Learning from 6,000 Projects

Lightweight Cross-Project Anomaly Detection

Natalie Gruska
Queen's University

Andrzej Wasylkowski
Saarland University

Andreas Zeller
Saarland University

Defect in Conspire 0.20

```
static int dcc_listen_init (…) {
    dcc->sok = socket(…);
    if (…) {
        while (…) {
            … = bind (dcc->sok, …);
        }
        /* with a small port range, reUseAddr is needed */
        setsockopt (dcc->sok, …, SO_REUSEADDR, …);
    }
    listen (dcc->sok, …);
}
```

should be called *before bind()*

Defect in Conspire 0.20

```
static int dcc_listen_init (…) {
    dcc->sok = socket(…);
    if (…) {
        while (…) {
            … = bind (dcc->sok, …);
        }
        /* with a small port range, reUseAddr is needed */
        setsockopt (dcc->sok, …, SO_REUSEADDR, …);
    }
    listen (dcc->sok, …);
}
```

should be called *before bind()*

```
bind < listen
setsockopt < listen
setsockopt < bind
```

Missing!

Anomaly Detection

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

Anomaly Detection

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

Anomaly Detection

```
bind < listen
setsockopt < listen
```

?

We need more examples!

Cross-project Anomaly Detection

Knowledge base

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

Program

```
bind < listen
setsockopt < listen
```

Cross-project Anomaly Detection

Knowledge base

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

```
bind < listen
setsockopt < listen
setsockopt < bind
```

Program

```
bind < listen
setsockopt < listen
```

Cross-project Anomaly Detection

Knowledge base

Program

bind < listen
setsockopt < listen

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen
setsockopt < bind

bind < listen
setsockopt < listen

Cross-project Anomaly Detection

Knowledge base

Program

- Goal: Learn from thousands of other projects

Lightweight Parser: Focus Languages

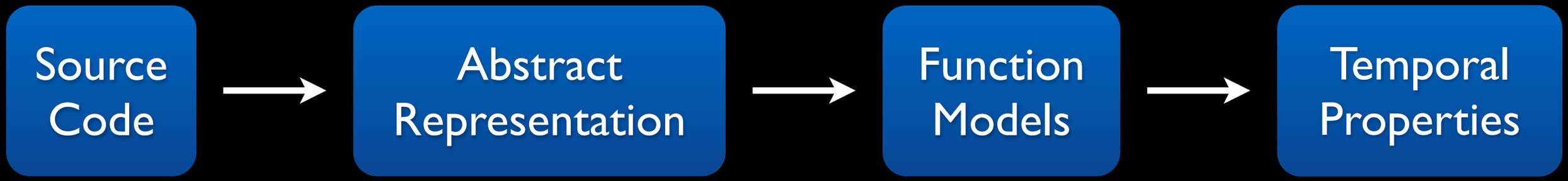
Java C++ C PHP Javascript

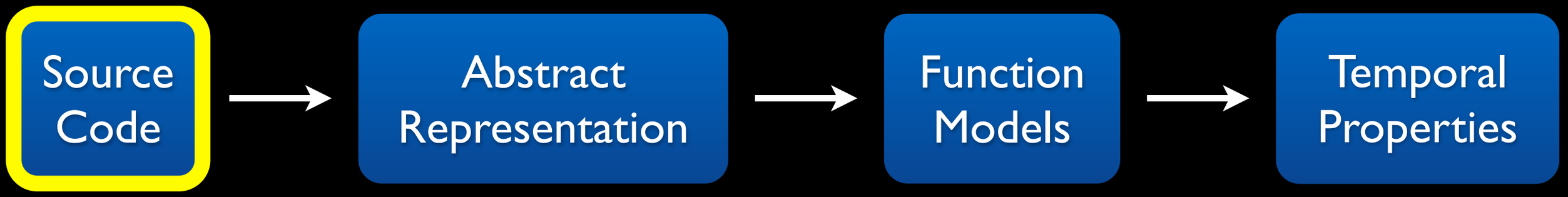
similar syntax:

```
{...}          ;          foo()
```

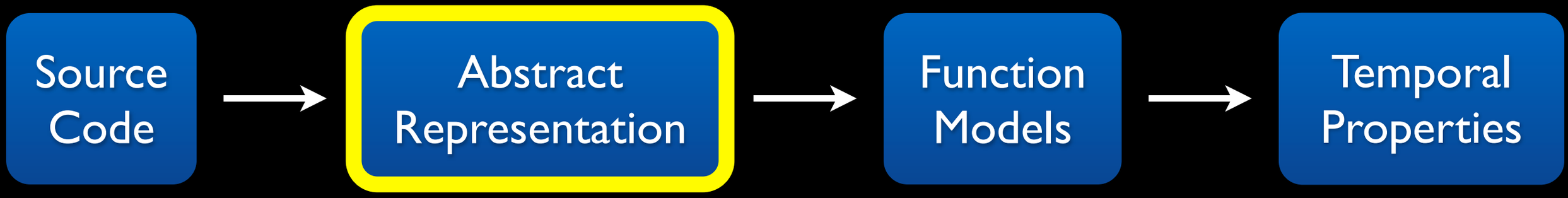
similar keywords:

```
while    if    switch    return
```

