CHAPTER

1

Responsible Data Science

Data science is an interdisciplinary field that combines elements of statistics, computer science, and information technology to generate useful insights from the increasingly large datasets that are generated in the normal course of business. Data science helps organizations capture value from their data, reducing costs and increasing profits, and also enables completely new types of endeavors, such as powerful information search and self-driving cars. Sometimes, data science projects can go awry, when the predictions made by statistical and machine learning algorithms turn to be not just wrong, but biased and unfair in ways that cause harm. History has shown that the dual good and evil nature of statistical methods is not new, but rather a characteristic that was present from nearly the moment that they were conceived. However, by adjusting and supplementing statistical and machine learning methods and concepts, we can diagnose and reduce the harm that they may otherwise cause.

In popular and technical writing, these issues are often captured by the general term "ethical data science." We use that term here, but we also use the more general phrase "responsible data science." Ethics can refer in some usages to narrow "rules of the road" that pertain to a particular profession, such as real estate or accounting. Our goal here is broader than that: presenting a framework for the practice of data science that is ethical, but not in a narrow sense: it is *responsible*.

The Optum Disaster

In 2001, the healthcare company Optum launched Impact-Pro, a predictive modeling tool. Impact-Pro was an early success for predictive analytics (predating the term *data science*), and a decade later, Steven Wickstrom, an Optum VP, touted its use cases. For healthcare providers, it could "support steerage to appropriate programs" and "identify members [patients] with gaps in care, complications, and comorbidities." Optum termed these *care opportunities* in one document (i.e., opportunities for more revenue), but they are also of interest to those concerned with cost management: the correct early intervention in a health problem can cost significantly less than more drastic action later. For insurers, information on health risks for specific groups and individuals could be used to set premiums more accurately than is possible using traditional underwriting criteria.

DEFINITION DATA SCIENCE **We use the term *data science* broadly to cover the process of understanding and defining a problem, gathering and preparing data, using statistical methods to answer questions, fitting models and assessing them, and deploying models in an organizational setting. We consider artificial intelligence (AI) to be part of data science, and we also consider the "science" component of data science to be important.**

In 2019, though, a research team found that the tool was fundamentally flawed. For one important group—African Americans—the tool consistently underpredicted need for healthcare. The reason? The tool was essentially built to predict future spending on healthcare, and prior spending was a key predictor for that goal. And prior spending is a function not just of need, but also of ability to pay for and gain access to healthcare. Relative to other ethnic groups in the United States, African Americans have been (and continue to be) less insured, are less able to access healthcare, and possess fewer financial resources for covering healthcare expenses. In Optum's data, therefore, African Americans had less prior spending and, hence, less predicted future need. As a result, African Americans were less targeted for preventive intervention and necessary follow-up healthcare than were other people with similar health profiles. Neither the model nor the data provided to it were able to account for the unanticipated and overlooked societal inequities lurking beneath.

Optum was blindsided. The company thought it had built a tool that was a winner on all fronts: improving health outcomes by being smarter about required follow-up care, and managing costs better in the bargain. Instead, it found itself the focus of widespread bad publicity and was pilloried for creating a product that exacerbated racial bias and widened the healthcare gap faced by African Americans. New York state regulators opened an investigation, and the controversy continued into 2020. At the time of writing, Optum continues to market Impact Pro.

In this case, and in many others, the original intent for using the algorithm was good: good for healthcare providers by optimizing the allocation of scarce resources, and good for patients by ensuring that patients with the greatest needs had those needs met. But good intentions plus smart *artificial intelligence* (AI) led to disaster.

DEFINITION ARTIFICIAL INTELLIGENCE **We use the term *artificial intelligence* generally, to cover both statistical and machine learning methods for prediction with structured numeric data and text, as well as image and voice recognition and synthesis. In this book, we think of AI as having underlying algorithms or models. When discussing solutions for reducing the harms of AI, changing these underlying algorithms or models will be one of the main focal points.**

Interestingly, the scenario of good statistics being ill-used is not new. In fact, statistics as a field has a long history of being used for nefarious purposes or causing unintended harms.

Jekyll and Hyde

Let's begin with a look back over a century in history to a classic work of fiction that serves as a metaphor for the issues we face with data science today. In his gothic tale *The Strange Case of Dr. Jekyll and Mr. Hyde*, Robert Louis Stevenson describes two characters. Dr. Jekyll is an analytical man of science, a great asset to society, and a doer of good deeds. However, there is a repulsive, cruel side to him in the form of a separate character, Mr. Hyde, who gets "released" from time to time. The evil Mr. Hyde, in his times of release, tramples a young girl, commits murder, and more. The phrase "Jekyll and Hyde" has come to represent something that has two contradictory but inextricably linked natures—one respected and upright, the other base and evil.

The dual nature of humanity—good and evil combined in the same package—is a universal theme in literature. As humans carry their intelligence into the artificial realm, this duality has come with it.

Artificial intelligence has taken on this Jekyll and Hyde character trait. The enormous benefits brought by AI are evident: it has been a major force powering economic growth over the last several decades. Most aspects of life and industry now incorporate AI approaches in some way. Here are just a few examples:

- When you apply for a loan or a credit card, it is an algorithm that judges whether the application should be approved. This speeds the process, lowers the cost of providing credit, and, by making the process more scientific, standardizes decisions and expands access to credit among the truly creditworthy.

- When you use Facebook, Instagram, Twitter, or other social media services, the ads you see are optimized by an algorithm to be those most likely to get you to respond. This "microtargeting" makes them more relevant to you and, more importantly, makes it possible to provide these social media services at no charge to the user.
- Criminals are often caught on camera at or near the scene of a crime, and facial recognition and identification algorithms make it much more likely that they will be identified and caught.

In each of these cases we can point to a related "Mr. Hyde" that lurks in the background.

