STUTERN

Intro to Data Science

SGA07_DATASCI

3rd December 2019

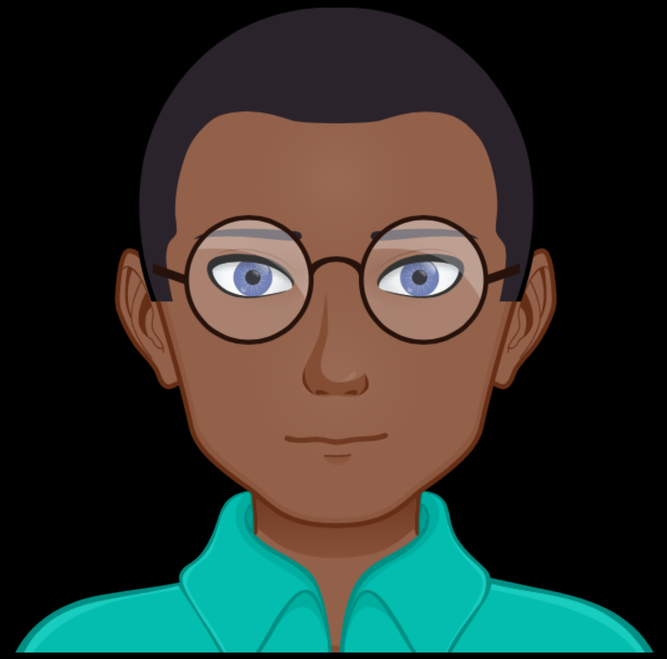
Module Overview

- What is Data Science
- Who is a Data Scientist
- The Data Science Process
- Choose a Project Domain

STUTERN

About Me

- Undergrad in Elect/Elect Unilag
- Postgrad in Data Analytics UoW
- Data Scientist @ Josla
- IBM GEP 2017 - Present



Akinlabi Ajelabi

Outcome

After this Module, you will;

- Clearly define data science
- Understand the components that make up data science
- Find motivation to become a data scientist
- Identify a methodology to deliver as a data scientist
- Review case studies of data science applications
- Review application areas for your data science project

What is Data Science

> 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.
>
> **Wikipedia**

What is Data Science

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.

Random Guy on Quora

What is Data Science

Data Science is the process of using data to understand different things - understand the world

Data Science is to translate data into a story that presents insights for strategic decisions

IBM

Data Science is to use data to validate any hypothesis or a model

Data Science is to uncover insight/trends hiding behind data

What is Data Science

> 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.
>
> **Ravi Kannan from Microsoft Research**

What is Data Science

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.

PWC

What is Data Science

> Data Science is to solve problems through data analysis using an appropriate method and to effectively communicate results to relevant stakeholders.
>
> **Rachel Schutt from Doing Data Science**

STUTERN

Definition Components

Solve Problems

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.

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.

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.

