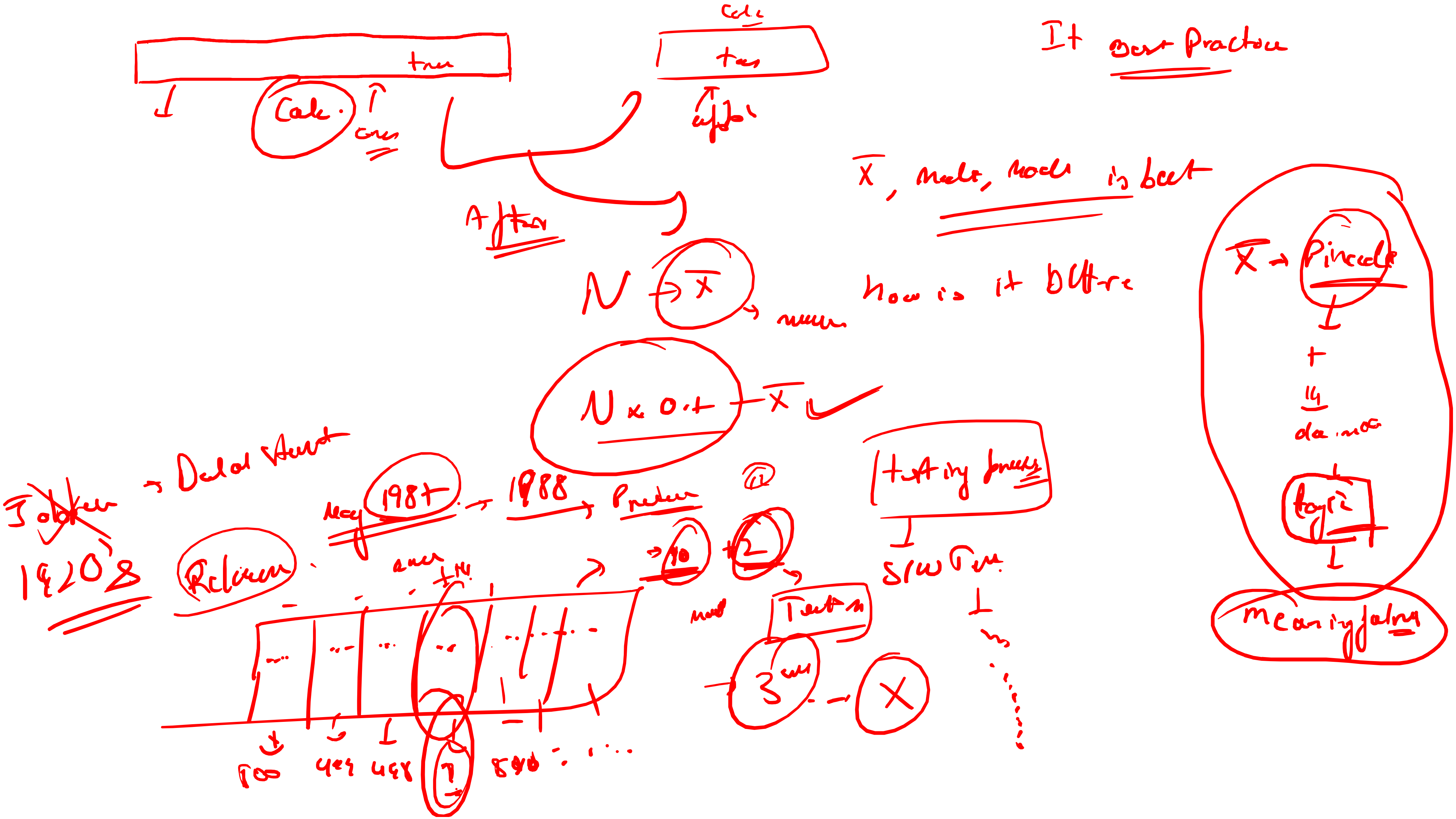


It best Practice
how is it better
After
Release
1987
1988
meaningful

→ Entropy ⇒ Splitting criterion → Information Gain (I.G.)

