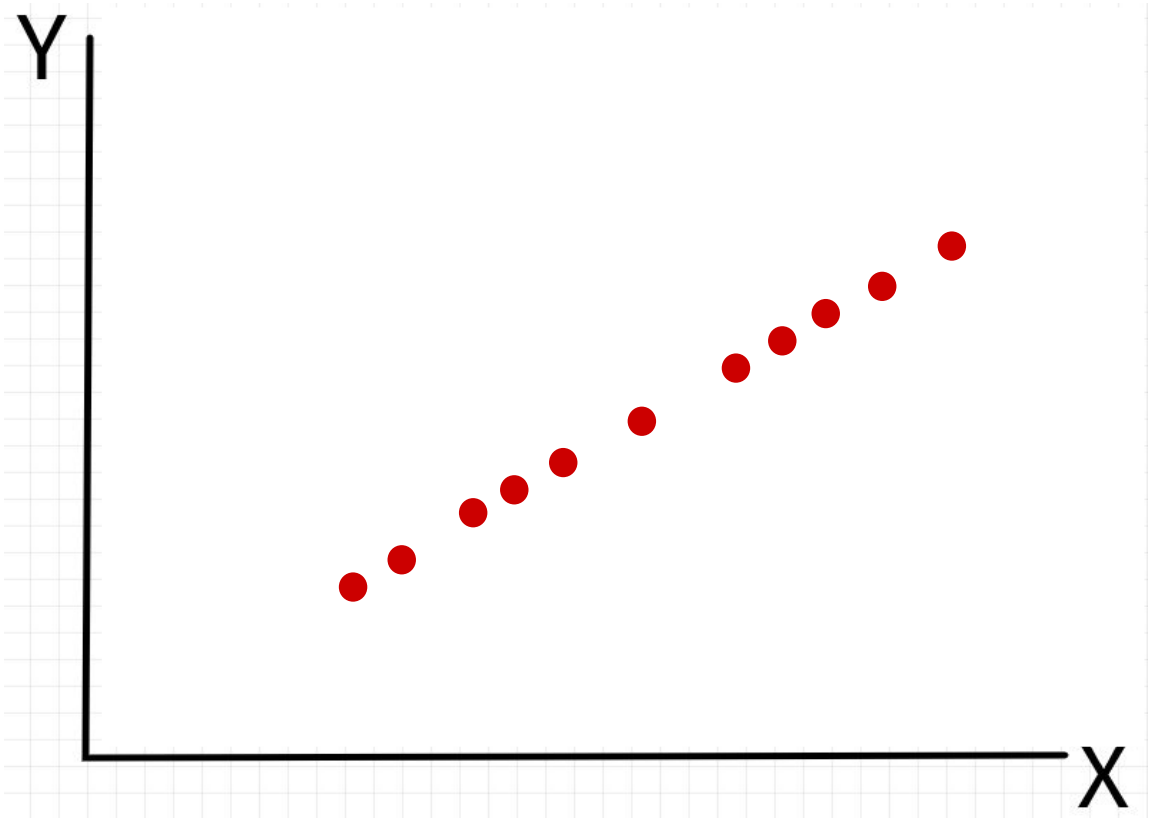
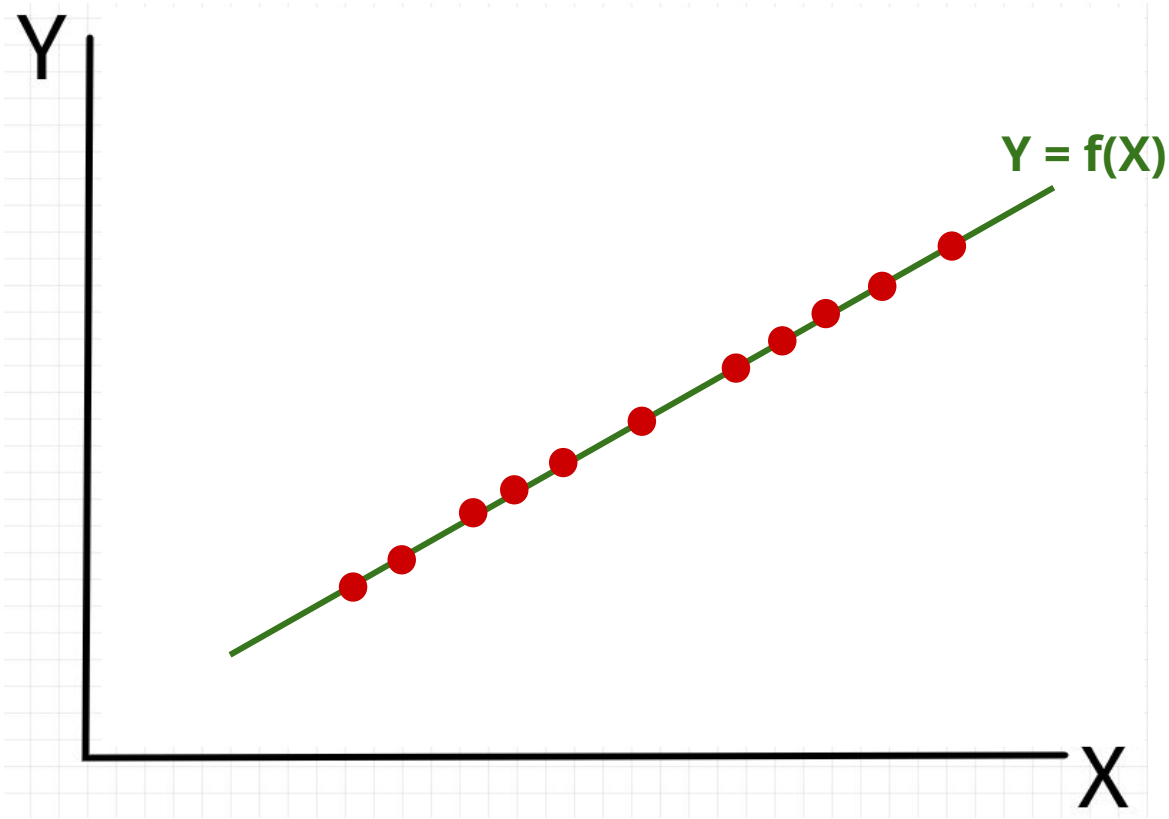
Introduction

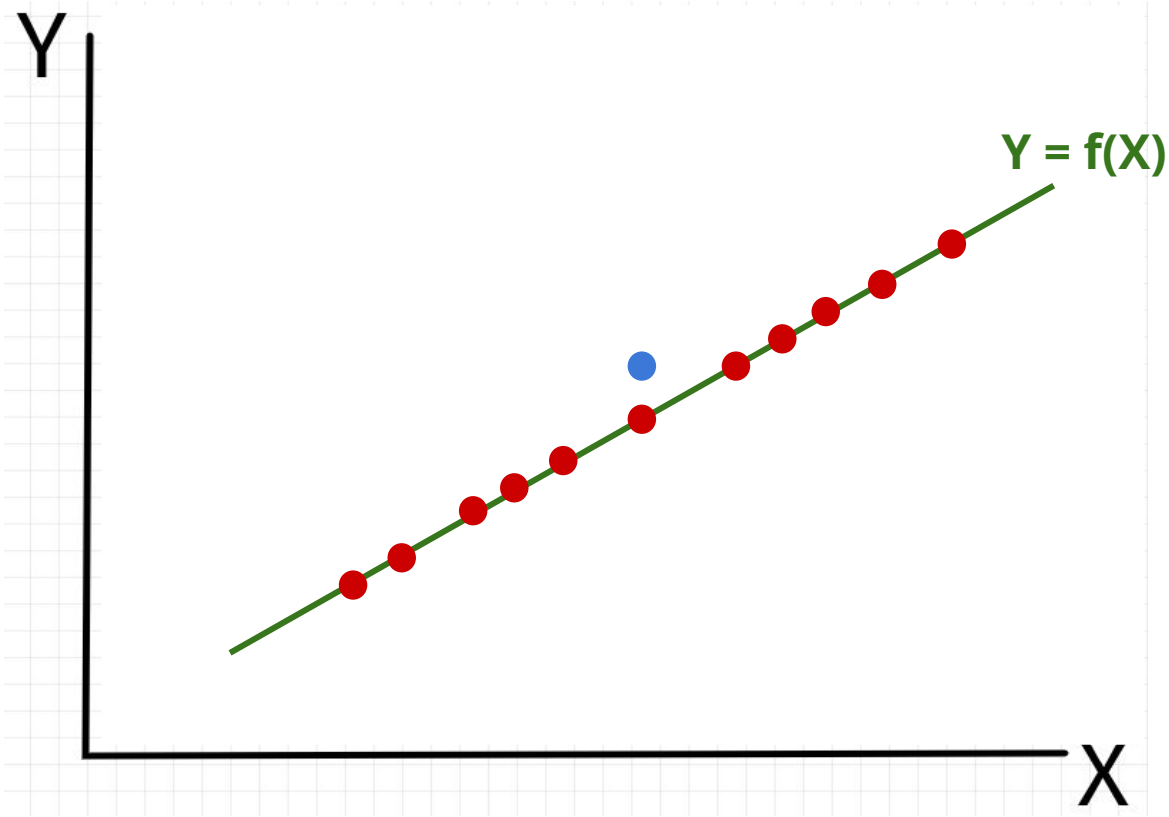
Boston University CS 506 - Lance Galletti

