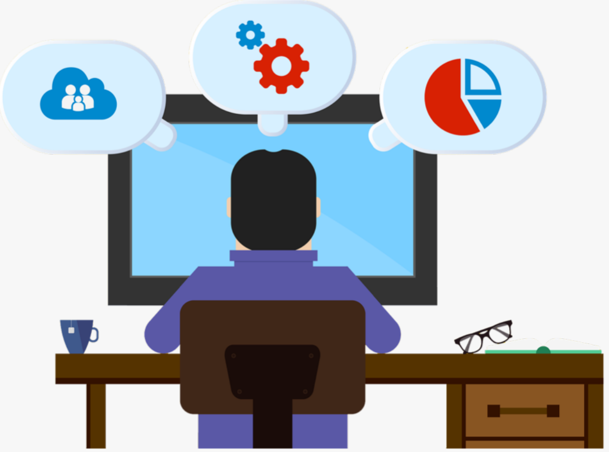


Software Developers' Work Habits and Expertise

Sebastian Baltes

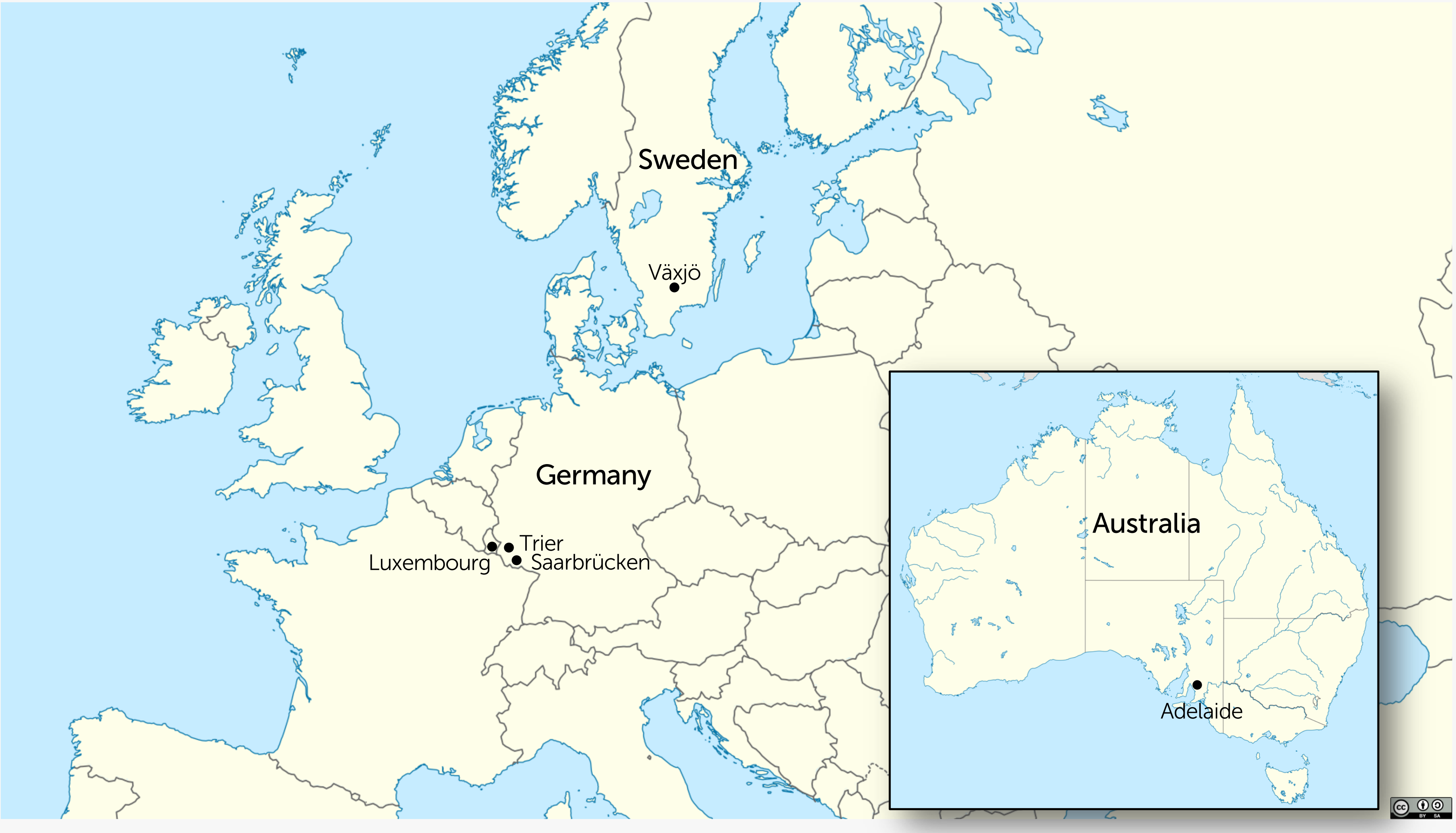
@s_baltes

empirical-software.engineering

THE UNIVERSITY *of* ADELAIDE

Interaction

My Background



Evidence-based Practice through Practice-based Evidence

Studying Developers’ Work Habits

Observe
Describe
Explain

Software Developers’ Work Habits ⇒ Expand knowledge

- → Derive requirements for better tool support
- → Identify possible process improvements
- → Communicate findings back to practitioners

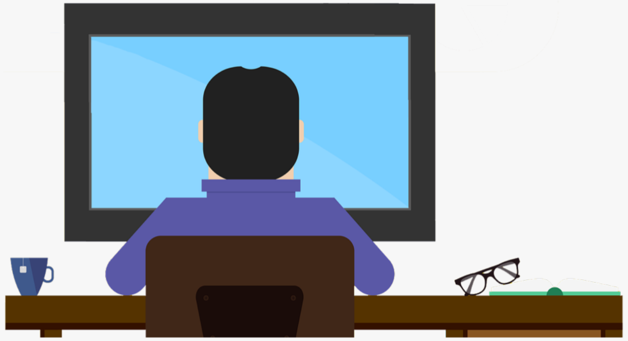
Habits?

A habit is a „**settled tendency** or **usual manner of behavior**“

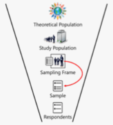
https://www.merriam-webster.com/dictionary/habit

Personal habits

Work habits



Studied Habits



Issues in Sampling Software Developers

Methodology

airbnb

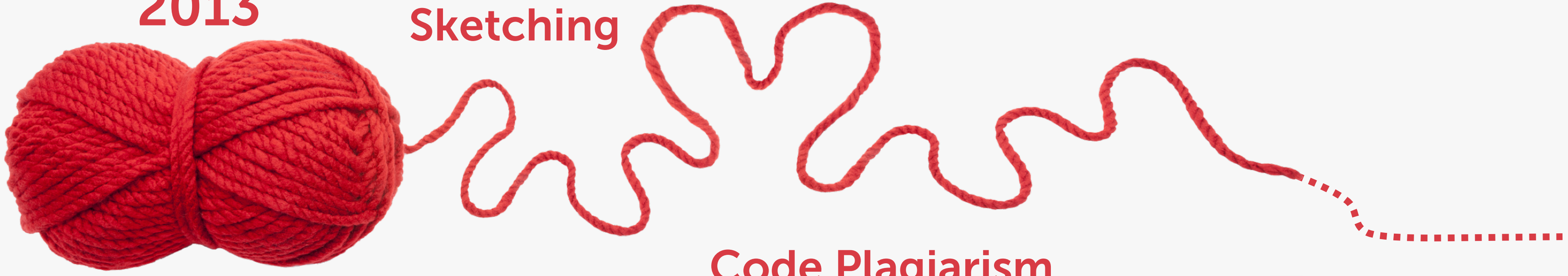
Constructing Urban Tourism Space Digitally

