Benchmarking Memetic Feature Selection

Amir Esseghir, Gilles Goncalves And Yahia Slimani

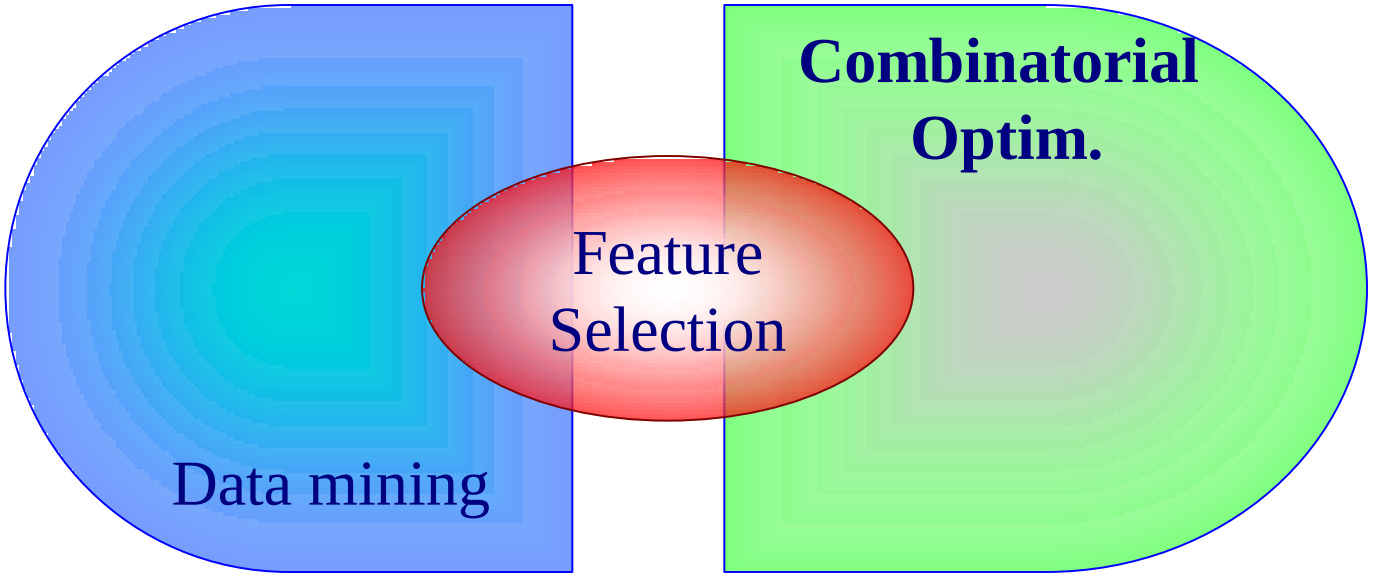
23 June 2010

HAIS-2010

Outline

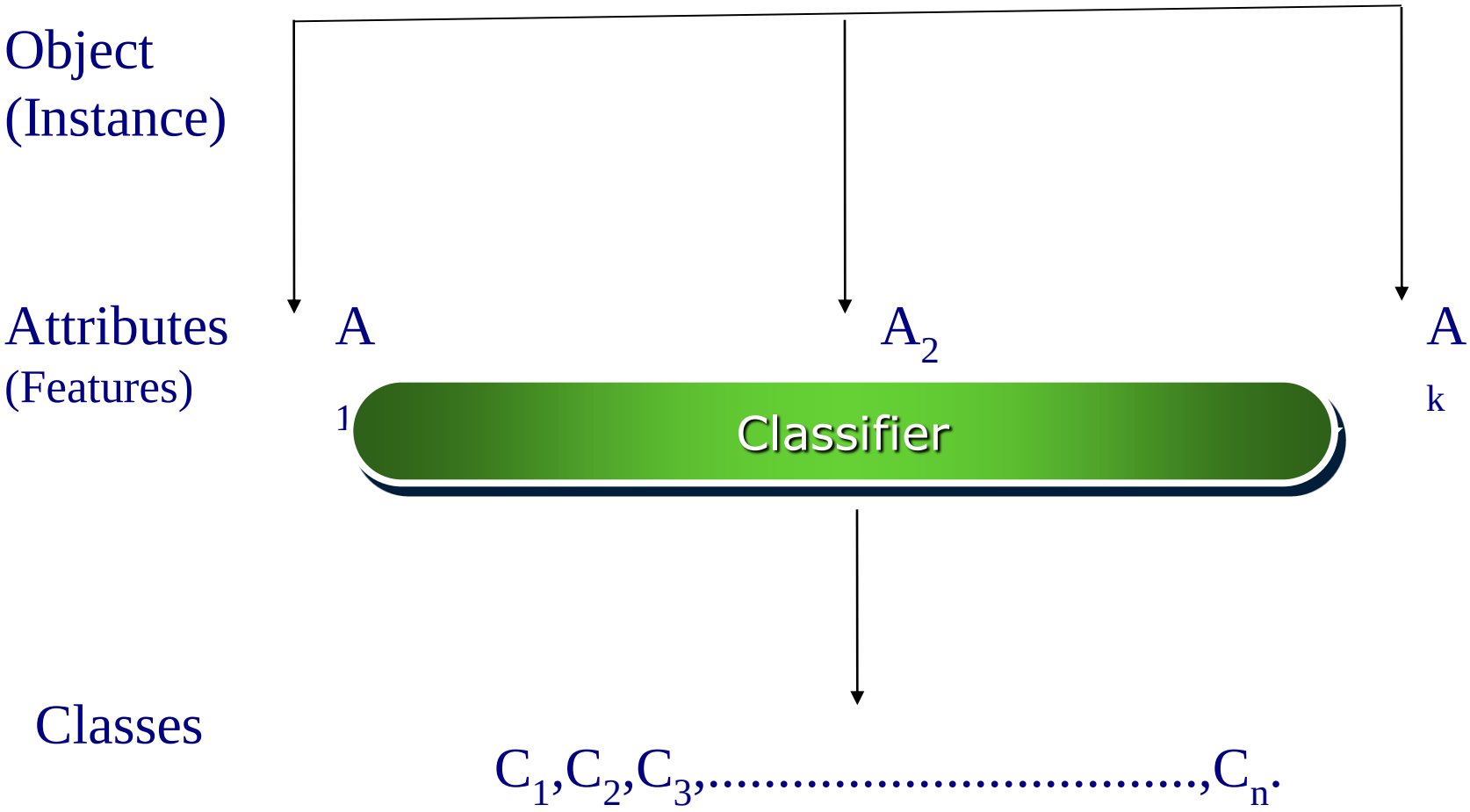
1. Introduction
2. Feature Selection
3. Proposed approach
4. Conclusion and persp.

Introduction



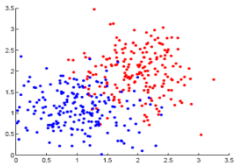
What are the best inputs that can improve classification accuracy ?

Classification



Feature Selection (FS)

Expert



Context

Q

How to select the more representative Features?

