

Artificial Data Sets based on Knowledge Generators: Analysis of Learning Algorithms Efficiency

Joaquin Rios-Boutin

Albert Orriols-Puig

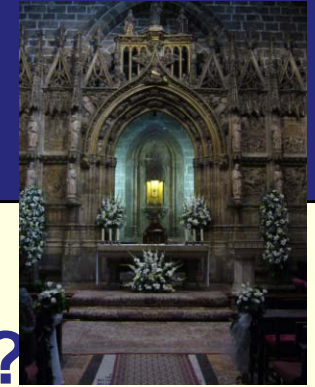
Josep-Maria Garrell-Guiu

**Grup de Recerca en Sistemes Intel·ligents
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
{jrios, aorriols, josepmg}@salle.url.edu**



HIS 2008

Motivation



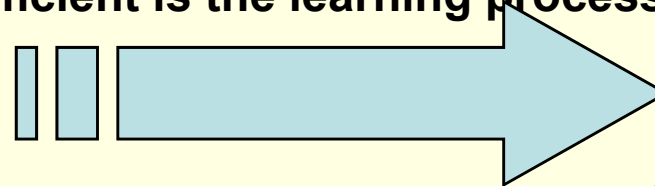
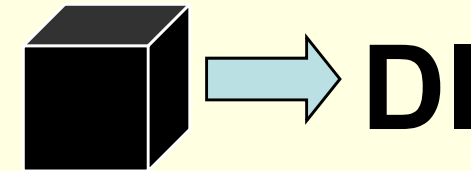
► What is the Holy Grail of Machine Learning?

- Find the right Learning Algorithm to every Problem
- Real Problems are black boxes

- We don't know which knowledge is contained

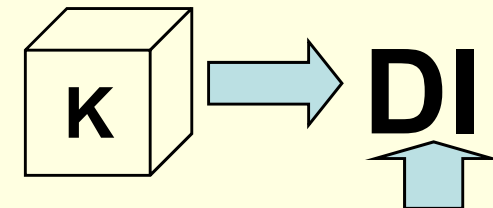
- We can't answer:

- When to stop training?
- How much efficient is the learning process?



– Artificial Problems:

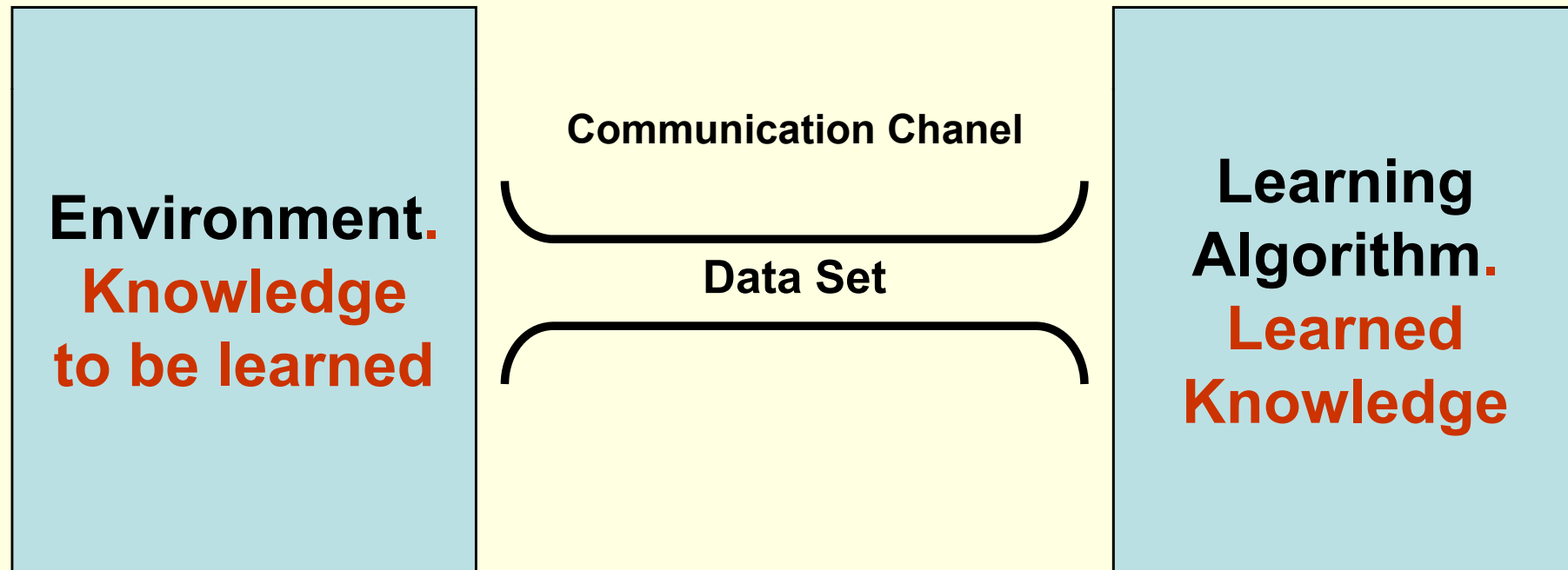
- Knowledge-driven
- Property-driven



Complex.Met.

