Microsoft: DAT210x Programming with Python for Data Science   <u>**Help**</u>

General Guidelines

🔖 **Bookmark this page**

General Guidelines

Everything just mentioned only applies if you're given the creative freedom to go out and collect your own data! If you are given data to work with, all you're really able to do is come up with nifty ideas for combining existing features 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 meet the goals 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* is 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 mileage, year manufactured, # cylinders, has turbo, etc.) and car costs, introducing features like car color or air freshener scent probably won't do you that much good. In order for machine learning to do its job of finding a relationship in your data, *in your data, a relationship must exist*.

POWERED BY
OPEN**edX**