From Galton's Pride to Du Bois's Pursuit: The Formats of Data-Driven Inequality

Colin Koopman [iD]
University of Oregon

Abstract
Data increasingly drive our lives. Often presented as a new trajectory, the deep immersion of our lives in data has a history that is well over a century old. By revisiting the work of early pioneers of what would today be called data science, we can bring into view both assumptions that fund our data-driven moment as well as alternative relations to data. I here excavate insights by contrasting a seemingly unlikely pair of early data technologists, Francis Galton and W.E.B. Du Bois. Galton, well known for his contributions to eugenics, was first and foremost a tinkering technician of measure. There are numerous domains of science over which Galtonian conceptions retain considerable influence, presumably without his pride in racial inequality. A more viable, because more egalitarian, alternative for the present can be found in the early data work of Du Bois.

Francis Galton is today remembered as an eminent prince of science if not also as a wild crank who, at the end of the 19th century, helped spawn the grandiose political disaster of eugenics, the ripples of which would continue to be felt far into the 20th century. More than prince or crank, Galton was first and foremost a tinkering technician of measure. He was engineer of a variety of means of tallying and quantifying, if not also at times a major theorist of all that measure adds up to. All of Galton's projects, from his scientific contributions to statistics to his politics of eugenics, depended on technologies of measure. Stephen Jay Gould (1996 [1981]) observes that, 'Quantification was Galton's god,

Corresponding author: Colin Koopman. Email: koopman@uoregon.edu
TCS Online Forum: https://www.theoryculturesociety.org/

and a strong belief in the inheritance of nearly everything he could measure stood at the right hand' (p. 108). Galton (1879) himself was equally explicit: 'until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science' (p. 149).

In his 1992 essay 'Galton's Regret' the anthropologist of science Paul Rabinow considers Galton's own dissatisfaction at his attempts to leverage his beloved measures toward the scientific implementation of racism. His project in eugenics, essentially an attempt to give racism the inflection and imprimatur of biological and statistical science, certainly persisted even beyond his death. Yet it never really gained the scientific stature which Galton had hoped for it. Rabinow (1996 [1992]) observes that this failure 'constituted for the Victorian founder of eugenics a major disappointment' (p. 114) – Galton's regret.

Rabinow (1996 [1992]) also observes how Galton's unrealized dream persists in ongoing efforts to articulate the high-tech bio-sciences of today, namely genetics and genomics, to 'older cultural understandings of race, gender, and age' (p. 127). Rabinow's concern was that when we translate the bio-sciences onto cultural configurations steeped in long histories of inequality, we risk producing new iterations of those inequalities even where we explicitly want to be pursuing egalitarian goals. In returning to Galton once again, I explore how Rabinow's concern remains salient across an expanding field of scientific and cultural practices. I here seek to animate that concern for other domains of data-driven science extending beyond the pre-eminent bio-sciences of genetics and genomics.[1]

Galton is eminently approachable as a figure in the history of our present, to draw on a methodological term of art of Michel Foucault's, whose work has been a serious influence for Rabinow and also for myself (Foucault, 1995 [1975]: 31). There are numerous domains of science over which Galtonian conceptions of measure retain considerable influence. Not least among these is a host of efforts in contemporary data science deeply reliant upon the kind of informational infrastructure Galton helped to develop. As critical data studies scholars have shown, such recent deployments of data science risk a bevy of injustices, variously conceptualized as 'automated inequality' (Eubanks, 2018), 'default discrimination' (Benjamin, 2019), 'algorithmic oppression' (Noble, 2018), and 'discriminating data' (Chun, 2021).

As one should expect of any history of the present, Galton remains a complicated predecessor for our 21st-century data politics. His efforts in the 1880s and 1890s anticipated (but by no means fully realized) what I have elsewhere conceptualized as the emergence of full-scale 'informational persons' and their 'infopolitics' in the 1920s and 1930s (Koopman, 2019a). Our data politics today are steeped in nearly one hundred years of infopolitical formats that fasten us to all manner of systems of datafication (see Koopman, 2021). Galton's information technology is part of the pre-history of how such a politics of data first emerged in the generation following him.

In what follows, I focus specifically on a Galtonian legacy that remains with us today. This legacy continues to take pride in data but has also tried to pay its dues by disowning Galton's regret. This should not sound altogether unfamiliar. It is the self-image operative in many contemporary deployments of data-driven inquiry, including much data science: a deep reliance on data, a use of those data for projects in (among other things) social amelioration, and an innocent disavowal of such odious bigotries as racism.

The contemporary self-image of innocent data innovation needs to be interrogated through the critical distinction between structural racism (sometimes referred to as institutional racism) and attitudinal racism (sometimes simply called racial prejudice).[2] This distinction demands of us that we interrogate our practices not just for overt racist attitudes (of which the contemporary virtuous citizen is presumably not possessed) but also for covert structures that reproduce racialized patterns of inequality. A concept of structural racism alerts us to the possibility that ongoing racial inequality is not just an unwanted after-image of centuries of intentional racial domination but is more pressingly an integral part of the practical functioning of our social institutions.

Leveraging this distinction, I here interrogate the possibility that among the many elements of our inegalitarian social structure are information technologies that directly contribute to the reproduction of inequalities. It is a crucial question for a highly technologized society such as ours whether technologies are significant components of social structure.[3] I here respond to this question by taking it up with respect to the specific issue of how social structure reproduces the distribution of racial inequalities through information technologies.

A summarizing frame for my approach is offered in André Brock's *Distributed Blackness* when he asks us to consider 'how whiteness structures application design' in such a way that 'the coders and engineers of Silicon Valley could be disabused of the notion that they are creating applications and software for "everyone" rather than for themselves' (Brock, 2020: 38, 242).[4] If Silicon Valley is not quite a hotbed of attitudinal racism, its products nonetheless offer ample evidence of a widespread and unacknowledged presumption of whiteness in its users, its platforms, and its very data structures. But since this presumption is hardly ever given overt recognition, it is difficult to locate it as an attitude. Thus Silicon Valley's 'racism-without-racists' is more particularly manifest in the technological dimensions of its structural racism (Brock, 2020: 155). Therefore, I argue, we need to ask whether, and more importantly *how*, data technology is a crucial component in the structural maintenance of inequalities of race.

