MIT BDA Module 1 Unit 1 Video 4 Transcript

Speaker key

DS: David Shrier

HY: HapYak

DY: Let's talk about the five Rs of data quality – relevancy, recency, range, robustness and reliability.

Relevancy.

HY: Relevancy

DS: Is the data relevant to the problem that I'm working with? We've found in our work that this device, the smart phone, or, frankly, even what's known as a feature phone or the dumb phone, is an incredibly powerful source of information about people, about where they are, about what they like, about where they're going. And they'll let you access that information if you give them something of value.

00:00:47

You may be familiar with the app, Waze

HY: Waze

DS: which was acquired by Google. Waze takes data from millions of people all over the world and converts it into information about where's there a traffic jam, what's the best route to go to work, is there a police trap up ahead that I should be aware of? By giving people something important and relevant for their daily use, people are in turn willing to share their most valuable and most sensitive information. In fact, when researchers have studied what information do people value the most, what do we consider the most private, where I am right now is the most valuable piece of information that I have. But I'm willing to give it to Waze because Waze is giving me something back, something useful that might let me save five minutes on my commute.

Recency.

HY: Recency

DS: How recently was the data generated? People often make the mistake when working with data, particularly big data, of thinking that, well, I've acquired the data; now I'm good to go. Point in fact, data is constantly changing. It is a river, not a rock, and so knowing the recency of your data is incredibly important for understanding its utility.

Range.

HY: Range

DS: How narrow or wide is the scope of my data? You may have a set of very, very fine detail data on 200 people, or on a building, or on an entire country. Understanding your data range will be important for developing hypotheses about what the data is telling you.

00:02:27

Robustness.

HY: Robustness

DS: How robust is my data? Another way to think about this is, what's the signal to noise ratio? All too often, we'll have a giant mass of data and we're trying to extract out some critical meaning from it. Tools such as those embedded within the Bandicoot library can help you pull the signal out of the noise. You'll learn more about this toolkit in the coming modules, but if your data set isn't robust to begin with, it doesn't matter how much analysis you do. You're not going to be able to get a true signal.

Reliability.

HY: Reliability

DS: How accurate is the data that you're working with? Is it from a satellite or GPS? Is it from a census? Is it from a survey? In our experience, we've found a great deal of difficulty in working with survey data because it's so often inaccurate. People tell you a combination of what they think, what they think you want them to think, and what they think their friends will think of them when answering a survey. So we prefer to look for other sources of data that tend to provide a more reliable signal.

HY: Ensuring that the data you collect applies to the research question you are asking refers to which of the five Rs?

Range

Reliability

Relevancy

Recency

Robustness

Which of the five Rs is required to understand how well your data correlates to current conditions?

Relevancy

Recency

Range

Robustness

Reliability

When developing your hypotheses, it is important to understand your data range.

True

False

If a data analyst cannot find a pattern in a big data sample due to excessive noise that obscures the signal, it would be best described as having poor ____.

Relevancy

Recency

Range

Robustness

Reliability

In MIT's experience, survey data has been found to be a highly reliable signal.

True

False

DS: One of the skills that you're going to want to develop in your career as a data analyst is how to either find or identify relevant data that can solve the particular problem that you're addressing, or how to create it. Often we find in our work there's some data that's provided by our partner or collaborator, but then we have to go out and create new data sets to address a particular question. So, understanding the scope of the problem that you're trying to solve and the relevancy of the data you have available will tell you what your data gaps are and how you can build a more relevant data set.

00:04:18

Another important concept for you to understand is the idea of ephemeral data and durable data. So to get into this idea, let's think about what's known as the data decay rate. How long does it take for 50% of your data to become irrelevant or useless or out of date? One example is someone's home address. On average in the United States people move once every seven years. Let's think about that for a second. What that means is about 15% to 18% of your data goes out of date every

single year. You have to constantly refresh this data. The decay rate, or half-life, of home address data in the United States is about three, three and a half years. This is a critical number for you to understand when you're looking at a data set. How rapidly does that information go out of date?

Ephemeral data is very short-lived data. This might be information like the minute-to-minute price of a stock that's actively traded on the stock exchange, or my location if I'm driving a fast car down a highway. Durable data is data that changes very slowly – the location of this building. It's been around since 1984, and it's probably going to be around for another 30 years at least. So understanding if you have ephemeral data, which is very rapidly changing and rapidly decaying, or durable data, which is slow to change, is one of the dimensions of data that is helpful in understanding data quality.

00:05:57

HY: Differentiate between ephemeral and durable data, and provide your own example of each.

Ephemeral data is short-lived and has a high data decay rate, while durable data is long-lived. An example of ephemeral data could be the height of floodwater, whereas an example of durable data wouldn't be able to change as quickly, such as the location of a capital city.

DS: The last idea that I want to share with you when thinking about data quality is data refresh. So, if I've got, let's say, information about my car travelling down the highway, how do I refresh that information? What are sources of data that can constantly update fast decay data? So, when you're working with human factors data, with social analytics, it's always important to keep in mind what utility am I providing the end user who is in turn providing me this data.