

## DS CAT1 SOLUTIONS

### 1. Describe Term Data Science? with its Need?

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

Data science is a field that involves using statistical and computational techniques to extract insights and knowledge from data. It encompasses a wide range of tasks, including data cleaning and preparation, data visualization, statistical modelling, machine learning, and more.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

Need :

- With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
- Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
- Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
- Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.
- It enables you to take better and faster decisions
- It helps you to recommend the right product to the right customer to enhance your business
- Allows to build intelligence ability in machines
- Data Science can help you to detect fraud using advanced machine learning algorithms

2. | Explain Data Science Process in detail?
5. | Write Short note on Data Science Process Lifecycle?
6. | Explain any Two Phases of Data Science Process?

Data science is not a one-step process such that you will get to learn it in a short time and call ourselves a Data Scientist. It's passes from many stages and every element is important. One should always follow the proper steps to reach the ladder. Every step has its value and it counts in your model. Buckle up in your seats and get ready to learn about those steps.

- **Problem Statement:** No work start without motivation, Data science is no exception though. It's really important to declare or formulate your problem statement very clearly and precisely. Your whole model and it's working depend on your statement. Many scientist considers this as the main and much important step of Date Science. So make sure what's your problem statement and how well can it add value to business or any other organization.
- **Data Collection:** After defining the problem statement, the next obvious step is to go in search of data that you might require for your model. You must do good research, find all that you need. Data can be in any form i.e unstructured or structured. It might be in various forms like videos, spreadsheets, coded forms, etc. You must collect all these kinds of sources.
- **Data Cleaning:** As you have formulated your motive and also you did collect your data, the next step to do is cleaning. Yes, it is! Data cleaning is the most favorite thing for data scientists to do. Data cleaning is all about the removal of missing, redundant, unnecessary and duplicate data from your collection. There are various tools to do so with the help of programming in either R or [Python](#). It's totally on you to choose one of them. Various scientist have their opinion on which to choose. When it comes to the statistical part, R is preferred over Python, as it has the privilege of more than 12,000 packages. While python is used as it is fast, easily accessible and we can perform the same things as we can in R with the help of various packages.
- **Data Analysis and Exploration:** It's one of the prime things in data science to do and time to get inner Holmes out. It's about analyzing the structure of data, finding hidden patterns in them, studying behaviors, visualizing the effects of one variable over others and then concluding. We can explore the data with the help of various graphs formed with the help of libraries using any programming language. In R, GGplot is one of the most famous models while Matplotlib in Python.
- **Data Modelling:** Once you are done with your study that you have formed from data visualization, you must start building a hypothesis model such that it may yield you a good prediction in future. Here, you must choose a good algorithm that best fit to your model. There different kinds of algorithms from regression to classification, SVM( Support vector machines), Clustering, etc. Your model can be of a [Machine Learning](#) algorithm. You train your model with the train data and then test it with test data. There are various methods to do so. One of them is the K-fold method where you split your whole data into two parts, One is Train and the other is test data. On these bases, you train your model.
- **Optimization and Deployment:** You followed each and every step and hence build a model that you feel is the best fit. But how can you decide how well your model is performing? This where optimization comes. You test your data and find how well it is performing by checking its accuracy. In short, you check the efficiency of the data model and thus try to optimize it for better accurate prediction. Deployment deals with the launch of your model and let the people outside there to benefit from that. You can also obtain feedback from organizations and people to know their need and then to work more on your model.

OR

OR

OR

OR

OR

**1. Business Understanding:** The complete cycle revolves around the enterprise goal. What will you resolve if you do no longer have a specific problem? It is extraordinarily

essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

**2. Data Understanding:** After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

**3. Preparation of Data:** Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

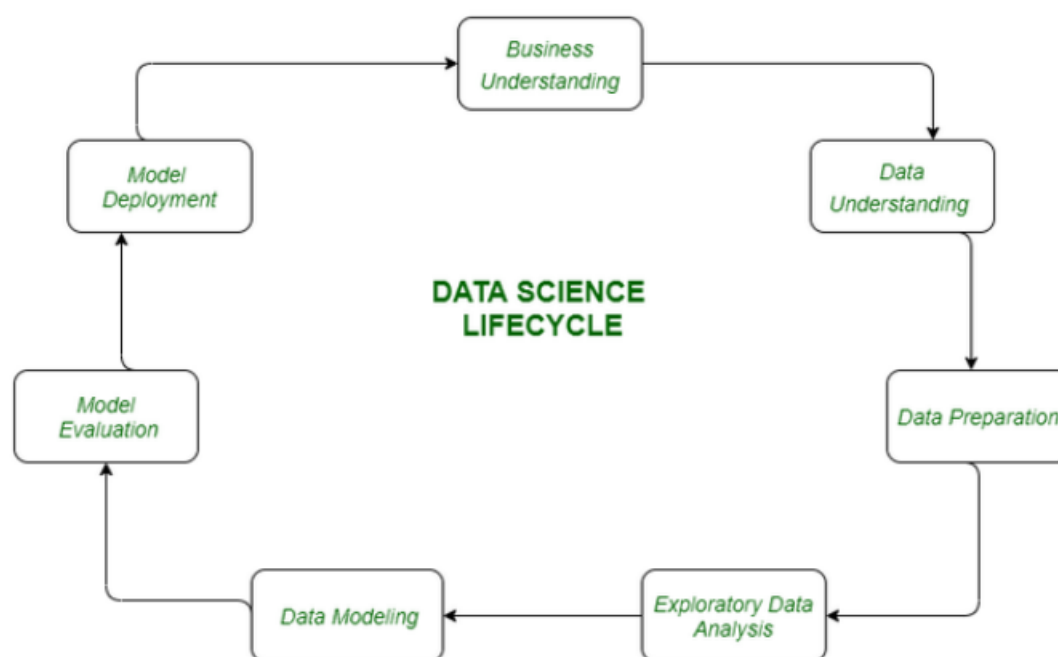
**4. Exploratory Data Analysis:** This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

**5. Data Modeling:** Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

**6. Model Evaluation:** Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously

thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

**7. Model Deployment:** The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.



3. Explain first Phases Of Data Science Process?

4. | Define term: 1. Data Science  
| 2.Data Exploration

## Data Science:

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

Data science is a field that involves using statistical and computational techniques to extract insights and knowledge from data. It encompasses a wide range of tasks, including data cleaning and preparation, data visualization, statistical modelling, machine learning, and more.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

## Data Exploration:

In the data science process, data exploration is leveraged in many different steps including preprocessing, modeling, and interpretation of the results. Data exploration, also known as **exploratory data analysis** (EDA), provides a set of simple tools to achieve a basic understanding of the data.

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.



Compare Discrete Data and continuous data.

BASIS FOR COMPARISON	DISCRETE DATA	CONTINUOUS DATA
Meaning	Discrete data is one that has clear spaces between values.	Continuous data is one that falls on a continuous sequence.
Nature	Countable	Measurable
Values	It can take only distinct or separate values.	It can take any value in some interval.
Graphical Representation	Bar Graph	Histogram
Tabulation is known as	Ungrouped frequency distribution.	Grouped frequency distribution.
Classification	Mutually Inclusive	Mutually Exclusive
Function graph	Shows isolated points	Shows connected points
Example	Days of the week	Market price of a product
Usage	Suitable for counting and categorizing data	Suitable for measuring and comparing data
Definition	Data that can only take specific values or countable values	Data that can take any value within a certain range

## Summarize importance of linear algebra in Data Science

### \* Importance of linear algebra in data science : →

Our Technology is advance to the point where it is today thanks to the number of fields including data science, AI and ML, Robotics and computer vision. As you begin to learn more about this technologies you will run into a number of words that are used frequently in connection with them.

Terms such as support vector machine, <sup>Lagrange</sup> ~~range~~ multiplier and ridge regression.

why imp.?

#### ① Linear algebra can help you to understand statistic better :

For ML to effectively organized and integrate data statistics are crucial, you must first understand how linear algebra functions if you want to better comprehend statistical concepts.

#### ② Data Science success is largely depend on linear algebra :

#### ③ Improved ML algorithms are made <sup>possible</sup> by the use of linear algebra :

#### ④ Data science and Machine learning prediction :

learning linear algebra helps you to develop the intuition or awareness that is so crucial to machine learning and data science. More view points will be available from you now.

your ability to think broadly and unconsciously will improve as a result of your study of matrices & vectors

Linear algebra is very important when it comes to data science. From the notations being used to define the algorithms to their actual implementation, linear algebra is the basis of machine learning.



