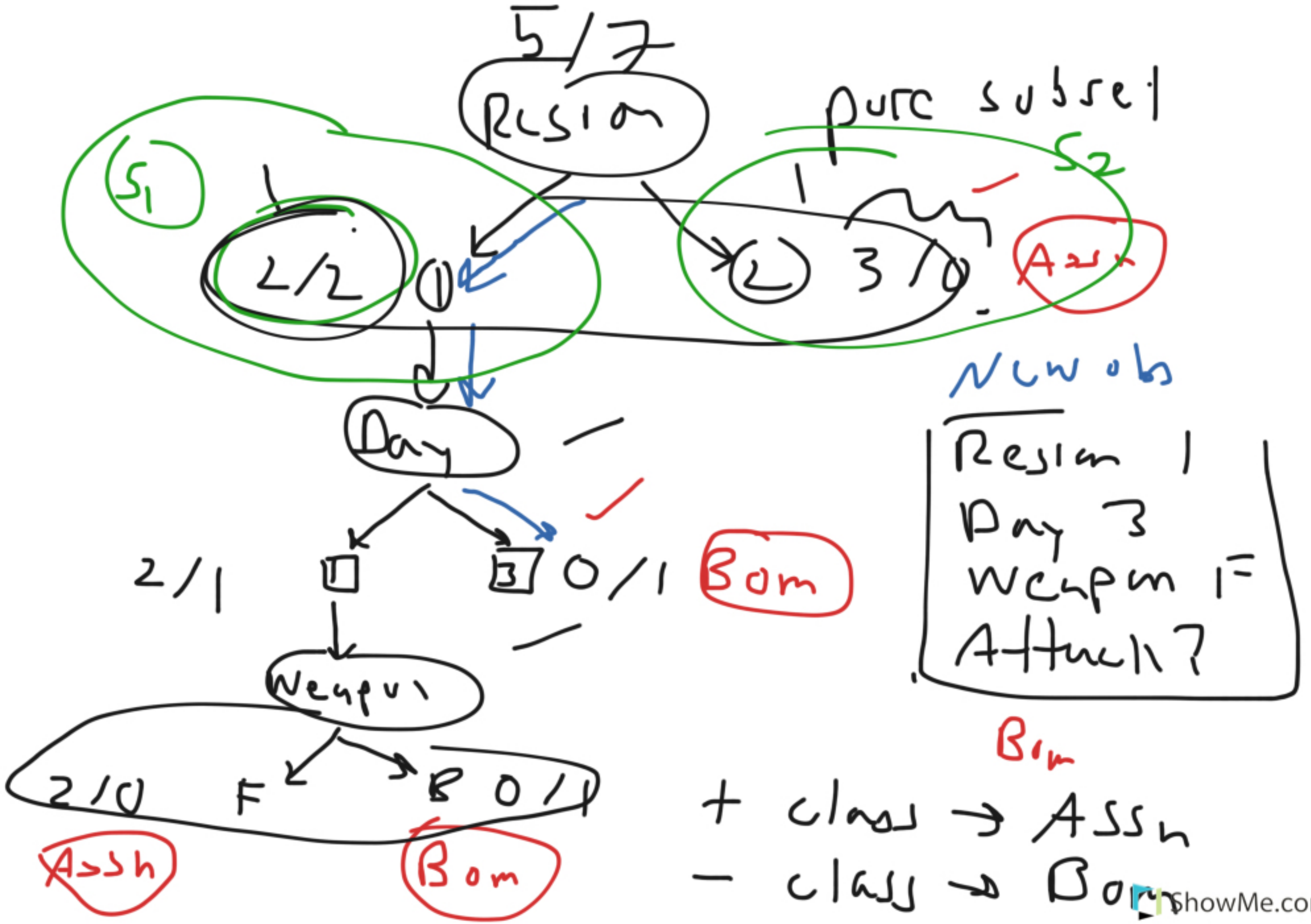
Region, Day, Weapon
(1, 2) (1, 2, 3, ..., 20) (F, B)

feature

Attack type - target
(Assn., Bom)


5/7
Region
pure subset
S1
S2
2/2
3/0
Assn
New ob
Region 1
Day 3
Weapon F
Attack?
Day
2/1
0/1
Bom
Weapon
2/0
F
B 0/1
Assn
Bom
+ class → Assn
− class → Bom

purity/entropy

- entropy of a subset in a tree tells us how much uncertainty each subset contains

