1. Data cleaning

missing values, noise, inconsistencies

attributes

i) Ignore the tuple

Temp.	GPS	speed	Auto/Man	Signal
(30°C	..	60	Auto	Signal)
(~~31°c~~ 90°c	.. ..	—	—	)

tuples

ii) fill in manually

iii) filling in using global constant

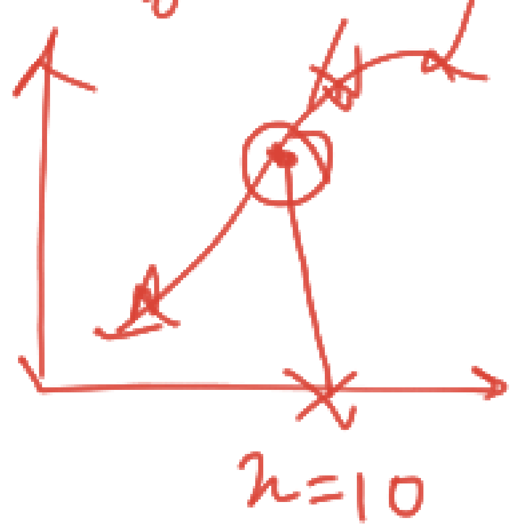
$\infty \longrightarrow 1999$

iv) filling in with mean value

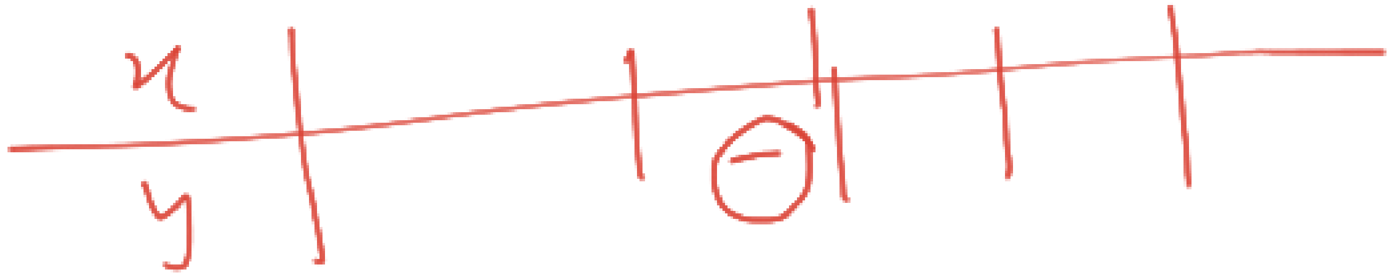
v). categorized $25^\circ - 35^\circ$

vi) most-probable value

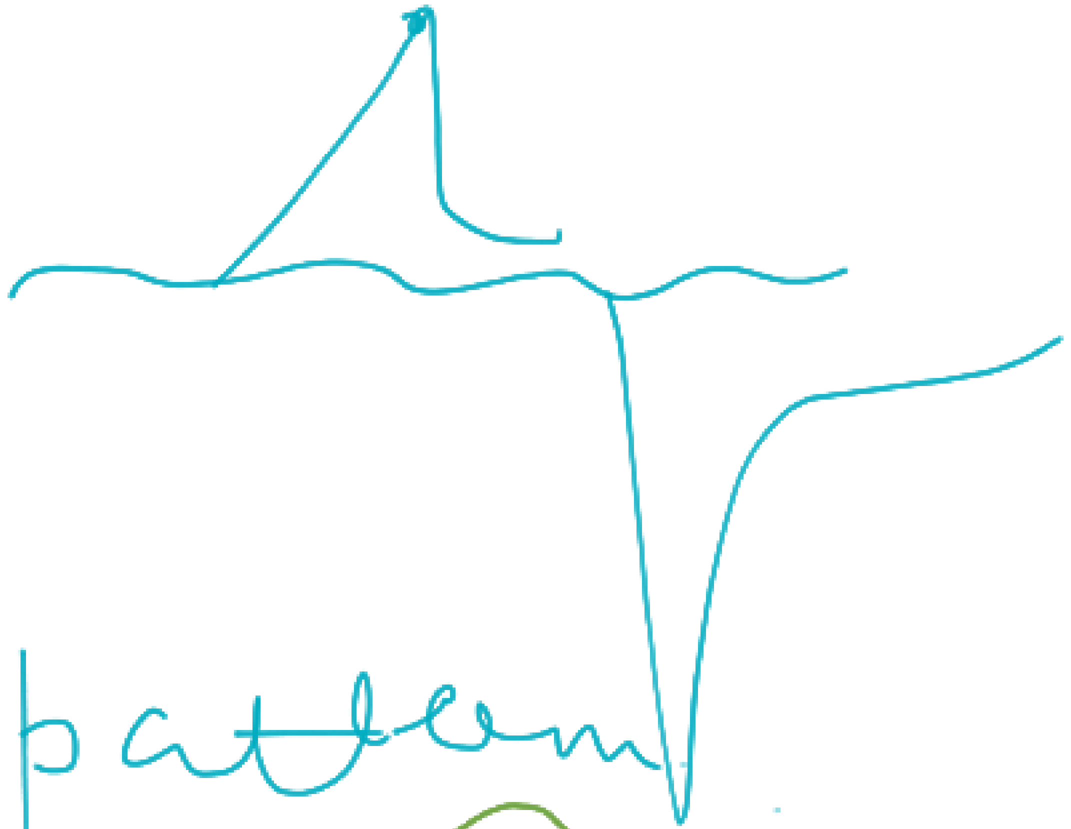
Regression



interpolation



Noise Elimination



meaningless data /

a deviation from normal pattern

20, 21, 25, 30, 31, 91, 34, 38, 25, 21, 20, 20

i) Binning

median

Bin1 : 20, 21, 25, 30, 31 → 23

Bin2 : 31, 34, 38, 91 → 36

Bin3 30, 25, 21, 20, 20 → 20.5

25

34

21

Replacing by mean of each bin

