Pantheon

Visualizing the Structure and Dynamics of Global Cultural Production

Amy (Zhao) Yu
B.A.S. Computer Science
B.S.E. Finance, Operations and Information Management
University of Pennsylvania, 2008

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 of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

June 2014
© Massachusetts Institute of Technology, 2014. All rights reserved.

Author

Amy (Zhao) Yu
Program in Media Arts and Sciences
May 16, 2014

Certified By

Dr. César A. Hidalgo
Assistant Professor in Media Arts and Sciences
Program in Media Arts and Sciences

Accepted By

Dr. Pattie Maes
Professor of Media Technology, Associate Academic Head
Program in Media Arts and Sciences
Pantheon

Visualizing the Structure and Dynamics of Global Cultural Production

Amy (Zhao) Yu

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
On May 16, 2014
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

Abstract

Cultural information influences the human experience across all scales - it is embodied in the institutions, beliefs, and technologies that shape our daily lives, it influences the advancement of industrial capacities, public services, and governance, and it impacts the trajectory of global development. While previous efforts have introduced various attempts for quantifying culture, we lack more direct measures of global cultural production—measures that enumerate the cultural contributions that have emerged throughout time, connected to geographies and languages. This thesis introduces a new approach for quantifying culture through a dataset and visualization platform for creating and visualizing metrics of global cultural production, defined as the cultural accomplishments that have broken the barriers of space, time and language. To develop measures of cultural production, I generate a dataset with the 11,340 biographies present in more than 25 languages on Wikipedia—the largest multilingual encyclopedia on the planet—and link each biography to a cultural domain, place of birth, and time period. We make this data available to public audiences via Pantheon, a platform for exploring the dataset through dynamic and interactive visualizations. I validate the dataset using accepted measures of recognition in various cultural domains, and I evaluate Pantheon by testing the significance of the patterns that it reveals within the structure and dynamics of global cultural production. Ultimately, the goal of this dataset and visualization engine is to motivate further work that builds on the measures of cultural achievements introduced by this thesis.

Advisor

Dr. César A. Hidalgo
Assistant Professor in Media Arts and Sciences
Program in Media Arts and Sciences
Pantheon
*Visualizing the Structure and Dynamics of Global Cultural Production*

Amy (Zhao) Yu

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
On May 16, 2014
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

Advisor

**Dr. César A. Hidalgo**
Asahi Broadcast Corporation Career Development Professor
Assistant Professor in Media Arts and Sciences
Massachusetts Institute of Technology

Reader

**Ethan Zuckerman**
Director, Center for Civic Media
Principal Research Scientist
Massachusetts Institute of Technology

Reader

**Dr. Hanspeter Pfister**
Director, Institute of Applied Computational Science
An Wang Professor of Computer Science
Harvard University
Acknowledgements

This thesis results from two years of collaboration and conversation with many wonderful people who have supported, challenged, and inspired me throughout my journey at the Media Lab. I am grateful to everyone who I have met along the way, and helped shape my time thus far at MIT. I especially would like to thank the following:

My advisor, César Hidalgo, thank you for your unlimited encouragement, feedback, and support over the course the entire project. Your advice and guidance were invaluable for cultivating the vision of Pantheon, and I look forward to continuing with you on the journey ahead.

The Pantheon team, your contributions were instrumental in making our vision into a reality. I am grateful to Shahar Ronen, Tiffany Lu, and Defne Gurel for your efforts in curating and wrangling the Pantheon dataset, and Ali Almossawi and Deepak Jagdish for contributing your design perspective and helping to shape Pantheon’s evolution. I would like to express my deepest thanks to Kevin Hu for the countless hours spent together building, coding, and collaborating on all parts of Pantheon – thank you for your boundless energy, dedication, and encouragement along the way.

My readers, Ethan Zuckerman and Hanspeter Pfister – thank you for supporting me, challenging me, and inspiring me over the past year. I also thank Alexander Lex for his insight, comments, and suggestions.

Macro Connections, thank you for your advice, your support, and most of all, your friendship. I am truly lucky to be a part of such a talented, creative, and intellectually promiscuous group. I’m always learning new things from each and every one of you.
Charles Murray, thanks for providing the Human Accomplishments dataset and adding another dimension to Pantheon’s visualizations.

Other members of the MIT Community who have supported me in this work over the last two years: the Data Centric Projects class for the feedback in the early stages of the project and Rahul Bhargava for his advice on all things coding and data related, Marco Tempest and Signe Fleishmann, for their feedback and support as part of the Media Lab Director’s Fellows program, and J. Nathan Matias and Valentin Heun for their helpful comments in preparation for Crit Day.

My family & friends, both near and far – thank you for your unconditional love and encouragement in all of my pursuits. And last but certainly not least, my better half, Andrew Mao, thank you for your unrelenting support in all aspects of our life together. You are my rock and my fortress and this thesis is dedicated to you with love.
Table of Contents

ACKNOWLEDGEMENTS .................................................................................................................. 5

1. INTRODUCTION .......................................................................................................................... 10
   1.1. DEFINING CULTURE .............................................................................................................. 10
   1.2. QUANTIFYING CULTURE ..................................................................................................... 11
   1.3. VISUALIZING HISTORICAL CULTURAL PRODUCTION .......................................................... 13

2. THE DATA ...................................................................................................................................... 17
   2.1. DATA COLLECTION ................................................................................................................ 17
   2.2. BIASES & LIMITATIONS ....................................................................................................... 20
   2.3. MEASURES OF GLOBAL CULTURAL PRODUCTION ............................................................... 24
   2.4. DATA VALIDATION ................................................................................................................ 27

3. PANTHEON: DESIGN & DEVELOPMENT ..................................................................................... 35
   3.1. INTERFACE DESIGN EVOLUTION ......................................................................................... 35
   3.2. VISUALIZATIONS .................................................................................................................. 38
       Treemaps .................................................................................................................................. 40
       Matrices .................................................................................................................................... 41
       Scatterplots ............................................................................................................................... 42
       Maps .......................................................................................................................................... 43
   3.3. OTHER PAGES AND FEATURES .......................................................................................... 45
       Rankings .................................................................................................................................... 45
       People profiles ......................................................................................................................... 47
       Tutorials .................................................................................................................................... 48
       Mobile layout ............................................................................................................................ 50
   3.4. TECHNICAL DETAILS .......................................................................................................... 50
       Database .................................................................................................................................... 50
       Application Framework ............................................................................................................. 51

4. LAUNCH & USAGE ....................................................................................................................... 53
   4.1. GLOBAL IMPRESSIONS ......................................................................................................... 53

5. ANALYSIS & IMPLICATIONS ....................................................................................................... 56
   5.1. THE NESTEDNESS OF CULTURAL PRODUCTION ................................................................. 56
5.2. Diversity of Cultural Production ................................................................. 61
5.3. Dynamics of Global Cultural Production .................................................... 63

6. **Conclusion** ........................................................................................................... 64
   6.1. Future Directions ............................................................................................... 64

7. **Appendix** ............................................................................................................. 66
   7.1. Cultural Domain Classifications ...................................................................... 66
   7.2. Historical Popularity Index Derivation ............................................................ 69
   7.3. Pantheon MongoDB Document Structure ....................................................... 70
   7.4. Seven Principles of Meteor ............................................................................. 71

8. **References** ........................................................................................................... 72
## Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nightingale's Rose</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>Nightingale's Rose, revisited</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>Cultural Domain Classification Hierarchy</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>Wikipedia Presence of Chilean Presidents</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>Wikipedia Presence of Academy Award Winners for Best Actress</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Wikipedia Presence of Top Scorers of the Spanish Liga</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>Cumulative Number of Individuals with at least N Wikipedia Language Editions</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>Validation using Formula 1 drivers</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>Validation Using Tennis Players</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>Validation Using Swimmers</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>Becoming Pantheon</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>Pantheon 1.0 Main Interface Layout</td>
<td>37</td>
</tr>
<tr>
<td>13</td>
<td>Visualizations</td>
<td>39</td>
</tr>
<tr>
<td>14</td>
<td>Treemap Examples - by Place of Birth, by Domain</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>Matrix Visualization of World Cultural Production, 4000BC - 2010</td>
<td>41</td>
</tr>
<tr>
<td>16</td>
<td>Scatterplots in Pantheon</td>
<td>43</td>
</tr>
<tr>
<td>17</td>
<td>Maps in Pantheon</td>
<td>44</td>
</tr>
<tr>
<td>18</td>
<td>Rankings within Pantheon</td>
<td>46</td>
</tr>
<tr>
<td>19</td>
<td>Michael Faraday's Pantheon Profile</td>
<td>47</td>
</tr>
<tr>
<td>20</td>
<td>Nora's Tutorial</td>
<td>49</td>
</tr>
<tr>
<td>21</td>
<td>Pantheon Mobile</td>
<td>50</td>
</tr>
<tr>
<td>22</td>
<td>Meteor.js Framework</td>
<td>51</td>
</tr>
<tr>
<td>23</td>
<td>Pantheon Page Views, March - April 2014</td>
<td>53</td>
</tr>
<tr>
<td>24</td>
<td>Global Usage of Pantheon, by user session</td>
<td>54</td>
</tr>
<tr>
<td>25</td>
<td>Most Viewed Pages</td>
<td>55</td>
</tr>
<tr>
<td>26</td>
<td>Matrix View on Pantheon</td>
<td>57</td>
</tr>
<tr>
<td>27</td>
<td>Nestedness of Cultural Production - Occupations</td>
<td>59</td>
</tr>
<tr>
<td>28</td>
<td>Nestedness of Cultural Production - Industries</td>
<td>60</td>
</tr>
<tr>
<td>29</td>
<td>Scatterplot of Countries (of birth) by Number of Characters and Diversity</td>
<td>62</td>
</tr>
</tbody>
</table>
1. Introduction

In this thesis, I address the challenge of measuring and visualizing global cultural production by presenting a carefully curated dataset containing more than eleven thousand culturally relevant biographies including all countries in the world and much of recorded history. All of these biographies are mapped to a cultural domain using a normalized three level hierarchical classification that we introduce (see Appendix). Also, 95% of all biographies are mapped to a country of birth and 100% to a year of birth. The biographies in this dataset represent the entire set of biographies with a presence in more than 25 language editions of Wikipedia (as of May 2013), representing the largest curated dataset of global cultural production available to date. To quantify the relative importance of each biography, the dataset also incorporates information on the number of languages in which each biography has a presence and the number of page views received by each Wikipedia page in each language. Using this dataset, we developed a visualization engine, Pantheon, to reveal the patterns within the data and open the data for exploration by audiences around the globe. This work aims to motivate future research exploring the quantitative study of historical cultural patterns by allowing audiences to visualize cultural production and highlighting the potential of human achievements.

