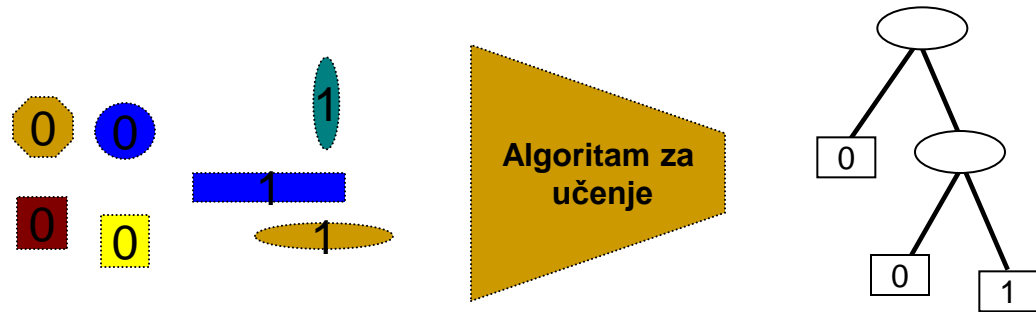


# Strojno učenje



□ WEKA

# WEKA

## Zbirka algoritama za strojno učenje (JAVA, GPL)

- ❑ Pred-procesiranje podataka (re-sampling, filtriranje: atributi, primjeri)
- ❑ Nadzirano, nenadzirano učenje
- ❑ Klasifikacija, regresija, “clustering”, asocijativna pravila,.....
- ❑ Algoritmi: Decision trees, Rule learning, NaiveBayes, NN, Bayesian Nets, SVM, Random Forest....
- ❑ Meta-learning
  - ❑ različite sheme (boosting, bagging, stacking, cost sensitive) i tehnike za kombiniranje više modela ili algoritama za učenje
- ❑ Experimenter

# WEKA

Resursi za učenje ([www.cs.waikato.ac.nz](http://www.cs.waikato.ac.nz))

- **WEKA manual**
  - stable and developer version

## Knjiga:

- “*Data Mining – Practical Machine Learning Tools and Techniques*“, Morgan Kaufmann, 2nd edition; by Ian Witten & Eibe Frank

