EMPIRICAL SOFTWARE ENGINEERING (VERSION 2.0) AND DATA MINING

LASER SUMMER SCHOOL ,EMPIRICAL SE

SEPT 5-11, 2010, ELBA ISLAND, ITALY

Tim Menzies,
WVU, USA, tim@menzies.us, http://menzies.us

Download from http://unbox.org/wisp/var/timm/10/laser

Version history
V1: Aug18 '10
V1a: Aug28 '10
V1b: Sept02 '10
V1c: Sept11'10

Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

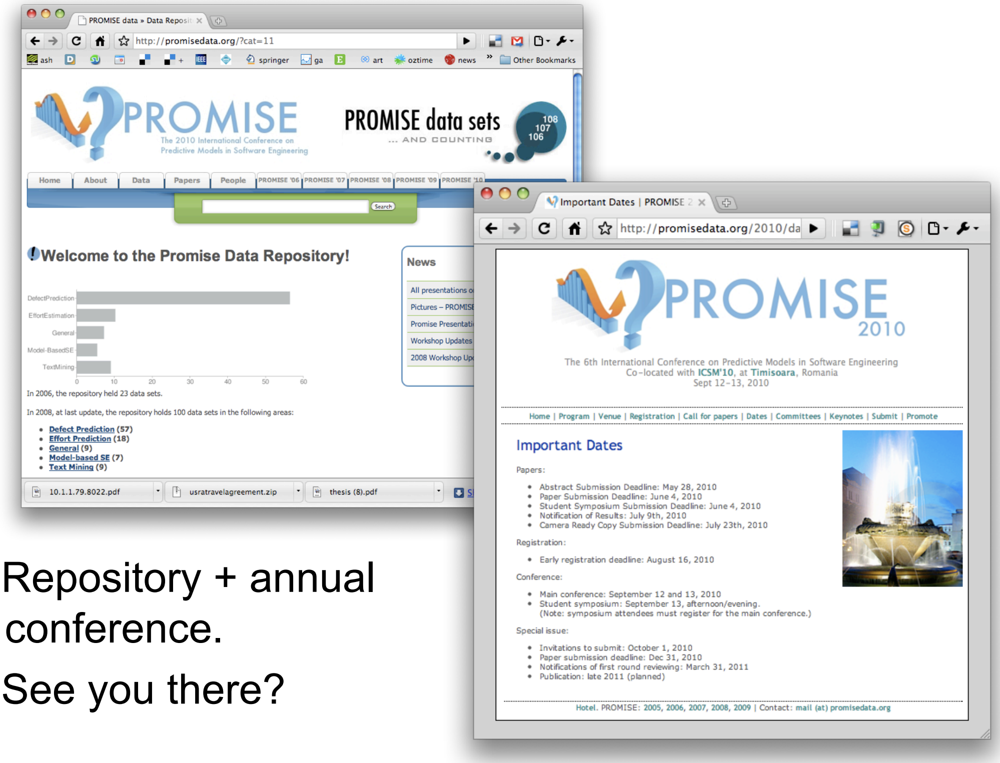
Change log

- Version 1: Aug 18, 2010
  - Version 1a: Aug 28, 2010
    - 2 more slides on "why empirical SE v2.0"
  - Version 1b: Sept 2: minor edits
  - Version 1c: minor edits
  - Version 1d: mew conclusion

About the author

- Dr. Tim Menzies (tim@menzies.us) has worked on advanced modeling + AI since 1986.
  - PhD from Uni. New South Wales, Sydney, Oz
  - Assoc/prof at WVU CS &EE
- Former research chair for NASA
- Author of190 refereed papers: http://menzies.us/papers.php
- Co-founder and organizer of the PROMISE conferences on repeatable experiments in SE
- For more, see http://menzies.us

http://promisedata.org/data



Repository + annual conference.

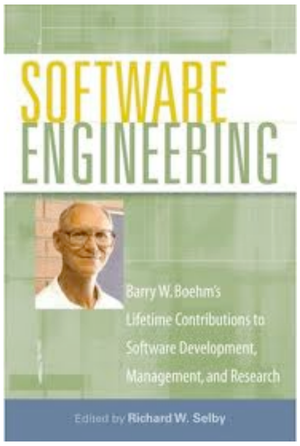
See you there?

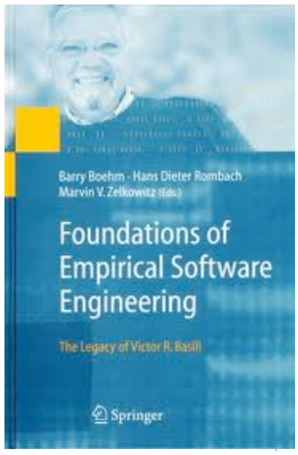
New trend

- Ph.D. students, finishing up their studies, using PROMISE to archive their data
  - E.g. 30 new OO data sets from Marian Jureczko
- Are you next?

Digressions

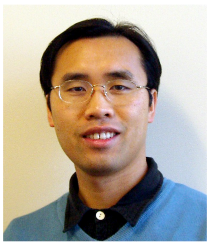
- References and further reading:
  - shown in blue.
- The following material has more Barry Boehm references than Victor Basili
  - Only cause I've been working with Barry on effort estimation & value-based SE.
  - To redress that imbalance, see
    - Forrest Shull, Carolyn Seaman, Marvin Zelkowitz, "Victor R. Basili's Contributions to Software Quality," *IEEE Software, vol. 23, no. 1, pp. 16-18, Jan./Feb. 2006,*
    - *Or*

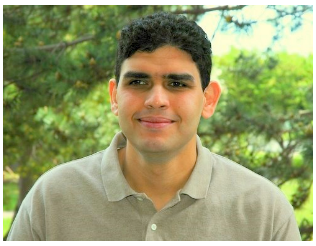




For other view on DM + SE

- ICSE 2010 Tutorial T18 Tuesday, 4 May 2010 (afternoon)
- Mining Software Engineering Data
  - Ahmed E. Hassan: Queen's University, Canada
  - Tao Xie: North Carolina State University, USA





- Tutorial Slides:
  - https://sites.google.com/site/asergrp/dmse/dmse-icse08-tutorial.ppt?attredirects=0

DATA MINING & SE (OVERVIEW)

Road map

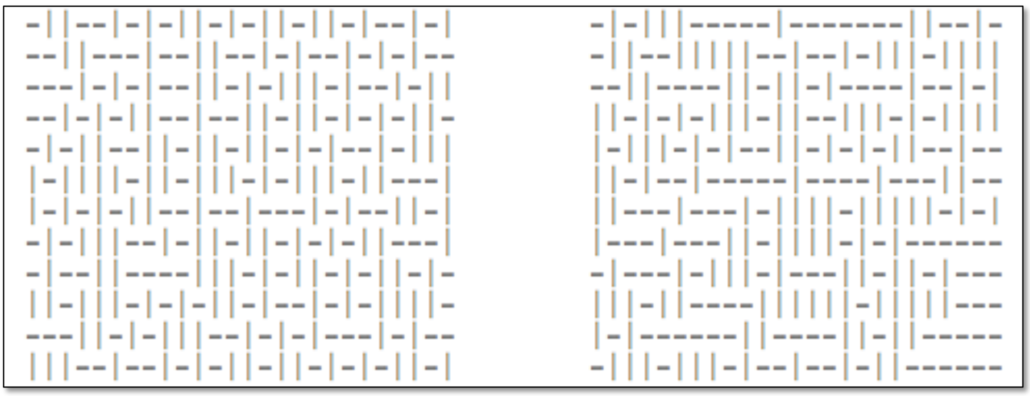
1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

Definition

- Finding patterns in (lots of) data
  - Diamonds in the dust
- Combines statistics, AI, visualization, ….
- Synonyms
  - Machine learning
  - Business intelligence
  - Predictive analytics
- The art of the approximate scalable analysis
  - Bigger is better
- Used for… anything
  - The review of current beliefs w.r.t. new data is the hallmark of human rationality.
  - It is irrational <u>NOT </u>to data mine.

Exercise #1

- One these these things is not like the other
  - One was generating by selecting "-" or "|" at random, 300 times.
- Which one?



Exercise #2

- A little experiment from http://www.youtube.com/v/vJG698U2Mvo&hl=en_US&fs=1&rel=0

- Rules
  - No one talks for the next 4 minutes
  - If you know what is about to happen, see (1)

- This is a selective attention test
  - Count the number of times the team with the white shirt passes the ball.

What have we learned?

- Lesson #1:
  - Algorithms can be pretty dumb
  - If they don't focus on X, they see any Y, at random.
- Lesson #2:
  - Humans can be pretty dumb
  - If they mono-focus on X, you can miss Y
- Maybe, any induction process is a guess
  - And while guessing can be useful
  - Guesses can also be wrong
- Lets us a create community of agents, each with novel insights and limitations
  - Data miners working with humans
  - Maybe in combination, we can see more that separately

Wikipedia:
List of cognitive biases
http://en.wikipedia.org/wiki/List_of_cognitive_biases
- 38 decision making biases
- 30 biases in probability
- 18 social biases,
- 10 memory biases

Applications

- Effort estimation
- Defect prediction
- Optimization of discrete systems
- Test case generation
- Fault localization
- Text mining
- Temporal sequence mining
  - Learning software processes
  - Learning APIs
- Etc
- Welcome to Empirical SE, Version 2.0

Applications

- Effort estimation
- Defect prediction
- Optimization of discrete systems
- Test case generation
- Fault localization
- Text mining
- Temporal sequence mining
  - Learning software processes
  - Learning APIs
- Etc
- Welcome to Empirical SE, Version 2.0

Data mining applications explored by me since 2007.

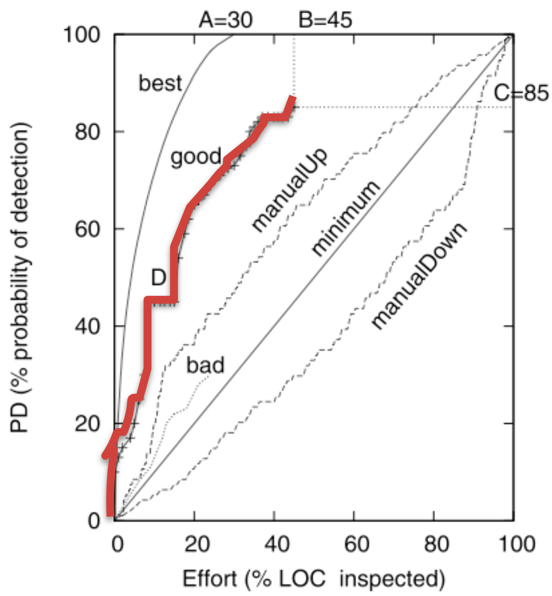
A career in data mining is a very diverse career, indeed

Application: Effort estimation

- Can we predict development effort (time * staff)?
- E.g. using linear regression; effort = a*KLOC^b c
  - Boehm, B.W. 1981 Software Engineering Economics
  - Boehm, B.W., Clark, Horowitz, Brown, Reifer, Chulani, Madachy, R., and Steece, B. 2000 Software Cost Estimation with Cocomo II
  - Sunita Chulani, Barry W. Boehm, Bert Steece: Bayesian Analysis of Empirical Software Engineering Cost Models IEEE Trans. Software Eng. 25(4): 573-583 (1999)
- E.g. using analogy
  - Describe past projects according to N dimensions
  - Float all known projects in an N-dimensional space
  - To estimate a project, insert into that space; query its nearest neighbors
  - For the classic estimation via analogy, see
    - Martin J. Shepperd, Chris Schofield: Estimating Software Project Effort Using Analogies IEEE Trans. Software Eng. 23(11): 736-743 (1997)
  - For 12,000+ variants to that process, see
    - Fig1 of http://menzies.us/pdf/10stable.pdf
- E.g. using other methods:
  - See 154 variants in http://menzies.us/pdf/10stable.pdf

Application: Defect Prediction

- Limited QA budgets, can't check everything.
  - Where should we place our inspection effort?
- For a review, see Section Two of
  - http://menzies.us/pdf/10which.pdf
- Practical value:
  - How to inspect less, and find more bugs



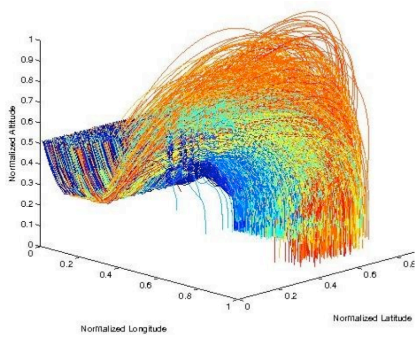
Application:
Optimizations of discrete systems

- Standard numeric optimizers assume continuous, possibly even linear, equations

- Data miners much happier to work in discrete spaces.

- What factors predict for landing closest to the target?
    - State-of-the-art optimizer
    - Simulated annealing
    - the TAR3 data miner
    - TAR3 45 times faster, found better solutions

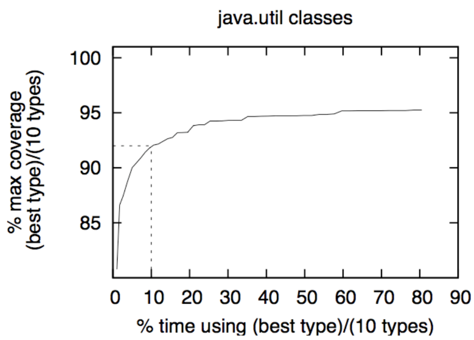
http://menzies.us/pdf/10keys.pdf



Application: Test Case Generation

- NIGHTHAWK: A genetic algorithm that mutates sequences of method calls in order to maximize code coverage.
- RELIEF: a data mining technique to find "interesting features"
    - Same attribute same values in all classes?
        - Boring
    - Same Attribute, different values in different classes?
        - Interesting
- RELIEF found that 90% of NIGHTHAWK's mutators were "boring"
    - Order of magnitude speed up in test generation
- James H. Andrews, Tim Menzies, Felix C.H. Li, "Genetic Algorithms for Randomized Unit Testing," IEEE Transactions on Software Engineering, 25 Mar. 2010.

Rank	Gene type t	$avgMerit$
1	numberOfCalls	85
2	valuePoolActivityBitSet	83
3	upperBound	64
4	chanceOfTrue	50
5	methodWeight	50
6	numberOfValuePools	49
7	lowerBound	44
8	chanceOfNull	40
9	numberOfValues	40
10	candidateBitSet	34



Application: Fault Localization

- 100,000 JAVA methods
  - In a matrix T*D
  - T = “terms” = all the method calls in each method
  - D = “documents” = all the methods

- Bug report
  - Replace text with just the method calls it mentions
  - Add edited report as row D+one in the matrix
  - Compute similarity of D+one to other rows (cosine similarity)
  - The actual buggy method is in the closest 100 methods
  - Use relevancy feedback to narrow down the search

- Gregory Gay, Sonia Haiduc, Andrian Marcus Tim Menzies: On the use of relevance feedback in IR-based concept location ICSM 2009: 351-360

Application: Text Mining

- 80% of data in organizations is unstructured
  - Not in databases, or XML schemas
  - But in the natural language of (say) Word documents

- Given enough of these seemingly unstructured documents, structures can be discovered

- E.g.
  - Thousands of natural language bug reports from NASA
  - Used “feature reduction” to find the top 100 most important words
  - Used standard data mining to learn predictors for defect severity from that top-100
  - Tim Menzies, Andrian Marcus: Automated severity assessment of software defect reports. ICSM 2008: 346-355

Application: Temporal Sequence Mining

- Learning software process descriptions
  - No more prescriptions of what we think goes on inside software projects
  - Lets look at see at what actually happens
    - Li, Mingshu and Boehm, Barry and Osterweil, Leon and Jensen, Chris and Scacchi, Walt "Experiences in Discovering, Modeling, and Reenacting Open Source Software Development Processes", Unifying the Software Process Spectrum, Lecture Notes in Computer Science, 2006, page 449 to 462
- Learning APIs from method sequence calls
  - Tao Xie and Jian Pei. MAPO: Mining API Usages from Open Source Repositories. In *Proceedings of the 3rd International Workshop on Mining Software Repositories (MSR 2006), Shanghai, China, pp. 54-57, May 2006*
- Learning patches from method sequence calls
  - Suresh Thummalapenta and Tao Xie. Mining exception-handling rules as sequence association rules. In ICSE '09: Proceedings of the 31st International Conference on Software Engineering, pages 496– 506, Washington, DC, USA, 2009. IEEE Computer Society.
- Obtaining sequence miners:
  - https://illimine.cs.uiuc.edu/
  - Another tool set is at http://himalaya-tools.sourceforge.net/
  - See more tools at https://sites.google.com/site/asergrp/dmse/resources

Application: etc etc etc

- Data mining + SE a very active area
  - PROMISE conference
  - Mining Software Repository conference
- See also
  - ESEM conference
  - Search-based software engineering
- Hint: to get ahead of the curve…
  - … learn sequence mining
- Welcome to Empirical SE, version 2.0

Empirical SE, Version 2.0

- Open Science movement
  - Open Data
    - Everyone places their data on-line, all the time
  - Open Access publishing
    - Death to subscription-based services
- Shneiderman, B. (2008) "Science 2.0" Science 319(5868):1349-50
  - Science meets web 2.0
  - International team of researchers posting and analyzing data
  - Research at internet speed
- Anda, Markus et al (*) distinguish between
  - **Case studies:** that collect new context variables from project data
  - **Experiments**: that explore case study data
  - Currently, very few case studies generating publicly available data
    - But very many researchers wanting to experiment on that data
    - Perfect setting for data mining
- (*) Bente Anda Audris Mockus and Dag I.K. Sjoberg. Experiences from replicating a case study to investigate reproducibility of software development. In First International Workshop on Replication in Empirical Software Engineering Research, ICSE'09,

Q: Why Empirical SE 2.0?
A: Case study results may not generalize

- What is true at one site,
  - May not be true for another
  - E.g. local sites have different goals, different biases, that changes what is "best" for that site
    - II, P. G., Menzies, T., Williams, S., and El-Rawas, O. 2009. Understanding the Value of Software Engineering Technologies. In *Proceedings of the 2009 IEEE/ACM international Conference on Automated Software Engineering(November 16 - 20, 2009)*
  - E.g. general policies perform worse than locally generated policies
    - Menzies, T., Williams, S., Boehm, B., and Hihn, J. 2009. How to avoid drastic software process change (using stochastic stability). In *Proceedings of the 31st international Conference on Software Engineering (May 16 - 24, 2009*
- So we need to audit the conclusions of one case study w.r.t. to data taken from other sites.
- Data mining is one technology that can (at least partially) automate that audit process

Q: Why Empirical SE 2.0? A: Sharing is a good thing

- WC= Within- company data
  - Locally collected, locally applied
- CC= Cross- company data
  - Collected elsewhere, applied here.
- Filtered CC works nearly as well as WC
  - Turhan, B., Menzies, T., Bener, A. B., and Di Stefano, J. 2009. On the relative value of cross-company and within-company data for defect prediction. *Empirical Softw. Engg. 14, 5 (Oct. 2009), 540-578*
  - Ekrem Kocaguneli, Gregory Gay, Tim Menzies, Ye Yang, Jacky Keung, When to Use Data from Other Projects for Effort Estimation, *IEEE ASE 2010*
- So if ever you are doing new work,
  - and lack local data,
  - you can apply other people's data
- But only if it is available
  - Open data !!!

Q: Why Empirical SE 2.0 A: Changing nature of data

- In the 21st century
  - we can access more data collected by <u>others</u> than we can ever can collect by <u>ourselves</u>.
- In the 20th century,
  - research was focused on case studies where researchers collected special purpose data sets for their particular questions.
- In the 21st century,
  - much research is devoted to experimentation with the data generated by the case studies,
  - possibly investigating hypotheses not originally considered when the data was collected.
  - Data mining is one way to experiment with data.

Q: Why Empirical SE 2.0?
A: Increasing pace of change

- New developments are radically changing SE: open source toolkits, agile development, cloud-based computing, etc.

- 20th century Empirical SE used "big science"
  - Research questions, data collection, analysis took years
  - Big science is too slow to keep up with changes to contemporary SE. e.g.
    - Increasing pace of organization change at NASA was fatal to the "big science" approach of Victor Basili's Software Engineering Laboratory (*)
    - V. Basili, F. McGarry, R. Pajerski, and M. Zelkowitz. Lessons learned from 25 years of process improvement: The rise and fall of the NASA software engineering laboratory. In Proceedings of the 24th International Conference on Software Engineering (ICSE) 2002, Orlando, Florida, 2002.

- Data mining is one response to the open and urgent issue of
  - how to reason faster about SE data.

Q: Why Empirical SE 2.0?
A: Changing nature of SE theories

- 20th century SE: the struggle for the single theory
  - E.g. Boehm's COCOMO effort estimation project
  - E.g. SEI capability maturity model [130];

- 21st century: faster pace = more diversity
  - Less likely that there exists a single over-arching grand theory of SE
- Recent reports [1,2,3,4,5] say that while such generality may elude us, we can still find the special lessons that work best on the local projects
  1. Rombach A. Endres, H.D.A Handbook of Software and Systems Engineering: Empirical Observa- tions, Laws and Theories. Addison Wesley, 2003.
  2. B. Kitchenham D. Budgen, P. Brereton. Is evidence based software engineering mature enough for practice & policy? In 33rd Annual IEEE Software Engineering Workshop 2009 (SEW-33), Skvde, Sweden, 2009.
  3. B. A. Kitchenham, E. Mendes, and G. H. Travassos. Cross- vs. within-company cost estimation studies: A systematic review. IEEE Transactions on Software Engineering, pages 316–329, May 2007.
  4. Tim Menzies and Forrest Shull. The quest for convincing evidence. In A. Oram and G. Wilson, editors, Making Software: What Really Works, and Why We Believe It. O'Reilly, 2010.
  5. H. Gall E. Giger T. Zimmermann, N. Nagappan and B. Murphy. Cross-project defect prediction. In ESEC/FSE'09, August 2009.

- Data mining is one way to rapidly find and verify the special practices that best work on the local projects.

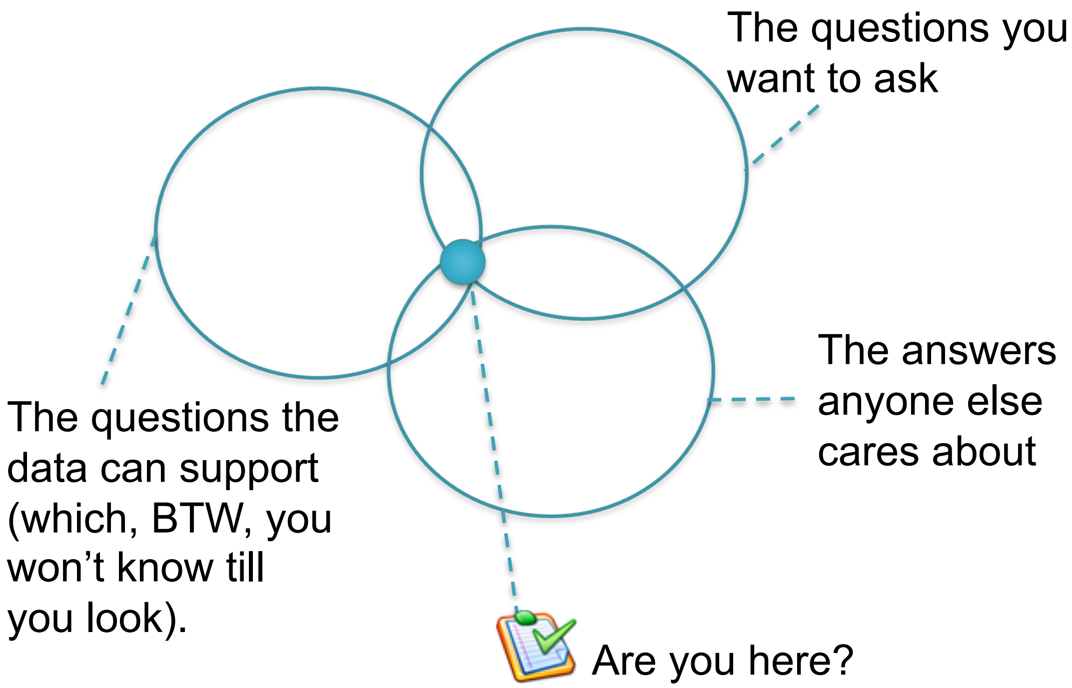
Q: Why Empirical SE 2.0? A: Changing nature of data analysis

- A contemporary empirical SE paper might explore gigabytes of core dumps looking for the method calls that lead to a crash.
- Faced with such large and complex data, analysis methods are becoming more intricate; e.g.
  - Model trees for multi-model data
  - Latent Dirichlet allocation (LDA) for document clustering
  - Mining sequences to learn exception handling rules
- It is now possible to find new insights in old data, just by applying a new analysis method.
  - E.g. see later, the "W" tool

Why Data Mining for SE?

- Natural tool to help a community:
  - racing to keep up with the pace of change in SE;
  - while finding and verifying the special theories that work best on local projects …
  - … from a new kind data sources …
  - … using a large menagerie of new data analysis tools.

Empirical Science 2.0 adjusts its questions to the available data



Coming next…

- Enough generalities
- Details of using a data mining tool suite
  - The "WEKA"

DIGRESSION #1

Are we spending our time on things that matter?

- Objects,
- Aspects,
- Pair programming,
- Design patterns
- ….

id	features	relative weight
1	Personnel/team capability	3.53
2	Product complexity	2.38
3	Time constraint	1.63
4	Required software reliability	1.54
5	Multi-site development	1.53
6	Doc. match to life cycle	1.52
7	Personnel continuity	1.51
8	Applications experience	1.51
9	Use of software tools	1.50
10	Platform volatility	1.49
11	Storage constraint	1.46
12	Process maturity	1.43
13	Language & tools experience	1.43
14	Required dev. schedule	1.43
15	Data base size	1.42
16	Platform experience	1.40
17	Arch. & risk resolution	1.39
18	Precedentedness	1.33
19	Developed for reuse	1.31
20	Team cohesion	1.29
21	Development mode	1.32
22	Development flexibility	1.26

Source: Boehm 2000. Regression results from 161 projects.

Are we spending our time on things that matter?

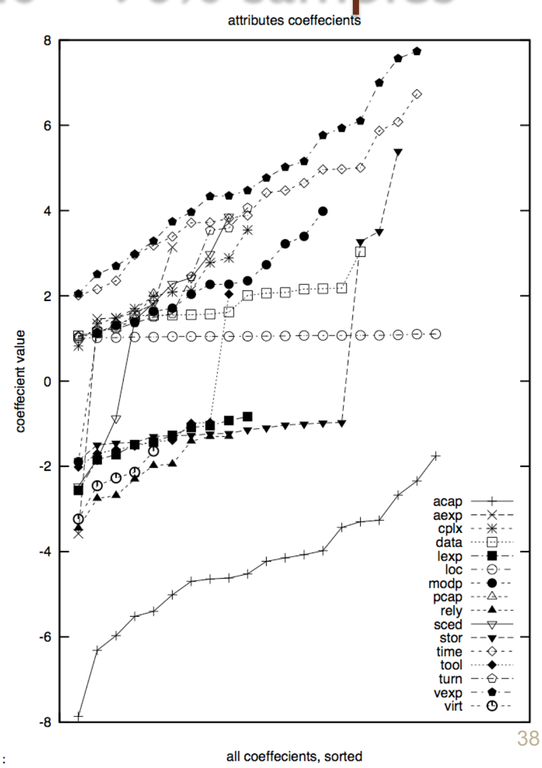
- Objects,
- Aspects,
- Pair programming,
- Design patterns
- ….

id	features	relative weight
1	Personnel/team capability	3.53
2	Product complexity	2.38
3	Time constraint	1.63
4	Required software reliability	1.54
5	Multi-site development	1.53
6	Doc. match to life cycle	1.52
7	Personnel continuity	1.51
8	Applications experience	1.51
9	Use of software tools	1.50
10	Platform volatility	1.49
11	Storage constraint	1.46
12	Process maturity	1.43
13	Language & tools experience	1.43
14	Required dev. schedule	1.43
15	Data base size	1.42
16	Platform experience	1.40
17	Arch. & risk resolution	1.39
18	Precedentedness	1.33
19	Developed for reuse	1.31
20	Team cohesion	1.29
21	Development mode	1.32
22	Development flexibility	1.26

Source: Boehm 2000. Regression results from 161 projects.

Once is not enough: Stability studies: 20 * 90% samples

- 20 experiments, using 66% of the data (selected at random)
- Linear regression:
  - Effort = b_0 + sum of $b_{i} * x_i$
  - Followed by a greedy back-select to prune dull variables
- Results
  - LOC influence stable
  - Some variables pruned away half the time
  - Large ranges (max – min)
  - Nine attributes even change the sign on their coefficients



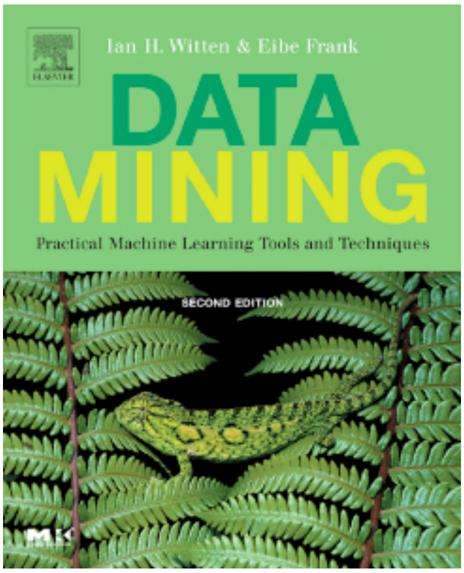
DATA MINING TOOLS (GUIDED TOUR OF "WEKA")

Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

WEKA

- Machine learning/data mining software written in Java
  - Used for research, education, and applications
  - Complements Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) Ian H. Witten, Eibe Frank, Morgan Kaufmann June 2005 525 pages ISBN 0-12-088407-0
- Main features
  - Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - Graphical user interfaces (incl. data visualization)
  - Environment for comparing learning algorithms



Access

- WEKA is available at
  http://www.cs.waikato.ac.nz/ml/weka
- Also has a list of projects based on WEKA
- WEKA contributors:

Abdelaziz Mahoui, Alexander K. Seewald, Ashraf M. Kibriya, Bernhard Pfahringer , Brent Martin, Peter Flach, Eibe Frank ,Gabi Schmidberger ,Ian H. Witten , J. Lindgren, Janice Boughton, Jason Wells, Len Trigg, Lucio de Souza Coelho, Malcolm Ware, Mark Hall ,Remco Bouckaert , Richard Kirkby, Shane Butler, Shane Legg, Stuart Inglis, Sylvain Roy, Tony Voyle, Xin Xu, Yong Wang,

Zhihai Wang

Data Files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

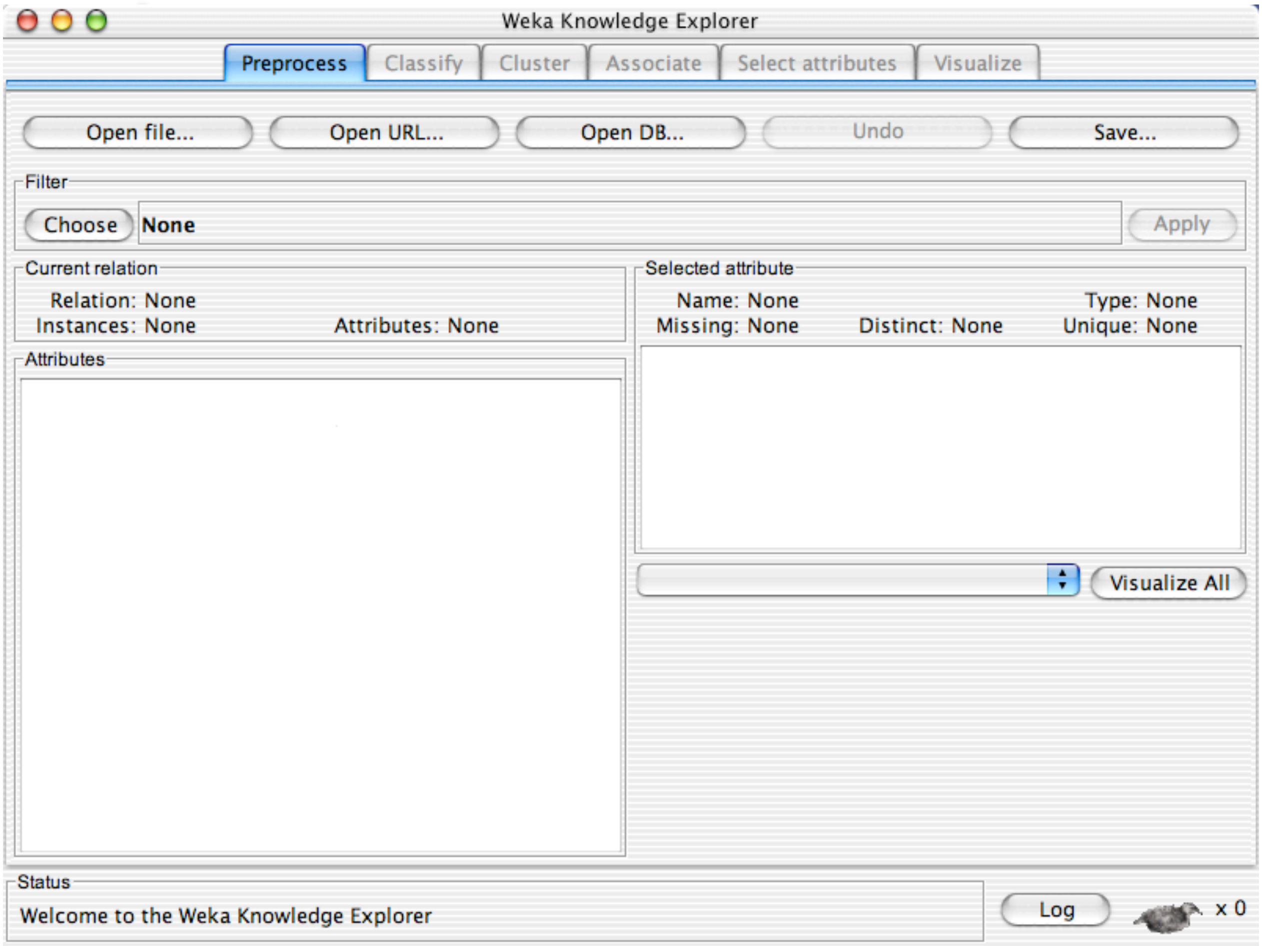
numeric attribute

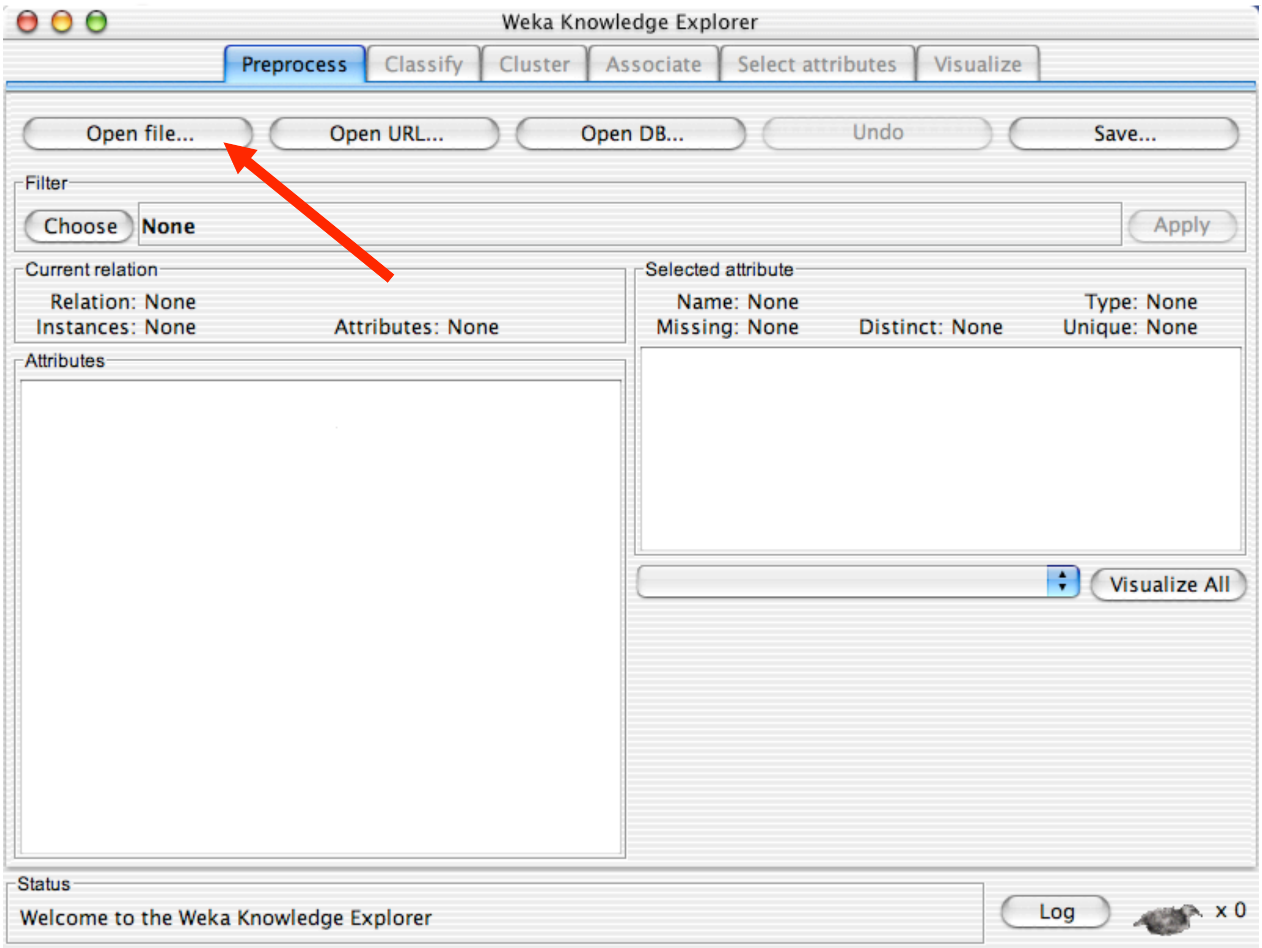
nominal attribute

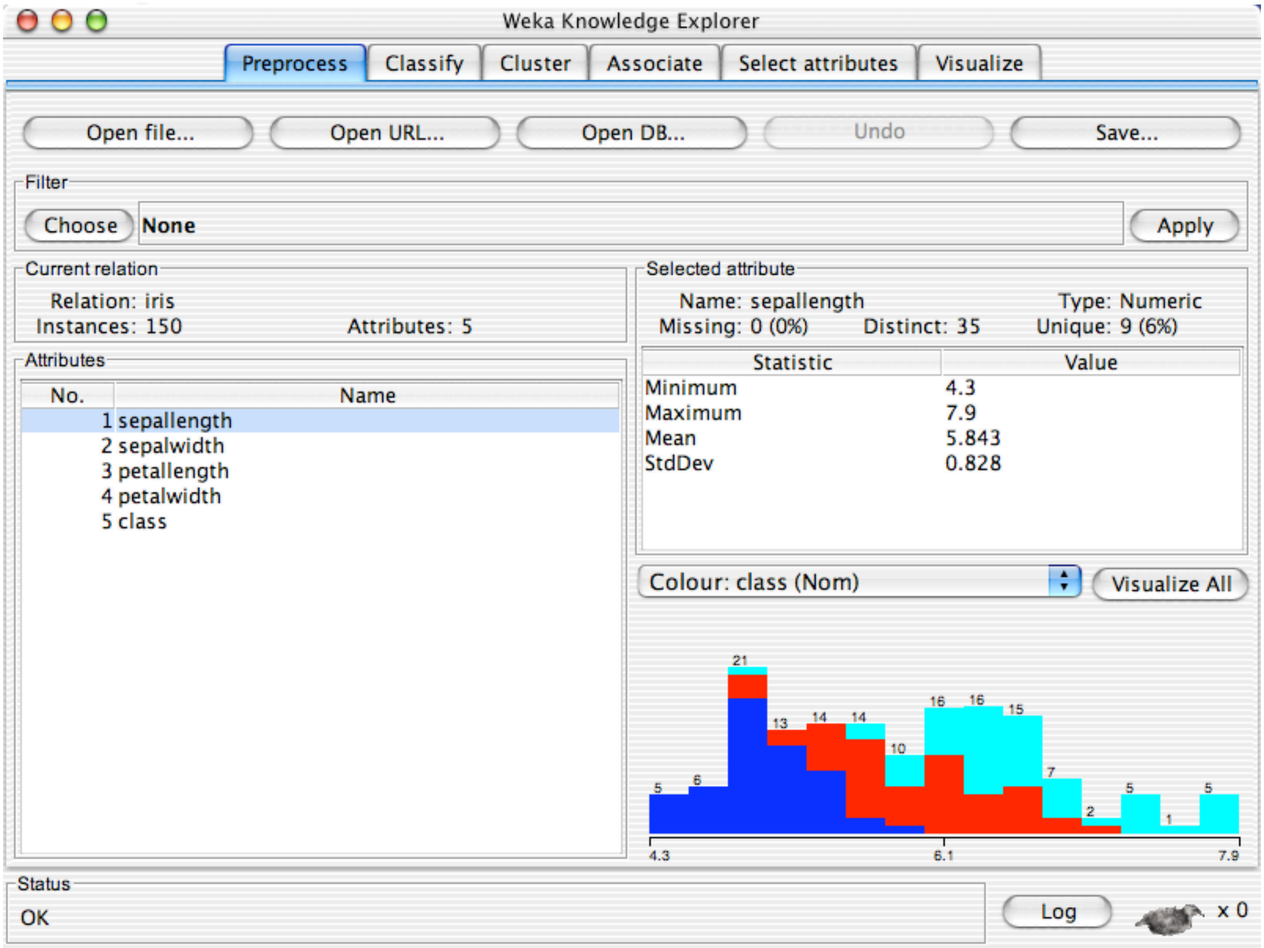
Flat file in ARFF format

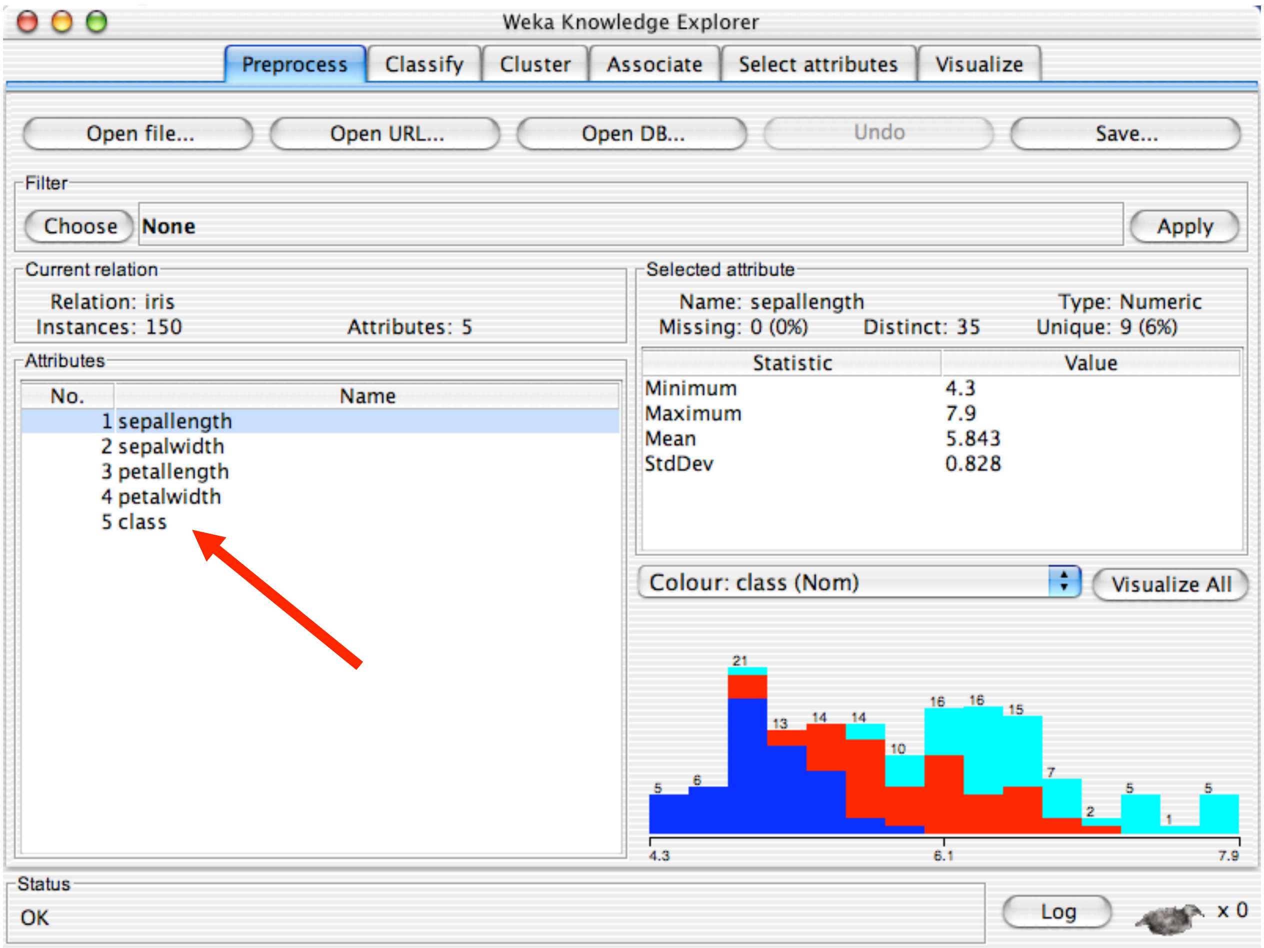
Explorer: pre-processing

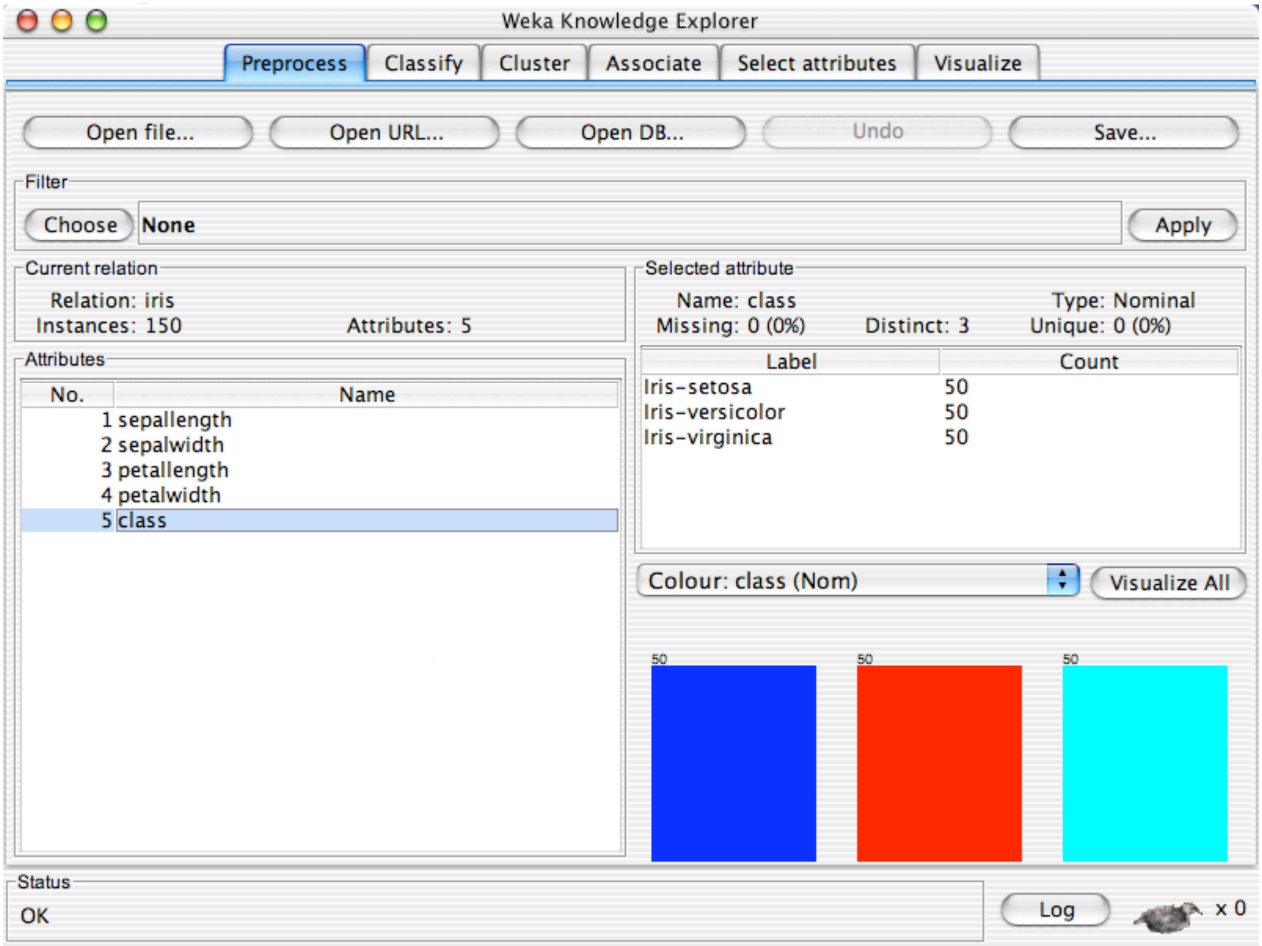
- Source
  - Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
  - Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools
  - Called "filters"
  - Discretization, normalization, resampling, attribute selection, transforming and combining attributes, …

