Clone matching Javascript to Stackoverflow posts

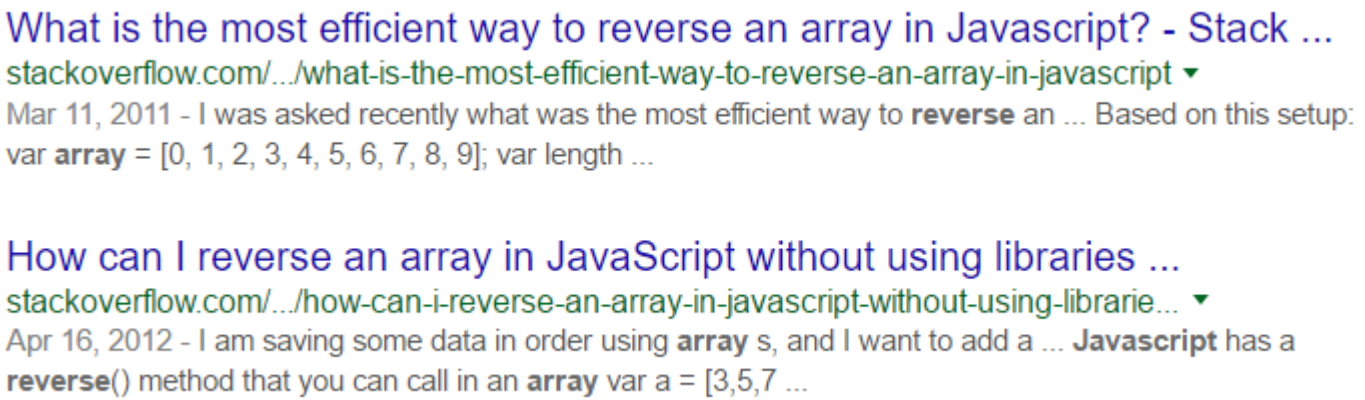
Zored Ahmer

zbahmer@cim.mcgill.ca

ECSE 611 Final Project

This programmer wants to reverse an array

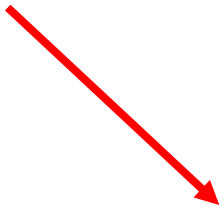
So, they google ‘reverse array javascript’ and Google gives a few answers

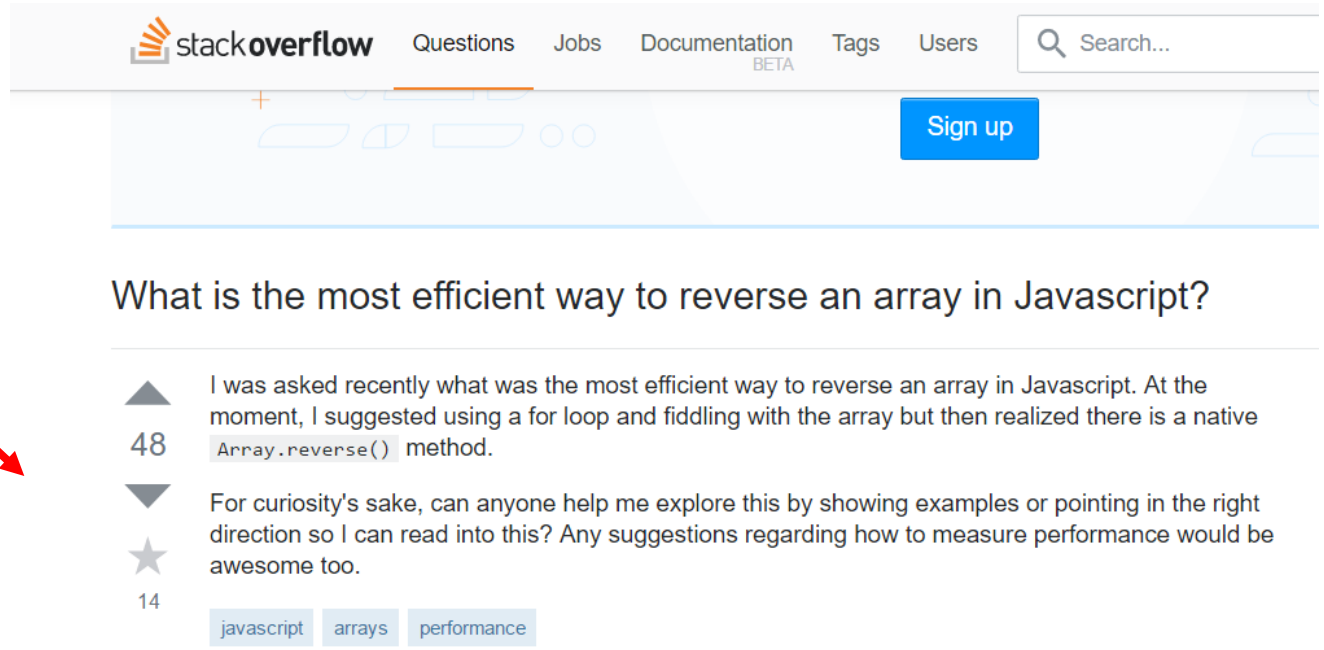


Google gives them a link to a handy Q&A website

What is the most efficient way to reverse an array in Javascript? - Stack ...
stackoverflow.com/.../what-is-the-most-efficient-way-to-reverse-an-array-in-javascript
Mar 11, 2011 - I was asked recently what was the most efficient way to **reverse** an ... Based on this setup: var **array** = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]; var length ...

How can I reverse an array in JavaScript without using libraries ...
stackoverflow.com/.../how-can-i-reverse-an-array-in-javascript-without-using-librarie...
Apr 16, 2012 - I am saving some data in order using **array** s, and I want to add a ... **Javascript** has a **reverse**() method that you can call in an **array** var a = [3,5,7 ...





http://stackoverflow.com/q/5276953

Based on this setup:

54

```
var array = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9];
var length = array.length;
```

`Array.reverse();` is the first or second slowest!

The benchmarks are here: http://jsperf.com/js-array-reverse-vs-while-loop/5

http://stackoverflow.com/a/9113136

54

Based on this setup:

```
var array = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9];
var length = array.length;
```

`Array.reverse();` is the first or second slowest!

