

# **An Introduction to the WEKA Data Mining System**

Zdravko Markov  
Central Connecticut State University  
[markovz@ccsu.edu](mailto:markovz@ccsu.edu)

Ingrid Russell  
University of Hartford  
[irussell@hartford.edu](mailto:irussell@hartford.edu)

# Agenda

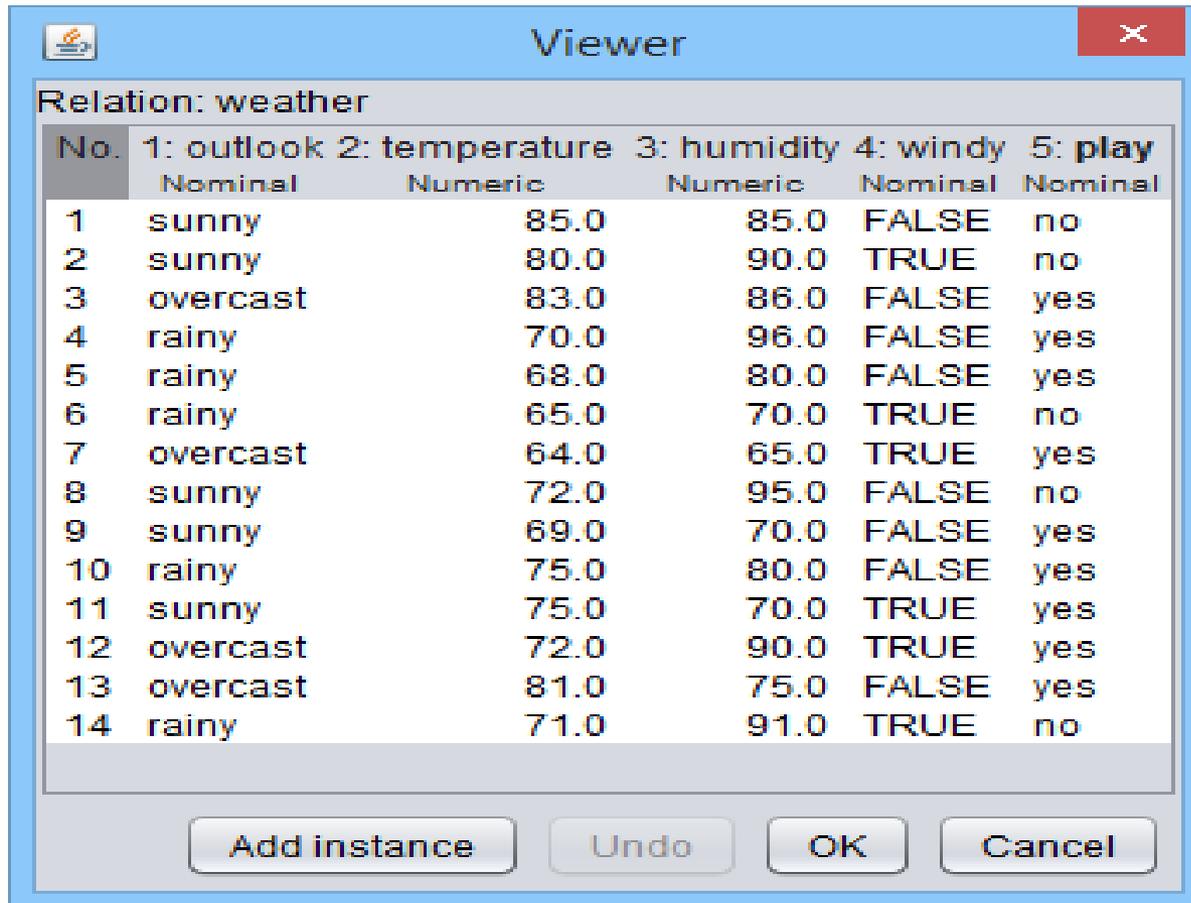
- Data Mining
- Weka Project
- Basic functionality of Weka by example
- Weka for document classification and clustering

# Database management systems (DBMS), Online Analytical Processing (OLAP) and Data Mining

Area	DBMS	OLAP	Data Mining
<b>Task</b>	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
<b>Type of result</b>	Information	Analysis	Insight and Prediction
<b>Method</b>	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
<b>Example question</b>	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

# Example of DBMS, OLAP and Data Mining: Weather data

Assume we have made a record of the weather conditions during a two-week period, along with the decisions of a tennis player whether or not to play tennis on each particular day. Thus we have generated **tuples** (or examples, instances) consisting of values of four **independent variables** (outlook, temperature, humidity, windy) and one **dependent variable** (play).



Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Buttons: Add instance, Undo, OK, Cancel

# Data-Base Management System

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

- What was the temperature in the sunny days? {85, 80, 72, 69, 75}
- Which days the humidity was less than 75? {6, 7, 9, 11}
- Which days the temperature was greater than 70? {1, 2, 3, 8, 10, 11, 12, 13, 14}
- Which days the temperature was greater than 70 and the humidity was less than 75?  
The intersection of the above two: {11}

# OLAP: Multidimensional Model (Data Cube)

Dimensions:

- Time: Week 1={1, 2, 3, 4, 5, 6, 7}, Week 2={8, 9, 10, 11, 12, 13, 14}
- Outlook: {sunny, rainy, overcast}

Unit: play (yes/no)

9 / 5	sunny	rainy	overcast
Week 1	0 / 2	2 / 1	2 / 0
Week 2	2 / 1	1 / 1	2 / 0

→ if outlook = overcast then play = yes

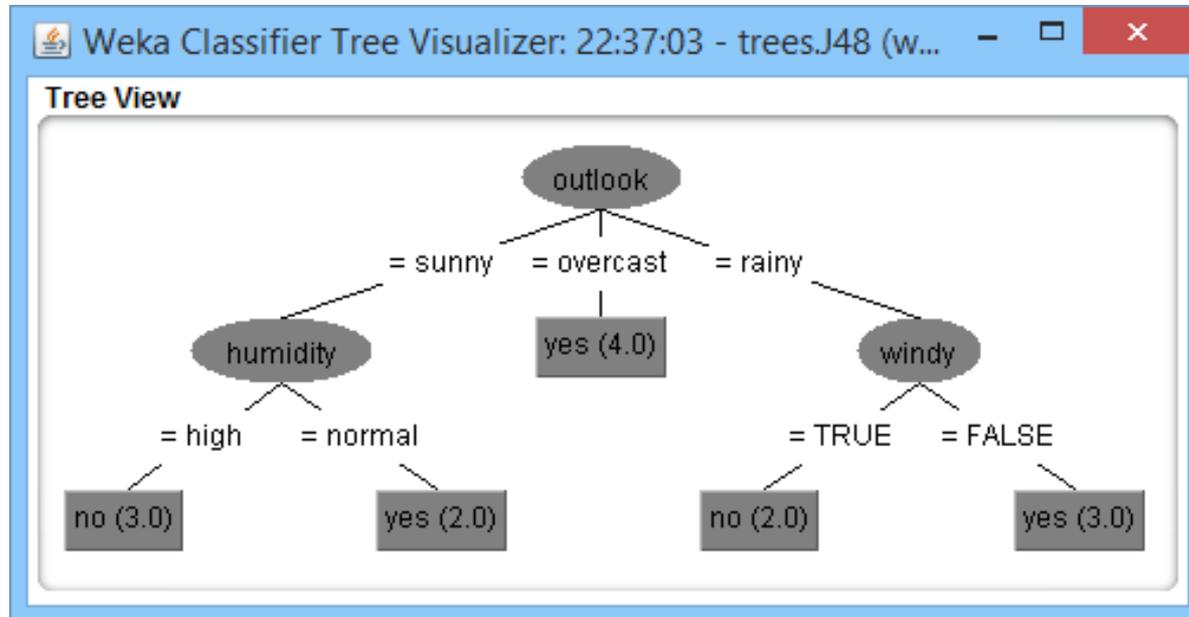
# Data Mining: Association Rules

**Discretize numeric attributes** (data pre-processing stage in data mining). Group the temperature values in three intervals (hot, mild, cool) and humidity values in two (high, normal).

1. humidity=normal windy=false 4 ==> play=yes (4, 1)
2. temperature=cool 4 ==> humidity=normal (4, 1)
3. outlook=overcast 4 ==> play=yes (4, 1)
4. temperature=cool play=yes 3 ==> humidity=normal (3, 1)
5. outlook=rainy windy=false 3 ==> play=yes (3, 1)
6. outlook=rainy play=yes 3 ==> windy=false (3, 1)
7. outlook=sunny humidity=high 3 ==> play=no (3, 1)
8. outlook=sunny play=no 3 ==> humidity=high (3, 1)
9. temperature=cool windy=false 2 ==> humidity=normal play=yes (2, 1)
10. temperature=cool humidity=normal windy=false 2 ==> play=yes (2, 1)

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

# Data Mining: Decision Tree and Rules



If outlook = overcast then yes

If humidity = normal and windy = false then yes

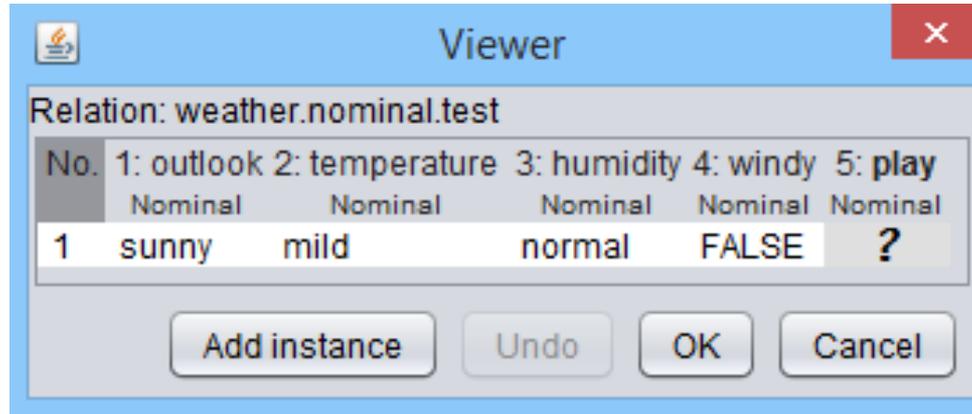
If temperature = mild and humidity = normal then yes

If outlook = rainy and windy = false then yes

If outlook = sunny and humidity = high then no

If outlook = rainy and windy = true then no

# Data Mining: Prediction



$P(\text{play}=\text{yes} \mid \text{outlook}=\text{sunny}, \text{temperature}=\text{mild}, \text{humidity}=\text{normal}, \text{windy}=\text{false}) = 0.8$

$P(\text{play}=\text{no} \mid \text{outlook}=\text{sunny}, \text{temperature}=\text{mild}, \text{humidity}=\text{normal}, \text{windy}=\text{false}) = 0.2$

# Weka Project



Machine Learning Group at the University of Waikato

**WEKA**  
The University of Waikato

Project Software Book Publications People Related

## Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

We have put together several free online courses that teach machine learning and data mining using Weka. Check out the **website for the courses** for details on when and how to enrol. The videos for the courses are available **on Youtube**.

Yes, it is possible to apply Weka to **big data**!

---

Getting started	Further information	Developers
<ul style="list-style-type: none"><li>• Requirements</li><li>• Download</li><li>• Documentation</li><li>• FAQ</li><li>• Getting Help</li></ul>	<ul style="list-style-type: none"><li>• Citing Weka</li><li>• Datasets</li><li>• Related Projects</li><li>• Miscellaneous Code</li><li>• Other Literature</li></ul>	<ul style="list-style-type: none"><li>• Development</li><li>• History</li><li>• Subversion</li><li>• Contributors</li><li>• Commercial licenses</li></ul>

75%

# 12 Years ago ...

## KDnuggets : News : 2005 : n13 : item2

**SIGKDD Service Award** is the highest service award in the field of data mining and knowledge discovery. It is given to one individual or one group who has performed significant service to the data mining and knowledge discovery field, including professional volunteer services in disseminating technical information to the field, education, and research funding.

The **2005 ACM SIGKDD Service Award** is presented to **the Weka team** for their development of the freely-available Weka Data Mining Software, including the accompanying book Data Mining: Practical Machine Learning Tools and Techniques (now in second edition) and much other documentation.

The Weka team includes **Ian H. Witten** and **Eibe Frank**, and the following major contributors (in alphabetical order of last names): Remco R. Bouckaert, John G. Cleary, Sally Jo Cunningham, Andrew Donkin, Dale Fletcher, Steve Garner, Mark A. Hall, Geoffrey Holmes, Matt Humphrey, Lyn Hunt, Stuart Inglis, Ashraf M. Kibriya, Richard Kirkby, Brent Martin, Bob McQueen, Craig G. Nevill-Manning, Bernhard Pfahringer, Peter Reutemann, Gabi Schmidberger, Lloyd A. Smith, Tony C. Smith, Kai Ming Ting, Leonard E. Trigg, Yong Wang, Malcolm Ware, and Xin Xu.

The Weka team has put a tremendous amount of effort into continuously developing and maintaining the system **since 1994**. The development of Weka was funded by a grant from the New Zealand Government's Foundation for Research, Science and Technology.

The **key features** responsible for Weka's success are:

- it provides many different algorithms for data mining and machine learning
- it is open source and freely available
- it is platform-independent
- it is easily usable by people who are not data mining specialists
- it provides flexible facilities for scripting experiments
- it has kept up-to-date, with new algorithms being added as they appear in the research literature.

# 12 Years ago ...

## **KDnuggets : News : 2005 : n13 : item2 (cont.)**

The Weka Data Mining Software has been downloaded **200,000 times** since it was put on SourceForge in April 2000, and is currently downloaded at a rate of 10,000/month. The Weka mailing list has over **1100 subscribers in 50 countries**, including subscribers from many major companies.

There are **15 well-documented substantial projects** that incorporate, wrap or extend Weka, and no doubt many more that have not been reported on Sourceforge.

Ian H. Witten and Eibe Frank also wrote a **very popular book "Data Mining: Practical Machine Learning Tools and Techniques"** (now in the second edition), that seamlessly integrates Weka system into teaching of data mining and machine learning. In addition, they provided **excellent teaching material** on the book website.

This book became one of the most popular textbooks for data mining and machine learning, and is **very frequently cited in scientific publications**.

Weka is a **landmark system in the history of the data mining and machine learning** research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of Weka was released 11 years ago). Other data mining and machine learning systems that have achieved this are individual systems, such as C4.5, not toolkits.

Since Weka is freely available for download and offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems. Weka also became one of the favorite vehicles for data mining research and helped to advance it by making many powerful features available to all.

**In sum, the Weka team has made an outstanding contribution to the data mining field.**

# Now ...

Machine Learning Group at the University of Waikato

Project Software **Book** Publications People Related

## Data Mining: Practical Machine Learning Tools and Techniques

Machine learning provides an exciting set of technologies that includes practical tools for analysing data and making predictions but also powers the latest advances in artificial intelligence. We have written a book that provides a highly accessible introduction to the area but also caters for readers who want to delve into the more mathematical techniques available in modern probabilistic modeling and deep learning approaches. **Chris Pal** has joined **Ian Witten**, **Eibe Frank**, and **Mark Hall** for the fourth edition, and his expertise in probabilistic models and deep learning has greatly extended the book's coverage. To make room for the new material, we now provide an online appendix on the Weka software. It is an extended version of a brief description of Weka included as an appendix in the book. The book continues to provide references to Weka implementations of algorithms that it describes. The **Weka MOOCs** provide activities similar to the tutorial exercises in the 3rd edition. We now also provide information on other software: the computational ecosystem for machine learning has grown enormously since we have written the third edition in 2011. A table of contents for the fourth edition, indicating where we have added new material, can be found [further down this page](#).

