Big Data in Business
Professor Raghavendra Rau
27th February 2023

Introduction

In this lecture, I would like to talk about big data and how it is being increasingly used in business. Examples of how business uses big data are everywhere. Some of the most popular and famous examples include Target, Facebook, Farecast, Google Flu Trends, Cambridge Analytica, Amazon, and Netflix.

Target's use of big data was first described by Charles Duhigg in an article in the New York Times in 2012 and later detailed in his book *The Power of Habit*. As Duhigg describes it, new parents are like the holy grail to retailers. The reason is because once consumers' shopping habits are ingrained, it's incredibly difficult to change them. Think of yourself going to a grocery store. You buy milk, eggs, bread almost on autopilot. You don't have to think about where to go to buy any of your everyday items. But that makes you a very sticky customer. You will not easily move to another supermarket because you don't know where any of the items are in those supermarkets.

But there's one point in your life when everything is in flux. That is when you have a baby for the first time. At that point, you're doing things you've never done before. You're searching for things you never searched for before. But a supermarket cannot wait till you actually announce the birth of your baby because by then you already inundated with ads from other retailers about all the baby products you might need. It needs to figure out whether you're going to have a baby *before* you actually have the baby. Then if it persuades you to start shopping in their supermarket before you have the baby, for example if you're buying prenatal vitamins or maternity clothing, there's a good chance your sticky behaviour will keep you there even after you've had the baby.

So Target threw in an enormous amount of resources to predicting whether a woman would have a baby or not. In his article, Duhigg describes the sheer amount of data Target tried to capture on people who were shopping at Target and using their loyalty cards. It then used this data to predict whether a woman was going to have a baby well before the baby was born. What was especially interesting is that Target was even apparently able to predict whether you were going to have a baby before *you* knew that you were pregnant. For example, perhaps you were buying larger quantities of unscented lotion or loading up on supplements like calcium, magnesium and zinc, for you, that might be because you could no longer bear the scent of your regular lotion but for Target, it was a change in your shopping habits that might predict that you were pregnant.

Duhigg also describes how a customer in a Target outside Minneapolis was upset demanding to see the manager. He was clutching coupons that had been sent to his daughter. "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" The manager looked at the mailer. It was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

That story captures the essence of big data I'm going to be talking about in this lecture. or business utilizers an enormous amount of data the spot correlations that will help them predict a specific outcome.

Data Vs. Big Data: What Is the Difference?

So what is the difference between data and big data? Where does the big part come in? Many authors who have authored books about this (Harkness, Mayer-Schoenberger and Cukier, 2013) describe it as the ability to deal with the whole population rather than a sample drawn from that population. To term this another way, to deal with a sample size N=all rather than a sample size = n.

How Do We Handle Data?

Let's start with data (i.e. sample data). The overriding requirement here is for clean data. We need to doublecheck and clean the data of any errors – and that is indeed possible in a relatively small database. As the database gets larger, not so much. But even before we end up cleaning the sample, we need to collect the sample. And that has problems in itself. How do we choose a sample? The precision of our inferences increases with randomness, not sample size. A very large sample chosen in a biased manner may be completely wrong. As examples, consider the famous examples of electoral polling mishaps: The Dewey defeats Truman headline in the New York Post, 1948, or the Trump election, 2016, both of which were missed by the pollsters. What happened? Sampling biases. If you pick up the phone to ask people about their voting intentions, you would necessarily only be sampling people who have phones or answer their phones if you do. But if voters who go to vote are the ones who do not have phones or who do not answer their phones, the eventual outcomes will be completely unpredicted. And looking at finer samples (married women with 2 children who might vote Tory) decreases randomness even more.

Why then do we collect samples? That's because we have an idea in our heads about the way the world works. We have a causal hypothesis – that is we have an idea about what causes what. And then we collect a sample to test that hypothesis. Usually, that means this data cannot be used for any other tests. The reason is because collecting a sample is expensive and so we custom design the sample we collect to the hypothesis we need to test. Sample testing is all about causal inference.

