked to drink.
Digital communication technologies are changing the way we build our collective memory by changing the way we communicate and store information. On one hand we have the effect of a global village, where cultures are starting to assimilate with one another. On the other hand we have the effect of how cheap and easy it is to communicate and to store information. My grandmother’s personal pictures will live only inside her shoebox, but my personal pictures have been copied so many times in different servers (Facebook, Flickr, Twitter, blogs, and other websites) that at least some of them might outlive me.

The digital age seems to be the time when the lines between communicative memory and cultural memory become blurry. More and more, messages, such as family stories, personal opinions, and personal points of view, which traditionally formed part of our communicative memory, are broadcasted into the digital world allowing everyone to revisit and reuse them. What will be the effect of this new widespread availability of information resources? Most of the questions regarding the effect that the new digital technologies will have in every sphere of our society are still open. This work adds to our understanding of this question by documenting a strong connection between the predominant communication technology of a time period and the size and composition of human collective memory.

Here, we used two large biographical datasets to test the consequences of the ideas of Innis, McLuhan, and Eisenstein regarding the effect of communication technologies in human collective memory. First, we studied the effect on 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 began to grow faster than the global population. Second, we documented a strong connection between the predominant communication technol-
ogy of a time period and the occupations of the biographies recorded in the Pantheon 1.0 dataset. 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 prior to the invention of the movable type press).

The introduction of printing, however, enabled the emergence of a new cultural elite, populated by scientists and artists. Similar shifts happened with the introduction of radio and cinema, and television, which gave rise to an elite of performers, including actors, musician, and sports players. 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.

The study of collective memory can also help shed some light on the question of how is it that humans are able to build on top of each other’s knowledge to create technologies that are far more complex than anything we can create individually [8, 26]. In this context, our collective memory is the result of the process of accumulating culture [21]. Here we have connected our ability to embody information in physical systems with our ability to accumulate cultural information by showing the effect of two communication revolutions in the size of our collective memory. Both transitions, the printing press and the public sphere, affected our collective memory through different mechanisms. On one hand, printing decreased the cost of encoding information [11, 12], and on the other, the public sphere enabled people to interact with a variety of sources [18] effectively increasing the number of people we can acquire knowledge from, our “cultural demonstrators” [13]. Our species’ ability to accumulate knowledge
has already been connected in with the number of available cultural demonstrators [13,26].

But while our study helps provide evidence in favor of the theories 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 [23].

Going forward, however, hypotheses 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.
Appendix A

Aggregation of Pantheon 1.0

Table A.1: Description of Pantheon 1.0 categories and the aggregation used in our analysis. \( N \) is the number of people in each occupation.

