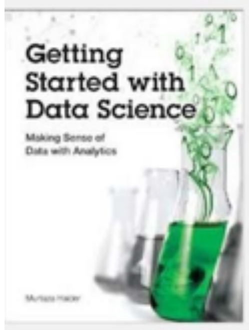
Course Text Book: 'Getting Started with Data Science' Publisher: IBM Press; 1 edition (Dec 13 2015) Print.

Author: Murtaza Haider



Prescribed Reading: Chapter 1 Pg. 12-15

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 use, consensus evades the notions 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 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*."

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

Skills Network

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

Author: Murtaza Haider

Prescribed Reading: Chapter 12 Pg. 529-531

Establishing Data Mining Goals

The first step in data mining requires you to set up goals for the exercise. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the expected level of accuracy and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost-benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, the cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals.

Selecting Data

The output of a data-mining exercise largely depends upon the quality of data being used. At times, data are readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys. The type of data, its size, and frequency of collection have a direct bearing on the cost of data mining exercise. Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

Preprocessing Data

Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from further consideration. At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. Data should be subject to checks to ensure integrity. Lastly, you must develop a formal method of dealing with missing data and determine whether the data are missing randomly or systematically.

If the data were missing randomly, a simple set of solutions would suffice. However, when data are missing in a systematic way, you must determine the impact of missing data on the results. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. Therefore, you must consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it.

Transforming Data

After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, variables may need to be transformed to help explain the phenomenon being studied. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

Often you need to transform variables from one type to another. It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual. This could help capture the non-linearities in the underlying behaviors.

Storing Data

The transformed data must be stored in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist. During data mining, new variables are created, which are written back to the original database, which is why the data storage scheme should facilitate efficiently reading from and writing to the database. It is also important to store data on servers or storage media that keeps the data secure and also prevents the data mining algorithm from unnecessarily searching for pieces of data scattered on different servers or storage media. Data safety and privacy should be a prime concern for storing data.

Mining Data

After data is appropriately processed, transformed, and stored, it is subject to data mining. This step covers data analysis methods, including parametric and non-parametric methods, and machine-learning algorithms. A good starting point for data mining is data visualization. Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in developing a preliminary understanding of the trends hidden in the data set.

Later sections in this chapter detail data mining algorithms and methods.

Evaluating Mining Results

After results have been extracted from data mining, you do a formal evaluation of the results. Formal evaluation could include testing the predictive capabilities of the models on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an "in-sample forecast". In addition, the results are shared with the key stakeholders for feedback, which is then incorporated in the later iterations of data mining to improve the process.

Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

Skills
Network

Big Data and Data Mining Lesson Glossary

Welcome! This glossary contains many of the terms in this lesson. 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
Analytics	The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights.	Data Scientists at New York University
Big Data	Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations.	How Big Data is Driving Digital Transformation
Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.	What is Hadoop?
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.	Introduction to Cloud
Chief Data Officer (CDO)	An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts.	How Big Data is Driving Digital Transformation
Chief Information Officer (CIO)	An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation.	How Big Data is Driving Digital Transformation
Cloud Computing	The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis.	Introduction to Cloud
Cloud Deployment Models	Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available to users, including public, private, and hybrid models.	Introduction to Cloud
Cloud Service Models	Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings.	Introduction to Cloud
Commodity Hardware	Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware.	What is Hadoop?
Data Algorithms	Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently.	Cloud for Data Science
Data Replication	A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures.	What is Hadoop?
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.	Data Scientists at New York University

Term	Definition	Video where the term is introduced
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own.	Data Scientists at New York University
Digital Change	The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers.	How Big Data is Driving Digital Transformation
Digital Transformation	A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes.	How Big Data is Driving Digital Transformation
Distributed Data	The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis.	What is Hadoop?
Hadoop	A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications.	Data Scientists at New York University
Hadoop Distributed File System (HDFS)	A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance.	What is Hadoop?
Infrastructure as a Service (IaaS)	A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them.	Introduction to Cloud
Java-Based Framework	Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
Map Process	The initial step in Hadoop’s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks.	What is Hadoop?
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.	Introduction to Cloud
On-Demand Self-Service	The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers.	Introduction to Cloud
Rapid Elasticity	The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use.	Introduction to Cloud
Reduce Process	The second step in Hadoop's MapReduce model is where results from the mapping process are aggregated and processed further to produce the final output, typically used for analysis.	What is Hadoop?
Replication	The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures.	What is Hadoop?
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.	Introduction to Cloud
Skills Network Labs (SN Labs)	Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development.	Cloud for Data Science

Term	Definition	Video where the term is introduced
Spilling to Disk	A technique used in memory-constrained situations where data is temporarily written to disk storage when memory resources are exhausted, ensuring uninterrupted processing.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
STEM Classes	Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science.	Data Scientists at New York University
Variety	The diversity of data types, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges.	Foundations of Big Data
Velocity	The speed at which data accumulates and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics.	Foundations of Big Data
Veracity	The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness.	Foundations of Big Data
Video Tracking System	A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports.	How Big Data is Driving Digital Transformation
Volume	The scale of data generated and stored is driven by increased data sources, higher-resolution sensors, and scalable infrastructure.	Foundations of Big Data
V's of Big Data	A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid generation, scale, diversity, quality, and value of data.	Foundations of Big Data

Deep Learning and Machine Learning Lesson Glossary

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
Artificial Neural Networks	Collections of small computing units (neurons) that process data and learn to make decisions over time.	Artificial Intelligence and Data Science
Bayesian Analysis	A statistical technique that uses Bayes' theorem to update probabilities based on new evidence.	Applications of Machine Learning
Business Insights	Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns.	Generative AI and Data Science
Cluster Analysis	The process of grouping similar data points together based on certain features or attributes.	Neural Networks and Deep Learning
Coding Automation	Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks.	Generative AI and Data Science
Data Mining	The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown.	Artificial Intelligence and Data Science
Decision Trees	A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions.	Applications of Machine Learning
Deep Learning Models	Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets.	Generative AI and Data Science
Five V's of Big Data	Characteristics used to describe big data: Velocity, volume, variety, veracity, and value.	Neural Networks and Deep Learning
Generative AI	A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data.	Generative AI and Data Science
Market Basket Analysis	Analyzing which goods tend to be bought together is often used for marketing insights.	Neural Networks and Deep Learning
Naive Bayes	A simple probabilistic classification algorithm based on Bayes' theorem.	Applications of Machine Learning
Natural Language Processing (NLP)	A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots.	Generative AI and Data Science
Precision vs. Recall	Metrics are used to evaluate the performance of classification models.	Applications of Machine Learning

Term	Definition	Video where the term is introduced
Predictive Analytics	Using machine learning techniques to predict future outcomes or events.	Neural Networks and Deep Learning
Synthetic Data	Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training.	Generative AI and Data Science