Learning a Variable-Clustering Strategy for Octagon from Labeled Data Generated by a Static Analysis

Kihong Heo[1], Hakjoo Oh[2], Hongseok Yang[3]

Seoul National University[1]
Korea University[2]
University of Oxford[3]

SAS 2016 @Edinburgh

Long Term Goal

- Self-evolving static analysis by learning big data
  - data : similar codes, old versions, user-feedbacks, bug reports, test results, etc
  - mature in other fields : amazon …

Bugzilla

stackoverflow

sourceforge

Big Data

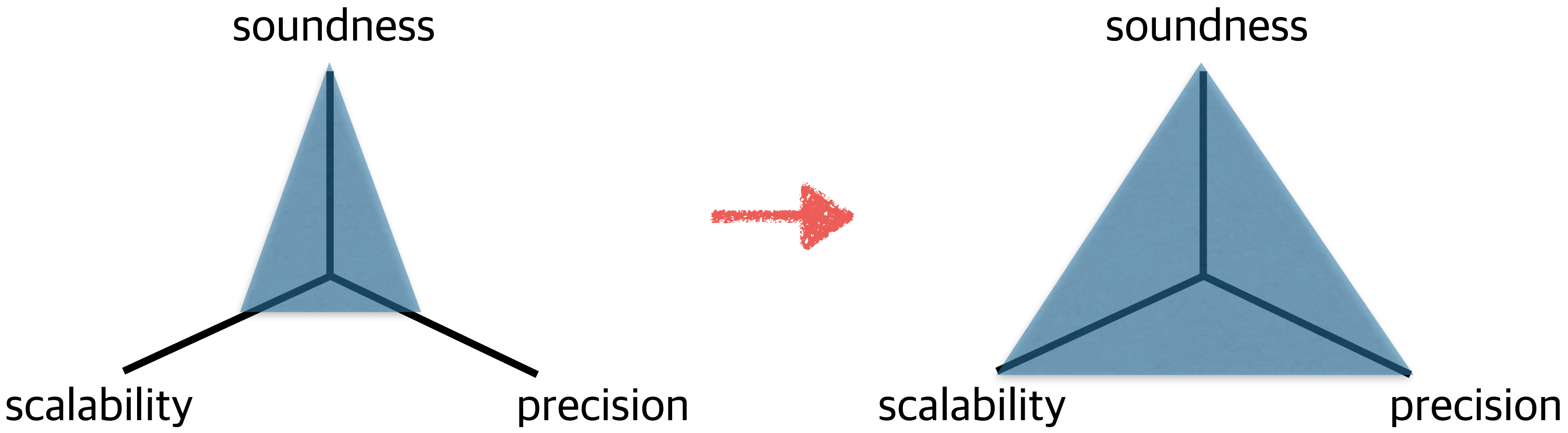
+

Sparrow

The Early Bird

Static Analyzer

Long Term Goal



$$F \in Pgm \times \underline{\Pi} \to \mathcal{A}$$

- Finding a good abstraction for adaptive static analysis
  - **Machine Learning** (learner) + **Static Analysis** (teacher)
  - e.g.) **relation**, context, flow, etc

Relational Analysis

- Tracking relationships among variables
  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```

