Promet in vreme

We are currently at a stage where we cannot imagine having a question and to not have an answer for it with a simple click. For whatever pops in our head there is already an article, image, video, news… Google has all the answers. However, this is not a one-way street, every click, every link, every site that we open is stored, every location that we visit is remembered. The world is currently driven by data. Data has power, and all the corporate giants already know that, and they know how to use it. So often in a hurry to find out what we want, we click the "I agree" button without even a glance at what it is asking us, so there is an enormous possibility if that window pops on the small screen in our car to click "I agree". McKenzie & Company already started the plan how to use that.

So, what is the plan? From every new signed car deal car companies can offer special applications, for example applications for tolls on the road, time saving tips such as information on free parking spots, information on markets and restaurants. For all these applications the customer pays to use them or gets them for free when they are ad-supported, but at the same time the customer gives his or her information to the car company. Additionally, all the car companies which would like to be part of this monetization would have to produce very well-equipped cars to get as much information as possible from the customer. The cars would have to have sensors for the current condition of the vehicle, special information transporters for the latest road changes, biological sensors, so the car can recognize who is driving it and in what condition he or she is, for example: blood pressure, heart rate, sobriety and so on. Considering that, through many applications on our smartphones we are already giving that information, there is a great probability that many people will be ready to give that information for the baits that the car stores will offer. However, there is one catch. For their plan for car data monetization to succeed they would be sharing the driver's information with third parties, like insurance companies, auto repair companies and so on. This is an ethically very questionable and may also be an issue regarding General Data Protection Regulation (GDPR). If data is obtained despite these obstacles, we are talking about enormous data, which will not only be hard to examine, but if that data is false, in my opinion can have extreme consequences, and in case of self-driving cars it can be even life threatening. As additional burden on this project is that this data will be very desired for many insurance companies which they can easily abuse, and extremely unsafe in cases of cyber-attacks. Obtaining data would be hard but managing it can be even harder for big data cases, and we can see how true this is with one example of big data.

In Netherlands on the Dutch highway there were 20.000 road sensors counting passing vehicles each minute and measuring their speeds. Their aim was to analyze the data for 4 years from 2010 to 2014 and in that time, they collected 80TB data. However, not all this data is usable. Although there are 20.000 sensors on this road the data between them is extremely hard to

compare even when they are only 250 meters apart, very often because of change in driving speed of the passing vehicles. So, the next logical simpler step would be to use the data from one sensor, but 98% of the sensors are missing at least 1 minute data. At the end even with 20.000 sensors they must estimate the number of vehicles in the existing gaps. Therefore, big data is not equal to quality data. Analyzing such enormous databases is extremely hard, even if we want to check a small chunk of this database is an enormous task, therefore that must be automatic process. In this process most of the data will be seen as noise and very little proportion will be seen as useful information. Perhaps, going a step further with less sensors but more powerful ones which would not only count and measure speed but will also check registration numbers on the passing cars, would be a more productive and exact way of completing these statistics. That way the missing data from one sensor can be easily found in the next sensor just by comparing the registration numbers. A second solution would be to set more sensors on a same location and if one of them losses connection or malfunctions there would be still parallel copies from the data measured by the other sensors. However, it is hard to predict all the possible measurement mistakes that can happen before starting.

Collecting data for 4 years will certainly have gaps and it will be hard to analyze. Therefore, we can only imagine what can happen if we want to use data collected over a period of 50 years. Technology changes, people who control it change, laws which have influence on the analyzed field change, requirements for documentation change. This is the case with Slovenian meteorological data followed from 1961 to 2011. In this period all the data passed from being stored for many years on simple paper to digitalized forms, from being measured only 3 times a day to being measured every 10 minutes, from being dependent from one person only to automatic measuring stations. So much can happen in one field in 50 years in Slovenia even the number of measuring stations had changed many times. In my opinion this is a lot of change for one data base, there is no constant time frame between measurements, there is no constant location, instrument or person and there is not enough dense network of measuring stations to fill in the gap for the period of missing data in one location. This database is only useful if we are interested in seeing the big picture as simple check of temperature increasing trough a period of 50 years in which the smaller time gaps in one station are not noticeable. Such analysis that includes 50 years of data collection is extremely hard work, the hardest part being to harmonize the data which was formed in so many ways.

With the constant development of technology and the increasing number of different sensors is easier to obtain data, but once we have an abundance of it, is extremely hard to manage it. Big data can be very useful if we are very careful how we are acquiring that data. Sometimes if the process of obtaining is not prudently planned, we can even get a 100TB data that will be extremely hard to analyze.

“aha” moments:

1. Big data can be very useful but extremely hard to analyze.
2. Very often if we have big data it is hard to see data and analyzes as two separate parts.
3. Very careful planning is crucial before such enormous project, because there are much more time and funds invested.