What is the Vocabulary of Flaky Tests?

@gustavopinto

Gustavo Pinto

Breno Miranda

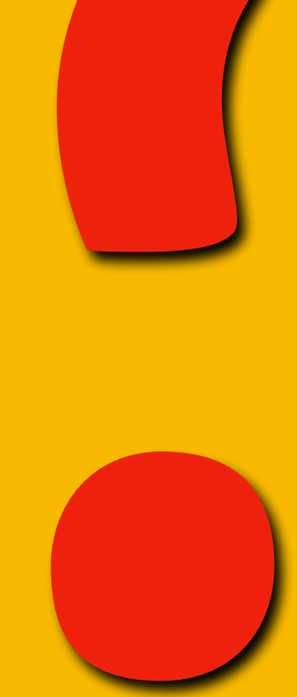
Supun Dissanayake

Marcelo d'Amorim

Christoph Treude

Antonia Bertolino

What is a flaky test?



@gustavopinto

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

Runs: 1/1 x Errors: 0 x Failures: 0

@gustavopinto

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

Runs: 1/1 x Errors: 1 x Failures: 0

@gustavopinto

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

What do we know about flaky tests?

Luo @ FSE 2014

Thorve @ ICSME 2018

Empirical-based

@gustavopinto

What do we know about flaky tests?

Luo @ FSE 2014

Thorve @ ICSME 2018

Bell @ ICSE 2018

Lam @ ICST 2019

Empirical-based

Dynamic detection of flakiness

@gustavopinto

What do we know about flaky tests?

Luo @ FSE 2014

Thorve @ ICSME 2018

Bell @ ICSE 2018

Lam @ ICST 2019

Empirical-based

Dynamic detection of flakiness

Static detection of flakiness

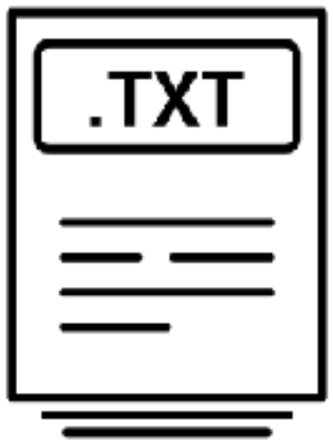
?

@gustavopinto

How to statically detect if a test code is flaky

@gustavopinto

Flaky tests



Tokenized Flaky test

Machine Learning algorithms

Ranking of code identifiers associated with flakiness

Vocabulary of Flaky tests

@gustavopinto

Flaky tests

Tokenized Flaky test

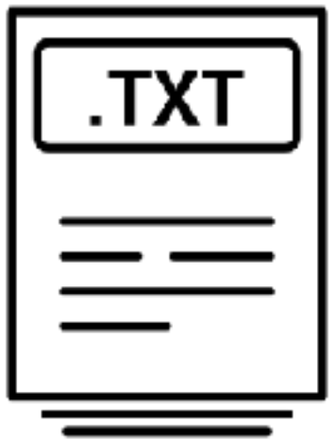
Machine Learning algorithms

Ranking of code identifiers associated with flakiness

Vocabulary of Flaky tests

@gustavopinto

Flaky tests



Tokenized Flaky test

Machine Learning algorithms

Ranking of code identifiers associated with flakiness

Vocabulary of Flaky tests

@gustavopinto

Flaky tests

Tokenized Flaky test

Machine Learning algorithms

Ranking of code identifiers associated with flakiness

Vocabulary of Flaky tests

@gustavopinto

Flaky tests

Tokenized Flaky test

Machine Learning algorithms

Ranking of code identifiers associated with flakiness

Vocabulary of Flaky tests

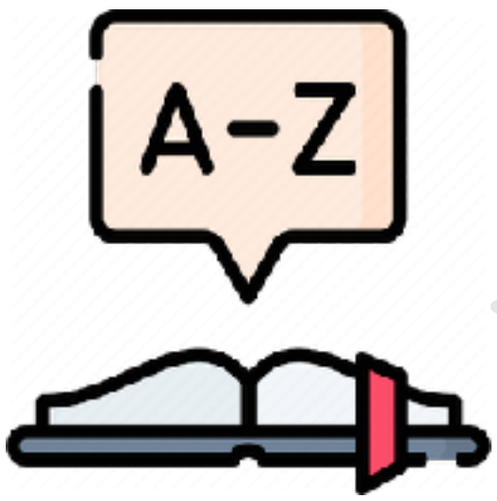
@gustavopinto

Flaky tests

Tokenized Flaky test

Machine Learning algorithms

Ranking of code identifiers associated with flakiness



Vocabulary of Flaky tests

@gustavopinto

Is this test flaky?

```
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

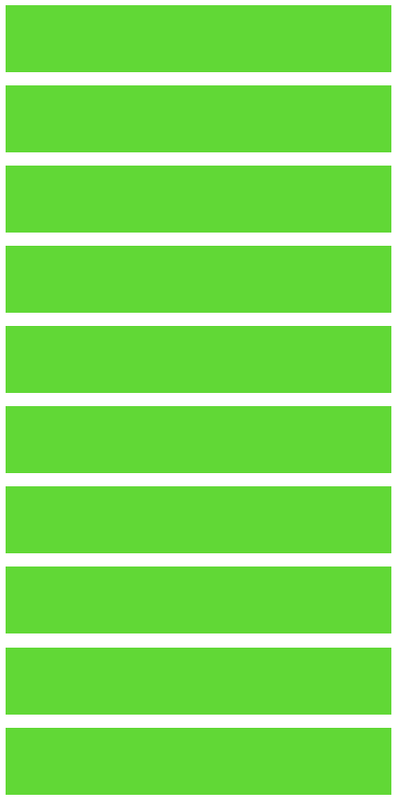
@gustavopinto

Is this test flaky?

