01236057 Data Mining

เหมืองข้อมูล

วัชระ ฉัตรวิริยะ

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

watchara.ch@kmitl.ac.th

คำอธิบายรายวิชา

- แนวคิดและเทคนิคต่างๆ ในการทำเหมืองข้อมูล ประสิทธิภาพและข้อดีข้อเสียของอัลกอริทึมต่างๆ ที่ใช้ในการทำเหมืองข้อมูล กระบวนการเตรียมข้อมูล การหารูปแบบที่เกิดขึ้นบ่อย การวิเคราะห์ความสัมพันธ์ การหากฎความสัมพันธ์ การจำแนกประเภท การทำนาย การวิเคราะห์กลุ่ม โครงข่ายประสาทเทียม การทำเหมืองข้อมูลกับข้อมูลพิเศษ ข้อมูลสื่อผสม ข้อมูลเครือข่ายสังคม ข้อมูลเชิงพื้นที่ การประยุกต์ใช้และทิศทางของการทำเหมืองข้อมูล

- Data mining concepts and techniques; efficiency, pros and cons of data mining algorithms; data preprocessing; frequent pattern mining; association rules; classification; prediction; cluster analysis; neural network; mining special data: multimedia data, social network data, and spatial data; data mining applications and trends.

เรียนทำไม

- เอาไปใช้อะไรได้
- งานตำแหน่งอะไรใช้วิชานี้
- ตัวอย่างการฝึกอบรมภายนอก

รูปแบบการเรียน

- บรรยายแนวคิดสำคัญ และคณิตศาสตร์พื้นฐาน
- บรรยายการใช้ซอฟต์แวร์ WEKA
- บรรยายการใช้ซอฟต์แวร์ Matlab
- บรรยายการวิเคราะห์ข้อมูล
- แนะนำการเขียนรายงาน และการนำเสนอ
- ฝึกใช้ WEKA สำหรับการวิเคราะห์ข้อมูล
- ฝึกใช้และการเขียนโปรแกรม Matlab สำหรับการวิเคราะห์ข้อมูล
- ฝึกเขียนโปรแกรม Matlab และการทำงานร่วมกับ MySQL

ผลลัพท์การเรียน

- เข้าใจแนวคิดสำคัญ
- เข้าใจหลักการคณิตศาสตร์ที่ใช้
- สามารถใช้ซอฟต์แวร์ WEKA
- สามารถวิเคราะห์ข้อมูลพื้นฐาน
- สามารถเขียนรายงานและนำเสนอ
- สามารถใช้และเขียนโปรแกรม Matlab สำหรับการวิเคราะห์ข้อมูล
- สามารถใช้และเขียนโปรแกรม Matlab และ MySQL สำหรับการวิเคราะห์ข้อมูล

การตัดเกรด

- จำนวนนักศึกษา xx คน
- A :
- B+:
- B :
- C+:
- C :
- D+:
- D :

Vocabulary?

- Model Analysis
- Statistical Analysis
- Data Analytic

- Machine Learning
- Big Data
- IoT

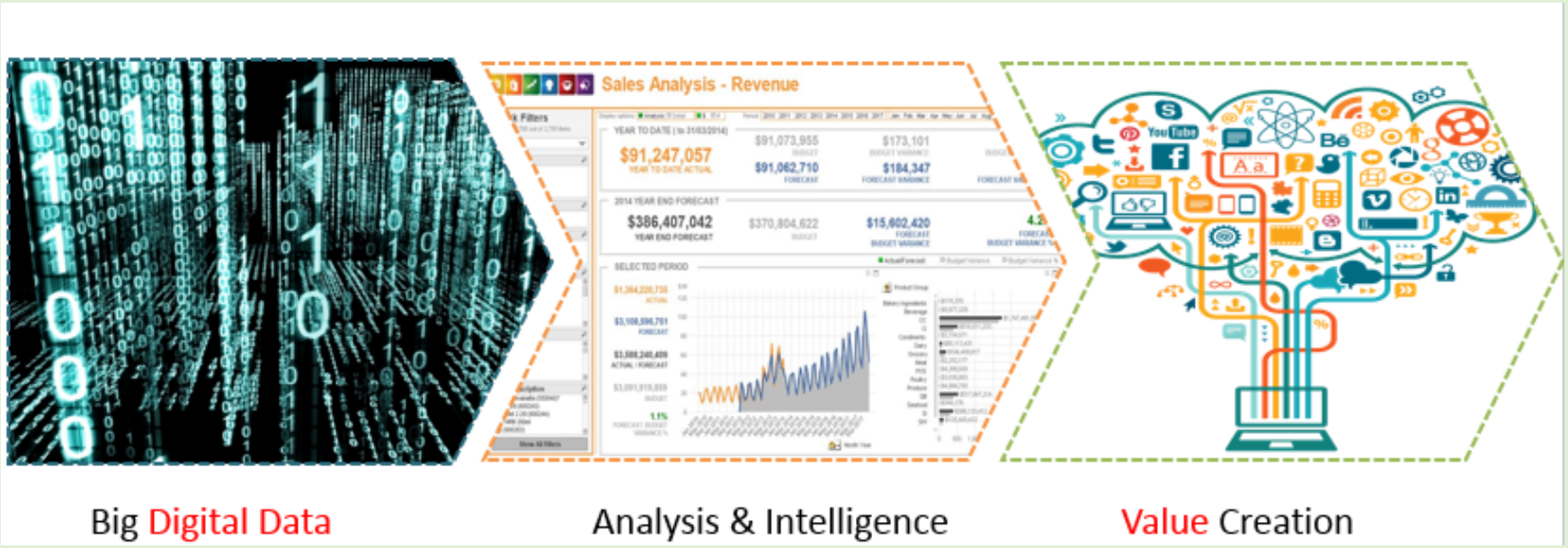
- Artificial Intelligent
- Rule-based
- Soft Computing

- Digital Economy
- Industry 4.0

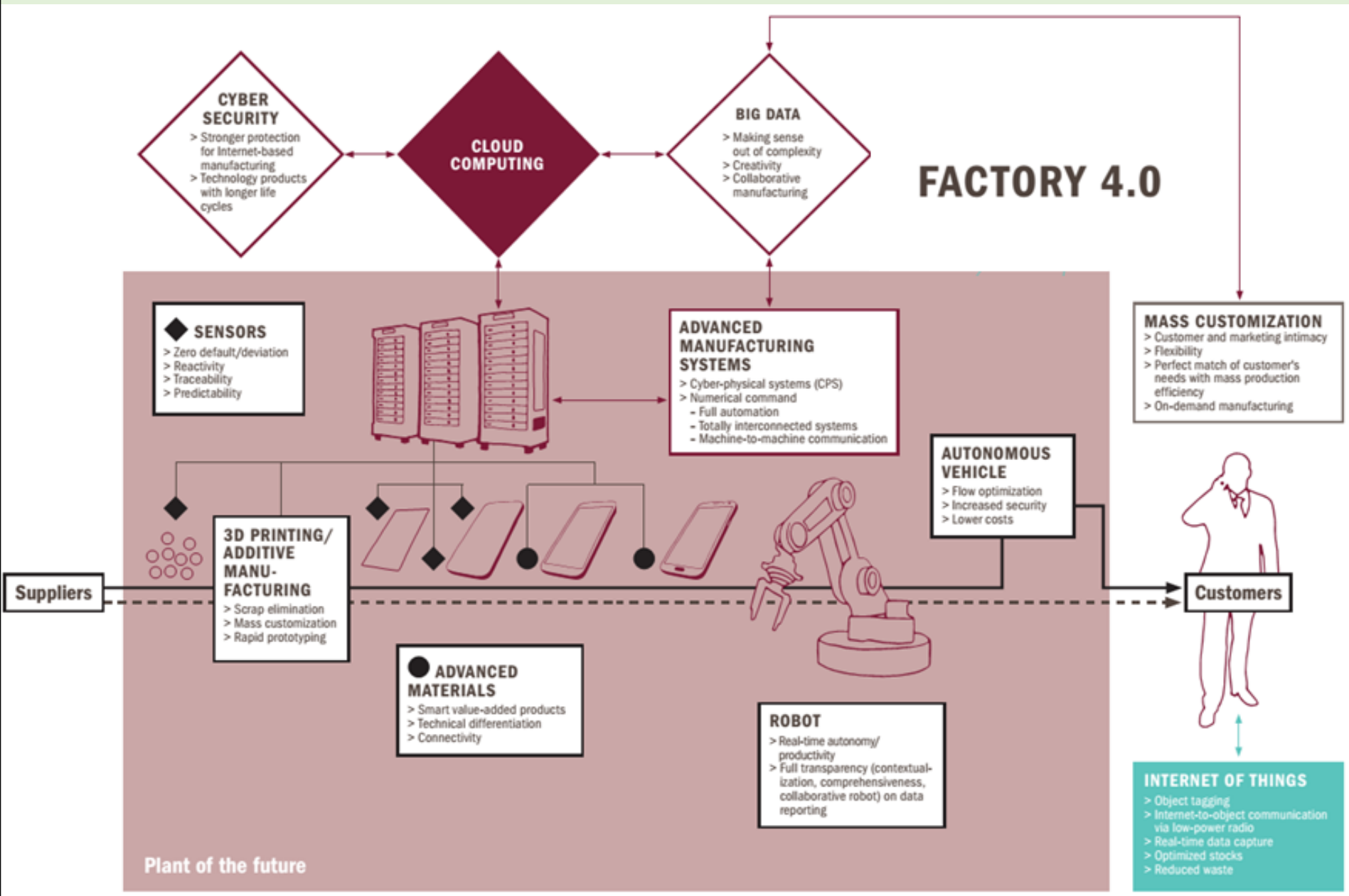
- Cloud Computing

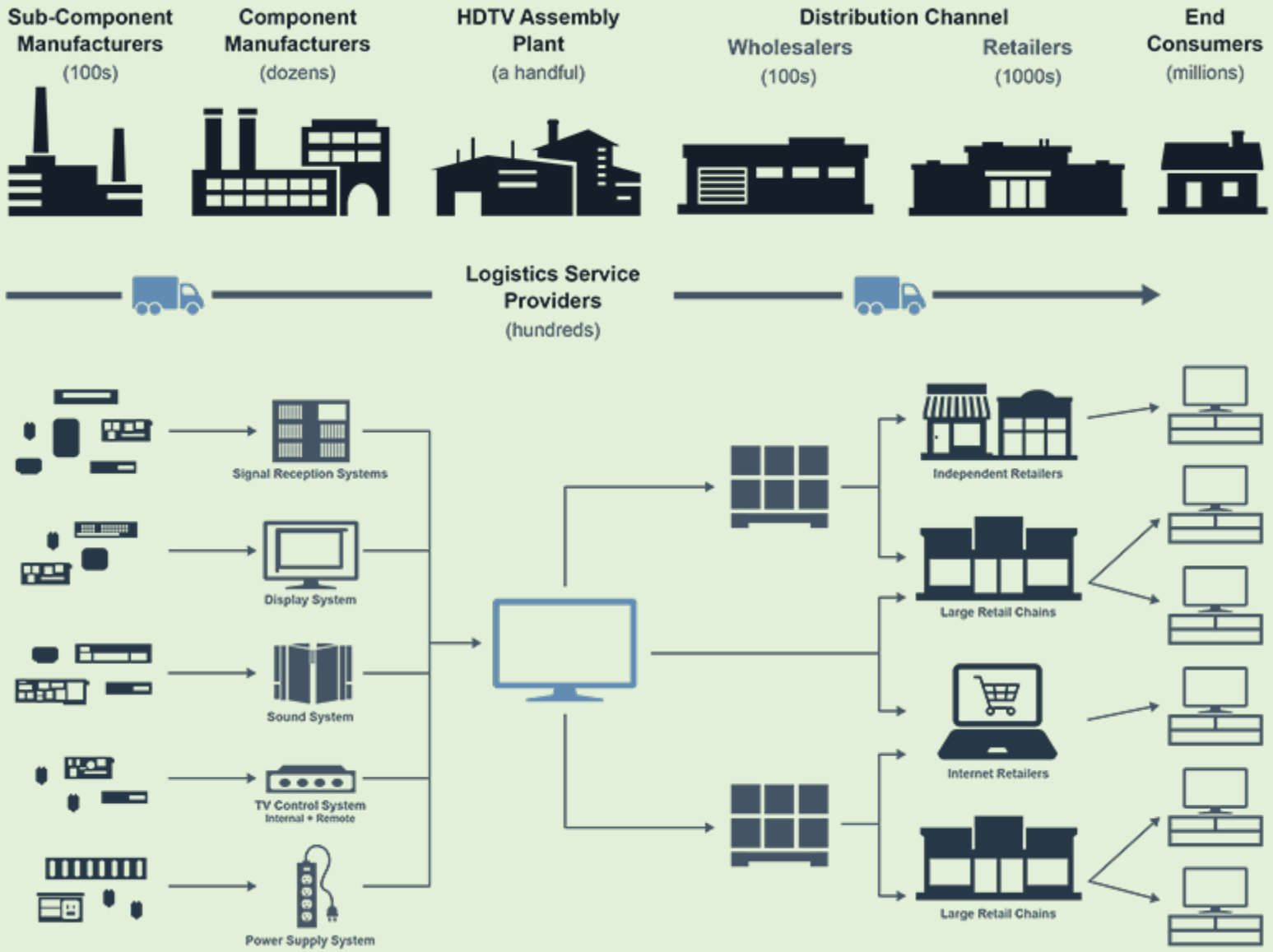
เศรษฐกิจดิจิทัล (Digital Economy)

Driven by Data

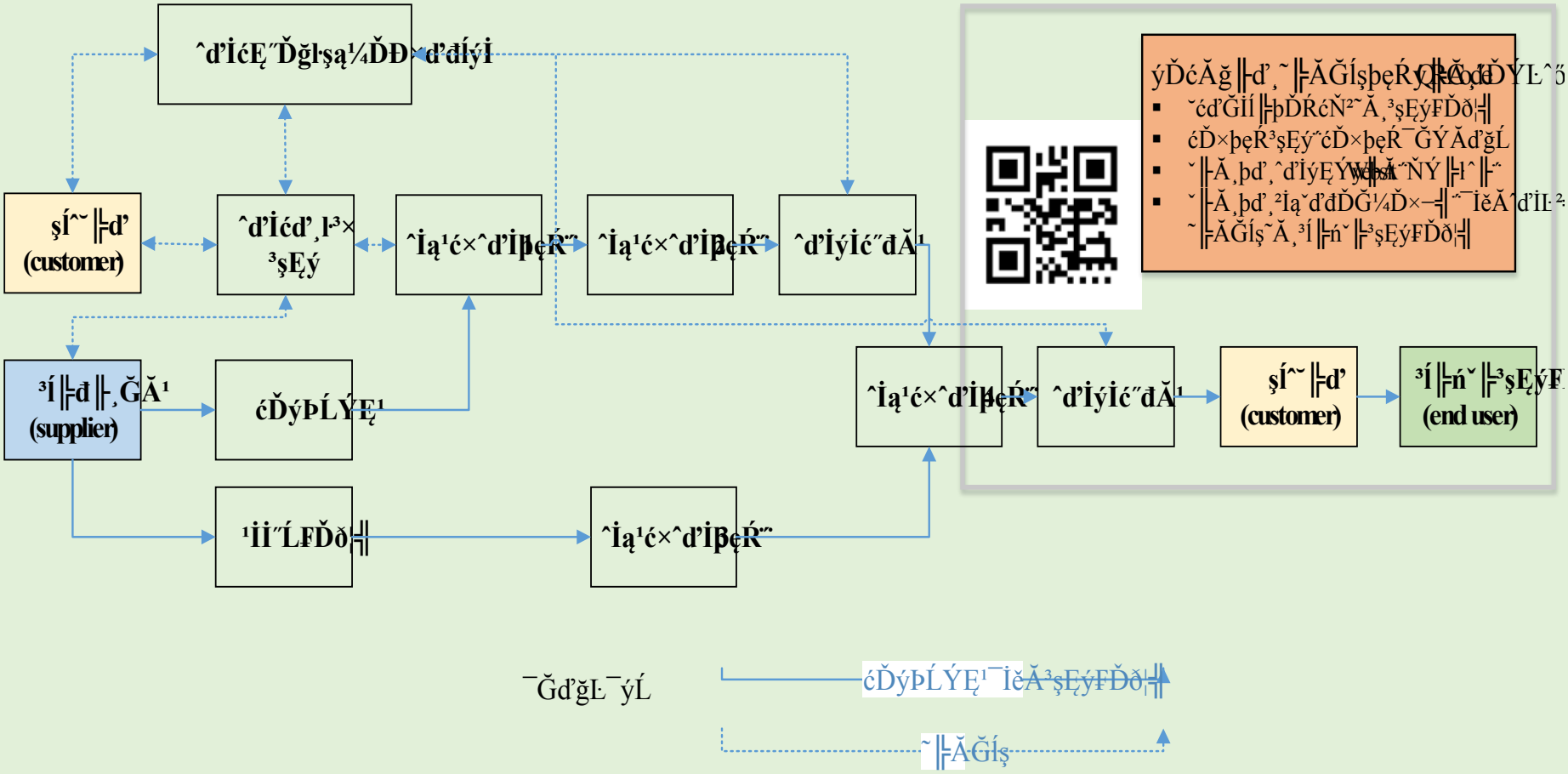


Driven by Value Creation

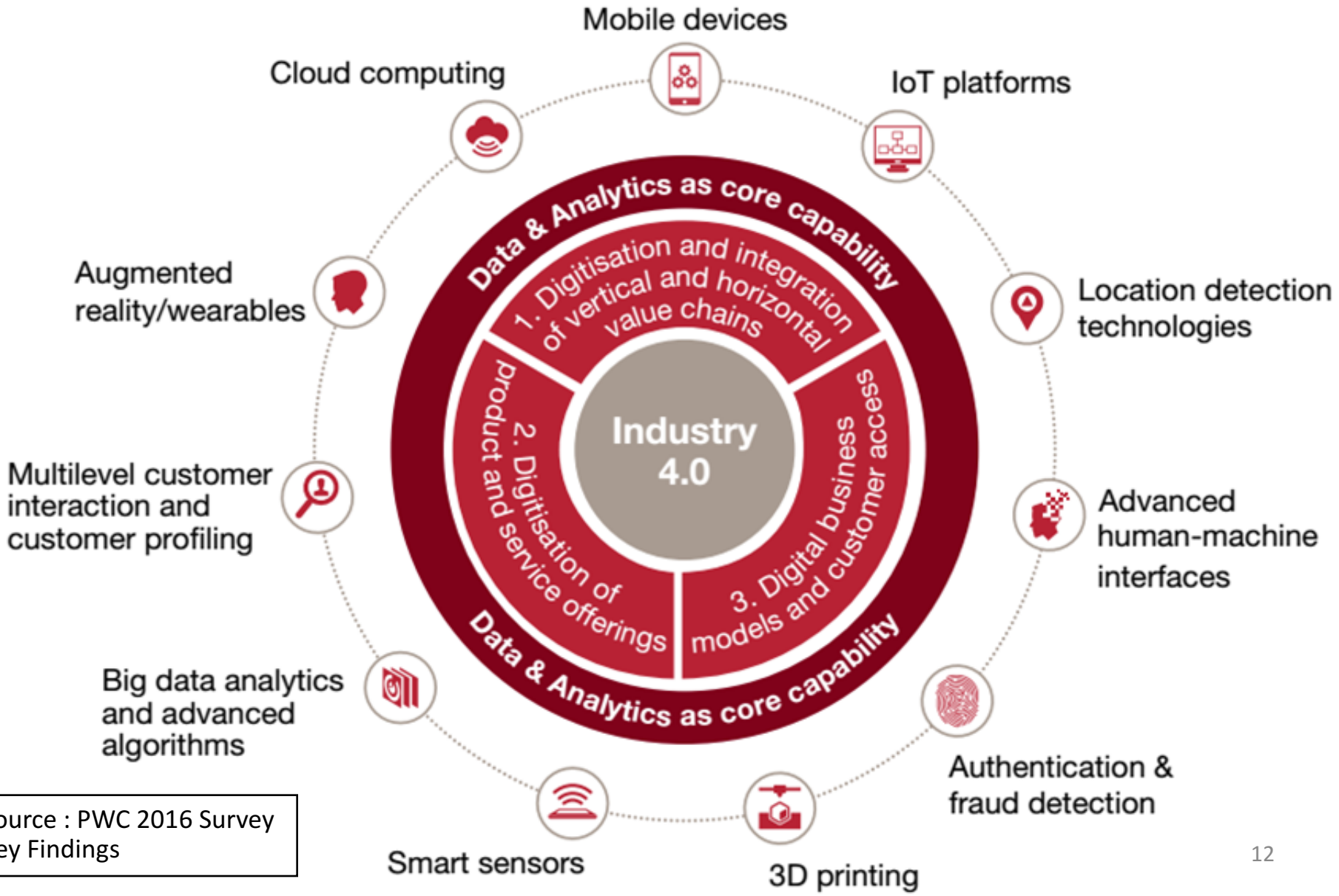

CYBER SECURITY
> Stronger protection for Internet-based manufacturing
> Technology products with longer life cycles
CLOUD COMPUTING
BIG DATA
> Making sense out of complexity
> Creativity
> Collaborative manufacturing
FACTORY 4.0
SENSORS
> Zero default/deviation
> Reactivity
> Traceability
> Predictability
ADVANCED MANUFACTURING SYSTEMS
> Cyber-physical systems (CPS)
> Numerical command
- Full automation
- Totally interconnected systems
- Machine-to-machine communication
MASS CUSTOMIZATION
> Customer and marketing intimacy
> Flexibility
> Perfect match of customer's needs with mass production efficiency
> On-demand manufacturing
AUTONOMOUS VEHICLE
> Flow optimization
> Increased security
> Lower costs
Suppliers
3D PRINTING/ ADDITIVE MANU-FACTURING
> Scrap elimination
> Mass customization
> Rapid prototyping
Customers
ADVANCED MATERIALS
> Smart value-added products
> Technical differentiation
> Connectivity
ROBOT
> Real-time autonomy/ productivity
> Full transparency (contextual-ization, comprehensiveness, collaborative robot) on data reporting
INTERNET OF THINGS
> Object tagging
> Internet-to-object communication via low-power radio
> Real-time data capture
> Optimized stocks
> Reduced waste
Plant of the future


Sub-Component Manufacturers
(100s)
Component Manufacturers
(dozens)
HDTV Assembly Plant
(a handful)
Distribution Channel
Wholesalers
(100s)
Retailers
(1000s)
End Consumers
(millions)
Logistics Service Providers
(hundreds)
Signal Reception Systems
Display System
Sound System
TV Control System
Internal + Remote
Power Supply System
Independent Retailers
Large Retail Chains
Internet Retailers
Large Retail Chains

การสร้างช่องทางการสื่อสารกับผู้ใช้ผลิตภัณฑ์



Industry 4.0 Framework



Source : PWC 2016 Survey Key Findings

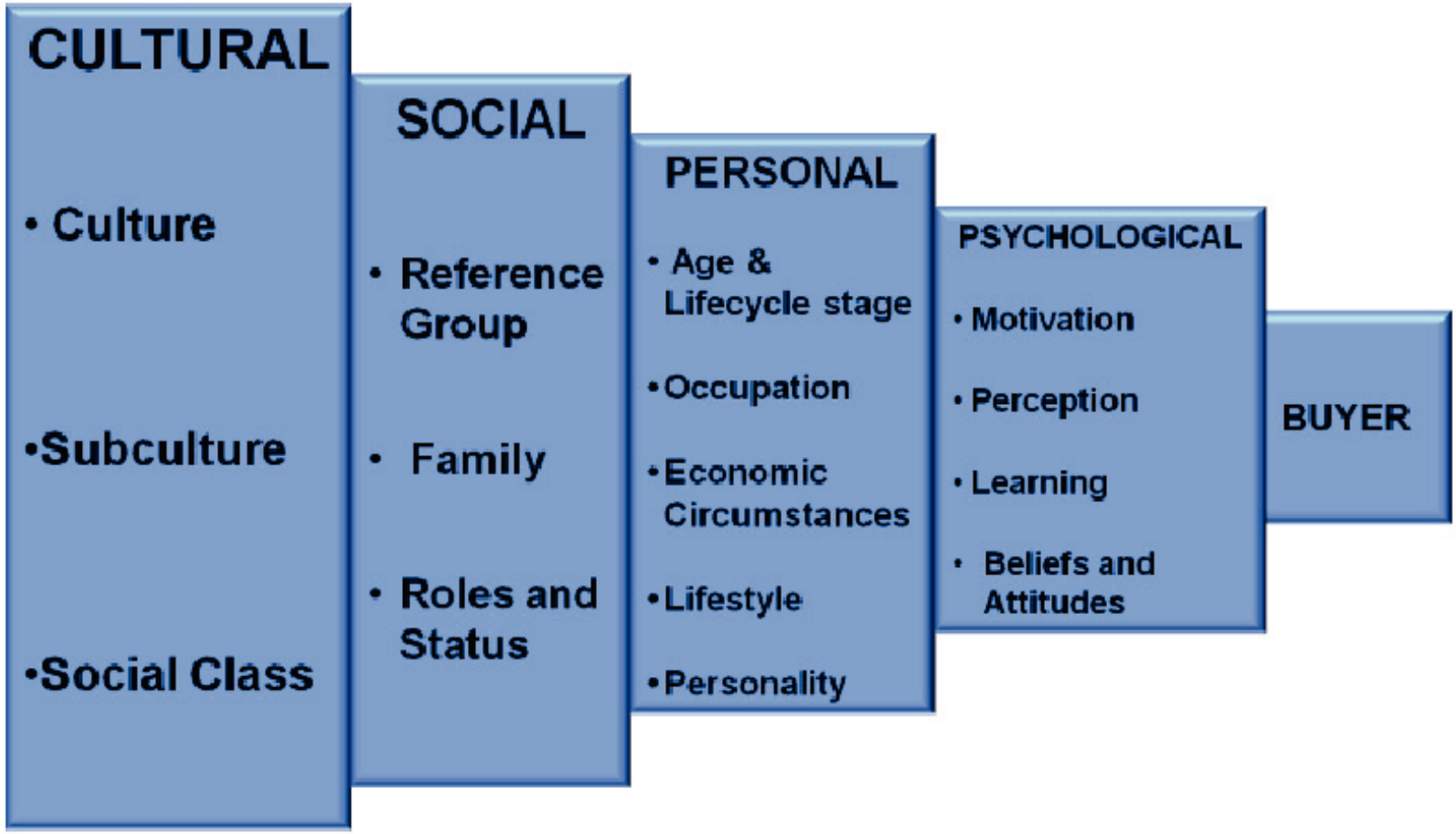
Questions?

What can be expected from Data Analytic?

