Data Portraiture and Topic Models

Aaron Zinman, Doug Fritz
Fluid Interfaces Group
MIT Media Lab
Cambridge, MA 02138
{azinman, doug}@media.mit.edu

Abstract

The art world uses the tradition of portraiture to semantically compress relevant information about their subjects into a single art work. We believe this tradition can be extended into the digital realm by the use of topic modeling to compress individuals' information into generative data portraits. Data portraiture has strong implications for navigating large social spaces, collaborative systems, and self reflection. We briefly showcase our works thus far to highlight potential directions in this emerging field.

1 Introduction

In cyberspace, we are bodiless. Despite the obvious and long desired advantages of removing race, gender, age, and other non-mental attributes from online interactions [7], the physical body remains a powerful force in face-to-face interactions. Stereotyping allows society to function as a whole [9], and minute physical gestures are important for efficient communication [10], trustworthiness [6], and expression of identity [4]. In the art world, portraiture has been long-standing and rich tradition that exploits our ability to recognize these physical properties to obtain a multidimensional gestalt of character, form, and function [2]. We believe that carrying over this tradition into the digital realm can help individuals not only make better sense of strangers in the online spaces they inhabit, it can help organizations to understand the information flow within, facilitate better collaboration, and function ego-centrically as a digital mirror to better understand ourselves.

Data-driven portraiture, or simply data portraiture, is the generative process of creating visual representations of the self based on abstract data rather than facial features. These representations can function as a body for where we are bodiless, facilitating navigation of large quantities of human-centered information. Topic modeling plays a special role in data portraiture for its ability to aggregate large amounts of unorganized information into high-level categories. These categories and associated weighted vectors allows data portraiture to significantly advance in utility and power.

In this paper we briefly outline related works and our contributions of data portraiture using topic modeling.

2 Presentation and interpretation of data

In post-Renaissance western Europe, portraiture was reserved as a way for the rich and powerful to encapsulate their accomplishments and status. Men would be painted with their weapons, symbolic or real. Noblemen differentiated themselves through clothing, stance, and scene. The potential meaning derived from a work become a mixture of projection of the subject through the lens of the artist. Artists have the benefit of human reasoning and expressive capacities to cast the subject in the light of their choosing. They can meaningfully add objects to the scene, alter expression on the micro-level, and even change the overall colors of the scene to reflect a desired mood. Like topic models, artists carefully perform semantic compression of their subjects.

In data-driven portraiture, we do not always have the luxury of human intervention with each generated portrait. Nor should we; we gain the digital advantage of presenting extracted meaning from data on a large scale. However, much like the artist injects subjectivity into their portraits with every brush stroke, the data artist does the same. In selecting the choice of algorithm, data sets, stop words, and eventual visual representation, the data miner injects their choices into a domain that is often viewed as authoritative. We must take special care to ensure that the resulting presentation reflects the same amount of ambiguity inherent in the compression. This is difficult in the abstracted domain because we cannot easily rely on the pre-existing categories and stereotypes that we normally use to infer unknown attributes of others [9]. Instead, we must pay careful attention to the tools of the abstract domain; visually this is often color, shape, and typography. Color effects, metaphors from the physical world, and cultural traditions can unexpectedly assign and alter meaning to different parameterizations of these abstract classes. For example, many, but not all cultures interpret the color red with danger, violence, and passion. Seemingly arbitrary choices like the hue range in a color spectrum can also affect interpretation of purely scientific data, despite the common brightness and saturation levels [8]. See [8] for a good source of issues and guidelines in visualizing scientific data.

Despite these issues amongst others, carefully designed data portraits can enable and facilitate exciting new applications. In a world increasingly overflowing with information, reliable methods of presenting raw and aggregated data has become a central necessity. Because of its semantic power and unsupervised nature, topic modeling has been especially appealing to us in generating data portraits. We now present a collection of our works to demonstrate what we believe to be an exciting direction.