```
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

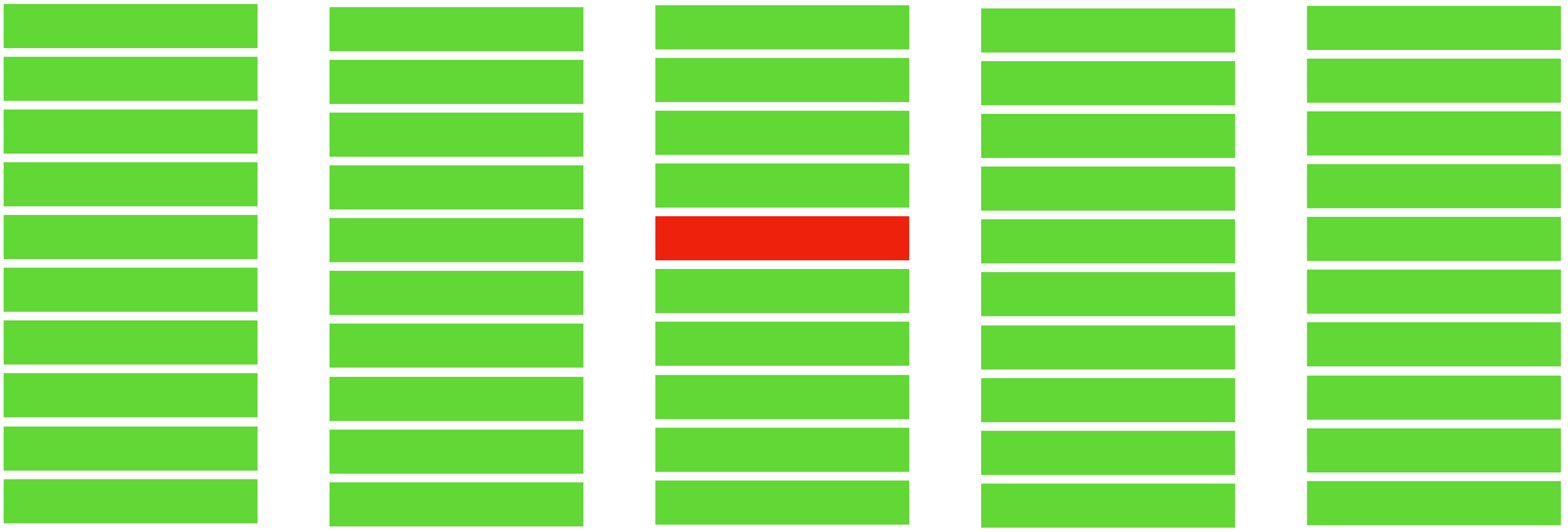
Runs: 1/1 x Errors: 0 x Failures: 0

@gustavopinto



10 executions

@gustavopinto



@gustavopinto

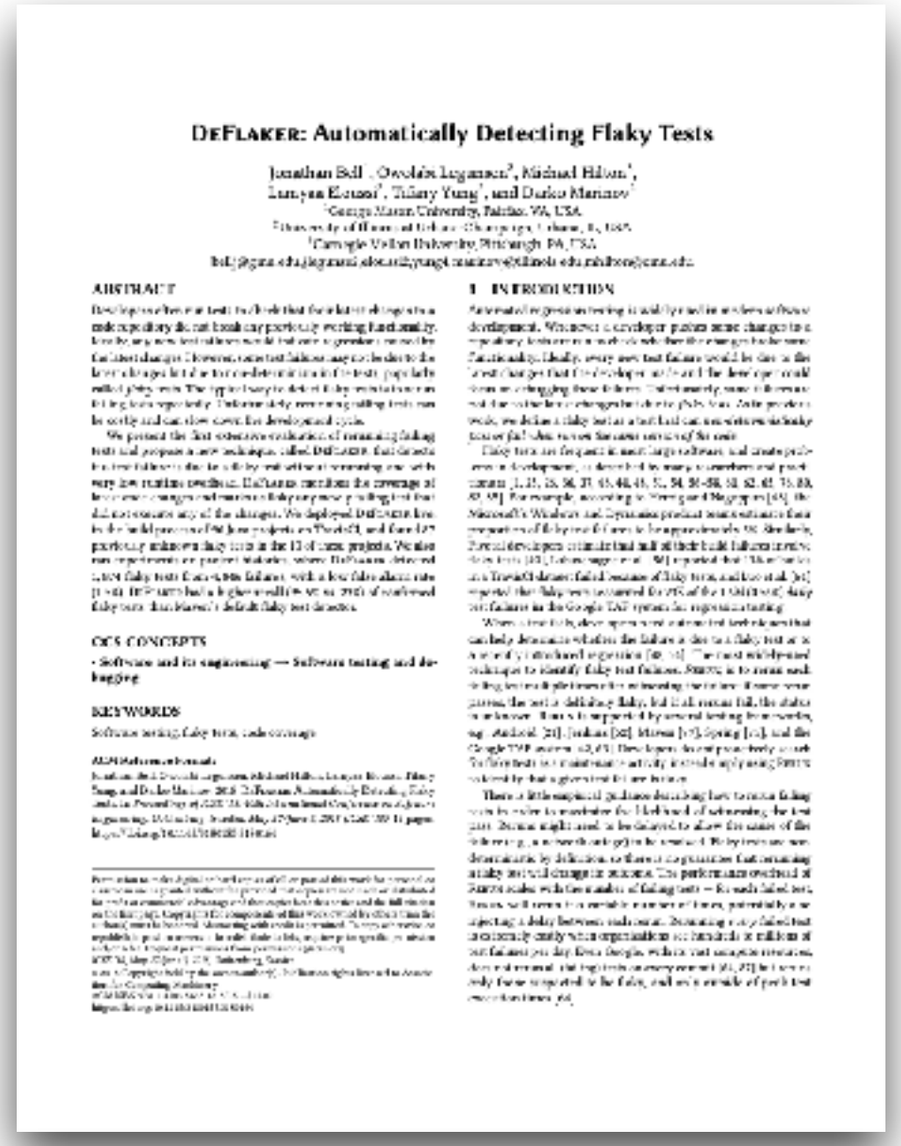
@gustavopinto

GitHub Project
achilles
alluxio
ambari
assertj-core
checkstyle
commons-exec
dropwizard
hadoop
handlebars
hbase
hector
httpcore
jackrabbit-oak
jimfs
logback
ninja
okhttp
oozie
orbit
oryx
spring-boot
togglz
undertow
wro4j
zxing

@gustavopinto

GitHub Project
achilles
alluxio
ambari
assertj-core
checkstyle
commons-exec
dropwizard
hadoop
handlebars
hbase
hector
httpcore
jackrabbit-oak
jimfs
logback
ninja
okhttp
oozie
orbit
oryx
spring-boot
togglz
undertow
wro4j
zxing

The same studied here



DeFlaker: Automatically Detecting Flaky Tests

Jonathan Bell[1], Owolabi Legunsen[2], Michael Hilton[1],
Lamyaa Eloussi[2], Tifany Yung[1], and Darko Marinov[1]
[1]George Mason University, Fairfax, VA, USA
[2]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[3]Carnegie Mellon University, Pittsburgh, PA, USA
{bell,gyung}@gmu.edu,{legunse2,eloussi2}@illinois.edu,mhilton@cmu.edu,marinov@illinois.edu

