PROFESSOR: All right. We'll back up and start again since I just turned on my microphone. I started with the observation that for most of recorded history people thought qualitatively, not quantitatively. They didn't know what statistics were. They must have had some intuitive sense, for example, that if you're old, you're more likely to have bad hearing than if you're young. If you're old you're more likely to die than if you're young. Things like that.

But they were just anecdotes, and they had no careful way to go from statements about individuals to statements about populations or expectations. This changed in about the middle of the 17th century, changed fairly dramatically, when an Englishman named John Graunt published something called *The Natural and Political Observations Made Upon the Bills of Mortality.* This was the first work in recorded history that actually used statistics. And what he did is he looked at the fairly comprehensive statistics of when people died in the city of London. And from that, attempted to produce a model that could be used to predict the spread of the plague.

And it turns out, he was pretty good at it. And that just changed the way people started to think. And since that time, people have used statistics both to inform and, unfortunately, to mislead. Some have willfully used statistics to mislead, others have merely been incompetent. And that gets me to what I want to talk about today, which is statistics.

This is often attributed to Mark Twain but, in fact, he copied it from Benjamin Disraeli, who said, "There are three kinds of lies: lies, damned lies, and statistics." And there is, unfortunately, a lot of truth in that. More recently, in the '50s Darrell Huff wrote a wonderful book, I recommend it, called *How to Lie with Statistics.* Now

here's a quote from the book, "If you can't prove what you want to prove, demonstrate something else and pretend that they're the same thing. In the daze that follows the collision of statistics with the human mind, hardly anyone will notice the difference." Alas, that seems to be true.

So what I want to do today is talk about a few ways in which one can be fooled into drawing inappropriate conclusions from statistical data. Now I trust that you will use this information only for good. And it will make you a better consumer and purveyor of data rather than a better liar, but that's up to you. All right. So let's start with the first way out thing I want to talk about. And it's important to always remember that no matter how good your statistics are, statistical measures don't tell the whole story. We've seen examples of this already earlier in the term.

There are an enormous set of statistics that can be extracted from a data set. And by carefully picking and choosing among them, it's possible to convey almost any impression you want about the same data set. The best antidote, of course, is to look at the data itself. In 1973, the statistician John Anscombe published a paper with this set of data. Four different examples where he gave values for x and y, nothing very sophisticated.

The interesting thing is that in many ways, the statistics for these four data sets are very similar. They have the same mean, the same median, the same variance for x and y both, the same correlation between x and y. And even if we use linear regression to get a fit, we get very similar things. So let's look at it. I actually wrote some code to do that.

So this just reads in the data set that we just had up, and then plots some things with it. And so let's look at it. So what you can see is, you have these in your hand out, we have four graphs, four plots. And I won't show you them all because they're all are identical. What they all give me is a mean of 7.5 and a whole bunch of things, same to 17 decimal places or so, the same median and the same linear fit, y equals 0.5x plus 3.

So I might write a paper in which I would tell you that well, from a statistical sense,

these data sets are all the same. Every statistical result I ran said these things are indistinguishable. Well, is it true that these data sets are indistinguishable? Well, one way to look at it is to actually plot the points. So we can run it with true, which says in addition to plotting the data, put points on them in addition to plotting the statistics. And now we see something fairly dramatic.

So for example, figure (2) and figure (1) don't look a lot like each other, when we actually look at the data. Figure (3) looks quite different and figure (4) is remarkably different. Compare figure (4) to, say, figure (1). Again, these are all in your hand out. So the moral is pretty simple here, and it's one we looked at before, which is, don't ever ignore the data. Don't just look at statistics about the data, try and find a way to look at the data itself.

Of course, it's easy enough to look at the data, but how do you look at it? And the next thing to remember is that pictures can be deceiving. There can be no doubt about the utility of graphics for quickly conveying information. However, when used carelessly or maliciously, a plot can be highly misleading. So let's look-- go back to the PowerPoint, and look at this plot of housing prices in the Midwest.