	a	b	c	i
a	0	∞	∞	∞
b	∞	0	∞	∞
c	∞	∞	0	∞
i	∞	∞	∞	0

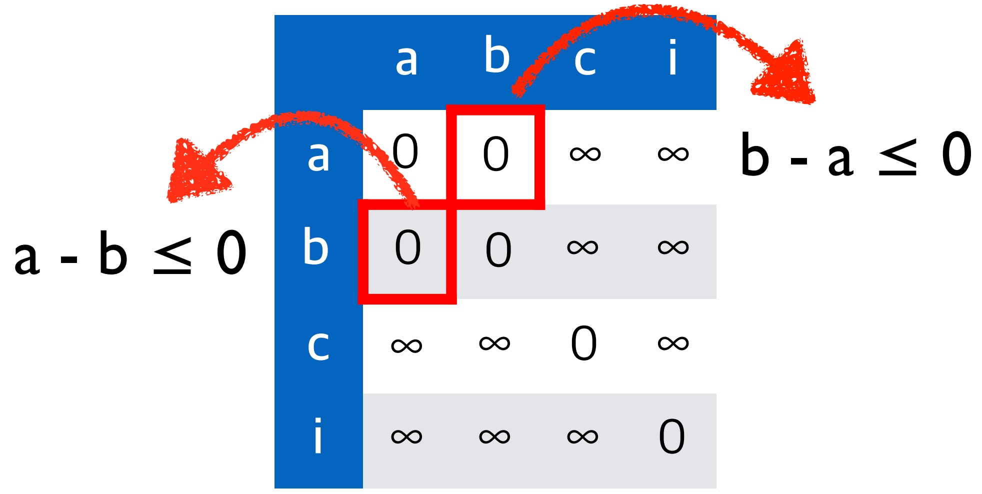
{a, b, c, i}

*Consider x-y ≤ c only, for simplicity

Relational Analysis

- Tracking relationships among variables
  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                 // User input
for (i = 0; i < b; i++) {
  assert (i < a);                // Query 1
  assert (i < c);                // Query 2
}
```



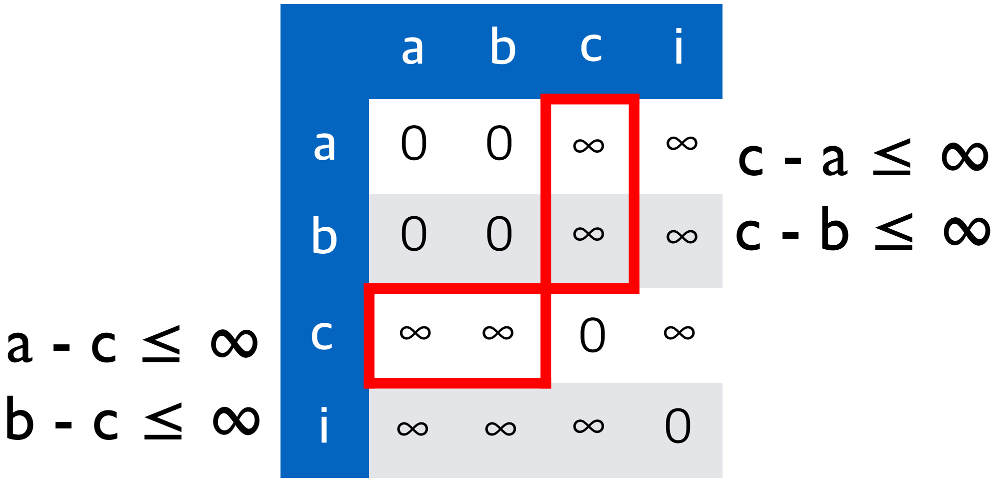
	a	b	c	i
a	0	0	∞	∞
b	0	0	∞	∞
c	∞	∞	0	∞
i	∞	∞	∞	0

{a, b, c, i}

Relational Analysis

- Tracking relationships among variables
  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```



	a	b	c	i
a	0	0	∞	∞
b	0	0	∞	∞
c	∞	∞	0	∞
i	∞	∞	∞	0

c - a ≤ ∞
c - b ≤ ∞

a - c ≤ ∞
b - c ≤ ∞

{a, b, c, i}

Relational Analysis

- Tracking relationships among variables

  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();            // User input
for (i = 0; i < b; i++) {
  assert (i < a);           // Query 1
  assert (i < c);           // Query 2
}
```

	a	b	c	i
a	0	0	∞	∞
b	0	0	∞	-1
c	∞	∞	0	∞
i	∞	∞	∞	0

i - b ≤ -1

{a, b, c, i}

Relational Analysis

- Tracking relationships among variables
  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```

	a	b	c	i
a	0	0	∞	-1
b	0	0	∞	-1
c	∞	∞	0	∞
i	∞	∞	∞	0

i - a ≤ -1

{a, b, c, i}

Relational Analysis

- Tracking relationships among variables
    - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```

	a	b	c	i
a	0	0	∞	-1
b	0	0	∞	-1
c	∞	∞	0	∞
i	∞	∞	∞	0

i - c ≤ ∞

{a, b, c, i}

Relational Analysis

- Tracking relationships among variables
  - e.g.) octagon analysis : $(\pm x) - (\pm y) \leq c$

```
int a = b;
int c = input();                  // User input
for (i = 0; i < b; i++) {
  assert (i < a);                 // Query 1
  assert (i < c);                 // Query 2
}
```

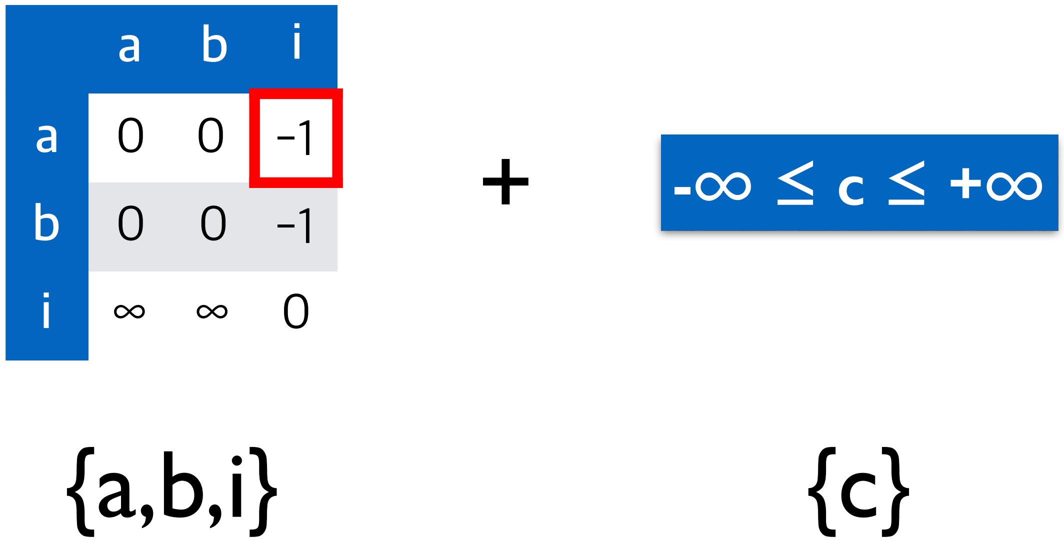
	a	b	~~c~~	i
a	0	0	~~∞~~	-1
b	0	0	~~∞~~	-1
~~c~~	~~∞~~	~~∞~~	~~0~~	~~∞~~
i	∞	∞	~~∞~~	0

