**Question:**Can you tell me more about this course?

**Answer:** Yes. This is a very *high level* look at five key libraries used in Python for machine learning. The libraries are covered very briefly to give you a quick look into what they do. My goal is to show those new to data science why Python is the gold standard in the applied machine learning space and the very basics on how to use them.

There is no depth to any of the lessons in this course. This is truly an overview.  I'm simply going to show the very basics of each library.  Keep in mind there are hundreds of libraries and I'm taking some poetic license on what I believe are the top ones. (I've asked around though and some top data scientists agree with me)

**Question:** Are you saying I have to learn all these libraries?

**Answer:** Yes. Maybe not NLTK. If you niche is going to be NLP or sentiment analysis then you'll need to know NLTK. Keep in mind the great thing about the Python ecosystem is it's ease of use. You'll see very shortly some tasks... like importing libraries are done exactly the same regardless of which library we are working with. You'll learn about the acronyms in the course.

**Question:** I'm having a hard time putting everything together. Can you tell me what a data scientist or machine learning engineer does?

**Answer:** We often associate the word data scientist with the PhD types. Most of us that build models for a company are called applied machine learning engineers.  The applied word simply means we work in the real world not at a University. We spend most of our time massaging data.

Data is dirty and we have to clean it really well before we can model it.  Additionally, our models are temperamental.  For example, if we have a column in our data set that has a wide range. Let's use salary as an example. The salaries are often listed like this: (90,000, 120,000) The first would be 90K and the second would be 120K. Our models should be able to handle this but in the real world they can't so we have to normalize it or place that data into an acceptable range.

Since most companies have tons of data all of our models are supervised. Supervised means we use an existing data set to create our models.  Once we cleanse our data we build models that are very much black boxes to most. We simply decide what model is best for a given scenario and then point our model at that data and wait for it to spit our a result. Now, obviously it's much more complicated but that's what we do all day.

**Question:** Do I need a PhD in order to be a Data Scientist?

**Answer:** You do not.  A highly skilled Python programmer is as valuable if not more valuable than most data scientists.  As machine learning engineers we aren’t writing the algorithms, we are building, tuning and evaluating pre-built models.

**Question:** Is the career outlook for data scientist exaggerated?

**Answer:** I don't believe so. *By 2018 the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge.*

-- McKinsey Report Highlights the Impending Data Scientist Shortage 23 July 2013

**Question:**I'm worried about the amount of math I'll need. How much will I need?

**Answer**:  If you're planning on authoring the algorithms then it's pretty heavy. It is true that data science requires an understanding of statistics and probability because most of the predictive modelling techniques are based on these concepts but applied predictive modelers don’t have to be math wizards.  With the right mix of logic and common sense, one can go far as a data scientist even if one has moderate mathematical abilities.

**Question:** Do machine learning engineers and data scientists get to build models all day?

**Answer:** I wish. They don't. You'll spend most of your time working with data. In the real world data is dirty. Even if you're dealing with structured data it will often be filled with holes.

*Insight:* There are two kinds of data. There is structured and unstructured data. Data in a relational database is structured. A text file sitting on a file system is unstructured data.  It's much easier working with structured data because before you begin building your models you need the data to have structure.

**Question:** What's a real world salary for a data scientist?

**Answer:**In the US a junior level position could pay up to 80K.  Those in the mid to upper levels should be making around 150K.  The salary obviously depends on a lot of factors. Location and niche play a large role.  If your niche is deep learning or NLP and you aren't above 150K then regardless of location you are either under-payed or aren't that skilled.

**Question:** How long will it take me to be a senior level data scientist or senior level machine learning engineer?

**Answer:**A long time. Regardless of your career choice mastery takes a long time.  I would suggest five years of experience with intense study before you put senior on your resume.

**Question:** Are their any careers that could short circuit the learning curve?

**Answer:** Yes. If you're a developer or database administrator and have spent your career working with data then some things will just make sense. In the past I've done data mining. Data mining and supervised predictive modeling are basically the same thing.  If your a mathematician or statistician then you're ahead of the curve.  Do keep in mind that data is about 80% of your job.

**Question:** Are there any certifications you'd recommend for data scientist or machine learning engineers?

**Answer:** Unfortunately not. I'm sure there will be a couple soon but there's nothing you can put on a resume right now that would get a recruiters attention.

**Question:** What can I do right now to put myself in a position to become a data scientist or machine learning engineer?

**Answer:**  Here's a suggested course of action.

* Become a voracious learner of all things machine learning.
* Enter many Kaggle competitions. It doesn't matter where you place.
* Find data that isn't common and create and publish your models to github.
* Learn Python and the associated libraries for machine learning.
* Create a blog about your learning experience and about data science.
* This is a great free resource: http://datasciencemasters.org/

**Question:** What about a master's or PhD?

**Answer:**  There are a few really good master's programs and a few of them are online.  The Berkeley program would be my first choice and the University of Illinois my second. However, I'm very apprehensive of any course that's not using Python as the core language.

**Question:** What language should I learn for applied machine learning?

**Answer:**  That's an easy one. Python. Sure, there's a lot of R still out there but the majority of real world models are being built in Python.