Module-1
Defining Data Science and What Data Scientists Do

- ➢ In Module 1, you delve into ***some fundamentals of Data Science***.
- ➢ In lesson 1, you listen to how other professionals define what data science is to them and the paths they took to consider data science as a career for themselves.
- ➢ You explore data scientists' different roles, how data analysis is used in data science, and how data scientists follow certain processes to answer questions with that data.
- ➢ Moving on to Lesson 2, the focus shifts to the daily activities of data scientists.
- ➢ This encompasses learning about real-world data science problems professionals solve, the skills and qualities needed to be a successful data scientist and opinions on how "big data" relates to those skills.
- ➢ You also learn a little about various data formats data scientists work with and algorithms used to process data.

Learning Objectives

- Define data science and its importance in today's data-driven world.
- List some paths that can lead to a career in data science.
- Summarize the advice seasoned data science professionals give to data scientists who are just starting out.
- Articulate why data science is considered the most in-demand job in the 21st century.
- Describe what a typical day in the life of a data scientist looks like.
- Define some of the commonly used terms in data science.
- Identify some of the key qualities of a successful data scientist.

Welcome to the Course

Course Introduction

🎯 Course Objective

- Welcome to your journey into a **thriving career in Data Science**.
- **Why Data Science?**
    - 📈 Rapid growth due to:
        - Abundance of electronic data
        - Enhanced computing power
        - AI advancements
        - Proven business value

Industry Outlook (U.S. Focus)

- 🔍 **Projected Growth**: 35% (US Bureau of Labor Statistics)
- 💵 **Median Annual Salary**: ~$103,000

🎓 Who Is This Course For?

- Beginners curious about Data Science
- No prior knowledge or degree is required
- 📈 Business Managers & Executives seeking to make **data-driven transformations**

💡 What You'll Learn

- Role and skills of a data scientist
- Key technologies and concepts:
    - 🔢 Big Data
    - 🧠 Artificial Intelligence (AI)
    - 📘 Machine Learning (ML)
    - 🧠 Deep Learning
- How data science:
    - Tells compelling stories 📖
    - Informs and transforms businesses 🏢

🧰 Learning Resources

- 🎥 Instructional videos
- 👩‍🏫 Interviews with Data Science professionals
- 📖 Readings and glossaries
- ✅ Practice assessments
- 📝 Summary videos for each lesson
- 📁 Final case study + quiz
- 📂 Final peer-reviewed project: **Explore real job listings**

🧭 Course Structure

📦 3 Core Modules + 1 Optional Module

📘 Module 1: Introduction to Data Science

- 🔍 What is Data Science?
- 👨‍💼 Role of a Data Scientist
- 🧠 Skills required to succeed
- 📁 File types & handling
- 📚 Topics and algorithms overview
- 👨‍🔧 Qualities of a good data scientist
- 💡 Importance of Big Data

☁️ Module 2: AI & Big Data in Transformation

Lesson 1:

- ☁️ Big Data + Cloud Computing = Digital Transformation
- 🛠️ Tools and techniques
- 🗂️ Data mining

Lesson 2:

- 🤖 Introduction to Artificial Intelligence
- 🧠 Concepts: Machine Learning & Deep Learning
- 🧰 How AI powers Data Science

🌍 Module 3: Applications of Data Science

- 🧭 Explore various real-world applications
- 🏥 Business, healthcare, finance, marketing, etc.
- 📊 Practical data science activities

🧩 Optional Module: Data Literacy

- 📚 Introduction to **Data Ecosystem**
- 📡 Sources of data
- 🏢 Databases vs Data Warehouses vs Data Marts vs Data Lakes
- 🔄 ETL Process (Extract, Transform, Load)
- 🛠️ Data Pipelines

💬 Support & Interaction

- 💬 Discussion Forums:
  - Ask questions
  - Find answers
  - Connect with peers and support staff

🏁 Course Outcome

- ✅ Complete the course
- 🏅 Earn a Certificate
- 🚀 Take your next step into a **Data Science Career**

📚 Pathways After This Course

- IBM Data Science Professional Certificate
- Introduction to Data Science (Specialization)
- Key Technologies for Business
- IBM AI Foundations for Business

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

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

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

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

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:

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

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: Infographics on roadmap

Data Literacy for Data Science (Optional):

Understanding Data

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

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

Professional Certificate Career Support

Coursera Community and Career Support

As a Data Science learner on Coursera, you have access to networking opportunities in Coursera's Professional Certificate Community and Data Science forums. Talk about what you're learning, ask questions, find peers to work with on projects, and share your career goals.

