CHAPTER 3

The Statistics of the Deviant

In 1915, a young graduate from Cambridge boarded a ship to return home to India. With him was a complete run of the most dynamic, the most exciting periodical of the day, focused on the science for understanding the present and building better futures through data. The journal, *Biometrika*, embodied a new data-driven approach to biological and social problems, including questions of race and heredity—a great furthering of the dreams of Quetelet. The graduate, Prasanta C. Mahalanobis, saw in them technologies both colonial and a future independent India could adapt, to know itself better in statistical ways, to guide its economic and social development. As the great institutionalizer of statistics in India, Mahalanobis transformed these methods while tempering at times their hubris. The new sciences Mahalanobis brought with him emerged from the intersection of data, Darwinism, and a crisis of confidence in imperial Britain. These sciences produced mathematical statistics, born amid British power—and great fears of decadence and decline.

Moral panics can create new sciences. Late nineteenthcentury elite Britons were consumed with worries about their empire's degeneracy. "The time appears to have arrived," wrote Florence Nightingale in 1858, "when by the British race alone must the integrity of that Empire be upheld."[1] Declining birthrates among the elite, "limited population," alcoholism, and failures abroad all spoke to an empire in crisis. A new field, eugenics, and the statistics to support it provided one way not only to diagnose society, but to attempt to cure it—a way to make Britain great again.

A Crisis of the Modern Era

When the gentleman-scholar Francis Galton inspected his fellow men of Victorian Britain, he found them wanting: "We want abler commanders, statesmen, thinkers, inventors, and artists," he wrote in an article called "Hereditary Talent and Character." "The natural qualifications of our race are no greater than they used to be in semi-barbarous times," even though "the conditions amid which we are born are vastly more complex than of old." Modern civilization was all too much. "The foremost minds of the present day seem to stagger and halt under an intellectual load too heavy for their powers."[2] Genius was needed, but was in short supply. Education would never be enough, for there were simply not enough gifted men and women, not enough born geniuses, to confront the complexity of the times. England needed more geniuses, more people of extraordinary talent. They needed, Galton decided, to be bred.

The writings of Galton's illustrious if infamous cousin, Charles Darwin, offered a way forward.[3] In his *Origin of Species*, Darwin used the example of human breeding of domestic animals, such as show pigeons and pedigree dogs, to motivate his account of natural selection. Just as human breeders select features they desire, certain features of animals are selected for, as it were, over time in particular environmental niches. Humans, in Galton's estimation, underestimated their own power to affect their species. "The power of man over animal life," Galton explained, "is enormously great. It would seem as though the physical structure of future generations was almost as plastic as clay, under the control of the breeder's will." Not just physical traits, but equally the mind could be altered: "It is my desire to show more pointedly than—so far as I am aware—has been attempted before, that mental qualities are equally under control."[4]

Galton soon coined the term "eugenics" to describe the conscious effort to improve the quality of human beings—and national races of human beings in particular. Eugenics quickly became central to many left- and right-wing political programs across Europe, the United States, and the world. Racist to its core, to be sure, Galton's primary focus nevertheless was class. His suggestions often were whimsical, especially in comparison with the forced sterilizations and genocides associated with eugenic programs to come outside of Britain:

> Let us, then, give reins to our fancy, and imagine a Utopia—or a Laputa, if you will—in which a system of competitive examination for girls, as well as for youths, had been so developed as to embrace every important quality of mind and body, and where a considerable sum was yearly allotted to the endowment of such marriages as promised to yield children who would grow into eminent servants of the State.[5]

Unlike many philosophers and economists of his time, Galton was fundamentally anti-egalitarian.[6] "I object to pretensions of natural equality," he wrote. "I have no patience with the hypothesis occasionally expressed, . . . that babies are born pretty much alike, and that the sole agencies in creating differences between boy and boy, and man and man, are steady application and moral effort."[7] All people were not created equal, Galton insisted, and not all market agents had comparable mental capacities. Liberal political thought and liberal economics were just wrong in his estimation.