{a, b, ~~c~~, i}

Selective Relational Analysis

- Selectively tracking relationships among variables
  - within the same cluster

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```



	a	b	i
a	0	0	-1
b	0	0	-1
i	∞	∞	0

+

$-\infty \leq c \leq +\infty$

{a,b,i} {c}

PLDI'14

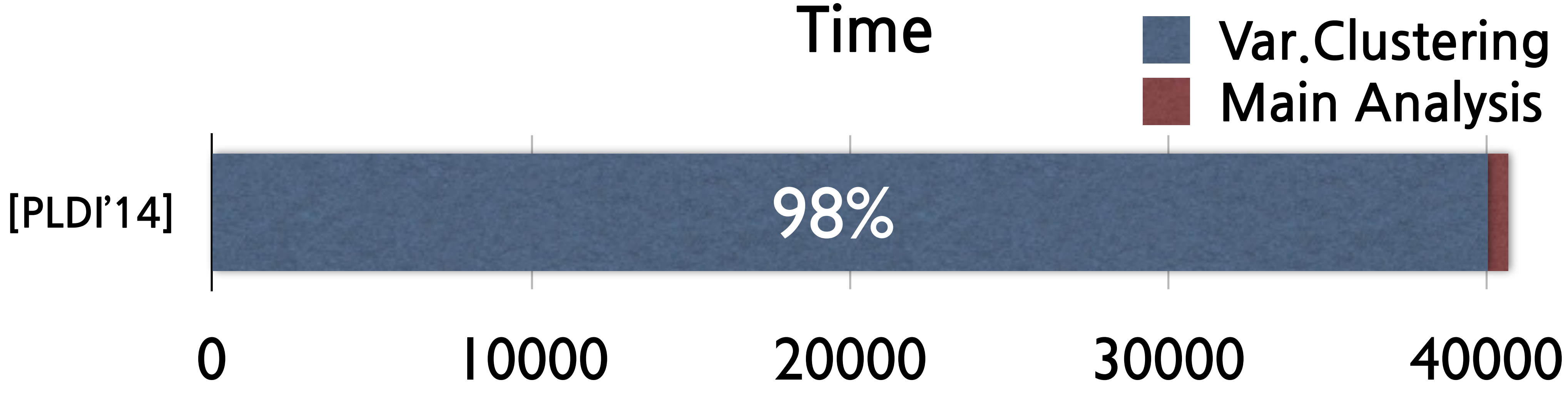
Previous Solution

- Variable clustering by impact pre-analysis
  - estimating the impact of relationships
  - more scalable than the baseline Octagon analysis
  - more scalable & precise than other clustering methods

PLDI'14

Problem

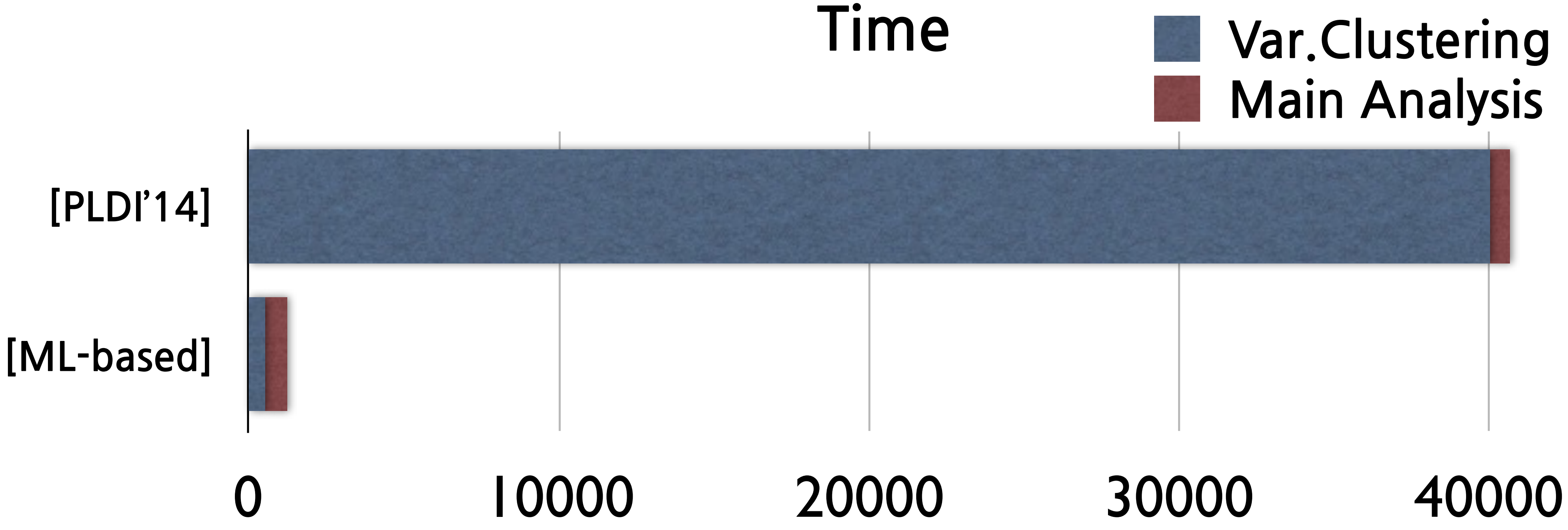
- Variable clustering by impact pre-analysis
  - fully relational pre-analysis as an online estimator
  - e.g.) 17 open source benchmarks (~100KLOC)



