Pointer Analysis

Rupesh Nasre.

Outline

- Introduction
- Pointer analysis as a DFA problem
- Design decisions
- Andersen's analysis, Steensgaard's analysis
- Pointer analysis as a graph problem
    - Optimizations
- Pointer analysis as graph rewrite rules
- Applications
- Parallelization
    - Constraint based
    - Replication based

What is Pointer Analysis?

```
a = &x;
b = a;
if (b == *p) {
    ...
} else {
    ...
}
```

What is Points-to Analysis?

a = &x;

a points to x

b = a;

if (b == *p) {

  …

} **else** {

  …

}

What is Points-to Analysis?

a = &x;

a points to x

b = a;

a and b are aliases

if (b == *p) {

...

} **else** {

...

}

What is Points-to Analysis?

```
a = &x;
b = a;
if (b == *p) {
    ...
} else {
    ...
}
```

a points to x

a and b are aliases

Is this condition always satisfied?

What is Points-to Analysis?

```
a = &x;
b = a;
if (b == *p) {
    …
} else {
    …
}
```

a points to x

a and b are aliases

Is this condition always satisfied?

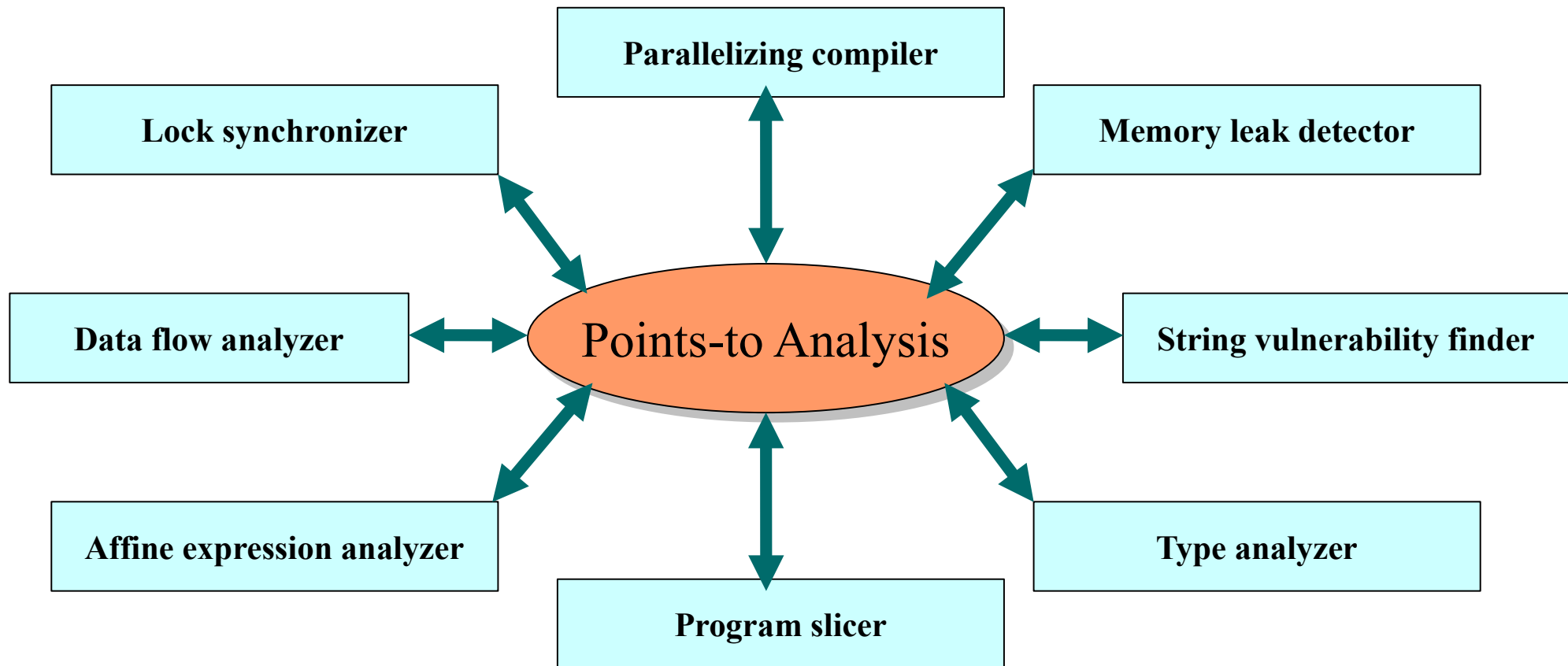Pointer Analysis is a mechanism to **statically** find out run-time values of a pointer.

Why Points-to Analysis?

- for Parallelization
  - fun(p) || fun(q)
- for Optimization
  - a = p + 2;
  - b = q + 2;
- for Bug-Finding
- for Program Understanding
- ...
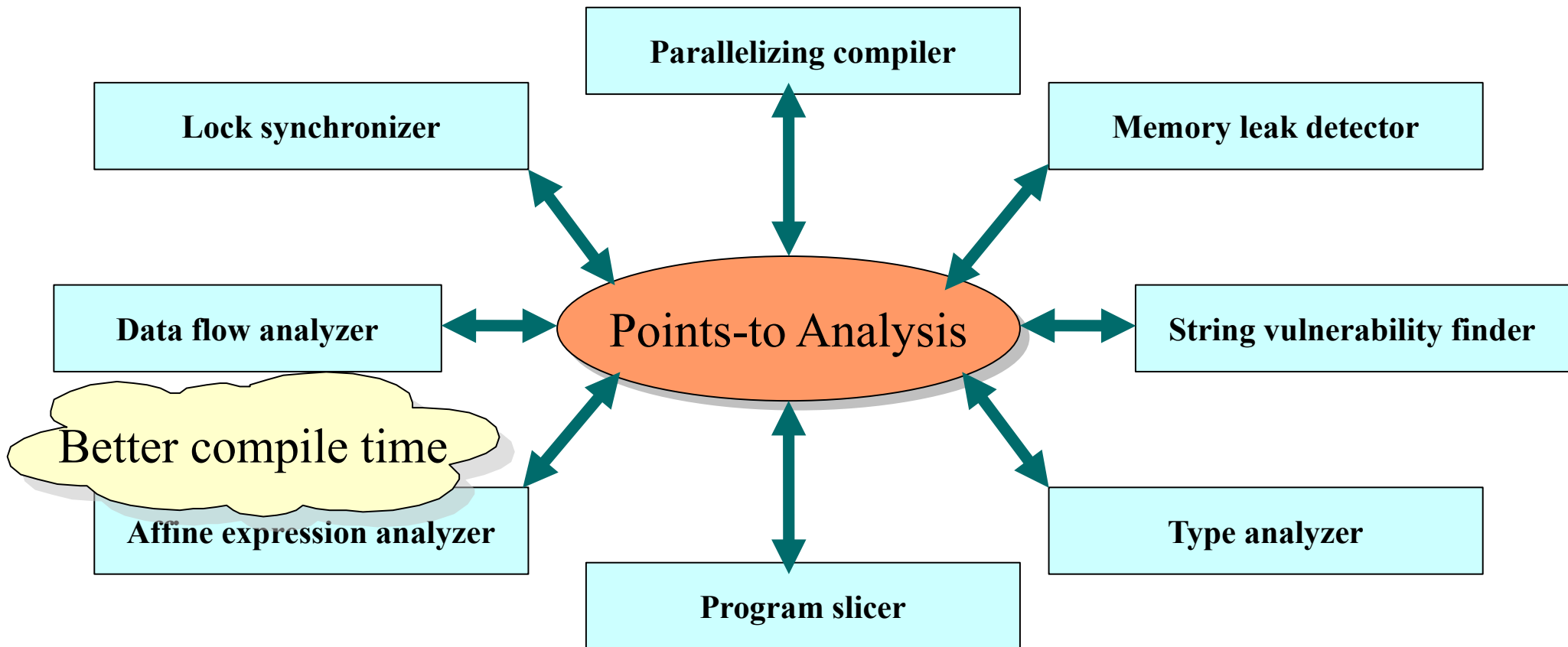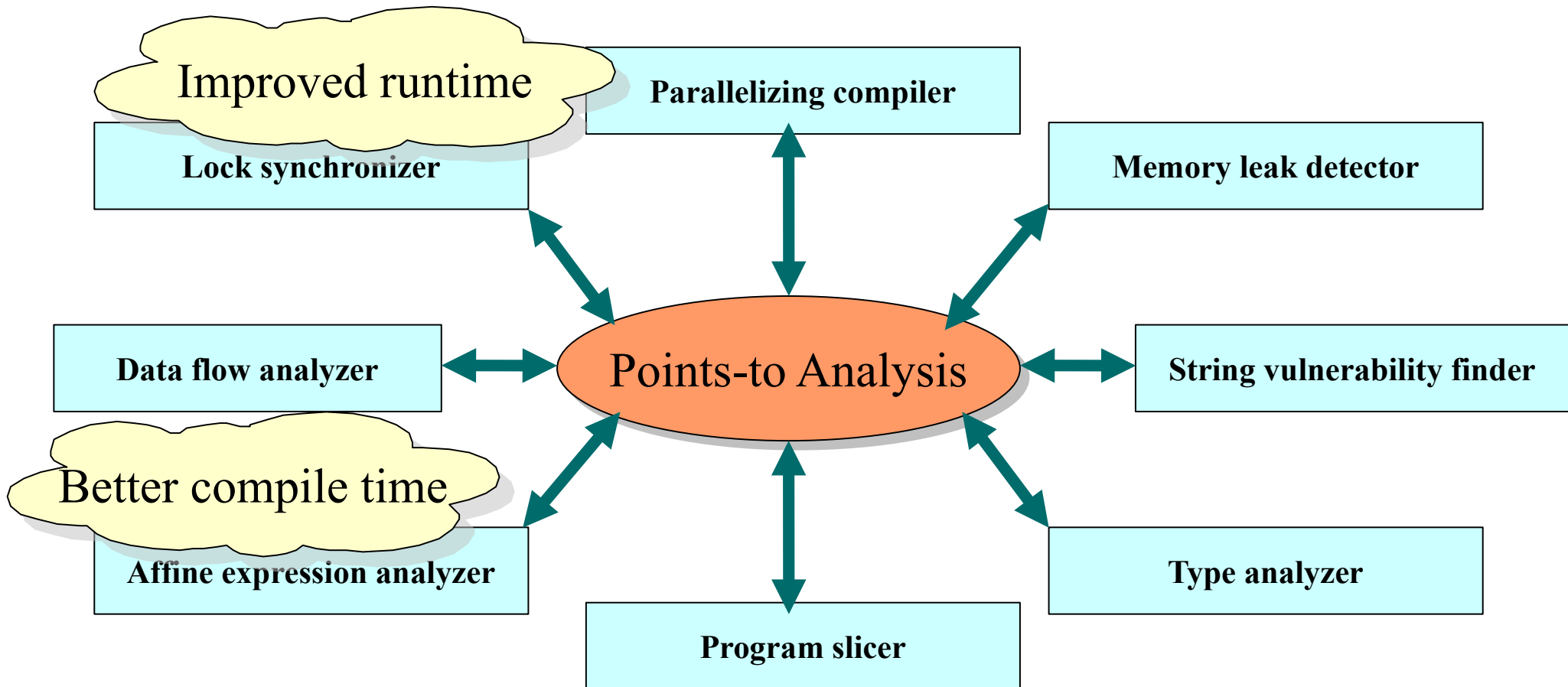
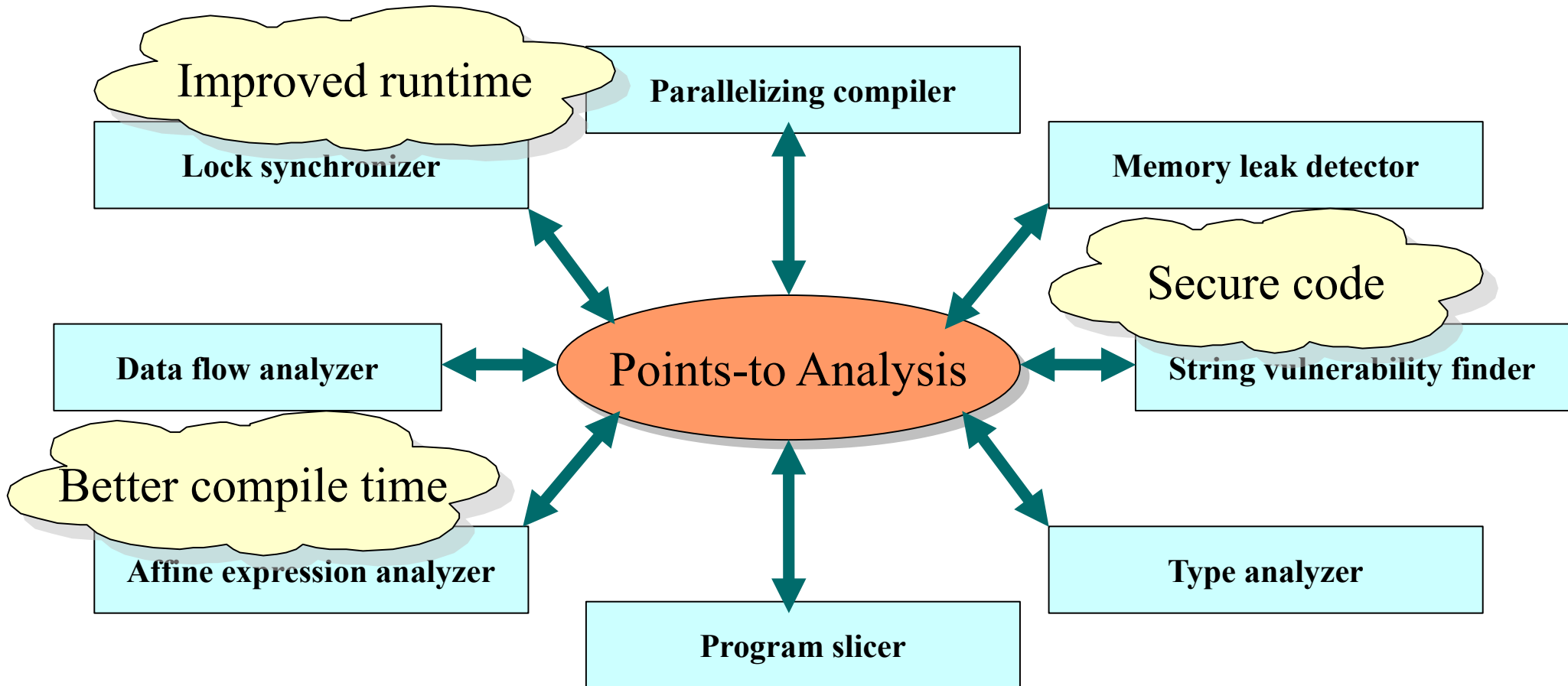**Clients of
Points-to Analysis**

