

We turn now to the role of diversity in helping collections of people make accurate predictions. We might think that a collection of moderately informed and knowledgeable individuals would not be able to make accurate predictions, but that's not necessarily true. Substantial evidence suggests that diverse collections of people often predict rather well. If not, then how could we ever expect markets and democracies to function? After all, what is a stock price or an election outcome but a prediction? What is a choice in an election or a referendum but a prediction of what a politician will do or what forces the referendum will put into play? Sure, prices and outcomes don't have to be perfectly accurate for markets and democracies to work, but they can't be consistently far off the mark. Otherwise, we'd have far more stock market crashes and wrestler governors. Not that those don't exist. They do. But they're not so widespread that we've abandoned either markets or democracies. Some even think that this market based democracies represent the end of history.

Stock market prices and election outcomes are predictions by huge numbers of people. In addition to these society-level predictions, we also make lots of predictions in small and moderately sized groups – in juries, management teams, boards of directors, bargaining teams, and department faculties. Many, but not all of the real world examples that we consider in this chapter and the next describe a reasonable large number of people, a crowd, making a prediction. Somehow, a thousand people making an accurate collective prediction seems more amazing than three people doing so. But the logic, in the case of three or three thousand remains the same; it relies on equal parts individual accuracy and collective diversity. For that reason, our not so real world examples will include only a few people.

In this chapter, we consider several types of information aggregation models. These will not be based on our frameworks. Instead, these are representative of the models typically constructed by social scientists. We get to why we're doing this in a few pages. First, we need to understand what information aggregation is and how it differs from predictive model aggregation. In information aggregation, people get "signals" about the outcome. We can think of these signals as information. We can also think of them as predictions but we see why that involves sloppy thinking and hand waving - two things we're trying to avoid. We take up predictive model aggregation in the next chapter.

In both this chapter and the next, we emphasize individual diversity and

the central role it plays in collective accuracy. The only way a collective prediction can be accurate if the individual predictors are not diverse is if all of the individual predictors are accurate. But then we have a result that says good individual predictors make accurate collective predictions. That result is not at all surprising. Yet, we have many examples in which individuals are not individually accurate but the collection of them is. That mystery is what we seek to explain.

Surowiecki proposes three necessary conditions for a collection of people to make accurate predictions. These are that people have diverse predictive models, that people are independent – they are not allowed to influence on another, and that the prediction process be decentralized – that people not communicate with one another. All three of these conditions require a diversity of predictive models. If people are not independent, they are not likely to have diverse predictive models. And if the process that aggregates predictions is not decentralized, then participants likely share predictive models thereby lessening the diversity of the models used. Thus, in a way, the third condition can be seen as implied by the second.

To provide some grist for our cognitive mills, we start with some examples in which collections of people make accurate predictions. Some of these are anecdotes, but most present systematic evidence that collections of people can make accurate predictions. This distinction matters. We might put forth a theory that Men are from Venus and Women are from Mars. If this theory accurately predicted correctly the behavior of only ten percent of men and women, it would be consistent with millions of anecdotes. Hence, we need systematic evidence. After looking at the evidence, we ask whether the models from the information aggregation models from social science prove up to the task of explaining these examples. The answer is no!

## Examples of Wise Crowds

Many of the specific examples in which collections of people predict correctly seem almost unbelievable. We relate some from Surowiecki's *The Wisdom of Crowds* here, but see his book for even more. These examples are not laboratory experiments with ten subjects; they come from Las Vegas, county fairs, and game shows. Evidence also comes from the stock market. Stock market prices encapsulate the predictions of a crowd of people about future

dividend streams. These predictions can be freakishly accurate. In 1986, the stock price of Morton Thiokol's fell soon following the 1986 Space Shuttle Challenger disaster. Only later was it discovered that the O-rings that Morton Thiokol manufactured were primarily to blame. We can dismiss this accurate pricing as pure luck, but in repeated settings markets have proven to be quite capable predictors.<sup>1</sup> The futures for the price of orange juice determined by market traders has often proven more accurate than weather forecasts of the likelihood of freezes.<sup>2</sup> Obviously market prices are not perfect predictors. Market bubbles and crashes occur with some regularity. But, as Surowiecki notes, markets create bubbles partly because they have no known end date, which creates an incentive for people who trade on trends to ride an increasing wave of prices.