So we've got a plot here, and we've got years 2006, 2007 and then in 2008 and 9, we have quarters. You may remember that there was an event in the housing market in 2008 precipitating the global financial crisis. And if we look at this, what impression do we get about housing prices in the mid west during this period? Well I would get the impression that they are remarkably stable. You publish this and you say, OK, look they really haven't changed very much. Maybe they've gone down a little, but nothing very serious.

If we compare that to this plot, which is exactly the same data, and now I ask you about housing prices in the Midwest. Well, what you might tell me is it's-- they're remarkably unstable. And in fact, there was clearly some sort of horrible event here. Exactly the same data, two plots both of which are truthful, but which give a radically different impression about what happened. The chart on the right was designed-- on the right in your handout-- was designed to show that they were highly unstable.

So what's the difference? What trick did I use to produce these two plots? Yeah?

AUDIENCE: In the more stable one, you used the logarithmic scale. And then here only has selected numbers so--

PROFESSOR: So that was certainly one trick that I performed. That in the first chart, I plotted the y-axis logarithmically, which always make things look like they are changing less than if I plot it linearly. And in this chart, I used a linear plot. Go ahead.

AUDIENCE: You see a much narrower scale on the second plot. So that the magnitude of the difference is much less compared to the magnitude of the whole graph of the scale.

PROFESSOR: So that's the other thing I did is, if you look at it, I sort of cheated. I had full years here and then I went to quarters. So in part of my chart, the resolution on the x is pretty wide is a whole year, and then part of it's on a quarter. And, not surprisingly, since we know that housing prices change seasonally, they're different in the spring than in the winter, once I start plotting quarters, even if there had not been a crash, it would have looked much less stable in the out years, because I changed the resolution on the x-axis.

I didn't lie. You can tell reading the legend that I did that. But I sure could have fooled a lot of people with these charts. Here's another nice example of statistics. So this plots-- this is from a paper by two professors, I think from Yale, shows what you can do if you're in the Ivy league, that plots initials against GPA for students at Yale. And so you can see that if your first name starts with the letter A, I think it was first initials, your GPA is considerably higher than if it starts with C or D. And if your parents weren't nice enough to give you an A name, you could hope they at least gave you a B name. And you certainly don't want them to give you a C or a D name.

So if you're Charlene or David you could have a real problem. I have to say my first child was named David. His GPA might have been somewhere in there. All right -- clearly it matters. Well, what tricks did I perform here?

AUDIENCE: There is very little disparity in x. You are going from 3.32 to--

PROFESSOR: Right. So what I did here is I made the range on the y-axis very small, ranging from 3.32 to 3.38, not a big difference. However, because that's the whole thing, it looks like it's a big difference. You will often see this when you, say, look at things in newspapers where, in fact, someone has manipulated one of the axes to make things look big or small. If I had ranged this from a GPA of 0.0 to a GPA of 4.0, the highest GPA at Yale, this difference would have looked very tiny. But I didn't -- actually it wasn't I. I actually copied this from their paper.

This was the way they presented it in their paper. Because they were trying to argue in the paper that your name had a big influence on your life. And they used many statistics including your grades. And so they actually formatted it kind of like this to try and give you this impression. Now later we'll see another statistical sin they committed in this paper, which, basically, was designed to show that in your name was destiny. And they had many other things. If you're a baseball player, and your name starts with K, you're more likely to strike out. Because K is the symbol for strikeouts in a baseball score book, a lot of implausible things.

All right. Moving right along, probably the most serious and common statistical error is the one known as Garbage In Garbage Out. And it's so common that people typically refer to it by its acronym GIGO -- Garbage In Garbage Out. A classic example of this occurred and, we could look at more recent examples, but I don't want to offend any of you. In 1840, United states census showed that insanity among free blacks and mulattoes was roughly 10 times more common than insanity among enslaved blacks or mulattoes. And the conclusion from this was obvious, I guess.

US Senator, former Vice President, and later, Secretary of State John C. Calhoun concluded from the census and I quote, "The data on sanity revealed In this census is unimpeachable. From it, our nation must conclude that the abolition of slavery would be to the African, a curse." Because after all, if you freed them from slavery, they would all go insane. That's what the statistics reported, he said. Now never mind it was soon clear that in fact that census was riddled with errors. And John Quincy Adams, a former Vice President and Massachusetts resident responded to

Calhoun and said, no that's a ridiculous conclusion. The census is full of errors.

