Data Science and Machine Learning Engineering

 [(https://github.com/learn-co-curriculum/dsc-data-science-and-machine-learning-engineering)](https://github.com/learn-co-curriculum/dsc-data-science-and-machine-learning-engineering)
[(https://github.com/learn-co-curriculum/dsc-data-science-and-machine-learning-engineering/issues/new)](https://github.com/learn-co-curriculum/dsc-data-science-and-machine-learning-engineering/issues/new)

Introduction

In this lesson, we'll learn about how the role of Machine Learning Engineer fits into the broader data science job market.

Objectives

- Explain the difference between a Data Scientist and a Machine Learning Engineer
- Explain the difference between deploying models and productionizing models
- Think about how your skillset fits into the data science job market

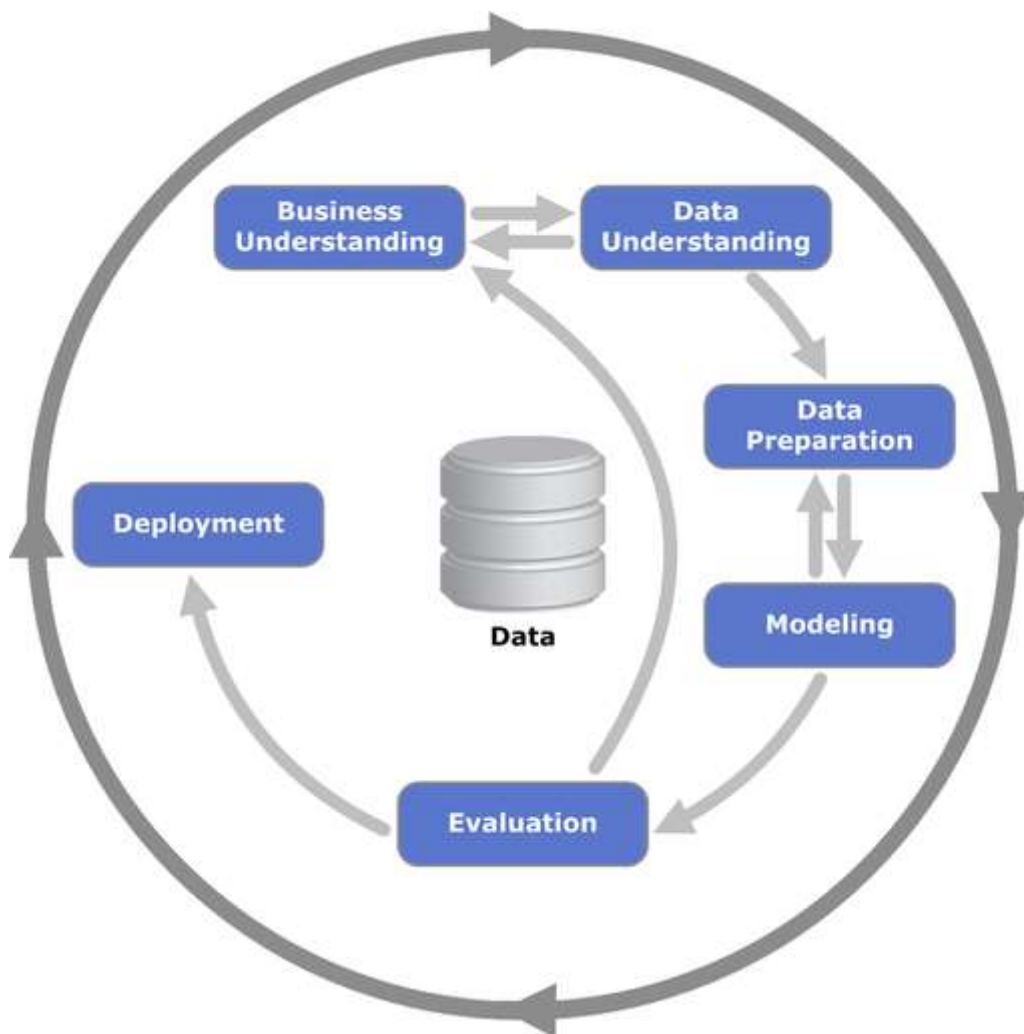# Data Scientist vs. Machine Learning Engineer

At large, established tech companies, data-related roles have specialized into separate Data Scientist and Machine Learning Engineer categories.

