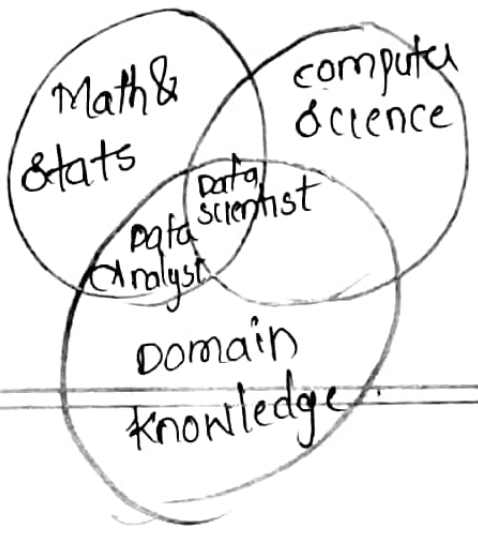
Datascience Using R programming

Assignment-1.

1. What is datascience? How it is different from Data Analysis and business intelligence?

Ans Datascience is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Datascience practitioners apply machine learning algorithms to numbers, text, images, video, audio and more to produce artificial intelligence systems to perform tasks that ordinarily require human intelligence.



Datascience is a field that deals with extracting meaningful information and insights by applying various algorithms, processes, scientific methods from structured and unstructured data. this field is related to big data and one of the most demanded skills currently

DataAnalysis is the technique of observing, transforming, cleaning and modelling raw facts and figures with the purpose of developing beneficial information and acquiring profitable conclusions.

8. What is Datascience process? Explain.

Datascience ~~could a~~ process helps datascientists use the tools to find unseen patterns, extract data, and convert information to actionable insights that can be meaningful to the company. This aids companies and businesses in making decisions that can help in customer retention and profits further, a datascience process helps in discovering hidden patterns of structured and unstructured raw data. the process helps in turning a problem into a solution by treating the business problem as a project

The six steps of the datascience process are as follows:

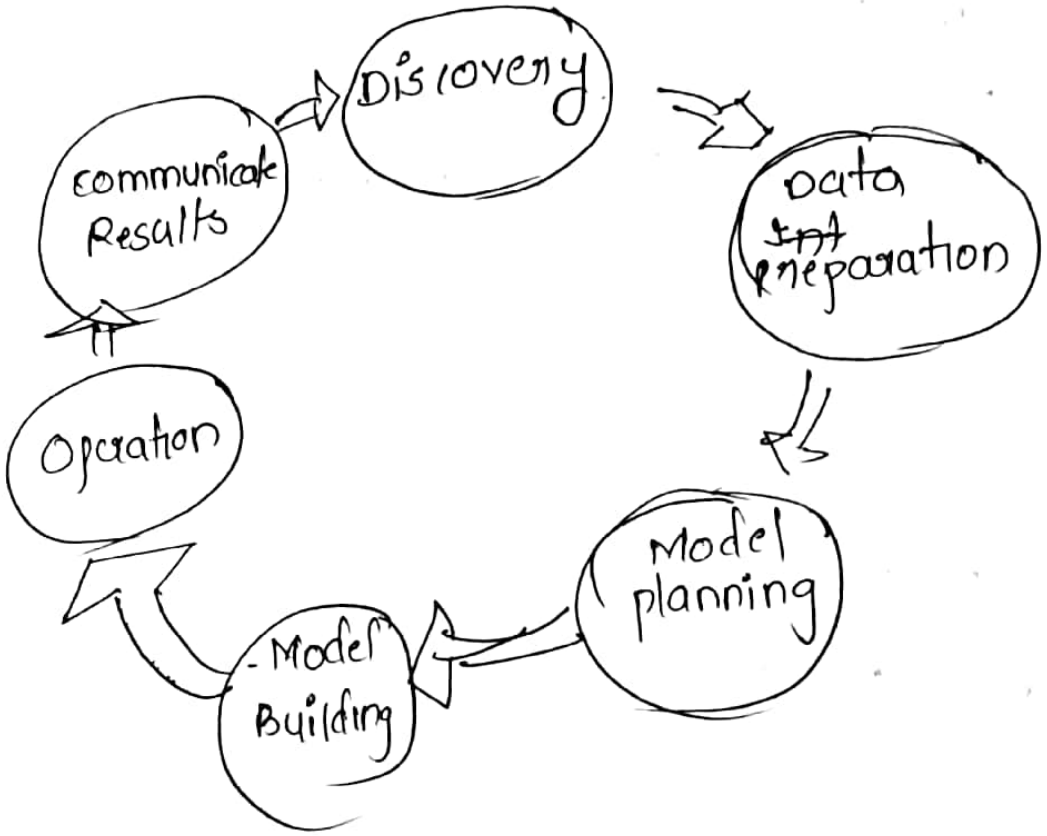
1. Frame the problem
2. collect the raw data needed for your problem.
3. process the data for analysis
4. explore the data

5. perform in-depth analysis
6. communicate results of the analysis.