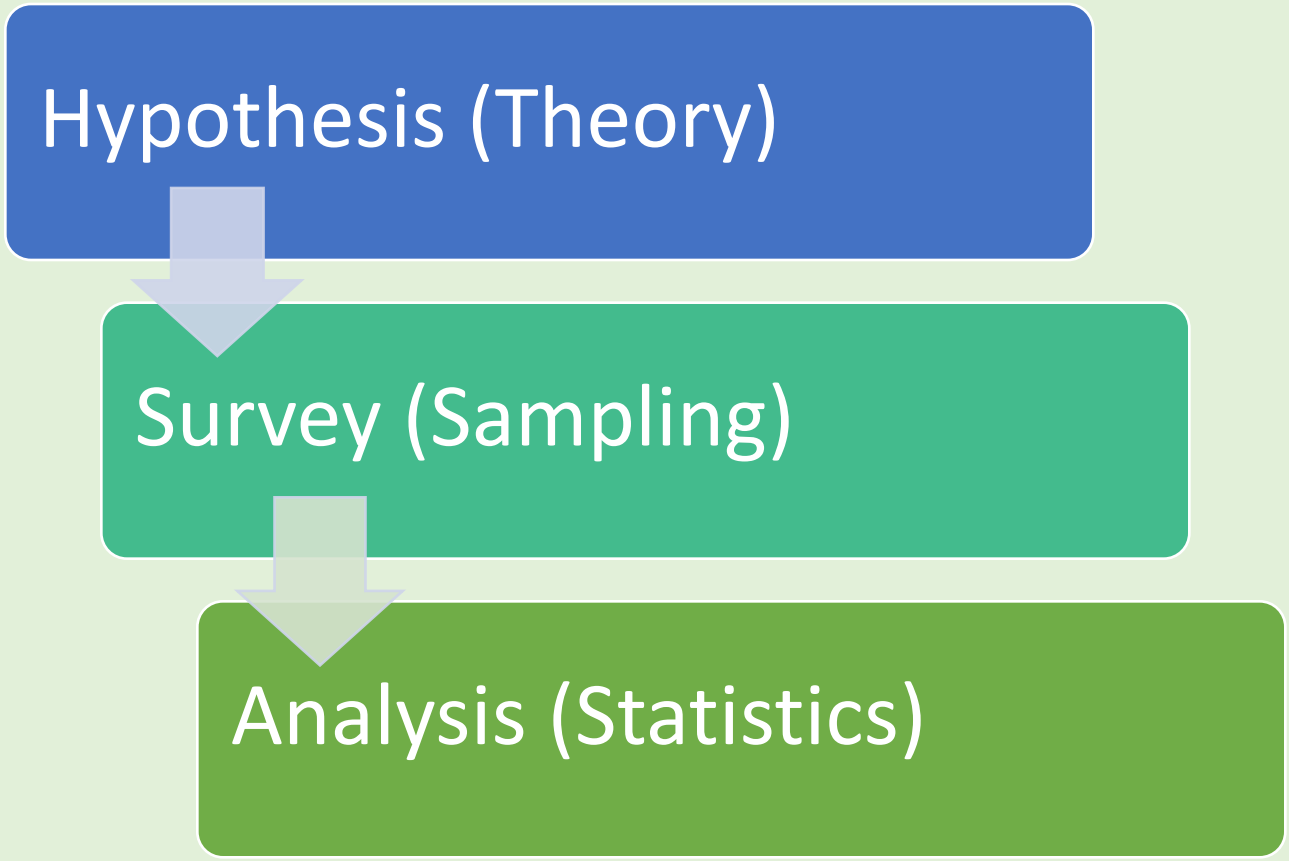
Methodology behind.

Tools and systems

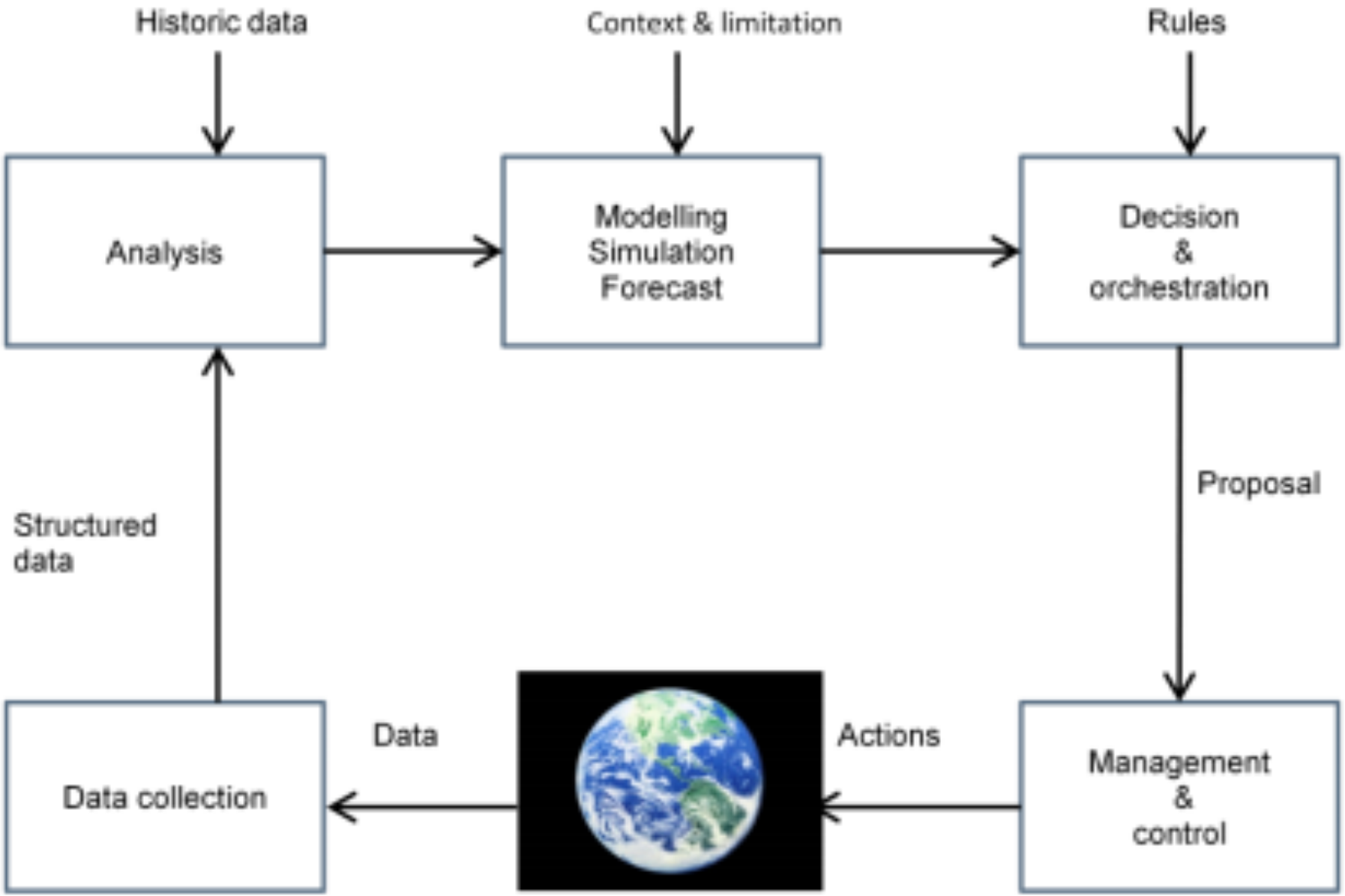
Customer Behavior Theory



Customer Behavior Analysis



Big data value chain



Source: See IBM, 2013

Data Science

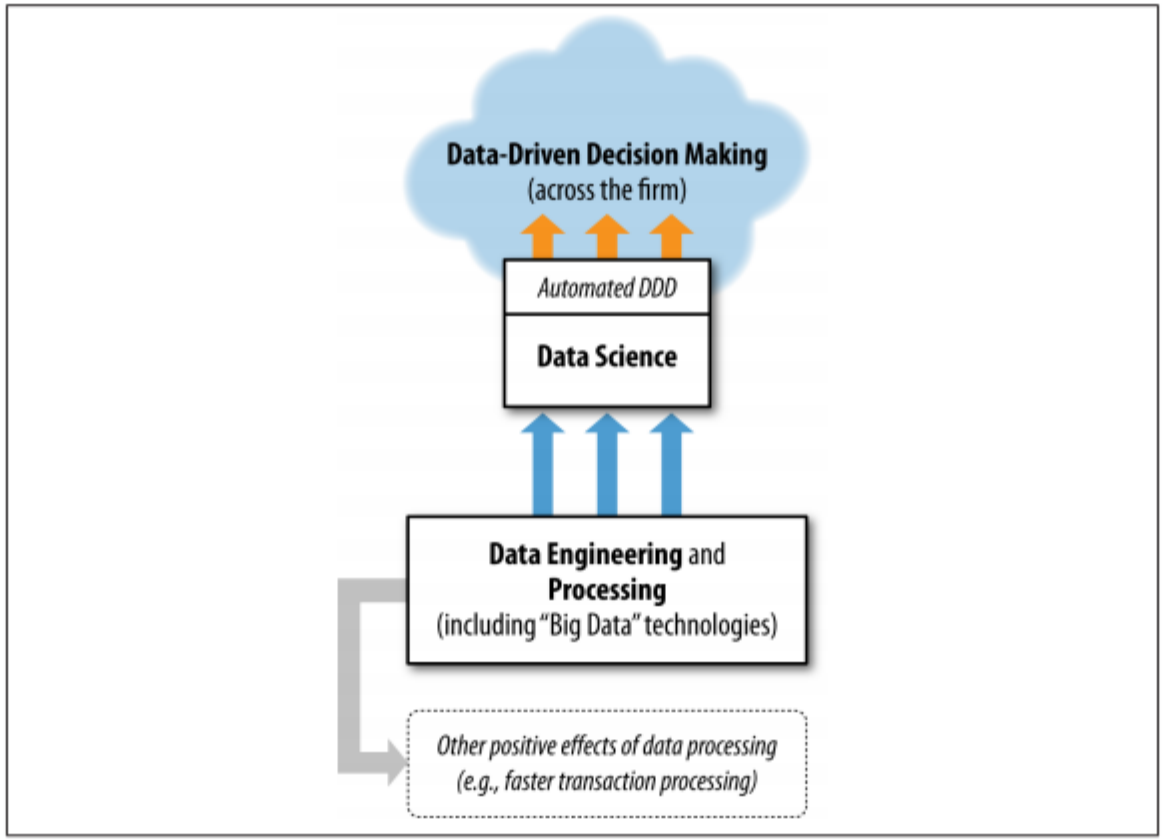
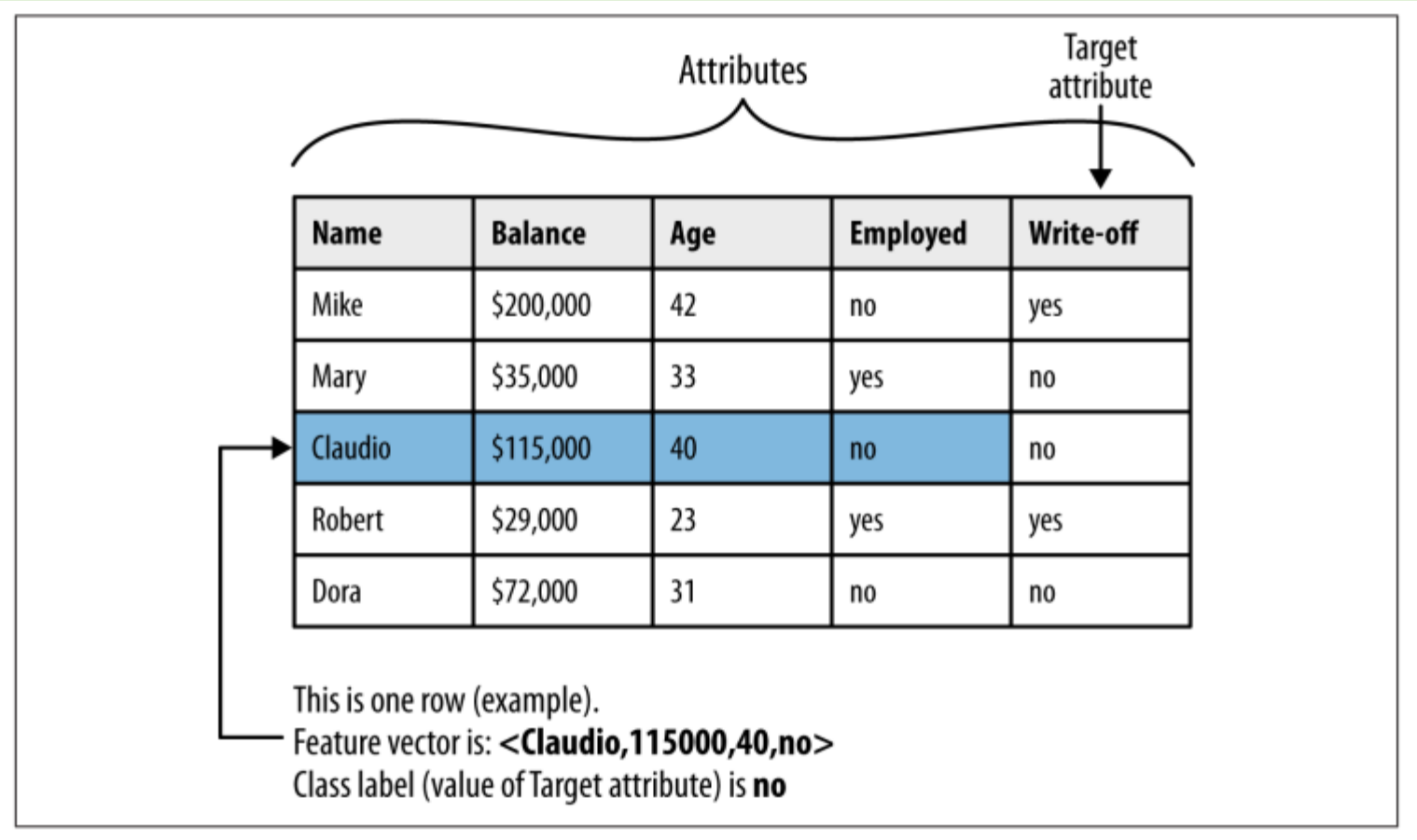


Figure 1-1. Data science in the context of various data-related processes in the organization.

Foster Provost & Tom Fawcett, Data Science for Business, O'Reilly, 2013



Foster Provost & Tom Fawcett, Data Science for Business, O'Reilly, 2013

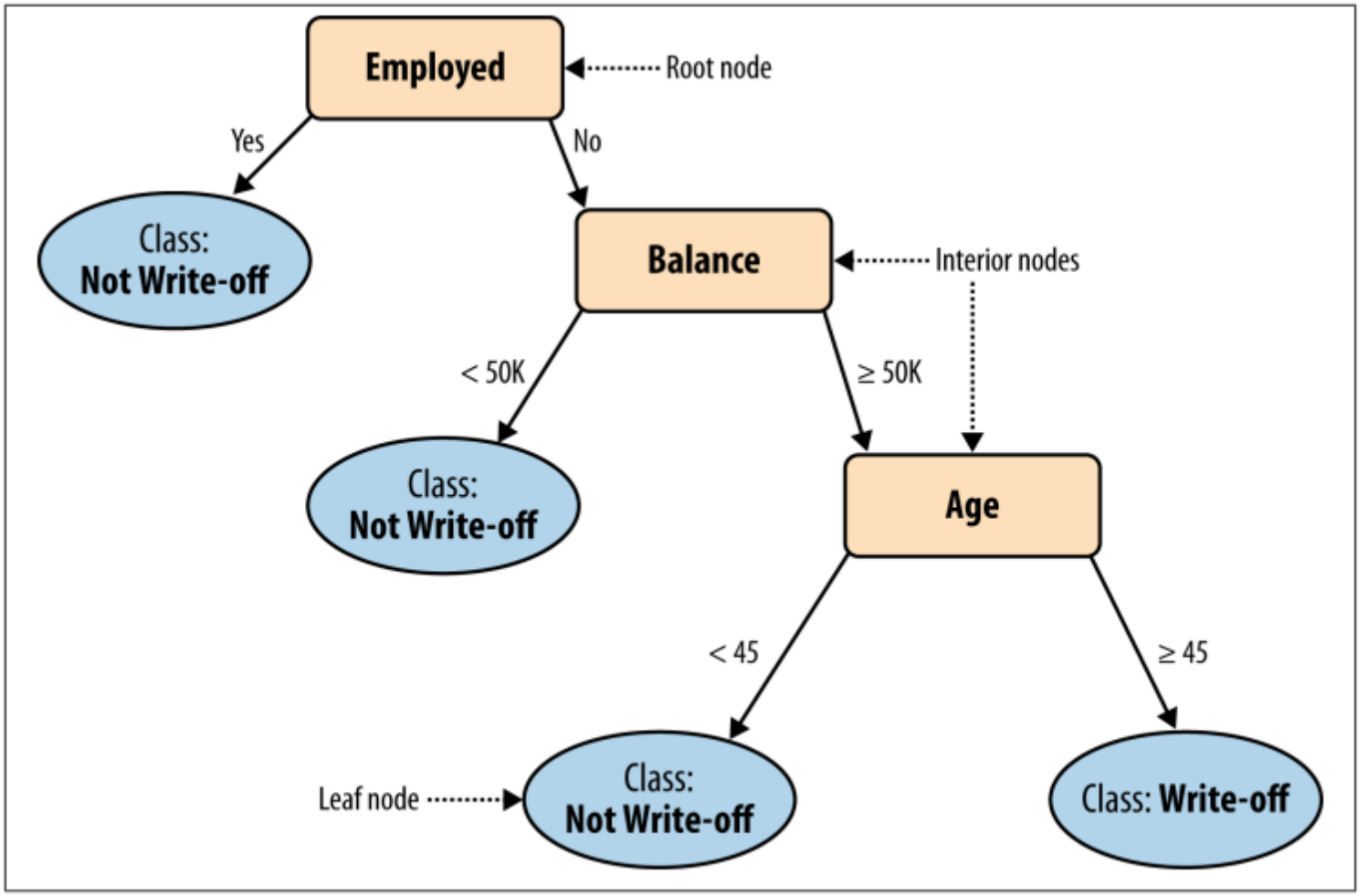
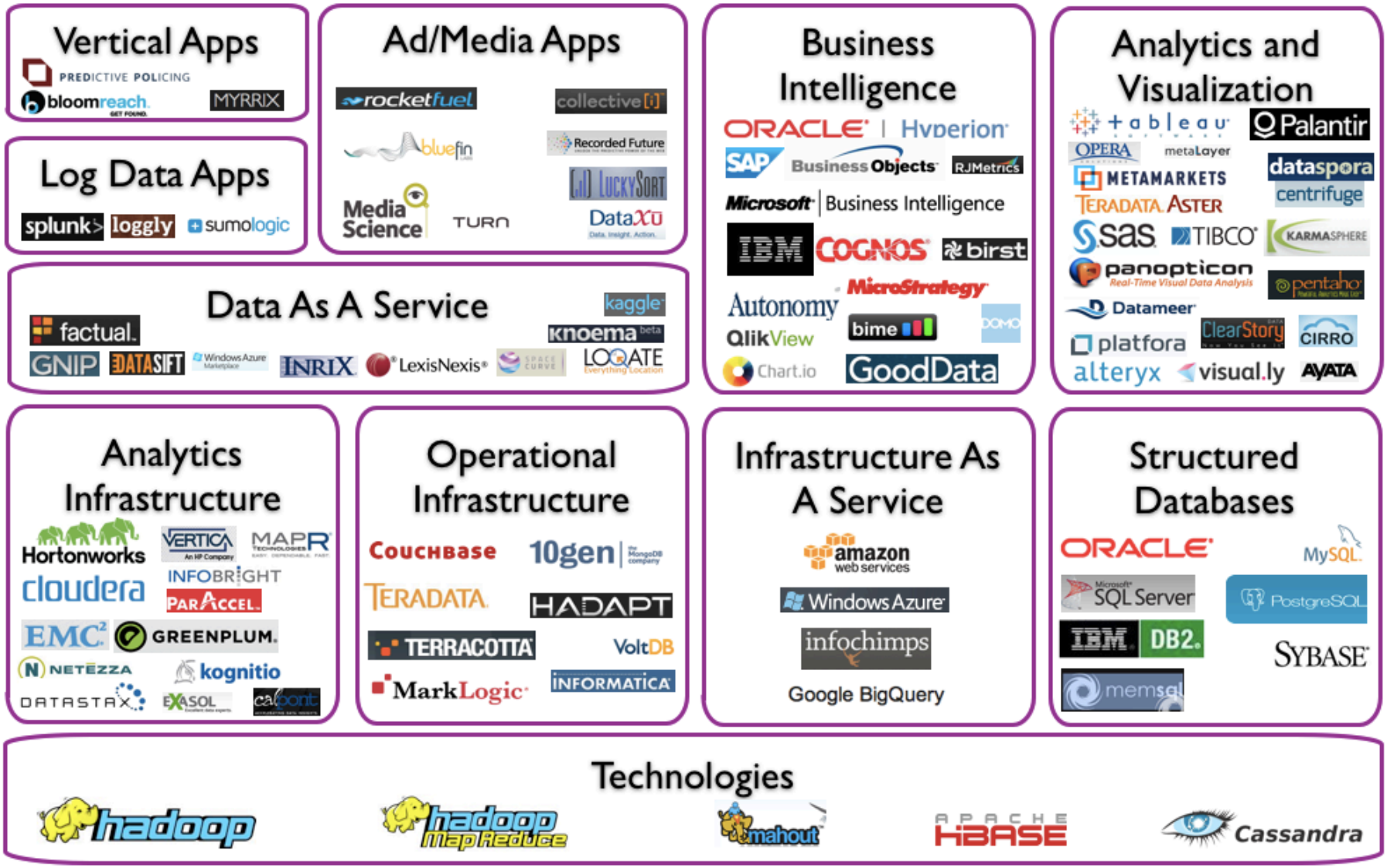


Figure 3-10. A simple classification tree.

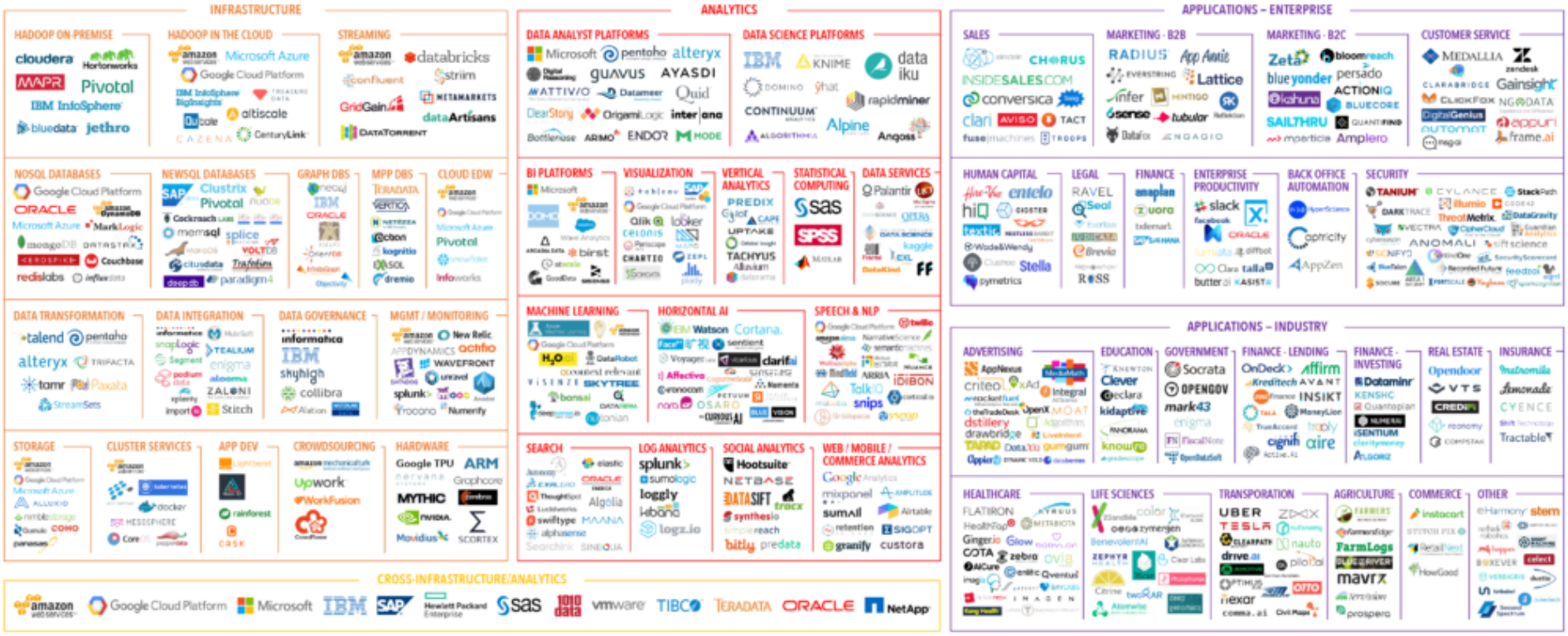
Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

BIG DATA LANDSCAPE 2017



Last updated 4/5/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data Analytic

Model Analysis

Statistical Analysis

Data Analytic (Machine Learning)

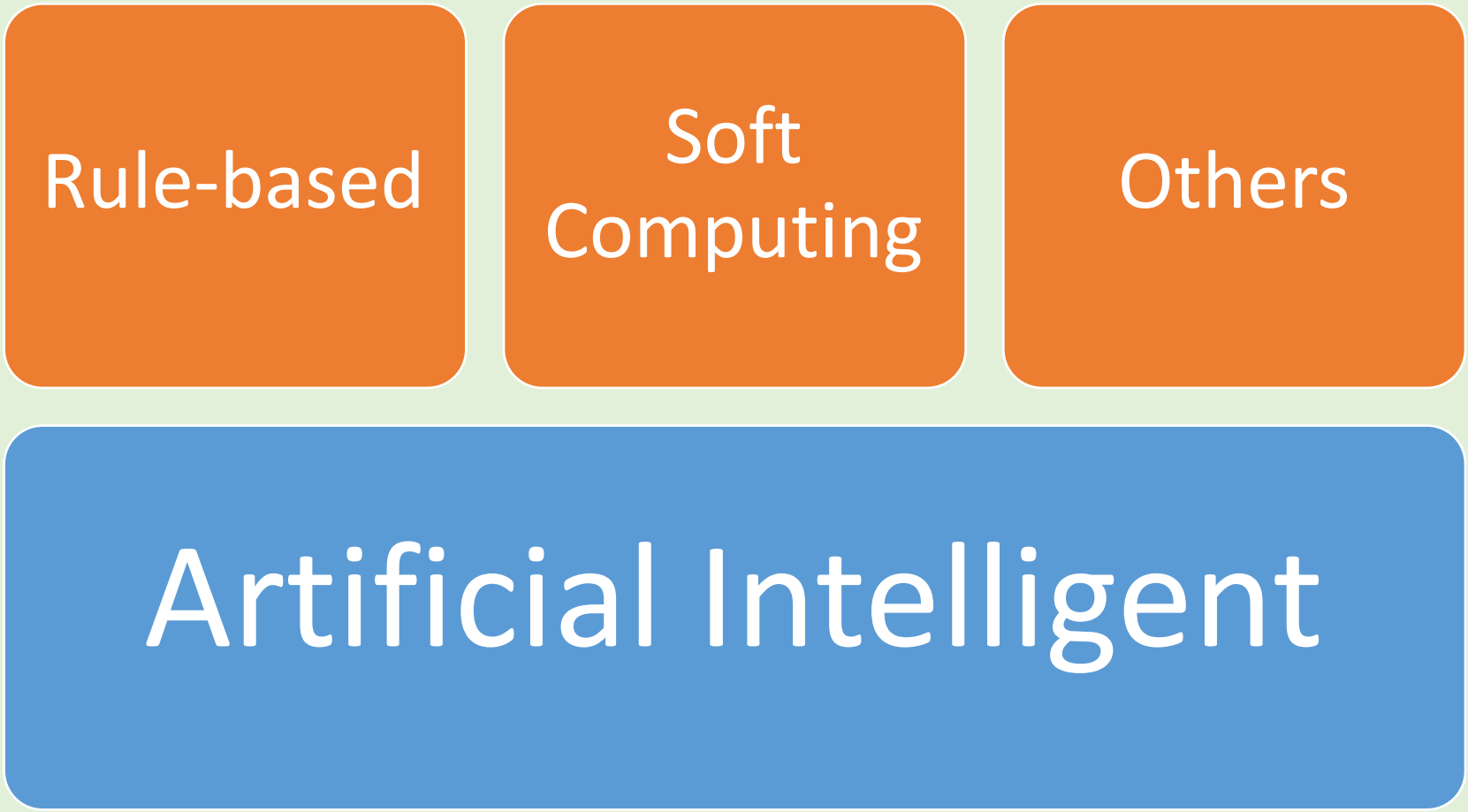
The initiated model is a general form of computation capable of the system behavior interaction.

