Data Mining III – Lesson 1

Tamara B. Sipes, Ph.D.

Lesson 1 Overview

- Introduction
- Course Description and Goals
- Data Mining III Overview
- Prerequisite Refresher:
  - Data Mining I (required)
  - Data Mining II (recommended)
  - Data Preparation for Data Mining (required)
- Weka Refresher

Introduction

- Data Mining III combines and builds on everything you have learned in Data Mining I, Data Mining II and Data Preparation for Data Mining
- Prerequisites: Data Prep, Data Mining I
- Recommended: Data Mining II
- We will go through several data mining projects, beginning to the end, planning and executing each of the steps of the data preparation, analysis, learning and modeling, finishing with a predictive/descriptive model that produces the best evaluation scores.

Course Description and Goals

- This class is designed to give you an in-depth knowledge of practical data mining and predictive modeling
- Necessary to have acquired the necessary theoretical knowledge of data mining and machine learning techniques (such as in DM I, and possibly DM II)
- Absolutely essential to have gone through Data Preparation for Data Mining class as well.

Course Description and Goals

- A challenging class that is sure to make you a solid data miner ready to attack those complex real-life data mining tasks that are awaiting you.
- Goals:
  - **G**ain hands-on experience of the whole process.
  - **O**btain the ability to find your way through the decision paths awaiting.
  - **A**cquire the knowledge and experience to take on real life modeling tasks without any guidance.
  - **L**earn how to deal with extra difficult datasets (hint: re-sampling and ensemble learners).

Data Mining III Overview

- Lesson 1: Introduction, Prerequisites, Description, Goals, Refreshers
- Lesson 2 & 3: Hands-On Case Studies, Step by Step Data Preparation and Modeling of Real-World Data
- Lesson 4 & 5: Heading Towards More Difficult Datasets: Various Issues and How to Deal with them
- Lesson 6: Ensemble Modeling: Why and How
- Lesson 7: Summary, Conclusions, Tips and Other Helpful Guidance

Prerequisite Refresher

- You have taken:
  - Data Mining I
  - Data Mining II (possibly)
  - Data Preparation for Data Mining
- Let's see what we need to remember in order to proceed with this class

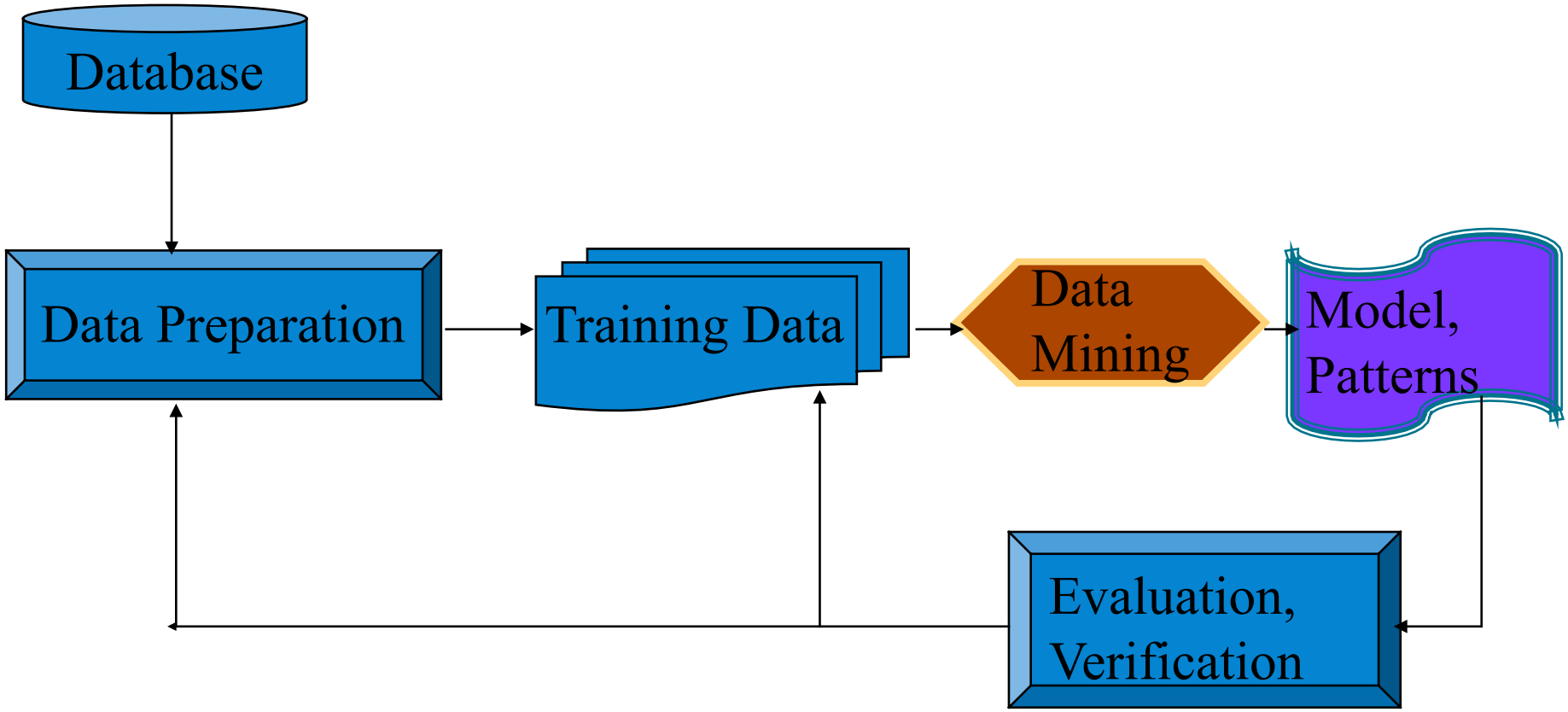
Data Mining I Summary

Data Mining Defined

- Data mining (knowledge discovery in databases) can be defined as:

  Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in databases

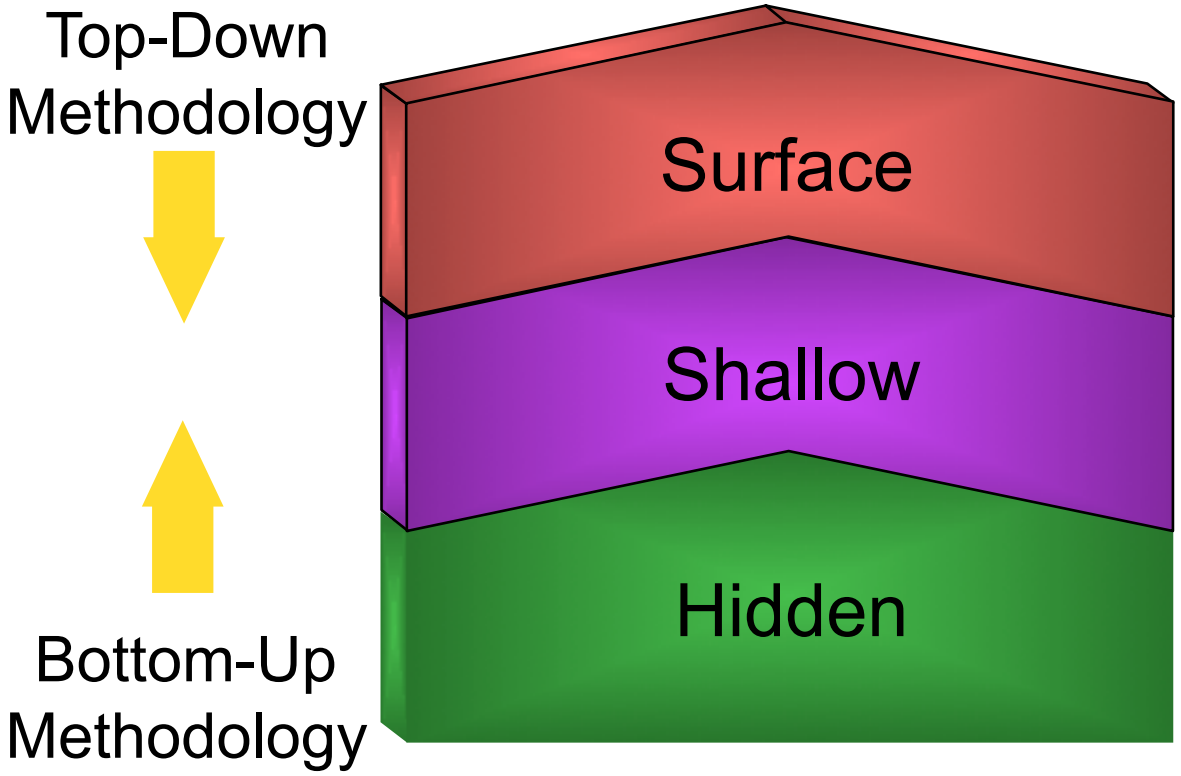
High Level View of Data Mining



What is DATA MINING?

- *Extracting or "mining" knowledge from large amounts of data*
- Data -driven discovery and modeling of hidden patterns (we never new existed) in large volumes of data
- Extraction of novel, potentially extremely useful patterns from data