1.1. Defining Culture

In a broad sense, culture is the information that is created and transmitted by individuals through non-genetic means [1]. Norms and beliefs are part of this information, but culture involves more than the inheritance of views and behaviors. The cultural information that we pass on from generation to generation can be as simple as last names [2] or as complex as the knowledge of rocketry that was required to put Neil Armstrong on the moon. Cultural information can be transmitted orally, but is largely embodied in works of art, literature and artifacts [1]. The ability to amass knowledge and information, and physically embody this into artifacts, is an ancient defining trait of our species.
Culture contributes to the overall progress of civilization, and represents the process by which our society expands the set of possible human activities. Transcontinental flight, instant video communication, vaccine therapy and refrigeration are among the many accomplishments that hinge on culture and have expanded the range of humanly possible activities in recent centuries. These types of accomplishments also contribute to the development of nations – for example, the development of the phone, light bulb and airplane cannot be considered an irrelevant aspect of the development of the United States, or for that matter, the development of the countries around the world that would later adopt these inventions. Similarly, the creation of writing, the saddle and the wheel in ancient times were not irrelevant accomplishments of the civilizations that spearheaded their development, even though there is consensus that the accumulation of this knowledge did not translate into economic growth [3]. Nevertheless, the cultural legacies that have been produced over time underwrite the expansion of our human capacities, and drive advancements that are essential in the overall progress of our societies.

This thesis addresses culture from the narrow perspective of global cultural production, defined as the cultural expressions that have surpassed linguistic, temporal, and geographic boundaries. Cultural production spans across a diverse range of domains, and is not necessarily tied to indicators of economic productivity. The legacies of Aristotle's Organon, Francis Bacon’s Novo Organum, Newton’s mechanics, Laplace’s mathematics, the Beatles, and the Nouvelle Vague are all expressions that contribute to historical cultural production.

1.2. Quantifying Culture

Historians, political scientists, linguists, computer scientists and economists have previously made various attempts to define measures of culture. The dominant paradigm for quantifying culture in the last half century utilizes survey methods and self-reported questionnaires – however, these methods primarily focus on the measurement of cultural values, and scholars have questioned the validity of these metrics in the context of studying global culture [4].
Previous literature on measuring cultural accomplishment is based primarily on texts written by domain experts, providing a limited view of selected cultural domains. Charles Murray’s work on *Human Accomplishments* contributes an inventory of 4,002 significant individuals, but this inventory only captures individuals within the domains of arts and sciences [5]. More recent methods introduced in the budding field of *Culturomics* have utilized digitized texts to study cultural trends from a linguistic perspective [6]. Also, recent efforts have focused on structuring Wikipedia data [7] and quantifying the impact of individuals across a more diverse set of cultural domains [8]. Table 1 provides a non-exhaustive comparison surveying a sample of the various datasets currently used for quantifying culture, in comparison with the dataset that Pantheon introduces.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Size</th>
<th>Domains</th>
<th>Time</th>
<th>Global?</th>
<th>Multi-lingual?</th>
<th>Interactive Visualizations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Values Survey [10]</td>
<td>Values</td>
<td>85,000 respondents</td>
<td>Religion, Institutions</td>
<td>1981-present</td>
<td>Yes (87 countries)</td>
<td>Yes (20 languages)</td>
<td>Yes – bar charts &amp; maps</td>
</tr>
<tr>
<td>Human Accomplishment [5]</td>
<td>Individuals and events</td>
<td>4,002 significant individuals, 1,560 significant events</td>
<td>Arts &amp; Sciences</td>
<td>800BC-1950</td>
<td>Yes (~30 countries &amp; regions)</td>
<td>Yes (~6 languages)</td>
<td>No</td>
</tr>
<tr>
<td>Who’s Bigger [8]</td>
<td>Entities (Individuals, events, places, things)</td>
<td>843,790 individuals</td>
<td>All domains known in Wikipedia</td>
<td>All time periods covered in Wikipedia</td>
<td>Yes</td>
<td>No (English only)</td>
<td>Yes – line graphs on whosbigger.com</td>
</tr>
<tr>
<td>Pantheon</td>
<td>Individuals</td>
<td>11,340 notable individuals</td>
<td>All domains</td>
<td>4000BC – 2010</td>
<td>Yes (194 countries)</td>
<td>Yes (280 languages)</td>
<td>Yes</td>
</tr>
</tbody>
</table>
All of the aforementioned datasets capture and quantify specific aspects of culture or measure cultural production at a limited scope. These data do not characterize the production or development of culture using metrics that link cultural production across all languages, cultural domains, geographies, and time periods. This thesis builds upon the previous work on characterizing the production of cultural accomplishments by contributing (1) a dataset that includes linkages to a wider range of cultural domains, time periods, geographies, and languages, and (2) a suite of dynamic visualizations to open the data for exploration by a global audience.

1.3. Visualizing Historical Cultural Production

“The purpose of visualization is insight, not pictures.”

- Ben Shneiderman

Data visualization, the process of encoding information using visual representations of data, is a powerful tool for sharing and discovering valuable insights from large datasets. Visualization amplifies cognition by allowing us to visually record and analyze data in order to find patterns, develop hypotheses, and communicate findings [12]. The tasks supported by visualization naturally lend potential in the context of storytelling – structuring information to explain the connections and patterns within data. Historically, static visualizations have commonly been used to facilitate both analysis and storytelling. For example, Florence Nightingale’s rose graphs visually represented the mortalities during the Crimean War and allowing her to analyze the causes of death and communicate the story behind needless deaths caused by poor hospital conditions and a lack of patient welfare. These graphs (Figure 1) allowed her to share the data that she meticulously collected with politicians without formal statistical training, and communicate her message to lobby for sanitary reform and improved conditions in hospitals. Nightingale used the visualization “to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears” [13].
Figure 1: Nightingale’s Rose

Florence Nightingale’s original rose graph, showing causes of mortality in red (wounds), blue (infectious diseases) and black (other causes). The radius of each slice is proportional to the number of deaths.

Image Source: [14]

Transforming information into data visualizations creates a medium for opening and sharing both data and insights with broader audiences. However, static visualizations present limited, pre-composed views of the data. These fixed representations are only ideal when the visualization aims to present a specific perspective, or when the image is published on a static medium. Dynamic visualizations allow users to interact with the data, empowering them to explore, browse, create their own questions, and search for answers. Interactive visualizations engage audiences and allow users to discover the stories imbedded within data – whether they are browsing the data or looking for a specific insight [15]. Figure 2 shows various views that are enabled within an interactive reproduction of Nightingale’s Rose – the addition of interaction allows the viewer to see an overview of the entire dataset, or create custom views by filtering on specific dimensions, such as time or cause of death.
Figure 2: Nightingale’s Rose, revisited

Interactive views of the causes of mortality during the Crimean war – (a) by Preventable (blue) diseases and wounds (red), (b) up to December 1984.

Image Source: [14]

The evolution of the Internet as both a communication medium and publishing platform has created a fertile ecosystem for interactive data visualization. We now have a proliferation of digital data to be explored [16], freely accessible web-standard technology platforms and libraries for crafting visualization tools supported by open-source development communities, and a publishing medium that can instantaneously reach a global audience. A number of technologies have emerged for developing web-based visualization tools. D3.js, or Data-Driven Documents, is a popular Javascript library that enables the manipulation of web documents and data within the browser [17]. D3 binds data to elements within the Document Object Model (DOM), and facilitates transformations to the web document using the data - however, it does not generate pre-defined visualizations or templates to directly load external datasets. This has inspired tools built on top of d3, such as d3plus [18], NVD3 [19], Crossfilter [20], and Cubism [21] – all of which are libraries that provide customizable, reusable interactive visualizations for displaying, exploring, and experimenting with data. Finally, services such as Google Chart Tools [22] and DataWrapper [23] now allow for users to create interactive charts without prior programming experience.

The richly integrated ecosystem of web-based visualization tools has accelerated the expansion of visualization for storytelling, in which graphical and interactive elements are
used to create both author-driven and reader-driven experiences. Segel & Heer define a spectrum of visualization that has evolved to provide tools for both discovering and communicating stories by supporting analysis as well as presentation [24]. Author-driven approaches have become prevalent in mainstream online journalism, with news organizations such as the New York Times [25] and Bloomberg [26] using interactive visualization to convey and support structured narratives.

However, data stories can differ from the controlled sequential structure of traditional narrative and invite users to pose new questions and discover alternative explanations through interactivity. Platforms such as the Observatory of Economic Complexity [27] and DataViva [28] follow a reader-driven approach, by allowing a high degree of interactivity and providing little to no prescribed ordering or messaging to structure user interactions. This type of approach opens the opportunity for users to verify and explore the data themselves, and use the platform to craft their own stories and present the insights discovered within the data. Indeed, some of the prior work on quantifying culture has incorporated this approach by using basic interactive visualizations to allow general audiences to explore the relevant datasets. The Culturomics project is associated with the Google Ngram Viewer [11], which provides a customizable interactive line graph to visualize the popularity of specific phrases over time, while the Who is Bigger website [29] also provides similar visualizations for their data on Wikipedia impressions and revisions. These prior platforms and tools demonstrate how interactive visualization opens data to audiences from various disciplines and backgrounds, allowing individual users the ability to analyze the data interactively and share the insights they gain. By building tools that allow both analysis of data and presentation of stories, we create more opportunities for users to engage with the data and broaden the impact of the stories that can be told. Therefore, Pantheon contributes an integrated suite of interactive visualizations, providing a diverse range of customized views to open the exploration of historical cultural production.
2. The Data

**pan·the·on**

: a group of illustrious or notable persons or things

- Merriam-Webster Dictionary

2.1. Data Collection

In order to measure the impact and development of cultural expressions, it would be ideal to use data that encompasses all cultural expressions, across all languages, with high spatiotemporal resolution and disaggregated across a standardized cultural classification schema. Since no such dataset at this scale currently exists, we create the dataset for Pantheon using the data available on Freebase and Wikipedia - both of which are open-source, collaborative, multi-lingual knowledge bases freely available online to the general public. While previous efforts have also produced structured datasets based on Wikipedia [7, 8], Pantheon introduces a new dataset with a wider scope – with linkages across all language editions, time periods, cultural domains, and geographies. While there are certainly considerable limitations to Wikipedia and Freebase, they are currently the only datasets available that represent large, domain-independent repositories of collaboratively edited human knowledge, and past research has demonstrated the reliability of these collaborative knowledge bases [30, 31].