Interdisciplinary Research

2013

Sketching

Expertise Development

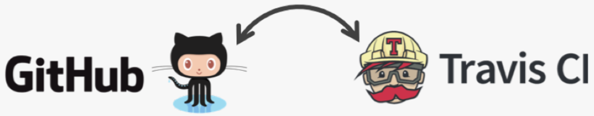


Code Plagiarism

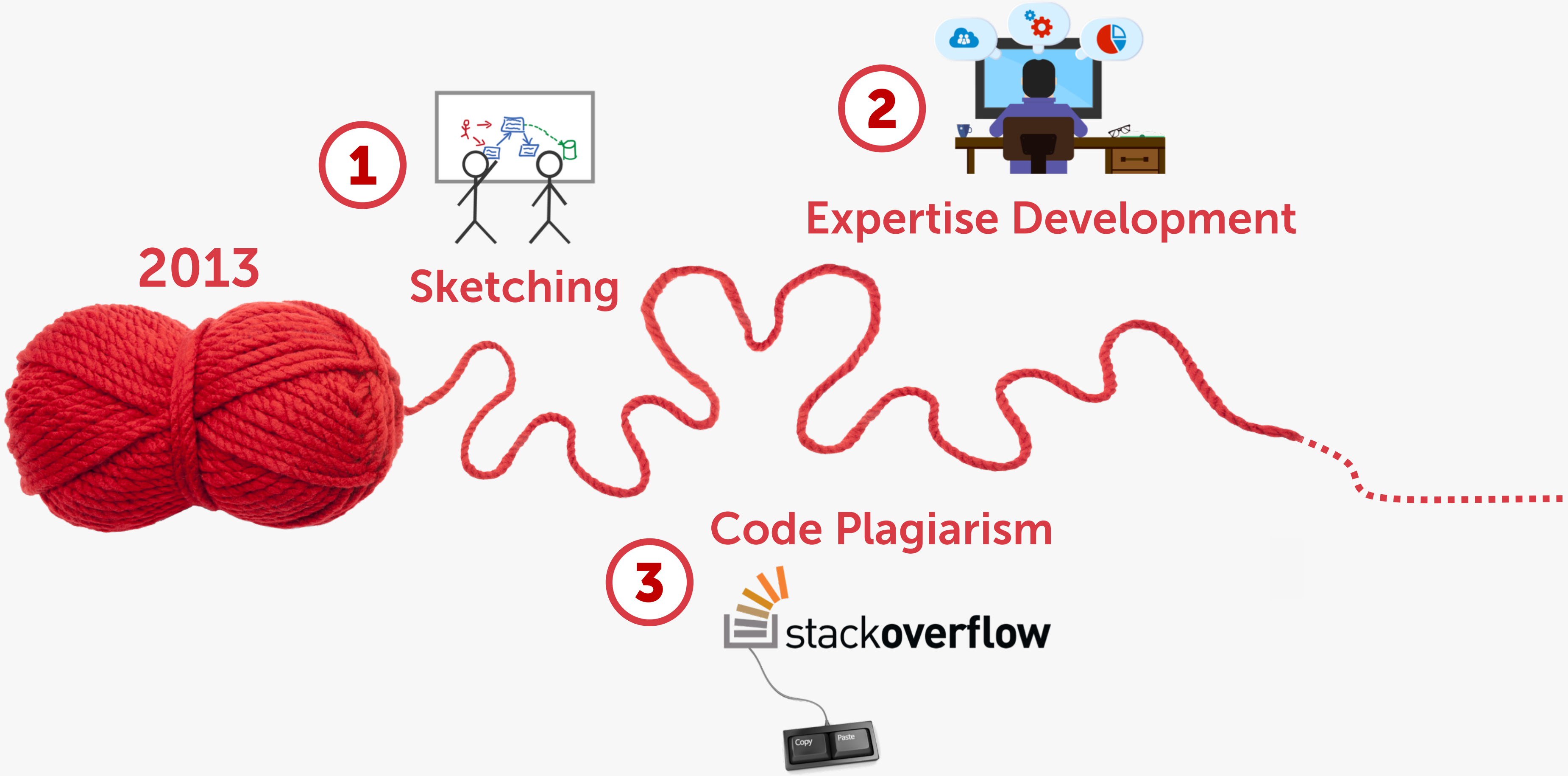
Regular Expressions

RegViz

Continuous Integration



Overview of this Talk



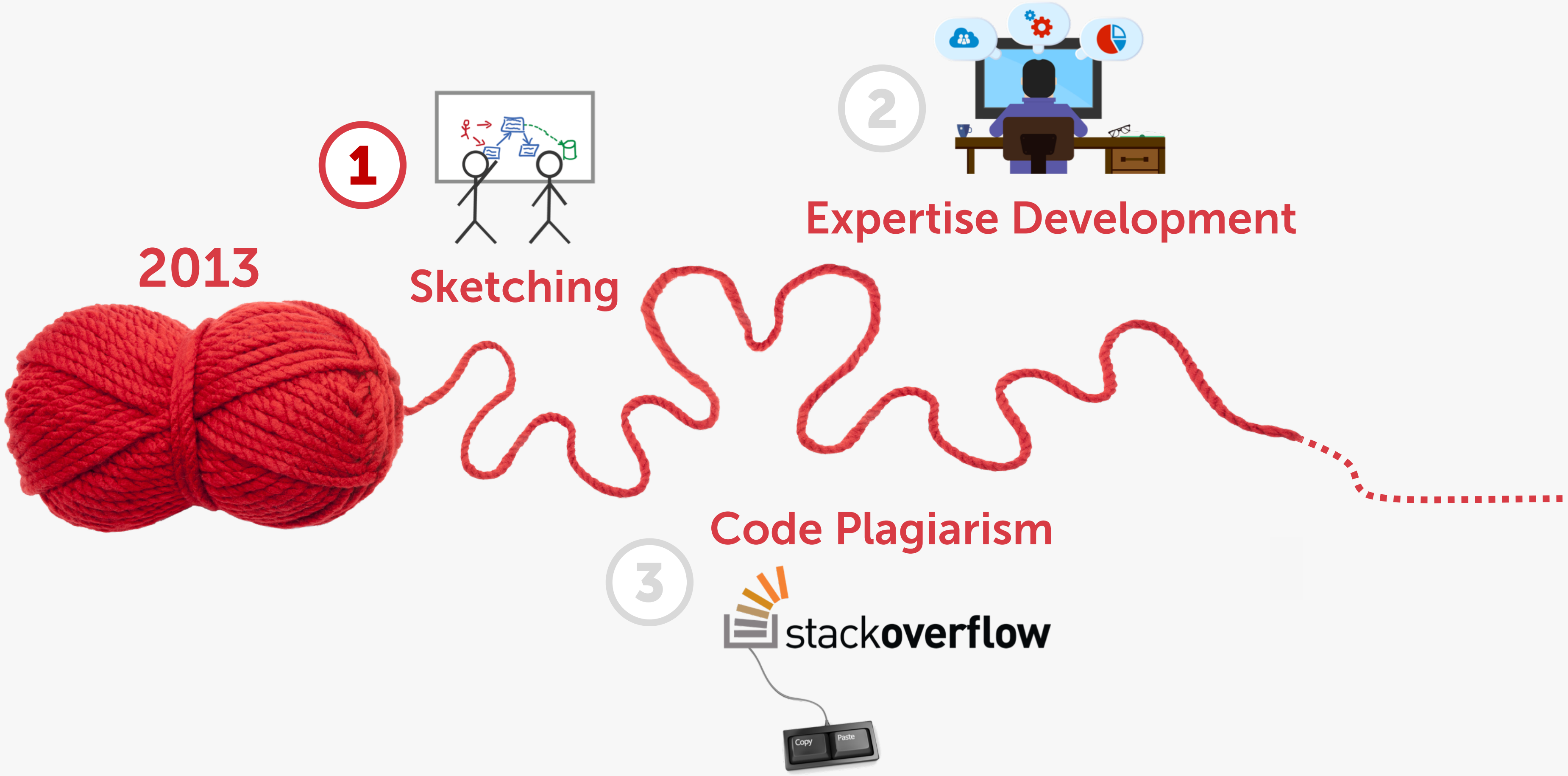
2013

1 Sketching

2 Expertise Development

3 Code Plagiarism

Overview of this Talk

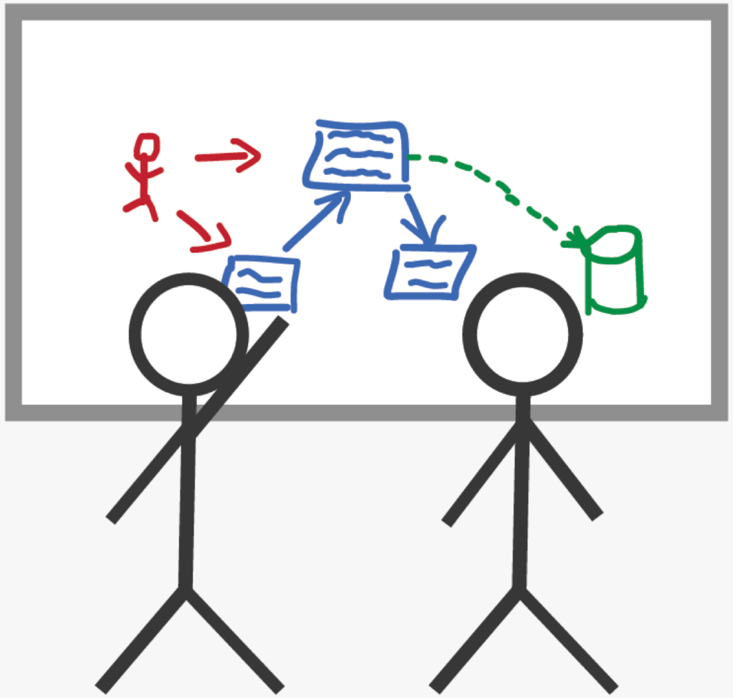


2013

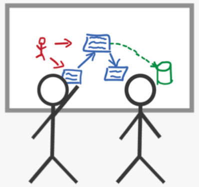
1 Sketching

2 Expertise Development

3 Code Plagiarism



Sketching



Research Questions

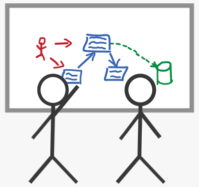
Questions:

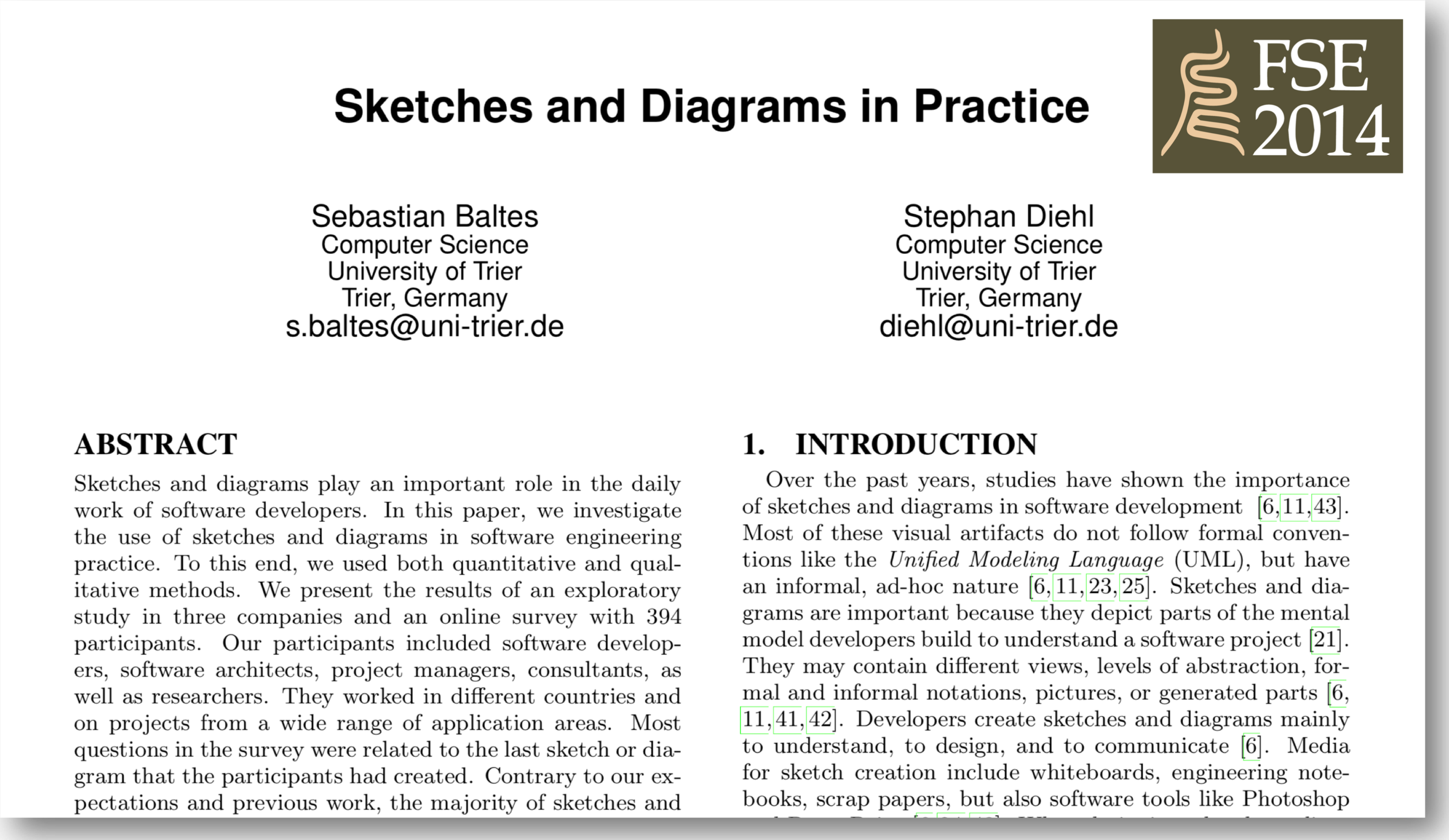
How and **why** do software practitioners use sketches and diagrams?
How are they related to **source code**?
How can we provide better **tool support**?

Approach:

Field study, online survey, lab study, formative tool evaluations

Sketching





Sketches and Diagrams in Practice

Sebastian Baltes
Computer Science
University of Trier
Trier, Germany
s.baltes@uni-trier.de

Stephan Diehl
Computer Science
University of Trier
Trier, Germany
diehl@uni-trier.de

ABSTRACT

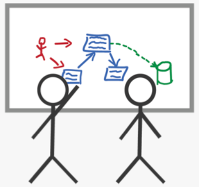
Sketches and diagrams play an important role in the daily work of software developers. In this paper, we investigate the use of sketches and diagrams in software engineering practice. To this end, we used both quantitative and qualitative methods. We present the results of an exploratory study in three companies and an online survey with 394 participants. Our participants included software developers, software architects, project managers, consultants, as well as researchers. They worked in different countries and on projects from a wide range of application areas. Most questions in the survey were related to the last sketch or diagram that the participants had created. Contrary to our expectations and previous work, the majority of sketches and

1. INTRODUCTION

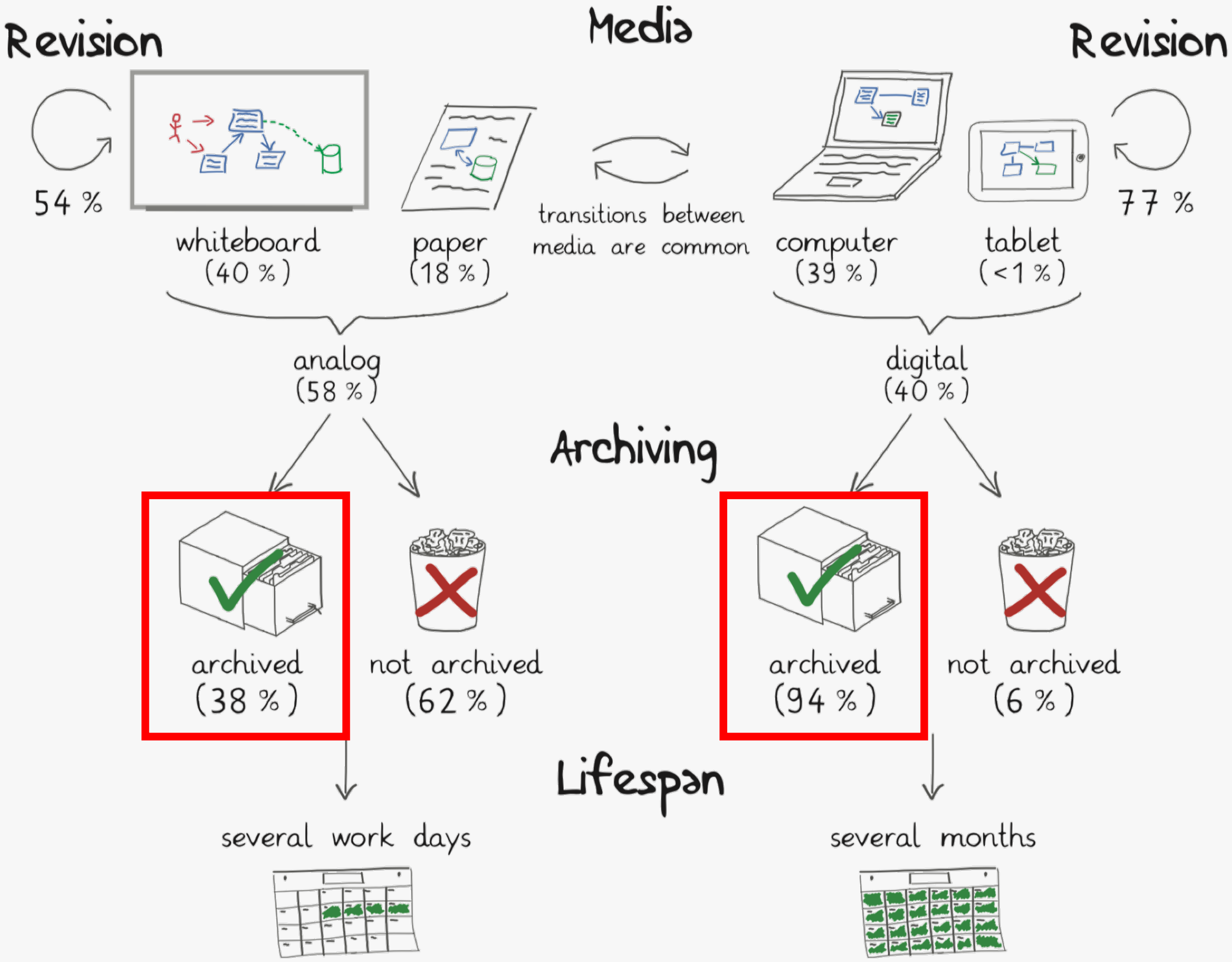
Over the past years, studies have shown the importance of sketches and diagrams in software development [6,11,43]. Most of these visual artifacts do not follow formal conventions like the *Unified Modeling Language* (UML), but have an informal, ad-hoc nature [6,11,23,25]. Sketches and diagrams are important because they depict parts of the mental model developers build to understand a software project [21]. They may contain different views, levels of abstraction, formal and informal notations, pictures, or generated parts [6,11,41,42]. Developers create sketches and diagrams mainly to understand, to design, and to communicate [6]. Media for sketch creation include whiteboards, engineering notebooks, scrap papers, but also software tools like Photoshop

https://empirical-software.engineering/projects/sketches/

Sketching



Sketches and Diagrams in Practice



Revision

54 %

Media

whiteboard (40 %)

paper (18 %)

transitions between media are common

computer (39 %)

tablet (< 1 %)

