

Mining Evolutionary Model

MEM

Rida E. Moustafa
And
Edward J. Wegman

George Mason University

Email:

{rmoustaf,ewegman}@galaxy.gmu.edu

Phone:703-993-1680

Interface 2000

April,4

Mining Evolutionary Model

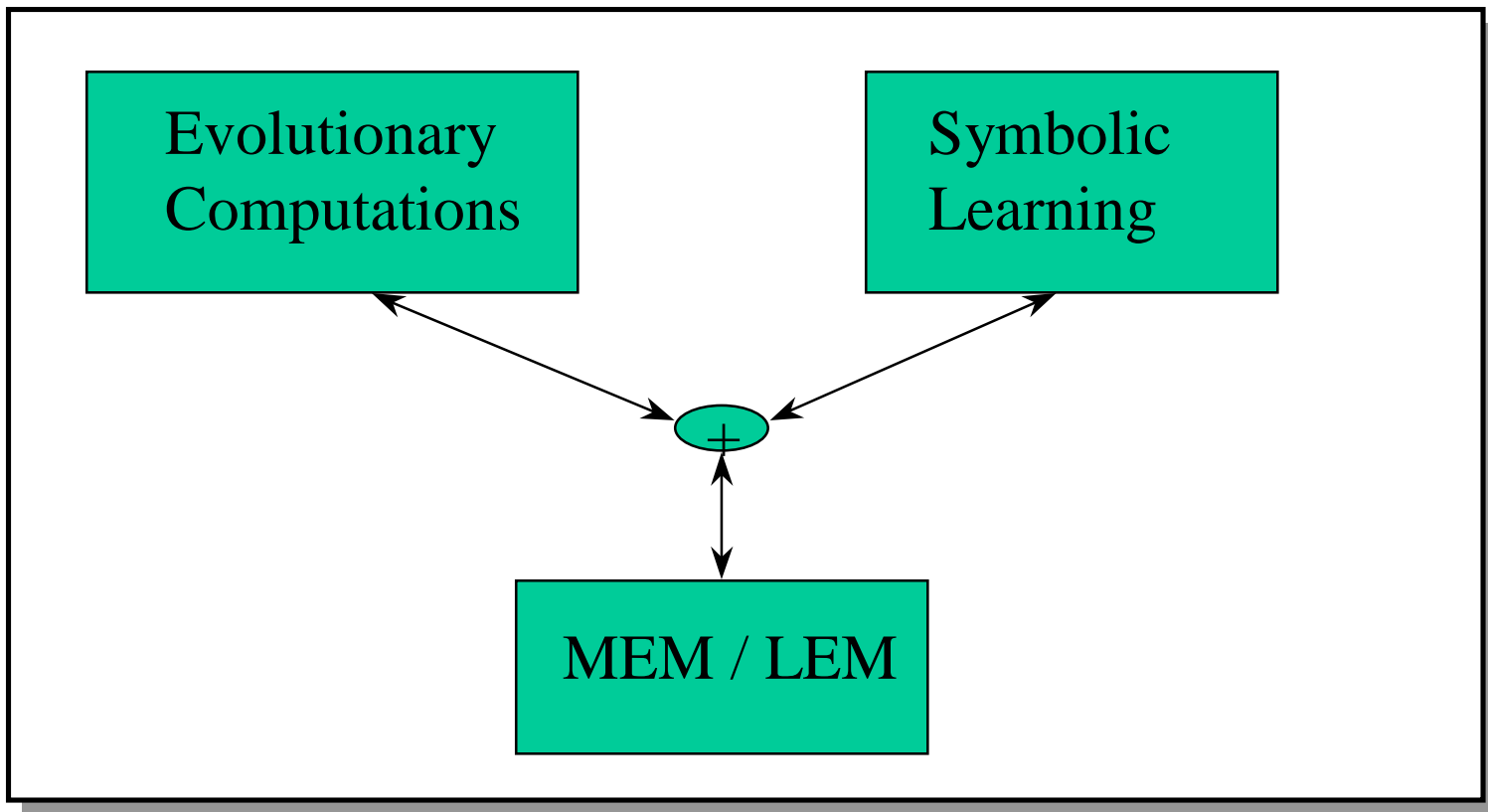
MEM

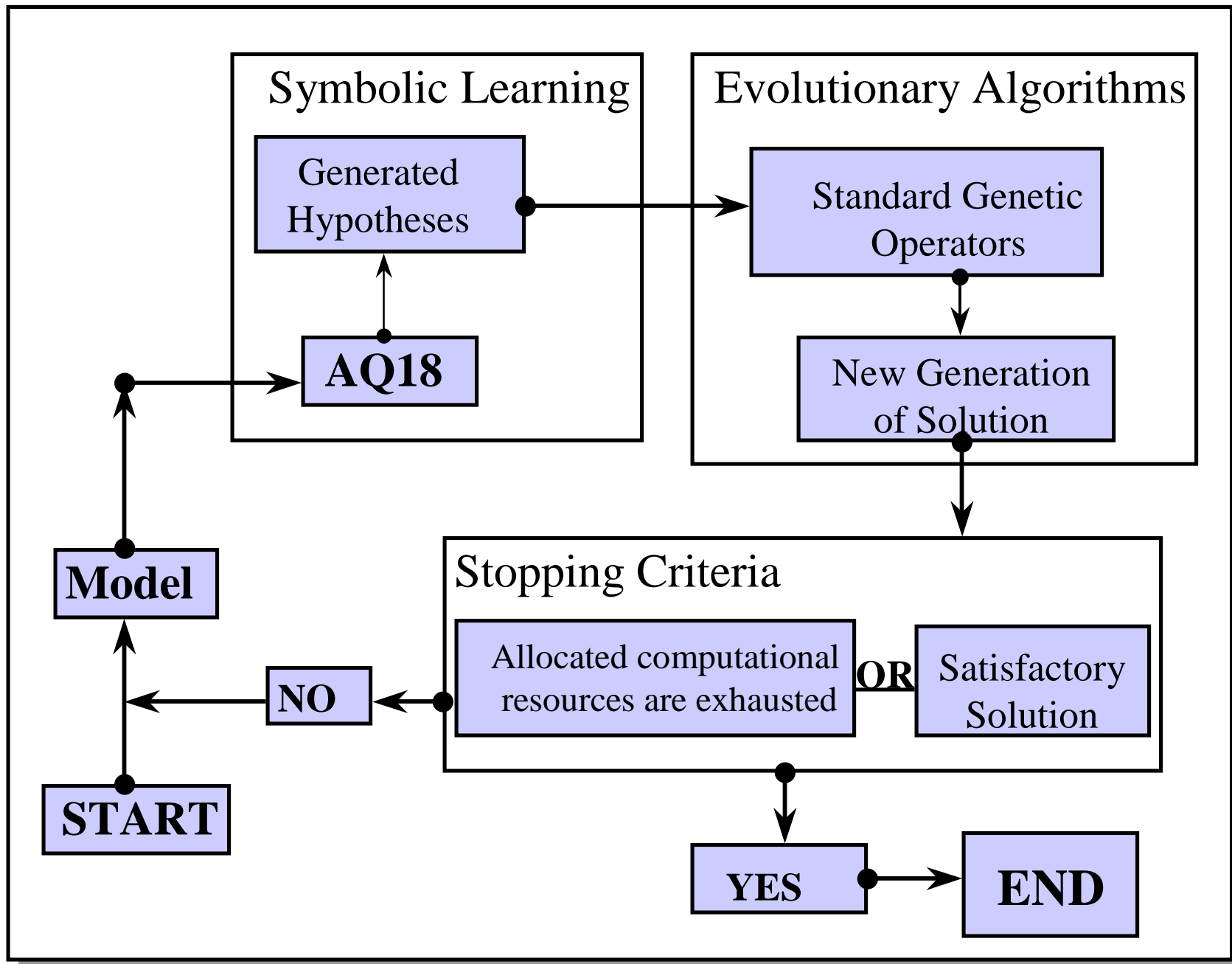
Talk Outline

