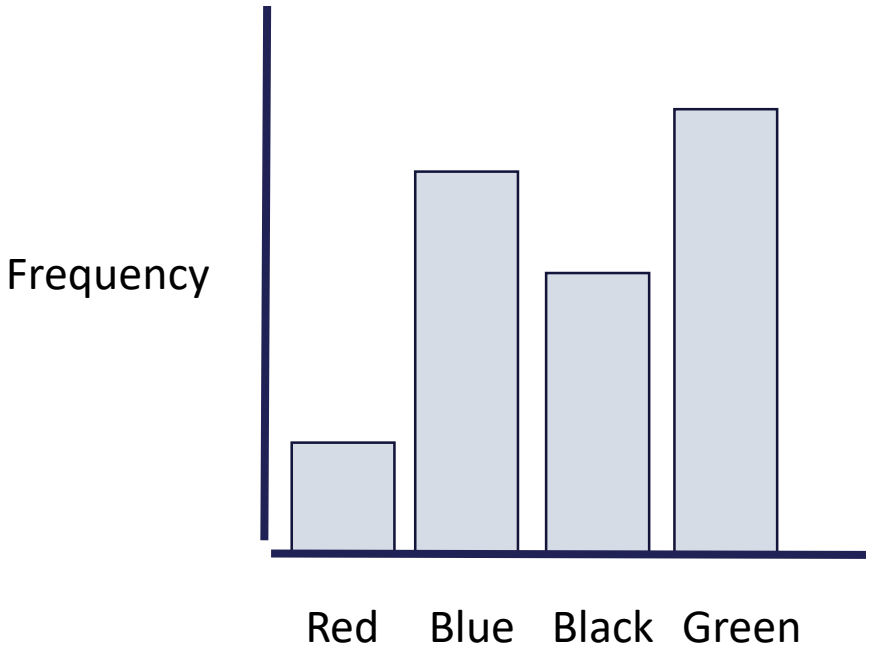
Synthetic tabular data

Tabular data

Colour	Age	Income	Car make	Nr. Cards
Red	65	50000	Ford	5

Tabular data – categorical variables

Colour	Age	Income	Car make	Nr. Cards
Red	65	50000	Ford	5

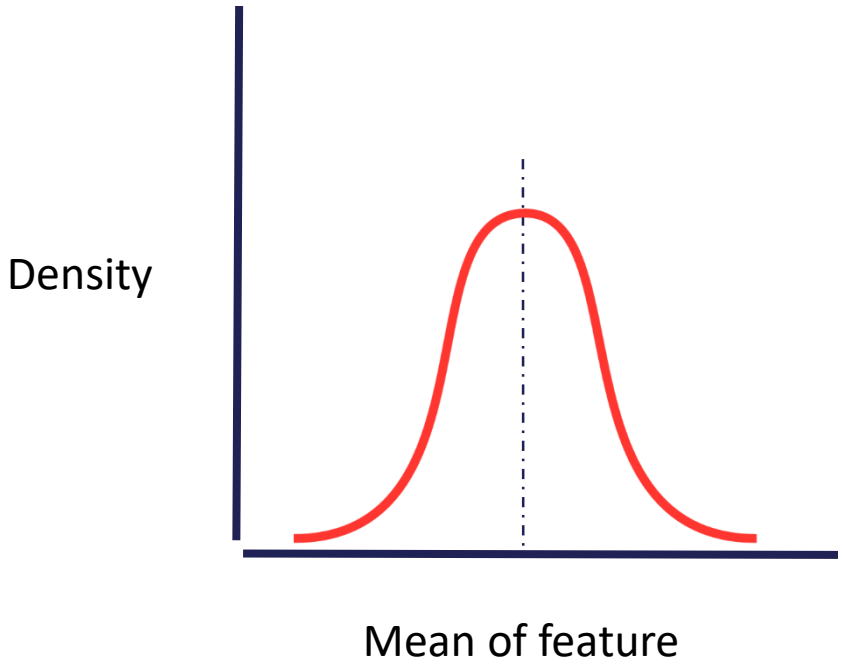


Sample based on frequency

Colour	Age	Income	Car make	Nr. Cards
Blue			Ford	
Red			Fiat	
Black			Ford	
Green			VW	
Green			Ford	