The benchmarks are here: http://jsperf.com/js-array-reverse-vs-while-loop/5

Temporary swap:

First variation:

```
function temporarySwap(array)
{
    var left = null;
    var right = null;
    var length = array.length;
    for (left = 0, right = length - 1; left < right; left += 1, right -= 1)
    {
        var temporary = array[left];
        array[left] = array[right];
        array[right] = temporary;
    }
    return array;
}
```

http://stackoverflow.com/a/9113136

Ideally you would put in a link every time you take some code from the internet

```
//Uses swaps to reverse an array
//Reference: http://stackoverflow.com/a/9113136
function reverse(array)
{
    var left = null;
    var right = null;
    var length = array.length;
    for (left = 0, right = length - 1; left < right; left += 1, right -= 1)
    {
        var temporary = array[left];
        array[left] = array[right];
        array[right] = temporary;
    }
    return array;
}
```

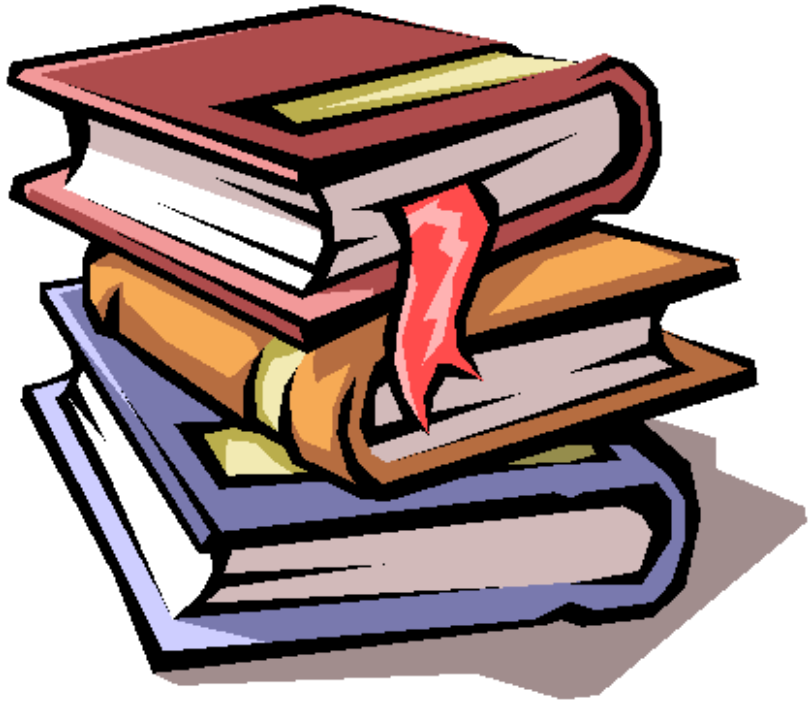
Ideally you would put in a link every time you take some code from the internet

```
//Uses swaps to reverse an array
//Reference: http://stackoverflow.com/a/9113136
function reverse(array)
{
    var left = null;
    var right = null;
    var length = array.length;
    for (left = 0, right = length - 1; left < right; left += 1, right -= 1)
    {
        var temporary = array[left];
        array[left] = array[right];
        array[right] = temporary;
    }
    return array;
}
```

..but maybe you forgot or didn't think to do it

There are problems with not keeping track of code taken from the internet

Possible copyright issues



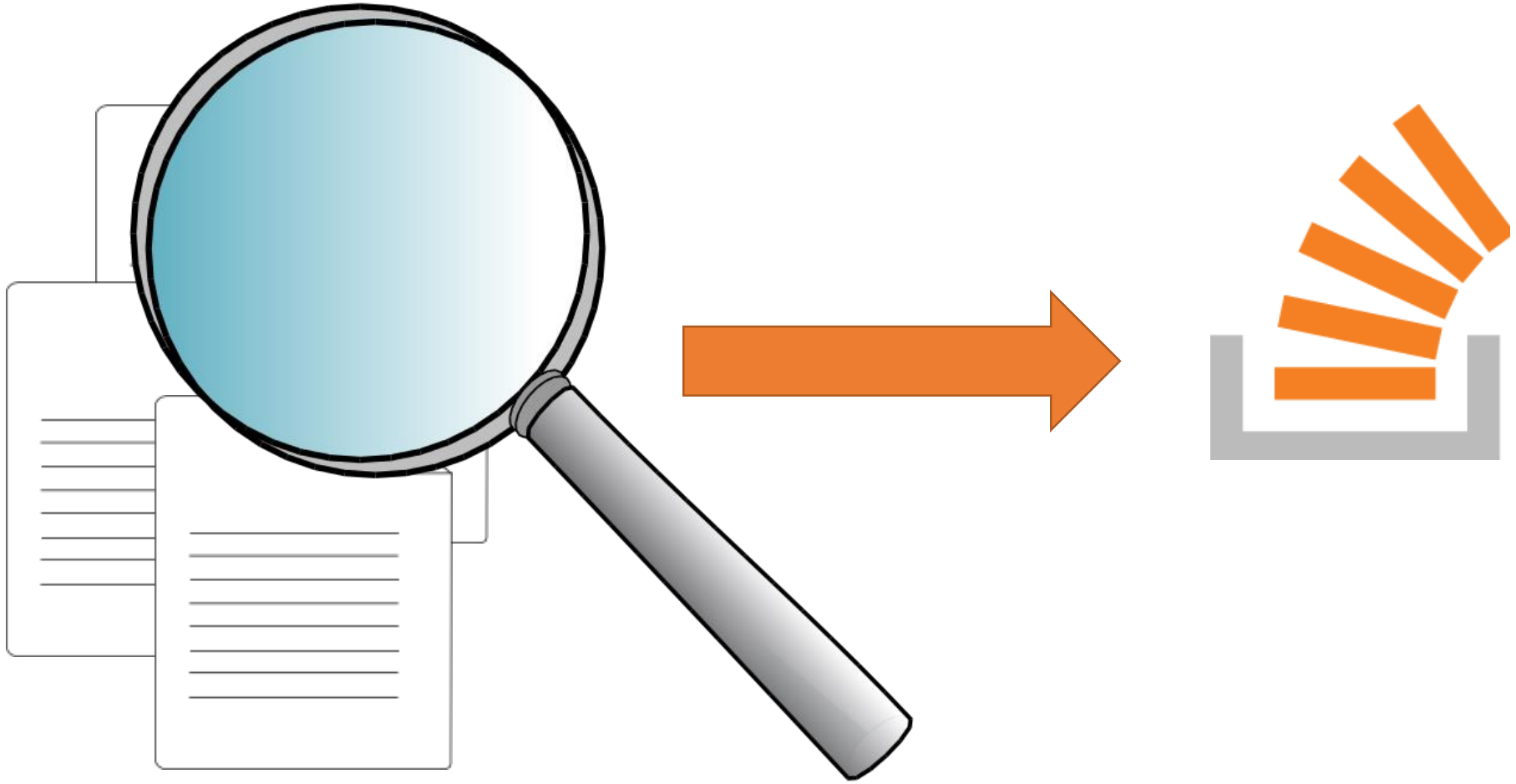
Missing Documentation

Photo: Steven DePolo[1]