Revision

77 %

analog (58 %)

digital (40 %)

Archiving

archived (38 %)

not archived (62 %)

archived (94 %)

not archived (6 %)

Lifespan

several work days

several months

Purpose

Designing (75 %) 1

Explaining (60 %) 2

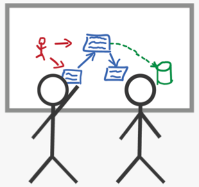
Understanding (56 %) 3

Analyzing Requirements (45 %)

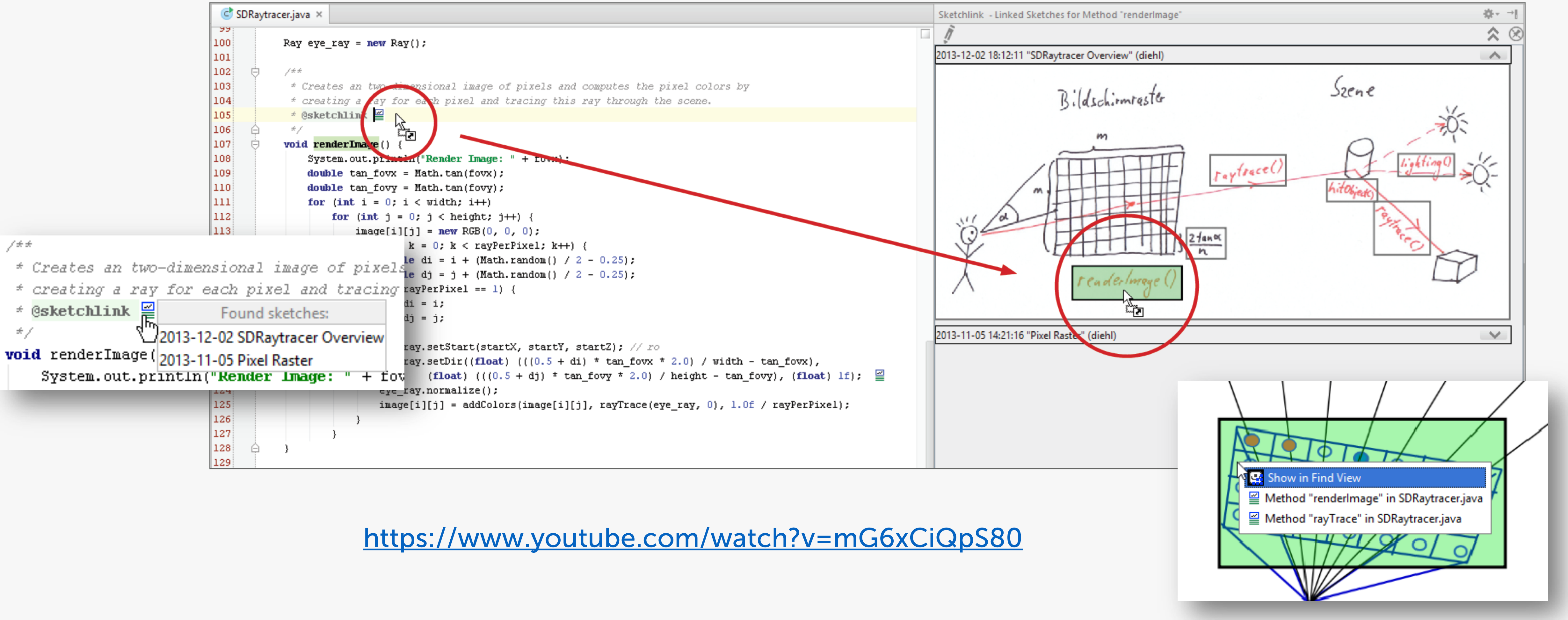
Relation to Source Code

47 % of the sketches are rated as helpful for others to understand the related source code artifacts.

Sketching



SketchLink



SDRaytracer.java

```
        Ray eye_ray = new Ray();

        /**
         * Creates an two-dimensional image of pixels and computes the pixel colors by
         * creating a ray for each pixel and tracing this ray through the scene.
         * @sketchlink
         */
        void renderImage() {
            System.out.println("Render Image: " + fovx);
            double tan_fovx = Math.tan(fovx);
            double tan_fovy = Math.tan(fovy);
            for (int i = 0; i < width; i++)
                for (int j = 0; j < height; j++) {
                    image[i][j] = new RGB(0, 0, 0);
                    k = 0; k < rayPerPixel; k++) {
                    le di = i + (Math.random() / 2 - 0.25);
                    le dj = j + (Math.random() / 2 - 0.25);
                    rayPerPixel == 1) {
                    di = i;
                    dj = j;

                    ray.setStart(startX, startY, startZ); // ro
                    ray.setDir((float) (((0.5 + di) * tan_fovx * 2.0) / width - tan_fovx),
                        (float) (((0.5 + dj) * tan_fovy * 2.0) / height - tan_fovy), (float) 1f);
                    eye_ray.normalize();
                    image[i][j] = addColors(image[i][j], rayTrace(eye_ray, 0), 1.0f / rayPerPixel);
                }
            }
        }
```

```
/**
 * Creates an two-dimensional image of pixels
 * creating a ray for each pixel and tracing
 * @sketchlink
 */
void renderImage(
    System.out.println("Render Image: " + fov
```

Found sketches:

2013-12-02 SDRaytracer Overview

2013-11-05 Pixel Raster

Sketchlink - Linked Sketches for Method "renderImage"

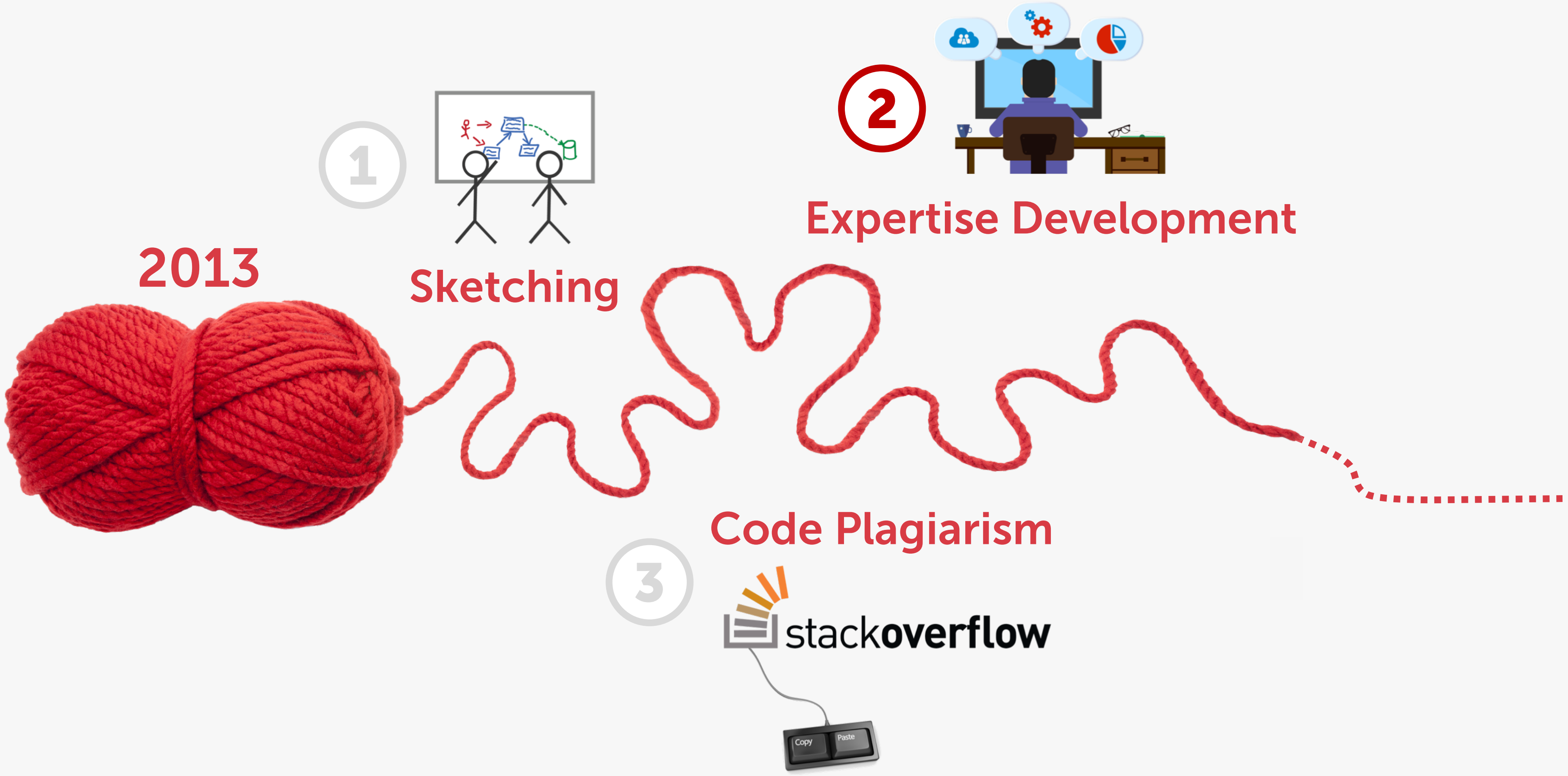
2013-12-02 18:12:11 "SDRaytracer Overview" (diehl)

2013-11-05 14:21:16 "Pixel Raster" (diehl)

Show in Find View
Method "renderImage" in SDRaytracer.java
Method "rayTrace" in SDRaytracer.java