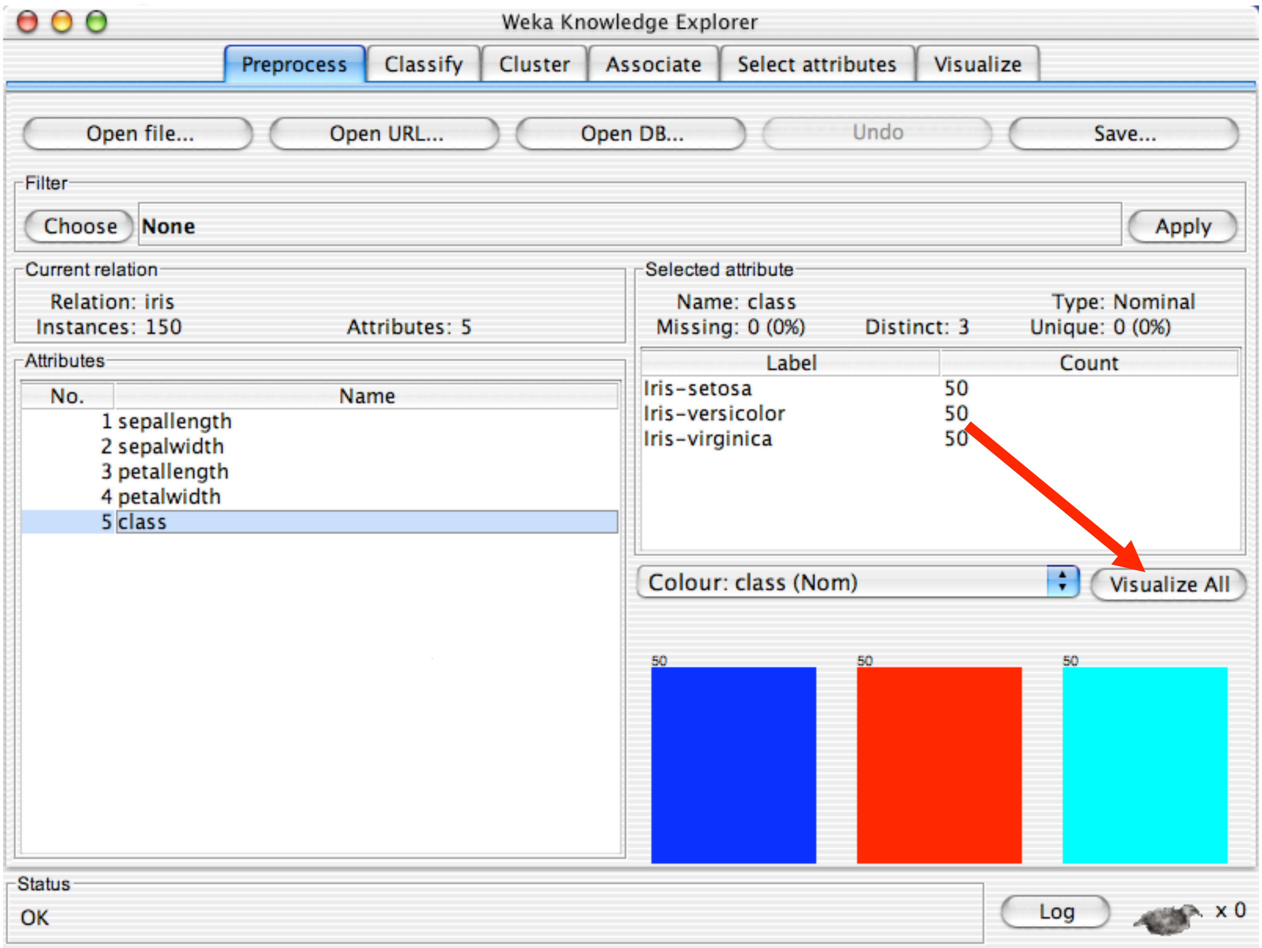

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Attributes
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

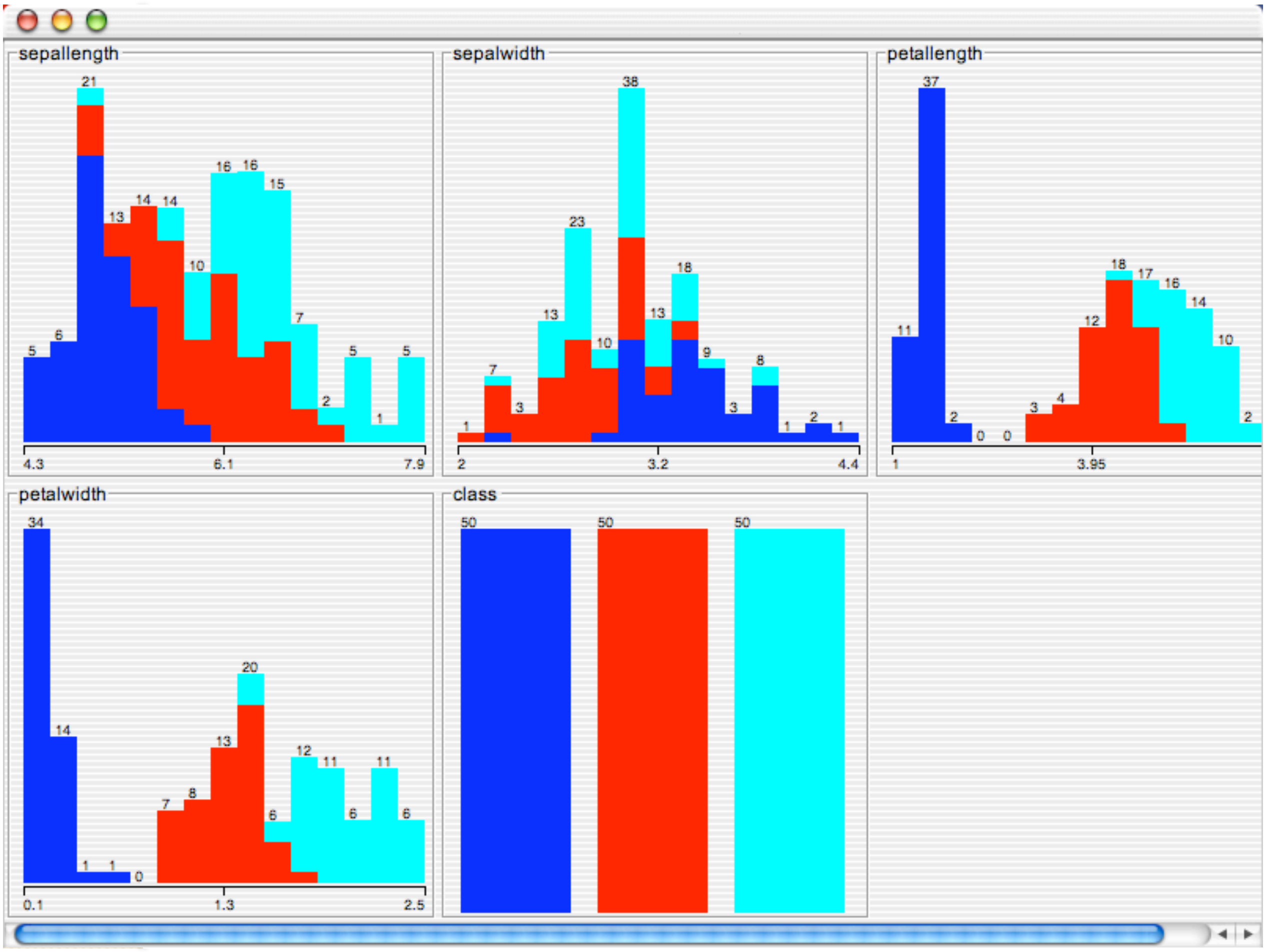

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Attributes
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

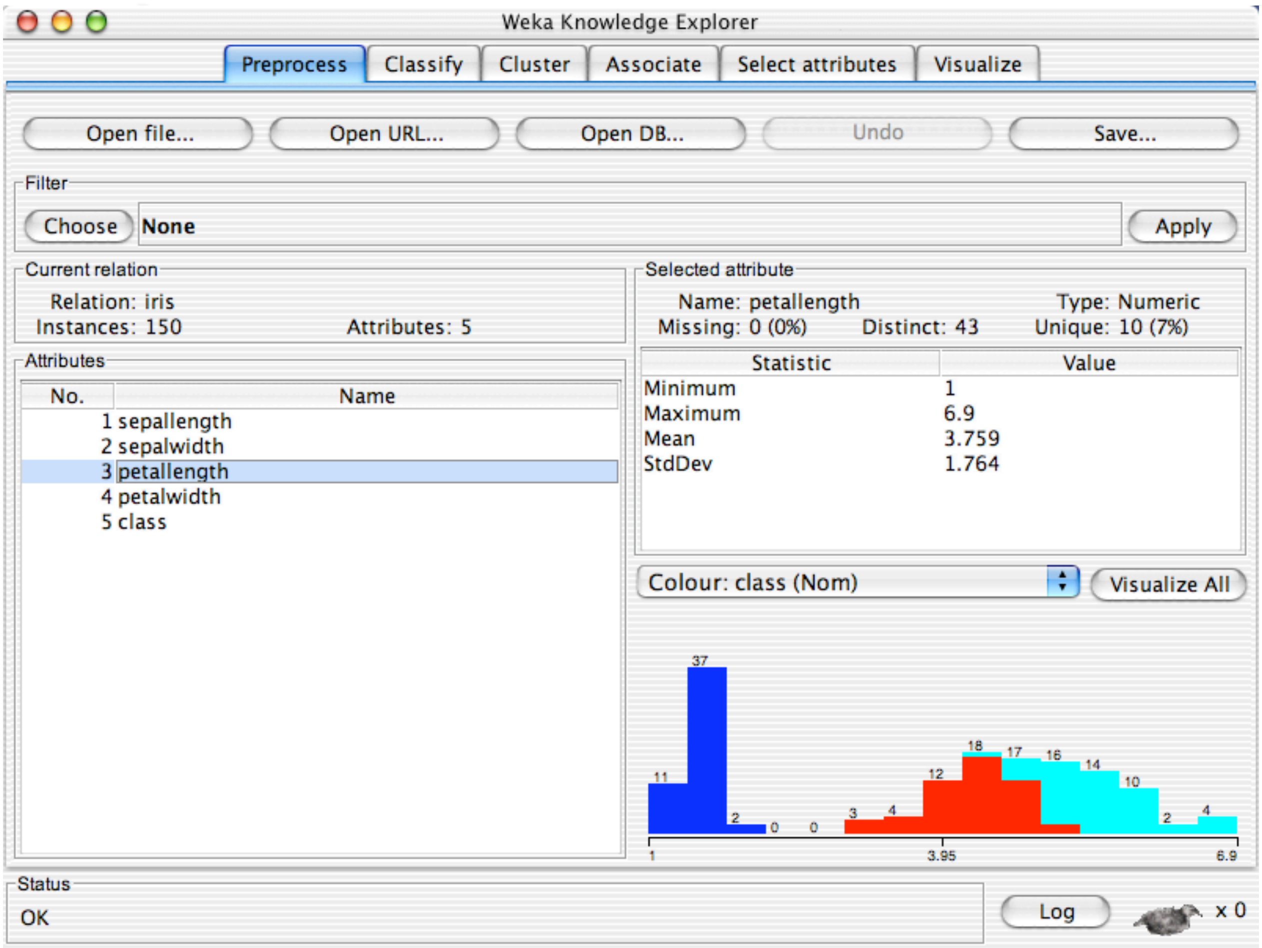

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Statistic
Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Colour: class (Nom)
Visualize All
21
5
6
13
14
14
10
16
16
15
7
2
5
1
5
4.3
6.1
7.9
Status
OK
Log
x 0

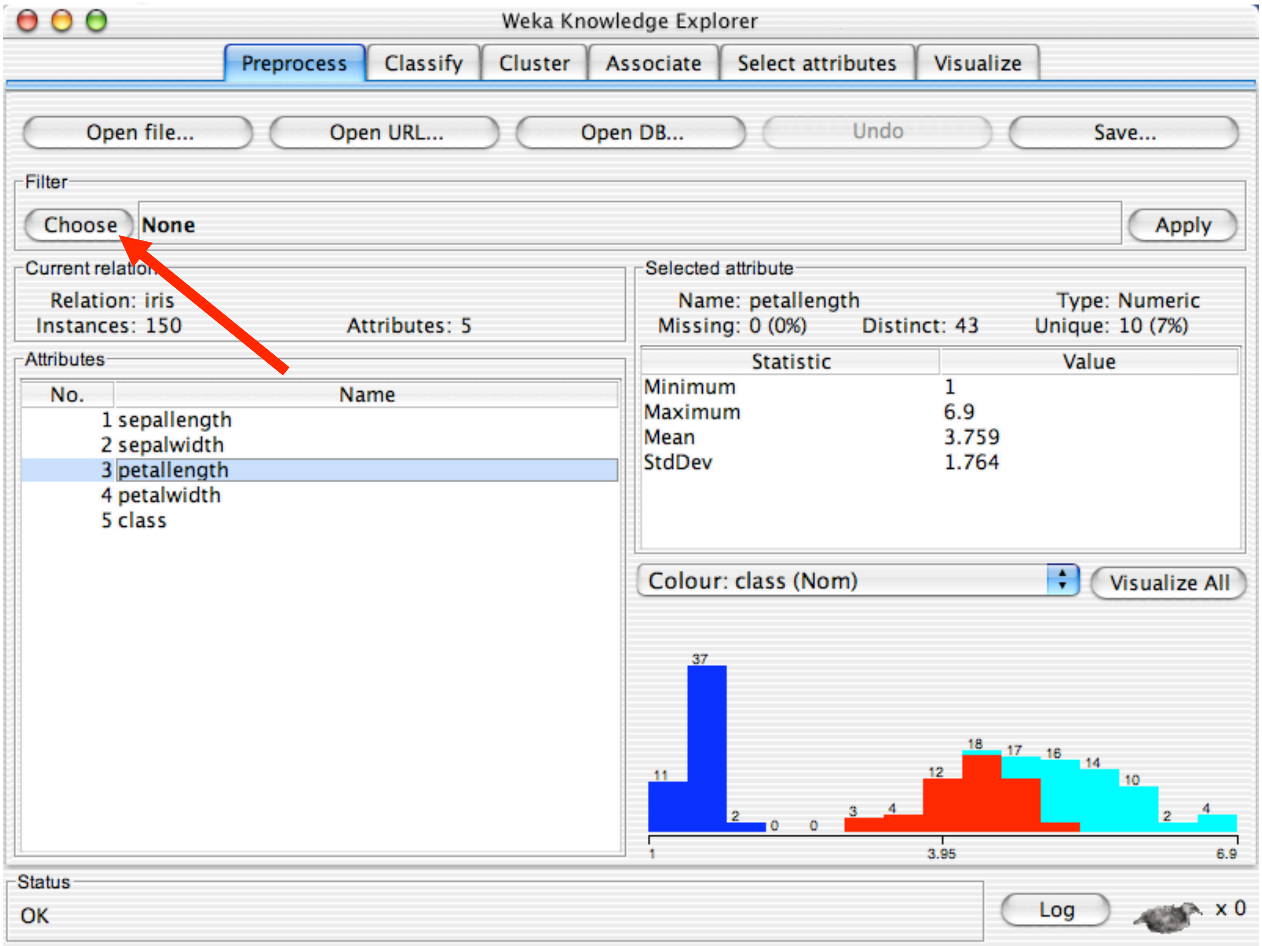

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Statistic
Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Colour: class (Nom)
Visualize All
21
5
6
13
14
14
10
16
16
15
7
2
5
1
5
4.3
6.1
7.9
Status
OK
Log
x 0

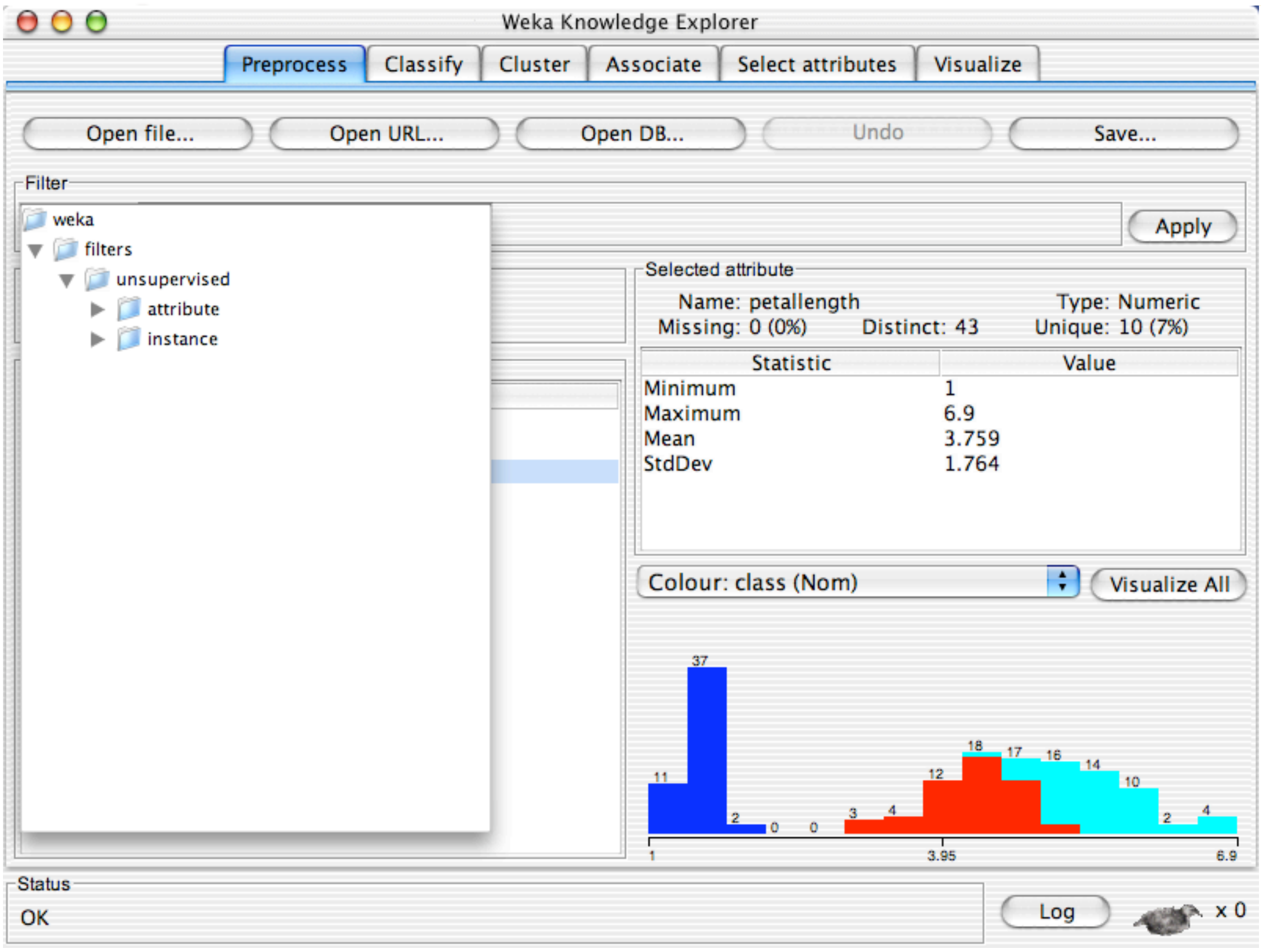

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
Colour: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0

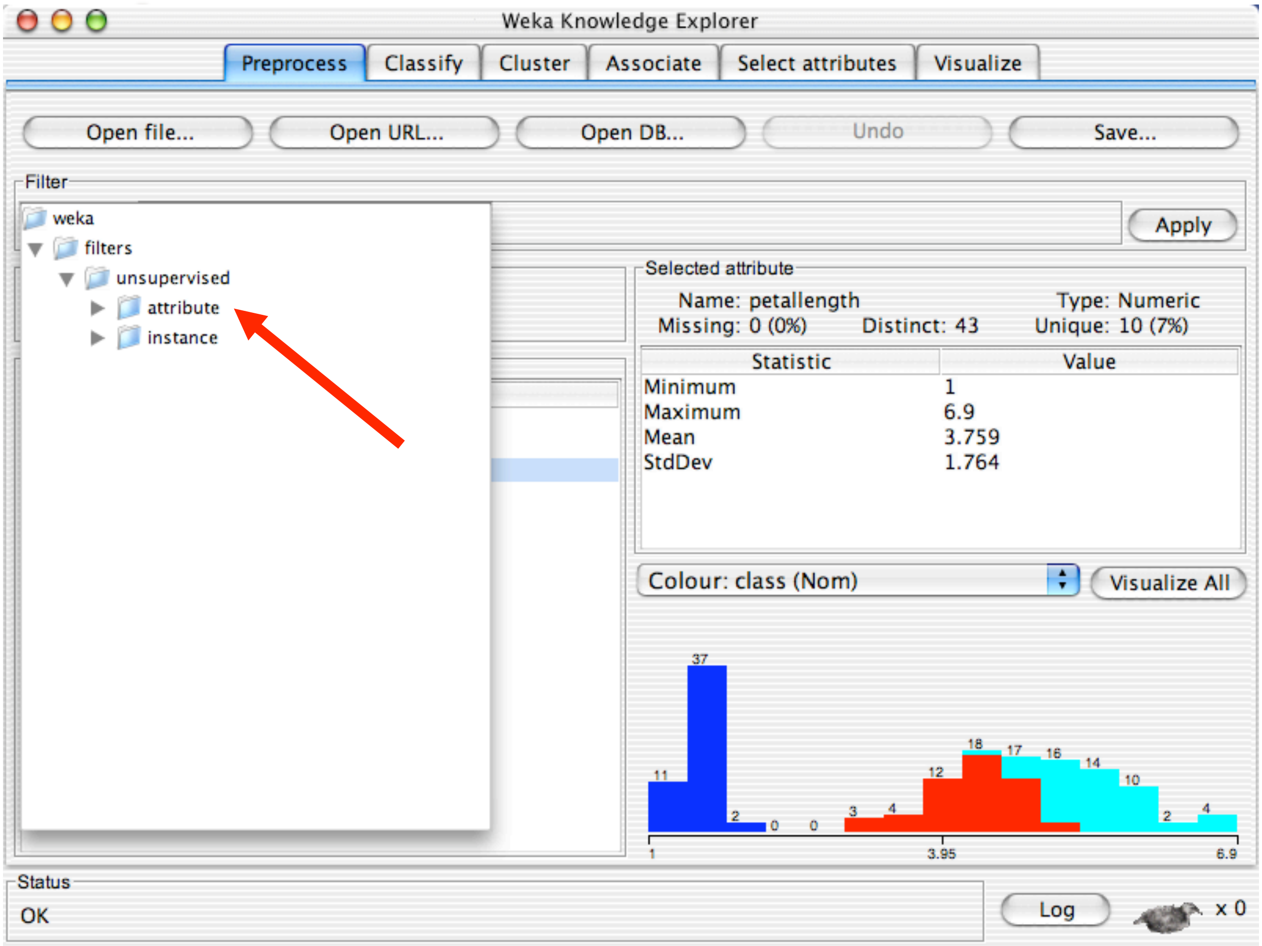

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
Colour: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0

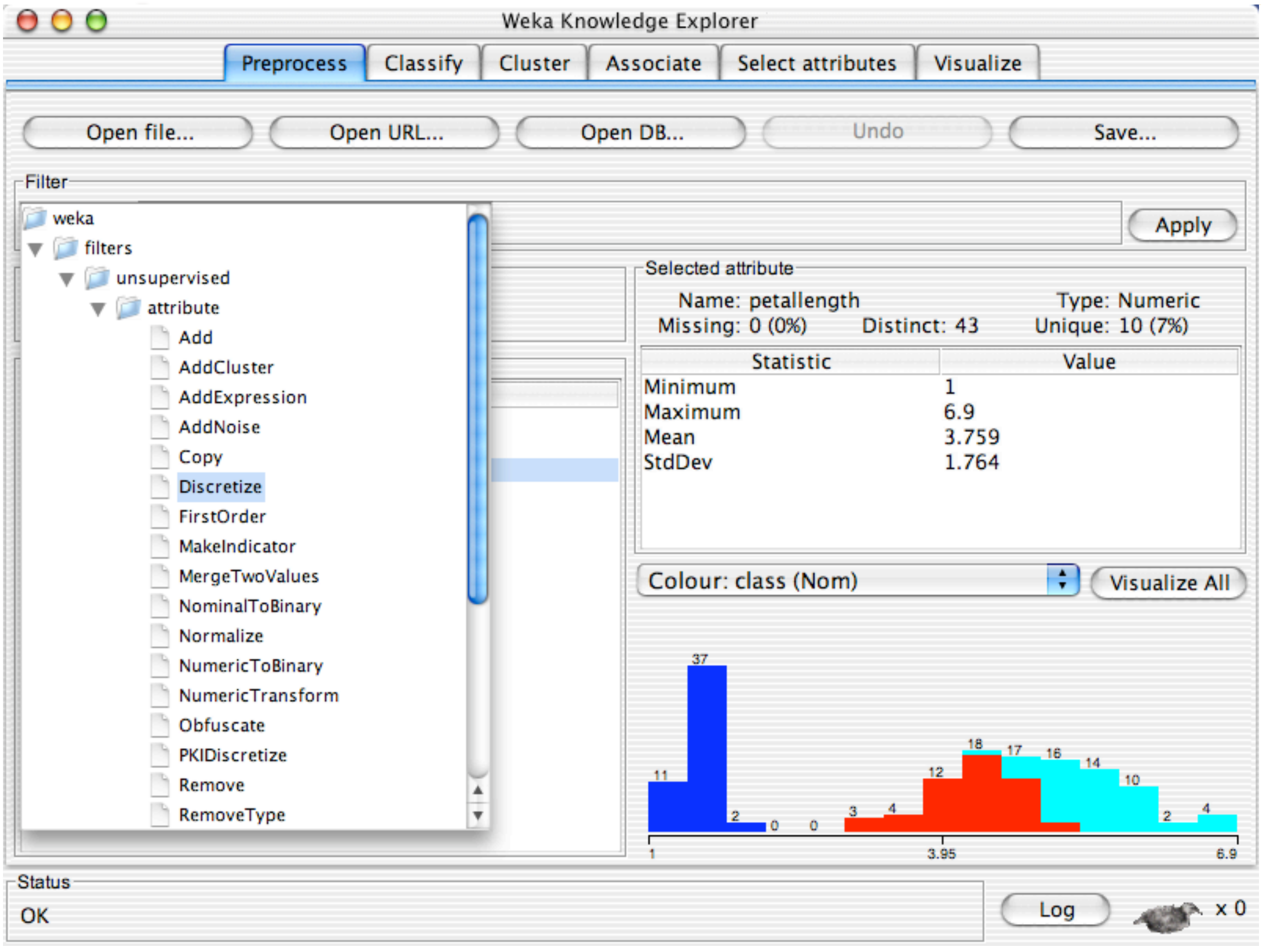

sepallength
4.3
6.1
7.9
sepalwidth
2
3.2
4.4
petallength
1
3.95
petalwidth
0.1
1.3
2.5
class
50
50
50

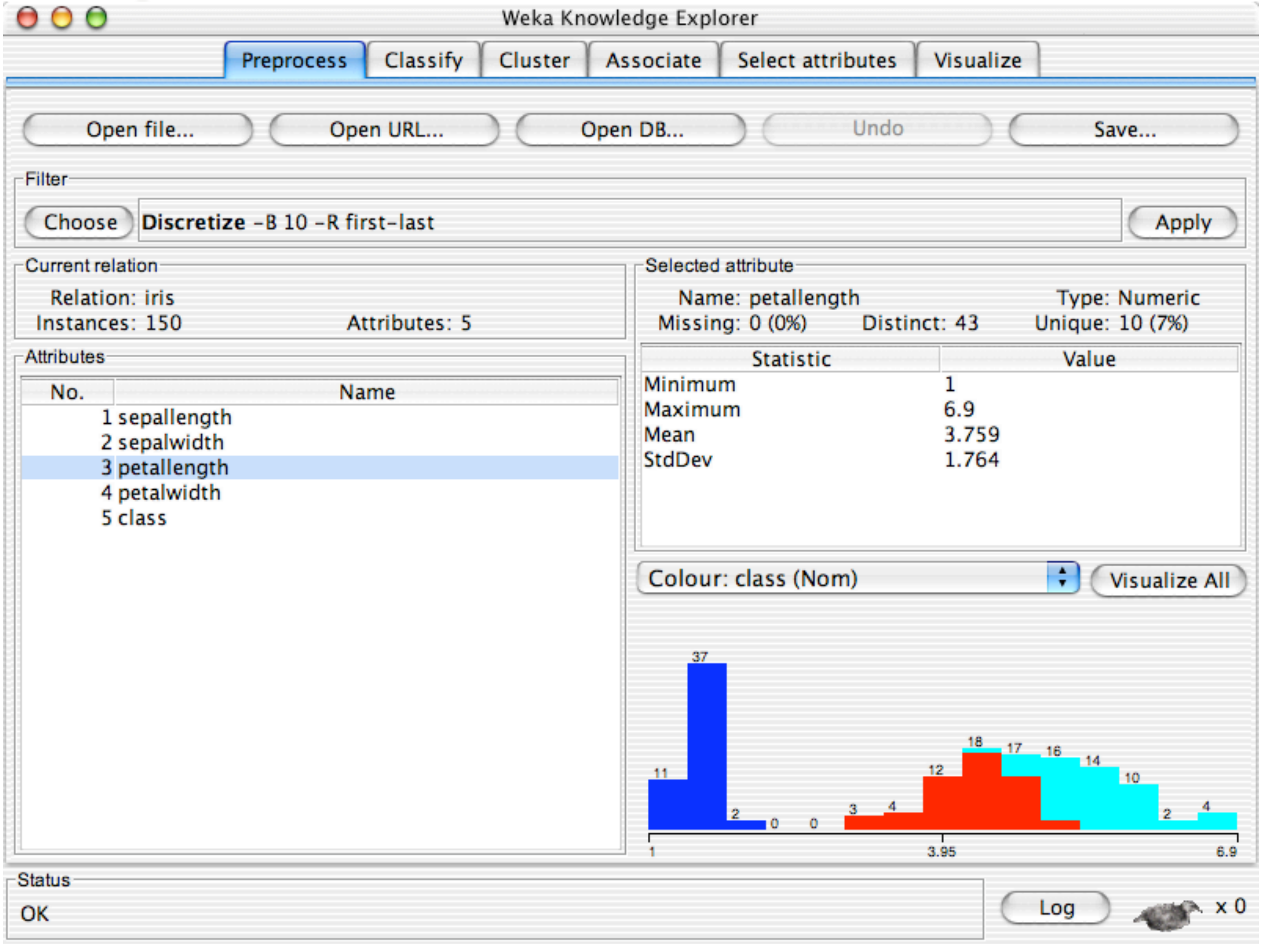

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

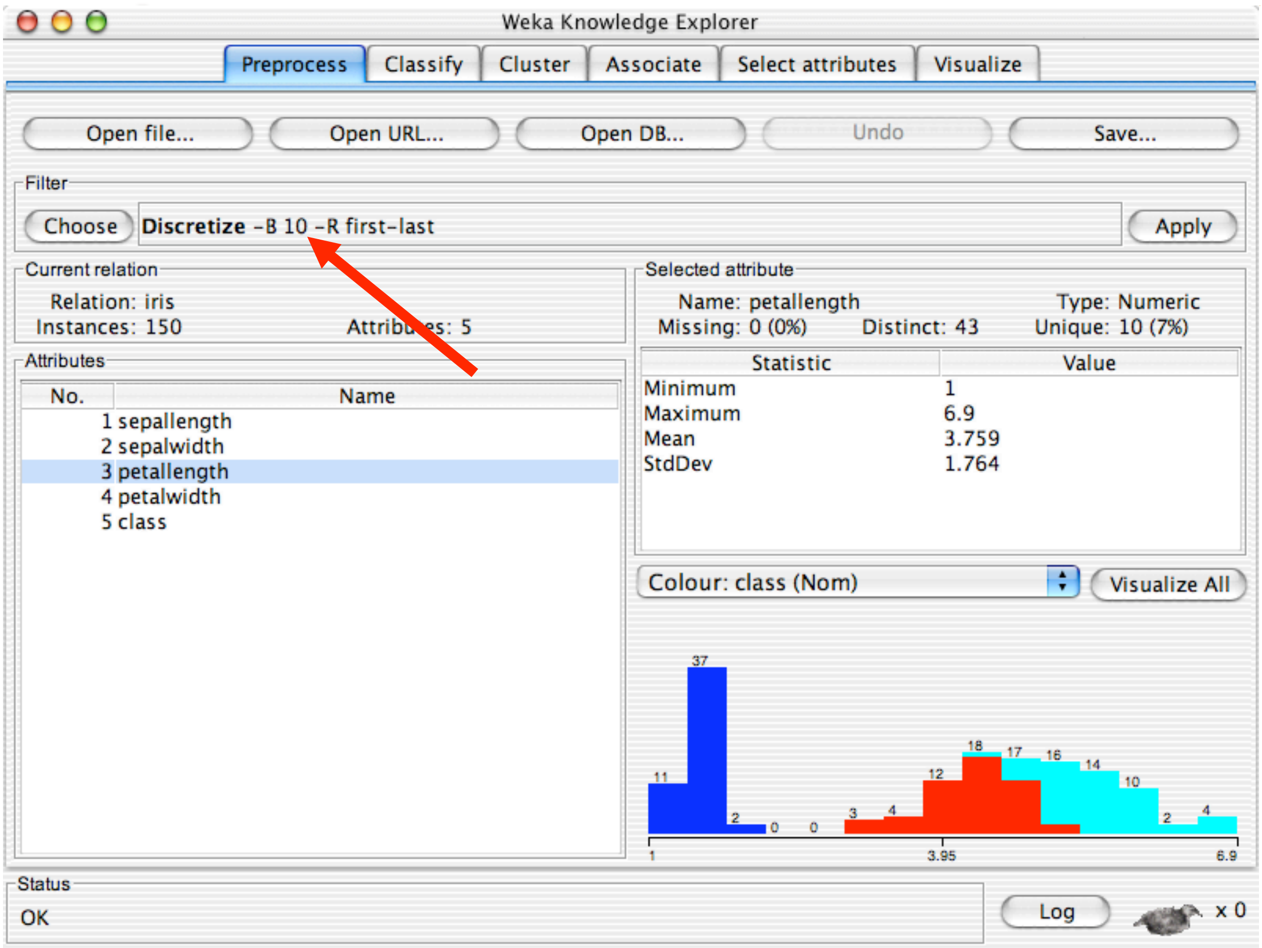

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
instance
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
instance
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

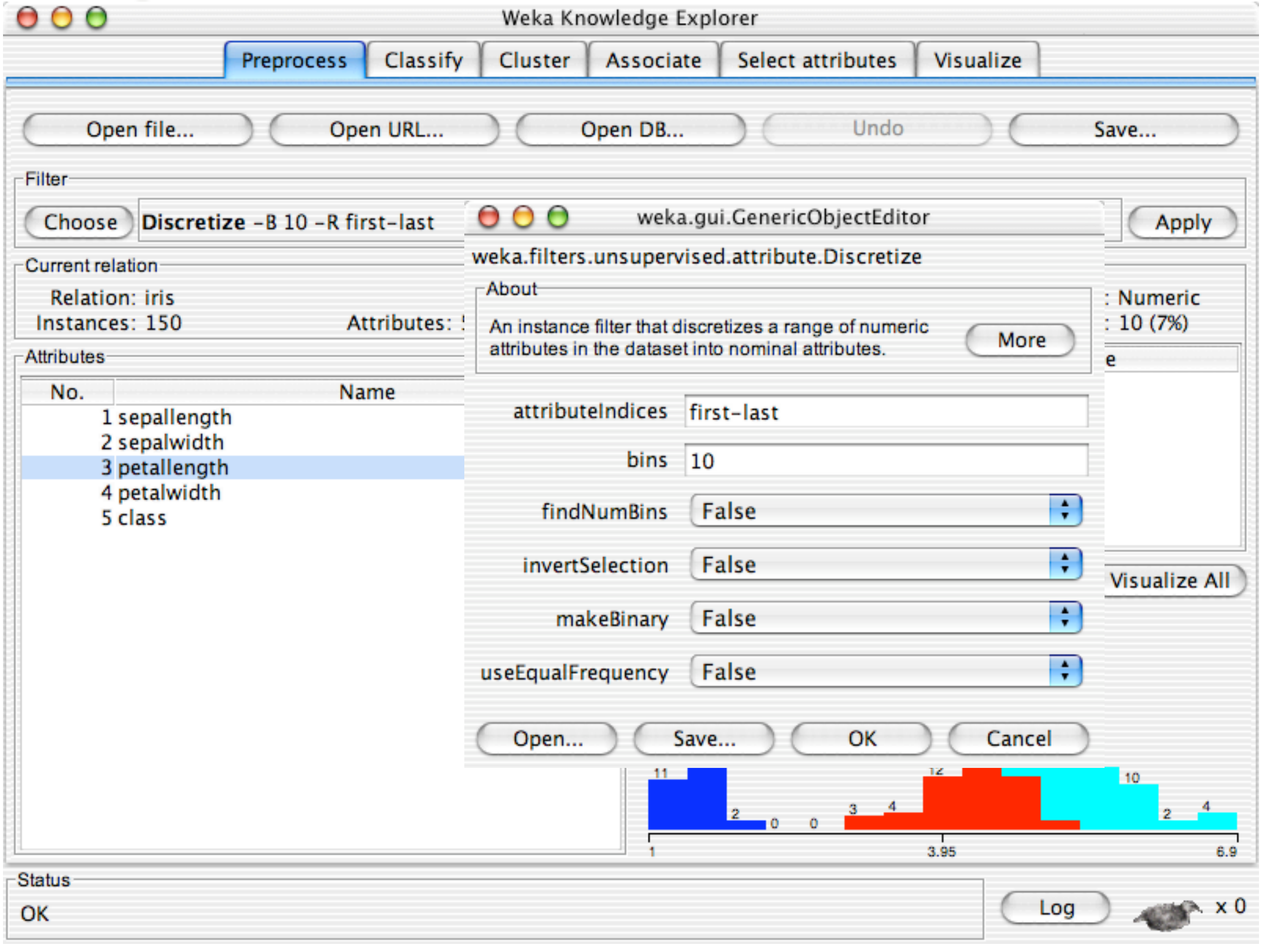

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
Add
AddCluster
AddExpression
AddNoise
Copy
Discretize
FirstOrder
MakeIndicator
MergeTwoValues
NominalToBinary
Normalize
NumericToBinary
NumericTransform
Obfuscate
PKIDiscretize
Remove
RemoveType
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