Data Science

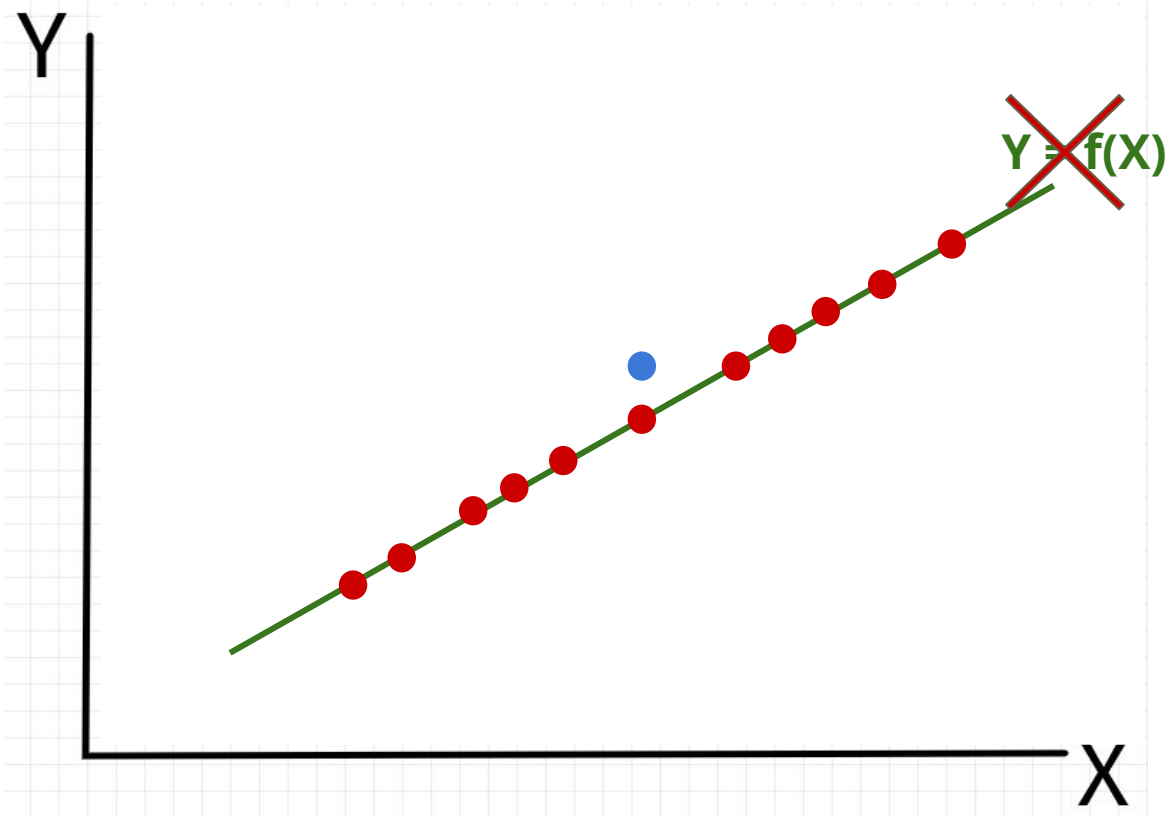
- Collection of methods and tools that allow for extracting knowledge from data
- Cross-disciplinary:
  - Math
  - Statistics
  - Computer Science
  - Domain Expertise
- Know what you don’t know!


Y
X

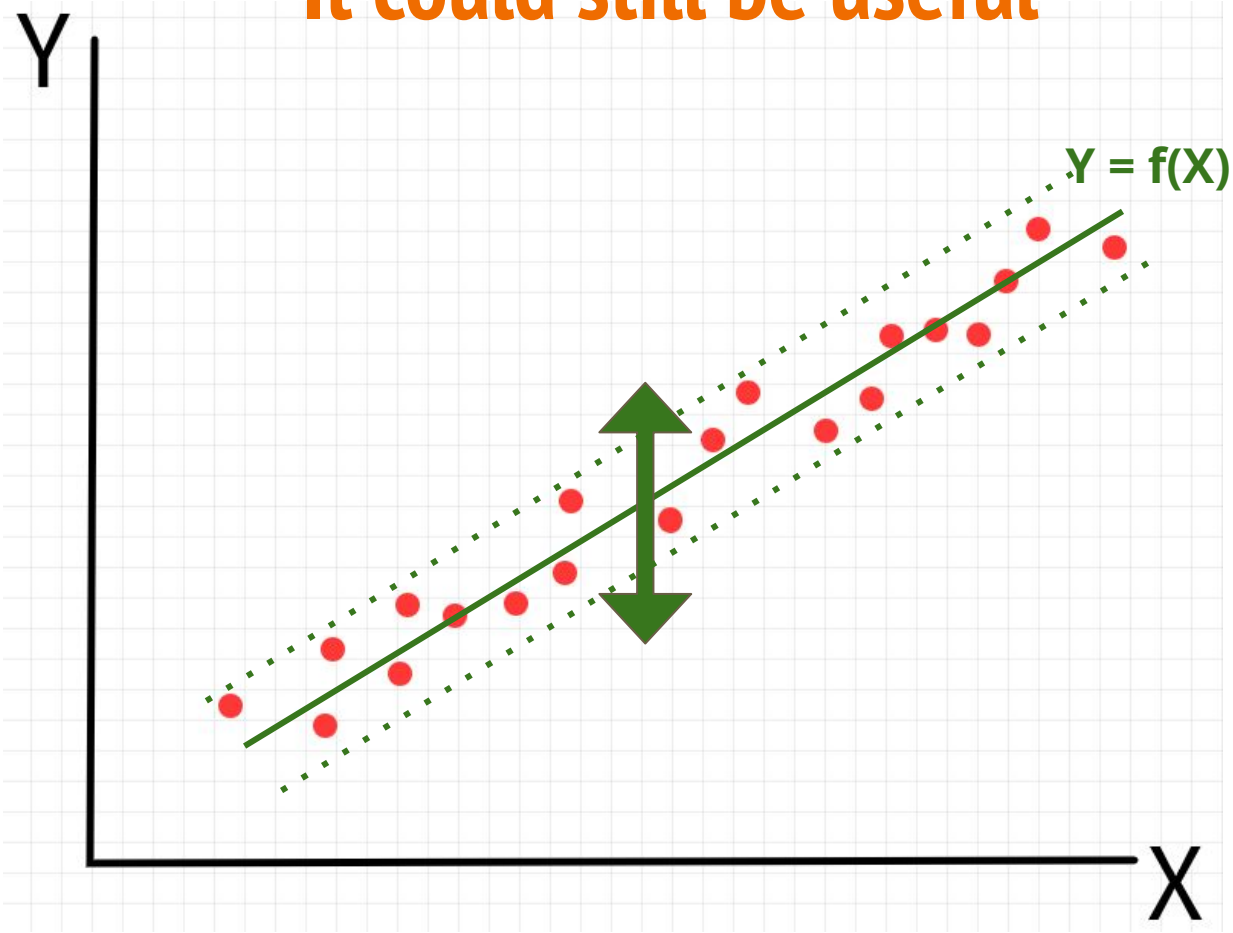

Y
Y = f(X)
X


Y
Y = f(X)
X

Your hypothesis is wrong, but it still might be useful



It could still be useful



Confirmation Bias

In a class just like this one, imagine playing the following game…

Confirmation Bias

I announce “(2, 4, 6) follows the rule”.