As the datascience process stages help in converting raw data into monetary gains and overall profits, any data scientist should be well aware of the process and its significance.

Steps in Datascience process.

A datascience process can be more accurately understood through datascience online courses and certifications on data science.



Step 1: Framing the problem.

Before solving a problem, the pragmatic thing to do is to know what exactly the problem is. Data questions must be first translated to actionable business questions. people will more than often give ambiguous inputs on their issues. And, in this first step, you will have to learn to turn those inputs into actionable outputs.

A great way to go through this step is to ask questions like:

- Who the customers are?
- How to identify them?
- What is the sale process right now?
- Why are they interested in your products?
- What products they are interested in?

Step 2: Collecting the Raw Data for the Problem

After defining the problem, you will need to collect the requisite data to derive insights and turn the business problem into a.

probable solution. the process involves thinking through your data and finding ways to collect and get the data you need.

Step3: processing the Data to Analyze
After the first and second steps, when you have all the data you need, you will have to process it before going further and analyzing it

Step4: exploring the Data
In this step, you will have to develop ideas that can help identify hidden patterns and insights. You will have to find more interesting patterns in the data, such as why sales of a particular product or service have gone up or down

Step5: performing In-depth Analysis.
This step will test your mathematical, statistical and technological knowledge, you must use all the data science tools to crunch the data successfully and discover every insight you can.

Step 6: Communicating Results of this Analysis

After all these steps, it is vital to convey your insights and findings to the sales head and make them understand their importance. It will help if you communicate appropriately to solve the problem you have been given.

3. List out the applications of Eigen vectors and eigen values in Datoiscience?

Ans Applications of eigen ~~values~~ vector in ~~Data~~ science

If V is a vector that is not zero, than it is an eigen vector of a square matrix A if Av is a scalar multiple of v. this condition should be written as the equation:

$$Av = \lambda v$$

In the above equation λ is scale known as the eigenvalue

1. Communication systems

Eigenvalues were used by claude shannon to determine the theoretical limit to how much information can be transmitted through a communication medium like your telephone line or through the air. this is done by calculating the eigen vectors and eigen values of the communication channel and then water filling on the eigenvalues.

2. Designing bridges

The natural frequency of the bridge is the eigen value of smallest magnitude of a system that models the bridge.

3) Designing car stereo system

Eigen value analysis is also used in the design of the car stereo systems, where it helps to reproduce the ~~viola~~ vibration of the car due to the music

