Welcome to the Johns Hopkins Data Science

Track. I'm incredibly excited to tell you a

little bit about the track and about where you're going to be going

over the next nine months. My name is Jeff Leek, and I'm a professor in the Johns Hopkins Bloomberg School ofmmmmmmmdmmmmmmm

Public Health. I thought I'd lead off this introductory

video with a quote by one of my favorite US Presidents,

Teddy Roosevelt. He said it's not the critic who counts. It's not the person who points out how the

person who's actually doing things is doing them

wrong or messing up. It's the person who's actually trying to

get things done, even when there are obstacles in the

way. And a lot of data science right now is

being able to push through a lot of the difficulties that you have when you're dealing with either large or messy

data. It includes collecting the data clean them

up and then building new announced techniques that

exploring new information about that data. And so, all of those steps are a little

bit complicated and sometimes it opens you to criticism when you're trying to do something new and

interesting. And so I wanted to lead with a quote that

said it's important to strive the valiantly do these sorts of things, even if you're going to take some

criticism. So the key challenge in data science is

actually really nicely summed up in this quote by

Dan Myer. He says, ask yourselves, what problem you,

have you ever solved, ever, that was worth solving, where you knew all of

the given information in advance? Where you didn't have a surplus of

information and have to filter some of it out, or you didn't have insufficient

information and have to go find some? And so, I think that this is a kind of a

critical quote because, in data science, this is

usually what is going on. You're either in a situation where you

really don't have enough data to answer the question that

you're interested in, and you have to go out and try to search for

it, find it on the web, or find it in other

places. Or you're in a situation where you are

overwhelmed with a surplus of data and you have to filter out all of the irrelevant information to try to narrow in on your

question. And you'll notice that I said question in

both of those cases. And I think this goes to the heart of our

philosophy about data science. We're interested in answering questions

with data. We think the question should come first

and then the data should follow after. And that actually makes it more

challenging, because sometimes, you can answer a

question with some data but you might not be able to

answer your question with some data. So this track is about refocusing on

answering the question that you're interested in solving with the

data that you have. So I thought I'd tell you a little bit

about the instructors that you'll be hearing from throughout the

course of this, course track. So we are all faculty in the Johns Hopkins Blumberg School of Public Health in the

Biostatistics Department. And you could say that we all do data intensive statistics in biology

and medicine. Brian Caffo works on the statistics of

brain, analyzing brain imaging data. And I work on the statistics of analyzing

genomics data. And Roger Peng works on the statistics of

analyzing fine particulate matter. All of us work on problems where the data aren't always clean and nice and easy to

handle. All of us work on problems where the

questions that we want to answer are complicated and you have to

break them down into parts. And all of us, sort of, work on questions

where we're very passionate about trying to get the right answer so that we

can help people in human health. But the techniques that you're going to be learning about are not exclusive to

biology and medicine. That's just one area where there's been a recent upsurge in the amount of data

that's available. So why data science? Why should you take this program? This is a cover of The Economist now. It's a little bit old I guess ancient

history from a couple of years ago. But it talks about the data deluge and

it's really true. Over the last several years data has

become much, much cheaper to collect. It's much easier to store. And there's so many free computing tools

out there right now, that you can actually do something with this entire data deluge

that's sort of assaulting all different areas of

science and business. So the other thing is that you've probably

heard the term big data. And so we'll hear a little bit more about

what we think about big data throughout the course of this particular course, the Data

Scientist's Toolbox. But big data is, sort of a new frontier in

the sense that, we have data in areas that we didn't used

to have that data. We didn't have access to information about

GPS coordinates from cars from everybody in

the entire world. It wasn't possible to sequence everybody's

genome. And now that's all possible. So we have access to this data and it allows us to answer questions we never

could before. So, it's an incredibly exciting time, and

you're somebody who can get in there and use that data to answer

those questions. So why statistical data science? You'll notice that we're, all of your instructors are biostatistics professors

and so this will, this data science track will

obviously have a little bit of a statistical bend. I think that that's appropriate given that statistics is the science of learning from

data. So, data is very, very, it's very rare

that you'll get a data set where all of the answers are really

clear, and there's no uncertainty. In any case where there is uncertainty, that's where statistics comes and plays a

role. So, this is a again, a little bit older

New York Times article now, but it talks about

how the key word for a lot of graduates to open

the door for a lot of jobs, is to learn about

statistics. So why are you lucky? You're lucky because this moment, right

now, in time is sort of like the moment that Jeff Bezos

discovered the internet. He got into building a internet company at

the time when there was this explosive growth

in internet usage and it just opened the door for the opportunity to build something amazing and

