Introduction to data science concepts

Dhikrullah (OD)

Outline

- What is data science
- What is not data science
- Why learn data science
- Data science and related fields
- Data science in real world
- Data science workflow
- Course objectives and outlines
- Course logistics
- Getting started

What is data science

Data science employs specific methods, algorithms, and processes to derive insights from both structured and unstructured data.

Data science combines scientific methods from mathematics, statistics, and computer science to perform data analytics.

Data science is the application of computational and statistical techniques to address or gain insights into some real world problems.

Data science vs Data analytics

Data science is a broader field that focuses on seeking to answer specific questions by observing trends and exploring heterogeneous sources of data.

The more specialized field of data analytics is a component of the bigger process. From massive databases, data scientists seek to derive **insights and identify patterns**. Unlike data science, data analytics is concerned with finding answers and gaining insights to **existing questions**.

While data analysis focuses on exploring new perspectives on the **known**, data science is specifically combined with the **unknown**.

Data science vs Artificial intelligence

- Artificial Intelligence (AI) is a field of computer science that deals with training machines to be able to mimic the human cognitive activities.
- AI trains machine to be able to see, hear, think, and provide solutions to common problems humans frequently encounter.
- While the goal of AI is to create models that mimic human intelligence, the focus of data science is still on creating models that make use of statistical insights.

Data science vs Machine learning

- Machine learning is a subset of AI that deals with training machines to learn and improve through experiences - without essentially programming them explicitly.
- While ML seeks to learn from data and create estimates and predictions, data science/analytics ultimately seeks to uncover patterns.
- To be precise, data scientists utilize ML as one of their tools to help them draw conclusions from their data.
- ML thus fits into the bigger ideas of data science.

Common skills across different roles

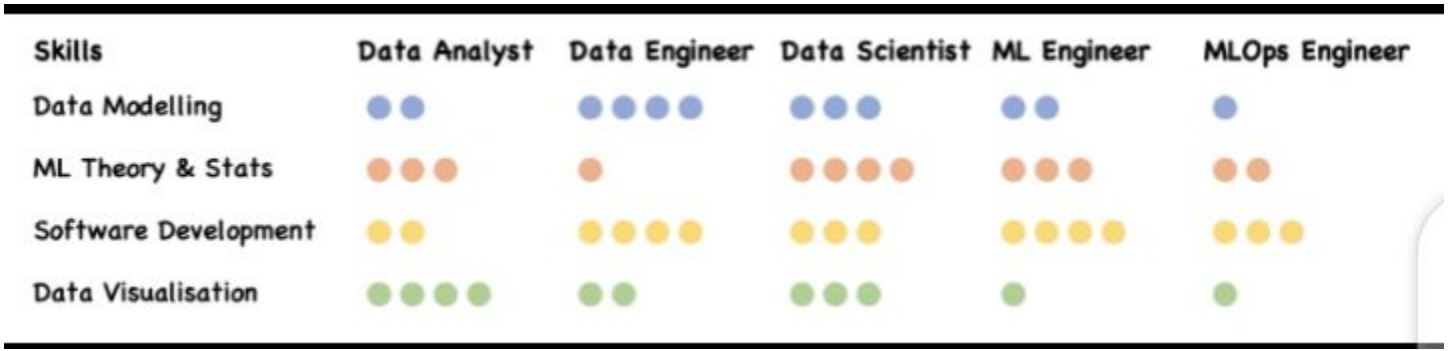


Image source: Towards AI

What then is data science?

