# An Assessment of Type-3 Clones as Detected by State-of-the-Art Tools

Rebecca Tiarks, Rainer Koschke and Raimar Falke

SCAM'09

9th IEEE International Working Conference

on Source Code Analysis and Manipulation

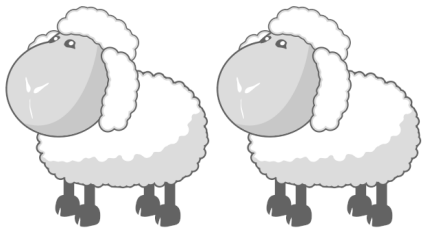
September 20th 2009

# Clone Types

I

```
i = length (l);
if  ( i < 3 ) { ... }
```

```
i = length (l);
if  ( i < 3 ) { ... }
```

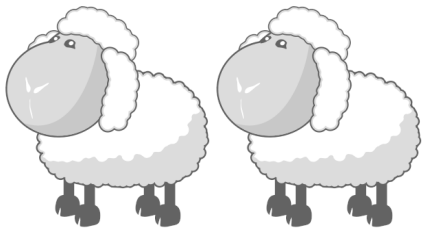


# Clone Types

I

```
i = length (l);
if  ( i < 3 ) { ... }
```

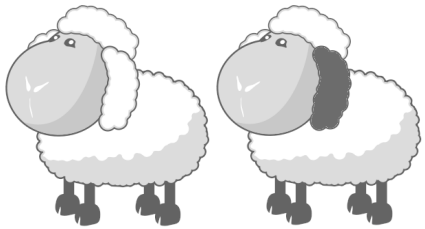
```
i = length (l);
if  ( i < 3 ) { ... }
```



II

```
i = length (l);
if  ( i < 3 ) { ... }
```

```
k = length (j);
if  ( k < x) { ... }
```

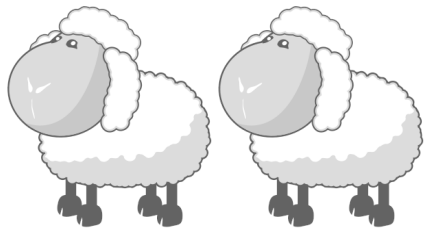


# Clone Types

I

```
i = length (l);
if  ( i < 3 ) { ... }
```

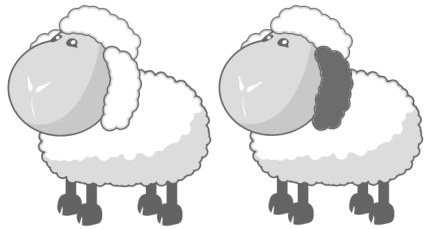
```
i = length (l);
if  ( i < 3 ) { ... }
```



II

```
i = length (l);
if  ( i < 3 ) { ... }
```

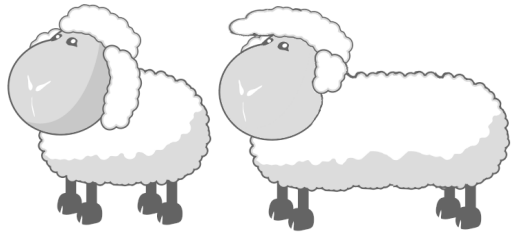
```
k = length (j);
if  ( k < x) { ... }
```



