Usages of Generalization in CBR

Eva Armengol

eva@iiia.csic.es

Artificial Intelligence Institute (IIIA)

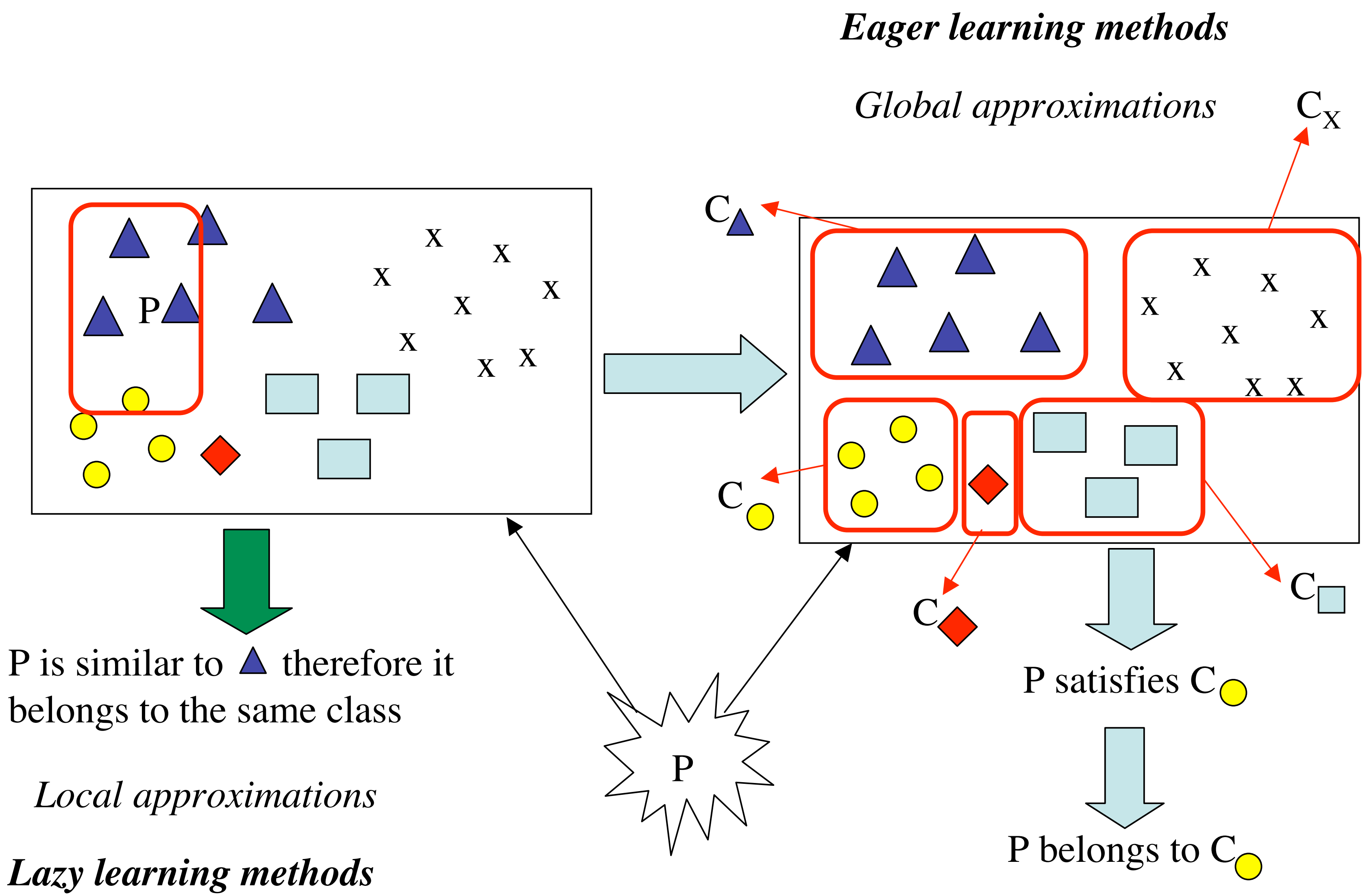
Spanish Research Council (CSIC)

Barcelona

Outline

- Eager vs Lazy learning methods for classification
- Precedents
- Lazy Induction of Descriptions (LID)
- Usages of generalization

Eager vs Lazy learning methods for Classification



Eager vs Lazy Learning Methods for Classification

- Eager learning methods
  - Global approximations of concepts (generalizations)
  - Global models are used for classifying new problems
- Lazy learning methods
  - Local approximations of concepts
  - Generalizations are not used for classifying new problems

How generalizations could be used in a lazy method?

Precedents

- PROTOS (Bareiss and Porter, 1987)
  - Generalizations are used to define categories of cases
  - Each category is represented by an *exemplar*
  - New problems are compared with the exemplars
- Generalized cases (Bergmann & Stahl, 1998)
  - A case represent a part of the solution space
  - Cases are clustered according to the solution space
    - Point case, Constant/Functional solution generalized case, Dependent/Independent alternative solution generalized case
- INRECA Project (1992-1995) (Manago, Bergmann, et al)
  - Combines decision trees with CBR

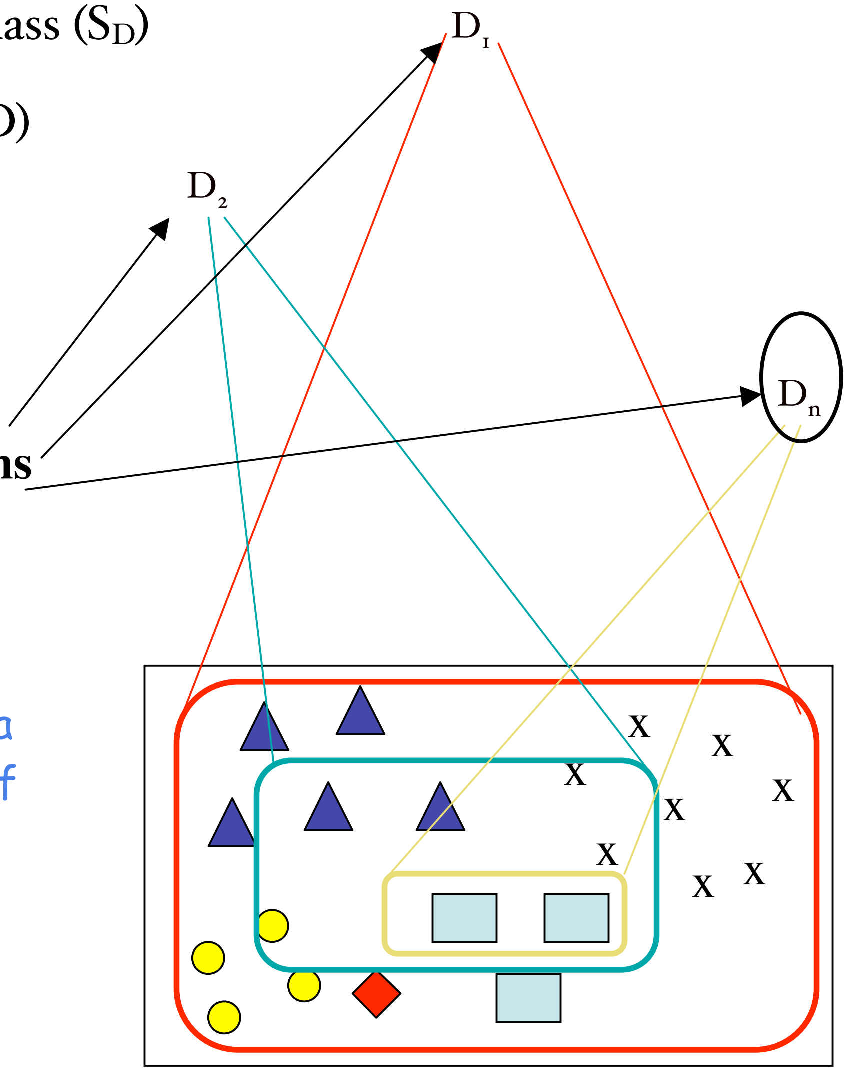
Lazy Induction of Descriptions (Armengol and Plaza, 2001)