ABSTRACT

1 INTRODUCTION

CCS CONCEPTS

• Software and its engineering → Software testing and debugging

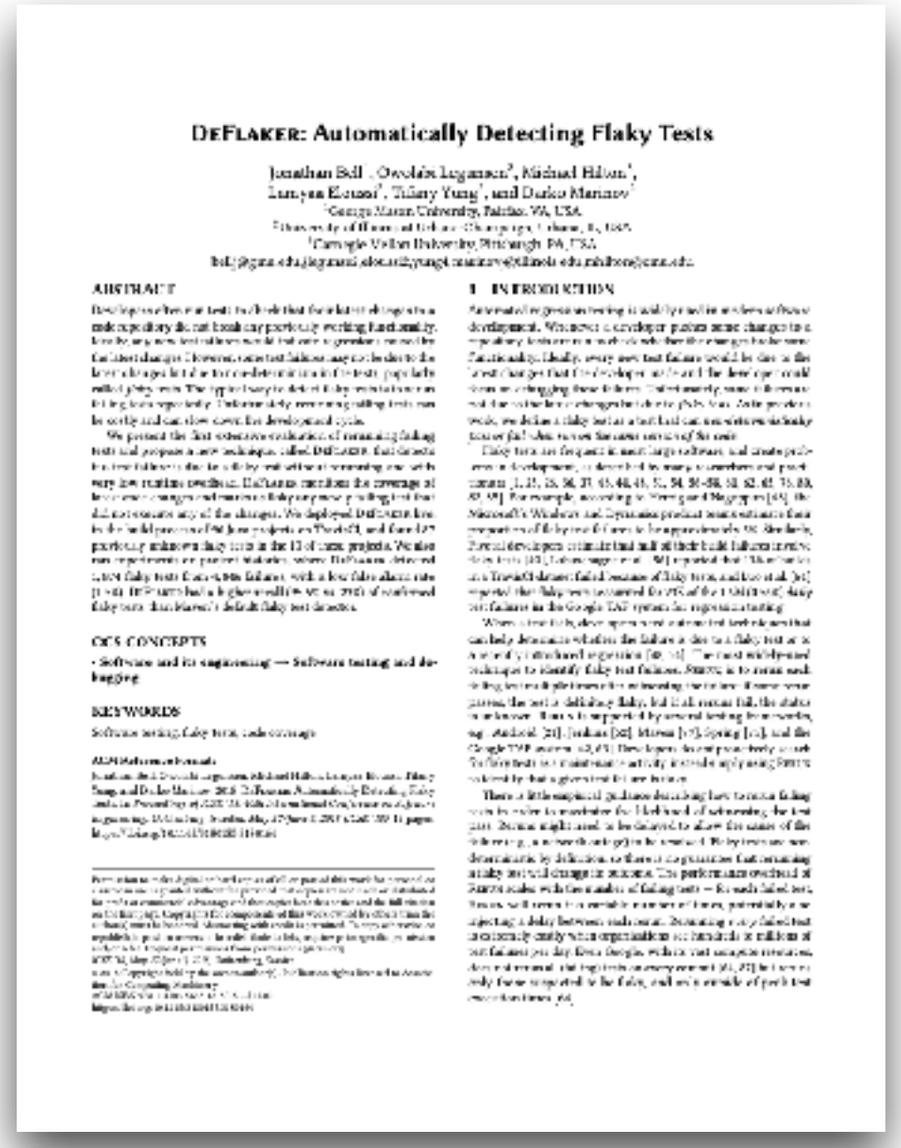
KEYWORDS

Bell @ ICSE 2018

@gustavopinto

GitHub Project
achilles
alluxio
ambari
assertj-core
checkstyle
commons-exec
dropwizard
hadoop
handlebars
hbase
hector
httpcore
jackrabbit-oak
jimfs
logback
ninja
okhttp
oozie
orbit
oryx
spring-boot
togglz
undertow
wro4j
zxing

The same studied here



DeFlaker: Automatically Detecting Flaky Tests

Jonathan Bell[1], Owolabi Legunsen[2], Michael Hilton[1],
Lamyaa Eloussi[2], Tifany Yung[1], and Darko Marinov[2]
[1]George Mason University, Fairfax, VA, USA
[2]University of Illinois at Urbana-Champaign, Urbana, IL, USA
[1]Carnegie Mellon University, Pittsburgh, PA, USA
{bell,gyani,eloussi,tyung}@gmu.edu,{legunse2,marinov}@illinois.edu,mhilton@cmu.edu

ABSTRACT

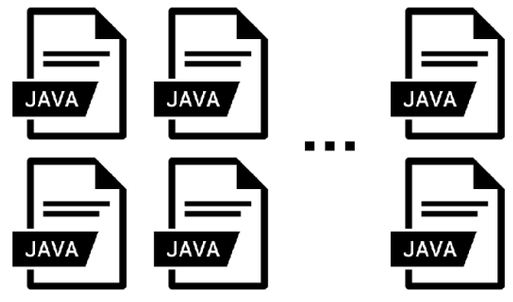
1 INTRODUCTION

CCS CONCEPTS

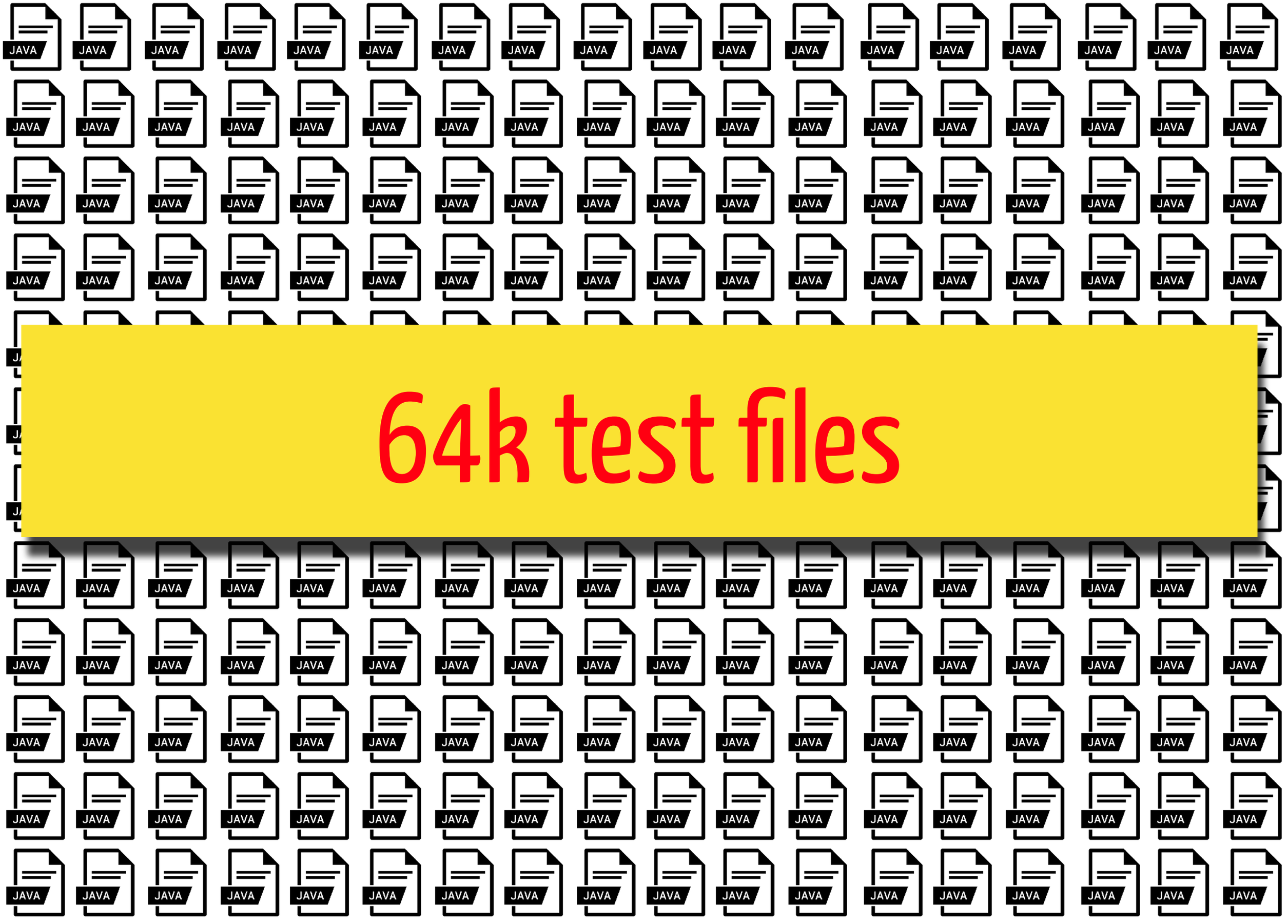
• Software and its engineering → Software testing and debugging

