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

# **Welcome to the Course**

## **Course Introduction**

**Data Science: A Simple Explanation** Data science is like being a detective, but instead of solving crimes, data scientists solve problems using data. They collect, analyze, and interpret large amounts of information to find patterns and insights that can help businesses make better decisions. Imagine you have a big box of puzzle pieces (the data), and your job is to put them together to see the whole picture (the insights).

**Engaging Example** Think of data science as cooking a delicious meal. You start with various ingredients (data) like vegetables, spices, and proteins. By mixing them in the right way (analyzing the data), you create a tasty dish (insights) that can satisfy your guests (business needs). Just like a chef needs to know how to combine flavors, a data scientist needs to know how to work with different types of data to uncover valuable information.

* **Course Overview**: This introductory course explores the field of data science, its importance, and the role of data scientists in organizations.
* **Key Topics**:
  + Definition of data science and its relevance in today's data-driven world.
  + Skills and qualities needed to become a successful data scientist.
  + Interaction of Big Data and Cloud Computing.
  + Concepts of artificial intelligence, machine learning, and their applications in data science.
  + Overview of data literacy and the data ecosystem.
* **Learning Objectives**:
  + Understand what data science is and why it matters.
  + Explore various career paths in data science.
  + Gain insights from experienced data science professionals.
* **Course Structure**: The course is divided into three main modules, plus an optional module, with instructional videos, readings, and assessments.

## **Course Syllabus**

This course provides an introduction to the field of data science, including its fundamental concepts, various career paths, and essential skills. It explores what data science is and what data scientists do and offers advice for those interested in pursuing a career in this exciting field.

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

**Defining Data Science**

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

**What Do Data Scientists Do?**

* A Day in the Life of a Data Scientist
* Data Science Skills & Big Data
* Working on Different File Formats
* Data Science Topics and Algorithms
* Discussion Prompt: Introduce Yourself
* Reading: What Makes Someone a Data Scientist?

**Data Science Topics**

**Big Data and Data Mining**

* How Big Data is Driving Digital Transformation
* Introduction to Cloud
* Cloud for Data Science
* Foundations of Big Data
* Data Scientists at New York University
* What is Hadoop?
* Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
* Reading: Data Mining

**Deep Learning and Machine Learning**

* Artificial Intelligence and Data Science
* Generative AI and Data Science
* Neural Networks and Deep Learning
* Applications of Machine Learning
* Reading: Regression
* Lab: Exploring Data using IBM Cloud Gallery

**Applications and Careers in Data Science:**

**Data Science Application Domains**

* How Should Companies Get Started in Data Science?
* Old Problems with New Data Science Solutions
* Applications of Data Science
* How Data Science is Saving Lives
* Reading: The Final Deliverable

**Careers and Recruiting in Data Science**

* How Can Someone Become a Data Scientist?
* Recruiting for Data Science
* Careers in Data Science
* Importance of Mathematics and Statistics for Data Science (only name change)
* The Report Structure
* Reading: Infograph on roadmap

**Data Literacy for Data Science (Optional):**

**Understanding Data**

* Understanding Data
* Data Sources
* Working on Varied Data Sources and Types
* Reading: Metadata

**Data Literacy**

* Data Collection and Organization
* Relational Database Management System
* NoSQL
* Data Marts, Data Lakes, ETL, and Data Pipelines
* Considerations for Choice of Data Repository
* Data Integration Platforms

# **Defining Data Science**

## **Lesson Overview**

In this lesson, “Defining Data Science,” you begin your journey with an introduction to Data Science. Through the videos in this lesson, you will learn what data science is, the data scientist’s role in an organization, and what makes a skilled data scientist. You will hear from experts on how to acquire these skills.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| “What is Data Science” video | Hear from data science experts in the field explaining what data science is to them. |
| “Fundamentals of Data Science” video | This animated video touches upon some of the core attributes of data science, such as data analysis, varied sources of data, the data science process, the qualities of a good data scientist, and the role of a data scientist in an organization. |
| “The Many Paths to Data Science” video | Hear from graduate students and professionals discuss what led them into the field and why data science is a good fit for them. |
| “The sexiest job in the 21st Century” reading | Read an excerpt from the “Getting Started with Data Science” textbook and learn about the qualities of data science that attract people to the profession. |
| Practice quiz | Test your understanding of the previous reading. |
| “Advice for New Data Scientists” video | Hear from professor and author Dr. Murtaza Haider, PhD, an associate professor from the Ted Rogers School of Management, give his perspective on how to gain a competitive analysis in the data science field. |
| Practice quiz | Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson. |
| Glossary | Use this glossary of terms to review the terminology presented in this lesson. |
| Graded quiz | Test your knowledge from this lesson by taking the graded quiz. |