This Work

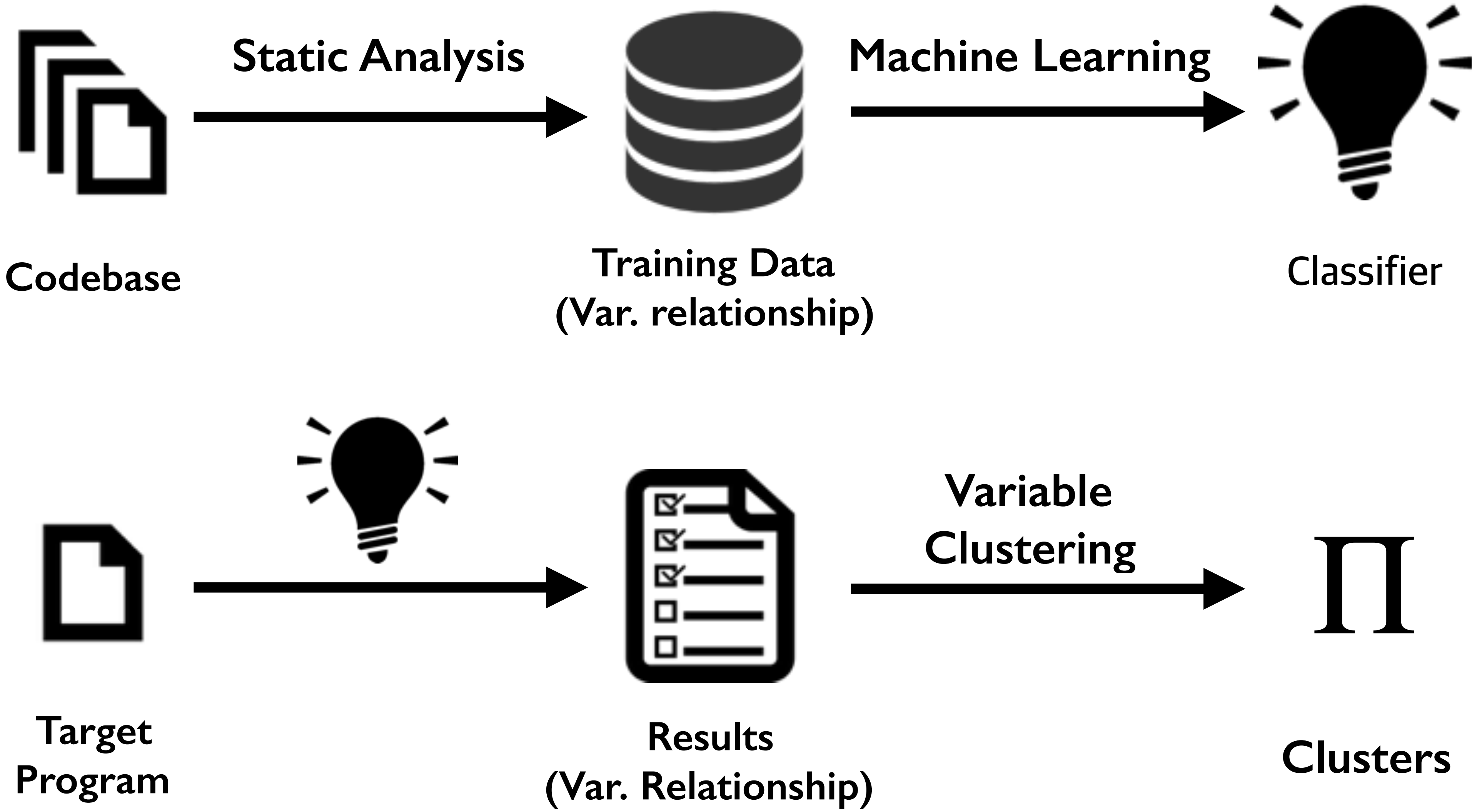
New Solution

- Learning a variable-clustering strategy from big data
  - fully relational pre-analysis as an offline teacher
  - 33x faster yet similarly precise



Big Picture

- Learning a variable-clustering strategy from big data



Static Analysis

Machine Learning

Codebase

Training Data
(Var. relationship)

