1. Welcome to Big Data Analytics for Healthcare. I'm Jimeng Sun, associate professor in college computing at Georgia Tech. I'm here at Children's Healthcare, Atlanta to tell you a little bit about this course. First, let me tell you a little bit about me. I was born in Beijing, China. >> [COUGH] >> No? I'll skip ahead. My research area is in healthcare analytics and data mining. Before Georgia Tech, I worked at IBM TJ Watson Research Center in Healthcare Informatics. I also enjoy swimming. In this lesson, we'll discuss what you'll do for this course and what you will learn, and why you should care.

2. This course exists at the intersection of house care and big data. On the one hand, we have house care, we talk about house care applications as well as house care data. On the other hand, we have data science and big data analytics. We'll talk about different algorithms. And systems for processing and analyzing big data. We'll focus on the intersection between these two and how one is applied to the other.

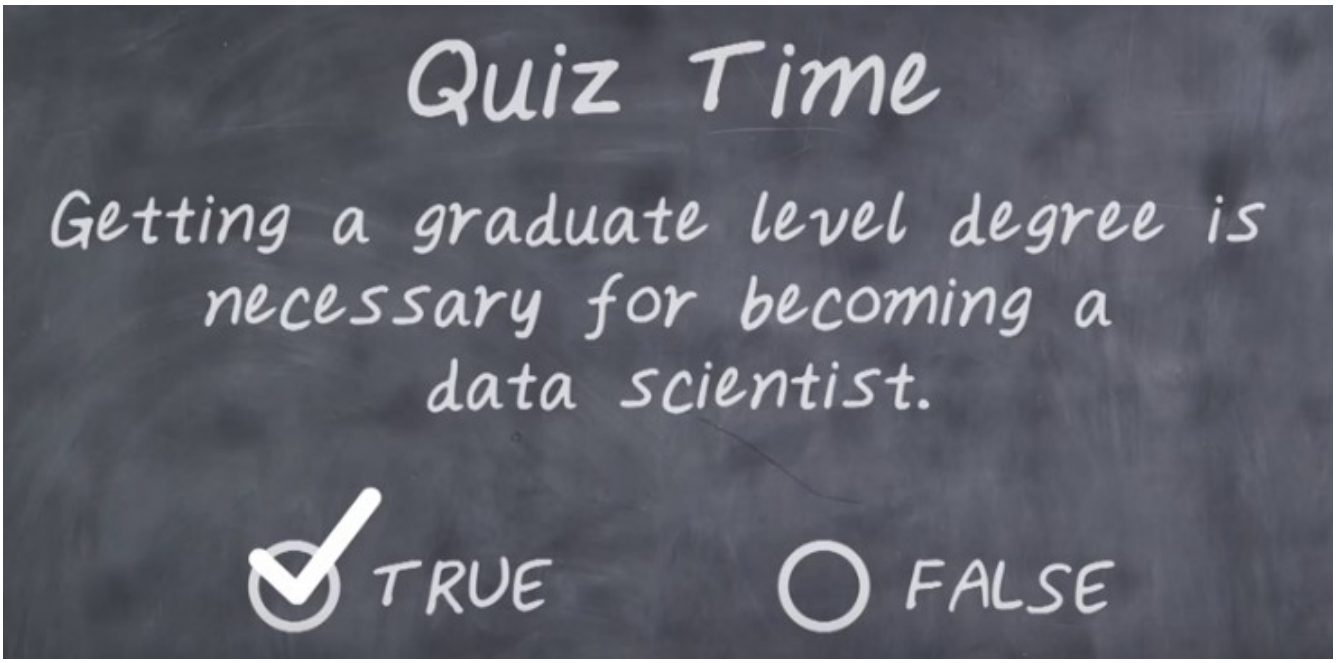
3. Our learning goals for this class are understanding healthcare data, understanding different analytics algorithms, and understanding big data systems. >> What will I be able to do once I complete this course? >> You'll be able build models on healthcare data. For example, models for individual disease risk prediction, recommending treatments, cluster a patient into groups with common characteristics, and find similar patients. [SOUND] >> How will I be assessed? >> You do have few homework assignments, using big data tools. You do a project involving building a system to analyze healthcare data, writing a report, and giving a presentation.

