CYBAEA Viewpoint

March 2012

When Big Data Matters

Summary

Big Data is a hot topic at the moment, but does it matter?

This note provides an introduction for the business manager and gives a framework for determining when Big Data is relevant to your problems.

About CYBAEA

CYBAEA is the "more than analytics" company that not only delivers the insights from your data but also help you execute on that knowledge to deliver commercial results.

Visit www.cybaea.net

Luton office

960 Capability Green
Luton LU1 3PE
United Kingdom

London office

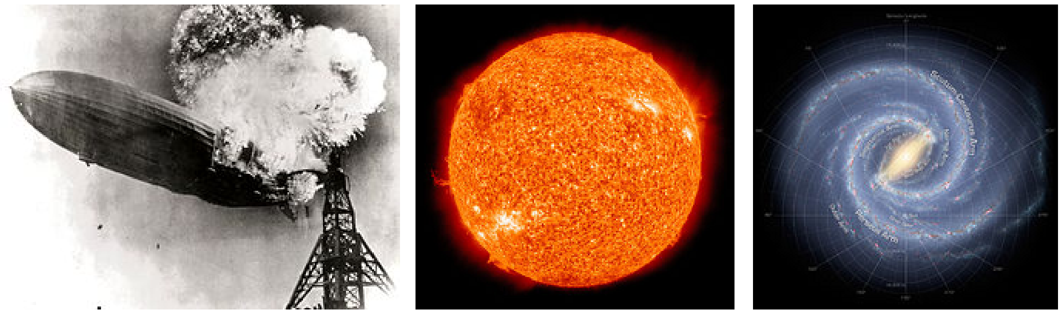
2nd Floor
Berkeley Square House
Berkeley Square
London W1J 6BD
United Kingdom

Big Data: hype or reality?

Big Data is a buzzword, but is it real: does it address real business issues or is it just an excuse to sell more computers, software, and consulting services?

We argue that it is real and it does matter, but only in some well-defined circumstances: it is not a universal solution or requirement to every problem. We provide a framework for determining where the Big Data applications are within your work and where traditional approaches apply.

Big can be a qualitative as well as a quantitative difference. The gas in the ill-fated Hindenburg airship, the gas that formed our Sun, and the gas that formed the Milky Way galaxy were just lumps of hydrogen atoms (with varying impurities[1]). The difference was in the number of atoms. But that difference in numbers made the three structures into different things. You simply cannot look at them in the same way. If you try to model the galaxy in the way you model a balloon you will fail.



Just a bunch of hydrogen atoms: when "big" makes a qualitative difference.

Back to planet Earth, the early years of this millennia saw an explosion of popular books and articles about the science of networks[2], which was, and continues to be, a subject of intense research. Networks are interesting, but the main point we will take here is that the network is fundamentally, qualitatively, a different thing from its constituent parts, and it can not be understood from just understanding the parts.

The canonical example from this time was inspired by the widespread power outage in north-eastern America on August 14, 2003, which affected an estimated 55 million people, then the second-most widespread failure of an electrical distribution network in history.

The components of electrical networks are well understood. Copper wires, transformers, generators, and the like have been studied for decades, if not centuries. Mr Maxwell had published the necessary mathematical equations by 1862. We know how to solve them and our calculations agrees with what we observe to a ridiculous 10 decimal digits[3], the limit, it should be noted, being our ability to observe accurately, not our ability to calculate precisely.

[1] Our sun is about 1/4 Helium, but it does not matter for this discussion: other stars are formed from almost pure hydrogen and they are not balloons or galaxies either.

[2] Perhaps the three most popular books are listed below; our review of the first is on the web site.
Barabasi A-L. 2002. *Linked: The New Science of Networks.*
Buchanan M. 2002. *Nexus: Small Worlds and the Groundbreaking Science of Networks.*
Watts DJ. 2003. Six Degrees: *The Science of a Connected Age.*

[3] The measurement of the so-called fine structure constant is one of the most accurate known to science and has an error of about 0.3 parts per billion. Technically, this one is computed from a successor to Maxwell's equations known as QED.

But put all of these components together in a network, and we are unable to really understand or predict extreme events in the system, such as the cascade that led to the 2003 outage. (I say this with with the greatest respect to the people building and maintaining power grids who are some of the most awesome engineers I have ever met and worked with. The failure to understand is because we are looking at the wrong things and because the right things are inherently difficult.)

The human brain is a network of relatively well-understood nerve cells, and yet we do not understand consciousness, perhaps the most obvious feature of our minds. The weather system can be considered a large scale application of the well-understood laws of thermodynamics, but long term weather forecasts are still beyond us.

