3. *The Science of Many Models*

Nothing is less real than realism. Details are confusing. It is only by selection, by elimination, by emphasis that we get to the real meaning of things.

—Georgia O'Keeffe

In this chapter, we take a scientific approach to motivate the many-model approach. We begin with the Condorcet jury theorem and the diversity prediction theorem, which make quantifiable cases for the value of many models in helping us act, predict, and explain. These theorems may overstate the case for many models. To show why, we introduce categorization models, which partition the world into boxes. Using categorization models shows us that constructing many models may be harder than we expect. We then apply this same class of model to discuss model granularity—how specific our models should be—and help us decide whether to use one big model or many small models. The choice will depend on the use. When predicting, we often want to go big. When explaining, smaller is better.

The conclusion addresses a lingering concern. Many-model thinking might seem to require learning a lot of models. While we must learn some models, we need not learn as many as you might think. We do not need to master a hundred models, or even fifty, because models possess a one-to-many property. We can apply any one model to many cases by reassigning names and identifiers and modifying assumptions. This property of models offers a counterpoise to the demands of many-model thinking. Applying a model in a new domain requires creativity, an openness of mind, and skepticism. We must recognize that not every model will appropriate to every task. If a model cannot explain, predict, or help us reason, we must set it aside.

The skills required to excel at one-to-many differ from the mathematical and analytic talents many people think of as necessary for being a good modeler. The process of one-to-many involves creativity. It is to ask: *How many uses can I think of for a random walk?* To provide a hint of the forms that creativity takes, at the end of the chapter we apply the geometric formula for area and volume as a model and use it to explain the size of supertankers, to criticize the body mass index, to predict the scaling of metabolisms, and to explain why we see so few women CEOs.

Many Models as Independent Lies

We now turn to formal models that help reveal the benefits of many-model thinking. Within those models, we describe two theorems: the Condorcet jury theorem and the diversity prediction theorem. The *Condorcet jury theorem* is derived from a model constructed to explain the advantages of majority rule. In the model, jurors make binary decisions of guilt or innocence. Each juror is correct more often than not. In order to apply the theorem to collections of models instead of jurors, we interpret each juror's decision as a classification by a model. These classifications could be actions (buy or sell) or predictions (Democratic or Republican winner). The theorem then tells us that by constructing multiple models and using majority rule we will be more accurate than if we used one of the constituent models. The model relies on the concept of a *state of the world,* a full description of all relevant information. For a jury, the state of the world consists of the evidence presented at trial. For models that measure the social contribution of a charitable project, the state of the world might correspond to the project's team, the organizational structure, the operational plan, and the characteristics of the problem or situation the project would address.

Condorcet Jury Theorem

Each of an odd number of people (models) classifies an unknown state of the world as either true or false. Each person (model) classifies correctly with a probability *p* > image, and the probability that any person (model) classifies correctly is statistically independent of the correctness of any other person (model).

Condorcet jury theorem: A majority vote classifies correctly with higher probability than any person (model), and as the number of people (models) becomes large, the accuracy of the majority vote approaches 100%.

Ecologist Richard Levins elaborates on how the logic of the theorem applies to the many-model approach: “Therefore, we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results, we have what we can call a robust theorem, which is relatively free of the details of the model. Hence our truth is the intersection of independent lies.”[1] Note that here he aspires to a unanimity of classification. When many models make a common classification, our confidence should soar.

Our next theorem, the *diversity prediction theorem*, applies to models that make numerical predictions or valuations. It quantifies the contributions of model accuracy and model diversity to the accuracy of the average of those models.[2]

Diversity Prediction Theorem

Many-Model Error = Average-Model Error − Diversity of Model Predictions

image

where M_i equals model *i*’s prediction, image equals the average of the model’s values, and *V* equals the true value.

The diversity prediction theorem describes a mathematical identity. We need not test it. It always holds. Here is an example. Two models predict the number of Oscars a film will be awarded. One model predicts two Oscars, and the other predicts eight. The average of the two models’ predictions—the many-model prediction—equals five. If, as it turns out, the film wins four Oscars, the first model’s error equals 4 (2 squared), the second model’s error equals 16 (4 squared), and the many-model error equals 1. The diversity of the models’ predictions equals 9 because each differs from the mean prediction by 3. The diversity prediction theorem can then be expressed as follows: 1 (the many-model error) = 10 (the average-model error) − 9 (the diversity of the predictive models).