3 Our work

3.1 Landscape of Words

Landscape of Words is a visualization created by Aaron Zinman and Alex Dragulescu that centered around applying Latent Dirichlet Allocation [1] to all of Twitter. It was constructed when the service had not quite reached the mainstream recognition it has today, and was meant to allow newcomers to become familiar with the service. Inspired by earlier work with landscape metaphors and LSA [3], we represent each topic as a "mountain" whose height is proportional to the total number of tweets that fall under that topic. Here we quickly run into problems of symmetric hyper-parameters in the model, where most topics are of similar likelihood (according to the model). Multi-dimensional scaling of Kullback–Leibler divergences grouped related topics. The number of topics was artificially reduced to not overload the landscape. Trade-offs in model completeness versus navigation are an inherent problem in topic modeling. The landscape is annotated using heat maps to display topic popularity across time (Figure 1, center), compare individuals, and social networks (Figure 1, right). Using a common model across all of Twitter keeps the location and contents of the topics static, thereby making it easier to compare like-entities. We reserve a small disconnected area in the corner for Tweets that have low-probability assignments to the model, an important feature when presenting data via topic models.

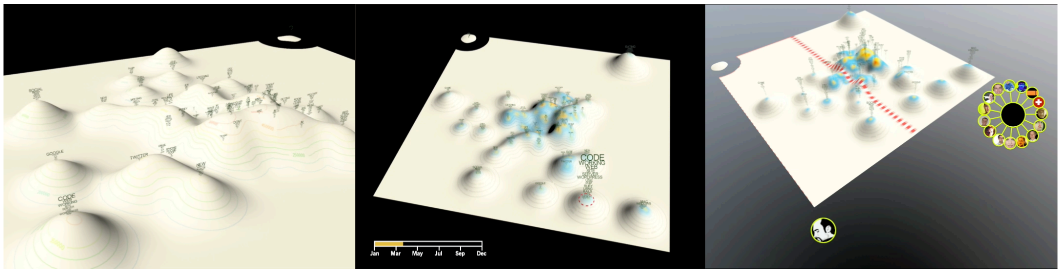
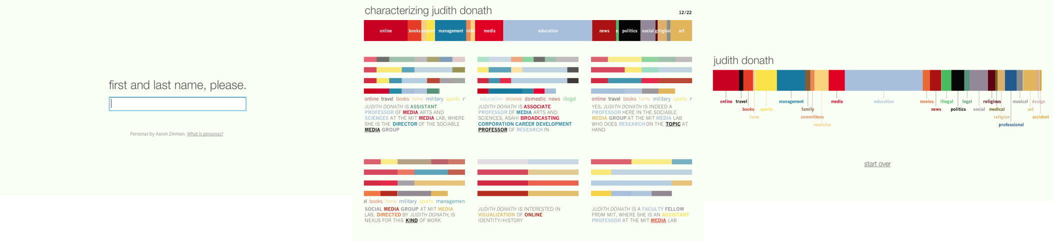


Figure 1: Landscape of Words. Each "mountain" represents a topic, annotated with its most probable words. Heat maps display popularity over time (center) or facilitate comparisons amongst individuals and groups (right).

3.2 Personas

Personas is an art piece designed for the Metropath(ologies) exhibit recently on display at the MIT Museum. Museum-goers approach the terminal which asks the question “first and last name, please.” as shown in Figure 2. Upon receiving a name, it uses Yahoo!’s BOSS service to find web pages that contain characterizing statements about the individual by searching phrases such as “first+last {is,was,will be}”. Results are filtered for the sentence containing the viewer’s name. Inference using LDA is performed on the sentences, saving the results of each iteration of the Gibbs sampler. A similar process was used to generate the training corpus, by searching two million names generated in proportion to the US census. The inference is then animated for the user, allowing them to reflect on the information given, the oscillations inherent in the machine trying to make sense of the data, and inspect the attributions. Meanwhile, an aggregated DNA-like strip of the document-topic vector is continuously adjusted as the results change. Finally, this aggregated strip is presented as the data portrait of the user. It is meant for them to reflect on issues of privacy, identity, and faith in data mining. Despite the tones of critique inherent in the piece, it is still a valuable exercise in data portraiture with topic modeling. Ported to the web, Personas has generated over 1.5 million portraits in just over one month. Commentary across Twitter and various social media has anecdotally demonstrated that everyday users admire the relative simplicity and legibility despite the lack of obvious utility. However, commentary has also highlighted the dangers of data portraiture using topic modeling. Despite being shown the raw data and the topic assignments as the inference process iterated, individuals attributed ESP-like properties to its conclusions.



