Name: Zhiao Zhou Net id: zz1749

Nowadays there is always a debate about whether theories or models are still useful for science since it’s now the age of big-data, with which we could easily obtain a conclusion without making a hypothesis or speculating but let the statistic algorithm do everything. Like Chris Anderson wrote in his famous article “The End of Theory: The Data Deluge Makes The Scientific Method Obsolete” in 2008, “With enough data, the numbers speak for themselves.” (Anderson, C., 2017, June 03)[1] And I believe this could also be found in urban informatics areas. However, from my perspective of view, big-data could answer some of the urban issues which is the reason why it is so charming, but not all of them where there are still some occasions in which we would need a model or theory or there would be several risks. The more we learn, the harder we find a model to explain.

First of all, before bringing the topic to why data analysis could be theory-free, another fact we should face is that theory isn’t easy to be applied even if we want to. Nowadays the data we are talking about isn’t the same thing as 10 years ago or even 5 years ago. We used to store our data sets on disks but nowadays they are in the cloud in that there are so many of them being created each day. Information is not easily a matter of multi-dimensional taxonomy but of dimensionally agnostic statistics.[1] So it’s much harder now to apply a model or theory to the data because it’s changing every day and it’s nearly impossible to classify them in order. Accordingly, we would have to focus on the data mathematically first.

Data scientists could easily use massive amounts of data to reach a consequence without knowing any underlying theory or models. All they have to know is the correlation, using statistic algorithms. For example, Google has the best searching and advertisement system in the world which were founded only by mathematics without knowing any underlying searching or advertisement culture and conventions.[1] If the statistic algorithm indicated that the website should be on the top, then that’s it, with which Google could translate all of the languages in this universe as long as there would be adequate data sets.

Google even proved big-data could be much better than traditional ways. Five years ago, Google announced a remarkable achievement in Nature, one of the top scientific journals over the globe. They managed to track the spread of influenza across the USA even faster than the Centers for Disease Control and Prevention by finding a correlation between what people searched for online and whether they had flu symptoms, while none of the team members knew about medical science. (Harford, T., 2014, March 28)[2] In Urban Informatics areas, we could use massive demographic or real-life data to learn what kind of risk factors lead to different diseases such as diabetes then we could control people’s lifestyle to reduce the huge burden on the health system.

So this is what the way is: correlation is enough and we don’t have to know anything about the causation. However, this is also the reason why theory and model are still important in the field of data analysis. Sometimes correlation can’t be enough.

First, going ahead without knowing certain kind of underlying mechanism provides a possibility that the correlation might crash. In the example of Google’s flu tracking system, after providing an accurate account of flu outbreaks for four years, Google’s system showed that there was a severe outbreak while the CDC’s data indicated that Google’s estimates of the spread of flu-like illnesses were overstated by nearly a factor of two. How could this happen since the system had been working well for such a long time? Chances are that back in that time people were so worried about the disease so a lot of healthy people tended to search it online which hadn’t been eliminated by Google.[2] In addition, before the Worldwide financial crisis in 2008, nobody would ever predict it because every quantitative financial model was working perfectly. Quantitative finance is another field where models are utilized without any underlying mechanisms so it is common that a model turns out to fail especially in some capital markets under control of governments. Consequently, without knowing the underlying mechanism under the correlation means that whether, when or how the correlation is going to break down is also unknown.

Second, there could be selection biases. The data sample might not be randomly drawn then the model is biased towards people who are in the sample. For example, Netflix movie recommendations model is biased to people who have the time and interest in rating a bunch of movies online. Of course, some might think that we could fix this by cleaning the data. However, at most of the time, there are huge bunches of data that it is impossible to take care them. And sometimes because they are too large, some data scientists seem to be convinced that the sampling problem isn’t worthwhile to worry about while it actually is.

Third, sometimes there might not be enough data at all. “Since good statistics on urban concentration do not exist even today for substantial parts of the world, and hardly exist for any part during most of the time since cities have been in existence, we are forced to rely on whatever credible evidence can be found and so can reach only broad conclusions concerning early periods and only approximations for recent times.” (Davis, K., 1955) [3] For example, in many developing countries which are witnessing a huge process of urbanization like my hometown Shanghai, there isn’t so many urban data collected like New York has so it would be hard for an urban planner in Shanghai to use data to motivate urban research. They would have to use models or theories from other big cities in the world for references.

What more, statisticians are now trying to develop new statistical models as well because data is growing so fast that the old one might not be viable for new data. And they have to use the old statistical theory as the basis so in other words, data science cannot abandon the utilization of the old theories or models no matter they are mathematical or practical.

In a word, urban data scientists now could attach importance more to the data sets than the theories or models but there are still situations where a certain type of theory is needed including statistical or industrial theories. It’s better if they would be developed together. Conversely, big-data could also make an impact on theories that have guided many types of research so far.

Most importantly, theories were based on hypothesis and speculation first in a lab, then proved so probably it would lack a little bit accuracy since the hypothesis is tested by probability. However, when it comes to the age of big-data, the city itself could be the real-time laboratory for urban scientists where they could make tests even every second of time. Accordingly, this could gradually calibrate and correct the old theories or even prove that they are wrong under some circumstances. Because usually, a theory is hypothesis-based but chances are that the hypothesis could be wrong. Newton's Laws could be a perfect example here since for hundreds of years, people are convinced that his laws could account for any physical phenomena.

We all know that all physical laws were developed under rigorous mathematical deduction after certain hypothesis and this model was actually invented by Newton. However, in his laws, he made a wrong hypothesis that the all the space and time in our universe are flat which could actually be warped due to high velocity or gravitation. And this was found thanks to the big-data analysis in physics. (Newton's laws of motion., 2017, August 28)[4]

This is also applied in urban informatics field. Big-data analysis might change the theories which have been guiding us.

References

1. Anderson, C. (2017, June 03). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Retrieved September 09, 2017, from <https://www.wired.com/2008/06/pb-theory/>
2. Harford, T. (2014, March 28). Big data: are we making a big mistake? Retrieved September 11, 2017, from <https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0>
3. Davis, K. (1955). The origin and growth of urbanization in the world. *American Journal of Sociology*, 60(5), 429-437.
4. Newton's laws of motion. (2017, August 28). Retrieved September 11, 2017, from https://en.wikipedia.org/wiki/Newton%27s\_laws\_of\_motion