- Lazy learning method
- Useful for classification tasks
- LID handles objects represented as feature terms
- LID builds a generalization that can be interpreted as representative of a set of cases

Lazy Induction of Descriptions (LID)

Function LID (p, D, S_D, C)
  if stopping-condition (S_D) then return class (S_D)
    else f_d := Select-leaf (p, S_D, C)
      D' := Add-path (π(root (p), f_d), D)
      $S_{D'}$:= $\boxed{\text{Discriminatory-set}}$ (D, S_D)
      LID (p, D', $S_{D'}$, C)
  end-if
end-function

Similitude terms



The similitude term of a set of cases is a generalization formed by the features of p assessed as relevant for classifying p

LID

- LID is a lazy learning method useful for classification tasks
- Given a new problem *p* the outcome of LID is
  - A classification for *p*
  - A similitude term that contains the features that have been assessed as the most relevant for classifying *p*

The similitude term (a generalization) is not used for solving new problems

Usages of the generalization

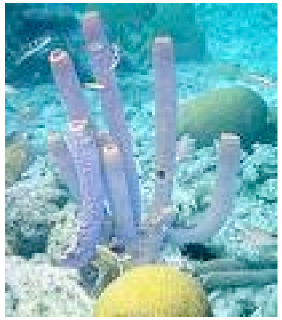
- Generalization as symbolic similarity
- For building partial domain models
- Generalization as explanation

Generalization as Symbolic Similarity

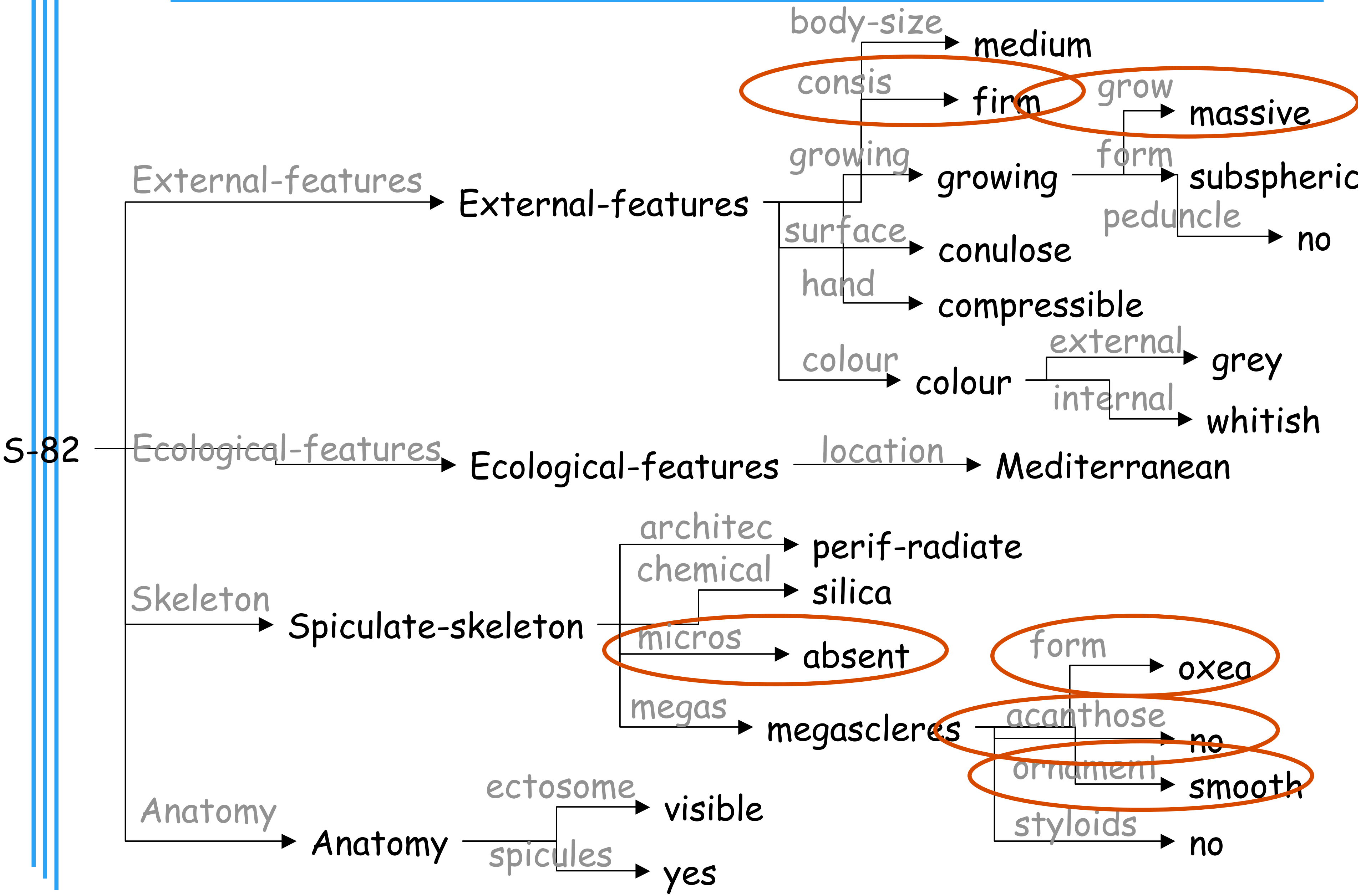
Cases/features	a_1	a_2	a_3	a_4	classes
c_1	0	1	0	0	
c_2	1	1	0	0	
c_3	0	1	0	1	
c_4	0	1	1	0	
p	1	1	1	0	

important features	similitude terms	
a_2	$a_2 = 1$	p is similar to c_1, c_2, c_3 and c_4
a_4	$a_2 = 1$ and $a_4 = 0$	p is similar to c_1, c_2 and c_4
a_3	$a_2 = 1$ and $a_4 = 0$ and $a_3 = 1$	p is similar to c_4

