Data Science Skills Exposure Hackathon Guidelines

The following guidelines and resources were created to aid the planning of a data science skills exposure hackathon, as part of the hackathon programme initiative of the IAU Office of Astronomy for Development under the "Knowledge and Skills from Astronomy" flagship in collaboration with IDIA and DARA Big Data, who are all partners of the Hack4dev project. These guidelines are the result of the experience gained and lessons learned from running hackathon events since 2020. Other hackathon events may take different formats, these are simply guidelines which can be adapted accordingly.

Background

The hackathons that have been carried out as part of the programme have targeted university students (both undergraduate and postgraduate) as well as young professionals who have a keen interest in data science and machine learning and are equipped with Python programming skills (a basic level is sufficient). However, the content of the hackathon would benefit anyone looking to gain exposure to data science and machine learning techniques, as well as hands-on experience working on a project that makes use of these skills. Events such as these can be organised by anyone, be it a faculty member, student, postdoc, or administrator at a tertiary education institution, a high-school teacher, or a member of a coding group/club. All the resources necessary to run a hackathon are available in the 'Resources' section below. These resources are described in the guidelines along with other suggestions, details and points to consider when implementing such an event.

With the technological advances resulting from the Fourth Industrial Revolution (4IR), there is now a vast amount of data that can be valuable in both academic and industry pursuits. The growing need for individuals who can analyse and interpret large amounts of data has led to a surge in opportunities in the data science field. From data engineers, to machine learning specialists, these skills are now highly sought after by most companies in industry, and are becoming increasingly valuable in academia. The demand for professionals in the field is therefore high and is expected to continue to grow in the future. Gaining exposure to the skills and techniques of the field can therefore be a crucial opportunity for students and working professionals to access the abundance of opportunities in the data science space.

The guidelines are set out in the following sections:

1. What is a Data Science Skills Exposure (DSSE) hackathon?
2. Requirements overview
3. Actions required in order to implement a data science skills exposure hackathon event
4. A detailed guide for the organising team

5. Resources that aid the running of a DSSE hackathon

NB: All code in the hackathon tutorials is written in Python and are therefore aimed at individuals who have experience in Python programming.

What is a Data Science Skills Exposure (DSSE) hackathon?

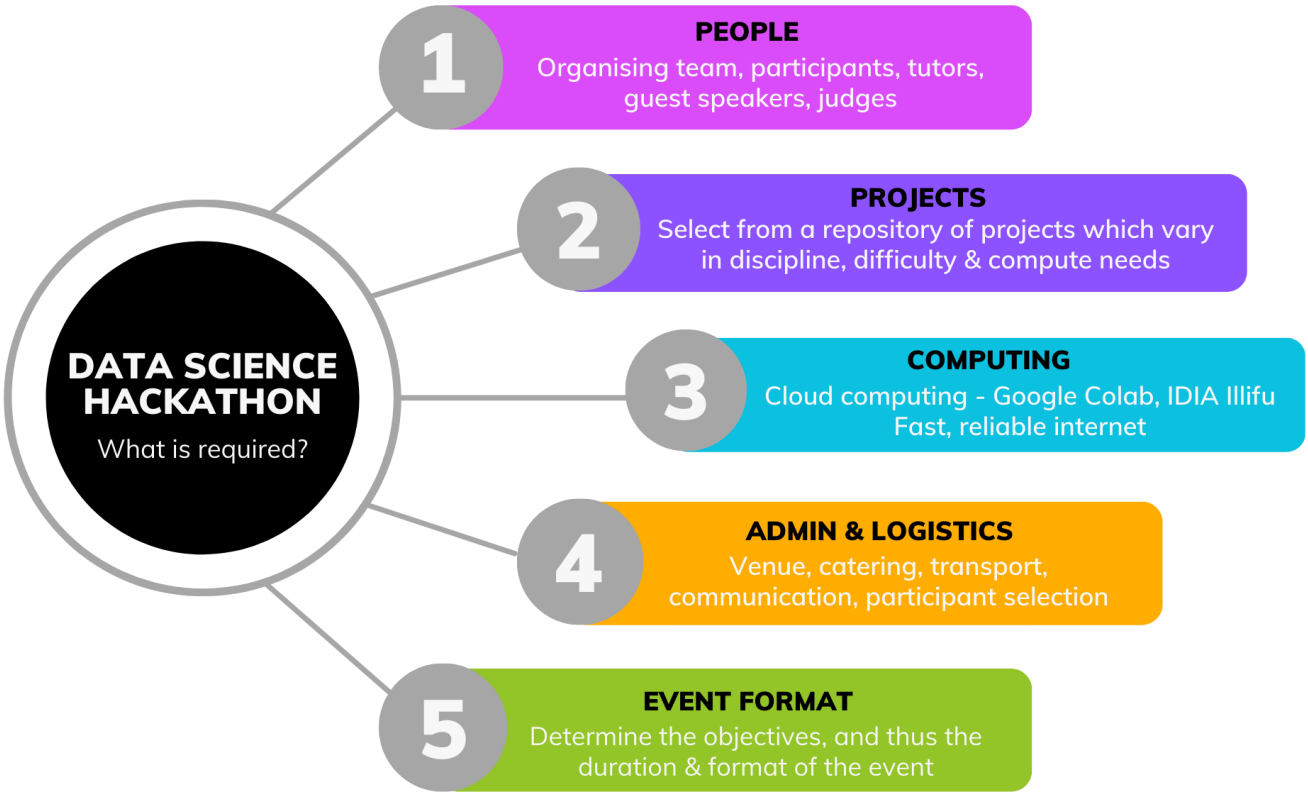
The aim of a DSSE hackathon is to provide exposure to data science and machine learning, and some of the techniques involved, in a friendly and supportive environment. Participants at the hackathon work on interesting real-world problems which are data-driven. The hackathons are usually aimed at final-year undergraduate students, postgraduate students, as well as young professionals in Science, Technology, Engineering & Mathematics (STEM) fields who have a keen interest in data science and machine learning, and who would benefit from further exposure and hands-on experience. There is a diverse range of projects ("hacks") which are available for hackathons, some related to development, e.g. flood detection using satellite images, and others related to academia, industry and commercial applications.