All → Select → Noisy, irrelevant, redundant

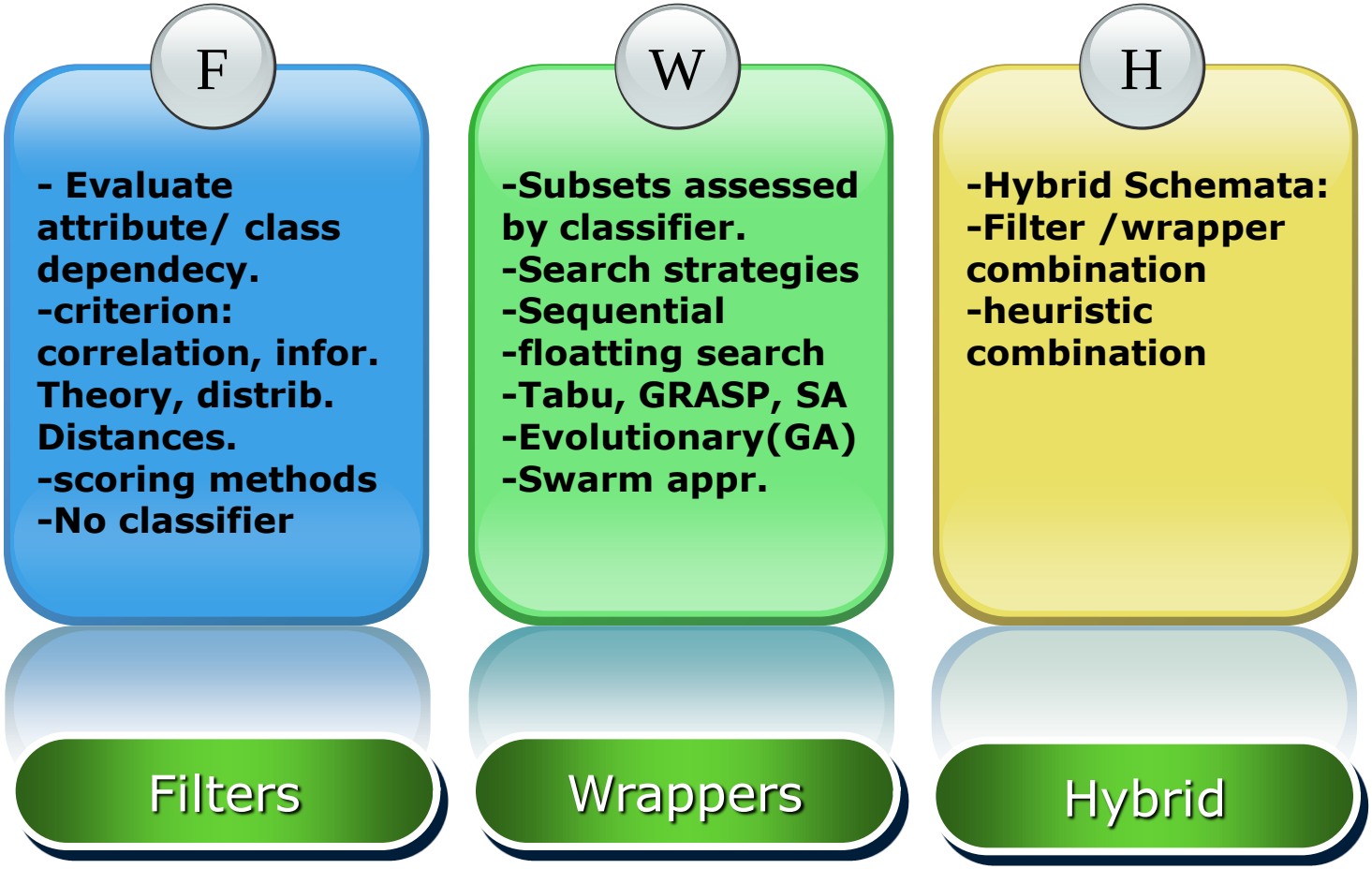
few → discard → Relevant

R

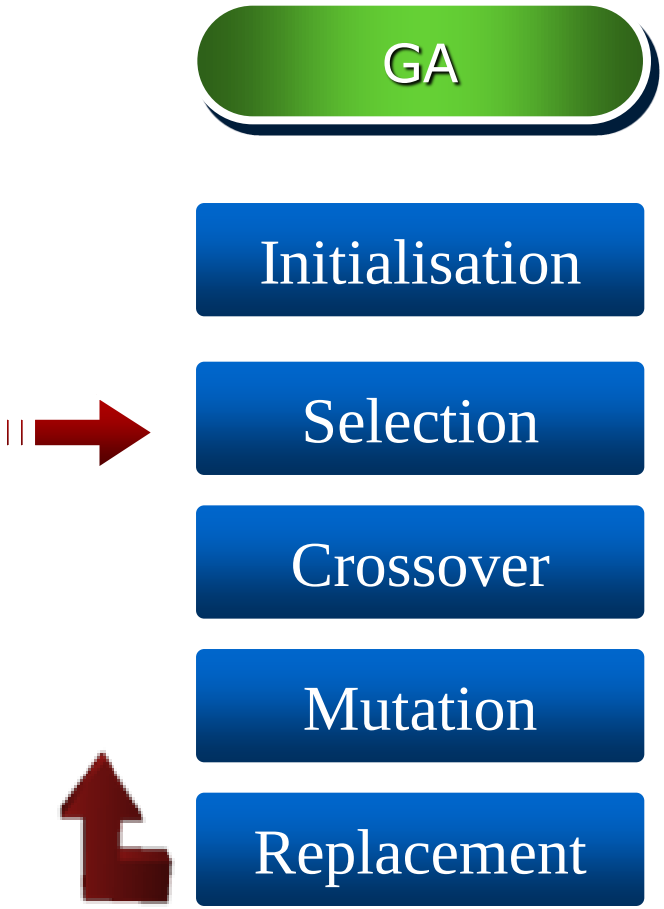
Exploring a search space of 2^n subsets?

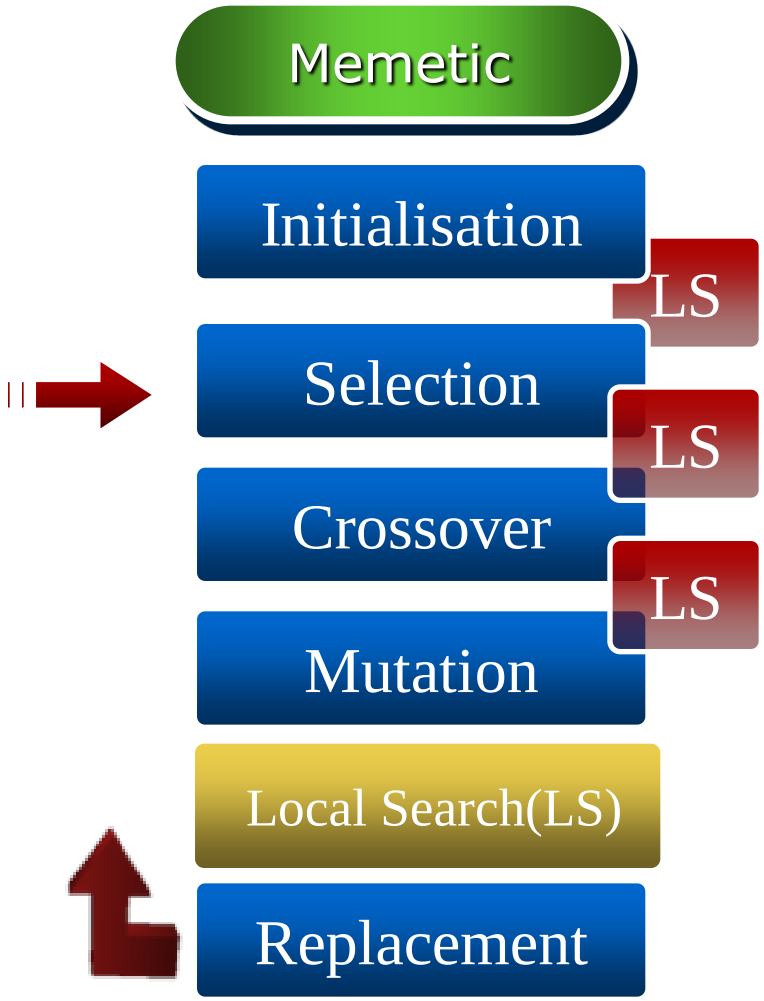
Combinatorial nature of the FS problem

Existing Approaches



GA vs Memetic Algorithms



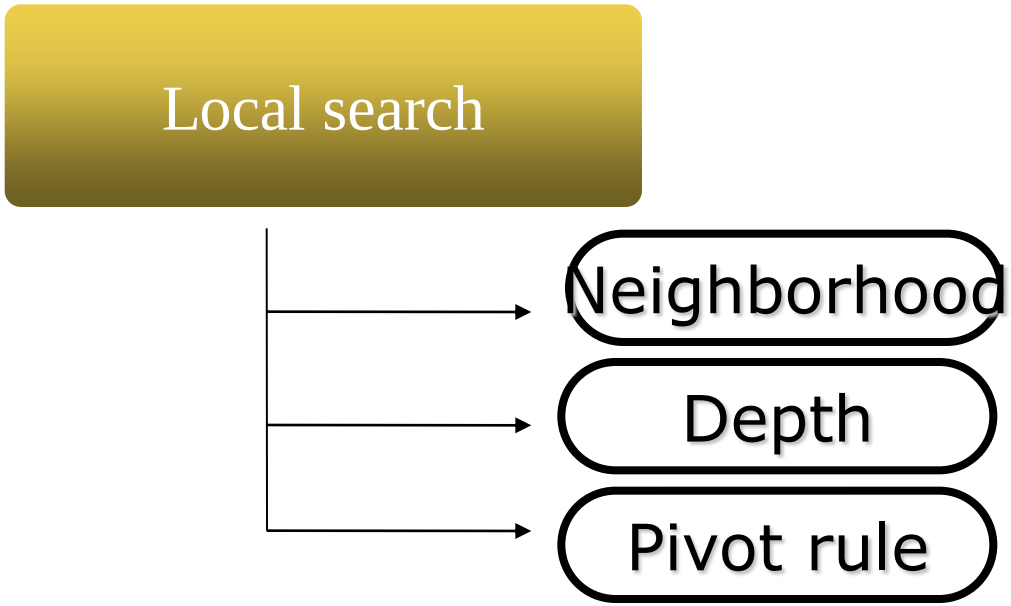


Why memetic search?

- Enhancing GA capabilities:
  - **exploration**: evolutionary Operators
  - **intensification**: Local search
- LS refines new solutions: exchanging them by improved ones.
- LS: ability to implement a specific problem Knowledge (meme).
  - Neigborgood exploration
- Numerous integration alternatives.
- Successful application of GA in FS modeling

Local Search

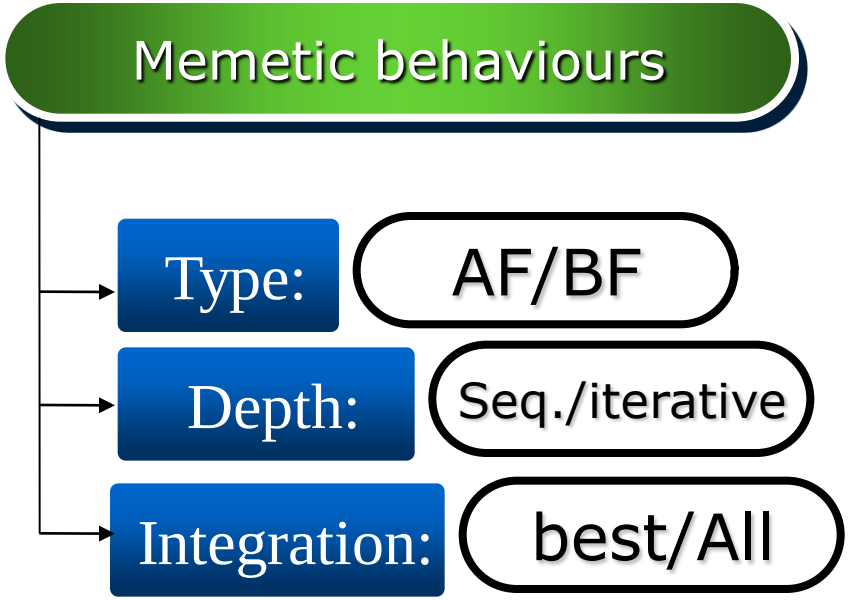
- **Idea**: improve the current solution by exploring its neighborhood



Proposed memetic Schemata

BitFlip:BF
-Add /remove a feature at a time
-Hamming distance= 1

Attrib. flip
- Replace an attribute by another attribute



Q **What is the more effective schema?**

Solution: binary representation
(0/1) Selected / Not selected

Experiments

- Data sets: 3 UCI benchmarks
  - Dimensions: [57..2000]
- Evaluation criterion:
  - Error rate(%)
  - Test methods: Hold out(Search), CV(validation)
- Experiments:
  - Run at least 10 times.
  - Reported results: Mean; St. Dev, *t*-Test

Results

		SpamBase		Arrhythmia		Colon cancer		
		Gain	Rank	Gain	Rank	Gain	Rank	
Seq.	BF(Best)	8,37%	8	1,93%	8	31,60%	4	
	AF(Best)	26,18%	3	9,45%	3	28,83%	5	
	BF(All)	12,12%	6	3,97%	6	26,69%	6	
	AF(All)	38,30%	2	13,89%	2	44,94%	2	
Iter.	BF(Best)	9,44%	7	1,98%	7	14,42%	7	
	AF(Best)	24,79%	4	6,48%	4	33,28%	3	
	BF(All)	14,91%	5	4,96%	5	13,34%	8	
	AF(All)	40,77%	1	17,21%	1	56,6%	1	

Results(2)