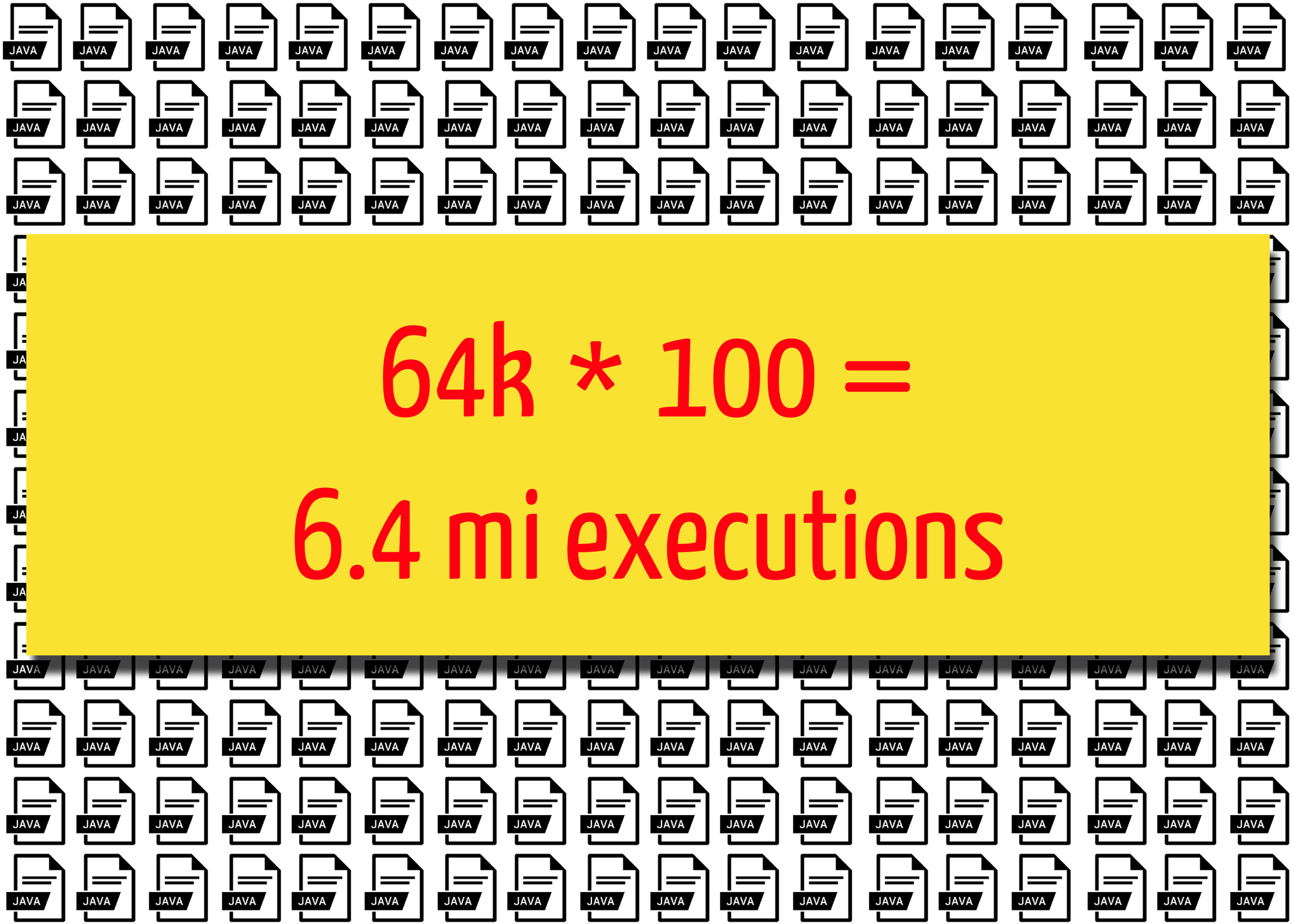
KEYWORDS

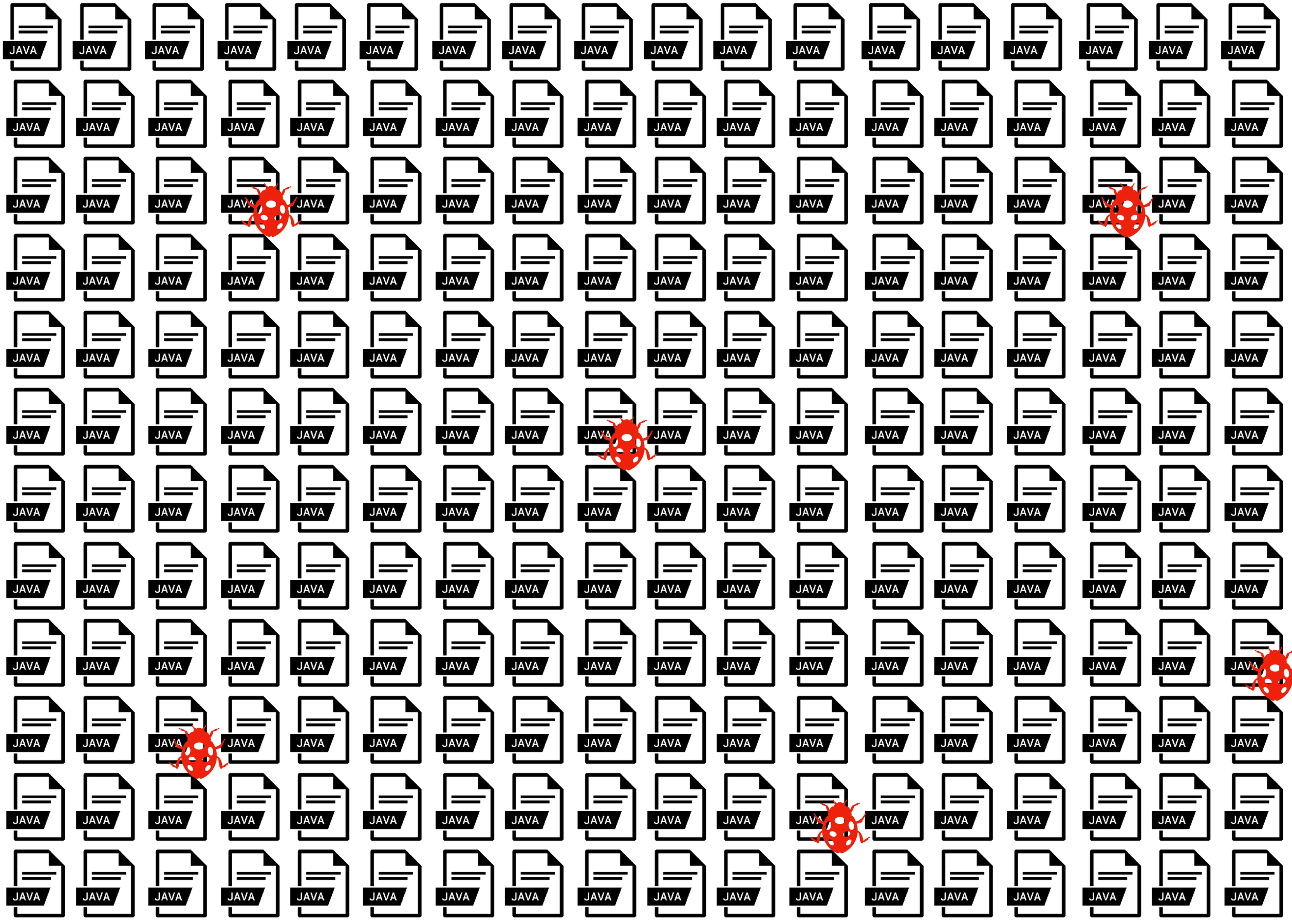
Bell @ ICSE 2018



@gustavopinto


64k test files
@gustavopinto


64k * 100 =
6.4 mi executions
@gustavopinto


JAVA
@gustavopinto






@gustavopinto

@gustavopinto

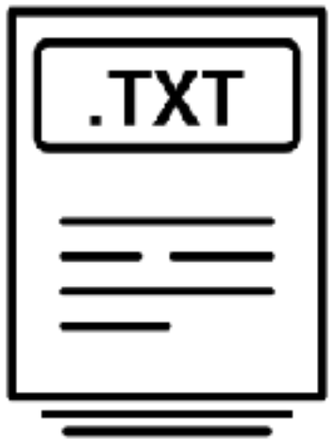

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

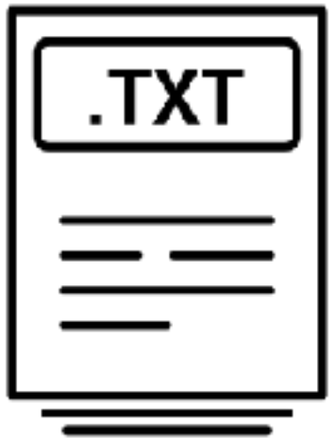
```java
public class TestIdentifyEncoder {
    @Test
    public void testCodingEmptySrcBuffer() throws Exception {
        final WritableByteChannelMock channel = new WritableByteChannelMock(64);
        final SessionOutputBuffer outbuf = new SessionOutputBufferImpl(1024, 128);
        final HttpTransportMetricsImpl metrics = new HttpTransportMetricsImpl();

        final IdentityEncoder encoder = new IdentityEncoder(channel, outbuf, metrics);
        encoder.write(CodecTestUtils.wrap("stuff"));

        final ByteBuffer empty = ByteBuffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final String s = channel.dump(Consts.ASCII);

        Assert.assertTrue(encoder.isCompleted());
        Assert.assertEquals("stuff", s);
    }
}
```

```
public class test identify encoder {
    @Test
    public void test coding empty src buffer() throws exception {
        final writable byte channel mock channel = new writable byte channel mock(64);
        final session output buffer outbuf = new session output buffer impl(1024, 128);
        final http transport metrics impl metrics = new http transport metrics impl();

        final identity encoder encoder = new identity encoder channel outbuf metrics
       encoder.write(codec test utils wrap("stuff"));

        final byte buffer empty = byte buffer.allocate(100);
        empty.flip();
        encoder.write(empty);
        encoder.write(null);
        encoder.complete();

        outbuf.flush(channel);
        final string s = channel.dump(consts.ascii);

        assert.assert true(encoder.is completed());
        assert.assert equals("stuff", s);
    }
}
```



test identify encoder coding empty src buffer byte channel mock writable session output outbuf session http transport metrics impl identity codec utils wrap buffer allocate flip write complete flush dump consts ascii is completed assert equals