Since people are the source of cultural information, we use globally known individuals as a proxy for cultural production. We derive our dataset of cultural production from Freebase’s entity knowledge graph and add metadata from Wikipedia accessible through its API. Freebase organizes information as uniquely identified entities with associated types and properties defined by a structured data ontology. Therefore, to identify globally known
historical and current cultural figures, we first determined a list of individuals through Freebase’s database of all entities classified as Persons [32]. The individuals are filtered based on the availability of personal and demographic details, including birthdate, birthplace, gender, and occupation. We link each individual to their relevant entry within English Wikipedia using their unique Wikipedia article id, and to obtain the linkages to all other Wikipedia language editions through the Wikipedia API. The set of individual ids represent a snapshot of Persons as of May 2013, since the dataset of individuals was obtained in early 2013. We supplement the data with monthly page view data from the Wikipedia data dumps [33] for page views for each individual from Jan. 2008 through Dec. 2013, across all language editions of Wikipedia.

Since no globally standardized classification system currently exists for cultural production, we introduce a normalized hierarchy of cultural domains, industries and occupations, classifying cultural domains at three levels of aggregation. Figure 3 shows an example drilling down into the classification tree. To create this classification hierarchy, we use raw data on individual occupations from Freebase to create a normalized listing of occupations – for example, we map “Entrepreneur”, “Business magnate”, and “Business development” to the normalized occupation of “Businessperson”. We grouped normalized occupations into a second-tier classification (called industries), and top-level cultural domains. A table containing the full classification hierarchy is provided in the Appendix. We associate individuals within the dataset with a cultural domain based on the occupation that best encompasses their primary area of cultural contribution. Thus, we link all individuals to exactly one occupation – for example, Barack Obama is a politician, and Leonardo Da Vinci is an inventor. By assigning individuals to only one primary cultural domain, we avoid double-counting individuals within the dataset, but introduce the limitation of restricting the contribution of polymaths to one singular domain. The challenge of fairly distributing the cultural impact of polymaths will be left for future consideration.
In terms of location assignment, we attribute individuals to a place of birth by country, based on current political boundaries to standardize the methodology for associating individuals with countries. Birthplaces were obtained by scraping both Freebase and Wikipedia, and further refined by using fuzzy location matching and geocoding within the Yahoo Placemaker [34] and Google Maps geocoding APIs [35], and by manual curation. First, we normalize the raw data from Freebase indicating the city of birth by latitude and longitude using fuzzy location matching available within the geocoding APIs. Using the coordinates obtained through the APIs, individuals are then mapped to countries based on present-day geographic boundaries using reverse geocoding API available on geonames.org. For example, individuals born in Moscow during the Soviet Union era are associated with Russia. Using present-day boundaries allows for a consistent basis for matching individuals to countries, and mitigates the technological limitation of the lack of existing historical
geocoding APIs for attributing geographic boundaries using latitude, longitude, and time. Historically, birthplace is a fairly suitable way of associating individuals to countries, however, given the increase of human mobility over time and the net migration gains experienced by developed regions [36], future refinement of the dataset may include consideration for improving the attribution of individuals to the most relevant geographic area from which their cultural contribution emerges.

2.2. Biases & Limitations

As with all large data collection efforts, this is an effort that is coupled with limitations and biases. These are limitations that should be considered carefully when interpreting the results. This dataset should be interpreted narrowly, as a view of global cultural production that emerges from the multilingual expression of historical figures in the Wikipedia as of May 2013. The main biases and limitations of the dataset come from:

1. The use of Wikipedia as a data source.
2. The use of place of birth to assign locations.
3. The use of biographies as proxies for cultural production.
4. Other technical limitations.

1. The use of Wikipedia as a data source

The data is limited by the set of people who contribute to Wikipedia. Wikipedia editors are not considered to be a representative sample of the world population, but a sample of publicly minded knowledge specialists that are willing and able to dedicate time and effort to contribute to the online documentation of knowledge. Wikipedia editors have an English Bias, a Western Bias, a gender bias towards males, and they tend to be highly educated and technically inclined. They are also more prevalent among developed countries with Internet access [37]. However, by using data from all Wikipedia language editions we are effectively reducing a bias that would favor the local culture of English speakers. As an example, we note that there is only one American Football Player in the dataset: O.J. Simpson. Certainly, his global notoriety is not purely from his football career, showing that
the use of many languages reduces the English bias of the dataset (famous American Football players, such as Peyton Manning, Tom Brady and Joe Montana all have a large presence in the English Wikipedia, but fail to meet the L>25 threshold). In comparison, the dataset contains over 1,000 soccer players – showing that soccer is a sport that is integrated within global culture.

Wikipedia also has a considerable bias in the inclusion of people from different categories. This bias could be the result of the notability criteria in Wikipedia being more lenient for individuals in some categories than others, or from the Wikipedia community. For instance, Spanish Wikipedia has entries for 78% of the players in the Chilean soccer club Unión Española, and 43% of the players from the second division team Magallanes, but only for 5.5% of the faculty at MIT. This suggests that the inclusion criteria in Wikipedia might be more lenient for athletes than academics, since being a player in a second division team in Chile is more likely to pass the notability criteria required by Wikipedia Editors than being a faculty at MIT.

Finally, Wikipedia also has a notable recency bias. Figures 4-6 illustrate and document this bias, showing three datasets of individuals with a similar level of accomplishment. The categories considered are Chilean presidents, top scorers of the Spanish Liga, and the winners of the Academy Award for best actress. All three data series demonstrate a positive correlation between the number of languages where these individuals have a presence and the year of their birth, indicating that the dataset is biased in favor of more recent individuals. The recency bias is interesting in the sense that it also exemplifies a bias of the Internet as new media. Certainly, individuals whose fame coincided with the existence of the Internet will be overrepresented in Wikipedia. Therefore, the bias could be interpreted as a combination of both, a bias for recency, but also, a bias from the media available at the time in which the individual became known for a cultural contribution.
Figures 4-6 illustrate the recency bias of Wikipedia, shown in the positive correlation between number of Wikipedia language editions and birthyear, for individuals with similar accomplishment levels.
2. The use of place of birth to assign locations

Individuals were assigned to geographic locations using their place of birth, based on present-day political boundaries. Since neither Freebase nor Wikipedia have clean, normalized data on places of birth, country assignments were complemented with geocoding APIs for normalization and manual curation (to correct for errors in API and completeness). Place of birth is one way of assigning a location to an individual that allow us to assign locations in a comprehensive and consistent manner. Yet, there are biases and limitations that need to be considered when using this location assignment method. An important limitation is the inability to account for individuals who became globally known after immigrating to another country. Would Neruda, Picasso or Hemingway be as famous if they had not participated of the Parisian art scene? The place where an individual was born is often different from the place where that individual made his or her more important contributions. In some case, the contributions are made in a number of places, and the use of birthplace is unable to capture where the contributions were made. This is particularly true for athletes who migrate to the world’s most competitive leagues, or artists that move to the artistic centers of their time. In this dataset, such individuals are not represented since programmatically geo-coding birthplaces is more consistent than registering the place where each individual made his or her more significant contribution, which can only be found through the unstructured data buried in historical narratives.

3. The use of biographies as proxies for cultural production

Using biographies to approximate cultural production excludes accomplishments where a clear connection between a cultural expression and its creator is not apparent, or when a cultural accomplishment cannot be attributed to single individuals. For example, consider collective enterprises where the accomplishments are the results of teams and not isolated individuals. Examples of accomplishments that are likely to get excluded include the works of music bands or orchestras, or the products produced by a firm, where the accolades collected from accomplishments are connected to a firm, or brand, rather than to an individual.
4. Other Technical Limitations

Other biases and limitations include the volatility of Wikipedia and other online APIs, which make the results presented here imperfectly reproducible. For example, the Yahoo Placemaker API, which was used for mapping individuals to countries by birthplace, has been deprecated recently and is no longer publicly available. Also, the set of included individuals is static and does not reflect events after early 2013 - as such, culturally impactful individuals who only recently rose to global prominence, including Pope Francis and Prince George of Cambridge, are excluded from this dataset.

2.3. Measures of Global Cultural Production

Global cultural production is measured by estimating the number of notable people associated with geographic areas based on present-day political boundaries. The illustriousness of historical characters is estimated using two measures. The simpler of the two measures, denoted as $L$, is the number of different Wikipedia language editions that have an article about a historical character. The documentation of an individual in multiple languages is a good first approximation for their notability because it indicates accomplishments that have broken successive language barriers. This helps to differentiate between global culture—information that has broken many linguistic barriers and local culture, which is information that is contained primarily in a few language groups.

Another measure, called the Historical Popularity Index (HPI), is a more nuanced metric for global cultural impact. HPI augments the simple metric of $L$ by including information on the time since birth of the historical character (A, calculated by 2013-birthyear, as a proxy for breaking the barrier of time), an $L^*$ measure that adjusts $L$ by accounting for the concentration of pageviews among different languages (to discount characters with pageviews mostly in a few languages), the coefficient of variation (CV) in pageviews (to discount characters that have short periods of popularity), and the number of non-English Wikipedia pageviews ($V^{NE}$) to further reduce any English bias. In addition, to dampen the recency bias of the data, HPI is adjusted for individuals known for less than 70
years. Formula 1 provides the full calculation for HPI, a detailed derivation of HPI is provided in the Appendix.

\[ HPI = \begin{cases} 
\ln(L) + \ln(L^*) + \log_4(A) + \ln(p^{NF}) - \ln(CV) & \text{if } A \geq 70 \\
\ln(L) + \ln(L^*) + \log_4(A) + \ln(p^{NF}) - \ln(CV) - \frac{70-A}{7} & \text{if } A < 70 
\end{cases} \]  

(1)

Table 2 shows the top ten people with the highest HPI for a few selected periods (before the year 500, between 500 and 1200, between 1200 and 1500, between 1500 and 1750, between 1750 and 1850, between 1850 and 1900, between 1900 and 1950). The table shows, that the most notable biographies for each period are associated primarily with well-known historical characters. Going forward, Historical Popularity Index (HPI) will be used as a proxy for an individual’s notability.

