Chapter Title: New Data? The Role of Statistics in DH
Chapter Author(s): TAYLOR ARNOLD and LAUREN TILTON

Book Title: Debates in the Digital Humanities 2019
Book Editor(s): Matthew K. Gold and Lauren F. Klein
Published by: University of Minnesota Press

Stable URL: https://www.jstor.org/stable/10.5749/j.ctvg251hk.27

PART III][*Chapter 24*

New Data? The Role of Statistics in DH

TAYLOR ARNOLD AND LAUREN TILTON

In an emerging terminological shift, digital humanists are increasingly reframing their objects of study as "humanities data." In "Big? Smart? Clean? Messy? Data in the Humanities," Christof Schöch compares the state of computational research in the humanities to the deluge of large, unstructured datasets that have become objects of study across academia and industry. Along the same lines, IEEE recently organized three workshops on "Big Humanities Data." Building off of this shift in terms, Miriam Posner in her piece "Humanities Data: A Necessary Contradiction" explores both why this terminology is needed and why it may be difficult for humanists to consider their evidential sources as akin to those of the sciences and social sciences. We also chose to use the term "humanities data" in our book: *Humanities Data in R.* Rather than emphasizing contradictions among the disciplines, we sought to find common ground across them. We set out to show how statistics—the organization, analysis, interpretation, and presentation of data—is a fundamental interlocutor and methodological approach for the digital humanities. In this chapter we take the latent argument of our book and make it explicit.

Despite the constant refrain suggesting that digital humanities should take better account of computer science, it is actually statistics that provides a set of approaches for exploring, analyzing, and thinking critically about data in order to posit new findings and questions related to our fields of study.[1] In this chapter, we argue that three areas of statistics—exploratory data analysis (EDA), data visualization, and statistical computing—are central to the digital humanities. We focus first on how EDA is obscured, but undergirds quantitative work in the digital humanities. We then show how data visualization offers a theoretical framework for understanding graphics as a form of argument with which the field should contend. Finally, we turn to the role of programming libraries to suggest we consider statistical computing in lieu of one-off tools.

While often not named, exploratory data analysis permeates the digital humanities. EDA, described in 1977 in John Tukey's seminal work by the same name, is a subfield of statistics offering a conceptual structure for drawing information from data sources. It often evokes new hypotheses and guides statisticians toward appropriate techniques for further inferential modeling. Tukey championed these techniques after becoming concerned that a substantial amount of statistical analysis was conducted by blindly applying models without first considering their appropriateness. His work, and its subsequent expansions over the past forty-five years, helps lay out an approach for thinking critically about data analysis in the humanities.

EDA is concerned with exploring data by iterating between summary statistics, modeling techniques, and visualizations. Summary statistics offer a quick description of the data using techniques such as mean, median, and range. A common approach is the five-number summary, which gives the minimum, lower quartile, median, upper quartile, and maximum values for a given variable. These compactly describe a robust measure of the spread and central tendency of a variable while also indicating potential outliers. Statistical models such as linear regression offer more complex summary statistics that describe general multivariate patterns in the data. Visualizations, as Turkey argued, can then be used to compactly augment summary statistics. Histograms, for example, give more detailed descriptions of the distributions of numeric variables. Scatterplots and box plots show patterns between variables, a strategy particularly important when simultaneously analyzing multivariate data. EDA undergirds quantitative work in the digital humanities, particularly text analysis, and yet is rarely explicitly named as a key method or identified as a powerful approach in the field.

Our work on Photogrammar offers an example of how EDA reshaped a digital public humanities project by leading to new inferences and questions. Photogrammar is a web-based site for organizing, searching, and visualizing 170,000 photographs taken from 1935–1945 under the auspices of the U.S. Farm Security Administration and Office of War Information (FSA-OWI). After acquiring the data from the Library of Congress, we turned to EDA using the programming language R. Calculating, assessing, and then reanalyzing summary statistics of photographs by state and by year revealed new and unexpected patterns, immediately shifting our object of study (see Figure 24.1; Arnold et al.). With an immediate signal from the data, we then used visualizations to convey our new arguments about the collection (see Figure 24.2). The new findings, revealed through EDA, changed the techniques of Photogrammar and opened up a new set of questions about the expanded visions of the archive, the federal government, and photographers.

