**Your IBM Cloud Feature Code:442dc2dec1c243b251549409c03a6565**

# Top 15 Data Science Applications for

# the Future

[**https://www.knowledgehut.com/blog/data-science/top-data-science-applications-for-future**](https://www.knowledgehut.com/blog/data-science/top-data-science-applications-for-future)

# Data Science for Finance: Benefits, Applications, Examples

[**https://www.knowledgehut.com/blog/data-science/data-science-for-finance**](https://www.knowledgehut.com/blog/data-science/data-science-for-finance)

Data science practices data mining, analysis, machine learning, and other such methods on structured, semi-structured, and unstructured data.

The capabilities of data science are endless. Almost all sectors, such as healthcare, finance, insurance, IT, pharmaceutical, manufacturing, energy, human resource, industries, marketing, etc., can leverage the benefits data science caters to every business.

**Week 1**

What is data science?

* 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.

Old Problem, New problems, Data Science Solutions

Organizations ultimately use data science to discover optimum solutions to existing problems. For example, In transport, Uber collects real-time user data to discover how many drivers are available, if more are needed, and if they should allow a surge charge to attract more drivers.  Uber uses data to put the right number of drivers in the right place, at the right time, for a cost the rider is willing to pay.

It takes gathering a lot of data, cleaning and preparing it, and then analyzing it to gain the insight needed to develop better solutions for today's enterprises. How do you get a better solution that is efficient? You must: Identify the problem and establish a clear understanding of it.  Gather the data for analysis. Identify the right tools to use and develop a data strategy. Case studies are also helpful in customizing a potential solution. Once these conditions exist and available data is extracted, you can develop a machine learning model. It will take time for an organization to refine best practices for data strategy using data science, but the benefits are worth it.

Data Science Topics and Algorithms

Using complicated machine learning algorithms does **not** always guarantee achieving a better performance. Occasionally, a simple algorithm such as k-nearest neighbour can yield a satisfactory performance comparable to the one achieved using a complicated algorithm. It all depends on the data.

# Cloud for Data Science

It allows you to bypass the physical limitations of your personal computer and the systems you are using.

Cloud is a godsend for data scientists. Primarily because you're able to take [the] your data, take your information and put it in the Cloud, put it in a central storage system. It allows you to bypass the physical limitations of the computers and the systems you're using and it allows you to deploy the analytics and storage capacities of advanced machines that do not necessarily have to be your machine or your company's machine. Cloud allows you not just to store large amounts of data on servers somewhere in California or in Nevada, but it also allows you to deploy  very advanced computing algorithms and the ability to do high-performance computing using machines that are not yours. Think of it as you have some information, you can't store it, so you send it to storage space,

let's call it Cloud, and the algorithms that you need to use, you don't have them with you. But then on the Cloud, you have those algorithms available. So What you do is you deploy those algorithms on very large datasets and you're able to do it even though your own systems, your own machines, your own computing environments were not allowing you to do so. So Cloud is beautiful. The other thing that Cloud is beautiful for is that it allows multiple entities to work with same data at the same time. You can be working with the same data that your colleagues in say  Germany and another team in India

and another team in Ghana, they are collectively working and they're able to do so because the information, and the algorithms, and the tools, and the answers, and the results, whatever they needed is available at a central place, which we call Cloud. Cloud is beautiful. 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. IBM offers the IBM Cloud, Amazon offers Amazon Web Services or AWS, and Google offers Google Cloud platform. IBM also provides Skills Network labs or SN labs to learners registered at any of the learning portals on the IBM Developer Skills Network, where you have access to tools like Jupyter Notebooks and Spark clusters so you can create your own data science project and develop solutions. With practice and familiarity, you will discover how the Cloud dramatically enhances productivity for data scientists.

**Week 2**

# What Makes Someone a Data Scientist?

* What is data science?

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.

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.

* How does it differ from statistics?
* What makes someone a data scientist?

# Foundations of Big Data

**Big Data:** 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

**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. Every 60 seconds, hours of footage are uploaded to YouTube which is generating data. Think about how quickly data accumulates over hours, days, and years.

**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.

The world population is approximately seven billion people and the vast majority are now using digital devices; mobile phones, desktop and laptop computers, wearable devices, and so on. These devices all generate, capture, and store data -- approximately 2.5 quintillion bytes every day

**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.

 Let's think about the different types of data; text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the Internet of Things.

