

*Data Scientist: The Sexiest Job of the 21st Century*

- Summary: Data scientists make discoveries from big data and analyze them by communicating the concise data with storytelling visually and verbally. Companies scour for quality data scientists due to high volumes of disorganized information, but there is a shortage of them. Most data scientists have a mixed background, with skills including hacking, analyzing, communicating, and advising. They also have fundamentals in math, statistics, probability, and computer science. For example, the idea of using machine learning to connect recommendations on LinkedIn significantly boosted the company's reach and activity.
- First of all, I have never heard of the platform Hadoop, but it seems like a useful tool for data analysis. Seeing this article was written almost nine years ago, data science has grown into its own section in computer science. Machine learning relies on big data, making data science a hot topic. This article paints the perspective that companies are in desperate need of data scientists' qualities and fill a huge deficit in the company as well as demonstrates how to find top notch talent and how to keep them.

*What Data Scientists Really Do*

- Summary: In the tech industry, data scientists perform analytics on data and use experiments to grow. In other fields such as medicine, data scientists use data to comprehend cancer research. In general, data scientists collect data, then clean, visualize, and communicate data. Simply organizing data is very time consuming. Having critical thinking and quantitative skills are more important than techniques. Categories or specialization in data science are also beginning to form, with those in business, decision making, and machine learning. The lack of regulation among gathering data is stirring questions about ethics due to harmful effects such as biases.
- I was shocked to find that arranging data takes up more time than analyzing data. Maybe it should not be much of a surprise, as computer programmers spend most of their time debugging their code, not actually writing the code. Ethics surrounding data has become a great concern, as I have also gotten curious about products that use AI and could be recording our every move.

*Big Data: The next frontier for innovation, competition and productivity*

- Summary: Big data is a big quantity of datasets, and data scientists capture, store, manage, and analyze them. In order to improve machine learning algorithms, user's data is often saved without consent, which may feel like an interruption to their privacy, particularly health and financial records. Anyhow, it can enhance the productivity and quality of companies. Currently, companies face shortages of people with talents in analyzing and managing because there needs to be the right tools and incentives in order to drive competition and innovation. Meanwhile, the economy and the users should be

taken into consideration. By addressing public concerns, creating transparency and easy access, especially in the public sector, data science can reduce search time and improve quality. Data is affecting every industry in a major way, but not all current data is storable. Training data requires experimentation to adapt to needs and variabilities to improve performance. Data science should support human decisions with algorithms to improve decision making, minimize risks, and provide valuable insights. Data security is also crucial, as decisions as to how competitive and sensitive data should be kept private remains questionable. Data breaches expose personal confidential info. Currently, leaders do not have an understanding of value in big data, which harms the ability to use big data to make better decisions. Policymakers need to protect privacy and provide data security by mitigating talent through education, immigration, and communication.

- There are so many aspects in order to make data science useful and ethical. Not much was new, but balancing growth and protecting privacy will be a great challenge for both companies and the government.

### *Introduction to Machine Learning and Deep Learning*

- Summary: Machine learning is different from traditional coding, in that the computer learns from input and output data. AI is non-biological systems designed to perform human intelligence, such as natural language, speech, and image recognition. Machine learning designs AI, as computers learn from models, representations, and rules. Deep learning, a subcategory of machine learning, learns from neural networks. AI is then split into AGI (artificial general intelligence) and narrow AI. Under machine learning there is supervised, unsupervised, and reinforcement learning. In supervised learning, there is semi-supervised and self-supervised learning. Supervised learning uses regression analysis and classification models, or a combination. Unsupervised learning has PCA models, which defines the variance in the dataset and forms another component axis coordinate system. Autoencoders are deep neural networks using decoders and encoders. Clustering is another method of classification, with prototype, density, and hierarchical clustering. Semi-supervising uses labelled data to label unlabelled data. In short, supervised develops predictive models to predict information. Data is represented in two ways: tabular and unstructured data. Tabular data is organized, like in a chart. Unstructured data is raw information, which can be images.
- I had always thought AI, machine learning, and deep learning were the same thing. All the ways of analyzing data are new to me. There are linear and nonlinear data organizations. especially reinforcement learning. The example explained chess, which I am not familiar with. The steps to train models, or the predictive modelling pipeline makes sense with structured and unstructured data, with real world applications.

# Introduction to RStudio