Tabular data – numerical variables

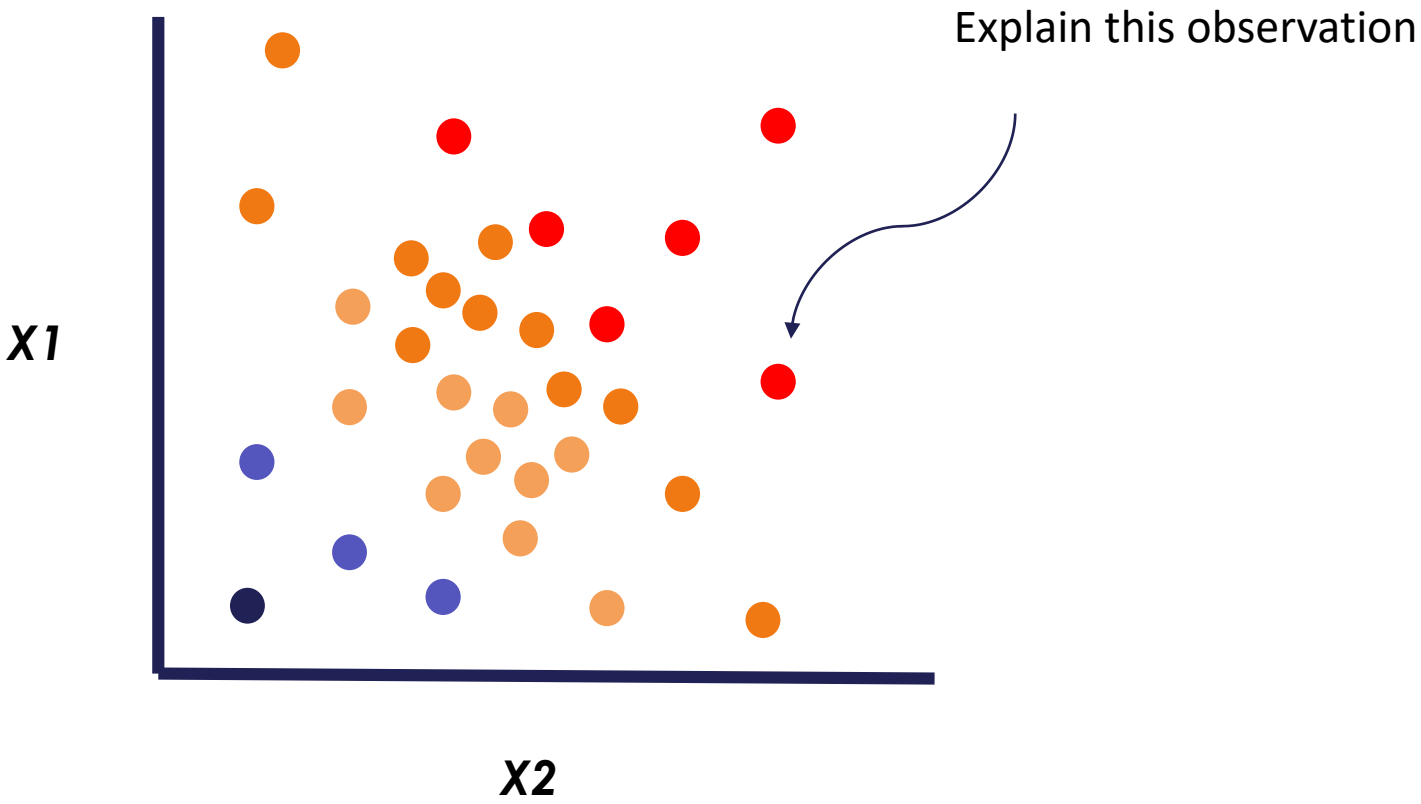
Colour	Age	Income	Car make	Nr. Cards
Red	65	50000	Ford	5



