Week 1: Introduction to Data Science

Day 1: Tuesday 3rd December 2019

The definitions that surround the profession

Lesson 1: What is Data Science

From the pre-course, I hope you are duly motivated towards the prospect of becoming a data scientist. I should warn that while it has been highly publicised as the sexiest job of the 21st century, it requires a balance in the discipline of being systematic and flexibility to be artistic.

Let's look at the definition of data science from a number of top organisations and individuals around the world:

Wikipedia:

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems"

Random Guy on Quora: x

Data Science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies. Mining large amounts of structured and unstructured data to identify patterns can help an organisation rein in costs, increase efficiencies, recognise new market opportunities and increase the organisation's competitive advantage.

Ankit Rathi on Medium:

Data Science is a field where we apply 'science' to available 'data' in order to get the 'patterns' or 'insights' which can help a business to optimise operations or improvise decisions.

IBM:

- Data Science is the process of using data to understand different things - understand the world
- Data Science is to use data to validate any hypothesis or a model
- Data Science is to uncover insight/trends hiding behind data
- Data Science is to translate data into a story that presents insights for strategic decisions
- Data Science is a process and system to extract data from various forms (structured/unstructured)
- Data Science is the study of data and its properties - like biological science is the study of biology or physical science is the study of physics
- Data Science is an attempt to work with data too find answers to questions

By: Akinlabi Ajelabi

Ravi Kannan from Microsoft Research:

Data Science is a subset of of Computer Science that involves the use of computer to understand and extract usable information from massive data in applications for natural science, commerce, social network and other fields.

PWC:

Data Science is the combination of data and analytic tools to give the ability to create the insight that will allows solve the new challenges (urbanisation, accelerating rate of population growth, climate change, global shift in economic & political power and technology explosion) and opportunities of today and tomorrow.

For the purpose of this course, I will like for us to go with the definition from Rachel Schutt from her book "Doing Data Science". She says

"Data Science is to solve problems through data analysis using an appropriate method and to effectively communicate results to relevant stakeholders."

Let's look to break down the definition into its component parts:

- **Solve problem:** Problems are at the heart of what every entity (either a person or business) does. Value is only created when the benefit of a solution to a problem far out weighs the cost of offering the solution. As a data scientist, you will be solving problem, support those solving problems or looking for new problems to solve.
- **Data analysis:** To solve any problem data must be collected. This data might need some form of cleaning, transforming or modelling to become useful toward developing a solution. As a data scientist, you will be required to logically analyse data with the goal to discover useful insight that can lead to optimal solution.
- **Appropriate method:** To apply logic is to be systematic in ones approach to reasoning solutions to a given problem. This systematic approach allows for a consistent way to create order from chaos that usually emanates from problem definition. As a data scientist, you will iteratively apply a fairly standardised method to move from problem definition to solution delivery.
- **Effectively communicate:** However, a solution that is not effectively communicated is as good as to have no solution at all. This is because the benefit of a solution can only make an impact when it is delivered in a concise manner and within the overall context of its application. As a data scientist, you will be required to use various tools to effectively communicate your solutions to whoever can derive value from It.
- **Relevant stakeholders:** Whoever derives value from the solutions delivered is considered a stakeholder. This can either be primary, secondary or tertiary depending on the degree of influence and interest towards the solutions. As a data scientist, you are responsible for the optimal experience of your stakeholders as their interact with you as regards the solution delivered.

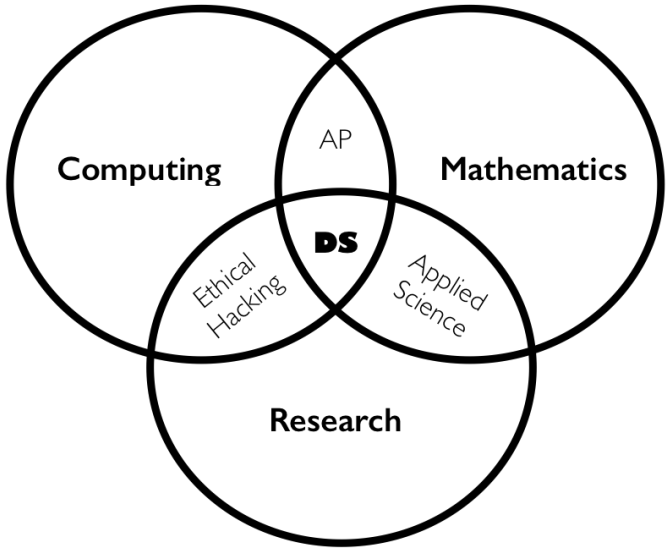
Lesson 2: Who is a Data Scientist

By now you know data science is about solving problems, using appropriate methods, effectively communication and interaction with stakeholders. All in the bid to have a better understanding of the world around us and provide insights for strategic decision making.

Furthermore, it has become an important profession that has garnered a lot of interest as a profession. More importantly, in both academia and industry there has been a surge in the search for a data scientist over the past decade. This is because there is a data deluge being realised from the rise of taking all aspects of our life and turning them to data. This is obvious from how we can quantify our friendship with likes on Facebook, document our random thoughts with Twitter or even leverage the use of micro-sensors to track our fitness using Fitbit.

However, to be able to take on the opportunity that these data present various skillset need to be developed. These skillsets are largely categorised into three fields whose intersection connotes the penultimate attributes of a data scientist

Data Science Venn diagram



*AP : Algorithmic Programming

- **Research** is creative and systematic work undertaken to increase the stock of knowledge and its use to develop new applications.
- **Mathematics** is the study of quantity, structure, space and change to formulate new conjectures that has helped evolve our standards of reasoning as humans.
- **Computing** is activities that involves the use of computers to manage, process and communicate information for various processes.

