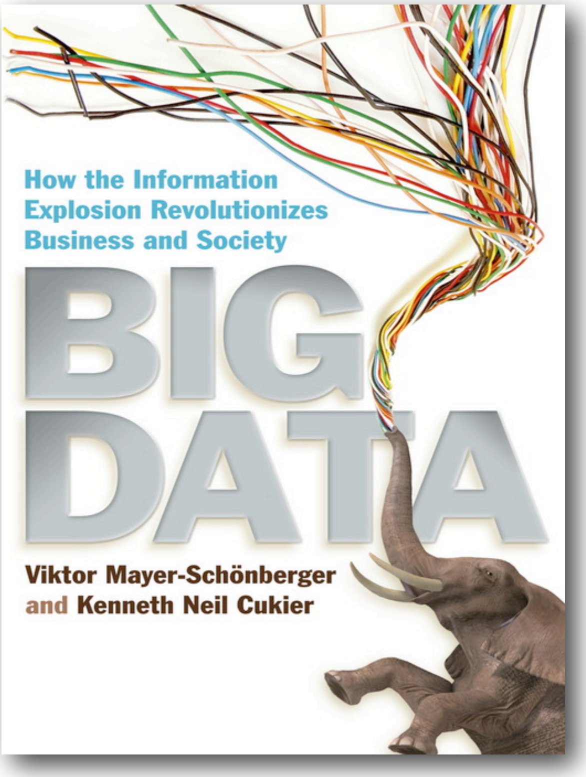
Big data - a revolution that will transform how we live work and think

by Vikto Mayer-Schonberger and Kenneth Cukier


How the Information Explosion Revolutionizes Business and Society
BIG DATA
Viktor Mayer-Schönberger and Kenneth Neil Cukier

THE BOOK IN A NUTSHELL

Big data will change the way we live, work and think. Big data is based on the ability to harness information in novel ways to produce useful insights and services of value. This big data will become a critical source of economic value and a source of innovation.

We still have only just scratched the surface over what big data can do (not that it is a perfect elixir).

THE BOOK

Google published a paper in Nature claiming that they could predict the spread of flu having analyzed 50m search terms and then run 450m different mathematical models. In 2009, their model was more accurate and faster at predicting the spread than government statistics.

The explosion of data

There has been a dramatic increase in the amount of data. When the Gutenberg press was invented there was a doubling of information stock every 50 years. Information now doubles every three years. Big data analysis has been made possible by three technological advances: increased datafication of things, increased memory storage capacity and increased processing power.

Oren Etzioni of Farecast took big data files of airline ticket prices relative to days before the flight - so it was able to calculate the optimum time for flight purchase. it crunches 200bn flight price records to make its predictions, saving passenger an average of $50 a flight. Microsoft eventually bought the company for $110m and integrated it into Bing.

When the Sloan Digital Sky survey began in 2010, there was more data in the first few weeks than had been collected over the entire history of astronomy. By 2010 is had amassed 140TB of information. In 2016 a new telescope will come on stream which will acquire the same amount of data in just 5 days. Likewise, the CERN particle physics laboratory in Switzerland collects less than 0.1% of the information that is generated during its experiments.

When scientists first decoded the human genome in 2003, it took a decade of work to sequence 3bn base pairs. Now a single facility achieve the same in just one day.

On the stock market, 7m stocks are traded each day. Google produces more than 24 petabytes of data a day. Facebook gets 10m new photos loaded up every hour. There were 400m tweets a day in 2012.

The power of prediction

One of the key benefits of analysis of this big data is predictions in the form of probabilities - be it engineering failure, financial failure or health failure.

At BP's Cherry Point refinery in Blaine, Washington, it constantly monitors temperatures, stress etc on all its pipes. By collecting such data it is able to predict in advance any stress points that could prevent a major issue.

Aviva predicts likelihood of illness by cross comparing data over a wide range of lifestyle data (such as sports, TV viewing, websites used and income) which led to a *predictive model to help them more quickly assess an individuals health risks without needing to have a full medical.*

Target analyzed buying habits of pregnant women and identified 24 products that they used as 'indicators' that a person is pregnant (e.g. unscented lotions around 3 months, and Magnesium and Zinc at 14 weeks), allowing them to target relevant and timely offers.

UPS has 60,000 vehicles it relies on. It now has sensors mounted on critical parts to anticipate failure without wholesale changes to every vehicle, saving UPS millions of dollars a year.

Facebook likewise learned that the the biggest predictor of an action (be it posting a comment, a like etc) was if their friends had previously done so - so it changed its layout to make people's actions more visible.

Improved decision making

Big data will increasingly be used as the primary default mechanism for many decisions as it increases accuracy and reduces irrelevant influences. Eric Brynjolfsen at MIT's sloan school found that companies that excelled at data driven decision making found their productivity was 6% higher than those who do not emphasize empirical judgment making.