A conventional, tech-related hackathon is an event at which computer programmers, and sometimes others involved in software development, come together to innovate and problem-solve over a short period of time, typically two to three days. The competitive aspect is usually a big factor of traditional hackathons and can result in a tense, high-pressure environment, but this need not be the case! See this resource created by Open Data Day on how to run a successful hackathon of the conventional kind. The focus of a DSSE hackathon is on learning rather than competing or solving a difficult problem. The desired outcome is to pique or deepen interest in the field of data science and machine learning and boost the confidence of participants in their ability to enter this field. This wider outlook on their study and/or career options could then be a stepping stone to allow for access to a greater number of job/internship opportunities going forward, as well as create a pipeline into programmes which are more technically advanced, such as the Deep Learning Indaba in Africa.

Summary of the objectives of a DSSE hackathon:

- Provide exposure to techniques used in data science and machine learning through hands-on experience working with real-world projects
- Broaden study/career prospects
- Encourage peer learning
- Development of skills such as leadership, teamwork and presenting
- Encourage further participation of females in the field of data science

Requirements Overview



What are the actions required in order to run a DSSE hackathon event?

Below is a checklist of the suggested actions required in order to implement a DSSE hackathon. The actions checklist is chronological and colour-coded according to the 5 sections of the requirements diagram above.
NB: Detailed guidelines for these actions follow in the next section.

- ☐ Establish a team of organisers and tutors for the event
- ☐ Determine your target audience and objectives for the event
- ☐ Based on your objectives, determine the duration of your event, i.e. a 3-day hackathon or a 5-day school/workshop.
- ☐ Prepare a list of budget items and secure any necessary funding. *NB: A minimal budget, i.e. making provision for just catering and prizes, can be used in cases where funding is limited.*