Use large data set to derive system behavior model through learning algorithm

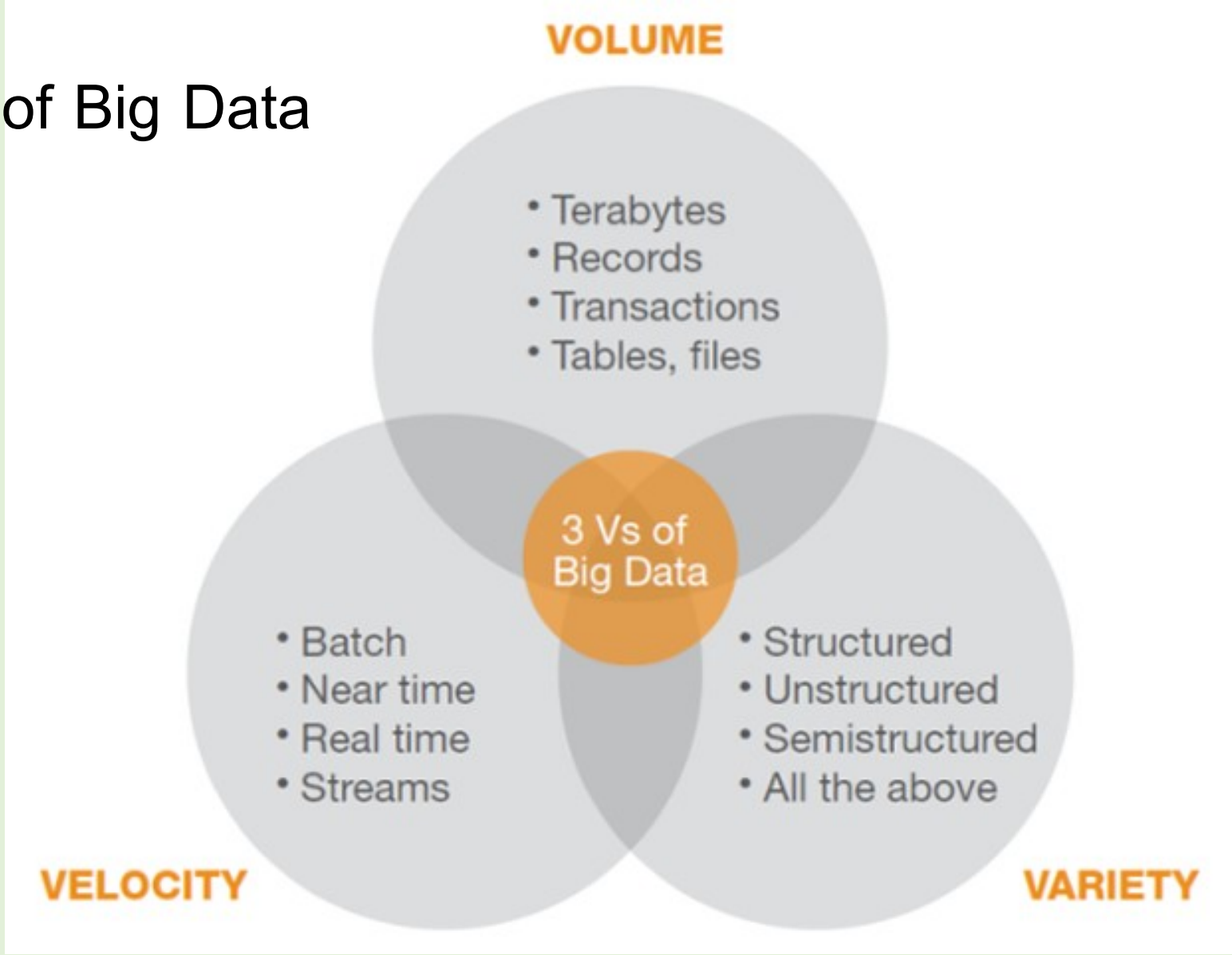
The new input is applied to the system model, and the trained model should response with similar output of the real system.

Intelligent

- Detection
- Identification
- Verification
- Classification
- Clustering
- Recognition
- Vision

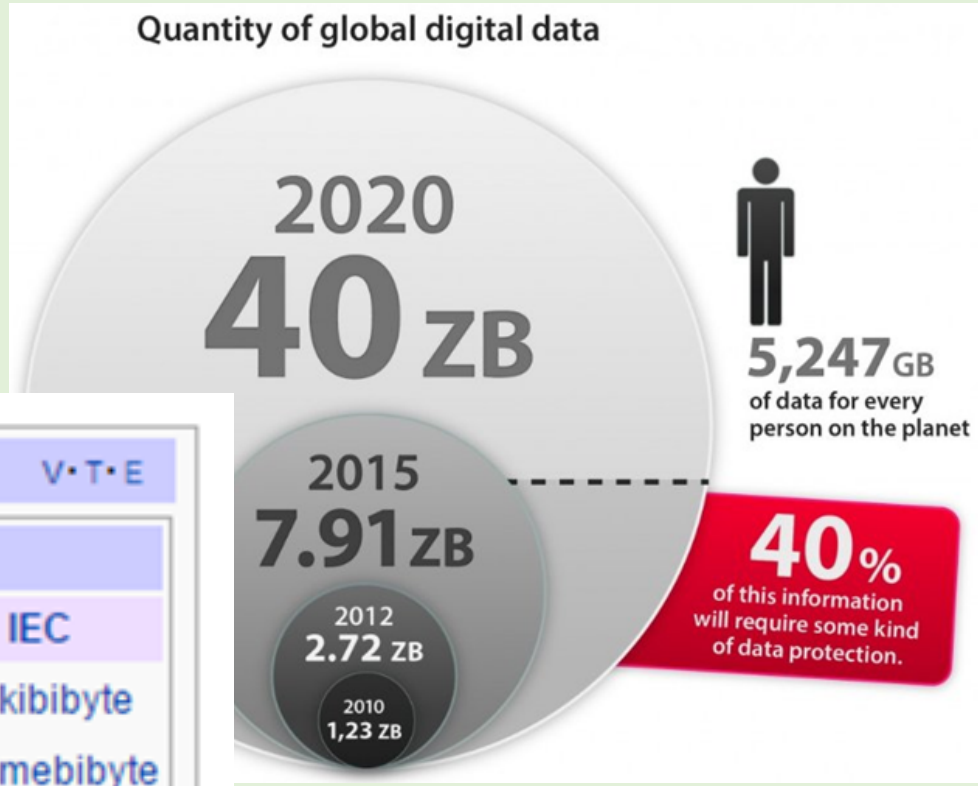

Rule-based
Soft Computing
Others
Artificial Intelligent

3 V of Big Data



https://www.linkedin.com/pulse/where-does-big-data-start-end-denis-sproten

How big is big?



Multiples of bytes V·T·E

Decimal			Binary				
Value	Metric		Value	JEDEC		IEC	
1000	kB	kilobyte	1024	KB	kilobyte	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MB	megabyte	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GB	gigabyte	GiB	gibibyte
1000^4	TB	terabyte	1024^4	–	–	TiB	tebibyte
1000^5	PB	petabyte	1024^5	–	–	PiB	pebibyte
1000^6	EB	exabyte	1024^6	–	–	EiB	exbibyte
1000^7	ZB	**zettabyte**	1024^7	–	–	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	–	–	YiB	yobibyte

Orders of magnitude of data

Data Mining Model

- Classification
- Clustering
- Association
- Times Series

Big Data examples

- Network data traffic logs
- Financial transactions
- Stock market feed
- CCTV VDO steam
- Manufacturing sensor feed
- Medical sensor feed
- Weather sensor feed

Network Traffic Data log

- 0.1 K byte generated every network activities
- various formats of traffic logs
- updated instantly (streaming)
- Need to analyze various network treats and make decision response in real time
- ~ 1000 MB per day for a 1000 users company
- 90 day traffic data log analysis - 90 GB

Health Data log

- 0.5 K (50x10x10) byte generated every second
- fixed formats of data
- updated instantly (streaming)
- Need to analyze various treats and make decision response in real time
- 40 MB (0.5x60x60x24) per day for one user

Data Mining Development Environments

Application

- Rapidminer
- Weka
- Orange
- KNIME

Programming

- R
- Python
- Matlab

Data Mining Architecture

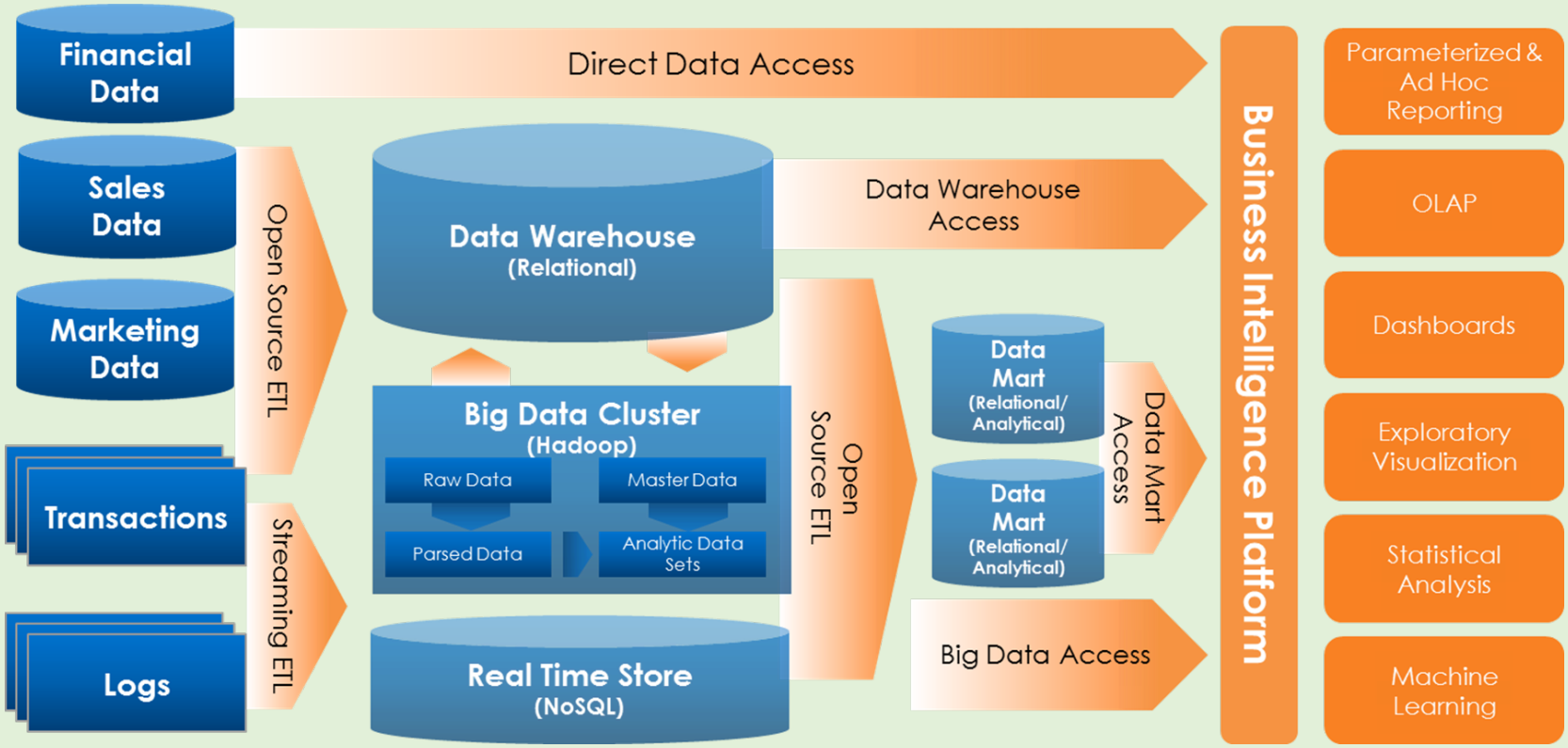
Hardware

- Cluster Computing Unit
- Large memory
- Large storage
- Interconnecting Network

Software

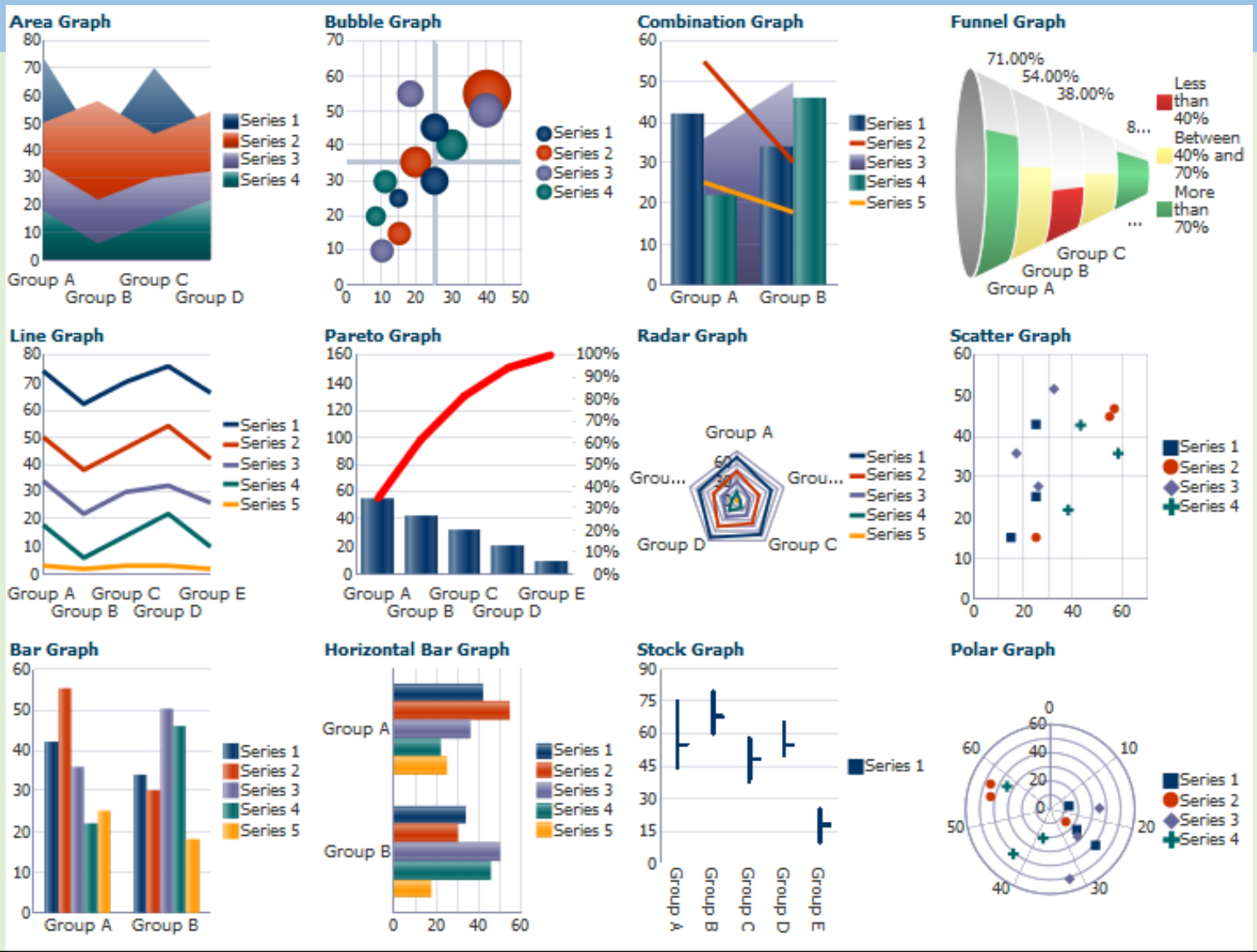
- Structured Database
- un-Structured Database
- Data warehouse
- Learning & Analysis
- Visualization

Big Data Intelligent Architecture



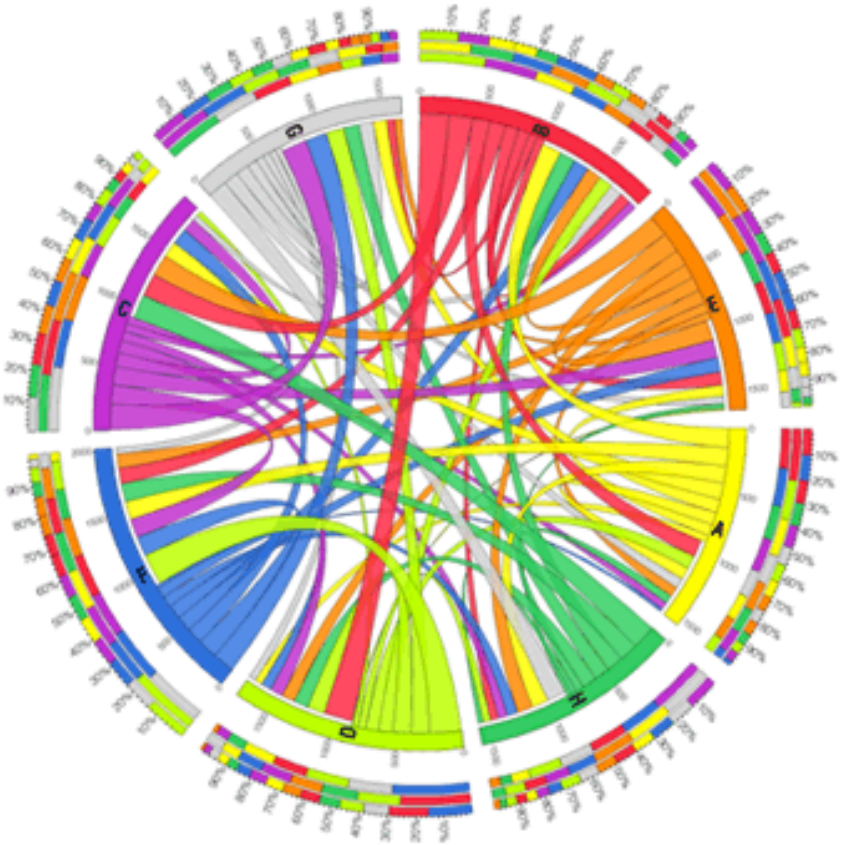
http://www.directbi.com/architecture.html

Data Visualization



Data Visualization

	A	B	C	D	E	F	G	H
A	54	133	157	94	88	141	167	133
B	49	113	111	113	202	53	7	92
C	66	130	69	162	123	62	106	117
D	60	138	49	85	98	98	122	87
E	53	88	15	91	91	20	69	127
F	118	32	62	139	135	95	60	64
G	114	108	73	44	103	139	37	145
H	74	110	84	120	9	41	45	131



Knowledge Discovering in Database (KDD) process

Selection

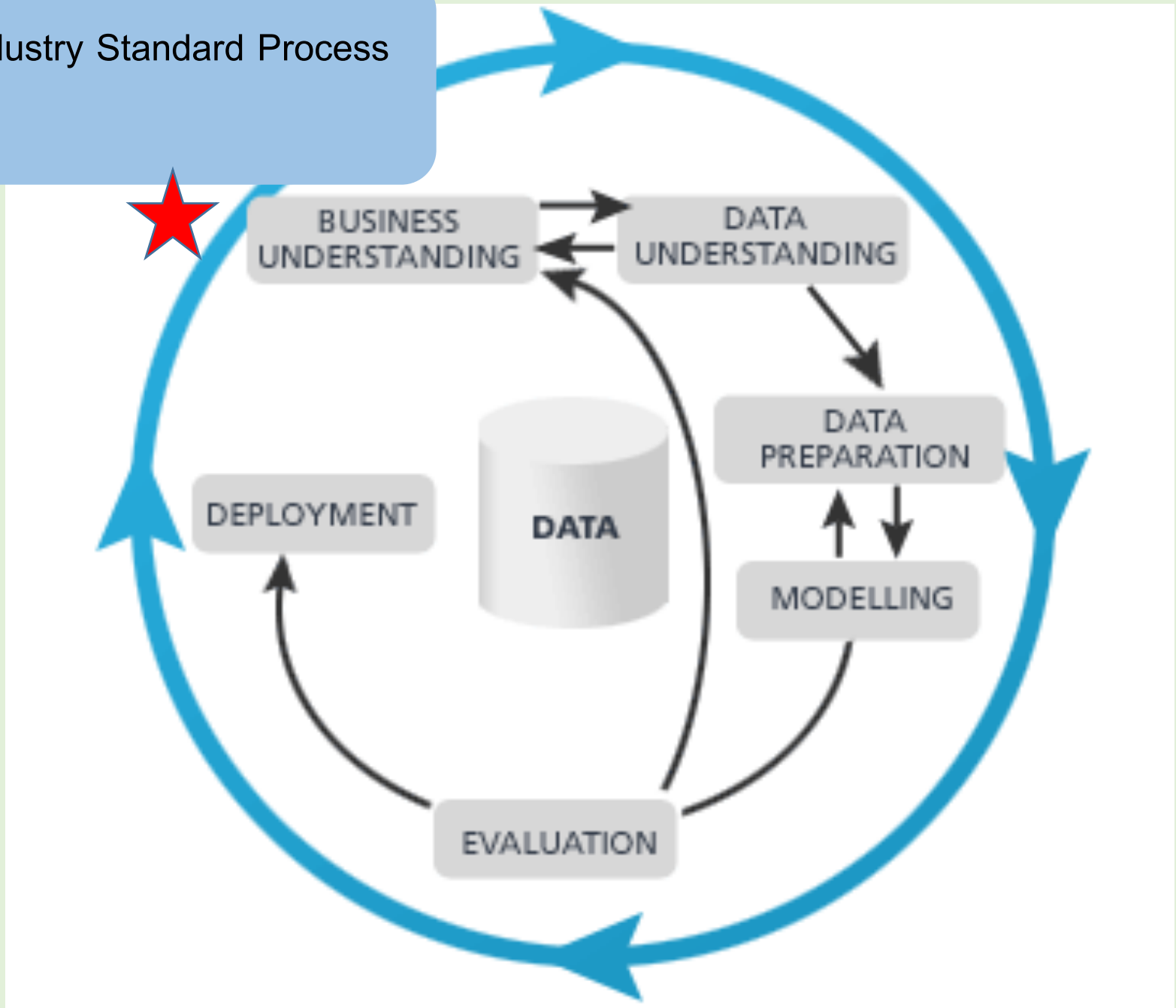
Pre-processing

Transformation

Data Mining

Interpretation/Evaluation

Cross Industry Standard Process (CRISP)



Data Mining Tasks

- Anomaly Detection (Outliner/change/deviation detection)
- Association rule learning (Dependency model)
- Clustering (Similarity)
- Classification (Unknown class samples)
- Regression (Function formulation)
- Summarization (Representation, Visualization)

What has been happening during the last six months with sales?

How do their gross sales figures compare to their target sales figures?

	September	October	November	December	January	February
Gross sales	$5,280,000	$5,501,000	$5,469,000	$5,480,000	$5,533,000	$5,554,000
Target sales	$5,280,000	$5,500,000	$5,729,000	$5,968,000	$6,217,000	$6,476,000
Ad costs	$1,056,000	$950,400	$739,200	$528,000	$316,800	$316,800
Social network costs	$0	$105,600	$316,800	$528,000	$739,200	$739,200
Unit prices (per oz.)	$2.00	$2.00	$2.00	$1.90	$1.90	$1.90

Do you see a pattern in Acme's expenses?

What do you think is going on with these unit prices? Why are they going down?

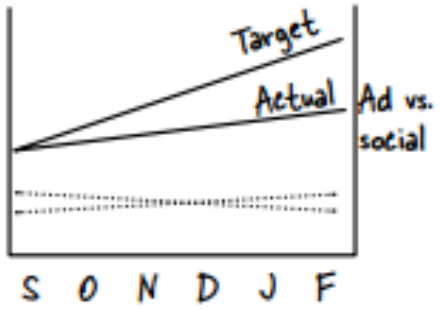
Acme Cosmetics Analytical Report

Context

MoisturePlus customers are tween girls (where tweens are people aged 11–15). They're basically the only customer group. Acme is trying out reallocating expenses from advertisements to social networking, but so far, the success of the initiative is unknown. We see no limit to potential sales growth among tween girls. Acme's competitors are extremely dangerous.

Interpretation of data

Sales are slightly up in February compared to September, but kind of flat. Sales are way off their targets. Cutting ad expenses may have hurt Acme's ability to keep pace with sales targets. Cutting the prices does not seem to have helped sales keep pace with targets.



Recommendation

It might be that the decline in sales relative to the target is linked to the decline in advertising relative to past advertising expenses. We have no good evidence to believe that social networking has been as successful as we had hoped. I will return advertising to September levels to see if the tween girls respond. **Advertising to tween girls is the way to get gross sales back in line with targets.**

This is the stuff we got from the CEO at the beginning.

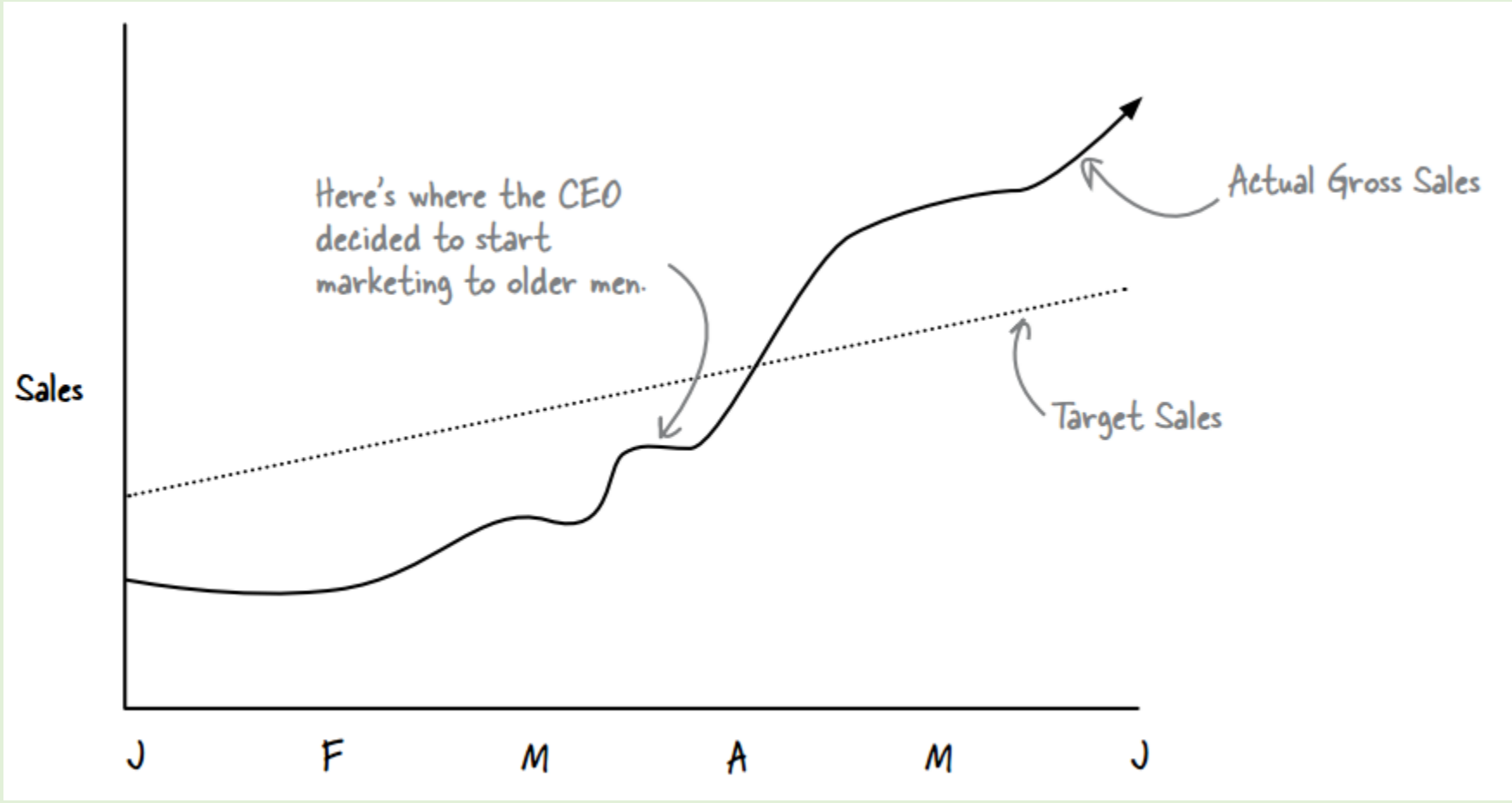
Here's the meat of your analysis.