Entropy subset $S = \{1, \ldots, N\}$

$$H(S) = -p(+)\log_2 p(+) - p(-)\log_2 p(-)$$

$$H(S_1) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1 \text{ bit}$$

$p(+) = \frac{2}{4}$, $p(-) = \frac{2}{4}$

Assn , Bor

$H(S_2) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - 0\log(101)$

$p(+) = \frac{3}{3}$ $p(-) = 0$

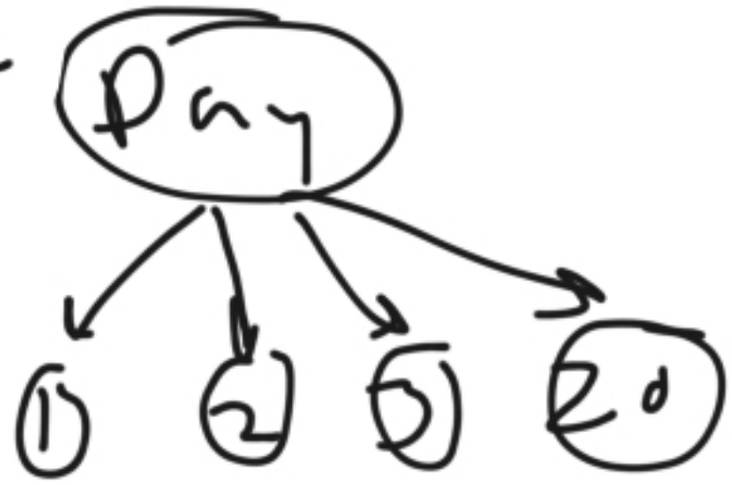
11
0