- [ ] Secure a venue for the event: e.g. A school, tertiary institution or a private venue potentially offered by a local partner/funder.
- [ ] Ensure that participants have access to computers or that they can bring their own.
- [ ] Determine which hackathon project/s (hacks) you will be running at your event
- [ ] Ensure that participants will have access to the required computing resources, i.e. such as cloud computing.
- [ ] Confirm tutors for the hackathon
- [ ] Secure catering for the event (lunch and tea breaks)
- [ ] Decide on selection criteria for participants of the event and make these clear in the call for applications and event advertising
- [ ] Send out a call for applications at least 6 weeks prior to the start of the event
- [ ] Identify potential speakers: speakers could give an overview of data science, machine learning, or applications of these in industry/academia.
- [ ] Decide on selection criteria for participants and notify successful applicants at least two weeks before the event and send out any necessary preparatory resources
- [ ] Conduct a tutor training session on the chosen hackathon projects at least two weeks prior to the event and a check-in meeting the week before the event
- [ ] Secure prizes for the top 2 or 3 teams
- [ ] Assemble a judging panel
- [ ] Finalise the event programme at least one week prior to the event and send out to all invited speakers and participants.
- [ ] After the event, send out a feedback survey to assist in the analysis of the impact of the event and aid in the planning of future events
- [ ] Share resources that participants can use to further their knowledge of data science and machine learning

Detailed Guidelines:

People:

There are many people involved in the planning and running of a hackathon event, from the organising team, tutors, tech support to the participants themselves. Below are a few guidelines on this most important resource and the relevant responsibilities/requirements.

Organisers

It is advised that a dedicated team responsible for planning and implementing the hackathon should be established from the onset. You may require a team of a few individuals, where each is able to carry out particular tasks, e.g.
Administrative - Event advertising, participant communication

Technical - Sourcing speakers and tutors, access to compute infrastructure, deciding on a project
Logistical - Venue, transport, catering

Participants

The target audience of a skills exposure hackathon could consist of high school or university students, working professionals, university lecturers, or even a mixture thereof! The only requirements are a working knowledge of Python programming and a keen interest in data science and machine learning.

- Determine your target audience who would benefit from the planned event, i.e. would the school/workshop/hackathon be targeted at students (if so, at what level), educators, working individuals who would benefit from upskilling, or a combination thereof?
- Participants should have some level of Python programming experience. In cases where the target audience does not meet this requirement, please see the links to learning resources for Python in the "Resources" section below which can be shared with participants ahead of the event.
- If a selection process is necessary, the following selection criteria can be used to determine the successful applicants:
    - Their level of interest in data science and machine learning
    - A strong motivation for attending the event
    - What they hope to learn/gain by attending the event, i.e. to ensure that what they hope to learn/gain is in line with what is being covered in the introductory talks or the hackathon projects.

Tutors

In order to run a DSSE hackathon where participants receive ample guidance, tutors are required to assist participants with the hackathon tutorials and the hackathon task. A ratio of one tutor for every two hackathon teams (consisting of 4-5 participants) is favourable to ensure that tutors are not overloaded and that participants receive assistance timeously.

- Tutors should be familiar with Jupyter Notebooks, have ample Python programming experience and some prior experience in the application of machine learning techniques (if it is used in the chosen project/s).
- Tutors must be willing to commit ~ 3-5 hours of preparation before the event to go through the hackathon tutorials and ~ 8-10 hours per day during the hackathon. The nature of the hackathon is such that it demands a lot of work from participants in a short space of time; a similar demand will therefore also be experienced by the tutors should the participants need assistance.

- Tutors should work through project tutorials and respective training videos from ~2 weeks prior to the event to ensure that they understand the content and are able to run the tutorials successfully.
- Tutors should familiarise themselves with the computing platform on which the hackathon project/s will be run and ensure that they are able to successfully run all tutorials.
- In the case of virtual events, tutors are encouraged to use platforms such as Slack, Zoom, Discord and WhatsApp groups in order to aid participants.
- Due to hackathon time constraints, it is recommended that tutors assist participants in understanding the purpose/function of the code in the tutorials, rather than focusing on syntax.
- Tutors should also urge their teams to complete the project tutorials in the time allotted according to the event programme so that sufficient time remains to work on the hackathon task.
- Tutors may need to keep an eye on teams throughout the event to ensure that they are working well together and to encourage team members who are not contributing as much, which may stem from a lack of confidence.

Judges

The judging panel may comprise members of the organising team, the event tutors, the invited speakers for the event or even specially invited guest judges from the institution at which the hackathon is hosted.