Placement of Points-to Analysis

Placement of Points-to Analysis

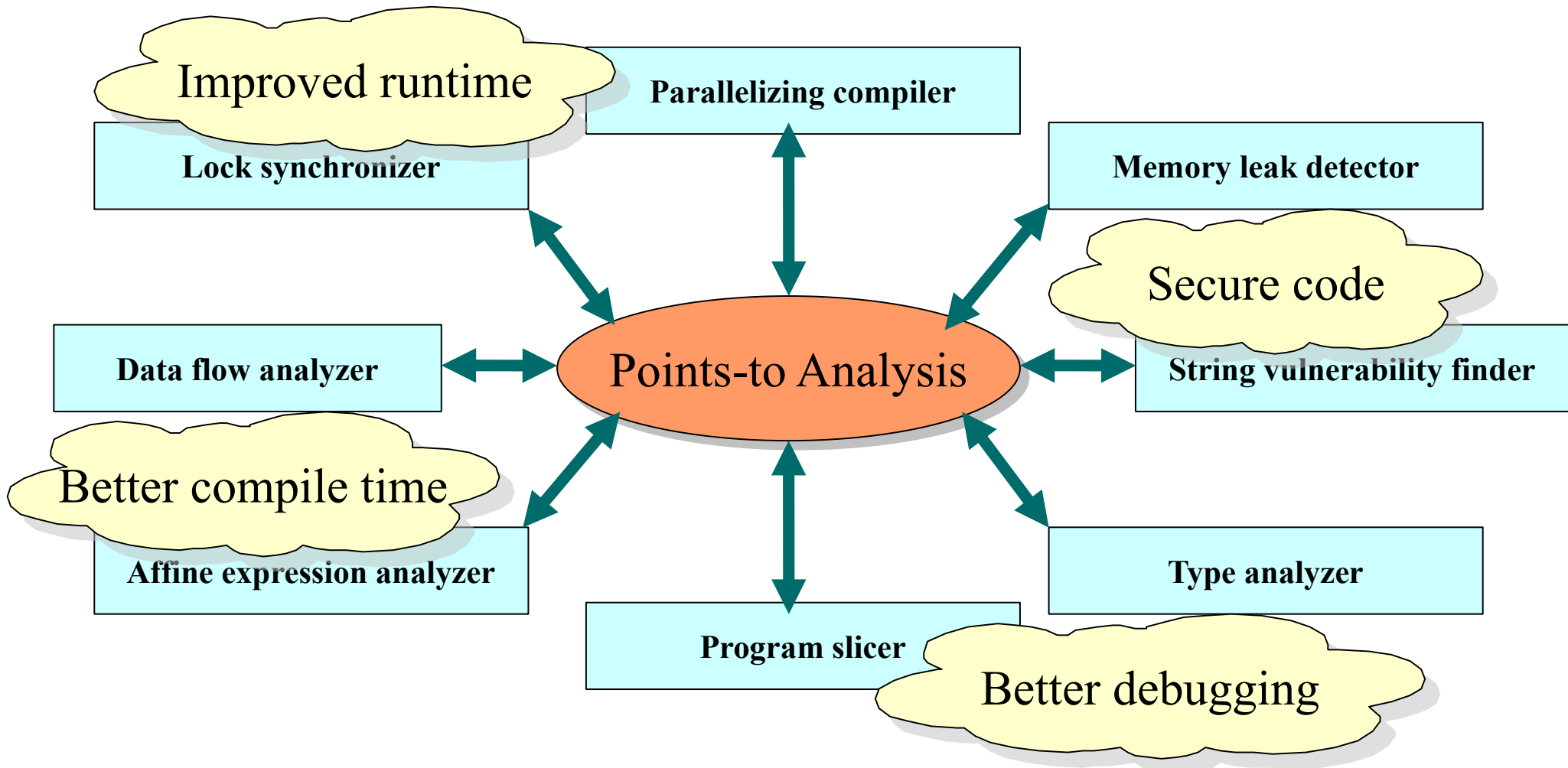# Placement of Points-to Analysis

Improved runtime

Parallelizing compiler

Memory leak detector

Lock synchronizer

Data flow analyzer

Points-to Analysis

String vulnerability finder

Better compile time

Affine expression analyzer

Type analyzer

Program slicer

Placement of Points-to Analysis

Improved runtime

Parallelizing compiler

Lock synchronizer

Memory leak detector

Secure code

Data flow analyzer

Points-to Analysis

String vulnerability finder

Better compile time

Affine expression analyzer

Program slicer

Type analyzer

Placement of Points-to Analysis

Improved runtime

Parallelizing compiler

Lock synchronizer

Memory leak detector

Secure code

Data flow analyzer

Points-to Analysis

String vulnerability finder

Better compile time

Affine expression analyzer

Type analyzer

Program slicer

Better debugging

13

Learning Outcomes

- To apply data-flow analysis and its variants on input programs and collect relevant information

  – Apply Andersen's analysis to an input program.

  – Apply Steensgaard's analysis to a program.

  – Compute points-to information with precision in-between Andersen's and Steensgaard's analyses.

  – Apply various analysis dimensions to compute different points-to information for the same program.

  – Define realizable facts.

- To design and implement analyses for new problems

Points-to Analysis

A C program can be normalized to contain only four types of pointer-manipulating statements or constraints.
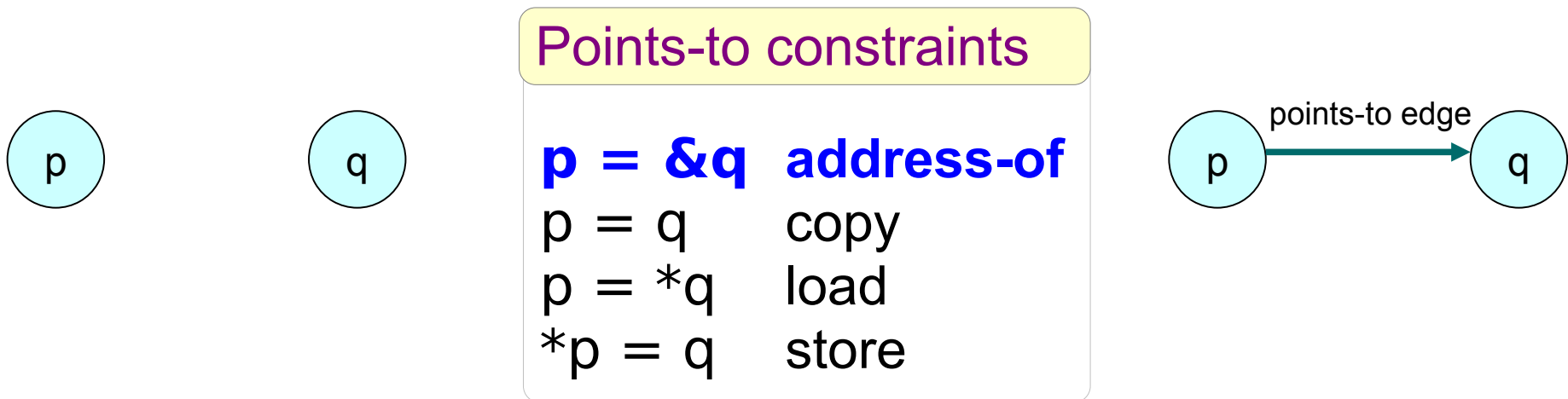
- NULL is treated as a special pointee.

- Uninitialized or dangling pointers need special treatment (for now assume their absence).

Points-to constraints

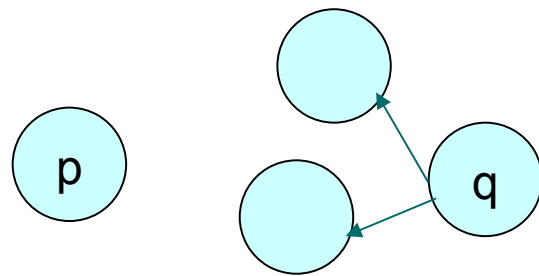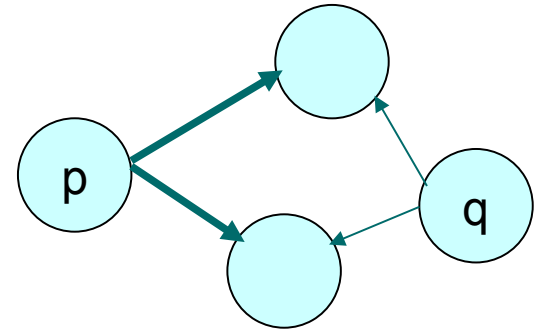|	
p = &q	address-of
p = q	copy
p = *q	load
*p = q	store

Points-to Analysis

A C program can be normalized to contain only four types of pointer-manipulating statements or constraints.

