- Hi, again.

provided by their data science teams.

Because companies take action based on these insights,

data science teams need to be experts in their practice

to ensure those insights are well-reasoned.

You've likely just begun hearing more

from the media about data science

and from employers about the demand for data scientists

so it might seem like data science came out of nowhere.

However, data science has been around for a very long time.

Scientists have always used data to gain insight

based on observations, so why then is data science

suddenly on the rise?

The answer lies in two things.

First, our ability to collect data in real time

has ballooned with data coming from a variety of places

including real time environmental sensors,

websites, smart phones, and a variety of other sources.

In turn, this influx of data has increased demand

for large scale data processing.

This data growth combined with the advances in storage,

networking, and computing at scale

has brought us to a new era of data science.

Many dynamic data driven applications in this new era

build upon data driven predictions to support decisions,

just like the Amazon book prediction example we discussed.

It is nearly impossible to find an industry,

scientific discipline, or engineering endeavor today

that is not impacted by data science.

One need only look at the major trends

in smart cities, precision medicine,

energy management, and smart manufacturing

to see how it is shaping our economy today,

and all these fields are looking for expert

in a combination of advanced data analytics,

the traditional modeling, and simulations.

We started by saying that we are collecting

more data than ever before, but how much data

and in what form are we really talking about?

Let's take a look.

The data can include anything from user preferences

and purchasing history on websites

to scientific data from remote sensors and instruments

and personal health data from variable devices

and social media data related to customer satisfaction,

political trends, health epidemics,

law enforcements and terrorists activities,

as well as medical data from drug trials,

treatment options, and patient population.

This is probably already sounding like a lot of data,

but we could look at this differently.

If you look at just one minute on the internet

we'll begin to fully grasp the massive size of data produced

and data stored every minute.

Every minute, 204 million emails are sent,

200,000 photos are uploaded

and 1.8 million likes are generated on Facebook.

On YouTube, 2.78 million videos are viewed

and 72 hours of video are uploaded.

It is not any different for scientific data.

HPWREN, the High Performance

Wireless Research and Education Network

that only connects sensors in San Diego,

Riverside, and Imperial Counties,

Thank you for making the first step

into your data science journey in Python.

Before we start Python programming in week two,

we'll start our course by reviewing what data science is

and how we conduct data science research.

By the end of this video, you will be able to

describe what modern data science is,

explain why data science is the key

to getting value out of data,

and where the growing interest for it comes from,

and list a recommended set of skills for a data scientist.

We have all heard it.

Data science turns data into insights or even actions,

but what does that really mean?

Data science can be thought of

as the basis for empirical research,

for data is used to inform our hypotheses

and provide observations.

In many cases, this data is used either

by businesses or by scientists

to inform their understanding of a phenomenon.

Because there are often large troves of data

which we can mine for insights,

we often call this big data.

Insight is a term we use to refer

to the data product of data science.

It is extracted from a diverse amount of data

through a combination of

exploratory data analysis and modeling.

The questions we ask are sometimes quite specific,

but sometimes it takes looking at the data

and patterns in it to come up with a specific question.

Another important point to recognize

is that data science is not a static, one-time analysis.

It involves a process where the models we generate

lead to insights and those insights are then improved

by gathering further empirical evidence, or simply, data.

For example, a book retailer like amazon.com

can constantly improve the model

of a customer's book preferences

using the customer demographics,

his or her previous purchases

and prior book reviews by the customer.

Their models also likely take into account

the similarity of customers to detect common interests.

The book retailer can also use this information

to predict which customers are likely to like a new book

and take action to market the book to those customers.

This is where we see insights being turned into action.

As we have seen in the book marketing example,

using data science and analysis of the past

and current information, data science generates actions.

This is not just an analysis of the past,

but rather generation of an actionable information

for the future.

This is what we call a prediction

and it is very similar to a weather forecast.

When you decide what to wear based on

the forecast for the day, you're taking action

based on insight delivered to you.

Just like this, business leaders and decision makers

take action based on the evidence

collect 30 terabytes of data annually.

in the order of terabytes,

but petabytes are becoming more common to our daily lives.

CERN's Large Hadron Collider

generates 15 petabytes of data a year.