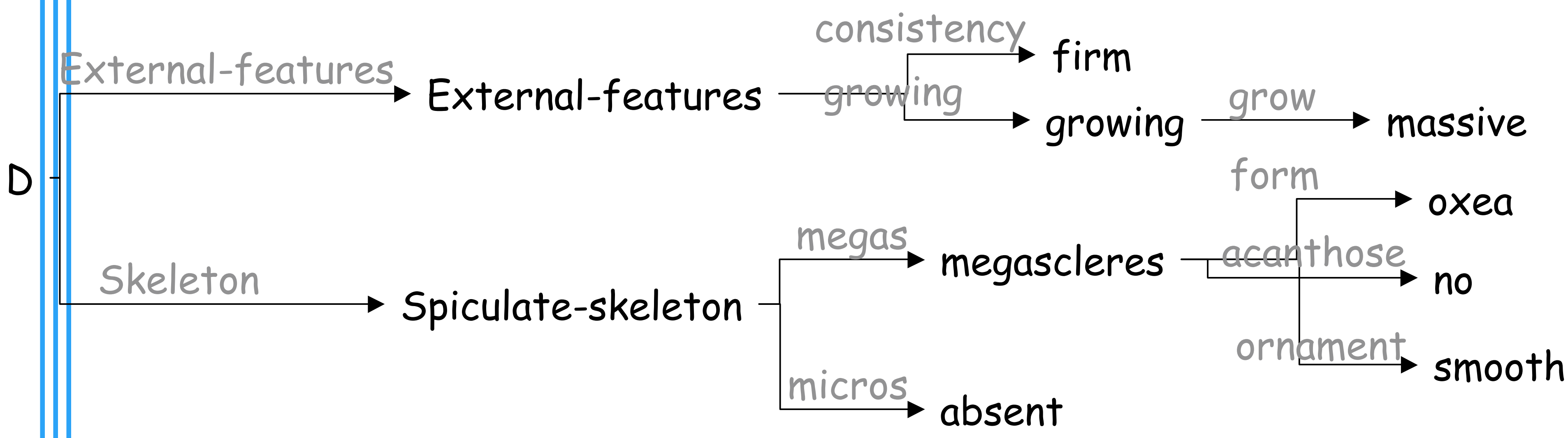
Classification of Marine Sponges



Example (I) : Marine Sponge Classification



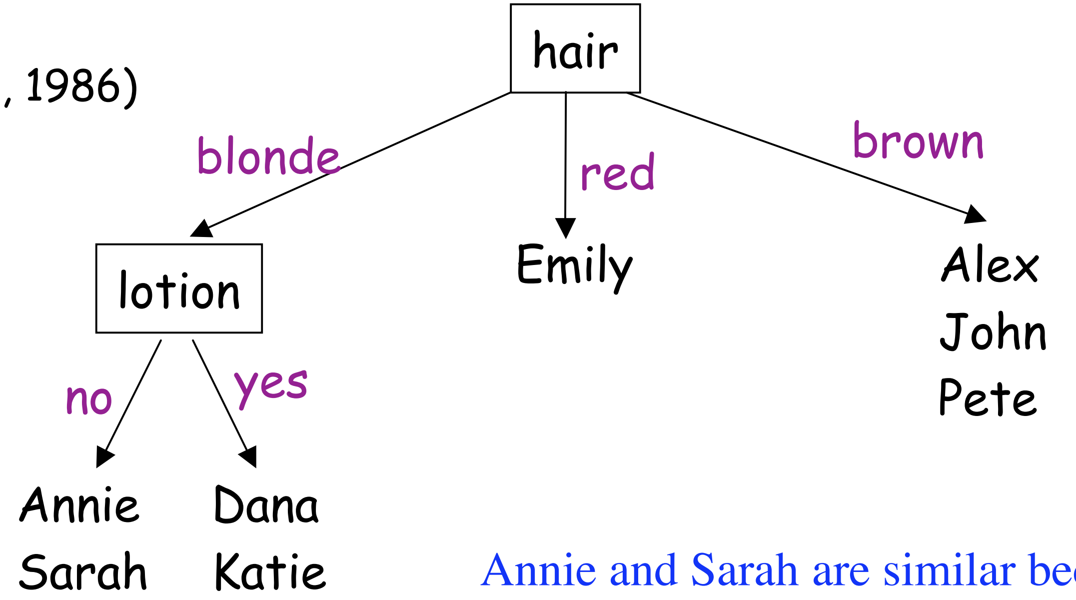
Example (II) :Marine Sponge Classification



There are 30 precedents in the case base that share with *sponge-82* the features in the similitude term. All them belong tot the *Astrophoridae* order

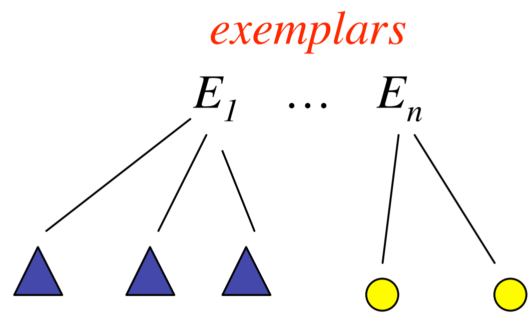
Generalization as Symbolic Similarity

ID3 (Quinlan, 1986)



Annie and Sarah are similar because both are blonde and they do not use lotion

PROTOS (Kolodner, 1993)



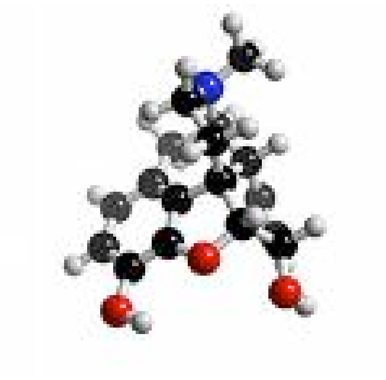
Summary : Generalization as symbolic similarity

- Generalizations can be interpreted as symbolic similarities because they contain aspects that are shared by a subset of examples of a class

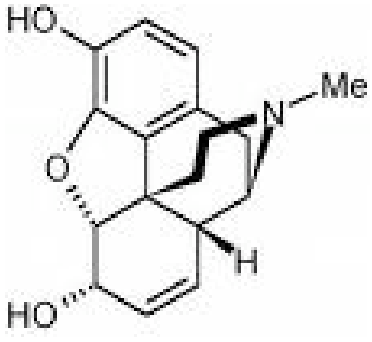
Usages of the generalization

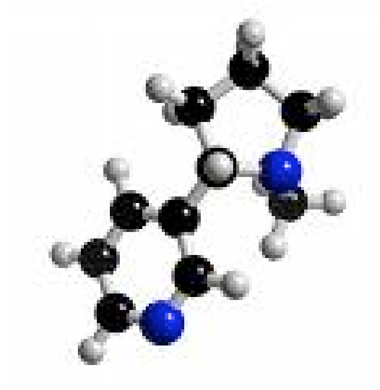
- Generalization as symbolic similarity
- For building partial domain models
- Generalization as explanation

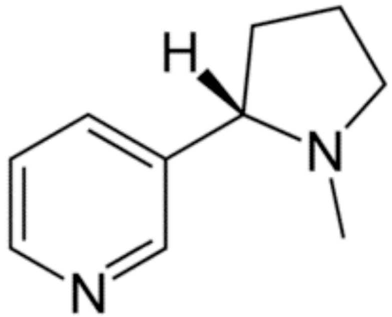
Predictive Toxicology



DDT







nicotine

Example: Predictive Toxicology

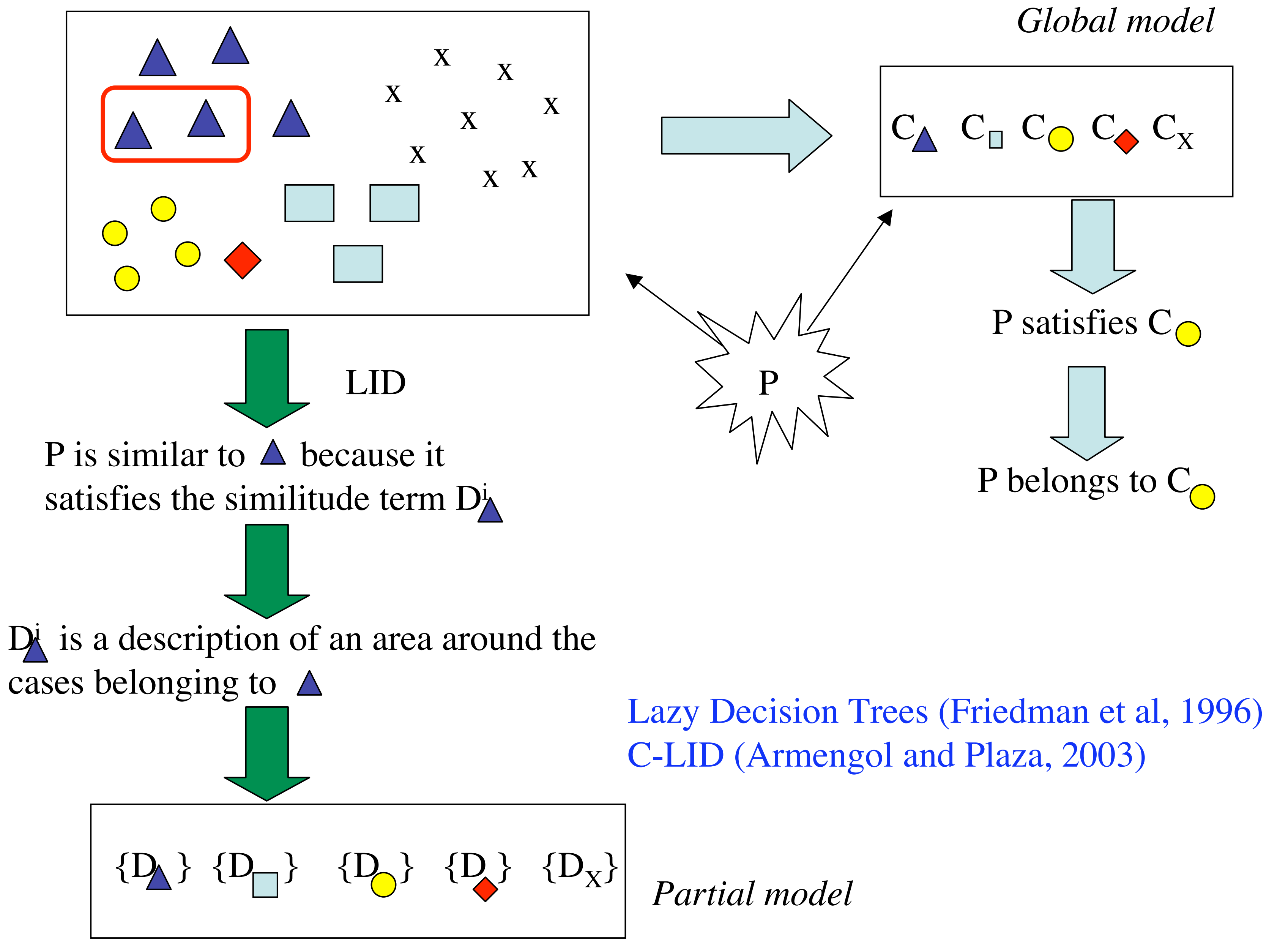
- The problem
  - New chemical compounds have to be deeply tested before introduce them in the market
  - The goal is to determine the potential carcinogenesis of new compounds
  - There are standard protocols to establish when a chemical compound is carcinogenic
    - Short-term experiments (90 days)
    - Long-term experiments (2 years)
    - High cost
    - Sometimes experiments are inconclusive

