Big Data (Vs: Velocity, Volume, Variety, Veracity, and Value.)

In this digital world, everyone leaves a trace. From our travel habits to our workouts and entertainment, the increasing number of internet connected devices that we interact with on a daily basis record vast amounts of data about us. There’s even a name for it: Big Data. Ernst and Young offers the following definition: “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.” There is no one definition of Big Data, but there are certain elements that are common across the different definitions, such as velocity, volume, variety, veracity, and value. These are the V's of Big Data. Velocity is the speed at which data accumulates. Data is being generated extremely fast, in a process that never stops. Near or real-time streaming, local, and cloud-based technologies can process information very quickly. Volume is the scale of the data, or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure. Variety is the diversity of the data. Structured data fits neatly into rows and columns, in relational databases while unstructured data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers, and video. Variety also reflects that data comes from different sources, machines, people, and processes, both internal and external to organizations. Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more. Veracity is the quality and origin of data, and its conformity to facts and accuracy. Attributes include consistency, completeness, integrity, and ambiguity. Drivers include cost and the need for traceability. With the large amount of data available, the debate rages on about the accuracy of data in the digital era. Is the information real, or is it false? Value is our ability and need to turn data into value. Value isn't just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction. The main reason that people invest time to understand Big Data is to derive value from it. Let's look at some examples of the V's in action. Velocity: Every 60 seconds, hours of footage are uploaded to YouTube which is generating data. Think about how quickly data accumulates over hours, days, and years. Volume: The world population is approximately seven billion people and the vast majority are now using digital devices; mobile phones, desktop and laptop computers, wearable devices, and so on. These devices all generate, capture, and store data -- approximately 2.5 quintillion bytes every day. That's the equivalent of 10 million Blu-ray DVD's. Variety: Let's think about the different types of data; text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the Internet of Things. Veracity: 80% of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights. The data must be categorized, analyzed, and visualized. Data Scientists today derive insights from Big Data and cope with the challenges that these massive data sets present. The scale of the data being collected means that it’s not feasible to use conventional data analysis tools. However, alternative tools that leverage distributed computing power can overcome this problem. Tools such as Apache Spark, Hadoop and its ecosystem provide ways to extract, load, analyze, and process the data across distributed compute resources, providing new insights and knowledge. This gives organizations more ways to connect with their customers and enrich the services they offer. So next time you strap on your smartwatch, unlock your smartphone, or track your workout, remember your data is starting a journey that might take it all the way around the world, through big data analysis, and back to you.

What is Hadoop

Traditionally in computation and processing data we would bring the data to the computer. You'd wanna program and you'd bring the data into the program. In a big data cluster what Larry Page and Sergey Brin came up with is very simple is they took the data and they sliced it into pieces and they distributed each and they replicated each piece or triplicated each piece and they would send it the pieces of these files to thousands of computers first it was hundreds but then now it's thousands now it's tens of thousands. And then they would send the same program to all these computers in the cluster. And each computer would run the program on its little piece of the file and send the results back. The results would then be sorted and those results would then be redistributed back to another process. The first process is called a map or a mapper process and the second one was called a reduce process. Fairly simple concepts but turned out that you could do lots and lots of different kinds of handle lots and lots of different kinds of problems and very, very, very large data sets. So the one thing that's nice about these big data clusters is they scale linearly. You had twice as many servers and you get twice the performance and you can handle twice the amount of data. So this was just broke a bottleneck for all the major social media companies. Yahoo then got on board. Yahoo hired someone named Doug Cutting who had been working on a clone or a copy of the Google big data architecture and now that's called Hadoop. And if you google Hadoop you'll see that it's now a very popular term and there are many, many, many if you look at the big data ecology there are hundreds of thousands of companies out there that have some kind of footprint in the big data world. (music)

