# Task One – Narrative

(Plan)

Chosen Scenario

NHS Covid-19 Test Data & Vaccine Data

Approaches used in big data analysis

* Modelling
  + Advantages
    - Allows for easy understanding and use of the data we have
  + Disadvantages
    - Overly Approximate relationships can lead to inaccurate or vague conclusions
* Visualization
  + Advantages
    - Easier to spot trends for the average person as well as data scientists
  + Disadvantages
    - Inaccurate sets of data can badly affect the visualisation of the data
* Machine Learning
  + Advantages
    - More data often means better learning and more accurate predictions or diagnostics <https://ieeexplore.ieee.org/abstract/document/7906512>
  + Disadvantages
    - Traditionally, ML algorithms were designed with “the assumption that entire dataset can fit in memory”. <https://ieeexplore.ieee.org/abstract/document/7906512>
* Data Mining
  + Advantages
    - …
  + Disadvantages
    - …

Sources:

<https://www.sciencedirect.com/science/article/pii/S0268401219302890>

(Actual Research Review)

## Introduction

In this research review, I will consider four approaches to big data analysis. Each method will be explained, advantages and disadvantages of each will be considered as well as fundamental limitations all in the context of a problem scenario. The problem scenario chosen is the processing of vaccination and Covid-19 data during the ongoing pandemic done by the NHS in the UK.

## Data Modelling

In big data, Data modelling is stated to be “methods of fast and cost-effective mathematical analysis with approximate relationships between variables” (Galetsi, Katsaliaki, & Kumar, 2020). In our problem scenarios’ case, this can be finding links in our data set between the severity of covid cases, and the vaccinations. The data, once modelled, could indicate that people that have had both doses of a vaccine and a booster dose have a reduced chance of hospitalisation following catching covid, for example.

Data Modelling has the advantage of allowing for easy understanding of data and the relationship between it and other data entities. It can help companies understand their data better to establish how it should use and for what purpose (Mullins, 2021)

## Data Visualization

Another approach is Data visualisation. This can be simply defined as “Creating tables, images, diagrams, and other intuitive ways to understand data” (Galetsi, Katsaliaki, & Kumar, 2020). In context, this can be interpreted to be graphs to show the current daily positive cases for covid, for example or it also could be a box diagram to illustrate the vaccine uptake in the population including first, second, and booster doses.

An advantage of this method is that it makes it much easier for the average person to understand. Other methods tend to be much less accessible to the public as the trends are less obvious often appearing in numbers rather than being illustrated as seen here. On the flipside, trends can be skewed due to inaccurate data. This is true for most big data techniques but especially so here as it is much more obvious in graphs and charts and can appear more skewed than in other approaches. Furthermore, some can only look at the important data without taking into consideration the bigger picture. This can lead to biased results in cases.

Overall, this method’s strengths lie in the ease of illustrating information and sharing it with almost anyone which is important when thinking about Covid-19 data. However, its limitations are that it is sensitive to inaccurate data which must be removed before the data can be processed. It can also lead to biased results as discussed formerly.

## Machine Learning

Thirdly, we have the machine learning approach. It is defined as “AI aimed to design algorithms that allow computers to evolve behaviours based on empirical data” (Galetsi, Katsaliaki, & Kumar, 2020). These algorithms “learn from data and provide data driven insights, decisions, and predictions” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017). Such Algorithms could be used to diagnose the public. They could enter their symptoms and be told likely positive or negative and be told to take a test.

An advantage of ML is the “presumption that algorithms can learn better with more data” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017). This is perfect for big data as they have access to large datasets and therefore can make more accurate predictions and insights. This comes with drawbacks, however. Traditional ML algorithms “were designed for smaller datasets” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017) assuming that “the entire dataset can fit in memory”. In big data, this is not practical nor true as the datasets are simply too sizable.

This method’s strengths lie in the fact that the more data it is supplied with. The better the predictions are. This is important when predicting data for any application including our own scenario. However, while the memory can cause a limitation as discussed previously, some of these have been alleviated with “paradigms” and “frameworks” such as MapReduce and Hadoop (L'Heureux, Grolinger, Elyamany, & Capretz, 2017)

## Data Mining

Finally,

## Summative Table

# Task Two – Analysis

## Introduction?

……

## Section I: Data Summary, Understanding and Visualisation

### Task 1; Missing Values:

…….

### Task 2; Summary Statistics:

……..

# Bibliography

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big Data analytics in health sector: Theorical Framework, Techniques and Prospects . *International Journal of Information Management* , 206-216.

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017, April 20 ). *Machine Learning with Big Data: Challenges and Approaches.* Retrieved from IEEE Explore: https://ieeexplore.ieee.org/abstract/document/7906512

Mullins, C. S. (2021, May 3). *The Importance of Data Modeling in a Big Data World.* Retrieved from Database, Trends and Applications: https://www.dbta.com/Editorial/Think-About-It/The-Importance-of-Data-Modeling-in-a-Big-Data-World-145915.aspx