Table 2: Notable Biographies by Time Period

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Aristotle</td>
<td>• Muhammad</td>
<td>• William Shakespeare</td>
<td>• Wolfgang Amadeus Mozart</td>
<td>• Adolf Hitler</td>
<td>• Che Guevara</td>
<td>• Mother Teresa</td>
</tr>
<tr>
<td>• Plato</td>
<td>• Charlemagne</td>
<td>• Shakespeare</td>
<td>• Mozart</td>
<td>• Albert Einstein</td>
<td>• Martin Luther King, Jr.</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Jesus Christ</td>
<td>• Genghis</td>
<td>• Michelangelo</td>
<td>• Napoleon</td>
<td>• Vincent van Gogh</td>
<td>• Walt Disney</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Socrates</td>
<td>• Khan</td>
<td>• Columbus</td>
<td>• Bonaparte</td>
<td>• Sigmund Freud</td>
<td>• Jean-Paul Sartre</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Alexander the Great</td>
<td>• Saladin</td>
<td>• Da Vinci</td>
<td>• Bach</td>
<td>• Pablo Picasso</td>
<td>• Bob Marley</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Confucius</td>
<td>• Leidias I</td>
<td>• Michelangelo</td>
<td>• Galileo</td>
<td>• Salvador Dalí</td>
<td>• Jimi Hendrix</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Julius Caesar</td>
<td>• Avicenna</td>
<td>• Caesar</td>
<td>• Galileo</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Homer</td>
<td>• Ali</td>
<td>• Leonardo da Vinci</td>
<td>• Immanuel Kant</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Pythagoras</td>
<td>• Li Bai</td>
<td>• da Vinci</td>
<td>• Descartes</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Archimedes</td>
<td>• Francis of Assisi</td>
<td>• Raphael</td>
<td>• Jean d’Arc</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Aristotle</td>
<td>• Muhammad</td>
<td>• William Shakespeare</td>
<td>• Wolfgang Amadeus Mozart</td>
<td>• Adolf Hitler</td>
<td>• Che Guevara</td>
<td>• Mother Teresa</td>
</tr>
<tr>
<td>• Plato</td>
<td>• Charlemagne</td>
<td>• Shakespeare</td>
<td>• Mozart</td>
<td>• Albert Einstein</td>
<td>• Martin Luther King, Jr.</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Jesus Christ</td>
<td>• Genghis</td>
<td>• Michelangelo</td>
<td>• Napoleon</td>
<td>• Vincent van Gogh</td>
<td>• Walt Disney</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Socrates</td>
<td>• Khan</td>
<td>• Columbus</td>
<td>• Bonaparte</td>
<td>• Sigmund Freud</td>
<td>• Jean-Paul Sartre</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Alexander the Great</td>
<td>• Saladin</td>
<td>• Da Vinci</td>
<td>• Bach</td>
<td>• Pablo Picasso</td>
<td>• Bob Marley</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Confucius</td>
<td>• Leidias I</td>
<td>• Michelangelo</td>
<td>• Galileo</td>
<td>• Salvador Dalí</td>
<td>• Jimi Hendrix</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Julius Caesar</td>
<td>• Avicenna</td>
<td>• Caesar</td>
<td>• Galileo</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Homer</td>
<td>• Ali</td>
<td>• Leonardo da Vinci</td>
<td>• Immanuel Kant</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Pythagoras</td>
<td>• Li Bai</td>
<td>• da Vinci</td>
<td>• Descartes</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
<tr>
<td>• Archimedes</td>
<td>• Francis of Assisi</td>
<td>• Raphael</td>
<td>• Jean d’Arc</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
<td>• Salvador Dalí</td>
</tr>
</tbody>
</table>

The dataset is restricted to the 11,340 biographies with a presence in more than 25 different languages ($L>25$). The choice of the $L>25$ threshold is guided by a combination of criteria, based on the structure of the data and the limits of manual data curation. Figure 7 shows the cumulative distribution of biographies on a log scale, as a function of the number of languages in which each of these has a presence. Most of the 997,276 total biographies
surveyed have a presence in a few languages, such that an \( L > 25 \) threshold is a high mark that can help filter the most notable of these biographies. Also, 95\% of individuals passing this threshold have an article in at least 6 of the top 10 spoken languages worldwide\(^1\), demonstrating the global reach of the individuals within the Pantheon dataset. Since there is no clean, corrected, structured data on this set of biographies, our ability to associate each individual to a cultural domain, time, and location also sets a limit on the size of the data that can be corrected by manual curation. The \( L > 25 \) threshold provides a dataset that although large (\( N = 11,340 \)), is still manageable for a team of researchers to check and curate.

\[ \text{Figure 7: Cumulative Number of Individuals with at least N Wikipedia Language Editions} \]

\(^1\) Top 10 spoken languages (by number of speakers worldwide): Chinese, English, Hindi, Spanish, Russian, Arabic, Portuguese, Bengali, French, Indonesian [53].
Finally, I emphasize that the use of Wikipedia as a source of cultural data has limitations, and the results presented here are not a general view of global culture, but represent the narrow view of cultural production that is provided by the data available in this media. Despite its limitations, Wikipedia provides a useful lens to study the development of global culture, even if this lens has optical aberrations. The measures of cultural production from these data indicate that the most notable individuals of the antiquity (prior to 500 AD) are Aristotle, Jesus, Socrates, Plato and Alexander the Great, whereas the most illustrious individuals born in the first half of the 19th century are Marx, Darwin, Picasso, Einstein, and Freud. The individuals that bubble up to the top indicate that the dataset contains information that although imperfect, is relevant. The names in the dataset are the individuals who have become known across multiple languages, geographies, and whose legacies have withstood the test of time – thus, they represent a considerable quantity of the historical cultural production of that is globally known to the world.

2.4. Data Validation

Are the metrics used within Pantheon indicative of cultural production, or are they simply measures of fame? Following an approach similar to that of Murray (2003), we use external measures of individual accomplishments to validate the dataset. Accomplishment metrics are most widely available for the domain of sports, since the achievements of individuals can be quantitatively expressed through measures such as number of championship titles won or points scored. In order to validate the Pantheon dataset as a source for measuring the impact of global cultural production, the metrics generated are characterized with respect to accepted measures of accomplishment within select domains. Formula-1 drivers, tennis players, and swimmers are used as case studies to examine the validity of the proposed metrics.

Formula One Racecar Drivers

First we examine the subset of the dataset containing the top 25 Formula-1 drivers, according to the number of languages in which they have a presence in Wikipedia. For each
of these drivers we created an additional dataset with the number of Grand Prix Wins, Championships Won, Podiums (number of times in the top 3), Starts, a dummy variable for Killed in Action, and the Time (in years) since the last F1-race. These variables are used to construct statistical models that explain the multilingual presence of each driver within Wikipedia as well as each driver’s Historical Popularity Index. Since Grand Prix Wins, Championships and Podiums are highly collinear—and hence not statistically significant when used together—only Podiums are used in the final model. Finally, the variables are used within an exponential function, since the function should always be positive (neither L nor HPI can be negative).

The first model in Figure 8a explains 87% of the variance in the number of languages in which each Formula-1 driver has a presence in the Wikipedia, showing that for Formula-1 drivers the number of languages in the Wikipedia accurately tracks accomplishments discounted by time. In contrast, when analyzing the same variables with the Historical Popularity Index, we find a model (see Figure 8b) that explains 92% of the variance in the Historical Popularity Index for each Formula-1 driver. The improved fit suggests that the corrections introduced by HPI enhances the L metric and contributes an improved characterization of accomplishment for this sample of individuals.
Figure 8: Validation using Formula 1 drivers

### Formula 1

**Part a.**

*Model*

\[ L = A e^{B_1 x_1} e^{B_2 x_2} e^{B_3 x_3} e^{B_4 T} \]

Where and

- \( x_1 = \text{Podiums} \)
- \( x_2 = \text{Starts} \)
- \( x_3 = \text{Killed in Action} \)
- \( T = \text{Time since last F1 race} \)

<table>
<thead>
<tr>
<th>( B_1 )</th>
<th>( B_2 )</th>
<th>( B_3 )</th>
<th>( B_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0031</td>
<td>0.0005</td>
<td>0.1647</td>
<td>-0.0116</td>
</tr>
</tbody>
</table>

**R\(^2\) = 87\%**

### Historical Popularity Index

**Part b.**

*Model*

\[ L = A e^{B_1 x_1} e^{B_2 x_2} e^{B_3 x_3} e^{B_4 T} \]

Where and

- \( x_1 = \text{Podiums} \)
- \( x_2 = \text{Starts} \)
- \( x_3 = \text{Killed in Action} \)
- \( T = \text{Time since last F1 race} \)

<table>
<thead>
<tr>
<th>( B_1 )</th>
<th>( B_2 )</th>
<th>( B_3 )</th>
<th>( B_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0005</td>
<td>0.0001</td>
<td>0.0146</td>
<td>0.0044</td>
</tr>
</tbody>
</table>

**R\(^2\) = 92\%**
Tennis Players

Next, a similar analysis is performed for Tennis Players. The Tennis player subset focuses on the top 45 Tennis players according to the number of languages in the Wikipedia and augmented by additional data on each individual - the number of weeks he/she spent as number one in the ATP or WTA, the number of Grand Slams wins, the top rank ever obtained, the gender (Female = 1), and the Age minus 24 (assuming 24 is peak age for tennis players).

For the number of language presences in Wikipedia ($L$), we consider the most parsimonious model—using Grand Slam Wins, Gender and Time— which explains 58% of the variance in the multilingual presence of each of these individuals in the Wikipedia (Figure 9a). This shows that once again, the number of languages in Wikipedia is a good proxy for individual accomplishments. When we considered HPI, we perform a similar analysis, but introduce a variable for the number of weeks that an individual is ranked as number 1 (as ranked by the Association of Tennis Professionals for men, and the Women’s Tennis Association, for women) instead of using Age-24. In this case, we find an improved model that explains 66% of the variation in HPI. This further supports the use of HPI as an appropriate proxy for accomplishment, since HPI tracks the degree of achievement for tennis players.
Figure 9: Validation Using Tennis Players

Model

\[ L = A e^{B_1 x_1 e^{B_2 x_2} e^{B_3 T}} \]

where:

- \( A = 88.9 \)
- \( B_1 = 0.016 \)
- \( B_2 = 0.0072 \)
- \( B_3 = 0.0543 \)

\[ x_1 = \text{Grand Slam Wins} \]
\[ x_2 = \text{Gender (Female = 1)} \]
\[ T = \text{Age} - 24 \]

a. Predicted from Grand Slam Wins, Age & Gender

b. Historical Popularity Index (MPH)

\[ L = A e^{B_1 x_1 e^{B_2 x_2} e^{B_3 x_3}} \]

where:

- \( A = 17.18 \)
- \( B_1 = 0.0043 \)
- \( B_2 = 0.0081 \)
- \( B_3 = -0.0286 \)

\[ x_1 = \text{Grand Slam Wins} \]
\[ x_2 = \text{Number of Weeks at #1} \]
\[ x_3 = \text{Gender (Female = 1)} \]
Swimmers

Finally, we perform a similar analysis considering Olympic swimmers born after 1950. In this case, the models use the total number of medals, gold medals, gender and time from last medal. In Figure 10a, we show the best parsimonious model for the data, which is obtained for gold medals, gender and time from last medal. This model explains 74% of the variance observed in the total number of languages that a swimmer has a presence in the Wikipedia, demonstrating that this measure is a good proxy for measuring accomplishment for swimmers.

