**WHAT IS DATA SCIENCE?**

* If you have data, and you have curiosity, and you're working with data, and you're manipulating it, you're exploring it, the very exercise of going through analyzing data, trying to get some answers from it is data science.

**FUNDAMENTALS OF DATA SCIENCE**

* When the data has revealed its insights, the role of the data scientist becomes that of a storyteller, communicating the results to the project stakeholders. Data scientists can use powerful data visualization tools to help stakeholders understand the nature of the results, and the recommended action to take.
* Data scientists can analyze structured and unstructured data from many sources, and depending on the nature of the problem, they can choose to analyze the data in different ways. Using multiple models to explore the data reveals patterns and outliers; sometimes, this will confirm what the organization suspects, but sometimes it will be completely new knowledge, leading the organization to a new approach.
* So, what is the process of data science? Many organizations will use data science to focus on a specific problem, and so it's essential to clarify the question that the organization wants answered. This first and most crucial step defines how the data science project progresses.
* Data scientists use data analysis to add to the knowledge of the organization by investigating data, exploring the best way to use it to provide value to the business.
* Data science can help organizations understand their environments, analyze existing issues, and reveal previously hidden opportunities.
* Data analysis isn't new. What is new is the vast quantity of data available from massively varied sources: from log files, email, social media, sales data, patient information files, sports performance data, sensor data, security cameras, and many more besides.

**THE MANY PATHS OF DATA SCIENCE**

* Data science didn't exist until 2009, 2011. Someone like DJ Patil or Andrew Gelman coined the term. Before that, there was statistics.

**ADVICE FOR NEW DATA SCIENTIST**

* My advice to an aspiring data scientist is to be curious, extremely argumentative and judgmental. Curiosity is absolute must. If you're not curious, you would not know what to do with the data. Judgmental because if you do not have preconceived notions about things you wouldn't know where to begin with. Argumentative because if you can argument and if you can plead a case, at least you can start somewhere and then you learn from data and then you modify your assumptions and hypotheses and your data would help you learn. And you might start at the wrong point. You may say that I thought I believed this, but now with data I know this. So, this allows you a learning process. So, curiosity being able to take a position, strong position, and then moving forward with it. The other thing that the data scientist would need is some comfort and flexibility with analytics platforms: some software, some computing platform, but that's secondary. The most important thing is curiosity and the ability to take positions. Once you have done that, once you've analyzed, then you've got some answers. And that's the last thing that a data scientist need, and that is the ability to tell a story. That once you have your analytics, once you have your tabulations, now you should be able to tell a great story from it. Because if you don't tell a great story from it, your findings will remain hidden, remain buried, nobody would know. But your rise to prominence is pretty much relying on your ability to tell great stories. A starting point would be to see what is your competitive advantage. Do you want to be a data scientist in any field or a specific field? Because, let's say you want to be a data scientist and work for an IT firm or a web-based or Internet based firm, then you need a different set of skills. And if you want to be a data scientist in the health industry, then you need different sets of skills. So figure out first what you're interested, and what is your competitive advantage. Your competitive advantage is not necessarily going to be your analytical skills. Your competitive advantage is your understanding of some aspect of life where you exceed beyond others in understanding that. Maybe it's film, maybe it's retail, maybe it's health, maybe it's computers. Once you've figured out where your expertise lies, then you start acquiring analytical skills. What platforms to learn and those platforms, those tools would be specific to the industry that you're interested in. And then once you have got some proficiency in the tools, the next thing would be to apply your skills to real problems, and then tell the rest of the world what you can do with it.

**OLD PROBLEMS... NEW PROBLEMS... DATA SCIENCE SOLUTIONS**

* How do you get a better solution that is efficient? You must: Identify the problem and establish a clear understanding of it. Gather the data for analysis. Identify the right tools to use. Develop a data strategy. Case studies are also helpful in customizing a potential solution. Once these conditions exist and available data is extracted, you can develop a machine learning model. It will take time for an organization to refine best practices for data strategy using data science, but the benefits are worth it.
* Toronto Transportation Commission has made great strides in solving an old problem with traffic flows, restructuring those flows in and around the city. Using data science tools and analysis, they have: Gathered data to better understand streetcar operations, and identify areas for interventions Analyzed customer complaints data Used probe data to better understand traffic performance on main routes Created a team to better capitalize on big data for both planning, operations and evaluation By focusing on peak hour clearances and identifying the most congested routes, monthly hours lost for commuters due to traffic congestion dropped from 4.75 hrs. in 2010 to 3 hrs. in mid-2014.
* Uber collects real-time user data to discover how many drivers are available, if more are needed, and if they should allow a surge charge to attract more drivers. Uber uses data to put the right number of drivers in the right place, at the right time, for a cost the rider is willing to pay.
* However, all organizations ultimately use data science for the same reason—to discover optimum solutions to existing problems