- Use of computational methods (Predictive Toxicology Challenge, 2001)
    - To reduce the experimental time
    - To build a model of carcinogenesis

Our approach (Armengol and Plaza, 2004)

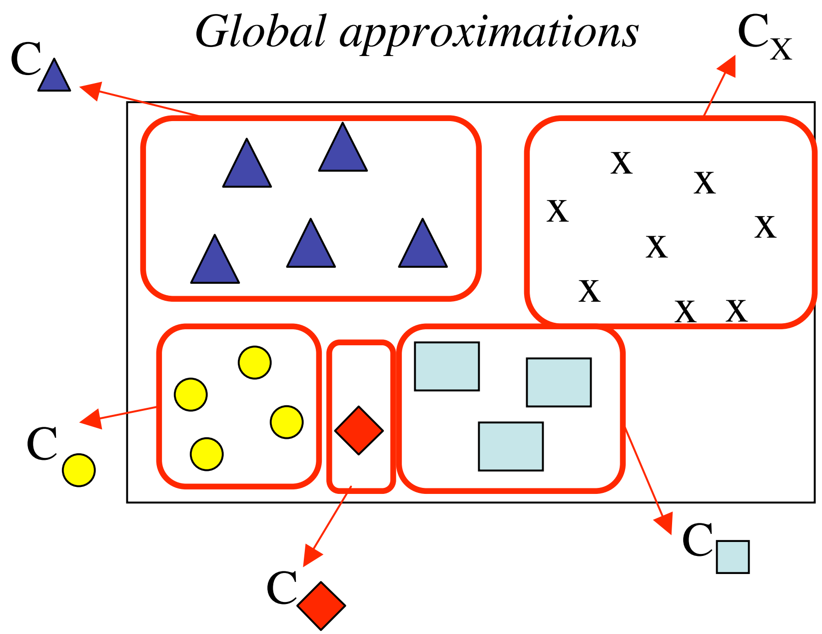
- To use lazy techniques for characterizing different classes of chemical compounds
  - LID and C-LID
- Why?
  - It is difficult to build a general description of the solution classes
  - Lazy techniques do not built intensional descriptions of the solution classes

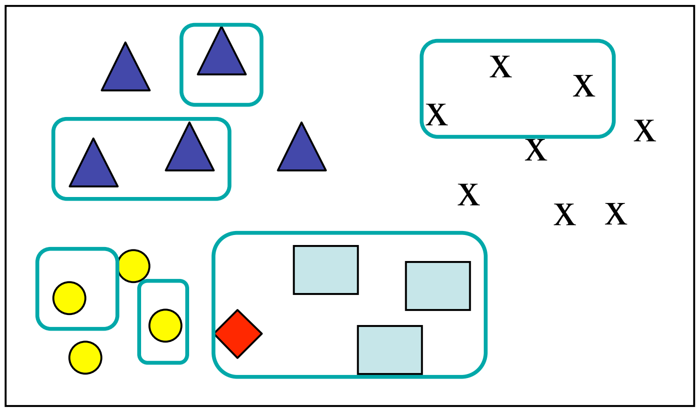
Generalization for building (partial) Domain Models



Lazy Decision Trees (Friedman et al, 1996)
C-LID (Armengol and Plaza, 2003)

Complete domain model
(from eager learning methods)





Partial domain model
(from lazy learning methods)

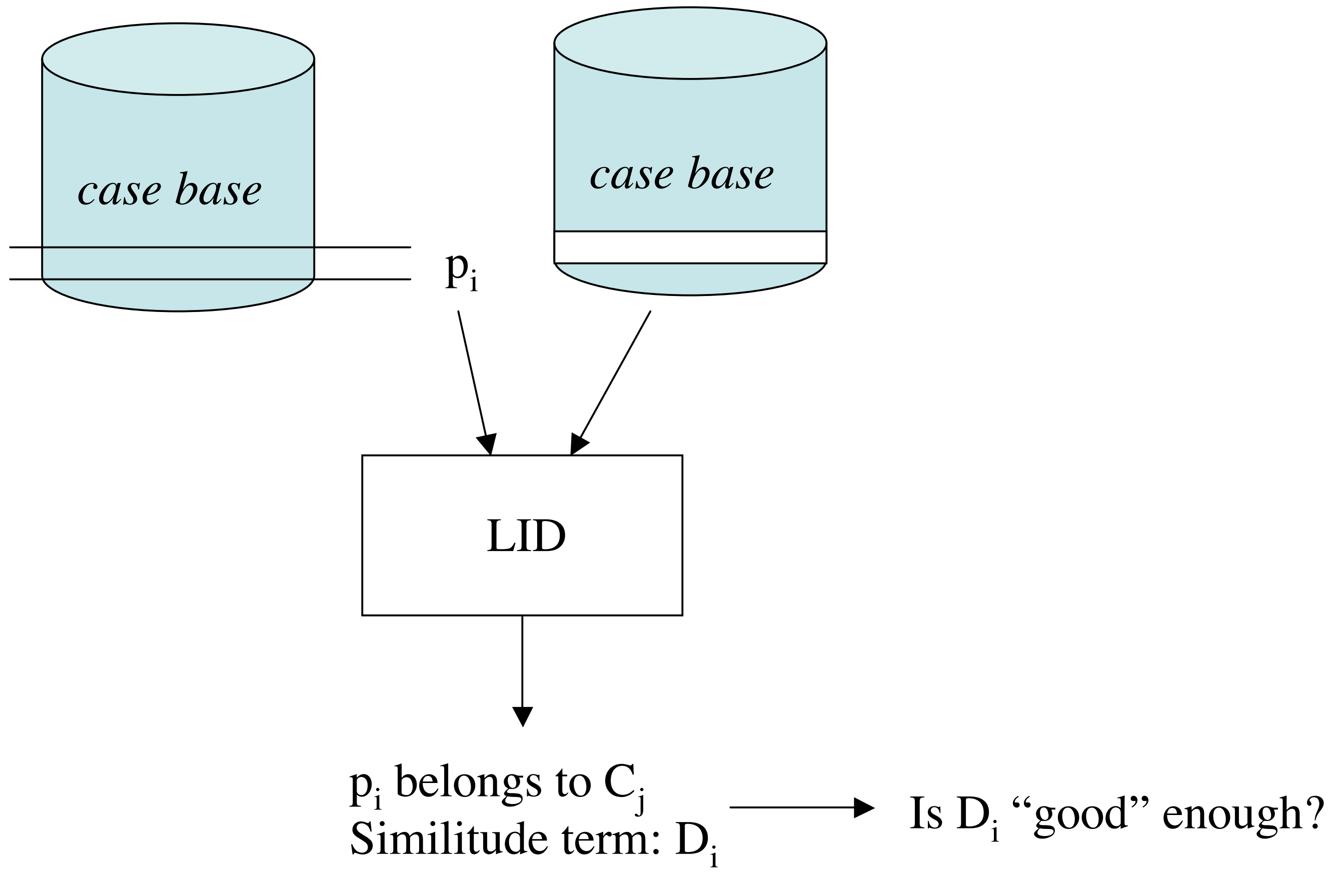
Using C-LID

- Goal: to use the similitude terms generated by LID for analyzing the compounds of the Toxicology dataset
  - Lazy process

