# What is Data Science?

A Simple Introduction for the Art Institute by
Rahul Rajeev

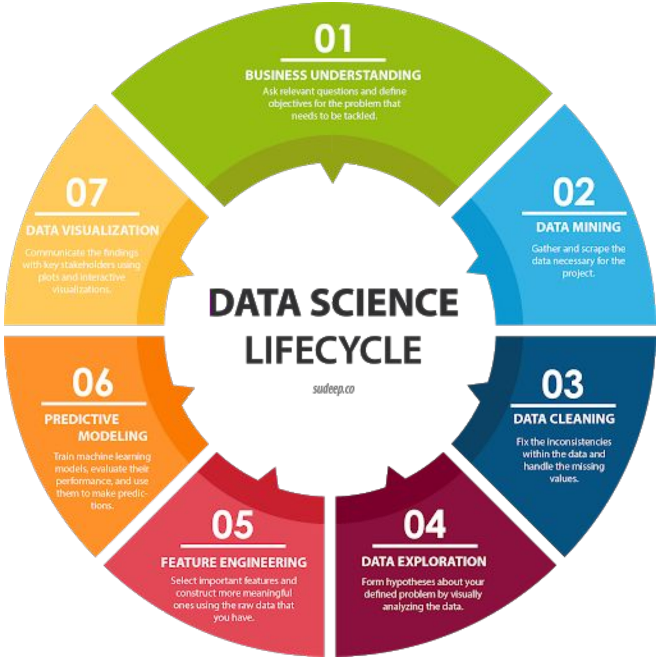
# Defining Data Science

- The study of trends and relationships between different sources of information
- Examples:
  - Geographic areas of racism
  - Predicting the results of an election
  - Do jokes increase happiness?
  - Popularity of certain sports teams
- Each of these examples take one source of information and compares it with another, seeking a certain relationship

# Data Science Life Cycle

The study of data science typically follows a number of steps from business understanding to data visualization. Each step is crucial towards a research project's success.

Some steps require more time than others and some require the use of programming while others don't. The main idea is that they follow this order in order to maintain integrity, but there are always opportunities to travel back and forth, should issues arise.



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

**Figure 1: The Data Science Life Cycle**
Credits: sudeep.co

# Business Understanding

The first and foremost step is to understand the trend you are trying to discover or a certain problem you are trying to solve. It's just like trying to picture the image you are going to replicate with art. Without a general idea of what you are trying to paint, draw, or illustrate, it's impossible to move forward.

Data science is typically used to answer the following questions:

1. How much or how many? (predicting future sales based on past art pieces)
2. Which category? (categorizing a painting based on its style)
3. Which group? (seeing whether the painting fits within an existing style)
4. Does this make sense? (looking for art that doesn't belong)
5. Which option should be taken? (recommended art styles)

This is where you should identify the central objectives of the project you are going to start.

# Data Mining

The next step is to gather the data required for your study. You can find data almost anywhere, on websites, in databases (places that store data), surveys, etc.

These are the collection methods data scientists use to collect data:

- Data querying - collecting data from a database using queries
    - Queries are a very formal way of describing the process of asking the database for specific sets of data, similar to how you do an art piece search based on the artist, the style, and the time period.
- Scraping - sifting through web pages to grab certain statistics
- Tracking - finding user engagements and interactions with a mobile app
    - Google Analytics sets up events and prompts that helps track how users engage with features on the app

# Data Cleaning

Following data collection, we should filter out the inconsistencies that always appear. It's like doing janitor work on data.

Data cleaning is usually the step in the data science process that takes “50% to 80% of the time” (Sudeep Agarwal).

Data cleaning looks out for the following inconsistencies in a table of data with rows and columns.

- Labeling within the same column (description of painting vs paint, which are the same, but will be considered different in the dataset)
- Types of data (for a certain category, table entries might be numbers, and others might be words or “strings”)
- Misspellings of certain categories (male vs. Male, romanticism vs romanticim)
- Missing data (we searched for a certain quantity and it wasn't recorded across a couple of the sources of information)

# Data Exploration

With a squeaky clean data set, although a data set with absolutely no inconsistencies is very hard to achieve, you can finally move on to predicting trends you might find in the data. This is the brainstorming of the actual analysis of the data.

A preliminary assessment of trends can be made here first, by taking a small portion of your data set. This small portion of data is called a subset which can be used to plot a graph or form an interactive visualization to explore the trend.

The predictions can be then used to form hypothesis, a prediction based on the data and the problem you are looking to solve.

For example, if you are trying to predict an art piece's price, you can plot prices of previous art pieces with a similar style against the release date and see if you can find a trend.

# Feature Engineering

In data science, a feature is a measurable property of the observed event. If we were predicting the quality of an art piece, a number of features we should analyze are the artist's hours of sleep or their recognition in the most recent years.

These are the two typical tasks of feature engineering:

- Feature selection - cutting down the numbers of properties we are analyzing, too many properties will make data analysis complicated
- Feature construction - creating new features or merging existing features
  - If we have multiple entries of the artist's hours of sleep for each day of the week, we can reduce it to an average hours of sleep per week.

# Predictive Modeling

This is where your data and feature engineering finally come into play. Predictive modeling of your data set requires machine learning, the process of teaching your computer to recognize certain trends in data given a subset which we call a training set.

It’s similar to how when we search about a certain topic on Youtube, eventually the videos we are interested in pop up on your homepage. All it requires is a subset of videos you watch (for example your subscriptions), and continuous training from watching specific videos from your subscription list and the videos you search for.

# Data Visualization

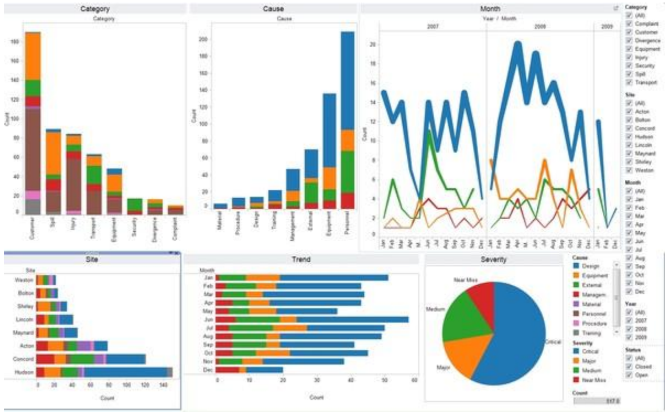


**Figure 2: Examples of Good Data Visualization**
Credits: Kristi Pelzel, Medium

This is where you display the results of your predictive modeling. The final communication of your data in a simple yet aesthetically pleasing way. You want to present the results to the stakeholders of the company or to demonstrate that you have proven a certain trend or predicted a certain value.

Although the data cycle seems to conclude here, this is actually the start of the next cycle, going back into business understanding.

Evaluate the success of your data science process and see where you can maybe improve it.

# Questions?

Understanding the Data Science Life Cycle
Good Data Visualization Examples