JUDITH DONATH IS ASSISTANT PROFESSOR OF MEDIA ARTS AND SCIENCES AT THE MIT MEDIA LAB, WHERE SHE IS THE DIRECTOR OF THE SOCIABLE MEDIA GROUP

JUDITH DONATH IS ASSOCIATE PROFESSOR OF MEDIA ARTS AND SCIENCES, ASAHI BROADCASTING CORPORATION CAREER DEVELOPMENT PROFESSOR OF RESEARCH IN

YES, JUDITH DONATH IS INDEED A PROFESSOR HERE IN THE SOCIABLE MEDIA GROUP AT THE MIT MEDIA LAB WHO DOES RESEARCH ON THE TOPIC AT HAND

SOCIAL MEDIA GROUP AT MIT MEDIA LAB, DIRECTED BY JUDITH DONATH, IS NEXUS FOR THIS KIND OF WORK

JUDITH DONATH IS INTERESTED IN VISUALIZATION OF ONLINE IDENTITY/HISTORY

JUDITH DONATH IS A FACULTY FELLOW FROM MIT, WHERE SHE IS AN ASSISTANT PROFESSOR AT THE MIT MEDIA LAB

Figure 2: Screenshots of Personas. Upon entering a name (left), characterizing snippets of the user and its inference in the model are animated (center). Finally the resulting weighted vector is displayed (right).

3.3 Defuse

Defuse is a project currently under development that aims to create top-down navigation for browsing newspaper comments online. Online commenting in large sites is a space that is ripe for data portraiture, as condensing the history of posters helps contextualize their messages for unfamiliar users. Currently topic modeling is being used to generate clusters of discussion topics and conversational style, which are then used to showcase users by their past topical history as a gestalt for their interest patterns over time. We plan to use topic modeling as a summarization tool for building navigable models of heavy participants.

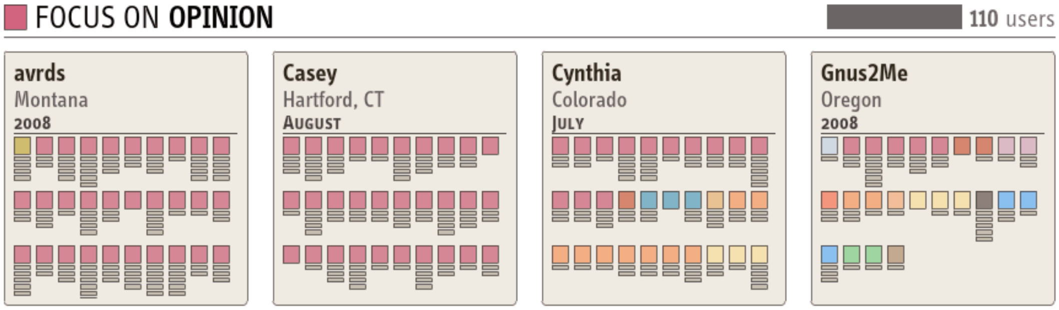


Figure 3: Data portraits of NYTimes commenters. Each colored box represents a message and its attributed topic, augmented with tick marks in proportion to the number of recommendations they

received from the community. Users are clustered by the topic most attributed to their posts which is currently human labeled.

3.4 ConnectUs

ConnectUs is a visualization to better incorporate our digital personas into physical interactions. The project optically tracks and identifies participants moving around in a space. A priori, participants bind their physical identity to their online social networking accounts. Available data is used to create a common topic model. An ego-centric data portrait follows participants as they physically move, deforming as they come into contact with others in the room to show similarity. Topics unique to those in conversation are highlighted to function as a digital impetus for social interaction.