↳ measure of Purity

↓

degree of randomness

↓

In the node, there exist multiple classes

↳ Impure ↑ Entropy ↑

↓

Measure of Impurity

⇒ I.G. = 1 − Entropy

⇒ $Entropy_{node} \Rightarrow -P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3 - \cdots - P_n \log_2 P_n$

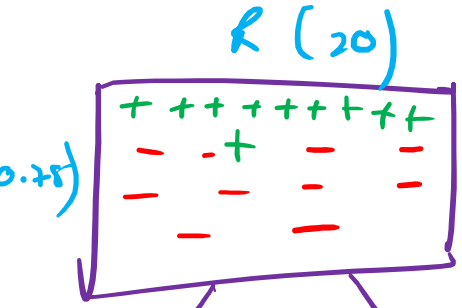
I.G. of the split ⇒ weighted I.G. = $w_1 IG_1 + w_2 IG_2$

→ Split on weight based on Entropy / IG

$I.G._A \Rightarrow 1 - (-P_+ \log_2 P_+ - P_- \log_2 P_-)$

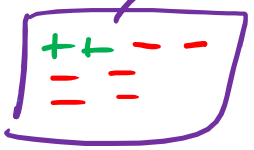
$\Rightarrow 1 - (-0.25 \log_2 0.25 - 0.75 \log_2 0.75)$

$\Rightarrow 0.189$

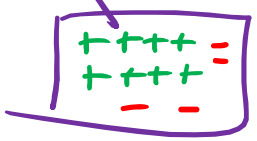


$P(+ve) = 0.25$

$P(-ve) = 0.75$



A(8)



B(12)

$I.G._B = 1 - (-P_+ \log_2 P_+ - P_- \log_2 P_-)$

$= 1 - (-0.67 \log_2 0.67 - 0.33 \log_2 0.33)$

$\Rightarrow 0.086$

$P(+ve) = 0.67$

$P(-ve) = 0.33$

$\Rightarrow$ I.G. on the split

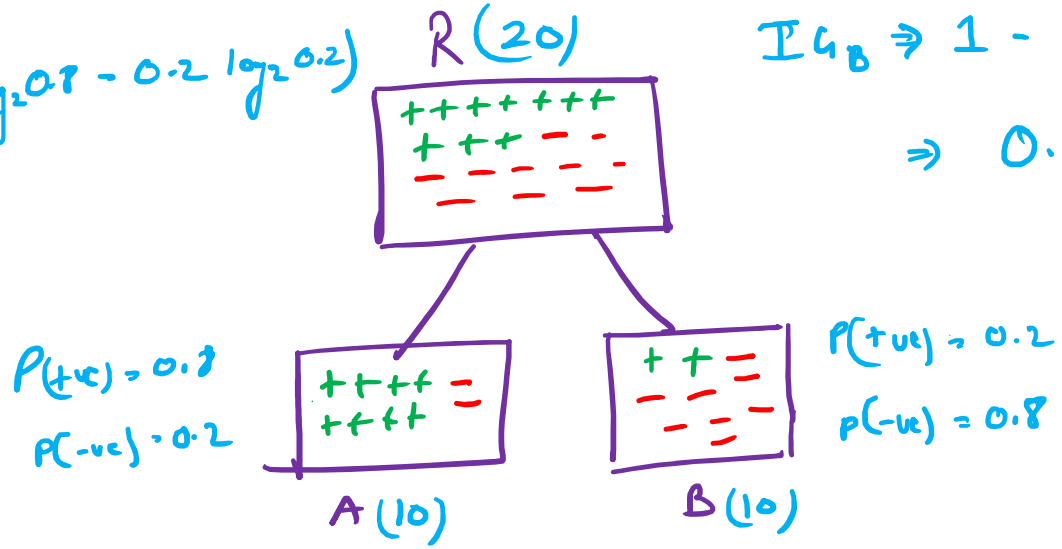
$\Rightarrow I.G._{weight} \Rightarrow W_A IG_A + W_B IG_B$

$\Rightarrow \frac{8}{20} * 0.189 + \frac{12}{20} * 0.086 \Rightarrow 0.127$

→ Split on class based on Entropy / IG.