Your conclusion might be different.

It's a good idea to state your and your clients' assumptions in your report.

A simple graphic to illustrate your conclusion.

Date	Vendor	Lot size (units)	Shipping ZIP	Cost
9/1/08	Sassy Girl Cosmetics	5253	20817	$75,643
9/3/08	Sassy Girl Cosmetics	6148	20817	$88,531
9/4/08	Prissy Princess	8931	20012	$128,606
9/14/08	Sassy Girl Cosmetics	2031	20817	$29,246
9/14/08	Prissy Princess	8029	20012	$115,618
9/15/08	General American Wholesalers	3754	20012	$54,058
9/20/08	Sassy Girl Cosmetics	7039	20817	$101,362
9/21/08	Prissy Princess	7478	20012	$107,683
9/25/08	General American Wholesalers	2646	20012	$38,102
9/26/08	Sassy Girl Cosmetics	6361	20817	$91,598
10/4/08	Prissy Princess	9481	20012	$136,526
10/7/08	General American Wholesalers	8598	20012	$123,811
10/9/08	Sassy Girl Cosmetics	6333	20817	$91,195
10/12/08	General American Wholesalers	4813	20012	$69,307
10/15/08	Prissy Princess	1550	20012	$22,320
10/20/08	Sassy Girl Cosmetics	3230	20817	$46,512
10/25/08	Sassy Girl Cosmetics	2064	20817	$29,722
10/27/08	General American Wholesalers	8298	20012	$119,491
10/28/08	Prissy Princess	8300	20012	$119,520
11/3/08	General American Wholesalers	6791	20012	$97,790
11/4/08	Prissy Princess	3775	20012	$54,360
11/10/08	Sassy Girl Cosmetics	8320	20817	$119,808
11/10/08	Sassy Girl Cosmetics	6160	20817	$88,704
11/10/08	General American Wholesalers	1894	20012	$27,274
11/15/08	Prissy Princess	1697	20012	$24,437
11/24/08	Prissy Princess	4825	20012	$69,480
11/28/08	Sassy Girl Cosmetics	6188	20817	$89,107
11/28/08	General American Wholesalers	4157	20012	$59,861
12/3/08	Sassy Girl Cosmetics	6841	20817	$98,510
12/4/08	Prissy Princess	7483	20012	$107,755
12/6/08	General American Wholesalers	1462	20012	$21,053


Here's where the CEO decided to start marketing to older men.
Actual Gross Sales
Target Sales
Sales
J
F
M
A
M
J

WEKA : Weather & Play

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Weka : Interesting Data Set

- Weather and Play
- Diabetes
- CPU
- Beast Cancer
- Glass
- Contact Lenses
- Credit
- Iris

Machine Learning Model

Supervised Learning (Classification)

Unsupervised Learning (Clustering)

Classification

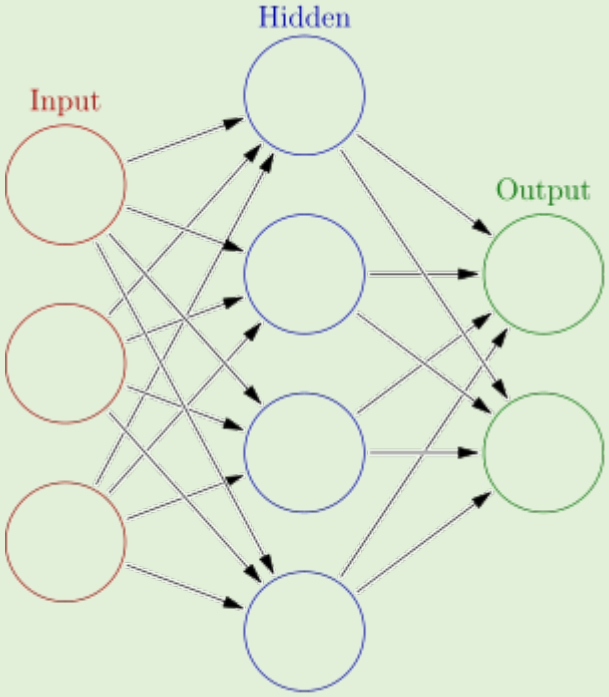
Pre processing (Data Preparation)

- Features Extraction
- Feature Selection
- Dataset Partitioning
- Classes Assignment

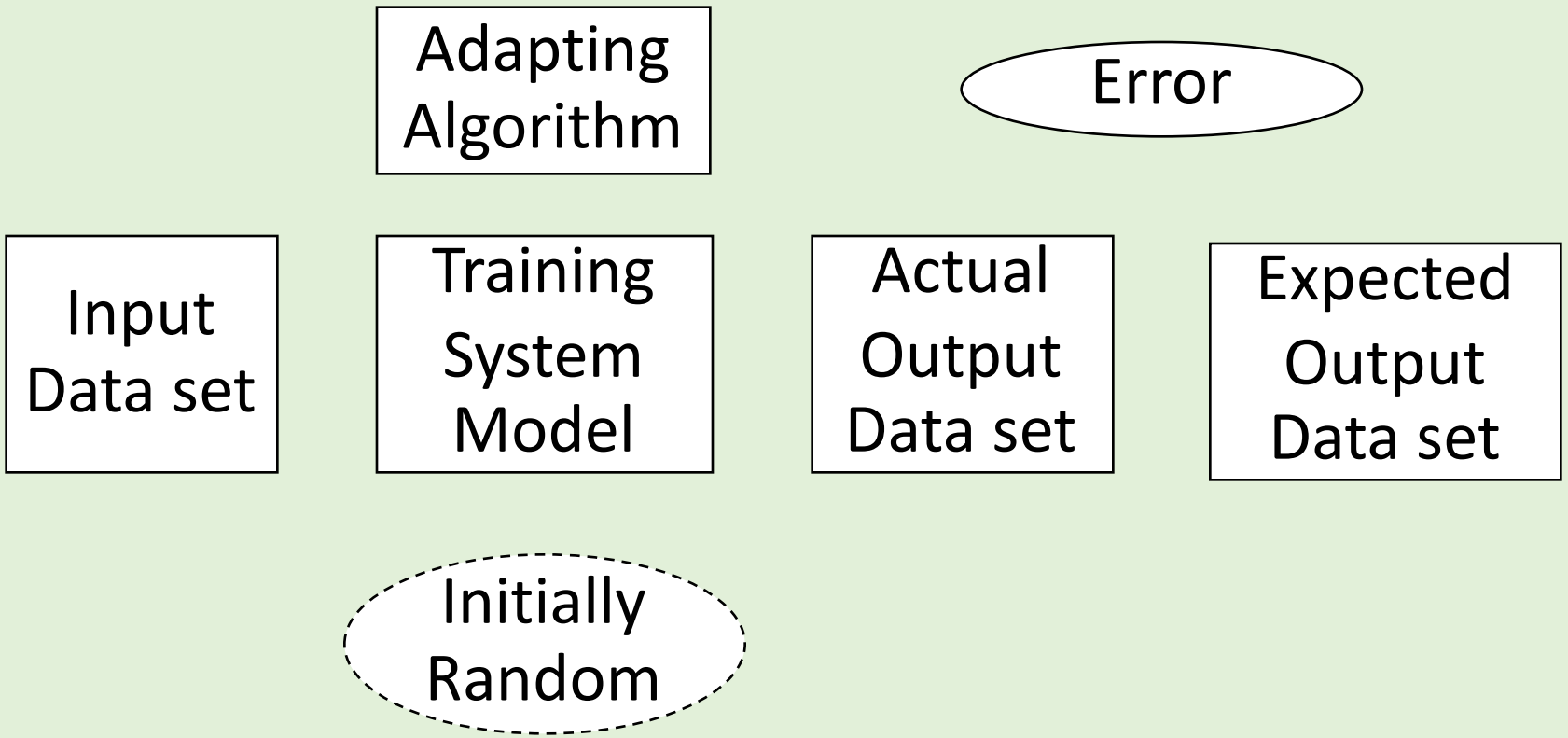
Training-Verification

Post processing

Neural Network



Neural Network (Learning step)



Neural Network (Learning step)

Acceptable Error,
No progress

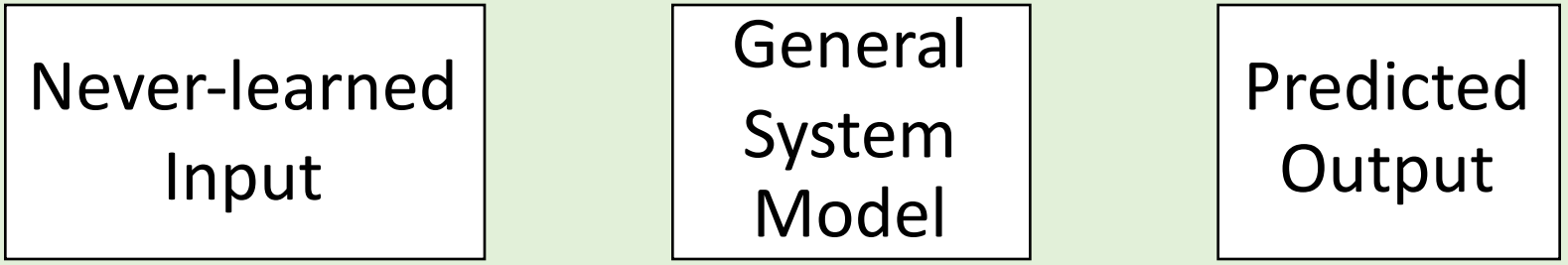
Input
Data set

Learned
System
Model

Actual
Output
Data set

Expected
Output
Data set

Neural Network (Learning step)



What can go wrong with your model?

- Gold mine problem
  - Throw away raw mine
- Landmines problem
  - Dig away the landmines
- Gold mining in landmines field

				ตู้ขายของอัตโนมัติ								
หมายเลข	ย่านที่ตั้ง	สัดส่วนคนหนุ่มสาว	จำนวนคนเดินผ่าน	ยอดขายรายเดือน	ยอดรวมสินค้า1	ยอดรวมสินค้า2	ยอดรวมสินค้า3	ยอดรวมสินค้า4	ค่าเช่า	ค่าไฟ	ระยะเวลาศูนย์บริการ	ค่าใช้จ่ายเดินสินค้า
1	A	1	500	5	1	2	1	1	3000	1000	5	1000
2	A	1	800	6	1	2	2	1	6000	1000	10	2000
3	B	1	300	9	2	3	3	1	3000	1000	5	1000
4	B	2	400	7	1	2	2	2	4000	1000	10	2000
5	B	2	400	9	2	3	2	2	3000	1000	5	1000
6	C	1	100	5	1	2	1	1	3000	1000	10	2000
7	C	2	300	9	1	3	2	3	4000	1000	5	1000
8	C	2	400	7	1	2	3	1	3000	1000	20	2000
9	C	2	100	8	1	3	1	3	3000	1000	20	2000
10	C	3	200	7	1	3	2	1	6000	1000	20	2000

Linear System

$Y = aX + b$			
Y	a	X	b
3	1	1	2
4	1	2	2
5	1	3	2
6	1	4	2
7	1	5	2

Non Linear System

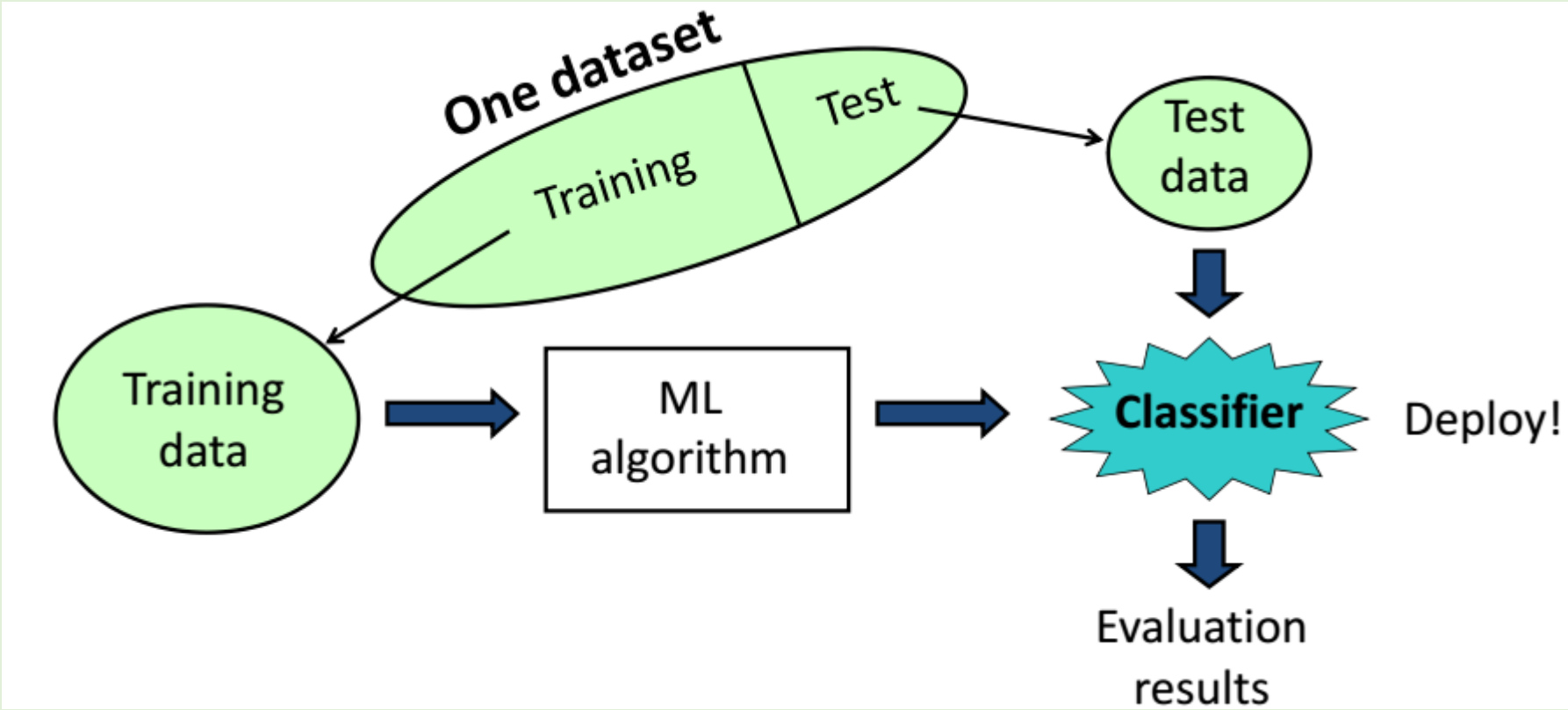
Y = aX^2 + b			
Y	a	X	b
3	1	1	2
6	1	2	2
11	1	3	2
18	1	4	2
27	1	5	2
3	1	-1	2
6	1	-2	2
11	1	-3	2
18	1	-4	2
27	1	-5	2

Normalized input/output

- Mean 0
- Variances +/- 1

- Normal Noise

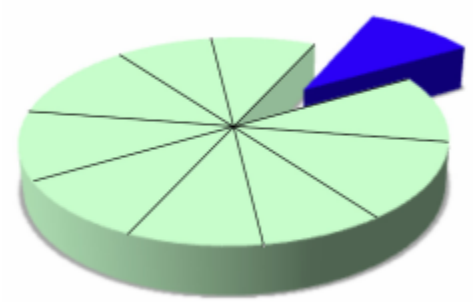
Training and Test Dataset



Validation

10-fold cross-validation

- Divide dataset into 10 parts (folds)
- Hold out each part in turn
- Average the results
- Each data point used once for testing, 9 times for training



Stratified cross-validation

- Ensure that each fold has the right proportion of each class value

Learning

- Feature Extraction - Dimension Reduction
- Curse of Dimensional
- Overfitting

Rules - IF Then - Decision Tree

- Generate rule or program due to complete possible input and outcomes
- What if not complete outcomes?

Weather & Play

Instants

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

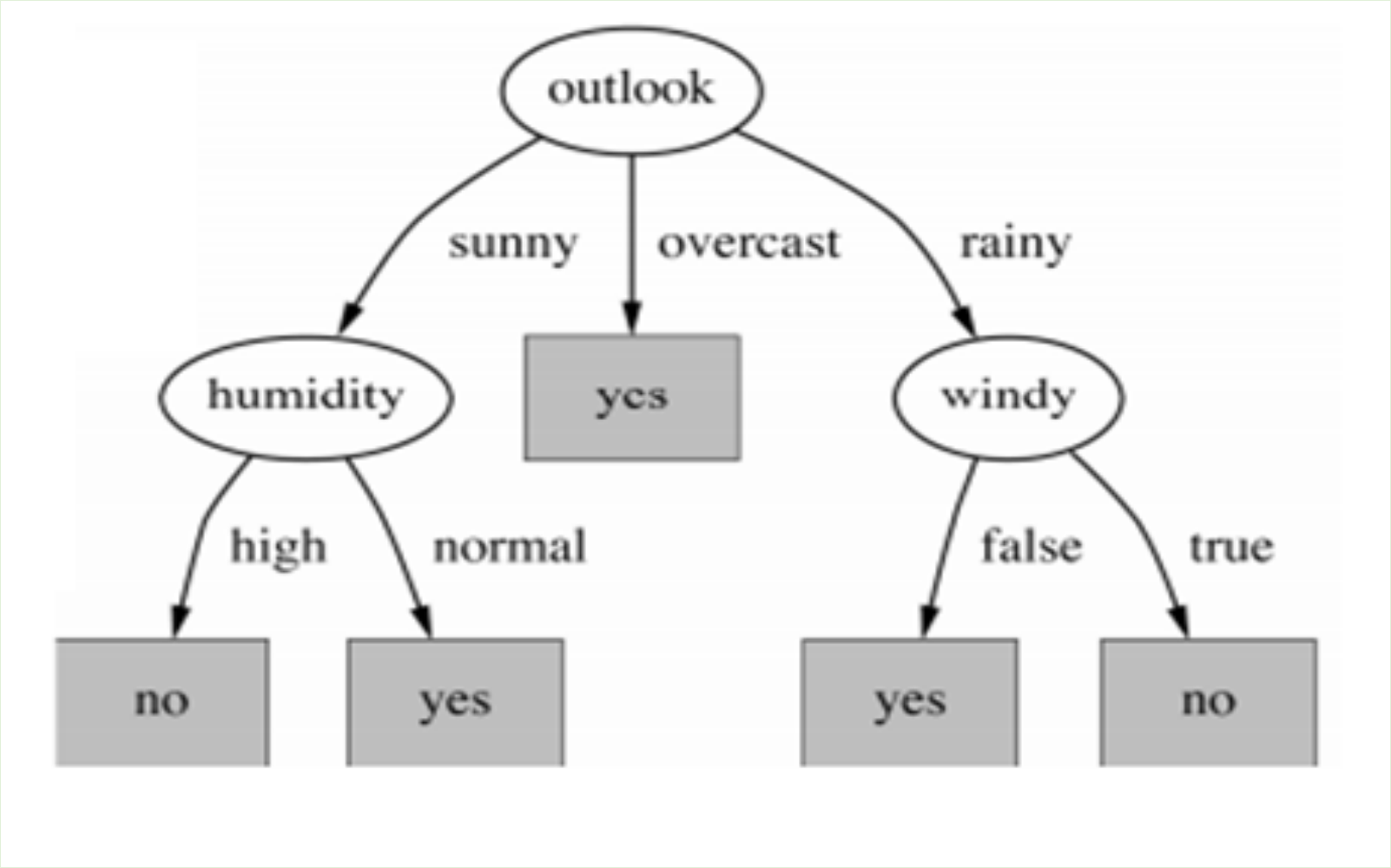
Attributes

Outcomes

Weather & Play

Outlook	Sunny	Overcast	Rainy
Temp	Hot	Cool	Mild
Humidity	High	Normal	
Windy	True	False	

Decision Tree



Weather & Play

		Numeric		Nominal	
Temp	35C	Hot	Mild	Cool
Humidity	40%	High	Normal	
Windy	5	True	False	

Weather & Play

		Numeric		Normalized
Temp	35C	-0.2	0.1	0.15
Humidity	40%	-1	+1	
Windy	5	+1	-1	

Weather & Play

14 Instants

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

3x3x2x2 =36 Possible

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

A new day:

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

Naïve Bayes

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Naïve Bayes

Outlook			Temperature			Humidity			Wind			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

A new day:

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

Naïve Bayes

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes]$$
$$\times \Pr[Temperature = Cool \mid yes]$$
$$\times \Pr[Humidity = High \mid yes]$$
$$\times \Pr[Windy = True \mid yes]$$
$$\times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/ × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

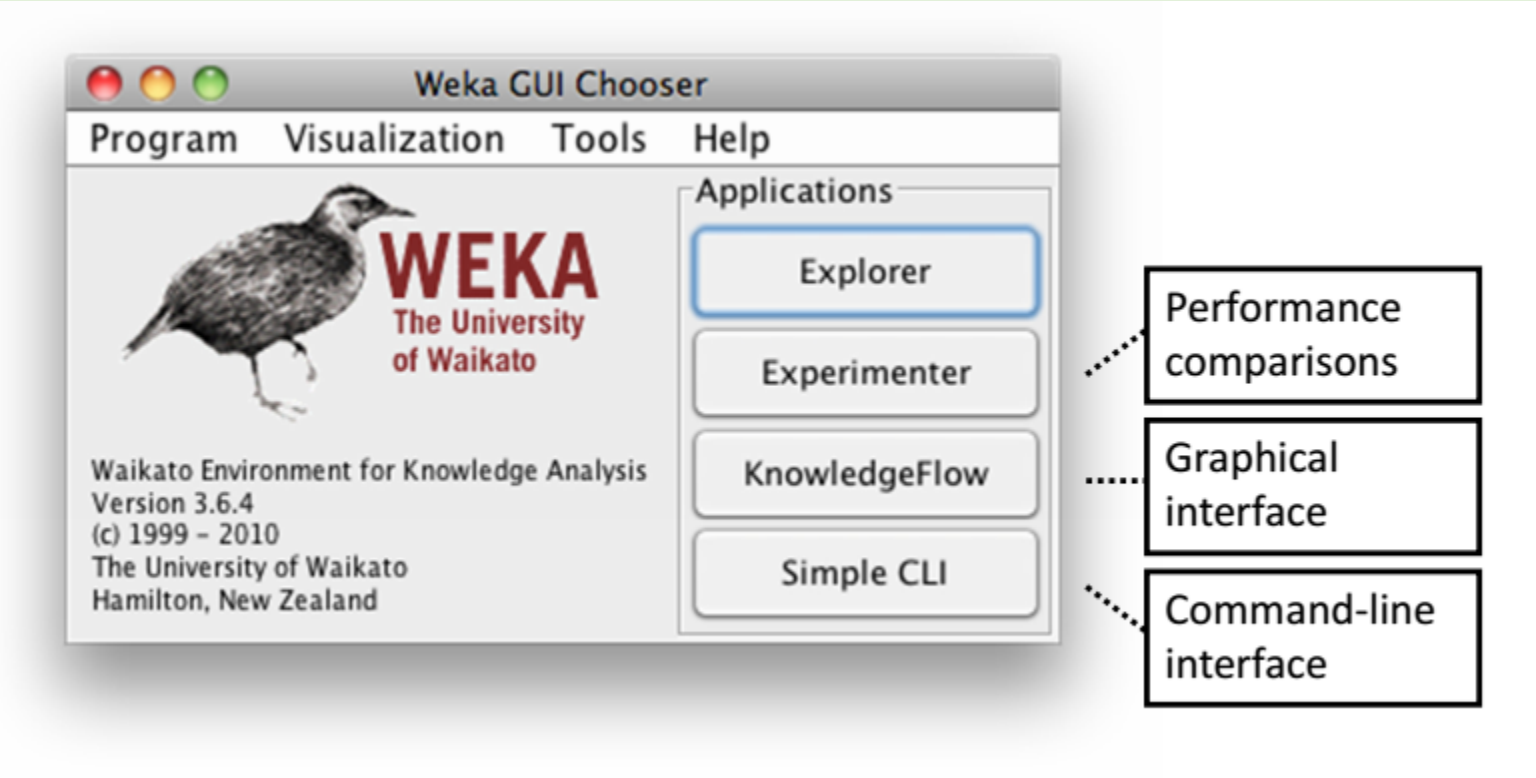
Weka : Introduction

Machine learning algorithms for data mining tasks

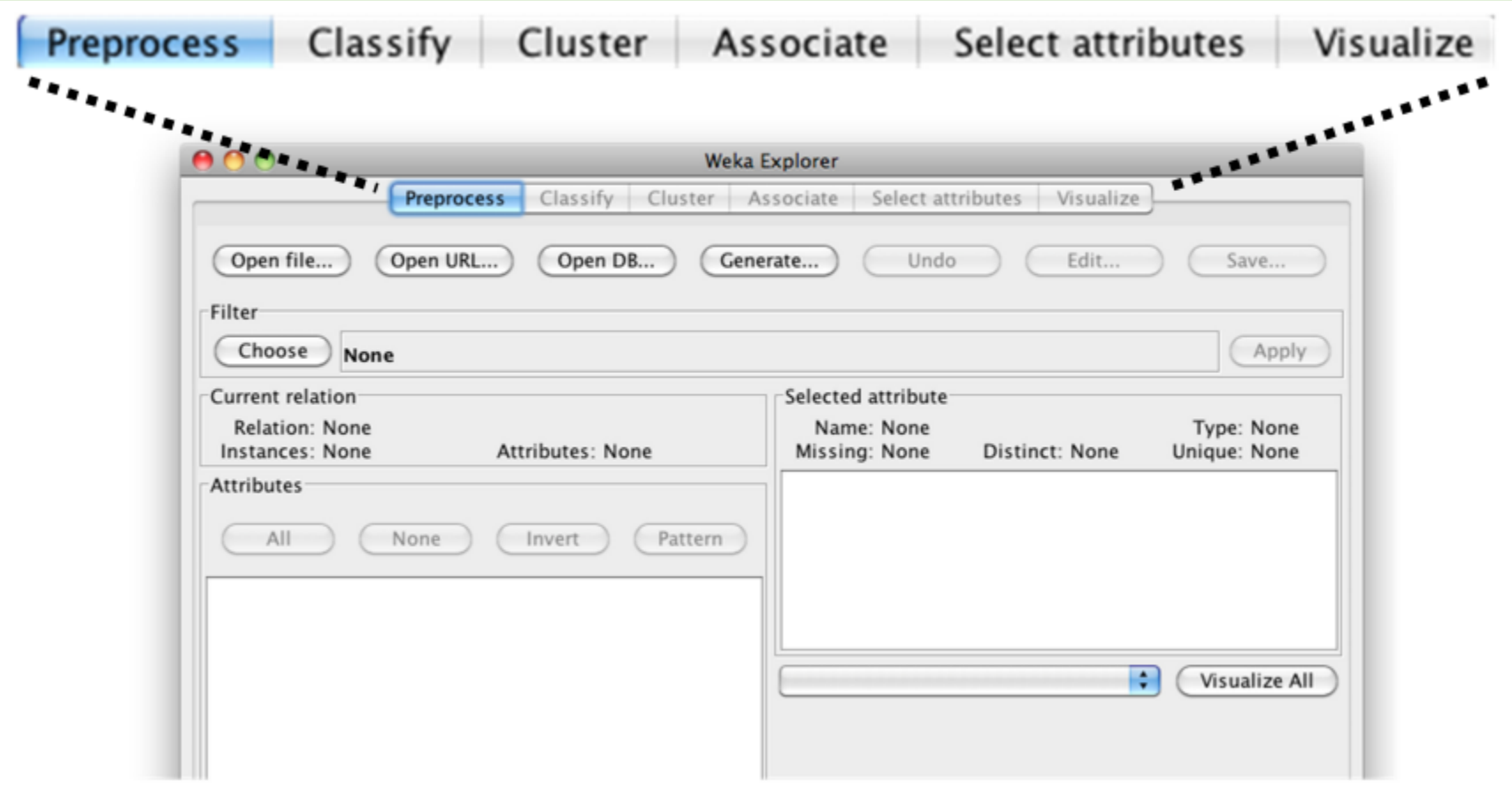
- 100+ algorithms for classification
- 75 for data preprocessing
- 25 to assist with feature selection
- 20 for clustering, finding association rules, etc