According to a report by IDC,

sponsored by a big data company called EMC,

digital data will grow by a factor of 44

until the year 2020.

This is a growth from .8 zettabytes in 2009

to 35.2 zettabytes.

A zettabyte is one trillion gigabytes.

That is 10 to the power of 21.

The effects of it will be huge.

Think of all the time, cost, and energy that will be used

to store and make sense of such an amount of data.

The next era will be yottabytes, that is 10 to the power 24,

and brontobytes, that is 10 to the power 27,

which is really hard to imagine for most of us at this time.

This is also what we call data at an astronomical scale.

The choice of Milky Way Galaxy in the middle of the circle

is not just for aesthetics.

That is what we would see if we were to scale out

10 to the power 21 times into the universe.

Cool, isn't it?

The bottom line is that all of these sources

point to an exponential growth in data volume and storage.

While many of us are excited

by the opportunities offered by big data,

this rapid growth also comes

with a number of management and analysis challenges,

least of which is information overload.

Our challenges isn't just to manage the data

but to try to see how everything is connected.

Finding the connections

between the kinds of data sets we've discussed

has the potential to lead to interesting discoveries.

Such an endeavor requires proper use

of data management, data driven methods,

scalable tools for dynamic coordination

and scalable execution,

and a skilled interdisciplinary workforce.

This is where you come in the picture.

By putting your time into skills and programming in Python,

statistics, machine learning, and big data,

you will be ready to take on

some of the technical challenges in data science

like drug effectiveness analysis,

crime pattern detection, and self-driving cars.

As a summary, a data science team often comes together

to analyze situations or answer questions

in business or science which no single person

could solve on their own.

There are lots of moving parts to the solution,

but in the end all these parts should come together

to provide actionable insight based on data science.

Being able to use evidence based insight in your decisions

is more important now than ever.

This MicroMasters will provide you

with related technical skills

on Python programming, statistical analysis,

machine learning and big data tools to make this happen.

Leo and I look forward to giving you

the fundamental data analysis skills in Python

that you will use throughout the entire MicroMasters.

We use HPWREN data collected from weather stations

throughout San Diego County

for wildfire monitoring and modeling.

This consists of daily amount of

half a gigabyte environmental sensor data

and four gigabytes of camera data throughout 18 stations.

This may not sound like a lot,

but this is just one system for three counties.

NASA's MODIS, or Moderate Resolution

Imaging Spectroradiometer

is a satellite that has imaging instruments

on two satellites called Aqua and Terra.

MODIS instruments on these satellites

capture images of the entire surface of Earth

every one to two days, acquiring data in 36 spectral events.

This equals 40 science products

and produce 600 gigabytes of data per day

which equals 219 terabytes of data per year.

It's not that different in precision medicine.

One of the key promises in precision medicine

comes from using individuals' genetic profile

to guide decisions regarding prevention,

diagnosis, and treatment of disease.

Genome sequencing is only one part of data

and it needs to be augmented with treatment data,

medical histories, and other biomedical data.

According to a Fast Company article in 2016,

only the genome sequences of people

who will be diagnosed with cancer

was predicted to equal four exabytes.

That's a lot of data.

Other large volume data sources in scientific research

comes from LIGO, Deep Space Network, and Protein Data Bank.

LIGO, the Laser Interferometer

Gravitational-Wave Observatory,

is a data source that led to

the gravitational wave discovery in 2016.

The experiment provides large scale physics and observatory

to detect cosmic gravitational waves.

Deep Space Network, which is NASA's network

of large antennas and communication sites

located in several countries

that are used to support space missions

and research asteroids and planets,

updates its data stores

with real time data every five seconds.

Another research product is the Protein Data Bank,

which is a repository of information

about 3-D structures of large biological molecules,

which is important for research

on human health and disease and drug development.

Management and analysis of such scientific data sets

is a huge challenge for modern scientific research,

and in there you heard me say words that start with peta,

exa, and even yotta to define a size,

but what does that all really mean?

For comparison, 100 megabytes

will hold a couple of encyclopedias.

A DVD is around five gigabytes,

and one terabyte would hold around 300 hours

of good quality video.

A data oriented business currently collects data