THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 5

HOW DATA GOES BAD by Chris Wild

Hello, everyone. Today we'll start with the question, how does data get to be bad? Here's a simple example.

It concerns arm spans versus heights. You may recognise this famous drawing by Leonardo da Vinci. If we were to graph arm span against height, what would we expect the graph to look like?

We know about people's basic shapes. They come in roughly the same proportions. We'd expect someone twice as tall to have about twice the arm span, half as tall to have half the arm span, et cetera. If this happened exactly, then all the points would fall exactly on a straight line. But people aren't exactly the same shape. So we'd expect something more like this.

Here's data on 30,000 New Zealand school students. The measurements were recorded by the students themselves. You'll see all sorts of strange patterns. Here is the pattern we'd expect. But here is a second rugby ball shape. What's that? Look at its centre. It's at roughly the same height value, but half the arm span value of the main shape.

What does that tell us? These children were doing half arm spans. Despite clear instructions to do whole arm spans, which included this photograph, they did this.

And then there are the stripes. Can you see a vertical stripe and a horizontal stripe at 100 centimetres? Those are the children who rounded their values to one metre. At a certain age, getting to be one metre tall is a big deal, which might explain some of this.

And then there are a lot of other more subtle rounding stripes as well, especially at the tens -- 120, 130, 140, 160, and so on.

So what's the point of all this? The point is that there are a lot of strongly visible patterns in this plot that have nothing to do with the real world of human body shapes. They are artefacts, artificial patterns caused by deficiencies in the data collection process. They have nothing to do with the real subject under investigation.

We're not looking at this, but something more like this. Most of these patterns were caused by a bad measurement processes that involved kids being kids. We know that there are a lot of things wrong with this data because we know a lot about arm spans, heights, and children. But if we were investigating something we knew almost nothing about, we'd probably have no idea that we were looking at artefacts, rather than facts, aspects of the real relationship.

Lessons.

Artefacts are artificial persons caused by deficiencies in the data collection process. No data collection processes are ever perfect, so in data analysis we're always struggling to distinguish between facts and artefacts.

Artefacts can be quite persuasive, especially when delivered in proficiently presented graphics. And humans are quite good at coming up with explanations for why the world really had to be that way.

The type of deficiency we've just shown was in the measurement prices. The distortions we showed were small, however, compared with some of those in the readings for this week. Next we'll show how a selection process can distort a relationship.

This is a graph of breathalyser and blood alcohol readings from people stopped by police on New Zealand roads. With the procedures used at the time, drivers with a breath reading below a lower limit (325) were under the limit and sent on their way without a blood test.

- 25% of those with borderline readings (between 325 and 375) had to have a blood test.

- Above 375, they were over the limit unless they chose to have a blood test, and the blood test showed them to be under the limit.
- 68% of those between 375 and 525 had a blood test.

- 27% of those over 525 had a blood test as well.

The graph shows breath alcohol versus blood alcohol for everyone who had a blood test.

Now, I claim that this gives a distorted picture of the relationship between breath alcohol and blood alcohol. I'll demonstrate using a computer simulation. Here we've generated a set of 4,000 readings to form a computer-generated population. This plot shows the relationship between their breath alcohol and blood alcohol values. The trend is linear. Here are randomly selected people to have blood tests according to the percentages and the real data. I've coloured the ones who got selected orange.

Here's the relationship between breath alcohol and blood alcohol for the selected people. Notice that the trend is now curved. Both the relationship and the trend for people who had blood tests, in orange, look quite different from the way they looked in the parent population.

The lesson is the process by which data gets into our data set can lead to a distorted picture of reality. This is a big problem with many data streams sourced from the internet, where no one knows what sorts of biasing filtering processes have affected what data ended up getting recorded and stored.

Data goes bad from two main causes.

- Bad measurement processes, measurements that do not actually measure what they are meant to,and
- bias, the process by which some things get recorded while others don't, can cause systematic biases.

Additionally, all sorts of things can go wrong when data is being entered and moved around on computers.

Real data is also full of holes, values that are missing for various reasons. If the people we have values for tend to be different from those whose values have gone missing, then we automatically have bias.

But why is all this so important? Garbage in, garbage out.

Unfortunately, we may not always recognise when our data is garbage. When the data's in, it is too late to do anything about it. The best protection against bad data is to design good data collection processes at the outset, that is before the data is collected. This brings us to the end of this video. Next time we'll talk about the problems in determining cause and effect.