* **Data science** turns data into insights or even actions, but what does that really mean?
* Think of it asbasis for empirical research where data is used to inform hypotheses + provide observations.
* In many cases, this data is used either by businesses or scientists to inform their understanding of a phenomenon.
* B/c there are often large troves of data which we can mine for insights, we often call this **big data**.
* **Insight** refers to the **data product** of data science + is extracted from a diverse amount of data through a combination of exploratory data analysis + modeling.
* The questions we ask are sometimes quite specific + sometimes it takes looking at the data + patterns in it to come up w/ 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 + those insights are then improved by gathering further empirical evidence, or simply, data.
* Ex: Amazon can constantly improve the model of a customer's book preferences using customer demographics, their previous purchases + prior book reviews by the customer.
* The models also likely take into account the similarity of customers to detect common interests.
* They can also use this info to predict which customers are likely to like a new book + take action to market the book to those customers.
* This is where we see insights being turned into action.
* Using data science + analysis of the past (historical) + current (near-real time) info, *data science generates actions* = generation of actionable info for the future = a **prediction**
* When you decide what to wear based on the weather forecast for the day, you're taking action based on insight delivered to you.
* Similarly, business leaders + decision makers take action based on evidence provided by their data science teams.
* B/c companies take action based on their insights, data science teams need to be experts in their practice to ensure those insights are well-reasoned.
* Data science has been around for a very long time + have always used data to gain insight based on observations
* Data science is suddenly on the rise due to 2 things.
* Our **ability to collect data** **in real-time** has ballooned, w/ data coming from a variety of places, including real-time environmental sensors, websites, smart phones, + a variety of other sources.
* In turn, this influx of data has increased demand for **large scale data processing**.
* Data growth combined w/ the advances in storage, networking, + 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, like Amazon book predictions.
* 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, + smart manufacturing to see how it is shaping our economy today, + all these fields are looking for experts in a combination of advanced data analytics, traditional modeling, + simulations.
* Collected data can include anything from user preferences + purchasing history on websites, to scientific data from remote sensors + instruments + personal health data from variable devices, to social media data related to customer satisfaction, political trends, health epidemics, law enforcements + terrorists activities, as well as medical data from drug trials, treatment options, + patient population.
* Every minute, 204 million emails are sent, 200k photos are uploaded+ 1.8 million likes are generated on Facebook.
* On YouTube, 2.78 million videos are viewed + 72 hours of video are uploaded.
* It is not any different for scientific data. HPWREN, the High Performance Wireless Research + Education Network that only connects sensors in San Diego, Riverside, + Imperial Counties, collect 30 TB of data annually.
* We use HPWREN data collected from weather stations throughout San Diego County for wildfire monitoring + modeling
* This consists of a daily amount of 1/2 GB of environmental sensor data + 4 GB of camera data throughout 18 stations.
* This may not sound like a lot, but this is just 1 system for 3 counties.
* NASA's MODIS, or Moderate Resolution
* Imaging Spectroradiometer
* is a satellite that has imaging instruments
* on two satellites called Aqua + 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
* + 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, + treatment of disease.
* Genome sequencing is only one part of data
* + it needs to be augmented w/ treatment data,
* medical histories, + other biomedical data.
* According to a Fast Company article in 2016,
* only the genome sequences of people
* who will be diagnosed w/ 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, + 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 + observatory
* to detect cosmic gravitational waves.
* Deep Space Network, which is NASA's network
* of large antennas + communication sites
* located in several countries
* that are used to support space missions
* + research asteroids + planets,
* updates its data stores
* w/ 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 + disease + drug development.
* Management + analysis of such scientific data sets
* is a huge challenge for modern scientific research,
* + in there you heard me say words that start w/ peta,
* exa, + 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,
* + one terabyte would hold around 300 hours
* of good quality video.
* A data oriented business currently collects data
* 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, + energy that will be used
* to store + make sense of such an amount of data.
* The next era will be yottabytes, that is 10 to the power 24,
* + 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 + storage.
* While many of us are excited
* by the opportunities offered by big data,
* this rapid growth also comes
* w/ a number of management + 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
* + scalable execution,
* + a skilled interdisciplinary workforce.
* This is where you come in the picture.
* By putting your time into skills + programming in Python,
* statistics, machine learning, + big data,
* you will be ready to take on
* some of the technical challenges in data science
* like drug effectiveness analysis,
* crime pattern detection, + 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
* w/ related technical skills
* on Python programming, statistical analysis,
* machine learning + big data tools to make this happen.
* Leo + I look forward to giving you
* the fundamental data analysis skills in Python
* that you will use throughout the entire MicroMasters.
* End of transcript. Skip to the start.