Calhoun, being a very patient person, explained to Adams the following, "There were so many errors, that they balanced one another out, and led to the same conclusion just as much as if they were all correct." There were just enough errors that you could be OK. Well, what was he relying on? What should he have said if he wanted to make this statement more mathematically precise?

What he was basically implying is that the measurement errors are unbiased and independent of each other and, therefore, almost identically distributed on either side of the mean. I see a typo, I might as well fix it. Might as well make it big enough to fix it. That's interesting.

If he had made this much more precise statement, then you could have had an meaningful discussion-- assuming it was possible to have a meaningful discussion with John Calhoun, which is perhaps dubious-- about whether or not, in fact, the errors are independent. Because if they're not, if, for example, they represent bias in the people compiling the data, then you cannot rely upon statistical methods to say that they'll balance each other out. You remember way back in Gauss' time, Gauss talked about this when he talked about the normal distribution. And said, well, if we take these astronomical measurements and we assume our errors are independent and normally distributed, then we can look at the mean and assume that that's close to the truth.

Well, those are important assumptions which in this case turned out to be not correct. And, in fact, it was later shown that the errors did not balance each other out nicely. And in fact, today you can say that no statistical conclusion can be drawn from that. On the other hand recently, the US National Research Council, perhaps the most prestigious academic organization in the United states, published a ranking of all universities in the country. And it was later shown that it was full of garbagy input. And they did extensive statistical analysis and published it on data that turned out to be just wrong. And it was very embarrassing.

Now the good news is MIT came out near the top of this analysis. And the bad news

is we can't conclude that it actually should have been near the top because who knows about the quality of the data, but kind of embarrassing. All right. Moving right along, another very common way to lie with statistics is to exploit what's called the cum hoc ergo propter hoc fallacy. So anyone here study Latin? Bunch of techy-- Oh OK. Well what does it mean?

AUDIENCE: With this therefore, because of this?

PROFESSOR: Boy, your Latin is good. Either that or you just know statistics. But I have to say that was the most fluent translation I've had it all the years I've asked this question. I hit the relay man-- the relay woman on the throw. All right. Yes, with this, therefore because of this. I don't know why but statisticians, like physicians and attorneys, like to show off by phrasing things in Latin. So for example, it is a statistical fact that college students, including MIT students, who regularly attend lectures have higher GPAs than students who attend lectures only sporadically.

So that would tell us that those of you in the room are likely to have a higher GPA than the various students in 6.00 who are not in this room. I hope it's true. Now if you're a professor who gives these lectures, what you want to believe it's because the lectures are so incredibly informative, that we make the students who come much smarter and, therefore, they do better. And so we'd like to assume causality. Because I give beautiful lectures and you choose to come, you will get a better grade in 6.00. Well, yes, there's a correlation. It's unquestionably true, but causation is hard to jump to.

For example, maybe it's the point that students who bother to come to lecture also bother to do the problem sets, and are just more conscientious. And whether they came to lecture or not, the fact that they're more conscientious would give them better GPAs. There's no way I know to separate those two things, other than doing a controlled experiment, right? Maybe kicking half of you out of lecture every day and just see how it goes. But it's dangerous but again, you can read things like the faculty newsletter, which will talk about how important it is to come the lecture because you'll do better. Because whoever wrote that article for the faculty

newsletter didn't understand this-- this fallacy, or was just thinking wishfully.

Another nice example, one that was in the news not too long ago, has to do with the flu. This was the cases of flu in New York State in recent years. And you'll notice that there was a peak in 2009, and that was the famous swine flu epidemic, which I'm sure you all remember. Now, if you look at this carefully, or even not too carefully, you'll notice a correlation between when schools are in session and when the flu occurs. That in fact, during those months when schools are in session, there are more cases of flu than in the months when school is not in session, high schools, colleges, whatever. Quite a strong correlation, in fact.

This led many to conclude that going to school is an important causative factor in getting the flu. And so maybe you shouldn't have come to the lectures because you would have just gotten the flu by doing so. And in fact because of this, you had many parents not sending their kids to school during the swine flu epidemic. And in fact, you had many schools closing in some communities because of the swine flu epidemic.

