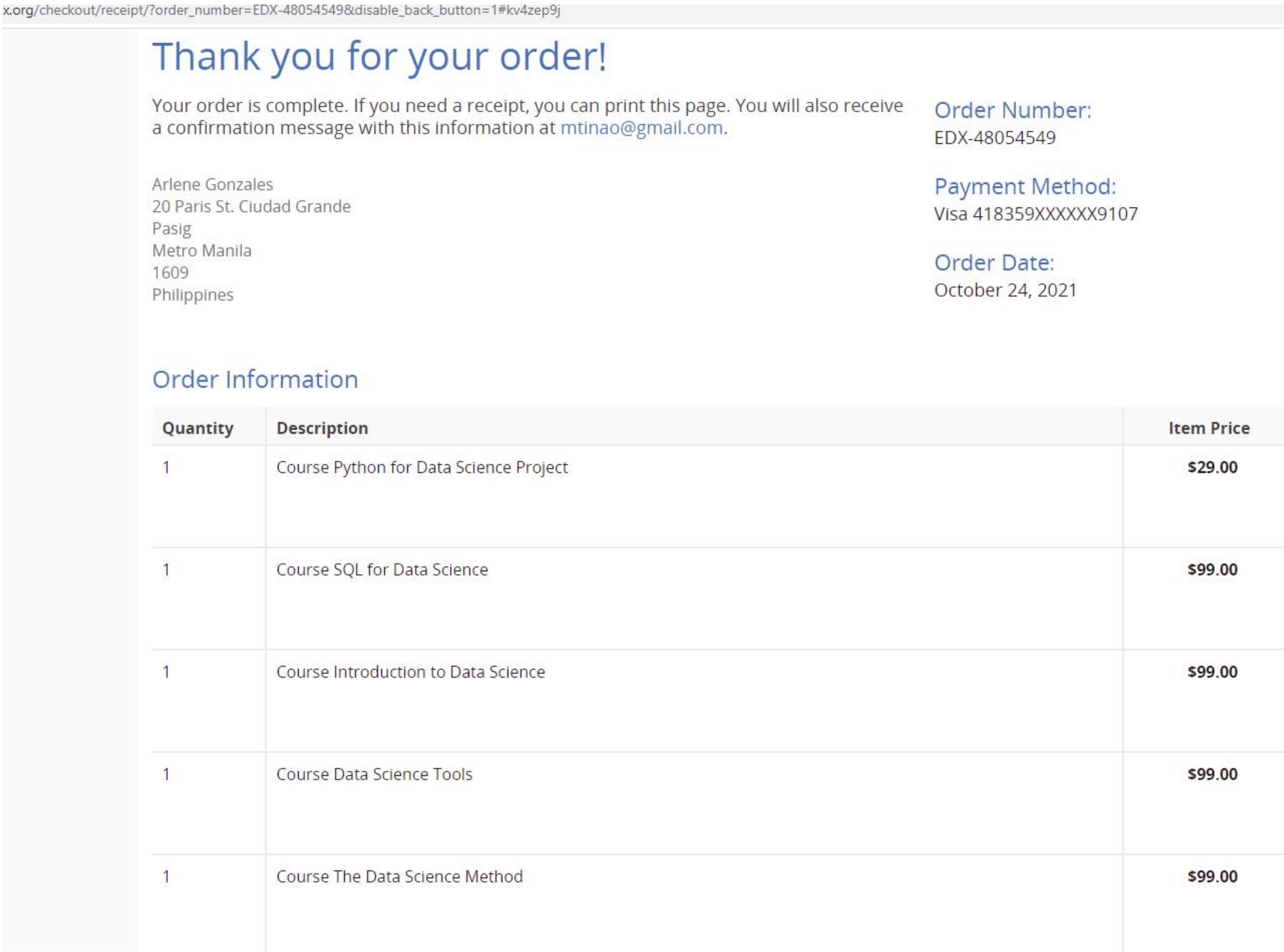


x.org/checkout/receipt/?order_number=EDX-48054549&disable_back_button=1#kv4zep9j

Thank you for your order!

Your order is complete. If you need a receipt, you can print this page. You will also receive a confirmation message with this information at mtinao@gmail.com.

Arlene Gonzales
20 Paris St. Ciudad Grande
Pasig
Metro Manila
1609
Philippines

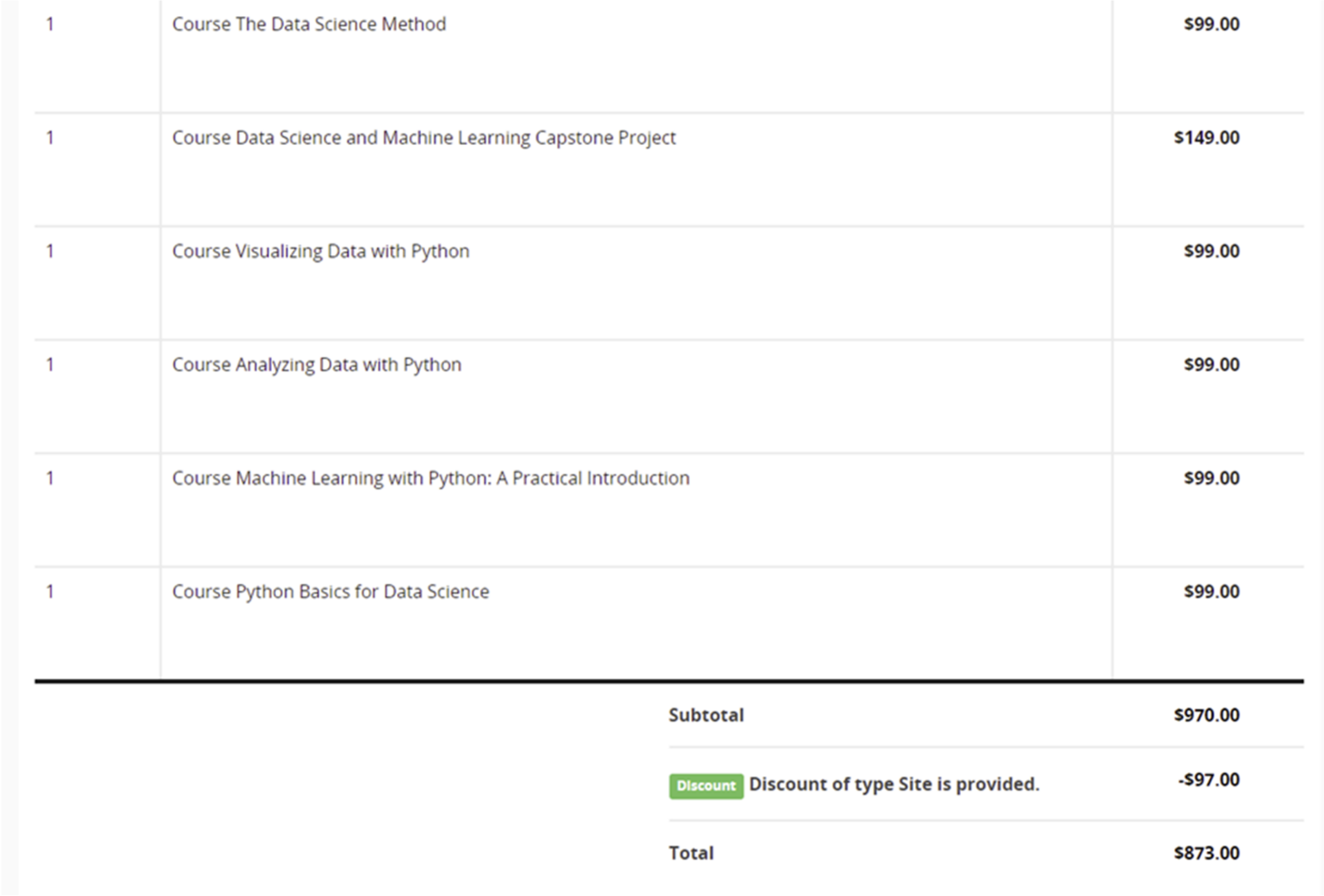
Order Number:
EDX-48054549

Payment Method:
Visa 418359XXXXXX9107

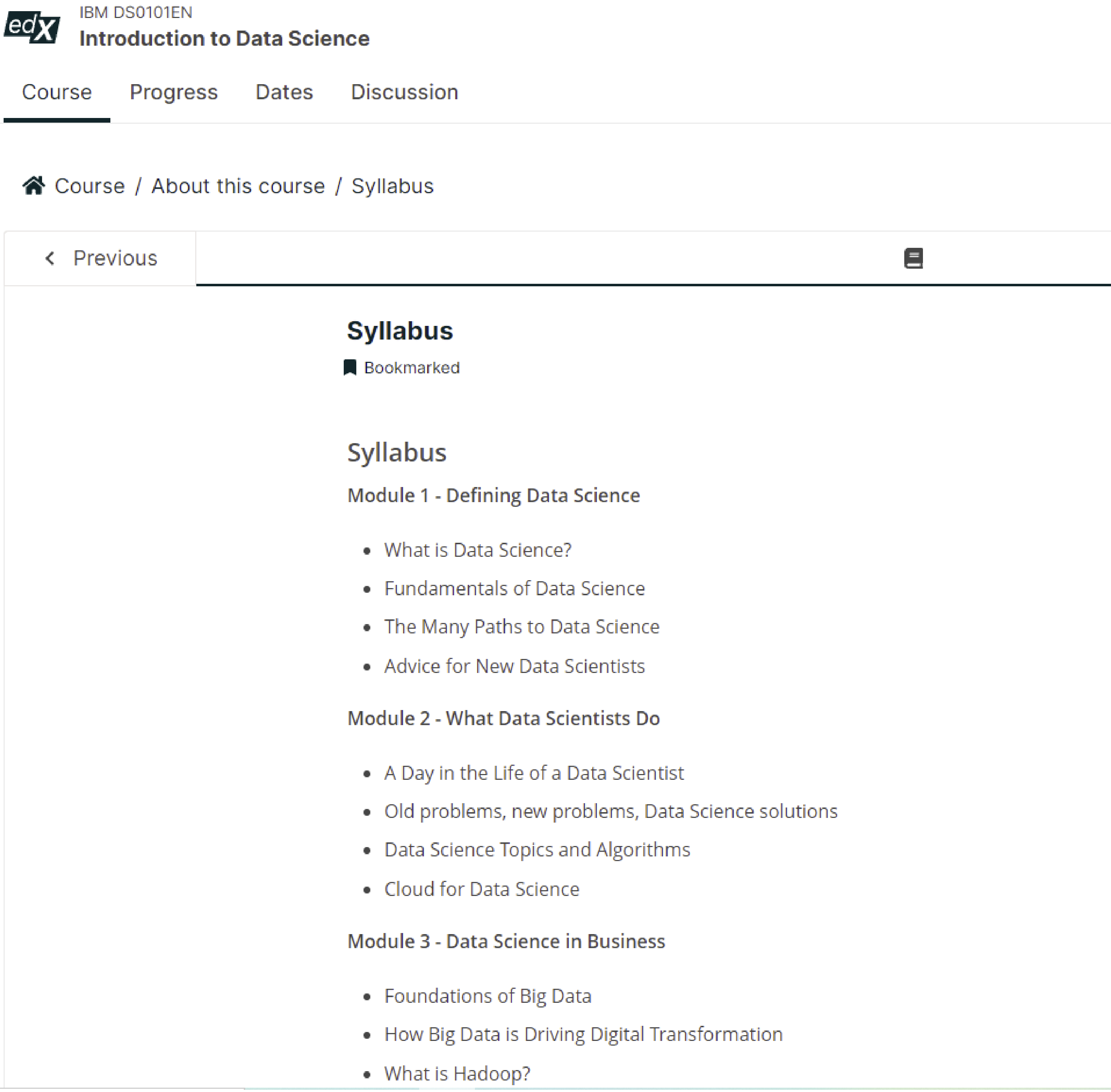
Order Date:
October 24, 2021

Order Information

Quantity	Description	Item Price
1	Course Python for Data Science Project	**$29.00**
1	Course SQL for Data Science	**$99.00**
1	Course Introduction to Data Science	**$99.00**
1	Course Data Science Tools	**$99.00**
1	Course The Data Science Method	**$99.00**



1	Course The Data Science Method	**$99.00**
1	Course Data Science and Machine Learning Capstone Project	**$149.00**
1	Course Visualizing Data with Python	**$99.00**
1	Course Analyzing Data with Python	**$99.00**
1	Course Machine Learning with Python: A Practical Introduction	**$99.00**
1	Course Python Basics for Data Science	**$99.00**
	Subtotal	**$970.00**
	Discount **Discount of type Site is provided.**	**-$97.00**
	Total	**$873.00**



IBM DS0101EN

Introduction to Data Science

Syllabus

Syllabus

Module 1 - Defining Data Science

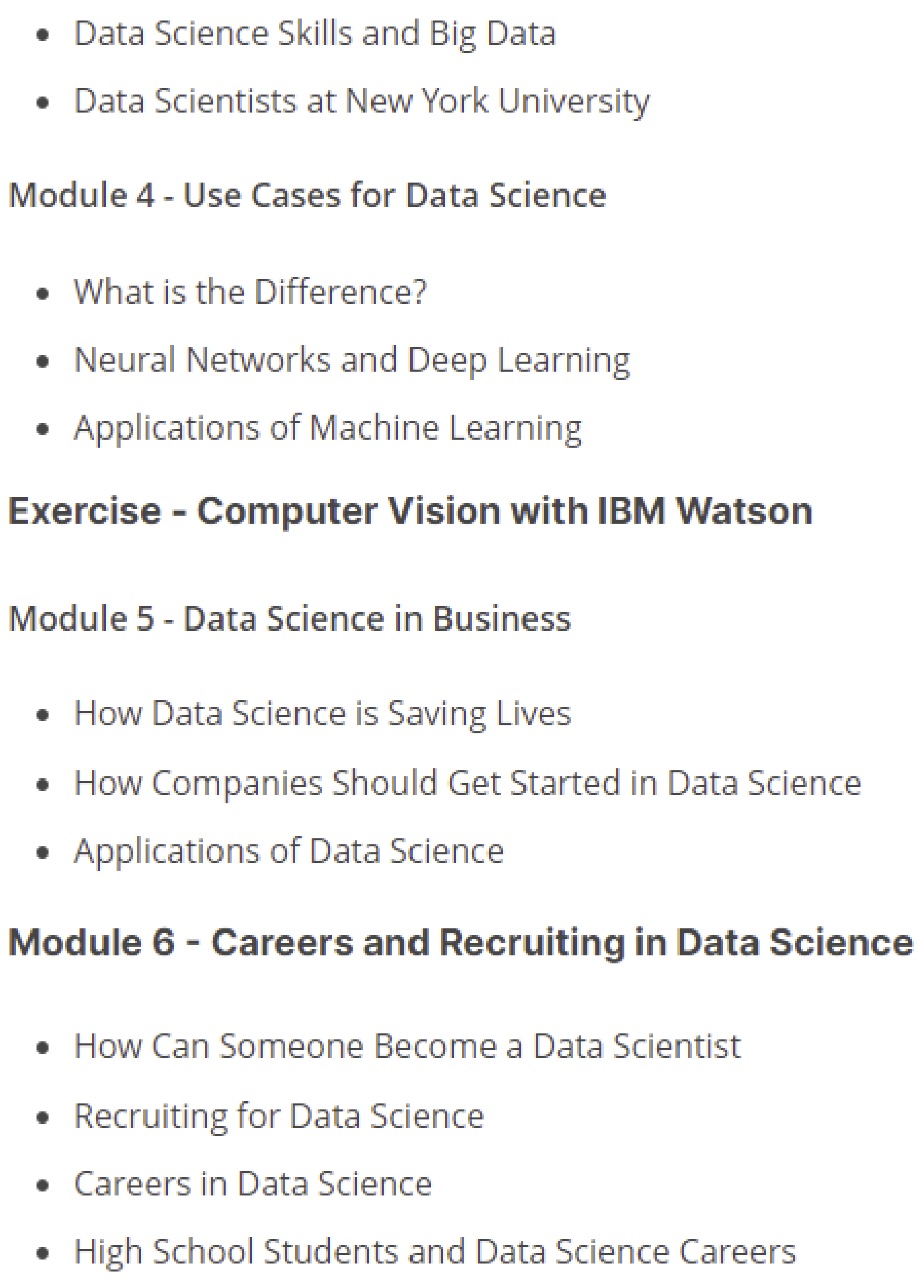
- What is Data Science?
- Fundamentals of Data Science
- The Many Paths to Data Science
- Advice for New Data Scientists

Module 2 - What Data Scientists Do

- A Day in the Life of a Data Scientist
- Old problems, new problems, Data Science solutions
- Data Science Topics and Algorithms
- Cloud for Data Science

Module 3 - Data Science in Business

- Foundations of Big Data
- How Big Data is Driving Digital Transformation
- What is Hadoop?



- Data Science Skills and Big Data
- Data Scientists at New York University

Module 4 - Use Cases for Data Science

- What is the Difference?
- Neural Networks and Deep Learning
- Applications of Machine Learning

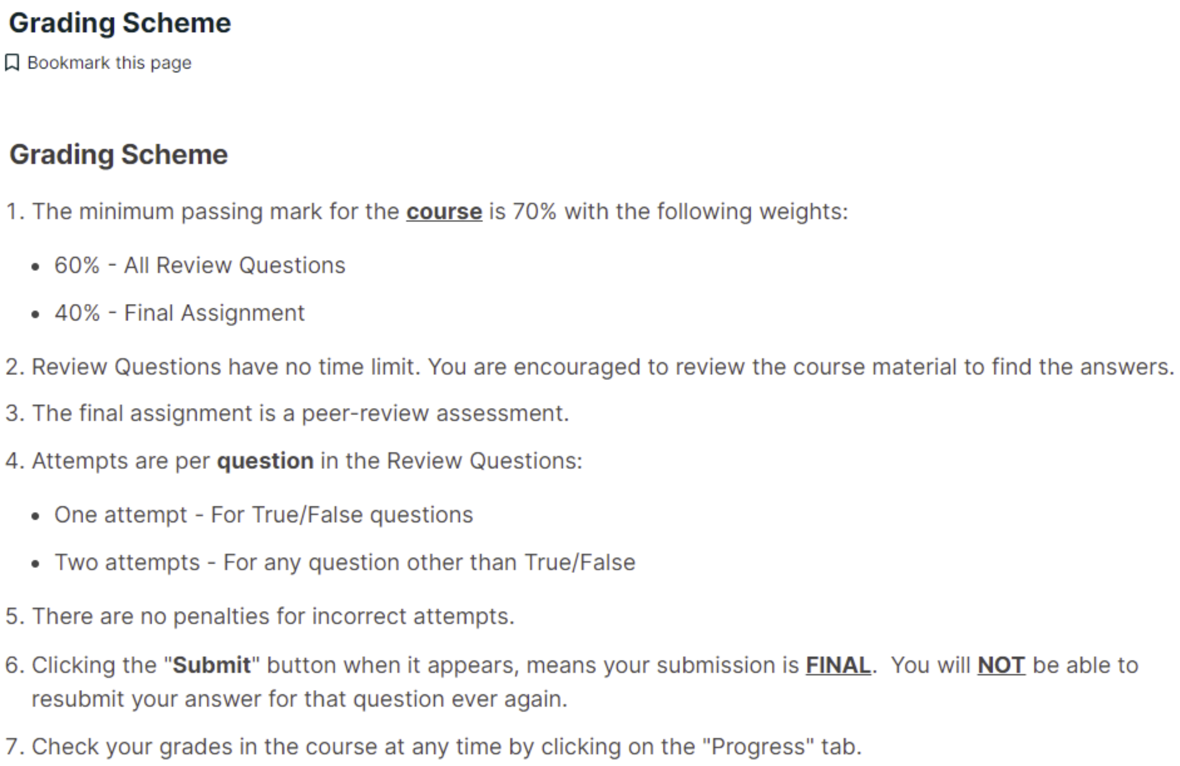
Exercise - Computer Vision with IBM Watson

Module 5 - Data Science in Business

- How Data Science is Saving Lives
- How Companies Should Get Started in Data Science
- Applications of Data Science

Module 6 - Careers and Recruiting in Data Science

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



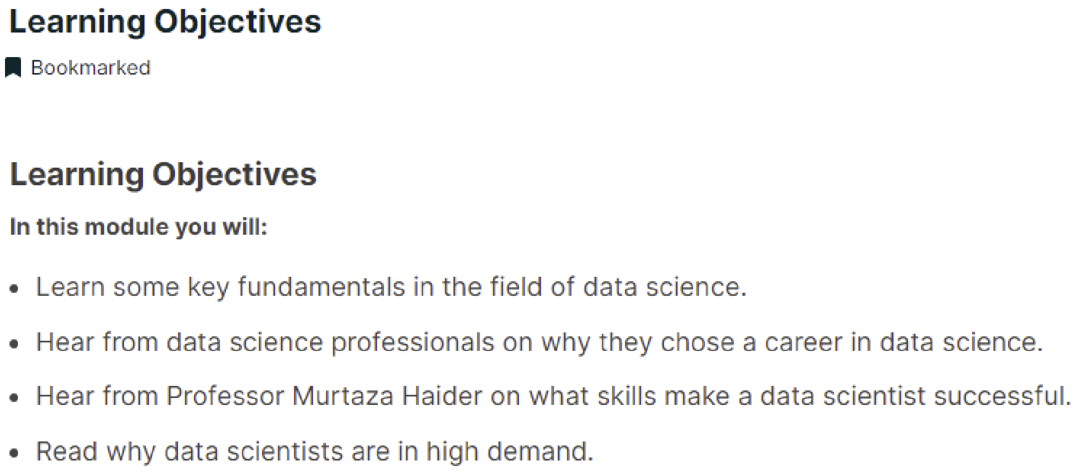
Grading Scheme

Bookmark this page

Grading Scheme

1. The minimum passing mark for the **course** is 70% with the following weights:
   - 60% - All Review Questions
   - 40% - Final Assignment
2. Review Questions have no time limit. You are encouraged to review the course material to find the answers.
3. The final assignment is a peer-review assessment.
4. Attempts are per **question** in the Review Questions:
   - One attempt - For True/False questions
   - Two attempts - For any question other than True/False
5. There are no penalties for incorrect attempts.
6. Clicking the "**Submit**" button when it appears, means your submission is **FINAL**. You will **NOT** be able to resubmit your answer for that question ever again.
7. Check your grades in the course at any time by clicking on the "Progress" tab.

Module 1



Learning Objectives

Bookmarked

Learning Objectives

In this module you will:

- Learn some key fundamentals in the field of data science.
- Hear from data science professionals on why they chose a career in data science.
- Hear from Professor Murtaza Haider on what skills make a data scientist successful.
- Read why data scientists are in high demand.

Module 1 – Defining Data Science

Data Science is a process, not an event. It is the process of using data to understand different things, to understand the world. For me is when you have a model or hypothesis of a problem, and you try to validate that hypothesis or model with your data.
Data science is the art of uncovering the insights and trends that are hiding behind data.
It's when you translate data into a story. So use storytelling to generate insight.
And with these insights, you can make strategic choices for a company or an institution.

Data science is a field about processes and systems to extract data from various forms of whether it is unstructured or structured form. Data science is the study of data.
Like biological sciences is a study of biology, physical sciences, it's the study of physical reactions.

Data is real, data has real properties, and we need to study them if we're going to work on them. Data Science involves data and some science. The definition or the name came up in the 80s and 90s when some professors were looking into the statistics curriculum, and they thought it would be better to call it data science.

But what is Data Science? I'd see data science as one's attempt to work with data,
to find answers to questions that they are exploring. In a nutshell, it's more about data than it is about science. If you have data, and you have curiosity, and you're working with data,
and you're manipulating it, you're exploring it, the very exercise of going through analyzing data, trying to get some answers from it is data science.

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

Fundamentals of Data Science

Everyone you ask will give you a slightly different description of what Data Science is, but most people agree that it has a significant data analysis component. Data analysis isn't new. What is new is the vast quantity of data available from massively varied sources: from log files, email, social media, sales data, patient information files, sports performance data, sensor data, security cameras, and many more besides. At the same time that there is more data available than ever, we have the computing power needed to make a useful analysis and reveal new knowledge.

Data science can help organizations understand their environments, analyze existing issues, and reveal previously hidden opportunities. Data scientists use data analysis to add to the knowledge of the organization by investigating data, exploring the best way to use it to provide value to the business.

So, what is the process of data science? Many organizations will use data science to focus on a specific problem, and so it's essential to clarify the question that the organization wants answered. This first and most crucial step defines how the data science project progresses. Good data scientists are curious people who ask questions to clarify the business need. The next questions are: "what data do we need to solve the problem, and where will that data come from?". Data scientists can analyze structured and unstructured data from many sources, and depending on the nature of the problem, they can choose to analyze the data in different ways. Using multiple models to explore the data reveals patterns and outliers; sometimes, this will confirm what the organization suspects, but sometimes it will be completely new knowledge, leading the organization to a new approach. When the data has revealed its insights, the role of the data scientist becomes that of a storyteller, communicating the results to the project stakeholders. Data scientists can use powerful data visualization tools to help stakeholders understand the nature of the results, and the recommended action to take. Data Science is changing the way we work; it's changing the way we use data and it’s changing the way organisations understand

the world.

