MIT EXperimental Learning

MIT BDA Module 1 Unit 1 Video 3 Transcript

Speaker key

DS: David Shrier

HY: Hapyak

DS: We're here at the MIT Media Lab in Kent Larson's group, City Science, where we've built an urban modelling tool, an entire section of Cambridge, Massachusetts out of Legos. We've taken this, the simple children's toy, and turned it into a powerful data visualization and data modelling tool.

So, why would we do that? Other than playing with the over six million Legos that are in the building, we find that people find data confusing, mysterious, distancing. Many of the people that you are going to be interacting with in your career – for example, senior decision makers, politicians, urban planners – may have varying degrees of readiness to work with data. So in the case of the CityScope we've taken a very complex set of information about people, transportation, energy flows, economic data, and we've put it into the context of a children's toy, one that you can actually interact with. So, our CityScopes allow non-experts to literally stand around the table and see the effects of different urban interventions.

00:01:24

So, we're able to take an incredible mass of ones and zeros and convert it into meaningful information to support powerful decisions. You might be asking the question, what is big data? A classic definition of big data is sufficient volume, variety and velocity.

So, volume, right. Big data is by its very nature big. We are talking terabytes of data, not bytes or kilobytes or megabytes of data. Even gigabytes of data would be arguably what we'd simply call data or even small data.

Variety. So, variety refers to a number of different dimensions of data, right. This model sitting next to me, for example, one square kilometer around the MIT campus has a dozen different dimensions of categories of data, and within that a number of different pieces of information.

Velocity. Big data frequently is used to describe data that's changing rapidly. It has a very short data decay rate. In other words, you've got information that's coming at you very quickly. So, to recap the characteristics of big data, you're looking at volume, variety and velocity.

00:02:46

So, what good is big data? I mean, other than people investing millions into IT budgets, what does it really get you? Big data, when it's converted into information, which is what you will learn how to do over the course of this class, can be used for things like improving personal health of an entire population of people, developing novel financial trading strategies that can move markets, fixing transportation grids in complex urban environments, addressing issues of public safety. The number of applications for big data are practically infinite. One way to think about it is, what problem are you solving for? A number of societal problems can be addressed with conversion of big data into big information.

There are many different sources of big data when you get specifically into the realm of social analytics and social physics.

HY: The number of different data sources you have refers to which of the three V's?

a. Volume
b. Variety
c. Velocity
d. None of the above

DS: When we're looking at big data about people, we tend to want to look at a selection of sources. So one of those is online social media like Baidu or Twitter or Facebook. We might want to look at the smart phone, a very rich source of big data when properly aggregated.

We might even want to look at credit card information. So, as you'll hear from Professor Pentland and his colleagues, we've been able to take a bank's own credit card information to develop a better financial model, one that's more predictive about your future credit behavior. This can be used for financial access. We could give people who have never had a credit card before the opportunity to get credit through analyzing geospatial data, big data.

00:04:42

Another source of big data might be an app that you create. Part of why we're providing you the Funf library is because you can embed those tools into an app which itself can then become a source of big data. You'll want to make sure that you get permission from your users to acquire this data, and we'll talk about data ethics and data privacy, but the fact is people are willing to give this information if they get utility out of it.

When you think about big data, think about a cycle, right. This is not a linear process, this is not a waterfall. Rather, it's a cycle that feeds on itself. At the top of the cycle is collection. You're acquiring information perhaps from one of these sources that we've described. Then the data moves to pre-processing. So it comes in as a mass of ones and zeros. We need to begin to normalize it into data models that we can use.

From pre-processing we go to hygiene. Hygiene is a way of extracting some of the noise out of your data, cleaning it up so that you have a more clear signal. It's important when you're engaged in data hygiene that you don't throw out your original raw source data because you might be confusing signal and noise. You don't want to throw out the information that you're actually trying to extract out.

00:06:02

From hygiene we go to analysis. Now you're doing a first order look at what do I have, what am I working with, can I begin to extract some patterns out of this data? This leads us to visualization. So, Kent Larson's CityScope tool is one of several data visualization tools that can help you understand what data am I working with, what's it telling me, what's the information that I can extract from this data?

Visualization then leads us to interpretation. I've decided on what the signal is that's coming out of my data. What does it mean? If the population in a certain area decreases by 5% year-over-year does that mean people are moving away? Does that mean there are people who are perhaps older who are dying off? Does that mean that on a net basis young couples are not having enough children so they're not replacing the attrition that comes from old age? Does it mean that the government is relocating people because they're going to be putting in a dam and the entire valley will be flooded? There are a lot of reasons why a population in an area might change 5% from year to year. So, applying interpretation is where higher-order data analysis occurs.

Finally, we get to intervention. Once you've acquired your data, pre-processed it, cleaned it, analyzed it, visualized it, developed interpretations of it, you then can begin to say 'OK, what are we going to do with this data'?

HY: Select the answer which has the steps of the big data processing cycle in the correct order.

a. Collection, pre-processing, hygiene, analysis, visualization, interpretation, intervention.
b. Pre-processing, hygiene, collection, visualization, analysis, interpretation, intervention.
c. Visualization, collection, pre-processing, hygiene, analysis, interpretation, intervention.
d. None of the above.

DS: For example, we looked at the data for a large city in Africa. We looked at all of the data for approximately 3 million people, and we determined that the way the bus routes were laid out – so one source of data, the linear path of the bus – did not match up with where people lived and where they worked. So, we were able to take that big data about where people were and where they worked and combine it with the small data of the bus route information, and make recommendations about hey, let's change the bus routes. This made everyone's commute time 10% faster, because it decreased traffic in the wrong areas. It decreased congestion. It increased the speed from which someone got from their home to their office or other place of work.