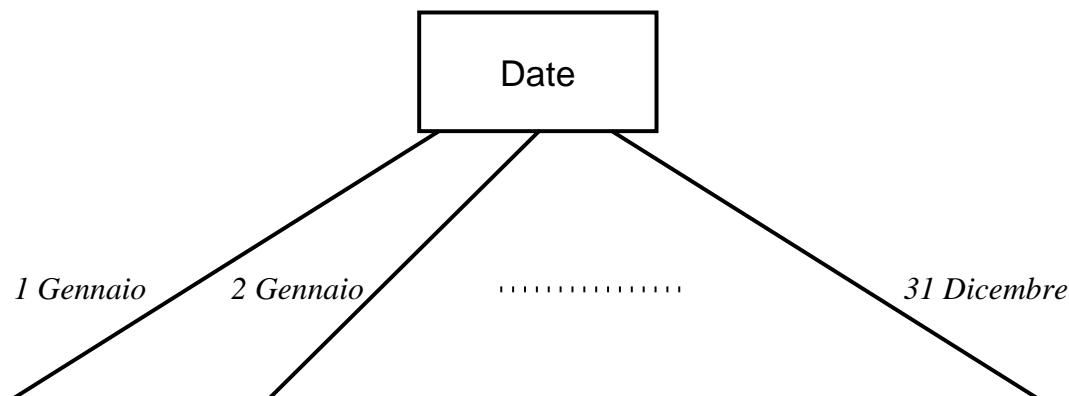


## Selezione Attributo Ottimo: Problema

**Problema:** Il *Guadagno* favorisce troppo attributi che possono assumere tanti valori diversi.

Esempio: se al problema di decidere quando giocare a tennis si aggiunge un attributo che consiste nella data del giorno considerato (es. *Data*="11 Novembre"), allora l'attributo *Date* è quello che avrà guadagno massimo (ogni giorno costituirà un sottoinsieme diverso e puro, quindi con entropia a 0), anche se in realtà non è significativo.



## Selezione Attributo Ottimo: Gain Ratio

Per rimediare a questo problema si definisce il *Gain Ratio*:

$$\text{GainRatio}(S, A) = \frac{\text{Guadagno}(S, A)}{\text{SplitInformation}(S, A)}$$

dove

$$\text{SplitInformation}(S, A) = - \sum_{v \in \text{Valori}(A)} \frac{|S_{A=v}|}{|S|} \log_2 \left( \frac{|S_{A=v}|}{|S|} \right)$$

La *SplitInformation* misura quanti, e quanto uniformi sono, i sottoinsiemi generati dall'attributo  $A$  a partire dall'insieme  $S$ .

*SplitInformation* corrisponde alla entropia di  $S$  dati i valori di  $A$ .

Si osservi che il termine *SplitInformation* sfavorisce attributi che suddividono  $S$  in molti sottoinsiemi tutti della stessa cardinalità.

## Selezione Attributo Ottimo: Gain Ratio

*GainRatio* non risolve tutti i problemi. Infatti potrebbe succedere che attributi significativi, ma che assumono qualche valore in più rispetto agli altri, potrebbero essere sfavoriti.

Di solito la strategia utilizzata è la seguente:

1. si calcola il *Guadagno* per ogni attributo;
2. si calcola la media dei guadagni calcolati;
3. si selezionano SOLO gli attributi che hanno *Guadagno* al di sopra della media;
4. si sceglie, fra gli attributi selezionati, quello che ha *GainRatio* maggiore.