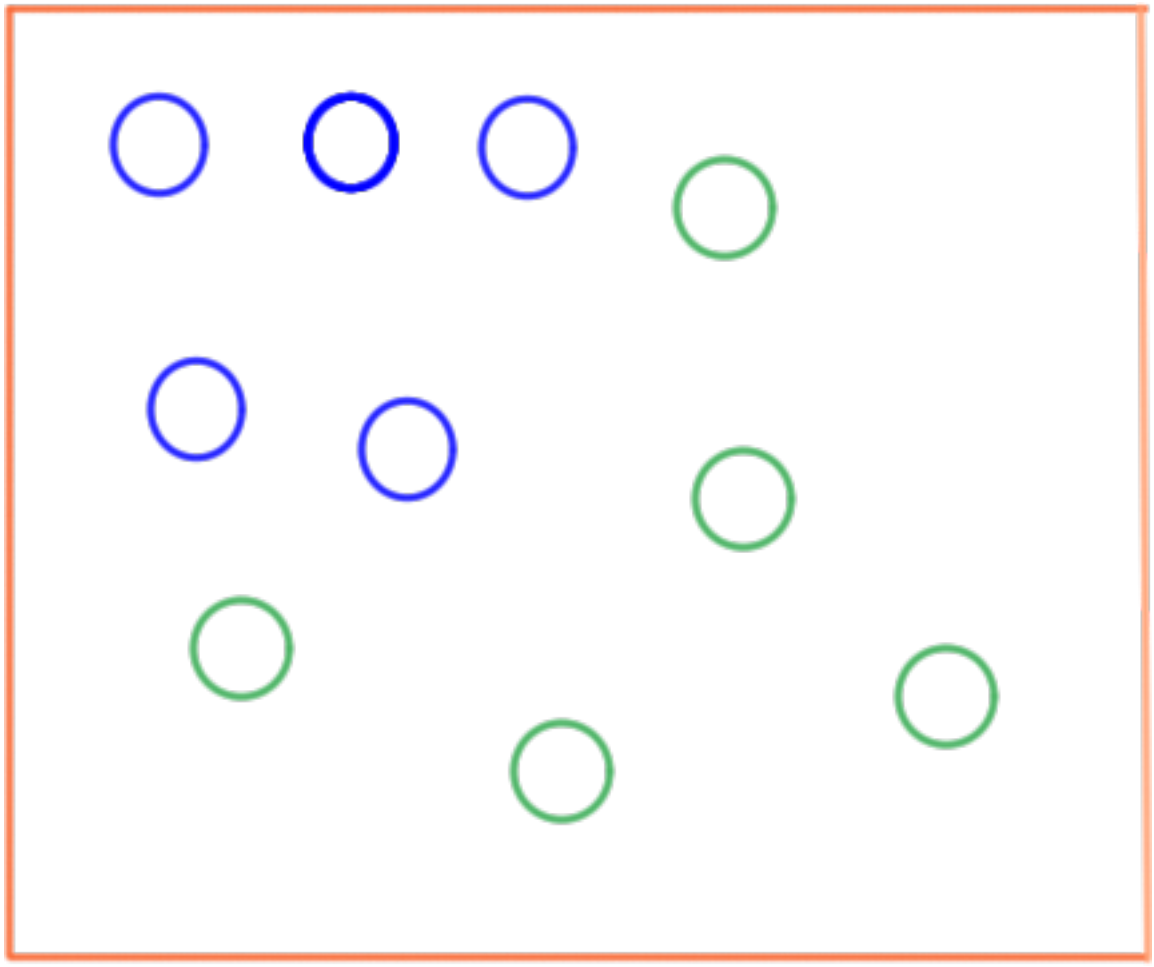
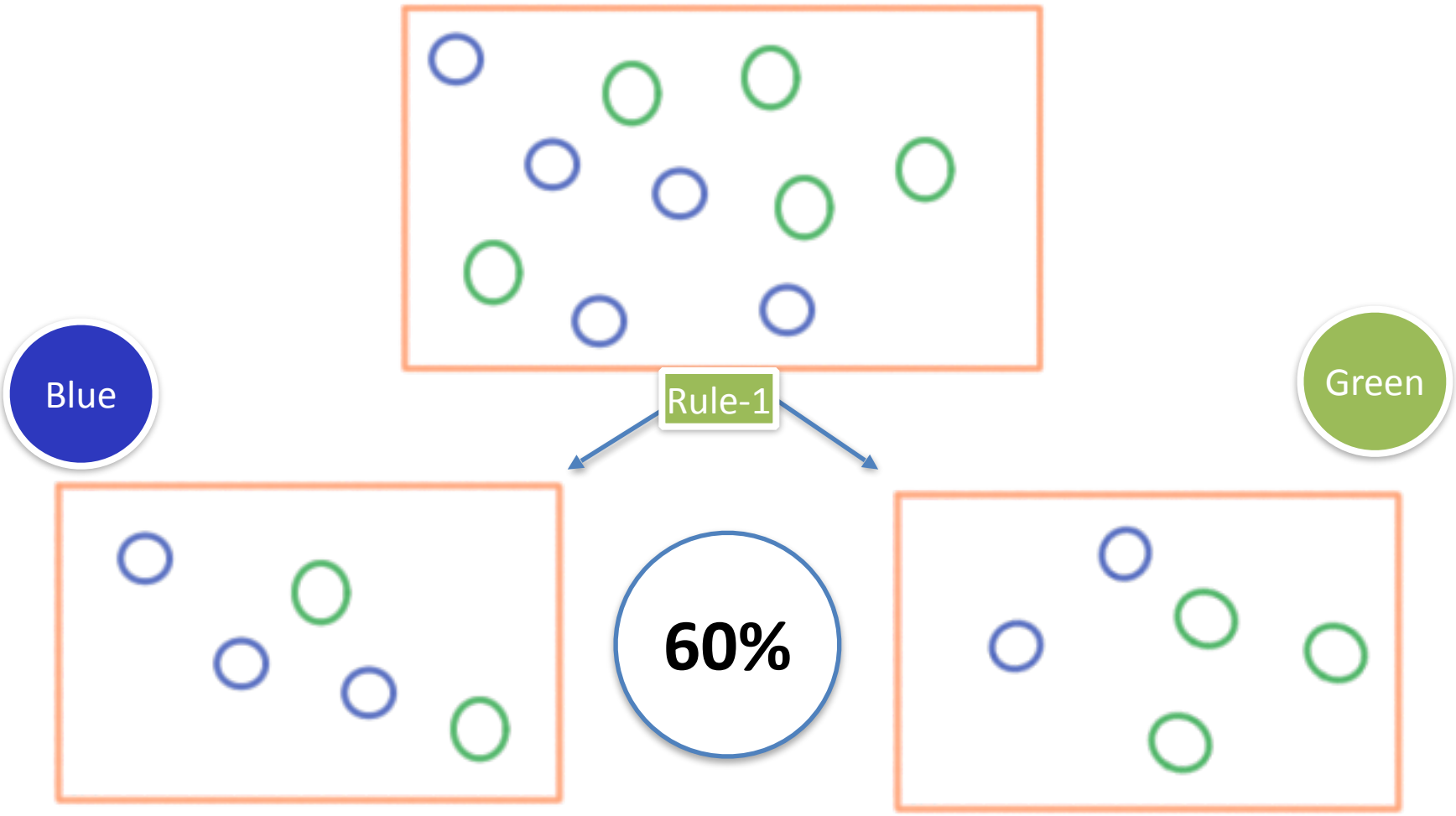
Decision Trees