huge and wonderful. And sort of, that's the right, that's what

the time is right now for data. It sort of there's an explosive growth of data in every possible area you can

imagine. And so it's the opportunity right now to

sort of jump on a rocket and, and find out something interesting, and, and

sort of carry it off into a, a really major

endeavor. You're also lucky because tools and

competitions and websites have all been developed around

the idea of helping to learn about data, but also

getting involved in projects that have super high

profile results. So, one example is the Heritage Health

Prize, which I'm showing you a picture of here. The Heritage Health Prize was a $3 million contest for people who could analyze data

and come up with a better predictor of who would be admitted to a hospital in another

year. So you can see that's a huge amount of

money that's being invested in these ideas of algorithm development and data science of

prediction. So it gives you an exciting opportunity to

get involved in projects that, sort of, weren't happening

five or ten years ago. This course track will focus almost

exclusively on the use of the R programming language. And so I thought it was appropriate to

talk a little bit about why we like R so much. So we like R obviously, because we all use

it. But it's also sort of increasingly the most commonly used language for data

science. There are other languages that are also

very useful. And we won't be talking about them a lot

in this course but they're obviously good

complements to the R programming language. Like, Python, in this class we'll be

focusing on R because it has a broad range of packages that

allow you to go from the rawest of raw files, all the way

to interactive reports and documents and web apps that you can share

with your collaborators. So, some more reasons why we might use r

is because it's free, it has a comprehensive set of packages, like I

mentioned, for all the processes that are involved in

data science. It has one of the best development environments of any programming language,

in our studio. It also has an amazing ecosystem of

developers. And what I mean by that is there are a lot of people that are developing our

packages. And they're also available to get in touch

with on mailing lists or by email or on stack

overflow. And so it's really possible to learn about the cutting edge of packages that are

being developed. There also very easy to install and play

nicely together, which is a, a feature that doesn't always happen in a lot of the

languages that are used for data science. So the next thing I thought I would

mention is, who is a data scientist? So, we're going to be talking about data

science a lot. And I thought I'd mention that some people

that I think are data scientists, that might not, either label themselves that

way or have other people label them that way. So the first is Daryl Morey, who'd the

general manager of the Houston Rockets basketball

team in the US. So he uses data to analyze basketball players and transactions and

making trades. And so I would consider him to be a data

scientist Because he's a person who uses data to

answer questions about basketball. Another data scientist that you may, or

may not, have heard of, is Hilary Mason. So, she used to be the Chief, Data Scientist at Bentley, and now she's at

Accel Partners. And so, she uses data to answer all sorts

of questions about mining the web, and understanding that way that

humans interact with each other through social

media. So, again she might not label herself a

data scientist, but I think the way that she

uses data, is a evocative of the sort of ideas, that we would like to convey in this data

sciences track. If you're taking this course, you probably

know who Daphne Koller is. She's the CEO of Coursera. But she's also another person who's using

all the data they're collecting through Coursera

to better, to improve the way that we do educational delivery

and educational assessment at this huge scale that

Coursera is providing. And finally, Nate Silver is one of the

most famous data scientists, or statisticians

in the world today. So he used a large amount of totally free

public data to make predictions about who would win elections

in the United States, and was remarkably

accurate. I'm going to finish with him, as the da-,

as the last data scientist I'm going to

illustrate because it's so amazing that he could use public free data and create

such an amazing product that so many people read about,

and are excited about. So our goal is to teach you about a bunch

different skills that will be useful for you as a

data scientist. So, this is a Venn Diagram and some

statisticians and data scientists don't like Venn Diagrams but I'm going to

get, show you one anyway. And so, this Venn Diagram has Data Science

at the, sort of, the center of this Venn diagram that

intersects several different skills. So, if you look right here there's data science and it involves three

different components. There's hacking skills, there's math and statistics knowledge, and there's

substantive expertise. And so our data science track will focus a

little bit on each of these, but it will primarily focus on math and statistics knowledge and hacking

skills. And so math and statistics knowledge sort

of speaks for itself. We're going to teach you a little bit about math and a little bit about

statistics. Hacking skills involves two different

components. One thing is we're going to teach you a

little bit about computer programming or at least

computer programming with R, which will allow you to access

data and play around with it and analyze it,

plot it. But hacking skills also has another

component to it which is the ability to go out and answer

questions for yourself. One key component of a data scientist job

right now is that most of the answers aren't already

outlined in the textbook. This is all new stuff that's happening. So what are the major skills of being a data scientist is being to go to Google,

