Data Science Unit 1

Introduction to Data Science

Before we start...

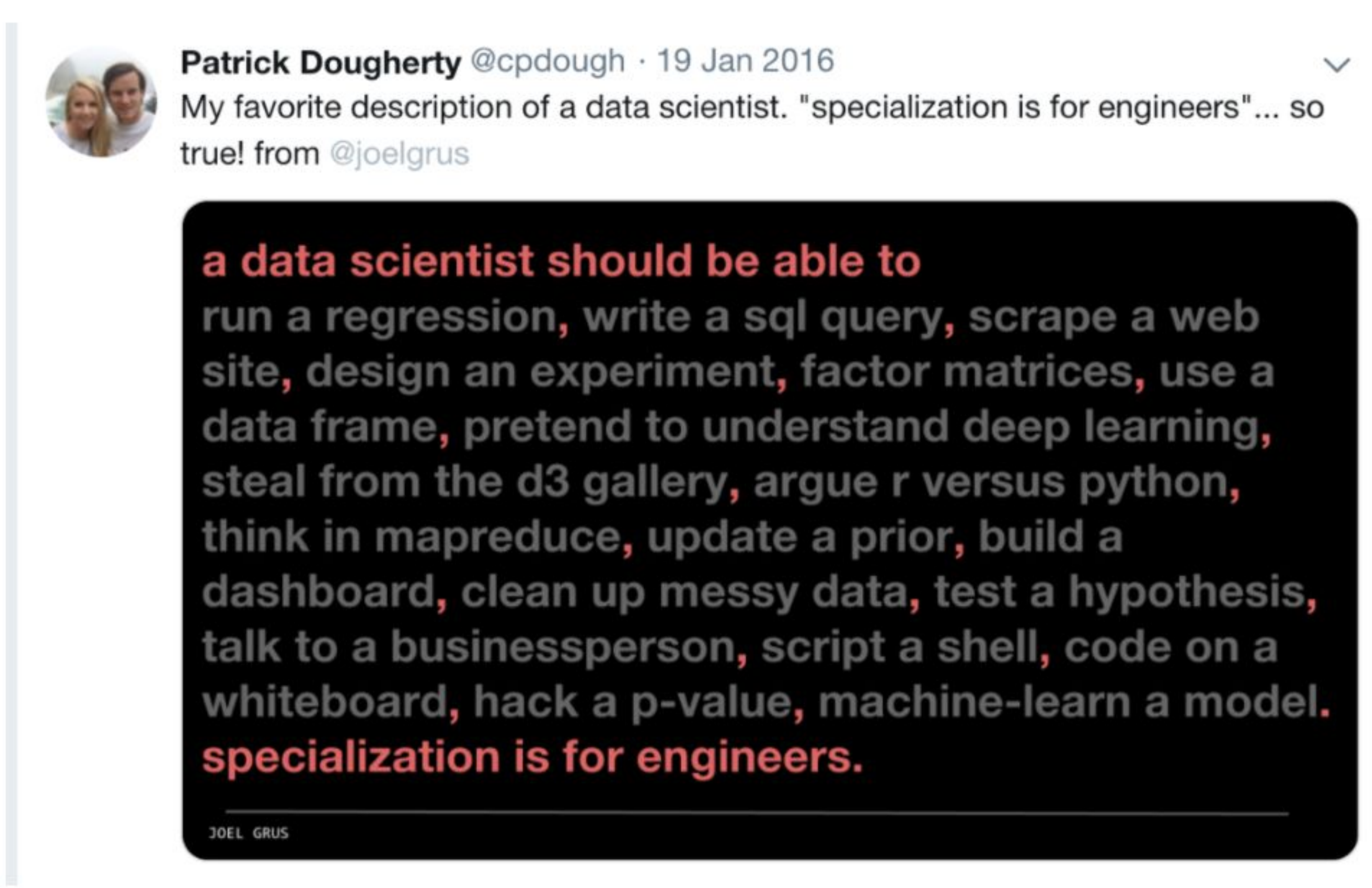
- Make sure you are comfortable
- Have water and maybe a strong coffee handy
- If you need a break...take it!
- If you need a stretch - please go ahead!
- Please mute yourselves if you are not talking
- Have your video on at all times

...and let’s get started!

