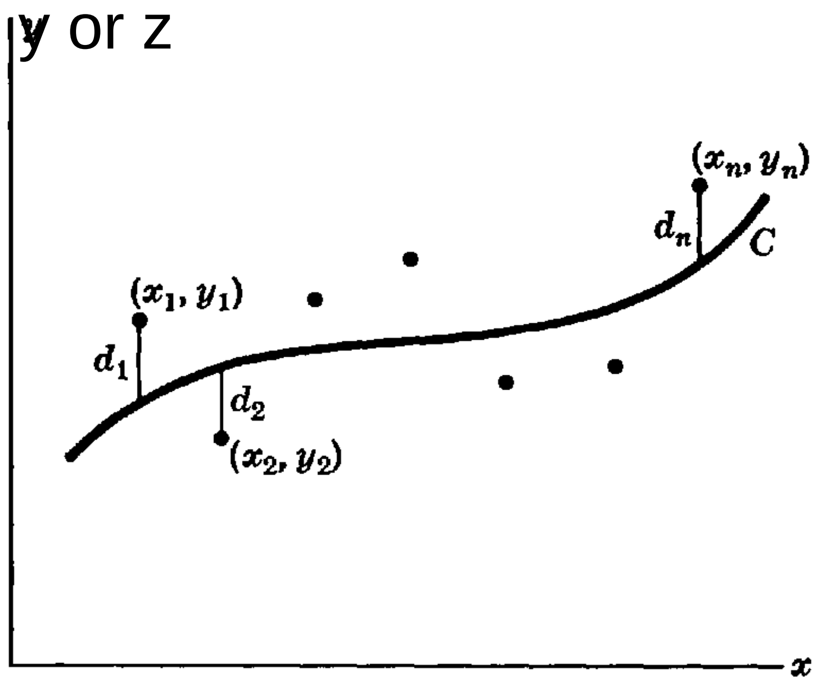


Deviation - error

- Given, x_i and y_i
- K number of data points
- Error = $y_i - f(x_i)$
- How many errors?
- K errors

Sum of Squared errors

$$d_1^2 + d_2^2 + \cdots + d_k^2 = \text{ a minimum}$$

$$\sum_{i=1}^{K} (y_i - f(x_i))^2 \longrightarrow \text{minimize}$$

Root mean square error

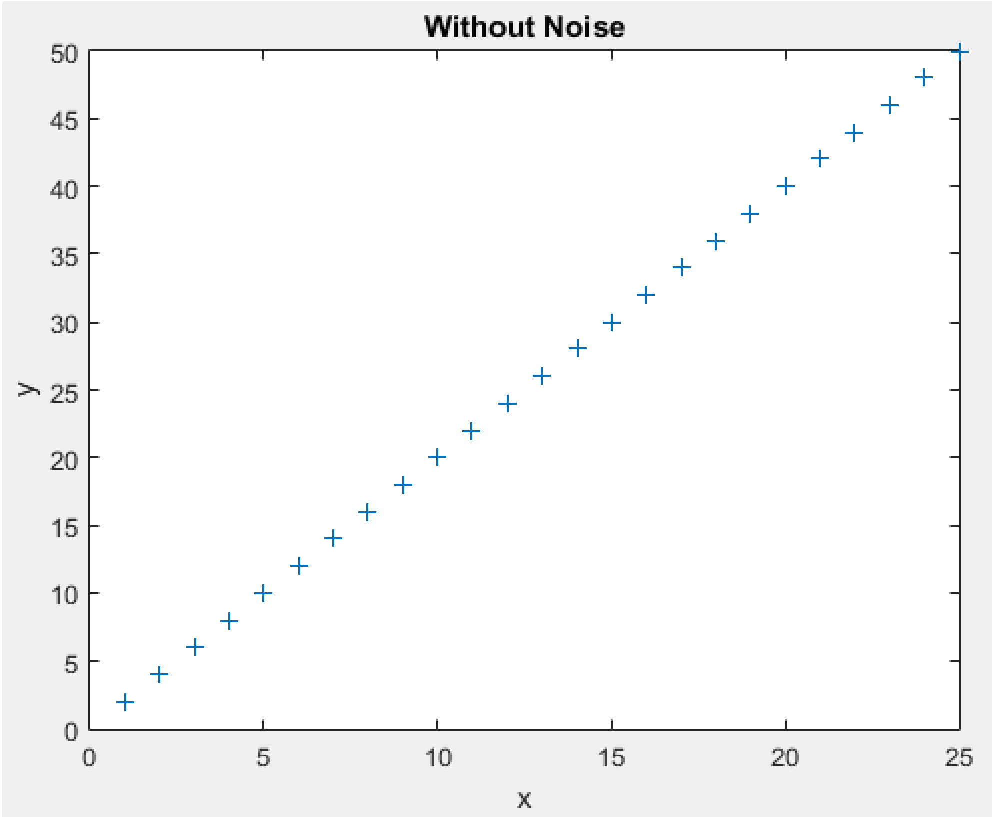
$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(y_i - f\left(x_i\right)\right)^2}$$

Data points generated: t=2x

X	y
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16
9	18
10	20
11	22
12	24
13	26
14	28