Data visualization, another subfield of statistics closely aligned with EDA, is also making theoretical and applied contributions to the digital humanities. Tukey's work on EDA called for constructing visualizations, such as scatterplots, histograms, and box plots, to understand patterns within multivariate datasets. The field of data visualization builds on his work, treating visualizations as their own object of study.

Jacques Bertin's *Semiology of Graphics,* William Cleveland's texts *The Elements of Graphing Data* and *Visualizing Data,* and Leland Wilkinson's *The Grammar of*

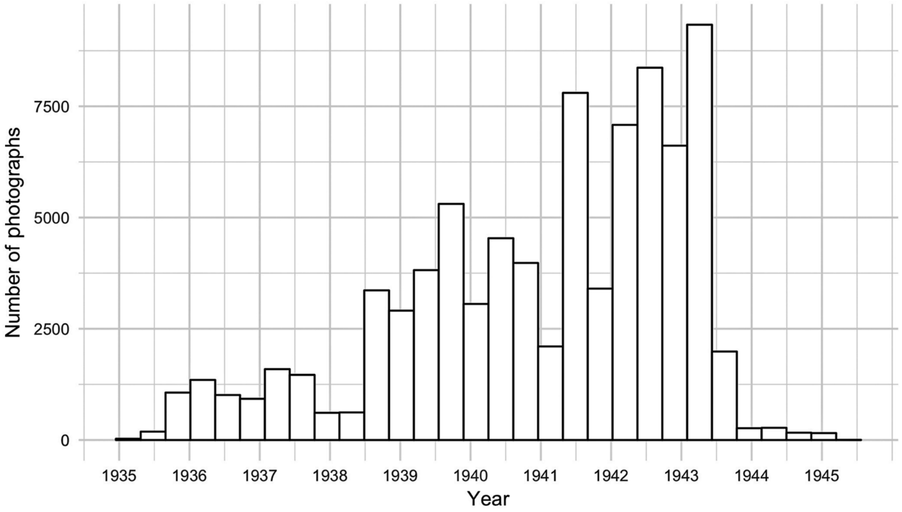

Figure 24.1. Histogram showing number of photographs taken over time in the FSA-OWI collection.

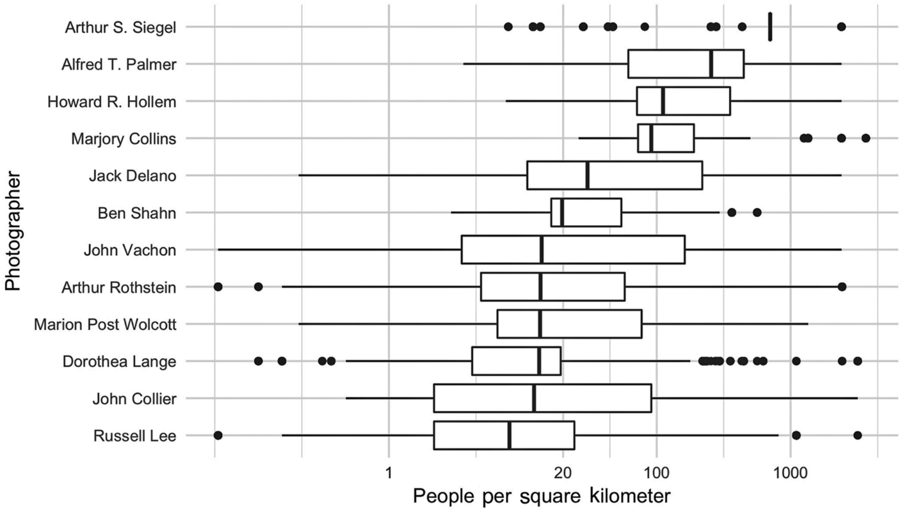

Figure 24.2. Box plot showing the population density, from the 1940 U.S. Census, of counties where FSA-OWI photographers captured photographs.

Graphics are foundational works in statistical graphics. Bertin argued that graphics "constitutes one of the basic sign-systems conceived by the human mind for the purposes of storing, understanding, and communicating essential information" (2). He showed explicitly how graphics can serve as their own form of argument and evidence. Cleveland built on Bertin's work by theorizing a distinction between graphing, the process of visualizing raw data, and fitting, the process of visualizing transformed data and statistical models. Wilkinson extended Cleveland's theory by distinguishing between the mathematical abstraction of a graph and the physical manifestation of a rendered graphic. He then set out to describe the fundamental units that comprised a visualization.