Well, let's think about it. Just as you could use this correlation to conclude that going to school causes the swine flu, you could have also used it to prove that the flu causes you to go to school. Because more people are in school when the flu season is at it's height. And therefore, it's the growth of flu that causes people to go to school. That's an equally valid statistical assumption from this data. Kind of a weird thing but it's true, right?

Just as we could conclude that having a high GPA causes people to come to the lecture. You look at your GPA every morning, and if it's high enough, you come to lecture, otherwise, you don't. You could draw that conclusion from the data as well. The issue here that you have to think about is whether or not there is what's called a lurking variable, some other variable that's related to the other two, and maybe that's the causative one.

So for example, a lurking variable here is that the school season coincides with or the non-school season, maybe I should say, coincides with the summer. And in fact,

if you study the flu virus in a lab, you will discover that it survives longer in cold weather than in hot and humid weather. When it's cold and dry, the flu virus will survive for a longer time on a surface than it will when it's warm and humid.

And so in fact, maybe it's the weather, not the presence of schools, that causes the flu to be more virulent during certain times of the year. In fact, it's probably likely true. So there is a lurking variable that we have to consider, and maybe that's the causative factor.

Now, this can actually lead to some really bad decisions in the world. I'm particularly interested in issues related to health care and public health. In 2002, roughly 6 million American women were taking hormone replacement therapy, in the belief that this would substantially lower their risk of cardiovascular disease. It was argued that women over a certain age or of a certain age, if you took extra hormones, they were less likely to have a heart attack. This belief was supported by several published studies in highly reputable journals in which they showed a strong correlation between being on hormone replacement therapy and not having cardiovascular disease.

And this data had been around a while and, as I said, by 2002 in the US, roughly 6 million women were on this therapy. Later that year, the Journal of the American Medical Society published an article asserting that in fact being on this therapy increased women's risk of cardiovascular disease. It made you more likely to have a heart attack. Well, how could this have happened? After the new study came out, people went back and reanalyzed the old study and discovered that the women in that study who'd been on hormone replacement therapy were more likely than the other women in the group to have also better diet and be on a better exercise regimen.

In fact, they were women who were more health conscious. So there were the lurking variables of diet and exercise and other things that were, in fact, probably the causative factors in better health, not the replacement therapy. But there was this lurking variable that had not been discovered in the initial analysis of the data.

So what we saw is that taking hormone replacement therapy and improved cardiac health were coincident effects of a common cause, that is to say being health conscious. Kind of a strange thing but a true and sad story.

All right. Moving right along, another thing to be cautious of is non-response bias and related problem of a non-representative sample. You'll probably recall that when I first started talking about statistics and the use of randomness, I said that all statistical techniques are based upon the assumption that by sampling a subset of a population, we can infer things about the population as a whole. And that's true, typically, because if random sampling is used, you can make assumptions that the distribution of results from the random sample, if it's large enough, will be the same as a distribution of results from the whole population. And that's why we typically want to sample randomly.

And so for all the simulations we looked at, we used random sampling to try and ensure that a small number of samples would give us something representative of the population. And then we use statistical techniques to answer the question about how many random samples we needed. But those techniques were only valid if the samples were indeed random. Otherwise, you can analyze it to your heart's content and any conclusions you've drawn are likely to be fallacious. Unfortunately, many studies, particularly in the social sciences, are based on what is often called a convenience sampling.

So for example, if you look at psychological-- psychology journals, you'll find that many psychological studies use populations of undergraduates for their studies. Why did they do this? Is it because they believe that undergraduates are representative of the population as a whole? No. It's because they're captive. They have to agree to participate, right? It's a convenience if you happen to be at a university to do your experiments on undergraduates.

And so they do that and then they say, well, the undergraduates are just like the population as a whole. You may have observed that at least at this institution, the undergraduates are probably not representative of the population as a whole. A

well-known example of what you can do with this occurred during World War II. Whenever an allied plane would return from a bombing run over Germany, the plane would be inspected to see where flak had hit it. So the planes would fly over to drop bombs, the Germans would shoot flak at the planes to try and knock them out the air. They'd come back to England, they'd get inspected, they'd say, well the flak hit this part of the plane more often than that part of the plane on average.

