The data science process is a systematic approach to solving a data-related problem which involves identifying a problem, collecting data and building a model that helps to solve the problem.

It consists of 7 steps:

1. Defining the problem: Understanding what the problem is, what questions you have, and what answers you need. you want to achieve by building a model.
2. Collecting data: In order to answer your problem statement, you need to collect relevant information.
3. Preparing the data: The data that you collect might contain some irrelevant or duplicated information, it might not be in the right format. You need to clean it and structure it.
4. Exploring the data: To understand the material you are working with. Identify potential relationships between features, patterns in the data. This not only helps you to better understand the material you are working with, it also helps you decide on the most suitable model to analyse the data.
5. Building the model: There are different types of statistical and machine learning algorithms that can be used to answer different questions/ solve different problems. For example, you might use regression models to predict house prices, or classification models to predict if an image is a benign or cancerous tumour. Based on the problem statement and the data that you have, you will choose a suitable model. This model will be trained on the available data and identify patterns, so that it is able to make predictions when presented with new data.
6. Evaluating the model: Once the model is trained, you will want to know how good the model is, for example, how precise it is. evaluate how well it makes predictions based on suitable evaluation metrics, for example, precision.
7. Deploying the model: If the model is good, you will put it into production, monitor and maintain it. Make updates as needed to improve accuracy.
8. Communicating results: For a model to be useful, its findings should be communicated in a clear and engaging way to stakeholders, so it can help decision-making.

I’ve brought an example to illustrate the data science process and hopefully make it more engaging. Since I’m a big fan of cartoons and comics, we’re going to use superheroes to bring the data science process to life.

Here I have a deck of Top Trump cards. You may be familiar with the game top trumps; to be clear, we are not going to be playing the game. We are using the cards because they are a useful way of visualising our data - the superheroes and their abilities.

So let’s begin. For our example, I would like you to imagine that you exist in the Marvel universe and you have a superhero agency, so you manage different heroes. Your agency is pretty successful and you manage many heroes - represented by this deck of cards. You collect information about your heroes - their abilities, which villains they have fought against, won and lost against.

Now imagine it’s a very busy day at the agency - all your heroes have been sent off to fight villains. Suddenly you get an emergency call that the villain Loki has appeared in Munich. You need to send a hero to defend the city, but all your superheroes are busy.

The only available heroes are 3 rookies you recently signed. They are not experienced in battle and they have never fought Loki. You don’t want to put them in unnecessary danger, but you need to send someone, so who do you send?

To help you make a decision, you decide to go through the data science process and train a model that will tell you which rookie you should send.

1. Define the problem: Which rookie should I send to fight Loki? Which rookie is likely to beat Loki in a fight?
2. You have data on all the established superheroes at the agency - their top trumps characteristics and their records of winning or losing against loki.
3. You only want to train your model on relevant data, so you exclude the year when they made their debut
4. EDA? 95% of heroes have strength between 25 and 35, so if above that, it’s a particularly strong hero
5. Based on your problem, you want a model that tells you if a hero is predicted to win against Loki or lose, so a good choice is a decision tree. It’s beyond the scope of this 10 min presentation to discuss in detail the maths behind the decision tree, it’s enough to know that it studies the features and the labels, and learns to predict which features are strong predictors for winning or losing against Loki. You train your model on 90% the heroes and this is how it looks…
6. On the remaining 10%, you evaluate your model. It has a 100% accuracy on the test set, so you’re pretty confident that it’s a model that will make good predictions.
7. Now you deploy the model by using it to decide which rookie to send… Nova

* [CRISP-DM](https://www.datascience-pm.com/crisp-dm-2/) stands for Cross Industry Standard Process for Data Mining. It’s an industry-standard methodology and process model that’s popular because it’s flexible and customizable. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
* Unlike CRISP-DM, [OSEMN](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584986) is not an iterative process. Since it omits business-oriented phases, it’s ideal for projects focusing on exploratory research, and is often used by research institutions and public health organizations. With OSEMN, you’re more interested in what the data have to say, as opposed to asking specific questions. Obtain Data, Scrub Data, Explore Data, Model Data, Interpret Results

<https://www.springboard.com/blog/data-science/data-science-process/#:~:text=The%20data%20science%20process%20is,presenting%20the%20solution%20to%20stakeholders>.

<https://www.geeksforgeeks.org/data-science-process/>