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

In this lesson, you have learned:

* 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 a little math, a little science, and a lot of curiosity about data.
* New data scientists need to be curious, judgemental and argumentative.
* Why data science is considered the sexiest job in the 21st century, paying high salaries for skilled workers.

**How do you get a better solution that is efficient?**

* 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

***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 date. 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.***

In this lesson, you have learned:

* 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.
* Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.

According to Erst and Young: “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”

There is no one definition of Big Data, but there are certain elements that are common across the different definitions, such as velocity, volume, variety, veracity, and value. These are the V's of Big Data.

Velocity is the speed at which data accumulates. Data is being generated extremely fast, in a process that never stops. Near or real-time streaming, local, and cloud-based technologies can process information very quickly. Volume is the scale of the data, or the increase in the amount of data stored. Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.

Variety is the diversity of the data.

Structured data fits neatly into rows and columns, in relational databases while unstructured

data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers, and video.

Variety also reflects that data comes from different sources, machines, people, and processes,

both internal and external to organizations. Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more.

Veracity is the quality and origin of data, and its conformity to facts and accuracy.

Attributes include consistency, completeness, integrity, and ambiguity.

Drivers include cost and the need for traceability.

With the large amount of data available, the debate rages on about the accuracy of data

in the digital age. Is the information real, or is it false? Value is our ability and need to turn data into value.

Value isn't just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction. The main reason that people invest time to understand Big Data is to derive value from it.

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benefits of new technologies. This digital change integrates digital technology into all areas of an organization resulting in fundamental changes to how it operates and delivers value to customers. It is an organizational and cultural change driven by Data Science, and especially Big Data. The availability of vast amounts of data, and the competitive advantage that analyzing it brings, has triggered digital transformations throughout many industries.

***Big data is data that is large enough and has enough volume and velocity that you cannot handle it with traditional database systems.***

Data mining is the process of analyzing dense volumes of data to find patterns, discover trends, and gain insight into how that data can be used. Data miners can then use those findings to make decisions or predict an outcome.

* **Big data** refers to data sets that are so massive, so quickly built, and so varied that they defy traditional analysis methods such as you might perform with a relational database. A new knowledge and insights are becoming available to everyone. Big data is often described in terms of five V's; velocity, volume, variety, veracity, and value.
* **Data mining** is the process of automatically searching and analyzing data, discovering previously unrevealed patterns. It involves preprocessing the data to prepare it and transforming it into an appropriate format. Once this is done, insights and patterns are mined and extracted using various tools and techniques ranging from simple data visualization tools to machine learning and statistical models.
* **Machine learning** is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it is learned without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms. Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.
* **Deep learning** is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can label and categorize information and identify patterns. It is what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct.
* **A neural network** in AI is a collection of small computing units called neurons that take incoming data and learn to make decisions over time. Neural networks are often layer-deep and are the reason deep learning algorithms become more efficient as the data sets increase in volume, as opposed to other machine learning algorithms that may plateau as data increases.
* **Data Science** is the process and method for extracting knowledge and insights from large volumes of disparate data. It's an interdisciplinary field involving mathematics, statistical analysis, data visualization, machine learning, and more. It's what makes it possible for us to appropriate information, see patterns, find meaning from large volumes of data and use it to make decisions that drive business.
* **AI** includes everything that allows computers to learn how to solve problems and make intelligent decisions.

In healthcare, data scientists use predictive analytics developed from data mining, data modeling, statistics, and machine learning to find the best options for patients. This type of predictive analytics examines all known factors for a disease, including gene markers, associated conditions, and environmental factors. It then recommends appropriate tests, suitable trials, and any suggested treatments. Developing more sophisticated big data analytics capabilities helps healthcare organizations move from basic descriptive analytics towards predictive insights, thanks to data science.

In the field of Disaster Preparedness, the ability to save lives using Data Science tools has been under development for many years. The use of predictive analytics tools is improving and providing new data analysis in a multitude of ways, alerting populations to danger faster than ever before. Earthquakes, hurricanes & tornados, floods, and volcanic eruptions can be predicted with the help of data science.

* The length and content of the final report will vary depending on the needs of the project.
* The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
* The report should present a thorough analysis of the data and communicate the project findings.

