

Term Paper Report  
On  
**Insightful world of Data – Data science**  
Submitted To  
Amity University, Uttar Pradesh



In Partial Fulfilment of The Requirements for The Award of The Degree  
Of  
Bachelor of Technology  
In  
Computer Science and Engineering  
By  
**Gaurav Jain A2305220595**  
Under the Guidance of  
**Ms. Shweta Bhardwaj**  
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY  
AMITY UNIVERSITY UTTAR PRADESH

## DECLARATION

I **Gaurav Jain**, Student of B. Tech (3 CSE6-X) hereby declare that the project titled **“INSIGHTFUL WORLD OF DATA – DATA SCIENCE”** Which is submitted to the Department of Computer Science and Engineering, **Amity School of Engineering and Technology**, Amity University, Noida in Partial Fulfillment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

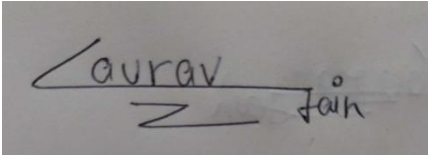
The Author attests that permission has been obtained for the use of any copy righted material appearing in the report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Date: July 6. 2021

Gaurav Jain

A2305220595

3CSE-6X (2020-2024)

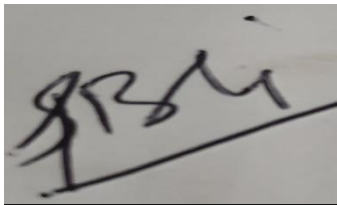
A photograph of a handwritten signature in black ink on a light-colored surface. The signature reads 'Gaurav Jain' with a stylized underline.

Signature of Student

## **CERTIFICATION**

This is to certify that Mr. Gaurav Jain student of B.Tech. in Computer Science Engineering has carried out the work presented in the project of the Term paper entitle "**INSIGHTFUL WORLD OF DATA – DATA SCIENCE**" as a part of First year program of Bachelor of Technology in Computer Science and Engineering from Amity University, Noida, Amity University Uttar Pradesh under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or diploma to this university or elsewhere.

A handwritten signature in black ink, appearing to read 'Shweta Bhardwaj', is written over a horizontal line.

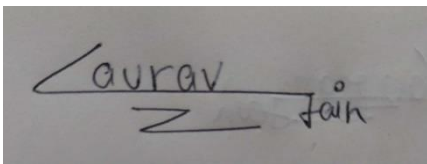
**Ms. Shweta Bhardwaj**

Department of Computer Science and Engineering

ASET. Noida

## ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I would like to thank Prof. (Dr) Sanjeev Thakur, Head of Department-CSE, and Amity University for giving me the opportunity to undertake this project. I would like to thank my faculty guide Ms. Shweta Bhardwaj who is the biggest driving force behind my successful completion of the project. She has been always there to solve any query of mine and also guide me in the right direction regarding the project. Without her help and inspiration, I would not have been able to complete the project. Also, I would like to thank my batchmates who guided me, helped me, and gave ideas and motivation at each step.

A photograph of a handwritten signature in black ink on a light-colored surface. The signature reads 'Gaurav Jain' with a stylized underline.

Gaurav Jain

## **TABLE OF CONTENT**

<b>Abstract</b>	<b>6</b>
<b>Introduction</b>	<b>6</b>
<b>History of Data Science</b>	<b>7</b>
<b>Life Cycle of Data Science</b>	<b>8</b>
<b>Skill Set Required for Data Science</b>	<b>10</b>
<b>Application of Data Science</b>	<b>22</b>
<b>Future of Data science</b>	<b>22</b>
<b>Conclusion</b>	<b>23</b>

# The Insightful World of Data – Data Science

Gaurav Jain, [gj979986@gmail.com](mailto:gj979986@gmail.com), 91+7011579022

**Abstract** - Data Surrounds us , the world is comprised of Data. Everything we do is connected with data, in a similitude data is spine of the present world yet that Data is convoluted and unstructured. If there is a way, we can utilize that/create insights from that we will accomplish a lot. We will have an Ocean of information that we can utilize it in a fruitful manner, and to accomplish that we have Data Science. The issue is that world is Developing at an exponential rate, as the world entered the period of enormous information and data, Traditional methods and procedures are presently not adequate to deal with it, better approaches for handling and processing data are developing since data is advancing as well. Data science is an emerging ambidextrous field that Combines traditional strategies like Statistics and mathematics with Computer Science. The fundamental point of Data science is to convert big sets of both unorganized and organized data into valuable information/Insights that can help organizations with settling on powerful and profitable data-driven choices. It encases around Collecting Data for Analysis and Processing, performing advanced Data Analysis, Data Visualization, Machine learning algorithms, notion of Deep learning and introducing the results to uncover patterns and empower organization, to draw informed insights. Since Data is Everywhere so application of Data science is diverse, from aiding associations to medical care, travel, government, social media, web-based media and so on. The aim of this paper is to introduce and explain Data Science and to introduce its advantages in genuine application in different fields

