Springboard

Machine Learning Engineering Career Track

What Data To Collect?

What Kind Of Datasets Do I Look For?

Besides leveraging the existing data repositories we provided, finding the right dataset(s) might require a little bit of “google-fu” and creativity. Many big cities (like New York) are opening their data to the public, so check on their websites. Don’t forget, the web is a very large repository of data in and of itself. Any public API from places like StackOverflow can provide fresh and unique data. Firehose APIs provide real time data and you can collect these from places like Twitter to see the latest trends. This will give real time data about current events happening locally or around the world, and how people might be reacting.

This article also gives a lot of good pointers:
https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b

What Are Some Recommended Datasets?

Here are a few great sources for large datasets that are appropriate for this course:

- fast.ai research datasets collection
- Google dataset search

- AWS open datasets repository
- Uber Movement
- Yelp dataset

Consider searching Kaggle, one of the most popular sources for datasets. Additionally, Quandl, US Government Open Data, Data is Plural, and UCI Machine Learning Repository are all great places to locate data as long as they meet the requirements below. If you have any questions, reach out to your mentor. Once you have picked datasets, your goal is to narrow them down to the ONE idea that you'll be working on.

How To Leverage Open Data Foundation?

Throughout the world, many organizations under the Open Data Foundation are seeking to open data as much as they can. This report from Jan 2019, gives you a better of the progress being made to make all of this data available and ready to use by anybody:
https://www.datafoundation.org/the-state-of-the-union-of-open-data-ed-3

The advantage of those datasets is that they are free to license and ready to use. The downside might be that they are not always well curated, and are likely to require a lot more data cleaning, as well as some domain knowledge expertise.

Using those kinds of datasets will give you a very close experience at what you might experience in the real world, especially in fields that rely on a lot of government generated data such as traffic patterns, income, revenue, or air quality. Government data sources do not have a good reputation for producing clean and immediately useable data, but they are an invaluable well to draw from.

https://index.okfn.org/dataset/
https://www.nsf.gov/data/

What About Big Data Sets?

We encourage you to use a large dataset so you can scale your model in a realistic way. But it is not required.

Ideally, your dataset should have at least 15K-20K samples at the bare minimum. In this course, since we'd like to see you build large-scale applications, we encourage you to go for larger datasets. We encourage you to choose something that's at least 8GB in size or has at least one million samples.

If you choose a small dataset, know these can be just as challenging as a large one, due to the fact that it might be hard to generalize.

If you choose to use a smaller dataset, you might have to do more prep work, leverage data augmentation tactics, and/or find additional datasets to pair smaller datasets with.

What About Live Data Processing?

Live data is great to work with, but you need to be able to archive a certain amount of data in order to train your model on a good size dataset. Evaluation can happen against a replay of archived live data or an actual live data segment like the previous week. If you want to work with a live data source, be sure to discuss the complications around this with your mentor soon.

What If I Can't Find An Existing Dataset?

If you are collecting the data on your own, either via an existing app or web scraping, make sure you budget for extra time. You'll need to not only collect the data but architect the data storage in an appropriate fashion. It will also require some additional discipline, testing of the collection process to minimize potential heavy bias, or other complications that make the model build more challenging.

Can I Use a Dataset from Kaggle?

You're welcome to use a dataset from Kaggle in your project! Many Kaggle competitions now provide large, complex datasets, including those for computer vision and NLP. If you'd like to use a large Kaggle dataset, make sure your mentor approves.

Just because you use a dataset from Kaggle, doesn't mean that you have to solve exactly the same problem that they ask. Here are a couple of ways you could use the dataset differently for your capstone project:

- Is there a different problem than the one asked in the competition that the dataset can be used to solve?
- Could you combine it with other datasets to solve a different problem, even if that problem is similar?

Your mentor has the final word on whether a Kaggle competition is appropriate for a Capstone Project, so please make sure to get their explicit approval for the dataset you'd like to use.

Springboard

Can I use a Private Dataset from my Employer or Another Source?

Many students use proprietary data from their employer to work on their capstone projects, which is perfectly fine. **We don't require that you share the raw data** with anyone. However, there are a few things you'll need to consider:

1. **Ensure you have the right permissions:** Your mentor is here to guide you through your project. They can only do that effectively if they can look at your code, summarized results, and charts. They may not need to directly access the data, but there are considerations to make.
    a. Springboard still requires that you turn in a project report and a slide deck based on your analysis and place it publicly on GitHub.
    b. If your employer or the people who are providing you the raw data are not comfortable with these requirements, you may need to rethink your project topic. It's your responsibility to ensure that private data is handled appropriately and securely. Please check with the legal team at your employer to see if you need approval in writing in the form of a legal contract or a Non-Disclosure Agreement (NDA).
2. **Start data collection early:** Even if you have the requisite permissions, please make sure to start the data collection process early and have a realistic idea of how soon you can get the data.
    a. Many companies have elaborate processes around data access and extraction, so sometimes, students have become stuck for weeks or months waiting around for their project data to become available.
    b. Ensure that you follow good privacy and security practices. For example, anonymize the data where appropriate. In some cases, you may be legally required to anonymize it (e.g. healthcare data). Please work with the legal and security teams at your employer to ensure you're always in compliance.

If you have any questions about whether or not you can use proprietary data for your Capstone Project, feel free to email your student advisor!