Before it can be useful, raw data must pass through various Data Science task categories, such as data management, data integration and transformation, data visualization, model building, model deployment, and model monitoring and assessment. To do these tasks, you need data asset management, code asset management, execution environments, and development environments.

The Data Science Task Categories are: Data Management, Data Integration and Transformation, Data Visualization, Model Building, Model Deployment, and Model Monitoring and Assessment. Data Science Tasks are supported by Data Asset Management, Code Asset Management, Execution Environments, and Development Environments.

* Data management is the process of collecting, persisting, and retrieving data securely, efficiently, and cost-effectively.
* Data Integration and Transformation, is the process of Extracting, Transforming, and Loading data. This is called “ETL” (Extracting, Transforming and Loading). Some of this data is distributed in multiple repositories. For example, a database, a data cube, and flat files. Use the Extraction process to extract data from these numerous repositories and save to a central repository like a Data Warehouse.
* Data Warehouses are primarily used to collect and store massive amounts of data for data analysis.
* Data Transformation is the process of transforming the values, structure, and format of data. After extracting the data, the next step is to transform the data.
* Data visualization is the graphical representation of data and information. You can use visualization to represent data in the form of charts, plots, maps, animations, etc. And data visualization conveys data more effectively for decision-makers. It is a crucial step in the data science process. Various forms of data visualizations include a bar chart, which compares the size of each component, a treemap, which displays hierarchy data, a line chart, which plots a series of data points over time, and a map chart, which displays data by location.
* Model building is where you train the data and analyze patterns with machine learning algorithms. The system ‘learns’ how to provide predictions or decisions by itself. You can then use this model to make predictions on new, unseen data. Model building can be done using a service called IBM Watson Machine Learning.
* Model deployment is the process of integrating a developed model into a production environment. In model deployment, a machine learning model is made available to third-party applications via APIs. Business users can access and interact with the data through these third-party applications. And this helps them make data-based decisions.
* Model monitoring and assessment run continuous quality checks to ensure a model’s accuracy, fairness, and robustness. Model monitoring uses tools like Fiddler to track the performance of deployed models in a production environment.
* Model assessment uses evaluation metrics like the F1 score, true positive rate, or the sum of squared error to understand a model's performance. A well-known example is the IBM Watson Open scale, which continuously monitors deployed machine learning and deep learning models. It will improve the accuracy and quality of your predictions.
* Code asset management provides a unified view where you manage an inventory of assets. When you want to develop a model, you may need to update it, fix bugs, or improve code features incrementally.
* Data asset management, also called digital asset management (DAM), is the organizing and managing of important data collected from different sources. DAM is performed on a DAM platform that allows versioning and collaboration. DAM platforms also support replication, backup, and access right management for the stored data.
* Development Environments, also called Integrated Development Environments, or “IDEs”, provide a workspace and tools to develop, implement, execute, test, and deploy source code. IDEs like IBM Watson Studio provide testing and simulation tools to emulate the real world so you can see how your code will behave after it is deployed. An execution environment has libraries to compile the source code and system resources that execute and verify the code.
* Fully integrated visual tools like IBM Watson Studio and IBM Cognos Dashboard Embedded cover all the previous tooling components, and can be used to develop deep learning and machine learning models.

Data management tools are MySQL, PostgreSQL, MongoDB, Apache CouchDB, Apache Cassandra, Hadoop File System, Ceph, and elastic search.

Data integration and transformation tools are Apache AirFlow, KubeFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, and NodeRED.

Data Visualization tools are Pixie Dust, Hue, Kibana, and Apache Superset.

Model deployment tools are Apache PredictionIO, Seldon, Kubernetes, Redhat OpenShift, Mleap, TensorFlow service, TensorFlow lite, and TensorFlow dot JS. Model monitoring tools are ModelDB, Prometheus, IBM AI Fairness 360, IBM Adversarial Robustness 360 Toolbox, and IBM AI Explainability 360. Code asset management tools are Git, GitHub, GitLab, and Bitbucket. And finally, data asset management tools are Apache Atlas, ODPi Egeria, and Kylo.

