Week1: Defining Data Science and What Data Scientists Do

Course Syllabus

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

Defining Data Science and What Data Scientists Do

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

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

Data Science Topics

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

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

Applications and Careers in Data Science:

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

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

Data Literacy for Data Science (Optional):

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

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

Lesson Overview: Defining Data Science

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

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

What is data Science? [video]

- It is the process of using data to understand different things, to understand the world. Data science can extract data from various forms of whether it is unstructured or structured form.
- 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.
- 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.

Fundamental of Data Science [video]

Good data scientists are curious people who ask questions to clarify the business need. 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

Quiz

- Data science is described as a multidisciplinary field that requires a combination of skills, including subject matter expertise, programming, and communication abilities.
- The increasing demand for data scientists and analytics professionals due to the digital revolution and the need to analyze big data for decision-making.

Advice for new Data Scientists

- aspiring data scientist is to be curious, extremely argumentative and judgmental.

Data Science Lesson Summary:

- **What is Data Science?**
  - It's the study of data to understand the world around us.
  - It involves uncovering insights and trends hidden in data.
  - It leverages recent advancements in data access and computing power.
  - It's similar to detective work, uncovering hidden information.
- **Data Science Process:**
  - Involves defining the problem, data collection, analysis, pattern recognition, storytelling, and visualization.

- **Skills of a Data Scientist:**
  - Curiosity to explore data and ask relevant questions.
  - Strong argumentation skills to explain findings and convince others.
  - Good judgment to guide the analysis process.
  - Versatility with programming, statistics, communication, and domain knowledge.
  - Comfort with math and storytelling abilities.
  - Backgrounds can be diverse (economics, engineering, medicine etc.).
- **Learning Path:**
  - Identify your strengths and interests in a particular field.

    - Master data analysis techniques relevant to your chosen field.
    - Select tools commonly used in your industry.
    - Apply your skills to solve real-world problems.
- **The Future of Data Science:**
    - Jobs will evolve with technological advancements and new data roles emerging.
    - Certifications will be increasingly important to demonstrate skills.
    - Core skills like logical thinking, using algorithms, and methodical analysis will remain critical.
    - Data collection accuracy and careful model analysis will be paramount for success.

Quiz: Defining of data science

Questions:

1. Imagine you're working for a retail company that wants to optimize its product offerings and marketing strategies. In this scenario, you would use Data Science for:
2. What is the role of data analysis in Data Science and how does it contribute to decision-making?
3. In a healthcare context with patient data, medical histories, and treatment outcomes, Data Science can be applied to:
4. Considering an individual with a marketing background transitioning to data science, how might their marketing experience contribute to their data science journey?
5. You have just started your career as a data scientist. Which of the following skills should you develop to succeed as a data scientist? You should:

Answers:

1. Analyzing customer purchase data to identify trends and tailor product recommendations.
2. Data analysis involves gathering insights from data and helps make informed decisions.
3. Analyzing patient data for personalized treatment plans.
4. Their marketing background might assist in interpreting data to generate actionable insights.
5. Cultivate curiosity, develop strong positions, and learn to communicate insights effectively through storytelling.

Defining Data Science Lesson Glossary

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

Term	Definition	Video where the term is introduced
Algorithms	A set of step-by-step instructions to solve a problem or complete a task.	What is Data Science?
Model	A representation of the relationships and patterns found in data to make predictions or analyze complex systems retaining essential elements needed for analysis.	What is Data Science?
Outliers	When a data point or points occur significantly outside of most of the other data in a data set, potentially indicating anomalies, errors, or unique phenomena that could impact statistical analysis or modeling.	What is Data Science?
Quantitative analysis	A systematic approach using mathematical and statistical analysis is used to interpret numerical data.	Many Paths to Data Science
Structured data	Data is organized and formatted into a predictable schema, usually related tables with rows and columns.	What is Data Science?
Unstructured data	Unorganized data that lacks a predefined data model or organization makes it harder to analyze using traditional methods. This data type often includes text, images, videos, and other content that doesn't fit neatly into rows and columns like structured data.	What is Data Science?

Quiz

Questions:

1. You are a data scientist about to start a new project. What would one of your key roles be?
2. When did the term "data science" come into existence and who is credited with coining the term?
3. As an aspiring data scientist, what primary qualities should you possess to succeed in the field?

Answers:

1. Asking questions to clarify the business need
2. 2009-2011, DJ Patil or Andrew Gelman
3. Curiosity and storytelling skills.

Lesson2: What do data Scientists Do?

Lesson Overview: What Do Data Scientists Do?

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

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

Understanding different types of file formats

Delimited text files:

- Files used to store data as text.
- Each value is separated by a delimiter.

Delimiter – a sequence of one or more character for specifying the boundary between independent entities or values. **ตัวคั่น (Delimiter)** หมายถึง อักขระหรือลำดับอักขระที่ใช้เพื่อแยกส่วนข้อมูลที่แตกต่างกันออกจากกันในไฟล์ข้อความ มักใช้ในไฟล์ข้อมูลที่ไม่มีโครงสร้างตายตัว เช่น comma, tab, colon, vertical Bar, Space

. TSV and .CSV

TSV ย่อมาจาก Tab Separated Values เป็นรูปแบบไฟล์ที่ใช้สำหรับจัดเก็บข้อมูลแบบตาราง (เช่น ข้อมูลที่สามารถแสดงในรูปแบบตาราง) ไฟล์ TSV คล้ายกับไฟล์ CSV แต่ใช้อักขระแท็บ (\t") เป็นตัวคั่นแทนเครื่องหมายจุลภาค

ตัวอย่างการใช้งานตัวคั่น:

- **ไฟล์ CSV (Comma Separated Values):** ใช้เครื่องหมายจุลภาค (,) เป็นตัวคั่นเพื่อแยกคอลัมน์ในแต่ละแถว ตัวอย่างเช่น:

```
ชื่อ,นามสกุล,อายุ
สมชาย,มาลา,30
หญิง,ทองคำ,25
```

- **ไฟล์ TSV (Tab Separated Values):** ใช้แท็บ (\t) เป็นตัวคั่นเพื่อแยกคอลัมน์ในแต่ละแถว ตัวอย่างเช่น:

```
ชื่อ	นามสกุล	อายุ
สมชาย	มาลา	30
หญิง	ทองคำ	25
```

Microsoft Excel Open XML Spreadsheet (XLSX):

- XML-based spreadsheet format by Microsoft. (**Structured format** with rows and columns)
- Contains multiple worksheets with rows and columns forming cells that hold data.
- Open file format accessible by most applications.
- Secure as it cannot store malicious code.