Classifier

Variable
Clustering

Π

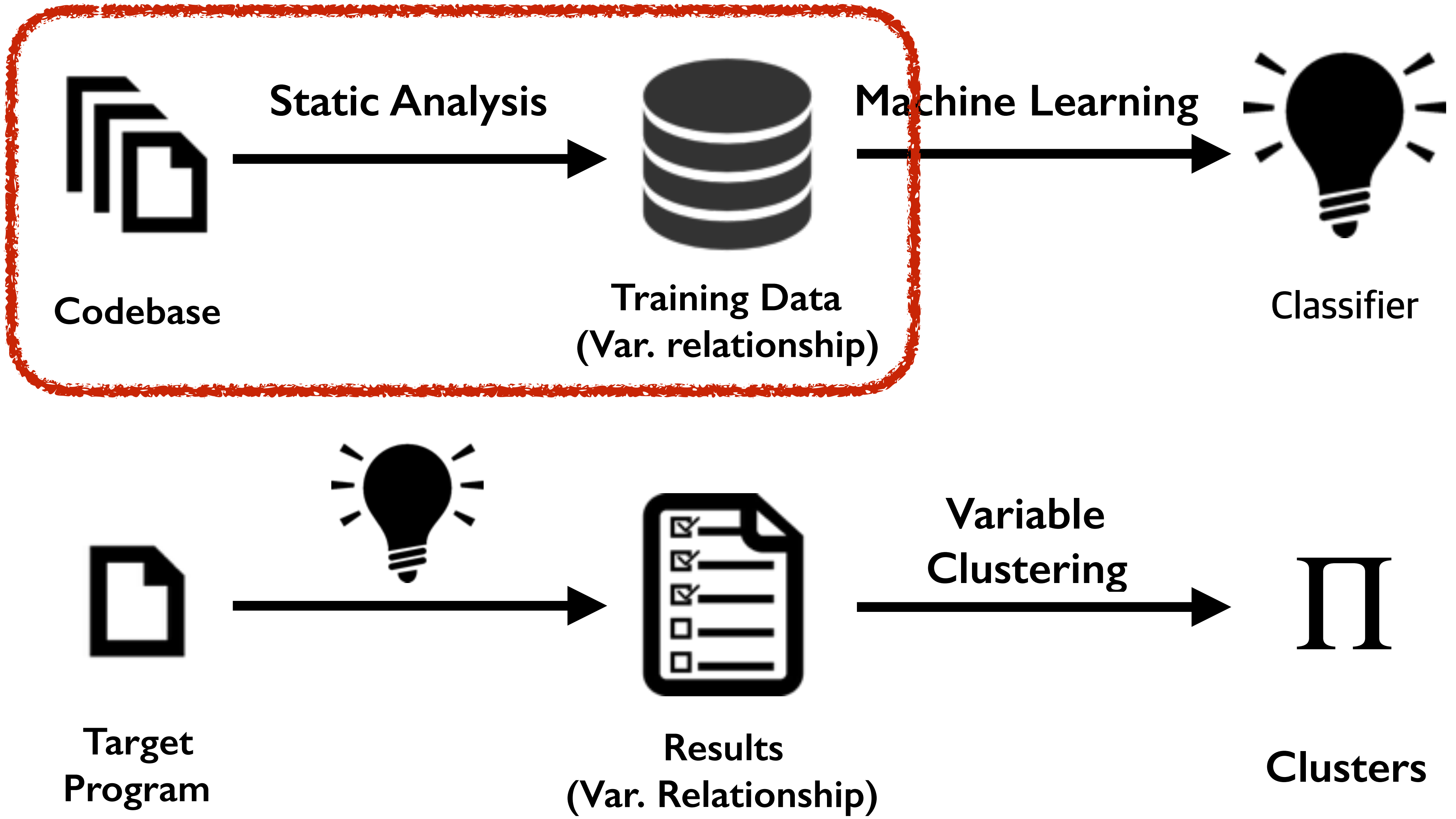
Target
Program

Results
(Var. Relationship)

Clusters

Big Picture

- Learning a variable-clustering strategy from big data



Training Data

- Pairs of two variables with label {⊕, ⊖}

  - ⊕: precise (< +∞), ⊖: imprecise (= +∞)

```
int a = b;
int c = input();                 // User input
for (i = 0; i < b; i++) {
  assert (i < a);                // Query 1
  assert (i < c);                // Query 2
}
```

	a	b	c	i
a	0	0	∞	-1
b	0	0	∞	-1
c	∞	∞	0	∞
i	∞	∞	∞	0

Octagon Analysis

⊕ : {(a,b), (a,i), (b,a) …}

⊖ : {(a,c), (b,c), (c,a) …}

Training Data

- Automatically generated by impact pre-analysis[PLDI'14]
  - fully relational, yet more scalable than the full octagon

```
int a = b;
int c = input();                // User input
for (i = 0; i < b; i++) {
  assert (i < a);               // Query 1
  assert (i < c);               // Query 2
}
```