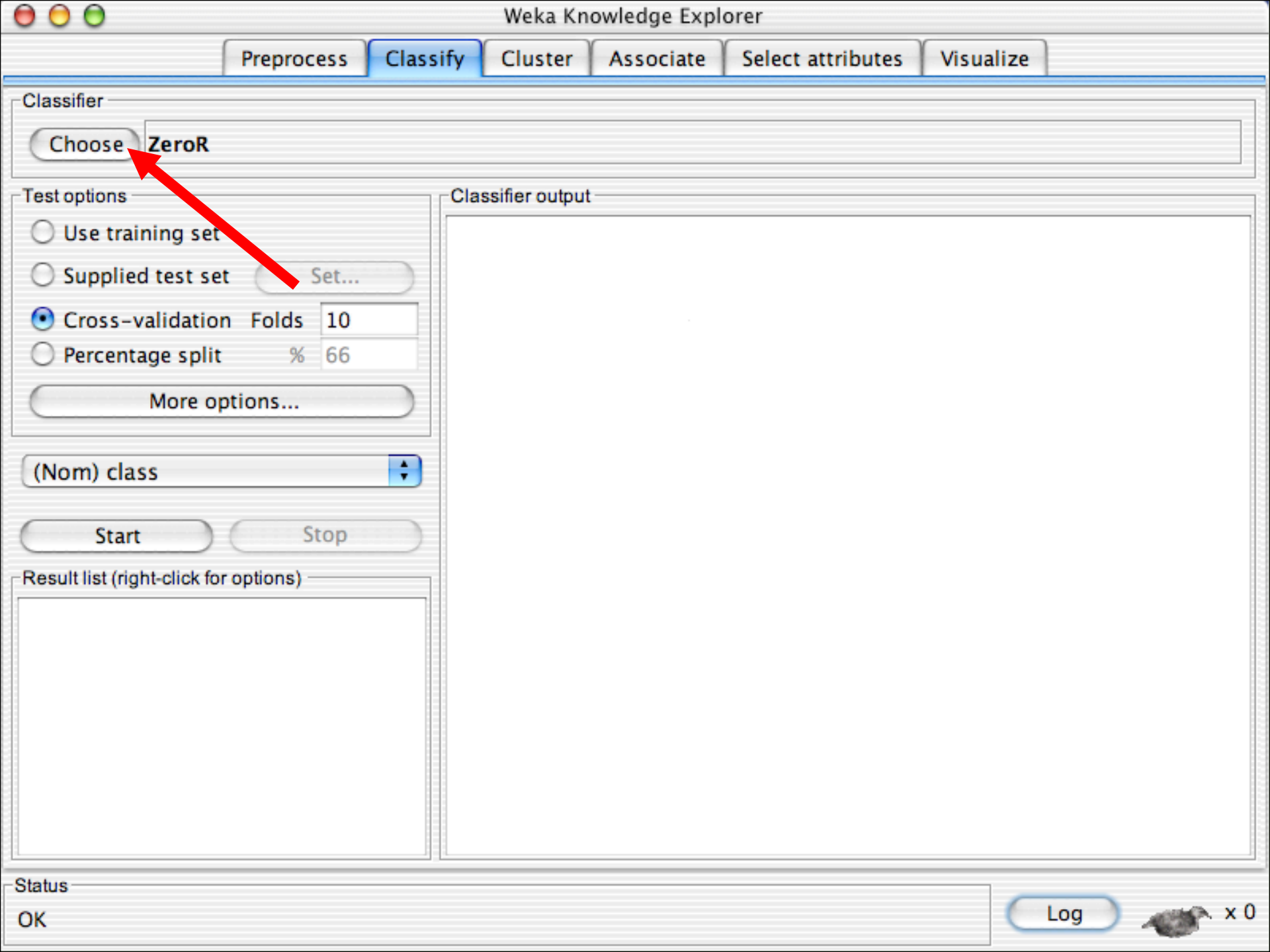
Weka : Basic Steps

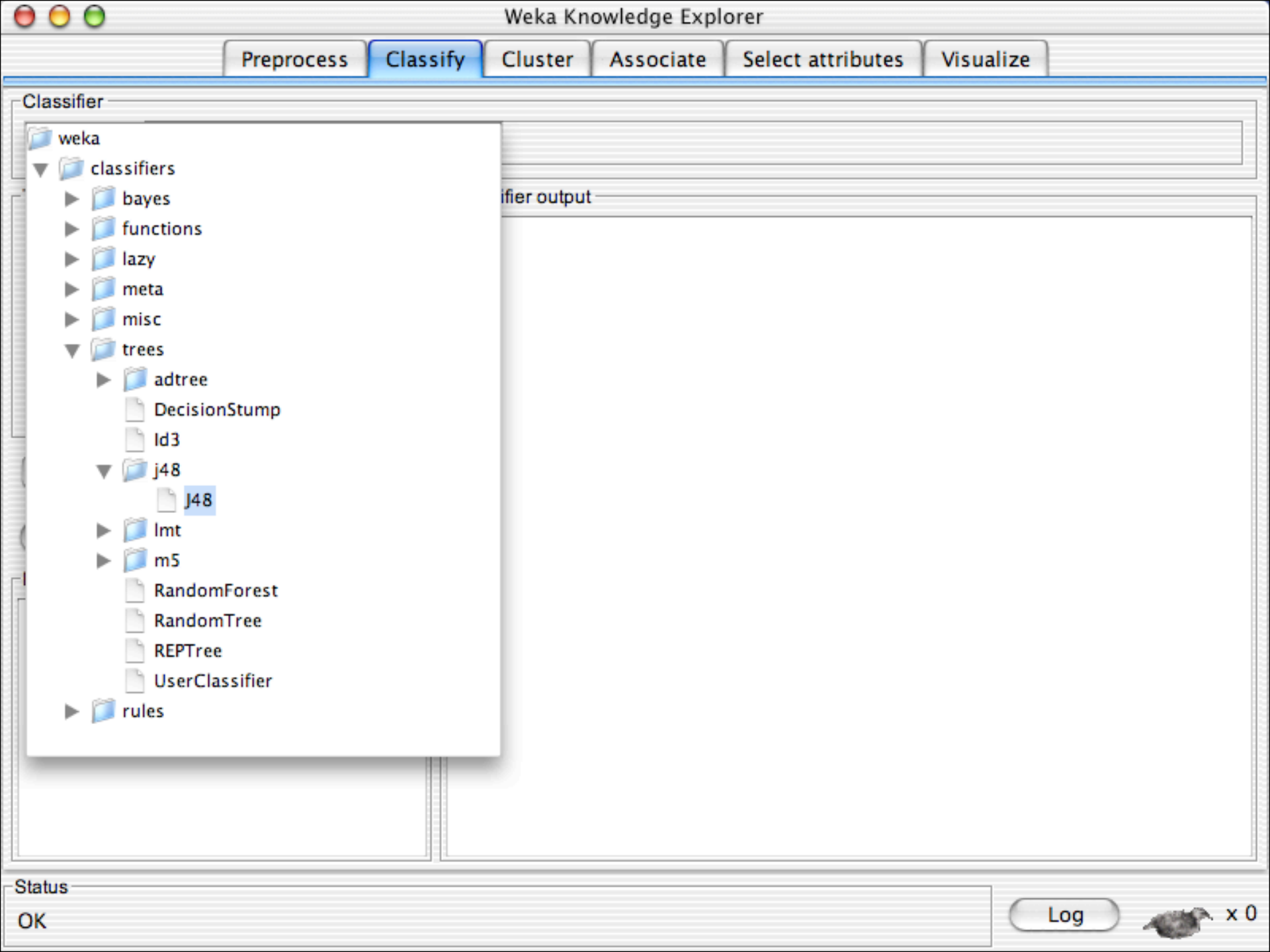
- Pay attention to [Explorer]

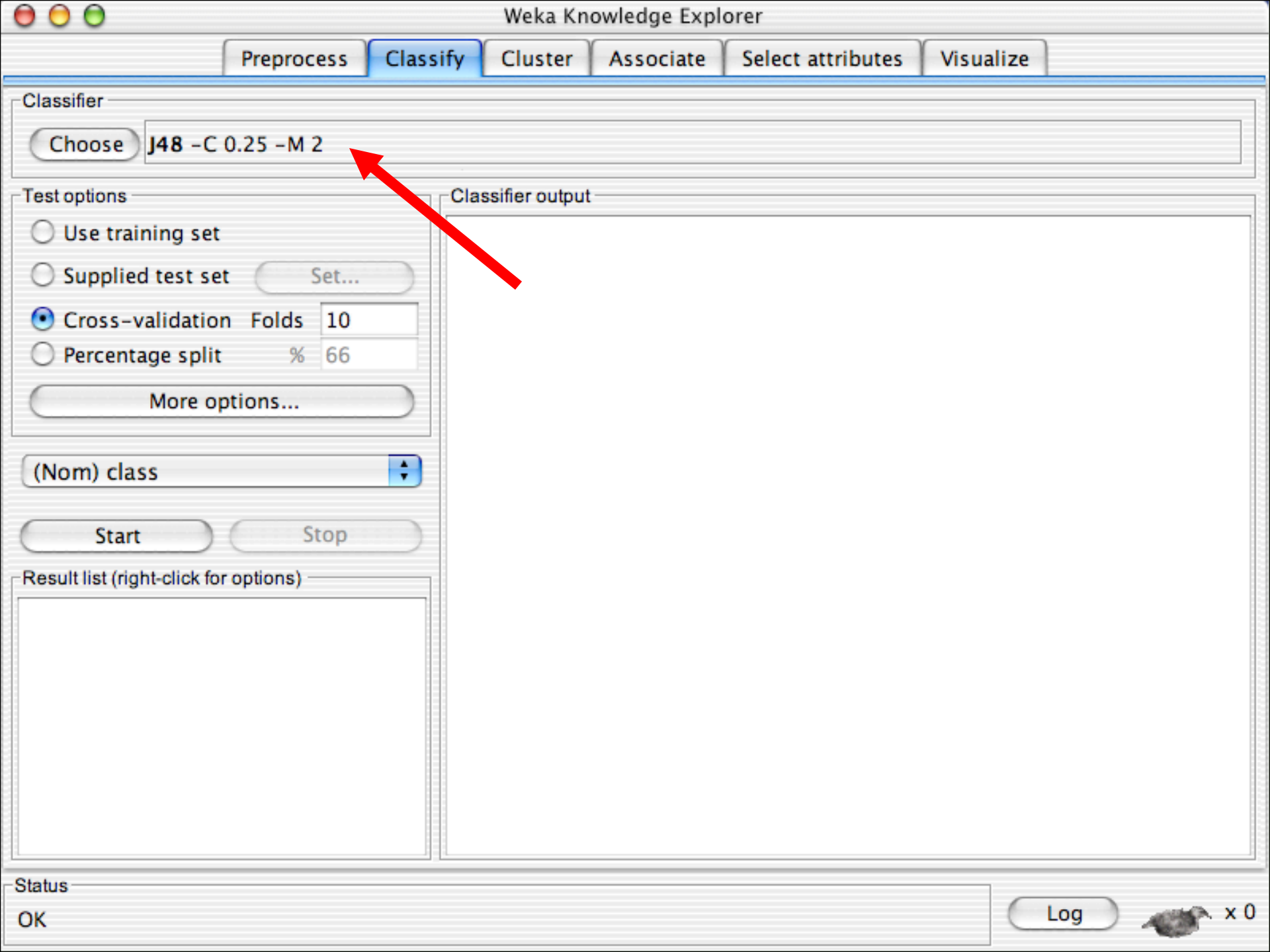


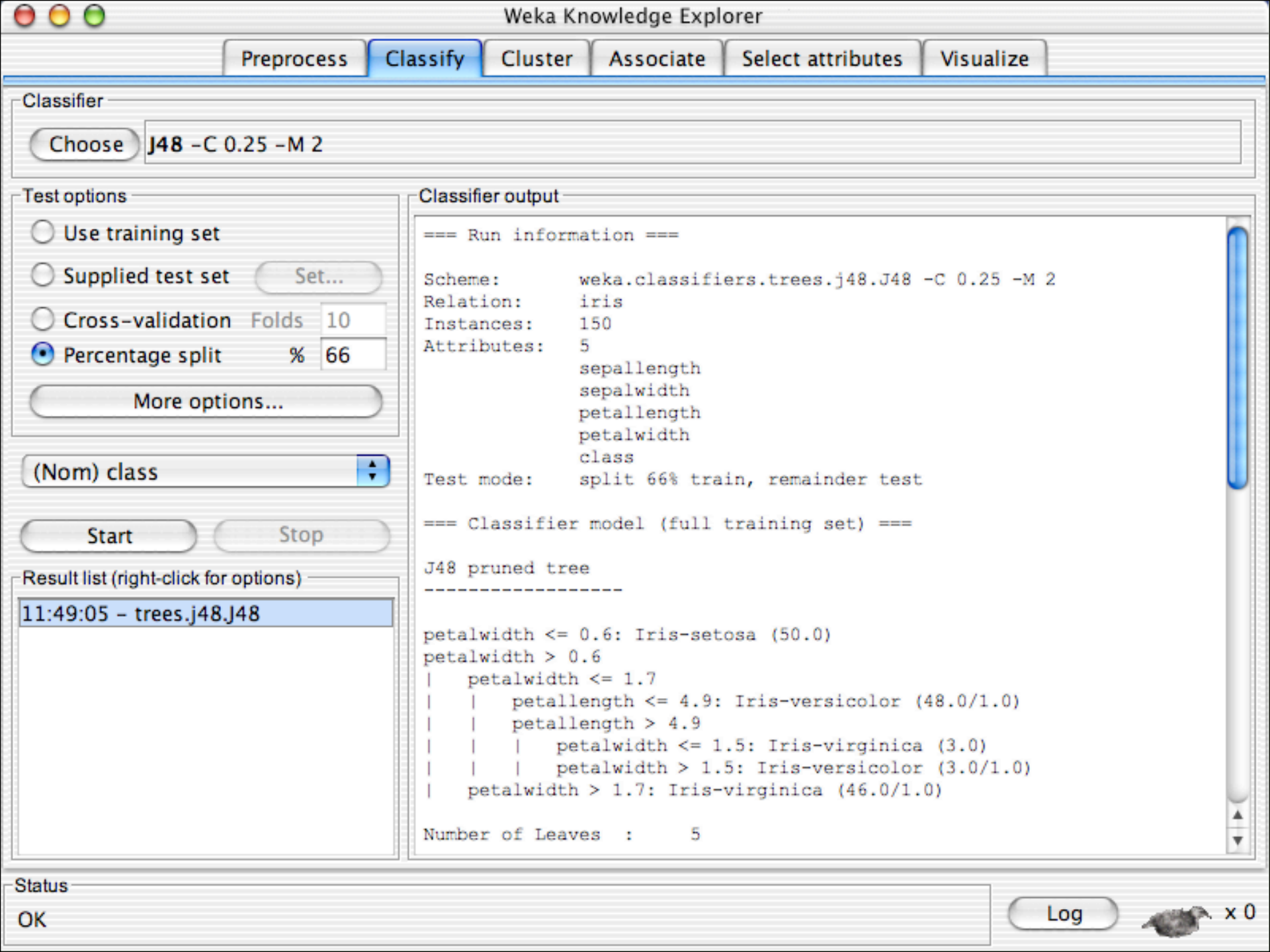
WEKA: Tasks

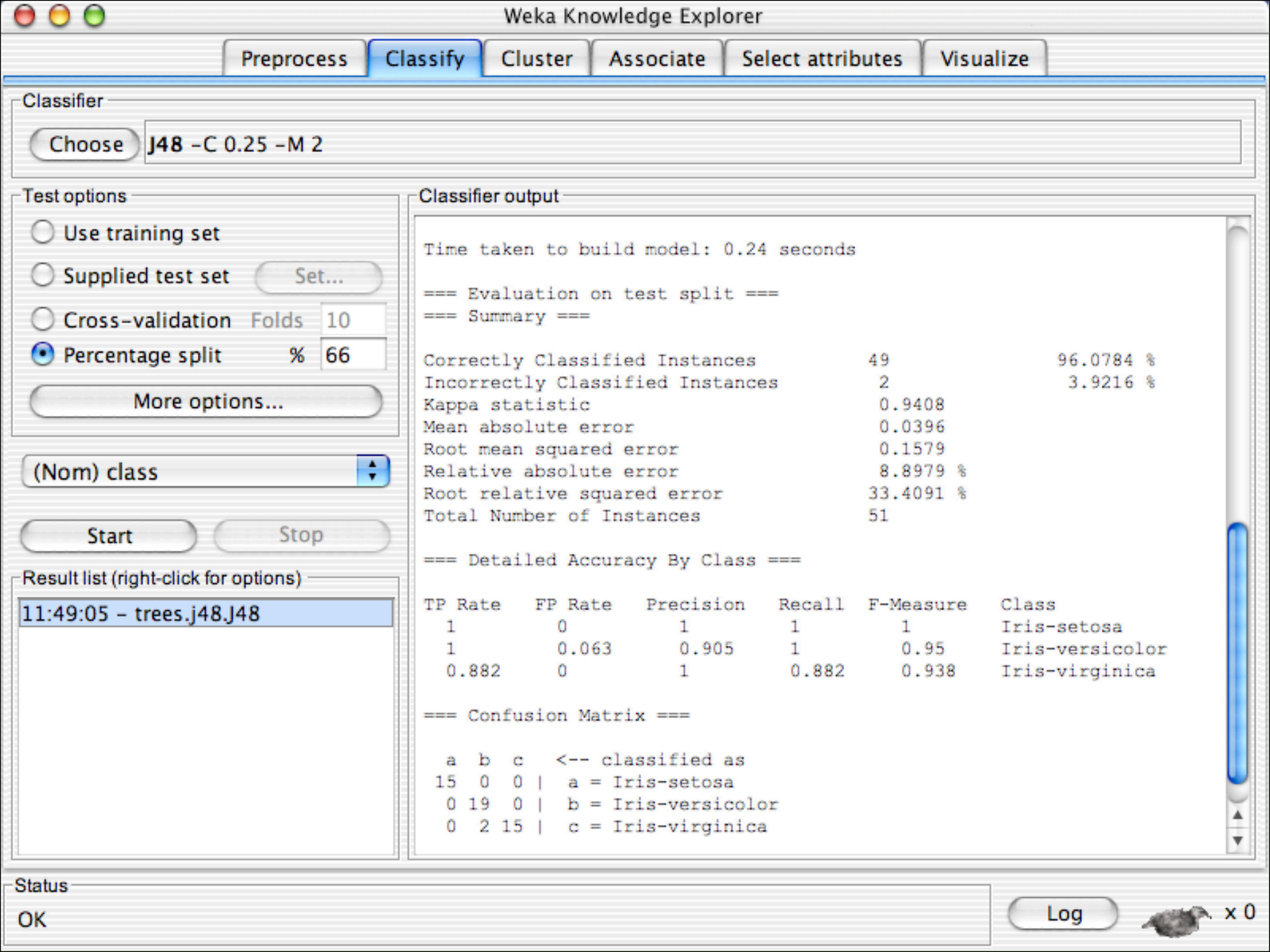

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All

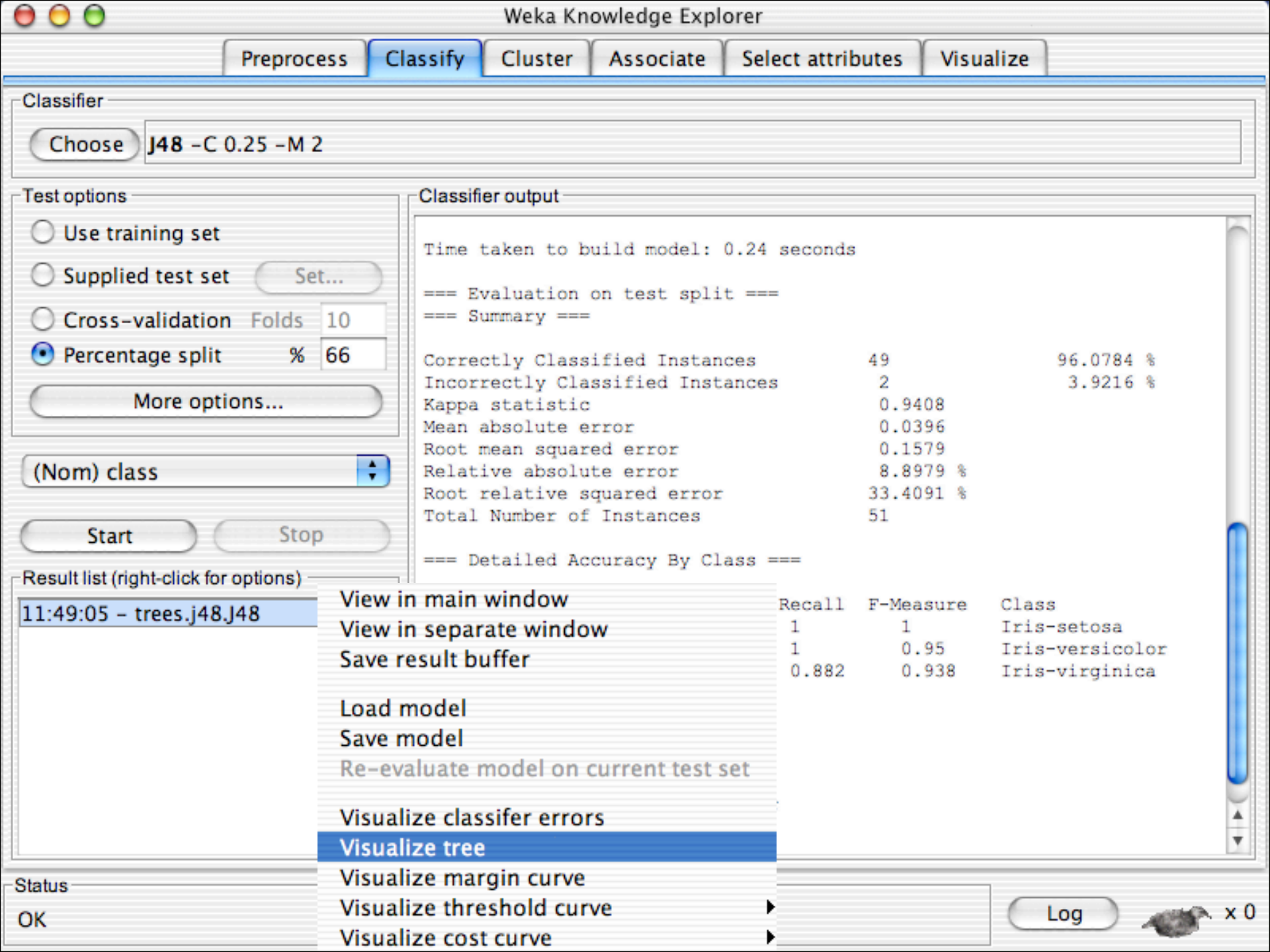

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for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

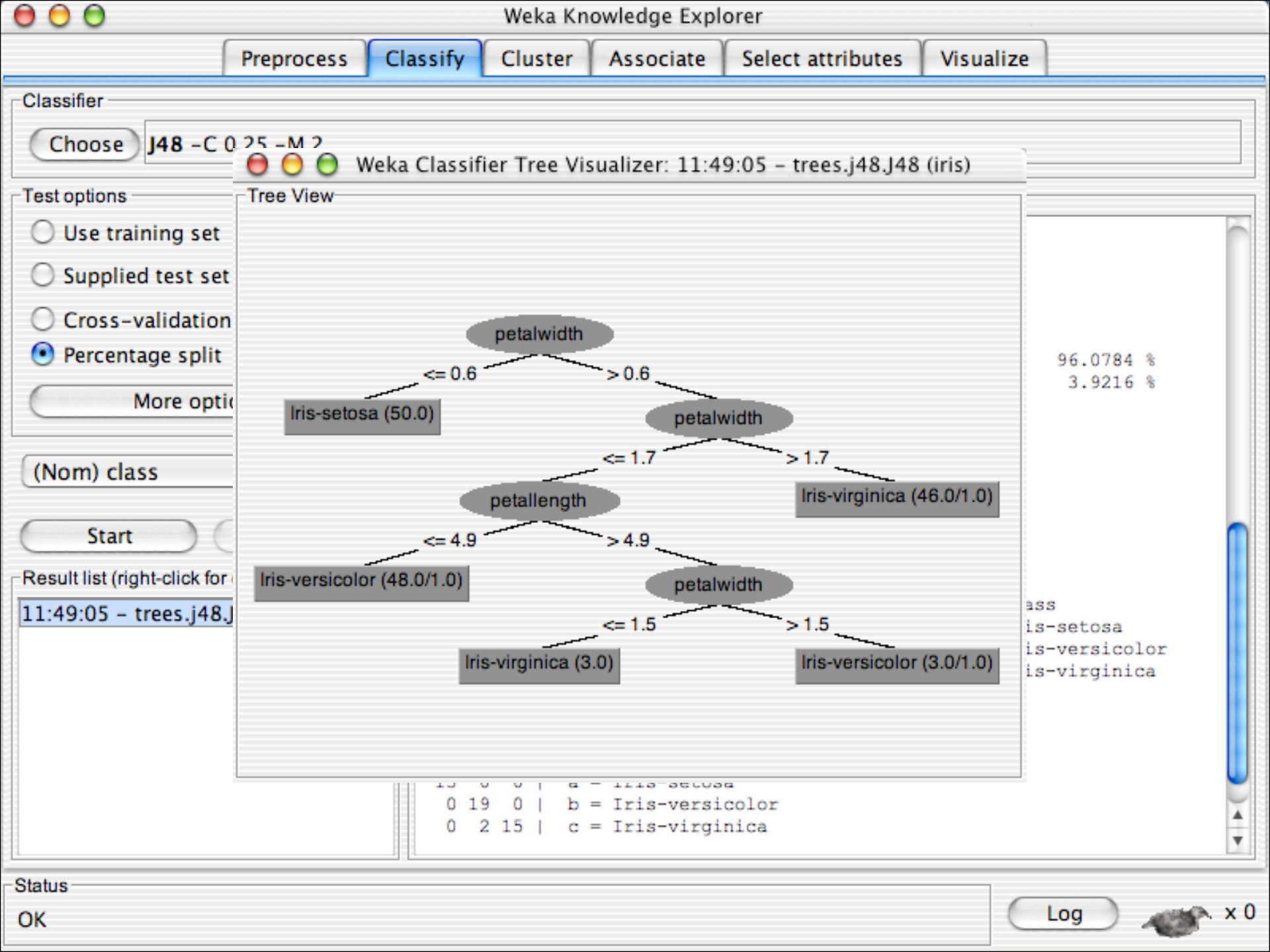

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
adtree
DecisionStump
Id3
j48
J48
lmt
m5
RandomForest
RandomTree
REPTree
UserClassifier
rules
ifier output
Status
OK
Log
x 0