Predictive markets other than the stock market include the Hollywood Stock Exchange (HSX), the Iowa Electronic Markets, and Tradesports.com.<sup>3</sup> On HSX, people buy stocks that pay dividends depending on ticket sales for movies. People also buy bonds on future ticket sales of movies for certain stars. Angelie Jolie, for instance, might sell for more than Ben Stern. A test showed that HSX performed as well as a leading expert in predicting the revenues for fifty movies released between March and September of 2000. HSX's predictions were off by an average of 31%, while the predictions of a leading expert, Brandon Gray of Box Office Mojo, erred by 27% on those same fifty movies.<sup>4</sup>

Later, we present evidence showing the accuracy of the Iowa Electronic Markets and the sports betting lines. This evidence is important. We could always find two or three cases where a crowd of people miraculously predicted an outcome correctly, but that does not prove crowds do so on average any more than my father's hole-in-one at Mullenhurst Golf Course in Yankee Springs, Michigan, in the mid 1990s would be proof that he's on a par, or

---

<sup>1</sup>Keep in mind if the predictive task is too hard, no one will do much better than a dartboard. See Tetlock, Philip (2005) *Expert Political Judgment: How Good is it? How Can we Know?* Princeton University Press

<sup>2</sup>Wolfers, Justin and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives* 18(2):107-26. Roll, R. 1984. "Orange Juice and Weather." *American Economic Review*. 74(5):861-80.

<sup>3</sup>HSX and IEM use real money as do Tradesports.com and Betfair.com. Other prediction markets like Ideosphere.com, use virtual money.

<sup>4</sup>See Pennock, Lawrence, Giles, and Nielson, in *Science* 291: 987-988, February 9 2001 (Letters).

under par as the case may be, with Tiger Woods. He's not.

Systematic evidence convinces social scientists, but that doesn't mean anecdotes don't have their place. Anecdotes are more captivating – both times I've asked crowds of students to guess my weight, their average guess has been within a pound . Within a pound! When I weighed 194, they guessed 193 on average. The next year, they predicted me a leaner 186. I weighed 185. Anecdotes such as this provide us with a point of analytic entry and intellectual motivation. The anecdote that leads off Surowiecki's book plays both roles. Here's the short version. In 1906, Francis Galton analyzed predictions of the weight of a young steer by attendees of the West of England Fat Stock and Poultry Exhibition. The average of 787 contestants' estimates of the weight was 1,197 pounds. The actual weight of the steer was 1,198 pounds. There was no consensus about how tasty the steer was. Galton wouldn't be happy with Surowiecki's analysis because he advocated the median as the predictor not the mean. Some crazy person can too easily manipulate the mean.

Despite Galton's objection, we keep Surowiecki's mean based analysis in mind as we walk through the information aggregation models of how collections of people make accurate predictions. Social scientists have devoted substantial effort to modeling how collections of people predict in economic and political settings. Those models highlight the information that people possess and the incentives they have to truthfully reveal it. The interesting thing is that none of those models explain the steer, but the model that we present in the next chapter - based on our predictive model framework - does. We might add that lest we get too carried away with this last example, guessing the weight of a steer is not that difficult. As a former amateur cattleman (my wife and I owned nine steers for a short time), I can say from experience that a person can estimate the weight of a steer within a hundred pounds without mastery of Galois Theory or Differential Equations. It's like guessing the weight of really really big people. Galton's was not an experiment in which random people on the street were asked to guess the viscosity of a mixture of Jell-o and motor oil in some obscure unit of measure. People know what a pound is. People know that steers weigh more than men but less than elephants. Nevertheless, as Thomas Schelling once wrote if you are in the mood to be amazed...

## Information Aggregation Models

We now look at standard social science models of information aggregation. The distinction between information aggregation and predictive model aggregation is subtle and often blurred. If we ask people to name the capital of Oregon, then we are asking them for information, for facts (not a prediction). If we ask those same people to predict the outcome of the next U.S. Senate race in Oregon, we are asking them to apply models to their information to predict the future.

In information aggregation models, people get signals about the answer. The various types of information aggregation models make one of three assumptions about what people know. These end up being assumptions about the signals, but we describe them here in terms of the information they provide. The models assume that people either know the answer with some probability, know pieces of the answer, or get blurry signals of the answer. The first assumption states that some people know the answer and others do not. The process of aggregation allows the people who know the answer to reveal it. The second assumption states that people know parts of the answer and that those parts can be aggregated to reveal the answer. As Aristotle puts it in what some call the “summation argument” – what Jeremy Waldron more sublimely refers to as “the doctrine of the wisdom of the multitude”:<sup>5</sup>

*For each individual among the many has a share of excellence and practical wisdom, and when they meet together, just as they become in a manner one man, who has many feet, and hands, and senses, so too with regard to their character and thought. Hence the many are better judges than a single man of music and poetry, for some understand one part, and some another, and among them they understand the whole.*

-Aristotle, *Politics*, Book 3, Chapter 11.