**Keywords:** Data Science, Data analysis, big data, Machine Learning, Deep Learning, Mathematics, Data visualization

## I. INTRODUCTION

We talk about data being DNA of today's world but what exactly Data is? In Simple words Data is Just Facts or figures, or information from which Interpretation can be drawn, moreover these Facts are everywhere so in today's world the main problem is not how to collect data, but how to extract useful, important insights from it. As Discussed above, the answer to all this is Data Science. As Defined above fundamental point of Data science is to convert big sets of both unorganized and organized data into useful information/Insights that can help organizations with settling on powerful and profitable data-driven choices . In simple words making data useful is what Data science is. The magic of data science occurs in 5 General steps – **OSEMN** Framework. **O** stands for **Obtain Data**. **S** stands for **Scrub Data**. **E** stands for **Explore Data**. **M** stands for **Model Data**. **N** stands for **Interpreting Data**. To perform these magical steps Different roles/responsibilities are present in Data science that are, Data Scientist, Data Engineer, Data Architect, Data Analyst. For these roles to perform those magical steps certain skills are required for example Knowledge of Mathematics and statistics, Programming languages, Database languages, Data visualization platforms, Machine Learning, AI, Communication skills, storytelling skills, etc. We will Discuss about all these things later on. The aim of this paper is to: Present a Short Summary of the History of Data Science, Explain the Life cycle of Data Science, Presenting the req skill set, outline the benefits and various application of data Science in real life, and will discuss about its future.

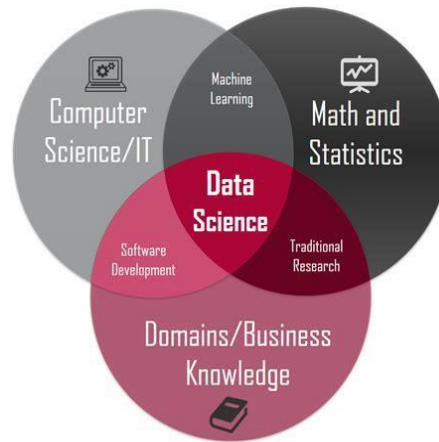


Figure 1: Data Science

## II. HISTORY OF DATA SCIENCE

As traditional ways of processing data required a lot of time and were inefficient, a solution was required which would be faster and more efficient. so now let us look how data science emerged as a solution of this problem:

### 1. Inspection of the data potential

- **1962:** John W Turkey Published his famous article “*The future of data analysis.*” In which he intricates the Correlation b/w data analysis and statistics
- **1974:** Peter Naur authored the *Concise Survey of Computer Methods*, Using the term “Data science,” repeatedly.

### 2. Further Research on potential of data

- **1977:** The IASC (International Association of Statistical Computing) was found. Their mission was to link traditional Statistical methodology with modern computer technology. It is the same period when Turkey’s second article was also published which presented the hypotheses for testing and data analysis
- **1989:** In this year the Knowledge Discovery in Databases organized its 1<sup>st</sup> Workshop.

### 3. Ignition of Data Science

- **1994:** This era ignited the beginning of Data Science as Businesses started getting attracted to Data science as it became popular, and they realize the true potential Data science holds.
- Business week ran the cover story, *Database Marketing*, revealing the ominous news companies had begun collecting enormous amounts of personal information, with plans to initiate new marketing campaigns.

### 4. Practice of Data Science:

- **2001:** SaaS (Software-as-a-Service) was created. A forerunner to using Cloud-based applications.
- **2002:** The International Council for Science: Committee on Data Science and Technology Began Publishing the *Data Science Journal*, a publication focused on issues such as description of data system, applications, and legal issues.