How Do We Handle Big Data?

In contrast, big data is all about analysing the population. the population need not be large. For example, in their book Freakonomics, Levitt and Dubner analyze the population of sumo wrestling matches, a not super large population for evidence of cheating. But it is the entire population, not a sample.

The Domesday book is another example of analyzing the population. After William the Conqueror invaded England in 1066, he decided to find out who held what property throughout England. So, until his death, data collectors fanned out across England collecting data on people's houses, pigs, cows, and other livestock.

What are the characteristics for big data? Above all, it is messy data (it varies in quality, is collected at different times around the world, is kept in a wide variety of places). So at best, we can only hope for general directions vs. making precise causal inferences. in other words, we measure correlations not causations. But we may be able to infer things that we cannot do with standard causal inference analysis.

Where Is the Big Data Coming From?

A lot of it comes from using data from other sources. For example, before Google came along, machine translation between languages was really difficult. What researchers tried to do was to quantify the rules for language and then put them into computer code. As you quickly realise, the problem with languages is that it is full of exceptions. What makes sense in one situation will not necessarily work in a different situation. It is exceedingly difficult to understand how to code for irony or jokes. What did Google Translate do that was different? Basically Google sucked up all the documents it had on the web plus all the books it had digitalised over the years and fed them in as unstructured data into a computer algorithm. The algorithm was not given instructions on what these pages actually meant. The data was messy, it was full of slang, misspellings, grammatical mistakes, and assorted other problems. But there was a lot of data. The algorithm figured out what the probabilities were for correspondences between words and their meanings in different languages. I'll talk a little bit more about that next time but for now it's important to emphasise that this is another example of a side effect of big data. The computer doesn't understand what the phrase is - it just understand what words are probabilistically likely to occur next to each other.

A new data sources are coming on board all the time. Examples include satellite photos of car parking lots revealing information about whether supermarket sales are growing or falling per quarter, Google traffic maps

showing the impending invasion of Ukraine, and of course, smart phones.

Smartphones

Thanks to smartphones, today's average individual has at her fingertips a thousand times more computing power than was necessary to send man to the moon, more information than the best library used to contain, and more communicating power than any propaganda machine ever dreamed of possessing. More importantly, smartphones convinced most human beings to wear tracking devices, once reserved only for convicted felons on parole. Not only is it now possible to know whom we have talked to, but also where we have been, near whom, and for how long. Smartphones can track what searches we carried out, what books we bought, what vacations we shopped for, what dreams we have, and which diseases we fear to have. The latest iterations of smart devices (Oura rings, Apple watches, Fitbits) add biometric information to the mix. These allow phones to capture additional information that we may not know ourselves. The smartphone knows when you are stressed, when you are low on sugar, and when you like a person of the same or opposite sex. In so doing, smartphones have enabled a degree of constant surveillance, a panopticon, that even Bentham or Orwell would have struggled to conceive. Smartphones have been augmented by the Internet of things which comes with near ubiquitous coverage by a huge number of sensors scattered in our houses or in our neighborhoods that gather an even larger amount of information on us, in many cases, without us even being aware of the extent of the information being gathered.

I note here that while this surveillance poses very serious political problems, it does create enormous opportunities to eliminate the frictions of financing. These frictions are largely related to the asymmetry of information. Adverse selection and moral hazard problems are intrinsically linked to the inability to observe some individual characteristics or some actions, respectively. In fact, adverse selection used to be called "hidden information" and moral hazard "hidden actions". If neither information nor actions can be hidden, the financing frictions related to asymmetry of information, which have dominated the finance literature for the better part of the second half of the 20th century, are eliminated as I discussed in my first lecture in this series in November 2022.

New Forms of Data

So now we have lots of different types of data. Examples include:

- Geospatial: Study of mobility and location-based proximity
- Sociometric: Study of social relationships, especially preferences, within groups
- Psychometric: Study of a person's mental state, personality and thought processes
- Biometric: Study of a person's biological characteristics

Why Are We Collecting So Much Data?