Partitioning Data & Majority Vote

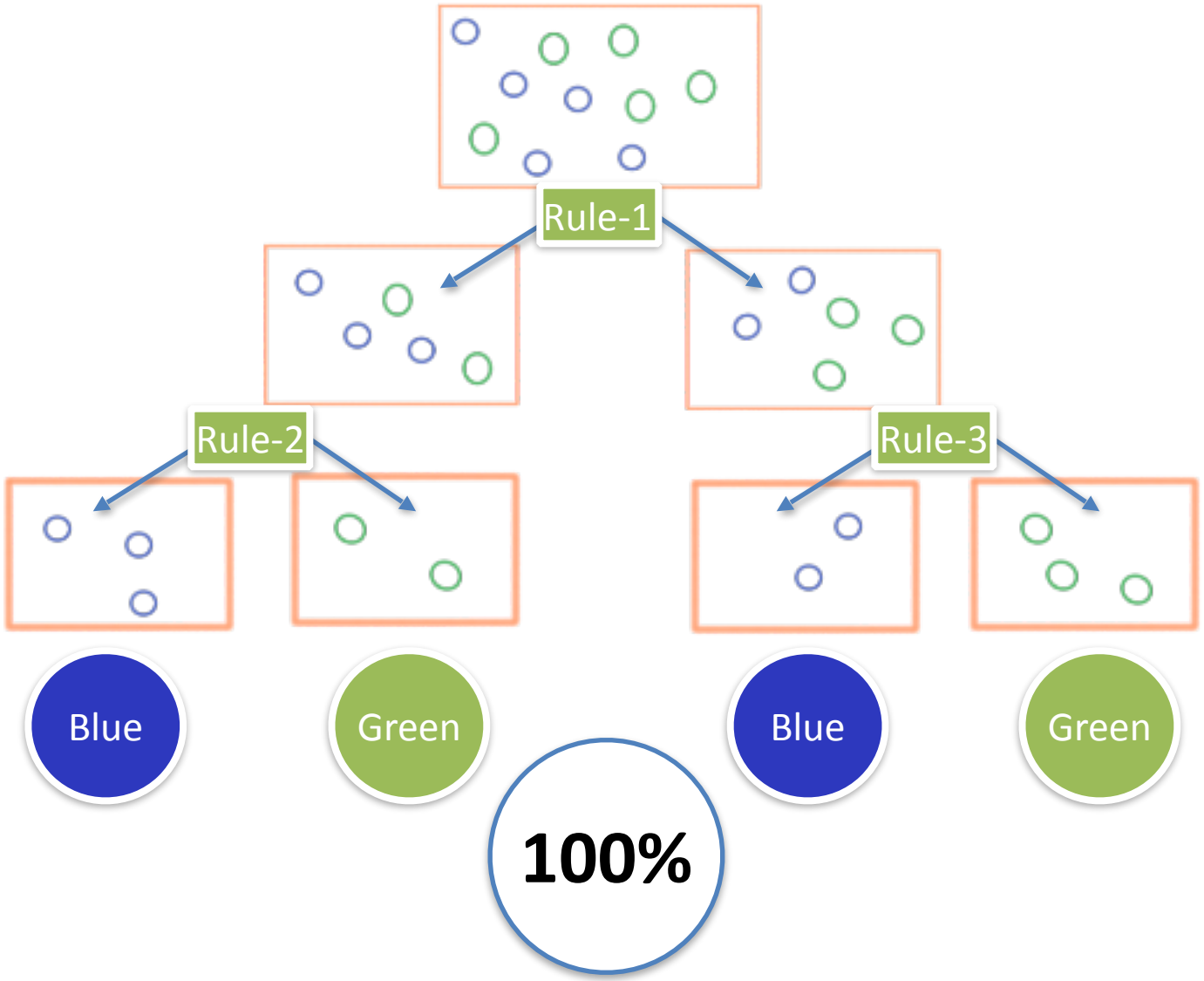
No Split :: Majority Vote :: Accuracy



One Split



One More Split