$IG_A = 1 - (-0.8 \log_2 0.8 - 0.2 \log_2 0.2)$

$\Rightarrow 0.28$



$IG_B \Rightarrow 1 - (-0.2 \log_2 0.2 - 0.8 \log_2 0.8)$

$\Rightarrow 0.28$

I.G. on split

$I.G._{class} = W_A IG_A + W_B IG_B$

$\rightarrow \frac{10}{20}(0.28) + \frac{10}{20}(0.28) = 0.28$

Feature	I.G.
Height	0.127
Class	0.28

Best split ⇒

→ Chi-Square (χ^2) → quality comparison

→ Mostly when more categorical variables are present

↳ Purity Index

↳ Homogeneity

$$\chi^2 = \frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}}$$

↑ χ^2 ↑ Homogeneity

⇓

Pure Node

$$\chi = \sqrt{\frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}}}$$

colors of 2 cap

C_1 (Red) C_2 (Red)

Same = Same

↓

high value of Chi-Square

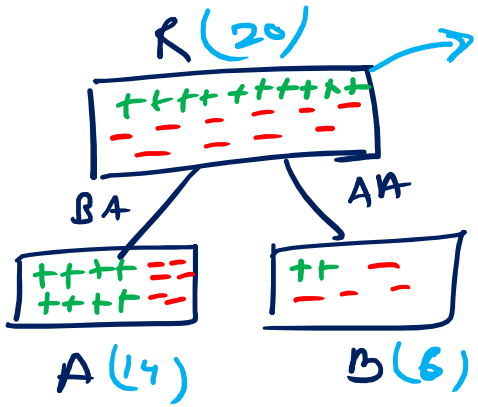
Chi → Wrong Red

Bright Red

↓

low value of Chi-Sq.