https://www.youtube.com/watch?v=mG6xCiQpS80

Overview of this Talk

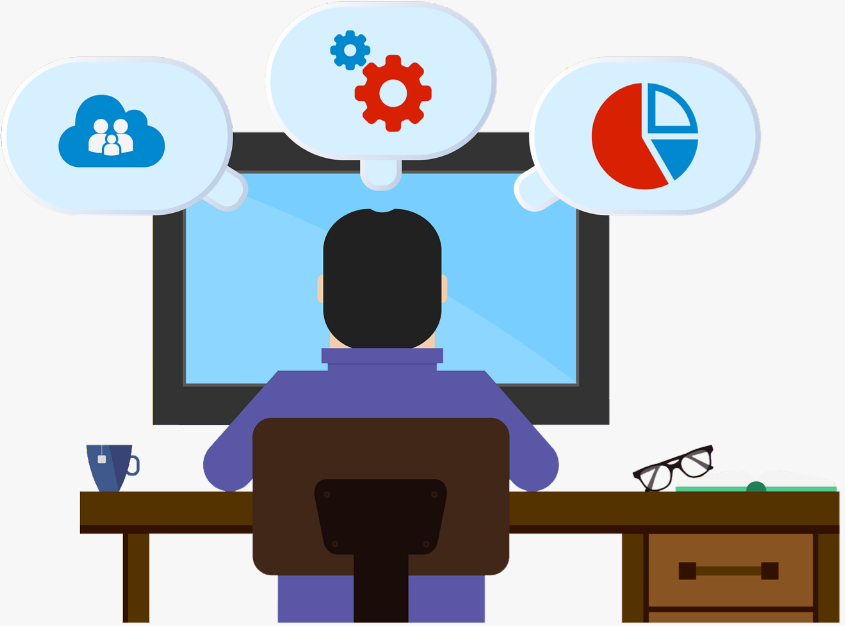


2013

1 Sketching

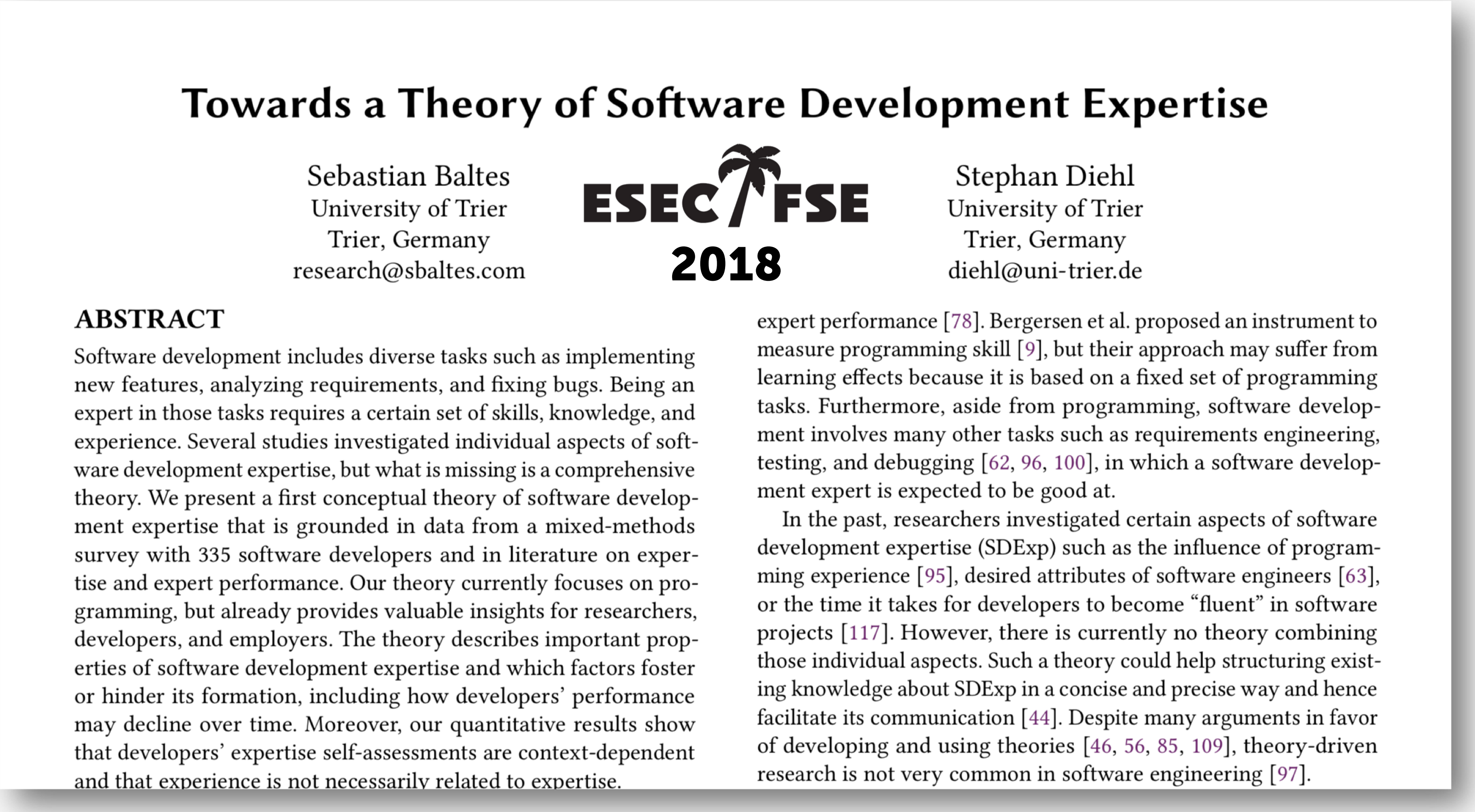
2 Expertise Development

3 Code Plagiarism



Expertise Development

Expertise Development



Towards a Theory of Software Development Expertise

Sebastian Baltes
University of Trier
Trier, Germany
research@sbaltes.com

Stephan Diehl
University of Trier
Trier, Germany
diehl@uni-trier.de

ESEC/FSE 2018

ABSTRACT

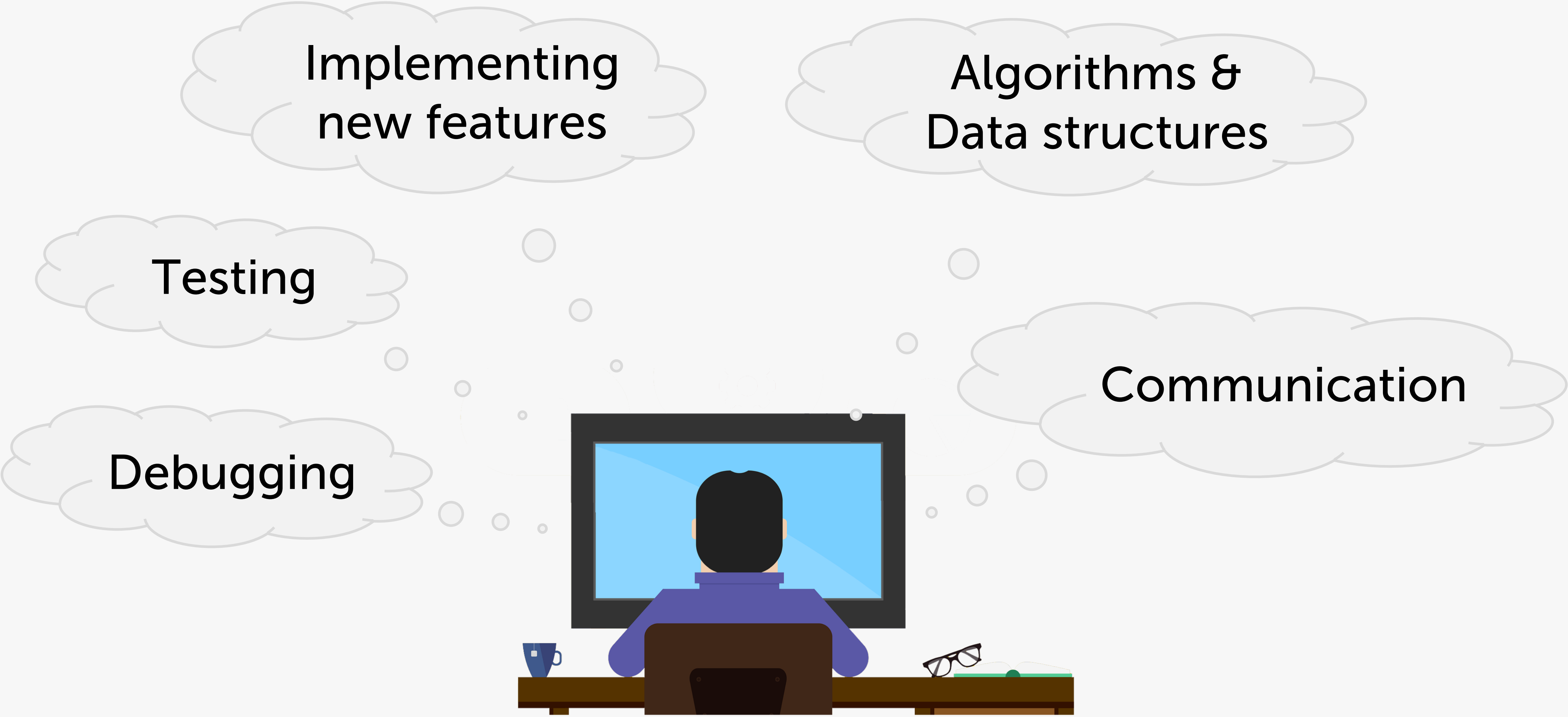
Software development includes diverse tasks such as implementing new features, analyzing requirements, and fixing bugs. Being an expert in those tasks requires a certain set of skills, knowledge, and experience. Several studies investigated individual aspects of software development expertise, but what is missing is a comprehensive theory. We present a first conceptual theory of software development expertise that is grounded in data from a mixed-methods survey with 335 software developers and in literature on expertise and expert performance. Our theory currently focuses on programming, but already provides valuable insights for researchers, developers, and employers. The theory describes important properties of software development expertise and which factors foster or hinder its formation, including how developers' performance may decline over time. Moreover, our quantitative results show that developers' expertise self-assessments are context-dependent and that experience is not necessarily related to expertise.

expert performance [78]. Bergersen et al. proposed an instrument to measure programming skill [9], but their approach may suffer from learning effects because it is based on a fixed set of programming tasks. Furthermore, aside from programming, software development involves many other tasks such as requirements engineering, testing, and debugging [62, 96, 100], in which a software development expert is expected to be good at.

In the past, researchers investigated certain aspects of software development expertise (SDExp) such as the influence of programming experience [95], desired attributes of software engineers [63], or the time it takes for developers to become "fluent" in software projects [117]. However, there is currently no theory combining those individual aspects. Such a theory could help structuring existing knowledge about SDExp in a concise and precise way and hence facilitate its communication [44]. Despite many arguments in favor of developing and using theories [46, 56, 85, 109], theory-driven research is not very common in software engineering [97].

https://empirical-software.engineering/projects/expertise/

Software Development Expertise?



Implementing new features

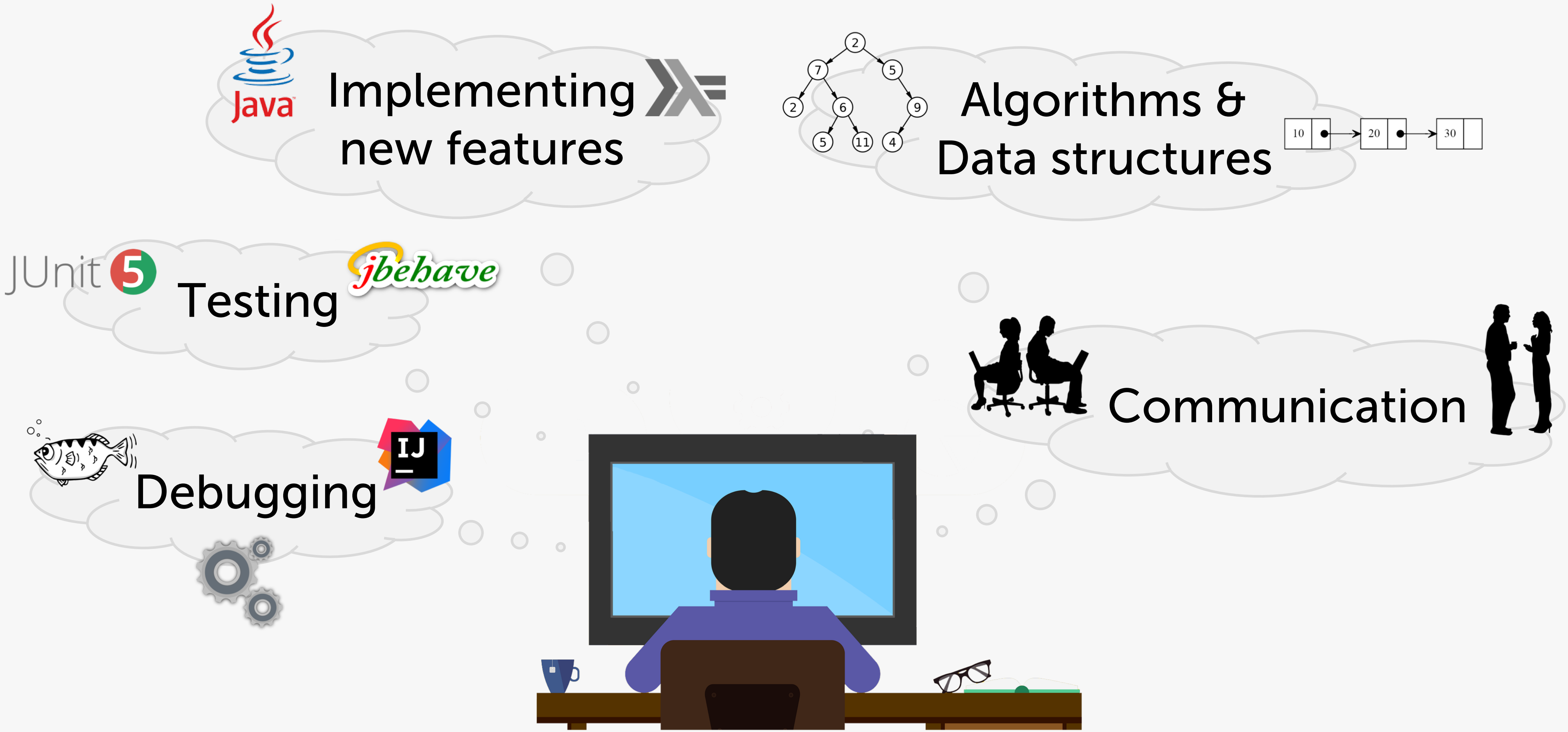
Algorithms & Data structures

Testing

Communication

Debugging

Software Development Expertise?



