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. The map as a medium was developed first, alongside modern notions of states and borders throughout the fifteenth and sixteenth centuries.[1] States had new reasons to formally define borders through a codified text and recognizable medium – the map. Other forms of data visualization, like bar and line charts, were developed over time to similarly convey graphically what language had to stretch to 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 idea of translating statistical concepts in graphical ideas remained the same. Most fundamentally, the idea of *measurement* to produce a graphic solidified during this time, assisted by both new tools and new schools of thought to guide them. As the age of Exploration, colonization, and the Enlightenment as a whole took form, so too did units of distance and time to measure and quantify colonizers' discoveries. First there emerged mediums to quantify raw distances and quantities, like maps for navigation, but

[1]

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

soon followed more elaborate and abstracted forms of describing space, time, and everything in between. Topographical maps involving contour plots emerged in 1584 but became more common in the 1700s. The first known scatter plot emerged in 1833. John Snow in 1854 created his famous maps of the London cholera outbreaks, at a time when the map as a medium was familiar but the application of such a medium to convey statistical trends was not. The transformation of scales to accomodate non-uniform data became prominent in (1863, 1869). All of these methodological innovations were accompanied by technologies like color printing and logistical projects like nationwide censuses that cemented visualization as a canonical form of understanding data. Besides a brief period that some scholars call the "dark ages" of modern data visualization, discoveries like these progressed steadily from 1600 onward.

Today, the accepted canon of data visualization pioneers has coalesced around Edward Tufte, Jacques Bertin, and John W. Tukey.[3] Bertin, in his *Semiology of Graphics*, established a conceptual backbone that related visual elements directly to trends within data, from which other graphical "grammars" (including ggplot2, as mentioned below) would one day develop. John Tukey, in his *The Future of Data Analysis* (1962) and *Exploratory Data Analysis* (1977), argued for data analysis as a branch of statistics distinct from mathematical statistics and argued that recognizable and reproducible visualizations were key parts of understanding data. Edward Tufte, in *The Visual Display of Quantitative Information* and many other texts, established data visualization as a

These three writers, along with the

To summarize, the dominant trend of data visualization has been to express graphically statistical arguments, or to gesture somewhat to

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 innovations to have been made. But I hope to demonstrate that Du Bois' work provides ideas that are still

Du Bois himself was born in Massachusetts in 1868, attending an integrated public school as a child

[3] There are of course many, many scholars that laid the foundation for modern data visualization besides these. Some examples are....

before going to Fisk University in Tennessee.[4] 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

ggplot2

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.

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

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, or "geoms", along with mappings between data and features in these geomeetric objects.
3. A coordinate system to translate conceptual *x/y* or *radian/theta* values to pixel values.
4. A theme, or aesthetic styling that are not directly related to the data itself.

A demonstration of this logic is shown in Block 1. The plot is initialize with `ggplot()`, then an arbitrary number of geometries can be added with the `geom_*` syntax, a scale is used to adjust how items in the data are mapped to aesthetics, and finally a theme is added for styling. The strength of `ggplot2` lies in how the individual components of a graphic are separated by a plus sign (+), or in other words that visual components can be written directly through syntactical components. Like a linguistic grammar, this grammar allows for the recombination and reuse of its component "words." One could substitute the point geometry layer `geom_point` with a density map layer `geom_density`, or the provided "classic" theme with a black-and-white theme `theme_bw` or even a custom user-created theme.[5]

Listing 1 An example of the grammar of graphics, in code.

```r
library(ggplot2)
ggplot(iris, aes(x =  Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point() +
  geom_smooth() +
  scale_color_manual(values = c("red", "blue", "green")) +
  theme_classic()
```

[5] A gallery of such themes can be seen here.