→ Split on Performance based on Chi-Square



$\frac{10}{10} = 1:1$

Expected for the child

Exp(+ve) = 3
Exp(-ve) = 3
Act(+ve) = 2
Act(-ve) = 4

χ is calc for node per class

$$\chi = \sqrt{\frac{(Act - Exp)^2}{Expect.}}$$

E(+ve) = 7
E(-ve) = 7
A(+ve) = 8
A(-ve) = 6

$$\chi_{A_{+ve}} = \sqrt{\frac{(8-7)^2}{7}} \Rightarrow 0.38$$

$$\chi_{A_{-ve}} = \sqrt{\frac{(6-7)^2}{7}} \Rightarrow 0.38$$

$$\chi_{B_{+ve}} = \sqrt{\frac{(2-3)^2}{3}} \Rightarrow 0.57$$

$$\chi_{B_{-ve}} = \sqrt{\frac{(4-3)^2}{3}} \Rightarrow 0.57$$

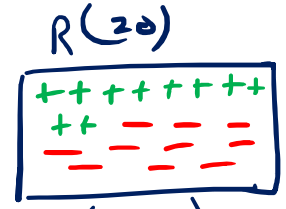
χ_{split} = Sum of χ of all the leafs

$\Rightarrow \chi_{A+ve} + \chi_{A-ve} + \chi_{B+ve} + \chi_{B-ve}$

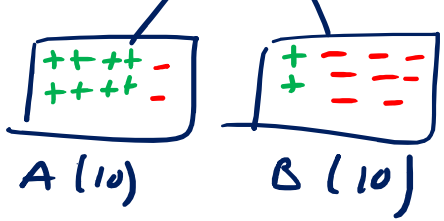
$\Rightarrow 0.38 + 0.38 + 0.57 + 0.57$

$\Rightarrow$ **1.9 - χ_{perf-}**

→ Split on Class based on Chi-Square



→ 1 : 2



E (true) = 5
E (-ve) = 5
A (true) = 8
A (-ve) = 2

E (true) = 5
E (-ve) = 5
A (true) = 2
A (-ve) = 8

$$\chi = \sqrt{\frac{(A-E)^2}{E}}$$

$$\chi_{A(true)} = \sqrt{\frac{(8-5)^2}{5}}$$

$$\Rightarrow 1.34$$

$$\chi_{A(-ve)} = \sqrt{\frac{(2-5)^2}{5}}$$

$$\Rightarrow 1.34$$

$$\chi_{B(true)} = \sqrt{\frac{(2-5)^2}{5}}$$

$$\Rightarrow 1.34$$

$$\chi_{B(-ve)} = \sqrt{\frac{(8-5)^2}{5}}$$

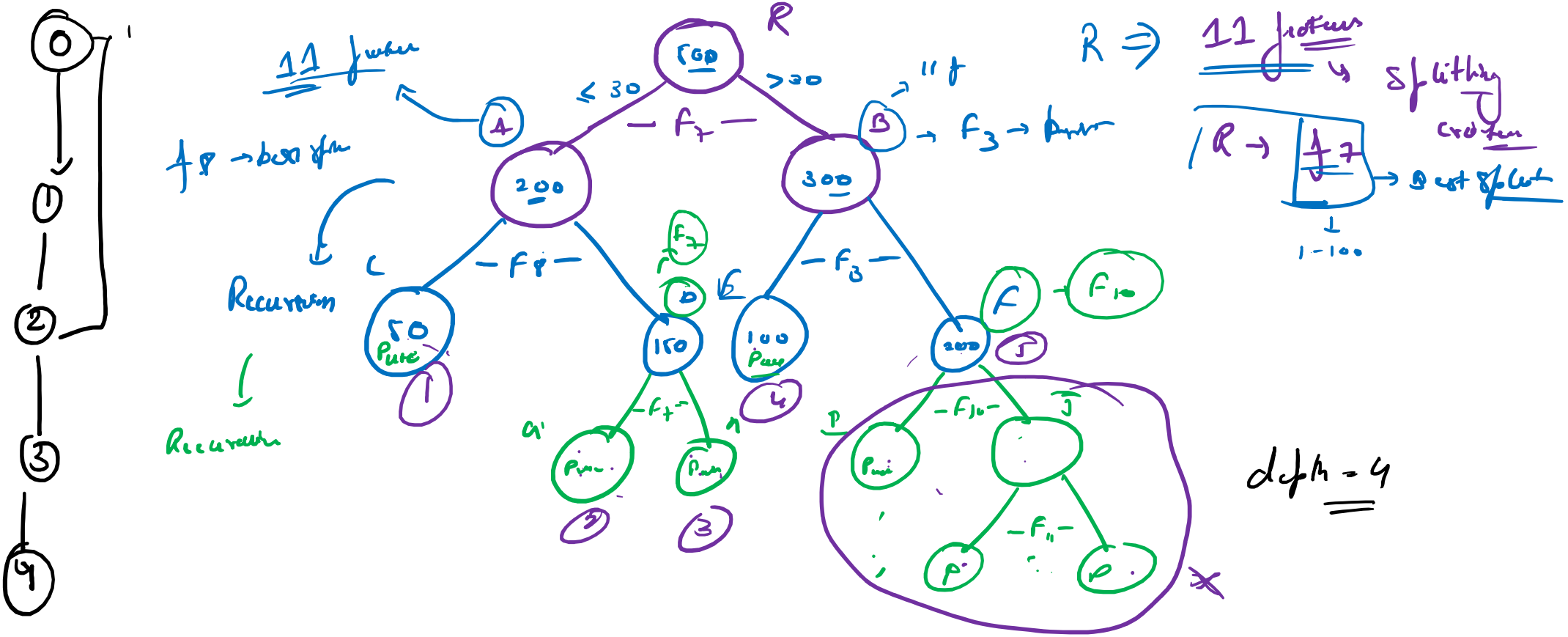
$$\Rightarrow 1.34$$

$$\chi_{Class} = 1.34 + 1.34 + 1.34 + 1.34$$

$$\Rightarrow 5.36$$

feature	χ
Perf	1.9
Class	5.36

Best split ⇒


0
1
2
3
4
R
11 functions
≤ 30
>30
200
300
50
150
100
200
Recursion
Recursion
R ⇒ 11 functions
Splitting
1-100
dfh = 4

→ (Hyperparameters ⟹) → MLA can be customized on the behaviour of execution

↳ Defaults

→ depth = 2, 3, 7 ...

↳ Performance / quality

↓↓

Test, Train

→ Min sample for a node to split = 12, 10, 8

→ Min ⁴ for a Terminal Node = 8, 5, 11

→ Max no. of Terminal nodes = 5

Pruning → DS.

