EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 8 UNIT 1
Video 6 Transcript

MIT BDA Module 8 Unit 1 Video 6 Transcript

Speaker key

AS: Arek Stopczynski

HY: Hapyak

AS: This is the final module. I hope this has been an incredible journey. Now let's think about studying populations. In social analytics I think there is very often a danger of focusing on studying the data rather than the actual population that is of interest. Our lives, human lives, are extremely complex and it's very rare or never that a single channel of data can actually capture the full complexity of our lives, where we are, how do we move around, who do we meet, what are we interested in, and so on and so forth.

In many cases big data looks at very few bits, but for millions and millions of people. Here we are asking a slightly different question. What happens if we have deep data? If we take a smaller population in the order of thousands, but we really try to cut across multiple multiple channels, almost all of them, if we could we would absolutely get everything. Not quite possible, but we are trying to cut through communication, physical behavior, physical interaction, interests, psychological traits and a lot more. What happens is we are starting in that case to focus or to have tools to actually focus on the population rather than to be limited by whatever data we have because this becomes the limiting factor. What data we have is what questions we can ask. If we have multiple multiple channels we can study the same population based on those multiple channels. We can see whether certain questions can be answered using certain subsets of those channels. We can even see how consistent are our answers based on different channels.

00:02:01

If we're saying that someone is an extrovert and interacts with many many people, is it true on all these channels or is it just that this person is really liking to talk over the phone, but it's not actually reflected in how they behave in the physical world or how they interact on social networks? So, to create such research where we can actually look at multiple channels and study population we are building something that we call Living Labs, and Living Labs is the concept of taking a population that live their lives in a natural way and then instrumenting people themselves, and the environment for us to be actually able to capture this data from multiple channels. And one of such studies is the Copenhagen Network Study.

HY: Studying a population across multiple channels allows you to see how consistent your answers are.

True

Correct, well done. Having access to different channels allows you to check if your answers hold true across multiple channels and, if so, they can be considered consistent.

False

Incorrect. Having access to different channels allows you to check if your answers hold true across multiple channels and, if so, they can be considered consistent.

AS: You might have heard me referencing the Copenhagen Network Study in my other videos. So now I would like to give you a slightly deeper overview of what it is and why did we actually build it. So, the Copenhagen Network Study is a big research project that we did in Denmark with the PI Sune Lehmann and the idea was let's really try to capture as much as possible about a single population.

So what we did, we wanted to look at students and their lives, but very importantly we wanted to look at an important period of every student's life, which is their freshman year. So we actually started following our population even right before they joined the university.

So, those people would come from their own cities or towns up to Copenhagen to join the Technical University of Denmark and they would start interacting with each other. They would start building friendships, social ties. They would start getting grades. They would be in a totally new environment for many of them and we wanted to capture that to really see how those things unfold and how the ties are created the population.

00:04:07

So, how can we capture the data? Well, we have seen in the other videos that personal sensors is a way that one can capture it. So yes, we actually bought 1,000 mobile phones and we handed them out to our students. They were all running Funf library with an application on top of that. Of course our students knew what we were doing. We're very explicit in communicating what we are collecting, how they can opt out, but also where they can see their own data and not just raw data because giving the user just the raw data may create a false transparency.

HY: Why do you think providing the participants with their own raw data could create a sense of false transparency?

Thank you for your reflection. Please continue watching to hear why this is the case.

AS: I can see my data. This is what you have about me. That's great, but what does it really mean? So we actually provided access to the features that we're extracting from the data such as this is how we see you moving, this is how we see you forming friendships, this is how we see your social interactions network and so on and so forth.

So being very transparent, very open towards students, what are we doing? Why are we doing that? We were able to record over two years of extremely high-resolution behavioral data about his freshman population of around 1,000 individuals. Of course, as we have seen in the other videos, what we mean by high resolution behavioral data is location, social interactions, telecommunication, but in addition to behavioral data we also collected survey data such as psychological traits, answers to questions related to health, related to sleep patterns and other things that we wanted to ask the user to see how they are reflected in the behavioral data.

00:05:54

But that's not all. As I mentioned in the video about interdisciplinary research, there are many many more different disciplines involved in this study. So, we actually had an anthropologist on the ground and this lady actually started university with those freshman students. She has already been doing here PhD and this was her work. She actually started and partying with those students and going to the technical courses when she was like, I don't really understand what's going on, but I'm powering through. I participate in the groups. And this was extremely important for us because this allowed us to have this insight about, is what we are seeing in the data actually a reflection of the behavior of this population? If we are seeing that Wednesday evenings are amazing in terms of social interactions, why is it? Is it something that is schedule-driven? Is it a recurring party that we're seeing? Is it people just spontaneously coming together? What does it actually mean?

And those bits of insight that were gathered on the ground allowed us to understand not just the data quality, not just the biases that we might have in the data, because for example we asked, do people focus, remember and focus that they are being observed? Because this is a huge thing in the study that you are running. You know we handed out phones and we want, within the spirit of Living Labs, for this population to act and behave naturally, but you can think that oh now I have a software that is recording 24/7 of my behavior in this extreme resolution. Will it change people's behavior? Will they actually remember and focus on that?

00:07:47

And, something that was very encouraging from our anthropologist, we got, no, this is absolutely not the case. People use their phones as their primary phones - that was part of the deal – and they do not focus on the fact that there's anything being recorded about them. Again, we had to balance that with the privacy of the participants because we want them to, at some level to remember or to remind them from time to time, but during the most of the observation we want it exactly to be as natural as possible with the recording of the data fading into the background, and this creating a true Living Lab.

HY: When participants in a study are aware of the fact that they are continually being observed, this may create what is known as the observer's paradox. Considering the nature of the Copenhagen Network Study, what adverse effects do you think this could have had on the data collected and results generated?

a. **Students could have altered or limited their behaviors and interactions.**
b. Students could have been more willing to participate in the study.
c. Students could have remained completely unaffected by the nature of the study.

This awareness could have led to students altering (or even limiting) their behaviors and interactions, or choosing not to use the provided phones as their primary phones, during the course of the study, thus resulting in an inaccurate reflection of the behavior of that population.

AS: Now the question, what can we actually learn from that? Well, we're still learning. I would even say that we are just scratching the surface of what is in those millions of man-hours of the data that we have recorded. But, some of the insights that we are getting already; I mentioned in the other

video about how can we study temporal dynamics of social networks? This machinery, this mathematical machinery has actually been developed because we are faced exactly with the problem of recording very high resolution data, social data, social interactions, that we couldn't really crack. We couldn't find the proper structures. It was too rich, too dynamic. So we had to develop entirely new machinery to analyze that, but we are also learning a lot about team performance. You might have heard me mentioning this Strength of the Strongest Ties research and this was also part of this Living Lab. We were looking at how people work in teams and what influences team performance.

00:09:27

We are learning about epidemics. You might have heard me talking about can we target small populations and actually vaccinate or monitor people in a very targeted way to stop the outbreaks, and this is, again, part of the learning that we are getting out of the Copenhagen Network Study. So, huge Living Labs, Living Lab, where we are studying multiple channels for a well-defined population as this population live their lives.