SemCluster:

Semantic Clustering of Programming Assignments

David Perry, Dohyeong Kim, Roopsha Samanta, Xiangyu Zhang

PURDUE UNIVERSITY

It seems like Everyone wants to be a programmer

Programming for Everybody (Getting Started with Python)

657,068

An Introduction to Interactive Programming in Python

581,043

Introduction to Computer Science

515,476

Learn to Programming: The Fundamentals

198,566

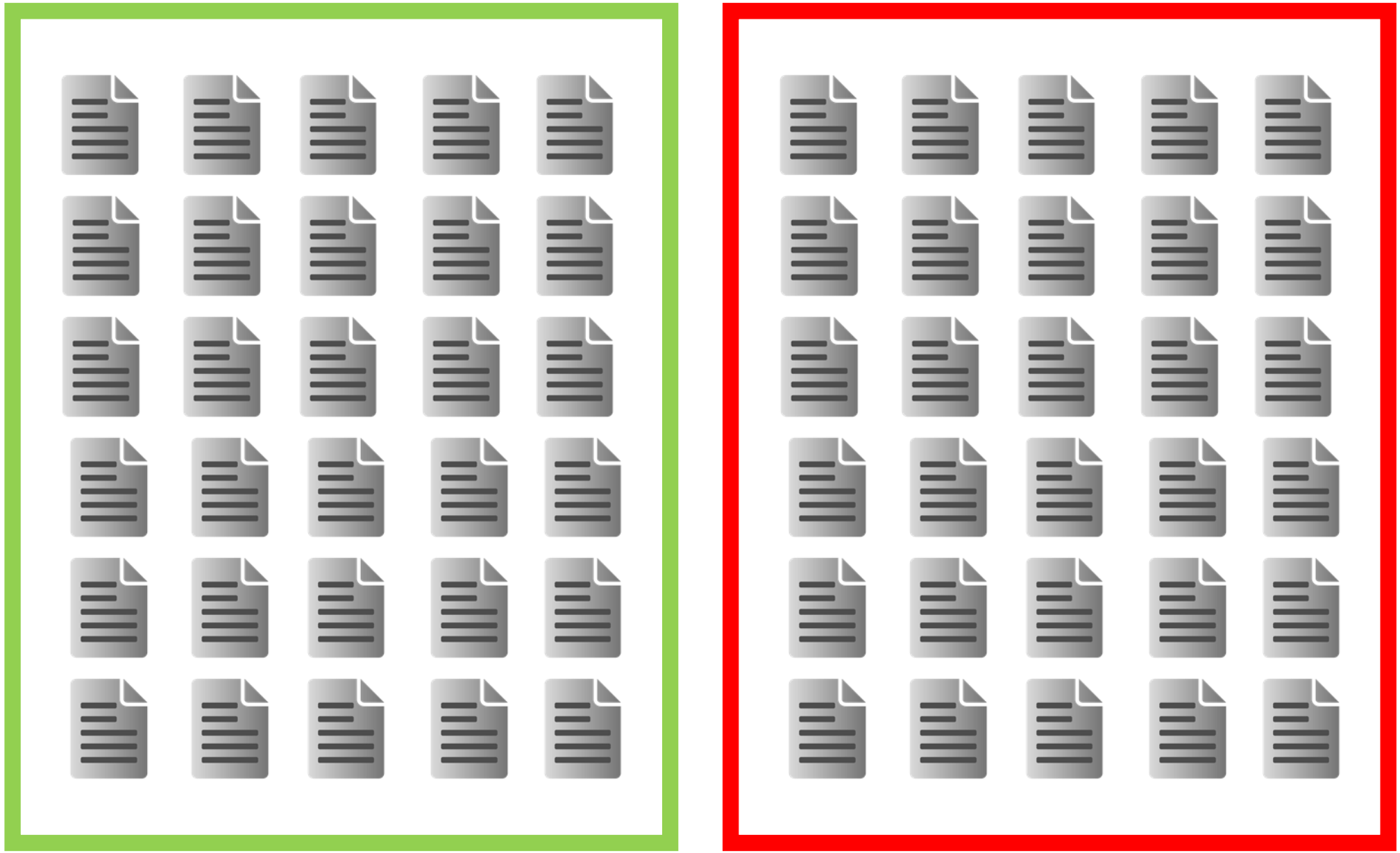
Introduction to Computer Science and Programming

