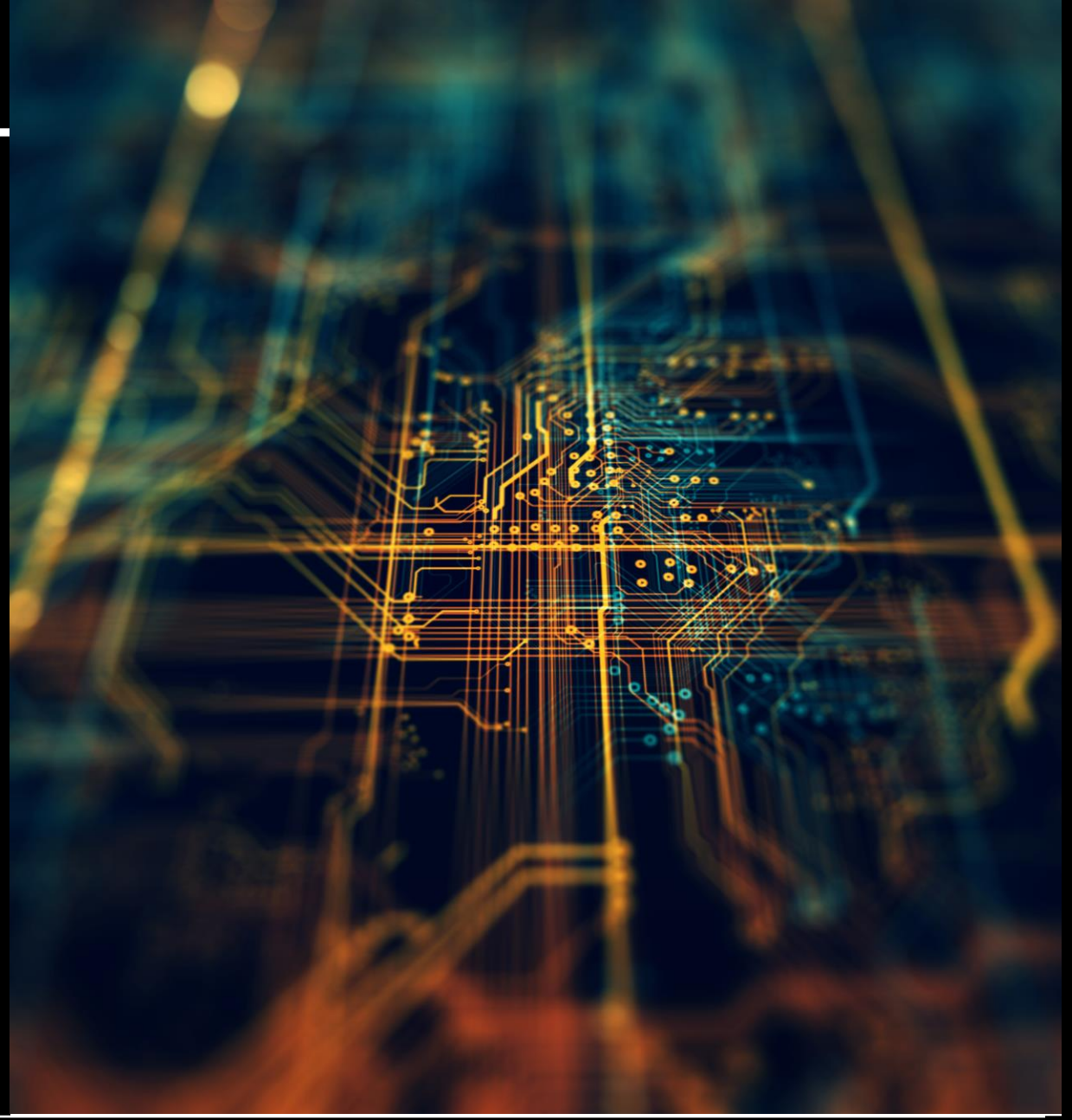
AI MODEL DEVELOPMENT

CMP-5366
DATA MANAGEMENT AND MACHINE LEARNING OPERATIONS

BIRMINGHAM CITY UNIVERSITY

KACPER POPIS – 23161791
KACPER.POPIS@MAIL.BCU.AC.UK



OUTLINE

