EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 3 UNIT 1
Video 4 Transcript

MIT BDA Module 3 Unit 1 Video 4 Transcript

Speaker key

AS: Arek Stopczynski

AS: Social analytics is amazingly exciting and it's exciting because worlds collide. You have social sciences colliding with the data-driven approaches and computer science and physics and engineering together to answer the most interesting, the most important questions you can be answering about societies. This collaboration between all those different domains that becomes really, really powerful in both academia and in industry doesn't come for free. It actually takes time, effort and skill to make projects like that work.

00:00:52

In the Copenhagen Network Study where we handed out 1,000 phones to our students to collect high resolution behavioral data for slightly over two years, this project has been collaboration between departments of Economics, Public Health, Anthropology, Philosophy, Computer Science, Network Science, Physics. Making this actually work has been a huge effort, totally worth it but it was a very explicit challenge that had to be addressed, both in the planning part of the project and the deployment but also as an ongoing challenge when the collaboration was actually happening.

So, now, let's go slightly deeper about how can we actually make things such as that work? One was to think about interdisciplinary research is to think about data process as pipelines. Each step in the pipeline takes the data and produces features of higher order and you chain those steps together to arrive at different answers to questions. And each step in the pipeline has an owner, someone who is responsible for that but very importantly, researchers from different domains they actually jump in and plug into the pipeline and are responsible for different blocks at different parts or different points of this pipeline.

00:02:14

Let's take an example. Imagine that public health researchers want to study whether we can spot people who are getting depressed by seeing that they are starting to stay home more often. For them, the question of all the raw data processing that leads to this answer is mostly irrelevant in the sense they are not really interested in how, technically, it has been done. They are focused and interested in their very particular question.

As we have seen in previous modules, to arrive at the home location this is actually an entire pipeline of data analysis. We collect raw GPS data. In this raw GPS data we find locations, we map the locations to a common idea across multiple weeks or months and out of this we actually find what is a home location or most likely home location? What is work location? And once we have that we can compute how much time every day this user has been staying home. And this is the feature that public health researchers that we are mentioning are now ready to take and plug into their model and their analysis

and their research and their understanding. And everything that leads to that point is probably owned by a researcher from different discipline.

Things, questions such as how do we translate raw GPS data into staying at home, is of interest to computer science researchers, network researchers even physicists or machine-learning people and they are the ones responsible for that and they are the ones excited about this part of the research. And the results that they are producing now become applicable to the research of other researchers from other domains.

00:04:02

And you can think about exactly making interdisciplinary research work as having this pipeline of data with researchers from different domains plugging into different parts of this pipeline. While thinking about interdisciplinary research as pipelines can be very effective and can actually make it possible for such research to happen, it also comes at cost. And probably the biggest cost here is the possibility that researchers in different compartments will lose the global view of the data and will not be aware of what the actual data that has been produced for them actually means.

For example, we want to measure how much time people are spending together? So we collect Bluetooth data and we actually process it and we arrive at user A, has been spending that many hours with user B. And if the researcher from public health domain or anthropology or any other domain now takes that and runs with it, the question remains, does this data actually capture people sitting at the desks next to each other and does it change what we should be thinking about in terms of the data? Can we actually spot those problems and how do we build this knowledge into the pipeline?

And this takes an explicit effort. Once you start making compartments you're immediately risking that researchers will start to misunderstand what the data from the other parts of the pipeline actually means. So the explicit effort to make sure that everyone and everything is in sync is documentation and having someone who is the system integrator or a programme manager who actually manages that.

00:05:56

So the documentation is relatively straightforward since we have pipelines and steps in the pipelines, we have owners of those steps. They should be the ones producing the actual documentation that is not only about how to produce the data, how to access the data but also how to interpret the data. Is my time spent together, capturing people sitting next to each other or are we removing that explicitly?

By the way, think how we could remove the problem of people sitting at the desks next to each other, to remove this data from time spent together. And there's also this role, this very important role of the system integrator, someone who's very interdisciplinary, who might be not deep in any particular part of the pipeline but who can understand across the pipeline and can actually make sure that everyone is in sync and understands different parts of what's going on.