WEB – site: [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

## Video-predavanja

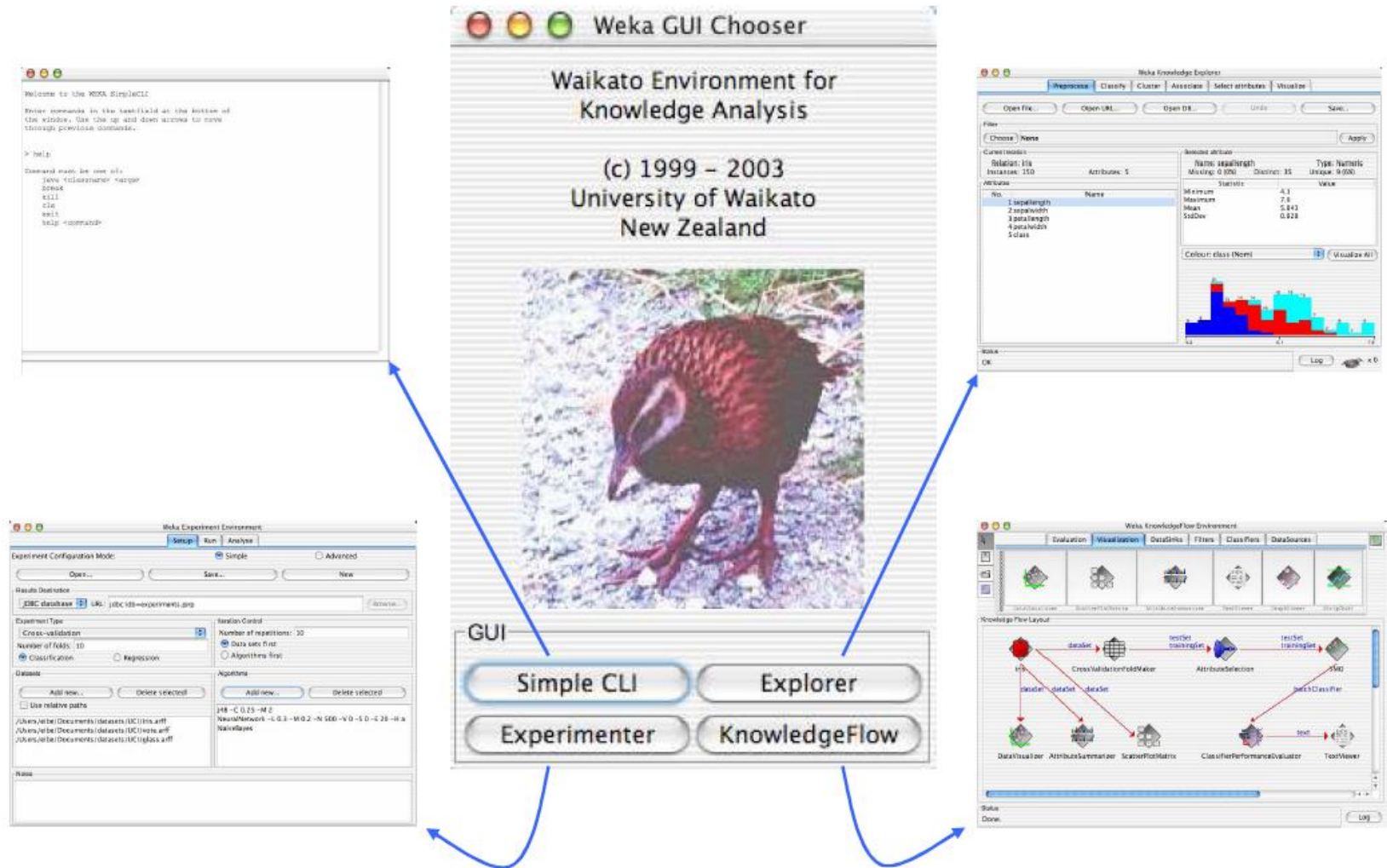
- Ian Witten: MOOC – Data Mining with Weka - YouTube

# WEKA -download

[www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)

- ❑ Requirements (Java 5.0+)
- ❑ Download (stabilna verzija 3.6.x + manual !!)
- ❑ Documentation, FAQ
- ❑ Tutorials (npr. <http://maya.cs.depaul.edu/~classes/ect584/WEKA/index.html>)
- ❑ Datasets (možete downloadati **UCI Irvine ML datasets** – samo neke – ali već u arff formatu !)
- ❑ .....
- ❑ Related Projects .....

# WEKA -download



# WEKA – format ulaznih podataka

- arff; csv, **c4.5, binary,... + 10tak drugih formata**

- **Moj\_problem.arff – format podataka**

Komentari – bilo gdje: % - ispred teksta, komentari nisu ograničeni količinom

% **1. Title: moj\_problem**

% **Author: Tom Smuc**

% **2. Sources:**

**@relation** moj\_problem                      <= ime datoteke ili problema

<= prazna linija

**@attribute** x1 numeric                      <= prvi atribut - numerički

**@attribute** x2 {plavo, bijelo, "crveno"}                      <= 2. atribut – kategorički

.....

