What does a Data Scientist do? By Shesh

Introduction

Created in the 1960's, data science is a field under the umbrella groups of science and information technology that delves into fields of statistics, programming, and machine learning. Combining these fields and techniques uncovers patterns and relationships within the data to better inform decision-making (Igual & Seguí, n.d.), solve real-world problems and construct evidence-based decisions. This essay will not only explore the decorated skills of data scientists, it will be a case-study investigating examples where data scientists excelled in their respective sectors of employment and expertise, from government, industry and research. Understanding how these cases contributed to their organisation's growth will help other data scientists in their journey towards success. The first case study investigates how data science aided the Chinese government's responses to the Covid-19 pandemic. Moreover, the next case study elucidates how data scientists have aided Johnson & Johnson (J&J) in the pharmaceutical industry to improve sales, efficiency of drug development and understand the importance of adapting to the market to meet customer needs. Lastly, within environmental research, data science techniques are used by the Environmental Defence Fund (EDF) to provide insight within the natural world for conservation and sustainability.

.

Government

During the Covid-19 pandemic, China was impacted extraordinarily. The sector within the Chinese government known as The Chinese Centre for Disease Control and Prevention (CDC) utilised their data science team to collect and interpret data to provide insights for outbreak predictions, root causes of Covid-19 and help guide policy decisions (Wu et al., 2020). The data scientists within the sector created an 'end-to-end data science system' (Luo et al., n.d, p.121) known as DEEPEYE (Luo et al., n.d.), which followed the data science methodology of data collection, data integration and data cleaning. The data collection was done when the team downloaded data from other countries' disease control and prevention centres. They connected to other data sources like population, temperature data and statistics and then finally provided with trajectory data. The data integration process looked at the team uniting the various types of data into 'predefined relational tables' (Luo et al., n.d, p.125). Finally, data cleaning went through the process of cleaning data errors like duplicates and missing values. This provides a concrete approach to data analysis to ensure efficiency, accuracy and consistency, additionally it provides opportunities for improvement in the data analysis process (Foroughi & Luksch, 2018). Moreover, using machine-learning and data visualisation, the team provided an analysis of the pandemic from sources like the public health agencies, news outlets and social media to track the spread of the virus for creating new public health policies. DEEPEYE was created for the purpose of simplifying complex data regarding Covid-19 cases and its effect on the public

into a structured approach, so the government can make informed decisions to combat the pandemic. Due to China's larger population, the data scientists highlighted the importance of data analysis at a large scale which in turn brought new eyes to use similar models for other purposes or sectors not just health care like the financial industry regarding credit scoring and retail industry to personalised marketing and improved customer experience (BernardMarr, 2016). The data scientists from the DEEPEYE team were highly qualified to work in the government sector as they possessed technical skills for a data scientist like computer programming and algorithms, analytical skills to investigate complex data sets, apply statistical methods to create a predictive model and communicative skills to socialise in a productive team (Ismali, N.A., & Zainal Abidin, n.d). Data scientists in these roles met strict requirements to work for the Chinese government. They needed a relevant healthcare or data science degree, outbreak investigation or research experience, proficiency in both Mandarin and English, and knowledge of the local healthcare system and cultural norms (Chinese Center for Disease Control and Prevention, n.d.).

Industry

In this world of digital and medical innovation, data science is an essential technique used by the pharmaceutical industry, as the field of pharmacy depends on data science to increase product development and success (The Role of Data Science in the Pharmaceutical Industry, n.d.). J&J is America's biggest pharmaceutical company that manufactures healthcare and pharmaceutical products. To provide for 175 counties where their products are sold (Fortune, 2022), data science is required to analyse the large amounts of data they compile to uncover patterns to simplify their operation while eradicating errors (Karpatne et al., 2017). In their drug-development plans, the company utilises computational modelling and machine-learning algorithms to analyse the large chemical and biological data to identify potential drug targets. For example, during the process of Covid-19 vaccine development, computational modelling helped create predictive models to provide an insight to the potential effects of the vaccine to further improve the vaccine's design (Role of Computer Modeling in Vaccine Development, 2022). Moreover, the machine-learning algorithms analyse data from clinical trials during drug-development to evaluate the safety and efficiency of the products. This is seen in the trials of the Covid-19 vaccine where during manufacturing, the machine-learning algorithms in the data generated predicted errors to help identify areas of improvement and ensure the safety of candidates (The COVID-19 Data Plan: 3 Innovative Ways Johnson & Johnson Is Using Data Science to Fight the Pandemic, 2021). The data science team utilises forecasting and predictive analysis to not only predict the trends during the distribution process, but provide an insight into customer behaviour and preferences from data sources like social media and customer surveys. This highlights how data scientists are of great value and importance for J&J as they develop effective treatments, optimise manufacturing processes and ensure efficient distribution of

pharmaceutical products (The COVID-19 Data Plan: 3 Innovative Ways Johnson & Johnson Is Using Data Science to Fight the Pandemic, 2021). Thus, aligning them with the values of the J&J company policies where their vision is about “impacting lives with better products and solutions" (Johnson, 2022, p.1). The team of data scientists have skills that use statistical analysis, machine-learning and programming with languages like Python and SQL. Moreover, they have experience within the healthcare system to help run the government health care system. Finally, without the communicative and collaborative skills, data scientists within the J&J company will work with other sectors like engineers, pharmacists and business leaders to develop data-cantered solutions.

