Course Syllabus

**Defining Data Science and What Data Scientists Do**

* Defining Data Science
* What is Data Science?
* Fundamentals of Data Science
* The Many Paths to Data Science
* Advice for New Data Scientists
* Data Science: The Sexiest Job in the 21st Century

**What Do Data Scientists Do?**

* A day in the Life of a Data Scientist
* Old problems, new problems, Data Science solutions
* Data Science Topics and Algorithms
* What is the cloud?
* What Makes Someone a Data Scientist?

**Data Science Topics**

* Foundations of Big Data
* How Big Data is Driving Digital Transformation
* What is Hadoop?
* Data Science Skills & Big Data
* Data Scientists at New York University
* Data Mining
* Quiz: Data Mining

**Deep Learning and Machine Learning**

* What's the difference?
* Neural Networks and Deep Learning
* Applications of Machine Learning
* Regression
* Quiz: Regression

**Data Science in Business**

* Applications of Data Science
* How Data Science is Saving Lives
* How Should Companies Get Started in Data Science?
* Applications of Data Science
* The Final Deliverable
* Quiz: The Final Deliverable

**Careers and Recruiting in Data Science**

* How Can Someone Become a Data Scientist?
* Recruiting for Data Science
* Careers in Data Science
* High School Students and Data Science Careers

**The Report Structure**

* The Report Structure
* Quiz: The Report Structure
* Final Assignment

**What is Data Science?**

Data Science is a process, not an event. It is the process of using data to understand different things, to understand the world. For me is when you have a model or hypothesis of a problem, and you try to validate that hypothesis or model with your data. Data science is the art of uncovering the insights and trends that are hiding behind data. It's when you translate data into a story.

We use storytelling to generate insight. And with these insights, you can make strategic choices for a company or an institution. Data science is a field about processes and systems to extract data from various forms of whether it is unstructured or structured form.

Data science is the study of data. Like biological sciences is a study of biology, physical sciences, it's the study of physical reactions. Data is real, data has real properties, and we need to study them if we're going to work on them.

Data Science involves data and some science. The definition or the name came up in the 80s and 90s when some professors were looking into the statistics curriculum, and they thought it would be better to call it data science. But what is Data Science? I'd see data science as one's attempt to work with data, to find answers to questions that they are exploring.

In a nutshell, it's more about data than it is about 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.

Data science is relevant today because we have tons of data available. We used to worry about lack of data. Now we have a data deluge. In the past, we didn't have algorithms, now we have algorithms. In the past, the software was expensive, now it's open source and free. In the past, we couldn't store large amounts of data, now for a fraction of the cost, we can have gazillions of datasets for a very low cost. So, the tools to work with data, the very availability of data, and the ability to store and analyze data, it's all cheap, it's all available, it's all ubiquitous, it's here. There's never been a better time to be a data scientist.

**Fundamentals of Data Science?**

Everyone you ask will give you a slightly different description of what Data Science is, but most people agree that it has a significant data analysis component. 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.

While there is more data available than ever, we have the computing power needed to make a useful analysis and reveal new knowledge. Data science can help organizations understand their environments, analyze existing issues, and reveal previously hidden opportunities. 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.

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. Good data scientists are curious people who ask questions to clarify the business need. The next questions are: "what data do we need to solve the problem, and where will that data come from?".

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.

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 Science is changing the way we work; it's changing the way we use data and it’s changing the way organizations understand the world.

**The Many Paths to 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. And I didn't want to be any of those. I want to be in business. And then I found data science a heck of a lot more interesting. I studied statistics, that's how I started.

I went through many different stages in my life where I wanted to be a singer and then a doctor. And then I realized that I was good at math. So I chose an area that was focused on quantitative analysis. And from then I do think that I wanted to work with data. Not necessarily data science as it's known today. The first time that I had contact with data science, when I was my first year as a mechanical engineering. And strategic consulting firms, they use data science to make decisions, it was my first contact with data science. I had a complicated problem that I needed to solve, and the usual techniques that we had at that time couldn't help with that problem.

I graduated with a math degree in the worst possible time, right after the economic crisis, and you actually had to be useful to get a job. So I went and got a degree in statistics. And then I worked enough jobs that were called data scientist that I suddenly became one. My undergraduate degree was in business, and I majored in politics, philosophy, and economics. And then I did a Masters in business analytics at New York University at the Stern School of Business.