The logic of the theorem relies on opposite types of errors (pluses and minuses) canceling each other out. If one model predicts a value that is too high and another model predicts a value that is too low, then the models exhibit predictive diversity. The two errors cancel, and the average of the models will be more accurate than either model by itself. Even if both predict values that are too high, the error of the average of those predictions will still not be worse than the average error of the two high predictions.

The theorem does not imply that any collection of diverse models will be accurate. If all of the models share a common bias, their average will also contain that bias. The theorem does imply that any collection of diverse models (or people) will be more accurate than its average member, a phenomenon referred to as the *wisdom of crowds*. That mathematical fact explains the success of ensemble methods in computer science that average multiple classifications as well as evidence that individuals who think using multiple models and frameworks predict with higher accuracy than people who use single models. Any single way of looking at the world leaves out details and makes us prone to blind spots. Single-model thinkers are less likely to anticipate large events, such as market collapses or the Arab Spring of 2011.[3]

These two theorems make a compelling case for using many models, at least in the context of prediction. The case may be too compelling, however. The Condorcet jury theorem implies that with enough models, we would almost never make a mistake. The diversity prediction theorem implies that if we could construct a diverse set of moderately accurate predictive models, we can reduce our many-model error to near zero. As we see next, our ability to construct many diverse models has limits.

Categorization Models

To demonstrate why the two theorems may overstate the case, we rely on *categorization models*. These models provide micro-foundations for the Condorcet jury theorem. Categorization models partition the states of the world into disjoint boxes. Such models date to antiquity. In *The Categories*, Aristotle defined ten attributes that could be used to partition the world. These included *substance*, *quantity*, *location*, and *positioning*. Each combination of attributes would create a distinct category.

We use categories any time we use a common noun. “Pants” is a category; so are “dogs,” “spoons,” “fireplaces,” and “summer vacations.” We use categories to guide actions. We categorize restaurants by ethnicity—Italian, French, Turkish, or Korean—to decide where to have lunch. We categorize stocks by their price-to-earnings ratios and sell stocks with low price-to-earnings ratios. We use categories to explain, as when we claim that Arizona’s population has grown because the state has good weather. We also use categories to predict: we might forecast that a candidate for political office with military experience has an increased chance of winning.

We can interpret the contributions of categorization models within the wisdom hierarchy. The objects constitute the data. Binning the objects into categories creates information. The assigning of valuations to categories requires knowledge. To critique the Condorcet jury theorem, we rely on a *binary categorization model* that partitions the objects or states into two categories, one labeled “guilty” and one “innocent.” The key insight will be that the number of relevant attributes constrains the number of distinct categorizations, and therefore the number of useful models.

Categorization Models

There exists a set of objects or states of the world, each defined by a set of attributes and each with a value. A **categorization model**, M, partitions these objects or states into a finite set of categories $\{S_1, S_2, \ldots, S_n\}$ based on the object’s attributes and assigns **valuations** $\{M_1, M_2, \ldots, M_n\}$ for each category.

Imagine we have one hundred student loan applications, half of which were paid back and half of which were defaulted. We know two pieces of information for each loan: whether the loan amount exceeded $50,000, and whether the recipient majored in engineering or the liberal arts. These are the two attributes. With two attributes we can distinguish between four types of loans: large loans to engineers, small loans to engineers, large loans to liberal arts majors, and small loans to liberal arts majors.

A binary categorization model classifies each of these four types as either repaid or defaulted. One model might classify small loans as repaid and large loans as defaulted. Another model might classify loans to engineers as repaid and loans to liberal arts majors as defaulted. It seems plausible that each of these models could be correct more than half the time, and that the two models might be approximately independent of each other. A problem arises when we try to construct more models. There exist only sixteen unique models that map four categories into two outcomes. Two of those models classify all loans as repaid or defaulted. Each of the remaining fourteen has an exact opposite. Whenever the model classifies correctly, its opposite model classifies incorrectly. Thus, of the fourteen possible models, at most seven can be correct more than half the time. And if any model happens to be correct exactly half of the time, then so must its opposite.

The dimensionality of our data limits the number of models we can produce. At most we can have seven models. We cannot construct eleven independent models, much less seventy-seven. Even if we had higher-dimensional data—say, if we knew the recipient's age, grade point average, income, marital status, and address—the categorizations that relied on those attributes must yield accurate predictions. Each subset of attributes would have to be relevant to whether the loan was repaid and be uncorrelated with the other attributes. Both are strong assumptions. For example, if address, marital status, and income are correlated, then models that swap those attributes will be correlated as well.[4] In the stark probabilistic model, independence seemed reasonable: different models make independent mistakes. When we unpack that logic with categorization models, we see the difficulty of constructing multiple independent models.