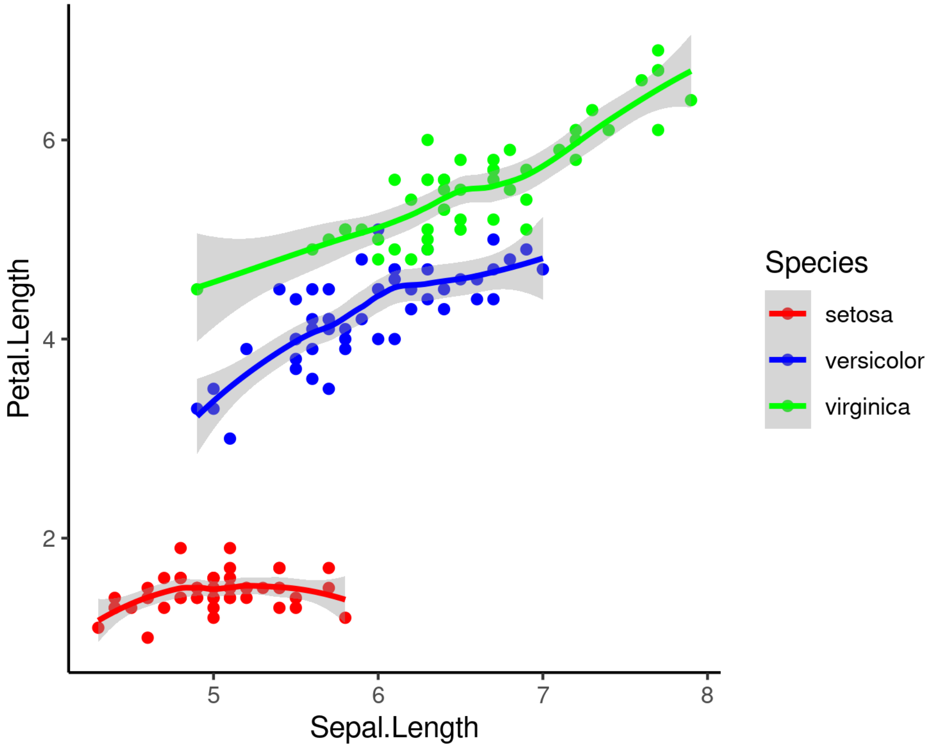

Figure 1: The rendered output of the `ggplot2` code.

The power of `ggplot2` lies in taking a chaotic array of plots and abstracting them into variations of a few recognizable forms. The ggplot2 philosophy holds that every graphic can be decomposed into the grammar of graphics, no matter how many odd shapes or colors it might have. Thus, in both the "vocabulary," or available geometries, and the "grammar," `ggplot2` has acted as a unifying tool for statisticians and data scientists.

But as much as it is unifying and expansive, `ggplot2` is still limited. As an example, consider force-network graphs, in which points are animated with positions determined by a simulation. At each step of the simulation the points' next resulting placement is calculated as a result of the points' previous placements, based on attraction towards linked points and repulsion from non-linked points. Points then become closer and further from each other on the graph depending on how "linked" they are, turning the graph into a sort of conceptual map. Such a graphic has been made popular in recent years by visualization tools like d3.js, but does not fit paradigmatically in the ggplot2's grammar of graphics. The original specification of the grammar of graphics assumed static, non-animated plots; where would animated graphs like force networks fit in? What would be a force network's scale, or the mapping between data and positional coordinates, given that nodes in a force network are updated with every "tick" of the simulation and the position of one node is determined by other nodes instead of attributes in the data?

Force networks are just one example of ggplot2's general limitation. By consolidating graphics into a few recognizable forms and rules, ggplot2 leaves out others. In some cases, like not including animated forms, this is mostly "accidental." It doesn't contradict the logic of ggplot2 to extend it with the axis of time, and animation could very well have just been left out of the `ggplot2` package simply to limit the package's scope. Other cases, like having positions of geometries not depend on data itself but be determined through a random or simulated process, are not part of the `ggplot2` logic and contradict existing aspects.

The community-prescribed solution both of these limitations has been to extend `ggplot2` with additional functionality and in the process rework the `ggplot2` grammar. The `gganimate` package was created to handle transitions, and is now maintained by one of the maintainers of the ggplot2 core library.[6] Other libraries, like `ggigraph`, `gggraph`, `geomnet`, and others have extended ggplot2's functionality to visualize

These extensions take a pragmatic position on the grammar of graphics

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 as a standard form in academic texts. Data visualization has to be expressive, stretching beyond the unified abstractions to more diverse (if chaotic) forms as well.

Methods

`ggdubois`: An extension for ggplot2

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

[6] the `gganimate` package and other animation packages in R create animations notably in a limited and performance-inefficient way compared to many tools. These animations by simply appending many individual animations together, whereas web-based and OpenGL-based animation tools can use optimized rendering engines to create animations faster and with smaller file sizes. This is an issue I hope will be improved on for R in the future, but as it is not directly relevant to my project I will leave this to a footnote.