Framework

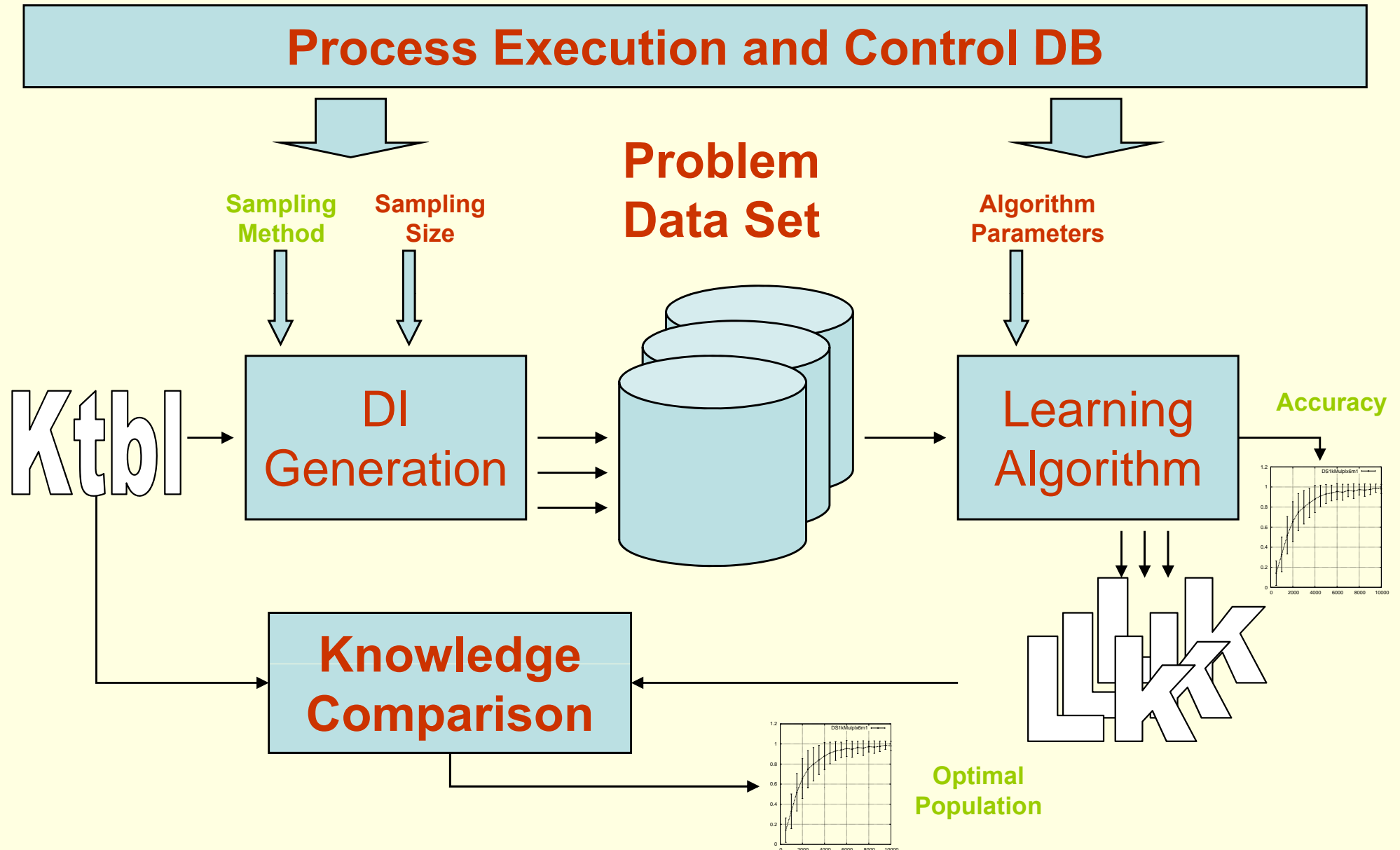
Machine Learning as a Communication System