IAN H. WITTEN • EIBE FRANK • MARK A. HALL

**DATA MINING**  
Practical Machine Learning Tools and Techniques  
THIRD EDITION

IAN H. WITTEN & EIBE FRANK

**DATA MINING**  
Practical Machine Learning Tools and Techniques  
SECOND EDITION

FOURTH EDITION

**DATA MINING**  
Practical Machine Learning Tools and Techniques

DATA MINING  
PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES WITH JAVA IMPLEMENTATIONS

It is certainly one of my favourite data mining books in my

50%



**WEKA**  
The University  
of Waikato

Waikato Environment for Knowledge Analysis  
Version 3.8.1  
(c) 1999 - 2016  
The University of Waikato  
Hamilton, New Zealand

**Applications**

- Explorer
- Experimenter
- KnowledgeFlow
- Workbench
- Simple CLI

- Preprocess
- Classify
- Cluster
- Associate
- Select attributes
- Visualize

- Open file...
- Open URL...
- Open DB...
- Generate...
- Undo
- Edit...
- Save...

**Filter**

Choose  Apply

**Current relation**

Relation: weather.symbolic      Attributes: 5  
Instances: 14                      Sum of weights: 14

**Attributes**

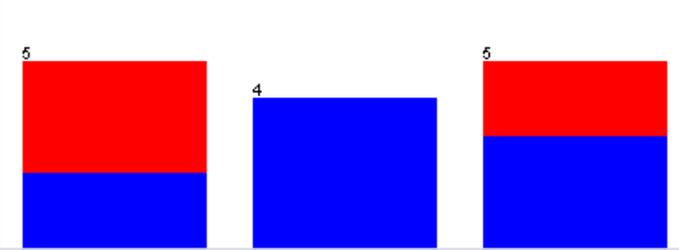
No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

**Selected attribute**

Name: outlook      Type: Nominal  
Missing: 0 (0%)      Distinct: 3      Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom) Visualize All



**Status**

OK Log x 0

*Machine Learning, Data and Web Mining*  
by Example  
(“learning by doing” approach)

- Data preprocessing and visualization
- Attribute selection
- Classification (OneR, Decision trees)
- Prediction (Nearest neighbor)
- Model evaluation
- Clustering (K-means)
- Association rules

# Data preprocessing and visualization

## Initial Data Preparation (Weka data input)

- Raw data (Japanese loan data)
- Web/Text documents (Department data)

# Data preprocessing and visualization

Japanese loan data (a sample from a loan history database of a Japanese bank)

Clients: s1,..., s20

- Approved loan: s1, s2, s4, s5, s6, s7, s8, s9, s14, s15, s17, s18, s19
- Rejected loan: s3, s10, s11, s12, s13, s16, s20

Clients data:

- unemployed clients: s3, s10, s12
- loan is to buy a personal computer: s1, s2, s3, s4, s5, s6, s7, s8, s9, s10
- loan is to buy a car: s11, s12, s13, s14, s15, s16, s17, s18, s19, s20
- male clients: s6, s7, s8, s9, s10, s16, s17, s18, s19, s20
- not married: s1, s2, s5, s6, s7, s11, s13, s14, s16, s18
- live in problematic area: s3, s5
- age: s1=18, s2=20, s3=25, s4=40, s5=50, s6=18, s7=22, s8=28, s9=40, s10=50, s11=18, s12=20, s13=25, s14=38, s15=50, s16=19, s17=21, s18=25, s19=38, s20=50
- money in a bank (x10000 yen): s1=20, s2=10, s3=5, s4=5, s5=5, s6=10, s7=10, s8=15, s9=20, s10=5, s11=50, s12=50, s13=50, s14=150, s15=50, s16=50, s17=150, s18=150, s19=100, s20=50
- monthly pay (x10000 yen): s1=2, s2=2, s3=4, s4=7, s5=4, s6=5, s7=3, s8=4, s9=2, s10=4, s11=8, s12=10, s13=5, s14=10, s15=15, s16=7, s17=3, s18=10, s19=10, s20=10
- months for the loan: s1=15, s2=20, s3=12, s4=12, s5=12, s6=8, s7=8, s8=10, s9=20, s10=12, s11=20, s12=20, s13=20, s14=20, s15=20, s16=20, s17=20, s18=20, s19=20, s20=30
- years with the last employer: s1=1, s2=2, s3=0, s4=2, s5=25, s6=1, s7=4, s8=5, s9=15, s10=0, s11=1, s12=2, s13=5, s14=15, s15=8, s16=2, s17=3, s18=2, s19=15, s20=2



# Data preprocessing and visualization

Attribute-Relation File Format (ARFF) - <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>

The image shows a screenshot of a Microsoft Internet Explorer browser window displaying the ARFF format documentation page. The browser's address bar shows the URL <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>. The page title is "Attribute-Relation File Format (ARFF)". The content includes a date "April 4th, 2006", a link to the WekaDoc Wiki, and a list of versions: 3.4.x and 3.5.x. Below this, there is a date "April 1st, 2002" and a paragraph explaining that an ARFF file is an ASCII text file describing a list of instances sharing a set of attributes. It mentions the Weka machine learning software and the Weka machine learning software project at the University of Waikato. Another paragraph explains that the explanation was cobbled together by Gordon Paynter and Eibe Frank, and has been edited by Richard Kirkby. The "Overview" section states that ARFF files have two distinct sections: the Header and the Data. The Header contains the name of the relation, a list of attributes, and their types. An example of a standard IRIS dataset is provided, showing the title, sources, and the ARFF format for the IRIS dataset. The Notepad window, titled "LoanData - Notepad", shows the ARFF format for a "LoanData" dataset. The header section includes attributes: ID (numeric), sex (f,m), married (n,y), age (numeric), money (numeric), pay (numeric), months (numeric), buy (pc,car), emp (y,n), lastemp (numeric), area (good,bad), and approved (y,n). The data section contains 20 lines of data, each representing a loan instance with values for the attributes listed in the header.

**Attribute-Relation File Format (ARFF)**

April 4th, 2006

This documentation is superseded by the [WekaDoc Wiki](#). Version specific documentation is available there:

- [3.4.x](#)
- [3.5.x](#)

April 1st, 2002

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Weka Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the [Weka machine learning software](#). This document describes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the document written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date attributes, and sparse instances).

This explanation was cobbled together by Gordon Paynter (gordon.paynter at ucr.edu) from the Weka 2.1 ARFF description, email from L. Myrealbox.com) and Eibe Frank (eibe at cs.waikato.ac.nz), and some datasets. It has been edited by Richard Kirkby (rkirkby at cs.waikato.ac.nz). If you're interested in seeing the ARFF 3 proposal.

## Overview

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. An example of a standard IRIS dataset looks like this:

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%   (a) Creator: R.A. Fisher
%   (b) Donor: Michael Marshall (MARSHALL@PLU@io.arc.nasa.gov)
%   (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

**LoanData - Notepad**

```
@relation LoanData

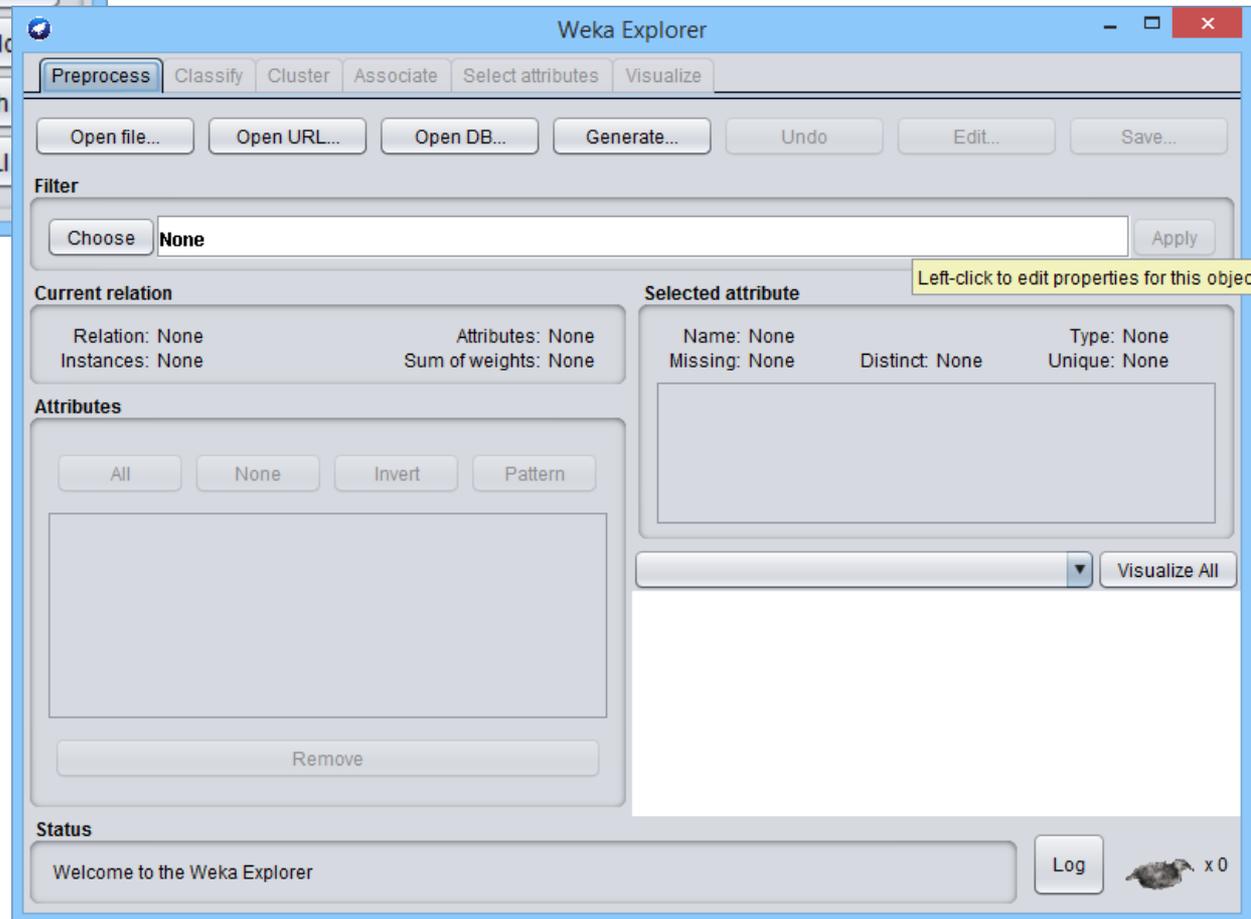
@attribute ID numeric
@attribute sex {f,m}
@attribute married {n,y}
@attribute age numeric
@attribute money numeric
@attribute pay numeric
@attribute months numeric
@attribute buy {pc,car}
@attribute emp {y,n}
@attribute lastemp numeric
@attribute area {good,bad}
@attribute approved {y,n}

@data

1, f, n, 18, 20, 2, 15, pc, y, 1, good, y
2, f, n, 20, 10, 2, 20, pc, y, 2, good, y
3, f, y, 25, 5, 4, 12, pc, n, 0, bad, n
4, f, y, 40, 5, 7, 12, pc, y, 2, good, y
5, f, n, 50, 5, 4, 12, pc, y, 25, bad, y
6, m, n, 18, 10, 5, 8, pc, y, 1, good, y
7, m, n, 22, 10, 3, 8, pc, y, 4, good, y
8, m, y, 28, 15, 4, 10, pc, y, 5, good, y
9, m, y, 40, 20, 2, 20, pc, y, 15, good, y
10, m, y, 50, 5, 4, 12, pc, n, 0, good, n
11, f, n, 18, 50, 8, 20, car, y, 1, good, n
12, f, y, 20, 50, 10, 20, car, n, 2, good, n
13, f, n, 25, 50, 5, 20, car, y, 5, good, n
14, f, n, 38, 150, 10, 20, car, y, 15, good, y
15, f, y, 50, 50, 15, 20, car, y, 8, good, y
16, m, n, 19, 50, 7, 20, car, y, 2, good, n
17, m, y, 21, 150, 3, 20, car, y, 3, good, y
18, m, n, 25, 150, 10, 20, car, y, 2, good, y
19, m, y, 38, 100, 10, 20, car, y, 15, good, y
20, m, y, 50, 50, 10, 30, car, y, 2, good, n
```

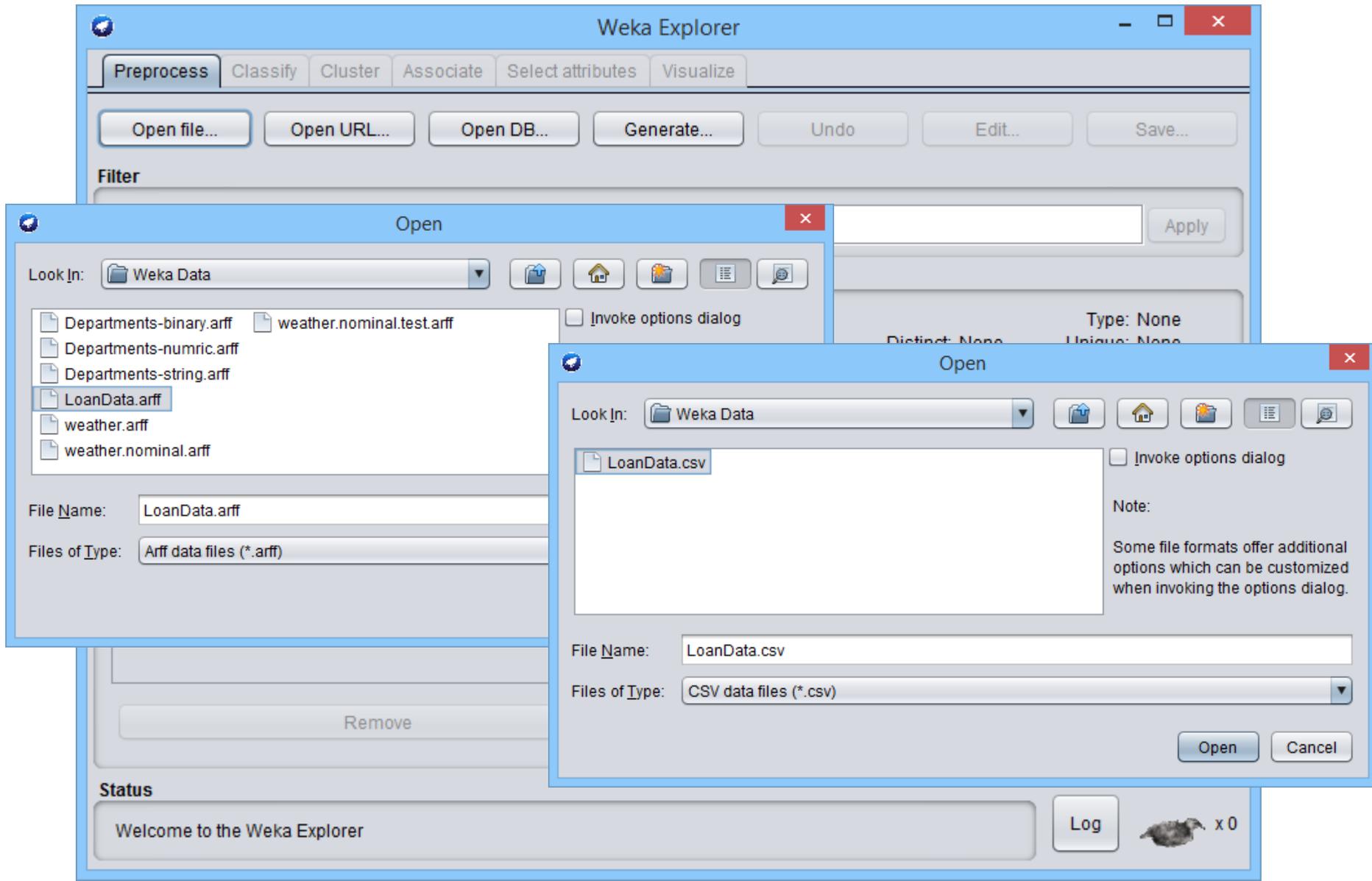
# Data preprocessing and visualization

Run Weka and select the Explorer



# Data preprocessing and visualization

Load data into Weka – ARFF format or CSV format (click on “Open file...”)