And so they would reinforce the skin of those parts of the plane where they expected the flak to hit, to try and make the plane less likely to be damaged in future runs, or the planes in general. What's wrong with this? Yeah?

AUDIENCE: They're not getting the planes that dropped?

PROFESSOR: Exactly. What they're not sampling is the planes that never made it back from the bombing, ooh, that never made it back from the bombing runs because they weren't there to sample. And in fact maybe it's the case that what they were doing was reinforcing those parts of the planes where it didn't matter if you got hit by flak because it wouldn't cause a plane to crash, and not reinforcing those parts of the planes that were most vulnerable to being damaged by flak. They did a convenient sampling, they drew conclusions, and they probably did exactly the wrong thing in what they chose to reinforce in the airplanes.

This particular error is called non-response bias where you do some sort of survey, for example, and some people don't respond and, therefore, you ignore what they would have said. It's perhaps something we see when we do the underground guide to Course 6. In fact I should point out that it's now online. And it would be good if each of you would go and rate this course, rate the lectures, rate the TAs, et cetera. We actually do read them and it makes a difference in how we teach the course in subsequent terms.

But there's clearly a bias. You know, maybe only the people who really feel strongly about the course, either positively or negatively, bother to fill out the survey. And we draw the conclusions that there's a bimodal distribution, and nobody thinks it's kind of mediocre, because they don't bother responding. Or maybe only the people who

hate the course respond, and we think everybody hates the course. Who knows. It's a big problem.

We see it, it's a big problem today with telephone polls, where you get more convenient sampling and non-representative samples, where a lot of polls are done using telephones. By law, these pollsters cannot call cell phones, so they only call land lines. How many of you have a land line? Let the record show, nobody. How many of your parents have a land line? Let the record show, pretty much everybody. Well, that means your parents are more likely to get sampled than you when there's a poll of, say, who should be nominated for president.

And so any of these polls that are based on telephones will be biased. And, unfortunately, their poll may just say, a telephone sample, and people may not realize the implication of that. That whole part of the population is under sampled.

There are lots of examples of this. All right. Moving along, another problem we often see is data enhancement. It's easy to read much more into data than it actually implies, especially when viewed out of context. So on April 29, 2009, CNN reported that quote, "Mexican health officials suspect that the swine flu outbreak has caused more than 159 deaths and roughly 2,500 illnesses." It was pretty scary stuff at the time, and people got all worried about the swine flu.

On the other hand, how many deaths a year do you think are attributable to the conventional seasonal flu in the US? Anyone want to hazard a guess? 36,000. So 36,000 people a year, on average, will die from the seasonal flu, which sort of puts in prospective that 159 deaths from the swine flu maybe shouldn't be so terrifying.

But again, people typically did not report both of those. Another great statistic, and accurate, is that most auto accidents happen within 10 miles of home. I'm sure many of you have heard that. So what does that mean? Almost nothing. Most driving is done with 10 miles-- within 10 miles of home. And besides that, what does home mean in this context?

What home means is the registration address of the car. So if I were to choose to

register my car in Alaska, does that mean I'm less likely to have an accident driving around MIT? I don't think so. Again, it's a kind of a meaningless. Another aspect of this is people often extrapolate from data.

So we can look at an example of internet usage. This is kind of a fun one too. So what I've plotted here is the internet usage in the United states as a percentage of population. And I plotted of this from-- starting at 1994. And the green line, or actually, the blue line there are the points and the green line is a linear fit. If you looked at my code, you'd see I was using polyfit with a 1 to get a line to fit, and you can see it's a pretty darn good fit. So people actually looked at these things and used this to extrapolate internet usage going forward.

So we can do that. Now, we'll run the same code with the extrapolation turned on. And so figure (1) is the same figure (1) as before, same data, same fit. And here's figure (2). And you'll notice that as of last year about 115% of the US population was using the internet, probably not true. It may be possible in sports to give 110%, but in statistics it's not.

