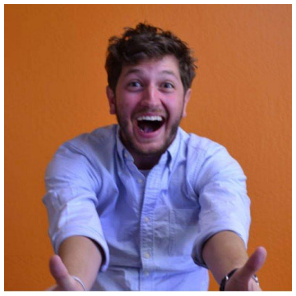
Intro to Data Science

Lesson 1 Notes

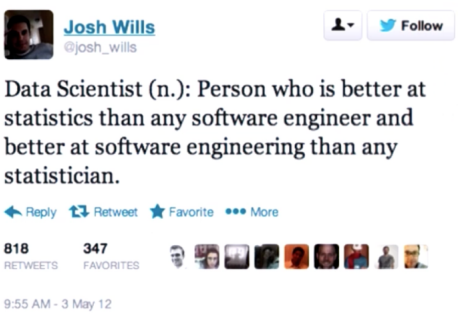
Introduction

Introduction to Data Science



Hi, and welcome to Introduction to Data Science. My name's Dave, and I'll be the instructor for this course. I've worked as a data scientist in Silicon Valley, most recently at a small company called Yub and before that at a company called TrialPay. I'm formally trained as a physicist, and I originally became interested in data scientist because I love the idea of improving the quality of people's lives or building really cool products by using data and mathematics. In this lesson, we'll discuss data science at a high level. Together we'll find out what data science is and discuss what skills are required to be a data scientist. We'll also hear from a bunch of other data scientists about interesting projects they worked on. And discuss how data science is being used to solve a bunch of different problems. This lesson in particular is going to be a little bit different than the others. We're not going to build as much. I think it's important to understand data science at high level before we dive into the details. Alright, well I'm really excited about this course, so why don't we get started.

What Is a Data Scientist



Copyright © 2014 Udacity, Inc. All Rights Reserved.

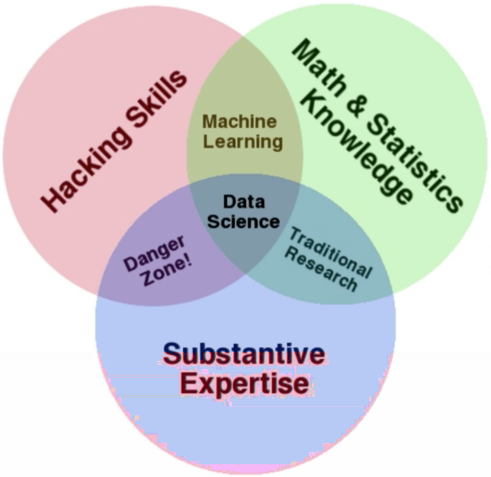
People have many different conceptions of what data scientists do. Some might say that a data scientist is just a data analyst who lives in California. While others might say that a data scientist is a person who's better at statistics than any software engineer, and better at software engineering than any statistician. As you can see, definitions vary wildly from place to place, and from person to person.

Quiz: What Is a Data Scientist

So before we get started, let me ask you a question. What do you think data scientists do in their day-to-day work? Type your thoughts in the text box below. Don't worry, there are no right or wrong answers and this quiz will not be graded.

Answer:

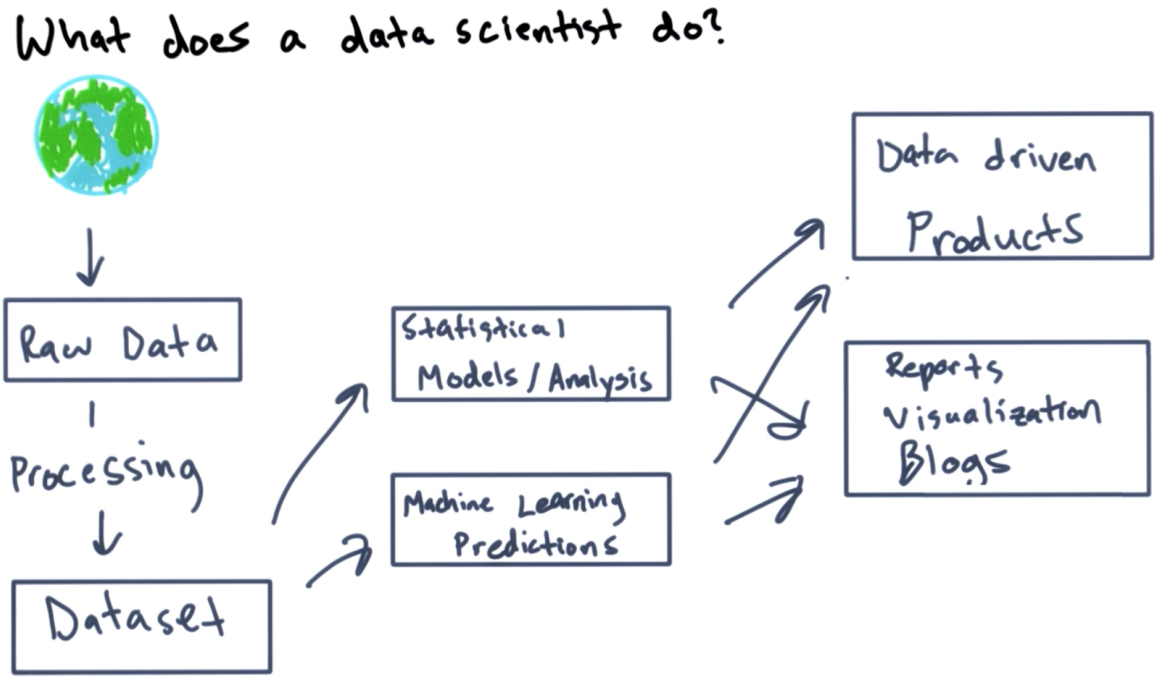
Let me tell you my perspective. From personal experience, data scientists today are people who have a blend of many different skills. This Venn diagram shows a definition of a data scientist that I like a lot. A data scientist is someone who knows mathematics and statistics, which allows them to identify interesting insights in a sea of data. They also have the programming skills to code up statistical models and get data from a variety of different data sources. Furthermore, a data scientist is someone who knows how to ask the right questions and translate those questions into a sound analysis. After doing the analysis, they have the communication skills to report their findings in a way that people can easily understand. In other words, data scientists have the ability to perform complicated analysis on huge data sets. Once they've done this, they also have the ability to write and make informative graphs to communicate their findings to others.



What Does a Data Scientist Do

Here are some things that a data scientist may do in his or her daily work. They might wrangle data. That is, collect data from the real world, process it, and create a data set that can be analyzed. Once they have a data set, they may analyze trends in the existing data or try to make data driven predictions about the future using the data at hand.

Copyright © 2014 Udacity, Inc. All Rights Reserved.



Based on this models or predictions, they cannot only build data driven products but also communicate their findings to those other data scientists and the general public. File visualizations, reports or blog posts. But hey, this is my point of view. Why don't we talk to some other data scientists and hear their thoughts.

Pi Chuan Introduction

My name is Pi-Chaun Chang. My background, so I've been doing Computer Science ever since college. I did a PhD CS PhD at Stanford, and I worked at Google for four years. Now I'm in a startup called AltSchool.

Pi Chuan - What Is Data Science

The term data science is actually pretty new to me as well, even though now I think about it, I have been doing data science since the day I was at in Taiwan doing a Masters in speech recognition. The way we would understand speech is to collect a lot of data and understand how to model things like a phoneme in speech. And how to understand people's language processing requires a lot of data collection as well. And at Google, which is a company that collects a lot of data, I also did personalization there which requires a lot of data to understanding a person's behavior. So, that to me is data science. Using data to build a useful model or to understand a particular pattern that is useful, then later on, for other software applications.

Copyright © 2014 Udacity, Inc. All Rights Reserved.

Gabor Introduction

So my name is Gabe Savo, I work at Twitter, and I am data scientist. I actually come from a background that's towards natural sciences. I did statistical physics, and I have a PhD in statistical physics. And anything that goes with that. Obviously, I was looking at like a lot of big, big systems as an interaction of, of very small very small entities composing of these systems. And and later on I did complex network research. So that means that we have interactions again. Imagine like a big gas composed of molecules. But instead here we have like, the humans interacting with, with each other through social networks. Through mobile communication networks. So that was the main focus of my research all year.

Gabor - What is Data Science

That's a great question. So what do data scientists do? I think it's it's really hard to to pinpoint exactly what they do because it's going to be tailored to their actual application area that they work. But in general, what they do is they take data and they find meaning in the data. And what the meaning is going to be really geared towards what they would like to explain. So it could be that a particular company, if they are looking at a company or, or the project. If they are looking for some, some particular signal or something. I think in general in my mind what data science does is use is use this data. Data sciences uses data to essentially explain and perhaps predict behavior be it human behavior or even the behavior of a more machine generated system, anything could be like that.