# Data preprocessing and visualization

Converting data formats through Weka (click on “Save...”)

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active. The 'Current relation' section shows 'Relation: LoanData' with 'Instances: 20' and 'Attributes: 12'. The 'Attributes' section contains a list of 12 attributes with checkboxes:

No.	Name
1	<input checked="" type="checkbox"/> ID
2	<input type="checkbox"/> sex
3	<input type="checkbox"/> married
4	<input type="checkbox"/> age
5	<input type="checkbox"/> money
6	<input type="checkbox"/> pay
7	<input type="checkbox"/> months
8	<input type="checkbox"/> buy
9	<input type="checkbox"/> emp
10	<input type="checkbox"/> lastemp
11	<input type="checkbox"/> area
12	<input type="checkbox"/> approve

The 'Selected attribute' section shows 'Name: ID' with 'Missing: 0 (0%)' and 'Distinct: 20'. Below it, a table lists statistics for the selected attribute:

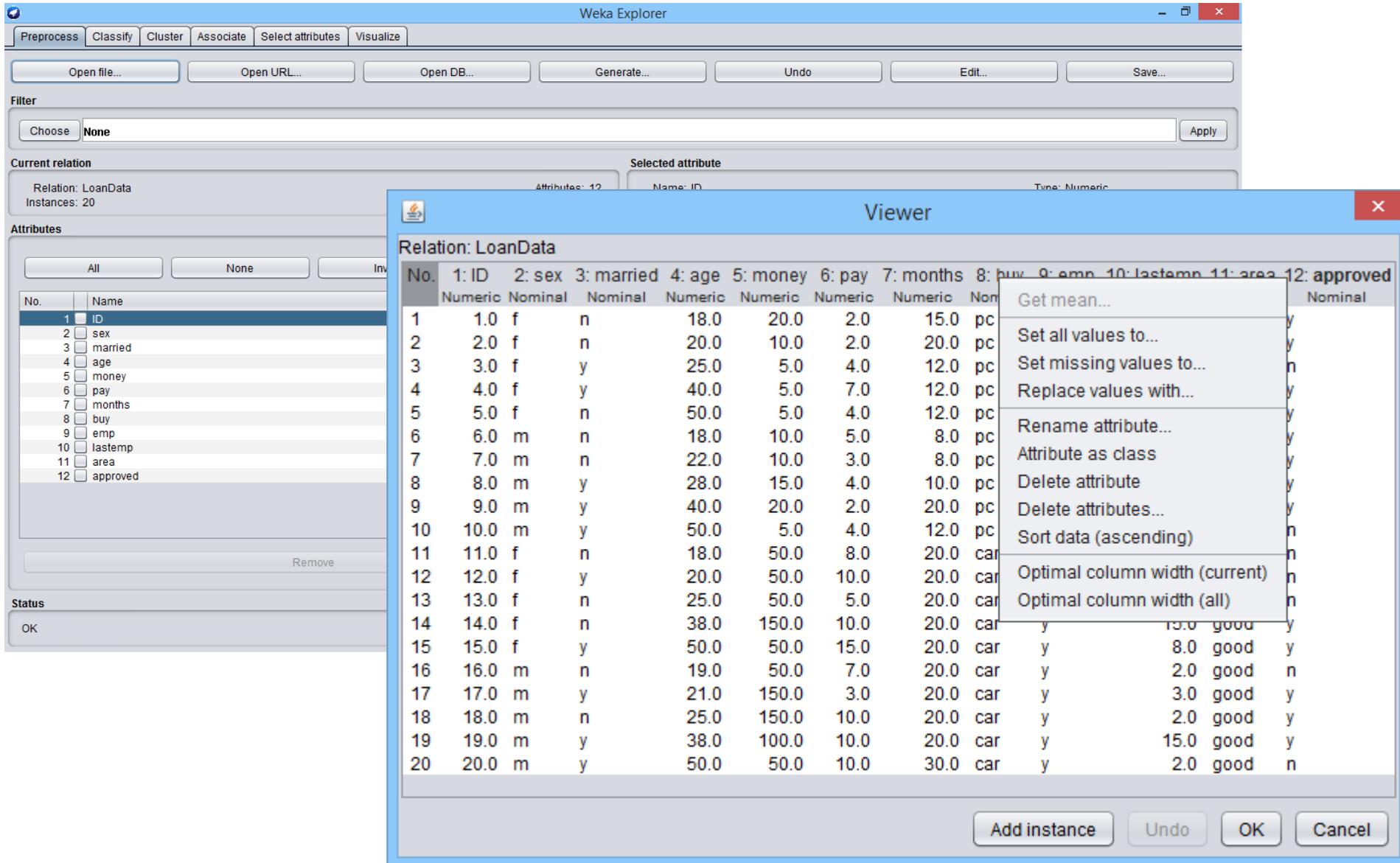
Statistic	Value
Minimum	1
Maximum	2
Mean	1
StdDev	0

This is a 'Save' dialog box from Weka Explorer. The 'Save in:' field is set to 'Weka Data'. The file list shows 'LoanData'. The 'File name:' field contains 'LoanData.csv'. The 'Files of type:' dropdown is set to 'CSV data files'. The 'Save' button is highlighted.

This is another 'Save' dialog box from Weka Explorer. The 'Save in:' field is set to 'Weka Data'. The file list shows 'LoanData'. The 'File name:' field contains 'LoanData.arff'. The 'Files of type:' dropdown is set to 'Arff data files'. The 'Save' button is highlighted.

# Data preprocessing and visualization

Editing data in Weka (click on "Edit...")



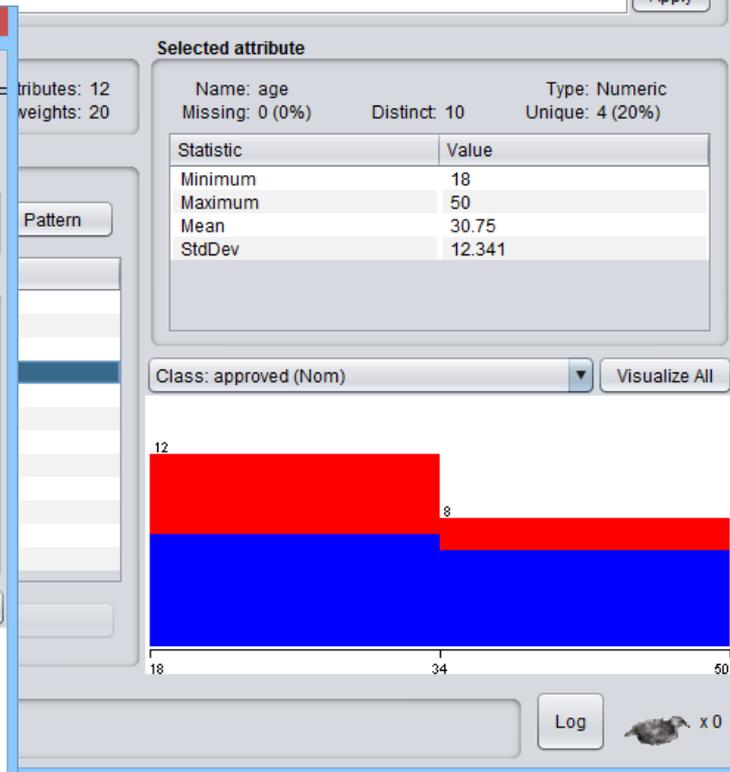
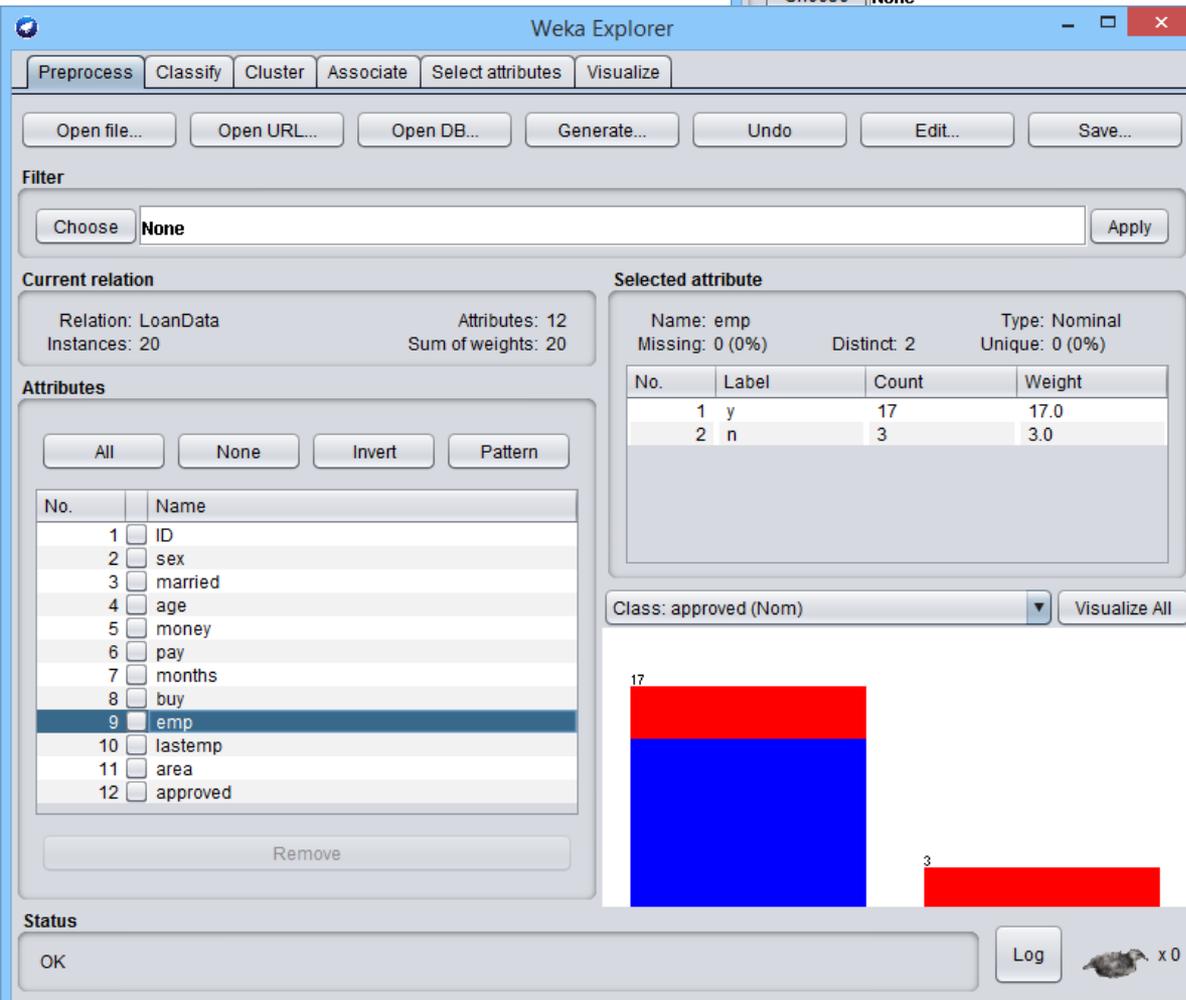
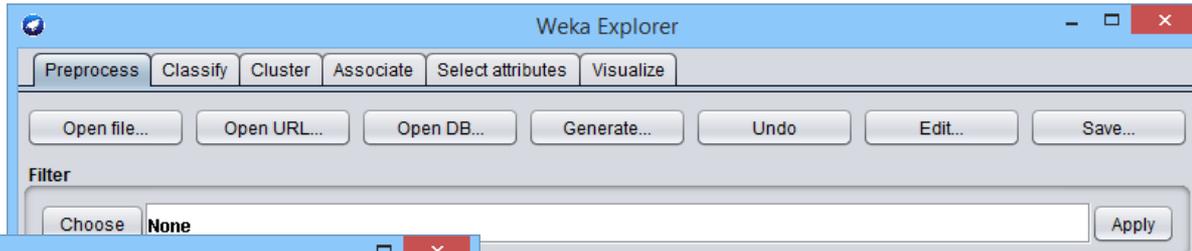
The screenshot shows the Weka Explorer interface with the 'Viewer' dialog box open. The 'Viewer' dialog displays a table of data for the 'LoanData' relation. A context menu is open over the table, listing various actions such as 'Get mean...', 'Set all values to...', 'Set missing values to...', 'Replace values with...', 'Rename attribute...', 'Attribute as class', 'Delete attribute', 'Delete attributes...', 'Sort data (ascending)', 'Optimal column width (current)', and 'Optimal column width (all)'. The table data is as follows:

No.	1: ID	2: sex	3: married	4: age	5: money	6: pay	7: months	8: buy	9: emp	10: lastemp	11: area	12: approved
	Numeric	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal
1	1.0	f	n	18.0	20.0	2.0	15.0	pc	y	15.0	good	y
2	2.0	f	n	20.0	10.0	2.0	20.0	pc	y	15.0	good	y
3	3.0	f	y	25.0	5.0	4.0	12.0	pc	n	15.0	good	n
4	4.0	f	y	40.0	5.0	7.0	12.0	pc	y	15.0	good	y
5	5.0	f	n	50.0	5.0	4.0	12.0	pc	y	15.0	good	y
6	6.0	m	n	18.0	10.0	5.0	8.0	pc	y	15.0	good	y
7	7.0	m	n	22.0	10.0	3.0	8.0	pc	y	15.0	good	y
8	8.0	m	y	28.0	15.0	4.0	10.0	pc	y	15.0	good	y
9	9.0	m	y	40.0	20.0	2.0	20.0	pc	y	15.0	good	y
10	10.0	m	y	50.0	5.0	4.0	12.0	pc	y	15.0	good	y
11	11.0	f	n	18.0	50.0	8.0	20.0	car	y	15.0	good	y
12	12.0	f	y	20.0	50.0	10.0	20.0	car	y	15.0	good	y
13	13.0	f	n	25.0	50.0	5.0	20.0	car	y	15.0	good	y
14	14.0	f	n	38.0	150.0	10.0	20.0	car	y	15.0	good	y
15	15.0	f	y	50.0	50.0	15.0	20.0	car	y	8.0	good	y
16	16.0	m	n	19.0	50.0	7.0	20.0	car	y	2.0	good	n
17	17.0	m	y	21.0	150.0	3.0	20.0	car	y	3.0	good	y
18	18.0	m	n	25.0	150.0	10.0	20.0	car	y	2.0	good	y
19	19.0	m	y	38.0	100.0	10.0	20.0	car	y	15.0	good	y
20	20.0	m	y	50.0	50.0	10.0	30.0	car	y	2.0	good	n

