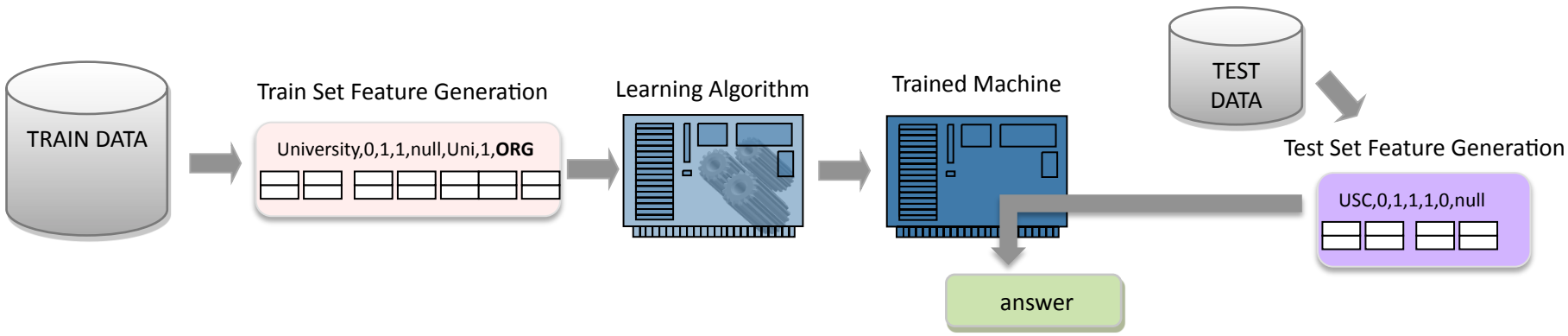
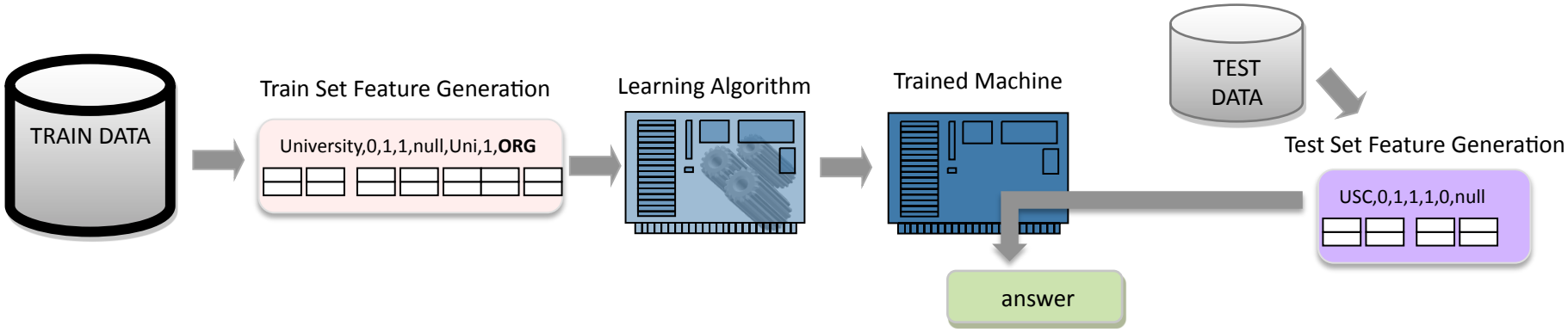
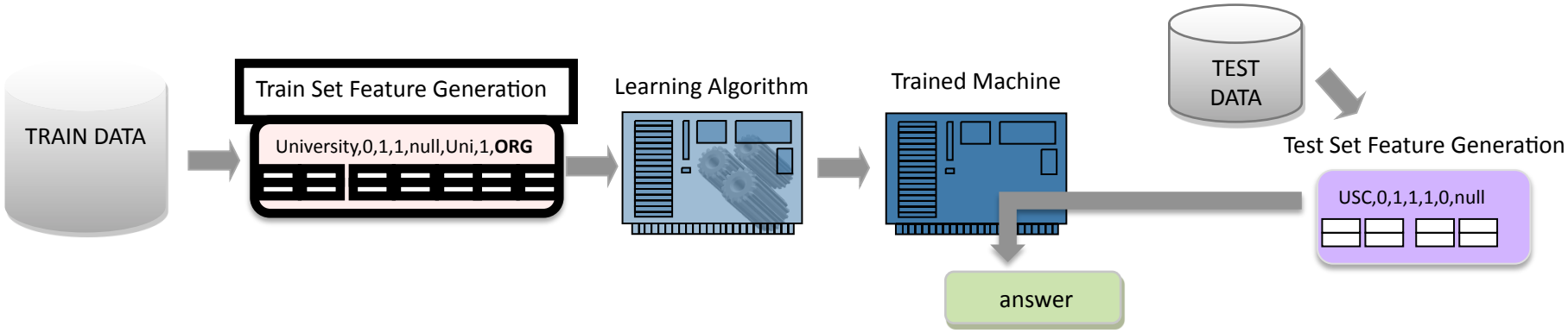
NE System Overview

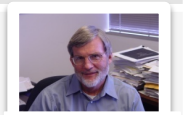

TRAIN DATA
Train Set Feature Generation
University,0,1,1,null,Uni,1,ORG
Learning Algorithm
Trained Machine
TEST DATA
Test Set Feature Generation
USC,0,1,1,1,0,null
answer



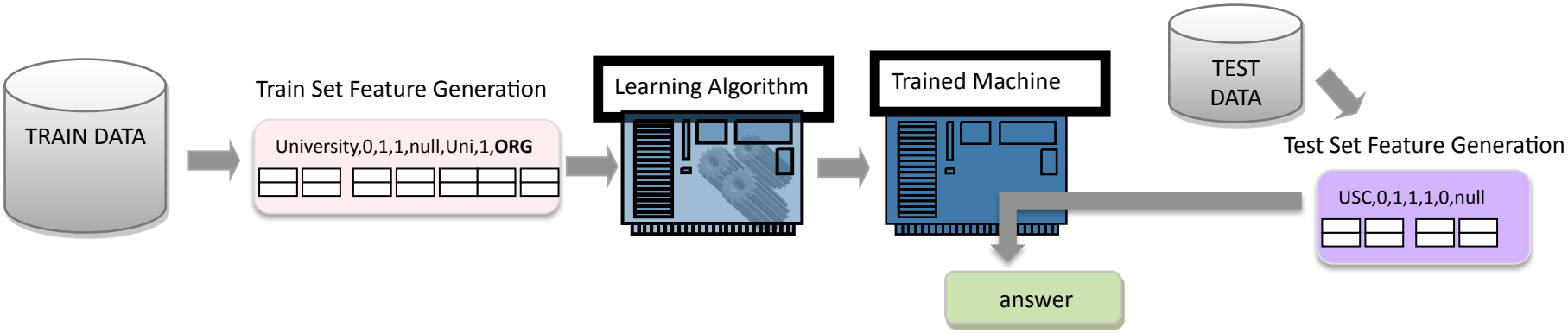
Given

example	class
	PERSON
Microsoft	ORGANIZATION
	PERSON
	LOCATION
ISI Information Sciences Institute	ORGANIZATION
	LOCATION
	OTHER

