Machine Learning Engineering Career Track

Frequently Asked Questions

How complex does a project need to be?

This goal of this course is to give you the skills to not only design and create an ML/DL application, but also deploy it to production using the latest engineering tools and techniques.

ML/DL contains a wide swath of techniques, and not everything in there can apply to the problem you have chosen. The more complex techniques you use, the more attractive your project will be to employers. But, it's important to balance that with a practical approach of producing a real, scalable application. Some questions you can ask yourself:

- What's the technique that best applies to this problem?
- How easily can this technique be deployed to production?
- Are the results of this technique "good enough" to meet the business requirements of the solution?

In the real world, it **can sometimes be better** to use a simple logistic regression approach that's accurate enough, highly performant and can be deployed quickly as a production application, than a really complex deep learning approach that's difficult to deploy or maintain. You'll have to make similar decisions and tradeoffs about your capstone project. When you're interviewing for your MLE role, you'll be expected to justify these decisions and tradeoffs.

Work with your mentor to determine a problem and approach that meets the requirements of this course and makes you both feel confident that you'll be able to do well.

Is it better to stick to particular areas or applications of ML/DL?

You can pick any problem and dataset you like (with mentor approval) that will let you demonstrate your competency with the entire process of creating ML systems. However, we have a few guidelines that we recommend, based on the material we cover in the course and the skills that are in demand.

1. **Deep Learning:** If you decide to do a project involving Deep Learning, we recommend you stick to a problem in **Computer Vision** (also called image processing) or **Natural Language Processing** (NLP). We cover these two areas deeply in the course, with examples, so you'll be able to use them to guide your project. You're welcome to try other areas or other kinds of data (audio, video, etc.), but you might have to figure out a few things on your own. If you're the kind that welcomes a challenge, go for it!

2. **Natural Language Processing (NLP):** Many of our students are excited about doing NLP projects, since it's an extremely high-demand area of ML/DL application. Many companies have vast amounts of data lying around in text documents and communications that they'd love to get insights into.

   NLP has a long history with many techniques to represent and analyze language. However, in the last few years, DL has revolutionized the field. This was solidified by Google moving all of its machine translation tools (e.g. Google Translate) to deep learning. As a result, we strongly recommend that if you choose to do a capstone project in NLP, you use deep learning based methods in your project.

3. **"Old-school" Machine Learning:** You're also welcome to build a capstone project using traditional ML techniques (regression, SVMs, random forests) rather than deep learning. However, in order for these projects to meet the bar of this course, we encourage you to really focus on the engineering and deployment aspects of these projects, including making sure your project can work with really large datasets, uses the latest tools/techniques (e.g. SparkML or TensorFlow), and has a strong engineering component.

Springboard

What kinds of cloud resources are available for capstone projects?

As part of your capstone project, you'll be expected to carry out part of your development and deployment on a cloud platform. At Springboard, we're working on providing you with as many cloud resources as possible, but we recommend that you use them cautiously and begin your project by using free options to prototype your work. This reflects the real world: In an actual job, there will be budget and resource constraints, so it helps to use fewer resources earlier on.

We've written up an entire set of guidelines on the available cloud resources and how to use them. Check them out here.

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, it 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 (or even a similar one)?

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.

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, even if they don't have access to the actual data.
    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!