Post-Completion Career Support Services for Professional Certificates

Completing a Professional Certificate on Coursera unlocks access to a private Professional Certificate Alumni Resources community, which provides exclusive career support resources, including:

- Step-by-step guide to ensure your success at every stage of your job search.
- 1 year of free access to Big Interview's expert video lessons, resume builder, and interactive interview practice tools (a $79/month value).
- A network and support of fellow completers of Coursera Professional Certificates.
- Various special offers, such as career coaching, webinars, and more.

After completing your Professional Certificate, you'll get an email about accessing these career support resources.

Questions? You can also contact the Coursera Careers Team by emailing career-support@coursera.org.

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 data science experts explain 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 fits 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?

📌 Definition & Essence

- 📊 **Data Science is a process**, not a one-time event.
- 🧠 It uses **data to understand the world**, uncover insights, and test models or hypotheses.
- 🎯 The core of data science is **formulating questions** and **seeking answers through data**.
- 🧪 Like other sciences (biology, physics), **Data Science is the study of data**.
- 🧩 It involves **data** and **science**—a blend of curiosity, logic, and computation.

🔍 Purpose & Function

- ✅ To **validate hypotheses or models** using data.
- 📖 To **translate data into stories** and insights.
- 💡 To **generate actionable insights** to drive strategic decision-making for companies or institutions.
- 🗺️ Helps us navigate and understand complex systems through data interpretation.

🛠️ Processes & Tools

- ⚙️ Data Science is built on **processes and systems** that work with:
  - 📄 **Structured Data** (e.g., spreadsheets, SQL databases)
  - 🌐 **Unstructured Data** (e.g., text, images, video)
- 🧰 Involves **data collection, cleaning, exploration, modelling, and interpretation**.
- 📈 Utilizes **statistics, machine learning, and programming** to analyze data.
- 📚 Originated as a **rebranding of academic statistics** during the 1980s–90s.

🌍 Why It Matters Today

- 🌊 We're experiencing a **data deluge** (huge amounts of data everywhere).
- 🤖 Algorithms are now more advanced and widely available.
- 💻 Software tools (Python, R, Jupyter, etc.) are mostly **open-source and free**.
- ☁️ **Data storage is cheaper and more scalable** than ever before.
- 📡 Data Science is highly relevant today because of:
  - Ubiquity of data
  - Accessibility of tools
  - Low cost of storage and computation

👷 Who is a Data Scientist?

🕵️ A **curious explorer** of data.

🧰 Someone who **manipulates and explores** data to answer meaningful questions.

📖 A **storyteller who speaks the language of data**.

🔁 In a Nutshell

Data Science is more about **data** than **science**.

If you have **data** + **curiosity** + **the drive to explore**, you're already doing data science.

🗓️ There's **never been a better time** to be a data scientist.

Fundamentals of Data Science

📌 Overview of Data Science

- Everyone gives a slightly different definition of **Data Science** 🤔.
- Common agreement: it has a **significant data analysis component** 📈.
- **Data analysis isn't new**, but:
  - 🧠 What's **new** is:
    - **Massive quantity** of data 📦
    - **Variety** of data sources 🌐

🗂️ Sources of Data

- Log files 🗃️
- Email ✉️
- Social media 📱
- Sales data 🛒
- Patient information files 🏥
- Sports performance data ⚽
- Sensor data 📟
- Security cameras 🎥
- And many more…

🖥️ Technological Advancements

- **More data** than ever 📊
- **Improved computing power** allows:
  - Deeper analysis 💻
  - Discovery of **new knowledge** 🔍

💼 Purpose of Data Science in Organizations

- Understand their **environment** 🌍
- Analyze **existing issues** 🛠️
- Reveal **hidden opportunities** 🎯
- Add to **organizational knowledge** 🧩

❓ The Data Science Process

1. **Clarify the question** the organization wants answered 🎯
   - ✅ **Crucial first step**
   - Determines the **direction** of the project
   - Good data scientists are **curious** and ask questions 🤓
2. **Determine the needed data**:
   - What data is required? 📄
   - Where will the data come from? 🌐
3. **Analyze the data**:
   - Both **structured** and **unstructured** data 📂🗃️
   - Different analysis methods depending on the **problem** 🧠
   - Use of **multiple models** to:
     - Discover patterns 📈
     - Detect outliers ⚠️
     - Validate or challenge assumptions 💡

