

DISCRIMINATING DATA

CORRELATION, NEIGHBORHOODS, AND THE NEW
POLITICS OF RECOGNITION

WENDY HUI KYONG CHUN

MATHEMATICAL ILLUSTRATIONS BY ALEX BARNETT

THE MIT PRESS
CAMBRIDGE, MASSACHUSETTS
LONDON, ENGLAND

© 2021 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in ITC Stone and Avenir by New Best-set Typesetters Ltd.

Library of Congress Cataloging-in-Publication Data

Names: Chun, Wendy Hui Kyong, 1969- author. | Barnett, Alex, 1972- illustrator.

Title: Discriminating data : correlation, neighborhoods, and the new politics of recognition / Wendy Hui Kyong Chun ; mathematical illustrations by Alex Barnett.

Description: Cambridge, Massachusetts : The MIT Press, [2021] | Includes bibliographical references and index.

Identifiers: LCCN 2021000481 | ISBN 9780262046220 (hardcover)

Subjects: LCSH: Big data—Social aspects. | Artificial intelligence—Social aspects. | Privacy, Right of.

Classification: LCC QA76.9.B45 C57 2021 | DDC 005.7—dc23

LC record available at <https://lccn.loc.gov/2021000481>

10 9 8 7 6 5 4 3 2 1

To all the students I taught at Brown, Penn, Chicago, and Simon Fraser, in particular the first four at the Digital Democracies Institute: Amy, Carina, Hannah, and Julia.

And to my first teachers: Jeannie, Maria, Ernie, and Bert.

1

CORRELATING EUGENICS

The Cambridge Analytica scandal exemplified social media's perceived threat to democratic institutions and processes. Cambridge Analytica—a data firm hired by the 2016 Donald Trump and Ted Cruz presidential campaigns and funded by Republican hedge fund and machine learning pioneer Robert Mercer—allegedly altered the results of that year's U.S. election and UK Brexit referendum. The immodest statements made by Cambridge Analytica CEO Alexander Nix partly fueled these allegations: during a 2016 speech at the Concordia Summit, Nix claimed responsibility for Cruz's success in that year's primaries. As Nix explained, Cruz was both generally unliked and unrecognized at the beginning of his primary campaign, but Cruz's embrace—through Cambridge Analytica—of behavioral science (psychographics), addressable ad technology, and big data powered his steady rise; in the end, he came in second only to Trump. Specifically, Cambridge Analytica, which had created “a profile of every adult in the United States of America,” targeted and swung “persuadable” voters for Cruz.¹

Cambridge Analytica's celebration of big data and the data firm's exaggerated claims were the norm during the first decades of the twenty-first century, the “century of big data.” *The Economist* proclaimed data “the oil of the digital era”—“the world's most valuable resource”; IBM promised that big data analytics would offer “insights without limits.”² Fox News

declared: “‘Big data’ will blow your mind and change the 21st century.”³ Bloomberg, Oracle, and numerous other organizations proclaimed that big data would “disrupt” everything.⁴ The 2020 documentary *The Social Dilemma* claimed that, through big data, social media platforms dominated users and turned them into marionettes.⁵

Big data’s power was said to be based on correlation, but this was not correlation’s first rodeo. Along with linear regression and other foundational statistical methods, correlation was developed by early twentieth-century biometric eugenicists, who were eager to breed a better “human crop.” By investigating the historical ties between big data and eugenics, we will see that the two are linked together by a fundamentally undisruptive view of the future. But, as we will also see later in the chapter, even though both have sought to make the future repeat a highly selective and discriminatory past through correlation (so that *ground truth* = *deep fake*), they differ in several important respects. In the transition from eugenics to data analytics, the focus group moved the nation to the neighborhood/tribe; the goal shifted from uplift to escape; and homophily (the notion that similarity breeds connection) went from aspiration to axiom.

CULTIVATING HUMANS

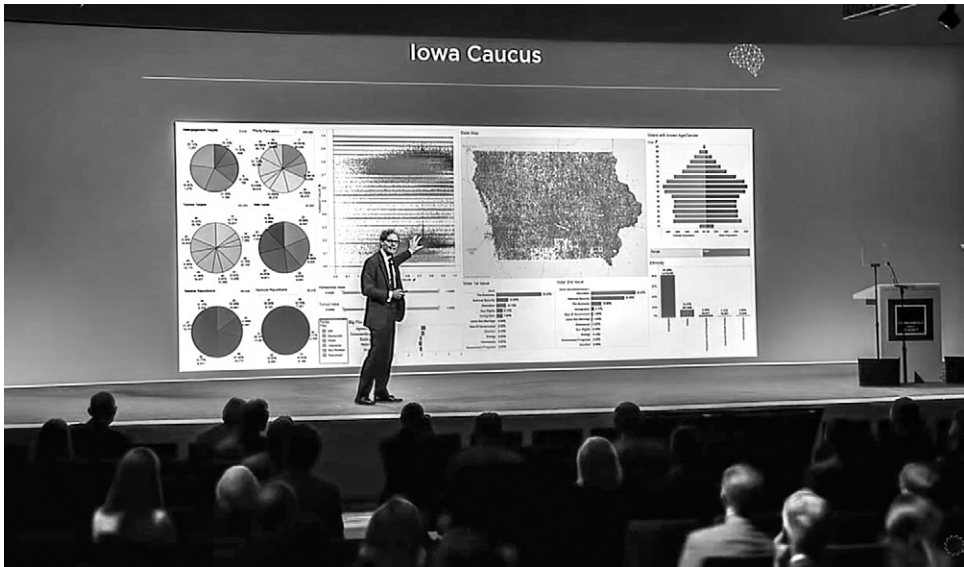
Key to Cambridge Analytica’s success was psychographics because, Nix contended during his Concordia presentation, “it’s personality that drives behavior and behavior that obviously influences how you vote.”⁶ Psychographics superseded demographics, geographics, and economics, with their crude assumptions “that all women should receive the same message because of their gender, or all African Americans because of their race, or all old people, or rich people or young people to get the same message because of their demographics.” White men, the group Nix actually targeted for Cruz, was tellingly missing from this list. As will be discussed further in this chapter, Nix’s “all X” formulation inadvertently revealed that his “solution” to identity-based politics and advertising was not to dissolve these demographic categories, but rather to further segment them, based on “personality.” To determine a person’s personality, Cambridge Analytica deployed a “long form quantitative instrument to probe the underlying traits that inform personality.” The data firm scored

a person's personality using the five-factor OCEAN model, where O = openness ("how open you are to new experiences"); C = conscientiousness ("whether you prefer order and habits and planning in your life"); E = extroversion ("how social you are"); A = agreeableness ("whether you put other peoples' needs and society and community ahead of yourself"); and N = neuroticism ("a measurement of how much you tend to worry").

Nix offered an extended example of likely Iowa caucus participants and the U.S. Second Amendment as proof. "For a highly neurotic and conscientious audience," he asserted, as he showed an image of a white woman with "professional hair," you needed a rational yet fear-based message: "The threat of a burglary and the insurance policy of a gun is very persuasive." In contrast, "For a closed and agreeable audience, these are people who care about tradition and habits and family and community," he explained as he displayed an image of a smiling middle-aged white male: "This could be the grandfather who taught his son to shoot and the father who will in turn teach his son. . . . Talking about these values is going to be much more effective in communicating your message" (figure 10). To decide whom to target and which ads to produce,



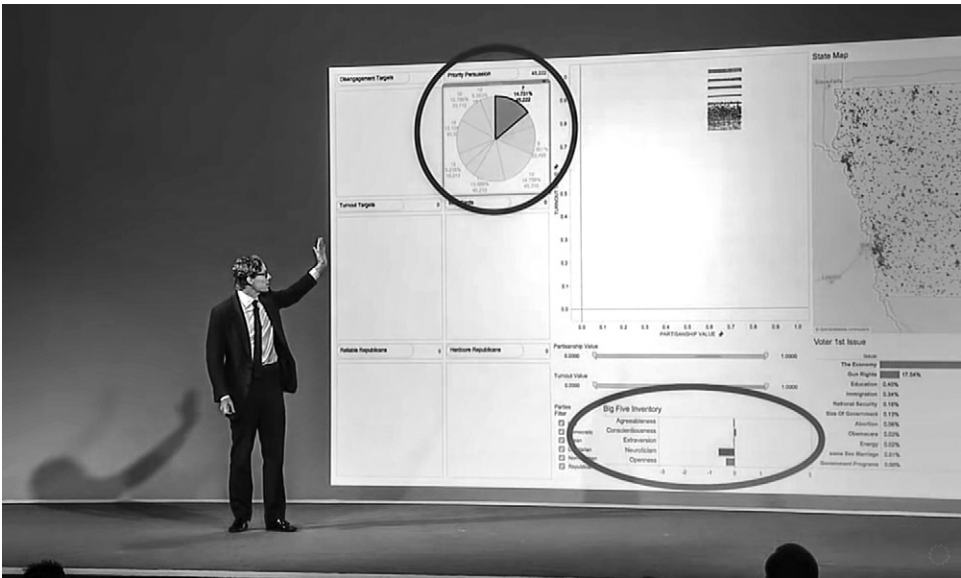
10 Still frame of Cambridge Analytica's psychographic messaging, from Nix's 2016 presentation at the Concordia Summit, <https://youtu.be/n8Dd5aVXLCc>.



11 Still frame of Cambridge Analytica's Iowa data dashboard, from Nix's 2016 presentation at the Concordia Summit, <https://youtu.be/n8Dd5aVXLCc>.

Cambridge Analytica created a “data dashboard” (figures 11 and 12) for each state, which sorted people by party and likelihood to vote.