## Illustrate importance of Statistics in Data Science.

### \* Statistics of data science: →

Statistical skill needed to perform data science work. DS required a mixture of technical skills such as python, R ~~for~~ programming languages as well as softskills including communication and attention to detail.

Important Skills Data scientist need in order to strengthen statistical abilities:

- ① Data manipulation.
- ② Critical thinking and attention to detail.
- ③ Quantity.
- ④ Organization.
- ⑤ Innovation and problem solving.

#### ① Data manipulation: →

Using excel, R, stata and other programs. Data scientists have the ability to clean and to organization large datasets.

#### ② Critical thinking and attention to detail:

Using linear regression data scientist extract and model relationships between dependent and independent variable.

#### ③ Quantity:

The desire to solve the complex puzzles, data scientist to design data plots and explore assumptions. They also discovered patterns by using advance data visualization.

#### ④ Organization:

Data Scientist are updated with information from various sources and ongoing project opportunities with budget and time constraints. Data scientist perform efficiently when they are mainly statistical function.

#### ⑤ Innovation and problem solving:

Above and beyond pure computation and basic data analysis data scientist use applied statistics to point abstract finding to real world problem.

When the data is big and unorganised, statistics plays a powerful role in that situation. When a company uses statistics to find insights, it makes the tedious task look minimalist and easy in front of the big and buffer information that was provided earlier.

Some ways in which Statistics helps in Data Science are:

1. **Prediction and Classification:** Statistics help in prediction and classification of data whether it would be right for the clients viewing by their previous usage of data.
2. **Helps to create Probability Distribution and Estimation:** Probability Distribution and Estimation are crucial in understanding the basics of machine learning and algorithms like [logistic regressions](#).

Cross-validation and LOOCV techniques are also inherently statistical tools that have been brought into the Machine Learning and [Data Analytics](#) world for inference-based research, A/B and hypothesis testing.

3. **Pattern Detection and Grouping:** Statistics help in picking out the optimal data and weeding out the unnecessary dump of data for companies who like their work organised. It also helps spot out anomalies which further helps in processing the right data.
4. **Powerful Insights:** Dashboards, charts, reports and other [data visualizations types](#) in the form of interactive and effective representations give much more powerful insights than plain data and it also makes the data more readable and interesting.
5. **Segmentation and Optimization:** It also segments the data according to different kinds of demographic or psychographic factors that affect its processing. It also optimizes data in accordance with minimizing risk and maximizing outputs.

## Contrast Nominal data and Ordinal data

There are two types of qualitative data. They are:

Nominal is used to label variables without any order or quantitative value.

ex. ① color of hair, gender, profession

ordinal data have natural number ordering where a number is present in some kind of order by their position on the scale.

ex. ① when companies asked for feedback, experience or satisfaction on a scale of 1 to 10.

② Ranking of people.

- **Nominal data** denotes labels or categories (e.g. blonde hair, brown hair).
- Nominal data are categorical, the categories being mutually exclusive without any overlap.
- The categories of nominal data are purely descriptive, that is, they do not possess any quantitative or numeric value. Nominal data can never be quantified.
- Nominal data cannot be put into any definite order or hierarchy. None of the categories can be greater than or worth more than one another.
- The mean of nominal data cannot be calculated even if the data is arranged in alphabetical order.
- The mode is the only [measure of central tendency](#) for nominal data.
- In most cases, nominal data is alphabetical.
- Some examples of nominal data are:

Which state do you live in? (Followed by a drop-down list of names of states)

Hair Color (black, brown, grey, blonde)

Preferred mode of Public Transport (bus, tram, train)

Employment Status (employed, unemployed, retired)

Literary Genre (comedy, tragedy, drama, epic, satire)

- **Ordinal data** refers to data that can be categorized and also ranked according to some kind of order or hierarchy (e.g. low income, medium income, high income).

Ordinal data is a kind of qualitative data that groups variables into ordered categories. The categories have a natural order or rank based on some hierarchical scale.

