Decision Trees

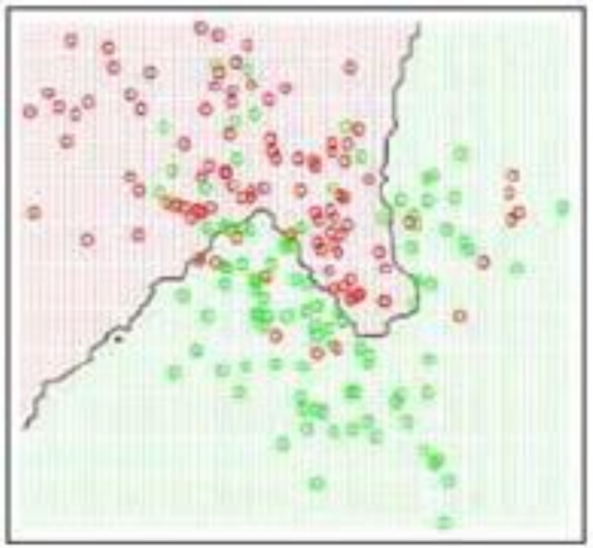
Conceptual Overview

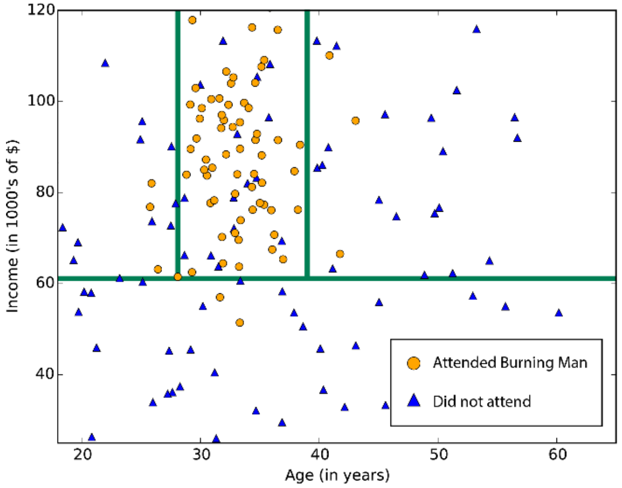
Decision Trees | Statistical Decision Theory | K-Nearest Neighbor

- Statistical Decision Theory
  - The best prediction of Y at an point X=x is the conditional mean. (L2 loss)

- knn
  - At each point x, approximate y by averaging all y_i with input x_i near x
  - Near x = k nearest neighbors
  - Locally constant approximation

- Decision Tree
  - At each point x, approximate y by averaging all y_i with input x_i near x
  - Near x = Region in which x lies | Find the region optimally
  - Locally constant approximation
    - M5 variant of decision tree embeds linear regression in each leaf





Decision Trees

Versatility

- Can be used for classification, regression & clustering
- Effectively handle missing values.
- Can be adapted to streaming data.

Predictive Accuracy

- Not so great.
- But : Bagging, Boosting, Random Forests

Interpretability

- Easy to understand / present / visualize
- Human interpretable rules
- Allow post processing: Rules systems

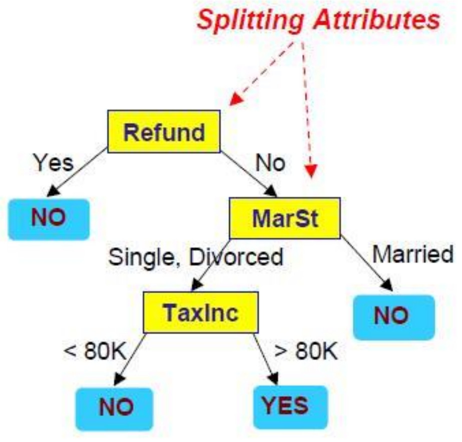
Model Stability

- High Variance: Strong dependence on training set.
- But : Bagging, Boosting, Random Forests

Building & Using Trees

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

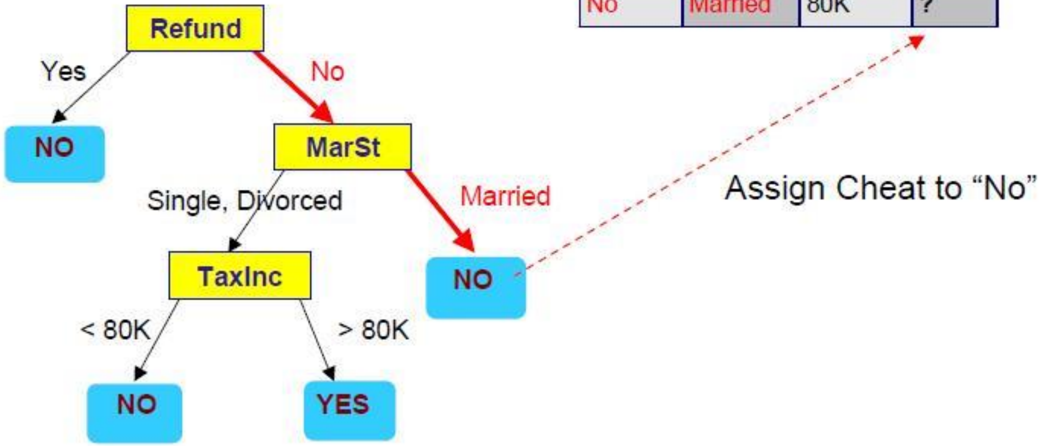
Training Data



Model: Decision Tree

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Build

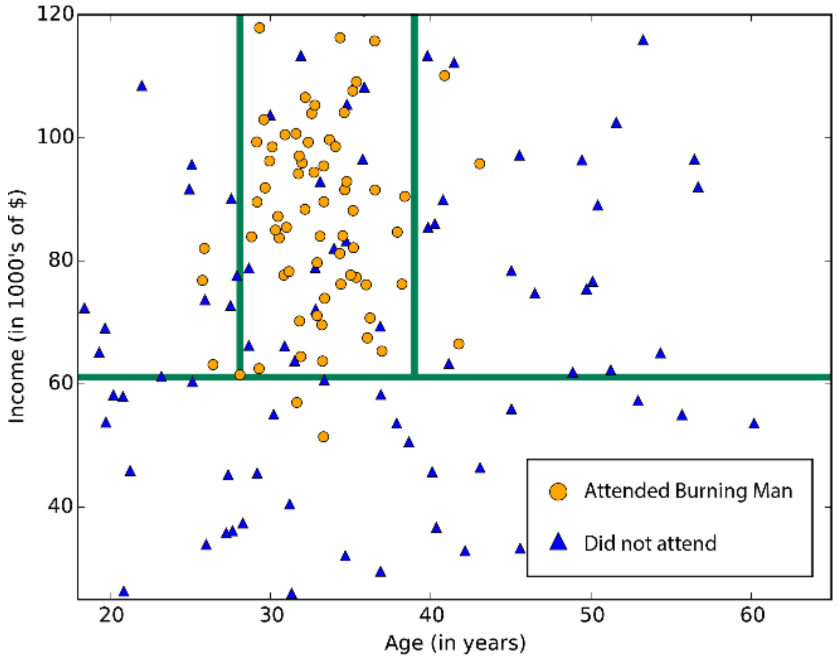
- Think : "If, Then" rules specified in the feature space.
- Greedily divide (binary split) the feature space into distinct, non-overlapping regions

Use

- Every observation mapped to a leaf node assigned the label most commonly occurring in that leaf (Classification)
- Every observation mapped to a leaf node assigned the mean of the samples in the leaf (Regression)
- "Natural" clustering given the target variable.