Aristotle’s interests run to poetry and music not the price of Microsoft stock or the capabilities of politician, but his insight applies in those cases as well. He proposes that because each person knows a part of the solution, the collection of people can know the whole. This whole as sum of the parts logic may explain some wise crowds but it fails to describe people making

---

<sup>5</sup>Waldron, Jeremy. 1995. “The Wisdom of the Multitude: Some Reflections on Book 3, Chapter 11 of Aristotle’s *Politics*.” *Political Theory* 23(4):563-84.

predictions based on models.

The third common assumption, that people see a blurry picture of reality, captures situations in which something blocks or distorts are viewed. With some bending and stretching this can be framed as an assumption about predictions. We can think of the distortions as errors in predictive models, but this assumes that the noise is an addition to the true state, to the right answer. That's not likely to be true of actual predictions. That they're perfectly accurate but they just have extra stuff added in (or left out). Thus, this last assumption only captures situations in which people do not see the outcome or the event clearly.<sup>6</sup>

In all of the real world examples above, people predict a future or unknown event: a future price or a current weight. People do not just recall bits of information, though sometimes that does happen – we consider a case involving a game show in a few pages – but it is the exception and not the rule in market and political settings. Some defend the social science models of information aggregation by claiming that their abstraction – the assumption that people get signals – allows them to be recast as models of people who make predictions. That can be done. But, as we shall see, doing so implicitly builds in diverse predictive models. Otherwise, how could these people receive different signals?

In what follows we consider some of the standard assumptions about signals using toy models of information aggregation. Yes, these models reveal the limits of the information aggregation approach, but that's not why we study them. We're not constructing straw men to knock them down. To the contrary, we want to pull them off their sticks and walk them along a yellow brick road toward relevance. By looking at these simple models, we can begin to understand how to add information. Adding one plus one is easy. Adding information in the form of signals is a bit harder. Adding predictive models is harder still and almost impossible without first learning to add signals. Thus, we first look at models, that in the end, we find lacking. But by doing so, we have the expertise - the tools - necessary to understand how information aggregates.

---

<sup>6</sup>For that reason, Lu Hong and I refer to these as *generated signals*. Hong, Lu, and Scott E. Page. n.d. "Interpreted and Generated Signals." Unpublished.

## A Million Tiny Pieces

In our first toy information aggregation model, each person in a crowd has partial information about the correct answer to a question. They aggregate their partial information by voting. We describe this model using yet another example from *The Wisdom of Crowds*, one that involves the game show *Who Wants to Be a Millionaire*. On the show, a contestant must choose from among four possible answers. If the contestant is correct several times in a row, she can win one million dollars. On a particular question, if the contestant is stumped, she can use “lifelines.” One lifeline is to call a friend. This friend is presumably an expert, not a buddy from junior high. Another lifeline allows the contestant to poll the audience. The audience at a game show consists of few editors of the Encyclopedia Britannica or University of California at Berkeley faculty.

Data from the show reveals that the friends (who we can assume were called because they were ostensibly experts) proved correct only two-thirds of the time. We can assume that the experts only get the call on hard questions. No one phones a friend to learn the number of people in the Jackson Five. To quote the 1980s rock icon Meatloaf: on difficult questions, two out of three ain’t bad. And yet this level of accuracy pales in comparison to the accuracy of the audience’s prediction. When polled, the audience predicted correctly nine times out of ten. Nine times out of ten is far better than ain’t bad. It is amazing, astounding, and some might even say magical.

Though some think that the crowd’s accuracy emerges from some deep and mysterious process, we should know better. There’s no mystery here. Mistakes cancel one another out, and correct answers, like cream, rise to the surface. To explain how that occurs, we construct a model (big surprise).

Let’s suppose that the question concerns the four members of The Monkees, a made-for-television band from the 70s.

**Question** *Which person from the following list was not a member of the Monkees?*

Identify the Non Monkee

- |                     |
|---------------------|
| (A) Peter Tork      |
| (B) Davy Jones      |
| (C) Roger Noll      |
| (D) Michael Nesmith |

We can assume that the Monkees are not familiar to everyone (and thank goodness for that!). Those people who can identify the three Monkees on the list correctly identify Roger Noll as not being a Monkee. They'd vote for him. (Few of these people could probably further identify Roger as a Stanford Economist and all around good guy.) We can next assume that those people who do not know any of the Monkees choose randomly from among the four names (more on that later). Now things get a little trickier. For those people who can identify two of the Monkees, we assume they randomize between the two names that they do not know. And for those people who know only one of the Monkees, we assume they randomize between the other three names.