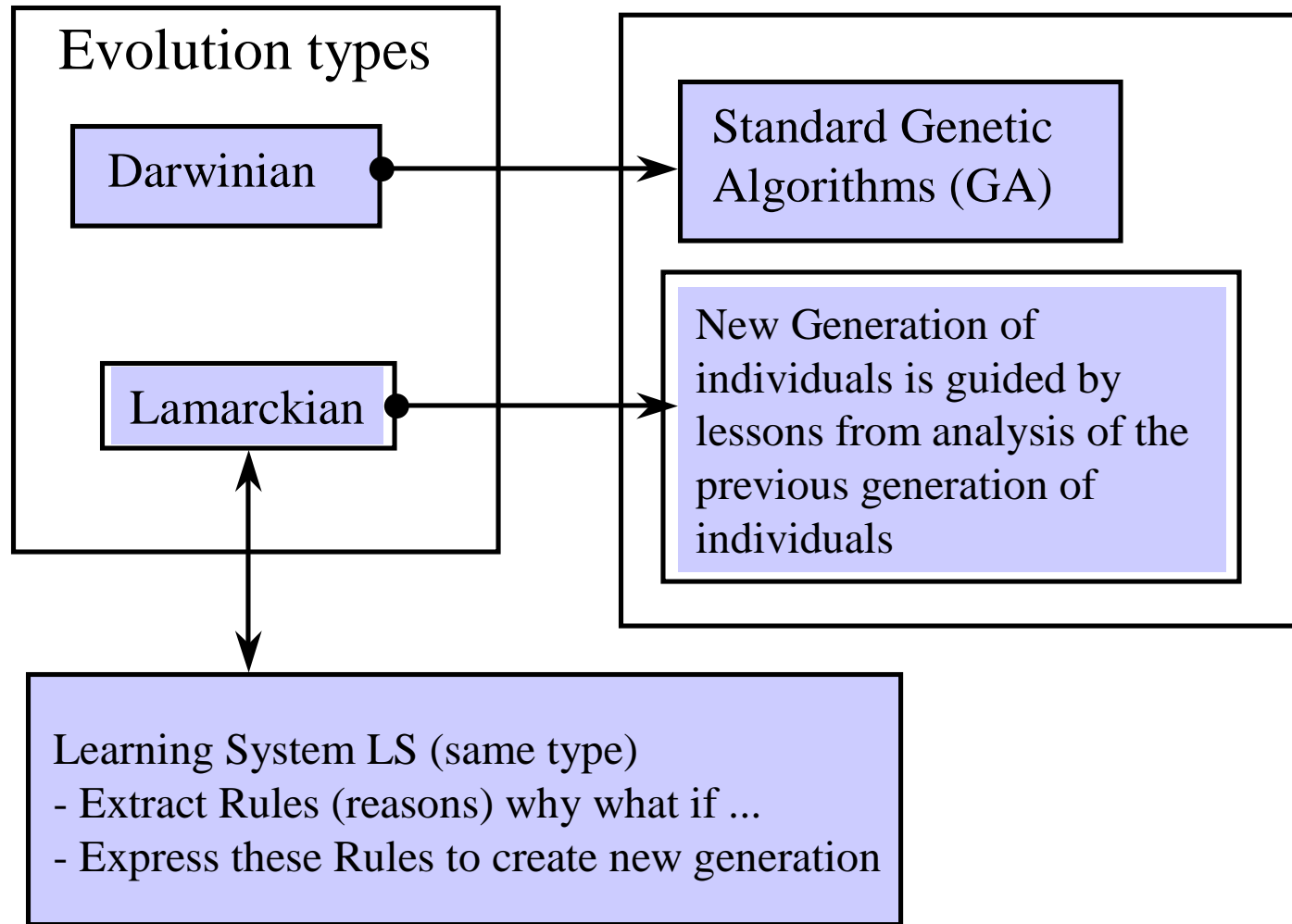
- MEM Theory.
- Evolutionary computation.
- Multidimensional Scaling.
- Gene Measurements.
- Results and Future work.