Points-to constraints

p = &q	**address-of**
p = q	copy
p = *q	load
*p = q	store

Points-to Analysis

A C program can be normalized to contain only four types of pointer-manipulating statements or constraints.

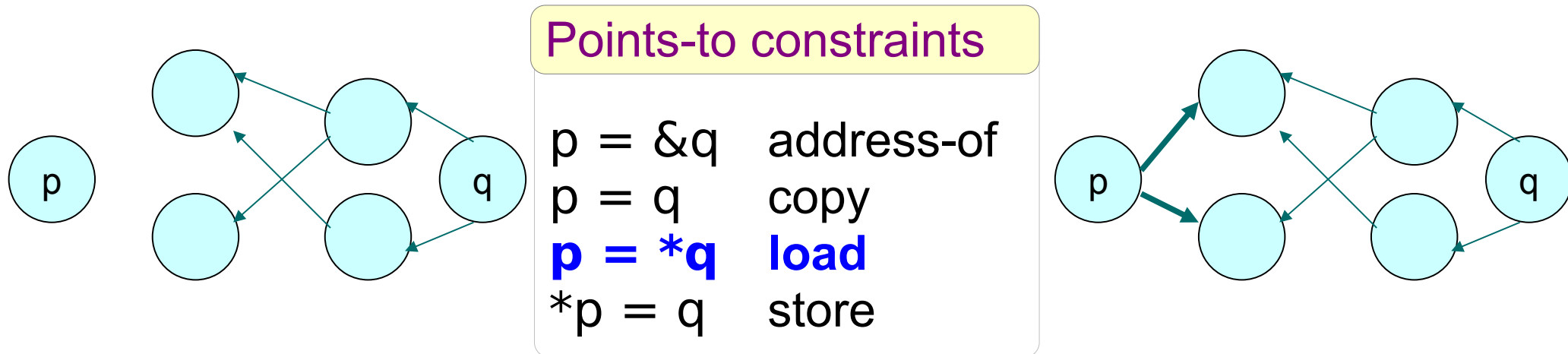Points-to constraints

p = &q	address-of
p = q	**copy**
p = *q	load
*p = q	store

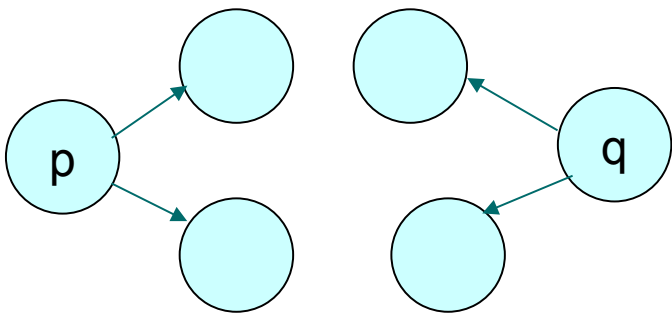# Points-to Analysis

A C program can be normalized to contain only four types of pointer-manipulating statements or constraints.

Points-to constraints

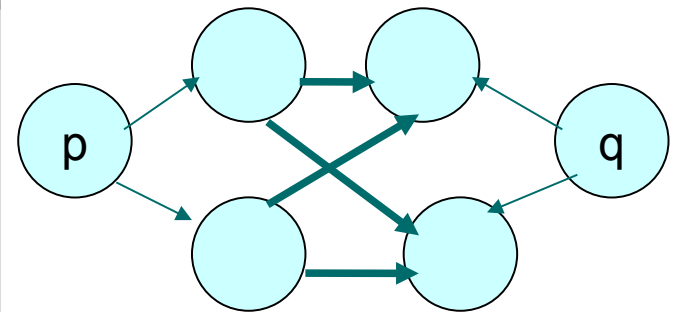|	
p = &q	address-of
p = q	copy
p = *q	**load**
*p = q	store

Points-to Analysis

A C program can be normalized to contain only four types of pointer-manipulating statements or constraints.

Points-to constraints

p = &q	address-of
p = q	copy
p = *q	load
*p = q	store

Definitions

- **Points-to analysis** computes points-to information for each pointer.

- **Alias analysis** computes aliasing information for all pointers.

- Aliasing information can be computed using points-to information, but not vice versa.

- Clients often query for aliasing information, but storing it is expensive $O(n^2)$, hence frameworks store points-to information.

- If $a \to x$, x is often called a **pointee** of a.

Points-to information

$a \to \{x, y\}$
$b \to \{y, z\}$
$c \to \{z\}$

Aliasing information

	a	b	c
a	--	Yes	No
b	--	--	Yes
c	--	--	--

Nomenclature

- Pointer analysis: Ambiguous usage in literature. We will use it to refer to both points-to analysis and alias analysis.

- In the context of Java-like languages, it is called reference analysis.

- Also called as heap analysis.

Algebraic Properties

- Aliasing relation is reflexive, symmetric, but not transitive.

- Points-to relation is neither reflexive, nor symmetric, not even transitive.

- The points-to relation induces a restricted DAG for strictly typed languages.

Cyclic Dependence

- ## Call graph ↔ function pointers

  - We build a call graph without running pointer analysis (conservative). Once we have points-to information, some edges in the call graph reduce. This can further reduce points-to information.

- ## Optimization ↔ points-to information

  - We can feed initial points-to information for performing optimizations (e.g., DCE). After removal of some statements, points-to information can become more precise.

As a DFA

a = &x: gen{a → x} and kill
a = b: gen{a → x} if {b → x} and kill
a = *p: gen{a → x} if {p → b → x} and kill
*p = a: gen{b → x} if {p → b and a → x} and
             kill{b → x} if {p → b and b → x} iff
             p's points-to set is a singleton.

             // based on the *In* sets and
             // internal killing

In(B) = U Out(P) where P ∈ Pred(B)
Out(B) = Gen(B) U (In(B) – Kill(B))

```
        b = &y
        p = &a      B0
        b = a
        if c
       /      \
  B1  a = b    *p = b   B2
      b = &z    a = &x
       \      /
         B3
```

gen(B0) = {p→a, b→x if a→x}   kill(B0) = {ptsto(p), ptsto(b)}
gen(B1) = {a→x if b→x, b→z}   kill(B1) = {ptsto(a), ptsto(b)}
gen(B2) = {a→x,m→n if p→m and   kill(B2) = {ptsto(ptsto(p)), ptsto(a)}
             b→n and m ≠ a}
gen(B3) = { }   kill(B3) = { }
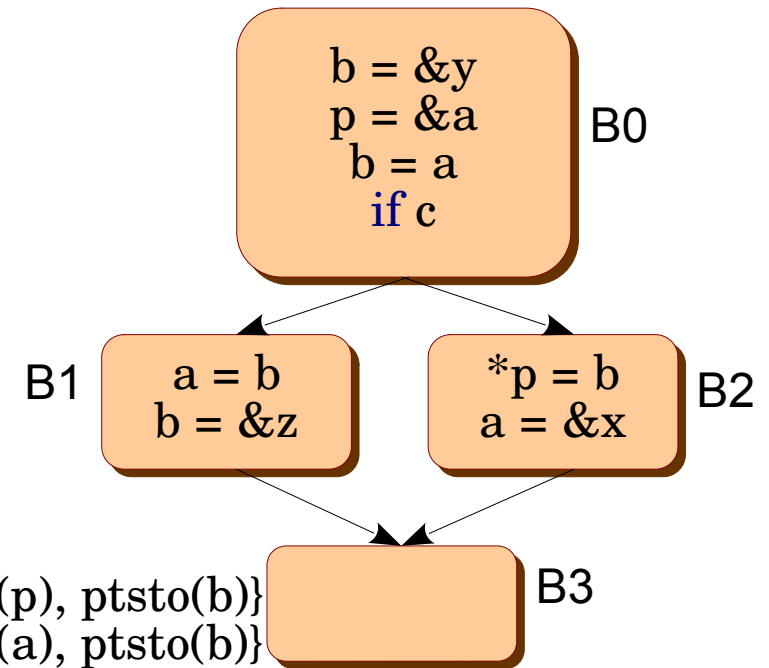
Note: p→a is really p→{a}.

	in1	out1	in2	out2	in3	out3
B0	{}	{p→a}	{}	{p→a}	{}	{p→a}
B1	{}	{b→z}	{p→a}	{p→a,b→z}	{p→a}	{p→a,b→z}
B2	{}	{a→x}	{p→a}	{p→a,a→x}	{p→a}	{p→a,a→x}
B3	{}	{}	{a→x, b→z}	{p→a,a→x,b→z}	{p→a,a→x,b→z}	{p→a,a→x,b→z}

As a DFA: Notes

- Gen and Kill are dynamic (not fixed before analysis).

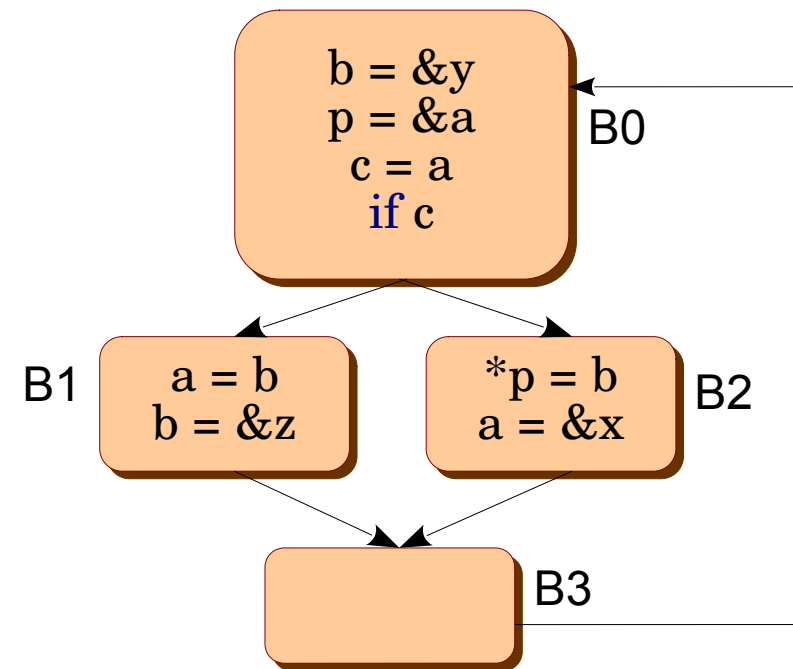- Gen/Kill and Points-to Information are cyclically dependent.

Classwork

a = &x: gen{a ⇀ x} and kill
a = b: gen{a ⇀ x} if {b ⇀ x} and kill
a = *p: gen{a ⇀ x} if {p ⇀ b ⇀ x} and kill
*p = a: gen{b ⇀ x} if {p ⇀ b and a ⇀ x} and
         kill{b ⇀ x} if {p ⇀ b and b ⇀ x} iff
         p's points-to set is a singleton.