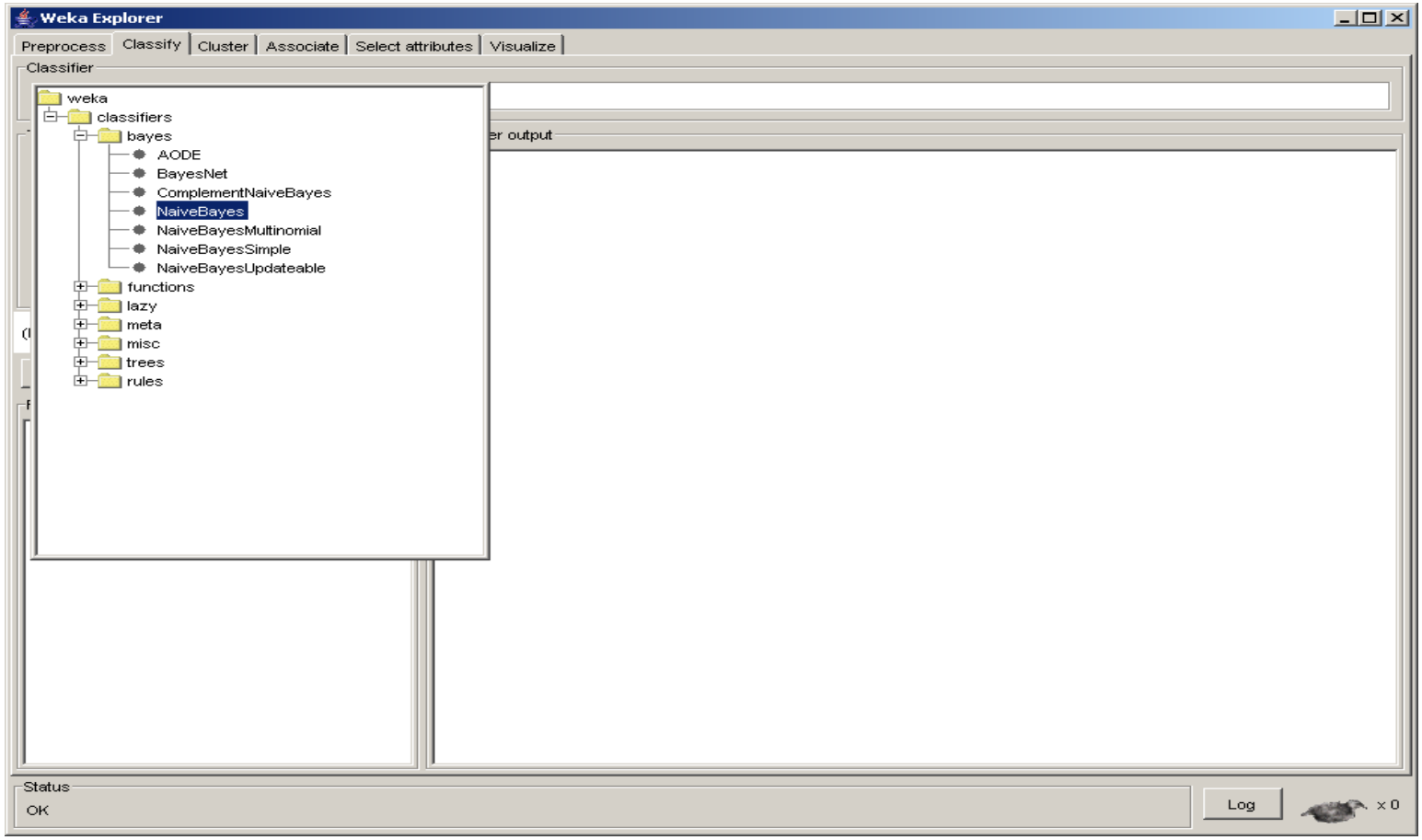


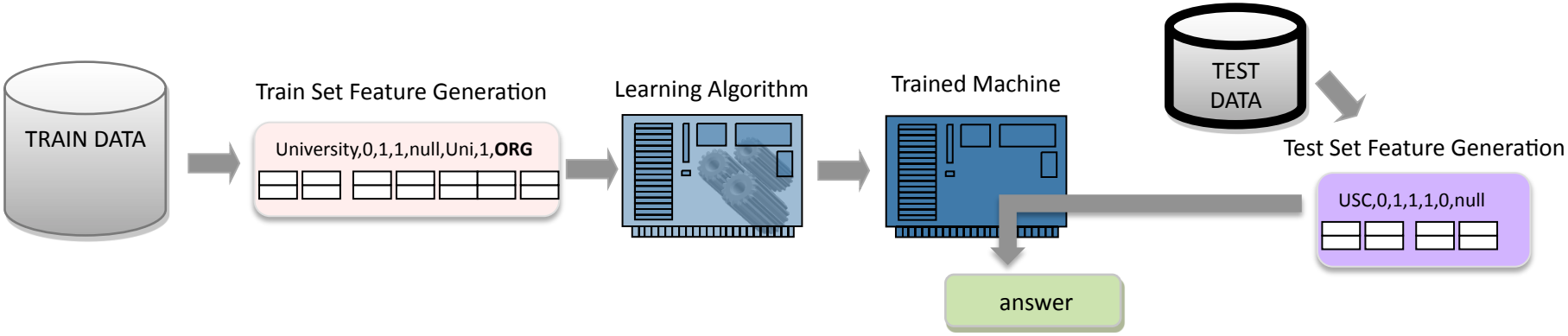
example	Cap.	inDicPer	inDicOrg	inDicLoc	NP	class
	1	1	1	0	1	PERSON
Microsoft	1	0	1	0	0	ORGANIZATION
	1	1	0	0	1	PERSON
	1	0	0	1	1	LOCATION
ISI Information Sciences Institute	1	0	1	0	0	ORGANIZATION
	1	1	0	1	1	LOCATION
	0	0	0	0	0	OTHER

nx**m** matrix, where **n** is number of examples, **m** is number of features+class label

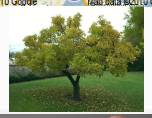


Choose a machine learning classifier from Weka

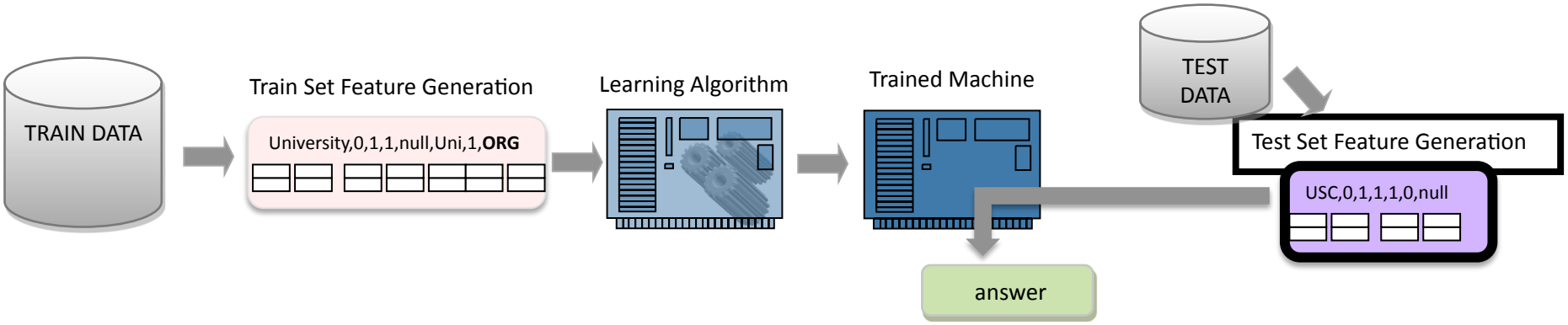




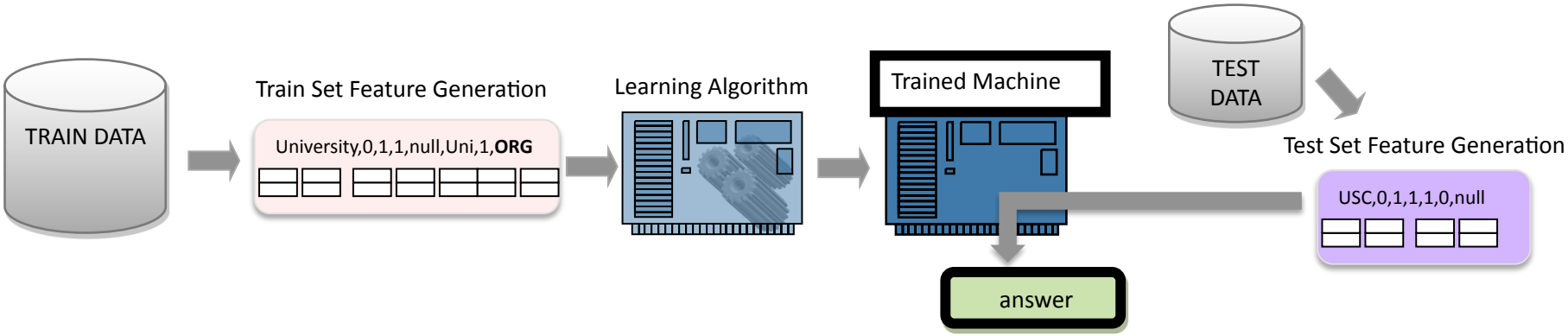
Given

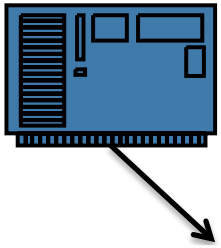
example	class
	?
	?
	?
IBM	?
	?
	?