Research

With a vast volume of data and the need to integrate from numerous sources, methods like data visualising and machine-learning to analyse and interpret the data from the research in environmental science for conservation and sustainability (Science, n.d). The EDF is an organisation working to protect the environment and human health. The EDF uses techniques like machine-learning, data integration and statistical modelling, to analyse environmental data. For instance, the EDF green lighted a project -MethaneSat in 2022- a satellite mission used to identify and quantify methane emissions from human resources throughout the world (Benmergui et al., 2018). By using machine-learning and atmospheric modelling the EDF identified and quantified methane emissions, to combat climate change. Also, the data scientists used machine-learning algorithms to recognise correlations within the methane emission from oil and gas operations, livestock and landfills that will be difficult to identify for humans (Sheng et al., 2020). Moreover, atmospheric modelling is employed by the data science team to track the movement of methane emissions over time, space and around the earth’s atmosphere (Propp et al., 2017). This ultimately aids EDF to create policies to reduce greenhouse emissions. As climate change becomes a serious threat over the years, the data scientists are of value for the organisation by providing the tools and techniques to collect and analyse the large and complex datasets, which further helps the rest of the world adopt more sustainable practices. To work in this role the data science team required strong data science analytical skills and knowledge in programming tools like Python and SQL to clean, manipulate and analyse large complex data sets. Also, machine-learning and statistical modelling is a vital skill to identify patterns and analyse trends within the data to identify patterns in the environmental data and communicate it in a clear way. As well as having proficient knowledge in environmental science regarding ecosystems, sustainability and climate change. Lastly, communicative skills to convey ideas and work with other members in the team to come to a census and inform complex ideas to policy makers, stakeholder and the general public.

Conclusion

Data science is a revolutionary field that combines coding languages and machine-learning with statistical knowledge to create insights. Through these elements, data scientist’s role in society are critical for the success of organisations and industries. For example, within the Chinese

government data scientists from the CDC created a data system known as DEEPEYE to combat Covid-19 cases in China. Moreover, the data scientist team in J&J in the pharmaceutical industry used modelling techniques during drug development of its vaccine. Finally, research by the EDF used data science in their MethanSAT project to detect changes within the human sourced methane emissions, to fight climate change.

References

Benmergui, J. S., Wofsy, S. C., Gautam, R., & Hamburg, S. (2018). MethaneSAT: A learning satellite for detecting and quantifying methane sources. NASA ADS, 2018, A43R3442. *https://ui.adsabs.harvard.edu/abs/2018AGUFM.A43R3442B/abstract*

BernardMarr. (2016, April 3). The Amazing Ways Big Data Is Used In China - DataScienceCentral.com. Data Science Central. *https://www.datasciencecentral.com/the-amazing-ways-big-data-is-used-in-china/*

Foroughi, F., & Luksch, P. (2018). Data Science Methodology for Cybersecurity Projects. ArXiv:1803.04219 [Cs]. *https://arxiv.org/abs/1803.04219*

Chinese Center for Disease Control and Prevention. (n.d.). En.chinacdc.cn. Retrieved April 20, 2023, from https://en.chinacdc.cn/

Fortune. (2022). Johnson & Johnson Company Profile. Fortune. *https://fortune.com/company/johnson-johnson/*

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Transactions on Knowledge and Data Engineering, 29(10), 2318–2331. *https://doi.org/10.1109/tkde.2017.2720168*

Igual, L., & Seguí, S. (n.d.). Undergraduate Topics in Computer Science Introduction to Data Science. Retrieved April 20, 2023, from *http://repository.psa.edu.my/xmlui/bitstream/handle/123456789/2016/2017_Book_IntroductionToDataScience.pdf?sequence=1&isAllowed=y*

Ismail, N. A., & Zainal Abidin, W. (n.d.). Data Scientist Skills. Advance Informatics School, University of Technology Malaysia, Kuala Lumpur.

Luo, Y., Tang, N., Li, G., Li, W., Zhao, T., & Yu, X. (n.d.). DEEPEYE: A Data Science System for Monitoring and Exploring COVID-19 Data. Retrieved April 20, 2023, from

https://web.archive.org/web/20220615213457id_/http:/sites.computer.org/debull/A20june/p121.pdf

Propp, A. M., Benmergui, J. S., Turner, A. J., & Wofsy, S. C. (2017). MethaneSat: Detecting Methane Emissions in the Barnett Shale Region. NASA ADS, 2017, A32D06. https://ui.adsabs.harvard.edu/abs/2017AGUFM.A32D..06P/abstract

Role of Computer Modeling in Vaccine Development. (2022, August 3). News-Medical.net. https://www.news-medical.net/health/Role-of-Computer-Modeling-in-Vaccine-Development.aspx

Science, J. D., SFSU Institute for Geographic Information. (n.d.). Introduction to Environmental Data Science. In bookdown.org. https://bookdown.org/igisc/EnvDataSci/

Sheng, H., Irvin, J., Munukutla, S., Zhang, S., Cross, C., Story, K., Rustowicz, R., Elsworth, C., Yang, Z., Omara, M., Gautam, R., Jackson, R. B., & Ng, A. Y. (2020). OGNet: Towards a Global Oil and Gas Infrastructure Database using Deep Learning on Remotely Sensed Imagery. ArXiv:2011.07227 [Cs]. https://arxiv.org/abs/2011.07227

The COVID-19 Data Plan: 3 Innovative Ways Johnson & Johnson Is Using Data Science to Fight the Pandemic. (2021, January 13). Content Lab U.S. https://www.jnj.com/innovation/how-johnson-johnson-uses-data-science-to-fight-covid-19-pandemic

The Role of Data Science in the Pharmaceutical Industry. (n.d.). Data Science Degree Programs Guide. Retrieved April 20, 2023, from https://www.datasciencedegreeprograms.net/industries/pharma/

Wu, J., Wang, J., Nicholas, S., Maitland, E., & Fan, Q. (2020). Application of Big Data Technology for COVID-19 Prevention and Control in China: Lessons and Recommendations. Journal of Medical Internet Research, 22(10), e21980. https://doi.org/10.2196/21980