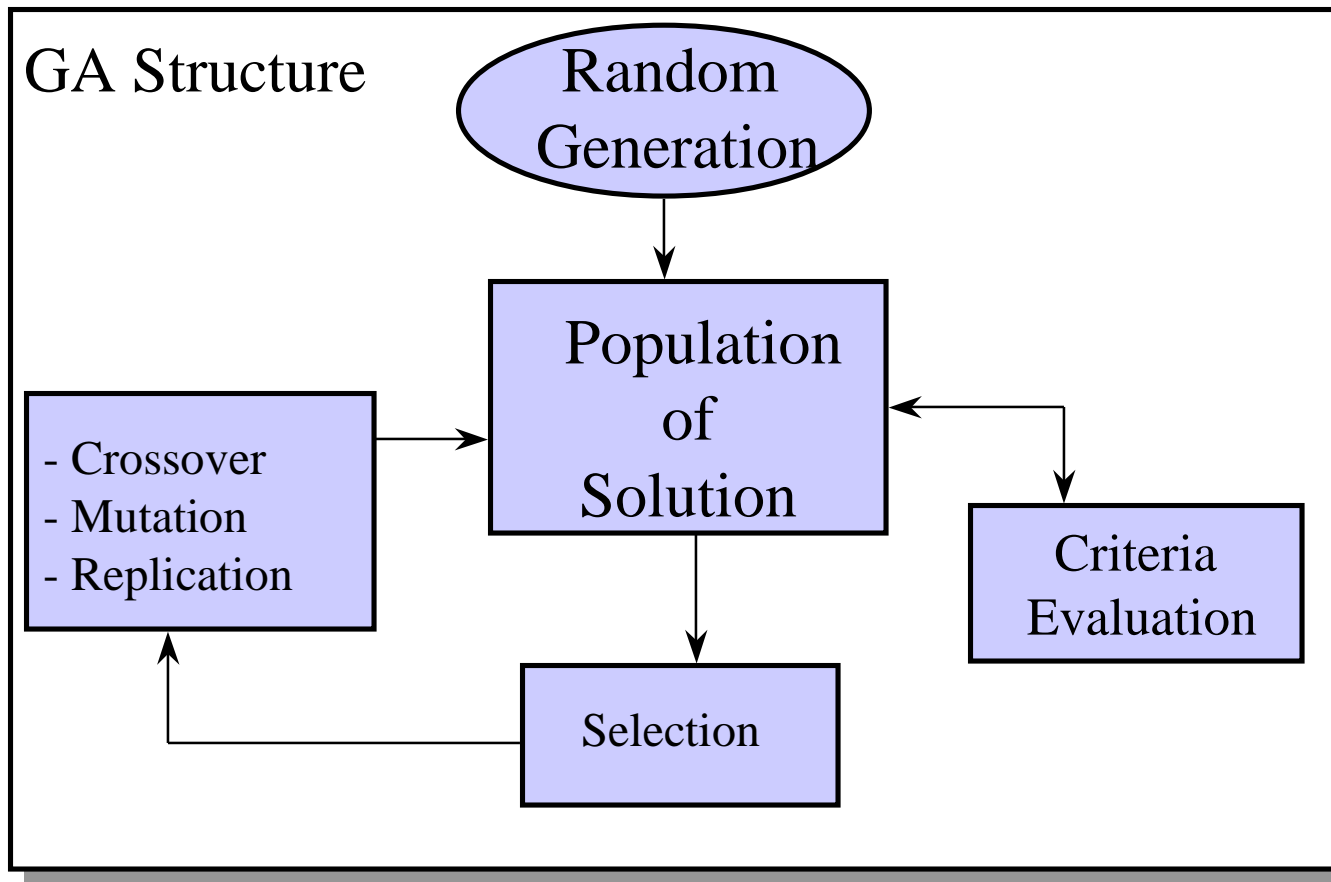
Mining Evolutionary Model

MEM





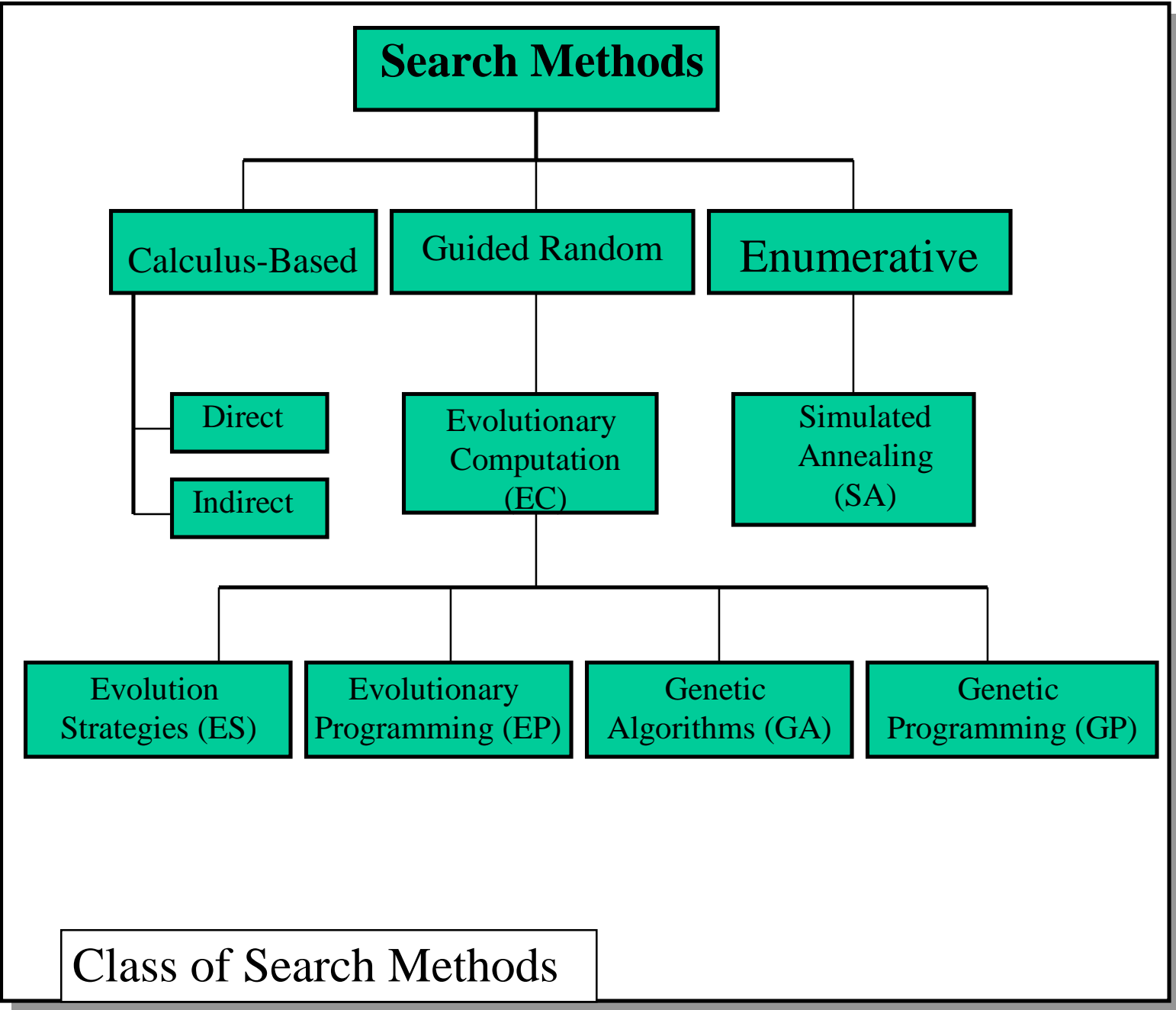




Replication: 11110011 → 11110011

Mutation : 11110011 → 10110011

Crossover $\begin{array}{l} \curvearrowright 11110011 \\ \curvearrowleft 10111101 \end{array} \Big| \rightarrow \begin{array}{l} 11110101 \\ 10111011 \end{array}$



The Problem

Mining Evolutionary Model:MEM/LEM

- Take 2-point recombination data.

Gene Loci	% Recombination
A,B	10
A,C	5
B,C	15

CASE I

Simple Case

Gene Loci	% Recombination
A,B	50
A,C	15
A,D	38
A,E	8
B,C	50
B,D	13
B,E	50
C,D	50
C,E	7
D,E	45

CASE II: Five Gene
Location (Posed Problem)

Goal : Indicate the relation between recombination percentage and interval length

Idea:

- Permute {A,B,C} S.T. $|A-B|=10$; $|A-C|=5$; $|B-C|=15$
- The recombination percentage corresponds to an absolute distance Between the relevant gene loci.

Sol: Set $A=0$ you can find the rest easy by inspection (+/- 10,5)

The Solution is not unique



No Exact Solution:

Assume for example you have :

$$|A-B|=50; |A-D|=38; |B-D|=13$$

Translate $A=0 \rightarrow \rightarrow B = (+/-50); D = (+/- 38)$
 $\rightarrow \rightarrow |B-D|=13$ can not hold !!!

- The data for Shorter lengths is more reliable [Russell 1986]
- $\rightarrow \rightarrow$ We must analyze the data set in more Statistical Fashion !!

Error Metric Approaches:

1. Simply strike out the larger percentages and work only with smallest
2. Weight the smaller distances preferentially.
3. Represent the distance-percentage relation for genes: $|G_{-I} - G_{-J}| = \%IJ$
4. Error = $\sum [((G_{-I} - G_{-J}) / \%IJ)^2 - 1]^2 ; I \neq J$.

Important Notice : The Error Proves Metric Space

1. If a coordinate solution is exact \rightarrow Error is Zero.
2. $E(G_{-I} - G_{-J}) = E(G_{-J} - G_{-I}) \rightarrow$ Symmetric.
3. $E(G_{-I} - G_{-J}) + E(G_{-J} - G_{-K}) \leq E(G_{-I} - G_{-K}) \rightarrow$ Triangle Inequality.