The Many Paths to Data Science

Data science didn't really exist when I was growing up. It's not something that I ever woke up And said, I want to be a data scientist when I grow up. No, it didn't exist. I didn't know I would be working in data science. When I grew up, there isn't that field called data science. And I think it's really new.

Data science didn't exist until 2009, 2011. Someone like DJ Patil or Andrew Gelman coined the term. Before that, there was statistics. And I didn't want to be any of those. I wanted to be in business. And then I found data science a heck of a lot more interesting.

I studied statistics, that's how I started. I went through many different stages in my life where I wanted to be a singer and then a doctor. And then I realized that I was good at math. So I chose an area that was focused on quantitative analysis. And from then I do think that I wanted to work with data. Not necessarily data science as it's known today.

The first time that I had contact with data science, when I was my first year as a mechanical engineering. And strategic consulting firms, they use data science to make decisions. So it was my first contact with data science.

I had a complicated problem that I needed to solve, and the usual techniques that we had at the time couldn't help with that problem.

I graduated with a math degree in the worst possible time, right after the economic crisis, and you actually had to be useful to get a job. So I went and got a degree in statistics. And then I worked enough jobs that were called data scientist that I suddenly became one.

My undergraduate degree was in business, and I majored in politics, philosophy, and economics. And then I did a master's in business analytics at New York University at the Stern School of Business. When I left my undergrad, the first company I joined, it turned out that they were analyzing electronic point of sale data for retail manufacturers.

And what we were doing was data science. But we only really started using that term much later. In fact, I'd say four or five years ago is when we started calling it analytics and data science.

I had several options for my internship here in Canada. And one of the options was to work with data science. I used to work with project development. But I think that was a good choice. And then I start my internship with data science.

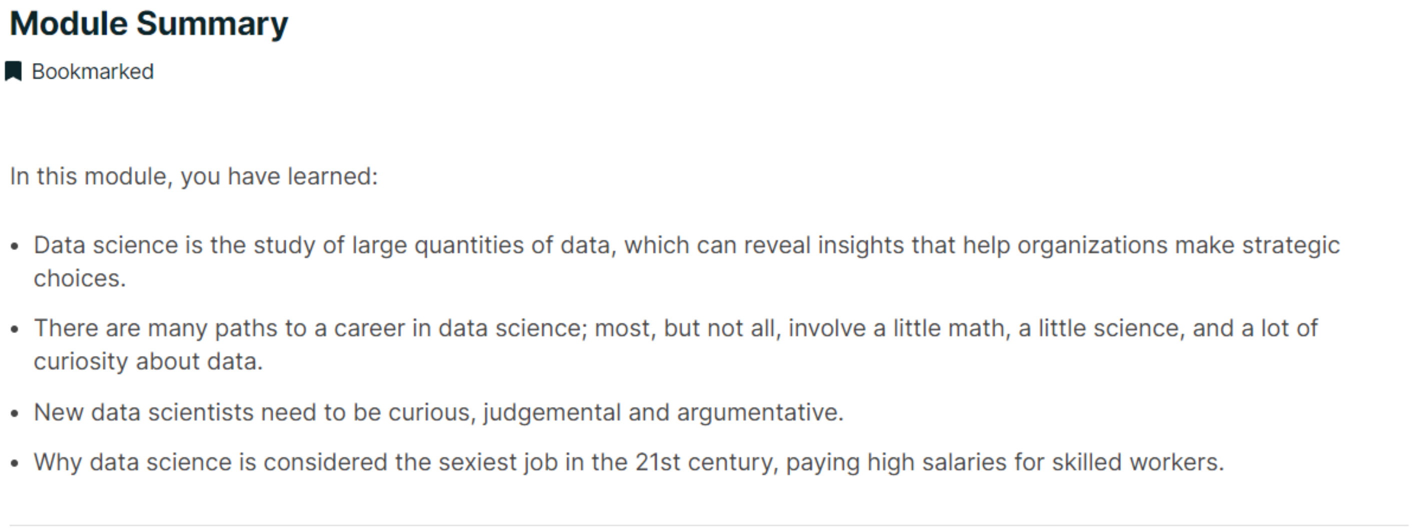
I'm a civil engineer by training, so all engineers work with data. I would say the conventional use of data science in my life started with transportation research. I started building large models trying to forecast traffic on streets, trying to determine congestion and greenhouse gas emissions or tailpipe emissions. So I think that's where my start was. And I started building these models when I was a graduate student at
the University of Toronto. Started working with very large data sets, looking at household samples of, say, 150,000 households from half a million trips. And that, too, I'm speaking from mid 90s when this was supposed to be a very large data set, but not in today's terms.
But that's how I started. I continued working with it.
And then I moved to McGill University where I was a professor of transportation engineering.
And I built even bigger data models that involved data and analytics.

My advice to an aspiring data scientist is to be curious, extremely argumentative and judgmental. Curiosity is absolute must. If you're not curious, you would not know what to do with the data. Judgmental because if you do not have preconceived notions about things you wouldn't know where to begin with. Argumentative because if you can argument and if you can plead a case, at least you can start somewhere and then you learn from data and then you modify your assumptions and hypotheses and your data would help you learn. And you might start at the wrong point. You may say that I thought I believed this, but now with data I know this. So, this allows you a learning process. So, curiosity being able to take a position, strong position, and then moving forward with it. The other thing that the data scientist would need is some comfort and flexibility with analytics platforms: some software, some computing platform, but that's secondary. The most important thing is curiosity and the ability to take positions. Once you have done that, once you've analyzed, then you've got some answers.
And that's the last thing that a data scientist need,
and that is the ability to tell a story.

That once you have your analytics, once you have your tabulations, now you should be able to tell a great story from it. Because if you don't tell a great story from it, your findings will remain hidden, remain buried, nobody would know. But your rise to prominence is pretty much relying on your ability to tell great stories.

A starting point would be to see what is your competitive advantage. Do you want to be a data scientist in any field or a specific field? Because, let's say you want to be a data scientist and work for an IT firm or a web-based or Internet based firm, then you need a different set of skills. And if you want to be a data scientist in the health industry, then you need different sets of skills. So figure out first what you're interested, and what is your competitive advantage.

Your competitive advantage is not necessarily going to be your analytical skills. Your competitive advantage is your understanding of some aspect of life where you exceed beyond others in understanding that. Maybe it's film, maybe it's retail, maybe it's health, maybe it's computers. Once you've figured out where your expertise lies, then you start acquiring analytical skills. What platforms to learn and those platforms, those tools would be specific to the industry that you're interested in. And then once you have got some proficiency in the tools, the next thing would be to apply your skills to real problems, and then tell the rest of the world what you can do with it.



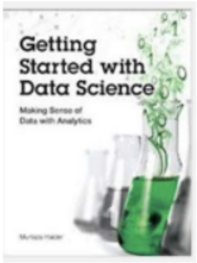
Module Summary

In this module, you have learned:

- Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
- There are many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
- New data scientists need to be curious, judgemental and argumentative.
- Why data science is considered the sexiest job in the 21st century, paying high salaries for skilled workers.

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

Author: Murtaza Haider


Getting Started with Data Science
Making Sense of Data with Analytics
Murtaza Haider

Prescribed Reading: Chapter 1 Pg. 4

Data Science: The Sexiest Job in the 21st Century

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

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

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

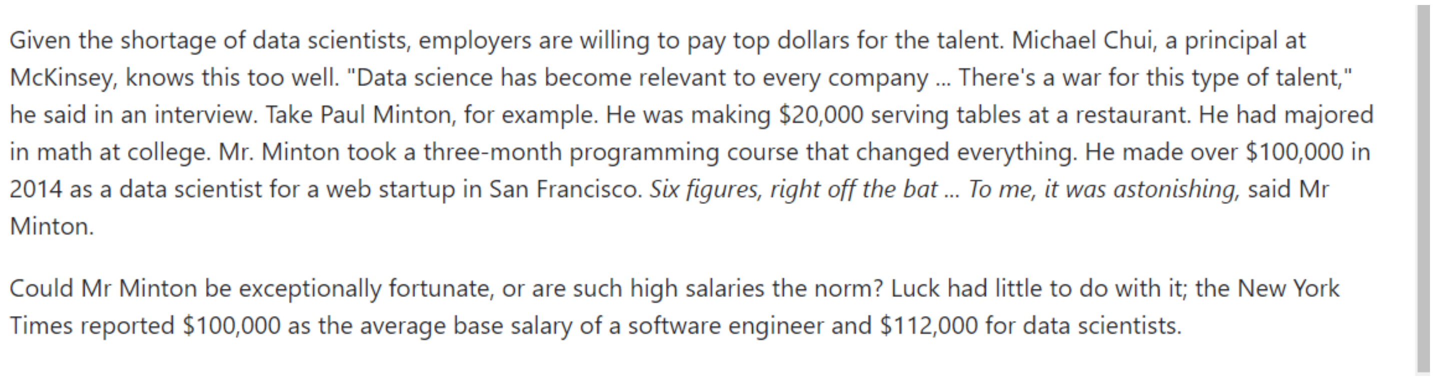
Because the digital revolution has touched every aspect of our lives, the opportunity to benefit from learning about our behaviors is more so now than ever before. Given the right data, marketers can take sneak peeks into our habit formation. Research in neurology and psychology is revealing how habits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

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

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

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

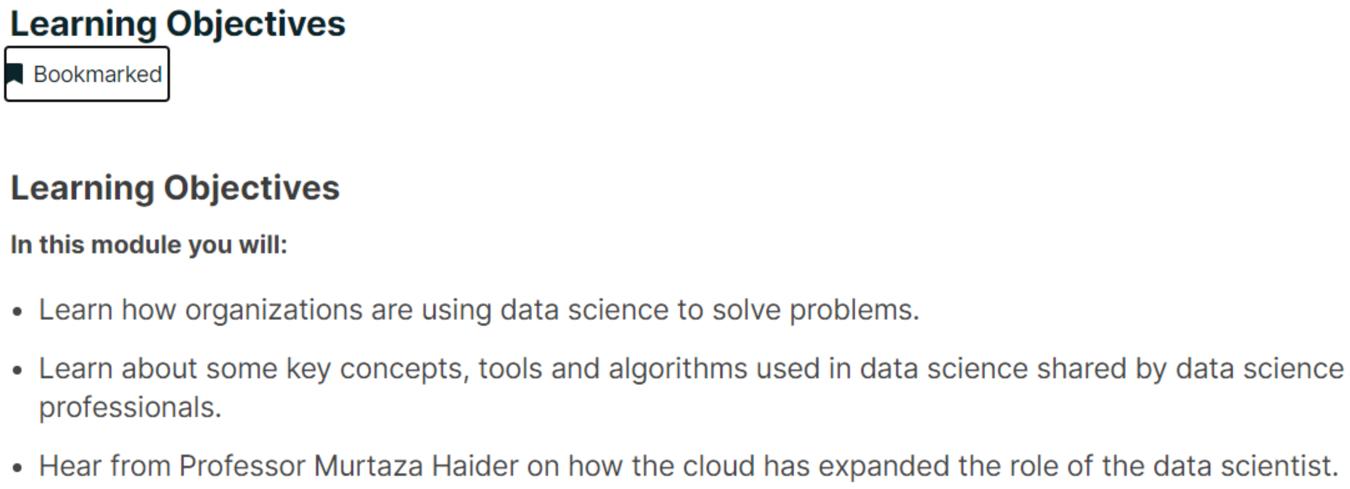
Bernard reported on Walmart, which turned to crowd-sourcing for its analytics need. Walmart approached Kaggle to host a competition for analyzing its proprietary data. The retailer provided sales data from a shortlist of stores and asked the competitors to develop better forecasts of sales based on promotion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



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

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

Module 2 What Data Scientist Do



Learning Objectives

Bookmarked

Learning Objectives

In this module you will:

- Learn how organizations are using data science to solve problems.
- Learn about some key concepts, tools and algorithms used in data science shared by data science professionals.
- Hear from Professor Murtaza Haider on how the cloud has expanded the role of the data scientist.

A Day in the Life of a Data Scientist

I've built a recommendation engine before as part of a large organization and worked through all types of engineers and accounting for different parts of the problem. It's one of the ones I'm most happy with because ultimately, uiI came up with a very simple solution that was easy to understand from all levels,

from the executives to the engineers and developers.

Ultimately, it was just as efficient as something really complex,

and they could have spent a lot more time on.

Back in the university,

we have a problem that we wanted to predict algae blooms.

This algae blooms could cause a rise in

toxicity of the water and it could cause problems through the water treatment company.

We couldn't like predict with our chemical engineering background.

So we use artificial neural networks to predict when these blooms will occur.

So the water treatment companies could better handle this problem.

In Toronto, the public transit is operated by Toronto Transit Commission.

We call them TTC. It's one of

the largest transit authorities in the region, in North America.

And one day they contacted me and said, "We have a problem."

And I said, "Okay, what's the problem?"

They said, "Well, we have complaints data,
and we would like to analyze it, and we need your help."
I said, "Fine I would be very happy to help."
So I said, "How many complaints do you have?"
They said, "A few." I said,
"How many?" Maybe half a million.
I said, "Well, let's start working with it."
So I got the data and I started analyzing it.
So, basically, they have done a great job of keeping
some data in tabular format that was unstructured data.
And in that case, tabular data was when the complaint arrived,
who received it, what was the type of the complaint,
was it resolved, whose fault was it.
And the unstructured part of it was the exchange of e-mails and faxes.
So, imagine looking at
how half a million exchanges of e-mails and trying to get some answers from it.
So I started working with it.
The first thing I wanted to know is why would people complain
and is there a pattern or is there some days when there are more complaints than others?
And I had looked at the data and I analyzed it in all different formats,
and I couldn't find the impetus
for complaints being higher on a certain day and lower on others.
And it continued for maybe a month or so.
And then, one day I was getting off the bus in Toronto,
and I was still thinking about it.
And I stepped out without looking on the ground,
and I stepped into a puddle, puddle of water.
And now, I was sort of ankle deep into water,
and it was just one foot wet and the other dry.
And I was extremely annoyed.
And I was walking back and then it hit me,

and I said, "Well, wait a second.
Today it rained unexpectedly,
and I wasn't prepared for it.
That's why I'm wet, and I wasn't looking forward."
What if there was a relationship between
extreme weather and the type of complaints TTC receives?
So I went to the environment Canada's website,
and I got data on rain and precipitation,
wind and the light.
And there, I found something very interesting.
The 10 most excessive days for complaints.
The 10 days where people complain the most were the days when the weather was bad.
It was unexpected rain,
an extreme drop in temperature,
too much snow, very windy day.
So I went back to the TTC's executives and I said,
"I've got good news and bad news."
And the good news is,
I know why people would complain excessively on certain days.
I know the reason for it. The bad news is,
there's nothing you can do about it.

Old Problems, New Problems, Data Science Solutions (3:56)

Organizations can leverage the almost unlimited amount of data now available to them in a growing number of ways.
However, all organizations ultimately use data science for the same reason—to discover optimum solutions to existing problems.
Let's take a look at three examples of data science providing innovative solutions for old problems.
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.

In a different transport related data science effort, the Toronto Transportation Commission has made great strides in solving an old problem with traffic flows, restructuring those flows in and around the city.

Using data science tools and analysis, they have:

Gathered data to better understand streetcar operations, and identify areas for interventions

Analyzed customer complaints data Used probe data to better understand traffic performance on main routes Created a team to better capitalize on big data for both planning, operations and evaluation

By focusing on peak hour clearances and identifying the most congested routes, monthly hours lost for commuters due to traffic congestion dropped from 4.75 hrs. in 2010 to 3 hrs. in mid-2014.

In facing issues in our environment, data science can also play a proactive role.

Freshwater lakes supply a variety of human and ecological needs, such as providing drinking water and producing food.

But lakes across the world are threatened by increasing incidences of harmful cyanobacterial blooms.

There are many projects and studies to solve this long-existing dilemma.

In the US, a team of scientists from research centers stretching from Maine to South Carolina is developing and deploying high-tech tools to explore cyanobacteria in lakes across the east coast.

The team is using robotic boats, buoys, and camera-equipped drones to measure physical, chemical, and biological data in lakes where cyanobacteria are detected, collecting large volumes of data related to the lakes and the development of the harmful blooms.

The project is also building new algorithmic models to assess the findings.

The information collected will lead to better predictions of when and where cyanobacterial blooms take place, enabling proactive approaches to protect public health in recreational lakes and in those that supply drinking water.

Such interdisciplinary training prepares the next generation of scientists to address societal issues with the proper modernized data science tools.

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.

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 (3:53)

Start of transcript. Skip to the end.

I really enjoy regression.

I'd say regression was maybe one of the first concepts that I, that really helped me understand data so I enjoy regression.

I really like data visualization.

I think it's a key element for people to get across their message to people that don't understand that well what data science is.

Artificial neural networks.

I'm really passionate about neural networks because we have a lot to learn with nature so when we are trying to mimic our, our brain I think that we can do some applications with this behavior with this biological behavior in algorithms.

Data visualization with R. I love to do this.

Nearest neighbor.

It's the simplest but it just gets the best results so many more times than some overblown, overworked algorithm that's just as likely to overfit as it is to make a good fit.

So structured data is more like tabular data things that you’re familiar with in Microsoft Excel format.

You've got rows and columns and that's called structured data.

Unstructured data is basically data that is coming from mostly from web where it's not tabular.

It is not, it's not in rows and columns.

It's text.

It's sometimes it's video and audio, so you would have to deploy more sophisticated algorithms to extract data.

And in fact, a lot of times we take unstructured data and spend a great deal of time and effort to get some structure out of it and then analyze it.

So if you have something which fits nicely into tables and columns and rows, go head.

That's your structured data.

But if you see if it's a weblog or if you're trying to get information out of webpages and you've got a gazillion web pages, that's unstructured data that would require a little bit more effort to get information out of it.

There are thousands of books written on regression and millions of lectures delivered on regression.

And I always feel that they don’t do a good job of explaining regression because they get into data and models and statistical distributions.

Let's forget about it.

Let me explain regression in the simplest possible terms.

If you have ever taken a cab ride, a taxi ride, you understand regression.

Here is how it works.

The moment you sit in a cab ride, in a cab, you see that there's a fixed amount there.

It says $2.50.

You, rather the cab, moves or you get off.

This is what you owe to the driver the moment you step into a cab.

That's a constant.

You have to pay that amount if you have stepped into a cab.

Then as it starts moving for every meter or hundred meters the fare increases by certain amount.

So there's a... there's a fraction, there's a relationship between distance and the amount you would pay above and beyond that constant.

And if you're not moving and you're stuck in traffic, then every additional minute you have to pay more.

So as the minutes increase, your fare increases.

As the distance increases, your fare increases.

And while all this is happening you've already paid a base fare which is the constant.

This is what regression is.

Regression tells you what the base fare is and what is the relationship between time and the fare you have paid, and the distance you have traveled and the fare you've paid.

Because in the absence of knowing those relationships, and just knowing how much people traveled

for and how much they paid, regression allows you to compute that constant that you didn't know.

That it's $2.50, and it would compute the relationship between the fare and and the distance and

the fare and the time.

That is regression.

Cloud for Data Science (3:22)

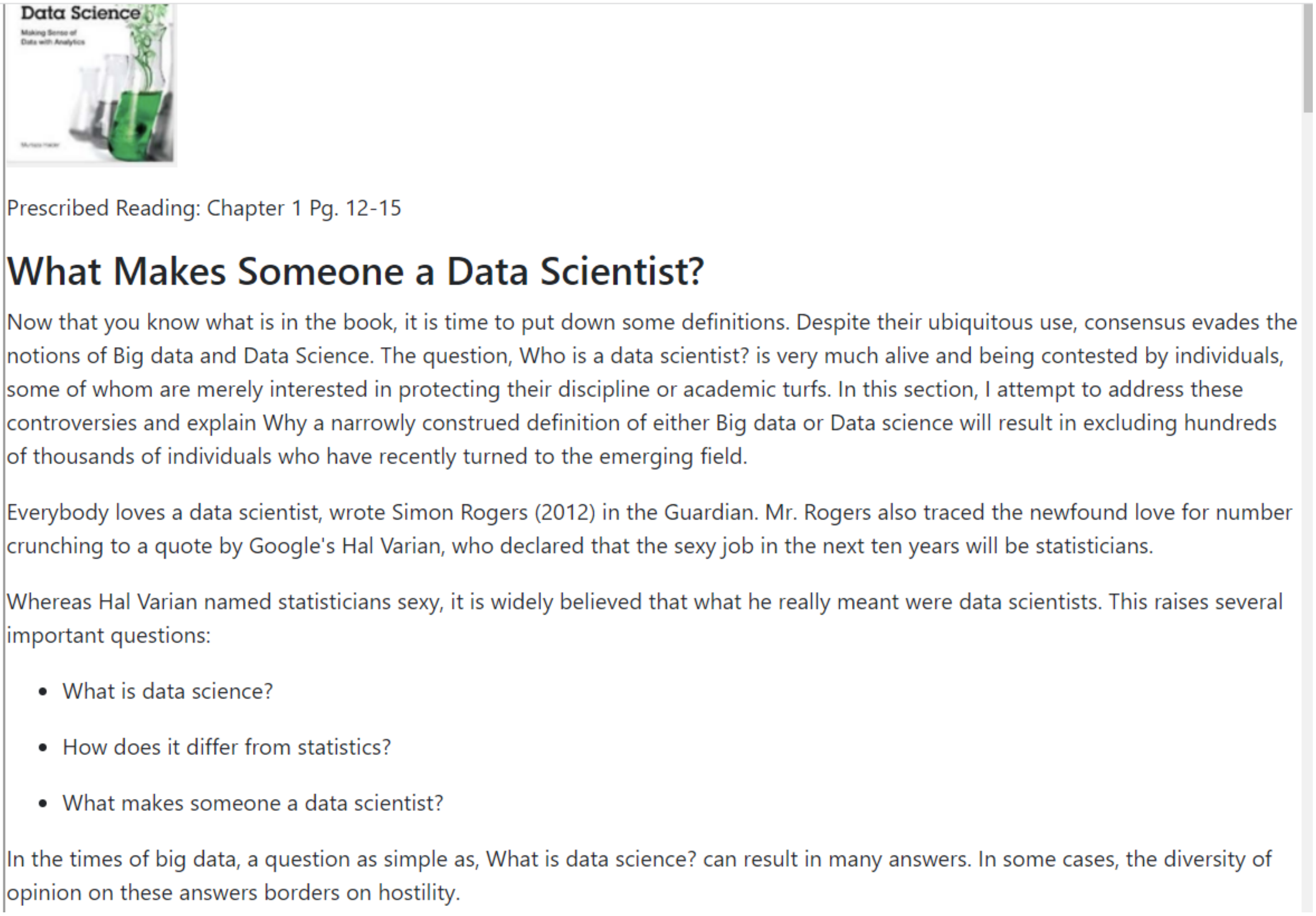
Cloud is is is-- it's a godsend for data scientists, primarily because you're able to take the, or you take your data, take your information and put it in the cloud, put it in the central storage system. It allows you to bypass the physical limitations of the computers and the systems you're using and 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. And 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. So, and 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 and the algorithms that you need to use, you don't have them with you but then on the cloud you have those, those algorithms available. So what you do is you deploy those algorithms on very large data sets. 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. And the other thing that cloud is beautiful for is that it allows multiple entities to work with same data at the same time. So you can be working with the same data that your colleagues in say, Germany, and another team in India, and another team in in Ghana. They are collectively working and they are 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. So 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.
Register to 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.

Cloud for Data Science

Cloud is is is-- it's a godsend for data scientists, primarily because you're able to take the, or you take your data, take your information and put it in the cloud, put it in the 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. And 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. So, and 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 and the algorithms that you need to use, you don't have them with you but then on the cloud you have those, those algorithms available. So what you do is you deploy those algorithms on very large data sets. 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. And the other thing that cloud is beautiful for is that it allows multiple entities to work with same data at the same time. So you can be working with the same data that your colleagues in say, Germany, and another team in India, and another team in in Ghana. They are collectively working and they are 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. So 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. Register to 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.



Prescribed Reading: Chapter 1 Pg. 12-15

What Makes Someone a Data Scientist?