The main differences between Nominal Data and Ordinal Data are:

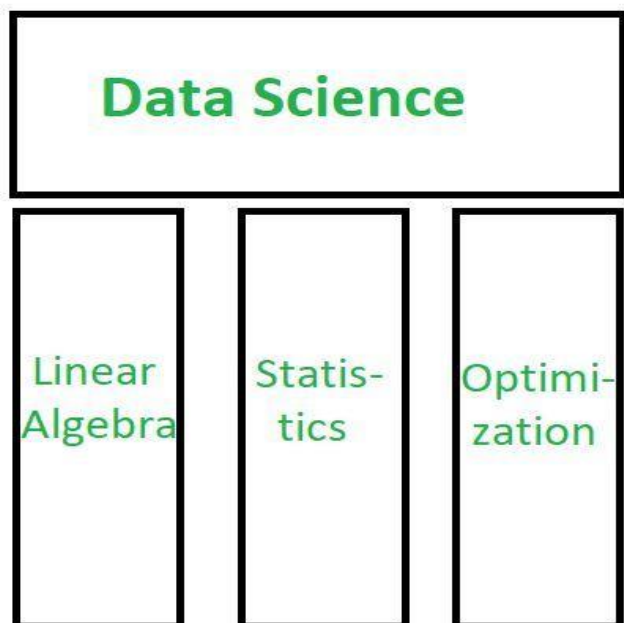
- While Nominal Data is classified without any intrinsic ordering or rank, Ordinal Data has some predetermined or natural order.
- Nominal data is qualitative or categorical data, while Ordinal data is considered “in-between” qualitative and quantitative data.
- Nominal data do not provide any quantitative value, and you cannot perform numeric operations with them or compare them with one another. However, Ordinal data provide sequence, and it is possible to assign numbers to the data. No numeric operations can be performed. But ordinal data makes it possible to compare one item with another in terms of ranking.
- Example of Nominal Data – Eye color, Gender; Example of Ordinal data – Customer Feedback, Economic Status

Example of Ordinal data – Rate education level according to:

- Elementary
- High School
- College
- Graduate
- Post-graduate

## Explain Importance of Optimization for a Data Science perspective.

From a mathematical foundation viewpoint, it can be said that the three pillars for data science that we need to understand quite well are **Linear Algebra**, **Statistics** and the third pillar is **Optimization** which is used pretty much in all data science algorithms. And to understand the optimization concepts one needs a good fundamental understanding of linear algebra.



### What's Optimization?

Wikipedia defines optimization as a problem where you maximize or minimize a real function by systematically choosing input values from an allowed set and computing the value of the function. That means when we talk about optimization we are always interested in finding the best solution. So, let say that one has some functional form (e.g. in the form of  $f(x)$ ) that he is interested in and he is trying to find the best solution for this functional form. Now, what does best mean? One could either say he is interested in minimizing this functional form or maximizing this functional form.

A basic understanding of optimization will help in:

- More deeply understand the working of machine learning algorithms.
- Rationalize the working of the algorithm. That means if you get a result and you want to interpret it, and if you had a very deep understanding of optimization you will be able to see why you got the result.
- And at an even higher level of understanding, you might be able to develop new algorithms yourselves.



Matrices are the building blocks of data science. They appear in various avatars across languages. From numpy arrays in Python, to dataframes in R, to matrices in MATLAB.

The matrix in its most basic form is a collection of numbers arranged in a rectangular or array-like fashion. This can represent an image, or a network or even an abstract structure.

Conventionally, the number of rows in a matrix is denoted by **m** and the number of columns by **n**. Since a rectangle's area is *height* x *width*, we denote a matrix's size by **m** x **n**. Thus if the matrix was to be called **A**, it would be written notationally as

$$A_{m \times n} = A_{3 \times 4} = \begin{pmatrix} 2 & 5 & 7 & 8 \\ 1 & 2 & 3 & 1 \\ 4 & 5 & 0 & 1 \end{pmatrix}$$

Matrix notation

Here m=3 and n=4. Thus there are 12 elements in the matrix A. A square matrix is one which has **m=n**.

$$A_{3 \times 3} = \begin{pmatrix} 2 & 5 & 7 \\ 1 & 2 & 3 \\ 4 & 5 & 0 \end{pmatrix}$$

Square matrix

A matrix with just one row is called a **row matrix** and a matrix with just one column is called a **column matrix**.