**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?

80% of data is considered to be unstructured and we must devise ways to produce reliable and ccurate insights. The data must be categorized, analyzed, and visualized.

Data Scientists today derive insights from Big Data and cope with the challenges that these massive data sets present. The scale of the data being collected means that it’s not feasible to use conventional data analysis tools.

 However, lternative tools that leverage distributed computing power can overcome this problem.

Tools such as Apache Spark, Hadoop and its ecosystem provide ways to extract, load, analyze,

and process the data across distributed compute resources, providing new insights and knowledge.

This gives organizations more ways to connect with their customers and enrich the services

they offer

**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.

# What is Hadoop?

Traditionally in computation and processing data we would bring the data to the computer. You'd wanna program and you'd bring the data into the program.  In a big data cluster  what Larry Page and Sergey Brin came up with is very pretty simple is they took the data and they sliced it into pieces and they distributed each and they replicated each piece or triplicated each piece and they would send it

the pieces of these files to thousands of computers first it was hundreds but then now it's thousands

now it's tens of thousands. And then they would send the same program to all these computers in the cluster. And each computer would run the program on its little piece of the file and send the results back. The results would then be sorted and those results would then be redistributed back to another process. The first process is called a map or a mapper process and the second one was called a reduce process.

Yahoo hired someone named Doug Cutting who had been working on a clone or a copy of the Google big data architecture and now that's called Hadoop.

At the bottom of data science you see probability and statistics. You see algebra, linear algebra you see programming and you see databases.

what we've seen is that the combination of traditional [technique] areas computer science probability, statistics, mathematics all coming together in this thing that we call Decision Sciences.

According to Dr. White, most of the components of data science, such as probability, statistics, linear algebra, and programming, have been around for many decades but now we have the computational capabilities to apply combine them and come up with new techniques and learning algorithms.

# How Big Data is Driving Digital Transformation

[Music]

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.

Netflix moved from being a postal DVD lending system to one of the world’s foremost video

streaming providers, the Houston Rockets NBA team used data gathered by overhead cameras

to analyze the most productive plays, and Lufthansa analyzed customer data to improve its service.

Organizations all around us are changing to their very core.

Let’s take a look at an example, to see how Big Data can trigger a digital transformation,

not just in one organization, but in an entire industry.

In 2018, the Houston Rockets, a National Basketball Association, or NBA team, raised their game

using Big Data.

The Rockets were one of four NBA teams to install a video tracking system which mined

raw data from games.

They analyzed video tracking data to investigate which plays provided the best opportunities

for high scores, and discovered something surprising.

Data analysis revealed that the shots that provide the best opportunities for high scores

are two-point dunks from inside the two-point zone, and three-point shots from outside the

three-point line, not long-range two-point shots from inside it.

This discovery entirely changed the way the team approached each game, increasing the

number of three-point shots attempted.

In the 2017-18 season, the Rockets made more three-point shots than any other team in NBA

history, and this was a major reason they won more games than any of their rivals.

In basketball, Big Data changed the way teams try to win, transforming the approach to the

game.

Digital transformation is not simply duplicating existing processes in digital form; the in-depth

analysis of how the business operates helps organizations discover how to improve their

processes and operations, and harness the benefits of integrating data science into

their workflows.

Most organizations realize that digital transformation will require fundamental changes to their

approach towards data, employees, and customers, and it will affect their organizational culture.

Digital transformation impacts every aspect of the organization, so it is handled by decision

makers at the very top levels to ensure success.

The support of the Chief Executive Officer is crucial to the digital transformation process,

as is the support of the Chief Information Officer, and the emerging role of Chief Data

Officer.

But they also require support from the executives who control budgets, personnel decisions,

and day-to-day priorities.

This is a whole organization process.

Everyone must support it for it to succeed.

There is no doubt dealing with all the issues that arise in this effort requires a new mindset,

but Digital Transformation is the way to succeed now and in the future.

# Data Science Skills & Big Data

According to Dr. White, his students, who are mostly aspiring data scientists, need to learn many tools such as Python, UNIX commands, pandas, and Jupyter notebook.

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

**Data Mining**

## **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.

## Deep Learning and Machine Learning

The term 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. The concurrent development of enormous compute power in distributed networks and new tools and techniques for data analysis means that organizations now have the power to analyze these vast data sets. 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.