The majority of `ggdubois` contains additional geometries, or geoms, through which new shapes in the `ggplot2` framework can be created. In some cases this required only adding a preprocessing step and then modifying a combination of existing "geoms," or geometries, in ggplot2. In other cases, this required reaching into the `grid` package over which `ggplot2` itself was built, and adding completely new shapes for `ggplot2`.

Besides new geometries, `ggdubois` contains two additional tools. The first is a theme, which can be applied to any `ggplot2` object and will replace existing the styling of that plot's text, colors, and outline with aesthetics inspired by W. E. B. Du Bois' work for the 1900 Paris Exposition. The second is a color palette, which is used in the theme but can also be used separately to color arbitrary graphics and plots with R.

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.

- one where a theoretical result is derived probabilistically in tandem with the data (continually updated), instead of assuming that a point value for a parameter exists independent of the data at hand
- repeated, continually updated inference <-> critical theory and self-critique
    - MCMC?

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, packaged into two datasets. The first dataset contains time-series data at the county level for Georgia, containing decennial data from 1970 to 2010 on race, educational attainment, housing ownership, and employment. This dataset roughly mirrors much of Du Bois' own scope as seen in his graphics for the 1901 Paris Exposition. The second dataset contains nationwide county-level data measured in 2017, and measures median household income, the unemployment rate, the child poverty rate, the population of color, the amount of particulate matter of less than 2.5 μm in the atmosphere, the "rent burden" or the average proportion of income spent on rent, the high school graduation rate, and the Gini index for income inequality (B19083). This dataset aims to extend Du Bois' commentary on socioeconomic inequality to the national level, with data Du Bois did not have access to in both subject and scope.(reword) More detailed descriptions of these two datasets can be found in Appendix 1. For the rest of this paper, the first dataset will be referred to as `georgia` and the second will be referred to as `demographics`.

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

The `ggdubois` package contains seven visualization extensions, a theme, and a color palette.

geom_scaledmap

The first custom geom that my package contributes is `geom_scaledmap`. Adapted from Plate 42 of *Data Portraits*, this geom receives one continuous variable and one categorical or discrete variable as an input, as well as the coordinates of a geographic polygon that they are associated with. The geometry then plots the polygon once for each value of the discrete variable, scaling the polygon according to the value of the continuous variable.

An example of this is shown below. From the `georgia` dataset described above, year from 1970 to 2000 is used as a discrete variable and the proportion of high school graduates is used as a continuous variable.[7] Georgia is plotted four times, one for each specified year, and scaled to the relative size of proportion of high school graduates.

Listing 2 geom_scaledmap()

```r
ggplot(georgia, aes(x = year, y = high_school_grads)) +
  geom_scaledmap()
```

This is fairly different from a traditional map of a continuous variable, which often colors sections of a polygon instead of repeatedly drawing a polygon with modifications. Traditional maps would tell viewers about the geographic distribution of a particular process,

Like many of the geoms shown below, this

`geom_spike`

The second geom that my package contributes is called `geom_spike`. This receives again one continuous variable and one categorical or discrete variable as an input. It then creates set of concurrent circles, with the distance between these concentric circles being proportional to the supplied continuous variable. "Spikes" extend from outer layers into inner ones.

Listing 3 geom_spike()

```r
ggplot(georgia, aes(x = year, y = high_school_grads)) +
  geom_spike()
```

This geom was inspired by plate 22 of *Data Portraits*,

A quirk about this geom and the `geom_spiral` function that follows is that in order to be understood within the `ggplot2` framework, the coordinate system must explicitly be specified as polar and not the default cartesian coordinate system. In the grammar of graphics, aesthetic mappings from data to coordinates and the mapping of coordinates to pixel positions are handled separately.

[7] Note that, as as idiomatic in `ggplot2`, the information encoding geometric columns is automatically detected by `geom_scaledmap`.

`geom_spiral`

`geom_spiralpath`

`geom_wrappedbar`

`geom_bibar`

`geom_wovenbar`

`theme_dubois`

`dubois_pal`

Analysis

Conclusion

Bibliography

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.