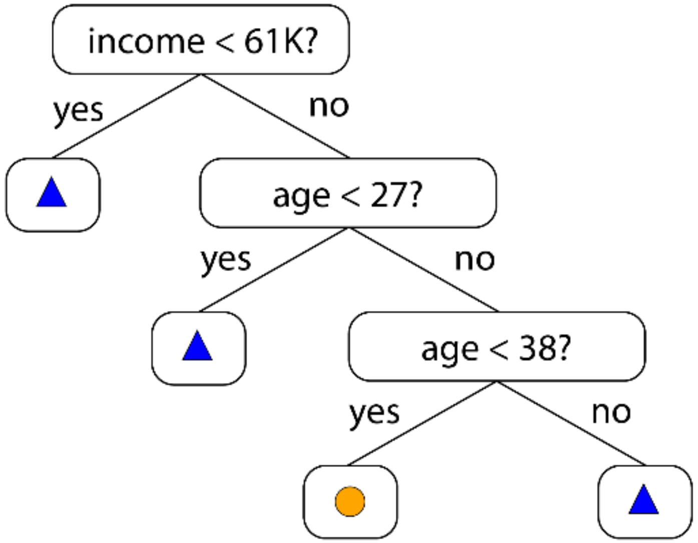
Decision Trees : Continuous splitting of the feature spaces

- In Feature Space



- The feature space contains all data
- Divided regions contain “homogeneous” data subsets
- Region boundaries define regions *(homogeneous data)*

- As a Tree



- Root contains all data
- Leaves contain “homogeneous” data subsets
- Paths along branches define leaves *(homogeneous data)*

Decision Trees : Key Variations

- How to split?
  - What criteria should be used to evaluate a split?
  - What is the split trying to achieve?
  - How do you measure the homogeneity of a subset?
  - In Classification / Regression
  - Supervised Clustering

- When to stop splitting (Avoiding overfitting)
  - Maximum depth / height
  - Minimum number of nodes
  - Grow & Prune
    - Complexity Parameter : Penalty parameter for # nodes

- Other Variations
  - Handling missing values
    - Different category, surrogate splits etc.
  - More than two child nodes
    - One variable appears only once in the tree

- Algorithm names
  - CART
  - C4.5
  - C5.0
  - CHAID
  - ID3
  - …

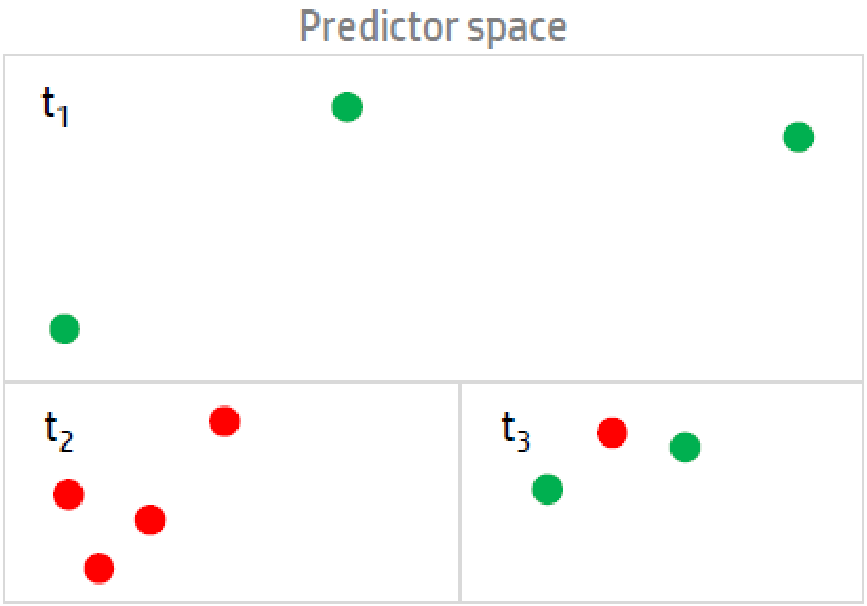
Choosing the Split - Classification

What is a good split?

- Among all possible splits *(all features, all split points)*
- Which split maximizes gain / minimizes error *(Greedy)*
- Information Gain / Impurity reduction.

Choosing feature, split-point

- Cluster “homogeneous” data (subset of data)
- What is a good split measure?
  - Classification Error $1-\max_j p_j$
  - Gini Index $p_1(1-p_2)+p_2(1-p_1)$
  - Entropy $p_1\log(p_1)+p_2\log(p_2)$



$$i(t_1) = 1 - \max\{p_g, p_r\} = 1 - \max\left\{\frac{3}{3}, \frac{0}{3}\right\} = 0$$

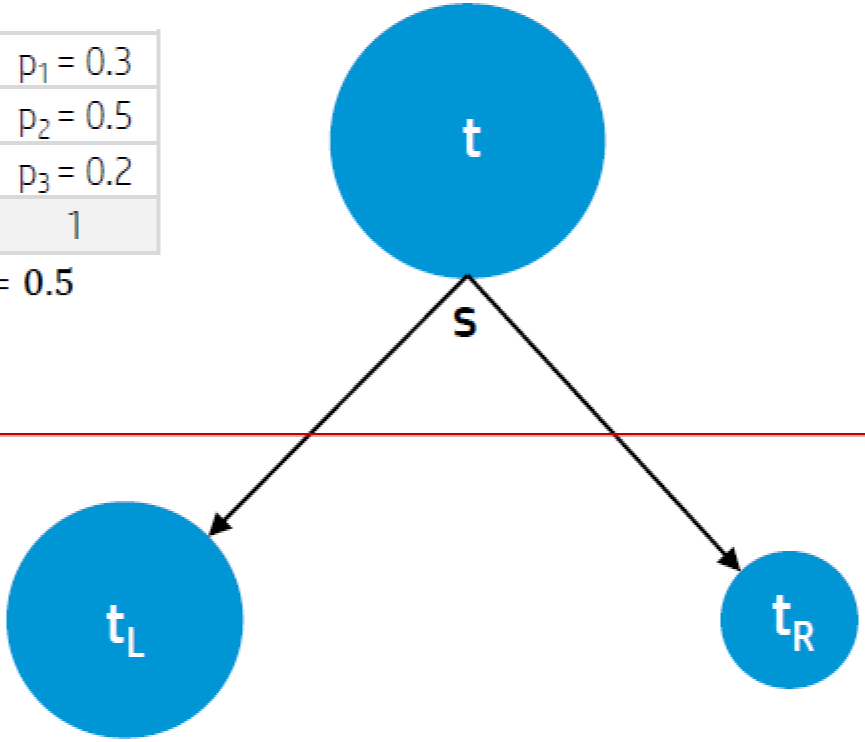
$$i(t_2) = 1 - \max\{p_g, p_r\} = 1 - \max\left\{\frac{0}{4}, \frac{4}{4}\right\} = 0$$