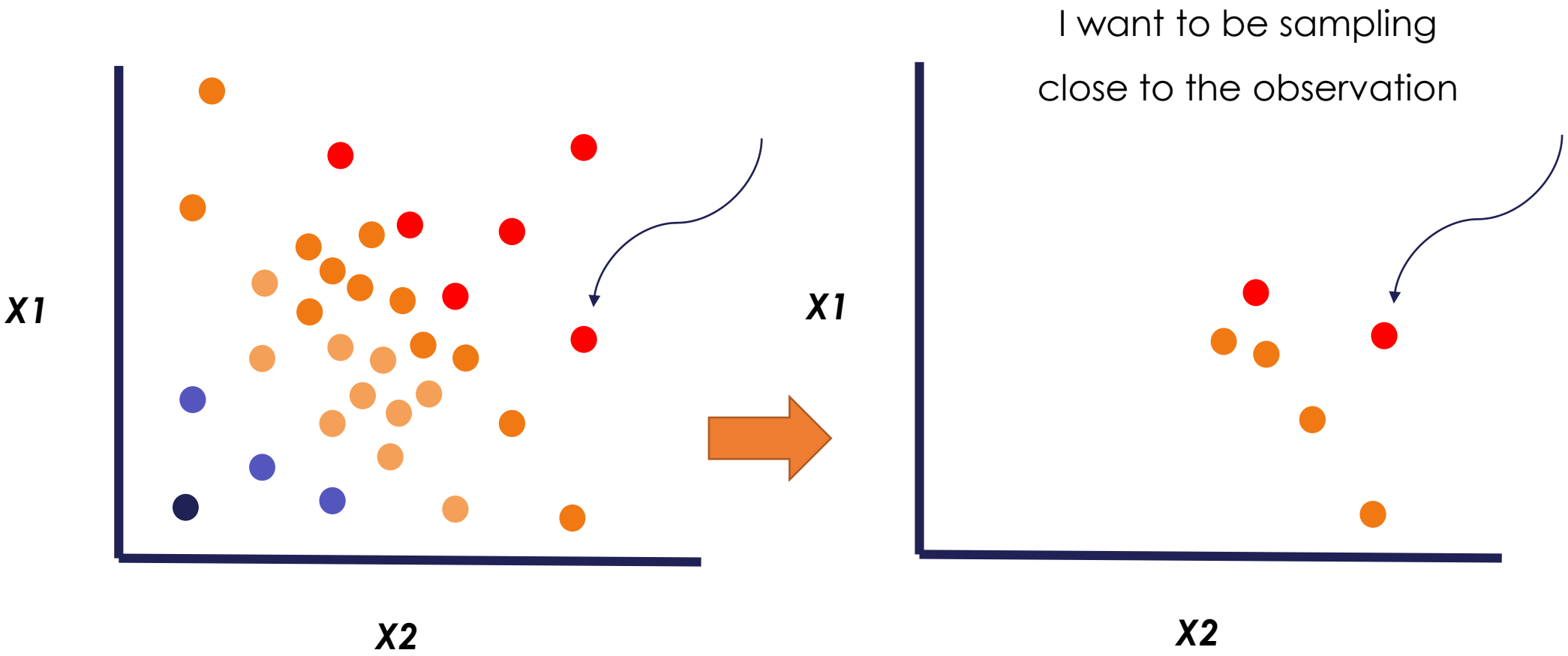
Sample based
N~(0,1)

Colour	Age	Income	Car make	Nr. Cards
Blue	65	51000	Ford	5
Red	63	59000	Fiat	5.2
Black	60	55000	Ford	5.9
Green	58	45000	VW	6
Green	55	47000	Ford	4