This argument challenges the familiar and comforting self-image we in the data society tend to indulge. Restricting ourselves to only the most odious form of racism, that of attitudinal racism, we tell ourselves that we can keep separate what Galton could not. We tell ourselves we are not racist after the model of Galton's despicable attitudes that motivated his programs in eugenics. We think of our pride as unencumbered from prejudice. And with that we take our duty to be discharged. We move forward in confidence that we are deploying our information technologies without exacerbating racial inequality. What is stunning about this self-image is simply that we also know for a fact that even where racial prejudice is rejected, there nevertheless persist deep racial inequalities. My aim here is to consider the possibility that one explanation for this is that we have yet to seriously interrogate other of Galton's prides that we have inherited. By historically excavating these aspects of Galton's work below I hope to open up space for broader visions of how we remain unwantingly and unwittingly entangled in racial inequality.

For we do have alternative precedents who exemplify quite different data pursuits that arc toward what I shall call 'data equality'. One such precedent was a transatlantic contemporary of Galton's, W.E.B. Du Bois. In considering this pairing, it is important that we regard Du Bois's strategies as more than just a counterpoint to Galton's pride. Du Bois's work more importantly forms a challenge to our own ongoing presumptions of

innocence about the data-driven inequalities we unwittingly perpetrate. By resolutely facing this challenge, his work models how to fulsomely pursue equality in our deployments of data, and specifically how such pursuits require vigilant attention to what I have elsewhere called the 'formats' that condition all information.[5]

In presenting contrastive formatting technology in Galton and Du Bois, the primary exhibits I will offer are informational apparatus that each crafted for international expositions in 1884 and 1900, respectively. These exhibits add further detail to Shannon Mattern's (2020) recent survey of the 'spectacles of data' showcased at world's fairs from the late 19th to mid-20th centuries. When we witness such spectacular data apparatus, their clean functionality easily dazzles us into the comforting belief that they are neutral. To disabuse ourselves of such pretenses, we must consider not just those past data programs that continue to constitute the present but also those that have for too long been left idle and underutilized. The Foucauldian media-archaeological methods I deploy here aim at precisely this insofar as they help sharpen tensions between Galton's data-driven displays and Du Bois's underacknowledged efforts at manifesting different data.[6]

Galton's Pride

Galton's failures in his many attempts to marry measure to race was his prime regret. But measure ever remained his pride. He sought to apply his gauges to nearly all that he looked upon. And surely the greatest of Galton's (1884b) passions for what he called 'grasp and measure' was that of data concerning the influence of heredity (p. 180). In this, Galton was without doubt indulging in that temper of his age most famously associated with his older half-cousin, the evolutionary naturalist Charles Darwin (Galton was born in 1822, Darwin in 1809). The difference between the two is fine but crucial. Whereas Darwin seemed largely content to establish heredity as a vector of influence on the organism and the species, Galton relentlessly pursued the precise measurement of hereditary influence. This pursuit led Galton out of psychological and anthropological science proper into that erstwhile quasi-science of eugenics. For if one could measure degrees of hereditary influence, then one was in a position to submit those measures to comparison, and to detect thereby presumed hereditary differentiations. If Darwin taught his Victorians that the self is not wholly self-made, then Galton tried to teach them that the self is almost wholly made by others, but specifically by parentage.

'Parentage' was the title of the first chapter of Galton's autobiography, published in 1908, some 27 months before he died. After ten pages spent recounting his parents, grandparents, great-grandparents, uncles, aunts, and even his famous half-cousin (he notes of Charles's father Erasmus Darwin that, 'His hereditary influence seems to have been very strong' [Galton, 1908: 7]), Galton (1908) comes to his point: 'The general result of the foregoing is that I acknowledge the debt to my progenitors of a considerable taste for science, for poetry, and for statistics; also . . . a rather unusual power of enduring physical fatigue without harmful results' (p. 11). Galton's autobiography neatly performed his theory that family mattered much in the making of the self. And for Galton, all that matters can be measured, so that we can say just how much it matters.

Before considering in more detail below Galton's vision for the measure of family, I turn first to exhibiting some of the more general features of his technologies of measure.

These technologies remain with us in more ways than we commonly think. What I excavate from them is an account of how a series of technical and discursive projects intersected in an informational infrastructure capable of producing pursuits of inequality that remain resilient more than a century later. For Galton's technologies of measure constitute a technical apparatus that remains very much at the heart of contemporary efforts in fields such as data science. Today's data science computations may run on clustered microprocessors, but their basic operations remain informed by Galton.

Consider one of Galton's small but deep contributions to statistical reasoning, namely the concept of percentile ranking (he called them 'centiles').[7] Today it is routine for us to ask which percentile we fall into among a population measured on a standard scale (such as a GRE [Graduate Record Examination] score). These percentile rankings mean something to us almost automatically. Simple as the idea is, it is easy to forget that somebody had to come up with this method of classing scores. As Galton (1908) himself said: 'All this is an old tale now, but I had to take a great deal of trouble before it was clearly thought out and well tested' (p. 268). That it was Galton who took the trouble to do so is quite telling of his admiration for distinction. Galton was less interested in how high (or low) a person scored on a test than in how high (or low) their scores were relative to other test-takers. And ultimately, Galton's interest in statistics had less to do with averages represented in the bulging middle of a curve and more to do with deviations from the mean where the superlative comes into view.

Our inheritance of a statistical apparatus that relies much on such of Galton's technical innovations as percentile rankings, or statistical correlation as later perfected by his admirer and fellow eugenicist Karl Pearson, is a story already well known.[8] Critical media theorist Wendy Chun (2021) has recently offered a new perspective on this matter by emphasizing how Galton's statistical correlation serves to 'mak[e] the present and future coincide with a highly curated past' (p. 52). My argument here builds on insights about temporal closure from Chun and others.[9] But I refocus away from algorithmic correlation to gain a clearer view of other Galtonian innovations that I believe reach more deeply into our technological horizon. My working hypothesis is that Galton's persistence into our present may be better grasped through techniques less complicated than his statistical ingenuities.

I focus here on how today's data technologies rely on a series of technical operations that Galton was keenly aware of but which many data scientists today routinely neglect. This involves apparatus for the collection and normalization of those data that algorithms will later on be able to process. Alongside Galton's contributions to the algorithmic processing of correlations are equally important technologies for what I am calling 'formatting'.[10]

The concept of the format overlaps more frequently discussed figures for information technology, such as the algorithm and the measure. Yet the format differs in crucial ways. Concerning algorithms, it is not inconsequential that formats (i.e. data structures) must precede the manufacture of any data as such, and therefore necessarily precede any data processing (e.g. by algorithms).[11] Both matter much, but only algorithms garner the attention of today's tech boosters and critical data studies scholars alike.[12] Considering measures, we find in this case a more proximate technical operation to formats.[13] But formats have a wider scope, in that they concern not just quantitative structuring (e.g.

scales for a mental test) but also taxonomical (e.g. racial categorization schema) and even purely conventional (e.g. rules for human names) data structures.