## **What is Data Science**

Data science is like being a detective, but instead of solving crimes, you're uncovering stories hidden in data. Imagine you have a big box of puzzle pieces (the data). Your job is to put those pieces together to see the bigger picture. This process involves asking questions, forming ideas (or hypotheses), and then using the data to find answers. Just like a detective uses clues to solve a mystery, data scientists use data to understand trends and insights that can help businesses make smart decisions.

For example, think about a bakery that wants to know which pastries are the most popular. A data scientist would collect sales data, analyze it, and discover that chocolate croissants sell best on weekends. This insight helps the bakery decide to bake more chocolate croissants for those busy days. So, data science is all about exploring data, asking questions, and finding answers that can lead to better choices.

**Summary of Data Science:**

* **Definition**: Data science is the process of using data to understand and solve problems. It involves creating models or hypotheses and validating them with data.
* **Purpose**: The goal is to uncover insights and trends from data, translating them into stories that inform strategic decisions for businesses or institutions.
* **Importance**: In today's world, there is an abundance of data available, and the tools to analyze it are more accessible and affordable than ever.
* **Relevance**: Data science is crucial because it helps organizations make data-driven decisions, find patterns, and predict outcomes.

## **Fundamentals of Data Science**

**In simple terms:** Data Science is the process of using data to find answers to questions or solve problems. It involves collecting data, analyzing it, and then using the insights gained to make decisions or predictions.

**Real-life examples:**

1. **Online Shopping:**
   * When you shop online, websites collect data about what you view, what you buy, and even what you leave in your cart. Data scientists analyze this data to understand customer preferences. For example, if many people are looking at a specific type of shoe but not buying it, the company might decide to lower the price or offer a discount to encourage purchases.
2. **Healthcare:**
   * Hospitals collect data on patient health records, treatments, and outcomes. Data scientists analyze this data to find patterns. For instance, they might discover that patients with a certain condition respond better to a specific treatment. This insight can help doctors make better treatment decisions for future patients.

**summary of** **Data Science**:

* **Definition**: Data Science is the process of using data to answer questions and solve problems.
* **Process**: It involves collecting data, analyzing it, and extracting insights to help organizations make informed decisions.
* **Analogy**: Think of data scientists as detectives who piece together information (data) to reveal a clear picture that guides actions.
* **Applications**: Data Science is used in various fields, such as online shopping (to understand customer behavior) and healthcare (to improve patient treatments).

## **The many paths to Data Science**

1. **What is Data Science?**: Data science is a new field that combines statistics and technology to analyze data and make decisions. It became popular around 2009-2011.
2. **Career Path**: Many people come to data science from different backgrounds, like engineering, business, or statistics. The speaker mentions how they started with a math degree and later got into data science through various jobs.
3. **Real-World Applications**: Data science is used in many industries, like retail and transportation, to solve complex problems. For example, it helps predict traffic patterns or analyze sales data.
4. **Learning and Growth**: The speaker emphasizes that they learned through experience and internships, showing that practical experience is important in this field.

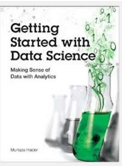
In summary, data science is about using data to find insights and solve problems, and many people from different fields can become data scientists.

* **Data Science Emergence**: Data science is a relatively new field that gained recognition around 2009-2011, combining statistics and technology to analyze data.
* **Diverse Backgrounds**: Individuals from various educational backgrounds, such as engineering, business, and mathematics, can enter the field of data science.
* **Practical Applications**: Data science is applied in different industries, like retail and transportation, to solve complex problems and make data-driven decisions.
* **Learning Journey**: The speaker highlights the importance of gaining practical experience through internships and jobs to develop skills in data science.

