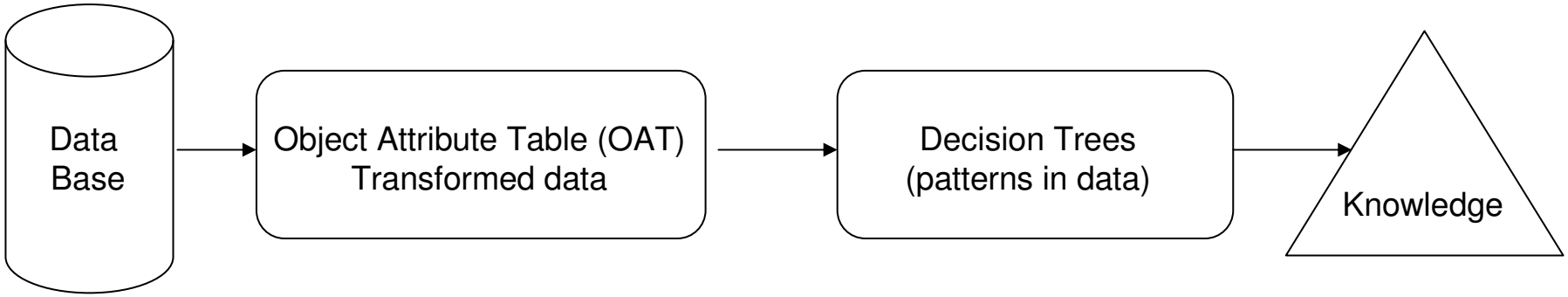
Implementation of Decision Trees using R

Margaret Miró-Julià,
Arnau Mir and Monica J. Ruiz-Miró
University of the Balearic Islands
Palma de Mallorca, SPAIN

Data vs. Knowledge

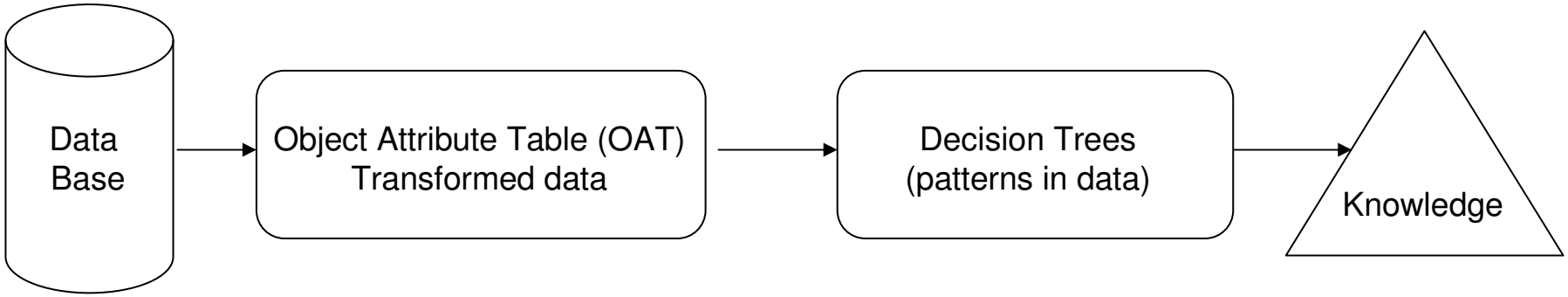
A large collection of unanalyzed facts from which conclusions may be drawn



The psychological result of perception and learning and reasoning
Confident understanding of the data together with the ability to use it for a specific purpose

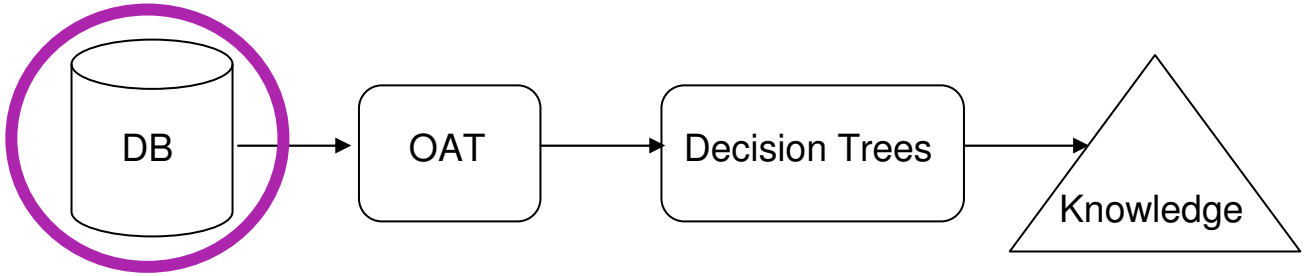
Implementation of Decision Trees using R

