Pantheon: Visualizing Historical Cultural Production

Amy (Zhao) Yu*, Kevin Zeng Hu†, Deepak Jagdish‡, Cesar A. Hidalgo§
MIT Media Lab, Massachusetts Institute of Technology

ABSTRACT

We introduce Pantheon, a dataset and visualization platform quantifying cultural accomplishments that have broken the barriers of space, time and language. The Pantheon dataset connects the 11,340 biographies available in more than 25 languages in Wikipedia with a cultural domain, place of birth, and time period. We present this data through an online data visualization platform supporting the exploration of the dataset. In this poster we describe the Pantheon dataset and visualization platform, both of which are available at http://pantheon.media.mit.edu.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; J.4 [Computer Applications] Social and Behavioral Sciences—Sociology

1 INTRODUCTION & RELATED WORK

For centuries, scientists and data visualization artists have attempted to quantify culture. More than two centuries ago Joseph Priestley created detailed timelines of history and biographies. More recent efforts, however, have been powered by the emergence of computational tools and the availability of digital records and printed books. These new efforts have breathed new life to efforts looking to quantify and visualize culture.

Some recent efforts have built on the use of biographies as a proxy for culture. Others have relied on the use of natural language and have focused on the evolution of words. An example of the former is Charles Murray's work on *Human Accomplishments*. In *Human Accomplishments*, Murray contributed an inventory of 4,002 significant individuals from the arts and sciences [1]. Another example is *Who's Bigger?*, a publication that ranked individuals and entities based on their popularity in the English Wikipedia [2]. Examples of the latter involve the use of digitized texts to study cultural trends from a linguistic perspective [3].

The development of these datasets has also been accompanied by efforts to visualize the data. An example here is the *Google Ngram Viewer* [4], which provides a customizable interactive line graph that can be used to visualize the popularity of specific phrases over time. The *Who is Bigger* website [5] also provides a similar visualization capacity. Yet, despite these efforts, there is plenty of room to improve our ability to visually explore patterns of historical cultural production.

Pantheon contributes a new dataset and visualization engine to this field. The Pantheon dataset has the advantage of being based on multilingual expressions extracted from more than 200 Wikipedia language editions, helping identify individuals that are globally rather than locally famous. Moreover, the Pantheon data links each biography to its place of birth, date of birth, and cultural domain (such as fine arts, singers, or astronauts). The latter is a novel advancement that opens the possibility to create richer visualizations exploring the evolution of cultural domains over space and time. In this poster we describe the creation of the Pantheon dataset and visualization engine.

2 DATA METHODOLOGY

2.1 Pantheon Dataset

2.1.1 Collection

Pantheon contributes a new comprehensive dataset of cultural production across geographies, time periods, and languages. We created the Pantheon dataset using Freebase and Wikipedia, both open-source, collaborative, multi-lingual knowledge bases freely available online. In Pantheon we use globally known individuals as a proxy for cultural production. To identify famous historical figures, we first collected a list of individuals and their demographic details through Freebase's database of all entities classified as Persons [6]. Then we linked each individual to their Wikipedia article ID, and obtained the linkages to all other Wikipedia language editions through the Wikipedia API. The set of individual IDs represent a snapshot of figures that as of May 2013 have a presence in more than 25 language editions of Wikipedia--11,340 biographies in total. Finally, we supplemented the data with monthly page view data from Jan. 2008 through Dec. 2013, across all language editions of Wikipedia.

Since there is no globally standardized classification system of cultural production, we introduce a normalized hierarchy of cultural domains, classifying cultural domains at three levels of aggregation. In terms of location assignment, we attribute individuals to a place of birth by country, based on current political boundaries for consistency through time. Birthplaces were obtained by scraping both Freebase and Wikipedia, and further refined by using fuzzy location matching and geocoding with the Yahoo Placemaker and Google Maps geocoding APIs, and by manual curation.

2.1.2 Validation

Pantheon measures fame, but fame can be strongly correlated with accomplishment in a world where the market of fame is relatively fair. To validate Pantheon's measures of individual fame as measures of accomplishment we regress Pantheon's Historical Popularity Index (HPI) against external measures of accomplishment in cultural domains where these are available (mainly individual sports). We find that Pantheon's metrics strongly correlate with the professional success of Formula-1 drivers (R^2=92%), tennis players (R^2=66%), and swimmers (R^2=70%). While these examples are not exhaustive across all cultural domains, they show that the measures of fame approximate metrics of cultural production across different domains.

2.2 Human Accomplishments Dataset

* ayu@media.mit.edu
† kzh@media.mit.edu
‡ djagdish@media.mit.edu
§ hidalgo@media.mit.edu

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

Pantheon's visualization engine also includes Charles Murray's Human Accomplishment dataset [1]. The Human Accomplishment dataset includes distinguished figures in the arts and sciences, each mapped to a location and time period, and ranked based on the level of attention given to the individual in qualified sources for the subset of cultural domains surveyed. Further detail can be found in *Human Accomplishment* [1].