From these data dashboards, Cambridge Analytica culled a “persuasion” group of about 45,000 Iowans who mattered to Cruz because they were definitely going to caucus, but they needed to be moved “a little more toward the right” if they were going to support him. After determining the mean personality type of the group to be “very low in neuroticism, quite low in openness and slightly conscientious,” the firm identified a subset who cared about gun rights versus gun control—a highly divisive, emotionally charged, and well-funded issue in the United States. In his presentation, Nix revealed that the Cruz campaign used their insights to drive not only their ads but also their field operations. These ads could target actual individuals since, Nix bragged, Cambridge Analytica had “somewhere close to four to five thousand data points on every adult in the U.S.” What Nix did not state, and what Carole Cadwalladr and Emma Graham-Harrison of the *Guardian* later revealed, was that his firm had harvested 50 million Facebook profiles unbeknownst to their “owners,”



12 Still frame of Cambridge Analytica's microtargeting, from Nix's 2016 presentation at the Concordia Summit, <https://youtu.be/n8Dd5aVXLCc>.

through a quasi-legal deal with then Cambridge University researcher Aleksandr Kogan, to produce their data dashboards.⁷

As many researchers have emphasized, the claims made by Cambridge Analytica need to be taken with four to five thousand grains of salt—there is much that cannot be known about the actual impact of Cambridge Analytica on the 2016 U.S. presidential elections. Indeed, following the elections, Cambridge Analytica itself told reporters that it was impossible to verify its claims, and the firm could not offer a single case as proof.⁸ Even if it could—and we accepted the 2016 presidential election as evidence—that single case would represent a sample size of one. Further, the efficacy of targeted political ads using the OCEAN model is still in question; Facebook and other social media effectively use metrics other than OCEAN to “prime” users.⁹

The rush to attribute the generally unforeseen victory of Donald Trump in 2016 to Cambridge Analytica also rewrites history. As late as November 4, 2016—the day before the election—Hillary Clinton, not Donald Trump, was heralded as the big data candidate. Countless articles

documented Trump's disregard for data, and his preference to "go with his gut." *Politico* and many other news outlets praised Clinton's data guru, Elan Kriegel, as "precise and efficient, meticulous and effective."¹⁰ Kriegel, who had formerly worked in the Obama "cave," had created a tool that could calculate the "cost per flippable delegate." According to Jeremy Bird, a consultant who worked with both Obama and Clinton, Obama consulted Kriegel before every decision about strategy in battleground states, and Elan "was never wrong." Kriegel's work, it was claimed, was even more precise for the Clinton campaign—she was thus sure to win. As *Politico* surmised: "Now, with Donald Trump investing virtually nothing in data analytics during the primary and little since, Kriegel's work isn't just powering Clinton's campaign, it is providing her a crucial tactical advantage in the campaign's final stretch. . . . As millions of phone calls are made, doors knocked and ads aired in the next nine weeks, it is far likelier the Democratic voter contacts will reach the best and most receptive audiences than the Republican ones [will]."¹¹

Right.

Clearly, Clinton lost and Trump won. That we know. We also know, in retrospect, that Clinton's models were overfitted to the previous presidential campaign: they had, for instance, presumed a significant African American voter turnout, even after the controversy over Clinton's 1996 "superpredators" comment. In other words, they had ignored experience and specific events in their formal conceptualization of voters.

Rather than dismissing Cambridge Analytica's claims as snake oil or sorting through the election voter data to assess the actual impact of Cambridge Analytica's "special sauce," however, the more pressing task is to answer these two questions:

To what extent did Cambridge Analytica get some things right not simply because it "discerned" what was out there, but because the data firm sought to create it as well?

And what world are we living in that Cambridge Analytica's claims seem plausible?

The goal of the firm's advertisements was to create transformational, "red pill" experiences: to have users go "down the rabbit hole" by following ads, carefully "breadcrumbs" across different sites and spread by their friends and others "like them."¹² Identification—or targeting—was

the first function within the program: the others were recognition (mutual identification) and conversion.

As Cambridge Analytica whistleblower Christopher Wylie explained to Carole Cadwalladr in 2018, the data firm's "information operations," which were inspired by the U.S. military's doctrine of "five-dimensional battle space," correlated culture with politics. Wylie (the self-described "gay Canadian vegan who somehow ended up creating 'Steve Bannon's psychological warfare mindfuck tool'") formerly researched fashion trends. He told Bannon, then chief executive officer of Trump's 2016 presidential campaign, that, based on his prior research, "politics was like fashion." Trump was like a pair of Ugg boots, and the goal was to find the inflection point that moved people from thinking these were "'Ugh. Totally ugly' to the moment when everyone is wearing them." And Bannon believed this message, Wylie explained, because he adhered to the Breitbart doctrine "that politics is downstream from culture, so to change politics you need to change culture."¹³ Cambridge Analytica took this doctrine one step further by arguing that, to change culture, you had to change individuals. Personalities were key to changing culture.

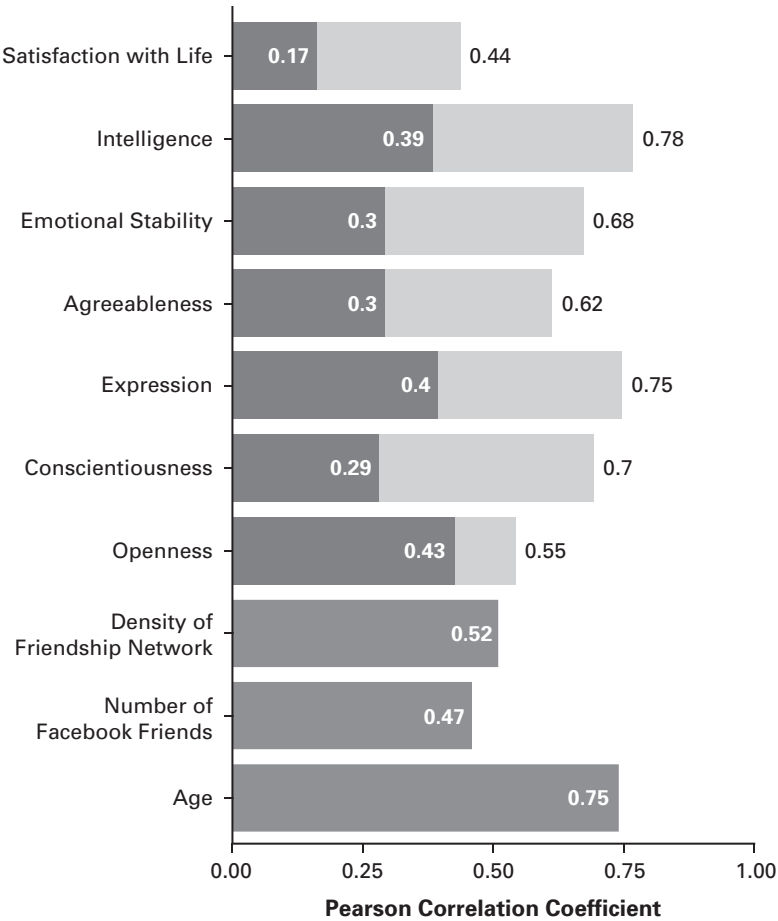
Maybe.

If we have learned anything from Cambridge Analytica, however, it is that finding and exploiting clusters is key to cultivating individual behavior—to fashioning change. "Personalization" works at the levels of individual actions, "latent" factors, and "bespoke" network neighborhoods—all at the same time. As chapter 3 elaborates, recommendations and social media "feeds" for a particular individual do not depend solely on that individual's history. If they did, they would be very limited in scope. Predictive data analytics for Internet users work—if and when they do—not by treating every Internet user like a unique snowflake, but rather by segregating users into "neighborhoods" or petri dishes based on their slightly odd or deviant—that is, "authentic"—likes and dislikes. Individuals are formed and identified by their so-called neighbors.¹⁴ Cambridge Analytica claimed to have discovered proxies that revealed a person's race, sexual orientation, political leanings, and so on: preferring an American car, for example, strongly indicated a possible Trump voter.¹⁵ Again, these proxies were sought in relation to inflection points—points at which curves would bend in new directions. "Culture" is not simply a noun, but also a

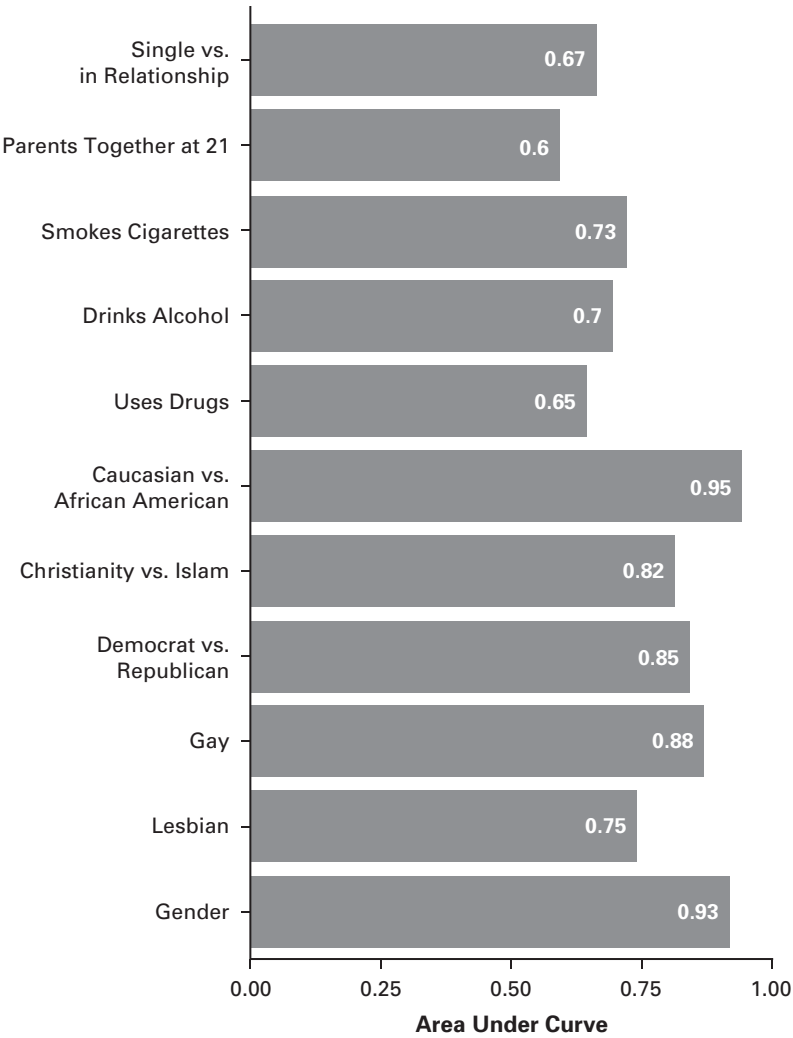
verb. To culture is to cultivate: “to propagate, grow or develop . . . under artificial conditions or in a nutrient medium.”¹⁶ “Culture,” “colony,” and “colonization” are all derived from the Latin *colere*, “to cultivate or worship.” A *colonus* was a settler: a Roman soldier-farmer, who was posted in foreign or hostile territory and who seized land by enclosing or settling on it. Cultures are and depend on invasive separations.