Here are the examples submitted by one of the participants:

- (2, 4, 3) -> NO
- (6, 8, 10) -> YES
- (1, 3, 5) -> YES

After which, they proceed to write down their hypothesized rule. Would you have wanted to try more examples? If so, which and for what reason?

Confirmation Bias

Let’s take a poll:

A. (100, 102, 104)
B. (5, 7, 9)
C. (1, 2, 3)

I guessed this because it helps us see if there is any upper limit to the rule

Confirmation Bias

Challenges of Data Science:

- A set of examples may not always be representative of the underlying rule
- There may be infinitely many rules that match the examples provided
- Rules and/or examples may change over time

So Data Science is VERY DIFFICULT!!! All models are wrong but some are useful

Confirmation Bias

Positive Examples VS Negative Examples

reinforces hypothesis disproves hypothesis

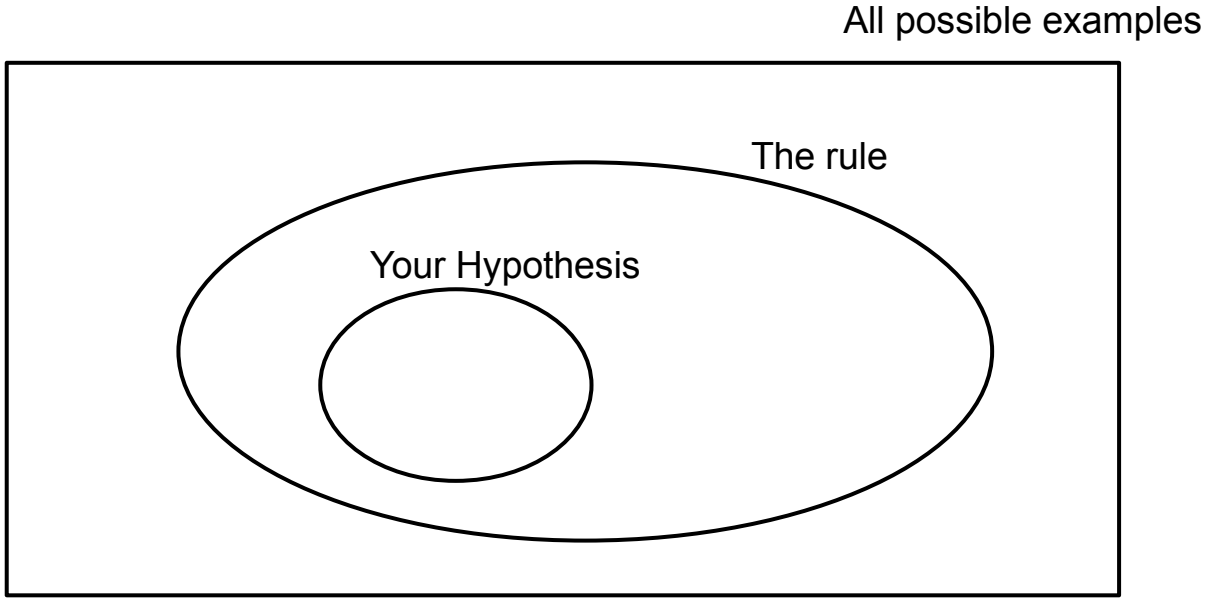
assuming the hypothesis h is (x, x+2, x+4) which type of examples are the following:

- (2, 4, 3)
- (6, 8, 10)
- (1, 3, 5)

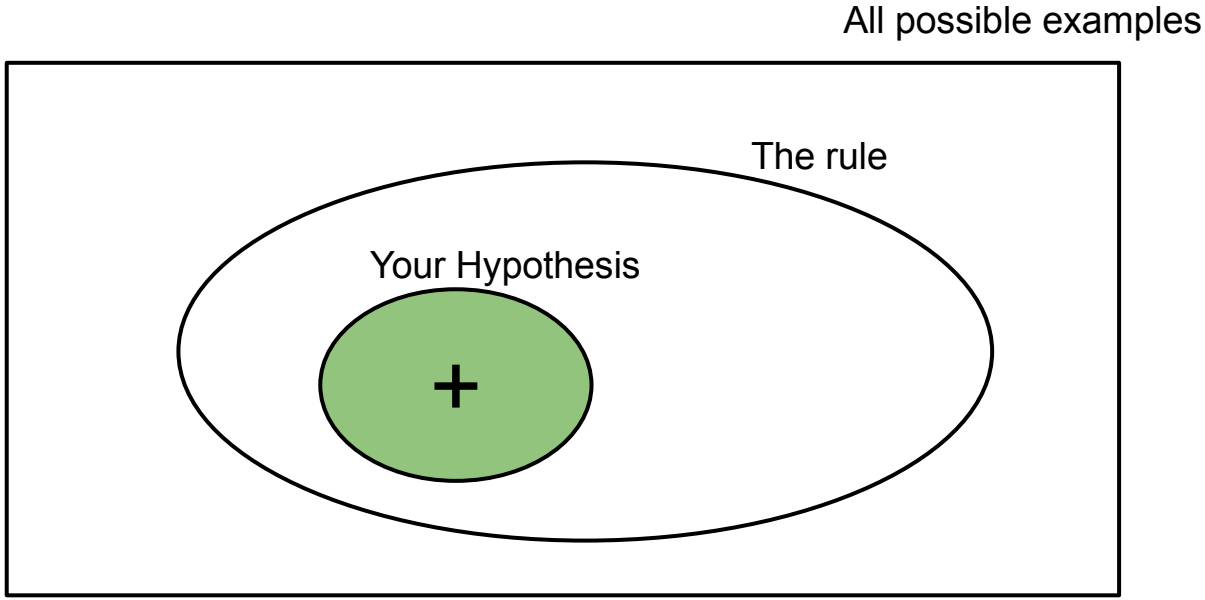
Confirmation Bias

- Both positive and negative examples can falsify a hypothesis
- Tendency to choose positive ones over negative ones

