What is Data Science?

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

• Better decisions (should we choose A or B)

• Predictive analysis (what will happen next?)

• Pattern discoveries (find pattern, or maybe hidden information in the data)

Where is Data Science Needed?

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

Examples of where Data Science is needed:

• For route planning: To discover the best routes to ship • To foresee delays for flight/ship/train etc. (through predictive analysis)

• To create promotional offers

• To find the best suited time to deliver goods

• To forecast the next years revenue for a company

• To analyse health benefit of training

• To predict who will win elections

Data Science can be applied in nearly every part of a business where data is available. Examples are:

• Consumer goods

• Stock markets

• Industry

• Politics

• Logistic companies

• E-commerce

How Does a Data Scientist Work? A Data Scientist requires expertise in several backgrounds:

• Machine Learning

• Statistics

• Programming (Python or R)

• Mathematics

• Databases

A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.

Here is how a Data Scientist works:

1. **Ask the right questions** - To understand the business problem. 2. **Explore and collect data** - From database, web logs, customer feedback, etc.

3. **Extract the data** - Transform the data to a standardized format. 4. **Clean the data** - Remove erroneous values from the data. 5. **Find and replace missing values** - Check for missing values

and replace them with a suitable value (e.g. an average value). 6. **Normalize data** - Scale the values in a practical range (e.g. 140 cm is smaller than 1.8 m. However, the number 140 is larger than 1.8. - so scaling is important).

7. **Analyze data, find patterns and make future predictions**. 8. **Represent the result** - Present the result with useful insights in a way the "company" can understand.

What is Data?

Data is a collection of information.

One purpose of Data Science is to structure data, making it interpretable and easy to work with.

Data can be categorized into two groups:

• Structured data

• Unstructured data

Unstructured Data

Unstructured data is not organized. We must organize the data for analysis purposes.

 

Structured Data

Structured data is organized and easier to work with.



Evolution of Data Science: Growth & Innovation

• The term “data science” — and the practice itself — has evolved over the years. • In recent years, its popularity has grown considerably due to innovations in data collection, technology, and mass production of data worldwide.

• Gone are the days when those who worked with data had to rely on expensive programs and mainframes.

• The proliferation of programming languages like Python and procedures to collect, analyze, and interpret data paved the way for data science to become the popular field it is today.

• Data science began in statistics.

• Part of the evolution of data science was the inclusion of concepts such as machine learning, artificial intelligence, and the internet of things.

• With the flood of new information coming in and businesses seeking new ways to increase profit and make better decisions, data science started to expand to other fields, including medicine, engineering, and more.

**Origins, Predictions, Beginnings**

• We could say that data science was born from the idea of merging applied statistics with computer science.

• The resulting field of study would use the extraordinary power of modern computing. Scientists realized they could not only collect data and solve statistical problems but also use that data to solve real-world problems and make reliable fact-driven predictions.

**1962: American mathematician** John W. Tukey first articulated the data science dream. In his now-famous article "The Future of Data Analysis," he foresaw the inevitable emergence of a new field nearly two decades before the first personal computers. While Tukey was ahead of his time, he was not alone in his early appreciation of what would come to be known as "data science." Another early figure was Peter Naur, a Danish computer engineer whose book Concise Survey of Computer Methods offers one of the very first definitions of data science:

"The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

**1977:** The theories and predictions of "pre" data scientists like Tukey and Naur became more concrete with the establishment of The International Association for Statistical Computing (IASC), whose mission was "to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

**1980s and 1990s:** Data science began taking more significant strides with the emergence of the first Knowledge Discovery in Databases (KDD) workshop and the founding of the International Federation of Classification Societies (IFCS). These two societies were among the first to focus on educating and training professionals in the theory and methodology of data science (though that term had not yet been formally adopted).

It was at this point that data science started to garner more attention from leading professionals hoping to monetize big data and applied statistics.

**1994:** BusinessWeek published a story on the new phenomenon of "Database Marketing.” It described the process by which businesses were collecting and leveraging enormous amounts of data to learn more about their customers, competition, or advertising techniques. The only problem at the time was that these companies were flooded with more information than they could possibly manage. Massive amounts of data were sparking the first wave of interest in establishing specific roles for data management. It began to seem like businesses would need a new kind of worker to make the data work in their favor.

**1990s and early 2000s:** We can clearly see that data science has emerged as a recognized and specialized field. Several data science academic journals began to circulate, and data science proponents like Jeff Wu and William S. Cleveland continued to help develop and expound upon the necessity and potential of data science.

**2000s:** Technology made enormous leaps by providing nearly universal access to internet connectivity, communication, and (of course) data collection.

**2005:** Big data enters the scene. With tech giants such as Google and Facebook uncovering large amounts of data, new technologies capable of processing them became necessary. Hadoop rose to the challenge, and later on Spark and Cassandra made their debuts.

**2014:** Due to the increasing importance of data, and organizations’ interest in finding patterns and making better business decisions, demand for data scientists began to see dramatic growth in different parts of the world.