- Loan approval algorithms, it turns out, are prone to "redlining" just as humans are, blocking whole neighborhoods from credit, rather than making decisions on the basis of individual characteristics. Moreover, unlike humans, algorithms, if they are not transparent, are resistant to moral suasion and are hard to correct.
- The economic efficiencies wrought by microtargeting of ads is offset by the unease many people feel about being "surveilled." What's more, algorithmic curation of content feeds, seeking to maximize user engagement, drives users towards content that is provocative, inflammatory, and often fabricated. Even without actively provoking, these same recommender algorithms that underpin social media companies also enable political extremists to coalesce and take action.

Computer image recognition algorithms that have been so helpful to law enforcement facilitate dramatic erosions of privacy: one company has scraped the Web and built a database of billions of tagged face images, allowing individuals to upload images of people and find out who they are. When these facial recognition approaches are deployed by law enforcement, the harm resulting from erroneous identifications is magnified, *especially* for darker skinned individuals who are more likely to be falsely identified by these approaches. Sometimes, the negative Mr. Hyde aspect is only weakly counterbalanced by a good Mr. Jekyll. The science of image and voice synthesis has introduced the world to destructive "deep fakes": fabricated videos of people (usually political figures or celebrities) saying things they never said. Individuals or organizations bent on sowing discord or disinformation, or inciting violence have already used deep fakes for these aims. The plus side of the technology is comparatively minimal: better avatars for video games and production efficiencies for Hollywood, which needn't hire so many actors. The public has been highly exposed to these failures (possibly more than to the successes) through public controversies and popular science journalism and books. The good and evil sides to AI are now widely recognized, but this is not the first time that statistics has gone over to the "dark side." Indeed, some of the most foundational

breakthroughs in statistical methodology were motivated by goals we now recognize as morally reprehensible.

Eugenics

Turn back the clock to 1886, the very year *The Strange Case of Dr. Jekyll and Mr. Hyde* was published. This was also the year that the famous British statistician Francis Galton published his article "Regression Towards Mediocrity in Hereditary Stature," referring to the tendency of very tall and very short parents to have children closer to average height. This phenomenon gave us the phrase "regression to the mean."

Galton, Pearson, and Fisher

Galton, in addition to his seminal work on regression, also made contributions in correlation and survey methods. His half-cousin was Charles Darwin, and Galton was much taken with Darwin's *The Origin of Species*. Galton thought that, with the help of statistical methods, the evolution of humans could be guided in a positive and useful way. He coined the term *eugenics*, focused much of his research and scientific publications on eugenics, and became the Honorary President of the British Eugenics Society.

Karl Pearson, who contributed to statistics the correlation coefficient, principal components, the (increasingly maligned) p-value, and much more, was a protégée of Galton who assumed the Galton Chair of Eugenics at the University of London. Pearson saw the ideal society as:

> an organized whole, kept up to a high pitch of internal efficiency by insuring that its numbers are substantially recruited from the better stocks, and kept up to a high pitch of external efficiency by contest, chiefly by way of war with inferior races.

R.A. Fisher (design of experiments, discriminant analysis, F-distribution) joined official committees to promote eugenics and, in his "Genetical Theory of Natural Selection," focused on eugenics and what he saw as the need for the upper classes to boost their fertility. The guiding philosophy among the first generation of eugenicists was suppression of reproduction among the "unfit" and encouragement of reproduction among the "fit."

Ties between Eugenics and Statistics

The close ties between eugenics and statistics dissolved as statistics branched out in the service of all scientific disciplines, and eugenics itself was discredited through its close association with Nazi Germany. Now, all who study statistics are

familiar with regression, correlation, and the various "lettered" tests: the t-test, F-test, and the chi-square test. Few, however, know that the founding fathers of statistics (they were all men) were also the founding fathers of eugenics: the "science" of manipulating society and individuals to produce a superior race.

Many of the statistical methods that were developed in the service of eugenics are sound and have survived the test of time. The genetic theories and social policies that motivated the founding fathers of statistics are but a long-faded shadow in the eyes of modern statisticians, but they remain a jarring reminder that illumination and truth often come bundled with a measure of darkness.

Another popular application of statistics over a century ago was the supposed correlation of physical features with criminal tendencies; this was part of the pseudoscience of physiognomy. At the time, the presumption of such a connection between appearance and criminality was generally accepted. A quick read of some Sherlock Holmes detective stories, the first of which was written in 1892, gives a flavor for how "criminal types" tended to have certain facial features. For example, a sinister criminal in *The Man with the Twisted Lip* has "a shock of orange hair, a pale face disfigured by a horrible scar, which, by its contraction, has turned up the outer edge of his upper lip, a bulldog chin, and a pair of very penetrating dark eyes."

Unlike eugenics, this application of statistics is not dead. AI approaches have recently been used to infer autism, trustworthiness, and even the criminality of individuals from facial images. A recent Chinese study reported the use of AI to successfully distinguish between criminal faces and noncriminal faces. The authors, Xiaolin Wu and Xi Zhang, assembled two sets of photos.

- **Criminals**: One source was a city police department; the other was wanted posters.
- **Noncriminals**: A set of photos taken from the internet of males meeting certain criteria: no facial hair, no markings or scars, etc.

You can see a sample of the faces in their article "Automated Inference on Criminality Using Face Images," published in the Cornell arXiv prepress service (November 13, 2016, revised May 26, 2017) at `arxiv.org/abs/1611.04135`.

Wu and Zhang reported that four different classifiers that they built—logistic regression, k-nearest-neighbors (KNN), support vector machines (SVM), and convolutional neural networks (CNN)—all performed well in distinguishing the criminal images from the noncriminal images. Considerable controversy ensued, and the authors claimed to have been taken completely off guard by the storm of criticism they received. They felt compelled to issue rebuttals to their critics, which you can read, along with the original article, at the source noted above. Other Chinese researchers have gone about solving similar problems more subtly, publishing a number of research articles in recent years on subjects such as ethnicity detection for minority groups (mainly Uighur people),

facial detection for social credit applications, and even research on constructing simulated facial imagery from DNA samples.