In this session we will...

- **Understand** what data science is and who a data scientist is
- **Understand** the differences between classical programming and machine learning
- **Discuss** the types of machine learning
- **Understand** algorithms and algorithm training
- **Familiarise** yourself with machine learning terms and definitions

What is a Data Scientist?



Patrick Dougherty @cpdough · 19 Jan 2016

My favorite description of a data scientist. "specialization is for engineers"... so true! from @joelgrus

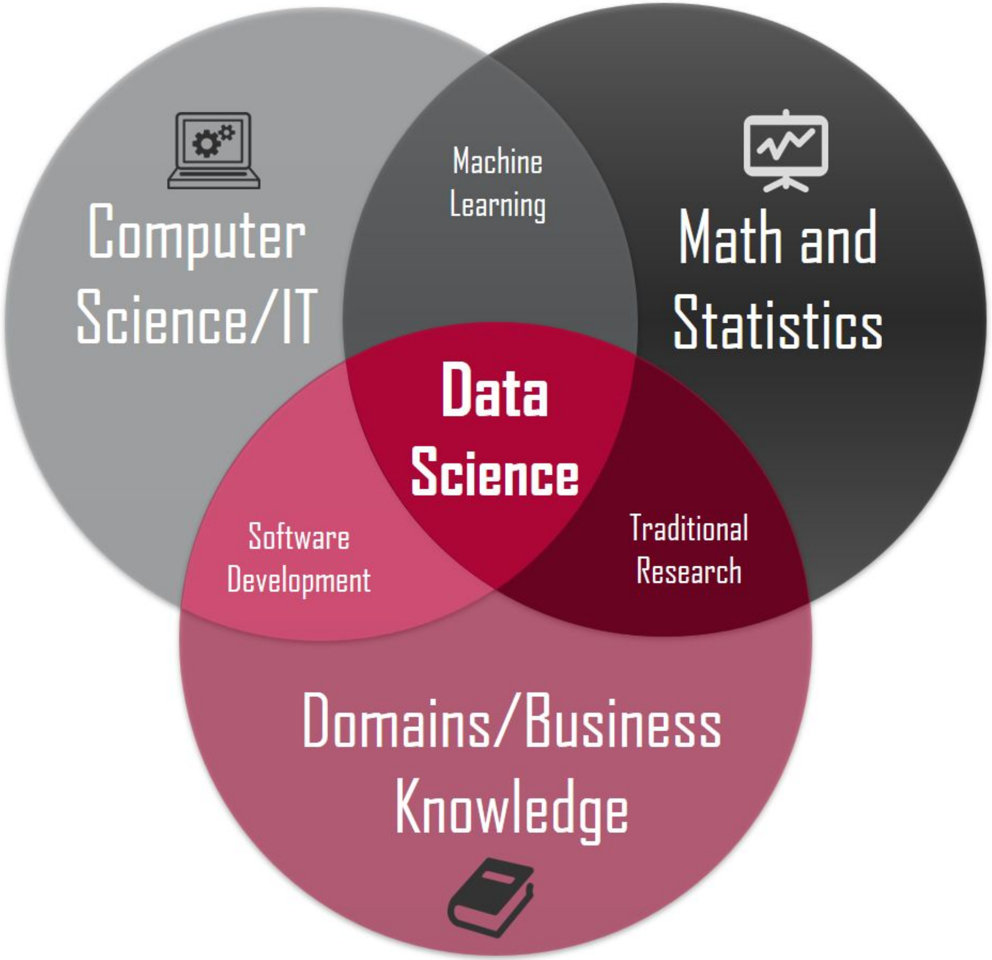
a data scientist should be able to

run a regression, write a sql query, scrape a web site, design an experiment, factor matrices, use a data frame, pretend to understand deep learning, steal from the d3 gallery, argue r versus python, think in mapreduce, update a prior, build a dashboard, clean up messy data, test a hypothesis, talk to a businessperson, script a shell, code on a whiteboard, hack a p-value, machine-learn a model.

specialization is for engineers.