Although Cambridge Analytica did not reveal how it used Facebook likes to determine personality, computational social scientists Michal Kosinski and David Stillwell, and computer scientist Thore Graepel, whose work partly inspired Cambridge Analytica, explained how this microtargeting works in their influential 2013 study “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.”¹⁷ They revealed how easy it was to predict latent user attributes (identity categories) such as “sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender” based on then publicly available Facebook likes. As the listed attributes makes clear, Kosinski, Stillwell, and Graepel sought to create a model that would estimate a range of characteristics. The researchers could do so because more than 58,000 Facebook users had completed the myPersonality Facebook questionnaire the researchers had circulated—and by doing so had given the researchers access to information in the their Facebook profiles.

To produce their estimates, the three researchers first posited boundary-making traits such as “political views,” “parents stayed together until the individual was 21,” “ethnic background,” and “intelligence,” which they “measured” using various methods, including the five-factor OCEAN model, intelligence tests, and visual examination of user profiles and online survey answers. They then created a vast but sparse user-like matrix comprising all likes associated with each user. Next, they decomposed this matrix using singular value decomposition (SVD), which reduces a matrix of data points into a series of vectors, ranked by how much they explain the original data set (described in greater detail in “Proxies” after chapter 2) to determine the hundred most significant components. Then, using these most significant components, they created linear regression models to predict numeric attributes, such as personality and age, and logistic regression models to predict dichotomous values, such as male versus female or Christian versus Muslim (figures 13 and 14).



13 Prediction accuracy for linear regression models. Redrawn from Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences* 110, no. 15 (2013): 5804.



14 Prediction accuracy for dichotomous logistic regression models, redrawn from Kosinski, Stillwell, and Graepel, "Private Traits and Attributes Are Predictable," 5803.

The logistic regression models (designed to predict dichotomous values) presumed the existence of separate and opposed binary categories (male versus female; Christian versus Muslim); these models were given two samples from each category and then assessed on their ability to predict who belonged in each (see figure 14). The goal was to put every sample into its right category by adhering to a strict either-or logic.

The accuracy of these models varied greatly, with the most accurate being for Caucasian versus African American and male versus female (figure 14). Not surprisingly, the highly structured dichotomous value models gave the best results. The least accurate predictions were those linked to numeric attributes (figure 13): satisfaction with life, conscientiousness, emotional stability, and agreeableness (these numeric attributes were assessed using the Pearson correlation coefficient, explained below in “Correlation” by Alex Barnett; figure 17). Finally, they produced tables of the most predictive likes—that is, those with the highest weighted average or the most extreme frequencies of classes—for certain traits (figure 15).

Based on this, Kosinski, Stillwell, and Graepel claimed that, by knowing as few as one like, they could determine a user’s related “intimate” trait. For example, given how highly correlated liking the Wu-Tang Clan band was for male heterosexuality, liking it would “give away” a user’s sexual orientation; and given a similarly high correlation, liking Sephora would “give away” a user’s low IQ score. Although this study justified its research in terms of “warning” users of possible privacy violations, it clearly showed how to cluster users in order to estimate their “latent” characteristics. Further, it revealed which categories were best for predictably separating users. The researchers stressed that the most significant likes for any given category did not simply or literally reflect that category: among male users, “Britney Spears” was a more popular and “revealing” like for “male homosexuality” than “Being Gay.” Their analysis discovered subcultural style cues, which signaled group membership to those in the know (for more on this, see chapter 4).

Crucially, the predictions were trained on carefully curated data, which determined both the coefficients of the regression models and their significant components. The models were then tested on their ability to predict this meticulously pruned past. This is not specific to these

Trait		Selected Most Predicted Likes	
Sexual Orientation	<div>IQ</div> <div>High</div>	The Godfather Mozart Thunderstorms The Colbert Report Morgan Freeman's Voice The Daily Show Lord of the Rings To Kill a Mockingbird Science Curly Fries	Jason Aldean Tyler Perry Sephora Chi Bret Michaels Clark Griswold Bebe I Love Being a Mom Harley Davidson Lady Antebellum
	<div>Homosexual Males</div>	No H8 Campaign Kathy Griffin Kurt Hummel Glee Human Rights Campaign Mac Cosmetics Adam Lambert Ellen DeGeneres Juicy Couture Sue Sylvester Glee Wicked The Musical	X Games Nike Basketball Bungie WWE Sportsnation Wu-Tang Clan Foot Locker Shaq Bruce Lee Being Confused After Waking Up From Naps
	<div>Homosexual Females</div>	Girls Who Like Boys Who Like Boys Rupauls Drag Race No H8 Campaign Gay Marriage Human Rights Campaign The L Word Sometimes I Just Lay In Bed and Think About Life Not Being Pregnant Gay Marriage Tegan And Sara	Lipton Brisk Yahoo Adidas Originals Foot Locker WWE Inbox 1 Makes Me Nervous Thinking Of Something And Laughing Alone I Just Realized Immature Spells I'm Mature Did You Get A Haircut No It Grew Shorter Nike Women

15 Postpredictive likes for dichotomous categories, redrawn from Kosinski, Stillwell, and Graepel, "Private Traits and Attributes Are Predictable," Table S-1.

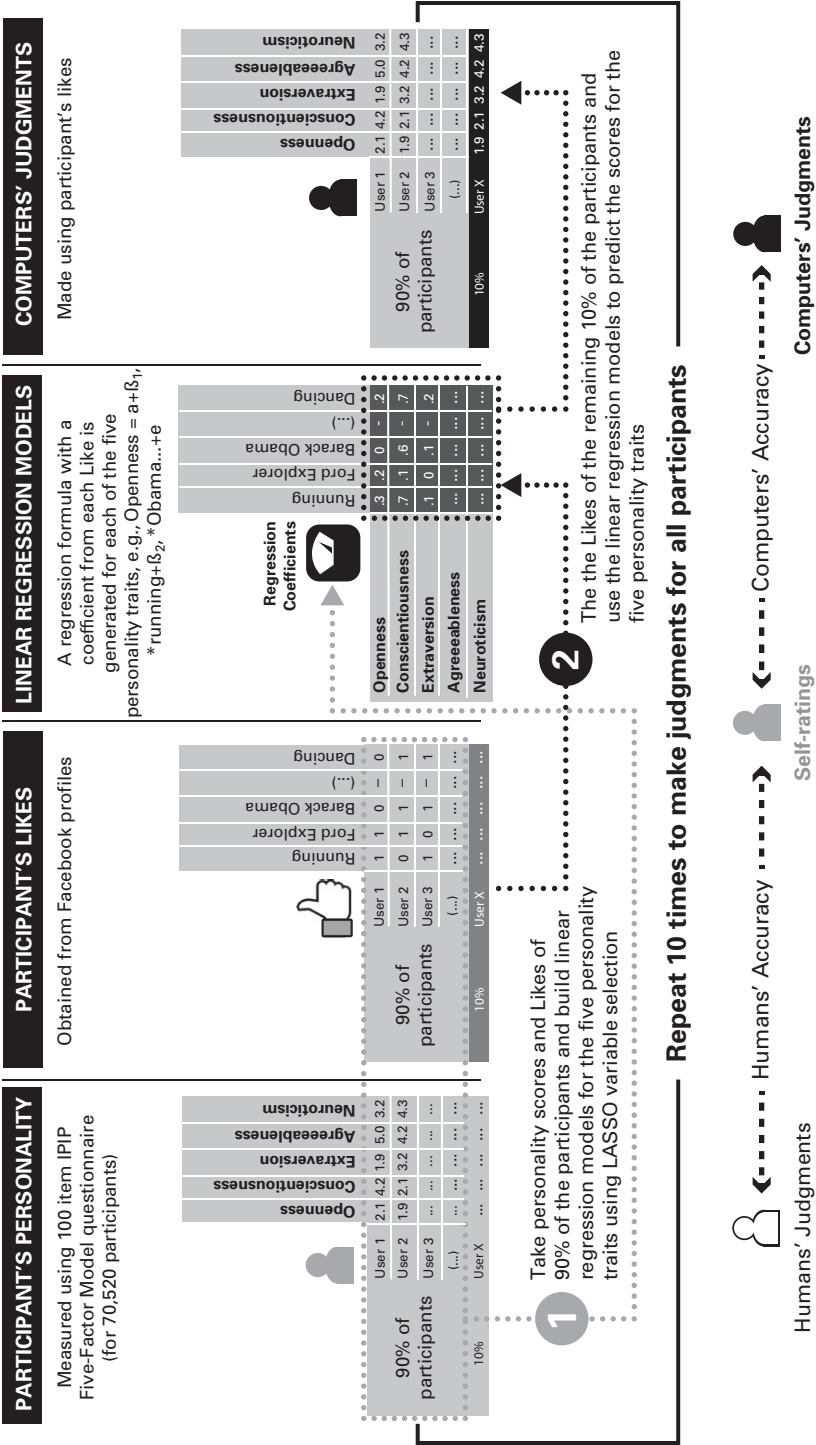
models or these types of models. Even more explicitly predictive algorithms, which tolerate higher bias and lower variance in order to avoid “overfitting,” are verified as correct if they predict the past correctly, for they are usually cross-validated using past data that are hidden during the training period or out of sample data, similarly drawn from the past.¹⁸ Wu Youyou, Michal Kosinski, and David Stillwell used this cross-validation test in their 2015 follow-up study, “Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans.”¹⁹ Unlike the

first study, the later study emphasized the importance of cross-validation. It used 90 percent of its data as a training set to build a linear regression model for predicting personality type, and then tested it against the remaining 10 percent for verification (figure 16).

Using this form of verification, standard for machine learning algorithms and models, means that if the captured and curated past is racist and sexist, these algorithms and models will only be verified as correct if they make sexist and racist predictions, especially if they rely on problematic measures such as standard IQ tests. Tellingly, low IQ in the 2013 study was found to be highly correlated with liking “I Love Being a Mom.”

