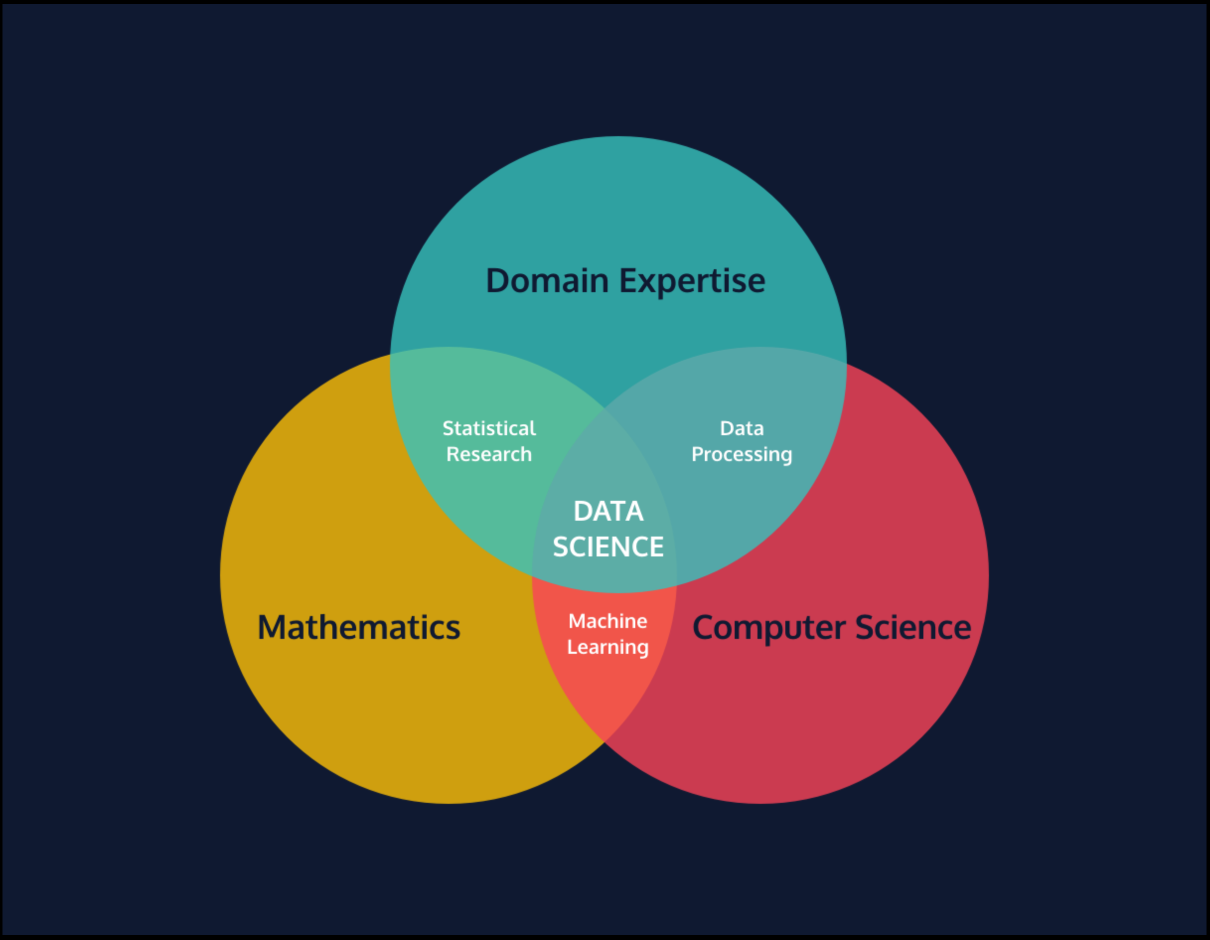
Data Scientist
Codecademy
Getting Started with Data Science

Introduction

Our world contains massive amounts of data. Data is the building blocks for information. And information can carry meaning – from a click telling us what someone likes, to toxins in the water signaling a health concern. But data as a collection of records is meaningless unless we do something with it. That’s where data science comes in.

Data Science enables us to take data and transform it into meaningful information that can help us make decisions. Data science is interdisciplinary and combines other well-known fields such as probability, statistics, analytics, and computer science.



Statistics

A fundamental part of data science is statistics.

Statistics is the practice of applying mathematical calculations to sets of data to derive meaning. Statistics can give us a quick summary of a dataset, such as the average amount or how consistent a dataset is.