JOEL GRUS

Intro to Data Science



Activity

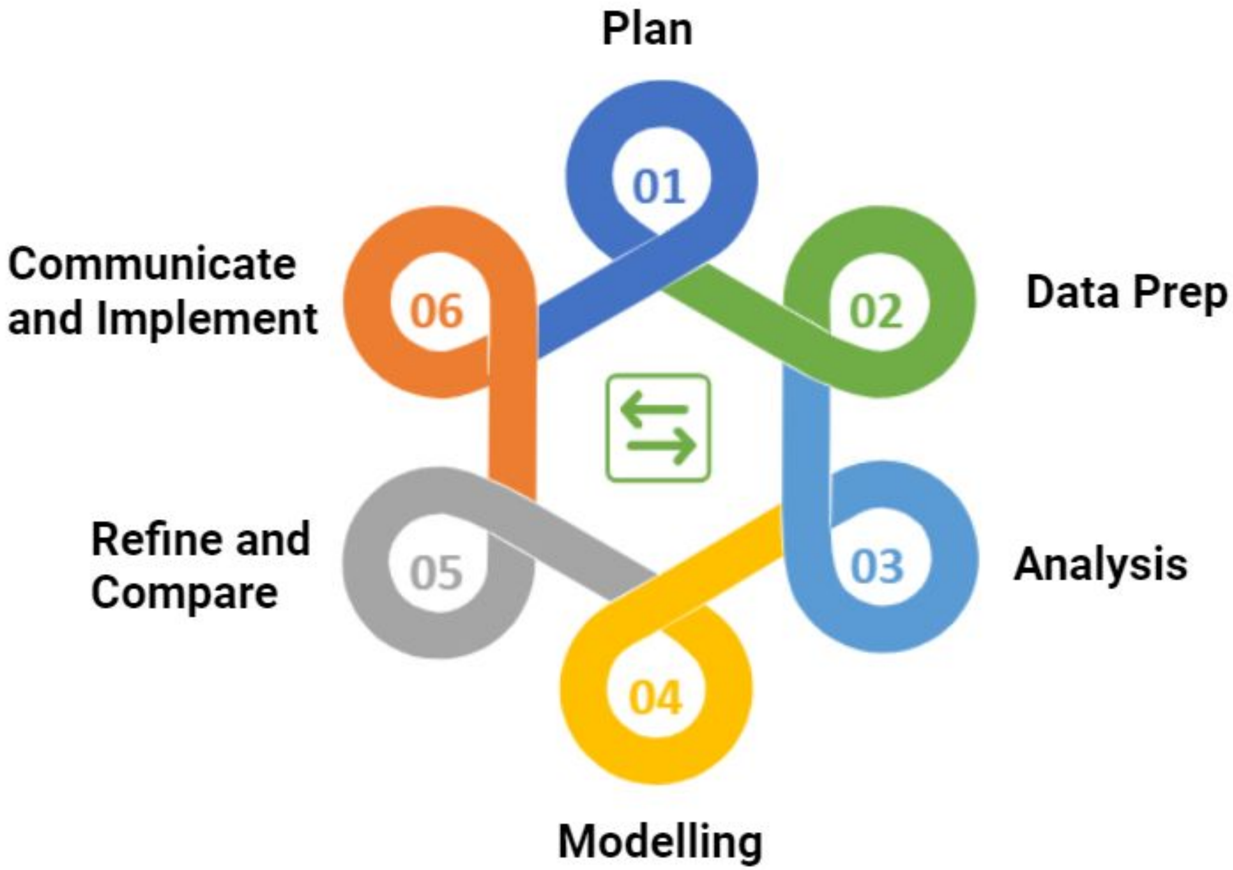
Give an example of a product or service you think utilises data science

How Is Data Science Used

Steve Jobs, 1955 - 2011

Data Science Workflow

Recall the Data Analytics Lifecycle



Activity

Business Scenario

You work for a real estate company interested in using data science to determine the best properties to buy and resell. Specifically, your company would like to identify the characteristics of residential houses that estimate their sale price and the cost-effectiveness of doing renovations. Using the analytics life cycle, describe the activities that you would carry out in each stage.