Possible Test Cases:

1. Set A=0 → 4-dimension Optimization Problem
2. Set A .NEQ. 0 → 5-dimension Optimization Problem.

Our result here shown for 5-D case:

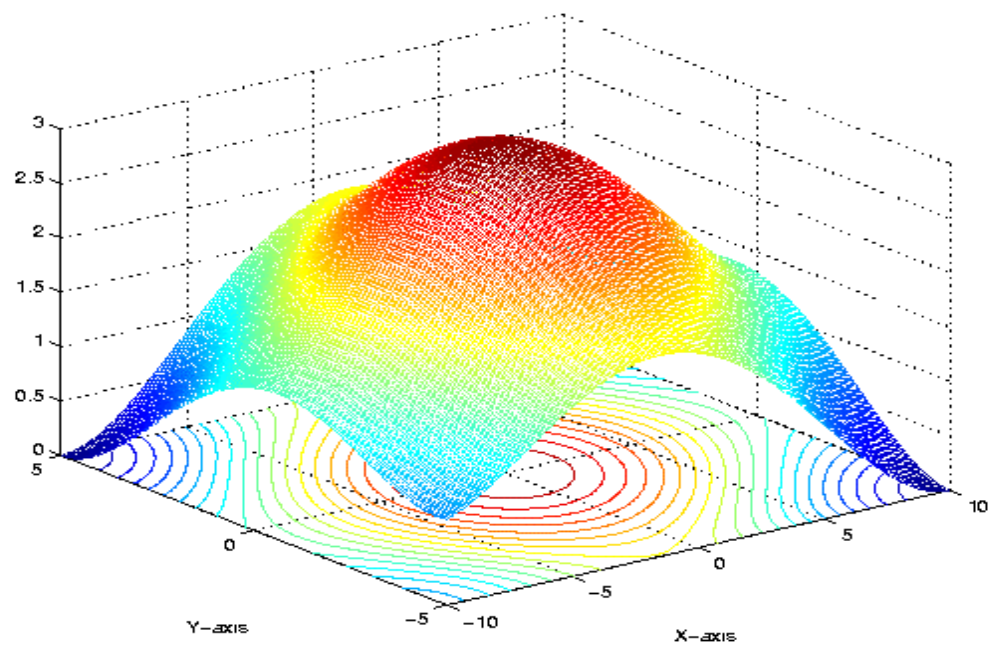
Landscape: $[-50,50]^5 \rightarrow (10^6)$ point representations

Fitness Function is : $\text{Fitness}=1/\text{Error}(G_I, G_J)$

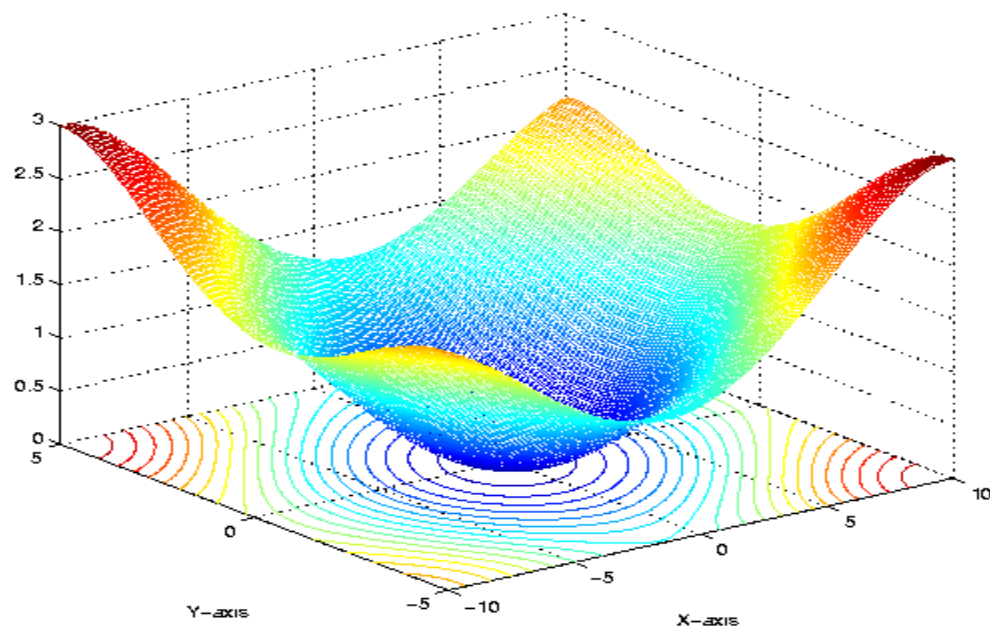
Min_Error → Max_Fitness

(which we look for) !!!