We next imagine a hypothetical crowd of one hundred people, seven of whom know all three Monkees listed, ten of whom know two of the Monkees, fifteen of whom know only one of the Monkees, and sixty eight of whom know none of the Monkees. On average, the individuals in this crowd are not well informed. Less than ten percent know the answer and more than two thirds have no clue.

Let's now have these people vote. Roger Noll gets seven votes from the seven people who know the answer. And, on average, he also gets five votes from ten people who know two Monkees on the list because these ten people randomize between just two names. Roger and one another person, who we can assume is random. Roger also gets five votes from the fifteen people who know only one Monkee on the list and randomize among three names. Finally, he receives (again, on average) seventeen, or one fourth, of the sixty-eight votes from people who have no clue as to the correct answer. If we sum these votes, we get that, on average, Roger Noll gets thirty-four votes.<sup>7</sup>

If, as assumed, people are randomizing, each of the other three names should split the remaining sixty-six votes and get about twenty-two votes.

---

<sup>7</sup>The formal calculation goes as follows:

$$7 * (100\%) + 10 * (50\%) + .5 * (33.3\%) + 68 * (25\%) = 34$$

Thus, Roger Noll should win. Thirty-four is larger than twenty-two. The crowd of not so wise people is wise. Even more amazingly, Roger Noll might win *even if no one in the crowd knows the correct answer*. To see this, suppose that each person knew that the correct answer was either Roger Noll or one of the other names. If every person votes for either Roger or one of the other names (chosen randomly), then Roger gets on average one half of the votes. Each of the others gets a sixth of the votes. Here, the crowd knows something – the identity of the person who is not the Monkee – that no one in the crowd knows.

In practice, Roger Noll wouldn't always be the prediction. The audience predicted correctly only nine times out of ten not ten times out of ten. As so many votes are random, another person might randomly get more votes. Just as it is possible to flip a coin ten times and get eight heads, it is possible for the random choices to disproportionately favor one of the wrong choices. The probability of this type of error can be answered using statistics. With the numbers from this example the probability that the crowd would be wrong would be in the ballpark of 10%.<sup>8</sup> A second reason why Roger Noll might not be selected is that the people who do not know the answer may not pick randomly. They may suffer from a common bias that leads them to predict one of the other names frequently. Correlated errors might arise if a previous question had been about seaman, in which case Davy Jones might be thought of as a comical reference to Davy Jones' locker (a grave injustice) and this might lead to a correlated error.

This model is simple and elegant, but it only goes partway toward explaining the wisdom of crowds. It explains how, if some members of the crowd have the correct information and others do not, then the incorrect information can cancel out through randomness. It characterizes some, but not all, of Aristotle's logic using mathematics. Unfortunately, it explains none of the examples that began the chapter. In none of those cases did some members of the crowd know the correct answer. There were not, so far as we know, some subset of people who knew the exact weight of the steer at the fair Galton visited, the outcome of the horse races, or the outcomes of elections. I'm also pretty sure no one in my class had ever seen my step on a scale.

---

<sup>8</sup>An exact calculation of the probability that the audience predicts correctly is tedious but not difficult.

## Regional Sales

A second type of information aggregation model applies to situations in which each member of the crowd knows a part of the answer. It completes the logic of Aristotle. As an example, imagine a business that has sales staffs that serve five subunits representing Europe, Asia (including Australia), Africa, North America, and South America. Each has a manager who knows his or her own region's sales, but has little or no idea about sales from the other regions. The task of the managers is to guess aggregate sales. On the surface, this appears easy. If each manager reveals the sales from his or her region and those sales are totaled, this gives the correct answer. But, the logic is a bit more complicated than that because the sales managers are predicting total sales not just revealing regional sales.