Artificial neural networks, often referred to simply as neural networks, take inspiration from biological neural networks, although they work quite a bit differently.

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.

Now that you have a broad understanding of the differences between some key AI concepts, there is one more differentiation that is important to understand that between Artificial Intelligence and Data Science.

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. Data Science can use many of the AI techniques to derive insight from data. For example, it could use machine learning algorithms and even deep learning models to extract meaning and draw inferences from data.

There is some interaction between AI and Data Science, but one is not a subset of the other. Rather, Data Science is a broad term that encompasses the entire data processing methodology while AI includes everything that allows computers to learn how to solve problems and make intelligent decisions.

Both AI and Data Science can involve the use of big data. That is, significantly large volumes of data.

#### Question

Neural networks have been around for decades, but due to religious reasons, people decided not to develop them any more because a neural network mimics the brain in the way it learns the data.



False.



True.

### Correct

Correct. They were abandoned for some time because they were computationally very expensive.

#### Question

Which of the following are use cases for deep learning?



Predicting the prices of houses using features such as number of bedrooms, square footage, and proximity to amenities.



Speech recognition.



Classifying images at a large scale.

Netflix uses machine learning to recommend movies to you based on movies that you have already watched and liked or disliked.



True.



False.

### 1.

Question 1

Regression is a statistical technique developed by Blaise Pascal.

**1 / 1 point**



False.



True.

**Correct**

### 2.

Question 2

What did the author's research reveal about proximity to large shopping centres?

**1 / 1 point**



The author discovered that houses located more than 2.5 kms to shopping centres sold for less than the rest.



The author discovered that proximity to large shopping centres didn't have any significant impact on the prices of housing units.



The author discovered that houses located more than 5 kms to shopping centres sold for less than the rest.



The author discovered that proximity to large shopping centres had a nonlinear impact on the housing prices.

**Correct**

Correct.

### 3.

Question 3

"What are typical land taxes in a house sale?" is a question that can be put to regression analysis.

**1 / 1 point**



False



True.

**Correct**

Correct.

# Lesson Summary

In this lesson, you have learned:

* The differences between some common Data Science terms, including Deep Learning and Machine Learning.
* Deep Learning is a type of Machine Learning that simulates human decision-making using neural networks.
* Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
* How to use regression to analyze data.

**Week 3**

# How Data Science is saving lives

Using Data Science techniques to understand and analyze the large data sets available

today has a huge impact on human lives.

It can provide targeted information to help healthcare professionals give the best treatment

to patients, or help predict natural disasters so that people can prepare early, and much

more besides.

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.

Every individual physician has their own store of knowledge gained from their studies, interests,

and experiences.

Data science systems that use predictive analytics ensure that all physicians can also access

the latest information about the disease, tests, and treatment plans, tailored to their

specific patient.

With this type of system, every physician has access to the same knowledge, and the

best options can be consistently offered, improving patient outcomes.

For example, a study by the Boston Consulting Group and AdvaMedDx, an industry association

of medical diagnostics companies, examined the barriers to the adoption of potentially

lifesaving diagnostic tests for patients with a specific cancer and a particular gene marker.

The study discovered that the biggest factor in the patient being offered a specific test

was the patient’s oncologist, who may or may not have known about the test and its

relationship to the gene marker.

By providing extra information through data science tools, physicians can be made aware

of the most helpful tests and treatments for a specific patient.

There are many opportunities to explore other ways to mine data, such as from electronic

medical records for different types of medical research.

Schools such as the NorthShore University HealthSystem in suburban Chicago, a leader

in the implementation of Electronic Medical Records (EMR) systems, now offer guidance

on data mining.

It is the first healthcare provider in America to be awarded the highest level of EMR deployment

for both inpatient and outpatient care.

This remarkable effort has generated much-anonymized data available for innovative analytics research.

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.

Large, high-quality data sets can be used to predict the occurrence of numerous types

of natural disasters, which can be the difference between life and death for thousands of people.

Earthquakes, hurricanes & tornados, floods, and volcanic eruptions can be predicted with

the help of data science.

Recent research at the University of Warwick in the UK used social media content such as

photos and keywords to track the development of floods, hurricanes and other weather events.

When added to the information recorded by scientists and weather stations, this type

of data can be used to improve the predictions for localised weather events.

Because the real benefit of this knowledge is so important, schools are starting to include

this type of data science education in their curriculum.