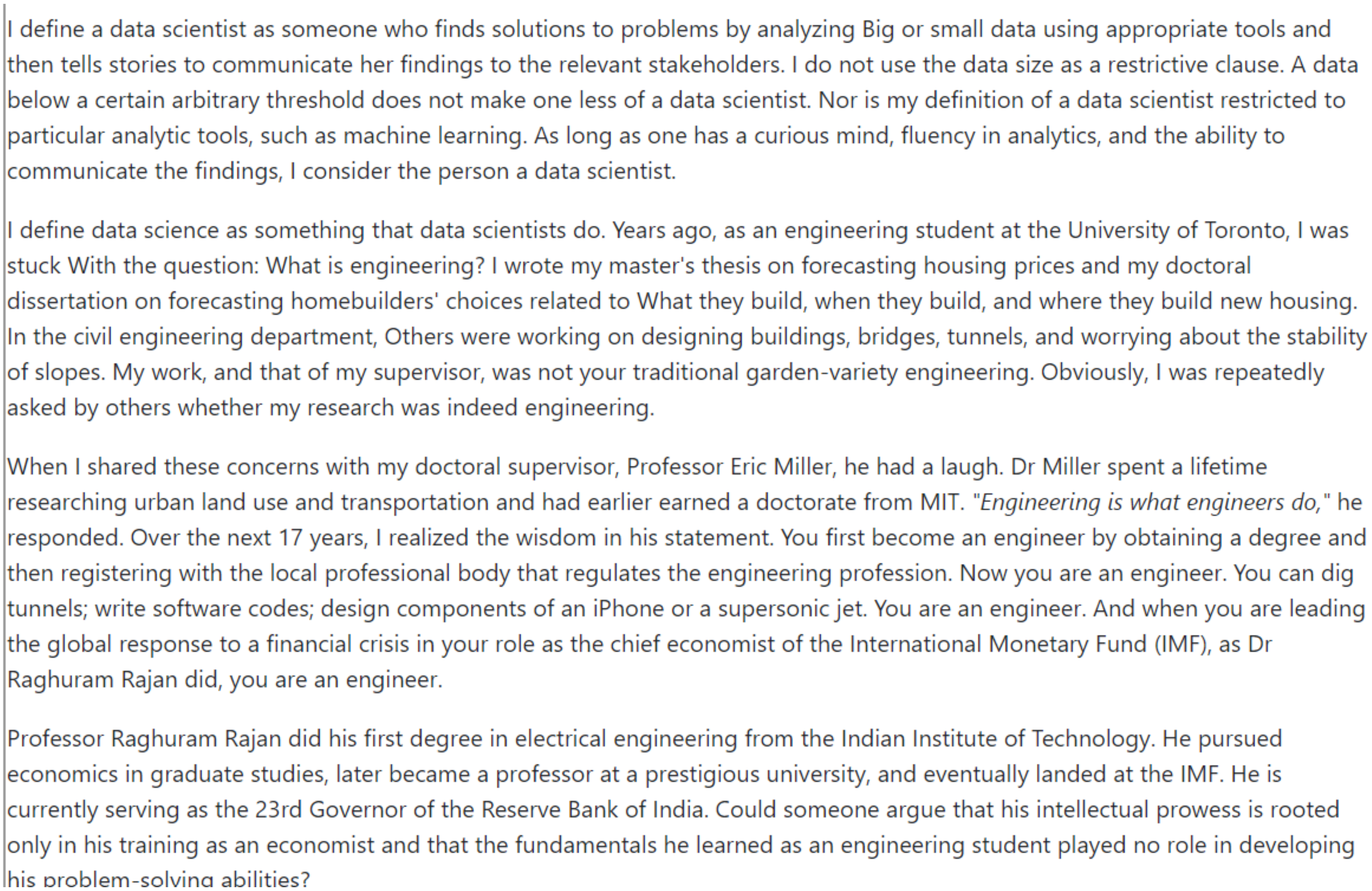
Now that you know what is in the book, it is time to put down some definitions. Despite their ubiquitous use, consensus evades the notions of Big data and Data Science. The question, Who is a data scientist? is very much alive and being contested by individuals, some of whom are merely interested in 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.

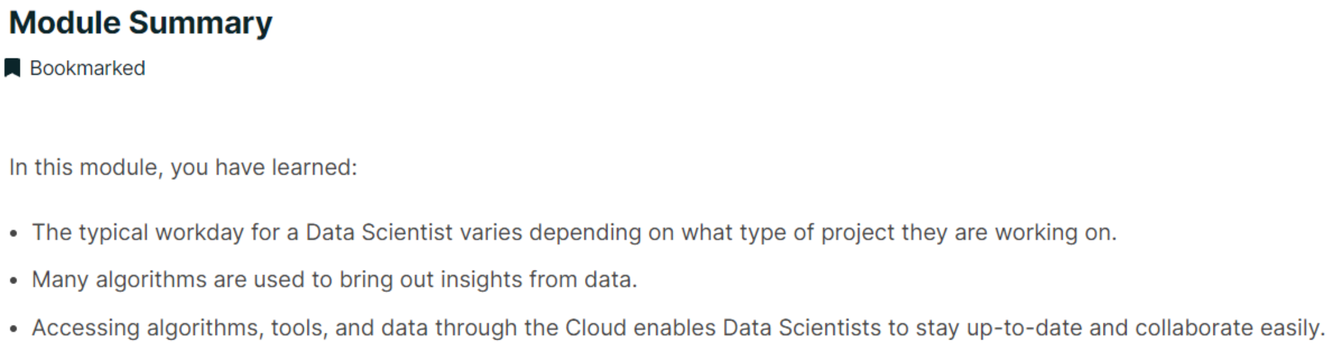


I define a data scientist as someone who finds solutions to problems by analyzing Big or small data using appropriate tools and then tells stories to communicate her findings to the relevant stakeholders. I do not use the data size as a restrictive clause. A data below a certain arbitrary threshold does 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.

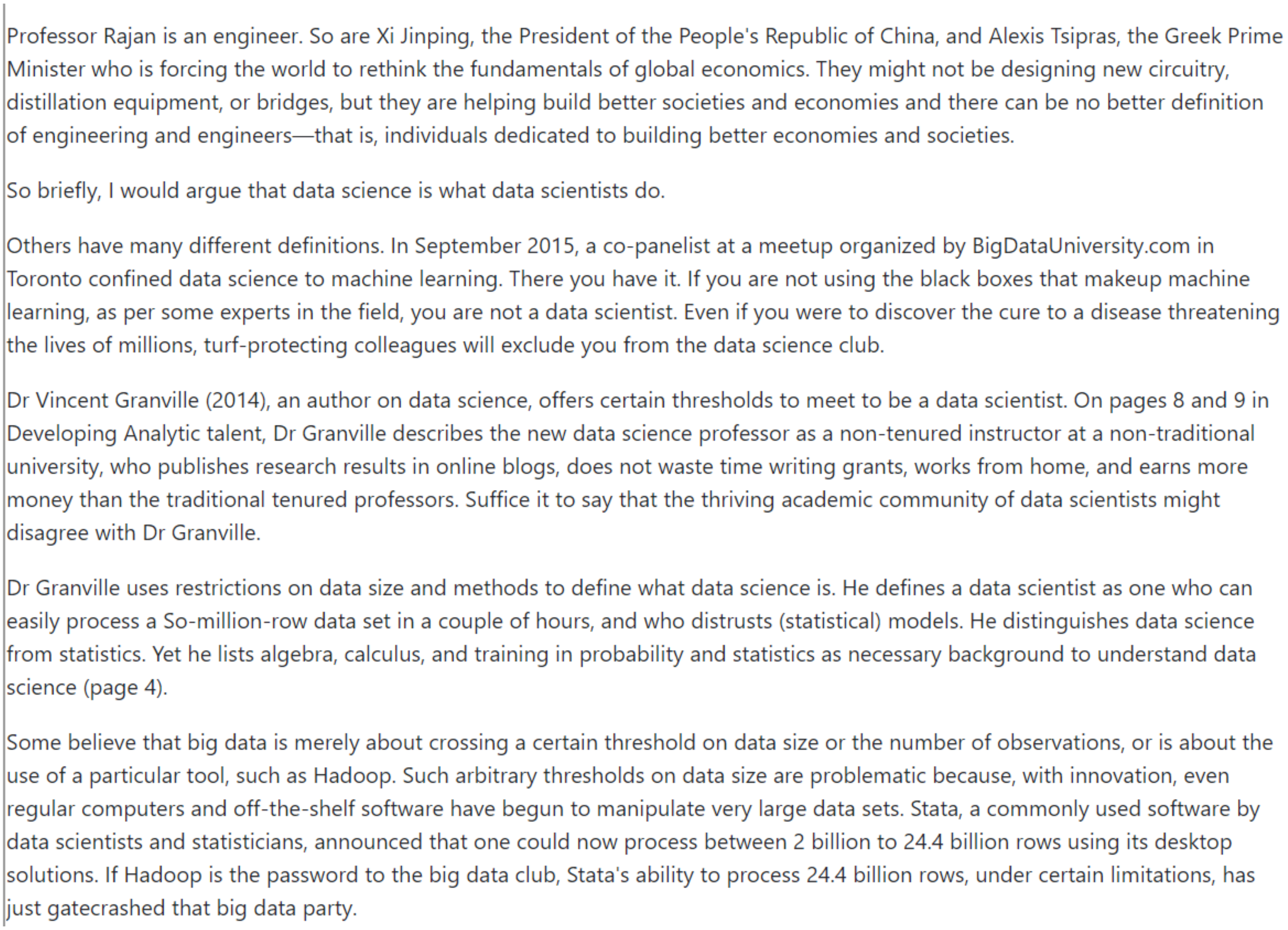
Professor Raghuram Rajan did his first degree in electrical engineering from the Indian Institute of Technology. He pursued economics in graduate studies, later became a professor at a prestigious university, and eventually landed at the IMF. He is currently serving as the 23rd Governor of the Reserve 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?



Module Summary

In this module, you have learned:

- The typical workday for a Data Scientist varies depending on what type of project they are working on.
- Many algorithms are used to bring out insights from data.
- Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.



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.

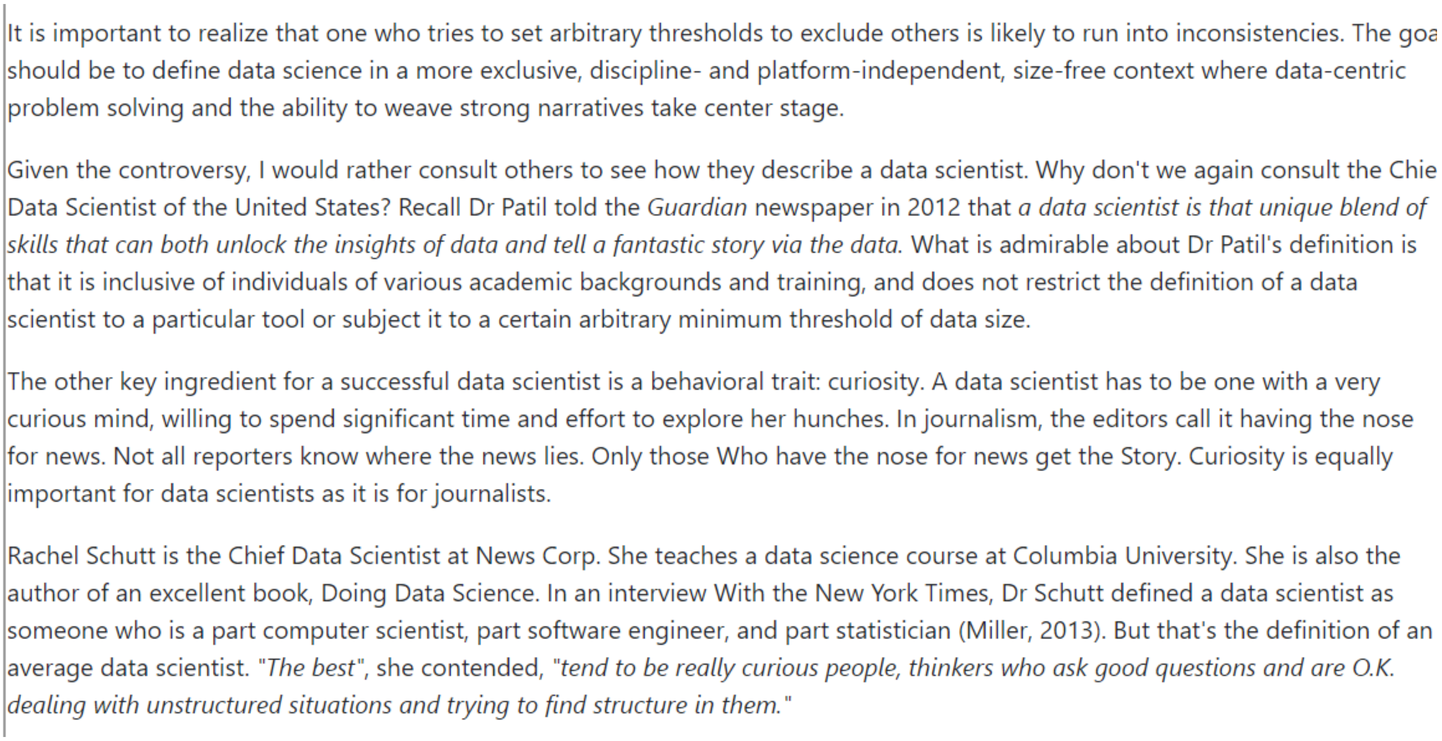
So briefly, I would argue that data science is what data scientists do.

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.



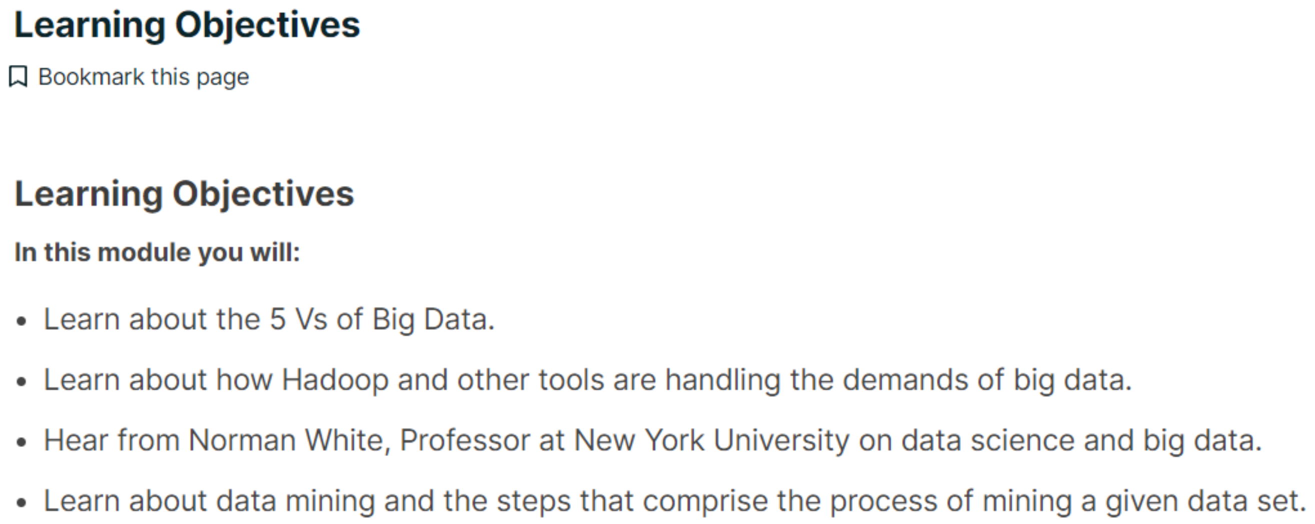
It is important to realize that one who tries to set arbitrary thresholds to exclude others is likely to run into inconsistencies. The goa should be to define data science in a more exclusive, discipline- and platform-independent, size-free context where data-centric problem solving and the ability to 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 Chie 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.*"

Module 3



Learning Objectives

Learning Objectives

In this module you will:

- Learn about the 5 Vs of Big Data.
- Learn about how Hadoop and other tools are handling the demands of big data.
- Hear from Norman White, Professor at New York University on data science and big data.
- Learn about data mining and the steps that comprise the process of mining a given data set.

Foundations of Big Data

In this digital world, everyone leaves a trace.

From our travel habits to our workouts and entertainment, the increasing number of internet connected devices that we interact with on a daily basis record vast amounts of data about us.

There’s even a name for it: Big Data.

Ernst and Young offers the following definition: “Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines.

It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”

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

These are the V's of Big Data.

Velocity is the speed at which data accumulates.

Data is being generated extremely fast, in a process that never stops.

Near or real-time streaming, local, and cloud-based technologies can process information very quickly.

Volume is the scale of the data, or the increase in the amount of data stored.

Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.

Variety is the diversity of the data.

Structured data fits neatly into rows and columns, in relational databases while unstructured data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers, and video.

Variety also reflects that data comes from different sources, machines, people, and processes, both internal and external to organizations.

Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more.

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

Attributes include consistency, completeness, integrity, and ambiguity.

Drivers include cost and the need for traceability.

With the large amount of data available, the debate rages on about the accuracy of data in the digital era.

Is the information real, or is it false?

Value is our ability and need to turn data into value.

Value isn't just profit.

It may have medical or social benefits, as well as customer, employee, or personal satisfaction.

The main reason that people invest time to understand Big Data is to derive value from it.

Let's look at some examples of the V's in action.

Velocity: 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: 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.

That's the equivalent of 10 million Blu-ray DVD's.

Variety: 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: 80% of data is considered to be unstructured and we must devise ways to produce reliable and accurate 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, alternative 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.

So next time you strap on your smartwatch, unlock your smartphone, or track your workout, remember your data is starting a journey that might take it all the way around the world, through big data analysis, and back to you.

How Big Data is Driving digital Transformation

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.

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 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.
Fairly simple concepts
but turned out that you could do
lots and lots of different kinds of
handle lots and lots of different kinds of problems
and very, very, very large data sets.
So the one thing that's nice about these big data clusters
is they scale linearly.
You had twice as many servers
and you get twice the performance
and you can handle twice the amount of data.
So this was just broke a bottleneck
for all the major social media companies.
Yahoo then got on board.
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.
And if you google Hadoop you'll see that
it's now a very popular term
and there are many, many, many
if you look at the big data ecology
there are hundreds of thousands of companies out there

that have some kind of footprint
in the big data world.
(music)
Most of the components of data science have been around for
many, many, many decades.
But they're all coming together now
with some new nuances I guess.
At the bottom of data science
you see probability and statistics.
You see algebra, linear algebra
you see programming
and you see databases.
They've all been here.
But what's happened now is we
now have the computational capabilities
to apply some new techniques - machine learning.
Where now we can take really large data sets
and instead of taking a sample
and trying to test some hypothesis
we can take really, really large data sets
and look for patterns.
And so back off one level from hypothesis testing
to finding patterns that maybe will generate hypotheses.
Now this can bother some very traditional statisticians
and gets them really annoyed sometimes
that you know you're supposed to have a hypothesis
that is not that is independent of the data
and then you test it.
So once some of these machine learning techniques started
were really the only thing
the only way you can analyze

some of these really large
social media data sets.
So what we've seen is that the combination
of traditional areas computer science
probability, statistics, mathematics
all coming together in this thing that we call
Decision Sciences.
Our department at Stern
I'll give a little plug here
we happen to have been very well situated
among business schools
because we're one of the few business schools
that has a real statistics department
with real PhD level statisticians in it.
We have an operations management department
and an information systems department.
So we have a wide range of computer scientists
to statisticians, to operations researchers.
And so we were perfectly positioned
as a couple of other business schools were
to jump on this bandwagon and say; okay
this is Decision Sciences.
And Foster Provost who's in my department was
the first director of the NYU Center for Data Science.
(music)
Four years ago maybe five years ago.
I mean, I feel this is one of those cases
where you can just to Google
and search for
data science and see how often it occurred
and you'll see almost nothing

and then just a spike.

The same thing you would see with big data

about seven or eight years ago.

So data science is a term I haven't heard of

probably five years ago.

(music)

The first question is what is it?

And I think

faculty and everybody is still trying to

get their hands around exactly what is

business analytics and what is data science.

We certainly know

the components of it.

But it's morphing and changing and growing.

I mean the last three years

deep learning has just been added into the mix.

Neural networks have been around for 20 or 30 years.

20 years ago I would teach neural networks in a class

and you really couldn't do very much with them.

And now some researchers have come up with

multi-layer neural networks

in Toronto in particular the University of Toronto.

And that technology is now rapidly expanding

it's being used by Google, by Facebook, by lots of companies.

Data Science Skills and Big Data

I'm Norman White, I'm a Clinical Faculty Member

in the IOMS Department,

Information, Operations and Management Science

Department here at Stern.

I've been here for a long time (laughs),
since I got out of college, pretty much.
I'm sort of a techy, geeky kind of person.
I really like to play with technology in my spare time.
I'm currently Faculty Director
of the Stern Center for Research Computing,
in which we have a private cloud
that runs lots of different kinds of systems.
Many of our faculty or PhD students who need
specialized hardware and software will come to us,
we'll spin up a machine for them, configure it,
I'll help them and advise on them.
A lot of the data scientists, or virtually all
the data scientists at Stern use our facilities.
And their PhD students use them a lot.
(music)
I have an undergraduate degree in Applied Physics
and while I was an undergrad I took a number
of economics courses, so I ended up deciding
to go to business school, but I had,
this was in the early days of computers (laughs)
and I had gotten interested in computers.
I came to Stern, which was then NYU Business School downtown
and they had a little computer center,
and I decided that I was gonna learn
two things while I was there.
One, I was gonna learn how to program.
I had taken one programming course in college.
And I was gonna learn how to touch type.

I never did learn how to touch type (laughs).
Or maybe I did but I've forgotten now,
and back to two finger pecking.
But I became a self taught programmer,
and then I took a number of courses at IBM
because I eventually came the director
of the computer center while I was getting my PhD
in Economics and Statistics at Stern.
In 1973, the school formed a department called
Computer Applications and Information Systems
and I was one of the first faculty members
in the department and I've been here ever since (laughs).
(music)
My typical Monday is, I usually get in around 11 o'clock
and I do my email at home first,
but I come in and I have two classes on Monday.
I have a class on design and development
of web based systems at six o'clock.
Two o'clock, I have a dealing with data class.
The class is based on Python notebooks,
so we start with the basics of Unix and Linux,
just to get the students used to that.
We move onto some Python, some regular expressions,
a lot of relational databases, some Python Pandas,
which is sort of like R for Python, lets you do
mathematical and statistical calculations in Python.
And then I end up with big data,
for which, as you probably know, I'm an evangelist.
The students I have, weekly homeworks.

I put them in teams and they have to do a big project

at the end of the term, and they do some really cool things.

(music)

Yes, in fact, the whole course

is taught using Jupyter notebooks.

Every student has their own virtual machine

on Amazon Web Services, so we pre configure all the machines

and they get a standard image that has all of the materials

for the course either loaded on it or in a Jupyter notebook,

there are the commands to download it

or update the server with the right software.

So everybody is in the same environment,

it doesn't matter what kind of,

whether they have a Mac or a Windows machine

or how old it is, everybody can do everything in the class.

Data Scientists at NYU

Everybody knows how to program,

at least a little bit.

They all have a little bit of programming background

at least, and some of them have a lot.

Some of them are Masters of Science and Computer Science,

some of them are MBA students who've come in

from technical fields and programmed every day.

And others are ones who maybe took

a programming course in college four or five years ago

but at least they can think computationally,

which I think is the most important thing that they need.

(music)

Data science and business analytics have become
very hot subjects in the last four or five years.
We have new tools, we have new approaches,
and we have lots and lots of data that traditional
techniques just couldn't really store and handle.
I think the word is out.
I think at this point, at first,
companies and employers understood the need,
especially in certain fields.
I can remember talking to a major bank three years ago
about big data and there was one little group in the bank
where one person had a little effort
in putting a little cluster together.
Now that same bank has five or six major big data clusters
and they're putting all of their credit card data in it
and they're grinding it upside down and sideways,
using all sorts of data science kinds of techniques.
Two years ago, or was it last year, I think,
our undergraduate dealing with data course
had 28 students in it.
This year it has 140.
So that means that the parents
are now beginning to get the word,
because one thing we understand with our undergrads
is the parents who are paying very hefty tuitions,
they, you know, they tell their sons and daughters,
"You know, you should be an accountant," right?
Or, "You should go into financial services,
"or into marketing, 'cause this is where the money is."

Now, they're getting the word that

maybe you should take some more STEM classes in high school

and be ready to go into data science

or go into fields where analytics

has become more and more important.

(music)

It depends on who you are (laughs).

I have my own definition of big data.

My definition of big data is data that is large enough

and has enough volume and velocity

that you cannot handle it with traditional database systems.

Some of our statisticians think big data

is something you can't fit on a thumb drive.

Big data, to me, was started by Google.

When Google tried to figure out how they were,

when Larry Page and Sergey Brin wanted to, basically,

figure out how to solve their page rank algorithm,

there was nothing out there.

They were trying to store all of the web pages in the world,

and there was no technology, there was no way to do this,

and so they went out and developed this approach,

which has now become, Hadoop has copied it,

but this is where all these large,

big data clusters are found.

But big data has now also expanded into,

how do you analyze?

