JOHN GUTTAG: Hello, everybody. Well, here we are at the last lecture. We're going to finish talking about statistical sins and then do a little bit of a wrap-up.

Let's look at a hot topic-- global fiction-- or global warming, fact or fiction. You've done a problem set related to temperatures in the US. Here is a plot generally accepted of the change in temperatures on the planet between 1880 and 2014.

Now, if we look at this plot, we could see this commits one of the statistical sins I complained about on Monday, that look where it's starting the y-axis, way down here at 55. And you remember, I told you to beware of charts for the y-axis doesn't start at 0. So maybe the people who are trying to claim about global warming are just deceiving us with this trick of the axis.

So here's what happens when you put it at 0. And as you can see-- or barely see-- this axis runs from 0 up to 110 as the average temperature. And as you can see quite clearly, it's hardly changed at all.

So what's the deal here? Well, which is a more accurate presentation of the facts? Which conveys the accurate impression?

Let's look at another example, maybe a little less controversial than climate change-- fever and flu. It's generally accepted that when you get the flu you might run a fever. So here is someone who had the flu. And this is plotting their fever from the beginning to its peak. And it does appear, if we were to fit a curve to this, it would look pretty much like that.

On the other hand, if we assume that somebody's temperature could range between 0 and 200, we can see that, in fact, your temperature doesn't move at all when you get the flu. So the moral is pretty clear, I think. Even though on Monday I talked about being suspicious when people start the y-axis too far from 0, you should truncate it to eliminate totally preposterous values.

No living person has a temperature of 0 degrees Fahrenheit. So again, don't truncate it just to make something look like it isn't, but don't expand it to deceive either. Let's return to global warming.

This is a chart that was actually shown on the floor of the US Senate by a senator from Texas, who I shall not name. And obviously, the argument here was that, well, sure global warming bounces up and down. But if we go back, we can see here, the date is 19-- can we see it? I can see it. Maybe 1986, I think.

You can see that the argument here is, in fact, if you fit a trend line to this, as he's done, it hasn't changed at all. And so even though we've had a lot of carbon emissions during this period, maybe global warming is not actually happening. This is in contradiction to the trend I showed before.

Well, what's going on here? This is a very common way that people use statistics poorly. They confuse fluctuations with trends. What we see in any theories of data-- time series, or other series-- you always have fluctuations. And that's not to be confused with the trend.

And in particular, what you need to think about when you're looking at a phenomenon is choose an interval consistent with the thing that's being considered. So we believe that climate change is something that happens over very long periods of time. And it's a little bit silly to look at it on a short period of time.

Some of you may remember two years ago, we had a very cold winter here. And there were people who were saying, well, that shows we don't have global warming. Well, you can't really conclude anything about climate change looking at a year, or probably not even looking at 10 years or 20 years. It's a very slow phenomenon.

On the other hand, if you're looking at the change in somebody's heart rate, seeing if they have a heart condition, you probably don't want to look at it over a 10-year period. So you have to decide what you're doing and find an interval that lets you look at the trends rather than the fluctuations.

Any rate, maybe even if we're having global warming, at least the Arctic ice isn't melting, though apparently, I read in the paper this morning they found a huge crack in it. So this was reported in the *Financial Post* on April 15, 2013. You can read it yourself. But the basic import of it is they took the period from April 14, 1989 to April 15, 2013 and said, look, it's not changing. In fact, the amount of arctic ice is unchanged.

Well, what's the financial-- not the financial-- what's the statistical sin being committed here? If we look at this data, this is an anomaly chart. I think you saw one of these in one of the problems sets, where you fix something at 0 and then you show fluctuations relative to that. So here, it's the Arctic ice relative to a point.

And what we see here is that if you go and choose the right date-- say this one in 1989-- and you come over here and you choose the right date in 2013-- say this one-- you can then draw a line and say, oh, look, it hasn't changed. This is something people frequently do, is they take a whole set of data, and they find two points that are consistent with something they believe. And they draw a line between those two points, fit a curve to those two points, and draw some conclusion.