The methods used by Kosinski and colleagues and Cambridge Analytica—correlation, linear and logistic regression, and factor analysis—stem from twentieth-century eugenics. The five-factor OCEAN model is the product of controversial and discredited eugenicists such as Charles Spearman, Hans Eysenck, and Raymond Cattell. They developed and used factor analysis, based initially on principal component analysis (PCA; see figure 37 by Alex Barnett in “Proxies, or Reconstructing the Unknown” after chapter 2) and correlation to “classify” raced and gendered groups according to intelligence, among other personality traits.²⁰ The “O” in OCEAN, “openness,” was initially labeled “intellect,” which means that those responding to Cruz’s “from father to son: from the birth of our nation” Second Amendment ads would once have been labeled “low intellect.”²¹ In the “five-factor” world, personality traits or factors were, and still are, considered “physiological.” According to Robert McCrae and Geert Hofstede, the “five-factor model” was “unique in asserting that traits have only biological bases”²²—an assertion that provided the basis for researchers using the model to frame personality within a biometric evolutionary schema.²³

These attempts to more finely “resolve” human groupings based on personality reinforce racial boundaries. Indeed, the images Cambridge Analytica used in its psychographic messaging were deeply raced, gendered, and classed (figure 10). Thus not only were the images of “persuadables” in figure 10 both white, the tagline “since the birth of our nation” riffed off D. W. Griffith’s racist *Birth of the Nation*, a 1915 silent film that valorized the Ku Klux Klan. Psychographics created connections not across races, but rather divisions within them. Although Nix argued



against all women receiving the same message according to their gender or all African Americans according to their race, he did not argue for messages that would cross gender or racial boundaries. He admitted that Cruz's challenge in Iowa was having "his voice heard" by "a largely homogeneous audience" of "largely white, middle-aged, male, conservatives, in support of the economy and the Second Amendment."

According to Christopher Wylie, to more effectively influence people, Cambridge Analytica took an "intersectional" approach to racial identity. Steve Bannon, he explained, was the only straight man he talked to about feminist intersectional theory, a methodology developed by women of color feminists, most notably legal scholar and critical race theorist Kimberlé Crenshaw, to explain how black women are at the "crossroads" of race, gender, and class. Crenshaw's analysis focused on how broad identity categories, such as female and black, often effectively exclude black women, and how this exclusion could be best addressed through a coalitional understanding of identity.²⁴ Cambridge Analytica perverted Crenshaw's method and sought to augment differences and exclusions, both real and perceived, based on values and "personality traits," within a racially homogenous space. Whereas Crenshaw started with how feminism and black empowerment movements often fail to address the needs and concerns of black women—for example, how funding agencies for support centers presume that rape victims are white middle-class women—in order to overcome these problems, Cambridge Analytica sought to find "neighborhoods" within identity categories such as "gunning white men" to better target and transform individuals.

Put most bluntly: in an attempt to destroy any and all senses of commonality, "communities" are being planned and constructed based on divisions and animosities. Instead of ushering in a post-racial, post-identitarian era, these social networks perpetuate angry microidentities through "default" variables and axioms. By using data analytics, individual differences and similarities are actively sought, shaped, and instrumentalized in order to capture and shape social clusters. Networks are neither unstructured masses nor endless rhizomes that cannot be cut or traced. Because of their complexities, noisiness, and persistent inequalities, networks provoke control techniques to manage, prune, and predict. This method—pattern discrimination 2.0—makes older, deterministic,

or classically analytic methods of control through direct discrimination seem innocuous.

Welcome to the swarming of the segregated neighborhood, spread through eugenic methods to cultivate futures based on mythical pasts.

CORRELATION, CORRELATION, CORRELATION

The ground beneath our feet is shifting. Old certainties are being questioned. Big data requires fresh discussion of the nature of decision-making, destiny, justice. A worldview we thought was made of causes is being challenged by a preponderance of correlations. The possession of knowledge, which once meant an understanding of the past, is coming to mean an ability to predict the future.

—Viktor Mayer-Schönberger and Kenneth Cukier, 2014²⁵

I felt like a buccaneer of Drake's days—one of the order of men "not quite pirates, but with decidedly piratical tendencies." . . . I interpreted . . . Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment. It was Galton who first freed me from the prejudice that sound mathematics could only be applied to natural phenomena under the category of causation. Here for the first time was a possibility—I will not say a certainty, of reaching knowledge—as valid as physical knowledge was then thought to be—in the field of living forms and above all in the field of human conduct.

—Karl Pearson, 1934²⁶

Correlation grounds big data's so-called revolutionary potential. As *Wired* editor Chris Anderson infamously declared in his 2008 editorial "The End of Theory," big data proved that "correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."²⁷ Less controversially, policy researcher Viktor Mayer-Schönberger and journalist Kenneth Cukier, in their popular 2014 book *Big Data: A Revolution That Will Transform How We Live, Work and Think*, asserted that, by replacing causality with "simple correlations," big data "challenges our most basic understanding of how to make decisions and comprehend reality."²⁸ Indeed, by substituting "what" for "why," they claim that big data and correlation have changed the direction of knowledge: it is no longer about understanding the past, but rather about grasping the future.

Not surprisingly, big data, also formally called “data analytics,” was immediately dismissed as “hype”: the latest in a long line of technoutopian (and dystopian) fads. Google Flu Trends, for example, was shown to be wildly inaccurate—predicting double the number of actual cases.²⁹ Although understanding the limits of data analytics is important, simply dismissing it as “hype” or celebrating its “missed” predictions as evidence of human unpredictability is dangerous. The gap between prediction and actuality should not give rise to snide comfort especially since random or “diverse” recommendations are often deliberately seeded in order to provoke spontaneous behavior.³⁰ Further, big data posed and still poses fascinating computational problems—How do we analyze data we can read only once, if at all?—and the plethora of correlations it documents raises fundamental questions about causality. If almost anything can be shown to be real, if almost any correlation can be discovered, how do we know what is true? The “pre-big data” example of the “Super Bowl predictor” nicely illustrates this dilemma: one of the “best” (most consistently correct) predictors of the U.S. stock market has been which football conference wins the Super Bowl: if a team from the original National Football League wins, it will most likely be a bull market; if a team from the original American Football League, most likely a bear market.³¹ Moreover, calling a new technology “hype” is hardly a profound criticism. Hype is part and parcel of new technologies, and demos of future technologies seem to elicit more praise or condemnation than everyday experiences of already existing ones (the Valley lives and dies by the demo).³² Thus to understand the impact of the “data deluge,” we need to move beyond celebrating or dismissing big data toward comprehending the force of its promise—or, more precisely, the ways it undermines the promise of promise. As philosopher Jacques Derrida has argued, a promise that is “automatically kept” is no promise at all, but rather “a computer, a computation.”³³ Perhaps, but computations do not automatically execute themselves, and actual computers fail all the time—something we know from experience, but, surprisingly, *not* in theory.

Again, this is not the first time that correlation has been heralded as revolutionary. More than a century ago, biometric eugenicists Francis Galton and Karl Pearson “discovered” correlation in their attempts to determine heredity. As quoted in this section’s second epigraph, Pearson

described feeling like “a buccaneer” on the edge of plunder and discovery because correlation expanded knowledge beyond causality and promised to make mathematically comprehensible living beings and human behavior. Pearson’s hyperbolic rhetoric foreshadows twenty-first-century big data hype. Correlation’s eugenicist history matters, not because it predisposes all uses of correlation towards eugenics, but rather because when correlation works, it does so by making the present and future coincide with a highly curated past. Eugenicists reconstructed a past in order to design a future that would repeat their discriminatory abstractions: in their systems, learning or nurture—differences acquired within a lifetime—were “noise.” The important point here is that predictions based on correlations seek to make true disruption impossible, which is perhaps why they are so disruptive.

The differences between twenty-first-century big data and twentieth-century eugenics, as the end of this chapter explains in greater detail, also matter. The move from statistics to data science signals a difference in purpose and focus. As philosopher of science Ian Hacking has pointed out, the term “statistics” comes from “state,” and national statistics testify to a state’s “problems, sores and gnawing cankers.”³⁴ Data science, in contrast, by focusing on the governmental interests of corporations and states through “network neighborhoods” or “clusters,” outlines possible “homophilic escapes” from national populations. For the twentieth-century eugenicists, homophily was an aspiration: they wanted to create a world in which like people automatically reproduced with like. In data analytics, homophily is a given, an axiom. Nightmares of global destruction and dreams of segregated “escape” have displaced narratives of impending racial doom. So how did we get here, and what is correlation anyway?

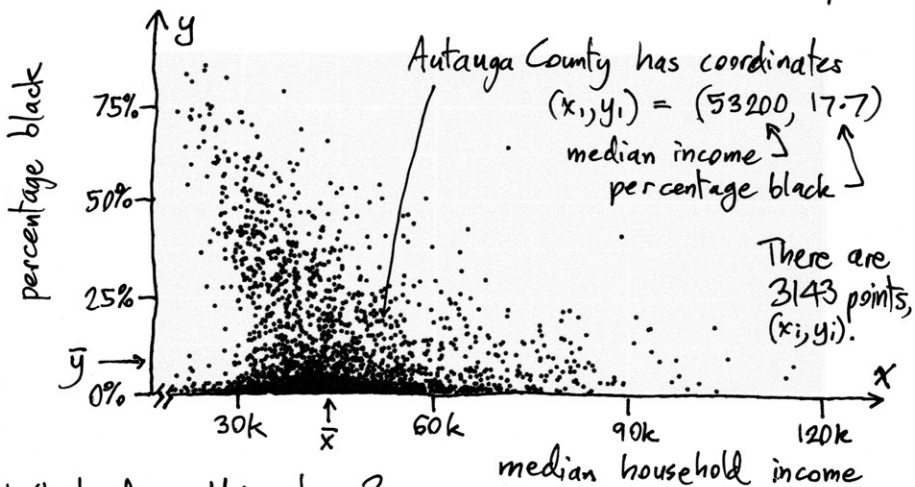
SPURRING CORRELATIONS

Most basically, correlation measures how two or more variables vary together. If variables increase and decrease in step, they are highly (positively) correlated; if they vary in opposite directions, they are negatively correlated (see figure 17).

Highly correlated variables are thus considered to be “proxies” of each other: by tracking one variable, you can capture the other. Correlations

CORRELATION

There are $n=3143$ counties in the US, and lots of publicly available data about them. (Here we use the "countyComplete" data in the "openintro" package for the free statistical software "R". Most data is from 2010.) Counties are indexed $i=1$ to 3143. Eg, $i=1$ is Autauga County, AL. Let's plot on the x-axis median household income, vs the y-axis the percentage of the county population that is black. This is a "scatter plot":



What does this show?