    <u>Function</u> C-LID (p, D, S_D, C)
    <u>if</u> p satisfies some similitude term <u>then</u> <u>return</u> class
    <u>else</u> LID (p, D, S_D, C)
    <u>end-if</u>
    <u>end-function</u>

  - Eager process:
    - LID with leave-one-out method to generate similitude terms
    - To select a subset of similitude terms
    - Analyze the case-base using the selected similitude terms

The eager process of C-LID



- *compound*
  - main-group = *acyclic-unsaturated*
    - main-group = *butane*

2 positive examples
0 negative examples

- *compound*
  - p-radicals = *position-radical*
    - radicals = *compound*
      - main-group = *epoxyde*

4 positive examples
0 negative examples

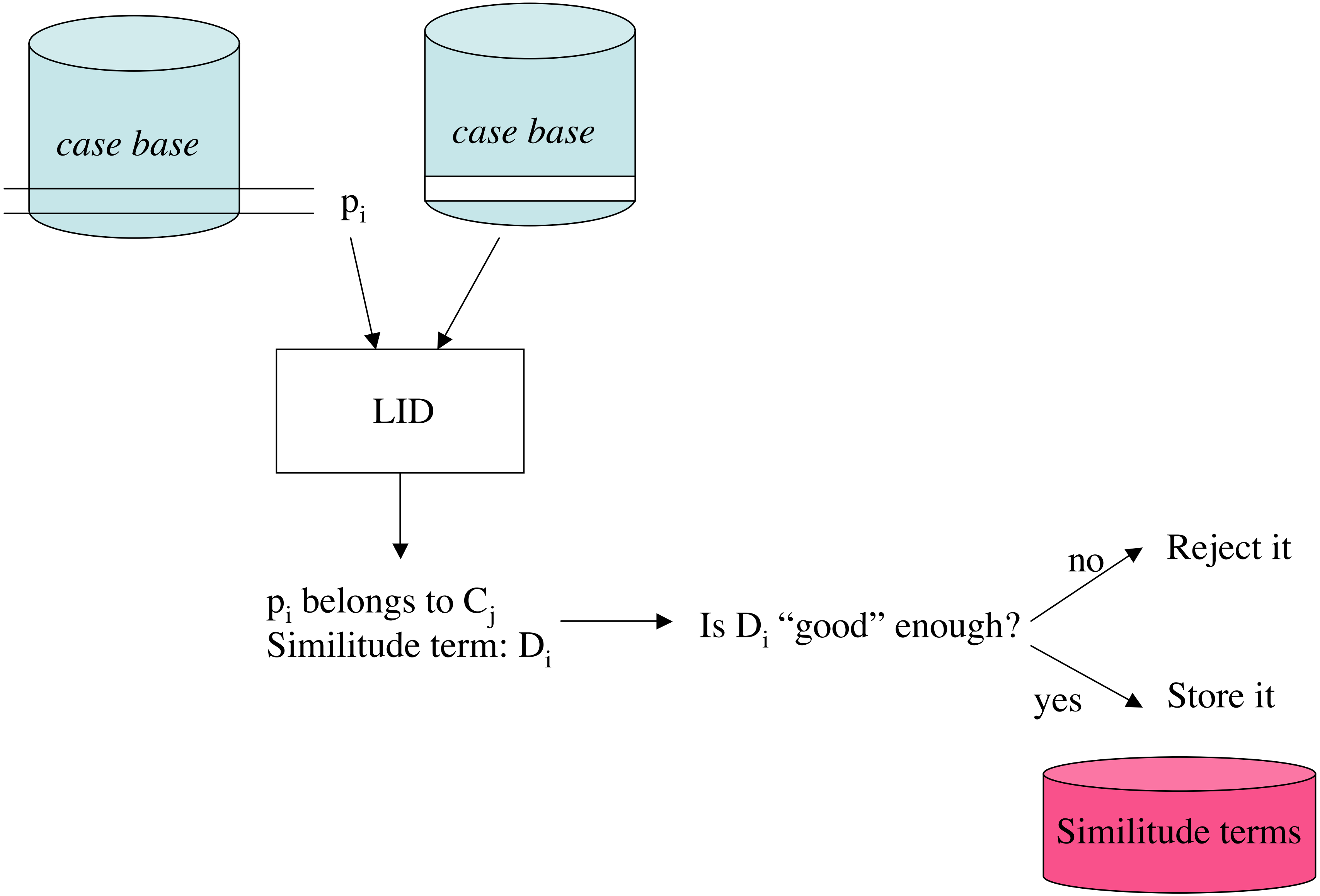
- *compound*
  - p-radicals = *position-radical*
    - radicals = *compound*
      - main-group = *amine*

15 positive examples
40 negative examples

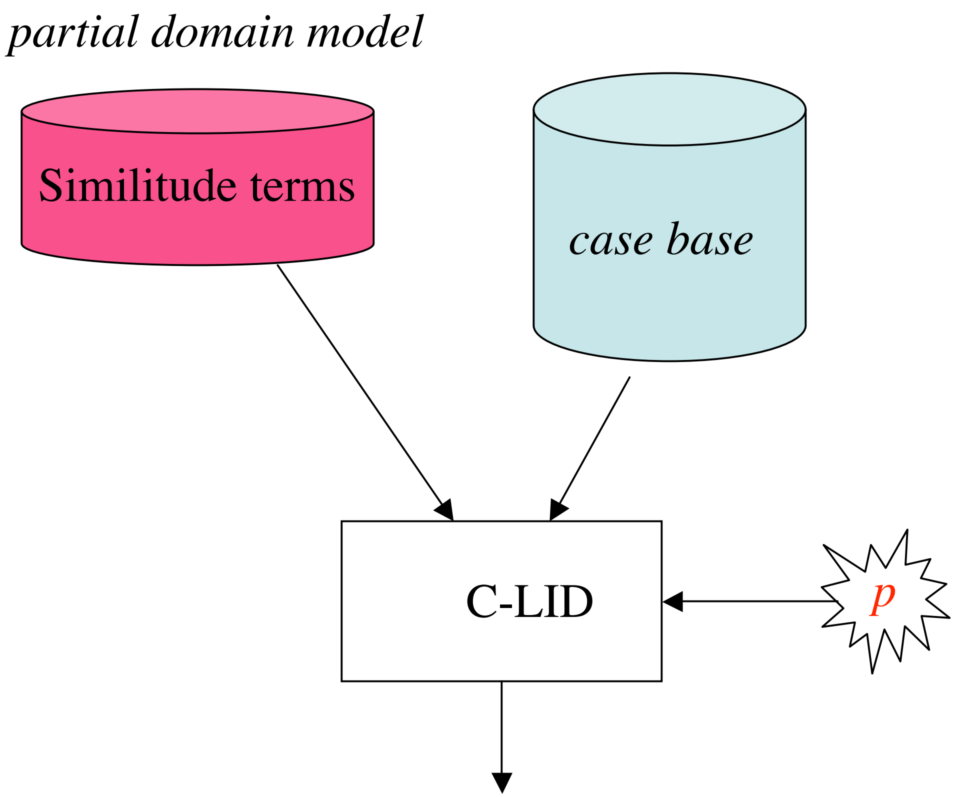
- *compound*
  - p-radicals = *position-radical*
    - radicals = *compound*
      - main-group = *bromine*

6 positive examples
2 negative examples

The eager process of C-LID



The eager process of C-LID



- <u>if</u> exist similitude term satisfying p <u>then</u> use it <u>else</u> LID
- use LID. If the solution has not enough support then use similitude terms
- use both similitude terms and LID

Summary: building partial domain models

- Eager learning methods produce complete domain models in the sense that class descriptions satisfy all known examples
  - In complex domains these descriptions could be too generals

- Using lazy learning methods we can obtain partial domain models since class descriptions are satisfied by a subset of examples of each class
  - In complex domains these descriptions could not be discriminatory