Ethical Problems in Data Science Today

The problems with data science and AI today share one common theme with those of statistical eugenics: human bias. In 1900, evidence-based science was in its infancy. Galton, Pearson, and Fisher shared the common prejudice of the day that people's characteristics and capabilities were genetically determined, with race and sex being key factors. The statistical methods they developed helped them explore and quantify this prejudice. At no point did their statistical work cause them to question their beliefs.

Some of the ethical problems with AI today have, at their root, similar prejudices, often unspoken. Rarely are they expressed as a specific intentional feature of a model. More often, they come in via the data used to train a model, which the model then magnifies and perpetuates at scale. We will see later the example of a résumé-rating algorithm that was led astray by training data, in which men were ranked highly while women were not.

We will not attempt a scholarly or legal definition of ethics in data science; to do so with precision would entangle us unnecessarily in endless argument. (The European Union's General Data Protection Regulation [GDPR] is close to 90 pages.) However, we can say that for a data science approach to be responsible and ethical, one or both of the following ought to be addressed:

- **Bias**: An algorithm that makes predictions for people of a certain race (or religion, ethnic group, gender, belief, or other grouping characteristic) systematically differently than for others is considered biased.
- **Unfairness**: An algorithm that makes predictions in ways that deny due process, deprive people of property or liberty (even temporarily) without transparency or human review, or make decisions that appear intemperate or capricious or aid undemocratic governments in oppression is perceived as unfair.

Bias and unfairness overlap, of course. An algorithm that produces biased predictions would usually be considered unfair. On the other hand, an algorithm may produce biased predictions that, due to an innocuous modeling task or the bias working in favor of underprivileged groups, are generally deemed fair. Both bias and unfairness are subjective, though bias has clear-cut legal implications that we will discuss in Chapter 3, "The Ways AI Goes Wrong, and the Legal Implications." Our goal is not to engage in philosophical debate about bias and unfairness and what constitutes either. Rather, we take the view that, whatever your exact definition of bias and unfairness, these issues require much

more attention in data science projects than they tend to receive. Our goal is to provide the guidance and tools to facilitate this.

You will note that "making biased or unfair predictions" lies at the center of this description of responsible data science. Let's now look at how algorithms make predictions.

Predictive Models

We've been speaking generally of data science and AI; let's be more concrete.

Before there was AI, there were predictive models—statistical models that predict an outcome (customer spending, whether an insurance claim is fraudulent, whether a loan will be repaid, etc.). The earliest predictive models have their roots in linear regression (we talked earlier of Galton's contributions) and discriminant analysis (likewise, for Fisher).

Most analysts are familiar with the standard linear regression model:

$$Y = a + b_1x_1 + b_2x_2 \ldots + b_nx_n + \text{error}$$

In this model, the effect of predictor variables (x_1, x_2, x_n) on the outcome (Y) is reflected in their coefficients (b_1, b_2, . . .b_n). For example, suppose Y is an individual's spending with a company, and x_1 is their income. The effect that an individual's income, x_1, has on their spending, Y, can be readily seen in the coefficient b_1. This might be useful if you are doing a research study about consumer spending and want to be able to distinguish the different effects of different predictor variables.

From Explaining to Predicting

Historically, the main purpose of regression was as suggested earlier—to shed light on a relationship between predictor variables and an outcome variable. In this case, the emphasis is not on predicting individual cases but rather on understanding the overall relationship. The focus is on the predictor coefficients (the *b*'s in the earlier equation). For many decades, predicting individual outcome values was a secondary goal, though not unheard of (e.g., using regression to forecast a time series such as product demand). However, this focus on using predictive models to examine underlying relationships in the data has since shifted to using predictive models solely to leverage the predictions that they generate.

NOTE SCARCE DATA **The classical statistical methods of regression and discriminant analysis were developed in the era of scarce data. Typically, data would be gathered in a study or experiment and then analyzed. The cost of gathering data was often high, and there might be hard limits on how much data could actually be obtained. Therefore, all the available data would be used in fitting a model, and**

statistical analysis methods had to squeeze all possible information out of a sample. In traditional statistical methods, information about the variability in the data is used to calculate metrics (e.g., R^2, standard errors, t-statistics, F-statistics) that help us understand how well the model fits the data. But getting a model that better fits the data at hand does not necessarily help you extrapolate beyond those data.

Suppose you catch a bunch of fish in a net and then proceed to make a detailed description of what you've caught. Now, you want to derive from those details a picture of what all fish are like. If the fish in the net were gathered according to a rigorous sampling plan, this will allow inference to the larger population of fish. However, going into too much detail about the fish in the net won't necessarily translate into an accurate picture of the fish beyond the net. In fact, it can distort the picture if you end up modeling random noise rather than meaningful signal.

In the era of plentiful data, however, there is a simpler solution: hold out some of the data where the outcome is known (i.e., refrain from using it to fit the model), and see how well your model predicts those outcomes.

Predictive Modeling

Companies and other organizations now have a wealth of transactional and other data produced in the routine course of business, and many opportunities to capture value from the data. The earliest widespread use of predictive modeling was credit scoring, which is essentially a general-purpose prediction about whether a person will repay a loan.

Credit scores do not require predictive models. The credit score company Equifax started in 1899 as Retail Credit Company. Founded by the Woolford brothers, it collected data on individuals in a community, information that could be used to determine whether to offer a person credit. In the age of computerized statistical modeling, automated predictive models allowed credit scoring to scale. In the predictive modeling process, a model (e.g., linear regression) is fit to (or "trained on") data where the outcomes, the Ys, are known. The model is then applied (or "scored") to additional "holdout" data where the outcome is known and is used to predict those outcomes. Note that the model is fit only to the training data; when it is scored to the holdout data, it is not refit to those data. The performance of the model is then assessed on the holdout data where we know the actual outcomes: how well did it do?

The analyst can choose among many different model types and can also choose different ways of implementing those models. With linear regression, for example, the analyst chooses which predictor variables to include and whether to include derived variables that account for interaction among the predictor variables. Different models and different implementations of the same model can all be trained and assessed on the holdout data, and the best-performing model can be the final model for deployment (see Figure 1.1).