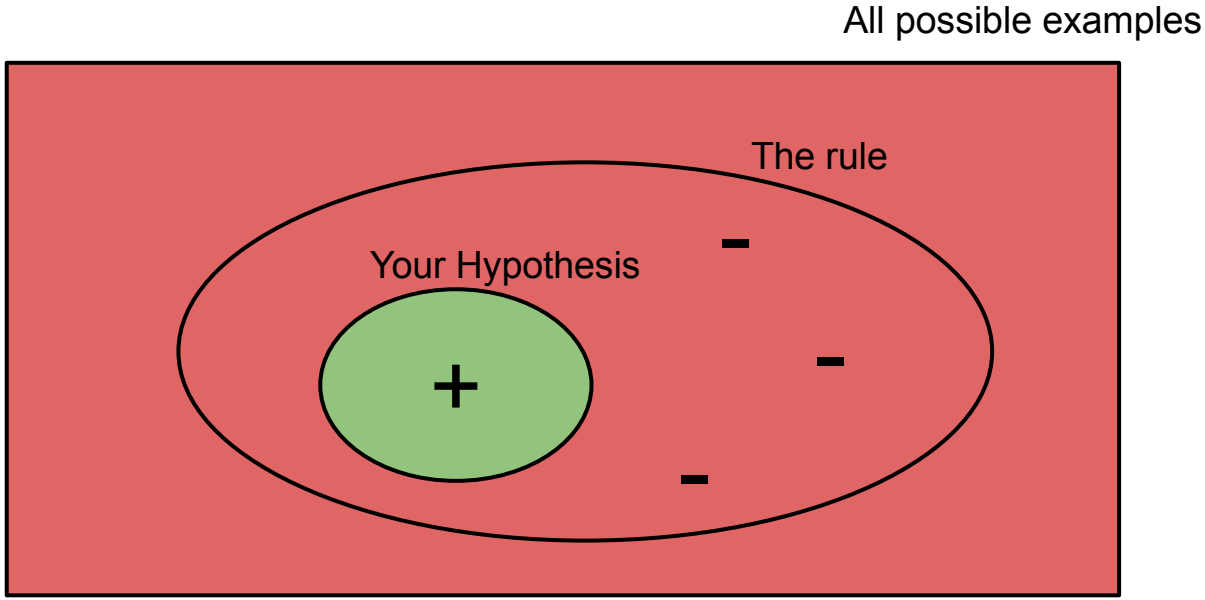
Confirmation Bias



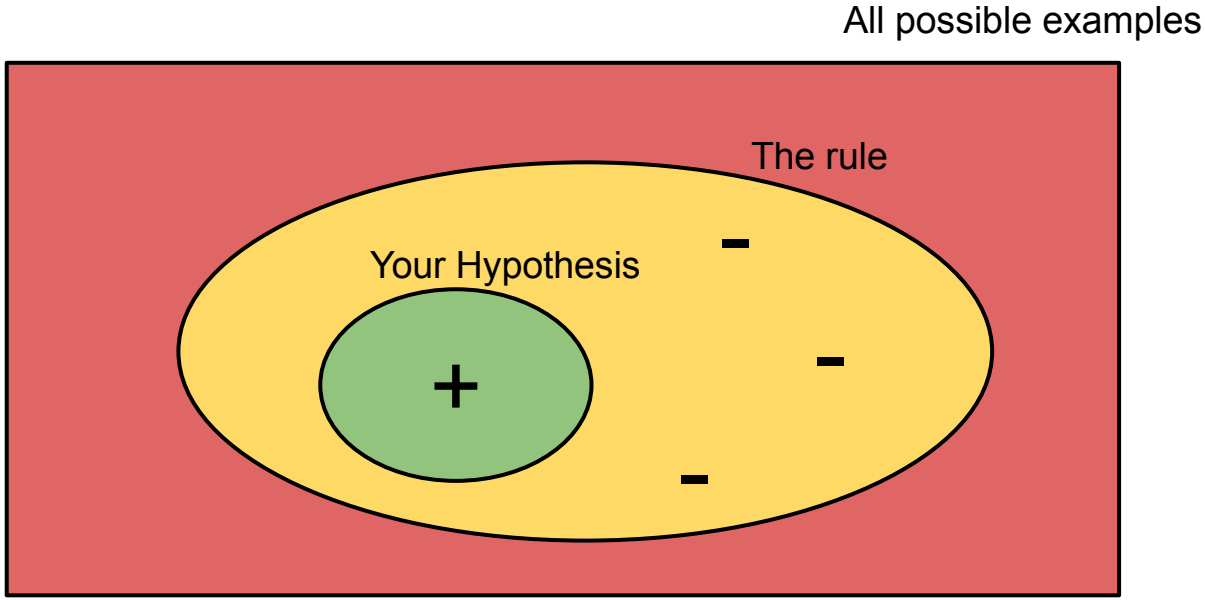
Confirmation Bias



Confirmation Bias



Confirmation Bias



Confirmation Bias

Let’s take a poll:

A. (100, 102, 104)
B. (5, 7, 9)
C. (1, 2, 3)

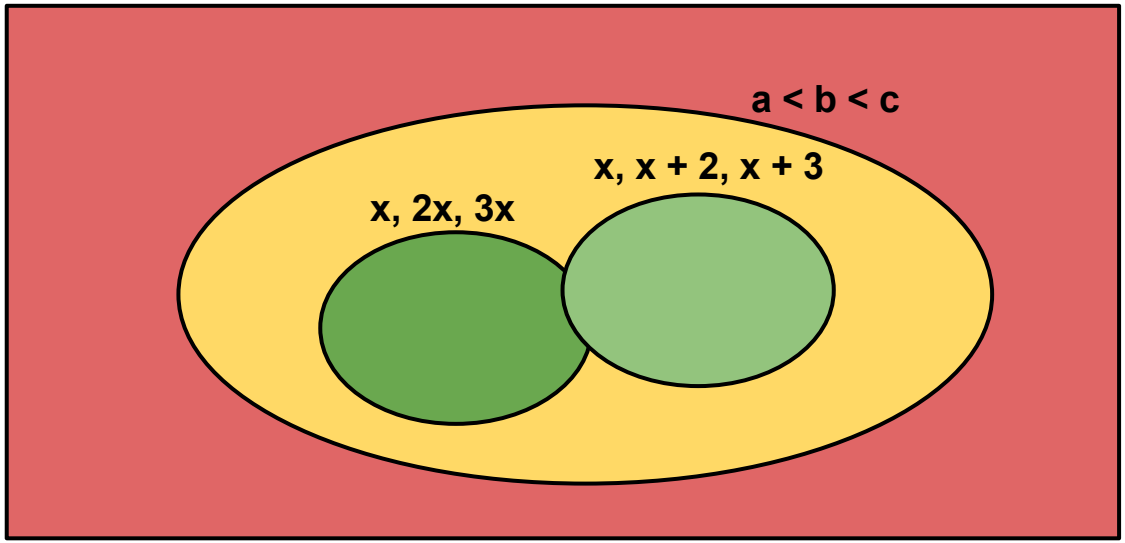
Confirmation Bias

The rule was (a < b < c).
If you only tried positive examples of either (x, x + 2, x+4) or (x, 2x 3x) you would only get confirmation.

For reference, this exercise was first introduced by Wason P.C in 1960 as part of a journal in experimental psychology.

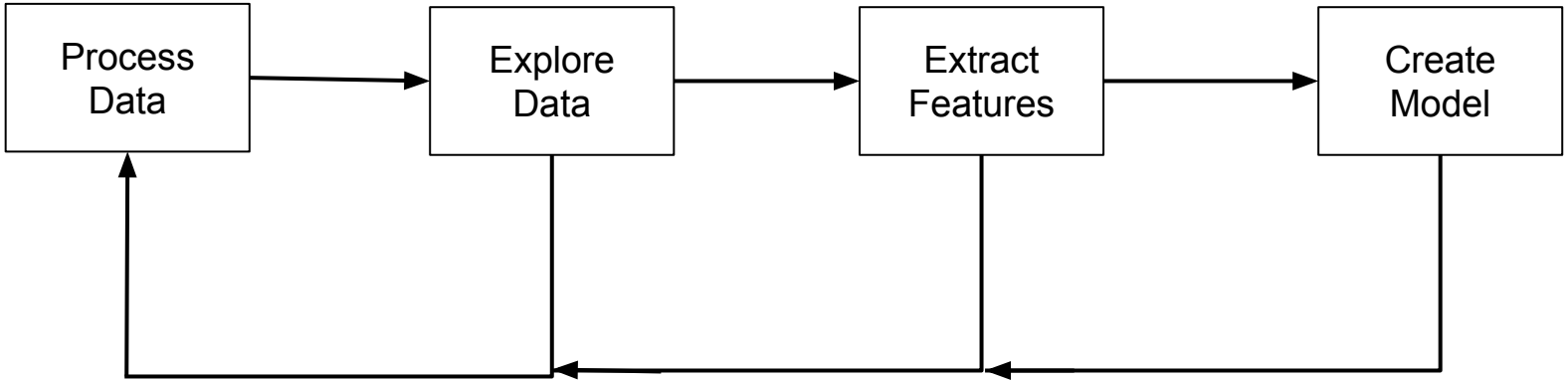
Confirmation Bias

All possible examples



Predict a student’s gpa

Data Science Workflow (simplified)