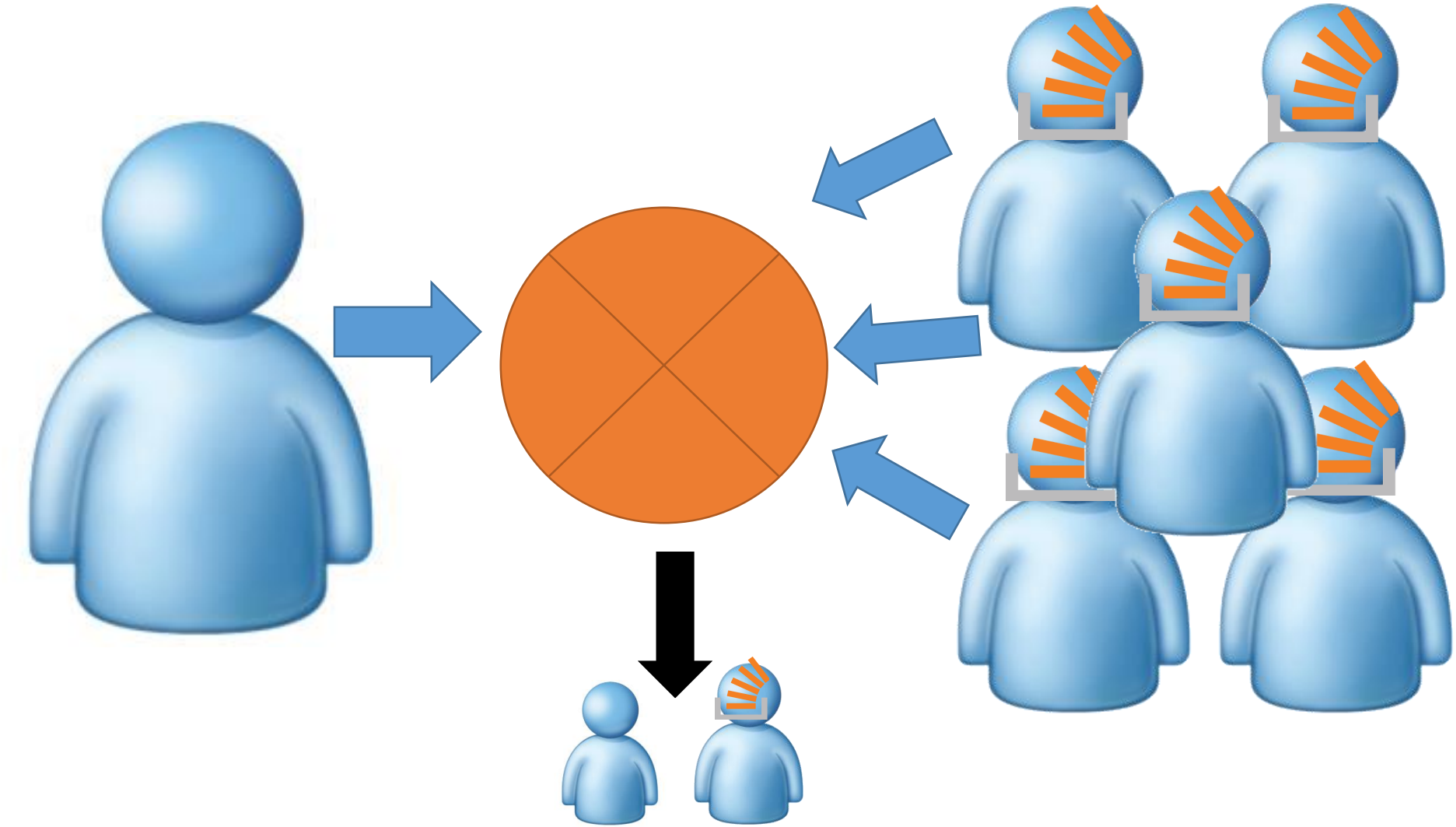
Aged Code

[1]https://www.flickr.com/photos/stevendepolo/4528758992

So, we might want to examine our code base and see how much of it came from StackOverflow



Idea: Can we use clone detection between our code and Stackoverflow code snippets?



There are some immediate concerns with this approach

```
var newObject = JSON.parse(JSON.stringify(oldObject));
```
[1]

Code snippets could be too small to be matched

```
var clonedObject = {
  knownProp: obj.knownProp,
  ..
}
```
[2]

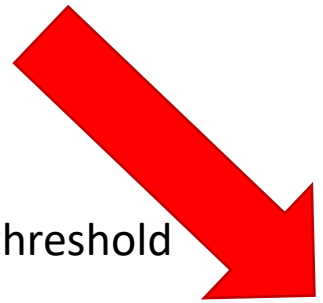
Snippets might be human readable but throw parse errors

[1]http://stackoverflow.com/a/4591639
[2]http://stackoverflow.com/a/5344074

There are some immediate concerns with this approach

```
var newObject = JSON.parse(JSON.stringify(oldObject));
```
[1]

Code snippets could be too small to be matched



Set too small a threshold

```
var clonedObject = {
  knownProp: obj.knownProp,
  ..
}
```
[2]

Snippets might be human readable but throw parse errors

[1]http://stackoverflow.com/a/4591639
[2]http://stackoverflow.com/a/5344074

There are some immediate concerns with this approach

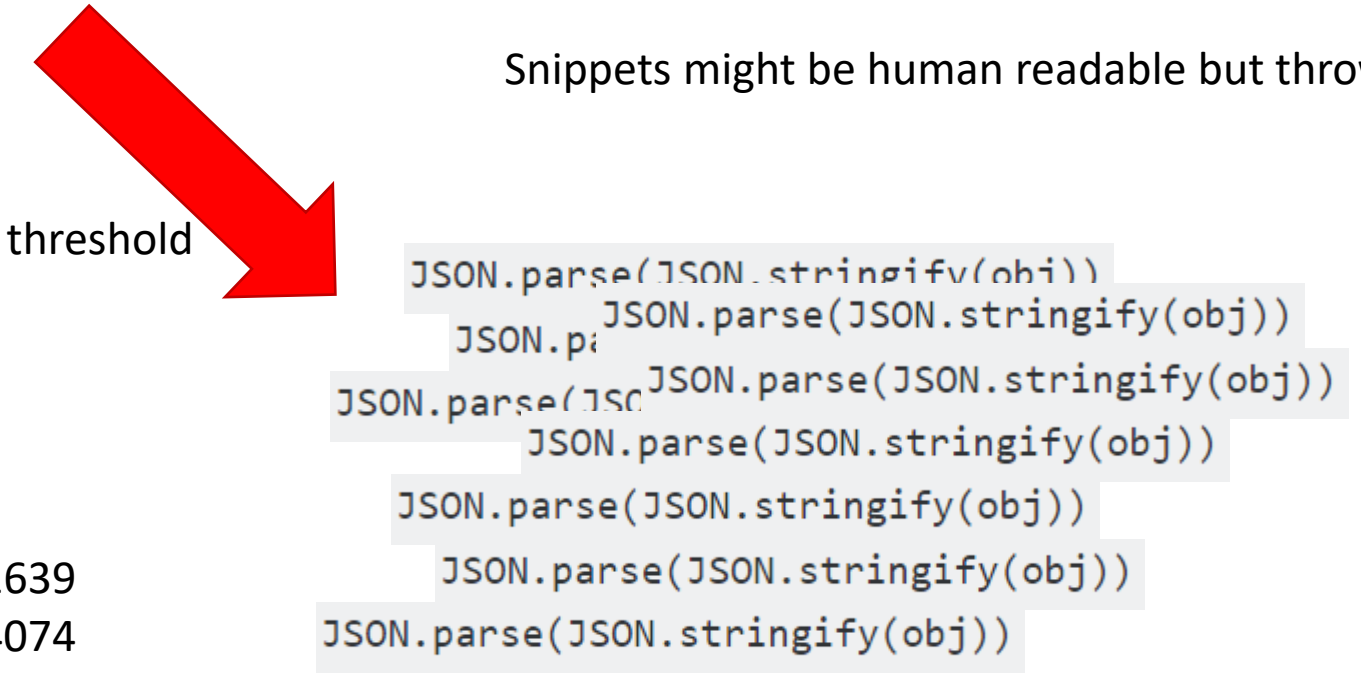
```
var newObject = JSON.parse(JSON.stringify(oldObject));
```
[1]