4) Electrical engineering

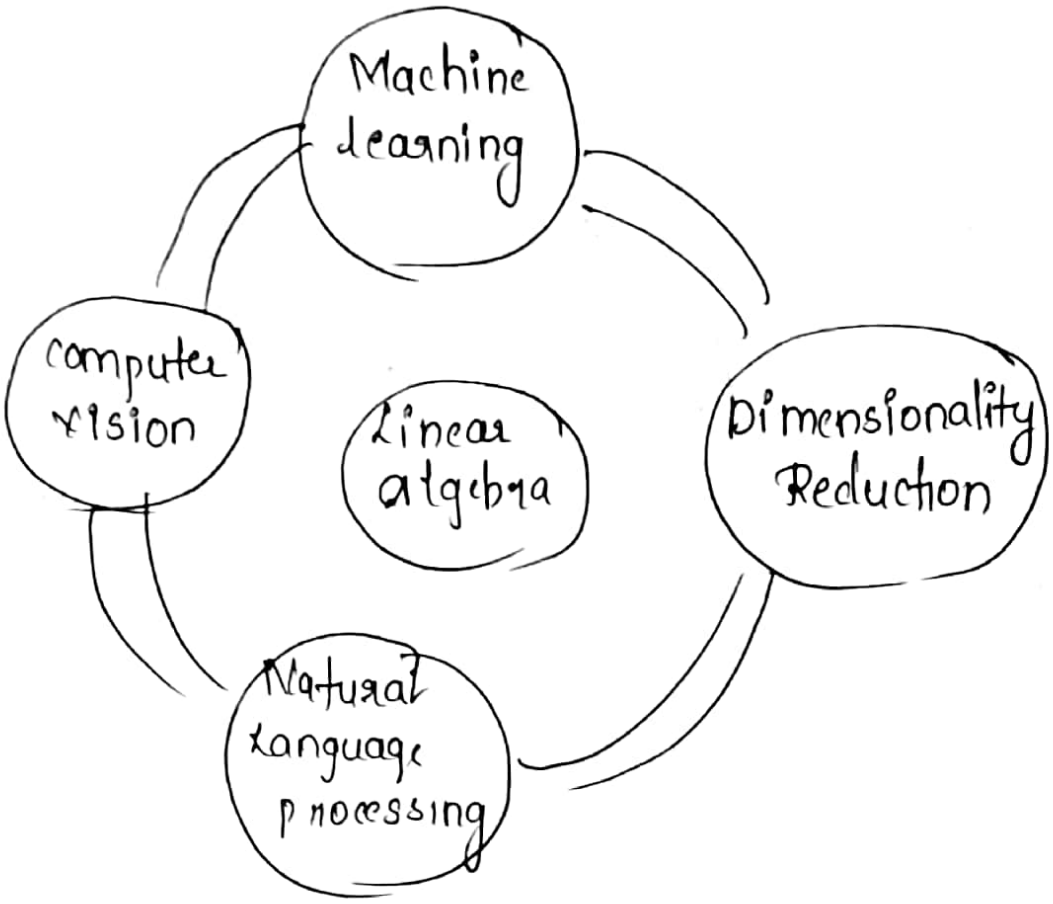
The application of eigenvalues and eigen vectors is useful for decoupling three-phase systems through symmetrical component transformation.

4) Why linear Algebra is significant in
(5) Datascience & How linear Algebra applied in Datascience.

Ans Linear-Algebra-A powerful tool for Datascience

Analysis of data is an important task in data managements systems. Many mathematical tools are used in data analysis. A new division of datamanagement has appeared in machine learning. linear algebra, an optimal tool to analyze and manipulate the data Datascience's a multidisciplinary subject that uses scientific methods to process the structured and unstructured data to extract the knowledge by applying suitable algorithms and systems.

Applications of Linear Algebra in Data Science



Linear Algebra in Machine Learning

1. Loss functions

=> You must be quite familiar with how a model, say a Linear Regression model, fits a given data.

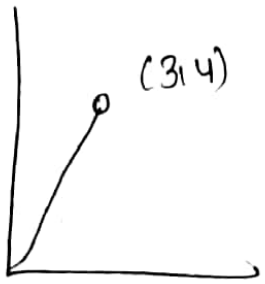
=> You start with some arbitary prediction functions

=> Use it on the independent features of the data to predict the output

=> calculate how far off the predicted output from the actual output

- L2 Norm: Also known as the Euclidean Distance. L2 Norm is the shortest distance of the vector from the origin as shown by the red path in the figure below:

Euclidean Distance or L2 Norm



L2 Norm of Vector $V = (v_1, v_2 \cdots v_n)$

$$||V||_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

2. Regularization

Regularization is a very important concept in data science. It's a technique we use to prevent models from overfitting

Regularization is actually another application of the Norm.

3. Covariance Matrix

Bivariate analysis is an important step in data exploration to study the relationship b/w pairs of variables

covariance or correlation is measures used to study relationships b/w ~~or~~ two continous variables.

⇒ use these calculated values to optimize your prediction function using some strategy like Gradient Descent

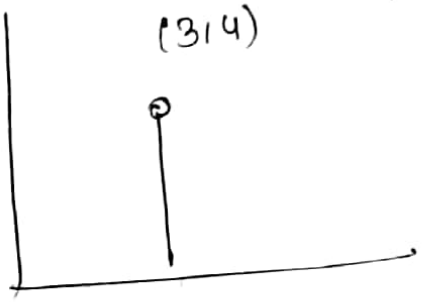
A loss function is an application of the vector Norm in Linear Algebra The Norm of a vector can simply be its magnitude There are many types of Vector norms I will quickly explain two of them:

- L1 Norm: Also known as the Manhattan Distance or Taxicab Norm. the L1 Norm is the distance you would travel if you went from the origin to the vector if the only permitted directions are parallel to the axes of the space.