Let's do an example. Suppose that the actual sales for each region are as follows:

### Actual Sales

<i>Region</i>	<i>Sales in Region</i>
Asia	60K
Europe	50K
Africa	95K
North America	75K
South America	40K
<i>Total Sales</i>	320K

How then should the managers make predictions? We consider two scenarios. Under the first scenario, the managers have some knowledge of past sales. Let's suppose that in the past, the average sales for each region have been 50K and that total sales have averaged 300K. This assumptions means that each manager would predict total sales to be 300K plus or minus the difference between actual sales in his or her region and 50K. The Asian manager would predict  $300K + (60K - 50K) = 310K$ . We can call this the Past Sales Scenario. The predictions for each of the five managers would be as follows:

### Managers' Predictions Under Past Sales Scenario

<i>Region</i>	<i>Predicted Total Sales</i>
Asia	310K
Europe	300K
Africa	345K
North America	325K
South America	290K
<i>Average Prediction</i>	314K

The aggregate prediction shown in the bottom line of the table equals 314K. The prediction is not perfectly accurate because sales were higher than expected. Their reliance on average past sales lowers their prediction. Nevertheless, the direction of their prediction is correct. Sales were above average; they predicted above average sales. Under the Past Sales Scenario, this will always be true. The predicted direction will always be correct. It just won't go far enough.

In the second scenario, we assume that the managers have no past sales data to rely on. They therefore assume that the other regions will have sales identical to their own. We call this the Correlated Sales Scenario. The Asian manager predicts sales of 360K and the European manager predicts sales of 300K. The predictions for all six regions are shown in the table below:

### Managers' Predictions Under Correlated Sales Scenario

<i>Region</i>	<i>Predicted Total Sales</i>
Asia	360K
Europe	300K
Africa	570K
North America	450K
South America	240K
<i>Average Prediction</i>	320K

This prediction is exactly correct. And if the managers use this rule, it always will be. The logic behind this result is easy to follow; it's just based on averaging the predictions.<sup>9</sup> Although this rule will always produce the correct

---

<sup>9</sup>Each manager predicts five times his region's sales then the average of these predictions is the sum of the five region's sales.

answer, the managers may not choose to use it if they have knowledge of the past. The manager of the African market, who had sales of 95K, would know that these were high sales figures. He would probably not expect every other region to do as well. At the same time, he might expect some correlation between the other regions' sales and his own. So his prediction might be somewhere in between his prediction in the Past Sales Scenario and in the Correlated Sales Scenario as might everyone else's predictions. The resulting collective prediction might then lie somewhere in between 314K and 320K. A prediction that, we might add, is also remarkably accurate.<sup>10</sup>

Though we've framed this as prediction, this example primarily involves aggregating diverse information. Each person knows a piece of the answer and these pieces can be pulled together. This logic might explain some situations, but it remains inadequate to address others. In this example, each manager knows a part of the relevant information, while in Galton's steer example it doesn't hold true-if it did one person would have to know the weight of the hooves, another the weight of the tail, another the weight of the head, and so on until all of the parts were covered.

That's not to say that this way of aggregating information doesn't work. It does. The logic is powerful and useful. Suppose that you are a manager. There may be an instance when you need to recall some piece of information buried deep in the company files, such as which of two product designs had been less costly to produce. You could search for the information in your company's files, or you could quickly poll your subordinates. You could send out a quick group email and ask people which product had been cheaper to produce. Those people that recall the correct answer will provide you with the correct information. Those that do not recall will (you hope) randomize, and in the aggregate you'll find the correct information.

Staying in this hypothetical managerial role, suppose that you need to know the total number of employees who have called in "sick" the day before a three day weekend. You could ask each of the directors of the divisions under you to report how many people called in sick from their divisions. They may be unwilling to tell you this information because it reflects badly on them. You might instead ask them to predict how many people they thought called in sick company wide. In this case as well, people may have

---

<sup>10</sup>If the managers weight past and current sales differently the collective prediction could lie outside this range.

an incentive to lie, but less of one. The average of their predictions might well be a good estimate. We can make that assumption using the logic of the second model. The second model is useful, but it does not capture predictions in their richest form.

## The Gravity of Truth

Our last two models differ only slightly. In the first, we assume that signals are discrete – heads or tails, yes or no; the second assumes that they are continuous, that they can take any real value, like the weight of a person or a steer. Both of these models are taught in probability classes and used by economists and political scientists to explain why markets and democracies work as well as they do. Most people find these probability models confusing and not much fun. They requires making lots of calculations using  $p$ 's and  $(1 - p)^2$  and the like. Sometimes to make sense of things, we have to trudge through ugly calculations.

The first model we consider relies on two possible discrete signals. One is accurate. The other is inaccurate. The probability that a person gets the accurate signal equals three fourths. To frame the model, we assume the predictive task is to determine whether sweaters just shipped in from Guatemala are made of wool ( $W$ ) or of artificial fibers ( $A$ ). For the purposes of this example, we assume that all of the sweaters in question are, in fact, knitted from artificial fibers, despite the “100% wool” tags sewn inside the back of their collars. (We can assume the tags are made of wool in the event the sweaters are made of artificial fibers, so the tags - technically speaking - are accurate.)