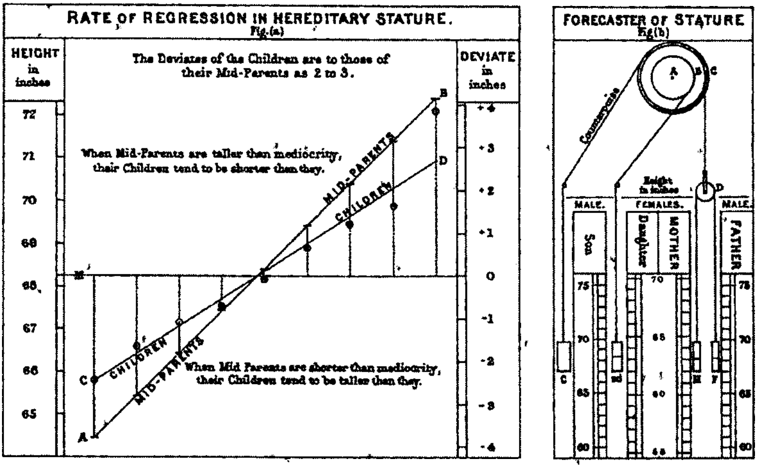
We associate eugenics and scientific racism with the far right, with Nazis. Things were otherwise around 1900. Many progressives as well as conservatives up to World War II saw science as capable of improving the human lot by improving the human race; indeed, one proponent noted, belief in eugenics offered "a perfect index of one's breadth of outlook and unselfish concern for the future of our race."[8] Statistical sciences were to replace the bigotries of old with evidence-based new sciences of human improvement: accounts of natural human hierarchies that moved easily from description to prescription.[9]

To improve the species, Galton needed to explore the wellsprings of talent and human excellence. Nurture was no explanation. Using biographical dictionaries of great men and women, Galton began investigating the density of talent and genius within families. In his long study *Hereditary Genius* of 1869, Galton studied prominent families and compared historical states with those around the earth. Despite the large number of his cases, his approach was intuitive and anecdotal. He generally argued that modern peoples were all lesser than ancient Greeks and that non-European peoples—he called them "races"—lesser than European ones.

While the approach in his book is largely anecdotal, Galton drew on Quetelet's normal curve to support his new ideas of ranking people and races. Quetelet used the normal curve to understand the qualities of a group as a whole. Galton used the same curve to understand variation within a group. Quetelet might seek the mean stature of Englishmen. Galton sought to understand the extremes of stature. His quarry was talent, not height, but he applied the same tools to one and other. The French sociologist Alain Desrosières explains, Galton used the normal curve as "a law of deviation allowing individuals to be classified, rather than as a law of errors."[10] What astronomers saw as errors to be eliminated, Galton saw as individuals to be ranked and classified. Every child getting test scores listing their percentile performance lives in the world Galton helped create.

And yet there was a major sticking point to all this investigation of excellence in distinguished families. Extremely tall people had tall children, but, on the average, those children were not as tall as their parents, reverting toward a population average height. Similar observations describe a wide range of human and animal traits. For breeders of animals—human or otherwise—this was a puzzle, one that would limit attempts to breed supposedly superior human beings. How to understand it? The answer would come from Galton's reworking of Quetelet's enthusiastic applications of the normal curve.

Why do offspring of tall parents tend not to be as tall as their parents, and more generally why do the attributes of a human group stay nearly constant over time? Galton came to explain both phenomena through what he called "regression," mathematically capturing the "tendency of that ideal mean filial type to depart from the parent type, 'reverting' towards what may be roughly and perhaps fairly be described as the average ancestral type."[11] With his statistical investigation, he discovered a powerful mathematical relationship between the amount of reversion of offspring and the extent of their parents' deviation from the mean. He not only showed the relationship to be linear, but also undertook what we would call today, thanks to Galton, the linear regression applied to



Francis Galton. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246–63. Plate IX.

the data, finding the coefficients of a simple linear equation like $y = mx+b$.