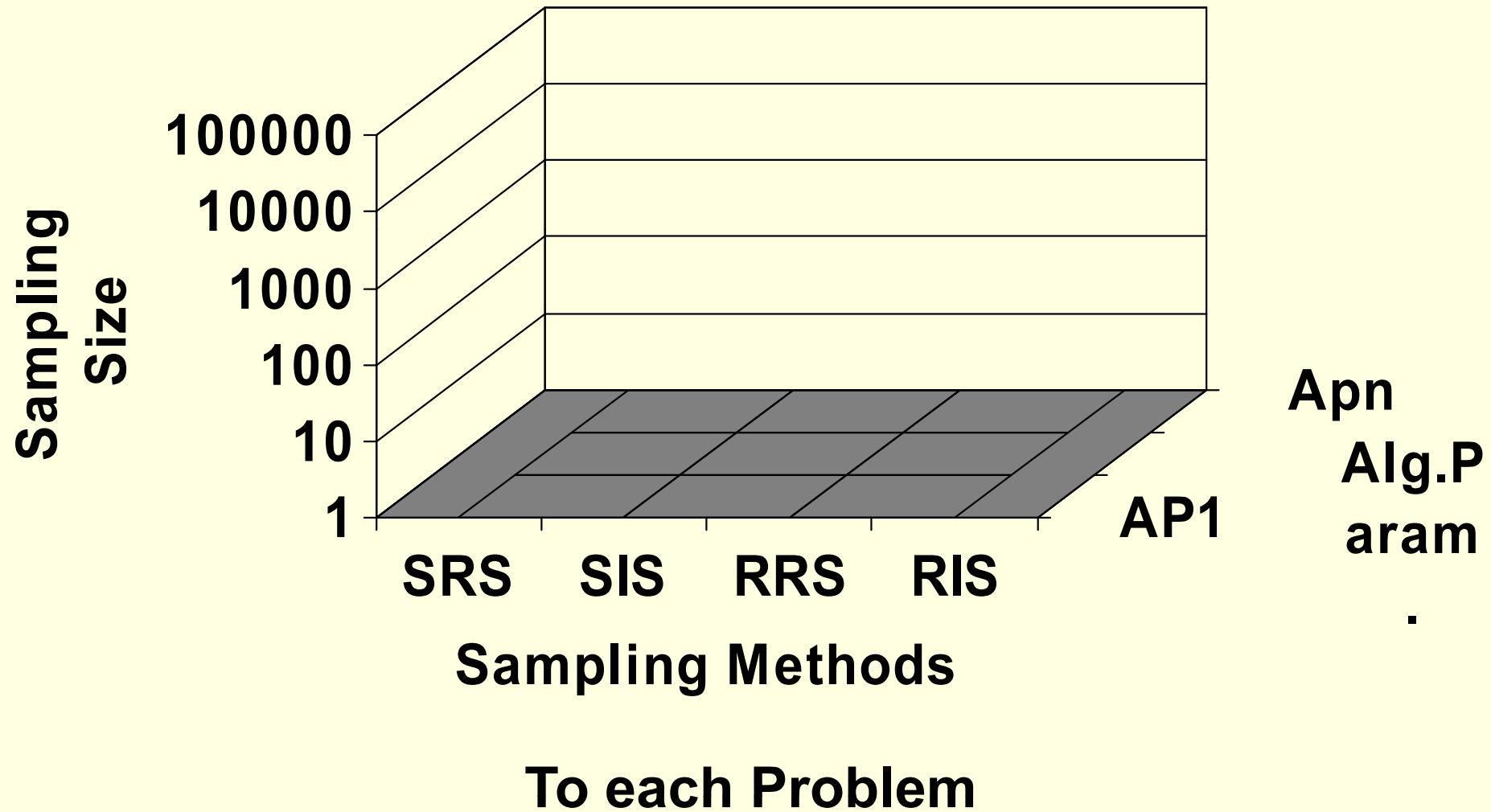
Outline

1. **Algorithm Evaluation Methodology Definition**
2. **Methodology Implementation**
3. **Experiment Description**
4. **Results and Analysis**
5. **Conclusions and Further Work**

1 Algorithm Evaluation Process



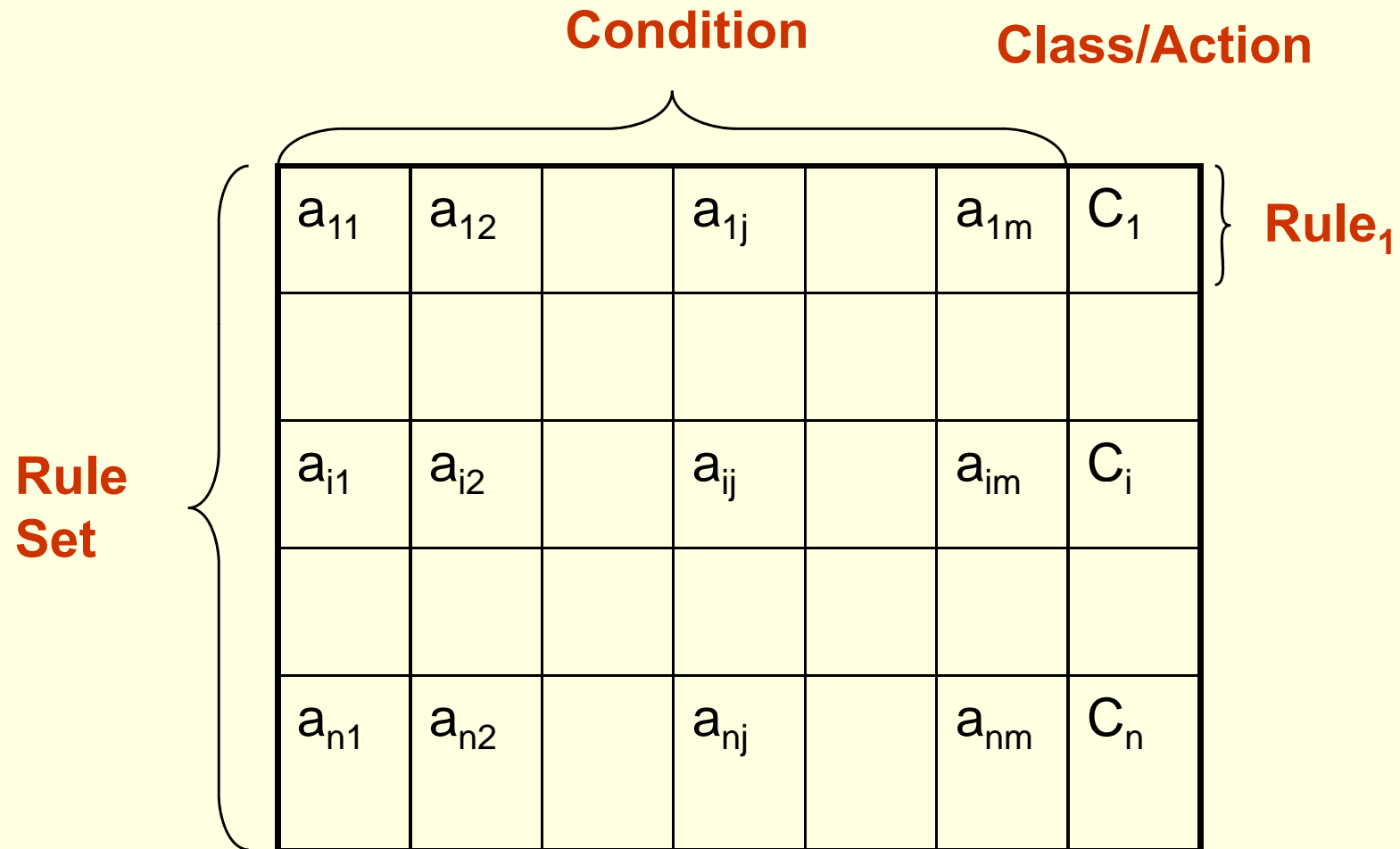
1 Algorithm Evaluation Process Dimensions



Outline

- 1. Algorithm Evaluation Methodology Definition**
- 2. Methodology Implementation**
- 3. Experiment Description**
- 4. Results and Analysis**
- 5. Conclusions and Further Work**