Types of Data

Types of Data - Records

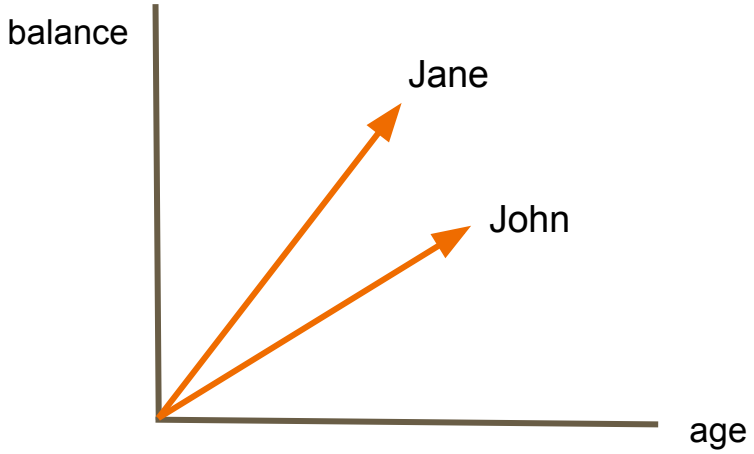
m-dimensional points / vectors

Example: (name, age, balance) -> (“John”, 20, 100)

Types of Data - Records

m-dimensional points / vectors

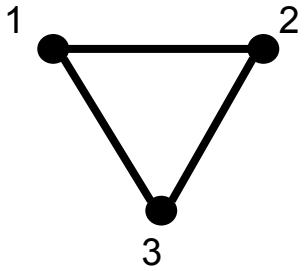
Example: (name, age, balance) -> (“John”, 20, 100)



Types of Data - Graphs

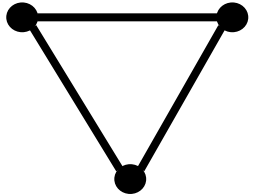
Nodes connected by edges

Example:



Adjacency Matrix

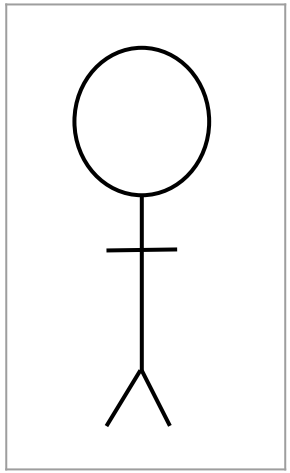
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$



