Longitude Explorer Prize

Datasets

Top tips when using datasets!

- When looking for great datasets, you want to identify data which does not have many inaccuracies or irrelevant parts to it. **However,** there is no such thing as perfect data. The process of collecting and interpreting/analysing the data introduces bias. So it is important to understand your data and think about what these biases could be when you work with a new dataset.
- By trying to avoid having too many rows or columns in your dataset, this can make it easier to use.
- 'Cleaning' data such as having to remove and edit sections can be very time consuming with large datasets so finding datasets that have already been 'cleaned' can help save you time!

What does open source mean?

Open source means that the data or software has been made freely available for use. What this also means is that anyone can add and make changes to it. Often, this can actually be beneficial as other people can spot mistakes or vulnerabilities in the original code or data, but it is also important to recognise that there is some risk attached to using open source data and software: https://www.pcworld.com/article/197789/open_source_safe.html

Open source datasets

General:

- https://github.com/awesomedata/awesome-public-datasets - through GitHub you can access many public datasets, this is a list which combines lots of them. A benefit of GitHub is that with many of the datasets there is also a technical repository where you can see how others have used the data, follow their notes and learn how to analyse/explore the data using different approaches.
- https://data.gov.uk/ - a data source where the UK government publishes data, including from the central government, local authorities and public bodies to help you build products and services.
- https://www.ukdataservice.ac.uk/ - the UK's largest collection of social, economic and population data resources. This is a useful resource as it also contains guides and advice around using datasets.

Living Better & Living Longer:

- https://digital.nhs.uk/data-and-information - NHS digital collects and publishes data from the UK health services
- https://vizhub.healthdata.org/gbd-compare/ - a useful health data visualisation tool

Living Greener:

- https://www.metoffice.gov.uk/research/climate/maps-and-data/data/index - National and regional climate statistics for the UK

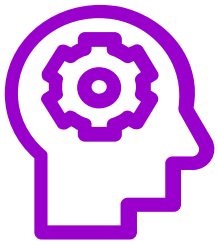
nesta Challenges

Living Together:

- https://www.ordnancesurvey.co.uk/business-government/tools-support/open-data-support - a set of digital maps of the UK with information about areas e.g. addresses, location names, transport networks, crime and pollution

Machine Learning:

- https://www.kaggle.com/datasets - Kaggle offers interesting data sets, but also resources for growing your knowledge and practicing skills around ML
- https://registry.opendata.aws/ - each dataset has a description and usage examples which can help you to use them



Other helpful resources:

- A free tool that introduces machine learning by providing hands-on experiences for training machine learning systems and building things with them: https://machinelearningforkids.co.uk/#!/welcome
- Machine Learning for Kids - activities developed by IBM: https://www.ibm.org/activities/machine-learning-for-kids
- Machine Learning and AI: Find Datasets: https://guides.library.cmu.edu/machine-learning/datasets
- Preparing Your Dataset for Machine Learning: 8 Basic Techniques That Make Your Data Better: https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/

Software

Free services offered:

- **Microsoft Azure machine learning studio:** https://studio.azureml.net/ **(a free tier of use) -** Build your cloud-based skills with these developer tools and learning resources. Choose anonymous Guest Access, or sign in with your work or school account
- **Google AI Platform:** https://cloud.google.com/ai-platform **(use of the AI platform is free) -** AI Platform makes it easy for machine learning developers, data scientists, and data engineers to take their ML projects from ideation to production and deployment
- **IBM Cloud AI:** https://cloud.ibm.com/catalog?category=ai **(free options for AI services) -** Explore the broad portfolio of managed services for infrastructure, developer tools, and more to build your apps on the public cloud

Open source:

- **Jupyter notebook:** https://jupyter.org/ - A very useful tool which this nonprofit organisation created to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages
- **Google Colabs:** https://colab.research.google.com/notebooks/welcome.ipynb - a similar option to Jupyter but integrated with Google Drive which can be very helpful
- **Stack Overflow:** https://stackoverflow.com/- a user enabled/crowdsourced community tool which can be used for fixing bugs and troubleshooting code

- **TensorFlow: https://www.tensorflow.org/ -** an open source ML tool, available in Python, C++, Haskell, Java, Go, Rust, and JavaScript

Other helpful resources:

- 10 Best Artificial Intelligence Software (AI Software Reviews In 2020) https://www.softwaretestinghelp.com/artificial-intelligence-software/
- Medium.com is an essential source of information and is where a lot of useful information is published. The *Towards Data Science* publication pages are great for any aspiring data scientists/AI technologists. A good article to start with would be: https://towardsdatascience.com/machine-learning-zero-to-hero-everything-you-need-in-order-to-compete-on-kaggle-for-the-first-time-18644e701cf1
- A data labelling service from amazon. It is important to understand and consider whether the data useful and what ethical and quality associated issues may arise from using it: https://www.mturk.com/
- Udacity and Coursera are both great online learning websites which can get you up and running with machine learning and help you start working using different ML techniques on a range of different open source datasets