This is what we call cherry picking, I guess from the notion that when you go to pick cherries you only want to pick the right ones, leave the others to ripen for a bit on the tree. It's really bad. And it's something that, unfortunately, the scientific literature is replete with, where people look at a lot of data, and they pick the points that match what they want to prove.

And so as you can see, while the trend is quite clear, you could prove almost anything you wanted by selecting two points very carefully. I could also show that it's crashing much faster than people think it is by picking these two points. If I wanted to argue that it's catastrophic, I'd pick those two points and say, look at that, it's disappearing at an incredible rate. So you can lie in either direction with this data by careful cherry picking.

As a service to you, I know the holidays are coming and many of you have not bought presents for your parents, so here's a modest gift suggestion, that the family that shoots together something or other. Well, all right, so we can ask, is this a good gift? Well, probably.

We can look at this statistic. It's not dangerous at least. We see that 99.8% of the firearms in the US will not be used to commit a violent crime. So guns apparently are not actually dangerous, or at least not in the hands of criminals.

Well, let's look at this. How many privately owned firearms are there in the US? And anyone want to guess who hasn't looked ahead? Yeah.

AUDIENCE: 400 million.

JOHN GUTTAG: 400 million. 340 million people and 400 million guns is the guess, more than one per person.

You certainly are the right order of magnitude. I think it's about 300 million, but it's hard to count them. Maybe this doesn't count water pistols.

So if you assume there are 300 million firearms and 0.2% of them are used to commit a violent crime in every year, we see that how many crimes is that? 600,000. So in fact, it's not necessarily very meaningful to say that most of them are not used to commit a crime.

Well, let's look at another place where we look at a statistic. Probably most of you don't even remember the scary swine flu epidemic. This was a big headline. And people got so scared of the swine flu they were doing things like closing schools to try limit the spread of the flu. New York City closed some schools because of it, for example.

So is this a scary statistic? Well, maybe, but here's an interesting statistic. How many deaths per year are from the seasonal flu in the US-- the ones we try and prevent with a flu shot? 36,000. So what we see is that, it doesn't make a lot of sense to panic over 159 in the light of this number. So the point here for both this and the issue about the firearms is that context matters.

Yeah, I love this cartoon. A number without context is just a number. And numbers by themselves don't mean anything.

So to say that there were 159 deaths from the swine flu is not very meaningful without some context. To say that only 0.2% of firearms are used to commit a violent crime is not very meaningful without context. Whenever you're presenting a statistic, reading about a statistic, and you just see a number that seems comforting or terrifying, try and put some context around it.

So a related thing is relative to what? Suppose I told you that skipping lectures increases your probability of failing this course by 50%. Well, you would all feel great, because you're here. And you would be laughing at your friends who are not here, because figuring that will leave much better grades for you.

What does this mean, though? Well, if I told you that it changed the probability of failing from a half to 0.75, you would be very tempted to come to lectures. On the other hand, if I told you that it changed the probability from 0.005 to 0.0075, you might say, the heck with it, I'd rather go to the gym.

Again this, is an issue. And this is something that we see all the time when people talk about percentage change. This is particularly prominent in the pharmaceutical field. You will read a headline saying that drug x for arthritis increases the probability of a heart attack by 1% or 5%. Well, what does that mean?

If the probability was already very low, increasing it by 5%, it's still very low. And maybe it's worth it not to be in pain from arthritis. So talking in percentages is, again, one of these issues of it doesn't make sense without the context. In order to know what this means, I need to know what regime I'm in here in order to make a intelligent decisions about whether to attend lecture or not. It goes without saying, you have all made the right decision.

So beware of percentage change when you don't know the denominator. You get a percentage by dividing by something. And if you don't know what you're dividing by, then the percentage is itself a meaningless number.

