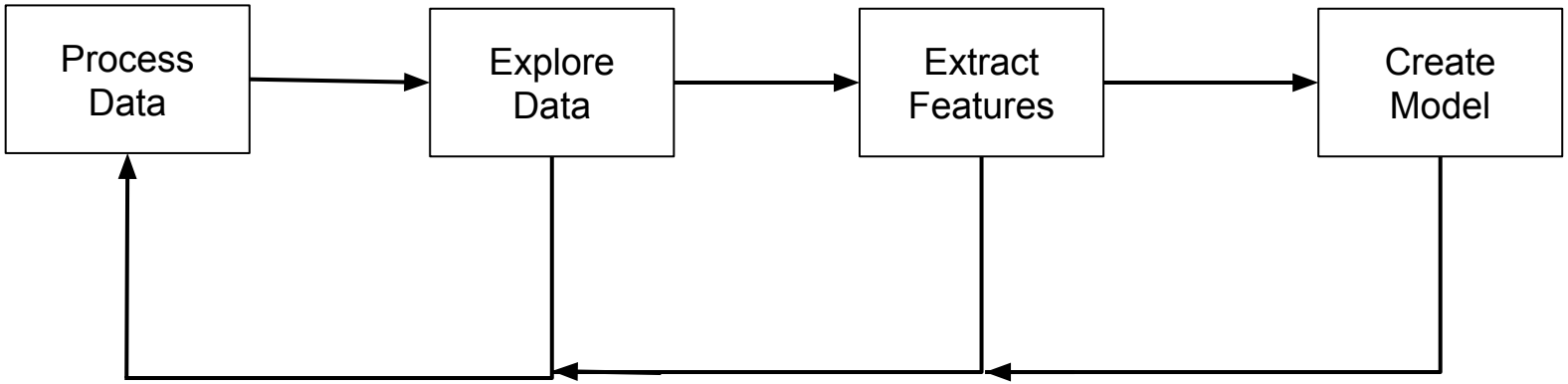
Introduction

Boston University CS 506 - Lance Galletti

Predict a student’s gpa

Data Science Workflow (simplified)



Types of Data

Types of Data - Records

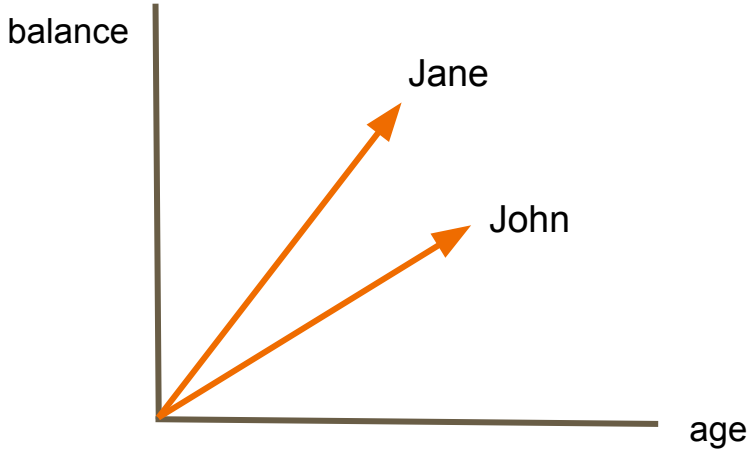
m-dimensional points / vectors

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

Types of Data - Records

m-dimensional points / vectors

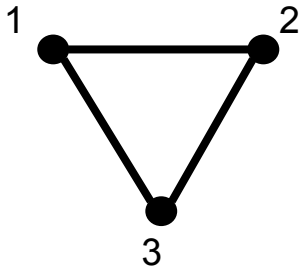
Example: (name, age, balance) -> (“John”, 20, 100)



Types of Data - Graphs

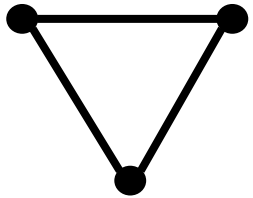
Nodes connected by edges

Example:



Adjacency Matrix

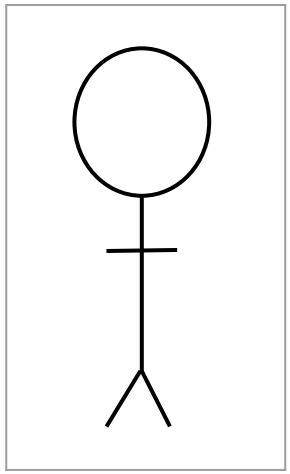
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$