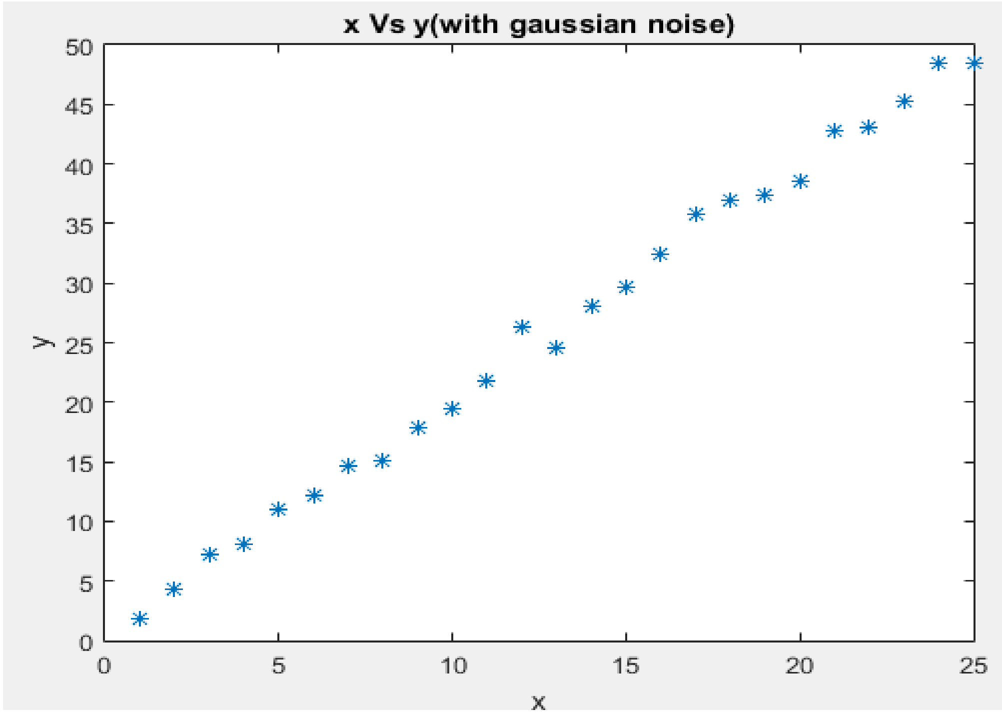
x	y
15	30
16	32
17	34
18	36
19	38
20	40
21	42
22	44
23	46
24	48
25	50

t=2x points plotted



$t = 2x$ + Gaussian noise ($\sigma = 2$)

All the points are disturbed



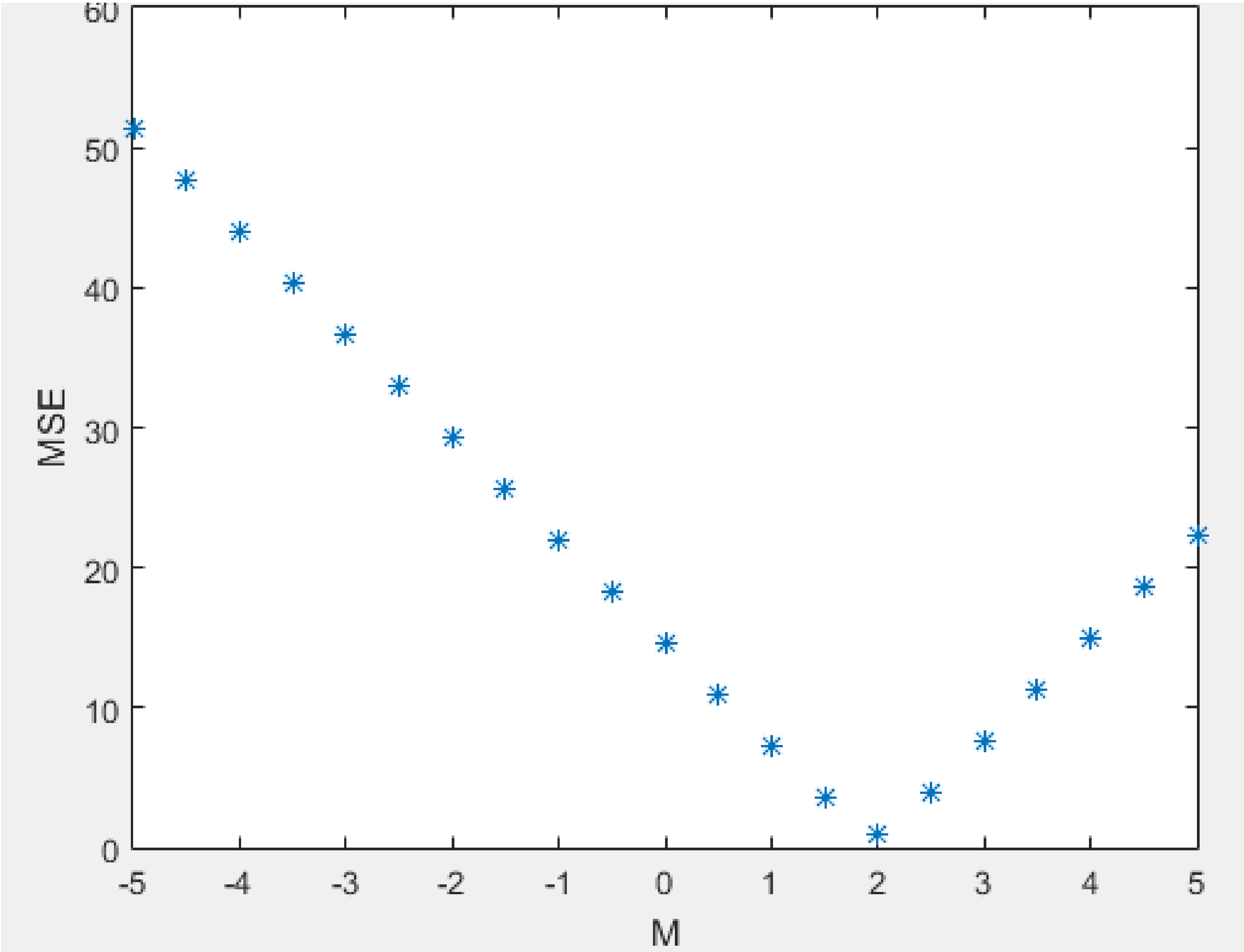
Algorithm

- We have decided to fit using y=m.x
- x=[---, ---, ---, ---, …., ---]
- t=[---, ---, ---, ---, …., ---]
- Choose m