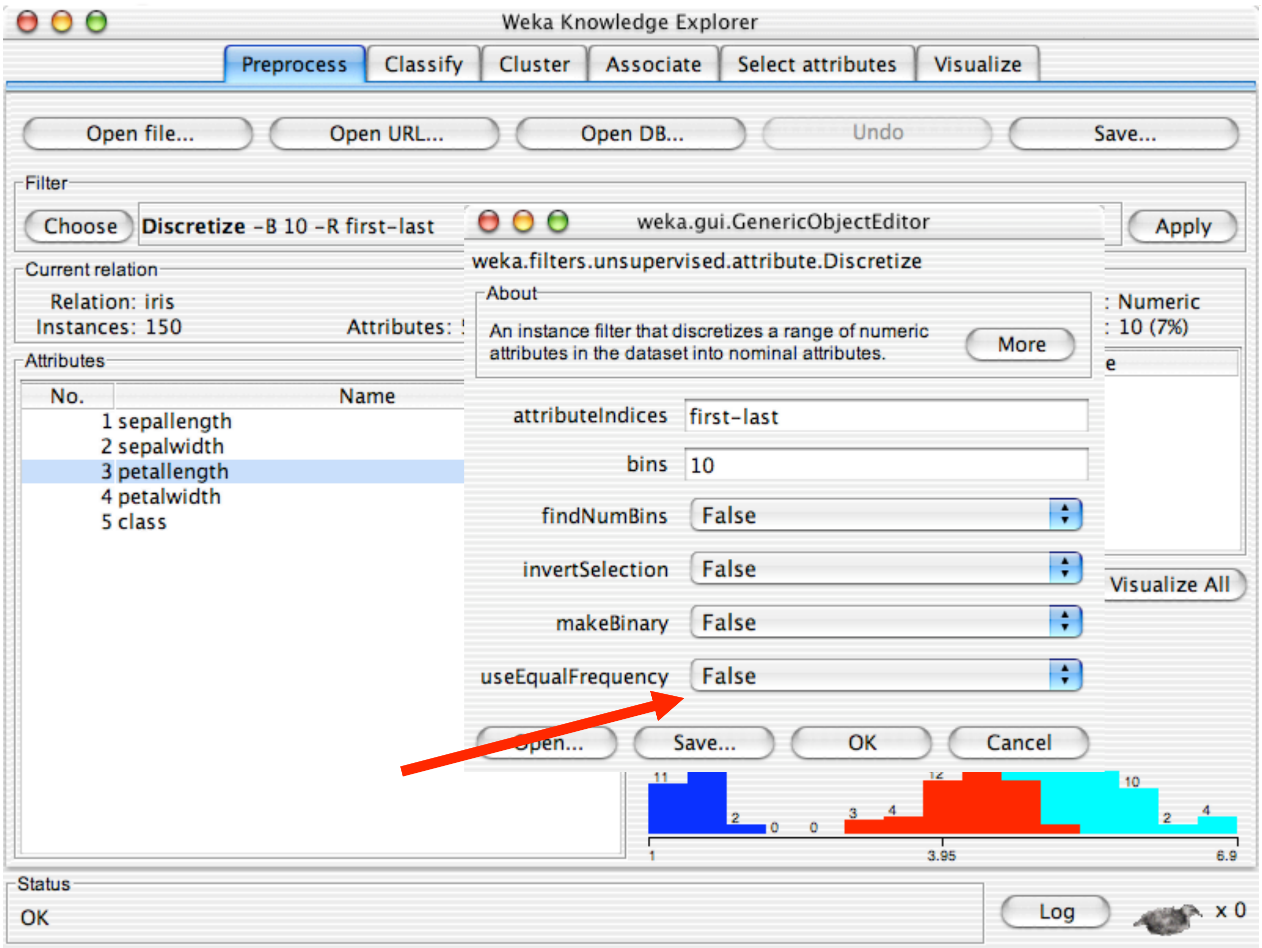


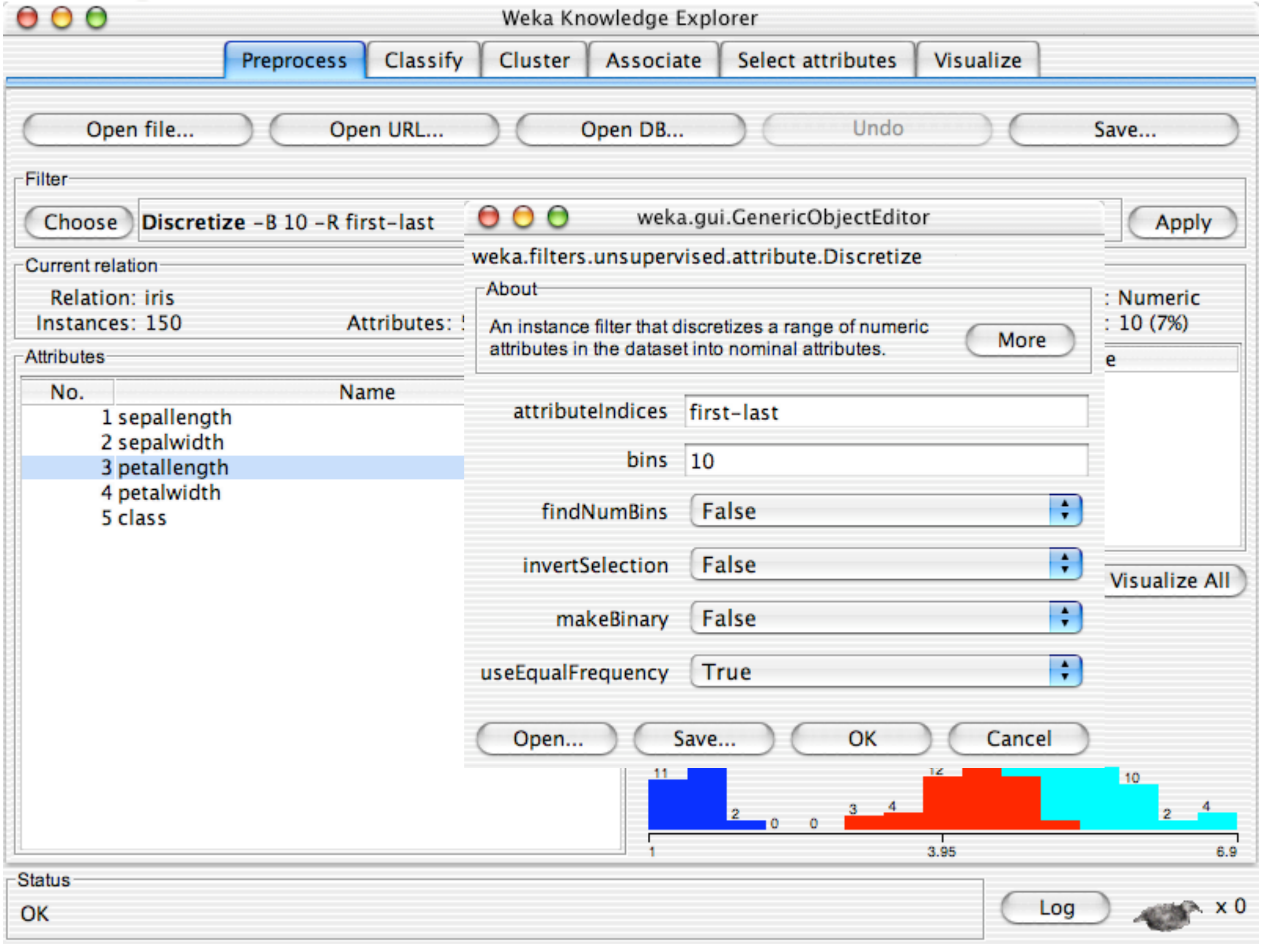
Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

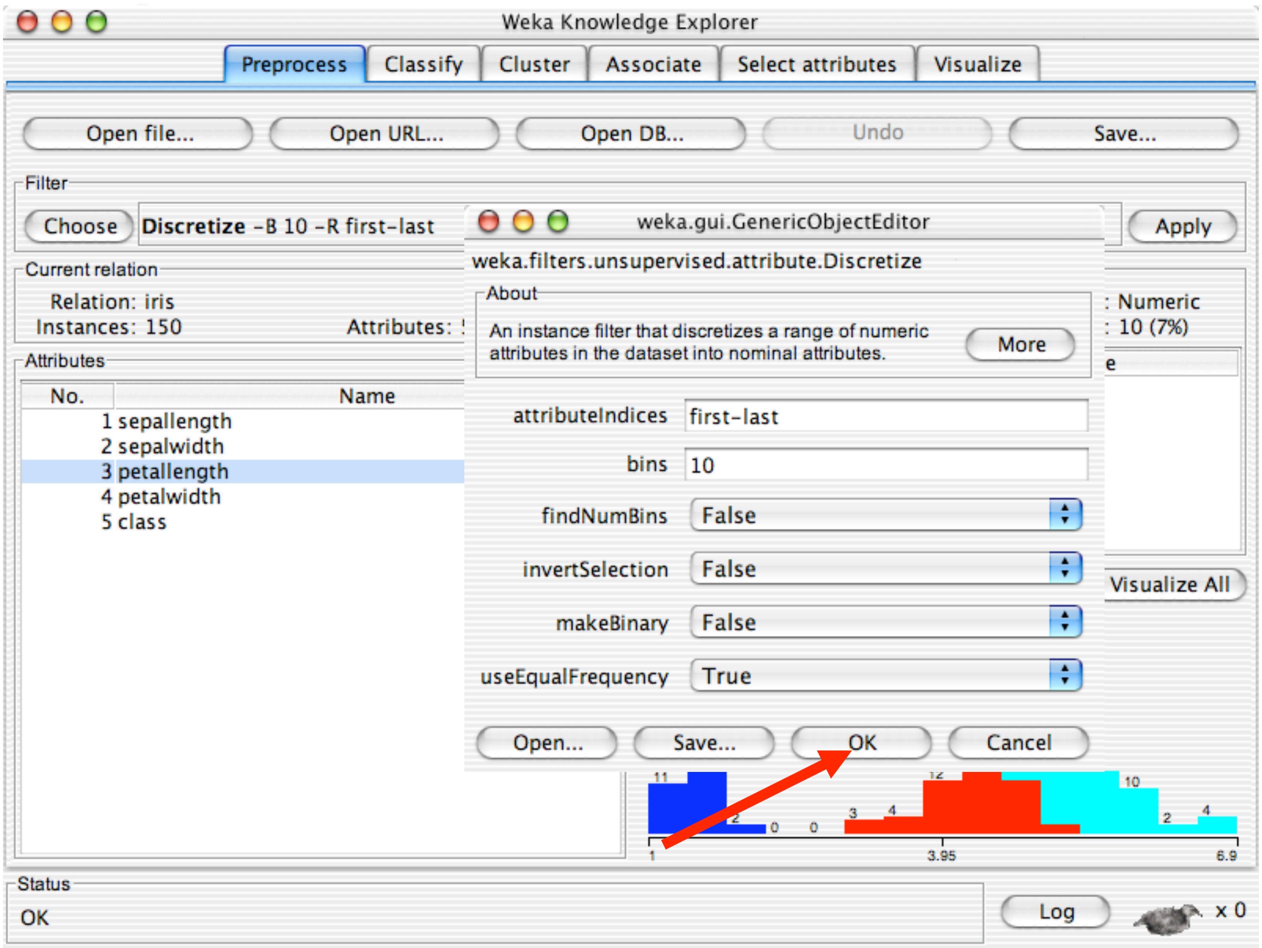


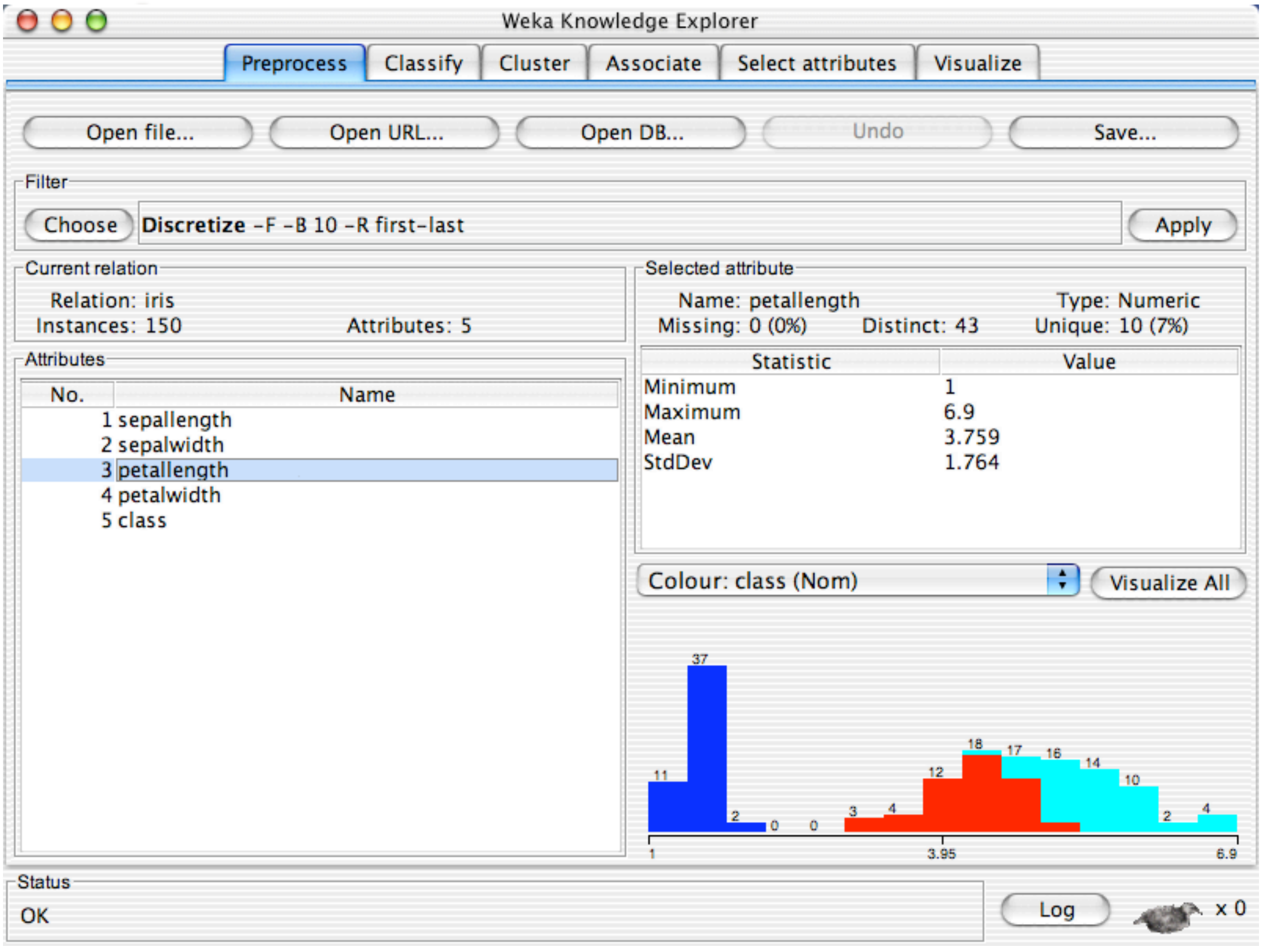
Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

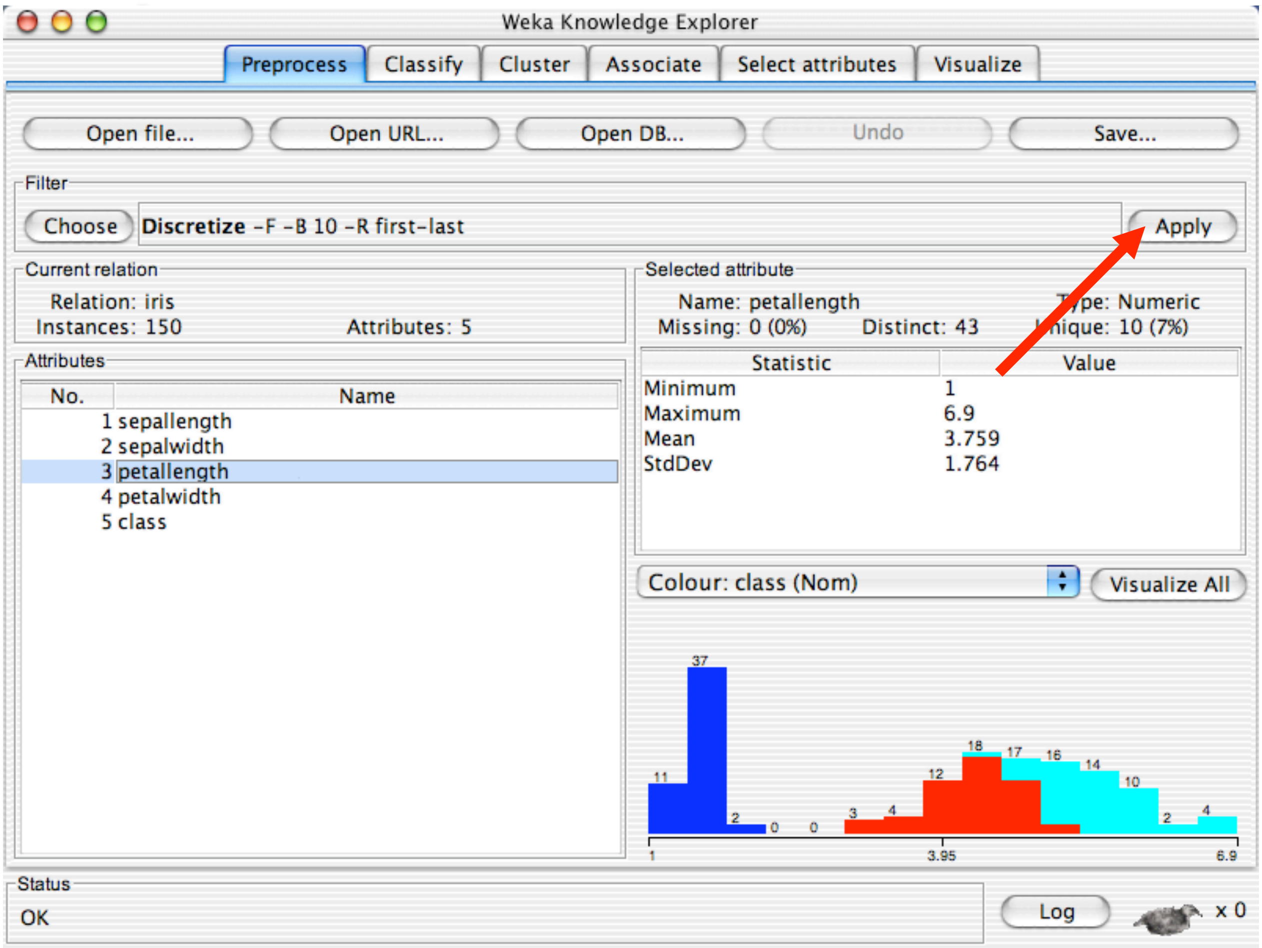

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices first-last
bins 10
findNumBins False
invertSelection False
makeBinary False
useEqualFrequency False
Open...
Save...
OK
Cancel
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
: Numeric
: 10 (7%)
Visualize All
Status
OK
Log
x 0

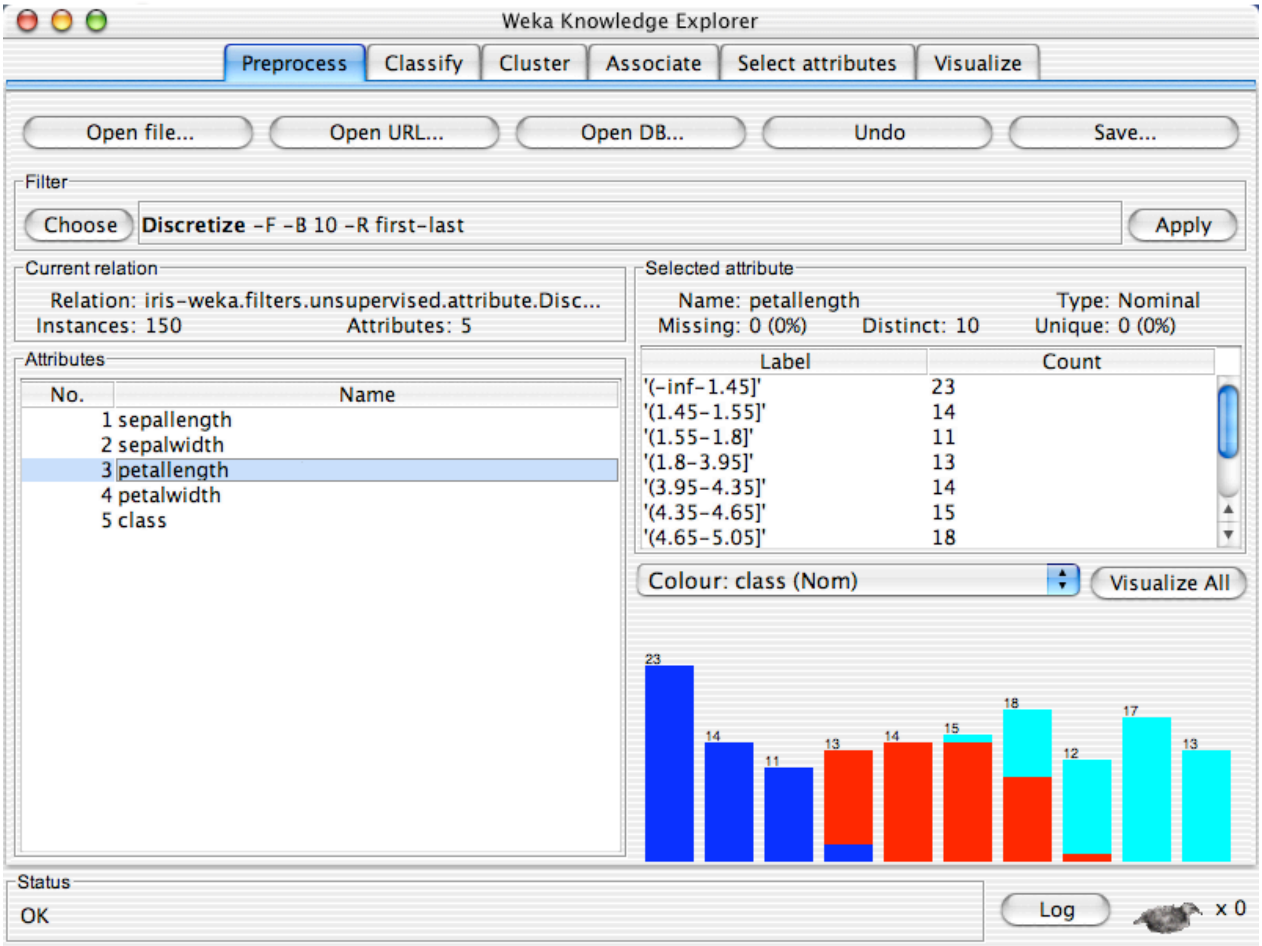

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices first-last
bins 10
findNumBins False
invertSelection False
makeBinary False
useEqualFrequency False
Open...
Save...
OK
Cancel
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
: Numeric
: 10 (7%)
Visualize All
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
Status
OK
Log
x 0

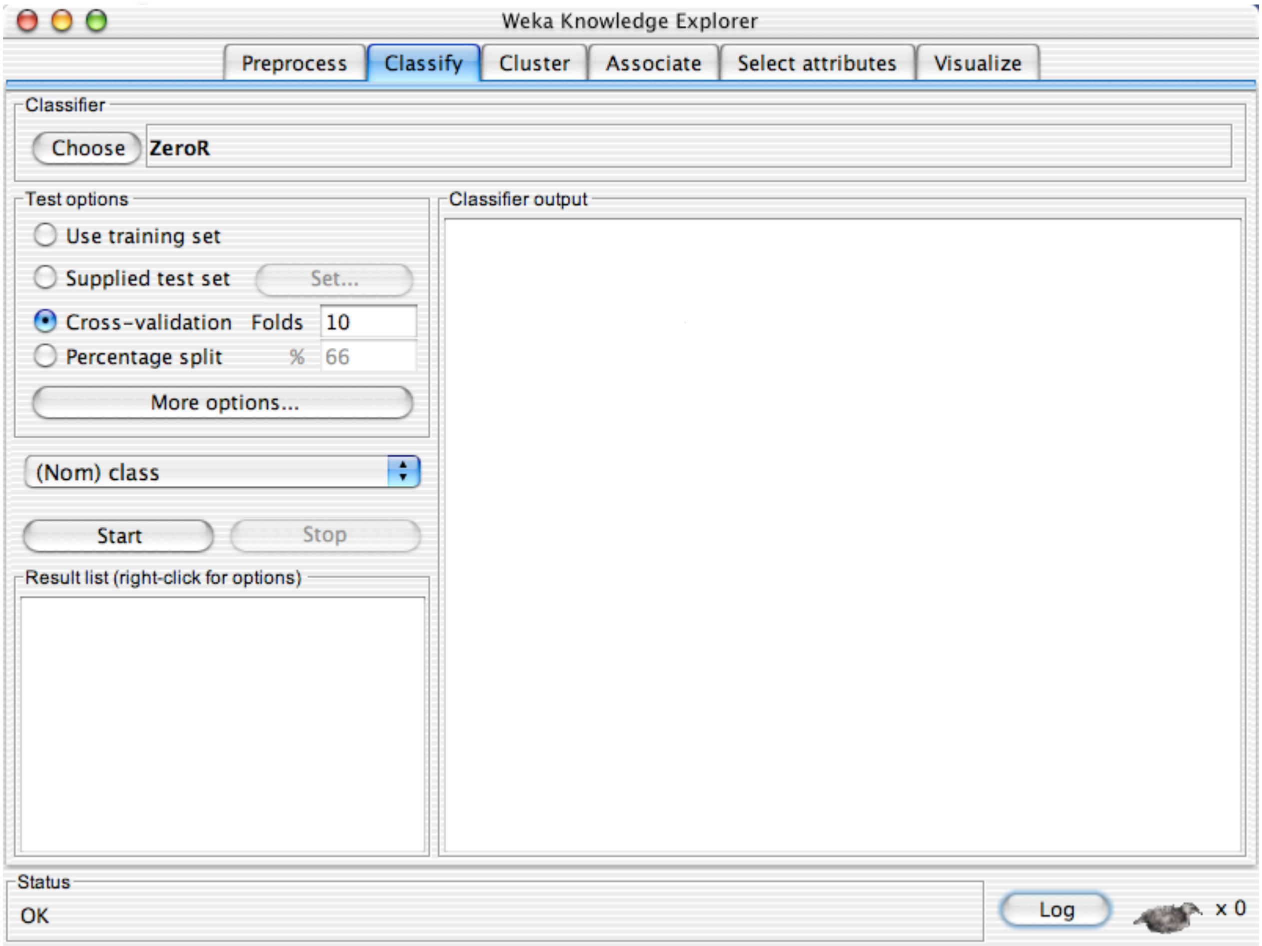

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -F -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -F -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Missing: 0 (0%)
Distinct: 43
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum 1
Maximum 6.9
Mean 3.759
StdDev 1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

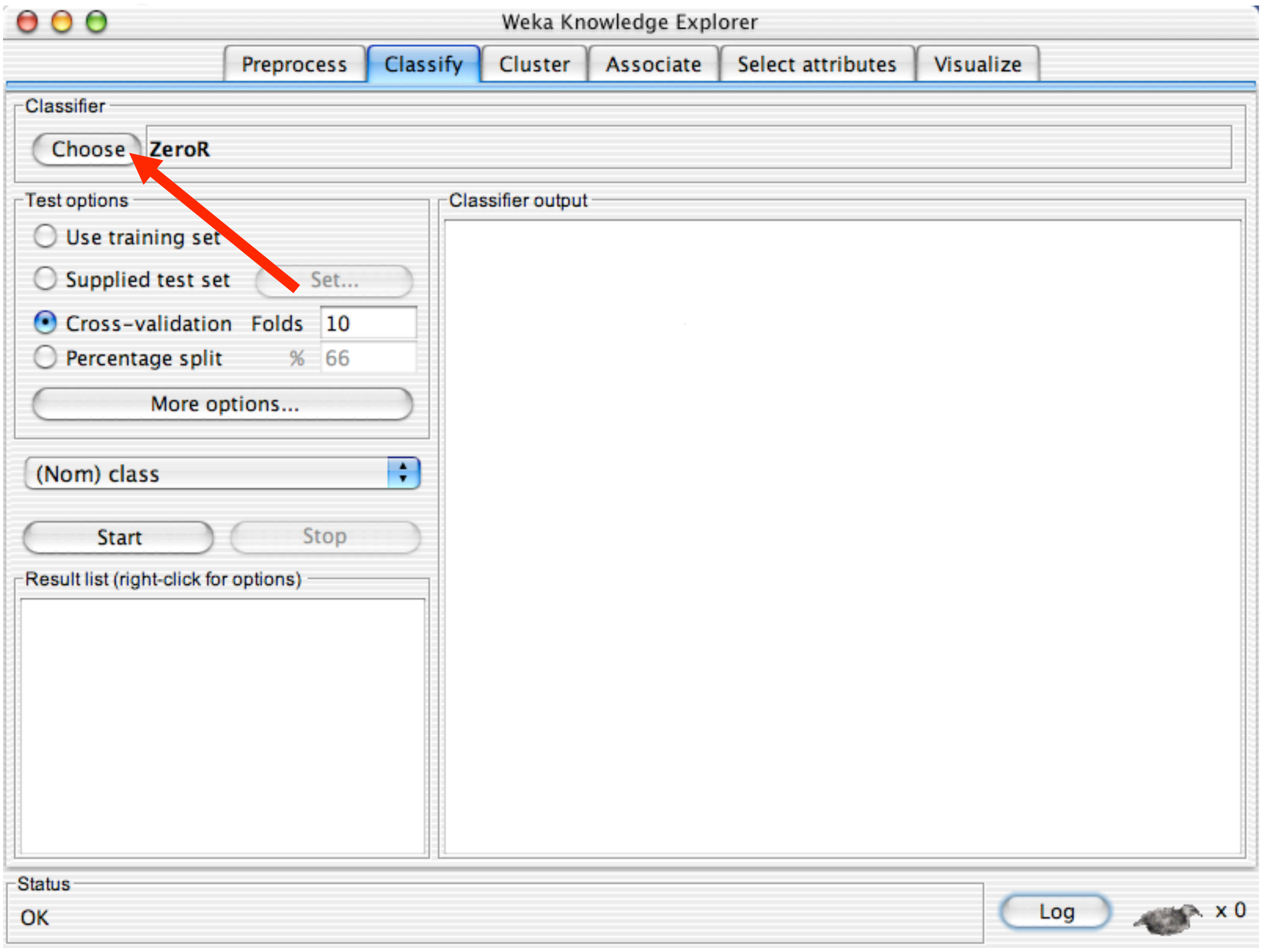


Explorer: building "classifiers"

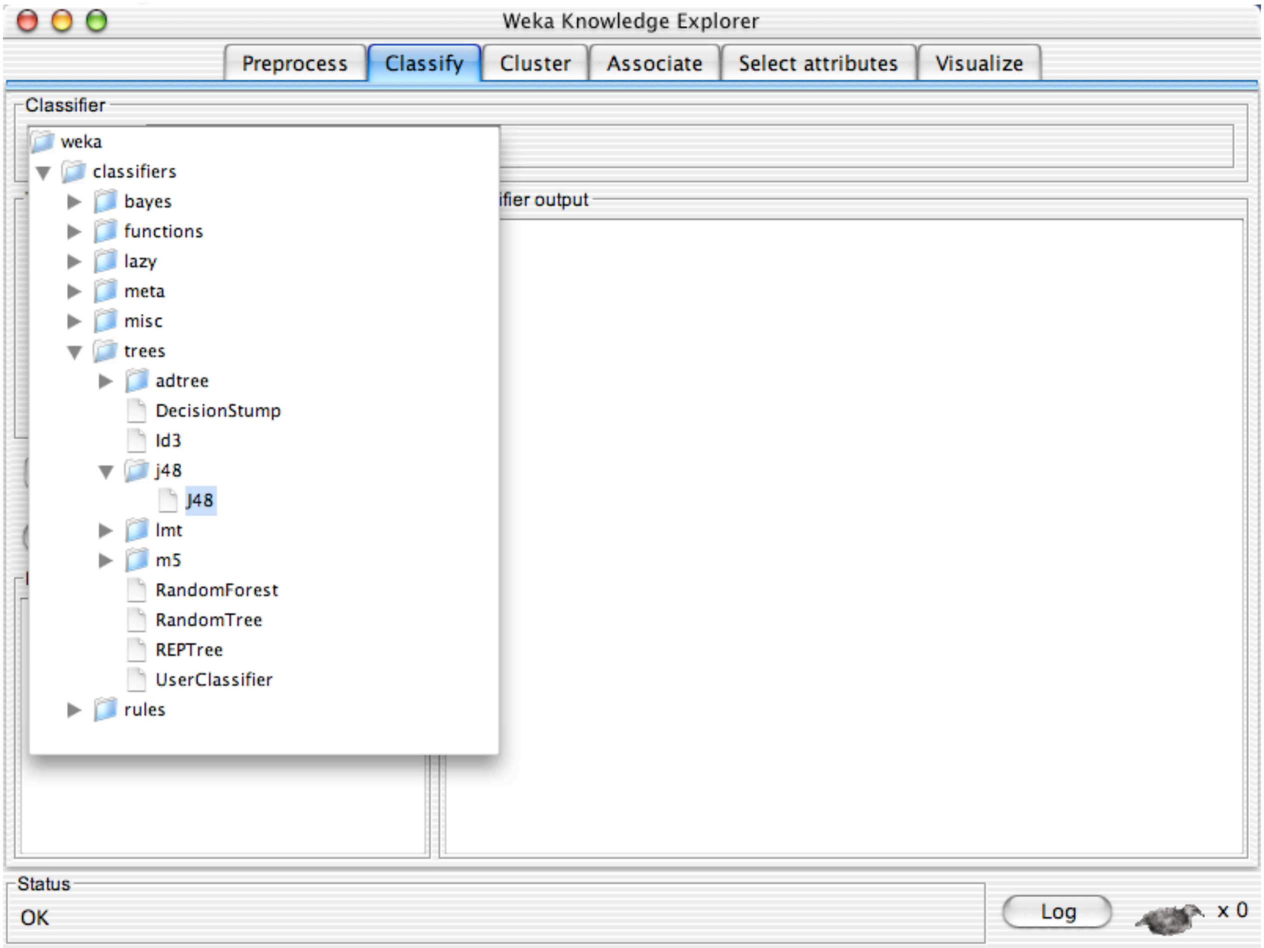
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, …
- "Meta"-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, …

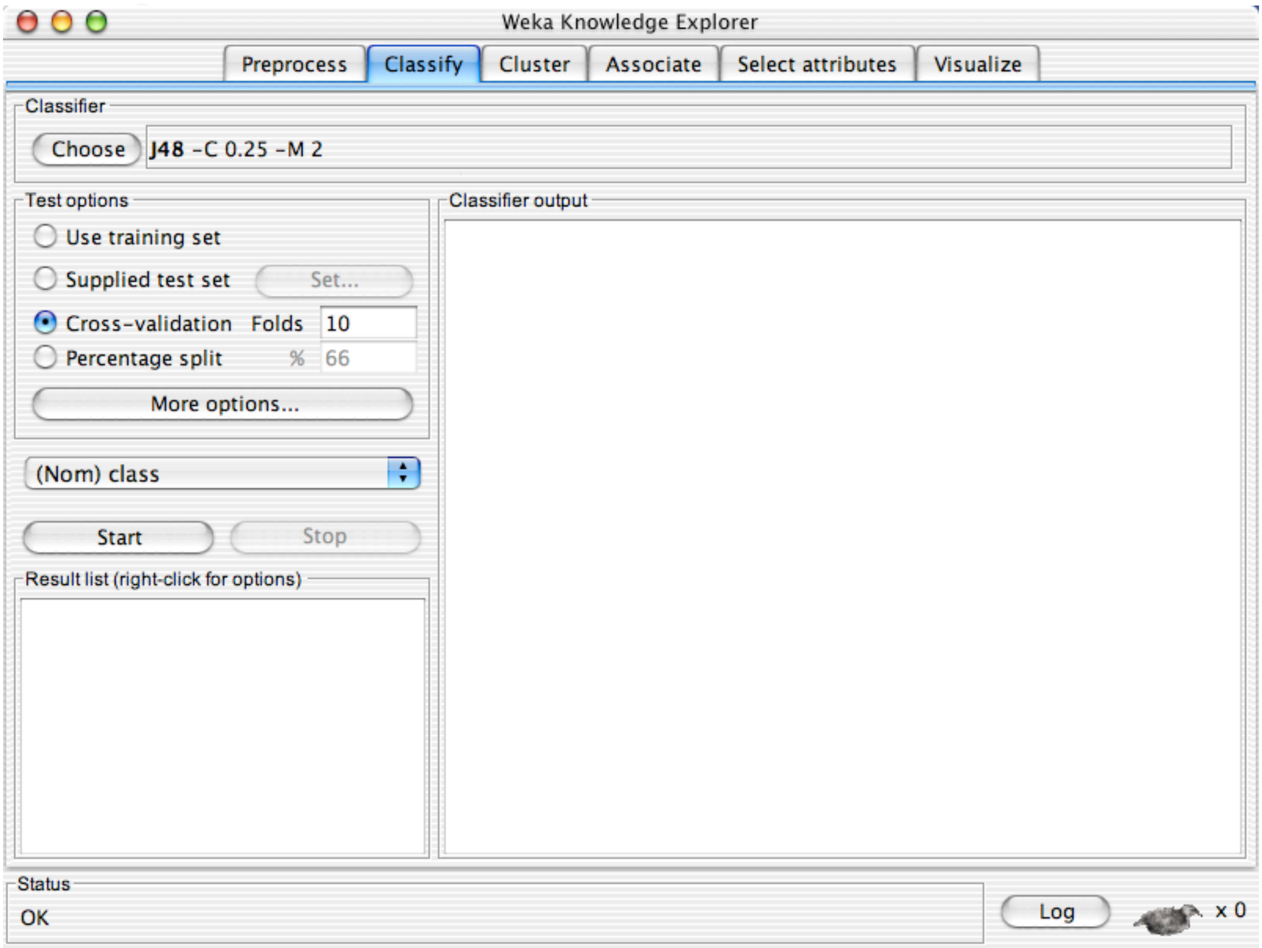


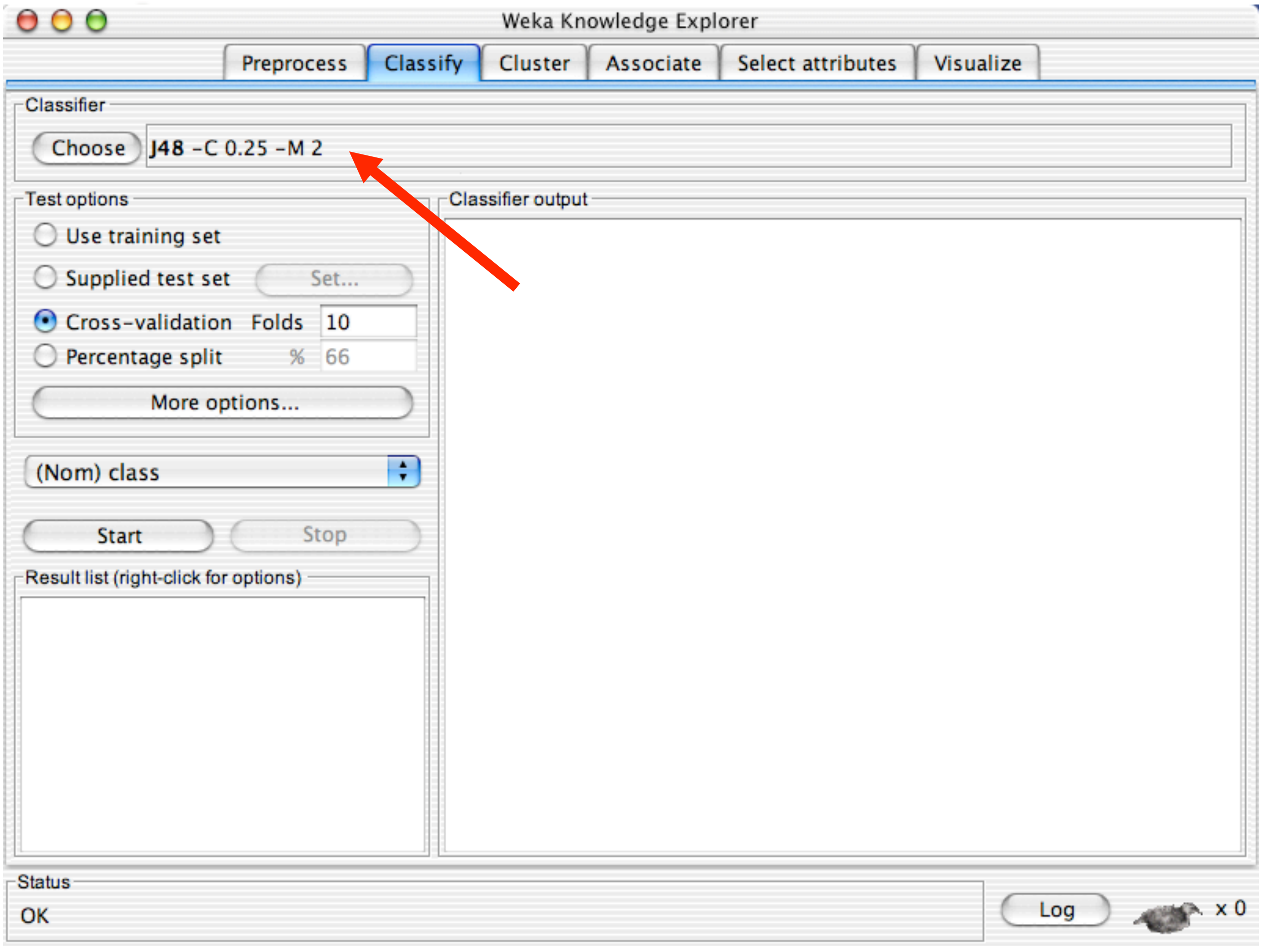
Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