The core technological apparatus for formatting, or the structuring of data, that is found throughout Galton's many projects is the deceptively simple one of the printed blank form.[14] A sheet of paper with rules and lines, a handful of pre-printed words describing categories, and a set of clearly defined blank areas (usually emphasized by underlining or a series of periods forming a kind of proto-under-line) which one is meant to fill out. We, and our various researchers (doctors, teachers, scientists), fill out the forms and we become thereby consistent correlates of the forms. We become persons in forms or in-form-ational persons. In Galton's day none of this had yet to consolidate. Galton's forms are for the most part still rudimentary prototypes of the complex, efficient, and rigorous devices that forms would soon become in the decades just after his death. He was part of the story of this consolidation, even if only a rather early part.

Perhaps one of Galton's most famous deployments of the form technology was at his Anthropometric Laboratory. First set up at London's International Health Exhibition in 1884, attendees paying threepence could walk through the laboratory, a 6-foot by 36-foot 'long narrow enclosure', to take a series of tests measuring such powers as eyesight, reaction time, span of arms, force of blow, and so on (Galton, 1908: 245). The exhibition closed the following year and Galton moved the laboratory to the South Kensington Museum where it operated for another six years. Most of the mental measures Galton was registering were by then already a well-known facet of German measurement psychology. Though interestingly enough, Galton himself had still devised much of the apparatus used in the tests, including 'small whistles with a screw plug for determining the highest audible note' (Galton, 1908: 247). But the most critical apparatus in operation there was Galton's deceptively humble blank form. The seemingly simple idea of giving a person a series of tests of bodily and perceptual ability assumed a new meaning once experts, scientists, and cranks of all kinds began recording those results on standardized printed forms. As Galton (1892) was finally closing up his laboratory he found it 'a great consolation' to receive 'on the very day that I began to dismantle it, the proof sheets of the register, and other forms in many respects like my own, that are to be used in the laboratory at Dublin' (p. 32). Galton (1885) reported that at the Health Exhibition alone 'the number of persons measured in the laboratory . . . was no less than 9,337, and each of them in 17 different ways' (p. 206). Each of them left with their own copy of the just-produced record of themselves as recorded on a printed blank form (see Figure 1).

Galton (1892) and his assistants retained a carbon copy of each record. And why? 'The data collected at my laboratory have been of service in many ways' (p. 32). In what ways? Galton calculated correlations among different bodily attributes (between lengths of different limbs, between stature and strength). These calculations led him to some of his important contributions to statistical theory, as noted above. But beyond statistics, Galton developed an interest in the use of these records as an index of who a person is.

For Galton, who a person is must be figured as an index of their heredity. In one among many discussions of what we now call the nature–nurture debate, Galton asserted that the signs of 'membership' in 'race' are 'partly personal, partly ancestral'. Yet he immediately rejoined that 'we need not trouble ourselves about the personal part' (Galton, 1919 [1883]: 212). All that matters is heredity. Put simply, Galton took family, or heredity, to be the foundation of all of the most interesting distinctions between persons.

Figure 1. Anthropometric lab record blank, from Galton (1885: 219).

It is with this in view that he could lament that, 'Owing to absence of data and the want of inquiry of the family antecedents of those who fail and of those who succeed in life, we are much more ignorant than we ought to be of their relative importance' (Galton, 1919 [1883]: 212). That was Galton in his 1883 *Inquiries into Human Faculty and its Development*.

Already in his 1869 *Hereditary Genius* Galton had himself performed a number of such inquiries with respect to the measure of intelligence. There he offered detailed studies of 'a large body of fairly eminent men' in a survey of every English Judge from 1660, every Statesman from the time of George III (about one hundred years later), and each Premier from roughly the same period, as well as a hodgepodge of Commanders, Writers, Painters, Musicians, Scholars (the capitalizations of these statuses is Galton's). These studies were intended to yield a picture of the family background, class status, and racial membership of all of these successful persons. The obvious premise of the project, given that Galton proposes it as a measure of hereditary genius, was that 'high reputation is a pretty accurate test of high ability' (Galton, 1892 [1869]: 2). Galton was not measuring genius itself, but rather took reputation as its proxy. With this in hand, he was able to show that genius (meaning reputation) is inherited along family lines in the form of 'the existence of a law of distribution of ability in families' (Galton, 1892 [1869]: 309).

The year immediately following his expression of lament for a lack of family data, Galton produced a striking instrument designed to fill the lack. A small but deep set of published printed blanks from 1884 remains one of the finest anthropometric paradigms for the practices of informatics that would in later decades fasten persons as subjects of data. Galton's *Record of Family Faculties, Consisting of Tabular Forms and Directions for Entering Data, with an Explanatory Preface* (Galton, 1884a) was published by

Macmillan and Co. of London. The volume was produced in a big but slim format: consisting of nearly 70 sheets measuring 8 inches by 11 inches, with a fly-leaf appended between the title page and page 1 on which is reproduced the advertisement in which Galton offered his £500 in prizes to British subjects who provide him with 'the best Extracts from their own Family Records'.[15] Galton (1884a) opens his 13-page introduction with this promise: 'This book is designed for those who care to forecast the mental and bodily faculties of their children, and to further the science of heredity' (p. 1). Explaining the potential value of such a science, Galton (1884a) asserts that, 'it is possible to foresee much of the latent capacities of a child in mind and body, of the probabilities of his future health and longevity, and of his tendencies to special forms of disease, by a knowledge of his ancestral tendencies' (p. 1). Yet, he observes, 'the advance of the science of heredity is seriously delayed through the want of such data', referring precisely to those data that such of his projects as the book itself and the Anthropometric Laboratory aimed to solicit, normalize, and store (Galton, 1884a: 2).[16]

Galton's book is essentially a data collection machine consisting of some 50 pages of printed blank forms on which a form-filler would write the names of their ancestors (Galton, 1884a: 14), then children (Galton, 1884a: 15), and then describe in detail each of the known family members in terms of vital information (birthdate, birthplace), bodily measures (height, hair color, 'general appearance'), mental measures ('mental powers and energy', eyesight), notes on 'character and temperament', and finally medical information regarding ailments, major illnesses, and causes of death (1884a: 16–60; see Figure 2). It was a book of printed blanks in which one would give an account of something presumed important about oneself. It was a series of pre-defined boxes into which one fit oneself, one's spouse, one's children, and one's parents.