1. ***Predict y=[---, ---, ---, ---, …., ---]***
2. ***Find out error***
3. ***e=[---, ---, ---, ---, …., ---]***
4. ***Generate squared error by squaring the elements of e***
5. ***se=[---, ---, ---, ---, …., ---]***
6. ***Compute the mean of se (MSE)***

- Change m; repeat 1 to 6
- Plot MSE versus m; choose the m corresponding to minimum MSE

Graph: MSE versus m



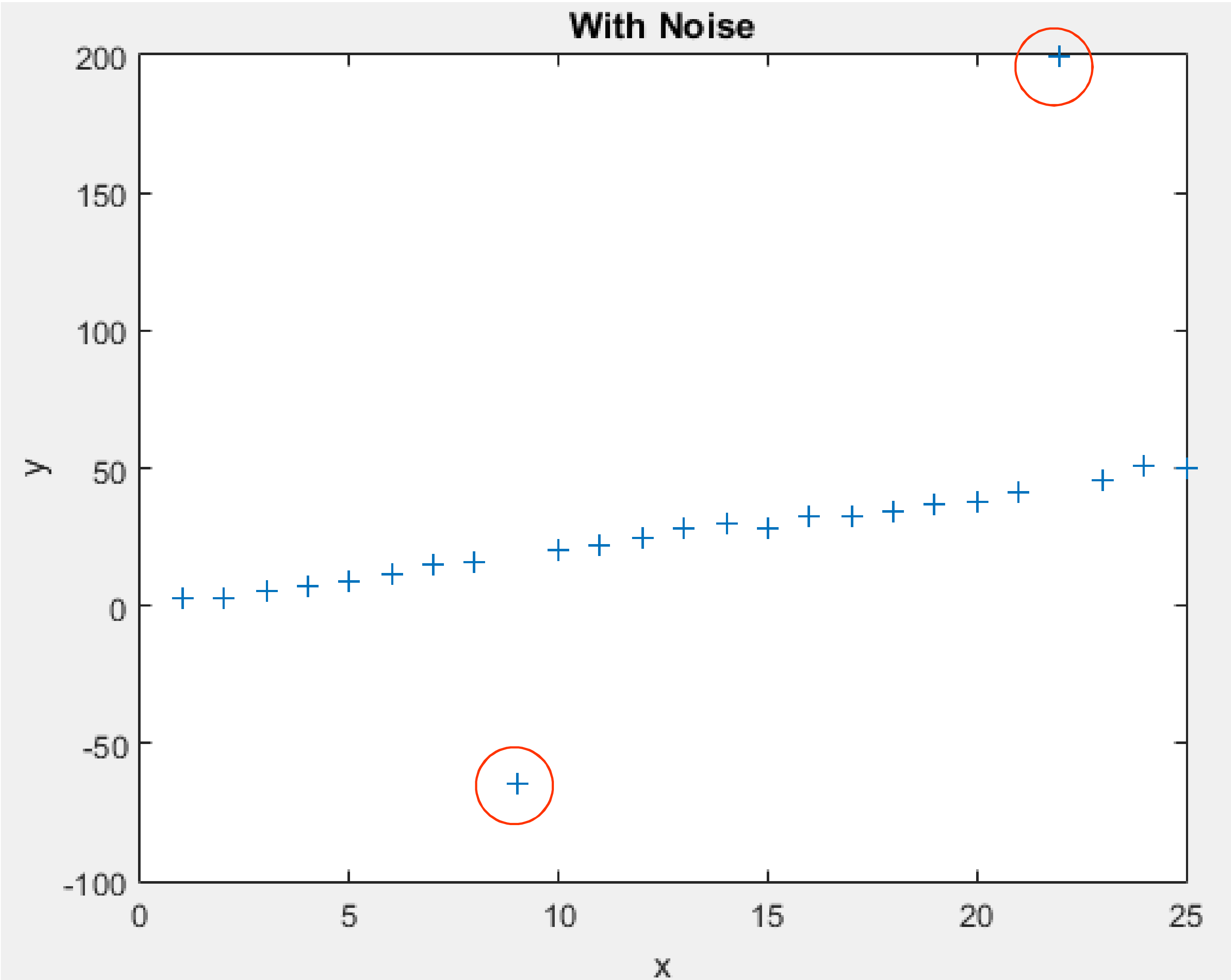
Data points t=2x + noise

x	y
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16
9	-65
10	20
11	22
12	24
13	26
14	28