When I left my undergrad, the first company I joined, it turned out that they were analyzing electronic point of sale data for retail manufacturers. And what we were doing was data science. But we only really started using that term much later. In fact, I'd say four or five years ago is when we started calling it analytics and data science. I had several options for my internship here in Canada. And one of the options was to work with data science. I used to work with project development. But I think that was a good choice. And then I start my internship with data science.

I'm a civil engineer by training, so all engineers work with data. I would say the conventional use of data science in my life started with transportation research. I started building large models trying to forecast traffic on streets, trying to determine congestion and greenhouse gas emissions or tailpipe emissions. So I think that's where my start was. And I started building these models when I was a graduate student at the University of Toronto. Started working with very large data sets, looking at household samples of, say, 150,000 households from half a million trips.

And that, too, I'm speaking from mid 90s when this was supposed to be a very large data set, but not in today's terms. But that's how I started. I continued working with it. And then I moved to McGill University where I was a professor of transportation engineering. And I built even bigger data models that involved data and analytics. And so I would say, yes, transportation research brought me to data science.

**Advice for New Data Scientists**

My advice to an aspiring data scientist is to be curious, extremely argumentative, and judgmental:

1. **Curiosity** is an absolute must. If you're not curious, you would not know what to do with the data.
2. **Judgmental** because if you do not have preconceived notions about things you wouldn't know where to begin with.
3. **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 may start at the wrong point. You may say that I thought I believed this, but now with data I know this.
4. **Comfort & 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
5. **Storytelling ability:** Once you have your analytics, 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, our findings will remain hidden, buried and nobody would know.

Your rise to prominence is pretty much relying on your ability to tell great stories. A starting point would be to see what your competitive advantage is. 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, for lets say, 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.

## Data Science: The Sexiest Job in the 21st Century

In the In the data-driven world, data scientists have emerged as a hot commodity. The chase is on to find the best talent in data science. Already, experts estimate that millions of jobs in data science might remain vacant for the lack of readily available talent. The global search for skilled data scientists is not merely a search for statisticians or computer scientists. In fact, the firms are searching for well-rounded individuals who possess the subject matter expertise, some experience in software programming and analytics, and exceptional communication skills.

Our digital footprint has expanded rapidly over the past 10 years. The size of the digital universe was roughly 130 billion gigabytes in 1995. By 2020, this number will swell to 40 trillion gigabytes. Companies will compete for hundreds of thousands, if not millions, of new workers needed to navigate the digital world. No wonder the prestigious Harvard Business Review called data science the sexiest job in the 21st century.

A report by the McKinsey Global Institute warns of huge talent shortages for data and analytics. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

Because the digital revolution has touched every aspect of our lives, the opportunity to benefit from learning about our behaviors is more so now than ever before. Given the right data, marketers can take sneak peeks into our habit formation. Research in neurology and psychology is revealing how habits and preferences are formed and retailers like Target are out to profit from it. However, the retailers can only do so if they have data scientists working for them. “For this reason, it is like an arms race to hire statisticians nowadays”, said Andreas Weigend, the former chief scientist at Amazon.com.

There is still the need to convince the C-suite executives of the benefits of data and analytics. It appears that the senior management might be a step or two behind the middle management in being informed of the potential of analytics-driven planning. Professor Peter Fader, who manages the Customer Analytics Initiative at Wharton, knows that executives reach the C-suite without having to interact with data. He believes that the real change will happen when executives are well-versed in data & analytics.

SAP, a leader in data and analytics, reported from a survey that 92% of the responding firms in its sample experienced a significant increase in their data holdings. At the same time, three-quarters identified the need for new data science skills in their firms. Accenture believes that the demand for data scientists may outstrip supply by 250,000 in 2015 alone. A similar survey of 150 executives by KPMG in 2014 found that 85% of the respondents did not know how to analyze data. Most organizations are unable to connect the dots because they do not fully understand how data and analytics can transform their business, Alwin Magimay, head of digital and analytics for KPMG UK, said in an interview in May 2015.

Bernard Marr writing for Forbes also raises concerns about the insufficient analytics talent. There just aren’t enough people with the required skills to analyze and interpret this information-transforming it from raw numerical (or other) data into actionable insights-the ultimate aim of any Big Data-driven initiative, he wrote. Bernard quotes a survey by Gartner of business leaders of whom more than 50% reported the lack of in-house expertise in data science.

