Data Fellowship Coach Hackathon Guide

Structure of Hackathon - For Coach

A hackathon consists of four parts: opening ceremony, hacking, judging, and closing ceremony. It is important for the coach to set the **tone** of the hackathon by setting expectations in the opening ceremony.

The aim of these hackathons is to help prepare apprentices for their synoptic project, where they will have to handle unseen data and complete this on their own.

This new structure allow apprentices to build confidence in

- Encouraging apprentices to be creative
- Exploring and handling unclean data and unclear outlines
- Stretching their current knowledge by practising new skills gained from bootcamp
- Working as a team

Run of Show (Schedule)

1. Opening ceremony (~15 mins)
2. "Hacking" - apprentices build project
3. Add as many breaks as needed (for coach to set)
4. Team presentation & judging
5. Judging debate (to decide winner)
6. Closing ceremony (~15 mins)

Ceremony - For Delivery

Before Opening Ceremony

- Set up music as people enter in
- Think about what tone you want to set for the hackathon
- Have slides up to introduce hackathon and datasets

Opening Ceremony

What is a hackathon?

A hackathon is an event for participants to build projects from scratch within teams, developing new knowledge and stretching their existing skills in a time frame, which is usually 24 hours. Projects created in hackathons are often prototypes and achieve real-world impact. One of the most famous examples is a hackathon project called Workflow, built at MHacks, which went on to the now known Apple Shortcuts app.

Set expectations

True to the nature of hackathons, we want the apprentices in teams to practise creating a project from start to finish with unseen datasets. This is a learning opportunity for the apprentices to develop and practise new skills gained from the bootcamp. We're not looking for a polished project, but how the apprentices apply the techniques learnt to a new project.

Please have a look at the diagram on the slide with the 'how to build a project'. Encourage the apprentices to focus on a small achievable goal for their project that is meaningful on its own but can be scalable. Only after this can they aim for a bigger goal rather than having a grand goal to start with and not having a complete project at the end of the hackathon.

What to do?

The apprentices will build a project based on one of three themes (or more) decided by the coach. From here, apprentices must pick one of the three datasets given in and try to adhere to their chosen theme/s. For more details, see Datasets and Themes.

Closing Ceremony

- Use the slides to present the winners, and dependent on coaches' judgement, to also present runners up, and/or highly commended.
- Be clear what the team did well (i.e. the ideal thing to do for synoptic project)
- Time set aside to gather thoughts/wrap up with feedback from apprentices
- Complete the Session Attendance Log

Themes and Datasets

Themes

The three themes, decided by the coach, replace the previous project briefs and can be adapted to fit with the company's mission or related to the apprentices' industry. This change should increase the apprentices' confidence in their own autonomy.

The three themes are:

- Upskilling (placeholder - feel free to use your own themes)
- Environmental (placeholder)
- Wildest/Funniest (placeholder)

Datasets

The three datasets used can be the ones listed below, or for a more catered hackathon, datasets provided by the apprentices' organisation.

If groups find it difficult to come up with their own data plans and aims, use the project briefs below based on the dataset chosen to give guided questions.

--IMPORTANT--
The datasets marked with **Project Brief, Guided Questions** or **Outline** could be given as additional information to help guide the apprentices ONLY if they are stuck. Datasets with **Project Brief** are further categorised by **DA** (Data Analytics Bootcamp) and **DS** (Data Science Bootcamp) to clarify the level of the project briefs.

1. Olympics - **Outline**
2. Netflix Movies and TV Shows - **Guided Questions**
3. COVID-19 World Vaccination Progress - **Guided Questions**
4. The Museum of ModernArt (MoMA) Collection - **Outline**
    a. Artworks Dataset
    b. Artists Dataset
5. Chinook Database - **Outline** (SQL Database)
    a. You can access this database using pgadmin. Please follow the instructions below:
        i. URL: http://delivery-pgadmin.multiverse.io/
        ii. Username: pgadmin4@docker
        iii. Password: pgadmin4
6. NOAA - **Project Brief DA & DS**
7. Weatherbase - **Project Brief DA & DS**
8. Open Postcode Geo - **Project Brief DA**
9. Commercial and Industrial Property Stats - **Project Brief DA** (*really old dataset*)
10. London Datastore - **Project Brief DA & DS**
11. Land Registry - **Project Briefs DS**
12. ONS - **Project Briefs DS**
13. Prescribing data - **Project Briefs DS**
14. Mental Health Statistics - **Project Briefs DS**

Judging Criteria

Use the judging criteria to give more actionable feedback and help apprentices understand how the synoptic projects will be assessed. Remember to use WWW (what went well) and EBI (even better if). Judging criteria is based on the synoptic project.