$$i(t_3) = 1 - \max\{p_g, p_r\} = 1 - \max\left\{\frac{2}{3}, \frac{1}{3}\right\} = 0.33$$

Impurity = Classification Error Rate

Class 1	$n(t_1) = 60$	$p_1 = 0.3$
Class 2	$n(t_2) = 100$	$p_2 = 0.5$
Class 3	$n(t_3) = 40$	$p_3 = 0.2$
Total	$n(t) = 200$	1

$$i(t) = 1 - (0.5) = 0.5$$



Class 1	$n(t_1) = 10$	$p_1 = 0.07$
Class 2	$n(t_2) = 100$	$p_2 = 0.66$
Class 3	$n(t_3) = 40$	$p_3 = 0.27$
Total	$n(t) = 150$	1

$$i(t_L) = 1 - 0.66 = 0.33$$

Class 1	$n(t_1) = 50$	$p_1 = 1.0$
Class 2	$n(t_2) = 0$	$p_2 = 0.0$
Class 3	$n(t_3) = 0$	$p_3 = 0.0$
Total	$n(t) = 50$	1

$$i(t_R) = 1 - 1.0 = 0$$

$$\frac{150}{200} \times 0.33 + \frac{50}{200} \times 0 = 0.25$$

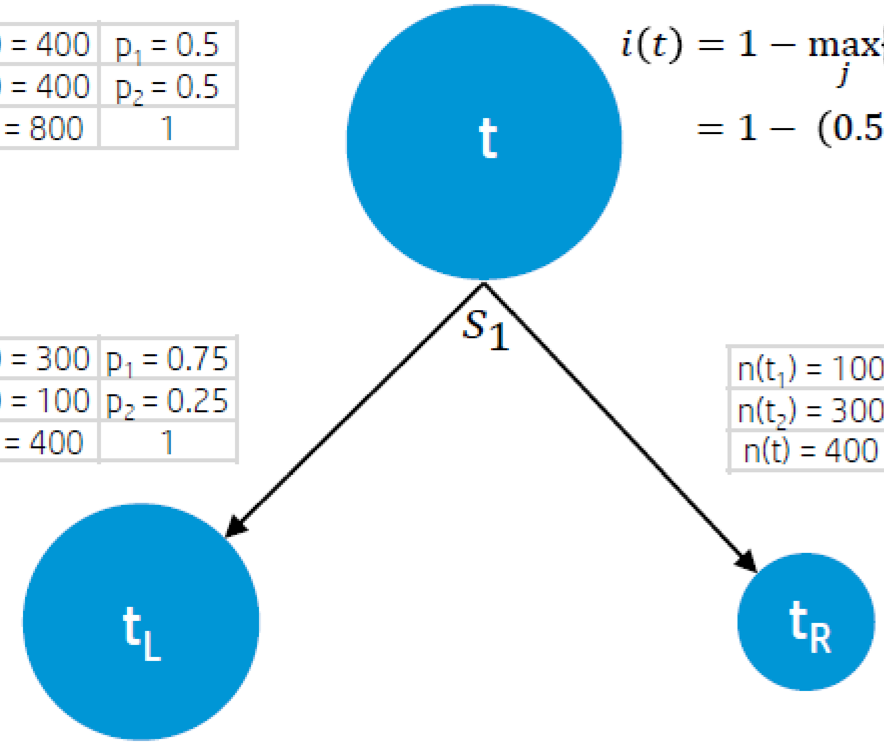
$$\Delta\, i(s,t) = 0.5 - 0.25 = 0.25$$

maximize $\{$ Information Gain $\}$

Impurity = Classification Error Rate (cont'd)

Split s_1

$n(t_1) = 400$	$p_1 = 0.5$
$n(t_2) = 400$	$p_2 = 0.5$
$n(t) = 800$	1



$$i(t) = 1 - \max_j\{p_j\}$$
$$= 1 - (0.5) = 0.5$$

$n(t_1) = 300$	$p_1 = 0.75$
$n(t_2) = 100$	$p_2 = 0.25$
$n(t) = 400$	1

$n(t_1) = 100$	$p_1 = 0.25$
$n(t_2) = 300$	$p_2 = 0.75$
$n(t) = 400$	1

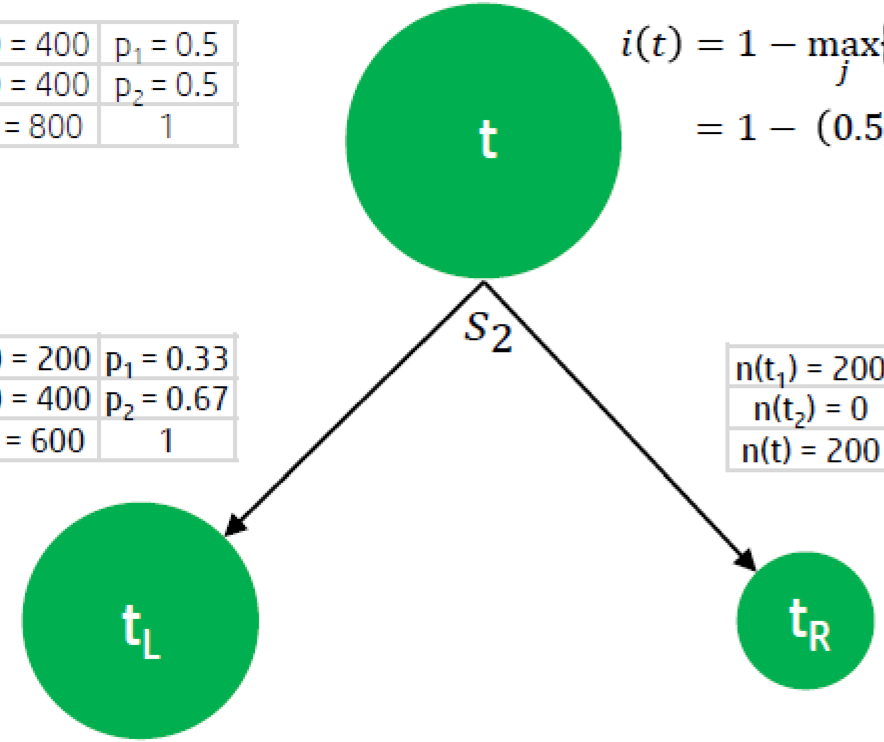
$$i(t_L) = 1 - 0.75 = 0.25$$

$$i(t_R) = 1 - 0.75 = 0.25$$

$$\Delta\, i(s,t) = 0.5 - \left[\frac{400}{800} \times 0.25 + \frac{400}{800} \times 0.25\right] = 0.25$$