Bin 1 ⟶ 25.4

Bin 2 ⟶ 45

Bin 3 ⟶ 23.2

~~X~~ 91

(ii) Regression

computationally expensive

better solution

iii) clustering

Red / Blue

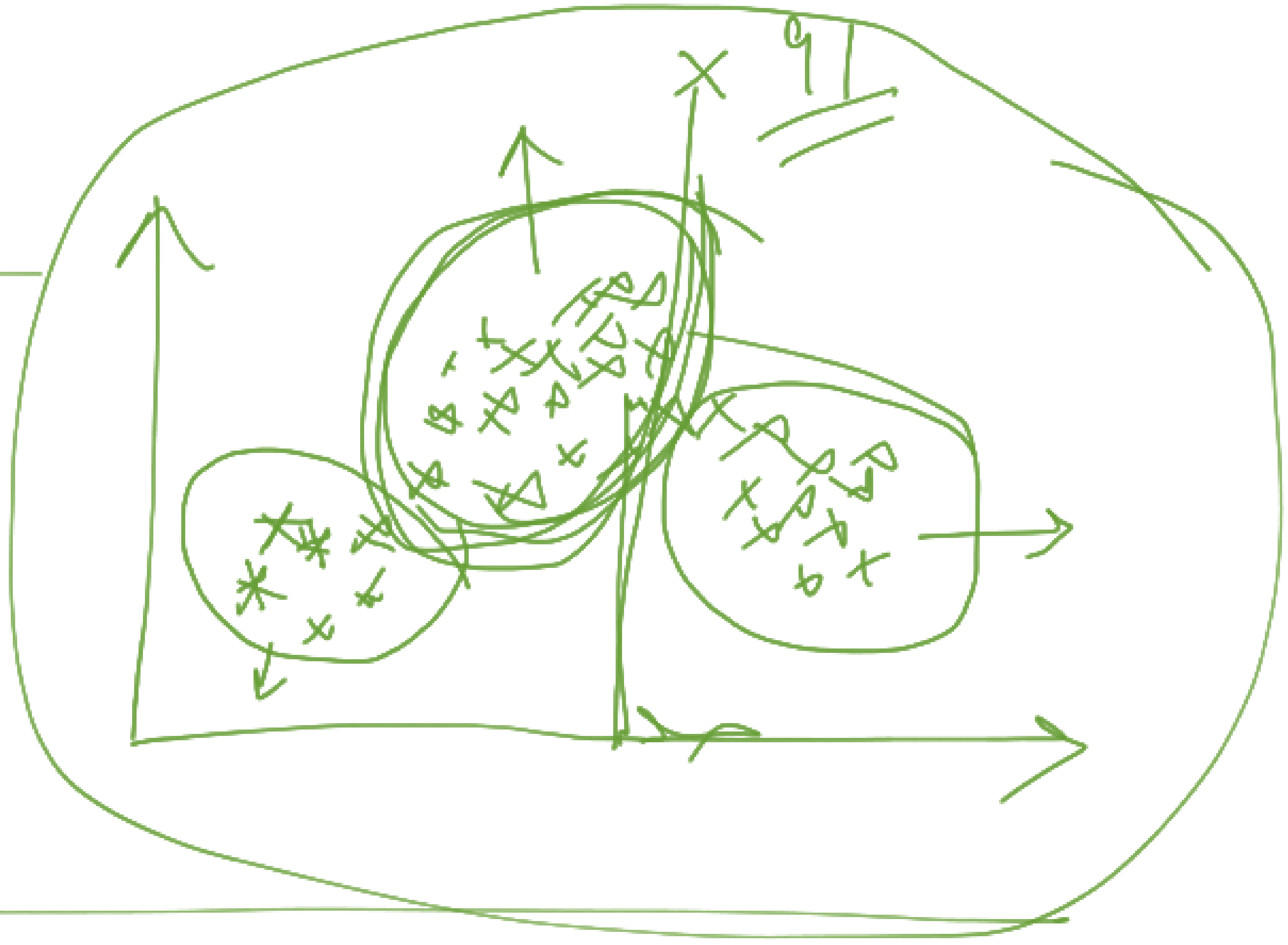
Run 1 — Red

Run 2 — Blue

Run 3 — Red

X Training

✓ Test



P.T.O

2. Data Integration

i) Redundancy

ii) Conflict Resolution

iii) Data Structure mismatch

A = 2
B = 4

A = 3
B = 6

$\boxed{B = 2A}$

A B C X Z

or

A C X Y Z

A	B	C

n

X	Y	Z

Correlation Coefficient (Pearson's coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{N} (a_i b_i) - N\bar{A}\bar{B}}{N \sigma_A \sigma_B}$$

$$\frac{\sum a_i}{N} = \bar{A}$$

or, $\sum a_i = \bar{A} \cdot N$

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$

$$\frac{\sum a_i}{N} = \bar{A}$$

$$= \frac{\sum_{i=1}^{N} (a_i b_i - a_i \bar{B} - b_i \bar{A} + \bar{A}\bar{B})}{N \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{N} a_i b_i - \bar{B} \sum_{i=1}^{N} a_i - \bar{A} \sum_{i=1}^{N} b_i + \bar{A}\bar{B} \sum_{i=1}^{N} 1}{N \sigma_A \sigma_B}$$

$$(1 + 1 + 1 + \cdots) \to N$$

$$= \frac{\sum_{i=1}^{N} a_i b_i - N\bar{A}\bar{B} - N \cdot \cancel{\bar{A}\bar{B}} + N \cancel{\bar{A}\bar{B}}}{N \sigma_A \sigma_B} = \frac{\sum_{i=1}^{N} a_i b_i - N \bar{A}\bar{B}}{N \sigma_A \sigma_B}$$

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \cdot \sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i - \bar{A})} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (b_i - \bar{B})}}$$

$$= \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{N} (a_i - \bar{A})} \sqrt{\sum_{i=1}^{N} (b_i - \bar{B})}}$$