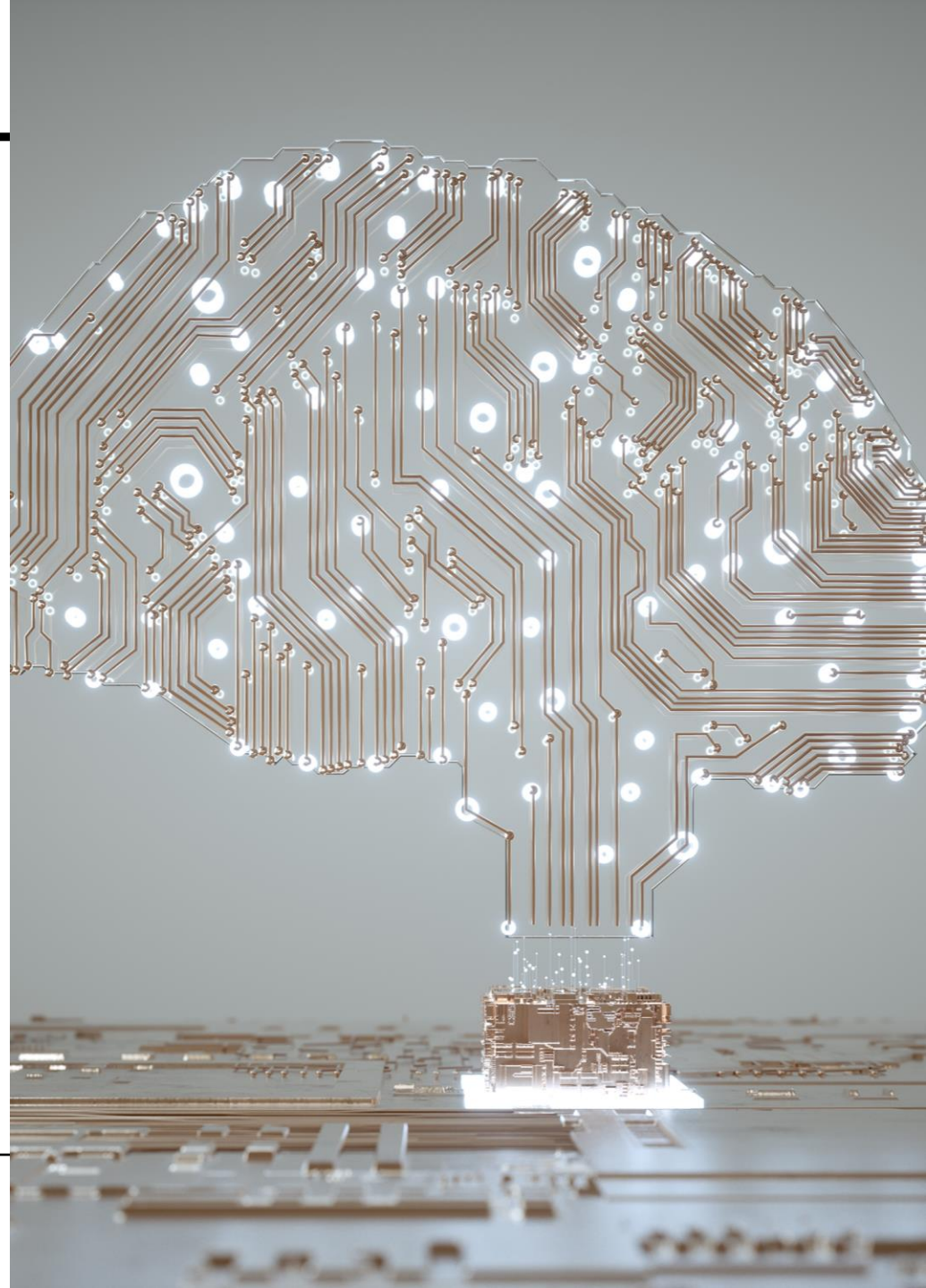
- Introduction
- Candidate 1 Dataset (Dataset A)
- Candidate 2 Dataset (Dataset B)
- Selected Dataset & Data Storage Plan
- MLOps Pipeline Plan
- Self Reflection
- Conclusion & Further Plans