Wilkinson constructed a formal language for describing statistical visualizations by separating out the mathematical specification of a graphics system from the aesthetic details of its assembly and display. He named each component of the visualization, moving from the original data to the output, a layer in his formal *Grammar of Graphics* system. Examples of layers include picking the scale of the plot, choosing the size of points, and computing summary statistics. Take, for instance, a histogram of the number of photos taken each month in Photogrammar. One of the layers in Wilkinson's grammar would be calculating the heights of the bins. His goal was to make explicit the action of determining the height, which in turn makes the modeling assumptions explicit. Wilkinson's formal language explicates what assumptions are being made, where these assumptions are being made, and how the original data has been modified to create the output. Furthermore, implementations of this language as an object-oriented system, as in Hadley Wickham's ggplot2 library, are able to construct rules for fitting together a small number of classes and relationships between them to create an "almost unlimited world of graphical forms" (1).

The theoretical formulation of graphics provided by statistics offers concepts ideas for the digital humanities to draw from and grapple with. The work of Wilkinson and his interlocutors provides an alternative foundation for graphics that is a critical, but overlooked, area in our conversations about visual forms of knowledge production. Since Wilkinson's approach separates each action into small discrete steps, humanities scholars can analyze and critique the parts individually. Understanding these theoretical frameworks allows for critical applications of graphics engines such as ggplot2—gg stands for grammar of graphics—and D3.js. Finally, a dialogue between statistics and the humanities offers another framework from which to debate, construct, and critique how visualizations are approached and implemented in the digital humanities.

The subfield of statistical computing, which is concerned with the design and implementation of software for working with data, also offers methodological inventions for DH practice. Many tools written specifically for applications in the humanities such as Palladio, SHANTI, TOME, and WordSeer have the potential to unlock and analyze latent patterns in humanities data. While we are beginning to see a shift, a major funding priority continues to be using one-off tools for a specific task such as Voyant, or omnibus tools that attempt to address multiple analytical approaches in one toolkit such as Palladio. These tools can serve as a great introduction to analytical approaches in the field such as text analysis while offering out-of-the-box visualizations. Yet, they are limited because they only allow users to access a small set of predefined summary statistics, models, and visualizations. One would be hard-pressed to find a tool that provides the flexibility humanists need. As a result, while existing tools provide useful insights and can be particularly powerful for teaching, the full range of summarization processes called for by EDA is difficult to accomplish within the scope of most DH tools. Stand-alone programs lock users into a specific set of programmed summary statistics and visualizations, making the iterative exploration of data quite limited. In the process, the plethora of tools obscures the field's dependency on and adoption of EDA from statistics.

In contrast, libraries written to function within general-purpose programming languages allow users to move between many different visualization and visualization engines while simultaneously cleaning and manipulating the underlying dataset. They thus help avoid privileging preconceived notations of what may be of interest in a particular set of data. Popular programming languages for working with data, such as R and Python, have thousands of user-contributed libraries for data analysis, manipulation, and visualization.[2] These libraries all operate on the same basic internal data structures, allowing for easy interoperability between libraries. Statisticians are able to maintain flexibility while conducting data exploration as a direct result of these libraries that operate with a single programming language, rather than using disconnected one-off tools.

Using a general-purpose statistical programming language greatly increases the available set of methodological approaches to studying humanities data. Many statistical techniques such as unsupervised learning and dimension reduction are only available with use of a language such as R or Python. Topic modeling, for example, is such a ubiquitous method in the digital humanities for understanding the structure of a corpus of texts that newcomers would not be remiss for thinking that it is the only such method. Unsupervised learning, however, can be applied to learn and visualize different kinds of latent structures in texts. For example, spectral clustering, an unsupervised method for categorizing data, produces a complete hierarchical structure of a corpus. This hierarchy can be visualized in many ways and is arguably a far more appropriate method for studying many corpora. Unlike latent Dirichlet allocation, the primary technique for topic modeling, spectral clustering requires no external tuning parameters, always returns the same result, and does not require the user to specify the number of latent topics. Techniques such as the singular value decomposition, or dimension reduction in general, offer further approaches to visualizing the structure of texts without forcing documents or topics into discretized buckets.