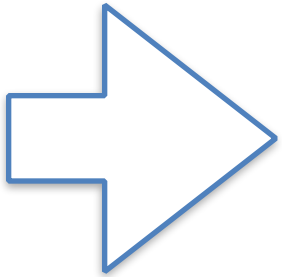


Making & Choosing Rules

Quantifying Better Splits

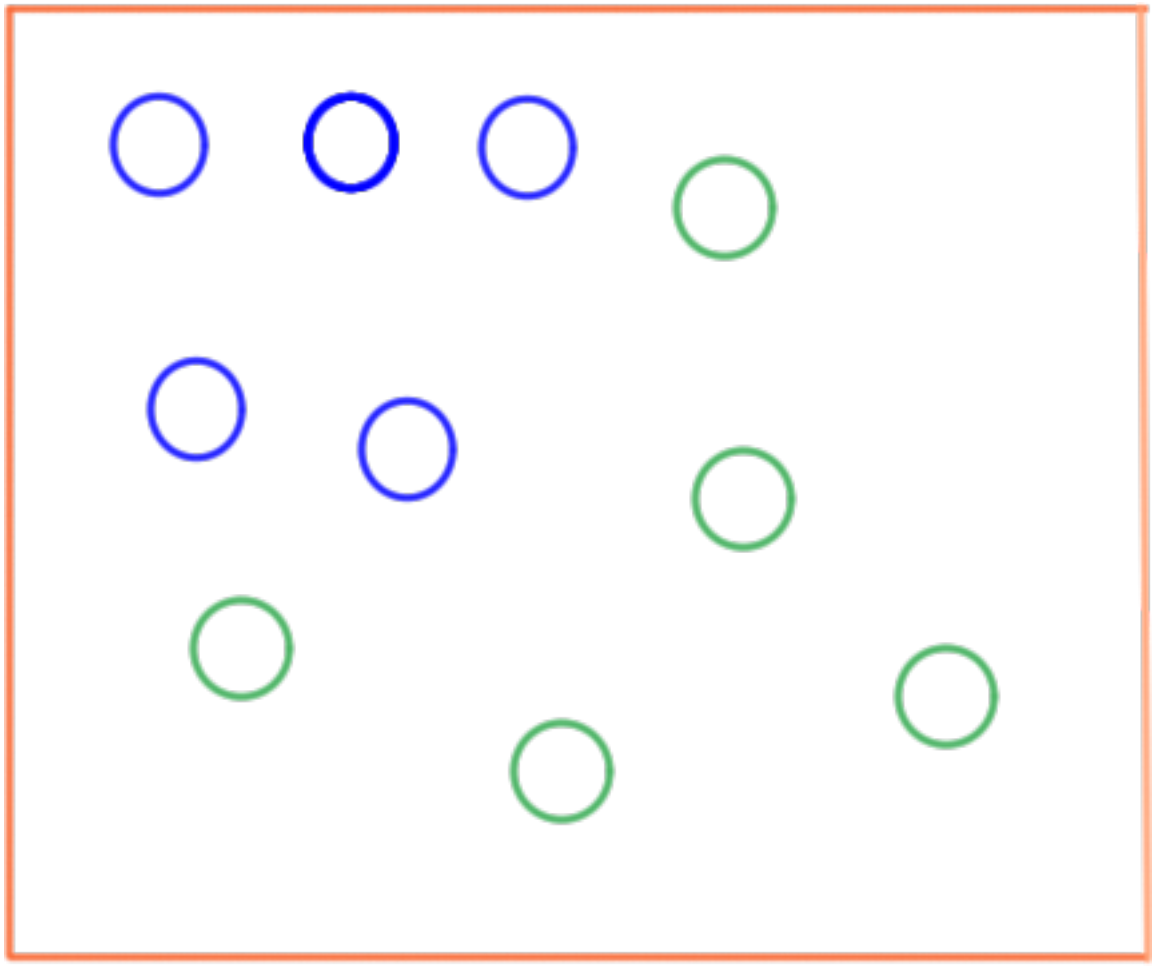
$$Entropy H = -\sum_{i=1}^{k} p_i * log(p_i)$$