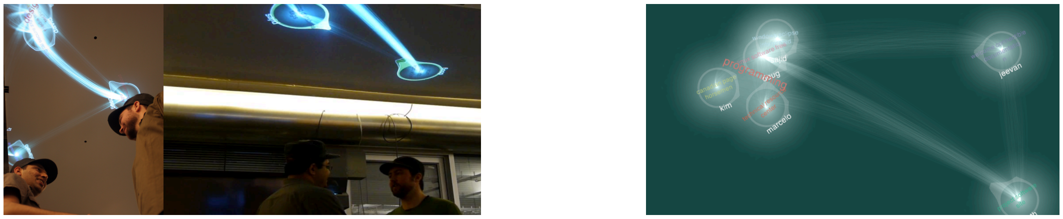


Figure 4: Showing the meeting of two people and the resulting deformation of their compressed digital portraits.

3.5 ThemeStream

ThemeStream is a visualization of the major themes of one's personal RSS feeds. Zooming around a 3D interface, it provides differing levels of detail depending upon the camera view, alternating between high-level top-down views, and focused contextual views. Much in the way an artist paints the same model from different perspectives, creating varying views of data, attribution, and detail often reveals more than any one method could provide. It uses topic modeling to compress individual time slices and OpenCalais for theme extractions.

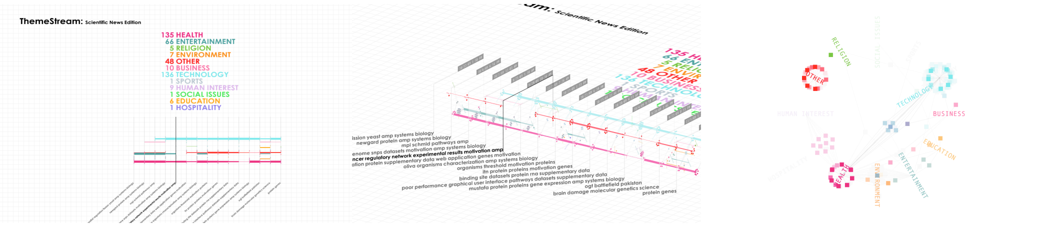


Figure 5: Each of the different views are transformations of the same consistent layout. The top view (left) exposes major fluctuations in time and major topic frequency changes as they flow from one time slice to the next. The three-quarters view (center) reveals a localized meta-view, displaying the topics popular within each time slice. Finally, the front view (right) accesses each of the individual items, which can be opened and viewed in full.

4 Conclusion

We believe that topic modeling can symbiotically provide useful semantic compression techniques for data portraiture. We further believe the presented works are unique contributions that can help illuminate potential directions for collaboration between artists and data miners. Such collaboration can lead to breakthroughs in navigation of large-scale social data, advancement in collaborative systems, and a better ego-centric understanding of our digital lives.

Acknowledgments

We would like to thank the sponsors of the MIT Media Lab for their generous support.

References

[1] Blei, D., Ng, A. & Jordan, M. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, **3**:993–1022.

[2] Brilliant, R. (1991) *Portraiture*. Cambridge, MA: Harvard University Press.

[3] M. Chalmers. (1993) Using a landscape metaphor to represent a corpus of documents. In A. Frank and I. Campari, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, pp. 377–390. Springer-Verlag.

[4] Donath, J. (2007) Virtually trustworthy. Science, **317**(5834):53-4.

[5] Donath, J. (2001) Mediated Faces. In M. Beynon, C.L. Nehaniv, K. Dautenhahn (Eds.). Cognitive Technology: Instruments of Mind Proceedings of the 4th International Conference

[6] Handy, C. (1995) Trust and the virtual organization. Harvard Business Review, **7**(3):40-50.

[7] Hiltz, S.R. & Turoff, M. (1978) The network nation: Human communication via computer. Reading, Massachusetts: Addison Wesley.

[8] Rogowitz, B.E. & Treinish, L.A. (1996). How NOT to Lie with Visualization. Computers in Physics, **10**(6):268–273.

[9] Simmel, G. (1910) "How is society possible?" *American Journal of Sociology,* **16**:372-391.

[10] Zebrowitz, L. (1997) *Reading Faces*. Westview Press, Boulder, CO.