Split s_2

$n(t_1) = 400$	$p_1 = 0.5$
$n(t_2) = 400$	$p_2 = 0.5$
$n(t) = 800$	1



$$i(t) = 1 - \max_j\{p_j\}$$
$$= 1 - (0.5) = 0.5$$

$n(t_1) = 200$	$p_1 = 0.33$
$n(t_2) = 400$	$p_2 = 0.67$
$n(t) = 600$	1

$n(t_1) = 200$	$p_1 = 1.00$
$n(t_2) = 0$	$p_2 = 0.00$
$n(t) = 200$	1

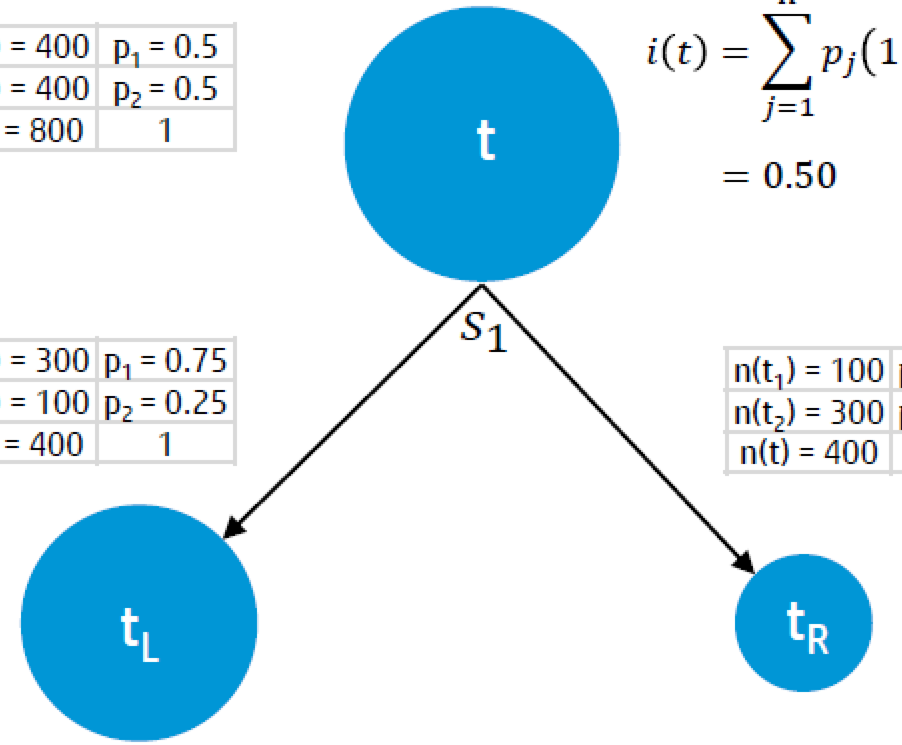
$$i(t_L) = 1 - 0.67 = 0.33$$

$$i(t_R) = 1 - 1.00 = 0.0$$

$$\Delta\, i(s,t) = 0.5 - \left[\frac{600}{800} \times 0.33 + \frac{200}{800} \times 0.0\right] = 0.25$$

Impurity = Gini Index

n(t$_1$) = 400	p_1 = 0.5
n(t$_2$) = 400	p_2 = 0.5
n(t) = 800	1



$$i(t) = \sum_{j=1}^{K} p_j(1-p_j)$$

$$= 0.50$$

n(t$_1$) = 300	p_1 = 0.75
n(t$_2$) = 100	p_2 = 0.25
n(t) = 400	1

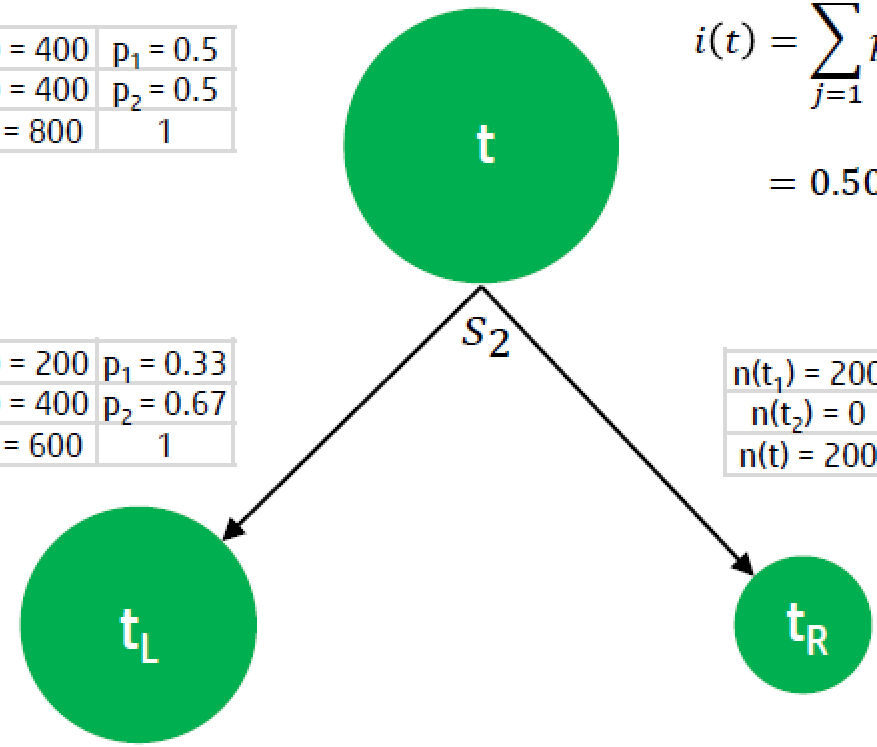
n(t$_1$) = 100	p_1 = 0.25
n(t$_2$) = 300	p_2 = 0.75
n(t) = 400	1

