

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?

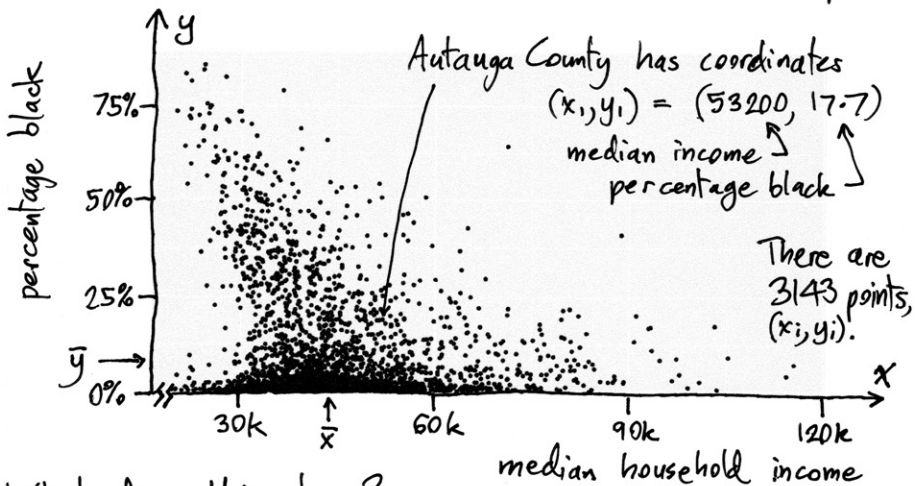
SPURRING CORRELATIONS

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

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

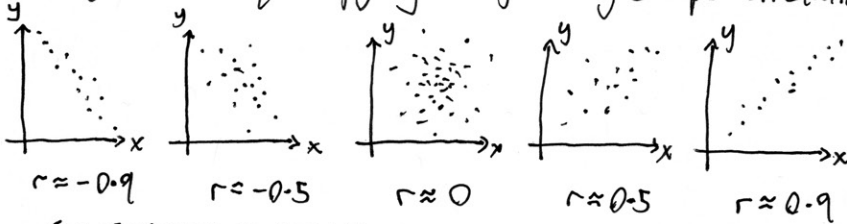
This (perhaps scary) formula involves two familiar quantities:

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 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.

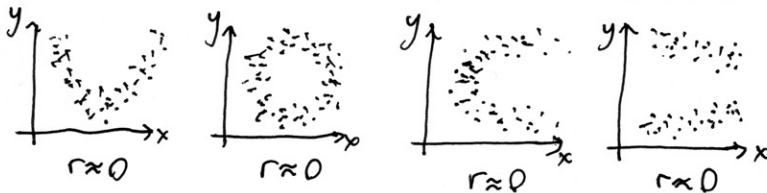
r is good at quantifying the following example correlations:



← STRONGER NEGATIVE

STRONGER POSITIVE →

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



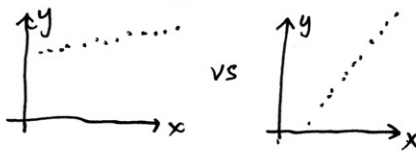
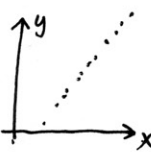
in each case there are correlations, but r cannot tell you this fact! It is insensitive to bimodality (the indicator of segregation earlier).

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 Cowsls 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

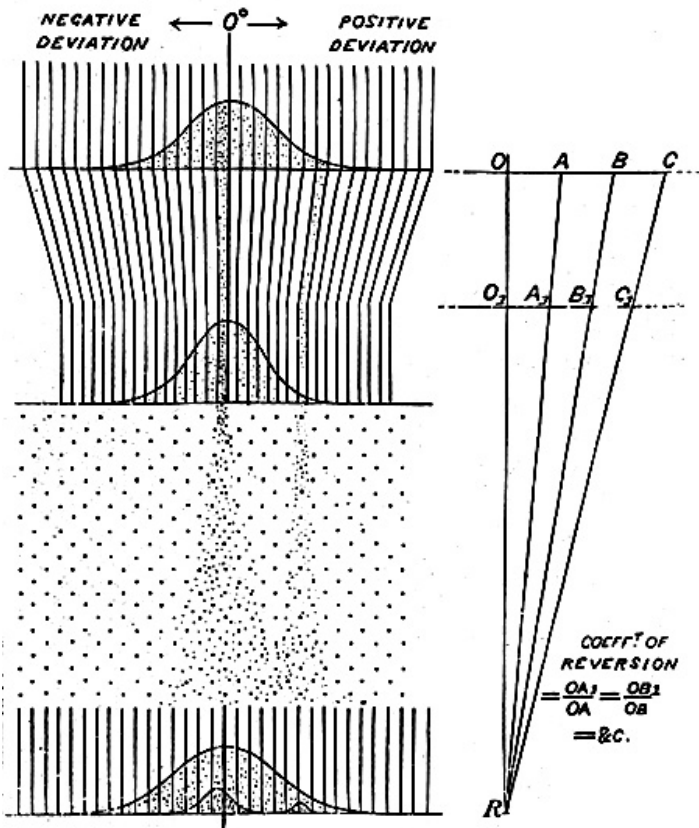
of financial discrimination explored in detail by Oscar Gandy, U.S. credit ratings correlate with race, or more precisely racism.⁵¹ As political scientist Ira Katznelson, policy researcher Richard Rothstein, and many others have shown, U.S. government policies such as the New Deal, Social Security, inexpensive 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.”