3 Visualization Selection and Design

The Pantheon website allows for dynamic, interactive visualization of the Pantheon and Human Accomplishments datasets. We included four visualization types (treemaps, matrices, scatterplots and maps) each corresponding to a specific type of question answered by the Pantheon dataset.

Treemaps: 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. The treemap in Figure 1 shows the shares of each domain within the cultural production of the United Kingdom.

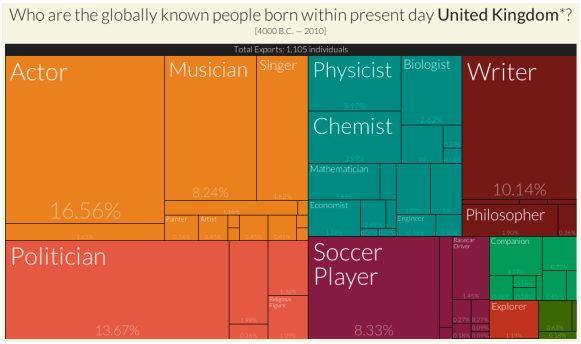

Figure 1: Treemap of the globally known people from the UK

Matrices The matrix view gives an overview of the underlying structure of large, high-dimensional datasets by simultaneously encoding the interactions between many variables. The rows of the matrix show the countries, and the columns show cultural domains. The sorting order of the matrix can be updated interactively. 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. Figure 2 below shows the matrix of domains for the top 40 countries overall.

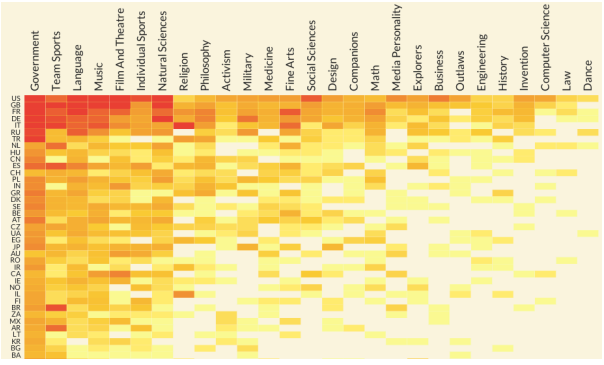

Figure 2: Matrix visualization of the top 40 countries

Scatterplots Scatterplots compare 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 to indicate whether the entity on the y-axis or the entity on the x-axis produce more individuals within the filter parameters. Figure 3 shows a log scale scatterplot comparing Explorers and Chemists.

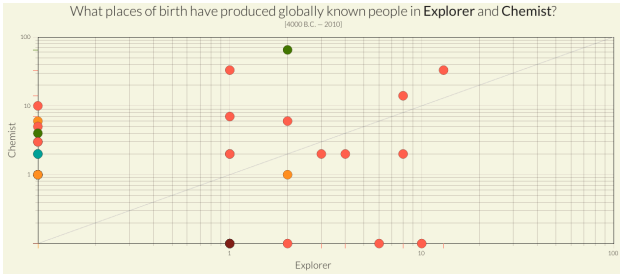

Figure 3: Scatterplot comparing Explorers and Chemists

Maps The map visualization shows geographic distribution of a selected cultural domain, encoded by color hue and intensity. For example, Figure 4 shows a map of individuals in the Sports domain, for all time periods.

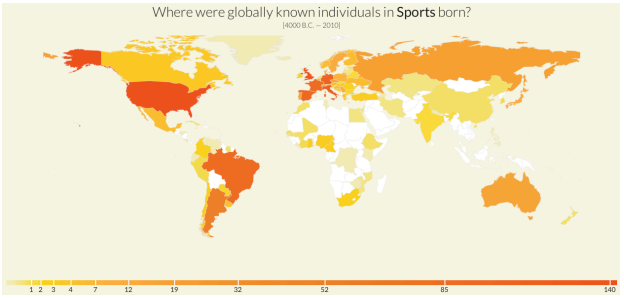

Figure 4: Map visualization of individuals in Sports

4 Conclusion and Future Work

This poster presents Pantheon, a new dataset and visualization portal for cultural production. The future potential of Pantheon includes extending the existing visualizations and incorporating new datasets to further enable the quantitative study of historical cultural patterns.

Acknowledgments

We thank Charles Murray for the *Human Accomplishments* dataset, and Ali Almossawi, Shahar Ronen, Tiffany Liu, Andrew Mao, and Defne Gurel for their contributions. This work is supported by funding from the MIT Media Lab Consortia.

References

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

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

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

[4] Google. (2013) Google Books Ngram Viewer. [Online]. https://books.google.com/ngrams

[5] Data Science Lab, Stony Brook University. (2013) Who is Bigger? [Online]. http://www.whoisbigger.com/

[6] Google. (2012, Nov) Freebase Data Dumps. [Online]. https://developers.google.com/freebase/data