The human brain is a network of relatively well-understood nerve cells, and yet we do not understand consciousness, perhaps the most obvious feature of our minds.

The list goes on. The network is fundamentally different from its parts.

But you can just sample….

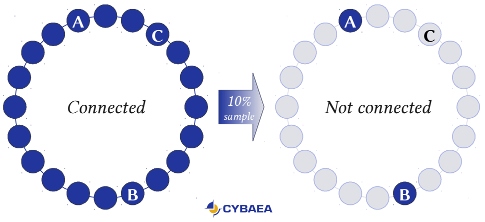
One objection to Big Data we often hear goes something like this: "*If you select a sample of the data, and the sample is good and truly random, then you can do all your analysis on the smaller set thereby avoiding all the problems of Big Data.*"

It is true that for many problems you can indeed sample, and it is a well understood statistical technique. If all you are doing with your data is a generalized linear model on a few categorical variables, then go ahead and sample to your heart's content.

But how do you sample a network? We did some work with a mobile telephone company where we were interested in understanding if, and under what circumstances, when you leave the network for a competitor your friends leave with you. The data was somewhat over 10 million current and recent subscribers with information about their usage, spend, and "calling circles": who did they call and how much, which was our definition of "friend" for this purpose.

We found that there were tipping points when a predicable and quantifiable part of a sub-network of customers had left, a large proportion of the remaining network would also leave. Intuitively you can understand why: if your friends or colleagues are all with this other company and love it and get cheap or free calls to each other, but not to you, then there is both a social and economic pressure for you to switch.

Consider if we had taken a random sample of customers, say 1 million, for our analysis. Now the network of friends is typically no longer connected and therefore not a network we can identify. To see why, consider a naive (and unrealistic for this application) network where each person is connected only to his two neighbours, as in the figure below. In this scenario, A and B are connected through two chains of nine friends.



After sampling, A and B are no longer connected

But if I take out 9/10 of the sample, then A and B appear very far removed. And if C is the trigger for the whole circle to leave, then I have lost my ability to predict.

To be fair, there are other ways of sampling a network. If we had looked not at customers, but at their calls we could have sampled those and rebuilt a reasonable representation of the network[4]. But that means looking not at 10 million customers, but at perhaps 10 billion call detail records: hardly the way to avoid Big Data.

When Big Data matters

First, Big Data always matters when you are looking at the whole rather than the parts, and when that whole is a different 'thing' from the parts. A nation is more than a group of people, a city more than a bunch of houses.

First, Big Data always matters when you are looking at the whole rather than the parts, and when that whole is a different 'thing'

[4] It works for this problem: we have compared results.

When you have data on people and houses but want to know about nations and cities then you are into Big Data territory.

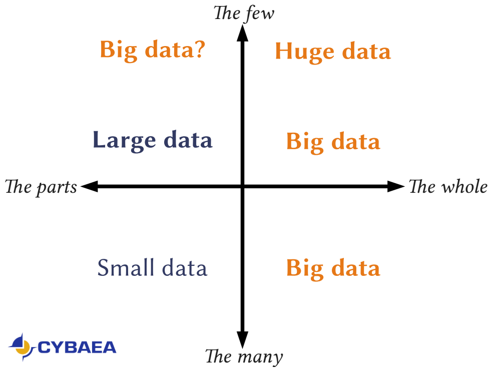
Second, Big Data *may* matter when you are looking at the few rather than the many

Second, Big Data *may* matter when you are looking at the few rather than the many. In applications like fraud detection the signal may be so small that you miss it, or miss too much of it, if you sample. Whenever you are looking for 'all' or 'every' or for 'the largest' or some other superlative you are into at least large data territory. (If you are looking for the 'mean', the 'typical' or anything like that, you are looking at aggregates of the many, and you do not need Big Data.) The canonical 'hello world' example for Hadoop, a popular Big Data tool, is counting the frequency of every word in a text corpus; 'every' being the key word here.

Anything involving a slice of a power law is firmly top right corner of our 'When Big Data matters' model

However, for many practical applications in this second territory you can get away with some sampling or else the raw data just is not that big to start with. In either case, you are looking at what we might call 'large data', by which we mean data that is manageable on, say, a large computer or an in-house cluster using what is essentially classical analysis techniques and tools.

When Big Data matters



In summary, our rough guide to when Big Data matters look somewhat like the illustration above. When you are looking at the whole (and the whole is a different thing from the parts) you are in big data territory; when you are looking for the few among the many you typically do not need big data techniques though your data may be large; and if you are looking for small signals in the whole, you may be into truly huge data.

