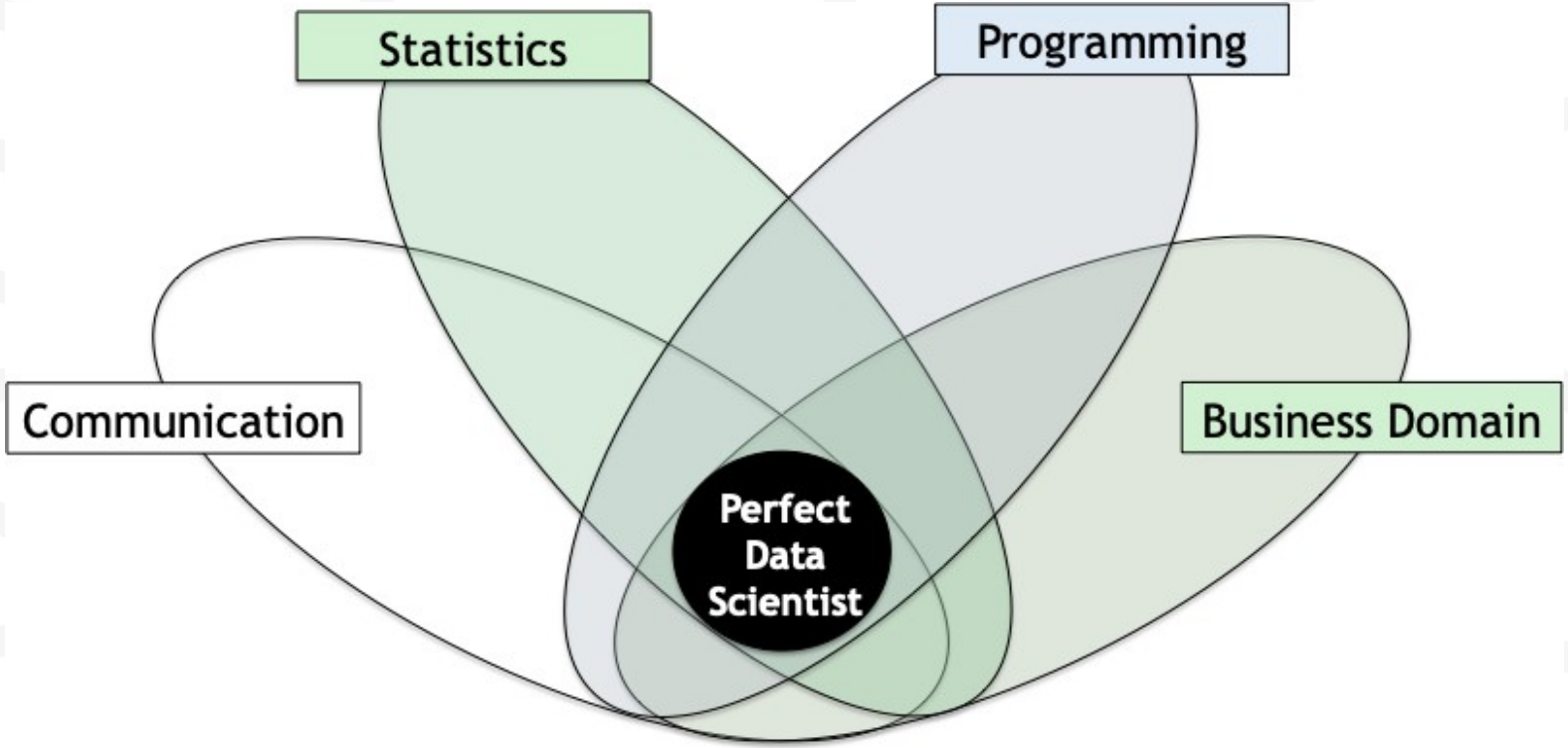
The Data Science Process & Problem Definition

Data Science: a Multidisciplinary Blend of Skills

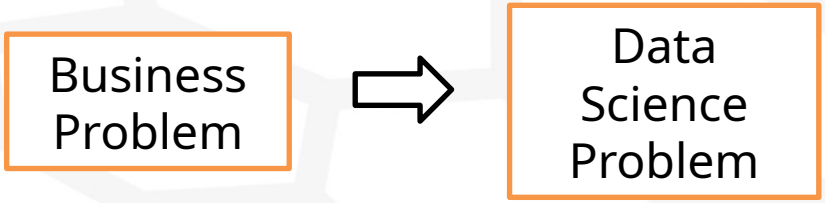


Data Science Process

Don’t do Data Science …

Solve Business Problems

“A good data science problem should be relevant, specific and unambiguous. It should align with the business strategy.”