Goal



Bringing Down
Entropy of the system

No Split

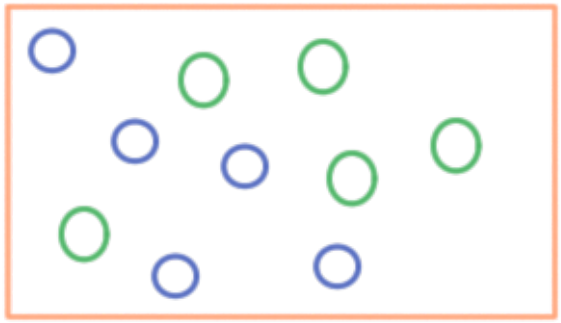


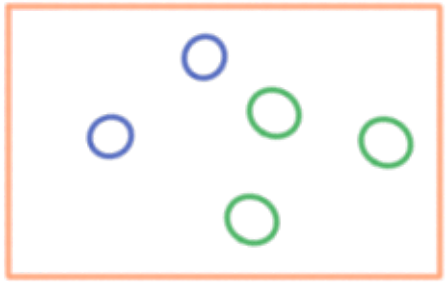
Entropy = -[0.5 * log (0.5) + 0.5 * log (0.5)]

=log(2)

=1

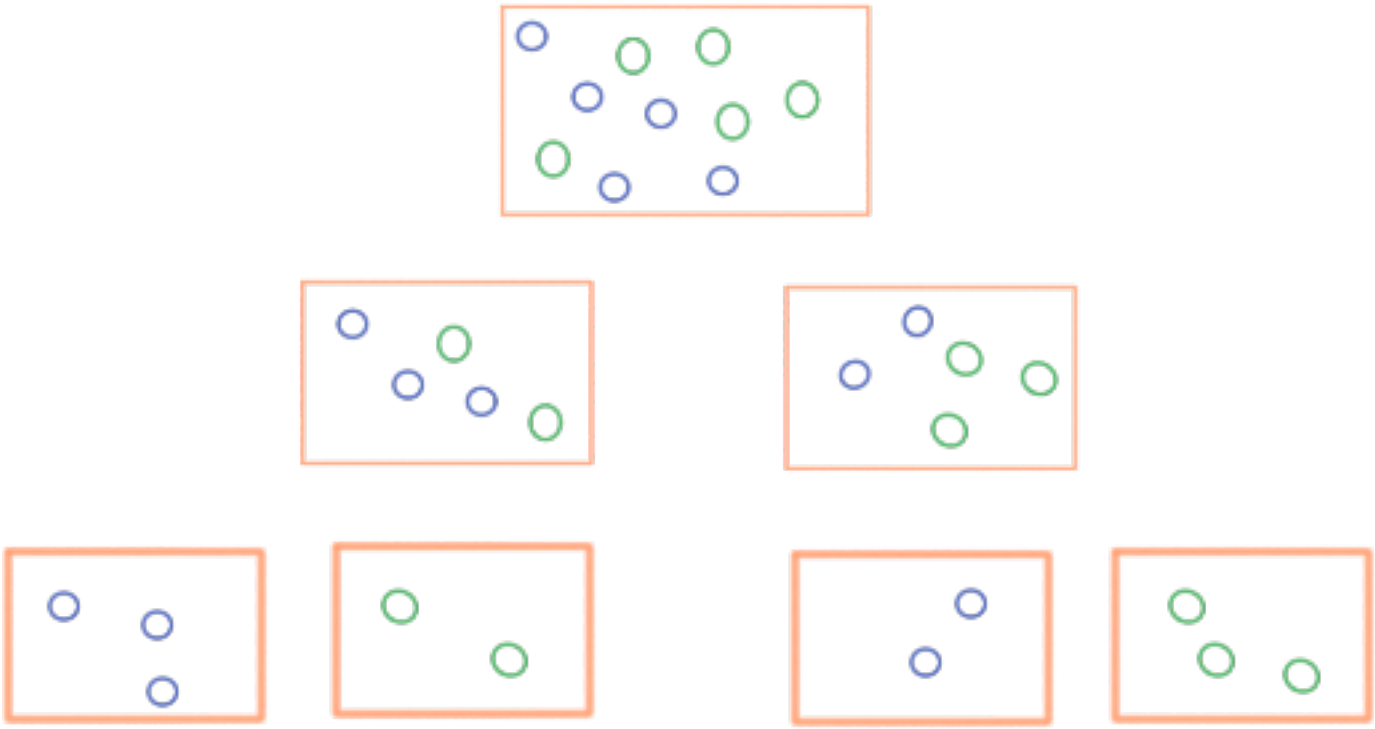
One Split





Entropy = -0.5 * [0.6 *log(0.6) + 0.4*log(0.4)] -0.5 * [0.6 *log(0.6) + 0.4*log(0.4)]
=0.97 Slight Improvement!

