Weka

- Weka e' una suite di strumenti di data mining
- Comprende strumenti per :
  - Apprendimento di alberi di decisione e di regole di produzione
  - Regressione
  - Scoperta di regole associative
  - Clustering
- Liberamente scaricabile (insieme al codice sorgente Java) da: http://www.cs.waikato.ac.nz/ml/weka/index.html
- Collegato al libro: Witten e Frank

Weka

- Contiene l'implementazione in Java dei principali algoritmi di Data Mining:
  - Clustering: k-means, EM, Cobweb
  - Regole associative: apriori
- Utilizza un formato chiamato ARFF (Attribute Relation File Format) per i file di ingresso

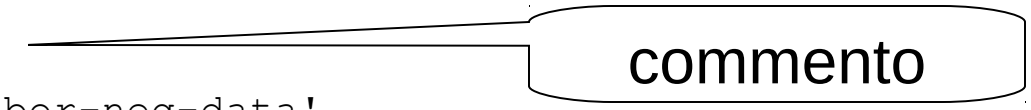
ARFF

- Un unico file
- Due sezioni:
  - Intestazione
  - Dati
- Intestazione: definizione degli attributi
  - @relation <nome del dataset>
  - @attribute <nome attr> {<val1>, <val2>, …, <valn>}
    - oppure
  - @attribute <nome attr> real
- L'ultimo attributo indica la classe

File labor.arff: intestazione

```
% labor.arff
@relation 'labor-neg-data'

@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
```



Sezione dati

- E’ costituita dal tag @data seguito dalle descrizioni degli esempi, una su ogni riga (terminata da a capo)
- Ogni esempio e’ descritto dalla lista dei valori per ciascun attributo separati da virgole
- Ogni valore corrisponde all’attributo che si trova nella stessa posizione nell’intestazione
- Le dichiarazioni @relation, @attribute e @data sono case insensitive.

Labor.arff: dati

```
@data
1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'
3,3.7,4,5,'tc',?,?,?,?,'yes',?,?,?,?,'yes',?,'good'
2,2,2.5,?,?,35,?,?,6,'yes',12,'average',?,?,?,?,'good'
3,6.9,4.8,2.3,?,40,?,?,3,?,12,'below_average',?,?,?,?,'good'
...
```

Gli apici sono necessari solo se le stringhe contengono spazi

ARFF

- Altri formati per gli attributi:
  - integer, string, date
- integer: l'attributo puo' assumere solo valori interi
- string: attributi stringa consentono di avere attributi contenenti arbitrari valori testuali.
  - Questo e' utile in applicazioni di text mining perche' possono essere creati dataset con attributi stringa che possono poi essere manipolati attraverso filtri di Weka (anche scritti dall'utente)
  - In weka, gli attributi stringa sono trattati come attributi nominali con molti valori possibili

ARFF

- date: attributi data. Hanno la forma:

  @attribute <name> date [<date-format>]

- dove <date-format> e' una stringa opzionale che specifica come i valori data debbano essere interpretati e stampati. Il formato di default e' il formato ISO-8601 "yyyy-MM-dd'T'HH:mm:ss"
- Esempio

```
@RELATION Timestamps
@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"
@DATA
"2001-04-03 12:12:12"
"2001-05-03 12:59:55"
```