## 5. A new period of Data Science

- **2008:** This Phase provide us the tittle Data Science. This title soon became a buzzword and a piece of the language. The word was given by Jeff Hammerbacher and DJ Patil.
- **2013:** IBM shared insights showing 90% of the data in the world had been created inside the last 2 years

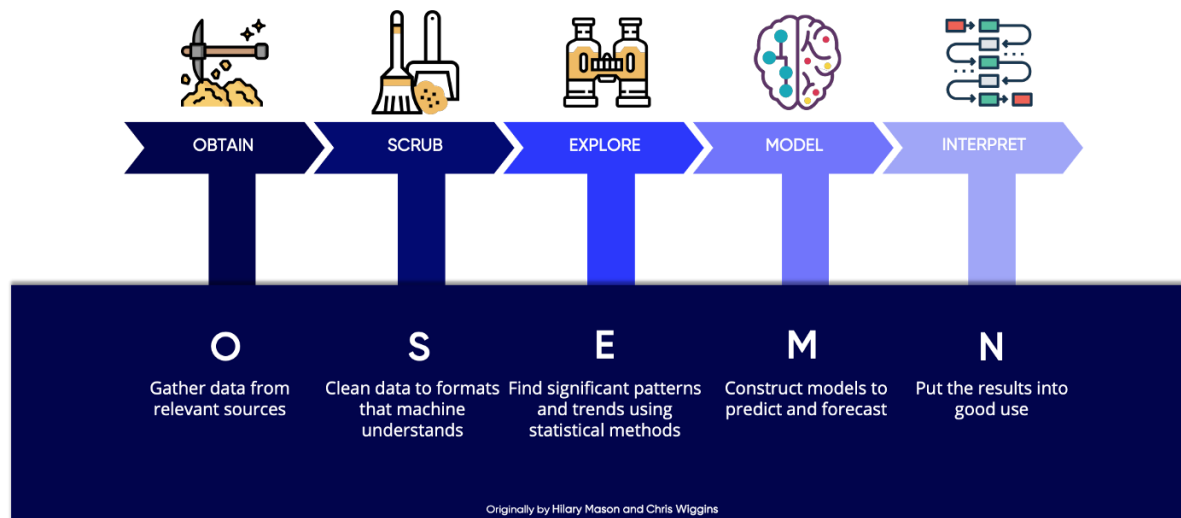
## 6. Data Science in Demand

- **2015:** This stage was the blooming stage of data science. Google and Microsoft started using Speech recognition, deep learning techniques and Apple gave credit to data Science for its increase insales which had dramatic increment of 49%.
- Bloomberg's Jack Clark, wrote it had been a landmark year of AI, ML.

In the Past ten years, Data science has grown drastically from including businesses and organizations to being used by government, engineers and even astronauts. Although it did not receive a warm welcome, the field has come a long way. Today it helps firms to increase their profit and help mankind in many ways

## III. LIFE CYCLE OF DATA SCIENCE

### Data Science Process



In this segment, we will Examine about How we Convert Raw data into Useful data. These steps of Conversion are Known as Life Cycle of Data Science. This Cycle works in an **OSEMN Framework**. Before we begun with the cycle it is important to comprehend the problem we are trying to solve, and this step is known as:

- Business Requirement:** In this step we just identify and comprehend the problem. We do that by identifying Central objectives and by identifying variables that needs to be predicted.

After that we can begin our **OSEMN Framework**

- O is for – Obtain Data / Data Acquisition:** After Business Requirement, in this we gather the data we need from available Data Sources. This Whole Process of Gathering Data is known as Data mining. We can obtain data from different databases and some Web APIs. Using Different skills. We will Discuss about the skill set in the next section of the paper.



2. **S is for – Scrub Data / Data Processing:** The data that we gathered is raw and unfiltered and is a huge mess. So, in this step we clean the Data, the process is known as **Data Wrangling**. In this we first transform data into desired and generalized format, then replace the missing data and remove corrupt and unnecessary data / values.
3. **E is for - Explore Data / Data Exploration:** Here we understand the data, form patterns in it to retrieve useful insights, whole procedure is known as **Data analysis and Data Visualization**. Analysis and visualization are very similar. Analysis is done by inspecting the data, making relation in it to retrieve insights. Visualization is done by using Different Graphs (Histogram, bar, etc.) to identify trends in the data.
4. **M is for - Model Data / Modelling:** This is the stage “*Where the magic happens*”. After Exploring Data, we construct models, that will predict, perform Different tasks, and will solve the problem more accurately. Here model training occurs in which insights are split into Training and Testing data set. Then we create Model using Training Data Set and evaluate it with testing Data Set using **Machine Learning Algorithms**.
5. **N is for – Interpreting Data / Deployment:** This stage refers to the presenting data to users with no technical background, hence deployment of the model into the real world. Soft skills such as Communication, storytelling skills, etc., are required. Here we deliver answers to the issues we faced when we started the cycle.