In this model, each of three product testers picks up a different sweater and gets a signal as to its composition. What is this signal? No bells ring or sirens go off. So we have to assume that the product testers have skin allergies to wool. Each rubs a sweater on her arm. If she breaks out in hives, she believes the sweater is wool. Otherwise, she believes the sweater to be made of artificial fibers. A skin test is not always accurate. We might expect that about one fourth of the time, product testers do not break out when testing a wool sweater, and that about one fourth of the time they do break out when testing an acrylic sweater because the acrylic has been in close contact with wool. Given these assumptions, 75% of the time a person gets the signal  $A$  for artificial fiber and 25% of the time, a person gets the signal  $W$  for wool.

In the language of probability theory, these signals can be said to be correct with probability three fourths.

We further assume that these signals are *independent*: the signal that one product tester receives depends in no way on the signals that the other receives. Formally, this property is called *independence conditional on the state of the world*. In our example, the state of the world is the stuff of which the sweaters are made. Assuming independence (conditional on the state), implies diversity, and lots of it. If people reacted to the sweaters the same way, then they would get the same signal. Therefore, to get diverse signals, they must either react differently, or they must test different sweaters.

Using a little math, we can compute the probabilities for all possible combinations of signals. Here come the detailed mathematical calculations that engineers love but cause poets to skip entire paragraphs. Call our three product testers Howard, Mita, and Rick. One possibility is that all three sweater testers get the signal  $A$ . By assumption, each gets the correct signal with probability  $3/4$ . As these signals are independent, the probability that both the first and second product testers get the signal  $A$  equals  $(3/4) \cdot (3/4)$ , and the probability that all three get the signal  $A$  equals  $(3/4) \cdot (3/4) \cdot (3/4)$ . Similar calculations for all possible combinations of signals are provided in the table below.

## Individual Signals and Collective Predictions

<i>Signals</i>			<i>Collective Prediction</i>	<i>Probability of Outcome</i>
<i>Howard</i>	<i>Mita</i>	<i>Rick</i>		
A	A	A	A	$\frac{3}{4} * \frac{3}{4} * \frac{3}{4} = \frac{27}{64}$
A	A	W	A	$\frac{3}{4} * \frac{3}{4} * \frac{1}{4} = \frac{9}{64}$
A	W	A	A	$\frac{3}{4} * \frac{1}{4} * \frac{3}{4} = \frac{9}{64}$
W	A	A	A	$\frac{1}{4} * \frac{3}{4} * \frac{3}{4} = \frac{9}{64}$
A	W	W	W	$\frac{3}{4} * \frac{1}{4} * \frac{1}{4} = \frac{3}{64}$
W	W	A	W	$\frac{1}{4} * \frac{1}{4} * \frac{3}{4} = \frac{3}{64}$
W	A	W	W	$\frac{1}{4} * \frac{3}{4} * \frac{1}{4} = \frac{3}{64}$
W	W	W	W	$\frac{1}{4} * \frac{1}{4} * \frac{1}{4} = \frac{1}{64}$

To show the wisdom of crowds we assume that given their signals, Howard, Mita, and Rick vote on the purchase lot of sweaters. When voting they just reveals the signals they received.<sup>11</sup> The table above shows that these collective predictions are correct in the first four cases and incorrect in the last four cases. The table also shows that the first four cases are far more likely than the latter four. A little addition reveals that the probability that the crowd (three is, after all, a crowd) makes the correct prediction equals  $\frac{54}{64}$ , or about 84%. This probability exceeds the probability that each person was correct individually, which was only 75%.

The collection of people predicts more accurately because a force pulls the group toward the correct answer. The underpinnings of this force can be seen best with a metaphor. Imagine two rooms. One has a door marked *A*, for artificial fibers, and the other has a door marked *W*, for wool. Imagine

---

<sup>11</sup>That assumption is not as innocuous as it may seem. In some models, rational people may choose not to vote informatively, but here they would. See Feddersen, Timothy and Wolfgang Pesendorfer. 1997. "Voting Behavior and Information Aggregation in Elections with Private Information." *Econometrica* 65(5):1029-58.

now a long line of people standing outside the two doors. One of the two rooms represents the correct answer. Suppose that each person is handed a card that tells them the correct door with probability  $p$  and the incorrect door the rest of the time. Assume  $p$  is greater than one half. These signals are passes to get into the rooms. To enter the room with the door marked  $A$  requires an  $A$  pass. To enter the room with the door marked  $W$  requires a  $W$  pass. After ten people have entered the rooms, the expected number of people who have entered the correct room will be  $10p$ . The expected number of people who enter the incorrect room will be  $10(1 - p)$ . More people will have entered the correct room than the incorrect room.