	a	b	c	i
a	0	0	∞	-1
b	0	0	∞	-1
c	∞	∞	0	∞
i	∞	∞	∞	0

Octagon Analysis

$$\gamma(\bigstar) = \mathbb{Z}$$

$$\gamma(\top) = \mathbb{Z} \cup \{+\infty\}$$

$\oplus$: {(a,b), (a,i), (b,a) …}

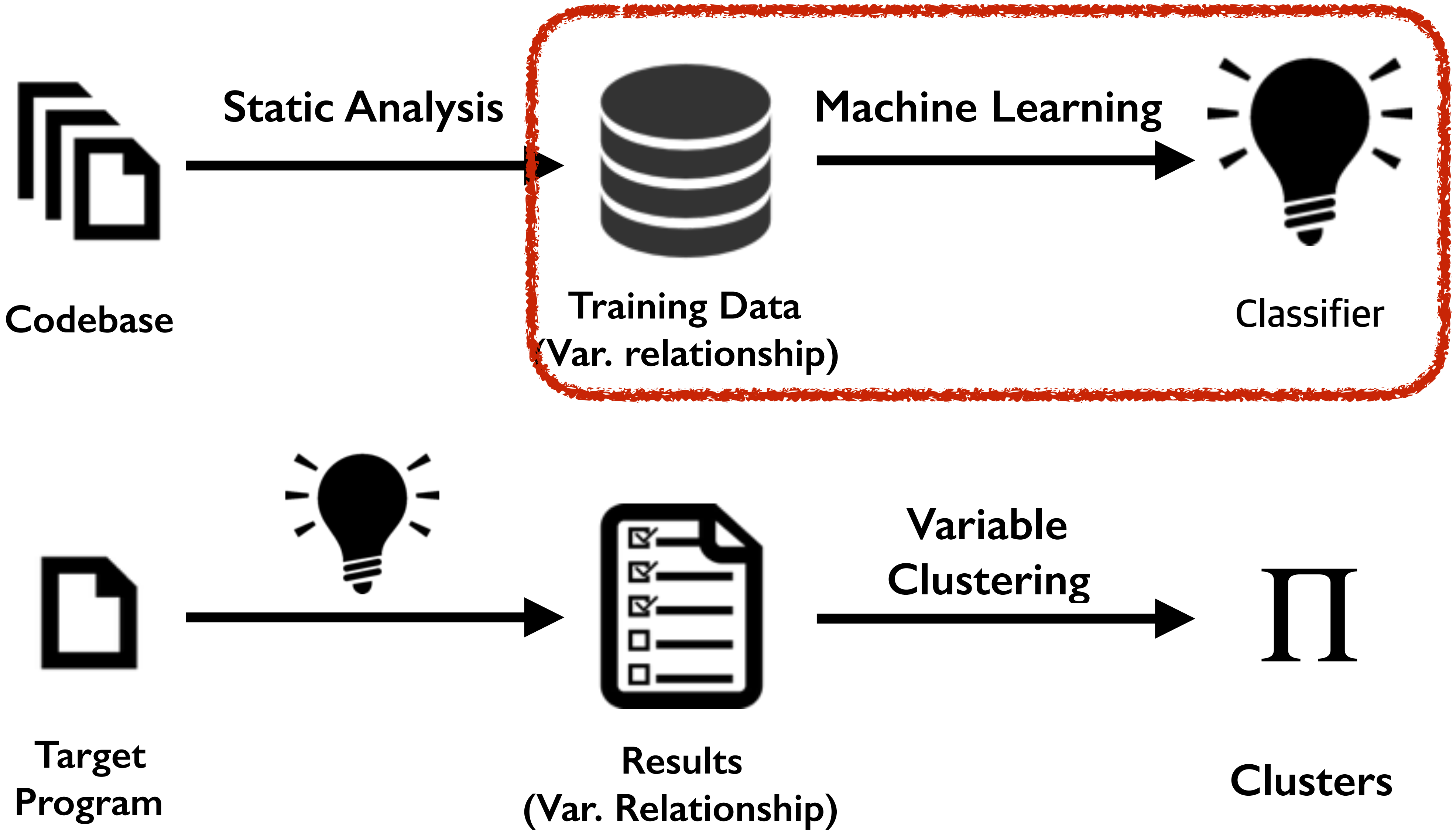
$\ominus$: {(a,c), (b,c), (c,a) …}

	a	b	c	i
a	★	★	⊤	★
b	★	★	⊤	★
c	⊤	⊤	★	⊤
i	⊤	⊤	⊤	★

Impact Pre-analysis

Big Picture

- Learning a variable-clustering strategy from big data



Features

- 30 Features of variable pairs
  - boolean predicate of (x,y) in program P

(Positive situations for Octagon)

- `x=y+k` or `y=x+k`
- `x<=y+k` or `y<=x+k`
- `x=malloc(y)` or `y=malloc(x)`
- `x[y]` or `y[x]`
- …

(Negative situations for Octagon)

- `x=cy` or `y=cx` (`c` != 1)
- `x=yz` or `y=xz`
- `x=y/z` or `y=x/z`
- …

(General syntactic features)

- x or y is a field
- x and y represent sizes of arrays
- x or y is the size of a const string
- x or y is a global variable
- …

(General semantic features)

- x or y has a finite interval
- x or y is a local var in a recursive function
- x, y are not accessed in the same function
- …

Features

- Importance of features by Gini Index
  - negative & general > positive & domain-specific

(Positive situations for Octagon)

- x=y+k or y=x+k
- x<=y+k or y<=x+k
- x=malloc(y) or y=malloc(x)
- x[y] or y[x]
- …

(Negative situations for Octagon)

- x=cy or y=cx (c != 1)
- x=yz or y=xz
- x=y/z or y=x/z
- …

(General syntactic features)

- x or y is a field
- x and y represent sizes of arrays
- x or y is the size of a const string
- x or y is a global variable
- …

(General semantic features)

- x or y has a finite interval
- x or y is a local var in a recursive function
- x, y are not accessed in the same function
- …

