Decision Tree

* Powerful mechanism to analysis large amounts of data
  + To describle the main characteristic of data
  + To classify data into disjoints classes
  + To generalize for predicting new data

# Entropy

* Show uncertainty of the dataset.
* Average Entropy
* Entropy of the tree

# Overfitting

* At the beginning of the tree, enough data to find good “general question”. Go down -> less data -> too specific question
* Strategies: prepruning (stop when growing) and post-pruning (remove bad node after grow full tree)
* Criteria: use some validation data, statistical significance test or set bound ( number of nodes, size of node, inf gain)
* Pre-pruning:
  + Stop when H not improve or require H a minimum improvement
  + Use minimum Description Length MLD: trade-off between accuracy and complexity
* Post-pruning:
  + Using validation data. For all nodes N in tree, replace with leaf (T\_n), compute classification accuracy of T\_n on validation data

# Binary Tree

* Question more attributes more effective. But data is split rapidly
* Maximum gain
* Find best binary question answer. So we have possible subset S, questions
  + Optimal subset