Note that the class is unknown for the examples of the test data



example	Cap.	inDicPer	inDicOrg	inDicLoc	NP
	1	1	1	0	1
	0	0	0	1	0
	1	1	0	0	1
IBM	1	0	0	1	1
	1	0	1	1	1
	0	1	0	0	0





example	Cap.	inDicPer	inDicOrg	inDicLoc	NP	Predicted Answer		True Answer
	1	1	1	0	1	LOCATION	+	LOCATION
	0	0	0	1	0	LOCATION	−	OTHER
	1	1	0	0	1	PERSON	+	PERSON
	1	0	0	1	1	ORGANIZATION	+	ORGANIZATION
	1	0	1	1	1	OTHER	−	ORGANIZATION
	0	1	0	0	0	OTHER	−	OTHER

$$\text{Precision} = \frac{\#\ \text{correct identified NEs}}{\#\ \text{identified NEs}}$$

$$\text{Recall} = \frac{\#\ \text{correct identified NEs}}{\#\ \text{gold standard data}}$$

NE Feature Generation

Features (1)

- **Contextual**
  - current word W_0
  - words around W_0 in [-3,…,+3] window
- **Part-of-speech tag** (when available)
- **Orthographic**

initial-caps	*all-caps*	*all-digits*
roman-number	*contains-dots*	*contains-hyphen*
acronym	*lonely-initial*	*punctuation-mark*
single-char	*functional-word**	*URL*

- **Word-Type Patterns**

functional	*lowercased*	*quote*
capitalized	*punctuation mark*	*other*

- **Left Predictions**
  - the tag predicted in the current classification for W_{-3}, W_{-2}, W_{-1}

**functional-word is preposition, conjunction, article*

Features (2)

- **Bag-of-Words**
  - words in [-5,…,+5] window

- **Trigger words***
  - for person (*Mr., Miss., Dr., PhD.*)
  - for location (*city, street*)
  - for organization (*Ltd.*, *Co.*)

- **Gazetteers**
  - names of cities, countries, villages, streets
  - names of organizations
  - person first name
  - person surname

* put each type of trigger words and gazetteers in separate files, because you can treat them as separate features

Features (3)

- Length in words of the entity being classified
- Pattern of the entity with regard to the type of constitutent words
- **For each classs**
  - whole NE is in gazetteer
  - any component of the NE appears in gazetteer
- **Suffixes** (length 1 to 4)
- Previous word is an article
- Previous word is a noun
- More idea on features:
  - http://www.cnts.ua.ac.be/conll2002/ner/
  - http://www.cnts.ua.ac.be/conll2003/ner/

Collecting External Resources

Gazetteer Collection Method 1

- Yago contains over 2 million entities (like persons, organizations, cities among others)

- Download Yago from:

  http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

- Extract from the relevant relations all named entities

  Ex.

  – *X* **born in** *Y*, where *X* is a person and *Y* is a location
  – *X* **works for** *Y*, where *X* is a person and *Y* is a person or an organization

Gazetteer Collection Method 2

Madonna (entertainer)

Person

From Wikipedia, the free encyclopedia

Madonna (born **Madonna Louise Ciccone**; August 16, 1958) is an American recording artist, actress and entrepreneur. Born in Bay City, Michigan, and raised in Rochester Hills, Michigan, she moved to New York City in 1977, for a career in modern dance. After performing as a member of the pop musical groups Breakfast Club and Emmy, she released her self-titled debut album, *Madonna*, in 1983 on Sire Records.

A series of hit singles from her next studio albums, *Like a Virgin* (1984) and *True Blue* (1986), gained her global recognition. They established her as a pop icon, for pushing the boundaries of lyrical content in mainstream popular music and imagery in her music videos, which became a fixture on MTV. Her recognition was augmented by the film *Desperately Seeking Susan* (1985) which widely became seen as a Madonna vehicle, despite her not playing the lead. Expanding on the use of religious imagery with *Like a Prayer* (1989), Madonna received positive critical reception for her diverse musical productions, while at the same time was criticised by religious conservatives and the Vatican. In 1992, Madonna founded the Maverick corporation, a joint venture between herself and Time Warner. The same year, she expanded the use of sexually explicit material in her work, beginning with the release of the studio album *Erotica*, followed by the publishing of the coffee table book *Sex*, and starring in the erotic thriller *Body of Evidence*, all of which received negative responses from conservatives and liberals alike.

In 1996, Madonna played the starring role in the film *Evita*, for which she won a Golden Globe Award for Best Actress in Motion Picture Musical or Comedy. Madonna's seventh studio album, *Ray of Light* (1998), became one of her most critically acclaimed, recognized for its lyrical depth. During the 2000s, Madonna released four studio albums – namely *Music* (2000), *American Life* (2003), *Confessions on a Dance Floor* (2005) and *Hard Candy* (2008) – all of which debuted at number one on the *Billboard* 200. Departing from Warner Bros. Records, Madonna signed an unprecedented $120 million dollar contract with Live Nation in 2008.

According to the International Federation of the Phonographic Industry, Madonna has sold more than 200 million albums worldwide.[1] She is ranked by the Recording Industry Association of America as the best-selling female rock artist of the 20th century, and the second top-selling female artist in the United States, behind Barbra Streisand, with 64 million certified albums.[2][3] *Guinness World Records* listed her as the world's most successful female recording artist of all time. In 2008, *Billboard* magazine ranked Madonna at number two, behind only The Beatles, on the "*Billboard* Hot 100 All-Time Top Artists", making her the most successful solo artist in the history of the chart. She was also inducted into the Rock and Roll Hall of Fame in the same year. Considered to be one of the most influential women in contemporary music, Madonna has been known for continually reinventing both her music and image, and for retaining a standard of autonomy within the recording industry. She is recognized as an influence among numerous music artists.

Madonna

Madonna at the premiere of *I Am Because We Are* in 2008.

Background information

Birth name	Madonna Louise Ciccone
Also known as	Madonna Ciccone, Madonna Louise Veronica Ciccone
Born	August 16, 1958 (age 51) Bay City, Michigan, United States
Genres	Pop, rock, dance
Occupations	Singer, songwriter, record producer, dancer, actress, film producer, film director, fashion designer, author, entrepreneur

Gazetteer Collection Method 2

- Step 1: Check if identified NE exists in Wikipedia
- Step 2: Extract the first 2-3 sentences
- Step 3: Pull the nouns matching the expression

  X is Y, Z

  X is Y and Z

- Step 4: Extract the information from the infobox
- Step 5: Verify in WordNet whether the found concepts are hyponyms of person, location, organization

  (*Madonna* is an <u>*artist*</u>, <u>*actress*</u>)