- A correlation between race & poverty: the points lean leftwards as one moves up. ($\approx 4k$ less per 10% increase)
- Counties with income $> 70k$ are almost all $< 20\%$ black. Thus income can be a surrogate for race.
- Poor counties are segregated: for incomes $< 30k$, the distribution is "bimodal", very white ($< 3\%$) or black ($> 30\%$).

So, a scatter plot can tell many stories. However, often only Pearson's "correlation coefficient" is given, which mathematically is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

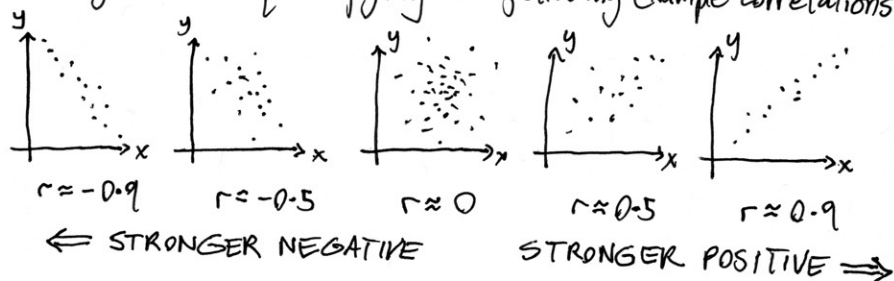
This (perhaps scary) formula involves two familiar quantities:

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the mean of the income over countries.

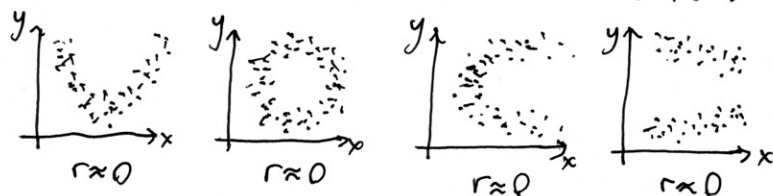
$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of the percentages.

For our data $\bar{x} \approx 44k$, $\bar{y} \approx 9\%$, and these are shown on the plot.

r is good at quantifying the following example correlations:



However there are many interesting & informative "nonlinear" correlations that r is oblivious to:



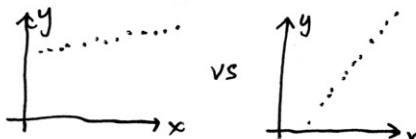
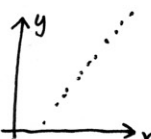
↳ in each case there are correlations, but r cannot tell you this fact! It is insensitive to bimodality (the indicator of segregation earlier).

Returning to our county income and percentage black data, what is r ? It turns out to be $r \approx -0.22$, which is negative (as expected from the overall downwards slope), but would be interpreted as very weak.

This shows the limitation of the correlation coefficient: it fails to capture the many aspects that a glance at the full scatter plot can show. One must look at the data rather than trust r .

Notes:

- you do not need to handle the formula for r : all statistical software has it built in.
- r lies between -1 and $+1$, and tells you the strength of the linear correlation, not to be confused with the strength of the effect (which r does not tell you).

Eg.  vs  Both have $r \approx 1$, but in the 2nd case y changes much faster with x .

- scatter plots can be 3D too with (x_i, y_i, z_i) data, or even higher dimension, but it is hard to picture!
- a better analysis of county data might "weight" each point by the county population.
- nonlinear correlations (bimodality, etc) can be found by using, eg, powers of variables, x^2 , x^3 , etc.

are most often used to uncover latent or hidden variables. In the Kosinski, Stillwell, and Graepel 2013 study, tracking the like “I Love Being a Mom” supposedly captured intelligence. Such correlation tracking provides the basis for Anderson’s assertion that theory is dead, or Mayer-Schönberger and Cukier’s that correlation gives us the future rather than the past.

Many researchers who deploy data-driven techniques have qualified or critiqued these broad proclamations of the death of causality. As sociologists Josh Cows and Ralph Schroeder explain, instead of either correlation or causality alone, what is necessary are “mixed methods” that combine correlational exploratory practices with causal explanatory research.³⁵ This is because, left unattended, big data methods often reinvent the wheel by “discovering” well-known latent correlations (that many gay men of a certain age like Britney Spears, to return to an example referenced earlier), or they produce an inordinate number of spurious correlations that defy basic concepts such as gravity or photosynthesis. Further, causality is often needed to solve problems—vaccines, for example, depend on mechanistic understandings of virus structure and behavior.

In addition, correlations often raise as many questions as they supposedly answer. For example, social scientists Nicholas Christakis and James Fowler’s much cited and disputed 2007 study of friendship data, which recycled data from the Framingham Offspring Study (begun in 1971), concluded that social, rather than physical, proximity to one or more persons who are obese matters most in predicting the likelihood of someone becoming obese.³⁶ Obesity, that is, spreads like a virus through social networks. This study was criticized not only for its conclusions but also for its conflation both of obesity with viruses and of viral spread with homophily (the tendency of individuals who are like each other to act similarly in the same context). As statisticians Cosma Shalizi and Andrew Thomas point out, it is mathematically difficult to separate habit from contagion.³⁷ Further, other seemingly contradictory correlations were also documented. Another study found that zip code and property value were strong proxies for obesity.³⁸ Further, spurious correlations arrived at using big data are not accidental; indeed, drawing on mathematical theory, theoretical computer scientists Cristian Calude and Giuseppe Longo have shown that, because of their size alone, all big data analyses must be riddled with such correlations.³⁹ And, for that matter, spurious correlations abound in

small data sets as well, the classic example being the Super Bowl market indicator mentioned earlier.⁴⁰

Traditionally, causality cuts through multiple correlations in order to find the things that really matter. As defined within the quantitative social sciences, causality depends on three conditions: (1) correlation; (2) the cause preceding the effect; and (3) the absence of a third variable that could explain the correlation.⁴¹ This definition draws from the more technical Wiener–Granger test for causality, commonly used in econometrics and neuroscience to determine if two variables, X and Y , are causally related. Y is said to be Wiener–Granger causal if it improves the prediction of X in a statistically significant way.⁴² In synchronous network models, simulations and parsimony are used to determine truth.⁴³

Spuriousness, however, is not the sole or even the main problem with correlations. As Cathy O’Neil and others have shown, correlations can perpetuate inequality. Those building what O’Neil has called “weapons of math destruction” use correlations and proxies to compensate for ignorance or lack of evidence. Since they cannot directly access the behavior they are most interested in, they use proxies as stand-ins: “They draw statistical correlations,” O’Neil tells us, “between a person’s zip code or language patterns and her potential to pay back a loan or handle a job. These correlations are discriminatory, and some of them are illegal.”⁴⁴ That is, correlations can serve as proxies for unknown or protected categories—categories that were deliberately hidden or unrecorded in an attempt to ensure equal treatment.⁴⁵

Proxies that uncover the obvious consequences of discrimination often work—they effectively target groups. As O’Neil notes, “rich people buy cruises and BMWs. All too often, poor people need a payday loan.” Because of this, “investors double down on scientific systems that can place thousands of people into what appear to be the correct buckets. It’s the triumph of Big Data.”⁴⁶ As this example makes clear, these models not only “discover” the effects of discrimination; they also automate and perpetuate them for they exploit, rather than remedy, inequalities. These correlations are at the heart of what communications scholar Oscar Gandy, writing in 2009, eight years before O’Neil, identified as “technologies of rational discrimination”: unless there is a clear determination not to discriminate, Gandy explained, these technologies perpetuate inequality by

creating and comparing “analytically generated groups in terms of their expected value or risk.”⁴⁷ That homophily drives these groups and correlations “that work” is no accident. As we will see in chapter 2, homophily, based on historical trends and actions, does in fact explain some behavior; the future does at times repeat the past. But this raises at least two interesting questions: In a dynamic world dominated by change, under what circumstances and to what end do some things seemingly repeat? And how does the ephemeral endure through our habitual actions? As philosophers as diverse as the Buddha and Gilles Deleuze, and as molecular biologists have shown, we live in a world of constant change—no two things are exactly alike, not even ourselves at different moments in time. Recognition always entails misidentification—an obscuring of present or future differences to past acquaintance.

Correlations, again, do not simply predict certain actions; they also form them. Correlations that lump people into categories based on their being “like” one another amplify the effects of historical inequalities. A signature quality of a weapon of math destruction is that the weapon “itself contributes to a toxic cycle and helps sustain it.”⁴⁸ Virginia Eubanks in *Automating Equality* offers a classic example of this: the Allegheny Family Screening Tool (AFST), used by Allegheny County, Pennsylvania, to determine the risk of child abuse and neglect.⁴⁹ Since the AFST training set was drawn from families who access public services, the use of public services itself became classified as a risk factor. This, like the Chicago police’s heat list, which lumped together murderers and murdered as “likely to be involved in a homicide,” erased the difference between victim and perpetrator. Children’s involvement with protective services became evidence of their likelihood as adults to abuse or neglect their own children. Families with private insurance or who used private services, such as therapists and nannies, on the other hand, were not included in the training data set and thus not flagged.⁵⁰ O’Neil also points to the unfair impact that credit ratings can have when factored into hiring programs. Produced by licensed agencies and more informal data brokers, and based on individual actions and increasingly social networks, these ratings are not simply proxies for responsibility: people who live from paycheck to paycheck have trouble maintaining their credit ratings during hard times, unlike those who are wealthy. Given the U.S. history

