EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

# Big Data and Social Analytics certificate course

**MODULE 7 UNIT 1**
**Video 1 Transcript**

# MIT BDA Module 7 Unit 1 Video 1 Transcript

**Speaker key**

AS: Arek Stopczynski

HY: Hapyak

AS: Hello everyone, we're almost at the end of this course, I hope you're doing great and having lots of fun. Right now I would like to talk about applications of big data in healthcare. And of course the topic is huge because data, data, data, it drives healthcare and it's expanding even more from understanding individual behavior, to studying small populations, to studying entire countries, to studying global state of, for example, epidemics. Here I want to focus a little bit of what is really the bleeding edge and what is possible or will be possible in few months, maybe few years, by really the ability to capture high-resolution, high-quality data about populations. In one of the first modules, we talked about studying human mobility and that it's a very important part of human behavior. It actually has a huge application in healthcare, in understanding the state and the dynamics of diseases.

HY: Amy Wesolowski et al. on using call detail records

AS: One of the great articles has been published recently by Amy Wesolowski et al. and it talks about using CDR's, Core Detail Records that you should have encountered by now in the course, to study mobility of people in Kenya, but to understand how malaria spreads. So looking at 15 million of people and how they move around and correlating that with sources and things of malaria parasite, Amy and her colleagues tried to understand where does the disease originate and where it is travelling to, giving us the hope to be able to contain and to actually understand the dynamics of the disease at the level of the entire country. Studying human mobility gives us understanding of disease dynamics, not only at the level of a country but even at the global level.

HY: Dirk Brockmann and Dirk Helbing paper

00:02:11

AS: Dirk **Brockmann** and Dirk **Helbing** studied the mobility patterns using flight connections, basically looking how many people are travelling between the airports to understand the dynamics of real world diseases and how they spread around the globe. And they found a really surprizing thing in that by modifying our notion of distance from pure geographic distance, how many kilometres or miles are between cities, to how many passengers actually travel or what fraction of passengers from one airport travels to the other airport and taking that as a distance measure they were able to predict and to show the spread of the diseases in a beautiful wave like form where there is the origin at the center and all the cities that are at a given distance understood as this effective distance, as they call it, the disease arrives exactly at the same time.

So this can be helpful not only to make prognosis about when will disease that originated in certain place arrive to a different place, but also do the opposite, to understand or to identify the point of origin of the disease.

00:03:31

So if we have seen the disease in this city at this time, and this city at this time, and this city at this time, what was the origin? Which is the city that has the right distance to all those places? Distance understood as effective distance. So we can actually understand where did this thing start.

HY: Dirk Brockmann and Dirk Helbing's research on human mobility and the spread of disease introduced and highlighted the importance of the notion of effective distance. In the context of disease spread, why would it be useful to know the city of origin?

Using the wave of effective distance to discover the city of origin, in the case of the spread of disease, would allow for more effective and concentrated vaccination, monitoring, and containment of the disease, beginning with the identified origin city. This could then allow governments to limit the extent that the disease propagates globally.

AS: Where there is a global scale of countries or the entire globe, there is also the microscale, the scale of small populations such as kindergartens, or schools, or workplaces, or small neighborhoods. And this scale is also amazingly important to understand if we want to think about the healthcare, specifically about epidemics spreading and even stopping that. Schools and workplaces, those are the places where many many people get infected and this is where the disease can really propagate very quickly and reach different parts of the population. We've all experienced that, if one kid in school is sick then it's very likely that almost everyone in school will get sick, same with the workplaces. That's why it's always so important to make sure that if you're sick you don't go to work, you actually stay home because the moment you enter work it's very easy for the disease to propagate.

Now, those complex populations as we have been discussing in some of the previous modules, we actually have tools right now to study them with amazing resolution using personal sensors, but also using social networks or telecommunication networks. So one of the questions that we asked recently was, if we have a data set about a single population that we know their physical interactions of the participants but we also know the social network and telecommunication network, can we use the data from easy-to-obtain networks such as Facebook or CDR's to actually find who should we vaccinate or who should we monitor in this population so we can understand and contain the disease?

00:05:40

So the idea is very simple, physical interactions data is expensive to capture, it's extremely rich but you need phones or you need badges, you need different sensors and it's very costly to do it at scale and for a period of time. But networks such as Facebook or telecommunication networks, they are relatively easily available, the data is at least there, the question is about the access to that. But this is problem that can be solved. We asked, by having those three different types of the networks, can we simulate the disease spreading on the physical interactions network and see whether we can actually predict who should we target with vaccination or monitoring using the other two networks.

And what we found out is, yes, people that are central in the telecommunication or social networks are also the central people in the physical network.

And this is very important because this means that simply monitoring these people will allow us to see that there is an epidemic outbreak in the population much earlier than if we just randomly looked at people in the population. And also if the disease requires a very close contact, think droplet or touch, we can actually stop it very effectively by vaccinating in a targeted way the key people in this population. Key being the most central in the social and telecommunication networks.

HY: The potential and impact of big data applications in healthcare cannot be overstated. However, not all big data applications are successful. Can you think of a failed application of big data in healthcare?

Thank you for your reflection. Continue watching to hear about Google Flu Trends - a well-known example of a failed application of big data in healthcare.

00:07:11

AS: Using big data, whether it's in healthcare or any other domain is not risk free. There are pitfalls and there are things that can go wrong and the things, even when your data is huge, may simply not work or even stop working.

There is a case of Google Flu Trends, this service is no longer active. But when it was active, Google went ahead to build flu nowcasting, basically a system that would be able to say what are the current levels of flu based on the search terms. The idea was extremely simple but also powerful. Based on the history data of search terms, Google went to find the terms that would closely correlate with CDC reports of flu levels and basically use those terms to see what are the levels of flu in the future. And while this worked at the beginning, at some point the model basically started over estimating the flu levels by huge margins.

HY: David Lazer et al. on what went wrong with Google Flu Trends

AS: There's a wonderful paper by David Lazer et al. that talks about exactly that, what went wrong, what could have gone wrong with Google Flu Trends, and why suddenly this big data approach, that assumes the more data that we can get the better we will get, stopped working at some point.

00:08:46

And one of the possible problems here is, I think that this is very important when we are talking about using big data in different applications, Google is a search engine. It's not a research project, so changes to the way results are displayed will also change the data that gets generated and then can be used, for example, for nowcasting flu. So whenever you work with the system there's always this risk that the changes that are perfectly valid changes driven by the product requirements, we need to show different search results, for example, in terms of Google Flu Trends. Or we want to show actual cards that show the disease after you look for the symptoms. They will change how the data gets generated and what kind of data gets generated, and if the research model that you're building does not account for those changes in the product, you are risking that basically at some point your research part will stop working because now the data gets generated in a very different manner.

And this is a very good case study to think, especially if you are thinking about applying big data analysis in an industry, be it healthcare, be it HR, be it anything else, if those systems are not designed specifically to generate this data, but the data is an outcome of some other product-related activities, you always are risking that you will invalidate your research by making changes into a product.