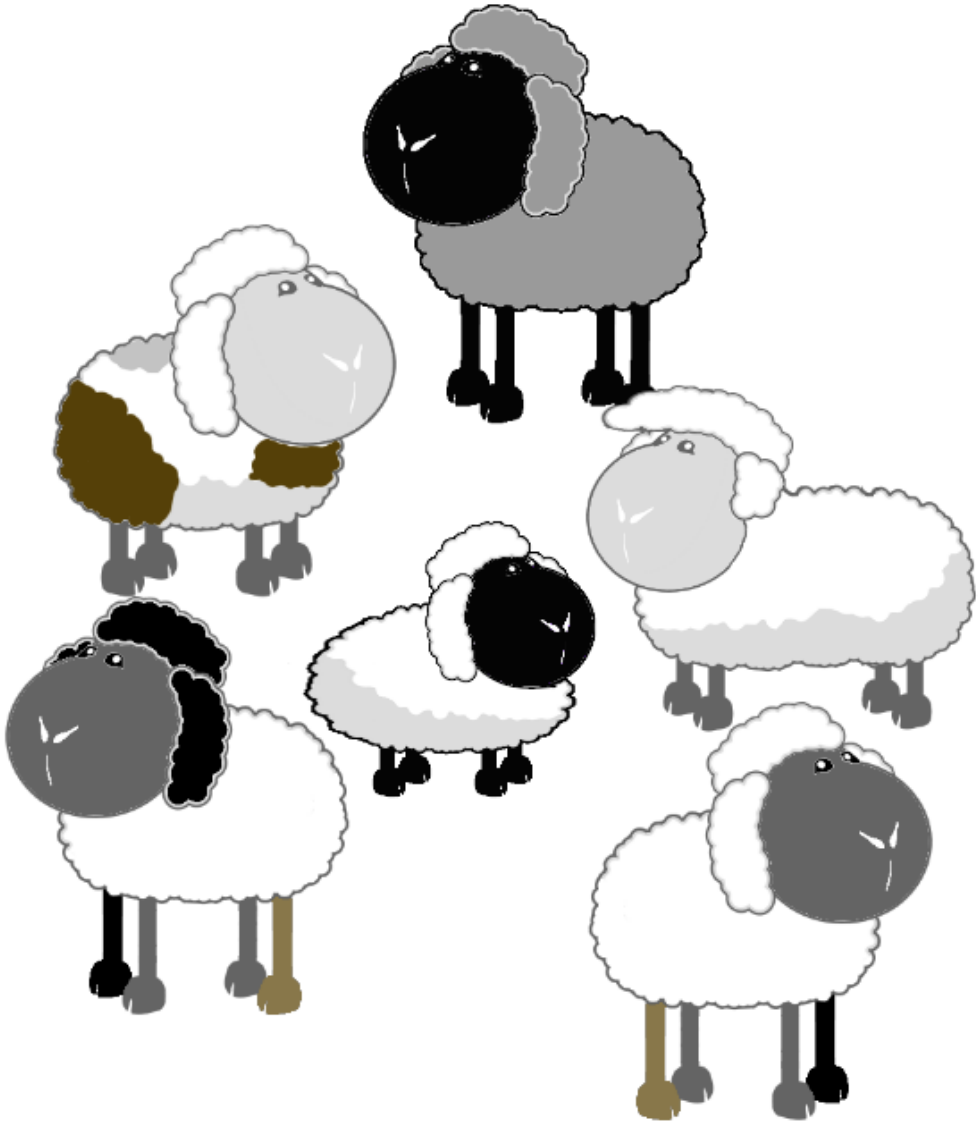
III

```
i = length (l);
if  ( i < 3 ) { ... }
```

```
i = length (l);
f ();
if  ( i < 3 ) { ... }
```



# Problem



=

?

# Research Questions

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics

Q1: How can we classify detected type-3 clones in terms of syntactic differences? And how often do they occur?

# Research Questions

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics

Q1: How can we classify detected type-3 clones in terms of syntactic differences? And how often do they occur?

Q2: What common abstractions can be assigned to the clones?

# Research Questions



Q1: How can we classify detected type-3 clones in terms of syntactic differences? And how often do they occur?

Q2: What common abstractions can be assigned to the clones?

Q3: Are there any code characteristics that indicate that a tool-suggested clone is a real type-3 clone from a human's perspective?

# Systems

| System | Language | KLOC |
|---|---|---|
| wget | C | 16 |
| Javadoc | Java | 19 |
| bison | C | 19 |
| Ant | Java | 35 |
| snns | C | 115 |
| JDTCore | Java | 148 |
| Swing | Java | 204 |
| postgresql | C | 235 |

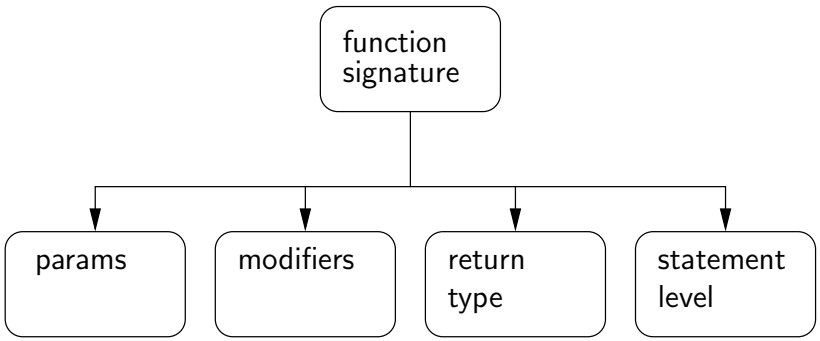
Table: Systems

# Data

- five tools based on different detection algorithms
- all configured with default settings $\longrightarrow$ comparability
- from Bellon Benchmark:
  - CLAN Merlo
  - Duplix Krinke
- from our tools:
  - clast
  - cscope
  - ccdiml
- **total amount of type-3 clones:** 391 628 **oracled:** 751