and go to Stack Overflow, and go to one of the

other sites and look up what you need to learn and figure out what answers you

know and what answers you don't know, and then

figuring out how you can use the information you

have to answer the question that you'd like to

answer. So another reason obviously is jobs. That might be the reason you're taking

this course track. And so you can see this is a plot of

listings are for data science jobs over time and of

course it's exploding. And we'll talk a little bit about why you

shouldn't extrapolate, necessarily, from your data forever, but

it does suggest that data science is a hot area that's growing and I

think obviously we're very excited about it and hope

you're excited about it too. So this course, Data Scientist's Toolbox,

will continue with lectures on the following

three things. First, we're going to introduce you to the

course track. Then we're going to tell you a little bit

about getting the tools that you need to get set up and get installed,

hopefully get you over that hump. And then we're going to give you the basic background on data science sort of

writ large, so that you'll be ready to jump into any of the individual classes and really

take off. Looking forward to seeing you in the rest

of the class.

+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

In week two of course, we're going to be covering a bunch of software that you're

going to install that will constitute the data

scientist's toolbox, as we described it for this

course's track. So the first question you might ask

yourself is, what software do you need? Well to know what, software you need, you

have to know what exactly a data scientist is

going to do. So, in this course sequence we're going to

talk about all the different components of

being a data scientist. So we're going to start with defining a

question of interest, and then identifying the ideal data set try to

ident, to answer that question. Determining if that data is even

accessible, a lot of times the ideal dataset isn't even

available. Then ways that you can go out and actually

obtain the data whether it's from a database, or from a website, cleaning the

data up so that it can be processed and

analyzed. Performing some sort of exploratory

analysis, including making plots and clusterings so

you can identify patterns that you didn't know

about before hand in the data set. Performing statistical prediction or

modeling to try to, build a sort of an intuition about what's going to happen

in the next sample you might take. Interpreting your results, challenging

them. Then synthesizing them and writing them up

in reproducible ways that can be shared with

other people. Finally, we're going to talk about

distributing results through things like interacting graphics,

also through right ups and presentations, and finally

through interactive apps built on top of R. So the main workhorse of data science in

terms of this data science track is the R

programming language. There are other alternative languages that

are also really great for data science, but

we're going to be focusing on R, since it's one

of the most widely used languages. And it's widely supported by a large group

of developers. Who can contribute new packages all the

time that can improve and extend the

functionality of R. We'll be installing this in the second

week of the class. We'll do most of our coding in RStudio. RStudio is an Integrated Development

Environment, an IDE for R. It's actually one of the best IDE's I

think for many other languages as well in terms of

data science. The R IDE is free as well just like the

language R, and so we will be downloading this IDE and setting

it up again the second week of class. The interface looks something like this. And we'll talk a lot more about this in

the second week and later on in the rest of

the class. But you can see here in the top left-hand

corner I've got a file. So this is a new .R file that's going to contain some code that we're going to be

writing in. So we can write that code, here in the

file at the prompt and then down here, you see

a console. So we'll be entering sort of a commands at the command line down here in this

console. And then over here you can see other information you might be interested in

looking at. See plots you recently made, the packages

that you have loaded, or help files for specific functions that

you might be interested in. There are a lot of other really nice

functions that come with Rstudio, and we'll be talking about

those more throughout the class. The primary type of file that we'll be

interacting with, for the most part in this class, is an R

script. So, an R script is a file with the

extension .R, and so it's just a, actually a text

file. Except the text file contains bits of R

code, so here it's you can see a comment. So this isn't actually executed but R you

could include that so that people can understand what's

happening in the code. And then there're things like functions

and so forth which we'll be talking about a lot more when

we're coding. If this seems intimidating to look at this

function right now you should worry about it when you get

through R programming. You'll be a wizard and be able to do

things much more complicated than this. The other thing that we'll be using is R

markdown documents. So, reproducible research involves

creating documents that can be reproduced. In other words, they can be rerun and

produce the exact same numbers that you got when you

did your analysis. And the primary vehicle for doing that is

through markdown and R markdown. So this is a file with an extension ,RMD

and this .RMD file has a very structured

format of text file. And so we'll talk a lot more about what

that format is later but you could take this structured file and you

can knit it to html with this button here. And you actually create an html file that

will actually be formatted very nicely. So for example, what you type in text

looks like this, and it turns into a nice bulleted list in

HTML, once you knit HTML. And we'll talk a lot more about how that

file works later in the class. We're going to talk about how we are

going to do distributed version control with Github

and Git. So, part of this class will be setting up your Github account and creating a

portfolio, of all the different things that you do

throughout the course track, that then you can share with

employers. Or you can share and contribute to other

projects, so that you can get your name out there in the

data science community. We're going to running most of the

commands from the shell or from the command line

interface. So this is a command line interface, it

doesn't look like much right here. You can see that there's a prompt up her

in the top left. And we're going to be entering commands as

text prompts. And those commands will then execute,

allowing the programs that we're going to be

talking about. So, there's a brief tour of all the tools that we're going to be using in this

class.