Quiz: Basic Data Scientist Skills

Just to recap, let me ask you a quick question. What does it mean for a data scientist to have substantive expertise and why is it important? Type your answer in the box below. Don't worry, your response won't be graded.

Basic Data Scientist Skills

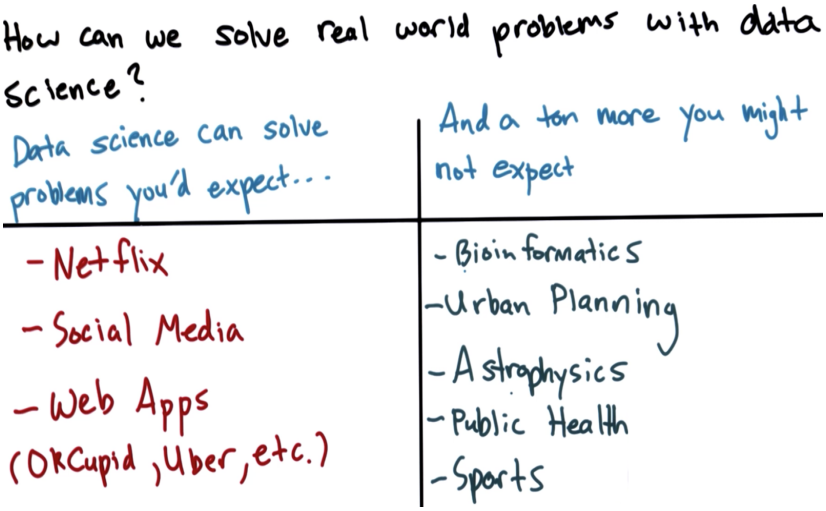
What does it mean for a data scientist to have 'substantive expertise' and why is it important?

Answer:

Copyright © 2014 Udacity, Inc. All Rights Reserved.

As we discussed earlier in this lesson, a data scientist needs to have substantive expertise. What does that mean? Well typically it means that a data scientist knows which questions to ask, can interpret the data well and understands the structure of the data. You can imagine that a data scientist needs to know about the problem that solving. For example, if you are solving an online advertising problem, you want to make sure you understand what types of people are coming to your website. How they are interacting with the website and what different data means that can help you ask the right questions like, Are people falling off and not completing our ads at a certain point in the flow, or do people complete more ads at a certain time of the day? You would also, then, be very familiar with how the data is stored and structured. And that could help you work more efficiently and more effectively. This is why it's important for a data scientist to have substantive expertise. It's important to note that data scientists usually work in teams. So it's normal for data scientists to be stronger in some areas and weaker in others. So even if you, as a data scientist, don't have a tons of substantive expertise, if you have great hacking skills or know a lot of statistics you can still be a valuable member of a data science team.

Problems Solved by Data Science



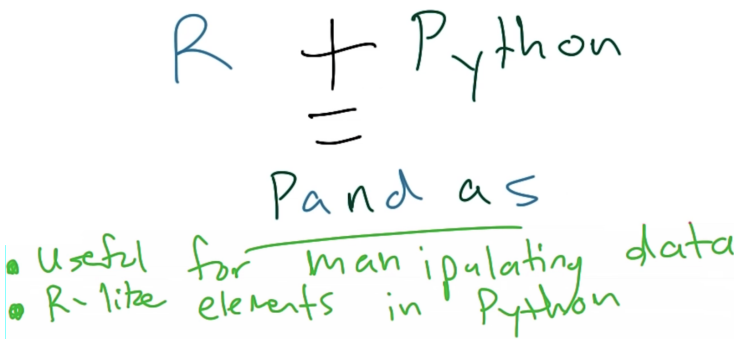
Now that you have a better Idea of what data science is, and what data scientist do, let's talk about how data science can be applied across a wide spectrum of industries. You might have signed up for this class under the notion that if you become a data scientist, you'll end up working for a Silicon Valley startup. Well it's true that most tech companies do employ data scientists. Data science can also be used to solve problems in many different fields. What are some examples of the types of problems being solved using data science? Well for one, Netflix uses collaborative filtering algorithms to recommend movies to users based on things they've previously watched. Also, elements of many popular social media websites are powered by data science. Things like recommending new connections on Linkedin, constructing your Facebook newsfeed or suggesting new people to follow on Twitter. Many other online services or apps such as dating websites like OKCupid or ride sharing services like Uber uses the vast amount of user data available to them, not only need to customize and improve the user experience but also to publish interesting anthropological findings regarding people's behavior in the offline world on their blogs.

Copyright © 2014 Udacity, Inc. All Rights Reserved.