1

Global improvement

-All memetic schema enhances GA

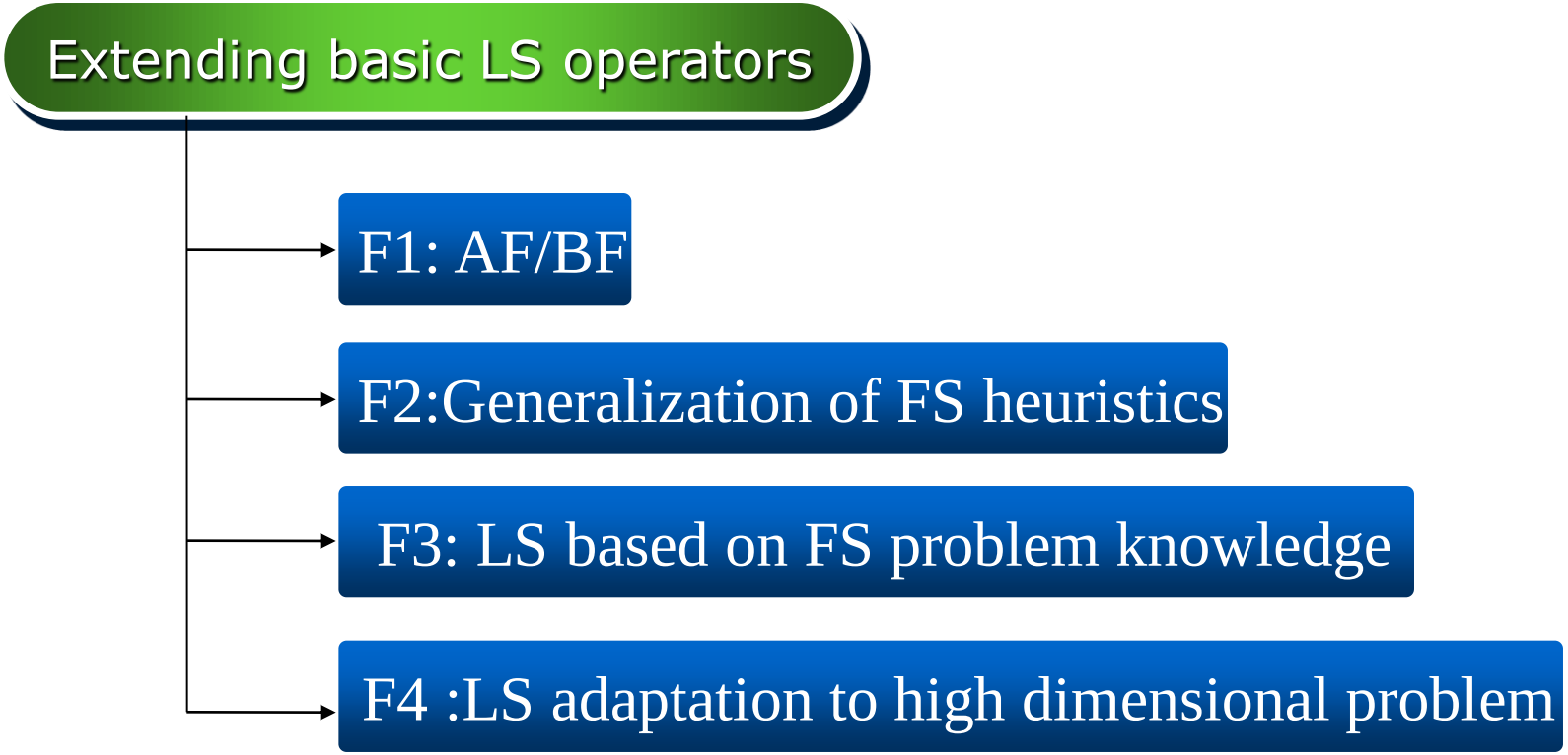
2

AF vs BF
-Suepriority of AF schema over BF
-Top 3

3

-Some sequential schemata outpermors iterative LS
-search improvements for large problem dimensionalities

Current Work



Conclusion

- FS problem is a multi-disciplinary reseach field,

and the combinatorial issue stills one of the more attractive.

- We investigate 8 memetic schemata.
  - Global improvement over GA
  - Some schemata are more intersting.
- Hybrid modeling is an effective way to tackle FS problems.

Perspectives

- Investigation of new Neighborhood structures.
- Study behavioral aspects of LS with differnt meta-heuristics
- Adaptation and application to high dimensional problems.
  - (Gene expression data).
  - Proteomics

Thank You !

Email: messeghir@sfr.fr

Questions?

Remarks!

UNIVERSITÉ D'ARTOIS

				VALIDATION ERROR%						
LS applied to GA		Measures	Fitness	ANN	NB	CPU (s)	# Attrib.	# Eval	Gain%GA	RANK
No LS (GA)		Mean:	9,32%	10,74%	7,81%	14774,13	15,04	1089		
		Sd:	0,95%	1,45%	1,10%	14759,99	3,15	0		
		t-test	0	0	0	0	0	-		
SEQ	BitFlip(Best)	Mean:	8,54%	11,10%	8,24%	35672,26	13,87	3089	8,37%	8
		Sd:	1,02%	2,04%	1,27%	36614,23	2,94	0	-	
		t-test	-18,92	3,71	4,95	22,08	-12,67	-	-	
	AttribFlip(Best)	Mean:	6,88%	9,61%	7,26%	32705,52	15,96	3089	26,18%	3
		Sd:	1,15%	2,25%	1,49%	28955,4	4,88	0	-	
		t-test	-37,43	-8,41	-5,85	18,74	5,31	-	-	
	BitFlip(all)	Mean:	8,19%	10,58%	8,23%	91471,7	15,45	11289	12,12%	6
		Sd:	0,98%	1,98%	1,50%	40881,06	3,2	0	-	
		t-test	-29,95	-1,7	7,22	31,64	1,49	-	-	
	AttribFlip(all)	Mean:	5,75%	8,95%	6,73%	95083,75	17,6	11289	38,30%	2
		Sd:	0,69%	1,93%	1,50%	45427,35	4,47	0	-	
		t-test	-73,7	-10,69	-23,21	44,9	85,05	-	-	
Iterative	BitFlip(Best)	Mean:	8,44%	9,98%	7,49%	36313,91	15,61	3171,17	9,44%	7
		Sd:	1,17%	1,75%	1,56%	34137,85	4,38	32,04	-	
		t-test	-20,55	-6,81	-4,72	54,25	8,08	-	-	
	AttribFlip(Best)	Mean:	7,01%	9,62%	7,27%	35944,87	14,87	3209,43	24,79%	4
		Sd:	1,01%	2,10%	1,44%	38229,41	3,63	33,37	-	
		t-test	-45,78	-11,65	-9,22	23,64	-3,54	-	-	
	BitFlip(all)	Mean:	7,93%	10,35%	7,70%	128806,75	15,35	14180,5	14,91%	5
		Sd:	0,93%	1,50%	1,18%	112289,05	3,1	418,7	-	
		t-test	-54,13	-3,5	-2,06	55,51	2,61	-	-	
	AttribFlip(all)	Mean:	5,52%	7,69%	6,30%	106967,85	17,6	12911	40,77%	1
		Sd:	0,49%	1,22%	1,31%	49985,26	2,91	137,52	-	
		t-test	-151,28	-31,84	-22,96	51,35	85,05	-	-	

Table 2. Data set: SpamBase (57 Attrib.)

				VALIDATION ERROR%						
LS applied to GA		Measures	Fitness	ANN	NB	CPU (s)	# Attrib.	# Eval	Gain%GA	RANK
No LS (GA)		Mean:	6,52%	6,58%	12,19%	31079,73	23,18	3032		
		Sd:	2,92%	3,14%	4,41%	22980,17	7,45	0		
		t-test	0	0	0	0	0	-		
SEQ	BitFlip(Best)	Mean:	4,46%	6,28%	10,32%	38454,95	20,55	5032	31,60%	4
		Sd:	2,05%	2,97%	4,59%	30297,6	6,6	0	-	
		t-test	-20,87	-2,11	-7,8	5,73	-5,45	-	-	
	AttribFlip(Best)	Mean:	4,64%	7,06%	10,06%	41058,14	21,41	5032	28,83%	5
		Sd:	2,25%	2,96%	3,23%	37162,49	8,55	0	-	
		t-test	-14,38	5,19	-15,01	7,69	-4,35	-	-	
	BitFlip(all)	Mean:	4,78%	5,96%	11,49%	200004,95	22,05	13232	26,69%	6
		Sd:	2,96%	2,98%	2,96%	102160,5	6,34	0	-	
		t-test	-17,19	-5,06	-4,23	22,81	-3,15	-	-	
	AttribFlip(all)	Mean:	3,59%	5,74%	11,49%	198320,55	22,05	13232	44,94%	2
		Sd:	1,62%	2,59%	3,12%	102691,3	6,78	0	-	
		t-test	-24,16	-4,14	-8,47	40,91	-3,15	-	-	
Iterative	BitFlip(Best)	Mean:	5,58%	6,29%	11,47%	39703,43	20,91	5073,74	14,42%	7
		Sd:	2,25%	3,17%	4,39%	30596,31	7,36	30,84	-	
		t-test	-5,31	-1,42	-1,37	12,2	-3,23	-	-	
	AttribFlip(Best)	Mean:	4,35%	6,94%	11,84%	39066,78	20,09	5068,09	33,28%	3
		Sd:	2,17%	2,81%	3,44%	31127,5	6,31	10,33	-	
		t-test	-15,9	2,48	-3,93	6,94	-6,98	-	-	
	BitFlip(all)	Mean:	5,65%	7,02%	11,38%	249361,3	18,45	15199,5	13,34%	8
		Sd:	2,86%	3,59%	3,86%	142791,7	6,51	711,49	-	
		t-test	-4,35	1,71	-5,02	57,21	-12,59	-	-	
	AttribFlip(all)	Mean:	2,83%	6,28%	9,89%	213912,1	20,2	13790	56,60%	1
		Sd:	1,88%	3,27%	3,86%	122469,18	7,03	66,38	-	
		t-test	-21,43	-1,33	-23,24	29,99	-6,06	-	-	

Table 3. Data set: Colon cancer (2000 Attrib.)