# Data preprocessing and visualization

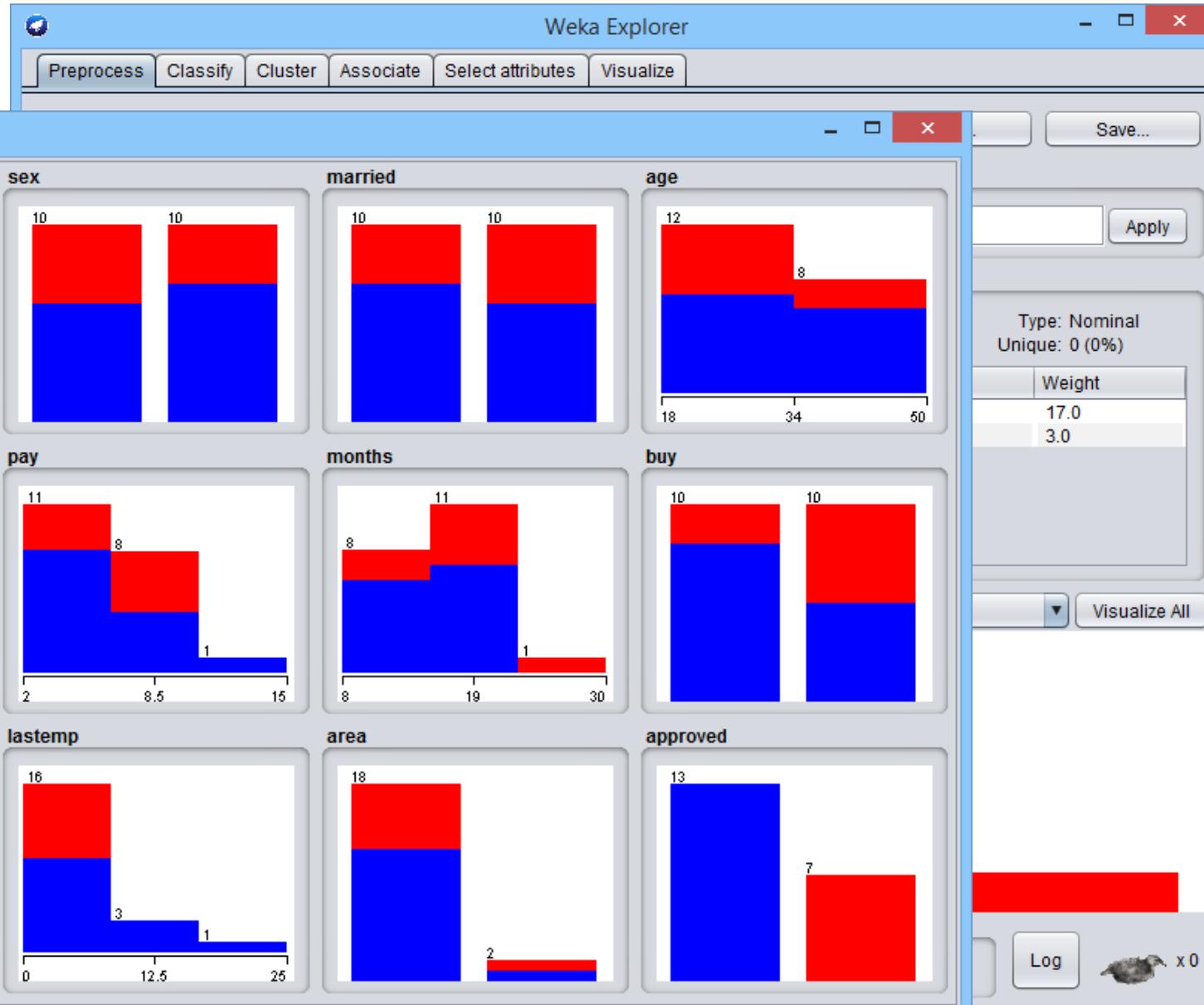
## Examining data

- Attribute type and properties
- Class (last attribute) distribution



# Data preprocessing and visualization

Click on “Visualize All”



# Data preprocessing and visualization

Click on “Visualize” tab, double-click on a plot to see the 2D projection of the instance space

The image shows the Weka Explorer interface. The main window has tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The Visualize tab is active, displaying a Plot Matrix with columns for money, pay, months, buy, emp, and lasttemp, and rows for lasttemp, emp, and buy. A detailed plot window titled "Weka Explorer: Visualizing LoanData" is open, showing a 2D scatter plot of LoanData. The X-axis is labeled "money (Num)" with values 5, 77.5, and 150. The Y-axis is labeled "emp (Nom)" with values 5 and 150. The plot shows data points colored by the "approved" attribute (Nom), with blue points representing "no" and red points representing "yes". A "Jitter" slider is visible, and a "Class colour" section is at the bottom. The status bar at the bottom shows "OK" and "Log" buttons.

# Data preprocessing and visualization

Using filters: click on “Choose” in the “Filter” window, select “Discretize”

The screenshot shows the Weka Explorer interface. The top menu bar includes Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. Below the menu bar are buttons for Open file..., Open URL..., Open DB..., Generate..., Undo, Edit..., and Save... The Filter window is open, showing a tree view of filters. The 'Discretize' filter is selected under the 'attribute' folder. The 'Selected attribute' section shows 'Name: lastemp', 'Missing: 0 (0%)', 'Distinct: 9', and 'Type: Numeric'. A table of statistics is displayed:

Statistic	Value
Minimum	0
Maximum	25
Mean	5.5
StdDev	6.732

The 'Class: approved (Nom)' dropdown is set to 'approved (Nom)'. A histogram shows the distribution of the 'lastemp' attribute, with a red bar at the top (value 16) and a blue bar at the bottom (value 3). The x-axis ranges from 0 to 25, and the y-axis ranges from 0 to 16. The status bar at the bottom shows 'OK' and a 'Log' button.

# Data preprocessing and visualization

Click in the “Discretize” in the Filter window and choose parameters, then click on “Apply”

The screenshot shows the Weka Explorer interface. The 'Filter' window is open, displaying the 'Discretize' filter configuration. The 'attributeIndices' field is set to 'first-last', 'bins' is set to 2, and 'useEqualFrequency' is set to True. The 'Visualize All' button is visible, and a bar chart below it shows the distribution of the 'lastemp' attribute after discretization. The bar chart has two bars: a taller red bar on the left and a shorter blue bar on the right. The red bar has a count of 11 and the blue bar has a count of 9.

**Selected attribute**

Name: lastemp  
Missing: 0 (0%)  
Distinct: 2  
Type: Nominal  
Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-2.5]'	11	11.0
2	'(2.5-inf)'	9	9.0

Class: approved (Nom) [Visualize All]

Bar chart showing the distribution of the 'lastemp' attribute after discretization. The x-axis represents the discretized bins, and the y-axis represents the count. The first bin (red) has a count of 11, and the second bin (blue) has a count of 9.

Note how the plot of “lastemp” changed.

# Data preprocessing and visualization

## Web/Text documents - Department data

School of Arts & Sciences Departments - Microsoft Internet Explorer

Address: <http://www.artsci.ccsu.edu/Departments.htm>

Central Connecticut State University

### Departments

[Department Chairs, Locations, Phone Numbers](#)

<u><a href="#">Anthropology</a></u>	<u><a href="#">History</a></u>
<u><a href="#">Art</a></u>	<u><a href="#">Mathematical Sciences</a></u>
<u><a href="#">Biological Sciences</a></u>	<u><a href="#">Modern Languages</a></u>
<u><a href="#">Chemistry</a></u>	<u><a href="#">Music</a></u>
<u><a href="#">Communication</a></u>	<u><a href="#">Philosophy</a></u>
<u><a href="#">Computer Science</a></u>	<u><a href="#">Physics/Earth Sciences</a></u>
<u><a href="#">Criminal Justice</a></u>	<u><a href="#">Political Science</a></u>
<u><a href="#">Design</a></u>	<u><a href="#">Psychology</a></u>
<u><a href="#">Economics</a></u>	<u><a href="#">Sociology</a></u>
<u><a href="#">English</a></u>	<u><a href="#">Theatre</a></u>
<u><a href="#">Geography</a></u>	

[ [A&S Home](#) ] [ [A-Z Directory](#) ] [ [Departments](#) ] [ [About Us](#) ]

page last updated: 10/27/04  
Comments, suggestions: [aswebmaster@ccsu.edu](mailto:aswebmaster@ccsu.edu)

Music - Microsoft Internet Explorer

Address: <http://www.artsci.ccsu.edu/Departments/Music.html>

## The School of Arts and Sciences

### Central Connecticut State University

# Music

Students majoring in music may pursue either a BS in Music education degree, the professional degree that certifies them to teach music in the public schools, or a BA in music, with specializations in either performance, music history, theory/composition, or jazz studies. Full-time and associate faculty are active in the United States and abroad performing, conducting, and presenting scholarly papers. The department's computer lab is equipped with MIDI keyboards and the industry's leading music software. The Music Department is the New England center for Orff Schulwerk training and the host for Connecticut's middle school/high school music festival and the Summer Music Institute, a national in-service program for music educators.

PROGRAMS OF STUDY: BS, BA, MS

DEPARTMENT CHAIR  
[Daniel D'Addio](#)

**Location:** Welte Hall 101  
**Phone:** 832-2900

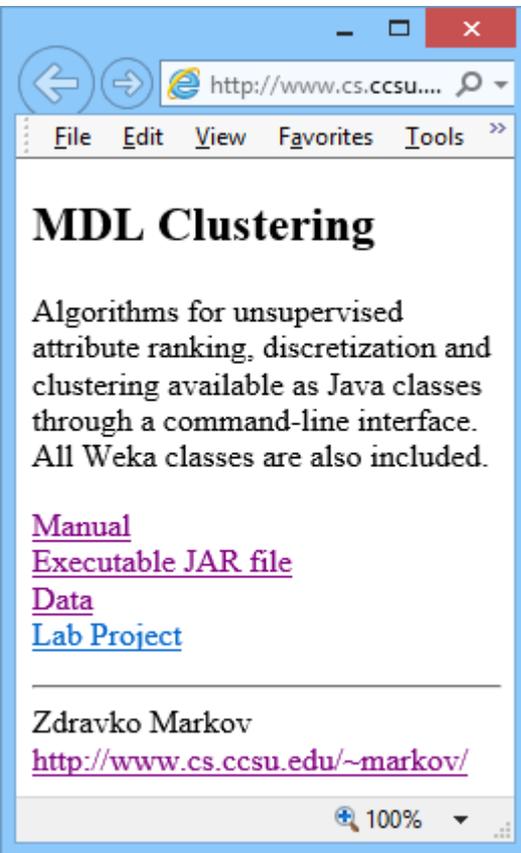
[Department Website](#)

# Data preprocessing and visualization

## Department data document collection

<http://www.cs.ccsu.edu/~markov/MDLclustering>

<http://www.cs.ccsu.edu/~markov/MDLclustering/data.zip>

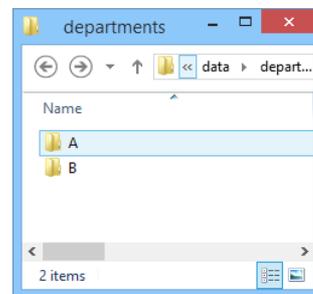


**MDL Clustering**

Algorithms for unsupervised attribute ranking, discretization and clustering available as Java classes through a command-line interface. All Weka classes are also included.

[Manual](#)  
[Executable JAR file](#)  
[Data](#)  
[Lab Project](#)

Zdravko Markov  
<http://www.cs.ccsu.edu/~markov/>

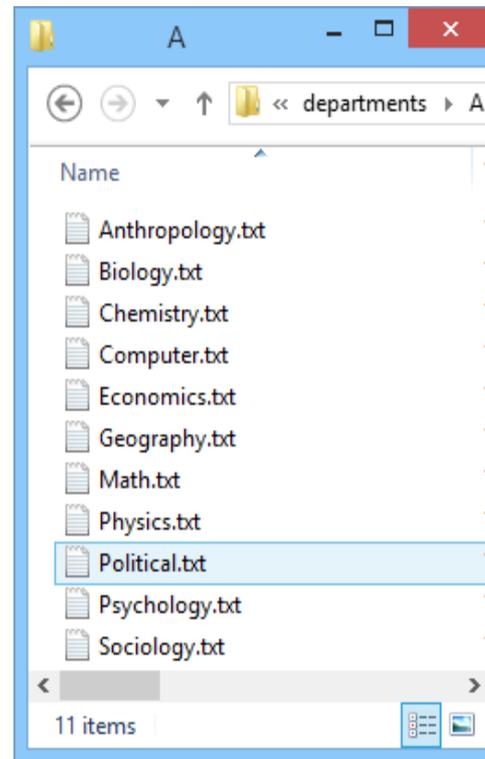


departments

Name

- A
- B

2 items

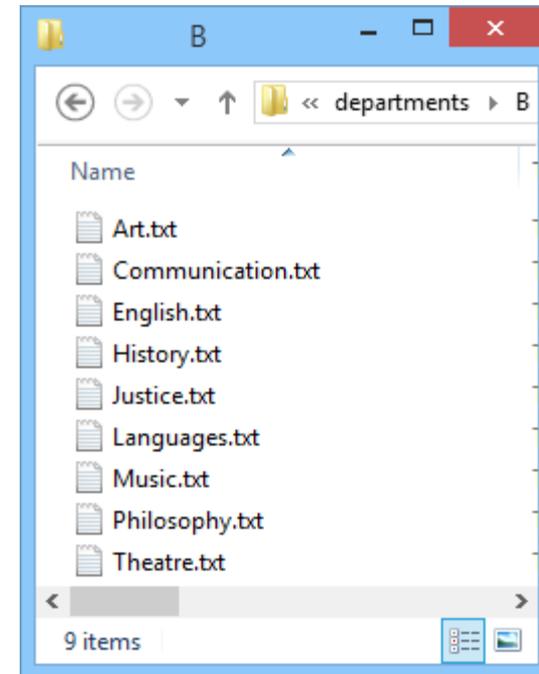


A

Name

- Anthropology.txt
- Biology.txt
- Chemistry.txt
- Computer.txt
- Economics.txt
- Geography.txt
- Math.txt
- Physics.txt
- Political.txt
- Psychology.txt
- Sociology.txt

11 items



B

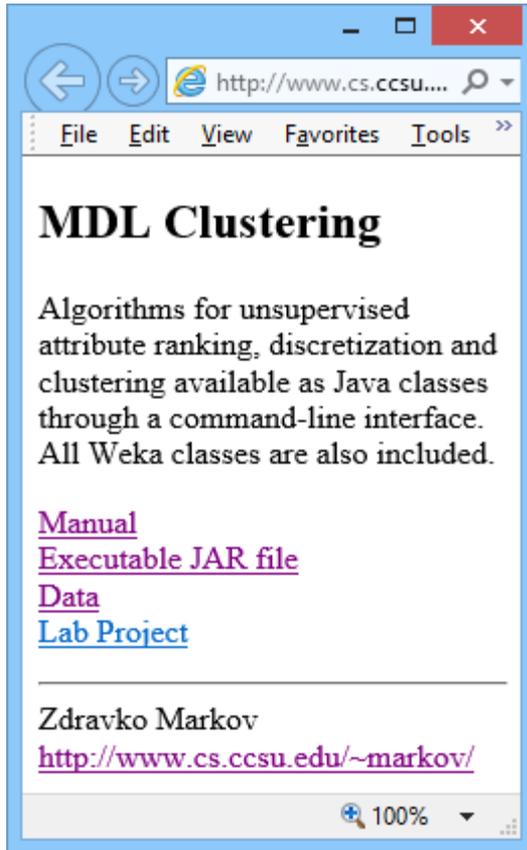
Name

- Art.txt
- Communication.txt
- English.txt
- History.txt
- Justice.txt
- Languages.txt
- Music.txt
- Philosophy.txt
- Theatre.txt

9 items

# Data preprocessing and visualization

## Department data: Create ARFF file



A screenshot of a web browser window. The address bar shows the URL <http://www.cs.ccsu.edu/~markov/MDLclustering/MDL.jar>. The page title is "MDL Clustering". The main content includes a description of algorithms for unsupervised attribute ranking, discretization, and clustering, and a list of links: Manual, Executable JAR file, Data, and Lab Project. At the bottom, the author's name "Zdravko Markov" and his website URL <http://www.cs.ccsu.edu/~markov/> are displayed.