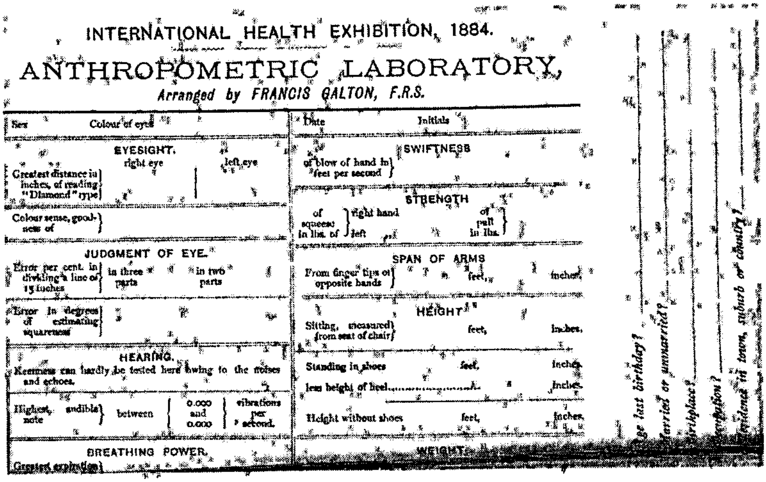
Galton was modeling facets of the process of generation, so his initial work with reversion involved only treating the parental heights as the x's and only the children's heights as the y's, for he was looking at a unidirectional biological process. But he soon realized that his process of regression could be detached from its biological mooring and used on a vast array of data. In investigating the process of "reversion," Galton had unknowingly hit on a much broader concept, namely that of statistical regression.

Correlation and Data

Galton did far more than introduce a powerful new approach to modeling data and making predictions from data. Quetelet studied society. Galton studied individuals in a distribution. He wanted better techniques to know and rank individuals and to know and rank races. In studying relations between pairs of attributes, such as the height of parent and child, Galton also introduced "co-relation," or as we would now term it: the correlation.

While governments were producing an ever-increasing number of statistics, they failed to accumulate enough of the data that most interested Galton—detailed investigations of the "chief physical characteristics" of a wide selection of the population, qualities such as "Keenness of Sight; Colour-Sense; Judgment of Eye; Hearing; Highest Audible Note; Breathing Power; Strength of Pull and Squeeze; Swiftness of Blow; Span of Arms; Height, standing and sitting; and Weight."[12] So challenging was collecting this data that Galton set up an Anthropometric Laboratory at the International Health Exhibition of 1884 in South Kensington. The laboratory measured 9,337 people in seventeen ways. He explained that "periodical measurements" would be useful to families in tracking their individual development, and to "discover the efficiency of the nation as a whole and in its several parts." Such records "enable us to compare, schools, occupations, residences, races, &c."[13] The data produced would continue to be studied well into the twentieth century. Galton's anthropometry, historian of psychology Kurt Danziger explains, "defin[ed] individual performances as an expression of innate *biological* factors, thereby sealing them off from any possibility of social influence."[14]

Galton's style enabled a dramatic new approach to understanding human differences. Following Quetelet, analysis of data could reveal the commonalities and range of quantifiable human behavior and attributes. And following Galton, each individual could be placed and ranked within those ranges: the top 5 percent, the bottom 10 percent. Inspired by


INTERNATIONAL HEALTH EXHIBITION, 1884.

ANTHROPOMETRIC LABORATORY,

Arranged by FRANCIS GALTON, F.R.S.

Sex · Colour of eyes · Date · Initials

EYESIGHT. right eye · left eye

Greatest distance in inches, of reading "Diamond" type

Colour sense, goodness of

JUDGMENT OF EYE.

Error per cent. in dividing a line of 15 inches · in three parts · in two parts

Error in degrees of estimating squareness

HEARING.