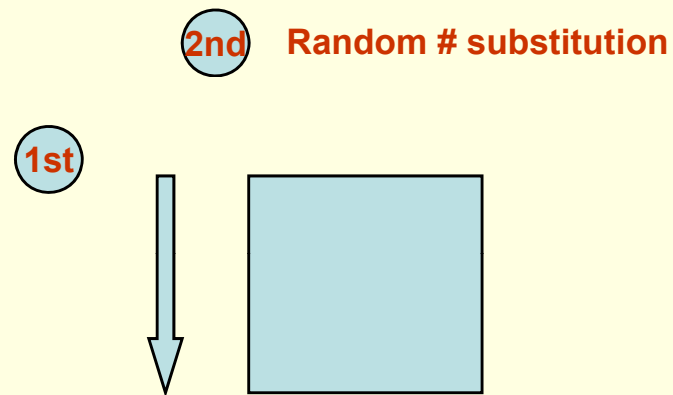
2 Knowledge Representation



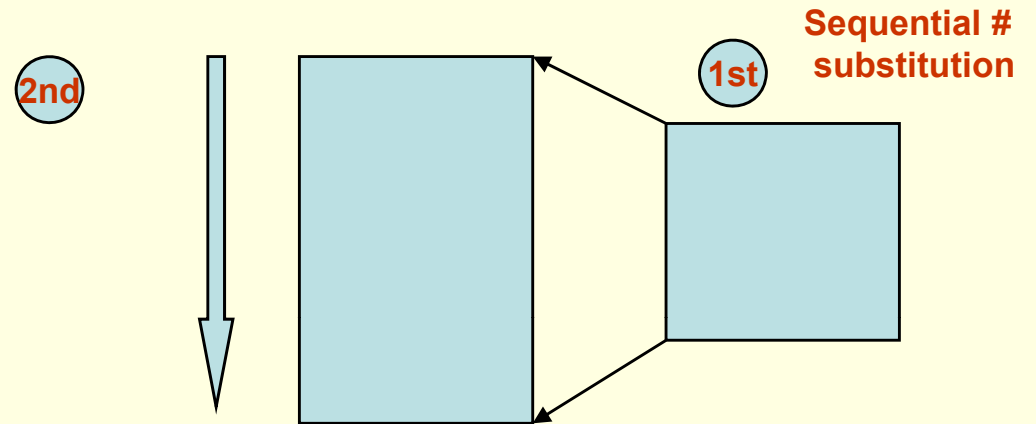
$$a_{ij} = \{0, 1, \#\} \quad C_i \in \mathbb{N}$$

2 Sampling Methods

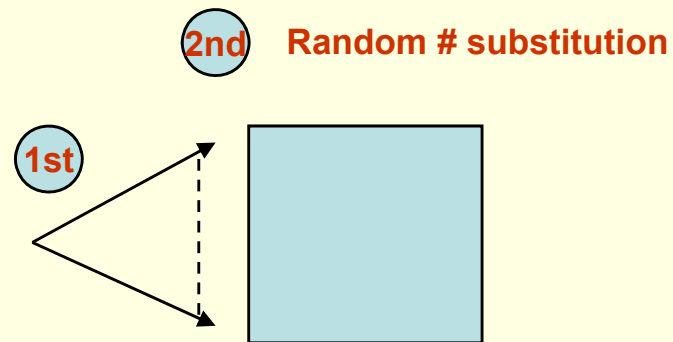
SRS Sequential Rule Selection



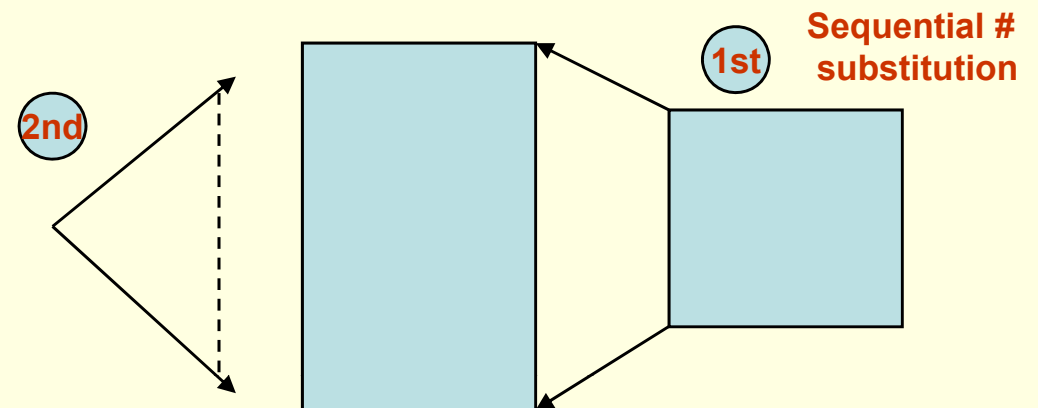
SIS Sequential Instance Selection



RRS Random Rule Selection



RIS Random Instance Selection



2 Problems to learn and Learning Algorithm

Mux6 Mux11

0	0	#	#	#	0	0
0	0	#	#	#	1	1
0	1	#	#	0	#	0
0	1	#	#	1	#	1
1	1	0	#	#	#	0
1	1	1	#	#	#	1

Parity5

0	0	0	0	0	0
0	0	0	0	1	1
0	0	0	1	0	1
0	0	0	1	1	0
1	1	1	1	0	0
1	1	1	1	1	1

XCS

Position5 Position11

0	0	0	0	0	0
0	0	0	0	1	1
0	0	0	1	#	2
0	0	1	#	#	3
1	#	#	#	#	5

Parity5-3

0	0	0	0	0	#	#	#	0
0	0	0	0	1	#	#	#	1
0	0	0	1	0	#	#	#	1
0	0	0	1	1	#	#	#	0
1	1	1	1	0	#	#	#	0
1	1	1	1	1	#	#	#	1