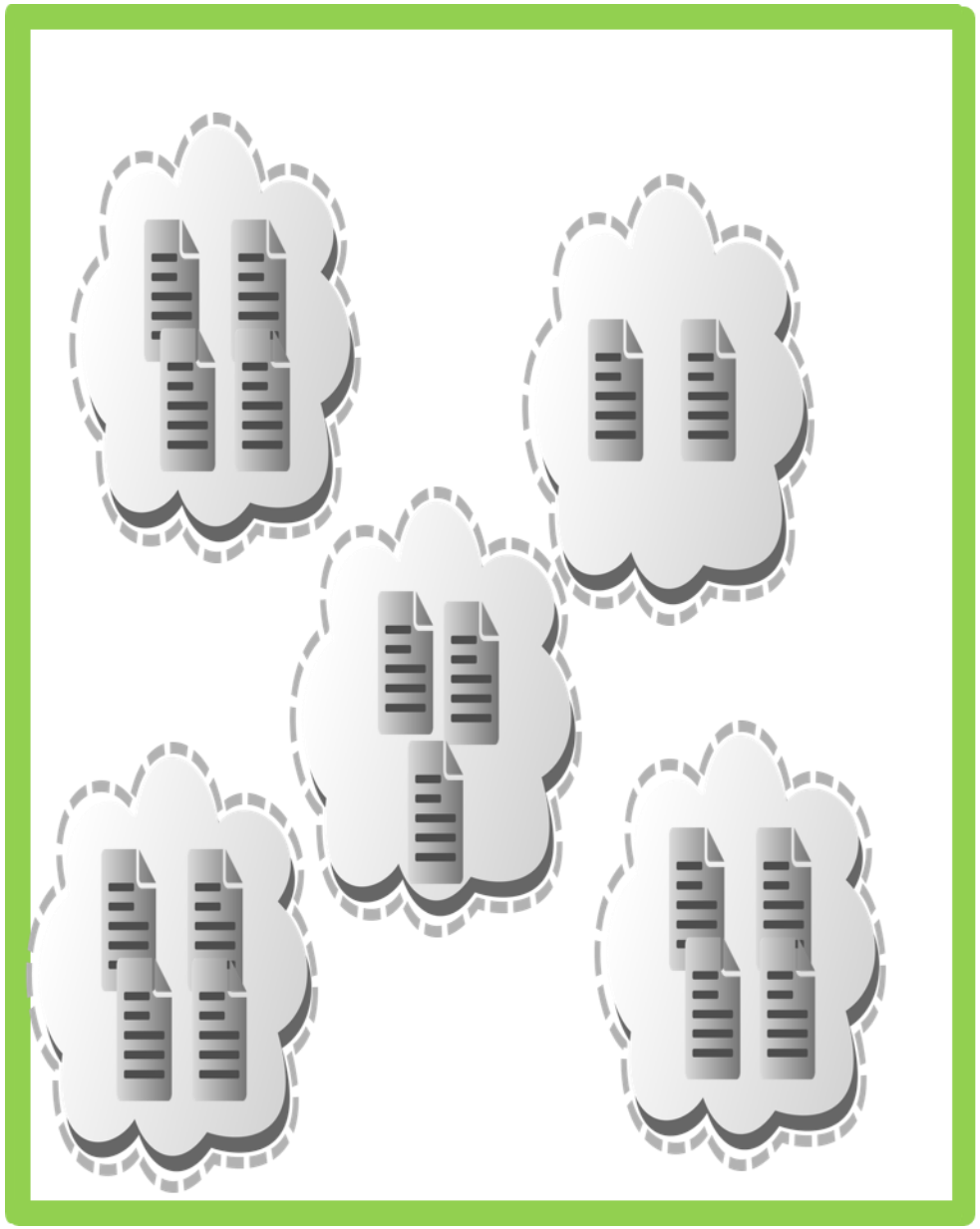
157,431

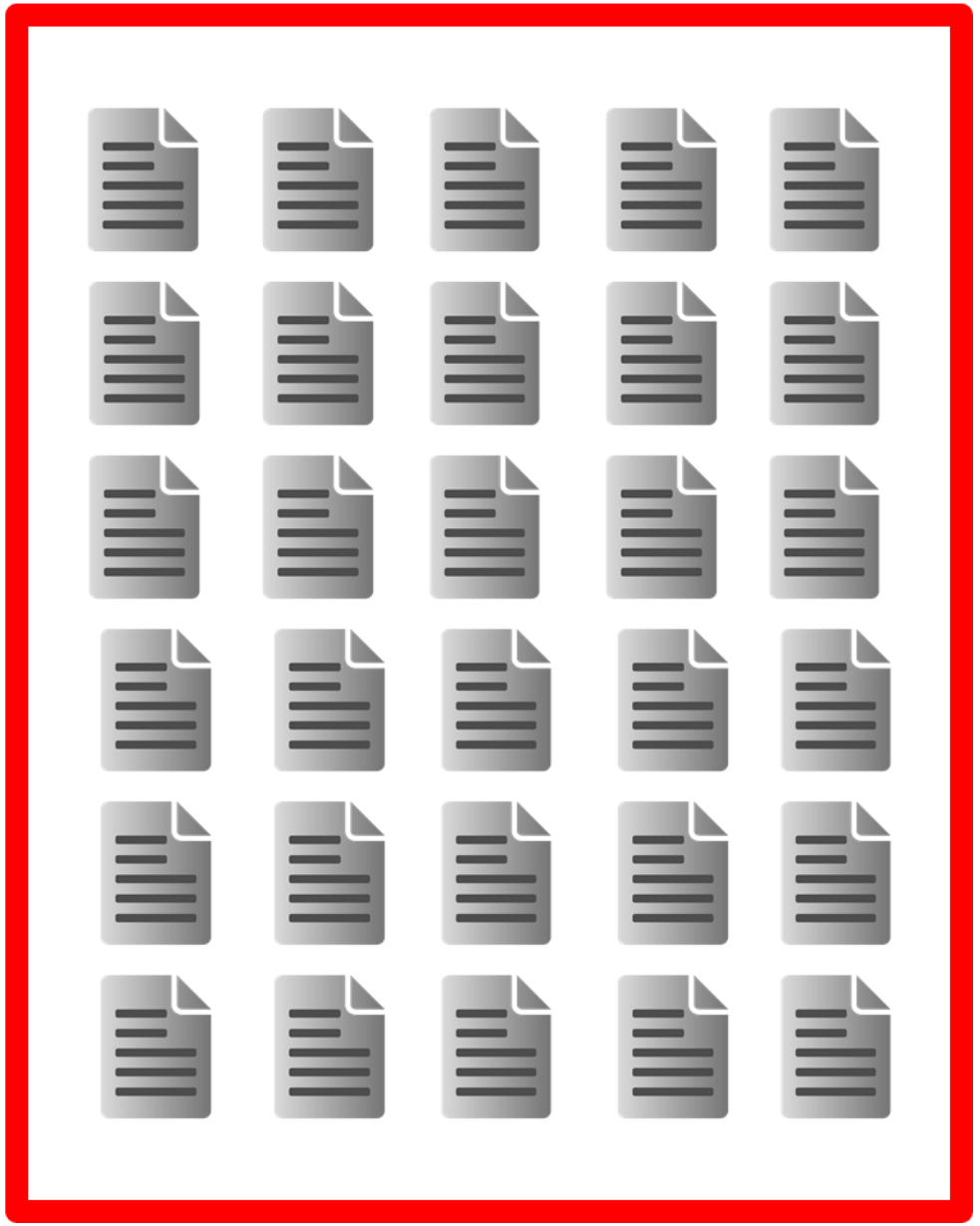
Current State-of-the-art

- Solution analysis, Feedback/repair