Attempts to construct a collection of diverse, accurate models encounter a similar problem. Suppose that we want to build an ensemble of categorization models that predict unemployment rates across five hundred mid-size cities. An accurate model must partition cities into categories such that within a category the cities have similar unemployment rates. The model must also predict unemployment accurately for each category. For two models to make diverse predictions, they must categorize cities differently, predict differently, or do both. Those two criteria, though not in contradiction, can be difficult to satisfy. If one categorization relies on average education level and a second relies on average income, they may categorize similarly. If so, the two models will be accurate but not diverse. Creating twenty-six categories using the first letter of each city's name will create a diverse categorization but probably not an accurate model. Here as well, the takeaway is that in practice "many" may be closer to five than fifty.

Empirical studies of prediction align with that inference. While adding models improves accuracy (they have to, given the theorems), the marginal contribution of each model falls off after a handful of models. Google found that using one interviewer to evaluate job candidates (instead of picking at random) increases the probability of an above-average hire from 50% to 74%, adding a second interviewer increases the probability to 81%, adding a third raises it to 84%, and using a fourth lifts it to 86%. Using twenty interviewers only increases the probability to a little over 90%. That evidence suggests a limit to the number of relevant ways of looking at a potential hire.

A similar finding holds for an evaluation of tens of thousands of forecasts by economists regarding unemployment, growth, and inflation. In this case, we should think of the economists as models. Adding a second economist improves the accuracy of the prediction by about 8%, two more increase it by 12%, and three more by 15%. Ten economists improve the accuracy by about 19%. Incidentally, the best economist is only about 9% better than average—assuming you knew which economist was best. So three random economists perform better than the best one.[5] Another reason for averaging many and not relying on the economist who has been best historically is that the world changes. The economist who performs at the top today may be middling tomorrow. That same logic explains why the US Federal Reserve relies on an ensemble of economic models rather than just one: the average of many models will typically be better than the best model.

The lesson should be clear: if we can construct multiple diverse, accurate models, then we can make very accurate predictions and valuations and choose good actions. The theorems validate the logic of many-model thinking. What the theorems do not do, and cannot do, is construct the many models that meet their assumptions. In practice, we may find that we can construct three or maybe five good models. If so, that would be great. We need only read back one paragraph: adding a second model yields an 8% improvement, while adding a third gets us to 15%. Keep in mind, these second and third models need not be better than the first model. They could be worse. If they are a little less accurate, but categorically (in the literal sense) different, they should be added to the mix.

One Big Model and the Granularity Question

Many models work in theory and in practice. That does not mean that they are always the correct approach. Sometimes we are better off constructing a single large model. In this section, we put some thought into when we should use each approach and along the way take up the *granularity question* of how finely we should partition our data.

To take on the first question, of whether to use one big model or many small ones, recall the uses of models: to *reason, explain, design, communicate, act, predict,* and *explore*. Four of these uses—to reason, explain, communicate, and explore—require simplification. By simplifying, we can apply logic allowing us to explain phenomena, communicate our ideas, and explore possibilities.

Think back to the Condorcet jury theorem. Within it, we could unpack logic, explain why an approach that uses many models was more likely to produce a correct result, and communicate our findings. Had we constructed a model of jurors with personality types and described the evidence as vectors of words, we would have been lost in a mangle of detail. Borges elaborates on this point in an essay on science. He describes mapmakers who make ever more elaborate maps: “The Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that this vast Map was useless.”

The three other uses of models—to *predict, design,* and *act*—can benefit from high-fidelity models. If we have BIG data, we should use it. As a rule of thumb, the more data we have, the more granular we should make our model. This can be shown by using categorization models to structure our thinking. Suppose first that we want to construct a model to explain variation in a data set. To provide context, suppose that we have an enormous data set from a chain of grocery stores detailing monthly spending on food for several million households. These households differ in the amount they spend, which we measure as variation: the sum of the squared differences between what each family spends and average spending across all households. If average spending is $500 a month and a given family spends $520, that family contributes 400, or 20 squared, to the total variation. Statisticians call the proportion of the variation that a model explains the model’s R^2.