Galton's family record book serves up a metaphor for a new kind of person that most citizens of democratic bureaucracies would become over the next 50 years: informational persons. One vehicle that would deliver them there was an analytical science of the data of heredity – here encapsulated in the formatting technology of query forms. In the final paragraph of his introduction Galton (1884a) writes: 'The scientific importance of each investigation will, however, be soon appreciated by the author of it, for his researches will lay bare many far-reaching biological bonds that tie his family into a connected whole, whose existence was previously little suspected' (p. 13). These bonds in data, he asserts, give rise to 'the conviction that no man stands on an isolated basis, but that he is a prolongation of his ancestry in no metaphorical sense' (Galton, 1884a: 13). One turns the page and the truth of it stares right out. Not in a poetic metaphor, but in a table carefully drawn-up and printed blank for the reader to fill in, one is and becomes an 'Index to Ancestors' (Galton, 1884a: 14).

Information-Technological Structural Racism

At the core of so many of Galton's projects was progressive hereditarianism. Hereditarianism is the idea that talents, traits, and features (such as intelligence) are primarily inherited – nature not nurture. This is an idea that goes against the grain of the democratic egalitarianism that was beginning to find a foothold in Galton's day. Galton recognized the conflict. 'It is in the most unqualified manner that I object to pretensions

Figure 2. Blank for 'Mother's Father' from family development record (Galton, 1884a: 22).

of natural equality,' he once wrote (Galton, 1892 [1869]: 12). Such hereditarianism seems to invite a bleak determinism in which we have little role in to play in the dramas of our own lives – one cannot nurture intelligence in oneself but must simply wait for nature to unfurl. But Galton cultivated a progressive branch of hereditarianism that countered this tragic conclusion with a program of uplift. If we cannot nurture the

individual, perhaps we can nurture nature itself. Thus was conceived eugenics – good breeding for heritable traits.[17] The term offends nearly everyone today. Yet the idea itself under proximate headings continues to compel many, as Rabinow (1996 [1992]) already observed in his discussion of Galton. The recent tumult over the upstart science of behavioral genetics is only the latest vivid example in our contemporary moment.[18]

Yet even after eugenics has been appropriately denounced, there remains an extremely complicated ethics in our inheritance of Galton's other pursuits of inequality in which race breeding was not the explicit and leading theme. These are pursuits which, like so many of our own today, cannot be simply dismissed for bearing the immoralities of attitudinal racism. These other pursuits, both in Galton and in ourselves, we do not disclaim so readily. Indeed we often altogether miss their ethical and political complexity.

Coming to terms with these more complicating configurations is crucial insofar as contemporary data science is today's inheritor of a sizable algorithmic apparatus and formatting technology that Galton in his day helped to perfect. Today's high-performance computing over super-scaled data using machine-learning methods may seem a far cry from Galton's pencil-and-paper statistics and his long-form printed blanks. But both of those Galtonian technologies are conceptual infrastructures upon which contemporary informational storage and processing rely. Without Galton (or someone else having done exactly what he did) there would be no data science today.

By archaeologically excavating such Galtonian layers of our technological present, we can see how, despite the widespread admonition of Galton's racist attitudes among today's technological elites, certain of his information technologies are nevertheless among the social structures scaffolding contemporary racial inequalities just insofar as these inequalities are today partly operationalized through information technology. Considering contemporary data sciences in light of the more complicated history I have been recounting, I take my lead from critical data studies scholars who have articulated the political and ethical dangers of what Ruha Benjamin (2019) calls 'default discrimination' and what Virginia Eubanks (2018) calls 'automated inequality'. If these accounts are right, we who are ensconced in data are under an obligation (by our own anti-racist lights) to actively pursue equality in our design of our data technologies.

Benjamin describes how algorithmic decision-making, for example that facilitated by predictive policing algorithms, is discriminatory not by intention but by design. This is because these systems are programmed in a way that 'builds upon already existing forms of racial domination' (Benjamin, 2019: 81). For example, even those who are most committed to fairness in these technologies often 'still use the crime rate as the default measure of whether an algorithm is predicting fairly, when that very measure is a byproduct of ongoing regimes of selective policing' (Benjamin, 2019: 82). The result can only be the deepening of entrenched disparities, which in the context of the measure of crime in the US are chiefly disparities of unequal treatment by race. Benjamin (2019) further details how even do-good high-tech projects aimed at 'technological benevolence' often serve to reproduce already embedded forms of discrimination despite their intention to counter it (pp. 137–59).

Eubanks focuses her analysis on projects located at the intersection of social science and social welfare. With a particular eye to family services work, Eubanks (2018) details the construction of what she calls 'the digital poorhouse' (p. 12). An exemplar is the

Allegheny Family Screening Tool (AFST) employed by a social services agency in rust-belt Pennsylvania. The directors of the agency, she reports, 'see little downside to data collection because they understand the agency's role as primarily supportive, not punitive' (Eubanks, 2018: 165). Theirs is a project of state-sponsored uplift. Yet similar to predictive policing algorithms, these systems are beset by the logic of self-fulfilling prophecies. In this case, 'a family scored as high risk by the AFST will undergo more scrutiny than other families' (Eubanks, 2018: 169). Higher scrutiny generates more data and more opportunities for data-based alarms. The intention might be support, but the design promises to mete out, and unequally, punishments.

Just as the statistical and technical infrastructure upon which data science relies stretches far back into the past, the automated discriminations that all this computation has produced are by no means a 21st-century invention that can be cheerily chalked up to being a beta-version mistake of nascent science. Our data-driven injustices have a longer and deeper history than we would like to believe. That longer history is of interest, because in it we can witness not only early glimmers of the present, but also the fractures and contests in light of which alternatives were available. Some of these alternatives may be viable again today.