For instance, the University of Chicago Graham School offers a Master of Science course in

Threat and Response Management.

Data science tools enable organizations to analyse vast quantities of data from widely

different sources, and present that information in a way that allows data scientists to gain

new knowledge, in some cases, saving hundreds of lives

# How Should Companies Get Started in Data Science?

At the end of the day, for businesses, they know one thing, that if they are unable to measure something, they are unable to improve it. And if they are unable to measure their costs, they are unable to reduce them.

If they're unable to measure their profits, they are unable to increase them. So the first thing a company has to do is to start recording information, start capturing data, data about costs.

And the differentiate it by labor costs and material cost, the cost to how much it cost to sell one product and the total cost. And then you look at the revenue, where's your revenue coming from?

Is 80% of your revenue coming from 20% of your customers? Or is it the other way around?

So first thing first, start capturing data. Once you have data, then you can apply algorithms and analytics to it. So the first thing to do would be to capture data. If you're not capturing it, start capturing it. If you're capturing it, archive it. Do not overwrite on your old data thinking you don't need it anymore. Data never gets old. Data is always relevant, even if it's 100 years old, 200 years old.

It is relevant to you and and your firm and your success. So keep data, capture it, archive it, make sure nothing goes to waste. Make sure there's a consistency. So someone 20 years later trying to nderstand, that data should be able to do so, so have proper documentation. Do it now. Put the best practices for data archiving in place the moment you start a business. And if you're already in business and you haven't done it, do it now.

 >> Start measuring things. Too many companies haven't measured things properly for a decade and, then they decide, they want data science. Data science inside a company

is only going to be as valuable as the data collected. Garbage in, garbage out is a rule in any sort of analysis.

>> If something is not measured, it's very difficult to improve it or to change it. So the very first step is measurement. If companies have existing data, then they should start looking at it and

cleaning it. If they don't have existing data, then they need to start collecting it.

>> I think to look for a team who love to work as a data scientist.

>> The first stop is to have employees, that they are interested on data science.

because if you don't have interest in your company, you will not have engagement.

>> Companies should remember, that it's key to have a team.

So it's not one data scientist, but a team of them, that each of them have strengths in different areas of data science.

# Applications of Data Science

Data science and big data are making an undeniable impact on businesses, changing day-to-day operations, financial analytics, and especially interactions with customers. It's clear that businesses can gain enormous value from the insights data science can provide. But sometimes it's hard to see exactly how. So let's look at some examples. In this era of big data, almost everyone generates masses of data every day, often without being aware of it. This digital trace reveals the patterns of our online lives. If you have ever searched for or bought a product on a site like Amazon, you'll notice that it starts making recommendations related to your search.

This type of system known as a recommendation engine is a common application of data science.

Companies like Amazon, Netflix, and Spotify use algorithms to make specific recommendations derived from customer preferences and historical behavior. Personal assistants like Siri on Apple devices use data science to devise answers to the infinite number of questions end users may ask.

Google watches your every move in the world, you're online shopping habits, and your social media.

Then it analyzes that data to create recommendations for restaurants, bars, shops, and other attractions based on the data collected from your device and your current location. Wearable devices like Fitbits, Apple watches, and Android watches add information about your activity levels,

sleep patterns, and heart rate to the data you generate. Now that we know how consumers generate data, let's take a look at how data science is impacting business. In 2011, McKinsey & Company said that data science was going to become the key basis of competition. Supporting new waves of

productivity, growth, and innovation. In 2013, UPS announced that it was using data from customers, drivers,

and vehicles, in a new route guidance system aimed to save time, money, and fuel. Initiatives like this support the statement that data science will fundamentally change the way businesses compete and operate. How does a firm gain a competitive advantage? Let's take Netflix as an example. Netflix collects and analyzes massive amounts of data from millions of users, including which shows people are watching at what time a day when people pause, rewind, and fast-forward, and which shows directors and actors they search for. Netflix can be confident that a show will be a hit before filming even begins by analyzing users preference for certain directors and acting talent, and discovering which combinations people enjoy. Add this to the success of earlier versions of a show and you have a hit. For example, Netflix knew many of its users had streamed to the work of David Fincher.

They also knew that films featuring Robin Wright had always done well, and that the British version of

House of Cards was very successful. Netflix knew that significant numbers of people who liked Fincher also liked Wright. All this information combined to suggest that buying the series would be a good investment for the company. They were right. It was a huge hit. Thanks to data science, Netflix knows what people want before they do.

