Course 1: What is Data Science

Week 1:

Lesson 1:

Data Science?

- Translate data into a story
- Field about processes and systems to extract unstructured and structured data
- Data is used to find answers to questions

Fundies:

- Clarify question org wants answered
- Find business need: what data do we need to solve the problem?

Advice:

- Be curious
- Comfort with analytical platforms
- Be able to tell a great story
- Find an industry I'm interested in

Sexiest Job Article

- Digital universe:
    - 1995: 130 billion GB
    - 2020: 40 trillion GB

Lesson 1 Summary:

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

Lesson 2:

A day in the life:

- Built recommendation engine
- Sort through complaints for travel company

Old / New Problems, Data Science Solutions:

- Uber
    - Places more drivers in places where drivers are needed
- Routing
    - Reducing hours lost during commutes from commuters
- Environment
    - Predicting cyanobacterial blooms with complex algorithms

Topics and Algorithms:

- Regression
    - Imagine a Taxi Ride:
    - The moment you sit in a cab ride, in a cab, you see that there's a fixed amount there. It says $2.50. You, rather the cab, moves or you get off. This is what you owe to the driver the moment you step into a cab. That's a constant. You have to pay that amount if you have stepped into a cab. Then as it starts moving for every meter or hundred meters the fare increases by certain amount. So there's a... there's a fraction, there's a relationship between distance and the amount you would pay above and beyond that constant. And if you're not moving and you're stuck in traffic, then every additional minute you have to pay more. So as the minutes increase, your fare increases. As the distance increases, your fare increases. And while all this is happening you've already paid a base fare which is the constant. This is what regression is. Regression tells you what the base fare is and what is the relationship between time and the fare you have paid, and the distance you have traveled and the fare you've paid. Because in the absence of knowing those relationships, and just knowing how much people traveled for and how much they paid, regression allows you to compute that constant that you didn't know. That it was $2.50, and it would compute the relationship between the fare and the distance and the fare and the time. That is regression.
- Data visualization
    - Using R
- Neural networks

- Structured & Unstructured Data
- Nearest Neighbor

Cloud for Data Science:
- Ability to do high performance computing without using your own computer but using something else
- Multiple entities can work with the same data
- Instant access to open source technologies
- Programs:
    - Jupyter
    - Spark clusters???

What makes someone a Data Scientist?:
- Someone who finds solutions to problems by analyzing big or small data sets using tools then appropriately tells a story to communicate findings to stakeholders
- They are curious, that's the key link
- Part CS, part Software Engineer, part Statistician

Lesson 2 Summary:
- 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.
- Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.

Week 2:

Lesson 1:

Foundations of Big Data:
- Everyone leaves a trace
- "Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value."
- Five V's
    - Velocity
        - Speed which data accumulates

        - Fast and never stops
        - Every 60 seconds, hours of footage are uploaded to youtube
    - Volume
        - Scale of data, increase of data stored
        - World population: 7 billion with vast majority using digital devices
        - 2.5 quintillion bytes a day are generated
    - Variety
        - Diversity of data
        - Structured and unstructured data
        - Different sources: machines, people, processes
        - Texts, pictures, films, sound, health, anything in IoT
    - Veracity
        - Quality and origin of data
        - Consistency, completeness, accuracy
        - 80% of data is unstructured
            - Need a reliable and efficient way to analyze and then visualize and communicate to others
    - Value
        - Ability to turn data into value
    - Apache Spark / Hadoop

Hadoop?
- Slice data and send it to many different computers with the same program executed on each slice
- All returns to one place, sorted, then analyzed

How is BIg Data Driving Digital Transformation?
- Organizational and cultural change
- Netflix
- Houston rockets
    - Raised game using big data
    - Mined raw data from games using overhead cameras
    - Changed the way teams tried to win
- Lufthansa

Skills & Big Data
- Just an interview honestly

Data Minig:
- Set goals
    - Identify key questions to answer
    - Costs and benefits
- Select Data
    - Identify sources of data and plan data collection initiatives

    - Data, size, frequency of collection all are important to consider
- Preprocess Data
    - Raw data is messy with lots of irrelevant data or missing data
    - Expunge irrelevant attributes of data
    - If missing data is random, meh oh well. If its systematic, you need to find out why and if it will affect your data. Findings relying on individuals' incomes with some leaving it out may leave systematic biases in the analysis
- Transform data
    - Determine the appropriate format to store data
    - Reduce number of attributes needed to explain phenomena
    - If multiple forms of income exist, compiling them all under an aggregate income would be beneficial and more representative of individual incomes
    - Transforming variables from one type to another is common
    - Principal Component Analysis
- Store data
    - Store data to make it conducive for mining
    - Unrestricted and immediate read/write privileges to data scientist
    - Data safety and privacy is a prime concern for storing data
- Mine Data
    - Parametric and non-parametric mining methods, ML algorithms, data visualization
    - Develop a way to demonstrate trends hidden in data sets
- Evaluate Mining Results
    - Formal evaluation of results after mining
    - Test predictive capabilities of models on observed data
    - Results are shared with key stakeholders for feedback
- Last two parts are iterative processes which develop improved algorithms and improve quality of results generated