Although these libraries require significant technical skill to apply them to any particular dataset, the ability to move nimbly between multiple methods and tools can give humanists the flexibility and nuance they seek. For this reason, we should consider modifying or building on top of extant general-purpose libraries, many of which are open-source libraries. By freeing the time that is now spent appropriating existing frameworks and libraries, DHers can concentrate on developing new modifications for extracting humanities knowledge. As an example of how this can work successfully, consider Lincoln Mullen's tokenizers package, which simplifies the process of turning raw text into tokenized text within the R programming language. Humanists can now apply dozens of other visualizations and models to their own textual data by way of other well-maintained R packages.

Making use of general-purpose software libraries requires an important shift in DH pedagogy. To fully utilize statistical analysis, one needs to be proficient in programming with a language such as R or Python. One does not need to create R packages, but rather one should be able, ideally, to read and implement existing code to explore, manipulate, and analyze humanities data. Otherwise, users are constrained to exploring a small preset selection of visualizations and have limited avenues for cleaning and summarizing their data through point-and-click tools. Writing in a programming language also allows for summarizing and publicizing analyses with documented code. Such an approach helps other scholars replicate our studies on datasets as well, providing transparency and reproducibility. At the same time, external tools are useful for running efficient interactive visualizations, and their limited toolsets are often ideal for initial data analysis and curated, public-facing projects. However, they are not well positioned to be the only mechanism for running the iterative experimentation required by exploratory data analysis or to give the full range of possibilities for data visualization.

While it is important for digital humanists to become conversant in the theory and application of data analysis, statistical visualizations, and high-level programming languages, collaborations with statisticians and other computational area experts will be instrumental for pushing advanced computational work in the digital humanities. No one person can be expected to understand such a wide range of disciplines brought together in the digital humanities in the depth required to develop the innovative insights and methods that are the promise of the field. Rather, the digital humanities should welcome statistics to the table, and this begins by better acknowledging the critical role of statistics in the field.

NOTES

1. The focus on tools is exemplified by National Endowment for the Humanities Office of Digital Humanities grant applications and awards, which have a significant focus on tool building.

2. The history of R in particular is deeply entwined with exploratory data analysis and data visualization. It was originally developed as the S programming language in the mid-1970s at Bell Labs by Rick Becker, John Chambers, and Allan Wilks (Becker and Chambers). It was built to support the kind of fast, iterative data analysis favored by John Tukey, who also had an appointment at Bell Labs during the same time. The ideas behind EDA and the formalized structure for statistical graphics all filter up to the R programming environment.

BIBLIOGRAPHY

Arnold, Taylor, and Lauren Tilton. *Humanities Data in R.* New York: Springer International, 2015.

Arnold, Taylor, Lauren Tilton, Stacey Maples, and Laura Wexler. "Uncovering Latent Metadata in the FSA-OWI Photographic Archive." *Digital Humanities Quarterly* 11, no. 2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000299/000299.xml.

Becker, Richard, and John Chambers. *S: An Interactive Environment for Data Analysis and Graphics.* Boca Raton, Fla.: CRC Press, 1984.

Bertin, Jacques. *Semiology of Graphics: Diagrams, Networks, Maps.* Madison: University of Wisconsin Press, 1983.

Cleveland, William. *The Elements of Graphing Data.* Monterey, Calif.: Wadsworth Advanced Books and Software, 1985.

Cleveland, William. *Visualizing Data.* Hobart Press, 1993.

Drucker, Joanna. *Graphesis: Visual Forms of Knowledge Production.* Cambridge, Mass.: Harvard University Press, 2014.

Mullen, Lincoln. "tokenizers: A Consistent Interface to Tokenize Natural Language Text." R package version 0.1.4., 2016.

Posner, Miriam. "Humanities Data: A Necessary Contradiction." June 25, 2015, http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities* 2, no. 3 (2013), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

Tukey, John Wilder. *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley, 1977.

Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis.* Berlin: Springer Science & Business Media, 2009.

Wilkinson, Leland. *The Grammar of Graphics.* Berlin: Springer Berlin Heidelberg, 1999.