				VALIDATION ERROR%						
LS applied to GA		Measures	Fitness	ANN	NB	CPU (s)	# Attrib.	# Eval	Gain%GA	RANK
No LS (GA)		Mean:	17,14%	13,38%	17,33%	158683,53	85,79	1311		
		Sd:	0,90%	1,13%	1,31%	72380,76	17,63	0		
		t-test	0	0	0	0	0	-		
SEQ	BitFlip(Best)	Mean:	16,81%	13,86%	17,09%	490897,85	93,7	3311	1,93%	8
		Sd:	0,91%	1,11%	1,30%	222623,21	14,37	0	-	
		t-test	-23,47	6,69	-2	26,47	11,23	-	-	
	AttribFlip(Best)	Mean:	15,52%	14,03%	17,29%	519275,2	92,25	3311	9,45%	3
		Sd:	1,08%	1,67%	1,59%	347110,37	46,42	0	-	
		t-test	-65,25	2,69	-0,42	17,8	2,1	-	-	
	BitFlip(all)	Mean:	16,46%	14,02%	17,36%	1785624,68	82,74	11511	3,97%	6
		Sd:	1,24%	1,91%	1,75%	898120,57	20,56	0	-	
		t-test	-36,87	7,77	0,22	37,76	-3,46	-	-	
	AttribFlip(all)	Mean:	14,76%	14,22%	16,83%	1892140,75	89,4	11511	13,89%	2
		Sd:	0,80%	2,25%	1,30%	1141857,26	42,76	0	-	
		t-test	-71,41	7,94	-5,5	33,02	11,46	-	-	
Iterative	BitFlip(Best)	Mean:	16,80%	14,40%	16,59%	525793,95	91,2	3516	1,98%	7
		Sd:	0,91%	1,50%	1,67%	290212,63	26,49	55,2	-	
		t-test	-16,38	11,1	-8,17	43,48	5,22	-	-	
	AttribFlip(Best)	Mean:	16,03%	14,16%	16,76%	468881,85	93,9	3512,5	6,48%	4
		Sd:	1,03%	1,78%	1,87%	315436,97	56,97	40,3	-	
		t-test	-26,71	5,14	-6,36	28,42	1,19	-	-	
	BitFlip(all)	Mean:	16,29%	13,79%	16,67%	2591560,62	83,1	15471,95	4,96%	5
		Sd:	1,15%	1,50%	1,52%	1343356,45	22,25	600,1		
		t-test	-24,61	5,56	-7,31	51,56	-8,75	-	-	
	AttribFlip(all)	Mean:	14,19%	14,91%	16,61%	1384673,95	51,65	13731,5	17,21%	1
		Sd:	0,92%	1,30%	1,38%	763942,71	32,57	272,29	-	
		t-test	-138,18	20,26	-7,89	172,14	-11,26	-	-	

Table 4. Data set: Arrhythmia (279 Attrib.)

LS-Operator		Order of Complexity	Parameters
SEQ	BitFlip(Best)	$\Theta(N)$	N: number of features m : mating pool size
	AttribFlip(Best)	$\Theta(N^2)$	
	BitFlip(all)	$\Theta(N.m)$	
	AttribFlip(all)	$\Theta(N^2.m)$	
Iterative	BitFlip(Best)	$\Theta(N.d)$	d: local search depth
	AttribFlip(Best)	$\Theta(N^2.d)$	
	BitFlip(all)	$\Theta(N.m.d)$	
	AttribFlip(all)	$\Theta(N^2.m.d)$	

Table 1. Complexity of local search operators


LOG
O


LOG
O

Diagram

1

-Boost existing FS approaches
-explore new strategies (ACO - PSO)

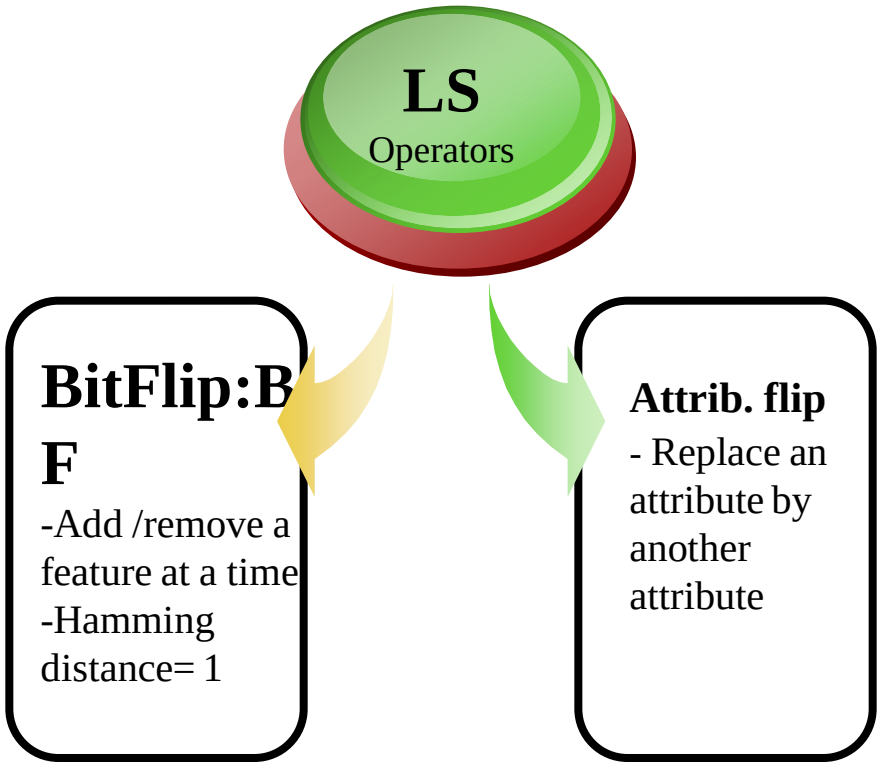
2

-Distributed version
Of retained model
-study dynamics and collaboration behavior

3

-Optimize Embedded Strategies
PSO/ACO
SVM-RF

Proposed memetic Schemata



Memetic behaviours

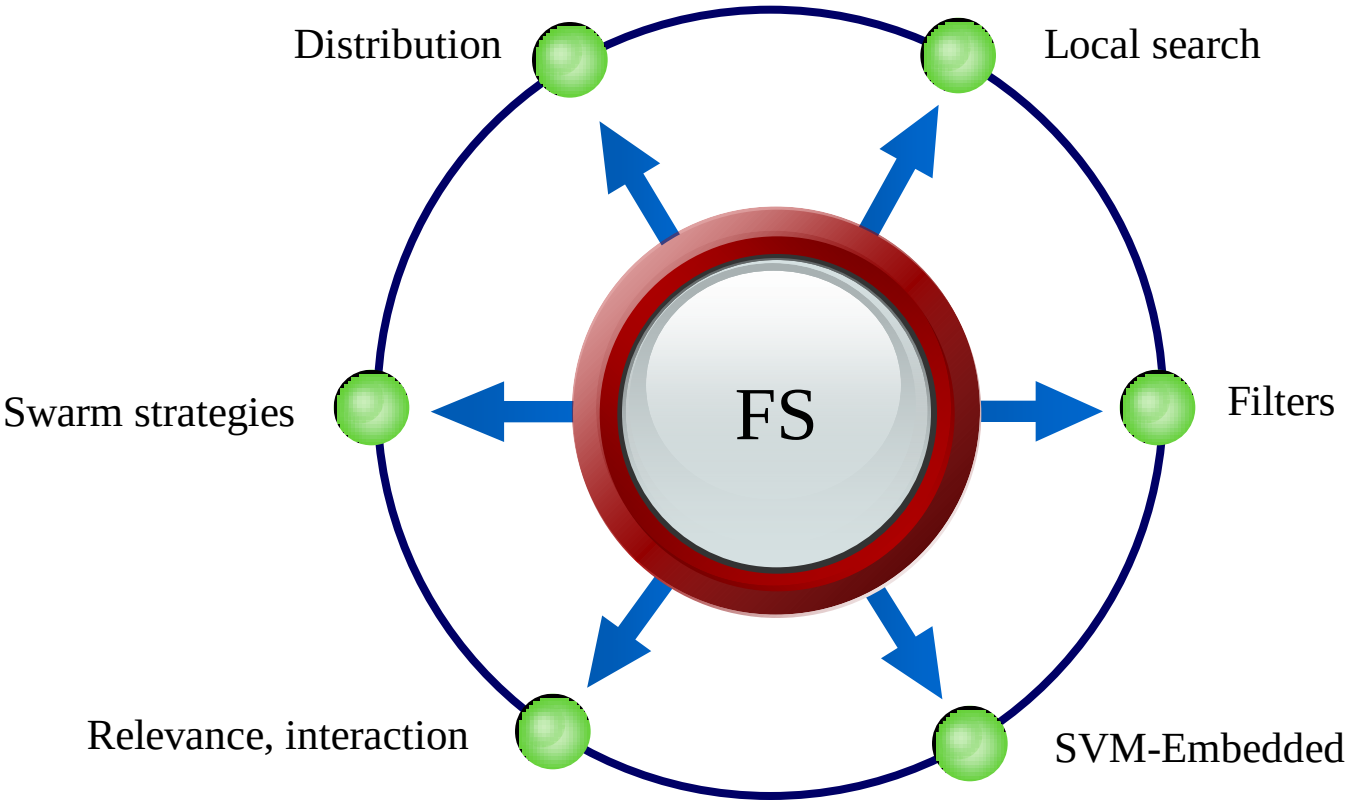
Type:

Application

Integration:

Solution: binary representation
(0/1) Selected / Not selected

Research directions



Diagram

1

-Boost existing FS approaches
-explore new strategies (ACO - PSO)

2

-Distributed version
Of retained model
-study dynamics and collaboration behavior