$i(t_L) = 0.38$

$i(t_R) = 0.38$

$$\Delta\, i(s,t) = 0.50 - \left[\frac{400}{800} \times 0.38 + \frac{400}{800} \times 0.38\right] = 0.12$$

n(t$_1$) = 400	p_1 = 0.5
n(t$_2$) = 400	p_2 = 0.5
n(t) = 800	1



$$i(t) = \sum_{j=1}^{K} p_j(1-p_j)$$

$$= 0.50$$

n(t$_1$) = 200	p_1 = 0.33
n(t$_2$) = 400	p_2 = 0.67
n(t) = 600	1

n(t$_1$) = 200	p_1 = 1.00
n(t$_2$) = 0	p_2 = 0.00
n(t) = 200	1

$i(t_L) = 0.44$

$i(t_R) = 0.0$

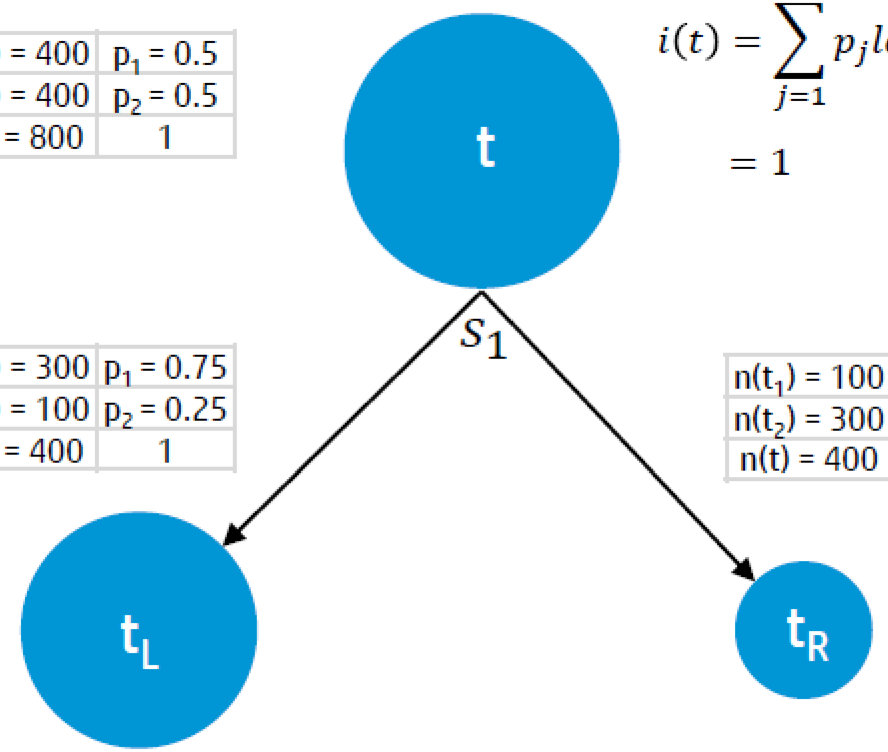
$$\Delta\, i(s,t) = 0.50 - \left[\frac{600}{800} \times 0.44 + \frac{200}{800} \times 0.0\right] = 0.17$$

Impurity = Cross Entropy

n(t_1) = 400	p_1 = 0.5
n(t_2) = 400	p_2 = 0.5
n(t) = 800	1

$$i(t) = \sum_{j=1}^{K} p_j log\left(\frac{1}{p_j}\right)$$

$$= 1$$



n(t_1) = 300	p_1 = 0.75
n(t_2) = 100	p_2 = 0.25
n(t) = 400	1

n(t_1) = 100	p_1 = 0.25
n(t_2) = 300	p_2 = 0.75
n(t) = 400	1

$i(t_L) = 0.81$

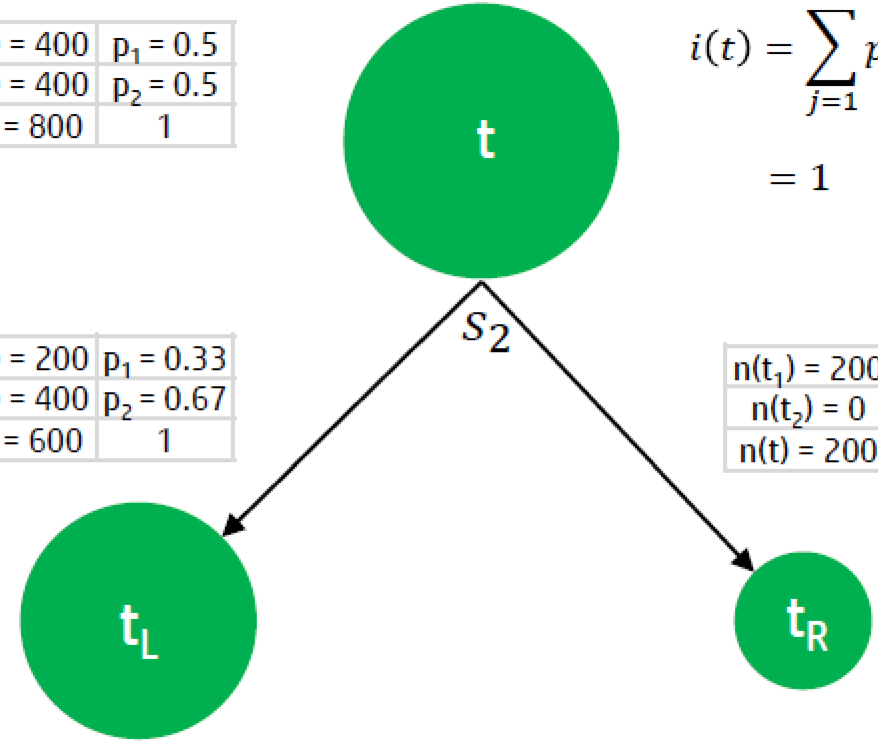
$i(t_R) = 0.81$

$$\Delta\, i(s,t) = 1\ - \left[\frac{400}{800} \times 0.81 + \frac{400}{800} \times 0.81\right] = 0.19$$

n(t_1) = 400	p_1 = 0.5
n(t_2) = 400	p_2 = 0.5
n(t) = 800	1

$$i(t) = \sum_{j=1}^{K} p_j log\left(\frac{1}{p_j}\right)$$

$$= 1$$



n(t_1) = 200	p_1 = 0.33
n(t_2) = 400	p_2 = 0.67
n(t) = 600	1

n(t_1) = 200	p_1 = 1.00
n(t_2) = 0	p_2 = 0.00
n(t) = 200	1

$i(t_L) = 0.91$

$i(t_R) = 0.0$

$$\Delta\, i(s,t) = 1\ - \left[\frac{600}{800} \times 0.91 + \frac{200}{800} \times 0.0\right] = 0.3175$$