Extensible Markup Language (XML):

- Human and machine-readable markup language with rules for data encoding. (**Self-descriptive and platform-independent**)
- Self-descriptive for transmitting information online.
- Similar but distinct from HTML (no predefined tags). (**More flexible than HTML for data exchange**)
- Platform and programming language independent, simplifying data sharing between various systems.

 Portable Document Format (PDF):

- Developed by Adobe for document presentation independent of software, hardware, and operating systems. (**Ensures consistent formatting across devices**)
- Ensures consistent viewing across devices (common for legal/financial documents).
- Can also be used for data entry in forms.

 JavaScript Object Notation (JSON):

- Text-based open standard for transmitting structured data over the web. (**Lightweight and easy to use**)
- Language-independent, readable by any programming language.
- Easy to use, compatible with most browsers, and suitable for sharing various data types (including audio/video).
- Popular for APIs and web services to return data. (**Efficient for web communication**)

Data Science Glossary (English-Thai)

Term	Definition (English)	Video
Comma-separated values (CSV) / Tab-separated values (TSV)	Commonly used format for storing tabular data as plain text where either the comma or the tab separates each value.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Data file types	A computer file configuration is designed to store data in a specific way.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Data format	How data is encoded so it can be stored within a data file type.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Data visualization	A visual way, such as a graph, of representing data in a readily understandable way makes it easier to see trends in the data.	หัวข้อและอัลกอริทึมของดาต้าไซเอนส์ (Data Science Topics and Algorithms)
Delimited text file	A plain text file where a specific character separates the data values.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Extensible Markup Language (XML)	A language designed to structure, store, and enable data exchange between various technologies.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Hadoop	An open-source framework designed to store and process large datasets across clusters of computers.	อะไรที่ทำให้ใครสักคนเป็นนักวิทยาศาสตร์ข้อมูล (What Makes Someone a Data Scientist)
JavaScript Object Notation (JSON)	A data format compatible with various programming languages for two applications to exchange structured data.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Jupyter notebooks	A computational environment that allows users to create and share documents containing code, equations, visualizations, and explanatory text. (See Python notebooks)	ทักษะด้านวิทยาศาสตร์ข้อมูลและบิ๊กดาต้า (Data Science Skills & Big Data)

Nearest neighbor	A machine learning algorithm that predicts a target variable based on its similarity to other values in the dataset.	ทำงานกับรูปแบบไฟล์ที่แตกต่างกัน (Working on Different File Formats)
Neural networks	A computational model used in deep learning that mimics the structure and functioning of the human brain's neural pathways. It takes an input, processes it using previous learning, and produces an output.	

Summary: What Do Data Scientists Do?
Congratulations! You have completed this lesson. At this point in the course, you know:

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

Week2

1. **Big Data and Data Mining**

Digital Transformation: Embracing Data Science for Success

Digital transformation is the integration of digital technology into all areas of a business, fundamentally changing how it operates and delivers value to customers. It's driven by data science, especially big data, and the vast amounts of data available.

Example: Houston Rockets and Big Data: In 2018, the NBA's Houston Rockets used big data to improve their game. They analyzed video tracking data to identify the most effective plays for scoring. Surprisingly, the data showed that two-point dunks from within the two-point zone and three-point shots from beyond the arc were more effective than long-range two-point shots. This discovery transformed their approach, leading to an increase in three-point attempts.

Digital Transformation Impacts

- **Process:** In-depth analysis leads to process improvement and integration of data science.
- **Employees:** Changes in processes may affect employee roles and responsibilities.
- **Customers:** Improved services and offerings tailored to customer needs.
- **Culture:** Organizational culture shift towards embracing digital technologies.

Key Players in Digital Transformation

- **Top Executives:** Crucial support from CEO, CIO, and emerging CDO roles.
- **Departmental Executives:** Support from executives controlling budgets, personnel, and priorities.

Digital Transformation: A Necessity for Success

Digital transformation is essential for organizational success in the present and future. While it requires effort, it's a necessary step for businesses to thrive.

Introduction to cloud

บทสรุปเกี่ยวกับ Cloud Computing

Cloud Computing คืออะไร

Cloud Computing คือการนำเสนอบริการคอมพิวเตอร์ตามความต้องการผ่านทางอินเตอร์เน็ต ผู้ใช้สามารถเข้าถึงทรัพยากรต่างๆ เช่น เครื่องคอมพิวเตอร์ พื้นที่จัดเก็บข้อมูล และซอฟต์แวร์ต่างๆ ได้เหมือนมีศูนย์ข้อมูลส่วนตัว โดยจ่ายเฉพาะค่าบริการที่ใช้งานจริง

