EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 2 UNIT 1
Video 3 Transcript

MIT BDA Module 2 Unit 1 Video 3 Transcript

Speaker key

AS: Arek Stopczynski

HY: HapYak

00:00:00

AS: Hello everyone, I hope you are enjoying the course so far, and right now I would like to talk to you about personal sensors, and really by personal sensors I mean mostly smartphones. It is estimated that, right now, over three billion people have smartphones, and by 2021, over six billion people will have smartphones. And each one of these is not just communication device, it's a powerful sensing device that travels with us as we go about our lives and can sense so much about our behavior. And this offer us an opportunity to tap into that, but of course we need to remember about the privacy of the users and be very respectful towards the user. So it's an extreme opportunity for us to study human behavior with unprecedented high resolution. So, let's now dive in and actually look what type of data can we sense using these phones.

Certain types of data that we can sense with the phones include location. Those phones have GPS, Wi-Fi, Bluetooth, so they can actually pinpoint our place in the world with an extreme resolution down to probably room level, at this point. We can also see who are we communicating with, using text metadata, call metadata, we can understand our physical activity, whether I'm standing or sitting right now, how many steps I'm taking every day. We can look at our physical interactions, who are we meeting, using, for example, Bluetooth scanning. And we can understand things about our interests, because, again, those phones, they don't only sense, we actually use them. So it's the question of the content we are looking at, the videos, the music, the news that we are reading, and things such as screen going on and off. We actually started looking at students at lectures, how their screen turns on and off, because we can look at it as a proxy of how focused students are at lectures.

00:02:07

HY: Select the option which has the terms matching their correct descriptions.

a. **Location (where we are in the world); physical activity (how we move); physical interactions (who we meet face-to-face); telecommunication (who we contact).**
b. Location (where we are in the world); physical activity (how we move); physical interactions (who we meet face-to-face); telecommunication (what applications and media we consume).
c. Location (how we move); interests (what applications and media we consume); physical interactions (who we meet face-to-face); telecommunication (who we contact).
d. All of the above.

AS: There is this richness of the data that we can get of the phones. But the data can be seen as crude oil. It actually needs to be refined so it can be used for something. There is a pipeline that you want

to build when you are working with data, especially high-resolution behavioral data. And some of these common steps that we are looking at include cleaning of the data. So remove all the outliers or the readings that simply do not make sense. For example, when I travel on the planes, very often it would place me in a random city around the world, simply because the Wi-Fi on the plane has been recorded in that particular spot. There is the step of binning. So phones can be producing data in an erratic way, and we want to normalize it both in time and in space. After binning we want to integrate the data from multiple sensors. For example, location can be sensed using Wi-Fi, GPS, Bluetooth. We want to arrive to a representation that is common and independent of the actual sensor from which the data was gathered. And then finally, we want to actually translate the data into higher-level features, as we'll see in the other video where we dive deeper into the details of this process. But data, just your lat-long, understood as location, is not yet useful.

00:03:31

We want to extract something bigger, such as what is your home location or how active you are. So let's now try to look slightly deeper into two types of the data and behaviors that we can sense, and those include mobility and social interactions.

Human mobility, so understanding how humans move, at multiple scales. It might be at the scale of room between room or town to town or even country to country. There's something common about how do we want to approach understanding human mobility as we study the behavior in high resolution. So the first step, as we gather the data from the phone is, of course, cleaning. So we want to reject those outliers. And from time to time, the GPS on the phone will say, I have no idea where this person is, and will return (0,0). And those are easy outliers to actually reject. But sometimes, what we will see is that, based on Wi-Fi access points, some of those Wi-Fi access points might be moving, like people's personal hotspots, and this will produce mobility of my own phone that is actually untrue.

So there is this step, sometimes it might be pretty complex, sometimes it might be simple, but of actually cleaning the data to arrive at a clean representation of where have I been. After we arrive at that representation, we usually want to normalize the data in the sense of sampling. Especially with the phones, it's very common that they might be producing data in irregular intervals.

00:05:07

For example, Android phones right now, they produce location data every couple of minutes when they actually update their internal state to understand where you are and then so they can present you what weather is around you or what interesting places are around you. But if you turn on navigation, suddenly a data point about your location is produced every few seconds. And this might create problems for certain algorithms, so we want to clean the data very often and actually split it into regular intervals. So for example, at most, one reading per minute. And after we have that, we want to integrate data from multiple places, such as Wi-Fi, Bluetooth, GPS, so we arrive at the representation that is effectively time, latitude and longitude.

