EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 3 UNIT 1
Video 3 Transcript

MIT BDA Module 3 Unit 1 Video 3 Transcript

Speaker key

AS: Arek Stopczynski

HY: Hapyak

AS: So, now, let's talk about testing on yourself. When you start working with high resolution personal data, which you'll quickly discover is, it is complex; multiple channels, multiple scales. For example, the question, where am I right now? Simple question but the answer is, I'm in the US, I'm in Cambridge, I'm at MIT Media Lab, I'm in this amazing room right here. At what scale are we answering this question? And this understanding of what are we actually trying to answer with the data, it's a very complex problem.

00:00:55

And it's probably one of the biggest problems I ever encountered in social analytics, which is we are trying to arrive and explain concepts that we, as humans, understand. So, for example, we want to answer a question about mobility. How mobile people are or how social they are, or how physically active they are, whether they have big social circles or small social circles. All those things we have, as people, intuitive understanding of but we need to make sure that the data we are capturing and the way we are analyzing that actually reflects that.

For example, if we are studying social interactions and our phones sense that two users are close, in physical proximity to each other, what does it really mean? Does it mean close is hundreds of meters? Are they just in the same building? Are they within ten meters or are they actually facing to each other? Does this data imply that there's interaction between the users or does it just imply that people are close to each other but maybe totally facing in opposite directions and never speaking to each other?

And the interpretation of that will actually change the things you can do with the data and the things you should do with the data. When attacking the data quality problem it's usually a good idea to start by testing on yourself. Before you do any kind of deployment, before you start collecting data of your users, just run it on yourself. Collect the data about your own day, about your own behavior and see how this corresponds to your own understanding of your day.

00:02:39

Maybe it would be helpful for you to keep a journal and actually write down where have you been, when have you left the room, if that's the resolution you are after? Or maybe you want to check it with your calendar at work if this is the data that you are interested in but think about is it reflecting, is the data I'm collecting reflecting my life as I would think about it as a researcher? And be very, very particular about any problems that you spot. Don't just forget about them or think they are small. If

the data is not a perfect reflection of your life and as you, as a researcher, would like to see it, think about why not? Think about what are the problems? What are the biases that you seeing there?

00:03:21

And once you do it for yourself, do it for a small group of people for whom you can gather some ground truth data. And this ground-truth data may be because you have access to their calendars or maybe you will interview them or maybe you can actually ask them to keep a little diary of their activities because this will give you the comfort of being able to verify that what you are seeing actually makes sense the way you want to think about the data but also, very quickly, you will spot that people actually have behaviors that can influence the data they produce in those very, very funny or unexpected ways.

00:03:59

For example, I always have location on, on my phone. I just never bother to turn it off. The moment I deployed it on a bunch of users, suddenly they were like, oh, the application never wakes up when I enter the university and this is because they would turn off their location when they didn't explicitly need it. And this type of the problem is only something you can spot after you actually deploy it on someone else than yourself.

So, start with yourself and then expand it a little bit to a set of users that you can trust, that you can actually gather some extra ground truth from but also that they will provide this larger cut of human behaviors that may influence and change how you look at the data.

HY: For optimal results in data processing it is recommended that you first test on _ and later test on _. The objective with this testing exercise is to compare _ with _ and thereby solve problems in your data.

a. **Yourself; a small group; ground truth; collected data.**
b. Ground truth; collected data; a small group; yourself.
c. A small group; yourself; ground truth; collected data.