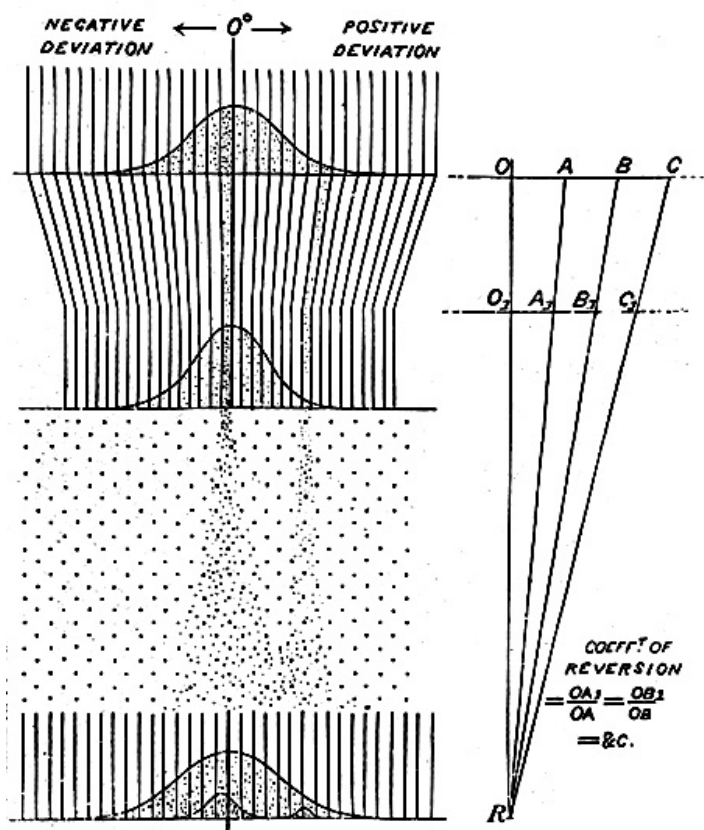
of financial discrimination explored in detail by Oscar Gandy, U.S. credit ratings correlate with race, or more precisely racism.⁵¹ As political scientist Ira Katznelson, policy researcher Richard Rothstein, and many others have shown, U.S. government policies such as the New Deal, Social Security, inexpensive mortgages, and the G.I. Bill concentrated wealth in the hands of white Americans.⁵² Weapons of math destruction automate and amplify past inequalities through their baseline correlations.⁵³

The problems with correlations are neither new nor limited to big data and weapons of math destruction, however. Based on eugenic reconstructions of the past and cultivated to foreclose the future, correlation contains within it the seeds of manipulation, segregation and misrepresentation.

REDISCOVERING OUR EUGENIC FUTURE

British eugenicists developed correlation and linear regression, key to machine learning, data analytics, and the five-factor OCEAN model, at least a century before the advent of big data. Although methods for linking two variables preceded his work, Francis Galton is widely celebrated for “discovering” correlation and linear regression, which he first called “linear reversion.” Second cousin of Charles Darwin, Galton is also considered the progenitor of the five-factor model and the “father” of eugenics, which, in Karl Pearson’s paraphrase, he defined as “the science of improving stock, not only by judicious mating, but by all the influences which give the more suitable strains a better chance” and which Galton agreed in a Cambridge lecture was “the study of those agencies which under social control may improve or impair the racial qualities of future generations, either physically or mentally.”⁵⁴ Correlation was key to “proving” that these agencies were natural rather than social. Correlation was never simply about discovering similarities, but also about cultivating physical similarities in order to control the future. Correlation provided the basis for eugenics’ “universal laws.”

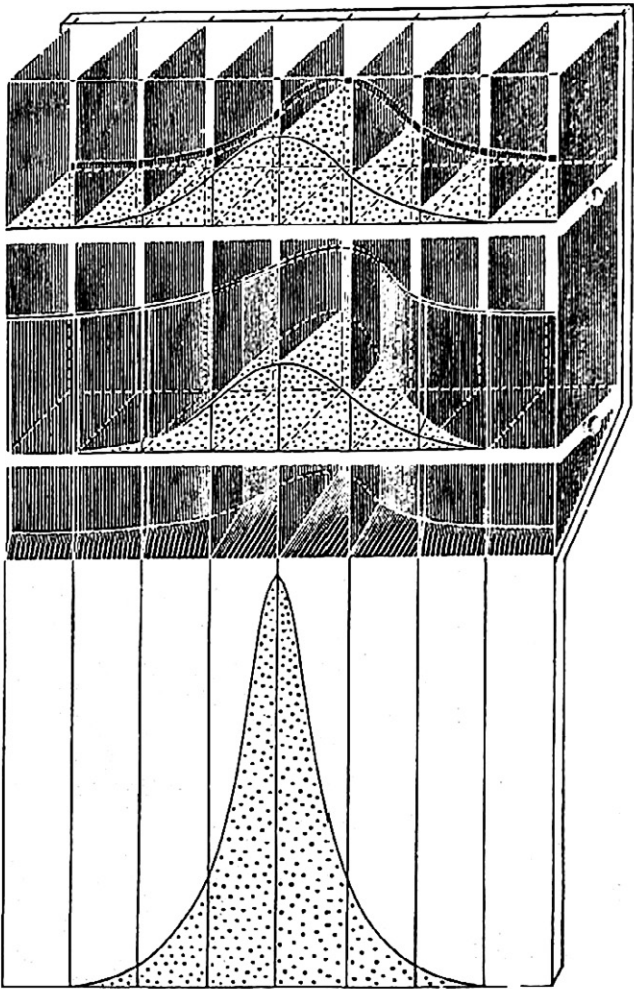
As Ruth Cowan and other historians of science have shown, Galton developed regression and correlation while studying heredity in humans and plants and the identification of criminals.⁵⁵ His fascination with the inheritance (or not) of genius (based on his undergraduate experiences at Cambridge University with the offspring of various famous families)



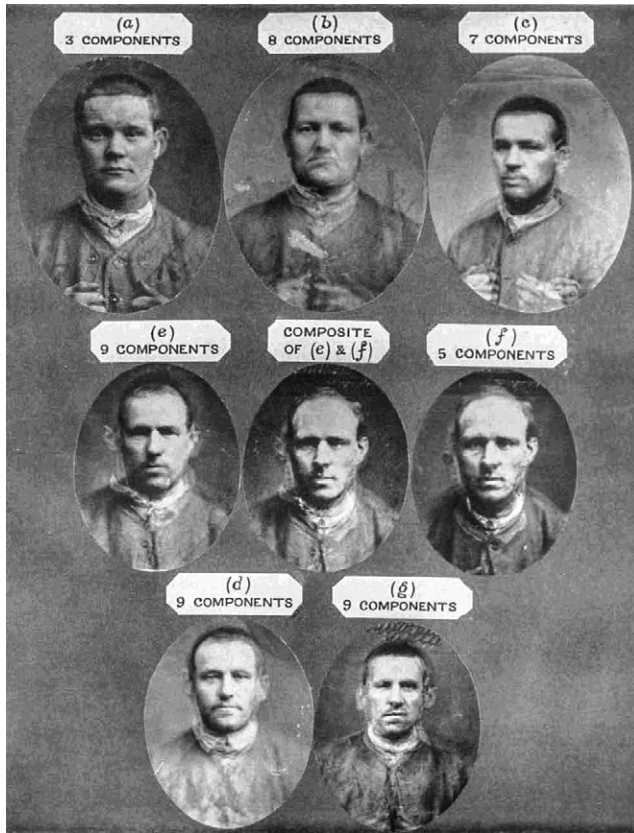
18 Galton's diagram of linear reversion. Karl Pearson, *The Life, Letters and Labours of Francis Galton*, vol. 3a, *Correlation, Personal Identification and Eugenics* (Cambridge: Cambridge University Press, 1930), 9.

moved him to write *Hereditary Genius*, first published in 1869.⁵⁶ Galton developed a "law of inheritance," expressed as a mathematical formula to quantify the contribution of each generation to the next. He first produced what would become linear regression while studying the variation in size between sweet pea and human parents and their offspring.

Figures 18 and 19 reveal Galton's overriding concern with deviation in offspring and its transmission to future generations. A biometrician rather than a Mendelian, Galton believed that all traits were distributed along a normal curve within a population, rather than determined by genes.⁵⁷ Exceptions, such as genius, were statistical outliers and thus located at the ends of the curve, in the fourth quartile. Since Galton wanted to preserve

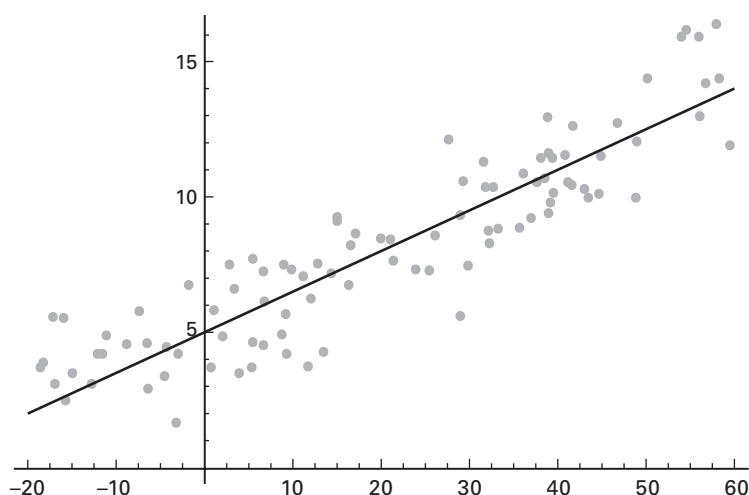


19 Galton's diagram explaining the influence of natural selection on reversion. Karl Pearson, *The Life, Letters, and Labours of Francis Galton*, 3a:10.



20 Francis Galton's "Criminal Composites," c. 1878. Plate XXVII from Karl Pearson, *The Life, Letters and Labours of Francis Galton*, vol. 2, *Researchers of Middle Life* (Cambridge: Cambridge University Press, 1924), 286.

and amplify "good" deviations, his curve tracked how deviations from the norm changed from one generation to the next (figure 18). To explain the effect of natural selection, he employed tubes, which he angled to produce more or less sharp bell curves, and therefore more or fewer outliers (figure 19). According to Galton, his graphs proved that offspring were "reverting" (later, "regressing") to an ancestral mean. He initially thought that only spontaneous deviations ("sports") induced through natural selection, could change the ancestral norm. This notion of a primordial mean also influenced his experiments with photography, in which he overlaid multiple exposures of criminals, alcoholics, and Jewish boys,



21 Standard linear regression, created by Joshua Cameron.

among many others, in order to reveal the archetype embedded within these individuals (figure 20, further discussed in chapter 4).

Galton's linear reversion thus differed significantly from the now standard linear regression. In tracking how generations deviated from the norm, his goal was to maximize "good" deviation. In contrast, linear regression seeks to minimize standard deviations and is most simply expressed by the equation $y = mx + b$, where m is the slope of the line mapping x onto y (figure 21), y is the dependent variable, and x is the independent variable.