Gazetteer Collection Method 3

- DBpedia contains structured information from Wikipedia

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT * WHERE {
?subject rdf:type <http://dbpedia.org/class/yago/CapitalsInEurope>.
?subject rdfs:label ?label.
?subject rdfs:comment ?abstract.
FILTER (lang(?label) = "en" && lang(?abstract) = "en")
} LIMIT 20
```

Results: Browse Go! Reset

Class	Instance
Place	462,000
Person	364,000
Organization	148,000
Resource (overall)	1,667,000

SPARQL results:

http://wiki.dbpedia.org/Datasets#h18-11

subject	label	abstract
:Andorra_la_Vella	"Andorra la Vella"@en	"Andorra la Vella is the capital of the Co-principality of Andorra, and is located high in the east Pyrenees between France and Spain. It is also the name of the parish that surrounds the capital. The principal industry is tourism, although the country also earns foreign income from being a tax-haven. Furniture and brandies are local products."@en
:Bratislava	"Bratislava"@en	"Bratislava is the capital of Slovakia and, with a population of about 429,000, also the country's largest city. Bratislava is in southwestern Slovakia on both banks of the Danube River. Bordering Austria and Hungary, it is the only national capital that borders two independent countries, Bratislava and Vienna are two of the closest European national capitals to each other, at less than 60 kilometres (37 mi) apart. Bratislava is the political, cultural, and economic centre of Slovakia."@en
:Gibraltar	"Gibraltar"@en	"Gibraltar is a British overseas territory located on the southern end of the Iberian Peninsula at the entrance of the Mediterranean, overlooking the Strait of Gibraltar. The territory itself is a peninsula of 6.843 square kilometres (2.642 sq mi) whose isthmus connects to the north with Spain. The Rock of Gibraltar is the major landmark of the area and gives its name to the densely populated town, home to almost 30,000 Gibraltarians."@en

Other Gazetteer Sources

- The 2000 U.S. Census data

  http://www.rdfabout.com/demo/census/

- Freebase

  http://www.freebase.com/schema/people

- Linked Data Sets

  http://esw.w3.org/DataSetRDFDumps

  …

