# Chapter 4

# Many to One: The Intersection of Independent Lies

*If I had a hammer, I'd hammer in the morning.*

*I'd hammer in the evening, all over this land.*

Lee Hays and Pete Seeger

Models help explain, predict, understand, and modify our world. Yet, as we have learned: all models are wrong (some more so than others). We apply multiple models to correct for those errors. To paraphrase Box and Draper, *though any one model will be wrong, many can be useful.*

In this chapter, we explore the value of multiple models highlighting nine reasons: *to leverage diverse representations*, *to identify distinct logics*,*to identify the boundaries of the possible*, *to combine models*, *to prevent overfitting*, *to capture phenomena at multiple levels*, *to cope with nonstationarity*, *to understand a non stationary world* and *to explore new possibilities*. These nine reasons only partly align with the more general reasons to model. We can bundle these nine reasons into two broader categories *accuracy* and *robustness*.

We begin the chapter with a discussion of why physicists often need only a single model and then show the value of many models in two historical cases: The Bay of Pigs and the 2008 Global Financial Crisis. We then elaborate on the nine reasons for many models. The bulk of the chapter is devoted to an analysis of how how many models improve prediction. We cover three results: the *Diversity Prediction Theorem*, the *Many Models Outperform the Average* and the *Categorization Prediction Theorem.* These results provide statistical justifications for multiple models. Each result can be expressed as an equality or an inequality that reveals that multiple models are better than one model at prediction. In the final part, we discuss how many models enhance robustness in decision making and discuss why many models instead of one big model.

## Unreasonably Effective Single Models

The many model approach advocated here differs form the single model approach we learn in high school science. In the physical sciences, a single model often suffices. To calculate force, we can apply $F = ma$. To calculate the relationship between the volume and pressure of a gas, we use Boyle's law. We can use one model, because the models work so well. In fact, they're *unreasonably effective* (Wigner1960). Quantum theory can predict phenomena to nine decimal points. [26]

Social science models are not unreasonably effective. Models of economies, political systems, or violent behavior explain only modest amounts of the variation that exists in the world and identify relatively few factors whose effects have large magnitudes (Ziliak and McCloskey 2008). The limited success of social science models can be partly attributed to the complexity of the task. People are sophisticated and multi-dimensional. We march to our own drummers. Even though billions of people exist, we interact in small to moderate sized groups, so we lack the big numbers of physical systems. Even more vexing to the modeler, we learn. We adapt. We do crazy things. We're socially influenced. Thus, not only do we exhibit variation in behavior, that variation doesn't cancel out because behavior is correlated.

The difference in the accuracy of social science models and physical model is evident scatterplots. Figure 4 shows data for income as a function of IQ alongside Boyle's original data relating volume and pressure. IQ is clearly not an unreasonably accurate predictor of income. To explain more of the variation in IQ, we need more variables and more models.

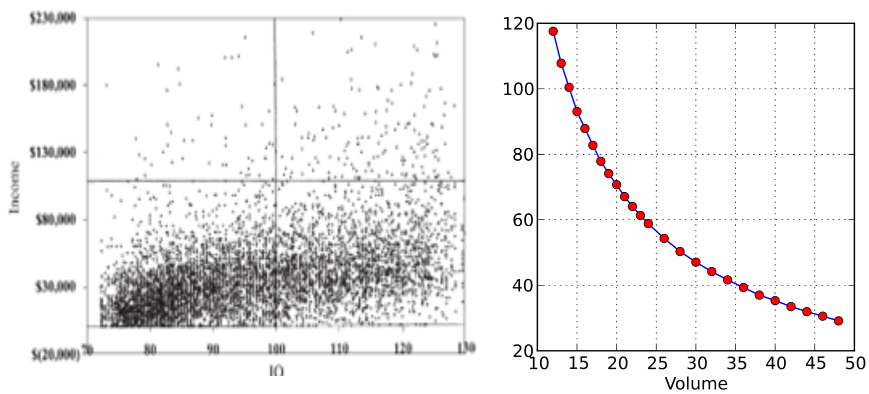


Figure 4.1: Income as a Function of IQ (Zagorsky 2007) and Pressure as a Function of Volume (Boyle's Law)

# A War Avoided, A Crash Endured

On April 17, 1961, a CIA trained paramilitary group landed on the shores of Cuba in a failed attempt to overthrow Fidel Castro's communist regime. Now called the Bay of Pig's Invasion, the landing led to heightened tension between the United States and the Soviet Union. Later that year, Soviet Premier Nikita Kruschev moved short range nuclear missiles to Cuba. In the eyes of many, this act pushed the world to the brink of nuclear war. President John F Kennedy reacted by blockading Cuba. The Soviet Union backed down. A crisis was averted.

How do we explain the events? We could write a rational actor model, boil the crisis down to its essence. Kennedy had a small set of options. He could start a nuclear war. He could invade Cuba militarily. He could impose a blockade. He chose the blockade. He chose a good action.

To explain why Kennedy and his advisors chose a blockade and why it worked using a *rational actor /game theory* model, we assume that Kennedy chooses an action by thinking through how the Soviets would respond. The logic goes as follows: the blockade would starve the Cubans. This would force the Soviet Union to choose one of two actions: to back down or to launch missiles. Assuming the Soviet Union's leadership valued human life, they would back down. Not backing down would result in the deaths of millions of people.

The game theory model includes two primary actors, a set of possible actions each can take, the likely outcomes of those actions, and the actors' preferences over outcomes. Though parsimonious, to first approximation, the model captures events. The model reveals the a central logic of the situation and reveals why Kennedy chose the blockade (he knew the Soviet's would then have to choose to remove the missiles).

Though it captures the main facts, the rational actor model doesn't provide a full explanation. For example, why didn't the Soviets hide the missiles? In *Essence of Decision: Explaining the Cuban Missile Crisis* Graham Allison contrasts the rational actor model with two other models: an *organizational process model* and a *governmental process model*.

These other models can explain deviations from the rational actor model. The organizational process model explains that the missiles weren't hid because of a lack of organizational capacity. Further, an organizational model can explain Kennedy's choice of blockade not as the rational action but as the only possible action. The United States Air Force probably lacked the capacity to wipe out all of the missiles in a surgical strike. One missile left over would be one too many.

