Intro to Data Science

Matt Speck, Data Science Instructor at General Assembly

About Me

- My name is Matt Speck and I am a co-instructor in General Assembly's Data Science Immersive Program. Before coming to General Assembly as an instructor, I was a student in the course. Before that I worked in Bolivia as an intern with a non-profit called Engineers Without Borders (EWB). From our offices in the city of La Paz, I traveled through Bolivia, doing monitoring and evaluation of ongoing infrastructure projects in rural areas. Before EWB, I studied physics at Fordham University. I realized in college that I wanted to develop technical skills that could be applied to a variety of problems in a variety of careers. Data science was the perfect fit!

GA

Objectives

- Learn buzzwords like ‘data science’ and ‘machine learning'
- Understand how data is used to solve problems
- Work through a data science problem together to understand the standard workflow
- Figure out where to start learning data science

What is a data science problem?

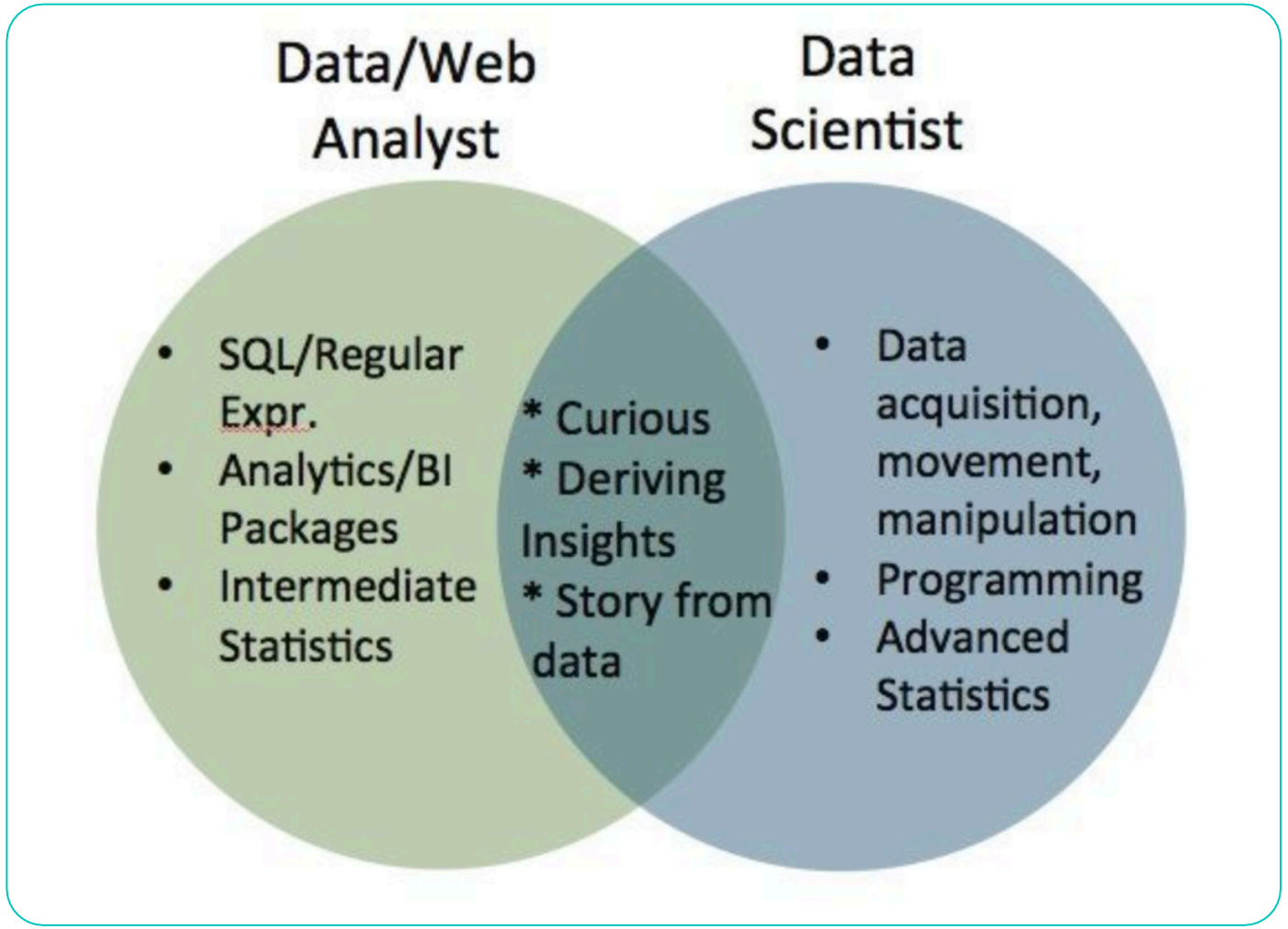
- The first thing we'll do is follow this link: https://quickdraw.withgoogle.com/#

What is a data science problem?