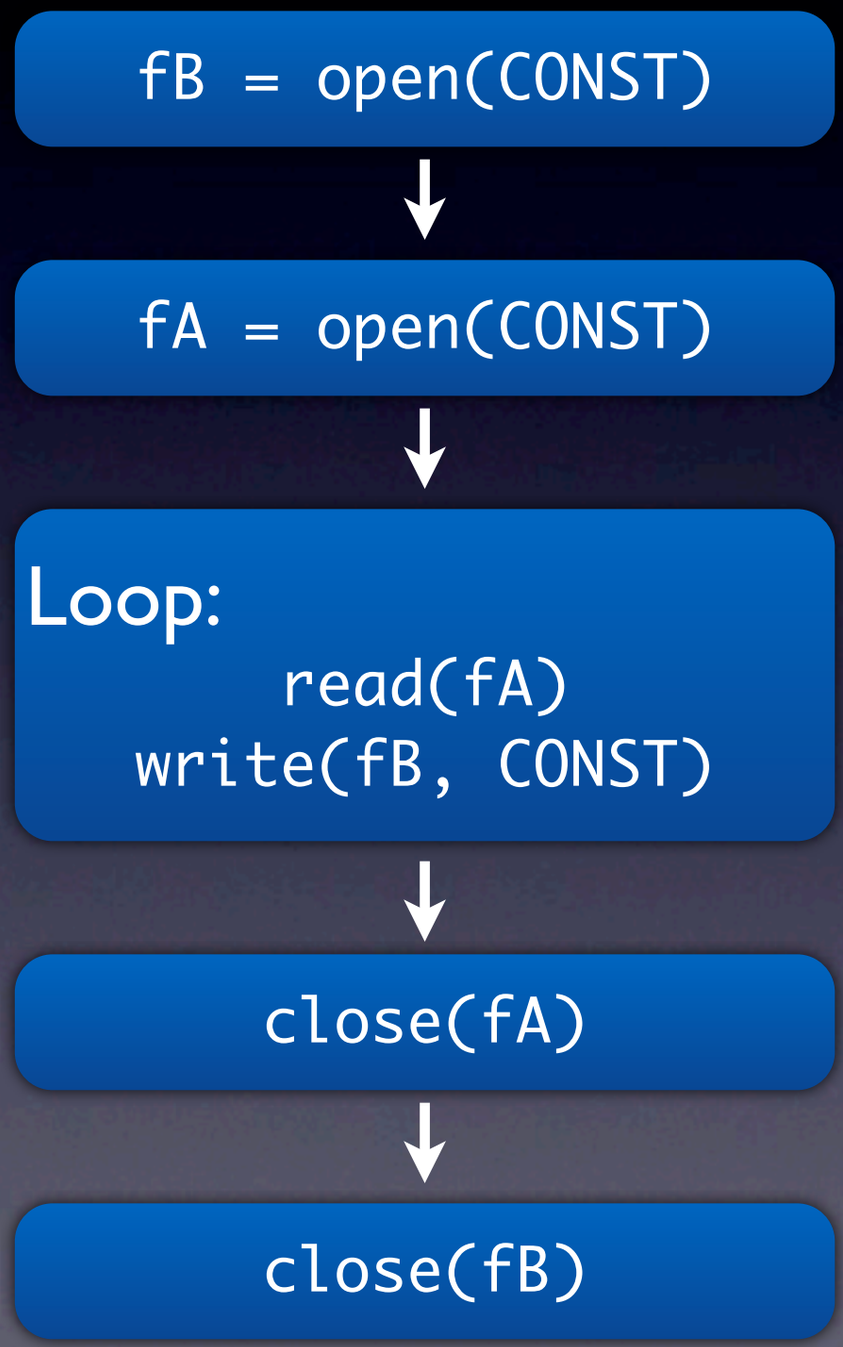

Source Code
Abstract Representation
Function Models
Temporal Properties

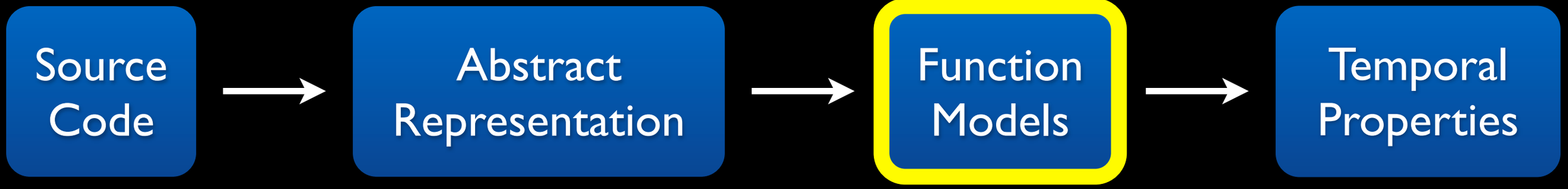



```
void foo () {
  int fA;
  int fB = open(“newFile”);
  fA = open(“myFile”);
  while(j > 3){
    read(fA);
    write(fB, “Hello”);
  }
  close(fA);
  close(fB);
}
```




```
void foo () {
  int fA;
  int fB = open(“newFile”);
  fA = open(“myFile”);
  while(j > 3){
    read(fA);
    write(fB, “Hello”);
  }
  close(fA);
  close(fB);
}
```



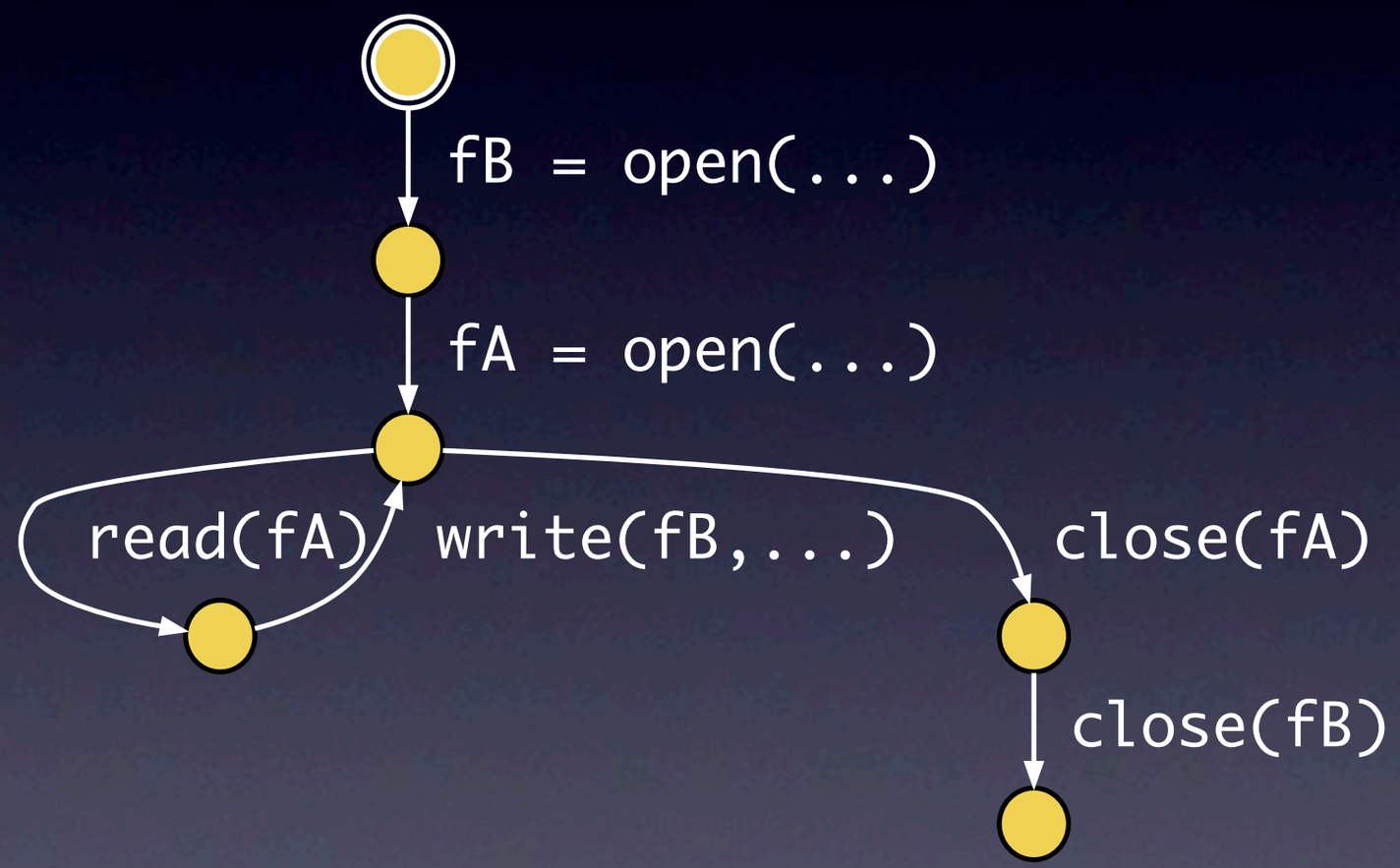



```
fB = open(CONST)
fA = open(CONST)
```

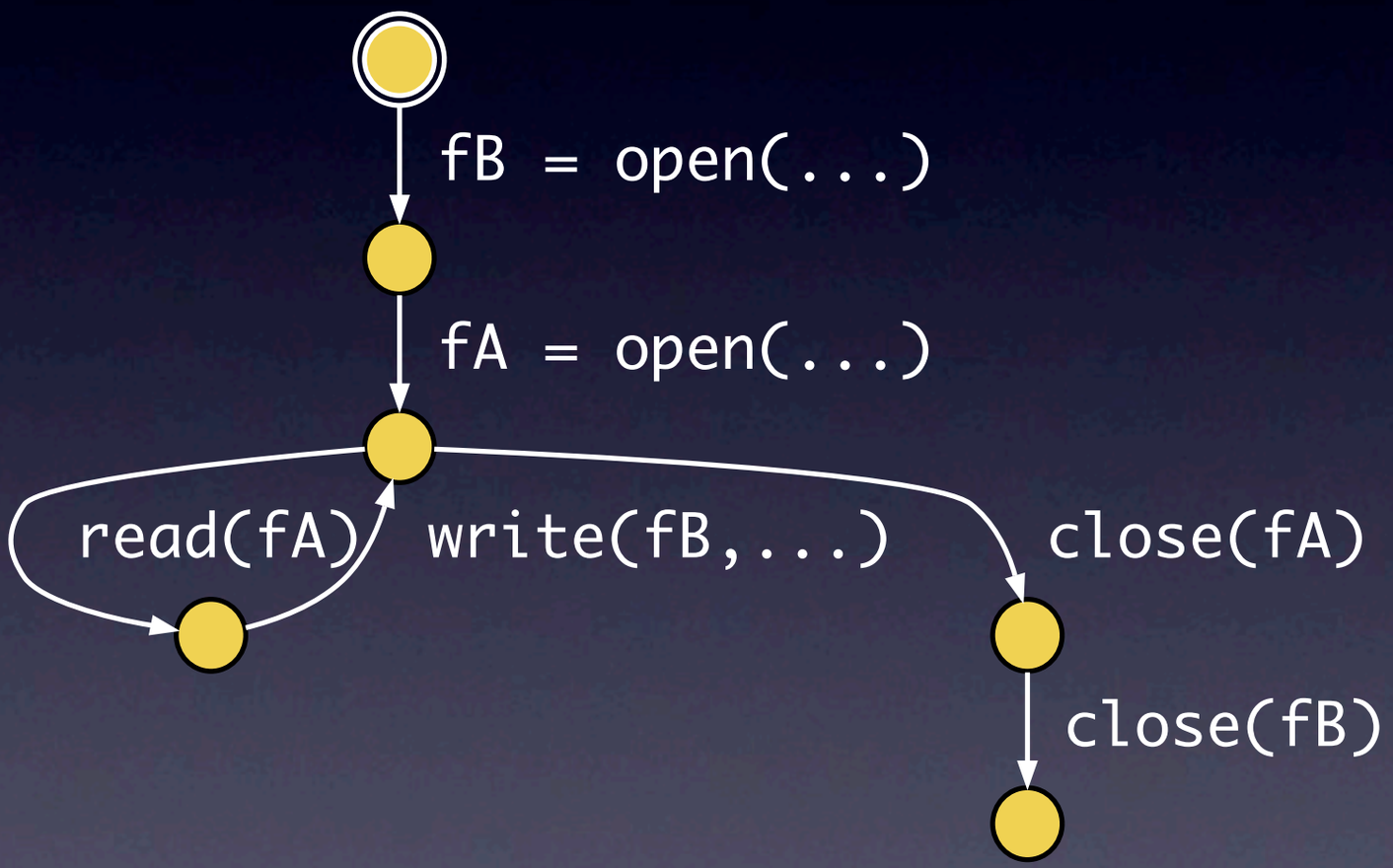
```
Loop:
    read(fA)
    write(fB, CONST)
close(fA)
close(fB)
```