One more split



Entropy = 0 . No further split necessary

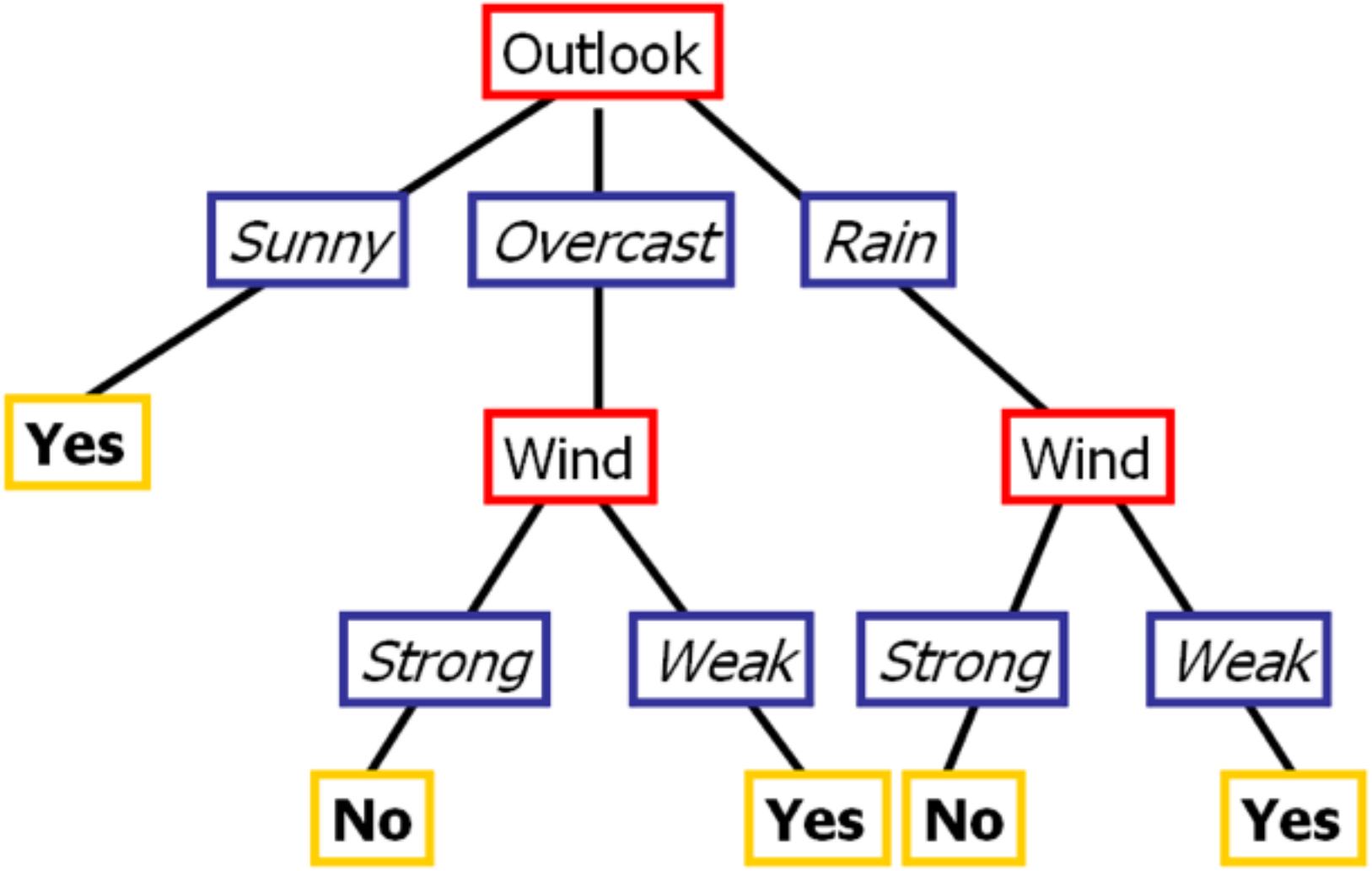
Another Measure : Gini Impurity

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

