What is Data Science

* + Intro to Data Science
    - What is data science?
      * IN general, data science is understood to be the practice of extracting insights from data to guide decision making. To achieve this, data scientists leverage a combination of computer skills, statistical understanding, and subject matter expertise
      * Put plainly, it's the application of scientific techniques to data to solve problems.
      * Data scientists are not always referred to by the same title. In fact, there are dozens of titles out there. Here are a few common roles that may use data science techniques:
        + Data scientist.
        + Data analyst.
        + Business intelligence analyst.
        + Data visualization expert.
        + Researcher.
        + Engineer.
        + Statistician.
        + Developer.
      * First off, it’s important to understand the skills involved. Data scientists commonly combine:
        + Business intelligence.
        + Machine learning modeling.
        + Database design/big data.
        + Applied math and statistics.
        + Programming.
        + Study design.
      * Data Science Workflow
        + Frame: Develop a hypothesis-driven approach to your analysis
        + Prepare: Select, Import, explore, and clean your data

In Step 2, we prepare our data. To start, we'll want to select and import the data we intend to use. We might do this by asking a series of questions:

What data is provided, if any?

Does this data describe the problems we're trying to solve?

Are there enough data points to adequately explore the issue?

Is this data set trustworthy? How was it collected?

Do we need to go out and collect more data?

If so, where should we collect additional data?

* + - * + Analyze: Structure, visualize, and complete your analysis

For Step three, we'll analyze our data. We might do things like:

Structure, segment, or isolate parts of our data,

Visualize our data using various methods,

Calculate and compare our selections.

Depending on the problem we're trying to solve (and the type of data we're working with), our analysis could take many different forms. When in doubt, think back to your original goal in step one - what is your hypothesis? What are you trying to prove or measure from this data?

To do this, you might start by calculating some basic statistics or visualizing relationships between specific variables. In later lessons, we'll cover how to perform these steps in Python.

* + - * + Interpret: Create recommendations and business decisions from your data

Once we've visualized and analyzed our data, we'll be able to interpret it. But what does that mean?

In step four, we'll use our analysis to answer our hypothesis, and then use this answer to make predictions about future data. This allows us to create meaningful, real world recommendations!

So how does this work? We might:

Recall our Planet Express hypothesis,

Review our prior visualizations and calculations,

Create a model that demonstrates how our analysis answers our original hypothesis,

Test this model on a sample set of new data to see whether it continues to hold true,

Refine our model until we're confident that we can make recommendations from it!

Our hypothesis from step one should guide our approach and analysis in steps two and three. In step four, our goal is to create a general model that interprets our data and resolves our hypothesis. Once we've tested this model on new data, we'll be able to use it to make predictions and justify future decisions.

* + - * + Communicate: Present insights from your data to different audiences

No matter how brilliant your model is, or how illuminating your findings are, they won't matter if no one listens. You'll need to help your audience understand your process, care about your results, and persuade them to take any next steps.

In other words, presenting solutions with data involves many of the same considerations as any other type of field. You'll want to do things like:

Consider your audience,

Create powerful visuals,

Tell a meaningful story,

Practice your presentation beforehand,

Be prepared to answer questions.

So how do you do this, exactly? What does a compelling data-driven presentation look like?

A few key components to consider, include:

Summarizing your findings,

Labeling all plots and visualizations

Restating your hypothesis and initial assumptions,

Describing your data and your process,

Explaining your model's strengths and limitations,

Providing an appropriate degree of disclosure for your audience (especially when dealing with proprietary data or sensitive user information).

* + - * Reliable and Reproducible
        + Our problem-solving framework will help you produce results that are reliable (so that your findings are more accurate) and reproducible (so that others can follow your steps and achieve the same results).
        + Note that, depending on the problem you’re trying to solve, the data science workflow will not always be linear. You may have to repeat some steps before drawing any conclusions!
        + We'll explore these steps through a project example in the following slides.