While we're sort of talking about medical things, let's look at cancer clusters to illustrate another statistical question. So this is a definition of a cancer cluster by the CDC-- "a greater-than-expected number of cancer cases that occurs in a group of people in a geographic area over a period of time." And the key part of this definition is greater-than-expected.

About 1,000 cancer clusters per year are reported in the US, mostly to the Centers for Disease Control, but in general to other health agencies. Upon analysis, almost none of them pass this test. So the vast majority, some years all of them, are deemed actually not to be cancer clusters.

So I don't know if-- has anyone here seen the movie *Erin Brockovich?* Subsequent analysis showed that was actually not a cancer cluster. It's a good movie, but turns out statistically wrong. This, by the way, is not a cancer cluster. This is a constellation.

So let's look at a hypothetical example. By the way, the other movie about cancer clusters was the one set in Massachusetts. What was the name? *A Civil Action.* Anyone see that? No. That was a cancer cluster.

Massachusetts is about 10,000 square miles. And there are about 36,000 cancer cases per year reported in Massachusetts. Those two numbers are accurate. And the rest of this is pure fiction.

So let's assume that we had some ambitious attorney who partitioned the state into 1,000 regions of 10 square miles each and looked at the distribution of cancer cases in these regions trying to find cancer clusters that he or she could file a lawsuit about. Well, you can do some arithmetic. And if there are 36,000 new cancer cases a year and we have 1,000 regions, that should say that we should get about 36 cancer cases per year and per region.

Well, when the attorney look at the data, this mythical attorney, he discovered that region number 111 had 143 new cancer cases over a three-year period. He compared that to 3 times 36 and said, wow, that's 32% more than expected. I've got a lawsuit. So he went to tell all these people-- they lived in a cancer cluster.

And the question is, should they be worried? Well, another way to look at the question is, how likely is it that it was just bad luck? That's the question we've always ask when we do statistical analysis-- is this result meaningful, or is it just random variation that you would expect to see?

So I wrote some code to simulate it to see what happens-- so number of cases, 36,000, number of years, 3. So all of this is just the numbers I had on the slide. We'll do a simulation. We'll take 100 trials.

And then what I'm going to do is for t in the range number of trials, the locations, the regions, if you will, I'll initialize each to 0, 1,000 of them, in this case. And then for i in the range number of years times number of cases per year, so this will be 3 times 36,000. At random, I will assign the case to one of these regions. This is the random. Nothing to do with cancer clusters, just at random, this case gets assigned to one of the 1,000 regions.

And then I'm going to check if region number 111 had greater than or equal to 143, the number of cases we assumed it had. If so, we'll increment the variable num greater by 1, saying, in this trial of 100, indeed, it had that many. And then we'll see how often that happens.

That will tell us how improbable it is that region 111 actually had that many cases. And then we'll print it. Does that makes sense to everyone, that here I am doing my simulation to see whether or not how probable is it that 111 would have had this many cases?

Any questions? Let's run it. So here's the code we just looked at. Takes just a second. That's why I did only 100 trials instead of 1,000.

I know the suspense is killing you. It's killing me. I don't know why it's taking so long. We'll finish. I wish I had the *Jeopardy* music or something to play while we waited for this. Anna, can you home some music or something to keep people amused? She will not.

Wow. So here it is. The estimated probability of region 111 having at least 1 case-- at least 143 cases-- easier to read if I spread this out is 0.01. So it seems, in fact, that it's pretty surprising-- unlikely to have happened at random.

Do you buy it? Or is there a flaw here? Getting back to this whole question. Yes.

AUDIENCE: I think it's flawed because first off you have to look at the population. That is more important.

JOHN GUTTAG: You have to look at what?

AUDIENCE: Population as opposed to like the number of areas, because when you get past the Boston area, you'd expect a--

JOHN GUTTAG: Let's assume that, in fact, instead of by square miles-- let's assume the populations were balanced.