## IV. SKILL SET REQUIRED FOR DATA SCIENCE

In this segment, we will Discuss about the Essential skill set Required for executing the Life cycle. Basically, there are 2 types of Skill set. That are **CORE** and **SOFT** skills.

### CORE:

#### 1. Mathematics

- **Statistics and Probability:** The Knowledge of the concept of High School Statistics, +
  - A] Mean
  - B] Median
  - C] Mode
  - D] Variance
  - E] Standard Deviation

#### Intermediate Level

- A. Probability distribution
- B. p-value
- C. Correlation coefficient and the covariance matrix
- D. MSE
- E. R2 Score
- F. Bayes Theorem
- G. A/B testing
- H. Monte Carlo Simulation

### **Advance level**

- A. Q-Q Plot
- B. Chebyshev's inequality
- C. Discrete and Continuous Distribution
- D. Log Normal Distribution
- E. Power Law Distribution
- F. Box Cox Transform
- G. Poisson Distribution
- H. Application of non-Gaussian Distribution

- **Multivariable Calculus:** The Knowledge of the concept of High School and *university level* calculus + Basics like (**Integration-Differentiation-Differential Equation**)

- A. Function of Several variables
- B. Step and Sigmoid Functions
- C. Logit Functions
- D. ReLU (Rectified Linear Unit) Functions
- E. Cost Functions
- F. Plotting of Functions
- G. Minima and Maxima of a Function

- **Linear Algebra:** Knowledge of

- A. Vectors
- B. Matrices
- C. Transpose of a matrix
- D. Inverse of matrix
- E. Determinant of matrix
- F. Dot Product
- G. Eigenvalues
- H. Eigenvectors

2. **Programming Languages:** You should have some Basic knowledge of the Following languages, on top of that knowledge on the packages associated with these languages is a must. [ as these packages will be helpful in Data wrangling, Data visualization, Data Analysis, and Machine Learning]

- **Python**

- A] NumPy
- B] Pandas
- C] Matplotlib
- D] Seaborn
- E] Scikit-learn
- F] PyTorch
- G] Scrubadub
- H] TensorFlow

- **R**
  - A] Tidyverse
  - B] Dplyr
  - C] Ggplot2
  - D] Caret
  - E] Stringr
- **Skills in other Database languages and Visualization Tools:**
  - A] Excel
  - B] Tableau
  - C] Hadoop
  - D] SQL
  - E] Spark
  - F] Power BI

3. **Data Wrangling Skills:** As discussed in the previous segment { refer... segment III of the Paper, i.e., **S is for – Scrub Data / Data Processing** }, So we know that this is the process where the crude data gets cleansed in various steps.  
**We can use Hadoop or spark by scripting or by using Python and R as Scripting materials.**

**FOR EXAMPLE:** We have a Report of Dallas Police Shooting . Here we need to clean the names of police officers for the sake of Privacy. We will use Pandas and Scrubadub as Scripting in Python and will have a Single Generalized format {CSV}

---

### Loading CSV File using Pandas

---

```
import pandas as pd
df = pd.read_csv('~\Dallas_Police_Officer-Involved-Shootings.csv')
df.head(n=5)
```

	Case #	Date	Location	Subject Deceased, Injured, or Shoot and Miss	Subject Weapon	Subject(s)	Officer(s)	Grand Jury Disposition	Attorney General Forms URL	Summary U
0	44523A	02/23/2013	3000 Chihuahua Street	Injured	Handgun	Curry, James L/M	Patino, Michael L/M; Fillingim, Brian W/M	No Bill	NaN	<a href="http://dallas.gov">http://dallas.gov</a>
1	121982X	05/03/2010	1300 N. Munger Boulevard	Injured	Handgun	Chavez, Gabriel L/M	Padilla, Gilbert L/M	No Bill	NaN	<a href="http://dallas.gov">http://dallas.gov</a>
2	605484T	08/12/2007	200 S. Stemmons Freeway	Other	Shotgun	Salinas, Nick L/M	Poston, Jerry W/M	See Summary	NaN	<a href="http://dallas.gov">http://dallas.gov</a>
3	384832T	05/26/2007	7900 S. Loop 12	Shoot and Miss	Unarmed	Smith, James B/M; Dews, Antonio B/M; Spearman,...	Mondy, Michael B/M	NaN	NaN	<a href="http://dallas.gov">http://dallas.gov</a>
4	244659R	04/03/2006	6512 South	Injured	Hands	Watkins, Caleb B/M	Armstrong, Michael	No Bill	NaN	<a href="http://dallas.gov">http://dallas.gov</a>