10 minutes

Plan

- **Identify the business/product objectives.**
  - The customer tells us their business goals are to accurately predict prices for houses (so that they can sell them for as large a profit as possible) and to identify which kinds of features in the housing market would be more likely to lead to foreclosure and other abnormal sales (which could represent more profitable sales for the company).
- **Identify and hypothesise goals and criteria for success.**
  - Deliver a presentation to the real estate team.
  - Write a business report discussing results, procedures used, and rationales.
  - Build an API that provides estimated returns.
- **Create a set of questions to help you identify the correct data set.**
  - Can you think of questions that would help this customer deliver on their business goals?
  - What sort of features or columns would you want to see in the data?

Data Prep

Common considerations when preparing our data include:

- Ensuring data is clearly defined and structured
- Check and clean data formatting as needed

Most data will not come perfectly clean and ready to use. Cleaning data is normally the most time-consuming task a data scientist faces.

Analyse

Data scientists often check for their data the:

- Mean
- Median
- Standard deviation
- Specific frequency counts
- Distribution
- Existence of Outliers

Model

Look to predict a value we are interested in, for example:

- House price (Regression)
- Number of rooms (Classification)

Refine and Compare

Develop Recommendations and Decisions

- Did you reject or fail to reject your hypothesis?
    - What does this mean for your project?
    - What does this mean for your client?
- Were your questions answered?
    - Which ones?
    - What do you need to do to answer the ones that weren't?
- Do your findings support any business recommendations, actions, or decisions?
    - Is there further supportive analysis?
    - How do your data support these recommendations?