Keenness can hardly be tested here owing to the noises and echoes.

Highest audible note · between 0.000 and 0.000 vibrations per second.

BREATHING POWER.

Greatest expiration

SWIFTNESS

of blow of hand in feet per second

STRENGTH

of squeeze in lbs. of right hand · left · of pull in lbs.

SPAN OF ARMS

From finger tips of opposite hands · feet, · inches.

HEIGHT

Sitting, measured from seat of chair · feet, · inches.

Standing in shoes · feet, · inches.

less height of heel · inches.

Height without shoes · feet, · inches.

WEIGHT.

Age last birthday? · Married or unmarried? · Birthplace? · Occupation? · Residence in town, suburb or country?

Francis Galton, *Anthropometric Laboratory; Arranged by Francis Galton, FRS, for the Determination of Height, Weight, Span, Breathing Power, Strength of Pull and Squeeze, Quickness of Blow, Hearing, Seeing, Colour-Sense, and Other Personal Data* (London: William Clowes, 1884), 13.

Galton's work in observing large numbers of human beings, mental tests, for example, emerged from the effort to place each person amid the range of measured human capacities. And entire sciences of examining large numbers of "subjects" in statistical ways emerged in its wake. "A new method for justifying psychological knowledge claims had become feasible" with the work of Galton and his intellectual successor Karl Pearson, explains the historian Danziger. "To make interesting and useful statements about individuals it was not necessary to subject them to intensive experimental or clinical exploration. It was only necessary to compare their performance with that of others, to assign them a place in some aggregate of individual performances."[15] And it didn't take long for an approach to become big business. While pioneers like Galton struggled to get data at an adequate scale, a vast appetite for such inquiries would soon open, especially in the United States after the First World War.[16]

Above all, Galton revealed how surveying a mass of people makes recognizing—and targeting—the individual possible. Lots of data about lots of people allows scientists, marketers, militaries, spies to better know you—and target you. We live in such a world, where our individuality is quantified in reference to all other users of the internet, and where ad-serving algorithms exploit this quantification of difference to compete for our attention.

Institutionalizing Biometrics

The indefatigable Galton did not himself institutionalize his new statistical approach. He likewise did not have the mathematical skills to make it rigorous. Drawing on Galton's ideas and financial support, his intellectual heir, Karl Pearson, worked at both. A descendant of Quakers, freethinker, mathematician, socialist, feminist, and eugenicist, Pearson had a "grand vision, the creation of a statistical biology as the basis of effective eugenics and, concomitantly, the development of a mathematical statistics that could be applied to virtually all areas of human knowledge"—in the words of his biographer Theodore Porter.[17] Superior mathematical statistics would enable its expansion to the whole range of phenomena Quetelet dreamed of, an entire spectrum of social reform.[18] His field-building institutionalized eugenics and an imperious new statistical approach to social and political programs, with the help of patrons such as Galton and, of all things, the Worshipful Company of Drapers.

To do all this Pearson required data, labor to process that data, and new mathematics.[19] As he noted while giving a prestigious lecture, "the work is essentially the result of a co-operative investigation extending over a number of years, and depending upon a body of collaborators" who produced and analyzed "the extensive data on which my results entirely depend."[20] Pearson toiled for decades with a cadre of workers to bring his projects to fruition; a generation of great statisticians worked under and with him and changed how we all use data. Pearson ran multiple laboratories, including separate biometric and eugenic laboratories, with distinct projects, methods, staff, and funding.[21] With the help of two women assistants in particular, Alice Lee and Ethel Elderton, he amassed a wide range of data for a wide range of statistical investigations, and published results based on them, mostly in journals he founded and ran.