generation, grading

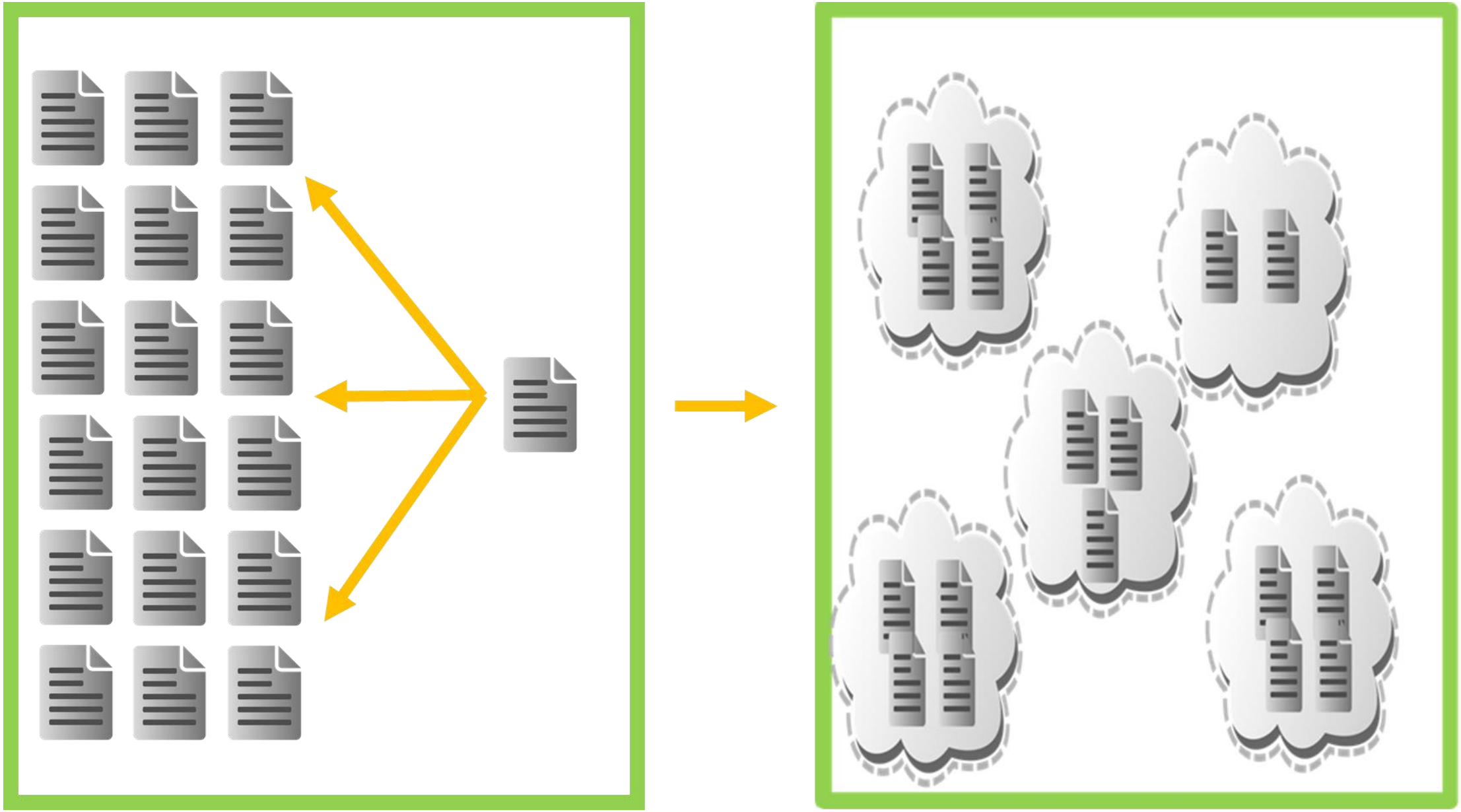
- Existing solutions from PL and SE communities
  - Clara
  - OverCode