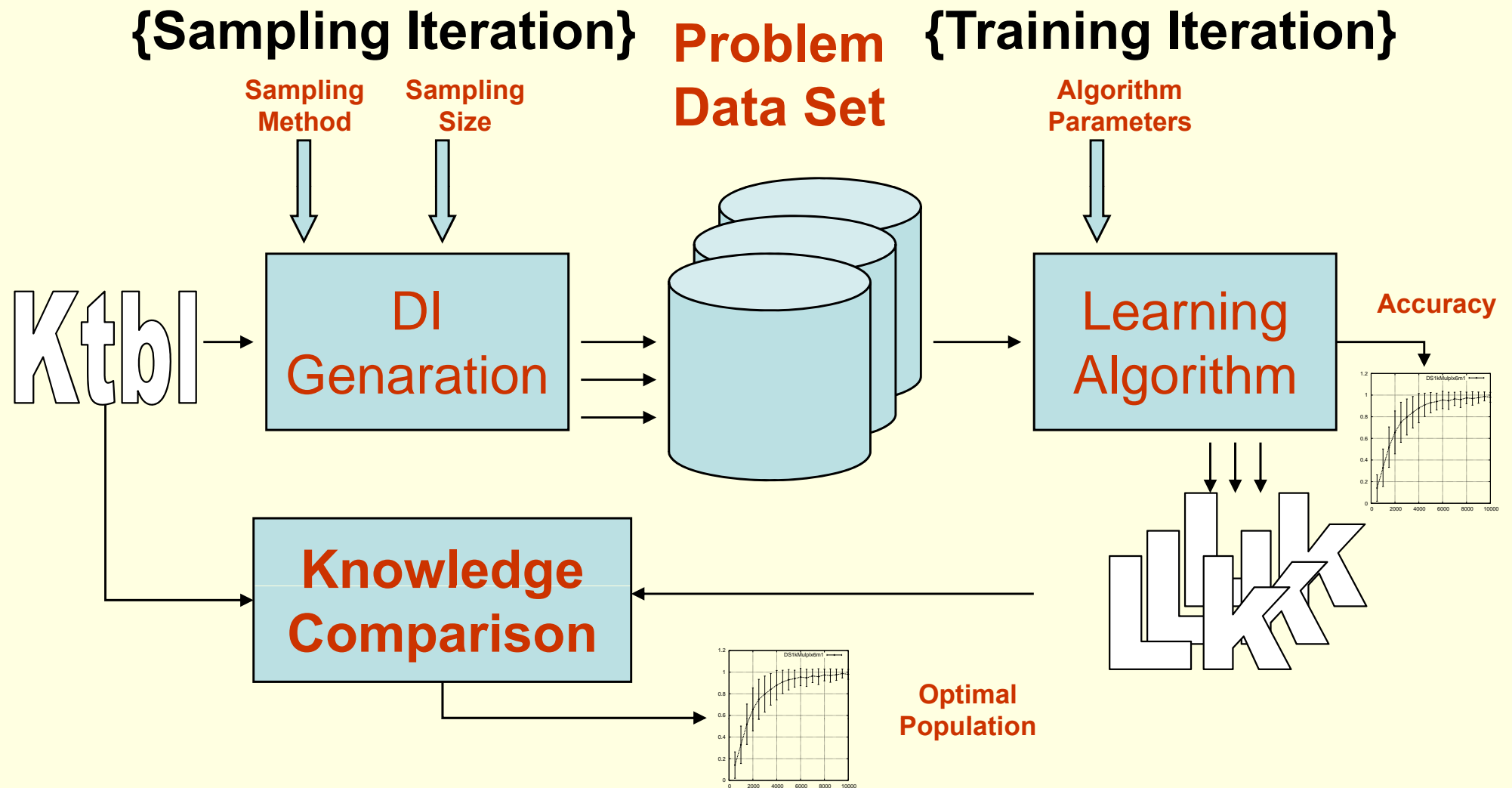
2 Problem Properties

- ▶ Optimal Rule Sets
 - Complete
 - Non overlapped
 - Irreducible
- ▶ Why?
 - Simple structure of knowledge complexity
 - Very known artificial problems

Outline

- 1. Algorithm Evaluation Methodology Definition**
- 2. Methodology Implementation**
- 3. Experiment Description**
- 4. Results and Analysis**
- 5. Conclusions and Further Work**

3 Sampling and Learning Iteration



3 Output Results and Iteration Reduction

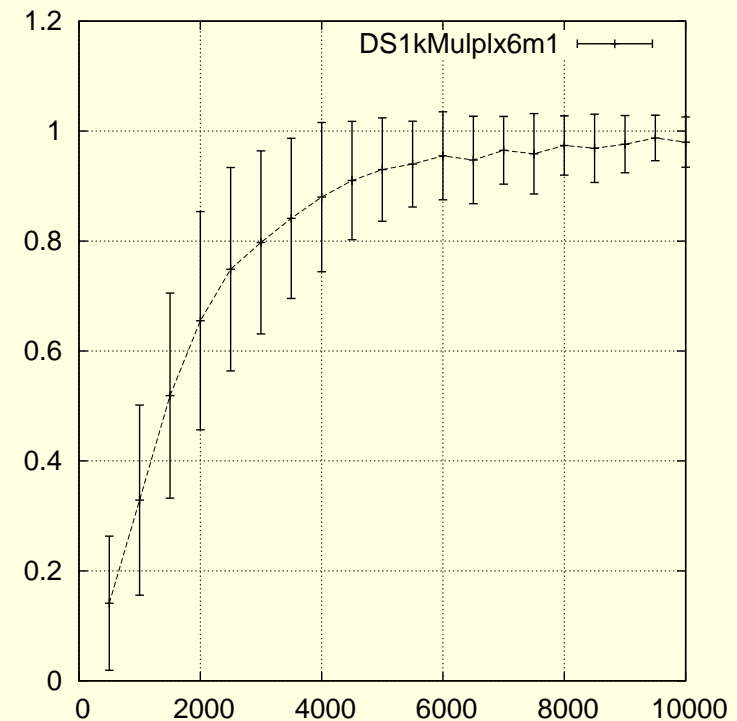
▶ Output Results

- 2 Plots to every Problem, Sampling Method, Sampling Size and Algorithm Parameters.