Judging criteria: These criteria should be determined by the judging panel and shared with the hackathon teams prior to preparing their presentations. Possible criteria include:

- ➢ Creativity
- ➢ Teamwork
- ➢ Presentation skills
- ➢ Clarity of slides and graphs
- ➢ A demonstration of an understanding of the techniques used

Speakers

If suitable, a few presentations can be given at the opening of the event. These are usually introductory/overview talks on data science and machine learning, or the application thereof in academia and/or industry. Whether given by fellow students, academic staff or industry experts in the field, the talks contribute to the participant's awareness of the many applications of data science and machine learning.

The number and duration of talks should be limited such that they do not eat into one of the

most important resources in a hackathon - time! Participants will require sufficient time to work through their tutorials and tackle the hackathon task.

Tech Support

It is recommended that individuals who are able to provide technical assistance are on standby at the chosen venue, should there be any issues related to the computers participants are working on, the internet connection, projector setup or anything tech-related.

Hackathon Projects *("Hacks")*:

The projects that have been created for DSSE hackathons (or for anyone who wishes to use them for learning/teaching purposes!) are based on real-world problems and make use of real data. They cover a diverse range of topics relating to development, scientific research, or commercial/industry applications. They also vary in their level of difficulty and the computing power necessary to run the code, i.e. some may be suitable for high school learners while others require more advanced knowledge of programming and machine learning, and some may run with adequate compute power from the average laptop while others may require access to more compute power, e.g. the use of cloud computing. These details can be found in each of the project descriptions on the respective GitHub page. All project tutorials are written in Python.

SKA Data Challenge

The SKA planned for 4 data challenges to prepare the scientific community for the challenges that the SKA will present. We took the 1st data challenge and turned it into a tutorial format suitable for early career researchers. From the tutorials, you will learn the following:

- Source finding (RA, Dec)
- Source property characteristics
- Source classification (Star-forming galaxies, Active galactic nuclei)

SKA Data Challenge 1 - GitHub
Introduction Video

Image Classification of Galaxies

This project makes use of galaxy images from the GalaxyMNIST dataset. The dataset contains 10,000 images which have been

classified through the Galaxy Zoo citizen science project. This dataset will be used in order to build an image classifier model! We will also take a look at how we could perform this classification using unsupervised learning methods in cases where we don't have access to labelled data.

Galaxy Classification - GitHub
Introduction Video

Apple Classification

This project makes use of infrared spectra data for three apple cultivars in order to classify between bruised or sound apples. The following is covered in the tutorials:

- Tutorial 1 - Data cleaning & visualisation
- Tutorial 2 - Baseline calculation
- Tutorial 3 - Feature engineering and selection
- Tutorial 4 - Classification of apples using machine learning

Apple Classification - Github
Introduction Video

Sentiment Analysis of Tweets

This project involves Natural Language Processing (NLP) and sentiment analysis of Twitter data. NLP involves giving computers the ability to understand text and spoken words similar to how humans can. The tutorials will demonstrate how to collect and clean Twitter data, perform sentiment analysis using existing toolkits and finally, perform sentiment analysis using machine learning. The dataset we will use consists of 2000 tweets collected from a dataset called Sentiment140, which contains 1.6 million general tweets and their corresponding sentiment labels.

Sentiment Analysis of Tweets - GitHub
Introduction Video

Rooibos Tea Classification

In this project you will make use of chemical signatures to classify rooibos tea. The tutorials will cover the following:

- Data visualisation
- Data correlation

- Classification of tea using statistics
- Classification using machine learning

Rooibos Tea Classification - GitHub
Introduction Video

Flood Detection using Remote Sensing

This project makes use of Earth Observation data from the Sentinel satellite. The multi-spectral image data is used in order to determine land coverage, be it vegetation, water etc. This allows one to determine regions of flooding after a natural disaster.

The tutorials cover the following:

- Introduction to optical satellite imagery
- Data preparation and clustering methods
- Image segmentation

Flood Detection - GitHub

Search for Extraterrestrial Intelligence

The goal of the Voyager Tutorial is to take you through the "SETI Pipeline", that is the method used by the Breakthrough Listen team to search for alien techno-signatures! You will take real data gathered by Breakthrough Listen at the Green Bank Telescope in West Virginia, run it through a few algorithms, and see if you can find an alien or two!

Search for Extraterrestrial Intelligence - GitHub

Introduction Video

Web Scraping & Image Classification

In this challenge you will learn how to web-scrape images from Google and use them to train/test a machine learning model. The aim is to come up with an image classification problem (cats vs dogs, people vs trees, Trump vs an orange Cheeto etc), web-scrape the images and then use ML for the classification.