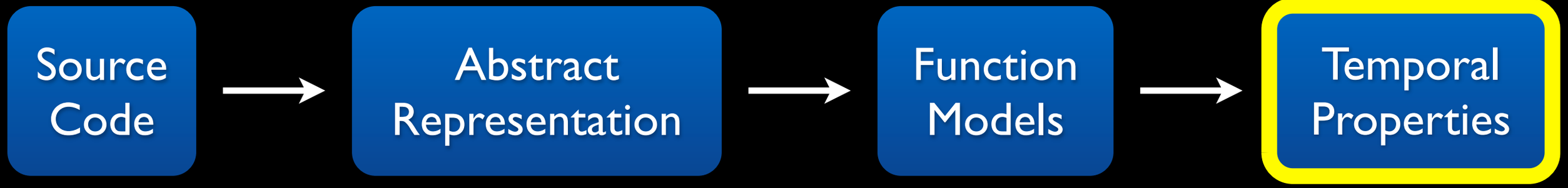
Source Code → Abstract Representation → **Function Models** → Temporal Properties

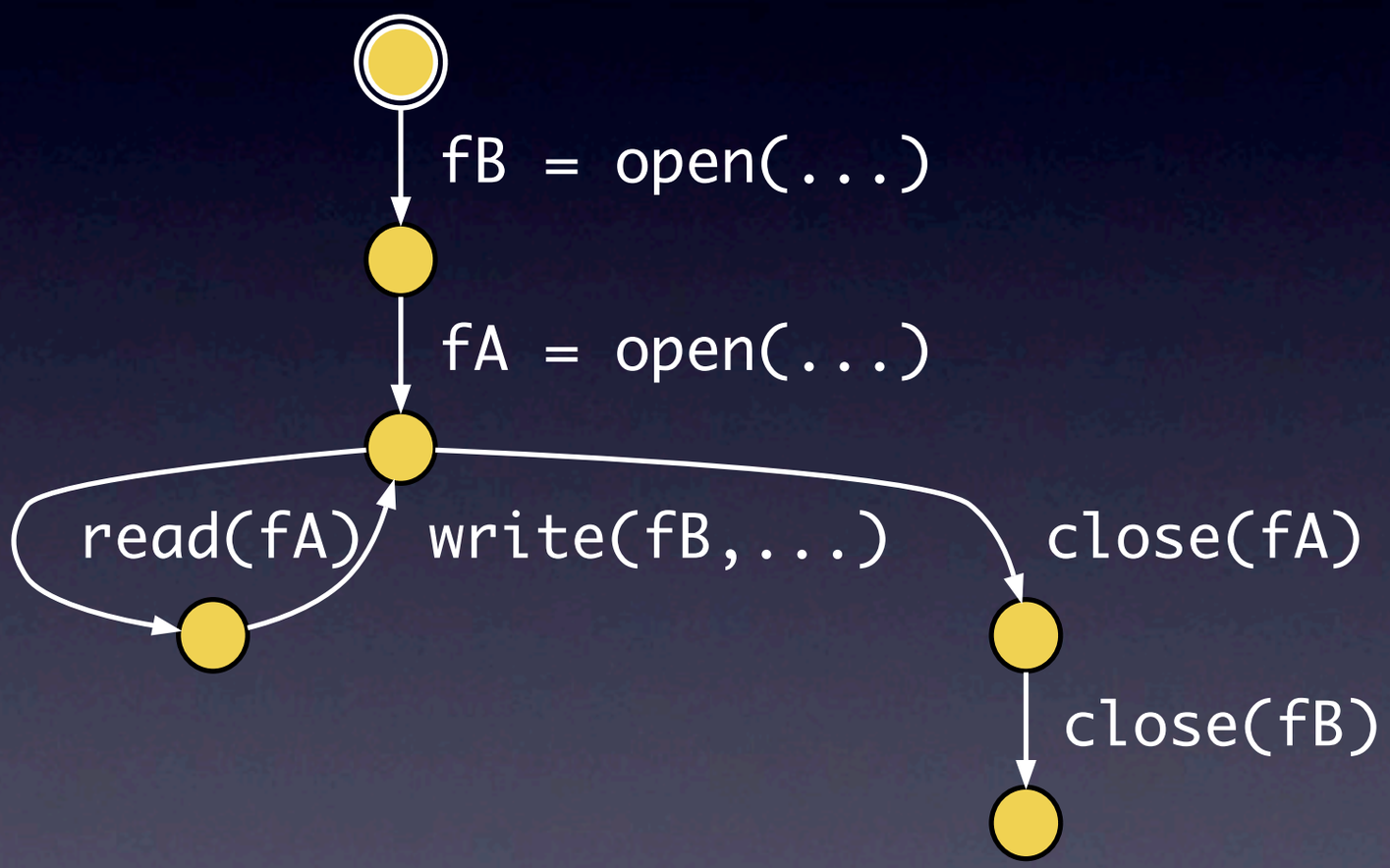
```
fB = open(CONST)
fA = open(CONST)
Loop:
    read(fA)
    write(fB, CONST)
close(fA)
close(fB)
```



Source Code → Abstract Representation → Function Models → Temporal Properties







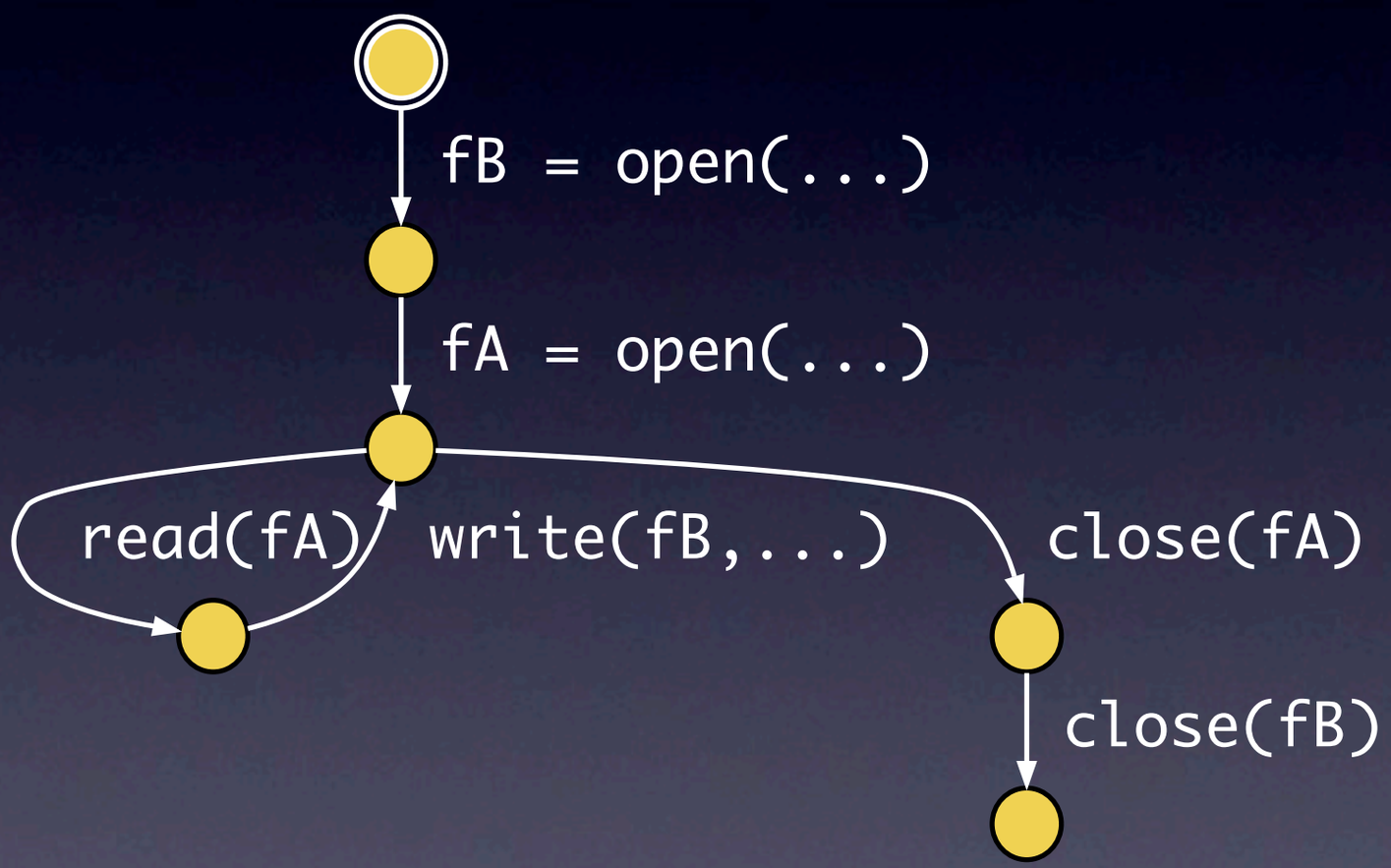
fB:

```
open < write
write < write
write < close
open < close
```