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

**Course Text Book: 'Getting Started with Data Science' Publisher: IBM Press; 1 edition (Dec 13 2015) Print.**

**Author: Murtaza Haider**



Prescribed Reading: Chapter 1 Pg. 4

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

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.

**For further reading**: [Is Data Science Still the Sexiest Job in the 21st Century?](https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century)  
-by Thomas H. Davenport and DJ Patil  
Harvard Business Review

## **Advice for new Data Scientists**

Curiosity and Storytelling in Data Science: A Simple Explanation

In the world of data science, being curious is like being a detective. You need to ask questions and dig deep into the data to uncover hidden insights. If you're not curious, you might miss out on important information that could help you understand the bigger picture. Think of it like exploring a new city; if you don’t ask questions or look around, you might not discover the best spots!

Once you've gathered and analyzed your data, the next step is to tell a story with it. This means taking the findings and presenting them in a way that others can understand and appreciate. Imagine you have a beautiful painting; if you just keep it hidden, no one will see it. But if you share it and explain what it represents, people can connect with it. In data science, your ability to tell a compelling story from your data can make your findings impactful and memorable.

Here's a brief summary of the key points:

* **Curiosity**: Essential for data scientists to explore and understand data. It drives the process of asking questions and seeking insights.
* **Judgment and Argumentation**: Having preconceived notions helps in forming hypotheses, which can be tested and modified based on data findings.
* **Storytelling**: After analyzing data, it's crucial to communicate the results effectively. A compelling story helps others understand and appreciate the insights derived from the data.
* **Competitive Advantage**: Identifying your unique strengths or interests in a specific field can guide your learning and skill development in data science.

## **Lesson Summary**

Data science is the practice of using data to understand and solve problems. Think of it as a way to make sense of the huge amounts of information we have in the world. Data scientists are like detectives who look for clues in data to uncover insights that can help businesses or organizations make better decisions.

How Does It Work?

1. **Finding the Problem:** First, data scientists identify what question or problem they want to solve. For example, a company might want to know why sales are dropping.
2. **Collecting Data:** Next, they gather data from various sources, like customer surveys, sales records, or social media.
3. **Analyzing Data:** After collecting the data, they analyze it to find patterns or trends. This is like piecing together a puzzle to see the bigger picture.
4. **Telling a Story:** Finally, they present their findings in a way that others can understand, often using visuals like charts or graphs. This helps people see the insights and make informed decisions.

Why is it Important?

Data science is important because it helps organizations make decisions based on facts rather than guesses. By understanding data, companies can improve their products, enhance customer experiences, and ultimately achieve better results.

Example: Improving Customer Experience in a Retail Store

* **Identifying the Problem:** A retail store notices a decline in customer satisfaction and sales. They want to understand why customers are unhappy and how to improve their experience.
* **Collecting Data:** The store collects data from various sources:
  + Customer feedback surveys
  + Sales data
  + Online reviews
  + In-store foot traffic patterns
* **Analyzing Data:** Data scientists analyze this information to find patterns. For instance, they might discover that:
  + Customers are unhappy with long wait times at checkout.
  + Certain products are frequently out of stock.
  + Online reviews mention poor customer service.
* **Telling a Story:** The data scientists present their findings to the management team, highlighting the key issues affecting customer satisfaction. They might use visualizations to show trends over time, making it easier for the team to understand the data.
* **Implementing Changes:** Based on the insights, the store decides to:
  + Hire more staff during peak hours to reduce wait times.
  + Improve inventory management to ensure popular products are always in stock.
  + Provide additional training for employees to enhance customer service.
* **Measuring Success:** After implementing these changes, the store continues to collect data to measure customer satisfaction and sales. They find that both have improved, demonstrating the effectiveness of using data science to solve real-world problems.

This example shows how data science can be applied in a practical setting to enhance customer experience and drive business success. If you have a specific industry or scenario in mind

Here's a concise summary of data science:

* **Definition:** Data science is the study of data to uncover insights and solve problems. It combines statistics, programming, and domain knowledge.
* **Process:**
  1. **Identify the Problem:** Determine what needs to be solved.
  2. **Collect Data:** Gather relevant information from various sources.
  3. **Analyze Data:** Look for patterns and trends in the data.
  4. **Communicate Findings:** Present insights in an understandable way, often using visuals.
* **Importance:** Data science helps organizations make informed decisions, improve products, and enhance customer experiences by relying on data rather than assumptions.