Image Classification

Pulsar Classification

In this challenge you are tasked with building a classifier to separate out real astronomical signals from man-made radio frequency interference (RFI). The astronomical signals that you're looking for come from pulsars, the ultra-dense relics of exploded stars. The dataset is available in two formats: (i) as a set of eight numerical features per sample suitable for classification using random forests, SVMs, neural nets etc. and (ii) as a set of images that show the data that the numerical features are drawn from, which are suitable for CNN based classification. Your hack challenge is to build the best classifier that you possibly can, remembering that we want as many correctly classified pulsars as possible, with as little contamination from RFI as possible.

Pulsar Classification

Technical Requirements:

Of course, a data science hackathon event necessitates access to computers, a reliable internet connection and, possibly, to computing resources with greater processing power than that found on the average laptop computer, depending on the chosen hackathon project/s.

Computers

Hackathon teams should ideally be given access to a computer lab for the event or asked to bring in their own devices. Even though the participants work through the tutorials and tasks in teams, it is most beneficial for each participant to have access to a computer/laptop so that each participant has the ability to run and manipulate the code first-hand and enhance their learning.

Internet

In the case of virtual hackathons, or in-person hackathons where the code will be run on a cloud computing platform, participants should have access to a reliable internet connection with speeds of at least 4 Mbps

Cloud computing

- Should the chosen hackathon project indicate that the use of cloud computing is recommended, participants may make use of the Google Colab computing platform. Colab is a free Jupyter notebook environment that runs entirely in the cloud. It does not require any prior setup and the notebooks that you create can be simultaneously edited

by others. Colab supports many machine learning libraries which can be easily loaded in the Jupyter notebook.

- Organisers may also contact the Inter-university Institute for Data Intensive Astronomy (IDIA - hack4dev@idia.ac.za) for possible computing support through the use of their research cloud facility. It is this cloud computing technology that was used to support the schools and the hackathons of this programme. Virtual machines on the Ilifu cloud are made available for the hackathons to fulfil their computing needs and participants work on the hackathon project on a Jupyter notebook interface.

Projector and screen

This equipment, along with the necessary audio-visual setup may be necessary for any introductory talks included in the event, as well as for the team presentations session.

Format:

Data Science Skills Exposure hackathons can be run over a period of 1-3 days, depending on the average level of Python and data science/ML experience of the participants attending, as well as the hackathon projects chosen. Some projects are less technical and could be run in 1 day, whereas others which make use of deep learning, for example, would require more time to allow for participants to work through the tutorials and grasp concepts. The hackathon may also be held as part of a larger event, such as a conference, workshop, or school at which lessons and/or talks related to data science and machine learning are delivered. Hackathons are intense by nature, in that participants will cover a lot of content and problem-solving in a short space of time. It is therefore important to follow a well-structured programme. An example programme can be downloaded from the 'resources' section below.

- If time allows, a small number of presentations which provide an overview/introduction to data science and/or machine learning and its applications may be included in order to ensure that all participants are aware of basic terminology and concepts, as well as the value of these skills, be it in academia or industry. It is recommended that presentations be kept to a minimum so as to not take attention and focus away from the hackathon activities.

- At the start of the hackathon participants will work through tutorials that have been created for each project, before tackling the project 'task', while receiving support from tutors who are familiar with the material and concepts.
- Participants work in teams of 4-5 individuals. It is recommended that teams be assigned such that each team is as diverse as possible (e.g. in terms of skill level, academic background/working profession, the tertiary institution at which they study, nationality if applicable, gender, etc). Having diverse teams will usually allow for better peer learning and the development of soft skills such as teamwork, communication and collaboration.
- Nearing the end of the event, teams may prepare short presentations (~5 - 10 minutes for each team) summarising their hackathon project task, the data they worked with, their

methods, results and conclusions.

Presentation guidelines: the expectations with regard to the team presentations should be clearly outlined before the teams begin their preparations. In addition to discussing solutions, showing their understanding of the techniques used, or demonstrating the accuracy of their solutions, it is also useful for teams to discuss the challenges faced and possible ways to tackle these going forward.

- The judging panel may each score the teams or identify their top three choices. After deliberation, a winning and runner-up team is identified and certificates and prizes may then be awarded where possible.