Controlling the growth of the Tree

Default & tuning rules

↳

Whenever a node becomes pure the growth will stop

↓

Splitting can reach to a single row in one node

Only Training

2, 12	3, 12	7, 12
2, 10	3, 10	7, 10
2, 9	3, 9	7, 9

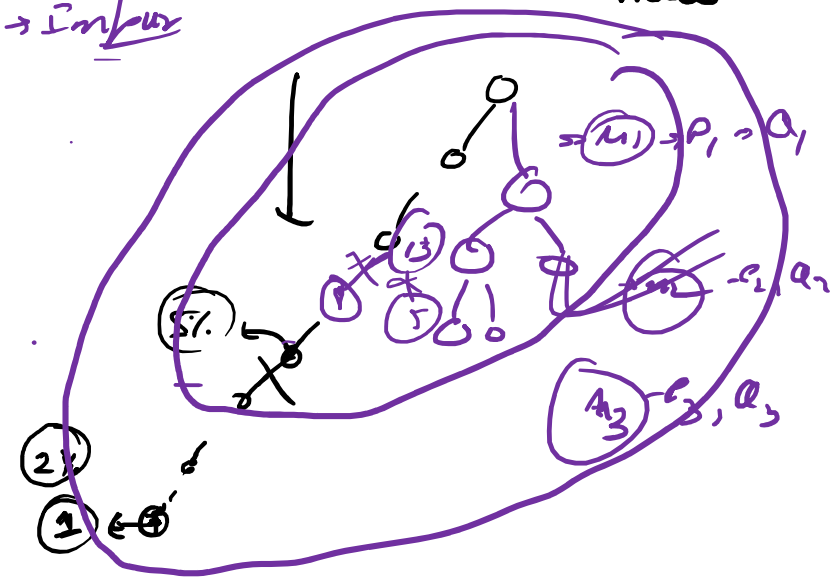
7, 12, 6