### **The Final Deliverable**

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. The data scientist should then use the insights to build the narrative to communicate the findings. In academia, the final deliverable is in the form of essays and reports. Such deliverables are usually 1,000 to 7,000 words in length. In consulting and business, the final deliverable takes on several forms. It can be a small document of fewer than 1,500 words illustrated with tables and plots, or it could be a comprehensive document comprising several hundred pages. Large consulting firms, such as McKinsey and Deloitte,I routinely generate analytics-driven reports to communicate their findings and, in the process, establish their expertise in specific knowledge domains.

Let's review the "United States Economic Forecast", a publication by the Deloitte University Press. This document serves as a good example for a deliverable that builds narrative from data and analytics. The 24-page report focuses on the state of the U.S. economy as observed in December 2014. The report opens with a **grabber** highlighting the fact that contrary to popular perception, the economic and job growth has been quite robust in the United States. The report is not merely a statement of facts.

In fact, it is a carefully crafted report that cites Voltaire and follows a distinct theme. The report focuses on the **good news** about the U.S. economy. These include the increased investment in manufacturing equipment in the U.S. and the likelihood of higher consumer consumption resulting from lower oil prices.

The Deloitte report uses time series plots to illustrate trends in markets. The GDP growth chart shows how the economy contracted during the Great Recession and has rebounded since then. The graphic presents four likely scenarios for the future. Another plot shows the changes in consumer spending. The accompanying narrative focuses on income inequality in the U.S. and refers to Thomas Pikkety's book on the same. The Deloitte report mentions many consumers did not experience an increase in their real incomes over the years, while they still maintained their level of spending. Other graphics focused on housing, business, and government sectors, international trade, labor, and financial markets, and prices. The appendix carries four tables documenting data for the four scenarios discussed in the report.

Deloitte's "United States Economic Forecast" serves the very purpose that its authors intended. The report uses data and analytics to generate the likely economic scenarios. It builds a powerful narrative in support of the thesis statement that the U.S. economy is doing much better than most would like to believe. At the same time, the report shows Deloitte to be a competent firm capable of analyzing economic data and prescribing strategies to cope with the economic challenges.

Now consider if we were to exclude the narrative from this report and presented the findings as a deck of PowerPoint slides with eight graphics and four tables. The PowerPoint slides would have failed to communicate the message that the authors carefully crafted in the report citing Piketty and Voltaire. I consider Deloitte's report a good example of storytelling with data and encourage you to read the report to decide for yourself whether the deliverable would have been equally powerful without the narrative.

Now, let us work backward from the Deloitte report. Before the authors started their analysis, they must have discussed the scope of the final deliverable. They would have deliberated the key message of the report and then looked for the data and analytics they needed to make their case. The initial planning and conceptualizing of the final deliverable is therefore extremely important for producing a compelling document. Embarking on analytics, without due consideration to the final deliverable, is likely to result in a poor-quality document where the analytics and narrative would struggle to blend.

**Congratulations! You passed!**

**Grade received** 100%

**Latest Submission Grade** 100%

**To pass** 66% or higher

Go to next item

### 1.

Question 1

The Untied States Economic Forecast is a publication by McKinsey University Press.

**1 / 1 point**



True.



False.

**Correct**

Incorrect.

### 2.

Question 2

The report discussed in the reading successfully did the job of using data and analytics to generate the likely economic scenarios.

**1 / 1 point**



False.



True.

**Correct**

Correct.

### 3.

Question 3

According to the reading, it is recommended that a team waits until the results of analytics are out before they can decide on the final deliverable.

**1 / 1 point**



True



False

**Correct**

Correct. In order to produce a compelling narrative, initial planning and conceptualizing of the final deliverable is of extreme importance.

# Lesson Summary

In this lesson, you have learned:

* Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
* That companies can start on their data science journey by capturing data. Once they have data, they can begin analysing it.
* Some ways that data is generated by consumers.
* How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
* The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

## **Careers and Recruiting in Data Science**

How Can Someone Become a Data Scientist?