Getting data was hard work. In 1903, a plague pit was opened in London. Less than a week later, "one of my workers, Mr. S. M. Jacob, had with unwonted energy 'begged' the whole of the crania & skeletons" for Pearson's work.[22] Most data acquisition was more prosaic. Aiming to extend Galton's studies of the inheritance of physical and mental capacities, Pearson and his team placed requests in magazines read by headmasters and teachers asking them to record a multitude of observations on pairs of siblings and to rank them intellectually. They sent out 6,000 forms and got back some 4,000 from a wide range of schools (see illustration). "The absolute classification and tabling has been a work of great labour," Pearson explained, thanking a team of exceptional women: "Miss Alice Lee, D.Sc.; Miss Marie Lewenz, M.A., Miss E. Perrin, Miss Mary Beeton and Miss Margaret Notcutt" before noting that the "chief labour of computing has fallen upon Dr Alice Lee."[23]

Processing data was arduous and expensive, even with the help of new machines. Galton supported the Eugenics Lab, and in 1903, the Worshipful Company of Drapers granted Pearson £500 for his Biometric Laboratory, which allowed him to begin paying Alice Lee, who had previously

K. PEARSON 163

APPENDIX IB.

DATA PAPER FOR COLLATERAL HEREDITY INVESTIGATIONS.

B. SISTER-SISTER SERIES.
(Whole, not half sisters.)

No. in whole series.
(Not to be filled in.)

Please return this Paper to Professor KARL PEARSON, F.R.S., University College, London.

School:

Observer:

Date:

No. in School Series

Place a cross against the class of each sister under as many headings as possible, except under III and VIII. Please read first the General Directions.

	ELDER SISTER.	YOUNGER SISTER.
Name ...		
Age ...		
District of Home ...		

I. PHYSIQUE:

	Very Strong.	Strong.	Normally Healthy.	Rather Delicate.	Very Delicate.	Athletic.	Non-Athletic.
ELDER SISTER ...							
YOUNGER SISTER ...							

II. ABILITY: (a) *General Scale.*

	Quick Intelligent.	Intelligent.	Slow Intelligent.	Slow.	Slow Dull.	Very Dull.	Inaccurate-Erratic.
ELDER SISTER ...							
YOUNGER SISTER ...							

(b) HANDWRITING: (*See Back.*)

	Very Good.	Good.	Moderate.	Poor.	Bad.	Very Bad.
ELDER SISTER ...						
YOUNGER SISTER ...						

(c) WORK:

		Classics.	Modern Languages.	History.	Mathematics.	Descriptive Science.	Drawing.	Singing, Music.
ELDER SISTER ...	Good at ...							
	Best at ...							
	Likes best							
YOUNGER SISTER	Good at ...							
	Best at ...							
	Likes best							

(d) GAMES OR PASTIMES:

	ELDER SISTER.	YOUNGER SISTER.
Likes ...		
Good at ...		

III. HEAD MEASUREMENTS:

	Length.	Breadth.	Height.	a.	b.	c.
ELDER SISTER ...						
YOUNGER SISTER ...						

(a), (b), (c), Indices (not to be filled in).

IV. HAIR:

	Red.	Fair.	Brown.	Dark.	Jet Black.	Smooth.	Wavy.	Curly.
ELDER SISTER ...								
YOUNGER SISTER ...								

V. EYES:

	Light.	Medium.	Dark.
ELDER SISTER ...			
YOUNGER SISTER ...			

VI. RELATIVE CAPABILITIES: This is *only* to be filled in in those cases wherein the two sisters fall into the same class.

	Physique, stronger in	More Athletic.	Ability, greater in	Handwriting, better in	Hair, darker in	Eyes, darker in
ELDER SISTER ...						
YOUNGER SISTER ...						

VII. CHARACTER, ETC.:

	Noisy.	Quiet.	Self-conscious.	Unself-conscious.	Self-assertive.	Shy.	Conscientiousness. Keen.	Conscientiousness. Dull.	Popular.	Unpopular.	Temper. Quick.	Temper. Good-natured.	Temper. Sullen.
ELDER SISTER ...													
YOUNGER SISTER ...													