A lot of people assume that more data is better data. You may have even heard the phrase "data is a new oil." Business assume that they will get valuable insights. Regulators assume that it reduces asymmetric information. For example, companies listed on the stock market are required to file quarterly financial reports with the stock market regulator, which are then made public. Banks and investment funds must comply with stringent reporting obligations. Companies operating in certain sectors (pharmaceuticals, health care, education, air travel) are required to provide additional information. The problem of course is that most human beings do not have either the time or the ability to interpret the massive amount of data produced by smartphones.

How Do We Make Inferences from Data?

Remember the economist or business cannot see what you think, only what you do. For example, if I offer you an apple or a banana at the same price, and you pick the banana, what do I infer? That you like the banana. But what are you reacting to? The price? Not necessarily. Suppose you prefer:

- apples > bananas
- organic > regular and
- ripe > green.

How do you choose between green conventional bananas and ripe organic apples? What dimension is more

important to you? What happens if there are more dimensions (how and where it was grown, its sugar content, nutritional value, and shelf life)? We are very bad at processing more than a few dimensions. But big data assumes that by counting all the possible dimensions of the data, they can come up with deep patterns that explain our behaviour, patterns that we might not ourselves be aware of, will emerge from the data.

We Need to Process the Data to Make Inferences

But unfortunately, it is not quite so simple to get inferences from big data. We need a clear stream or steps.

1. We need a standard language to compare preferences.
2. We need to better match preferences along multiple dimensions.
3. We need an effective way to comprehensively capture our preferences.

Let's take each of these in turn.

Element 1: Developing A Data Ontology

How do we search for something on say, a shoe website, like Zappos? Zappos list literally hundreds of different types of shoes and makes and sizes and colours. how does it have so much information on issues? the answer is collected data about each characteristic of the shoe in a specialised language called tags. Every tag has to be the same across shoe and each tag must refer to only one characteristic of the shoe. For example, you can have a tag for the colour of the shoe, you can have a tag for the size of the shoe, and so on. This is easy in a specialized market place such as shoes, computers cover washing machines, hard drives, and cameras.

It is much more difficult in a general website like YouTube. How does the YouTube algorithm work? It looks for relevance of the video, your engagement with the video, the quality of the video, and personalizes the choice to your viewing behavior - videos you have watched in the past, or the videos you have typically watched. All that information comes from the information the uploader provides in the metadata, in the title, in the name of the video file and so on. And even then, many videos never get watched, however good they may be.

Element 2: Multi-dimensional Preference Matching

The second element consists of constructing algorithms to evaluate sets of multiple preferences and their relative weights and to identify best matches. This is the same technology used to manage our photo collections to find pictures with certain faces/features, or to have Siri or Alexa “understand” voice commands, or to make our smart watches detect signs of a dangerous heart condition. Basically, preference data is just a data stream forming a particular pattern, so we can adapt ordinary pattern-matching algorithms to this job.

Take the example of Netflix. What did Netflix want to use big data for? It wanted to find new shows to commission (House of Cards) or figure out how much to pay for new films. The challenge was obtaining ratings on the films that viewers watched. And it was worse because at that time, Netflix was not a streaming service. It shipped out DVDs by mail, so it had to figure out which DVDs to stock. It had to decide which viewer would get the DVD if there was only one in stock (the answer to the last one is that new viewers would get priority because Netflix was interested in retaining them and finding out their viewing habits; older viewers would get another video on their list. But of course, viewers figured that out and would cancel and sign up again with different names, or would share DVDs among friends and so on).

Netflix designed the Cinematch system to predict which films a viewer would want to watch based on the video ratings it received from some viewers. But it faced problems.