A real data scientist, the high-end data scientists, are mostly PhDs.They often come out of physics, out of statistics,they have to have a computer science background, they have to have a math background,

they have to know about databases and statistics and probability and all that stuff.However, if you're coming into a data science team, I think the first skills you need is you need to know how to program,

at least have some computational thinking, so having taken a programing course, you need to know some algebra, at least up to analytics, geometry, and hopefully some calculus, some basic probability, some basic statistics, I mean really have to understand the difference and different statistical distributions, and database. I mean, one of the easiest places to start is relational databases, which stores lots and lots of our data so people can first walk before they can run by at least understanding about computers and databases and how we store things and if you understand relational databases nowadays you can still, just with that understanding, use big data clusters as if they were just a big relational database. You don't have to really have understand the whole MapReduce programming model. But then, as you go further up in the field, then you have to know a lot of computer science theory and statistics, it's really, and probability, it's really the intersection of them that the high end data scientists, the PhD data scientists work with. (music)

Play video starting at :2:4 and follow transcript2:04 I do a lot of self-learning. I think everybody these days, I mean, I learned about Hadoop all by myself, I read some articles, I watched some videos, I thought, I played, although I'm a builder, I'm a tinkerer, so if I wanna figure out how to do something, I build it. I mean, my first HPC cluster I heard about this term a Beowulf cluster, I mean, yeah, what the hell's that? So I looked it up and said, oh, it's just a bunch of computers hooked together

with a TCP/IP network, that's pretty easy, so we get a grant from Citi Bank and we built a five thing cluster and I said, oh, well, that's HPC. I said, I had one of the first HPC clusters at the university, it was tiny but a lot of our researchers loved it because they could run stuff 40 and 50 times faster.

So I think one of the ways you learn things is you do them, you have to do them, and these online learning platforms especially now that we have things like IPython and Jupyter Notebooks and I guess Zeppelin means that you can actually go in and take some of these courses and you can do things right then and you can see them and feel them and play with them and, at that point, you know, you'll start to get your head around what is actually happening. Motivation is the key problem in all of these,

is how to keep people motivated and I think the badge system that the, what was it, Big Data University has, is one of the ways is how do you get people to keep going through. But if they want to, they can.

It's up to the individual to. So they have to understand what the goal is.

(music)

Play video starting at :4:5 and follow transcript4:05

The place it can't sit

is probably under the CIO, the Chief Information Officer.

CIOs current chief information officers in many companies

got there from an accounting background

or a finance background, they're clueless.

Sorry.

But they really, it has to come out of the research side.

So you'll find data scientists primarily in companies

that have some research agenda, pharmaceuticals,

finance, all of, any technology company.

If you look at, we can't keep some of our

PhD data scientists in our program,

they are now at Facebook,

they're at Linkedin, they're at Uber, they're at Lyft,

because the demand out there for the PhD level

data scientist is just unbelievable.

They make large amounts of money

and they're playing with problems

that are really, really neat.

How do you schedule the Uber cars?

You have enormous amounts of data.

# Lesson Summary

In this lesson, you have learned:

* Data Scientists need programming, mathematics, and database skills, many of which can be gained through self-learning.
* Companies recruiting for a Data Science team need to understand the variety of different roles Data Scientists can play, and look for soft skills like storytelling and relationship building as well as technical skills.
* High school students considering a career in Data Science should learn programming, math, databases, and, most importantly practice their skills.

## The Report Structure

### **The Report Structure**

Before starting the analysis, think about the structure of the report. Will it be a brief report of five or fewer pages, or will it be a longer document running more than 100 pages in length? The structure of the report depends on the length of the document. A brief report is more to the point and presents a summary of key findings. A detailed report incrementally builds the argument and contains details about other relevant works, research methodology, data sources, and intermediate findings along with the main results.

I have reviewed reports by leading consultants including Deloitte and McKinsey. I found that the length of the reports varied depending largely on the purpose of the report. Brief reports were drafted as commentaries on current trends and developments that attracted public or media attention. Detailed and comprehensive reports offered a critical review of the subject matter with extensive data analysis and commentary. Often, detailed reports collected new data or interviewed industry experts to answer the research questions.

Even if you expect the report to be brief, sporting five or fewer pages, I recommend that the deliverable follow a prescribed format including the cover page, table of contents, executive summary, detailed contents, acknowledgments, references, and appendices (if needed).

I often find the cover page to be missing in documents. It is not the inexperience of undergraduate students that is reflected in submissions that usually miss the cover page. In fact, doctoral candidates also require an explicit reminder to include an informative cover page. I hasten to mention that the business world sleuths are hardly any better. Just search the Internet for reports and you will find plenty of reports from reputed firms that are missing the cover page.