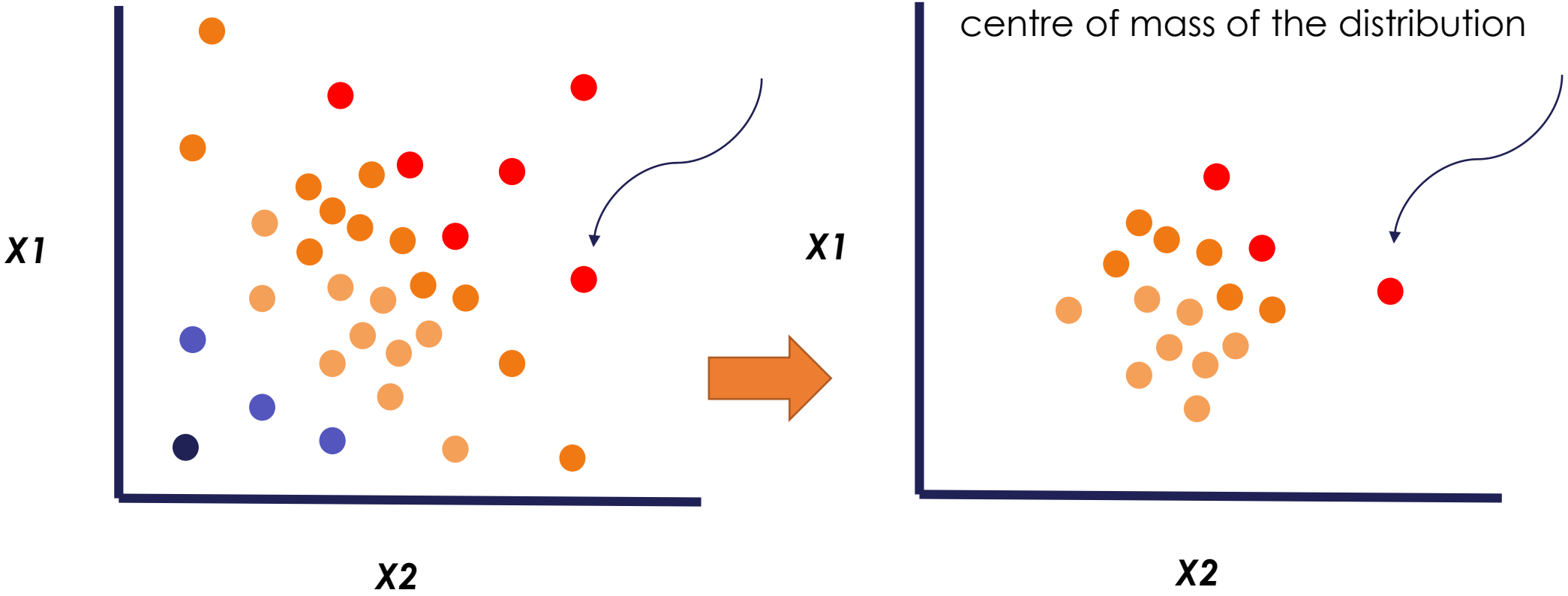
Tabular data - problematic



Tabular data - problematic

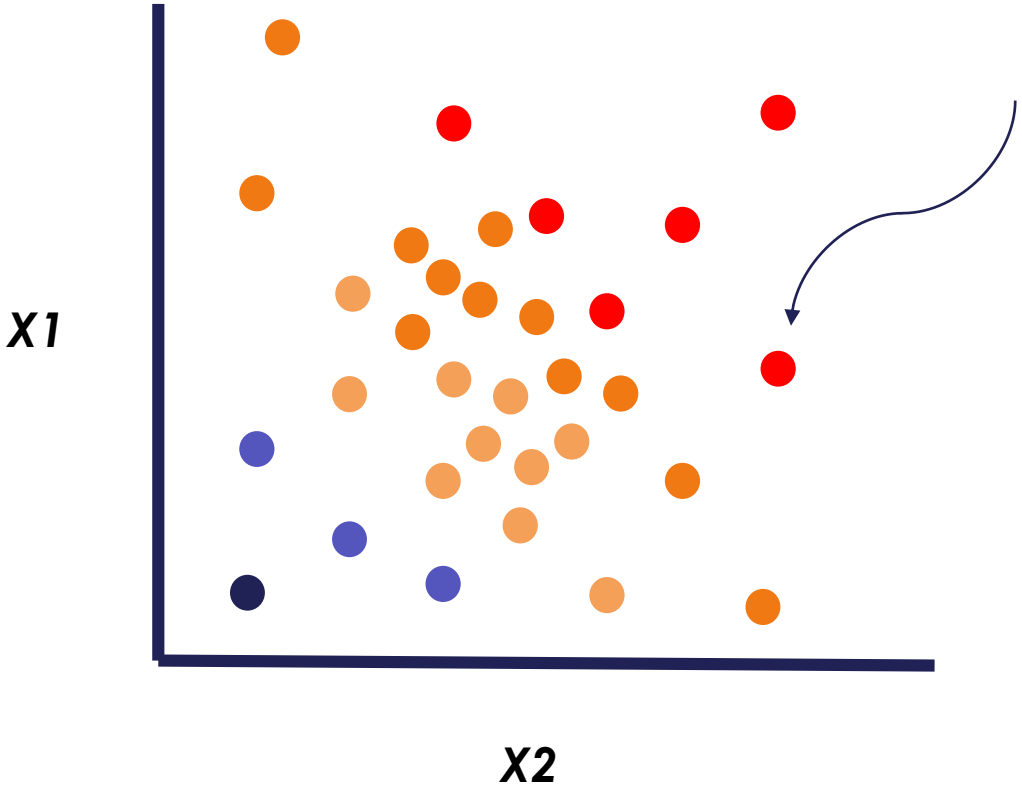


Tabular data - problematic