         // based on the *In* sets and
         // internal killing

In(B) = ∪ Out(P) where P ∈ Pred(B)
Out(B) = Gen(B) ∪ (In(B) − Kill(B))

b = &y
p = &a
c = a
if c

B0

B1
a = b
b = &z

*p = b
a = &x
B2

B3

26

Design Decisions

- Analysis dimensions

- Heap modeling

- Set implementation

- Call graph, function pointers

- Array indices

Time

Holy grail

Memory
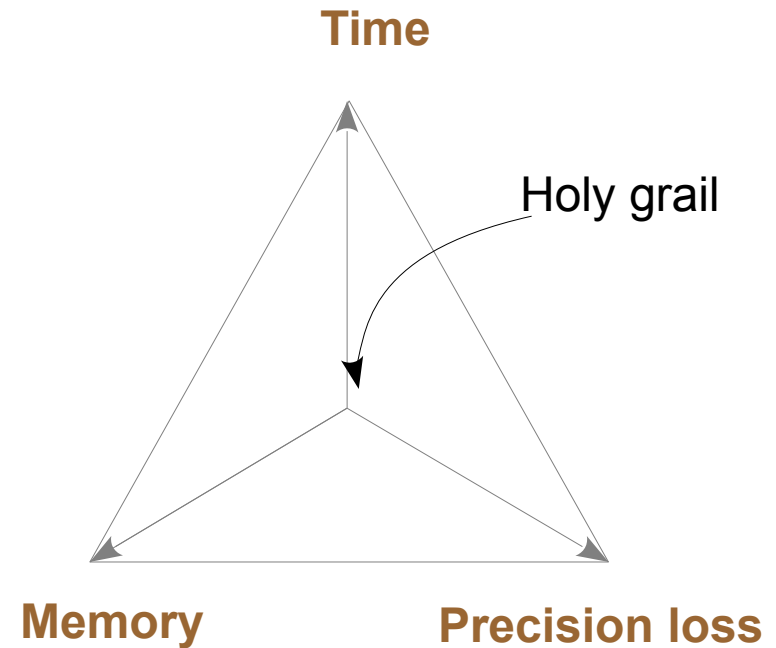
Precision loss

Analysis Dimensions

An analysis's precision and efficiency is guided by various design decisions.

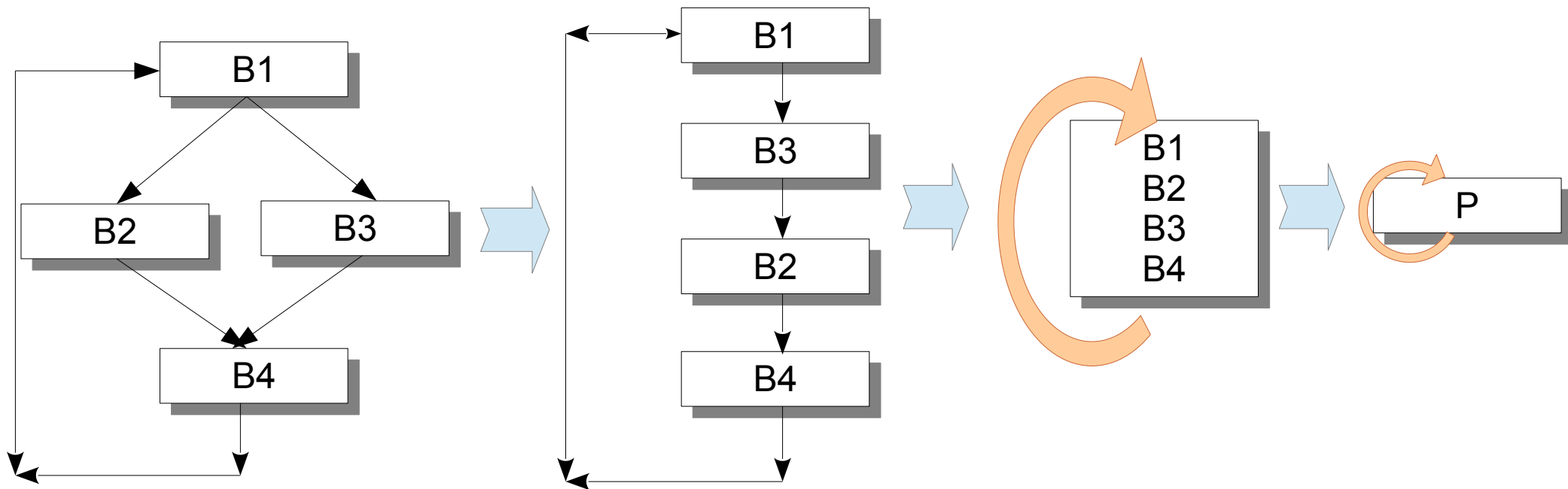- Flow-sensitivity
- Context-sensitivity
- Path-sensitivity
- Field-sensitivity

Flow-sensitivity

```
L0:  a = &x;
L1:  a = &y;
L2:  ...
```

Flow-sensitive solution: *at L1 a points to x, at L2 a points to y*
Flow-insensitive solution: *in the program a's points-to set is {x, y}*

Flow-insensitive analyses ignore the control-flow in the program.

Context-sensitivity

```
main() {                fun(int *a) {
    L0:  fun(&x);            b = a;
    L1:  fun(&y);        }
}
```

Context-sensitive solution:
b points to x along L0, b points to y along L1

Context-insensitive solution:
b's points-to set is {x, y} in the program

main
f   f
g g g g
...

Exponential Number of contexts

Along main-f1-g1, …
Along main-f1-g2, …
Along main-f2-g1, …
Along main-f2-g2, ...

Exponential time requirement

Exponential storage requirement

30

Context-sensitivity

```
main() {            fun(int *a) {
    L0:  fun(&x);       b = a;
    L1:  fun(&y);   }
}
```
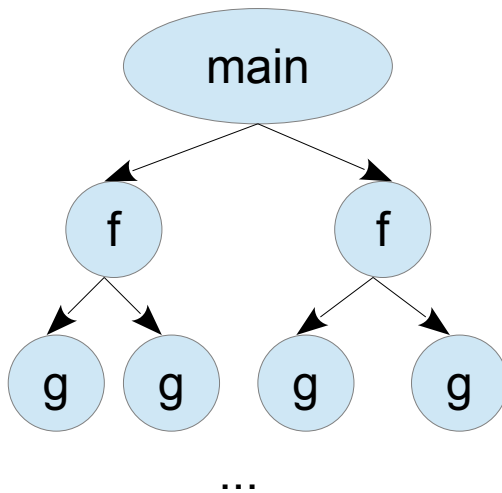
Context-sensitive solution:
b points to x along L0, b points to y along L1

Context-insensitive solution:

Inter-procedural ⟶ *b's points-to set is {x, y} in the program*

intra-procedural ⟶ *b's points-to set is {all address-taken variables}*

Path-sensitivity

```
if (a == 0)
    b = &x;
else
    b = &y;
```

Path-sensitive solution:
   b points-to x when a is 0, b points-to y when a is not 0

Path-insensitive solution:
   b's points-to set is {x, y} in the program

```
if (c1)
  while (c2) {
    if (c3)
      …
    else
      for (; c4; )
        ...
  }
else
  ...
```

c1 and c2 and c3, …
c1 and c2 and !c3 and c4, …
c1 and c2 and !c3 and !c4, …
c1 and !c2, …
!c1 ...
...

Field-sensitivity

```
struct T s;

s.a = &x;
s.b = &y;
```

Field-sensitive solution:
s.a points-to x, s.b points-to y

Field-insensitive solution:
s's points-to set is {x, y}

Aggregates are collapsed into a single variable.
 e.g., arrays, structures, unions.

This reduces the number of variables tracked during the analysis and reduces precision.

Andersen's Analysis

- Inclusion-based / subset-based / constraint-based analysis

- Flow-insensitive

For a statement $X = Y$,

   create a constraint $\mathrm{ptsto}(X) \supseteq \mathrm{ptsto}(Y)$

   where X is of the form {*a, a}, and

   Y is of the form {*a, a, &a}.

Solving these inclusion constraints results in the points-to solution.

Andersen's Analysis: Example

Program

```
a = &x;
b = &y;
p = &a;
c = b;
*p = c;
```

Constraints

$ptsto(a) \supseteq \{x\}$
$ptsto(b) \supseteq \{y\}$
$ptsto(p) \supseteq \{a\}$
$ptsto(c) \supseteq ptsto(b)$
$ptsto(*p) \supseteq ptsto(c)$

fixed-point

Pointers	Iteration 0	Iteration 1	Iteration 2
a	{ }	{x, y}	
b	{ }	{y}	
c	{ }	{y}	
p	{ }	{a}	
x	{ }		
y	{ }		

Imprecision

Andersen's Analysis: Modified Example

Program

```
a = &x;
b = &y;
p = &a;
*p = c;
c = b;
```

Constraints

$ptsto(a) \supseteq \{x\}$
$ptsto(b) \supseteq \{y\}$
$ptsto(p) \supseteq \{a\}$
$ptsto(*p) \supseteq ptsto(c)$
$ptsto(c) \supseteq ptsto(b)$

Order does not matter for correctness, but it does matter for efficiency.

fixed-point

Pointers	Iteration 0	Iteration 1	Iteration 2	Iteration 3
a	{ }	{x}	{x, y}	
b	{ }	{y}		
c	{ }	{y}		
p	{ }	{a}		
x	{ }			
y	{ }			

Andersen's Analysis: Classwork

Program

```
*p = c;
b = &y;
b = *p;
p = &a;
a = &x;
*p = c;
c = p;
c = &z;
```

Constraints

$ptsto(*p) \supseteq ptsto(c)$
$ptsto(b) \supseteq \{y\}$
$ptsto(b) \supseteq ptsto(*p)$
$ptsto(p) \supseteq \{a\}$
$ptsto(a) \supseteq \{x\}$
$ptsto(*p) \supseteq ptsto(c)$
$ptsto(c) \supseteq ptsto(p)$
$ptsto(c) \supseteq \{z\}$

fixed-point

Pointers	Iteration 0	Iteration 1	Iteration 2	Iteration 3
a	{ }	{x}	{a, x, z}	
b	{ }	{y}	{a, x, y, z}	
c	{ }	{a, z}	{a, z}	
p	{ }	{a}	{a}	
x	{ }			
y	{ }			
z				

37

Andersen's Analysis: Optimizations

- Avoid duplicates
  - Constraints and Pointer-equivalence

- Reorder constraints
  - Follow def-use chains.

- Process address-of constraints once

- Difference propagation
  - Use diff of points-to sets to propagate.

Andersen's Analysis: Complexity

- Total information computed (storage) = $O(n^2)$

- From each pointer

  To each other pointer

  Propagate O(n) information

  O(n) times

  $O(n^4)$ Naive
  This is incorrect.

- From each pointer

  To each other pointer

  Propagate O(n) information

  $O(n^3)$ Difference Propagation
  This is okay.

Open: Can you reduce the gap between storage and time complexities?