- Optimal Population
- Accuracy

▶ Iteration Reduction

- SIS Pure sequential
 - No Sampling Iteration Needed
- Problems without “don't care”
 - $SRS=SIS$ and $RRS=RIS$



3 Experimental Parameters

- ▶ Number of Problems = 6
- ▶ Number of Sampling Methods = 4
- ▶ Number of different Sampling Sizes = 4
- ▶ Number of different Algorithms Parameters Sets = 2
- ▶ Number of Sampling Iterations = 10
- ▶ Number of Training Iterations = 10
- ▶ Number of Data Sets Generated = 744
- ▶ Number of Training Process = 14880

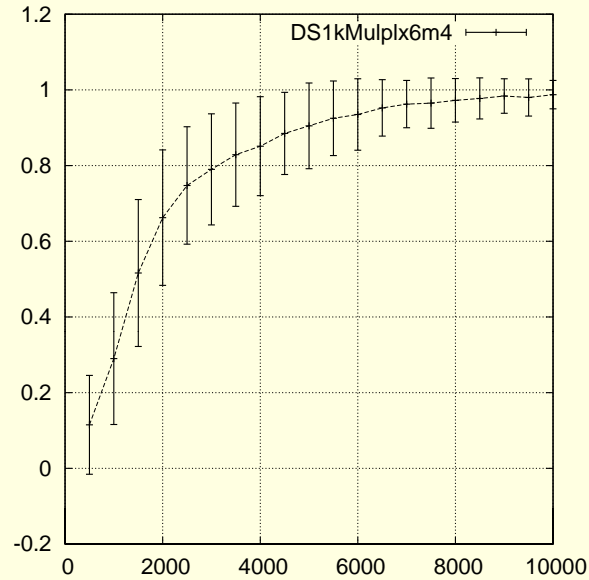
Outline

- 1. Algorithm Evaluation Methodology Definition**
- 2. Methodology Implementation**
- 3. Experiment Description**
- 4. Results and Analysis**
- 5. Conclusions and Further Work**

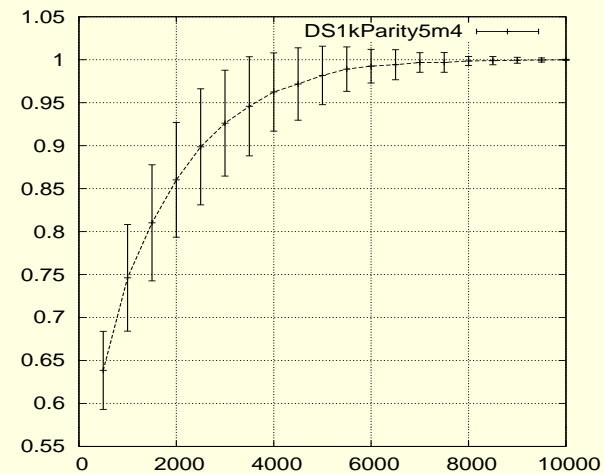
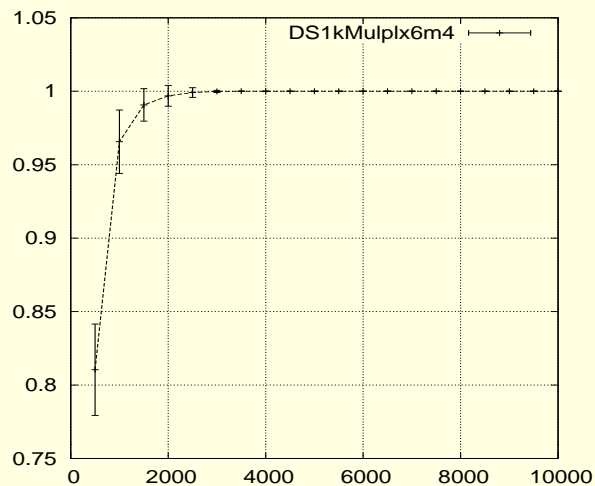
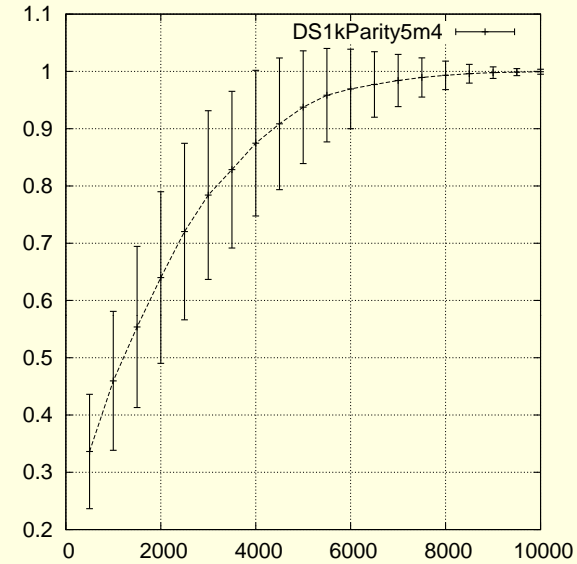
Problem Dimension

