Introducing Machine Learning Concepts

Over the next several pages, we'll introduce to several key concepts that will be necessary to understand before you build out a machine learning model:

1. The data science process
2. Training and evaluation
3. Evaluation metrics and choosing a model

The Data Science Process

First, let's begin with an overview of the typical **data science process**.

Key Phases of the Data Science Process

There are five key steps to the data science process:

1. **Collect data**. We need to be able to collect reliable data and ensure that we understand its origin, quality, and meaning.
2. **Prepare data**. The next phase in the data science process is data preparation. This can take up an enormous percentage of total time--estimates range upwards of 80-90%.
3. **Train a model**. In this phase, we apply appropriate algorithms to our data, resulting in a trained model. Depending on the algorithm, training may take a considerable amount of time. For example, certain applications of neural networks may take over a month to train fully. Generally, however, model training does not take that much time.
4. **Evaluate the model**. After we have trained a model, we want to ensure that it meets our expectations in terms of quality. Because the real world is always more complicated than our training data set, we want to make sure that the

results look reasonable on non-training before pushing a model out to production. This helps us avoid overfitting to the training data.

5. **Deploy the model**. Having a model is great, but having it available for use is the natural next step. Historically, deploying a model typically meant rewriting it into a "production" development language like C or C++. Today, it is easy to run a microservice in a language like R or Python and handle prediction needs.