# **DHQ: Digital Humanities Quarterly**

2022 Volume 16 Number 4

## The Banality of Big Data: A Review of *Discriminating Data*

Amanda Furiasse <afuriasse\_at\_gmail\_dot\_com>, Nova Southeastern University ២

#### Abstract

This review critically interrogates Wendy Chun's book *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition* (MIT Press, 2021) from the perspective of the digital medical or health humanities. The monograph's exploration of predictive machine learning and big data's propensity to encode segregation through their default assumptions about correlation raises important questions about machine learning's growing uses in fields, such as medicine and pharmacology, where the stakes of such digital experimentation are particularly high. Chun's exploration of the predictive processes by which data analytics replicates 20th-century eugenics discourses makes an important contribution to the field of digital medical ethics and also offers unique insight into the mechanisms by which digital humanities scholars can disrupt and challenge the use and application of such predictive programs.

## Introduction

On April 14, 2003, the International Human Genome Sequencing Consortium, a collaborative project in the United States between the National Human Genome Research Institute and the Department of Energy, announced the successful completion of the Human Genome Project. It entailed developing a complete sequence of the Human Genome, a task made possible by 800 interconnected Compag Alpha-based computers performing 250 billion sequence comparisons an hour [Bergeron and Chan 2004]. The torrent of data produced by the project not only revealed the incredible complexity animating human life but also and more importantly, demonstrated how machinic systems could be used to both reveal and even reverse engineer the patterns and trends driving such large data sets, giving humanity an almost god-like power over nature. As President Bill Clinton announced: "Today we are learning the language in which God created life. We are gaining ever more awe for the complexity, the beauty, the wonder of God's most divine and sacred gift. With this profound new knowledge, humankind is on the verge of gaining immense new power to heal" [The New York Times 2000]. Thus began the era now known as "big data" and with it the assumption that the analysis of large data sets would allow humanity to reveal and fundamentally alter the patterns and trends driving the natural world. Nearly two decades after the Genome Project's success, the promise of big data, however, seems more like a nightmare. In her new book Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition, Wendy Hui Kyong Chun delves into this nightmare and argues that heightened polarization during the era of big data is not an error or flaw within the system [Chun 2021]. Rather, Chun makes the argument that discrimination is in fact big data's primary product.

1

2

## **Correlating the Data**

As with Chun's previous contributions, *Control and Freedom* (2008), *Programmed Visions* (2011), and *Updating to Remain the Same* (2017), Chun investigates a fundamental paradox and contradictory tension at the heart of new digital technologies. Machine learning algorithms are often marketed as free from human bias and prejudice, yet they simultaneously encode segregation into the very logic by which they promise to reverse engineer life. As Chun explains, "We need to understand how machine learning and other algorithms have been embedded with human prejudice and discrimination, not simply at the level of data, but also at the levels of procedure, prediction, and logic" [Chun 2021, 16].

Prejudice then is not a problem that results from human error, but it structures the basic logic animating our machinic systems.

The monograph's argument unfolds around four foundational concepts in computing: correlation, homophily, authenticity and recognition. Correlation remains among the most crucial concepts undergirding nearly every aspect of existing Al systems. According to Chun, correlation is not just a conceptual category, but it constitutes an everyday practice whereby people are lumped into "categories based on their being 'like' one another amplifying the effects of historical inequalities" [Chun 2021, 58]. These inequalities are in turn naturalized with data organization systems making it appear as though they are innate or *sui generis* categories which already preexist in the world. As Chun warns, "correlation contains within it the seeds of manipulation, segregation and misrepresentation" [Chun 2021, 59]. As a result of their reliance on correlation, social networks create "microidentities" by default which instrumentalize and weaponize individual differences. Data analytics consequently reimagines eugenics discourses within a big data future where correlations are not only assumed to be predictive of future outcomes, but surveillance is assumed to be a necessary component of every human institution and one which will allow humanity to improve nearly every component of daily life.