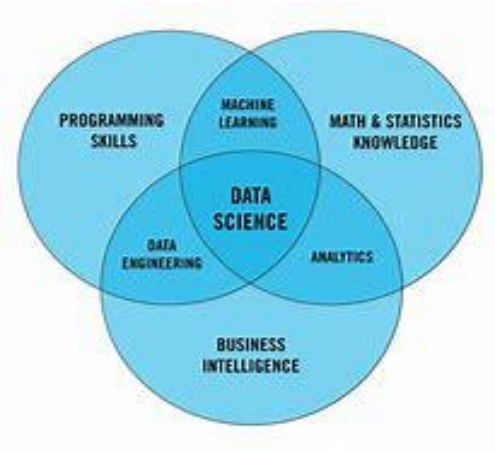


Image source: learn.co

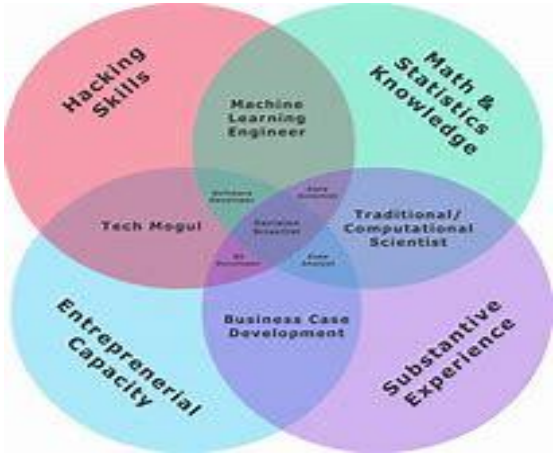


Image source: medium

Why learning data science

- Data science is revolutionizing companies and giving them edge in competitions.
- Data science cut across various industries because data is everywhere now
- Huge discrepancy in the demand and supply of Data Scientists
- Equip you with the skills necessary to analyze large amount of data.
- One of the hottest jobs of the 21st century due to the massive exponential amount of data.
- One of the most lucrative jobs out there today

Data science in real world - Healthcare

Data science is being used healthcare to identify and predict disease, and personalize healthcare recommendations.

“Using data from previous cancer patients, Oncora's software uses machine learning to generate individualized suggestions for current cancer patients. UT Health San Antonio and Scripps Health are two healthcare organizations that use the platform provided by the business. Their radiology team worked with Oncora data scientists to mine 15 years' worth of information from more than 50,000 cancer records on diagnosis, treatment regimens, results, and side effects. Oncora's algorithm learnt to recommend tailored chemotherapy and radiation regimens based on this data.”

E-commerce

By using different data science and analytics technique, data scientist can tailor ads placement to the right audience thereby ultimately adding value to the business.

“Taboola generates engagement opportunities for advertisers and digital properties using deep learning, AI, and massive datasets. Through the placement of adverts across a wide range of online publications and websites, its discovery platform generates additional revenue, viewership, and engagement. A new product or service can be introduced to readers through its discovery platform, along with news, entertainment, topical information, or advice. According to the company's website, it collaborates with publications like USA Today, Bloomberg, Business Insider, and MSN.”

Taboola

Government

Government can leverage the potential of data science and analytics to gain insights to tax patterns of its citizens and effectively tackle the long term debacle of tax evasion.

“According to one estimate, tax evasion costs the U.S. government $458 billion annually, so it seems sense that the IRS has updated its fraud-detection procedures for the digital age. To the chagrin of privacy activists, the agency has increased efficiency by creating multidimensional taxpayer profiles using information from several sources, including public social media data, metadata, email analysis, electronic payment trends, and more. The agency estimates individual tax returns based on such profiles; anyone with significantly divergent real and anticipated returns is reported for audits.”

Transportation and logistics

Transportation uses data science and data processing to **improve the efficiency of routes**. By analyzing driver behaviors, the industry can optimize delivery routes and reduce costs. By comparing data between different routes, companies can improve the efficiency of their operations by creating better transportation planning.

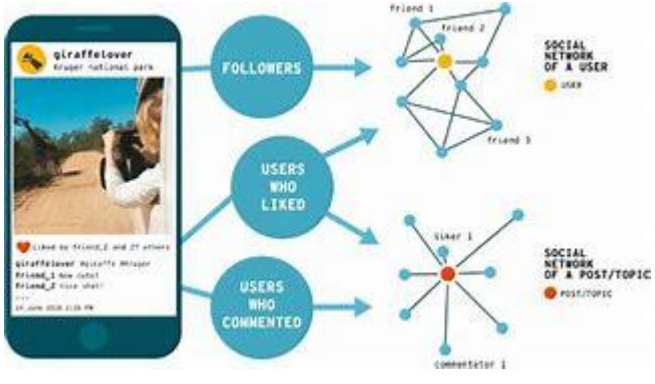
Using Uber as an example, "The data scientists at UberEats have a pretty straightforward objective: having hot food delivered as soon as possible. But to make it happen nationwide, you need staff meteorologists, complex statistical modeling, and machine learning. The team must forecast how every possible element, from storms to holiday surges, will affect traffic and cooking time in order to optimize the entire delivery process.

Social Media

Data scientists can use techniques such as cluster analysis to group individuals with common interests.