Allison's other model, the governmental process model, assumes that countries are not unitary actors. Both Kennedy and Kruschev made decisions with an eye toward politics. Kennedy had to worry about Republicans in congress as well as eventual re-election. Further, Kruschev needed to maintain a political base of support. This governmental process model can explain why Kruschev placed the missiles in Cuba. The act gave an appearance of strength.

This third model explains a central problem with the rational actor model. Had Kruschev been rational, he too should have reasoned ahead. If he had, he would have realized that Kennedy would put in a blockade, forcing Kruschev to remove the missiles. Kruschev therefore should never have entered the game.

We can see how each model explains some of the facts and fails to explain others. Each models is wrong. The actors may well have been trying to think and act rationally, but on some dimensions organizational and political capacities and constraints either determined or influenced behaviors and outcomes. By applying multiple models we can attain a more complete understanding.

In the Bay of Pig's, a crisis was avoided. That wasn't true in 2008, when global financial markets collapsed reducing total wealth (or what many thought was wealth) by trillions of dollars. The crash produced a four year long global recession. Why did the 2008 recession occur? Multiple plausible accounts have been put forward. Some claim that too much foreign investment from China led to a bubble in the United States real estate market. Others argue that investment banks were over leveraged – that their monetary obligations were too large relative to how much cash they had on hand. Moreover, many of their leveraged assists had been originated by mortgage banks with incentives to write as many loans as possible. Still others claim that the financial system was so complex that no one knew what was going on until the entire house of cards collapsed. Finally, some believed that the investment banks knew there was a bubble, and like Kennedy they looked down the game tree and knew that the government would bail them out because the banks were too big too fail.

If take these various accounts and turn them into models – even crude models – we can test whether they align with facts – to the extent facts are known. In a survey of twenty-one accounts of the crisis, Andrew Lo (2012) writes "we should strive at the outset to entertain as many interpretations of the same set of objective facts as we can, and hope that a more nuanced and internally consistent understanding of the crisis emerges in the fullness of time." He goes on to say that "Only by collecting a diverse and often mutually contradictory set of narratives can we eventually develop a more complete understanding of the crisis."

Lo advocates many model thinking. Multiple interdependent sequences of events and actions produced the global financial crises. These were occurring simultaneously. Money was flowing in from China. Loan originators were writing too many loans. Investment banks did increase leverage ratios. The financial instruments were complex. And banks did probably count on using political pressure to avoid bankruptcy.

Lo reasons that none of these models fully explains the crisis. Why would anyone put money into a system and contribute to a bubble leading to a global crisis? Even if loan originators were writing bad loans, they still had to distribute them to someone. And yes, leverage ratios did increase over 2002 levels but they were not that much higher than they had been ten years earlier. As for the notion that big banks wouldn't be allowed to fail and everyone knew that, while many banks received bailouts, Lehman Brothers failed. Again, all models are wrong but a collection of models deepens our understanding of events.

# Nine Reasons For Many Model Thinking

We now describe eight reasons for using many models. The first two reasons derive from more general benefits of diversity of thought. By having many models, we produce diverse representations and diverse logics. As mentioned, these arguments will be presented informally, i.e. without models.

> ## Reason #1 To Leverage Diverse Representations
>
> *Different models represent the world differently improving predictions, expanding the sets of potential casual forces and implications, and allowing for richer explorations.*

Suppose that we want to predict or explain why people vote. We might construct one model of voter turnout that relies on age. We might construct a second that relies on income as an explanatory variable. The first model would allow us to see patterns in turnout as a function of age, without accounting for income. The second model would allow us to see the effect of income. Each model would explain some of the variation in turnout. Each model could also be used to predict whether someone would vote. And each model could be used to help derive actions to increase turnout. If both models of turnout supported a similar action – say keeping polls open later – then we could have more confidence in that action.

The flow of information on the Internet might be modeled as a process in which individuals share links and videos. We might also model it as an evolutionary process. We might then ask whether information *mutates* – does it change over time, like in the telephone game where each person in a circle in turn whispers a story into the ear of the person on her right. As the story passes from person to person, errors accumulate. The final story can be different than the original story. It may even be nonsense.

Unlike the telephone game, on the Internet stories can be passed over multiple paths. A person may get the same information from two sources. This should reduce the potential for error. Yet, errors arise. And scientists have traced the mutations in messages through the web (Simmons et al 2011). These models of error propagation are borrowed from ecology. Using these models, computer scientists can predict the number of mutations. They can also approximate the *fitness* of those mutated messages by the frequency with which they are passed on to others. They can even investigate the correlation between mutations and fitness. Ironically,none of these three exercises: counting mutations, assigning fitness, and correlating fitness to number of mutations can be accomplished in biological settings.

These ecological models only capture a portion of what occurs on the web. They provide a logical structure within which to explore dynamics (how fast do ideas spread?), variation (at what rates do mutations occur?), selection (at what rates to mutations get passed along to others?), distributions (how many variants of a story persist?), and fitness (how accurate are the surviving stories?). These are all natural questions to a biologist, but they would not be necessarily all be natural questions to an economist, a physicist, a sociologist, or a psychologist. Those types of scientists would want other models.

When we choose a model, we implicitly choose a set of questions we can ask and (hopefully) answer. Ecological models enable us to ask questions about diversity. Economic models enable us to ask questions about equilibrium and efficiency. Physical models allow us to answer questions related to macro level dynamics. Each model lets us explore a problem in a different way.

## Reason #2 To Identify Distinct Logics

*Distinct models can rely on different causal logic leading to distinct predictions, explanations, and preferred actions or designs.*

Different models not only can include different variables and different representations, they can also rely on different logics. Suppose that we are trying to understand consumer behavior. Why do people choose a particular style of eyewear or clothing. We might assume that economically based behavior, that people choose clothing based on price and how it looks. We might alternatively assume that social pressure drives behavior, that people buy articles of clothing similar to those worn by others. Or, people may value low prices and be subject to social pressure. By working through each model independently, we can see the implications of each. We can also compare the likely effects of actions in each model. For example, in the more economic model cutting prices would lead to an increase in sales. In the social model, small changes in price may not meaningfully move sales figures.