At a minimum, the cover page should include the title of the report, names of authors, their affiliations, and contacts, the name of the institutional publisher (if any), and the date of publication. I have seen numerous reports missing the date of publication, making it impossible to cite them without the year and month of publication. Also, from a business point of view, authors should make it easier for the reader to reach out to them. Having contact details at the front makes the task easier.

"A table of contents (ToC)" is like a map needed for a trip never taken before. You need to have a sense of the journey before embarking on it. A map provides a visual proxy for the actual travel with details about the landmarks that you will pass by in your trip. The ToC with main headings and lists of tables and figures offers a glimpse of what lies ahead in the document. Never shy away from including a ToC, especially if your document, excluding cover page, table of contents, and references, is five or more pages in length.

Even for a short document, I recommend an "abstract" or an "executive summary". Nothing is more powerful than explaining the crux of your arguments in three paragraphs or less. Of course, for larger documents running a few hundred pages, the executive summary could be longer. An "introductory section" is always helpful in setting up the problem for the reader who might be new to the topic and who might need to be gently introduced to the subject matter before being immersed in intricate details. A good follow-up to the introductory section is a review of available relevant research on the subject matter. The length of the literature review section depends upon how contested the subject matter is. In instances where the vast majority of researchers have concluded in one direction, the literature review could be brief with citations for only the most influential authors on the subject. On the other hand, if the arguments are more nuanced with caveats aplenty, then you must cite the relevant research to offer adequate context before you embark on your analysis. You might use the literature review to highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

In the "methodology" section, you introduce the research methods and data sources you used for the analysis. If you have collected new data, explain the data collection exercise in some detail. You will refer to the literature review to bolster your choice for variables, data, and methods and how they will help you answer your research questions.

The results section is where you present your empirical findings. Starting with descriptive statistics (**see Chapter 4, "Serving Tables"**) and illustrative graphics (**see Chapter S, "Graphic Details" for plots and Chapter 10, "Spatial Data Analytics" for maps**), you will move toward formally testing your hypothesis (**see Chapter 6, "Hypothetically Speaking"**).

In case you need to run statistical models, you might turn to regression models (**see Chapter 7, "Why Tall Parents Don't Have Even Taller Children"**) or categorical analysis (**see Chapters 8, "To Be or Not to Be" and 2., "Categorically Speaking About Categorical Data"**). If you are working with time-series data, you can turn to Chapter 11, **Doing Serious Time with Time Series.** You can also report results from other empirical techniques that fall under the general rubric of data mining (**see Chapter 12, "Data Mining for Gold"**). Note that many reports in the business sector present results in a more palatable fashion by holding back the statistical details and relying on illustrative graphics to summarize the results.

The results section is followed by the discussion section, where you craft your main arguments by building on the results you have presented earlier.

The "discussion section" is where you rely on the power of narrative to enable numbers to communicate your thesis to your readers. You refer the reader to the research question and the knowledge gaps you identified earlier. You highlight how your findings provide the ultimate missing piece to the puzzle.

Of course, not all analytics return a smoking gun. At times, more frequently than I would like to acknowledge, the results provide only a partial answer to the question and that, too, with a long list of caveats.

In the "conclusion" section, you generalize your specific findings and take on a rather marketing approach to promote your findings so that the reader does not remain stuck in the caveats that you have voluntarily outlined earlier. You might also identify future possible developments in research and applications that could result from your research. What remains is housekeeping, including a list of references, the acknowledgment section (**acknowledging the support of those who have enabled your work is always good**), and "appendices", if needed.

## **Have You Done Your Job as a Writer?**

As a data scientist, you are expected to do thorough analysis with the appropriate data, deploying the appropriate tools. As a writer, you are responsible for communicating your findings to the readers. Transport Policy, a leading research publication in transportation planning, offers a checklist for authors interested in publishing with the journal. The checklist is a series of questions authors are expected to consider before submitting their manuscripts to the journal. I believe the checklist is useful for budding data scientists and, therefore, I have reproduced it verbatim for their benefit.

* Have you told readers, at the outset, what they might gain by reading your paper?
* Have you made the aim of your work clear?
* Have you explained the significance of your contribution?
* Have you set your work in the appropriate context by giving sufficient background (including a complete set of relevant references) to your work?
* Have you addressed the question of practicality and usefulness?
* Have you identified future developments that might result from your work?
* Have you structured your paper in a clear and logical fashion?