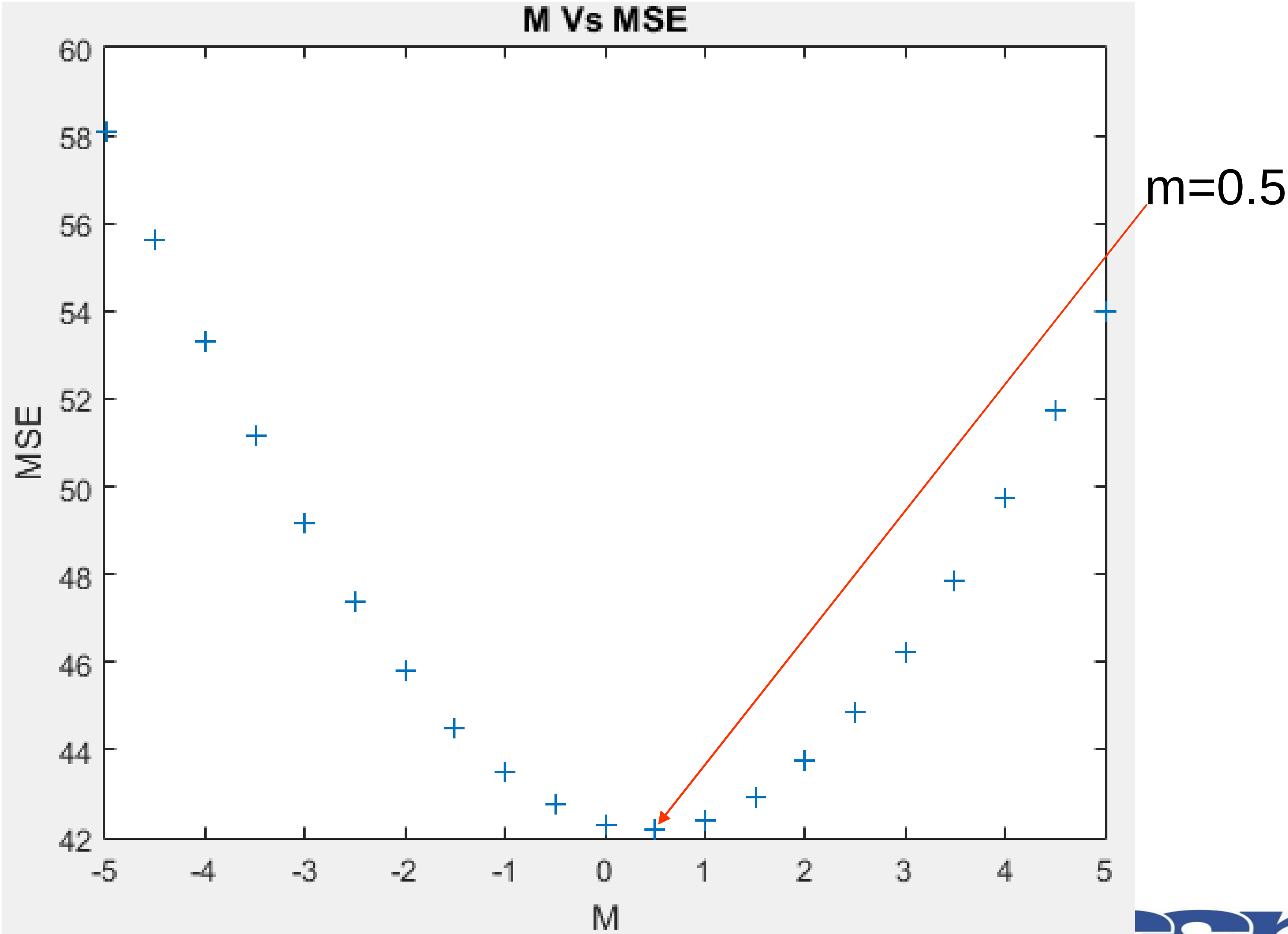
x	y
15	30
16	32
17	34
18	36
19	38
20	40
21	42
22	200
23	46
24	48
25	50

Noise: Two points are disturbed

t=2x points plotted with 2 points disturbed



Plot: MSE versus m



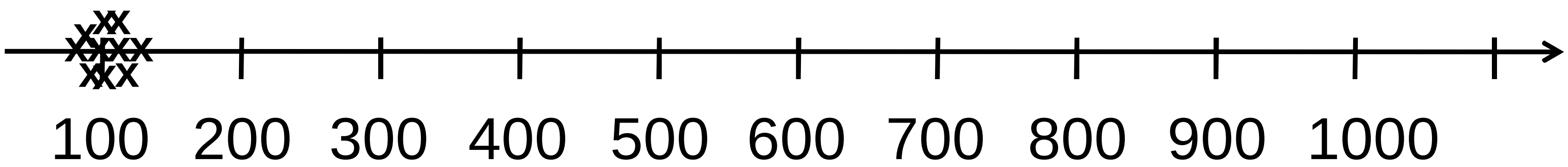
Observation

- All the points disturbed by noise (Gaussian)
- MSE works
- Just two points disturbed by noise (extreme values)
- MSE fails

How does the Arithmetic mean handle Outlier?[1]

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - 100, 120, 90, 110, 115, 125, 95, 105, 110, 100

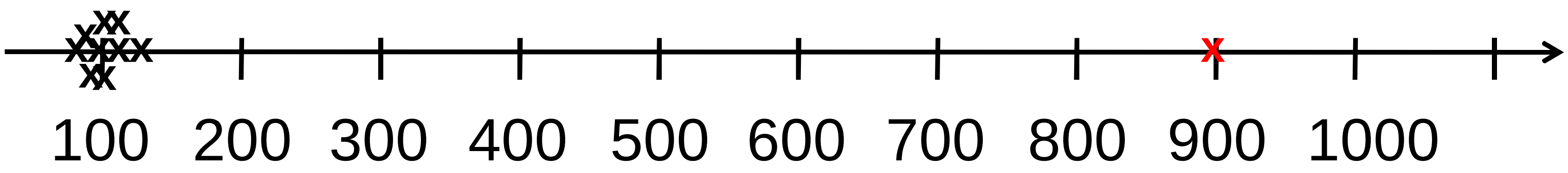
Arithmetic mean = Sum of above numbers/10 = 107



How does the Arithmetic mean handle Outlier?[2]

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - **<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100

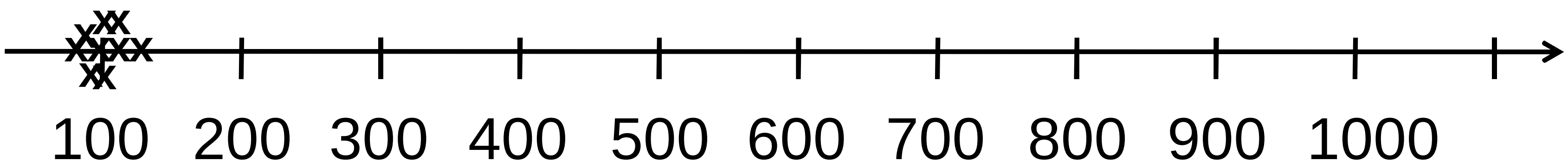
Arithmetic mean = Sum of above numbers/10 = 187