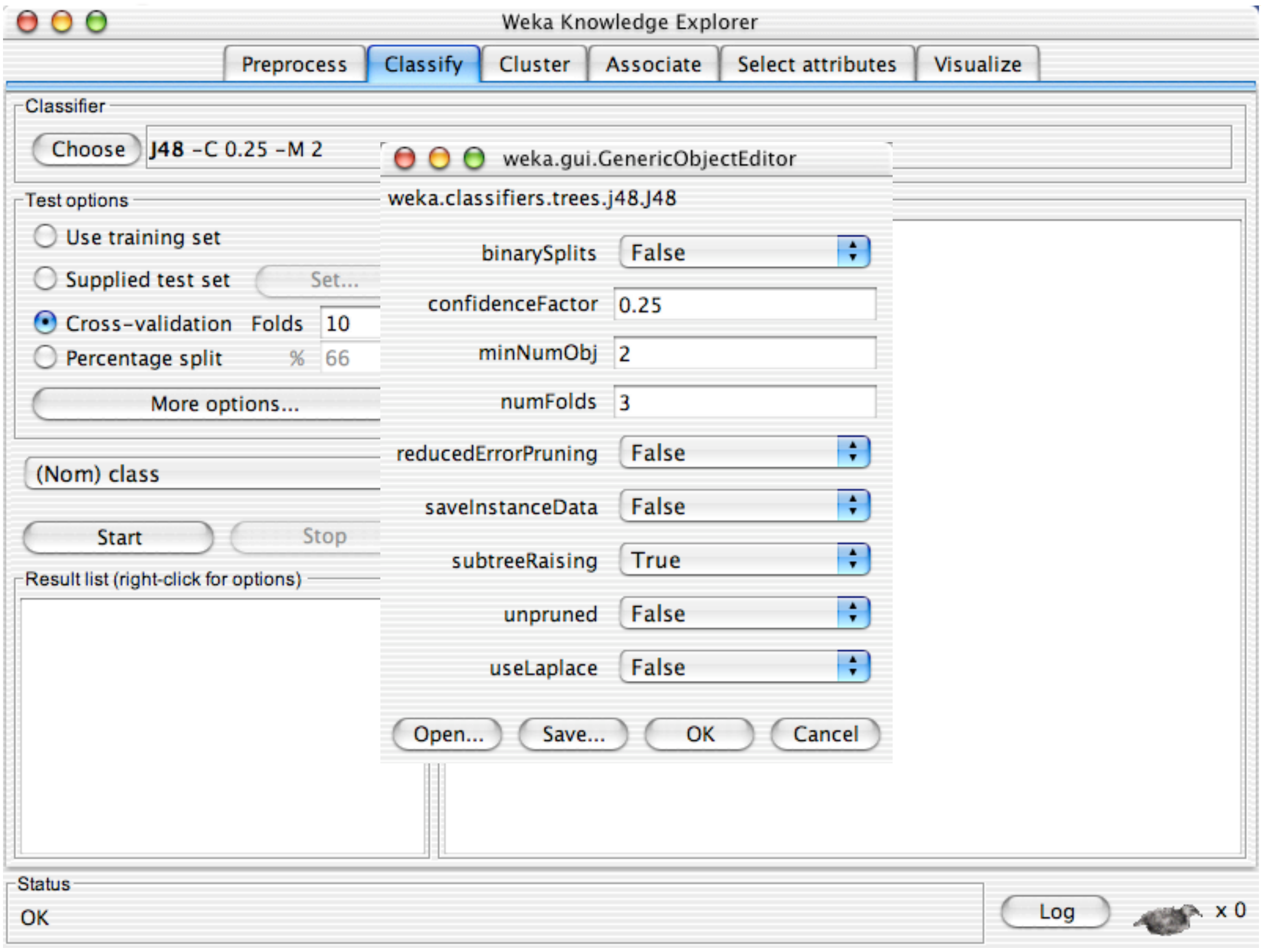


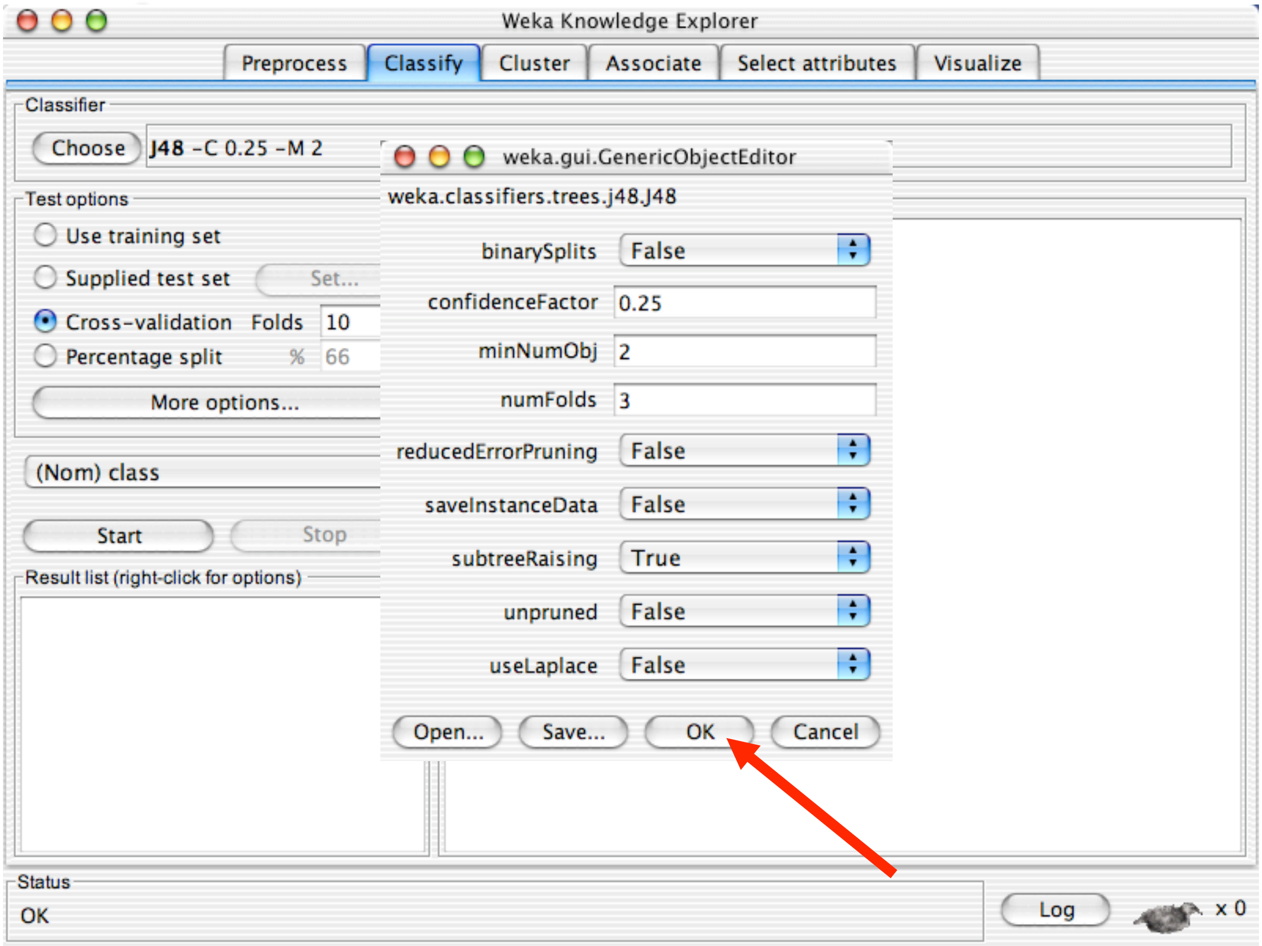
Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

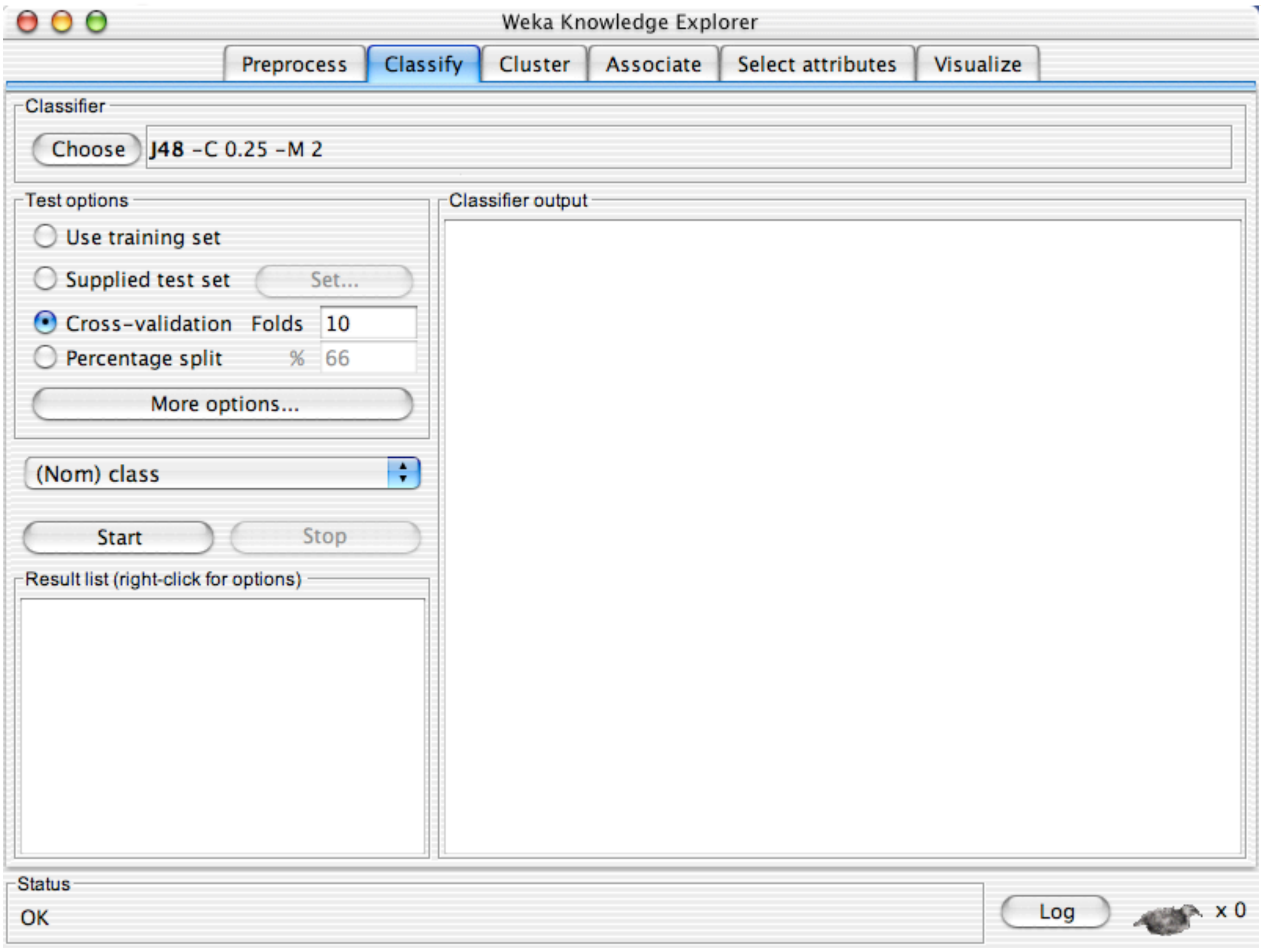

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
adtree
DecisionStump
Id3
j48
J48
lmt
m5
RandomForest
RandomTree
REPTree
UserClassifier
rules
ifier output
Status
OK
Log
x 0

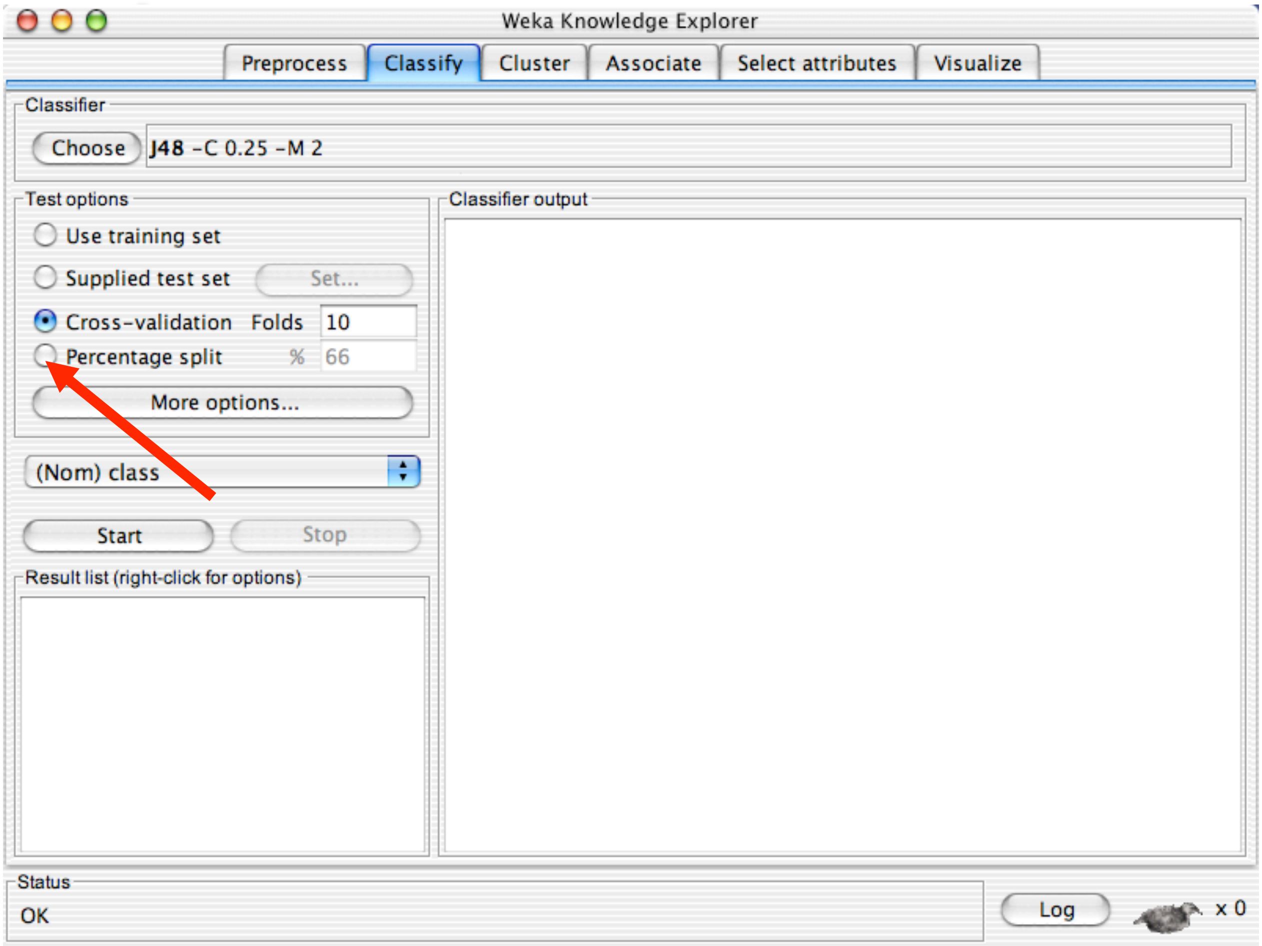

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

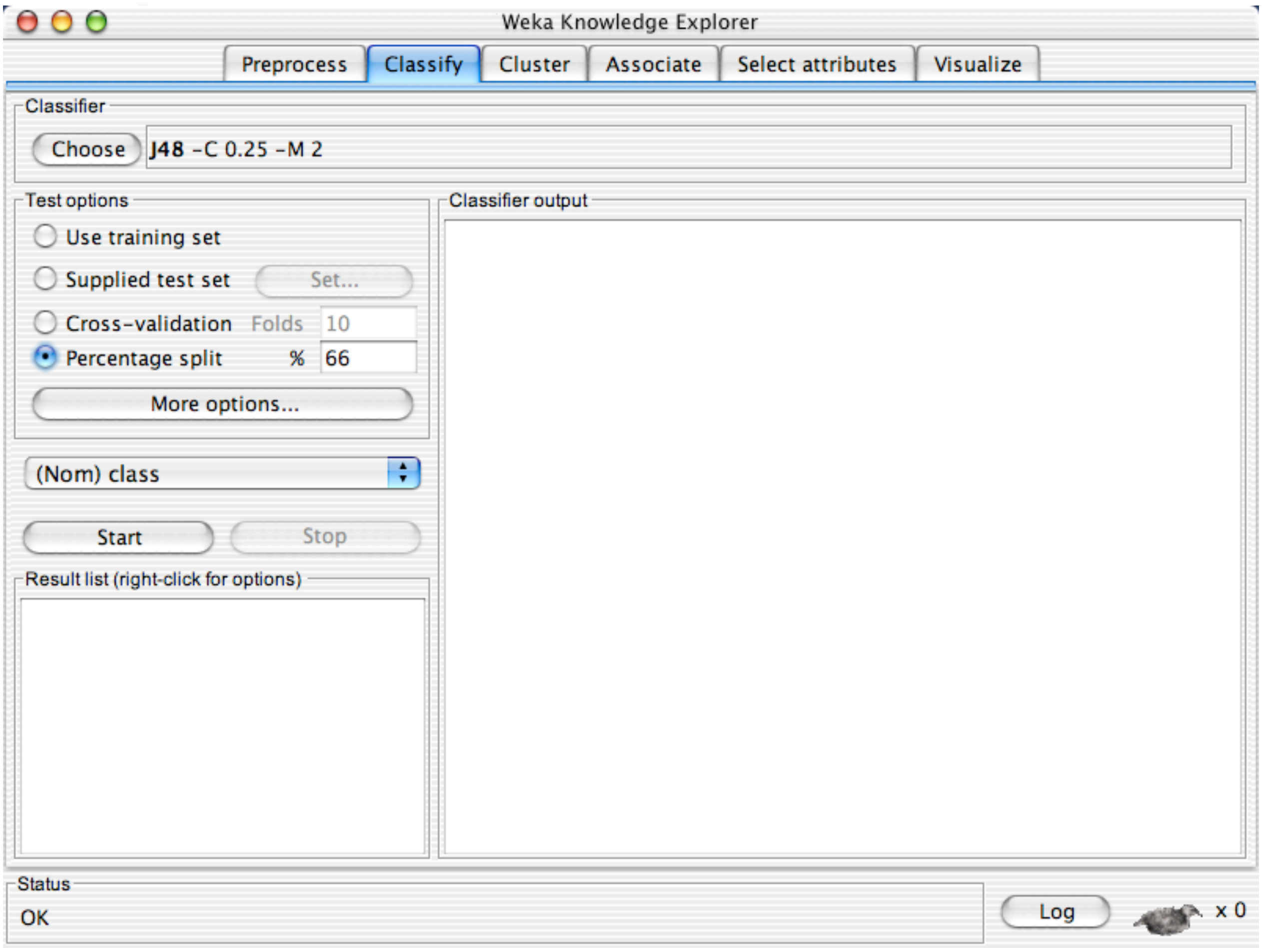

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

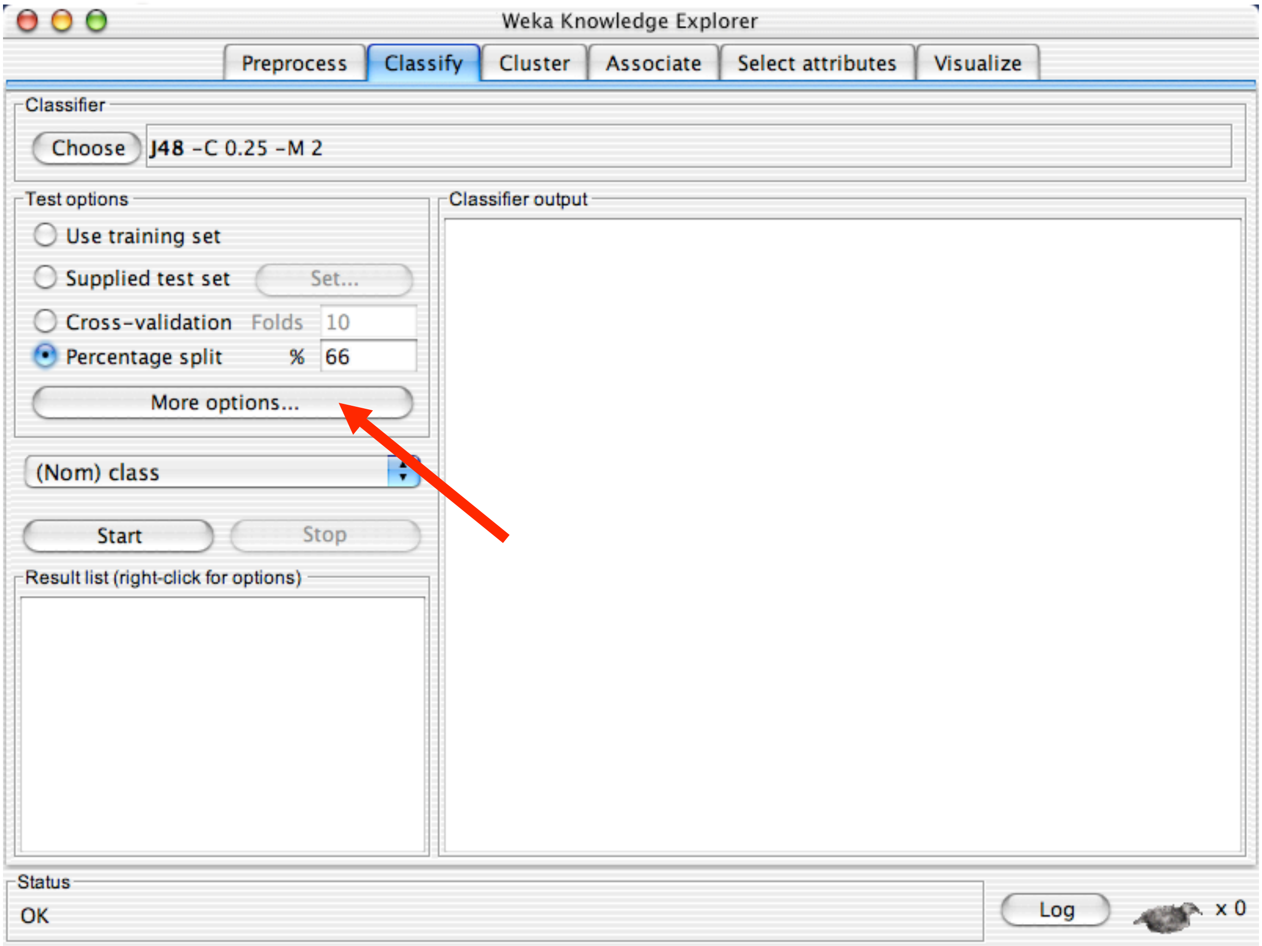

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
weka.gui.GenericObjectEditor
weka.classifiers.trees.j48.J48
binarySplits
False
confidenceFactor
0.25
minNumObj
2
numFolds
3
reducedErrorPruning
False
saveInstanceData
False
subtreeRaising
True
unpruned
False
useLaplace
False
Open...
Save...
OK
Cancel
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)
Status
OK
Log
x 0

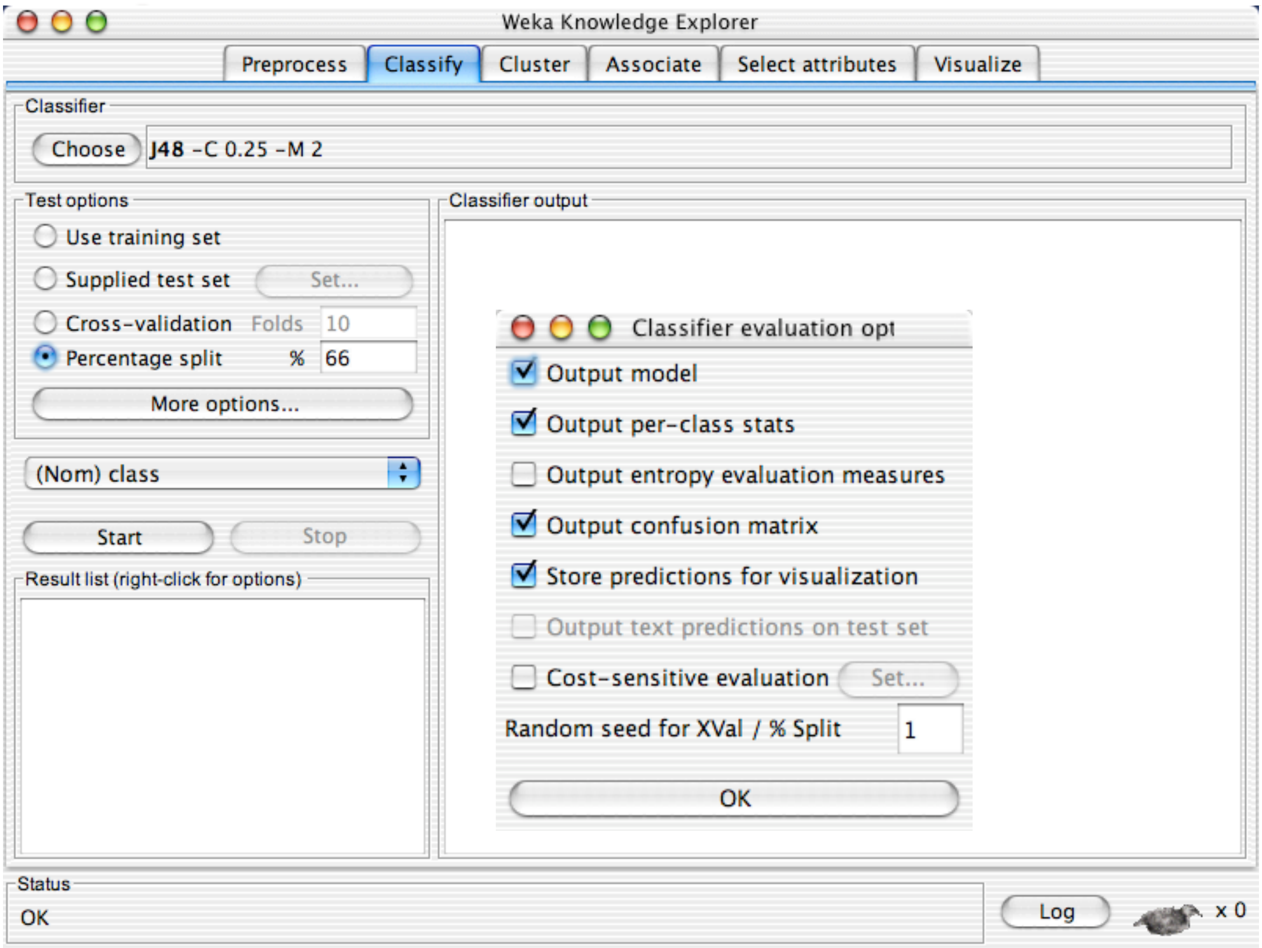

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
weka.gui.GenericObjectEditor
weka.classifiers.trees.j48.J48
binarySplits False
confidenceFactor 0.25
minNumObj 2
numFolds 3
reducedErrorPruning False
saveInstanceData False
subtreeRaising True
unpruned False
useLaplace False
Open...
Save...
OK
Cancel
Status
OK
Log
x 0

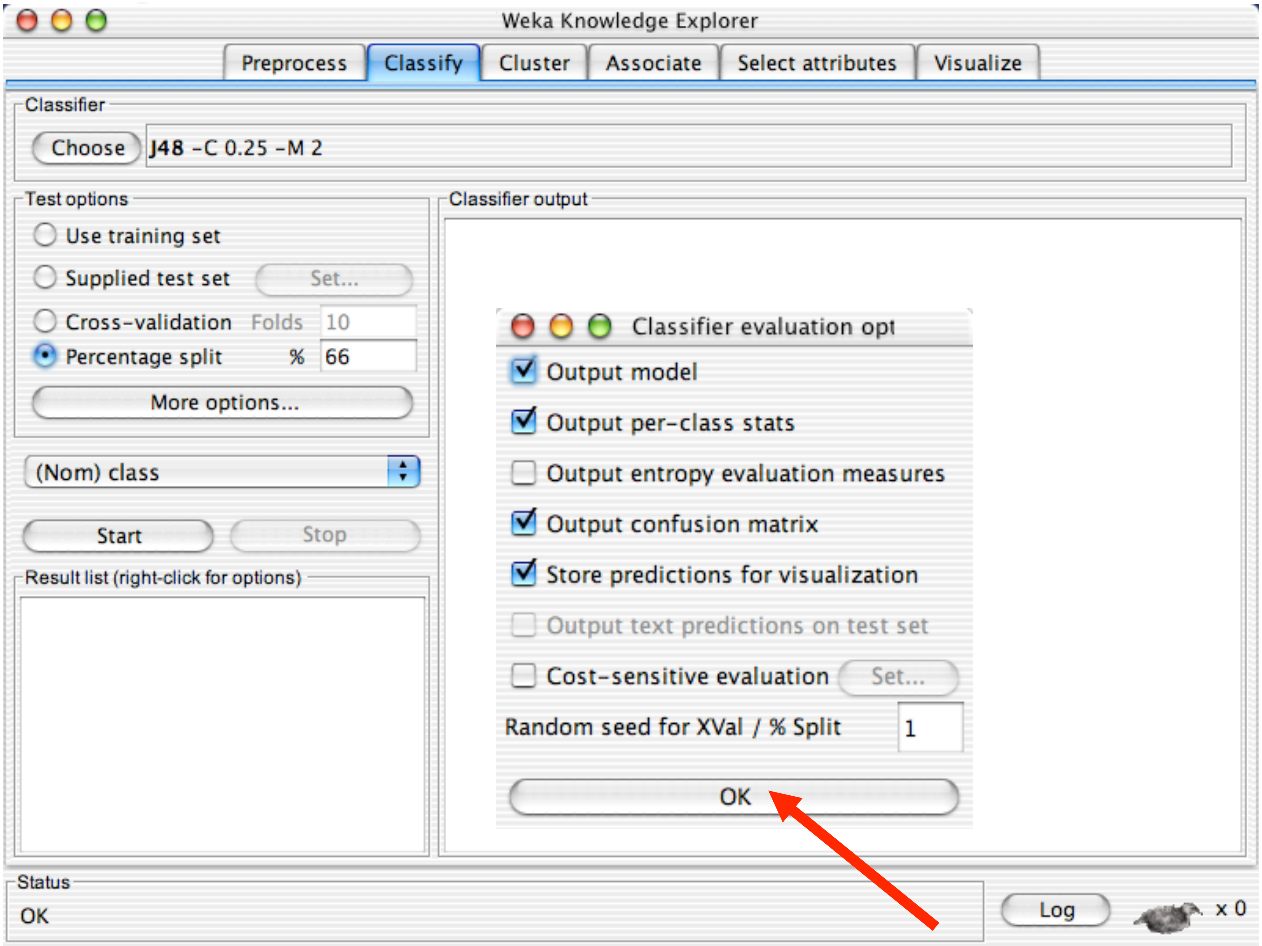

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

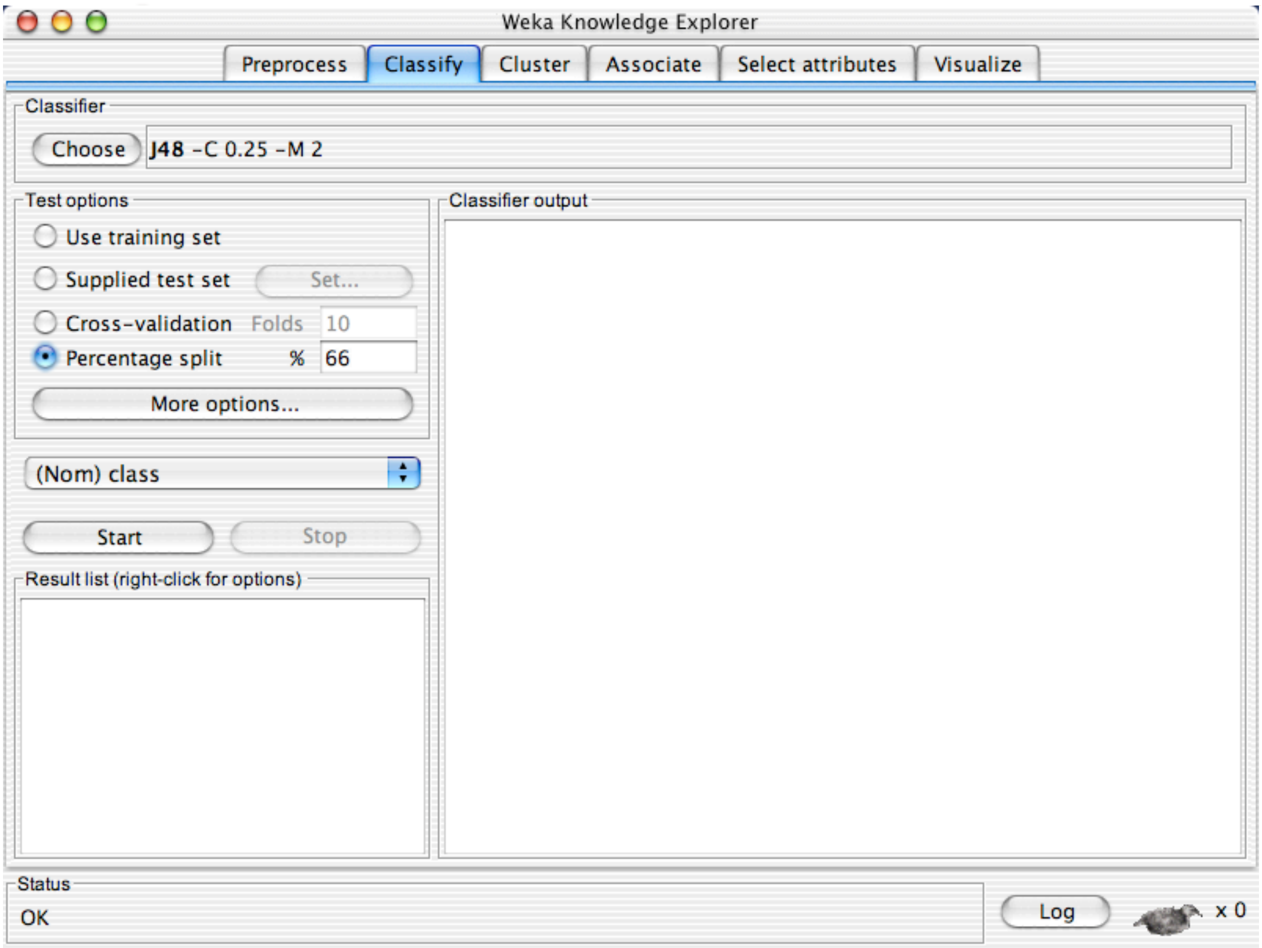

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

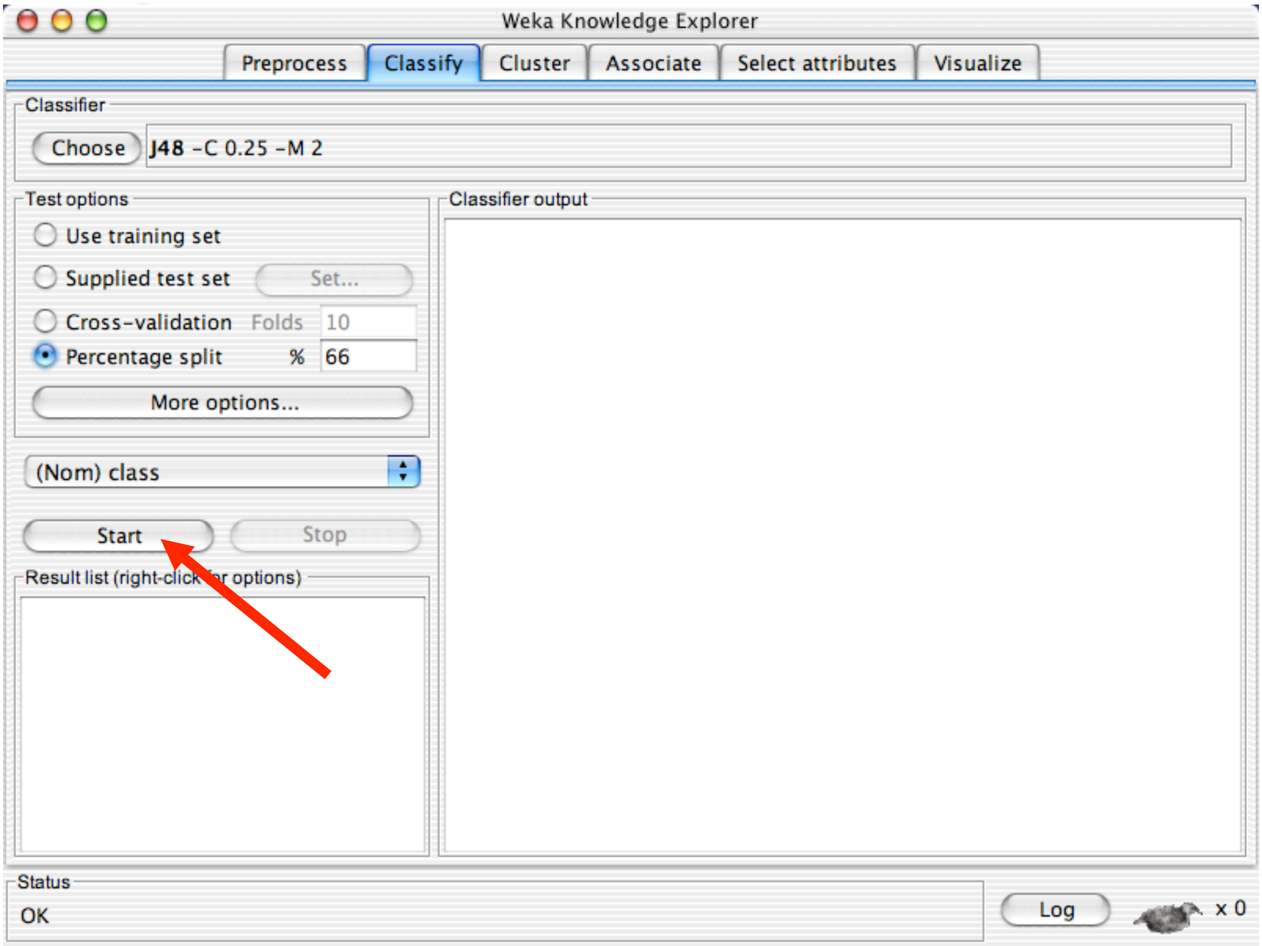

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

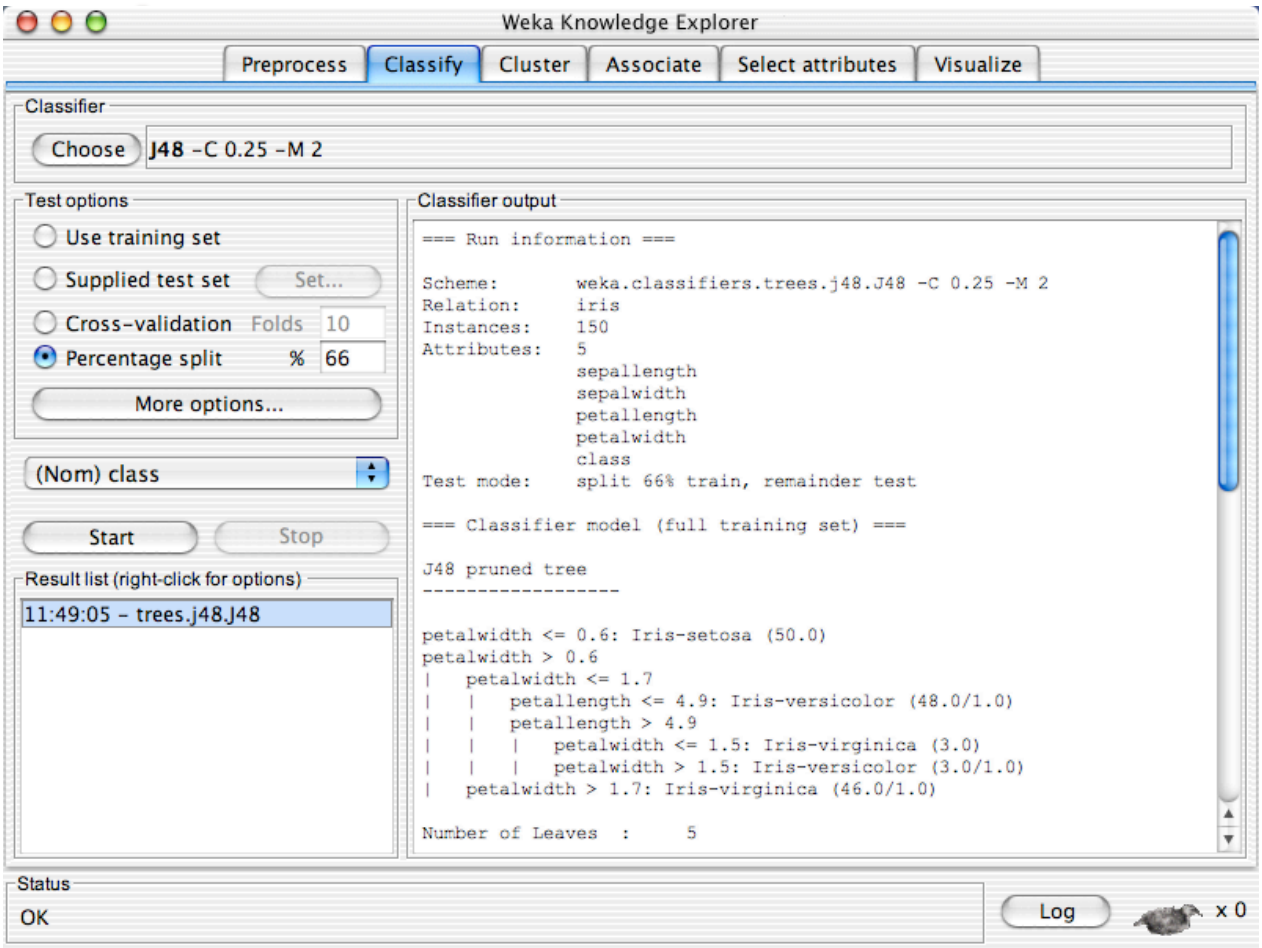

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

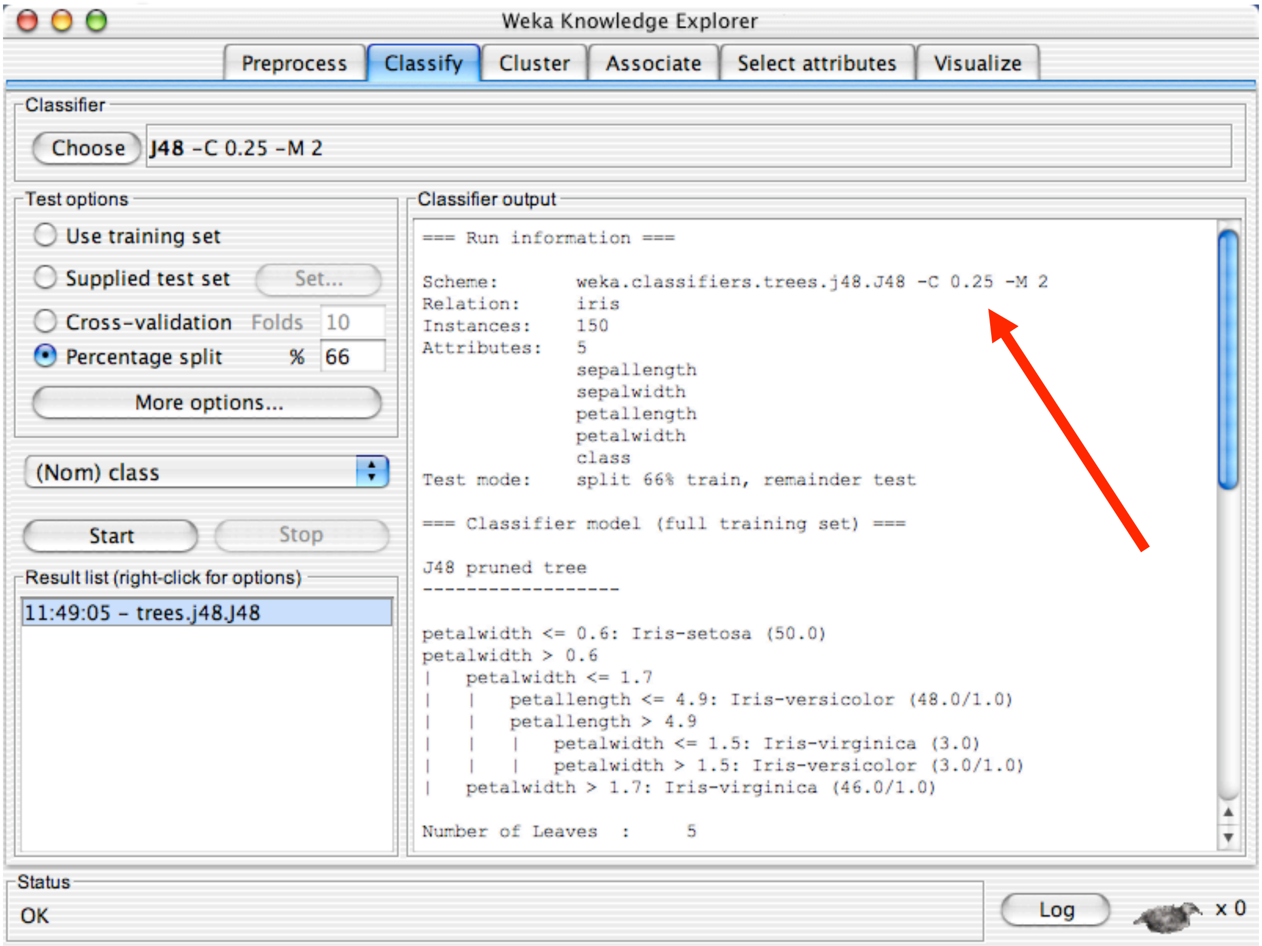

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Classifier evaluation opt
Output model
Output per-class stats
Output entropy evaluation measures
Output confusion matrix
Store predictions for visualization
Output text predictions on test set
Cost-sensitive evaluation
Set...
Random seed for XVal / % Split
1
OK
Status
OK
Log
x 0

