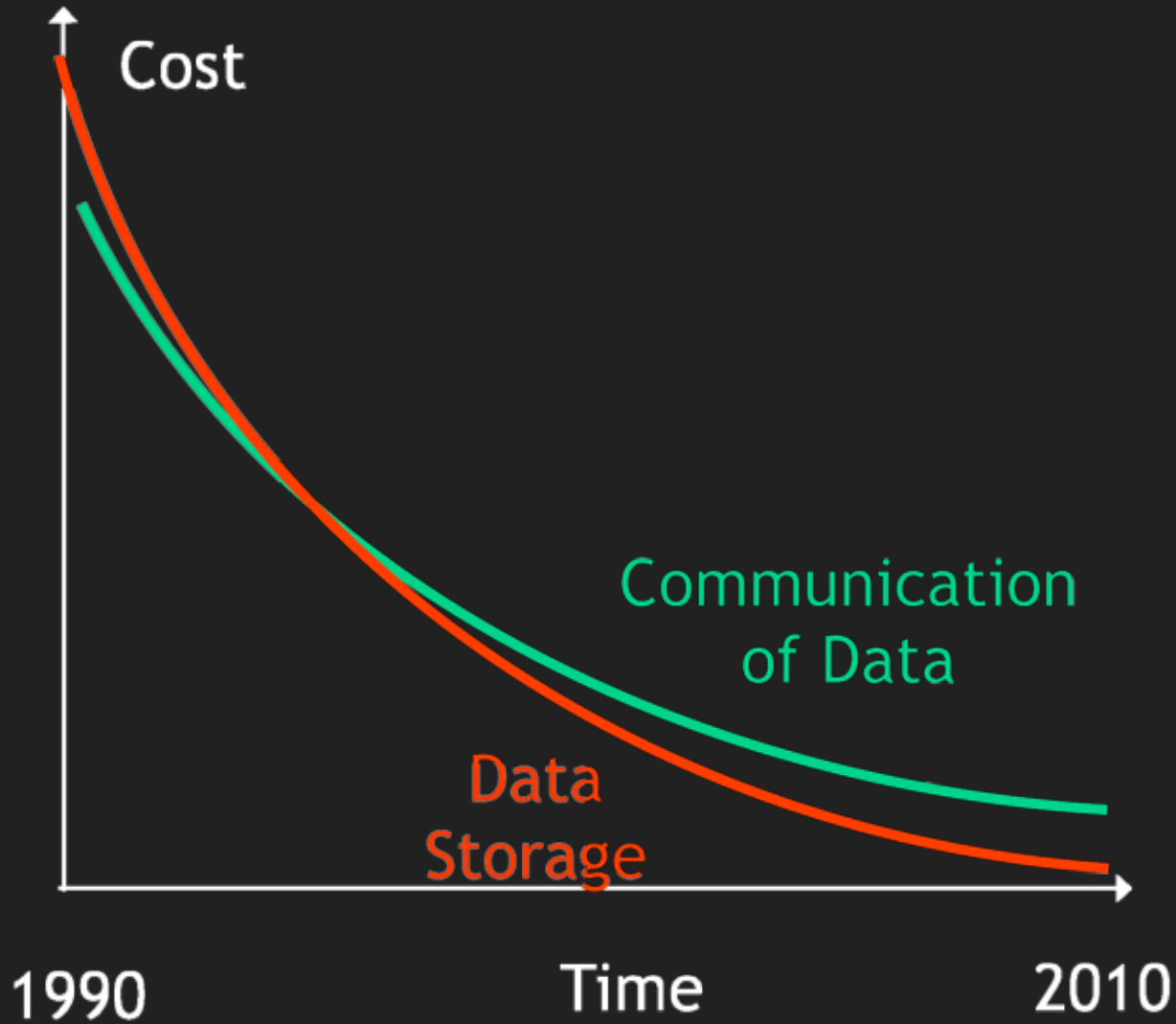


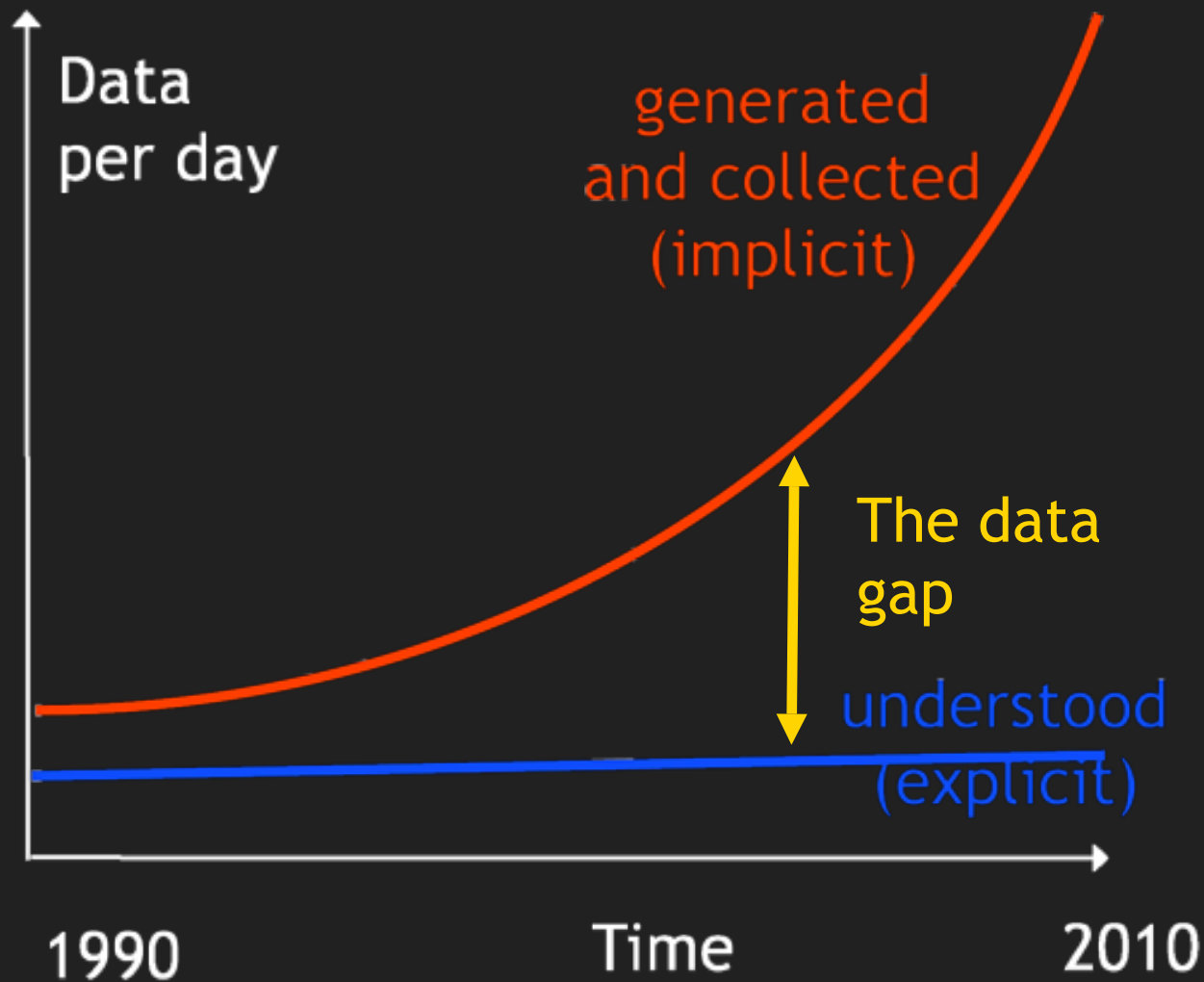
# Data Mining 101

George Tziralis, FOSS Conf, June 19 09, Athens, GR

# the facts



# the facts



the promise

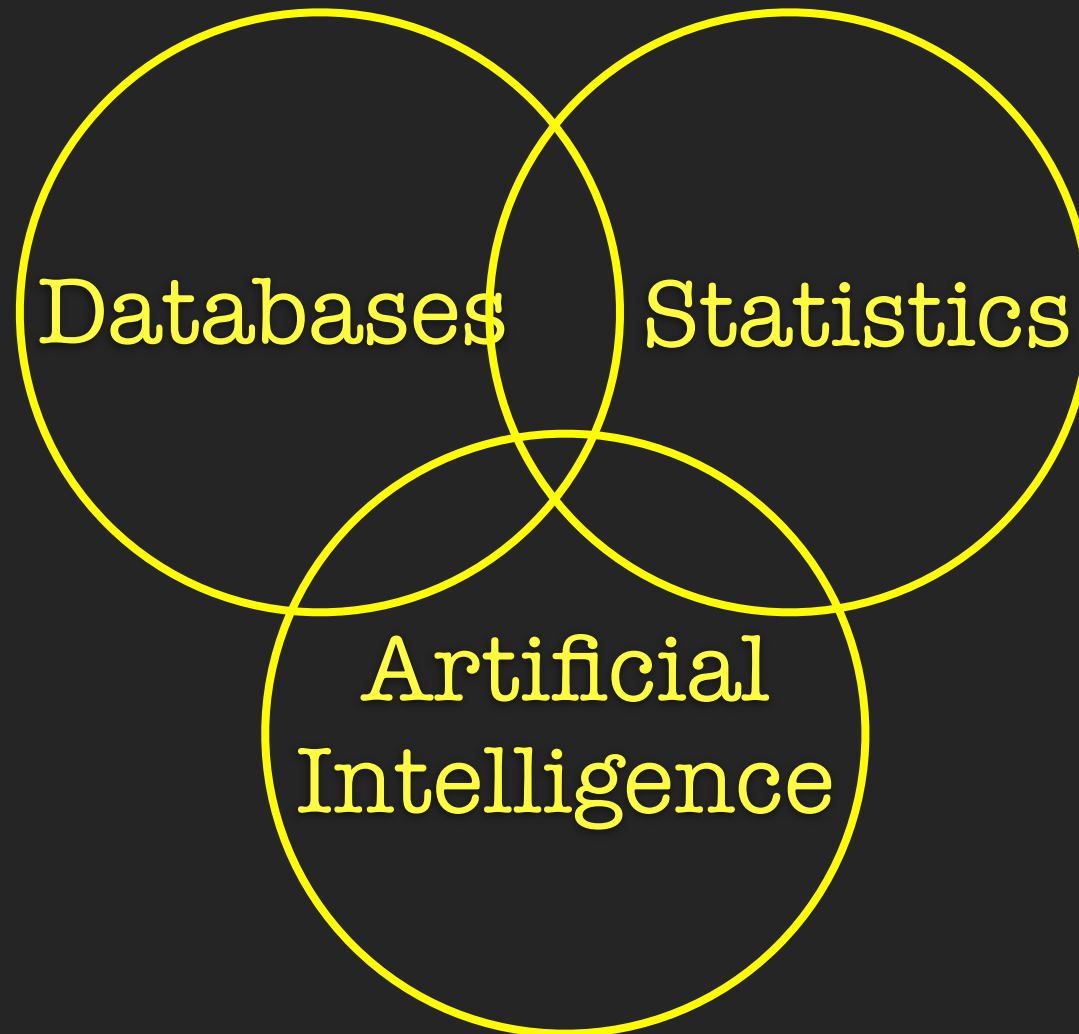
understand and take  
advantage of the world's  
information

the name

data mining:

statistics at speed, scale  
and simplicity

# what is



# the difference

- statistics: **define** a hypothesis, **then test**
- data mining: test **all possible** hypotheses
- is it possible? YES!

# the tasks

- classification
- association
- clustering
- prediction

# the process

- data **input** & exploration
- **preprocessing**
- data mining **algorithms**
- **evaluation** &  
intrepretation

# an example

| #  | color  | size | value | buy |
|----|--------|------|-------|-----|
| 1  | blue   | 5.32 | b     | no  |
| 2  | yellow | 8.57 | a     | yes |
| 3  | green  | 1.23 | c     | no  |
| 4  | yellow | 9.35 | c     | yes |
| 5  | red    | 5.99 | b     | yes |
| 6  | red    | 4.43 | b     | yes |
| 7  | green  | 6.21 | b     | no  |
| 8  | white  | 4.89 | a     | yes |
| 9  | black  | 5.15 | b     | no  |
| 10 | green  | 5.67 | b     | no  |

