
Reviewed by Spencer Dean Stewart (Purdue University)

Published on H-History-and-Theory (March, 2024)

Commissioned by Andrew J. Kettler (University of South Carolina)

It is an exciting time for the field of digital history. New journals have emerged to publish digital scholarship while established venues are beginning to feature digitally informed work.[1] A growing number of monographs and edited volumes that use digital methods, such as mapping and computational text analysis, have appeared in recent years with some winning major awards.[2] And textbooks and online tutorials serve as valuable resources for analyzing historical data.[3] Even if “distant reading” and quantification aren’t new to history, a growing interest in using large data sets to study broad historical trends demands greater reflection.[4] Jo Guldi’s new book, *The Dangerous Art of Text Mining: A Methodology for Digital History,* provides this and more by establishing a theoretical and methodological foundation for future work at the intersection of history and text analysis.

While the main audience for this book will likely be those interested in digital humanities and digital history (or what Guldi more precisely calls “text mining for historical analysis”), its scope and implications go beyond these fields. *The Dangerous Art of Text Mining* is an ambitious work that demonstrates how quantitative methods can inform the study of the past while also making a compelling case for how historical methods should inform the work of data science. Guldi outlines a middle ground based on “hybrid thinking,” where historians and data scientists can meet to think critically about how data is used and abused to understand the past and present.

The book is divided into three sections. Part 1, “Toward a Smarter Data Science,” is a call for data scientists in a “world without history” to approach data with the eye of a historian. Many principles covered here will be familiar to those with training in history, but these chapters are useful for two reasons: to reassure the skeptical historian that text mining need not abandon critical historical thinking and to demonstrate that historical
methods can enrich the work of data science. Data scientists need to recognize that historical data is incomplete and plagued with bias, prejudice, and falsehoods. Such issues require a level of familiarity with primary sources and the historical events these sources claim to represent. Extracting word counts is easy. Knowing what words to search for and how to interpret them is a challenge that historians are well equipped to address.

Although many of the issues outlined in this section can be resolved through closer collaboration between historians and data scientists, there remains one division that Guldi suggests will be difficult to overcome: the pressures of fantasy and prediction. Whereas the data scientist might use historical data to make predictions about the future or to uncover “natural laws” that govern how societies function, historians are instead interested in complexity, contingency, and the social construction of laws in the past. When done carefully, algorithms can be powerful tools for studying bias and omission. However, predictive thinking has often resulted in the naive implementation of oversimplified models to the detriment of society. It has stripped historical data of its complexity and nuance, thereby perpetuating past problems and prejudices into the present and future.[5]

Part 2, “The Hidden Dimensions of Temporal Experience,” brings together the philosophy of history with computational text analysis to ask how a theoretically driven approach to text mining might enrich historical scholarship. The chapters demonstrate that different approaches to “distant reading,” like topic modeling, word embeddings, and named entity recognition, can provide new insights into frequently asked questions about temporality, memory, influence, and modernity. Guldi succeeds in providing a gentle but thorough introduction to text mining by reviewing a wealth of secondary scholarship and applying these different methods to a study of British parliamentary debates and US congressional speeches.

Chapter 12 is a useful example. Here Guldi uses word embeddings, word counts, and close reading to analyze shifting congressional discourses on climate change in the United States from 1970 to 2009. These methods allow her to uncover how references to “environmentalist” changed over time, to identify moments of heightened hostility toward the environmentalist movement, and to propose possible explanations for growing skepticism about climate change. By combining qualitative and quantitative methods, Guldi pinpoints a moment circa 2001 when key individuals shifted their rhetoric from “extreme environmentalism” to “radical environmentalist,” a shift that was part of “an attempt to paint environmentalists as enemies of the state” (p. 397). Guldi’s rich engagement with text and method results in an innovative approach to studying the history of ideas and the politics of science.

Part 3, “Disciplinary Implications,” outlines Guldi’s vision for the future of digital history by first looking at historical linguistics, a field that overcame many harmful and inaccurate views of world languages through the use of data, analysis, and visualization. It is within this context of the potential value of computational analysis that Guldi envisions the emergence of “cyborg historians” (part traditional historian, part data scientist) who will advance research in both history and data science. This hybrid approach will require the field to rethink graduate education and to place greater value on collaboration and coauthorship.[6] It will also require us to think carefully about the role of critique as errors arise. To avoid controversies that have plagued other fields, such as literary studies, Guldi advocates for an approach to critique that is “rigorous enough to invite constant improvement, yet not so strict that it penalizes any single scholar for failing at collective standards that can only be realistically accomplished by the labor of many hands.” A productive critique, Guldi continues, should “be specific: not an indictment of quantitative methods as a whole,
or of the labor of all of the hands that touched the field” (p. 450).

A couple minor notes. First, readers will likely find some chapters to be more accessible than others. When used in the classroom, individual chapters are great at facilitating discussion and demonstrating the possibilities of text mining. But like other books in this genre, the methods can at times be overwhelming to the uninitiated. Relatedly, this isn’t a textbook. The purpose of the book isn’t to teach the reader how to apply computational methods to new datasets. Its value is instead as a work of theory and as a meditation “on the fit between code, questions, and data” (p. 448).

There is an ever-growing volume of digitized primary sources and born-digital data that are waiting to be mined for historical insights. The question remains: will historians take a backseat to data science, or will we actively contribute to ongoing research? *The Dangerous Art of Text Mining* provides an effective framework for the latter by not only demonstrating the value of text mining for historical analysis but also revealing how historians can contribute to broader debates about data and method.

Notes

[1]. New journals in digital history include *Current Research in Digital History* and *Journal of Digital History*. An example of existing journals recently featuring digital methods is the *American Historical Review* History Lab.


If there is additional discussion of this review, you may access it through the network, at https://networks.h-net.org/h-history-and-theory


URL: https://www.h-net.org/reviews/showrev.php?id=60563

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License.