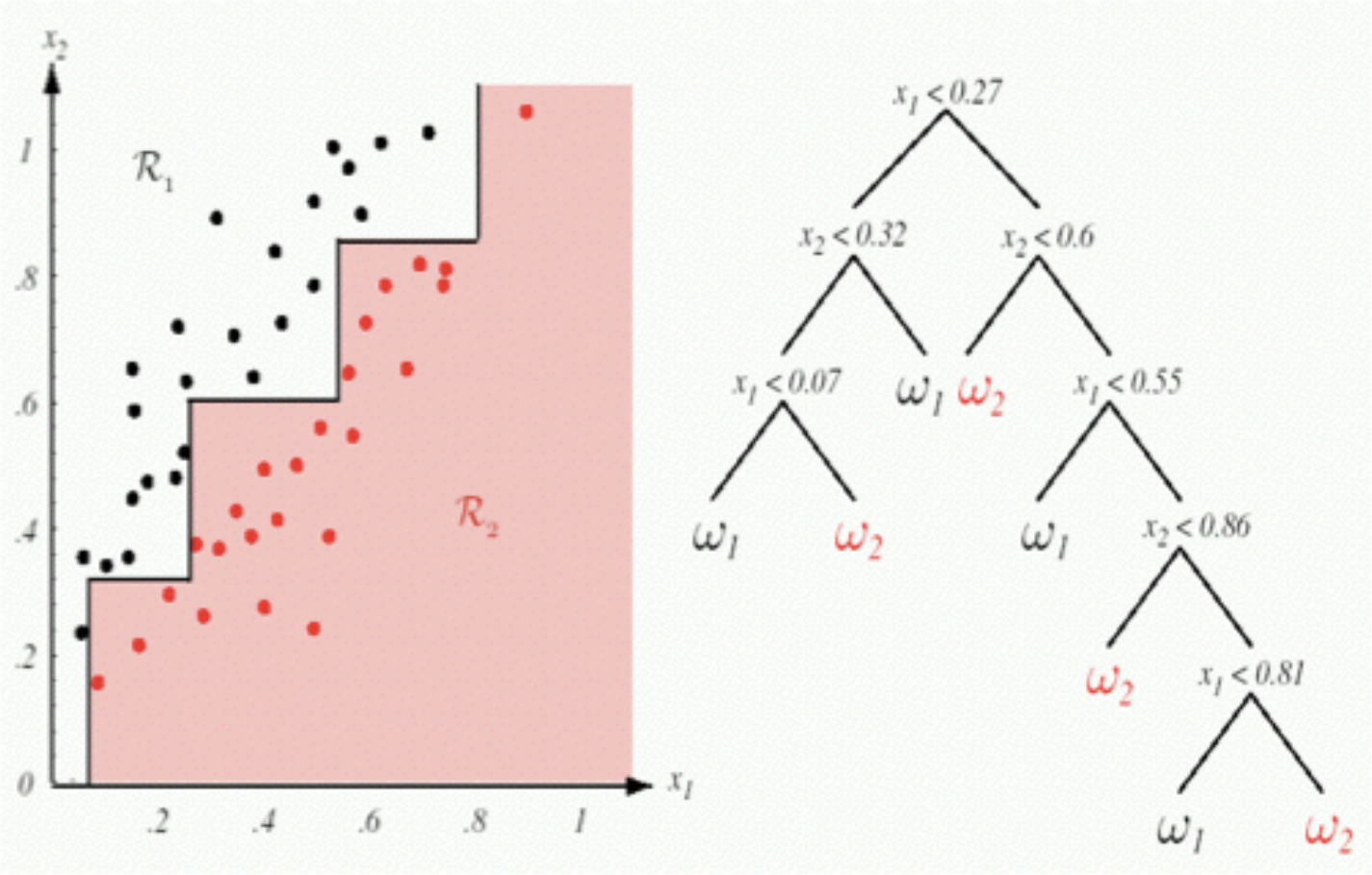
Making Rules: To Tennis or Not to Tennis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