4

5

6

All three other foundational concepts are conceptually linked to correlation. Homophily, for example, plays a crucial role in naturalizing correlations, making likeness seem like an obvious and innate way to group and organize data. The assumption that people gravitate toward things that are like them consequently becomes "a tool for discovering bias while simultaneously perpetuating those very biases in the name of 'comfort,' predictability, and common sense" [Chun 2021, 85]. Chun in turn stresses how the very feeling of comfort which homophily generates naturalizes these acts of discrimination. Like homophily, authenticity has become automated in that authentication of one's identity within systems of validation are assumed to be a necessary and nonnegotiable component of digitization. Before you log in, you must authenticate yourself and essentially prove yourself to be real. Algorithmic authenticity is rarely if ever challenged. Instead, we are "trained to be transparent" [Chun 2021, 24]. Finally, recognition makes discrimination possible, since it encompasses the process whereby we come to accept someone or something's status and authority as authentic.

#### Assessing the Costs of Algorithmic Performativity

Although Chun's four foundational concepts might organize all computational systems, they are ultimately dependent upon our performance of them. Throughout the monograph, Chun emphasizes that data analytics requires our performance of it, since it is only through our performance of the roles that data analytics assigns us that it comes to take on its meaning. Here, Chun makes the most significant and deeply troubling argument in the monograph. Through our active participation in it, data analytics transforms the world into a research laboratory where we are its main research subjects. The rhythms of life are consequently transformed into significant patterns that reveal some sort of deeper order that can be reconfigured to improve social outcomes. People who fail to follow their algorithmically-assigned roles are either dismissed as meaningless outliers or at worse categorized as deviant actors and subject to swift disciplinary action. However, deviance, as Chun warns, is also pre-scripted and performed, since the internet requires offensive and deviant content to hold people's attention. Deviations are consequently the very way by which users come to be authenticated and clustered [Chun 2021, 332].

For Chun, the transformation of society into an experimental lab is one which bears explicit similarities to one of the darkest and most troubling eras of human history: eugenics. While often dismissed as a relic of a bygone era in human history, eugenics is in fact increasingly coming to undergird contemporary institutions. From medicine to education, the assumption that people can be divided into certain groups based on certain perceived behavioral or assumed physical characteristics and then those groupings can be used to evaluate and predict future behaviors and outcomes is now the underlying logic which guides human decision-making and institutional structures. As Chun warns, "If twentieth-century eugenicists however defended their work against accusations that it experimented on humans, twenty-first-century data scientists openly embrace experimentation" [Chun 2021, 158]. Their unyielding belief in the inherent benevolent power of experimentation is why firms like Cambridge Analytica have been able to experiment on the public relatively free of oversight or regulatory pressures, despite the fact that the firm's claim that they were able to predict and shape voting patterns in the 2016 election was never verified and no evidence was ever provided to verify their claims of success.

While Cambridge Analytica might have not been able to prove their claims, that has not stopped nor mitigated data metrics' use in other fields such as medicine where the stakes of such reverse engineering experiments are particularly high. Postmarket surveillance, for example, was once a practice unique to digital technology companies but in recent years has become a standardized practice by which medical and pharmaceutical companies can expedite the lengthy and costly process of developing new drugs and increase their profitability. Chun echoes Olivia Banner's cautionary tale about medicine's increasing reliance on this practice in *Communicative Biocapitalism* [Banner 2017]. Once relegated to decades of costly and lengthy trials, today new drugs are commonly studied only after they have been approved for the public, leading to drugs later being pulled from the market after causing significant and lasting damage to patients. The adverse effects of OxyContin, for example, were only discovered years after its approval to the general public. Such examples demonstrate how our growing faith in the power of big data analytics in fields such as medicine is ultimately predicated on our underlying faith that the benefits of experimentation always outweigh its potential costs.