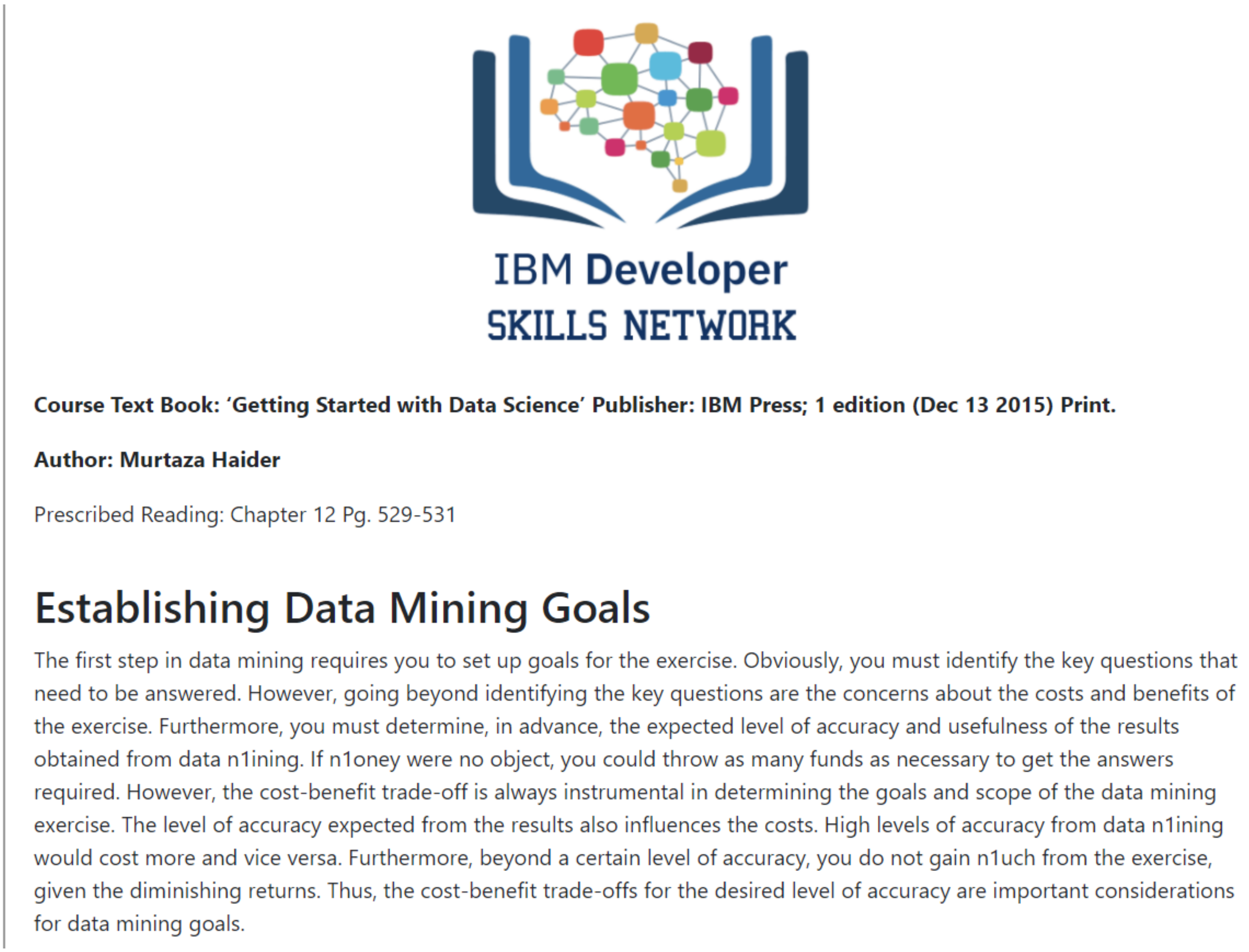
There are new analytical techniques

and statistical techniques for handling these

really, really, really large data sets.

We'll probably get to deep learning

at some point along here.



IBM Developer
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 n1ining. If n1oney 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 n1ining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain n1uch 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

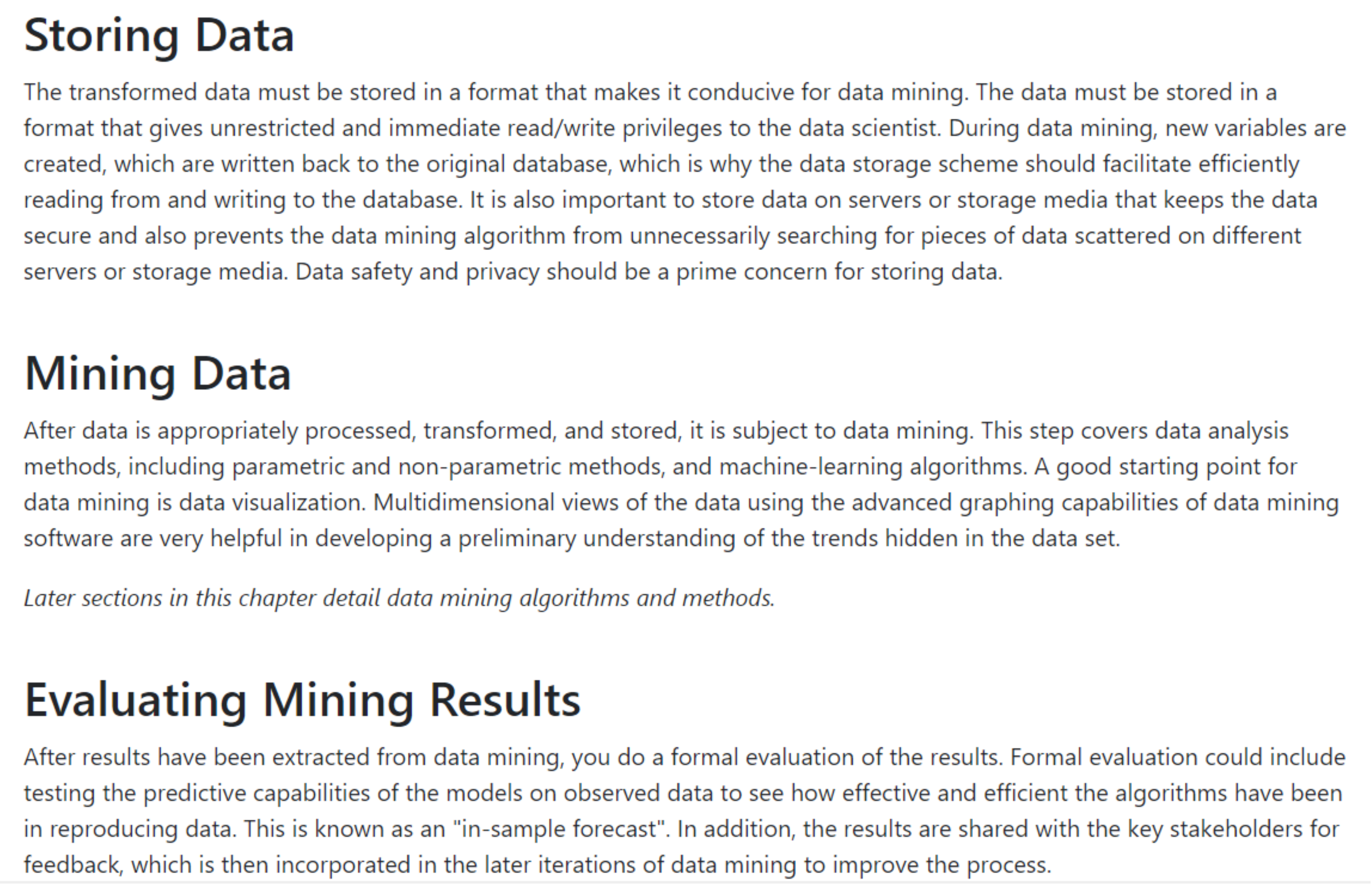
Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you identify the irrelevant attributes of data and expunge such attributes from 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.

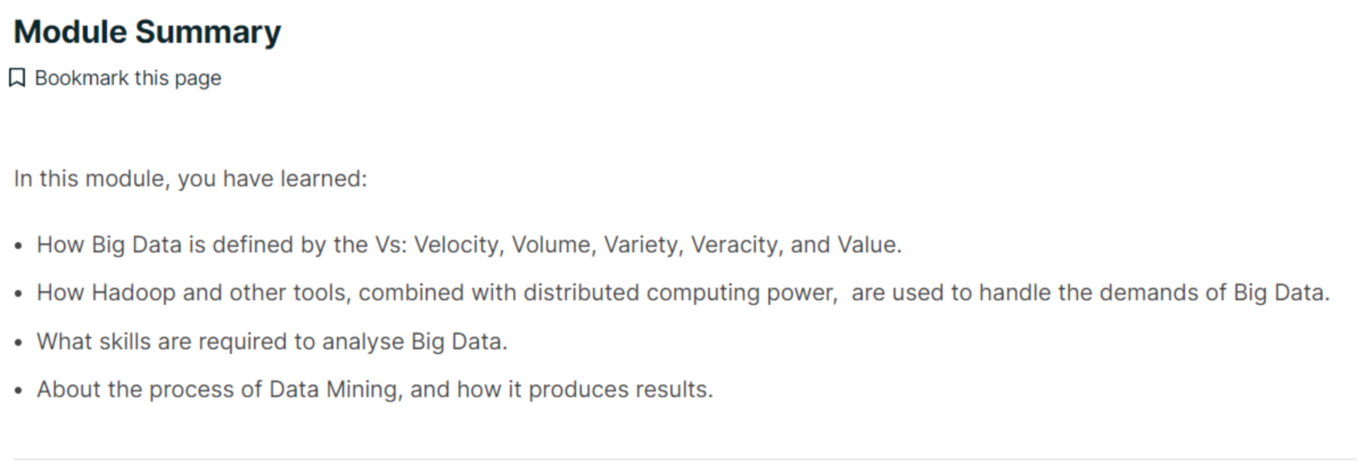
Mining Data

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.

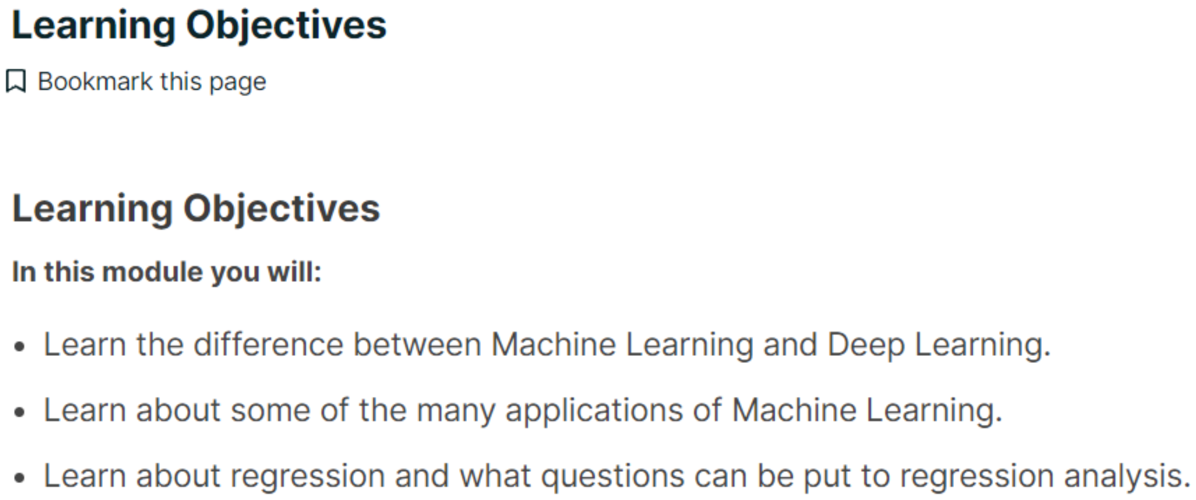


Module Summary

In this module, you have learned:

- How Big Data is defined by the Vs: Velocity, Volume, Variety, Veracity, and Value.
- How Hadoop and other tools, combined with distributed computing power, are used to handle the demands of Big Data.
- What skills are required to analyse Big Data.
- About the process of Data Mining, and how it produces results.

Module 4



Learning Objectives

Bookmark this page

Learning Objectives

In this module you will:

- Learn the difference between Machine Learning and Deep Learning.
- Learn about some of the many applications of Machine Learning.
- Learn about regression and what questions can be put to regression analysis.

What’s the Difference?

In Data Science, there are many terms that are used interchangeably, so let’s explore the most common ones.

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 analyse these vast data sets, and new knowledge and insights are becoming available to everyone.

Big data is often described in terms of five Vs - 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 has 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 layered deep and are the reason deep learning algorithms become more efficient as the datasets 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 intersection between AI and data science, but one is not a subset of the other.

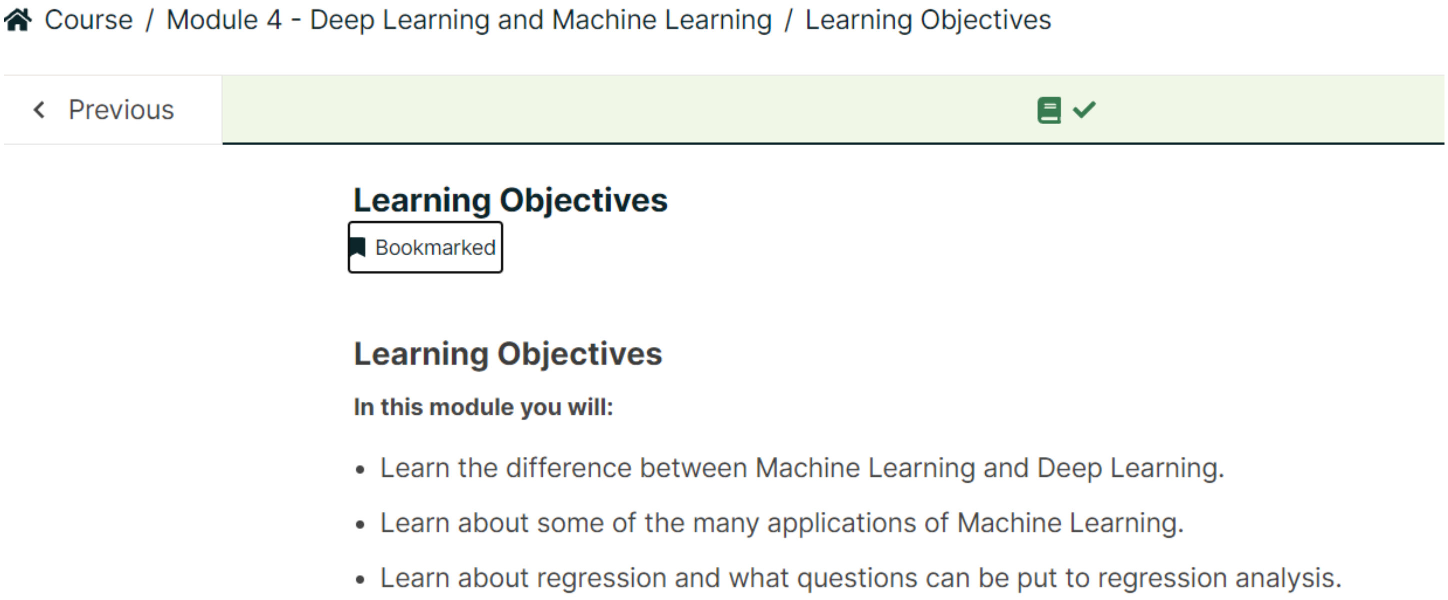
Rather, data science is a broad term that encompasses the entire data processing methodology.

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.


Course / Module 4 - Deep Learning and Machine Learning / Learning Objectives
Previous
Bookmarked

Learning Objectives

Learning Objectives

In this module you will:

- Learn the difference between Machine Learning and Deep Learning.
- Learn about some of the many applications of Machine Learning.
- Learn about regression and what questions can be put to regression analysis.

What’s the Difference?

In Data Science, there are many terms that are used interchangeably, so let’s explore

the most common ones.

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 analyse

these vast data sets, and new knowledge and insights are becoming available to everyone.

Big data is often described in terms of five Vs - 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 has 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 layered deep and are the reason deep learning algorithms become more efficient as the datasets 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.

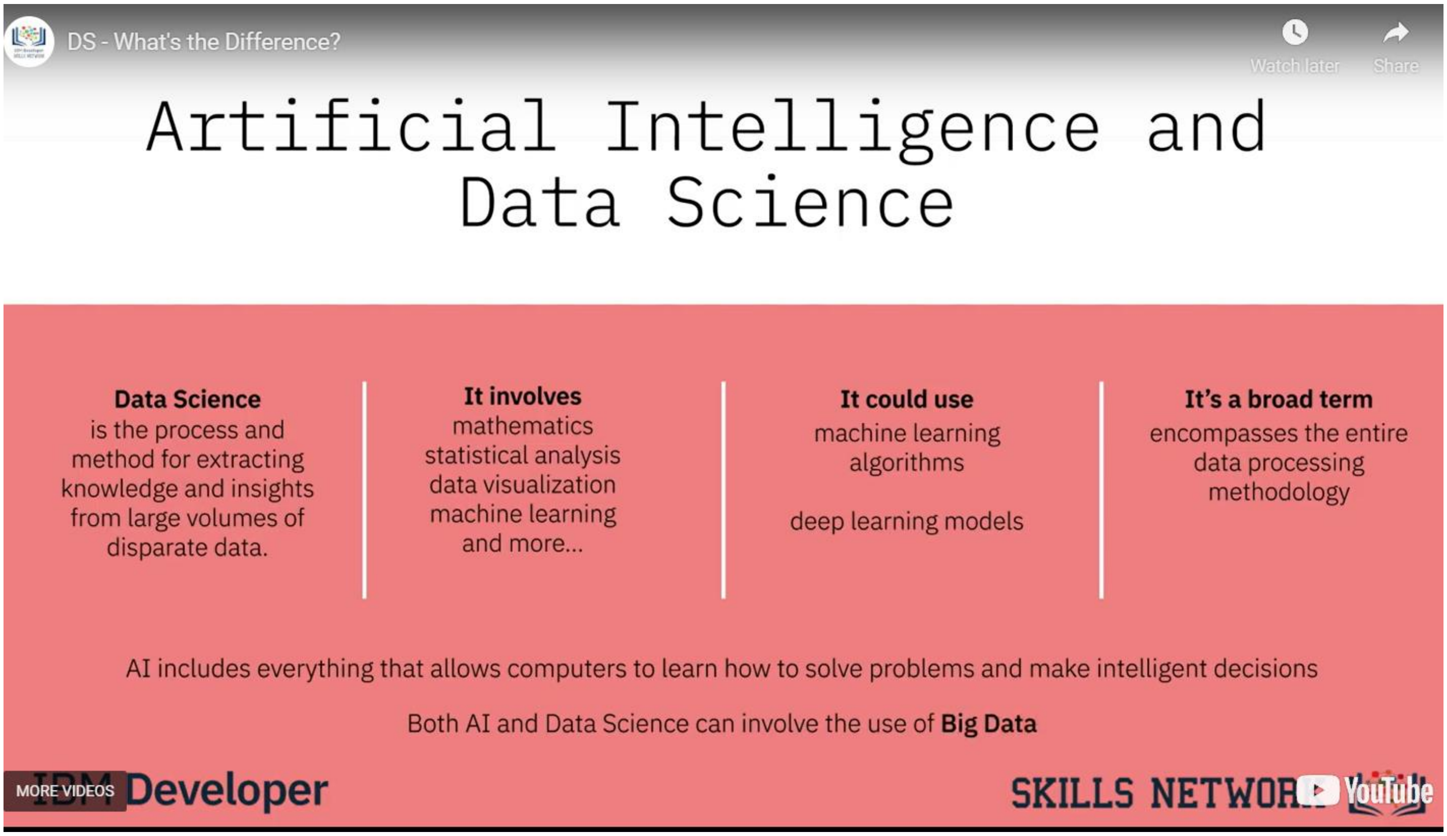
There is some intersection between AI and data science, but one is not a subset of the other.

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.



Artificial Intelligence and Data Science

Data Science is the process and method for extracting knowledge and insights from large volumes of disparate data.

It involves mathematics statistical analysis data visualization machine learning and more...

It could use machine learning algorithms

deep learning models

It's a broad term encompasses the entire data processing methodology

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**

Neural Network and Deep Learning

- It's a, I guess Computer Science's attempt
to mimic a real,
the neurons and how our brain actually functions.
So 20, 30 years ago a neural network
would have some inputs that would come in.
They would be fed into different processing nodes
that would then do some transformation on them
and aggregate them or something,
and then maybe go to another level of nodes
and finally some output would come out.
And I can remember training a neural network
to recognize digits, handwritten digits and stuff.
So a neural network is trying to use
a computer program that will mimic
how neurons, how our brains use neurons to process things,
brains to synapse, neurons to synapses
and building these complex networks that can be trained.
So a neural network starts out
with some inputs and some outputs
and you keep feeding these inputs in
to try to see what kinds of transformations
will get to these outputs,
and you keep doing this over and over and over again
in a way that this network should converge
so these input, the transformations
will eventually get these outputs.
The problem with neural networks was that
even though the theory was there
and they did work on small problems,

like recognizing handwritten digits and things like that,
they were computationally very intensive,
and so they went out of favor.
I stopped teaching them,
well, probably 15 years ago.
Then all of a sudden we started hearing about deep learning.
I heard the term deep learning.
This is another term that
when did you first hear it?
Fours years ago, five years ago?
So I finally said,
"What the hell is deep learning?
It's really doing all this great stuff.
What is it?"
I Google it and I find this is neural networks on steroids.
What they did was they just had more
multiple layers of neural networks
and they use lots and lots and lots
of computing power to solve them.
Just before this interview
I had a young faculty member in the marketing department
whose research is partially based on deep learning.
She needs a computer that has
a graphics processing unit in it
because it takes an enormous amount of matrix
and linear algebra calculations
to actually do all of the mathematics
that you need in neural networks,
but they are now quite capable.
We now have neural networks and deep learning
that can recognize speech, can recognize people.

If you're out there and getting your face recognized,
I guarantee that NSA has a lot of work
going on in neural networks.
The University, right now,
as Director of Research Computing,
I have some small set of machines,
down at our South Data Center,
and I went in there last week
and there were just piles and piles and piles
of cardboard boxes all from Dell with a GPU on the side.
Well, a GPU is a graphics processing unit.
There is only one application in this University
that needs 200 servers,
each with graphics processing units in it,
and each graphics processing unit
has the equivalent of 600 cores of processing,
so this is tens of thousands of processing cores.
That is for deep learning.
I guarantee.
(music)
Some of the first ones are speech recognition.
Yann LeCun who teaches the deep learning class at NYU
and is also the Head Data Scientist at Facebook,
comes into class with a notebook,
and it's a pretty thick notebook.
It looks a little odd because it's like this.
It's that thick because it has
a couple of graphics processing units in it
and then he will ask the class
to start to speak to this thing
and it will train while he's in class,

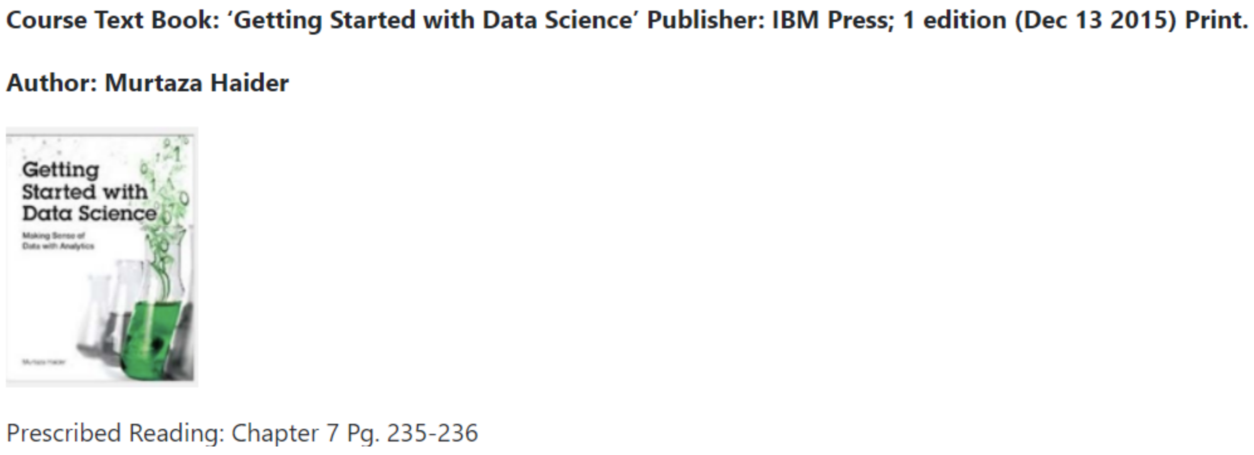
he will train a neural network to recognize speech.
So recognizing speech, recognizing people, images,
classifying images, almost all of the traditional tasks
that neural nets used to work on in little tiny things,
now they can do really, really large things.
It will learn, on it's own, the difference between
a cat and a dog and different kinds of objects.
It doesn't have to be taught.
It doesn't, it just learns.
That's why they call it deep learning,
and if you hear, he plays this.
If you hear how it recognizes speech and generates speech,
it sounds like a baby learning to talk.
You can just, you're like
(babbles)
All of sudden this stupid machine is talking to you
and learned how to talk.
That's cool.
(music)
You need to learn some linear algebra.
A lot of this stuff is based on matrix
and linear algebra, so you need to know how to do,
use linear algebra and do transformations.
Now, on the other hand,
there's now lots of packages out there
that will do deep learning
and they'll do all the linear algebra for you.
But you should have some idea
of what is happening underneath.
Deep learning, in particular,
needs really high powered computational power.

So it's not something that you're going to go out
and do on your notebook for, you could play with it,
but if you really want to do it seriously
you have to have some special computational resources.