# an example

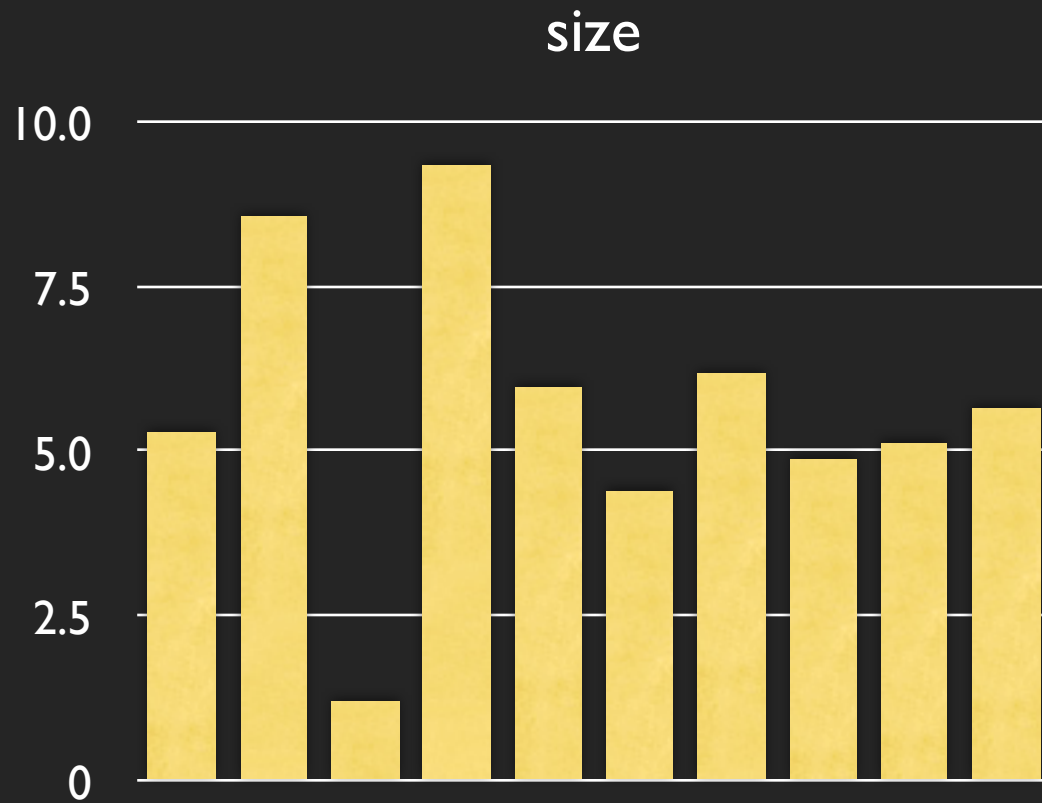
attribute

target

| #  | color  | size | value | buy |
|----|--------|------|-------|-----|
| 1  | blue   | 5.32 | b     | no  |
| 2  | yellow | 8.57 | a     | yes |
| 3  | green  | 1.23 | c     | no  |
| 4  | yellow | 9.35 | c     | yes |
| 5  | red    | 5.99 | b     | yes |
| 6  | red    | 4.43 | b     | yes |
| 7  | green  | 6.21 | b     | no  |
| 8  | white  | 4.89 | a     | yes |
| 9  | black  | 5.15 | b     | no  |
| 10 | green  | 5.67 | b     | no  |

instance

# so far



# now

- if `size = [4.0 - 7.0]` & `value = {b,c}`  
then `buy = no`

# now

- If color = yellow then buy = yes
- If color = red then buy = yes
- If color = white then buy = yes
- If color = green then buy = no
- If color = blue then buy = no
- If color = black then buy = no

ok, cool! but how?

# the tool



## **Weka**

Waikato Environment for Knowledge Analysis  
OSS, written in Java, providing API

# start



start -> explorer

# explore

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter  
Choose None Apply

Current relation  
Relation: contact-lenses  
Instances: 24 Attributes: 5

Selected attribute  
Name: age Type: Nominal  
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

| Label          | Count |
|----------------|-------|
| young          | 8     |
| pre-presbyopic | 8     |
| presbyopic     | 8     |

Class: contact-lenses (Nom) Visualize All

Attributes  
All None Invert

| No.                                 | Name                 |
|-------------------------------------|----------------------|
| <input checked="" type="checkbox"/> | 1 age                |
| <input type="checkbox"/>            | 2 spectacle-prescrip |
| <input type="checkbox"/>            | 3 astigmatism        |
| <input type="checkbox"/>            | 4 tear-prod-rate     |
| <input type="checkbox"/>            | 5 contact-lenses     |

Remove

Status  
OK Log x 0

open file -> data -> contact-lenses.arff

# .arff how-to

% ARFF file of the example's data

@relation testset

@attribute color {blue, yellow, green, red}

@attribute size numeric

@attribute value {a, b, c}

@attribute buy {yes, no}

@data

blue, 5.32, b, no

yellow, 8.57, a, yes

green, 1.23, c, no

...

# preprocess

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter

weka None

filters

supervised

unsupervised

attribute Attributes: 5

Add AddCluster AddExpression AddNoise ChangeDateFormat ClusterMembership Copy Discretize FirstOrder MakeIndicator MergeTwoValues NominalToBinary Normalize NumericToBinary NumericTransform Obfuscate

Remove

Selected attribute

Name: age Type: Nominal

Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

| Label          | Count |
|----------------|-------|
| young          | 8     |
| pre-presbyopic | 8     |
| presbyopic     | 8     |