Business leaders and subject matter experts need to work alongside data scientists to formulate a business problem and translate it to a data science problem.

Defining the Business Problem

Below are some things to consider when defining a business problem.

- What problem does the business have?
- What data do we need to solve this?
- What will success look like? How will we measure performance?
- How will the results be used?
- What constraints do we have?

Fraud Detection - Example

We are a credit card processing company and have had issues with fraudulent charges.

We want to be able to predict when a pending transaction is fraudulent.

Fraud Detection – Example

Question	Answer
What is the business problem?	
What data do we need to solve this?	
What will success look like?	
How will the results be used?	
What constraints do we have?	

Amount	Date	Type	...
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

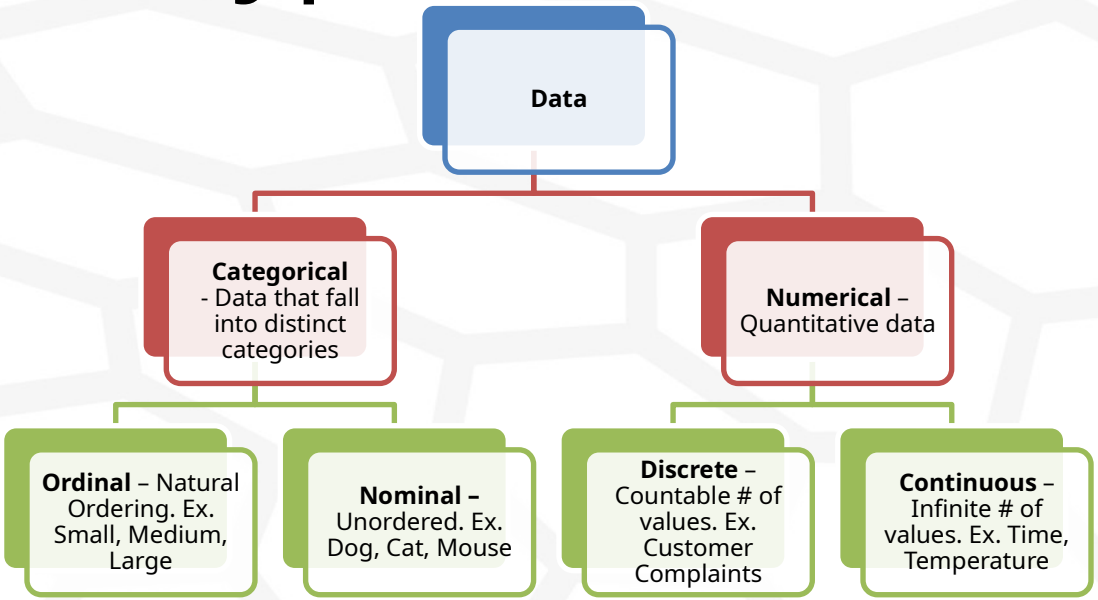
Data Science Problems

Once the business problem has been defined, we can work on translating the **business problem** to a **data science problem**.

Data science problems typically

- Categorize or group data
- Detect patterns (or anomalies)
- Show relationships between variables
- Predict outcomes

Types of Data



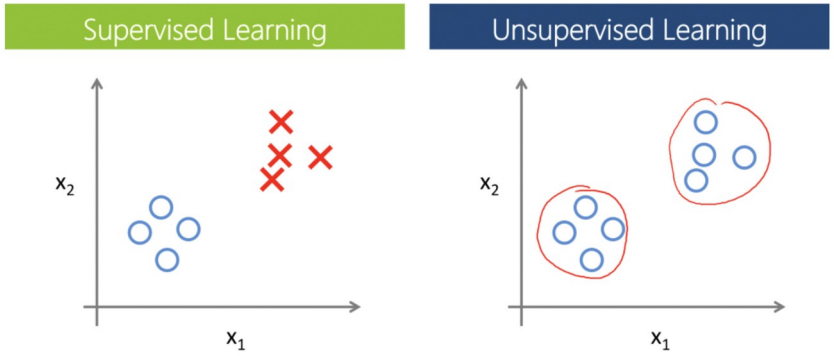
You need to know the type(s) of data you are dealing with before you can define your data science problem.