Applications of Machine Learning

Start of transcript. Skip to the end.
Everybody now deals with machine learning, but recommender systems are
certainly one of the major applications, classifications, cluster analysis, trying
to find some of the some of the marketing questions from 20 years ago,
market basket analysis, what goods tend to be bought together.
That was computationally a very difficult problem. I mean we are now
doing that all the time with machine learning. So predictive analytics is
another area of machine learning. We're using new techniques to predict
things that statisticians don't particularly like. Decision Trees,
Bayesian Analysis, Naive Bayes, lots of different techniques. The nice thing
about them is that in packages like R now, you really have to
understand how these techniques can be used and you don't have to know
exactly how to do them but you have to understand what their meanings are.
Precision versus recall and the problems of over
over sampling and overfitting so you can, someone who knows a little bit about
data science can apply these techniques, but they really need to know maybe not
the details of the technique as much as how, what the trade-offs are. So some
applications of machine learning in FinTech are probably a couple of
different things I can talk about there. One of them is recommendations, right? So
when you use Netflix, or you use Facebook, or a lot of different software services,
the recommendations are served to you. Meaning, "Hey, you are a user, you have
watched this show, so maybe you'd like to see this other show." Or, "You follow
this person, so maybe you should follow that other person." It's actually kind of
the same thing in FinTech. Because you've looked at, if you are an investment

professional, and because you have looked at this investment idea, it might be really cool for you to look at this other investment idea, which is kind of similar. It is a similar kind of asset, or it is a similar kind of company, or it is a similar kind of technique for doing the investment. So we can apply recommendations using machine learning throughout a lot of different parts of FinTech. Another one that people talk about and is important, especially in the retail aspects of banking and finance, is fraud detection; trying to determine whether a charge that comes through a credit card is fraudulent or not, in real time, is a machine learning problem. You have to learn from all of the transactions that have happened previously. And build a model. And when the charge comes through, you have to compute all this stuff, and say, "Yeah, we think that's okay." Or, "Hmm, that's not so that's not so good, let's route it to our fraud people to check."



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

Author: Murtaza Haider

Prescribed Reading: Chapter 7 Pg. 235-236

Chapter 7. Why Tall Parents Don't Have Even Taller Children

You might have noticed that taller parents often have tall children who are not necessarily taller than their parents and that's a good thing. This is not to suggest that children born to tall parents are not necessarily taller than the rest. That may be the case, but they are not necessarily taller than their own "tall" parents. Why I think this to be a good thing requires a simple mental simulation. Imagine if every successive generation born to tall parents were taller than their parents, in a matter of a couple of millennia, human beings would become uncomfortably tall for their own good, requiring even bigger furniture, cars, and planes.

Sir Frances Galton in 1886 studied the same question and landed upon a statistical technique we today know as regression models. This chapter explores the workings of regression models, which have become the workhorse of statistical analysis. In almost all empirical pursuits of research, either in the academic or professional fields, the use of regression models, or their variants, is ubiquitous. In medical science, regression models are being used to develop more effective medicines, improve the methods for operations, and optimize resources for small and large hospitals. In the business world, regression models are at the forefront of analyzing consumer behavior, firm productivity, and competitiveness of public and private sector entities.

I would like to introduce regression models by narrating a story about my Master's thesis. I believe that this story can help explain the utility of regression models.

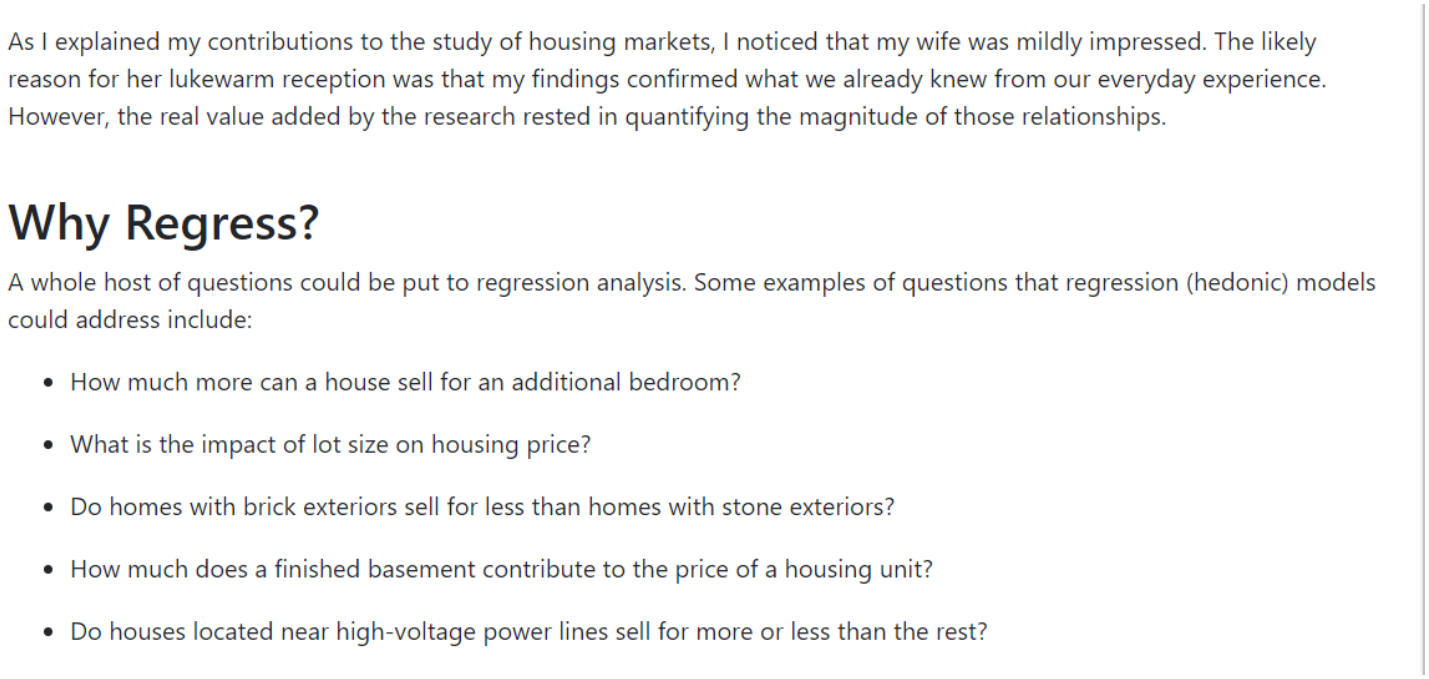
The Department of Obvious Conclusions

In 1999, I finished my Masters' research on developing hedonic price models for residential real estate properties. It took me three years to complete the project involving 500,000 real estate transactions. As I was getting ready for the defense, my wife generously offered to drive me to the university. While we were on our way, she asked, "Tell me, what have you found in your research?". I was delighted to be finally asked to explain what I have been up to for the past three years. "Well, I have been studying the determinants of housing prices. I have found that larger homes sell for more than smaller homes," I told my wife with a triumphant look on my face as I held the draft of the thesis in my hands

We were approaching the on-ramp for a highway. As soon as I finished the sentence, my wife suddenly turned the car to the shoulder and applied brakes. As the car stopped, she turned to me and said: "I can't believe that they are giving you a Master's degree for finding just that. I could have told you that larger homes sell for more than smaller homes."

At that very moment, I felt like a professor who taught at the department of obvious conclusions. How can I blame her for being shocked that what is commonly known about housing prices will earn me a Master's degree from a university of high repute?

I requested my wife to resume driving so that I could take the next ten minutes to explain to her the intricacies of my research. She gave me five minutes instead, thinking this may not require even that. I settled for five and spent the next minute collecting my thoughts. I explained to her that my research has not just found the correlation between housing prices and the size of housing units, but I have also discovered the magnitude of those relationships. For instance, I found that all else being equal, a term that I explain later in this chapter, an additional washroom adds more to the housing price than an additional bedroom. Stated otherwise, the marginal increase in the price of a house is higher for an additional washroom than for an additional bedroom. I found later that the real estate brokers in Toronto indeed appreciated this finding. I also explained to my wife that proximity to transport infrastructure, such as subways, resulted in higher housing prices. For instance, houses situated closer to subways sold for more than did those situated farther away. However, houses near freeways or highways sold for less than others did. Similarly, I also discovered that proximity to large shopping centers had a nonlinear impact on housing prices. Houses located very close (less than 2.5 km) to the shopping centers sold for less than the rest. However, houses located closer (less than 5 km, but more than 2.5 km) to the shopping center sold for more than did those located farther away. I also found that the housing values in Toronto declined with distance from downtown.

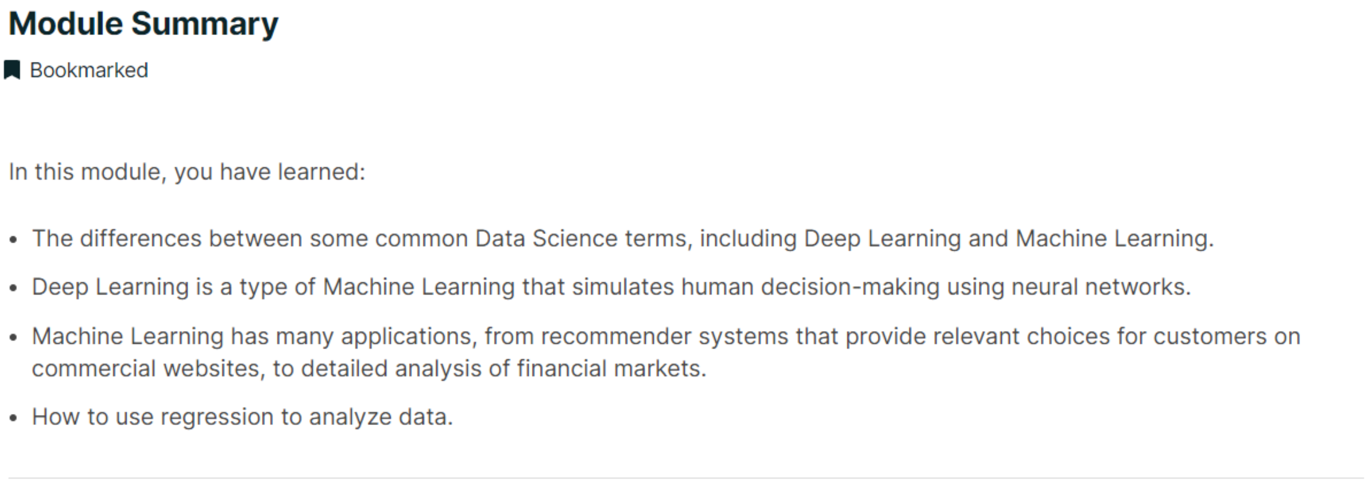


As I explained my contributions to the study of housing markets, I noticed that my wife was mildly impressed. The likely reason for her lukewarm reception was that my findings confirmed what we already knew from our everyday experience. However, the real value added by the research rested in quantifying the magnitude of those relationships.

Why Regress?

A whole host of questions could be put to regression analysis. Some examples of questions that regression (hedonic) models could address include:

- How much more can a house sell for an additional bedroom?
- What is the impact of lot size on housing price?
- Do homes with brick exteriors sell for less than homes with stone exteriors?
- How much does a finished basement contribute to the price of a housing unit?
- Do houses located near high-voltage power lines sell for more or less than the rest?

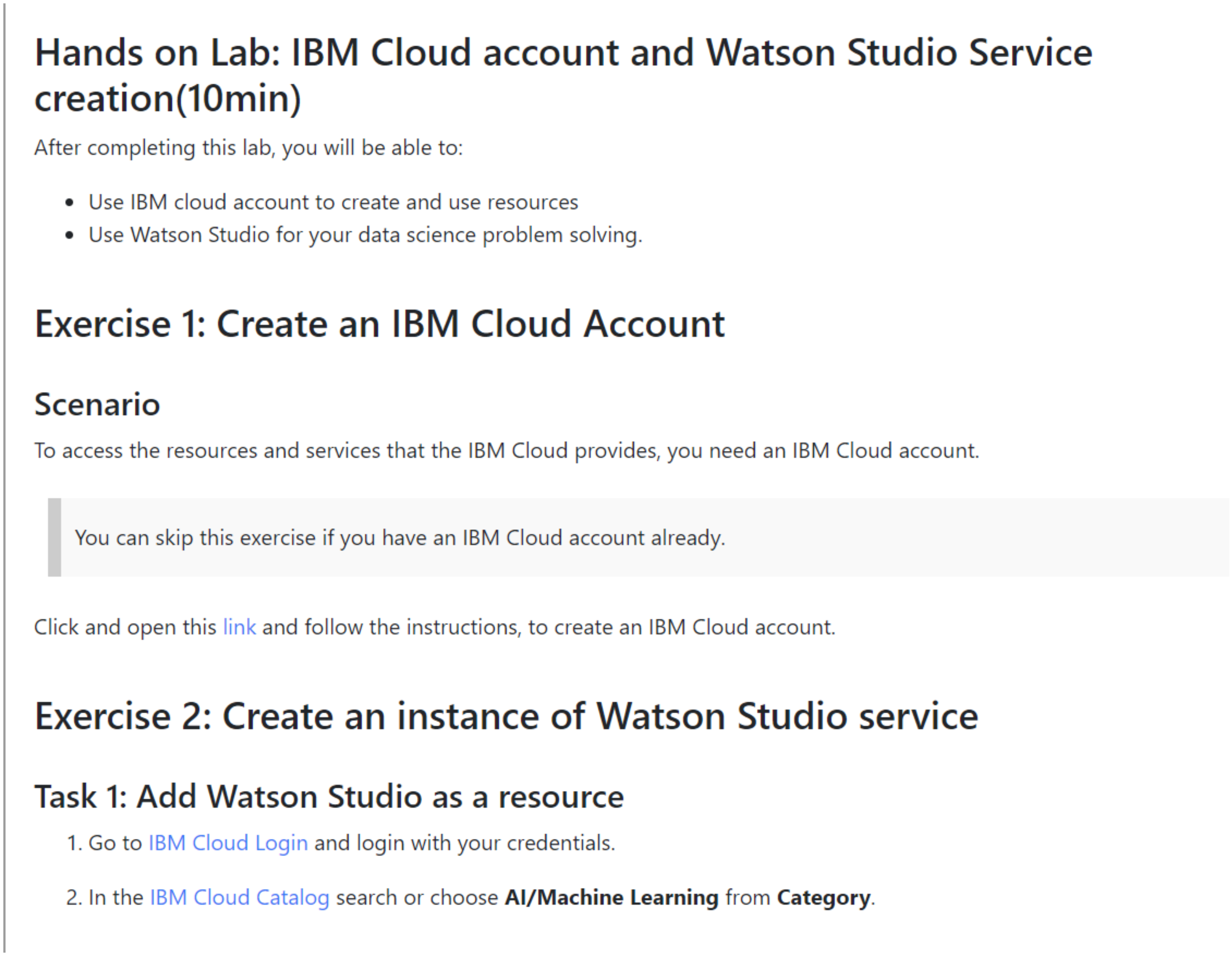


Module Summary

Bookmarked

In this module, 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.



Hands on Lab: IBM Cloud account and Watson Studio Service creation(10min)

After completing this lab, you will be able to:

- Use IBM cloud account to create and use resources
- Use Watson Studio for your data science problem solving.

Exercise 1: Create an IBM Cloud Account

Scenario

To access the resources and services that the IBM Cloud provides, you need an IBM Cloud account.

> You can skip this exercise if you have an IBM Cloud account already.

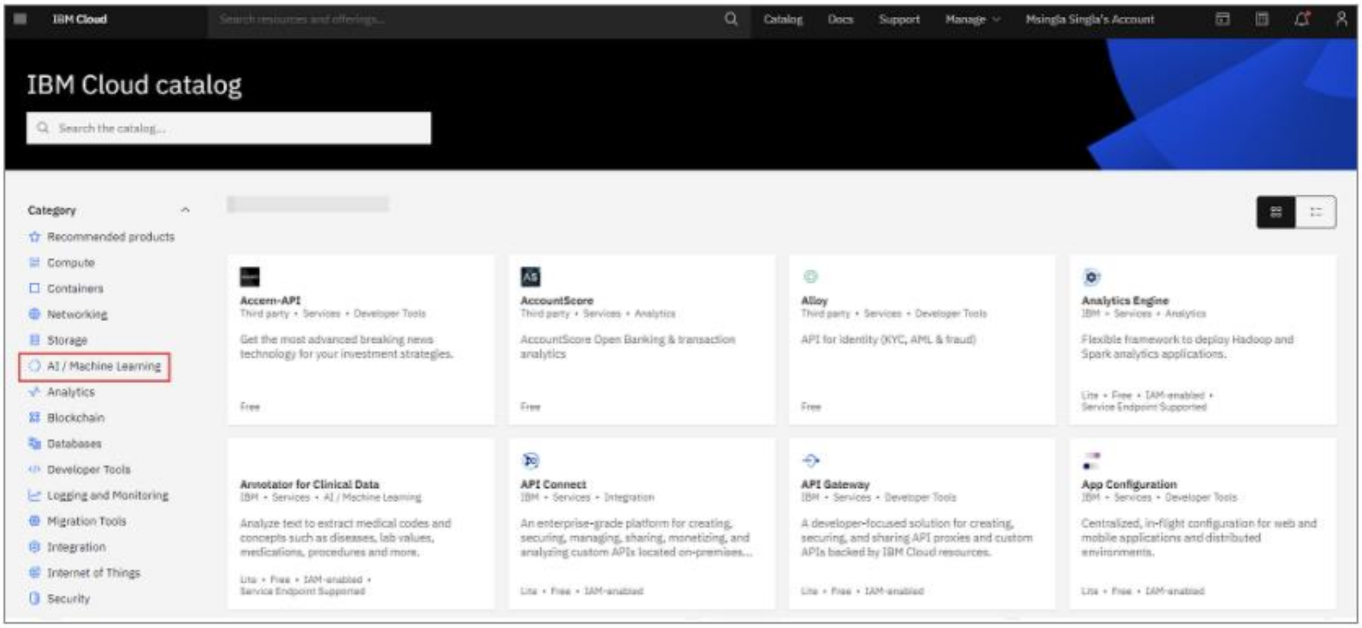
Click and open this link and follow the instructions, to create an IBM Cloud account.

Exercise 2: Create an instance of Watson Studio service

Task 1: Add Watson Studio as a resource

1. Go to IBM Cloud Login and login with your credentials.

2. In the IBM Cloud Catalog search or choose **AI/Machine Learning** from **Category**.



Accern-API
Third party • Services • Developer Tools
Get the most advanced breaking news technology for your investment strategies.
Free

AccountScore
Third party • Services • Analytics
AccountScore Open Banking & transaction analytics
Free

Alloy
Third party • Services • Developer Tools
API for identity (KYC, AML & fraud)
Free

Analytics Engine
IBM • Services • Analytics
Flexible framework to deploy Hadoop and Spark analytics applications.
Lite • Free • IAM-enabled • Service Endpoint Supported

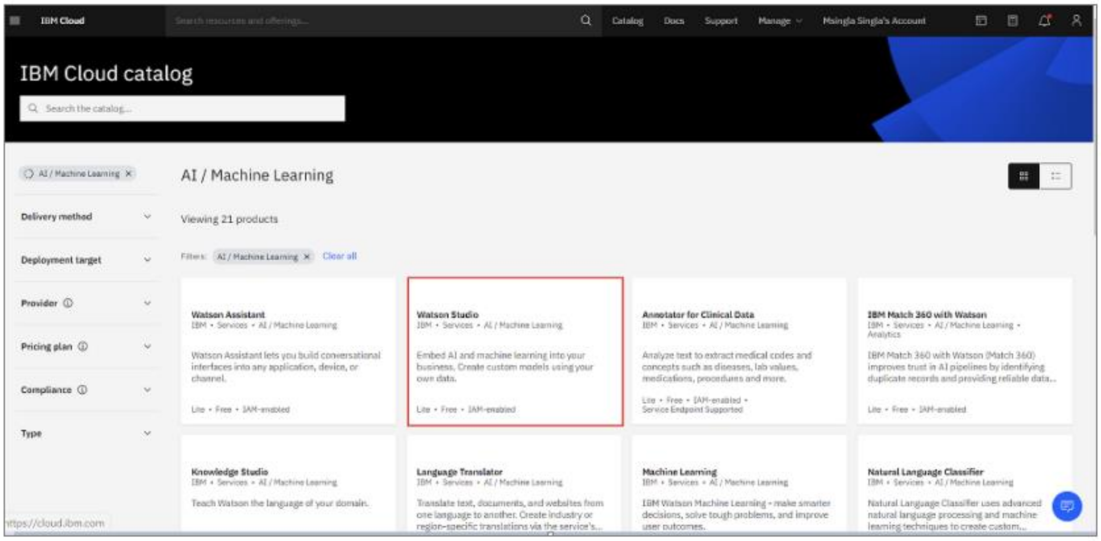
Annotator for Clinical Data
IBM • Services • AI / Machine Learning
Analyze text to extract medical codes and concepts such as diseases, lab values, medications, procedures and more.
Lite • Free • IAM-enabled • Service Endpoint Supported

API Connect
IBM • Services • Integration
An enterprise-grade platform for creating, securing, managing, sharing, monetizing, and analyzing custom APIs located on-premises...
Lite • Free • IAM-enabled

API Gateway
IBM • Services • Developer Tools
A developer-focused solution for creating, securing, and sharing API proxies and custom APIs backed by IBM Cloud resources.
Lite • Free • IAM-enabled

App Configuration
IBM • Services • Developer Tools
Centralized, in-flight configuration for web and mobile applications and distributed environments.
Lite • Free • IAM-enabled

and select **Watson Studio**.



Watson Assistant
IBM • Services • AI / Machine Learning
Watson Assistant lets you build conversational interfaces into any application, device, or channel.
Lite • Free • IAM-enabled

Watson Studio
IBM • Services • AI / Machine Learning
Embed AI and machine learning into your business. Create custom models using your own data.
Lite • Free • IAM-enabled

Annotator for Clinical Data
IBM • Services • AI / Machine Learning
Analyze text to extract medical codes and concepts such as diseases, lab values, medications, procedures and more.
Lite • Free • IAM-enabled • Service Endpoint Supported

IBM Match 360 with Watson
IBM • Services • AI / Machine Learning • Analytics
IBM Match 360 with Watson (Match 360) improves trust in AI pipelines by identifying duplicate records and providing reliable data...
Lite • Free • IAM-enabled

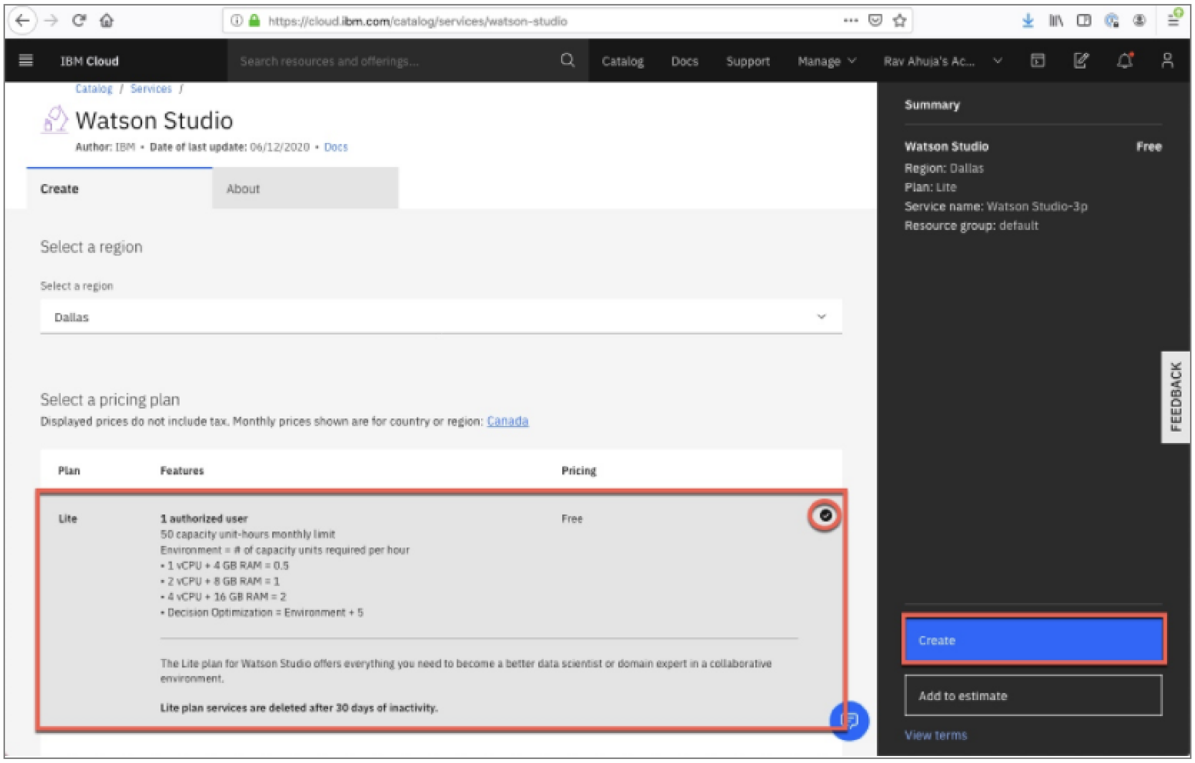
Knowledge Studio
IBM • Services • AI / Machine Learning
Teach Watson the language of your domain.

Language Translator
IBM • Services • AI / Machine Learning
Translate text, documents, and websites from one language to another. Create industry or region-specific translations via the service's...

Machine Learning
IBM • Services • AI / Machine Learning
IBM Watson Machine Learning - make smarter decisions, solve tough problems, and improve user outcomes.

Natural Language Classifier
IBM • Services • AI / Machine Learning
Natural Language Classifier uses advanced natural language processing and machine learning techniques to create custom...

3. On the Watson Studio page, select an appropriate region depending on where you are accessing from, verify that the **Lite** plan is selected, and then click **Create**.