Du Bois's Pursuit

In the earlier half of the 19th century Alexis de Tocqueville (1990 [1835]) famously accused democracy, as represented by America, of a limitless love of egalitarianism: 'for equality their passion is ardent, insatiable, incessant, invincible' (p. 97; second book, chapter I). A generation later, a rough contemporary of Galton's was both one of his age's greatest witnesses to America's unrealized passions for equality and also one if its greatest innovators of methods in the informational analysis of social inequalities. W.E.B. Du Bois is still today widely affirmed as one of our most powerful progenitors of racial equality.[19] Late in life he wrote of his own 'personal life crusade to prove Negro equality' (Du Bois, 1958, cited by Morris, 2020). As part of that crusade, Du Bois was also a pioneering sociologist of the late 19th and early 20th century, whose contributions to both quantitative and qualitative analysis long went unrecognized.[20]

Du Bois's early innovations in social science from the 1890s and 1900s were roughly contemporary with Galton's. Both men preceded the onrushing wave of datafication that would crash over democratic nations in the 1920s and 1930s. Each was equally prescient in sighting what the tides would soon bring. Where Galton's prescience offers us instruction in how technical innovations can be harnessed by dreams of inequality, Du Bois shows how the same can be put to work in pursuit of equality. Yet it is crucial that Du Bois's contribution not be taken as comforting today's data analysts and scientists with the thought that their work could be used for good as much as for ill. Such comfort indulges the false promise that data technologies can be neutral in the actual context of deployment. How could a technology be neutral in a context where its use is already structured by stark inequality? Adding mere neutrality to inequality can only result in inequality reiterated. Du Bois knew all of this quite well, and long before almost anyone else was even asking the question.

What is most compelling in Du Bois's data designs is that he not only pursued equality *with* data but that he also sought equality *within* the data themselves. In other words, Du Bois recognized racializing data as themselves a terrain upon which equality needed pursuing. His data work relied upon, and also amplifies, the idea that unless one explicitly and fervently pursues equality within the very parameters of people's datafication, then inequality is almost surely bound to follow. This strategy of *pursuing equality within data* needs to be distinguished from the much more familiar work of *using data as an instrument for equality*. For example, consider how datasets are frequently used to prove inequality through statistical analyses of discriminatory disparate impact (in employment, education, housing and more), but are far less commonly interrogated as potential sites of inequality at which the data structures themselves (even patently neutral ones) serve to reproduce extant inequalities, exclusions, and divisions.

The central idea I offer here of Du Bois as pursuing equality within data builds on Autumn Womack's (2022) recent redescription of Du Bois's work as 'an attempt to make data move' (p. 49). Making data move means 'to realign the presumed equation between black life and data while also insisting on a data regime that might allow us to perceive and receive social life' (Womack, 2022: 49). Womack's emphasis on moving data resonates with Saidiya Hartman's (2019) claim that Du Bois's 'figures, charts, and graphs aspired to be a moving picture of black life', and a 'story in motion' narrating 'a changing and variable entity' (p. 110). Womack's and Hartman's conceptions capture something crucial in Du Bois, which I describe as an effort in formatting, and reformatting, racial data. The stakes of racial data formats are clarified in Womack's (2022) refusal of 'an easy ideological distinction between documenting and experimenting', affirming instead that, 'in the face of black life, documenting was always already an experimental process, one that had the potential to reorder and disorder the easy bundling of data and black life' (p. 8). Where data formats are calcified such that they are no longer available as ongoing objects for inquiry, there do we lose the possibility for 'undisciplining data' in the context of experiments in documentation (Womack, 2022: 9). And with that we lose the very possibility of equality in data.

Consider some of the most forceful exemplifications of equality within data in Du Bois's early publications. I recounted above Galton's employments of his anthropometric forms at the 1884 International Health Exhibition in London. Just 16 years later, Du Bois was across the channel in Paris at the 1900 Exposition Universelle, where he curated the American Negro Exhibit in the Pavilion of Social Economy. Du Bois's presentation has recently been collected and reprinted in full color for the first time by Whitney Battle-Baptiste and Britt Rusert in their *W.E.B. Du Bois's Data Portraits* (Du Bois, 2018 [1900]).

What is most remarkable about the data portraits is their unabashed expression of positive facts of African-American progress and equality. In the words of Du Bois's collaborator on the exposition, Thomas Calloway (1901), the team's work was focused on 'compiling data and collecting material for an exhibit of the progress of the American negroes in education and industry' (p. 463). In Du Bois's (1900) own words, the exhibit offered 'a series of striking models of the progress of the colored people, beginning with the homeless freedman and ending with the modern brick schoolhouse and its teachers' (p. 576). By featuring African-American 'development' and 'progress', Du Bois and his

collaborators skillfully countered the discourse of the hereditarians, many of whom would have been present in Paris with their own exhibits (though Galton himself seems not to have been).[21]

Consider Plate 47, titled 'Illiteracy of the American Negroes compared with that of other nations' (see Figure 3). This is a classic bar chart with ten measures protruding from left to right, each bar labeled (in French). No exact percentages are given, and so the chart serves a purely comparative purpose. At the top of the chart are bars for 'Roumanie', 'Servie', and 'Russie', each indicating roughly the same level of illiteracy. Just below them, showing significantly less illiteracy, is a bar labeled 'Negroes, U.S.A.' followed by 'Hongrie' with only slightly less illiteracy. The last five bars are all Central and Western European nations with comparatively lower illiteracy, though that in Italy is not much lower than in Hungary, and Sweden (at the bottom of the chart) is shown to have remarkably higher literacy than even France (second from last).

In describing the exhibition the following year, Du Bois (1900) called attention to this particular chart as a model for the series (p. 577). The editors of the reprint note that Plate 47 serves to unambiguously 'correct misconceptions about the education of black Americans' (Battle-Baptiste and Rusert in Du Bois, 2018 [1900], caption to Plate 47). For it visually demonstrates that inequalities in education owe more to socio-historical factors than biological-racial causes. It cannot be overlooked that the argument made by this illustration relied upon the innovative work in data production through which Du Bois and his team made data move in ways that pre-existing datasets could not allow.

Du Bois's data portraits in Paris are not the only instance of his egalitarian science of data. Another instance, which I do not have the space to detail here, is an early criminological survey conducted in Georgia (see Du Bois, 1904). The study presents survey data by race of respondents' perceptions of the criminal justice system. The analysis, striking for its time, calls forth the possibility of alternative measures of justice based not so much on proportions of arrest per demographic segment but proportions of perception of fairness per demographic segment.