- It is often useful to allow for time at the end of the official hackathon for an “Ask Me Anything” session at which participants, organisers, tutors, invited speakers and judges can have discussions around the field of data science and machine learning, the hackathon itself, or perhaps just general career advice.

Administration:

- In cases where the event is not targeting a pre-selected group of individuals, e.g. all Honours and Masters students in the University’s Physics department, a poster/advert containing all relevant information can be created and disseminated via mailing lists and/or social media in order to attract suitable applicants
- A detailed application form should be created to ensure that all necessary information is captured for each applicant. This could include personal information, e.g. academic background/field or work (state the policy regarding the protection of personal information and ensure transparency regarding the use of this information), information regarding their existing knowledge/experience and, most importantly, their level of interest in data science and machine learning, motivation for attending such an event and what they hope to learn/gain. An example registration form can be downloaded from the ‘resources’ section below.

Optional Post-Event Actions:

- Follow up with participants after the hackathon to gather feedback and evaluate the success of the event. This feedback can be used to improve future hackathons. Feedback surveys may be used to determine the following:
    - The areas of data science and machine learning in which participants feel they have gained experience or knowledge
    - Aspects of the event that worked well and those that did not
    - Whether or not the event has impacted the participants’ view of their study and/or career options

  Example pre- and post-event surveys can be downloaded from the ‘resources’ section below

- Organisers may create an alumni network of participants who have attended the DSSE hackathon. This can be in the form of a mailing list, LinkedIn page, Facebook group, etc. Creating an alumni network can offer several benefits, including:
    - Networking opportunities: An alumni network can provide a platform for participants to connect with one another, share experiences and knowledge, and build professional relationships that can help them in their careers.
    - Access to useful information/resources: The network can allow for the sharing of resources such as employment/internship opportunities, mentoring programs, training materials, and industry events that can help participants to continue developing their skills, stay up-to-date with the latest trends of the field and advance in their careers.
    - Brand building: An alumni network can help to build the brand of the institution/organisation that hosted the skills development event, as participants who are satisfied with their experience are likely to spread the word and recommend the organisation to others.
    - Long-term impact assessment: Alumni networks can be a valuable source of feedback as participants can provide insights into what (if any) impact the event had on their academic path and/or career.
    - Building a data science/machine learning community: Creating an alumni network can help to create a sense of community among participants, support ongoing learning and career development.

DSSE Hackathon Resources:

In addition to the hackathon projects, the following resources have been collated in order to aid the implementation of a hackathon event.

Application Form - Word Document
Application Form - Google Form
Event Programme
Pre-event Survey - Word Document
Pre-event Survey - Google Form
Post-event Survey - Word Document
Post-event Survey - Google Form

- Project intro videos have been created for each hackathon project. These are "walkthrough" videos which provide an overview of the project and highlight any "pressure points" of the project and how to aid participants in tackling these.
- Recorded talks on topics in data science and machine learning are available on the Hack4dev YouTube channel and may be utilised at the event for the benefit of the participants.

- Participant preparatory resources: these resources may aid participants in their preparation ahead of the hackathon and may be circulated to all participants approximately two weeks prior to the event. The resources cover aspects of Python programming, as well as a basic introduction to certain machine learning concepts.
- There is a multitude of free Python programming and other Data Science learning resources available online. We have listed some of these resources below. These can be shared before or after the hackathon in order for participants to develop their Python skills or to enable and encourage participants to continue their journey into data science and machine learning.

OAD Astronomy & Data Science Toolkit
Zindi Learn
PyData Video Library
Kaggle Learn
Coursera Data Science Professional Certificate
Google Cloud Machine Learning and AI

Good luck organising your event! Let us know at hack4dev@idia.ac.za if you plan to run an event based on these resources and guidelines, or get in touch with us if you have any questions.

Download PDF Version

Below are links to some of the events that have been carried out as part of this programme.

UNZA Hackathon 2020
Space Generation Advisory Council (SGAC) All-Africa Hackathon 2020
UEM Hackathon 2021
Big Data Kenya 2021
Africa Women in Data Science 2022
NUST-DARA Data Science School 2022
Big Data Mauritius 2022

BITDN Hackathon 2022
UKZN Hackathon 2023
AfAS Hackathon 2023