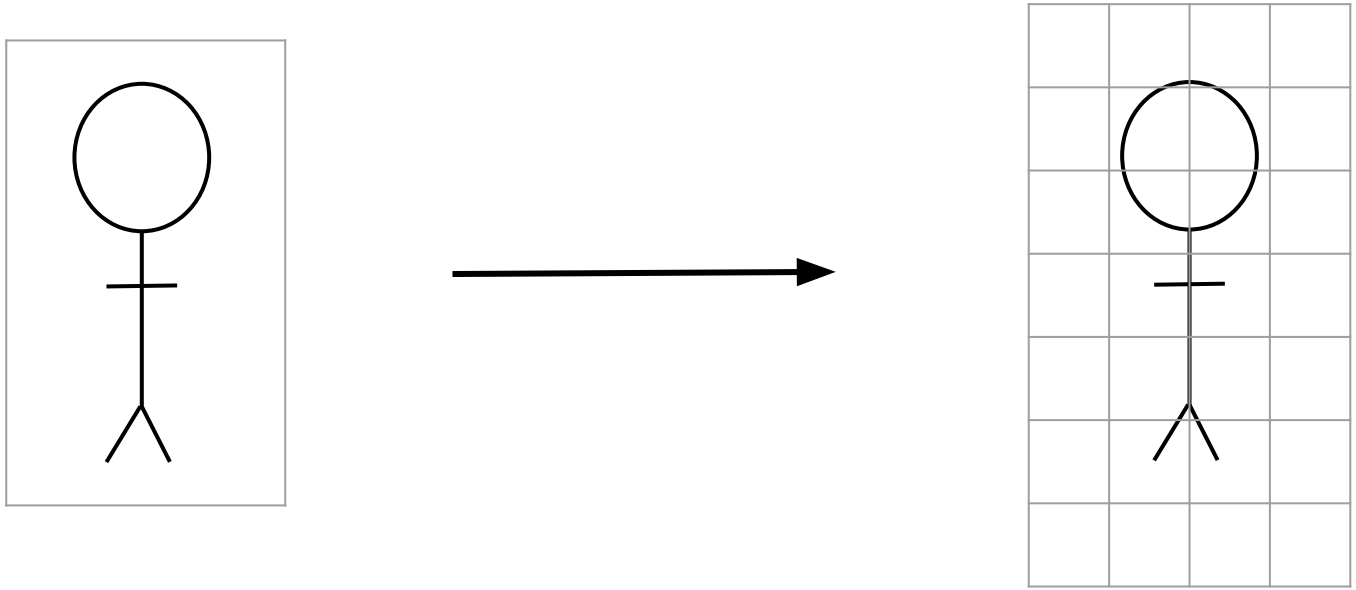
Adjacency List

1 : {2, 3}
2 : {1, 3}
3 : {1, 2}

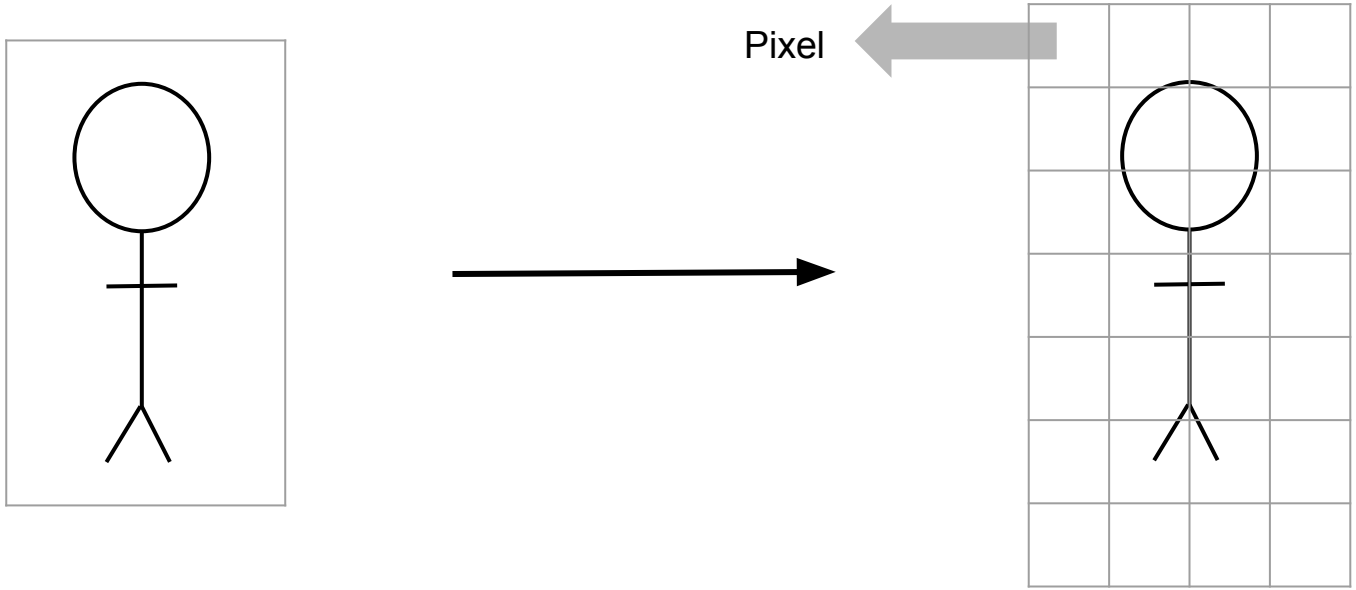
Types of Data - Images



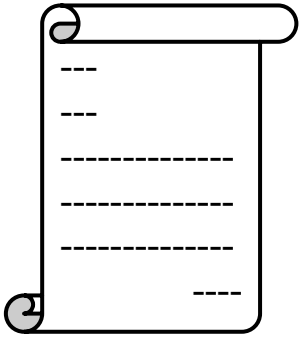
Types of Data - Images



Types of Data - Images

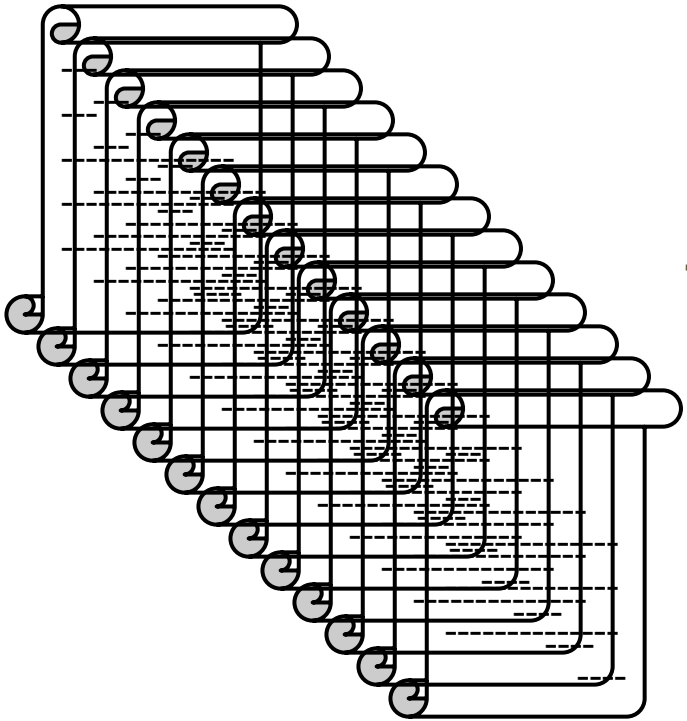


Types of Data - Text



List of words

Types of Data - Corpus of Documents



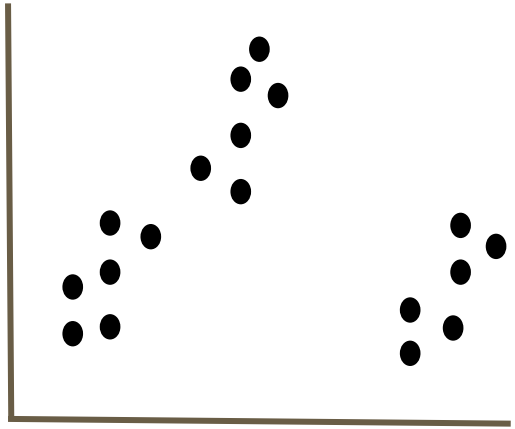
	w_1	w_2	…	w_m
D_1	1	0	…	1
D_2	0	0	…	0
…	…	…	…	…
D_n	1	1		1

Types of Learning

- Unsupervised Learning
- Supervised Learning

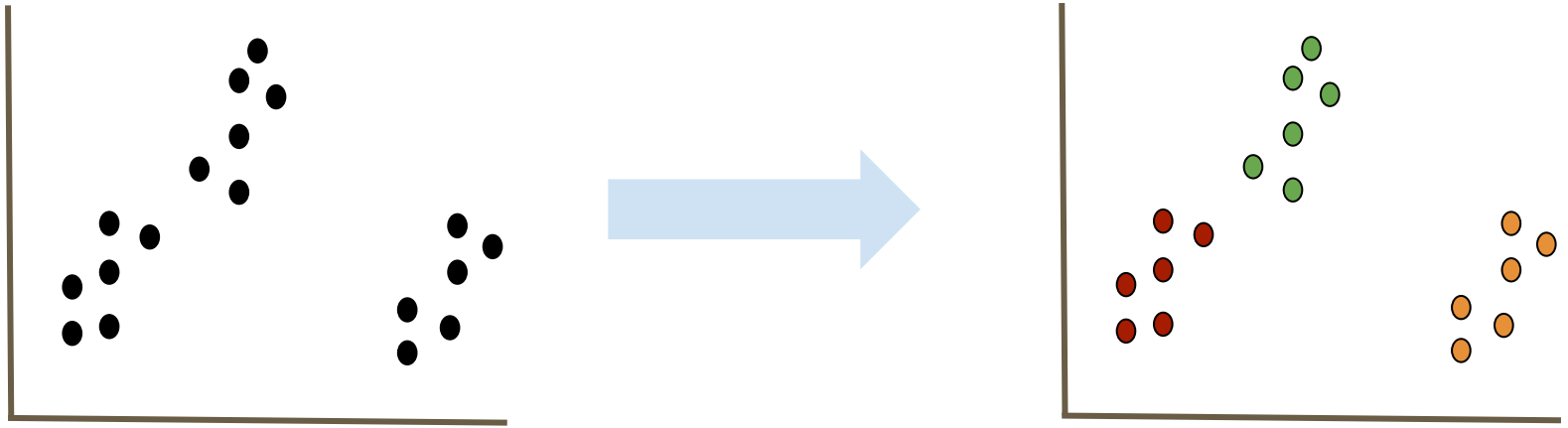
Unsupervised Learning

Goal: Find interesting structure in the data



Unsupervised Learning

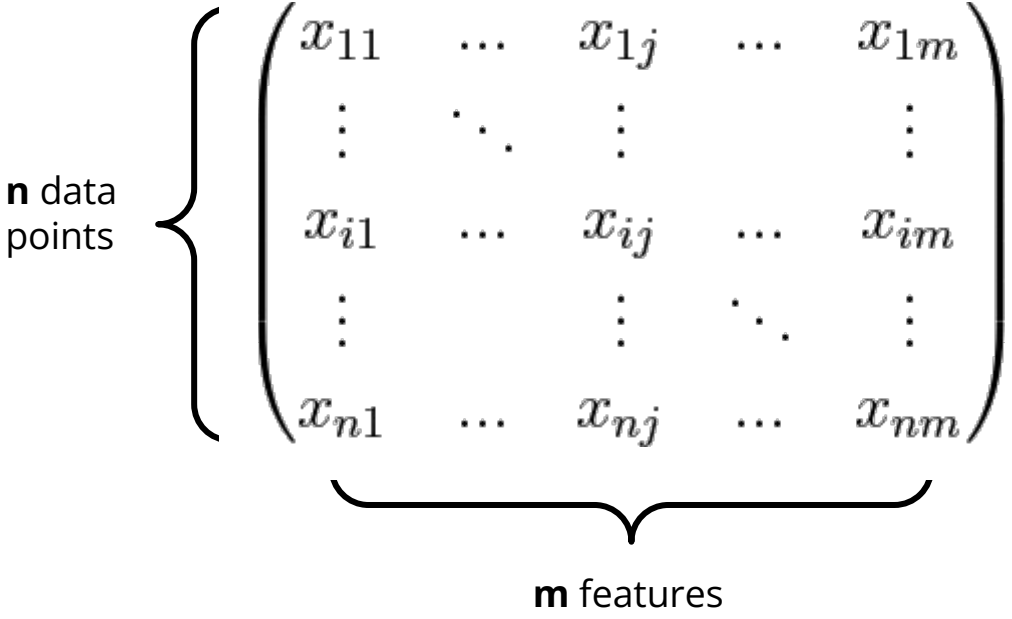
Goal: Find interesting structure in the data



This type of unsupervised learning is referred to as clustering

Unsupervised Learning

What are some linear algebraic properties of the matrix of data? What does that tell me about the data?



Unsupervised Learning

Dataset: Collection of Articles

Question: Are these articles covering the same topics?