Also striking is his 1899 study *The Philadephia Negro*, produced on the back of a gargantuan task of compiling a huge volume of data on an urban black population, the first study of its kind on any demographic in any city. This book offers insight into Du Bois's data collection methods, including those that may have been employed for the Paris exhibition the next year. Of particular interest are its appendix reproductions of the questionnaires or 'schedules' used to generate study data (Du Bois, 2007 [1899]: 276–86; see Figure 4). Du Bois's schedules are functionally quite like Galton's anthropometric records discussed above – their very construction is designed to elicit data inputs. But where Galton's forms belie formats in search of familial inequalities, Du Bois's formats are a formidable alternative in their function of tuning inquiry to the many modes of development possible within a politically neglected population.

One example of this that would resonate in Du Bois's presentations in Paris the next year are his studies of the growth of literacy rates among black Philadelphians (Du Bois, 2007 [1899]: 276, questions 9 and 10; see also discussion based on these data at Du Bois, 2007 [1899]: 64). He included in one footnote a table with a bar graph showing that the literacy rate of a sample black population in Philadelphia was above that in five European nations, and only slightly below that in Germany (Du Bois, 2007 [1899]: 68n8).
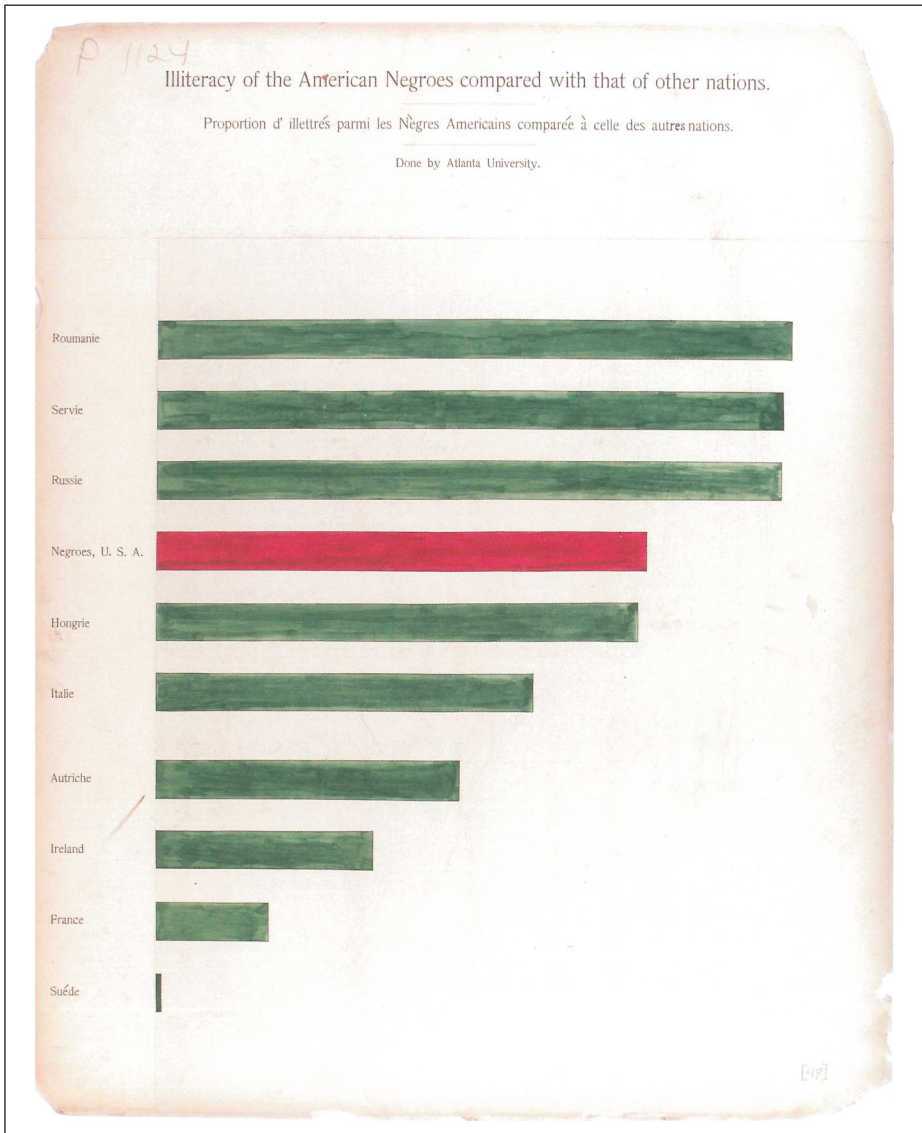
Figure 3. Item 47 from Du Bois's statistical charts exhibited at the 1900 Paris Exhibition, sourced from Library of Congress (LOT 11931, no. 47 (M) [P&P]); see also reprint in Du Bois (2018 [1900]), edited by Battle-Baptiste and Rusert.

In contrasting Du Bois's schedules to Galton's forms, what stands out is the former's commitment to collecting those data that would reveal not just the influence of an individual's heredity, but also the influences of their social environment. In this, Du Bois's data methods are designed to make space for evidence of those equalities among persons that Galton simply assumed away.

Figure 4. Selection from Schedule 2 from Appendix A to original edition of Du Bois's (1899) *The Philadelphia Negro* (p. 404).

From these and other of Du Bois's pursuits of equality follows a crucial imperative: *where we grasp and fasten persons through data, even when it is expressly for the sake of ameliorating social conditions, our very design of data must be actively and fervently trained on equality*, for otherwise it is just too hard to not reproduce inequality. This charge is crucial insofar as data-driven social projects are particularly susceptible to techno-structural inegalitarianism, a tendency that Du Bois spent a career witnessing, and which contemporary scholars like Benjamin and Eubanks have observed again. It is as if inequality is the default condition for data-driven social science, at least where the object of social inquiry is a society deeply riven by inequality.

Toward Equality in Data

There is equality and then there is equality. My argument is not just that data can and should be used to pursue equality, for this is an argument that is already widely familiar and hardly contested. Rather, my point is that those who do anything with data (including pursuing equality) need to be fervently attentive to ways in which inequalities may be

designed into their data. Today's critical data studies scholars have shown how the use of data in pursuit of equality can go awfully awry. Du Bois helps us further understand that the pursuit of equality *within* data is a condition of the pursuit of equality *with* data.

