Lecture 2: Introduction

1) Data Science

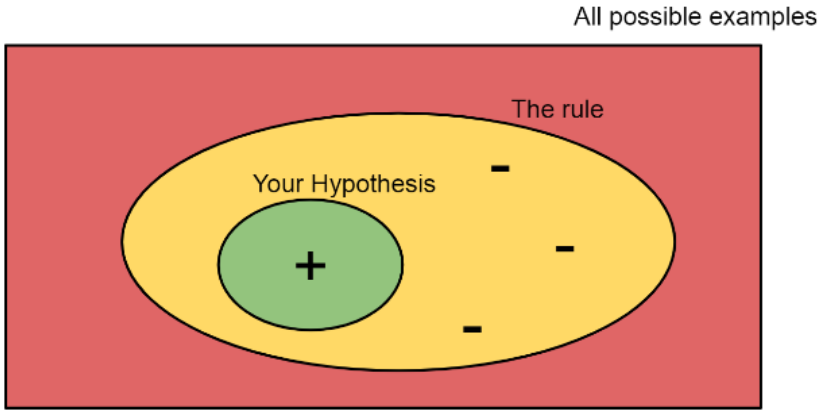
a) Collection of methods and tools that allow extracting knowledge from data

2) Confirmation Bias

a) Playing a game and announce: (2, 4, 6) follows the rule
   i) Examples submitted by one of the participants
      - (2, 4, 3) -> NO
      - (6, 8, 10) -> YES
      - (1, 3, 5) -> YES
   ii) Participants try to write down hypothesized rule

b) Challenges of Data Science
   i) Not all examples contribute to similar amounts of information
   ii) Set of examples may not always represent underlying rule
   iii) There may be infinitely many rules that match examples provided

c) Both positive and negative examples can falsify an example, but we have a tendency to choose positive ones over negative ones

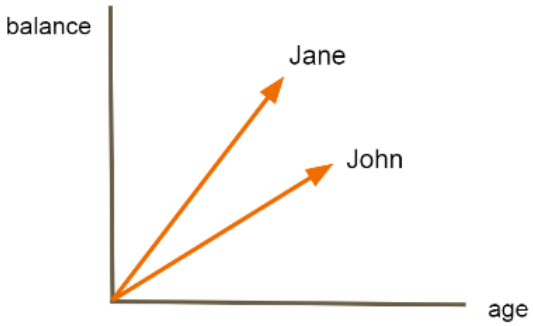


d) Rules was (a < b < c)
   i) If we only tried positive examples of either (x, x+2, x+4) and (x, 2x, 3x) we would only get confirmations

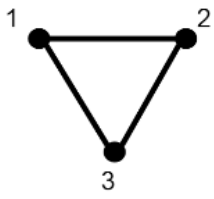
3) Type of Data

a) Records: m - dimensional points/vectors

i) Ex: (name, age, balance) -> ("John", 20, 100)

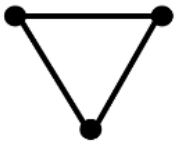


b) Graphs: nodes connected by edges



Adjacency Matrix

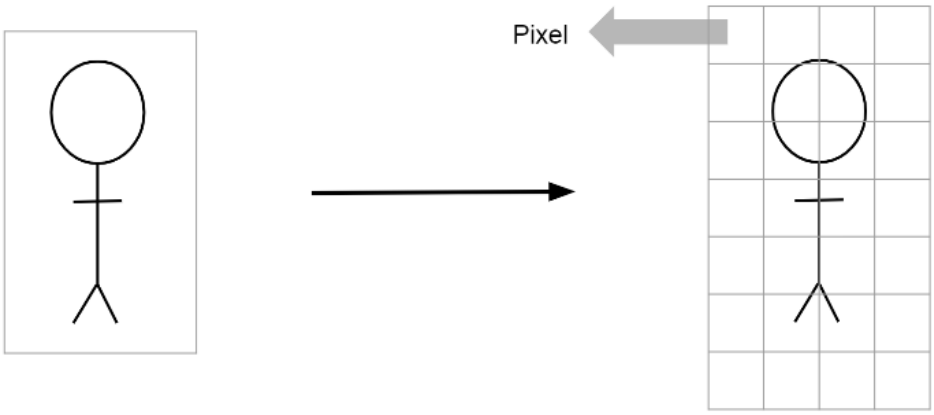
	1	2	3
1	0	1	1
2	1	0	1
3	1	1	0



Adjacency List

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

c) Images:

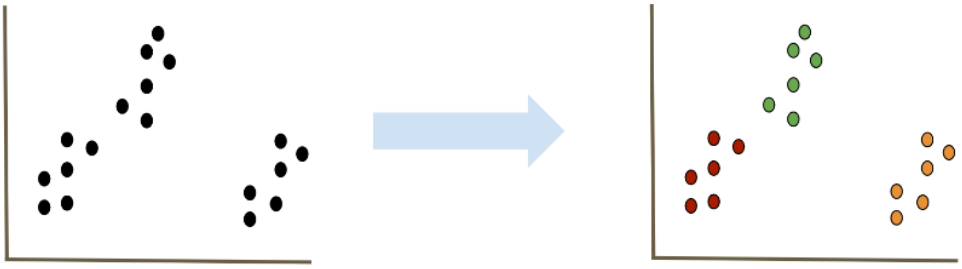


d) Text, Strings, Time Series: list of data at specific intervals of time

4) Types of Learning

a) Unsupervised Learning

i) Find interesting structure in data



This type of unsupervised learning is referred to as clustering

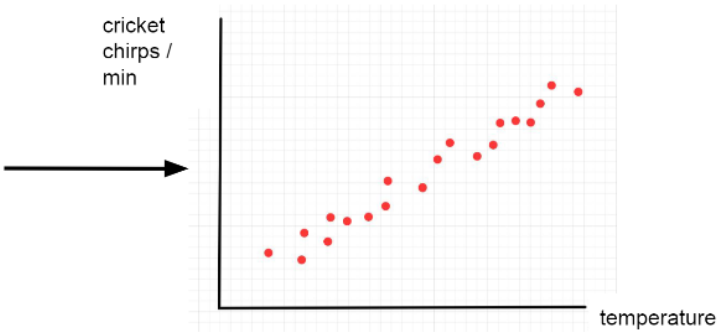
ii) Goals:

1) Better understand/describe data
   (a) Data exploration/visualization
   (b) Recommender Systems
2) Provide sensible defaults to missing data
   (a) Data preprocessing

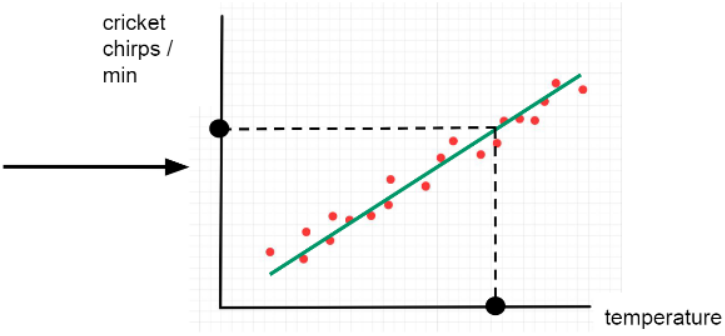
b) Supervised Learning

i) Ex1:

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



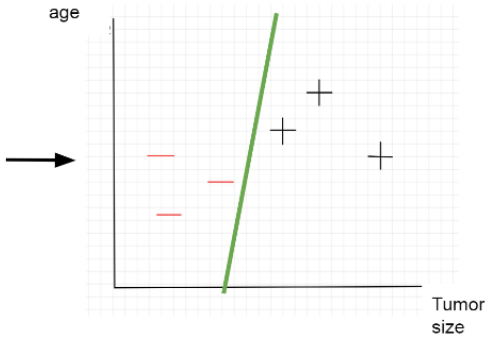
cricket chirps / min	temperature
10	40
5	37
17	53
55	103
40	78



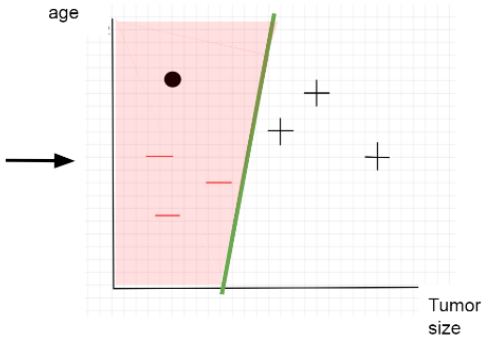
This type of supervised learning is referred to as regression

ii) Ex2:

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

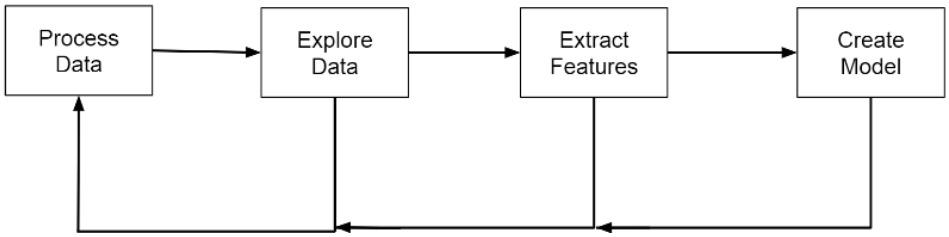


age	tumor size	malignant
20	12	0
22	15	1
47	20	1
59	2	1



This type of supervised learning is referred to as classification

5) Data Science Workflow



1) Data Processing:
    a) What should be done with data, cleaning data, etc
2) Exploratory Data Analysis
    a) Describe, conteculaize, and visualize data
        i) What can be predicted?
3) Feature Extraction
    a) What features can be extracted and what are the best ones?
4) Finding the right model
    a) What and who the model is intended for
    b) Success of current step depends entirely on work done in previous steps
        i) Garbage in and garbage out
    c) Is the model easy to explain and if it fails can you explain why?