Watson Studio

Author: IBM • Date of last update: 06/12/2020 • Docs

Select a region

Select a region: Dallas

Select a pricing plan

Displayed prices do not include tax. Monthly prices shown are for country or region: Canada

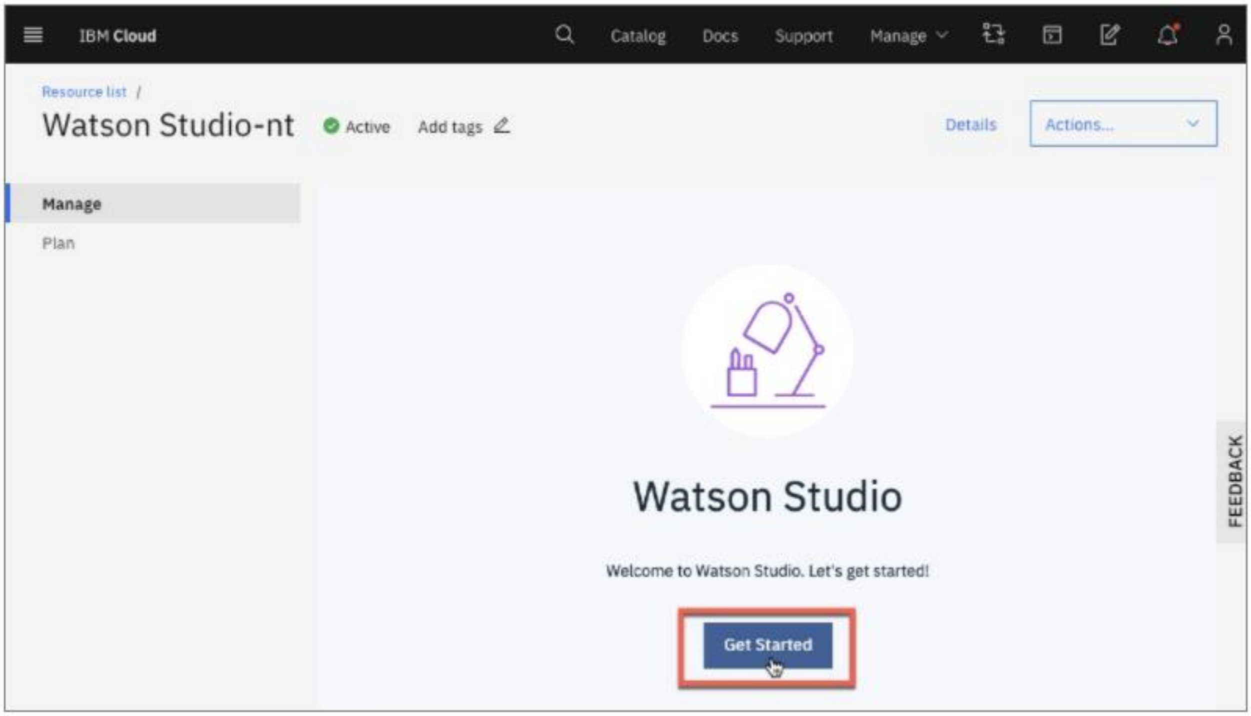
Plan	Features	Pricing
Lite	**1 authorized user** 50 capacity unit-hours monthly limit Environment = # of capacity units required per hour • 1 vCPU + 4 GB RAM = 0.5 • 2 vCPU + 8 GB RAM = 1 • 4 vCPU + 16 GB RAM = 2 • Decision Optimization = Environment + 5 The Lite plan for Watson Studio offers everything you need to become a better data scientist or domain expert in a collaborative environment. **Lite plan services are deleted after 30 days of inactivity.**	Free

Summary

Watson Studio Free

Region: Dallas
Plan: Lite
Service name: Watson Studio-3p
Resource group: default

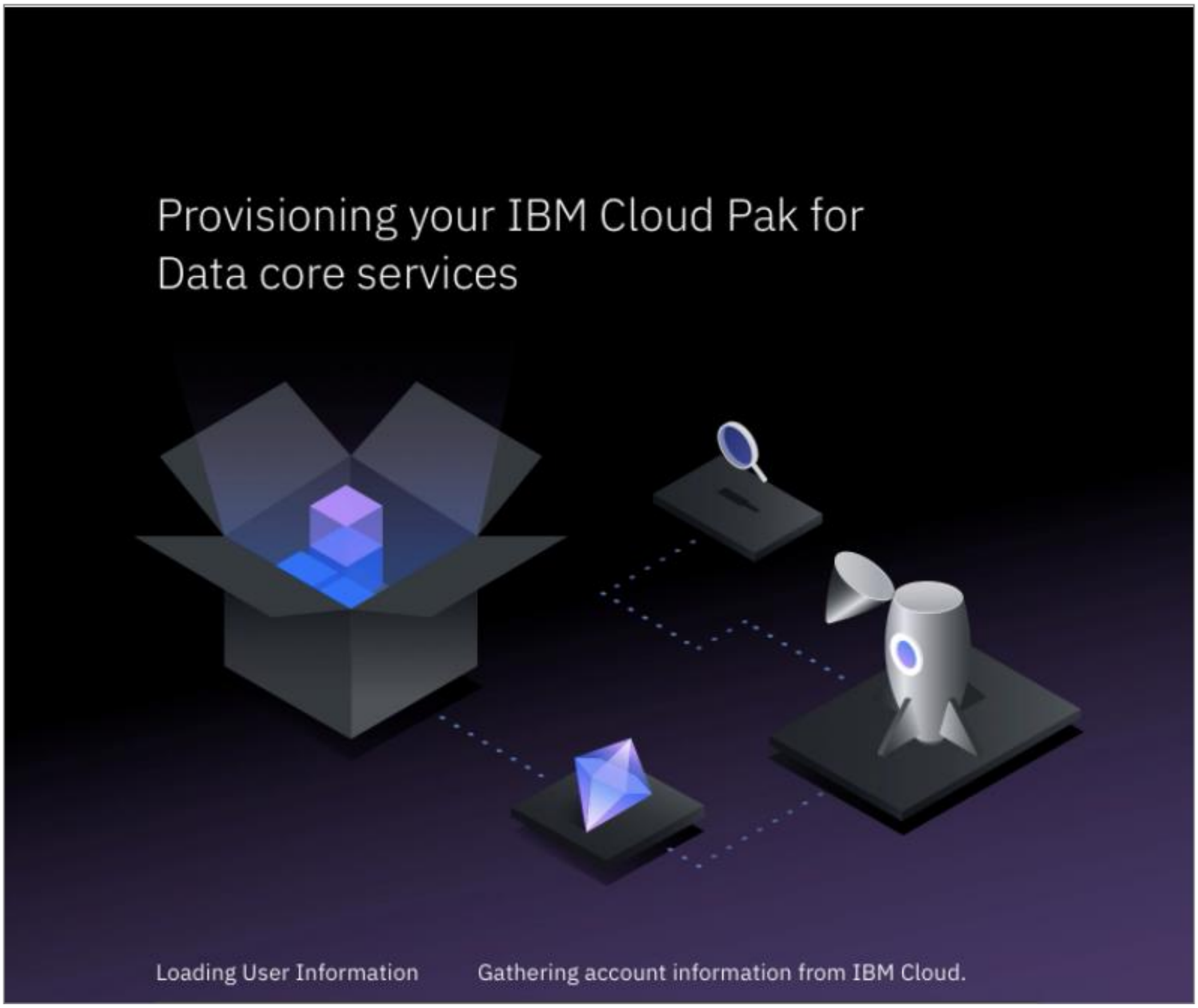
4. Once the Watson Studio instance is successfully created, click on **Get Started**.



Watson Studio

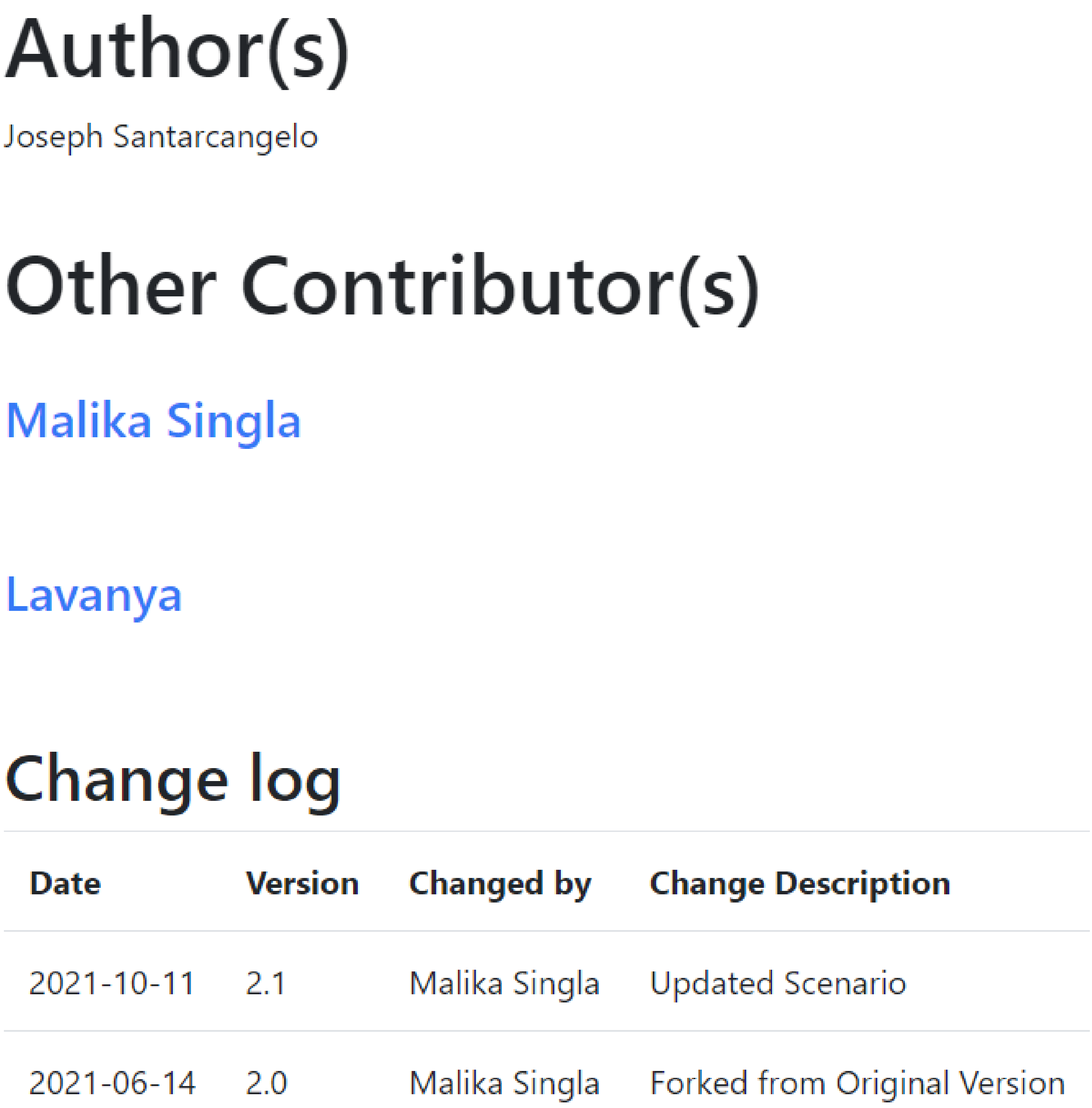
Welcome to Watson Studio. Let's get started!

5. Once you get started, the first time, your IBM Cloud Pak for Data core services is provisioned.



Provisioning your IBM Cloud Pak for Data core services

Loading User Information Gathering account information from IBM Cloud.



Author(s)

Joseph Santarcangelo

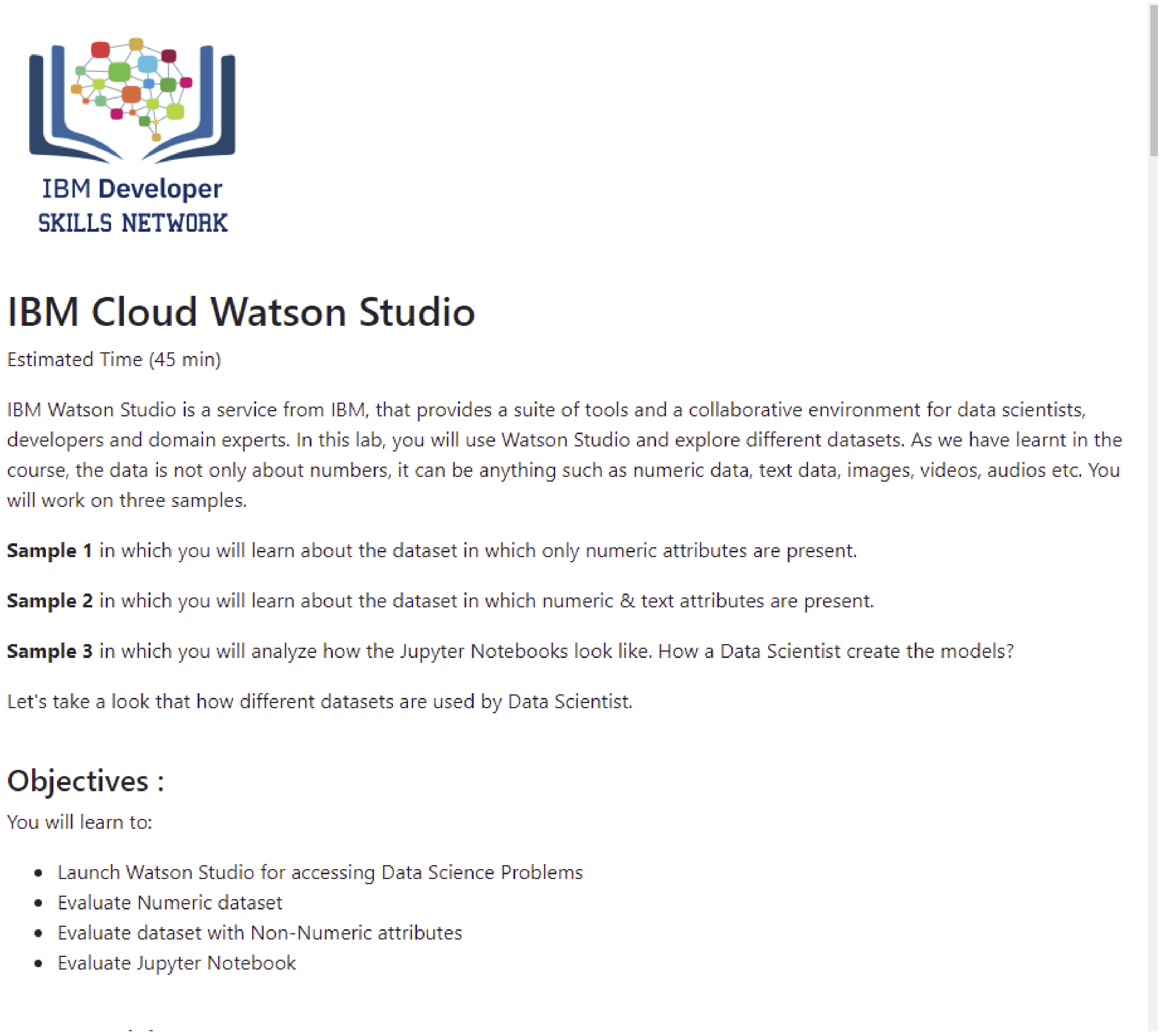
Other Contributor(s)

Malika Singla

Lavanya

Change log

Date	Version	Changed by	Change Description
2021-10-11	2.1	Malika Singla	Updated Scenario
2021-06-14	2.0	Malika Singla	Forked from Original Version



IBM Developer
SKILLS NETWORK

IBM Cloud Watson Studio

Estimated Time (45 min)

IBM Watson Studio is a service from IBM, that provides a suite of tools and a collaborative environment for data scientists, developers and domain experts. In this lab, you will use Watson Studio and explore different datasets. As we have learnt in the course, the data is not only about numbers, it can be anything such as numeric data, text data, images, videos, audios etc. You will work on three samples.

Sample 1 in which you will learn about the dataset in which only numeric attributes are present.

Sample 2 in which you will learn about the dataset in which numeric & text attributes are present.

Sample 3 in which you will analyze how the Jupyter Notebooks look like. How a Data Scientist create the models?

Let's take a look that how different datasets are used by Data Scientist.

Objectives :

You will learn to:

- Launch Watson Studio for accessing Data Science Problems
- Evaluate Numeric dataset
- Evaluate dataset with Non-Numeric attributes
- Evaluate Jupyter Notebook

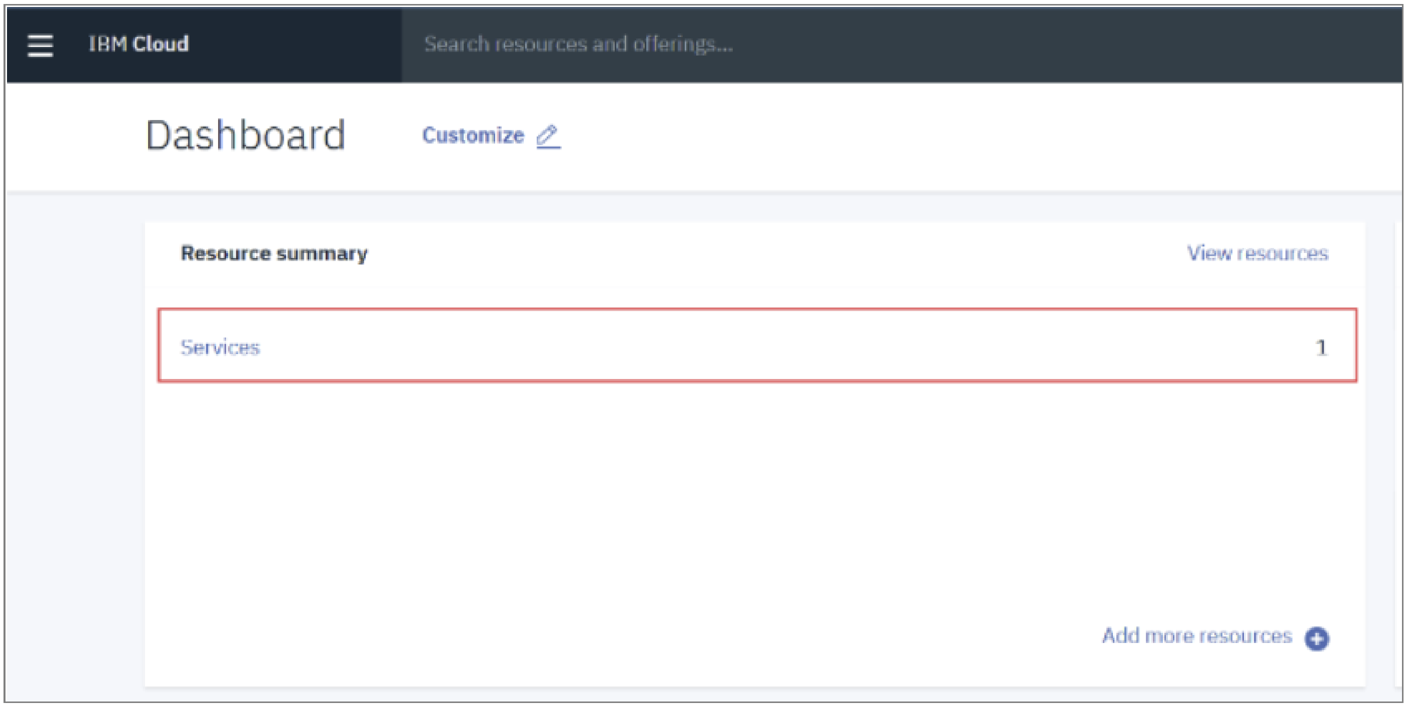
Pre-requisite:

Before you start, you need to have an IBM Cloud account. If not, follow the instructions given in the link

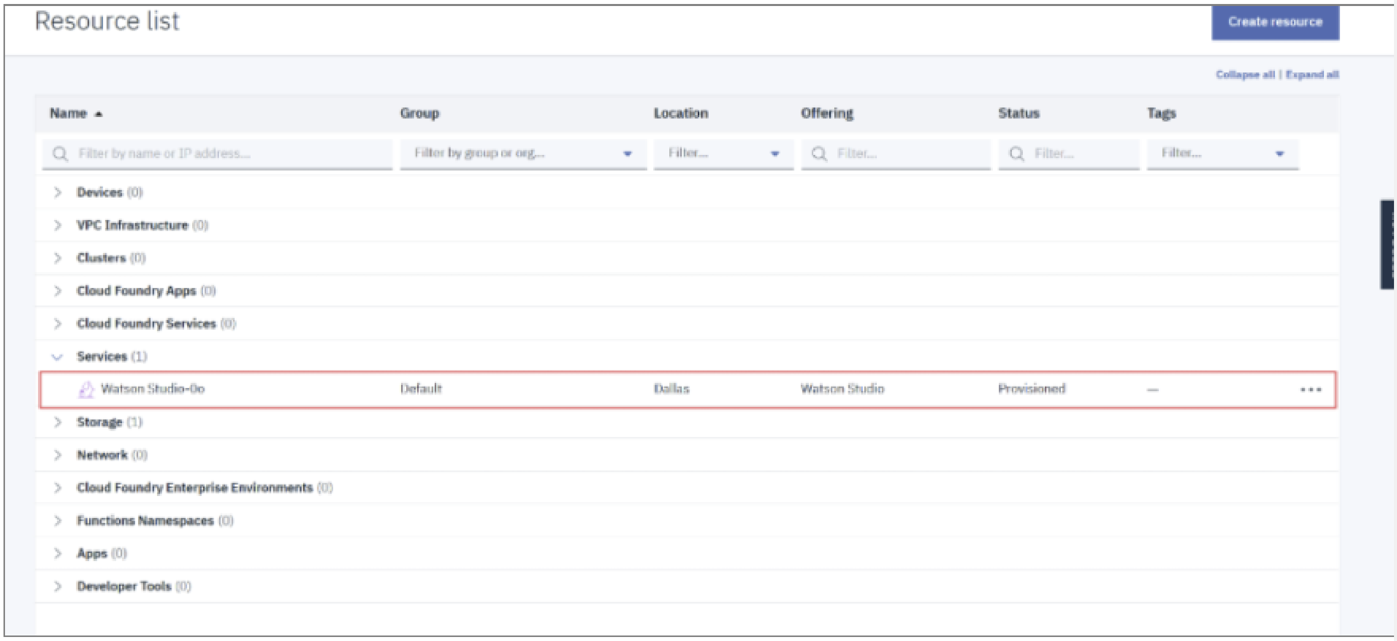
Exercise 1: Launch Watson Studio for accessing Data Science Problems

1. Login to IBM Cloud: https://cloud.ibm.com/login

2. Scroll down and click *Services* given in *Resource Summary*.



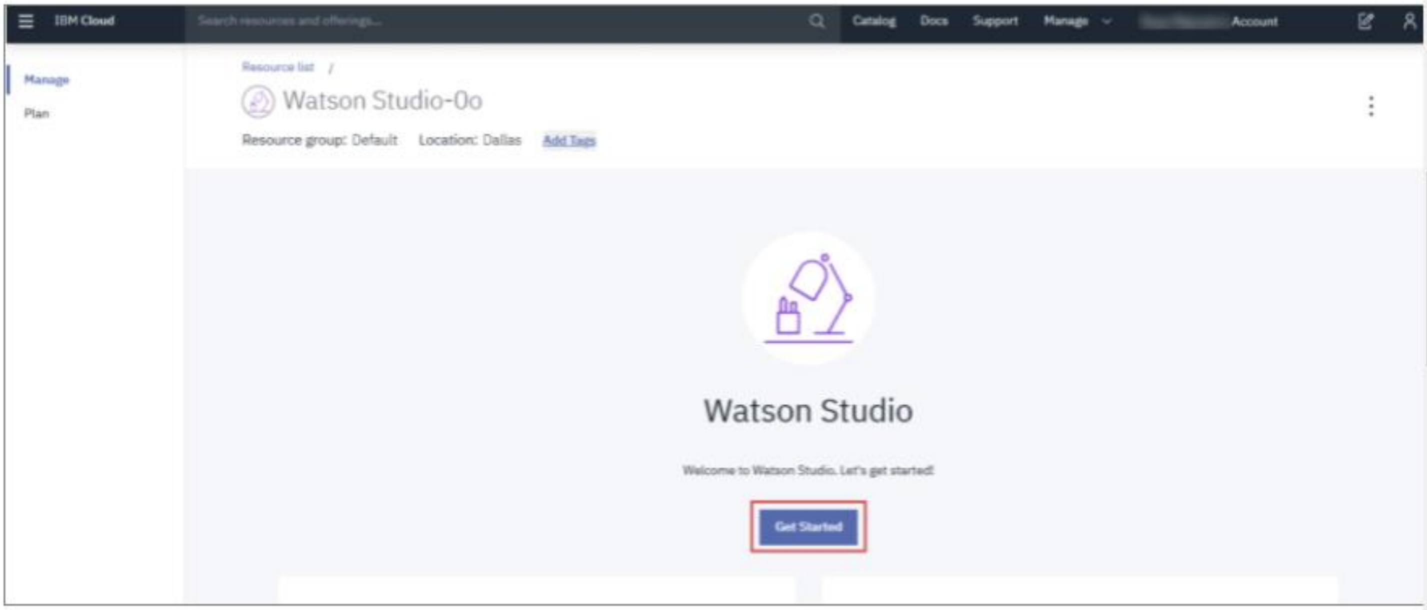
3. When you click on Services, all your existing services will be shown in the list. Click the Watson Studio service you created:


Create resource
Collapse all | Expand all
Filter by name or IP address...
Filter by group or org...
Filter...
Filter...
Filter...
Filter...