STATISTICS
The analyst states a question (supposition - intuition) explores the data and constructs a model.
The analyst proposes the model, which is validated



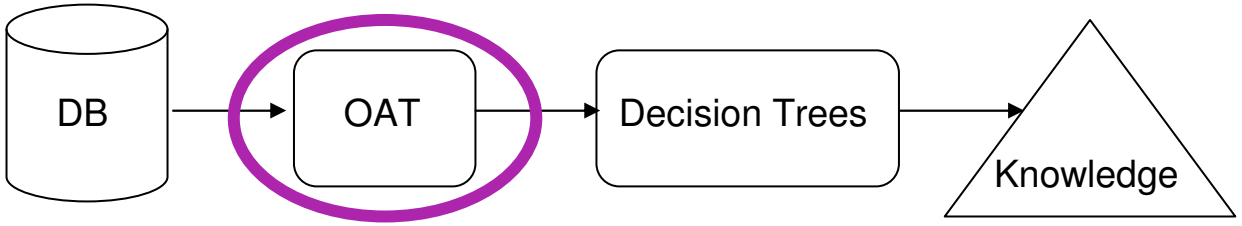
ARTIFICIAL INTELLIGENCE
The system generates models automatically by identifying patterns

Implementation of Decision Trees using R



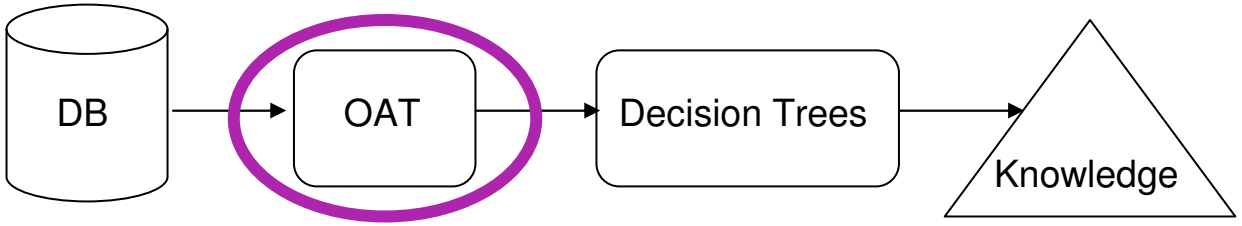
- Large amounts of data that must be structured
- Relational Database or table
  - Objects or rows
  - Attributes or columns

Implementation of Decision Trees using R



- An Object Attribute Table (OAT) is a structure that allows the description of a set of concepts in terms of a collection of objects described by the values of their attributes

Implementation of Decision Trees using R



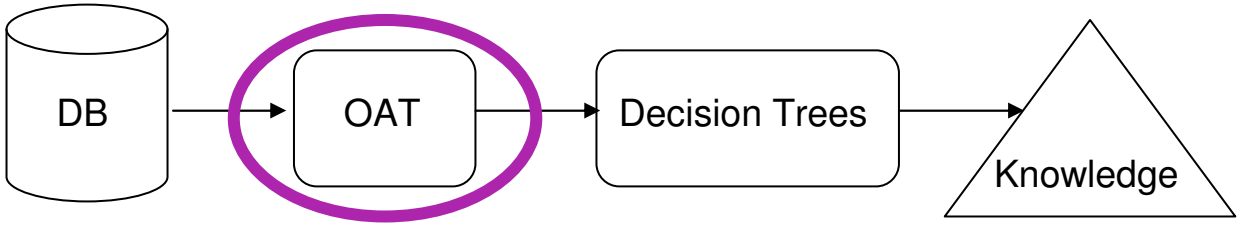
$C = \{c_x, c_y, \ldots, c_z\}$ set of concepts

$D = \{d_1, d_2, \ldots, d_m\}$ set of objects

$R = \{r_a, r_b, \ldots, r_g\}$ set of attributes

an Object Attribute Table (OAT) can describe a situation by means of the values of the attributes