Steensgaard's Analysis
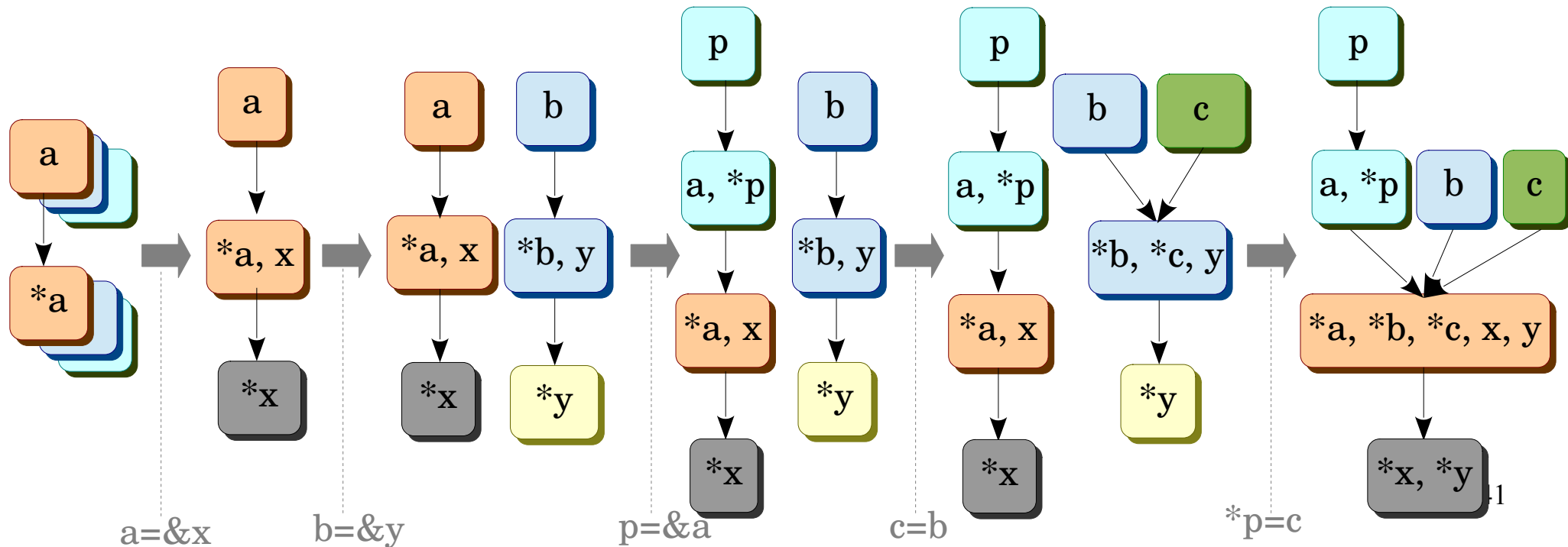
- Unification-based: single points-to edge per pointer.

- Almost linear time $O(m\alpha(m))$

- More imprecise

For a statement $\mathrm{p} = \mathrm{q}$, merge the points-to sets of p and q.

In subset terms, $\mathrm{ptsto(p)} \supseteq \mathrm{ptsto(q)}$ **and** $\mathrm{ptsto(q)} \supseteq \mathrm{ptsto(p)}$ with a single representative element.

Steensgaard's Hierarchy

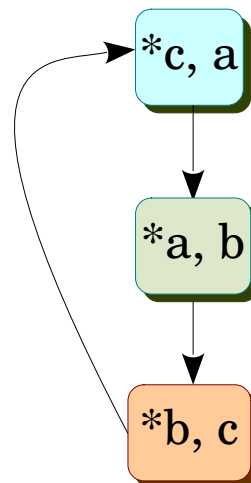| Program	Andersen's	Steensgaard's

a = &x;
b = &y;
p = &a;
c = b;
*p = c;
	a → {x, y}
	b → {y}
	c → {y}
	p → {a}
		a → {x, y}
		b → {x, y}
		c → {x, y}
		p → {a}

a=&x

b=&y

p=&a

c=b

*p=c

Classwork

Program

```
*p = c;
b = &y;
b = *p;
p = &a;
a = &x;
*p = c;
c = p;
c = &z;
```

Andersen's

a → {a, x, z}
b → {a, x, y, z}
c → {a, z}
p → {a}

Steensgaard's

```
a = &b
b = &c
c = &a
```

*c, a

*a, b

*b, c

- Thus, a cycle may be present at the bottom of the hierarchy.
- A cycle collapses the hierarchy beneath it.

Steensgaard's Hierarchy

- What is its structure?

- How many incoming edges to each node?

- How many outgoing edges from each node?

- Can there be cycles?

- What happens to $p = \&p$?

- What is the precision difference between Andersen's and Steensgaard's analyses?

- If for each P = Q, we add Q = P and solve using Andersen's analysis, would it be equivalent to Steensgaard's analysis?

Unifying Model Two

- Steensgaard's hierarchy is characterized by a single outgoing edge.

- Andersen's points-to graph can have arbitrary number of outgoing edges (maximum n).

- Number of edges in between the two provide precision-scalability trade-off.
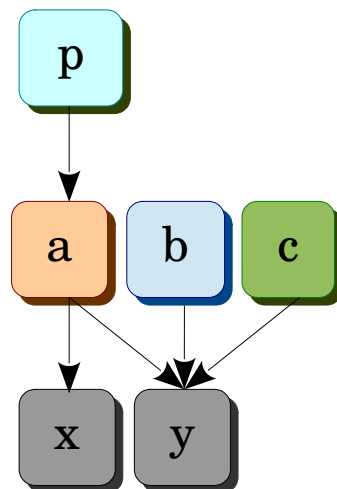
Unifying Model Two

Program	Andersen's	Steensgaard's

a = &x;
b = &y;
p = &a;
c = b;
*p = c;
	a → {x, y}
	b → {y}
	c → {y}
	p → {a}
		a → {x, y}
		b → {x, y}
		c → {x, y}
		p → {a}

Andersen's diagram:

p → a
a, b, c
a → x, b → y, c → y

x   y

Steensgaard's diagram:

p → (a, *p) b c

(a, *p), b, c → (*a, *b, *c, x, y)

*a, *b, *c, x, y → (*x, *y)

Unifying Model Two

Program	Andersen's	Steensgaard's

a = &x;
b = &y;
p = &a;
c = b;
*p = c;
	$a \rightarrow \{x, y\}$
	$b \rightarrow \{y\}$
	$c \rightarrow \{y\}$
	$p \rightarrow \{a\}$
		$a \rightarrow \{x, y\}$
		$b \rightarrow \{x, y\}$
		$c \rightarrow \{x, y\}$
		$p \rightarrow \{a\}$

Unifying Model Two

Program	Andersen's	Steensgaard's

a = &x;
b = &y;
p = &a;
c = b;
*p = c;
	$a \to \{x, y\}$
	$b \to \{y\}$
	$c \to \{y\}$
	$p \to \{a\}$
		$a \to \{x, y\}$
		$b \to \{x, y\}$
		$c \to \{x, y\}$
		$p \to \{a\}$
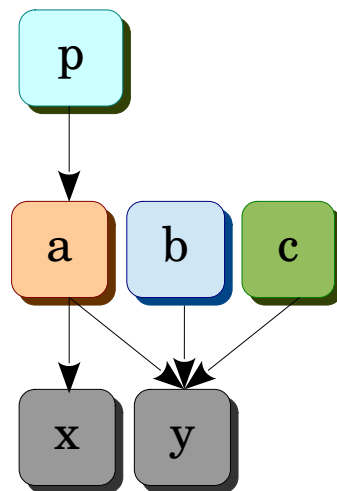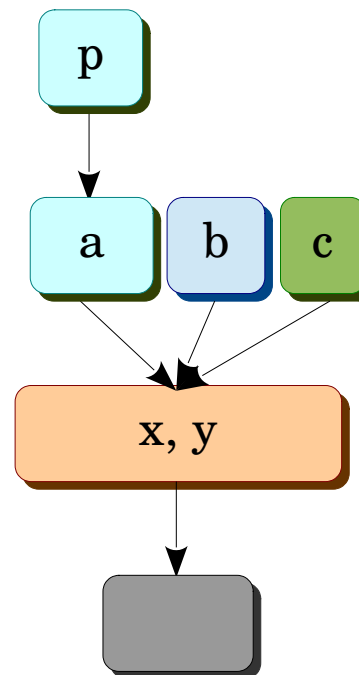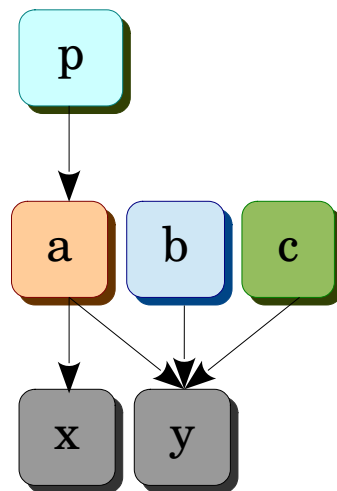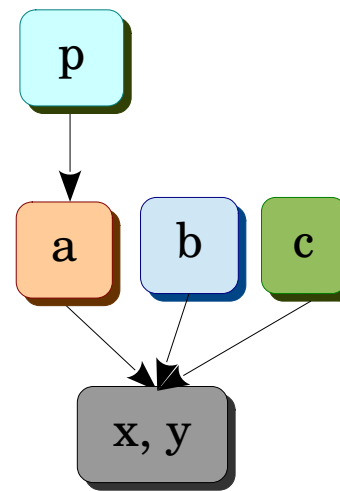
Unifying Model Two

Program	Andersen's	Steensgaard's

a = &x;
b = &y;
p = &a;
c = b;
*p = c;
a = &z;
	$a \to \{x, y, z\}$
	$b \to \{y\}$
	$c \to \{y\}$
	$p \to \{a\}$
		$a \to \{x, y, z\}$
		$b \to \{x, y, z\}$
		$c \to \{x, y, z\}$
		$p \to \{a\}$

Unifying Model Two

| Program	Andersen's	Steensgaard's	In between
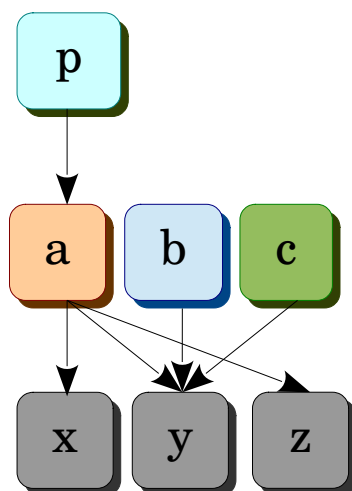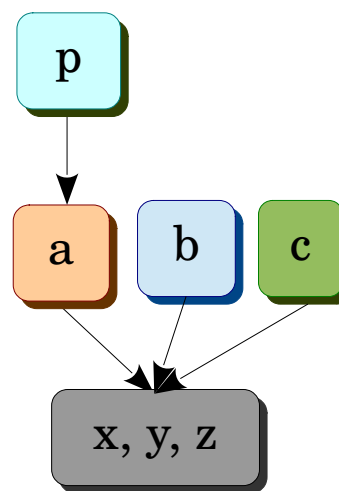
a = &x;
b = &y;
p = &a;
c = b;
*p = c;
a = &z;
	$a \rightarrow \{x, y, z\}$
	$b \rightarrow \{y\}$
	$c \rightarrow \{y\}$
	$p \rightarrow \{a\}$
		$a \rightarrow \{x, y, z\}$
		$b \rightarrow \{x, y, z\}$
		$c \rightarrow \{x, y, z\}$
		$p \rightarrow \{a\}$
			$a \rightarrow \{x, y, z\}$
			$b \rightarrow \{x, y\}$
			$c \rightarrow \{x, y\}$
			$p \rightarrow \{a\}$
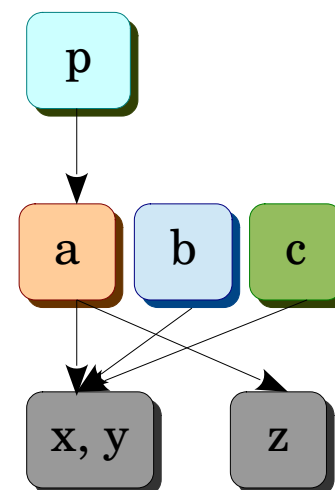
What if x and z are merged?