📢 Communicating Results

- Data scientists become **storytellers** 📖
- Communicate findings to **stakeholders** 👥
- Use **data visualization tools** 📊:
    - Help stakeholders **understand the results**
    - Provide **recommended actions**

🌐 Impact of Data Science

- Changing the **way we work** 💼
- Changing how we **use data** 📉➡️📈
- Transforming how **organizations understand the world** 🌎

The Many Paths to Data Science

🌍 1. The Emergence of Data Science

🔄 **Data Science didn't exist** as a formal field during the early years of the speakers' lives.

🗓️ Became recognized around **2009–2011**.

🧪 Coined/popularized by **DJ Patil** and **Andrew Gelman**.

📉 Before that, similar work was classified under:

- **Statistics**
- **Analytics**
- **Business Intelligence**

👴 Nobody "grew up wanting to be a data scientist" — the title and career path didn't exist.

👥 2. Personal Backgrounds & Career Paths

👩‍🎤 A. Speaker 1

🎓 Studied **Statistics**.

🧮 Realized strong math skills → pursued **quantitative analysis**.

🎤 Wanted to be a **singer**, then a **doctor**, before discovering a love for data.

💼 Didn't originally plan for data science — but discovered it later and found it fascinating.

👨‍💼 B. Speaker 2

- 📘 First exposed to data science during **1st year of Mechanical Engineering**.
- 💼 Saw its application in **strategic consulting firms**.
- 🧩 Had a **complex problem** that couldn't be solved with traditional techniques — data science helped.
- 🛠️ Applied data to solve **real-world challenges**.

👨‍🎓 C. Speaker 3

- 📈 Earned a degree in **Math** — during the **economic crisis**.
- 🧰 Needed practical skills to get a job → Pursued **statistics**.
- 🔁 Worked in several roles titled "Data Scientist" → became one by practice.
- 💼 Career evolution from **math → stats → data science**.

👩‍💼 D. Speaker 4

- 🎓 Bachelor's in **Business** with a major in:
  - 🏛️ Politics
  - 📚 Philosophy
  - 💰 Economics (PPE)
- 🎓 Master's in **Business Analytics** at **NYU Stern School of Business**.
- 🛒 First job involved analyzing **electronic point-of-sale (POS) data**.
- 🔍 Realized only **later** that this work was **data science**.
- 🕒 The term "data science" was adopted around **4–5 years ago** in that company.

👨‍🔧 E. Speaker 5

Had **multiple internship options** in Canada.

- 💡 Choose an internship in **Data Science** over **Project Development**.
- ✅ Feels it was a **good decision**.
- 🚀 That internship marked the **start of their data science career**.

👷 F. Speaker 6

- 🎓 Background in **Civil Engineering**.
- 🧱 Engineers naturally work with data.

- 🚦 Entry into data science began through **transportation research**:
  - 🚗 Forecasting **traffic**
  - 🔥 Predicting **greenhouse gas emissions**
  - 🚦 Measuring **congestion**
- 🏫 Started building large models as a **graduate student at the University of Toronto**.
- 📊 Worked with large datasets (mid-90s):
  - 👨‍👩‍👧‍👦 150,000 households
  - 🚘 500,000 trips
- 🏢 Became a **Professor at McGill University** in **Transportation Engineering**.
- 📈 Built **larger models** with integrated **data and analytics**.

🎯 3. Key Themes & Takeaways

- ❗ Data Science is **new and evolving**.
- 🔍 Discovered through:
  - Practical problems
  - Career transitions
  - Academic research
- 💡 Common traits among data scientists:
  - Strong **math and analytical skills** 🧠
  - Interest in **problem-solving** 🧩
  - Exposure through **interdisciplinary fields**
- 🧭 No one set path — people come from:
  - Engineering 👷
  - Business 👔
  - Statistics 📊
  - Philosophy & Politics 📘
  - Project Management 📋

Advice for New Data Scientists

🔍 Core Qualities of a Data Scientist

1. **Curiosity** 🧠
    - *Absolute must.*
    - Without curiosity, you won't know what to do with the data.
    - Curiosity fuels exploration and questioning.
2. **Judgmental** ⚖️
    - Having *preconceived notions* or beliefs gives you a starting point.
    - These hypotheses guide your investigation, even if proven wrong later.
3. **Argumentative** 🗣️
    - The ability to argue or plead a case is essential.
    - Take a *strong position* and then validate or modify it with data.
    - Enables a **learning process**:

      "I thought I believed this, but now, with data, I know this."