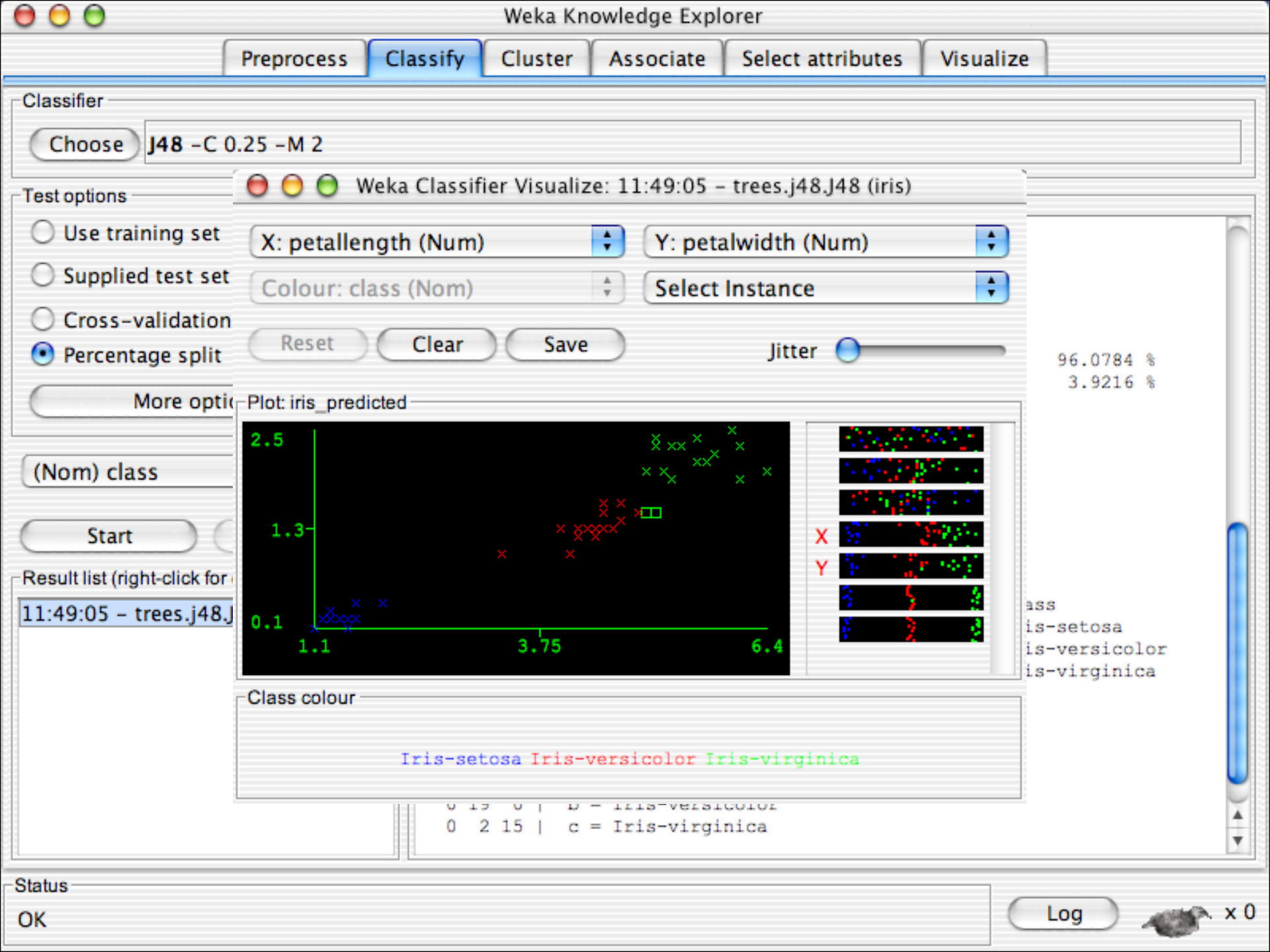

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click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

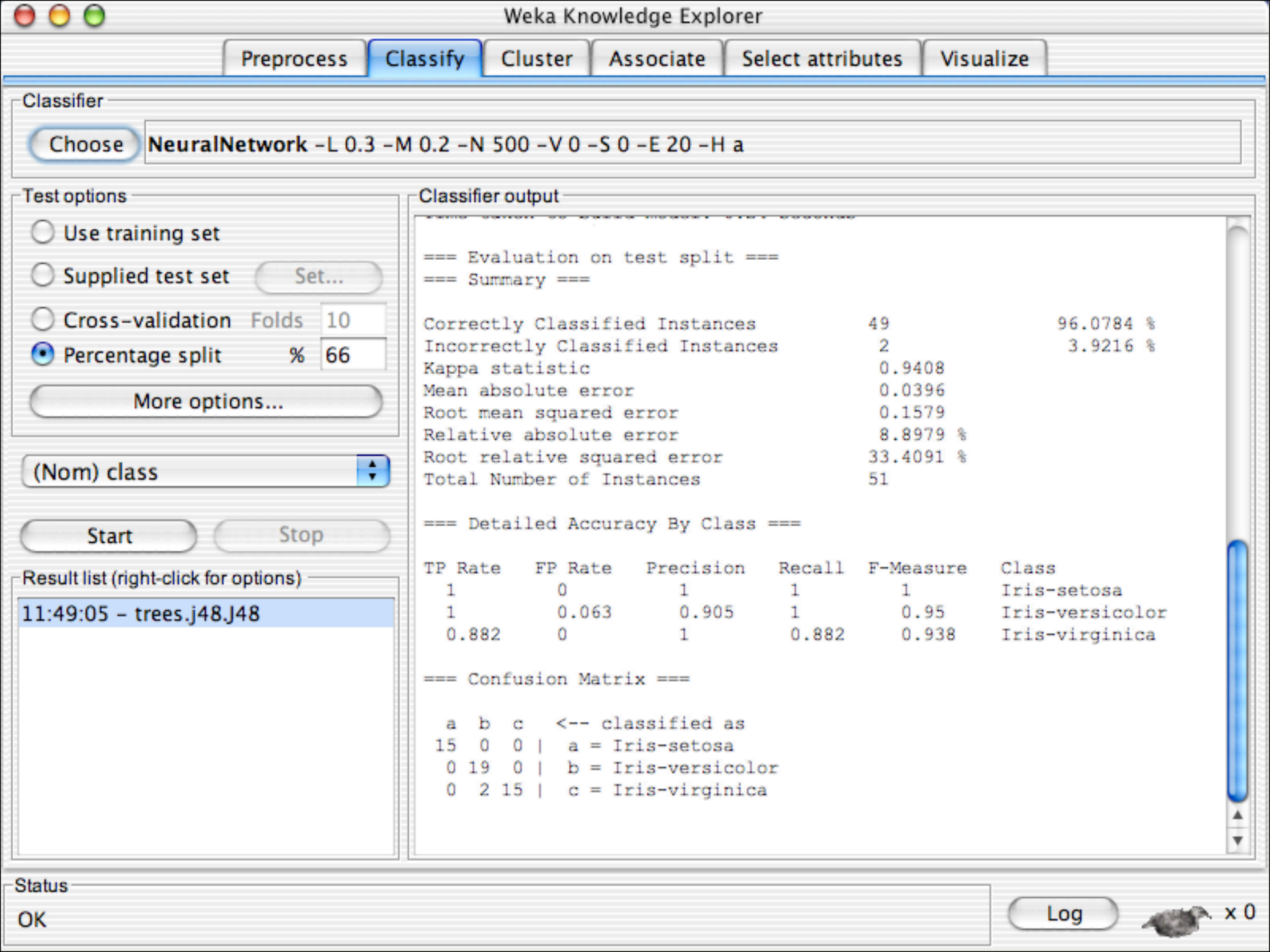

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)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

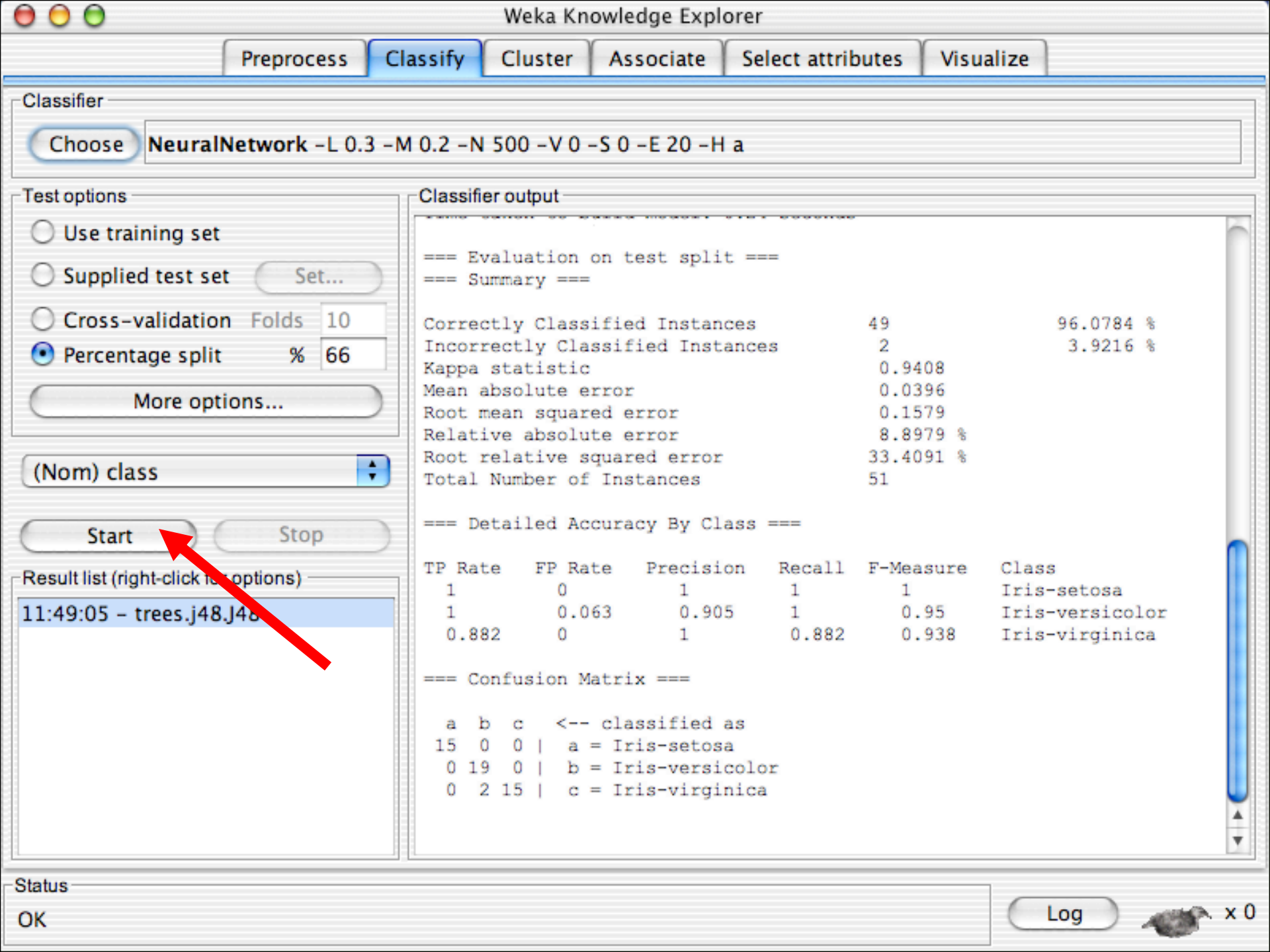

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)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


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Weka Classifier Tree Visualizer: 11:49:05 - trees.j48.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More optic
(Nom) class
Start
Result list (right-click for
11:49:05 - trees.j48.J
96.0784 %
3.9216 %
ass
is-setosa
is-versicolor
is-virginica
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
Cross-validation
Percentage split
More optic
(Nom) class
Start
Result list (right-click for
11:49:05 - trees.j48.J
Weka Classifier Visualize: 11:49:05 - trees.j48.J48 (iris)
X: petallength (Num)
Y: petalwidth (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: iris_predicted
2.5
1.3
0.1
1.1
3.75
6.4
X
Y
Class colour
Iris-setosa Iris-versicolor Iris-virginica
96.0784 %
3.9216 %
ass
is-setosa
is-versicolor
is-virginica
0 2 15 | c = Iris-virginica
Status
OK
Log
x 0

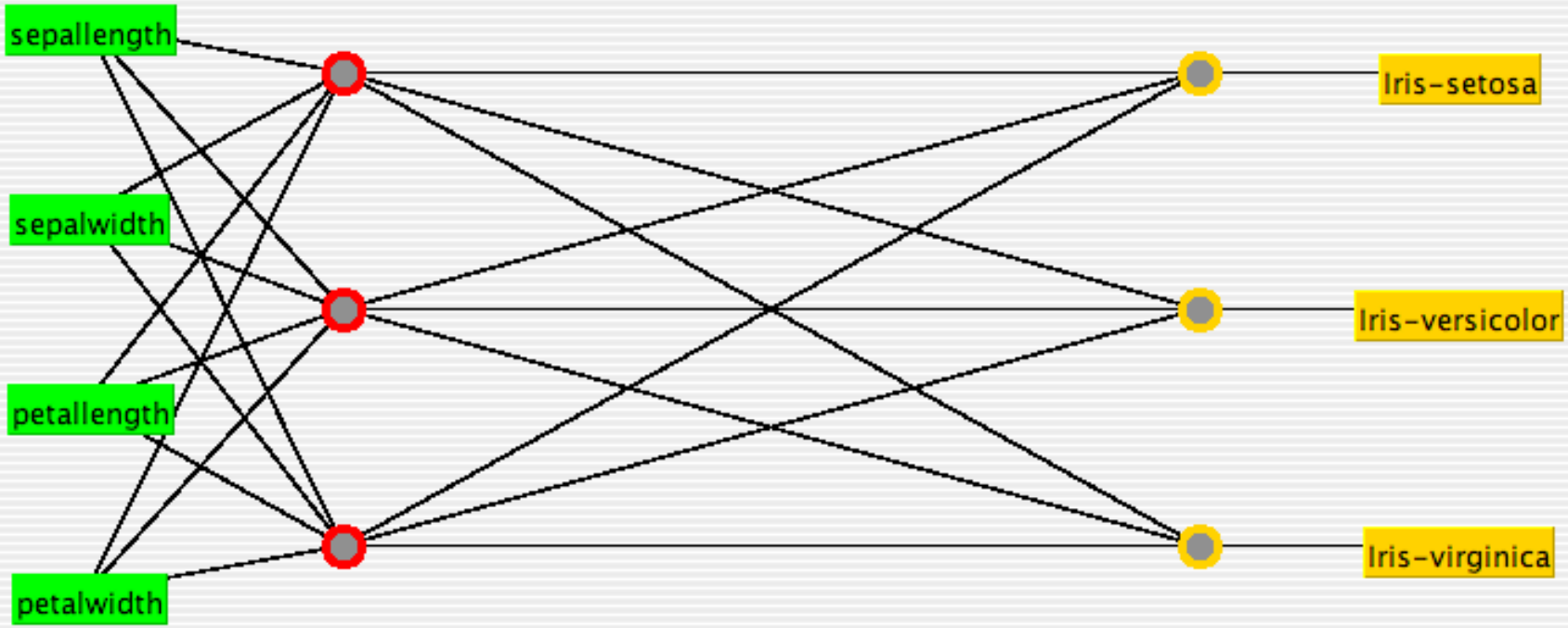

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
Classifier output
=== 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
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
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
=== 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

Neural Network



Controls

Start

Accept

Epoch 0

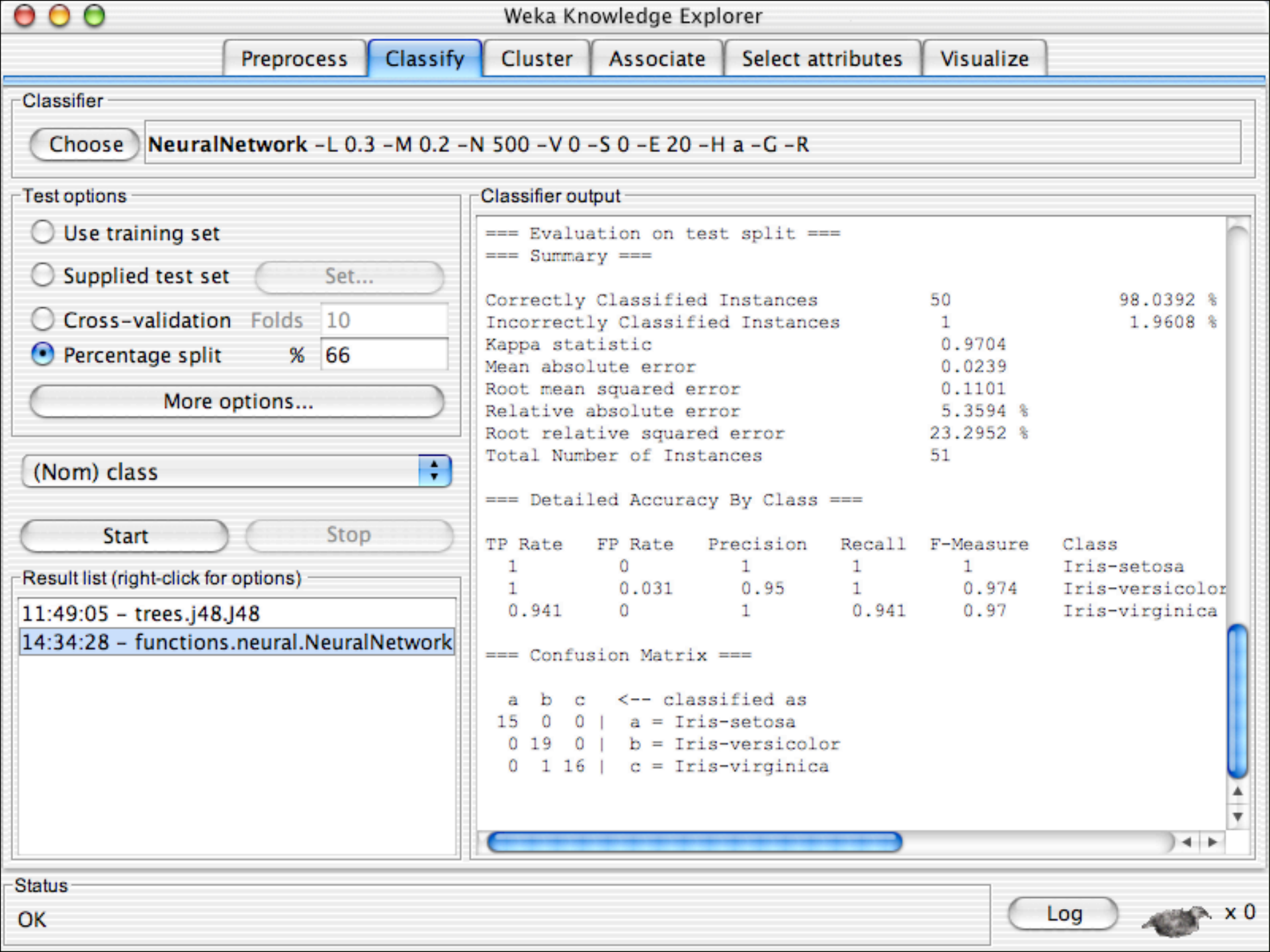
Num Of Epochs 500

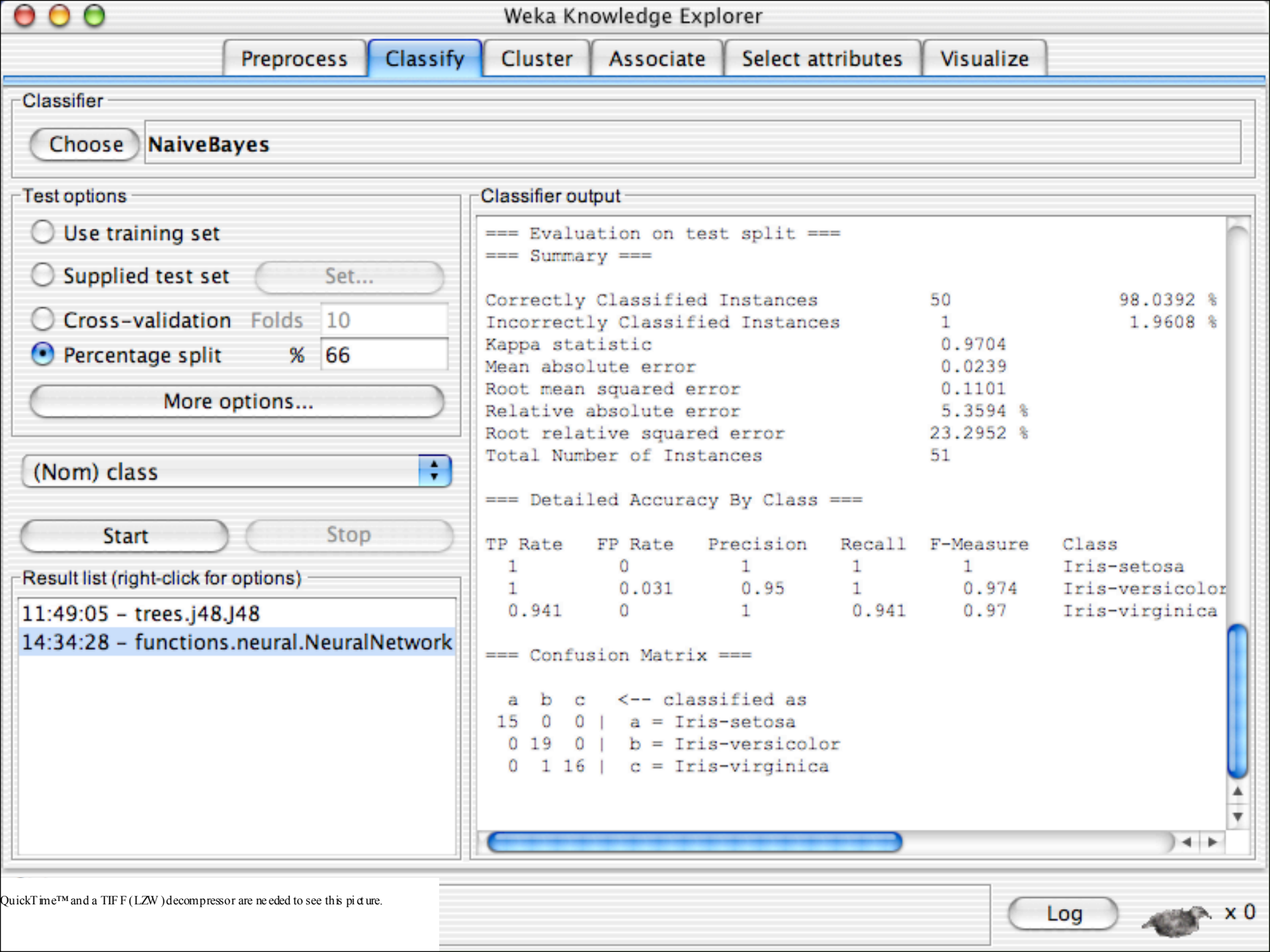
Error per Epoch = 0

Learning Rate = 0.3

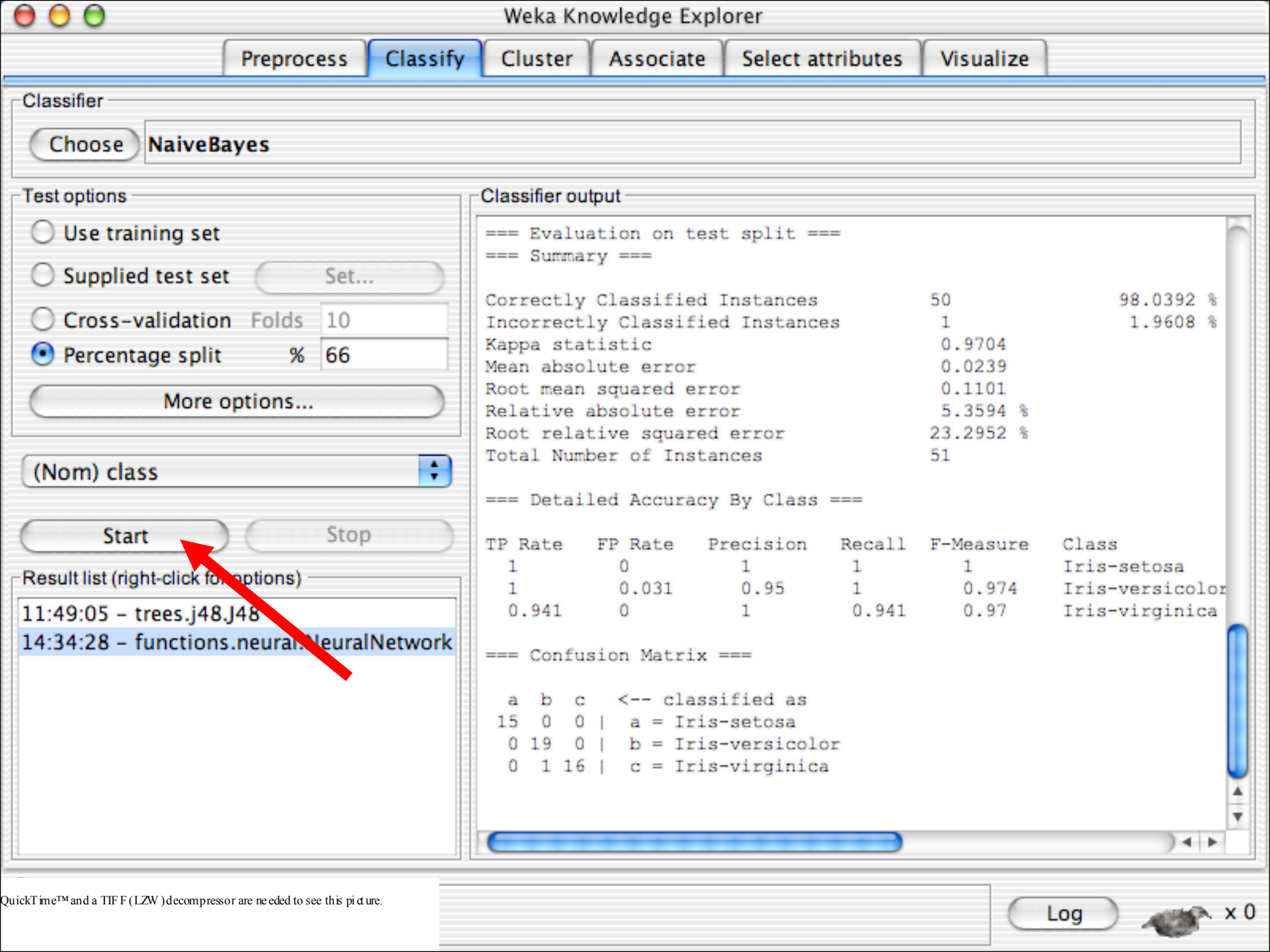
Momentum = 0.2

Building model on training data...