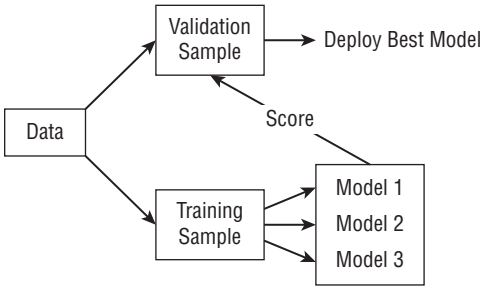

Figure 1.1: The use of data partitions to assess different models

Setting the Stage for Ethical Issues to Arise

When the goal of statistical analysis was to understand relationships among variables, the resulting model and its interpretation could be reviewed and vetted at comparative leisure. One can imagine analysts in a government office poring over regression printouts, or professors in a seminar picking apart a colleague's proposed model. The conclusions of the model might guide public policy or further research but were not to be implemented as decisions at an individual level.

When the statistical model moved into the realm of prediction, as opposed to explanation, the implications of a model became immediate and consequential for individuals. If you were issued a low credit score, it affected your life materially. If you were not told why you received the low score, it made matters worse by providing little opportunity for recourse.

While public concern surfaced about the transparency of credit scores and lack of recourse, the development and tuning of statistical algorithms to make these decisions was not considered much more than the employment of a tool. Back in the 1980s, when these methods saw wide adoption in credit scoring, little thought was given to developing a "responsible approach" to predictive models.

The first step in developing a responsible data science framework is to review the different types of statistical and machine learning models. Various types of models are handled differently, in particular transparent models versus black-box models.

Classic Statistical Models

Linear regression is a transparent model. As shown earlier, the effect of predictor variables on the outcome can be read from the coefficients. Linear regression is only one of a number of popular statistical and machine learning models. Let's briefly review the others, starting with the older classical statistical models.

Logistic Regression. While linear regression tries to predict a numeric value, logistic regression tries to estimate the probability that an observation belongs to a specific class or category. Most logistic regressions deal with binary outcomes, usually coded as 0 or 1. We can use a linear model to do this but not directly. A probability must lie between 0 and 1, and a standard linear regression does not assure this. With a couple of steps, though, we can get there. (Note, we're not showing all the math here.)

1. Instead of probabilities, we deal with odds, which are familiar to bettors. If the probability of an event is 25%, a bettor would say the odds are 1 *to* 3, or 1 *in* 4. Mathematically, we use 1 to 3, so a 25% probability equates to 33%:

$$\text{odds} : \text{odds} = p / (1 - p).$$

2. It turns out we can use odds in a linear model, if we take the natural log of the odds, or the logit:

$$\log(\text{odds}) = a + b_1 x_1 + b_2 x_2 \ldots + b_n x_n + \text{error}$$

3. The predictor coefficient estimates—the *b*'s —can get us back to a probability via exponentiation:

$$\text{probability of outcome} = 1 / \left(1 + e^{-(a + b_1 x_1 + b_2 x_2 \ldots)}\right)$$

Logistic regression remains a workhorse of predictive modeling. The coefficients it produces can illuminate the effects of individual predictors on the odds of the outcome, which makes generating explanations for humans relatively simple.

Discriminant Analysis. Discriminant analysis is the oldest classification technique, dating back almost a century to the work of R.A. Fisher. It was used in early credit-scoring algorithms, and, though it has been eclipsed somewhat in the data science field, it is very computationally efficient, so it can be used where rapid model-fitting is important.

Discriminant analysis calculates functions that, when applied to the predictor values, do the best job of separating the records into their different classes (i.e., classes that are very homogeneous and differ the most from one another). These functions have coefficients that reflect the contribution of individual predictors in separating the classes. They also yield a *classification score* for each record for each class. The classification scores for a record reflect how close a record is to the average predictor values for a class—the higher the score, the closer it is. The scores are based on "statistical distance," a standardized version of the squared difference between a record's predictor values and the predictor value average for the class.

Naive Bayes. Naive Bayes works only with categorical predictors. Returning to the concept of closeness, naive Bayes starts with the idea of finding records whose predictor categories match those of the record to be classified. Whatever class the majority of those records have is the predicted class for the new record. When there are more than a few variables, however, finding exact matches on predictor values is next to impossible. So, instead of this "exact Bayes" algorithm, we work with a "naive" version that considers overall probabilities of predictor values in the outcome classes. Roughly, the algorithm starts as follows:

1. Given a record to be classified for each outcome class, find the probability that each predictor value in the record to be classified occurs in the class. For example, among the outcome class "loan fails to pay," the category of "prior personal bankruptcy yes" likely has higher probability than among the outcome class "loan is current."
2. Multiply those predictor category probabilities together and by the proportion of records belonging to that class.
3. Repeat for the other outcome classes.

Each record gets assigned to whichever class has the highest such product of probabilities.

Naive Bayes is computationally efficient and, alone among predictive algorithms, has the ability to handle categorical variables directly. This is an advantage in cases where data have numerous categories that otherwise would require the creation of many dummy variables.

Black-Box Methods

***K*-Nearest Neighbors (KNNs).** KNN is intuitively simple. The KNN algorithm predicts outcomes for a given record by a two-step process:

1. Find records that are closest to (most like) the record to be predicted.
2. Find the average outcome among those nearby records, or find the majority class—*this is the prediction*.

"Nearness" between two records is typically measured using Euclidean distance (square root of the sum of squared differences) between the two vectors of predictor values, though other measures of distance should be used when the predictor values include categorical as well as numeric data.

For example, if you have two houses with the predictor attributes as listed in the table for Rooms, Acres, and Square Feet, with Value as an outcome, the Euclidean distance between House 1 and House 2 is the square root of 4.56 (considering only the predictor variables).