Code snippets could be too small to be matched

```
var clonedObject = {
  knownProp: obj.knownProp,
  ..
}
```
[2]

Snippets might be human readable but throw parse errors

Set too small a threshold



Too many results and a large number of false positives

[1]http://stackoverflow.com/a/4591639
[2]http://stackoverflow.com/a/5344074

Research Questions

%

How many SO snippets can we clone match against?

>

What thresholds should we compare with?

How many clones can we find in open source software?

Methodology

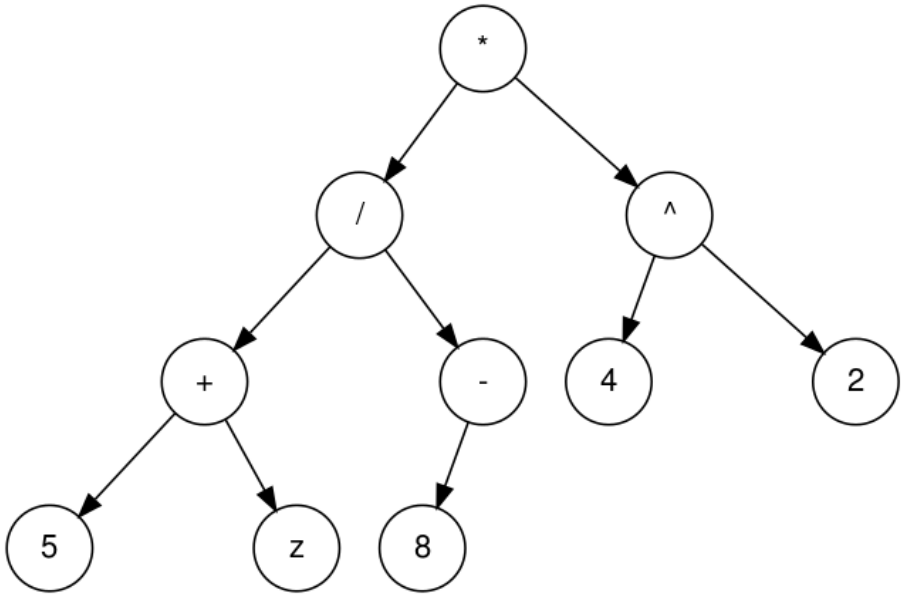
Methodology

Trying to sort the overflowing stack

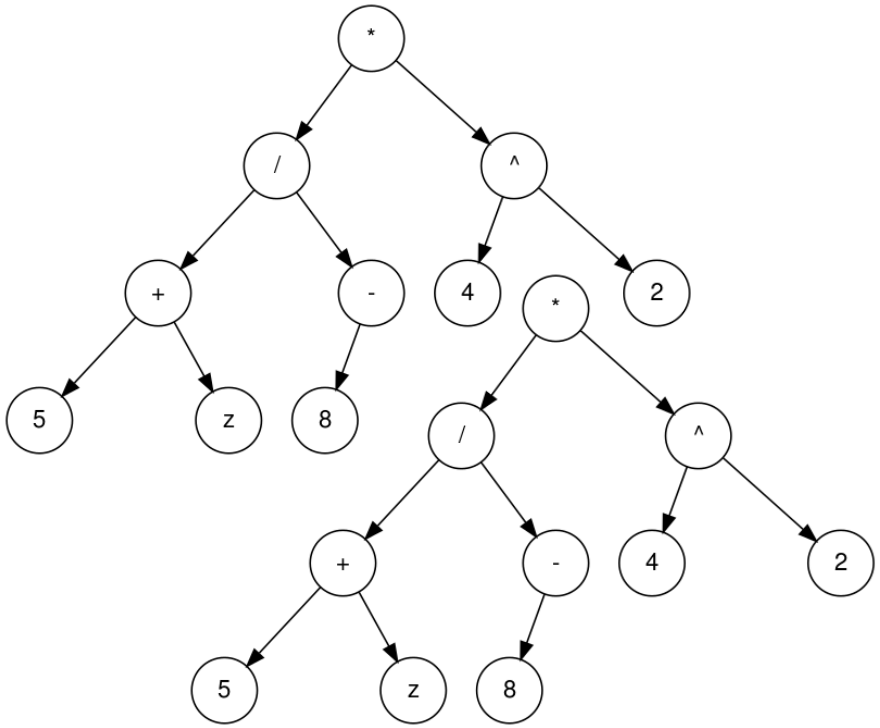
For clone detection, we used jsinspect, a tool by Daniel St. Jules