Previously, one had to take small samples of data because it was impossible to collate let alone process it all. Even though a random sample of 1,100 people has been shown to provide fairly robust statistical reliability, it starts falling apart when we want to look in depth at sub categories.

The other advantage of 'n=all' is it shows up correlations that would not appear under normal circumstances.

Walmart analyzed sales during hurricane periods and found a large upsurge in sales of PopTarts (so on future such occasions, they were able to capitalize on this consumer need and placed Pop Tarts close to checkouts).

Analysis of 'outlying data' of credit card transactions exposed a fraud ring in New Jersey. Likewise match fixing in Sumo wrestling was uncovered by analysis of big data.

The trouble with small data samples is one develops a hypothesis in advance and then tests against it. The reality is we are flawed decision makers often with an unconscious bias towards the interpretation of data. By analyzing big data the computer in unbiased and will find what it finds.

> *The film Moneyball depicts how big data outperforms human instinct. Rather than relying on 'intuition' and 'experience' (which is often unconsciously influenced by other irrelevant factors - like in this case a player's girlfriend or his swing), it relies on amassing a lot of clean data that has been uninfluenced by emotions and prejudice.*
>
> *Huffington Post, Gawker and Forbes rely on data to help them decide the top stories - not just editorial sentiment.*
>
> *Jeff Bezos got rid of his internal book reviewers at Amazon as he found hard data drove bigger sales.*

Big data now pulls into question the quality of decision making of 'experts'. Big data outperforms experts (cf computers versus chess masters) because a computer can gain the experience of 20 years in 2 minutes (whilst still remaining clean).

The messy nature of big data

Big data is often messy and incomplete (only 5% of all data is structured). But the sheer scale of data compensates for this lack of precision e.g. in a Vineyard, measuring the temperature with just one sensor once a day will be much less accurate than 100 sensors taking readings every minute.

> *Manhattan has 94,000 miles of underground cable, served by 51,000 manhole covers. Sometimes, these 300lb iron cover explode into the air unexpectedly reaching heights of several stories before crashing to the ground. The city wanted to discover ways of predicting which ones were likely to 'erupt'. The data was extremely messy as the records went back to 1880 and there was no set way that engineers recorded their findings. Through analysis of different streams of data they were able to identify the key factors that correlated with explosions so could focus their attention on those key areas.*
>
> *The bureau of labour statistics surveys 80,000 prices to calculate a retail price index. However it appears a few weeks later. Economists from MIT set up PriceStats. It uses software to pick out around 0.5m prices of products advertised on-line. It now collates similar data in more than 70 countries. Whilst the information is messy, it's fast, providing real time data which may organisations subscribe to and use to help them make decisions. indeed the Economist relies on PriceStat's inflation numbers rather than Argentina's own data bank.*

Correlation does not equal causation

The data merely shows what it is, but cannot explain why it is. Big data is about the '*what*' not the '*why*'. The trouble is we have a flawed processing system and often see causal links when none exist. By freeing us from trying to create these causal links we can see the data more clearly and allow the data to speak for itself. Thus correlations are discovered that one would never have looked for.

> *Albert-Laszlo Barabsi analyzed all the phone calls of 1/5th of a country's mobile phone calls over a four month period. He discovered that people with a lot of links, were not the ultimate 'connectors of groups, but it was those on the outside of groups (who connect between different groups) were the key to information transfer across a network.*
>
> *Amazon used to cluster their customers into similar groups and told each of them about the other purchases they had made. The trouble was this led to lots of illogical recommendations based on the assumption that they would be similar in other areas as well. They then realised it was better to make recommendations simply based on associations among the products themselves - so if another book (or toaster for that matter) correlated with purchase of a certain book, then that would be highlighted. This approach led to a significant uplift in sales. Today 1/3rd of all Amazon sales are driven by this recommendation system. Likewise for Netflix, 75% of its sales comes from similar recommendation system.*

Datafication

Datafication is the unearthing of data from seemingly undatafiable sources. The reality is these days almost anything can be datafied - from pressure points across a retail floor, through to measuring sleep patterns via our mobil phones. In 2009 Apple were granted a patent for collecting blood oxygenation levels, heart rate and body temperature from the earbuds connected to their iPhones. Likewise, GreenGoose have developed tiny movement sensors that can be put onto a pack of dental floss providing real behavioural data as opposed to claimed usage.

