Attributes with many values

Problem:

- If attribute has many values, $Gain$ will select it
- Imagine using $Date = Jun\_3\_1996$ as attribute

- So many values that it
  - Divides examples into tiny sets
  - Each set likely uniform → high info gain
  - But poor predictor…
- Need to penalize these attributes

One approach: Gain ratio

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

where S_i is subset of S for which A has value v_i

SplitInfo ≅ entropy of S *wrt* values of A

(Contrast with entropy of S *wrt* **target** value)

⇓ attribs with many uniformly distrib values

e.g. if A splits S uniformly into n sets

SplitInformation = $\log_2(n)$... = 1 for Boolean