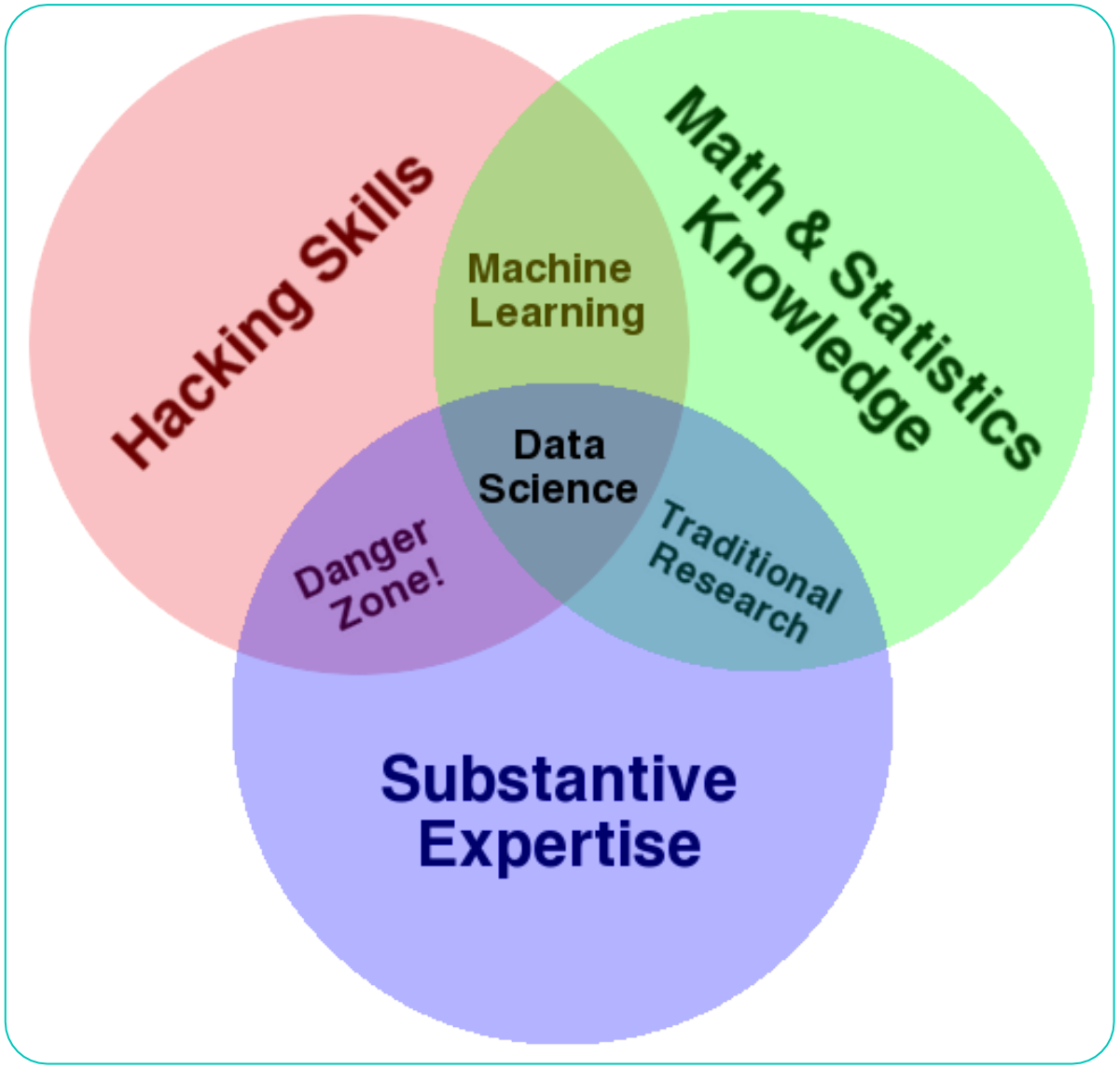
- What was the ‘data’ in Google Quick Draw?

What is a data science problem?