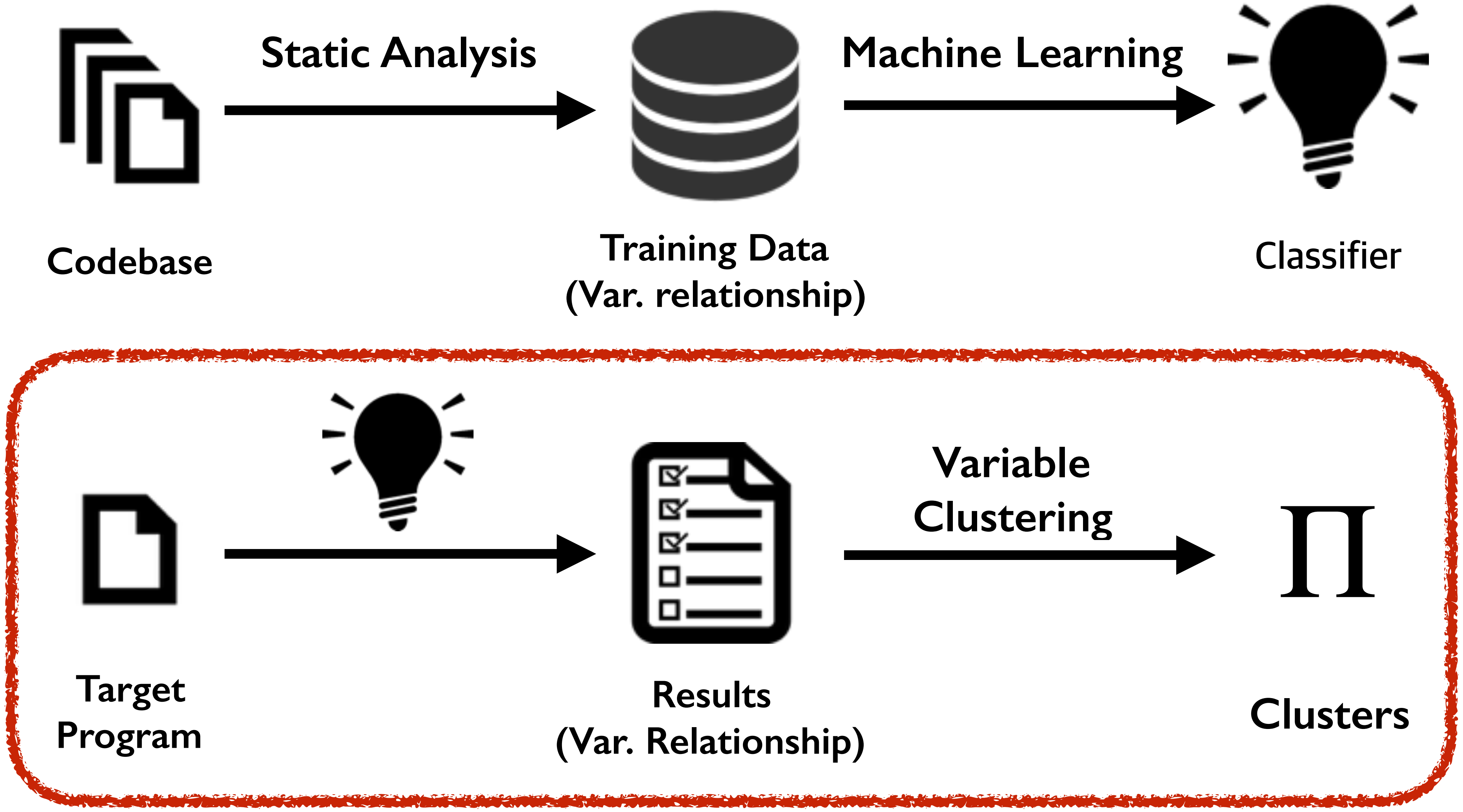
*Top 5 most important features

Classifier

- Learning a binary classifier $\mathcal{C} : Var \times Var \to \{\oplus, \ominus\}$
  - using an off-the-shelf ML algorithm: decision tree
- Why decision tree?
  - more expressive than linear models
  - e.g.) Octagon with logistic regression : 10~12x slower

Big Picture

- Learning a variable-clustering strategy from big data



Static Analysis

Machine Learning

Codebase

Training Data
(Var. relationship)

Classifier

Variable
Clustering

Π

Target
Program

Results
(Var. Relationship)

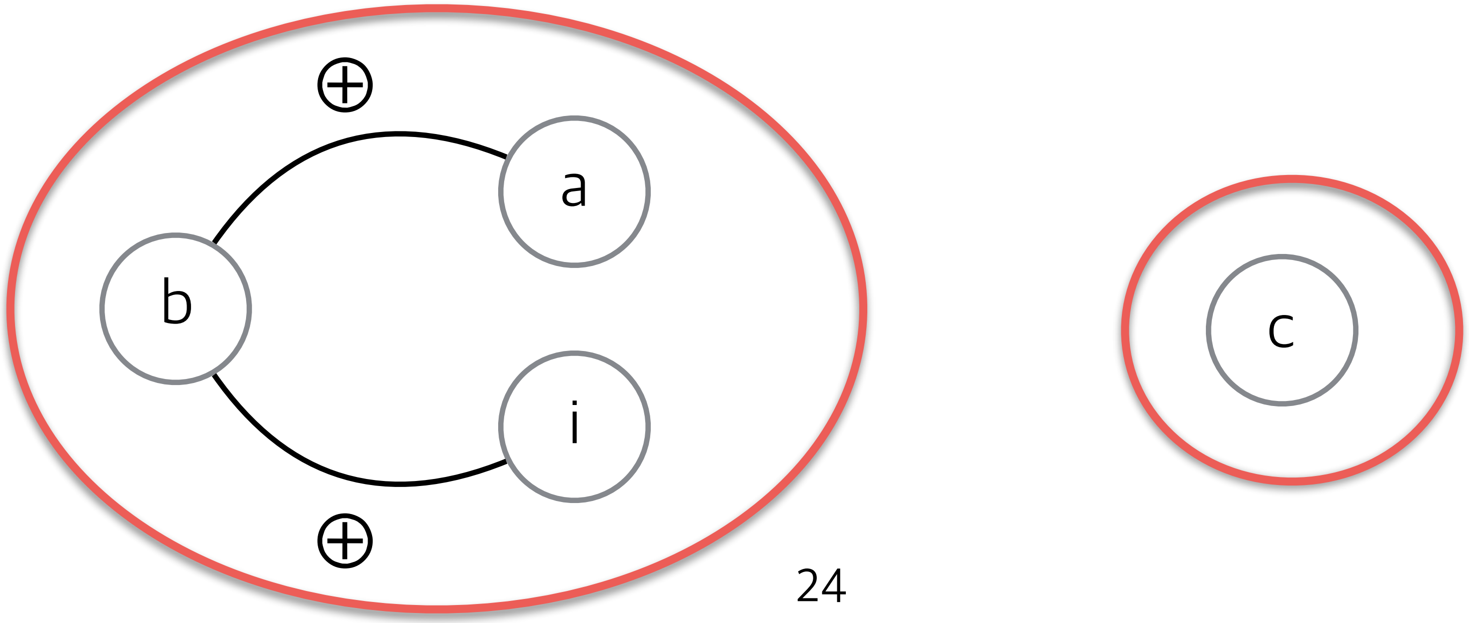
Clusters

Clustering Strategy

- ⊕-marked variable pairs in the same cluster
  - naturally covers transitive relationships

```
int a = b;
int c = input();            // User input
for (i = 0; i < b; i++) {
  assert (i < a);           // Query 1
  assert (i < c);           // Query 2
}
```

	C(x,y)
(a,b)	⊕
(a,i)	⊖
(b,i)	⊕
(a,c)	⊖
…	…



Experiments

- Implemented on top of Sparrow The Early Bird
  - sound & global analyzer
  - a buffer overrun detector for full C
- 17 open source benchmarks (~100KLOC)

Experimental Results

- Effectiveness (leave-one-out cross validation)