Resource list

Name ▲	Group	Location	Offering	Status	Tags
> Devices (0)					
> VPC Infrastructure (0)					
> Clusters (0)					
> Cloud Foundry Apps (0)					
> Cloud Foundry Services (0)					
∨ Services (1)					
Watson Studio-0o	Default	Dallas	Watson Studio	Provisioned	— •••
> Storage (1)					
> Network (0)					
> Cloud Foundry Enterprise Environments (0)					
> Functions Namespaces (0)					
> Apps (0)					
> Developer Tools (0)					

4. Click *Get Started*.



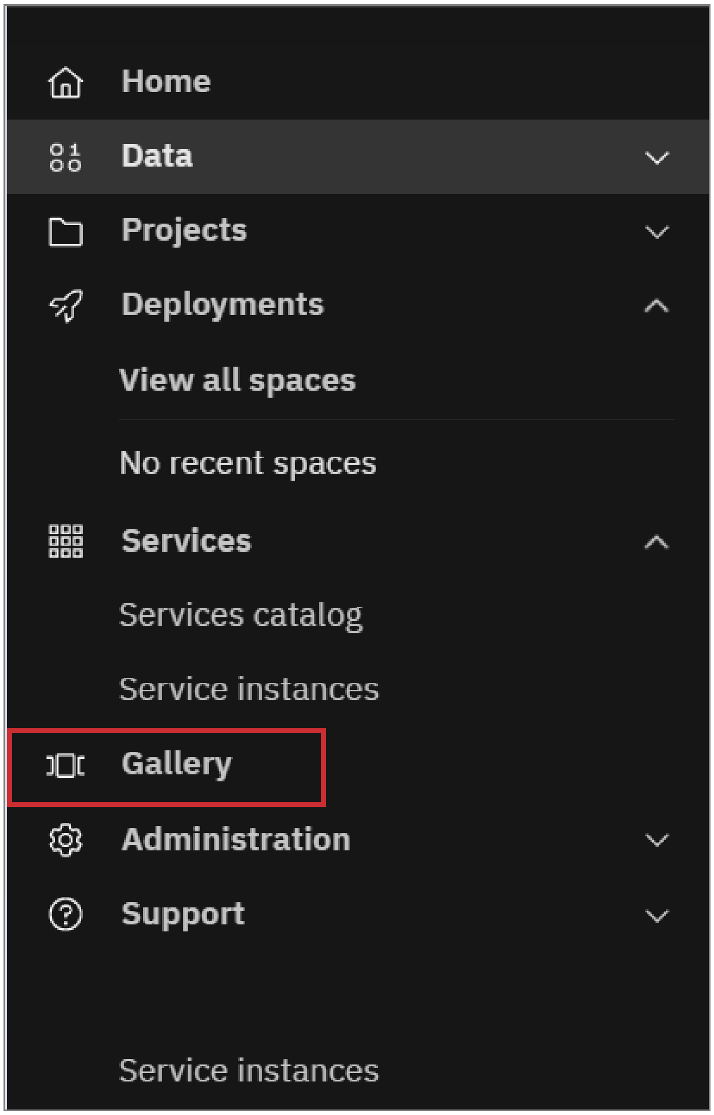
Watson Studio-0o

Resource group: Default Location: Dallas Add Tags

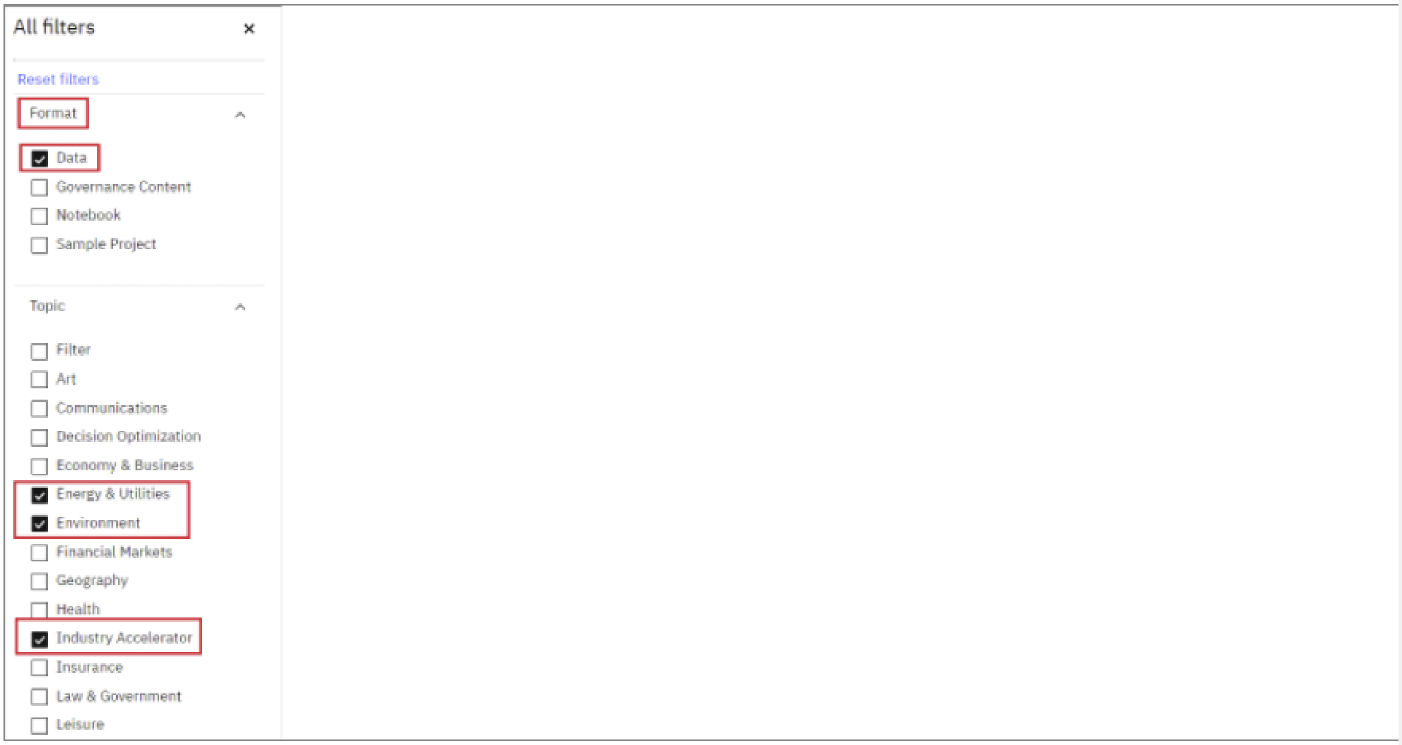
Watson Studio

Welcome to Watson Studio. Let's get started!

2. Click on *Gallery*.



3. Select *All Filters*. From *Format* select *Data* and from *Topic* select *Energy & Utilities, Enviornment and Industry Accelerator*



All filters ×

Reset filters

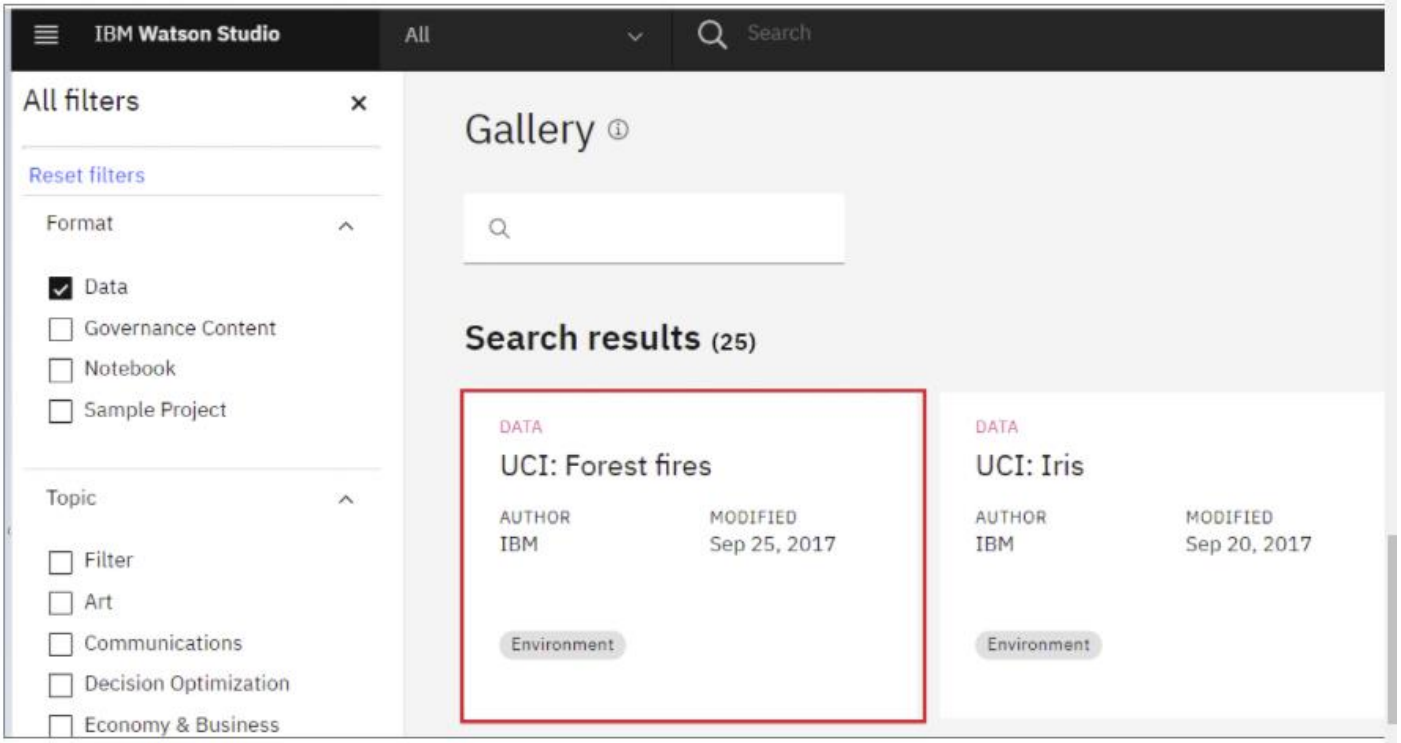
Format

☑ Data
☐ Governance Content
☐ Notebook
☐ Sample Project

Topic

☐ Filter
☐ Art
☐ Communications
☐ Decision Optimization
☐ Economy & Business
☑ Energy & Utilities
☑ Environment
☐ Financial Markets
☐ Geography
☐ Health
☑ Industry Accelerator
☐ Insurance
☐ Law & Government
☐ Leisure

4. Click on *UCI: Forest Fires*.



Search results (25)

DATA

UCI: Forest fires

AUTHOR IBM

MODIFIED Sep 25, 2017

Environment

DATA

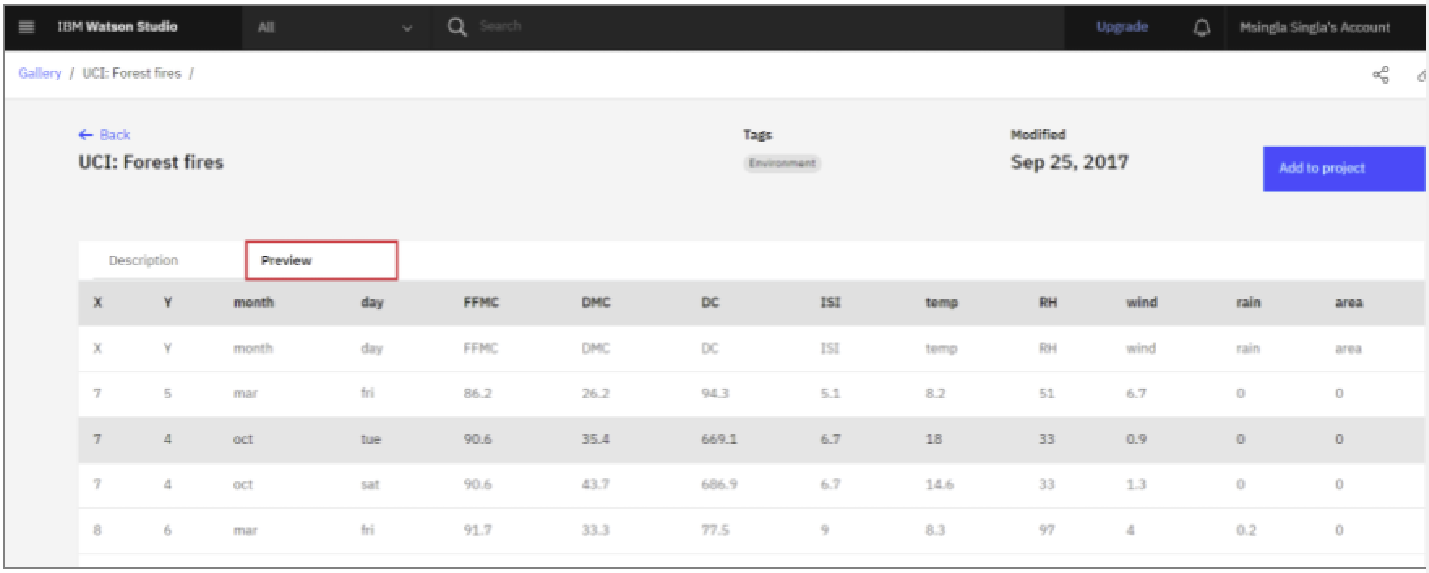
UCI: Iris

AUTHOR IBM

MODIFIED Sep 20, 2017

Environment

5. Preview the data using the *Preview* option.

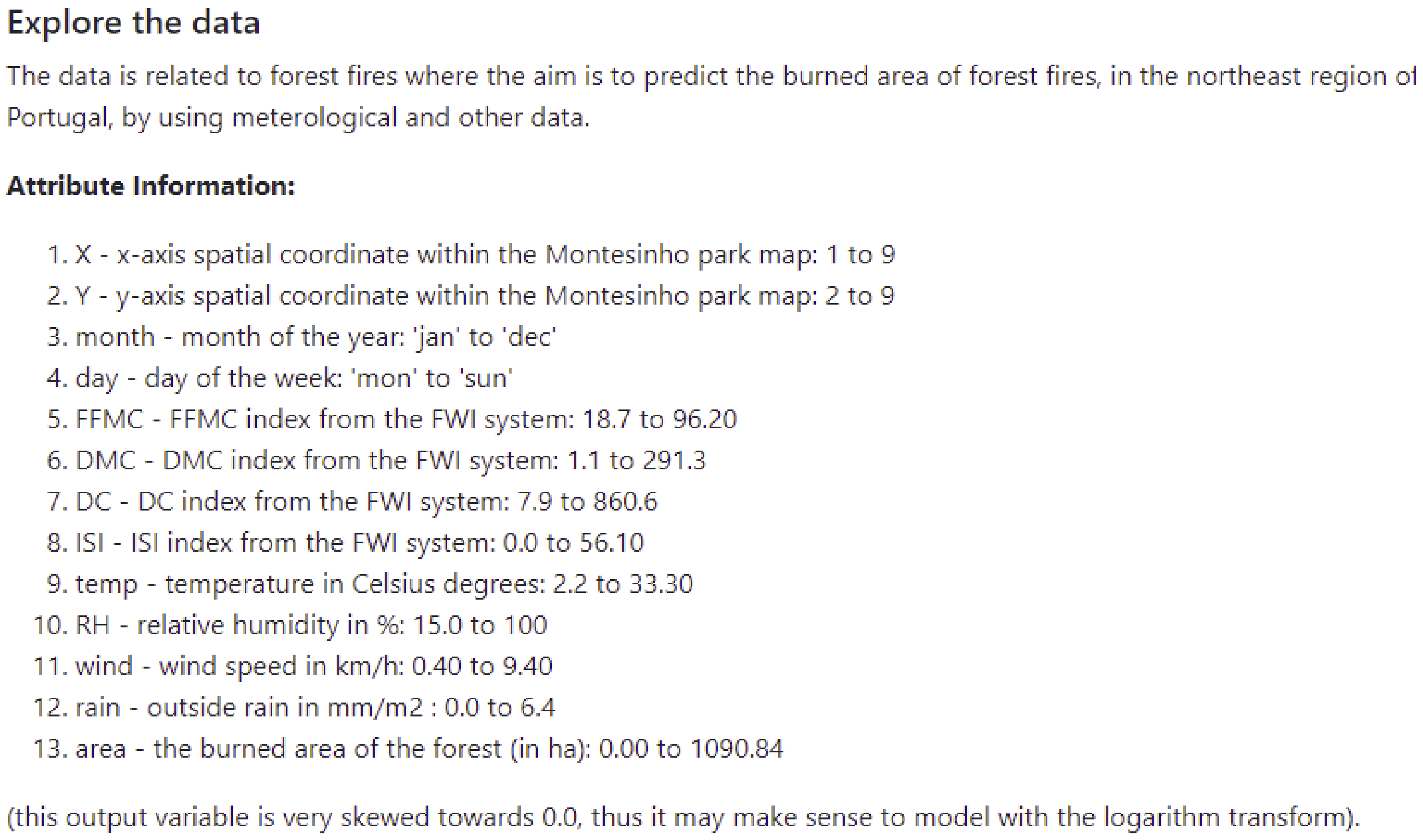


UCI: Forest fires

Tags Environment

Modified Sep 25, 2017

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0



Explore the data

The data is related to forest fires where the aim is to predict the burned area of forest fires, in the northeast region of Portugal, by using meterological and other data.

Attribute Information:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

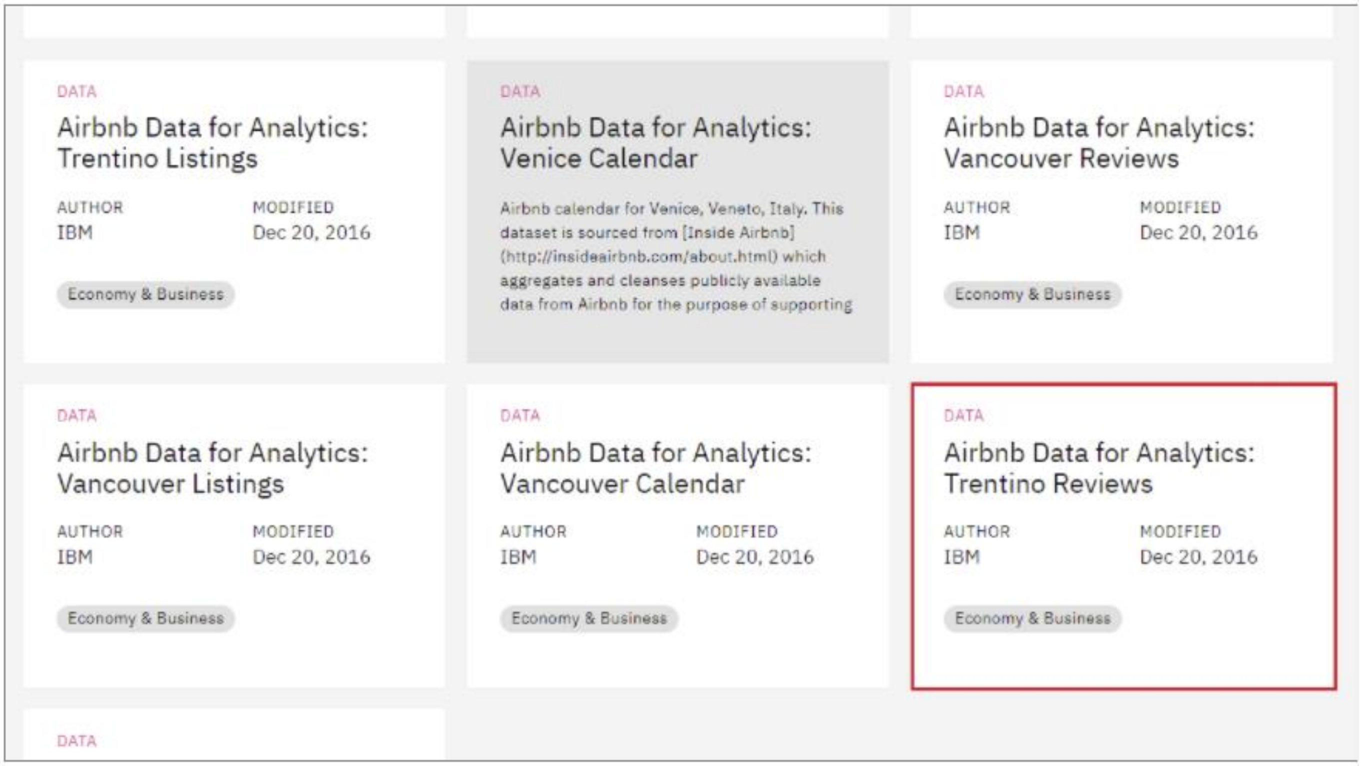
(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Exercise 2: Evaluate Non-Numeric dataset

The data doesn't have to be only based on numbers. Data can be text, images and other types as well. Let's look into data having text values.

1. Use the *All Filters*. From *Format* select *Data* and from *Topic* select *Economy and Business*.

You will get mutiple datasets given. Scroll down and select *Airbnb Data for Analytics: Trentino Reviews* (If you will not get the data using **Load More** option)



DATA
Airbnb Data for Analytics: Trentino Listings
AUTHOR IBM
MODIFIED Dec 20, 2016
Economy & Business

DATA
Airbnb Data for Analytics: Venice Calendar
Airbnb calendar for Venice, Veneto, Italy. This dataset is sourced from [Inside Airbnb] (http://insideairbnb.com/about.html) which aggregates and cleanses publicly available data from Airbnb for the purpose of supporting

DATA
Airbnb Data for Analytics: Vancouver Reviews
AUTHOR IBM
MODIFIED Dec 20, 2016
Economy & Business

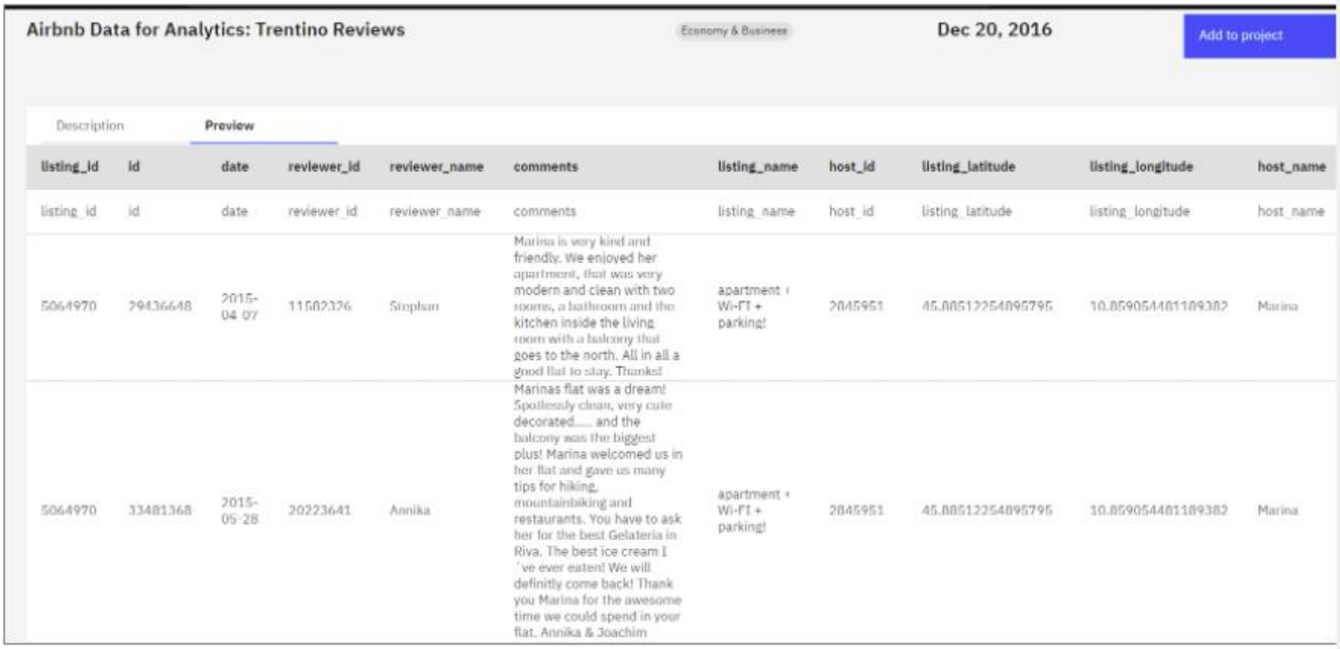
DATA
Airbnb Data for Analytics: Vancouver Listings
AUTHOR IBM
MODIFIED Dec 20, 2016
Economy & Business

DATA
Airbnb Data for Analytics: Vancouver Calendar
AUTHOR IBM
MODIFIED Dec 20, 2016
Economy & Business