49

Unifying Model One

- Steensgaard's unification can be viewed as equality of points-to sets.

- Thus, if $a = b$ merges their points-to sets and $b = c$ merges their points-to sets, then a and c become aliases!

- Remember: aliasing is not transitive.

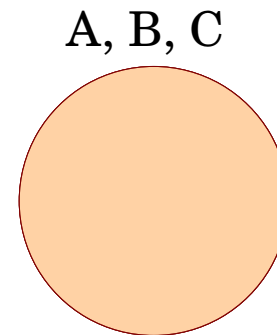- So, unification adds transitivity to the aliasing relation.

Unifying Model One

Andersen's
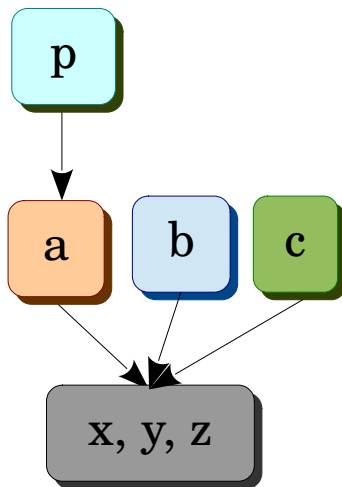
Steensgaard's

A   B   C

A, B, C
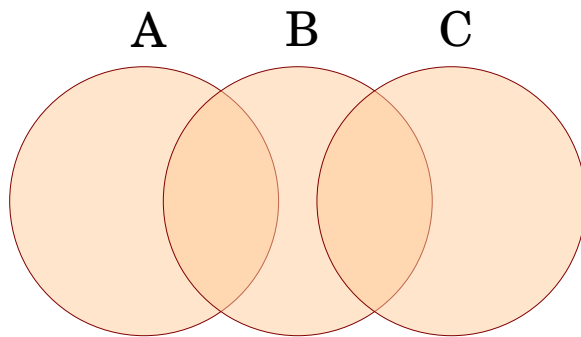
Aliasing is non-transitive

Aliasing becomes transitive

Back to Steensgaard's

- Aliasing relation is transitive.

- We know that it is also reflexive and symmetric.

- This means aliasing becomes an equivalence relation.

- Steensgaard's unification partitions pointers into equivalent sets.

p

a   b   c

x, y, z

All predecessors of a node form a partition.
The equivalence sets are {p}, {a, b, c}, {x, y, z}.
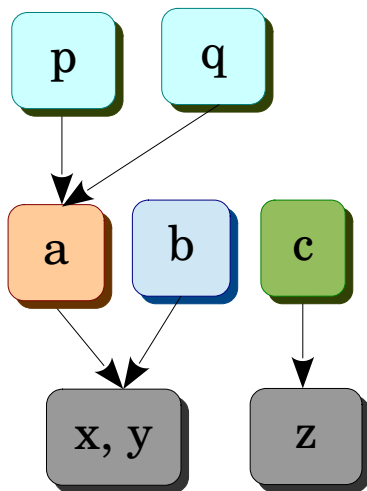
Back to Steensgaard's

- Aliasing relation is transitive.

- We know that it is also reflexive and symmetric.

- This means aliasing becomes an equivalence relation.

- Steensgaard's unification partitions pointers into equivalent sets.

All predecessors of a node form a partition.
The equivalence sets are {p, q}, {a, b}, {c}, {x, y}, {z}.

Realizable Facts

Statements	Andersen's points-to
a = &c	a → {b, c}
b = &a	b → {a, b, c}
c = &b	c → {b}
b = a	d → {a, b, c}
* b = c	
d = * a	

A realizability sequence is a sequence of statements such that a given points-to fact is satisfied.

The realizability sequence for b→c is a=&c, b=a.
The realizability sequence for a→b is c=&b, b=&a, *b=c.
Classwork: What is the realizability sequence for d → a?
Classwork: What is the realizability sequence for d → c?
a→b and b→c are realizable individually, but not simultaneously.

```
int *fun(int *a, int *b) {
    int *c;
    if (*a == *b) {
        c = b;
    } else {
        c = a;
    }
    return c;
}
int *g;
void main() {
    int *x, *y, *z, **w;
    int m = 0, n = 1;
    char *str;
    x = &m;
    y = &n;
    str = (char *)malloc(30);
    w = (int *)&str;
    if (m < n) {
        strcpy(str, "m is smaller\n");
        z = fun(y, x);
    } else {
        printf("m is >= n\n");
        w = &x;
        *w = fun(x, y);
    }
    printf("**w=%d\n", **w);
}
```
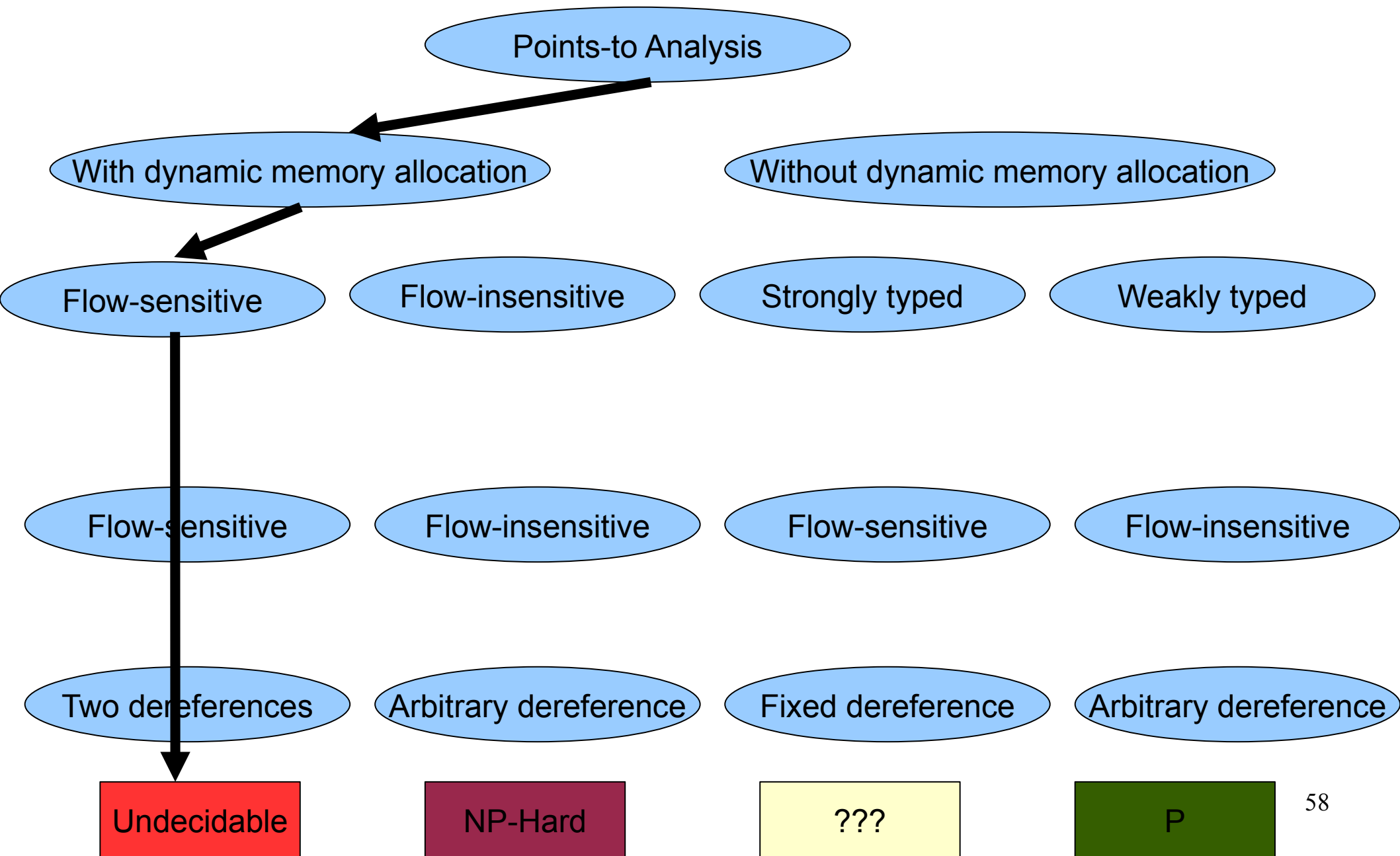
- How do we take care of malloc?
- How do we take care of type-casts?
- Find the set of normalized statements for intra-procedural pointer analysis.
- Perform intra-procedural Andersen's analysis.
- How do we take care of strcpy and printf? How about the global g?
- Perform inter-procedural context-insensitive Andersen's analysis.
- Perform Steensgaard's analysis.

55

Learning Outcomes

- To apply data-flow analysis and its variants on input programs and collect relevant information

  - Apply Andersen's analysis to an input program.

  - Apply Steensgaard's analysis to a program.

  - Compute points-to information with precision in-between Andersen's and Steensgaard's analyses.

  - Apply various analysis dimensions to compute different points-to information for the same program.

  - Define realizable facts.

- To design and implement analyses for new problems

Extra

Complexity of Points-to Analysis

Complexity of Points-to Analysis

Points-to Analysis

With dynamic memory allocation — Without dynamic memory allocation

Flow-sensitive — Flow-insensitive — Strongly typed — Weakly typed

Flow-sensitive — Flow-insensitive — Flow-sensitive — Flow-insensitive

Two dereferences — Arbitrary dereference — Fixed dereference — Arbitrary dereference