**DATA SCIENCE TOPICS & ALGORITHMS**

* Because in the absence of knowing those relationships, and just knowing how much people traveled for and how much they paid, regression allows you to compute that constant that you didn't know. That it's $2.50, and it would compute the relationship between the fare and and the distance and the fare and the time. That is regression.
* Let me explain regression in the simplest possible terms. If you have ever taken a cab ride, a taxi ride, you understand regression. Here is how it works. The moment you sit in a cab ride, in a cab, you see that there's a fixed amount there. It says $2.50. You, rather the cab, moves or you get off. This is what you owe to the driver the moment you step into a cab. That's a constant. You have to pay that amount if you have stepped into a cab. Then as it starts moving for every meter or hundred meters the fare increases by certain amount. So there's a... there's a fraction, there's a relationship between distance and the amount you would pay above and beyond that constant. And if you're not moving and you're stuck in traffic, then every additional minute you have to pay more. So as the minutes increase, your fare increases. As the distance increases, your fare increases. And while all this is happening you've already paid a base fare which is the constant. This is what regression is. Regression tells you what the base fare is and what is the relationship between time and the fare you have paid, and the distance you have traveled and the fare you've paid.
* **STRUCTURED DATA**: If you have something which fits nicely into tables and columns and rows, go head. That's your structured data.
* **UNSTRUCTURED DATA**: Basically, data that is coming from mostly from web where it's not tabular. It is not, it's not in rows and columns. It's text. It's sometimes it's video and audio, so you would have to deploy more sophisticated algorithms to extract data. And in fact, a lot of times we take unstructured data and spend a great deal of time and effort to get some structure out of it and then analyze it. (ie: Tweets, blog posts, pictures, numbers, and video)

**THE CLOUD FOR DATA SCIENCE**

* The advantage of working with Data Scientist Workbench is not only that you have the ability to work with these advanced algorithms into computing platforms, but you also have the ability to work with very large data set, because Spark is integrated and it's all in the Cloud.

**\*\*IBM provides SN Labs to learners registered at any of the learning portals on the IBM Developer Skills Network \*\*Follow Up Big Data University\*\***

* At the Big Data University which is an IBM initiative, we have these courses people can take and learn about data science. But at the same time, we provide these Cloud-based environment for not only analytics but also for working with big and small data. So, one of the products that is integrated with Big Data University is Data Scientist Workbench. Data Scientist Workbench is an internet-based solution. You log in, and the moment you log in, you now have access to some very advanced computing environments. As simple as R in RStudio, and data and algorithms to define the data set using OpenRefine, but also the ability to work with very large data sets using technologies like Spark. So, the advantage of working with Data Scientist Workbench is not only that you have the ability to work with these advanced algorithms into computing platforms, but you also have the ability to work with very large data set, because Spark is integrated and it's all in the Cloud. You don't have to maintain it. You don't have to download it. You don't have to worry about updating it. All is being done for you in the Cloud by the Data Scientist Workbench.
* Another thing Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. So, you can be working with the same data that your colleagues in, say, Germany, and another team in India, and another team in Ghana, they are collectively working and they're able to do so because the information, and the algorithms, and the tools, and the answers, and the results, whatever they needed is available at a central place which we call Cloud.
* Cloud allows you not just to store large amounts of data on servers somewhere in California or in Nevada, but it also allows you to deploy very advanced computing algorithms and the ability to do high performance computing using machines that are not yours. So, think of it as you have some information, you can't store it, so you send it to storage space, let's call it Cloud. And the algorithms that you need to use, you don't have them with you. But then, on the Cloud, you have those algorithms available. So, what you do is you deploy those algorithms on very large data sets and you're able to do it even though your own systems, your own machines, your own computing environment would not allow you to do so.
* Allows you to bypass the physical limitations of the computers and the systems you're using, and it allows you to deploy the analytics and storage capacities of advanced machines that do not necessarily have to be your machine or your company's machine.

**FOUNDATIONS OF BIG DATA**

* “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.”
* 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.
* 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.
* Data Scientists today derive insights from Big Data and cope with the challenges that these massive data sets present.

**The V's of Big Data: velocity, volume, variety, veracity, and value.**

* The main reason that people invest time to understand Big Data is to derive value from it.
* 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.
* 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.
* Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more.
* Variety also reflects that data comes from different sources, machines, people, and processes, both internal and external to organizations.
* Variety is the diversity of the data. (ie: Text, pictures, film, sound, health data, devices connected to IoT)
* Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.
* Volume is the scale of the data, or the increase in the amount of data stored (ie: world population approx 7 billion. Majority use mobile phones, dektop, laptop, wearables, etc. These devices generate, capture, & store data - approx 2.5 quintillion bytes/day)
* 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 (ie: Every 60 seconds, hours of footage are uploaded to YouTube which generates data)