> *Google has 'datafied' the written word from all the texts they are currently scanning. This has allowed the birth of culturonomics - the study of a culture through its literature. By datafying the texts it allows people to understand the most common words used at the time, when words were first used and the sentiment of the time. Thus only by turning the printed page into data could that latent value be accessed.*
>
> *UPS uses geo-location data to optimize its drivers' routes and to know exactly where every driver is at any time. In 2011 UPS shaved 30m miles off its driver routes, saving 3m gallons of petrol and 30,000 tonnes of CO_2. It also improved safety by reducing the number of turns across cross sections (which they found leads to greater collisions as well as saving fuel whilst waiting for gaps in the traffic).*

We are also seeing the datafication of people, and their relationships. Facebook's 'likes' have datafied sentiment but the rich data of all the personal interconnections provides a great source of analysis - Facebook's user base of 1bn represents 10% of the entire world's population - no other database has as much information about people and their interconnections. Whilst they have been very cautious about exploiting this data, the information that lies within it will have help us understand people, their relationships and societies.

> *Two hedge funds started analyzing tweets as signals for investment in the stock market which they now sell onto traders. MarketPsyche, for example have developed 18,864 indices across 119 countries that monitor minute by minute the emotional state of a nation.*

> *One company have shown a correlation between speed and volume of tweets for a new film and its likely success.*

The big value of big data

Data is increasingly becoming a major asset for an organisation (and could soon be considered a real asset to be costed onto the organization's balance sheet).

People are constantly finding new uses of this data, with an increasing number of data brokers springing up to help maximize the return from an organisations latent database.

> *As mobile phone revenues start to stagnate, they will turn more to releasing the business potential hidden within their data. For example Telefonica set up a separate company in 2012 called Telefonica digital insights to sell anonymous aggregated data to retailers.*

One of the biggest opportunities is the using of data for secondary uses.

> *Captcha was set up to prevent spambots by getting people to do something that humans find easy but computers find very hard. When relaunched as Recaptcha the second word people were asked to decipher was taken from digitized texts that was not clear what the word was. By using 200 million interactions each day they were able to provide correlated answers to questions that otherwise would have cost $1bn a year to achieve.*

There appears to be four ways of polishing the data diamond: Reuse (i.e allowing other people to see your data); Merging (i.e cross comparing a number of different data sets as sometimes the real value of data is best released when combined with other data), Combining (i.e. where build bigger data platforms from a number of different sources) and Twofers (i.e. The primary owner of the data finds a secondary use of the data):

> *Reuse: The Bank of England uses open sourced house price data to get a better sense of the economy at the ground level.*

> *Reuse: Mobile phone companies have massive access to geo-loco data which they sell to companies like AirSage. They crunch 15bn geo-loco records daily from cellphones to create real time traffic reports in over 100 cities across America.*

Reuse: Governments have historically been the most prolific keepers of large data. The only issue is they tend to be very slow at analyzing and releasing the data. When Obama came into office, he started an initiative that opened up the government data (www.data.gov), on the basis that outside firms could more efficiently and effectively do the data analysis. This has led to the release of 450,000 data sets across 172 agencies within 3 years. Fly on time.com was only made possible by the release of such data streams as it combined government flight data (scheduled departure versus actual departure) and published weather data to predict the likely delays by airport and weather event (e.g. Boston to La Guardia suffers 2x as long delays for fog as for snow).

Reuse: There has been a growth in companies who now 'broker' other company's data sets (such as Factual and Microsoft's Azure marketplace). Twitter for example has licensed two brokers to sell-on their data to others.

Merging: IBM merged data from Honda, Pacific Gas & Electric company - even the weather to predict the best places to install car charging points.

Merging: In New York, they used aggregated data across a very wide range of sources to predict which apartments in Manhattan which were more prone to overcrowding, fire and health risks. This data included the quality of recent external brickwork. This led to successfully identifying buildings at risk from 13% to 70%.

Re-combining: The Healthcare Cost Institute combined the data banks of a group of different health insurers to produce a more robust understanding of the insurance claims behaviour of 5bn claims from 33m people. It allowed them to spot trends their own individual data failed to show up - for example they found health costs in emergency rooms had risen by 11% whilst nursing facility prices had declined.

Re-combining: Inrix analyses data collected from over 100m BMW, Toyota and Ford cars (along with some commercial fleets and taxi's) and then combines this data with the driver's mobile phone data. It crunches all this data to predict traffic flows which it then sells back to fleets as well to the government.

Re-combining: UPS sold its data analysis unit to a private equity firm as they realised an independent company could aggregate data from other sources, creating even greater insight for UPS than just holding onto their own data set.

Twofers: Google collects street data from its photo cars to not only improve its maps but is also being used by them to develop their driver-less car.