	ROOMS	ACRES	SQ. FEET (000)	VALUE ($000)
House 1	8	1.5	3.2	450
House 2	10	1	3.8	670
Difference	–2	0.5	0.6	
Difference^2	4	0.25	0.36	Sum = 4.56

Now consider House 3, with 12 rooms, 2.5 acres, and 3,700 square feet. If you do the calculations, you will see that House 1 is closer to House 2 than to House 3. One important note: You can readily see that the Rooms variable dominates the calculation because its scale is greater. This is typically corrected for by converting each variable to a standardized counterpart (e.g., how many standard deviations it is away from the mean for that variable).

How many neighbors should you look at when calculating their mean (or predominant class)? The user sets this number, *k*. The larger *k* is, the less variable the predictions are, and the more they will come to resemble the overall average. A smaller *k* results in predictions that are more responsive to local conditions in the data, but are also more prone to fitting noise.

Decision Tree Ensembles. Tree methods produce a set of rules that are applied to predictor variables to yield predictions. For example, a rule for predicting whether a bank customer will apply for a new loan might look like "IF income > 110.5 and Education level ≤ 1.5 AND Family ≤ 2.5, THEN predict as nonapplier." These rules do not come from humans applying commonsense business judgment, but are derived by an algorithm that determines which rules yield the best predictions (i.e., do the best job of dividing customers into those who applied for the loan and those who did not). Tree algorithms generally proceed as follows:

1. Pick a predictor variable value (a "decision node"), say income = $61,000, and then divide the data into records where Income > $61,000 and records where Income ≤ $61,000. Measure how homogeneous the two partitions are in terms of loan application. Are those above the split mostly "appliers" and those below mostly "nonappliers"?
2. Repeat for all other values of the predictor, and record the split that yields the greatest homogeneity (also called *purity*) in the two partitions. Note that if Income ends up being the best discriminator, which it does, this yields an initial rule of "IF Income ≤ 110.5, classify as 1; otherwise, classify as 0."
3. Repeat for all the other predictors, noting which variable and split yields the greatest homogeneity, and implement that split.
4. In each of the two resulting partitions, repeat the previous process.
5. Keep going until you have fully pure final partitions (called *leaves* or *terminal nodes*).

Figure 1.2 shows the beginning of this process. The algorithm decided on Income = 110.5 as the optimal initial split (out of all possible splits of all predictor variables), and Education = 1.5 as the next split.

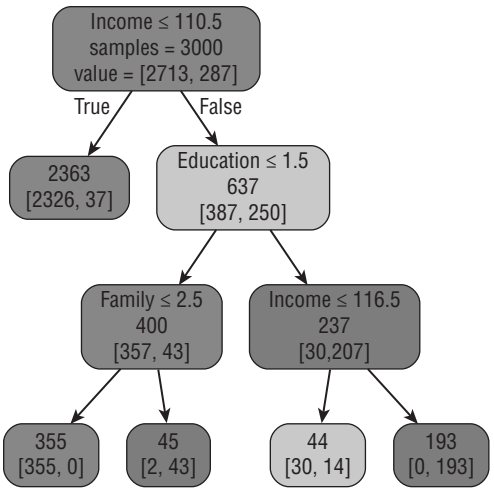

Figure 1.2: Initial stages of the splitting process for the bank customer data

Data Mining for Business Analytics by Shmueli, Bruce, & Patel, © 2020 John Wiley & Sons, Inc. Used by permission.

You can see that this process, carried to the end, yields increasingly meaningless rules for small segments of the data—rules that are fitting the detailed noise in the data, not the signal. One type of remedy is to stop the tree growth, or "prune" the tree back to a point where its rules generalize effectively to the holdout data. A popular twist to this is to use cross-validation, or sequential segregation of different parts of the data as holdout sets, so as not to be biased by the particularities of a single holdout set.

Another remedy is to forget about fine-tuning tree growth and simply to average the predictions of lots of different trees in an "ensemble" approach, taking advantage of the "wisdom of the crowd" (see note). This approach generally performs extremely well, but it loses the advantage of producing simple rules that can be presented effectively to non–data scientists. Ensembled trees have among the best performance of all predictive algorithms and are more popular than their simple tree counterparts.

Although simple trees produce easily interpretable rules, we classify decision trees here as a black-box method, due to the dominance of the ensembled version.

NOTE WISDOM OF THE CROWD AND THE OX **In his book *The Wisdom of Crowds*, James Surowiecki recounts how Francis Galton attended an event at a country fair in England where the object was to guess the weight of an ox.**

Individual contestants were relatively well informed on the subject (the audience was farmers), but their estimates were still quite variable. Nonetheless, the mean of all the estimates was surprisingly accurate—within 1% of the true weight of the ox (over half a ton). On balance, the errors from multiple guesses tended to cancel one another out. The crowd's average was more accurate than nearly all the individual estimates; hence, the name of Surowiecki's book. Diversity of estimates was a strength, not a drawback.

Artificial Neural Networks (ANNs). ANNs initially were an attempt to mimic neural processes in the brain. A useful introduction to neural networks is to reconsider linear regression, in which the coefficients are calculated via a formula that minimizes the squared differences between the predicted values and the actual values. Now imagine an algorithm that randomly chooses the coefficients, uses them to make a prediction, finds the error in the prediction, changes the coefficients to improve the prediction, finds the error again, and keeps repeating that process. Neural networks are an extension of that process.

Consider a simple neural net to predict whether a consumer will like or dislike a new food product. The manufacturer has made a number of samples, all with different combinations of salt and fat (the two ingredients deemed key to consumer acceptance), and has had them rated by consumer panels.

Neural networks are configured in different ways. This one starts with two small random numbers, one for each input (the input layer in the diagram), that are then each multiplied by, say, three weights—the w's in Figure 1.3. The three products for input 1 (Salt) are added to the three products for input 2 (Fat), which yields three "nodes" (neurons) in a "hidden layer" (the choice of three weights being somewhat arbitrary). The process is that the outputs of the three nodes within the hidden layer are then multiplied by three additional weights and summed in the output node, which is compared to a cutoff threshold to assign a classification. (Some additional steps have been omitted for brevity's sake; see Chapter 8, "Auditing for Neural Networks," for more details.)