Data Scientists typically run experiments and train models until they have found a solution that works. Once they have trained and validated the model, they typically then hand off productionization of the model to **Machine Learning Engineers**. Whereas the Data Scientist creates the basic prototype, the Machine Learning Engineer's job is to put that model into a production system in a performant and maintainable manner. Whereas Data Scientists at large companies focus on the "big picture" by finding solutions to business problems, Machine Learning Engineers focus on the details, implementing the solutions created by the Data Scientists in the best way possible. Data Scientists focus more on analytics and statistics, whereas Machine Learning Engineers will have a stronger command of backend engineering, data structures and algorithms, and software engineering overall. The following diagram lays out the relationship between different technical roles well:
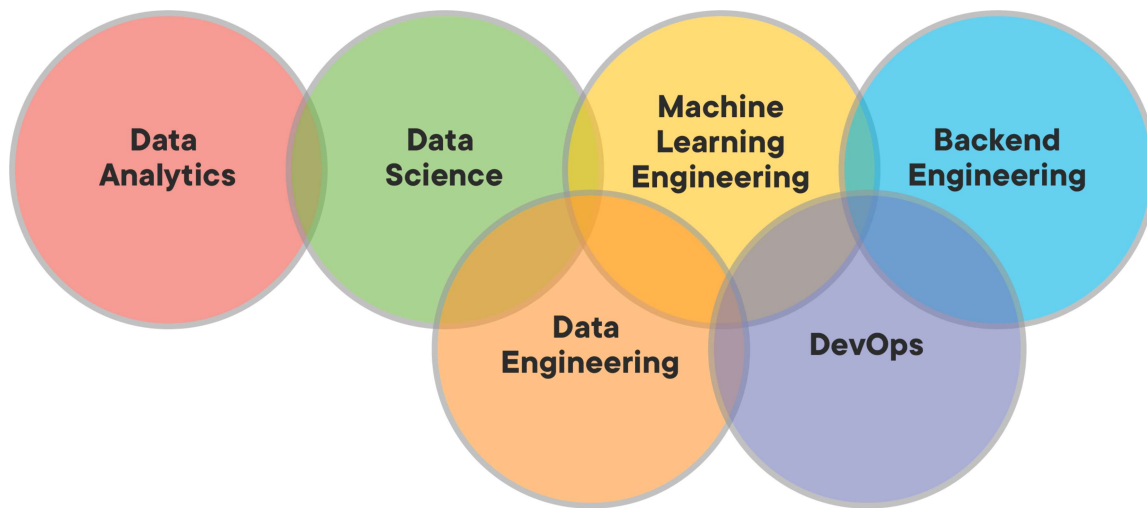
Deploying vs. Productionizing

If we think back to the CRISP-DM process model, Deployment is the final step. We have previously discussed model pickling and how to ensure that data science processes are reproducible, as part of the deployment process:

[(https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)](https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)

So, if a Data Scientist toolkit already includes deployment, what additional skills and techniques play into the **productionizing** of models that Machine Learning Engineers perform?

There is no bright-line distinction between deploying and productionizing, but in general deploying focuses on the bare minimum to get a model into a context where it is usable for the client, whereas productionizing considers broader factors such as scaling, maintenance, and security. A deployed model might be manually re-trained on the latest data once a month, whereas a productionized model might be continuously re-training as new data comes in. Productionized models might also have test coverage, metric tracking, and **more** ⬀ **(https://ml-ops.org/)** .

Data Science Skills and the Job Market

As you start to look at job postings, you probably notice that **a "Data Scientist" job can mean many, many different things**. In some companies, it means a data analyst or a DBA focused on databases or data pipelines. In others, it means someone with a scientific mindset skilled with running A/B tests. Yet others may be highly specialized machine learning roles in areas like NLP, Computer Vision, or Deep Learning -- and these roles may break down further into specializations focused on either research or implementation. As a Junior Data Scientist entering the workforce, it's most likely that you'll land in a generalist role, spending the first few years of your career working on various tasks that focus on all of these areas at least a little bit. Specialization happens later in your career. Out of the gate, the best thing that you can be is a **strong generalist**, with the demonstrated ability to contribute to many different sorts of projects that might be expected of a Data Science team.

Being a 'Scrappy' Data Scientist

Large companies with official Machine Learning Engineer roles are typically only looking for more-advanced candidates. For example, a Machine Learning Engineer posting from Amazon we found lists these required qualifications:

> · 3+ years of experience contributing to the architecture and design (architecture, design patterns, reliability and scaling) of new and current systems.
> · 4+ years of non-internship professional software development, data engineering, or machine learning experience
> · Bachelor's degree in Electrical Engineering, Computer Sciences, Mathematics or equivalent work experience
> · Experience deploying and running services in AWS and in engineering big-data solutions using technologies like Glue, S3, and Spark

In companies like that, there is less need for anyone in a Data Scientist role to have model productionizing skillsets, since the size of their data teams allows for this kind of specialization.

With small and medium-sized companies, it's much less likely that a role like this will exist. Data Scientists in smaller organizations are expected to be a bit more independent, and will likely have to "wear more hats".

For a data science role at a startup, it's a common expectation for their ata scientists to handle every part of a data science project. This means starting by interviewing key stakeholders and identifying the problem to be solved, followed by rapidly prototyping a solution until you've trained/tuned/validated a model that meets your standards, followed by actually putting that model in production!

This means that it's very important to be 'scrappy', and be able to handle anything that's thrown at you as a Data Scientist. Smaller companies often don't have the funds or the infrastructure for a separate Machine Learning Engineering team to handle the details of implementation. In this respect, being able to productionize a machine learning model can make you a much more attractive candidate to employers, and give you a competitive advantage!

Summary

In this lesson, we learned about the similarities and differences between Data Scientists and Machine Learning Engineers. We also learned why the ability to productionize a machine learning model is a crucial skill for data scientists at small companies, as well as how this skill can provide a competitive advantage when applying for jobs!