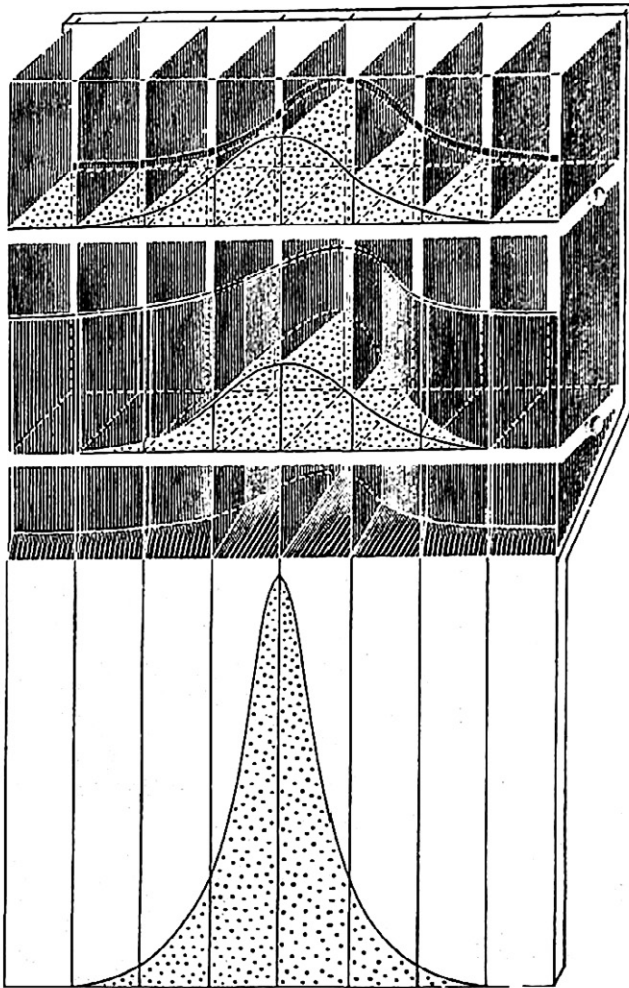
As Ruth Cowan and other historians of science have shown, Galton developed regression and correlation while studying heredity in humans and plants and the identification of criminals.⁵⁵ His fascination with the inheritance (or not) of genius (based on his undergraduate experiences at Cambridge University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



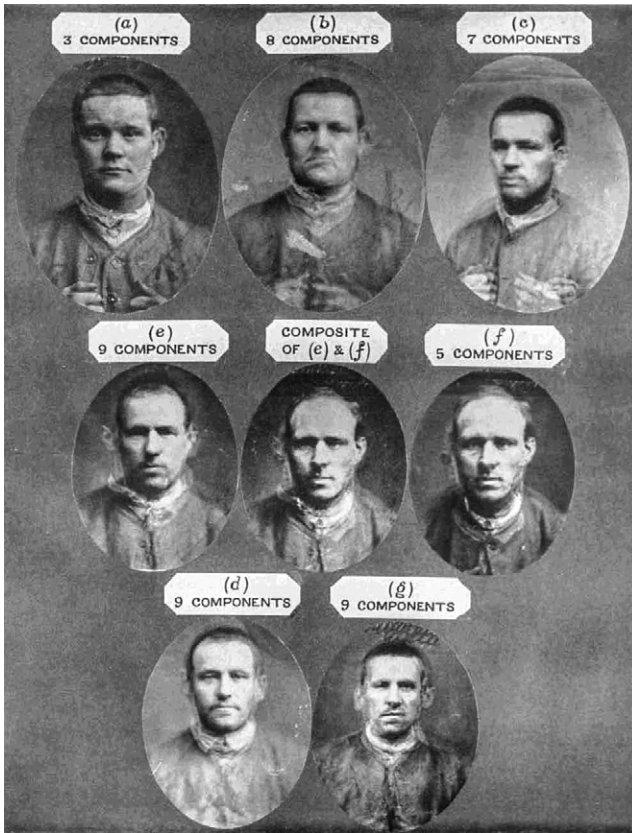
18 Galton's diagram of linear reversion. Karl Pearson, *The Life, Letters and Labours of Francis Galton*, vol. 3a, *Correlation, Personal Identification and Eugenics* (Cambridge: Cambridge University Press, 1930), 9.

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

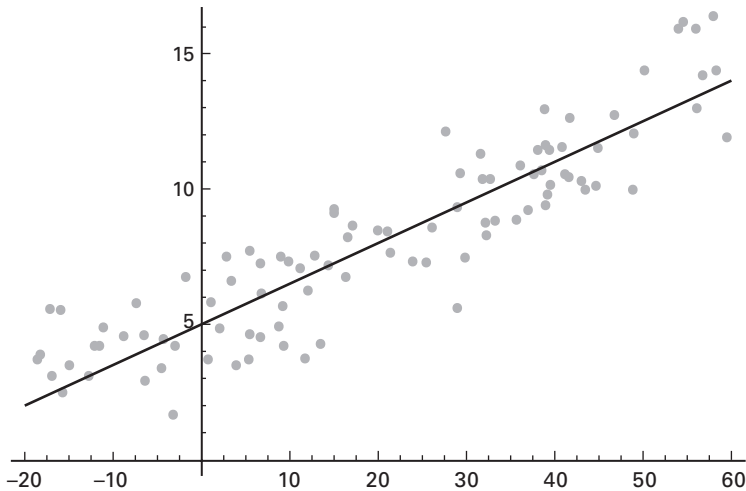


19 Galton's diagram explaining the influence of natural selection on reversion. Karl Pearson, *The Life, Letters, and Labours of Francis Galton*, 3a:10.



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. Eugenicians 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.