Undecidable — NP-Hard — ??? — P

59

Complexity of Points-to Analysis


```
                          Points-to Analysis

    With dynamic memory allocation      Without dynamic memory allocation

  Flow-sensitive    Flow-insensitive    Strongly typed    Weakly typed

  Flow-sensitive    Flow-insensitive    Flow-sensitive    Flow-insensitive

  Two dereferences  Arbitrary dereference  Fixed dereference  Arbitrary dereference

  Undecidable       NP-Hard            ???               P
```
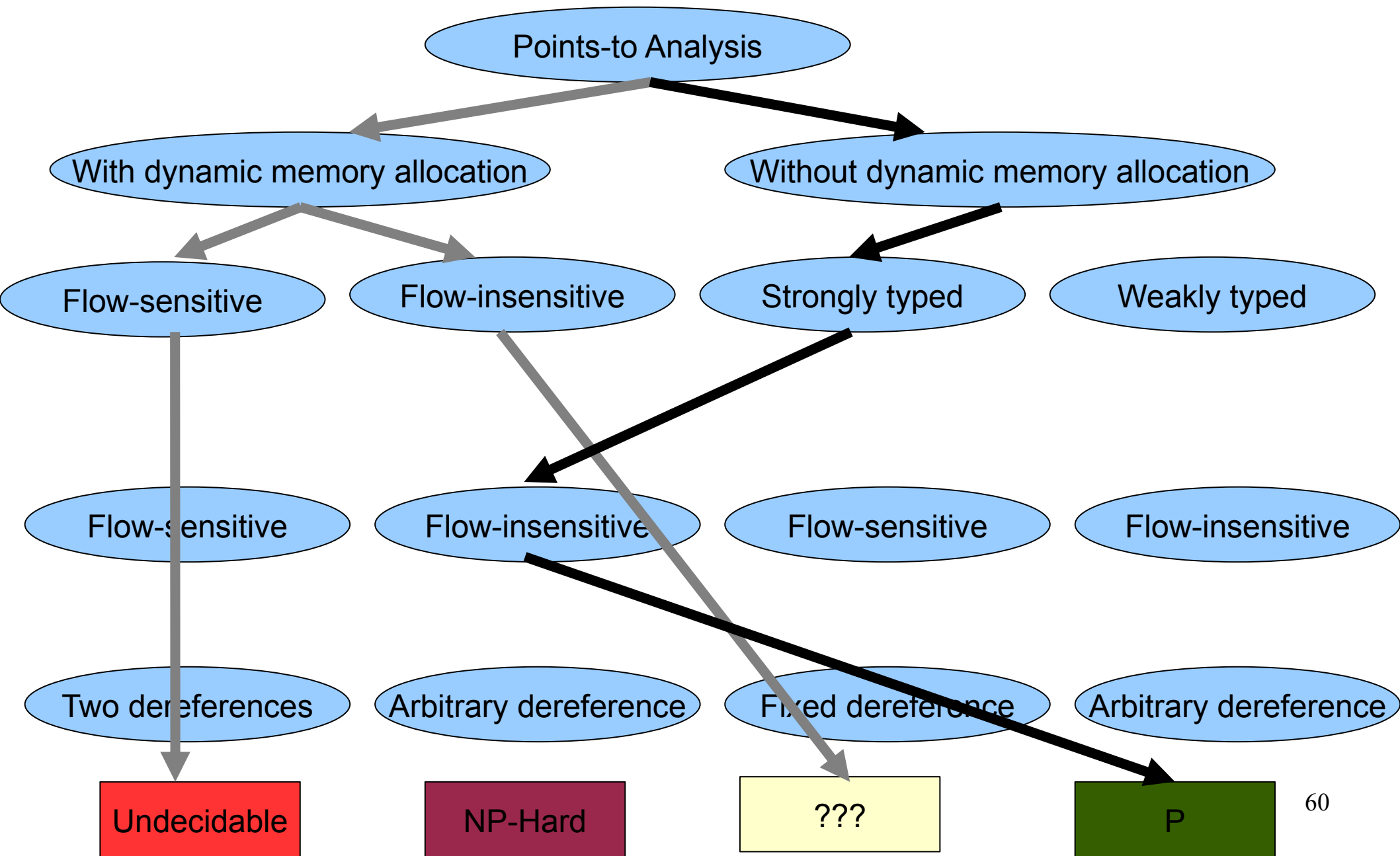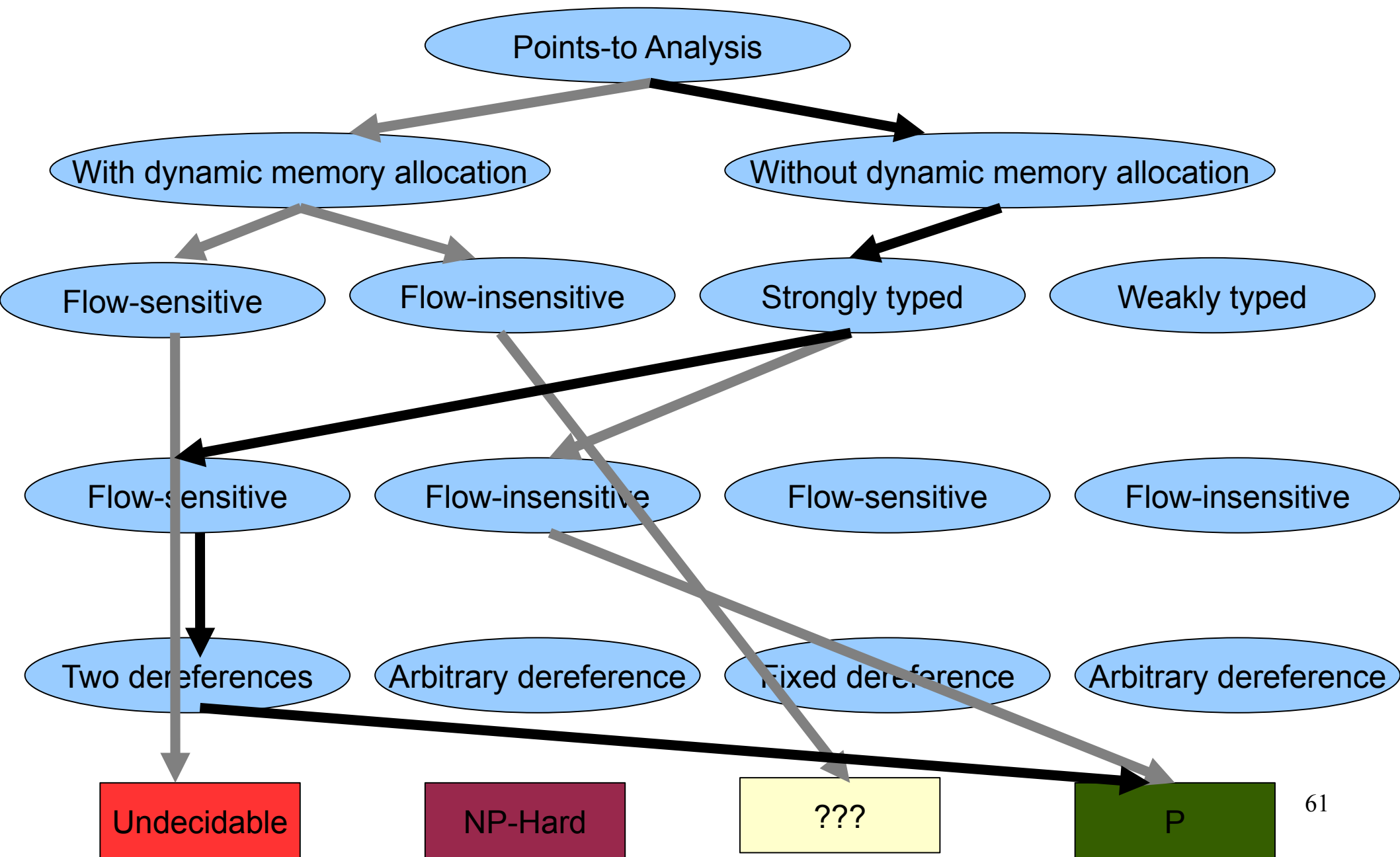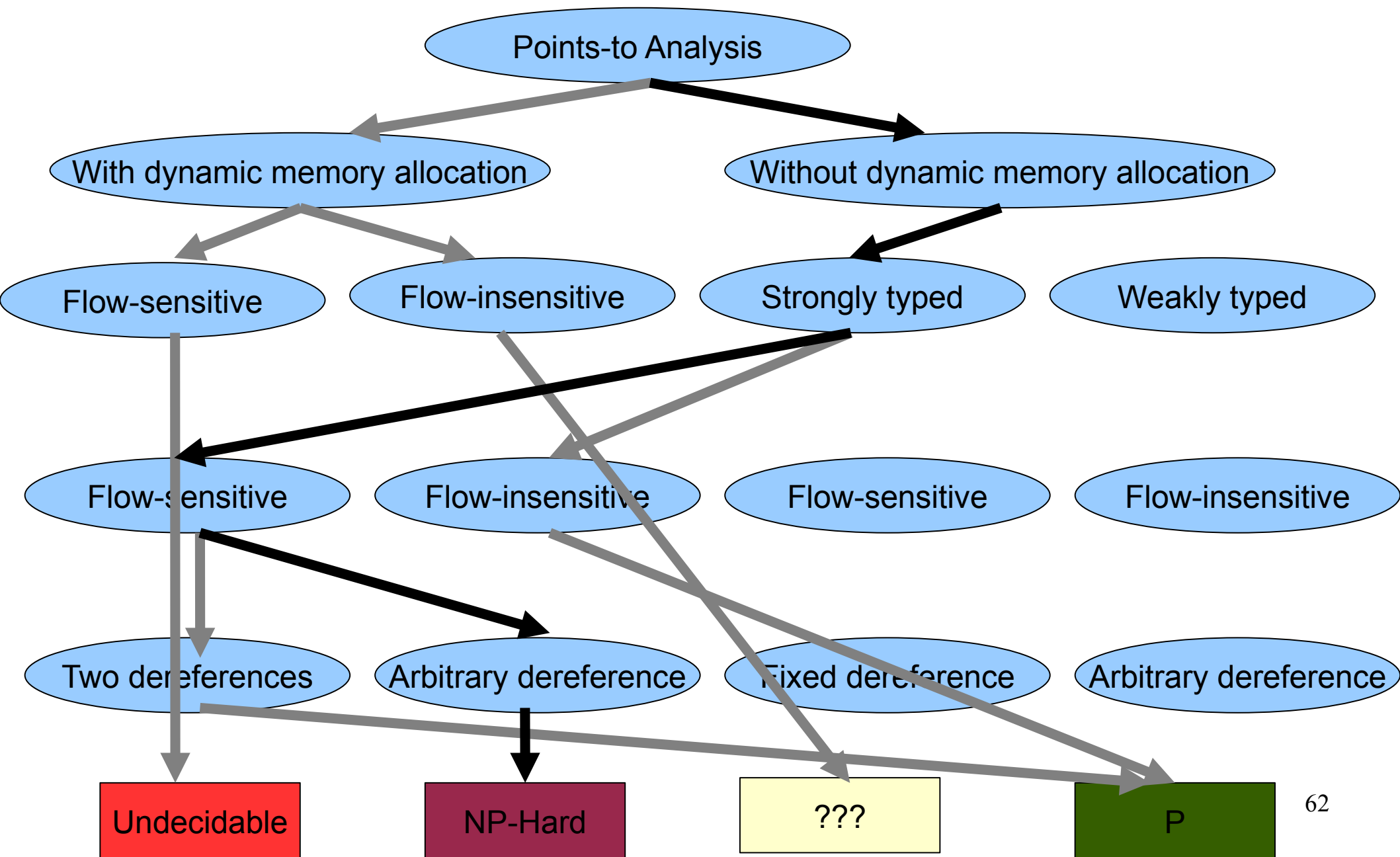
Complexity of Points-to Analysis

Complexity of Points-to Analysis

Points-to Analysis

With dynamic memory allocation | Without dynamic memory allocation

Flow-sensitive | Flow-insensitive | Strongly typed | Weakly typed

Flow-sensitive | Flow-insensitive | Flow-sensitive | Flow-insensitive

Two dereferences | Arbitrary dereference | Fixed dereference | Arbitrary dereference