Okay, so I know what you're thinking. So far, all of these seem like problems data scientists are expected to solve, but data scientists work in many domains. Data science concepts are integral in processing and modeling data in the field bioinformatics where scientists are working on projects like annotating genomes and analyzing data sequences. This past summer, data science for social good fellows in Chicago worked on a project attempting to solve Chicago's bus crowding issues using data. In addition, physicists use data science concepts when building a 100 terabyte database of astronomical data collected by the Sloan digital sky server. This one's cool. Analyzing electronic medical records allowed the city of Camden, New Jersey to save enormous amounts of money by targeting their efforts towards specific buildings accounting for a majority of emergency admission. Finally, NBA teams like the Toronto Raptors are installing a new technology, sports view cameras, on the basketball courts. They collect huge amounts of data on players movement and playing styles. This helps teams to better analyze game trends and improve coaching decisions. You've probably noticed by now that data science is making an impact in areas far and wide. Data science isn't simply a trendy new way to think about tech problems. It's a tool that can be used to solve problems in a variety of fields. Data scientists are working at Silicon Valley start-ups to enrich our online experiences, but they're also doing important work in our cities, in our laboratories and in our sports stadiums.

Pandas

As you can imagine, since data science is being deployed in such a wide range of fields. Data scientists use many different tools. Depending on the task at hand. One of the most versatile and ubiquitous is a Python package called Pandas. Which we'll use in order to handle and manipulate data during this course. You might wonder why we'll be using Pandas as opposed to another tool. Pandas allows us to structure and manipulate our data in ways that are particularly well suited for data analysis. If you happen to be familiar with the scripting language R, Pandas takes a lot of the best elements from R and implements them in python.

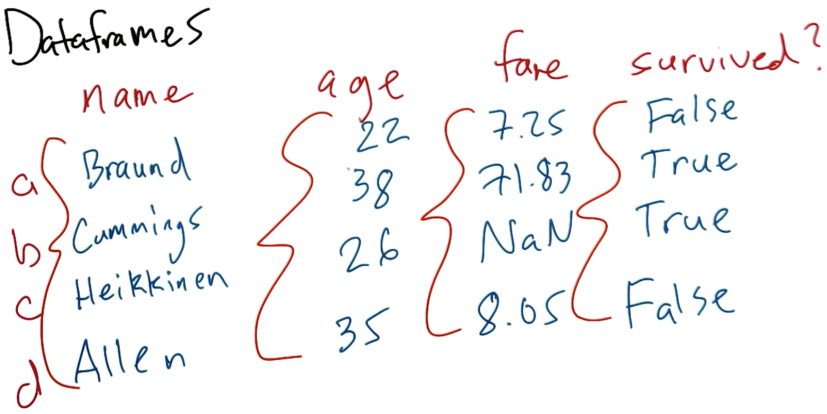


Dataframes

First of all, data in Pandas is often contained in a structure called a dataframe. A dataframe is a two-dimensional labeled data structure with columns which can be different types if necessary. For

Copyright © 2014 Udacity, Inc. All Rights Reserved.

example types like string, int, float, and Boolean. You can think of a dataframe as being similar to an Excel spreadsheet. We'll talk about making dataframes in a bit. For now, here's what an example dataframe might look like. Using data describing passengers on the Titanic, and whether or not they survived the Titanic's tragic collision with an iceberg. Note that there are numerous columns. Name, age, fare, and survived? These columns have different data-types. There are also some Not-a-Number entries which happen when we don't specify a value. There are a bunch of cool things we can do with this data frame. Let's jump to the command line. Say that I had already loaded this data into a data frame called DF. We can operate on specific columns by calling on them as if they were keys in a dictionary. For example, DFH and we can call on specific rows by calling a data frame objects loc method, and passing the row index as an argument, for example, df.loc('a').



Create a New Dataframe

Panda also allows us to operate on your data frame in a vectorized item by item way. What does it mean to operate on a data frame in a vectorized way? Well first let's create a new data frame. Note that first I want to create a dictionary where the keys are going to be my column names and the values are series corresponding to their values and then the indexes for the rows where these values should appear. In order to make a data frame, I can simply say df equals data frame of this dictionary d. Let's see what this data frame looks like. We can call dataframe.apply and pass in some arbitrary function. In this case, numpy.mean to perform that function on every single column in the data frame. So when we df.apply numpy.mean, what we get back is the mean of each column in our data frame df. There are also some operations that simply cannot be vectorized in this way, that is, take a numpy.mean as their input, and then return an array or a value. So we can also call map on particular columns or apply map on entire data frames. These methods also accept functions, but functions that take in a single value and return a single value. For example, if I were to type df1.map lambda x, x greater than or equal to 1, what I get back here is whether or not every single value in the column 1 is greater than or equal to 1. Now say that I were to call df.applymap lambda x, x greater than or equal to 1, what this function returns is whether or not every single value in the data frame df is greater than or equal to 1. This is just the tip of the iceberg when it comes to Panda's functionality. If you're interested to read more about what the library can do, you should check out the full documentation at the URL