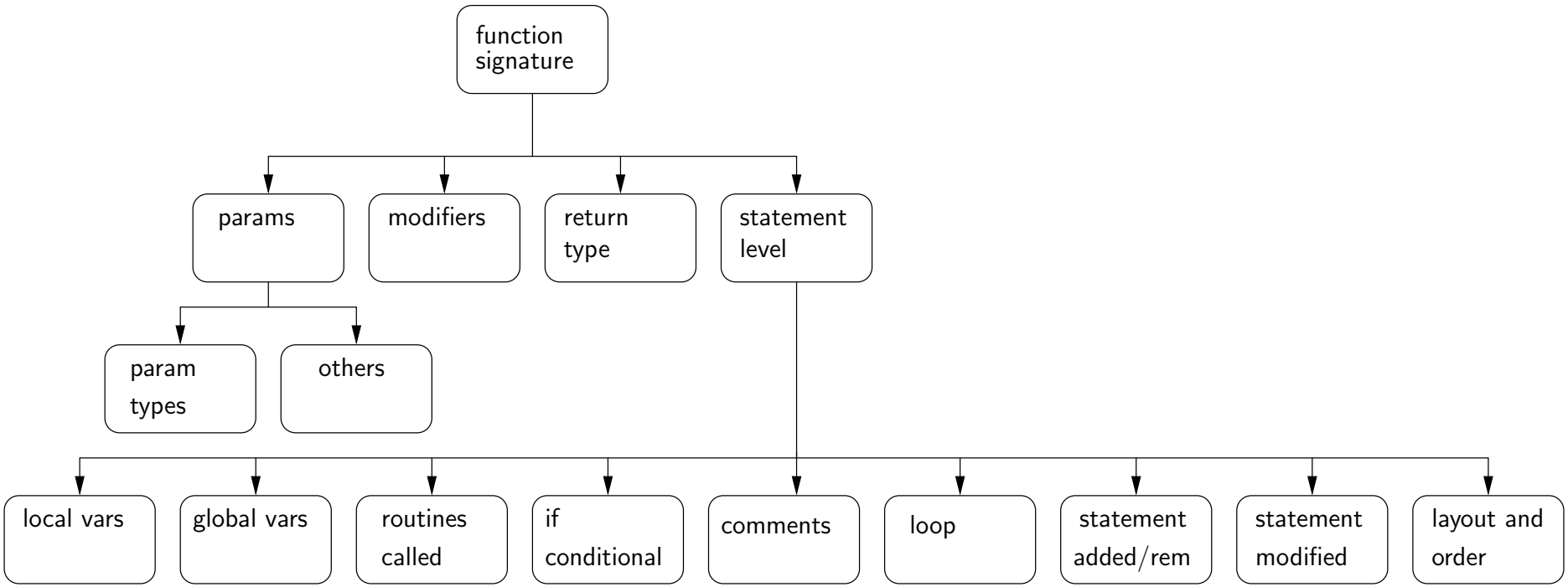
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