$\to$ cov

$\to$ SD

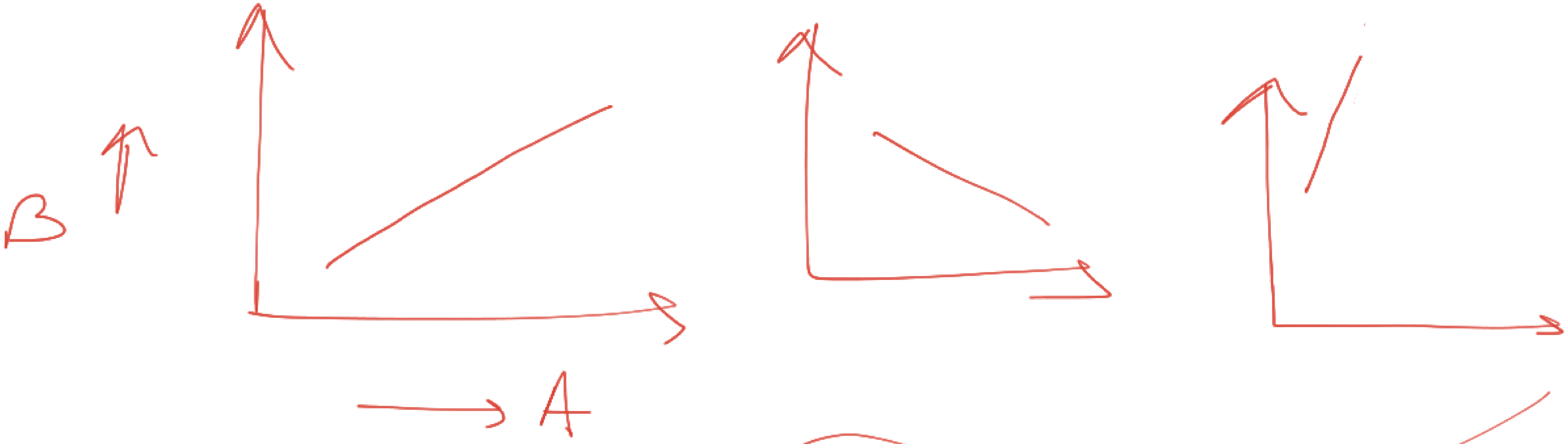
-1 to $+1$

If $r_{A,B} > 0 \longrightarrow$ A & B are positively correlated

$r_{A,B} < 0 \longrightarrow$ -vely correlated

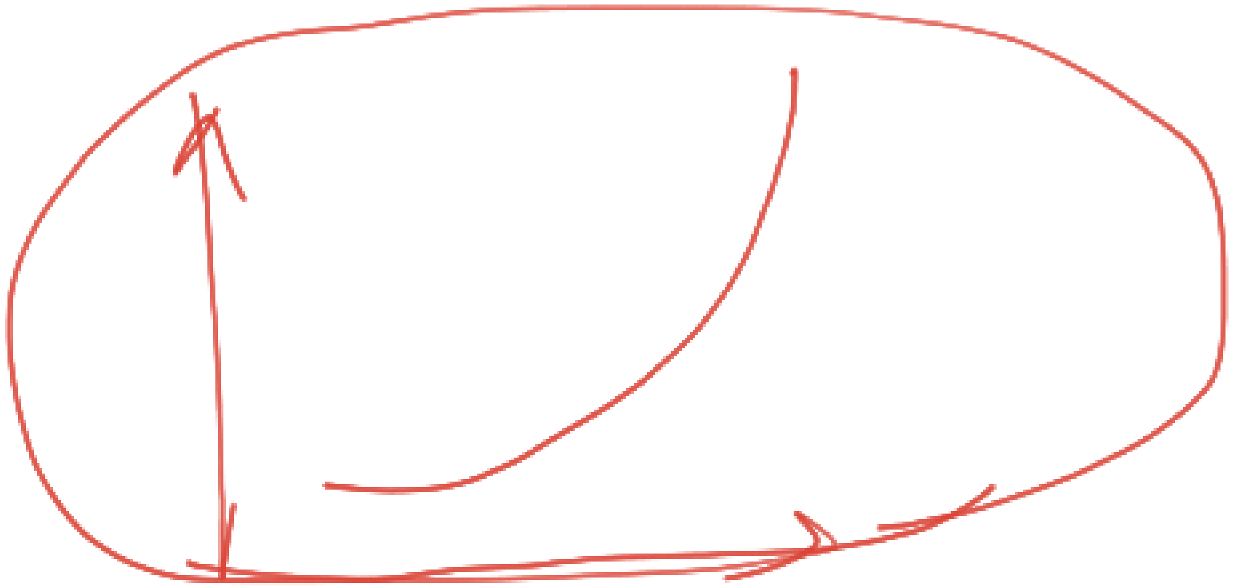
$r_{A,B} = 0 \longrightarrow$ A & B are independent.

is able to detect only linear correlation



linear correlation

$A = 2B$



$\longrightarrow$ X

22

A	2	3	4	5	8
B	4	6	8	10	16

$\bar{A} = 4.4$

$\bar{B} = 8.8$

$\sigma_A = 2.1$

$\sigma_B = 4.118$

$r_{A,B} = 13.358$

$$\sigma_B = \sqrt{\frac{1}{5}\left[\boxed{(4.8)^2 + (2.8)^2 + (0.8)^2 + (1.2)^2 + (7.2)^2}\right]}$$

84.8

$$\sigma_A = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{A})^2}$$

$$= \sqrt{\frac{1}{5}\left((-2.4)^2 + (-1.4)^2 + (-0.4)^2 + (0.6)^2 + (3.6)^2\right)}$$

$$\sigma_A = \sqrt{4.24}$$

$= 2.1$

$$r_{A,B} = \frac{(-2.4)\cdot(-4.4) + (-1.4)(-2.8) + (-0.4)(-0.8) + (1.2)(0.4) + (7.2)(3.6)}{5 \times \underline{2.1} \times \underline{4.118}}$$