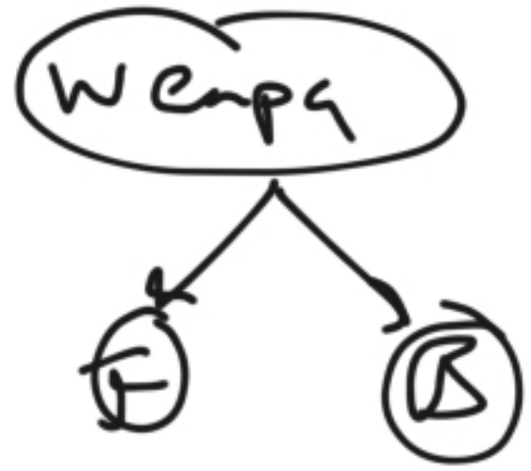
Information gain

$Gain(S, A) = H(S) - \sum_v \frac{|S_v|}{|S|} H(S_v)$

V - values of A
S - set of examples X

5/12





$Gain(S, Resim)$ $Gain(S, Day)$ $Gain(S, weather)$

I_1 I_2 I_3

$H(S) = -p(+)\log_2 p(+) - p(-)\log_2 p(-)$

$= -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.59$

$$Gain(S, Resim)$$

$$= H(S) - \frac{4}{7} H(S_1) - \frac{3}{7} H(S_2)$$