Unsupervised Learning

Goals:

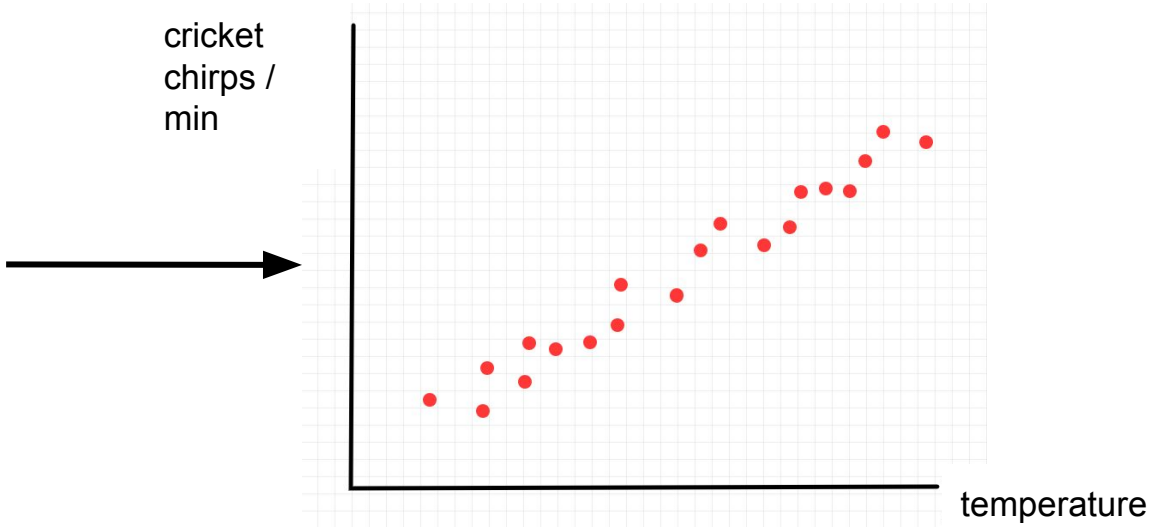
1. Better understand / describe the data
    a. Data exploration / visualization step
    b. Find anomalies
    c. Recommender Systems (similar users might be recommended the same things, emails similar to those marked as spam could be spam etc.)
2. Extract Features
3. Fill in gaps in data
    a. Data preprocessing step
4. Make learning algorithms faster
    a. Get rid of noise

Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78

Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78

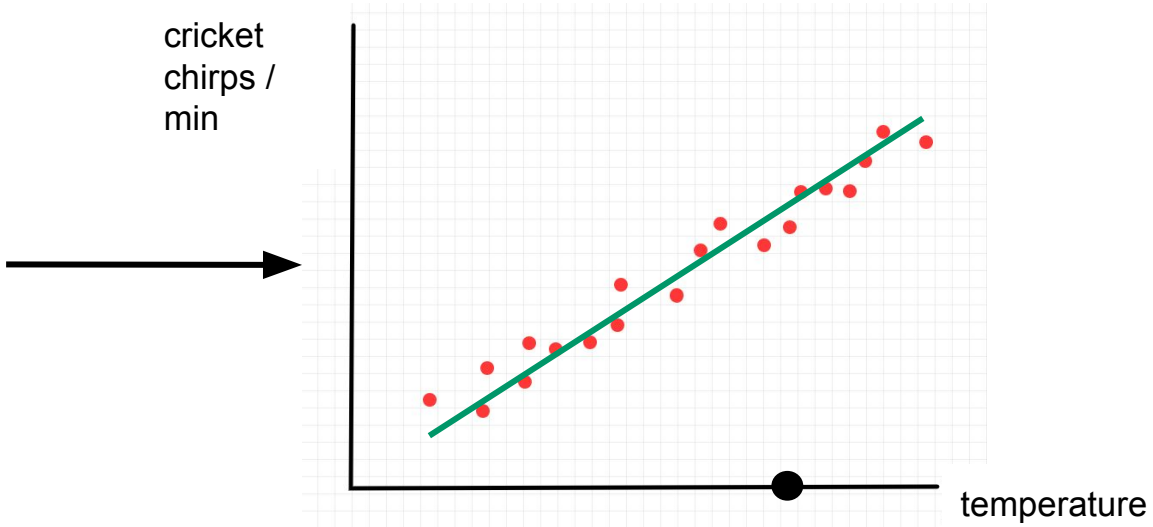


Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78

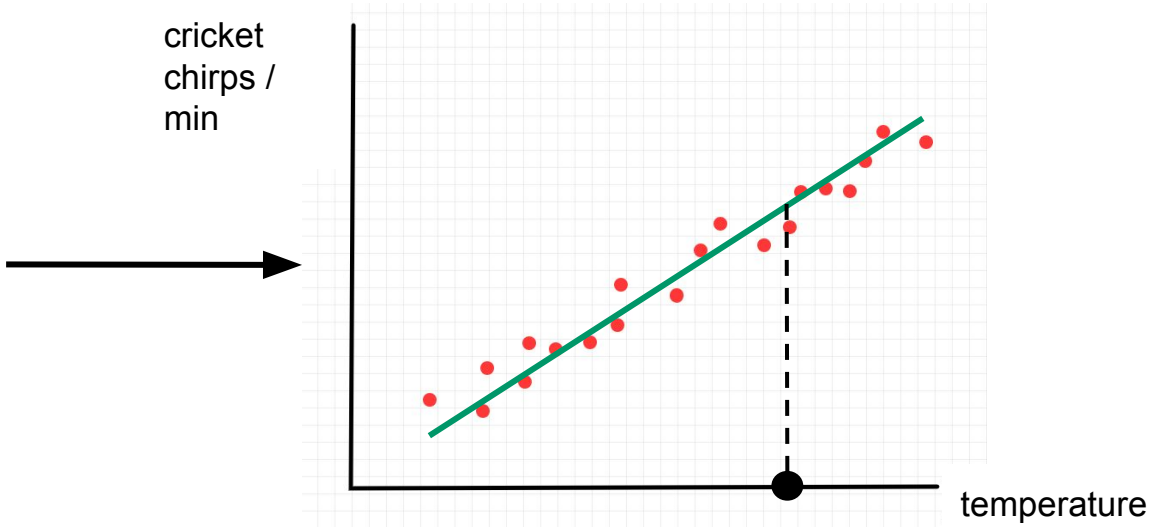
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



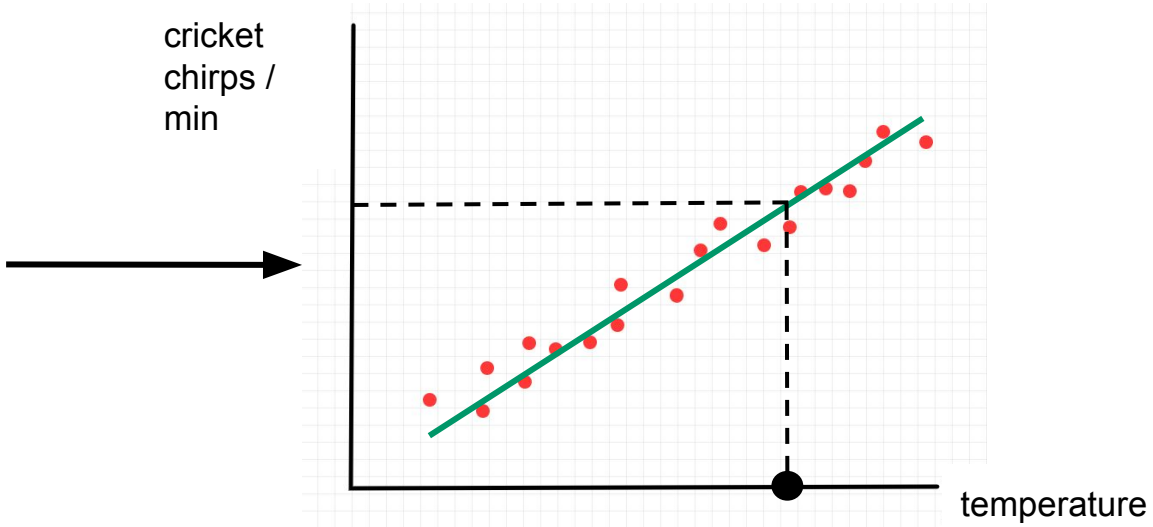
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