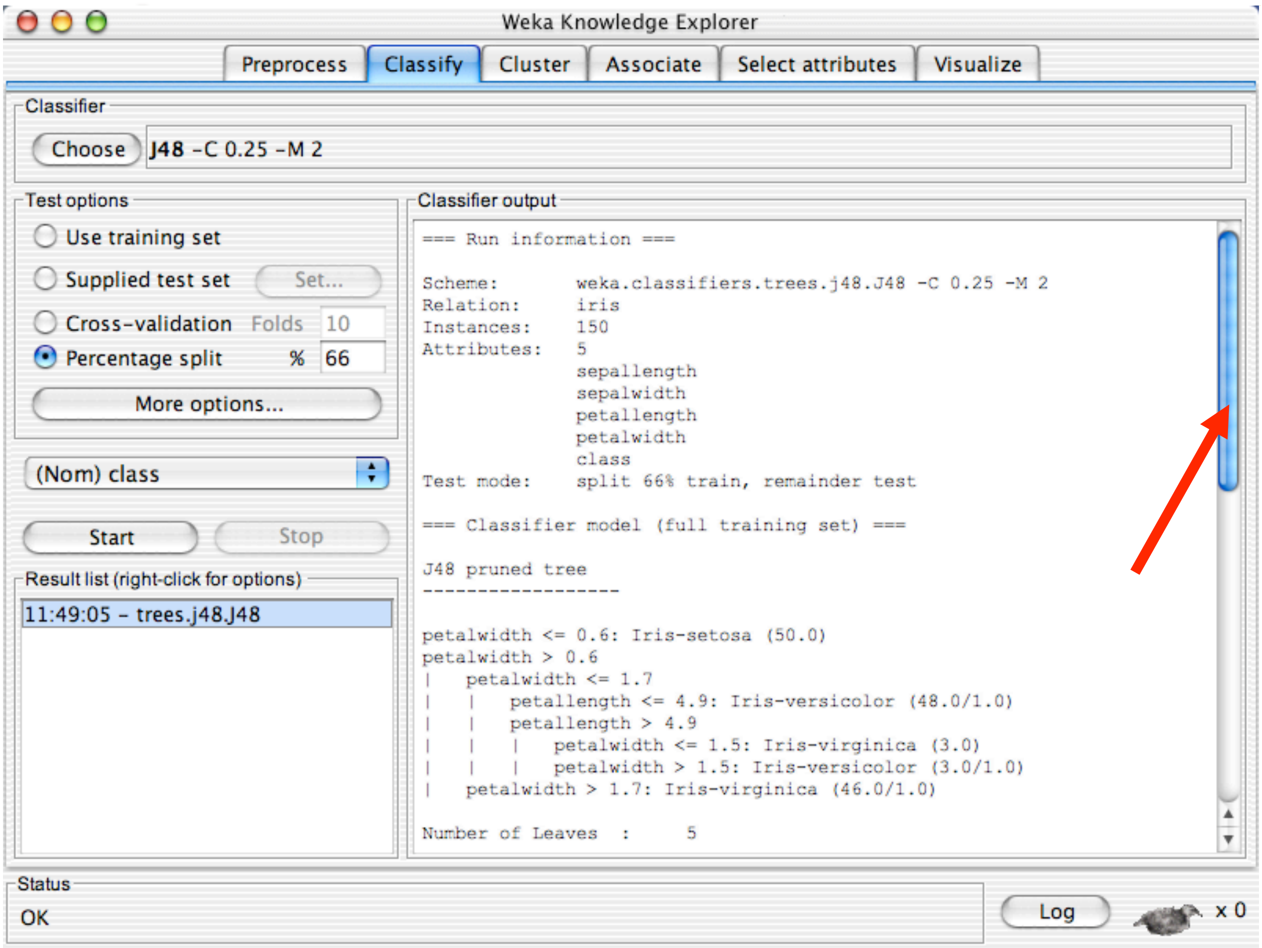

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Classifier evaluation opt
Output model
Output per-class stats
Output entropy evaluation measures
Output confusion matrix
Store predictions for visualization
Output text predictions on test set
Cost-sensitive evaluation
Set...
Random seed for XVal / % Split
1
OK
Status
OK
Log
x 0

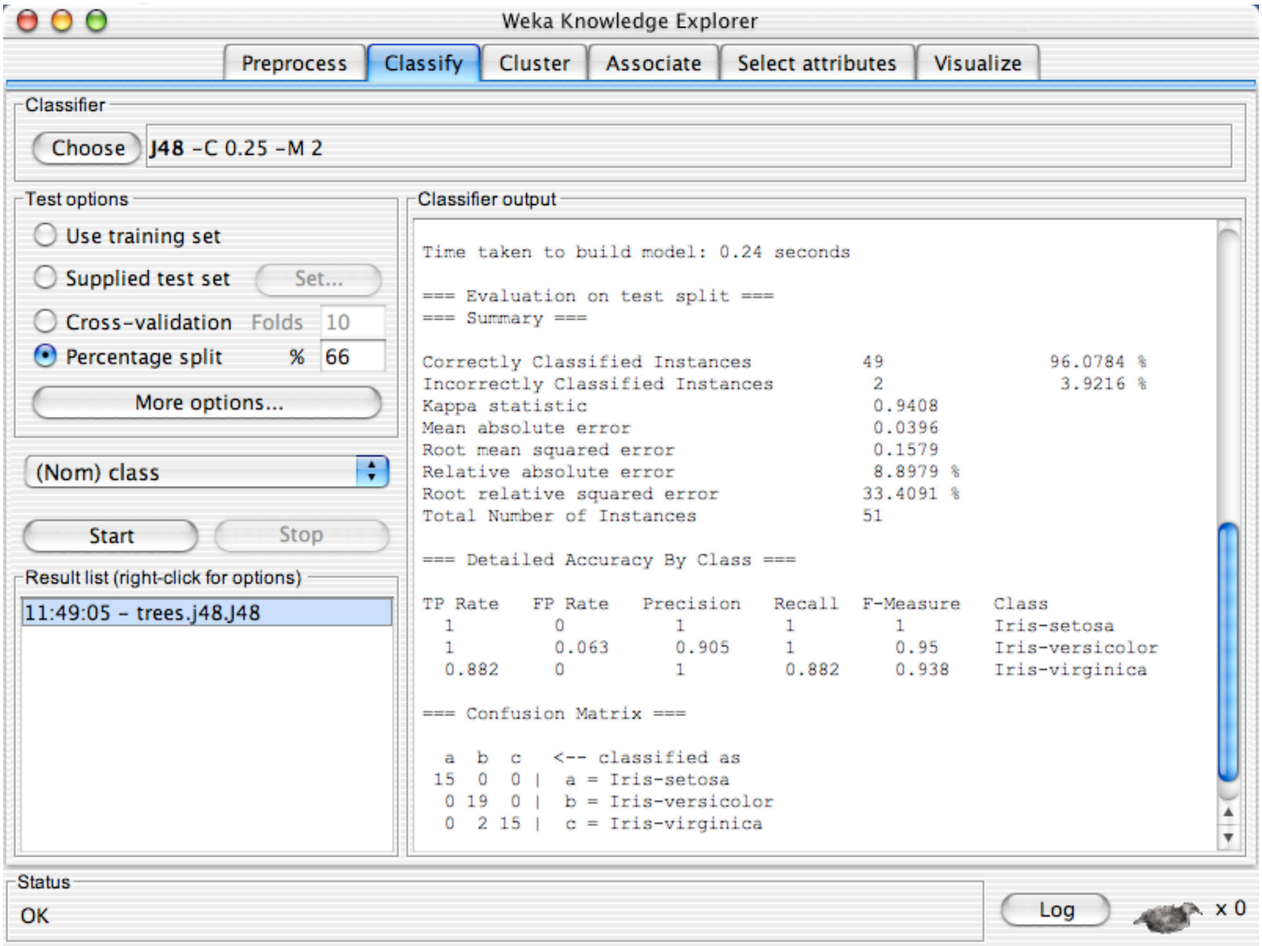

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

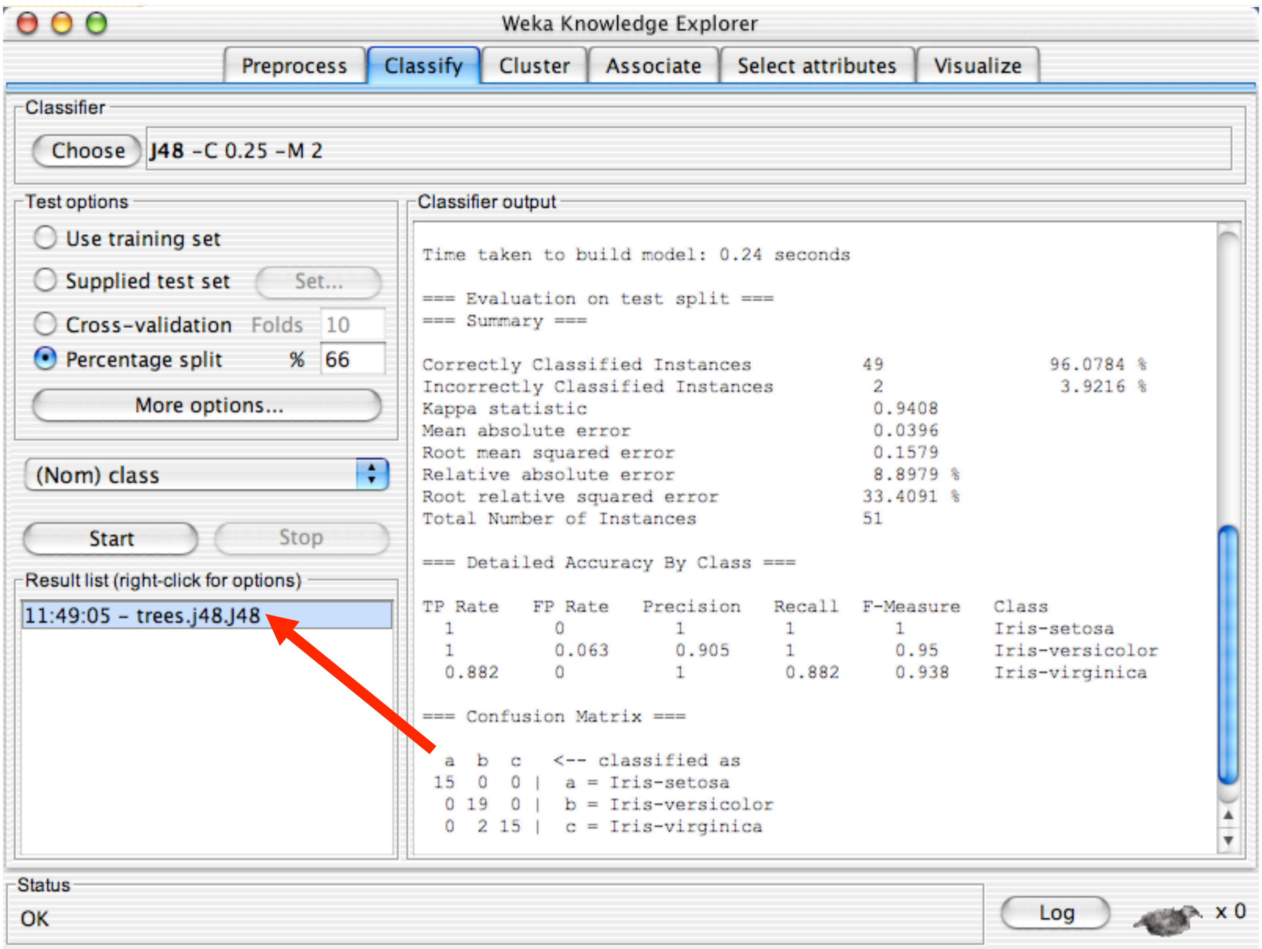

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
Classifier output
Status
OK
Log
x 0

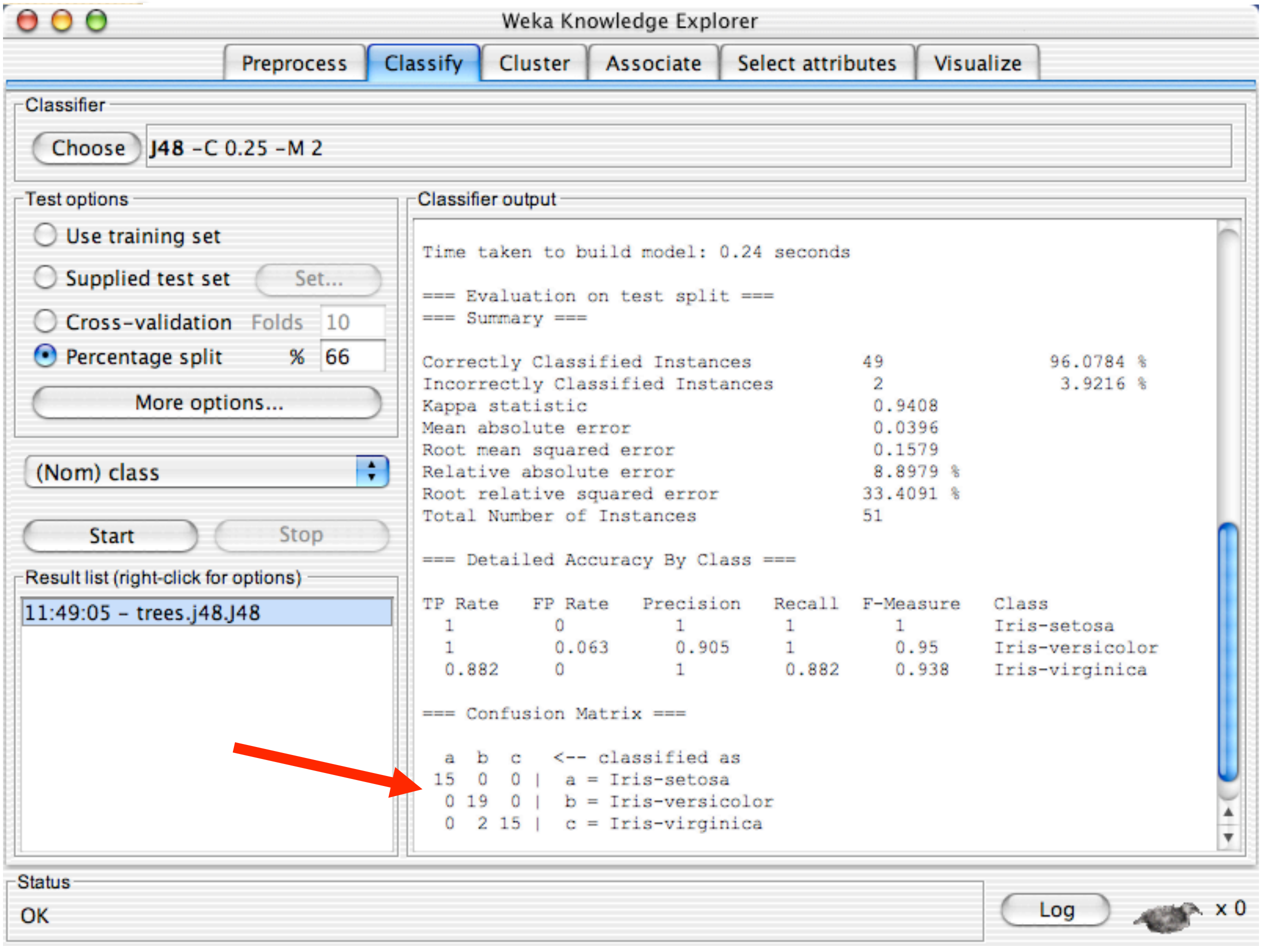

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: split 66% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Status
OK
Log
x 0

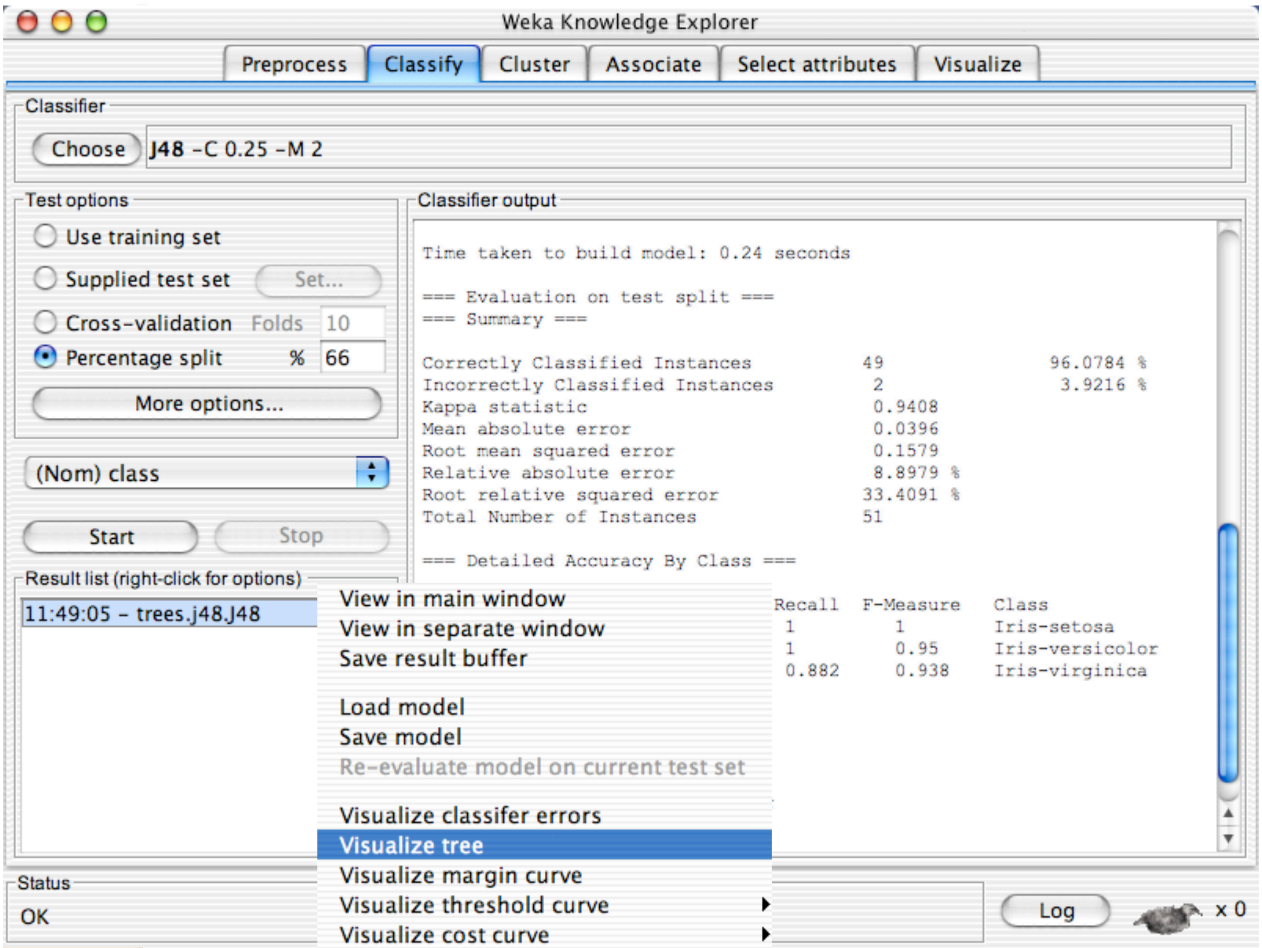

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: split 66% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Status
OK
Log
x 0

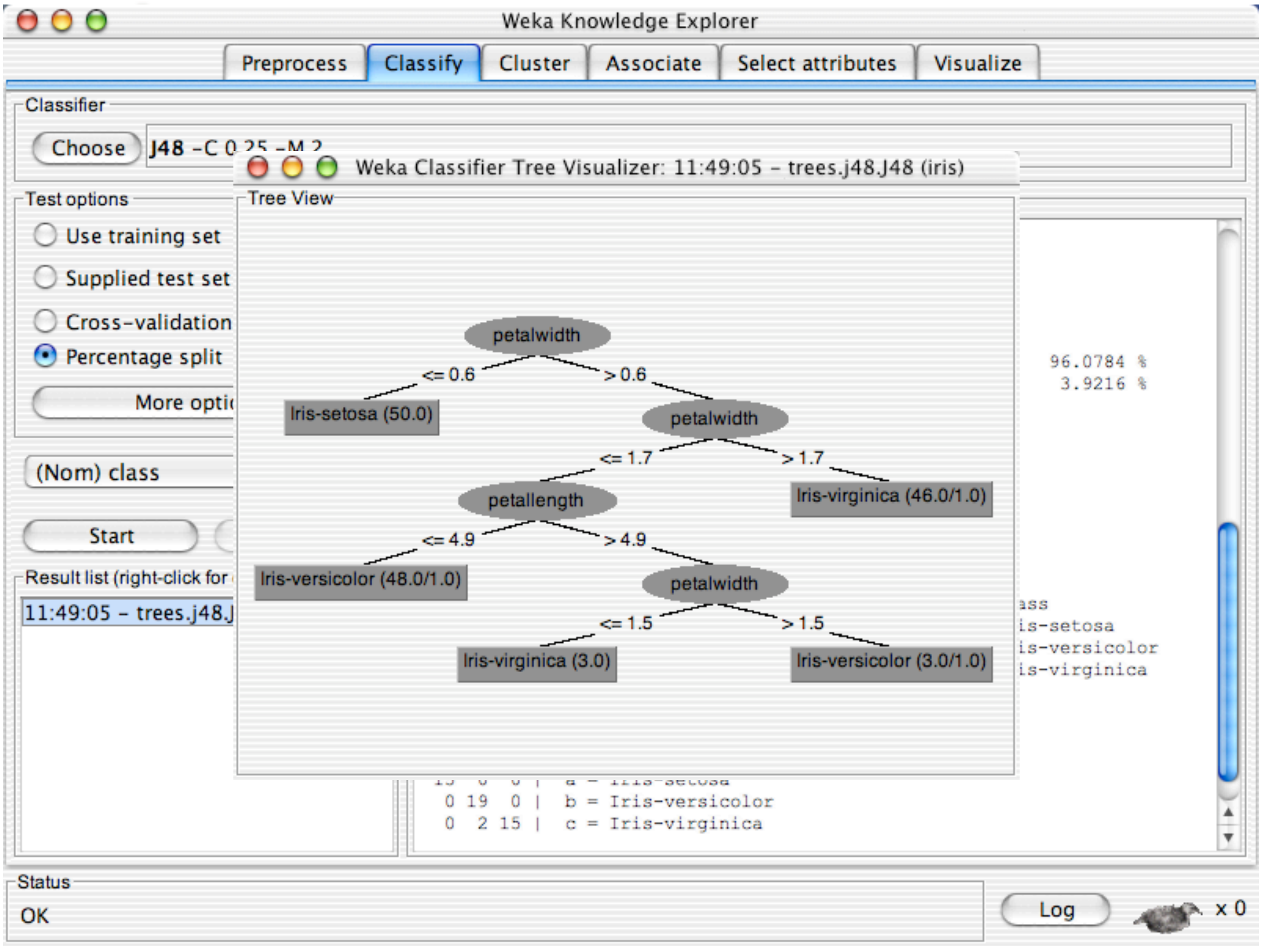

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: split 66% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Status
OK
Log
x 0


Weka Knowledge 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...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
Time taken to build model: 0.24 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 Iris-setosa
1 0.063 0.905 1 0.95 Iris-versicolor
0.882 0 1 0.882 0.938 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Status
OK
Log
x 0


Weka Knowledge 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...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
Time taken to build model: 0.24 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 Iris-setosa
1 0.063 0.905 1 0.95 Iris-versicolor
0.882 0 1 0.882 0.938 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Status
OK
Log
x 0

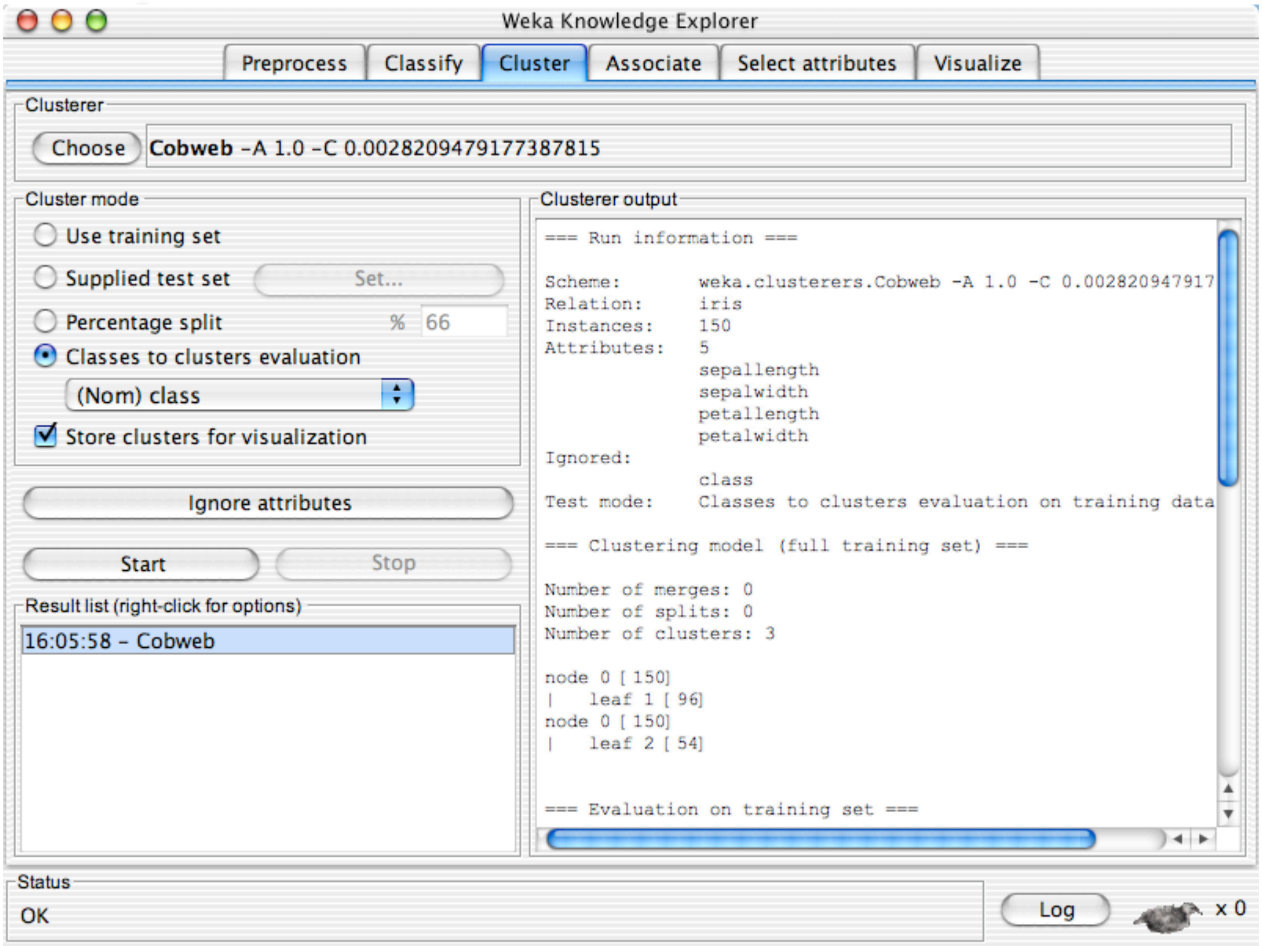

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
Time taken to build model: 0.24 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 Iris-setosa
1 0.063 0.905 1 0.95 Iris-versicolor
0.882 0 1 0.882 0.938 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Status
OK
Log
x 0


Weka Knowledge 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...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
Time taken to build model: 0.24 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 49 96.0784 %
Incorrectly Classified Instances 2 3.9216 %
Kappa statistic 0.9408
Mean absolute error 0.0396
Root mean squared error 0.1579
Relative absolute error 8.8979 %
Root relative squared error 33.4091 %
Total Number of Instances 51
=== Detailed Accuracy By Class ===
Recall F-Measure Class
1 1 Iris-setosa
1 0.95 Iris-versicolor
0.882 0.938 Iris-virginica
View in main window
View in separate window
Save result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifer errors
Visualize tree
Visualize margin curve
Visualize threshold curve
Visualize cost curve
Status
OK
Log
x 0



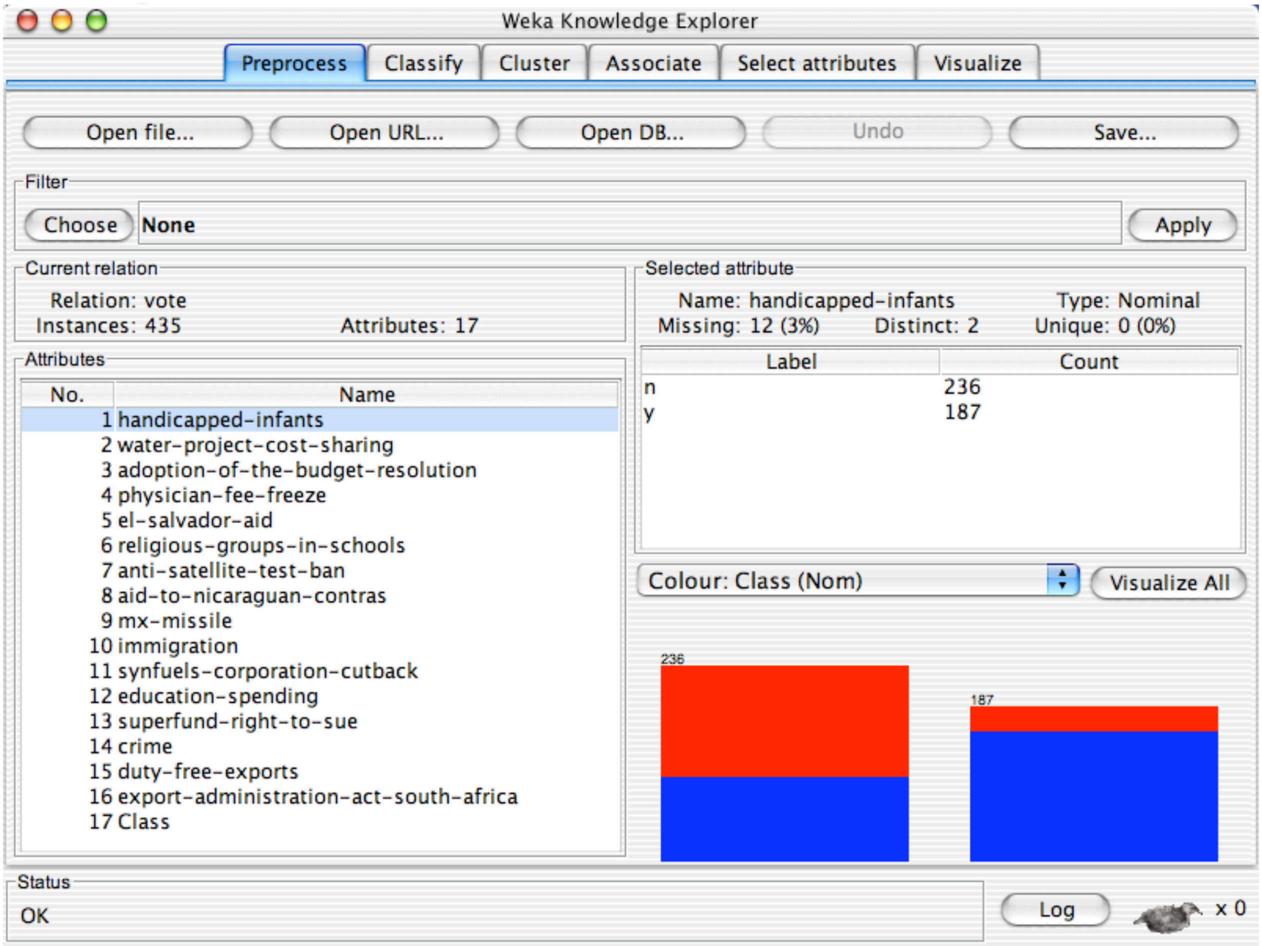
Explorer: clustering data

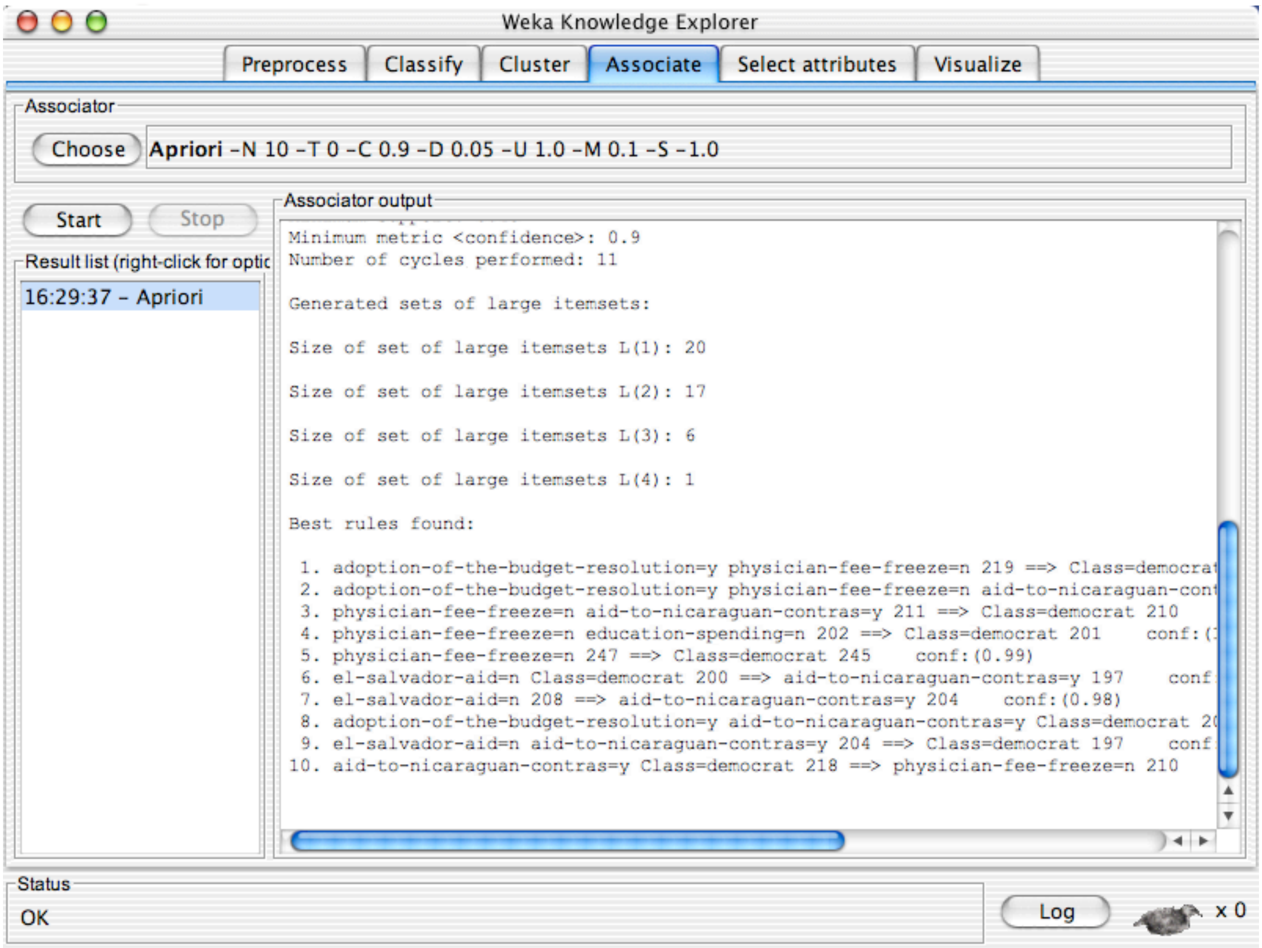
- WEKA contains "clusterers" for finding groups of similar instances in a dataset
- Implemented schemes are:
  - *k*-Means, EM, Cobweb, *X*-means, FarthestFirst
- Clusters can be visualized and compared to "true" clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution



Explorer: finding associations

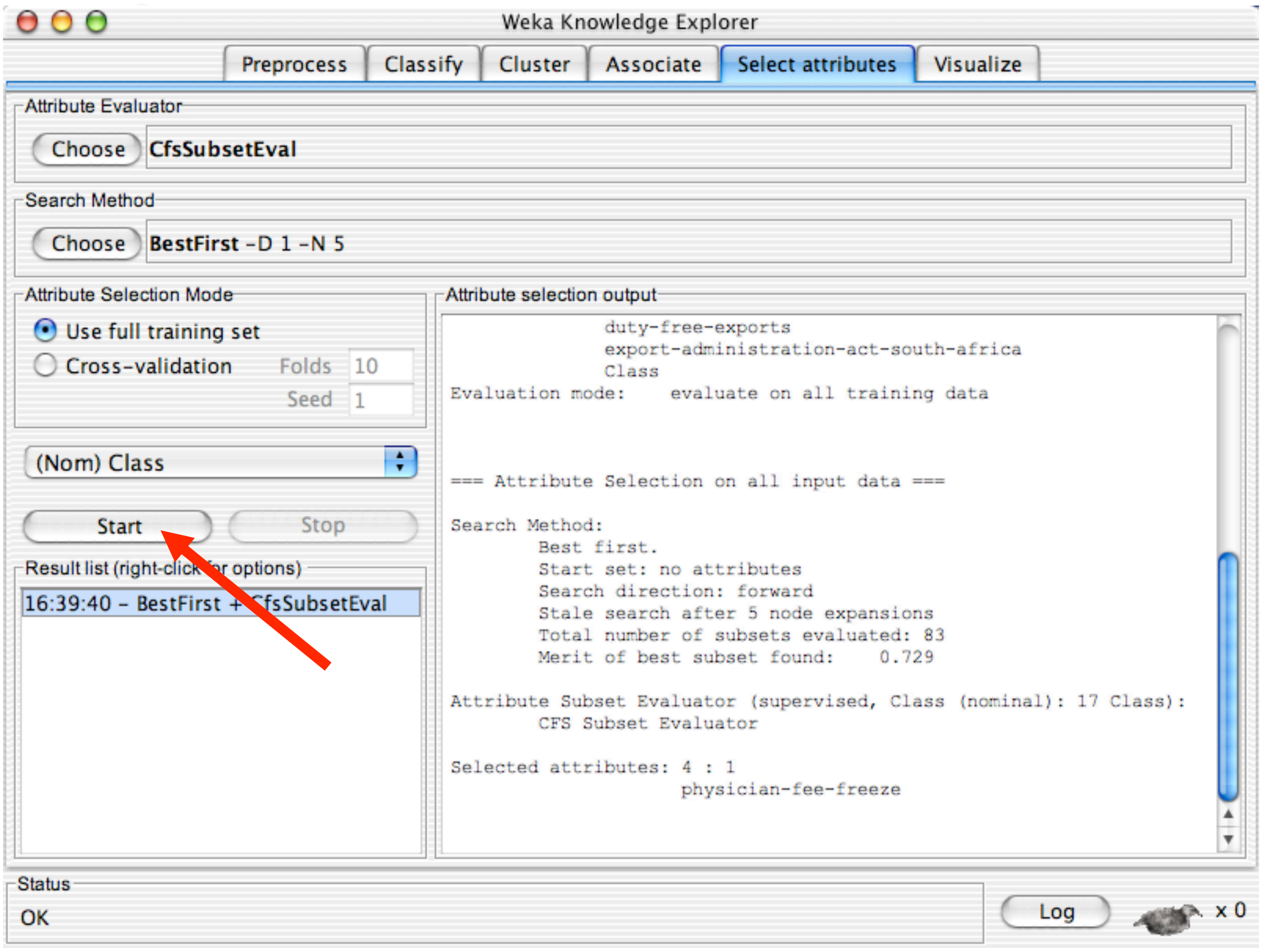
- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter ⇒ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have a given minimum support and exceed a given confidence

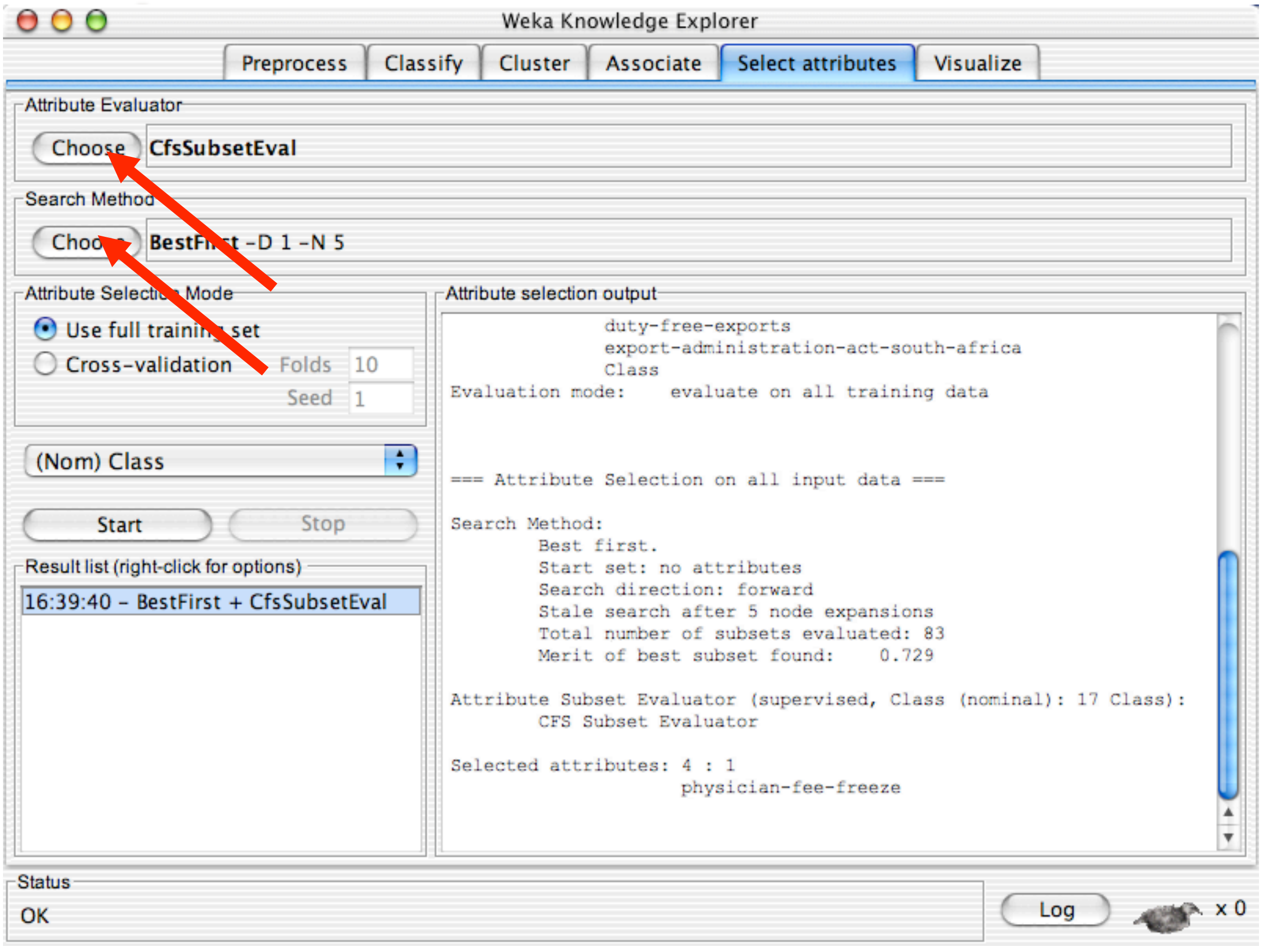

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: vote
Instances: 435
Attributes: 17
Attributes
No.
Name
1 handicapped-infants
2 water-project-cost-sharing
3 adoption-of-the-budget-resolution
4 physician-fee-freeze
5 el-salvador-aid
6 religious-groups-in-schools
7 anti-satellite-test-ban
8 aid-to-nicaraguan-contras
9 mx-missile
10 immigration
11 synfuels-corporation-cutback
12 education-spending
13 superfund-right-to-sue
14 crime
15 duty-free-exports
16 export-administration-act-south-africa
17 Class
Selected attribute
Name: handicapped-infants
Type: Nominal
Missing: 12 (3%)
Distinct: 2
Unique: 0 (0%)
Label
Count
n 236
y 187
Colour: Class (Nom)
Visualize All
236
187
Status
OK
Log
x 0