Business examples

Big Data may be relevant to you if:

1. You are looking at groups of people "spontaneously" acting together. Examples include predicting the next runaway best-seller that everybody suddenly wants to read or predicting which of a customer's friends will leave when she does, as in our mobile telecommunications churn example above. This falls into the bottom right corner of our 'When Big Data matters' illustration. Pretty much anything "social" involving humans fall into this category, which is why Facebook uses Big Data in a (apologies) big way.
2. You are looking at the network and its structures or behaviours. We looked at discovering the top influencers for a company: those people who's opinion about you, or your products or services, make a big difference on the general perception of your brand. We trawled through blogs, comments, purchases, interactions, and much more data to determine these key people and how their influence rippled through the network. Anything of this nature – anything involving the top end of a power law – is firmly top right corner of our 'When Big Data matters' model. Google is as much interested in how the web pages are linked together as in what they say, which is one reason Google is a Big Data pioneer.
   - Another example is from our work on quantifying brand perception: you know who buys what and when, and you perhaps know what is being said about you in media and on social sites, but pulling all of that together to a picture of how your brand is perceived: that is a complex Big Data problem (but unlike the previous example not a Huge Data problem because we were just looking at the perception, not directly trying to identify the influencers).
3. Your data is at the level of the parts, but the structure that you want to investigate is more than the parts and you are not sure in

advance how he parts interact. These are the houses/cities and people/nations type problems. Any random sample of the UK population would almost certainly miss the members of parliament (0.001% of the population) unless you knew a priori that they were important for your analysis and therefore skewed your sample.

- These problems are *hard*, because you may not recognise the extent of your ignorance of the behaviours in your data until after you have done the analysis.
- Examples include the power distribution network problem we mentioned on page 1 and much analysis in the social sciences.

4. We already mentioned fraud detection as an example of the top left corner in our 'When Big Data matters' model. They are large data problems because the signal is small and you can't afford to sample until you have isolated that signal. However, they are often not Big Data problems because as soon as you have identified the small signal, you can sample the rest of the data for the comparison.
5. Where fraud detection becomes more interesting and the problem moves from the left towards the right of the model is when you are looking to identify collusion. This is also applicable for many compliance applications like insider trading or policing Chinese Walls within an organization. With one insurance company we looked at identifying staged accidents, but more successful examples come from the casino (Gambling) industry where the casinos successfully cooperate with each other to prevent fraud[5] while still protecting confidential information about their "whales" and other customers.[6]

Dealing with Big Data

The reader will have noticed that we have so far avoided defining 'Big Data'. For our purposes we use the following terminology:

1. "Big Data Sets" usually refer to data that is too large to handle using traditional relational databases and is inefficient to analyse using non-distributed applications.
2. "Big Data" is the industry of managing, transforming, and analysing these big data sets, and the ecosystem of tools and techniques that this uses.
3. Often *not* included in Big Data is the exploitation part which is our obsession: how to deploy the insights from data (big or otherwise) back into the business, its processes and its people, and execute on those insights.

There is not doubt that there is much hype about Big Data at the moment. But that just means we need focus. Big Data is real and addresses real business problems, though it is not a universal solution or requirement to every problem. This briefing should have given you a better understanding of the types of problems where Big Data will deliver business value.

Big Data is not scary any more. For sure you have to approach it different from "small data", but we now have the tools, the infrastructure, and, perhaps most importantly, the techniques and the design and analysis patterns to handle very big data. And we understand how to exploit the insights from Big Data analysis within our business processes to deliver measurable value and sustained competitive advantage.

Until a few years ago, we had the excuse that Big Data was too hard and too expensive for all but academia and a few businesses. That is simply no longer the case: large and Big Data is for everyone. And those pioneers, companies like Walmart, Tesco, and Google who are the most successful in their industries, have shown us how data can be a sustained competitive advantage.

What is your Big Data opportunity? What are you waiting for?

These problems are *hard*, because you may not recognise the extent of your ignorance until after you have done the analysis.

There is not doubt that there is much hype about Big Data at the moment. But that just means we need focus.

Big Data is real and addresses real business problems, though it is not a universal solution or requirement to every problem.

About the author:
Allan Engelhardt is a Partner with CYBAEA who got his introduction to Huge Data while working on experiments at CERN, the European centre for high energy physics.
Since he left academia two decades ago he has worked primarily in CRM, Sales, and Marketing, in both consulting and interim management roles. He is a recognized visionary by industry analysts and more than a little obsessed with delivering commercial results for companies.

[5] Fraud and other undesired activities such as card counting which is not actually illegal.

[6] Sometimes branded as 'Big Data' are problems where speed of response is critical. This is effectively a third dimension to our model, and where real-time responses are needed it can have a massive impact on the approach. We have much experience in augmenting call centre processes with real-time decision support tools but we will leave this subject for another note.