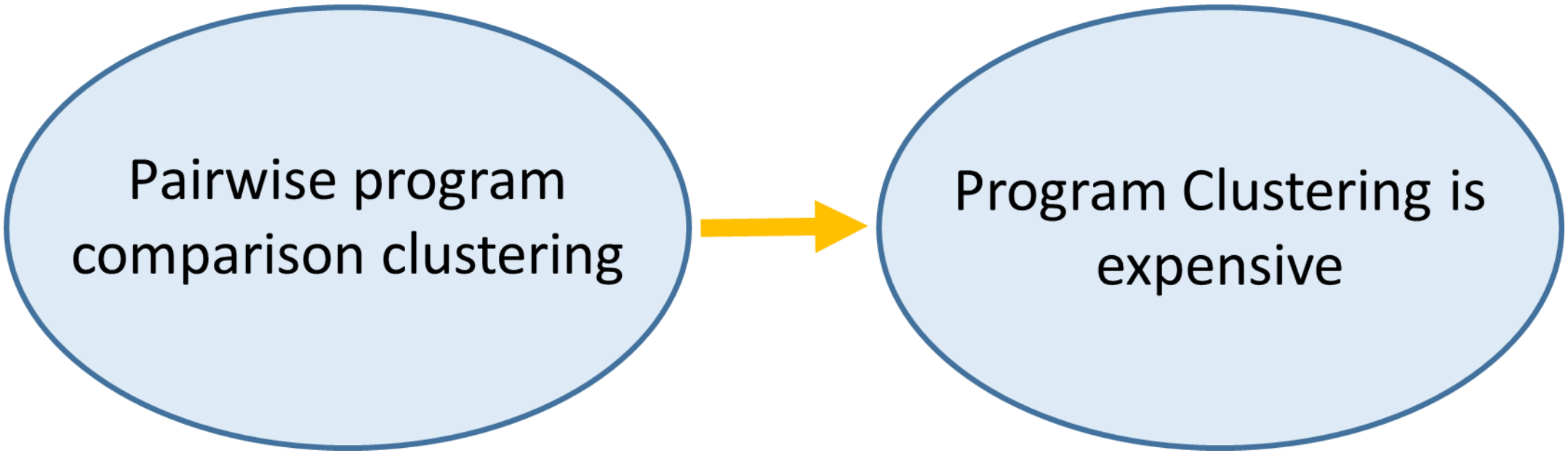






Problem: Pairwise program comparison-based clustering

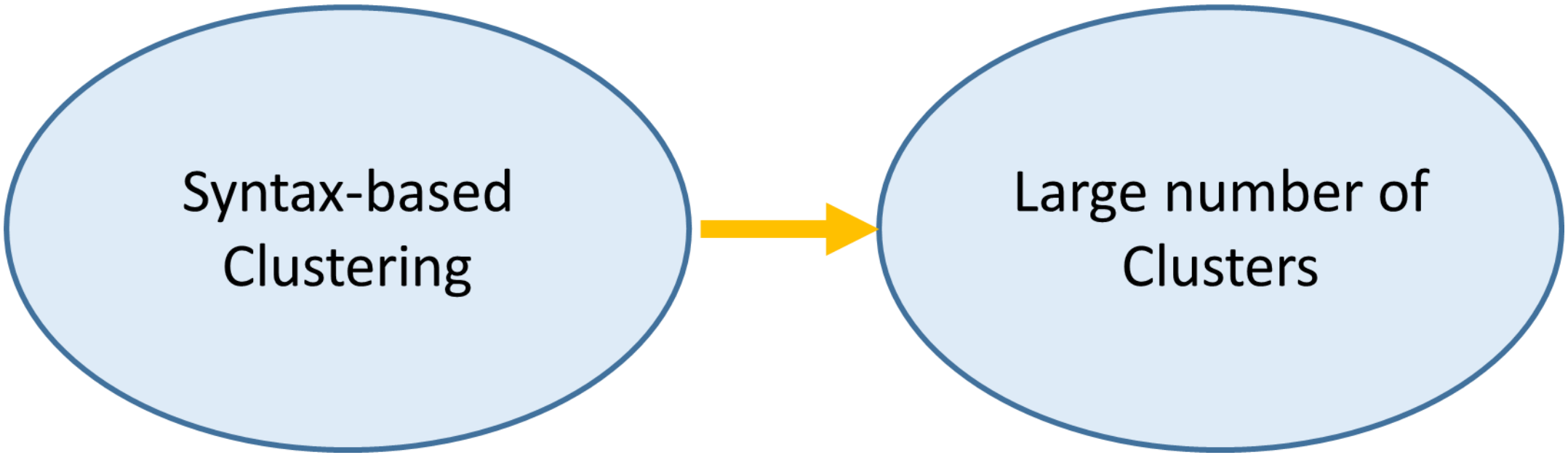



Pairwise program comparison clustering
Program Clustering is expensive

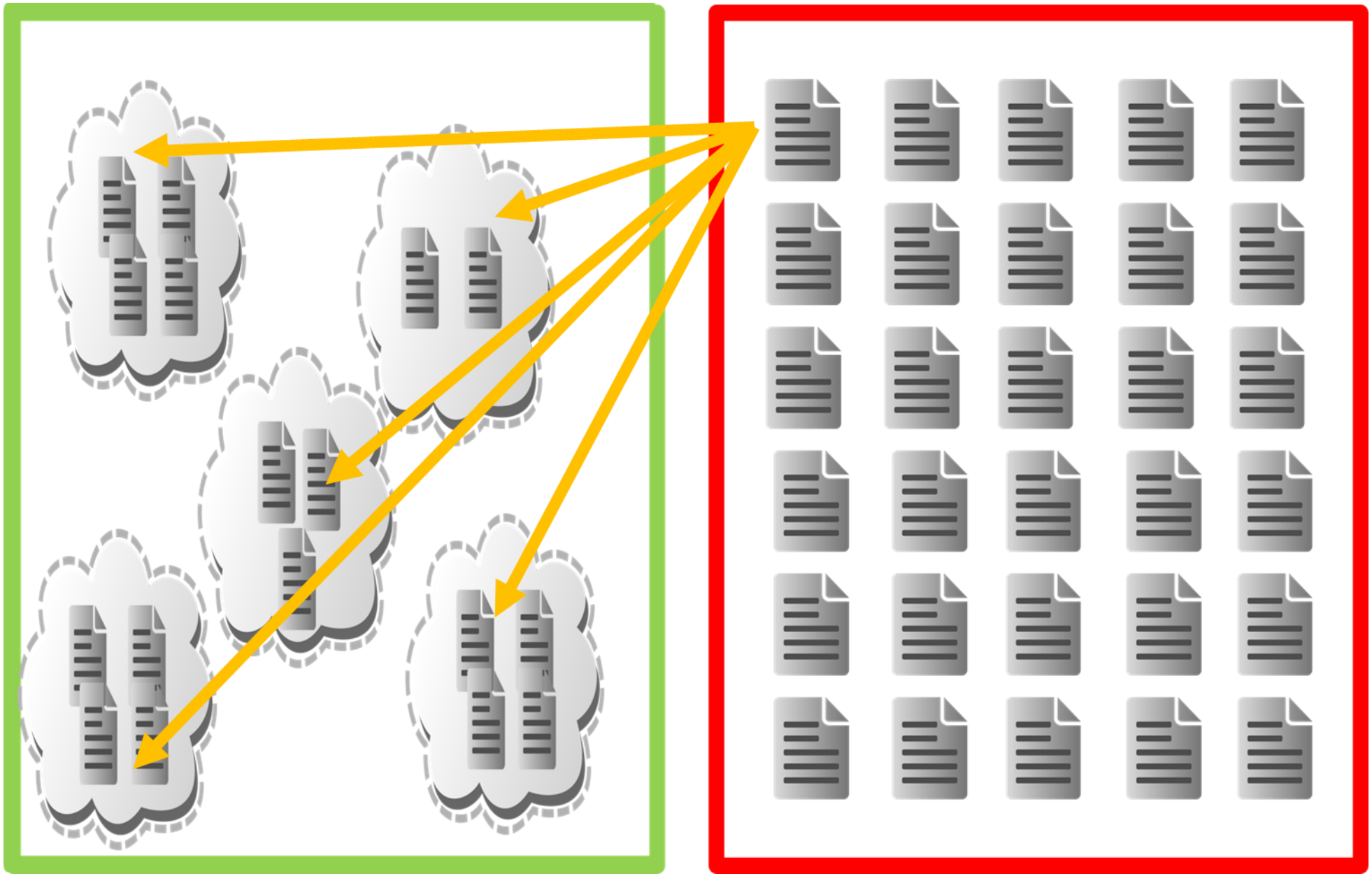
Problem: Syntax-based clustering

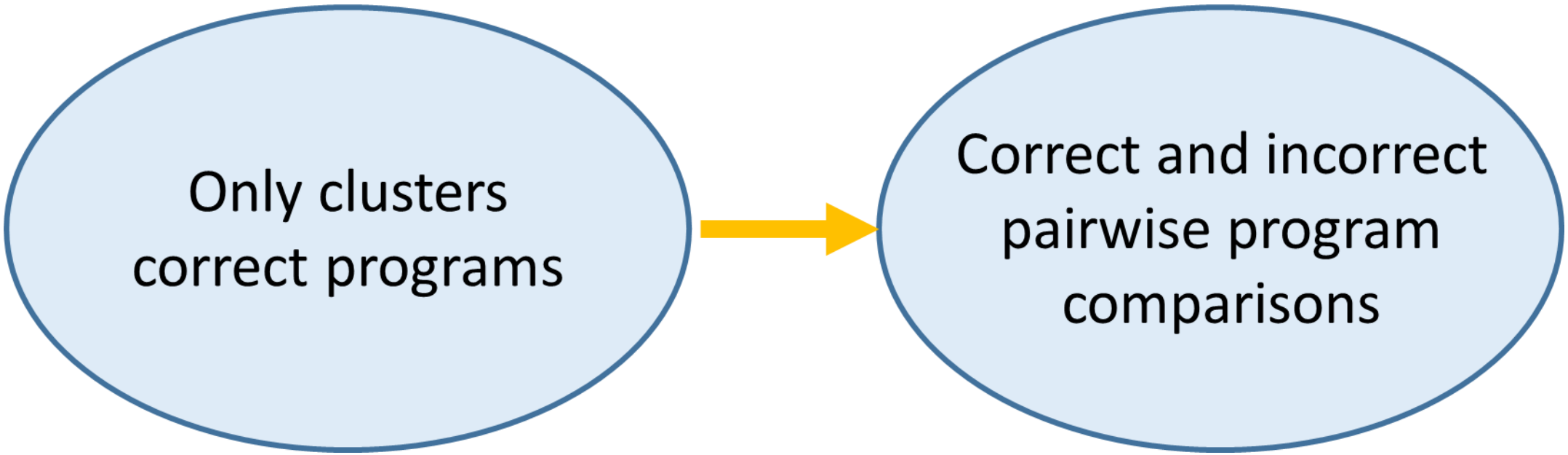
```
1 while(x < 10) {
2    if(y < 5) {
3       //do stuff
4       break;
5    }
6 }
```

