1.4 Big Data: Wat en Hoe?

Jeff Templon

“Wat en Hoe” is Dutch for “What and How”, and this title is borrowed from a series of very practical language guides I made frequent use of during my travels in the days before Google Translate existed. I hope that this piece can serve a similar function, to explain in a very practical fashion, What Big Data is, What it is good for, How it can be used, and How you can learn more.

What Is Big Data?

Big Data, in the most general sense, is the enormous amount of data stored in digital form on computers connected to networks. The network is important, because it enables the enormous amount of data to be collected (e.g., computers at Vodafone collecting the location data from all mobile phones connected to their 3G network) as well as to be accessed by others (e.g., TomTom traffic information being accessed by thousands of navigation devices spread across the country).

There are a lot of definitions for ‘Big Data’ in circulation, partly because of industry hype — if you call it ‘Big Data’ then it sells better. Below I will describe two ways of understanding Big Data, and they are both linked to what is called ‘exponential growth’ as displayed in Fig. 1. For example, if you start with two cents and add two cents each day, this is called linear growth. Exponential growth is if you double your money each day. In both cases, on the second day you will have four cents instead of two; on the second day you will have either six cents (linear) or eight cents (exponential). At the end of the month, this is 62 cents (linear) or 21 million euros (exponential). Big data is, in many cases, growing exponentially.

One way to see Big Data is how big it is. Taking digital music as an example, music in mp3 format started to become popular in 1997, when the WinAmp program was introduced. In those days, an average computer disk drive could hold about 400 songs of high-quality mp3 music (2 GB). Nowadays a home computer can hold more like one hundred thousand songs — assuming you don’t already have your disk filled up with digital photos, also a novelty in 1997. However, many people now use streaming music services like Spotify to listen to music; these services offer around 50 million songs. This is too much data to fit within a single computer, and this is another common definition of ‘Big Data’ — your data is big if it no longer fits on a single computer.

At Nikhef, about 50 large-capacity storage computers are networked together to hold our share of the data from the experiments we work on at CERN and elsewhere in the world. Taken together, these computers at Nikhef have enough storage capacity to hold three copies of the entire (all 50 million songs) iTunes store. Although Nikhef operates one of the largest data repositories, there are hundreds of others scattered across the globe, leading to yet another definition used for ‘Big Data’: your data is big if it no longer fits in the same building.

Why is there so much interest in Big Data?

Big Data is certainly interesting just because it is ‘big’, as the digital music example shows. In the early 1990s, when I went on vacation, I used to carry two small briefcase-sized cases full of cassette tapes with me. Now a mobile phone, a pair of speakers and an internet connection takes up much less space and provides access to much more music.

Big Data is, however, much more than this, as the large amounts of data allow us to do many new things that were previously difficult, if not impossible, to do. The very first examples mentioned in this article (mobile phones and traffic information) make a good example. The mobile phone location data, collected several times per minute, can be converted to a mobile phone ‘speed’, especially for those mobile phones located in cars moving along highways and streets. Taken together, these speeds show how fast traffic is moving. If the speed limit on a certain portion of the A9 is 100 km/hr but the phones are moving at 40 km/hr, this indicates a traffic jam, and that information is sent out to connected navigation devices.

In the past, traffic jams were deduced mainly by looking at traffic cameras on major highways. There was a significant delay before this information reached drivers, since the traffic report was broadcast only a few times per hour on the radio. The modern systems include all roads, and are updated every few minutes;
they also include other information like gasoline prices, lists of nearby restaurants, and weather reports, having collected this data from other sources on the internet.

**Big Data and machine learning**
Perhaps the most exciting (and sometimes frightening) aspect of Big Data has to do with how it has changed how we program computers. Until recently, in most cases scientists, or other programmers, ‘wrote a program’ and the computer used this program to produce some output. So you put a program ‘in’ to the computer and you get data back ‘out’.

There is another way to do it: you put the data ‘in’ to the computer, and you get the program back ‘out’. The program you get out is the one that would have generated the data you put in. This program can then be put back ‘in’ and used for tasks to generate similar data that you don’t have yet. This is called Machine Learning and can best be illustrated by example.

Suppose you have a lot of text that was originally written in English, and had been translated into Dutch. You have computer files containing the text in both English and Dutch. You can give these files to a Machine Learning system, and it will give you back a program that would have been able to translate the Dutch text from the English. This in itself is not so useful, as you already had the Dutch text, but you could now give it a different English text and your program will translate it into Dutch.