INTRODUCTION

The main purpose of this AI model is to train the model on real and factual data to have the most accurate prediction model as possible

Another purpose of this model would be to compare how well the claims made by the researchers of their respective reports hold up against the predictions made by the model

The storage plan for this project will also be explored and all the development steps for the model will be outlined to show how an AI model will be created



CANDIDATE 1 DATASET (DATASET A)

- As mentioned in Task Sheet 1, Dataset A isn't the dataset that was used in the scientific report of Candidate 1 (A)
- The dataset is one that is found on Kaggle, and has a large collection of video games and different features about them that can help with building an AI model
- More importantly, as this dataset is being used as a replacement for the dataset that was used in the report, the data that will be used from this dataset will be limited to the time outlined in the report (February 2006 – November 2016), and any data that exceeds or precedes this time will be dropped (not used)
- As this is a replacement dataset, the validity and accuracy of the models that will be developed will be compared to the models used in the scientific report – if this dataset were to be chosen

CANDIDATE 2 DATASET (DATASET B)

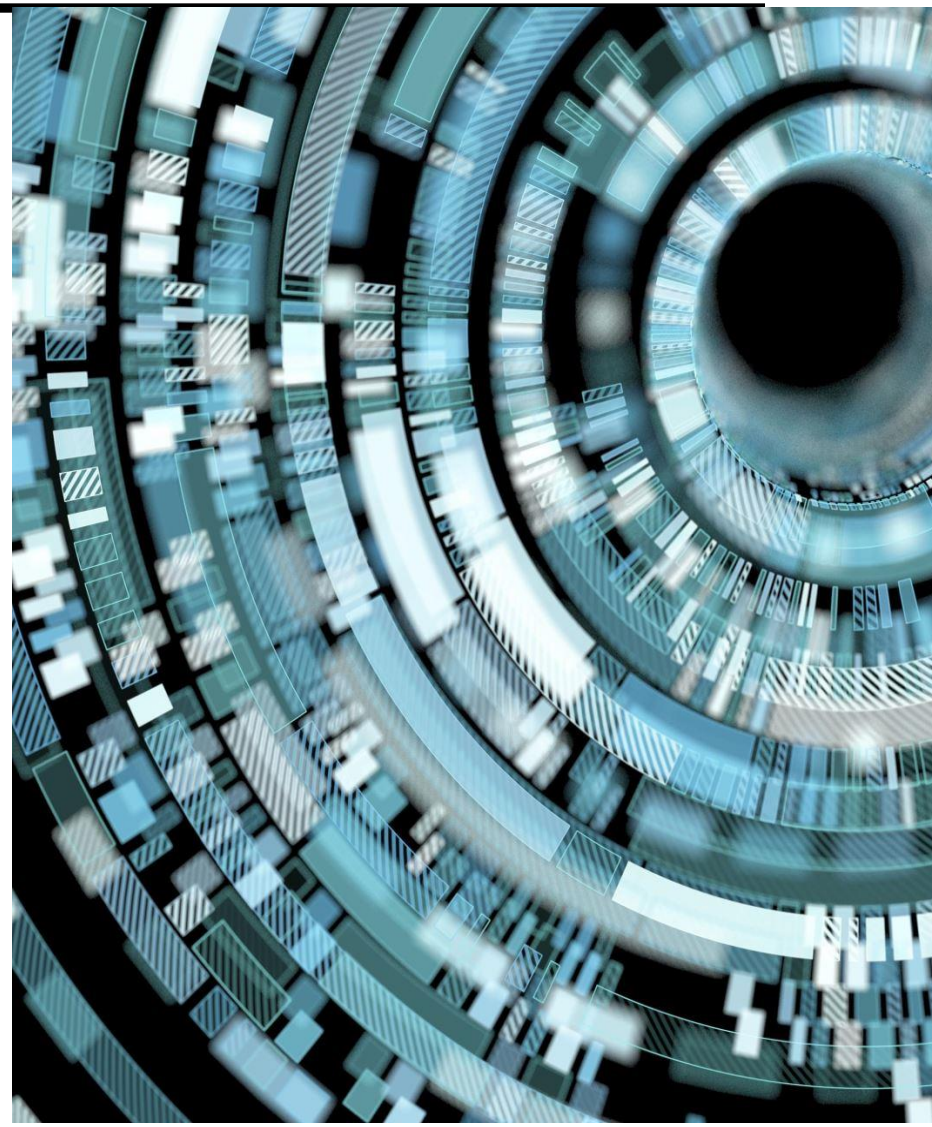
- As mentioned in Task Sheet 1, Dataset B is the dataset that was used in the scientific report of Candidate 2 (B)
- The dataset is one that is found on the page where the scientific report is posted (Figshare), and has a collection of 5 different tables with each table some unique features and some linked features
- The scientific report does not use its dataset to train and develop and AI model, instead its only findings that the researchers have done and any predictions that they have speculated based on the collected data – if this dataset was to be chosen, then the accuracy of the predictions made by the researchers will be tested against the model itself
- This dataset contains over 8 million rows per table which will provide the model will a large amount of data to be trained on and will be able to complete more complex predictions and provide a greater accuracy

