A day in the Life of a Data Scientist

I've built a recommendation engine before, as part of a large organization and worked through all types of engineers and accounted for different parts of the problem. It's one of the ones I'm most happy with because ultimately, I came up with a very simple solution that was easy to understand from all levels, from the executives to the engineers and developers. Ultimately, it was just as efficient as something really complex, and they could have spent a lot more time on.

Back in the university, we have a problem that we wanted to predict algae blooms. This algae blooms could cause a rise in toxicity of the water and it could cause problems through the water treatment company. We couldn't like predict with our chemical engineering background. So we use artificial neural networks to predict when these blooms will reoccur. So the water treatment companies could better handle this problem.

In Toronto, the public transit is operated by Toronto Transit Commission. We call them TTC. It's one of the largest transit authorities in the region, in North America. And one day they contacted me to analyse half a million of complaint data. I said, "Well, let's start working with it." So I got the data and I started analyzing it. So, basically, they have done a great job of keeping some data in tabular format that was unstructured data. And in that case, tabular data was when the complaint arrived, who received it, what was the type of the complaint, was it resolved, whose fault was it. And the unstructured part of it was the exchange of e-mails and faxes. So, imagine looking at how half a million exchanges of e-mails and trying to get some answers from it. So I started working with it.

The first thing I wanted to know is why would people complain and is there a pattern or is there some days when there are more complaints than others? And I had looked at the data and I analyzed it in all different formats, and I couldn't find [what] the impetus for complaints being higher on a certain day and lower on others. And it continued for maybe a month or so. And then, one day I was getting off the bus in Toronto, and I was still thinking about it. And I stepped out without looking on the ground, and I stepped into a puddle, puddle of water. And now, I was sort of ankle deep into water, and it was just one foot wet and the other dry. And I was extremely annoyed. And I was walking back and then it hit me, and I said, "Well, wait a second. Today it rained unexpectedly, and I wasn't prepared for it. That's why I'm wet, and I wasn't looking for it." What if there was a relationship between extreme weather and the type of complaints TTC receives?

So, I went to the environment Canada's website, and I got data on rain and precipitation, wind and the light. And there, I found something very interesting. The 10 most excessive days for complaints. The 10 days where people complain the most were the days when the weather was bad. It was unexpected rain, an extreme drop in temperature, too much snow, very windy day. So I went back to the TTC's executives and I said, "I've got good news and bad news." And the good news is, I know why people would complain excessively on certain days. I know the reason for it. The bad news is, there's nothing you can do about it.

# Old problems, new problems, Data Science solutions

Organizations can leverage the almost unlimited amount of data now available to them in a growing number of ways. However, all organizations ultimately use data science for the same reason—to discover

optimum solutions to existing problems. Let’s take a look at three examples of data science providing innovative solutions for old problems. In transport, 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.

In a different transport related data science effort, the 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 and 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, in facing issues in our environment, data science can also play a proactive role. Freshwater lakes supply a variety of human and ecological needs, such as providing drinking water and producing food. But lakes across the world are threatened by increasing incidences of harmful cyanobacterial blooms. There are many projects and studies to solve this long-existing dilemma. In the US, a team of scientists from research centers stretching from Maine to South Carolina is developing and deploying high-tech tools to explore cyanobacteria in lakes across the east coast. The team is using robotic boats, buoys, and camera-equipped drones to measure hysical, chemical, and biological data in lakes where cyanobacteria are detected, collecting large volumes of data related to the lakes and the development of the harmful blooms.

The project is also building new algorithmic models to assess the findings. The information collected will lead to better predictions of when and where cyanobacterial blooms take place, enabling proactive approaches to protect public health in recreational lakes and in those that supply drinking water. Such interdisciplinary training prepares the next generation of scientists to address societal issues with the proper modernized data science tools.

It takes gathering a lot of data, cleaning and preparing it, and then analyzing it to gain the insight needed to develop better solutions for today's enterprises. How do you get a better solution that is efficient? You must:

1. **Identify the problem** and establish a clear understanding of it.
2. **Gather the data** for analysis.
3. **Identify the right tools** to use and 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
4. **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.

# Data Science Topics and Algorithms

**Regression** helped me understand data.

**Data Visualization** is a key element for people to get across their message to people that don't understand data science well.

**Artificial neural networks**: we have a lot to learn with nature so when we are trying to mimic our brain I think that we can do some applications with this behavior with this biological behavior in algorithms.

**K-Nearest Neighbors**. It's the simplest but it just gets the best results so many more times than some overblown, overworked algorithm that's just as likely to overfit as it is to make a good fit.

**Structured data** is more like tabular data things that you’re familiar with in Microsoft Excel format. You've got rows and columns and that's called structured data.

**Unstructured data** is 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.

# Cloud for Data Science

Cloud is a godsend for data scientists. Primarily because it 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.

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. 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 datasets and you're able to do it even though your own systems, your own machines, your own computing environments were not allowing you to do so.

The other thing that Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. 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.

Using the Cloud enables you to get instant access to open-source technologies like Apache Spark without the need to install and configure them locally. Using the Cloud also gives you access to the most up-to-date tools and libraries without the worry of maintaining them and ensuring that they are up to data. The Cloud is accessible from everywhere and in every time zone. You can use cloud-based technologies from your laptop, from your tablet, and even from your phone, enabling collaboration more easily than ever before.

Multiple collaborators or teams can access the data simultaneously, working together on producing a solution. Some big tech companies offer Cloud platforms, allowing you to become familiar with cloud-based technologies in a pre-built environment. IBM offers the IBM Cloud, Amazon offers Amazon Web Services or AWS, and Google offers Google Cloud platform.