There are two types of statistics:

- Descriptive statistics
  - describe a dataset using mathematically calculated values, such as the mean and standard deviation.
- Inferential statistics
  - statistical calculations that enable us to draw conclusions about the larger populations.

Probability

Probability is the mathematical study of what could potentially happen.

In Data Science, probability calculations are used to build models. Models are able to help us understand data that has yet to exist – either data we hadn’t previously collected or data that has yet to be created. Data scientists create models to help calculate the probability of a certain action and then they use that probability to make informed decisions.

Programming

Programming is an essential part of data science apart from similar fields, like data analytics.

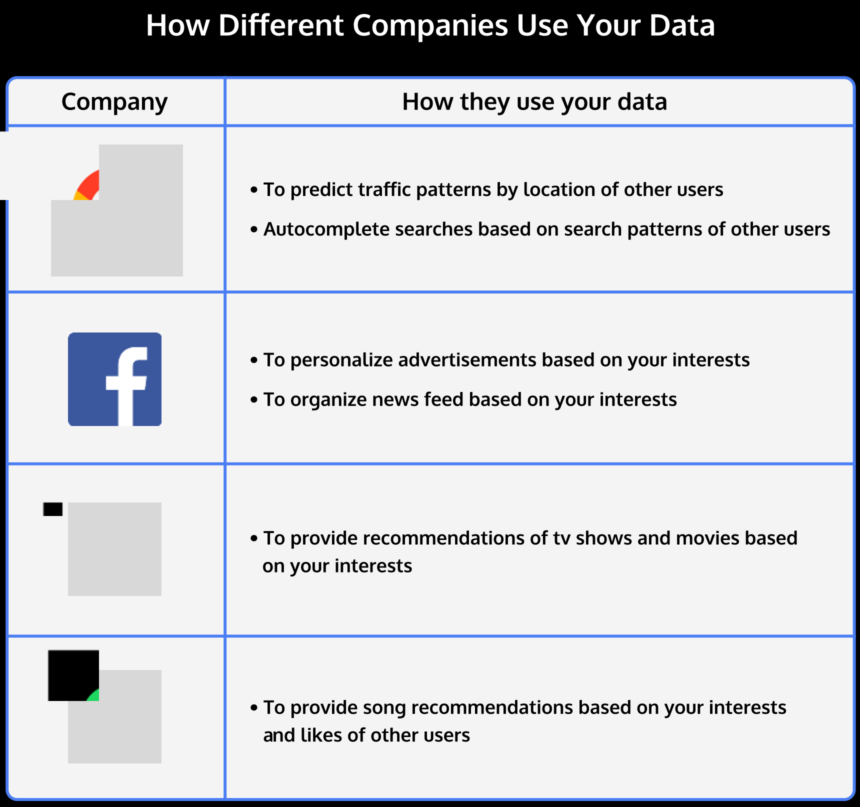
Programming is the practice of writing commands for a computer to execute. Computer science is the discipline of interacting with computation systems.

A computer program is a series of instructions that tells the computer to perform a certain task.

In Data science, programming allows us to hand the processing power over to the computers. Given the right commands, computers can process millions of data points in a matter of seconds. Within data science, programs will allow you to reproduce experiments by simply running the program again.

Domain expertise

Domain expertise refers to the particular set of knowledge that someone cultivates in order to understand their data. You may be able to crunch big numbers, but in order to understand their meaning, you’re going to need a lot of context. This context could come from research, teammates, or the knowledge gained from working in a particular industry.



How Different Companies Use Your Data

Company	How they use your data
	• To predict traffic patterns by location of other users • Autocomplete searches based on search patterns of other users
	• To personalize advertisements based on your interests • To organize news feed based on your interests
	• To provide recommendations of tv shows and movies based on your interests
	• To provide song recommendations based on your interests and likes of other users

Figure 1:google facebook netfilx spotify

Why is it called Data Science?

Professionals who do data science are driven by desire to ask questions. To answer these questions, data scientists use a process of experimentation and exploration to find answers. Like other scientific fields, data science follows the scientific method. But data science process also include steps particular to working with large digital datasets.

The process generally goes as follows:

- ask a question
- determine the necessary data

- get the data
- clean and organize the data
- explore the data
- model the data
- communicate your findings

Formulate a question

Data science is all about asking questions and using data to get answers.

While coming up with a question may seem relatively simple, coming up with a good question requires practice.