In the Kosinski, Stillwell, and Graepel 2013 study discussed earlier, y would be the degree of being an extrovert and x would be a particular SVD component comprised of relevant Facebook Likes (beerpong, Michael Jordan, Dancing were the most highly correlated Likes for extroversion) and m the weight given to that particular component. Linear regression is typically used to determine the best line between a scattered set of points, where "best" means the line that minimizes the distance between the data points and the projected line.

Galton's concept of correlation also emerged from Galton's dispute with French police detective Alphonse Bertillon regarding the best way to identify criminals. As further explained in chapter 4, Bertillon had

developed a system of nine measurements to supplement mug shots. Galton believed that some of Bertillon's nine measurements, such as the length of a person's arm and the length of the person's leg, were linked together and therefore redundant.⁵⁸ To prove these measurements were not independent, he produced a coefficient that linked these variables.⁵⁹ In this version of correlation—a version more commonly used in statistics—correlation is used to cut down on the number of variables involved, not to uncover “hidden” or latent variables.

Galton's facility with mathematics was intuitive, but limited. Tellingly, for example, he used quartiles rather than standard deviations. Karl Pearson made Galton's concepts more mathematically precise. Still in use today, the Pearson correlation coefficient provides a measure from -1 to $+1$ for a correlation by dividing the product of the variations of two variables by the product of their standard deviations (see “Correlation” by Alex Barnett; figure 17). Pearson updated Galton's law of ancestral heredity by arguing that, although the generations varied linearly, the influence of ancestors on their offspring diminished geometrically,⁶⁰ a conclusion he came to while studying the transmission of physical traits across generations and the differences between twins. Although not convinced that mental traits always corresponded to physical ones (as opposed to Galton, who was infatuated with phrenology and believed that skull size was a proxy for intelligence), Pearson was certain that physical and mental traits followed the same ancestral law. Diminishing skull size thus did not equal diminishing intelligence, but rather skull size and intelligence diminished in an analogous, geometrical fashion.⁶¹

Pearson also believed both natural and artificial selection could easily and continuously affect future generations: the past and future were linked linearly. In contrast, Mendelian eugenicists did not hold such a simple, progressivist view since regressive traits could reappear at any time and thus frustrate phenotype-based breeding. According to Charles Davenport, a U.S. Mendelian contemporary of Pearson's, one “defective” yet fecund individual, such as the infamous Max Juke, could have a profound impact on the population of a nation.⁶² Mendelian eugenicists thus sought to create “pure” bloodlines cleansed of “undesirable” traits, whether dominant or recessive, whereas biometricians viewed racial or national populations as inherently mixed and intermingled; there was no

“pure” breed, and positive deviations needed to be preserved and disseminated. Eugenists in both camps, however, held individuals responsible for the future: their behavior could either benefit or destroy the nation.⁶³ And both camps believed that nature triumphed over nurture, making eugenics central to breeding a “better” national future.

The biometricians’ belief in the geometrical law of ancestral heredity made cultivating a “better” future much easier for them than Mendelians. Ominously in light of what was to come decades later, Pearson asserted that correlation helped society move towards a “final solution of almost any social problem,” for it revealed how nature triumphed over nurture, how “selection of parentage is the sole effective process known to science by which a race can continuously progress.”⁶⁴ This conclusion is not surprising given their methodology: biometricians classified all similarities as “hereditary,” and all differences as “environmental.”⁶⁵ Since commonalities outweighed differences, Pearson asserted, “there is no real comparison between nature and nurture; it is essentially the man who makes his environment, and not the environment which makes the man.”⁶⁶ In terms of intelligence, he asserted that although “intelligence could be aided and trained . . . no training or education could create it. It must be bred.”⁶⁷ Programs to alleviate the appalling conditions of working-class Britons and to provide them with educational and medical support were therefore a waste of time and money. Thus Pearson, an avowed socialist, declared: “Give educational facilities to all, limit the hours of labour to eight-a-day—providing leisure to watch two football matches a week—give a minimum wage with free medical advice, and yet you will find that the unemployables, the degenerates and the physical and mental weaklings increase rather than decrease.”⁶⁸ Moreover, by suspending the work of natural selection, these social uplift programs threatened to destroy the English race: through them, the “unfit” multiplied at the expense of the “fit.”⁶⁹ In the nationalist view of biometric eugenics, every citizen was connected: natural and artificial selection operated at the level of the nation-state.

After Nazi Germany was defeated and the horrors of the Holocaust exposed, eugenics seemed to die away or to transform itself into genetics—only to reappear, as many saw it, in the form of genetic tests for birth defects, artificial insemination, and “designer babies.” In the late

twentieth century, historian of biology Nils Roll-Hansen described an “inescapable eugenics,” based on current progress in molecular genetic knowledge,⁷⁰ and sociologist Troy Duster contended that the modern resurgence of biological definitions of race have created a “backdoor to eugenics.”⁷¹ In contrast, sociologist Nikolas Rose argued that, because eugenics focused on the population, not the individual, genetic “improvements” to the individual are not eugenic.

Highlighting the reemergence of biometrics in the twenty-first century, this chapter and book enter this debate, in conversation with work on the resurgence of biometrics by new media researchers such as Jacqueline Wernimont, by asking: To what extent has eugenics reemerged—if it has—not simply or directly through the proliferation of genetic testing and manipulation, but also through biometric methods and predictions?⁷² And how have data analytics and machine learning been used to found a revised form of eugenics, in which discriminatory pasts, presents, and futures coincide? Again, to be clear, I am not claiming that the methods developed by biometric eugenicists are inherently eugenicist. As we will see in later chapters, correlation has been key to developing explanatory global climate change models; it is also mirrored in studies of ideology and ideology critique. Rather, I am asking:

To what extent do the current descriptions of correlation as unlocking the future reflect the twentieth-century celebrations of correlation and its confidence in eugenic solutions?

To what extent can understanding this mirroring help elucidate why and how the world of data analytics and machine learning, based on methods arising from these descriptions, feels so small and enclosed?

And how did a worldview that did not believe learning could happen—that intelligence could only be bred—become the basis for machine learning?

OUR EUGENIC FUTURE, AGAIN

In addition to treating correlation as inherently predictive, there are many similarities between twentieth-century eugenics and twenty-first-century data analytics. Both emphasize data collection and surveillance, especially of impoverished populations; both treat the world as a laboratory; and both promote segregation.

Eugenics and big data depend on surveillance, especially of the poor. Karl Pearson, Francis Galton, and Charles Davenport all argued that the future of eugenics depended on the gathering of national statistics. Since it aimed to show “how much harm is being done by some one course of action, and how much good by some other, and how closely connected social practices are with the future vigour of the nation,”⁷³ eugenics required detailed surveillance of human populations. Eugenacists thus collected data to produce charts documenting the transmission of traits (such as criminality). Their goal was to accumulate the “knowledge” necessary to foster reproduction of the “fit,” as well as to impair that of the “unfit,” either voluntarily or involuntarily. Eugenacists generally studied the poor in order to “save” the middle classes and the rich: by studying the transmission of “negative traits,” the middle classes could learn how to “marry intelligently” and how to segregate themselves from the “unfit.”⁷⁴ The research centers, chairs, and journals founded by these eugenacists—most notably, the Cold Spring Harbor Laboratory, the Galton Chair in National Eugenics (now the Galton Professor of Human Genetics), and the journal *Annals of Eugenics* (now *Annals of Human Genetics*)—still exist, although all now engage primarily in genetics.

Virginia Eubanks has linked twentieth-century eugenics to twenty-first-century data analytics and machine learning through their practices of surveillance. Eugenics she has revealed, “created the first database of the poor,”⁷⁵ and contemporary programs to automate public services programs have given rise to digital poorhouses, all too similar to the physical poorhouses of the nineteenth century, which imprisoned and punished the poor: “Marginalized groups face higher levels of data collection when they access public benefits, walk through highly policed neighborhoods, enter the health-care system, or cross national borders,” in what amounts to “feedback loop[s] of injustice.”⁷⁶ As the Allegheny Family Screening Tool example mentioned earlier illustrates, this data collection—ostensibly designed to help streamline public services—usually makes things more difficult for those these services are supposed to aid and places them under additional surveillance.

Data analytic methodologies, Eubanks warns, are not limited to the poor. As the term “training” implies, once “perfected,” machine learning

programs are meant to be let loose on the general public. As one of her informants cautioned: "Poor women are the test subjects for surveillance technology. . . . You should pay attention to what happens to us. You're next."⁷⁷ Analyzing this "progression" in her 2015 study "First They Came for the Poor: Surveillance of Welfare Recipients as an Uncontested Practice," policy analyst Nathalie Maréchal places Edward Snowden's leaks and post-911 surveillance next to the systematic infiltration and spying on African American communities, civil rights activists, and antiwar groups throughout the 1950s, 1960s, and early 1970s under the counter-intelligence program (COINTELPRO).⁷⁸ Indeed, the "progression" is part of the historical spread of control technologies: as historian Chandak Sengoopta has shown, fingerprinting started in colonial India as a way for the English to control and distinguish between "the natives," and the timetable, as visual studies scholar Nicholas Mirzoeff and Black studies and surveillance studies researcher Simone Browne have elaborated, originated on the Southern plantation—to give just two of many instances.⁷⁹ That reactionary publics in the twenty-first century would "draw from" the civil rights movement is thus to be expected, since those fighting for decolonization and civil rights were the first to battle these systems.