AUDIENCE: Then I also think it's flawed because I don't think the importance of block 111 having 143 is important. I think the importance is just one area having a higher--

JOHN GUTTAG: Exactly right. Exactly right. I'm sorry, I forgot my candy bag today. Just means there'll be more candy for the final.

What we have here is a variant of cherry picking. What I have done in this simulation is I've looked at 1,000 different regions. What the attorney did is, not in a simulation, is he looked at 1,000 different regions, found the one with the most cancer cases, and said, aha, there are too many here.

And that's not what I did in my simulation. My simulation didn't ask the question, how likely is it that there is at least one region with that many cases. But it asked the question, how likely is it that this specific region has that many cases. Now, if the attorney had reason in advance to be suspicious of region 111, then maybe it would have been OK to just go check that. But having looked at 1,000 and then cherry pick the best is not right.

So this is a simulation that does the right thing. I've left out the initialization. But what you can see I'm doing here is I'm looking at the probability of there being any region that has at least 143 cases. What this is called in the technical literature, what the attorney did is multiple hypothesis checking. So rather than having a single hypothesis, that region 111 is bad, he checked 1,000 different hypotheses, and then chose the one that met what he wanted.

Now, there are good statistical techniques that exist for dealing with multiple hypotheses, things like the Bonferroni correction. I love to say that name. But you have to worry about it. And in fact, if we go back to the code and comment out this one and run this one, we'll see we get a very different answer.

The answer we get is-- let's see. Oh, I see. All right, let me just comment this out. Yeah, this should work, right?

Well, maybe you don't want to wait for it. But the answer you'll get is that it's actually very probable. My recollection is it's a 0.6 probability that at least one region has that many cases. And that's really what's going on with this whole business of people reporting cancer clusters. It's just by accident, by pure randomness, some region has more than its share.

This particular form of cherry picking also goes by the name of the Texas sharpshooter fallacy. I don't know why people pick on Texas for this. But they seem to. But the notion is, you're driving down a road in Texas and you see a barn with a bunch of bullet holes in the wall right in the middle of a target.

But what actually happened was you had a barn. The farmer just shot some things at random at the barn, then got out his paint brush and painted a target right around where they happened to land. And that's what happens when you cherry pick hypotheses.

What's the bottom line of all these statistical fallacies? When drawing inferences from data, skepticism is merited. There are, unfortunately, more ways to go wrong than to go right. And you'll read the literature that tells you that in the scientific literature more than half of the papers were later shown to be wrong.

You do need to remember that skepticism and denial are different. It's good to be skeptical. And I love Ambrose Bierce's description of the difference here. If you had never read Ambrose Bierce, he's well worth reading. He wrote something called *The Devil's Dictionary,* among other things, in which he has his own definition of a lot of words.

And he went by the nickname Bitter Bierce. And if you read *The Devil's Dictionary,* you'll see why. But this, I think, has a lot of wisdom in it.

Let's, in the remaining few minutes, wrap up the course. So what did we cover in 6.0002? A lot of things. If you look at the technical, things were three major units-- optimization problems, stochastic thinking, and modeling aspects of the world.

But there was a big subtext amongst all of it, which was this. There was a reason our problem sets were not pencil and paper probability problems, but all coding. And that's because we really want, as an important part of the course, is to make you a better programmer.

We introduced a few extra features of Python. But more importantly, we emphasized the use of libraries, because in the real world when you're trying to build things, you rarely start from scratch. And if you do start from scratch, you're probably making a mistake.

And so we wanted to get you used to the idea of finding and using libraries. So we looked at plotting libraries and machine learning libraries and numeric libraries. And hopefully, you got a lot of practice in that you're a way better programmer than you were six weeks ago.

A little more detailed-- the optimization problems, the probably most important takeaway is that many important problems can be formulated in terms of an objective function that you either maximize or minimize and some set of constraints. Once you've done that, there are lots of toolboxes, lots of libraries that you can use to solve the problem.