Program	LOC	#Abs.Loc.	# Alarms			Time(s)		
			Itv	Impt	ML	Itv	Impt	ML
brutefir	103	54	4	0	0	0	0	0
consol	298	165	20	10	10	0	0	0
id3	512	527	15	6	6	0	0	1
spell	2,213	450	20	8	17	0	1	1
mp3rename	2,466	332	33	3	3	0	1	1
irmp3	3,797	523	2	0	0	1	2	3
barcode	4,460	1,738	235	215	215	2	9	6
httptunnel	6,174	1,622	52	29	27	3	35	5
e2ps	6,222	1,437	119	58	58	3	6	3
bc	13,093	1,891	371	364	364	14	252	16
less	23,822	3,682	625	620	625	83	2,354	87
bison	56,361	14,610	1,988	1,955	1,955	137	4,827	237
pies	66,196	9,472	795	785	785	49	14,942	95
icecast-server	68,564	6,183	239	232	232	51	109	107
raptor	76,378	8,889	2,156	2,148	2,148	242	17,844	345
dico	84,333	4,349	402	396	396	38	156	51
lsh	110,898	18,880	330	325	325	33	139	251
Total			7,406	7,154	7,166	656	40,677	1,207

Experimental Results

- Effectiveness (leave-one-out cross validation)

Program	LOC	#Abs.Loc.	# Alarms			Time(s)		
			Itv	Impt	ML	Itv	Impt	ML
brutefir	103	54	4	0	0	0	0	0
consol	298	165	20	10	10	0	0	0
id3	512	527	15	6	6	0	0	1
spell	2,213	450	20	8	17	0	1	1
mp3rename	2,466	332	33	3	3	0	1	1
irmp3	3,797	523	2	0	0	1	2	3
barcode	4,460	1,738	235	215	215	2	9	6
httptunnel	6,174	1,622	52	29	27	3	35	5
e2ps	6,222	1,437	119	58	58	3	6	3
bc	13,093	1,891	371	364	364	14	252	16
less	23,822	3,682	625	620	625	83	2,354	87
bison	56,361	14,610	1,988	1,955	1,955	137	4,827	237
pies	66,196	9,472	795	785	785	49	14,942	95
icecast-server	68,564	6,183	239	232	232	51	109	107
raptor	76,378	8,889	2,156	2,148	2,148	242	17,844	345
dico	84,333	4,349	402	396	396	38	156	51
lsh	110,898	18,880	330	325	325	33	139	251
Total			7,406	7,154	7,166	656	40,677	1,207
				-252	-240			

Experimental Results

- Effectiveness (leave-one-out cross validation)

Program	LOC	#Abs.Loc.	# Alarms			Time(s)		
			Itv	Impt	ML	Itv	Impt	ML
brutefir	103	54	4	0	0	0	0	0
consol	298	165	20	10	10	0	0	0
id3	512	527	15	6	6	0	0	1
spell	2,213	450	20	8	17	0	1	1
mp3rename	2,466	332	33	3	3	0	1	1
irmp3	3,797	523	2	0	0	1	2	3
barcode	4,460	1,738	235	215	215	2	9	6
httptunnel	6,174	1,622	52	29	27	3	35	5
e2ps	6,222	1,437	119	58	58	3	6	3
bc	13,093	1,891	371	364	364	14	252	16
less	23,822	3,682	625	620	625	83	2,354	87
bison	56,361	14,610	1,988	1,955	1,955	137	4,827	237
pies	66,196	9,472	795	785	785	49	14,942	95
icecast-server	68,564	6,183	239	232	232	51	109	107
raptor	76,378	8,889	2,156	2,148	2,148	242	17,844	345
dico	84,333	4,349	402	396	396	38	156	51
lsh	110,898	18,880	330	325	325	33	139	251
Total			7,406	7,154	7,166	656	40,677	1,207
				-252	-240			

Experimental Results

- Effectiveness (leave-one-out cross validation)

Program	LOC	#Abs.Loc.	# Alarms			Time(s)		
			Itv	Impt	ML	Itv	Impt	ML
brutefir	103	54	4	0	0	0	0	0
consol	298	165	20	10	10	0	0	0
id3	512	527	15	6	6	0	0	1
spell	2,213	450	20	8	17	0	1	1
mp3rename	2,466	332	33	3	3	0	1	1
irmp3	3,797	523	2	0	0	1	2	3
barcode	4,460	1,738	235	215	215	2	9	6
httptunnel	6,174	1,622	52	29	27	3	35	5
e2ps	6,222	1,437	119	58	58	3	6	3
bc	13,093	1,891	371	364	364	14	252	16
less	23,822	3,682	625	620	625	83	2,354	87
bison	56,361	14,610	1,988	1,955	1,955	137	4,827	237
pies	66,196	9,472	795	785	785	49	14,942	95
icecast-server	68,564	6,183	239	232	232	51	109	107
raptor	76,378	8,889	2,156	2,148	2,148	242	17,844	345
dico	84,333	4,349	402	396	396	38	156	51
lsh	110,898	18,880	330	325	325	33	139	251
Total			7,406	7,154	7,166	656	40,677	1,207
							x62	x2

Experimental Results

- Generalization : training only with small (<60KLOC) pgms

Program	LOC	Abs. Loc.	# Alarms			Time(s)		
			Itv	All	Small	Itv	All	Small
pies	66,196	9,472	795	785	785	49	95	98
icecast-server	68,564	6,183	239	232	232	51	113	99
raptor	76,378	8,889	2,156	2,148	2,148	242	345	388
dico	84,333	4,349	402	396	396	38	61	62
lsh	110,898	18,880	330	325	325	33	251	251
Total			7,406	3,886	3,886	413	865	898

+4%

Summary

Bugzilla

stackoverflow

sourceforge

+

Sparrow

The Early Bird

- Adaptive variable-clustering strategy for Octagon
  - Machine Learning (learner) + Static Analysis (teacher)
- 33x faster than a static-analysis-only approach

Summary

Bugzilla

stackoverflow

sourceforge

+

Sparrow

The Early Bird

- Adaptive variable-clustering strategy for Octagon
  - **Machine Learning** (learner) + **Static Analysis** (teacher)
- 33x faster than a static-analysis-only approach

Thank You