Known Flaky test

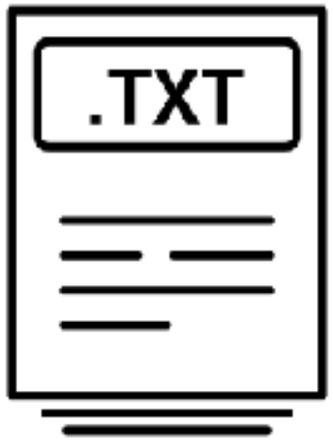


Tokenized Flaky test

Machine Learning algorithms

@gustavopinto

Known Flaky test



Tokenized Flaky test

Machine Learning algorithms

Weka

Nearest Neighbour

Support Vector Machine

Decision Tree

Naive Bayes

Random Forest

Implementation available at:
https://github.com/damorimRG/msr4flakiness/

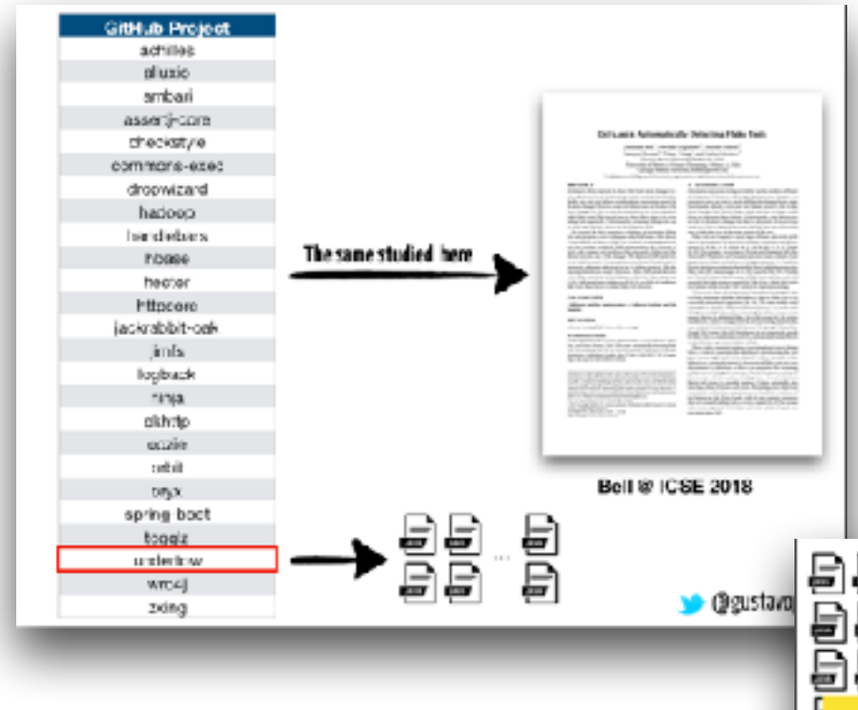
@gustavopinto

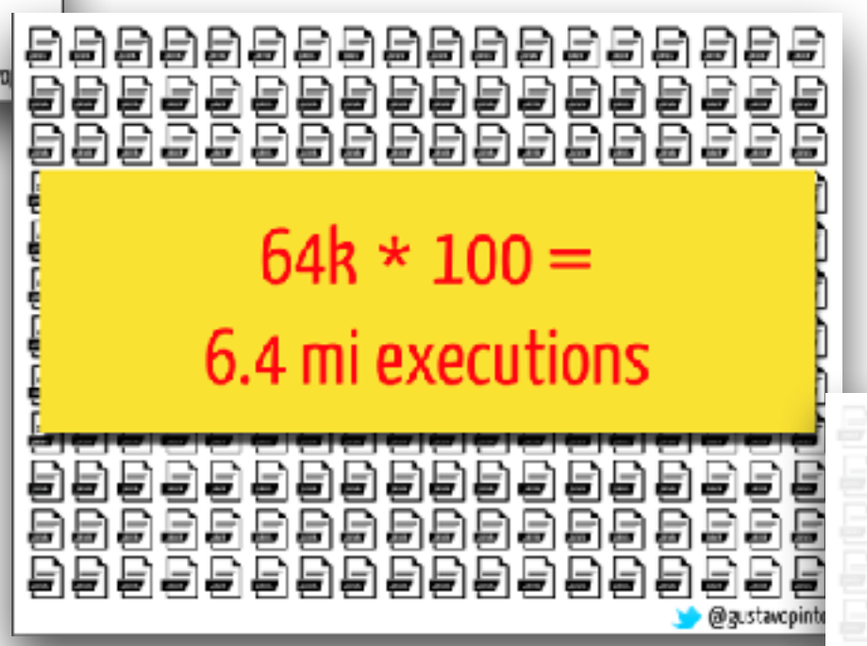
RQ1:

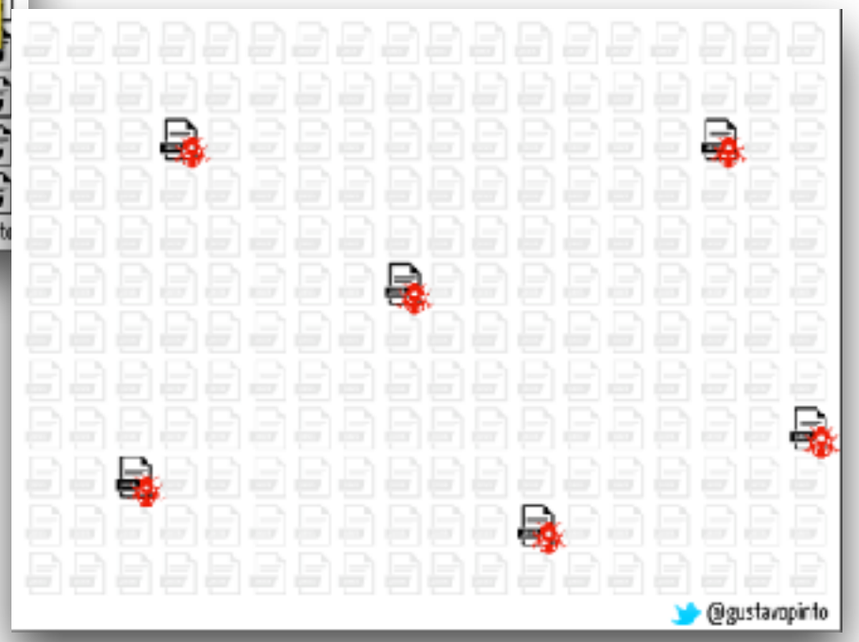
How prevalent are flaky tests?

@gustavopinto

RQ1: How prevalent are flaky tests?







@gustavopinto

RQ1: How prevalent are flaky tests?

Project	# Test	# Flaky	% Flaky
alluxio	3,034	12	0.4
hector	322	40	12.4
jackrabbit-oak	13,193	2	2
okhttp	1,682	19	19
undertow	609	2	2
wro4j	1,158	11	11

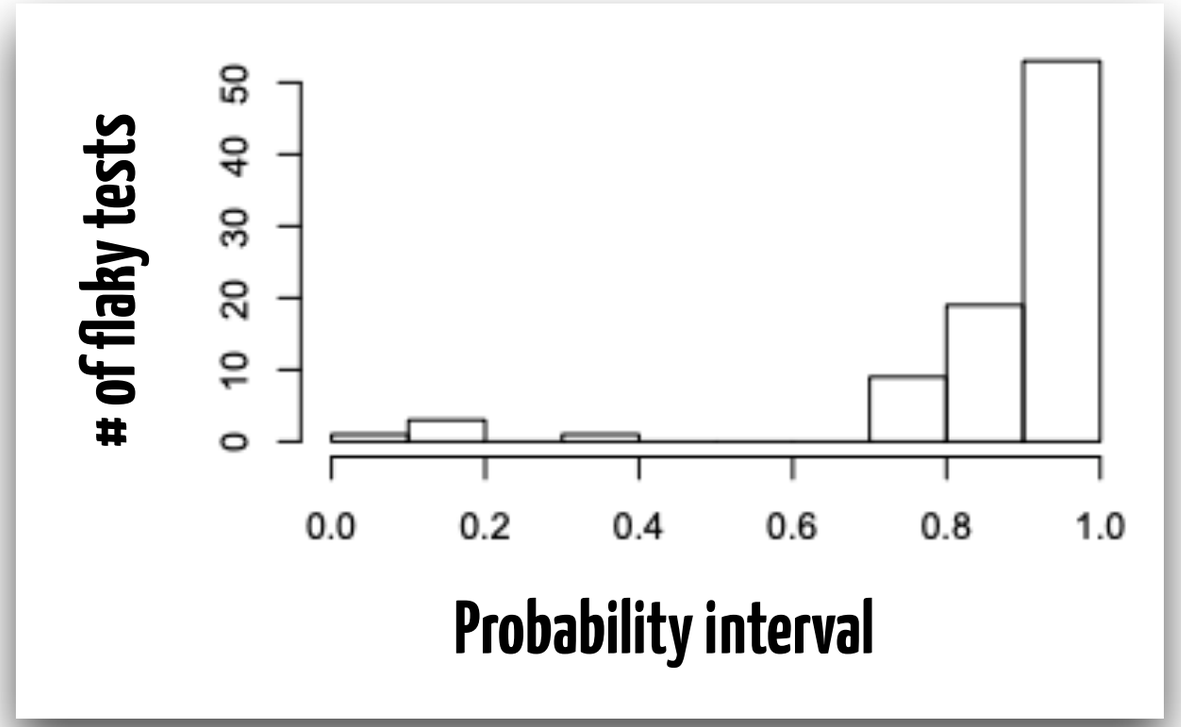
25% of the projects
have at least one flaky test

We found
86 flaky tests

@gustavopinto

RQ1: How prevalent are flaky tests?