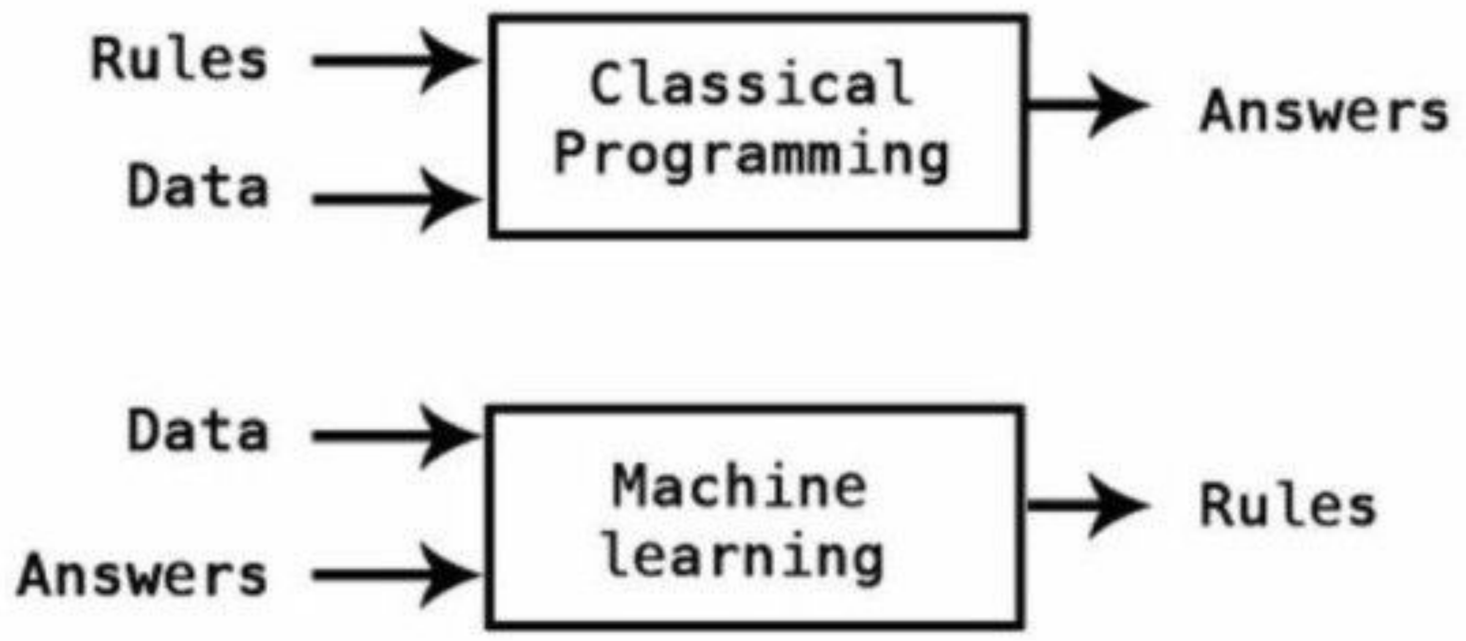
Communicate and Implement

Share the Results of Your Analysis

- Reaching a conclusion:
  - Seek guidance/interaction with subject matter experts (SMEs).
  - If those are not available, check with the data — are you coming to reasonable conclusions and predictions given what you've seen?
  - Do the next steps that you envision have any dependencies or corollary steps?
- What are some conclusions you can draw?
  - Conclusion: " "Properties with 3 or more bedrooms were twice as likely to sell than properties with 2 or less bedrooms"".
  - Recommendation: "We should target (buy and renovate) properties with 3 or more bedrooms."
  - Conclusion: "Other than the number of rooms I found no significant evidence that any other feature affected the odds of properties selling faster than others".

Introduction to Machine Learning

ML



Rules
Data
Classical Programming
Answers
Data
Answers
Machine learning
Rules

ML

Rules:

- **Multiply data by 5**
- **Add 7**

Data:

- **[3, 4, 5]**

Answer (computer generates answers based on rules and data)

- **[22, 27, 32]**

ML

Answers

- [22, 27, 32]

Data:

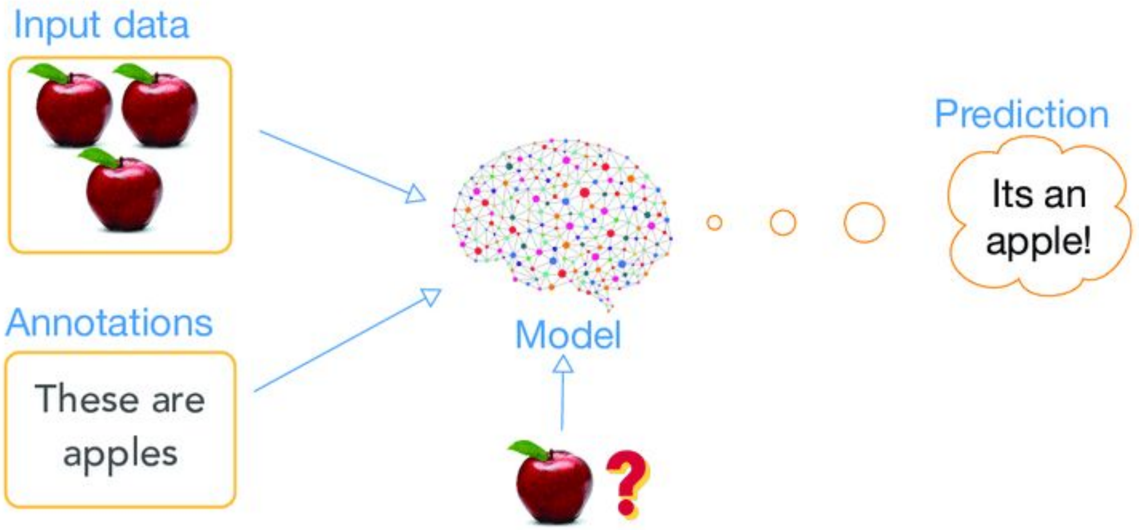
- [3, 4, 5]