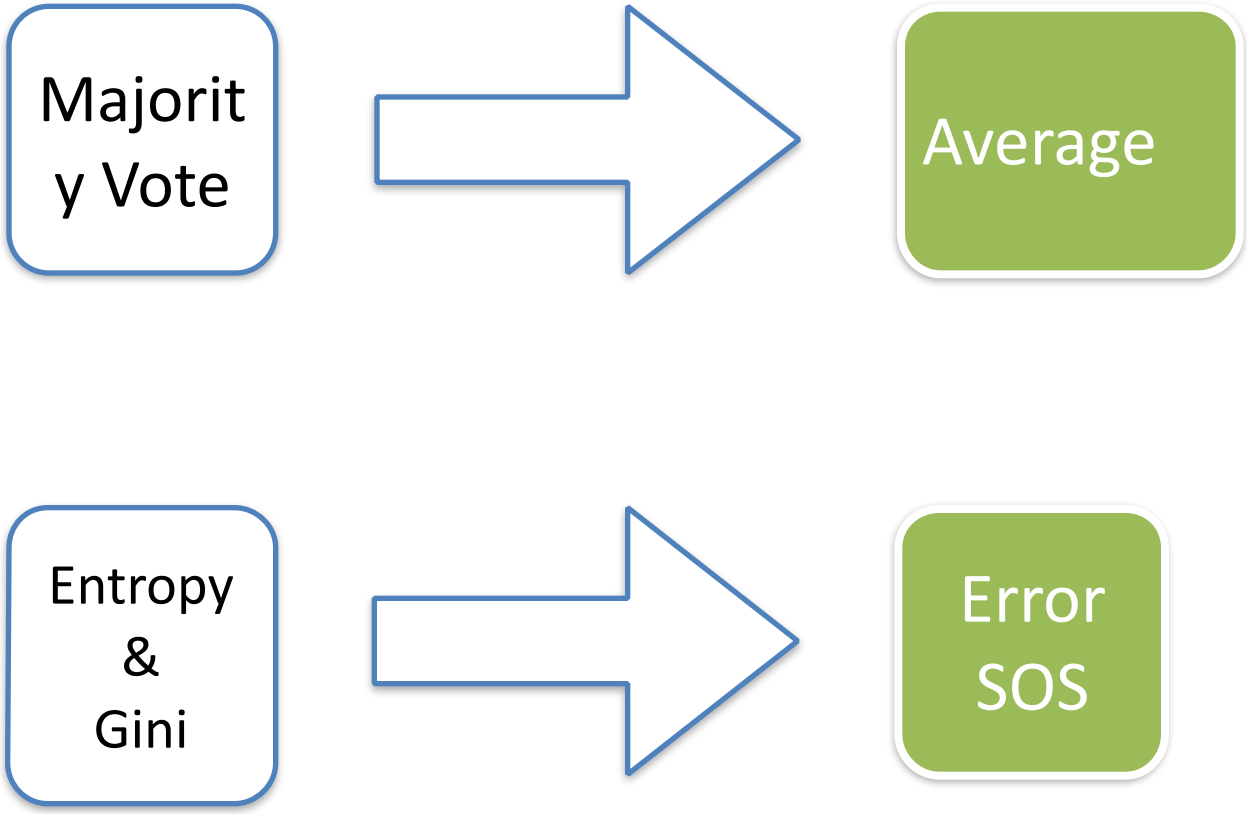
Selected Rules and Decision Tree



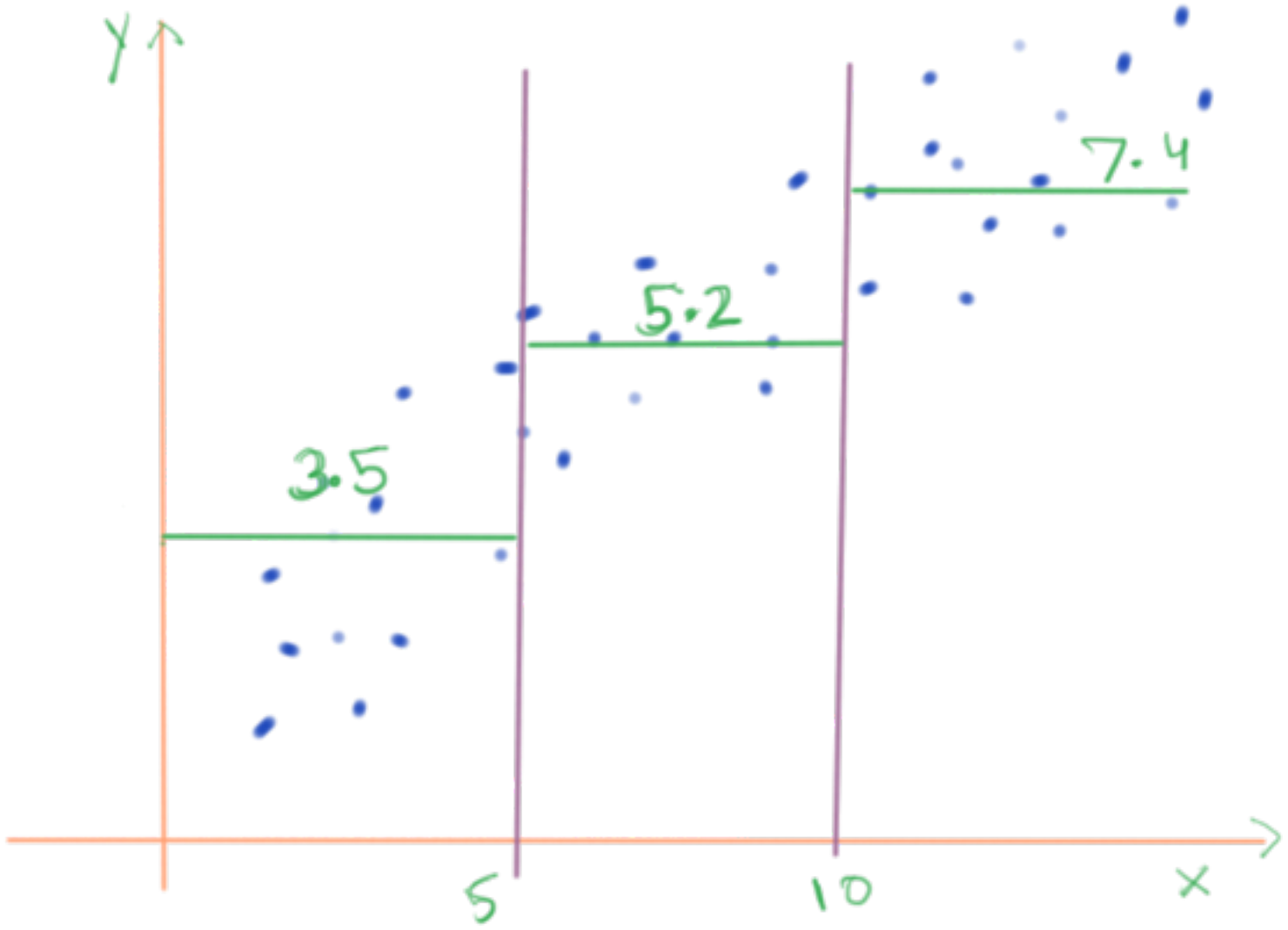
Rules with Continuous Predictors



Regression Tree



Regression Tree :Example



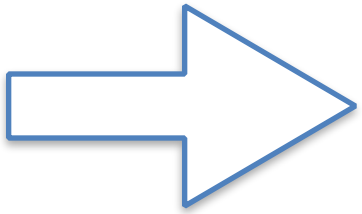
When to stop

- No split required for a homogenous group
- Decide a size for a group to be split