When we perform the analysis for Historical Popularity Index, we find a similar result, with the model explaining 70% of the variance observed in the HPI for swimmers. Figure 10b shows the second model, which shows that HPI is also an appropriate proxy for quantifying accomplishment for swimmers.
Figure 10: Validation Using Swimmers

(a.) Swimmers born after 1950

Swimmers

\[ R^2 = 74\% \]

Model

\[ L = Ae^{B_1 x_1} e^{B_2 x_2} e^{B_3 T} \]

where

- \( x_1 \): Gold Medals
- \( x_2 \): Gender (Female = 1)
- \( T \): Time since last Olympic Medal

and

- \( A = 25.5 \)
- \( B_1 = 0.0243 \)
- \( B_2 = -0.0014 \)
- \( B_3 = -0.001 \)

(b.) Swimmers

\[ R^2 = 70\% \]

Model

\[ L = Ae^{B_1 x_1} e^{B_2 x_2} e^{B_3 T} \]

where

- \( x_1 \): Gold Medals
- \( x_2 \): Gender (Female = 1)
- \( T \): Time since last Olympic Medal

and

- \( A = 12.96 \)
- \( B_1 = 0.0059 \)
- \( B_2 = -0.0548 \)
- \( B_3 = 0.0041 \)
Discussion

The case studies on Formula One racecar drivers, tennis players, and swimmers demonstrate the validity of the measures developed within Pantheon, as compared with accepted measures of accomplishment within these specific domains. While these case studies are not exhaustive across all cultural domains, they show that the measures used and visualized within Pantheon are effective metrics for characterizing cultural production across diverse sets of domains, time, and geography. Consider a Formula One racecar driver. Certainly, for a Formula One racer the number of Grand Prix won, or Championships, would be a better metric of accomplishment than the number of languages in the Wikipedia. Yet, since Grand Prix won is a metric that applies only for Formula-1 drivers, it cannot be used for basketball players, swimmers, musicians or scientists. While imperfect, the metrics based on the online presence of characters in diverse languages represent metrics of global notability that can be used for a wide range of cultural domains, and hence, can be used to generate datasets that include people from all domains of culture.
3. Pantheon: Design & Development

The validated dataset is used as the basis for the development of a web-based data visualization engine – Pantheon. Pantheon contributes a platform for creating dynamic visualizations that inspire the exploration of global culture, allowing different audiences to explore the data, construct visual stories, and provide feedback on the data and methodologies.

3.1. Interface Design Evolution

The interface design of Pantheon has evolved through numerous iterations in terms of both graphic design and feature specifications. The initial proof of concept (Figure 11.1) was a single page web application built using Spine.js, an MVC (model-view-controller) framework for building Javascript web applications. The initial app included a dynamic treemap built with d3.js, and allowed users to filter the dataset by country and time period, and view the shares of the cultural domains exported by selected countries and years. Simple mouseover interactions on the treemap showed details on how many individuals were present in each domain, and sliders allowed users to customize the treemap parameters. A backend data API was built with Python to create a data pipeline from a SQL database to the front-end web app. This proof of concept served as inspiration for the next iteration, as the platform began to take shape as the “Observatory of Global Culture”.

Figure 11: Becoming Pantheon

Screenshots from design iterations for Pantheon, moving from initial proof-of-concept (1), design prototype (2), The Observatory of Global Culture (3), to Pantheon 1.0 (4).

The design of the application was further enhanced (Figure 11.2) through introducing a framing question for the main visualization, new dimensions and entities for the treemaps, and a new visualization – the matrix – to add a different perspective for exploring the dataset. The Observatory continued to evolve (Figure 11.3), as we experimented with the layout and color scheme. In the process, we also broadened the visualizations to include scatterplots and maps, creating an integrated set of dynamic visualizations allowing users a diverse range of lenses with which to explore the data across various dimensions. The specific visualizations were chosen to show different views of the dataset that could be linked together – for example, the matrix visualization gives an overview perspective of the
data, allowing users to explore aggregate patterns of the dataset, while the map views allow
users to explore more details at the country or domain level and the scatterplots facilitate
the comparison of countries or domains.

In addition, we integrated Charles Murray’s Human Accomplishment dataset [5] within Pantheon, such that users can visually explore the dataset of close to 4,000 significant figures in the arts and sciences. The Human Accomplishment dataset consists of distinguished figures across domains (limited to the arts and sciences), each mapped to a location and time period. Each individual is ranked using an index score created based on the level attention given to the individual in literary records, reference works, encyclopedic sources, and other authoritative records that serve as qualified sources for the subset of cultural domains surveyed. Further detail on the methods and limitations behind this dataset can be found in Human Accomplishment [5].

Figure 12: Pantheon 1.0 Main Interface Layout

Interface elements within the finalized layout of Pantheon, v1: (a) Main navigation panel (b) Visualization navigation & parameter selection (c) Visualization panel (d) Top 10 rankings and sharing links
The final design of Pantheon 1.0 (Figure 11.4) consists of a layout designed for users to explore, experience, and share dynamic views of both the Pantheon and Human Accomplishments datasets. The navigation panel (Figure 12a) at the top of the page allows users to move between the different pages of Pantheon. The left hand side of the main navigation features pages that showcase the data at different levels of granularity. The visualizations provide dynamic views with the ability to aggregate and filter the data using different parameter selections, the rankings show rankings of different entities (countries, individuals, and domains), and people page shows individual profiles for biographies within Pantheon. The right hand side of the navigation points users to information about the creation of Pantheon – the methods, vision, and team. The main page, where users explore the Visualizations, features a right panel (Figure 12b) for navigating between different visualizations and selecting data parameters, the main visualization (Figure 12c), and a right panel (Figure 12d) with the top ten individuals within the current visualization and links allowing users to share their Pantheon experience via social media.

3.2. Visualizations

At its core, Pantheon is a visualization engine that allows for dynamic, interactive visualization of the cultural production dataset. Currently, four different types of visualizations are presented within Pantheon: (1) treemaps, (2) matrices, (3) scatterplots, and (4) maps.
These visualizations focus on different dimensions of the data and map to each of the following questions and entities:

<table>
<thead>
<tr>
<th>Question</th>
<th>Output Entities</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who are the globally known people born in present-day [selected country]?</td>
<td>Domains, People</td>
<td>Treemap</td>
</tr>
<tr>
<td>Where were globally known individuals in [selected domain] born?</td>
<td>Countries, People</td>
<td>Treemap, Map</td>
</tr>
<tr>
<td>How many globally known people are associated with each place of birth and domain?</td>
<td>Domains, Countries, People</td>
<td>Matrix</td>
</tr>
<tr>
<td>How do [selected country1] and [selected country2] compare in terms of globally known people?</td>
<td>Domains, People</td>
<td>Scatterplot</td>
</tr>
<tr>
<td>What places have produced globally known people in [selected domain1] and [selected domain2]?</td>
<td>Countries, People</td>
<td>Scatterplot</td>
</tr>
</tbody>
</table>
**Treemaps**

*Figure 14: Treemap Examples - by Place of Birth, by Domain*

On the left, the treemap visualizes the individuals born within the present day boundaries of the United Kingdom. Colors are mapped to top-level domains. On the right, a treemap of the countries producing Physicists is shown, colors are mapped based on continents.

Treemaps are spatially compact visualizations that were first introduced by Schneiderman in the early 1990s to map the file structure and file size of hard disks [38]. This type of visualization is effective for representing datasets with hierarchical structure. A treemap consists of a set of nested rectangles that encode information by color and area to visualize the proportional shares of a whole. Given the structure of the Pantheon dataset, the treemap is a useful view for visualizing the data aggregated by domain or country. The treemap by place of birth uses color to encode domains, and area to encode the share of each domain within a selected country’s cultural footprint – for example, Figure 14 includes a treemap showing the cultural domains of the globally known individuals born in the United Kingdom. The treemap by domain shows the countries that produce a selected cultural domain, coloring countries by continent, and encoding the share that each country represents by area. Figure 14 includes a treemap showing the shares of each birthplace of globally known physicists. Mouseover interactions on each sub-rectangle of the treemap shows a tooltip with the top five individuals (ranked by HPI) within the relevant country and domain, and a more detailed ranking table of all individuals is shown if the user clicks on a specific rectangle. Treemaps also link to related visualizations by either domain or country.
Matrices

Matrix visualization allows users to explore the underlying structure of large, high-dimensional datasets by simultaneously encoding the associations and interactions between many variables. Within a matrix, position, color and intensity are the visual variables used to encode information about the data [39]. In Pantheon, the rows of the matrix show the countries, and the columns show cultural domains. Figure 15 shows the matrix visualization of the full Pantheon dataset, from 4000BC through 2010, for the top 40 countries by diversity. The sort order of the matrix can be updated interactively, to order the columns and rows either by the count of individuals, or by a simple sorting on the name of the variable. The color encodes the presence or absence of different cultural domains by country, while the intensity of the color encodes the number of individuals within a specific country and domain. Each cell of the matrix reveals a tooltip on mouseover, with the top 5 individuals attributed to a specific country and domain. Users can also link to both the treemap and map visualizations for further detail by clicking a specific cell to access additional relevant views.

Figure 15: Matrix Visualization of World Cultural Production, 4000BC - 2010

The matrix visualization provides an overview of the dataset by visualizing all domains and countries for a specified time period.
Scatterplots

Scatterplots are two-dimensional, bivariate visualizations often used to show the correlation between variables or for regression analysis [40]. Within Pantheon, the scatterplots are used to facilitate the comparison of cultural domains or places of birth. Color is used to encode either country or domain information, whereas the position of the point on the plot indicates how many individuals are produced in the selected countries or domains. A reference line where x=y is provided, such that points above or below the line respectively indicate whether the entity on the y-axis or the entity on the x-axis produce more individuals within the selected parameters. Users can also adjust the visualization view by choosing whether to apply a linear or logarithmic scale to the axes, and whether to mirror the axes. Each point on the scatterplot also shows a tooltip with relevant individuals on hover. For example, Figure 16 shows (a) a comparison of Germany and the United Kingdom, and (b) a log-log scatterplot comparing Explorers and Chemists.
Figure 16: Scatterplots in Pantheon

![Scatterplot of Germany vs United Kingdom](image)

**How do Germany and United Kingdom compare in terms of number of globally known people? (4000 B.C. – 2010)**

**What places of birth have produced globally known people in Explorer and Chemist? (4000 B.C. – 2010)**

Examples of scatterplots showing country/country and domain/domain comparisons. Above, Germany and the United Kingdom are compared on a linear scale. Below, countries producing Explorers and Chemists are plotted on a log-log scale.