Undecidable | NP-Hard | ??? | P

62

Complexity of Points-to Analysis

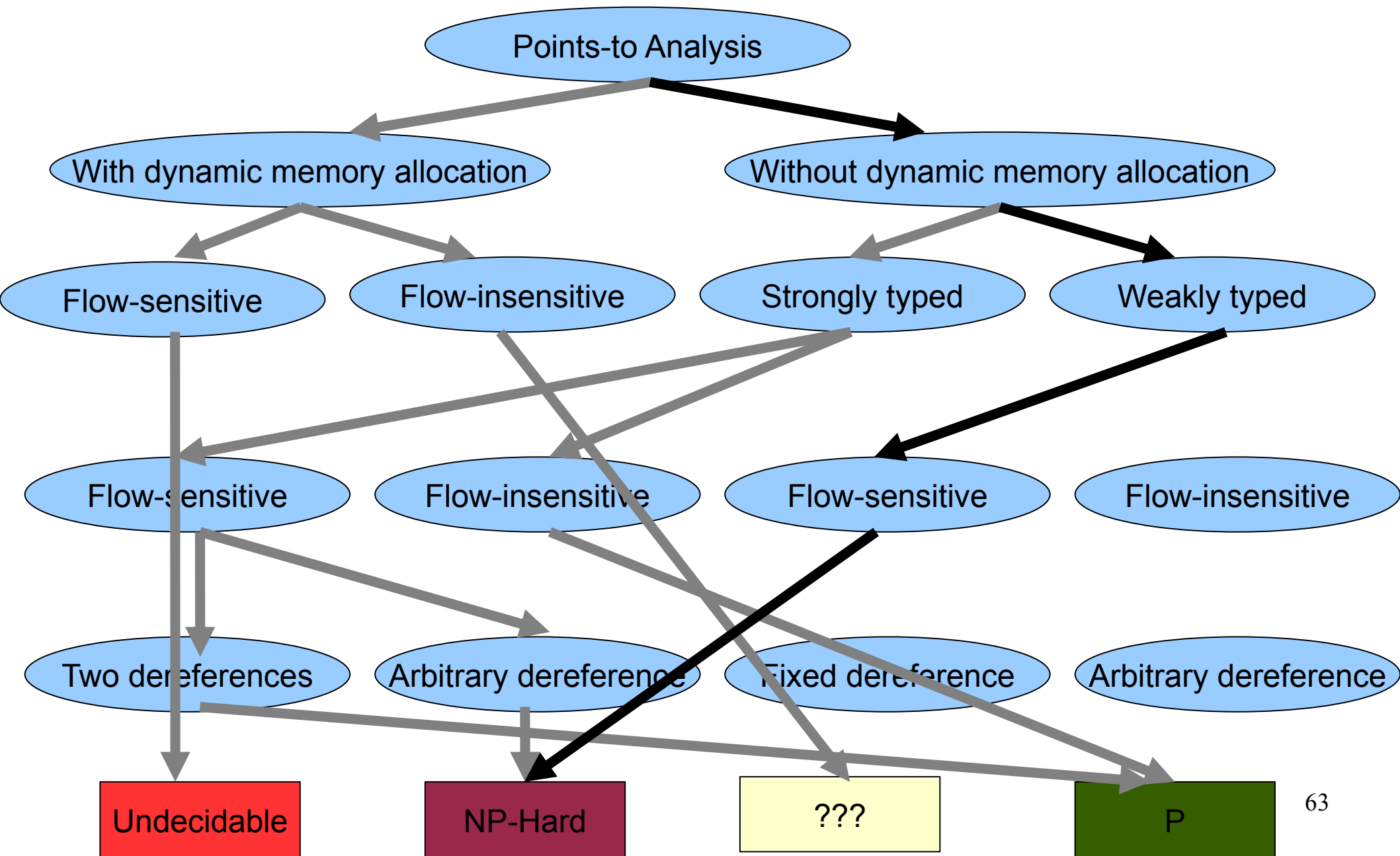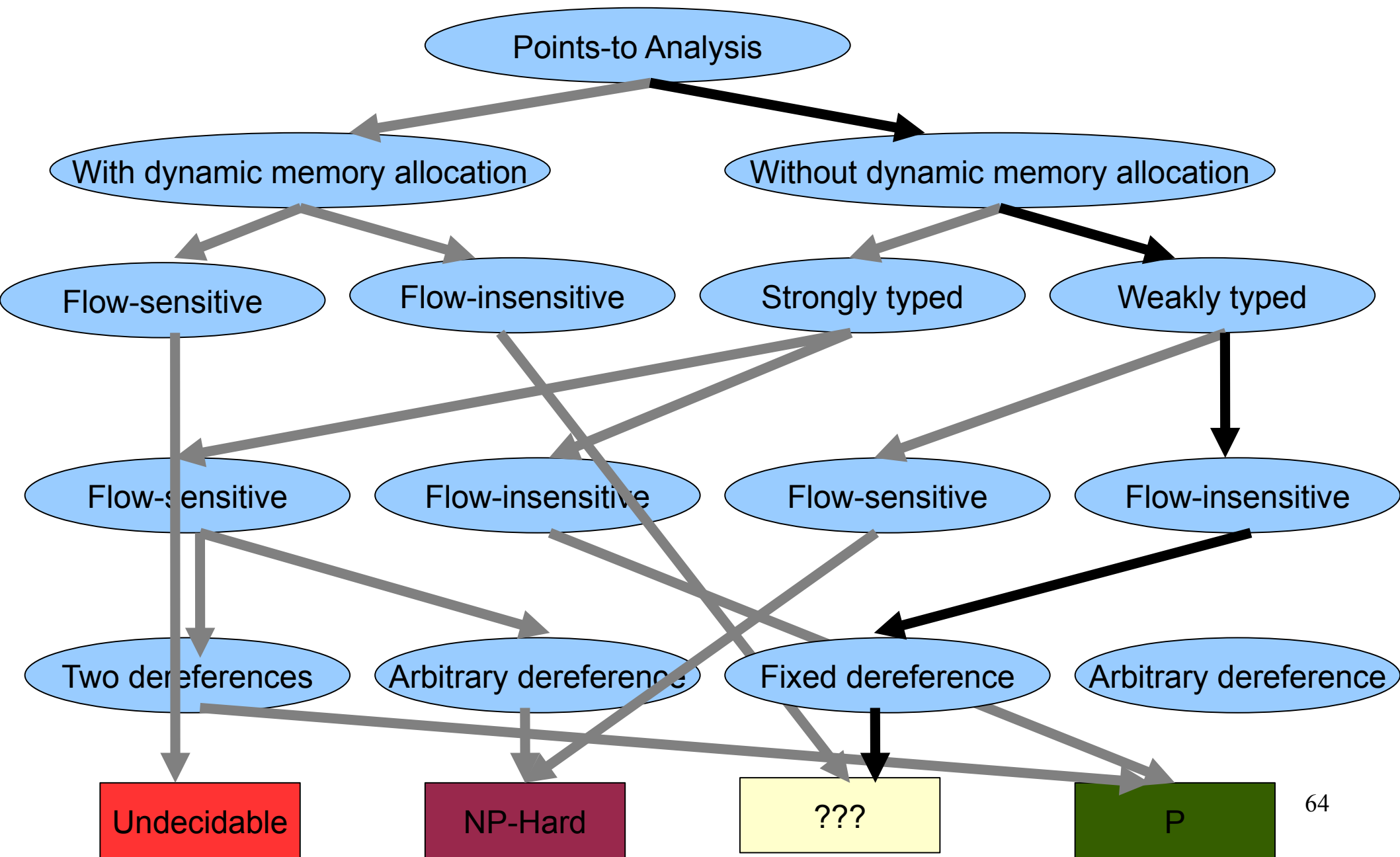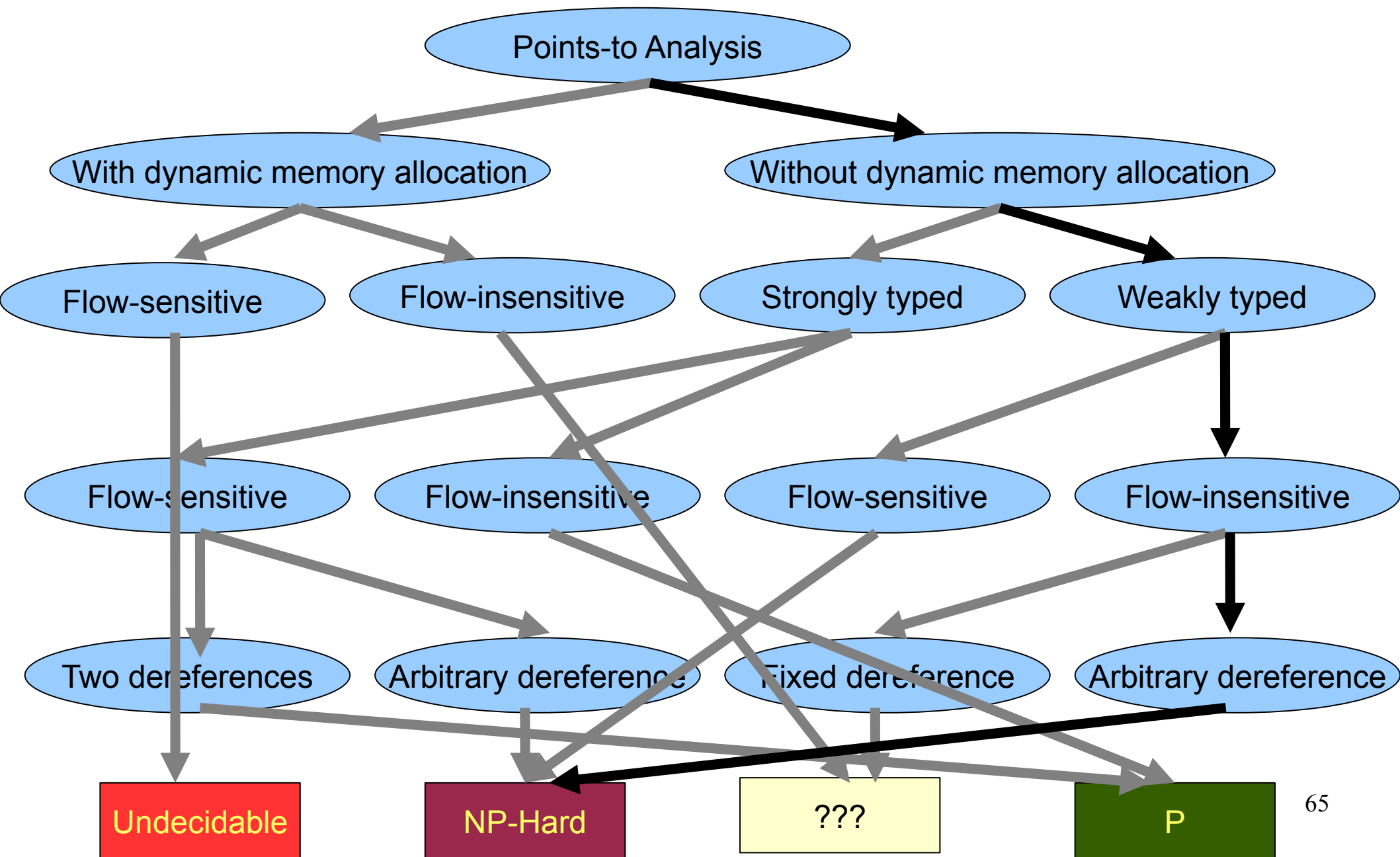# Complexity of Points-to Analysis

64

Complexity of Points-to Analysis

Related Work

Precision ← ➝

	Context-Sensitive	**Context-Insensitive**
Flow-Sensitive	Landi, Ryder 92 Choi et al. 93 Emami et al. 94 Reps et al. 95 Hind et al. 99 Kahlon 08	Zheng 98 Hardekopf, Lin 09
Flow-insensitive	Liang, Harrold 99 Whaley, Lam 04 Zhu, Calman 04 Lattner et al. 07	Andersen 94 Steensgaard 96 Shapiro, Horwitz 97 Fahndrich et al. 98 Das 00 Rountev, Chandra 00 Berndl et al. 03 Hardekopf, Lin 07 Pereira, Berlin 09 Mendez-Lojo 10
Surveys	Hind, Pioli 00 Qiang, Wu 06	

Precision