- Variable relationships
    - In data science, we want to know the effect that different things have on each other. One way is to think of the relationship between the different variables, like how x is related to y?
- Scope
    - Another thing to keep in mind while forming a question is scope. A question should be specific enough that we know it is answerable, but it shouldn’t be too specific to the point where no relevant data exist, and we are unable to draw any real conclusions.
- Context
    - Another thing to keep in mind is that a good question requires context. Gaining context requires doing research, such as looking at any relevant data that you already have.
    - For working on large database run SQL queries to find data. Professionals often use third party software such as Azure and Tableau to run their queries.

Determine the necessary data

After having a question, we make an educated guess about what the answer might be. This educated guess is known as hypothesis and helps to determine the data you need to collect.

It’s important to remember that the data collected can’t just be the data that we think is going to answer our question. We can’t carefully select for what we know will prove the hypothesis is correct – actually, the data needs to disprove the hypothesis.

???

In science, it’s actually impossible to prove that something is true. Rather, we try and show that we’re really, really confident that it’s not false. That’s because the only way we can say we’re 100% positive our hypothesis is correct is to collect all the data for an entire population – and that’s pretty much impossible!

To determine what data is necessary to collect:

- First we need to determine what data could disprove our hypothesis.
- Next we need to figure out how much data to collect.
  Since it is impractical to collect information for an entire population, we collect a sample set of data, a smaller amount of data that are representative of the entire population.
  How do we ensure that our sample is representative? We figure out the necessary number of samples that have similar descriptive statistics to the entire population.

Rule of thumb: the larger the sample size and the more diverse our dataset is, the more confident we will be in our results. We don’t want to go through the trouble of designing and running an experiment only to discover that we don’t have enough information to make a good decision!

Clean the Data

Data is typically organized in columns and rows. But raw data can actually come in a variety of file types and formats. This is especially true if we are getting data from elsewhere, like public data sets.

We as humans may be able to understand the organizing logic of a dataset, but computers are very literal. A missing value or unlabeled column can completely throw them off their game. Even worse – the program could still run, but the outcomes would be incorrect.

An important part of data science is to clean and organize the datasets, sometimes referred to as data wrangling. Processing a dataset could mean a few different things.

The Python library Pandas is great tool for importing and orgainizing datasets. We can use Pandas to convert a spreadsheet document, like a CSV, into easily readable tables and charts known as DataFrames. We can also use libraries like Pandas to transform our datasets by adding columns and rows to an existing table, or by merging multiple tables together!

Explore the Data

Exploring a dataset is a key step because it will help us quickly understand the data we're working with and allow us to determine if we need to make any changes before moving forward. Changes could include some additional dataset cleaning, collecting more data, or even modifying the initial question we're trying to answer.

There are two strategies to exploring data:

- Statistical calculations
  - When we first get a dataset, we can use descriptive statistics to get a sense of what it contains. Descriptive statistics summarize a given dataset using statistical calculations, such as the average, median, and standard deviation. We can immediately learn what are common values in our dataset and how spread out the dataset is.
  - We can use a Python module known as NumPy to calculate descriptive statistics values. NumPy (Numerical Python) supplies short commands to easily perform statistical calculations, like np.mean(), which calculates the mean of a dataset.
- Data visualization
  - The practice of data visualization enables us to see patterns, relationships, and outliers, and how they relate to the entire dataset. Visualizations are particularly useful when working with large amounts of data.
  - Python data visualization libraries like Matplotlib and Seaborn can display distributions and statistical summaries for easy comparison. The JS library D3 enables the creation of interactive data visualizations, which are useful for modeling different scenarios.

Modeling and Analysis

Models are abstractions of reality, informed by real data, that allow us to understand situations and make guesses about how things might change given different variables.

A model gives us the relationship between two or more variables. For example, we can build a model that relates the number of grade school children in a neighborhood with sales of toys, or a model that connects the amount of trucks that travel certain roads with the amount of a city's budget assigned to road maintenance.

Models allow us to analyze our data because once we begin to understand the relationships between variables, we can make inferences about certain circumstances. Why is it that the sales of toy increases as the number of grade school children grows? Well, maybe it’s because parents are buying their children more toys.

Models are also useful for informing decisions, since they can be used to predict unknowns. Once we understand a relationship between variables, we can introduce unknown variables and calculate different possibilities. If the number of children in the neighborhood increases, what do we predict will happen to the sales of toys?