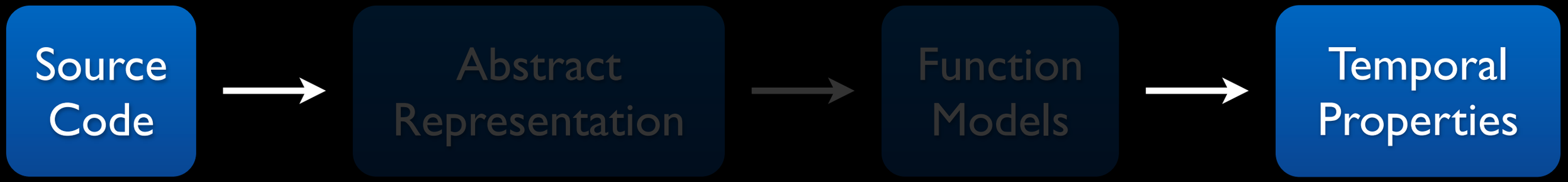
fA:

```
open < read
read < read
read < close
open < close
```

Source Code → Abstract Representation → Function Models → **Temporal Properties**



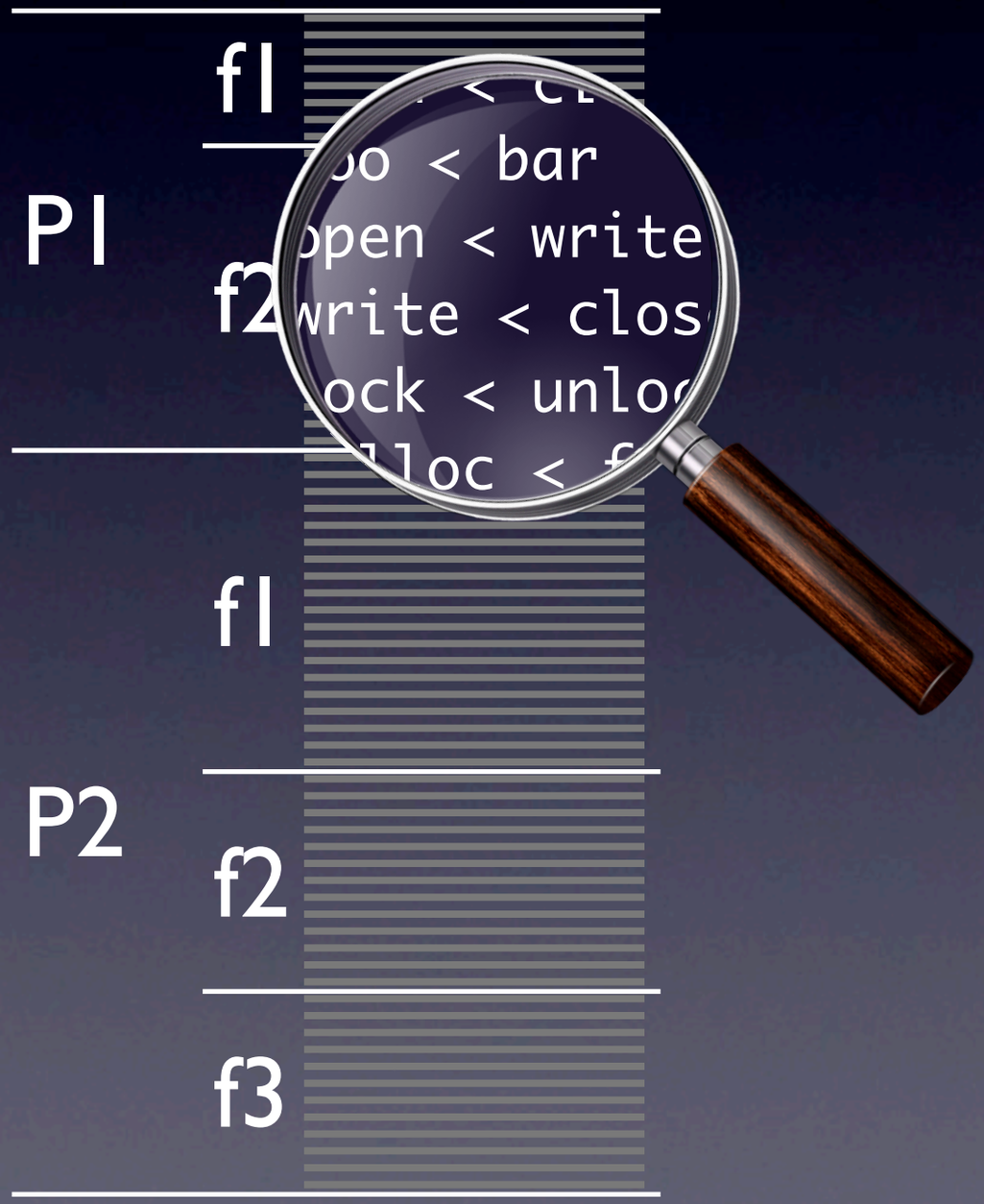

```
open < write
open < read
write < write
read < read
write < close
read < close
open < close
```

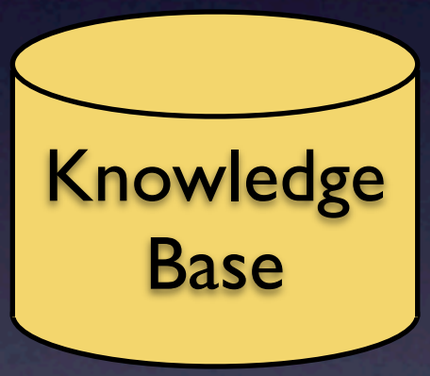



```
void foo () {
  int fA;
  int fB = open(“newFile”);
  fA = open(“myFile”);
  while(j > 3){
    read(fA);
    write(fB, “Hello”);
  }
  close(fA);
  close(fB);
}
```

```
open < write
open < read
write < write
read < read
write < close
read < close
open < close
```