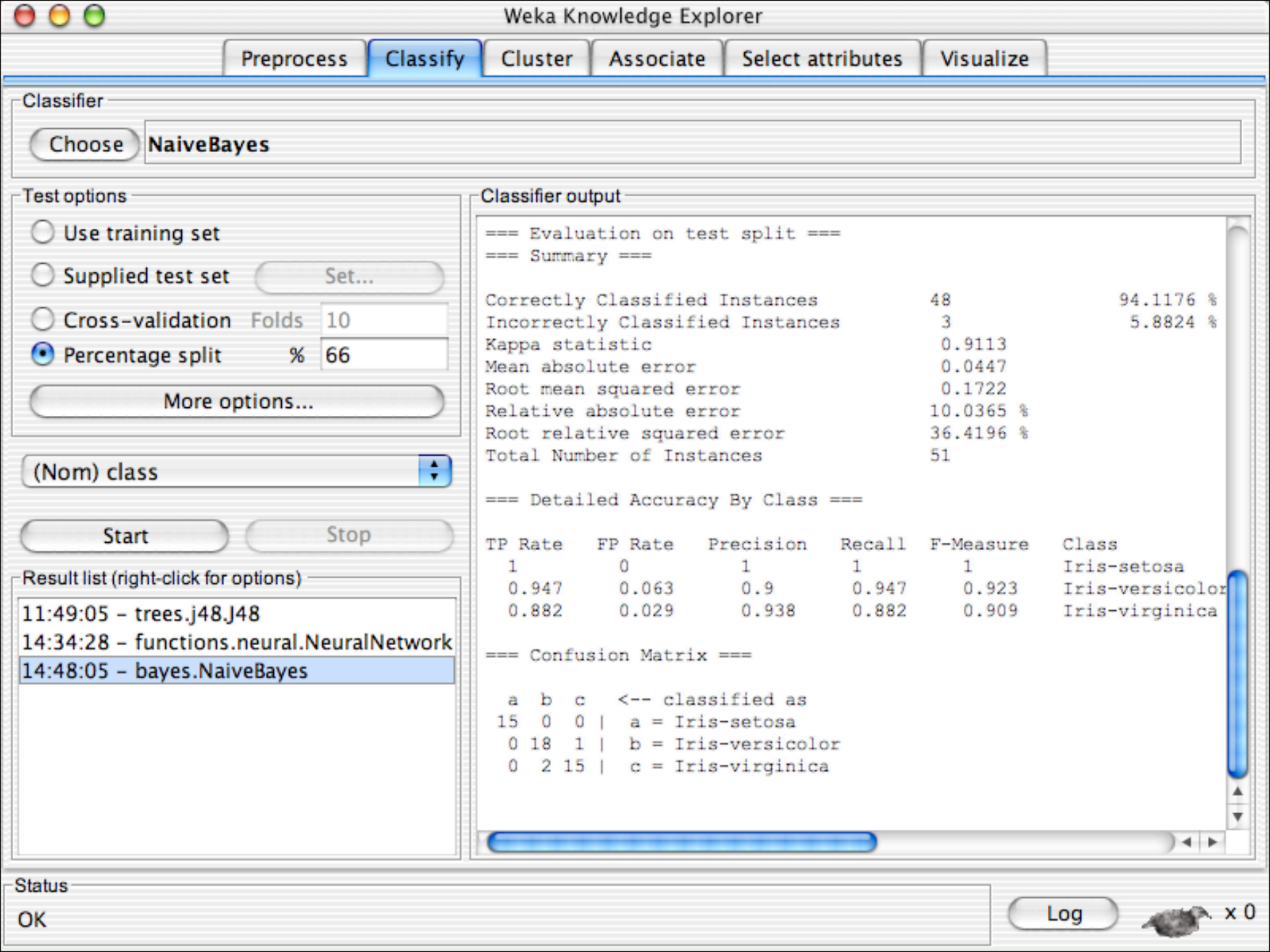

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
14:34:28 - functions.neural.NeuralNetwork
Classifier output
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 50 98.0392 %
Incorrectly Classified Instances 1 1.9608 %
Kappa statistic 0.9704
Mean absolute error 0.0239
Root mean squared error 0.1101
Relative absolute error 5.3594 %
Root relative squared error 23.2952 %
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 0.031 0.95 1 0.974 Iris-versicolor
0.941 0 1 0.941 0.97 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 1 16 | c = Iris-virginica
Status
OK
Log
x 0

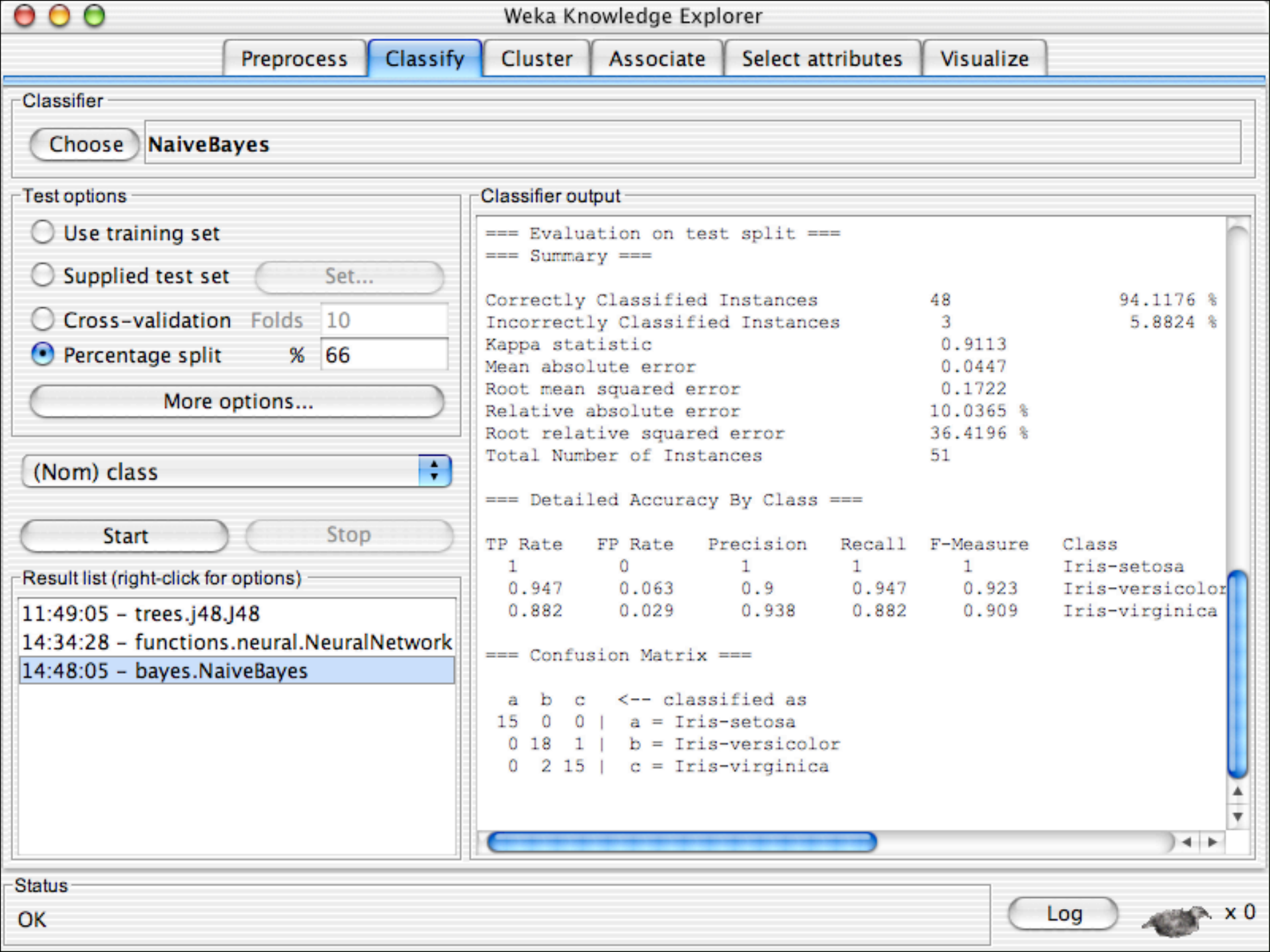


QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.



QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
14:34:28 - functions.neural.NeuralNetwork
14:48:05 - bayes.NaiveBayes
Classifier output
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 48 94.1176 %
Incorrectly Classified Instances 3 5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
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
0.947 0.063 0.9 0.947 0.923 Iris-versicolor
0.882 0.029 0.938 0.882 0.909 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 18 1 | 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
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:49:05 - trees.j48.J48
14:34:28 - functions.neural.NeuralNetwork
14:48:05 - bayes.NaiveBayes
Classifier output
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances 48 94.1176 %
Incorrectly Classified Instances 3 5.8824 %
Kappa statistic 0.9113
Mean absolute error 0.0447
Root mean squared error 0.1722
Relative absolute error 10.0365 %
Root relative squared error 36.4196 %
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
0.947 0.063 0.9 0.947 0.923 Iris-versicolor
0.882 0.029 0.938 0.882 0.909 Iris-virginica
=== Confusion Matrix ===
a b c <-- classified as
15 0 0 | a = Iris-setosa
0 18 1 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
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: Glass
Instances: 214 Attributes: 10

Attributes

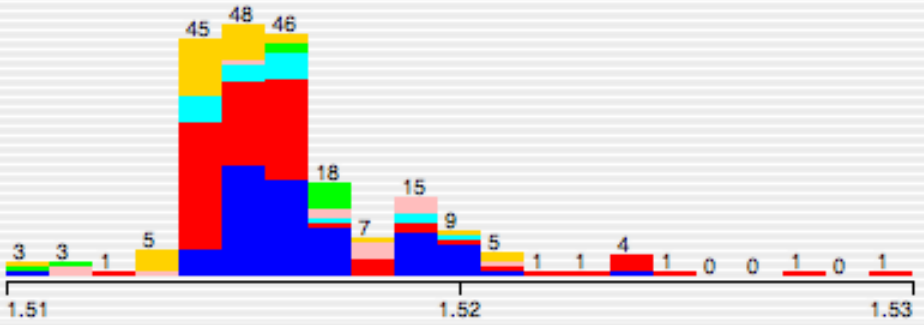
No.	Name
1	RI
2	Na
3	Mg
4	Al
5	Si
6	K
7	Ca
8	Ba
9	Fe
10	Type

Selected attribute

Name: RI Type: Numeric
Missing: 0 (0%) Distinct: 178 Unique: 145 (68%)

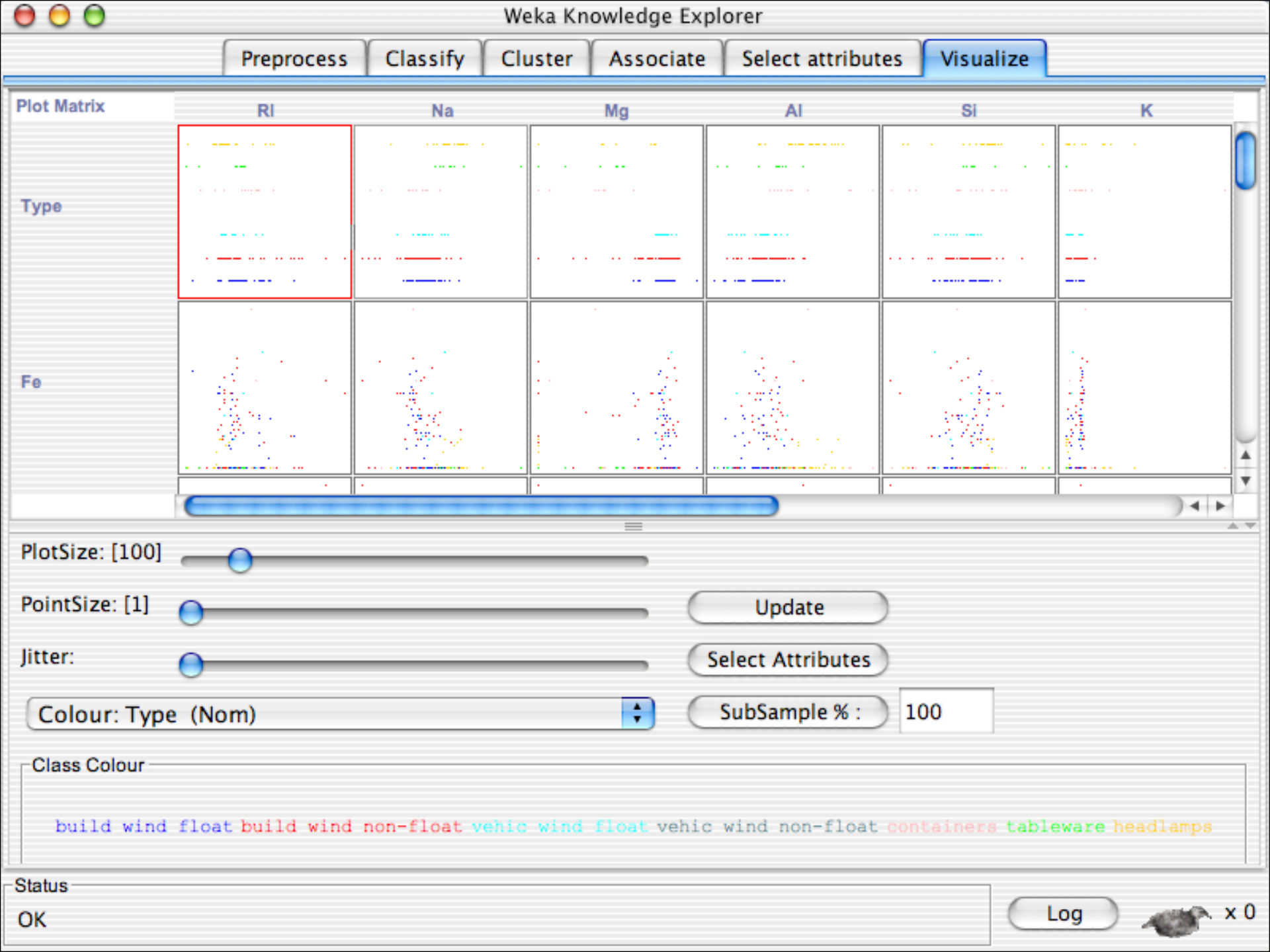
Statistic	Value
Minimum	1.511
Maximum	1.534
Mean	1.518
StdDev	0.003

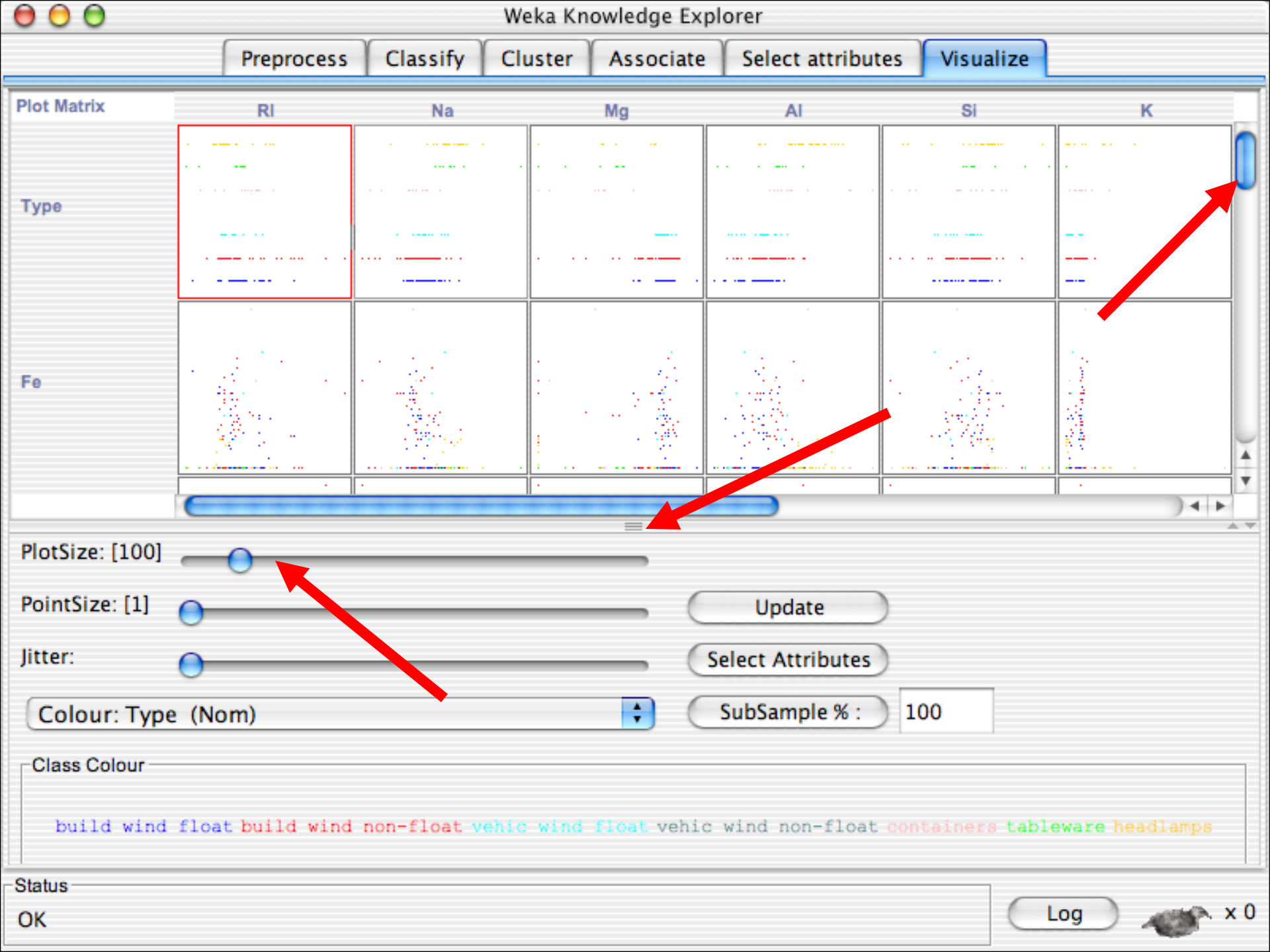
Colour: Type (Nom) | Visualize All



Status

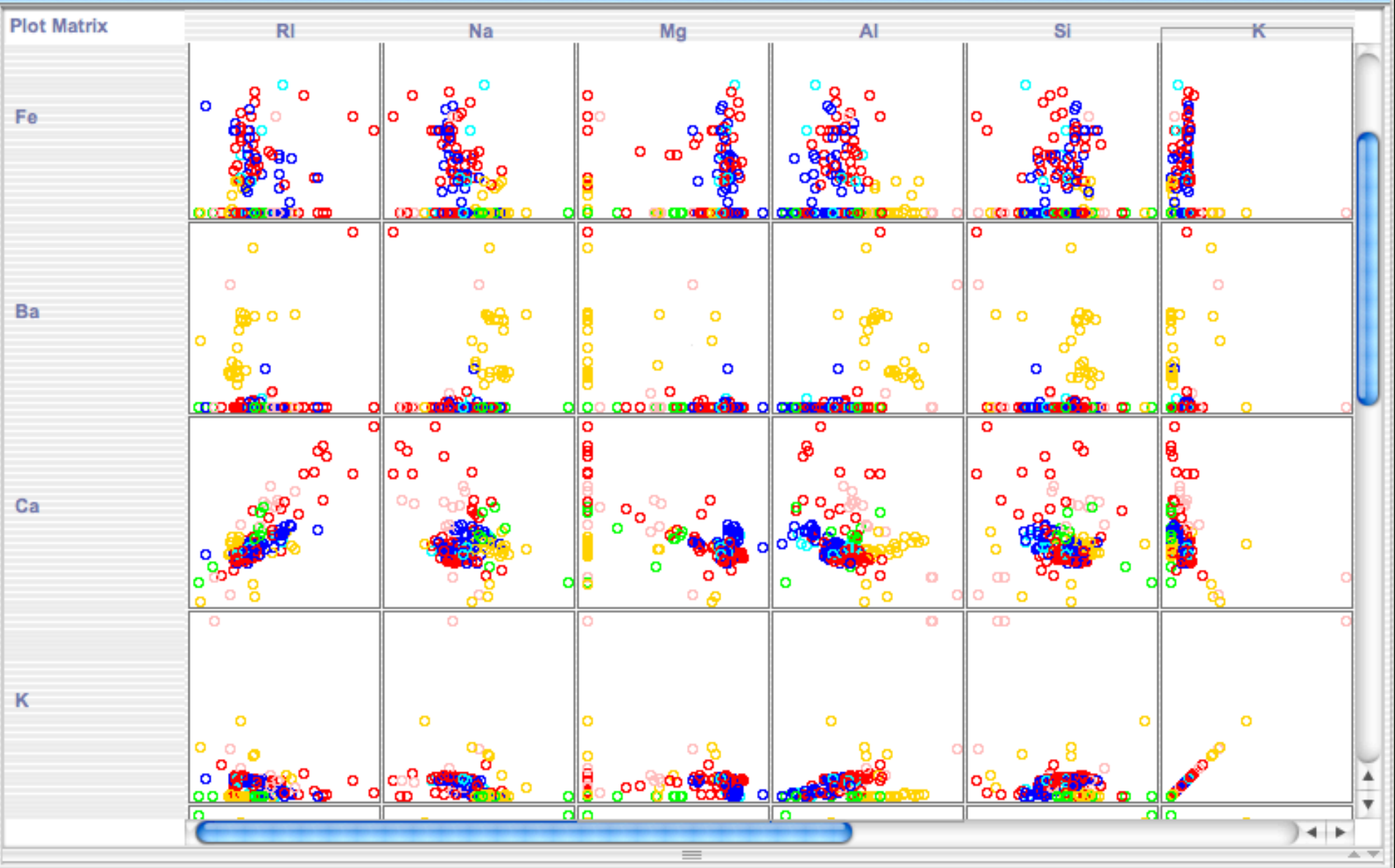
OK | Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Type
Fe
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: Type (Nom)
SubSample % :
100
Class Colour
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Type
Fe
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: Type (Nom)
SubSample % :
100
Class Colour
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
Status
OK
Log
x 0

Weka Knowledge Explorer

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



Status

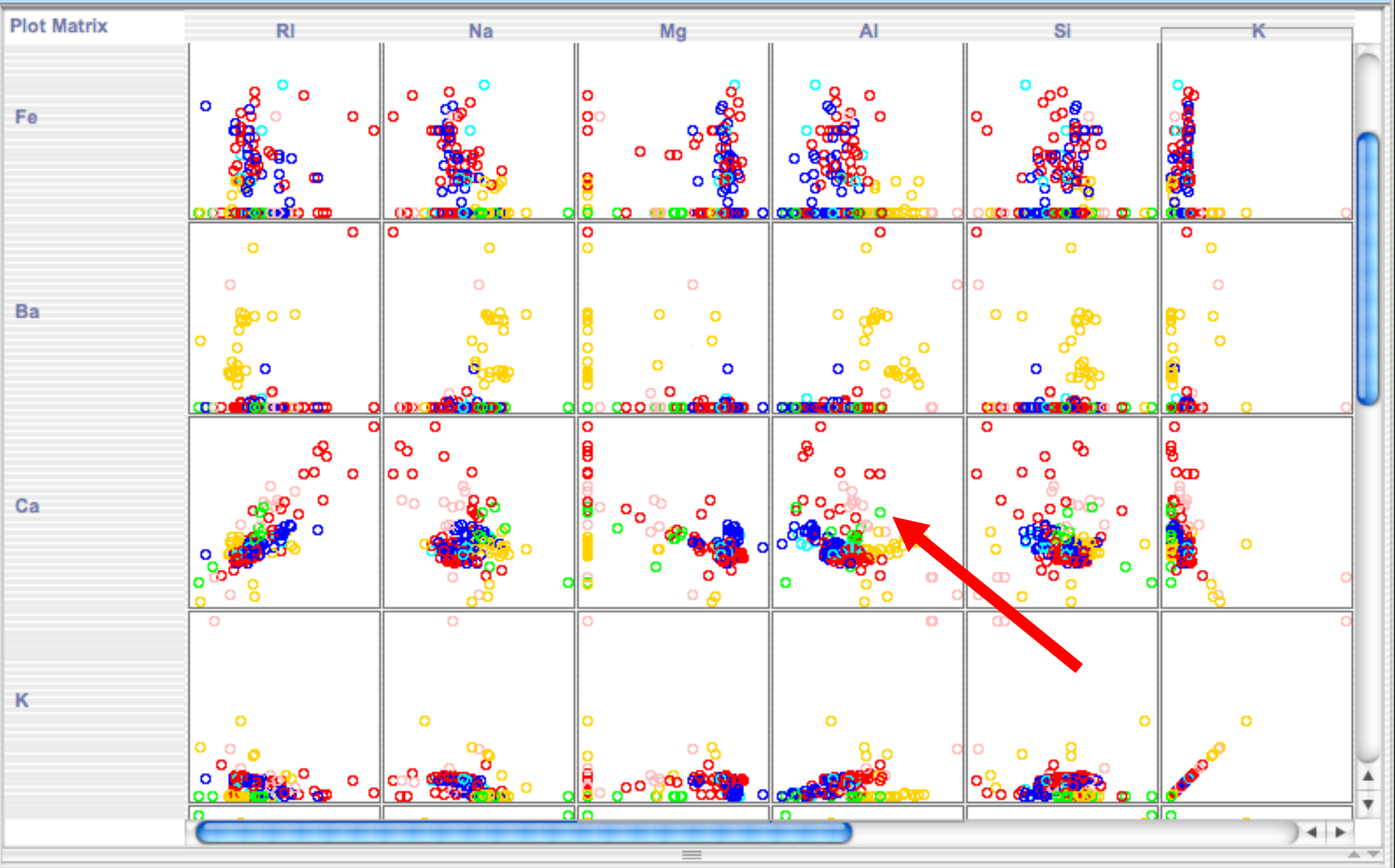
OK

Log

x 0

Weka Knowledge Explorer

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



Status

OK

Log

x 0

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

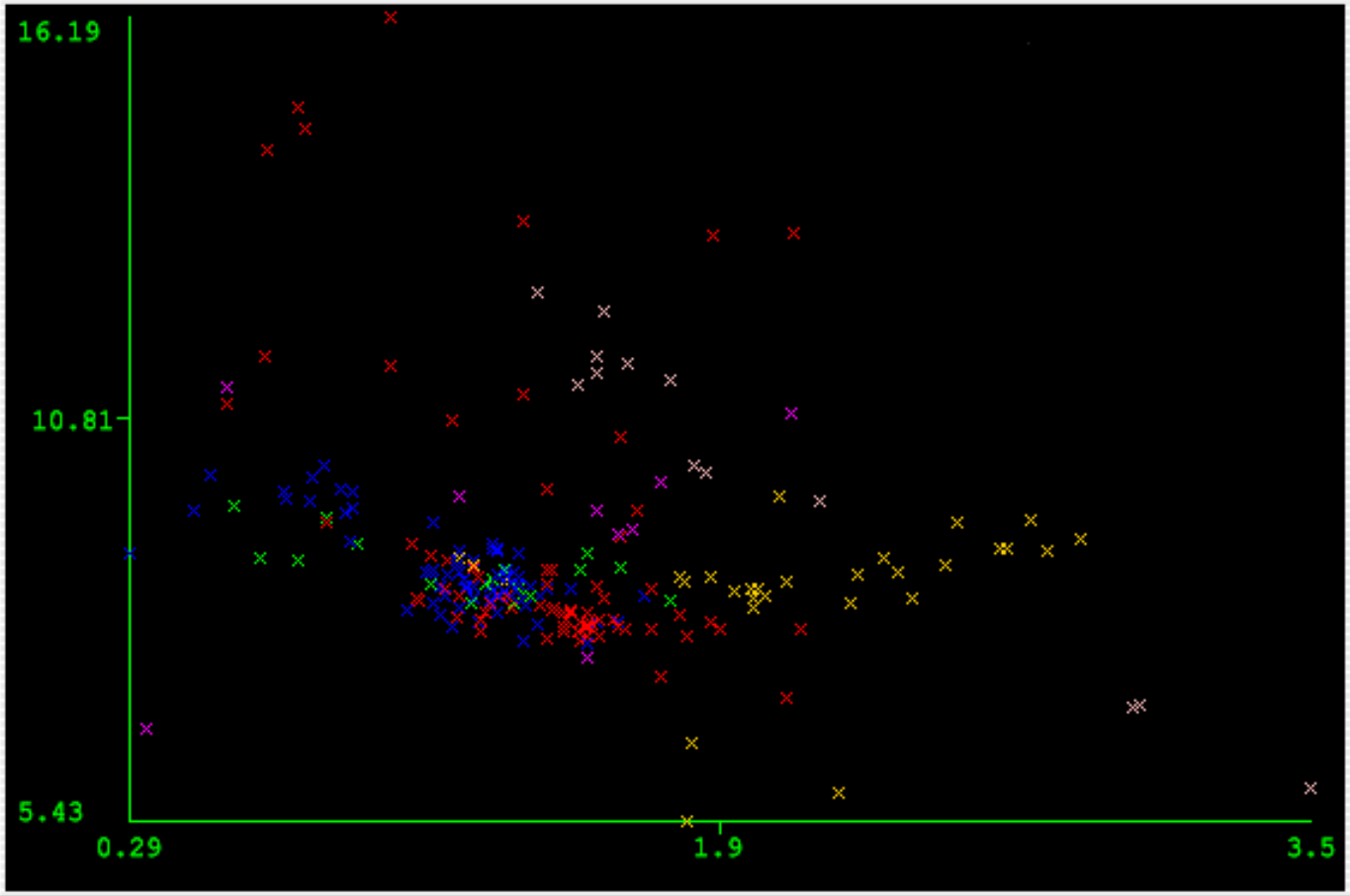
Colour: Type (Nom)