Sampling M. = RIS Sampling Size = 1000 Learning Alg. Param. = pDNC 0.2

Mux6



Parity5

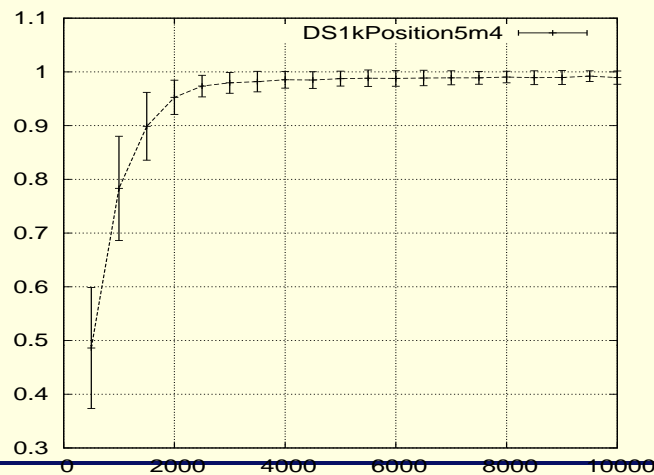
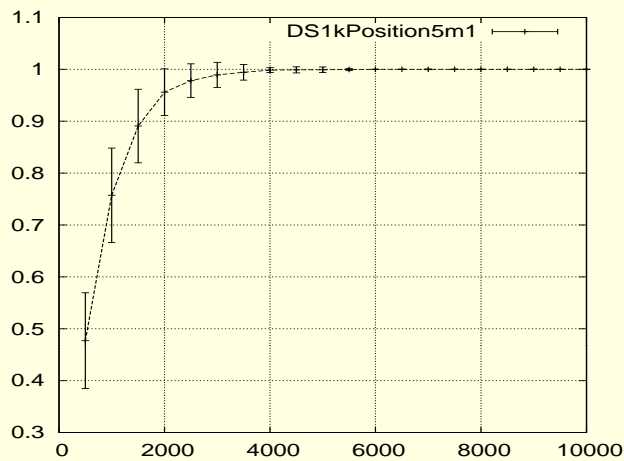
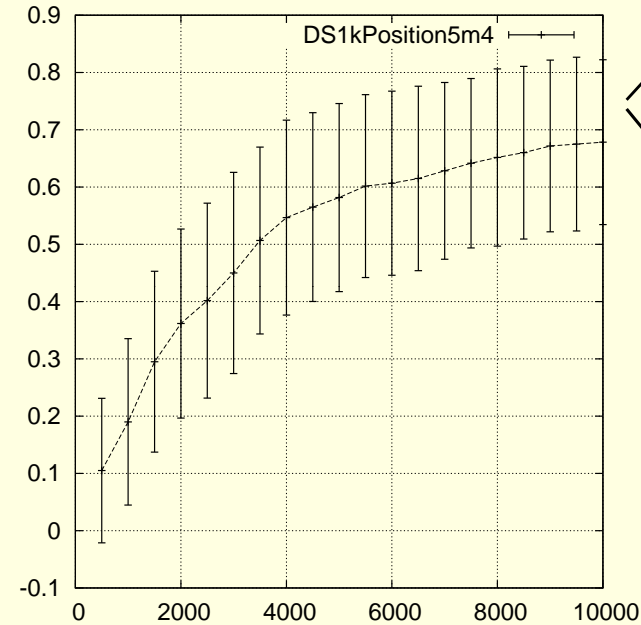
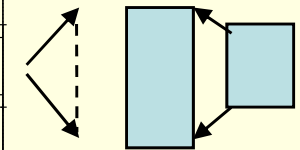
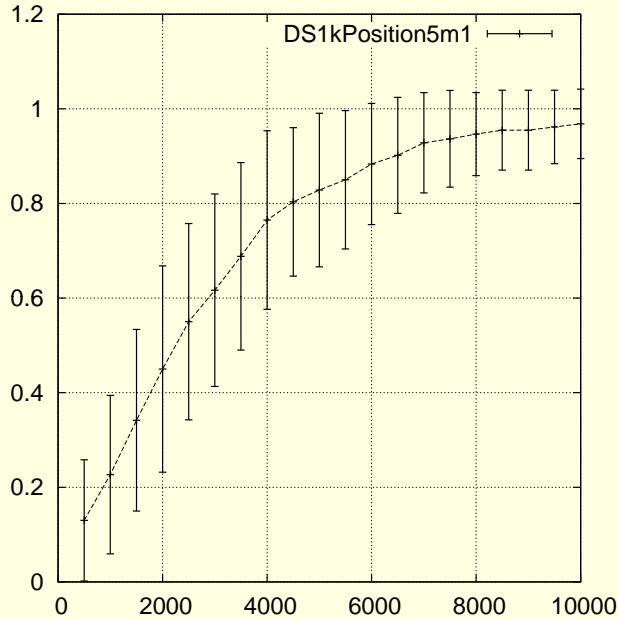
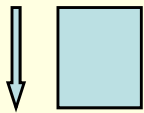


Sampling Method Dimension

Problem = Position5 Sampling Size = 1000 Learning Alg. Param. = pDNC 0.2

SRS Sequential Rule Selection

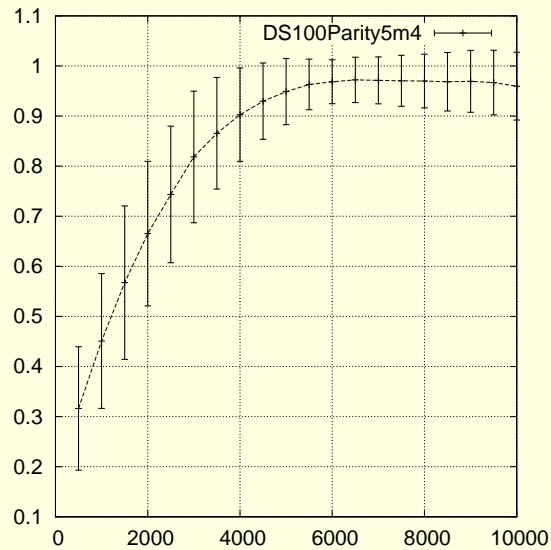
RIS Random Instance Selection



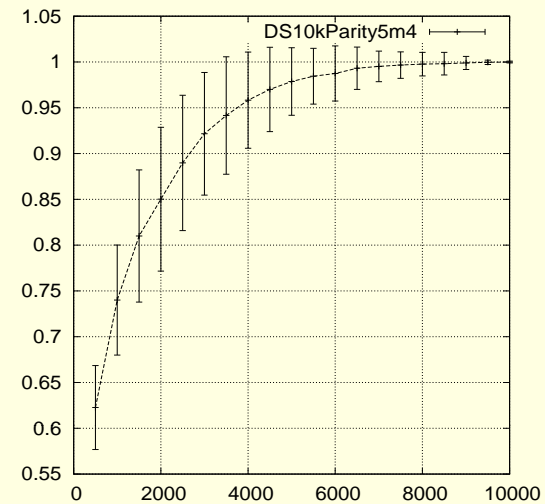
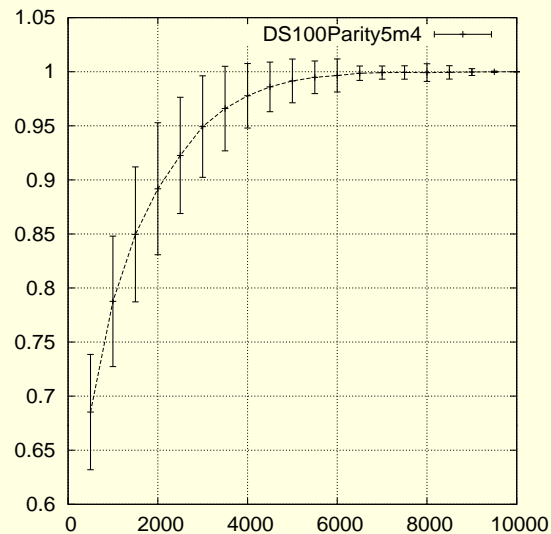
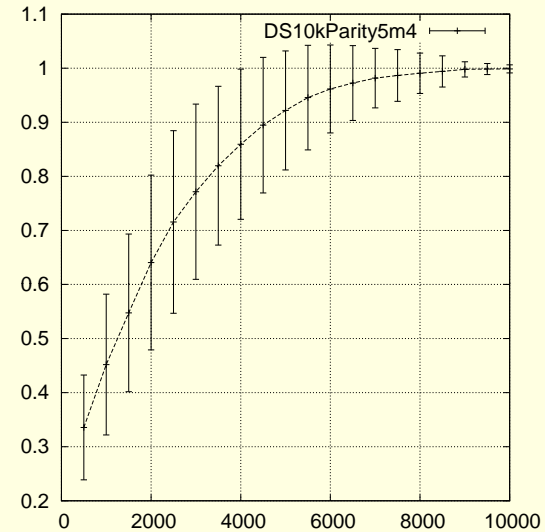
Sampling Size Dimension