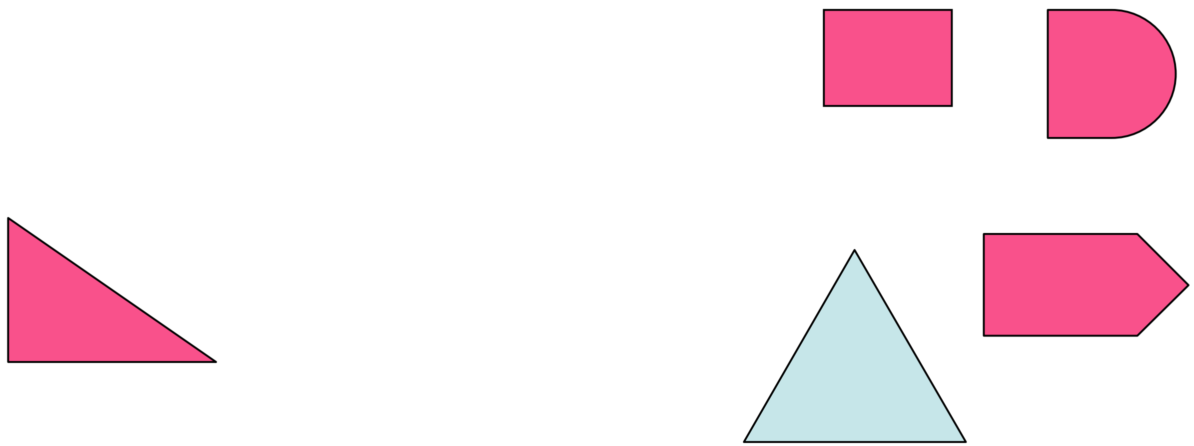
Usages of the generalization

- Generalization as symbolic similarity
- For building partial domain models
- Generalization as explanation

What to explain in lazy learning methods?

- To explain the Retrieve
  - Based on similarities

- To explain the Reuse
  - Based on similarities among the problem and the cases of each class
  - The user can easily understand the differences among the cases of each class



State of the art

- Given a problem, a CBR system retrieves the most similar case
  - Cases have a complicated structure (Doyle et al, 2003; McSherry, 2005)
- For some domains (e.g. Medicine) experts understand better a description of the differences between the problem and the retrieved cases
- The more explanatory cases are those close to the frontiers of classes (Doyle et al, 2004)
- An explanation should make explicit the contribution of each feature value to the classification of the problem (Nugent and Cunningham, 2005)
- Both the similarities and differences between problem and retrieved cases are useful for CBR explanations (McSherry, 2005)

Generalization as Explanation (Armengol and Plaza, 2004)

- Our approach: usage of symbolic similarities to explain the classification

1) Explanation of Retrieve
   - ✓ A symbolic description consisting of all that is shared by the problem and the retrieved cases

2) Explanation of Reuse
   - ✓ Cases are organized according to the class
   - ✓ A symbolic description for each class
   - ✓ Each symbolic description consists of all that is shared by the problem and the cases of a class

Generalization as Explanation (Armengol and Plaza, 2004)

- Our approach: usage of symbolic similarities to explain the classification

1) Explanation of Retrieve
   - ✓ A symbolic description consisting of all that is shared by the problem and the retrieved cases

2) Explanation of Reuse
   - ✓ Cases are organized according to the class
   - ✓ A symbolic description for each class
   - ✓ Each symbolic description consists of all that is shared by the problem and the cases of a class

- Similitude terms of LID
- Anti-unification concept

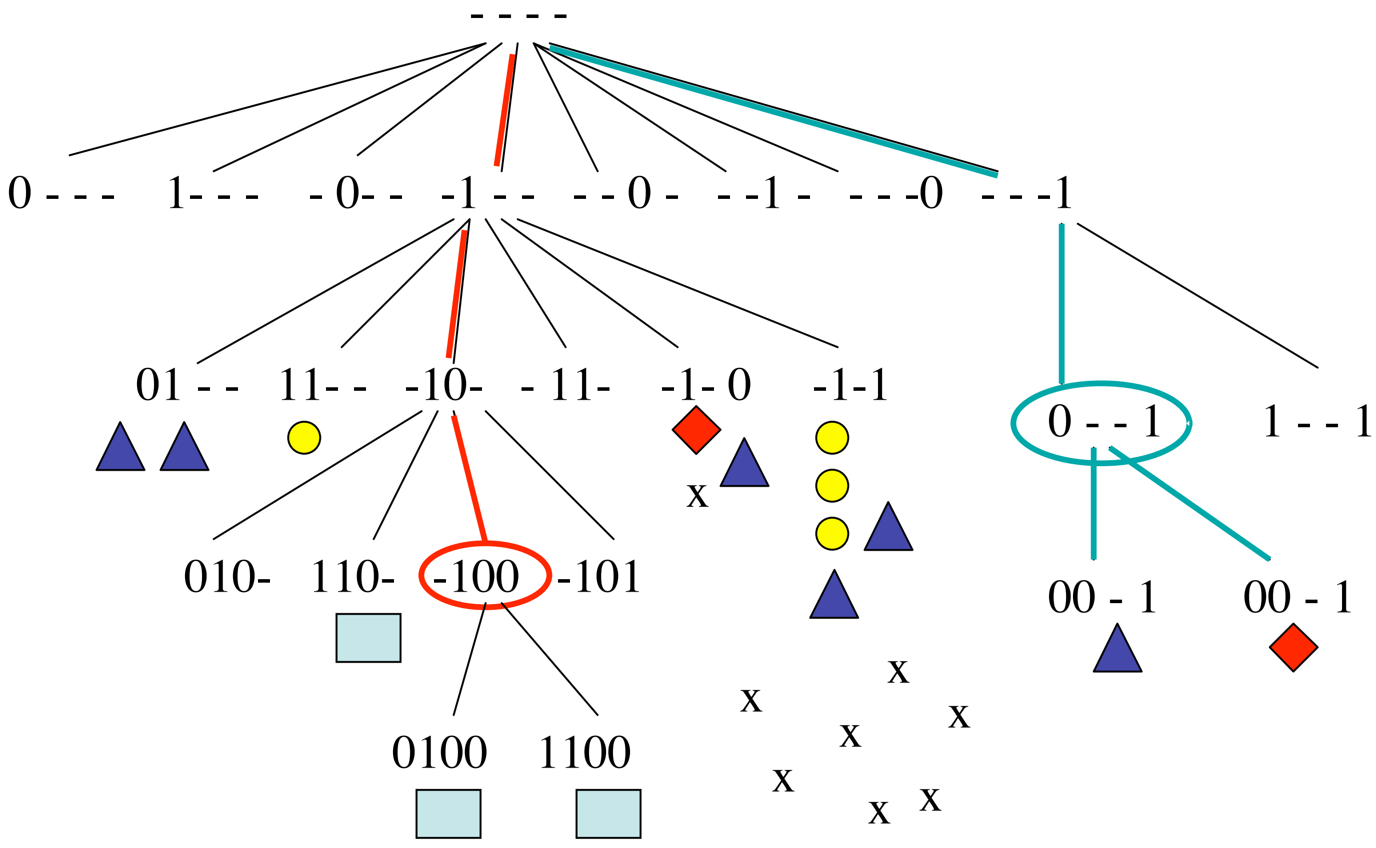
The Anti-unification concept

- *A generalization* is a description showing (some) aspects shared by a set of objects

- The most specific generalization (*anti-unification*) is a description showing *all* aspects shared by a set of objects