https://github.com/danielstjules/jsinspect

jsinspect is an AST comparison tool

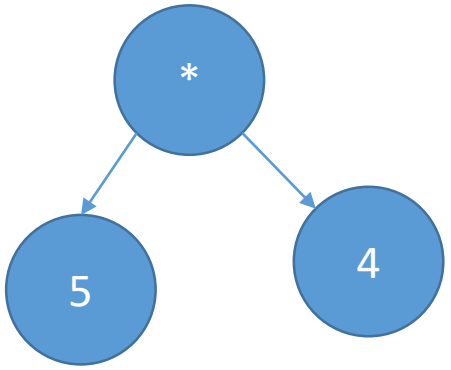


Creates Abstract Syntax Trees for Javascript code

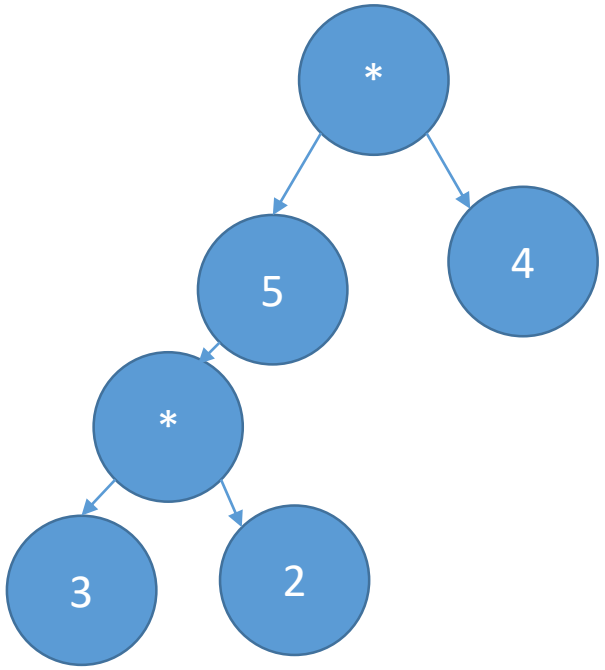


Compares the Abstract Syntax Trees to find clones

jsinspect has a threshold option and increasing it leads to more nodes compared



Small Threshold=Less Nodes Compared



Larger Threshold=More Nodes Compared

Comparing more nodes changes the results

Low Threshold

- More Matches
- More False Positives

High Threshold

- Less False Positives
- Less Matches

Used the Stackoverflow data dump

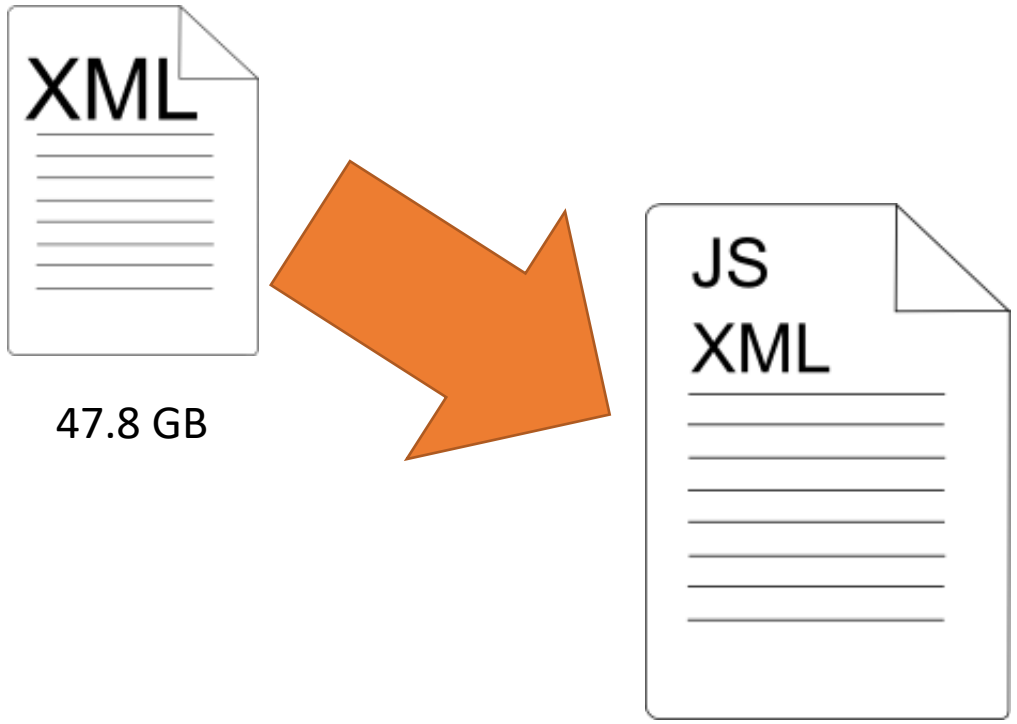
- Data up to December 16, 2016
- XML file, one line per post

```
<?xml version="1.0" encoding="utf-8"?>
<posts>
  <row Id="4" PostTypeId="1" AcceptedAnswerId="7" Cr
  <row Id="6" PostTypeId="1" AcceptedAnswerId="31" C
  <row Id="7" PostTypeId="2" ParentId="4" CreationDa
  <row Id="9" PostTypeId="1" AcceptedAnswerId="1404"
  <row Id="11" PostTypeId="1" AcceptedAnswerId="1248
  <row Id="12" PostTypeId="2" ParentId="11" Creation
  <row Id="13" PostTypeId="1" AcceptedAnswerId="357"
  <row Id="14" PostTypeId="1" CreationDate="2008-08-
```