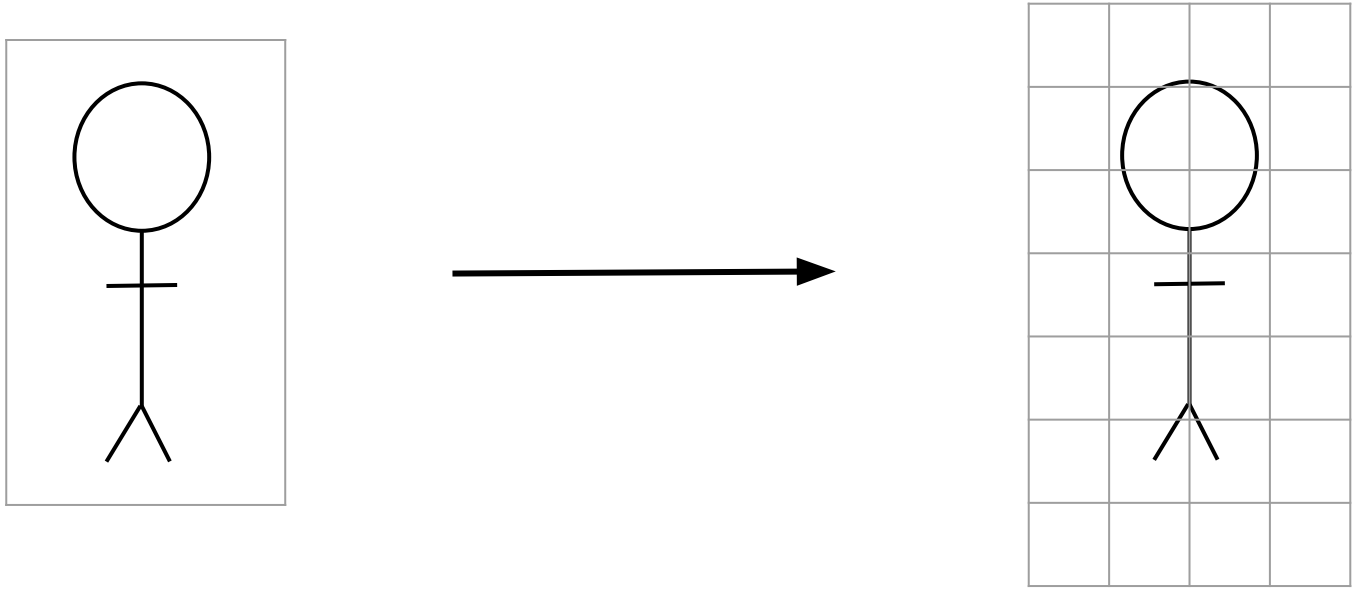
Adjacency List

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

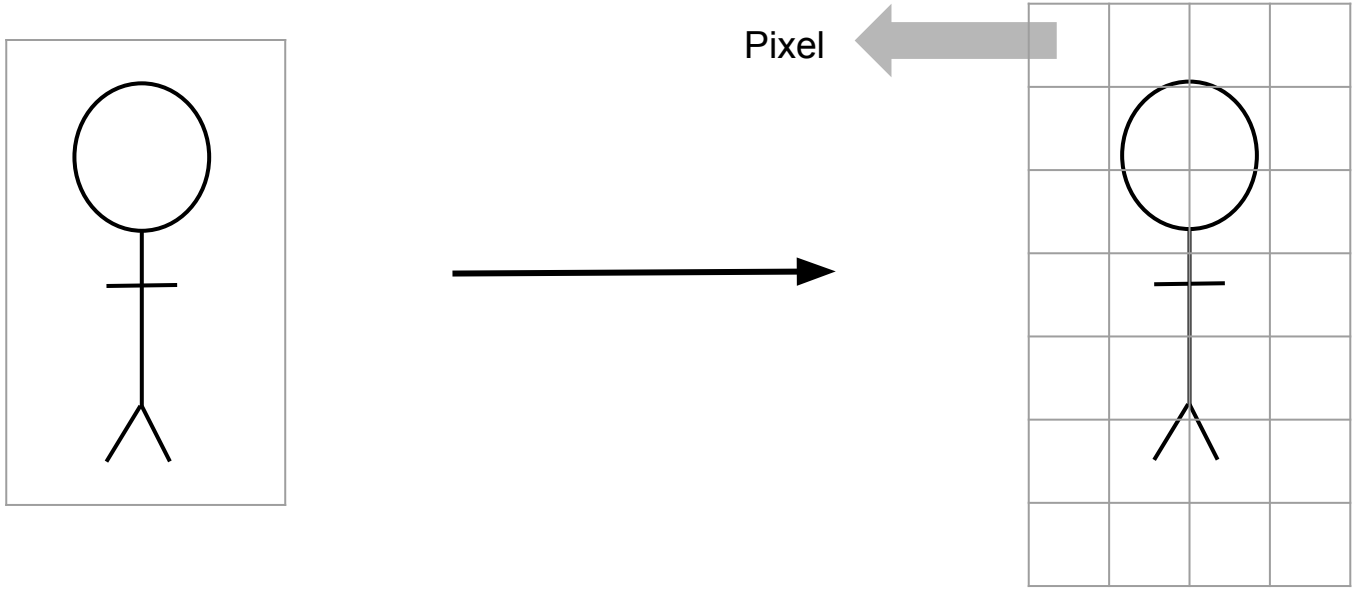
Types of Data - Images



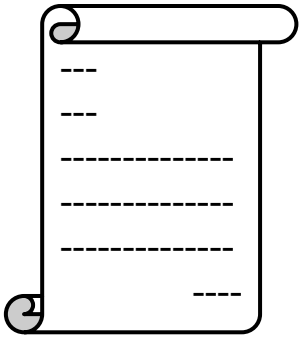
Types of Data - Images