Knowledge Base

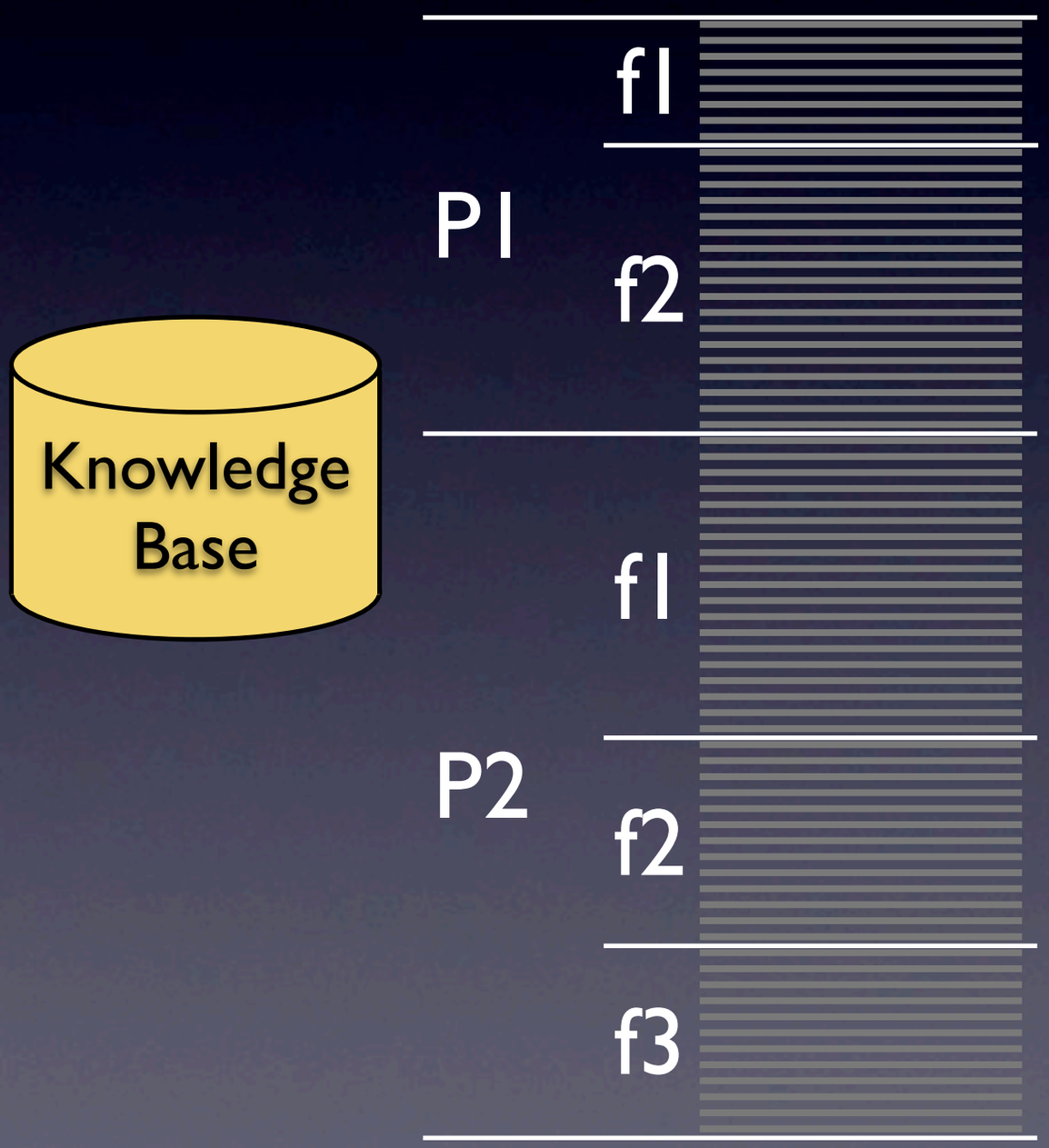




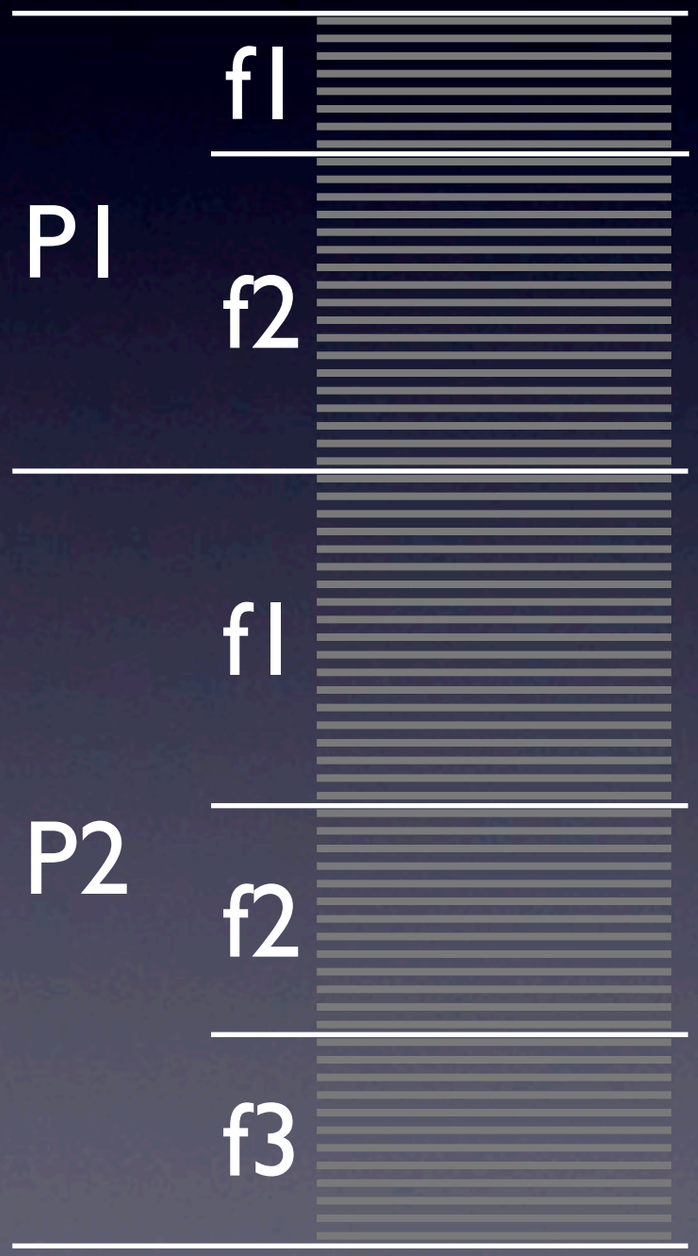
The Knowledge Base

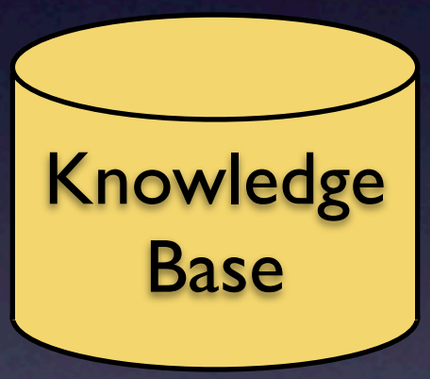
- Gentoo Linux distribution
- C projects
  - ≈6,000 projects
  - ≈200,000,000 lines of code
- ≈16,000,000 temporal properties
- Creation time: 18h (11s per project)

Using a Knowledge Base

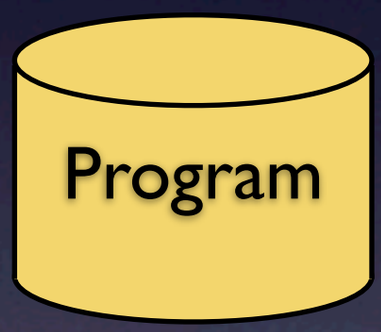


Using a Knowledge Base

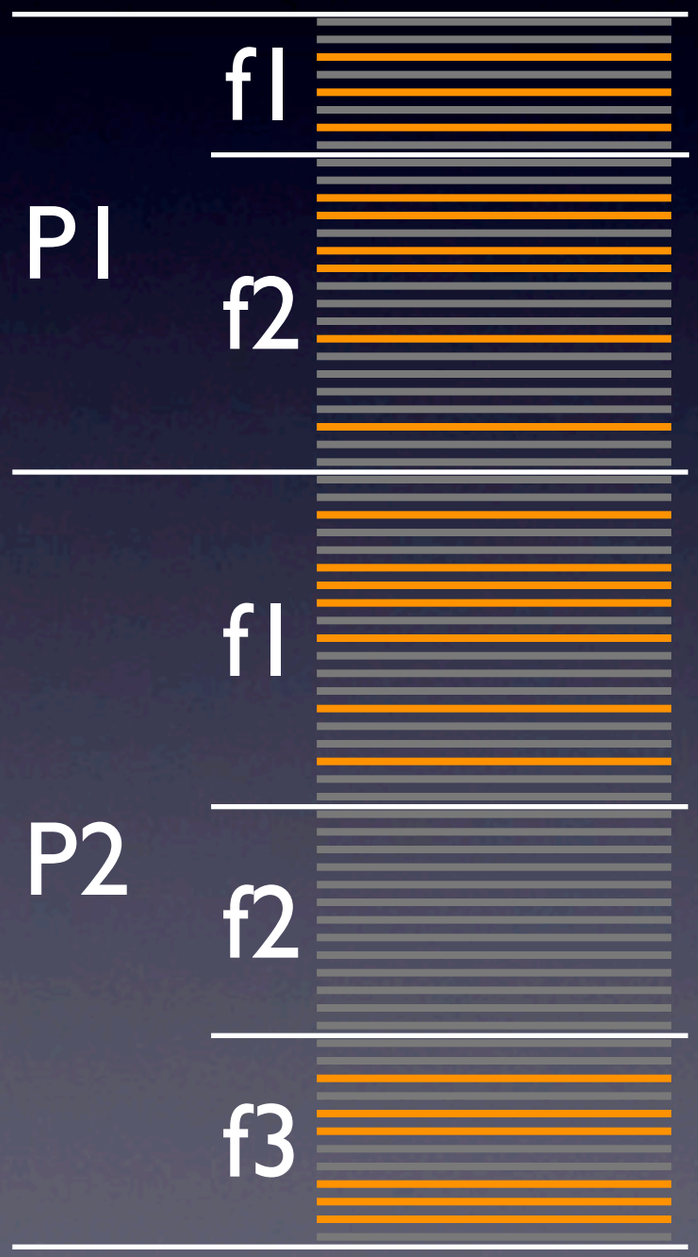


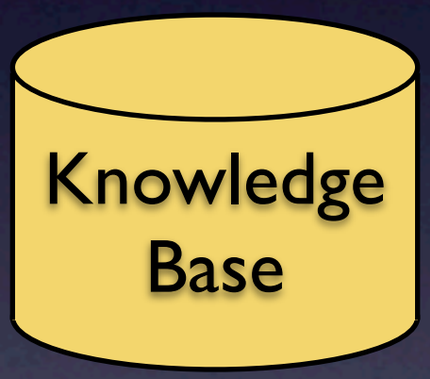



```
noths < inet_ntoa
memset < sizeof
bind < listen
setsockopt < listen
```