You wrote some optimization code yourself. But most of the time, we don't solve them ourselves. We just call a built-in function that does it. So the hard part is not writing the code, but doing the formulation.

We talked about different algorithms-- greedy algorithms, very often useful, but often don't find the optimal solution. So for example, we looked at k-means clustering. It was a very efficient way to find clusters. But it did not necessarily find the optimal set of clusters.

We then observed that many optimization problems are inherently exponential. But even so, dynamic programming often works and gives us a really fast solution. And the notion here is this is not an approximate solution. It's not like using a greedy algorithm. It gives you an exact solution and in many circumstances gives it to you quickly.

And the other thing I want you to take away is, outside the context of dynamic programming, memoization is a generally useful technique. What we've done there is we've traded time for space. We compute something, we save it, and when we need it, we look it up. And that's a very common programming technique.

And we looked at a lot of different examples of optimization-- knapsack problems, several graph problems, curve fitting, clustering, logistic regression. Those are all optimization problems, can all be formulated as optimization problems. So it's very powerful and fits lots of needs.

The next unit-- and, of course, I'm speaking as if these things were discrete in time, but they're not. We talked about optimization at the beginning. And I talk to an optimization last week. So these things were sort of spread out over the term.

We talked about stochastic thinking. And the basic notion here is the world is nondeterministic, or at least predictably nondeterministic. And therefore, we need to think about things in terms of probabilities most of the time, or frequently. And randomness is a powerful tool for building computations that model the world. If you think the world is stochastic, then you need to have ways to write programs that are stochastic, if you're trying to model the world itself.

The other point we made is that random computations-- randomness is a computational technique-- is useful even for problems that don't appear to involve any randomness. So we used it to find the value of pi. We showed you can use it to do integration.

There's nothing random about the value of the integral of a function. Yet, the easiest way to solve it in a program is to use randomness. So randomness is a very powerful tool. And there's this whole area of random algorithms-- research area and practical area that's used to solve non-probabilistic problems.

Modeling the world-- well, we just talked about part of it. Models are always inaccurate. They're providing some abstraction of reality. We looked at deterministic models-- the graph theory models. There was nothing nondeterministic about the graphs we looked at.

And then we spent more time on statistical models. We looked at simulation models. In particular, spent quite a bit of time on the Monte Carlo simulation. We looked at models based on sampling.

And there-- and also when we talked about simulation-- I really hope I emphasized enough the notion that we need to be able to characterize how believable the results are. It's not good enough to just run a program and say, oh, it has an answer. You need to know whether to believe the answer. And the point we made is it's not a binary question. It's not yes, it's right, no, it's wrong.

Typically, what we do is we have some statement about confidence intervals and confidence levels. We used two variables to describe how believable the answer is. And that's an important thing.

And then we looked at tools we use for doing that. We looked at the central limit theorem. We looked at the empirical rule. We talked about different distributions. And especially, we spent a fair amount of time on the normal or Gaussian distribution.

And then finally, we looked at statistical models based upon machine learning. We looked at unsupervised learning, basically just clustering, looked at two algorithms-- hierarchical and k-means. And we looked at supervised learning.

And there, we essentially focused mostly on classification. And we looked at two ways of doing that-- k-nearest neighbors and logistic regression. Finally, we talked about presentation of data-- how to build plots, utility of plots, and recently, over the last two lectures, good and bad practices in presenting results about data.

So my summary is, I hope that you think you've come a long way, particularly those of you-- how many of you were here in September when we started 6.0001? All right, most of you. Yeah, this, by the way, was a very popular ad for a long time, saying that, finally women are allowed to smoke, isn't this great. And Virginia Slims sponsored tennis-- the women's tennis tour to show how good it was that women were now able to smoke.