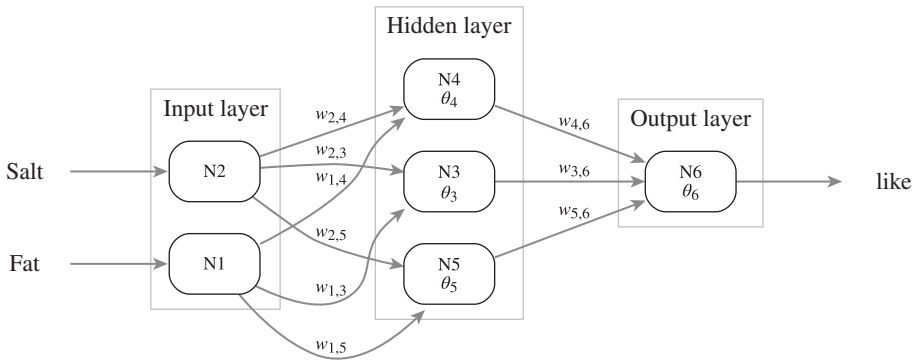

Figure 1.3: Neural network layout: a simple artificial neural network

To this point, predictions from this process are largely random and not worth much. However, the network then reverses itself (*back propagation*): the "error" in the final layer becomes an input and flows through the network in reverse, but with a slight adjustment in the weights. The cycle repeats, with the weights repeatedly adjusted in a way that causes the overall prediction error to continuously diminish.

Deep Neural Nets (Deep Learning). Neural nets have been through several peaks and valleys of popularity since they were first introduced in 1967. However, their popularity has skyrocketed ever since the advent of the "deep learning revolution," which began in 2012 when deep learning and deep neural nets (DNNs) shattered state-of-the-art performance records in image prediction and biomolecular target prediction. What makes a DNN distinct from an ANN? Depth, obviously, is one of the main characteristics. Unlike an ANN, a DNN will have multiple hidden layers (see Figure 1.4).

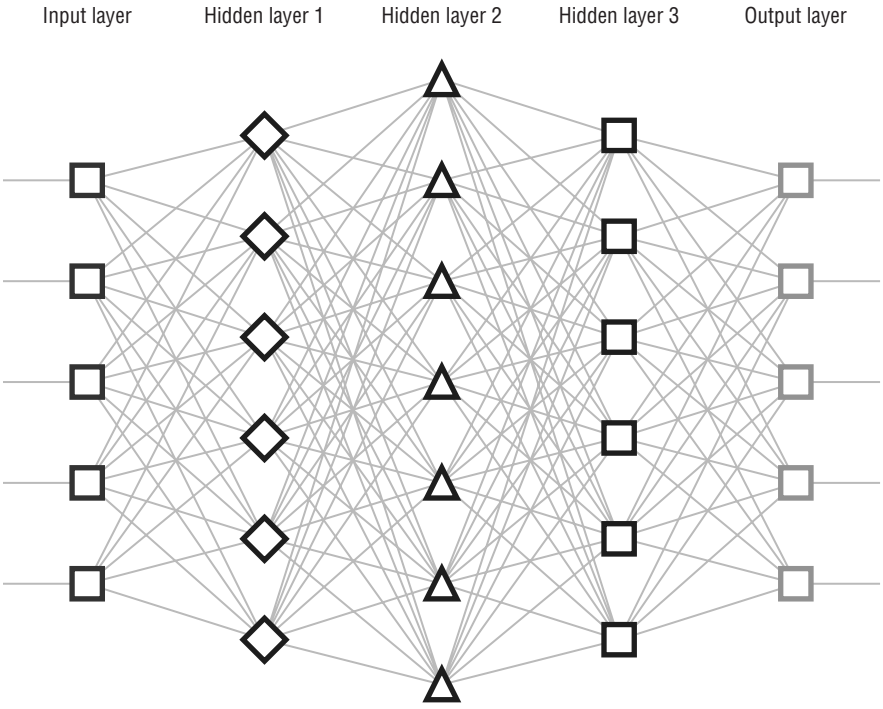

Figure 1.4: Diagram of a DNN

Much of the progress and excitement in AI since 2012 has been fueled by further developments in these DNNs. In fact, DNNs have almost completely supplanted ANNs in practice, with most data scientists using the term *neural networks* to describe DNNs rather than ANNs. We will follow this convention in this book and will confine discussion of ANNs to the technical reviews in

the current chapter and the one following. DNNs differentiate themselves from their shallow counterparts in that they contain more than one hidden layer. Each of these hidden layers creates an intermediate representation of the input data, with later layers creating progressively more abstract representations prior to making a prediction. This ability allows DNNs to learn progressively more abstract features from raw data, making them especially useful for modeling tasks that use unstructured data such as image or text data.

Consider the case of deep learning using a convolutional neural network (CNN), a popular variant of DNNs used for image recognition tasks. The many individual pixel values input into a CNN are repeatedly consolidated into larger conceptual groups in a lengthy and complex iterative process. The consolidations that are effective in correctly identifying labeled images (say, "car" as opposed to "person") are saved and used to help classify new images. Those consolidations, it turns out, are highly similar to the higher-level features (edges, shapes, textures, etc.) that humans see in images.

A CNN that learns the features in an image, however, lacks the deep breadth of human knowledge. While a well-trained CNN may achieve remarkable results within specific tasks, it can also easily go haywire when predicting on data that differ even slightly from the data it was trained on. In one case, a CNN was trained to distinguish between malignant and benign skin moles. The CNN did a very effective job with the labeled training data, but what it really did was distinguish between images that had a ruler next to the mole and images that did not. Dermatologists were more likely to place a ruler next to a mole they considered malignant, and the presence or absence of a ruler turned out to be the best predictor of whether a mole was malignant or benign.