3

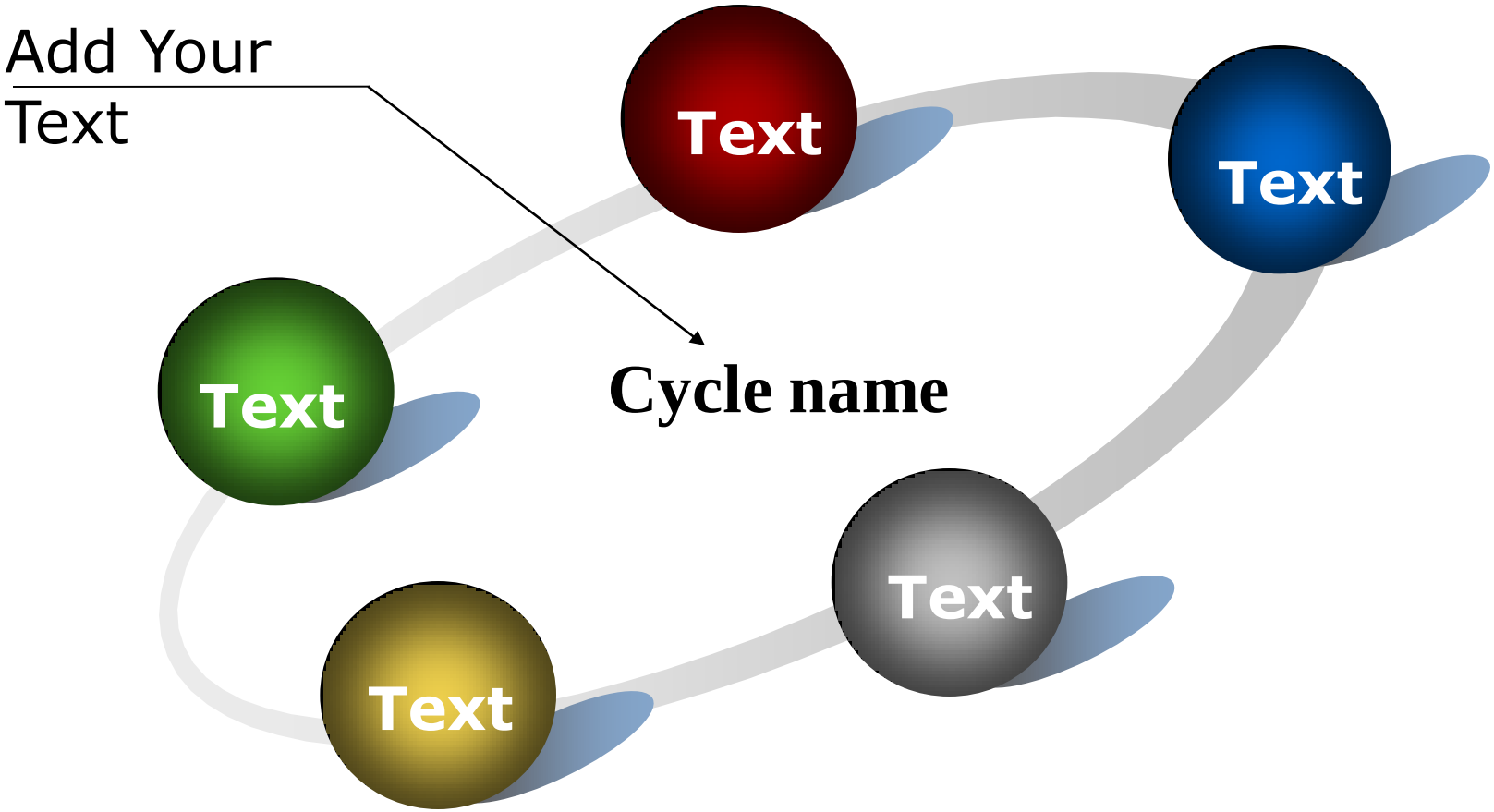
-Optimize Embedded Strategies
PSO/ACO
SVM-RF

Diagram

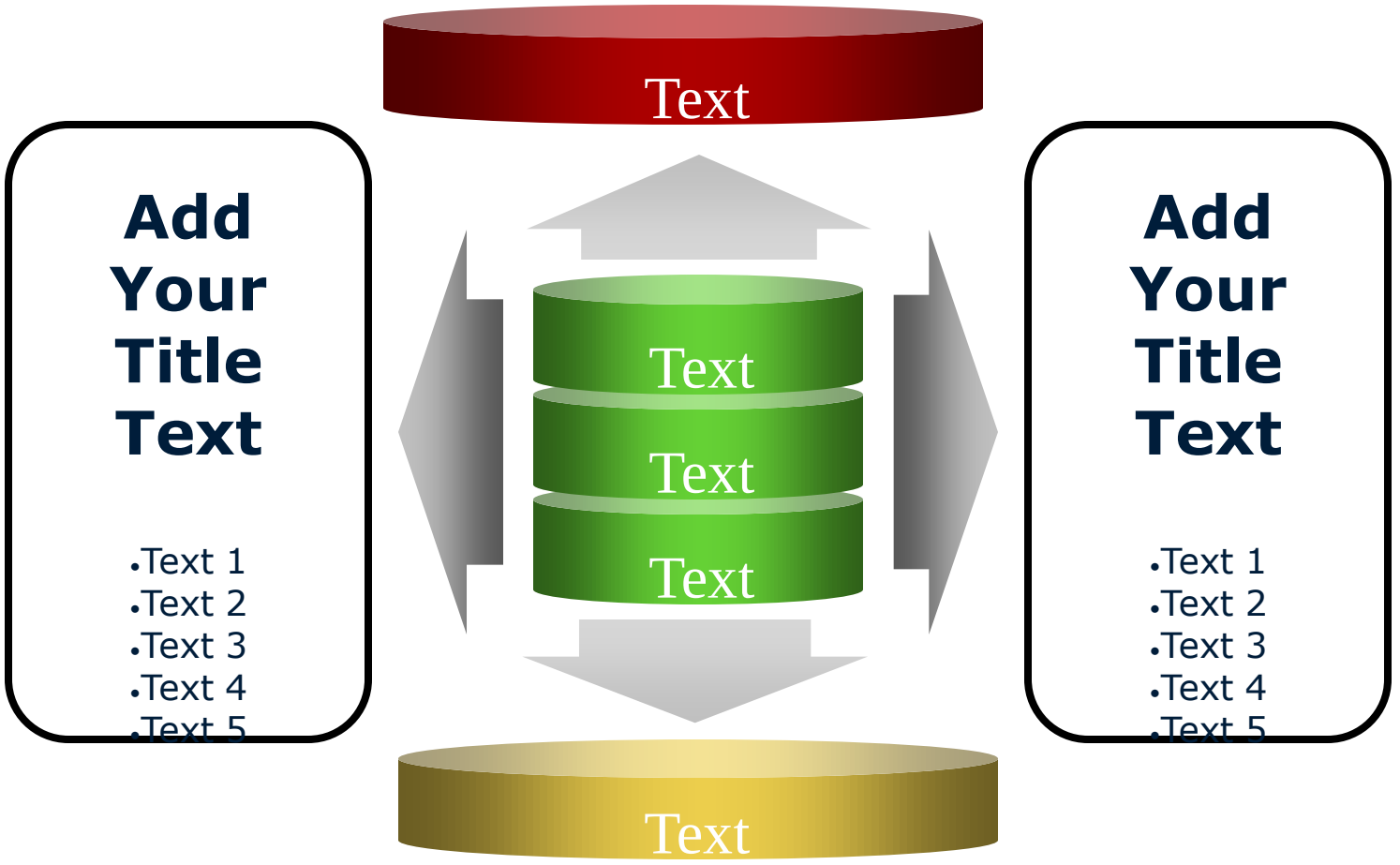
ThemeGallery is a Design Digital Content & Contents mall developed by Guild Design Inc.

ThemeGallery is a Design Digital Content & Contents mall developed by Guild Design Inc.