Cloud Computing มีลักษณะอย่างไร

- บริการตนเองตามความต้องการ (On-demand self-service) : ผู้ใช้สามารถเข้าถึงทรัพยากรต่างๆ บน Cloud ได้เองโดยง่าย ไม่ต้องติดต่อผู้ให้บริการ
- เข้าถึงได้ผ่านเครือข่ายกว้าง (Broad network access) : สามารถเข้าถึงทรัพยากรบน Cloud ได้จากอุปกรณ์ต่างๆ เช่น มือถือ แท็บเล็ต แล็ปท็อป และคอมพิวเตอร์ตั้งโต๊ะ
- การรวมกลุ่มทรัพยากร (Resource pooling) : ผู้ให้บริการ Cloud นำทรัพยากรมาใช้ร่วมกันเพื่อประโยชน์ของผู้ใช้หลายคน ช่วยให้ประหยัดต้นทุน
- ความยืดหยุ่นรวดเร็ว (Rapid elasticity) : ผู้ใช้สามารถเพิ่มหรือลดทรัพยากรตามความต้องการได้อย่างรวดเร็ว
- บริการตามการใช้งานที่วัดได้ (Measured service) : ผู้ใช้จ่ายเฉพาะทรัพยากรที่ใช้งานจริง

Cloud Computing มีรูปแบบการจัดวางระบบอย่างไร

- Public Cloud : ผู้ใช้ใช้บริการ Cloud ร่วมกับผู้อื่นบนโครงสร้างพื้นฐานของผู้ให้บริการ
- Private Cloud : องค์กรมีระบบ Cloud เป็นของตัวเอง ไม่ต้องใช้ร่วมกับผู้อื่น
- Hybrid Cloud : เป็นการผสมผสานระหว่าง Public Cloud และ Private Cloud

Cloud Computing มีรูปแบบบริการอย่างไร

- Infrastructure as a Service (IaaS) : ผู้ใช้สามารถเข้าถึงโครงสร้างพื้นฐาน เช่น เซิร์ฟเวอร์ เครือข่าย พื้นที่จัดเก็บข้อมูล โดยไม่ต้องดูแลระบบเอง
- Platform as a Service (PaaS) : ผู้ใช้สามารถเข้าถึงแพลตฟอร์มสำหรับการพัฒนาและติดตั้งแอปพลิเคชัน
- Software as a Service (SaaS) : ผู้ใช้สามารถใช้ซอฟต์แวร์บน Cloud ได้โดยตรง โดยไม่ต้องติดตั้งบนเครื่อง

สรุป

Cloud Computing เป็นเทคโนโลยีที่ช่วยให้ผู้ใช้เข้าถึงทรัพยากรคอมพิวเตอร์ได้อย่างง่ายดาย ช่วยลดต้นทุน และเพิ่มความคล่องตัว เหมาะสำหรับองค์กรธุรกิจทุกขนาด

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

Course Text Book: 'Getting Started with Data Science' Publisher: IBM Press; 1 edition (Dec 13 2015) Print.
Author: Murtaza Haider
Prescribed Reading: Chapter 12 Pg. 529-531

Establishing Data Mining Goals

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

Selecting Data

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

Preprocessing Data

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

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

Transforming Data

After the relevant attributes of data have been retained, the next step is to determine the appropriate format in which data must be stored. An important consideration in data mining is to reduce the number of attributes needed to explain the phenomena. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and

explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, variables may need to be transformed to help explain the phenomenon being studied. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.
Often you need to transform variables from one type to another. It may be prudent to transform the continuous variable for income into a categorical variable where each record in the database is identified as low, medium, and high-income individual. This could help capture the non-linearities in the underlying behaviors.

Storing Data

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

Mining Data

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

Evaluating Mining Results

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

สรุปภาษาไทย

สรุปการทำเหมืองข้อมูล (Data Mining)

1. กำหนดเป้าหมาย

- ต้องการตอบคำถามอะไร
- คุ้มทุนหรือไม่
- ความแม่นยำของผลลัพธ์

2. เลือกข้อมูล

- ข้อมูลที่มีอยู่แล้ว (เช่น ฐานข้อมูลลูกค้าร้านค้า)
- ต้องหาแหล่งข้อมูลเพิ่มเติม หรือเก็บข้อมูลใหม่

3. ปรับแต่งข้อมูล

- ทำความสะอาดข้อมูล (กำจัดข้อมูลผิดพลาด ไม่เกี่ยวข้อง)
- ตรวจสอบความถูกต้องของข้อมูล
- หาแนวทางจัดการกับข้อมูลที่หายไป

4. เปลี่ยนรูปแบบข้อมูล

- ลดจำนวนตัวแปร (โดยไม่สูญเสียข้อมูลสำคัญ)
- เปลี่ยนรูปแบบข้อมูล (เช่น รายได้จากหลายแหล่ง เป็นรายได้รวม)
- แปลงข้อมูลเป็นประเภท (เช่น รายได้น้อย ปานกลาง มาก)

5. จัดเก็บข้อมูล

- จัดเก็บข้อมูลให้นักวิทยาศาสตร์ข้อมูลเข้าถึงได้ง่าย
- เก็บข้อมูลอย่างปลอดภัย

6. วิเคราะห์ข้อมูล

- ใช้เครื่องมือค้นหาแนวโน้มในข้อมูล
- ใช้โมเดลต่างๆ ในการวิเคราะห์ข้อมูล

7. ประเมินผลลัพธ์

- ประเมินประสิทธิภาพของโมเดล
- นำเสนอผลลัพธ์ให้ผู้เกี่ยวข้อง
- ปรับปรุงกระบวนการวิเคราะห์ข้อมูล

Big Data and Data Mining: A Summary

This video explored the impact of big data on various aspects of society, from business to sports, and reviewed the key characteristics and challenges associated with big data. Here's a recap of the key points:

Big Data: Transforming Our World

- The vast amount of data available, known as big data, is driving significant changes in businesses, industries, and daily life.
- Organizations need fundamental changes in their approach to handle this data effectively.
- Big data allows us to gain real-time insights related to consumers, risks, profits, and performance, ultimately enhancing business value.