If  $p$  is close to one half, then it is possible that more people will have entered the wrong room provided there were not many people in the original line. Suppose though that a million people have now entered the rooms. Even if  $p$  is close to one half,  $1,000,000p$  will be much larger than  $1,000,000(1 - p)$ . more people will have entered the correct room. Statisticians explain this phenomenon using the *Law of Large Numbers*. As more independent signals get produced, the true value of  $p$  reveals itself. And if we assume that  $p$  is greater than one half, then a large crowd eventually gets the right answer.

A not so obvious implication of this reasoning is that we should be willing to trade off some accuracy from group size. Specifically, a group of three people each of whom gets the correct signal (independently) with probability  $\frac{3}{4}$  will not be as accurate as a group of eleven people each of whom gets the correct signal (again independently) with probability  $\frac{3}{5}$ . Bernie Grofman refers to this as the tradeoff between stupidity and group size. We'd be willing to sacrifice some accuracy if we could have more people in the group.<sup>12</sup>

This model might seem to explain the wisdom of crowds. The gravity of the truth wins out in the long run. If this seems all too convenient, it is. On what basis can we assume that people get independent signals? Though social scientists often assume this form of independence, why should we believe it to exist? Can each person in a crowd even get independent signals? We consider those questions at the end of this chapter and in the next one. For the moment, we need only recognize that this model implicitly assumes a tremendous amount of diversity across the signals. And by doing so, it makes the crowd wise. It's what some call a heroic assumption, but

---

<sup>12</sup>See Grofman, B., Owen, G., and Feld, S. "Thirteen theorems in search of the truth." *Theory and Decision*,. 15:261-278, 1983.

what he might call add hoc.

In this stark form, the model fails to apply to any of our earlier examples. None of those predictive tasks involved making a binary choice. Each required predicting a numerical value. No one received a correct signal of the weight of the steer with probability  $p$  (or an incorrect signal with probability  $(1 - p)$ ). Nevertheless, the model remains helpful. by showing how: *independent random errors cancel*. If we can find a way to guarantee that members of the crowd make random errors, then we have a wise crowd.

## The Averaging of Noise

Our final model also assumes that people get blurry signals but these take on real values. As per usual, we analyze this model in the context of an example. Suppose that a collection of people is assigned the task of determining whether or not McDonalds' policy is to keep their coffee at 170 degrees. If these people go to McDonald's and get coffee, provided that they do not go to the same restaurant at the same time, we might think that each receives something close to an independent signal conditional on the true state of the world, which in this case is McDonalds' policy.

If in fact the thermostats in McDonalds' coffee makers keep the coffee at 170 degrees, the distribution of temperatures may have a mean 170 degrees plus or minus small errors. For the cup of coffee purchased by the  $i$ th person, call this error  $E_i$ . This person's signal equals the true temperature as set by policy, call this  $T$ , plus the error term. Letting  $S_i$  denote the signal that the  $i$ th person receives, then we have that  $S_i = T + E_i$ . (Sorry about all of these subscripts, but we see in a moment why they are needed.) If these assumptions hold, a collection of McDonalds' coffee drinkers could, with high accuracy, uncover McDonalds' policy. Each person's belief of the true temperature equals the temperature set by policy plus a small error. The prediction from the crowd, which we assume here to be the average and denote by  $T^{Pred}$  equals the sum of the six predictions divided by six.

$$T^{Pred} = \frac{T + E_1 + T + E_2 + T + E_3 + T + E_4 + T + E_5 + T + E_6}{6}$$

Summing up all of the  $T$ 's this gives

$$T^{Pred} = T + \frac{E_1 + E_2 + E_3 + E_4 + E_5 + E_6}{6}$$

This prediction will be close to  $T$  provided the sum of the error terms is close to zero. If some of the individual error terms are negative and some are positive, the average of the error terms may well be smaller in absolute value than any of the individual error terms. If we assume that the errors have an average value of zero and that they are independent, then the average will be close to zero and with an even larger crowd of predictors the average will be even closer. This size of this reduction in error can be made formal using the *Law of Large Numbers*, but we won't bother. We want to focus on the underlying logic. Rousseau describes it quite accurately in his discussion of the difference between the general will and the will of all.