**Figure 3:** Dallas Police Shooting Report involving Police Officers

---

## Removing names using Scrubadub

---

```
import scrubadub

scrub = lambda x: scrubadub.clean(x.decode('utf-8'), replace_with='identifier')

df['Officer(s)'] = df['Officer(s)'].apply(scrub)
```

---

## Writing Cleansed data back to CSV

---

```
df.to_csv("~/Dallas_Police_Officer-Involved_Shootings.csv", encoding='utf-8', index = False)
```

	Case #	Date	Location	Subject Deceased, Injured, or Shoot and Miss	Subject Weapon	Subject(s)	Officer(s)	Grand Jury Disposition	Attorney General Forms URL	Summary UR
0	44523A	02/23/2013	3000 Chihuahua Street	Injured	Handgun	Curry, James L/M	{{NAME-0}}, {{NAME-1}}, {{NAME-2}}, {{NAME-3}},...	No Bill	NaN	http://dallaspc
1	121982X	05/03/2010	1300 N. Munger Boulevard	Injured	Handgun	Chavez, Gabriel L/M	{{NAME-6}}, {{NAME-7}}, {{NAME-2}}	No Bill	NaN	http://dallaspc
2	605484T	08/12/2007	200 S. Stemmons	Other	Shotgun	Salinas, Nick L/M	{{NAME-8}}, {{NAME-9}}	See Summary	NaN	http://dallaspc

Figure 4: Dallas Police Shooting Report after Data wrangling

This was just a small example to show the working of Data Wrangling, it is much Bigger and deeper

4. **Data Analysis and Visualization:** As Discussed in the Previous Segment {refer... Segment III of the paper, i.e., **E is for - Explore Data / Data Exploration**} Data analysis is where we comprehend the data and Form relations in it. We can use Excel, SQL, Python (Pandas are best in this).

**FOR EXAMPLE:** We have 2 files **Train.csv**, and **Test.csv** Which contains data of Passengers of titanic ship. It includes Survival intel, so we will analyze the survival rate of both genders from Train.csv and will predict in test.csv

Train.csv



train.csv

Test.csv



test.csv

Now using pandas and NumPy we will perform analysis-

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestClassifier #for model building
train_data = pd.read_csv("/input/titanic/train.csv")
train_data.head()
test_data = pd.read_csv("/input/titanic/test.csv")
test_data.head()
```

```
train_data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Figure 5: First 5 values of File Train.csv

```
test_data.head()
```

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

Figure 6: First 5 values of File Test.csv

```
women = train_data[train_data.Sex=="female"].Survived
women_survival_predict = sum(women)/len(women)
women_survival_predict

men = train_data[train_data.Sex=="male"].Survived
men_survival_predict = sum(men)/len(men)
men_survival_predict

print("% of women Survived: ", women_survival_predict)
print("% of men Survived: ", men_survival_predict)
```

```
% of women Survived: 0.7420382165605095
% of men Survived: 0.18890814558058924
```

Figure 7: Survival % of men and women

Insights that we Extracted from this is that 75% women and 18-19% men survived. Remember This was just a small example to show the working of Data analysis, it is much Bigger and deeper

**Data Visualization** as Discussed in the Previous Segment {refer... Segment III of the paper, i.e., **E is for - Explore Data / Data Exploration**} refers to forming patterns and comprehending trends in the data. Apart from packages programming languages offer, Some other skills such as matplotlib , SciPy, etc. can be used. Most popular being Tableau.

#### Knowledge of Graphs are important in this

- A] Histogram
- B] Bar charts
- C] Pie Charts

#### Advance Charts

- A] Waterfall Charts
- B] Thermometer Charts, etc.