Decision Tree

- Function Approximation formulation
- Choosing feature, split-point
  - Cluster “homogeneous” data (subset of data)
  - What is a good split measure?
    - Classification Error $1-\max_j p_j$
    - Gini Index $p_1(1-p_2)+p_2(1-p_1)$
    - Entropy $p_1\log(p_1)+p_2\log(p_2)$
  - CART, C4.5, CHAID, ID3 variants
- When to stop splitting (Avoiding overfitting)
  - Grow & Prune
  - Complexity Parameter : Penalty for # nodes

$$f(X) = \sum_{m=1}^{|T|} c_m \cdot 1_{(X\in R_m)} \quad \text{Decision Tree}$$

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j \quad \text{Linear Regression}$$

$$N_m = \#\{x_i \in R_m\}$$

$$\hat{c}_m = \frac{1}{N_m}\sum_{x_i\in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m}\sum_{x_i\in R_m}(y_i-\hat{c}_m)^2$$

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

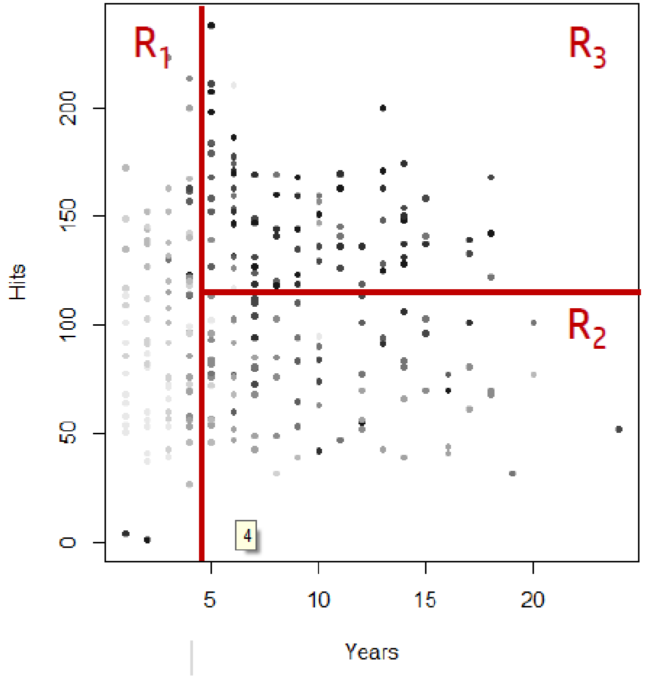
Choosing the Split - Regression

What is a good split?

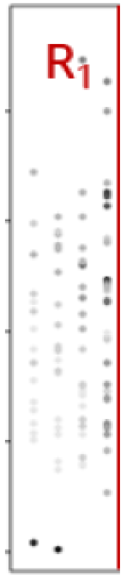
- Among all possible splits *(all features, all split points)*
- Which split maximizes gain / minimizes error *(Greedy)*
- Information Gain / Impurity reduction

Choosing feature, split-point

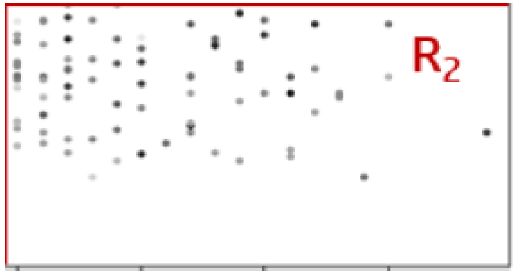
- Contain "homogeneous" data *(subset of data)*
- What is a good split measure?
- Squared Sum of Errors $\sum_{i \in L} (\hat{y}_L - y_{i,L})^2 + \sum_{i \in R} (\hat{y}_R - y_{i,R})^2$



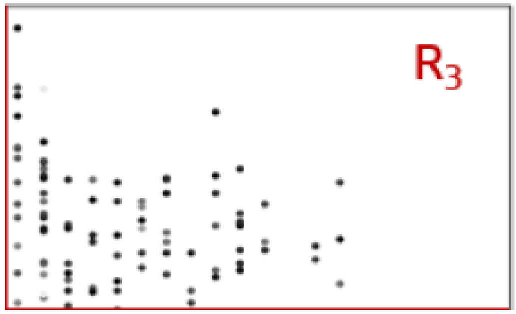
$\hat{y}_{R_1} = 226$



$\hat{y}_{R_2} = 465$



$\hat{y}_{R_3} = 949$



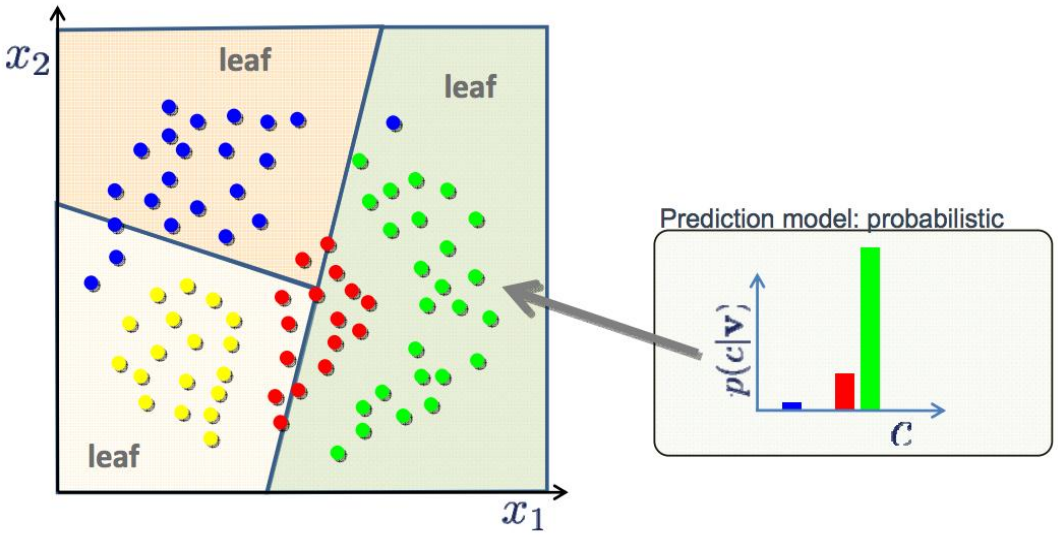
$$\sum_{i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i \in R_2} (y_i - \hat{y}_{R_2})^2 + \sum_{i \in R_3} (y_i - \hat{y}_{R_3})^2 =$$

$$\text{minimize} \left\{ \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \right\}$$

(Additional) Advantages of splitting

Splits : Branches for homogenizing data

- Alternative splits evaluated at build-time
- If an alternative split ~ actual split, use the alternative split at prediction time if variable missing.