**Congratulations! You passed!**

**Grade received** 100%

**Latest Submission Grade** 100%

**To pass** 66% or higher

Go to next item

### 1.

Question 1

The results section is where you present:

**1 / 1 point**



The empirical findings.



The conclusion.



The methods used.



R Squared.

**Correct**

Correct.

### 2.

Question 2

The discussion section is where you:

**1 / 1 point**



Introduce the research methods and data sources used for the analysis.



Highlight how your findings provide the ultimate missing piece to the puzzle.

**Correct**

Correct.



Rely on the power of narrative to enable numbers to communicate your important findings to the readers.

**Correct**

Correct.



Refer the reader to the research question and the knowledge gaps you identified earlier.

**Correct**

Incorrect.

### 3.

Question 3

According to the reading, what is an example of housekeeping?

**1 / 1 point**



Adding slide numbers.



Saving the report as a PDF file.



Adding headings to charts.



Adding a list of references.

**Correct**

Correct.

# Lesson Summary

In this lesson, you have learned:

* 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.

Final Assignment

Based on the videos and the reading material, how would you define a data scientist and data science? **(3 marks)**

A data scientist is someone who finds solutions to problems by analyzing Big or small data using appropriate tools and then tells stories to communicate his or her findings to the relevant stakeholders. A data scientist's duties can include developing strategies for analyzing data, preparing data for analyzing and exploring , visualizing data , building models with data using programming languages, such as Python and R, and deploying models into applications.

​Data science is a process of gathering a lot of structured and unstructured data from various sources, cleaning and preparing it, and then analyzing it to gain the insight needed to develop better solutions for businesses.

* **Your IBM Cloud Feature Code:442dc2dec1c243b251549409c03a6565**

**Final Assignment:**

**Question 1:** Based on the videos and the reading material, how would you define a data scientist and data science? **(3 marks)**

**Answer:** Data science incorporates computer science, mathematics, statistics, data analysis, and visualization skills. It relies on statistical methods and algorithms to extract structured and unstructured data insights. People who practice data science are called data scientists.

A Data Scientist’s job is to develop processes for collecting, cleansing, and storing data, mining those data for patterns, and using that to develop and deliver strategic solutions to key problems

**Question 2:** As discussed in the videos and the reading material, data science can be applied to problems across different industries. Give a brief explanation describing what industry you are passionate about and would like to pursue a data science career in? **(2 marks)**

I am curious about business and want to figure out how the best businesses work. I want to work in retail industry. As a data scientist, I will use the data science techniques for gathering information about the retail store performance as well as the customers and their demographics, behaviors, attitudes, and actions. It will help the businesses to reduce costs, improve decision making and explore ways to do more with data such as forecasting demand and footprint to optimise operational efforts. Also, a better understanding of consumer buying patterns to enhance sales.

**Question 3:** Based on the videos and the reading material, what are the **ten** main components of a report that would be delivered at the end of a data science project? **(5 marks)**

**Answer:** The ten main components of a report that would be delivered at the end of a data science project are:

The ten main components of a report that would be delivered at the end of a data science project are:

1. **Cover Page:**The cover page should include the title of the report, the names of authors, their affiliations, and contacts, the name of the institutional publisher (if any), and the date of publication.
2. **Table of content:**The Table of content includes lists of tables with headings and figures which indicates a glimpse of what lies ahead in the document.
3. **Introduction:**This section will contain the complete information about the project, in brief, to understand what is all about the project.
4. **Methodology:**​The methodology section introduces the research methods and data sources we used for the analysis.
5. **Result:** The results section is where we present our empirical findings with tables, illustrative graphics, and plots.
6. **Discussion: In** discussion section is where we familarise our thesis to our readers.
7. **C​onclusion: I**n this section where we highlight the main aspects of this project and how it will solve the real-world problem and also, the future scope of the project.
8. **References:**  The reference list is a list of all the sources used (and cited) in a project. Referencing allows us to acknowledge the contribution of other writers and researchers in our work.
9. **Acknowledgment:** the acknowledgment section acknowledging the support of those who have enabled your work is always good​.
10. **Appendices:** An appendix contains supplementary material that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem and/or is information which is too cumbersome to be included in the body of the paper.