DATA
Airbnb Data for Analytics: Trentino Reviews
AUTHOR IBM
MODIFIED Dec 20, 2016
Economy & Business

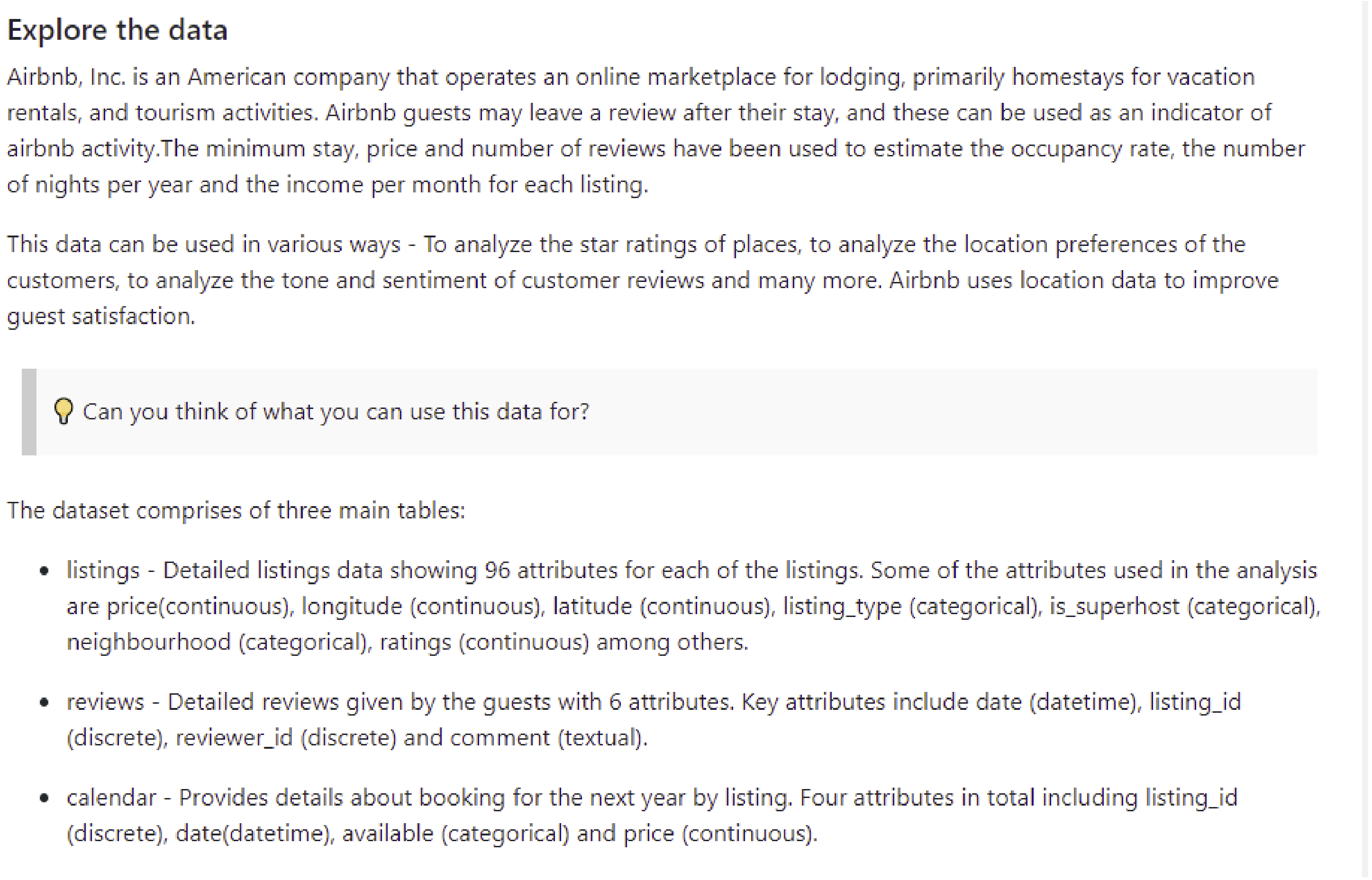
DATA

2. Preview the data using the *Preview* option.



Airbnb Data for Analytics: Trentino Reviews

listing_id	id	date	reviewer_id	reviewer_name	comments	listing_name	host_id	listing_latitude	listing_longitude	host_name
listing_id	id	date	reviewer_id	reviewer_name	comments	listing_name	host_id	listing_latitude	listing_longitude	host_name
5064970	29436648	2015-04-07	11502326	Stephan	Marina is very kind and friendly. We enjoyed her apartment, that was very modern and clean with two rooms, a bathroom and the kitchen inside the living room with a balcony that goes to the north. All in all a good flat to stay. Thanks!	apartment + Wi-FI + parking!	2845951	45.88512254895795	10.859054481189382	Marina
5064970	33481368	2015-05-28	20223641	Annika	Marinas flat was a dream! Spotlessly clean, very cute decorated..... and the balcony was the biggest plus! Marina welcomed us in her flat and gave us many tips for hiking, mountainbiking and restaurants. You have to ask her for the best Gelateria in Riva. The best ice cream I ´ve ever eaten! We will definitly come back! Thank you Marina for the awesome time we could spend in your flat. Annika & Joachim	apartment + Wi-FI + parking!	2845951	45.88512254895795	10.859054481189382	Marina



Explore the data

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Airbnb guests may leave a review after their stay, and these can be used as an indicator of airbnb activity.The minimum stay, price and number of reviews have been used to estimate the occupancy rate, the number of nights per year and the income per month for each listing.

This data can be used in various ways - To analyze the star ratings of places, to analyze the location preferences of the customers, to analyze the tone and sentiment of customer reviews and many more. Airbnb uses location data to improve guest satisfaction.

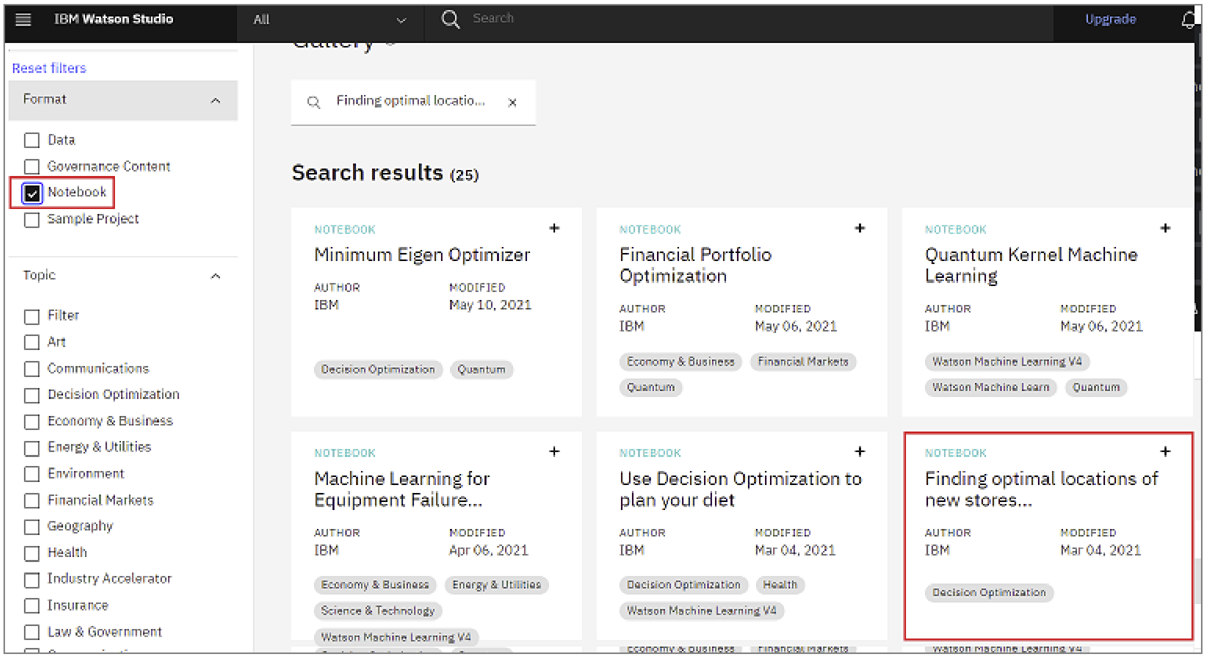
> 💡 Can you think of what you can use this data for?

The dataset comprises of three main tables:

- listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing_type (categorical), is_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.
- reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing_id (discrete), reviewer_id (discrete) and comment (textual).
- calendar - Provides details about booking for the next year by listing. Four attributes in total including listing_id (discrete), date(datetime), available (categorical) and price (continuous).

Exercise 3: Evaluate Jupyter Notebook

Use the *All Filters*. From *Format* select *Notebook* and select *Finding optimal locations of new stores using Decision Optimization*



Finding optimal locatio...

Format

□ Data
□ Governance Content
☑ Notebook
□ Sample Project

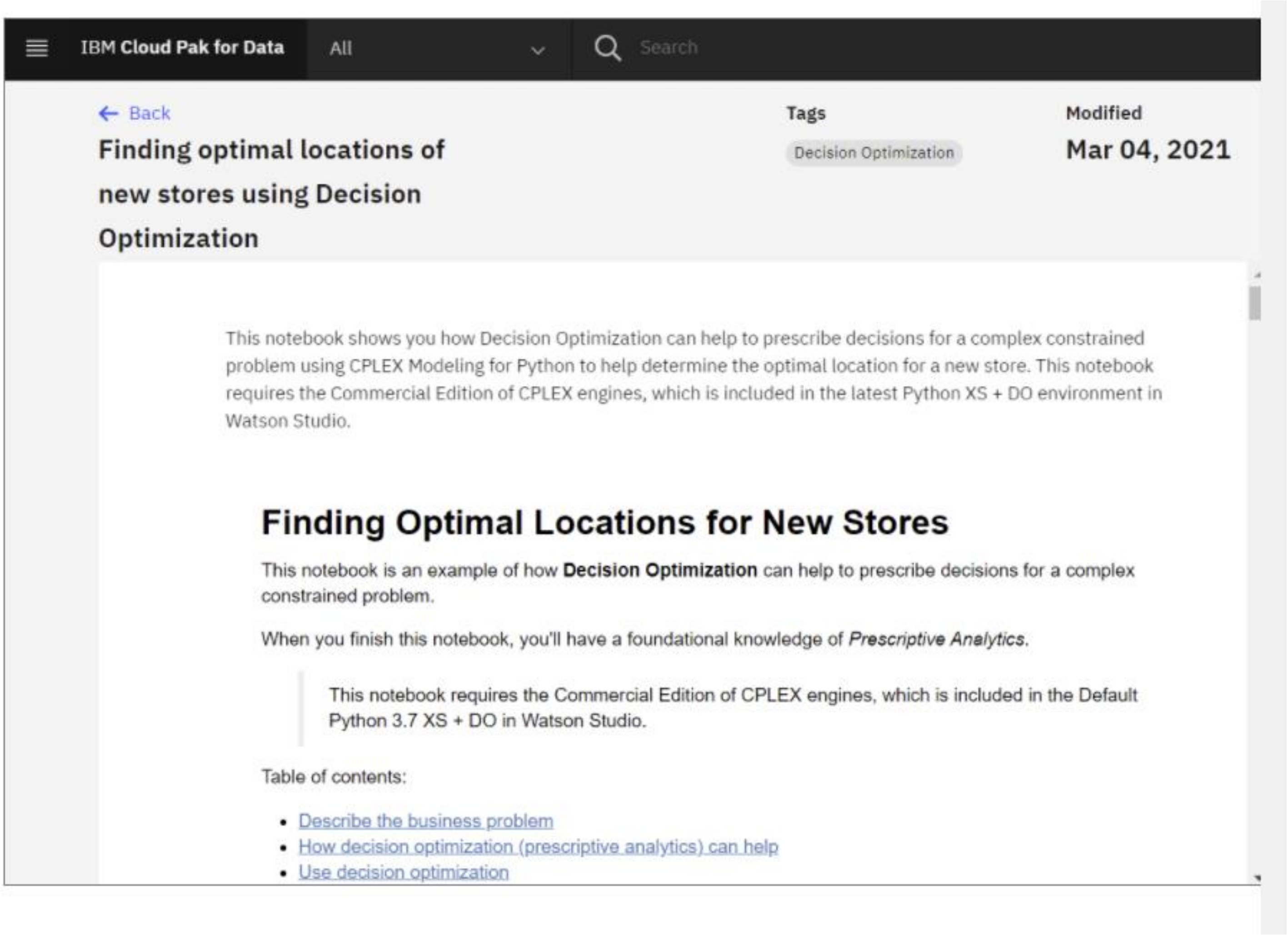
Topic

□ Filter
□ Art
□ Communications
□ Decision Optimization
□ Economy & Business
□ Energy & Utilities
□ Environment
□ Financial Markets
□ Geography
□ Health
□ Industry Accelerator
□ Insurance
□ Law & Government

Search results (25)

This notebook shows you how Decision Optimization can help to prescribe decisions for a complex constrained problem using Python to help determine the optimal location for a new store.

The objective is to minimize the total distance from libraries to coffee shops so that a book reader always gets to our coffee shop easily. It can be done by analyzing and displaying the location of the coffee shops on a map.



Finding optimal locations of new stores using Decision Optimization

Tags

Decision Optimization

Modified

Mar 04, 2021

This notebook shows you how Decision Optimization can help to prescribe decisions for a complex constrained problem using CPLEX Modeling for Python to help determine the optimal location for a new store. This notebook requires the Commercial Edition of CPLEX engines, which is included in the latest Python XS + DO environment in Watson Studio.

Finding Optimal Locations for New Stores

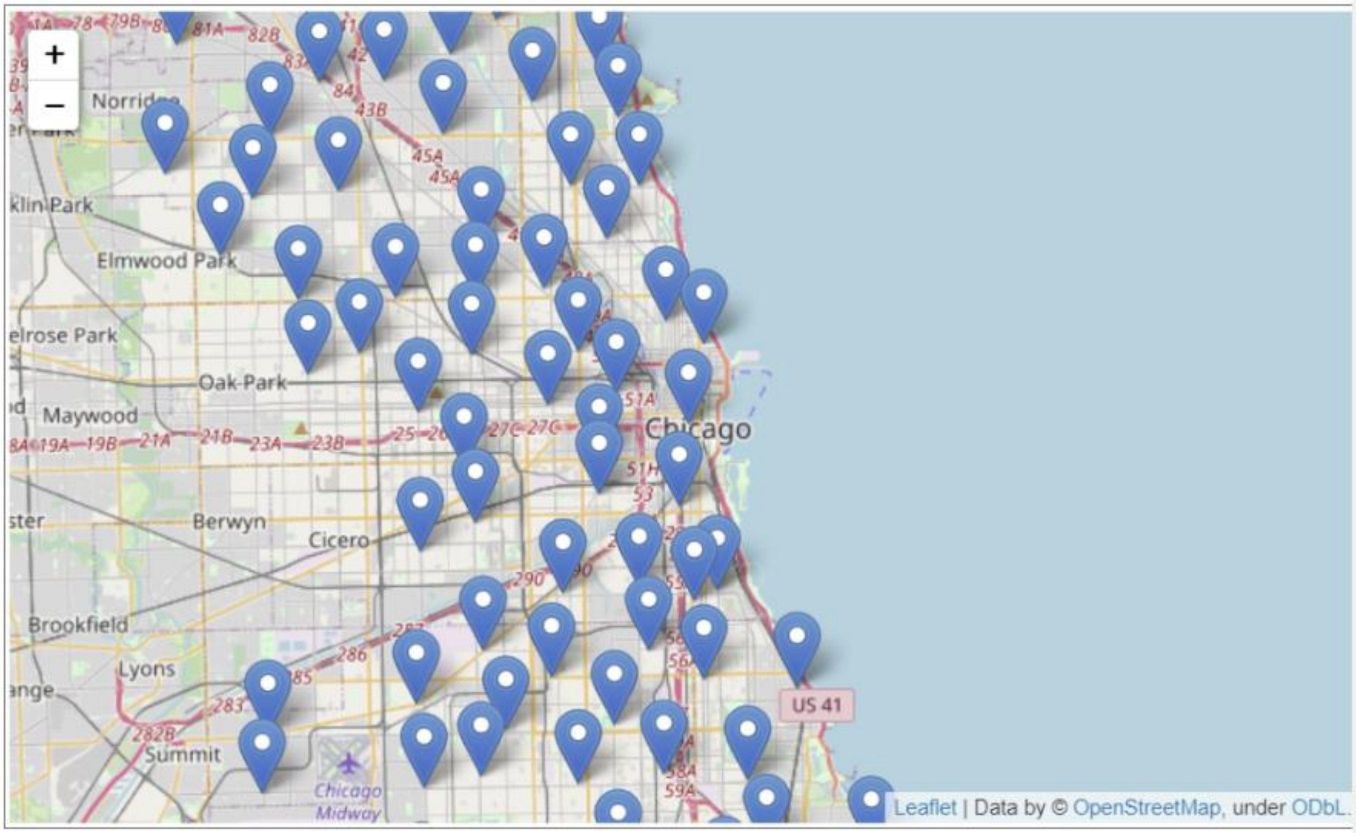
This notebook is an example of how **Decision Optimization** can help to prescribe decisions for a complex constrained problem.

When you finish this notebook, you'll have a foundational knowledge of *Prescriptive Analytics*.

> This notebook requires the Commercial Edition of CPLEX engines, which is included in the Default Python 3.7 XS + DO in Watson Studio.

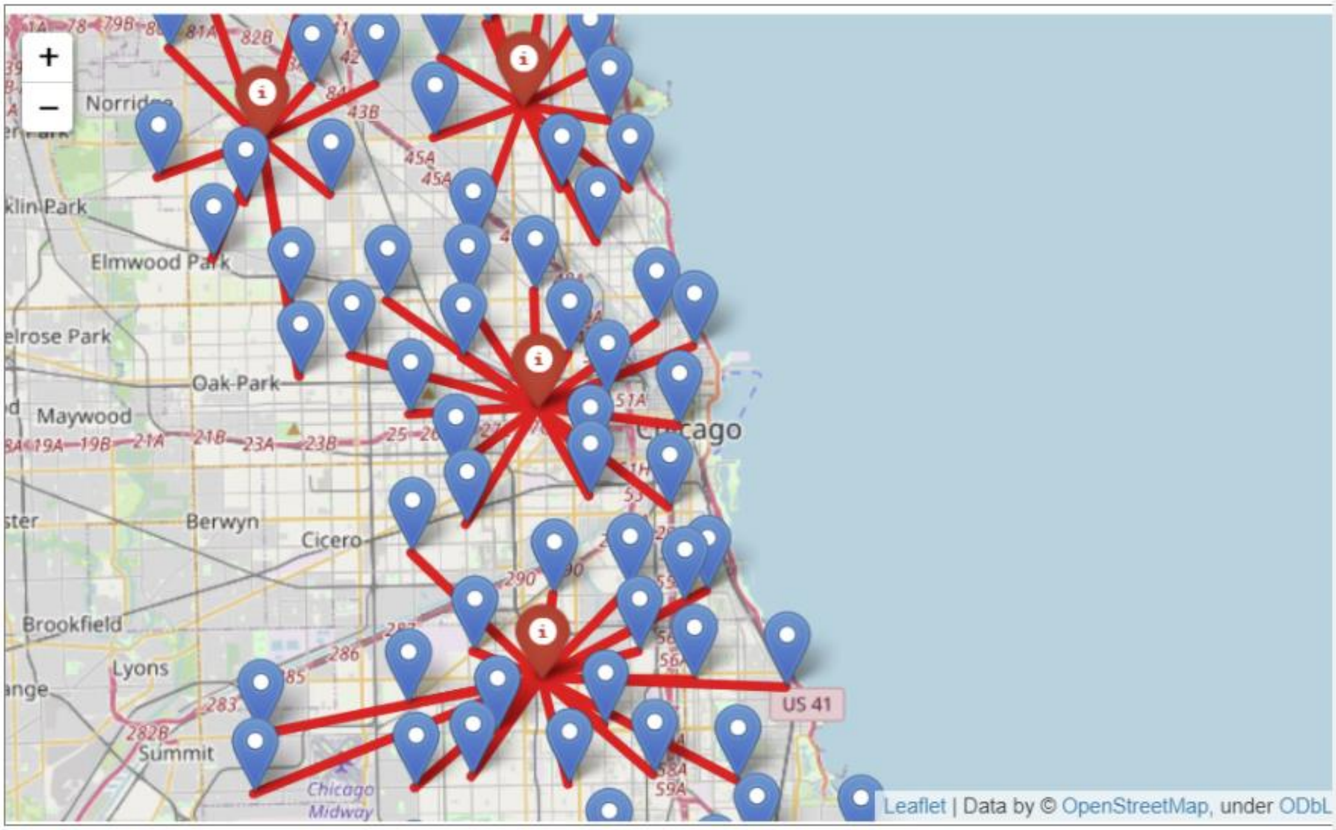
Table of contents:

When we validate the dataset, the locations on map are seperated.



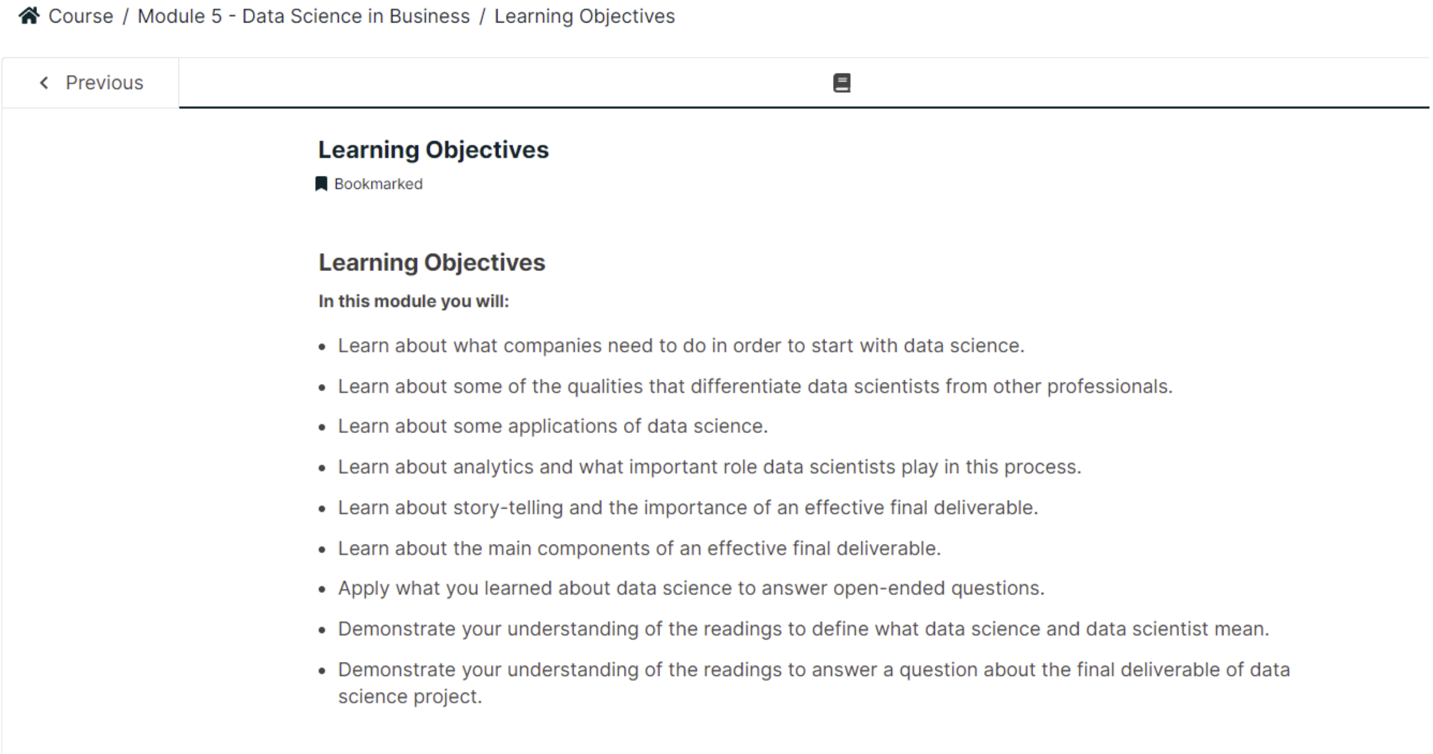
But it is impossible to determine where to ideally open the coffee shops by just looking at the map.

This is solved by an optimization model that will help us determine where to locate the coffee shops in an optimal way.



Summary

In this lab, you have learnt about how different datasets are available and how a data scientist create and predict the models using the model building in IBM Watson Jupyter Notebook.


Course / Module 5 - Data Science in Business / Learning Objectives
Previous
Bookmarked

Learning Objectives

Learning Objectives

In this module you will:

- Learn about what companies need to do in order to start with data science.
- Learn about some of the qualities that differentiate data scientists from other professionals.
- Learn about some applications of data science.
- Learn about analytics and what important role data scientists play in this process.
- Learn about story-telling and the importance of an effective final deliverable.
- Learn about the main components of an effective final deliverable.
- Apply what you learned about data science to answer open-ended questions.
- Demonstrate your understanding of the readings to define what data science and data scientist mean.
- Demonstrate your understanding of the readings to answer a question about the final deliverable of data science project.