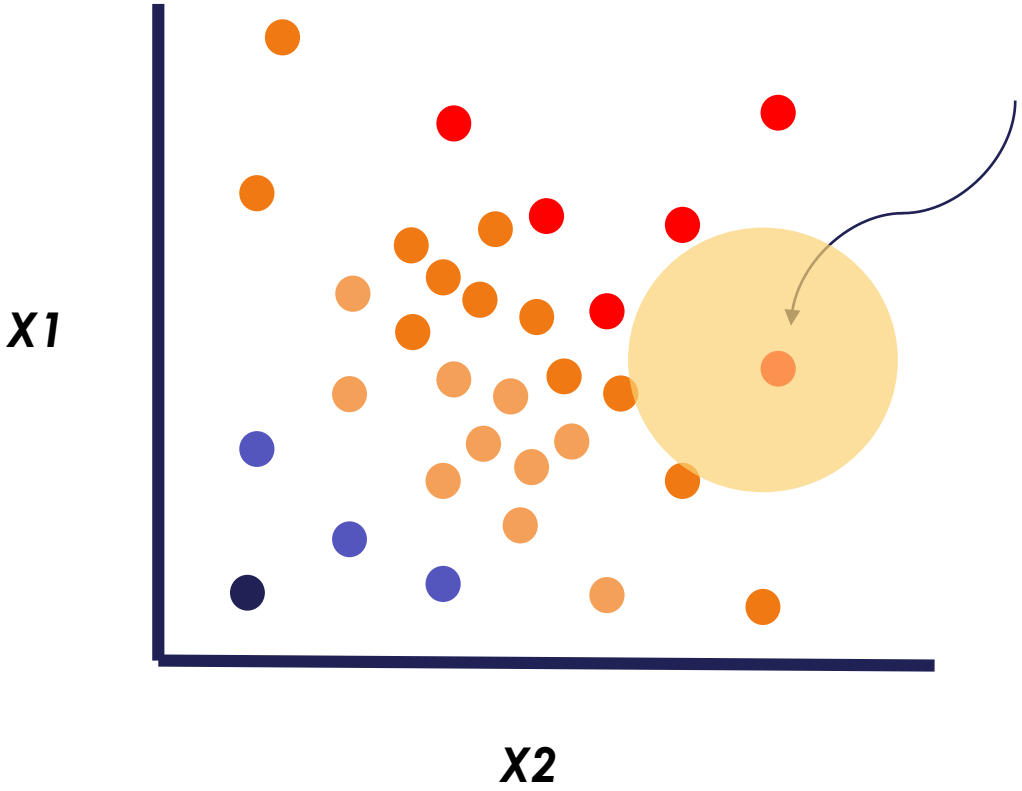
Yet, I will be sampling at the centre of mass of the distribution

Tabular data - problematic



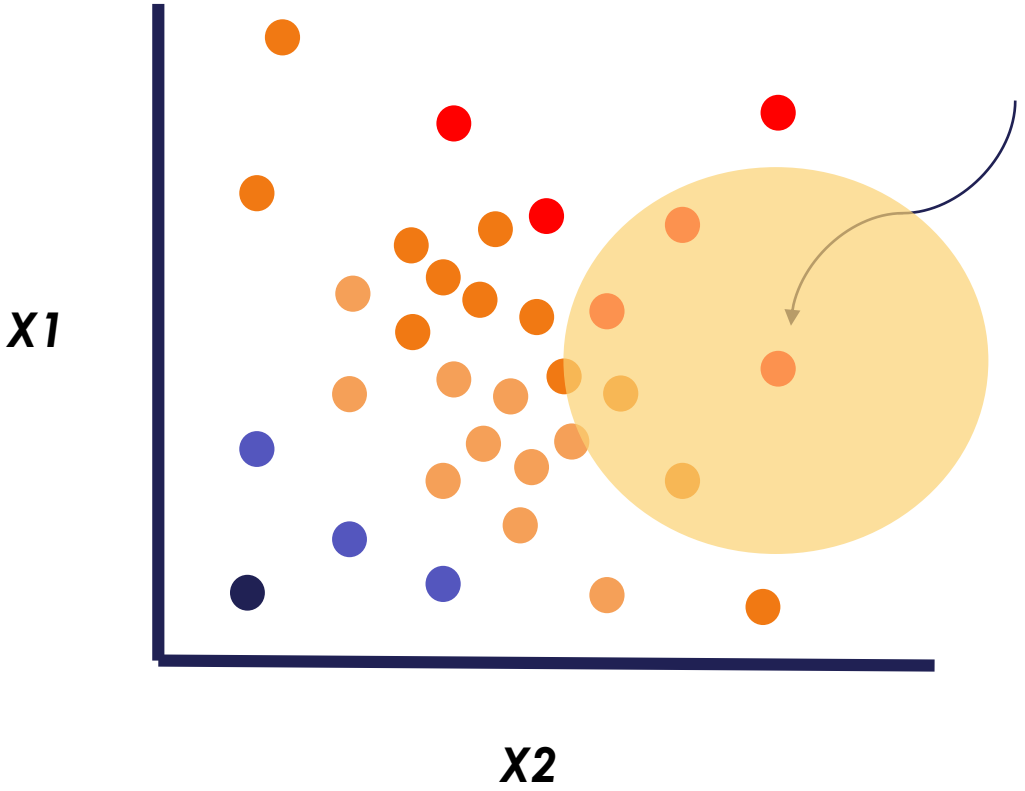
What is vicinity?