If the data had a total variation of 1 billion and a model explains 800 million of that variation, then the model has an R^2 of 0.8. The amount of variation explained corresponds to how much the model improves on the mean estimate. If the model estimates that a household will spend $600 and the household in fact spent $600, then the model explains all 10,000 that the household contributes to total variation. If the household spent $800 and the model says $700, then what had been a contribution of 90,000 to total variation ($(800-500)^2$) is now only a 10,000 contribution ($(800-700)^2$). The model explains image of the variation.

R^2: Percentage of Variance Explained

image

where $V(x)$ equals the value of x in X, image equals the average value, and $M(x)$ equals the model’s valuation.

In this context, a categorization model would partition the households into categories and estimate a value for each category. A more granular model would create more categories. This may require considering more attributes of the households to create those categories. As we add more categories, we can explain more of the variation, but we can go too far. If we follow the example of Borges's mapmakers and place each household in its own category, we can explain all of the variation. That explanation, like the life-sized map, would not be of much use.

Creating too many categories overfits the data, overfitting undermines prediction of future events. Suppose that we want to use last month's data on grocery purchases to predict this month's data. Households vary in their monthly spending. A model that places each household in its own category would predict that each household spends the same as in the previous month. That would not be a good predictor given monthly fluctuations in spending. By placing the household into a category with other similar households, we can use the average spending on groceries for similar households to create a more accurate predictor.

To do this, we think of each household's monthly purchases as a draw from a distribution (we will cover distributions in Chapter 5). That distribution has a mean and a variance. The objective in creating a categorization model is to construct categories based on attributes so that the households within the same category have similar means. If we can do that, one household's spending in the first month tells us about the other households' spending in the second month. No categorization will be perfect. The means of households within each category will differ by a little. We call this *categorization error*.

As we make larger categories, we increase categorization error, as we are more likely to clump households with different means into the same category. However, these larger categories rely on more data, so our estimates of the means in each category will be more accurate (see the square root rules in Chapter 5). The error from misestimating the mean is called the *valuation error*. Valuation error decreases as we make categories larger. One or even ten houses per category will not give an accurate estimate of the mean if households vary substantially in their monthly spending. A thousand households will.

We now have the key intuition: increasing the number of categories decreases the categorization error from binning households with different means into the same category. Statisticians call this *model bias*. However, making more categories increases the error from estimating the mean within each category. Statisticians refer to this as increasing the *variance* of the mean. The trade-off in how many categories to create can be expressed formally in the *model error decomposition theorem*. Statisticians refer to the result as the bias-variance trade-off.

Model Error Decomposition Theorem

The Bias-Variance Trade-off

Model Error = Categorization Error + Valuation Error

image

where $M(x)$ and M_i denote the model's values for data point x and category S_i and $V(x)$ and V_i denote their true values.[6]

One-to-Many

Learning models takes time, effort, and breadth. To reduce those demands, we take a *one-to-many* approach. We advocate mastering a modest number of flexible models and applying them creatively. We use a model from epidemiology to understand the diffusion of seed corn, Facebook, crime, and pop stars. We apply a model of signaling to advertising, marriage, peacock feathers, and insurance premiums. And we apply a rugged-landscape model of evolutionary adaption to explain why humans lack blowholes. Of course, we cannot take any model and apply it to any context, but most models are flexible. We gain even when we fail because attempts at creative uses of models reveal their limits. And it is fun.

The one-to-many approach is relatively new. In the past, models belonged to specific disciplines. Economists had models of supply and demand, monopolistic competition, and economic growth; political scientists had models of electoral competition; ecologists had models of speciation and replication; and physicists had models describing laws of motion. All of these models were developed with specific purposes in mind. One would not apply a model from physics to the economy or a model from economics to the brain any more than one would use a sewing machine to repair a leaky pipe.

Taking models out of their disciplinary silos and practicing one-to-many has produced notable successes. Paul Samuelson reinterpreted models from physics to explain how markets attain equilibria. Anthony Downs applied a model of ice cream vendors competing on a beach to explain the positioning of political candidates competing in ideological space. Social scientists have applied models of interacting particles to explain poverty traps, variation in crime rates, and even economic growth across countries. And economists have taken models of self-control based on economic principles to understand the functioning of the brain.[7]

One-to-Many: Higher Powers (X^N)

Creatively applying models requires practice. To provide a preview of the potential of the *many-to-one* principle, we take the familiar formula of a variable raised to a power, X^N, and apply it as a model. When the power equals 2, the formula gives the area of a square, when the power equals 3, it gives the volume of a cube. When raised to higher powers, it captures geometric expansion or decay.