Class: contact-lenses (Nom) Visualize All

Status OK Log x 0

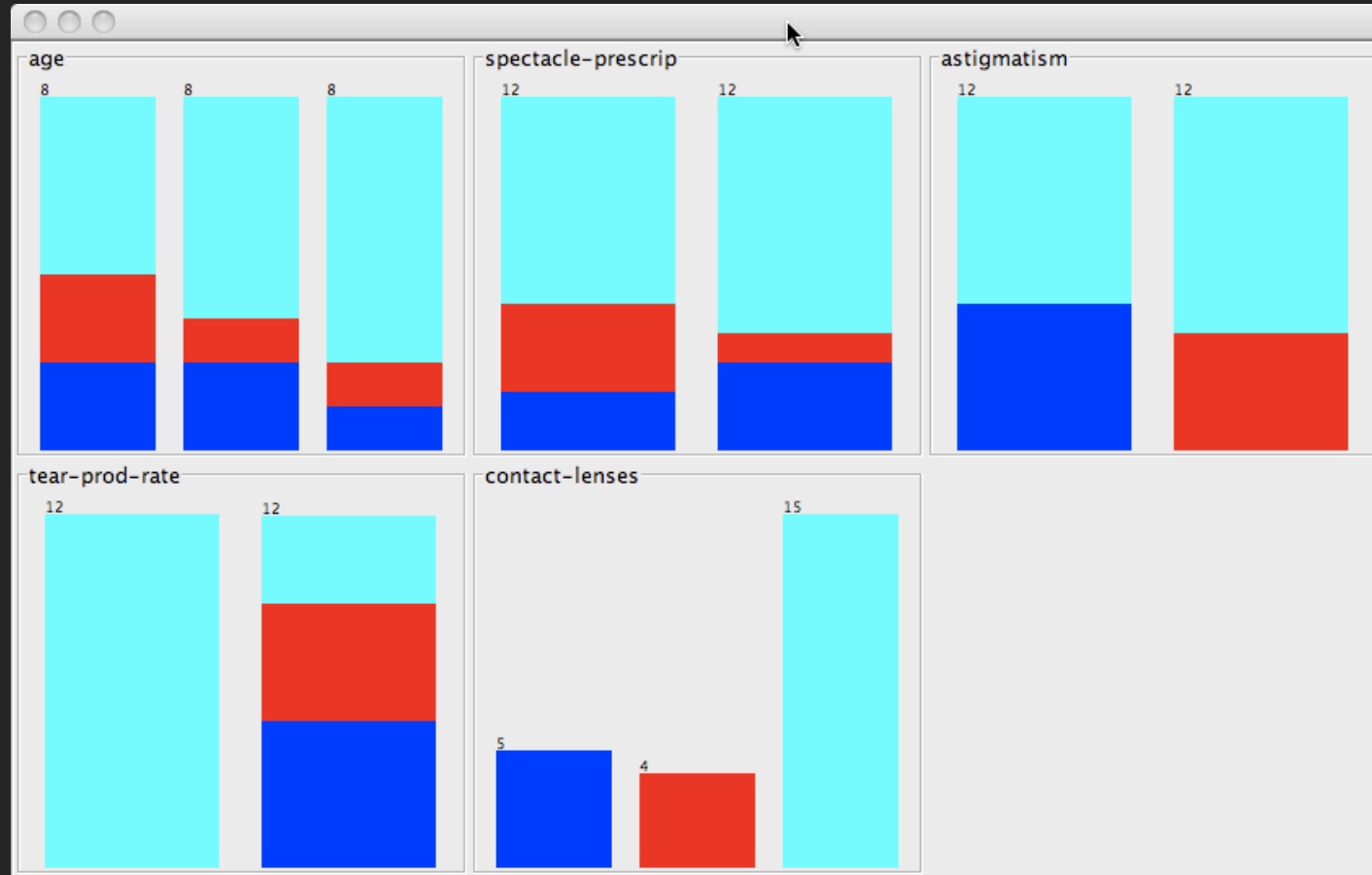
filter -> ... {tons of filters}

# visualize