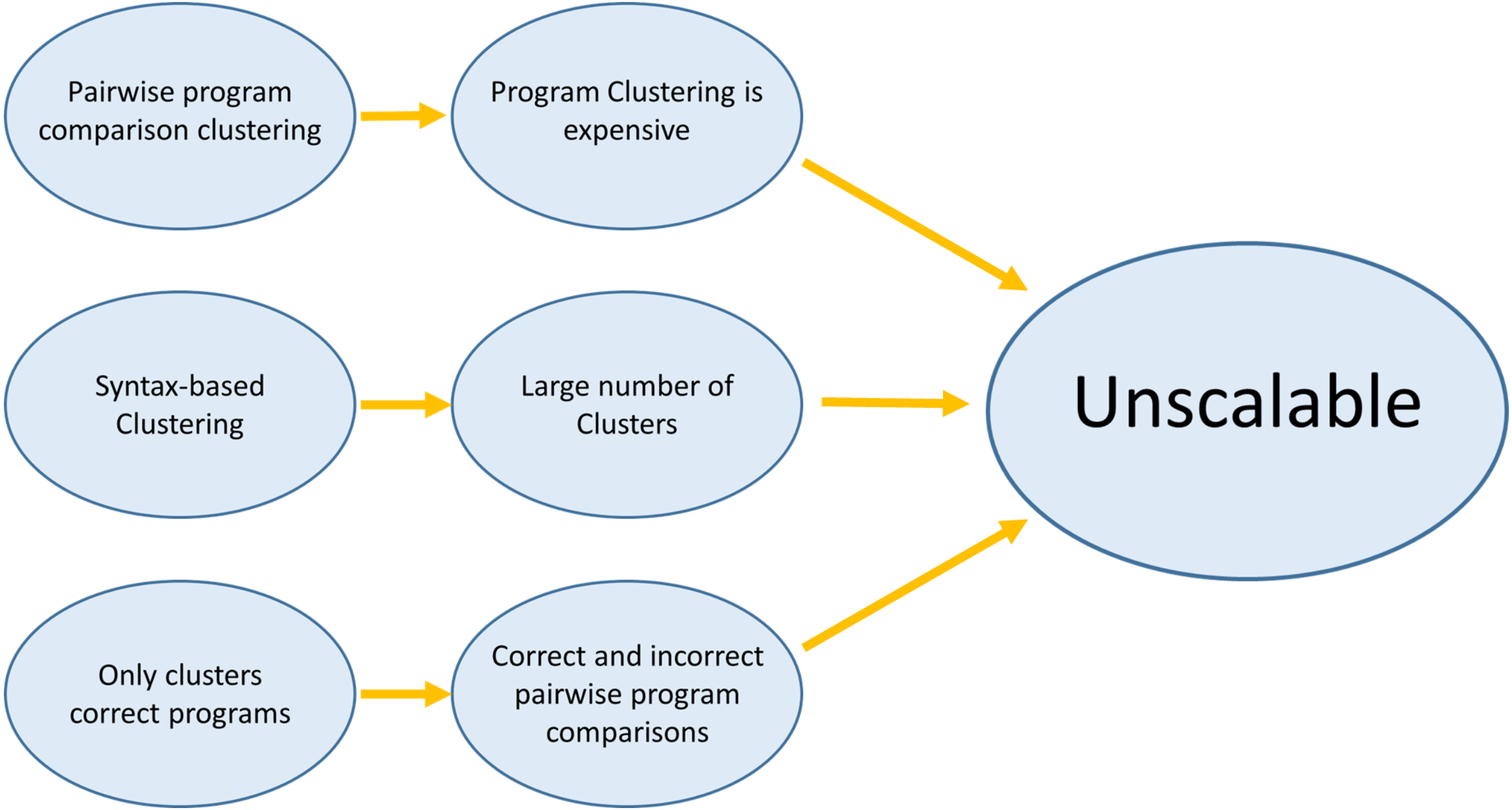
```
1 while(x < 10) {
2    if(y < 5 && bool == 1) {
3       //do stuff
4       bool = 0;
5    }
6 }
```


Syntax-based
Clustering
Large number of
Clusters

Problem: Only clusters correct programs




Only clusters correct programs
Correct and incorrect pairwise program comparisons


Pairwise program comparison clustering
Program Clustering is expensive
Syntax-based Clustering
Large number of Clusters
Only clusters correct programs
Correct and incorrect pairwise program comparisons
Unscalable