$$\text{Row Matrix} \Rightarrow A_{1 \times 3} = \begin{pmatrix} 2 & 5 & 7 \end{pmatrix}$$

$$\text{Column Matrix} \Rightarrow A_{3 \times 1} = \begin{pmatrix} 2 \\ 1 \\ 4 \end{pmatrix}$$

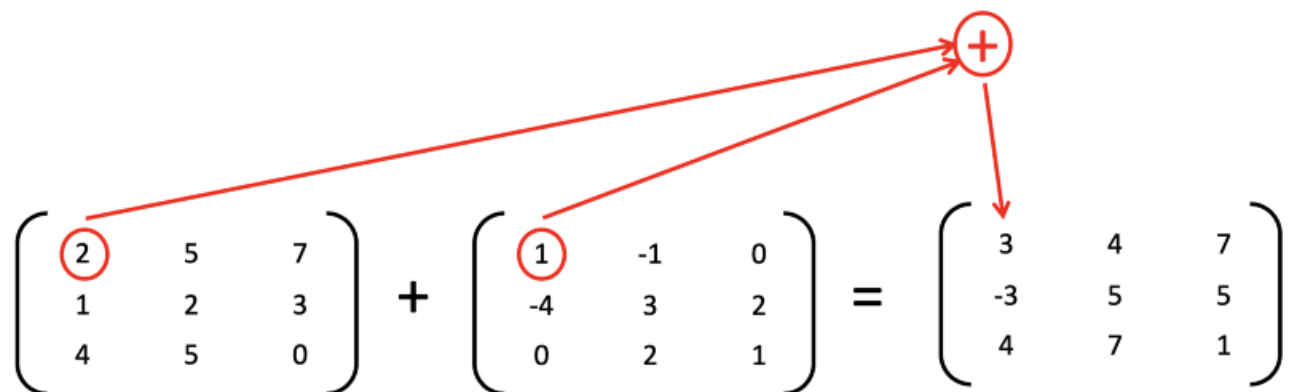
## What can we do with them?

Matrices just like numbers can be added, subtracted and multiplied. The division though is slightly nuanced. Not all matrices can be divided.

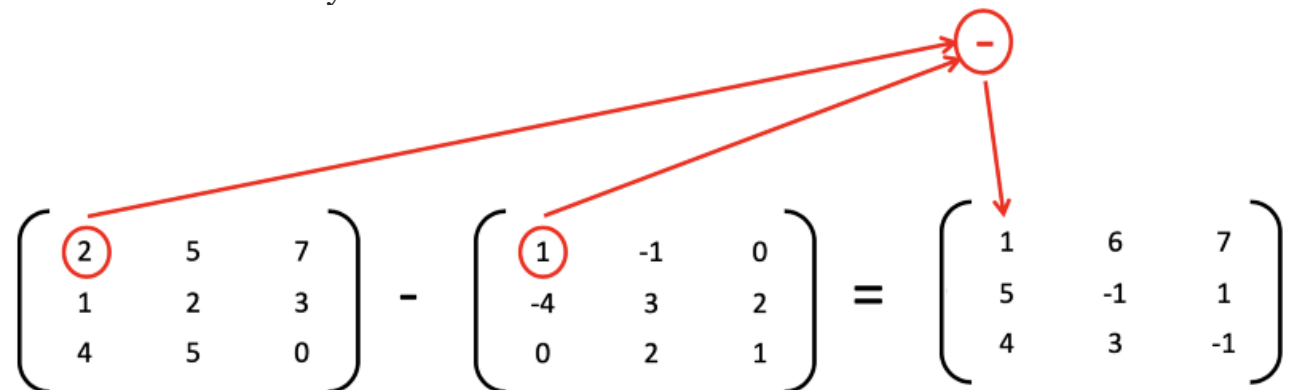
There are certain rules for even the addition, subtraction and multiplication.