Bernard reported on Walmart, which turned to crowd-sourcing for its analytics need. Walmart approached Kaggle to host a competition for analyzing its proprietary data. The retailer provided sales data from a shortlist of stores and asked the competitors to develop better forecasts of sales based on promotion schemes.

Given the shortage of data scientists, employers are willing to pay top dollars for the talent. Michael Chui, a principal at McKinsey, knows this too well. “Data science has become relevant to every company … There’s a war for this type of talent,” he said in an interview. Take Paul Minton, for example. He was making $20,000 serving tables at a restaurant. He had majored in math at college. Mr. Minton took a three-month programming course that changed everything. He made over $100,000 in 2014 as a data scientist for a web startup in San Francisco. Six figures, right off the bat … To me, it was astonishing, said Mr Minton.

Could Mr Minton be exceptionally fortunate, or are such high salaries the norm? Luck had little to do with it; the New York Times reported $100,000 as the average base salary of a software engineer and $112,000 for data scientists.

In the data-driven world, data scientists have emerged as a hot commodity. The chase is on to find the best talent in data science. Already, experts estimate that millions of jobs in data science might remain vacant for the lack of readily available talent. The global search for skilled data scientists is not merely a search for statisticians or computer scientists. In fact, the firms are searching for well-rounded individuals who possess the subject matter expertise, some experience in software programming and analytics, and exceptional communication skills.

Our digital footprint has expanded rapidly over the past 10 years. The size of the digital universe was roughly 130 billion gigabytes in 1995. By 2020, this number will swell to 40 trillion gigabytes. Companies will compete for hundreds of thousands, if not millions, of new workers needed to navigate the digital world. No wonder the prestigious Harvard Business Review called data science **the sexiest job in the 21st century**.

A report by the McKinsey Global Institute warns of huge talent shortages for data and analytics. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

Because the digital revolution has touched every aspect of our lives, the opportunity to benefit from learning about our behaviors is more so now than ever before. Given the right data, marketers can take sneak peeks into our habit formation. Research in neurology and psychology is revealing how habits and preferences are formed and retailers like Target are out to profit from it. However, the retailers can only do so if they have data scientists working for them. “For this reason, it is like an arms race to hire statisticians nowadays”, said Andreas Weigend, the former chief scientist at Amazon.com.

There is still the need to convince the C-suite executives of the benefits of data and analytics. It appears that the senior management might be a step or two behind the middle management in being informed of the potential of analytics-driven planning. Professor Peter Fader, who manages the Customer Analytics Initiative at Wharton, knows that executives reach the C-suite without having to interact with data. He believes that the real change will happen when executives are well-versed in data  
and analytics.

SAP, a leader in data and analytics, reported from a survey that 92% of the responding firms in its sample experienced a significant increase in their data holdings. At the same time, three-quarters identified the need for new data science skills in their firms. Accenture believes that the demand for data scientists may outstrip supply by 250,000 in 2015 alone. A similar survey of 150 executives by KPMG in 2014 found that 85% of the respondents did not know how to analyze data. Most organizations are unable to connect the dots because they do not fully understand how data and analytics can transform their business, Alwin Magimay, head of digital and analytics for KPMG UK, said in an interview in May 2015.

Bernard Marr writing for Forbes also raises concerns about the insufficient analytics talent. There just aren’t enough people with the required skills to analyze and interpret this information-transforming it from raw numerical (or other) data into actionable insights-the ultimate aim of any Big Data-driven initiative, he wrote. Bernard quotes a survey by Gartner of business leaders of whom more than 50% reported the lack of in-house expertise in data science.

Bernard reported on Walmart, which turned to crowd-sourcing for its analytics need. Walmart approached Kaggle to host a competition for analyzing its proprietary data. The retailer provided sales data from a shortlist of stores and asked the competitors to develop better forecasts of sales based on promotion schemes.

Given the shortage of data scientists, employers are willing to pay top dollars for the talent. Michael Chui, a principal at McKinsey, knows this too well. “Data science has become relevant to every company … There’s a war for this type of talent,” he said in an interview. Take Paul Minton, for example. He was making $20,000 serving tables at a restaurant. He had majored in math at college. Mr. Minton took a three-month programming course that changed everything. He made over $100,000 in 2014 as a data scientist for a web startup in San Francisco. Six figures, right off the bat … To me, it was astonishing, said Mr Minton.

Could Mr Minton be exceptionally fortunate, or are such high salaries the norm? Luck had little to do with it; the New York Times reported $100,000 as the average base salary of a software engineer and $112,000 for data scientists.