# **What Do Data Scientists Do?**

## **Lesson Overview: What Do Data Scientists Do?**

In the lesson "What Do Data Scientists Do?" you’ll dive into data science. The first video shows a day in the life of data scientists. You’ll also learn essential skills for becoming a good data scientist and why big data matters. You’ll explore handling different file types, study data science topics, and algorithms, and discuss the qualities that define a data scientist. The lesson ends with a summary video and a quiz to ensure you grasp this dynamic field.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| “A Day in the Life of a Data Scientist” video | Gain firsthand insights into the daily routines and challenges faced by data scientists, providing a practical glimpse into their roles. |
| “Data Science Skills and Big Data” video | Delve into the core skills required in the data science profession and understand the significance of big data in contemporary data analysis. |
| “Working on Different File Formats” video | Explore the intricacies of handling diverse file formats, a crucial skill for data scientists when dealing with various data sources. |
| “Data Science Topics and Algorithms” video | Dive into essential data science topics and algorithms that form the foundation of data analysis and decision-making. |
| Discussion Prompt: Introduce Yourself | Engage with fellow learners by introducing yourself, fostering a sense of community and collaborative learning. |
| “What Makes Someone a Data Scientist?” reading | Read an excerpt from "What Makes Someone a Data Scientist?" where the author addresses the ongoing debates surrounding the definition of data science and the elusive identity of a data scientist. |
| “Lesson Summary” video | Summarize and reinforce your understanding of the key concepts covered in the lesson, ensuring a comprehensive grasp of the material. |
| Practice quiz | Test your understanding of the previous reading. |
| Glossary | Use this glossary of terms to review the terminology presented in this lesson. |
| Graded quiz | Test your knowledge from this lesson by taking the graded quiz. |

## **A Day in the life of Data Scientist**

In the video, we learned about recommendation engines, which are systems that suggest products, services, or content to users based on their preferences and behaviors. Think of it like a friend who knows your taste in movies and always recommends the perfect film for you to watch. These engines analyze data, such as what you’ve liked or purchased in the past, to make personalized suggestions.

For example, if you often watch romantic comedies, a recommendation engine might suggest new romantic comedies that you haven’t seen yet. It uses patterns in your behavior to predict what you might enjoy next. This way, you spend less time searching and more time enjoying what you love!

Here's a brief summary of the content:

* The speaker shares experiences with **recommendation engines** and their applications in various fields.
* They discuss a project predicting **algae blooms** using artificial neural networks to help water treatment companies.
* Another example involves analyzing **complaints data** for Toronto Transit Commission (TTC) to find patterns related to weather conditions.
* The speaker discovered that complaints increased on days with **extreme weather**, linking it to user dissatisfaction.

This highlights the importance of data analysis in understanding patterns and improving services.

## **Data Science Skills and Big Data**

Data Science Skills

* **Programming**: Knowing languages like Python or R is crucial for analyzing data.
* **Statistics**: Understanding statistical methods helps in making sense of data and drawing conclusions.
* **Data Manipulation**: Skills in tools like SQL or libraries like Pandas allow you to clean and organize data.
* **Machine Learning**: Knowledge of algorithms helps in building models that can predict outcomes based on data.
* **Data Visualization**: Being able to present data clearly using tools like Tableau or Matplotlib is important for communicating findings.

Big Data

* **Definition**: Big data refers to extremely large datasets that traditional data processing software can't handle efficiently.
* **Characteristics**: It is often described by the "Three Vs":
  + **Volume**: The amount of data is massive.
  + **Velocity**: Data is generated and processed at high speeds.
  + **Variety**: Data comes in various formats (text, images, videos, etc.).

Importance

* **Insights**: Analyzing big data can reveal patterns and trends that help businesses make informed decisions.
* **Innovation**: It drives advancements in fields like healthcare, finance, and marketing by enabling data-driven strategies.