If we had even more models – say a model based on competing clothing manufacturers, or a model based on seasonal spending on clothing, we would gain an even more nuanced understanding of what drives sales as well as a variety of predictions of the impact on sales of a price drop. The highest and lowest of these estimates can be thought of as upper and lower bounds on what could occur. This is a third reason to have multiple models: to identify the set of possible outcomes.

## Reason #3 To Identify Possible Outcomes

*By having multiple models, we get a better understanding of the set of possible outcomes.*

Identifying possible outcomes aids decision making. Suppose that you have an offer to join a new start up selling kayaks. They offer you one percent of the company's value. If the company becomes worth one hundred million dollars, you'd get a million. If it becomes worth two million, you get twenty-thousand. You might construct one model based on national sales of kayaks. You might construct another model based on growth of the sporting good sector, and a third model based on data from all manufacturing startups. If all three models predict values below $50 million, you shouldn't expect to become a millionaire from the job. If one model predicted a value of $200 million, you might be.

Recombination is another reason for multiple models. We can can combine models of the same process to create a more elaborate model. We can combine a models of consumer behavior based on economic principles and our model of social pressure. We can also combine simple behaviors to create behavioral diversity. We might assume that some people act in their self-interests and that others are altruistic. We might assume that some people optimize and that others use rules of thumb. Each model can be thought of as a lego piece. We can take out one behavioral rule (rational self interest) and replace it with another (regret minimization). We can also consider behavior to be a weighted average of two models. People may learn by best responding to what has happened in the past, by taking the behavior they believe will perform best, or by taking a weighted average of the two (Camerer 2003).

We can also combine models that address different parts of a process. We can combine a learning model and a network model to produce a model of learning on networks. Similarly, we can combine a model of worker effort and a model of hierarchy and create a model of worker effort in a hierarchy.

## Reason #4 New Dimensions and Phenomena

*Additional models include new dimensions and can lead to new knowledge.*

When we apply a model to data, we estimate parameters. Prior to using the model, those parameters might not have been considered. If we apply a model of learning to production processes, we find that *learning curves*: the changes in productivity over time exhibit a regular pattern. The model focuses our attention on variable or set of variables, in this case rates of output over time, that was unknown because it was not contemplated. To use the more poetic phrasing of T.S. Eliot, it was"not known, because not looked for."[1]

## Reason #5 Multiple Models Can be Combined

*We can create new models by combining existing models. These models may cover the same or different phenomena.*

A many model approach prevents us applying a single model too broadly. In trying too hard align a model and reality, one can make either of two mistakes. One can bend and distort facts so that they fit the model. In Voltaire's *Candide*, the character of Dr. Pangloss views the world as"the best of all possible worlds. " Pangloss, based partly on the German mathematician Leibniz, insists on this belief regardless of what befalls him. Pangloss is not

[1]Excerpted from Eliot's gorgeous *Little Gidding* that celebrates a rich life.

blind to reality. Instead, he uses his intellectual powers to squeeze the square peg of reality into his round holed theory by constructing ever more elaborate rationalizations. He is a single model thinker.

Single model thinking fairs poorly empirically. In a two decade long study involving nearly three hundred participants - many of whom were experts – making more than twenty-five thousand predictions about world events, Philip Tetlock (2005) found that people who used multiple ways of thinking (he called these people *foxes*) were much better than people who used a single mental model (he called this people *hedgehogs*).[2]

One can also contort a model to fit reality. This is known an *overfitting*. If we overfit a model, then we run the risk of that model being a poor predictor of other, new instances. Statisticians call this *out of sample* failure. Suppose for example that we have data on the price and quantity for sales of oregano at a grocery store and using those data we predict the price per ounce.

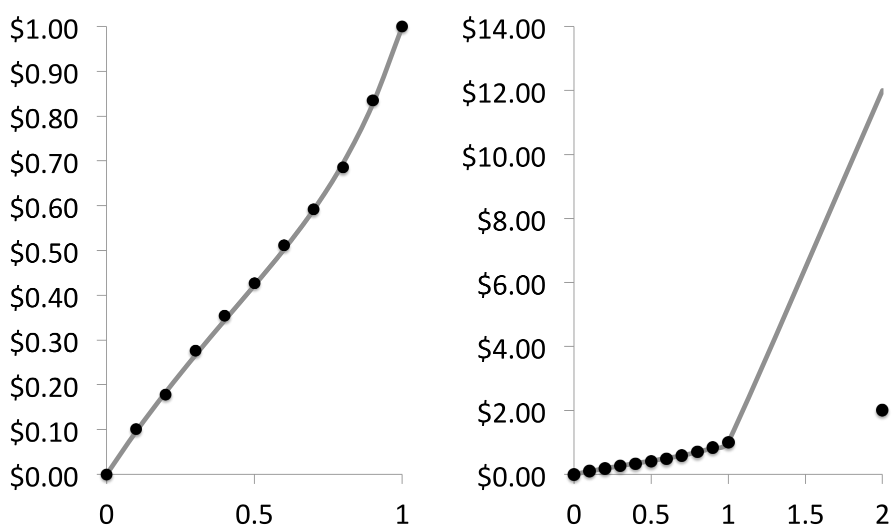


Figure 4.2: Overfitting Data

[2]Tetlock also found that one could predict better than his hedgehogs by dividing the possible events into three equally likely categories: *up*, *down*, and *unchanged* and randomly picking one of the three.

The data reveal that price of the oregano equals approximately one dollar per ounce, but due to rounding errors the data don't exactly fit on a straight line as can be seen in the left hand graph in figure 4. If we let $P$ equal the price and $Q$ equal the quantity, we can fit the data almost perfectly with the following polynomial equation:

$$P = Q - 50Q + 50Q^2 - 50Q^3 + 50Q^4$$

The equation overfits the data by adding three nonlinear terms that capture small deviations from linearity. The higher order terms that we added can produce huge predictive errors if we choose an independent variable that is outside the range of data used to fit the model. In our example, the non linear model predicts that two ounces of oregano would cost twelve dollars. This is shown in the right panel of figure 4. The actual cost would be approximately two dollars. If we move the amount further outside of the range, the errors produced by overfitting become even more extreme: the model predicts that ten ounces of oregano would cost forty-five thousand dollars!