Implementing new features

Algorithms & Data structures

Testing

Debugging

Communication

Expertise Development

Research Questions

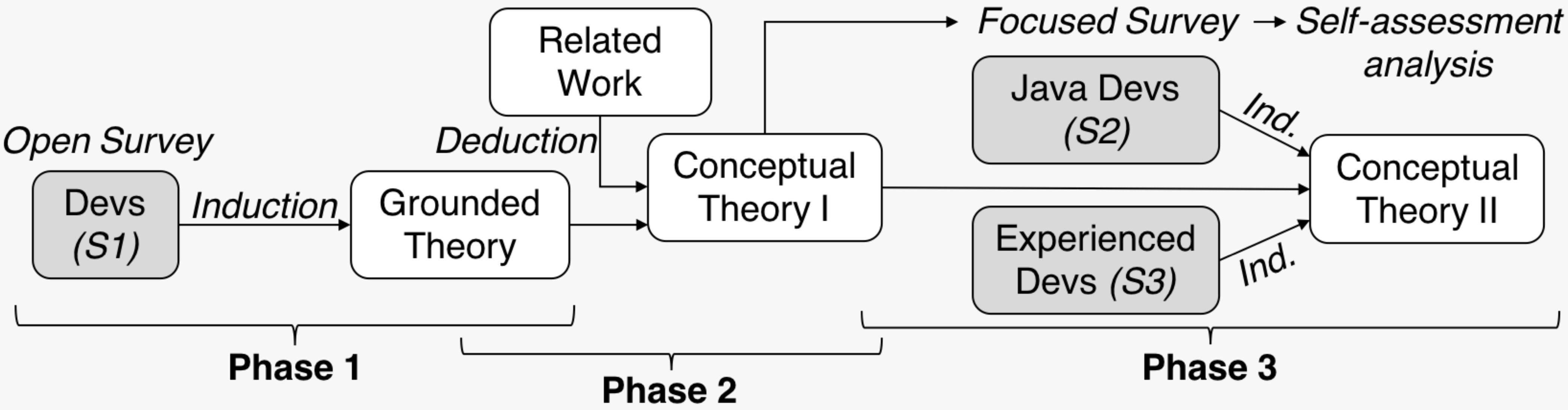
Questions:

How to **structure** all those expertise-related aspects? Which factors influence **expertise development** over time?

Approach:

Iterative theory building

Research Design



Open Survey

Devs *(S1)*

Induction

Grounded Theory

Related Work

Deduction

Conceptual Theory I

Focused Survey → *Self-assessment analysis*

Java Devs *(S2)*

Experienced Devs *(S3)*

Ind.

Ind.

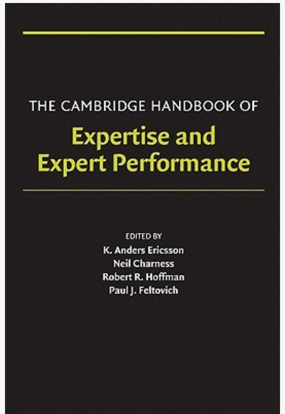
Conceptual Theory II

Phase 1

Phase 2

Phase 3

- **Induction:** 335 online survey participants in total
- **Deduction:** Main source *"Cambridge Handbook of Expertise and Expert Performance"*

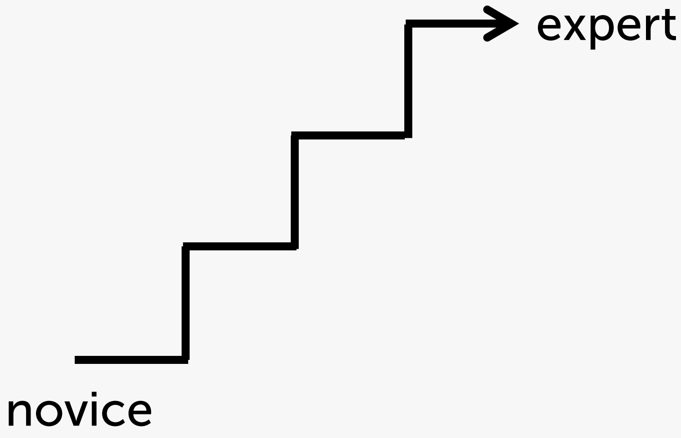


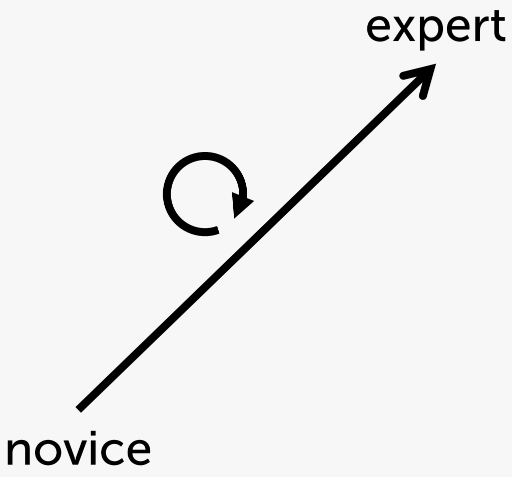
THE CAMBRIDGE HANDBOOK OF
Expertise and
Expert Performance

EDITED BY
K. Anders Ericsson
Neil Charness
Robert R. Hoffman
Paul J. Feltovich

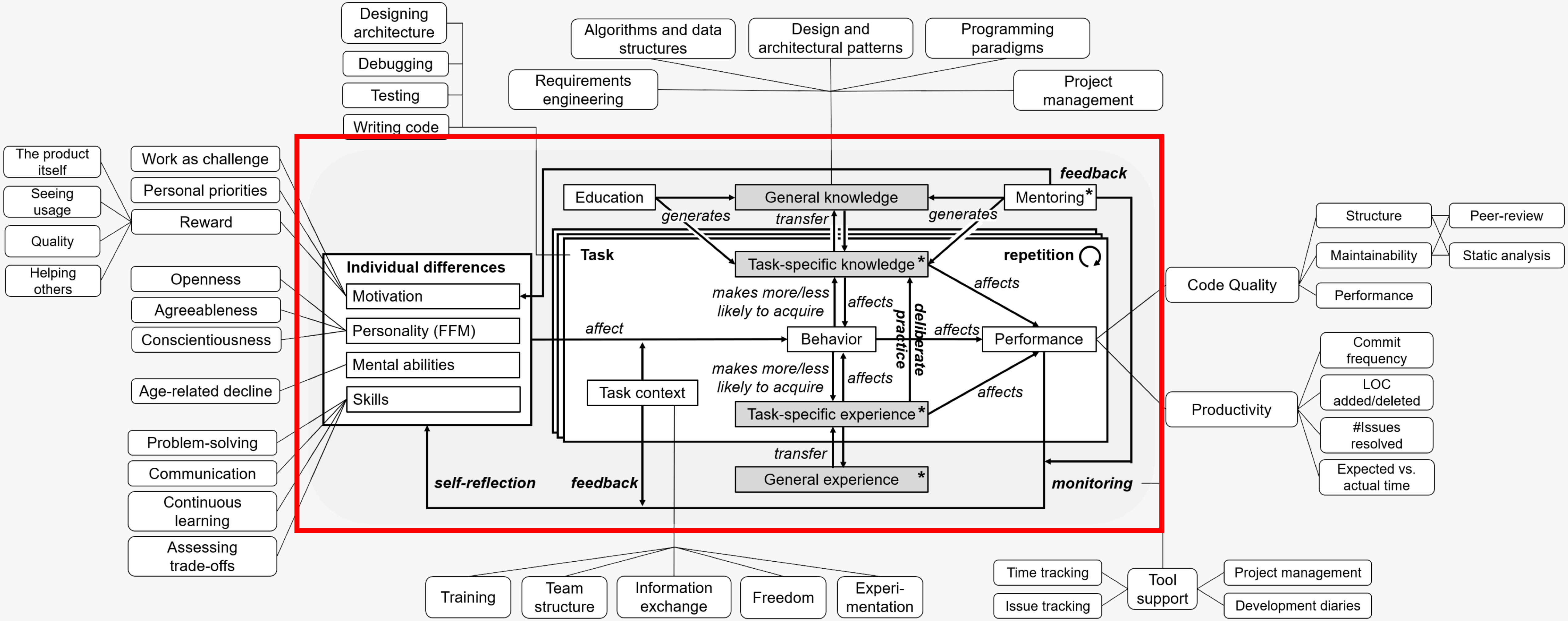
Our Expertise Model

- **Task-specific** (e.g., writing code, debugging, testing)
- Focuses on **individual developers**
- **Process** view (repetition of tasks)
- Notion of **transferable knowledge and experience** from related fields or tasks
- **Continuum** instead of discrete expertise steps

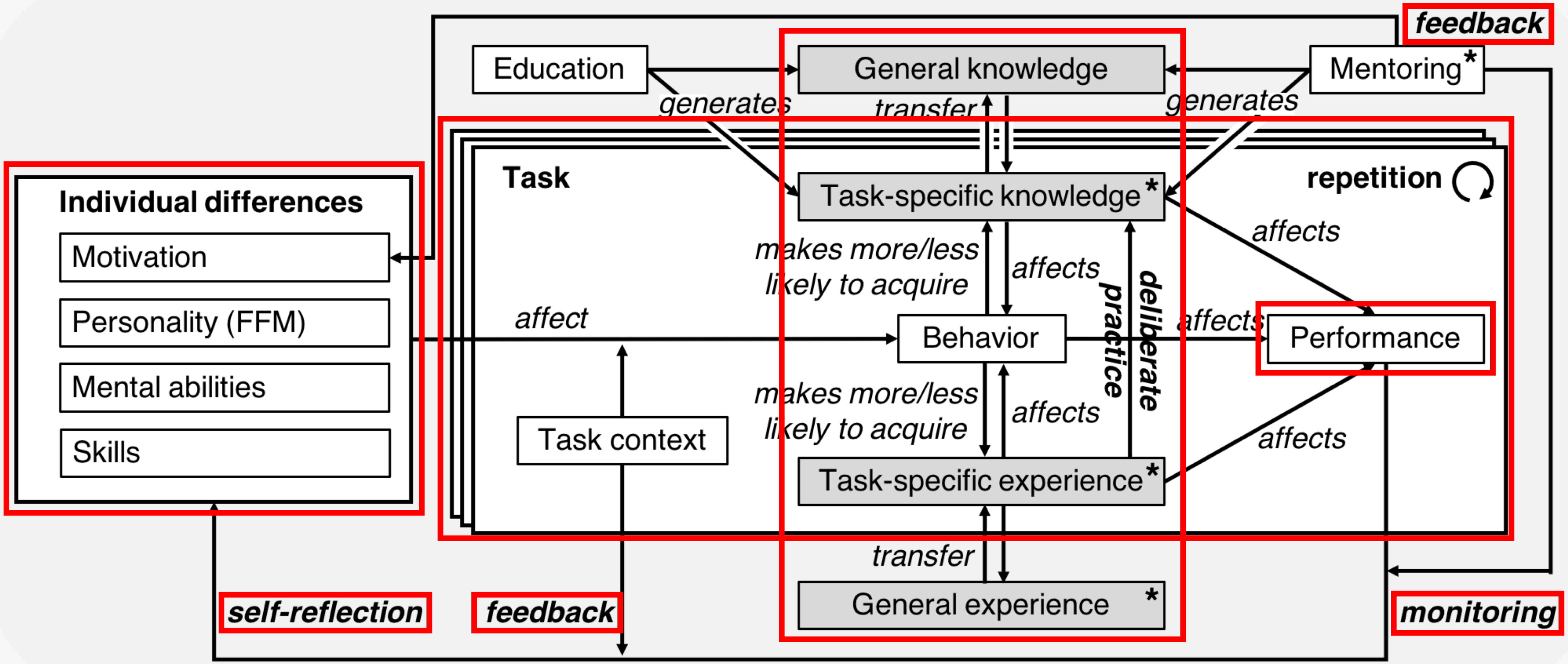




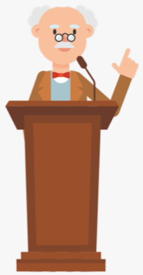
Conceptual Theory



Conceptual Theory



Summary



Researchers can...