Criteria:

- Data Discovery/Plan
    - Review of key fields used
    - Models considered
- Data Preparation
    - Ensure data quality/integrity
- Data Analysis
    - Building the report/analysis
    - Highlighting any obstacles that occur
- Modelling
    - Models, Charts, Report
- Refine and Compare
    - Testing, verification and validation
- Communication & Implement
    - Presentation skills
    - How they handled Q&As
    -

Additional Information for Datasets

Dataset 1 - Olympics

This dataset provides an opportunity to ask questions about how the Olympics have evolved over time, including questions about the participation and performance of athletes, different nations, and different sports and events.

Dataset 2 - Netflix Movies and TV Shows

Some of the interesting questions which can be performed on this dataset

a. Understanding what content is available in different countries
b. Identifying similar content by matching text-based features
c. Network analysis of actors / directors and find interesting insights
d. Is Netflix increasingly focusing on TV rather than movies in recent years?

Dataset 3 - COVID-19 World Vaccination Progress

Some potential questions to be asked:

a. Which country is using what vaccine?
b. In which country is the vaccination programme more advanced?
c. Where are more people vaccinated per day? But in terms of per cent of the entire population?

Dataset 4 - The Museum of ModernArt(MoMA) Collection

a. Artworks Dataset

The Artworks dataset contains 138,151 records, representing all of the works that have been accessioned into MoMA's collection and catalogued in our database. It includes basic metadata for each work, including title, artist, date created, medium, dimensions, and date acquired by the Museum. Some of these records have incomplete information and are noted as "not Curator Approved."

b. Artists Dataset

The Artists dataset contains 15,222 records, representing all the artists who have worked in MoMA's collection and have been catalogued in our database. It includes basic metadata for each artist, including name, nationality, gender, birth year, death year, Wiki QID, and Getty ULAN ID.

Dataset 5 - Chinook SQL Database

The Chinook data model represents a digital media store, including tables for artists, albums, media tracks, invoices and customers. Media related data was created using real data from an iTunes Library. Customer and employee information was manually created using fictitious names, addresses that can

be located on Google Maps, and other well-formatted data (phone, fax, email, etc.). Sales information is auto-generated using random data for a four year period.

Datasets 6 & 7 - NOAA & Weatherbase

DA Bootcamp:
The CMO for a travel company is looking to target 25-35-year-olds who are concerned about climate change. The company has set itself the goal of reducing air travel for customers by 37% in 2020. The executive sponsor has asked for your recommendation on the top 3 destinations for summer 2020, that does not involve a flight, based on the following criteria:

- Average monthly temperature above 22 degrees Celsius
- Average monthly rainfall less than 60mm
- Less than 4hrs from London

DS Bootcamp:
A travel company is concerned about the potential effect of climate change on their future bookings. Their major concerns centre around changes in average/maximum temperatures as well as the amount of rainfall experienced. Their current top five destinations are:

- Sydney, Australia
- Venice, Italy
- Paris, France
- The Dead Sea, Jordan
- Malmo, Sweden

The COO has asked you to investigate whether these two types of climate change will affect their ability to offer bookings to these destinations.
Extend your Project - Are there new territories to target, based on the climate changes identified?

Datasets 8 & 9 - Open Postcode Geo & Commercial and Industrial Property Stats (DA Bootcamp ONLY)

The CTO of an on-demand delivery startup based in Greater London is looking to open the company's first operational facility. The company uses a fleet of electric scooters to provide a "last-mile" delivery service to Fortune 500 companies in the city (e.g. important packages, laundry service, the CEO's sushi order). The executive sponsor has asked you to make a recommendation about where to open the first location, based on the following criteria

- Lowest rent
- Within 20 minutes of the City and Canary Wharf
- Low emission zone

Dataset 10 - London Datastore (DA Bootcamp ONLY, for DS see below)

The CEO of the Department of Health is putting together a spending recommendation for 2020. The department is particularly concerned about rising levels of obesity, access to nutrition and air quality across Greater London. You have been asked to put together a report analysing the top health issues per borough, based on the following criteria:

- Which boroughs should receive more investment?
- How would you prioritise spending across the city?
- What health issues should the Dept. of Health focus on?

Datasets 10, 11 & 12 - London Datastore, Land Registry & ONS (DS Bootcamp ONLY)

A Birmingham based property developer is looking to expand their portfolio into the London market rentals, their potential options (in order of preference) are:

- Flipping the property (renovate and sell within 3 month period) - assume a BTL with a minimum LTV of 70% and a mortgage with 4.5% annual interest
- Corporate (long term) rental - no management fees but £3500 placement fee
- Assured Shorthold Tenancy 10% rental income as management fee

Which option should the developer take in order to make the most profit over a three year period?
Extend your project - What are the three property characteristics they should prioritise in their search?

Datasets 13 & 14 - Prescribing Data & Mental Health Statistics (DS Bootcamp ONLY)

The QA Officer of NHS England is concerned that the future provision of Mental Health services will be compromised due to the currently used, but under-estimating, formula of:

- New funding = Old funding * 1.012

You have been asked to determine which of the London boroughs are likely to require the largest funding increases over a five year period.
Extend your Project - how would your recommendations differ if you were considering physical health