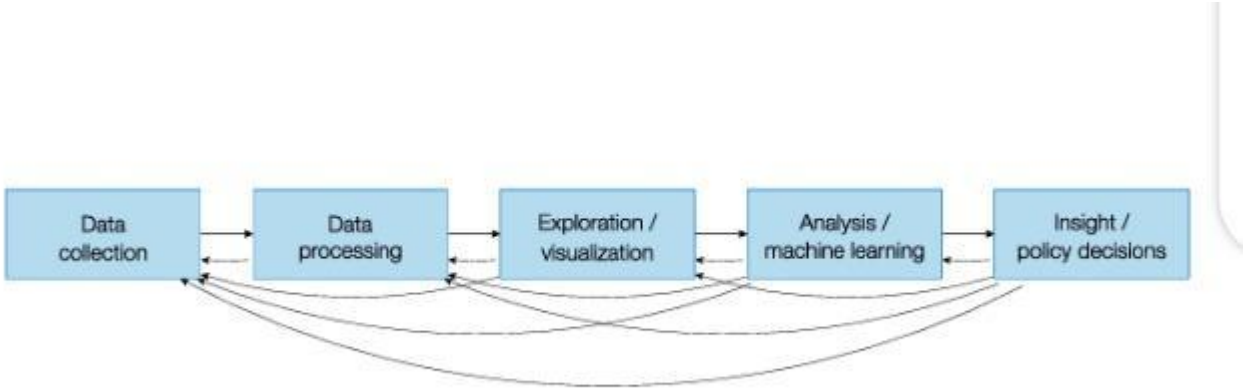
A real world example of this is seen in common social media platforms such as Facebook, Instagram, Twitter etc.

For example, using Tinder as a case study, they employ algorithms that prioritize matches between active users, users near each other and users who seem like each other's "types" based on their swiping history.



Data science workflow

Data science is the iterative cycle of designing a concrete problem, building a model to solve it, and evaluating what insights this provides for the real underlying problem. In order to accomplish this, several steps are taken.



Data science workflow

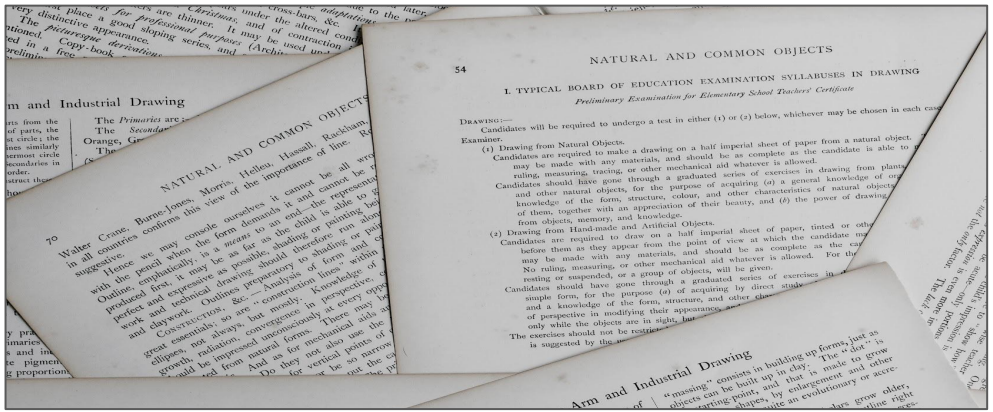
- Business Problem Definition
- Research
- Data Acquisition
- Data Manipulation
- Data Storage
- Exploratory Data Analysis
- Data Preparation for Modeling and Assessment
- Modeling
- Implementation

Business problem definition

Normally, defining the issue and accurately assessing the potential benefit it might have for an organization are nontrivial stages of a big data science project.

Research

Investigate what other businesses have done in a similar scenario. This is seeking out solutions that are practical for your business, even if it means modifying other ideas to fit the needs and resources specific to your organization.



Data acquisition

Data collection is a difficult process step that typically entails collecting unstructured data from many sources. As an illustration, it might entail creating a crawler to gather reviews from a website.

Data manipulation

Once the data is retrieved, for example, from the web, it needs to be stored in an easy to-use format.

Data storage and management

Once the data is processed, it sometimes needs to be stored in a database. Commonly, a relational database.

Exploratory analysis

The data exploration phase is necessary after the data has been cleaned and stored so that insights can be drawn from it. Understanding the data is the goal of this stage, which is typically accomplished using statistical methods and data charting. This is an excellent time to assess the problem definition's logic and viability.

Data preparation

This stage involves reshaping the cleaned data retrieved previously and using statistical preprocessing for missing values imputation, outlier detection, normalization, feature extraction and feature selection.

Modelling

This stage involves trying different models and looking forward to solving the business problem at hand. In practice, it is normally desired that the model would give some insight into the business. Finally, the best model or combination of models is selected evaluating its performance on a left-out dataset.

Implementation

In this stage, the data product developed is implemented in the data pipeline of the company. This involves setting up a validation scheme while the data product is working, in order to track its performance.

Course objectives

- Learn in-demand technical and problem solving skills
- Master programming skills
- Gain proficiency in data management
- Gain proficiency in statistical analysis and hypothesis testing
- Apply data science concepts and methods to solve real world problems
- Communicate findings and solutions to stakeholders/non technical audience effectively.