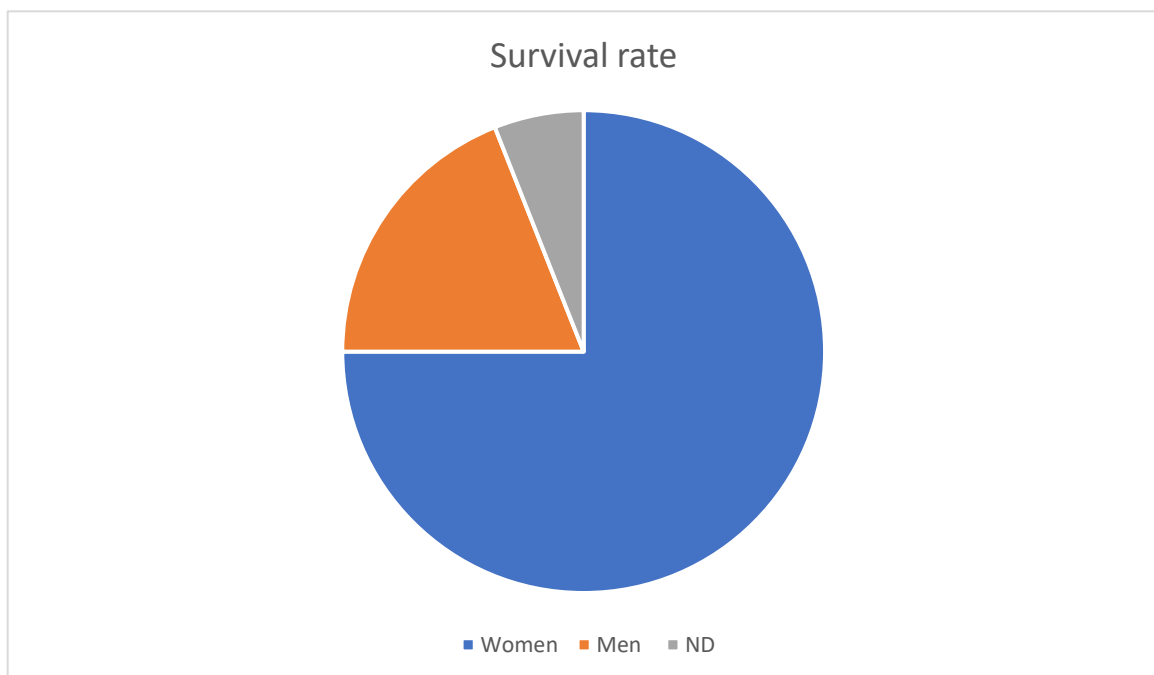


Figure 7: Pie Chart Visualization of the titanic analysis

5. **Machine Learning:** Machine Learning refers to those Computer algorithms that automatically improve through experience. It imitates the human learning. In certain way, building certain algorithms with using data to create dynamic learning. As Discussed in Segment III, It is used to create and train models. Now Skills required in this are :

**A. Supervised Machine Learning**

- i. Linear Regression
- ii. Logistic Regression
- iii. **Decision Tree**
- iv. SVM
- v. Naïve Bayes
- vi. kNN
- vii. **Random Forest**
- viii. Gradient Boosting algorithms
  - GBM
  - XGBoost
  - LightGBM
  - CatBoost

**B. Unsupervised Machine Learning:**

- i. K means
- ii. DBSCAN
- iii. PCA
- iv. Hierarchical Clustering

**C. Reinforcement**

- i. Deep Q Networks
- ii. Deep Deterministic Policy Gradient
- iii. A3C algo
- iv. Q Learning

**FOR EXAMPLE:** In the Continuation of the titanic ship data analysis now we will use certain packages and algorithms to create and teach a model which will predict who survived or not in test.csv [Here for model building, **Random Forest** and **Decision tree algorithms** are used ].

```
#model building and training
#training and testing data set
target = train_data["Survived"]
features = ["Pclass", "Sex", "SibSp", "Parch", "Embarked"]
X = pd.get_dummies(train_data[features]) #training data test #converting strings to numerical
X_test = pd.get_dummies(test_data[features])
X
X_test
```

	Pclass	SibSp	Parch	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	3	0	0	0	1	0	1	0
1	3	1	0	1	0	0	0	1
2	2	0	0	0	1	0	1	0
3	3	0	0	0	1	0	0	1
4	3	1	1	1	0	0	0	1
...	...	...	...	...	...	...	...	...
413	3	0	0	0	1	0	0	1
414	1	0	0	1	0	1	0	0
415	3	0	0	0	1	0	0	1
416	3	0	0	0	1	0	0	1
417	3	1	1	0	1	1	0	0