## **[Understanding Different Types of File Formats](https://www.coursera.org/learn/what-is-datascience/lecture/nVEjl/understanding-different-types-of-file-formats)**

In the video, we learned about various data file formats that data professionals often use. Think of these formats as different types of containers for information, each designed to hold and organize data in specific ways. For example, a CSV (Comma-Separated Values) file is like a simple list where each item is separated by a comma, making it easy to read and process. On the other hand, an XLSX file is like a multi-page notebook where you can have different sheets filled with organized rows and columns of data.

To illustrate, imagine you have a recipe book. If you write down each recipe on a separate page, that's similar to using an XLSX file. But if you jot down a shopping list with items separated by commas, that's like a CSV file. Each format has its own strengths and weaknesses, so understanding them helps you choose the right one for your needs.

Here's a brief summary of the topic on data file formats:

* **Data File Formats**: Different types of files used to store and organize data.
* **Common Formats**:
  + **Delimited Text Files**: Store data as text with values separated by delimiters (e.g., commas in CSV, tabs in TSV).
  + **XLSX**: Microsoft Excel format that allows multiple worksheets with organized rows and columns.
  + **XML**: A markup language for encoding data that is readable by both humans and machines.
  + **PDF**: A format for presenting documents consistently across devices.
  + **JSON**: A text-based format for transmitting structured data, widely used in web services.

Understanding these formats helps in choosing the right one for data storage and performance needs.

## **Data Science Topics and Algorithms**

Regression is like taking a taxi ride. When you get into a cab, there's a fixed starting fare, let's say $2.50. As the cab drives, the fare increases based on how far you go and how long you stay in the cab. So, if you travel more distance or spend more time in traffic, you pay more. Regression helps us understand the relationship between these factors: the base fare (the constant) and how the distance and time affect the total fare. It allows us to figure out these relationships even if we only know the total fare and not the details of the ride.

In simpler terms, regression is a way to find patterns in data. It helps us see how different things are connected, just like how your taxi fare is connected to the distance you travel and the time you spend in the cab.

Here's a brief summary of regression:

* **Definition**: Regression is a statistical method used to understand the relationship between different variables.
* **Taxi Ride Analogy**: Think of it like a taxi fare:
  + There’s a fixed starting fare (constant).
  + The fare increases based on distance traveled and time spent in the cab.
* **Purpose**: It helps identify patterns and relationships in data, allowing us to predict outcomes based on known variables.

# **What Makes Someone a Data Scientist?**

Now that you know what is in the book, it is time to put down some definitions. Despite their ubiquitous(सर्वव्यापी) use, consensus(एकमत) evades(टाळतो) the notions(कल्पना) of Big data and Data Science. The question, **Who is a data scientist?** is very much alive and being contested by individuals, some of whom are merely(फक्त) interested in protecting their discipline or academic turfs. In this section, I attempt to address these controversies and explain Why a narrowly construed definition of either Big data or Data science will result in excluding hundreds of thousands of individuals who have recently turned to the emerging field.

**Everybody loves a data scientist,** wrote Simon Rogers (2012) in the Guardian. Mr. Rogers also traced the newfound love for number crunching to a quote by Google's Hal Varian, who declared that ***the sexy job in the next ten years will be statisticians.***

Whereas 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 answers borders on hostility.

I define a data scientist as someone who finds solutions to problems by analyzing Big or small 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.

I define data science as something that data scientists do. Years ago, as an engineering student at the University of Toronto, I was stuck With the question: What is engineering? I wrote my master's thesis on forecasting housing prices and my doctoral dissertation on forecasting homebuilders' choices related to What they build, when they build, 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.

So briefly, I would argue that data science is what data scientists do.

Others have many different definitions. 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, offers certain thresholds to meet to be a data scientist. On pages 8 and 9 in Developing Analytic talent, Dr Granville describes the new data science professor as a non-tenured instructor at a non-traditional university, who publishes research results in online blogs, does not waste time writing grants, works from home, and earns more money than the traditional tenured professors. Suffice it to say that the thriving academic community of data scientists might disagree with Dr Granville.

Dr Granville 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."*

## **Lesson Summary**

**The Role of a Data Scientist: A Simple Explanation**  
Data scientists are like detectives of the digital world. They investigate data to find answers to important questions and solve real-world problems. For example, if many people complain about public transportation, a data scientist might look at weather data to see if bad weather caused those complaints. They use various tools and techniques to analyze data, much like a chef uses different ingredients to create a delicious dish.