***Currently, the most famous development environment data scientists are using is “Jupyter,” which emerged as a tool for interactive Python programming. Jupyter now supports more than a hundred different programming languages through “kernels.”***

Difference between Open source and Free software: The Open-Source Initiative (OSI) champions open source, while the Free Software Foundation (FSF) defines free software. Open source is more business focused, while free software is more focused on a set of values. ***(Python is open source, and R is free software.)***

* Libraries are a collection of functions and methods that allow you to perform many actions without writing the code.
* Scientific computing libraries (contain built-in modules providing different functionalities, which you can use directly. They are also called frameworks.): Pandas (Data structure and tools) and NumPy (Arrays and matrices).
* Data visualization: methods to communicate with others and display meaningful results of an analysis. Matplotlib (plots and graphs, most popular) and Seaborn (heat maps, time series and violin plots)
* Libraries usually contain built-in modules providing different functionalities that can be used directly.
* An application programming interface (API) allows communication between two pieces of software. API is the part of the library you see while the library contains all the components of the program. REST APIs allow you to communicate through the internet and take advantage of resources like storage, data, artificially intelligent algorithms, and much more.
* A data set is a structured collection of data. Data embodies information represented as text, numbers, or media such as images, audio, or video files. A tabular data set comprises a collection of rows containing columns that store the information. One popular tabular data format is "comma separated values," or CSV. A CSV file is a delimited text file where each line represents a row, and a comma separates data values.
* Hierarchical or network data structures are typically used to represent relationships between data. Hierarchical data is organized in a tree-like format, whereas network data is stored as a graph.
* Data Asset eXchange (DAX) provides a curated collection of open data sets, both from IBM Research and trusted third-party sources. DAX open data sets include tutorial notebooks that provide basic and advanced walk throughs for developers.
* The Model Asset eXchange (MAX) is a free open source repository for ready-to-use and customizable deep learning microservices.
* Jupyter stands for Julia, Python, and R. However, it supports many other languages. A Jupyter Notebook is a browser-based application that allows you to create and share documents containing code, equations, visualizations, narrative text links, and more. A Jupyter Notebook allows a Data Scientist to record their data experiments and results that others can reuse.
* Jupyter Lab is a browser-based application that allows you to access multiple Jupyter Notebook files, other code, and data files. Jupyter Notebooks can be used with cloud-based services like IBM and Google Colab. They don't require any installation on your local machine.
* A notebook kernel is a computational engine that executes the code contained in a Notebook file.

**Python Codes**

* You can create headings by adding a # sign before a **word** or a **phrase** (There are six levels of headings.)
* You can create bold text by adding two asterisks or underscores before and after a word, phrase, or a sentence.
* You can display text in italics by adding a single asterisk or underscore before and after a word, phrase, or a sentence.
* You can display text in both bold and italic style at the same time by adding a three asterisks or underscores before and after a word, phrase or a sentence.
* To display a clickable link without a name, enclose the link in angle brackets. <<https://skills.network/> >
* To attach an URL link to the word “Skills Network”, type as follows:

[Skills Network](<https://skills.network/>)

Image

* Images can be rendered in the following format
* ­To create tables, use: **hyphens** (----) for column headers, **pipes** | to separate each column, **Text on a new line** to separate each row

Eg: | Country Name | Capital |

| -------------| ------ |

| United States | Washington DC |

***Jupyter architecture implements a two-process model with a kernel and a client. The client is the interface offering the user the ability to send code to the kernel. It is the browser in a Jupyter Notebook. The kernel executes the code and returns the result to the client for display. Jupyter Notebooks represent your code, metadata, contents, and outputs.***

* Anaconda is a free and open-source distributor for Python and R, the top languages used in data science and machine learning.
* Google Colaboratory (or 'GoogleColab') is a free Jupyter notebook environment that runs entirely in the cloud.​ GoogleColab Jupyter notebooks execute on a browser, and GoogleColab projects are stored on Google Drive and GitHub. > <
* R is a statistical programming language. It is a powerful tool for data processing and manipulation, statistical inference, data analysis, and machine learning algorithm. It is also known for producing great visualizations and contains packages to handle data analysis without the need to install additional libraries.
* RStudio is a popular integrated development environment for developing and running the R language source code and programs. R studio includes: a syntax-highlighting editor that supports direct code execution and a place where you can keep a record of your work, a Console for typing R commands, a workspace and History tab that shows the list of R objects you created during your R session and the history of all previous commands, and finally, Files, Plots, Packages, and Help tabs.