Manhattan Distance or L1 Norm

L1 Norm of vector $V = (v_1, v_2, \ldots v_n)$

$$||V||_1 = |v_1| + |v_2| + \ldots + |v_n|$$



In this 2D space, you could reach the vector (3,4) by travelling 3 units along the y-axis and then 4 units parallel to the y-axis.

Support vector Machine classification

One of the most common classification algorithms that regularly produces impressive results. It is an application of the concept of vector spaces in linear Algebra.

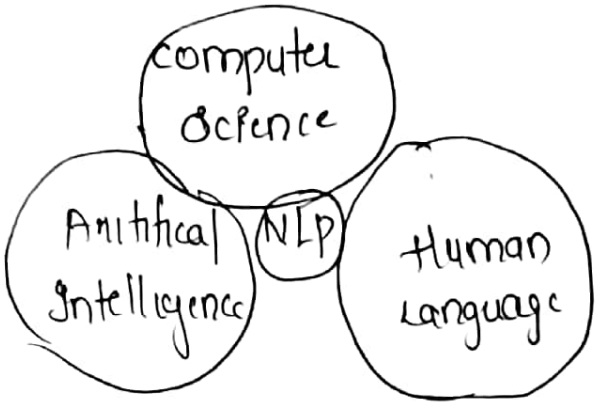
Support Vector Machine is a discriminative classifier that works by finding a decision surface. It is a supervised machine learning algorithm.

Dimensionality Reduction

You will often work with datasets that have hundreds and even thousands of variables That's just how the industry functions. Is it pratical to look at each variable and decide which one is more important? that doesn't really make sense. We need to bring down the number of variables to perform any sort of coherent analysis this is what dimensionality reduction is.

Natural Language processing (NLP)

NLP is a field of Artificial intelligence that gives the machine the ability to read, understand and derive meaning from human languages



Word Embeddings

Word Embeddings is a way of representing words as low dimensional vectors of numbers while preserving their context in the document These representations are obtained by training different neural networks on a large amount of text which is called a corpus.

Latent Semantic Analysis

Latent Semantic Analysis is one of the techniques of Topic Modelling. It is another application of Singular value Decomposition

Latent means 'hidden'. True to its name.

~~LSA att~~

computer vision

Another field of deep learning that is creating waves - computer vision.

Image Representation as tensors

A digital image is made up of small indivisible units called pixels

convolution and image processing

2D convolution is a very important operation in image processing. It consists of the below steps

1. Start with a small matrix of weights called a kernel
2. Slide this kernel on the 2D input data, performing element-wise multiplication
3. Add the obtained values and put the sum in a single output pixel.

6. Draw and explain the lifecycle of Datascience project?

A datascience life cycle is nothing but a repetitive set of steps that you need to take a complete and deliver a project to your client

Although the datascience projects and the teams involved in deploying and developing the model will be different, every datascience lifecycle will be slightly different in every other company