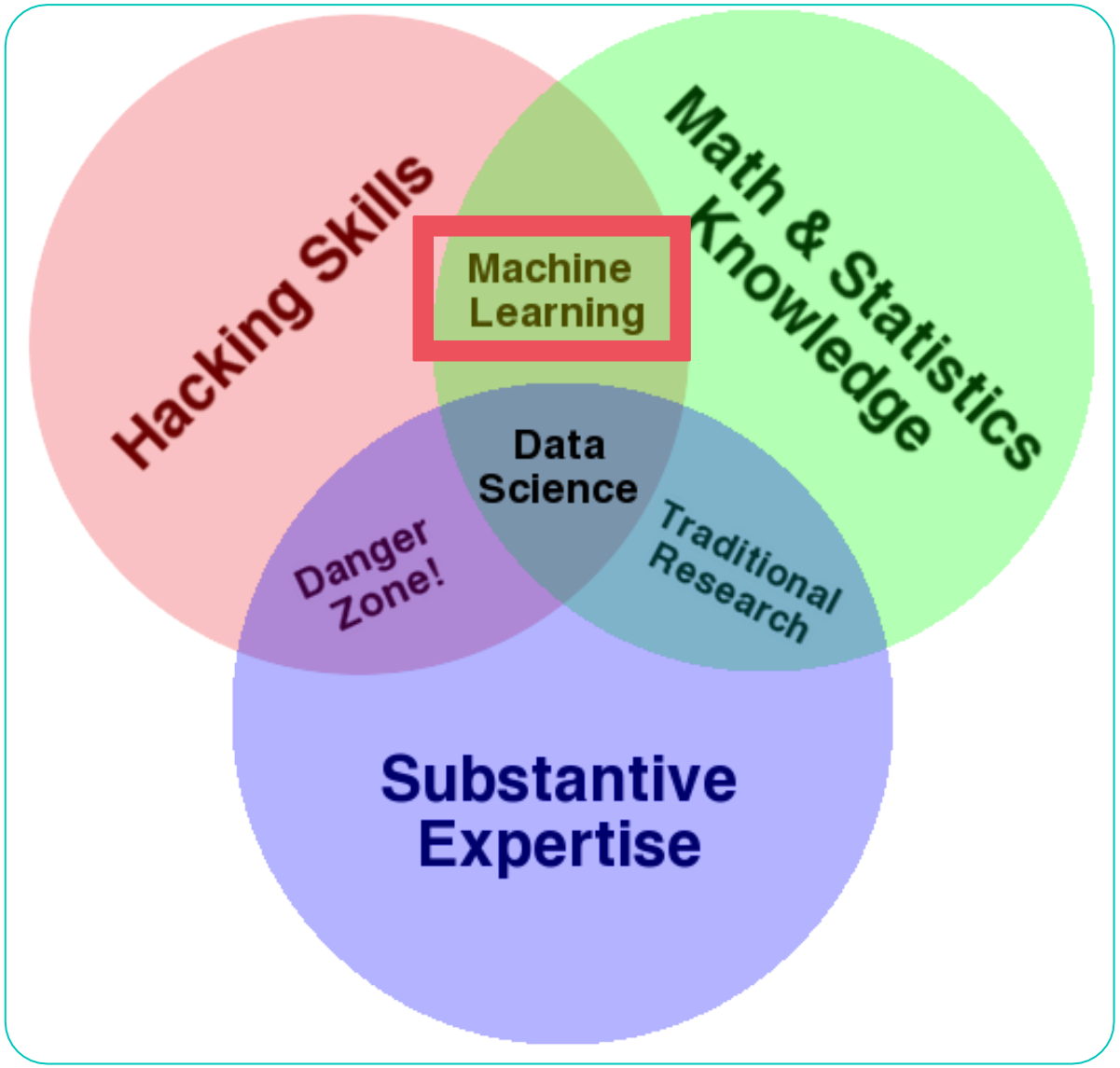
- What was the ‘data’ in Google Quick Draw?
- What was the ’science?’



Data Science vs. Data Analysis



What is Data Science?



What is Data Science?

Our Data Science Problem

- Situation: You are the new head of the data science team for senatorial candidate running in 2018. You're in charge of the entire data science process for a campaign that is just starting off.

- We're going to walk through how we might apply data science in a political situation like this.

Step 1: Identify the Problem

- What might be a / some problem statement(s) that you might be tasked with solving? In other words, how might we be able to use data in order to help the campaign?

Step 2: Acquire the Data

- What data do we need?
- How might we collect the data?

Step 2: Acquire the Data

- How might we collect the data?