Types of Data - Images

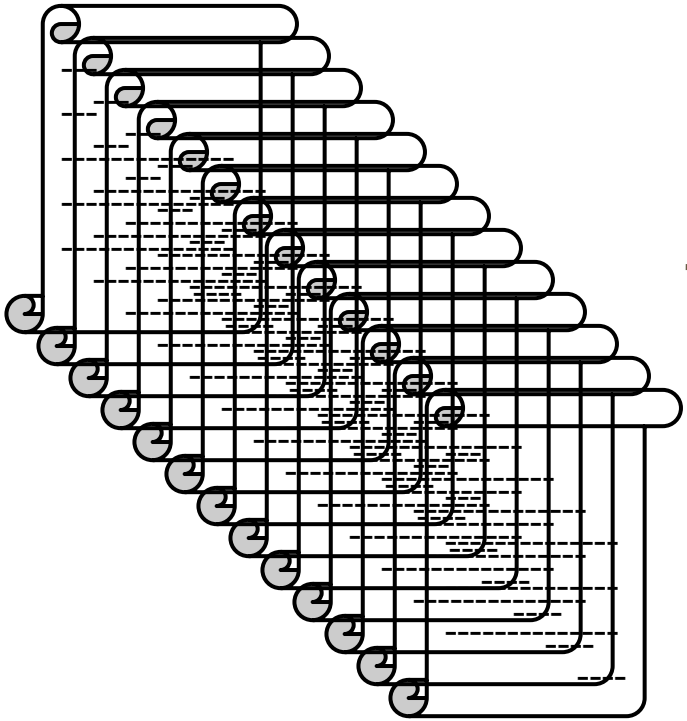


Types of Data - Text



List of words

Types of Data - Corpus of Documents



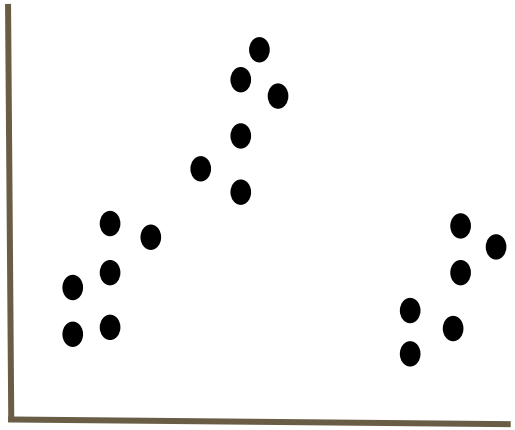
	w_1	w_2	…	w_m
D_1	1	0	…	1
D_2	0	0	…	0
…	…	…	…	…
D_n	1	1		1