Patterns

Pattern Extraction

- Collect statistics for patterns containing NEs

  Ex.

  – Jenny_**PER** <u>works</u>_O for_O IBM_**ORG** ._O
  – Sam_**PER** works_O for_O Microsoft_**ORG** ._O
  – Paul_**PER** Adams_**PER** <u>worked</u>_O for_O George_**PER** ._O

  – Jenny_**PER** <u>bought</u>_O an_O organge_O ._O
  – Yahoo!_**ORG** <u>bought</u>_O Overtrue_**ORG** ._O

- Extract verbs to the left and to the right of the NE

  Ex.

  – London_LOC **is**_O **located**_O in_O
  – John_PER **drinks**_O juice_O

Classifier Combination

Majority Voting

- Let C^1 … C^N be the set of classifiers that are induced by training N different learning algorithms L^1… L^N on a data set D consisting of feature vectors.
- Given a new instance, query classifiers C^1 … C^N and assign to the instance the class with the highest count

C^1	C^2	C^3	VOTING
LOC	LOC	PER	LOC
PER	ORG	PER	PER
ORG	LOC	PER	ORG

Expected question by Cris: **can we do weighted voting?**

- Look at weighted voting and voting with probability distribution (Diettrich, 1997)

10-fold Cross Validation

- Data is split into 10 approximately equal partitions
- Each partition is used in turn for testing while the remainder is used for training

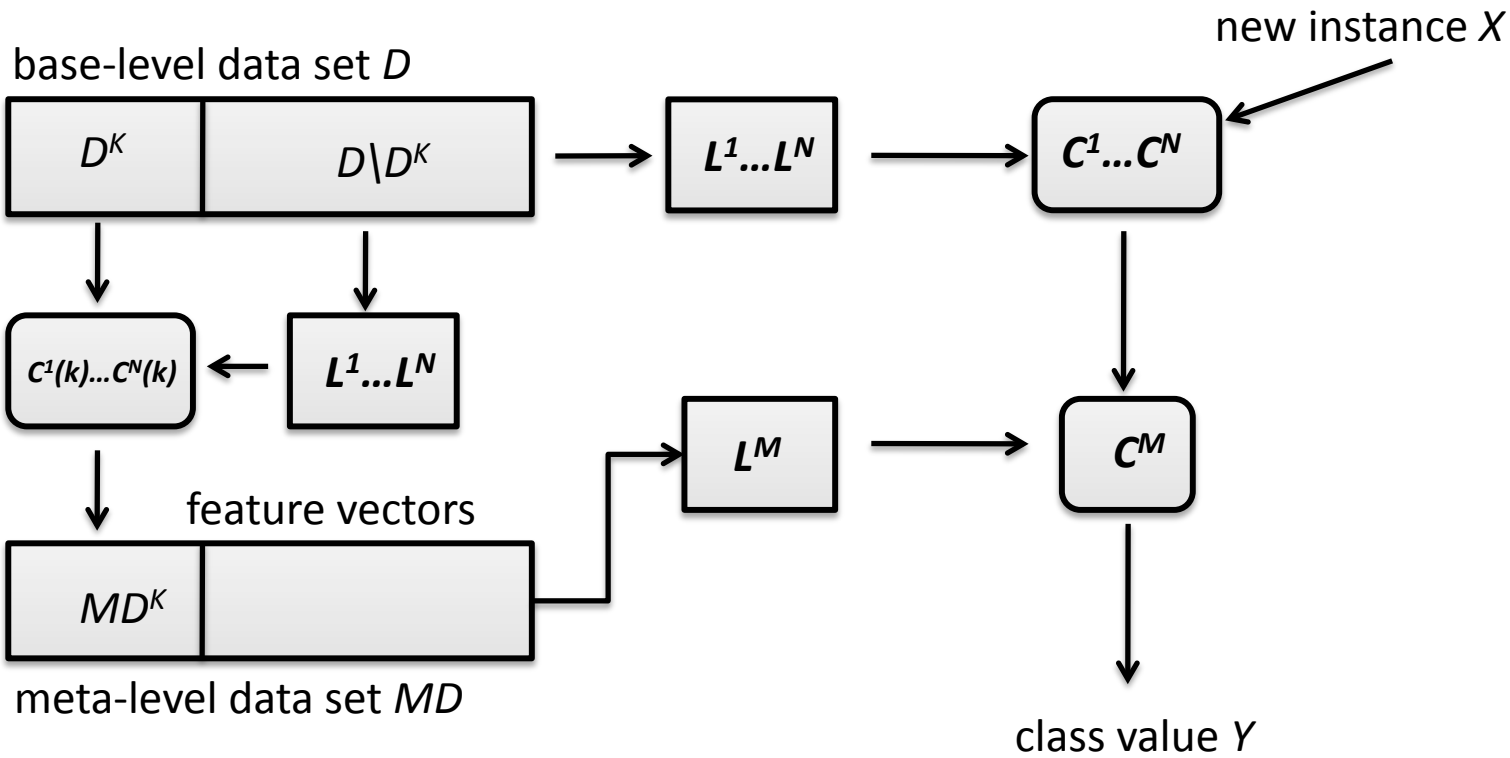
  9/10 of data is used for training

  1/10 of the data is used for testing

- Repeat the whole procedure 10 times
- Overall error rate is equal to the average of the error rates on each partition
- Finally generate the final classifier by learning from all of the data.

Stacking

- Learn a meta (level-1) classifier using the output of base-level (level-0) classifiers estimated via cross-validation



Boosting and Bagging

- Let $C^1 \ldots C^N$ be the set of classifiers that are generated by applying a single learning algorithm to *N* different versions of a given data set, rather than training *N* different algorithms.

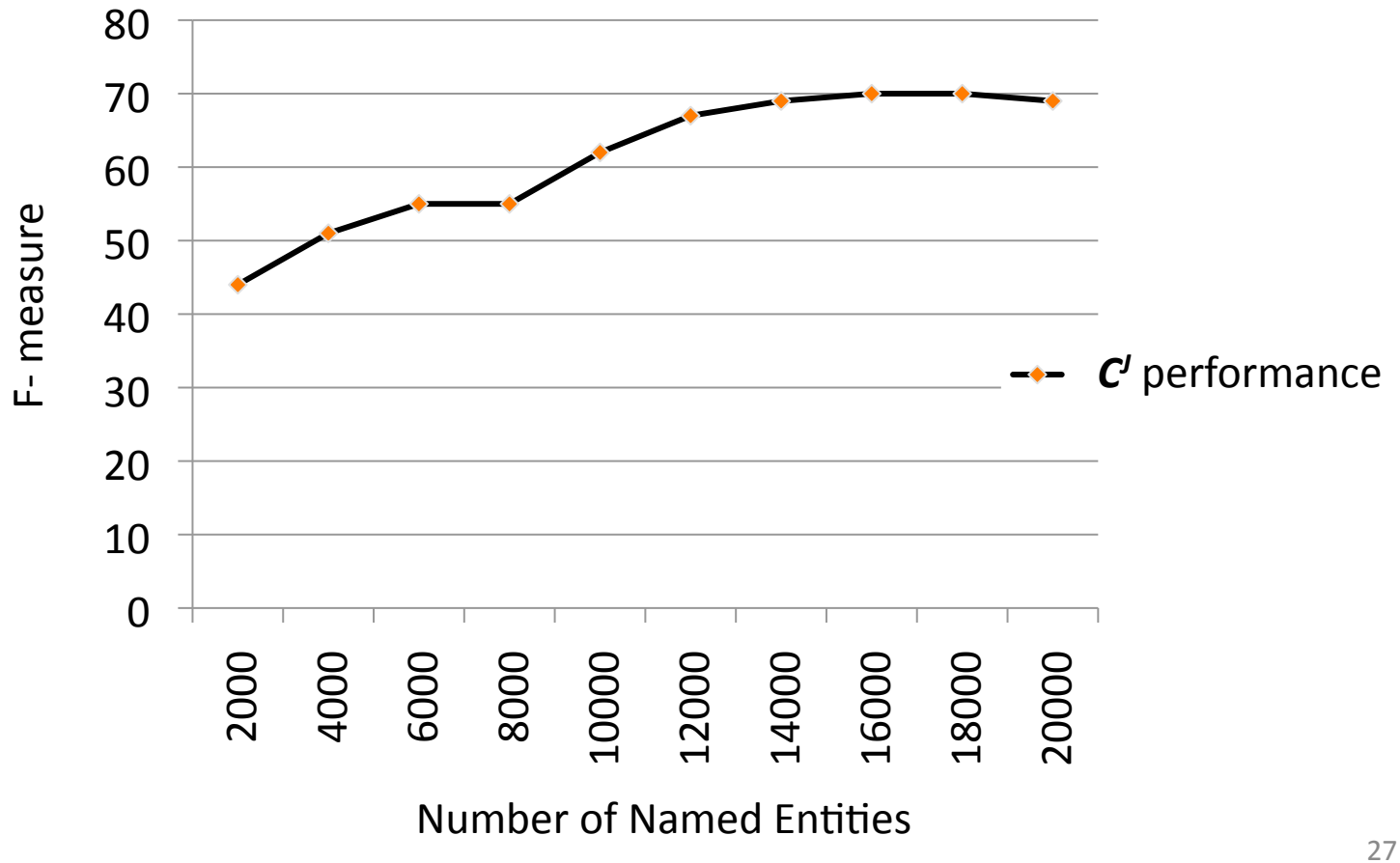
- Typically examples that are misclassified gain weight and examples that are classified correctly lose weight

- Relevant literature
  - Boosting (Freud and Schapire, 1996)
  - Bagging (Breiman, 1996)

Amount of Training Data