Key Characteristics of Big Data

There's no universally agreed-upon definition, but big data generally exhibits five key characteristics:

1. **Value:** The data must hold potential value for analysis.
2. **Volume:** The sheer size of the data is immense. Factors like increasing data sources and scalable infrastructure contribute to this volume.
3. **Velocity:** Data is generated continuously at an ever-increasing rate.
4. **Variety:** Data comes from diverse sources, including structured and unstructured formats.
5. **Veracity:** The data must be accurate and reliable, reflecting reality.

Cloud Computing: A Powerful Tool for Big Data

Cloud computing, which delivers computing resources on-demand, offers benefits for big data analysis:

- **Essential Characteristics:** Cloud computing offers five key features:
    - On-demand access to processing power, storage, and network resources.
    - Network access via the internet.
    - Resource pooling allows efficient sharing of resources among users.

    - Elasticity to scale resources up or down as needed.
    - Measured service – users only pay for what they use.
- **Benefits for Big Data:** Cloud computing addresses challenges like scalability, collaboration, accessibility, and software maintenance, making it valuable for data analysis.
    - Provides instant access to technologies without installation or configuration.
    - Offers automatic updates to software tools.

Popular Open-Source Big Data Tools

Several open-source tools help process big data:

- **Apache Hadoop:** Provides distributed storage and processing across computer clusters.
- **Apache Hive:** A data warehouse built on top of Hadoop, allowing data query and analysis.
- **Apache Spark:** A general-purpose processing engine for large datasets across various applications.

The Data Mining Process: Extracting Knowledge from Big Data

Data mining is a six-step process:

1. **Goal Setting:** Identify key questions to answer, considering costs and benefits.
2. **Data Selection:** Choose data sources or plan data collection initiatives.
3. **Preprocessing:** Clean the data by removing irrelevant or erroneous information.
4. **Transformation:** Determine the appropriate format for storing the data.
5. **Mining:** Analyze the data using machine learning algorithms.
6. **Evaluation:** Assess the effectiveness of the analysis and share results with stakeholders. This is an iterative process where findings inform future efforts.

In Conclusion

Big data, with its unique characteristics, is driving significant change across industries. Cloud computing provides the power and resources to handle big data effectively, and

open-source tools like Hadoop, Hive, and Spark facilitate efficient data mining. This summary provides a high-level overview. For a deeper understanding, refer to the full video or additional resources.

Big Data and Data Mining Lesson Glossary

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

Term	Definition	Video where the term is introduced
Analytics	The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights.	Data Scientists at New York University
Big Data	Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations.	How Big Data is Driving Digital Transformation
Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.	What is Hadoop?
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.	Introduction to Cloud

Term	Definition	Video where the term is introduced
Chief Data Officer (CDO)	An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts.	How Big Data is Driving Digital Transformation
Chief Information Officer (CIO)	An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation.	How Big Data is Driving Digital Transformation
Cloud Computing	The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis.	Introduction to Cloud
Cloud Deployment Models	Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available to users, including public, private, and hybrid models.	Introduction to Cloud
Cloud Service Models	Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings.	Introduction to Cloud
Commodity Hardware	Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware.	What is Hadoop?
Data Algorithms	Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently.	Cloud for Data Science

Term	Definition	Video where the term is introduced
Data Replication	A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures.	What is Hadoop?
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.	Data Scientists at New York University
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own.	Data Scientists at New York University
Digital Change	The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers.	How Big Data is Driving Digital Transformation
Digital Transformation	A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes.	How Big Data is Driving Digital Transformation
Distributed Data	The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis.	What is Hadoop?
Hadoop	A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications.	Data Scientists at New York University

Term	Definition	Video where the term is introduced
Hadoop Distributed File System (HDFS)	A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance.	What is Hadoop?
Infrastructure as a Service (IaaS)	A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them.	Introduction to Cloud
Java-Based Framework	Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
Map Process	The initial step in Hadoop’ s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks.	What is Hadoop?
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.	Introduction to Cloud
On-Demand Self-Service	The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers.	Introduction to Cloud
Rapid Elasticity	The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use.	Introduction to Cloud

Term	Definition	Video where the term is introduced
Reduce Process	The second step in Hadoop's MapReduce model is where results from the mapping process are aggregated and processed further to produce the final output, typically used for analysis.	What is Hadoop?
Replication	The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures.	What is Hadoop?
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.	Introduction to Cloud
Skills Network Labs (SN Labs)	Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development.	Cloud for Data Science
Spilling to Disk	A technique used in memory-constrained situations where data is temporarily written to disk storage when memory resources are exhausted, ensuring uninterrupted processing.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
STEM Classes	Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science.	Data Scientists at New York University
Variety	The diversity of data types, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges.	Foundations of Big Data

Term	Definition	Video where the term is introduced
Velocity	The speed at which data accumulates and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics.	Foundations of Big Data
Veracity	The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness.	Foundations of Big Data
Video Tracking System	A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports.	How Big Data is Driving Digital Transformation
Volume	The scale of data generated and stored is driven by increased data sources, higher-resolution sensors, and scalable infrastructure.	Foundations of Big Data
V's of Big Data	A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid generation, scale, diversity, quality, and value of data.	Foundations of Big Data

Week2 Part2

Lesson Overview: Deep Learning and Machine Learning

In this lesson, "Deep Learning and Machine Learning," you'll dive into the exciting concepts of artificial intelligence and data science. Throughout this module, you will explore various machine learning and deep learning aspects, gaining valuable insights and skills.