🛠️ Technical & Analytical Skills

4. **Comfort with Analytics Platforms** 💻
    - Knowledge of tools/software/computing platforms is important.
    - But it's **secondary** to curiosity and hypothesis-building.
5. Examples:
    - Programming languages (Python, R) 🐍
    - Data platforms (SQL, Hadoop, Spark) 🧱
    - Visualization tools (Tableau, PowerBI) 📊

📖 The Importance of Storytelling

5. **Tell a Great Story** 📚
    - After analysis and tabulation, communicate your findings effectively.

- Good storytelling ensures your work doesn't remain hidden.
- *"Your rise to prominence relies on your ability to tell great stories."*

🎯 **Career Direction & Competitive Advantage**

6. **Identify Your Field of Interest** 🧭
   - Decide if you want to work in a **specific domain** (e.g., health, IT, retail) or be a generalist.
7. **Know Your Competitive Advantage** 🏆
   - It may **not** be your analytical ability.
   - It could be your *deep understanding* of a certain domain:
     - Film 🎬
     - Retail 🛍️
     - Health 🏥
     - Technology 💾
8. **Acquire Domain-Specific Tools & Skills** 🛠️
   - Choose tools and platforms based on your target industry.
   - Example:
     - Health: Biostatistics, EMR systems
     - IT/Web: Web analytics, cloud computing

🚀 **Apply and Share**

9. **Apply Your Skills to Real Problems** 📈
   - Use real-world projects to demonstrate your abilities.
10. **Show the World What You Can Do** 🌎

- Publicly share your work and insights.
- Portfolio, blogs, GitHub, Kaggle 😺📁

✅ **Summary Checklist**

- Be curious 🔍
- Be judgmental ⚖️

- Be argumentative 🗣️
- Learn analytics tools 💻
- Master storytelling 📖
- Choose your industry 🎯
- Know your unique strength 🏆
- Get domain-relevant skills 🧰
- Solve real problems 📊
- Tell the world! 📢

Lesson Summary: Defining Data Science

🔍 What is Data Science?

- **Simple Definition**:
  Data science is the **study of data** to understand the world around us.
- **Purpose**:
  Uses data to uncover **insights and trends**.
  Referred to as an **art of uncovering hidden knowledge**.
- **Modern Advancements**:
  - Increased **data access** 📊
  - Enhanced **computing power** 🖥️
  - Leads to **deeper insights** and **knowledge extraction**
- **Analogy**:
  Like a **detective**, a data scientist uncovers secrets within data and translates them into **stories** that guide **strategic decisions**.

🧠 The Role of a Data Scientist in an Organization

- **Function**:
  Translates data into **actionable insights** for better decision-making.
- **Process Includes**:
  - Clarifying the **problem** ❓
  - **Collecting** data 📥
  - **Analyzing** data 🧮
  - Recognizing **patterns** 🔁
  - **Storytelling** with data 📖

- Data **visualization** 📈

- **Data Types**:
  - Structured data (e.g., databases)
  - Unstructured data (e.g., text, images)

🧰 Qualities of a Skilled Data Scientist

According to **Professor Murtez Aheter**, three key qualities:

1. **Curiosity** 🔎
   - Drives exploration
   - Leads to meaningful questions
2. **Argumentation** 💬
   - Building strong, evidence-based reasoning
   - Convincing others to shift perspectives
3. **Judgment** ⚖️
   - Helps in choosing the right starting point and approach

🛠️ Skills & Attributes Needed

- More than **statistics** or **computer science**
- **Multidisciplinary background** is valuable:
  - Economics, engineering, medicine, etc.
- Ideal data scientists:
  - Comfortable with **math** ➗
  - **Curious** and exploratory 🧐
  - Skilled in **storytelling** with data 📚
  - Can **communicate** insights 🗣️

📚 How to Acquire These Skills

1. **Identify your strengths** and interests
2. **Master data analysis** in your chosen domain
3. Use **industry-relevant tools** 🧪

4. Apply knowledge to solve **real-world problems** 🌍
    - Like solving a mystery using clues 🔍

🚀 Future of Data Science Careers

- **Evolving roles** with advancing technology
- Increased demand for **certification** 📜
- The continuous need for:
    - **Logical thinking** 🧠
    - Effective use of **algorithms** 🔢
    - **Methodical approaches** ✅
    - Accurate **data gathering** and **model analysis** 📊
- Goal: Achieve **successful business results** using data-driven decisions 💼📈

🎯 Conclusion

A data scientist combines **technical expertise**, **domain knowledge**, and **communication skills** to generate impactful insights. The field is dynamic, and success comes from **continuous learning** and **practical application**.