THE ANTI-UNIFICATION IS A SYMBOLIC SIMILARITY

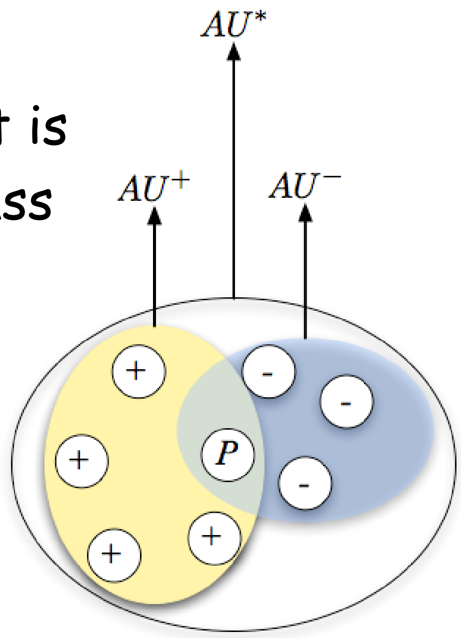
Example : Anti-unification



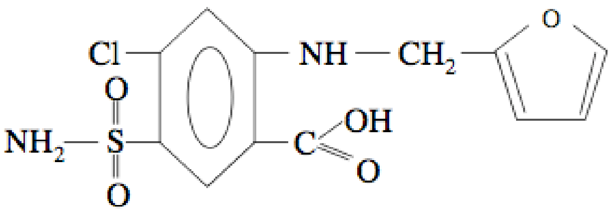
Explanation scheme

1) Explanation of Retrieve
   - ✓ A symbolic description consisting of all that is shared by the problem and the retrieved cases

2) Explanation of Reuse
   - ✓ Cases are organized according to the class
   - ✓ A symbolic description for each class
   - ✓ Each symbolic description consists of all that is shared by the problem and the cases of a class