**@attribute** xn numeric                      <= n-ti atribut – numerički

**@attribute** class {on,off}                      <= n+1-vi atribut – default – zadnji atribut = ciljni atribut

<= prazna linija

**@data**                      <= odavde pa do kraja datoteke su podaci – CSV format !

5.1, plavo,.... ,0.2,on

4.3, bijelo,.... ,2.2,off

4.3, ?,.... ,2.2,off

<= **?** Označava “nedostajuće” podatke (en. Missing data)

# WEKA – format ulaznih podataka

- iris.arff – problem dataset

**@RELATION iris**

<b>@ATTRIBUTE sepallength</b>	<b>numeric</b>
<b>@ATTRIBUTE sepalwidth</b>	<b>numeric</b>
<b>@ATTRIBUTE petallength</b>	<b>numeric</b>
<b>@ATTRIBUTE petalwidth</b>	<b>numeric</b>
<b>@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}</b>	

**@DATA**

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
.....
.....
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
```

# WEKA – format ulaznih podataka

## ■ iris.csv

Sepallength,sepalwidth,petallength,petalwidth,class

5.1,3.5,1.4,0.2,Iris-setosa

4.9,3.0,1.4,0.2,Iris-setosa

4.7,3.2,1.3,0.2,Iris-setosa

.....

.....

6.3,2.5,5.0,1.9,Iris-virginica

6.5,3.0,5.2,2.0,Iris-virginica

6.2,3.4,5.4,2.3,Iris-virginica

5.9,3.0,5.1,1.8,Iris-virginica



# WEKA – sa komandne linije

## WINDOWS

- ❑ CLASSPATH – environment varijabla
- ❑ set CLASSPATH=c:\Program Files\Weka-3-5\weka.jar

**U nekom vašem direktoriju .....**

**%prompt> java weka.classifiers.trees.J48**

**- Output : help on J48 (stabla odlučivanja – C4.5)**

**%prompt> java weka.classifiers.lazy.IB1**

**- Output : help on IB1 (1-nn !)**

**....**

**%prompt> java weka.classifiers.trees.J48 – t data\labor.arff**

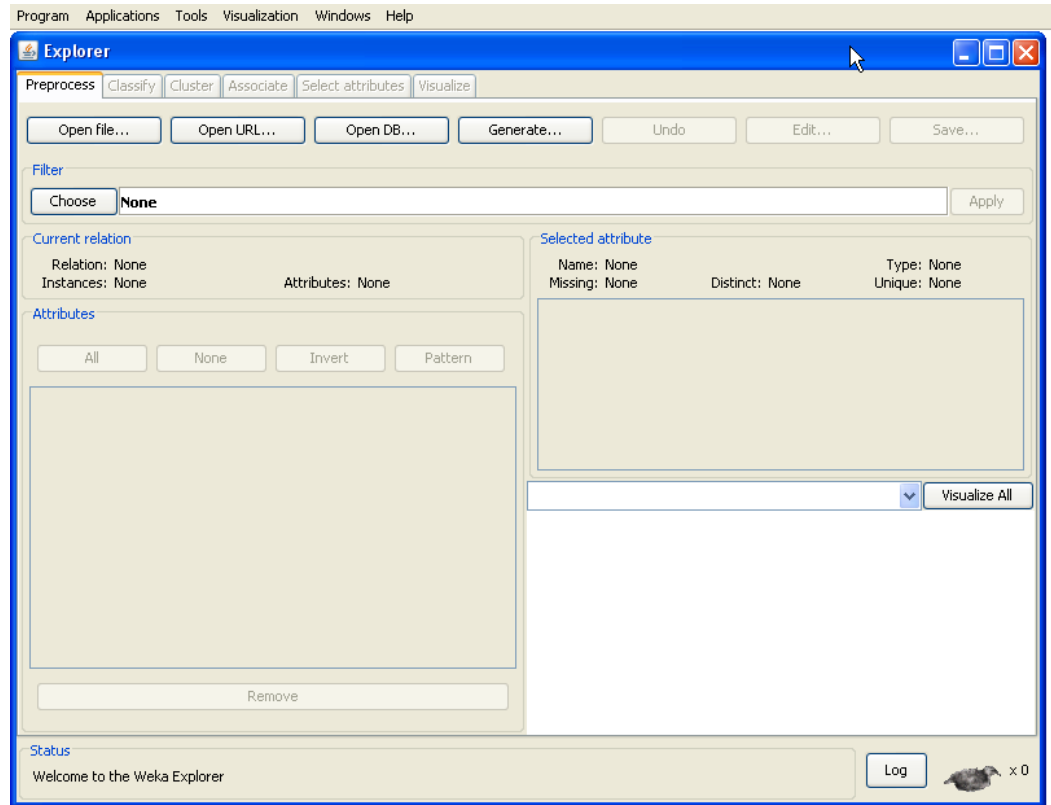
**Output: summary rezultata na “training” skupu**

**-x 10 (opcija koja pokreće 10-fold cross validation)**

# WEKA – GUI

## Applications

- **Explorer** →
- Experimenter
- KnowledgeFlow
- SimpleCLI



# WEKA – Explorer - preprocessing

## Filteri:

- ☐ Supervised

- ☐ Unsupervised

  - Za primjere (en. instances)

  - za attribute (varijable)

# WEKA – Explorer – Classification

## Decision Trees - output

=== Run information ===