To prevent overfitting, we could disallow higher order terms, i.e. restrict ourselves to a linear model. That solution prevents crazy predictions, but it would be flawed if the data do not fall on line. A more sophisticated solution uses a technique called *bootstrap aggregation* or *bagging* (Breiman 1996). Boostrap aggregation consists of three steps. First, we create multiple data sets by *bootstrapping* the data. Second, for each data set we construct a model, and third, we average those models. To *bootstrap* a data set, we create a new data set by randomly drawing data points from the original set until we have a set the same size as our original data set. This technique does not give us multiple replicas of the original data because we we draw points *with replacement*. That is, when we draw a data point, we return it to the original set. This will almost never result in an exact replica of the original data set. Some points will be chosen twice or even three times, and other points won't be chosen at all.

Bootstrap aggregation treats the actual data as one of many possible realizations and randomly creates a collection of other possible realizations. Each of these realizations gets its own model. Even if each model is overfit, so long as they're overfit in distinct and idiosyncratic ways, the average will have less error. We see why later in this chapter when we cover the *Diversity Prediction Theorem*.

## Reason #6 To Prevent Overfitting

*If ample data exists, a single model can be overfit. Using many models we avoids overfitting.*

Our next reason that we need multiple models is that our world exists at multiple scale. The phenomena that occur at each scale differ. In a story that has survived over two hundred years, a young person claims that the earth rests on the back of a giant elephant. A scientist asks the child what the elephant stands on, to which the child replies, "a giant turtle." Anticipating what's about to come next, the child quickly adds "don't even ask, it's turtles all the way down."[27]

If our world were turtles all the way, if it were self similar, than we would only need one model, and that model would apply at every level. But, it's not turtles all the way. Take the brain. It consists of *molecules* that form synapses. The synapses in turn form neurons. The neurons combine to form networks. The networks combine to form elaborate maps that can be studied with brain imaging. These maps exists on a scale below that of functional systems – such as the limbic system or the cerebellum. Those systems combine to form the brain itself. Each level exists at a different scale and produces distinct functionalities (see figure 4).

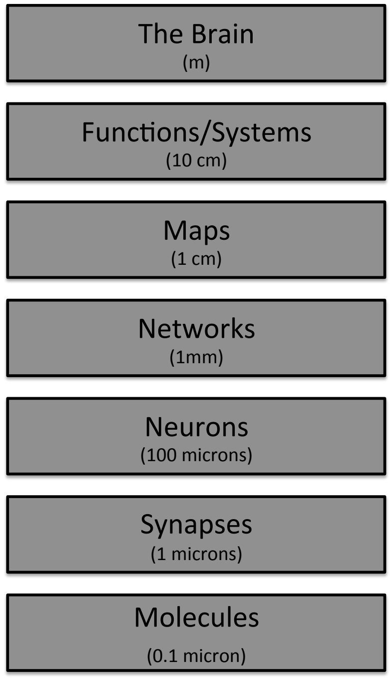


Figure 4.3: Levels in the Brain - modification of Figure 1 in Churchland and Sejnowski (1992)

Within the brain, interactions at one level produce phenomena at higher levels. Thus, any "model of the brain" must consist of many models. The models used to characterize the robustness of neuronal networks bear little resemblance to the molecular biology models used to explain brain cell function. Similarly, the network models do not include variables related to personality characteristics or emotion that can be found in psychological models. The brain is not turtles all the way down. It differs at each level, and at each level we need different models.

Similarly, financial systems consists of people, banks, regional and national financial institutions and markets, and the global financial system. Each level requires distinct models. A model of an individual investor in Muncie, Indiana won't be of much use for explaining the US Stock market of international exchange rates.

## Reason #7 Phenomena Differ By Level

*Many systems consists of multiple levels. Each level may require its own model*

Even if our single model works today or worked yesterday, it need not work tomorrow. This is the seventh reason for using multiple models: *the social world is non stationary.* Physical laws remain unchanged. We have no reason to worry about force not equalling mass times acceleration next week Wednesday. Engineers don't need to go back and check if the tensile strength of steel has changed over the past decade. But social scientists confront a world that changes day to day and week to week. How much people spend on clothing or what percentage of people vote or why people vote can change. For several elections during the latter part of the 20th century, one could predict the likely winner of United States Presidential Elections using only economic data. Those same models did not predict the 1992 election (Fair 2012). To guard against a model no longer working, we must apply many models, we must investigate reality through a variety of lenses.

## Reason #8 The Social World is Not Stationary

*The social world changes. Relationships and causal forces that held at one moment in time need not hold the next.*

Our last reason to use many models combines two of our reason to model: *to explore possible designs.* Design, as already mentioned, often include new features. To be relevant a feature must have some effect. That effect cannot be known. Evaluating two designs may require building two models.

> ## Reason #9 To Explore Alternative Designs
>
> *To explore alternative designs, we need many models.*

Think back to the economists trying to design an auction for the FCC. They built several models and analyzed the implications of each.

## Many Models and Prediction

To provide a more formal basis for the value of many models, we introduce the *Diversity Prediction Theorem*.[28] Imagine that we have a collection of models that each make a prediction or forecast of some future or unknown value.[3]

After a predicted value becomes known, we can assign an *error* to each mode equal to the difference between the value of the outcome and the model's prediction. Statisticians square the errors to make all errors positive and to punish larger mistakes.

Given a collection of models, we can sum up the squared errors for each model and divide by the number of models to arrive at the *average squared error*. We can then compute the squared error for the average of the collection of models by squaring the difference between the true value and the mean prediction of the various models. Call this the *many model*

[3] Some people distinguish between *forecasts* and *predictions* with the former being an extrapolation from past data or based on some existing state of the world and the latter being a judgment of some future novel event. Hence people refer to weather forecasts and technological predictions.