The connection to Big Data comes because, in many interesting cases, Machine Learning needs a lot of examples in order to ‘learn’ a program like the translator. As an example, one can teach a speech-recognition program to make the proper choice between confusing words: did the speaker say “go to”, “go too”, or “go two”? In order to learn a program that gets the right answer 99% of the time, you need to give it billions of examples. This was unthinkable in the pre-internet age, so people were still trying to write the translation program in the old-fashioned way. Now with so much written text on the web, and much of it already translated, it is possible (even relatively easy) to get billions of examples. Big Data is making things possible which were not possible before.

Translation programs, recommendation systems at web stores like Amazon, the little boxes that appear by people’s faces on your camera (how did the camera know that this part of the picture was a face?), web services like Evernote that can recognise words inside scanned images: all these systems are based on Machine Learning systems. Some of the most advanced Machine Learning systems are called ‘Neural Networks’ and are built to mimic the way neurons are connected in the human brain.

I mentioned that this aspect of Big Data could be frightening. More on this later on, but one of the problems is that the Machine Learning ‘output’ program is only as good as the data ‘input’.

The financial crisis of 2008 was a surprise to most; with all the financial data available, and claims being made that economic theory was no longer needed, we could use Big Data to understand everything: why did nobody predict the crisis? One of the main reasons is: the machine learning programs that were used to predict how well the market would do, were produced from input data on the previous ten or so years. Those years were economically quite calm, so the learned programs had no example data on what a financial crisis looked like, and hence had no way of being able to spot all the warning signs.

**Some example uses of Big Data**
Most of what we call ‘Big Data’ is publicly available on the web, so anyone can use it. One use I personally like is, checking how well weather forecasters are doing. In the pre-web times, you could find the weather forecast for the coming days, but they never included data on how well they predicted yesterday, or for this day in history, how much did the forecast and the actual weather differ, on average? If you were a determined person, you could keep a notebook every day of the forecast and the actual weather, the more cities and more forecasts you try, the more of a chore this becomes.

Once the weather services started making web sites where current temperatures, and the forecasts, were published, anybody could write a small program that picks up the forecast once per day, and the temperatures a few times a day, for as many cities as they liked. It turns out that the weather predictions are generally better than a) guessing, b) saying “tomorrow will be just like today” and c) “tomorrow’s weather will be the historical average high and low for today”. This is true for the day after tomorrow, but just barely true for today next week … and the weather services may like to publish weather forecasts for two weeks, but after 8 days, the forecasts are actually worse than just looking at the historical weather.

Also how well the forecast predicts the weather, depends on where you get your forecasts from. The national weather services are generally pretty good; local radio and tv stations are the worst. This was unexpected, as these stations get their weather predictions from the national services. So why are they worse? Because these local stations adjust the forecasts to be ‘wetter’. They do this because people in general prefer a pleasant surprise to an unpleasant
surprise, so you will make your audience happier if you are wrong about rain more often than wrong about sunshine.

The weather example is from 2002, and was discovered by a computer-science graduate student, who was doing Big Data before it was called Big Data. A more recent example comes from one of our colleagues, prof.dr. Barend Mons, who works at the Leiden University Medical Center. Prof. Mons’ group has played a leading role in collecting all available biomedical data into a single unified framework. Fig. 2 shows a graphical representation of a research result extracted from that data collection. The research question was to find genes (yellow boxes) involved in both ‘inflammation’ and ‘HIV disease progression’ (red boxes). The black boxes represent published papers linking genes to either HIV or inflammation. The two genes displayed in the figure have many published papers that link to either of the diseases, but never to both diseases in the same paper. This link is only discovered when big-data analysis techniques are used on the entire collection.

How can Big Data keep expanding exponentially?
Big Data has continued to expand because the number of connected things continues to increase, and because the underlying technology — especially in networks — allows more and more stuff to be put online.

When the World Wide Web was started, first all high-energy physics labs were connected together; the Web was invented at CERN, and Nikhef was the third site to join worldwide. Very quickly many other academic departments and research institutes joined the web, and commercial sites started to join. In 1992 there were only about 15,000 .com sites, now there are several hundred million. The expansion continued as people (not companies) started to put more and more things online, think of Facebook, Instagram, Twitter, and the many Google services. In recent years, the number of smartphones (see Fig. 3) has been increasing rapidly, adding the location history of millions of people (and millions of ‘selfies’) to the available Big Data. One might think that with all companies online, all people online, and everyone using a smartphone, the expansion would be over; the next wave is to connect ‘everything’ to the internet. This wave has started with home thermostats; now there are internet-connected washing machines, cars, garage doors, home lighting, pet collars; you can even have your doorbell send you an email if someone rings it. In most cases, cloud services are involved which means that there is a company that receives this data on its servers and can process it to learn things about the owners. You might be able to learn a lot about a neighborhood, or a person, by studying how often, and at what times, the doorbell is being rung. Even more if you could connect that data to a Facebook profile, and use cell-phone tracking data to find out exactly who it is that is visiting when the doorbell rings.