Asset name and type	Description
"Artificial Intelligence and Data Science" video	Get introduced to the captivating field of artificial intelligence and its role in data science.
"Generative AI and Data Science" video	Discover the exciting realm of generative artificial intelligence and its applications in data science.
"Neural Networks and Deep Learning" video	Explore the fundamentals of neural networks and delve into the depths of deep learning.
"Applications of Machine Learning" video	Uncover the real-world applications of machine learning and its impact on various industries.
"Regression" reading	Learn about regression analysis, a fundamental statistical technique used in machine learning.
Practice quiz	Test your understanding of the previous reading.
"Exploring Data using IBM Cloud Gallery" lab	Engage in hands-on exploration of data using the IBM Cloud Gallery, gaining practical experience in data analysis.
"Lesson Summary" video	Sum up your learning from this module with a concise lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you've understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.
Graded quiz	Test your knowledge from this lesson by taking the graded quiz.

This passage clarifies key terms in data science and artificial intelligence (AI).

- **Big Data:** Refers to massive, fast-growing, and diverse datasets that traditional data analysis methods cannot handle. It's often described by five characteristics: velocity (speed of data generation), volume (amount of data), variety (different data formats), veracity (data accuracy), and value (potential usefulness of data).
- **Data Mining:** The process of automatically analyzing data to discover hidden patterns. It involves preparing the data, transforming it into a usable format, and then extracting insights using various techniques like data visualization, machine learning, and statistical models.
- **Machine Learning:** A subset of AI that uses algorithms to analyze data and learn from it without explicit programming. Machine learning algorithms are trained on large datasets and make predictions based on what they learn.
- **Deep Learning:** A specialized form of machine learning that uses layered artificial neural networks to mimic human decision-making. Deep learning excels at labeling,

categorizing information, and identifying patterns. It allows AI systems to continuously learn and improve results.

- **Artificial Neural Networks:** Inspired by biological neural networks, they are collections of interconnected processing units (neurons) that learn to make decisions over time. Deep learning algorithms leverage these networks and become more efficient with larger datasets compared to other machine learning methods.
- **Data Science:** An interdisciplinary field that uses mathematics, statistics, data visualization, and machine learning techniques to extract knowledge and insights from large, diverse datasets. It helps us understand data, find patterns, and make data-driven decisions. Data science can utilize various AI techniques, including machine learning and deep learning, to analyze data.

Key takeaway: AI and data science are distinct fields. Data science focuses on extracting knowledge from data, while AI encompasses anything that enables computers to learn and make intelligent decisions. Both can leverage big data for their tasks.

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?

Deep Learning and Machine Learning: A Summary
This video recaps key concepts from the Deep Learning and Machine Learning lesson.
Key Terms in AI:

- **Artificial Intelligence (AI):** A branch of computer science focused on creating intelligent systems that mimic human capabilities.

- **Machine Learning (ML):** A subset of AI that uses algorithms to learn from data and make predictions without explicit programming.
- **Deep Learning (DL):** A subset of ML that utilizes layered neural networks inspired by the human brain for more complex decision-making.
- **Neural Networks:** Networks of interconnected processing units (neurons) that learn from data over time.
- **Generative AI:** A branch of AI focused on creating new data (images, music, code) that resembles existing data.

Data Science and Machine Learning:

- Data scientists leverage AI, particularly machine learning, to analyze data and extract insights.
- Machine learning algorithms are often used for:
    - Predictive analytics: Making predictions based on trends in data (e.g., fraud detection).
    - Recommendation systems: Suggesting products or services based on user preferences.

Regression Models:

- Regression is a statistical technique used in machine learning to identify the relationship between two or more variables.
- For example, a regression model could analyze how house price correlates with square footage and bedrooms.

In Conclusion:

- Generative AI creates new data, while deep learning and machine learning analyze existing data.
- Deep learning utilizes neural networks for complex pattern recognition.
- Data scientists use all these AI areas, powered by big data, to make data-driven predictions.

Deep Learning and Machine Learning Lesson Glossary

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

Term	Definition	Video where the term is introduced
Artificial Neural Networks	Collections of small computing units (neurons) that process data and learn to make decisions over time.	Artificial Intelligence and Data Science
Bayesian Analysis	A statistical technique that uses Bayes' theorem to update probabilities based on new evidence.	Applications of Machine Learning

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

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

Summary: Deep Learning and Machine Learning

Congratulations! You have completed this lesson. At this point in the course, you know:

- Big Data has five characteristics: velocity, volume, variety, veracity, and value.
- The five cloud computing characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.
- Data mining has a six-step process: goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.
- The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology to drive transformation.
- Deep learning utilizes neural networks to teach itself patterns in inputs and outputs. Machine learning is a subset of AI that uses computer algorithms to learn about data and make predictions without explicitly programming the analysis methods into the system.
- Regression identifies the strength and amount of the correlation between one or more inputs and an output.
- Skills involved in processing Big Data include the application of statistics, machine learning models, and some computer programming.
- Generative AI, a subset of artificial intelligence, focuses on producing new data rather than just analyzing existing data. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.

Week 3: Application and careers in Data Science
Part 1: Data Science Application Domains

Lesson Overview: Data Science Application Domains
In this lesson, "Data Science Application Domains," you'll embark on a journey to explore the vast and impactful realms where data science plays a pivotal role. This engaging module delves into various activities that shed light on the diverse applications of data science in our world today. You'll uncover how data science drives innovation and transformation across different sectors, from revolutionizing industries to saving lives. Dive into this lesson to discover the real-world applications that define the dynamic landscape of data science.

Asset name and type	Description
“How Should Companies Get Started in Data Science?” video	Gain insights into how organizations can embark on their data science journey effectively.
“Old problems with New Data Science Solutions” video	Discover how data science offers innovative solutions to age-old and real-world problems.
“Applications of Data Science” video	Explore the wide-ranging applications of data science across various industries and sectors.
“How Data Science is saving lives” video	In this video, you will learn about the life-saving potential of data science in healthcare and beyond.
“The Final Deliverable” reading	Dive into the details of what constitutes the final deliverable in data science projects.
Practice quiz	Test your understanding of the previous reading.
“Lesson Summary” video	Recap the essential takeaways from this module with a lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.