Types of Data Quiz

Age	Gender	Height	Number of Pets	Favorite Food	Cholestoral
22	F	63.5	0	Pizza	Low
15	M	69.2	1	Tacos	Normal
24	F	58.8	3	Ice Cream	Low
54	F	65.4	2	Enchiladas	Normal
35	M	70.3	5	Pasta	Normal
39	F	66.1	2	Fruit	Normal
40	M	68.4	1	Pizza	High
23	M	72.9	1	Oatmeal	Normal
32	M	73.3	2	Sushi	High
46	F	55.7	0	Popcorn	Normal
67	M	56.6	1	Chocolate	Low
82	F	54.1	3	Steak	Normal
39	F	52.8	2	Fried Chicken	Normal

What Type of Data Science Problem is it?

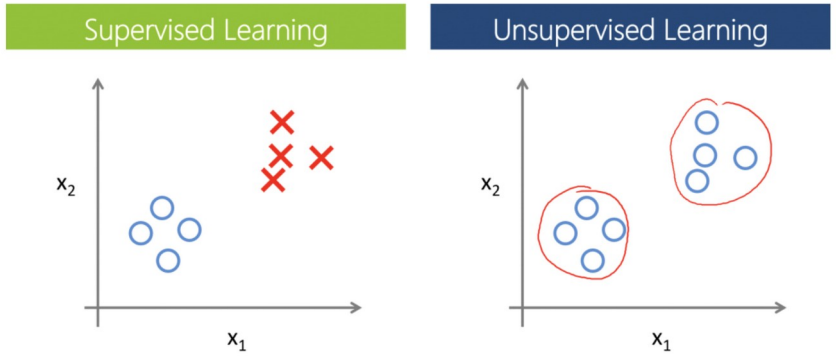
Supervised vs. Unsupervised Learning



Supervised learning involves building a model for predicting an output based one or more inputs.
Unsupervised learning involves learning relationships of inputs without a labeled output.

What Type of Data Science Problem is it?

X1 age	X2 income	Response Purchase?
18	5,000	No
20	20,000	No
22	19,000	No
24	32,000	No
28	40,000	Yes
32	35,000	Yes
35	50,000	Yes
42	55,000	Yes



X1 age	X2 income
18	5,000
20	20,000
22	19,000
24	32,000
28	40,000
32	35,000
35	50,000
42	55,000

Supervised learning involves building a model for predicting an output based one or more inputs.
Unsupervised learning involves learning relationships of inputs without a labeled output.

What Type of Data Science Problem is it?

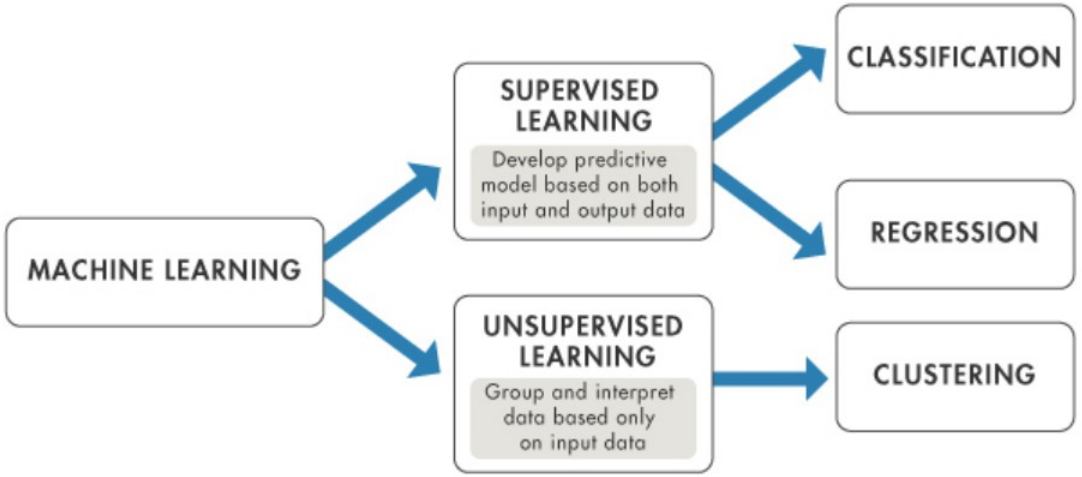


Image per MathWorks

Supervised learning can be broken up into classification and regression problems.
Unsupervised learning is generally a clustering problem.

Regression vs. Classification

Regression

Let's build a model to predict what the temperature will be tomorrow.

70° F

Classification

Let's build a model to predict whether or not it will rain tomorrow.

Yes

Regression involves making a prediction of a continuous output.
Classification involves making a prediction of a categorical or binary output.