$$\text{II}$$

$$0.59 - \frac{4}{7}(1) - \frac{3}{7}(0) = 0.026$$

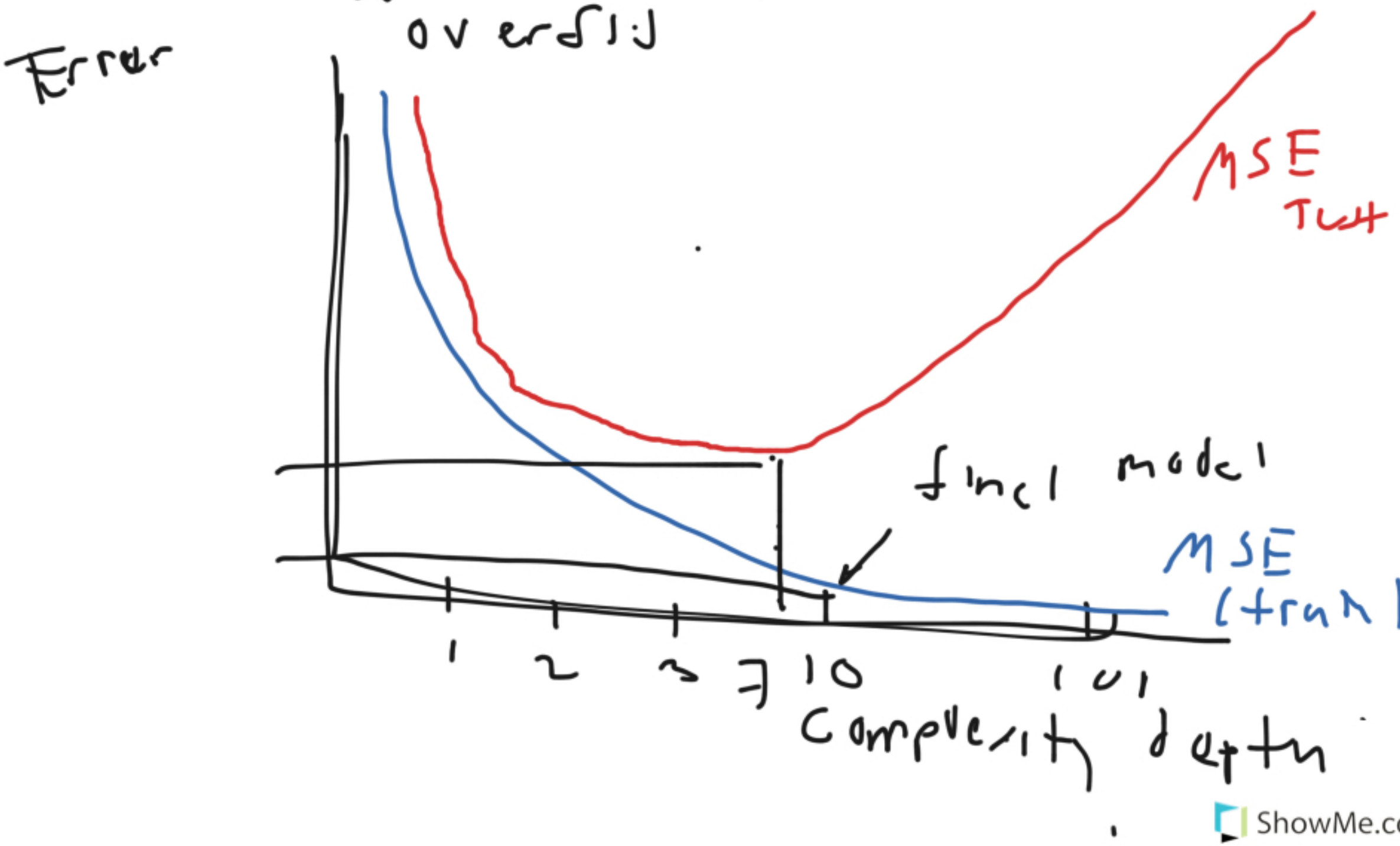
$$H(S_1) = 1$$

$$H(S_2) = 0$$

Random Forests

– an ensemble of decision trees

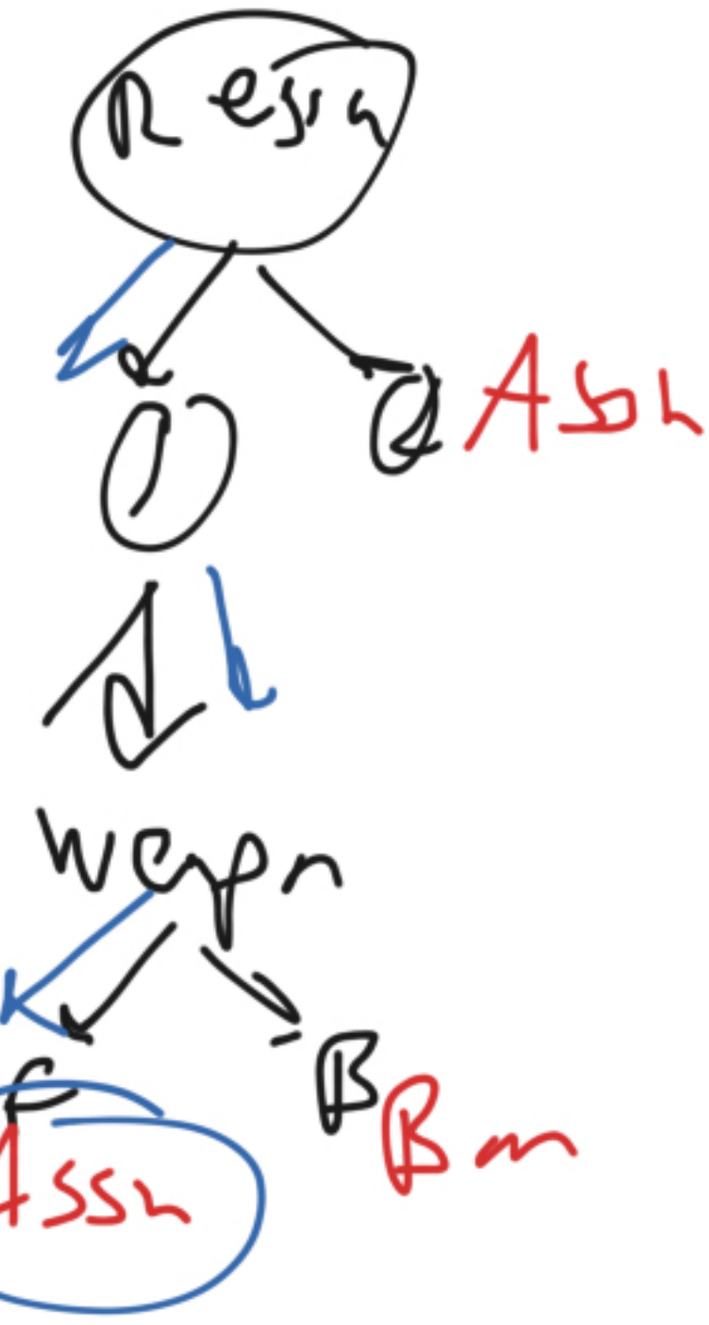
overfit



Region	Day	Weapon
1	1	F
.	.	.
.	.	.
2	2U	Bom

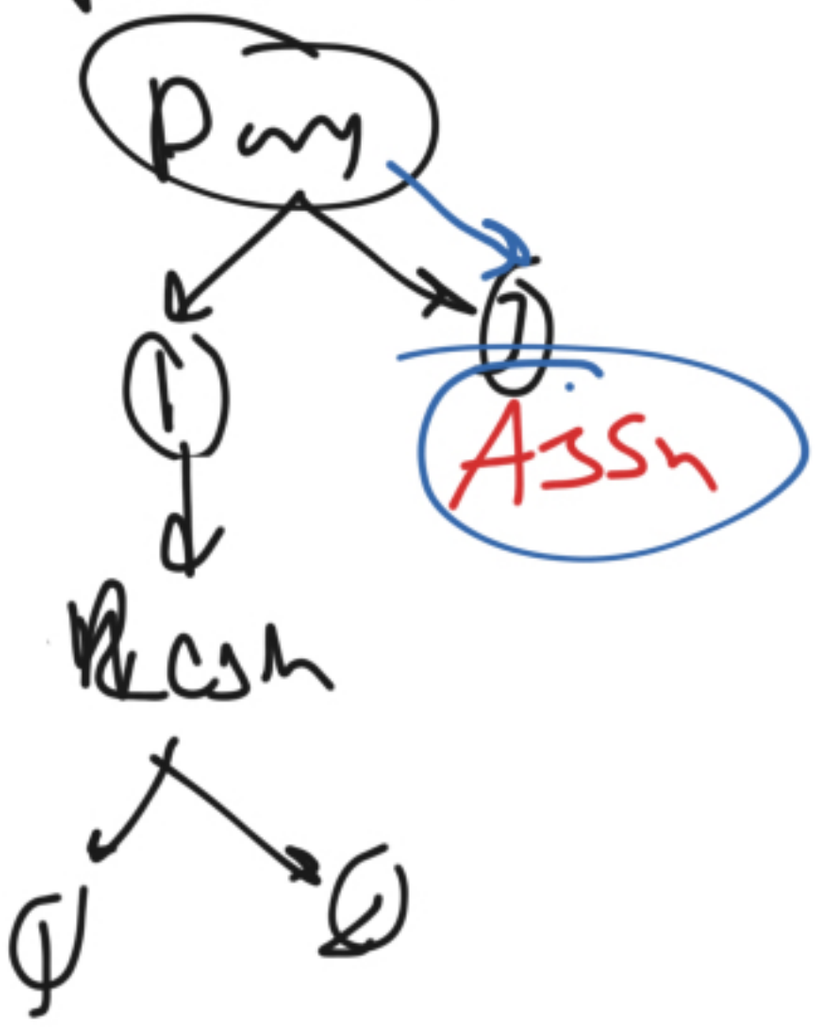
- S_1 (Region, Weapon)
- S_2 (Region, Day)
- S_3 (Region, Day)