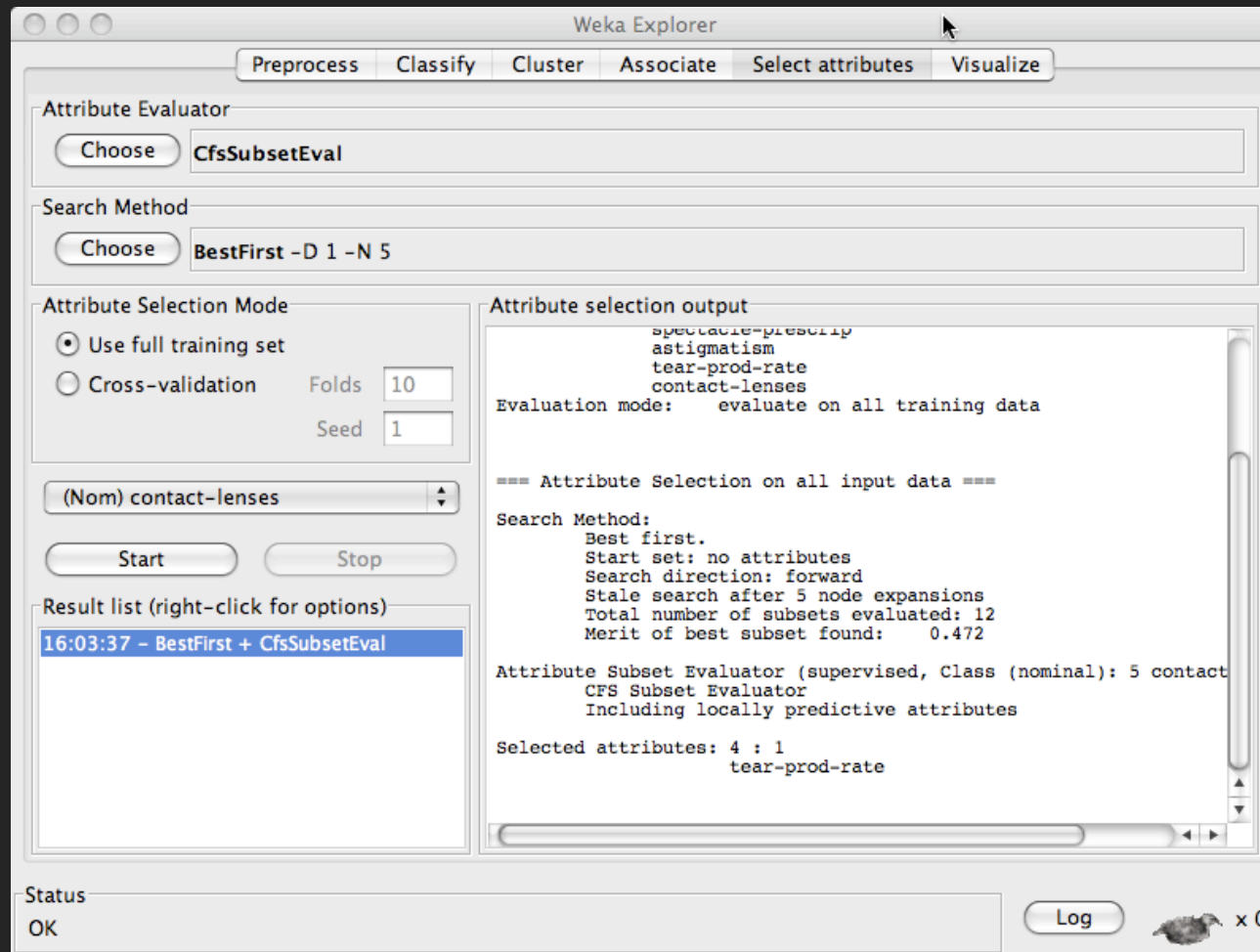
tab “visualize” (per target/class)

# visualize



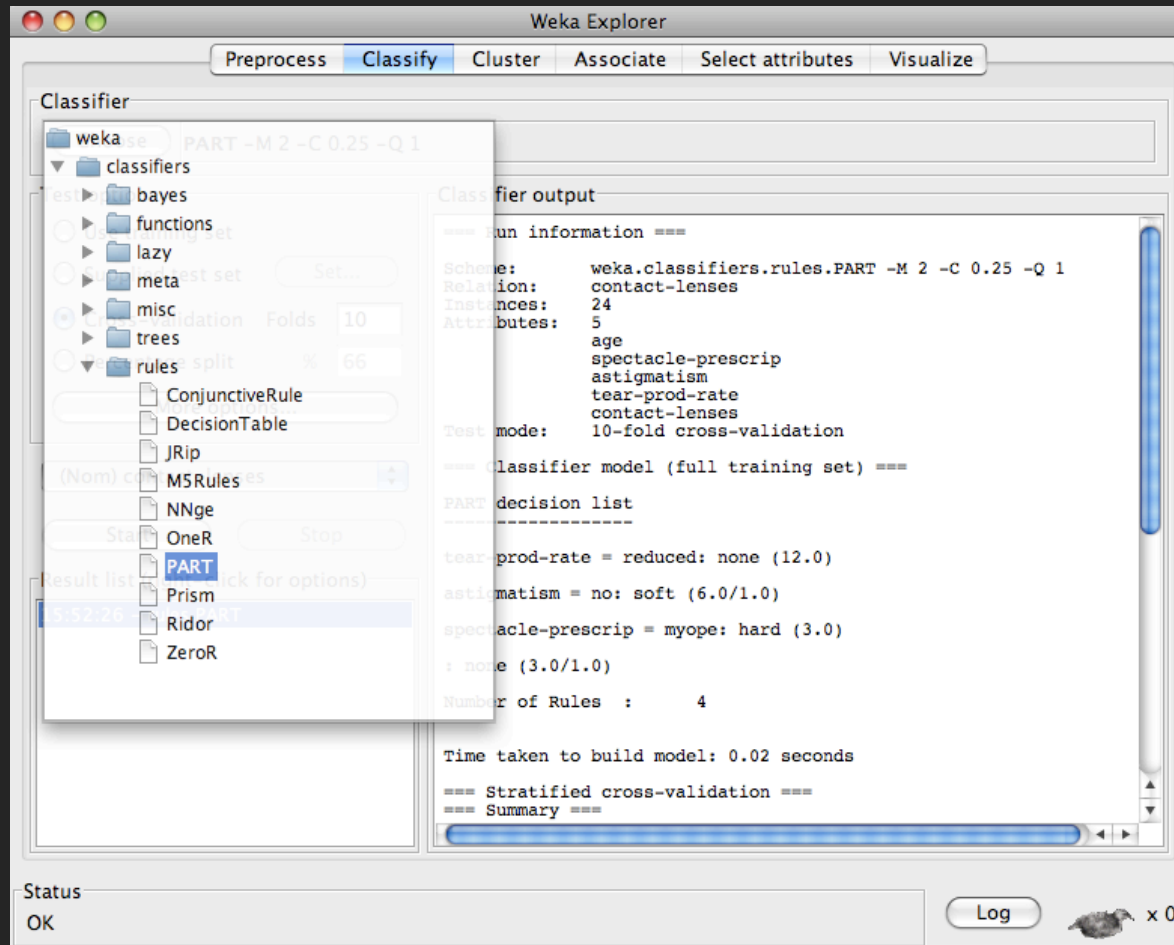
tab "preprocess" -> visualize all (per class)