$2.1 \to 2.059$

$4.118 \to 4.418$

$$= \frac{42.4}{42.39}$$

$$\simeq \underline{1}$$

+ve

-ve

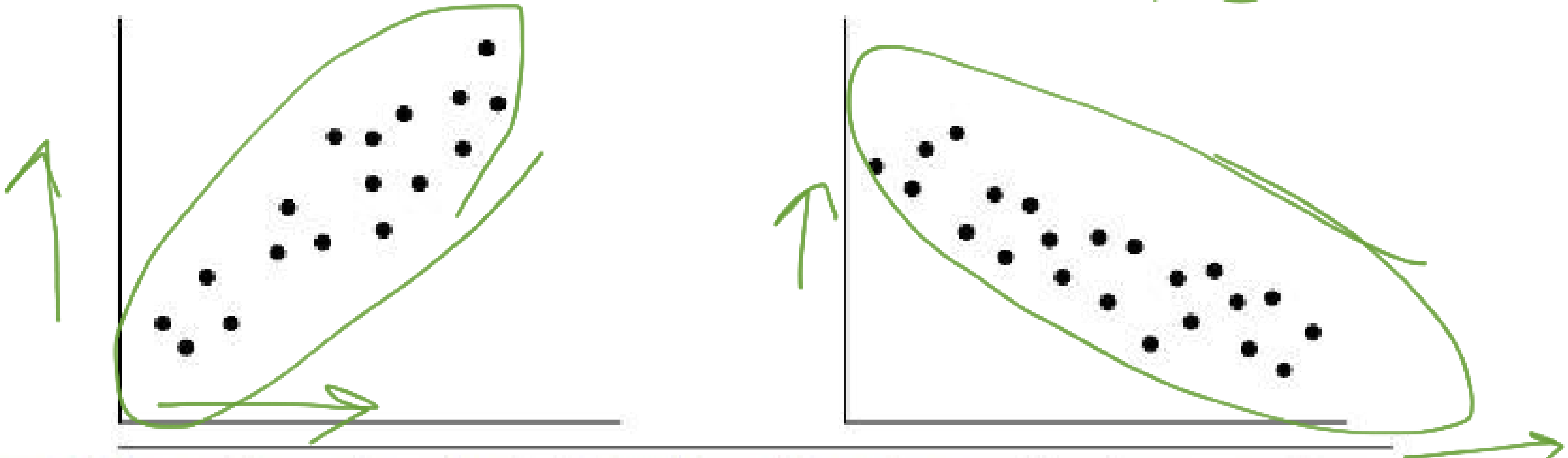


Figure 2.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

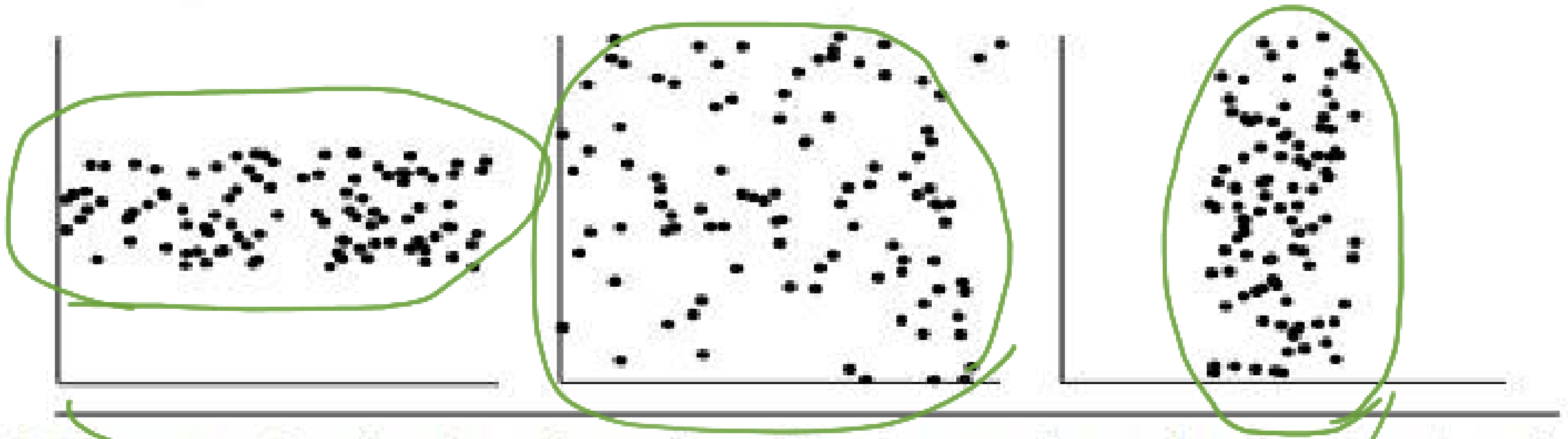


Figure 2.9 Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

No correlation.

Next

x^2 Data Normalization