Posts.xml

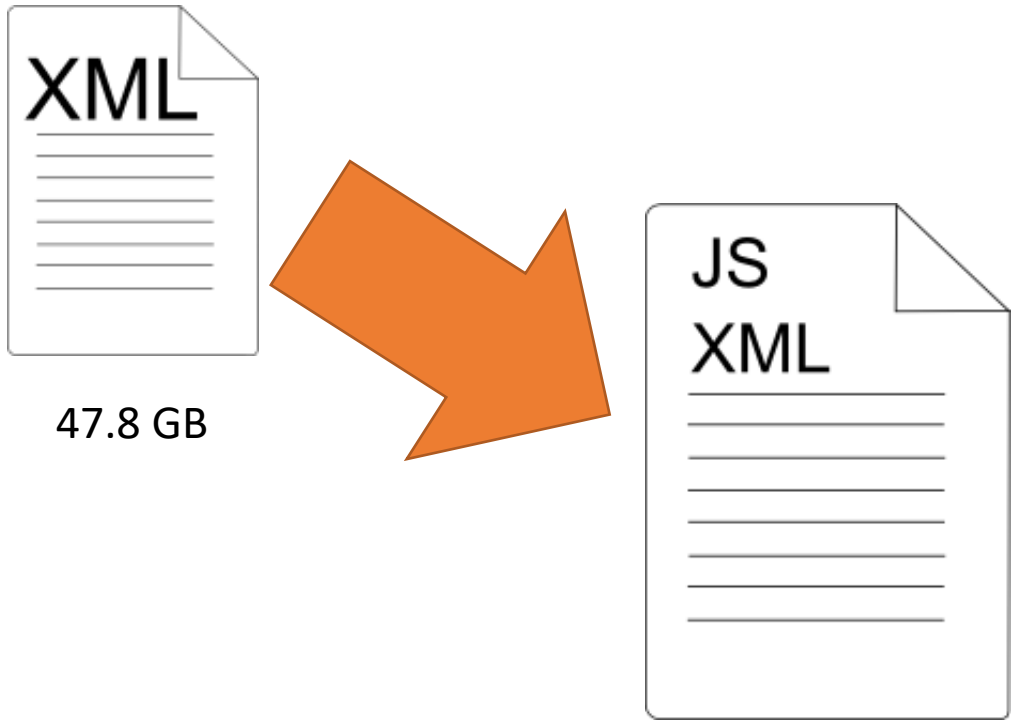
We need to reduce the SO dataset

- First Step:
  - Reduce just to posts tagged javascript



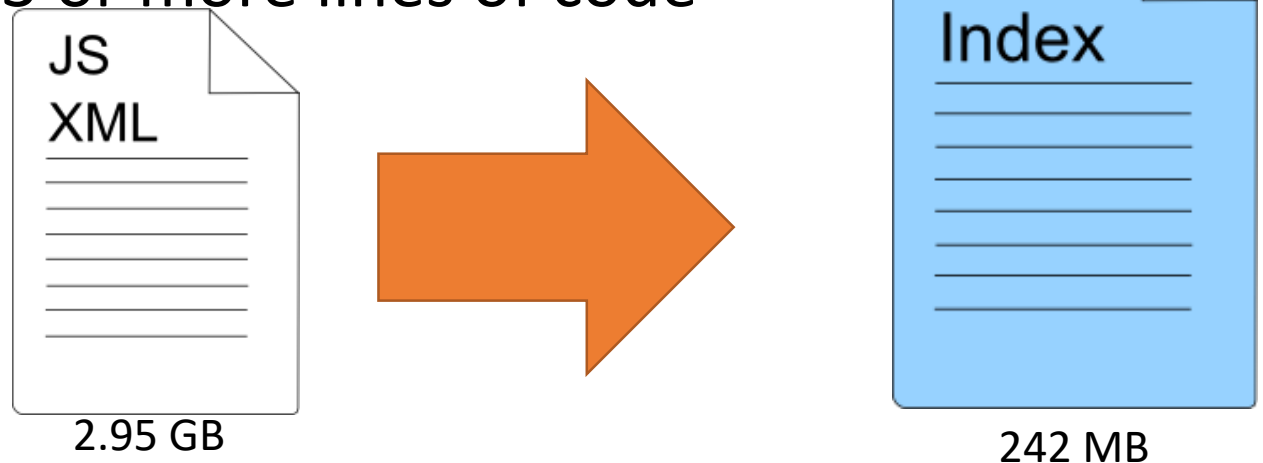
Our Data is still too big, so we apply more filters

- No posts tagged jQuery, html or css
- Accepted Answers only

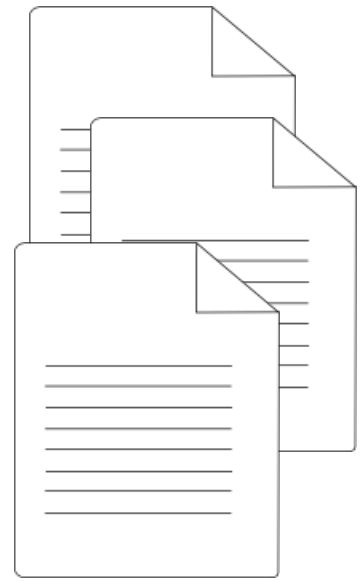


We need to reduce the SO dataset

- First Step:
  - Reduce just to Javascript Posts
- Second Step:
  - Reduce to just code snippets
  - Only use code blocks (no inline code)
  - Requirement: 3 or more lines of code



A single index was too big to directly compare against, so we split it into individual files.



```
1  //#{"OwnerUserId": "7441", "AcceptedAnswer": true,
2  function sayHello(name) {
3      var text = 'Hello ' + name;
4      var say = function() { console.log(text); }
5      say();
6  }
7  function sayHello2(name) {
8      var text = 'Hello ' + name; // Local variable
9      var say = function() { console.log(text); }
10     return say;
11 }
12 var say2 = sayHello2('Bob');
13 say2(); // logs "Hello Bob"
```

Part of a Snippet File

We filter out the unparsable snippets

- Use “new Function(snippet)” [1]
- Throws two kinds of errors:
  - SyntaxError
  - ReferenceError
- Jsinspect can still work on reference errors

[1]http://stackoverflow.com/a/15333480

After the parsing, we are left with 51.17% of the snippets

	Number of Snippets
SyntaxError	123330
ReferenceError	1078
No Error	128120
Total	252528

Table 1: Number of parsable snippets

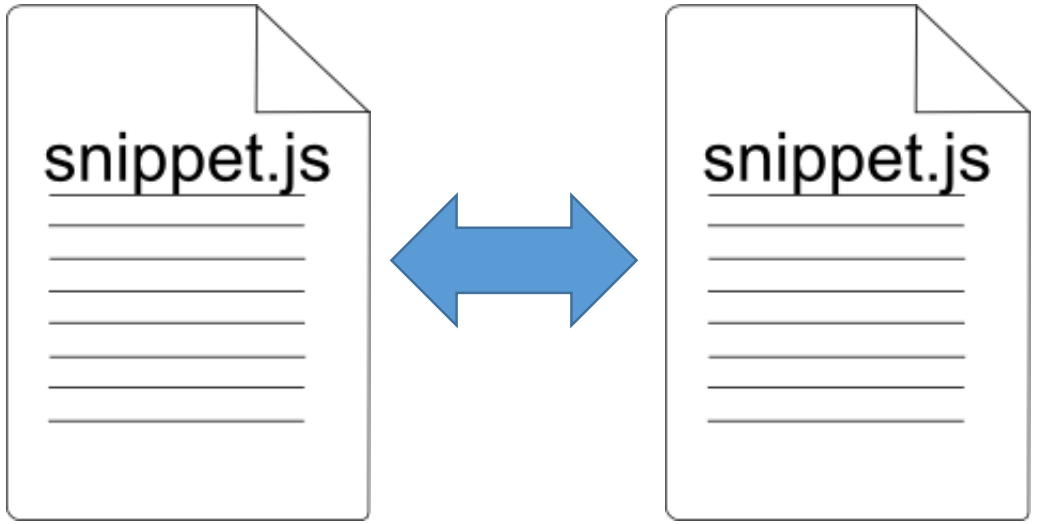
We get a total of 129198 snippets we can work with

But can we actually run clone detection on these snippets?

- And at what threshold?

Idea: Find a maximum possible threshold a snippet matches with itself

- Run jsinspect with the file compared to itself



Command: jsinspect snippet1.js snippet1.js

We can’t test all possible thresholds so try to group them

- Tried to group into thresholds: 110,90,70,50,30,20,15
- Any snippet that did not fit into those values, was classified as ‘below 15’