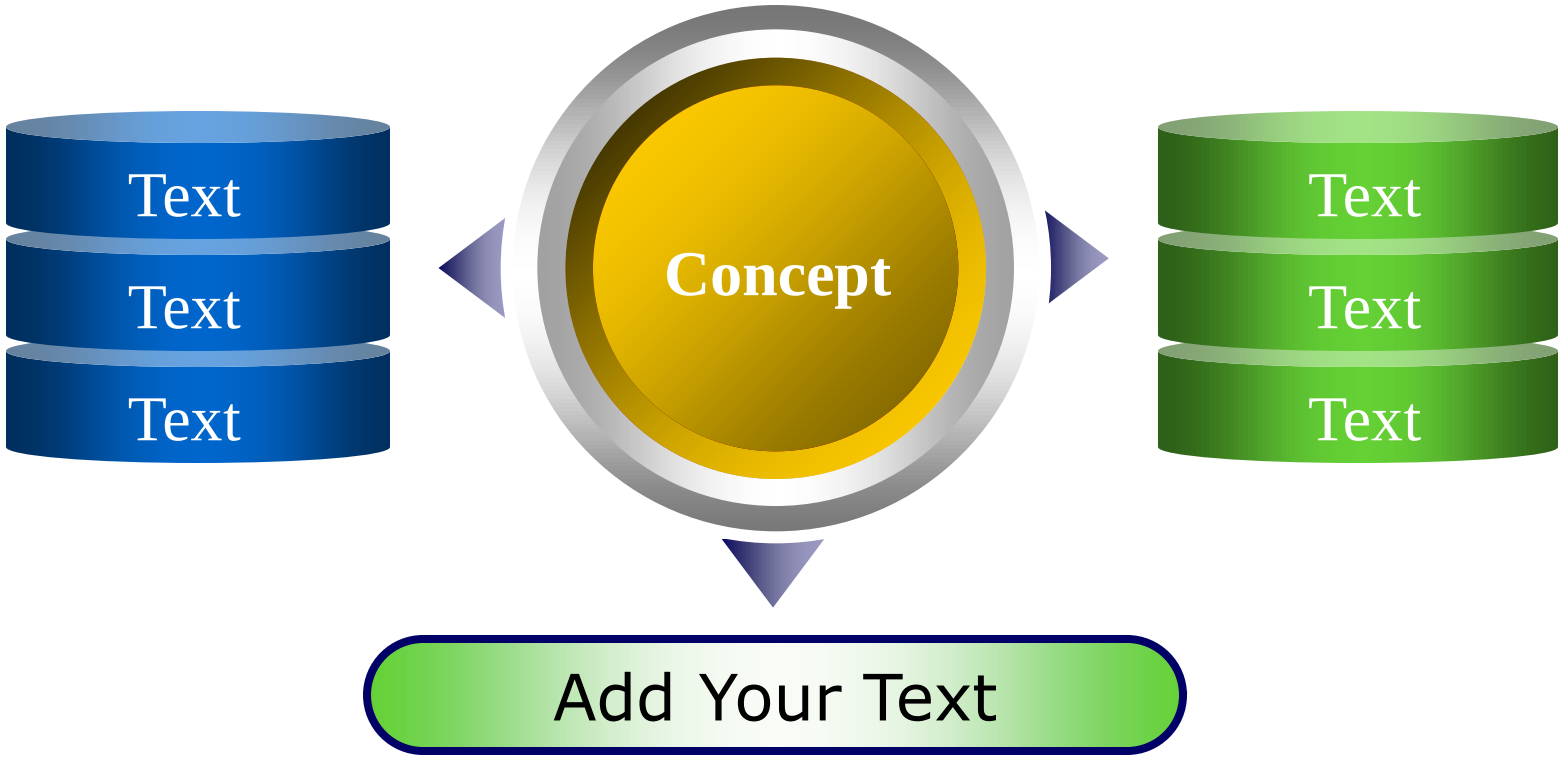
Cycle Diagram



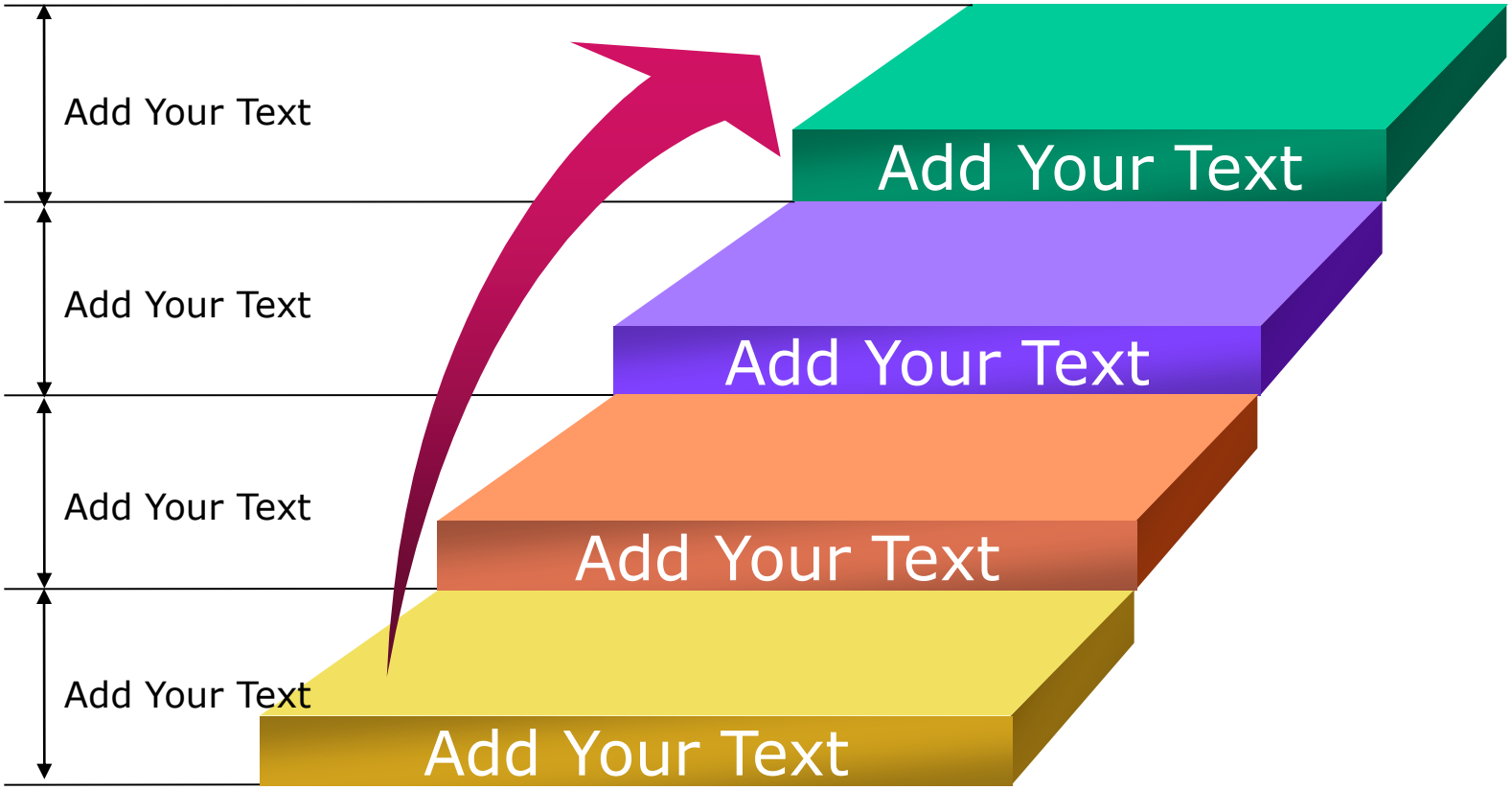
Diagram


Text
Add Your Title Text
•Text 1
•Text 2
•Text 3
•Text 4
•Text 5
Text
Text
Text
Add Your Title Text
•Text 1
•Text 2
•Text 3
•Text 4
•Text 5
Text

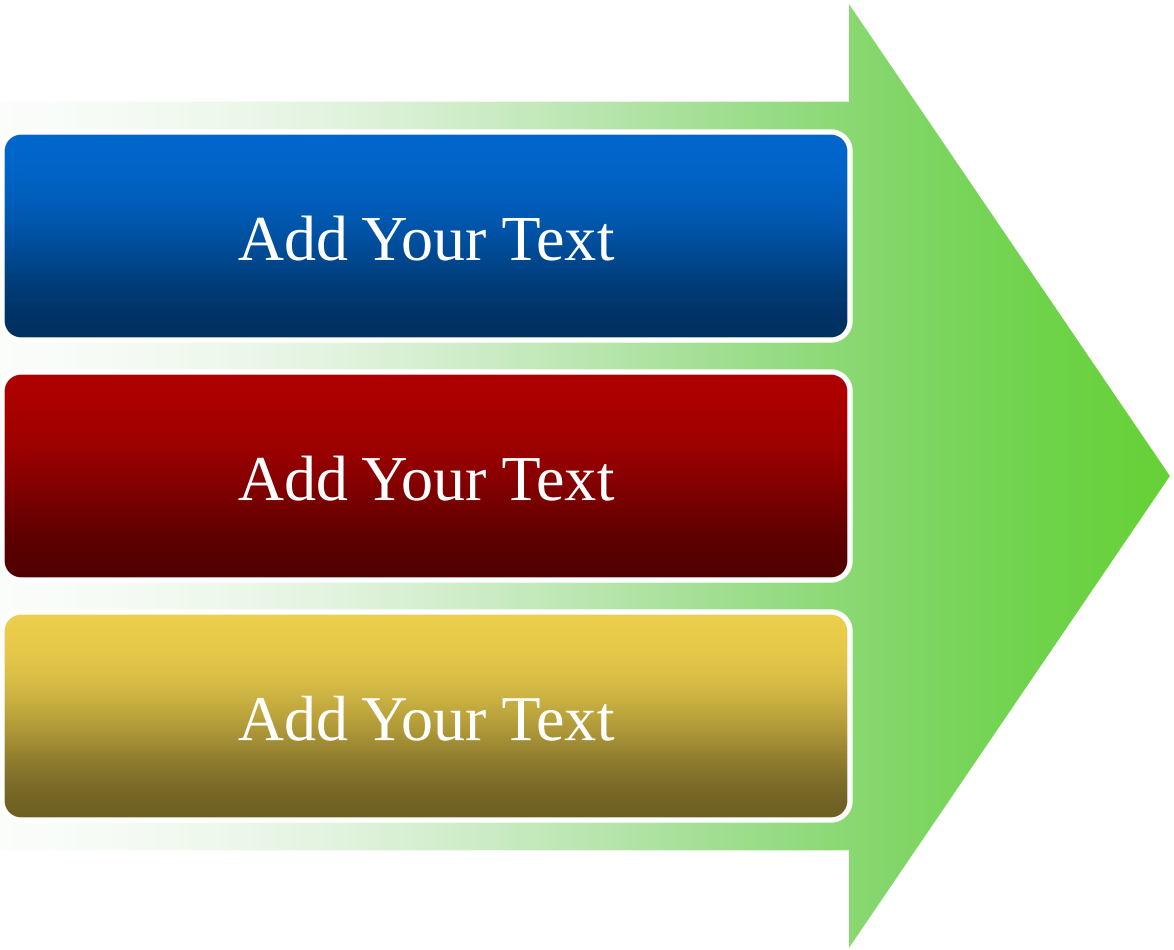
Diagram

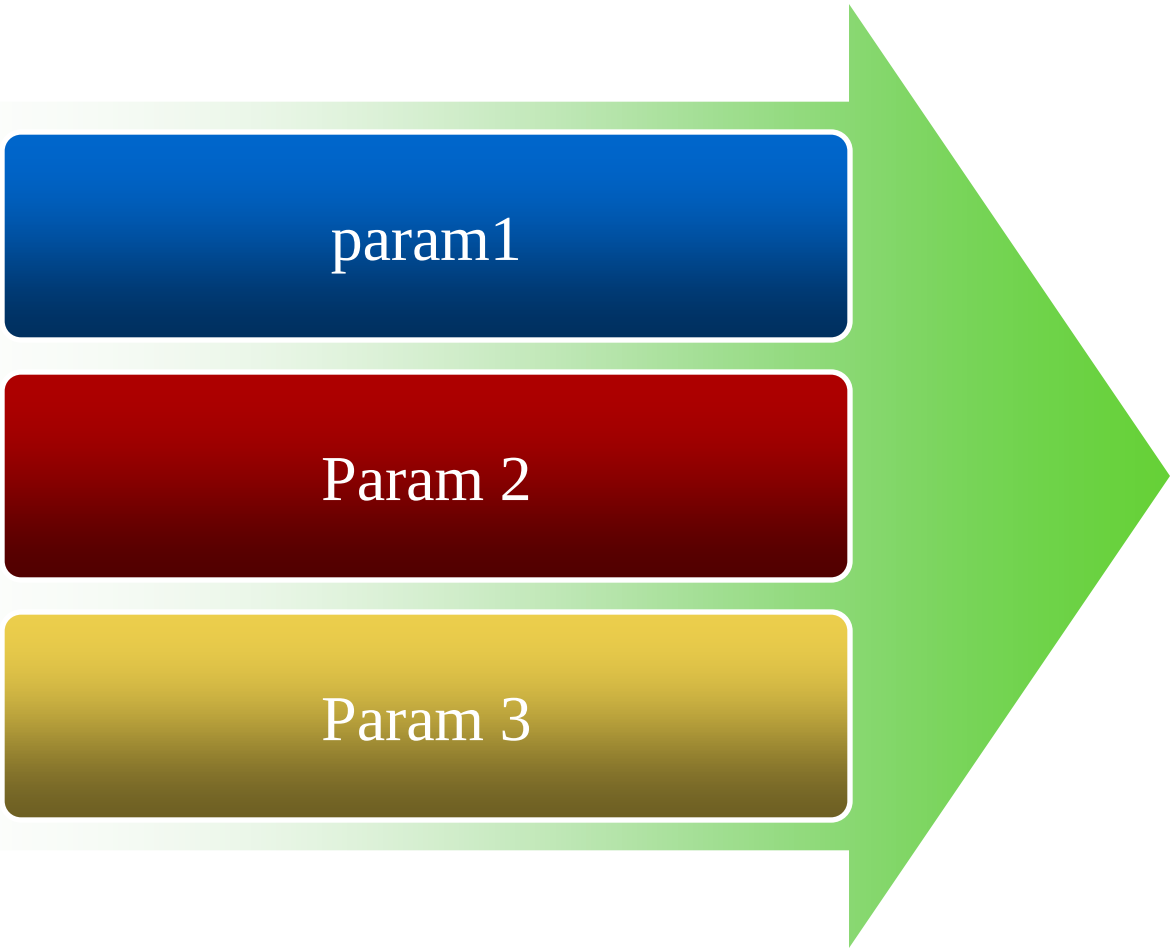


Diagram


Add Your Text
Add Your Text
Add Your Text
Add Your Text
Add Your Text
Add Your Text
Add Your Text
Add Your Text