4. Health care industry is huge, and there are a lot of data coming out of health care. U.S. health care is incredibly expensive. Overall spending is $3.8 trillion in U.S. per year. That's more than the value of ten biggest companies plus ten Beijing Olympics plus a Warren Buffet, and a Bill Gates. But does it have to be that way? There is massive waste in health care unfortunately. The estimated waste in U.S. healthcare alone is $765 billion. That's equivalent to the NASA's total budget for the past 50 years. Not only the cost is an issue, but also the quality of healthcare is poor. There are 200 to 400 thousand preventable deaths per year in the U.S. That's 1,000 people per day. Four preventable deaths occurred during this video. If we classify the preventable deaths against other causes of deaths, it will be the number three causes of deaths in the United States. After only heart disease and cancer. So, there are massive problems presented in modern health care. Including high costs, high waste, and low quality. How can big data help? The hope is big data can lead to better care and lower cost.

5. And big data for healthcare. People talk about the four v's. Volume. There's a massive amount of data, which gives analytics algorithms or systems a lot to act on. Variety. There's a variety of data that lets us connect lots of information sources together. Velocity, often times, this data, are coming in, in real time. Meaning that, data is coming in live, and needs to be processed, and analyzed, live. Veracity(正确度), there's a common problem with veracity. There's a lot of noise, a lot of missing data, a lot of errors, and a lot of [SOUND] false alarms.

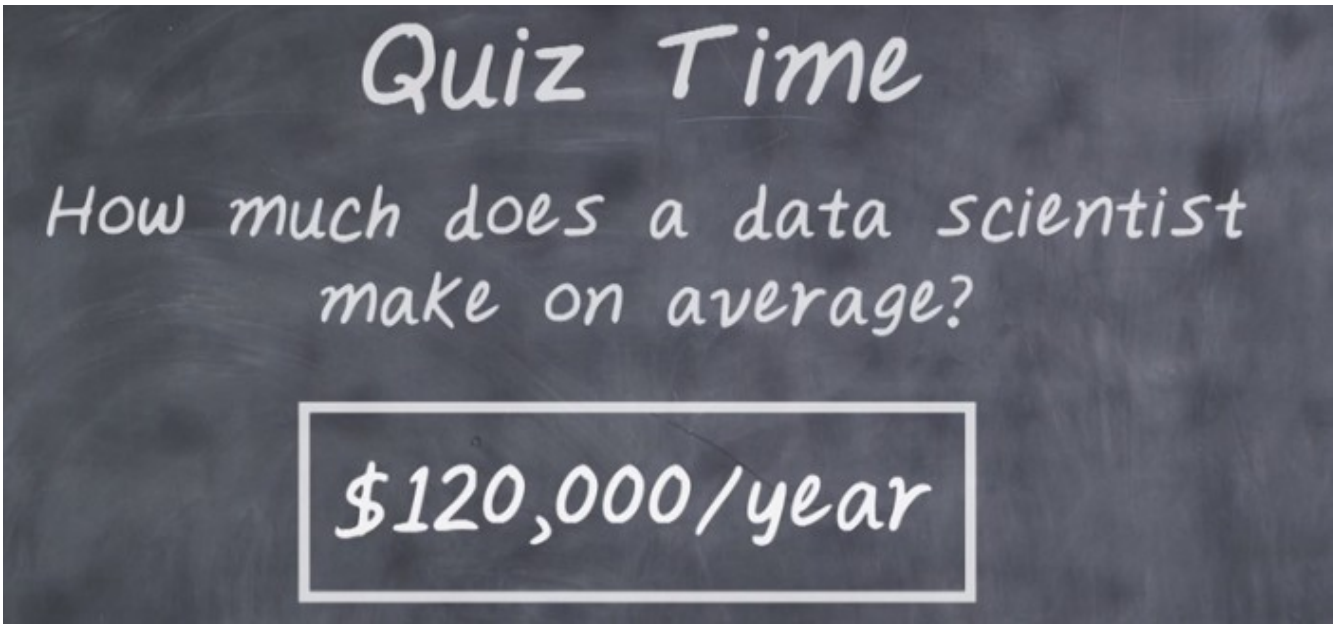
6. Healthcare generates a large volume of date. For example, for genomic data each human genome requires 200 gigabytes of raw data or 125 megabytes, if we store just snipes. For medical imaging data, a single fMRI is about 300 gigabytes. Medical imaging data generated in the US, per year was estimated to be 100 petabyte. That's a lot of data. Healthcare also generates a lot of different kind of informations. Such as clinical informations, including patient's demographics, diagnosis, procedure, medication, lab results, and the clinical notes. And patient generated data such as information coming out of arm body sensors and other devices that patients wear. And real time data sources such as blood pressure measures, temperature, heart rate, drug dispensing levels at intensive care units. Our main focus is on dealing with this wide variety of data. So we'll talk a lot about this later in this course.