What makes Du Bois such a compelling model for our contemporary moment is his adeptness at making data move in ways that pre-existing datasets given to his milieu would not permit. Du Bois built data inside of spaces and problems where it had not yet traveled, or at least had yet to travel well. Data analysts today are in an odd position of having at their disposal both high-performance processing apparatus and vast libraries of pre-existing data sources. Yet these sources are almost always designed for purposes other than those that animate their recycling. Expensive algorithms thus get run on what is all too often off-the-shelf trash. When this happens, any ostensible commitment to equality easily goes wanting. Du Bois's model shows how a commitment to equality needs to run throughout our information technology, all the way down to data design and formatting.

What, then, is the pursuit of equality in data? It involves resolute attentiveness to data formats, including, for instance, the relevant fields and permissible variables internal to any datafication. It involves unflagging focus on the dangers of innocent-seeming proxy fields for politically charged social categories. It involves explicit interrogation into whether the measuring instruments employed to make data are themselves reproductive of social conditions they might be charged to ameliorate, as exampled by racial bias in intelligence testing instruments, to take another case in which Du Bois anticipated later critical scholarship (Du Bois, 1987 [1920]).

Without pursuing egalitarianism within data as a counter to data's tendencies toward inequality, we leave those whose lives are disclosed by data too much exposed to the haunting hierarchies of manifold legacies of inequality. Yet there truly is nothing in the very idea of data that fosters inequality rather than equality. It is our choice whether we design and deploy databases, information architectures, and downstream algorithmic apparatus that generate or mitigate inequality. And yet such choices are deeply burdened by the histories in whose futures we remain buried, and therefore also by present social and technological contexts within which these choices become operative. While it may seem to some that it is easy to make choices for equality, the history of our present teaches us how hard that choice has been to maintain. The data that are given to us have so infrequently been moved to beat against the currents of our history.

Acknowledgements

ORCID iD

Colin Koopman [iD] https://orcid.org/0000-0003-0354-7522

Notes

1. My formulation here is expressly meant to imply that genetics and genomics have always been data-centric sciences, as I argue in Koopman (2019b).
2. See Carmichael and Hamilton (1967: 4) and Bonilla-Silva (2014: 2).
3. See Winner (1980) for one precedent of how to pose this question.
4. See also Chun (2021: 20) on 'whiteness as default'.
5. On the concept of 'formats' see my discussion in Koopman (2019a and 2021).
6. Among the media-archaeological and media-genealogical methodological precedents informing my approach here are studies by Vismann (2008 [2000]), Gitelman (2014), and Packer et al. (2023). For another recent media-archaeological approach to Du Bois see Bering-Porter (2022) – whereas I agree with Bering-Porter on the need for an interrogation of the mediatic conditions of Du Bois's data portraits, my approach here also pairs such an inquiry with a consideration of the normative values informing Du Bois's work, a consideration largely left to the side by Bering-Porter.
7. See Galton (1908: 268ff.) for his own account.
8. See Hacking (1990: 180ff.) for one example
9. See Sheehey (2019) for a related argument.
10. On formats see note 5 above.
11. I describe the contrast between algorithm and format in greater detail in Koopman (2021).
12. See on algorithms the summary account by Gillespie (2016); the full literature on this technical concept is truly enormous.
13. See on measure Beer (2016) and Brighenti (2018), both of which offer genealogies of measure that resonate with my study here.
14. See on the importance of blank forms Gitelman (2014: ch. 1).
15. This text is from the fly-leaf to Galton (1884a).
16. Galton (1884a) also here (p. 9) refers to the need for anthropometric laboratories (such as his own).
17. In Galton's (1884a) family records project, this logical next step for hereditarian inegalitarianism is already made plain in his treatment of certain data fields (namely 'birthplace' and 'residence') on his forms as proxies or 'partial indications' of race (p. 5).
18. See Harden (2021) and the profile by Lewis-Kraus (2021); for a different kind of example, see discussion by Russell (2018) of the continuing presence of race planning in the familial domain of assisted reproduction.
19. Given the widely divergent interpretations of Du Bois's work in the literature, I note that my presentation of the early Du Bois is one I take to be generally in line with Dawson's (2001: 15) characterization of him as a liberal radical egalitarian and Mills's (2018: 19–56) presentation of him as a black radical liberal.
20. On Du Bois's early work in what would later be understood as quantitative sociology, see Morris (2020: 20–21); on how the quantitative dimensions of Du Bois's research program are still too often overlooked today, see Morris (2015).
21. For his criticisms of Galton, much later in life, see Du Bois (c. 1955).

References

Battle-Baptiste, Whitney and Rusert, Britt (2018) Introduction. In: *W.E.B. Du Bois's Data Portraits: Visualizing Black America*. New York, NY: Princeton Architectural Press.

Beer, David (2016) *Metric Power*. New York, NY: Palgrave Macmillan.

Benjamin, Ruha (2019) *Race after Technology*. Medford, OR: Polity.

Bering-Porter, David (2022) Data as symbolic form: Datafication and the imaginary media of W.E.B. Du Bois. *Critical Inquiry* 48(2): 262–285.

Bonilla-Silva, Eduardo (2014) *Racism without Racists*. Lanham, MD: Rowman and Littlefield.

Brighenti, Andrea Mubi (2018) The social life of measures: Conceptualizing measure–value environments. *Theory, Culture & Society* 35(1): 23–44.

Brock, André, Jr., (2020) *Distributed Blackness: African American Cybercultures*. New York, NY: New York University Press.

Calloway, Thomas (1901) The Negro exhibit. In: *Report of the Commissioner-General for the United States to the International Universal Exposition, Paris, 1900*, vol. II. Washington, DC: Government Printing Office, pp. 463–467.

Carmichael, Stokely and Hamilton, Charles V (1967) *Black Power: The Politics of Liberation in America*. New York, NY: Vintage.

Chun, Wendy Hui Kyong (2021) *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press.

Dawson, Michael (2001) *Black Visions: The Roots of Contemporary African-American Political Ideologies*. Chicago, IL: University of Chicago Press.

Du Bois, William Edward Burghardt (1899) *The Philadelphia Negro: A Social Study*. Philadelphia, PA: University of Pennsylvania.

Du Bois, William Edward Burghardt (1900) The American Negro at Paris. *American Review of Reviews* 22(5): 575–577.

Du Bois, William Edward Burghardt (ed.) (1904) *Some Notes on Negro Crime, Particularly in Georgia*. Atlanta, GA: Atlanta University Press.

Du Bois, William Edward Burghardt (c. 1955). Equality of the races. W.E.B. Du Bois papers (MS 312), Special Collections and University Archives, University of Massachusetts Amherst Libraries, Available at: https://credo.library.umass.edu/view/full/mums312-b207-i018, p. 3 (labeled '14' in top-left corner), accessed 23 March 2023.