<table>
<thead>
<tr>
<th>Category</th>
<th>Occupation</th>
<th>Industry</th>
<th>Domain</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>Fashion designer</td>
<td>Design</td>
<td>Arts</td>
<td>10</td>
</tr>
<tr>
<td>Arts</td>
<td>Game designer</td>
<td>Design</td>
<td>Arts</td>
<td>4</td>
</tr>
<tr>
<td>Arts</td>
<td>Designer</td>
<td>Design</td>
<td>Arts</td>
<td>16</td>
</tr>
<tr>
<td>Arts</td>
<td>Comic artist</td>
<td>Design</td>
<td>Arts</td>
<td>24</td>
</tr>
<tr>
<td>Arts</td>
<td>Architect</td>
<td>Design</td>
<td>Arts</td>
<td>73</td>
</tr>
<tr>
<td>Arts</td>
<td>Photographer</td>
<td>Fine arts</td>
<td>Arts</td>
<td>12</td>
</tr>
<tr>
<td>Arts</td>
<td>Sculptor</td>
<td>Fine arts</td>
<td>Arts</td>
<td>21</td>
</tr>
<tr>
<td>Arts</td>
<td>Artist</td>
<td>Fine arts</td>
<td>Arts</td>
<td>88</td>
</tr>
<tr>
<td>Arts</td>
<td>Painter</td>
<td>Fine arts</td>
<td>Arts</td>
<td>178</td>
</tr>
<tr>
<td>Arts</td>
<td>Composer</td>
<td>Music</td>
<td>Arts</td>
<td>225</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Dancer</td>
<td>Dance</td>
<td>Arts</td>
<td>12</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Comedian</td>
<td>Film and theater</td>
<td>Arts</td>
<td>4</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Film director</td>
<td>Film and theater</td>
<td>Arts</td>
<td>177</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Actor</td>
<td>Film and theater</td>
<td>Arts</td>
<td>1193</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Musician</td>
<td>Music</td>
<td>Arts</td>
<td>381</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Singer</td>
<td>Music</td>
<td>Arts</td>
<td>437</td>
</tr>
<tr>
<td>Performing arts</td>
<td>Conductor</td>
<td>Music</td>
<td>Arts</td>
<td>11</td>
</tr>
<tr>
<td>Humanities</td>
<td>Historian</td>
<td>History</td>
<td>Humanities</td>
<td>48</td>
</tr>
<tr>
<td>Humanities</td>
<td>Critic</td>
<td>Language</td>
<td>Humanities</td>
<td>5</td>
</tr>
<tr>
<td>Humanities</td>
<td>Journalist</td>
<td>Language</td>
<td>Humanities</td>
<td>19</td>
</tr>
<tr>
<td>Humanities</td>
<td>Linguist</td>
<td>Language</td>
<td>Humanities</td>
<td>21</td>
</tr>
<tr>
<td>Humanities</td>
<td>Writer</td>
<td>Language</td>
<td>Humanities</td>
<td>955</td>
</tr>
<tr>
<td>Humanities</td>
<td>Philosopher</td>
<td>Philosophy</td>
<td>Humanities</td>
<td>281</td>
</tr>
<tr>
<td>Government</td>
<td>Judge</td>
<td>Government</td>
<td>Institutions</td>
<td>9</td>
</tr>
<tr>
<td>Government</td>
<td>Public worker</td>
<td>Government</td>
<td>Institutions</td>
<td>14</td>
</tr>
<tr>
<td>Government</td>
<td>Diplomat</td>
<td>Government</td>
<td>Institutions</td>
<td>36</td>
</tr>
<tr>
<td>Government</td>
<td>Nobleman</td>
<td>Government</td>
<td>Institutions</td>
<td>116</td>
</tr>
<tr>
<td>Government</td>
<td>Politician</td>
<td>Government</td>
<td>Institutions</td>
<td>2528</td>
</tr>
<tr>
<td>Government</td>
<td>Military personnel</td>
<td>Military</td>
<td>Institutions</td>
<td>223</td>
</tr>
<tr>
<td>Government</td>
<td>Pilot</td>
<td>Military</td>
<td>Institutions</td>
<td>9</td>
</tr>
<tr>
<td>Religion</td>
<td>Religious figure</td>
<td>Religion</td>
<td>Institutions</td>
<td>517</td>
</tr>
<tr>
<td>Science</td>
<td>Computer scientist</td>
<td>Computer science</td>
<td>Science &amp; tech.</td>
<td>34</td>
</tr>
<tr>
<td>Science</td>
<td>Engineer</td>
<td>Engineering</td>
<td>Science &amp; tech.</td>
<td>41</td>
</tr>
<tr>
<td>Science</td>
<td>Inventor</td>
<td>Invention</td>
<td>Science &amp; tech.</td>
<td>67</td>
</tr>
<tr>
<td>Science</td>
<td>Statistician</td>
<td>Math</td>
<td>Science &amp; tech.</td>
<td>4</td>
</tr>
<tr>
<td>Science</td>
<td>Mathematician</td>
<td>Math</td>
<td>Science &amp; tech.</td>
<td>157</td>
</tr>
<tr>
<td>Science</td>
<td>Physician</td>
<td>Medicine</td>
<td>Science &amp; tech.</td>
<td>143</td>
</tr>
<tr>
<td>Science</td>
<td>Geologist</td>
<td>Natural Sc.</td>
<td>Science &amp; tech.</td>
<td>10</td>
</tr>
<tr>
<td>Science</td>
<td>Astronomer</td>
<td>Natural Sc.</td>
<td>Science &amp; tech.</td>
<td>83</td>
</tr>
<tr>
<td>Science</td>
<td>Biologist</td>
<td>Natural Sc.</td>
<td>Science &amp; tech.</td>
<td>141</td>
</tr>
<tr>
<td>Science</td>
<td>Chemist</td>
<td>Natural Sc.</td>
<td>Science &amp; tech.</td>
<td>220</td>
</tr>
<tr>
<td>Science</td>
<td>Physicist</td>
<td>Natural Sc.</td>
<td>Science &amp; tech.</td>
<td>268</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-----</td>
</tr>
<tr>
<td>Science</td>