7. So that's why you should care about the health care side. What about the data science side? There's an article in Hartford Business Reviews named Data Scientist, The Sexiest Job in the 21st Century. Here are some main points in that article. Capitalizing on big data depends on hiring scarce data scientists. Scientists. What data scientists do is to make discoveries while swimming in big data. I all ready told you I enjoy swimming by the way. [SOUND] Data scientists realize that they face technical limitations, but they won't allow that to bog down their search for novel solutions. Data scientists actually created the systems that we will discuss in this course, like Hadoop and Spark.



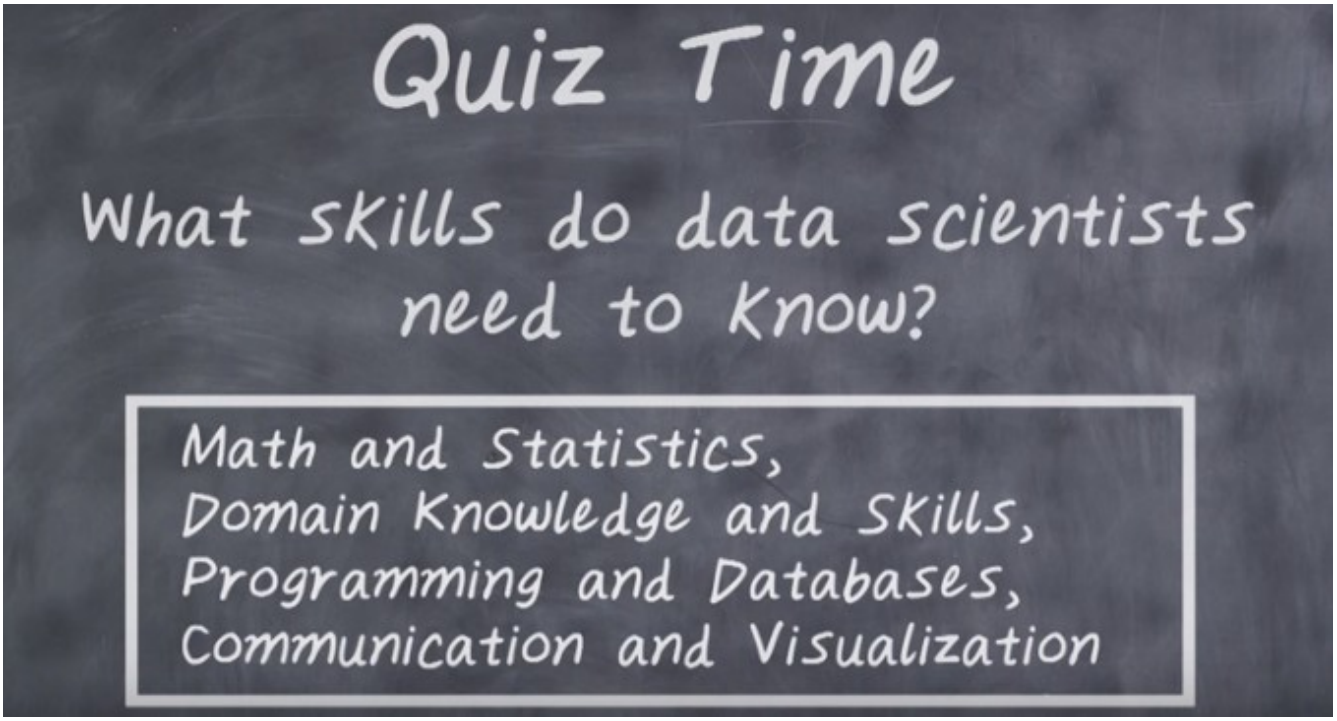
8. Now it's time for a quiz. First, true or false? Getting a graduate level degree is necessary for becoming a data scientist.

9. The answer is true.



10. Next, guess how much does a data scientist make, on average.

11. The answer is $120,000 per year. An experienced data scientist actually make over $150,000 year.



12. And finally, what skills do data scientists need to know?

13. The answer is: math and statistics, domain knowledge and skills, programming and databases, and communication and visualization.