*squared error.*

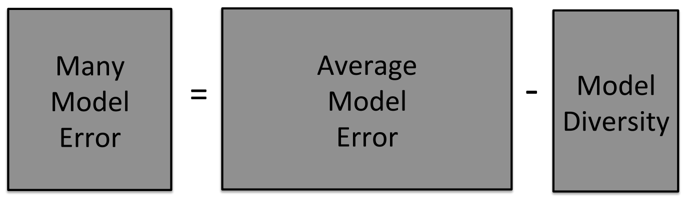


Figure 4.4: Graphical Representation of the Diversity Prediction Theorem

Last, we can define the *model diversity* of the models' predictions as the average squared difference between the models' predictions and the mean prediction of the models. The diversity captures the variation in the predictions. For example, if the mean prediction of seven weather forecasting models for the temperature on March 30th in Cleveland, Ohio equals sixty degrees, the diversity equals the average of the squared differences from each model's prediction to sixty.

We have defined three statistics relating to the model predictions: the *averaged squared error*, the *many model squared error*, and the *model diversity.* The *Diversity Prediction Theorem* combines all three in a single equation stating that the many model squared error equals the average squared error minus the model diversity of the models.

The *Diversity Prediction Theorem* is a mathematical identity. Given any collection of predictions and any true value, the equation holds. The equation holds if there exists two hundred models. It holds if there exists only two models.

## The Diversity Prediction Theorem

Given $N$ predictive models, let $\text{Model}_i$ denote the prediction of model $i$, *Many Models* denote the mean prediction of the $N$ models, and let *Truth* equal the true value of the outcome being predicted. The following equality always holds:

$$(\text{Many Models - Truth})^2 = \sum_{i=1}^{N} \frac{(\text{Model}_i - \text{Truth})^2}{N} - \sum_{i=1}^{N} \frac{(\text{Model}_i - \text{Many Models})^2}{N}$$

An example helps to reveal the intuition that underlies the result. Suppose we have two models that predict the number of Academy Awards (Oscars) that a film will win One model predicts two Oscars, and the other predicts eight. In this case, the mean of the two models' predictions equals five. Let's assume that the film wins four Oscars. The squared error for the average of the two models, the *many model error*, equals one. The first model's prediction is off by two, so its squared error equals four. The second model's prediction is off by four, for a squared error of sixteen. The *average squared error* of the models equal ten. Finally, each model differs from the average prediction by three, so the average squared difference from the mean prediction, the *diversity*, equals nine. We can write the *Diversity Prediction Theorem* as follows: *one* (the many model error) = *ten* (the average squared error) minus *nine* (the diversity).

Alternatively, suppose that both models predicted two Oscars. The average prediction of the models will again be two, so the squared error for the collective will be nine. Nine is also the averaged squared error for the two models. Finally, the models have no diversity. In this case, the *Diversity Prediction Theorem* would be written as follows: *nine* (the many model error) = *nine* (the average error) minus *zero* (the model diversity).

The logic for why averaging multiple diverse models produces more accurate predictions should now be clearer. If one model predicts a value that is too high and another model might predict a value that's too low, then the errors partly cancel, so the average is better. If the second model also predicts a value that's too high, then the error of the average of the two high predictions won't be worse than the average of the two high predictions.

Many model thinking advocates using multiple, distinct models. The *Diversity Prediction Theorem* relies on the models making distinct predictions. Different models make different predictions because they simplify the world in distinctly. Suppose that we want to predict the amount of leather in a cowhide.

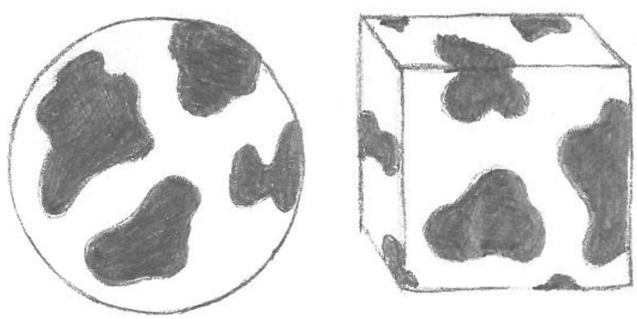

Figure 4.5: A Spherical and Cubic Cow

To make that estimate, we need to know the surface area of a cow. That's not a formula to be found in a geometry book so we need a model (or two). The book *Consider the Spherical Cow* provides one approach: calculate the surface area for a spherical cow, and use that as an approximation (Harte 1988). If our spherical cow had a radius of two feet, then its surface area would equal $16\pi$ or approximately fifty square feet.

The spherical cow is one model. We might also model a cow as a cube. We might then approximate the cow as four feet long by two feet wide by three feet high. This gives a surface area equal to fifty two square feet. The spherical cow and the cubic cow models produce similar but different predictions. Each model is wrong, but they're wrong in different ways because of the $\pi term$. The spherical and cubic cow models provides a nice example of how different models because they make different assumptions lead to dusting predictions.

Different predictions are a good thing. The *Diversity Prediction Theorem* implies that if diversity is positive (which it will be if any two models make different predictions), then the many model error must be strictly less than the average squared error. The average prediction of a collection of predictive models will be more accurate than a randomly selected model. *Many models outperform the average model.*

## Many Models Outperform the Average Model

Given any $N$ predictive models in which at least two models make different predictions, the following inequality holds:

$$(\text{Many Models - Truth})^2 < \sum_{i=1}^{N} \frac{(\text{Model}_i - \text{Truth})^2}{N}$$

Where *Many Models* denotes the mean prediction of all the models, $\text{Model}_i$ denotes the prediction of the model $i$, and *Truth* equals the true outcome value.

We do not have to test *Diversity Prediction Theorem* empirically. It is a mathematical identity. It always holds.

## The Wisdom of Crowds (of Models)

The *Diversity Prediction Theorem* implies that if a group or crowd consists of people applying diverse and accurate models, then the crowd will make accurate predictions, a phenomenon sometimes referred to as the *wisdom of crowds* (Suroweicki 2006). Examples of wise crowds abound. At a livestock exhibition in the West of England in 1906 seven hundred and eighty-seven people guessed the weight of steer. Their average guess was less than a pound from the steer's actual weight. In 2014, I asked forty-six students in a class at INSEAD to guess the number of cars per one thousand people in Latvia. Their average guess was three hundred and eighteen point six. The actual number was three hundred and nineteen.