Play video starting at :2:28 and follow transcript2:28

Most of the components of data science have been around for many, many, many decades. But they're all coming together now with some new nuances I guess. At the bottom of data science you see probability and statistics. You see algebra, linear algebra you see programming and you see databases. They've all been here. But what's happened now is we now have the computational capabilities to apply some new techniques - machine learning. Where now we can take really large data sets and instead of taking a sample and trying to test some hypothesis we can take really, really large data sets and look for patterns. And so back off one level from hypothesis testing to finding patterns that maybe will generate hypotheses. Now this can bother some very traditional statisticians and gets them really annoyed sometimes that you know you're supposed to have a hypothesis that is not that is independent of the data and then you test it. So once some of these machine learning techniques started were really the only thing the only way you can analyze some of these really large social media data sets. So what we've seen is that the combination of traditional areas computer science probability, statistics, mathematics all coming together in this thing that we call Decision Sciences. Our department at Stern I'll give a little plug here we happen to have been very well situated among business schools because we're one of the few business schools that has a real statistics department with real PhD level statisticians in it. We have an operations management department and an information systems department. So we have a wide range of computer scientists to statisticians, to operations researchers. And so we were perfectly positioned as a couple of other business schools were to jump on this bandwagon and say; okay this is Decision Sciences. And Foster Provost who's in my department was the first director of the NYU Center for Data Science. (music)

Play video starting at :5:3 and follow transcript5:03

Four years ago maybe five years ago. I mean, I feel this is one of those cases where you can just to Google and search for data science and see how often it occurred and you'll see almost nothing and then just a spike. The same thing you would see with big data about seven or eight years ago. So data science is a term I haven't heard of probably five years ago. (music)

Play video starting at :5:34 and follow transcript5:34

The first question is what is it? And I think faculty and everybody is still trying to get their hands around exactly what is business analytics and what is data science. We certainly know the components of it. But it's morphing and changing and growing. I mean the last three years deep learning has just been added into the mix. Neural networks have been around for 20 or 30 years. 20 years ago I would teach neural networks in a class and you really couldn't do very much with them. And now some researchers have come up with multi-layer neural networks in Toronto in particular the University of Toronto. And that technology is now rapidly expanding it's being used by Google, by Facebook, by lots of companies. (music)

Big Data driving digital transformation

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benefits of new technologies. This digital change integrates digital technology into all areas of an organization resulting in fundamental changes to how it operates and delivers value to customers. It is an organizational and cultural change driven by Data Science, and especially Big Data. The availability of vast amounts of data, and the competitive advantage that analyzing it brings has triggered digital transformations throughout many industries. Netflix moved from being a postal DVD lending system to one of the world’s foremost video streaming providers, the Houston Rockets NBA team used data gathered by overhead cameras to analyze the most productive plays, and Lufthansa analyzed customer data to improve its service. Organizations all around us are changing to their very core. Let’s take a look at an example, to see how Big Data can trigger a digital transformation, not just in one organization, but in an entire industry. In 2018, the Houston Rockets, a National Basketball Association, or NBA team, raised their game using Big Data. The Rockets were one of four NBA teams to install a video tracking system which mined raw data from games. They analyzed video tracking data to investigate which plays provided the best opportunities for high scores, and discovered something surprising. Data analysis revealed that the shots that provide the best opportunities for high scores are two-point dunks from inside the two-point zone, and three-point shots from outside the three-point line, not long-range two-point shots from inside it. This discovery entirely changed the way the team approached each game, increasing the number of three-point shots attempted. In the 2017-18 season, the Rockets made more three-point shots than any other team in NBA history, and this was a major reason they won more games than any of their rivals. In basketball, Big Data changed the way teams try to win, transforming the approach to the game. Digital transformation is not simply duplicating existing processes in digital form; the in-depth analysis of how the business operates helps organizations discover how to improve their processes and operations, and harness the benefits of integrating data science into their workflows. Most organizations realize that digital transformation will require fundamental changes to their approach towards data, employees, and customers, and it will affect their organizational culture. Digital transformation impacts every aspect of the organization, so it is handled by decision makers at the very top levels to ensure success. The support of the Chief Executive Officer is crucial to the digital transformation process, as is the support of the Chief Information Officer, and the emerging role of Chief Data Officer. But they also require support from the executives who control budgets, personnel decisions, and day-to-day priorities. This is a whole organization process. Everyone must support it for it to succeed. There is no doubt dealing with all the issues that arise in this effort requires a new mindset, but Digital Transformation is the way to succeed now and in the future.