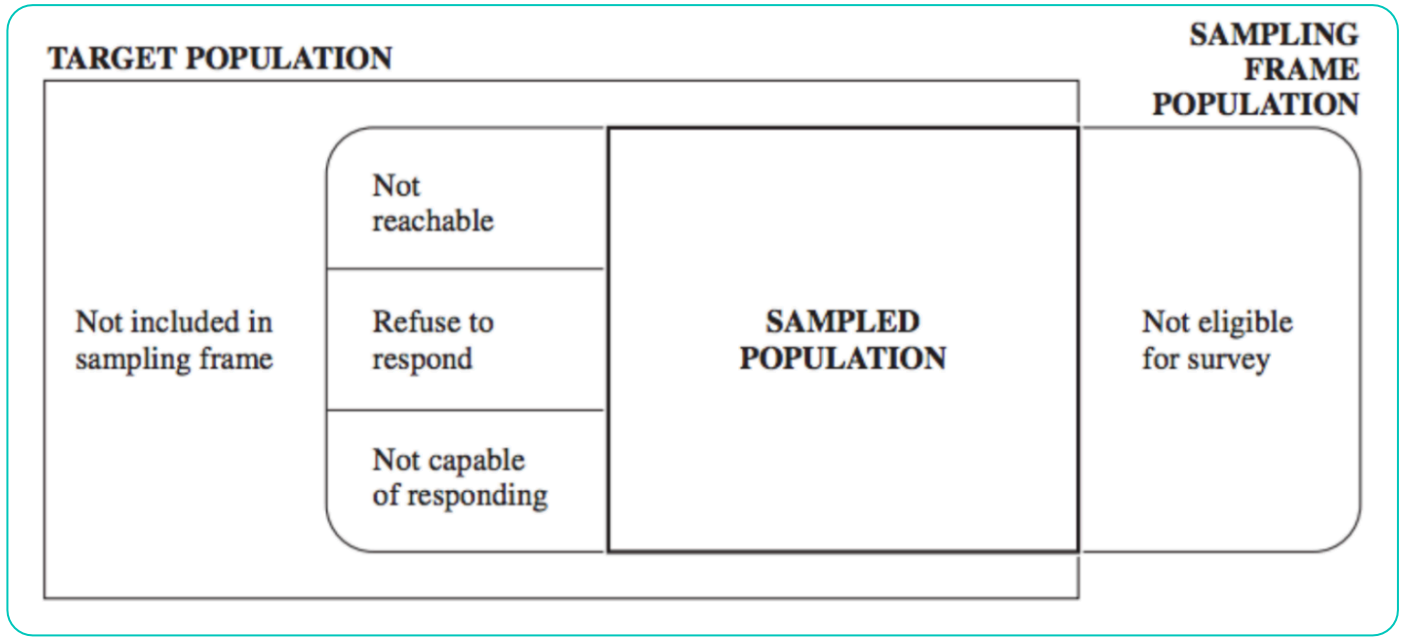
Step 2: Acquire the Data

- How might we collect the data?
  - Phone surveys
  - Canvassing
  - Online surveys

Step 2: Acquire the Data

- How might we collect the data?
  - Phone surveys
  - Canvassing
  - Online surveys
- What might be some limitations to these methods?