Answer (computer generates rules based on data and answers)

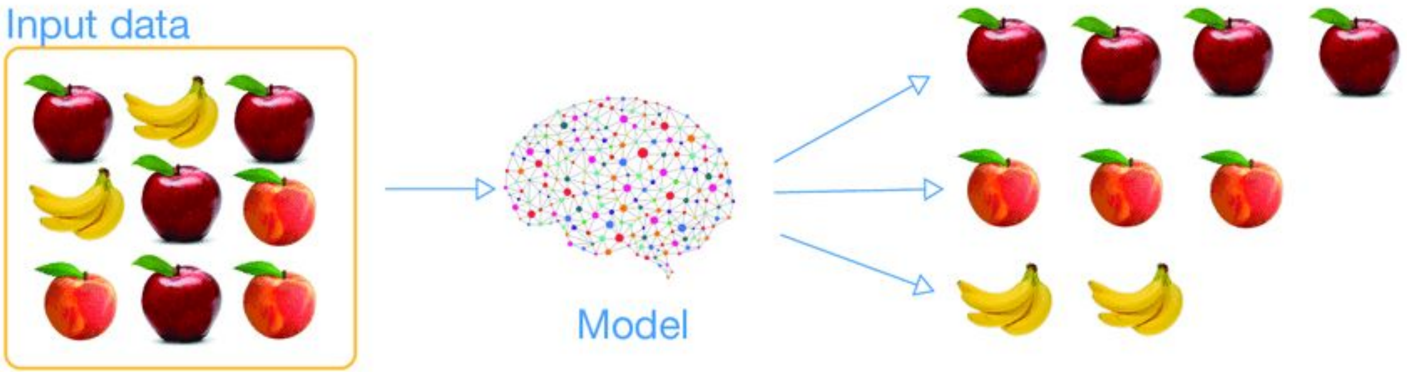
- Data * 5 + 7

Categories of ML

- **Supervised Learning**



- **Unsupervised Learning**



Supervised Learning

“The model is provided with both data (FEATURES) and the answers (TARGET). To put it simply, train the model using LABELLED data.”