7

8

9

10

11

### **Rethinking the Role of the Digital Humanities**

Chun's description of the apparent parallels between eugenics and big data surveillance mechanisms make the monograph a particularly important read for digital humanities scholars, particularly those who work in the digital medical or health humanities. In particular, Chun's five-step program to counter data-informed discrimination and repair "the mistakes of our discriminatory past" [Chun 2021, 2] offers a possible set of responses for scholars interested in working toward more socially and culturally conscious ways of designing and working with machinic systems. According to Chun, being more socially and culturally conscious ultimately entails critical interrogation of the historical and political conditions which propel and inform the logic underlying our algorithms and data structures. This process of interrogation, however, hinges on embracing our moral responsibility in participating in these experiments.

As Hannah Arendt once argued, eugenics was able to find a home in German society, because it was so banal and deflected individual moral responsibility to institutional structures [Arendt 1963]. In the case of Nazi officers like Eichmann, Arendt explains that they deflected moral accountability for their active participation in the murder and torture of millions of people by claiming that they were just following administrative commands and orders from higher up. Yet, Arendt aptly explains that one's moral responsibility only increases in such situations where you might not know the victim nor personally be the one who carried out the execution. Responding to Eichmann's defense, the judgment of the court ruled "On the contrary in general the degree of responsibility increases as we draw further away from the man who uses the fatal instrument with his own hands" [Arendt 1963, 247]. Likewise, are we just following the data? Put differently, if a decision is made on the basis of data metrics, who exactly bears moral responsibility for its outcomes?

Although machine learning and Al systems might have ushered in a new historical era constituted by the promise of reverse engineering social systems, this era is one which carries with it many of the same moral problems of the one that preceded it. Perhaps more is needed than just critical interrogation and awareness of the propensity for algorithms to reorganize social institutions around polarizing divides. Such new technological innovations reveal the need for a new moral system, one with the potential to hold individuals and organizations accountable for the outcomes of their digital experiments on social, political, and economic institutions. If board members and executives of a company were held morally responsible for the outcomes of their use and application of data metrics, would they be more likely to critically interrogate the assumptions behind those metrics? How exactly can scholars break up this synergistic nexus between the corporate boardroom, Congress, and laboratory?

Ultimately, Chun's monograph exposes the urgency for more investment in the digital humanities, particularly when it comes to the study of ethics. As the digital humanities comes to define itself as a field, ethics must remain a nonnegotiable and fundamental component of the field's central task and purpose. The development, creation, and implementation of a moral system which holds people accountable for the outcomes of their data metrics, modeling, and social experiments must take precedence in the field as the digital humanities works toward the development and application of more socially and culturally conscious ways of using and applying digital technologies. We might not know what those moral systems look like in practice today, but the field is already increasingly coalescing around critical assessments and interrogations of the material and human costs of digital technologies as awareness and evidence of digital media's costs continues to accumulate. The era of big data might have given humanity god-like abilities, yet

perhaps it is the task of the digital humanities to expose how those god-like powers have made us even more flawed and vulnerable than we were before.

#### **Works Cited**

Arendt 1963 Arendt, H. Eichmann in Jerusalem: A Report on the Banality of Evil. Penguin (1963).

- **Banner 2017** Banner, O. Communicative Biocapitalism: The Voice of the Patient in Digital Health and the Health Humanities. University of Michigan Press, Grand Rapids (2017).
- Bergeron and Chan 2004 Bergeron, B. and P. Chan. *Biotech Industry: A Global, Economic, and Financing Overview*. John Wiley & Sons, Hoboken (2004).
- Chun 2021 Chun, W. H. K. Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition. MIT Press, Cambridge (2021).
- **The New York Times 2000** The New York Times. "Text of the White House Statements on the Human Genome Project." New York (27 June 2000). https://archive.nytimes.com/www.nytimes.com/library/national/science/062700sci-genome-text.html.



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.