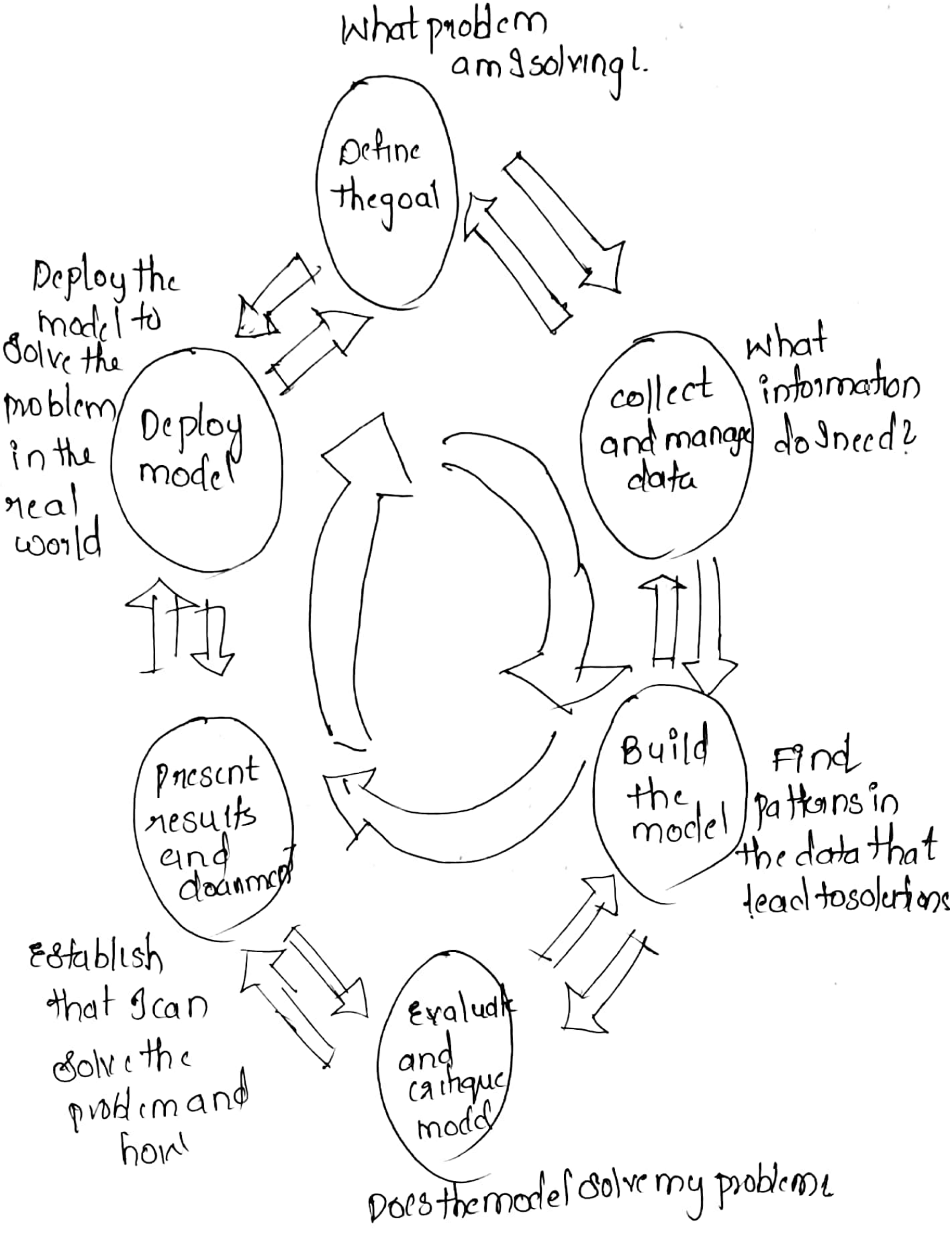
The ideal datascience environment is one that encourages feedback and iteration blw the datascientist and all other stakeholders.

This is reflected in the lifecycle of a datascience project

project

The lifecycle of a datascience project:

loops within loops



1. Defining the goal

The first task in data science project is to define a measurable and quantifiable goal. At this stage learn all that you can about the context of your project

- Why do the sponsors want the project in the first place? What do they lack, and what do they need?
- What are they doing to solve the problem now and why isn't that good enough?
- What resources you will you need: What kind of data and how much staff? Will you have domain experts to collaborate with and what are the computational resources?

Data collection and management

This stepencompasses identifying the data you need, exploring it, and conditioning it to be suitable for analysis This stage is often the most-time consuming step in the process. It's also one of the most important

- What data is available to me?
- Will it help me solve the problem?
- Is it enough?
- Is the data quality good enough?

Modeling

You finally get to statistics and machine learning during the modelling. Here is where you try to extract useful insights from the data in order to achieve your goals

Since many modeling procedures make specific assumptions about data distribution and relationships, there will be overlap and back-and-forth between the modeling stage and the data cleaning stage as you try to find the best way to represent the data

The most common datascience modeling tasks are these:

- classification
- Scoring
- Ranking
- clustering
- finding relations
- characterization

Model evaluation and critique

once you have a model, you need to determine if its meets your goals:

- Is it accurate enough for your needs? Does it generalize well?
- Does it perform better than "the obvious guess"? Better than whatever estimate you currently use?
- Do the results of the model make sense in the context of the problem domain?

Presentation and documentation

Once you have model that meets your success criteria, you'll present your results to your project sponsor and other stakeholders. You must also document the model for those in the organization who are responsible for using, running and maintaing the model once it has been deployed.

Model deployment and maintainence

Finally, the model is put into ~~actior~~ operation In many ~~ope~~ organizations this means the data scientist no longer has primary responsibility for the day-to-day operation of the model. But you still should ensure that the model will run smoothly and won't make disastrous unsupervised decisions. You also want to make sure