As we collect more data, our model can change, allowing us to draw new insights and get a better idea of what might happen in different circumstances. Our models can tell us whether or not an observed variance is within reason, is due to an error, or possibly carries significance. How would our understanding of our model change if in 2016 we discovered that the number of toys did not increase, but instead, decreased? We’d want to look for an explanation as to why this year did not fit the trend.

Models can be expressed as mathematical equations, such as the equation for a line. We can use data visualization libraries like Matplotlib and Seaborn to visualize relationships. If we pursue machine learning, we can use the Python package scikit-learn to build predictive models, such as linear regressions.

Communicating Findings

After we have done analyses, built models, and gotten some results, its time to communicate the findings.

Two important parts of communicating data are visualizing and storytelling.

- Visualizing
- Storytelling
    - It is an important part of data science because it gives meaning to the data. Contextualizing your findings and giving them a narrative draws people into our work enables us to convince them as to why they should make a certain decision.

Reproducibility and Automation

After finishing experiments and communicating our findings, it’s also important to remember that the scientific method requires that our work can be reproduced. In the field of data science, we call this quality reproducibility.

Reproducibility is important because it enables us to reuse and modify experiments, but it is also how the scientific community can confirm our analysis is valid. If our study produces results that no one can reproduce, it is likely that our results are invalid and the product of bias or error.

Another concept tied to reproducibility is the idea of automation. If we are creating reports, it’s most likely that we will be processing the same data at regular intervals. Rather than writing a new program each time, we can write a program that automats these processes, freeing up our time to do even more data science! Automation may be as simple as writing a Python program we can re-use all the way up to building machine learning models.

Data Science Applications

Some common applications of data science are:

- Reports
- Recommender systems
- Dynamic pricing
- Natural Language Processing

Similar to how data science is made up different disciplines, the applications of data science are far ranging. But in all of these situations, we will apply data science to find patterns, draw meaningful conclusions, and make decisions.

Reports

Reports are the most fundamental application of data science. A report is a document in which we present our process and findings. Reports are important because they enable those who work with data to translate numbers and calculations into accessible insights and recommendations for team members.

A report could take the form of a publication that circulated within our company, or an article that's published on the internet, or as part of a conference presentation.

A good report should have the following characteristics:

- Simple: non-technical team members will need to be able to quickly grasp findings
- Clear: language should be to the point
- Engaging Presentation: charts and presentations should be well designed

Five-Thirty-Eight is a popular statistics website that publishes reports on a range of topics, such as politics and sports.

Recommendation Engines

One of the more well-known applications of data science is using data and machine learning to build a recommender system, also known as a recommendation engine.

A recommender system is a type of content filtration system that seeks to predict what a user would be interested in consuming. These suggestions could come from the user's preferences that they've shared with the platform, like how Amazon suggests things we might want to buy based off of previous purchases. Other recommender systems work by looking at the preferences of people in your network, or those of people with similar demographics.

Recommender systems work for all types of information, from Spotify using it to recommend new artists to Netflix predicting what will be our next binge fest.

Dynamic Pricing

Another data science application is dynamic pricing. Ever go to buy an airplane ticket but then 5 minutes later the pricing changes? You have dynamic pricing to thanks for that.

Dynamic pricing, also known as surge pricing, is the practice of setting prices for products or services based on market demand. Companies that use dynamic pricing build algorithms that take into account competition, supply and demand, as well as other factors related to the specifics of the industry. Dynamic pricing exists across several industries, including transportation, entertainment, amusement parks, and professional sports.

The most common example is airline tickets. Airlines started to use computers to determine flight prices as early as the 1950s, taking into account the season, day of the week, and time of day when setting ticket prices. However, airlines have recently come under scrutiny for utilizing more robust dynamic pricing techniques. Several are now determining fares based on the buyer.

While dynamic pricing is increasingly common, it’s a tactic that is often seen to benefit the company more than the customer.

Natural Language Processing

Data science is also helpful in recognizing trends and patterns in language.

NLP is the application of programming and artificial intelligence to process and analyze text. It can be used for research purposes, to understand, for example, the grammatical structure of a text, to more creative pursuits such as generative poetry.

In NLP, our datasets are made up of examples of language usage, known as a corpus. After training on this dataset, a machine can then perform different functions, such as part of speech tagging, language translation, and sentiment analysis.

A common application of NLP is chatbots. While many chatbots follow predetermined scripts, more advanced ones can use NLP to enable a dynamic discourse. NLP enables a bot to continue learning as it talks, making it better at handling different and unexpected situations.