Again, you see this all the time when people are doing these projections. They'll say, fit some data, they extrapolate into the future without understanding why maybe that isn't a good thing to do. We saw that by when we were modeling springs, right? We could accurately project linearly until we exceeded the constant of elasticity at which point our linear model was totally broken.

So you always need to have some reason other than just fitting the data to believe that what you're doing makes actual sense. All right. The final one I want to talk about is what is typically called in the literature the Texas sharpshooter fallacy. And this is a little bit tricky to understand sometimes. Actually, is there anyone here from Texas? Oh good, so no one will be offended by this.

Well imagine that you're driving down some country road in Texas and that you see a barn. And that barn has six targets painted on it, and in the dead center of each target, you find a bullet hole. So you're driving, your pretty impressed, and you stop. And you see the owner of the barn and you say, you must be a damn good shot.

And he says, absolutely, I never miss. At which point the farmer's wife walks out and says, that's right there ain't a man in the state of Texas who is more accurate with a paint gun.

What did he do? He shot six bullets into the barn, and he was a terrible shot. They were all over the place. Then he went and painted a target around each of them. And it looked like he was a great shot. Now you might think that, well that's silly, no one would do that in practice. But in fact, it happens all of the time in practice.

A classic of this genre appeared in the magazine *New Scientist,* in 2001. And it reported that a research team led by John Eagles of the Royal Cornhill Hospital in Aberdeen, had discovered that and I quote, "Anorexic women are most likely to have been born in the spring or early summer between March and June. In fact, there were more than-- there were more than 13%-- there were 13% more anorexics born on average in those months, and 30% more anorexics, on average, in June."

Now, let's look at this worrisome statistic. Are any of you women here born in June? All right. Well, I won't ask about your health history. But maybe you should be worried, or maybe not. So let's look at how they did this study. You may wonder why so many of these studies are all studies about women's health. And then, perhaps, because they're all done by male doctors. Anyway, the team studied 446 women who had been diagnosed as anorexic. So if you divide that by 12 what you've discovered is that, on average, there should have been 37 women born in each of those months, of the 446.

And in fact, in June, there were 48 anorexic women born. So they said, well, how likely is this to have occurred simply by chance? Well as I am want to do in such occasions, I checked their analysis, and I wrote a little piece of code to do that. So trying to figure out what's the probability of 48 women being born in June, I ran a simulation in which I simulated 446 births and chose a month at random, and looked at the probability.

And let's see what it was when we run it. Oops, well we didn't want these graphs.

The probability of at least 48 births in June was 0.042. So in fact, pretty low. You might say, well, what's the odds of this happening by accident? Pretty small. Therefore, maybe we are really on to something. Maybe it has to do with the conditions of the birth and the weather or who knows what.

Well, what's wrong with this analysis? Well, one way to look at it is this analysis would have been perfectly valid if the researchers had started with a hypothesis that there are more babies born in June than in any other month-- more future anorexics born in June than in any other month, and then run this experiment to test it, and validated it. So if they had started with the hypothesis, and then from the hypothesis conducted what's called a prospective study then they would have, perhaps, valid reason to believe that the study supports the hypothesis.

But that's not what they did. Instead what they did is they looked at the data and then chose a hypothesis that matched the data, the Texas sharpshooter fallacy. Given that that was the experiment they performed, the right question to ask is not what is the probability that you had 48 future anorexics born in June, but what was the probability that you have 48 future anorexics born in at least one of the 12 months? Because that's what they were really doing, right? So therefore, we should really have run this simulation.

Similar to the previous one, again, these are in your hand out, but is there at least one month in which there were 48 births? And if we run that we'll see that the probability is over 40%, not so impressive as 4%. So in fact, we see that we probably shouldn't draw any conclusion. The probability of this happening by pure accident is almost 50%. So why should we believe that it's somehow meaningful.

Again, an example of the Texas sharpshooter fallacy that appeared in the literature and a lot of people fell for it. And if we had more time, I would give you many more examples, but we don't. I'll see you on Thursday, on Tuesday, rather. Two more lectures to go. On Tuesday, I'm going to go over some code that I'll be asking you to look at in preparation for the final exam. And then on Thursday, we'll wrap things up.