**Maps**

Thematic maps show the geographic dispersion of a dataset. Since all of the individuals within the dataset are associated with countries based on their location of birth and the present-day political boundaries, the map visualization shows the number of individuals within each country, encoded by color hue and intensity. The visualization is updated as users filter the data based on time period and cultural domain. Figure 17 shows different views within the Pantheon map, using the sports domain as an example. The map visualization presents a global view of the dataset, and allows users to detect geographic...
patterns of cultural domains – however, since the country areas are not uniform, a common criticism of this type of visualization is the visual emphasis of countries with larger land area. To partially remedy this in Pantheon, the map also includes a zoom feature to allow users to zoom and pan to specific regions of interest. Further detail for each country is provided on mouseover, and users can continue exploring the dataset with the relevant links to other visualizations.

**Figure 17: Maps in Pantheon**

*Map visualizations within Pantheon – the first map shows the map of the global production of individuals in Sports across all time, the second shows a zoomed view of Europe, with detail on the individuals produced from Belgium.*
3.3. Other Pages and Features

**Rankings**

To provide users with a table view of the dataset, Pantheon also includes a rankings page where users can access lists of entities (countries, people, or domains) and relevant statistics. Figure 18 shows the layout of the rankings page within Pantheon. Similar to the visualizations, users are able to filter the data that is shown by customizing a set of parameters such as time period, place of birth, and cultural domain. The statistics shown for places of birth include the total number of individuals, the percent of individuals that are female, the diversity in terms of number of unique occupations, a measure of the total number of individuals with at least 50 different language editions (i50), an H-index measure (similar to the measure of academic productivity, a country has index h if h individuals from that country have a presence in at least h language editions of Wikipedia), as well as HCPI—the historical cultural production index (aggregate HPI individuals). At the individual level, the attributes listed include their demographic details (gender, place of birth, occupation, birthyear), the metrics of impact (L, L* and HPI), and the relevant aggregated page view statistics (total, English and non-English, standard deviation). For domains, the table provides aggregated measures including total individuals and total number of countries. The table also includes a search feature for users interested in specific data items, and is sortable by any column variable.
Figure 18: Rankings within Pantheon

The rankings page within Pantheon, showing rankings of places of birth. Rankings are also supported for people and cultural domains.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Place of Birth*</th>
<th>Number of People</th>
<th>% Women</th>
<th>Diversity</th>
<th>ISD</th>
<th>H Index</th>
<th>HCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>2,344</td>
<td>77.0</td>
<td>388.0</td>
<td>85</td>
<td>426.733</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>United Kingdom</td>
<td>1,234</td>
<td>62.0</td>
<td>245.6</td>
<td>69</td>
<td>243.323</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>France</td>
<td>1,057</td>
<td>58.0</td>
<td>192.0</td>
<td>68</td>
<td>226.910</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Italy</td>
<td>782</td>
<td>48.0</td>
<td>240.0</td>
<td>66</td>
<td>262.704</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Germany</td>
<td>730</td>
<td>60.0</td>
<td>136.0</td>
<td>62</td>
<td>270.935</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Russia</td>
<td>534</td>
<td>44.0</td>
<td>81.0</td>
<td>58</td>
<td>212.687</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Japan</td>
<td>203</td>
<td>36.0</td>
<td>70.0</td>
<td>53</td>
<td>388.611</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Turkey</td>
<td>203</td>
<td>60.0</td>
<td>38.0</td>
<td>48</td>
<td>490.536</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Poland</td>
<td>198</td>
<td>36.0</td>
<td>38.0</td>
<td>45</td>
<td>387.222</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Austria</td>
<td>142</td>
<td>35.0</td>
<td>33.0</td>
<td>46</td>
<td>345.083</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>The Netherlands</td>
<td>126</td>
<td>54.0</td>
<td>35.0</td>
<td>45</td>
<td>308.713</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Greece</td>
<td>120</td>
<td>24.0</td>
<td>42.0</td>
<td>47</td>
<td>227.337</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Japan</td>
<td>117</td>
<td>31.0</td>
<td>24.0</td>
<td>41</td>
<td>355.035</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Sweden</td>
<td>114</td>
<td>30.0</td>
<td>25.0</td>
<td>42</td>
<td>266.118</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>India</td>
<td>118</td>
<td>28.0</td>
<td>30.0</td>
<td>43</td>
<td>271.434</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Brazil</td>
<td>112</td>
<td>19.0</td>
<td>19.0</td>
<td>40</td>
<td>267.311</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Canada</td>
<td>118</td>
<td>23.0</td>
<td>12.0</td>
<td>40</td>
<td>241.124</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>China</td>
<td>99</td>
<td>24.0</td>
<td>27.0</td>
<td>40</td>
<td>290.037</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Belgium</td>
<td>105</td>
<td>29.0</td>
<td>26.0</td>
<td>40</td>
<td>270.749</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Czech Republic</td>
<td>98</td>
<td>20.0</td>
<td>19.0</td>
<td>40</td>
<td>220.560</td>
<td></td>
</tr>
</tbody>
</table>
People profiles

Pantheon features individual profile pages for the biographies that are included within the dataset. The profile shows an automatically generated summary of the individual’s biography, a relevant picture (when available), and a listing of the relevant metrics and statistics for the selected individual. To allow users to further explore, the individual is also ranked among his or her contemporaries based on cultural domain, time period, and place of birth. To view other people, users can either “shuffle” the biographies to see a randomly selected individual, or use the search bar to select a particular biography to view. For example, Figure 19 below shows Michael Faraday's profile page within Pantheon, and his ranking amongst his contemporaries as a physicist born in the United Kingdom in the late 18th century.

Figure 19: Michael Faraday's Pantheon Profile
Tutorials

Past research on human-computer interaction has demonstrated the effectiveness of using communicative agents in the context of educating users [41], so in order to introduce the platform to new users, Pantheon incorporates a story-telling feature using guided tutorials. Through the homepage, users can select different tutorials led by a character persona that guide viewers by stepping through a series of views and pages to present a short vignette. Currently, three different vignettes are available to help users learn about the Renaissance, Exploration, as well as Pantheon as a visualization engine. Figure 20 shows the beginning of the vignette led by a character named Nora, who guides the user through the entire Pantheon site to introduce the various visualizations and pages that are available. Each vignette provides a guided, linear path through Pantheon to acclimate users to the different features of the site, while also presenting the user with a narrative that connects through various visualizations. Each tutorial highlights various areas of Pantheon as the user steps through the story. At anytime during the tutorial, the user may choose to exit in order to start their own exploration of Pantheon.
Figure 20: Nora’s Tutorial

Steps 1-3 of Nora’s Tutorial, guiding new users into Pantheon
Mobile layout

Given the increasing trend in mobile web traffic [42], a lightweight mobile browser version of Pantheon was implemented prior to the official launch to accommodate users accessing the platform through mobile devices. The mobile version introduces different interactions for navigation through the core pages and features of the site, such as the side panel for the main navigation, and touch events for all visualizations. The mobile version also excludes features that are better suited for desktop browsing – including the tutorials and large inset tooltips. Figure 21 below shows the various screens within the mobile version of Pantheon.

Figure 21: Pantheon Mobile

The mobile browser experience within Pantheon – the different screens show the homepage, navigation panel, visualization view, and rankings table view.

3.4. Technical Details

Database

Pantheon’s data is stored using MongoDB, an open source, document-based database system. Unlike relational databases, MongoDB stores data as documents in collections within each database, where each document is a JSON-style data structure composed of field-value pairs. MongoDB queries generate JSON objects by default, making it well suited
to power web applications built with Javascript. Also, it is currently the database that is supported by the Meteor web framework (see next section), and has numerous performance advantages over traditional SQL databases. All of Pantheon’s data is contained within three collections, based on the different entities that are used within the interface – people, countries, and domains. The document structure of each person, country and domain is provided in the Appendix.

**Application Framework**

Pantheon is built on a next-generation web development framework called meteor.js. Meteor is an open-source Javascript web framework that introduces innovative features for building data-intensive applications. The framework focuses on seven core principles [43], which are detailed in the Appendix. Meteor is a natural fit for the development of Pantheon, allowing for rapid prototyping and integration of existing open-source packages and libraries, and providing a seamless environment that synchronizes data between server and client without the need to develop standalone data APIs or manual callbacks to track data changes or updates. Figure 22 illustrates how different components of an application built with Meteor interact.

**Figure 22: Meteor.js Framework**
For Pantheon, Meteor is used on the server side to output a standalone application bundle, which includes all the source code, files and packages needed to run Pantheon. Node.js, a software platform for running scalable, server-side Javascript applications, is used to run the bundled Pantheon application, which connects to a MongoDB database that serves the data for the application. On the front-end (client side), users interact with the app through the browser, and the client-side Javascript handles and renders the relevant page templates (in a templating language called Spacebars, similar to Handlebars.js), based on user actions such as updating parameters to change the visualizations, or going to a specific URL address that routes to a particular template. The data on the client side is held in minimongo, Meteor’s client-side Mongo emulator, which caches data on the browser and updates through a subscription mechanism that proactively syncs the data between client and server based on changes from either the data on the server, or from the user on the client. For example, the data on the client is updated when the user changes the view or page on Pantheon, or when new data updates are pushed to the MongoDB on the server. This caching technique contributes to the latency compensation feature of Meteor. Finally, the client side also integrates a number of open-source Javascript libraries including d3plus (for visualizations), DataTables (for rendering ranking tables), and MathJax (for rendering mathematical notation). Taken together, these components encompass the entirety of the Pantheon application.
4. Launch & Usage

Pantheon was officially launched on March 14\textsuperscript{th}, 2014 and has since received substantial attention by the popular press, and within the general public. As of April 22\textsuperscript{nd}, 2014 Pantheon has received a total of 189,216 pageviews from a total of 145,604 users from 191 different countries across the globe (see Figures 23-24). The site activity has been punctuated by periods of high traffic, flowing from international media sources and social media activity. The launch of Pantheon coincided with coverage in New York Times Magazine [45], which resulted in the initial traffic spike and helped the platform gain attention from other news media worldwide. The second spike in traffic resulted from coverage from Fast Company on March 25, 2014 [46].