Arithmetic mean after removing outlier

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - ~~**900**~~, 120, 90, 110, 115, 125, 95, 105, 110, 100

Arithmetic mean = Sum of above numbers/9 = 107.8

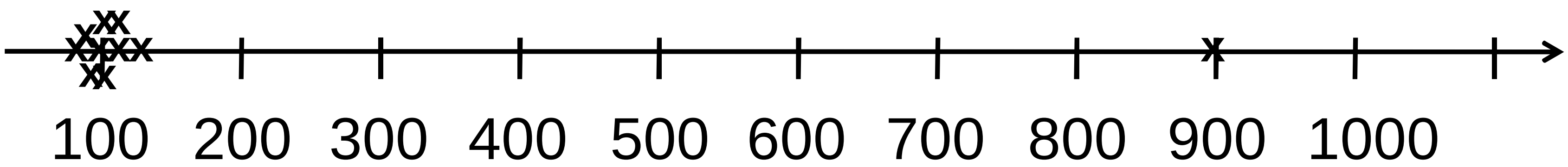


How does the SM handle Outlier?

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - 100, 120, 90, 110, 115, 125, 95, 105, 110, 100

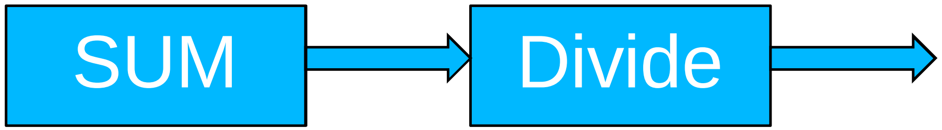
AM = Sum of above numbers/10 = 107

SM = sum of the squared numbers $\Rightarrow$ taking root = 107.5

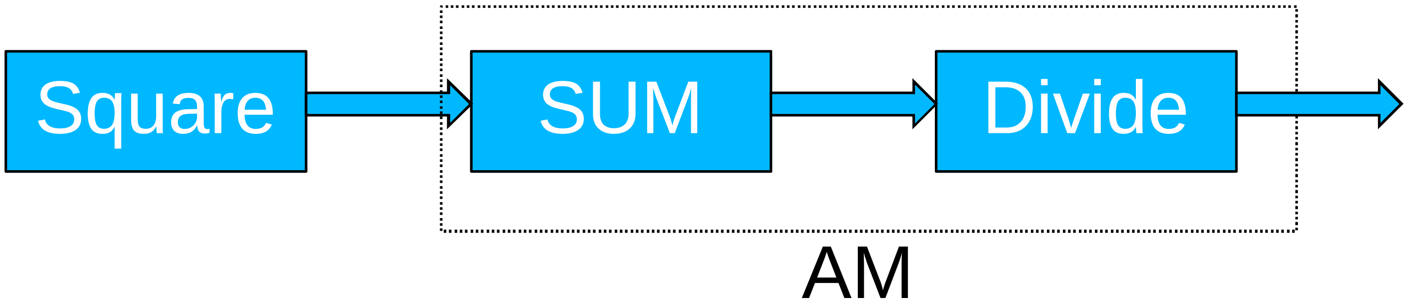


Squared mean

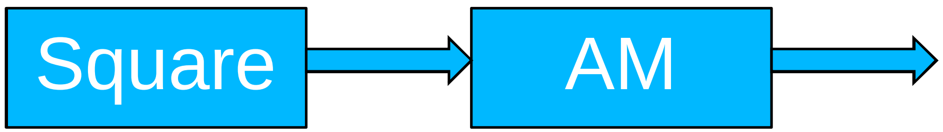
AM: Sum and divide



SM: Square, sum and divide



=

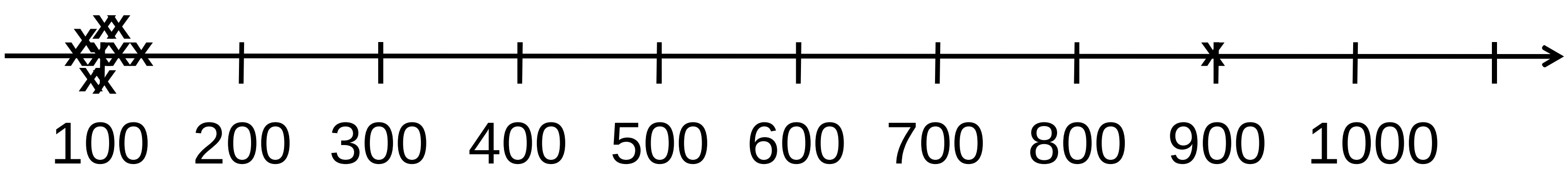


How does the SM handle Outlier?

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - **<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100

AM = Sum of above numbers/10 = 187

SM = sum of the squared numbers/10 $\Rightarrow$ taking root = 302.6

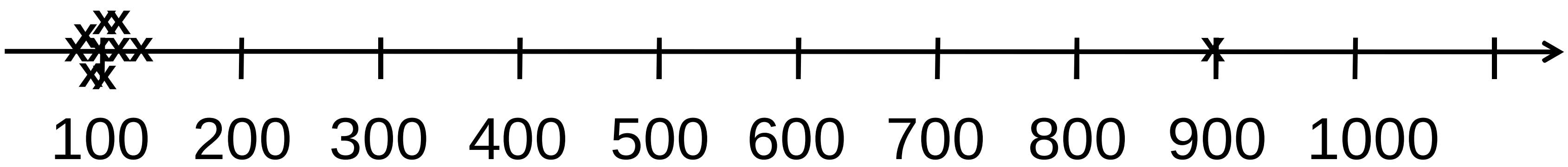


SM after removing outlier

- Consider a 1000 marks test.
- 10 students taking up the test.
- Their marks are
  - ~~**900**~~, 120, 90, 110, 115, 125, 95, 105, 110, 100