Types of Learning

- Unsupervised Learning
- Supervised Learning

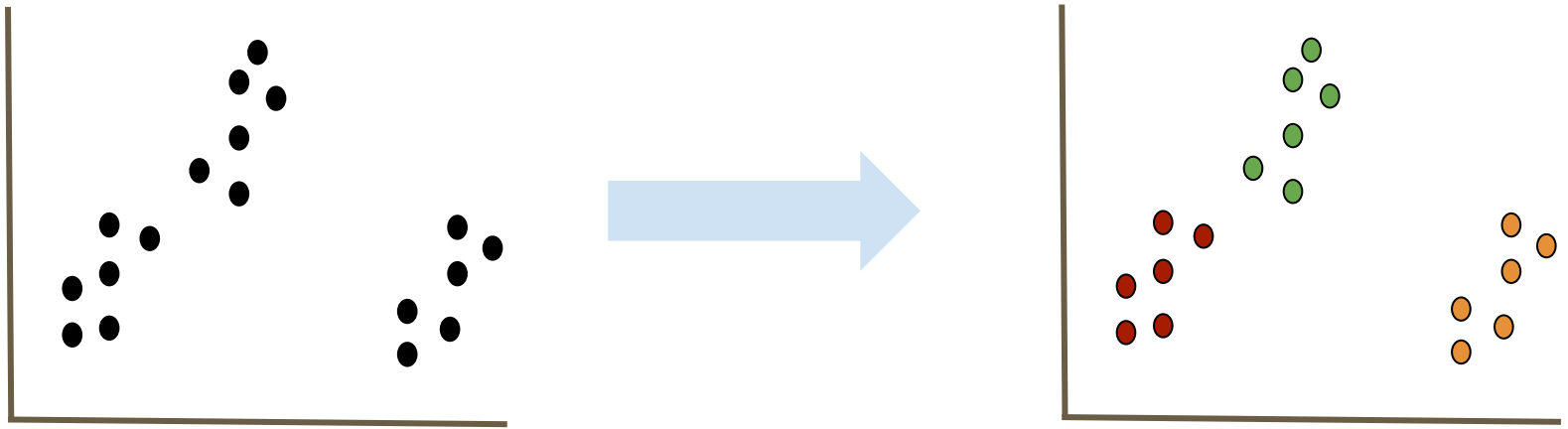
Unsupervised Learning

Goal: Find interesting structure in the data



Unsupervised Learning

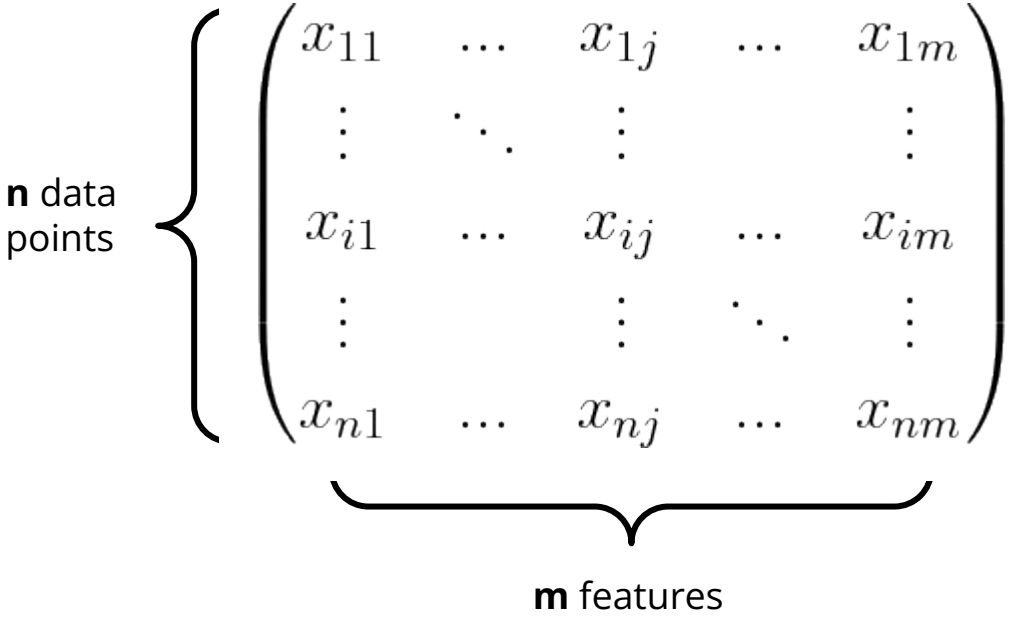
Goal: Find interesting structure in the data



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:

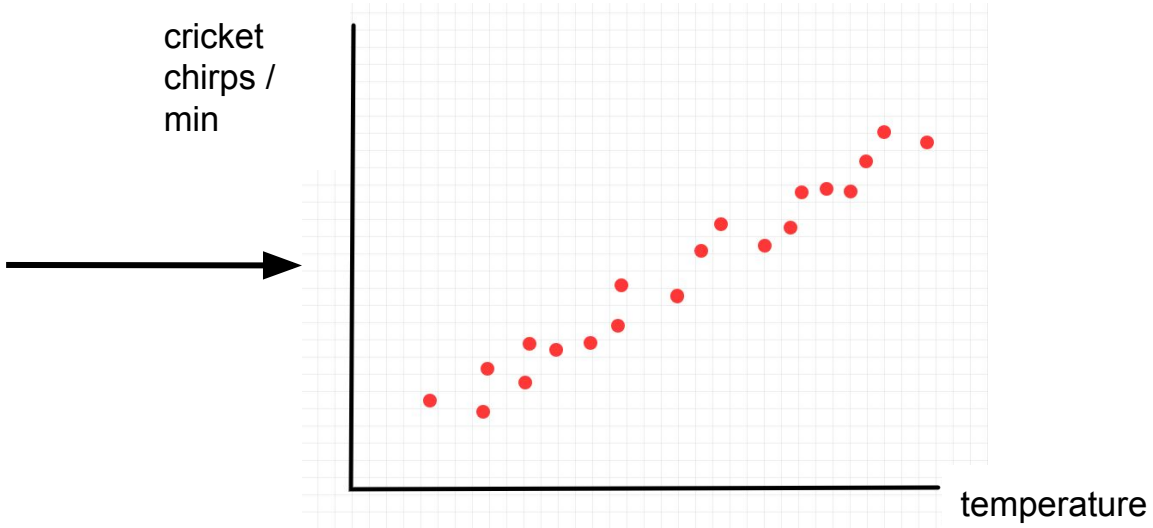
1. Better understand / describe the data
    a. Data exploration / visualization step
    b. Find anomalies
    c. Recommender Systems (similar users might be recommended the same things, emails similar to those marked as spam could be spam etc.)
2. Extract Features
3. Fill in gaps in data
    a. Data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