What Do Data Scientists Do?

Lesson Overview: What Do Data Scientists Do?

You'll dive into data science in the lesson "What Do Data Scientists Do?". 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 data scientists' daily routines and challenges, providing a practical glimpse into their roles.
"Data Science Skills and Big Data" video	Delve into the core skills required in data science and understand big data's significance 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 data analysis and decision-making foundation.
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.

A Day in the Life of a Data Scientist

🔍 1. Recommendation Engine at a Large Organization

- ✅ **Responsibility**: Built a recommendation engine from the ground up.
- 🤝 **Collaboration**: Worked with various stakeholders — engineers, developers, executives.
- 🧩 **Key Success**:
  - Delivered a **simple, elegant solution** that:
    - It was **easy to understand** across all technical levels.
    - It was **highly efficient** despite its simplicity.
    - Avoided unnecessary complexity — could’ve taken longer with a more complex approach.
- 🏆 **Proud Moment**: This solution stands out as a favorite due to its clarity, accessibility, and performance.

🌱 2. Algae Bloom Prediction at University

- 🎯 **Objective**: Predict harmful **algae blooms** that:
  - Increase **water toxicity** 💧
  - Disrupt **water treatment operations**
- 🧪 **Challenge**:

- Traditional **chemical engineering** methods were insufficient for prediction.

- 🧠 **Solution**:
  - Used **Artificial Neural Networks (ANNs)** 🤖
  - Predicted **timing and occurrence** of algae blooms.
  - Helped **water treatment companies** plan better and prevent potential issues.

- 🎓 **Significance**: Applied interdisciplinary thinking (chem eng + AI) to solve a real-world environmental issue.

🚌 3. Toronto Transit Commission (TTC) Complaints Analysis

📞 The Ask

- 🏙️ Organization: Toronto Transit Commission (TTC) — one of the largest transit agencies in North America.
- 📩 Issue: Analyzing **complaints data** from customers.
- 📊 Dataset:
  - ~ **500,000 complaints** 😳
  - Combination of **structured + unstructured data**.

📁 Data Breakdown

- 🧾 **Structured data**:
  - Date of complaint
  - Complaint recipient
  - Type of complaint
  - Resolution status

    - Attribution (who was responsible)

- ✉️ **Unstructured data**:
    - Email exchanges
    - Faxes
    - Free-text entries

🧪 Initial Exploration

- 🧠 Goal: Understand **why** and **when** people complain.
- 🗓️ Hypothesis: There might be patterns — certain days had **spikes** in complaint volume.
- 🧮 Tried:
    - Multiple data formats
    - Various statistical and exploratory techniques
- ❌ Initial Findings: No **clear patterns** or explanations for complaint surges.

🌧️ Breakthrough Moment

- 🌫️ Real-life experience: Stepped into a puddle on a rainy day — felt annoyed.
- 💡 Insight: Could **weather** be influencing complaints?
- 🔍 Investigation:
    - Collected **weather data** from Environment Canada:
        - Rain
        - Snow
        - Wind

        - Sudden temperature changes

📈 Findings

- 📊 Discovered a strong correlation:
    - **Top 10 complaint-heavy days** had:
        - **Bad weather**
        - Sudden **temperature drops**
        - **Unexpected rain**
        - **Heavy snow**
        - **Strong winds**
- 🧠 Insight: **External factors** like weather were **key drivers** of spikes in complaints.

🗣️ Communication to Executives

- 📢 Delivered:
    - ✅ **Good News**: Discovered **why** complaints spike on certain days.
    - ❌ **Bad News**: **TTC has no control** over weather conditions.
- 🧠 Implication:
    - Some operational challenges are **unpredictable** and **external**.
    - Importance of **contextual data** (like weather) in modelling human behavior.

🧩 Key Takeaways

- 🔁 Simplicity in design can be just as powerful as complex solutions.