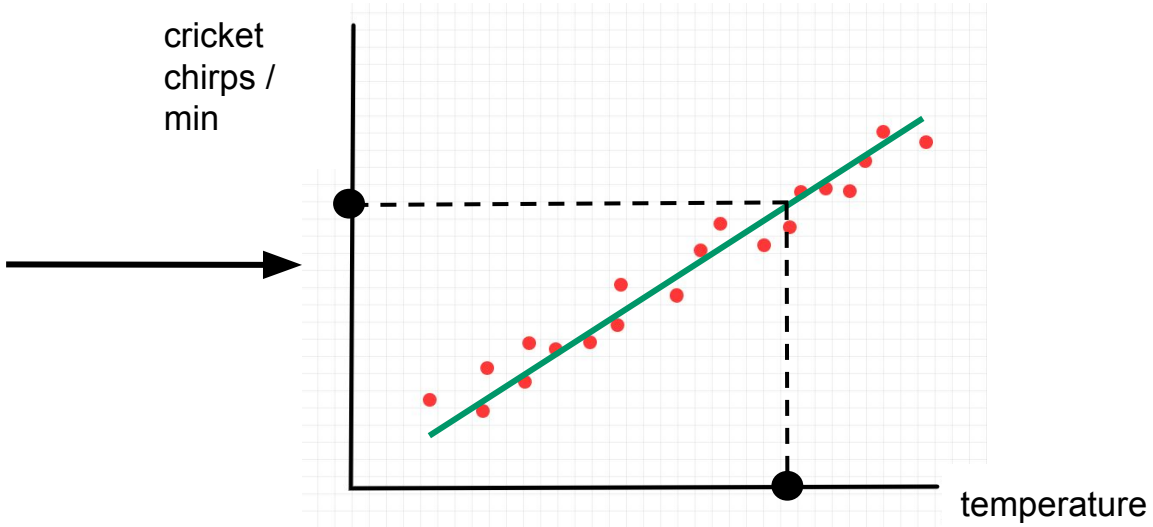
Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



Supervised Learning

cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



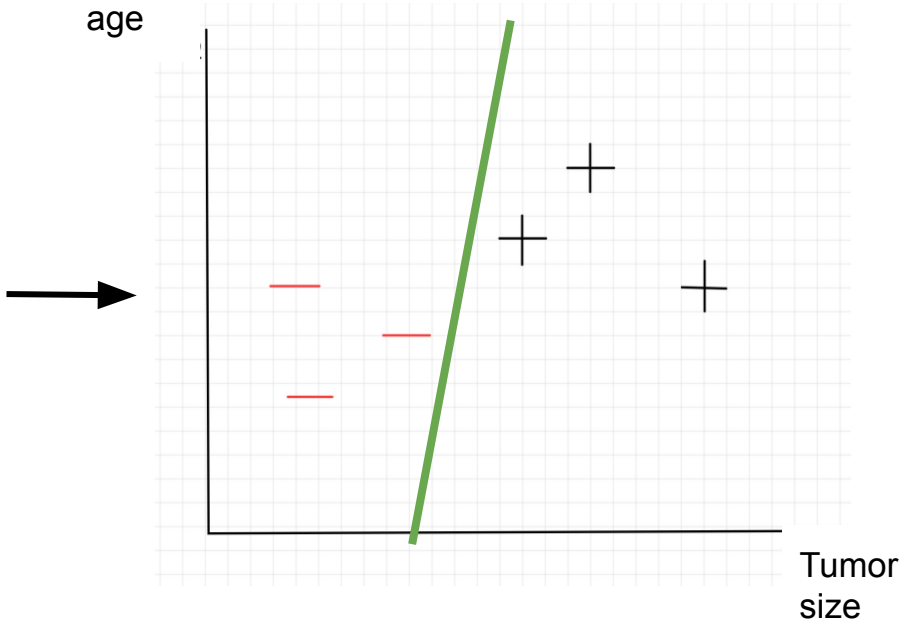
This type of supervised learning is referred to as regression

Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1

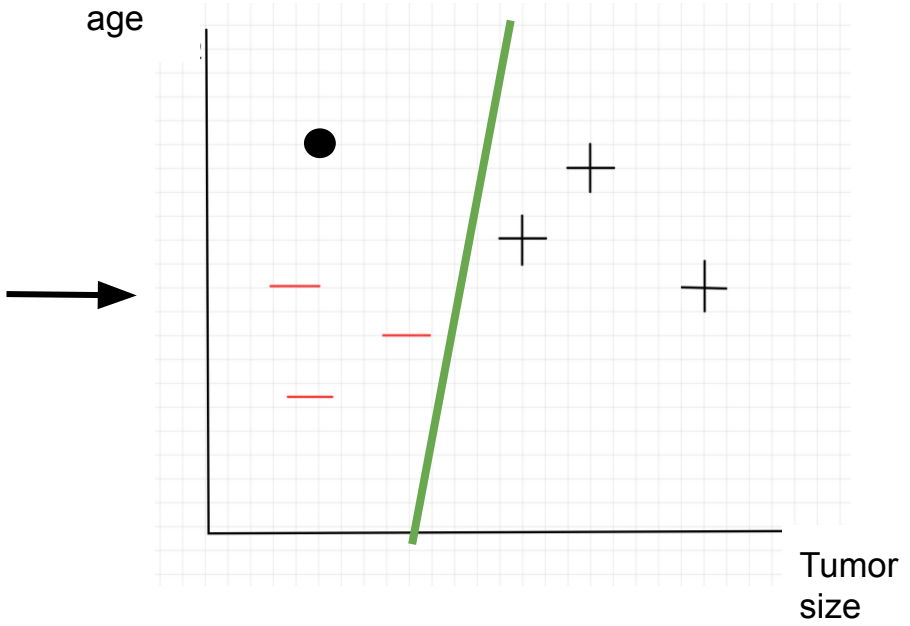
Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



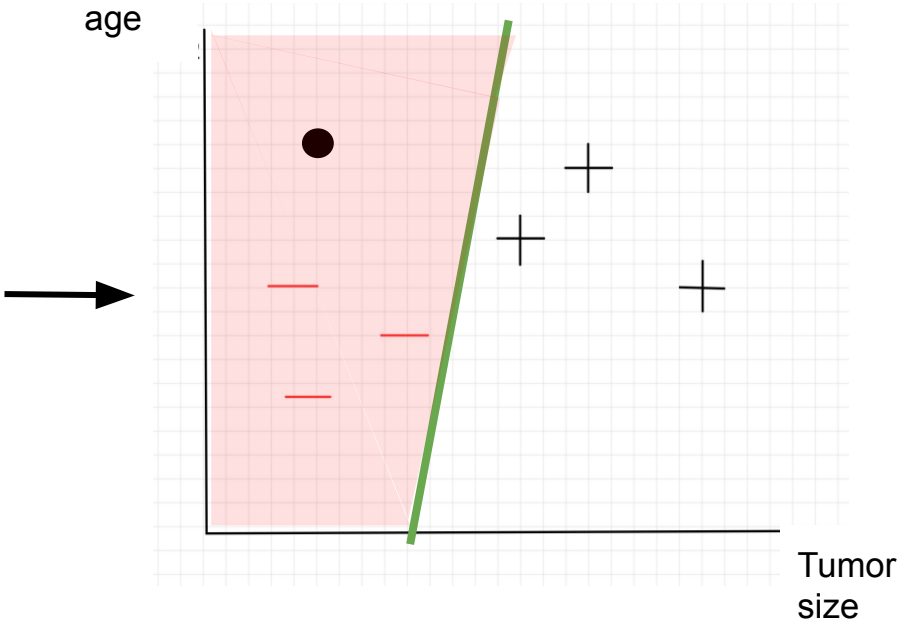
Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



Supervised Learning

age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



This type of supervised learning is referred to as classification