In this case, the DNN's mistake was obvious from a quick review of the images, and it is only our surmise, albeit a strong one, that it was a mistake. The presence or absence of a ruler was not specified as a variable; it was something the net learned on its own. There are no readily apparent coefficients and variables that allow us to quickly learn what the net was basing its decision on. We discuss DNNs and how to audit them for these sorts of issues in Chapter 8, "Auditing for Neural Networks."

Important Concepts in Predictive Modeling

Having reviewed the landscape of statistical and machine learning methods used in prediction, let's now review some additional concepts that are key to the predictive modeling paradigm.

Feature Selection

The key to success in prediction is identifying variables (features) that successfully predict an outcome of interest. Both traditional statistical models

and machine learning models are driven by the proper selection of predictor variables. Data, such as those drawn from company records of customers and transactions, are most often recorded in a tabular form, in which rows are cases (customers, transactions, patients) and columns are variables. Early on in the evolution of data mining, it was the analyst's job to examine these data and select useful predictor variables. For example, in predicting whether a loan will be repaid, useful variables are things like prior loan repayments, income, job status, age, etc.

As data mining evolved and began to be called data science or data analytics, methods and models also evolved, in which features began to play a much less transparent role in the modeling process. The increasingly widespread use of image and text data necessitated this change, as the high dimensionality of these data made creating handcrafted features practically infeasible. This is most evident in image recognition, in which the "features" are nothing but pixel values (integers) arrayed in a matrix. These must be reinterpreted as higher-level features such as edges and shapes, and even higher-level features such as eyes. Requiring a human to label the relevant features within every image or sentence in a text corpus defeats the whole purpose of machine learning. This task is now done with remarkable accuracy by DNNs.

Model-Centric vs. Data-Centric Models

Earlier, we grouped predictive modeling methods into classical statistical models and black-box machine learning models. A related classification scheme categorizes methods as model-centric or data-centric. Linear and logistic regression and discriminant analysis are all model-centric in the sense that they calculate a formula that is applied globally across all the data to make predictions. Of course, the models are derived from data, but we consider them model-centric due to the global formula. KNN, naive Bayes, classification and regression trees, and neural nets make predictions based on more local characteristics of the data. They are also referred to as *models*, but we put them in the data-centric category because the predictions are more responsive to characteristics of the data that are not global. Data-centric algorithms are more complex and flexible and are more prone to pay attention to random noise in the data.

Holdout Sample and Cross-Validation

To test how well a model predicts, we typically evaluate its performance with a holdout sample, which is a randomly selected sample of data drawn from the same source as the data used to fit the model. The holdout sample, also termed a *validation* or *test sample*, is not used in fitting the model. The data used to fit the model are termed the *training data*.

In some modeling processes, the holdout sample can be used sparingly to get a good idea how the model will perform with new data. In other processes, the holdout sample can be used iteratively to tune the model and to improve its performance. In such a case, it is common to use a modified holdout sample in a process called *cross-validation*.

In cross-validation, the holdout process is repeated for nonoverlapping partitions of the data until all the data have been used in validation. *K*-fold cross-validation involves taking multiple different divisions of the data into training and validation partitions such that *k* different nonoverlapping validation partitions are created and used. The training partitions are overlapping.

Overfitting

The use of a separate sample to evaluate model performance has an important benefit: it helps avoid overfitting, also known as "fitting the noise" instead of the signal. Consider the data shown in Figure 1.5 on advertising expenditure and sales revenue for different periods.

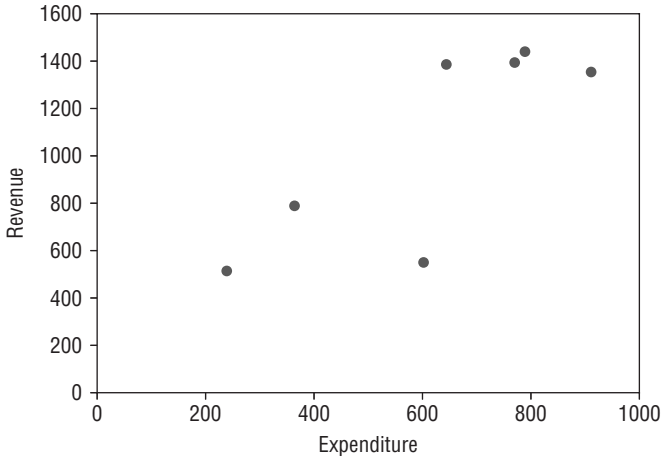

Figure 1.5: Advertising expenditure and revenue for different periods

Data Mining for Business Analytics by Shmueli, Bruce, & Patel, © 2020 John Wiley & Sons, Inc. Used by permission.

As advertising spending rises, so does sales revenue, in general, though the relationship does not closely follow a straightforward line or curve. We could fit a mathematically complex curve to the data, as shown in Figure 1.6.

This model fits the data perfectly—there is no error. We would have little confidence, however, that it would be a good fit to future data. We have fit the noise, not the signal.

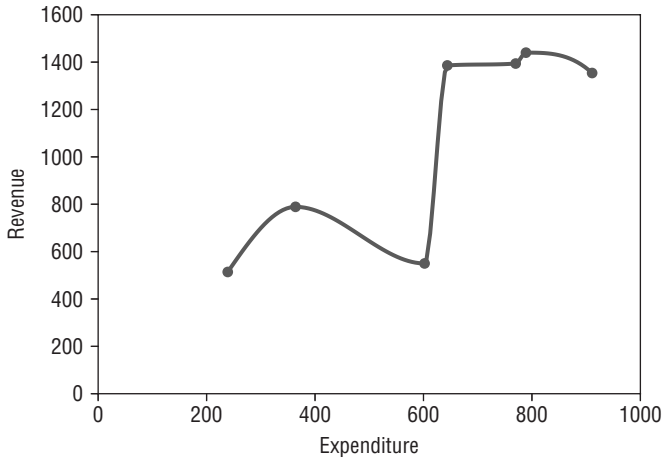

Figure 1.6: Smoothed curve closely fitting to the advertising and expenditure data

Data Mining for Business Analytics by Shmueli, Bruce, & Patel, © 2020 John Wiley & Sons, Inc. Used by permission.