VIII. GENERAL REMARKS. Add here any striking features of resemblance or dissimilarity in the sisters.

ELDER SISTER ...

YOUNGER SISTER ...

[On the back of the Schedule spaces were arranged for samples of the handwriting.]

21—2

Data Paper for Heredity Investigations, in Karl Pearson, "On the Laws of Inheritance in Man: II. On the Inheritance of the Mental and Moral Characters in Man, and Its Comparison with the Inheritance of the Physical Characters." *Biometrika* 3, no. 2/3 (1904): 131–90, at p. 163.

undertaken extensive calculations for him on a volunteer basis, as well as collaborating with him. "Her duties included reducing data, computing correlation coefficients, creating bar charts . . . and calculating a new kind of statistic"—chi-squared—as well as supervising calculators male and female.[24] Calculation with machines became so central to the work at Pearson's laboratory that one visitor noted a "preoccupation with mastery of details of calculation" that could obscure the new mathematical statistics.[25] Much of this labor resulted in major collections of printed tables. It is hard to appreciate today just how essential such tables were as computational infrastructure for the growth of mathematical statistics.

While his women co-workers were often preoccupied with the tedium of calculating, Pearson also encouraged their higher-level work and often published with them. He argued, for example, "that Miss Elderton be no longer spoken of as a clerk, but be made a Francis Galton Scholar. She is quite capable of doing original work." Besides their contributions to statistics, they could become leaders in local social work. "It is most desirable that people trained in the Eugenics Laboratory should pass into work in public or municipal service of some type, as in dealing with mental defectives or invalid children, etc. We shall thus develop into a training school for practical eugenic work."[26] The most prominent of these women, F. N. (Florence Nightingale) David, named for the famous health reformer, went on to a storied career in statistics, including as professor in California.

Inheritance and Social Policy

What did all this labor with data prove? Intelligence was inherited, and Britain was losing the game of intelligence: "we are ceasing as a nation to breed intelligence as we did fifty to a hundred years ago. The mentally better stock in the nation is not reproducing itself at the same rate as it did of old; the less able, and the less energetic, are more fertile than the better stocks." This had major implications for social reform, as the problem wasn't schools but the breeding stock. "No scheme of wider or more thorough education will bring up in the scale of intelligence hereditary weakness to the level of hereditary strength." The only "remedy" is "to alter the relative fertility of the good and the bad stocks in the community."[27]

For Pearson, statistics was to be central for a new eugenic socialism necessary for a modernity both industrial and a conflict of races. If the goals were eugenic planning for a superior race, however, they were not trivial impositions of a racist and classist belief system onto data. Investigating skulls led Pearson and his collaborator Alice Lee to deny any reliable correlation between cranial size and intelligence, and to deny that skulls demonstrate the innate lower intelligence attributed to women.[28]

Eugenic statistics told tough truths: "we have failed to realize that the psychical characters, which are, in the modern struggle of nations, the backbone of a state, are not manufactured by home and school and college; they are bred in the bone; and for the last forty years the intellectual classes of the nation. . . . have ceased to give us in due proportion the men we want to carry on the ever-growing work of our empire, to battle in the fore-rank of the ever intensified struggle of nations."[29] The pressing political issues of the day required superior eugenical knowledge:

> The whole problem of immigration is fundamental for the rational teaching of national eugenics. What purpose would there be in endeavouring to legislate for a superior breed of men, if at any moment it

> could be swamped by the influx of immigrants of an inferior race, hastening to profit by the higher civilisation of an improved humanity? To the eugenist permission for indiscriminate immigration is and must be destructive of all true progress.[30]

Like Galton, Pearson argued that the "struggle of nations" was simply too important to rest on false eugenical science: that struggle required better science.