Copyright © 2014 Udacity, Inc. All Rights Reserved.

contained in the instructor notes. Now, we know some of the very basics when it comes to handling the data, but how do we acquire the data that we wish to handle and analyze?

Lesson Project - Titanic Data

Alright. Before we get started on the class project, this lesson's assignment will allow you to get comfortable with the type of work that data scientists do using a small and classic data set. This data set describes the riders on the Titanic and a bunch of information about them. For example, what class they were in, whether they were male or female, how old they were, etcetera. Over the course of the assignment, you'll build a few different models. The models will start out simple, but they'll get increasingly complex. To see if using data science, we can predict who will survive and who won't survive the Titanic tragedy. This may sound complicated, but don't worry. We'll give you plenty of help. This assignment is meant to get you in the habit of thinking like a data scientist.

Class Project

Through the class project, you'll investigate New York City subway ridership as a data scientist might. First, you'll pull some publicly available data on subway ridership, and also on New York weather conditions, using the New York MTA website and the Weather Underground API. Then, you'll answer some questions about subway ridership using statistics and machine learning. Does the weather influence how many people ride the subway? Is the subway busier at certain times than others? Can we predict subway ridership? Finally, you'll develop some charts and graphics that communicate your findings, and synthesize everything into a cohesive write-up that your friends or family might find useful and informative. This may sound daunting, but we'll be going through this step by step, and learning how to use the necessary tools as we go along.

Pi Chuans Advice for Aspiring Students

So for me, the reason why I even come into this field of data science is my passion for something specific, which is natural language. Right, so I observe the people who work around me who knows about data science. I think one thing that's very important is they either have a passion for the particular data they're looking at. Like, you know, natural language, or like speech recognition kind of

Copyright © 2014 Udacity, Inc. All Rights Reserved.

data. Or some people are just very interested in patterns in data. Like when they see some data they would try to calculate the mean of the data, the variance of data. It comes natural to them because they want to find patterns to the data. So I think for anyone who wants to become a data scientist, it's good to think about, what kind of data you are interested in doing, and start with that. And then later on when you have this skill of analyzing data, you can apply it to any other kind of thing.

Gabors Advice for Aspiring Data Scientist

I think the, the most important thing that, that inspired me to start to start keep In mind is that they should have a very curious mind. They should have the ability to ask questions, to formulate these questions as it pertains to them, as they would see these questions being raised in their own lives. So if, if there's a problem they see with the pieces that they are working with or with the project that they're working with, they should try to ask these questions in terms of and how they can understand for themselves. And once they understand, once they know what is the gist of the question is, then they can go and use algorithms. Obviously it helps tremendously if you have experience about all the arguments that are out there to attack these questions. But I think the most important ability is that you should have the mindset as a data scientist, and you could obviously improve throughout your career in this if you have this inquisitive mindset, where you are trying to ask the right questions. While you are trying to, to see what is important, you should also have an overview of what kind of data can support your conclusions and draw conclusions with the help of these algorithms that you are going to use to solve these problems.

Recap of Lesson 1

To recap today's lesson, data science is a discipline that incorporates methods and expertise from a variety of different fields, to glean insights from large data sets, and then use those insights to build data driven products, or to create action items. There's no cookie cutter way to become a data scientist, but a lot of data scientists have a very strong background in mathematics and statistics, and also have the programming skills to handle these data sets and also to perform analysis. Currently, data science is most closely associated with the tech field, but more and more data science is being used all over the world in a variety of different fields. In order to solve new and old problems, more efficiently and effectively. Now I know what you're thinking. This sounds awesome, data science sounds really cool, I want to work on a project get me some data. Unfortunately, data seldom comes in formats that we want. It might be in a format we've never seen, it might come from a weird source. There might be missing or erroneous values. Because of this, any data scientist worth their salt is skilled at acquiring data. And then massaging it into a workable format. That's what we're going to be covering in the next lesson.

Copyright © 2014 Udacity, Inc. All Rights Reserved.