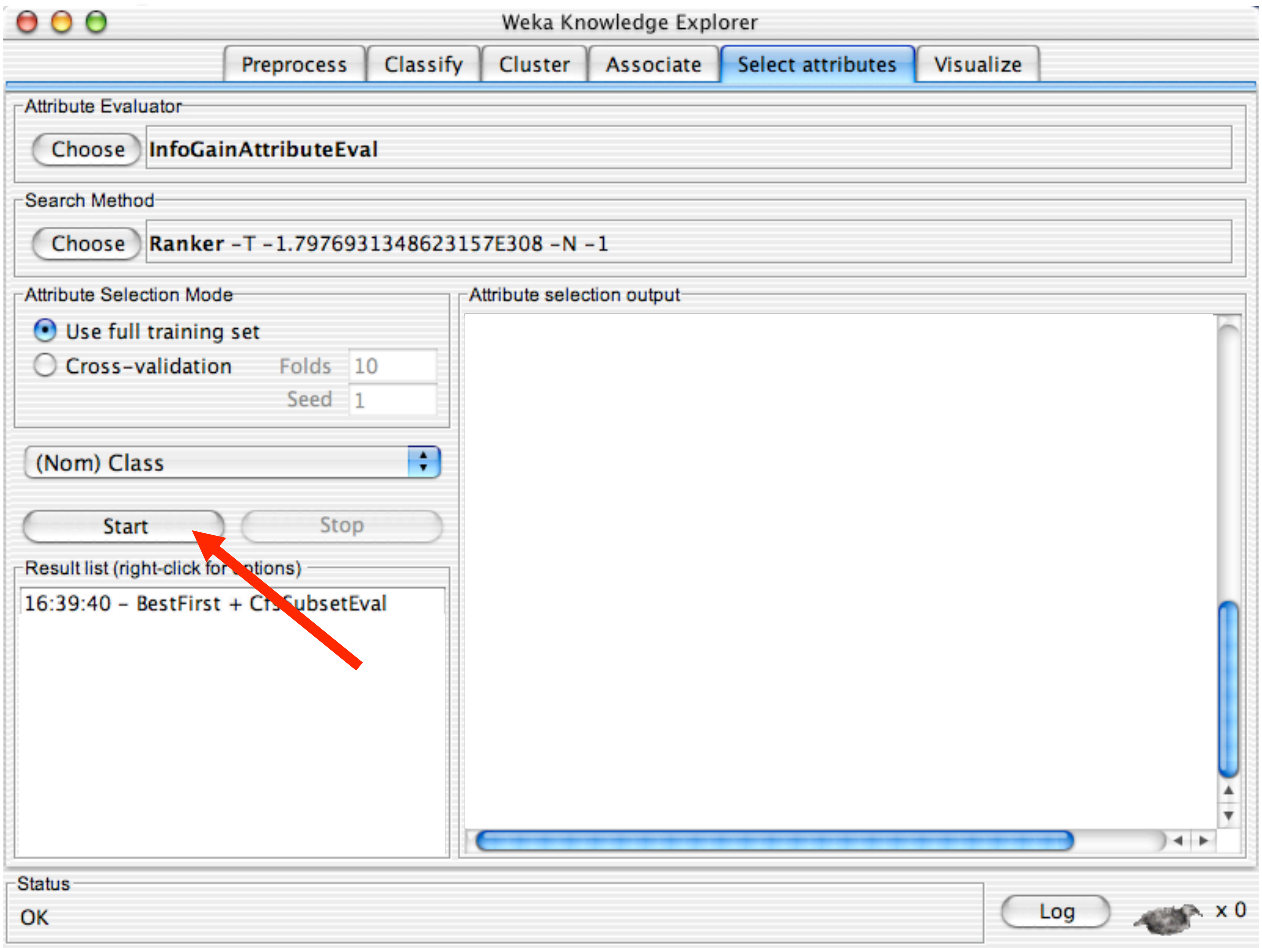

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for optic
16:29:37 - Apriori
Associator output
Minimum metric <confidence>: 0.9
Number of cycles performed: 11
Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 17
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1
Best rules found:
1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-cont
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(
5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 2
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210
Status
OK
Log
x 0

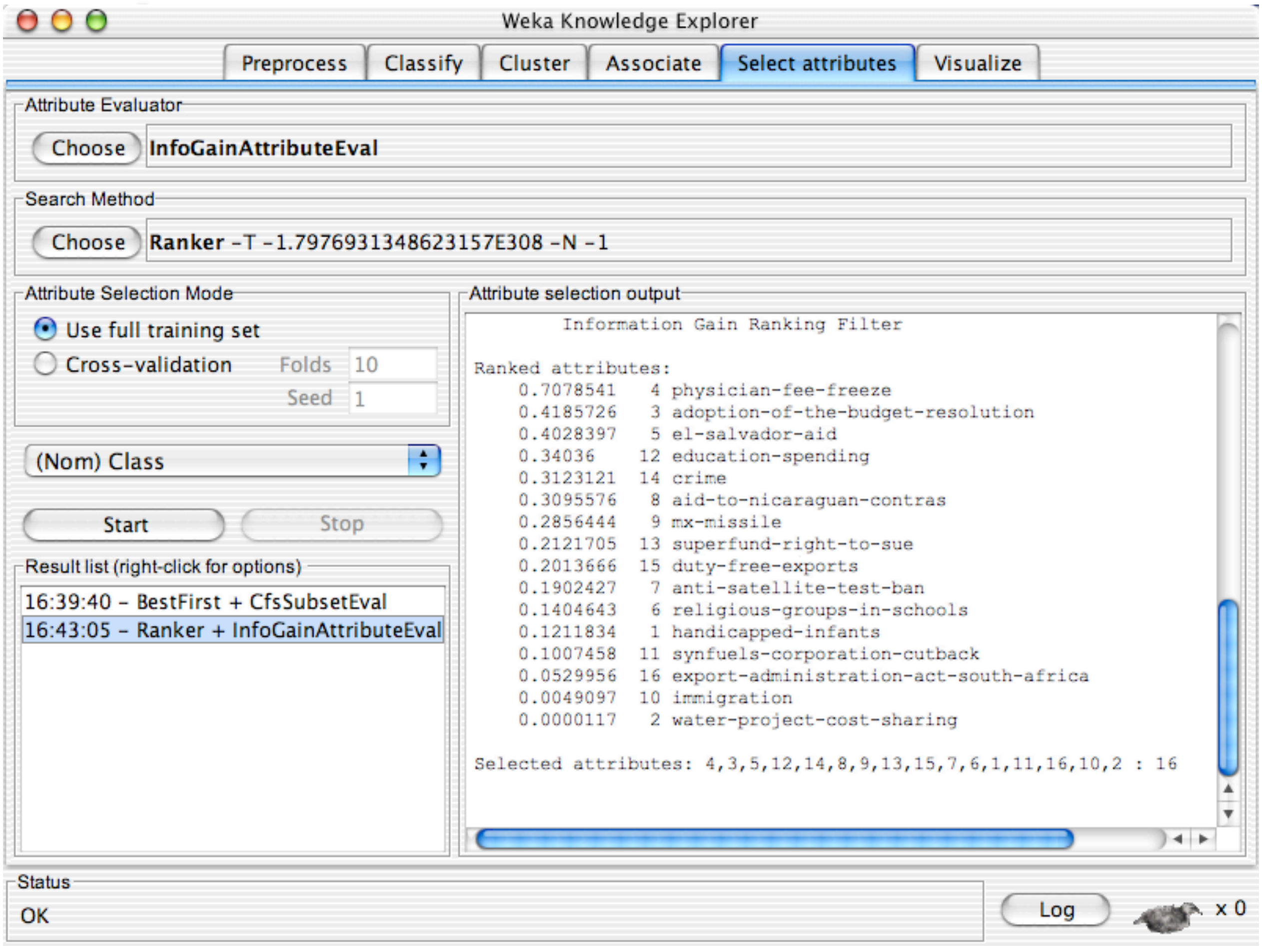
Explorer: attribute selection

- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - An evaluation method: correlation-based, wrapper, information gain, chi-squared, …
- Very flexible: WEKA allows (almost) arbitrary combinations of these two




Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Attribute selection output
duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data
=== 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: 83
Merit of best subset found: 0.729
Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator
Selected attributes: 4 : 1
physician-fee-freeze
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Attribute selection output
Status
OK
Log
x 0



Which attribute selector?

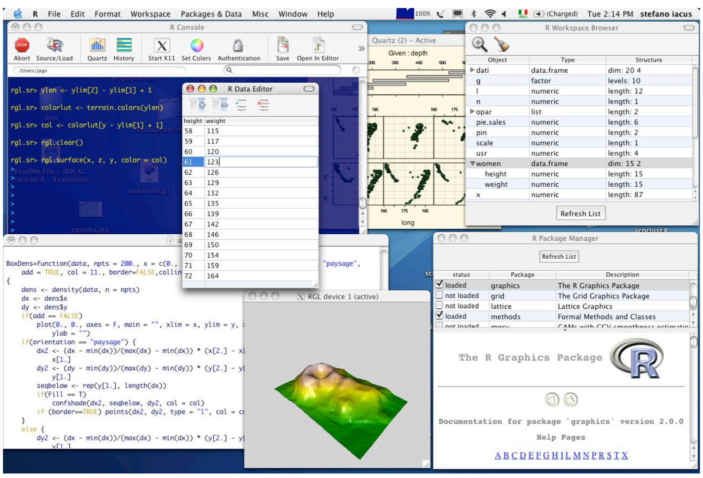
- Best: WRAPPER
  - Slow: O(2^N) search through all attribute combinations
  - The "wrapped" learner called to assess each combination
  - Some heuristics to prune the search; but does not scale
- If not WRAPPER
  - Use InfoGain / OneR for very big datasets
  - Use CFS otherwise
- Don't use PCA
  - This is an unsupervised selector
  - So it is uninformed on how dimensions help classification

Limitations

- Loads all data into ram prior to learning
  - Problem for large data sets
- Not good for complex experiments
- IMHO, discourages experimentation with new learners
  - The "WEKA effect"
    - Try every learner till something works
- Still, very useful for
  - Initial investigations
  - Learning data mining
  - Or as a sub-routine of other systems

Alternate tools: "R"

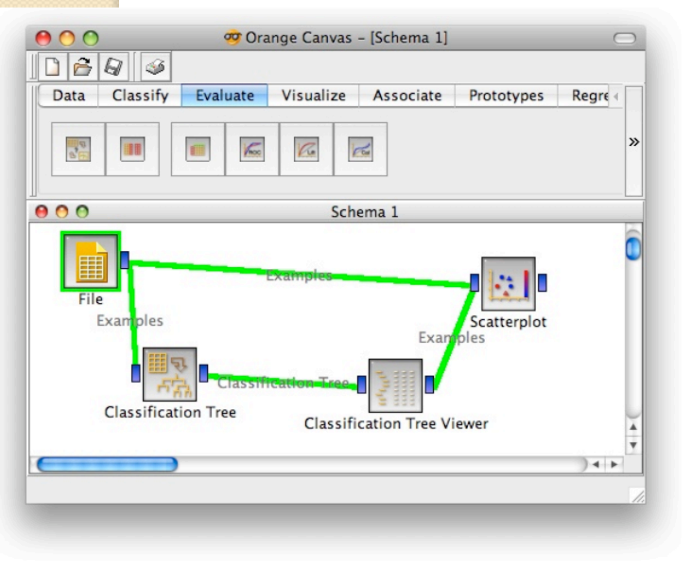
- Leading open-source system for statistical computing and graphics,
- http://www.r-project.org/



Alternate tools: Matlab

- For me: just say no
- Open science, open tools

Alternate tools: Orange

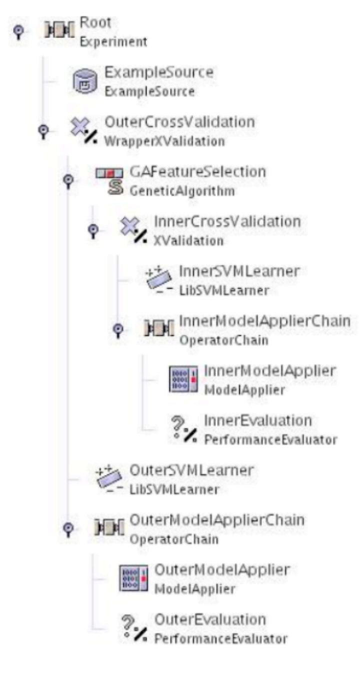


Written in Python

Simpler specification (but see WEKA's KnowledgeFlow Environment).

Also, less community support/debugging. So sometimes frustrated by random bugs

Alternate tools: RapidMiner



Experiments specified in an XML tree syntax

In theory, possible to share experimental descriptions

Alternate tools: OurMine

```
Java=$Base/lib/java
Weka="java -Xmx2048M -cp $Java/weka.jar "
Clusterers="java -Xmx1024M -jar $Java/Clusterers.jar "
Reducers="java -Xmx1024M -jar $Java/Reduce.jar "

nb() {
   local learner=weka.classifiers.bayes.NaiveBayes
   $Weka $learner -p 0 -t $1 -T $2
}

nb10() {
   local learner=weka.classifiers.bayes.NaiveBayes
   $Weka $learner -i -t $1
}

j48() {
   local learner=weka.classifiers.trees.J48
   $Weka $learner -p 0 -C 0.25 -M 2 -t $1 -T $2
}
```

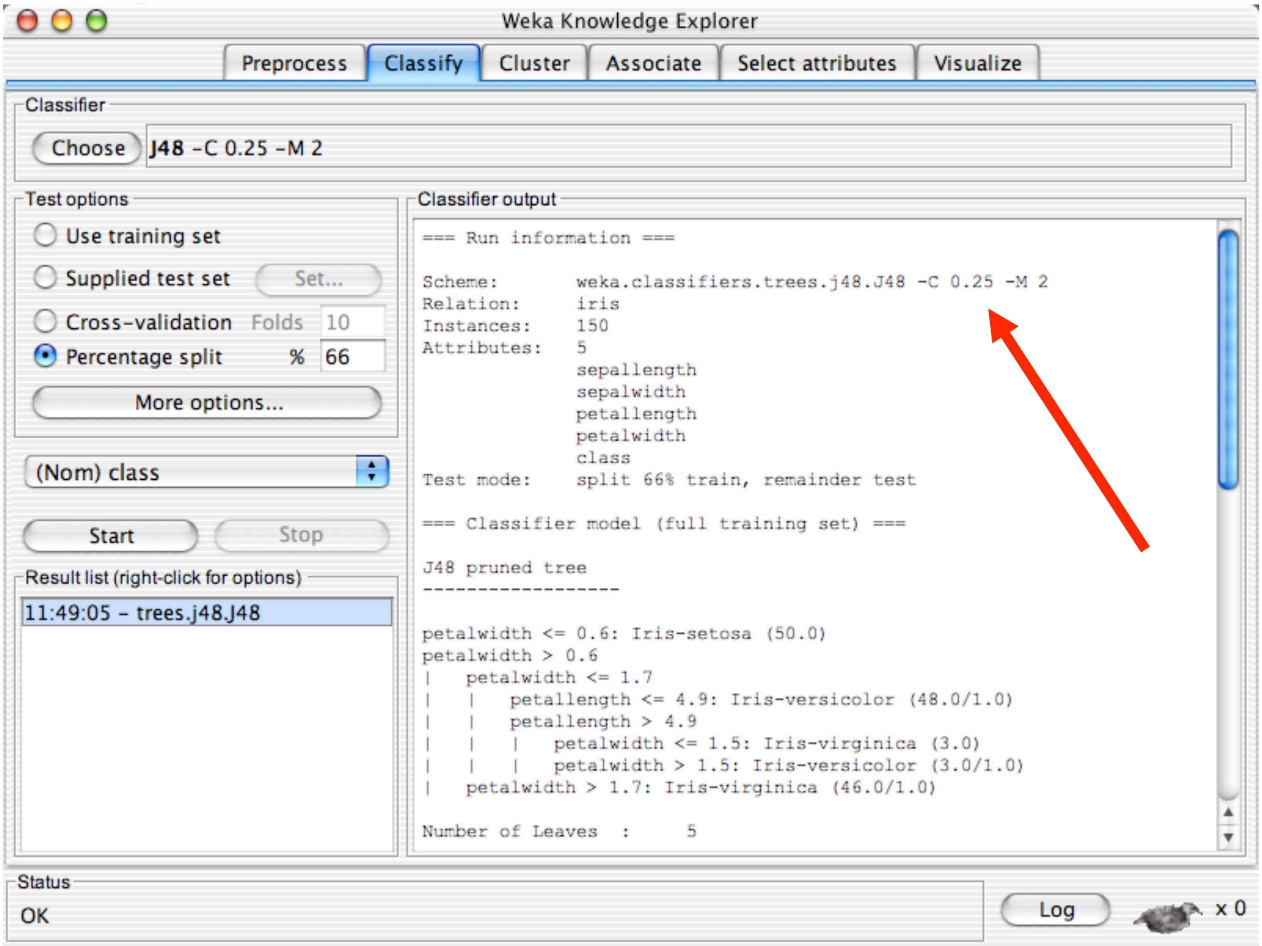
Adam Nelson, Tim Menzies, Gregory Gay,
Sharing Experiments Using Open Source Software,
Softw. Pract. Exper. 2011

Forget the visuals.

Make WEKA a sub-routine
inside Bash script

Now you can mix WEKA's JAVA with learners written in your
favorite language.

But how do you find the magic command strings?



Why go to all that trouble?

```
analysis1(){
   local origdata=$1
   local outstats=$2
   local nattrs="2 4 6 8 10 12 14 16 18 20"
   local learners="nb10 j4810 zeror10 oner10 adtree10"
   local reducers="infogain chisquared oneR"
   local tmpred=$Tmp/red
   echo "n,reducer,learner,accuracy" > $outstats

   for n in $nattrs; do
      for reducer in $reducers; do
          $reducer $origdata $n $tmpred
          for learner in $learners; do
            accur=`$learner $tmpred.arff | acc
            out="$n,$reducer,$learner,$accur"
            blabln $out
           echo $out >> $outstats
         done
      done
   done
}
```

Complex experiments, specified succinctly.

Experiments can now be reviewed, audited, by others.

Also, in 12 months time when Reviewer2 wants a tiny extension to the old paper, you don't have to remember all that clicking you did: just rerun the script.

Coming next…

- Enough details
- So many tools in WEKA, R, Rapid-Miner, Orange, OURMINE…
- The great secret
  - All those "different" tools do the same thing.
    - Carve up vector space.

DATA CARVING (THE CORE OPERATORS OF DM)

Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

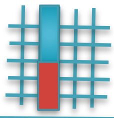
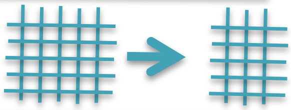
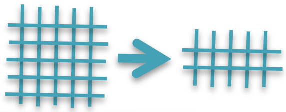
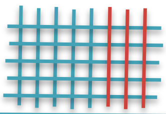
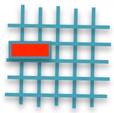
"Data Carving": A geometric view of data mining

- Data is like a block of marble,
  - waiting for a sculptor (that's you)
  - to find the shape within
- So "data mining" is really "data carving"
  - chipping away the irrelevancies
  - To find what lies beneath.

Four operators of data carving

- Each example is a row in a table
- What can can we do change the table geometry?

```
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,      85,85,FALSE,  no
sunny,      80,90,TRUE,   no
overcast,   83,86,FALSE,  yes
rainy,      70,96,FALSE,  no
rainy,      68,80,FALSE,  yes
rainy,      65,70,TRUE,   no
overcast,   64,65,TRUE,   yes
sunny,      72,95,FALSE,  no
sunny,      69,70,FALSE,  yes
rainy,      75,80,FALSE,  yes
sunny,      75,70,TRUE,   yes
overcast,   72,90,TRUE,   yes
overcast,   81,75,FALSE,  yes
rainy,      71,91,TRUE,   no
```

1. Clump 
2. Select columns 
3. Select rows 
4. Rotate (add new columns) 
5. Clump columns 

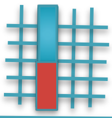
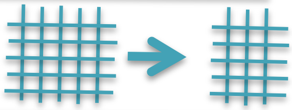
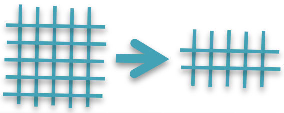
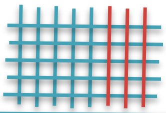
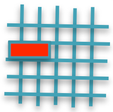
The field is called "data mining", not "algorithm mining"

- To understand data mining, look at the data, not the algorithms
- Why? We do data mining not to study algorithms.
  - But to study data
- Our results should be insights about data,
  - not trivia about (say) decision tree algorithms
- Besides, the thing that most predicts for performance is the data, not the algorithm,
  - Pedro Domingos and Michael J. Pazzani, On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Volume 29, number 2-3, pages 103-130, 1997

Table 1. Classification accuracies and sample standard deviations, averaged over 20 random training/test splits. "Bayes" is the Bayesian classifier with discretization and "Gauss" is the Bayesian classifier with Gaussian distributions. Superscripts denote confidence levels for the difference in accuracy between the Bayesian classifier and the corresponding algorithm, using a one-tailed paired t test: 1 is 99.5%, 2 is 99%, 3 is 97.5%, 4 is 95%, 5 is 90%, and 6 is below 90%.

Data Set	Bayes	Gauss	C4.5	PEBLS	CN2	Def.
Audiology	73.0 ± 6.1	73.0 ± 6.1^{6}	72.5 ± 5.8^{6}	75.8 ± 5.4^{3}	71.0 ± 5.1^{5}	21.3
Annealing	95.3 ± 1.2	84.3 ± 3.8^{1}	90.5 ± 2.2^{1}	98.8 ± 0.8^{1}	81.2 ± 5.4^{1}	76.4
Breast cancer	71.6 ± 4.7	71.3 ± 4.3^{6}	70.1 ± 6.8^{5}	65.6 ± 4.7^{1}	67.9 ± 7.1^{1}	67.6
Credit	84.5 ± 1.8	78.9 ± 2.5^{1}	85.9 ± 2.1^{3}	82.2 ± 1.9^{1}	82.0 ± 2.2^{1}	57.4
Chess endgames	88.0 ± 1.4	88.0 ± 1.4^{6}	99.2 ± 0.1^{1}	96.9 ± 0.7^{1}	98.1 ± 1.0^{1}	52.0
Diabetes	74.5 ± 2.4	75.2 ± 2.1^{6}	73.5 ± 3.4^{5}	71.1 ± 2.4^{1}	73.8 ± 2.7^{6}	66.0
Echocardiogram	69.1 ± 5.4	73.4 ± 4.9^{1}	64.7 ± 6.3^{1}	61.7 ± 6.4^{1}	68.2 ± 7.2^{6}	67.8
Glass	61.9 ± 6.2	50.6 ± 8.2^{1}	63.9 ± 8.7^{6}	62.0 ± 7.4^{6}	63.8 ± 5.5^{6}	31.7
Heart disease	81.9 ± 3.4	84.1 ± 2.8^{1}	77.5 ± 4.3^{1}	78.9 ± 4.0^{1}	79.7 ± 2.9^{3}	55.0
Hepatitis	85.3 ± 3.7	85.2 ± 4.0^{6}	79.2 ± 4.3^{1}	79.0 ± 5.1^{1}	80.3 ± 4.2^{1}	78.1
Horse colic	80.7 ± 3.7	79.3 ± 3.7^{1}	85.1 ± 3.8^{1}	75.7 ± 5.0^{1}	82.5 ± 4.2^{2}	63.6
Hypothyroid	97.5 ± 0.3	97.9 ± 0.4^{1}	99.1 ± 0.2^{1}	95.9 ± 0.7^{1}	98.8 ± 0.4^{1}	95.3
Iris	93.2 ± 3.5	93.9 ± 1.9^{6}	92.6 ± 2.7^{6}	93.5 ± 3.0^{6}	93.3 ± 3.6^{6}	26.5
Labor	91.3 ± 4.9	88.7 ± 10.6^{6}	78.1 ± 7.9^{1}	89.7 ± 5.0^{6}	82.1 ± 6.9^{1}	65.0
Lung cancer	46.8 ± 13.3	46.8 ± 13.3^{6}	40.9 ± 16.3^{5}	42.3 ± 17.3^{6}	38.6 ± 13.5^{3}	26.8
Liver disease	63.0 ± 3.3	54.8 ± 5.5^{1}	65.9 ± 4.4^{1}	61.3 ± 4.3^{6}	65.0 ± 3.8^{3}	58.1
LED	62.9 ± 6.5	62.9 ± 6.5^{6}	61.2 ± 8.4^{6}	55.3 ± 6.1^{1}	58.6 ± 8.1^{2}	8.0
Lymphography	81.6 ± 5.9	81.1 ± 4.8^{6}	75.0 ± 4.2^{1}	82.9 ± 5.6^{6}	78.8 ± 4.9^{3}	57.3
Post-operative	64.7 ± 6.8	67.2 ± 5.0^{3}	70.0 ± 5.2^{1}	59.2 ± 8.0^{2}	60.8 ± 8.2^{4}	71.2
Promoters	87.9 ± 7.0	87.9 ± 7.0^{6}	74.3 ± 7.8^{1}	91.7 ± 5.9^{3}	75.9 ± 8.8^{1}	43.1
Primary tumor	44.2 ± 5.5	44.2 ± 5.5^{6}	35.9 ± 5.8^{1}	30.9 ± 4.7^{1}	39.8 ± 5.2^{1}	24.6
Solar flare	68.5 ± 3.0	68.2 ± 3.7^{6}	70.6 ± 2.9^{1}	67.6 ± 3.5^{6}	70.4 ± 3.0^{2}	25.2
Sonar	69.4 ± 7.6	63.0 ± 8.3^{1}	69.1 ± 7.4^{6}	73.8 ± 7.4^{1}	66.2 ± 7.5^{5}	50.8
Soybean	100.0 ± 0.0	100.0 ± 0.0^{6}	95.0 ± 9.0^{3}	100.0 ± 0.0^{6}	96.9 ± 5.9^{3}	30.0
Splice junctions	95.4 ± 0.6	95.4 ± 0.6^{6}	93.4 ± 0.8^{1}	94.3 ± 0.5^{1}	81.5 ± 5.5^{1}	52.4
Voting records	91.2 ± 1.7	91.2 ± 1.7^{6}	96.3 ± 1.3^{1}	94.9 ± 1.2^{1}	95.8 ± 1.6^{1}	60.5
Wine	96.4 ± 2.2	97.8 ± 1.2^{3}	92.4 ± 5.6^{1}	97.2 ± 1.8^{6}	90.8 ± 4.7^{1}	36.4
Zoology	94.4 ± 4.1	94.1 ± 3.8^{6}	89.6 ± 4.7^{1}	94.6 ± 4.3^{6}	90.6 ± 5.0^{1}	39.4

The rest of this hour

1. Clump 
2. Select columns 
3. Select rows 
4. Rotate (add new columns) 
5. Clump columns 

Carving can be dangerous

- While carving the training data is recommended
  - It is a methodological error to carve the test data
- Whatever is learned from the training data
  - Should be assessed on "raw" (i.e. uncarved) test data

Clumping column data (a.k.a. discretization)

```
overcast, 64,65,TRUE, yes
Rainy,    65,70,TRUE, no
sunny,    69,70,FALSE,yes
sunny,    75,70,TRUE, yes
overcast, 81,75,FALSE,yes
rainy,    68,80,FALSE,yes
rainy,    75,80,FALSE,yes
sunny,    85,85,FALSE,no
overcast, 83,86,FALSE,yes
overcast, 72,90,TRUE, yes
sunny,    80,90,TRUE, no
rainy,    71,91,TRUE, no
sunny,    72,95,FALSE,no
rainy,    70,96,FALSE,no
```

- Learning = compression
  - Take a target concept that is spread out across all the data
  - Squeeze it together till it is dense enough to be visible.
- Discretization: clump together observations taken over a continuous range
  - into a small number of regions.
- E.g. "toddlers" If age =1,2,3
- Discretization improves the performance of a learner
  - Gives a learner a smaller space to reason about,
  - With more examples in each part of the space

Supervised Discretization

```
overcast, 64,65,TRUE, yes
Rainy,    65,70,TRUE, no
sunny,    69,70,FALSE,yes
sunny,    75,70,TRUE, yes
overcast, 81,75,FALSE,yes
rainy,    68,80,FALSE,yes
rainy,    75,80,FALSE,yes
sunny,    85,85,FALSE,no
overcast, 83,86,FALSE,yes
overcast, 72,90,TRUE, yes
sunny,    80,90,TRUE, no
rainy,    71,91,TRUE, no
sunny,    72,95,FALSE,no
rainy,    70,96,FALSE,no
```

- Standard method:
  - Find a break that most reduces class diversity either side of the break
  - Recurse on data:
    - above break,
    - below break
  - Fayyad and Irani, Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning IJCAI'93, pp1022-1027

Unsupervised Discretization

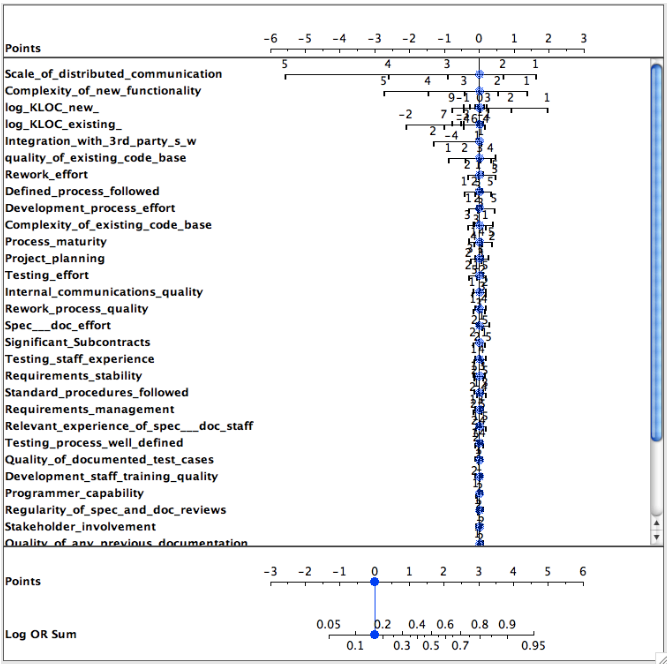
- Divide into “B” bins
  - (X – Min) / ((Max – Min)/ B)
  - B=3 or 10 very common
- Divide into P percentile groups
  - Each bins contains (say) 25% of the rows
- For Bayesian methods
  - Divide into groups of N items
  - Ying and Webb recommends N= sqrt(rows)
  - Ying Yang and Geoff Webb, Weighted Proportional k-Interval Discretization of Naïve Bayes classifeirs, PAKADD’03, p501-512, 2003

Select columns

- Occam's Razor - Entia non sunt multiplicanda praeter necessitatem. ("Entities should not be multiplied more than necessary").
  - the fewer features used to explain something, the better
- Log(OR):
  - Discrete every feature. For all pairs of target / other of size C1, C2 count frequency of range N1, N2 in each class
  - Log(odds ratio) = log((N1/C1) / (N2/C2)) > 0 if more frequent in target
  - “Pivots” are the ranges with high Log (OR)
  - Možina, M., Demšar, J., Kattan, M., and Zupan, B. 2004. Nomograms for visualization of naive Bayesian classifier. In*Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (Pisa, Italy, September 20 - 24, 2004)*
- InfoGain:
  - Use Fayyad Irani trick: assses each column by how well it divides up the data
  - Takes linear time : O(C)
- Wrapper:
  - Explore 2^C subsets of C columns: takes time $O(2^C)$
  - Call a learner on each subset
  - Use the columns that maximize learner performance
  - Not practical for large data sets
- For more, see Hall, M. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering. 15(3), November/December 2003

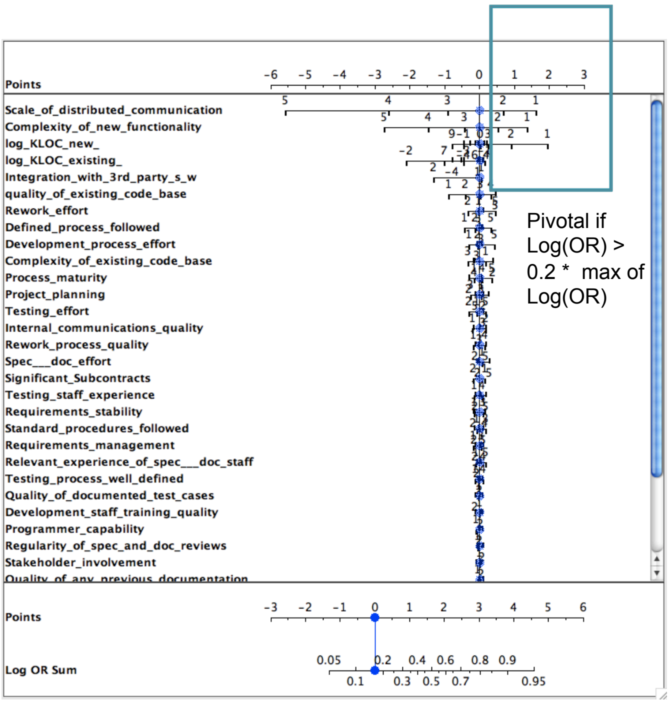
Select columns with log(OR)

- Data from Norman Fenton's Bayes Net
  - Project Data Incorporating Qualitative Factors for Improved Software Defect Prediction Norman Fenton, Martin Neil, William Marsh, Peter Hearty, Lukasz Radlinski and Paul Krause., PROMISE 2008
- Target class. worse defects
- Only a few features matter
- Only a few ranges of those features matter



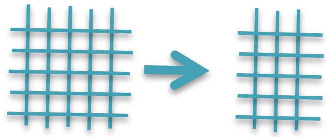
Select columns with log(OR)

- Data from Norman Fenton's Bayes Net
  - Project Data Incorporating Qualitative Factors for Improved Software Defect Prediction Norman Fenton, Martin Neil, William Marsh, Peter Hearty, Lukasz Radlinski and Paul Krause., PROMISE 2008
- Target class. worse defects
- Only a few features matter
- Only a few ranges of those features matter

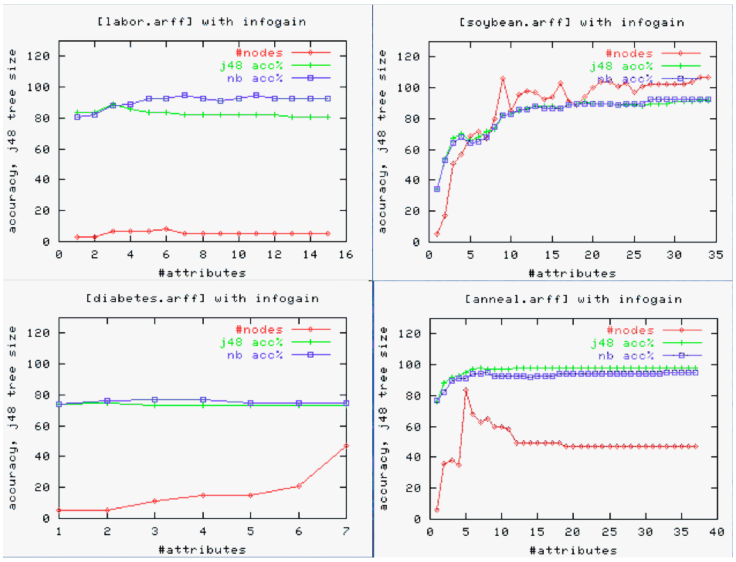


Pivotal if Log(OR) > 0.2 * max of Log(OR)

Select columns with InfoGain

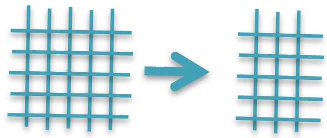


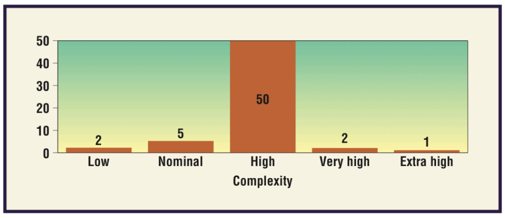
X-axis sorted by sum(-p*log(p))

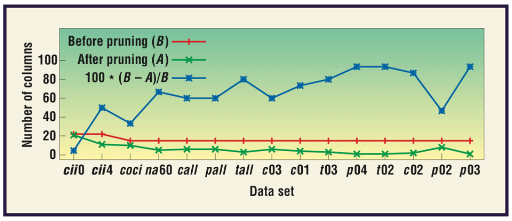


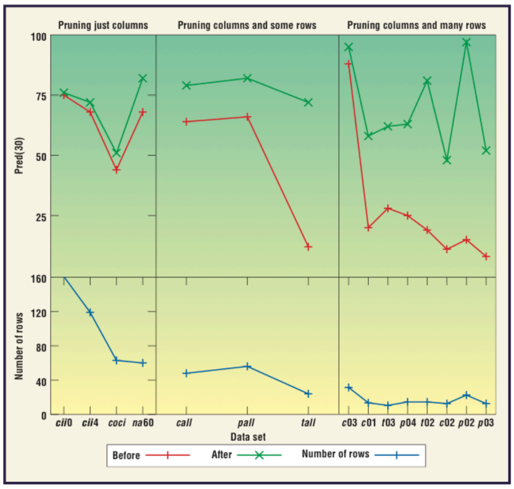
Simpler theories after column selection, work just as well as using everything

Select columns with WRAPPER



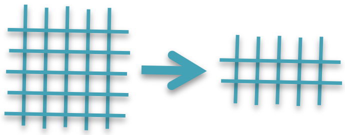






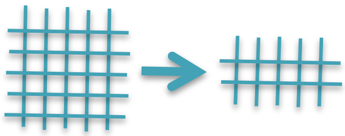
- Finding the Right Data for Software Cost Modeling Chen, Menzies, Port, Boehm, IEEE Software Nov/Dec 2005

Select rows

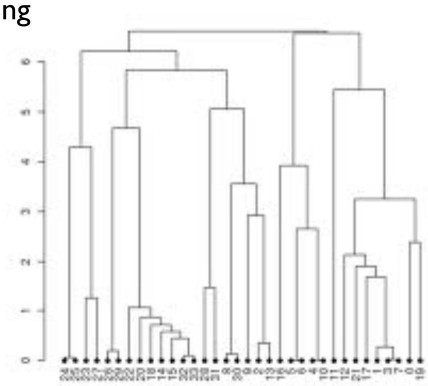


- Replace N rows
  - with M < N rows
  - that best exemplify the data
- Typical result:
  - Can throw out 80 to 90% of the rows without lossing accuracy
  - C. Chang, "Finding prototypes for nearest neighbor classifiers," IEEE Trans. on Computers, pp. 1179–1185, 1974.
- Benefits:
  - Outlier removal
  - Any downstream processing is faster
    - E.g. any $O(N^2)$ process is 100 times faster on N/10 of the data
  - Less errors in conclusions
    - Instance learner: classify according to nearest neighbors
    - If nearest neighbors further away, harder for data collection errors to cause wrong classifications
  - Easier to visualize
    - Fewer things to look at

Select rows

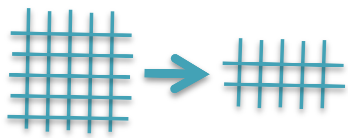


- Exponential time
  - Genetic algorithm to explore the 2^R subsets of rows.
    - When more rows than columns, even slower than the WRAPPER's $O(2^C)$ search
    - Y. Li, M. Xie, and T. Goh, "A study of project selection and feature weighting for analogy based software cost estimation," Journal of Systems and Software, vol. 82, pp. 241–252, 2009.
- Polynomial time: Greedy agglomerative clustering
  - Push every instance to its closest neighbor.
  - Build a synthetic example at each pair's median
  - Repeat for the synthetic points.
  - Prototypes are all nodes at level X of GAC tree
  - For R rows, $O(R^2)$



- TEAK = GAC plus …
  - Prune sub-trees with large variance
  - When to Use Data from Other Projects for Effort Estimation Ekrem Kocaguneli, Gregory Gay, Tim Menzies, Ye Yang, Jacky W. Keung, ASE 2010
- Linear-time
  - Rank ranges by frequency delta in different classes
  - Discard all rows that do not have the top R pivots

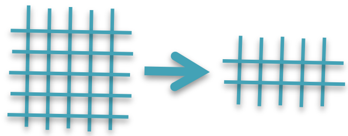
Select rows (with TEAK)



- To effort estimate a test instance, start at root of GAC tree
  - Move to nearest child
  - Stop at leaf or when sub-tree variance greater than super-tree
  - Estimate = median of instances in that sub-tree
- Compared with
  - linear regression,
  - neural nets,
  - analogy methods that use K=1,2,4 nearest neighbors (no variance pruning)
- Compared using
  - 20 * {shuffle rows, 3-way cross-val)
  - #wins - # losses (in a Wilcoxon, 95%)
  - Count number of times ranked first by this procedure
- Conclusion: row-selection using clustering + variance pruning is a good thing

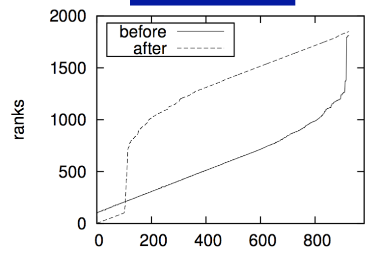
	TEAK	LR	NNet	Best(K)	k=1	k=16	k=2	k=4	k=8
MRE									
Cocomo81	▲								
Cocomo81e	▲								
Cocomo81o	▲								
Nasa93		▲							
Nasa93c2		▲							
Nasa93c5	▲								
Desharnais		▲							
Sdr	▲								
ISBSG-Banking	▲								
Count	6	3	0	0	0	0	0	0	0
Pred(25)									
Cocomo81	▲								
Cocomo81e			▲						
Cocomo81o	▲								
Nasa93		▲							
Nasa93c2		▲							
Nasa93c5	▲								
Desharnais		▲							
Sdr	▲								
ISBSG-Banking	▲								
Count	5	3	1	0	0	0	0	0	0
AR									
Cocomo81				▲					
Cocomo81e	▲								
Cocomo81o	▲								
Nasa93		▲							
Nasa93c2		▲							
Nasa93c5	▲								
Desharnais		▲							
Sdr	▲								
ISBSG-Banking	▲								
Count	6	3	0	1	0	0	0	0	0

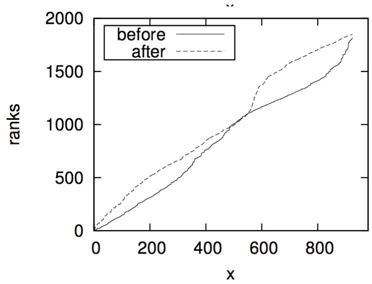
Select rows (with range pruning)



- For K in Klasses
  - Let NotK = Klasses – K
  - Let N1, N2 be number of rows with K and NotK classes
  - For C in columns
    - For R in range of column C
      - Let F1, F2 = frequency of C.R in K and NotK
      - Let x = F1 / N1 and y = F2 / N2
      - Let R.score = $x^2 / (x + y)$
        ;;; pivotal if R far more frequent in K than NotK
- Remove all rows without the top five pivots
  - If accuracy of reduced set decreases, then ABORT.
- For each instance, find distance needed to travel before a K=5 nearest neighbor algorithm changes the classification.
  - In the full data set
  - In the reduced data set
- Result:
  - Much charger to change classification in reduced data set
- Conclusion: if concerned about errors in data collection, use row selection (and less classes)

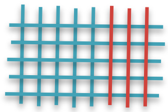
3 class system



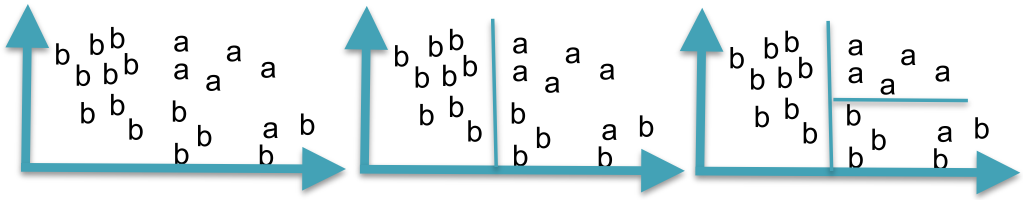


5 class system

Rotate
(add columns)



- Sometimes, the data's raw dimensions suffice for isolating the target concept..