*There is often a great difference between the will of all (what all individuals want) and the general will; the general will studies only the common interest while the will of all studies private interest, and is indeed no more than the sum of individual desires. But if we take away from these same wills the plusses and minuses which cancel each other out, the balance which remains is the general will.*

- Jean-Jacques Rousseau, *The Social Contract, Or Principles of Political Right*, 1762

This quote by Rousseau describes in words what social scientists explain with all of those mathematical symbols – *the errors cancel out*. Crowds can be wise if each person sees the true answer plus an error (possibly a large one) so long as these errors have mean zero and are independent. The independence assumption guarantees that with enough people, the errors cancel out one another. Could this happen? Sometimes. The case of the coffee at McDonald's may be just such a case. The cups of coffee represent generated signals. Each cup of coffee has a temperature of 170 degrees plus or minus some idiosyncratic effects – perhaps cups that were stored in a cool storeroom, or a server with warm hands. For another example, suppose that we asked people living at different spots on the earth to measure the luminosity of stars. Each person would measure the true luminosity plus or minus error terms owing to ambient light, humidity, and so on, so these signals too would be generated. In this case, they would even be normally distributed since each error is the sum of lots of little random terms.<sup>13</sup>

---

<sup>13</sup>This logic underpins the design of many modern multi-mirrored telescopes. Each of the mirrors can be thought of as a diverse, flawed telescope. When combined, the

As elegant and general as this last model seems does it describe the wisdom of crowds? Does it explain how people predict election outcomes, stock prices, the winner of sporting events, or much less the weight of a steer. If we apply this model to the steer example, why should we assume that people's guesses were draws from some distribution that had the correct mean and independent errors? No Ray Kroc trained employee was handing out pieces of paper with the true value of the steer (plus or minus some error term) at Galton's fair. Yet, the model lacks any explanation of the source of the signals. They're black boxed. The model fails to explain what goes on in the heads of the people making the predictions. The model contains no interpretations, and no predictive models. Instead, the model only describes signals. And as if by magic, each signal has the correct mean and is independent of the others. How does either of these things happen? *Does either happen?* Probably no expect in cases of generated signals. The averaging-of-noise model, therefore, falls as a model of prediction, though it is a good model of blurry vision.

## The Space Between

The three types of information aggregation all produce valuable insights. Let's quickly review. Crowds can predict the correct answer even if only a small set of people in the crowd know it (Model 1). As we may not know which people in the crowd know the answer, we may therefore choose to rely on crowds to reveal information. We also saw that if individuals who each know a piece of information and if they make predictions based on history or on an assumption that the other pieces are like their own, then they collectively make accurate predictions. Both of these models can explain cases where crowds predicted accurately, but neither fully explains our examples. We also say that if people receive independent, generated signals, be they discrete (Model 3) or continuous (Model 4), that their errors cancel. The collective prediction is highly accurate. The last two models apply neatly to situations in which people see the quality or the value plus or minus a small error. They apply less neatly to situations in which people make predictions based on models.

---

idiosyncratic blurriness from the various pictures cancels out creating a clear vision of a portion of the night sky.

Implicitly, all of these models assume some diversity. That's what we want to keep in mind. That even though we didn't know it. All of those assumptions about independence were also assumptions about diversity. However, the connections between the two concepts – diverse predictive models and independent predictions are not formalized. We are left to rely on our intuitions as to whether predictions are independent, and whether on average they are correct. In some cases, these statistical assumptions may be valid. In others, they may be hiding the great and powerful Oz behind a flimsy curtain. So, to understand the role of diversity in contributing to the wisdom of crowds we must unpack where those predictions come from and how they differ. We need to look at differences directly. Once we have, we can then come back to these models with a deeper understanding of when they make sense and when they don't.

For example, if we hope to understand the wisdom of crowds, we need a model in which people make predictions. To provide a preview of how that model might work, we return to the steer example. The estimates by fair goers were not the true weight plus an error. It wasn't as though they saw the steer standing on a scale from different angles, and as a result, each saw the true weight with some error. More likely, each had some primitive model of what a steer weighs. These models led to predictions of the steer's weight. The predictions were not naive shots in the dark. In 1906 people knew a lot about steers. The farmers at this exhibition probably categorized the steer based on its attributes: big head, thin haunches, tall at the shoulders, large barrel chest, and so on, and then made educated guesses. Based on Galton's data, they all had slightly different models. Otherwise, their predictions would have been the same.

But certainly, the diversity of their predictions cannot explain why they were collectively so accurate? Amazingly, combined with their moderate abilities, their diversity drove them to accuracy. So it would seem that we need both some level of individual accuracy and some amount of collective diversity for a crowd to be wise. But that's just crude intuition. We want to build a logic.