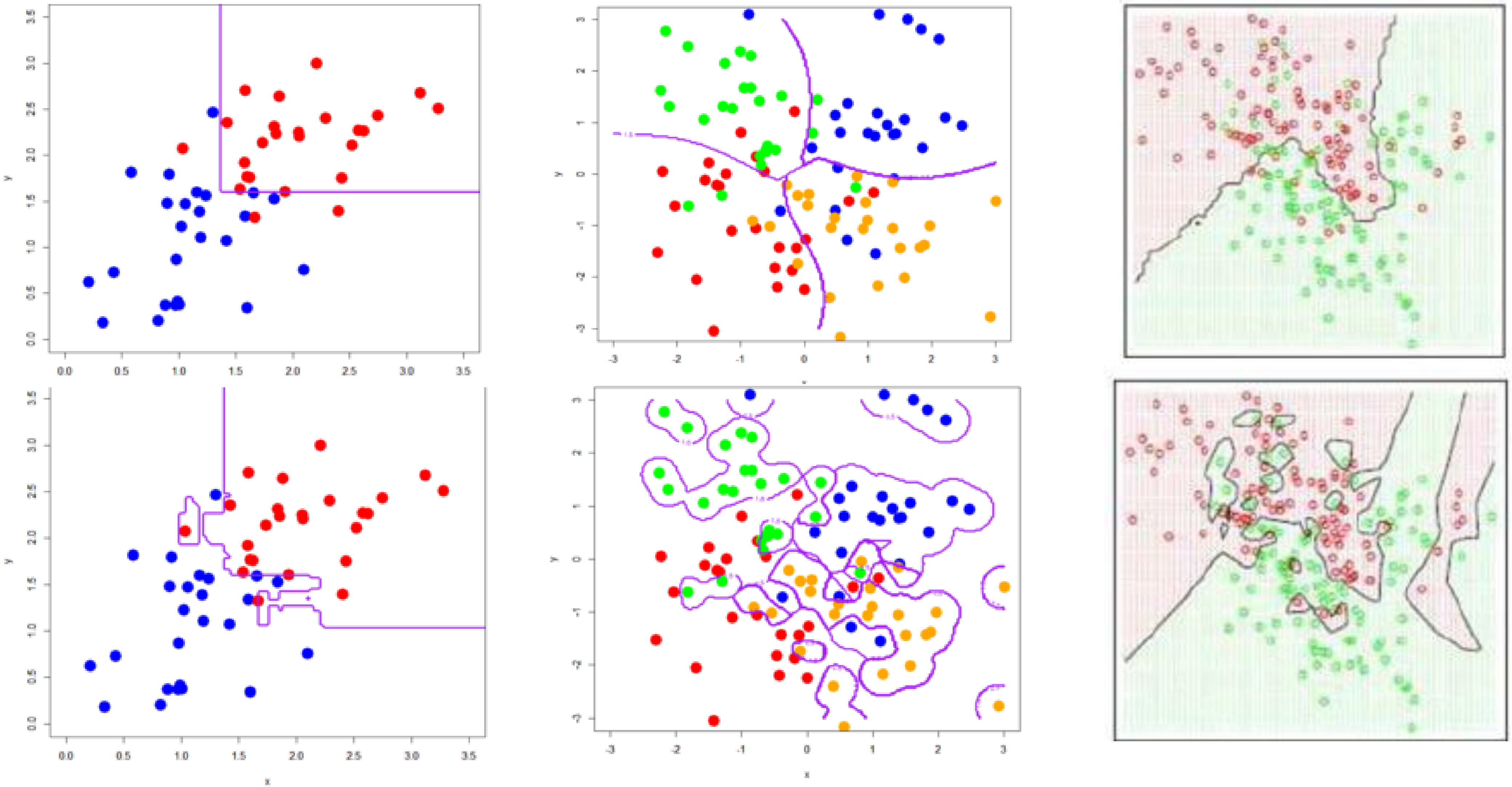
Surrogate splits handle missing values

- Alternative splits evaluated at build-time
- If an alternative split ~ actual split, use the alternative split at prediction time if variable missing.

Feature Importance

- Reduction in Optimization Criteria due to splits containing feature.
- Features which appear higher and more often more important.

All models must guard against Overfitting …



When to stop splitting?

When will we be "forced" to stop?

- When all nodes are pure (homogeneous leaves)
- These trees can be very deep : Overfitting
- Good trees don't over-fit !

Early Stopping

- Information Gain < Threshold
- Minimum Instances per Node
- Maximum Tree Depth

Building a good tree?

- Reduction / Gain in optimization criteria
- But tree building is greedy!
- Current split gain < Future split gain (Gotcha !)

Alternate

- Grow & Prune…

Split & Merge : Grow & Prune

Key !dea

- Grow deep trees first (Greedy split workaround)
- Prune low gain branches.

Cost Complexity Tradeoff

- Cost of pruning : Increase in Impurity
- Reduction in Complexity : Shorter trees, Fewer leaves

What is a good tree?