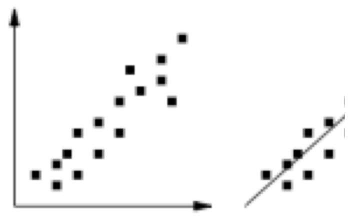


Theory 1:
if true, then "a"

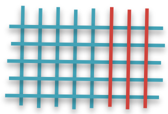
Theory 2::
if x > 1, then "a"

Theory 3::
if x > 1, and y > 1 then "a"

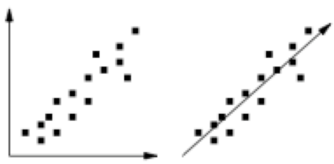
- But what if the target concept falls across and not along, the raw dimensions?



Rotate
(add columns)



- Synthesize a new dimension that combines the raw into something new
- Apply single-valued decomposition (SVD) to
  - the covariance matrix (principal component analysis, or PCA)
  - or the data table (latent semantic indexing, or LSI)
- PCA that produces a set of orthogonal "components"
  - Transforms C correlated variables into fewer uncorrelated "components".
  - Component[i] accounts for as much variability as possible.
  - Component[i+1] accounts for as much of the remaining variability as possible.

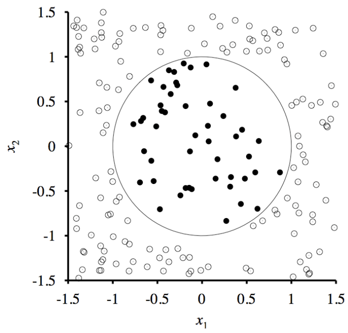


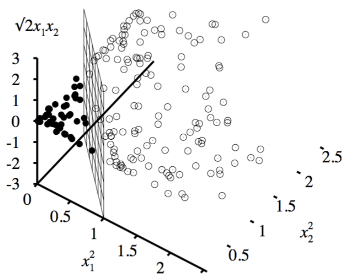
- Much easier to learn rules when dimensions match the data. E.g. a defect predictor:
- if comp[1] ≤ 0.180
  then NoDefects
  else if comp[1] > 0.180
      then if comp[1] ≤ 0.371 then NoDefects
      else if com[1] > 0.371 then Defects
- But it can be hard to explain that predictor:

  Comp[1] = 0.236*v(g) +0.222*ev(g)
  +0.236*iv(g) +0.241*n +0.238*v -0.086*l
  +0.199*d +0.216*i +0.225*e +0.236*b
  +0.221*t +0.241*lOCode +0.179*lOComment
  +0.221*lOBlank +0.158*lOCodeAndComment
  +0.163*uniq_Op +0.234*uniq_Opnd
  +0.241*total_Op +0.241*total_Opnd
  +0.236*branchCount

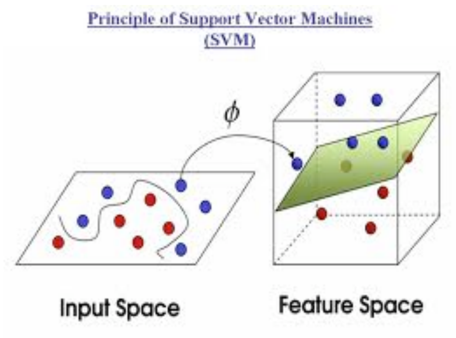
Rotate (add columns)

- Special transforms





- Support vector machines: construct a hyper-plane that separates classes



Clump rows (a.k.a. generalize)

- Ever notice that rows and rules have (nearly) the same syntax?
  - Age=young and wealth=rich and iq=high and class=happy
  - If age=old and wealth=rich and iq=high then happy
- But when we write rules, we only do it for frequently occurring patterns in the other rows
- "Clump rows" : replace them with a rule that covers many rows, but many only mention some of the columns
  - **If** age=old and wealth=rich **then** happy
- If you do this after clumping columns and selecting good rows and selecting good columns and (maybe) adding in good columns
  - Then the search space is very small
  - The exploring can be heavily biased by the other steps (e.g. look at great rows before dull ones)
  - And, hey presto, you've got a working data miner

Hints and tips (note: only my view)

- Always try clumping with discretization
  - So very simple
  - So experiment with / without discretization
- Always try column selection
  - Usually, massive reduction in the columns
- If the data won't fit in RAM,
  - try column selection first (use a linear-time approach)
  - then you can explore row selection by (say)
    - Era1: read first 1000 instances and apply row selection
    - Era[i+1]: read next 1000 records and ignore instances that fall close to the instances selected at Era[i]
- Try these last: PCA / Support vector machines
  - Benefits of PCA often achieved, or beaten by other column selectors
    - Hall, M. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans on Knowledge and Data Engineering. 15(3), November/December 2003
  - The FASTMAP heuristic FASTMAP, can do what PCA does, faster, scalable.
    - Faloutsos, C. and Lin, K. 1995. FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international Conference on Management of Data*
  - For text mining (PCA / LDA) vs TF*IDF never benchmarked

Coming next…

- Enough geeking
- What have you learned, that is useful, at the business level?
  - What can you <u>say</u> about how to do better SE?

GENERALITY (OR NOT)

Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

This hour

- Claim:
  - Current SE empirical practice asks for conclusions that are are external valid
    - apply to more than one domain
  - So far, such external valid conclusions are illusive
    - Despite decades of research.
- Implications:
  - The goal is wrong
  - Seek not for general theories
    - Only for the special lessons that work best on local projects.
- "W"
  - a baseline tool for generating those special lessons
  - Case-Based Reasoning vs Parametric Models Software Quality Optimization, Adam Brady, Tim Menzies, PROMISE 2010

What general lessons have we learned from all this data mining?

Only a small minority of PROMISE papers (11/64) discuss results that repeated in data sets from multiple projects

E.g. Ostrand, Weyuker, Bell PROMISE '08, '09

Same functional form

Predicts defects for generations of AT&T software

E.g. Turhan, Menzies, Bener PROMISE '08, '09

10 projects

Learn on 9

Apply to the 10th

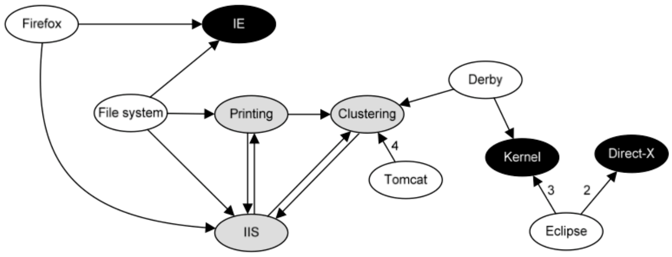
Defect models learned from NASA projects work for Turkish white goods software

Caveat: need to filter irrelevant training examples. See also

- When to Use Data from Other Projects for Effort Estimation Ekrem Kocaguneli, Gregory Gay, Tim Menzies, Ye Yang, Jacky W. Keung , ASE 2010
- B. Turhan, T. Menzies, A. Bener, and J. Distefano. On the relative value of cross-company and within- company data for defect prediction. Empirical Software Engineering, 68(2):278–290, 2009

What general lessons have we learned from all this data mining?

- The usual conclusion is that we learn that we can learn very little
- FSE'09: Zimmerman et
  - Defect models not generalizable
    - Learn "there", apply "here" only works in 4 of their 600+ experim
  - Opposite to Turhan'09
    - ?add relevancy filter



- ASE'09: Green, Menzies et al.
  - AI search for better software project options
  - Conclusions highly dependent on local business value proposition
- And others
  - TSE '01, '05: Shepperd et al
    - Any conclusion regarding "best" effort estimator varies by data sets, performance criteria, random selection train/test set
  - TSE'06: Menzies, Greenwald:
    - attributes selected by column selection vary wildly across projects

The gods are angry

- Fenton at PROMISE' 07 (invited talk)
  - "... much of the current software metrics research is inherently irrelevant to the industrial mix ..."
  - "... any software metrics program that depends on some extensive metrics collection is doomed to failure ..."
- Budgen & Kitchenham:
  - "Is Evidence Based Software Engineering mature enough for Practice & Policy?
  - Need for better reporting: more reviews.
  - Empirical SE results too immature for making policy.
  - B. Kitchenham D. Budgen, P. Brereton. Is evidence based software engineering mature enough for practice & policy? In 33rd Annual IEEE Software Engineering Workshop 2009 (SEW-33), Skvde, Sweden, 2009.
- Basili : still far to go
  - But we should celebrate the progress made over the last 30 years.
  - And we are turning the corner

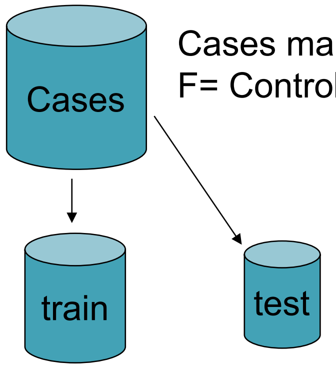
A new hope (actually, quite old)

- Experience factories
  - Method for find the special lessons that work for the local projects
- Basili'09 (pers. comm.):
  - "All my papers have the same form.
  - "For the project being studied, we find that changing X improved Y."
- Translation (mine):
  - Even if we can't find general models (which seem to be quite rare)….
  - … we can still research general methods for finding the special lessons that work best on the local projects

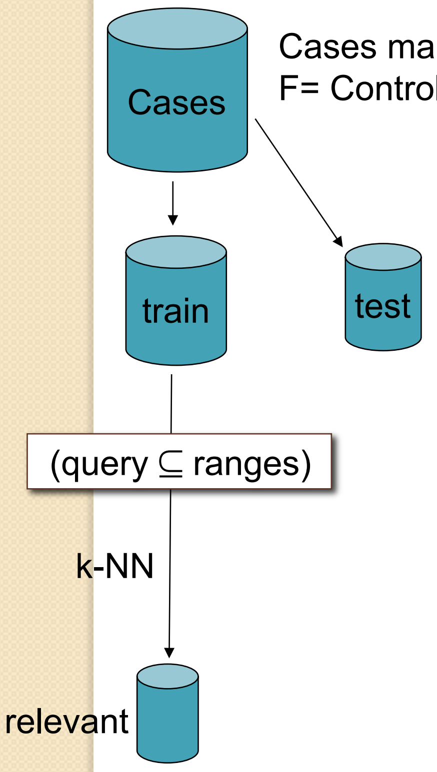
"W" + CBR: Preliminaries

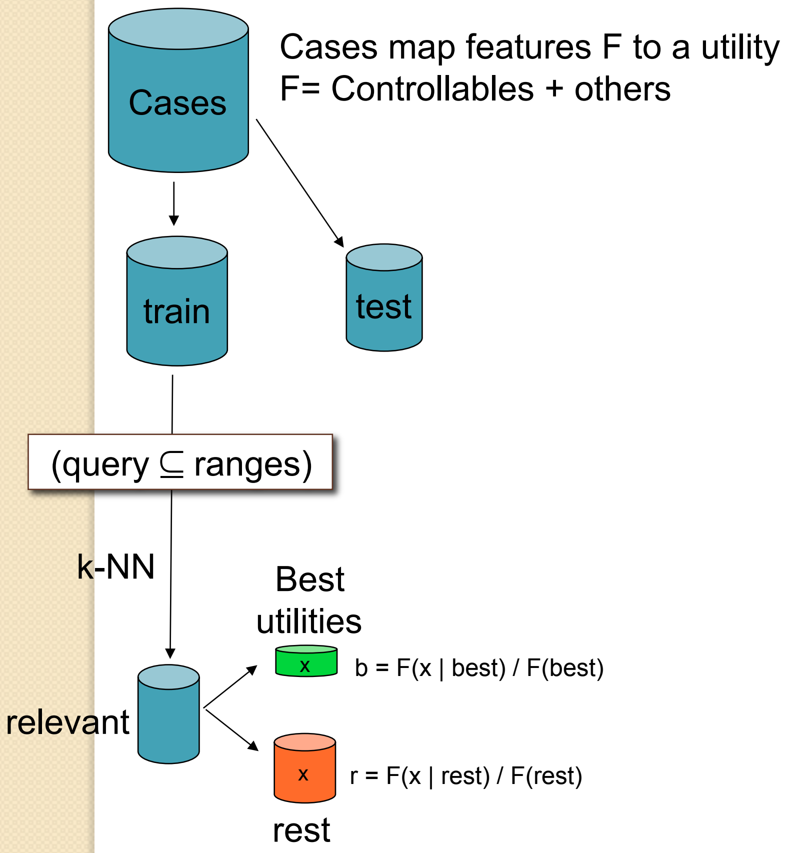
- "Query"
  - What kind of project you want to analyze; e.g.
    - Analysts not so clever,
    - High reliability system
    - Small KLOC
- "Cases"
  - Historical records, with their development effort
- Output:
  - A recommendation on how to change our projects in order to reduce development effort

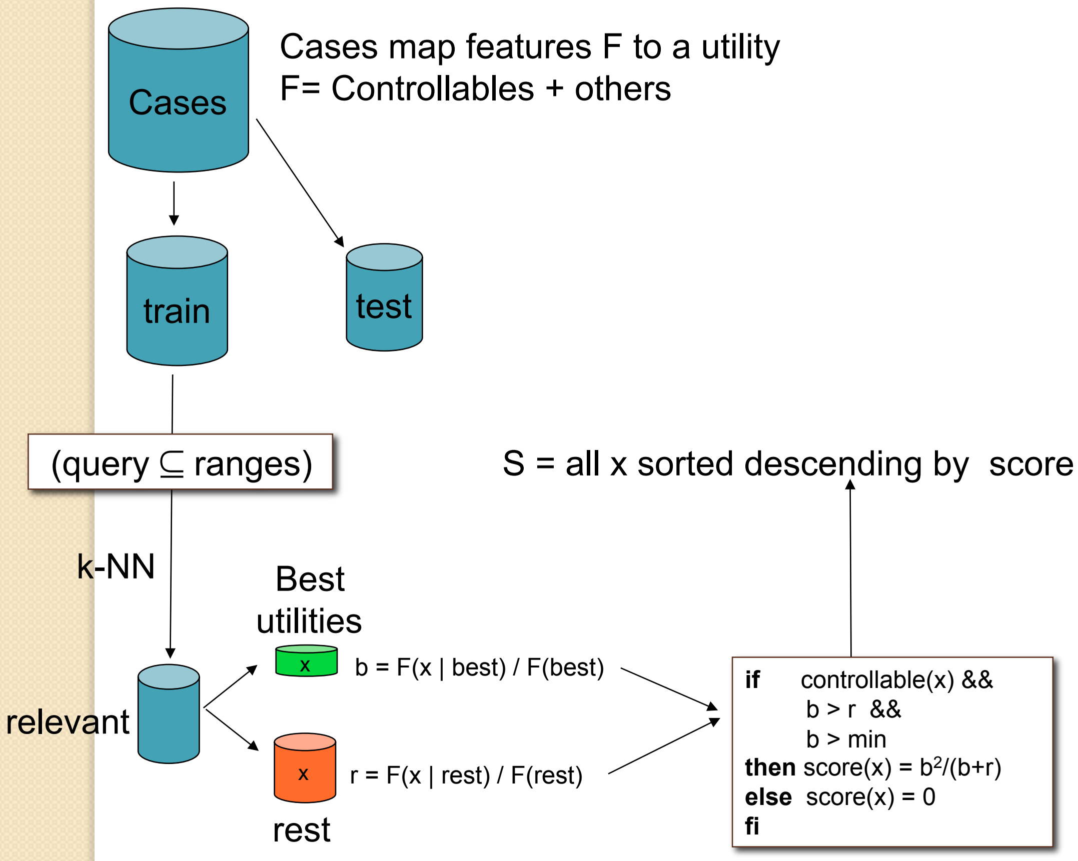
Cases map features F to a utility
F= Controllables + others

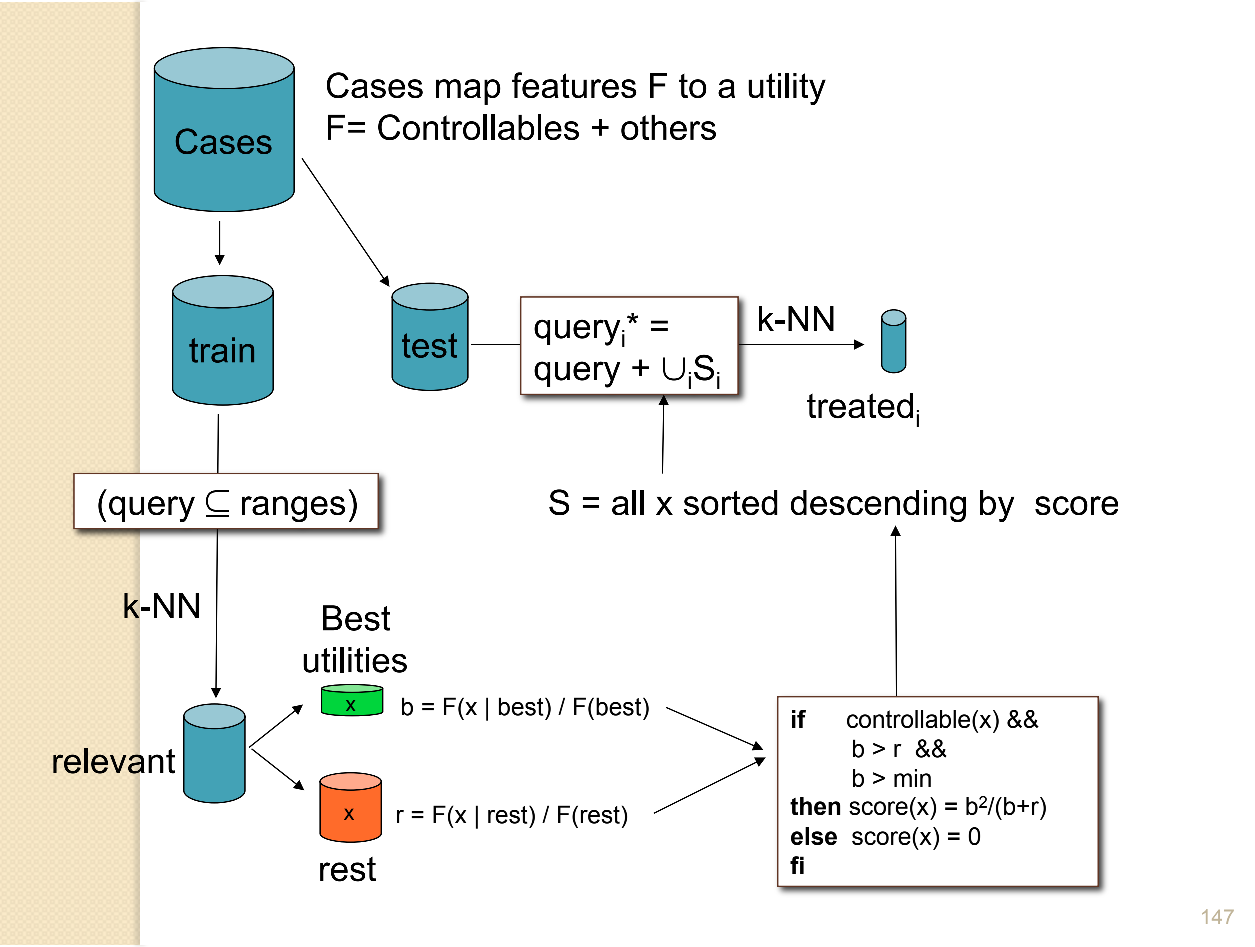


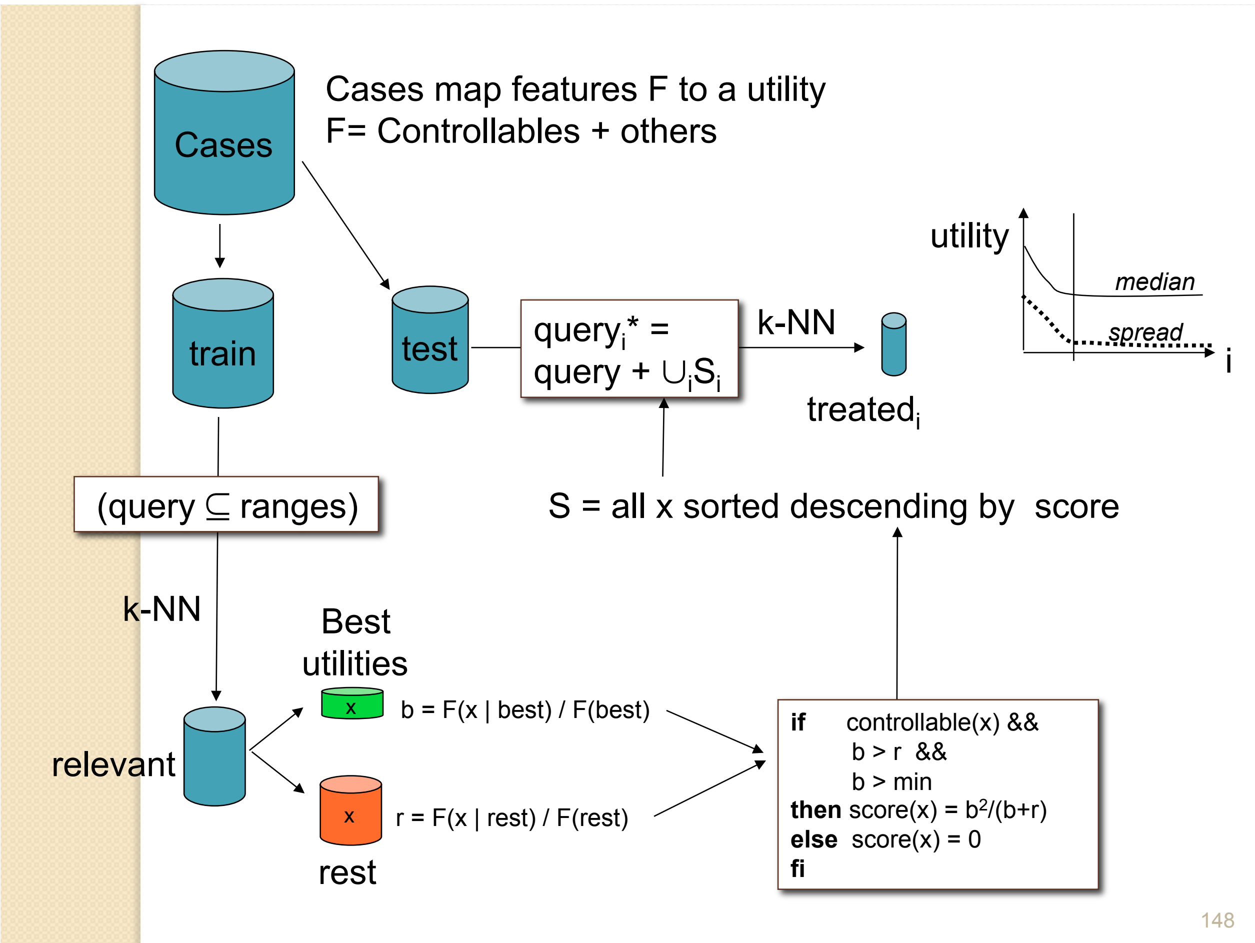
Cases map features F to a utility
F= Controllables + others

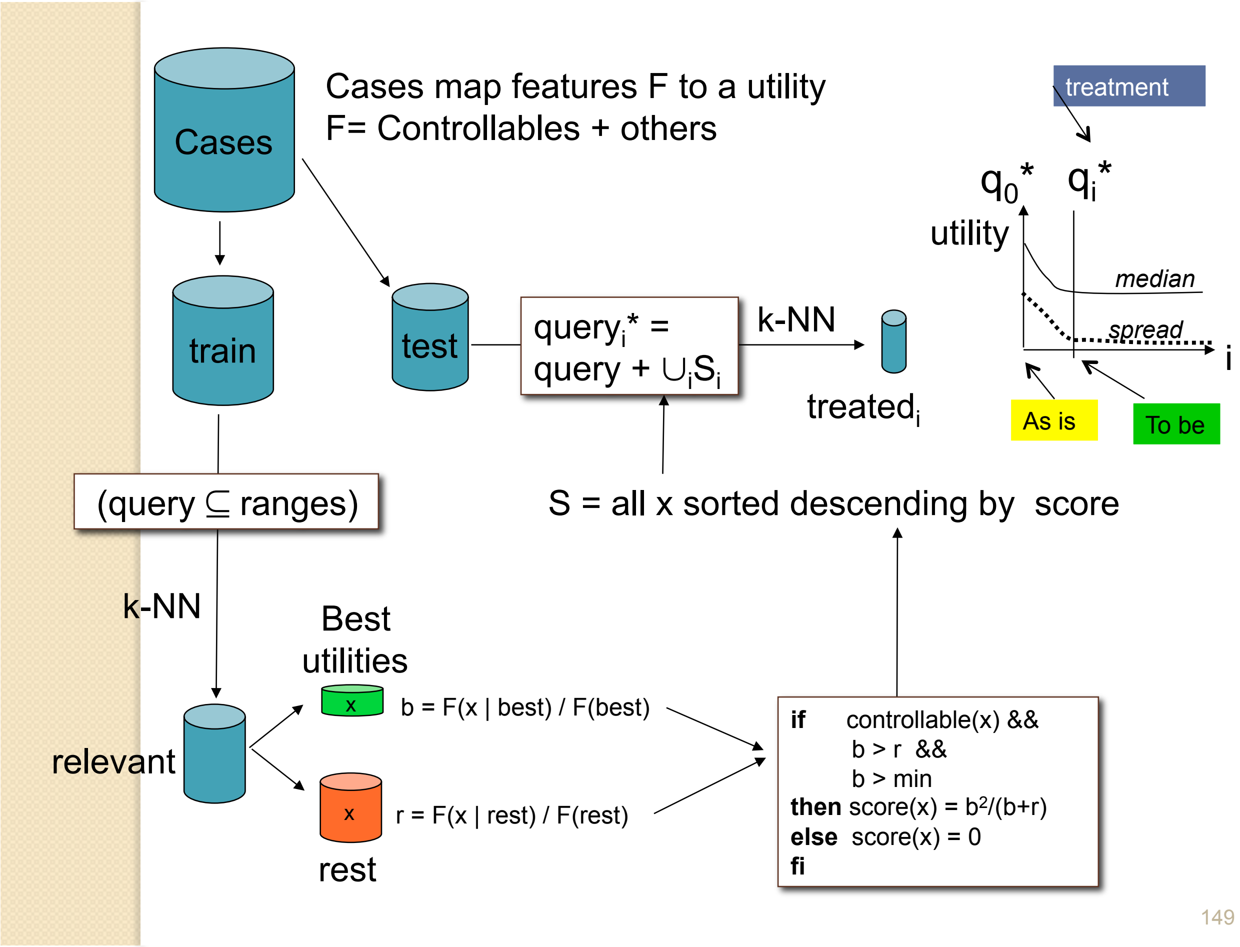



Cases map features F to a utility
F= Controllables + others
Cases
train
test
(query ⊆ ranges)
k-NN
relevant
Best
utilities
x
b = F(x | best) / F(best)
x
r = F(x | rest) / F(rest)
rest


Cases map features F to a utility
F= Controllables + others
Cases
train
test
(query ⊆ ranges)
S = all x sorted descending by score
k-NN
relevant
Best
utilities
x
b = F(x | best) / F(best)
x
r = F(x | rest) / F(rest)
rest
if controllable(x) &&
b > r &&
b > min
then score(x) = $b^2/(b+r)$
else score(x) = 0
fi

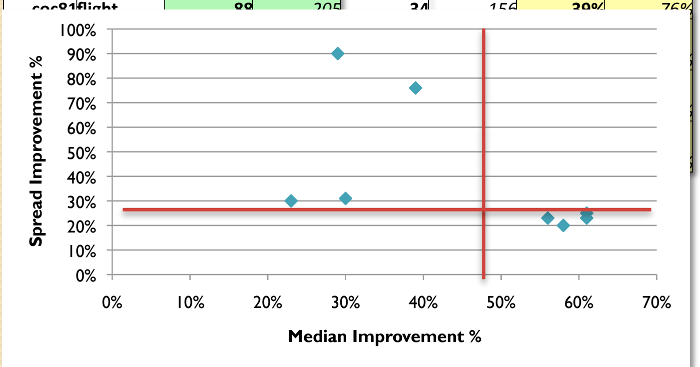

Cases map features F to a utility
F= Controllables + others
Cases
train
test
query$_i$* =
query + $\cup_i S_i$
k-NN
treated$_i$
(query ⊆ ranges)
S = all x sorted descending by score
k-NN
Best
utilities
relevant
x
b = F(x | best) / F(best)
x
r = F(x | rest) / F(rest)
rest
if controllable(x) &&
b > r &&
b > min
then score(x) = b^2/(b+r)
else score(x) = 0
fi


Cases map features F to a utility
F= Controllables + others
Cases
train
test
query$_i$* =
query + $\cup_i S_i$
k-NN
treated$_i$
utility
median
spread
i
(query ⊆ ranges)
S = all x sorted descending by score
k-NN
Best
utilities
relevant
x
b = F(x | best) / F(best)
x
r = F(x | rest) / F(rest)
rest
if controllable(x) &&
b > r &&
b > min
then score(x) = b^2/(b+r)
else score(x) = 0
fi



Results (distribution of development efforts in q_i*)

		X = as is		Y = to be		(X-Y) / X	
cases	*query*	*median*	*spread*	*median*	*spread*	*median*	*spread*
nasa93	ground	162	*349*	99	*80*	61%	*23%*
nasa93	flight	215	*398*	131	*100*	61%	*25%*
nasa93	osp	117.6	*396*	68	*79*	58%	*20%*
nasa93	osp2	170	*409*	95	*94*	56%	*23%*
coc81	flight	88	*205*	34	*156*	39%	*76%*



Cases from *promisedata.org/data*

Median = 50% percentile
Spread = 75% - 25% percentile

Improvement = (X - Y) / X
- X = as is
- Y = to be
- more is better

Usually:
- spread reduced to 25% of "as is"
- median reduction to 45% of "as is"

But that was so easy

- And that's the whole point
- Yes, finding the special lessons that work best on the local projects need not be difficult
- Strange to say…
  - There are no references in the CBR effort estimation literature for anything else than estimate = nearest neighbors
  - No steps beyond into planning , etc
  - Even though that next steps is easy

What should change?
$(q_j^* - q_o^*)$

cases	query	acap 3	apex 3	ltex 3	ltex 4	plex 3	pmat 3	pmat 4	sced 2	sced 3	stor 3	time 3	tool 3	# of Changes
nasa93	**ground**					100%	55%				85%			3
nasa93	**flight**					95%	70%				100%			3
nasa93	**osp**	95%	90%										100%	3
nasa93	**osp2**				100%			80%	85%					3
coc81	**flight**						60%					65%		2
coc81	**osp2**			55%	55%		65%			100%				4
coc81	**ground**						80%					100%		2
coc81	**osp**						65%			65%				2
Overall:		12%	11%	7%	19%	24%	49%	10%	11%	21%	23%	21%	13%	

Good news: improving estimates requires very few changes

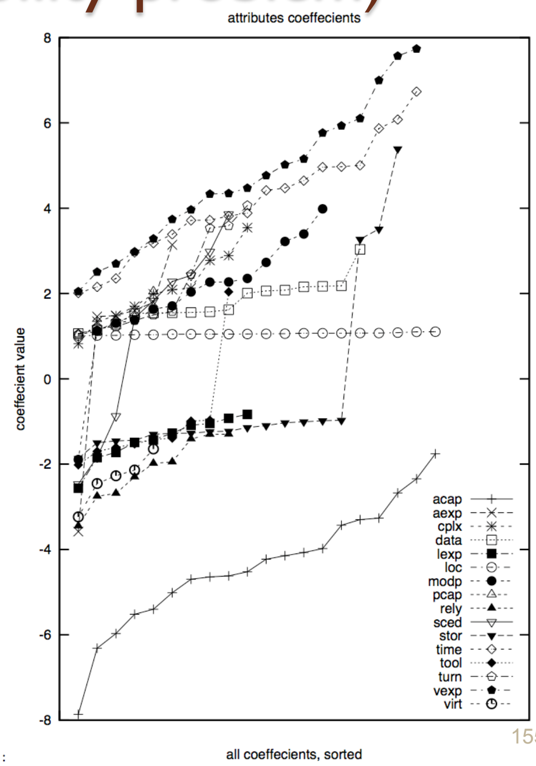
cases	query	*acap 3*	*apex 3*	*ltex 3*	*ltex 4*	*plex 3*	*pmat 3*	*pmat 4*	*sced 2*	*sced 3*	*stor 3*	*time 3*	*tool 3*	*# of Changes*
nasa93	**ground**					100%	55%				85%			3
nasa93	**flight**					95%	70%				100%			3
nasa93	**osp**	95%	90%										100%	3
nasa93	**osp2**				100%			80%	85%					3
coc81	**flight**						60%					65%		2
coc81	**osp2**			55%	55%		65%			100%				4
coc81	**ground**						80%					100%		2
coc81	**osp**						65%			65%				2
Overall:		12%	11%	7%	19%	24%	49%	10%	11%	21%	23%	21%	13%	

Not-so-good news: special lessons very local

cases	query	*acap 3*	*apex 3*	*ltex 3*	*ltex 4*	*plex 3*	*pmat 3*	*pmat 4*	*sced 2*	*sced 3*	*stor 3*	*time 3*	*tool 3*	*# of Changes*
nasa93	**ground**					100%	55%				85%			3
nasa93	**flight**					95%	70%				100%			3
nasa93	**osp**	95%	90%										100%	3
nasa93	**osp2**				100%			80%	85%					3
coc81	**flight**						60%					65%		2
coc81	**osp2**			55%	55%		65%			100%				4
coc81	**ground**						80%					100%		2
coc81	**osp**						65%			65%				2
Overall:		12%	11%	7%	19%	24%	49%	10%	11%	21%	23%	21%	13%	

Better than model-based approach (conclusion instability problem)

- 20 experiments, using 66% of the data (selected at random)
- Linear regression:
  - Effort = b_0 + sum of $b_i * x_i$
  - Followed by a greedy back-select to prune dull variables
- Results
  - LOC influence stable
  - Some variables pruned away half the time
  - Large ranges (max – min)
  - Nine attributes even change the sign on their coefficients



Q: Can we do better than "W"? A: Most certainly!

- "W" contains at least a dozen arbitrary design decisions
  - Which is best?
- But the algorithm is so simple
  - It should least be a baseline tool
  - Against which we compare supposedly more sophisticated methods.
  - The straw man
- Methodological advice
  - Before getting complex, get simple
  - Warning: often: my straw men don't burn

Certainly, we should always strive for generality

- But don't be alarmed if you can't find it.
- The experience to date is that,
  - with rare exceptions,
  - SE research does not lead to general models
- But that's ok
  - Very few others have found general models (in SE)
  - E.g. Turhan, Menzies, Ayse ESE journal '09
    - B. Turhan, T. Menzies, A. Bener, and J. Distefano. On the relative value of cross-company and within- company data for defect prediction. Empirical Software Engineering, 68(2): 278–290, 2009
  - E.g. Menzies et al ASE conference, 2010
    - When to Use Data from Other Projects for Effort Estimation Ekrem Kocaguneli, Gregory Gay, Tim Menzies, Ye Yang, Jacky W. Keung , ASE 2010
- Anyway
  - If there are few general results, there may be general methods to find the special lessons that work best on the local projects
    - Seek not "models as products"
    - But general "models to generate products"

Two definitions of "model"

- A hypothetical description of a complex entity or process.
  - Model as output from research machine
  - The "product" of research

- A plan to create, according to a model or models
  - Model of the research machine
  - The "generator" of products

- "W" is a general model generator.

If we can't find general models, is it science?

Popper '60: Everything is a "hypothesis"

- And the good ones have weathered the most attack
- SE "theories" aren't even "hypotheses"
- Karl Popper, Conjectures and Refutations, London: Routledge and Keagan Paul, 1963

Endres & Rombach '03: Distinguish "observations", "laws", "theory"

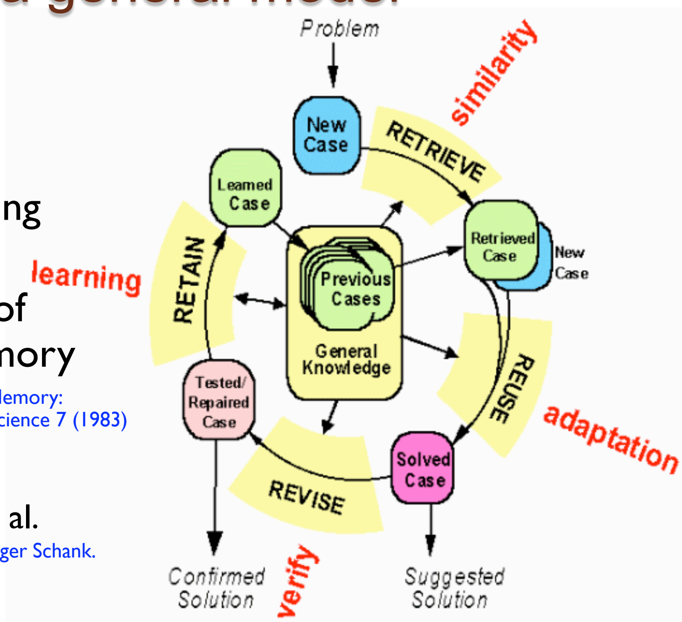
- Laws predict repeatable observations
- Theories explain laws
- Laws are either hypotheses (tentatively accepted) or conjectures (guesses)
- Rombach A. Endres, H.D. A Handbook of Software and Systems Engineering: Empirical Observa- tions, Laws and Theories. Addison Wesley, 2003.

Sjoberg '08 : 5 types of "theory":

- Building Theories in Software Engineering Dag I. K. Sjøberg, Tore Dyba Bente C. D. Anda and Jo E. Hannay, GUIDE TO ADVANCED EMPIRICAL SOFTWARE ENGINEERING2008,

1. Analysis (e.g. ontologies, taxonomies)
2. Explanation (but it is hard to explain "explanation")
3. Prediction (some predictors do not explain)
4. Explanation and prediction
5. "models" for design + action
   - Don't have to be "right"
   - Just "useful"
   - A.k.a. Endres & Rombach's "laws"?

Btw, constantly (re)building the special local models is a general model



- Case-based reasoning
- Kolodner's theory of reconstructive memory
  - Janet Kolodner, "Reconstructive Memory: A Computer Model," Cognitive Science 7 (1983)
- The Yale group
  - Shank & Riesbeck et al.
    - Riesbeck, Christopher, and Roger Schank. Inside Case-based Reasoning. Northvale, NJ: Erlbaum, 1989.
  - Memory, not models
  - Don't "think", remember

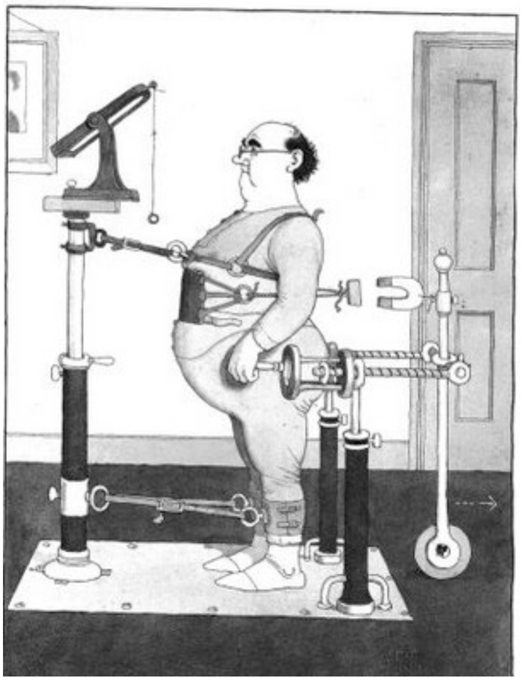
Kludges: they work

Ask some good old fashioned AI types

Minsky'86: "Society of Mind"
- The brain is a set of 1000+ kludges
  - Minsky, Marvin The Society of Mind, Simon and Schuster, New York. 1988.

Feigenbaum'83
- Don't take your heart attack to the Maths Dept.
  - Were they will diagnose and treat you using first principles
- Instead, go to the E.R room
  - Staffed by doctors who spent decades learning the quirks of drugs, organs, diseases, people, etc
  - Edward Feigenbaum and Pamela McCorduckThe Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World, Addison-Wesley (1983)

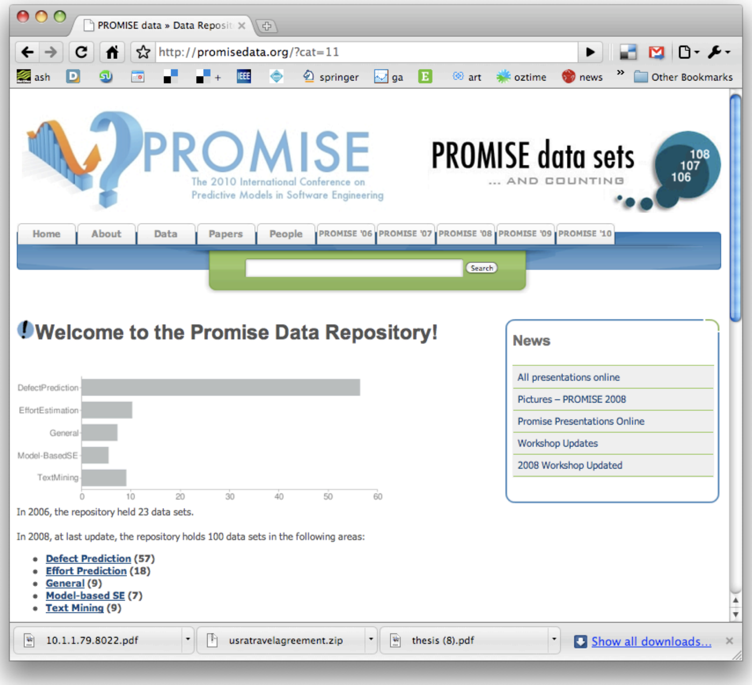
Seek out those that study kludges.
- You'll be treated faster
- You'll live longer



Disagree with me?

- Want to find some general conclusions on SE?
- Need to go somewhere to get a lot of data from different projects?

http://promisedata.org/data



Repository + annual conference. See you there?

Coming next…

- If all SE conclusions are biased by local conditions….
  - … Is this an enormous problem?
  - … Or a way to generate new insights?

BIAS (IS YOUR FRIEND)

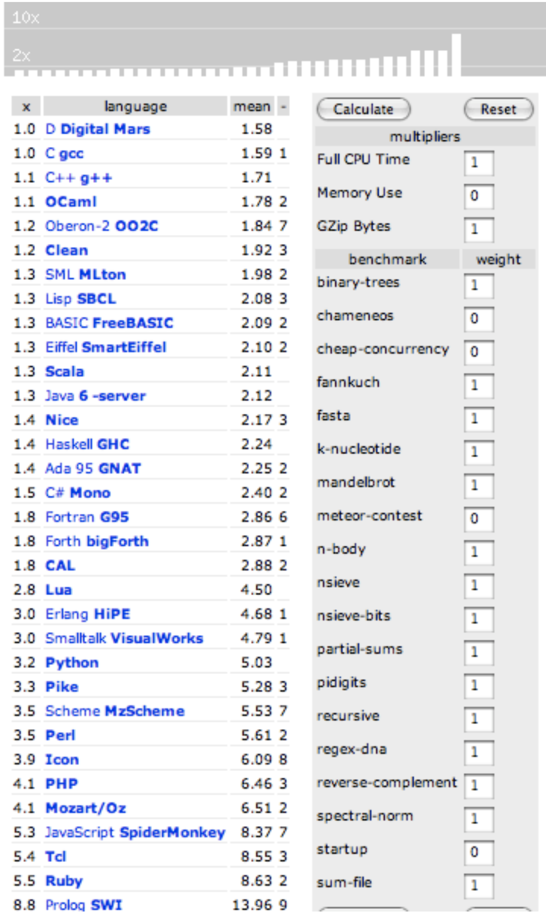
Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

Q: What is the "best" programming language?