The Final Deliverable [book]

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. The data scientist should then use the insights to build the narrative to communicate

the findings. In academia, the final deliverable is in the form of essays and reports. Such deliverables are usually 1,000 to 7,000 words in length.

In consulting and business, the final deliverable takes on several forms. It can be a small document of fewer than 1,500 words illustrated with tables and plots, or it could be a comprehensive document comprising several hundred pages. Large consulting firms, such as McKinsey and Deloitte, I routinely generate analytics-driven reports to communicate their findings and, in the process, establish their expertise in specific knowledge domains.

Let's review the "United States Economic Forecast", a publication by the Deloitte University Press. This document serves as a good example for a deliverable that builds narrative from data and analytics. The 24-page report focuses on the state of the U.S. economy as observed in December 2014. The report opens with a **grabber** highlighting the fact that contrary to popular perception, the economic and job growth has been quite robust in the United States. The report is not merely a statement of facts.
In fact, it is a carefully crafted report that cites Voltaire and follows a distinct theme. The report focuses on the **good news** about the U.S. economy. These include the increased investment in manufacturing equipment in the U.S. and the likelihood of higher consumer consumption resulting from lower oil prices.
The Deloitte report uses time series plots to illustrate trends in markets. The GDP growth chart shows how the economy contracted during the Great Recession and has rebounded since then. The graphic presents four likely scenarios for the future. Another plot shows the changes in consumer spending. The accompanying narrative focuses on income inequality in the U.S. and refers to Thomas Pikkety's book on the same. The Deloitte report mentions many consumers did not experience an increase in their real incomes over the years, while they still maintained their level of spending. Other graphics focused on housing, business, and government sectors, international trade, labor, and financial markets, and prices. The appendix carries four tables documenting data for the four scenarios discussed in the report.

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

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

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

Data Science Application Domains Lesson Glossary

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

Term	Definition	Video where the term is introduced
Arithmetic Models	Data science often uses Mathematical models to analyze data and predict outcomes.	Old problems, new data science solutions
Case study	In-depth analysis of an instance of a chosen subject to draw insights that inform theory, practice, or decision-making.	Old problems, new data science solutions

Term	Definition	Video where the term is introduced
Data mining	Extracting information from raw data, such as making decisions, predicting trends, or understanding phenomena.	How Data Science is Saving Lives
Data Science	The field involves collecting, analyzing, and interpreting data to extract valuable insights and make informed decisions.	Old problems, new data science solutions
Data Strategy	A plan that outlines how an organization will collect, manage, and use data to achieve its goals.	Old problems, new data science solutions
Predictive analytics	Using data, algorithms, models, and machine learning to make predictions.	How Data Science is Saving Lives

Part2 Careers and Recruiting in Data Science

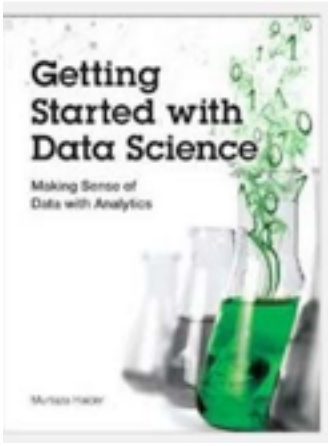
In the lesson "Data Science Application Domains," you'll embark on a journey to explore the vast and impactful realms where data science plays a pivotal role. This engaging module delves into various activities that shed light on the diverse applications of data science in our world today.

Dive into this lesson to discover the real-world applications that define the dynamic landscape of data science.

Asset name and type	Description
"How Can Someone Become a Data Scientist?" video	Discover the pathways to becoming a proficient data scientist, exploring the skills and knowledge required.
"Recruiting for Data Science" video	Gain insights into organizations' strategies and considerations for recruiting data science talent.
"Careers in Data Science" video	Explore the diverse career opportunities and roles available in the dynamic field of data science.
"Importance of Mathematics and Statistics for Data Science" video	Understand the fundamental role of mathematics and statistics in data science, emphasizing their significance.
"The Report Structure" reading	Delve into the intricacies of structuring reports within data science projects, enhancing your understanding of this essential aspect.
Practice quiz	Test your understanding of the previous reading.
"Infograph on Roadmap" reading	Explore an informative infographic detailing the roadmap for success in data science careers.

Asset name and type	Description
“Lesson Summary” video	Recap the essential takeaways from this module with a lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.
Grade Quiz	Evaluate your knowledge of data science in a business setting with this graded quiz.
“Data Science in Business” Reading	Summarize your learning journey in this module, reviewing the key takeaways and insights gained.

Course Text Book: ‘Getting Started with Data Science’ Publisher: IBM Press; 1 edition (Dec 13 2015) Print.
Author: Murtaza Haider



Prescribed Reading: Chapter 3 Pg. 60-62

The Report Structure

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

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

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

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

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

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

Even for a short document, I recommend an "abstract" or an "executive summary". Nothing is more powerful than explaining the crux of your arguments in three paragraphs or less. Of course, for larger documents running a few hundred pages, the executive summary could be longer.

