Revisiting Du Bois

A Data Visualization Project

Nathan Kim

Advised by Professor Elisa Celis

Introduction

Prolific and prominent Black writer, historian, professor, political activist, W.E.B. Du Bois has left a legend in many ways. One profound way as been in developing and articulating data visualizations many years before canonical pioneers of visualization like Edward Tufte, Jacques Bertin, or Stephen Few. My project seeks to explore his work with a critical and quantitative lens. I hope to bring to Du Bois' portraits the worlds of modern statistics, computing resources, and interactive visualization, but also to these fields bring Du Bois' humanist lens and the goal of visualizing data to create a better world.

To do so, my project will focus on two questions:

1. **How does Du Bois' approach to data visualization depart from canonical views of data and information?**

In addition to being a sociologist who meticulously drew and recorded social statistics for over fifty years, Du Bois was also a Black theorist of knowledge, an educator, a novel writer, and a historian. Some of these perspectives are visible in his approach to data visualization. While other data visualization pioneers like Edward Tufte try to objectively gauge the value of a dataset as the proportion to which it represents "truth," Du Bois recognizes visualization as a creative endeavor alongside (quite literally, in the case of the

Paris Exposition) photography and historical analysis. In extending Du Bois' data visualization through practical tools, I hope to also convey his artistic perspective on data visualization.

2. **How have the subjects of Du Bois' works in the 1900 Paris Exposition evolved over time?**

Du Bois was concerned with the "afterlives" of slavery and prospects of Black people in the American South after emancipation in 1865, and he provided answers by studing land ownership, occupations, income, and geographic concentration. How does the same topic of the afterlife of slavery in land ownership, occupations, income, and geographic dispersion look like in the contemporary age?

Background

Data visualization through the centuries

Modern data visualization began with modernity itself. In the 1600s, as writers

The map as a medium was developed alongside modern notions of states and borders throughout the fifteenth and sixteenth centuries.[1] Other forms of data visualization, like bar and line charts, were developed over time to similarly convey graphically what language itself could not express. The earliest known use of these graphics to convey statistical information appeared in 1644, when a Flemish astronomer named Michael Florent was tasked with representing many different distances. A table might have sufficed to convey the raw denotation of these [2].

The next three hundred years brought much change to what is known as data visualization, but the fundamental idea of translating statistical concepts in graphical ideas remained the same. Technological advancements in reproducible printing and color usage

Today, the accepted canon of data visualization pioneers has coalesced around Edward Tufte, Jacques Bertin, and John W. Tukey.

[1]

[2] As an example, see the discussion fo Michael Florent's 1644 graphic in Friendly, "Milestones in the History of Data Visualization."

Du Bois

The majority of Du Bois' work would come during the aforementioned "modern dark ages" of data visualization, during which some scholars consider few methodological, technological, or artistic innovations to have been made.

Du Bois himself was born in Massachusetts in 1868, attending an integrated public school as a child before going to Fisk University in Tennessee.[3] He attended Harvard University for a second degree beginning in 1888, and enrolled in graduate study at Harvard for sociology. After receiving his degree from Harvard, he began a highly prolific career with various positions at Wilberforce University, the University of Pennsylvania, Atlanta University, the Tuskegee Institute, the NAACP, and others. He died in 1963 in Ghana, while working on an encylopedia of Afria and the African diaspora, in exile from the U.S. for his Communist sympathies.

He is known today for his massive array of contributions to the fields of African American studies and sociology, ranging from the first sociological study of a Black community in *The Philadelphia Negro*, to histories like *Africa: Its Place in Modern History*, to more theoretical texts like *The Souls of Black Folk*, and finally to creative and personal pieces like *Dusk of Dawn* and *The Quest of the Silver Fleece*. Because of the breadth of his work,

Perhaps most substantially for my project, Du Bois also had unique views on statistics and quantitative information that made its way into its work. For instance, *The Philadelphia Negro* was one of the first studies to incorporate statistics into a sociological study, now a standard practice. More generally, Du Bois' views on statistics are ones that I hope to

In the seventy-five years since Du Bois' death, technological advacements have given

My project

[3] Many scholars attribute Du Bois' central interest in racism in the South to have begun during this time.

Methods

Extending ggplot2

The first contribution of my project is to create an R package extending W. E. B. Du Bois' data visualizations to present day. These include

R is a programming language widely used in statistics and data visualization. Its features of being open-sourced (and thus free and extensible), easier to learn and write compared to some oth er languages, and having many native data structures and functions for computing models have made it popular today.

One especially important extension in R is the ggplot2 library and the ecosystem of user-contributed software packages it has spawned. The ggplot2 package is popular for providing many utility functions for graphics in R, for example a function called `geom_smooth` for smoothed conditional means. These functions can often abstract away complex logic from users, so that (in the case of `geom_smooth`) a potentially complicated choice between loess and general aggression models can be hidden from the user that simply sees a best fit line.