+++++++++++++++++++++++++++++++++++++++++

This is a brief follow-up video to getting

help and it's about finding answers. The reason why there are two videos about

this is because it's such a critical skill in

data science. So you can see one of the three

fundamental sort of, skills you see in the Venn diagram here is

Hacking Skills. And the reason why it's one of the three fundamental skill is because almost

of none of the knowledge that you will need is already sort of set in standardized text

books. It's often scattered in a bunch of

different places and you have to be able to sort of synthesize it or find the

information that you need about. Whether it's about which data set you need

to be using, or the statistical analysis you need to be doing, or the R Package

that you need to be using. All of this is sort of scattered around,

and you have to be willing to do a little bit of hard work and elbow

grease to find it yourself. Obviously we'll tell you as much as we can in lectures, but we're necessarily limited

by the amount of time that we can lecture every week, and

so it's important to be able to find that

information yourself. So key, some key characteristics packers

are that they're willing to go out and find the

answers on their own even if it takes a little bit

of time or a little bit of effort. They're knowledgeable about where to find

those answers whether its Google, or stack over glow, or

cat. Cross validated or a message fo history of

mailing lists. They're unintimidated by new data types or

packages. It's very common as a data scientist to be

thrown a new kind of data or a new kind of Our package that you

need to learn very quickly to be able to analyze the data to being

unintimidated by that is important and then being unafraid to say that you

don't know the answer to a question. And so a key characteristic I would say

the way to summarize it up is being alive but

relentless here. Going after the answer and you just trying

to find it But you're very polite while doing

it. And so Google knows this too. In their hiring practices they're looking

for these sort of characteristics. The kind of people that will go after

these sorts of things as it's described in this article,

I'm linked to here. [SOUND] So an important question is where to look for, for different types of

questions. So for our programming you might want to

go straight to the the archive of the class

forums. Where the class you're taking will focus

on very specific questions or functions. And there'll be a large group of

interested people. You could read the manual or help files like I showed you in the getting help

lecture. You can search on the web. That's actually one of the best ways to do

it. You can ask a skilled friend. That's even better if you've got a person

that you know that already is a bit of a data

scientist. They can often help you out. And then you can post to the class forums

and try to get your answers. Remember to be specific with your

questions. You can also post to forums outside of the

class. The R Mailing List or Stackoverflow, if

you have R questions. For data analysis or statistics type

questions, you want to go to start again with the class forums, and then go

to the web or to friends. And then there's another outside forum

called CrossValidated where you can ask these types of

questions. Brother software you might have to go to

software specific websites. So forget HUB, they have a lot of

tutorials and nice information that you can use to get

answers there. [SOUND] So an important point to know is

that Googling data science questions isn't always the

easiest thing in the world. So, the best place to start with if you have a pretty general question is often in

the forums. And if people can direct you to where you

should be searching outside of the forums. Keep in mind that Stackoverflow with the

tag R is a really good place to get information

about R. And, so, you have to use this tag because

if you just use the letter R, it obviously is, sometimes,

a little bit hard to search for. You can also try the R mailing list for software questions or CrossValidated

for more general questions. Usually what I've found is that if I'm

going to work in Google, searching Google, I use

usually type something like the data type and then data

analysis or I type the data type and then the R

package. I found that data type R package often

works a little bit better than data type, data

analysis when Googling things. Another thing to keep in mind is that data analysis or data science is often called

something different depending on what kind of data you are looking at, so for example medical data it might be

called biostatistics. For data from the web it might be data

science. For data in computer vision it might be

machine learning or natural language processing for data

from text and so on. So, knowing what the right word to Google

is, is often half the battle. And so, you can often find that out by

posting to forums and people will let you know what the right

word to be googling is. [SOUND] Again the credits for this go to

Roger's Getting Help Video. And it was inspired by Eric Raymond's How

to Ask Questions the Right Way.