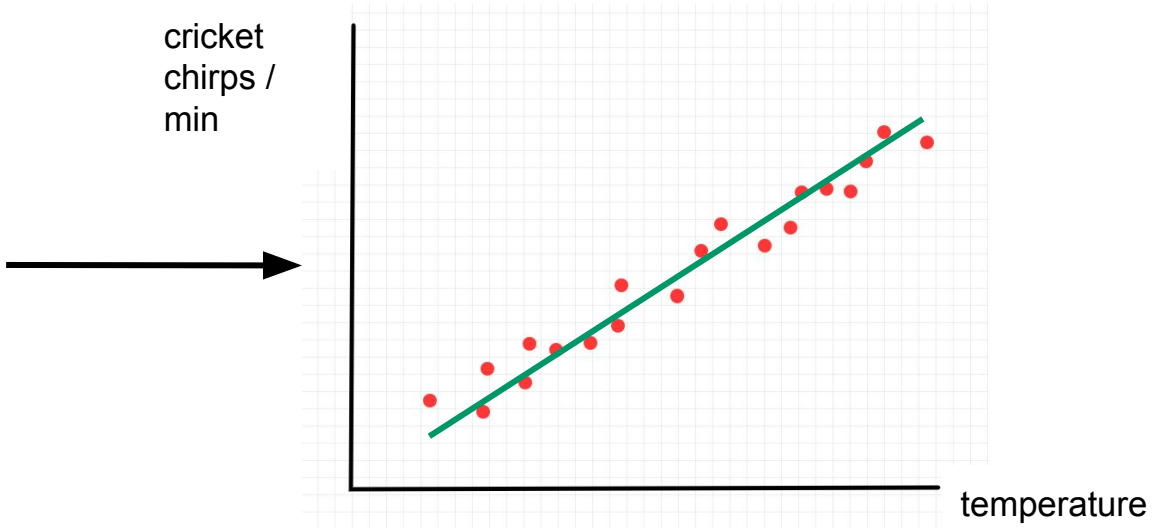
Supervised Learning

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



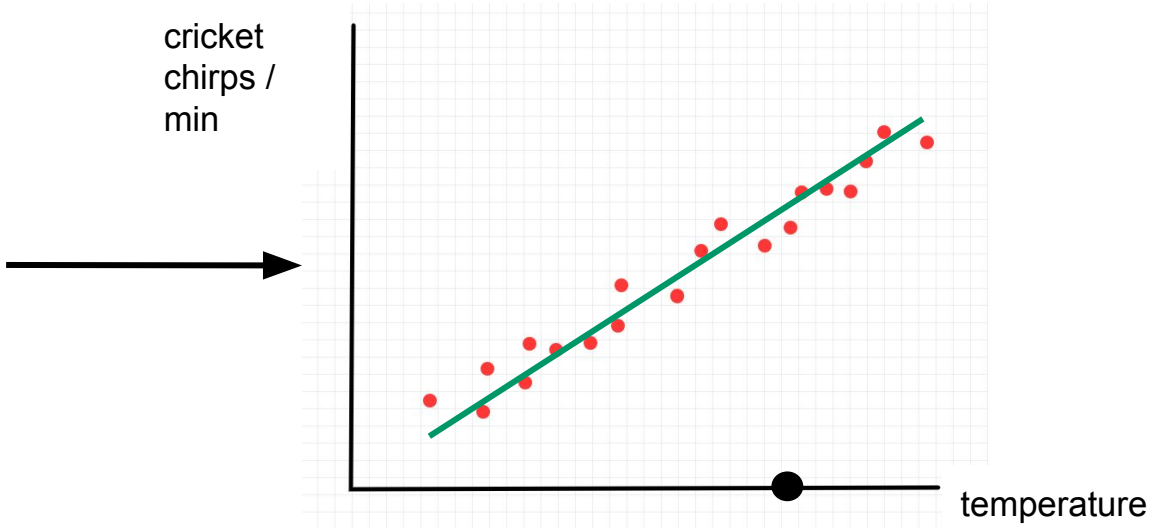
Supervised Learning

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



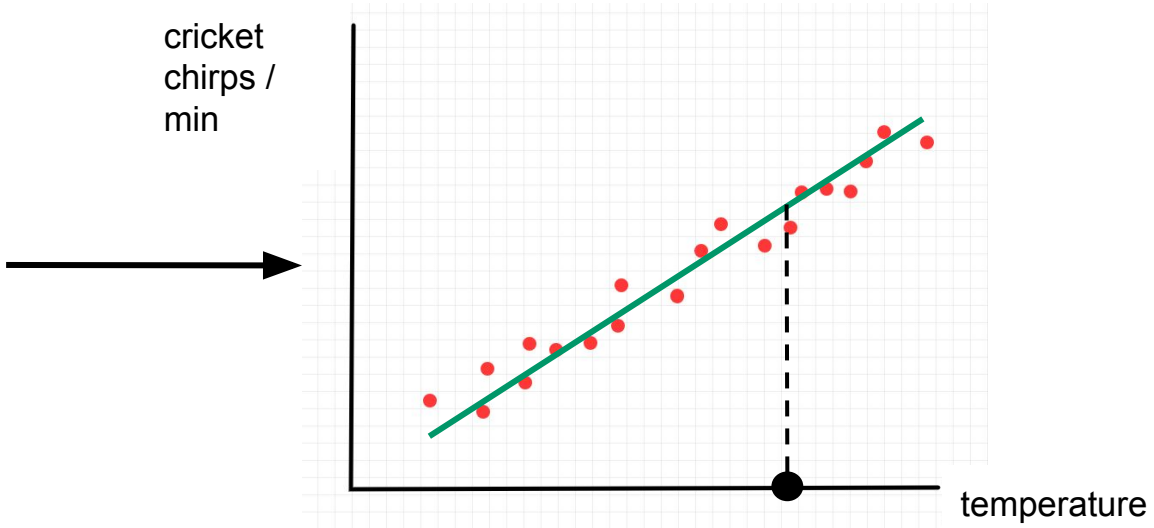
Supervised Learning

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