- Decide a minimum requirement for improvement for any split

Drawbacks & Remedies

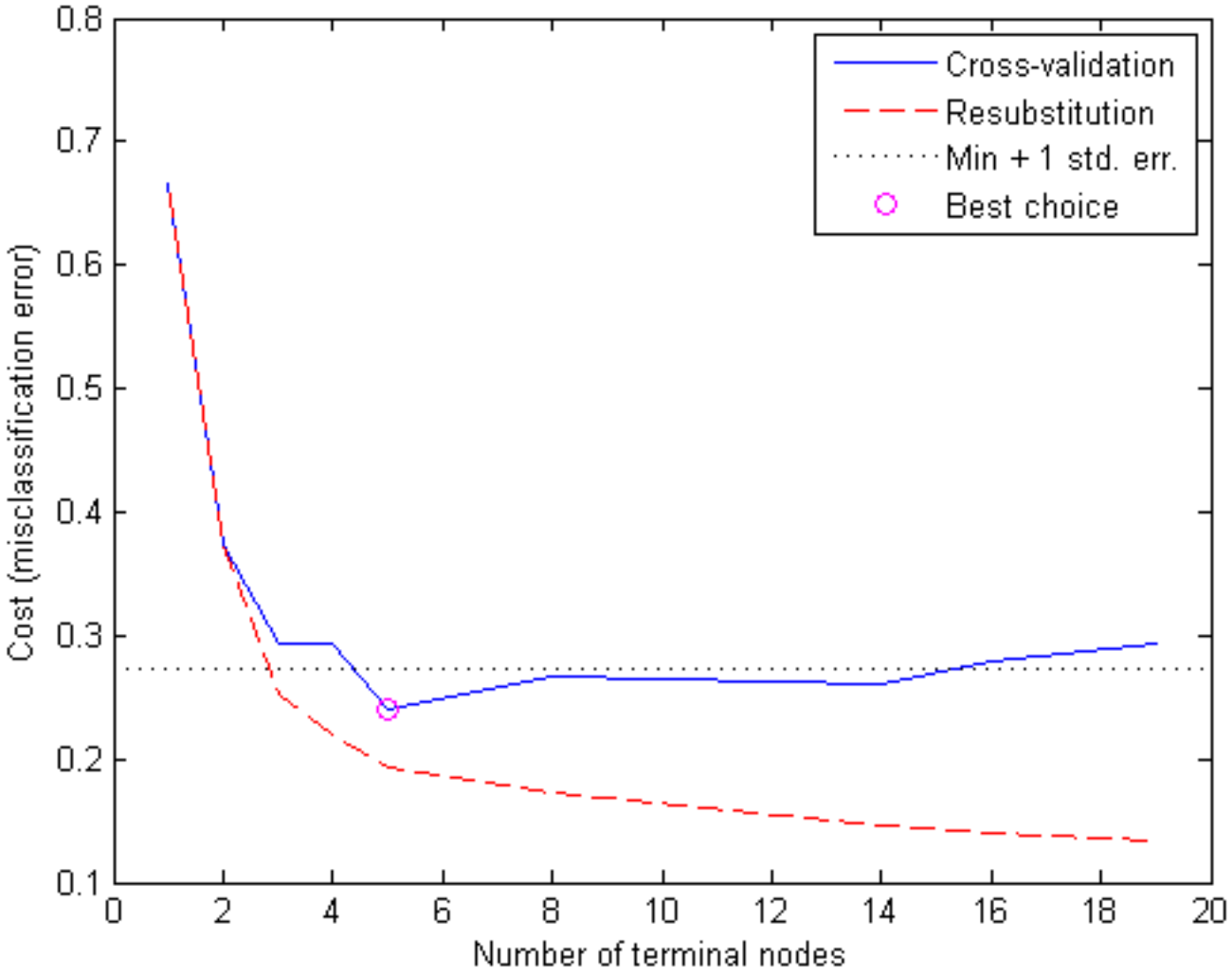
Drawbacks

High Bias / Overfit



Bad Generalisation

Cross-Validation : Pruning Your Tree



Random Forests

Noise in the data

From Observations

From Variables

Random Forest