Data Mining vs. Other Tools



Analytical Tools

SQL for simple queries and reporting

Statistics & OLAP for summaries and analysis

Data Mining for knowledge discovery

The Process of Data Mining

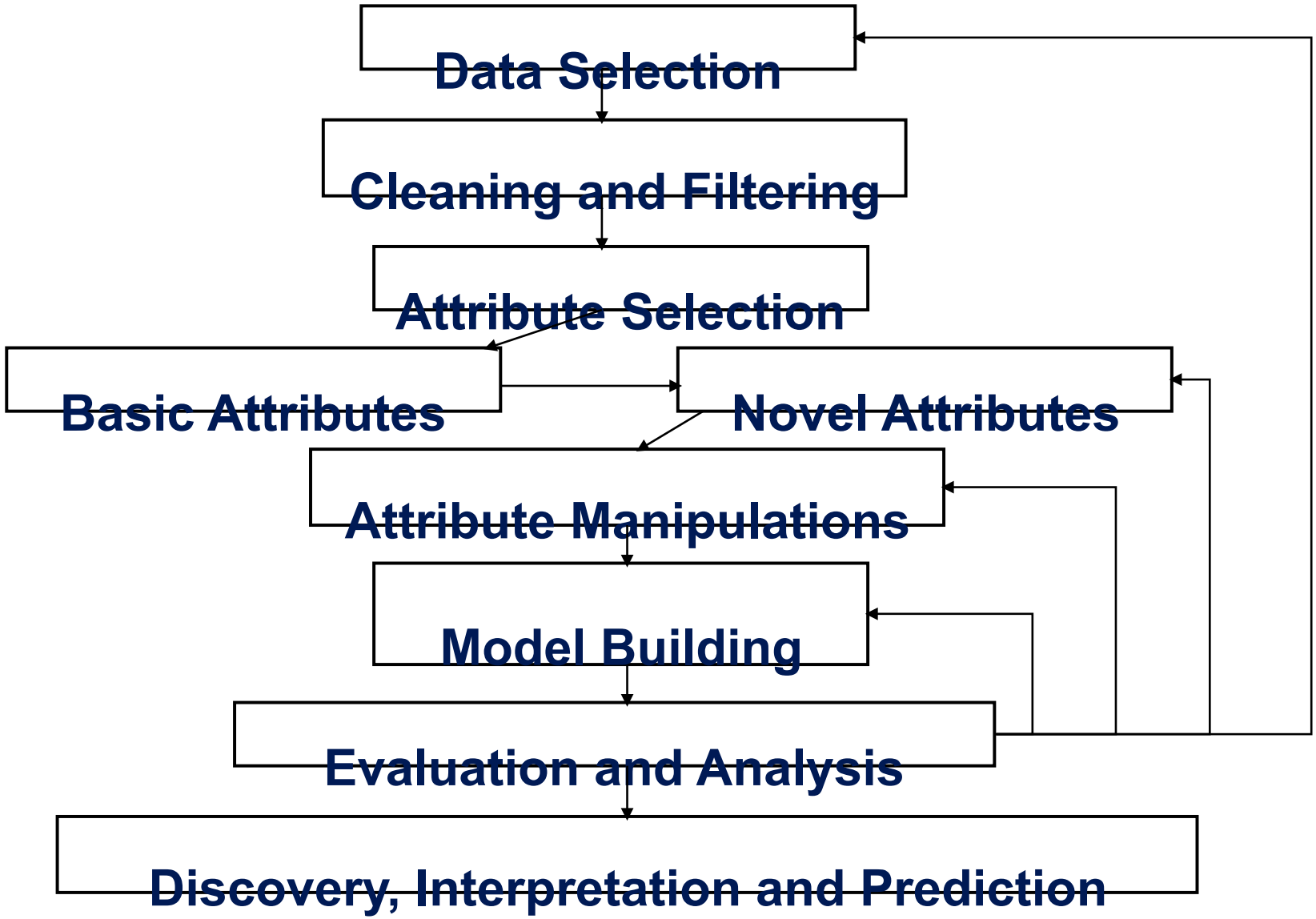
- Basic steps:
  - Data Acquisition
  - Data Preparation
  - Modeling
  - Evaluation
  - Feedback

CRISP-DM Methodology

- Cross Industry Standard Process for Data Mining
  - http://www.crisp-dm.org/
- Six Phases:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

The Details of the DM Process



The Data Mining Process

- Iterative Nature
- Exploratory Process
- Highly tailored to the dataset
- Need for Fine-Tuning
- Need for Model Revision from time to time

Types of Learning

- Supervised vs. unsupervised
- Incremental vs. non-incremental
- Descriptive vs. predictive

LEARNING AND MODELING METHODS

- Decision Tree Induction
- Regression Tree Induction
- Multi-variate Regression Tree
- Clustering
  - K-means, EM, Cobweb
- Neural network
  - Backpropagation
  - Recurrent
- IBL, Bayesian Learning, etc.

Basic Methods in depth

- 1R
- Decision Tree
- Instance-Based Learning
- Classification Rules: List, Table, Tree
- Numeric Prediction: Regression and Model Trees
- Clustering: k-means, EM, cobweb
- Association Rules
- Bayesian Learners

Evaluation

- Types of evaluation methods
- Numeric vs. nominal evaluation
- Feedback
- Comparing the models
- Useful tools: confusion matrix, ROC curves, etc.

Data Mining II Review

Advanced Data Mining Methods

- In depth presentation of ANNs
- Training and learning in ANNs
- Different ANN capabilities (tasks)
- ANN real-life applications
- Bayesian Learning
- HMM
- SVM
- Engineering the input and output

Types of ANNs

- Perceptrons
- Delta-Learning ANNs
- Backpropagation
- Multi-layered Neural Networks
- Radial-basis function (RBF) networks
- Recurrent Neural Networks

Engineering the input and output

- Modifying the input: attribute selection, discretization, transformations

- Modifying the output: combining classification models to improve performance
  - Bagging, boosting, stacking, error-correcting output codes

Combining multiple models

- Basic idea of "meta" learning schemes
  - build different "experts" and let them vote
- Advantage
  - often improves predictive performance
- Disadvantage
  - produces output that is very hard to analyze
- Schemes
  - Bagging, boosting, stacking
  - Can be applied to both classification and numeric prediction problems

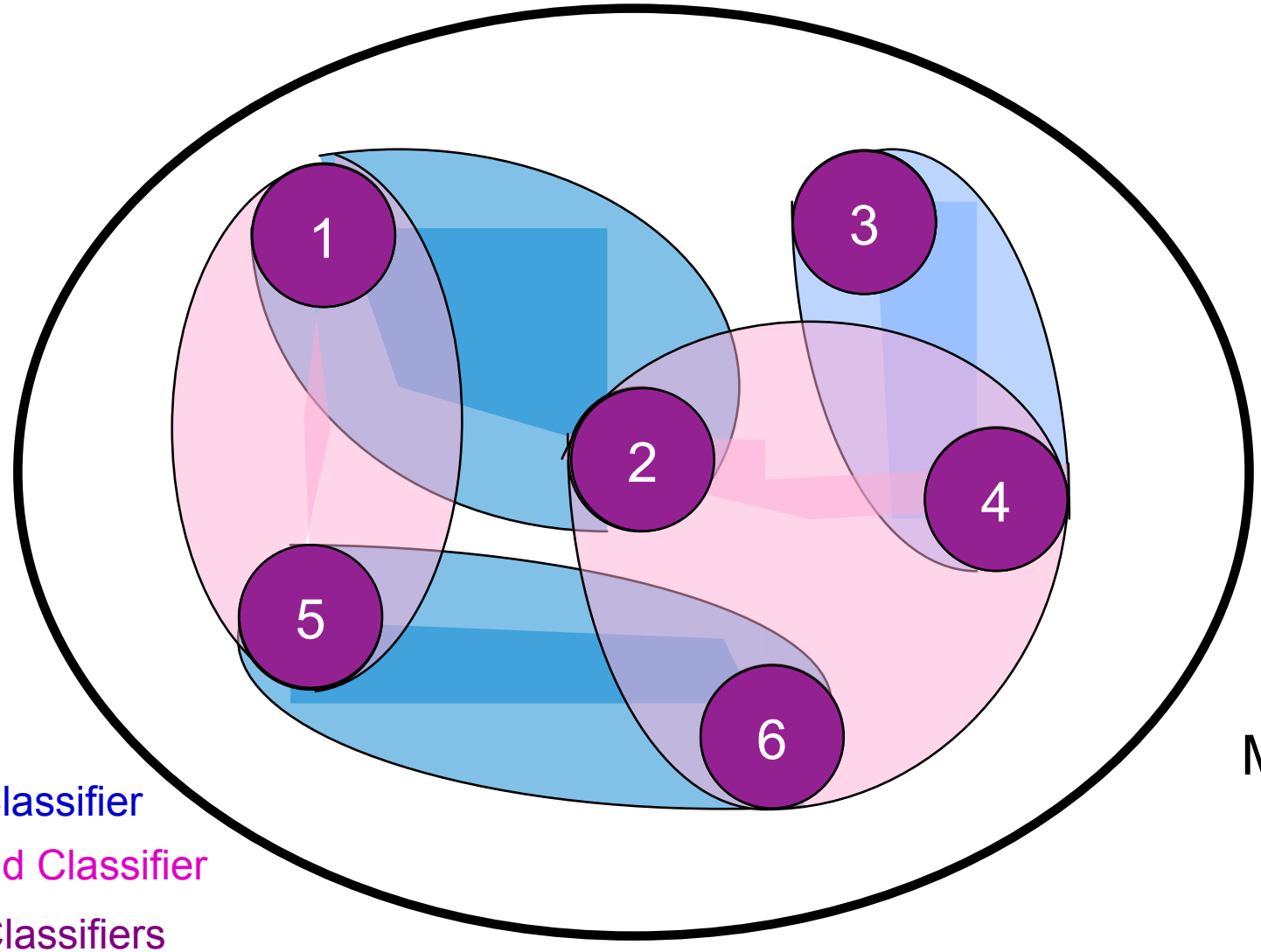
Combining Classifiers

- Given
  - Training data set *D* for supervised learning
  - Collection of inductive learning algorithms
  - Return: new prediction algorithm that combines outputs from collection of prediction algorithms
- Desired Properties
  - Guarantees of performance of combined prediction
    - ability to improve weak classifiers

Mixtures of Experts

- What Is A Weak Classifier?
  - One not guaranteed to do better than random guessing
  - Goal
    - combine multiple weak classifiers
    - get one at least as accurate as strongest
- Mixtures of Experts
  - “experts” express hypotheses
    - drawn from a hypothesis space

Improving Weak Classifiers



First Classifier

Second Classifier

Both Classifiers

Mixture Model

Summary

- ANNs, HMMs, SVMs very powerful tools
- Meta learner can produce even better results
- Bagging and boosting use the same method of aggregating different models together - voting
- Bagging, boosting and stacking can be applied to both classification and numeric prediction
- Bagging and boosting combine models of the same type, stacking – of different types

Data Preparation for Data Mining Run Through

Importance

- “Garbage in, garbage out”
- A crucial step of the DM process
- Need to know what to do with the “dirty data” – after it is collected and before it is mined

Data Preparation

- Could be the difference between a successful data mining project and a failure
- Could take 60-80% of the whole data mining effort

Definition

- Data Preparation is a process of cleaning, filtering and organizing the data for successful mining and modeling, by solving or avoiding problems in the data, and presenting the data to the modeling schema in the optimal way.

Good Data Preparation Practice

- When data is properly prepared and surveyed:
  - high quality modeling results are more likely
  - the quality of models produces will depend mostly on the content of the data, not so much on the modeler's expertise level.