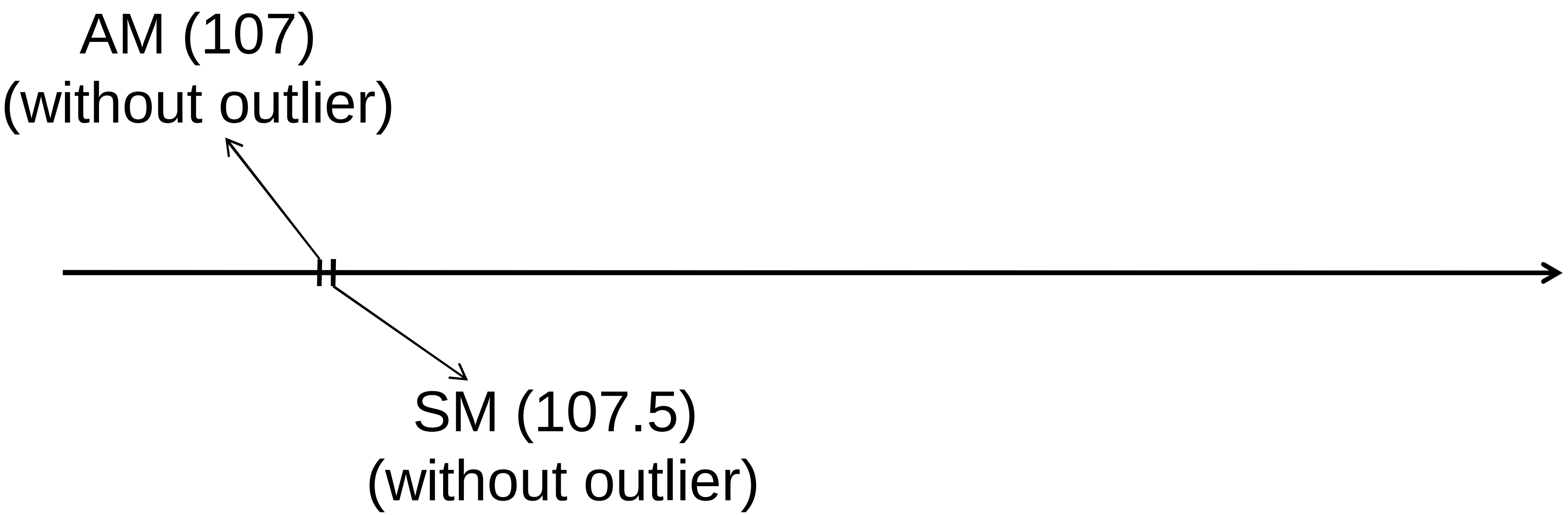
AM = Sum of above numbers/9 = 107.8

SM = sum of the squared numbers/9 $\Rightarrow$ taking root
  = 108.3



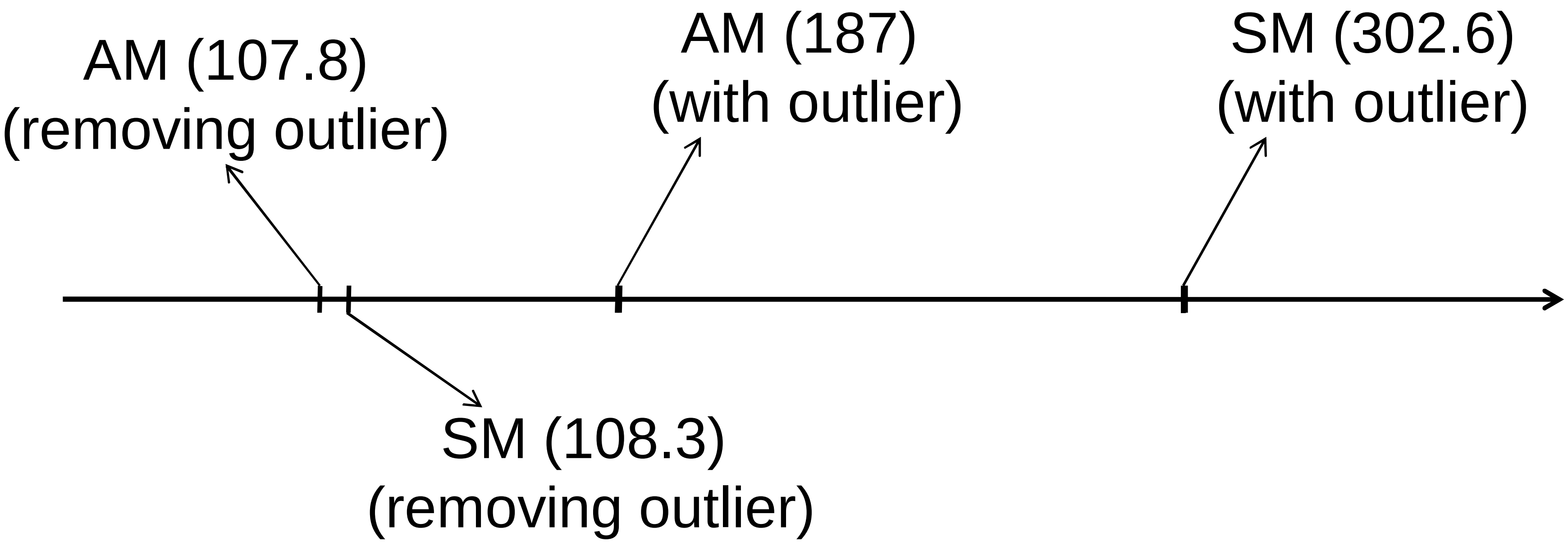
AM and SM without outlier

- Data = {**<u>100</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}



AM and SM with outlier

- Data = {**<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}



$$average \quad of \quad \{1,3,5,7,9\}$$

$$\mu = \frac{(1 + 3 + 5 + 7 + 9)}{5}$$

$$= \left(\frac{1}{5}\right)1 + \left(\frac{1}{5}\right)3 + \left(\frac{1}{5}\right)5 + \left(\frac{1}{5}\right)7 + \left(\frac{1}{5}\right)9$$