Figure 8: Conversion of String values to numerical so that model can work

```

model = RandomForestClassifier(n_estimators=150, max_depth=7, random_state=1)
model.fit(X, target)
predictions = model.predict(X_test)

output = pd.DataFrame({'PassengerId' : test_data.PassengerId, 'Survived' : predictions})
output

```

Out[11]:

	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	1
...	...	...
413	1305	0
414	1306	1
415	1307	0
416	1308	0
417	1309	1

418 rows × 2 columns

Figure 9: Predictions done by model



Final Result.csv



6. **Deep Learning:** Deep Learning is a subset of ML. It attempts to mimic human brain. Essentially it is a Neural Network with 3 or more layers.

Now skill set req for learning Deep learning is :

- 1) Neural Network, Loss Function, Optimizers –{Gradient Descent, SGD, Adagrad, RMSprop, Adam}
- 2) Artificial Neural Network → Deployment projects
- 3) Convolutional Neural Network → Transfer Learning (Example: Vgg16, Alexnet)→ Object Detection –{RCNN, Masked RCNN, SSD, YOLO} → Deployment Projects
- 4) RNN –{LSTM, GRU, Bidirectional, LSTM} → Word Embeddings → Word2vec → Encoders and Decoders, Attention Models → Transformers → BERT

7. **NLP (Natural Language Processing):** Being subset of ML, this field focuses on connection b/w Computers and Humans. It provides computers the ability to comprehend text and spoken words the same way human beings can. Allow us to retrieve information from text.

Now skill set req for learning NLP (similar to DL) :

- 1) Neural Networks, Loss Function, Optimizers- {Gradient Descent, SGD, Adagrad, RMSprop, Adam}
- 2) Text Preprocessing- {Gensim, Word2vec, AvgWord2vec}
- 3) Solve Machine Learning Usercases
- 4) Artificial Neural Network
- 5) Recurrent Neural Networks, LSTM, GRU
- 6) Text Preprocessing Level 3- Word Embeddings, Word2vec
- 7) Bidirectional LSTM RNN, Encoders and Decoders, Attention Models
- 8) Transformers
- 9) BERT

## **SOFT SKILLS:**

**Communication Skills** Discussed in the segment III of the paper {Refer... **N is for – Interpreting Data / Deployment**}, This skill helps in deploying model and answering the questions we once started with. Quite possibly the main skill set in data science. In order to explain the model to users with no technical background, you should have a smooth and a good set of communication skills, on top of that its subset which is **Storytelling skills** plays an important part, as it will be really helpful in explaining the model to users or stakeholders in an interesting manner. Most Importantly, in this field one should **never stop learning**, as data is ever lasting and ever evolving so there are new learnings, research and discoveries

occurring every day in this field. **Learning never stops here.** Even after getting involved in this field as one of its ‘roles’ {refer to **fig 10**} one must upgrade and update every day.

To do that, an individual can utilize platforms like **LinkedIn, GitHub, Kaggle etc.** Very useful for keeping yourself up-to-date about recent development in this fields.

ROLES	RESPONSIBILITIES
<b>Data Architect</b> 	Develops data architecture to effectively capture, integrate, organize, centralize and maintain data. Core responsibilities include: <ul style="list-style-type: none"> <li>✓ Data Warehousing Solutions</li> <li>✓ Extraction, Transformation and Load (ETL)</li> <li>✓ Data Architecture Development</li> <li>✓ Data Modeling</li> </ul>
<b>Data Engineer</b> 	Develop, test and maintain data architectures to keep data accessible and ready for analysis. Key tasks are: <ul style="list-style-type: none"> <li>✓ Extraction Transformation and Load (ETL)</li> <li>✓ Installing Data Warehousing Solutions</li> <li>✓ Data Modeling</li> <li>✓ Data Architecture Construction and Development</li> <li>✓ Database Architecture Testing</li> </ul>
<b>Data Analyst</b> 	Processes and interprets data to get actionable insights for a company. Responsibilities include: <ul style="list-style-type: none"> <li>✓ Data Collection and Processing</li> <li>✓ Programming</li> <li>✓ Machine Learning</li> <li>✓ Data Munging</li> <li>✓ Data Visualization</li> <li>✓ Applying Statistical Analysis</li> </ul>
<b>Data Scientist</b> 	Data analysis once data volume and velocity reaches a level requiring sophisticated technical skills. Core tasks are: <ul style="list-style-type: none"> <li>✓ Data Cleansing and Processing</li> <li>✓ Predictive Modeling</li> <li>✓ Machine Learning</li> <li>✓ Identifying Questions</li> <li>✓ Running Queries</li> <li>✓ Applying Statistical Analysis</li> <li>✓ Correlating Disparate Data</li> <li>✓ Storytelling and Visualization</li> </ul>