The “Dark Side” of Data

- Missing values (null = empty or something else)
- Dirty data (erroneous zip codes, etc.)
- Inconsistent values (different revisions)
- Duplicate values
- GIGO (Garbage In, Garbage Out)

Prerequisites

- Data understanding:
  - Descriptors, values, ranges, labels
- Data history
- Exploring the Problem Space
- Exploring the Solution Space
- Implementation Method Specification
- Familiarity with the Nature of the World

Exploring the Problem Space

- A crucial starting point
- Avoids any possible misconceptions and unrealistic expectations from DM project

- A MUST: *identify the right problem to solve*

Dealing with variables

- Variables as objects
- Removing variables
- Sparcity
- Monotonicity
- Increasing dimensionality
- Outliers
- Numerating categorical variables
- Anachronisms

Building mineable data sets

- Make things as easy for the tool as possible!
- Exposing the information content
- Getting enough data
- Missing and empty values
  - to fill in or to discard?
- Shape of the data set

Overview of Basic Data Preparation Techniques

- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation

Summing up

- Data preparation is a big issue for mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but still an active area of research

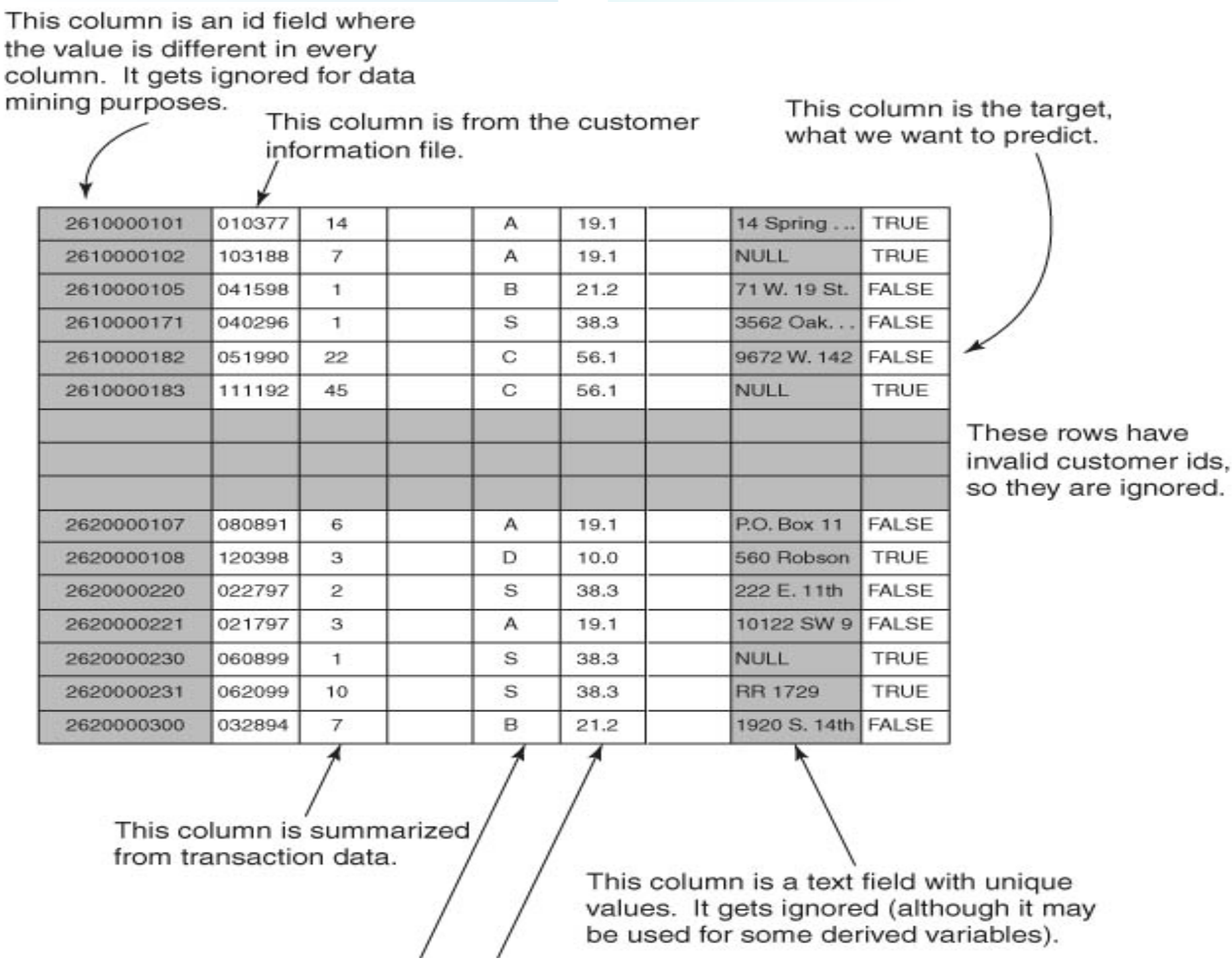
What the Data Should Look Like

- All data mining algorithms want their input in tabular form – rows & columns as in a spreadsheet or database

Reference	Time	Location	Count	Total Sales
1234	6/1/2000	San Diego	5	1700
1235	6/1/2000	San Diego	2	500
1236	6/1/2000	San Diego	3	1000
1234	6/2/2000	San Diego	5	1700
1235	6/2/2000	San Diego	2	500
1236	6/2/2000	San Francisco	3	1000
1240	6/2/2000	San Francisco	5	1500
1235	6/3/2000	San Francisco	2	500
1236	6/3/2000	San Francisco	3	1000
1240	6/3/2000	New York	10	3000
1235	6/3/2000	New York	2	500
1236	6/6/2000	New York	3	1000

What the Data Should Look Like

- Example: Customer Signature Data
  - Next slide
  - Continuous “snapshot” of customer behavior
  - Each row represents the customer and whatever might be useful for data mining



2610000101	010377	14		A	19.1		14 Spring . . .	TRUE
2610000102	103188	7		A	19.1		NULL	TRUE
2610000105	041598	1		B	21.2		71 W. 19 St.	FALSE
2610000171	040296	1		S	38.3		3562 Oak. . .	FALSE
2610000182	051990	22		C	56.1		9672 W. 142	FALSE
2610000183	111192	45		C	56.1		NULL	TRUE
2620000107	080891	6		A	19.1		P.O. Box 11	FALSE
2620000108	120398	3		D	10.0		560 Robson	TRUE
2620000220	022797	2		S	38.3		222 E. 11th	FALSE
2620000221	021797	3		A	19.1		10122 SW 9	FALSE
2620000230	060899	1		S	38.3		NULL	TRUE
2620000231	062099	10		S	38.3		RR 1729	TRUE
2620000300	032894	7		B	21.2		1920 S. 14th	FALSE

What the Data Should Look Like

- The columns
  - Contain data that describe aspects of the customer (e.g., sales $ and quantity for each of product A, B, C)
  - Contain the results of calculations referred to as *derived variables* (e.g., total sales $)

Num	Unit Price	Quantity	Total Price
12345	10.99	50	549.50
24357	21.95	7	153.65
87921	39.95	25	998.75

Data Assay and Data Survey

- The purpose of data assay: check that the data is coherent, sufficient, can be assembled into a needed format, and makes sense within the proposed framework
- Data survey is the process of taking a high-level overview to discover what is contained in the data set. The miner gains a very important insight into the nature of the data. Purpose of the Data Survey is to show and clarify shape, substance, structure, relationships and limits of the information in data.

Data Assay

- Assessment of quality of data for mining
- Leads to assembly of data sources to one file.

- How to get data and does it suit the purpose
- Main goal: miner understands where the data come from, what is there, and what remains to be done.
  - It is helpful to make a report on the state of data
- It involves miner directly - rather than using automated tools
  - After assay rest can be carried out with tools

Assessing Data - “Data assay”

- Data Discovery
  - Discovering and locating data to be used. Coping with the bureaucrats and data hiders.
- Data Characterization
  - What is it, the data found? Does it contain the information needed or is it mere garbage?
  - Domain experts
- Data Set Assembly
  - Making a “flat file” – a (ascii) table combining the data coming from different sources

Surveying the data

- The goal
  - to find the problem areas in the data, so that the mining can be planned optimally.
- Main tools

- **Confidence analysis**
- **Entropy analysis**
- **Analysis of sparsity and variability**
- **Cluster analysis**
- **Distribution analysis**

Basic Data Survey Procedure

- Estimate how well the data represents and covers the true population
- Analyze the entropy of and between the variables
- Try to explain the clusters
  - Check the mapping between input and output clusters.
- Check sparsity and uncertainty
- Check variable distributions
  - Try to explain the possible changes in the distributions.

Weka Refresher

- 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 is open source software issued under the GNU General Public License

Weka Reference

Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

FYI: Pentaho's live forum for Weka

- The open-source BI software company Pentaho has become major sponsor of Weka development and will take over the administration of Weka's Sourceforge site in the near future
- Pentaho also provides a live forum for interaction among Weka project community members:

http://forums.pentaho.org/forumdisplay.php?f=61

The Weka mailing list

- For Weka-related questions, comments, and bug reports to the Weka mailing list:

http://list.scms.waikato.ac.nz/mailman/listinfo/wekalist

- There is also the searchable mailing list archive:

https://list.scms.waikato.ac.nz/mailman/htdig/wekalist/

(Mirrors: news.gmane.org, and

Nabble (http://www.nabble.com/WEKA-f435.html).)

Weka Tutorials

- Numerous tutorials are given on the weka site!