```
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    labor-neg-data
Instances:   57
Attributes:  17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class
Test mode:   10-fold cross-validation
```

# WEKA – Explorer – Classification

## Decision Trees - output

=== Classifier model (full training set) ===

J48 pruned tree

-----

wage-increase-first-year <= 2.5: bad (15.27/2.27)

wage-increase-first-year > 2.5

| statutory-holidays <= 10: bad (10.77/4.77)

| statutory-holidays > 10: good (30.96/1.0)

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0 seconds

# WEKA – Explorer – Classification

## Decision Trees - output

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	42	73.6842 %
Incorrectly Classified Instances	15	26.3158 %
Kappa statistic	0.4415	
Mean absolute error	0.3192	
Root mean squared error	0.4669	
Relative absolute error	69.7715 %	
Root relative squared error	97.7888 %	
Total Number of Instances	57	

# WEKA – Explorer – Classification

## Decision Trees - output

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.7	0.243	0.609	0.7	0.651	0.695	bad
0.757	0.3	0.824	0.757	0.789	0.695	good

=== Confusion Matrix ===

```
a  b  <-- classified as
14  6 |  a = bad
 9 28 |  b = good
```

## Pred-procesiranje (podataka)

- ❑ Naći podatke (... mogu i vlastiti podaci)
  - Iris
  - zoo
  - Waveform
  
- ❑ Upoznavanje WEKA-e:
  - Filteri
    - ❑ sampliranje, spremanje....
  - Vizualizacija podataka:
    - ❑ summary stats, scatter plots...



## Klasifikacija

- ❑ 3 klasifikatora (možda i koji više...)
  - Decision trees (J48)
  - Naive Bayes
  - k-nn
- ❑ Određivanje točnosti klasifikatora
  - Training error
  - Training/test
  - Cross validation
  - ROC krivulje
- ❑ Vizualizacija modela

# WEKA eksperimenti = 5 dodatnih bodova !

## □ UCI datasets

- dostupni sa [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka)
- **Odabrati 4 skupa podataka** – po vlastitom nahođenju
- Napraviti krivulje učenja (ovisnost greške (XV) o veličini skupa za učenje za **4 algoritma: Naive Bayes(NB), Decision Tree (J48), k-nn (IBk), (neuralna mreža) MultiLayer Perceptron**
- veličina skupa za učenje za određivanje **krivulje učenja** – (10%, 20%, 40%, 80%, 100%)
- **Usporediti algoritme na svakom od ta 4 skupa**
- **Izvještaj u formi kratkog članka**
- **+ bodovi:** statistički ocijeniti značajnost razlika između algoritama
  - { 2 načina:
    - a. Koristiti WEKA-u +
      - članak T.G. Dietterich: Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms (google – CiteSeer – download)
    - b.
      - Naučiti koristiti Experimenter u WEKA-i (lakši put :-)