So what is this data worth? Facebook's floatation valuation was $104bn yet in pure financial/accounting terms, the company was just worth $6.3bn. This massive discrepancy is due to the perceived value of its information. Facebook has over 2.1 trillion pieces of 'moneterizable' content. The floatation price valued every customer to have a book value of $100. This starts to suggest that the old fashioned way of valuing a company is no longer adequate in the current big data world where such assets are not properly valued in company accounts. The value of such 'intangible assets' (which also include 'goodwill' and 'brands') has grown from an estimated

40% of the value of publicly traded companies in the 1980's to more than 75% in 2000.

The value of those who hold the data is increasing. For example in 2006, Microsoft purchased Farecast for $110m, but just two years later, Google paid $700m to acquire Farecast's data supplier.

Finding new uses of data exhaust

When we have completed a search we believe that is the end of the usefulness of that data. Google realized that people's detritus (called data exhaust) still had use.

> *Ebooks captures masses of data about the literary preferences of people and how they use the service - e.g. what sentences they underline, notes they make in the margins, how long they take to read a book, a chapter etc for example. They can then use this data to pass back to the publishers for further improvements.*
>
> *Online education programs like Udacity, Coursera and EdX track the performance of their online courses (e.g. dwell rates through to quality of answers given in quizzes) to constantly improve their offerings.*

There is so much untapped potential from the data that surrounds us. For example, cars have on average 40 microprocessors inside them these days, but only a tiny percentage of their data is being reused.

The developing areas of influence

There are three groups who are at the heart of the development of big data: The data owners; the data analysts (who convert it into useable information) and finally the big data entrepreneurs (who spot new uses that other people are blind to).

> *Data Analysts: Mastercard collects 65bn transactions from 1.5bn customers in 210 countries. They have set up a separate company called Mastercard Advisors which analyzes this data and then sells on useful knowledge to others. For example, it found that motorists who fill up around 4pm will often spend $35-$50 in a supermarket within the next hour. Selling this information to a gas station allows them to potentially take a share of that spend.*

Often organizations who own the data are too close to see the extended uses of the data - outsiders can see what others cannot. It took a hedge fund team, not Barnes & Noble to spot the potential of an online bookstore. Likewise it took a software engineer not Sotheby's to develop eBay. A physicist developed the algorithm to predict insurance claims and identify defective cars, whilst an actuary developed a model to predict biological responses to chemical compounds.

Indeed, the use of data will increasingly become a key competitive edge (or competitive disadvantage) as we move into a data collecting world. Those with the greatest access to data will win out against those who do not. So the likes of Google, Facebook and Amazon will continue to soar. The previous big corporate players like P&G, GE, Boeing, Nestle and Walmart will need to find new ways to start collecting more data. Apple's move into the i-phone brought them access to a whole new stream of data. They demanded from the airtime contractors that the data collected be passed onto them. Thus Apple has data from millions of people around the world on a minute by minute basis.

The risks of big data

In East Germany, 100,000 Stasi Police employees accumulated 39m index cards and 70 miles of documents. But this is a drop in the ocean compared to the current level of surveillance we are under. There are more than 30 cameras within 200 yards of the London apartment where George Orwell wrote 1984. Every phone call, every transaction and every click of the mouse is being recorded.

The essential point about big data is that change of scale leads to a change in state. It is now possible to even judge (and punish) people before the very act has taken place.

Aggregated health data is used to predict the likelihood of illnesses within people for which they have to pay a premium now. Parole boards use big data to predict likelihood of re-occurrence, and so impacts on a person's chance to be freed.

In Memphis, they use historic crime data by area and time to schedule police patrols.

In the US they are rolling out smart electricity meters that can record data up to every 6 seconds - so it knows your movements based on your electricity usage. What is more every electrical item has a different power signal - so it can detect the difference between a laptop and marijuana lights.

Many companies claim that the mass aggregation of data guarantees anonymity at the individual level. However quite a few examples have demonstrated that this is not true (e.g. AOL and Netflix). With enough data, one can pinpoint an individual. So Big data rather than hiding people's identity actually helps uncover it.

Big data is not the magic elixir for everything. Robert McNamara (US secretary of defense during the Vietnam war) believed he could tame the complexity of war via analysis of big data - but as history has shown, such a strategy was flawed.

CRITICISM

This felt to be one of the most significant books I have recently read. It’s an easy read and is full of great examples. Whilst any book becomes rapidly out of date, it still awakes us to the power of data and how much it will affect our lives.

That said, we should not blindly listen just to the data (cf Robert McNamara). If one just loses oneself in the macro data one can miss the holistic picture. Thus we need to combine the analytics of data with our creativity, sensitivity and humanity.

PAC

Skilled experienced facilitator

If you are looking for an enjoyable,
yet effective away-day
(be it brainstorming, vision &
values, strategic brand building or
teambuilding) maybe I can help?

Paul Arnold Consulting
PLANNING - FACILITATION - TRAINING
07768 775988 paul_arnold@me.com