Project	# Test	# Flaky	% Flaky
alluxio	3,034	12	0.4
hector	322	40	12.4
jackrabbit-oak	13,193	2	2
okhttp	1,682	19	19
undertow	609	2	2
wro4j	1,158	11	11



25% of the projects
have at least one flaky test

We found
86 flaky tests

70% (61 out of the 86)
passed more than 90%

@gustavopinto

RQ2:

How accurately can we predict test flakiness?

@gustavopinto

RQ2: Can we predict flakiness?

ML	Precision	Recall	F1	MCC	AUC
Random Forest	**0.99**	0.91	**0.95**	**0.90**	**0.98**
Decision Tree	0.98	0.88	0.89	0.77	0.91
Naive Bayes	0.93	0.80	0.86	0.74	0.93
Support Vector	0.93	**0.92**	0.93	0.85	0.92
Nearest Neighbour	0.97	0.88	0.92	0.85	0.93

Random Forest
achieved best precision

Tuning (e.g., # of trees)
Had no performance impact

@gustavopinto

RQ3:

What value do different features add to the classifier?

@gustavopinto

RQ3: What's the value of different features

Random Forest

Features	Precision	Recall	F1	MCC	AUC
All features	0.99	0.91	0.95	0.90	0.98
No stemming	0.99	0.91	0.95	0.90	0.98
No Stop W. removal	0.99	0.91	0.95	0.90	0.98
No Lowercasing	0.98	0.91	0.94	0.89	0.98
No Java Keywords	0.99	0.90	0.94	0.89	0.98

@gustavopinto

RQ3: What's the value of different features

Random Forest

Features	Precision	Recall	F1	MCC	AUC
All features	0.99	0.91	0.95	0.90	0.98
No stemming	0.99	0.91	0.95	0.90	0.98
No Stop W. removal	0.99	0.91	0.95	0.90	0.98
No Lowercasing	0.98	0.91	0.94	0.89	0.98
No Java Keywords	0.99	0.90	0.94	0.89	0.98

Support Vector Machine

Features	Precision	Recall	F1	MCC	AUC
All features	0.93	0.92	0.93	0.85	0.93
No stemming	0.93	0.92	0.93	0.85	0.93
No Stop W. removal	0.93	0.92	0.93	0.85	0.93
No Lowercasing	0.91	0.93	0.92	0.84	0.92
No Java Keywords	0.93	0.92	0.93	0.85	0.93

@gustavopinto

RQ3: What's the value of different features

Random Forest

Features	Precision	Recall	F1	MCC	AUC
All features	0.99	0.91	0.95	0.90	0.98
No stemming	0.99	0.91	0.95	0.90	0.98
No Stop W. removal	0.99	0.91	0.95	0.90	0.98
No Lowercasing	0.98	0.91	0.94	0.89	0.98
No Java Keywords	0.99	0.90	0.94	0.89	0.98

Support Vector Machine

Features	Precision	Recall	F1	MCC	AUC
All features	0.93	0.92	0.93	0.85	0.93
No stemming	0.93	0.92	0.93	0.85	0.93
No Stop W. removal	0.93	0.92	0.93	0.85	0.93
No Lowercasing	0.91	0.93	0.92	0.84	0.92
No Java Keywords	0.93	0.92	0.93	0.85	0.93

No performance impact

@gustavopinto

RQ4:

Which test code identifiers are strongly associated with test flakiness?

@gustavopinto

RQ4: What's the vocabulary?

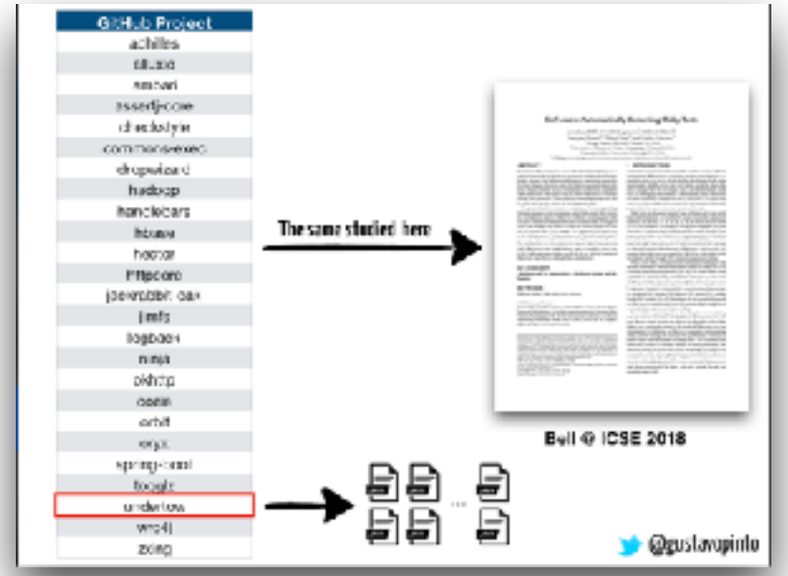
Features	inf. gain	Flaky # test	Flaky # projects	Non-Flaky # test	Non-Flaky # projects
job	0.20	524	2	4	1
table	0.14	406	4	8	2
id	0.14	552	9	52	4
action	0.13	387	3	8	2
oozie	0.13	274	1	0	0
services	0.13	371	2	7	1
coord	0.11	307	1	0	0
getid	0.11	287	4	1	1
coordinator	0.10	258	1	0	0
xml	0.10	147	2	6	2
workflow	0.09	207	1	0	0
getstatus	0.08	246	2	2	2
record	0.08	296	2	18	1
jpa	0.07	207	2	0	0
jpaservice	0.07	200	1	0	0
service	0.07	367	4	67	3