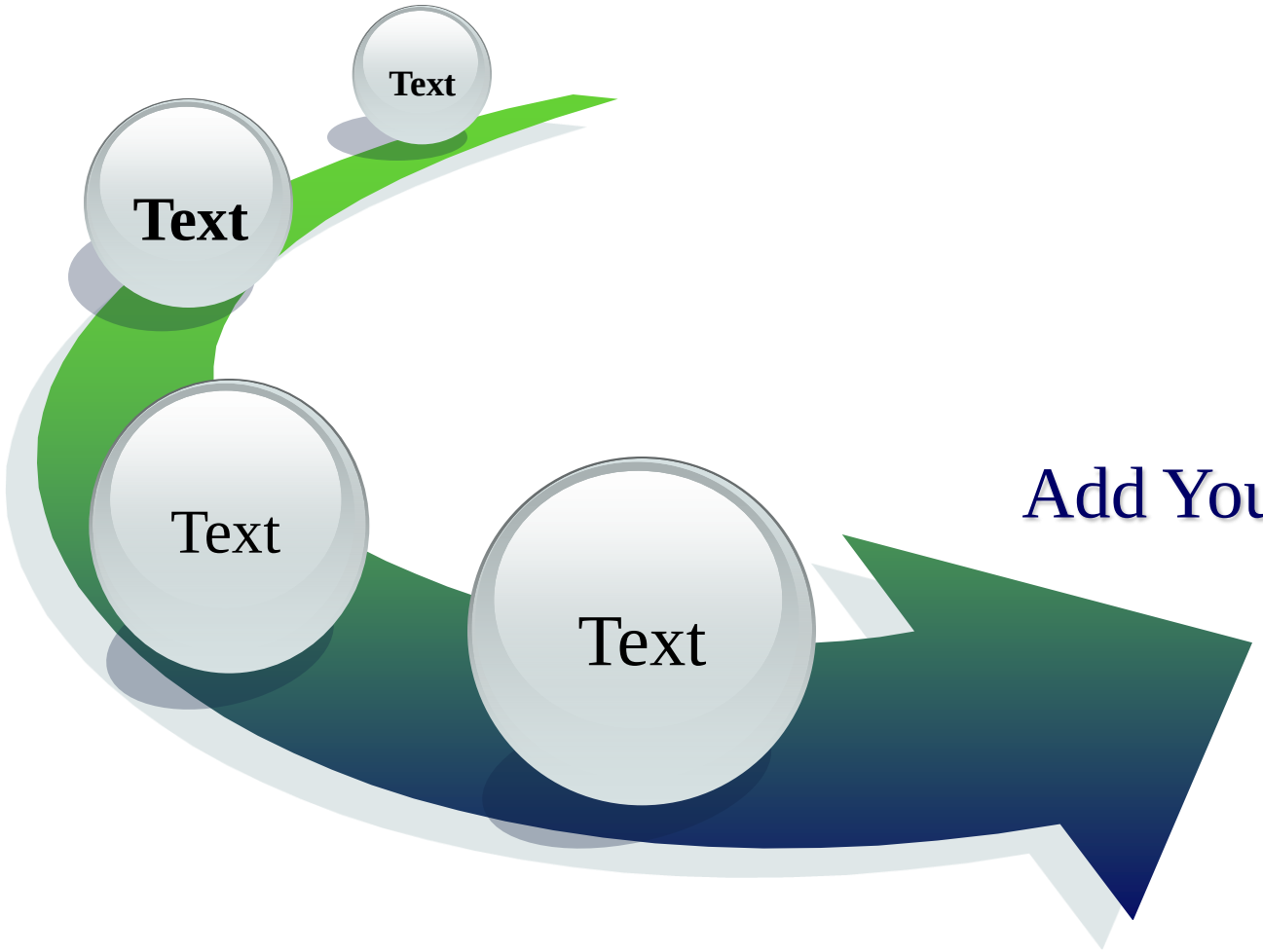
Diagram



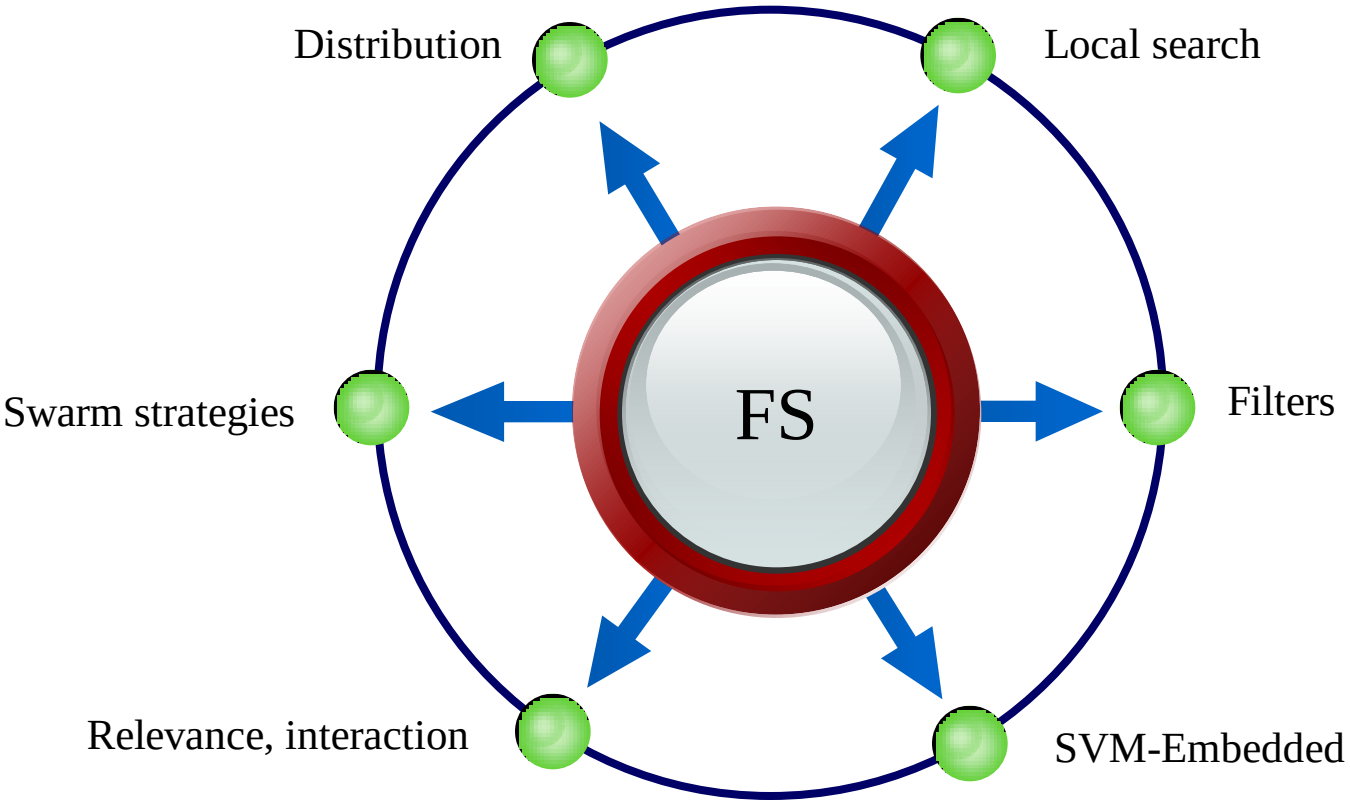

param1
Param 2
Param 3

Diagram

Diagram



Research directions



Diagram

1

ThemeGallery is a Design Digital Content & Contents mall developed by Guild Design Inc.

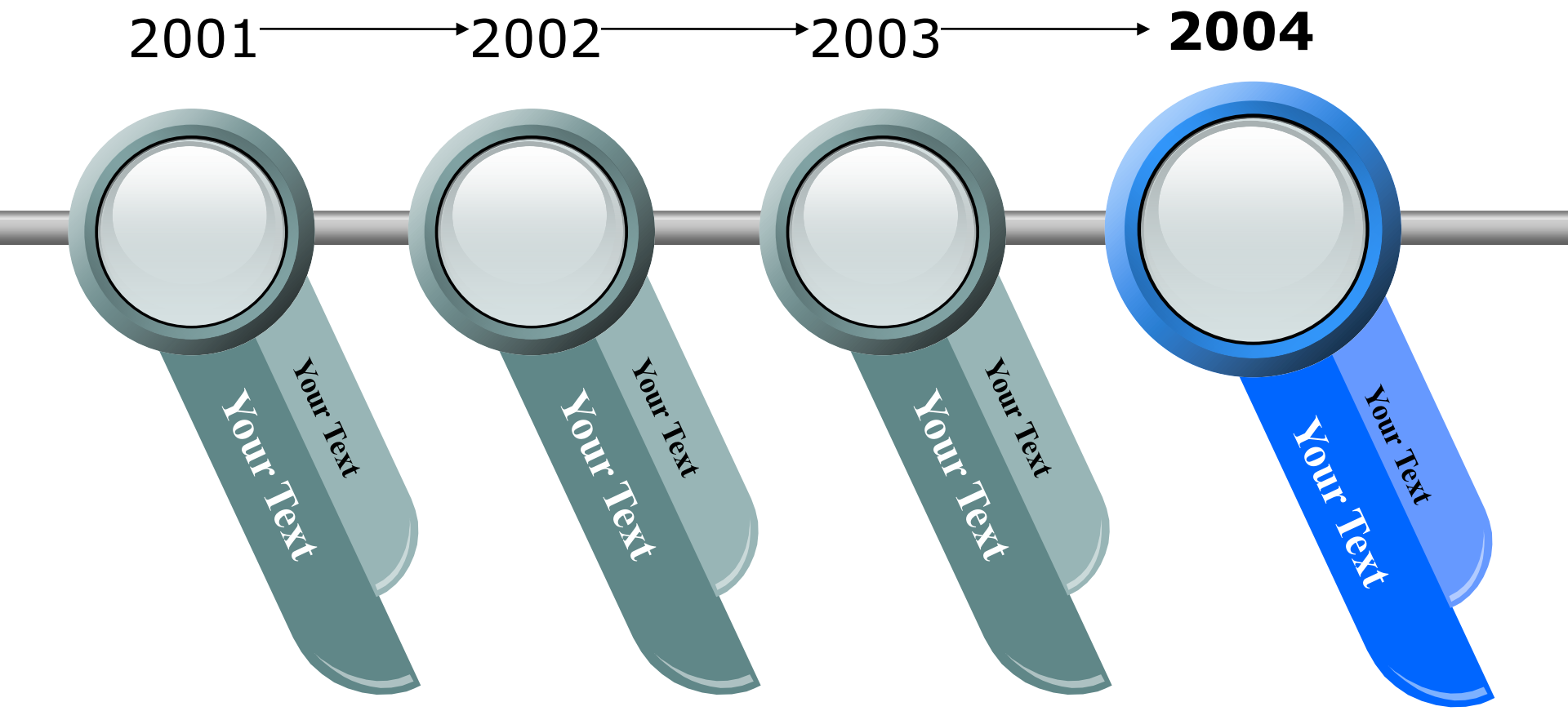
2

ThemeGallery is a Design Digital Content & Contents mall developed by Guild Design Inc.