Jsinspect errored on 1257 snippets but ran on 103,237

1257 Errors

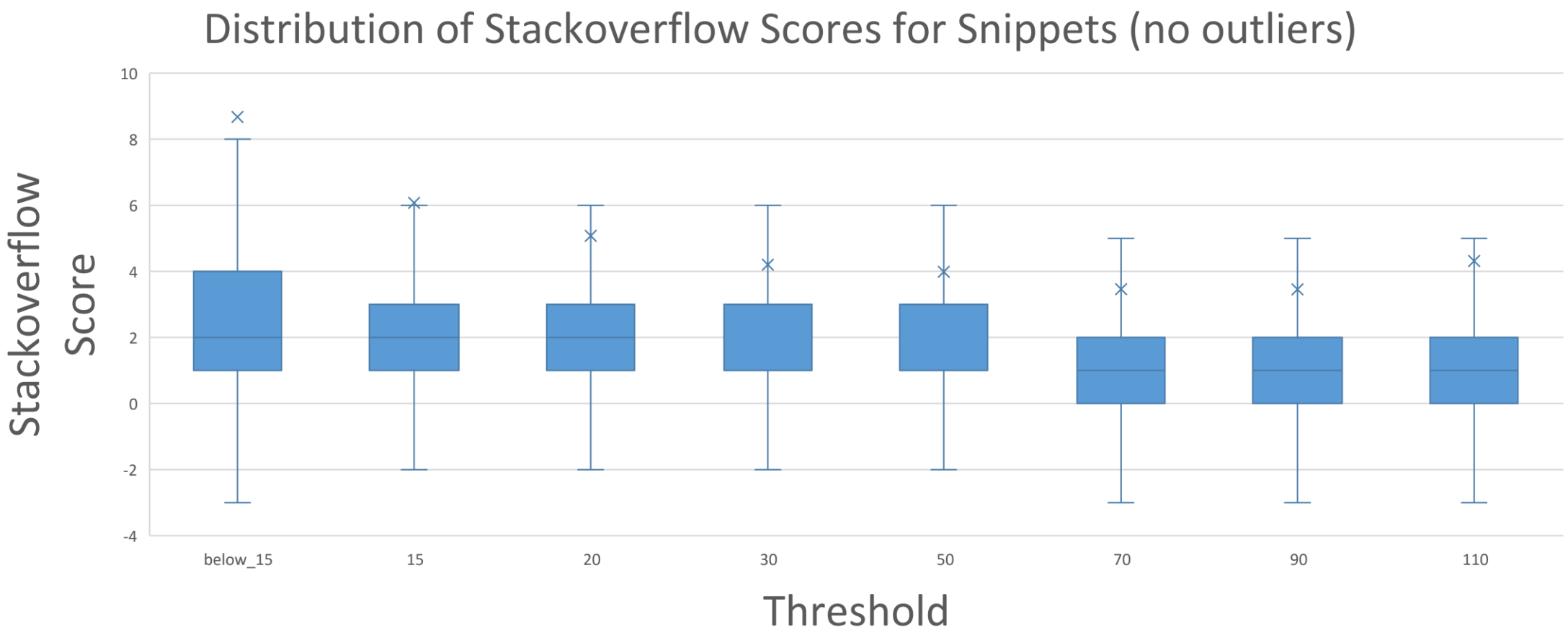
103,237 Ran

We were able to get thresholds for 40.8% of the snippets

Threshold	Number	Percentage(%) of successful snippets files (103,237)	Percentage(%) of total snippets (252,528)
110	7416	7.183471	2.936704
90	4358	4.221355	1.725749
70	7890	7.642609	3.124406
50	14992	14.52193	5.936767
30	29587	28.6593	11.71632
20	24487	23.71921	9.696746
15	14507	14.05213	5.744709
Below 15	24704	23.92941	9.782678
Total	103,237	100	40.88141

Table 2: Threshold splits for snippets

Lower thresholds tend to have higher votes



Research Questions 1 & 2 answered

- 40.8% of Stackoverflow accepted Answer javascript code snippets can be used for clone detection
- We can derive a maximum comparison threshold by self-comparing

Tried to find Stackoverflow Clones in two Open source projects

- Settings:
  - Threshold = 20
  - Identifiers = False
  - Literals = False
- Excluded files that could not self-match (at threshold 30)
- Looked at
  - Passport-Github
  - Mocha

simple, flexible, fun

Mocha, a Javascript test framework

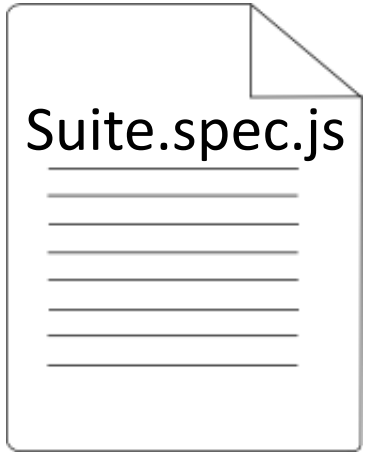
No clones found in Passport-Github

- Compared 21 files against Snippets in groups: 50,30,20

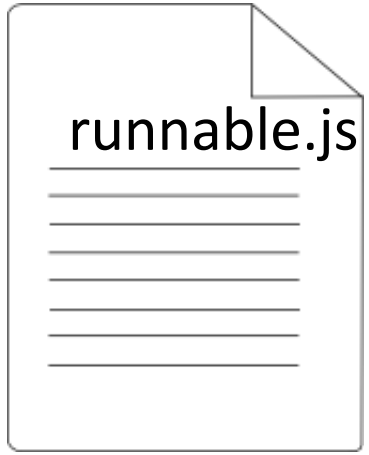
3 Mocha files had matches

- Looked at 132 JS files
- Compared to SO snippets with a threshold 50

9 matches



1 match



1 match

Example match from mocha.js

```
15542 exports._extend = function(origin, add) {
15543   // Don't do anything if add isn't an object
15544   if (!add || !isObject(add)) return origin;
15545
15546   var keys = Object.keys(add);
15547   var i = keys.length;
15548   while (i--) {
15549     origin[keys[i]] = add[keys[i]];
15550   }
15551   return origin;
15552 };
```

Mocha.js code

Example of a match in mocha.js

Source code of Node's `_extend` function is in here:
https://github.com/joyent/node/blob/master/lib/util.js

```
exports._extend = function(origin, add) {
  // Don't do anything if add isn't an object
  if (!add || typeof add !== 'object') return origin;

  var keys = Object.keys(add);
  var i = keys.length;
  while (i--) {
    origin[keys[i]] = add[keys[i]];
  }
  return origin;
};
```

share improve this answer

edited Jun 1 '14 at 8:16

answered Feb 23 '13 at 12:40

jimbojw

8,182 ● 5 ● 20 ● 36

Stackoverflow Post[1]

[1]http://stackoverflow.com/a/15040626

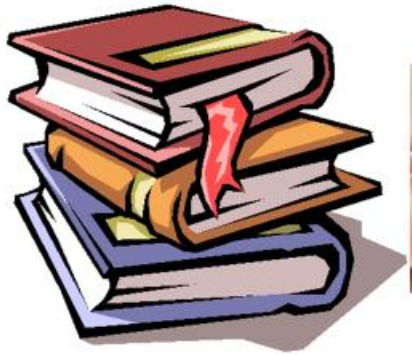
Future Work and Recommendations

- Look at larger snippet dataset
- Generate ASTs ahead of time for faster matching
- Augment with text similarity measures
  - Source Code Comments and Stackoverflow posts perhaps?