**MDL Clustering**

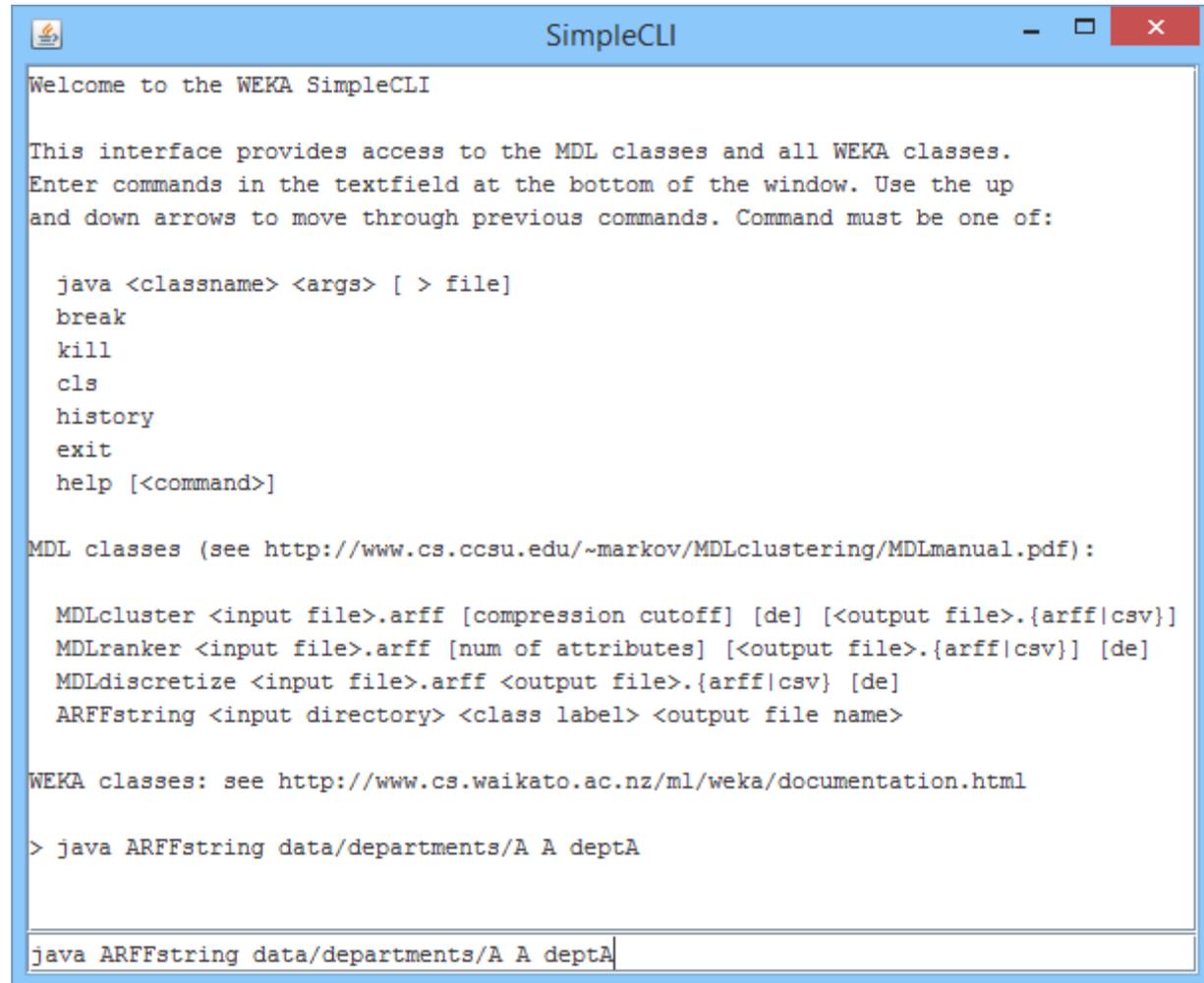
Algorithms for unsupervised attribute ranking, discretization and clustering available as Java classes through a command-line interface. All Weka classes are also included.

[Manual](#)  
[Executable JAR file](#)  
[Data](#)  
[Lab Project](#)

---

Zdravko Markov  
<http://www.cs.ccsu.edu/~markov/>

<http://www.cs.ccsu.edu/~markov/MDLclustering/MDL.jar>



A screenshot of the WEKA SimpleCLI window. The window title is "SimpleCLI". The text inside the window provides instructions on how to use the interface and lists available commands. The commands listed are: java, break, kill, cls, history, exit, and help. Below the commands, it lists MDL classes and WEKA classes. The MDL classes listed are MDLcluster, MDLranker, MDLdiscretize, and ARFFstring. The WEKA classes are mentioned with a reference to the WEKA documentation. At the bottom, a command is entered: `> java ARFFstring data/departments/A A deptA`.

```
Welcome to the WEKA SimpleCLI

This interface provides access to the MDL classes and all WEKA classes.
Enter commands in the textfield at the bottom of the window. Use the up
and down arrows to move through previous commands. Command must be one of:

java <classname> <args> [ > file]
break
kill
cls
history
exit
help [<command>]

MDL classes (see http://www.cs.ccsu.edu/~markov/MDLclustering/MDLmanual.pdf) :

MDLcluster <input file>.arff [compression cutoff] [de] [<output file>.{arff|csv}]
MDLranker <input file>.arff [num of attributes] [<output file>.{arff|csv}] [de]
MDLdiscretize <input file>.arff <output file>.{arff|csv} [de]
ARFFstring <input directory> <class label> <output file name>

WEKA classes: see http://www.cs.waikato.ac.nz/ml/weka/documentation.html

> java ARFFstring data/departments/A A deptA

java ARFFstring data/departments/A A deptA
```

# Data preprocessing and visualization

Department data: Create ARFF file in string format (using SimpleCLI)

1. Create file deptA with the files in folder data/departments/A with class label A:

```
java ARFFstring data/departments/A A deptA
```

2. Create file deptB with the files in folder data/departments/B with class label B:

```
java ARFFstring data/departments/B B deptB
```

3. Merge deptA and deptB into one file departments-string.arff

4. Add the following ARFF file header in the beginning of departments-string.arff:

```
@relation departments_string
@attribute document_name string
@attribute document_content string
@attribute document_class {A,B}
@data
```

# Data preprocessing and visualization

## Loading text data in Weka

- String format for ID and content
- One document per line
- Add class (nominal) if needed

```
departments-string.arff - Notepad
File Edit Format View Help
@relation departments_string

@attribute document_name string
@attribute document_content string
@attribute document_class {A,B}

@data

"Anthropology.txt","Anthropology Anthropology Anthropology c
cultural anthropology, physical anthropology, archaeology, a
subfields, concentrations are offered in biological anthropo
comparison. The anthropology major provides students with a
science background and prepares students for a range of care
marketing and international management. Through independent
with faculty doing research. Students regularly attend profe
and other discipline-related events. Special programs includ
archaeology, internships in applied anthropology, and partic
diversity training institutes. PROGRAM OF STUDY: BA DEPARTME
Diloreto Hall 110 Phone: 830-2610 Department Website ",A
"Biology.txt","Biology Biological Sciences The undergraduate
by the Department of Biological Sciences explore the discipl
undergraduate programs are available in medical technology a
interpretation; also available are specialized graduate prog
health sciences. Students preparing for various health and m
primarily by the department's pre-health professions advisor
require a research project or internship. Many laboratories,
experimental garden, controlled environmental rooms, cell cu
facilities, photosynthesis research laboratory, growth chamb
are available for research and instruction. PROGRAMS OF STUD
Ruth Rollin Location: Copernicus Hall 332 Phone: 832-2645 De
Information ",A
```

The screenshot shows the Weka Explorer interface with the 'StringToNominal' filter applied to the 'document\_name' attribute. The 'Current relation' section shows 3 attributes and 20 instances. The 'Attributes' section lists 'document\_name', 'document\_content', and 'document\_class'. The 'Class' is set to 'document\_class (Nom)'. A status message at the bottom indicates 'Attribute is neither numeric nor nominal.'

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose StringToNominal -R 1 Apply

Current relation: Relation: departments\_string, Instances: 20, Attributes: 3, Sum of weights: 20, Name: document\_name, Type: String, Missing: 0 (0%), Distinct: 20, Unique: 20 (100%)

Attributes: All None Invert Pattern

No.	Name
1	<input type="checkbox"/> document_name
2	<input type="checkbox"/> document_content
3	<input type="checkbox"/> document_class

Class: document\_class (Nom) Visualize All

Status: OK Log x 0

# Data preprocessing and visualization

## Converting a string attribute into nominal

Choose filters/unsupervised/attribute/StringToNominal, set attributeRange to 1, click on Apply

The screenshot shows the Weka Explorer interface with the 'StringToNominal' filter applied to the 'document\_name' attribute. The filter dialog is open, showing the 'attributeRange' set to 1. The 'Selected attribute' table shows the conversion of 'document\_name' from a string to a nominal attribute with 20 unique values. A bar chart below the table shows the distribution of these values, with 14 blue bars and 6 red bars.

**Selected attribute**

No.	Label	Count	Weight
1	Anthropology.txt	1	1.0
2	Biology.txt	1	1.0
3	Chemistry.txt	1	1.0
4	Computer.txt	1	1.0
5	Economics.txt	1	1.0
6	Geography.txt	1	1.0
7	Math.txt	1	1.0
8	Physics.txt	1	1.0
9	Psychology.txt	1	1.0
10	Statistics.txt	1	1.0
11	Mathematics.txt	1	1.0
12	Science.txt	1	1.0
13	History.txt	1	1.0
14	Art.txt	1	1.0
15	Music.txt	1	1.0
16	Sports.txt	1	1.0
17	Health.txt	1	1.0
18	Education.txt	1	1.0
19	Business.txt	1	1.0
20	Law.txt	1	1.0

Class: document\_class (Nom) Visualize All

Status: OK

# Data preprocessing and visualization

Converting text data into TFIDF (Term Frequency – Inverted Document Frequency) attribute format

- Choose filters/unsupervised/attribute/StringToWordVector
- Set the parameters as needed (see “More”)
- Click on “Apply”

The screenshot shows the Weka Explorer interface. The 'Filter' tab is active, and the 'StringToWordVector' filter is selected. The filter parameters are: -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-ha. The 'Current relation' shows 'departments\_string-weka.filte...' with 772 attributes and 20 instances. The 'Selected attribute' table shows the following data:

No.	Label	Count	Weight
1	Anthropology.txt	1	1.0
2	Biology.txt	1	1.0
3	Chemistry.txt	1	1.0
4	Computer.txt	1	1.0
5	Economics.txt	1	1.0
6	Geography.txt	1	1.0
7	Math.txt	1	1.0
8	Physics.txt	1	1.0

The 'Attributes' list on the left shows 'document\_name' selected. The 'Status' bar at the bottom shows 'OK'.

The screenshot shows the 'weka.gui.GenericObjectEditor' dialog box for the 'StringToWordVector' filter. The 'About' section states: 'Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.' The 'More' and 'Capabilities' buttons are visible. The 'Parameters' section includes:

- IDFTransform: False
- TFTransform: False
- attributeIndices: first-last
- attributeNamePrefix: (empty)
- debug: False
- dictionaryFileToSaveTo: -- set me --
- doNotCheckCapabilities: False
- doNotOperateOnPerClassBasis: False
- invertSelection: False
- lowerCaseTokens: False
- minTermFreq: 1
- normalizeDocLength: No normalization
- outputWordCounts: False
- periodicPruning: -1.0
- saveDictionaryInBinaryForm: False
- stemmer: Choose NullStemmer
- stopwordsHandler: Choose Null
- tokenizer: Choose AlphabeticTokenizer
- wordsToKeep: 1000

The 'Open...', 'Save...', 'OK', and 'Cancel' buttons are at the bottom.

# Data preprocessing and visualization

Make document\_class last attribute

- Choose filters/unsupervised/attribute/Copy
- Set the index to 2 and click on Apply
- Remove attribute 2

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Filter' section shows 'Copy - R 2' is applied. The 'Current relation' section shows 'Relation: departments\_string-weka.filte...' with 673 attributes and 20 instances. The 'Attributes' section shows a list of attributes, with 'document\_name' selected. The 'Selected attribute' section shows a table with columns 'No.', 'Label', 'Count', and 'Weight'. The 'Class' is set to 'document\_class (Nom)'. A bar chart at the bottom shows 20 bars, with the first 10 being blue and the last 10 being red.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **Copy - R 2** Apply

Current relation

Relation: departments\_string-weka.filte... Attributes: 673  
Instances: 20 Sum of weights: 20

Selected attribute

Name: document\_name Type: Nominal  
Missing: 0 (0%) Distinct: 20 Unique: 20 (100%)

No.	Label	Count	Weight
1	Anthropology.txt	1	1.0
2	Biology.txt	1	1.0
3	Chemistry.txt	1	1.0
4	Computer.txt	1	1.0
5	Economics.txt	1	1.0
6	Geography.txt	1	1.0
7	Math.txt	1	1.0
8	Physics.txt	1	1.0
9	...	1	1.0
10	...	1	1.0

Class: document\_class (Nom) Visualize All

Status

OK Log x 0

# Data preprocessing and visualization

- Change the attributes to nominal (use NumericToNominal filter)
- Save data on a file for further use

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Filter' section shows the 'NumericToNominal -R first-last' filter applied. The 'Current relation' section shows the relation 'departments\_string-weka.file...' with 673 attributes and 20 instances. The 'Attributes' section shows a list of attributes, with 'Academic' selected. The 'Selected attribute' section shows the details for the 'Academic' attribute, which is nominal with 2 distinct values and 0 missing values. The 'Class' is set to 'document\_class (Nom)'. The 'Visualize All' button is visible, and a bar chart is displayed below it, showing the distribution of the 'Academic' attribute across the two classes.

**Current relation**

Relation: departments\_string-weka.file...      Attributes: 673  
Instances: 20      Sum of weights: 20

**Attributes**

All   None   Invert   Pattern

No.	Name
1	document_name
2	Academic
3	Accreditation
4	Ali
5	American
6	Antar
7	Anthropology
8	BA
9	BS
10	Biological
11	Biology
12	Board
13	Brian
14	CCSU

Remove

**Selected attribute**

Name: Academic      Type: Nominal  
Missing: 0 (0%)      Distinct: 2      Unique: 0 (0%)

No.	Label	Count	Weight
1	0	17	17.0
2	1	3	3.0

Class: document\_class (Nom)      Visualize All

17

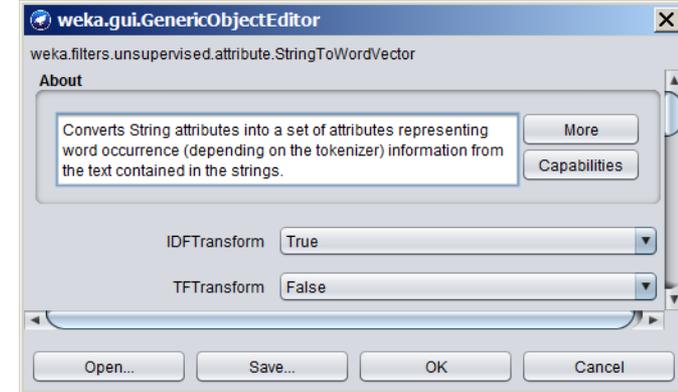
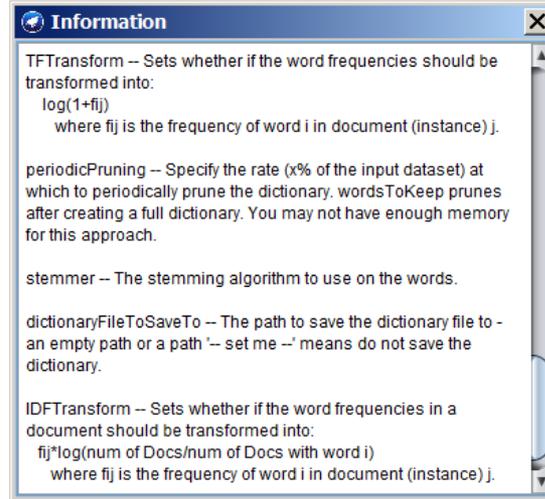
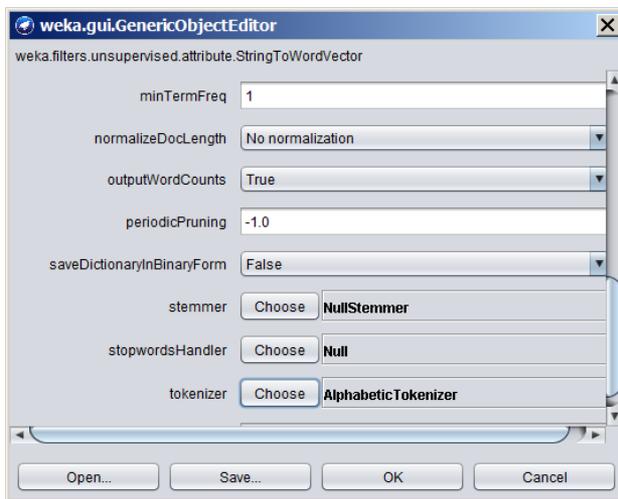
3

Status: OK      Log      x 0



# Data preprocessing and visualization

ARFF department data in TF and TFIDF format



Viewer

Relation: departments\_string-weka.filters.unsupervised.attribute.StringToNominal-R1-weka.filters.unsupervised.attribute.StringToWordVector-R2-W10...