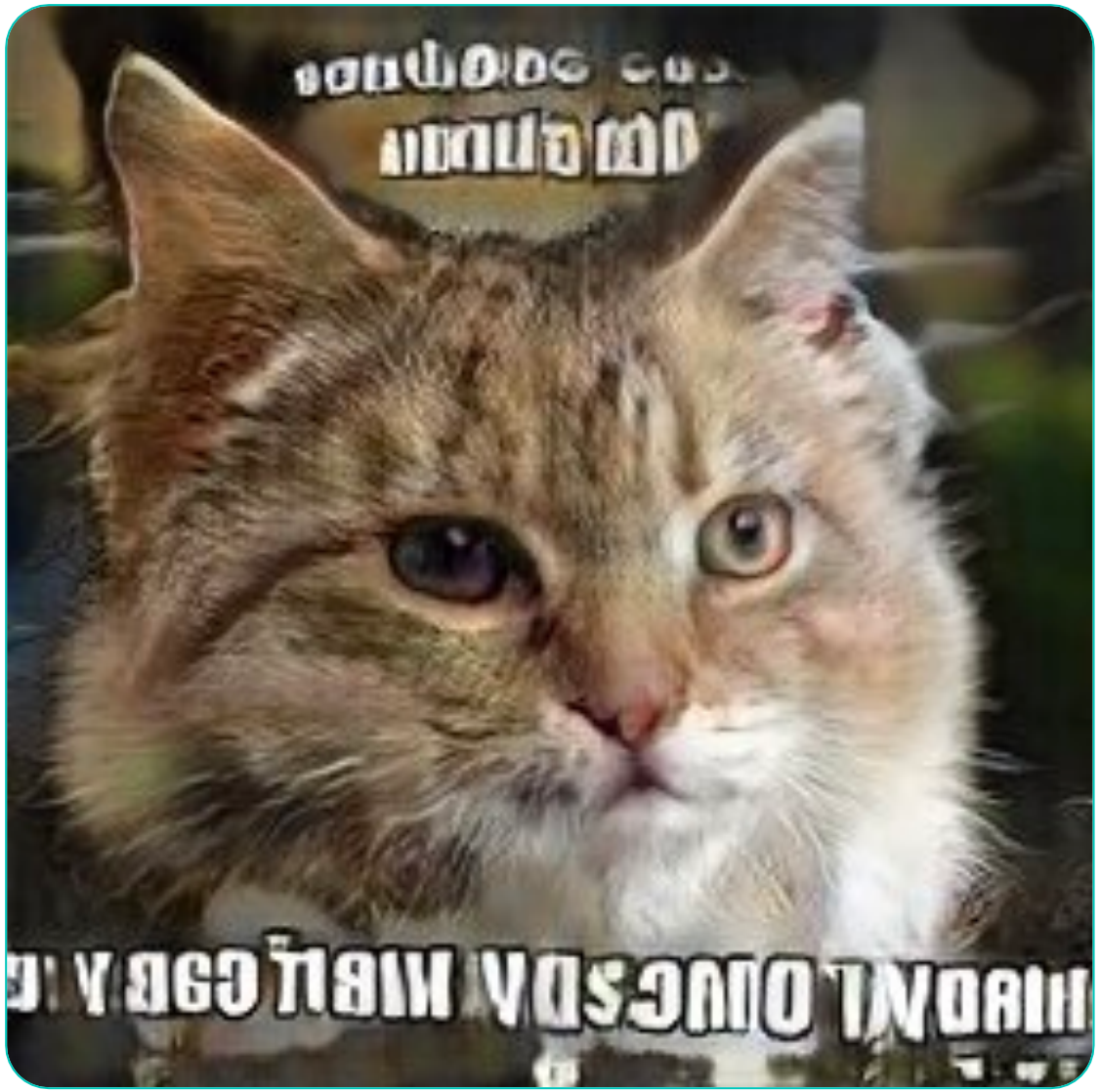
Sample vs. Population



- Target Population: What we're interested in
- Sample population: What we're able to get

Step 2: Acquire the Data

- Takeaway: We want the data to be **representative**. If our data doesn't represent what we're looking for, we might get weird / incorrect results.



Step 3: Explore the Data

- Look at distributions
- Verify data integrity

Step 3: Explore the Data

- Look at distributions
- Verify data integrity
  - Null values

Step 3: Explore the Data

- Look at distributions
- Verify data integrity
  - Null values
  - Incorrect values

Step 3: Explore the Data

- Look at distributions
- Verify data integrity
  - Null values
  - Incorrect values

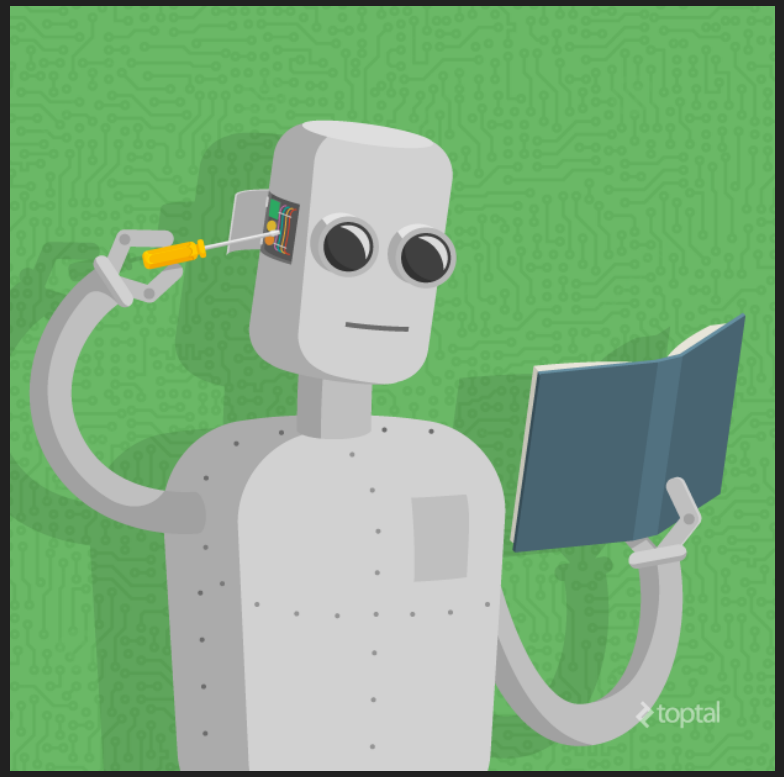
- This is often the lengthiest task in a data science project!