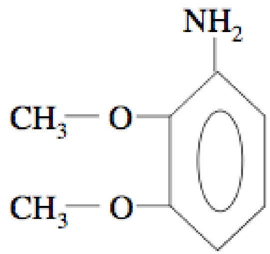


Example: Predictive Toxicology

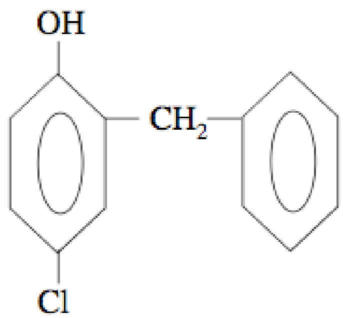


C-356: Furosemide

negative

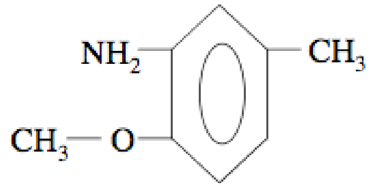


C-171: 2,4 - dimethoxyaniline



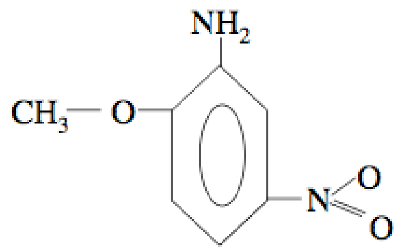
C-424: Benzyl - P - chlorophenol

positive

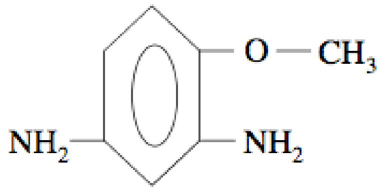


C-142: P-cresidine

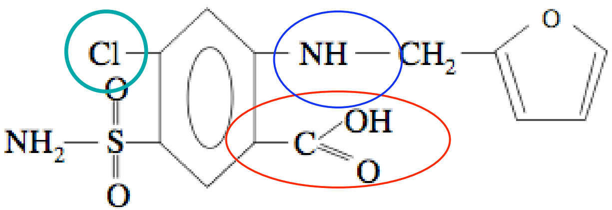
C-127: 5-nitro O- anisole



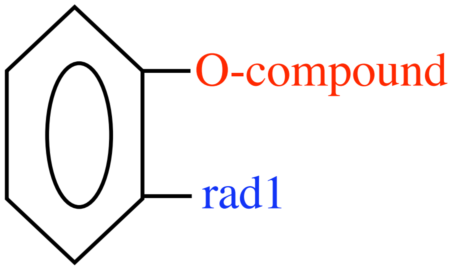
C-084: 2,4 - diamino anisole



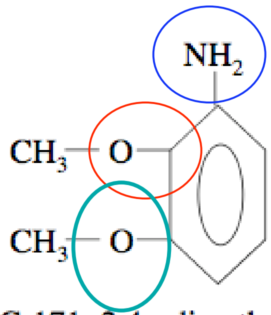
Example: AU*



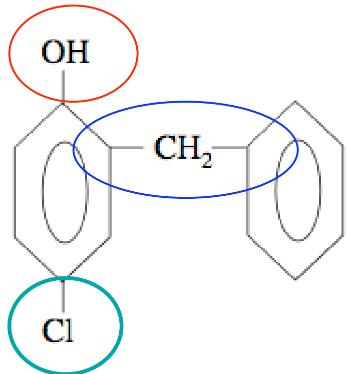
C-356: Furosemide



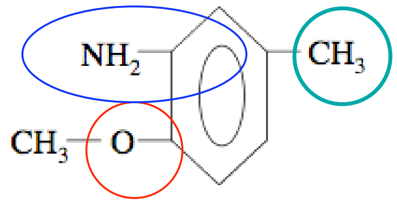
position? — rad2



C-171: 2,4 - dimethoxyaniline

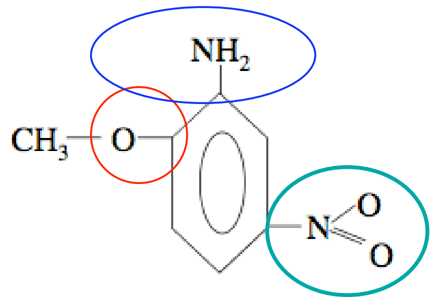


C-424: Benzyl - P - chlorophenol

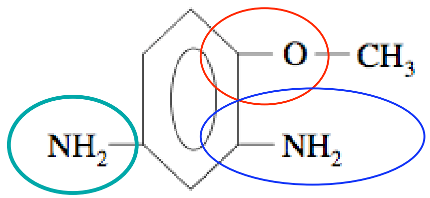


C-142: P-cresidine

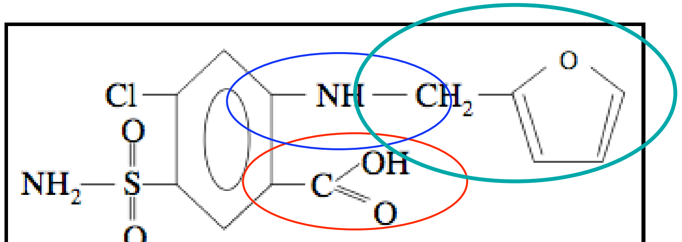
C-127: 5-nitro O- anisole



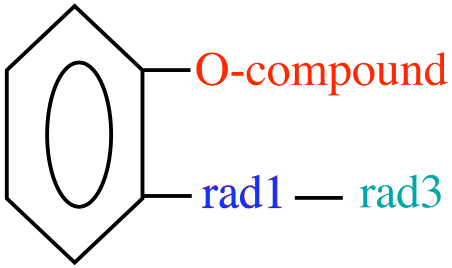
C-084: 2,4 - diamino anisole



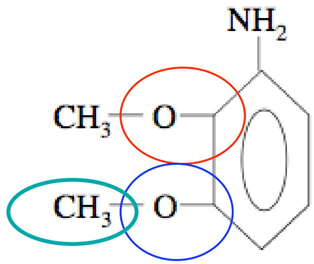
Example: AU$^-$



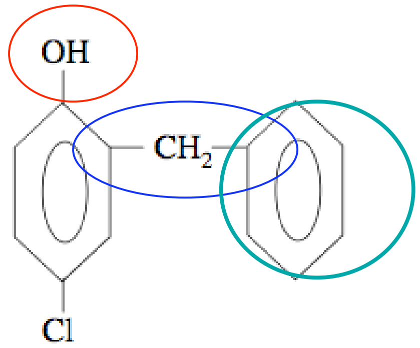
C-356: Furosemide



position? — rad2

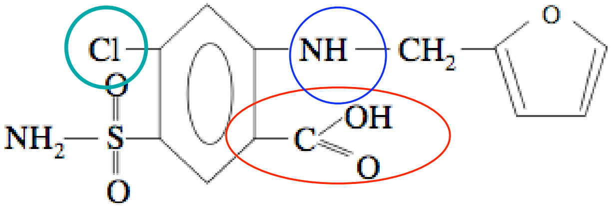


C-171: 2,4 - dimethoxyaniline

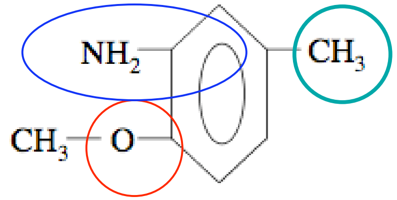


C-424: Benzyl - P - chlorophenol

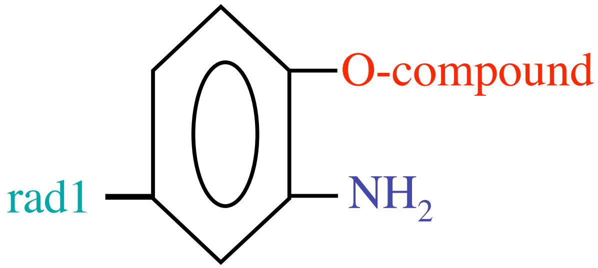
Example: AU^{+}



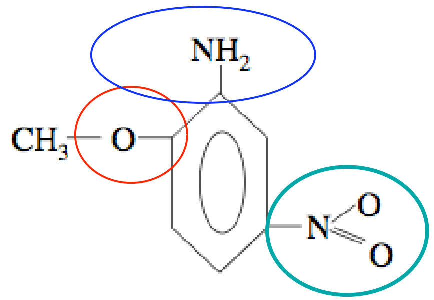
C-356: Furosemide



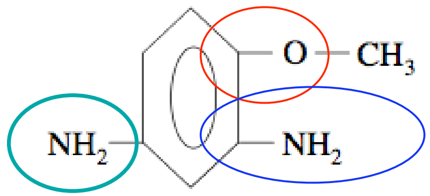
C-142: P-cresidine



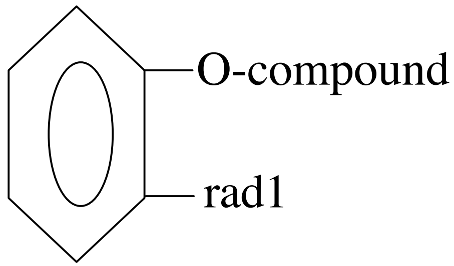
C-127: 5-nitro O- anisole



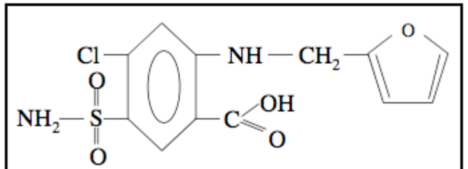
C-084: 2,4 - diamino anisole



AU * : anti-unification of retrieve set and problem

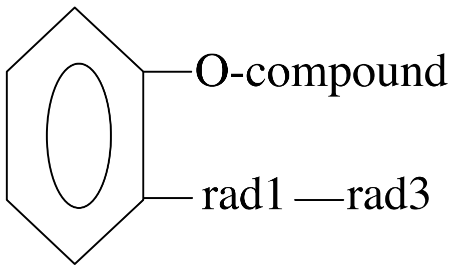


position? — rad2

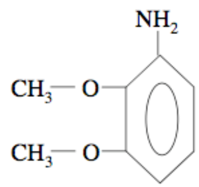


C-356: Furosemide

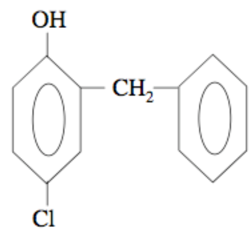
AU⁻ : anti-unification of negative cases and problem



position? — rad2

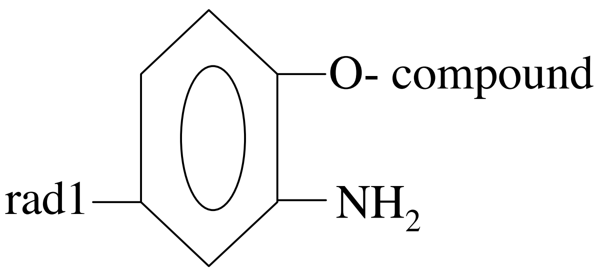


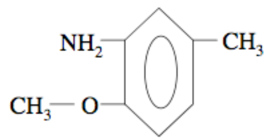
C-171: 2,4 - dimethoxyaniline



C-424: Benzyl - P - chlorophenol

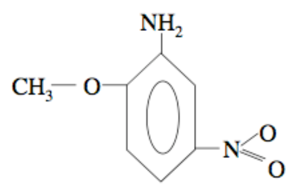
AU+ : anti-unification of positive cases and problem



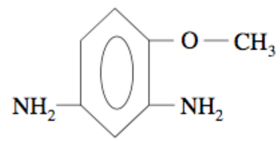


C-142: P-cresidine

C-127: 5-nitro O- anisole

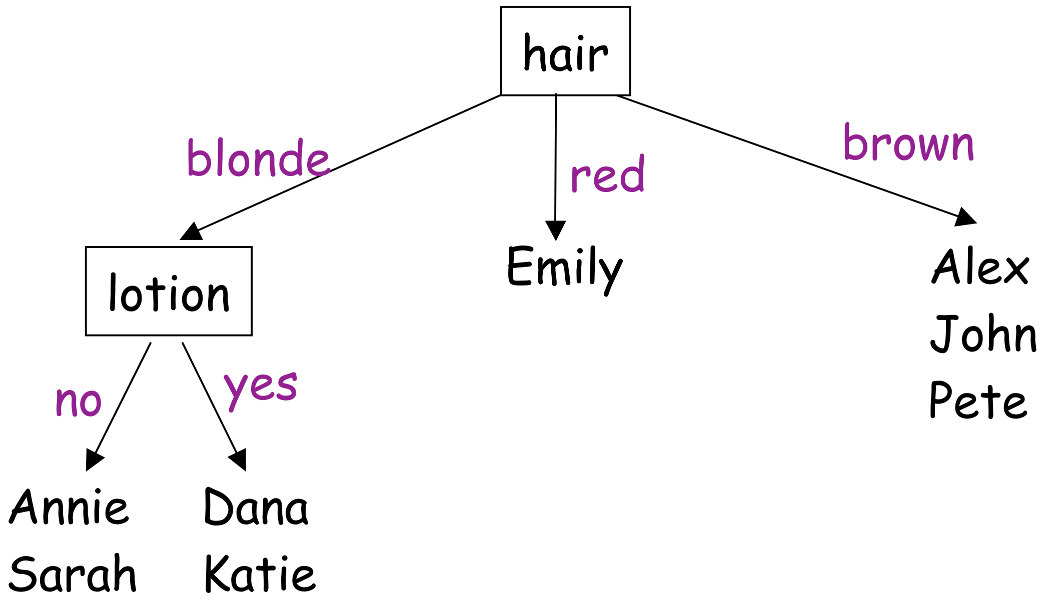


C-084: 2,4 - diamino anisole



Generalization as Explanation

ID3 (Quinlan, 1986)



The similitude term as explanation

1) Explanation of Retrieve
   - ✓ ~~A symbolic description consisting of all that is shared by the problem and the retrieved cases~~

   A symbolic description (the similitude term) with the features of the problem relevant for the classification (Armengol, 2007)

2) Explanation of Reuse
   - ✓ Cases are organized according to the class
   - ✓ A symbolic description for each class
   - ✓ Each symbolic description consists of all that is shared by the problem and the cases of a class

Anti-unification vs similitude term

- Similitude term of LID
  - Contains the features relevant for classification
  - Supervised data
- Anti-unification
  - Contains all that is common among a set of cases
  - It is independent on the similarity measure used for retrieval
  - Semi-supervised data
    - Explanation of clusters (Fornells et al., 2007)

Conclusions

- Generalizations are present in both eager and lazy learning methods
  - Eager learning methods build global approximations
  - Lazy learning methods build local approximations

- We propose three usages of generalization :
  - As symbolic similarity among cases since generalizations contain aspects shared by a set of objects
  - By storing the generalizations built by lazy learning methods we can obtain partial domain models
  - Generalizations can be interpreted as explanations of the system result

CSIC