No.	1: document_name	2: document_class	3: Academic	4: Accreditation	5: Ali	6: All	7: American	8: Antar	9: Anthropology	10: Available	11: BA
	Nominal	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	Anthropology.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	1.0
2	Biology.txt	A	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3	Chemistry.txt	A	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	Computer.txt	A	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
5	Economics.txt	A	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
6	Geography.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
7	Math.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
8	Physics.txt	A	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
9	Political.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
10	Psychology.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
11	Sociology.txt	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
12	Art.txt	B	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
13	Communication.txt	B	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
14	English.txt	B	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
15	History.txt	B	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	1.0
16	Justice.txt	B	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
17	Languages.txt	B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
18	Music.txt	B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
19	Philosophy.txt	B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
20	Theatre.txt	B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0

Buttons: Add instance, Undo, OK, Cancel

# Data preprocessing and visualization

## Student Projects

- [Preprocess.html](#)
- [Visualization.html](#)

# Attribute Selection

*Finding a minimal set of attributes that preserve the class distribution*

Attribute relevance with respect to the class – irrelevant attribute (*accounting*)

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Current relation' is 'departments\_string-weka.filte...' with 673 attributes and 20 instances. The 'Selected attribute' is 'accounting', which is a nominal attribute with 2 distinct values and 1 unique value (5%). A table shows the distribution of the 'accounting' attribute:

No.	Label	Count	Weight
1	0	19	19.0
2	1	1	1.0

The 'Attributes' list shows 'accounting' selected. A bar chart below the table visualizes the distribution, with a red bar for label '0' (count 19) and a blue bar for label '1' (count 1). The status bar at the bottom shows 'OK' and 'Log'.

IF accounting=1 THEN class=A (Error=0, Coverage = 1 instance → **overfitting** )

IF accounting=0 THEN class=B (Error=10/19, Coverage = 19 instances → **low accuracy**)

# Attribute Selection

Attribute relevance with respect to the class – relevant attribute (*science*)

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Current relation' is 'departments\_string-weka.filte...' with 673 attributes and 20 instances. The 'Selected attribute' is 'science', which is a nominal attribute with 2 distinct values and 0 missing values. The 'Attributes' list on the left shows 'science' selected. The 'Class' is 'document\_class (Nom)'. A bar chart at the bottom right shows the distribution of the 'science' attribute: 13 instances for label '0' (red bar) and 7 instances for label '1' (blue bar).

No.	Label	Count	Weight
1	0	13	13.0
2	1	7	7.0

IF science=1 THEN class=A (Error=0, Coverage = 7 instance)

IF science=0 THEN class=B (Error=4/13, Coverage = 13 instances)

# Attribute Selection (with document\_name)

The image shows two overlapping windows of Weka Explorer. The foreground window displays the results of an attribute selection process. The background window shows the filter settings for the 'StringToWordVector' filter.

**Weka Explorer (Foreground) - Attribute Evaluator**

Choose **CfsSubsetEval -P 1 -E 1**

**Search Method**

Choose **BestFirst -D 1 -N 5**

**Attribute Selection Mode**

Use full training set  
 Cross-validation Folds: 10 Seed: 1

(Nom) document\_class

Start Stop

**Result list (right-click for options)**

19:50:24 - BestFirst + CfsSubsetEval

**Attribute selection output**

```
=== Attribute Selection on all input data ===  
  
Search Method:  
Best first.  
Start set: no attributes  
Search direction: forward  
Stale search after 5 node expansions  
Total number of subsets evaluated: 4683  
Merit of best subset found: 0.652  
  
Attribute Subset Evaluator (supervised, Class (nominal): 673 document_class  
CFS Subset Evaluator  
Including locally predictive attributes  
  
Selected attributes: 1,362,368 : 3  
document_name  
research  
science
```

**Weka Explorer (Background) - Filter**

Choose **StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -C -N 0 -stemmer weka.core.stemmers.SnowballStemmer -stopw** Apply

**Selected attribute**

Name: document\_name Type: Nominal  
Missing: 0 (0%) Distinct: 20 Unique: 20 (100%)

No.	Label	Count	Weight
1	Anthropology.txt	1	1.0
2	Biology.txt	1	1.0
3	Chemistry.txt	1	1.0
4	Computer.txt	1	1.0
5	Economics.txt	1	1.0
6	Geography.txt	1	1.0
7	Math.txt	1	1.0
8	Physics.txt	1	1.0

Class: document\_class (Nom) Visualize All

Log x 0

# Attribute Selection (without document\_name)

Select document\_name and click on Remove

The screenshot shows the Weka Explorer interface. The 'Attributes' list is visible, with a 'Remove' button at the bottom. The 'Current relation' section shows 'Relation: departments\_string-weka.filte...' and 'Instances: 20'. The 'Attributes' section has buttons for 'All', 'None', 'Invert', and 'Pattern'. The list of attributes includes: Academic, Accreditation, Ali, American, Antar, Anthropology, BA, BS, Biological, Biology, Board, Brian, CCSU, and CFS.

No.	Name
1	Academic
2	Accreditation
3	Ali
4	American
5	Antar
6	Anthropology
7	BA
8	BS
9	Biological
10	Biology
11	Board
12	Brian
13	CCSU
14	CFS

The screenshot shows the Weka Explorer interface with the 'Attribute Evaluator' and 'Attribute Selection Mode' settings. The 'Attribute Evaluator' is set to 'CfsSubsetEval -P 1 -E 1'. The 'Search Method' is set to 'BestFirst -D 1 -N 5'. The 'Attribute Selection Mode' is set to 'Use full training set'. The 'Attribute selection output' window shows the results of the attribute selection process.

Attribute Evaluator: Choose CfsSubsetEval -P 1 -E 1

Search Method: Choose BestFirst -D 1 -N 5

Attribute Selection Mode:  Use full training set,  Cross-validation Folds: 10, Seed: 1

Attribute selection output: Attribute Subset Evaluator (supervised, Class (nominal): 672 document\_class, CFS Subset Evaluator, Including locally predictive attributes. Selected attributes: 7, 145, 236, 242, 282, 335, 361, 367, 458, 465, 511, 579, 648 : 13 BA, business, honors, include, making, professional, research, science, Internships, Languages, academic, history, special.



# Attribute Selection (ranking)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Attribute Evaluator: Choose GainRatioAttributeEval

Search Method: Choose Ranker -T-1.7976931348623157E308 -N-1

Attribute Selection Mode:  Use full training set,  Cross-validation, Folds: 10, Seed: 1

(Nom) document\_class

Start Stop

Result list (right-click for options): 20:03:40 - Ranker + GainRatioAttributeEv

Status: OK

Attribute selection output

Search Method: Attribute ranking.

Attribute Evaluator (supervi Gain Ratio feature e

Ranked attributes:

0.52115	361	research
0.44317	367	science
0.32239	648	special
0.32239	579	history
0.27955	7	BA
0.26675	465	Languages
0.26675	599	literature
0.26675	607	national
0.26675	551	diverse
0.26675	588	interdiscip
0.26675	458	Internships
0.26675	471	Maloney
0.26675	513	active
0.26675	528	composition

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Attribute Evaluator: Choose GainRatioAttributeEval

Search Method: Choose Ranker -T-1.7976931348623157E308 -N-1

Attribute Selection Mode:  Use full training set,  Cross-validation, Folds: 10, Seed: 1

(Nom) document\_class

Start Stop

Result list (right-click for options): 20:03:40 - Ranker + GainRatioAttributeEv

Status: OK

Attribute selection output

0.00172 150 center

0.00172 374 service

0.00172 372 select

0.00172 130 associate

0.00172 164 complete

0.00172 116 agencies

0.00172 268 leading

0.00172 395 summer

0.00172 366 schools

0.00172 272 linguistics

0 28 DEPARTMENT

0 29 Department

0 68 Phone

0 78 STUDY

0 14 CHAIR

0 101 Website

0 53 Location

Selected attributes: 361,367,648,579,7,465,599,607,551,588,458,471,513,528

Status: OK Log x 0

# Attribute Selection (explanation of ranking)

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **StringToWordVector -R 2 -W 1000 -prune-rate -1.0 -C -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.sto** Apply

**Current relation**  
Relation: departments\_string-weka.filters.unsu... | Attributes: 673  
Instances: 20 | Sum of weights: 20

**Attributes**  
All | None | Invert | Pattern

No.	Name
353	<input type="checkbox"/> range
354	<input type="checkbox"/> real
355	<input type="checkbox"/> regional
356	<input type="checkbox"/> regularly
357	<input type="checkbox"/> related
358	<input type="checkbox"/> relations
359	<input type="checkbox"/> relationship
360	<input type="checkbox"/> relativity
361	<input type="checkbox"/> require
362	<input checked="" type="checkbox"/> research
363	<input type="checkbox"/> rooms
364	<input type="checkbox"/> running
365	<input type="checkbox"/> scholarships
366	<input type="checkbox"/> school
367	<input type="checkbox"/> schools
368	<input type="checkbox"/> science
369	<input type="checkbox"/> sciences
370	<input type="checkbox"/> seat

Remove

**Selected attribute**  
Name: research | Type: Nominal  
Missing: 0 (0%) | Distinct: 2 | Unique: 0 (0%)

No.	Label	Count	Weight
1	0	12	12.0
2	1	8	8.0

Class: document\_class (Nom) | Visualize All

12 8

Status: OK | Log x 0

Attributes | Visualize

Generate... | Undo | Edit... | Save...

-prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-hand Apply

**Selected attribute**  
Name: Location | Type: Nominal  
Missing: 0 (0%) | Distinct: 1 | Unique: 0 (0%)

No.	Label	Count	Weight
1	1	20	20.0

Class: document\_class (Nom) | Visualize All

20

Status: OK | Log x 0

# Attribute Selection (using filters)

- Choose filters/supervised/attribute/AttributeSelection
- Set parameters to InfoGainAttributeEval and Ranker
- Click on Apply and see the attribute ordering

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **AttributeSelection** -E "weka.attributeSelection.GainRatioAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348"

Current relation

Relation: departments\_string-weka.filte... Attributes: 672  
Instances: 20 Sum of weights: 20

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> research
2	<input type="checkbox"/> science
3	<input type="checkbox"/> special
4	<input type="checkbox"/> history
5	<input type="checkbox"/> BA
6	<input type="checkbox"/> Languages
7	<input type="checkbox"/> literature
8	<input type="checkbox"/> national
9	<input type="checkbox"/> diverse
10	<input type="checkbox"/> interdisciplinary
11	<input type="checkbox"/> Internships
12	<input type="checkbox"/> Maloney
13	<input type="checkbox"/> active

Remove

Status

OK Log x 0

No.	Label	Count	Weight
1	0	12	12.0
2	1	8	8.0

Class: document\_class (Nom)

12

8

weka.gui.GenericObjectEditor

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes. More Capabilities

debug False

doNotCheckCapabilities False

evaluator Choose GainRatioAttributeEval

search Choose Ranker -T -1.7976931348623157E308 -N -1

Open... Save... OK Cancel

# Attribute Selection (using filters)

Choose filters/supervised/attribute/AttributeSelection and use CfsSubsetEval and BestFirst search. Then click on Visualize All

The screenshot displays the Weka Explorer interface. The 'Filter' section shows 'AttributeSelection' with 'CfsSubsetEval' and 'BestFirst' search methods. The 'Current relation' is 'departments\_string-weka.filters.un...' with 20 instances and 14 attributes. The 'Attributes' list shows 14 attributes, with 'BA' selected. The 'Selected attribute' table shows:

No.	Label
1	0
2	1

The 'Visualize All' window shows a bar chart for 'document\_class (Nom)'. The chart displays two bars: a blue bar with a value of 4 and a red bar with a value of 16. The total number of instances is 20.