Tree 1



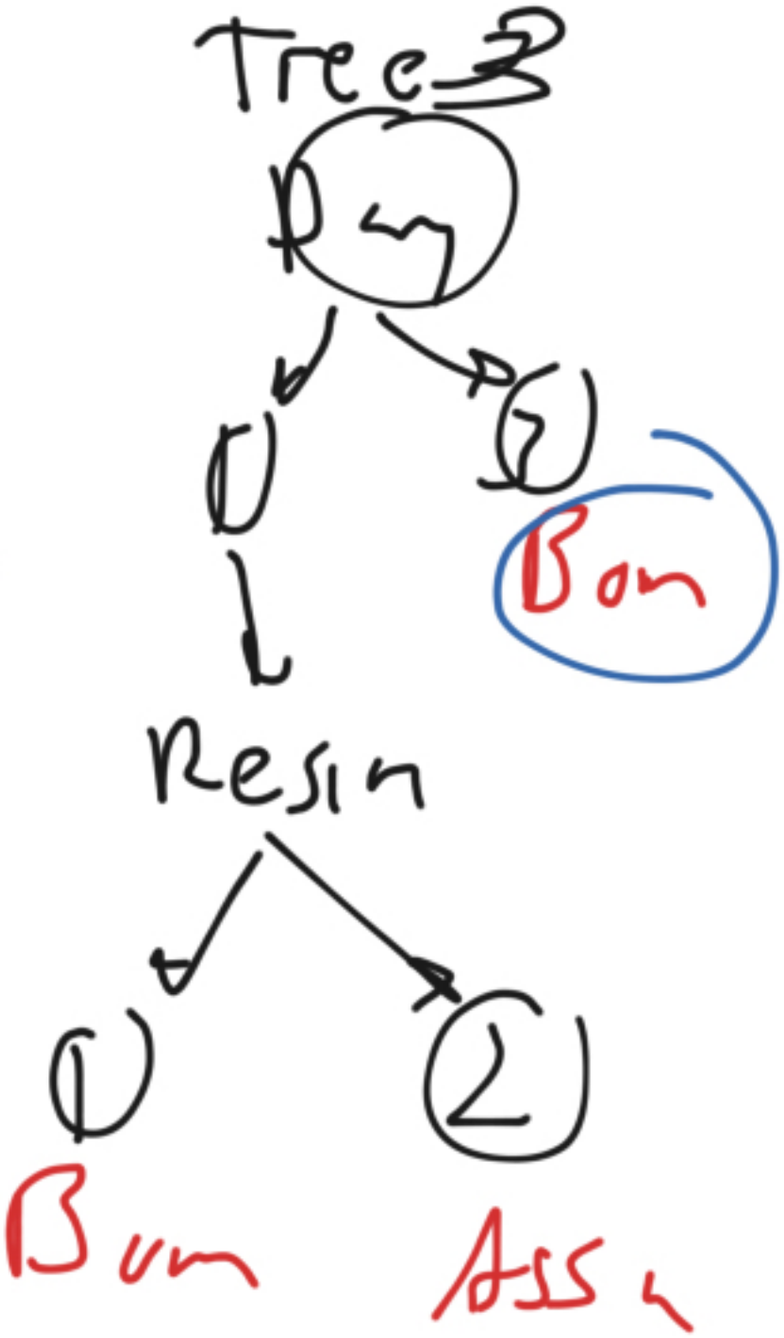
Assn

Tree 2



Assn

Tree 3



Bun

Resin | Bny 3 weyn t

$$P(\text{Ass}_n \mid \text{Trees}) = \frac{2}{3}$$

$$P(\text{Bus} \mid \text{Trees}) = \frac{1}{3}$$

Majority vote = Ass$_n$ ✓

2/3