# select attributes



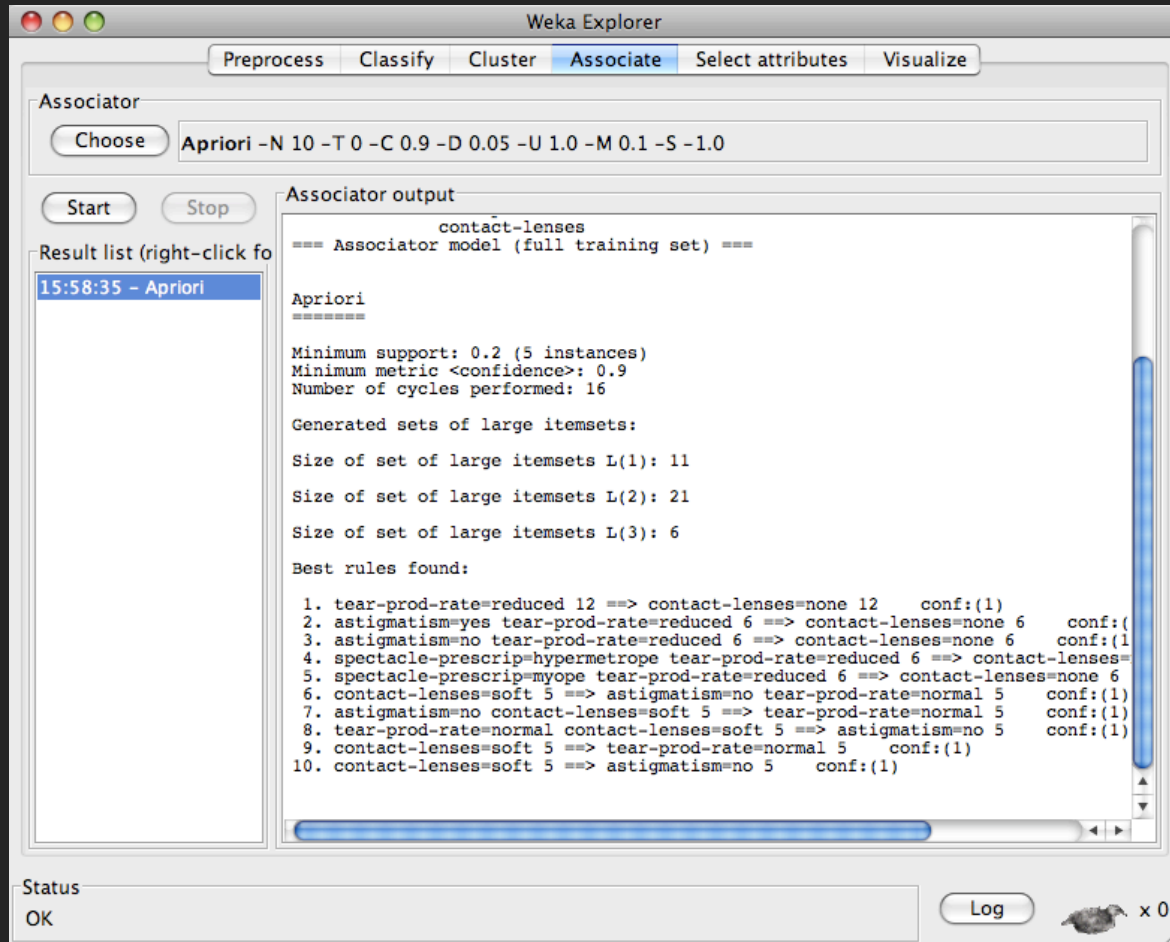
tab “select attributes” (default settings)

