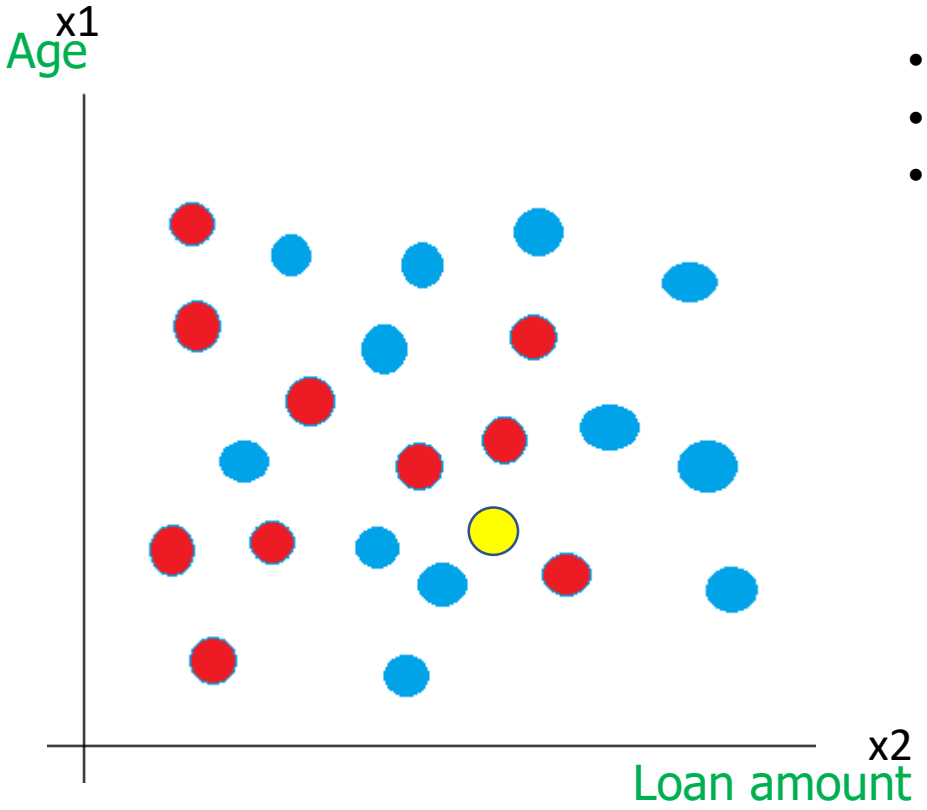
k-NN
k-Nearest Neighbours

k-NN – an illustration



- Let X1 and X2 be two variables eg: Age and Loan amount
- Red represents "loan defaulter"
- Blue represents "loan paid"

Predict the yellow record ?

loan defaulter / **loan paid**

Selecting the number of neighbours – optimum value for 'k'

- Inspect the data
- A large **'k'** value gives a better result
- Perform cross-validation with different 'k' values to get the best **'k'**
- Industry standard / Best practice methodology for the value of **k** is:

  3 <= k <= 10
- Select an odd 'k' to avoid tie and random selection

Exercise

Given the following dataset, predict if the given record of a customer will default the loan or not?
Take 'Neighbours' = 5

$(48-25)^2 + (91-40)^2$

Obs	Age	Loan_amt	Default	Distance
1	25	40	N	55.95
2	27	45	Y	50.57
3	35	70	N	24.70
4	37	68	Y	25.50
5	29	55	N	40.71
6	40	14	N	77.41
7	43	67	N	24.52
8	52	90	Y	**4.12**
9	34	58	Y	35.85
10	38	77	Y	17.20
11	37	85	Y	**12.53**
12	33	79	Y	19.21
13	27	20	N	74.04
14	28	16	N	77.62
15	26	10	N	83.93
16	41	90	Y	**7.07**
17	53	55	Y	36.35
18	49	80	N	**11.05**
19	47	67	N	24.02
20	40	77	Y	**16.12**
21	**48**	**91**	**????**	

Neighbours (k)	Default
4.12	Y
7.07	Y
11.05	N
12.53	Y
16.12	Y
17.20	
19.21	
24.02	
24.52	
24.70	
25.50	
35.85	
36.35	
40.71	
50.57	
55.95	
74.04	
77.41	
77.62	
83.93	

Prediction (48,91)
Y