$$= w_1 * 1 + w_2 * 3 + w_3 * 5 + w_4 * 7 + w_5 * 9$$

$$\Rightarrow w_1 = w_2 = w_3 = w_4 = w_5 = \frac{1}{5}$$

$$mean \quad square \quad average \quad of \quad \{1,3,5,7,9\}$$

$$\mu = \left(\frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2}{5} \right)$$

$$= \frac{(1*1 + 3*3 + 5*5 + 7*7 + 9*9)}{5}$$

$$= \left(\frac{1}{5}\right)1 + \left(\frac{3}{5}\right)3 + \left(\frac{5}{5}\right)5 + \left(\frac{7}{5}\right)7 + \left(\frac{9}{5}\right)9$$

$$= w_1 * 1 + w_2 * 3 + w_3 * 5 + w_4 * 7 + w_5 * 9$$

$$\Rightarrow w_1 = \frac{1}{5}; w_2 = \frac{3}{5}; w_3 = \frac{5}{5}; w_4 = \frac{7}{5}; w_5 = \frac{9}{5}$$

Weights of MSE

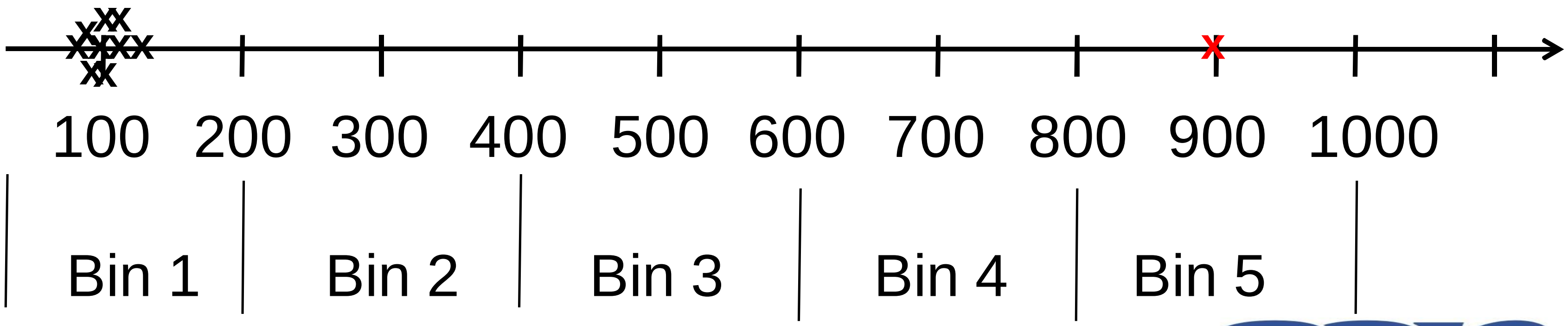
- Bigger numbers $\Rightarrow$ bigger weights
- If outliers happen to be larger number then big weight is allotted

Limitations/characteristics of MSE

- Errors are squared and summed
- Characteristics of squaring:
  - After squaring a big number becomes a bigger number
  - Errors occur in a range
  - Big errors are given more important compared to small errors

How does the weighted AM handle Outlier?

- Data = {**<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}
- Bin 1 – 9 numbers
- Bin 5 – 1number
- Probability of bin1=0.9 & prob of bin5=0.1



Weights

N number of data points

AM

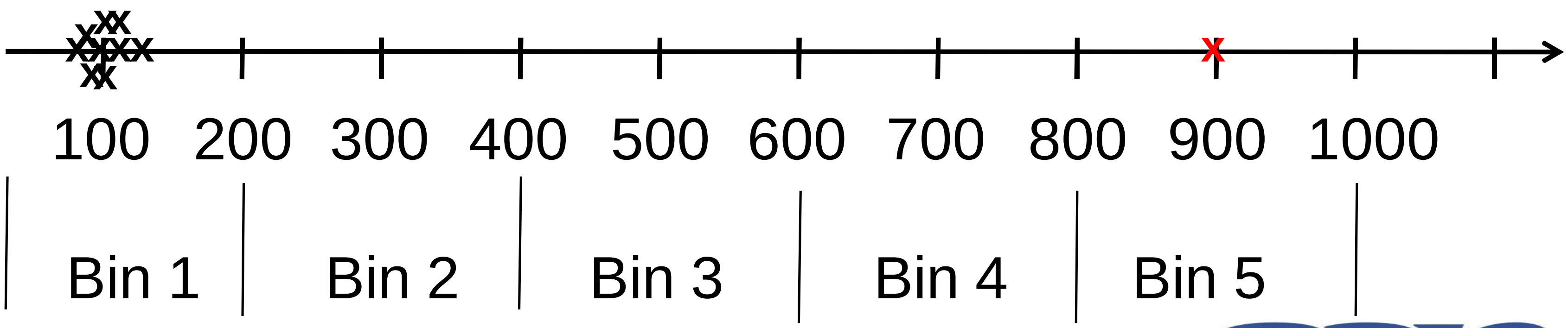
- Weights $w_1 = w_2 = .. = w_N = 1/N$
- $w_1 + w_2 + .. + w_N = 1$

Expectation

- Weights $w_1 = p_1$; $w_2 = p_2$; ...$w_N = p_N$
- $w_1 + w_2 + .. + w_N = p_1 + p_2 + .. + p_N = 1$