- Weka Primer:

http://weka.sourceforge.net/wekadoc/index.php/

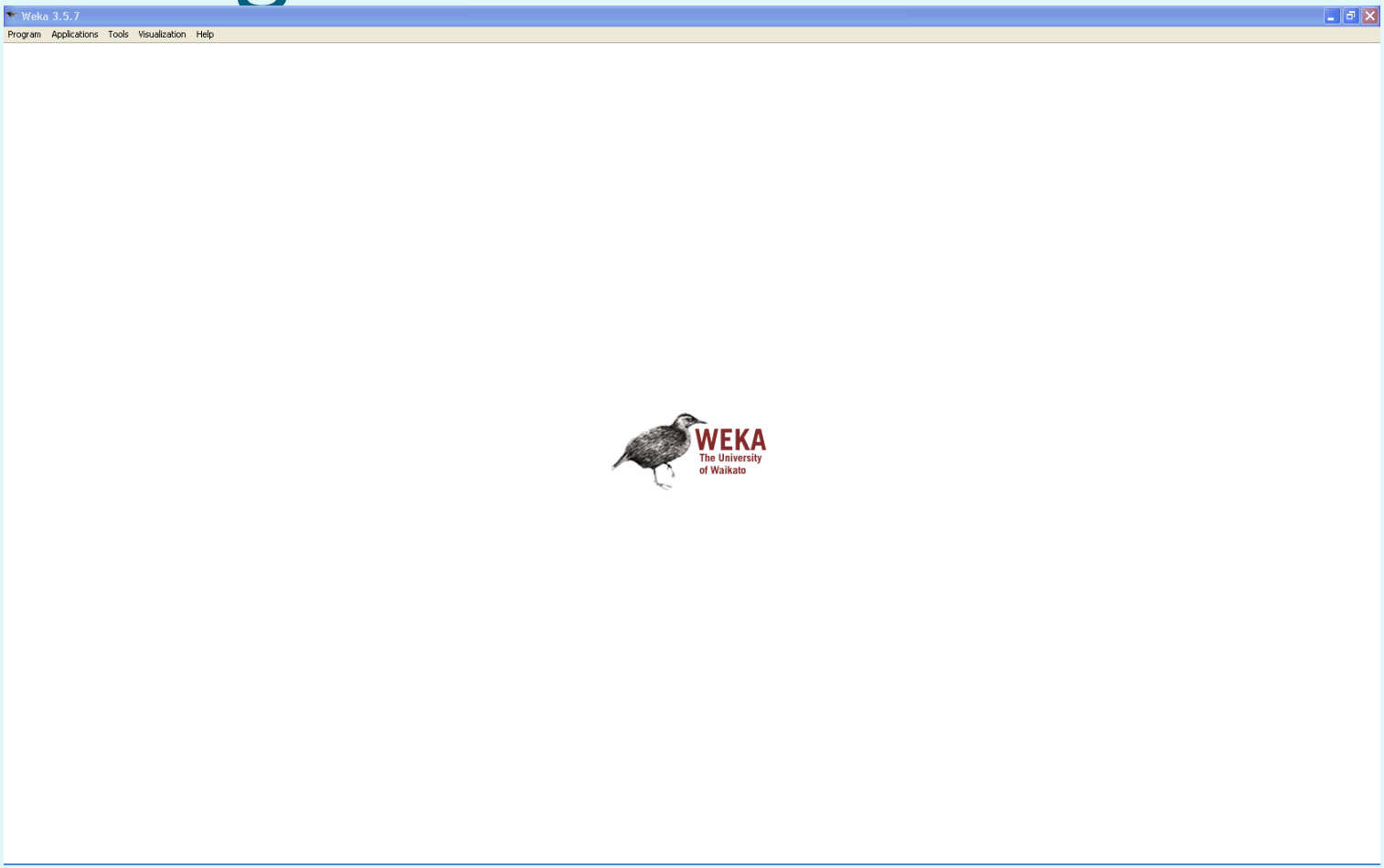
Weka Download

- Website:
  - http://www.cs.waikato.ac.nz/ml/weka/
- Download version 3.5.7 (or similar)
- SourceForge.net

High-Level Description

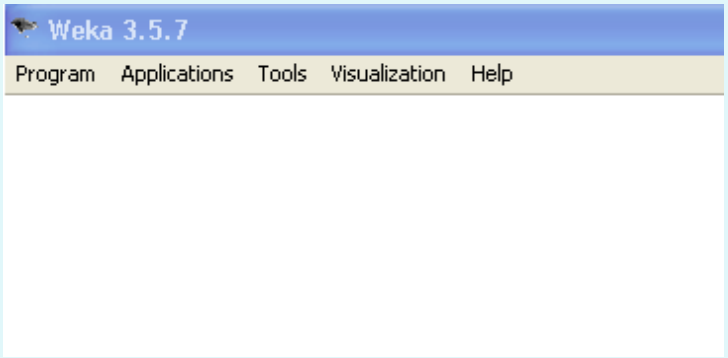
- Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization
- It is also well-suited for developing new machine learning schemes

Getting Started



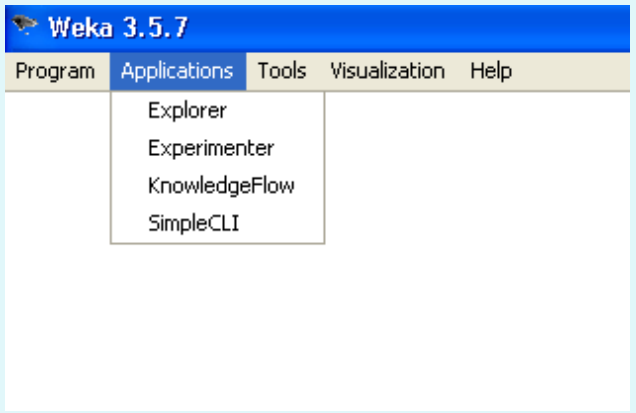
Menu

- Program
- Applications
- Tools
- Visualization
- Help



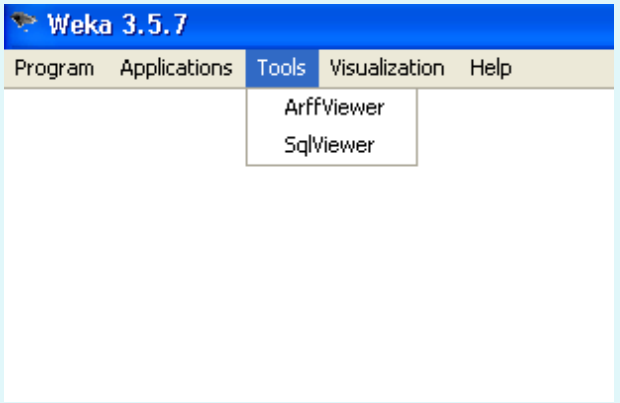
Applications

- Explorer
- Experimenter
- KnowledgeFlow
- SimpleCLI



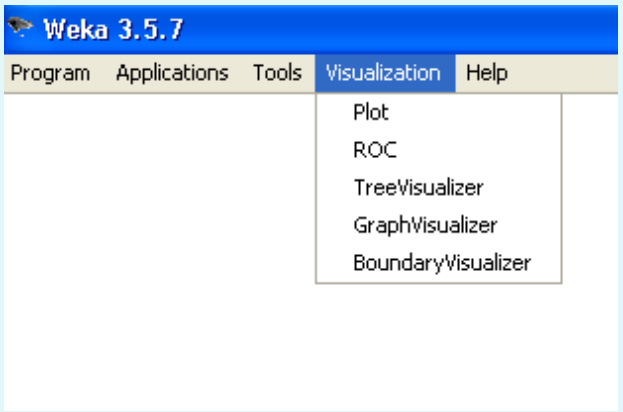
Tools

- ArffViewer
- SqlViewer



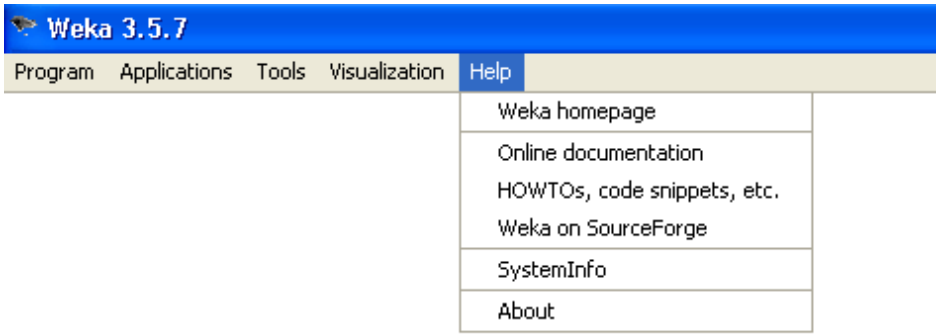
Visualization

- Plot
- ROC
- Tree Visualizer
- Graph Visualizer
- Boundary Visualizer



Help

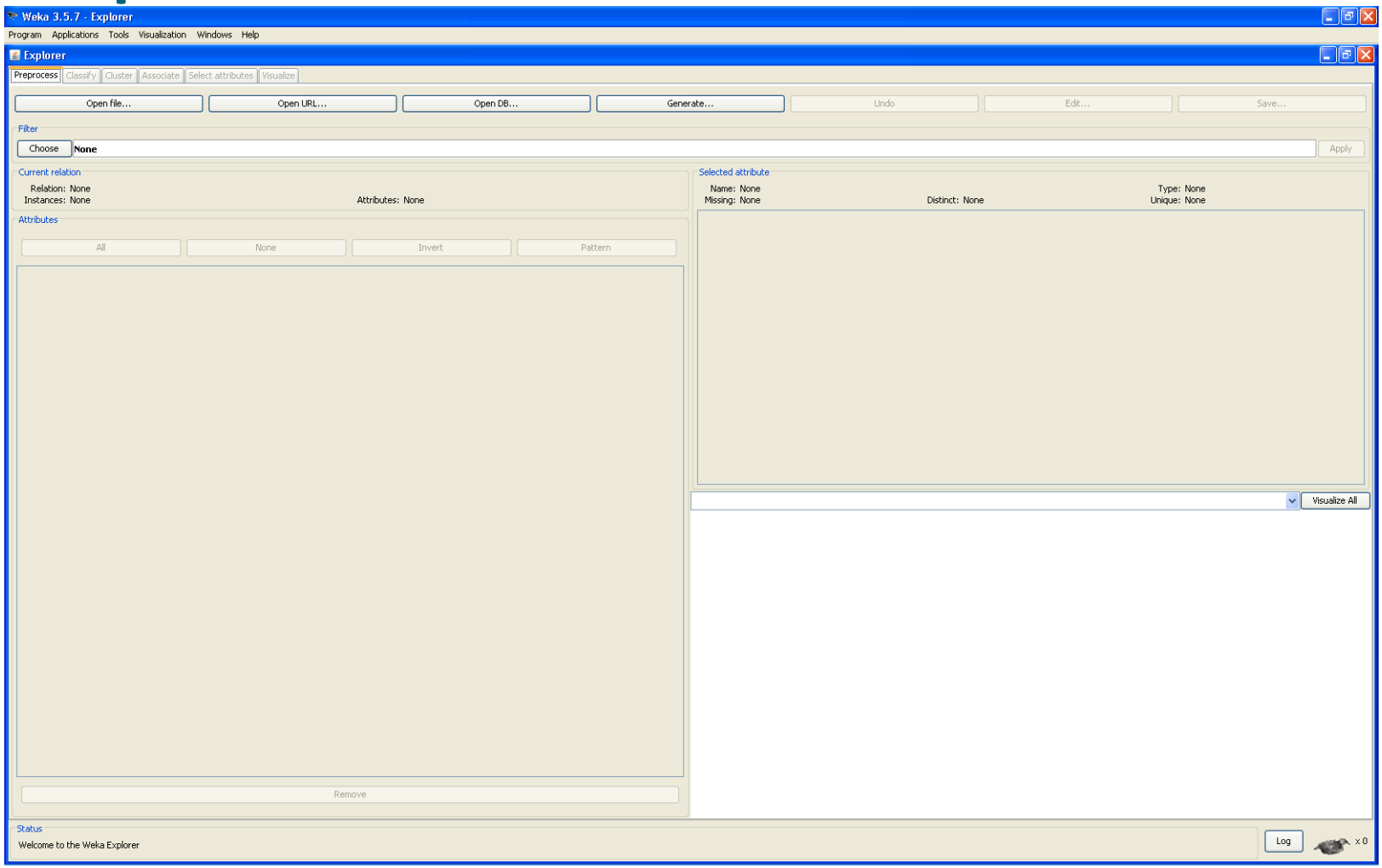
- Weka homepage
- Online documentation
- HOWTos, code snippets
- Weka of SourceForge
- SystemInfo
- About