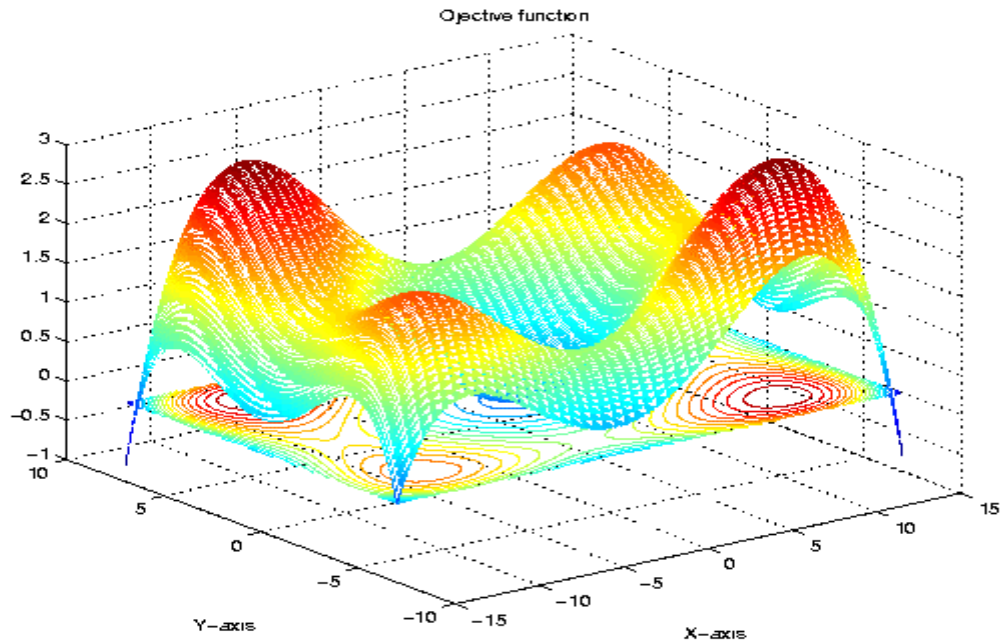
The Global Maximum



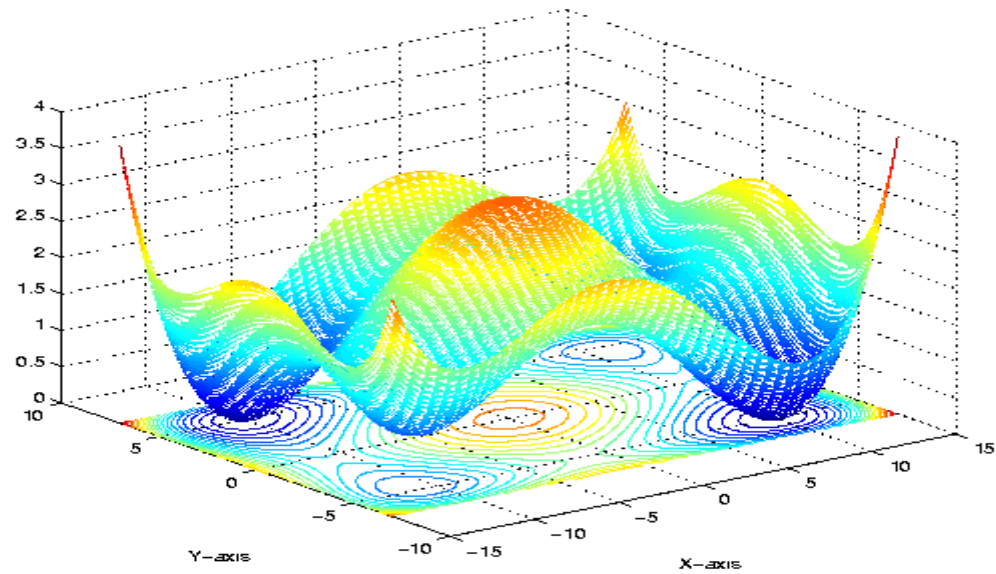
The Global Minimum

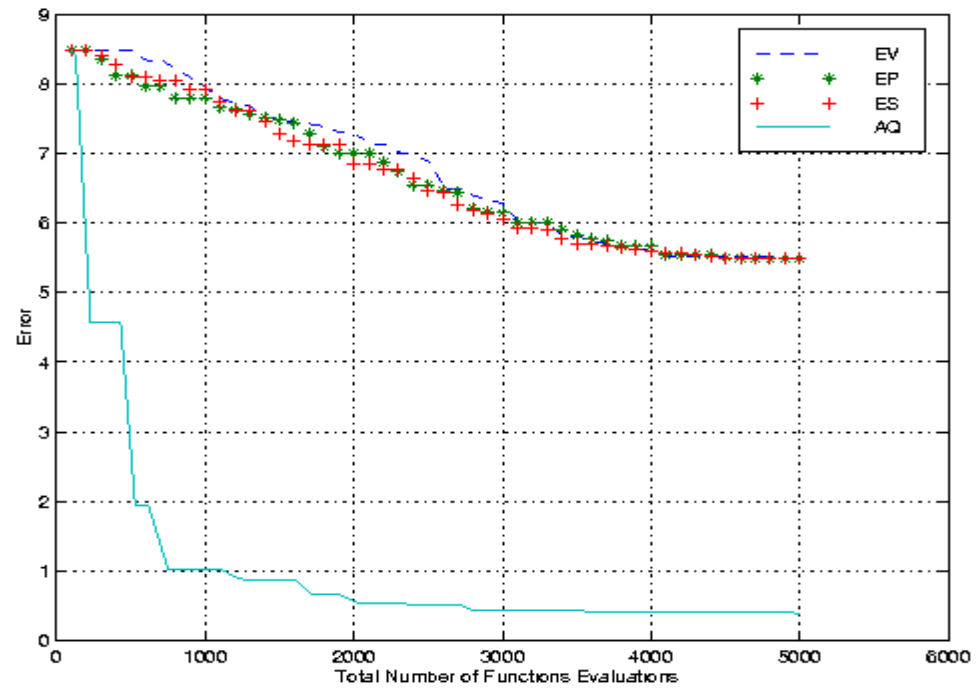


Multiple peaks: Different landscape

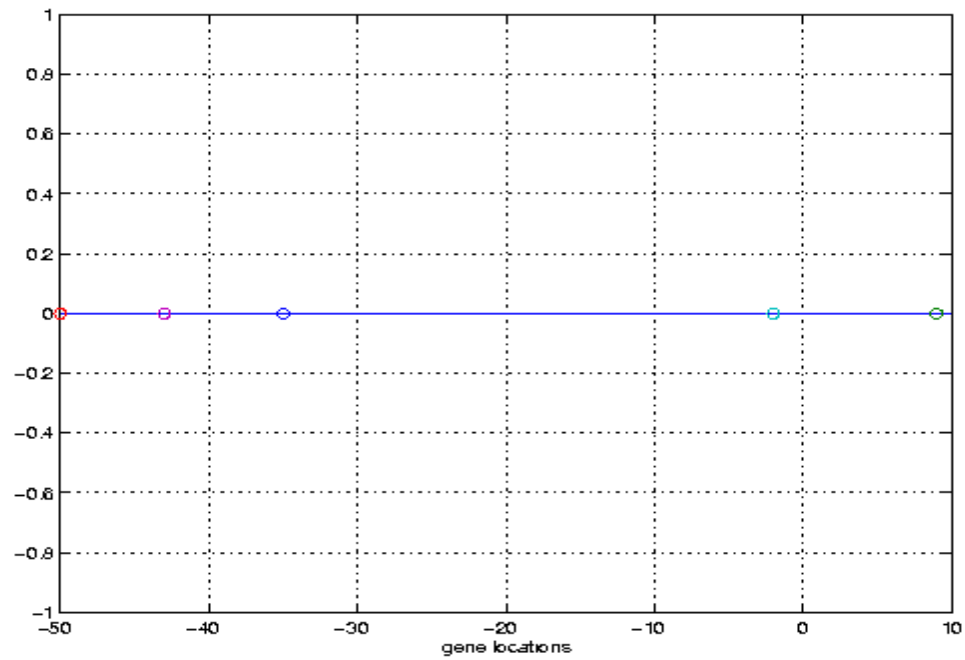


Multiple Optima (Minimum) Different landscape





Convergence Rate : Comparison



The Optimal Location of 5-genes

Algorithm	Generations	Error	The Optimal Gene locations				
EP	5000	5.48812	12.2196	15.9109	3.2243	1.7178	4.1080
EV	4910	5.51054	13.5779	12.7448	-1.3963	-0.7175	5.4734
ES	5000	5.48137	11.4165	14.9727	-4.0696	0.9398	3.0567
MEM	5010	0.387745	-35.0000	9.0000	-50.0000	-2.0000	-43.000

Results and Comparisons

Summary and Future Work:

1. Landscape has its own effect and should be chosen to include all possible solutions.
2. Knowing what you're looking for (Max/Min) and what the package will offer you, Then design your Fitness function .
3. MEM has more accurate Results than EV itself.
4. The code can handle up to 100-D and can be modified for higher.
5. The code easily Parallelize for reducing time complexity.