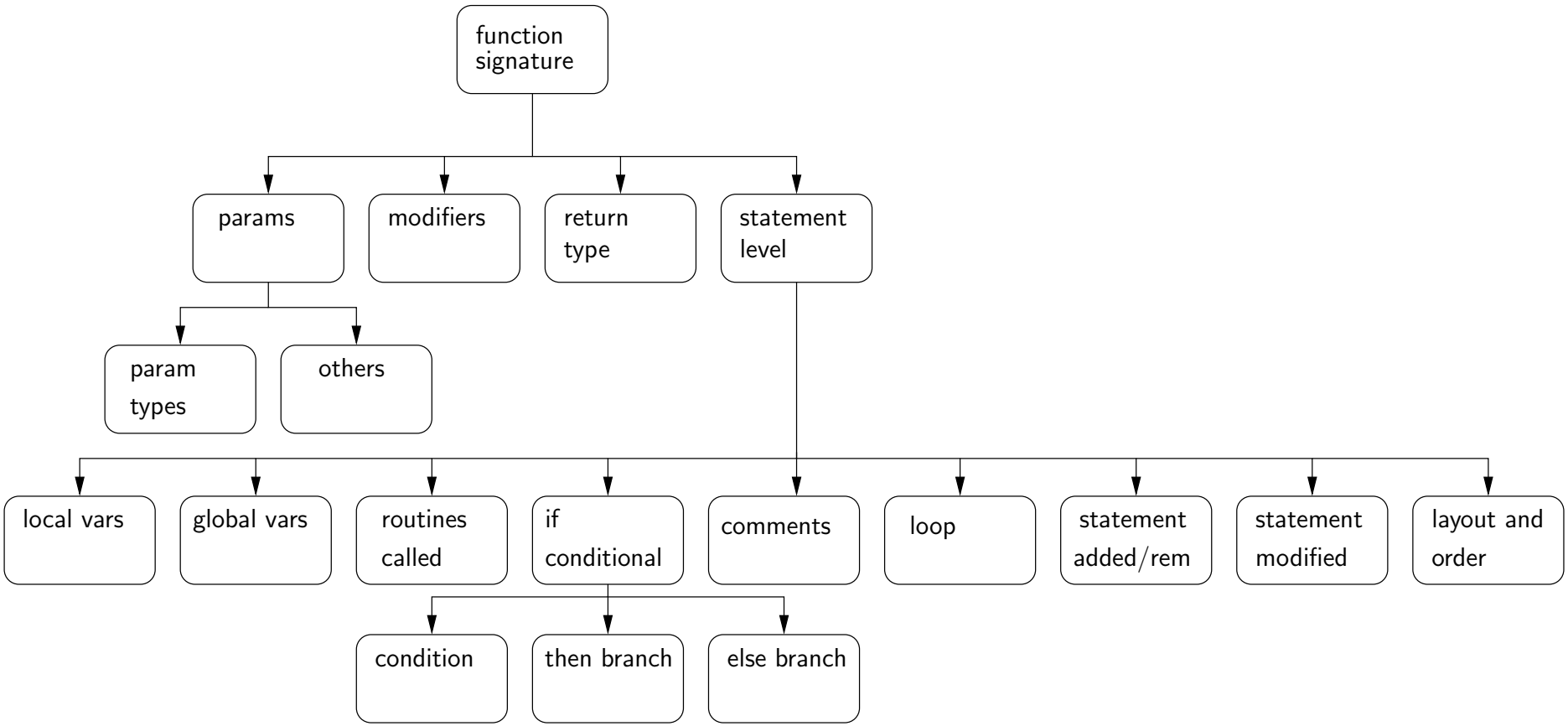
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