A Superior Science of Big Data

The social and biological sciences needed a remaking based in mathematics and in the production of data: "the loose qualitative or descriptive reasoning of the older biologists must give way to an accurate mathematico-statistical logic. The trained biologist may discover and tabulate facts, much as the physicist does today, but it will need the trained mathematician to reason upon them. The great biologist of the future will be like the great physicist of to-day, a mathematician trained and bred."[31] Many contemporary biologists disagreed, needless to say. Pearson extolled large-scale data collection and analysis rather than small-scale laboratory and experimental work.

What was true in biology was even more true in politics. Pearson noted with irritation the ease with which people opine on social questions: "every politician, every platform orator; who would hesitate to express even his opinion regarding a question in astronomical physics or cytology is ready with a decisive answer to each social problem that arises." But social problems were far harder than astronomical ones. "Social problems needed scientific answers. Every social problem belongs to a class embracing the hardest of all problems—it is vital not physical, it is biological, it is medical, it is statistical. It needs not less but far more investigation for its solution than any academic physical or biological problem."[32] Pearson's laboratories offered models for organizing political and social order along these new scientific lines.[33]

Correlation, Not Causation

Correlation, we are ever taught, doesn't equal causation. And to Galton's intellectual heir, that's why it was so exciting. Karl Pearson explained that he realized there was a "category broader than causation, namely correlation, of which causation was only the limit." Now more sciences could be made mathematical: "this new conception of correlation brought psychology, anthropology, medicine and sociology in large parts [*sic*] into the fields of mathematical treatment."[34] Correlations were particularly attractive in looking at sets of data with no clear causal relation. In studying evolution, correlation helped understand the processes of evolution without providing knowledge of its causes. Pearson believed fertility was strongly correlated with lower intelligence, lower morals. Correlation was essential, for example, to understand the reproductive policy a nation should follow if it was not to decline. Correlation, he proclaimed late in life, "has not only enormously widened the field to which quantitative and therefore mathematical methods can be applied, but it has at the same time modified our philosophy of science and even of life itself."[35]

Most of statistics in the twentieth century focused centrally upon causation, as we will see in the chapters to follow. But much of our current data revolution involves the reemergence of correlation as the most important tool in commerce, spycraft, and science. Whether in finding correlations or claiming expertise about the social world, a Pearsonian spirit pervades the data sciences.

New Data-Driven Racisms

From our point of view, all these figures appear largely backward racists and classists. And they were. Yet they were not hidebound traditionalists or conservatives. To the contrary: their science was central to part of their progressivism, to how they proposed to study social difference, and to foster a national unity that they believed to be undergirded by the best knowledge of their day. These new sciences would disrupt the conceptual foundations for the social order—even if they didn't ultimately change this order very much. Showing how radical technical *disruptions* often serve to reinforce existing inequalities will be a theme throughout this book.

Eugenicists saw few of their favored policies adopted as quickly as they would have liked, so some historians have dismissed the significance of the movement. Historian Robert Nye explains, "the long-term importance of a eugenics discourse in England was the way it transformed a narrow class outlook into a matrix of biomedical concepts claiming to represent the interests of the whole society, and which became an irresistible perspective for generations of educated Britons."[36] Eugenical ideas became default ideas for many in the educated classes; in Britain concerns with class predominated; in the United States race figured prominently. Eugenical ideas also shaped policy in Nazi Germany, with genocidal results.

Biometry, Race, and the Problems of Modern Society

"If modern civilisation is distinguished from all other civilisations by its scientific basis," Brajendranath Seal explained, "the problems that this civilisation presents must be solved by the methods of Science." In this opening address to the

1911 First Universal Races Congress, with W. E. B. Du Bois in attendance, Seal argued that the solutions to the pressing problems of race in the modern world required new sciences of humanity—not the old humanistic or philosophical methods of an Aristotle or a Machiavelli, but the new biometric sciences. "A scientific study of the constituent elements and the composition of races and peoples, of their origin and development, and of the forces that govern these, will alone point the way to a settlement of inter-racial claims and conflicts on a sound progressive basis," in the divided US, the restive British Empire, and the rest of the world.[37]