Step 4: Refine the Data

- Feature engineering: Making new predictors out of current ones
- Changing the format of your data

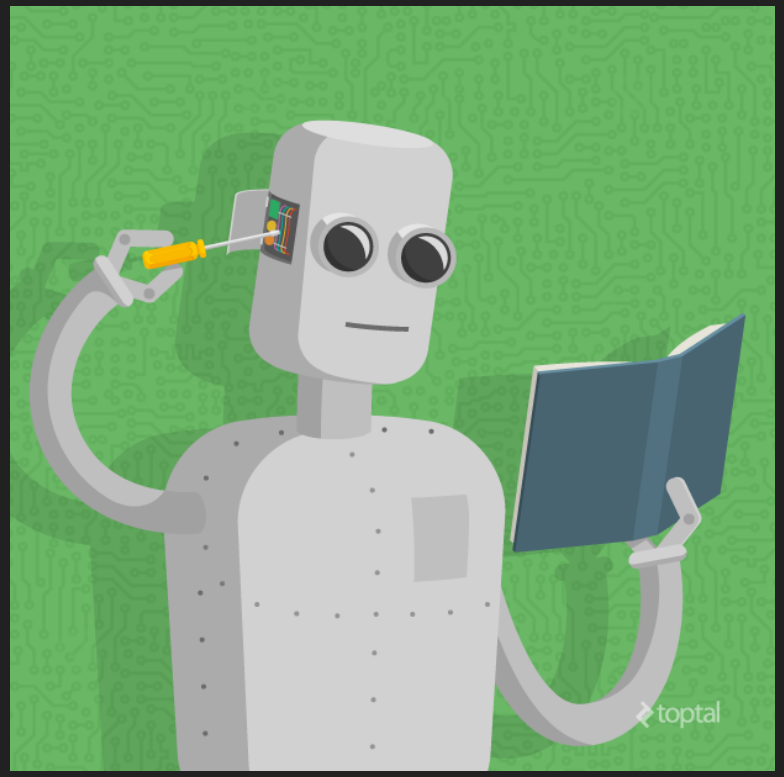
Step 5: Build the Model

- This is where ‘machine learning’ comes in.
- Machine learning: ‘a field of computer science that gives computers the ability to learn without being explicitly programmed.’ [Wikipedia, Arthur Samuel]