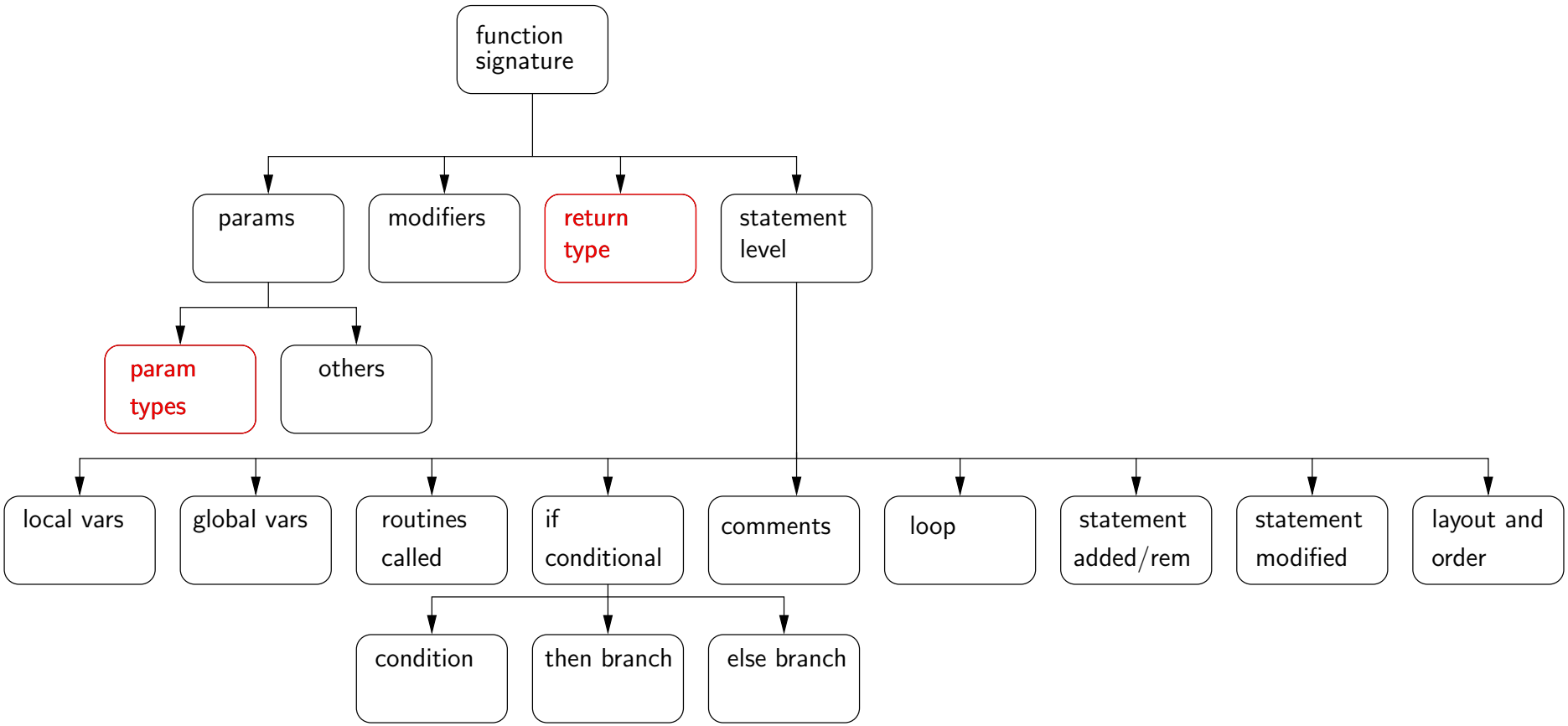
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



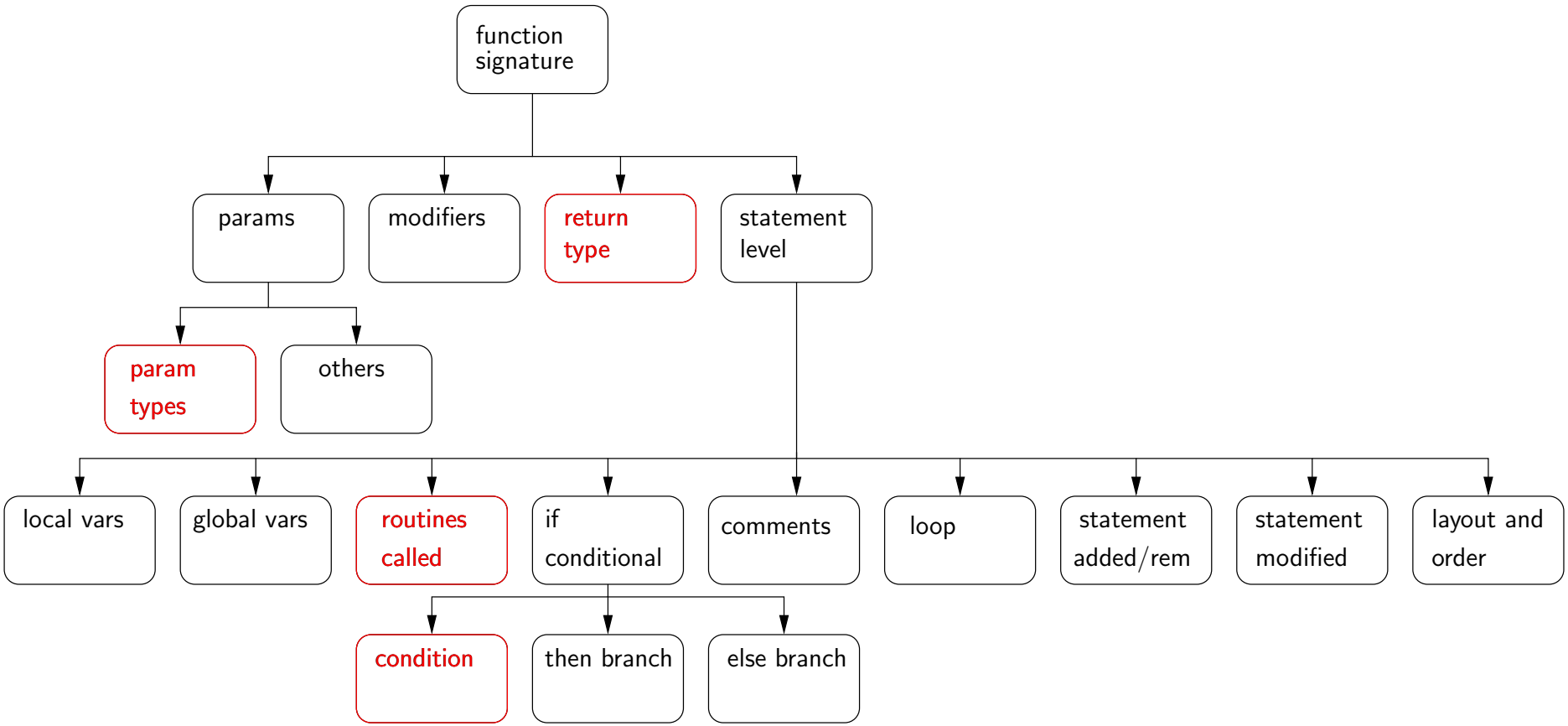
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