Tabular data - problematic



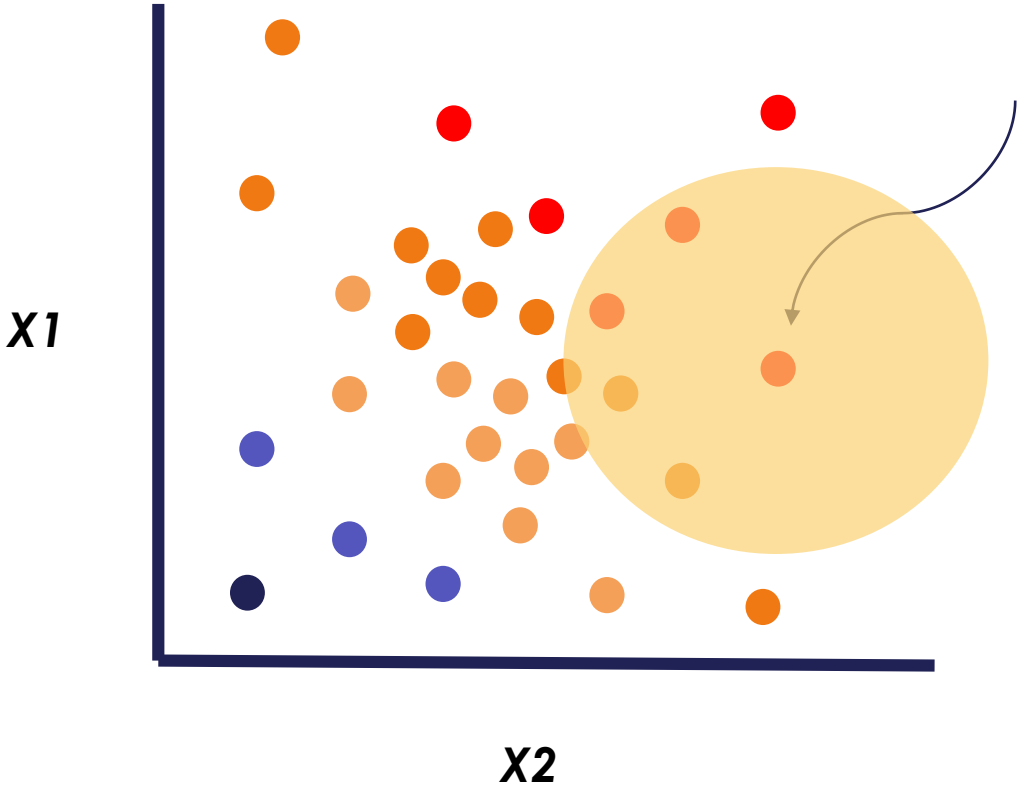
What is vicinity?

Tabular data - problematic



What is vicinity?