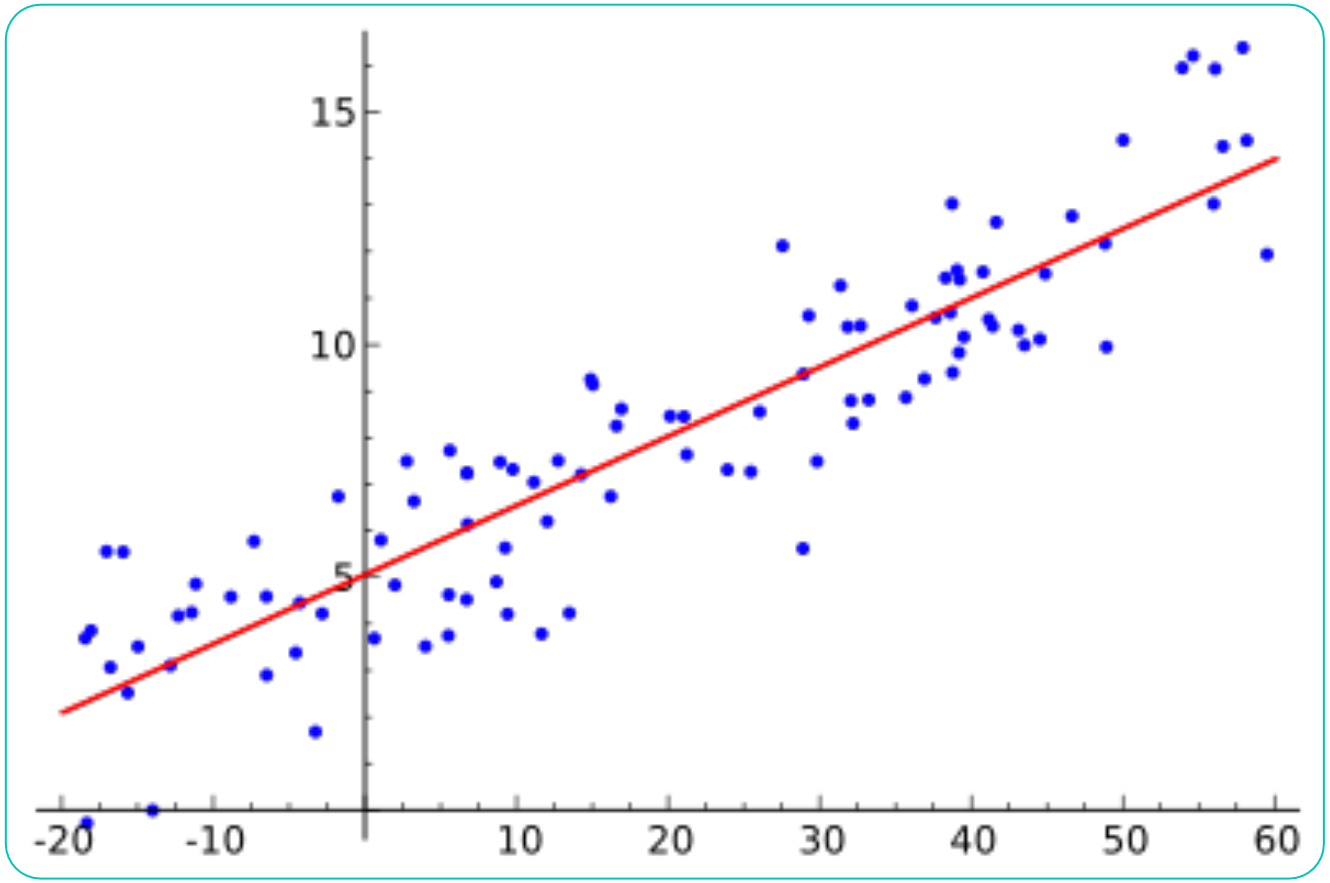
Step 5: Build the Model

- Examples of machine learning in our lives:
  - Netflix and Spotify recommendations
  - Texting autocomplete



Step 5: Build the Model

- A simpler example of machine learning: **linear regression**



Step 5: Build the Model

- We use 'machine learning' to create **models** that we can use to predict unknown **target variables** based on our **features.**

- Examples of **features**:
  - Voter registration
  - Voter race
  - Voter gender
- Examples of **targets**:
  - Whether a person will vote for our candidate
  - How much someone could be expected to donate

Step 5: Build the Model

- We can segment model types into four different categories:
  - **Classification** vs. **Regression**
  - **Supervised** vs. **Unsupervised**

Classification vs. Regression

- **Classification** models allow us to predict the **class** or **label** that an observation belongs to (a **discrete** target)
  - **E.g.** Dem vs. Repub, will vote vs. won't vote

Classification vs. Regression

- **Classification** models allow us to predict the **class** or **label** that an observation belongs to (a **discrete** target)
  - **E.g.** Dem vs. Repub, will vote vs. won't vote
- **Regression** models allow us to predict a **continuous** target
  - **E.g.** How much money a person will donate, how much engagement a social media post will get

Supervised vs. Unsupervised Learning

- **Supervised learning**: We know what we're trying to predict (we have a **target**)
- **Unsupervised learning**: We do not have a target

Supervised vs. Unsupervised Learning

- **Supervised learning**: We know what we're trying to predict (we have a **target**)
- **Unsupervised learning**: We do not have a target (huh?)

Supervised vs. Unsupervised Learning

- **Supervised learning**: We know what we're trying to predict (we have a **target**)
- **Unsupervised learning**: We do not have a target (huh?)
  - **Unsupervised learning** is used when we want to group our observations by similarity (think recommendations in Netflix)

Selecting our Model

- There are a lot of models available:
  - Linear Regression
  - Logistic Regression
  - K-Nearest Neighbors
  - Decision Trees and Random Forests
  - Support Vector Machines
  - Artificial Neural Networks

Selecting our Model

- When selecting a model, there a few things to consider:
  - Situation (classification vs. regression, supervised vs. unsupervised)
  - Interpretability vs. predictive power
  - Computational power
  - Time
  - Distributions of data

Step 5b. Separating Our Data

- Before we build our model on our data, we will do a '**train-test split**'. Any guesses on what this means?