The wisdom of crowds requires a large average error. If not, then the crowd is wise because it consists of only accurate predictors. In each of the two cases described, the average squared errors were large. In the contest to guess the weight of the steer, the average squared error was just less than three thousand. For the INSEAD students, the averaged squared error exceeded two hundred thousand. If the crowd is accurate (many model error is small) and the individuals who make the predictions are not (average model error is high), then model diversity must be high for the equation in the *Diversity Prediction Theorem* to hold (see figure 4.6).

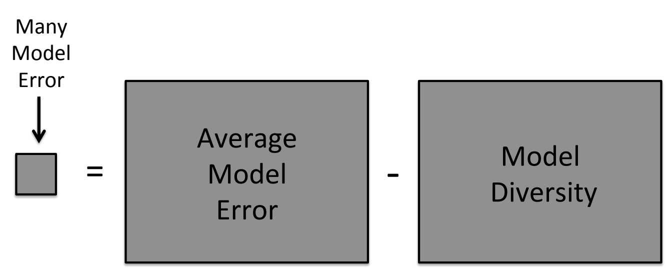


Figure 4.6: The Wisdom of Crowds: Diversity Prediction Theorem

It follows that a wise crowd implies diverse predictors (be they models or people). The logic does not apply in the other direction. We could have a many model prediction that's far off the mark even with diverse predictions. For this to occur, the average model error must also be large, and finally, the diversity cannot be large (at least relative to the average error). If diversity is relatively the same size as the average model error, then the many model prediction would be crowd. Thus, many models can still be inaccurate if the models are more inaccurate than they are different.[29]

## Diverse Representations and Diverse Logics

The *Diversity Prediction Theorem* gives us a core intuition as to why we want multiple models for prediction. By using multiple models, we get predictive diversity. And predictive diversity contributes to collective accuracy. What we have called predictive diversity can result from *diverse representations* or *diverse logics*. We can clarify the difference using *categorical models*.

Categorical models partition a set of entities into groups or bins. A formal *categorical model* consists of nothing more than taking a set of items, placing them in categories of similar items. The idea to use categories to makes sense of the world can be traced back to Aristotle. In *The Categories*, he creates ten categories which partition the world. His categories include *substance*: people or wood; *quantity*: two feet; *where*: in the kitchen; and *being in a position*: lying down.

We all use categorical models. We categorize restaurants by ethnicity: Italian, French, Turkish, or Korean We categorize people by professions: doctor, lawyer, teacher, or bricklayer. And, we classify countries by continent: Asia, Africa, Europe, North America, South America, or Australia. We then make predictions given those categories. "A Mexican restaurant, I bet they have good tacos."

We can construct a formal *predictive categorical model* as follows. We assume a set of objects. These could be cars, houses, people, countries or types of dessert. Associated with each object is a *value*. This could be the price of a car or the number of calories in a dessert.

In a*predictive categorical model*the objects are partitioned into categories in such a way that the predicted value for two objects in the same category is the same.

> ## Predictive Categorical Models
>
> *Given a set of $N$ objects, a* **predictive categorical model** *partitions the objects into categories $S_1, S_2, ...S_n$. For each category $S_i$, the model assigns a* **predicted value**, $\text{Pred}(S_i)$.

The accuracy of a categorical model depends on the effectiveness of the categorization (are objects with similar values placed in the same category) and on the accuracy of predictions within each category.

We are now in a position to identify two types of model diversity within categorical models. First, two models could rely on distinct categorizations. One model might categorize automobiles by year, another might categorize them by manufacturer. Distinct categorizations are an example of *diverse representations*. Second, two models could rely on the same categories but make distinct predictions. If so, provided that the models rely on the same data they must be based on *diverse logic*.[30] What we next want to see is how these two types of diversity can both produce more accurate predictions. To accomplish that we present a brief aside on the accuracy of categorical models that extends our earlier notion of average squared error.

### Accuracy of Categorical Models

To measure the accuracy of a *predictive categorical model*, we first calculate the *total variation* in the objects' values. The total variation equals the sum of the squared differences from the objects' values to the mean value. Imagine the total variation in the data as a box. A *predictive categorical model* explains some percentage of that variation. (If it doesn't, if the model increases the variation, you should get rid of or emend the model as it is less than useless!). The remainder of the box, the variation that's unexplained - the *total error*, has two components: the *categorization loss*: the variation that exists within the categorization and the *prediction error*: the squared difference between the mean value in each category and the predicted value for that category.[4]

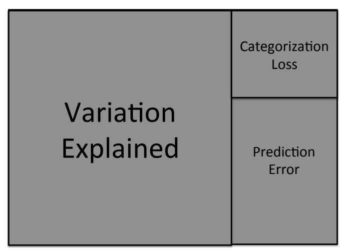


Figure 4.7: The Components of Total Variation for Categorical Models

We will refer this decomposition of error as the *Categorization Prediction Theorem.*

## Categorization Prediction Theorem

*Given a predictive model based on categorization,* model error *equals the sum of the categorization loss and the predictive error.*

*Model Error = Categorization Loss + Predictive Error*

An example clarifies the mathematics. Assume that we have a categorical model of

[4]If using a model to explain existing data and not to predict, the prediction error equals zero and all of the unexplained variation will be due to categorization loss.

housing prices in the Berry Hill neighborhood of Nashville, a leafy residential area populated by craftsman bungalows. Our model creates categories based on whether a house has been turned into a recording studio.

Consider the four bungalows, denoted by $A$,$B$,$C$, and $D$ in figure 4.8, along with their market values. Create two categories based on whether or not the bungalow contains a recording studio (denoted by a circle representing a CD or LP above the door). Bungalows $A$ and $B$ do not contain recording studios, so they belong to one category, while bungalows $C$ and $D$ do contain studios, so they belong to the second category.