Du Bois, William Edward Burghardt (1958) The early beginnings of the Pan-African movement, June 20, 1958. W.E.B. Du Bois papers (MS 312), University of Massachusetts Amherst Libraries.

Du Bois, William Edward Burghardt (1987 [1920]) Race intelligence (from *The Crisis*, July 1920). In: *Du Bois, Writings*, ed. Nathan Huggins. New York, NY: Library of America, pp. 1181–1183.

Du Bois, William Edward Burghardt (2007 [1899]). *The Philadelphia Negro: A Social Study*. Oxford: Oxford University Press.

Du Bois, William Edward Burghardt (2018 [1900]) *W.E.B. Du Bois's Data Portraits: Visualizing Black America*, ed. and with an Introduction by Whitney Battle-Baptiste and Britt Rusert. New York, NY: Princeton Architectural Press.

Eubanks, Virginia (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: Macmillan.

Foucault, Michel (1995 [1975]) *Discipline and Punish: The Birth of the Prison*, trans. A Sheridan. New York, NY: Vintage Books.

Galton, Francis (1879) Psychometric experiments. *Brain* 2: 149–162.

Galton, Francis (1884a) *Record of Family Faculties, Consisting of Tabular Forms and Directions for Entering Data, With an Explanatory Preface*. London: Macmillan and Co.

Galton, Francis (1884b) Measurement of character. *Fortnightly Review* 36: 179–185.

Galton, Francis (1885) On the anthropometric laboratory at the late international health exhibition. *Journal of the Anthropological Institute* 14: 205–218.

Galton, Francis (1892) Retrospect of work done at my anthropometric laboratory at South Kensington. *Journal of the Anthropological Institute* 21: 32–35.

Galton, Francis (1892 [1869]) *Hereditary Genius: An Inquiry into its Laws and Consequences*, 2nd edn. London: Macmillan and Co.

Galton, Francis (1908) *Memories of My Life*, 3rd edn. London: Methuen and Co.

Galton, Francis (1919 [1883]) *Inquiries into Human Faculty and its Development*, 2nd edn. New York, NY: E.P. Dutton and Co.

Gillespie, Tarleton (2016) Algorithm. In: Peters, Benjamin (ed.) *Digital Keywords*. Princeton, NJ: Princeton University Press, pp. 18–30.

Gitelman, Lisa (2014) *Paper Knowledge: Toward a Media History of Documents*. Durham, NC: Duke University Press.

Gould, Stephen Jay (1996 [1981]) *The Mismeasure of Man*, rev. edn. New York, NY: W.W. Norton.

Hacking, Ian (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Harden, Kathryn Paige (2021) *The Genetic Lottery: Why DNA Matters for Social Equality*. Princeton, NJ: Princeton University Press.

Hartman, Saidiya (2019) *Wayward Lives, Beautiful Experiments: Intimate Histories of Riotous Black Girls, Troublesome Women, and Queer Radicals*. New York, NY: W.W. Norton.

Koopman, Colin (2019a) *How We Became Our Data: A Genealogy of the Informational Person*. Chicago, IL: University of Chicago Press.

Koopman, Colin (2019b) Coding the self: The infopolitics and biopolitics of genetic sciences. *Hastings Center Report* 50(5): S6–S14.

Koopman, Colin (2021) The political theory of data: Institutions, algorithms, and formats in racial redlining. *Political Theory* 50(2): 337–361.

Lewis-Kraus, Gideon (2021) Can progressives be convinced that genetics matters? *The New Yorker*, 13 September, pp. 48–59.

Mattern, Shannon (2020) The spectacle of data: A century of fairs, *fiches*, and fantasies. *Theory, Culture, & Society* 37(7–8): 133–155.

Mills, Charles (2018) W.E.B. Du Bois: Black radical liberal. In: Brommell, Nick (ed.) *A Political Companion to W. E. B. Du Bois*. Lexington, KY: University Press of Kentucky, pp. 19–56.

Morris, Aldon (2015) *The Scholar Denied: W.E.B. Du Bois and the Birth of Modern Sociology*. Berkeley, CA: University of California Press.

Morris, Aldon (2020) The sociology of W.E.B. Du Bois as a weapon of racial equality. *Quaderni di Sociologia* 83(64): 11–24.

Noble, Safiya (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press.

Packer, Jeremy, de Villavicencio, Paula Nuñez, Monea, Alexander, Oswald, Kathleen, Maddalena, Kate, and Reeves, Joshua. (2023) *The Prison House of the Circuit: Politics of Control from Analog to Digital*. Minneapolis, MN: University of Minnesota Press.

Rabinow, Paul (1996 [1992]) Galton's regret. In: Rabinow, Paul (ed.) *Essays on the Anthropology of Reason* (reprint). Princeton, NJ: Princeton University Press, pp. 112–128.

Russell, Camisha (2018) *The Assisted Reproduction of Race*. Bloomington, IN: Indiana University Press.

Sheehey, Bonnie (2019) Algorithmic paranoia: The temporal governmentality of predictive polic-
    ing. *Ethics and Information Technology* 21: 49–58.
Tocqueville, Alexis de (1990 [1835]) *Democracy in America*, trans. Henry Reeve. New York, NY:
    Vintage Books.
Vismann, Cornelia (2008 [2000]) *Files: Law and Media Technology*, trans. G. Winthrop-Young.
    Stanford, CA: Stanford University Press.
Winner, Langdon (1980) Do artifacts have politics? *Daedalus* 109(1): 121–136.
Womack, Autumn (2022) *The Matter of Black Living: The Aesthetic Experiment of Racial Data,
    1880–1930*. Chicago, IL: University of Chicago Press.

Colin Koopman is Professor and Head of Philosophy at the University of Oregon, where he also
serves as Director of the interdisciplinary New Media & Culture program. He is currently working
on a book on the politics of data egalitarianism, tentatively titled *Data Equals*, recently profiled in
The New York Times Magazine (Mar. 26, 2023). His previous book is *How We Became Our Data:
A Genealogy of the Informational Person*. His writings on data politics have appeared in *The New
York Times*, *Political Theory*, *BioSocieties*, and elsewhere. Other writings on figures in the history
of philosophy such as Michel Foucault and William James have appeared in venues including
Critical Inquiry, *diacritics*, and *Journal of the History of Philosophy*.