What is a Predictive Model?

Real Life Observation

$$y = weight = f(height, age) + \epsilon$$

Our Model Approximating Real Life

$$\hat{y} = \hat{f}(height, age)$$

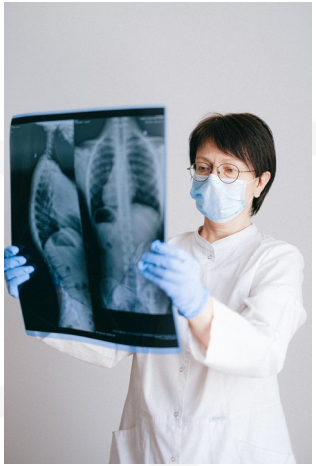
- Response
- Prediction
- Dependent Variable
- Output

Model

- Features
- Predictors
- Independent Variables
- Inputs

What Are the Use Cases?

- Supervised: Classification
  - Predicts a discrete category
    - Medical imaging, Natural Language Processing, image recognition
- Supervised: Regression
  - Predicts a numerical response
    - Home sale price, weather forecasting, blood sugar levels
- Unsupervised: Clustering
  - Groups together data based on similar characteris
    - Political polling, retail recommendation system



Get Specific on Problem Type

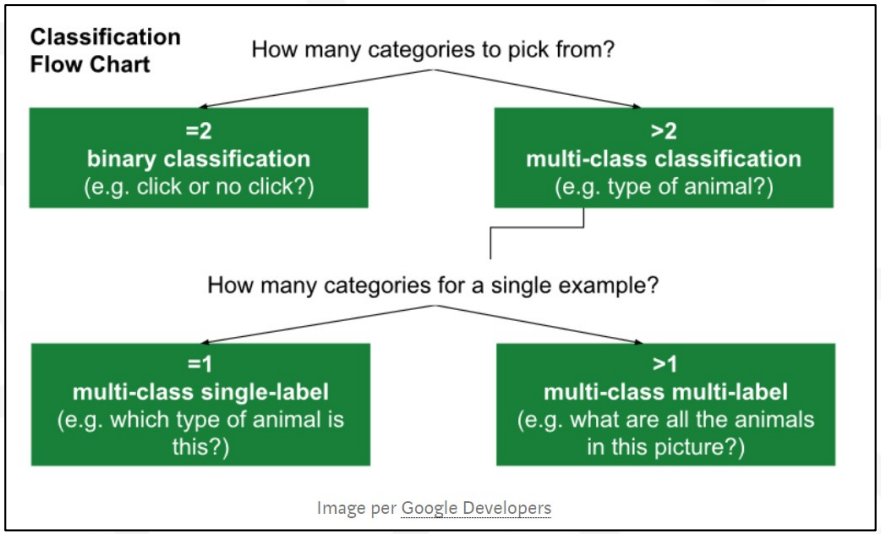


Image per Google Developers

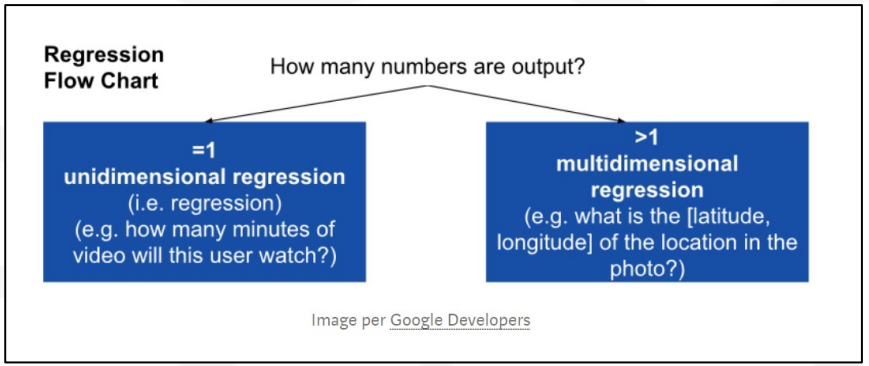


Image per Google Developers

Fraud Detection - Example

We are a credit card processing company and have had issues with fraudulent charges.

We want to be able to predict when a pending transaction is fraudulent.

What type of data science problem is this?

Data Science Problem Activity

Recap

- Data science is the process of using data analytics, statistics, and programming to solve business problems.

- Don't do data science; instead, solve business problems.

- When scoping your problem, think about the following:
  - The data needed to solve the problem
  - The desired outcome or intended usage
  - How success will be measured
  - Any constraints
  - The data science problem