Risks of Big Data
The situation just described makes it pretty clear that the more we get connected, and the more connected our data is, the less privacy we have. Even if the company you give your data to (doorbell, thermostat, cell phone location) is not doing anything shady with this data, they do have the data and it is connected to the internet. Lots of people who would like to misuse your data...
Getting the right answers

A short aside on using computers to process large amounts of data and make a decision: computers are very literal. They do exactly what you tell them to do. There is the old joke about a computer washing its hair: it runs out of shampoo, because the instructions say “apply shampoo; lather; rinse; repeat” — nowhere do the instructions say “stop and dry your hair” so the computer repeats until it runs out of shampoo. When you write a computer program, you have to get it exactly right — there are more than millions of ways to get the program wrong, and only a handful of ways to get it right. In the profit and loss case, it is pretty quickly clear whether the answer was right. In other cases, for example if you’re trying to predict something that won’t happen for several years, it might not be at all obvious whether your answer is correct. Unfortunately, some of the most important possible uses of big data are like this. Big data is a valuable tool, but we also need to have an understanding of what we’re trying to predict, and to be able to check whether the analysis makes sense.

are also connected to the internet. If the people you give your data to are not very careful about security, your data winds up in the wrong hands and can be misused.

A somewhat more subtle form of Big Data risk is called ‘targeting’ and already happens on a large scale. Because of all the Big Data present, there is no such thing as ‘the internet’. When you visit a website, a few pieces of information arrive with you, and in many cases this information is enough so that the website knows it is specifically you that is visiting. Via connected data from Facebook and Google, the website might have a good idea of your income and how much you could pay for a particular product, also of your interests and so how much you might want to buy that product. The website will present in some cases a higher price if it believes you can and will pay it.

These examples come from the cases of companies using your personal ‘Big Data’, or maybe coupling your data with that of similar people (like Amazon recommendations). There is a natural feedback at work to make sure that the companies are getting the right answer out of this Big Data: money. If the companies process the Big Data and use it to make decisions like what services to offer or products to produce, or to guess how much you might be willing to pay for a product, they had better get the right answer; if they don’t they will lose money, and if they lose enough they will go out of business.

A final risk of Big Data has already been presented earlier: how can we be sure that the results being presented, have been honestly presented without ‘fudging’? I refer here to the ‘wetter weather’ predictions. The average citizen can’t repeat a big-data analysis to check whether the numbers have been correctly presented — the best we can do is to check whether the numbers resemble the real weather (or economy, or expected lifespan, or unemployment rates).

How sure are we of what we know?

Recall the example given earlier about Machine Learning: that is when you give the computer data, and what you get back is a program that can compute similar things for you. Recall also that in order for the learned program to be really accurate, you need to have lots of data as input. There are a lot of questions:

1. Did you give your learning program enough data?
2. Did you give it the right data?
3. Did you write your learning program correctly? The program that turns your data into a program, is in itself a program, that might have mistakes in it.

Unfortunately the answers to these questions are usually not simple. You can answer 3) mostly by testing your program very well (see the text box “Getting the right answers”), but the answers to 1) and 2) require a good understanding of statistics. Companies selling you Big Data solutions usually leave this part out — they want to sell you the product, and you might not buy it if they tell you you need to spend months learning statistics before you can really understand what your Big Data analysis is telling you.

These problems can be illustrated with the following small-data situation: flipping a coin. You want to analyse coin flips to understand, what is the probability that a flipped two-euro coin will land on ‘heads’. Easy, right? You just take data on coin flips, calculate the percentage of flips that come up heads, and that gives you the probability. I just did this: I flipped a two-euro coin eight times; it came up heads 5 times, so the probability is 62%. Right?

This is certainly a case of question 1) above — not enough data. Basic statistics tells us that if I only flip the coin eight times, my estimate of the probability will very likely be off by as much as 18%, this is called the ‘uncertainty’ in the answer. It might also be a case of question 2) — I am trying to figure out the probability when flipping a two-euro coin, that is any coin, but I am only trying the one I had in my pocket — maybe this coin is bent or damaged so that it is different than most other two-euro coins.
For Big Data, question 2) can be a real problem, because the Big Data that is out there is very skewed in some cases. For example, there is a lot of current interest in analysing Twitter data for all sorts of reasons; a recent study tried to determine what was the happiest day of the year. However this analysis could only tell you how happy Twitter users are — recent studies show only about 20% of online adults use Twitter, and Twitter users are mostly below age 30.