# classify



tab “classify” -> rules -> PART -> start!

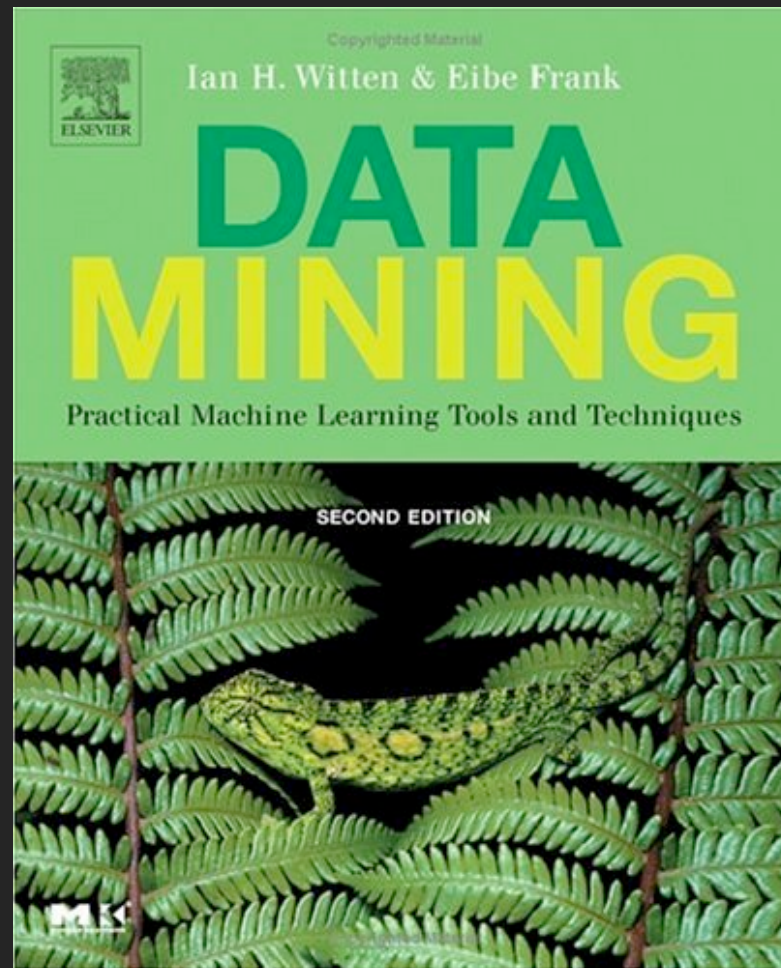
# associate



tab “associate” -> start! (default settings)

pls tell me more!

# the book



your data mining & data guide!

# Data Mining, a Course by Blog

[FRONT PAGE](#)[ABOUT](#)[LIVE](#)[LEADERBOARD](#)

## A Course by Blog, Lessons Learned

July 18, 2008 · No Comments

This semester, as you may know, I was privileged enough to teach (or, to be academically correct, to provide teaching assistantship, but I truly thank [my professor](#) for giving me the honor and full freedom to handle everything by myself) the course 'Information Extraction Algorithms' -pure Data Mining in practice- at the postgraduate program '[Applied Mathematical Sciences](#)' of the [National Technical University of Athens](#). And the whole course actually turned out to be a one-off experience, for both me and the 10 students that appeared to take it, by deciding to host the whole course process in a blog.

'A Course by Blog', a wordpress blog created on-the-fly during the first introductory lecture, finally ended up with 142 posts from 11 authors, 182

To search, type and hit enter

### DATA MINING, A COURSE BY BLOG

- [A Course by Blog, Lessons Learned](#)

July 18, 2008

This semester, as you may know, I was privileged enough to teach (or, to be academically correct, to provide teaching assistantship, but I truly thank my professor for giving me the honor and full freedom to handle everything by myself) the course 'Information Extraction Algorithms' -pure Data Mining in practice- at the postgraduate program [...]

thank you

[gtziralis@gmail.com](mailto:gtziralis@gmail.com)