1.1 Transcript: What is data science about and what is not about?

Data Science is about using some tools that make it possible to work creatively with data and to extract the required possible insight from data. It is not about number and/or equations (data) and its arithmetic's or basic science and experiments in the sophisticated laboratories (although they create vast amount of data).

In order to obtain the required possible insight from data one may view all available data, because collection of all may present an answer or representative of say a meaningful pattern while each individual may contain certain meaning. This process of putting all puzzle pieces together to obtain the required insight, demands creativity especially in using appropriate tools and technologies in data science.

Creativity in mining and excavating insights from raw data distinguishes data science from other disciplines in the information age so that it become one of the most important fields with high demands. Data science mines unstructured raw and hazy data, then finds order, meaning, and value in them. Those are important information for example to provide competitive edges and advantages for businesses, growth in security information for government and discover many vital information such as crime rate in the society however these are not easy to achieve without creativity in using tools which is a limiting factor. That is why data science is in high demands, for example according to a white paper published by the McKinsey Global Institute [see reference 1.] projected a demand between 140,000 to 190,000 for deep analytical talent positions and 1.5 million for more data-savvy managers in only United States in the next few years. As another example Harvard Business Review, stated that Data Scientist is the sexiest Job of the 21st Century [see reference 2] . The high demand has driven incomes of data scientists up with excellent salary: based on Glassdoor [see reference 3] data science offers a median base salary of over $116,000 while data from O'Reilly.com indicate a total remuneration of about $144,000 a year. These have made data science a compelling career opportunity.

With this in mind, how the role and expertise of a data scientist should be define so it matches best with current and future climate of the information age, how you are going to define data science, what tools are required, how is possible to work creatively with tools and technology to extract the required possible insight from data, what sort of expertise, skills and background are required? Is data science about pure numerical data, its math and statistics and does it require fundamentals sciences practiced in the research laboratories? Can I be a data scientist without becoming a mathematician or statistician and many more questions?

Based on above, you may define data sciences as the ability to use tools for coding in order to apply required analysis, which may include some math or statistics, to a practical problem in your field. While this is reasonable, you may also consider the fact that data science is the analysis of all diverse data. Data that a mathematician or statistician might throw them away as there is no mathematical methodology to analysis them. However, you as a data scientist will try to use all available data and the information that you have in order to obtain a deeper insight, make a sense of them and find an attractive solution for your clients. As you can see inclusive analysis, i.e., the analysis of all diverse data, is deeply embedded in data science and will be discussed through the course of this module.

There are two main distinct categories for the tools and technologies useful in data science: In one end of spectrum there are Open (source) tools, which are not sold by vendors, but they are combination of freely available open sources, programming languages and their intermediary products. For example

- Python programming language and libraries,
- R programming and statistical analysis tools

- Julia which are most popular for data analytics.

In the other end of the spectrum, there are proprietary tools and technologies that are auto-managed tools such as closed products, i.e., tools that you can buy, and you can start using it right out of the box. Excel could be a good and popular example. Other most popular proprietary tools for data analysis in this category are: • Qlike

- Tableau
- Looker
- Zoho Analytics

Many other vendors developed databases such as SPSS, offering various simple built in tools for customized analysis related to the data in their database. This would offer similar advantageous and disadvantageous to auto-managed tools and sometimes considered in the same category as auto-managed tools, which will be discussed in a later lecture this week.

The tools and technologies that are required in data science, may highlight the skills and expertise that are necessary for a data scientist, however it doesn't provide an answer on how to work creatively with data and associated tools in order to extract the required possible insight from data. One answer that you might think of is by visualising data so you as an expert in your field can see and recognise hidden patterns and shapes in your data, i.e., to consider data visualization like visual art in inspiring creativity in your work. This is very fascinating as this may lead you to a solution or motivating methodology by prompting you to brush up your analysis, machine learning or algorithms that you are developing based on your data visualization. Please note that there might not be an answer in all the situations that you may encounter, however data visualization would help you to use your tools and technology more effectively to get the best out of your data to highlight the bottlenecks and possible ways to remove them. These are important reasons for why data visualisation is so vital in data science and why you need data visualisation tools such as matplotlib to provide you these opportunities and potentials that are quite essential in success of your data analytic project.

We continue these discussions in other lectures however this introduction will get you up and rolling in practicing some of data science tools next.