What is Data Science

Contents

1 Defining Data Science and What Data Scientists Do

1.1 What is Data Science:

- A process to uncover the secrets behind data, Analysis and Exploration of Data.

1.2 Fundamentals of Data Science:

- Understand their environment.
- Analyse existing issues.
- Reveal previously hidden opportunities.
- Analysis of structured and unstructured data.

2 Data Science Topics

2.1 Foundation of Big Data:

- A process to uncover the secrets behind data, Analysis and Exploration of data.

2.2 Fundamentals of Data Science:

- Understand their environment.
- Analyse existing issues.
- Reveal previously hidden opportunities.
- Analysis of structured and unstructured data.

2.3 Foundations of Big Data?:

Big Data refers to the dynamic, large and disparate volumes of data being created by people, tool, and machines. It requires new, innovative, and scalable technology to collect, hist, and analytically process vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhance shareholder value

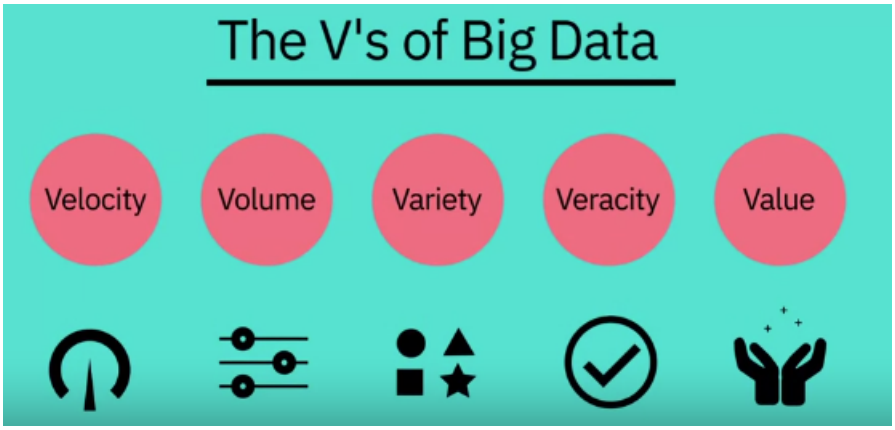

Figure 1: The V's of Big Data

- **Velocity:** Speed of Data Collection
- **Volume:** Amount of Data Collected
- **Variety:** Diversity / Type of Data
- **Veracity:** Quality and Origin of Data
- **Value:** Ability to turn Data into value

2.4 What is Hadoop?:

Big Data is divided into many smaller datasets and distributed over servers. Then the results will be processed through mapper and reduce processes.

2.5 Data Mining:

2.5.1 Goals:

- Costs
- Benefits
- Expected Accuracy
- Usefulness

2.5.2 Data Selection:

- Quality of Data
- Sources
- Types of Data
- Accuracy
- Reliability

2.5.3 Pre-Processing Data:

- Messy Data
- Irrelevant Data (Identify)
- Missing Information (Gather)(Determine Impact)
- Check Integrity

2.5.4 Transform Data:

- Format Data
- Data Reduction
- Principal Component Analysis (reduction of large dataset into small dataset).

2.5.5 Sorting Data:

- Unrestricted and immediate read
- Data Storage Scheme
- Store on servers
- Safety and Privacy

2.5.6 Mining Data:

- Data Analysis Model
- Parametric, and Non-Parametric methods
- Machine Learning Algorithms
- Data Visualisation
- Graping Capabilities

2.5.7 Evaluaing Results:

- Test Predictions
- In-Sample Forecast
- Stake Holders Feedbacks

3 Deep Learning and Machine Learning

3.1 What's the difference?

3.1.1 Big Data:

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

3.1.2 Data Mining:

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

3.1.3 Machine Learning:

Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence.

3.1.4 Deep Learning:

Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.

3.1.5 Neural Networks:

Neural Networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains.

3.2 Neural Networks and Deep Learning

Neural Network	Deep Learning
Nodes and Experimentation	Multiple Neural Networks
Mimics Brain	High Computation Power
Input - Output	
Process: Collection of Nodes	
Good for small problems	
Computation Intensive	
Part of Speech detection	Speech and Face recognition system

3.3 Applications of Machine Learning

- Cluster Analysis
- Predictive Analysis
- Decision Trees
- Recommendations
- Fraud Detection

3.4 Regression

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

4 Report Structure

According to the course material, a final deliverable in the form of a report, has the following 10 main components:

1. Cover page
2. Table of contents
3. Introductory section
4. Methodology section
5. Results section
6. Discussion section
7. Conclusion section
8. References
9. Acknowledgment
10. Appendix

5 Summary

- The typical work day for a Data Scientist varies depending on what type of project they are working on.
- Many algorithms are used to bring out insights from data.
- Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.
- The differences between some common Data Science terms, including Deep Learning and Machine Learning.
- Deep Learning is a type of Machine Learning that simulates human decision making using neural networks.
- Machine Learning has many applications, from recommender systems that provide relevant choices for customers on commercial websites, to detailed analysis of financial markets.
- How to use regression to analyze data.
- The length and content of the final report will vary depending on the needs of the project.
- The structure of the final report for a Data Science project should include a cover page, table of contents, executive summary, detailed contents, acknowledgements, references and appendices.
- The report should present a thorough analysis of the data and communicate the project findings