Effect of Training Data Size

- Study the effect of the number of examples used during training and the performance of the classifier C^J



Semi-Supervised Learning

Semi-supervised Learning

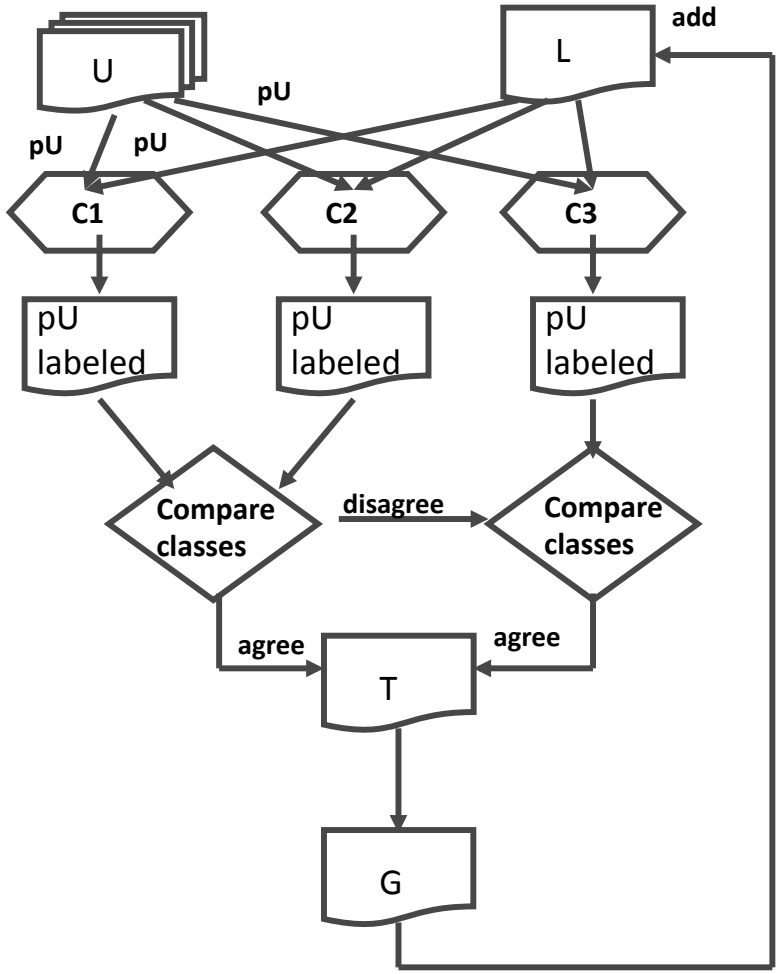
- Learn from a small amount of labeled data and a large amount of unlabeled data
- Methods:
  - Co-training [E.g. Blum & Mitchell, 1998; Collins, 1999; Pierce & Cardie, 2001; Steedman et al. 2003, Mihacea 2004]
  - Self-training [E.g. Banko & Brill, 2001, Nigham]
  - Active learning [E.g. Cohn et al. 1994; Lewis & Catlett 1994; Schohn & Cohn 2000; Shen 2004]
- Tasks which could be resolved: NE recognition, POS tagging, Parsing, …

Co-training / Self-training

- A set L of labeled training examples
- A set U of unlabeled examples
- Classifiers C_i

1. Create a pool of examples U'
   - choose P random examples from U
2. Loop for I iterations
   - Train C_i on L and label U'
   - Select G most confident examples and add to L
     - maintain distribution in L
   - Refill U' with examples from U
     - keep U' at constant size P

Co-training/Self-training



WEKA

Waikato Environment for Knowledge Analysis

Weka: Data Mining Software

- Collection of machine learning algorithms
  - open-source package written in Java
- Used for research, education and application
- Main features:
  - data pre-processing tools
  - learning algorithms
  - evaluation methods
  - graphical inference
  - environment for comparing learning algorithms

Weka: Data Mining Software

- Classification algorithms:
  - decision trees, linear classifiers, SVM, Naive-bayes, kNN
- Prediction algorithms:
  - regression (linear/SVM) , perceptron
- Meta-algorithms:
  - bagging, boosting (AdaBoost)

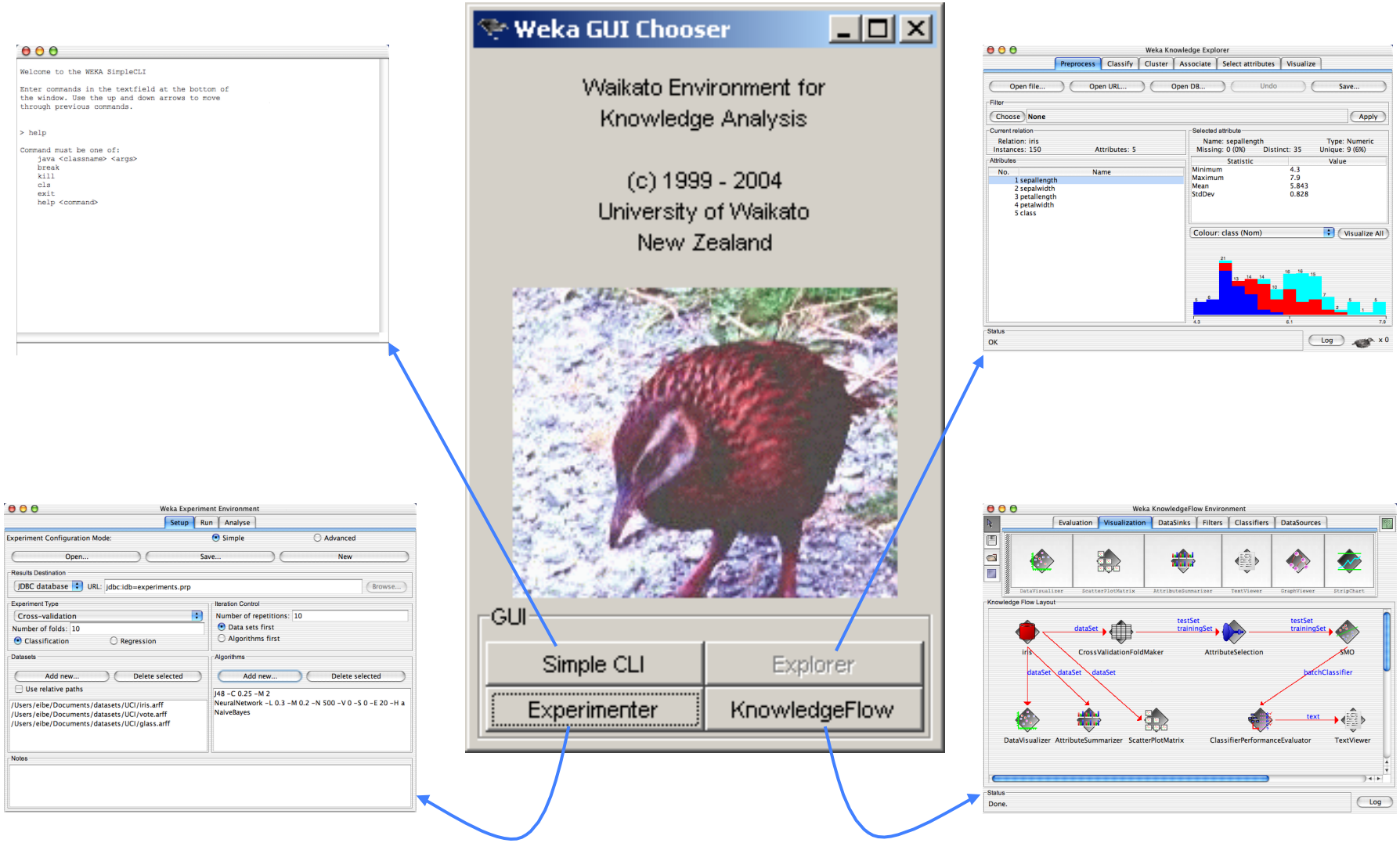
among others

Getting Started

- Install Weka software (on Linux):
  - Download link:
    - http://prdownloads.sourceforge.net/weka/weka-3-6-2.zip
    - Unzip the software
  - Requirement: Java 1.5 (or higher)
  - Invoke Weka command:
    - java -cp weka.jar <*weka-command*>

Weka GUI Chooser

java **-Xmx1000M** -jar weka.jar



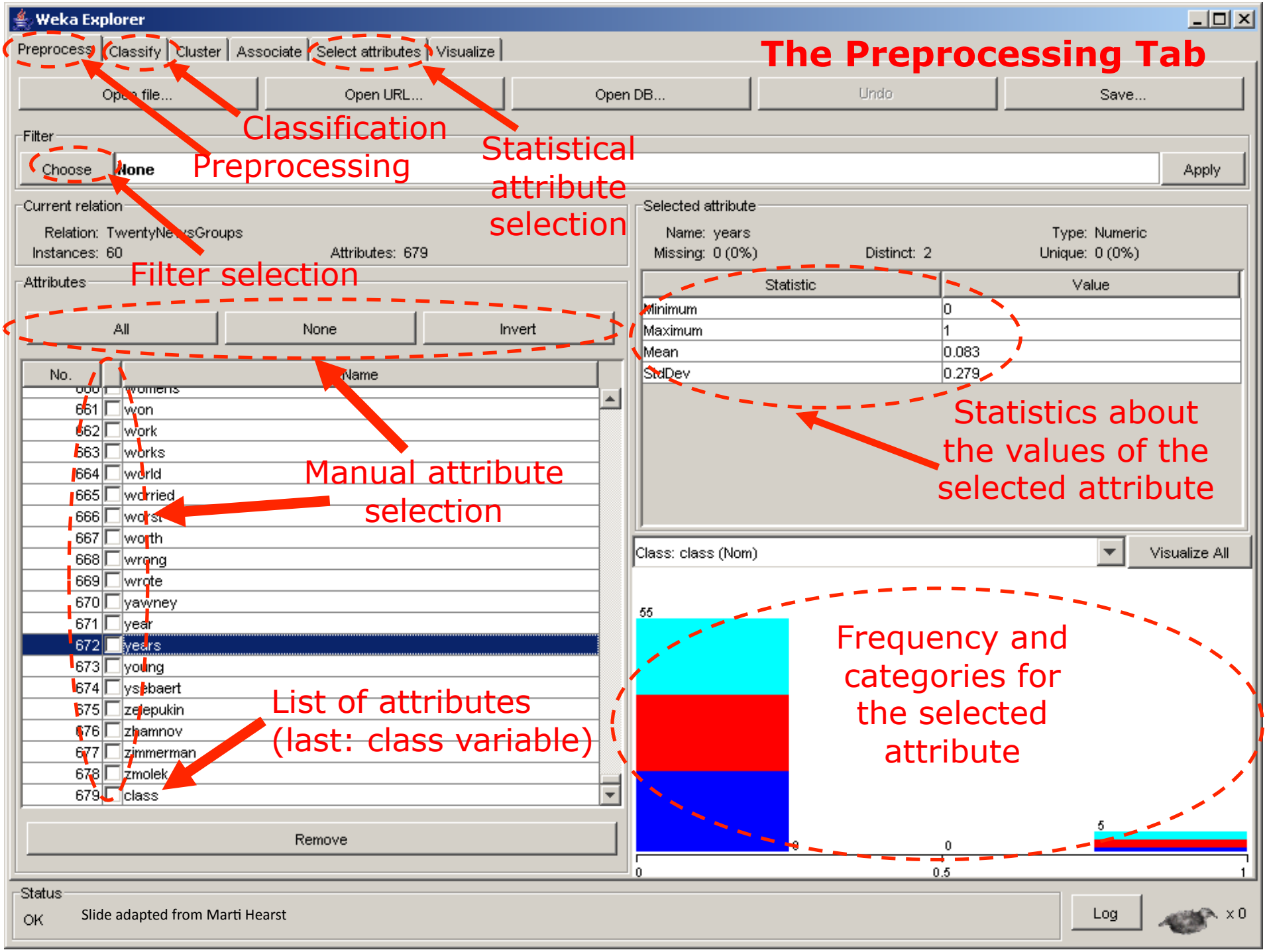
Data file format (.arff)