Lesson 1 Summary:

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

Lesson 2:

The Differences:

- Big Data

        - Data sets so large and develop so quickly they defy traditional analysis methods
        - Five V's
- Data Mining
    - Automatically searching and analyzing data
    - Discover undiscovered patterns
    - Data visualization
    - Machine learning
- Machine Learning
    - A subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it has learned, without being explicitly programmed
    - Trained with large sets of data
    - Learn from examples
    - Do not follow rules-based algorithms
- Deep Learning
    - Subset of machine learning
    - Uses layered neural networks to simulate human decision-making
    - Label and categorize information and identify patterns
    - Learns on the job and determines if decisions were correct
- Neural Networks
    - Takes inspiration from biological neural networks, although work differently
    - Collection of small computing units called neurons
    - More data = more efficient
- AI VS Data Science
    - Data Science = process and methods for extracting knowledge and insights from large volumes of data
        - Uses math statistical analysis, data visualization, machine learning, etc.
        - Broad term encompassing entire data processing methodology

Neural Networks & Deep Learning

- Deep Learning uses
    - Speech recognition
    - Facial/Image recognition
- Getting Started
    - Learn linear algebra
    - High computational power

Applications of Machine Learning

- Predictive analytics
- Recommendations (on social media platforms)
- Fraud detection

Why Tall Parents Don't Have Even Taller Children

- Let's use real estate as an example
    - An additional washroom adds more to a housing price than an additional bedroom
    - Proximity to transport infrastructure also increases housing prices
    - Houses close but not too close to malls also had higher prices
- Basically, its being able to observe how individual attributes of a greater whole result in what that greater whole becomes

Lesson 2 Summary:

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

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

Week 3:

Lesson 1:

How is Data Science saving lives?

- Targeted information for healthcare professionals to assess patients
    - Predictive analytics
- Predict natural disasters

How should companies get started with Data Science?

- If unable to measure something, cannot improve thing
- Start recording all sorts of information (Capturing data)
    - Always archive old data, don't get rid of it
    - Consistency, readable
- Apply algorithms and analytics to it

- Have a team od data scientists, not just one

Applications of Data Science
- Changing day to day operations
- Interactions -> generate recommendations
- Shopping, website traffic, activity, etc all contribute
- Netflix keeps track of plenty of data to introduce shows its consumers want before they do
    - Knew their base liked an actor, director, and house of cards, so made the decision to buy house of cards and capitalize on it

The Final Deliverable
- Analytics' purpose is to communicate findings to people who can make use of those findings
    - Tables and plots
- Storytelling is just as important as presenting the facts and data
- Discuss scope of final deliverable before analytics to have them blend well together in the final document

Lesson 1 Summary:
- Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- That companies can start on their data science journey by capturing data. Once they have data, they can begin analysing it.
- Some ways that data is generated by consumers.
- How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

Lesson 2:

How can someone become a Data Scientist?
- Programming, algebra, probability, statistics, and databases (relational)
- Learn things by doing them

Recruiting for Data Science
- Curiosity is on demand (tell a story, other social skills)
- Sense of humor??
- Technical skills

- Platforms:
    - Predictive analytics
        - R
        - Stada
        - Python
    - Unstructured
        - Python
    - Big Data
        - Hadoop
        - Spark

Lesson 2 Summary:

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

Lesson 3:

The Report Structure

- Structure
    - How many pages?
    - Short reports is to the point and presents a summary
    - Longer reports build the argument slowly and has details of other works and intermediate findings along with the main results
- Formatting
    - Cover page (minimum)
        - Title
        - Name of authors
        - Affiliations and contact details
        - Name of publisher
        - Date of publication
        - Table of contents
    - Executive summary
        - Research questions, thesis, hypothesis
    - Detailed contents
    - Results
        - Descriptive statistics
        - Illustrative graphics

        - Move toward formally testing the hypothesis
    - Discussion
        - Craft main arguments
        - Highlight how findings provide the missing piece to the puzzle
        - Often, only parts of the answer are discovered and with a list of caveats at that
    - Conclusion
        - Generalize specific findings
        - Promote findings
        - Identify future possible developments in research and applications
    - Acknowledgements
    - References
    - Appendices (if needed)
- Checklist to reference:
    - 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?

Lesson 3 Summary:

- The length and content of the final report will vary depending on the needs of the project.
- The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
- The report should present a thorough analysis of the data and communicate the project findings.