### Matrices Addition

The addition of two matrices  $A(m \times n)$  and  $B(m \times n)$  gives a matrix  $C(m \times n)$ . The elements of  $C$  are the sum of corresponding elements in  $A$  and  $B$


$$\begin{pmatrix} 2 & 5 & 7 \\ 1 & 2 & 3 \\ 4 & 5 & 0 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 0 \\ -4 & 3 & 2 \\ 0 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 7 \\ -3 & 5 & 5 \\ 4 & 7 & 1 \end{pmatrix}$$

Subtraction works similarly.


$$\begin{pmatrix} 2 & 5 & 7 \\ 1 & 2 & 3 \\ 4 & 5 & 0 \end{pmatrix} - \begin{pmatrix} 1 & -1 & 0 \\ -4 & 3 & 2 \\ 0 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 6 & 7 \\ 5 & -1 & 1 \\ 4 & 3 & -1 \end{pmatrix}$$

The thing to note here is that you can only add/subtract matrices with the same number of rows and columns i.e. **the same order (order = rows x columns)**

- Number of Rows of  $A$  = Number of Rows of  $B$
- Number of Columns of  $A$  = Number of Columns of  $B$

### Points to note

- Addition of matrices is **commutative** which means  $A+B = B+A$

- Addition of matrices is **associative** which means  $A+(B+C) = (A+B)+C$
- Subtraction of matrices is **non-commutative** which means  $A-B \neq B-A$
- Subtraction of matrices is **non-associative** which means  $A-(B-C) \neq (A-B)-C$
- The order of matrices A, B, A-B and A+B is always the same
- If the order of A and B is different, A+B, A-B can't be computed
- The complexity of addition/subtraction operation is  $O(m*n)$  where  $m*n$  is order of matrices

Multiplication though is slightly complicated

### Matrices Multiplication

The multiplication of two matrices  $A(m*n)$  and  $B(n*p)$  gives a matrix  $C(m*p)$ . Notice that for multiplication you do not need the rows/columns of A and B to be the same. You only need

- No. of Columns of A = No. of Rows of B
- Or, No. of Columns of B = No. of Rows of A.

To calculate the top-left element of the resulting matrix C, multiply elements of 1st row of A with 1st column of B and add them

$$\begin{bmatrix} -4 & 3 & 2 \\ 0 & 2 & 1 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 1 & 2 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 3 & -4 \\ 6 & 9 \end{bmatrix}$$

$m \times n : 2 \times 3 \quad \times \quad n \times p : 3 \times 2 \quad = \quad m \times p : 2 \times 2$

## Multiplication

### Points to note

- Multiplication of matrices is non-commutative which means  $A*B \neq B*A$
- Multiplication of matrices is associative which means  $A*(B*C) = (A*B)*C$
- Existence of  $A*B$  does not imply the existence of  $B*A$
- The complexity of multiplication operation ( $A*B$ ) is  $O(m*n*p)$  where  $m*n$  and  $n*p$  are an order of A and B respectively
- The order of matrix C computed as  $A*B$  is  $m*p$  where  $m*n$  and  $n*p$  are order of A and B respectively

Structured thinking is the process of creating a structured framework to solve an unstructured problem. As a problem-solving methodology, structured thinking involves dividing a large problem into smaller ones in order to solve the big problem faster and more efficiently.

## 5 STRUCTURED THINKING TECHNIQUES FOR DATA SCIENTISTS

1. Six Step Problem Solving Model
2. Eight Disciplines of Problem Solving
3. The Drill Down Technique
4. The Cynefin Framework
5. The 5 Whys Technique

### 1. Six Step Problem Solving Model

This technique is the simplest and easiest to use. As the name suggests, this technique uses six steps to solve a problem, which are:

1. Have a clear and concise problem definition.
2. Study the roots of the problem.
3. Brainstorm possible solutions to the problem.
4. Examine the possible solution and choose the best one.
5. Implement the solution effectively.
6. Evaluate the results.

This model follows the mindset of continuous development and improvement. So, on step six, if your results didn't turn out the way you wanted, go back to step four and choose another solution (or to step one and try to define the problem differently).

My favorite part about this simple technique is how easy it is to alter based on the specific problem you're attempting to solve.



## 2. Eight Disciplines of Problem Solving

The eight disciplines of problem solving offers a practical plan to solve a problem using an eight-step process. You can think of this technique as an extended, more-detailed version of the six step problem-solving model.

Each of the eight disciplines in this process should move you a step closer to finding the optimal solution to your problem. So, after you've got the prerequisites of your problem, you can follow disciplines D1-D8.