A new clustering approach

Program semantics vector representation

Large scale evaluation with excellent results

A new clustering approach

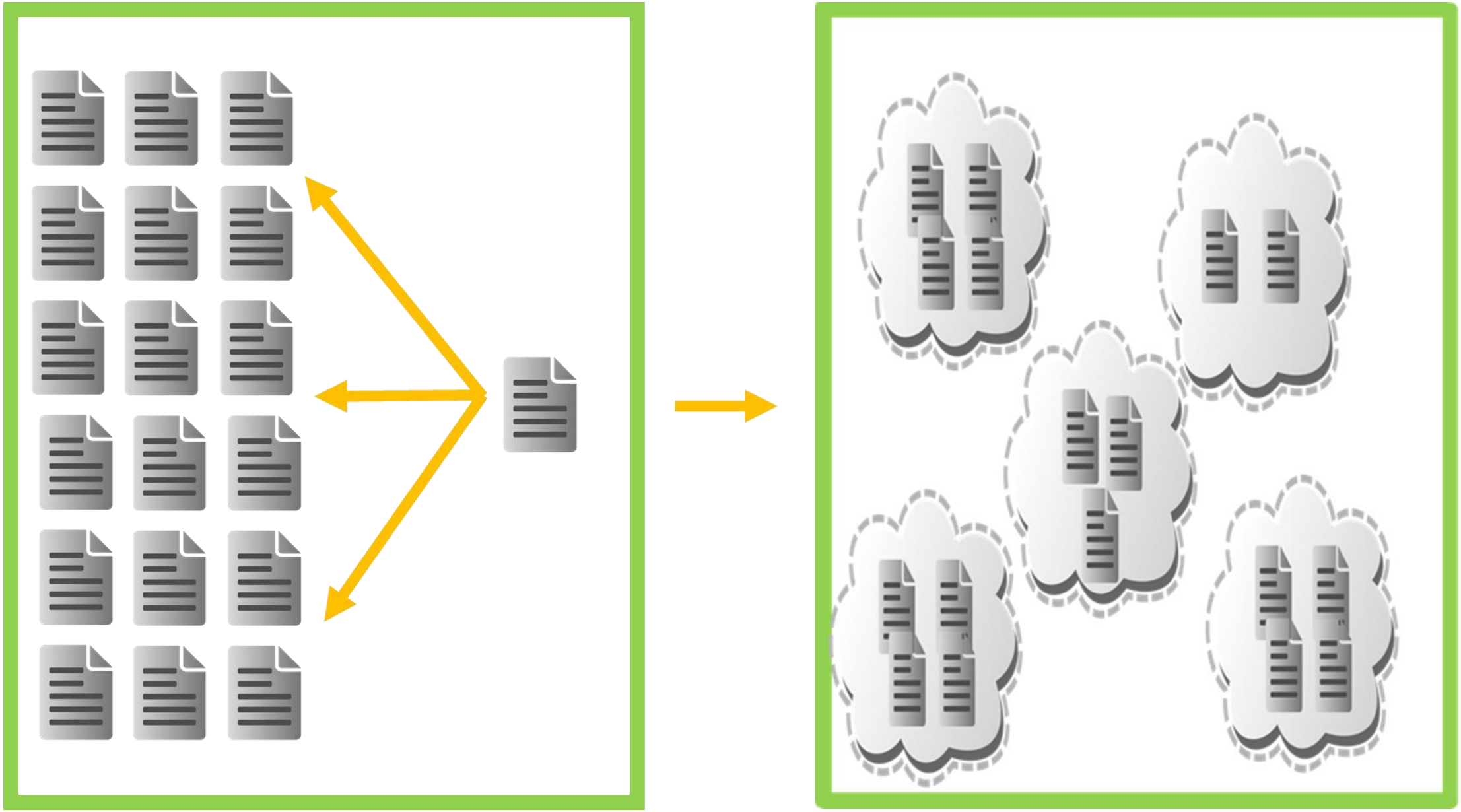
Program semantics vector representation

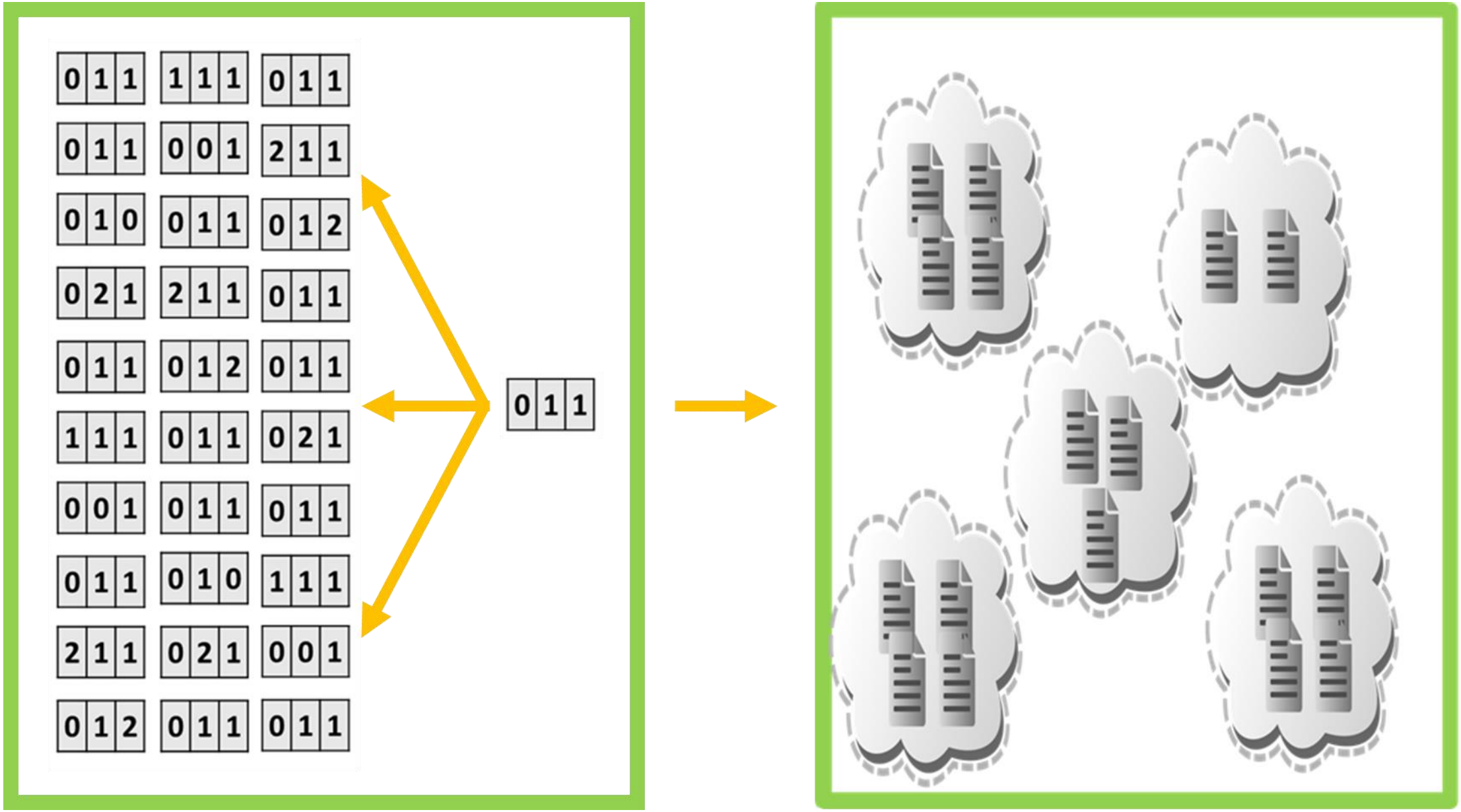
Large scale evaluation with excellent results

A new clustering approach

- Perform program comparisons with Euclidean distance

- Use existing machine learning techniques for clustering




0 1 1
1 1 1
0 1 1
0 1 1
0 0 1
2 1 1
0 1 0
0 1 1
0 1 2
0 2 1
2 1 1
0 1 1
0 1 1
0 1 2
0 1 1
1 1 1
0 1 1
0 2 1
0 0 1
0 1 1
0 1 1
0 1 1
0 1 0
1 1 1
2 1 1
0 2 1
0 0 1
0 1 2
0 1 1
0 1 1
0 1 1

A new clustering approach

Program semantics vector representation

Large scale evaluation with excellent results

```
1 while(x < 10) {
2     if(y < 5 && bool == 1) {
3         //do stuff
4         bool = 0;
5     }
6 }
```

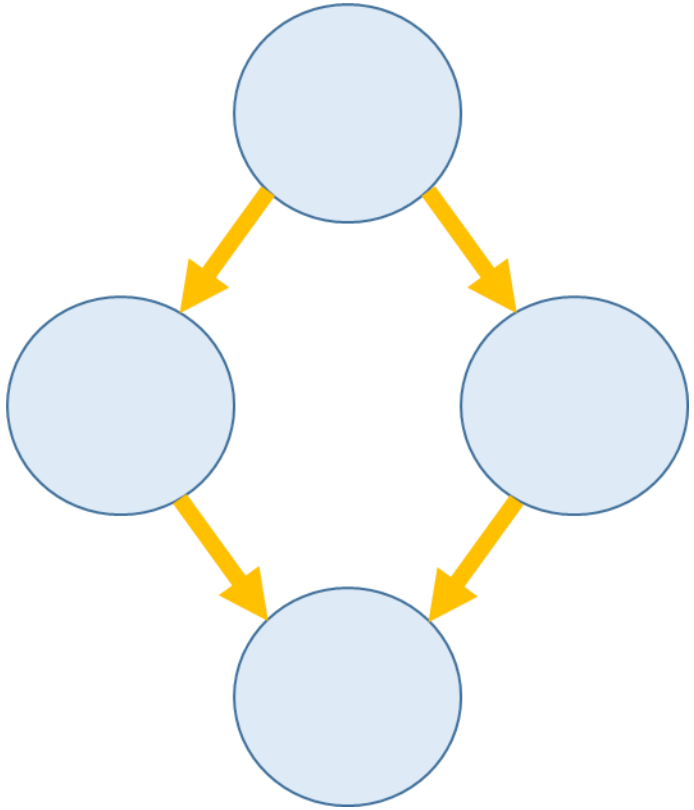
0	1	3	1	5

Control Flow Features (CFF)

