



Microsoft: DAT210x Programming with Python for Data Science



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

2. Data And Features > Lecture: Determining Features > General Guidelines



Bookmark

General Guidelines

Of course, everything just mentioned only applies if you're given the creative liberty to go out and collect your own data! If you are provided data to work with, then all you're really able to do is come up with nifty ideas for combining pre-existing features in order to form new and beneficial derivatives.

Your machine learning goal should be to train your algorithms instead of hard coding them. When it comes to deriving features, approach them the same way. Let your expertise and intuition guide you, by brainstorming what data you would need to collect to achieve the objective of your analysis. Think of your machine learning models as if they were small children who have absolutely no knowledge except what you train them with; what information would they need to know to make the right decisions?

There are no hard and fast rules when it comes to thinking up good features for your samples; but a rule does exist about *what to avoid*: garbage. If you collect details about your samples that you *know* to be statistically irrelevant to the domain of the problem you're trying to solve, you'll only be wasting your time and eroding the accuracy of your analysis. Garbage in, garbage out.

If you're trying to have machine learning model a regression relationship between various car features (MPG, comfort level, current milage, year manufactured, # cylinders, has turbo, etc.) and car costs, introducing features like car color and and car air freshener scent probably won't do you that much good. In order for machine learning to do it's job of finding a relationship in your data, *in your data, a relationship must exist*.

- ▶ 3. Exploring Data
- ▶ 4. Transforming Data
- ▶ 5. Data Modeling

© All Rights Reserved



© edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