```
@relation english_named_entity

@attribute position numeric
@attribute pos_tag { NN, NP, VB, DT}
@attribute word_length numeric
@attribute in_gazetteer { no, yes}
@attribute class { PER, LOC, ORG, MISC}

@data
3,DT,3,no,ORG
4,NP,10,yes,ORG
15,NP,6,yes,PER
7, NN,12,?,MISC
...
```

Missing value

Other attribute types:

- String
- Date



Slide adapted from Marti Hearst

The Classification Tab

Weka Explorer

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

Classifier

Choose ZeroR

Choice of classifier

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

Classifier output

Cross-validation: split the data into e.g. 10 folds and 10 times train on 9 folds and test on the remaining one

The attribute whose value is to be predicted from the values of the remaining ones.

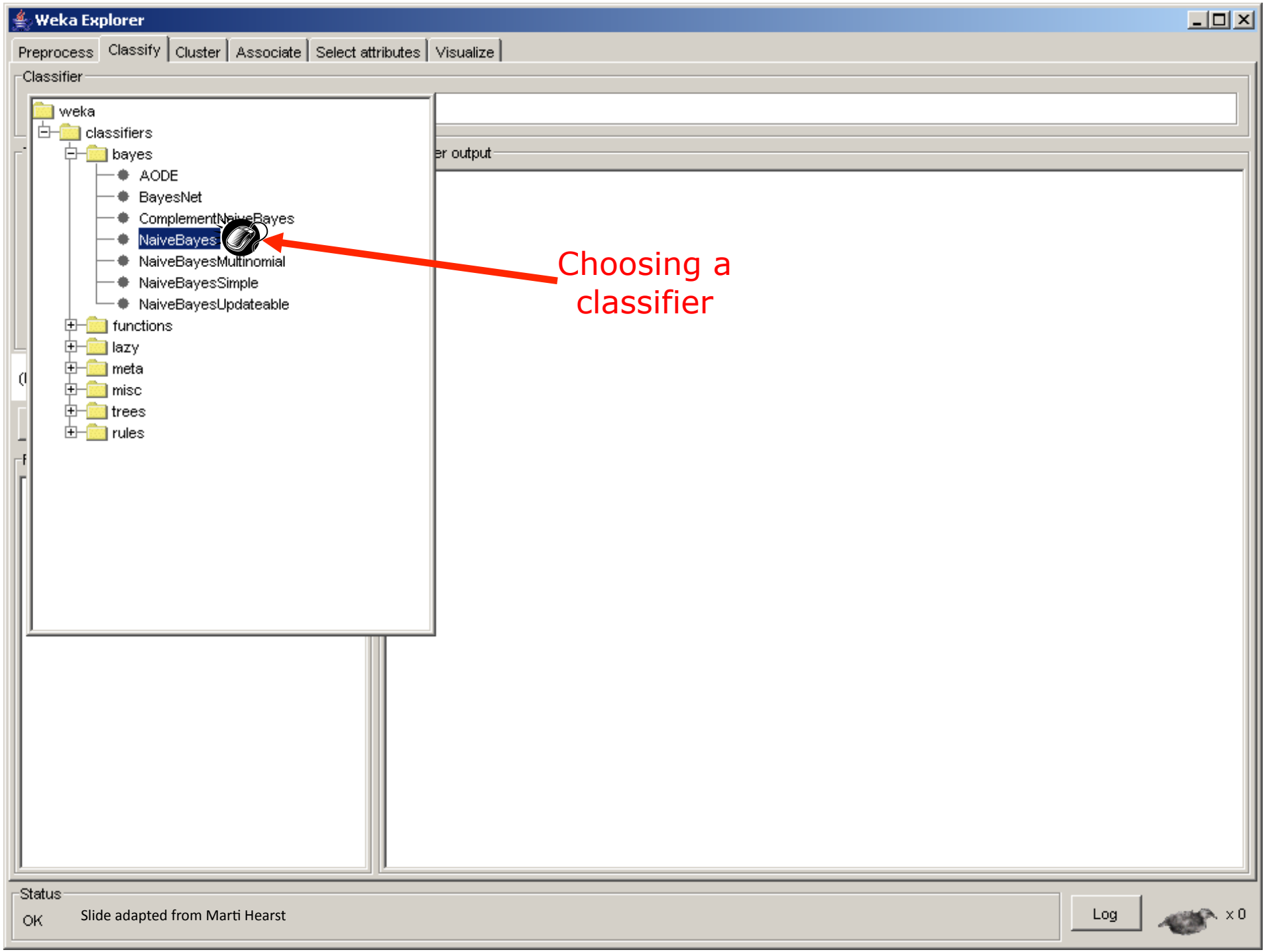
Default is the last attribute.

Status

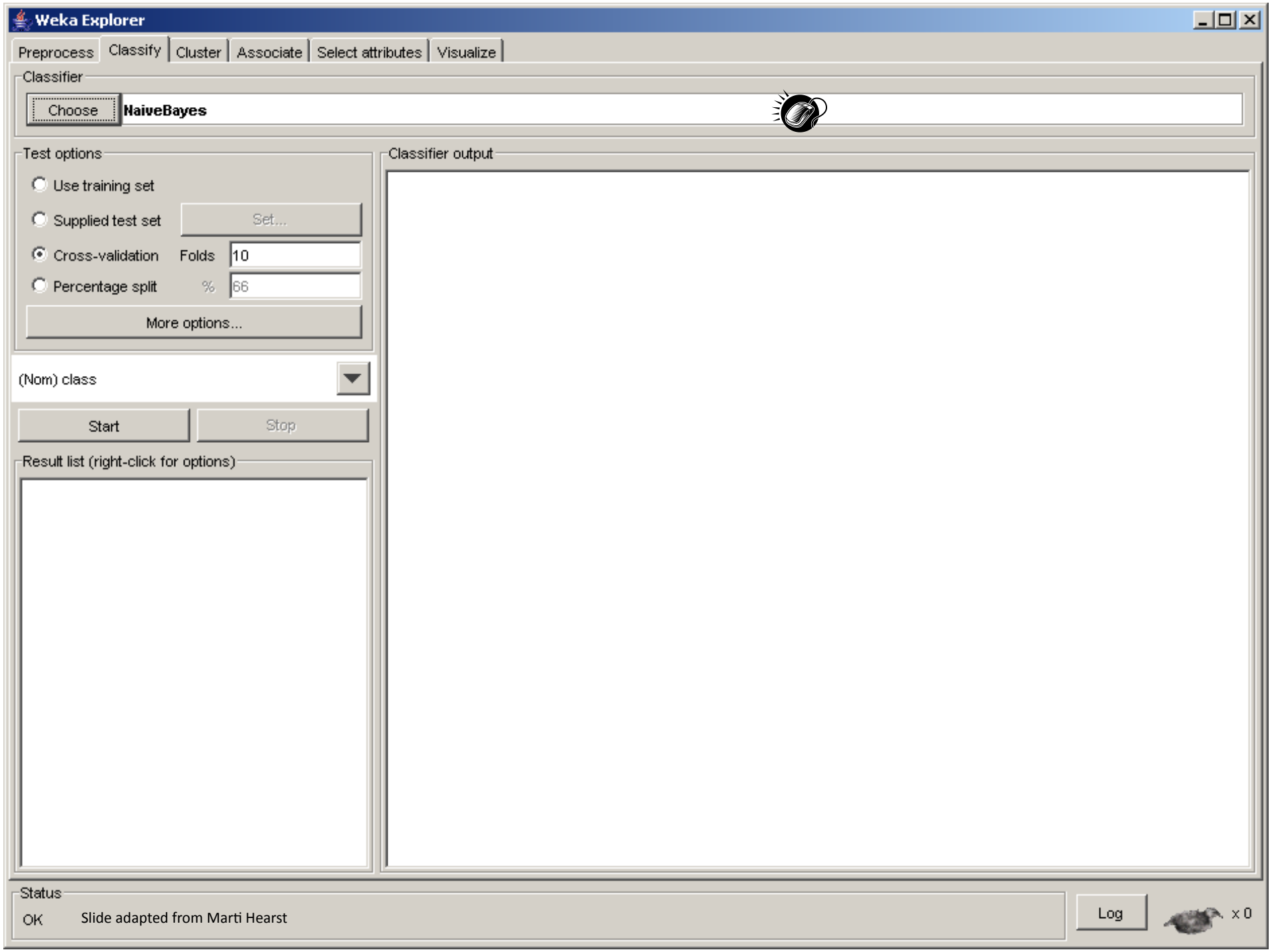
OK

Log x 0

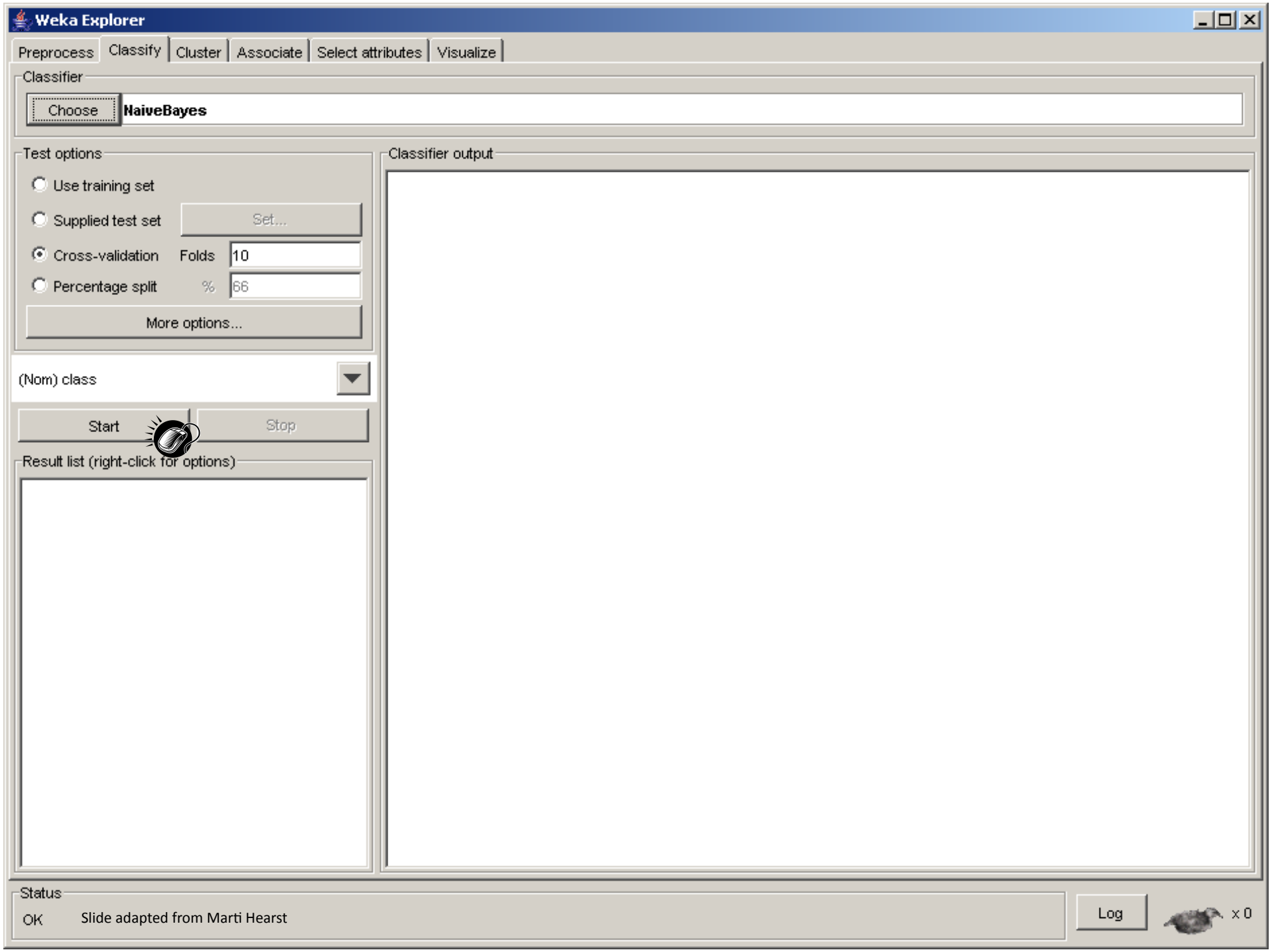
Slide adapted from Marti Hearst



Slide adapted from Marti Hearst



Slide adapted from Marti Hearst



Slide adapted from Marti Hearst

Weka Explorer

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

Classifier

Choose **NaiveBayes**

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) class

Start | Stop

Result list (right click for options)

Starts the classification

09:49:58 - bayes.NaiveBayes

Classifier output