- Encapsulates programmer's control flow decisions
- Number of inputs that trigger a specific path
  - path → SMT formula
  - Calculated using Model Counting
- One feature is calculated per test case

Control Flow Features

If x < 30



Bounds:	0 ≤ x ≤ 99	
Inputs:	X == 25	X == 75
CFFs:	30	70

Data Flow Features (DFF)

- Encapsulates programmer's data flow decisions
- Measures number of "value changes" during an execution
- Multiple features are calculated for each test case

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y = x * 2;
```

Inputs: x == 1, y == 1

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y = x * 2;
```

Inputs: x == 1, y == 1

1 → 3		
2		

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y = x * 2;
```

Inputs: x == 1, y == 1

1 → 3	3 → 4	
2	1	

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y = x * 2;
```

Inputs: x == 1, y == 1

1 → 3	3 → 4	3 → 8
2	1	1

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y++;
```

Inputs: x == 1, y == 1

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y++;
```

Inputs: x == 1, y == 1

1 → 3		
2		

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y++;
```

Inputs: x == 1, y == 1

1 → 3	3 → 4	
2	2	

Data Flow Features

```
1.  int x, y, z;
2.
3.  x = 3;
4.  y = 3;
5.
6.  x++;
7.  y++;
```

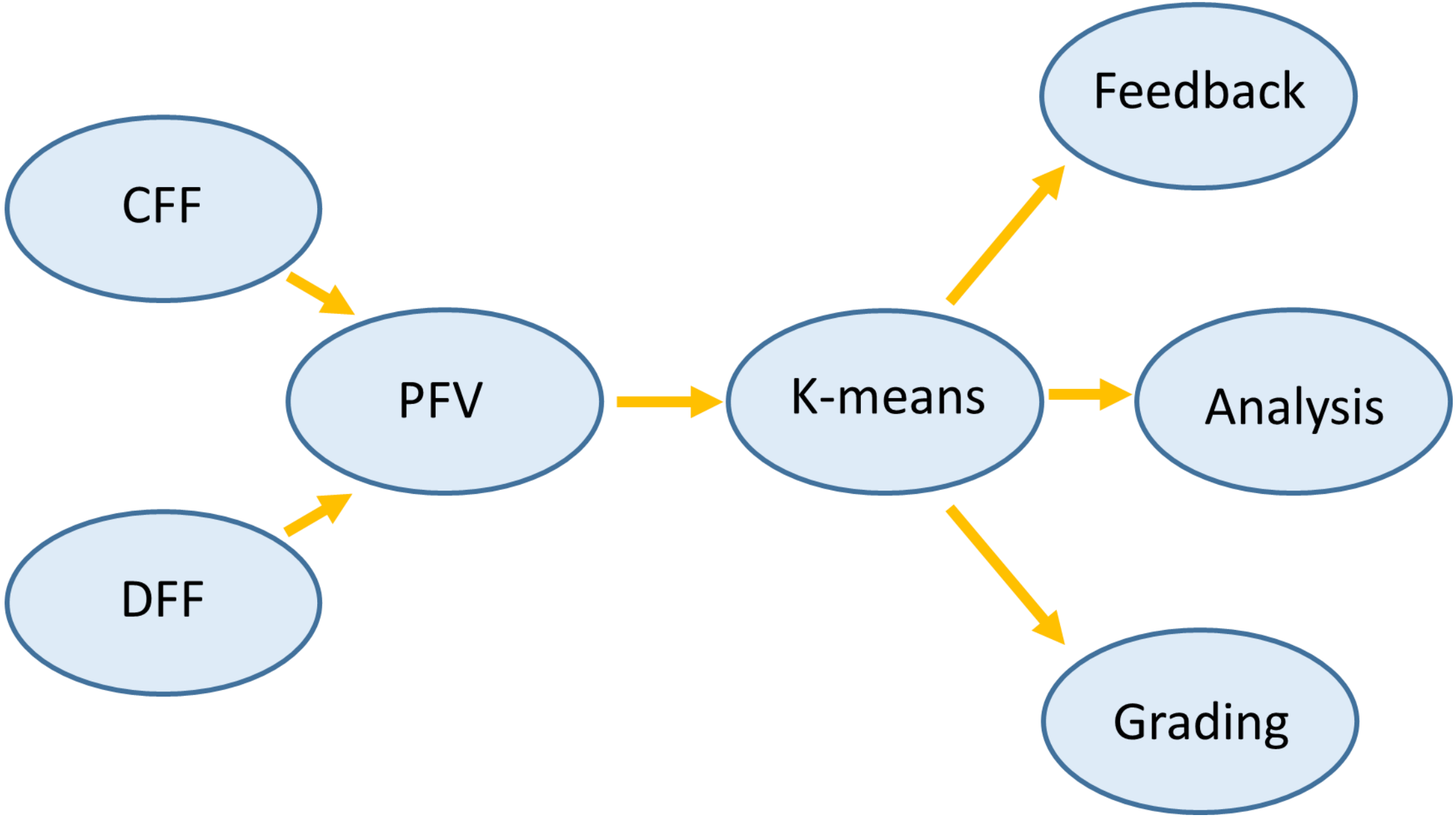
Inputs: x == 1, y == 1

1 → 3	3 → 4	3 → 8
2	2	0

Program Feature Vector (PFV)

- Concatenate Data Flow and Control Flow features

- Normalize weight of feature in the vector


CFF
DFF
PFV
K-means
Feedback
Analysis
Grading

A new clustering approach

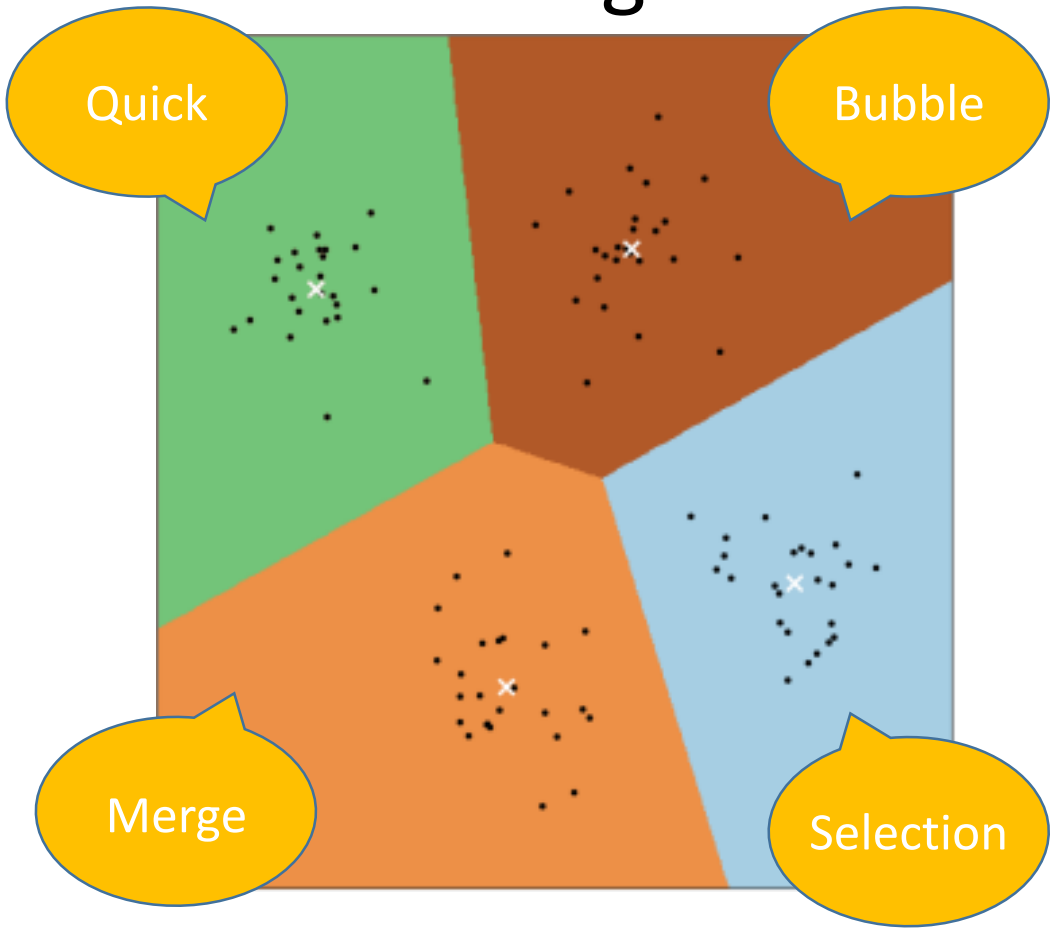
Program semantics vector representation

Large scale evaluation with excellent results

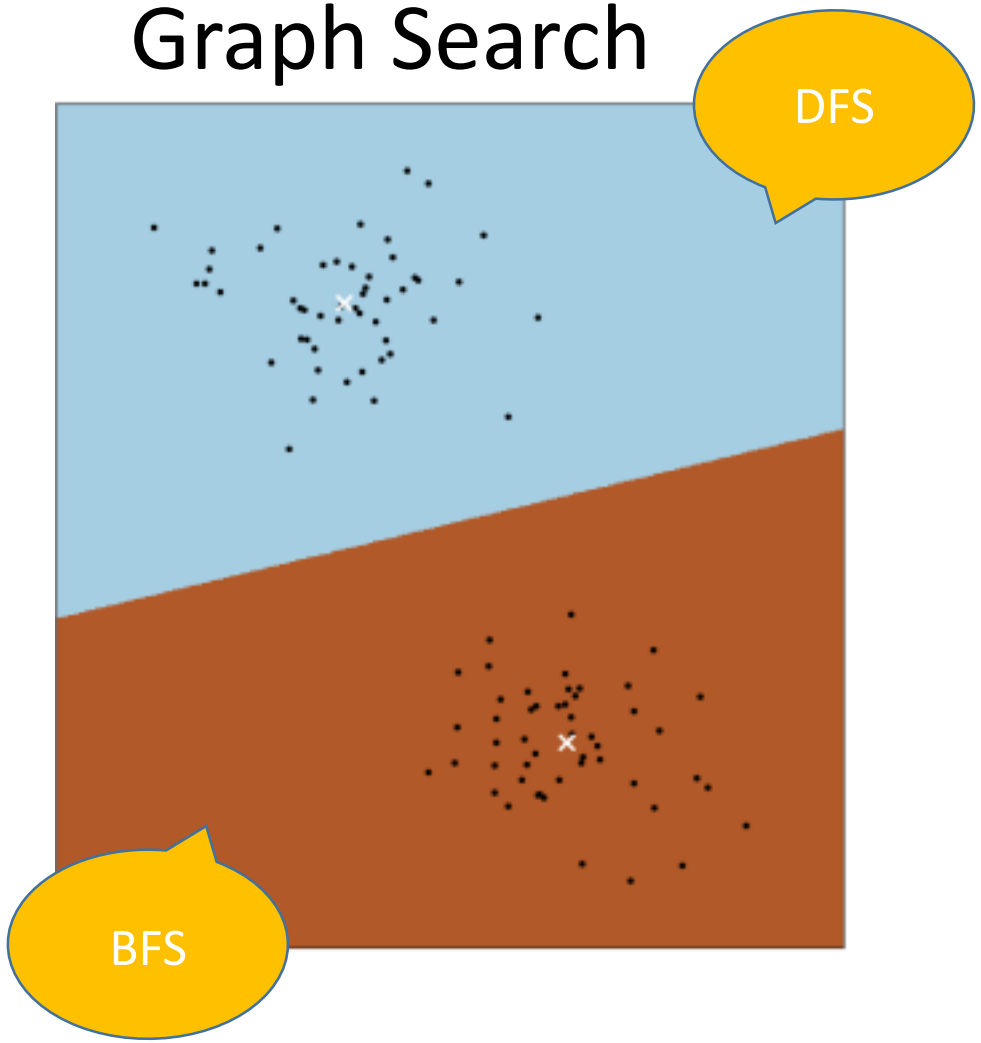
Neural Network Approach

Problem	Avg. LOC	# Subs	Clara	OverCode	DPE	SemCluster
COINS	38	1033	89	101	10	8
PRIME1	59	920	120	125	9	10
LAPIN	65	175	62	62	9	8
LCM	15	806	99	103	12	13
FibSum	14	1030	30	32	12	15

Sorting



Graph Search



Problem	SemCluster		DPE	
	In Cluster	Out Cluster	In Cluster	Out Cluster
COINS	83.2	9.4	85.2	7.3
PRIME1	80.9	12.5	77.2	14.2
LAPIN	87.7	11.3	88.9	10.8
LCM	79.1	18.2	77.4	20.2
FibSum	87.9	5.2	77.2	9.1

Technique	Small Cluster Size	High Accuracy	No Training Requirement
DPE	✓	✓	✗
SemCluster	✓	✓	✓

A new clustering approach

Program semantics vector representation

Large scale evaluation with excellent results

Future Work

Questions?