But the library is far more influential for contributing its eponymous *grammar of graphics*, or an extendable logic for how to make plots. Plots in every programming language can quickly become complex and syntactically verbose because of how many geometries and aesthetics of a plot there are to consider. ggplot2 was created as a response to this situation in R, creating a grammar from which almost any graphic could be greated. In this grammar, every plot can be decomposed into a few ingredients:

1. A dataset
2. One or more layers of geometric objects
3. A scale to determine how data should be mapped to positions or aesthetics
4. A coordinate system to translate conceptual *x/y* or *radian/theta* values to pixel values
5. A theme, or aesthetic styling that are not directly related to the data itself.

A demonstration of this logic is below. The plot is initialize with `ggplot()`, then an arbitrary number of geometries can be added with the `geom_*` syntax, and finally a theme is added for styling.

Listing 1 Demo of the grammar of graphics

```
library(ggplot2)
ggplot(iris, aes(x =  Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point() +
  geom_smooth() +
  theme_classic()
```

The power of ggplot2 lies in taking a chaotic array of plots and consolidates them into a few recognizable forms. In both the "vocabulary," or available geometries, and the "grammar,"

But this feature is also a limitation of ggplot2. Extending ggplot2 to fit new geometries is difficult, and users are left with the (albeit large) selection of features that the ggplot2 library maintainers have already created. Creating a

To return finally to Du Bois' perspectives on data visualization, it is often not enough to rely only on canonical forms of charts that we can see dryly in academic texts.

Bayesian analysis

My project seeks to extend Du Bois' works not only with modern graphical tools, but also with modern analysis tools. In much the same way, I hope to convey the spirit of Du Bois' analyses with the methodological advancements he did not have.

In my view, much of his perspective can be seen in nonparametric testing and analyses,

I will run a set of Bayesian hierarchical time-series model to investigate Du Bois' main research questions of property ownership, educational attainment,

Data and source code

Data used in the analysis and as a demonstration tool in the `ggdubois` package will come from several demographic sources. From the United States Census Bureau, I

All variables will be taken for Georgia

The source code for the `ggdubois` package can be found at https://github.com/18kimn/ggdubois, and the source code for this paper and the associated analyses can be found at https://github.com/

18kimn/revisiting-dubois.

`ggdubois`

Analysis

Conclusion

Bois, W. E. B. Du. *Black Folk Then and Now (The Oxford W.E.B. Du Bois): An Essay in the History and Sociology of the Negro Race*. Oxford University Press, 2014.

———. *The Philadelphia Negro: A Social Study: The Oxford W. E. B. Du Bois, Volume 2*. OUP USA, 2007.

———. *The Souls of Black Folk*. Oxford University Press, 2008.

Bois, William Edward Burghardt Du. *Darkwater: Voices from Within the Veil*. Harcourt, Brace and Howe, 1920.

———. *The World and Africa: An Inquiry Into the Part Which Africa Has Played in World History and Color and De: The Oxford W. E. B. Du Bois*. OUP USA, 2007.

Bois, William Edward Burghardt Du, and Henry Louis Gates. *Dusk of Dawn (the Oxford W. E. B. Du Bois)*. Oxford University Press, 2014.

"Emancipatory Empiricism: The Rural Sociology of W.E.B. Du Bois - Joseph Jakubek, Spencer D. Wood, 2018." Accessed September 15, 2021. https://journals.sagepub.com/doi/full/10.1177/2332649217701750?casa_token=B5xNMv-OdHgAAAAA%3AA6_BieR56fCKu381aEryc0gG_a7YnrGUNGPHh2ZbWuySKakXvEMwr3HTWtHNS5A8zpjTdezZZpb5Sw.

Friendly, Michael. "A Brief History of Data Visualization." In *Handbook of Data Visualization*, edited by Chun-houh Chen, Wolfgang Härdle, and Antony Unwin, 15–56. Springer Handbooks Comp.Statistics. Berlin, Heidelberg: Springer, 2008. https://doi.org/10.1007/978-3-540-33037-0_2.

———. "Milestones in the History of Data Visualization: A Case Study in Statistical Historiography."

In *Classification — the Ubiquitous Challenge*, edited by Claus Weihs and Wolfgang Gaul, 34–52. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer, 2005. https://doi.org/10.1007/3-540-28084-7_4.

Karduni, Alireza, Ryan Wesslen, Isaac Cho, and Wenwen Dou. "Du Bois Wrapped Bar Chart: Visualizing Categorical Data with Disproportionate Values," January 30, 2020. http://arxiv.org/abs/2001.03271.

Rabaka, Reiland. "The Racialization of Information: W.E.B. Du Bois, Early Intersectionality, and Social Information." In *Information and the History of Philosophy*. Routledge, 2021.

Tukey, John Wilder. *Exploratory Data Analysis*. Addison Wesley Publishing Company, 1970. https://books.google.com?id=hrJ5yAEACAAJ.