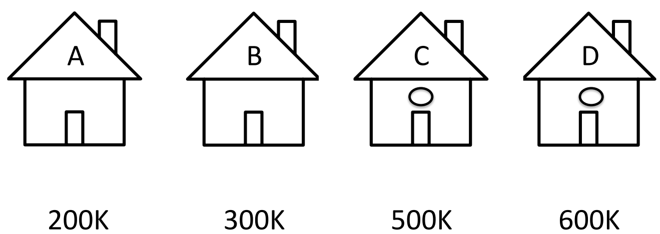


Figure 4.8: Four Bungalows and Their Market Values

We first compute the *total variation* in the prices of the bungalows. This tells us how our model could could explain. The **total variation** equals the sum of the squared differences from each value to the mean. The **mean** value of all four bungalows equals. \$400K, so total variation equals one hundred thousand.[5]

To calculate the *categorization loss*, we assume that we knew the true mean within each category. The means are \$250K for the first category and \$550K for the second category – the bungalows that have been turned into recording studios. The categories have different means. This implies that the categorization explains some of the variation. It does not

[5]Total Variation $= (200-400)^2 + (300-400)^2 + (500-400)^2 + (600-400)^2 = 100,000$

explain all of the variation. The categorization lumps together houses of different values. That remaining variation equals the **categorization loss**. In this example, categorization loss equals five thousand for each category.[6] The *categorization loss* equals the sum of these two numbers, or ten thousand, an amount equal to one-tenth of the *total variation.*. That's good. By creating two categories based on whether or not a bungalow contains a recording studio, we've explained ninety percent of the variation in house prices.

The categorization explains so much of the variation because we made the best possible prediction for each category. In practice, we wouldn't know those values. We would make predictions and those would be inaccurate. The differences between the predictions in each category and in the true values for the category equal the *predictive error.* For example, suppose that we predict \$300K for bungalows $A$ and $B$ and \$600K for bungalow $C$ and $D$. The predictive error equals the squared differences between the predictions for each category and the true mean. In the first category, the best prediction was \$250K but we predicted \$300K. For the second category, the best prediction was \$550K but we predicted \$600k. Therefore, predictive error equals ten thousand.[7]

Next, we can compute the *model error*, the squared differences between our predictions and the actual values. This equals \$20,000.[8] Notice that the *model error* equals the sum of *categorization loss* and *predictive error.* Figure 4.9 shows the decomposition of total variation into three parts: the variation explained, the categorization loss, and the predictive error.

To measure categorical model accuracy, we compute the *percentage of total variation explained* by the model. Statisticians call this *R-squared.* Using this measure, our model has an R-squared of 0.8.

[6] Categorization Loss A & B = $(200 - 250)^2 + (300 - 350)^2 = 5,000$
Categorization Loss C & D = $(500 - 550)^2 + (600 - 550)^2 = 5,000$

[7] Predictive Error A & B = $(300 - 250)^2 + (300 - 250)^2 = 5,000$
Predictive Error C & D = $(600 - 550)^2 + (600 - 550)^2 = 5,000$

[8] Model Error = $(200 - 300)^2 + (300 - 300)^2 + (500 - 600)^2 + (600 - 600)^2 = 20,000$

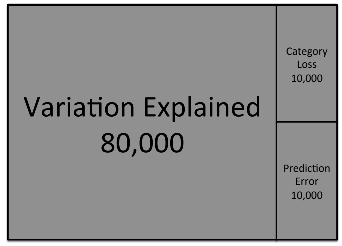


Figure 4.9: The Components of Total Variation For Housing Model

## $R^2$: The Percentage of Variance Explained

*Given a predictive model applied to a set of data, the* **R-Squared**,$R^2$, *equals the percentage of the total variation in the data explained by the model.*

$$R^2 = \frac{\textit{Variation Explained by Model}}{\textit{Total Variation in Data}}$$

We can now return to our discussion of the two possible types of diversity: representational and logical. Let's start with *logical diversity* as its impact can be seen rather easily. Suppose that we had a second predictive model that used the same categorization as our original model but that it relied on a different logical argument to make estimates within each category. We then know from the *Diversity Prediction Theorem* that within each category, that the average of the two models would have a lower squared error than the average squared error of the two models considered individually. Thus, having distinct logics given a common categorization will reduce errors.

Let's now turn to *representational diversity.* The argument here is more complicated and requires unpacking both *categorization loss* and *prediction error*. Suppose that we have multiple models, and each categorizes the objects differently. By using all of the categories, we can create a *refinement* of our original categorization[9] in effect, we are taking an intersection of the original categories.

To see how to create a refinement, return to our example of the bungalows. Our original model categorized based on whether the bungalow was a house or a studio. A second model might categorize bungalows by whether they are on a busy street or a quiet street, and a third model might characterize them based on whether they have new bathrooms or old bathrooms. Each of our three models divides the houses into two categories as shown in figure 4.10.

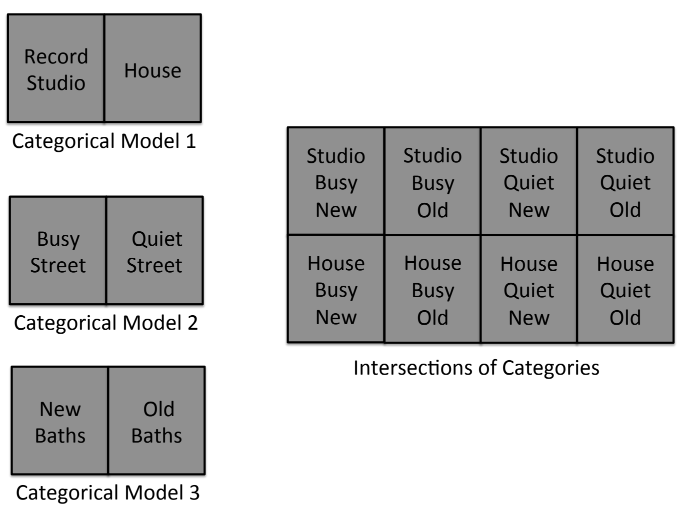


Figure 4.10: Three Categorizations and Their Intersection

The set of three models divides the houses into eight categories. The eight categories are produced by intersecting the three original categories. [10]. As before, we can calculate the categorization loss, and because we have *refined* the original categories, we necessarily lower category loss. Consider the new category denoted *Studio, Busy, New*.

[9]One categorization *refines* another if every category in the first is a subset of some category in the second.
[10]Earlier, we considered only four bungalows. Here, we assume we have a larger set of houses.