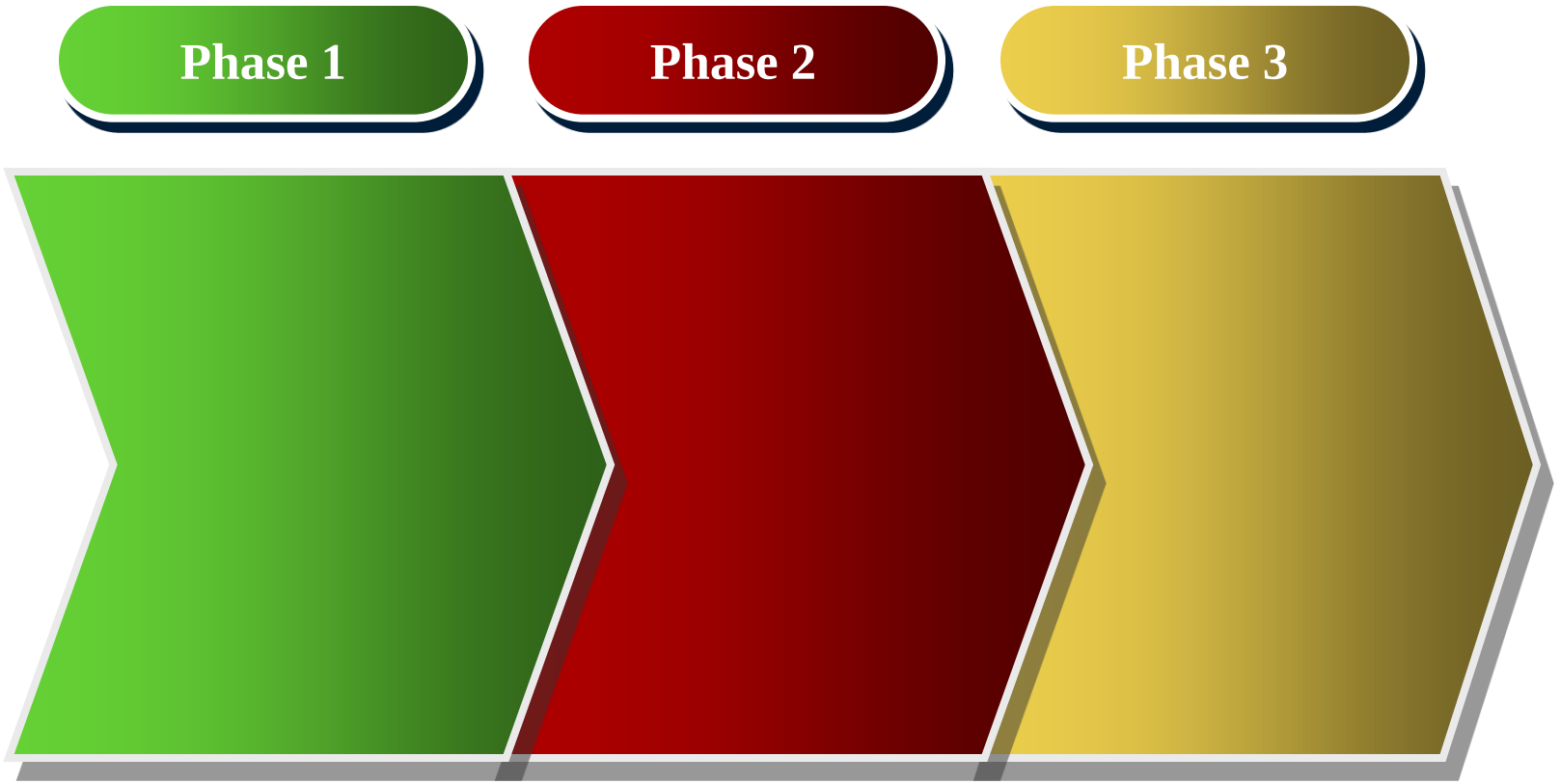
3

ThemeGallery is a Design Digital Content & Contents mall developed by Guild Design Inc.

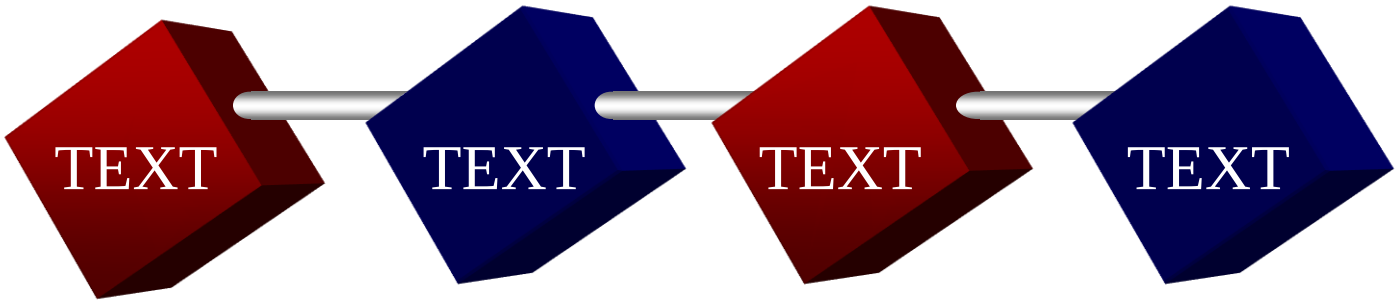
Diagram


2001
2002
2003
2004
Your Text
Your Text
Your Text
Your Text
Your Text
Your Text
Your Text
Your Text

Progress Diagram



Block Diagram

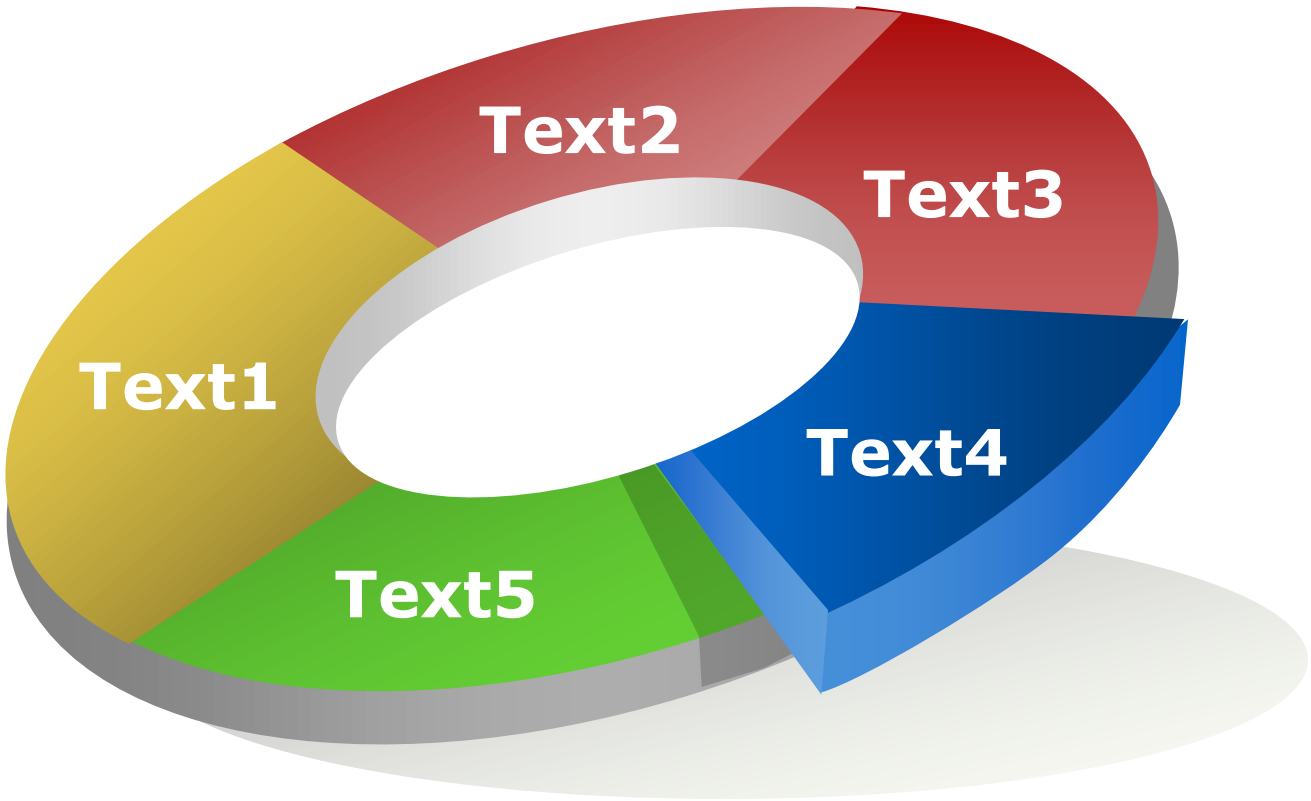


TEXT	TEXT	TEXT	TEXT

Table

	TEXT	TEXT	TEXT	TEXT	TEXT
<;;>					
Title B					
Title C					
Title D					
Title E					
Title E					
Title F					

3-D Pie Chart



Marketing Diagram

Add Your Text

Thank You !

Questions?

Remarks!

LOG
O