🌎 Interdisciplinary thinking (ENG + AI + environmental science) solves real problems.

💡 Real-world intuition (stepping into a puddle!) can spark major insights.

📊 Data without **context** (like weather) can lead to missed patterns.

🧠 Communicating insights clearly to both technical and non-technical stakeholders is **crucial**.

Data Science Skills & Big Data

👩‍🏫 **Profile: Norman White**

- **Name**: Norman White
- **Role**: Clinical Faculty Member
- **Department**: IOMS (Information, Operations, and Management Sciences)
- **Institution**: NYU Stern School of Business
- **Current Title**: Faculty Director, Stern Center for Research Computing
- **Tenure**: At Stern since finishing college (early 1970s)

💻 **Personal Interests & Background**

- Self-described as a **"techy, geeky"** individual
- Loves **playing with technology** in spare time
- Early interest in computers since the **early days of computing**

🎓 **Academic Background**

- **Undergraduate**: Degree in **Applied Physics**
    - I took several **Economics courses** during my undergrad
- **Graduate**:
    - PhD in **Economics and Statistics** from Stern

🖥️ Career Milestones

🏢 Early Roles

- Joined **NYU Business School** (then downtown)
- Worked at the **computer center** during studies
    - Learned to **program** (self-taught)
    - Attempted to learn **touch typing** (now back to two-finger pecking 😄)
    - Took courses at **IBM**
    - Became **Director of the Computer Center**

👩‍🏫 Department Formation

- In **1973**, NYU formed the **Computer Applications and Information Systems Department**
- Norman was one of the **founding faculty members**
- He has remained in the department ever since

☁️ Stern Center for Research Computing

- Norman is the **Faculty Director**
- The Center manages a **private cloud** infrastructure
- Supports:
    - **Faculty & PhD students**
    - Custom **virtual machines** spun up for specific hardware/software needs
- 🔍 Widely used by **Data Scientists** at Stern
    - Especially PhD students

🗓️ Typical Monday Schedule

Time	Activity
Morning	Email (handled at home before arriving)
11:00 AM	Arrives on campus
2:00 PM	**"Dealing with Data" class**
6:00 PM	**"Design and Development of Web-Based Systems"**

📚 “Dealing with Data” Course Overview

🛠️ Technologies & Topics Covered

- **Unix/Linux** basics
- **Python** programming 🐍
- **Regular Expressions** 🔍
- **Relational Databases** 🗄️
- **Python Pandas** 📊
    - Compared to **R**, it allows statistical/mathematical operations
- **Big Data** 🌐
    - Norman is a strong **advocate (evangelist)** of big data

📈 Course Format

- Weekly **homework assignments**
- **Team-based final projects**
    - Students create **innovative and practical solutions**

💡 **Teaching Infrastructure**

🧪 **Learning Platform**

- The entire course is taught using **Jupyter Notebooks** 📓

☁️ **Deployment**

- Each student receives a **personal virtual machine** on **AWS (Amazon Web Services)** 🌩️
- Standardized virtual machine images:
    - All software and course materials pre-loaded or accessible via Jupyter Notebook commands
- **Platform-Agnostic Setup**:
    - Consistent environment regardless of student's OS (Mac, Windows, old/new machine)

🎵 **Final Notes**

- Teaching style blends **technical rigour**, **real-world applications**, and **hands-on cloud-based computing**
- Passionate about equipping students with **modern data science tools** in a **uniform, scalable, and accessible** way

Understanding Different Types of File Formats

1. 📜 Delimited Text File Formats

- **Definition**: Text files where each line represents a record and values are separated by a **delimiter**.
- **Common Delimiters**:
    - Comma `,` → CSV (Comma-Separated Values)
    - Tab `\t` → TSV (Tab-Separated Values)
    - Others: Colon `:`, Vertical bar `|`, Space
- **Use Cases**:
    - CSV: Default format when delimiter is a comma.
    - TSV: Used when commas exist in text data (avoids confusion).
- **Structure**:
    - Each row = one record 📄

    - First row = column headers 🏷️
    - Columns may contain different data types (e.g., date, string, integer) 📊
- **Pros**:
    - Field values can be of any length.
    - Standard, simple, and supported by almost all applications 🛠️
    - Excellent for schema representation and data exchange 💬
- **Limitation**:
    - May lack strong data typing or advanced features.