But anyway, I know not everyone in this class is a woman. So just for the men in the room, you too could have come a long way. I hope you think that, if you look back at how you struggled in those early problems sets, I hope you really feel that you've learned a lot about how to build programs. And if you spend enough time in front of a terminal, this is what you get to look like.

What might be next? I should start by saying, this is a hard course. We know that many of you worked hard. And the staff and I really do appreciate it. You know your return on investment.

I'd like you to remember that you can now write programs to do useful things. So if you're doing a UROP, you're sitting in a lab, and you get a bunch of data from some experiments, don't just stare at it. Sit down and write some code to plot it to do something useful with it. Don't be afraid to write programs to help you out.

There are some courses that I think you're now well-prepared to take. I've listed the ones I know best-- the courses in course 6. 6.009 is a sort of introduction to computer science. I think many of you will find that too easy after taking this course. But maybe, that's not a downside.

6.005 is a software engineering course, where they'll switch programming languages on you. You get to program in Java. 6.006 is a algorithms course in Python and I think actually quite interesting. And students seem to like it a lot, and they learn about algorithms and implementing them.

And 6.034 is an introduction to artificial intelligence also in Python. And I should have listed 6.036, another introduction to machine learning in Python.

You should go look for an interesting UROP. A lot of students come out of this course and go do UROPs, where they use what they've learned in this course. And many of them really have a very positive experience. So if you were worried that you're not ready for a UROP, you probably are-- a UROP using what's been done here.

You can minor in computer science. This is now available for the first time this year. But really, if you have time, you should major in computer science, because it is really the best major on campus-- not even close, as somebody I know would say.

Finally, sometimes people ask me where I think computing is headed. And I'll quote one of my favorite baseball players. "It's tough to make predictions, especially about the future."

And instead of my predictions, let me show you the predictions of some famous people. So Thomas Watson, who was the chairman of IBM-- a company you've probably heard of-- and he said, "I think there is a world market for maybe five computers." This was in response to, should they become a computer company, which they were not at the time. He was off by a little bit.

A few years later, there was an article in *Popular Mechanics,* which was saying, computers are amazing. They're going to change enormously. Someday, they may be no more than 1 and 1/2 tons. You might get a computer that's no more than 3,000 pounds-- someday. So we're still waiting for that, I guess.

I like this one. This is, having written a book recently, the editor in charge of books for Prentice Hall. "I traveled the length and breadth of this country and talked with the best people. And I can assure you that data processing is a fad that won't last out the year."

MIT had that attitude for a while. For about 35 years, computer science was in a building off campus, because they weren't sure we were here to stay. Maybe that's not why, but that's why I interpret it.

Ken Olsen, an MIT graduate-- I should say, a course 6 graduate-- was the founder and president and chair of Digital Equipment Corporation, which in 1977 was the second largest computer manufacturer in the world based in Maynard, Massachusetts. None of you have ever heard of it. They disappeared. And this is in part why, because Ken said, "there's no reason anyone would want a computer in their home," and totally missed that part of computation.

Finally, since this is the end of some famous last words, Douglas Fairbanks, Sr., a famous actor-- this is true-- the last thing he said before he died was, "never felt better." Amazing. This was from the movie *The Mark of Zorro.*

Scientists are better. Luther Burbank, his last words were, I don't feel so good. And well, I guess not.

[LAUGHTER]

And this is the last one. John Sedgwick was a Union general in the Civil War. This is a true story. He was riding behind the lines and trying to rally his men to not hide behind the stone walls but to stand up and shoot at the enemy. And he said, "they couldn't hit an elephant at this distance." Moments later, he was shot in the face and died.

[LAUGHTER]

And I thought this was an apocryphal story. But in fact, there's a plaque at the battlefield where this happened, documenting this story. And apparently, it's quite true.

So with that, I'll say my last words for the chorus, which is I appreciate all your coming. And I guess you were the survivors. So thank you for being here.

[APPLAUSE]