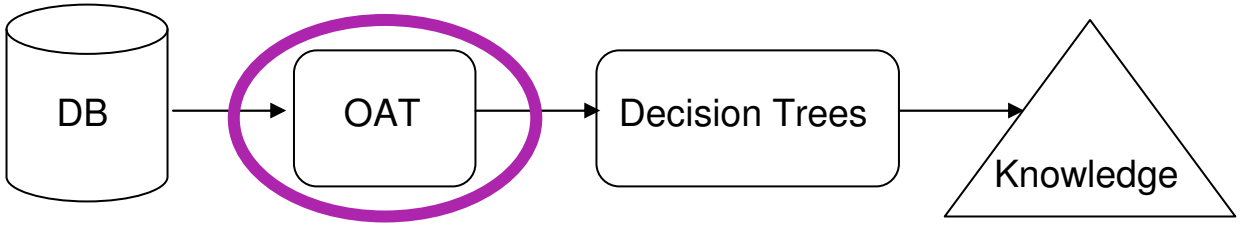
Implementation of Decision Trees using R



R

D		r_a	r_b	…	r_g	C
	d_1	a_1	b_1	…	g_1	c_y
	d_2	a_2	b_2	…	g_2	c_w
	d_3	a_3	b_3	…	g_3	c_x
	…	…	…	…	…	…
	d_m	a_m	b_m	…	g_m	c_w

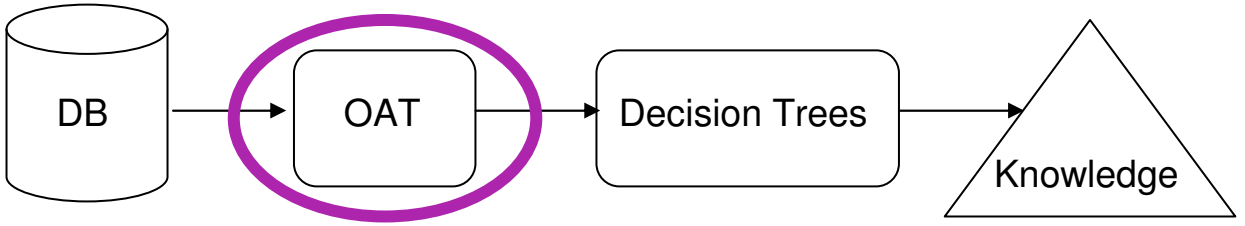
Implementation of Decision Trees using R



IMPORTANT FEATURES

- Type of data
  - Numerical: discrete or continuous
  - Categorical
- Number of objects and attributes
- Properties of the attributes: number of values, cost, frequency

Implementation of Decision Trees using R



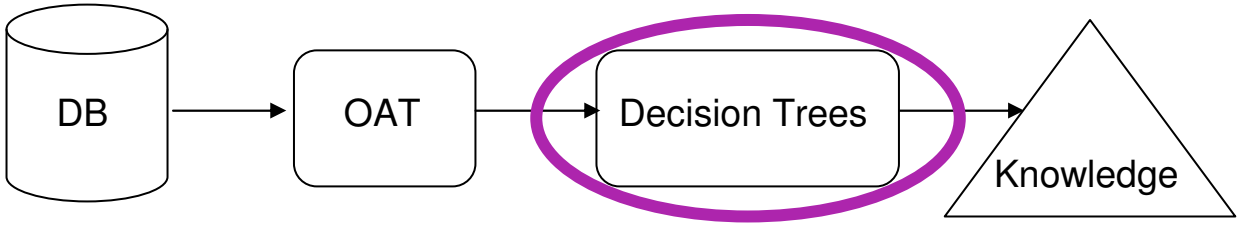
	r_1	r_2	r_3	C
d_1	0	0	1	c_1
d_2	1	0	1	c_3
d_3	0	0	0	c_1
d_4	0	1	1	c_3
d_5	1	1	1	c_2
d_6	0	1	0	c_3

Binary OAT

	r_1	r_2	r_3	C
d_1	0	0	a	c_1
d_2	1	0	b	c_3
d_3	0	0	a	c_1
d_4	0	1	c	c_3
d_5	1	1	c	c_2
d_6	0	1	b	c_3