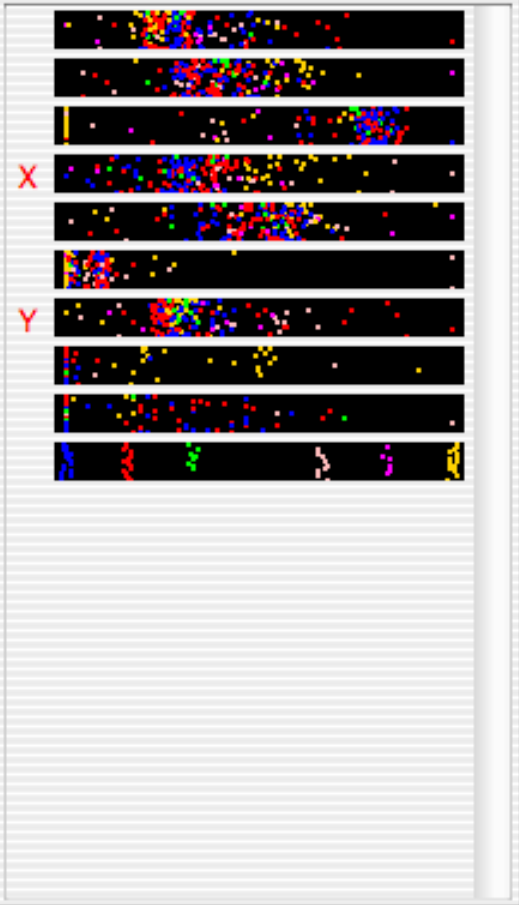
Select Instance

Reset | Clear | Save

Jitter

Plot: Glass





Class colour

build wind float

build wind non-float

vehic wind float

vehic wind non-float

containers

tableware

headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

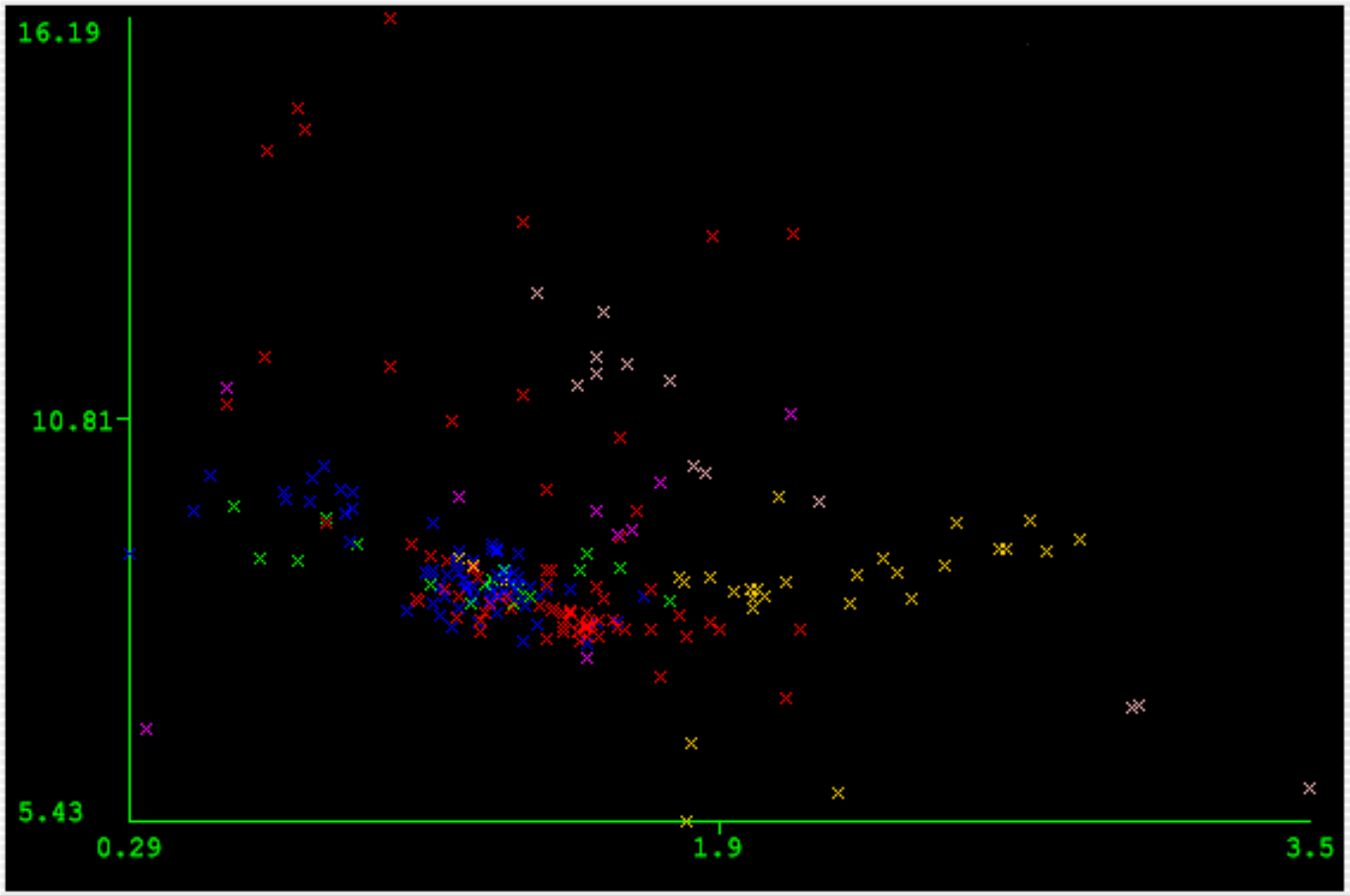
Colour: Type (Nom)

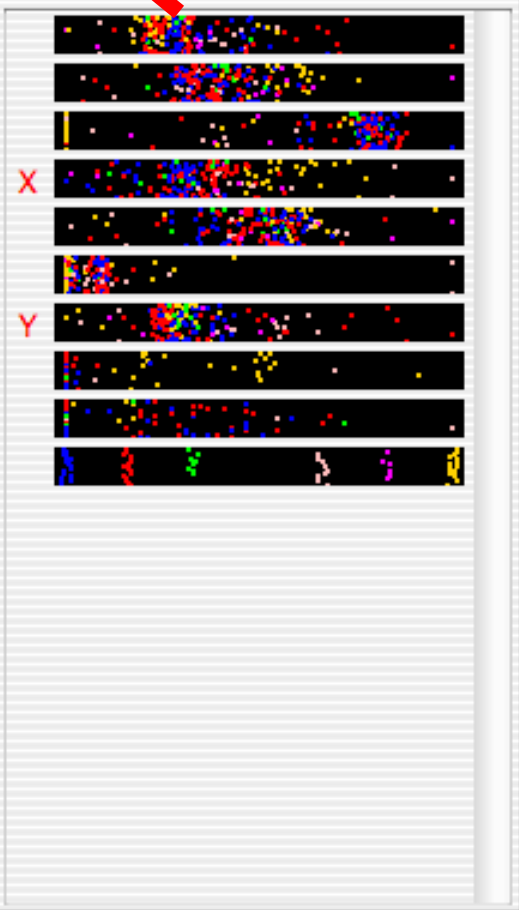
Select Instance

Reset | Clear | Save

Jitter

Plot: Glass





Class colour

build wind float

build wind non-float

vehic wind float

vehic wind non-float

containers

tableware

headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

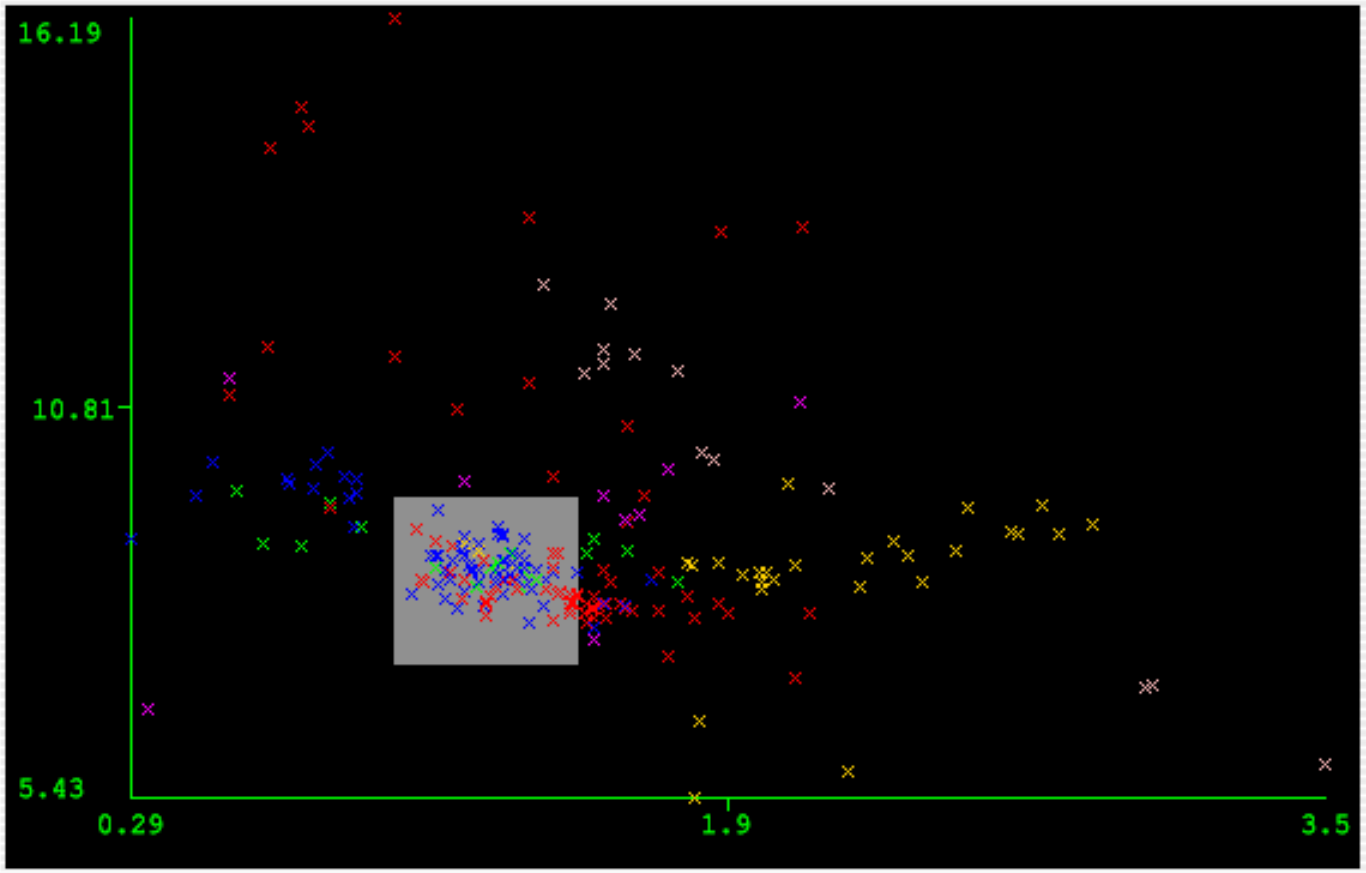
Colour: Type (Nom)

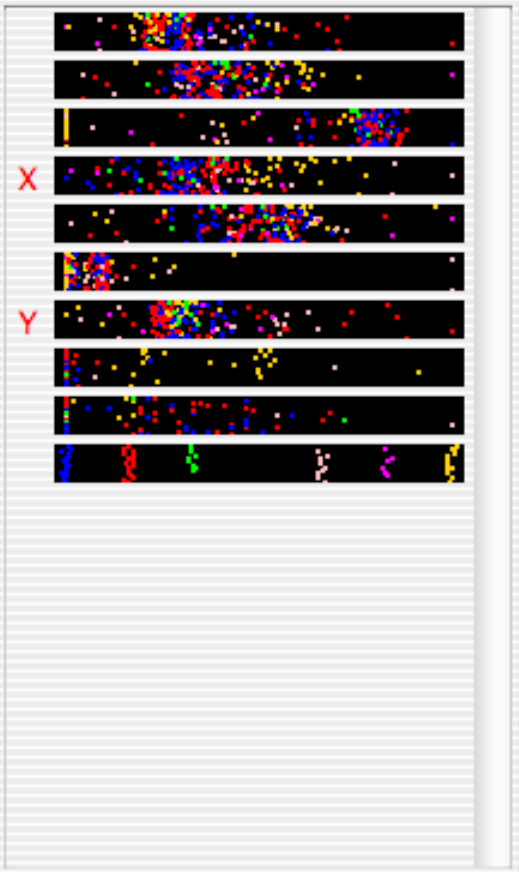
Rectangle

Submit | Clear | Save

Jitter

Plot: Glass





Class colour

build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

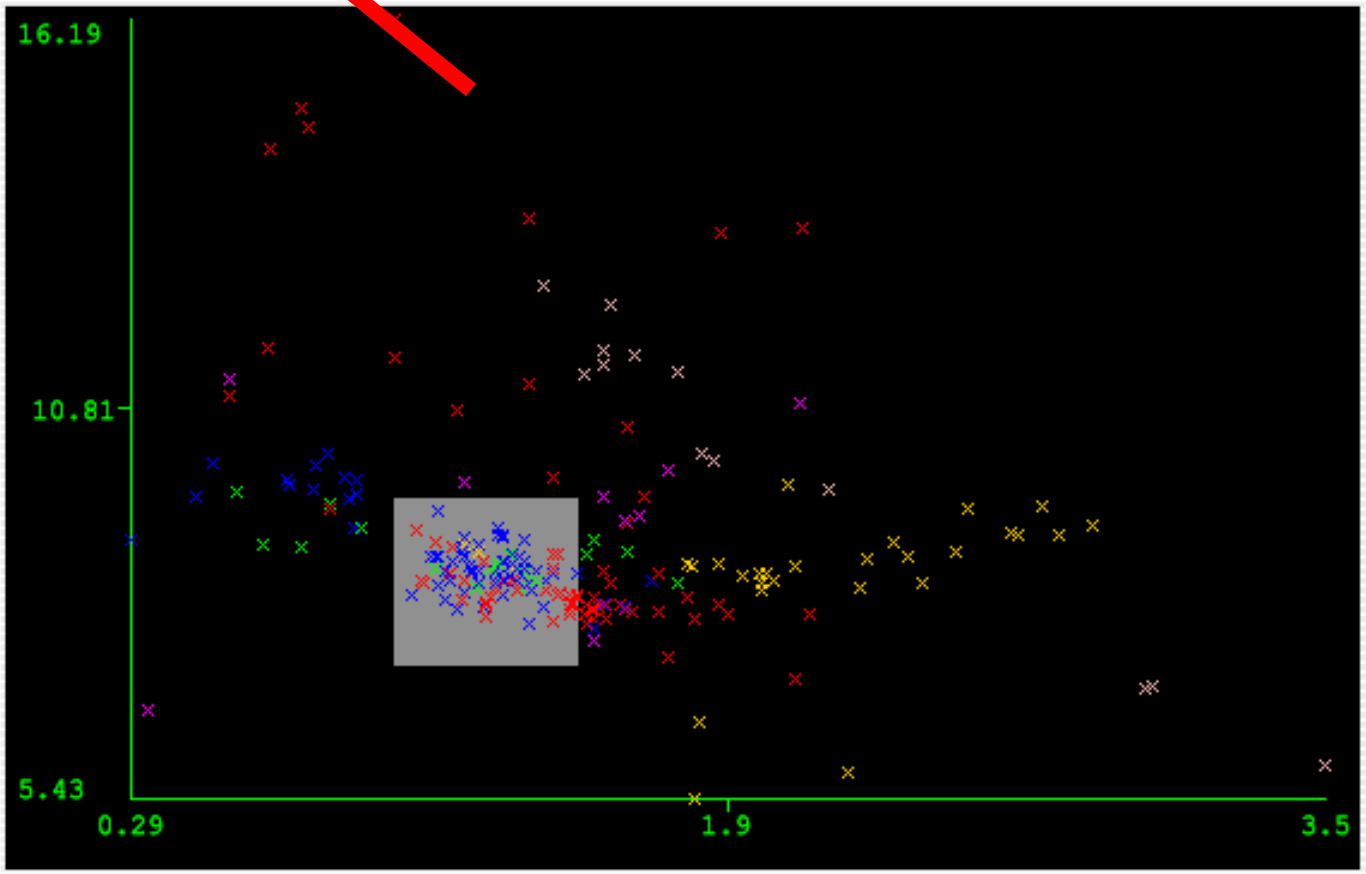
Colour: Type (Nom)

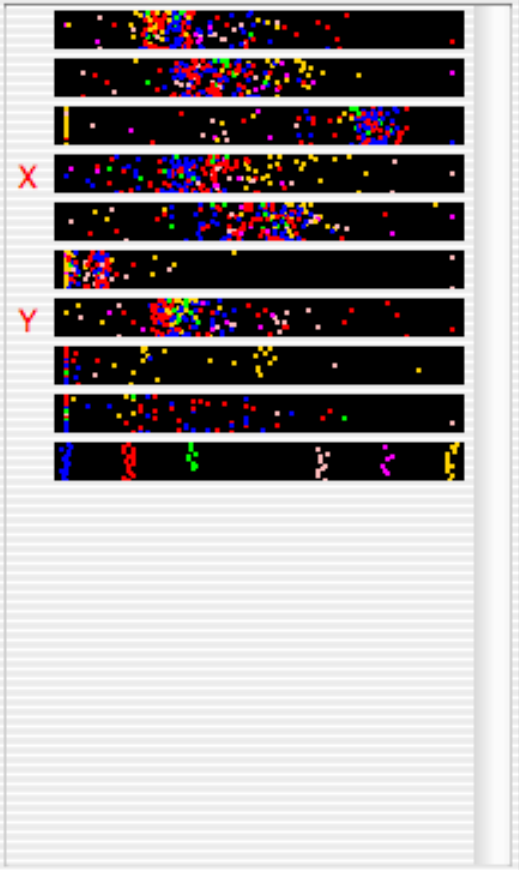
Rectangle

Submit | Clear | Save

Jitter

Plot: Glass





Class colour

build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps

Weka Knowledge Explorer: Visualizing Glass

X: Al (Num)

Y: Ca (Num)

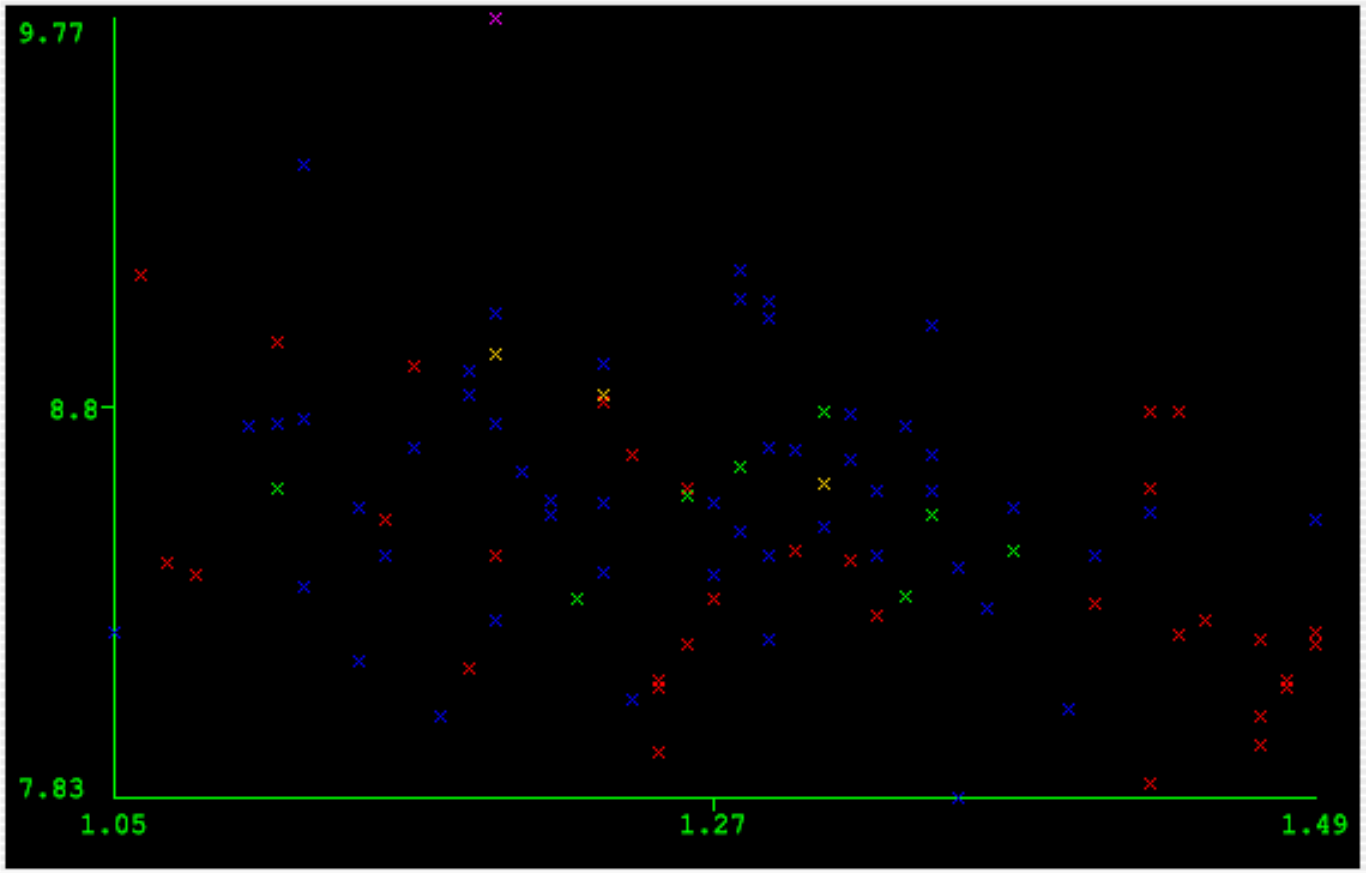
Colour: Type (Nom)

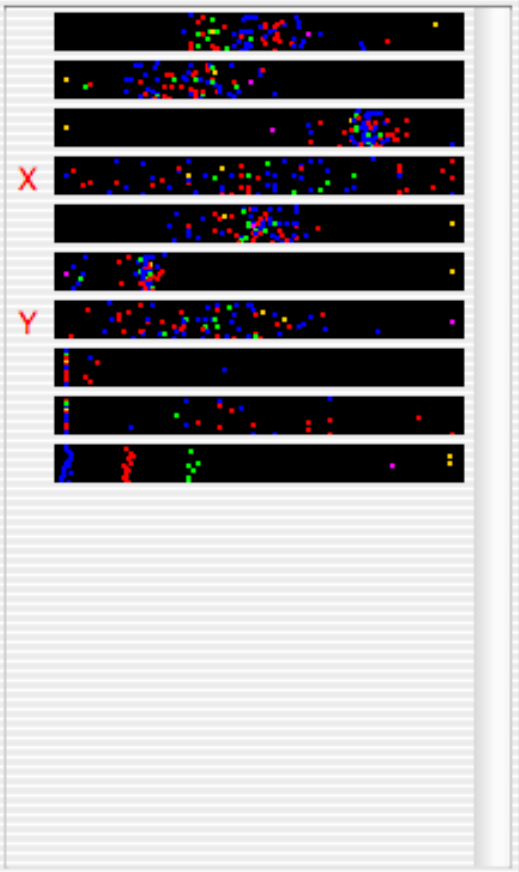
Rectangle

Reset | Clear | Save

Jitter

Plot: Glass





Class colour

build wind float
build wind non-float
vehic wind float
vehic wind non-float
containers
tableware
headlamps

Reference

- Ian H. Witten, Data Mining with Weka presentation, http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/
- Eibe Frank, Machine Learning with WEKA presentation,http://prdownloads.sourceforge.net/weka/weka.ppt