Both eugenics and big data use surveillance in order to experiment with humans. Eugenicists drew from the history of animal husbandry and agriculture to justify their goal to breed a better "human crop." Eugenics began with Francis Galton's Darwinian realization that humans were a species like all other animals: what applied to other animals and to plants therefore applied to humans. The eugenicists' insight that nature trumps nurture is said to have emerged from the work of "intelligent farmers and gardeners." In the words of Francis Galton: "I perceived that the importance ascribed by all intelligent farmers and gardeners to good stock might take a wider range. . . . All serious inquirers into heredity now know that qualities gained by good nourishment and by good education never descend by inheritance, but perish with the individual, whilst inborn qualities are transmitted. It is therefore a waste of labour to try so to improve a poor stock by careful feeding or careful gardening as to place it on a level with a good stock."⁸⁰ This crop or herd metaphor extended to eugenicists themselves. Responding to critics who accused eugenicists of engaging in unethical experiments, Pearson explained that

eugenicists were not farmers or owners, but “members of the herd.” No one was outside eugenics. Rather than manipulate their fellow humans, they, like medical professionals of “the higher type,” surveilled them to track experiments already in play. Eugenics was possible because humans themselves engaged in reproductive experiments “directly impossible for the eugenicist. This stock marries kin for six generations; those parents surfeit themselves with alcohol; there the tuberculous taint meets insanity; here the man of genius marries into his class; there he takes a woman of the people.” By observing and framing the world in this manner, eugenicists claimed that they were merely forming “an analytical record of . . . the biological laws which govern [a person’s] social development” in order to offer the basis from which “to predict what lines of conduct foster, what lines check national welfare.”⁸¹

Similarly, data and network scientists describe their work as revealing the inner workings of the human psyche via experimentation. Acclaimed network scientist and author Albert-László Barabási has claimed that network science, combined with “increasingly penetrating digital technologies,” places us in “an immense research laboratory that, in size, complexity, and detail, surpasses everything that science has encountered before” and that reveals “the rhythms of life as evidence of a deeper order in human behavior, one that can be explored, predicted, and no doubt exploited.”⁸² If twentieth-century eugenicists however defended their work against accusations that it experimented on humans, twenty-first-century data scientists openly embrace experimentation. Data scientist and journalist Seth Stephens-Davidowitz openly proclaimed in 2017 that big data “allows us to undertake rapid, controlled experiments.”⁸³ These experiments have moved from simple A/B testing to so-called contextual bandits to reinforcement learning: all techniques to “optimize” content based on users’ prior interactions.⁸⁴ Twentieth-century network scientists also emphasized the importance of technology, as do their twenty-first-century successors: digital media accelerate “normal” human experiments by placing them within technologically enriched petri dishes.

Twentieth-century eugenics and twenty-first-century data analytics also both promote or presume segregation. Historians have exposed the strong ties between eugenics and segregation: eugenics was the segregationists’ science, and segregation the eugenicists’ strategy.⁸⁵ Indeed,

eugenicists—and a significant proportion of noneugenicists—supported segregation as a more “humane” alternative to sterilization, which was nonetheless regularly practiced on African Americans and other minorities in the United States without their consent as late as the 1970s in some states.⁸⁶ The forced segregation of the “mentally backward” was the only legislative success of the British eugenicists.⁸⁷ According to eugenicists, the “unfit,” especially those who could physically pass as “fit,” like the “feeble-minded,” had to be removed from the general population, and “unfit” males and females had to be kept apart from each other in order to prevent national degeneration.⁸⁸

Segregation was embraced within the United States as a way to counter any possible equalizing and liberating results of the Civil War. As historian Grace Elizabeth Hale has detailed, segregation became the principal post-Civil War strategy to establish a “myth of absolute racial difference.”⁸⁹ Responding to public displays of African American affluence in public spaces such as trains and hotels, segregation reinforced white supremacy by making “race dependent on space.” It sought to contain racial identity and cement inequality so that those “who moved within spaces marked ‘colored’ were African American, and the difference—the inferiority of the black spaces—marked the difference—the inferiority of the black and even ‘almost white’ people.”⁹⁰ Segregation “train[ed] the ground of difference”⁹¹ and, by doing so, sought to create a world in which that difference was accepted and expected. Segregation was, and still is, a training program for racism.

Segregation is also a default within network neighborhoods, in which users are clustered into neighborhoods filled with people “like them.” In networks, similarity breeds connection. Not surprisingly, U.S. residential segregation is regularly used to justify this clustering, and the ties between homophily and U.S. residential segregation, as chapter 2 will show, run deep: the term “homophily” emerged from studies by sociologists Paul Lazarsfeld, Robert K. Merton, Patricia West, and Marie Jahoda on segregated and segregating U.S. housing projects. But homophily is not the only way segregated neighborhoods enter network and data science. Machine learning is filled with “neighborhood” methods used for pattern recognition, such as “K-nearest neighbor,” “K-means testing,” and “support vector machines” (SVMs). The “K-nearest neighbor

algorithm” draws boundaries between data points based on proximity; it presumes that those data points closest to one another geographically or topographically are of the same class. K-means testing similarly uses proximity to intuit the existence of clusters, or neighborhoods. As chapter 4 explains, support vector machines, a high-dimensional method for determining boundaries between data points, are based on the linear discriminant function, developed by statistician and eugenicist Ronald A. Fisher to determine racial and species difference in characteristics such as skull size. Given this, it is not surprising that the “dark secret” revealed by network science is often racism.⁹²

POST-EUGENICS?

The differences between early twentieth-century eugenics and early twenty-first-century data analytics matter. As sociologist Donald Mackenzie notes, modern statistics may have evolved from those of eugenics but both are social and historical products.⁹³ The move from nation to neighborhood, as well as the move from forced segregation to homophily—from discrimination to recognition—changes the equations. The compressed time period (from human generations to user clicks) also alters the ways in which the present, past, and future are now intertwined. Further, with data analytics, minorities are not only surveilled and used to determine algorithmic governance; they are also excluded from certain databases so that, even though they are overrepresented in certain databases, usually having to do with criminal justice, they are underrepresented in others.

What is most significant, however, is that the eugenicist aspiration—the reproduction and selection of like with like—has now become axiomatic. The task before eugenicists was, given “the custom that prevails in America and England of free selection of mates,”⁹⁴ how to make like breed with like. R. A. Fisher similarly focused on enhancing sexual selection in order to produce a eugenic future (see chapter 4). Once homophily becomes the default, however, the task is no longer how to make like breed with like, but rather how to use this “natural” proclivity to predict and shape human behavior. The normalization of homophily in the early twenty-first century is remarkable; indeed, as late as the mid-twentieth century,

it was not a given. As chapter 2 further explains, when Paul Lazarsfeld and Robert Merton coined the term “homophily” (the tendency of like individuals to associate and bond with one another), they also coined the term “heterophily” (the tendency of *unlike* individuals to associate and bond with one another), and they did not presume homophily to be “naturally” present. Rather, they asked: “What are the dynamic processes through which the similarity or opposition of values shapes the formation, maintenance, and disruption of close friendships?”⁹⁵ Homophily in their much-cited yet seldom read 1954 study “Friendship as Social Process” is only one instance of friendship formation.

This normalization of homophily also drives the other major difference between twentieth-century eugenics and twenty-first-century data analytics: the move from the nation to the neighborhood through the notion of individual preference. Again, Nikolas Rose has argued that twenty-first-century genetics is not eugenic because it focuses on the individual and the community, rather than on the nation. Eugenics, he contended, emphasized “the links established between population, quality, territory, nation and race,” exploring the evolutionary fitness of national populations rather than the health of individuals and taking as its territory the nation rather than the “domesticated spaces of family and community.”⁹⁶ In contrast, twenty-first-century biopolitics has focused on managing individual risks, not on mandating racial “cleanliness.”⁹⁷

The move to risk and individuals, as the first part of this chapter—and as the work of Oscar Gandy, among many others, has shown—reinforces and reinscribes racial discrimination. Race is acknowledged as a “boundary” within homophily, and the shift to the individual was itself endorsed by late twentieth- and early twenty-first-century eugenicists. During a 1984 interview—in which Raymond Cattell described working women and taxing the wealthy as “dysgenic,” and called for an end to immigration and for students to be taught how to marry intelligently, that is, eugenically—he also argued that eugenics should be based on the individual. By focusing on the individual, Cattell stressed, eugenicists could avoid becoming “sidetracked into all the emotional upsets that go on in discussions of racial differences.”⁹⁸

The move toward “Sovereign Individuals,” described in the introduction, enables a logic of escape from the nation, which the twentieth-century

eugenicists neither desired nor foresaw. In the world of the biometricians, members of national populations—assumed to be racially homogeneous—were inextricably intertwined: the fate of one person affected that of the others, hence the need to restrict others in order to help oneself. In the world of the “Sovereign Individual,” exit reigns supreme because, in the place of nationalism, there are “communities and allegiances . . . not territorially bounded. Identification . . . [is] precisely targeted to genuine affinities, shared beliefs, shared interests, and shared genes.”⁹⁹ The relationship between individuals and populations still matters, but the relevant group is now the “network neighborhood”—or the homophilic cluster, groupings that are based on kinship and specialized interests rather than on notions of equality.¹⁰⁰

The resurgence and revision of tribal rhetoric—what Jodi Byrd has called “tribal 2.0”—testifies to this shift.¹⁰¹ When Francis Galton and Karl Pearson used the term “tribe,” they used it in relation to “nation” (primitive tribes = primitive nations).¹⁰² Similarly, R. A. Fisher used “tribe”—more particularly, “barbarians”—to describe an ideal past state, in which sexual and natural selection coincided, and to which he thought future British society should aspire.¹⁰³ In the twenty-first-century, data scientists such as Cathy O’Neil underscore behavioral tribes within nation states or larger populations. O’Neil warns “with the relentless growth of e-scores, we’re batched and bucketed according to secret formulas, some of them fed by portfolios loaded with errors. We’re viewed not as individuals but as members of tribes, and we’re stuck with that designation.”¹⁰⁴ Lazarsfeld and Merton drew their terms “homophily” and “heterophily” from Karl Pearson’s work on associative mating and the ethnographic work of anthropologist Bronislaw Malinowski on the “savage Trobrianders.”¹⁰⁵

Although the resurgence of tribal rhetoric, which itself has deeply racial under- and overtones, has not completely undermined national interventions or identity, it has meant that national interventions happen through the functional equivalent of what Antonio Gramsci diagnosed as “hegemony,” albeit formed in reverse. But if “hegemony” once meant the creation of a majority by various minorities accepting a dominant worldview (such as the Greek city-states accepting Athenian values), majorities are now formed by bringing together angry or affectively charged minorities. As the examples of Cambridge Analytica and targeted

political ads reveal, the goal is to produce and maintain small charged clusters, in order to build majority support through “consolidation.”

Neighbors, however, are not neighborhoods, nor do invocations of tribes always have to erase “natives.” Instead, as Leanne Howe has argued, tribalogy produces transformative indigenous creation stories that “pull together all the elements of their tribe—meaning the people, land, and characters, and all their manifestations and revelations—and connect these in past, present, and future milieus.” Our task is to follow and amplify these stories—as Howe puts it, “where we go from here is only limited by our imagination.”¹⁰⁶