To start:

- Applications/Explorer

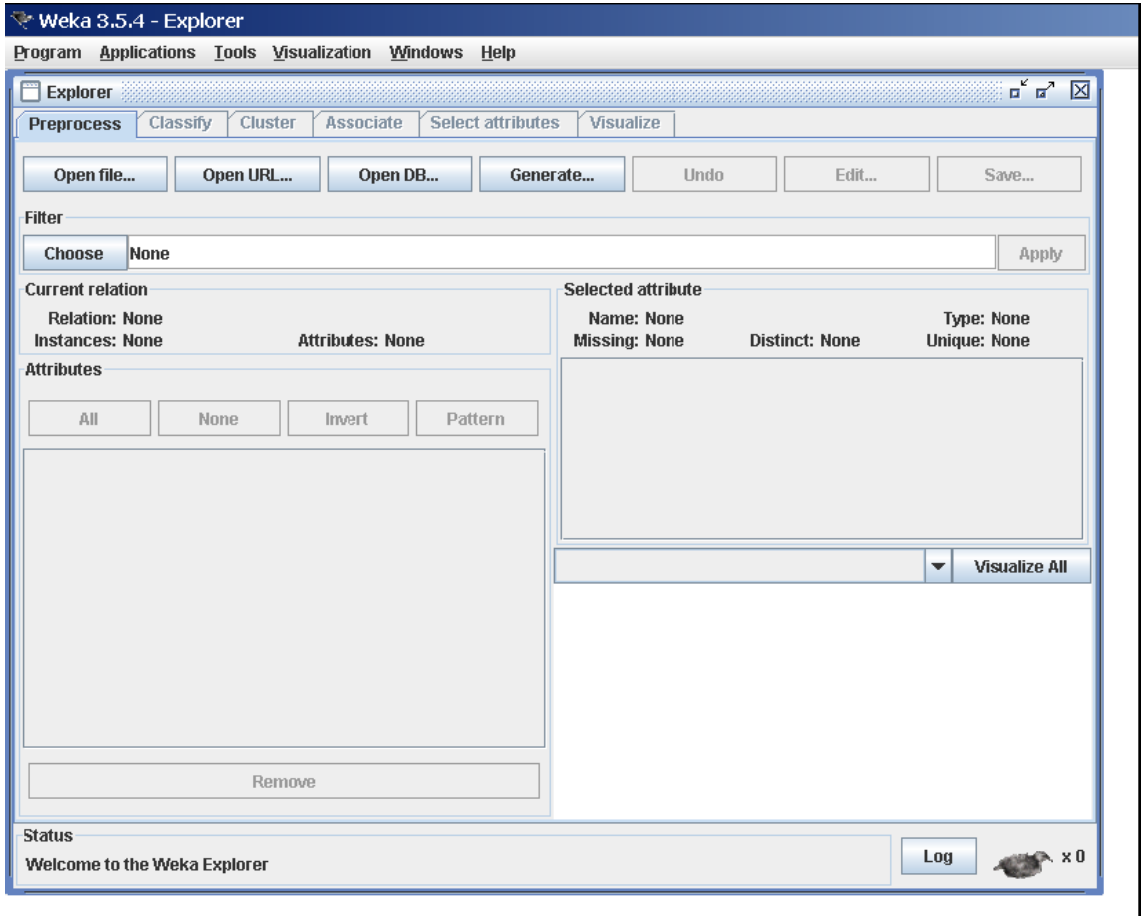
Explorer Screen



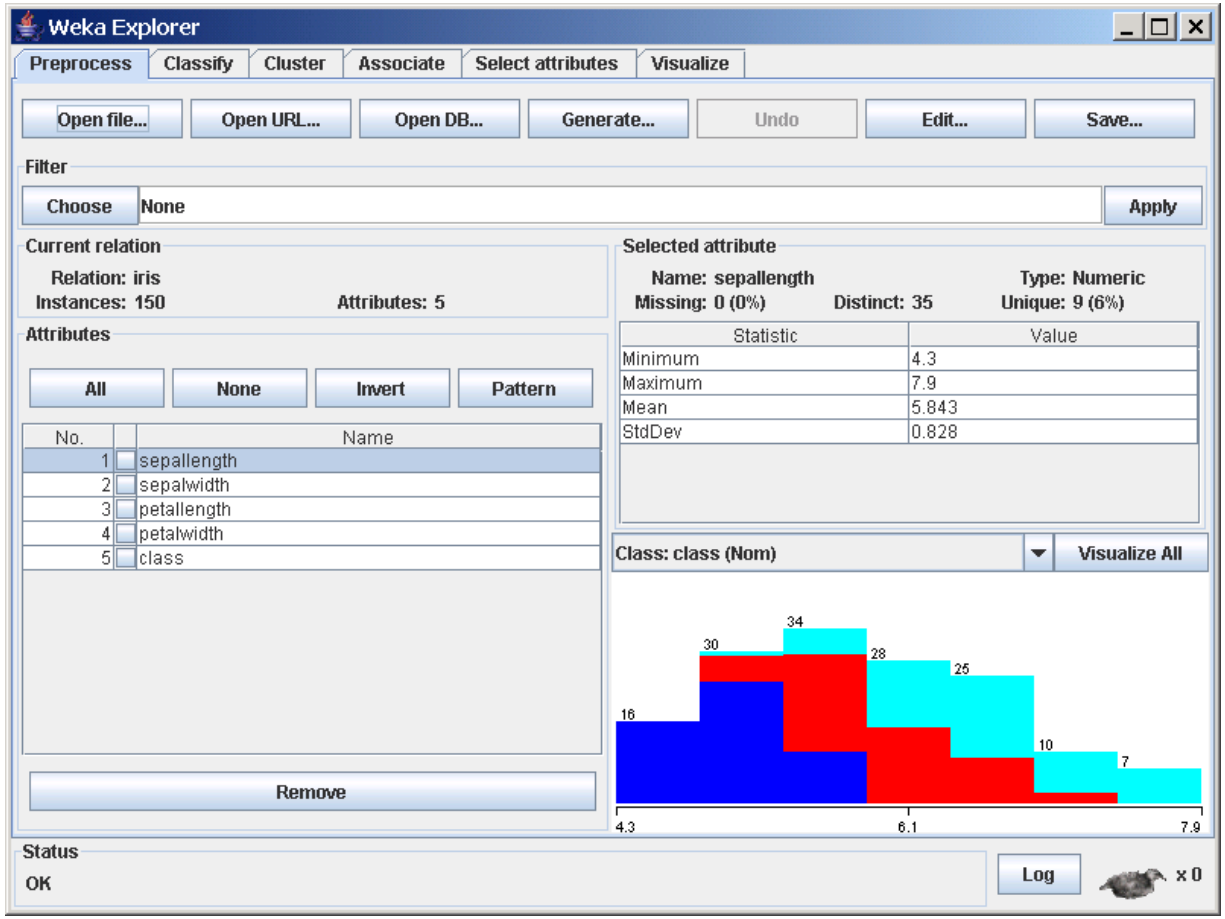
Explorer Panels

- Preprocess Panel
- Classifier Panel
- Cluster Panel
- Associate Panel
- Select Attributes Panel
- Visualize Panel