1. **D1:** Put together your team. Having a team with the skills to solve the project can make moving forward much easier.
2. **D2:** Define the problem. Describe the problem using quantifiable terms: the who, what, where, when, why and how.
3. **D3:** Develop a working plan.
4. **D4:** Determine and identify root causes. Identify the root causes of the problem using cause and effect diagrams to map causes against their effects.
5. **D5:** Choose and verify permanent corrections. Based on the root causes, assess the work plan you developed earlier and edit as needed.
6. **D6:** Implement the corrected action plan.
7. **D7:** Assess your results.
8. **D8:** Congratulate your team. After the end of a project, it's essential to take a step back and appreciate the work you've all done before jumping into a new project.

## 3. The Drill Down Technique

The drill down technique is more suitable for large, complex problems with multiple collaborators. The whole purpose of using this technique is to break down a problem to its

roots to make finding solutions that much easier. To use the drill down technique, you first need to create a table. The first column of the table will contain the outlined definition of the problem, followed by a second column containing the factors causing this problem. Finally, the third column will contain the cause of the second column's contents, and you'll continue to drill down on each column until you reach the root of the problem.

Once you reach the root causes of the symptoms, you can begin developing solutions for the bigger problem.

## 4. The Cynefin Framework

The Cynefin framework, like the rest of the techniques, works by breaking down a problem into its root causes to reach an efficient solution. We consider the Cynefin framework a higher-level approach because it requires you to place your problem into one of five contexts.

1. **Obvious Contexts.** In this context, your options are clear, and the cause-and-effect relationships are apparent and easy to point out.

2. **Complicated Contexts.** In this context, the problem might have several correct solutions. In this case, a clear relationship between cause and effect may exist, but it's not equally apparent to everyone.
3. **Complex Contexts.** If it's impossible to find a direct answer to your problem, then you're looking at a complex context. Complex contexts are problems that have unpredictable answers. The best approach here is to follow a trial and error approach.
4. **Chaotic Contexts.** In this context, there is no apparent relationship between cause and effect and our main goal is to establish a correlation between the causes and effects.
5. **Disorder.** The final context is disorder, the most difficult of the contexts to categorize. The only way to diagnose disorder is to eliminate the other contexts and gather further information.

## 5. The 5 Whys Technique

Our final technique is the 5 Whys or, as I like to call it, the curious child approach. I think this is the most well-known and natural approach to problem solving.

This technique follows the simple approach of asking “why” five times — like a child would. First, you start with the main problem and ask why it occurred. Then you keep asking why until you reach the root cause of said problem.

[Probability theory](#)[External link:open in new](#) is a branch of mathematics focusing on the analysis of random phenomena. It is an important skill for [data scientists](#) using data affected by chance.

With randomness existing everywhere, the use of probability theory allows for the analysis of chance events. The aim is to determine the likelihood of an event occurring, often using a numerical scale of between 0 and 1, with the number “0” indicating impossibility and “1” indicating certainty.

A classic example of this is a coin toss, where there can be two possible options: heads or tails. Here the possibility of flipping a head or a tail on a single toss is 50%. When conducting your own experiment you may find that the outcomes can vary. But if you continue flipping the coin, the outcome grows closer to 50/50.

Probability allows data scientists to assess the certainty of outcomes of a particular study or experiment. An experiment is a planned study that is executed under controlled conditions. When a result is not already predetermined, the experiment is referred to as a chance experiment. Conducting a coin toss twice is an example of a chance experiment.

Today's data scientists need to have an understanding of the foundational concepts of probability theory including key concepts involving probability distribution, statistical significance, hypothesis testing and regression.

Axioms mean a rule a principle that most people believe to be true. It is the premise on the basis of which we do further reasoning

There are three axioms of probability that make the foundation of probability theory-

#### *Axiom 1: Probability of Event*

The first one is that the probability of an event is always between 0 and 1. 1 indicates definite action of any of the outcome of an event and 0 indicates no outcome of the event is possible.

For any event E,  $0 \leq P(E) \leq 1$

$$P(event) = \frac{\text{count of outcomes in Event}}{\text{count of outcomes in Sample Space}}$$

#### *Axiom 2: Probability of Sample Space*

For sample space, the probability of the entire sample space is 1.

For Sample Space,  $P(S) = 1$

*Axiom 3: Mutually Exclusive Events*

And the third one is- the probability of the event containing any possible outcome of two mutually disjoint is the summation of their individual probability.

$P(A \cup B) = P(A) + P(B)$  for mutually exclusive events



Defend use of matrix factorization in Data Science.