There are existing overlaps across these fields. Such as an overlap between research and mathematics as led to origination of applied science which is the application of existing scientific knowledge to natural challenges. Also, overlap between research and computing as led to origination of hacking which is the applied skills of a computer expert to overcome a problem. Finally, the overlap of mathematics and computing has led to algorithmic programming which is a procedure written for a computer to follow precisely to solve a problem or reach a goal.

Therefore, as a data scientist you will be involved in daily activities that involves research into existing systems towards building new systems that leverages on computing to solve real-life challenges. In an industrial context, a data scientist

"is responsible to think and execute the data strategy within an organisation, from clearly defining the problem statement, to collecting data, to setting up the infrastructure to process and manage the data, to generating insights from the data, to communicating the insight for strategic decision that would impact business outcomes."

TASK: CREATE A DATA SCIENCE PROFILE OF YOURSELF.

- Rank your skill level for each criteria from 0 (low) to 10 (high)
  - Reading:
  - Critical Thinking:
  - Time Management:
  - Mathematics:
  - Computer Programming:
  - System Design:
  - Report Writing:
  - Listening:
  - Teamwork:
  - Curiosity (Asking Questions):
- Plot a bar chart to help visualise your data science profile

Resources:

1. https://www.youtube.com/watch?v=f9AqD83qHGg
2. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx
3. Doing Data Science: Straight talk from the frontline (Chapter 1: Introduction to Data Science) by Cathy O'Neil & Rachel Schutt
4. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Day 2: Thursday 5th December 2019

The underlying methodology to succeed in the profession

Lesson 1: The Data Science Process

As like any project, the goal of a data science project is its undertaking to achieve result or an outcome.

Following the definition uncovered for data science and the role of a data scientist, you will come to realise that for a data science project a systematic approach is often employed to navigate between uncertainty (where little is known of the problem at hand) and certainty (where an outcome is achieved). This systematic approach is called the data science process/methodology.

The data science process is applied across projects with varying degree of complexity. Examples of data science projects include:

1. Customer Intelligence: Churn Prediction for a Marketing Firm
2. Demand Prediction for a Taxi
3. Automated Machine Learning for an IT Company
4. Fraud/Anomaly Detection for a Bank
5. Recommendation System for an eCommerce Store
6. Creative AI for a Telecommunication Company

A typical data science process usually involves about 6 phases. Below is an overview of the phases I embark on my data science process

Data Science Process

1. **Data Sourcing:** Source data from internal systems and external source through traditional surveys, web scraping and database APIs
2. **Data Exploration:** Explore data collected to gain first intuition into relationships and structure
3. **Data Cleaning:** Clean the data to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data
4. **Data Transformation:** Transform the data by reduction and integration to form consolidated cubes for easy manipulation
5. **Data Mining:** Mine the data to discover intuitive patterns and knowledge that can be actionable towards a business objective
6. **Data Visualisation:** Communicate insights generated from data clearly and effectively through graphical representations

Consideration of time and resources spent on each phase of the process depends largely on the motivation of the project. For example, if the goal of a project is to engineer new data sources to improve the performance of an existing model then considerable time and resources will be spent in first phase of the process. Or, if the goal of a project is to develop a model that has high performance in accuracy and sensitivity in prediction from an organisation's legacy database then considerable time and resource would be spent in cleaning, transformation and mining (modelling).

In reality, the data science process is follows a more iterative than sequential approach as well as cuts across various organisational functions including business analysis, system architecture, software

engineering, product development and IT operations. You will be required to work with frontend units to clearly define the problem statement as part of the data sourcing phase or be required to work with architects and engineers to clearly define functional requirements for the infrastructure of an effective data pipeline.

Most importantly, every data scientist should always ensure that the result/outcome of a data science project is some form of data product. The product would either be in the form of a hardware or software artefact that is standalone or be integrated into an existing system. You may be required to make a business case or justification on how the product will be deployed to achieve tangible business outcomes of either increase revenue or cost reduction.

Lesson 2: Choose a domain for your project

Another caveat, while it true that most data science projects follow a certain process the successful application of this process is largely linked to the knowledge of the field that the data belongs which is known as Domain Knowledge.

Domain Knowledge in data science is like in software engineering in which your result/outcome/data product would only make an impact when there is consideration for domain (problem and information context) in which its used/operates/deployed.

Domain selection can either be as broad or as narrow depending on the motivation behind the project. While we will not speak much of the scale of domain selection, it is important to take note of your reality as regards how much you understand the problem or information context before you embark on implementing your data science process to given project.

TASK: FOLLOWING YOUR MOTIVATION REPORT ON YOUR FINAL PROJECT, CAN YOU STREAMLINE THE IMPACT OF YOUR PROJECT TO TACKLE ANY REAL CHALLENGE IN ANY OF THESE INDUSTRIES IN NIGERIA:

- Finance
- Energy
- Healthcare
- Security
- Education
- Agriculture
- Supply-chain

You are also required to identify at least one book, three journal and ten blog/news articles that can form a good foundation of domain knowledge for the final project.

Resources:

1. https://cdn.oreillystatic.com/en/assets/1/event/292/Practicing%20data%20science_%20A%20collection%20of%20case%20studies%20Presentation.pdf
2. https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492
3. https://towardsdatascience.com/minimum-viable-domain-knowledge-in-data-science-5be7bc99eca9