- Cold start problem: When a recommendation system lacks data to make predictions: Products newly added to the database and, more critically, for new users for whom no data is available.
- Popularity bias: Popular movies are recommended frequently while less popular, niche films are recommended rarely or not at all.
- Sparse data problem: The Netflix data set is both extremely large and relatively sparse, with only about 1% of the potential user-movie data points available.
- Noisy data problem: Ratings that people report do not reflect actual behavior

So, in October 2006, Netflix announced the Netflix prize. It was $1 million contest that invited the public to devise a recommendation algorithm that could beat Cinematch by at least a 10% improvement. The contest

would be open for five years. If no one won within a year, Netflix would award $50,000 to whoever made the most progress above a 1% improvement and would award the same amount each year until someone won the grand prize. Netflix released a database of 100 million customer ratings to anyone interested in cracking the code. Netflix received more than 160 submissions within two weeks of announcing the Netflix Prize. The winning entry (composed of two teams) eventually won the prize – but with a caveat – there were some films that were simply unpredictable regardless of the number of tags you applied. Examples? Napoleon Dynamite or Miss Congeniality. If you like either of these films, rest assured that Netflix doesn't know why you liked them.

Element 3: Understanding Us

The final element consists of asking whether big data can understand us? What does it need? It needs four pre-requisites:

- Huge volumes of data
- Frequent feedback for the system
- Self-adjusting capacity
- Freshness (dumping old data once it is stale)

Using all these prerequisites, let's take three examples to illustrate how this data can be used to predict our behavior for fun, profit, or other business related decisions.

Twitter/Foursquare data

In this example, Hristova and her colleagues (Hristova, 2016) collected a database of 38,000 users of Foursquare and Twitter and mapped 433K reciprocal connections between these users on Twitter. They mapped 550K check-ins to 42K venues in London and 3M user transitions between venues on Foursquare. So they had converted geographic place map into a social network map. they went on to show that it is possible to predict the growth of real estate prices using the development of diversity in a particular area.

Facebook data

David Stilwell, a colleague of mine at the Judge Business School and his co-authors developed an app called myPersonality. David, who was an undergrad in psychology at the time, created a personality questionnaire and shared with his around 50 Facebook friends. Friends of his friends, and then the friends of the friends of his friends, all joined the study. Soon, thousands of people were participating every day. Six *million* people took one of the myPersonality questionnaires. myPersonality participants could (but did not have to!) opt in to donate their Facebook data to the project. To do so, they would have to read and opt in to a consent statement, and then confirm their intent again in a pop-up dialog window shown by Facebook. About 40% of the users allowed their Facebook data to be used in David's research.

In a series of papers (Kosinski, Stillwell and Graepel, 2013, Markovikj et al, 2021), David and his colleagues showed that these easily accessible digital records of behavior, (Facebook Likes), could be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The For example, they were able to distinguish between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases.

Loan application data

In this example, the authors, Netzer, Lemaire, and Herzenstein examined Prosper – a peer-to-peer lending site. They wanted to show that borrowers, consciously or not, leave traces of their intentions, circumstances, and personality traits in the text they write when applying for a loan. On Prosper.com, potential borrowers write a brief description of why they need a loan and why they are likely to pay the lender back. Potential lenders decide whether to lend them the money. It turns out that 13% of borrowers defaulted on their loan. The authors use text-mining and machine-learning tools to automatically process and analyze the raw text in over 120 thousand loan requests.

To take an example from their paper, imagine you consider lending $2,000 to one of two borrowers on Prosper.com. The borrowers are identical in terms of their demographic and financial characteristics, the amount of money they wish to borrow, and the reason for borrowing the money. However, the text they provided when applying for a loan differs:

Borrower #1 writes “I am a hard working person, married for 25 years, and have two wonderful boys. Please let me explain why I need help. I would use the $2,000 loan to fix our roof. Thank you, god bless you, and I promise to pay you back.” Borrower #2 writes “While the past year in our new place has been more than great, the roof is now leaking and I need to borrow $2,000 to cover the cost of the repair. I pay all bills (e.g., car loans, cable, utilities) on time.”

Who do you think would be more likely to default?