A1: Eiffel! (of course)

A2: Depends on the bias



x	language	mean	-
1.0	D Digital Mars	1.58	
1.0	C gcc	1.59	1
1.1	C++ g++	1.71	
1.1	OCaml	1.78	2
1.2	Oberon-2 OO2C	1.84	7
1.2	Clean	1.92	3
1.3	SML MLton	1.98	2
1.3	Lisp SBCL	2.08	3
1.3	BASIC FreeBASIC	2.09	2
1.3	Eiffel SmartEiffel	2.10	2
1.3	Scala	2.11	
1.3	Java 6 -server	2.12	
1.4	Nice	2.17	3
1.4	Haskell GHC	2.24	
1.4	Ada 95 GNAT	2.25	2
1.5	C# Mono	2.40	2
1.8	Fortran G95	2.86	6
1.8	Forth bigForth	2.87	1
1.8	CAL	2.88	2
2.8	Lua	4.50	
3.0	Erlang HiPE	4.68	1
3.0	Smalltalk VisualWorks	4.79	1
3.2	Python	5.03	
3.3	Pike	5.28	3
3.5	Scheme MzScheme	5.53	7
3.5	Perl	5.61	2
3.9	Icon	6.09	8
4.1	PHP	6.46	3
4.1	Mozart/Oz	6.51	2
5.3	JavaScript SpiderMonkey	8.37	7
5.4	Tcl	8.55	3
5.5	Ruby	8.63	2
8.8	Prolog SWI	13.96	9

Calculate | Reset

multipliers	
Full CPU Time	1
Memory Use	0
GZip Bytes	1

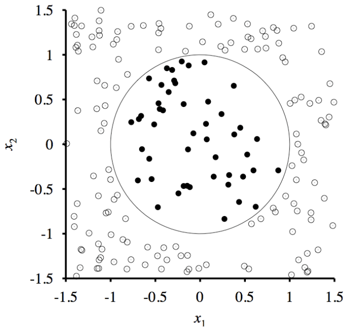
benchmark	weight
binary-trees	1
chameneos	0
cheap-concurrency	0
fannkuch	1
fasta	1
k-nucleotide	1
mandelbrot	1
meteor-contest	0
n-body	1
nsieve	1
nsieve-bits	1
partial-sums	1
pidigits	1
recursive	1
regex-dna	1
reverse-complement	1
spectral-norm	1
startup	0
sum-file	1

Bias is unavoidable

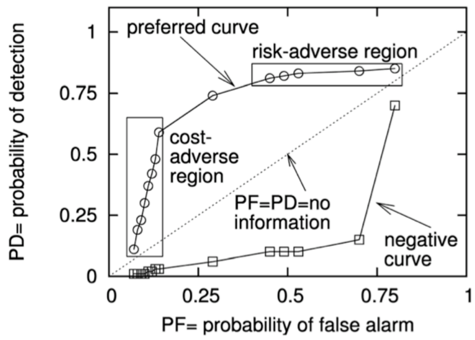
- Without bias
  - we can't assess relevance / irrelevance
- Without irrelevance,
  - we can't prune the data
- Without pruning,
  - we can't summarize
- Without summarization,
  - we can't generalize
- Without generalizing past experience
  - we can't predict the future
- So bias makes us blind (to some things)
  - But also, it lets us see (the future)

Sources of bias

- Sampling:
  - what data do you select in the pre-process?
- Language
  - E.g. if propositional, can't learn linear equations
- Search
  - When growing a model, what do you look at next?
- Over-fitting avoidance
  - When pruning a model, what is chopped first?
- Evaluation
  - Do you seek high accuracy? high support? What?



e.g. language bias. Hard to describe a circle if your language Is restricted to "Z op Value"



Different learners use different biases

- 48 learners, 320 combinations of biases
  - 48/320 = 15%
- Separate-and-conquer rule learningJ. FurnkranzArtificial Intelligence Review, 13, pages 3--54, 1999. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.4894

	Language Bias							Search Bias							Overfitting Avoidance		
	Static					Dyn.		Algorithm				Strategy					
Algorithm	Selectors	Literals	Synt. Restr.	Rel. Clichés	Rule Models	Lang. Hier.	Constr. Ind.	Hill-Climbing	Beam Search	Best First	Stochastic	Top-Down	Bottom-Up	Bidirectional	Pre-Pruning	Post-Pruning	Integrated
AQ	×							×	×			×					
AQ15	×							×	×			×				×	
AQ17	×						×	×	×			×					
ATRIS	×							×			×			×		×	
BEXA	×							×	×			×			×	×	
CHAMP	×	×	×				×	×	×			×			×		
CIPF	×						×	×				×				×	
CN2	×							×	×			×			×		
CN2-MCI	×						×	×	×			×			×		
CLASS	×									×		×					
DLG	×							×	×				×				
FOCL	×	×		×				×				×			×		
FOIL	×	×	×					×				×			×		
FOSSIL	×	×	×	×				×				×			×		
GA-SMART	×	×		×	×						×	×			×		
GOLEM		×	×					×					×				
GREEDY3	×							×				×				×	
GRENDEL					×			×				×					
GROW	×							×				×				×	
HYDRA	×	×						×				×					
IBL-SMART	×	×		×						×				×	×		
INDUCE	×	×						×	×			×					
I-REP, I²-REP	×	×	×	×				×				×					×
JoJo	×	×						×						×			
m-FOIL	×	×	×					×	×			×			×		
MILP	×	×	×								×	×			×		
ML-SMART	×	×		×				×	×	×		×			×		
NINA					×	×		×					×				
POSEIDON	×							×	×			×				×	
PREPEND	×							×				×					
PRISM	×							×				×					
PROGOL	×	×	×							×		×					
REP	×	×		×				×				×				×	
RIPPER	×							×				×				×	×
RDT					×			×				×					
SFOIL	×										×	×			×		
SIA	×										×		×			×	
SMART+	×	×		×	×			×	×	×	×	×			×		
SWAP-1	×							×						×		×	
TDP	×	×	×	×				×				×			×	×	

Bias can change conclusions

- Every data miner has its own bias
- Same data, different data miners, different conclusions
  - Changing biases changes what we best believe
- So, relativistic soup?
  - No basis to make policies, to plan for the future?
  - Data mining is a pack of lies?
    - No more than any other inductive generalization process

Nothing is "right", but some things are "useful"

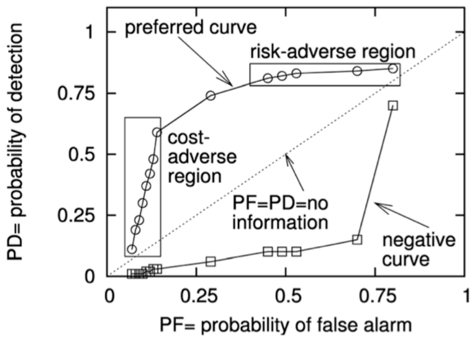
- Sure, one data set supports many theories.
  - But there are many many more theories that are unsupported.
- No model is *right*, but some things are *useful*
    - (perform well on test data)
    - George Box
- And many many many more ideas are *useless*
    - Can't make predictions
    - Not defined enough to support (possible) refutation

Embrace bias

- When reporting a conclusion, report the biases that generated it.
- Make it a first class modeling construct
- Example #1: "W"
  - Recall the sampling bias of "W"
  - Different biases (the query "q") lead to different conclusions
  - Case-Based Reasoning vs Parametric Models Software Quality Optimization, Adam Brady, Tim Menzies, PROMISE 2010
- Example #2: "WHICH"
  Defect prediction from static code features: current results, limitations, new approaches. Tim Menzies, Zach Milton, Burak Turhan, Bojan Cukic, Yue Jiang and Ayşe Bener Automated Software Engineering (2010) 17: 375-407, July 23, 2010. http://menzies.us/pdf/10which.pdf

Evaluation Bias #1 : AUC(Pd,Pf)

- Much research
- Little recent improvement:
  - Lessmann, S., Baesens, B., Mues, C., Pietsch, S.: Benchmarking classification models for software defect prediction: a proposed framework and novel findings. IEEE Trans. Softw. Eng. (2008)
- A shallow well?
  - And we've reached the bottom?

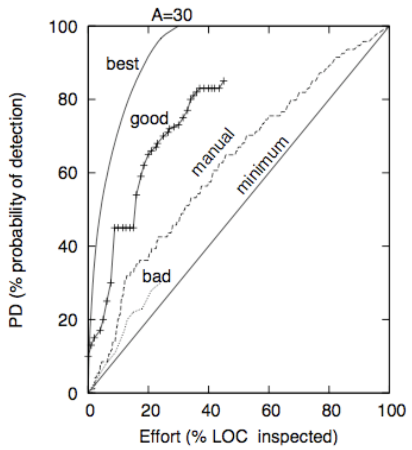


Evaluation Bias #2 : AUC(Pd,effort)

- Inspect fewest LOC to find the most bugs.

- Arisholm and Briand[2006]
  - E. Arisholm and L. Briand. Predicting fault-prone components in a java legacy system. In 5th ACM-IEEE International Symposium on Empirical Software Engineering (ISESE), Rio de Janeiro, Brazil, September 21-22, 2006. Available from http://simula.no/research/engineering/publications/Arisholm.2006.4.
  - For a budget-conscious team,
  - if X% of modules predicted to be faulty
  - But they contain ≤X% of the defects,
  - Then that defect predictor is not useful
  - i.e. their bias is pd>effort

- Operationalizing their bias:
  - Find modules triggered by the learner
  - Sort them in ascending order of size
  - Assume human inspectors find Δ of the defects in the triggered modules
  - Use ratoos of "best" effort-vs-pd curve
    - "best"only triggers on defective modules
    - Note: Δ cancels out



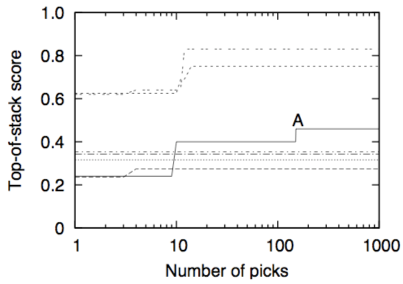
"bad" : worse than manual
"good" : beats manual

Implementing a bias-specific learner

- All learners have an search bias S and an evaluation bias E . e.g. C4.5:
  - S = infogain
  - E = pd, pf, accuracy, etc
- Note: usually, not(S = E)
- Question: What if we make S = E ?
  - Answer: "WHICH"

Implementing a bias-specific learner (more)

- Fuzzy beam search
1. Discretize all numeric features.
2. Sort all ranges using E on to a stack
3. Pick any 2 items near top-of-stack
4. Combine items, score them with E, insert them into the sorted stack.
5. Goto 3
- Note: no S and E is customizable
- But when to stop? (Use 200 picks)



Top of stack stabilizes quickly (UCI data).

Results:
10 random orderings * 3-way cross-val

- 10 sets of static code features from NASA, Turkish whitegoods
- "Rank" computed using Mann-Whitney U test (95%)
- E = AUC(effort, pd)
- Micro20: training on 20 defective + 20 non-defective

rank	treatment	median "best" %	2nd quartile, median, 3rd quartile
1	WHICH	87.3	
2	micro20	76.3	
3	NB	64.2	
3	manual	64.2	
4	C4.5	23.1	
4	jRip	17.7	
			50%

- WHICH destroys classic learners
  - Which were built to optimize accuracy
  - So bias changes everything
- BTW, once again a shallow well
  - we do not need much data to do it (40 examples).

Discussion

- Bias changes everything
- But this is not a problem
  - It is a research opportunity
- What biases are current in industrial SE?
  - How do they effect our conclusions?

Coming up…

- Let's focus on one particular bias
  - Evaluation

EVALUATION (DOES IT REALLY WORK?)

Road map

1. Data mining & SE (overview)
2. Data mining tools (guided tour of "WEKA")
3. Data "carving" (core operators of DM)
4. Generality (or not)
5. Bias (is your friend)
6. Evaluation (does it really work?)

Wolfgang Pauli: the conscience of physics

- The critic to whom his colleagues were accountable.
- Scathing in his dismissal of poor theories
  - often labeling it *ganz falsch*, utterly false.
- But "*ganz falsch*" was not his most severe criticism,
  - He hated theories so unclearly presented as to be
    - untestable
    - unevaluatable,
  - Worse than wrong
    - because they could not be proven wrong.
  - Not properly belonging within the realm of science,
    - even though posing as such.
  - Famously, he wrote of of such unclear paper:
    - "This paper is right. It is not *even* wrong."

Lesson: evaluation is important

So evaluation is important

- We saw above how "evaluation" actually became "the learning algorithm"
  - The "WHICH" experiment
- So evaluation is not some post hoc bolt,
  - Only to be explored as an after-thought once the work is done
  - Rather, it is an integral part of the work
  - Best to be get continual feedback from your algorithms as you go along
- BTW: to fail at a data mining Ph.D.
  - Plan to start evaluation in year3

Lesson: build the evaluation rig FIRST

Performance measures for continuous classes

- Absolute residual = AR = (actual – predicted)
- Relative error = RE = AR/actual
- Magnitude of relative error = MRE = abs(RE)
  - Can be surprisingly large (see next slide)
- MER = AR / predicted
- Median MRE, Median MER
- Mean MRE (severely deprecated)
  - Tron Foss, Erik Stensrud, Barbara Kitchenham, Ingunn Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Transactions on Software Engineering, vol. 29, no. 11, pp. 985-995, Nov. 2003*
- Pred(X) = percents of RE within X% of actual
  - E.g. if 80% of the predictions are with 30% of actual then Pred(30) = 80
  - Note Pred will not notice if a small number of predictions are really bad

Performance measures for discrete classes

a	b	c	<-- classified as
15	0	0	a= Iris-setosa
0	19	0	b=Iris-versicolor
0	2	15	c=Iris-virginica

consider "TRUE"= iris-virginica and FALSE= everything else

	Ground truth FALSE	TRUE
detector silent	A =34	B = 2
detector loud	C= 0	D = 15

accuracy	(A+D)/(A+B+C+D)	(34+15)/51	96%
recall (pd)	D/(B+D)	15/(2+15)	88%
false alarm (pf)	C/(A+C)	0/34	0%
precision	D/(C+D)	15/(15+0)	100%
f-measure	2*prec*pd/ (prec+pd)	2*1*0.88/ (1+0.88)	94%

Collect separately for each class.
Repeat 10 times (re-ordering data) * 10-way
Repeat for each learner * discretizer * x * y * ….

Instability and Precision

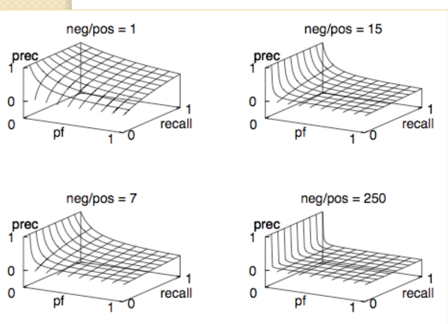
- Tim Menzies, Alex Dekhtyar, Justin S. Di Stefano, Jeremy Greenwald: Problems with Precision: A Response to "Comments on 'Data Mining Static Code Attributes to Learn Defect Predors", IEEE Transactions on Software Engineering, Volume 33, Number 9, September 2007

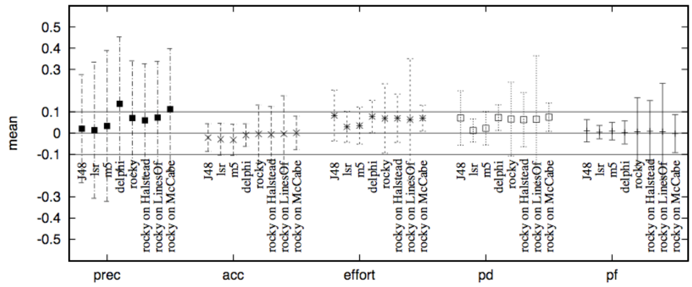
$$
\left.
\begin{aligned}
pd = recall &= \frac{D}{B+D}\\
pf &= \frac{C}{A+C}\\
prec = precision &= \frac{D}{D+C}\\
acc = accuracy &= \frac{A+D}{A+B+C+D}\\
selectivity &= \frac{C+D}{A+B+C+D}\\
neg/pos &= \frac{A+C}{B+D}
\end{aligned}
\right\}
$$

$$prec = \frac{D}{D+C} = \frac{1}{1+\frac{C}{D}} = \frac{1}{1+neg/pos \cdot pf/recall}$$

which can be rearranged to

$$pf = \frac{pos}{neg} \cdot \frac{(1-prec)}{prec} \cdot recall$$





Lesson: avoid precision when target class is rare

Strange tales of performance measures

		Truth 0	Truth 1	
Detector	0	A	B	Prec = D/(C+D) Acc = (A+D) / (A+B+C+D)
	1	C	D	PD = D/(B+D) PF = C / (A+C)
Detector	0	0	0	PF = PD = 1 (so detection does not preclude
	1	10	10	bad false alarm rates)
Detector	0	80	15	Acc = 85% (so when target is comparatively
	1	0	5	PD = 33% rare, Acc does not predict for PD)
Detector	0	100	0	Acc = 100% (so highly accurate predictors can
	1		0	PD = 0 miss everything)
Detector	0	0	10	PD = 80% (so PD does not predict
	1	50	40	Prec = 44% for precision)

Lesson: avoid Accuracy; consider both PD and Pf

Evaluation is time-consuming

```
analysis1(){
  local origdata=$1
  local outstats=$2
  local nattrs="2 4 6 8 10 12 14 16 18 20"
  local learners="nb10 j4810 zeror10 oner10 adtree10"
  local reducers="infogain chisquared oneR"
  local tmpred=$Tmp/red
  echo "n,reducer,learner,accuracy" > $outstats

  for n in $nattrs; do
    for reducer in $reducers; do
       $reducer $origdata $n $tmpred
       for learner in $learners; do
         accur=`$learner $tmpred.arff | acc
         out="$n,$reducer,$learner,$accur"
         blabln $out
        echo $out >> $outstats
       done
    done
  done
}
```

Learners * data sets * pre-processors
- Repeated 30 – 100 times for statistical validity

Time to run experiments
- Hours to days (first time)

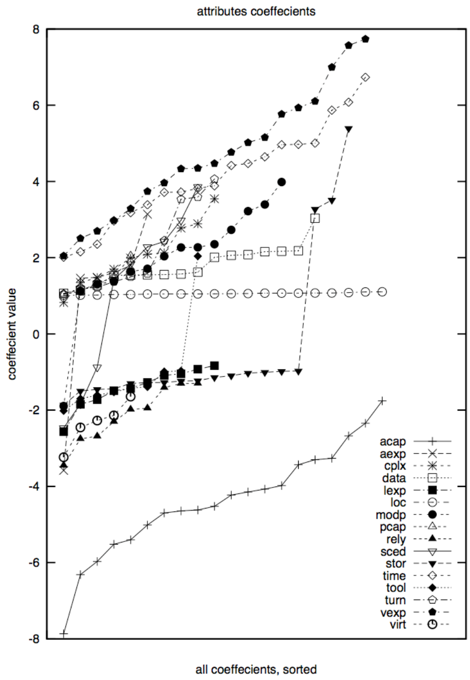
Then comes the “oh dear moment”
- Do it all again

1 masters = 20 days of CPU (for evaluation)

Lesson: start your evaluations ASAP

Variance problems (more)

- "Simple Software Cost Estimation: Safe or Unsafe?" by Tim Menzies and Zhihao Chen and Dan Port and Jairus Hihn. Proceedings, PROMISE workshop, ICSE 2005 2005 .Available fromhttp://menzies.us/pdf/05safewhen.pdf .
- 20 experiments, using 66% of the data (selected at random)
- Linear regression:
  - Effort = b_0 + sum of $b_i * x_i$
  - Followed by a greedy back-select to prune dull variables
- Results
  - LOC influence stable
  - Some variables pruned away half the time
  - Large ranges (max – min)
  - Nine attributes even change the sign on their coefficients



Lesson: avoid Accuracy; consider both PD and Pf

Evaluation (using hypothesis testing) is contentious

- Statistical significance tests of the form (H0 vs H1) are a 'potent but sterile intellectual rake who leaves . . . no viable scientific offspring'.
    - Cohen J. 1988. The earth is round ($p < .05$). American Psychologist 49: 997 – 1003.
- Consider one study showing that, using significance testing, estimates from multiple sources are no better than those from a single source.
    - How to explain 31 other studies where multiple sources out-performed single source by 3.4 to 23.4% (average = 12.5%).
    - Odds of that happening at random?
        - 2^31 < less than a billionth
        - Armstrong JS. 2007. Significance tests harm progress in forecasting. International Journal of Forecasting 23: 21 – 327.

Table: Error Reductions from Combining Ex Ante Forecasts

Study	Methods	Components	Criterion	Data	Situation	Validation Forecasts	Forecast Horizon	Percent error reduction
Levine (1960)	intentions	2	MAPE	annual	capital expenditures	6	1	18.0
Okun (1960)	"	2	"	"	housing starts	6	1	7.0
Landefeld & Seskin (1986)	"	2	MAE	"	plant & equipment	11	1	20.0
Armstrong et al. (2000)	"	4	RAE	"	consumer products	65	varied	5.5
Winkler & Poses (1993)	expert	4	Brier	cross-section	survival of patients	231	varied	12.2
Thorndike (1938)	"	4 to 6	% wrong	"	knowledge questions	30	varied	6.6
Makridakis et al. (1993)	"	5	MAPE	monthly	economic time series	322	1 thru 14	19.0
Richards & Fraser (1977)	"	5	"	annual	company earnings	213	1	8.1
Batchelor & Dua (1995)	"	10	MSE	"	macroeconomic	40	1	16.4
Kaplan et al. (1950)	"	26	% wrong	cross-section	technology events	16	varied	13.0
Zarnowitz (1984)	"	79	RMSE	quarterly	macroeconomic	288	1	10.0
Sanders & Ritzman (1989)	extrapolation	3	MAPE	daily	public warehouse	260	1	15.1
Makridakis & Winkler (1983)	"	5	"	monthly	economic time series	617	18	24.2
Makridakis et al. (1993)	"	5	"	"	"	322	1 thru 14	4.3
Lobo (1992)	"	5	"	quarterly	company earnings	6,560	1 thru 4	13.6
Schnaars (1986)	"	7	"	annual	consumer products	1,412	1 thru 5	20.0
Landefeld & Seskin (1986)	econometric	2	MAE	annual	plant & equipment	7	1	21.0
Clemen & Winkler (1986)	"	4	MAD	quarterly	GNP (real & nominal)	45	1 thru 4	3.4
Shamseldin et al. (1997)	"	5	MAPE	annual	rainfall runoff	22	1	9.4
Lobo (1992)	expert/extrap	2	MAPE		company earnings	6,560	1 thru 4	11.0
Lawrence et al. (1986)	"	3	"	annual monthly	economic time series	1,224	1 thru 18	10.7
Sanders & Ritzman (1989)	"	3	"	daily	public warehouse	260	1	15.5
Lobo & Nair (1990)	"	4	"	annual	company earnings	768	1	6.4
Landefeld & Seskin (1986)	intentions/econ	2	MAE	annual	plant & equipment	11	1	11.5
Vandome (1963)	extrap/econ	2	MAPE	quarterly	macroeconomic	20	1	10.1
Armstrong (1985)	"	2	"	annual	photo sales by country	17	6	4.2
Weinberg (1986)	expert/econ	2	"	cross-section	performing arts	15	varied	12.5
Bessler & Brandt (1981)	exprt/extrap/econ	3	"	quarterly	cattle & chicken prices	48	1	13.6
Fildes (1991)	"	3	MAE	annual	construction	72	1 & 2	8.0
Brandt & Bessler (1983)	"	6	MAPE	quarterly	hog prices	24	1	23.5
							Unweighted average	**12.5**

Lesson: Don't base conclusions on just hypothesis testing

Evaluation is humbling

- All that clever programming, then…
    - Then simpler ideas do as well, or better, than the more sophisticated
- Example
    - E.g. "Bayes" = simple correlation unaware learner
    - C4.5 = more sophisticated method, correlation aware
    - And no evidence here that the added complexity of C4.5 is better than dumb Bayes
    - Pedro Domingos and Michael J. Pazzani, On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Volume 29, number 2-3, pages 103-130, 1997

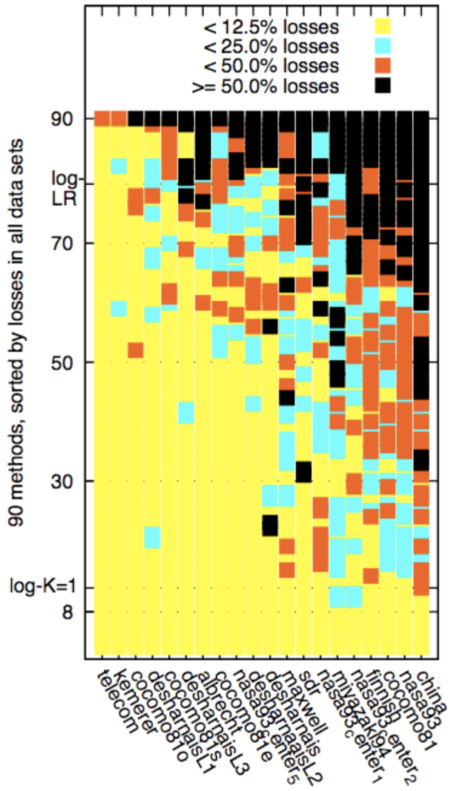
Table 1. Classification accuracies and sample standard deviations, averaged over 20 random training/test splits. "Bayes" is the Bayesian classifier with discretization and "Gauss" is the Bayesian classifier with Gaussian distributions. Superscripts denote confidence levels for the difference in accuracy between the Bayesian classifier and the corresponding algorithm, using a one-tailed paired t test: 1 is 99.5%, 2 is 99%, 3 is 97.5%, 4 is 95%, 5 is 90%, and 6 is below 90%.

Data Set	Bayes	Gauss	C4.5	PEBLS	CN2	Def.
Audiology	73.0 ± 6.1	73.0 ± 6.1^{6}	72.5 ± 5.8^{6}	75.8 ± 5.4^{3}	71.0 ± 5.1^{5}	21.3
Annealing	95.3 ± 1.2	84.3 ± 3.8^{1}	90.5 ± 2.2^{1}	98.8 ± 0.8^{1}	81.2 ± 5.4^{1}	76.4
Breast cancer	71.6 ± 4.7	71.3 ± 4.3^{6}	70.1 ± 6.8^{5}	65.6 ± 4.7^{1}	67.9 ± 7.1^{1}	67.6
Credit	84.5 ± 1.8	78.9 ± 2.5^{1}	85.9 ± 2.1^{3}	82.2 ± 1.9^{1}	82.0 ± 2.2^{1}	57.4
Chess endgames	88.0 ± 1.4	88.0 ± 1.4^{6}	99.2 ± 0.1^{1}	96.9 ± 0.7^{1}	98.1 ± 1.0^{1}	52.0
Diabetes	74.5 ± 2.4	75.2 ± 2.1^{6}	73.5 ± 3.4^{5}	71.1 ± 2.4^{1}	73.8 ± 2.7^{6}	66.0
Echocardiogram	69.1 ± 5.4	73.4 ± 4.9^{1}	64.7 ± 6.3^{1}	61.7 ± 6.4^{1}	68.2 ± 7.2^{6}	67.8
Glass	61.9 ± 6.2	50.6 ± 8.2^{1}	63.9 ± 8.7^{6}	62.0 ± 7.4^{6}	63.8 ± 5.5^{6}	31.7
Heart disease	81.9 ± 3.4	84.1 ± 2.8^{1}	77.5 ± 4.3^{1}	78.9 ± 4.0^{1}	79.7 ± 2.9^{3}	55.0
Hepatitis	85.3 ± 3.7	85.2 ± 4.0^{6}	79.2 ± 4.3^{1}	79.0 ± 5.1^{1}	80.3 ± 4.2^{1}	78.1
Horse colic	80.7 ± 3.7	79.3 ± 3.7^{1}	85.1 ± 3.8^{1}	75.7 ± 5.0^{1}	82.5 ± 4.2^{2}	63.6
Hypothyroid	97.5 ± 0.3	97.9 ± 0.4^{1}	99.1 ± 0.2^{1}	95.9 ± 0.7^{1}	98.8 ± 0.4^{1}	95.3
Iris	93.2 ± 3.5	93.9 ± 1.9^{6}	92.6 ± 2.7^{6}	93.5 ± 3.0^{6}	93.3 ± 3.6^{6}	26.5
Labor	91.3 ± 4.9	88.7 ± 10.6^{6}	78.1 ± 7.9^{1}	89.7 ± 5.0^{6}	82.1 ± 6.9^{1}	65.0
Lung cancer	46.8 ± 13.3	46.8 ± 13.3^{6}	40.9 ± 16.3^{5}	42.3 ± 17.3^{6}	38.6 ± 13.5^{3}	26.8
Liver disease	63.0 ± 3.3	54.8 ± 5.5^{1}	65.9 ± 4.4^{1}	61.3 ± 4.3^{6}	65.0 ± 3.8^{3}	58.1
LED	62.9 ± 6.5	62.9 ± 6.5^{6}	61.2 ± 8.4^{6}	55.3 ± 6.1^{1}	58.6 ± 8.1^{2}	8.0
Lymphography	81.6 ± 5.9	81.1 ± 4.8^{6}	75.0 ± 4.2^{1}	82.9 ± 5.6^{6}	78.8 ± 4.9^{3}	57.3
Post-operative	64.7 ± 6.8	67.2 ± 5.0^{3}	70.0 ± 5.2^{1}	59.2 ± 8.0^{2}	60.8 ± 8.2^{4}	71.2
Promoters	87.9 ± 7.0	87.9 ± 7.0^{6}	74.3 ± 7.8^{1}	91.7 ± 5.9^{3}	75.9 ± 8.8^{1}	43.1
Primary tumor	44.2 ± 5.5	44.2 ± 5.5^{6}	35.9 ± 5.8^{1}	30.9 ± 4.7^{1}	39.8 ± 5.2^{1}	24.6
Solar flare	68.5 ± 3.0	68.2 ± 3.7^{6}	70.6 ± 2.9^{1}	67.6 ± 3.5^{6}	70.4 ± 3.0^{2}	25.2
Sonar	69.4 ± 7.6	63.0 ± 8.3^{1}	69.1 ± 7.4^{6}	73.8 ± 7.4^{1}	66.2 ± 7.5^{5}	50.8
Soybean	100.0 ± 0.0	100.0 ± 0.0^{6}	95.0 ± 9.0^{3}	100.0 ± 0.0^{6}	96.9 ± 5.9^{3}	30.0
Splice junctions	95.4 ± 0.6	95.4 ± 0.6^{6}	93.4 ± 0.8^{1}	94.3 ± 0.5^{1}	81.5 ± 5.5^{1}	52.4
Voting records	91.2 ± 1.7	91.2 ± 1.7^{6}	96.3 ± 1.3^{1}	94.9 ± 1.2^{1}	95.8 ± 1.6^{1}	60.5
Wine	96.4 ± 2.2	97.8 ± 1.2^{3}	92.4 ± 5.6^{1}	97.2 ± 1.8^{6}	90.8 ± 4.7^{1}	36.4
Zoology	94.4 ± 4.1	94.1 ± 3.8^{6}	89.6 ± 4.7^{1}	94.6 ± 4.3^{6}	90.6 ± 5.0^{1}	39.4

Lesson: baseline your new method against a simpler alternative

Evaluation is humbling (2)

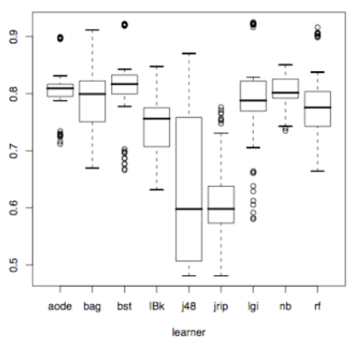
- 90 data miners
  - 9 learners with
  - 10 pre-processors
- 20 datasets
- (Win – Loss) results when one miner is compared to 89 others.
- Sum of five different performance measures
- And most miners perform about the same



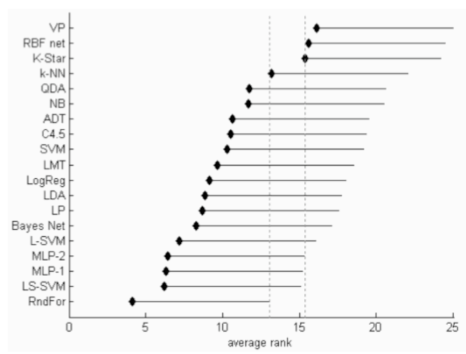
Lesson: beware "ceiling effects"

Evaluation is humbling (3)

- Left:
  - Y. Jiang, B. Cukic, and T. Menzies. Fault prediction using early lifecycle data. In ISSRE'07, 2007. Available from http://menzies.us/pdf/07issre.pdf.
- Right:
  - Lessmann, S., Baesens, B., Mues, C., Pietsch, S.: Benchmarking classification models for software defect prediction: a proposed framework and novel findings. IEEE Trans. Softw. Eng. (2008)



6/9 methods are "best"



14/19 methods are "best"

Lesson: most "improvements", aren't

No consensus on the "best" evaluation

	1999	2000	2001	2002	2003
Total number of papers	54	152	80	87	118
Relevant papers for our study	19	45	25	31	54
Sampling method [%]					
cross validation, leave-one-out	22	49	44	42	56
random resampling	11	29	44	32	54
separate subset	5	11	0	13	9
Score function [%]					
classification accuracy	74	67	84	84	70
classification accuracy - *exclusively*	68	60	80	58	67
recall, precision...	21	18	16	25	19
ROC, AUC	0	4	4	13	9
deviations, confidence intervals	32	42	48	42	19
Overall comparison of classifiers [%]	53	44	44	26	45
averages over the data sets	0	4	6	0	10
t-test to compare two algorithms	16	11	4	6	7
pairwise t-test one vs. others	5	11	16	3	7
pairwise t-test each vs. each	16	13	4	6	4
counts of wins/ties/losses	5	4	0	6	9
counts of *significant* wins/ties/losses	16	4	8	16	6

An overview of the papers accepted to International Conference on Machine Learning in years 1999—2003. The reported percentages (the third line and below) apply to the number of papers relevant for our study.

Janez Demsar: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7: 1-30 (2006)

- No global standard
- Advice:
  1. Study evaluation methods in current state-of-the-art papers
     - Copy them
  2. Avoid t-tests and their simplistic Gaussian assumptions
  3. Don't bother with results that report a (say) 4% improvement
  4. Be prepared to change the evaluation to make the reviewers happy
  5. Favor informative visualizations,
     - Use statistical tests as sanity checks on the conclusions form the visualization

Visualizations need not be elaborate

```
                                                                     == PERCENTILES ===
Rank   Treatment 0%                      50%                  100%   10  30  50  70  90
====   ========= ==                      ===                  ====   ==  ==  ==  ==  ==
   1   (M 3 K 3) |                        |               ---   *|   81  88  94 100 100
   1   (M 3 K 2) |                        |            ------   *|   76  88  94 100 100
   1   (M 3 K 1) |                        |            ---      *|   76  82  94 100 100
   1   (M 3 K 0) |                        |               ---   *|   81  88  94 100 100
   1   (M 2 K 3) |                        |                     *|   81  82  94 100 100
   1   (M 2 K 2) |                        |            ------   *|   76  88  94 100 100
   1   (M 2 K 1) |                        |            ---      *|   76  82  94 100 100
   1   (M 1 K 3) |                        |            ------   *|   76  88  94 100 100
   1   (M 1 K 2) |                        |            ------   *|   76  88  94 100 100
   1   (M 1 K 1) |                        |            ----     *|   76  85  94 100 100
   1   (M 1 K 0) |                        |            ------   *|   76  88  94 100 100
   1   (M 2 K 0) |                        |            ----  *   |   76  85  88 100 100
   2   (M 0 K 0) |                    --- |      *               |   41  49  65 100 100
   3   (M 0 K 3) |                 -------|   *                  |   35  50  59 100 100
   4   (M 0 K 2) |                  ------|   *                  |   38  50  59 100 100
   5   (M 0 K 1) |                 ------ |   *                  |   35  47  59 100 100
```

M,K: two magic params inside a NaiveBayes classifier handling low frequency counts
PD measurements in a 10*3 cross-val on IRIS
Rank set by a Mann-Whintey (95%(comparing each row to proceeding rows of the same rank

tim@menzies.us

WHAT HAVE WE LEARNED?

Fatal flaws in data mining for SE?

- Barbara Kitchenham et al, ESE journal, 2008
  - Replications can replicate stupid errors
- Vic Basili, LASER, 2010
  - If we give people our data, they can make stupid mistakes, cause they don't understand our context
- Well get back to this….

Data mining = a diverse and lucrative career

- Effort estimation
- Defect prediction
- Optimization of discrete systems
- Test case generation
- Fault localization
- Text mining
- Temporal sequence mining
  - Learning software processes
  - Learning APIs
- Etc

Data mining applications explored by me since 2007.

A career in data mining is a very diverse career, indeed

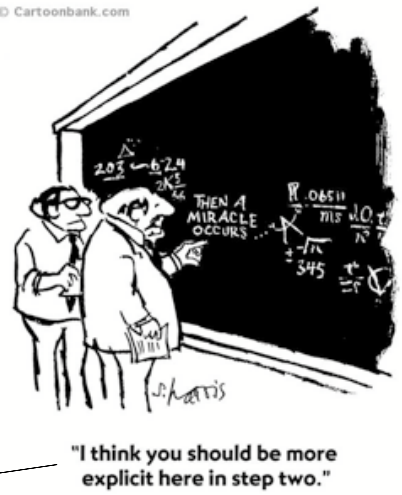
We need help

- A little experiment from http://www.youtube.com/v/vJG698U2Mvo&hl=en_US&fs=1&rel=0
- Rules
  - No one talks for the next 4 minutes
  - If you know what is about to happen, see (1)
- This is a selective attention test
  - Count the number of times the team with the white shirt passes the ball.

Data analysis deserves (much) more than zero pages

Easterbrook et al. (2007)

- 9 pages: selecting methods
- 3 pages: research questions
- 2 pages: empirical validity
- 2 pages: different forms of "empirical truth"
- 1 page: role of theory building
- 1 page: conclusions
- 1 page: data collection techniques
- 0 pages: data analysis
  - and then a miracle happens

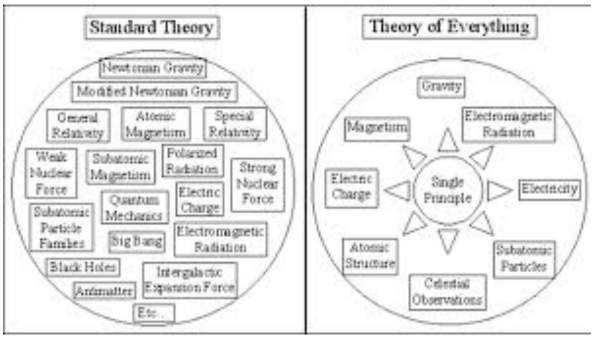


Don't just do data mining

- Be of the empirical research community
  - Go to LASER, ICSE, etc
  - Talk
- Find current hypothesis that of interest
  - E.g. max AUC(effort,pdf)
  - E.g. cross-vs-within data
  - E.g. TOE
  - E.g. text mining for structured reviews
    - Juristo, Menzies, 2011

Don’t do data mining once

- Continuous process monitoring
  1. Learn expectations
  2. Stale smell policy: when good ideas go bad
  3. Repair policies: how to modify old ideas (more mining)
  4. Escalation policy: recognize when you need to call for help
- Bt the way,
  - 1,2,3,4 can all be implemented by data miners.
- Welcome to TOE



Other Do-s and Don’ts

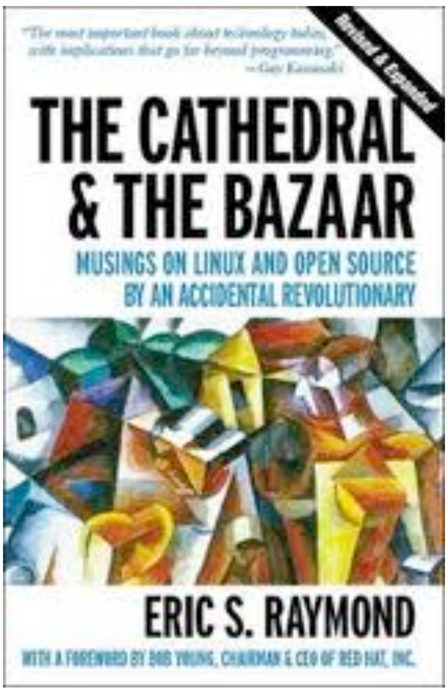
- Do learn about data mining
  - People make mistakes
  - Need communities of agents (human and otherwise)
  - New algorithms, old data, new insights
- Don’t used dumb data mining:
  - correlation, PCA??
  - Forgettaboutit
- Don’t quote old dumb studies:
  - E.g. Mccabe
- Do study stability:
  - 20 * 66% of the data
- Do model bias
  - Bias is where the business meets the learning

Exploit the crowd source advantage

- Crowd source
  - Join the community of people studying the data
  - Be there for them
- Lead, follow, or get out of the way
  - What's fair got to do with it? Its going to happen
  - Wolfgang Grieskamp from Microsoft, at Dagstuhl 2010

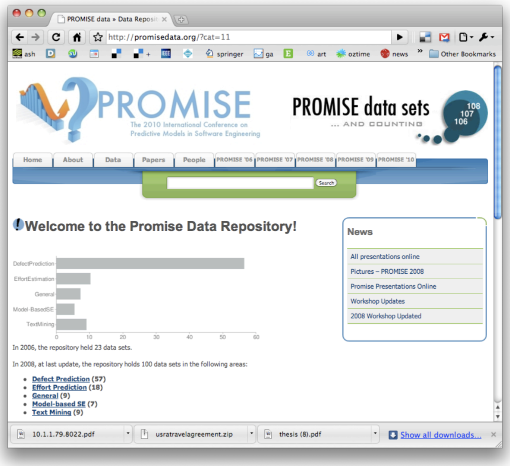
Open data initiatives

- Open source?
  That'll never work
- Menzies = bazaar!
- Are you the high priest in a cathedral?



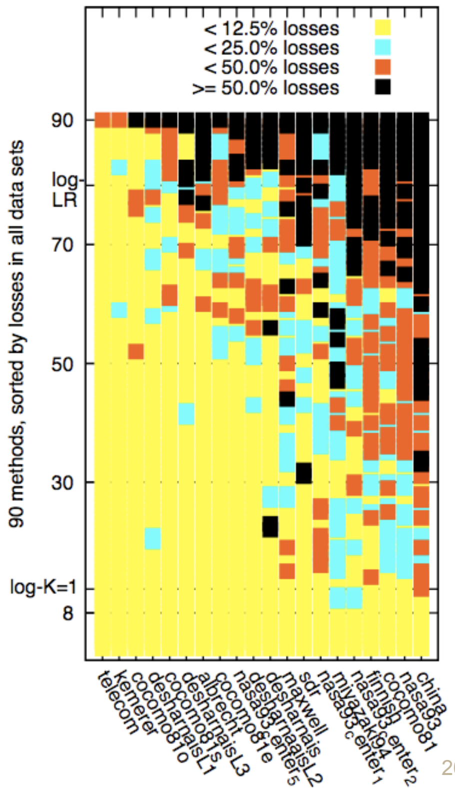
Lighten up!

- Do put data on the web
- Do collect data with "sunset clauses" (when it can go public)
  - The COCOMO experience
- Do collect data that joins
  - performance indicators
  - with things you can change
- Much inaccessible empirical data:
  - Data from the 152/ 154 MSR papers
  - ISERN 2007, ISERN 2008, ISERN 200
  - COCOMO-II
  - SEL
  - CeBase.org
- No propriety software
  - Static pages (no code that needs maintaining)
  - Password free sites (after the sunset)



Generate better results, faster

- Empirical SE results greatly lag the pace of innovation in the field.
- In too many cases
  - A trusted body of empirical results….
  - …Only appears after the innovative is already well on their way to obsolescence or standard practice.
- The generality of a result from any one case study is highly questioned. We urgently need:
  - Faster ways to learn local lessons
  - Faster ways to study data from multiple sources
- Can't always afford N people*Y years
  - Managers need answers yesterday
  - Funding bodies want progress
- Every time someone says "it depends"…
  - A grad student dies.



If a tree falls in a forest….

- Pooh and Piglet were walking together in the Thousand Acre Wood.
- The wind was blowing ferociously and the treetops were swaying.
- Somewhat disconcerted, Piglet asked Pooh, "What if a tree falls on us?"
- Pooh considered for a moment, before replying "What if it doesn't?"
- Barbara Kitchenham et al, ESEj, 2008
  - Replications can replicate stupid errors
  - Me: and sometimes, they don't
- Vic Basili, LASER, 2010
  - If we give people our data, they can make stupid mistakes, cause they don't understand the context
  - Me: and sometimes, they won't

Dude! Chill out!

By the way….

I am happy to report that there is no book called "data mining for dummies"