```
Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          41               68.3333 %
Incorrectly Classified Instances        19               31.6667 %
Kappa statistic                          0.525
Mean absolute error                      0.2062
Root mean squared error                  0.4493
Relative absolute error                 46.4007 %
Root relative squared error             95.3122 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.75      0.3       0.556     0.75     0.638     misc.forsale
  0.7       0.025     0.933     0.7      0.8       rec.sport.hockey
  0.6       0.15      0.667     0.6      0.632     comp.graphics

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  1  4 |  a = misc.forsale
  4 14  2 |  b = rec.sport.hockey
  8  0 12 |  c = comp.graphics
```

accuracy

different/easy class

all other numbers can be obtained from it

Status

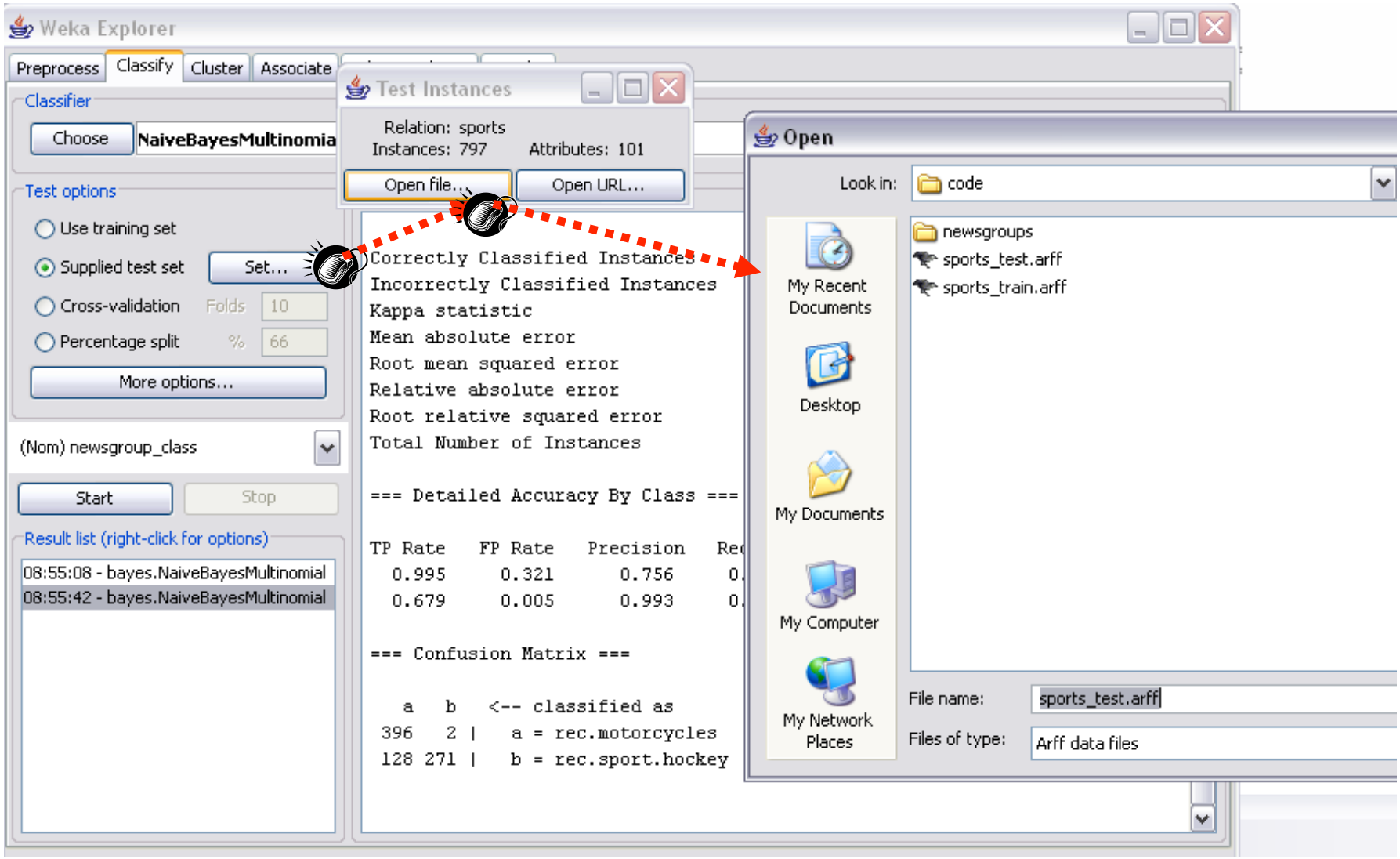
OK

Slide adapted from Marti Hearst

Log

x 0

Running on Test Set



Slide adapted from Marti Hearst

WEKA
Command Line

Weka specifications

- Train classifier on training data and output model
  - java -cp weka.jar <*classifier-function*> –t <*train-file*> -d <*trained-model*>
- Run trained classifier model on test data
  - java -cp weka.jar <*classifier-function*> –T <*test-file*> -l <*trained-model*>
- Specifying parameters:

  -t : training file (.arff)

  -T : test file (.arff)

  -d : output filename (trained classifier model)

  -l : input model (for testing)

  -K : number of nearest neighbors for kNN algorithm

  -h : help (check out other parameter options, etc.)

general parameters

Classifier-specific parameters

Example: *k*NN in Weka

- Train a classifier using 2NN algorithm
  - java -cp weka.jar

weka.classifiers.lazy.IBk	*Classifier-function in weka*
-t data/weather.arff	*Training file*
-K 2	*Algorithm parameter*
-d model.2nn	*Output model name*

- Run the trained classifier on test data
  - java -cp weka.jar

weka.classifiers.lazy.IBk	*Classifier-function in weka*
-T data/weather.arff	*Test file*
-l model.2nn	*Input model name*

Sample Weka output

=== Error on test data ===

Correctly Classified Instances	**13**	**92.8571 %**
Incorrectly Classified Instances	**1**	**7.1429 %**
Kappa statistic	**0.8372**	
Mean absolute error	**0.1333**	
Root mean squared error	**0.2333**	
Total Number of Instances	**14**	

More detailed output

- Classification labels for each instance (use "–p 1" option)
  - java -cp weka.jar weka.classifiers.lazy.Ibk -T data/weather.arff -l model.2nn -p 1

=== Predictions on test data ===

inst#	actual	predicted	error	prediction (outlook)
1	2:no	2:no		0.967 (sunny)
2	2:no	1:yes	+	0.5 (sunny)
3	1:yes	1:yes		0.967 (overcast)
4	1:yes	1:yes		0.967 (rainy)
5	1:yes	1:yes		0.967 (rainy)
6	2:no	2:no		0.967 (rainy)
7	1:yes	1:yes		0.967 (overcast)
8	2:no	2:no		0.967 (sunny)
9	1:yes	1:yes		0.5 (sunny)
10	1:yes	1:yes		0.967 (rainy)
11	1:yes	1:yes		0.5 (sunny)
12	1:yes	1:yes		0.967 (overcast)
13	1:yes	1:yes		0.967 (overcast)
14	2:no	2:no		0.967 (rainy)

Weka classification functions

- kNN: weka.classifiers.lazy.Ibk
- Decision trees: weka.classifiers.trees.J48
- Naïve Bayes: weka.classifiers.bayes.NaiveBayes
- AdaBoost: weka.classifiers.meta.AdaBoostM1

Additional Information

- General documentation:

  http://www.cs.waikato.ac.nz/ml/weka/

  http://prdownloads.sourceforge.net/weka/weka.ppt

- Command line doc:

  http://weka.wikispaces.com/Primer

Remember to send an e-mail to

kozareva@isi.edu

with subject CS544 homework

to obtain the train and development data for Assignment 1