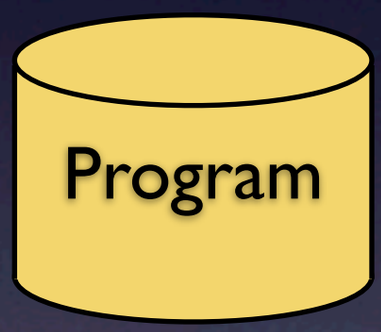


Using a Knowledge Base

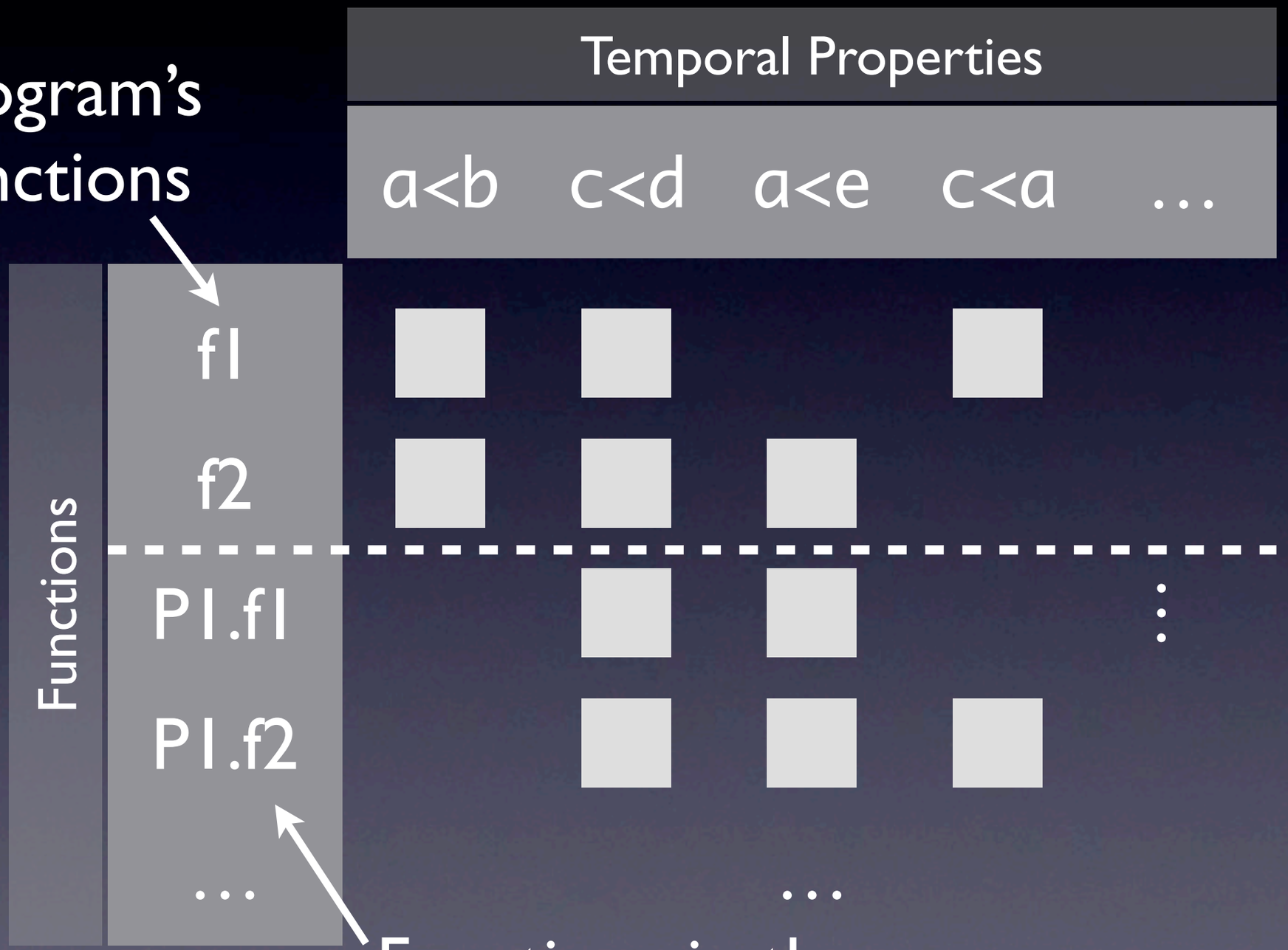




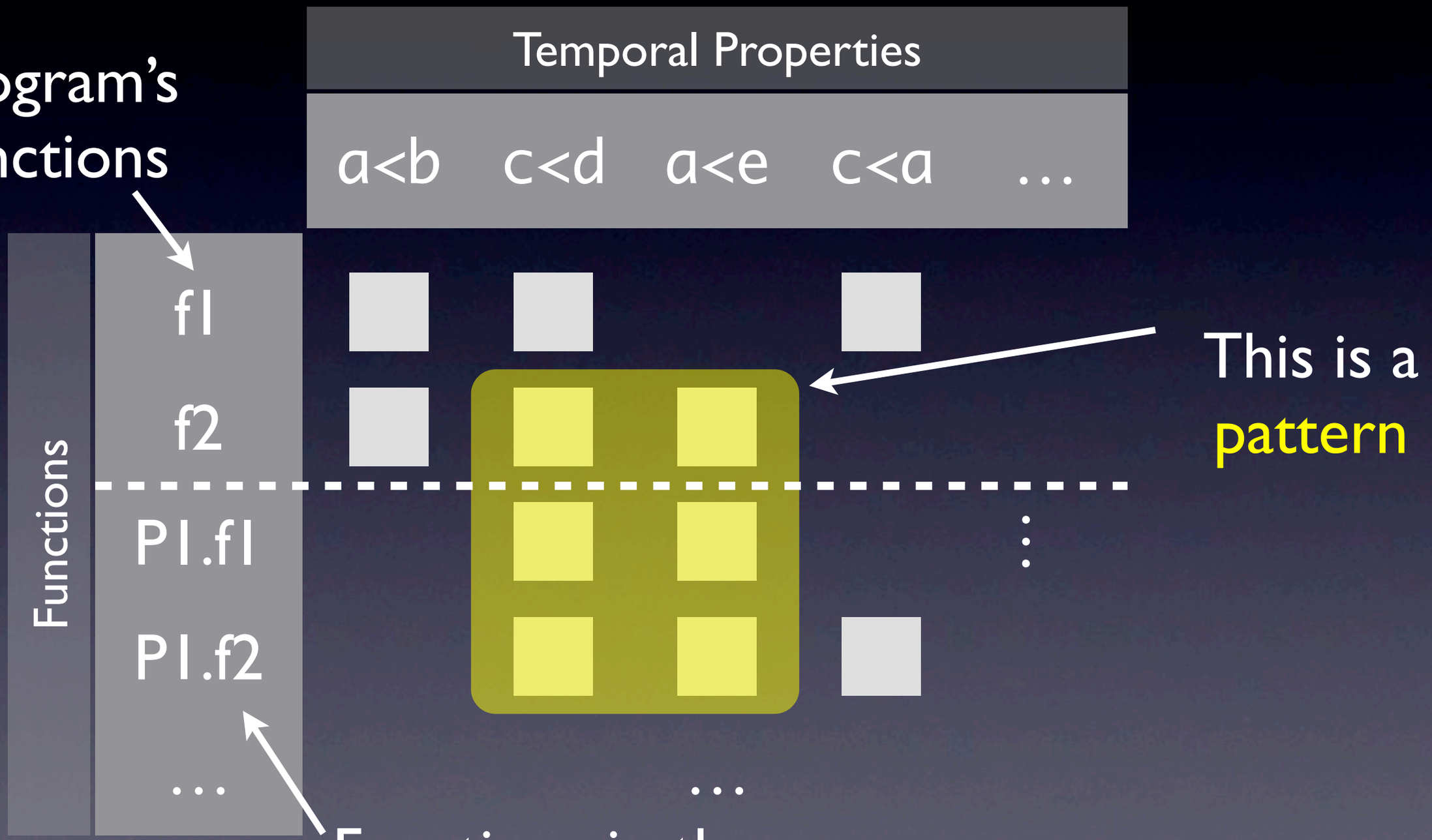

```
noths < inet_ntoa
memset < sizeof
bind < listen
setsockopt < listen
```