There are unfortunately a lot of Big Data analyses being made by people with no background in statistics, meaning that there are a lot of Big Data conclusions out there for which we have no good idea if these conclusions are correct. When you want to analyse Big Data in this way, you move from Big Data to data science; Fig. 4 illustrates the relationship of ‘hacking skills’ (in this case, is the researcher good at dealing with very large amounts of data), statistics knowledge, and expertise in a field: all three are needed to do data science well. The ‘danger zone’ in the diagram illustrates a problem often encountered in Big Data studies: someone applied an analysis using a lot of data (= hacking skills) to some particular problem (= substantive expertise in some field) but without a good understanding of principles of statistics and the math needed to use those principles; such studies can be very powerful generators of, well, nonsense.

**Nikhef and Big Data (Science)**

In our research in high-energy physics at Nikhef, we deal with essentially all of the issues described above, and have been dealing with them long before ‘Big Data’ became a buzzword. High-energy physics was one of the first scientific fields to generate and analyse large amounts of data, and for a long time we had more data than pretty much anybody else. Only relatively recently, we’ve been overtaken in data size by a few companies like Google and Facebook.

From the beginning, our analyses have been aimed at getting scientific conclusions from Big Data. We learned — sometimes the hard way — how to avoid making ‘bias’ errors (wrong data) and how to understand the uncertainties in our answers related to how much data we have ... the uncertainty only really goes to zero if you have a basically unlimited amount of data (which means an unlimited amount of money, which we don’t have!).

My group at Nikhef has played a leading role in developing the Worldwide Computing Grid for the CERN physics experiments, and as mentioned before, we run a large data center ourselves here at Nikhef — we have the hacking skills in Fig. 4, and together with our physicists, we have the substantive expertise (in High-Energy Physics). All of our physicists know the basic rules and concepts of statistical analysis, and some of us are real experts; two Nikhef physicists contributed a chapter to a recently published book called “Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods” [1].

I mentioned earlier how important it was to keep certain kinds of Big Data secure, and my group at Nikhef works on this problem too. The main problem is, how to make it possible to share certain types of data between a large group of researchers located all over the world, without making this data public to everybody. It turns out to not be so easy, especially since most things we could do to make the data more secure, makes the data also harder to use for research.

Machine Learning is used in many different applications in high-energy physics experiments and their analyses. A good example is how our analysis programs can figure out, for all the thousands of particles it detects in a collision, what kind of particles (e.g. proton, pion, muon, electron, …) they are. We teach these programs how to identify the particles using machine learning (and there is a whole chapter on machine learning in the book just mentioned).
Finally, Nikhef produces data scientists! There are many indications that data scientists — the people who know enough of all three areas in the diagram above to be able to get trustworthy conclusions out of Big Data — will be in short supply in the coming years.

In a variant of the data science diagram shown in Fig. 4 the word data science in the middle is replaced by ‘unicorn’. The reason being is that the author of the diagram claims that people who are good in all three areas are as scarce as unicorns. All of our PhD students master the substantive expertise and statistical methods; many master the large-scale computer skills too. This is why many of our students go on to jobs in fields related to Big Data. For example, most of the big-data group at KPMG Netherlands have worked at Nikhef at some point in their career; one of these former Nikhef PhD students, dr. Sander Klous, was just named Professor of Big Data Ecosystems for Business and Society at the University of Amsterdam.

Further reading
I promised to tell you how you could learn more about Big Data. Nate Silver’s book “The Signal and The Noise”[2] is a very readable description of the types of things you can (and cannot!) do with Big Data; a few of the examples presented here are described in much more detail in that book. A discussion of basic probability and statistics, presented mostly without equations and using only high-school math, with lots of examples, is John Allen Paulos’ “Innumeracy”[3]. Finally, “Thinking Fast and Slow”[4] is a recent book by Nobel Prize laureate Daniel Kahneman, who talks about how the human brain works, why the brain is very bad at thinking statistically and probabilistically, and how easy it is to trick the brain with large numbers. This book reinforces, from a completely different viewpoint, the message made here: in the Big Data age it is increasingly important to have many people trained not only in working with the Big Data, but also interpreting, analysing, understanding, validating, and checking all that data — it is all too easy to draw wrong conclusions from data, intentionally or unintentionally.

By the way — I continued flipping that two-euro coin, just to see what would happen. After flipping another sixteen times, I was at 12 heads and 12 tails, giving the expected result of 50% heads. Although, statistics tells us, this could still be off by as much as 10% due to ‘not enough data’!

Bibliography