It turns out that including the textual information in the loan significantly helps predict loan default and can have substantial financial implications. The authors find that loan requests written by defaulting borrowers are more likely to include words related to their family, mentions of god, the borrower’s financial and general hardship, pleading lenders for help, and short-term focused words. The authors further observe that defaulting loan requests are written in a manner consistent with the writing style of extroverts and liars.

The Pitfalls of Big Data

The biggest problem of big data is that to analyze it, we depend on correlations. There are too many variables to analyze causation. And that leads to a problem.

Let’s go back to 1854 London to see why this is a problem. Cholera, was ravaging London after its arrival from India in 1831. In his book, the Ghost Map, a brilliantly written account of the scientific method, Johnson describes how John Snow, a physician in Soho, observed how the disease appeared to spread from patient to patient within families, or via clothes and bedding belonging to somebody who had died of cholera, or to people who washed and laid out a body, Snow developed the theory that it was spread via the alimentary canal. He eventually heard of a case in Broad Street where there was a sudden upsurge in cholera cases with over 80 deaths registered within three days. He suspected that the water pump on Broad Street was the cause and had it removed. The cholera stopped.

This was causal reasoning. Unfortunately, correlation analysis would have given a completely different and equally plausible answer. The prevailing theory of the time for how cholera spread was miasma: bad air rose from the river and carried the illness with it. And the river Thames in the early nineteenth century was putrid and foul.

Enter William Farr. Farr collected data on where the deaths happened, and on factors that might explain the difference in death rates between districts. Farr analysed his data, combining the districts according to elevation above high water. The results were striking. The mortality rate for cholera was highest within 20ft of river level, with over ten people per thousand of the population dying of cholera that year. If you lived in a district 30–40ft above the river, and the death rate falls to around six per thousand, and so on, with the improvement becoming more gradual in higher districts. 340ft and more above river level, less than one person in a thousand died of cholera.

This would be conclusive proof, as far as big data was concerned, that the miasma theory was correct, and that eliminating the noxious gases would reduce the death toll. The problem today is of degree. We have more data and more correlations. But that does not mean we know what is going on. In the last two lectures in this series, I will first discuss how AI systems like ChatGPT use this data to make inferences about you. In the final lecture, I will expand more on the dark side of using all this technology to make unwarranted inferences.

References and Further Reading

Duhigg, Charles, 2012, How companies learn your secrets, New York Times magazine, February 16, 2012 (available at https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html).

Harkness, T., 2016, Big Data: Does Size Matter, Bloomsbury Sigma.

Hristova, D., Williams, M., Musolesi, M., Panzarasa, P., and Mascolo, C., 2016, Measuring Urban Social Diversity Using Interconnected Geo-Social Networks, WWW '16: Proceedings of the 25th International Conference on World Wide Web (available at https://dl.acm.org/doi/abs/10.1145/2872427.2883065).

Johnson, S., 2006, The Ghost Map: The Story of London's Most Terrifying Epidemic--And How It Changed

Science, Cities, and the Modern World, Riverhead books.

Kosinski, M,. Stillwell, D., and Graepel, T. 2013, Private traits and attributes are predictable from digital records of human behavior, Proceedings of the National Academy of Sciences, 110 (15), 5802-5805 (available at https://doi.org/10.1073/pnas.1218772110)

Levitt, S.D. and Dubner, S.J., 2006, Freakonomics: A Rogue Economist Explores the Hidden Side of Everything, Penguin.

Mayer-Schonberger, V. and Cukier, K, 2013, Big Data: A Revolution That Will Transform How We Live, Work and Think, John Murray.

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D., 2021. Mining Facebook Data for Predictive Personality Modeling. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(2), 23-26. (available at https://doi.org/10.1609/icwsm.v7i2.14466)

Netzer, Oded, Lemaire, Alain and Herzenstein, Michal, 2019, When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Columbia Business School Research Paper No. 16-83. (available at SSRN: https://ssrn.com/abstract=2865327)

Stillwell, D. J., & Kosinski, M. (2015). myPersonality Project website (available at https://sites.google.com/michalkosinski.com/mypersonality).