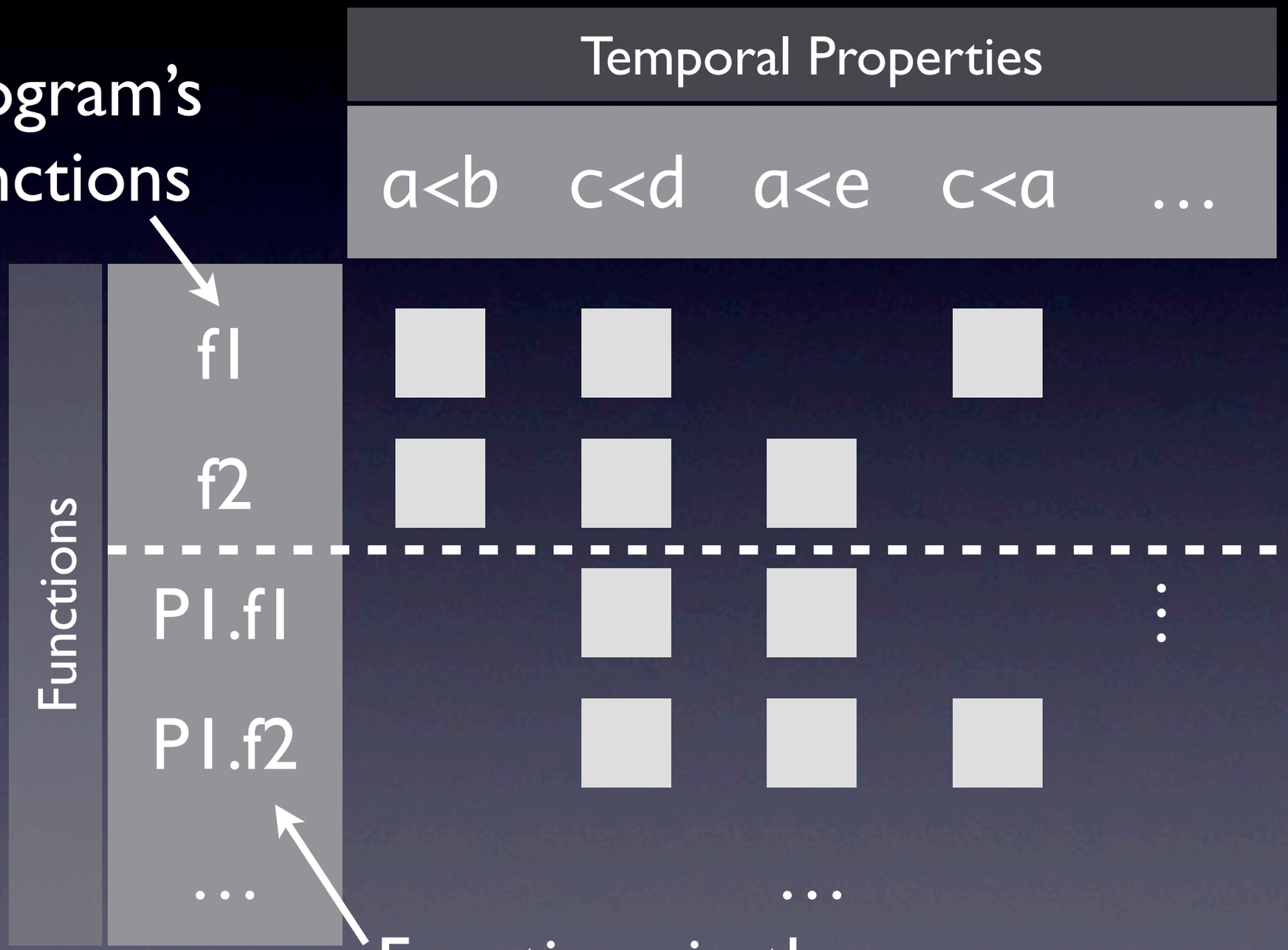
Finding Patterns



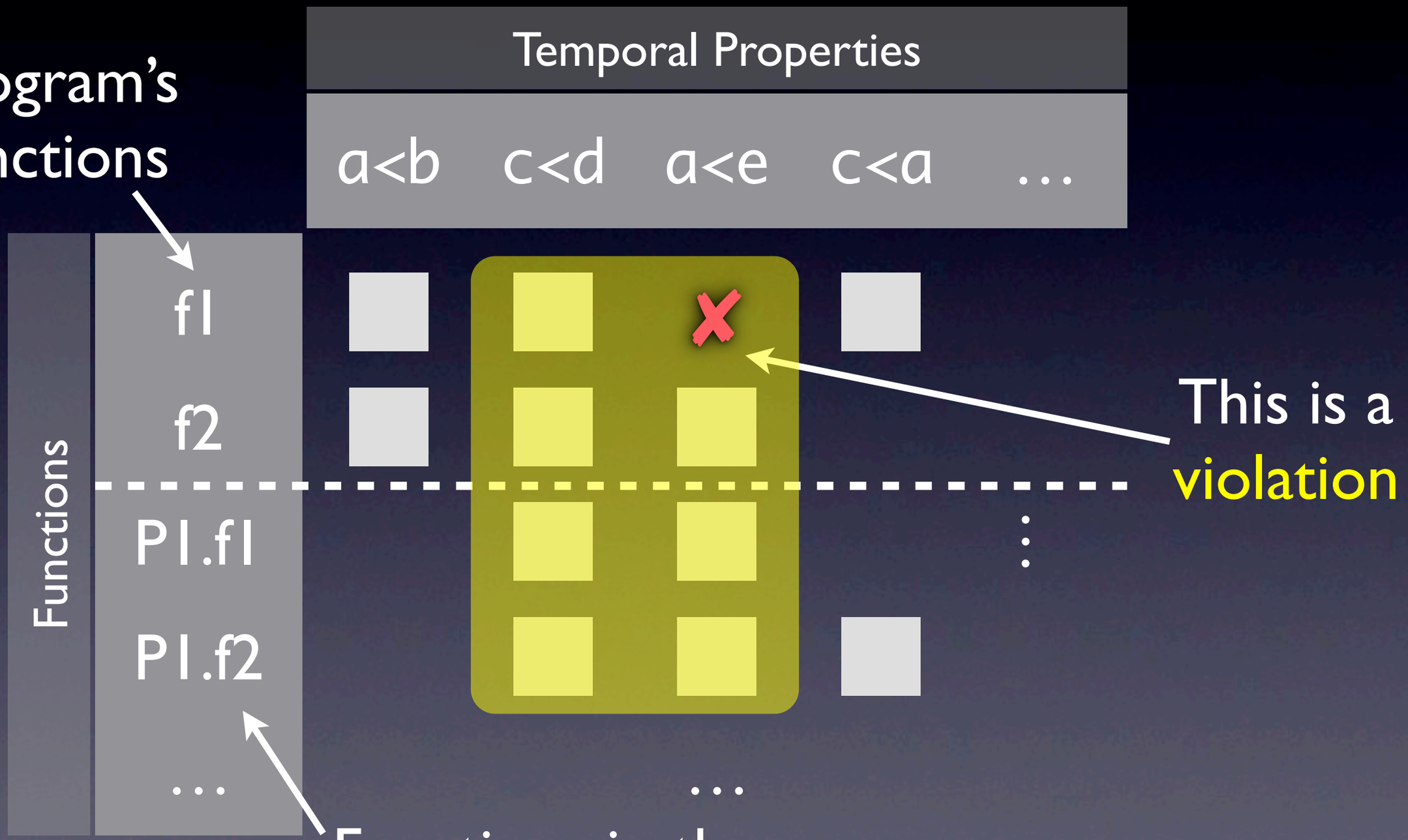
Finding Patterns



Detecting Violations



Detecting Violations



Evaluation

- 20 randomly chosen projects
  - Ran anomaly detection on each of them
    - Classify top 25%
    - Defects, code smells, false positives

20 Projects

tclxml-2.4.tar.gz
python-scw-0.4.7.tar.gz
dhcpdump-1.8.tar.gz
memcached-1.3.3.tar.gz
glade3-3.6.4.tar.bz2
cacao-0.95.tar.gz
psycopg-1.1.15.tar.gz
cksfv-1.3.13.tar.bz2
ggv-2.12.0.tar.bz2
gimp-2.6.6.tar.bz2
vdr-arghdirector-0.2.6.tar.gz
LDL-2.0.1.tar.gz
viewres-1.0.1.tar.bz2
Yap-5.1.3.tar.gz
xf86-video-savage-2.2.1.tar.bz2
daudio-0.3.tar.gz
httrack-3.43-4.tar.gz
concentration-1.2.tar.gz
mpich-1.2.7p1.tar.gz
otp_src_R13B.tar.gz