**An Engaging Example**  
Imagine you have a big puzzle, and each piece represents a piece of information. A data scientist takes those pieces, examines them closely, and figures out how they fit together to tell a story. They might use special tools, like Python, to help them organize and analyze the data. Just like a storyteller captivates an audience, data scientists share their findings in a way that helps others understand the insights they’ve uncovered.

concise summary of the key points about data science and the role of data scientists:

* **Definition**: Data science is the field that involves analyzing data to uncover insights and solve problems.
* **Role of Data Scientists**: They investigate data to find explanations for issues, such as customer complaints or environmental challenges.
* **Skills Required**: Data scientists use statistics, machine learning, and programming languages like Python to analyze data.
* **Types of Data**: They work with various data formats, including structured (like tables) and unstructured (like emails).
* **Importance**: Data scientists blend technical skills with storytelling to communicate their findings effectively.

**Glossary: What Do Data Scientists Do?**

Welcome! This alphabetized glossary contains many of the terms in this course. These terms are important for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

| **Term** | **Definition** | **Video where the term is introduced** |
| --- | --- | --- |
| Comma-separated values (CSV) / Tab-separated values (TSV) | Commonly used format for storing tabular data as plain text where either the comma or the tab separates each value. | Working on Different File Formats |
| Data file types | A computer file configuration is designed to store data in a specific way. | Working on Different File Formats |
| Data format | How data is encoded so it can be stored within a data file type. | Working on Different File Formats |
| Data visualization | A visual way, such as a graph, of representing data in a readily understandable way makes it easier to see trends in the data. | Data Science Topics and Algorithms |
| Delimited text file | A plain text file where a specific character separates the data values. | Working on Different File Formats |
| Extensible Markup Language (XML) | A language designed to structure, store, and enable data exchange between various technologies. | Working on Different File Formats |
| Hadoop | An open-source framework designed to store and process large datasets across clusters of computers. | What Makes Someone a Data Scientist |
| JavaScript Object Notation (JSON) | A data format compatible with various programming languages for two applications to exchange structured data. | Working on Different File Formats |
| Jupyter notebooks | A computational environment that allows users to create and share documents containing code, equations, visualizations, and explanatory text. See Python notebooks. | Data Science Skills & Big Data |
| Nearest neighbor | A machine learning algorithm that predicts a target variable based on its similarity to other values in the dataset. | Working on Different File Formats |
| Neural networks | A computational model used in deep learning that mimics the structure and functioning of the human brain’s neural pathways. It takes an input, processes it using previous learning, and produces an output. | A Day in the Life of a Data Scientist |
| Pandas | An open-source Python library that provides tools for working with structured data is often used for data manipulation and analysis. | Data Science Skills & Big Data |
| Python notebooks | Also known as a “Jupyter” notebook, this computational environment allows users to create and share documents containing code, equations, visualizations, and explanatory text. | Data Science Skills & Big Data |
| R | An open-source programming language used for statistical computing, data analysis, and data visualization. | Data Science Skills & Big Data |
| Recommendation engine | A computer program that analyzes user input, such as behaviors or preferences, and makes personalized recommendations based on that analysis. | A Day in the Life of a Data Scientist |
| Regression | A statistical model that shows a relationship between one or more predictor variables with a response variable. | Data Science Topics and Algorithms |
| Tabular data | Data that is organized into rows and columns. | A Day in the Life of a Data Scientist |
| XLSX | The Microsoft Excel spreadsheet file format. | Working on Different File Formats |

## **Summary: What Do Data Scientists Do?**

* Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
* There are many paths to a career in data science; most, but not all, involve math, programming, and curiosity about data.
* New data scientists need to be curious, judgemental and argumentative.
* Knowledgeable data scientists are in high demand. Jobs in data science pays high salaries for skilled workers.
* The typical work day for a Data Scientist varies depending on what type of project they are working on.
* Many algorithms are used to bring out insights from data.
* Some key data science related terms you learned in this lesson include: outliers, model, algorithms, JSON, XML. CSV, and regression.