- Use our theory to **design studies** on expertise development
- Adopt our **theory building** approach

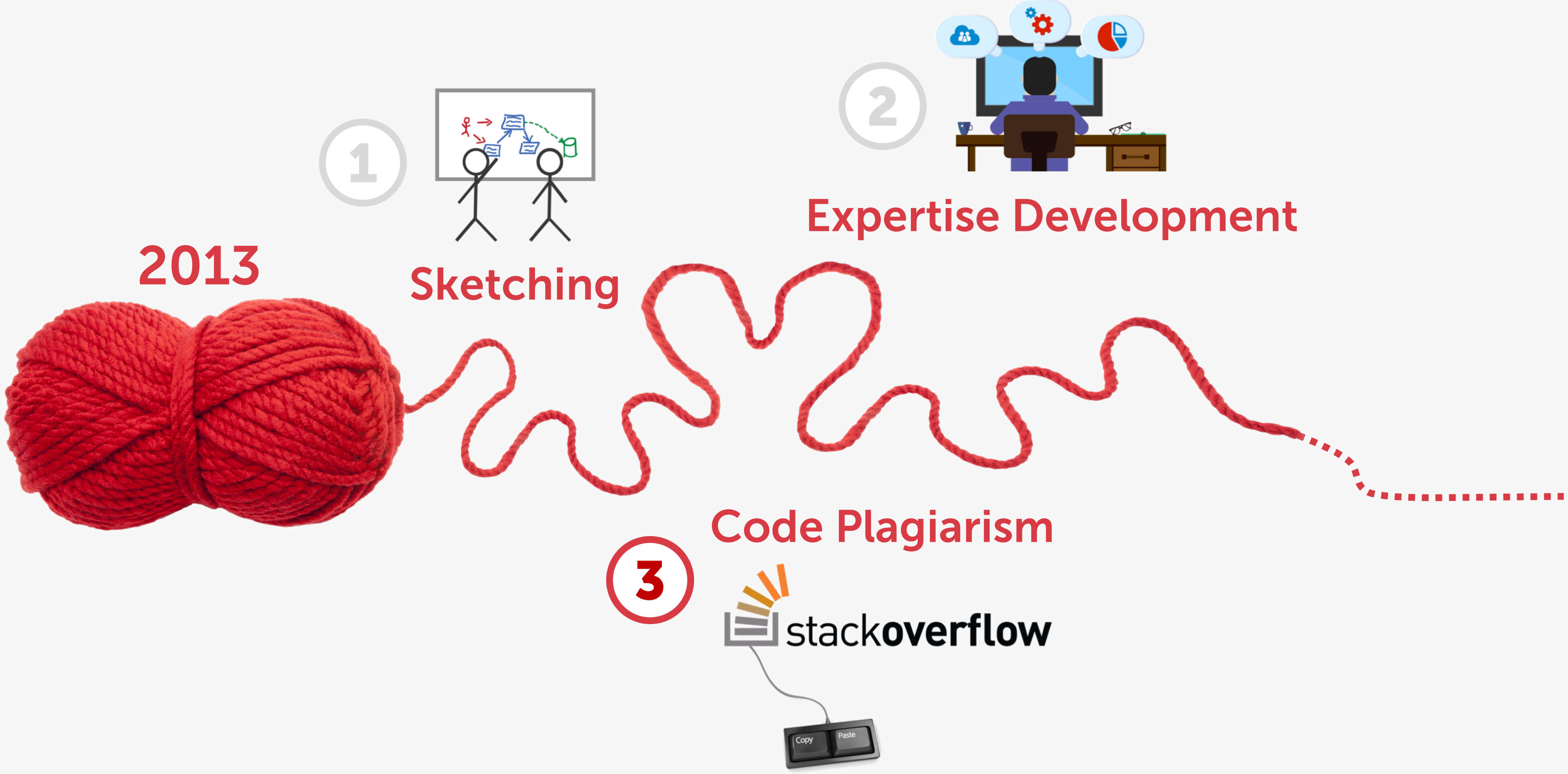
Developers can...

- Learn what other developers expect from **experts/mentors**
- Learn which **behaviors** may lead to becoming an expert

Employers can...

- Learn what **(de)motivates** employees and thus fosters or hinders expertise development
- Reflect on ideas to build a work environment **supporting self-improvement** of their staff

Overview of this Talk



2013

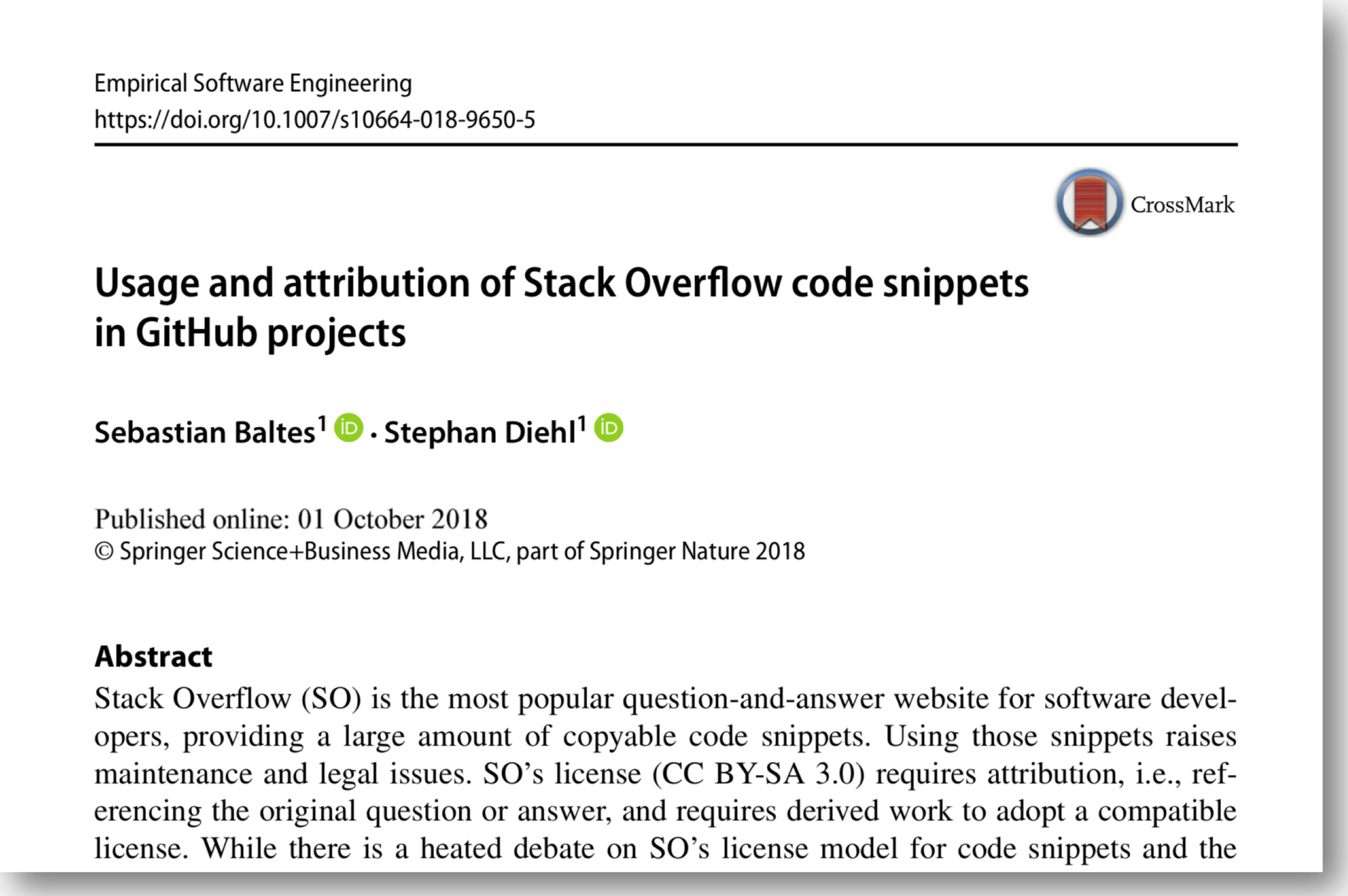
1 Sketching

2 Expertise Development

3 Code Plagiarism

Code Plagiarism

Code Plagiarism



Empirical Software Engineering
https://doi.org/10.1007/s10664-018-9650-5

CrossMark

Usage and attribution of Stack Overflow code snippets in GitHub projects

Sebastian Baltes[1] · **Stephan Diehl**[1]

Published online: 01 October 2018

Abstract

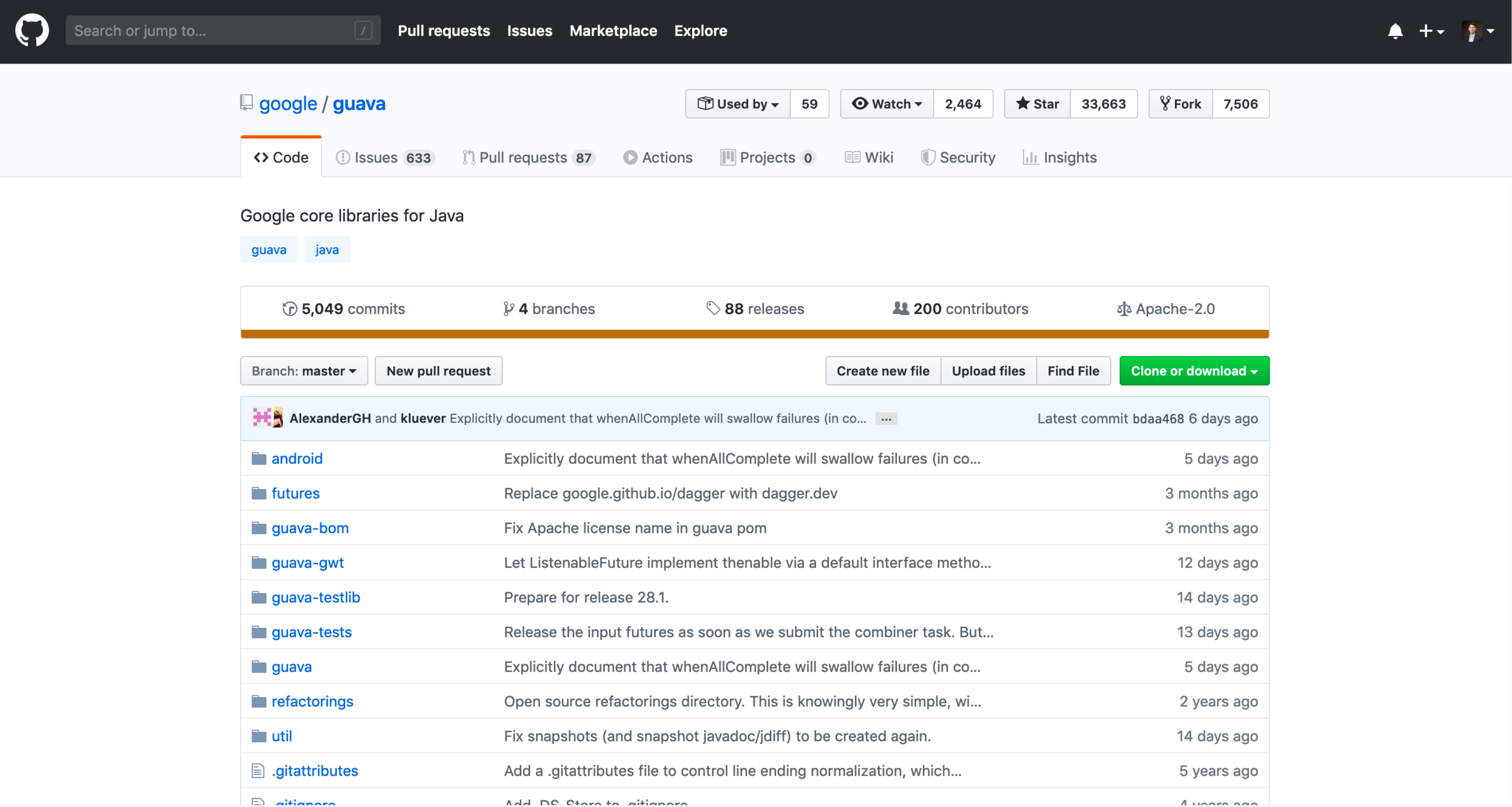
Stack Overflow (SO) is the most popular question-and-answer website for software developers, providing a large amount of copyable code snippets. Using those snippets raises maintenance and legal issues. SO's license (CC BY-SA 3.0) requires attribution, i.e., referencing the original question or answer, and requires derived work to adopt a compatible license. While there is a heated debate on SO's license model for code snippets and the

https://empirical-software.engineering/projects/snippets/

GitHub

- **Hosted version control** platform for (software) projects
- Features include access control, **collaboration features** such as **issue tracking**, wikis, gamification of development activity
- **Public** projects and **private** projects with up to three collaborators are **free**
- As of May 2019: **>37m users** and **>100m projects**