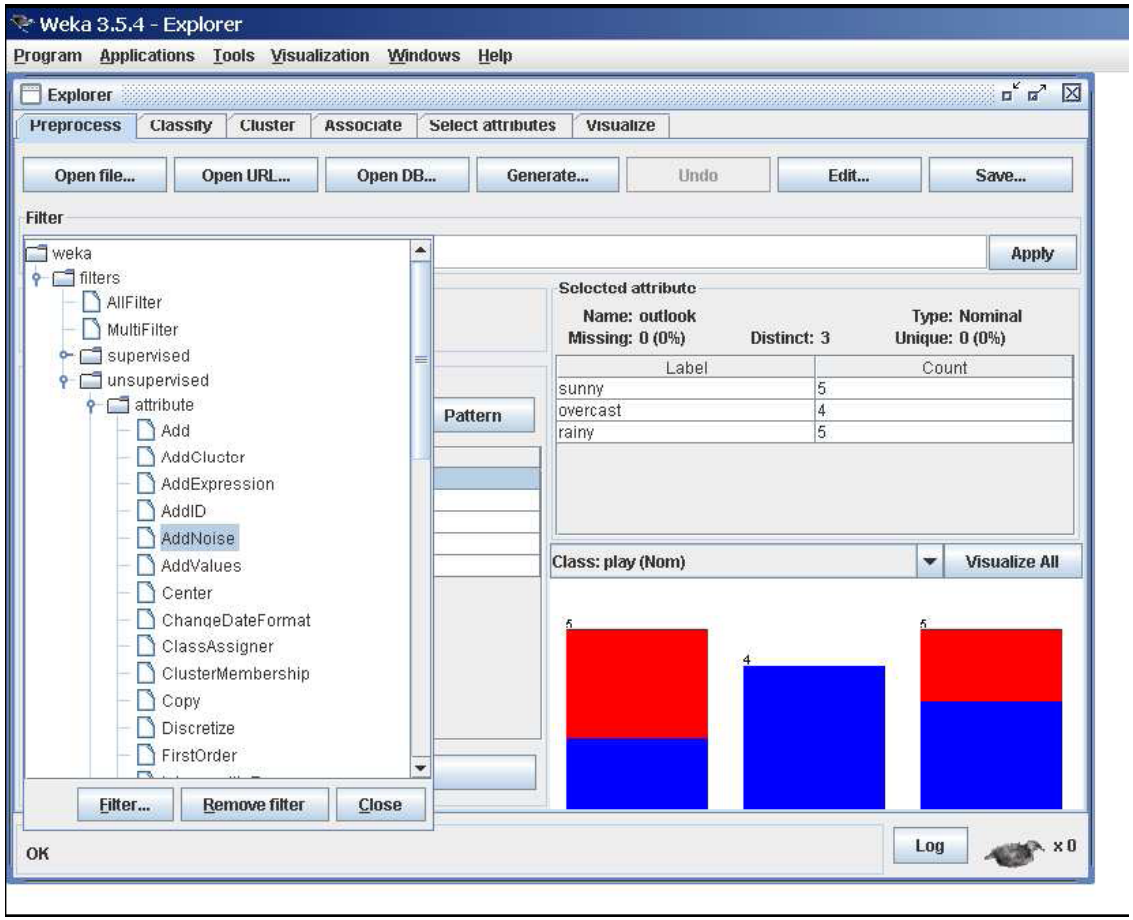
Loading a File



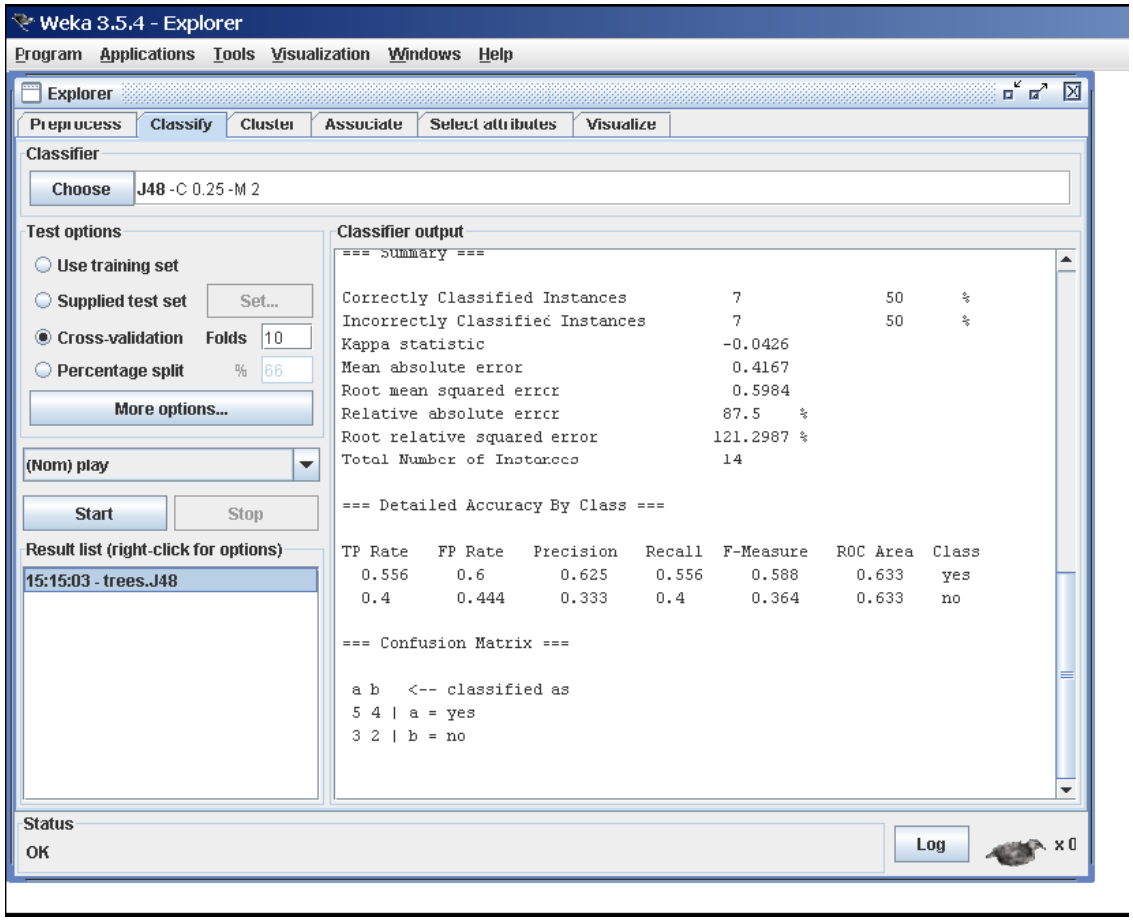
Iris.arff in Preprocess Tab



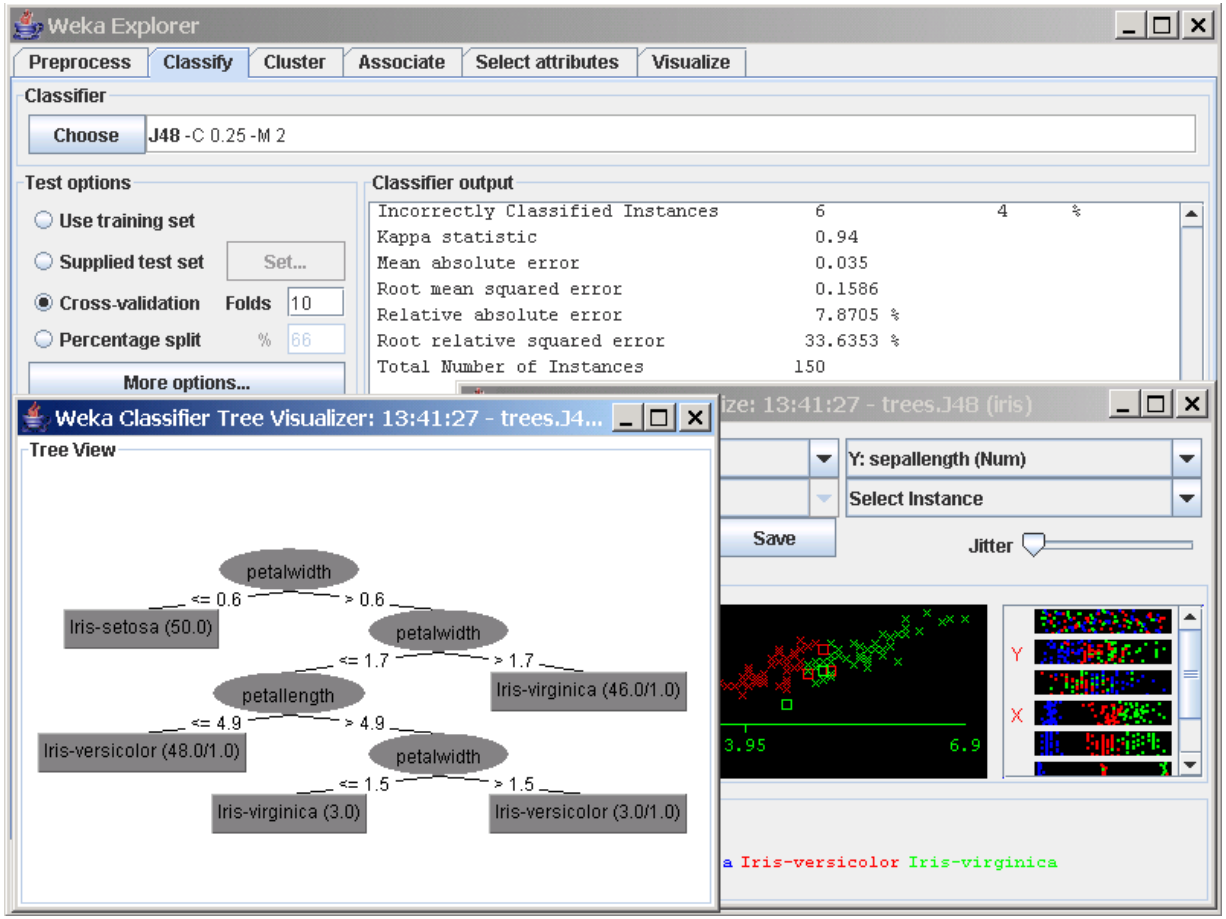
Applying Filters



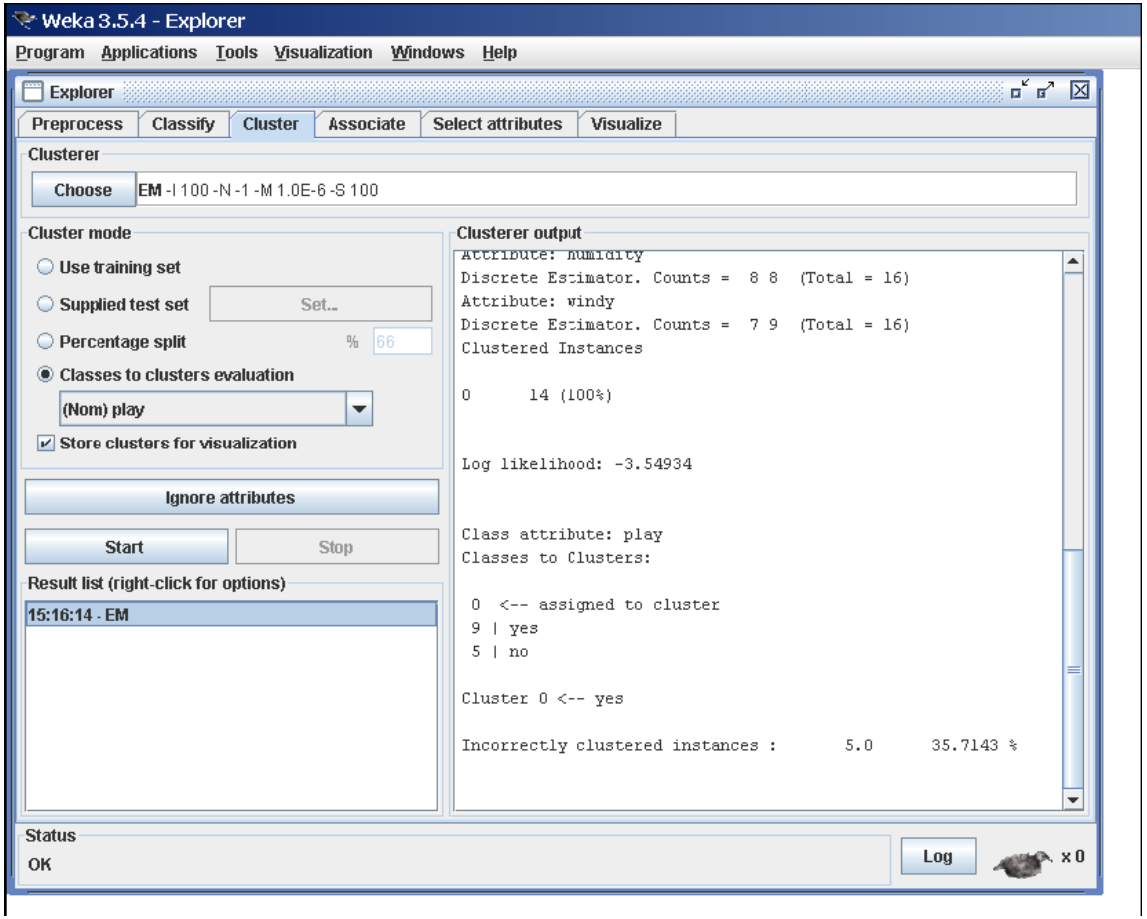
Classify Tab



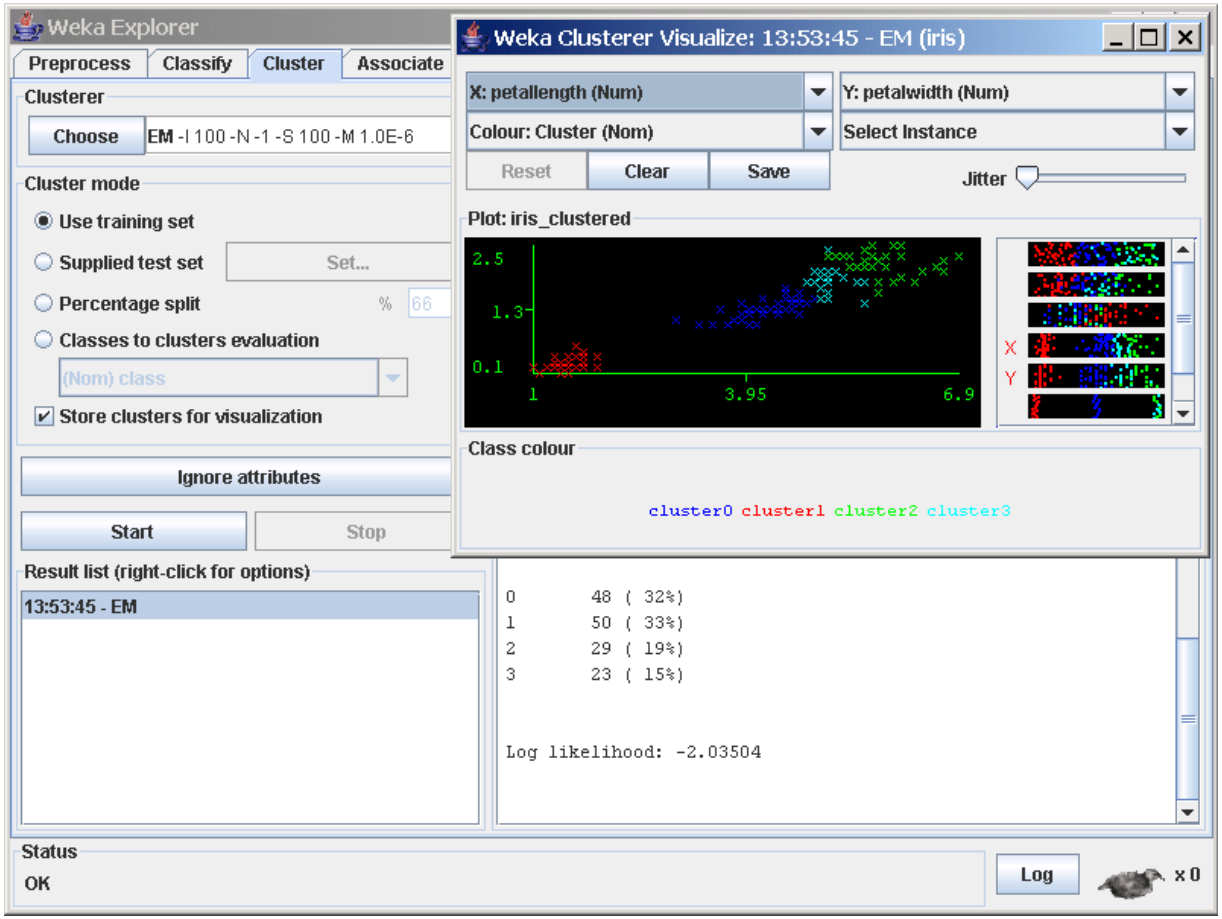
Classify Tab++



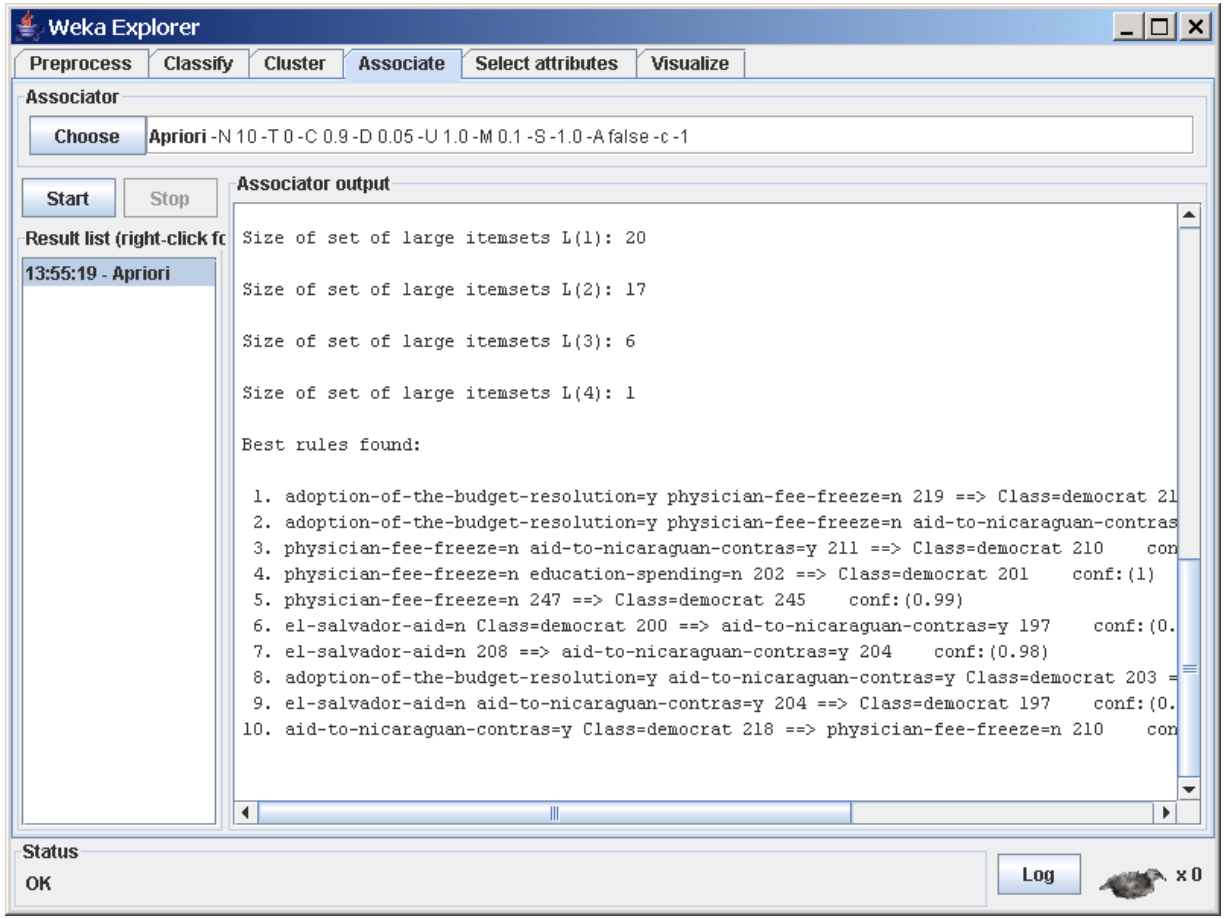
Cluster Tab


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Clusterer
Choose EM -I 100 -N -1 -M 1.0E-6 -S 100
Cluster mode
Use training set
Supplied test set Set...
Percentage split % 66
Classes to clusters evaluation
(Nom) play
Store clusters for visualization
Ignore attributes
Start Stop
Result list (right-click for options)
15:16:14 - EM
Clusterer output
Attribute: humidity
Discrete Estimator. Counts = 8 8 (Total = 16)
Attribute: windy
Discrete Estimator. Counts = 7 9 (Total = 16)
Clustered Instances
0 14 (100%)
Log likelihood: -3.54934
Class attribute: play
Classes to Clusters:
0 <-- assigned to cluster
9 | yes
5 | no
Cluster 0 <-- yes
Incorrectly clustered instances : 5.0 35.7143 %
Status
OK
Log
x 0

