edX　　　**Microsoft:** DAT210x Programming with Python for Data Science

2. Data And Features > Lecture: Features Premiere > Video

🔖
Bookmarks

▸ Start Here

▸ 1. The Big Picture

▾ **2. Data And Features**

Lecture: Features Premiere
Quiz　　　　　　　　　🖉

Lecture: Determining Features
Quiz　　　　　　　　　🖉

Lecture: Manipulating Data
Quiz　　　　　　　　　🖉

Lecture: Feature Representation
Quiz　　　　　　　　　🖉

Lecture: Wrangling Data
Quiz　　　　　　　　　🖉

Lab: Data and Features
Lab　　　　　　　　　🖉

Dive Deeper

Features Premiere

In the last section, you were exposed to the types of problems that are solvable with machine learning. It was also mentioned that "with enough good data, it is simply amazing what you can get a computer to do." Part of having good data is ensuring your data is properly organized so it can be processed to begin with. To be usable by SciKit-Learn, the machine learning library for Python you'll be using for this course, your data needs to be organized into matrix of **samples** and **features**:

	feature_0	feature_1	..	feature_n
sample_0	X	X	..	X
sample_1	X	X	..	X
..	..	..	..	..
sample_n	X	X	..	X

A sample is any phenomenon you can describe with quantitative traits and is represented by a row in your dataset. If you were building a dataset of 'companies', each sample would be details about a company. When it's said that you need a lot of *data*, what is generally meant is that you need *many samples*.

Features are those quantitative traits that describe your samples. They might be numeric or textual in nature. For example, *CompanyName* could be a textual feature in your 'companies' dataset, and different samples of it might result in values such as 'Microsoft', 'EdX', or 'Coding Dojo', etc.

If *CompanyRating* were a numeric feature in the same dataset, its value might be score between 1 and 5 for each company:

Companies

	CompanyName	Rating
0	Microsoft	5
1	EdX	5
2	Coding Dojo	5

POWERED BY
OPENedX