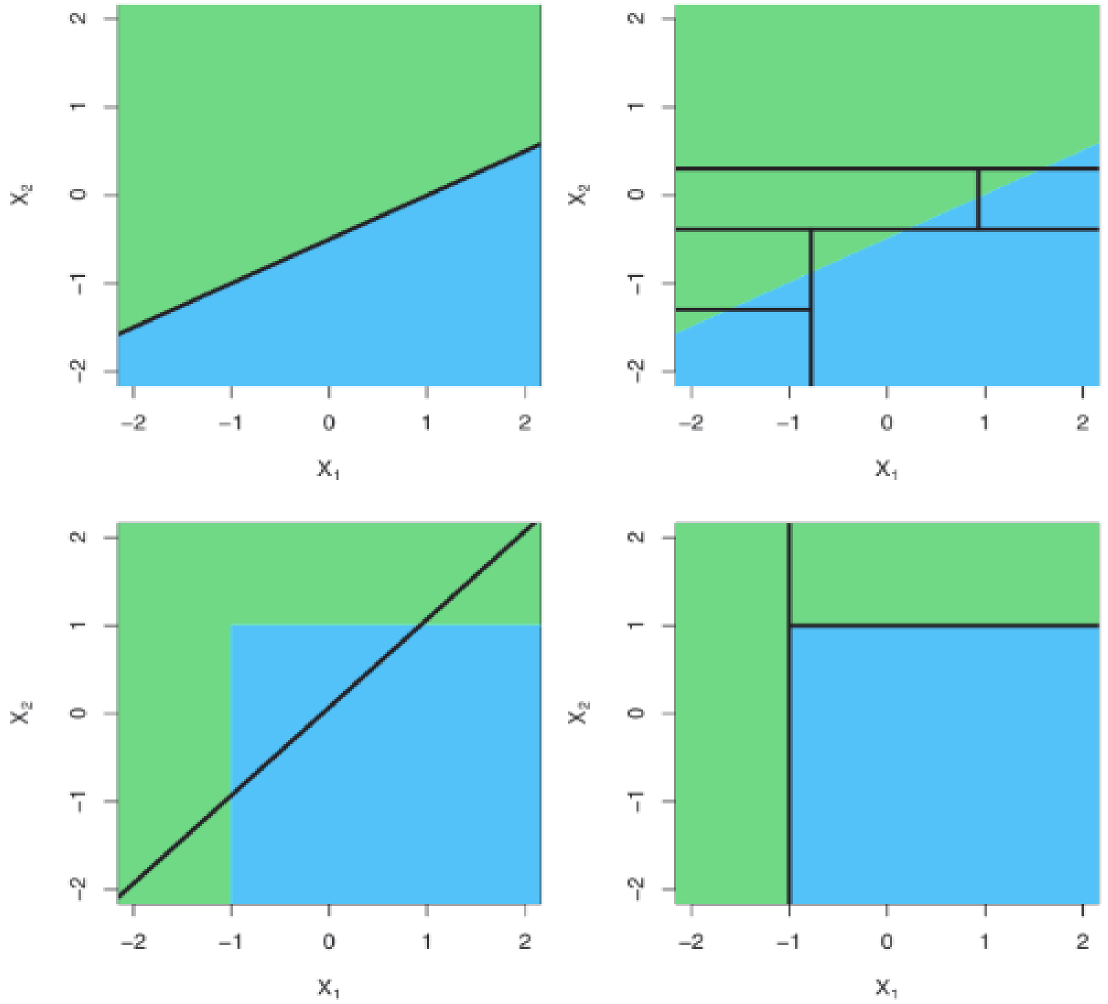
- When to stop pruning?
- Overfitting measure: number of leaves, depth of tree

Optimal Tradeoff

- Parameter trading off cost complexity
- Try different values: choose one based on performance on test data

Decision Trees vs. Linear Regression (Separating Hyperplane)

- Linear Regression
    - Linear: y is a linear combination of its features
    - The separating boundary is a hyperplane
- Decision Tree
    - The separating boundary is piecewise linear along one of the features
    - Keep splitting the feature spaces till variance in the dependent variable is low enough
- Y = f(X)



Decision Trees : Summary

Splits = Branching

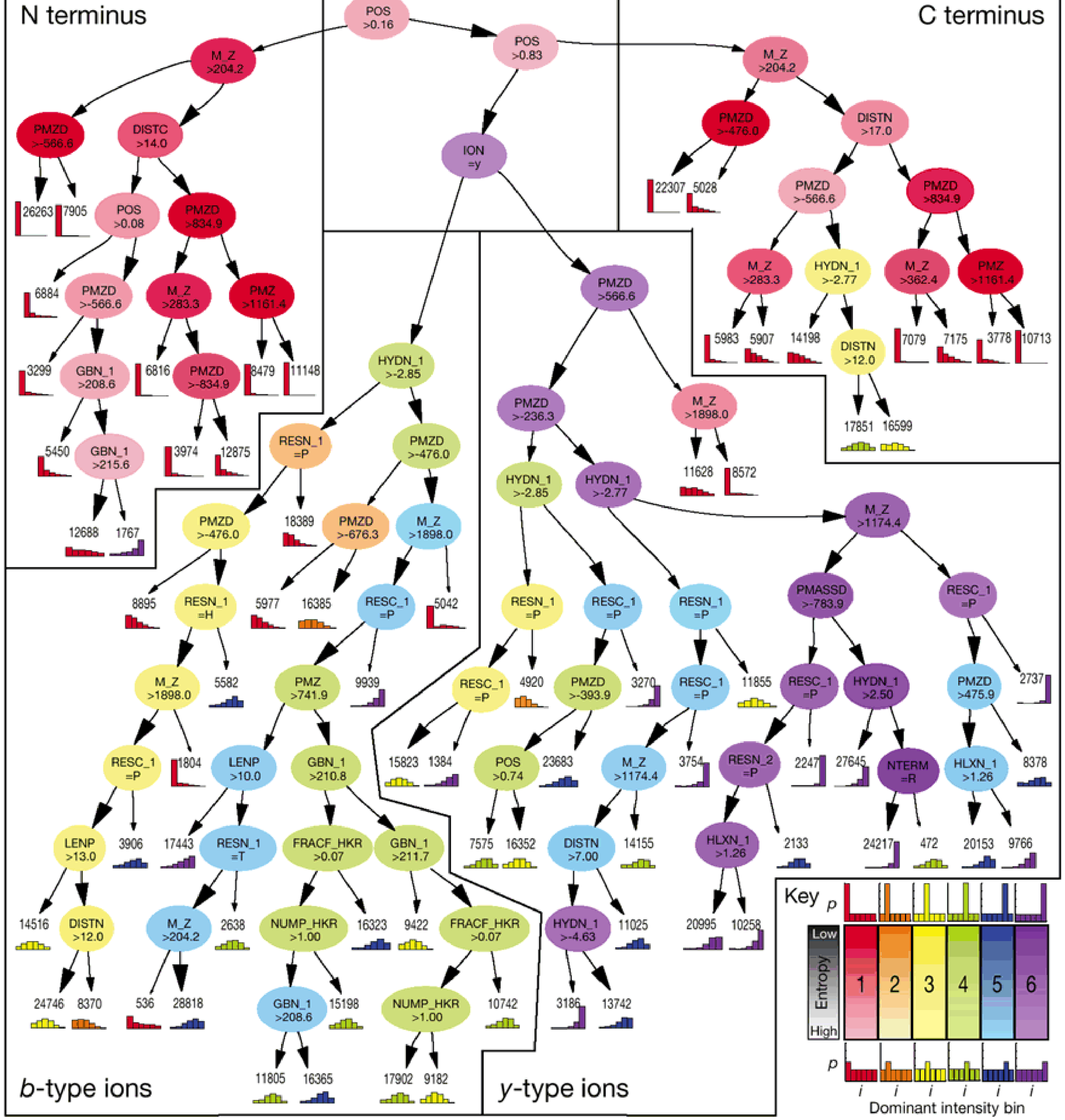
- Split = Feature, Split point

Information gain (Entropy) = Colour

- In the dominant intensity bin

Leaf Distribution = Data Homogeneity

- Some leaves are better than others



Q?

Praphul Chandra

Insofe

Bangalore, Hyderabad