<td>Political scientist</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>7</td>
</tr>
<tr>
<td>Science</td>
<td>Anthropologist</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>11</td>
</tr>
<tr>
<td>Science</td>
<td>Geographer</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>14</td>
</tr>
<tr>
<td>Science</td>
<td>Sociologist</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>15</td>
</tr>
<tr>
<td>Science</td>
<td>Psychologist</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>38</td>
</tr>
<tr>
<td>Science</td>
<td>Economist</td>
<td>Social Sc.</td>
<td>Science &amp; tech.</td>
<td>102</td>
</tr>
<tr>
<td>Sports</td>
<td>A. football player</td>
<td>Team sports</td>
<td>Sports</td>
<td>1</td>
</tr>
<tr>
<td>Sports</td>
<td>Hockey player</td>
<td>Team sports</td>
<td>Sports</td>
<td>2</td>
</tr>
<tr>
<td>Sports</td>
<td>Cricketer</td>
<td>Team sports</td>
<td>Sports</td>
<td>2</td>
</tr>
<tr>
<td>Sports</td>
<td>Baseball player</td>
<td>Team sports</td>
<td>Sports</td>
<td>5</td>
</tr>
<tr>
<td>Sports</td>
<td>Referee</td>
<td>Team sports</td>
<td>Sports</td>
<td>10</td>
</tr>
<tr>
<td>Sports</td>
<td>Basketball player</td>
<td>Team sports</td>
<td>Sports</td>
<td>71</td>
</tr>
<tr>
<td>Sports</td>
<td>Coach</td>
<td>Team sports</td>
<td>Sports</td>
<td>75</td>
</tr>
<tr>
<td>Sports</td>
<td>Soccer player</td>
<td>Team sports</td>
<td>Sports</td>
<td>1064</td>
</tr>
<tr>
<td>Other</td>
<td>Producer</td>
<td>Business</td>
<td>Business &amp; law</td>
<td>12</td>
</tr>
<tr>
<td>Other</td>
<td>Businessperson</td>
<td>Business</td>
<td>Business &amp; law</td>
<td>79</td>
</tr>
<tr>
<td>Other</td>
<td>Lawyer</td>
<td>Law</td>
<td>Business &amp; law</td>
<td>17</td>
</tr>
<tr>
<td>Other</td>
<td>Astronaut</td>
<td>Explorers</td>
<td>Exploration</td>
<td>32</td>
</tr>
<tr>
<td>Other</td>
<td>Explorer</td>
<td>Explorers</td>
<td>Exploration</td>
<td>70</td>
</tr>
<tr>
<td>Other</td>
<td>Social activist</td>
<td>Activism</td>
<td>Public figure</td>
<td>114</td>
</tr>
<tr>
<td>Other</td>
<td>Companion</td>
<td>Companions</td>
<td>Public figure</td>
<td>101</td>
</tr>
<tr>
<td>Other</td>
<td>Pornographic actor</td>
<td>Media personality</td>
<td>Public figure</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>Celebrity</td>
<td>Media personality</td>
<td>Public figure</td>
<td>21</td>
</tr>
<tr>
<td>Other</td>
<td>Chef</td>
<td>Media personality</td>
<td>Public figure</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>Magician</td>
<td>Media personality</td>
<td>Public figure</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>Presenter</td>
<td>Media personality</td>
<td>Public figure</td>
<td>19</td>
</tr>
<tr>
<td>Other</td>
<td>Model</td>
<td>Media personality</td>
<td>Public figure</td>
<td>30</td>
</tr>
<tr>
<td>Other</td>
<td>Pirate</td>
<td>Outlaws</td>
<td>Public figure</td>
<td>9</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>---------</td>
<td>---------------</td>
<td>---</td>
</tr>
<tr>
<td>Other</td>
<td>Mafioso</td>
<td>Outlaws</td>
<td>Public figure</td>
<td>13</td>
</tr>
<tr>
<td>Other</td>
<td>Extremist</td>
<td>Outlaws</td>
<td>Public figure</td>
<td>34</td>
</tr>
<tr>
<td>Other</td>
<td>Golfer</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>Snooker</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>Mountaineer</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>Gymnastic</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>7</td>
</tr>
<tr>
<td>Other</td>
<td>Martial arts</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>7</td>
</tr>
<tr>
<td>Other</td>
<td>Skater</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>Boxer</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td>Skier</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>17</td>
</tr>
<tr>
<td>Other</td>
<td>Swimmer</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>20</td>
</tr>
<tr>
<td>Other</td>
<td>Cyclist</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>29</td>
</tr>
<tr>
<td>Other</td>
<td>Chessmaster</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>30</td>
</tr>
<tr>
<td>Other</td>
<td>Wrestler</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>44</td>
</tr>
<tr>
<td>Other</td>
<td>Athlete</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>74</td>
</tr>
<tr>
<td>Other</td>
<td>Racecar driver</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>104</td>
</tr>
<tr>
<td>Other</td>
<td>Tennis player</td>
<td>Ind. sports</td>
<td>Sports</td>
<td>161</td>
</tr>
</tbody>
</table>
Bibliography