Other attribute selection results are shown in a grid of bar charts:

- BA**: 4 (blue), 16 (red)
- business**: 17 (blue), 3 (red)
- honors**: 18 (blue), 2 (red)
- include**: 17 (blue), 3 (red)
- making**: 18 (blue), 2 (red)
- professional**: 15 (blue), 5 (red)
- research**: 12 (blue), 8 (red)
- science**: 13 (blue), 7 (red)
- Internships**: 18 (blue), 2 (red)
- Languages**: 18 (blue), 2 (red)
- academic**: 18 (blue), 2 (red)
- history**: 17 (blue), 3 (red)
- special**: 17 (blue), 3 (red)
- document\_class**: 11 (blue), 9 (red)

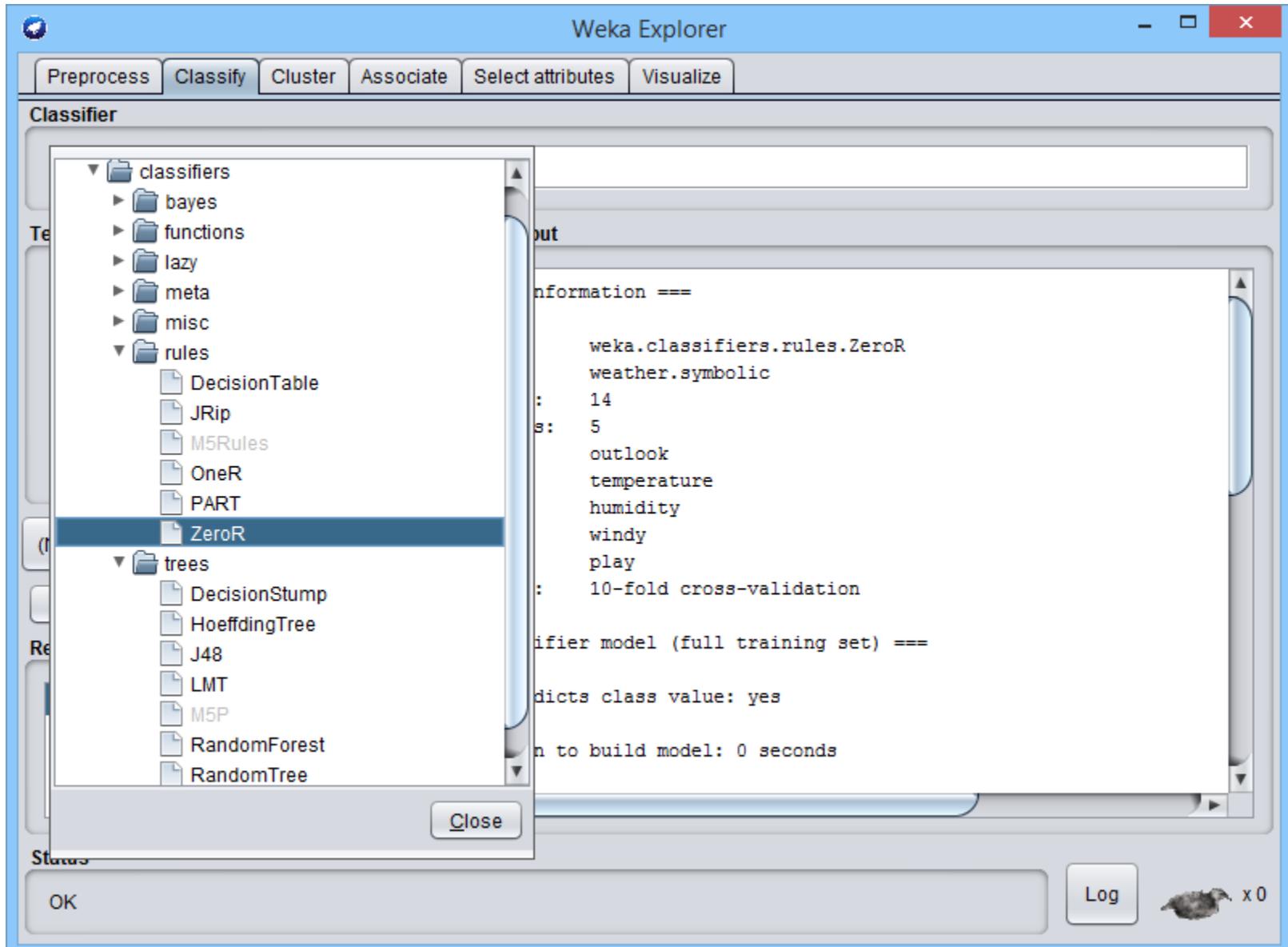
# Attribute Selection

## Student Projects

- [Attribute Selection.html](#)

# Classification – creating models (hypotheses)

*Mapping (independent attributes -> class)*



The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, and a list of classifiers is displayed. The 'ZeroR' classifier is selected. The 'Output' window shows the following text:

```
Information ===  
  
weka.classifiers.rules.ZeroR  
weather.symbolic  
: 14  
s: 5  
outlook  
temperature  
humidity  
windy  
play  
: 10-fold cross-validation  
  
Classifier model (full training set) ===  
predicts class value: yes  
Time to build model: 0 seconds
```

The status bar at the bottom shows 'OK' and 'Log' buttons, along with a small icon and the text 'x 0'.

# Classification – creating models (hypotheses)

## Inferring one-attribute rules - OneR

Weather data (weather.nominal.arff)

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	sunny	mild	high	FALSE	no
4	sunny	cool	normal	FALSE	yes
5	sunny	mild	normal	TRUE	yes
6	overcast	hot	high	FALSE	yes
7	overcast	cool	normal	TRUE	yes
8	overcast	mild	high	TRUE	yes
9	overcast	hot	normal	FALSE	yes
10	rainy	mild	high	FALSE	yes
11	rainy	cool	normal	FALSE	yes
12	rainy	cool	normal	TRUE	no
13	rainy	mild	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Attribute	Rules	Errors	Total error
outlook	sunny -> no overcast -> yes rainy -> yes	2/5 0/4 2/5	4/14
temperature	hot -> no mild -> yes cool -> yes	2/4 2/6 1/4	5/14
humidity	high -> no normal -> yes	3/7 1/7	4/14
windy	false -> yes true -> no	2/8 3/5	5/14

# Classification – OneR

The screenshot shows the Weka Explorer interface with the OneR classifier selected. The classifier output window displays the following information:

```
Scheme:      weka.classifiers.rules.OneR -B 6
Relation:    weather.symbolic
Instances:   14
Attributes:  5
              outlook
              temperature
              humidity
              windy
              play
Test mode:   evaluate on training data

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

outlook:
  sunny   -> no
  overcast -> yes
  rainy   -> yes
(10/14 instances correct)
```

The interface includes a top menu bar with options: Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The Classifier section shows 'OneR -B 6' selected. Test options include 'Use training set' (selected), 'Supplied test set', 'Cross-validation' (10 folds), and 'Percentage split' (66%). The result list shows '21:13:27 - rules.OneR'. The status bar indicates 'OK' and a 'Log' button.

# Classification – decision tree

Right click on the highlighted line in Result list and choose Visualize tree

The image shows the Weka Explorer interface with a decision tree classifier selected. The classifier is J48 - C 0.25 - M 2. The test options are set to Use training set, Cross-validation (10 folds), and Percentage split (66%). The result list shows a single entry: 21:19:33 - trees\_J48. The classifier output shows the pruned tree structure and its performance metrics.

**Classifier**  
Choose **J48 - C 0.25 - M 2**

**Test options**  
 Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66  
More options...

**Classifier output**  
==== Classifier model (full tree)====  
J48 pruned tree  
-----  
outlook = sunny  
| humidity = high: no (3.0)  
| humidity = normal: yes (2.0)  
outlook = overcast: yes (4.0)  
outlook = rainy  
| windy = TRUE: no (2.0)  
| windy = FALSE: yes (3.0)  
Number of Leaves : 5  
Size of the tree : 8  
Time taken to build model: 0 seconds

**Tree View**

```
graph TD
    outlook((outlook)) -- "= sunny" --> humidity((humidity))
    outlook -- "= overcast" --> yes40[yes (4.0)]
    outlook -- "= rainy" --> windy((windy))
    humidity -- "= high" --> no30[no (3.0)]
    humidity -- "= normal" --> yes20[yes (2.0)]
    windy -- "= TRUE" --> no20[no (2.0)]
    windy -- "= FALSE" --> yes30[yes (3.0)]
```

# Classification – decision tree

Top-down induction of decision trees (TDIDT, old approach know from pattern recognition):

- Select an attribute for root node and create a branch for each possible attribute value.
- Split the instances into subsets (one for each branch extending from the node).
- Repeat the procedure recursively for each branch, using only instances that reach the branch (those that satisfy the conditions along the path from the root to the branch).
- Stop if all instances have the same class.

ID3, C4.5, J48 (Weka): Select the attribute that minimizes the class entropy in the split.

# Classification – numeric attributes

weather.arff

**Weka Explorer**

Preprocess Classify Cluster Associate Select attributes Visualize

**Classifier**  
Choose **J48 -C 0.25 -M 2**

**Test options**  
 Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66  
More options...

**Classifier output**  
J48 pruned tree  
-----  
outlook = sunny  
| humidity <= 75: yes (2.0)  
| humidity > 75: no (3.0)  
outlook = overcast: yes (4.0)  
outlook = rainy  
| windy = TRUE: no (2.0)  
| windy = FALSE: yes (3.0)  
Number of Leaves : 5  
Size of the tree : 8  
Time taken to build model: 0 sec

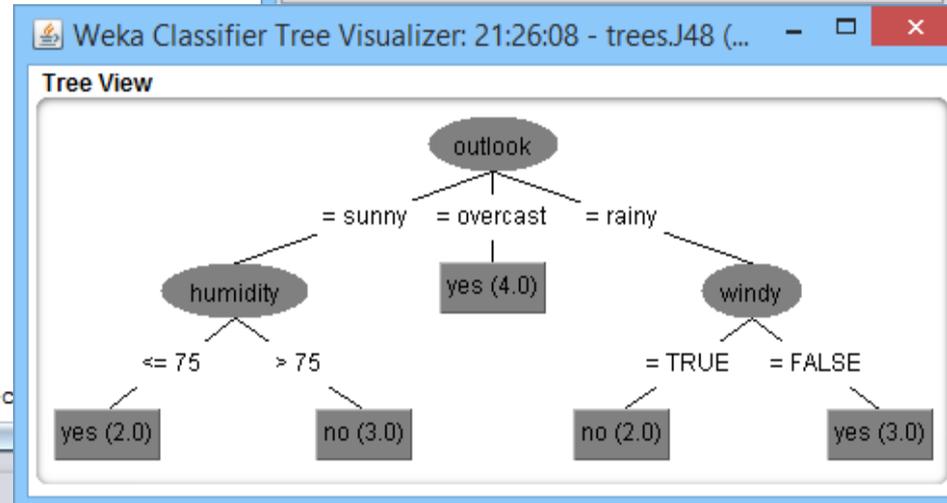
**Result list (right-click for options)**  
21:25:39 - trees.J48  
21:26:08 - trees.J48

**Status**  
OK

**Viewer**

Relation: weather

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no



# Classification – predicting class

Click on Set...

Click on Open file...

The screenshot shows the Weka software interface with several windows open:

- Test Instances**: A dialog box for loading data. It shows "Relation: weather..." and "Attributes: 5". The "Class" is set to "(Nom) play". Buttons for "Open file..." and "Open URL..." are visible.
- Open**: A file selection dialog showing the "data" directory. The file "weather.nominal.test.arff" is selected. The "Files of Type" is set to "Arff data files (\*.arff)".
- Classifier evaluation options**: A dialog box with various options checked, including "Output model", "Output per-class stats", "Output confusion matrix", and "Store predictions for visualization". The "Output predictions" are set to "PlainText".
- Classifier output**: A text window displaying the results of the classification process, including the decision tree structure and the test set predictions.
- weather.nominal.test.arff - Notepad**: A text editor window showing the contents of the loaded test set file.

**Classifier output**

```
outlook = sunny
| humidity = high: no (3.0)
| humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
```

Number of Leaves : 5  
Size of the tree : 8  
Time taken to build model: 0 seconds

=== Predictions on test set ===

inst#	actual	predicted	error	prediction
1	1:?	1:yes	1	

**weather.nominal.test.arff - Notepad**

```
@relation weather.nominal.test
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,mild,normal,FALSE,?
```

# Classification – predicting class

Right click on the highlighted line in Result list and choose Visualize classifier errors

Click on the square

The screenshot displays the Weka Explorer interface with several windows open:

- Weka Explorer:** The main window shows the 'Classify' tab. The classifier selected is 'J48 -C 0.25 -M 2'. The 'Test options' section has 'Supplied test set' selected. The 'Classifier output' window shows the following results:

```
outlook = sunny
| humidity = high: no (3.0)
| humidity = normal: yes (2.0)
| humidity = overcast: yes (4.0)
| humidity = rainy:
| windy:
RUE: no (2.0)
FALSE: yes (3.0)
ves : 5
ree : 8
build model: 0 seconds
n on test set ===
Time taken to test model on supplied test set: 0
=== Summary ===
Correctly Classified Instances 0 0 %
Incorrectly Classified Instances 1 100 %
```
- weather.nominal.test.arff - Notepad:** A Notepad window showing the ARFF file content:

```
@relation weather.nominal.test
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,mild,normal,FALSE,no
```
- Weka Classifier Visualize: 21:54:37 - trees.J48...:** A window for visualizing classifier errors. It shows 'X: outlook (Nom)' and 'Y: temperature (Nom)'. The plot is titled 'weather.symbolic\_predicted' and shows a 2D scatter plot with 'sunny', 'overcast', and 'rainy' on the x-axis and 'c', 'o', 'm', 'i', 'h', 'o' on the y-axis. A red square is visible on the plot.
- Weka: Instance info:** A window showing details for 'Instance: 1':

```
Plot : weka.classifiers.trees.J48
Instance: 1
outlook : sunny
temperature : mild
humidity : normal
windy : FALSE
prediction margin : -1.0
predicted play : yes
play : no
```

# Classification – predicting class

Click on Save

The image shows the Weka Classifier Visualize interface. The main window displays a plot titled "weather.symbolic\_predicted" with a scatter plot of data points. The X-axis is labeled "outlook (Nom)" and the Y-axis is labeled "temperature (Nom)". The plot shows a decision boundary separating the "sunny" and "rainy" regions. A red square represents the current instance being viewed.

Two windows are open in the foreground:

- Weka: Instance in...** displays the following information:

```
Plot : weka.classifiers.trees.J48
Instance: 1
  outlook : sunny
  temperature : mild
  humidity : normal
  windy : FALSE
prediction margin : -1.0
predicted play : yes
play : no
```
- weather.nominal.test.error.arff - Not...** displays the ARFF file content:

```
@relation weather.symbolic_predicted
@attribute outlook {sunny,overcast,rainy}
@attribute temperature {hot,mild,cool}
@attribute humidity {high,normal}
@attribute windy {TRUE,FALSE}
@attribute 'prediction margin' numeric
@attribute 'predicted play' {yes,no}
@attribute play {yes,no}

@data
sunny,mild,normal,FALSE,-1,yes,no
```

# Classification

## Student Projects

- [Classification.html](#)

# Prediction (no model, lazy learning)

test: (sunny, cool, high, TRUE, ?)

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

- K-nearest neighbor (IBk)  
*Take the class of the nearest neighbor or the majority class among K neighbors*

K=1 -> no

K=3 -> no

K=5 -> yes

K=14 -> yes (Majority predictor, ZeroR)

- Weighted K-nearest neighbor

K=5 -> undecided

no =  $1/1 + 1/2 = 1.5$

yes =  $1/2 + 1/2 + 1/2 = 1.5$

X	2	8	9	11	12	...	10
Distance(test,X)	1	2	2	2	2	...	4
play	no	no	yes	yes	yes	...	yes

- Distance is calculated as the number of different attribute values
- Euclidean distance for numeric attributes

# Prediction (no model, lazy learning)

The screenshot displays the Weka Explorer interface with several windows open. The main window shows the 'Classify' tab with the 'IBk' classifier selected. The 'Test options' section is set to 'Supplied test set'. The 'Classifier evaluation options' dialog is open, showing various output options. A Notepad window displays the ARFF file content for 'weather.nominal.test1.arff'. The 'weka.gui.GenericObjectEditor' dialog is open, showing the configuration for the 'K-nearest neighbours classifier' with 'KNN' set to 1 and 'nearestNeighbourSearchAlgorithm' set to 'LinearNNSearch'. The main window's output area shows the classifier's performance on the test set.

```
@relation weather.nominal.test
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,cool,high,TRUE,?
```

Instance-based classifier  
1 nearest neighbour(s) for classification  
Time taken to build model: 0 seconds  
Predictions on test set ===

Inst#	actual	predicted	error	prediction
1	1:?	2:no	0.938	

Evaluation on test set ===

# Prediction

## Student Projects

- [Prediction.html](#)

# Model evaluation – holdout (percentage split)

The image shows the Weka Explorer interface with the J48 classifier selected. The 'Test options' section is set to 'Percentage split' at 66%. A terminal window titled '11:35:56 - trees.J48' displays the following output:

```
=== Predictions on test split ===

inst#   actual   predicted error prediction
  1     1:y     2:n    +    1
  2     1:y     2:n    +    1
  3     2:n     2:n          1
  4     1:y     2:n    +    1
  5     1:y     2:n    +    1
  6     1:y     2:n    +    1
  7     2:n     2:n          1

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          2           28.5714 %
Incorrectly Classified Instances        5           71.4286 %
Total Number of Instances              7

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                0.000    0.000    0.000     0.000    0.000     y
                1.000    1.000    0.286     1.000    0.444     n
Weighted Avg.   0.286    0.286    0.082     0.286    0.127

=== Confusion Matrix ===

 a b  <-- classified as
 0 5 | a = y
 0 2 | b = n
```

The terminal window also shows the classifier model (full training set) as a pruned tree with the following structure:

```
emp = y
| buy = pc: y (8.0)
| buy = car
| | money <= 50: n (5.0/1.0)
| | money > 50: y (4.0)
emp = n: n (3.0)
```

# Model evaluation – cross validation

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

**Classifier**

Choose **J48 -C 0.25 -M 2**

**Test options**

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) approved

**Result list (right-click for options)**

11:38:00 - trees.J48

11:41:25 - trees.J48

**Status**

OK

11:41:25 - trees.J48

=== Predictions on test data ===

inst#	actual	predicted	error	prediction
1	2:n	2:n	1	
2	1:y	2:n	+	1
1	2:n	2:n	0.75	
2	1:y	1:y	1	
1	2:n	2:n	1	
2	1:y	1:y	1	
1	2:n	2:n	0.75	
2	1:y	1:y	1	
1	2:n	2:n	0.75	
2	1:y	1:y	1	
1	2:n	2:n	1	
2	1:y	1:y	1	
1	2:n	2:n	1	
2	1:y	1:y	1	
1	2:n	2:n	1	
2	1:y	1:y	1	
1	2:n	2:n	1	
2	1:y	1:y	1	

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	19	95	%
Incorrectly Classified Instances	1	5	%
Total Number of Instances	20		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Weighted Avg.	0.923	0.027	0.875	0.923	0.960	y
	1.000	0.077	0.875	1.000	0.933	n

=== Confusion Matrix ===

a	b	<-- classified as
12	1	a = y
0	7	b = n

**Classifier output**

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: LoanData

Instances: 20

Attributes: 12

ID

sex

married

age

money

pay

months

buy

emp

lastemp

area

approved

Test mode: 10-fold cross-validation

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

J48 pruned tree

-----

```

emp = y
| buy = pc: y (8.0)
| buy = car
| | money <= 50: n (5.0/1.0)
| | money > 50: y (4.0)
emp = n: n (3.0)
    
```

# Model evaluation – leave one out cross validation

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

**Classifier**

Choose **J48 -C 0.25 -M 2**

**Test options**

Use training set  
 Supplied test set (Set...)  
 Cross-validation Folds   
 Percentage split %   
More options...