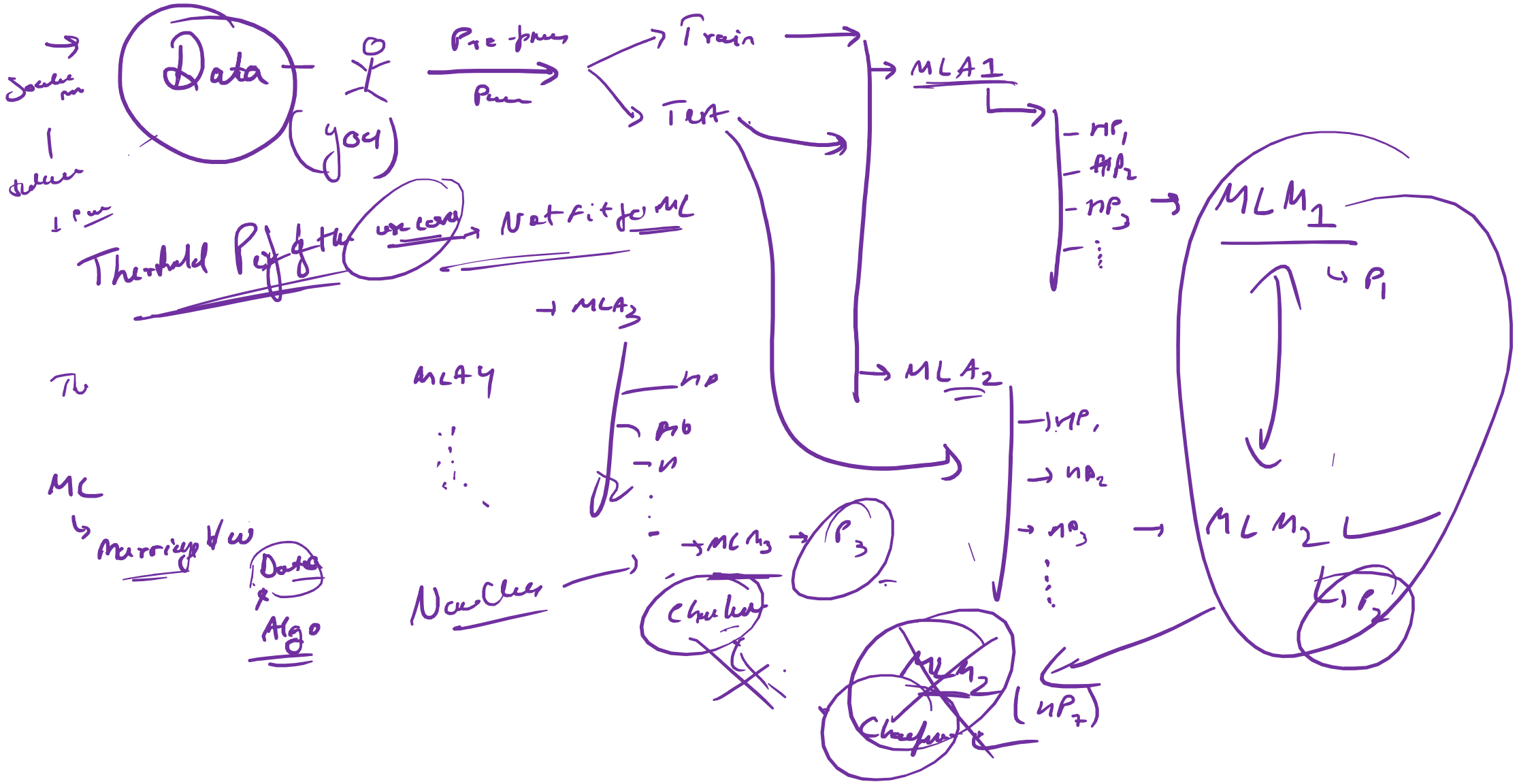
13 → Impure

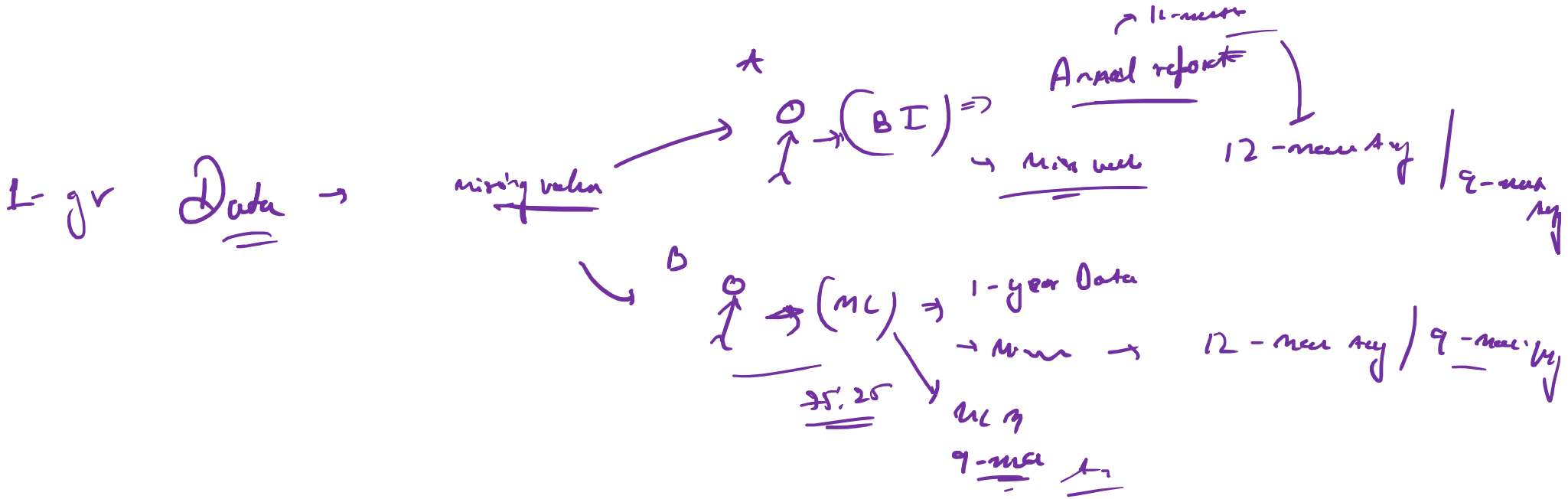
8

X

X




Data
Train
Test
MLA1
MLA2
MLA3
MLA4
MLM1
MLM2
Threshold
Not Fit for ML
Algo


1- yr Data
missing value
A
BI
Annual report
12 - mean Avg / 4 - mean Avg
B
ML
1-year Data
12 - mean Avg / 4 - mean Avg
35.25