Fig 10: Different Roles in Data Science

## V. APPLICATION OF DATA SCIENCE

Well, what is the importance of anything if it is not applicable in real life? 0 right. So, in this segment we will examine how those insights are used i.e., application of data science.

Well Application of Data Science is countless. It is used in finance, genetics, banking, medicine, business, and transportation for problems like financial trading, credit scoring, fraud detection, online advertising, direct marketing, internet search, recommendations for cross-selling, etc.

For example, In **Healthcare Industry**, Classification algorithms are used to detect cancer and tumors at an early stage using *image recognition*. In **Medicine and drug Development Industry** DS plays an important role. DS enables us to create drugs faster and with high precision using less resources, as algorithms will predict how a drug will react to the human body.

Now in **Image Recognition**, let's say you upload a picture with your mate on Instagram, then Instagram will give suggestions regarding the tags. This is done with the help of ML and DS. When an Image is Recognized, Data analysis is done on one's Instagram friends and after analysis, if the features of the faces which are present in the picture matches with someone else profile, then Instagram suggests auto-tagging.

We saw how Important DS is, and it is not only used for associations but also for helping mankind and making their life easier, Like in **Search engines** it is used to get quick searches and best results, in **Target recommendation** it is used in giving out recommended ads/products on the basis of user's search history and it is also used in day to day life in the form of Voice assistant (Siri, Alexa, Google Assistant). Using speech recognition to understand humans voice, convert them into textual data using NLP, Deep Learning and ML and then giving out the appropriate answers using data analysis.

## VI. FUTURE AND CONCLUSION

With the colossal expansion in data, there is and will be a consistent requirement for organizing and analyzing such large amount of data. Data Science is a field which can manage all this and foster valuable AI models that anticipate future outcomes.

With the pace this world is developing, it is safe to say that DS will have a bright future. As long as data remains in this world, the requirement for data science will be there. Data is ever evolving and everlasting. So, in future data science will be required more than ever but,

The working distributions of the roles will be spilt and talking about AutoML ( it is a process of automating the objectives of applying machine learning without being an expert ), will it replace or destroy DS? The answer is No, but it will modify the working of data science, it will bring a standard shift in the working of data science. Building models to do analysis wouldn't remain as their simple task but will shift to the true analysis of data, soft skills will become more and more important.

So, in a nutshell future of data science is technically a modification, it will not grow with the same rate but will change according to the changes in tech. We can say that Data science which is an emerging ambidextrous field that Combines traditional strategies like Statistics and mathematics with Computer Science to extract useful insights to help associations and mankind in different ways has a bright future and Data science will continue to evolve into new phases and stages depending on the need of the hour and the development of the tech.

moreover, as long as data exists Data science will exist.

## REFERENCES

- [1] Provost, F. and Fawcett, T. (2013). Data science for business: what you need to know about data mining and data-analytic thinking (1<sup>st</sup> ed.). Sebastopol: O'Reilly.
- [2] Van der Aalst, W. (2016). Data science in action. In W. van der Aalst, Process mining (pp. 3–23). Berlin; Heidelberg: Springer.
- [3] Foote, K. D. (2016). A brief history of data science. Retrieved 17. 10. 2019 from <https://www.dataversity.net/brief-history-data-science/#>.
- [4] <https://www.ibm.com/cloud/learn/data-science-introduction>
- [5] <https://www.jigsawacademy.com/blogs/data-science/history-of-data-science/>
- [6] YouTube Channel: Krish Naik-  
<https://www.youtube.com/user/krishnaik06/featured>
- [7] What is data science? [Blog]. (2019). from <https://intellipaat.com/blog/what-is-data-science/>.
- [8] Cole Nussbaumer Knafl (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals (1<sup>st</sup> ed). Wiley
- [9] Zahavi, J. (1999). Mining data for nuggets of knowledge. Retrieved 7. 10. 2019 from <https://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge/>.