![Figure 23: Pantheon Page Views, March - April 2014](image)

4.1. Global Impressions

The impact of the site has drawn strong feedback from Pantheon users worldwide. Many users helped improve the quality of the dataset by suggesting corrections to discrepancies within the data, or by helping to fill in unknown variables for individuals with missing data. Also, many users found the use of birthplace to be controversial and non-intuitive for assigning individuals to a particular geography. Since cultural accomplishment is strongly associated with national identity, some users interpreted birthplace to proxy for nationality – thus, the site was patched with messages highlighting the caveats behind our methodology for the location assignment of individuals.
In terms of the pages that received the most traffic in Pantheon (see Figure 25), the rankings page was the second most viewed page, followed closely by the treemap visualization. This somewhat surprising result was likely driven by the media attention that focused on the rankings of individuals within the dataset, instead of the significance of the aggregate patterns revealed through the visualizations. Also, based on the traffic statistics, the treemaps were by far the most popular visualization that users engaged with. We found that viewers’ reactions were markedly different depending on the page that they engaged with the most. Many users who were drawn to the individual rankings of people within the dataset focused mainly on the differences in ranking between people. Others, who explored the visualizations, used Pantheon to explore and share interesting aggregate trends they noticed in the dataset. Overall, Pantheon has sparked numerous discussions and debates on cultural accomplishment and perceptions of global notability.
<table>
<thead>
<tr>
<th>Rank</th>
<th>URL Path</th>
<th>Views</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/</td>
<td>84,546</td>
<td>44.68%</td>
</tr>
<tr>
<td>2</td>
<td>/rankings/</td>
<td>41,054</td>
<td>21.70%</td>
</tr>
<tr>
<td>3</td>
<td>/treemap/</td>
<td>36,820</td>
<td>19.46%</td>
</tr>
<tr>
<td>4</td>
<td>/people/</td>
<td>18,464</td>
<td>9.77%</td>
</tr>
<tr>
<td>5</td>
<td>/methods</td>
<td>4,773</td>
<td>2.52%</td>
</tr>
<tr>
<td>6</td>
<td>/matrix/</td>
<td>1,167</td>
<td>0.62%</td>
</tr>
<tr>
<td>7</td>
<td>/map/</td>
<td>1,143</td>
<td>0.60%</td>
</tr>
<tr>
<td>8</td>
<td>/scatterplot/</td>
<td>463</td>
<td>0.24%</td>
</tr>
<tr>
<td>9</td>
<td>/vision</td>
<td>311</td>
<td>0.16%</td>
</tr>
<tr>
<td>10</td>
<td>/team</td>
<td>216</td>
<td>0.11%</td>
</tr>
</tbody>
</table>
5. Analysis & Implications

Unlike other cultural datasets, the Pantheon data connects each biography with its date of birth, place of birth, cultural domain and notability. This allows us to go beyond the individual and explore the patterns of cultural production of the places associated with each cultural domain, and the dynamics of the cultural domains expressed for each time period. The aggregate patterns that emerge within the data demonstrate significant relationships between the ubiquity of cultural domains, the diversity of cultural production between countries, and the volume of known individuals associated with each place of birth. These significant relationships imply that cultural development is a structured process, and suggest that the evolution of a country’s cultural production is highly path dependent.

5.1. The Nestedness of Cultural Production

Pantheon’s matrix visualization shows an overview of the dataset by allowing users to view all data for all cultural domains and countries, aggregated across time. Figure 26 shows the pattern that emerges when sorting the view on the diversity of cultural production of countries and the ubiquity of the cultural domains (at the second level of aggregation, industries). The pattern that the user observes within Pantheon is referred to as “nestedness”, a phenomenon that has been previously studied in the context of ecology [47, 48] and economic development [49].
Figure 26: Matrix View on Pantheon

The complete matrix view within Pantheon, for all countries, all domains, and all time periods.
A deeper analysis reveals the robustness and significance of this nestedness pattern, across different levels of cultural domain classifications. The figures below represent the presence-absence matrices for countries and cultural domains – Figure 27 shows occupations (most disaggregate category) sorted by ubiquity and countries sorted by diversity, Figure 28 shows cultural industries (grouped occupations) and countries sorted by diversity. A cultural domain is present if a country produces at least one known individual in the specified category with more than 25 language editions in Wikipedia. The lines indicate the diversity (in gold) of each place of birth, and the ubiquity (in red) of each cultural domain – they serve as a visual guide to show where the presences of the matrix would fall in the case of perfect nestedness.

The matrices shown for both occupations and industries are highly nested. This means that the less frequent cultural domains, such as invention, are expressed only in the countries with an already diverse cultural footprint. We quantify the nestedness of the matrix using a measure called Nested Overlap Decreasing Fill [50, 51]. For this measure, NODF = 100 signifies perfect nestedness, while NODF = 0 signifies no nestedness. In the matrices shown in Figures 27-28, I find that NODF = 65.13 for occupations and NODF = 71.66 for industries. This is a highly statistically significant relationship – I test the nestedness of both country-occupation and country-industry matrices with two null models - ER: where presences are randomized within the matrix, and CE: where the probability of a presence in a particular cell in the matrix is the average of the probabilities of finding it in the particular row and column from the original matrix [48]. The nestedness of both country-occupation and country-industry matrices is significant at the 0.01 level, compared to both null models. This phenomena is analogous to the nestedness observed in industry-location matrices in the study of economic development [49], and that has also been observed in insular habitats [47] and mutualistic networks [48] in ecology. The nestedness observed has implications for understanding and predicting the trajectories of the cultural production of countries, by demonstrating that cultural production evolves through a structured process, where locations diversify from the most ubiquitous to least ubiquitous domains.
Figure 27: Nestedness of Cultural Production - Occupations
5.2. Diversity of Cultural Production

In addition to examining the presence and absence of cultural domains, we can also analyze the correlation between the number of individuals born within a particular country, and the diversity (number of different domains) that emerge from that location. The relationship between the diversity of countries (as places of birth) and the total volume of notable individuals produced is revealed by plotting countries by the number of total individuals that emerge and the diversity of the domains that are produced. Figure 29 shows this plot on a log-log scale, and a regression of the log-transformed diversity and population numbers implies that a 1% increase in individuals increases the expected diversity by 3.8%, a 10% increase in individuals increases the expected diversity by 21%. The relationship suggests a positive correlation between the diversity of a location and the number of individuals produced. Moreover, it shows that highly diverse locations will produce exponentially more in individuals than less diverse countries. This implies that the process of diversification requires an accumulation of cultural production that is nonlinear, such that the number of individuals needed attain higher levels of diversity in cultural production increases as the cultural production of a country becomes more diverse.
Figure 29: Scatterplot of Countries (of birth) by Number of Characters and Diversity

\[ \log(y) \sim \log(x)^b \]
\[ y \sim e^{\log(x)^{0.71}} \]

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| b        | 0.70771    | 0.01352 | 52.34    | 0.0000    |

Residual standard error: 0.4761 on 196 degrees of freedom
5.3. Dynamics of Global Cultural Production

The Pantheon dataset reveals the nestedness of cultural production and the relationship between the diversity of cultural production and the volume of known individuals associated with a location. Taken together, these results demonstrate that cultural production follows structured patterns - where the diversity of a location’s global cultural footprint is closely tied to its size (in terms of number of individuals), and the footprints of locations are nested, such that the less diverse locations produce almost perfect subsets of the set of domains produced by highly diverse locations. These patterns suggest that cultural development is path dependent, and that the accumulation of cultural production is exponential in relation to the diversity of outputs, such that the expansion of cultural production involves underlying challenges in both diversifying into less ubiquitous domains and building an accumulation of cultural outputs. Further investigation will reveal and confirm the dynamics that give rise to these structural patterns.
6. Conclusion

This thesis presents Pantheon, which incorporates both a new dataset for quantifying global cultural production, and a visualization platform for exploring cultural datasets to uncover insights about the structure and dynamics of cultural production. The dataset is validated using external measures of accomplishment and impact in various cultural domains, and the patterns revealed through visualization are analyzed for statistical significance. Moreover, Pantheon has also had global impact and reach, engaging users worldwide in the exploration of historical cultural production and sparking discourse on the measurement of cultural accomplishment. While there are certainly limitations to the data and methods, Pantheon contributes a quantitative view of culture that reveals insights into the process of global cultural production, and opens the dialogue for creating indicators that can facilitate the quantitative study of historical cultural patterns. Pantheon itself is an incomplete resource, however, it offers the opportunity for further exploration of our cultural legacies, and our human potential.

6.1. Future Directions

Pantheon’s dataset and visualizations open a number of considerable opportunities for future development. First, there are a number of data enhancements to consider for improving upon the existing dataset. For example, improving the geographic granularity of the data by cleaning the data to the city level, augmenting the dataset with additional fields such as date and place of death, and expanding the dataset by adjusting the overall threshold of \( L > 25 \) using improved, automated methods for data cleaning and matching. There is also the challenge of accounting for historical geography and attributing individuals with contributions in multiple domains. Pantheon, as a visualization tool, can be expanded with additional visualizations that add more dimensionality to the views that users can apply to the data. Finally, from a broader perspective, we can consider future efforts for expanding this quantitative approach for characterizing and exploring historical cultural patterns. Potential avenues for future research include connecting measures of cultural accomplishment with other indicators of societal progress, exploring the underlying
dynamics connecting cultural production with economic development, and quantifying the patterns of cultural production with the evolution of communication media. By building on the approach introduced by Pantheon, we can build quantitative stories to augment the historical narratives of our global cultural legacies.
### 7.1. Cultural Domain Classifications