Weights for this example

- Bin 1 – 9 numbers
- Bin 5 – 1number
- Weight for bin 5 = w
- Weight for bin 1 = 9w



Reduced weight for outlier

- Weight for bin1 numbers > weight for bin5 numbers
- 9 times bigger than bin5 weight
- 9 numbers in bin1 and 1 number in bin5
- $9x(9w)+w=82w=1 \Rightarrow 0.012$
- Data = {**<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}

=0.012*900 + 0.108*(120+90+110+115+125+95+105+110+100)

=115.6

Limitation of weighted AM

- Data = {**<u>900</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}

=0.012*900 + 0.108*(120+90+110+115+125+95+105+110+100)

=115.6

- Data = {**<u>1800</u>**, 120, 90, 110, 115, 125, 95, 105, 110, 100}

=0.012*1800 + 0.108*(120+90+110+115+125+95+105+110+100)

=126.4

What’s the problem?

$$weighted\ AM = \sum normal\_weight\ X\ normal\_data$$
$$+\ \ reduced\_weight\ X\ outlier$$

- Still it depends on outlier value

Lesson

- Our measure should not depend upon outlier

Our measure should not depend upon outlier

- How do you know 'something is outlier'?

Measure should not depend on data values

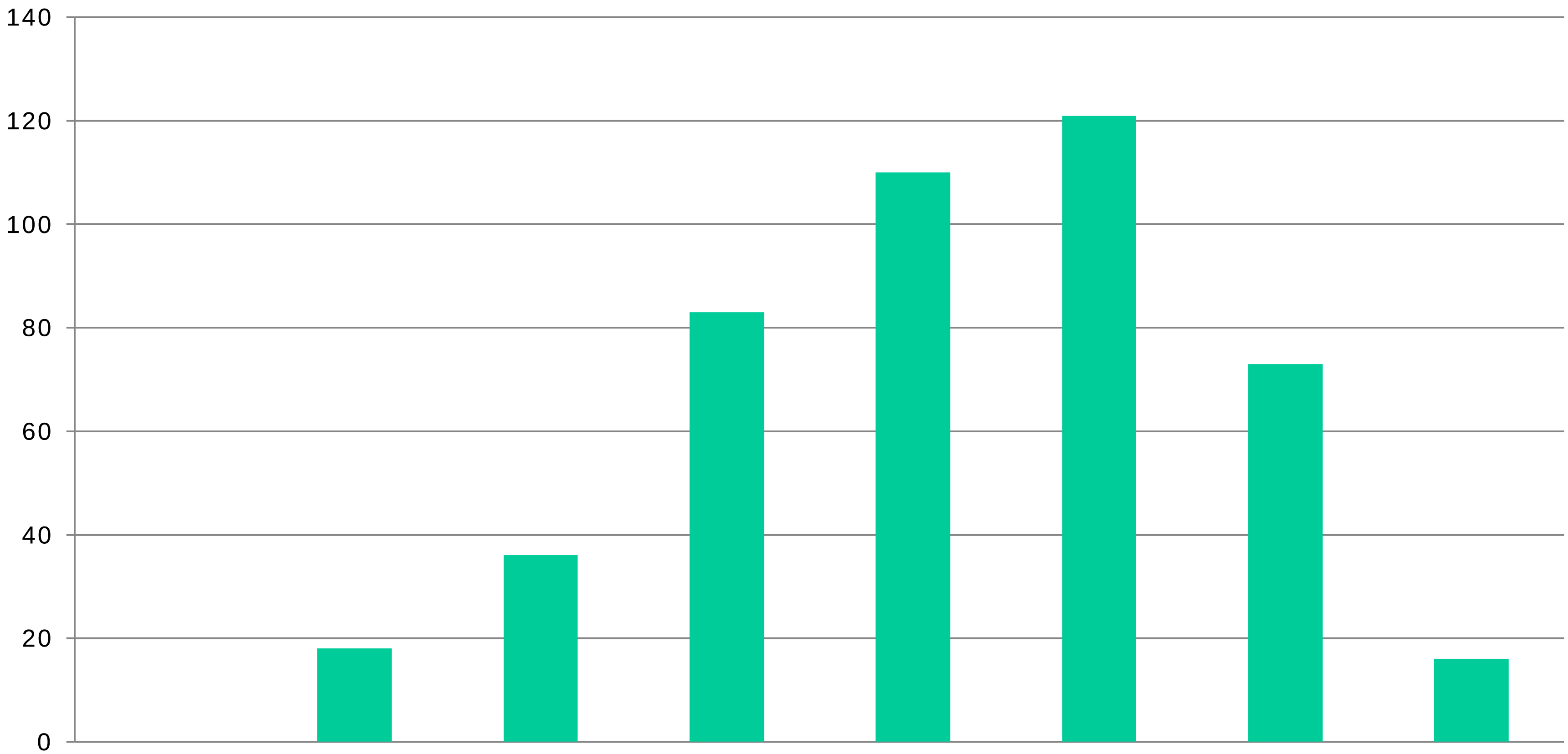
Ponder over the statement

Measure should not depend on data values

- Apart from data what else we can use
- Distribution (ore frequency) of the data

Frequency distribution

Value of data in X axis & frequency of data in Y axis



Lesson

- Our measure should not depend upon outlier
- Or simply data values should not be used

No more X axis

- We'll work with Y axis
- i.e. not with data values rather with frequency of data values

If we do not use data values then...

- Use their frequency distribution
- Assume marks of 457 students given to us
- Data = {90, 12, 155, 88, 65, ...76}
- Make frequency distribution out of this data

A measure works with Y axis i.e. frequency of occurrence

- Entropy
- Frequency of occurrence closely related with probability
- **Probability = normalized frequency distribution**

$$\sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{r} p_i \log_2 p_i$$