**Classifier output**

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2  
Relation: LoanData  
Instances: 20  
Attributes: 12  
ID  
sex  
married  
age  
money  
pay  
months  
buy  
emp  
lastemp  
area  
approved

Test mode: 20-fold cross-validation

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

J48 pruned tree  
-----  
emp = y  
| buy = pc: y (8.0)  
| buy = car  
| | money <= 50: n (5.0/1.0)  
| | money > 50: y (4.0)  
emp = n: n (3.0)

**Result list (right-click for options)**

- 11:38:00 - trees.J48
- 11:41:25 - trees.J48
- 11:43:23 - trees.J48

**Status**

OK

11:43:23 - trees.J48

=== Predictions on test data ===

inst#	actual	predicted	error	prediction
1	2:n	2:n		1
1	2:n	2:n	0.75	
1	2:n	2:n		1
1	2:n	2:n	0.75	
1	2:n	2:n	0.75	
1	2:n	2:n		1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	2:n	+	1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	1:y		1
1	1:y	1:y		1

=== Stratified cross-validation ===

=== Summary ===

Metric	Value	Total
Correctly Classified Instances	19	95
Incorrectly Classified Instances	1	5
Total Number of Instances	20	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.923	0.000	1.000	0.923	0.960	y
	1.000	0.077	0.875	1.000	0.933	n
Weighted Avg.	0.950	0.027	0.956	0.950	0.951	

=== Confusion Matrix ===

a	b	<-- classified as	
12	1	a = y	
0	7	b = n	

# Model evaluation – confusion (contingency) matrix

**Classifier**

Choose **J48 -C 0.25 -M 2**

**Test options**

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

**Classifier output**

```

inst#  actual  predicted error prediction
  1    1:yes  1:yes    0.667
  2    1:yes  2:no     + 0.75
  3    1:yes  1:yes    1
  4    1:yes  1:yes    1
  5    2:no   1:yes    + 0.667

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances      3      60  %
Incorrectly Classified Instances    2      40  %
Total Number of Instances          5

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  Class
                -----  -----  -
                0.750    1.000    0.750     0.750    0.750     yes
                0.000    0.250    0.000     0.000    0.000     no
Weighted Avg.   0.600    0.850    0.600     0.600    0.600

=== Confusion Matrix ===

a b  <-- classified as
3 1 | a = yes
1 0 | b = no
    
```

**Result list (right-click for options)**

- 11:38:00 - trees.J48
- 11:41:25 - trees.J48
- 11:43:23 - trees.J48
- 11:46:56 - trees.J48
- 11:47:21 - trees.J48

**Status**

OK  x0

predicted

actual

	yes	no
yes	3	1
no	1	0

predicted

actual

	yes	no
yes	TP	FN
no	FP	TN

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

# Model evaluation

## Student Projects

- [Evaluation.html](#)

# Clustering – k-means

Click on Ignore attributes

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

**Clusterer**

Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2

**Cluster mode**

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation (Nom) play
- Store clusters for visualization

**Ignore attributes**

**Result list (r)**

00:03:39

**Select items**

- outlook
- temperature
- humidity
- windy
- play**

**Status**

OK

**weka.gui.GenericObjectEditor**

weka.clusterers.SimpleKMeans

**About**

Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory 100

canopyMinimumCanopyDensity 2.0

canopyPeriodicPruningRate 10000

canopyT1 -1.25

canopyT2 -1.0

debug False

displayStdDevs False

distanceFunction Choose **EuclideanDistance -R first I**

doNotCheckCapabilities False

dontReplaceMissingValues False

fastDistanceCalc False

initializationMethod Random

maxIterations 500

numClusters 2

numExecutionSlots 1

preserveInstancesOrder False

reduceNumberOfDistanceCalcsViaCanopies False

seed 10

# Clustering – classes to clusters evaluation

Right click on Result list, select Visualize cluster assignments

Click on Save

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'SimpleKMeans' algorithm is used with parameters: -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -t2 10000. The 'Classes to clusters evaluation' mode is selected, with the 'play' attribute chosen for evaluation. The 'Clusterer output' pane shows the following results:

```
Time taken to build model (full training): 0.000 seconds
=== Model and evaluation on training set ===

Clustered Instances

0      10 ( 71%)
1       4 ( 29%)

Class attribute: play
Classes to Clusters:

0 1 <-- assigned to cluster
6 3 | yes
4 1 | no

Cluster 0 <-- yes
Cluster 1 <-- no

Incorrectly clustered instances :          7.0
```

The 'Result list' shows a single entry: '00:15:54 - SimpleKMeans'. A right-click context menu is open over this entry, with 'Visualize cluster assignments' selected. This opens the 'Weka Clusterer Visualize' dialog, which is configured to visualize the 'weather.symbolic\_clustered' plot. The 'Class colour' section is empty. A 'Notepad++' window is also open, displaying the ARFF file for the visualization:

```
@relation weather.symbolic_clustered
@attribute Instance_number numeric
@attribute outlook {sunny,overcast,rainy}
@attribute temperature {hot,mild,cool}
@attribute humidity {high,normal}
@attribute windy {TRUE,FALSE}
@attribute play {yes,no}
@attribute Cluster {cluster0,cluster1}

@data
0,sunny,hot,high,FALSE,no,cluster0
1,sunny,hot,high,TRUE,no,cluster0
2,overcast,hot,high,FALSE,yes,cluster0
3,rainy,mild,high,FALSE,yes,cluster0
4,rainy,cool,normal,FALSE,yes,cluster1
5,rainy,cool,normal,TRUE,no,cluster1
6,overcast,cool,normal,TRUE,yes,cluster1
7,sunny,mild,high,FALSE,no,cluster0
8,sunny,cool,normal,FALSE,yes,cluster0
9,rainy,mild,normal,FALSE,yes,cluster0
10,sunny,mild,normal,TRUE,yes,cluster0
11,overcast,mild,high,TRUE,yes,cluster0
12,overcast,hot,normal,FALSE,yes,cluster1
13,rainy,mild,high,TRUE,no,cluster0
```

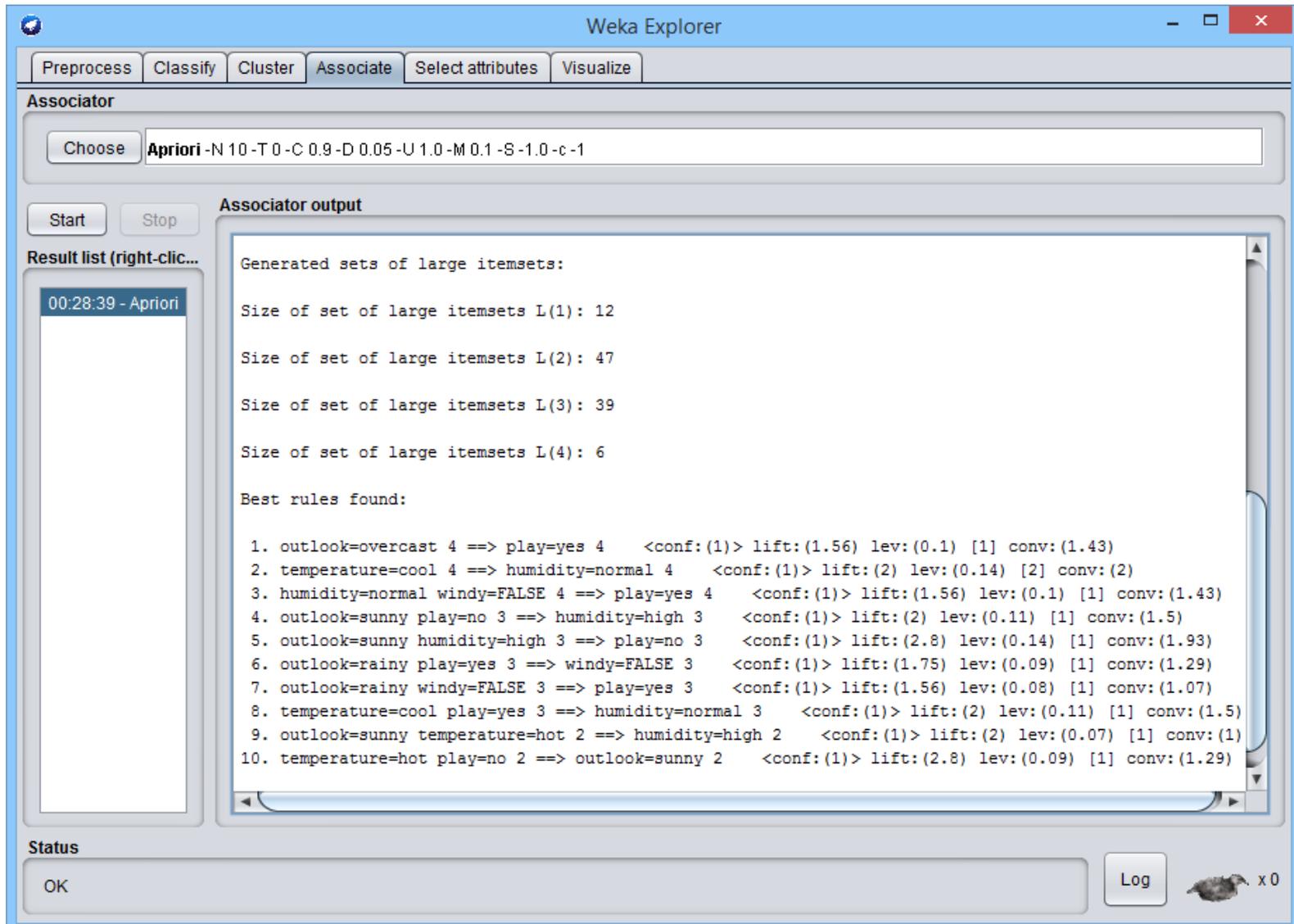
# Clustering

## Student Projects

- [Clustering.html](#)

# Association Rules ( $A \Rightarrow B$ )

- *Confidence* (accuracy):  $P(B|A) = (\# \text{ of tuples containing both } A \text{ and } B) / (\# \text{ of tuples containing } A)$ .
- *Support* (coverage):  $P(A,B) = (\# \text{ of tuples containing both } A \text{ and } B) / (\text{total } \# \text{ of tuples})$



The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Associator' section is set to 'Apriori' with parameters: -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1. The 'Associator output' pane displays the following results:

```
Generated sets of large itemsets:  
Size of set of large itemsets L(1): 12  
Size of set of large itemsets L(2): 47  
Size of set of large itemsets L(3): 39  
Size of set of large itemsets L(4): 6  
  
Best rules found:  
1. outlook=overcast 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)  
2. temperature=cool 4 ==> humidity=normal 4 <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)  
3. humidity=normal windy=FALSE 4 ==> play=yes 4 <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)  
4. outlook=sunny play=no 3 ==> humidity=high 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)  
5. outlook=sunny humidity=high 3 ==> play=no 3 <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)  
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)  
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)  
8. temperature=cool play=yes 3 ==> humidity=normal 3 <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)  
9. outlook=sunny temperature=hot 2 ==> humidity=high 2 <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)  
10. temperature=hot play=no 2 ==> outlook=sunny 2 <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)
```

The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

# Association Rules

## Student Projects

- [Association.html](#)

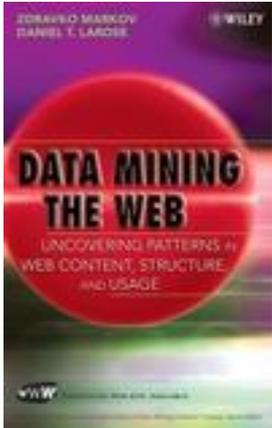
# Document classification and clustering

## Predict the class of the Theatre document

1. Create a training set – all departments excluding Theatre (data collection)
2. Use Binary, Term Frequency or TFIDF representation (data preprocessing)
3. Select a relevant subset of attributes (attribute selection)
4. Use J48, IBk, and Naïve Bayes (classification)
5. Evaluate all models by cross validation (model evaluation)
6. Choose the best model and predict the class of Theatre (prediction)
7. Cluster the training set with K-means compare the cluster centroids with Theatre

# Document classification and clustering

## Teaching resources and student projects based on Weka



- Zdravko Markov and Daniel T. Larose, *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, [Wiley](#) 2007 (free excerpts: Chapter 1, TOC, Index)
- Lecture slides: [dmw1.pdf](#), [dmw2.pdf](#), [dmw3.pdf](#), [dmw4.pdf](#), [dmw5.pdf](#)
- Data sets: <http://www.cs.ccsu.edu/~markov/dmwdata.zip>
- [Clustering.html](#)

<http://www.cs.ccsu.edu/~markov/MDLclustering/>

<http://www.cs.ccsu.edu/~markov/DMWprojects>

DMW Student Projects - Internet Explorer

<http://www.cs.ccsu.edu/>

File Edit View Favorites Tools Help

Zdravko Markov and Daniel T. Larose, *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, Wiley, April 2007.

---

### Student Projects

- [Web Document Classification](#)
- [Probabilistic Reasoning with Naïve Bayes and Bayesian Networks](#)
- [Relational Learning for Web Document Classification](#)
- [Web User Profiling](#)

---

Zdravko Markov  
<http://www.cs.ccsu.edu/~markov/>

<http://www.cs.ccsu.edu/>

File Edit View Favorites Tools >>

## MDL Clustering

Algorithms for unsupervised attribute ranking, discretization and clustering available as Java classes through a command-line interface. All Weka classes are also included.

- [Manual](#)
- [Executable JAR file](#)
- [Data](#)
- [Lab Project](#)

---

Zdravko Markov  
<http://www.cs.ccsu.edu/~markov/>

100%