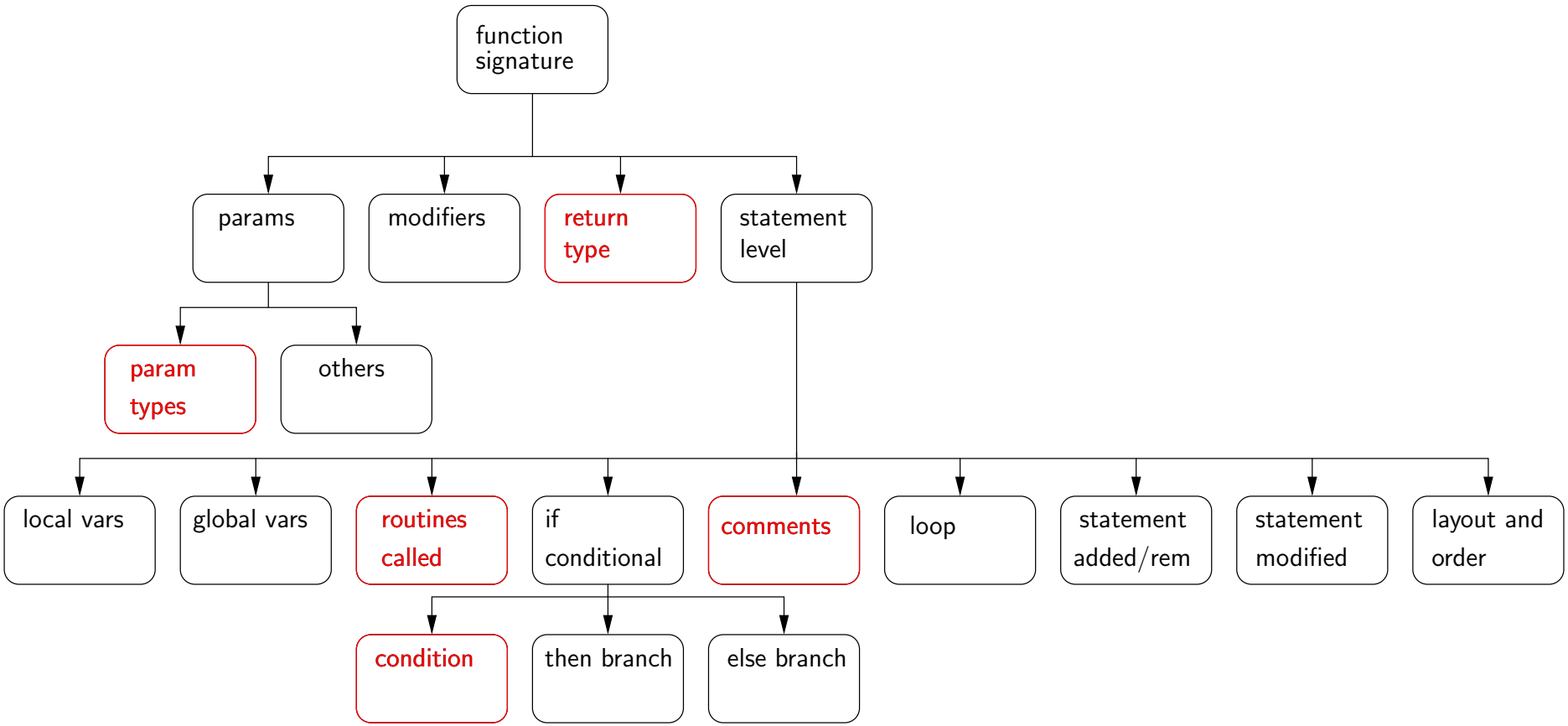
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