GitHub



google / guava

Google core libraries for Java

guava java

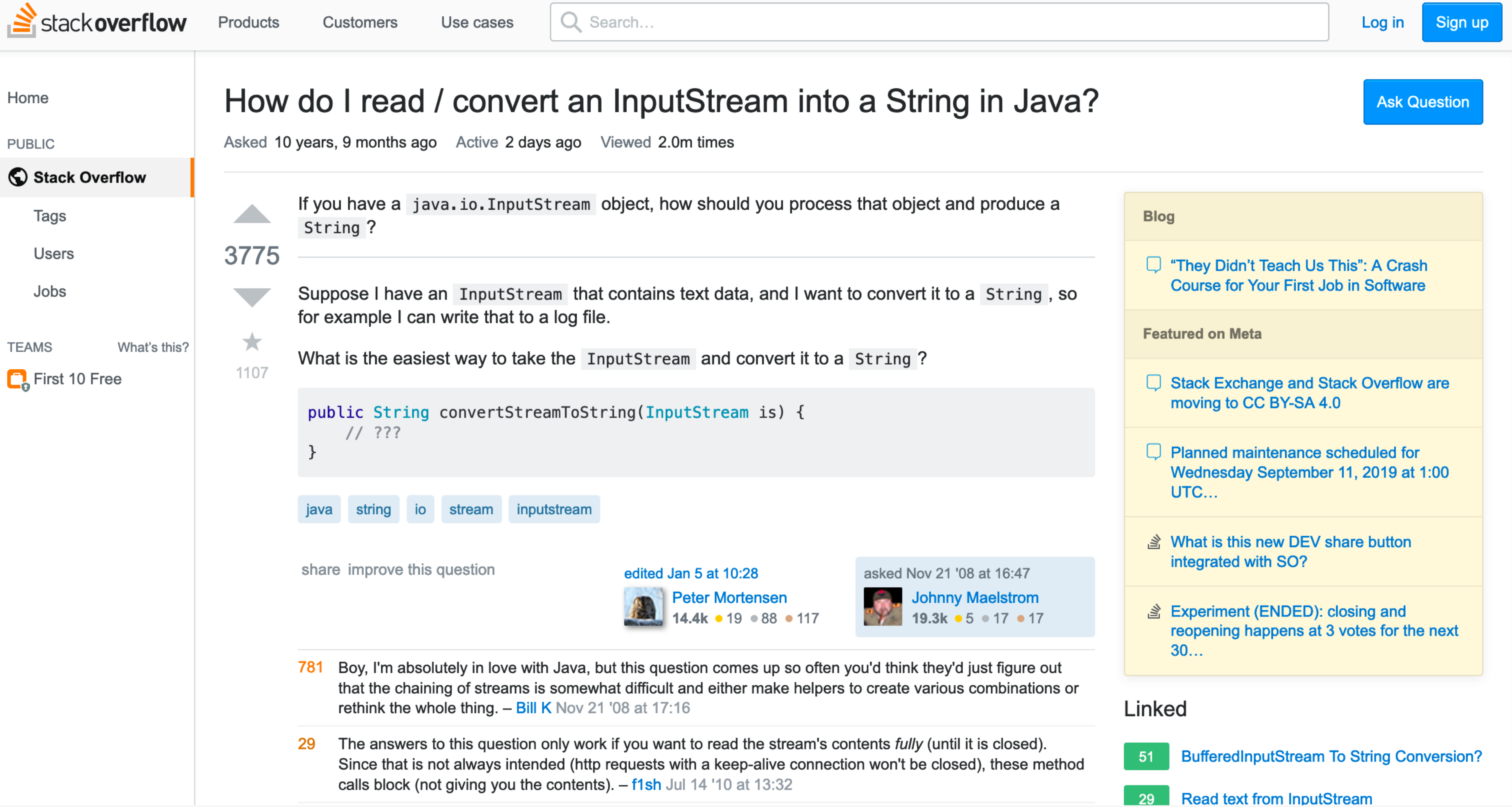
5,049 commits · 4 branches · 88 releases · 200 contributors · Apache-2.0

AlexanderGH and kluever Explicitly document that whenAllComplete will swallow failures (in co... … Latest commit bdaa468 6 days ago

android	Explicitly document that whenAllComplete will swallow failures (in co...	5 days ago
futures	Replace google.github.io/dagger with dagger.dev	3 months ago
guava-bom	Fix Apache license name in guava pom	3 months ago
guava-gwt	Let ListenableFuture implement thenable via a default interface metho...	12 days ago
guava-testlib	Prepare for release 28.1.	14 days ago
guava-tests	Release the input futures as soon as we submit the combiner task. But...	13 days ago
guava	Explicitly document that whenAllComplete will swallow failures (in co...	5 days ago
refactorings	Open source refactorings directory. This is knowingly very simple, wi...	2 years ago
util	Fix snapshots (and snapshot javadoc/jdiff) to be created again.	14 days ago
.gitattributes	Add a .gitattributes file to control line ending normalization, which...	5 years ago
.gitignore	Add .DS_Store to .gitignore	4 years ago

Stack Overflow

- **Question and answer** website for software developers
- Covers a **wide variety** of **programming-related topics**
- Posts can be commented, edited, and up-/down-voted
- **Gamification** through reputation points awarded for different kinds of contributions
- **Jobs** section for advertising employment opportunities
- As of June 2019 **>10.5m** registered users and **>17.7m questions**



How do I read / convert an InputStream into a String in Java?

Asked 10 years, 9 months ago Active 2 days ago Viewed 2.0m times

3775

1107

If you have a `java.io.InputStream` object, how should you process that object and produce a `String`?

Suppose I have an `InputStream` that contains text data, and I want to convert it to a `String`, so for example I can write that to a log file.

What is the easiest way to take the `InputStream` and convert it to a `String`?

```
public String convertStreamToString(InputStream is) {
    // ???
}
```

java string io stream inputstream

share improve this question

edited Jan 5 at 10:28
Peter Mortensen
14.4k ● 19 ● 88 ● 117

asked Nov 21 '08 at 16:47
Johnny Maelstrom
19.3k ● 5 ● 17 ● 17

781 Boy, I'm absolutely in love with Java, but this question comes up so often you'd think they'd just figure out that the chaining of streams is somewhat difficult and either make helpers to create various combinations or rethink the whole thing. – Bill K Nov 21 '08 at 17:16

29 The answers to this question only work if you want to read the stream's contents *fully* (until it is closed). Since that is not always intended (http requests with a keep-alive connection won't be closed), these method calls block (not giving you the contents). – f1sh Jul 14 '10 at 13:32

Blog

"They Didn't Teach Us This": A Crash Course for Your First Job in Software

Featured on Meta

Stack Exchange and Stack Overflow are moving to CC BY-SA 4.0

Planned maintenance scheduled for Wednesday September 11, 2019 at 1:00 UTC…

What is this new DEV share button integrated with SO?

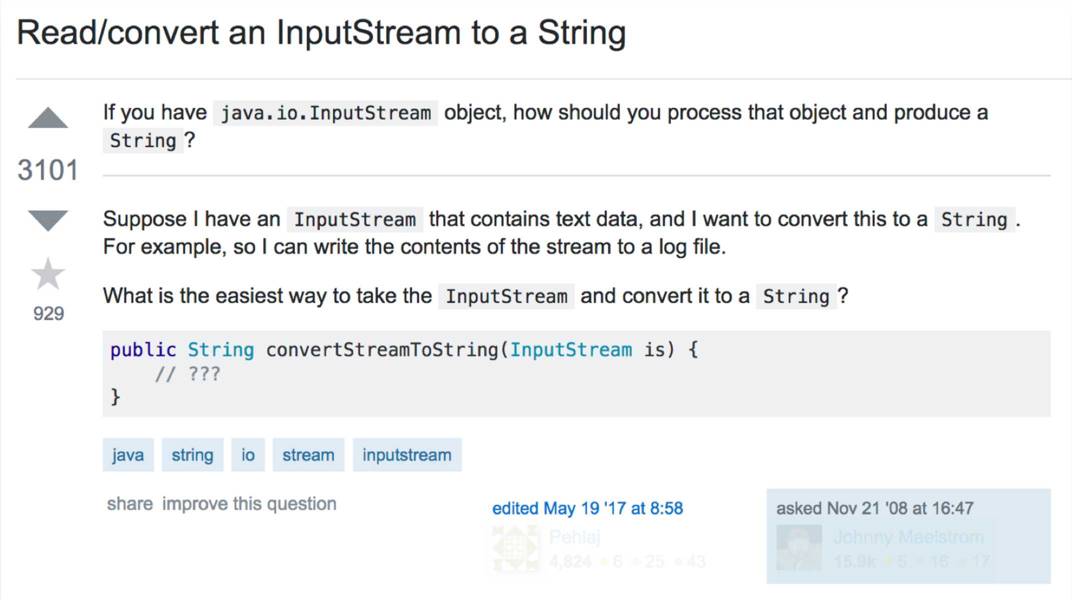
Experiment (ENDED): closing and reopening happens at 3 votes for the next 30…

Linked

51 BufferedInputStream To String Conversion?

29 Read text from InputStream

Example



Read/convert an InputStream to a String

3101

929

If you have `java.io.InputStream` object, how should you process that object and produce a `String`?

Suppose I have an `InputStream` that contains text data, and I want to convert this to a `String`. For example, so I can write the contents of the stream to a log file.

What is the easiest way to take the `InputStream` and convert it to a `String`?

```
public String convertStreamToString(InputStream is) {
    // ???
}
```

java string io stream inputstream

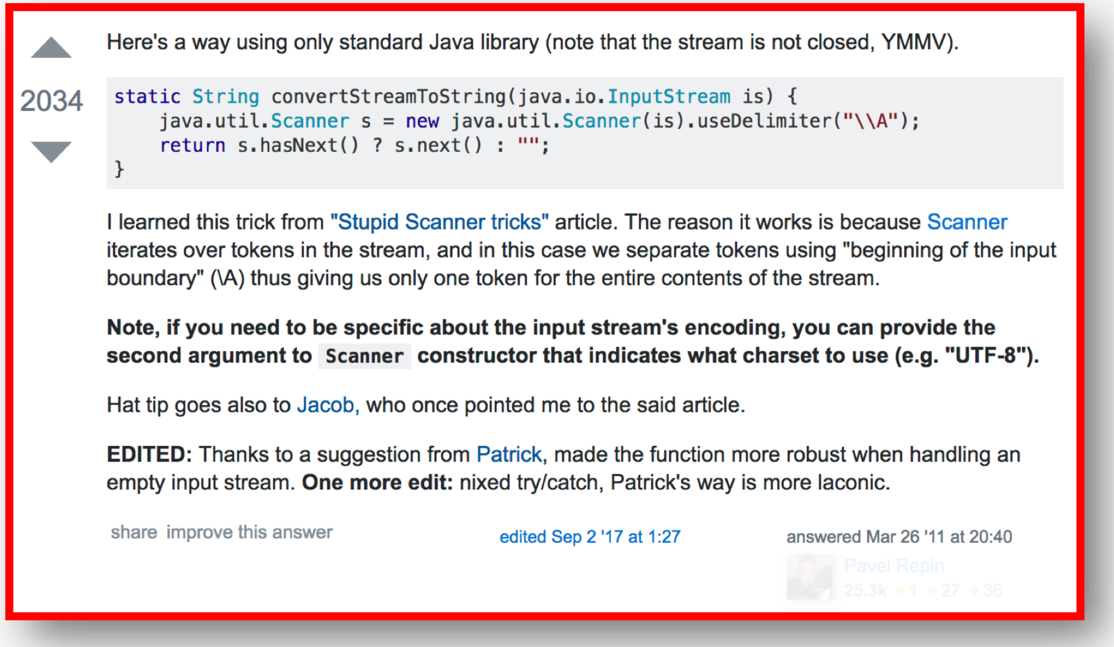
share improve this question

edited May 19 '17 at 8:58
Pehlaj
4,824 ●6 ●25 ●43

asked Nov 21 '08 at 16:47
Johnny Maelstrom
15.9k ●5 ●16 ●17

Question

https://stackoverflow.com/q/309424



2034

Here's a way using only standard Java library (note that the stream is not closed, YMMV).

```
static String convertStreamToString(java.io.InputStream is) {
    java.util.Scanner s = new java.util.Scanner(is).useDelimiter("\\A");
    return s.hasNext() ? s.next() : "";
}
```

I learned this trick from "Stupid Scanner tricks" article. The reason it works is because Scanner iterates over tokens in the stream, and in this case we separate tokens using "beginning of the input boundary" (\A) thus giving us only one token for the entire contents of the stream.

Note, if you need to be specific about the input stream's encoding, you can provide the second argument to `Scanner` constructor that indicates what charset to use (e.g. "UTF-8").

Hat tip goes also to Jacob, who once pointed me to the said article.

EDITED: Thanks to a suggestion from Patrick, made the function more robust when handling an empty input stream. **One more edit:** nixed try/catch, Patrick's way is more laconic.