Types of Supervised ML

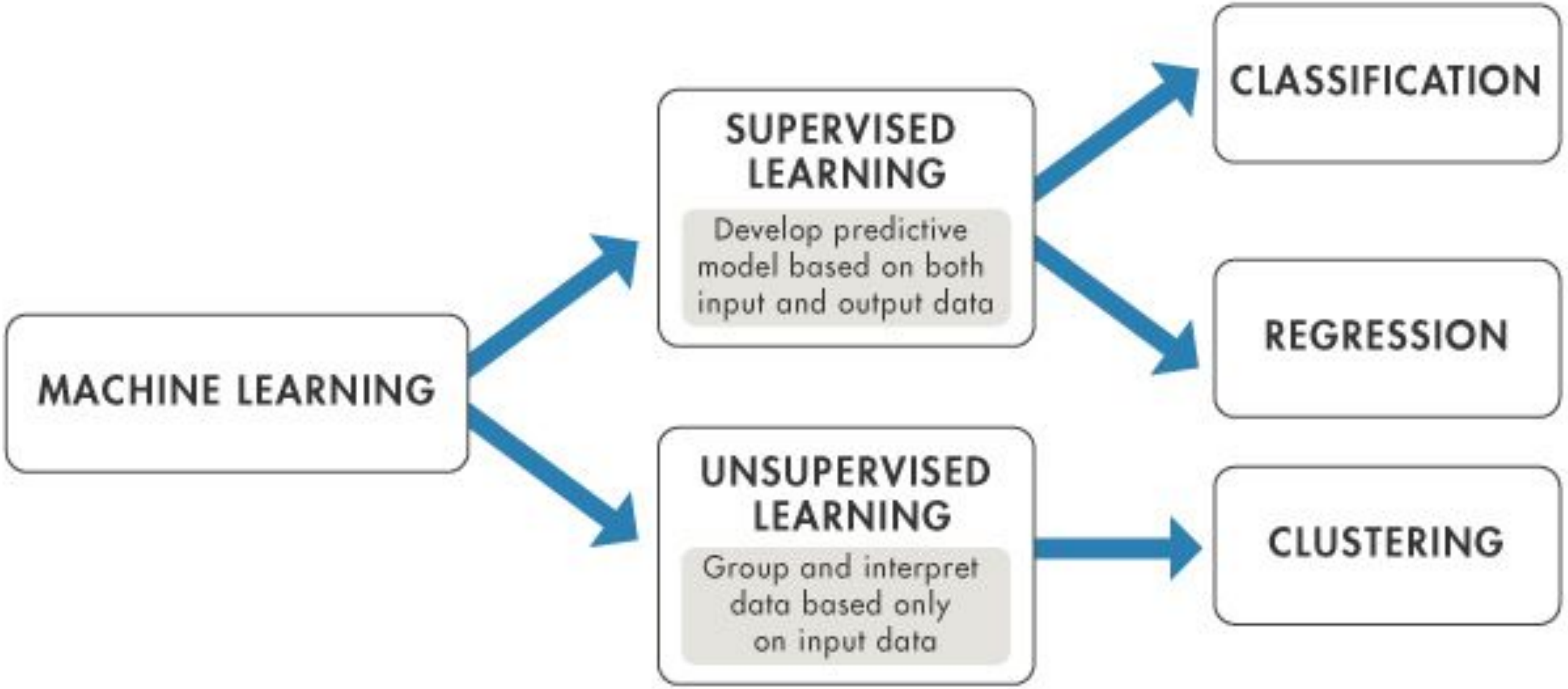
- **Regression**: The outcome we are trying to predict is a continuous value.

- **Classification**: The outcome we are trying to predict is categorical.

Unsupervised Learning

“The model is provided with only data (FEATURES) and it learns the interactions in the features, creating groups (CLUSTERS) in the process”

Machine Learning



MACHINE LEARNING

SUPERVISED LEARNING

Develop predictive model based on both input and output data

UNSUPERVISED LEARNING

Group and interpret data based only on input data

CLASSIFICATION

REGRESSION

CLUSTERING

Activity

Activity

- Open the folder ‘Ames_housing’- there is a dataset called “ames.csv” and a file called “description.txt”
- Your task is to have look at the data and sketch out answers for the following
    - What is a potential target in your data for a regression model?
    - What is a potential target in your data for a classification model?
    - (Extend) Could unsupervised learning be used within this data? How so?

10 minutes

Algorithms

Algorithms

“an algorithm is a finite sequence of well-defined, computer-implementable instructions, to solve a problem or to perform a computation”

Algorithm

Algorithm:

- **Multiply data by 5**
- **Add 7**

Equation:

Answer = 5 * data + 7

$y = mx + c$

Activity

Activity

Let's say we are a real estate agent looking to price a house using only its **square footage**. We know there are other features that can highly influence this outcome, but we are only focusing on square footage for now. **We know that, as square footage increases, so does price**.

Recently, we sold a house whose **square footage was 2,500** for about **£285,000** and an additional **£10,000** for stamp duties. Based on this information:

1. Generate an equation for house price using the square footage and stamp duties
2. Generate an algorithm for a computer to compute price of a house given this information

10 minutes

Activity

Activity

- In breakout rooms, think about a use case (if any) or potential use case of ML in your organization
- What type of ML would it be (supervised or unsupervised)? If supervised, is it going to be regression or classification?
- What benefits does your organization get with potential implementation of ML

10 minutes