Summary

There are problems with not keeping track of code taken from the internet

Possible copyright issues



Missing Documentation

Photo: Steven DePolo[1]

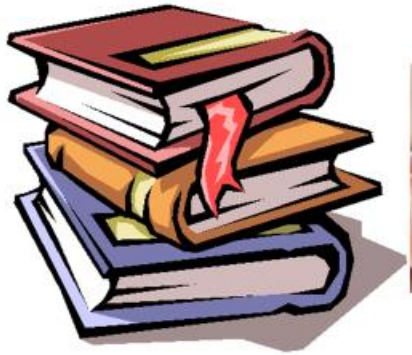
Aged Code

7

[1]https://www.flickr.com/photos/stevendepolo/4528758992

There are problems with not keeping track of code taken from the internet

Possible copyright issues



Missing Documentation

Photo: Steven DePolo[1]

Aged Code

[1]https://www.flickr.com/photos/stevendepolo/4528758992

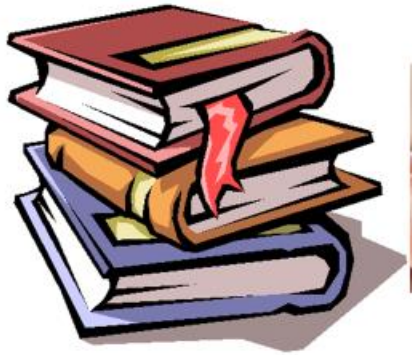
We were able to get thresholds for 40.8% of the snippets

Threshold	Number	Percentage(%) of successful snippets files (103,237)	Percentage(%) of total snippets (252,528)
110	7416	7.183471	2.936704
90	4358	4.221355	1.725749
70	7890	7.642609	3.124406
50	14992	14.52193	5.936767
30	29587	28.6593	11.71632
20	24487	23.71921	9.696746
15	14507	14.05213	5.744709
Below 15	24704	23.92941	9.782678
Total	103,237	100	40.88141

Table 2: Threshold splits for snippets

There are problems with not keeping track of code taken from the internet

Possible copyright issues



Missing Documentation

Photo: Steven DePolo[1]

Aged Code

[1]https://www.flickr.com/photos/stevendepolo/4528758992

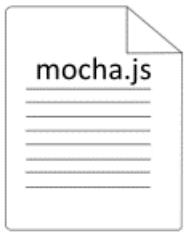
We were able to get thresholds for 40.8% of the snippets

Threshold	Number	Percentage(%) of successful snippets files (103,237)	Percentage(%) of total snippets (252,528)
110	7416	7.183471	2.936704
90	4358	4.221355	1.725749
70	7890	7.642609	3.124406
50	14992	14.52193	5.936767
30	29587	28.6593	11.71632
20	24487	23.71921	9.696746
15	14507	14.05213	5.744709
Below 15	24704	23.92941	9.782678
Total	103,237	100	40.88141

Table 2: Threshold splits for snippets

3 Mocha files had matches

- Looked at 125 JS files
- Compared to SO snippets with a threshold 50



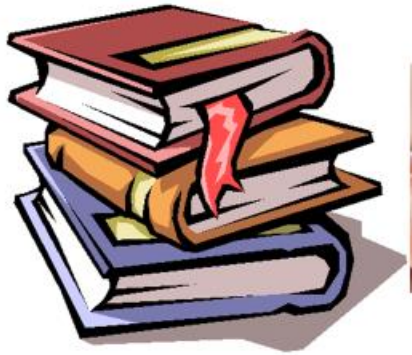
9 matches

1 match

1 match

There are problems with not keeping track of code taken from the internet

Possible copyright issues



Missing Documentation

Photo: Steven DePolo[1]

Aged Code

[1]https://www.flickr.com/photos/stevendepolo/4528758992

We were able to get thresholds for 40.8% of the snippets

Threshold	Number	Percentage(%) of successful snippets files (103,237)	Percentage(%) of total snippets (252,528)
110	7416	7.183471	2.936704
90	4358	4.221355	1.725749
70	7890	7.642609	3.124406
50	14992	14.52193	5.936767
30	29587	28.6593	11.71632
20	24487	23.71921	9.696746
15	14507	14.05213	5.744709
Below 15	24704	23.92941	9.782678
Total	103,237	100	40.88141

Table 2: Threshold splits for snippets

3 Mocha files had matches

- Looked at 125 JS files
- Compared to SO snippets with a threshold 50

9 matches

1 match

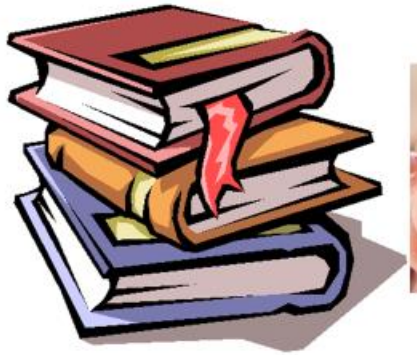
1 match

Future Work and Recommendations

- Look at larger snippet dataset
- Generate ASTs ahead of time for faster matching
- Augment with text similarity measures
  - Source Code Comments and Stackoverflow posts perhaps?

There are problems with not keeping track of code taken from the internet

Possible copyright issues



Missing Documentation

Photo: Steven DePolo[1]

Aged Code

[1]https://www.flickr.com/photos/stevendepolo/4528758992

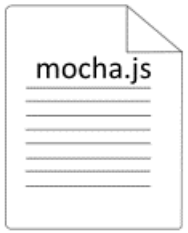
We were able to get thresholds for 40.8% of the snippets

Threshold	Number	Percentage(%) of successful snippets files (103,237)	Percentage(%) of total snippets (252,528)
110	7416	7.183471	2.936704
90	4358	4.221355	1.725749
70	7890	7.642609	3.124406
50	14992	14.52193	5.936767
30	29587	28.6593	11.71632
20	24487	23.71921	9.696746
15	14507	14.05213	5.744709
Below 15	24704	23.92941	9.782678
Total	103,237	100	40.88141

Table 2: Threshold splits for snippets

3 Mocha files had matches

- Looked at 125 JS files
- Compared to SO snippets with a threshold 50



9 matches

1 match

1 match

Future Work and Recommendations

- Look at larger snippet dataset
- Generate ASTs ahead of time for faster matching
- Augment with text similarity measures
  - Source Code Comments and Stackoverflow posts perhaps?

Zored Ahmer
zbahmer@cim.mcgill.ca

Extra Slides

Research Questions

Research Questions

- RQ1: How many StackOverflow Javascript snippets can be read by clone detection software?

Research Questions

- RQ1: How many StackOverflow Javascript snippets can be read by clone detection software?
- RQ2: At what thresholds should those snippets be compared against?

Research Questions

- RQ1: How many StackOverflow Javascript snippets can be read by clone detection software?
- RQ2: At what thresholds should those snippets be compared against?
- RQ3: How many files in OpenSource software (Passport-Github and Mocha) can be clone matched to SO snippets?

Limitations and Problems

Only looked at accepted answers

Not much open source code matched against

Jsinspect has a few problems