IBM also provides Skills Network labs or SN labs to learners registered at any of the learning portals on the IBM Developer Skills Network, where you have access to tools like Jupyter Notebooks and Spark clusters so you can create your own data science project and develop solutions. With practice and familiarity, you will discover how the Cloud dramatically enhances productivity for data scientists.

### What Makes Someone a Data Scientist?

Hal Varian named statisticians sexy, it is widely believed that what he really meant were data scientists. This raises several important questions:

• What is data science?  
• How does it differ from statistics?  
• What makes someone a data scientist?

In the times of big data, a question as simple as, what is data science? can result in many answers. In some cases, the diversity of opinion on these hidden answers’ borders on hostility.

I define a data scientist as **someone who finds solutions to problems by analyzing data using appropriate tools** and **then tells stories to communicate her findings to the relevant stakeholders**. I do not use the data size as a restrictive clause. A data below a certain arbitrary threshold does not make one less of a data scientist. Nor is my definition of a data scientist restricted to particular analytic tools, such as machine learning. As long as **one has a curious mind, fluency in analytics, and the ability to communicate the findings**, I consider the person a data scientist.

Years ago, as an engineering student at the University of Toronto, I wrote my master’s thesis on forecasting housing prices and my doctoral dissertation on forecasting homebuilders’ choices related to what, when and where they build new housing. In the civil engineering department, Others were working on designing buildings, bridges, tunnels, and worrying about the stability of slopes. My work, and that of my supervisor, was not your traditional garden-variety engineering. Obviously, I was repeatedly asked by others whether my research was indeed engineering.

When I shared these concerns with my doctoral supervisor, Professor Eric Miller, he had a laugh. Dr Miller spent a lifetime researching urban land use and transportation and had earlier earned a doctorate from MIT. “Engineering is what engineers do,” he responded. Over the next 17 years, I realized the wisdom in his statement. You first become an engineer by obtaining a degree and then registering with the local professional body that regulates the engineering profession. Now you are an engineer. You can dig tunnels; write software codes; design components of an iPhone or a supersonic jet. You are an engineer. And when you are leading the global response to a financial crisis in your role as the chief economist of the International Monetary Fund (IMF), as Dr Raghuram Rajan did, you are an engineer.

Professor Raghuram Rajan did his first degree in electrical engineering from the Indian Institute of Technology. He pursued economics in graduate studies, later became a professor at a prestigious university, and eventually landed at the IMF. He is currently serving as the 23rd Governor of the Reserve Bank of India. Could someone argue that his intellectual prowess is rooted only in his training as an economist and that the fundamentals he learned as an engineering student played no role in developing his problem-solving abilities?

Professor Rajan is an engineer. So are Xi Jinping, the President of the People’s Republic of China, and Alexis Tsipras, the Greek Prime Minister who is forcing the world to rethink the fundamentals of global economics. They might not be designing new circuitry, distillation equipment, or bridges, but they are helping build better societies and economies and there can be no better definition of engineering and engineers—that is, individuals dedicated to building better economies and societies.

There are many different definitions on Data Science. In September 2015, a co-panelist at a meetup organized by BigDataUniversity.com in Toronto confined **data science to machine learning**. There you have it. If you are not using the black boxes that makeup machine learning, as per some experts in the field, you are not a data scientist. Even if you were to discover the cure to a disease threatening the lives of millions, turf-protecting colleagues will exclude you from the data science club.

Dr Vincent Granville (2014), an author on data science, uses restrictions on data size and methods to define what data science is. He defines a data scientist as **one who can easily process a So-million-row data set in a couple of hours, and who distrusts (statistical) models**. He distinguishes data science from statistics. Yet he lists algebra, calculus, and training in probability and statistics as necessary background to understand data science (page 4).

Some believe that big data is merely about crossing a certain threshold on data size or the number of observations, or is about the use of a particular tool, such as Hadoop. Such arbitrary thresholds on data size are problematic because, with innovation, even regular computers and off-the-shelf software have begun to manipulate very large data sets. Stata, a commonly used software by data scientists and statisticians, announced that one could now process between 2 billion to 24.4 billion rows using its desktop solutions. If Hadoop is the password to the big data club, Stata’s ability to process 24.4 billion rows, under certain limitations, has just gatecrashed that big data party.

It is important to realize that one who tries to set arbitrary thresholds to exclude others is likely to run into inconsistencies. The goal should be to define data science in a more exclusive, discipline- and platform-independent, size-free context where data-centric problem solving and the ability to weave strong narratives take center stage.

Given the controversy, I would rather consult others to see how they describe a data scientist. Why don’t we again consult the Chief Data Scientist of the United States? Recall Dr Patil told the Guardian newspaper in 2012 that a data scientist is that **unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data.** What is admirable about Dr Patil’s definition is that it is inclusive of individuals of various academic backgrounds and training and does not restrict the definition of a data scientist to a particular tool or subject it to a certain arbitrary minimum threshold of data size.

The other key ingredient for a successful data scientist is a behavioral trait: curiosity. A data scientist has to be one with a very curious mind, willing to spend significant time and effort to explore her hunches. In journalism, the editors call it having the nose for news. Not all reporters know where the news lies. Only those Who have the nose for news get the Story. Curiosity is equally important for data scientists as it is for journalists.

Rachel Schutt is the Chief Data Scientist at News Corp. She teaches a data science course at Columbia University. She is also the author of an excellent book, Doing Data Science. In an interview With the New York Times, Dr Schutt defined a data scientist as someone who is a part computer scientist, part software engineer, and part statistician (Miller, 2013). But that’s the definition of an average data scientist. “The best”, she contended, “tend to be really **curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them**.”