STUTERN

Definition Components

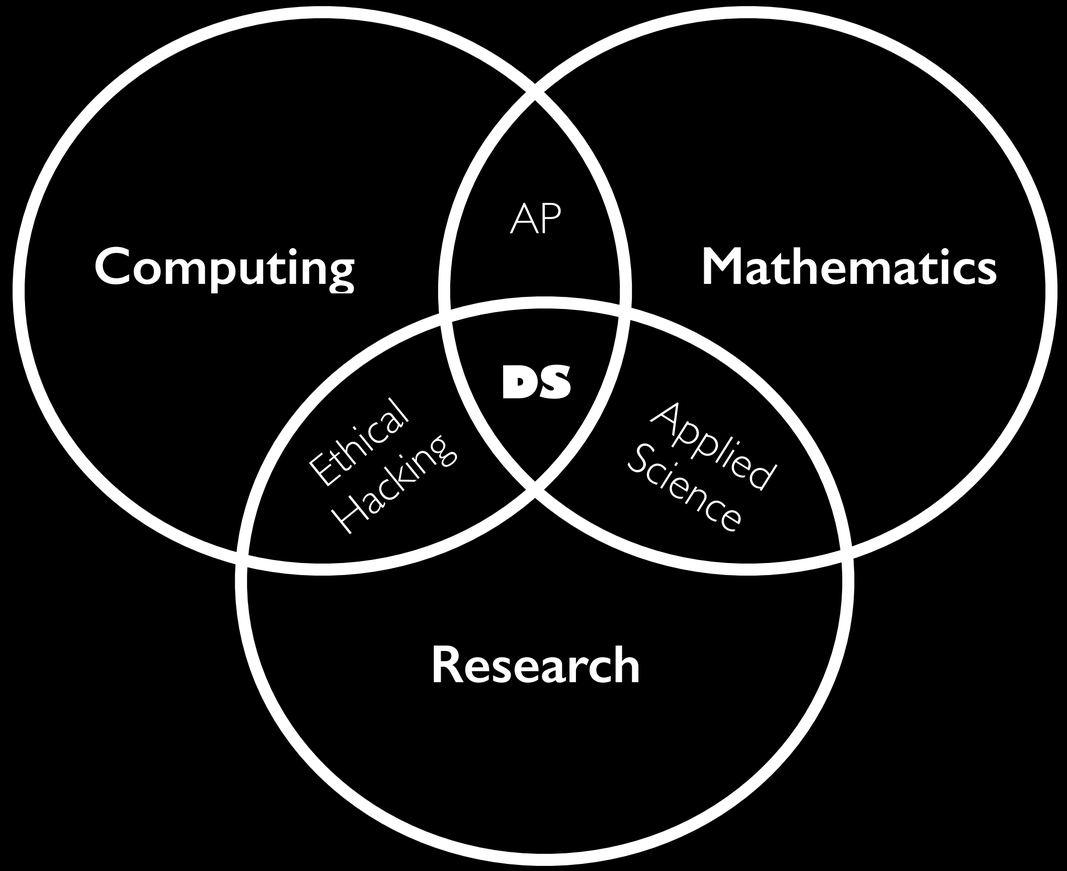
Effective Communication

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 an overall context

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.

Data Science Venn diagram



*AP : Algorithmic Programming

Who is a Data Scientist

Motivation:

- Data is the new oil
- Sexiest job of 21st century
- All about Datafication

> 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

STUTERN

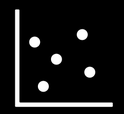
The Data Science Process

Sourcing

Source data from internal systems and external source through traditional surveys, web scraping and database APIs

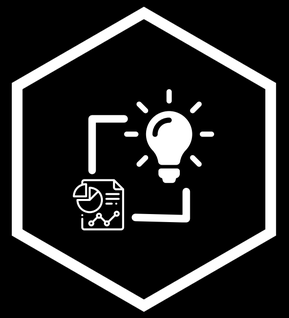
Exploration

Explore data collected to gain first intuition into relationships and structure



Cleaning

Clean the data to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data



Visualisation

Communicate insights generated from data clearly and effectively through graphical representations

Transformation

Transform the data by reduction and integration to form consolidated cubes for easy manipulation

Mining

Mine the data to discover intuitive patterns and knowledge that can be actionable towards a business objective

Case Studies

- Customer Intelligence: Churn Prediction for a Marketing Firm
- Demand Prediction for a Taxi
- Automated Machine Learning for an IT Company
- Fraud/Anomaly Detection for a Bank
- Recommendation System for an eCommerce Store
- Creative AI for a Telecommunication Company

STUTERN

Project Brief

50% of course is your project that validates you as a **Data Scientist!**

Finance

Energy

Healthcare

Education

- **Start** early (at least 75 hours ~ 2 weeks of working effort)
- **Solve** problems with a local context for application
- **Use** multiple data sources (online and field research)
- **Data** should not be too small (KB) or too big (GB)
- **Accompany** report should not be more than 7500 words, 10 pages
- **Deploy** your final product in a cloud environment

Recap/Summary

At the end of this Module, you should understand;

- Data Science is all about solving problems through analysis
- It involves research, computing and mathematics
- It's one of the most sort after and sexiest profession (right now anyways!)
- Requires you to employ a methodology for effective delivery
- Google & Uber are examples of companies borne out of data science applications
- Book keeping as regards your final project work product

Practice Lab

Create a simple table and bar chart to rank your skills as a data science.

Use the following criteria: from 0 (low) to 10 (high)

- Reading:
- Critical Thinking:
- Time Management:
- Mathematics:
- Computer Programming:
- System Design:
- Report Writing:
- Listening:
- Teamwork:
- Curiosity (Asking Questions):

Suggested Material

- https://www.youtube.com/watch?v=f9AqD83qHGg
- https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx
- Doing Data Science: Straight talk from the frontline (Chapter 1: Introduction to Data Science) by Cathy O'Neil & Rachel Schutt
- http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- https://cdn.oreillystatic.com/en/assets/1/event/292/Practicing%20data%20science_%20A%20collection%20of%20case%20studies%20Presentation.pdf
- https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492
- https://towardsdatascience.com/minimum-viable-domain-knowledge-in-data-science-5be7bc99eca9