Step 5b. Separating Our Data

- Before we build our model on our data, we will do a '**train-test split**'. Any guesses on what this means?

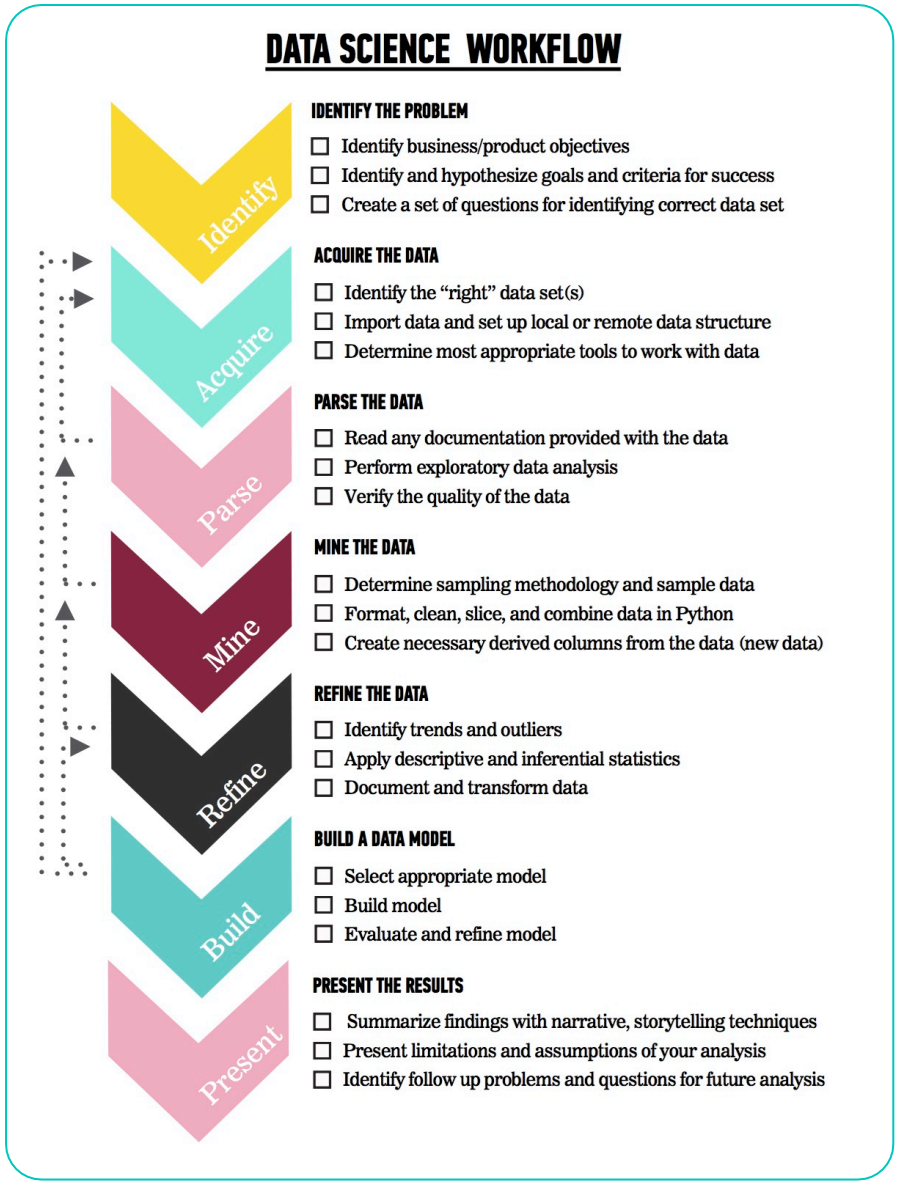
- We're going to take our data and divide it into two sets: **training data** which we'll **build** our model on, and **testing data** which we will **evaluate** it on
- Why might we want to do this?

Step 5b. Separating Our Data

- Before we build our model on our data, we will do a '**train-test split**'. Any guesses on what this means?
- We're going to take our data and divide it into two sets: **training data** which we'll **build** our model on, and **testing data** which we will **evaluate** it on
- Why might we want to do this?
  - Looking at how our model performs on data it has never seen before will give us a more objective evaluation of its true performance.

Step 5c. Making our Model Better

- Once we've trained and evaluated our model, we'll likely find it hasn't done well on the first try. Model building is an **iterative process**, and you may spend a lot of time tuning **parameters** and **hyperparameters**, as well as selecting new features, changing the model, gathering more data, etc.



Data Science is an Iterative Process

Next Steps:

coursera