Multivalued OAT

Implementation of Decision Trees using R



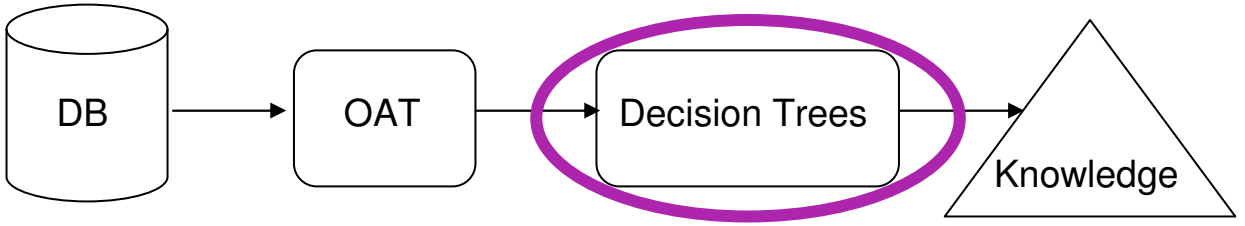
UIB-IK: knowledge acquisition tool to induce decision trees

- Binarization of the OAT
- Identification of the attribute basis: subsets of attributes that describe the concepts without contradiction

  (basis is formed by those attributes essential to the concept description)
- Generation of the tree (according to criteria)

Fiol-Roig, G. UIB-IK: A Computer System for Decision Trees Induction. LNCS **1609**, 601-611, 1999

Implementation of Decision Trees using R



Binarization

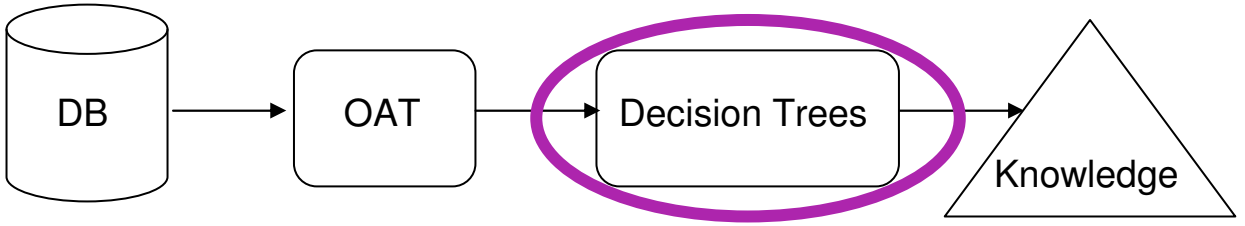
	r_1	r_2	C
d_1	1	a	1
d_2	1	b	0
d_3	2	a	0
d_4	3	c	1

	r_1^1	r_1^2	r_2^1	r_2^2	C
d_1	0	0	1	0	1
d_2	0	0	0	1	0
d_3	0	1	1	0	0
d_4	1	1	1	1	1

1 → 0 0 a → 1 0
2 → 0 1 b → 0 1
3 → 1 1 c → 1 1

Boolean algebra

Implementation of Decision Trees using R



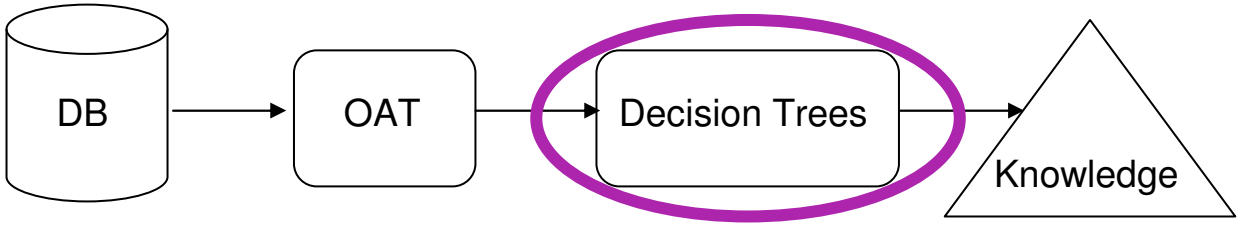
Attribute basis:

	r_1	r_2	C
d_1	3	a	1
d_2	1	b	0
d_3	2	a	0
d_4	3	c	1

$\{r_1\}$ is a basis

$\{r_1, r_2\}$ is a basis

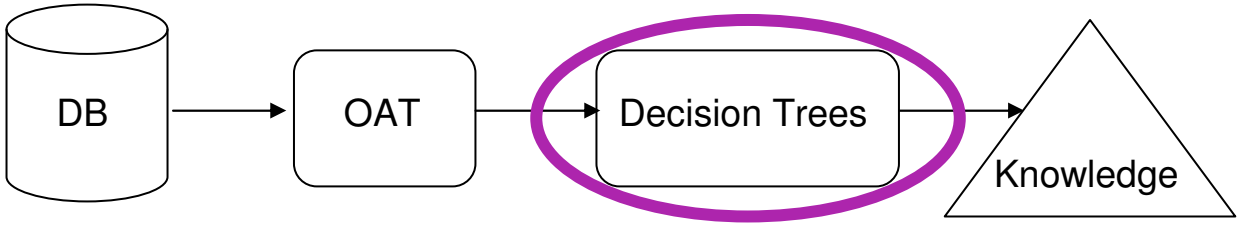
Implementation of Decision Trees using R



More than one basis, which one do we chose?

- Minimum cost, considering that each attribute of the OAT has an associated cost
- Minimum base, minimum number of attributes
- Fastest base, minimum number of questions

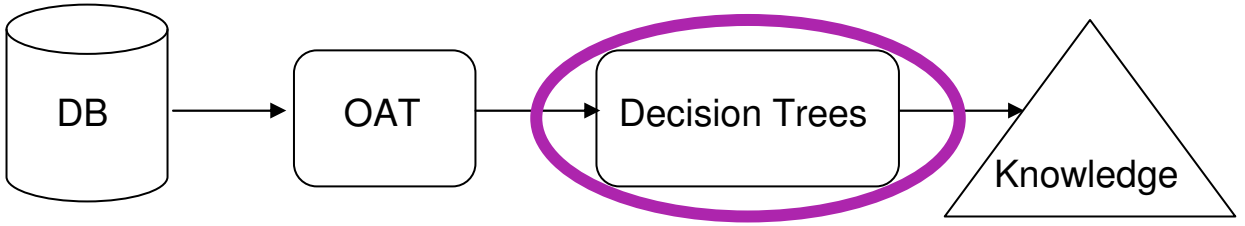
Implementation of Decision Trees using R



Decision tree: common knowledge structure where leaf nodes represent the concepts and branches represent conjunctions of features that lead to those concepts

UIB-IK generates decision trees depending on the basis selected

Implementation of Decision Trees using R



IMPROVEMENTS

- Multivalued algebra similar to the boolean algebra
- Problems in the implementation
- Discretization of the multivalued attributes in the OAT

Miró-Julià, M. and Fiol-Roig, G. An Algebra for the Treatment of Multivalued Information Systems. LNCS **2652**, 556-563, 2003

UIB

Implementation of Decision Trees using R

In order to carry out the improvements R was used

- To generated the discrete OAT, the range of attribute values was partitioned using R:
  - Intervals of the same size, subsets with the same number of attribute values
  - Intervals with the same relative frequency, subsets of attribute values that appear with the same frequency
  - Intervals with other statistical properties, subsets of attribute values with other statistical properties

R was easy to work with

Implementation of Decision Trees using R

R was also used to calculate the information gain due to attribute *K* in a recursive manner

$$Gain(K) = I(OAT) - E(K)$$

$$I(OAT) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad E(K) = p_i \times I(OAT_K)$$

Implementation of Decision Trees using R

Finally, subtables (nodes) were generated recursively with R as follows:

- Calculate information gain of the table
- Find attribute M that maximizes information gain (put in first column)
- Generate subtables, by grouping rows with same attribute values for M, eliminate M

Summary

- R makes the generation of the discrete OAT simple and easygoing
- The discretization is similar for numerical or categorical values of the attribute
- R allows for the generation of subtables in a recursive manner
- The results obtained encourage us to continue using R in Artificial Intelligence

Implementation of Decision Trees using R

I would like to thank

- Arnau and Ricardo for pointing out R's marvelous features and steering me in the right direction
- Monica for teaching me how to use R

Literature

- Fiol-Roig, G. UIB-IK: A Computer System for Decision Trees Induction. LNCS **1609**, 601-611, 1999.
- Miró-Julià, M. and Fiol-Roig, G. An Algebra for the Treatment of Multivalued Information Systems. LNCS **2652**, 556-563, 2003.
- Fiol-Roig, G. Learning from Incompletely Specified Object Attribute Tables with Continuous Attributes. Frontiers in Artificial Intelligence and Applications **113**, 145-152, 2004.