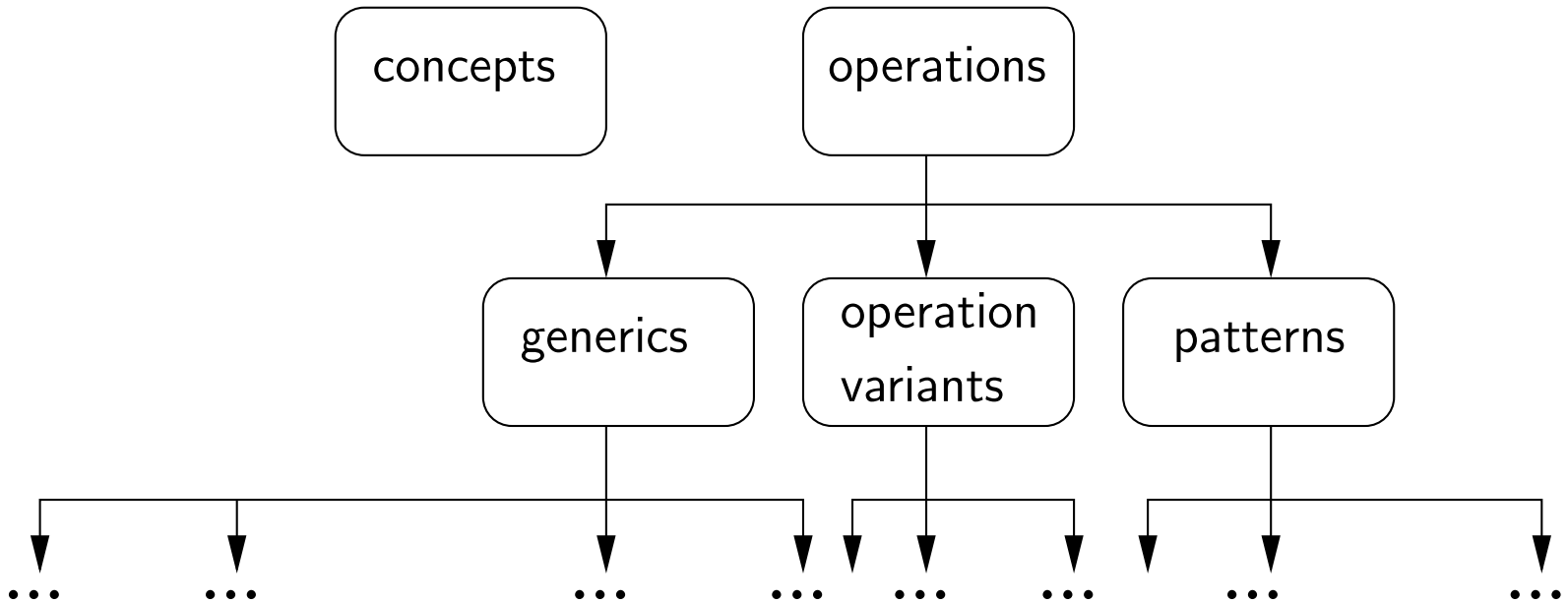
# Syntactic Classification

Q1:Syntactic Classification Q2:Common Abstractions Q3:Code Characteristics



# Common Abstractions





# Code Characteristics



- metrics used:
  - two text similarity measures
  - eight metrics per token types
  - metrics based on fragments length
- e.g., token per line, character per token, occurrence of token type, . . .

# Conclusion

- need to improve detection of type-3 clones
- tools using more abstraction should be able to detect and classify clones with minor syntactic differences
- text similarity can be used to filter false positives
- tools could use data mining techniques to improve detection

*"Abstraction based type-3 clones can help to improve the detection, maintenance, and removal of clones"*