Problem = Parity5 Sampling M.= RIS Learning Alg. Param. = pDNC 0.2

100



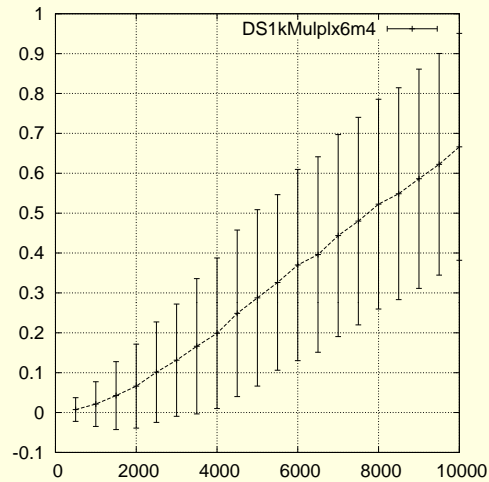
10000



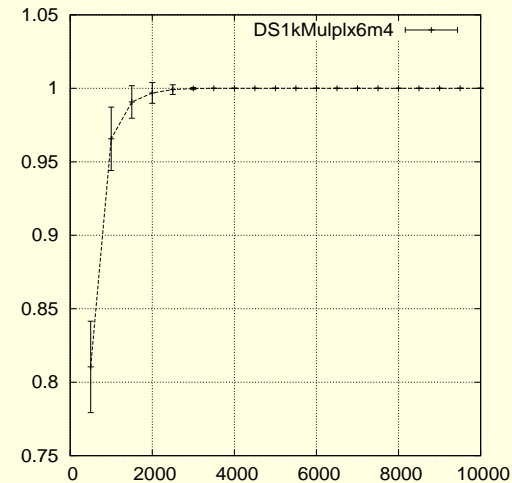
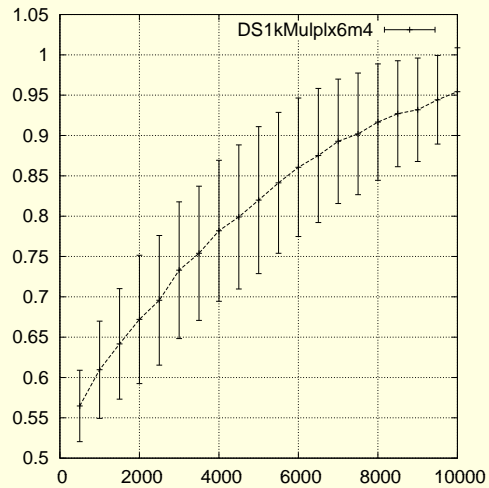
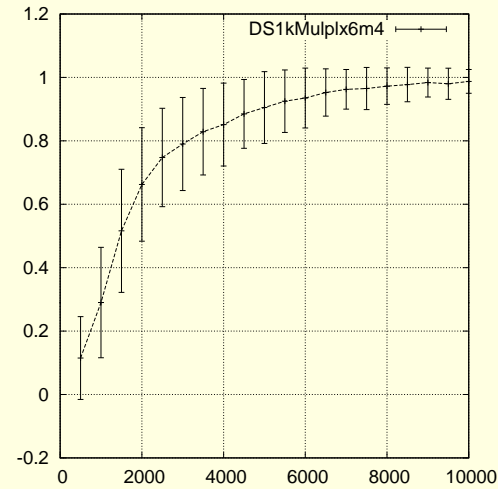
Parameter Algorithm Dimension

Problem = Mux6 Sampling M. = RIS Sampling Size = 1000

pDNC 0.8



pDNC 0.2



Outline

- 1. Algorithm Evaluation Methodology Definition**
- 2. Methodology Implementation**
- 3. Experiment Description**
- 4. Results and Analysis**
- 5. Conclusions and Further Work**

Conclusions and Further Work

▶ Conclusions

- Automatic Learning Algorithm Analyzer based on Artificial Data Sets
- Four dimensions comparisons
- Methodology Implementation, Experiment and Results Analysis

▶ Further Work

- Non ORS Problems
- Real Attributes
- Sampling Methods based on distance or transition matrix
- Multi Step Problems
- Different Learning Algorithms
- Different Knowledge representations
- Knowledge Covering Metrics
- Applying Data Set Complexity Metrics Suite

GRSI

▶ Artificial Data Sets based on Knowledge Generators: Analysis of Learning Algorithms Efficiency

Joaquin Rios Boutin, Albert Orriols-Puig, Josep-Maria Garrell-Guiu

{jrios, aorriols, josepmg}@salle.url.edu

▶ **GRSI** (Grup de Recerca en Sistemes Intel·ligents)

- <http://www.salle.url.edu/GRSI>

– Oriented to:

- ***CBR (Computer Based Reasoning) Algorithms***
- ***Evolutionary Computation Algorithms***
- ***Data Mining Technology Transfer***