An "introductory section" is always helpful in setting up the problem for the reader who might be new to the topic and who might need to be gently introduced to the subject matter before being immersed in intricate details. A good follow-up to the introductory section is a review of available relevant research on the subject matter. The length of the literature review section depends upon how contested the subject matter is. In instances where the vast majority of researchers have concluded in one direction, the literature review could be brief with citations for only the most influential authors on the subject. On the other hand, if the arguments are more nuanced with caveats aplenty, then you must cite the relevant research to offer adequate context before you embark on your analysis. You might use the literature review to highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

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

The results section is where you present your empirical findings. Starting with descriptive statistics (**see Chapter 4, "Serving Tables"**) and illustrative graphics (**see Chapter S, "Graphic Details" for plots and Chapter 10, "Spatial Data Analytics" for maps**), you will move toward formally testing your hypothesis (**see Chapter 6, "Hypothetically Speaking"**). In case you need to run statistical models, you might turn to regression models (**see Chapter 7, "Why Tall Parents Don't Have Even Taller Children"**) or categorical analysis (**see Chapters 8, "To Be or Not to Be" and 2., "Categorically Speaking About Categorical Data"**). If you are working with time-series data, you can turn to Chapter 11, **Doing Serious Time with Time Series.** You can also report results from other empirical techniques that fall under the general rubric of data mining (**see Chapter 12, "Data Mining for Gold"**). Note that many reports in the business sector present results in a more palatable fashion by holding back the statistical details and relying on illustrative graphics to summarize the results.
The results section is followed by the discussion section, where you craft your main arguments by building on the results you have presented earlier.

The "discussion section" is where you rely on the power of narrative to enable numbers to communicate your thesis to your readers. You refer the reader to the research question and

the knowledge gaps you identified earlier. You highlight how your findings provide the ultimate missing piece to the puzzle.

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

In the "conclusion" section, you generalize your specific findings and take on a rather marketing approach to promote your findings so that the reader does not remain stuck in the caveats that you have voluntarily outlined earlier. You might also identify future possible developments in research and applications that could result from your research.

What remains is housekeeping, including a list of references, the acknowledgment section (**acknowledging the support of those who have enabled your work is always good**), and "appendices", if needed.

Have You Done Your Job as a Writer?

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

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

- Have you structured your paper in a clear and logical fashion?

Careers and Recruiting in Data Science Lesson Glossary

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

Term	Definition	Video where the term is introduced
Adobe Spark	A suite of software tools that allow users to create and share visual content such as graphics, web pages, and videos.	Recruiting for Data Science
Analytical skills	The ability to analyze information systematically, logically, and organized.	Recruiting for Data Science
Chief information officer (CIO)	A business executive is responsible for an organization’ s information technology systems and tech-related initiatives.	How Can Someone Become a Data Scientist
Computational thinking	Breaking problems into smaller parts and using algorithms, logic, and abstraction to develop solutions. Often used but not limited to computer science.	How Can Someone Become a Data Scientist
Data clusters	A group of similar, related data points distinct from other clusters.	How Can Someone Become a Data Scientist

Term	Definition	Video where the term is introduced
Executive summary	Usually occurring at the beginning of a research paper, this section summarizes the important parts of the paper, including its key findings.	The Report Structure
High-performing computing (HPC) cluster	A computing technology that uses a system of networked computers designed to solve complex and computationally intensive problems in traditional environments.	How Can Someone Become a Data Scientist
Mathematical computing	The use of computers to calculate, simulate, and model mathematical problems.	Importance of Mathematics and Statistics for Data Science
Matrices	Plural for matric, matrices are a rectangular (tabular) array of numbers often used in mathematics, statistics, and computer science.	Recruiting for Data Science
Stata	A software package used for statistical analysis.	Recruiting for Data Science
Statistical distributions	A way of describing the likelihood of different outcomes based on a dataset. The "bell curve" is a common statistical distribution.	How Can Someone Become a Data Scientist
Structured Query Language (SQL)	A language used for managing data in a relational database.	Importance of Mathematics and Statistics for Data Science
TCP/IP network	A network that uses the TCP/IP protocol to communicate between connected devices on that network. The Internet uses TCP/IP.	How Can Someone Become a Data Scientist

Summary: Careers and Recruiting in Data Science

Congratulations! You have completed this module. At this point, you know that:

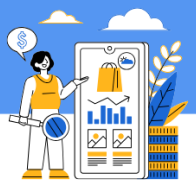
- Data Science helps physicians provide the best treatment for their patients , helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- Companies can start on their data science journey by capturing data. Once they have data, they can begin analyzing it.
- Everyone who uses the Internet generates mass amounts of data daily.
- Amazon and Netflix use recommendation engines, and UPS uses data from customers, drivers, and vehicles to use the drivers' time and fuel efficiently.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
- The report should present a thorough analysis of the data and communicate the project findings.
- Companies should look for someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.
- A clearly organized and logical report should communicate the following to the reader:
    - What they gain by reading the report
    - Clearly defined goals
    - The significance of your contribution
    - Appropriate context by giving sufficient background
    - Why this work is practical and useful
    - Conjecture plausible future developments that might result from your work

Data Science

A roadmap to your Data Science journey

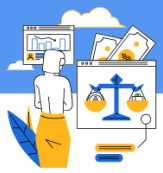
Personality Characteristics

Curiosity is key
Make sound arguments
Use good judgement
Familiarize with analytics platforms
Storyteller
Know your area of interests (such as healthcare or IT)



Many Paths

Diverse educational and career backgrounds
Exposure to data challenges sparked interest
Data science is adaptable across professions

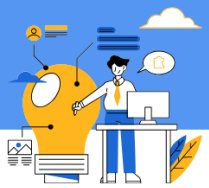


Data Literacy

Analyze both structured and unstructured data
Understand file formats
Database and SQL skills
Big Data, Cloud

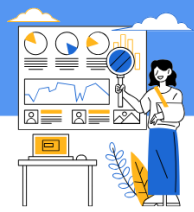
Tools & Techniques

Programming with Python and R
Hadoop
Python libraries: NumPy, pandas, scikit-learn
Data visualization tools
Machine learning algorithms
Data preprocessing techniques