<table>
<thead>
<tr>
<th>Domain</th>
<th>Industry</th>
<th>Occupation</th>
<th># Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>Dance</td>
<td>Dancer</td>
<td>12</td>
</tr>
<tr>
<td>Arts</td>
<td>Design</td>
<td>Architect</td>
<td>73</td>
</tr>
<tr>
<td>Arts</td>
<td>Design</td>
<td>Comic Artist</td>
<td>24</td>
</tr>
<tr>
<td>Arts</td>
<td>Design</td>
<td>Designer</td>
<td>16</td>
</tr>
<tr>
<td>Arts</td>
<td>Design</td>
<td>Fashion Designer</td>
<td>10</td>
</tr>
<tr>
<td>Arts</td>
<td>Design</td>
<td>Game Designer</td>
<td>4</td>
</tr>
<tr>
<td>Arts</td>
<td>Film And Theatre</td>
<td>Actor</td>
<td>1193</td>
</tr>
<tr>
<td>Arts</td>
<td>Film And Theatre</td>
<td>Comedian</td>
<td>4</td>
</tr>
<tr>
<td>Arts</td>
<td>Film And Theatre</td>
<td>Film Director</td>
<td>177</td>
</tr>
<tr>
<td>Arts</td>
<td>Fine Arts</td>
<td>Artist</td>
<td>88</td>
</tr>
<tr>
<td>Arts</td>
<td>Fine Arts</td>
<td>Painter</td>
<td>178</td>
</tr>
<tr>
<td>Arts</td>
<td>Fine Arts</td>
<td>Photographer</td>
<td>12</td>
</tr>
<tr>
<td>Arts</td>
<td>Fine Arts</td>
<td>Sculptor</td>
<td>21</td>
</tr>
<tr>
<td>Arts</td>
<td>Music</td>
<td>Composer</td>
<td>225</td>
</tr>
<tr>
<td>Arts</td>
<td>Music</td>
<td>Conductor</td>
<td>11</td>
</tr>
<tr>
<td>Arts</td>
<td>Music</td>
<td>Musician</td>
<td>381</td>
</tr>
<tr>
<td>Arts</td>
<td>Music</td>
<td>Singer</td>
<td>437</td>
</tr>
<tr>
<td>Business &amp; Law</td>
<td>Business</td>
<td>Businessperson</td>
<td>79</td>
</tr>
<tr>
<td>Business &amp; Law</td>
<td>Business</td>
<td>Producer</td>
<td>12</td>
</tr>
<tr>
<td>Business &amp; Law</td>
<td>Law</td>
<td>Lawyer</td>
<td>17</td>
</tr>
<tr>
<td>Exploration</td>
<td>Explorers</td>
<td>Astronaut</td>
<td>32</td>
</tr>
<tr>
<td>Exploration</td>
<td>Explorers</td>
<td>Explorer</td>
<td>70</td>
</tr>
<tr>
<td>Humanities</td>
<td>History</td>
<td>Historian</td>
<td>48</td>
</tr>
<tr>
<td>Humanities</td>
<td>Language</td>
<td>Critic</td>
<td>5</td>
</tr>
<tr>
<td>Humanities</td>
<td>Language</td>
<td>Journalist</td>
<td>19</td>
</tr>
<tr>
<td>Humanities</td>
<td>Language</td>
<td>Linguist</td>
<td>21</td>
</tr>
<tr>
<td>Humanities</td>
<td>Language</td>
<td>Writer</td>
<td>955</td>
</tr>
<tr>
<td>Humanities</td>
<td>Philosophy</td>
<td>Philosopher</td>
<td>281</td>
</tr>
<tr>
<td>Institutions</td>
<td>Government</td>
<td>Diplomat</td>
<td>36</td>
</tr>
<tr>
<td>Institutions</td>
<td>Government</td>
<td>Judge</td>
<td>9</td>
</tr>
<tr>
<td>Institutions</td>
<td>Government</td>
<td>Nobleman</td>
<td>116</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>Institutions</td>
<td>Government</td>
<td>Politician</td>
<td>2529</td>
</tr>
<tr>
<td>Institutions</td>
<td>Government</td>
<td>Public Worker</td>
<td>14</td>
</tr>
<tr>
<td>Institutions</td>
<td>Military</td>
<td>Military Personnel</td>
<td>223</td>
</tr>
<tr>
<td>Institutions</td>
<td>Military</td>
<td>Pilot</td>
<td>9</td>
</tr>
<tr>
<td>Institutions</td>
<td>Religion</td>
<td>Religious Figure</td>
<td>518</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Activism</td>
<td>Social Activist</td>
<td>114</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Companions</td>
<td>Companion</td>
<td>101</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Celebrity</td>
<td>21</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Chef</td>
<td>2</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Magician</td>
<td>4</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Model</td>
<td>30</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Pornographic Actor</td>
<td>11</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Media Personality</td>
<td>Presenter</td>
<td>19</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Outlaws</td>
<td>Extremist</td>
<td>34</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Outlaws</td>
<td>Mafioso</td>
<td>13</td>
</tr>
<tr>
<td>Public Figure</td>
<td>Outlaws</td>
<td>Pirate</td>
<td>9</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Computer Science</td>
<td>Computer Scientist</td>
<td>34</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Engineering</td>
<td>Engineer</td>
<td>41</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Invention</td>
<td>Inventor</td>
<td>67</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Math</td>
<td>Mathematician</td>
<td>157</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Math</td>
<td>Statistician</td>
<td>4</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Medicine</td>
<td>Physician</td>
<td>143</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Archaeologist</td>
<td>13</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Astronomer</td>
<td>83</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Biologist</td>
<td>141</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Chemist</td>
<td>220</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Geologist</td>
<td>10</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Natural Sciences</td>
<td>Physicist</td>
<td>268</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Anthropologist</td>
<td>11</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Economist</td>
<td>102</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Geographer</td>
<td>14</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Political Scientist</td>
<td>7</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Psychologist</td>
<td>38</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>Social Sciences</td>
<td>Sociologist</td>
<td>15</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Athlete</td>
<td>74</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
<td>-----------</td>
<td>----</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Boxer</td>
<td>14</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Chessmaster</td>
<td>30</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Cyclist</td>
<td>29</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Golfer</td>
<td>2</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Gymnast</td>
<td>7</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Martial Arts</td>
<td>7</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Mountaineer</td>
<td>5</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Racecar Driver</td>
<td>104</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Skater</td>
<td>9</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Skier</td>
<td>17</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Snooker</td>
<td>3</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Swimmer</td>
<td>20</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Tennis Player</td>
<td>161</td>
</tr>
<tr>
<td>Sports</td>
<td>Individual Sports</td>
<td>Wrestler</td>
<td>44</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>American Football Player</td>
<td>1</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Baseball Player</td>
<td>5</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Basketball Player</td>
<td>71</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Coach</td>
<td>75</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Cricketer</td>
<td>2</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Hockey Player</td>
<td>2</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Referee</td>
<td>10</td>
</tr>
<tr>
<td>Sports</td>
<td>Team Sports</td>
<td>Soccer Player</td>
<td>1064</td>
</tr>
</tbody>
</table>
7.2. Historical Popularity Index Derivation

For each biography $i$, we define:

$L_i = \text{Number of different languages editions of Wikipedia for biography } i$

$L^*_i = \text{Effective number of language editions for biography } i$

$L^*_i = \exp(H_i)$

where $H_i$ is the entropy in terms of Page Views

$$H_i = \sum_j \sum_i v_{ij}^* \ln \left( \frac{v_{ij}}{\sum_i v_{ij}} \right)$$

and $v_{ij} = \text{total page views of individual } i \text{ in language } j$

$A_i = 2013 - \text{Year of Birth}$

$CV = \text{Coefficient of variation in page views}$

$$CV_i = \frac{\sigma_i}{\mu_i}$$

$\sigma_i = \text{standard deviation in pageviews across all languages}$

$\mu_i = \text{average monthly pageviews}$

$v^{NE} = \text{total pageviews in non-English editions of Wikipedia}$

Using the above, the Historical Popularity Index (HPI) of an individual, $i$, is defined as:

$$HPI = \begin{cases} 
\ln(L) + \ln(L^*) + \log_4(A) + \ln(v^{NE}) - \ln(CV) & \text{if } A \geq 70 \\
\ln(L) + \ln(L^*) + \log_4(A) + \ln(v^{NE}) - \ln(CV) - \frac{70-A}{7} & \text{if } A < 70 
\end{cases}$$
7.3. Pantheon MongoDB Document Structure

Sample Person document:

```
{
    "_id" : ObjectId("53508ed2a845d7d4b60c59a9"),
    "en_curid" : 307,
    "name" : "Abraham Lincoln",
    "numlangs" : 131,
    "countryCode" : "US",
    "countryCode3" : "USA",
    "countryName" : "United States",
    "continentName" : "North America",
    "birthyear" : 1809,
    "birthcity" : "Hodgenville",
    "gender" : "Male",
    "occupation" : "POLITICIAN",
    "industry" : "GOVERNMENT",
    "domain" : "INSTITUTIONS",
    "dataset" : "pantheon",
    "TotalPageViews" : NumberLong(66145211),
    "L_star" : 5.801386687,
    "StdDevPageViews" : 586914.722,
    "PageViewsEnglish" : NumberLong(41477236),
    "PageViewsNonEnglish" : NumberLong(24667975),
    "AverageViews" : 504925.2748,
    "HPI" : 27.93858549
}
```

Sample Country document:

```
{
    "_id" : ObjectId("53508ed5a845d7d4b60c8703"),
    "countryCode" : "US",
    "countryCode3" : "USA",
    "countryName" : "United States",
    "continentName" : "North America",
    "dataset" : "pantheon"
}
```

Sample Domain document:

```
{
    "_id" : ObjectId("53508ed5a845d7d4b60c85f6"),
    "occupation" : "DANCER",
    "industry" : "DANCE",
    "domain" : "ARTS",
    "dataset" : "pantheon"
}
```
7.4. Seven Principles of Meteor

- Data on the wire – Data is sent over the network instead of HTML, the client renders the app using templates.
- One language – Javascript is used both on the client and server. Data is all in one format – JSON.
- Database everywhere – Provides an API to make the application database accessible from both client and server.
- Latency compensation – Simulates zero-latency between client and server
- Full stack reactivity – Data changes are seamlessly synchronized across all components of the application stack
- Embrace the ecosystem- integrated with the open source community, hundreds of self-contained, community-written and supported packages are available for Meteor developers to use and extend
- Simplicity equals productivity – Clean and simple APIs
8. References


https://www.sciencenews.org/article/florence-nightingale-passionate-statistician

http://understandinguncertainty.org/coxcombs

http://chimera.labs.oreilly.com/books/1230000000345


http://www.bloomberg.com/visual-data

[27] Alex Simoes. The Observatory of Economic Complexity. [Online].
http://atlas.media.mit.edu/


http://www.whoisbigger.com/


[41] Patrick Doyle, "When is a Communicative Agent a Good Idea?," Knowledge Systems Laboratory, Stanford University, Stanford, CA.


http://www.nytimes.com/2014/03/16/magazine/whos-more-famous-than-jesus.html


http://meta.wikimedia.org/wiki/Top_Ten_Wikipedias