---

2. 📊 Microsoft Excel Open XML Spreadsheet (.XLSX)

- **Definition**: XML-based file format used by Microsoft Excel.
- **Structure**:
    - **Workbook** (.xlsx file) contains multiple **worksheets** 📘
    - Each worksheet has **rows** and **columns** forming **cells** containing data.
- **Features**:
    - Open format → accessible to most applications 🔓
    - Supports Excel functions (formulas, charts, styles, etc.)
    - More **secure** than legacy formats (does not save malicious code) 🔐
- **Use Case**:
    - For structured, tabular data with formulas and formatting needs.

---

3. 🧩 Extensible Markup Language (XML)

- **Definition**: A **markup language** used to encode documents in a human- and machine-readable format.
- **Features**:
    - Self-descriptive tags 🏷️
    - No predefined tags (unlike HTML)
    - Platform and language-independent 🌐
    - Often used for **data interchange** across systems.

- **Use Case**:
    - Sending structured information across the internet.
    - Configuration files, web services, B2B communication.

---

4. 📄 Portable Document Format (PDF)

- **Definition**: A format developed by Adobe for document presentation independent of software, hardware, and OS.
- **Features**:
    - Consistent view across devices 📱💻🖥️
    - Commonly used for **legal**, **financial**, and **form-based documents** 📃
- **Use Case**:
    - Sharing read-only documents
    - Capturing fixed layouts with embedded fonts and graphics.

---

5. 🧬 JavaScript Object Notation (JSON)

- **Definition**: A text-based, open standard format for transmitting structured data over the web 🌐
- **Features**:
    - Language-independent 📜
    - Readable in any programming language 👨‍💻
    - Highly compatible with web technologies and browsers
    - Supports **structured**, **nested**, and even **binary (audio/video)** data 🎥🎧
- **Use Case**:
    - APIs and web services (default format)
    - Configuration files and data exchange in web apps.

📚 Summary

Format	Best For	Pros	Limitation
CSV/TSV	Simple tabular data	Widely supported, simple	No formatting, minimal metadata
XLSX	Rich spreadsheets	Secure, support formulas	Requires spreadsheet software
XML	Data interchange	Self-descriptive, cross-platform	Verbose, complex to parse
PDF	Read-only documents	Consistent appearance	Hard to extract structured data
JSON	Web data exchange	Lightweight, widely used in APIs	Limited data typing

---

➡️ **Next Topic**: Different Sources of Data 📡

Data Science Topics and Algorithms

📈 Regression

- ✅ **First Concept Many Understand in Data Science**
- 💡 **Regression is foundational** to understanding relationships in data.
- 📚 **Many books & lectures exist**, but they often complicate things with too much statistical jargon.
- 🚕 **Simple Analogy - Taxi Ride:**
  - **Base Fare** (e.g., $2.50) = **Constant**
  - **Fare increases** with:
    - 🛣️ Distance (e.g., per km)
    - ⏱️ Time (e.g., stuck in traffic)
  - 🧠 Regression finds:
    - Base fare (intercept)
    - Cost per distance (coefficient)
    - Cost per minute (coefficient)
  - 🔍 **Use Case**: Even without knowing the fare formula, regression can **derive the relationship** between inputs (distance, time) and output (fare).

📊 Data Visualization

- 🎨 **Loved by the speaker**
- 🗣️ **Key to communicating** with people who aren't deeply familiar with data science.
- 🧰 Preferred Tool: **R** for Data Visualization
- 🧠 Makes complex data **accessible and interpretable**.
- 📢 Essential for storytelling with data.

🧠 **Artificial Neural Networks (ANNs)**

- ❤️ **Strong Passion** for ANNs
- 🌿 Inspired by **biological behavior**, especially the **human brain**.
- 🧬 **Mimicking nature** through algorithms can unlock powerful applications.
- 🧪 There's a lot to learn from **natural intelligence** to improve artificial ones.

🤝 **Nearest Neighbor (KNN Algorithm)**

- 🧩 **Simple yet powerful**
- 🎯 Often gets **better results** than more complex models
- ⚖️ Complex models may **overfit**, while KNN often gives:
  - 🔁 **Consistent performance**
  - 🧪 **Less risk of overfitting**
- 🛠️ Works well in practice despite its simplicity.