SELECTED DATASET & DATA STORAGE PLAN

- The selected dataset for this AI model project will be candidate 2's dataset (Dataset B)
- One of the reasons why this dataset is chosen is because the dataset that is provided with the dataset comes from a factual background and contains large amounts of data (compared to the other dataset) which will allow the model to go into great detail
- Another reason why this dataset is being chosen is due to the other candidate's dataset not being made public therefore making the credibility of the report and the purpose of this model null
- For this AI project the most efficient approach (both time wise and processing wise) to storing the data would be with the ETL approach
- E – Extracting the data from the original source (which in this case is the Figshare website of the candidte's report) and downloading it to the device
- T – Each table from the dataset will undergo the preprocessing stage to make it usable by the model while also removing any outliers / invalid data
- L – Loading the **transformed** dataset into a SQL database (MariaDB) - while also maintaining the links between the tables – so the AI model can use the data

MLOPS PIPELINE PLAN

- As previously mentioned, the dataset will undergo the ETL process to allow for the data to be usable and have **ONLY** valid data be used on the model
- The next step after preparing the data for the model would be developing the actual AI model, for this project a variety of regression techniques will be utilised (like linear or random forest regression) and comparing the accuracy of each
- Once the AI model has been trained and is functional, the model will allow the user to provide various inputs to then predict statistics based on the user data



SELF-REFLECTION

- One of the major restraints / issues that I find in this project is the amount of data that is provided in the chosen dataset as each table contains millions of data which does have an impact on the execution time of different processes
- Another one for this project would be that I am utilising technologies that are new to me such as VM (Virtual Machines) and their environments and working and developing on Linux which slows down the project as there are new techniques and apporaches to solve issues

CONCLUSION & FURTHER PLANS

- Overall, I am satisfied with my choice in dataset as it provides a challenge for me alongside the usage of new technologies and applying my old knowledge to help navigate the different task sheets
- The next steps for this project will be to complete the training of the model and compare the different regression techniques to be able to justify a "best" technique
- After training the model, there will need to be an "access point" created for users to utilise the AI model – all these developments will be documented for all the task sheets

THANK YOU

Kacper Popis
23161791
kacper.popis@mail.bcu.ac.uk