Supertankers: Our first application considers a cubic supertanker whose length is eight times its depth and width, which we denote by S. As shown in figure 3.1, the supertanker has a surface area of $34S^2$ and a volume of $8S^3$. The cost of building a supertanker depends primarily on its surface area, which determines the amount of steel used. The amount of revenue a supertanker generates depends on its volume. Computing the ratio of volume to surface area, image, reveals a linear gain in profitability from increasing size.

image

Figure 3.1: A Cubic Supertanker: Surface Area = $34S^2$, Volume = $8S^3$

Shipping magnate Stavros Niarchos, who knew this ratio, built the first modern supertankers and made billions during the period of rebuilding that followed World War II. To give some sense of scale: the T2 oil tanker used during World War II measured 500 feet long, 25 feet deep, and 50 feet wide. Modern supertankers such as the *Knock Nevis* measure 1,500 feet long, 80 feet deep, and 180 feet wide. Imagine tipping the Willis (Sears) Tower in Chicago on its side and floating it in Lake Michigan. The *Knock Nevis* resembles a T2 oil tanker scaled up by a factor of a little over three. The *Knock Nevis* has about ten times the surface area as a T2 oil tanker and over thirty times the volume. A question arises as to why supertankers are not even larger. The short answer is that tankers must pass through the Suez Canal; the *Knock Nevis* squeezes through with a gap of a few feet on each side.[8]

Body mass index: Body mass index (BMI) is used by the medical profession to define weight categories. Developed in England, BMI equals the ratio of a person's weight (in kilograms) to her height in meters squared.[9] Holding height constant, BMI increases linearly with weight. If one person weighs 20% more than another person of the same height, the first person's BMI will be 20% higher.

We first apply our model to approximate a person as a perfect cube made up of some mixture of fat, muscle, and bone. Let M denote the weight of one cubic meter of our cubic person. The human cube's weight equals its volume times the weight per cubic meter, or $H^3 \cdot M$. Our cube's BMI equals $H \cdot M$. Our model reveals two flaws: BMI increases linearly with height, and given that muscle weighs more than fat, fit people have higher M and therefore higher BMIs. Height should be unrelated to obesity, and muscularity is the *opposite* of fatness. These flaws remain if we make the model more realistic. If we make a person's depth (thickness front to back) and width proportional to height using parameters *d* and *w*, then BMI can be written as follows: image The BMIs of many NBA stars and other athletes place them in the overweight category (BMI > 25), along with many of the world's top male decathletes.[10] Given that even moderately tall, physically fit people will likely have high BMIs, we should not be surprised that a meta-analysis of nearly a hundred studies with a combined sample size in the millions found that slightly overweight people live longest.[11]

Metabolic rates: We now apply our model to predict an inverse relationship between an animal's size and its metabolic rate. Every living entity has a metabolism, a repeated sequence of chemical reactions that breaks down organic matter and transforms it into energy. An organism's metabolic rate, measured in calories, equals the amount of energy needed to remain alive. If we construct cubic models of a mouse and an elephant, figure 3.2 shows that the smaller cube has a much larger ratio of surface area to volume.

image

Figure 3.2: The Exploding Elephant

We can model the mouse and the elephant as composed of cells 1 cubic inch in volume, each with a metabolism. Those metabolic reactions produce heat that must dissipate through the surface of the animal. Our mouse has a surface area of 14 square inches and a volume of 3 cubic inches, a surface-to-volume ratio of roughly 5:1.[12] For each cubic-inch cell in its volume, the mouse has five square inches of surface area through which it can dissipate heat. Each heat-producing cell in the elephant has only one-fifteenth of a square inch of surface area. The mouse can dissipate heat at seventy-five times the rate of the elephant.

For both animals to maintain the same internal temperature, the elephant must have a slower metabolism. It does. An elephant with a mouse's metabolism would require 15,000 pounds of food per day. The elephant's cells would also produce too much heat to be dissipated through its skin. As a result, elephants would smolder and then explode. The reason elephants do not blow up is that they have a metabolism roughly twenty times lower than that of mice. The model does not predict the rate at which metabolism scales with size, only the direction. More elaborate models can explain the scaling laws.[13]

Women CEOs: For our last application, we increase the exponent in the formula and use the model to explain why so few women become CEOs. In 2016, fewer than 5% of Fortune 500 companies had women CEOs. To become a CEO a person must receive multiple promotions. We can model those promotion opportunities as probabilistic events: a person has some probability of receiving a promotion. We further assume that to become CEO, a person must be promoted at each opportunity.