🗄️ **Structured vs. Unstructured Data**

🧮 **Structured Data:**

- 📊 Data in **rows and columns**
- 👨‍💼 Familiar format like **Microsoft Excel**
- 💡 Easy to analyze and model
- 🔎 Ideal for traditional data science models

🌐 **Unstructured Data:**

- 📄 Comes from sources like:
  - 🌍 Web pages
  - 📝 Text
  - 🎥 Video
  - 🔊 Audio
- 🔧 Requires **advanced techniques** to extract value
- 🏗️ Often involves converting it into **structured format** for analysis

---

🧠 **Key Takeaway on Regression**

- 🚕 **Taxi Fare Analogy**:
  - **Base Fare** = Constant (intercept)
  - **Per km & per minute** rates = Coefficients
  - 📊 Regression calculates the **underlying pattern** from observed data
- 🧮 Lets you **predict outcomes** when you know the relationships
- 📈 A cornerstone of predictive analytics

Lesson Summary: What Do Data Scientists Do?

📊 **Data Science: A Gateway to Innovation & Discovery**

🧠 **Role and Purpose of Data Scientists**

- Data is more than numbers – it's a gateway to **innovation**, **discovery**, and **endless possibilities**.
- Data scientists:
  - Investigate problems
  - Find explanations using data
  - Offer innovative, real-world solutions

🔍 **Real-World Examples**

1. **Public Transit Complaints in Toronto**
   - 👨‍🏫 *Dr. Murtaza Haider* found a link between **bad weather** and **increased transit complaints**.

- Demonstrates how data can uncover hidden patterns in everyday issues.

2. **Environmental Challenges**
   - Predicting **algae blooms** 🌱 to prevent **water toxicity** 💧.
   - Use of **artificial neural networks** 🤖 to assist **water treatment companies** in protecting ecosystems.
3. **Recommendation Engines**
   - 👩 *Norman White* built systems to simplify complex, cross-departmental problems.
   - Showcases data science as a problem-solving discipline.

🎓 Education & Skill Development

- 👩 *Dr. White's* classroom:
  - Python notebooks 🐍
  - Unix/Linux systems 💻
  - Relational databases 🗃️
  - Tools: Pandas 🐼
- 👩 *Dr. Vincent Granville* emphasized:
  - **Mathematics**: Algebra ➕, Calculus ∫, Probability & Statistics 📈
  - Differentiates:
    - **Statistician**: uses stats only
    - **Data Scientist**: blends stats with broader tools and skills

---

📈 Techniques & Tools

- **Statistical Models**
  - Regression: shows probable relationships (e.g., distance vs. gas usage 🚗 ⛽)
- **Machine Learning Algorithms**
  - Nearest neighbor algorithm for processing **big data** 🧠
- **Big Data Tools**
  - Hadoop 🐘: breaks traditional data limitations
  - Data size is evolving — "big data" is a fluid term

- 👨‍🏫 *Dr. Patel's* Perspective:
  - Data scientists aren't limited by:
    - Data size
    - Specific tools
  - Welcomes diverse academic backgrounds

💬 **Communication & Storytelling**

- Data scientists must:
  - Extract insights from data
  - Tell **compelling stories** to stakeholders 🎤
  - Combine **technical skill** with **communication finesse**

📁 **Data Types & Formats**

- **Sources**: Video 📽️, Audio 🎧, Unformatted Text 📝
- **Structures**:
  - Structured: Tables (rows & columns) 📊
  - Unstructured: Emails 📧, Logs 📜
- **Formats**:
  - Delimited text files (.csv) 📄
  - Spreadsheets (.xls/.xlsx) 📑
  - XML 🗒️
  - PDFs 📘
  - JSON 🔣

🌟 **What Makes an Exceptional Data Scientist?**

- 👨‍🏫 *Rachel Schutt* (Chief Data Scientist, News Corp):
  - Key trait: **Curiosity** 🔍

- Ideal blend of:
  - Computer Scientist 👨‍💻
  - Software Engineer 🛠️
  - Statistician 📉
- Skilled in turning **unstructured problems** into **structured insights**

🚀 **Conclusion: A Journey, Not Just a Job**

- Being a data scientist is:
  - A journey of **exploration**
  - A mission of **innovation**
  - A path of **storytelling**
- With skills, curiosity, and determination, data scientists navigate the vast data universe 🌌 to reveal the extraordinary ✨.

Summary: What Do Data Scientists Do?

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

- Data science studies 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 pay high salaries for skilled workers.
- The typical work day for a data scientist varies depending on the 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, models, algorithms, JSON, and XML. CSV, and regression.