- Between 69 and 595,664 SLOC (C only)
  (generated using David A. Wheeler's 'SLOCCount')
- 136 violations found in 11 projects
- Analysis time per project < 6 minutes
  (gimp < 18 minutes)

Defect in Conspire 0.20

```
static int dcc_listen_init (…) {
    dcc->sok = socket(…);
    if (…) {
        while (…) {
            … = bind (dcc->sok, …);
        }
        /* with a small port range, reUseAddr is needed */
        setsockopt (dcc->sok, …, SO_REUSEADDR, …);
    }
    listen (dcc->sok, …);
}
```

should be called *before bind()*

Defect in cksfv-1.3.13

```
static int find_file (…)
{
    DIR *dirp;
    struct dirent *dirinfo;
    …
    dirp = opendir(".");
    if (dirp == NULL)
    {
        …
    }
    while ((dirinfo = readdir(dirp)) != NULL)
    {
        …
    }
    rewinddir(dirp);
    return 1;
}
```

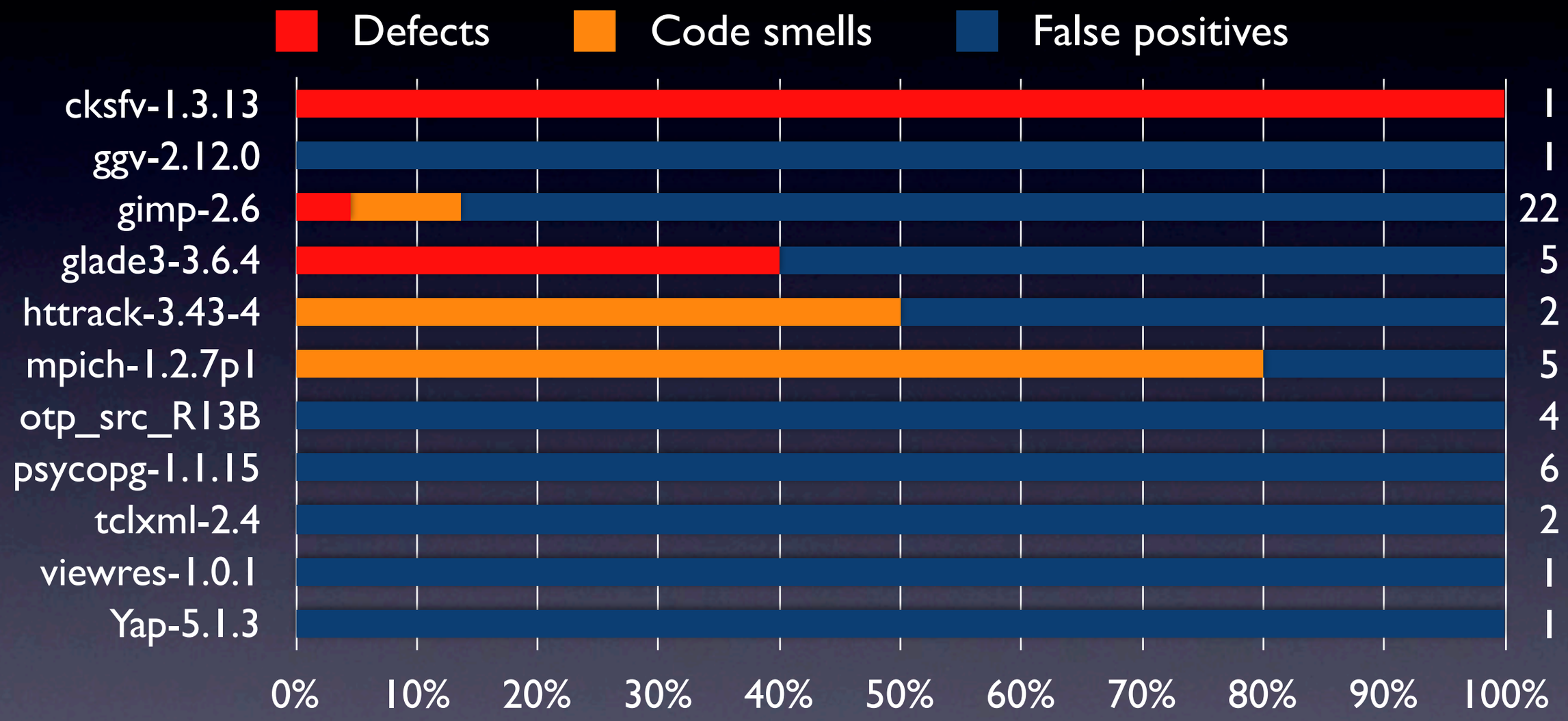
should call *closedir()* instead

Code smell in gimp-2.6.6

```
static gboolean gimp_page_selector_item_width_idle (…)
{
    GimpPageSelectorPrivate *priv = …;
    GtkTreeModel *model = GTK_TREE_MODEL (priv->store);
    GtkTreeIter iter;
    …
    for (… = gtk_tree_model_get_iter_first (model, &iter);
         …;
         … = gtk_tree_model_iter_next (model, &iter))
    {
        …
        gtk_tree_model_get (GTK_TREE_MODEL (priv->store),
            &iter, …, …, …);
        …
    }
    …
}
```

should be replaced with *model*

Violations



Global true positive rate: 22%

How to Improve Things?

- Use CTL formulas instead of temporal properties
- Explore API evolution
- Take user feedback into account

checkMyCode.org

http://www.checkmycode.org/index.php?action=result&token=x94e2k8xTZZ3bRZZwpv8LvEYTW4g8hOh

Google

Adnotacje ▾ RSS ▾

checkmycode.org

compare your code with the wisdom of linux

BETA

Home:

- Home/Analysis

About The Tool:

- Tutorial
- F.A.Q.
- Publications

Authors:

- Andrzej Wasylkowski
- Andreas Zeller

Contact:

- SE chair at Saarland University
- feedback@checkmycode.org

Share Share

Your result is now ready. Please click on a violation (red marked line) to see a detailed description for the violation.

```
1  void bar (void)
2  {
```

A call to closedir() is probably missing. There is a potential problem with a call to opendir().

>> Textual representation

DATA FLOW IN YOUR CODE:

- 1st arg of readdir() -> 1st arg of readdir()
- return value of opendir() -> 1st arg of readdir()

MISSING DATA FLOW:

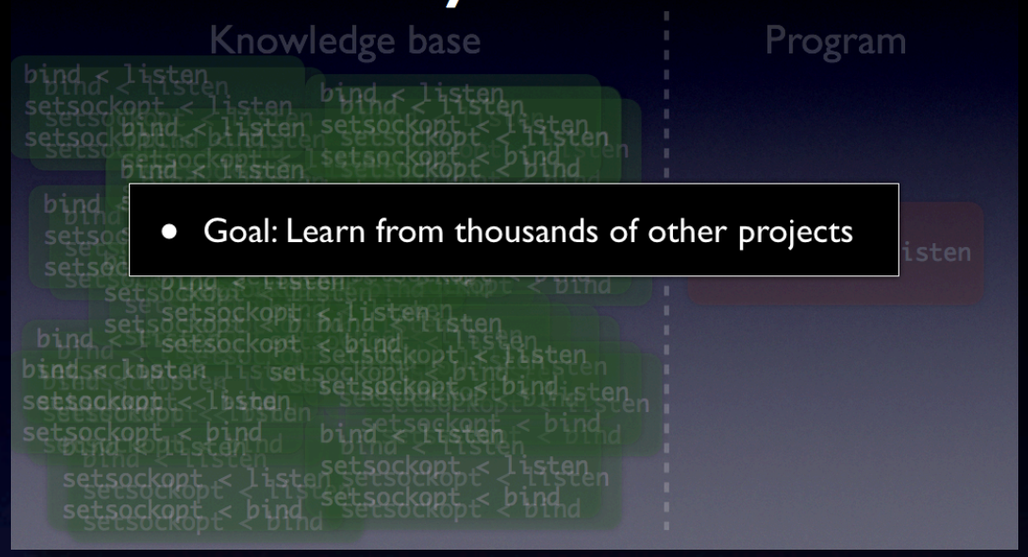
- return value of opendir() -> 1st arg of closedir()

>> Example code for valid data flow

- function '*fsdb_search_dir', line 54
- function 'dvipost_search', line 137
- function 'find_timebase', line 29

```
3      DIR *dirp;
4      struct dirent *dirinfo;
5
```

Cross-project Anomaly Detection



- Goal: Learn from thousands of other projects


```
void foo () {
  int fA;
  int fB = open("newFile");
  fA = open("myFile");
  while(j > 3){
    read(fA);
    write(fB, "Hello");
  }
  close(fA);
  close(fB);
}
```

```
open < write
open < read
write < write
read < read
write < close
read < close
open < close
```

Summary

Using a Knowledge Base