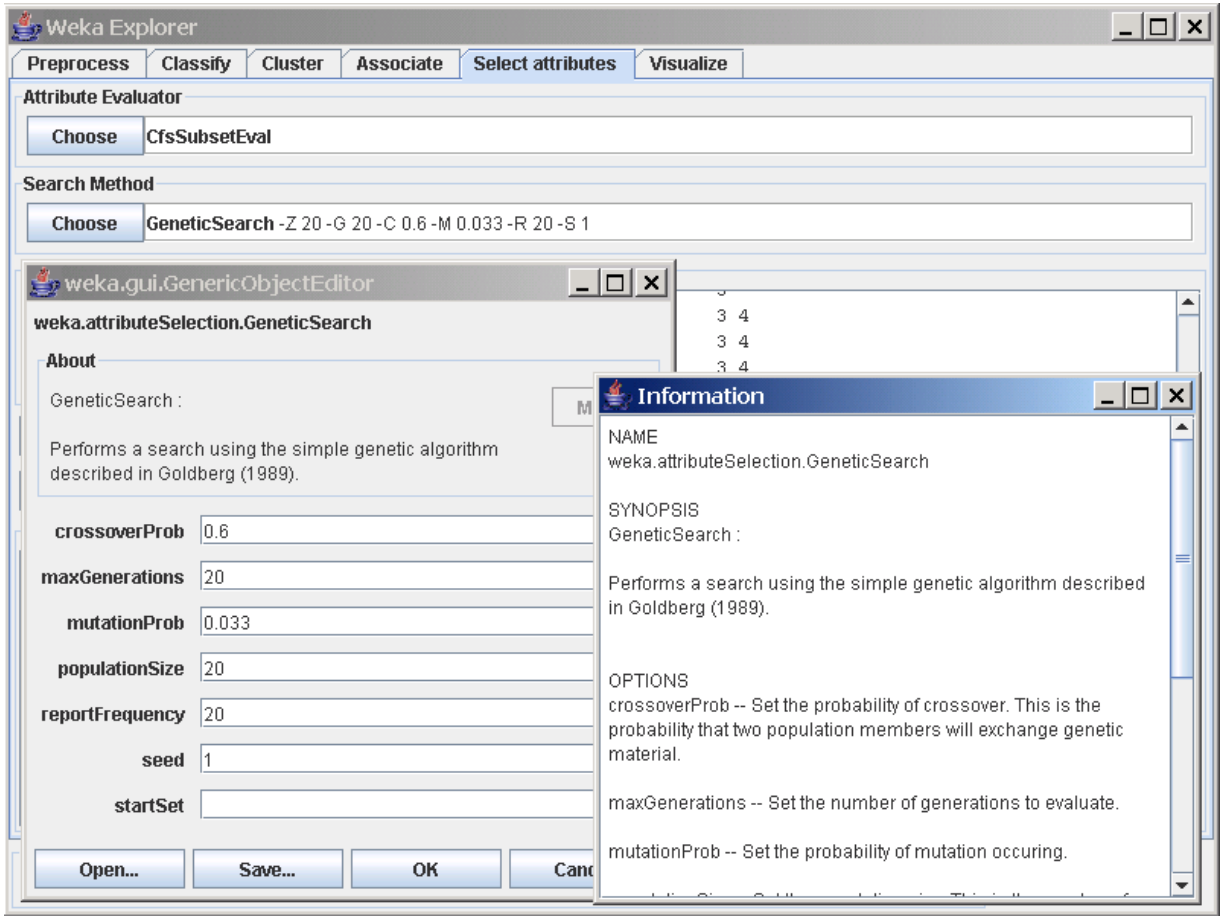
Cluster Tab+


Weka Explorer
Preprocess
Classify
Cluster
Associate
Clusterer
Choose
EM -I 100 -N -1 -S 100 -M 1.0E-6
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
13:53:45 - EM
Weka Clusterer Visualize: 13:53:45 - EM (iris)
X: petallength (Num)
Y: petalwidth (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris_clustered
2.5
1.3
0.1
1
3.95
6.9
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3
0 48 (32%)
1 50 (33%)
2 29 (19%)
3 23 (15%)
Log likelihood: -2.03504
Status
OK
Log
x 0

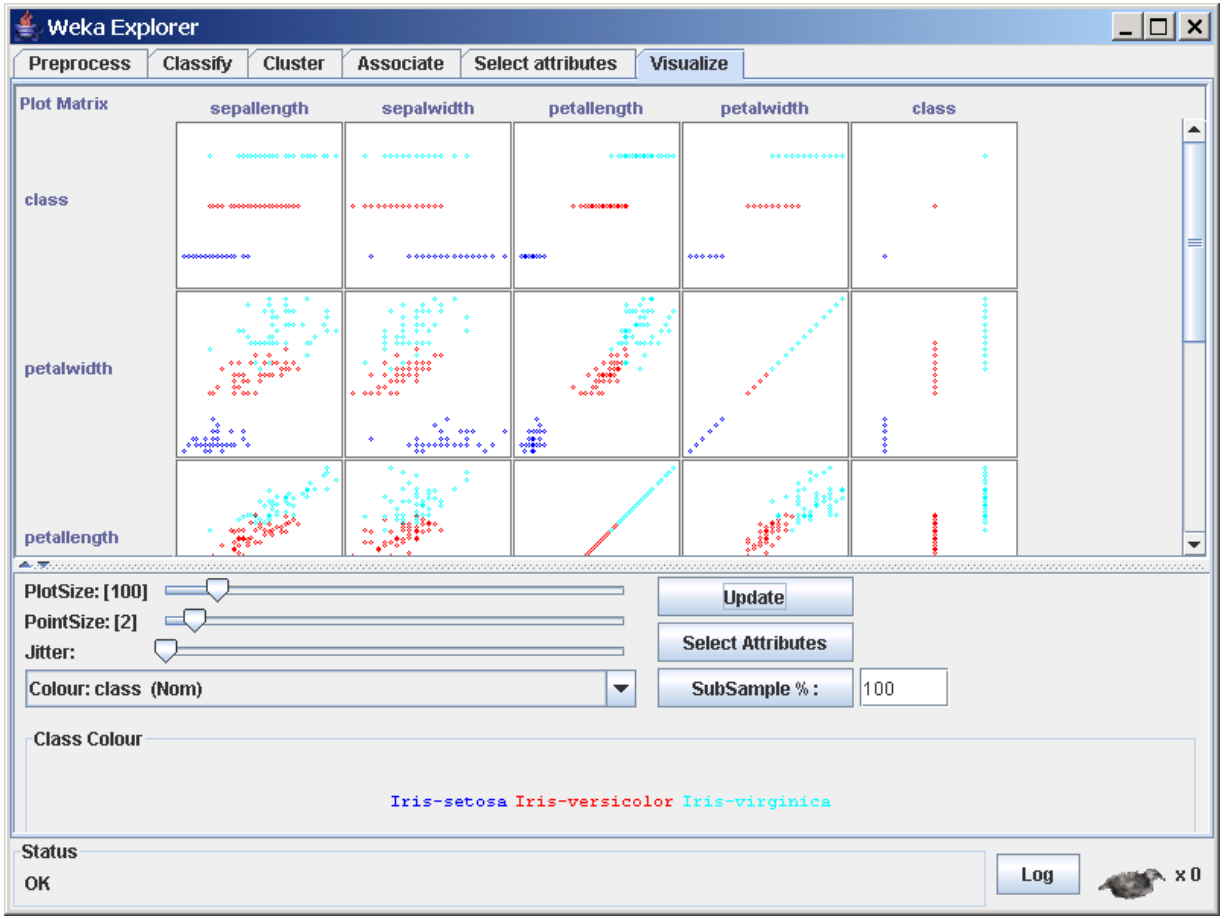
Associate Tab



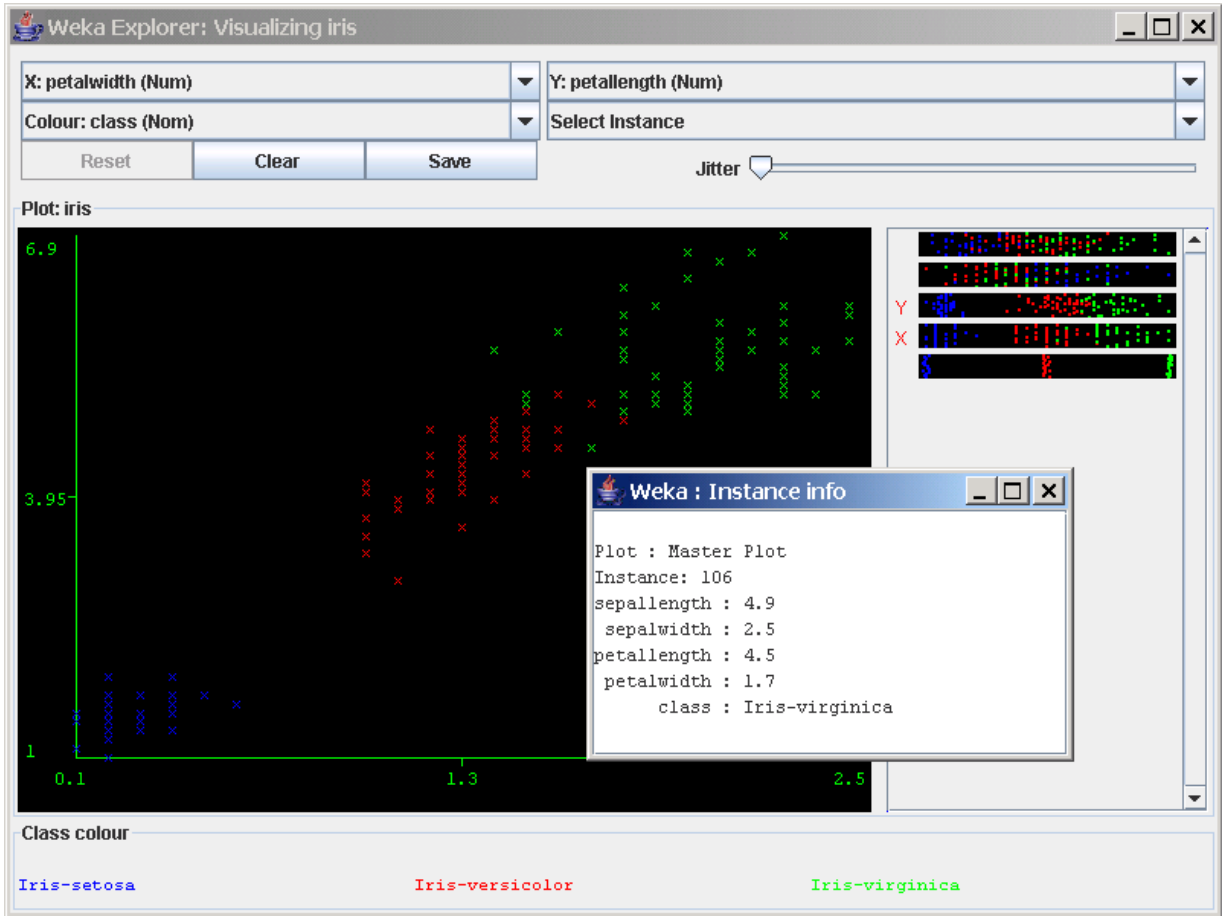
Select Attributes Tab



Visualize Tab



Visualize Tab - Details



Class Resource

- For more details, look up:

  ExplorerGuide-3.5.6.pdf

  under Class Resources

Next

- Lesson 2