OVERFITTING AND THE BIAS—VARIANCE TRADE-OFF

Overfitting is a constant theme in predictive modeling. Computer science and machine learning brought increasingly complex and data-centric models to the table, and these models are particularly prone to fitting the noise rather than the signal. The model-building process itself must incorporate ways to limit the complexity of models and curb the tendency to fit noise, without so severely curbing the model as to diminish useful predictive capability. The use of holdout data in evaluating model performance is often key to this balancing act—if a model is allowed to grow too complex, this will show up in decreased predictive accuracy with the holdout data.

In more technical terms, this is called the *bias-variance trade-off*. Models that underfit do a poor job of predicting, meaning the way they capture significant relationships between predictor variables and outcomes is *biased*. Models that are overfit will "jump around" in their estimates from one sample to another, due to noisy data; they have high *variance*.

Note: This use of *bias* is in its statistical sense—consistently producing error in a particular direction.

Unsupervised Learning

The models discussed to this point are supervised learning algorithms, where the goal is to make predictions based on learning from data where an outcome is known. There are other widely used data science algorithms that do not focus on learning to make predictions from labeled data.

Clustering

Clustering algorithms group together records that are similar to one another. The result is a limited number of clusters that are relatively homogeneous and unlike other clusters. Identifying customer segments is a popular application of clustering.

Recommender Systems

- Virtually all consumers are familiar with the results of recommender systems, which can be seen in product recommendations on Amazon and other digital merchants. They also operate less visibly in social media to display ads and content that you are most likely to be interested in. These recommender systems tailor the content feeds for individuals, seeking to maximize user engagement. Engagement is generated by content that is provocative and inflammatory, and that speaks to a user's concerns and biases. It may often be fabricated. Even without actively provoking, these same recommender algorithms that underpin social media companies also enable political extremists to coalesce and take action.

Our focus in this book is on models that make predictions, because it is in this area that so many concerns about ethical implementation and transparency have arisen, particularly as machines take over the decision-making process.

The Ethical Challenge of Black Boxes

With the arrival of big data and data mining (now commonly referred to as data science), the ability to make real-time automated decisions has also arrived. Knowing how a variable (e.g., income) affects an outcome (e.g., spending) has become comparatively less important—just being able to predict the outcome is often considered sufficient. The black-box methods described earlier share one attribute: They make it hard to discern the effects of individual predictor variables (features) on outcomes. Their impact on the outcome variable lies hidden inside the black box (see Figure 1.7).

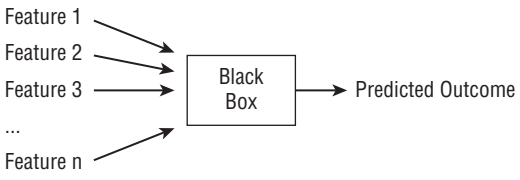

Figure 1.7: The black-box model conceals the effects of features.

Two Opposing Forces

Less than a generation ago, AI and machine learning consisted mainly of clever but still relatively prosaic extensions of traditional statistics in service of business goals—models that would predict the ability to repay loans or propensity to buy a product. Now, AI and machine learning are finally moving into realms that we would have considered science fiction less than a generation ago: cars that drive themselves, machines that carry on adaptive and seemingly intelligent conversations with you, algorithms that create videos of you saying things you never said, and much more.

Pressure for More Powerful AI

This quasimagical aspect of AI acts as an accelerant to the development of even more clever algorithms. Young data scientists and their visionary funders are more motivated by the prospect of developing something disruptive and exciting than they are by the possibility of contributing an extra 2% to a company's margin. This motivating pressure takes on a life of its own, widening and deepening the roles that these cutting-edge AI and machine learning models play in our economic and personal lives. Of course, the widening scope of AI approaches, along with the increased trust placed in them, can be harmful. Taken to the extreme, these models act in service of what researchers like AI Now Institute co-director Meredith Broussard refer to as *technochauvinism* and *technosolutionism*, or the belief that technical approaches alone are capable of solving complex social issues.

Public Resistance and Anxiety

At the same time (and perhaps as a result of increasing backlash against technosolutionism), the public's anxiety and wariness grow about AI's role in society with each new story of violations of privacy, due process, fairness, or the social compact. Regulatory countermeasures, such as the GDPR within the European Union or the recently introduced Algorithmic Accountability Act in the United States, have come about as a result. This anxiety is exacerbated by the ways in which governments like that in China, citing the need for regulation, have comprehensively constrained technology companies, leveraging AI's Big Brother capabilities while limiting its ability to act as a force for transparency and justice.

It is important to understand the grassroots nature of AI development, driven by the need to do things that are clever and cool. Much as water running downstream in a creek will easily outflank large boulders, advances in AI will not remain bottled up for long. Both the development of new AI technologies and the public's wariness of AI's ill effects will continue in opposition to each other.

Resolving that tension is not something that can be solved with a single decisive action. Finding the right "balance" of when, where, and how AI methods ought to be used is an ongoing issue that will require constant attention and management.

To prepare for a discussion on how this issue can be navigated, we will use the next chapter to dive deeper into the statistical and machine learning modeling methods that we have outlined so far.

SUMMARY

We began this chapter with a review of how a well-intentioned algorithm to predict an individual's need for healthcare services ended up significantly underpredicting this need among African Americans, whose access to healthcare as a group is already limited. We saw how statistics and data science have led a Jekyll and Hyde existence for more than a century, dating from when modern statistical methods were born, joined at the hip with eugenics. Good and bad are often combined in the same package.

Nowadays, automated decision-making and lack of transparency are the sources of danger associated with AI models. We introduced the distinction between interpretable models and black-box models that are not interpretable. Finally, we began our tour of predictive models, a tour that will continue in Chapter 2, "Background: Modeling and the Black Box Algorithm."