Foundational Skills

Statistical knowledge. Mathematics, Calculus, Linear Algebra
Exploratory data analysis
Select, train, and test models
Communication and presentation skills



Range of tasks

Build Recommendation Engines
Predictive Modeling
Data Analysis and Problem Solving
Identify Patterns
Utilize External Data Sources
Communication of Findings

IBM

Skills Network

Course Summary

Congratulations! You have completed this course. At this point, you know that:

- Data science is the practice of extracting valuable insights from vast datasets to guide strategic decision-making.
- Data science careers offer diverse paths, often involving mathematics, programming, and a curiosity for data exploration.
- Successful data scientists exhibit qualities like curiosity, critical judgment, and an aptitude for constructive argumentation.
- The data science field is characterized by high demand, resulting in attractive remuneration for skilled professionals.
- A Data Scientist's daily routine can vary significantly depending on the project's nature.
- A wide array of algorithms is available for extracting insights from data.
- Big Data plays a pivotal role in driving digital transformation across industries.
- Cloud computing is a fundamental technology in modern data science.
- Data mining techniques are essential for uncovering patterns and knowledge from data.
- Tools like Hadoop, HDFS, Hive, and Spark are employed for processing Big Data.
- Deep learning, machine learning, and regression are critical data science topics.
- Data science applications span diverse domains, solving complex problems.
- Companies can harness data science to address age-old challenges with innovative solutions.
- Data science contributes significantly to saving lives and improving various aspects of society.
- Careers in data science offer exciting opportunities, with mathematics and statistics being essential foundations.
- Reports in data science adhere to specific structures, and career roadmaps provide guidance.
- Case studies and projects offered practical application of the knowledge acquired during the course.

Week4: Understanding data and literacy data.
Lesson Overview: Understanding Data

In this lesson, "Understanding Data," you'll explore data basics through videos on data comprehension and sources. The reading covers metadata's role, and quizzes reinforce your understanding. The lesson ends with a summary video, ensuring you grasp essential data concepts.

Asset name and type	**Description**
“Understanding Data:” video	Gain foundational insights into data comprehension. The video will also help you explore data types, characteristics, and their importance in various fields of study.
“Data Sources” video	Explore the diverse origins of data in this video, uncovering where and how data is generated, collected, and utilized in different contexts.
“Working on Varied Data Sources and Types” video	This video equips you with the skills to effectively manage and analyze data from a wide range of sources and in various formats, ensuring adaptability in data handling.
“Metadata” reading	Read an excerpt on “Metadata” and discover the significance of metadata in data analysis through this reading, which encompasses three primary metadata types: technical, process, and business metadata.
Practice quiz	Test your knowledge of metadata concepts based on the previous reading.
“Lesson Summary” video	Recap the key points from this lesson.
Practice quiz	Take a practice quiz to assess your comprehension of data fundamentals.

ข้ามมา เป็น NoSQL

NoSQL

NoSQL คืออะไร?

NoSQL (Not Only SQL) คือประเภทของระบบจัดการฐานข้อมูลที่ออกแบบมาเพื่อรองรับการเก็บข้อมูลที่ไม่เป็นเชิงสัมพันธ์ (non-relational) โดยมีความยืดหยุ่นในการจัดเก็บข้อมูลมากกว่าฐานข้อมูลเชิงสัมพันธ์ (relational databases) เช่น SQL เนื่องจาก NoSQL สามารถรองรับข้อมูลที่หลากหลายรูปแบบและมีความสามารถในการปรับตัวตามปริมาณข้อมูลที่เพิ่มขึ้นอย่างรวดเร็ว

ประเภทของ NoSQL

1. Document Stores: จัดเก็บข้อมูลในรูปแบบของเอกสาร (documents) ที่มักใช้ JSON, BSON หรือ XML เป็นรูปแบบในการเก็บข้อมูล
    - ตัวอย่าง: MongoDB, CouchDB
2. Key-Value Stores: จัดเก็บข้อมูลในรูปแบบของคู่ค่า (key-value pairs) โดยที่แต่ละคีย์จะมีค่าเชื่อมโยงกับมัน
    - ตัวอย่าง: Redis, DynamoDB
3. Column-Family Stores: จัดเก็บข้อมูลในรูปแบบของตาราง แต่ละแถวในตารางจะมีคอลัมน์ที่ต่างกันได้
    - ตัวอย่าง: Cassandra, HBase
4. Graph Databases: จัดเก็บข้อมูลในรูปแบบของกราฟ (graph) ซึ่งโหนด (nodes) และขอบ (edges) จะเชื่อมโยงกันเพื่อแสดงถึงความสัมพันธ์
    - ตัวอย่าง: Neo4j, OrientDB

ข้อดีของ NoSQL (แบบสั้น ๆ)

1. ยืดหยุ่นสูง: เก็บข้อมูลได้ทั้งแบบไม่มีโครงสร้างและกึ่งโครงสร้าง
2. ขยายระบบง่าย: รองรับการขยายระบบในแนวนอน (เพิ่มเซิร์ฟเวอร์)
3. ประมวลผลเร็ว: ไม่ต้องใช้การสืบค้นเชิงสัมพันธ์
4. เก็บข้อมูลหลายประเภท: เช่น ข้อมูลเอกสาร, คู่คีย์-ค่า, ข้อมูลกราฟ
5. ทนทานต่อความล้มเหลว: มีการสำรองและกระจายข้อมูล
6. ประมวลผลแบบกระจาย: ทำงานบนหลายเซิร์ฟเวอร์พร้อมกัน
7. ใช้งานง่าย: ไม่ต้องออกแบบ schema ซับซ้อน
8. รองรับ Real-Time: เหมาะสำหรับการประมวลผลข้อมูลแบบเรียลไทม์