We assume fifteen promotion opportunities as a benchmark, as that corresponds to a promotion every two years on a thirty-year path to CEO. The weight of evidence reveals modest biases in favor of men, which we can model as men having a higher probability of being promoted.[14] We model this as a man's probability of promotion, P_M, being slightly larger than a woman's, P_W. If we benchmark these probabilities at 50% and 40%, respectively, then a man is nearly thirty times more likely than a woman to become CEO.[15] The model reveals how modest biases accumulate. A 10% difference in promotion rates becomes a 30-fold bias at the top. This same model provides a novel explanation for why a much larger percentage (about 25%) of college and university presidents are women. Colleges and universities have fewer administrative layers than Fortune 500 companies. A professor can become president in as few as three promotions: department chair, dean, and then president. Less bias accumulates over three levels. Thus, the larger proportion of women presidents need not imply that educational institutions are more egalitarian than corporations.

Summary

We began the chapter by laying logical foundations for the many-to-one approach using the Condorcet jury theorem and the diversity prediction theorem. We then used categorization models to show the limits of model diversity. We saw how many models can improve our abilities to predict, act, design, and so on. We also saw that it is not easy to come up with many diverse models. If we could, then we could predict with near perfect accuracy, which we know we cannot. Nevertheless, our goal will be to construct as many useful, diverse models as possible.

In the chapters that follow, we describe a core set of models. Those models make salient different parts of the world. They make different assumptions about causal interactions. Through their diversity they create the potential for productive many-model thinking. By emphasizing distinct parts of more complex wholes, each model contributes on its own. Each also can be part of an even more powerful ensemble of models.

As noted earlier, many-model thinking does require that we know more than one model. However, we need not know a huge number of models, so long as we can apply each model that we do know in multiple domains. That will not always be easy. Successful one-to-many thinking depends on creatively tweaking assumptions and constructing novel analogies in order to apply a model developed for one purpose in a new context. Thus, becoming a many-model thinker demands more than mathematical competence; it requires creativity as was evident in our many applications of our model of a cube.

Bagging and Many Models

Often we fit a model to a sample from an existing data set and then test that same model against the remainder of the data. Other times we fit a model to existing data and use that model to predict future data. This type of modeling creates a tension: the more parameters we include in our model, the better we can fit data and the more we risk overfitting. Good fit does not imply a good model. Physicist Freeman Dyson tells of Enrico Fermi's reaction to a piece of Dyson's research that had exceptional model fit. "In desperation I asked Fermi whether he was not impressed by the agreement between our calculated numbers and his measured numbers. He replied, 'How many arbitrary parameters did you use for your calculations?' I thought for a moment about our cut-off procedures and said, 'Four.' He said, ‘I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.’ With that, the conversation was over.”[16]

The estimates used to “wiggle the trunk” often include higher-order terms: squares, cubes, and fourth powers. This introduces a risk of large errors, because higher-order terms amplify. While 10 is twice as large as 5, 10^4 is 16 times as large as 5^4. The figure below shows an example of overfitting.

image

Overfitting and Out-of-Sample Error

The graph on the left shows (hypothetical) sales data from a company that manufactures industrial 3-D printers as a function of the number of site visits made (on average) per month by their sales team. The graph on the left shows a nonlinear best fit that includes nonlinear terms up to the fifth power. The graph on the right shows that the model predicts sales of 100 printers if sales visits reach 30. That cannot be correct if customers buy at most one 3-D printer. By overfitting, the model makes a huge error out of the sample.

To prevent overfitting, we could avoid higher-order terms. A more sophisticated solution known as *bootstrap aggregation* or *bagging* constructs many models. To bootstrap a data set, we create multiple data sets of equal size by randomly drawing data points from the original data. The points are drawn with replacement—after we draw a data point, we put it back in the “bag” so that we might draw it again. This technique produces a collection of data sets of equal size, each of which contains multiple copies of some data points and no copies of others.

We then fit (nonlinear) models to each data set, resulting in multiple models.[17] We can then plot all the models on the same set of axes, creating a *spaghetti graph* (see below). The dark line shows the average of the different models.

image

Bootstrapping and a Spaghetti Graph

Bagging will capture robust nonlinear effects, as they will be evident in multiple random samples of the data, while avoiding fitting idiosyncratic patterns in any single data set. By building diversity through random samples and then averaging the many models, bagging applies the logic that underpins the diversity prediction theorem. It creates diverse models, and as we know, the average of those models will be more accurate than the models themselves.