Suppose that the mean value for houses in that category equals \$500K. In computing the *categorization loss* for our original categorization based only on whether or not the house contained a studio, we relied on a mean of \$600K for all houses with a studio. For this subset of that category, the mean value equals only \$500K, so the *categorization loss* within this subset will be higher for the original categorization than for the refined categorization. Applying this same logic to all of the new subcategories and each of the three models reveals that *the categorization loss of the refinement model cannot exceed the categorization loss of any of the constituent models*. Therefore, diverse representations reduce categorization loss.

On the intersection of the categories, we might take the average prediction of the models. Applying the *Diversity Prediction Theorem*, we know that the *predictive error* of the average of the models must be less than or equal to the average *predictive error* of the models themselves. We therefore know that the average of a collection of categorical models predict more accurately than a randomly selected model from the collection. Many models are better than one.

## Robustness

We took a deep dive into prediction because we could derive algebraic expressions that demonstrate the value of multiple models. The other benefits of many models cannot be as cleanly articulated. They nonetheless action When we look back at the nine reasons, we see many of them imply more robust understanding. That's certainly true of the power of many models to *to identify distinct logics*,*to identify the boundaries of the possible*, *to prevent overfitting*, *to capture phenomena at multiple levels*, *to cope with nonstationarity*, *to understand a non stationary world* and *to explore new possibilities.*

To make a more direct link between these advantages of many models and robust understanding, we return to the question of financial stability. The 1929 Black Tuesday stock market crash led to the decade long Great Depression. The 2008 Home Mortgage Crises had a less extreme effect on the global economy but produced substantial real losses in wealth, income, and well being.

Governments and regulators would like to prevent these crashes. That requires understanding why crashes occur, how they can be prevented, and the will to intervene. Models improve our ability to carry out the first two of these tasks. The better the models, the more confidence an actor might have in action. However, to quote John F Kennedy (1956) "they cannot supply courage itself. For this each man must look into his own soul."

In seeking to prevent future crashes, governments and central banks turn to models, and to the extent possible they link those models to data. The models take many forms as they must cover many levels: individuals, banks, as well as entire financial infrastructure. Let's start with individual investors. Policy makers use rational choice models as a benchmark. They complement these with more psychologically based models as well as models that assume people follow simple strategies that are copied from other people or discovered through experimentation. This type of model might be implemented on a computer using *agent based modeling* (see Miller and Page 2007).

Policy makers also pay heed to models of traders. High frequency trading can produce large fluctuations in prices through the interactions of contingent rules. In formulating trading restrictions and oversight provision, policy makers look to models to help with design (Wellman 2013).

To determining the solvency of banks, policy makers use a variety of models from economics and finance. One such model, attributed to Merton (1969) and Markowitz (1952) provides a way to assign a risk to financial portfolios. The model assumes a known probability distribution over future market risk. That can be estimated but rarely known.

The need for many models becomes clear when one looks at how poorly many of the individual models perform. Regulators would like to know which banks are likely to fail. The (risk based) ratios of total capital to to assets on hand would seem a good predictor. Banks with many assets on hand (and lower ratios) would be thought to be more solvent. The data show less support than expected. Charts of the capital to asset ratios of failing and surviving banks shows little to no relationship (Haldene 2012).

The failure of any one measure explains why regulators use many models and many measures instead of only one. The number of measures has even increased. The Basel Accord of 1998, the first international financial regulatory agreement for prudent management of financial systems, defined five risk measures. The most recent agreement, Basel III, obliges banks to measure the risk of large individual loans. The thinking is that greater granularity leads to more accurate assessments.

The standard model for estimating the stability of the entire systems has been *stress tests*. Stress tests rely on multiple models to ascertain what might occur as the result of a large change in a single risk factor, say a change in interest rates or prices. They can also test solvency under a *scenario*: a change in multiple factors simultaneously based on some plausible event (Blaschke et al 2001).

To gauge the solvency of the entire system, many analysts now also use network models. Network models take many forms. Some include connections between the financial sector and households. Some include heterogeneity in firm sizes and portfolio allocations. Others are much simpler. These constellation of models reveals contradictory effects of greater connectedness. Connections allow risk to be diversified. At the same time, they provide routes for failure to spread across the system (Glasserman and Young 2015). The provide Munger's lattice of models on which to make the decision.

Regulators even rely on models of themselves. Using machine learning techniques, they can gauge the sentiment of their own conversations by feeding transcripts of their meetings into a computer learning algorithm to measure the sentiment and content of their discussions (Haldene 2015).

The reliance on many models begs the question of why someone wouldn't combine all of the models into a single grand model. The primary reason relates to two of the reasons we model: to understand the logic and to explain phenomena. Modelers, like map makers leave out some details and make others prominent. Both do so in order to produce understanding. Modelers justify simple models by referring to a Borges' (1975) story of the mapmakers who drew map of the same size as the country it represented. The elaborate map was of little value (and presumably difficult to fold).

Understandability is not the only reason. As already mentioned, with a single, elaborate model, we run the risk of overfitting. Overfitting can mean inaccurate predictions out of sample. In addition to overfitting, large models are also prone to under fitting. As a rule of thumb, for each each parameter in a regression, we need ten to twenty data points (Harrell 2001). A model with twenty variables would require four hundred data points. That's a minimum. In practice, we man need even more data. DeMiguel et al (2009) show that to make sufficiently accurate estimates of risks in order for the aforementioned Merton and Markowitz method of allocating risk to outperform a portfolio that consists of an equal amount of each of twenty-five assets would require two hundred and fifty years of data. A portfolio of fifty assets would require five hundred years of data.

We have now learned nine reasons for applying many models and, using models, shown how many models improve prediction, discussed the relationship between many models and robust understanding, and discussed why many simple models may be better than a single, elaborate model.

Two takeaways from this chapter are that *more are better* and *different is better*. More is better because all else equal, we'd rather have more models of a phenomenon than fewer.

Mmany models show us multiple logics and make us better able to predict, explain, act, design, and explore. Different is better, because distinct models produce unique categorizations and logics. Having multiple, diverse ways of thinking enhances robustness in understanding.