Embracing a eugenics program, Seal noted, the "study of genetic conditions and causes, of the biological, psychological, sociological forces at work, which have shaped and governed the rise, growth, and decadence of Races of Man, can alone enable us to guide and control the future evolution of Humanity by conscious selection in intelligent adaptation to the system and procedure of Nature."[38] And yet Seal distrusted the usual division of humans into races, and called for biometry to delineate properly the divisions of humanity, based on data. Seal was imbued with the approaches of Galton and Pearson: we must "adopt biometric methods in studying characters and variations," distrust averages, as "the range of variations in a character is as important an index as the character itself."[39]

A few years later Seal told Prasanta Mahalanobis, "You have to do work in India similar to that of Karl Pearson in England." In building institutions and pursuing biometric investigations, Mahalanobis did so. He brought the biometrical program to India, developed and challenged Pearson's methods, and founded mathematical statistics in India.[40]

Committed both to the acquisition of biometrical data and the development of ever more rigorous investigation of it, Mahalanobis came ultimately to turn highly problematic colonial data produced by the English into potent forms of nationalist knowledge as India secured its independence. In time, he made colonial data serve the new postcolonial Indian state.[41]

Following the aspirations of Seal, Mahalanobis sought techniques for discerning racial and caste mixture of various populations. Today he is best known for a measure of distance used in statistics that he first developed, "caste difference," as an alternative to Pearson's approach to the scientific study of racial difference. Unlike many of the racial theorists of their time, Seal and Mahalanobis stressed slow but real transformations over time. In his 1925 study of the Anglo-Bengals, Mahalanobis discerned dramatic but intelligible change. Caste had some transitory reality he claimed but "caste-synthesis" was well under way. "Intermixture within the province has gone on slowly and steadily even if imperceptibly and a larger Hindu Samaj has evolved which is not only not identical with the traditional society of Vedic or classic times but is in many respects even antagonistic."[42] The data analysis revealed the slow biological creation of a new Indian nation with real biological unity out of caste and sectarian division.

Aiding his approach to quantifying caste were powerful new tools for examining correlations among social groups at great scale. His new empirical techniques, he argued, both revealed this slow unification but equally the diversity of castes and tribes in India. In a massive data analysis, Mahalanobis and his collaborators undertook a data-driven clustering of castes and tribes in Uttar Pradesh.

Producing these analyses involved not just a team of human calculators, but also the use of "Mallock's Machine" housed in Cambridge in the United Kingdom.[43] Both in their empirical approach focused on calculating correlations at huge schools and in the use of new calculating devices, Mahalanobis and his team were doing data science long before data science. They found clear distinctions between

Brahmins, artisans, and tribal groups. And yet, for all the power of his techniques, Mahalanobis recognized the limits of the implications of these numerical differences. "To make further progress, it is necessary to take into consideration the social and cultural history of the tribes and castes, that is, the known ethnological evidence."[44] The failure to turn to such expert knowledge would plague—and indeed plagues to this day—too much data-driven science. No matter how powerful the algorithm or extensive the data, if one fails to embed this data analysis within broader forms of knowledge, scientific and humanistic alike, that so-called knowledge should be seen as incomplete at least, dangerous at worst.

Yearning for Causes—of Racial and Class Difference, for Example

Reviewing the 1911 First Universal Races Congress where the Indian intellectual Brajendranath Seal envisioned the coming together of nations, the American sociologist and delegate W. E. B. Du Bois drew out the most significant takeaways. "History illustrates these truths," he wrote on his notes, before quoting a distinguished speaker. "If we find an immense difference between the mind of some race" in Africa "and that of European race, we must seek the cause not in any difference of national qualities," but in external conditions. "It is not a difference of mentality in the race, but a difference of instruction, the same difference that we find to a greater or less extent, between the various classes of one and the same race or the different periods of its history."[45] Race and class differences must not be taken for granted, and differences in current intelligence ought not to be ascribed based on existing differences.