Tabular data - problematic

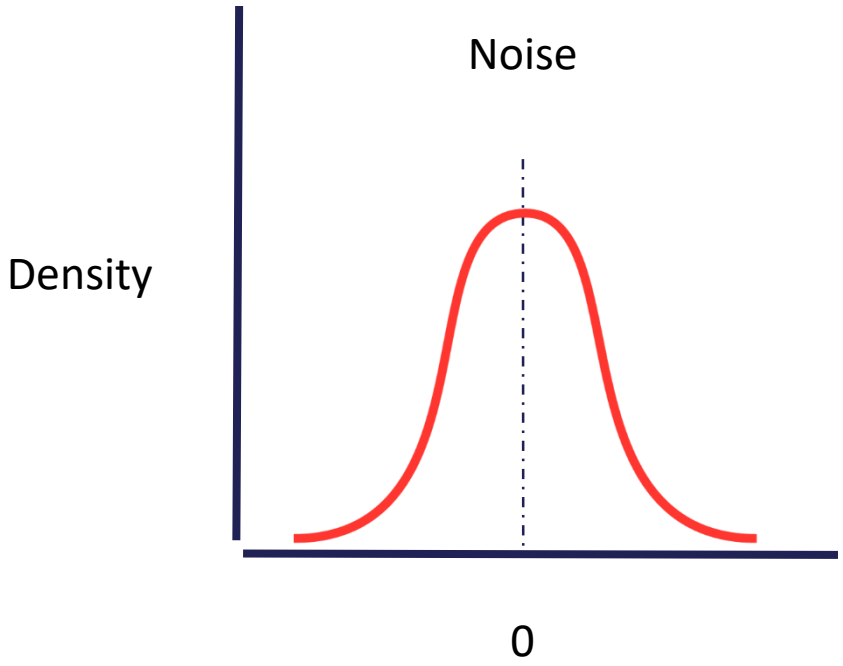


What is vicinity?

➔ Kernel width (arbitrary)

Tabular data – alternative sampling

Colour	Age	Income	Car make	Nr. Cards
Red	65	50000	Ford	5



Add noise extracted from N~(0,1)

Colour	Age	Income	Car make	Nr. Cards
Blue	65	51000	Ford	5
Red	63	59000	Fiat	5.2
Black	60	55000	Ford	5.9
Green	58	45000	VW	6
Green	55	47000	Ford	4

Tabular data – alternative sampling

Colour	Age	Income	Car make	Nr. Cards
Red	65	50000	Ford	5

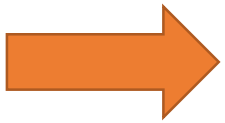
Original implementation ➔ sample at random from variable distribution.

Alternative (better) implementation ➔ Add random noise to the observation we want to explain.

LIME with tabular data

Colour	Age	Income	Car make	Nr. Cards
Blue	65	51000	Ford	5
Red	63	59000	Fiat	5.2
Black	60	55000	Ford	5.9
Green	58	45000	VW	6
Green	55	47000	Ford	4

Trained black box

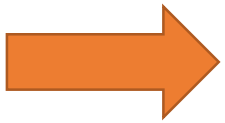


Predictions
y1
y2
y3
y4
y5

LIME with tabular data

Colour	Age	Income	Car make	Nr. Cards
Blue	65	51000	Ford	5
Red	63	59000	Fiat	5.2
Black	60	55000	Ford	5.9
Green	58	45000	VW	6
Green	55	47000	Ford	4

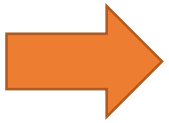
Trained black box



Predictions
y1
y2
y3
y4
y5

Colour	Age	Income	Car make	Nr. Cards	Predictions
0	65	51000	1	5	y1
1	63	59000	0	5.2	y2
0	60	55000	1	5.9	y3
0	58	45000	0	6	y4
0	55	47000	1	4	y5

Explainable model



Interpretations

LIME with tabular data

Colour	Age	Income	Car make	Nr. Cards
Blue	65	51000	Ford	5
Red	63	59000	Fiat	5.2
Black	60	55000	Ford	5.9
Green	58	45000	VW	6
Green	55	47000	Ford	4

Trained black box

Predictions
y1
y2
y3
y4
y5

Colour	Age	Income	Car make	Nr. Cards	Predictions
0	65	51000	1	5	y1
1	63	59000	0	5.2	y2
0	60	55000	1	5.9	y3
0	58	45000	0	6	y4
0	55	47000	1	4	y5

Explainable model

Interpretations

Considerations

- The categorical variables are **one hot encoded** for the explainable model.
- Numerical variables can be sampled from a Gaussian or lhs distribution. Alternatively, add noise to the observation.
- Option to discretize continuous features and treat them as categorical.

Train In Data

THANK YOU

www.trainindata.com