And after we have that, we have this raw mobility trace, but it's not yet the most interesting part. Because this is where the actual fun starts and we can start learning what do we know about this person from this behavior. So, for example, we have the trace and people moves around. We want to discover places. As I'm moving and then I stay somewhere, my phone starts producing the readings in

the same place, and we want to discover that this is a place, that this is where someone stopped. After we have places, not just transitions, we can try to merge them across longer periods of time to produce consistent IDs.

So, for example, every night I'm going home and we want the label of this place to be consistent across weeks and months. And we are using different algorithms to do that, one of which is, for example, density-based clustering, like DBSCAN. That basically looks, exactly as the name suggests, at the density of the points that we are observing. And after we have that, after we have places and then consistent places across time, we can start asking questions such as what are these places.

00:07:00

HY: Select the correct order in which these steps should be executed.

a. Binning; cleaning; transform into higher-level features; integrating across channels.
b. Cleaning; binning; transform into higher-level features; integrating across channels.
c. Cleaning; binning; integrating across channels; transform into higher-level features.
d. Integrating across channels; cleaning; binning; transform into higher-level features.

AS: And some of them are easy. For example, identifying your home location is pretty easy, this is where you are spending most of your nights, or your work location, where you are spending most of your days. But sometimes it's very tricky, such as distinguishing between gym and coffee shop. If they are right next to each other, we may never know whether you are actually going there to exercise, whether you're just staying there to grab some coffee.

HY: Why is it more difficult to determine what people do when they are exploring than when they are engaging in routine behavior such as being at work or at home?

Routine places like work and home are easy to spot as they are always the same and you spend a great deal of time there. They are stable and repeated patterns which are easy to discover. Places you explore such as particular shops or a gym are more challenging to identify as there are less observations that can be used for this identification.

AS: So that's human mobility.

And now let's think about social interactions, the other extremely important aspect of human life. Of course, social interactions, they happen in multiple ways. We call people, we text people, we talk to them on Facebook. But one of the really interesting ones that only is possible to study when we have high resolution behavioral data is physical interactions. So normally, or most commonly, we would discover those physical interactions, just physical proximity or even actually facing someone, using sensors on the phone such as Bluetooth. So phones can scan for Bluetooth, and if someone is within 30 feet, which is around 10 meters, we would know that there was the other device in this proximity.

And this gives us the idea of who is meeting with whom in physical space. And this data can be very noisy, so we want to clean it a little bit. For example, thresholding on the received signal strength indicator. So we don't pick up those very, very weak interactions because just someone happened to be passing by in the next room.

00:08:37

An interesting cleaning here is we are now building a view of the population, not just individual view of what you have seen, but how did the population look like. And to do that, we need to remember to synchronize the clock on the phones. Because even though most of us, or majority of us, have their clock on the phone set to automatically adjust, the actual differences in the population might be up to several minutes. And this will produce an inconsistent view of the population. So we need to remember to clean that.

After we are done with the cleaning, we want to do temporal binning again. Normally we would use the temporal resolution of around five minutes, and this is supported by social literature that says that meaningful interactions, they happen around few minutes. It may be two minutes, it may be ten minutes, but if you're just passing someone, if there's actual interaction you're having, it won't really be much shorter than five minutes. So we are using around five minutes for this type of the data.

Now, we also want to integrate the data across entire population. And here is a very nice trick that we can do because we do not expect false positives from Bluetooth. Which means if my phone is sensing someone, it means that someone, understood as a device, was actually there. So after we produced the matrix of interaction, we can force it to be symmetric. If I have seen someone, someone also has seen me. And such symmetric network is much easier to analyze, but also we are gaining the information because we can account for phone deciding not to scan for whatever reason or some data being lost or some imperfection in the actual scanning process.

00:10:27

HY: When analyzing an event (such as a physical interaction), why is it important to observe it independently from two or more perspectives?

We can use the information gained from these perspectives to fill out possible gaps in the data, as well as to improve data quality.

AS: So after we have that, what we are producing is a highly-temporal network of social interactions in the population, understood as triples, time, Person A and Person B. And we know that whenever we have a data point there, that Person A has been within certain physical proximity or even facing someone in that, time being T. Some of the questions that we can ask to this network are now around what communities are there, with whom people work, with whom do they play, and think how can we actually distinguish these two.

00:11:05

We can ask questions about team effectiveness, what makes people more productive. We can ask about how central people are and whether this actually helps in their lives. And also we can ask questions that are related to health, for example, how epidemics spread in the society and how we can stop them. And we'll dive deeper into that in the upcoming videos and I'm really excited to take a deeper look into that.