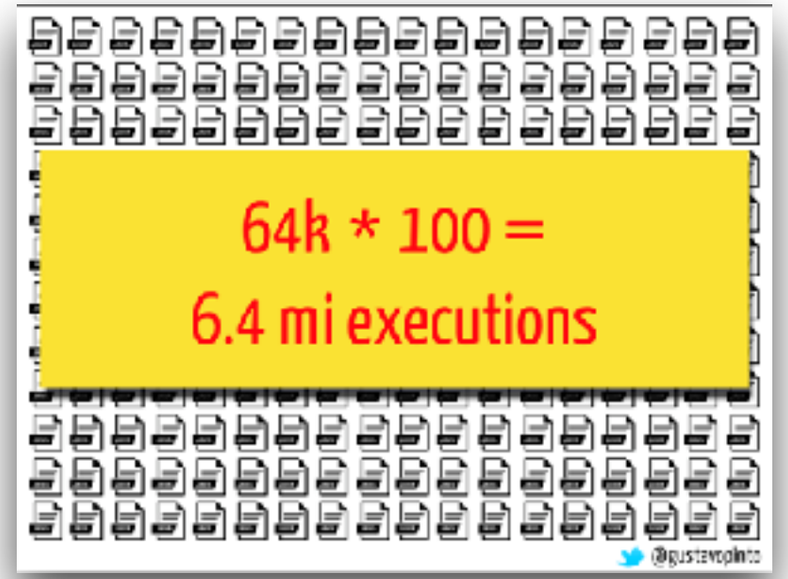
Many of them associated with

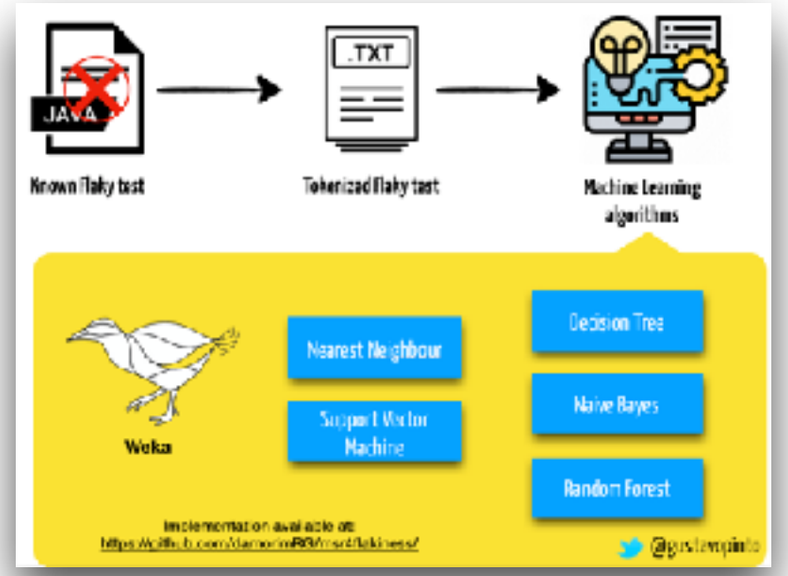
remote tasks

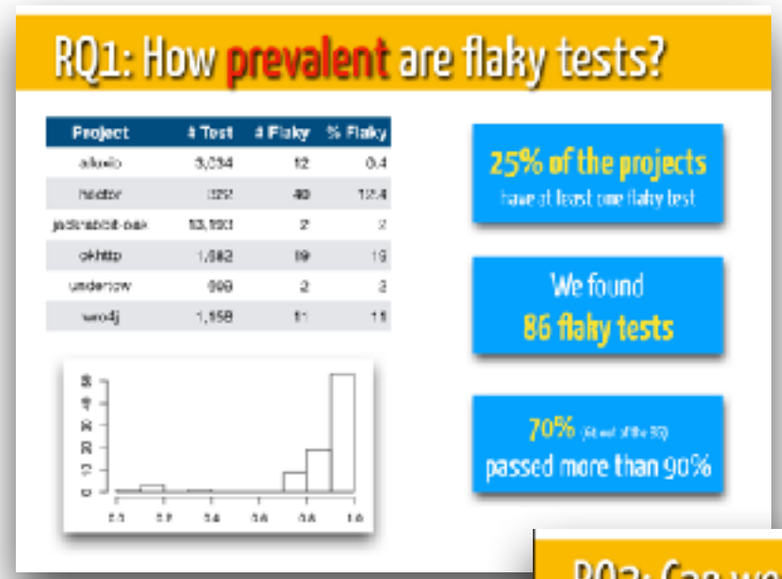
and/or

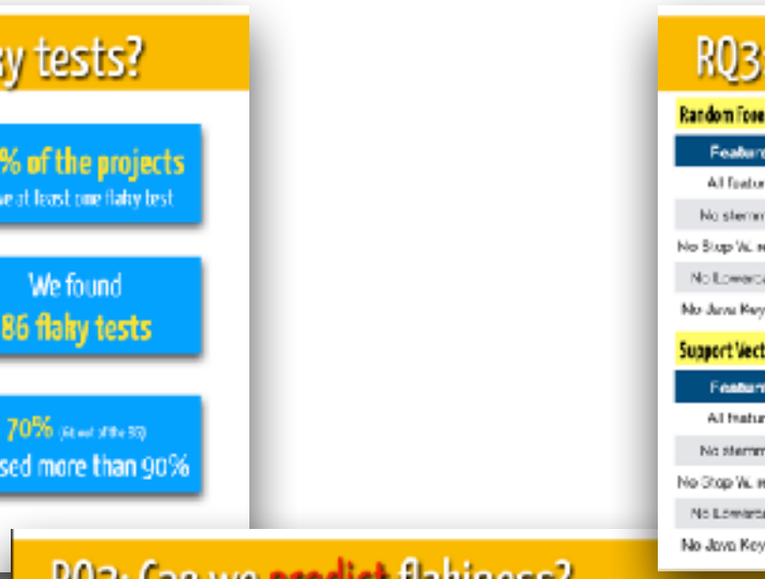
queue events

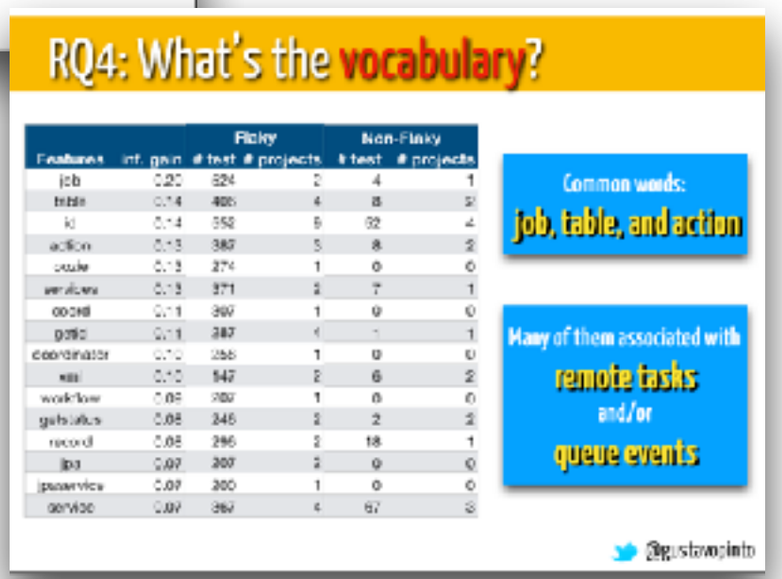












@gustavopinto