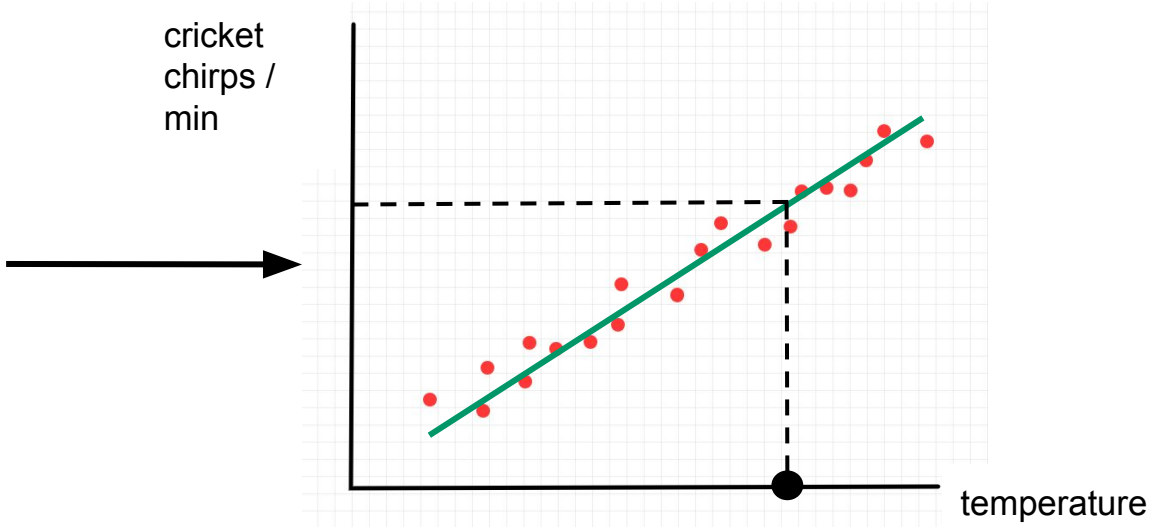
Supervised Learning

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



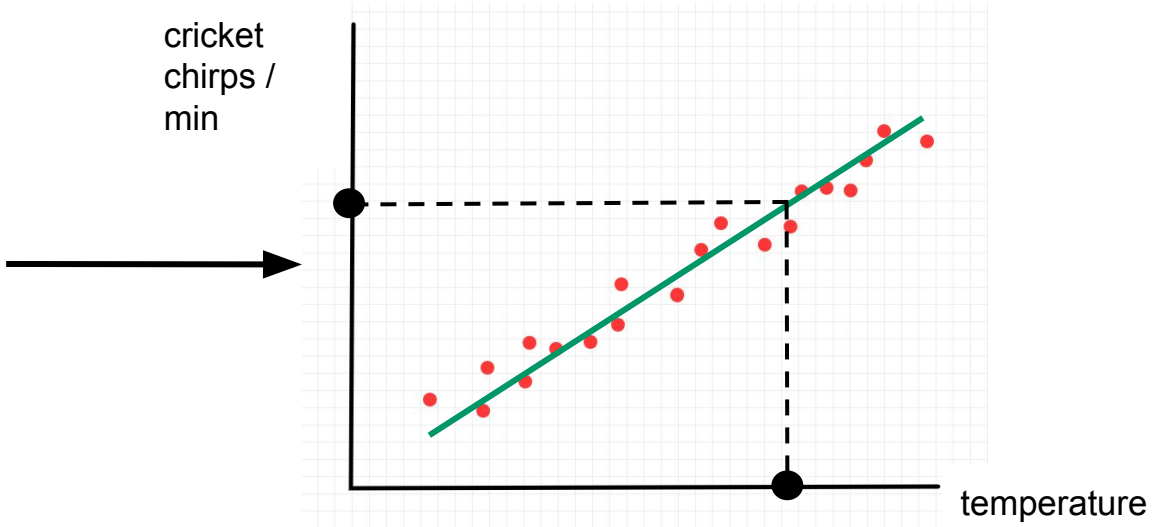
Supervised Learning

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



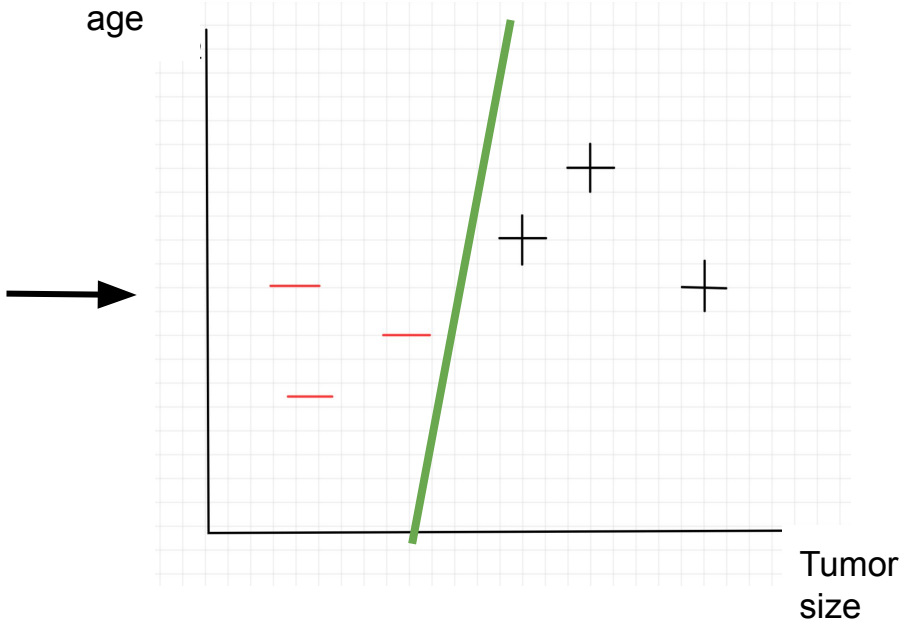
This type of supervised learning is referred to as regression

Supervised Learning

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

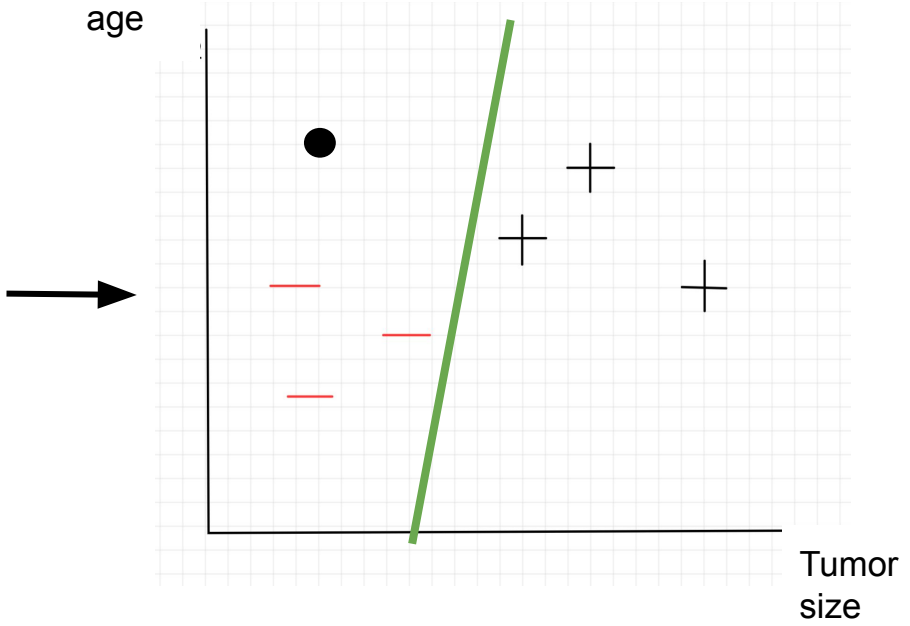
Supervised Learning

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



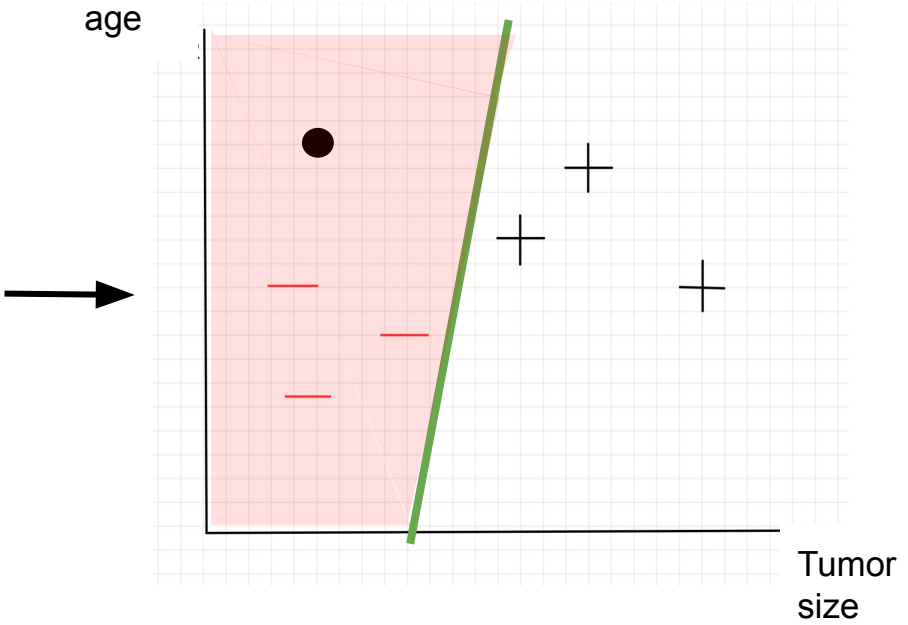
Supervised Learning

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