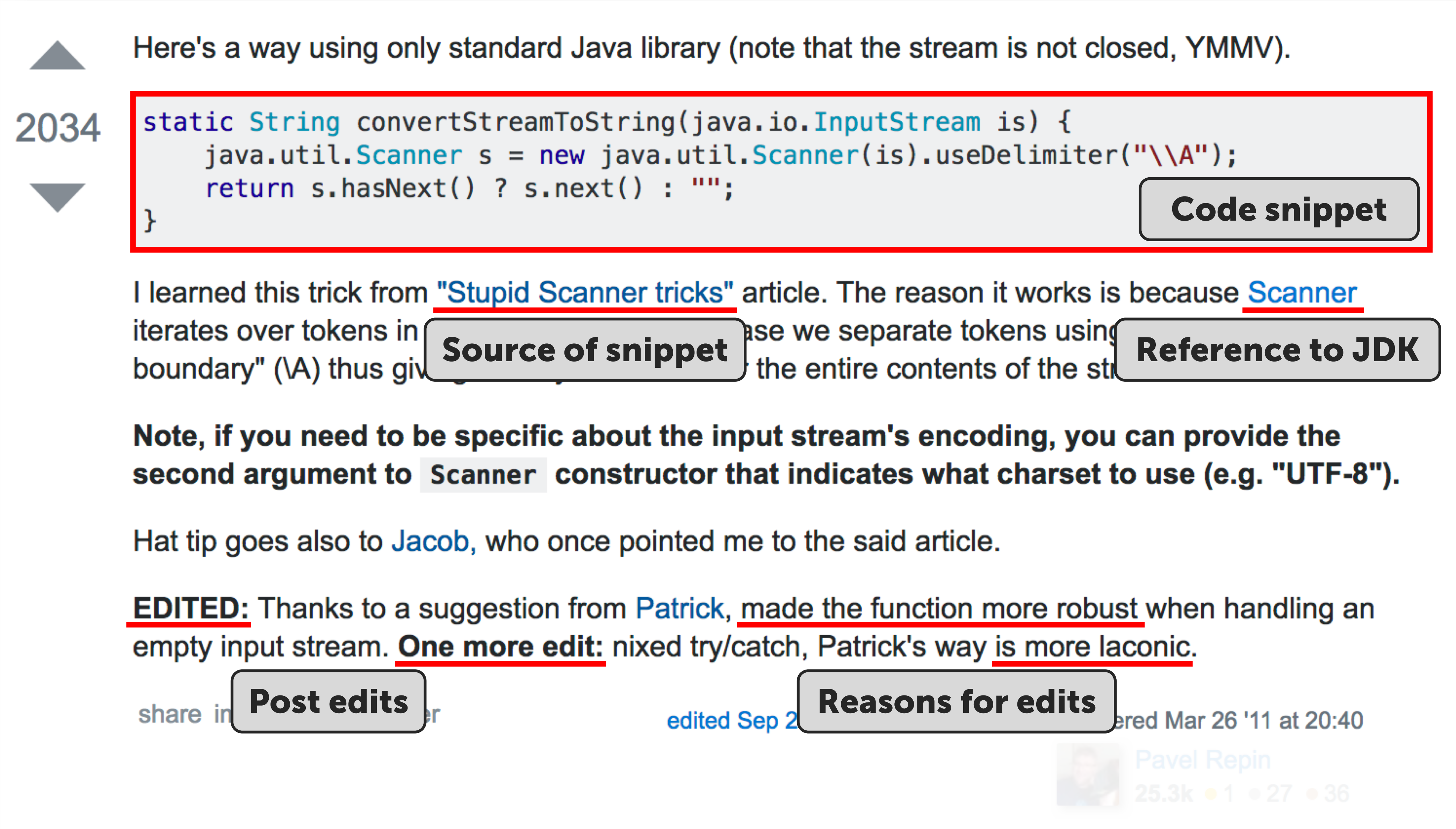
share improve this answer

edited Sep 2 '17 at 1:27

answered Mar 26 '11 at 20:40
Pavel Repin
25.3k ●1 ●27 ●36

Answer

https://stackoverflow.com/a/5445161



2034

Here's a way using only standard Java library (note that the stream is not closed, YMMV).

```
static String convertStreamToString(java.io.InputStream is) {
    java.util.Scanner s = new java.util.Scanner(is).useDelimiter("\\A");
    return s.hasNext() ? s.next() : "";
}
```

I learned this trick from "Stupid Scanner tricks" article. The reason it works is because Scanner iterates over tokens in ase we separate tokens using boundary" (\A) thus giv the entire contents of the st

Note, if you need to be specific about the input stream's encoding, you can provide the second argument to `Scanner` constructor that indicates what charset to use (e.g. "UTF-8").

Hat tip goes also to Jacob, who once pointed me to the said article.

EDITED: Thanks to a suggestion from Patrick, made the function more robust when handling an empty input stream. **One more edit:** nixed try/catch, Patrick's way is more laconic.

share in er

edited Sep 2 ered Mar 26 '11 at 20:40

Pavel Repin

25.3k ● 1 ● 27 ● 36

Code snippet

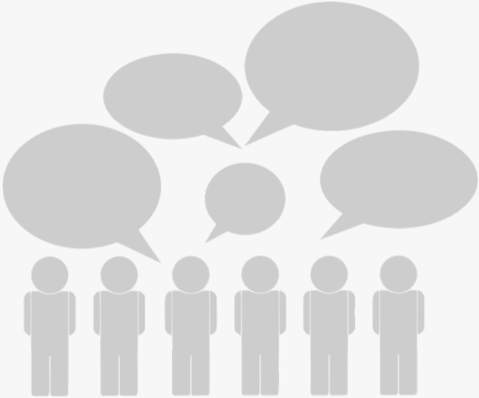
Source of snippet

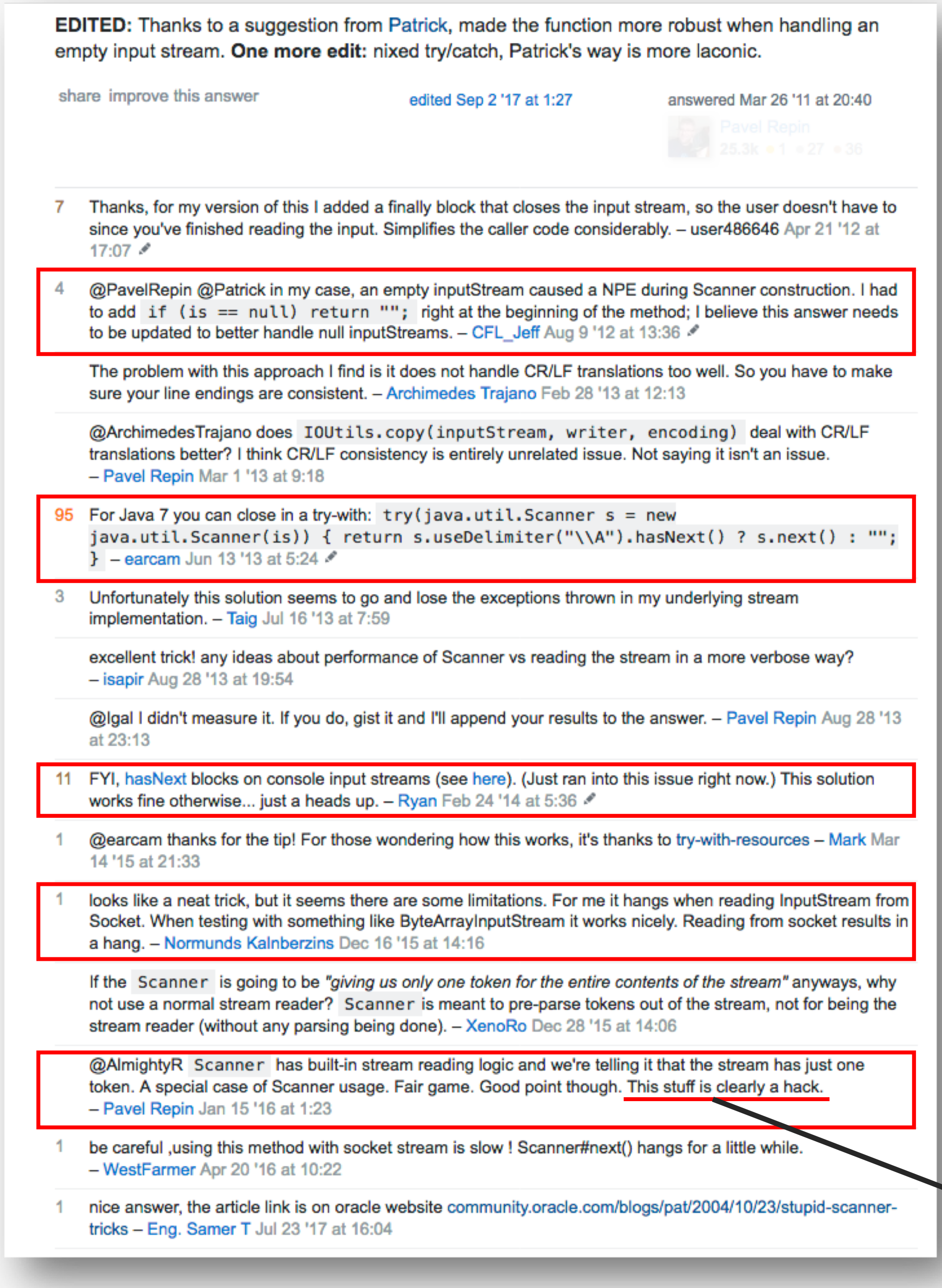
Reference to JDK

Post edits

Reasons for edits

Comments





EDITED: Thanks to a suggestion from Patrick, made the function more robust when handling an empty input stream. **One more edit:** nixed try/catch, Patrick's way is more laconic.

share improve this answer edited Sep 2 '17 at 1:27 answered Mar 26 '11 at 20:40

Pavel Repin
25.3k ●1 ●27 ●36

7 Thanks, for my version of this I added a finally block that closes the input stream, so the user doesn't have to since you've finished reading the input. Simplifies the caller code considerably. – user486646 Apr 21 '12 at 17:07

4 @PavelRepin @Patrick in my case, an empty inputStream caused a NPE during Scanner construction. I had to add `if (is == null) return "";` right at the beginning of the method; I believe this answer needs to be updated to better handle null inputStreams. – CFL_Jeff Aug 9 '12 at 13:36

The problem with this approach I find is it does not handle CR/LF translations too well. So you have to make sure your line endings are consistent. – Archimedes Trajano Feb 28 '13 at 12:13

@ArchimedesTrajano does `IOUtils.copy(inputStream, writer, encoding)` deal with CR/LF translations better? I think CR/LF consistency is entirely unrelated issue. Not saying it isn't an issue. – Pavel Repin Mar 1 '13 at 9:18

95 For Java 7 you can close in a try-with: `try(java.util.Scanner s = new java.util.Scanner(is)) { return s.useDelimiter("\\A").hasNext() ? s.next() : ""; }` – earcam Jun 13 '13 at 5:24

3 Unfortunately this solution seems to go and lose the exceptions thrown in my underlying stream implementation. – Taig Jul 16 '13 at 7:59

excellent trick! any ideas about performance of Scanner vs reading the stream in a more verbose way? – isapir Aug 28 '13 at 19:54

@Igal I didn't measure it. If you do, gist it and I'll append your results to the answer. – Pavel Repin Aug 28 '13 at 23:13

11 FYI, hasNext blocks on console input streams (see here). (Just ran into this issue right now.) This solution works fine otherwise... just a heads up. – Ryan Feb 24 '14 at 5:36

1 @earcam thanks for the tip! For those wondering how this works, it's thanks to try-with-resources – Mark Mar 14 '15 at 21:33

1 looks like a neat trick, but it seems there are some limitations. For me it hangs when reading InputStream from Socket. When testing with something like ByteArrayInputStream it works nicely. Reading from socket results in a hang. – Normunds Kalnberzins Dec 16 '15 at 14:16

If the `Scanner` is going to be *"giving us only one token for the entire contents of the stream"* anyways, why not use a normal stream reader? `Scanner` is meant to pre-parse tokens out of the stream, not for being the stream reader (without any parsing being done). – XenoRo Dec 28 '15 at 14:06

@AlmightyR `Scanner` has built-in stream reading logic and we're telling it that the stream has just one token. A special case of Scanner usage. Fair game. Good point though. This stuff is clearly a hack. – Pavel Repin Jan 15 '16 at 1:23

1 be careful ,using this method with socket stream is slow ! Scanner#next() hangs for a little while. – WestFarmer Apr 20 '16 at 10:22

1 nice answer, the article link is on oracle website community.oracle.com/blogs/pat/2004/10/23/stupid-scanner-tricks – Eng. Samer T Jul 23 '17 at 16:04

Bug report

Alternative solution

Bug report

Bug report

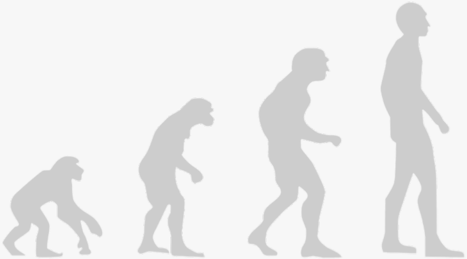
Comment by author

This stuff is clearly a hack.

Evolution

- Like other **software artifacts**, SO posts **evolve:**
  - Content of **17.3m** posts has been edited
  - **Bugs** in code snippets are fixed
  - **Clarifications** are added in text documenting the code
  - Snippets are **updated** to new language/library versions

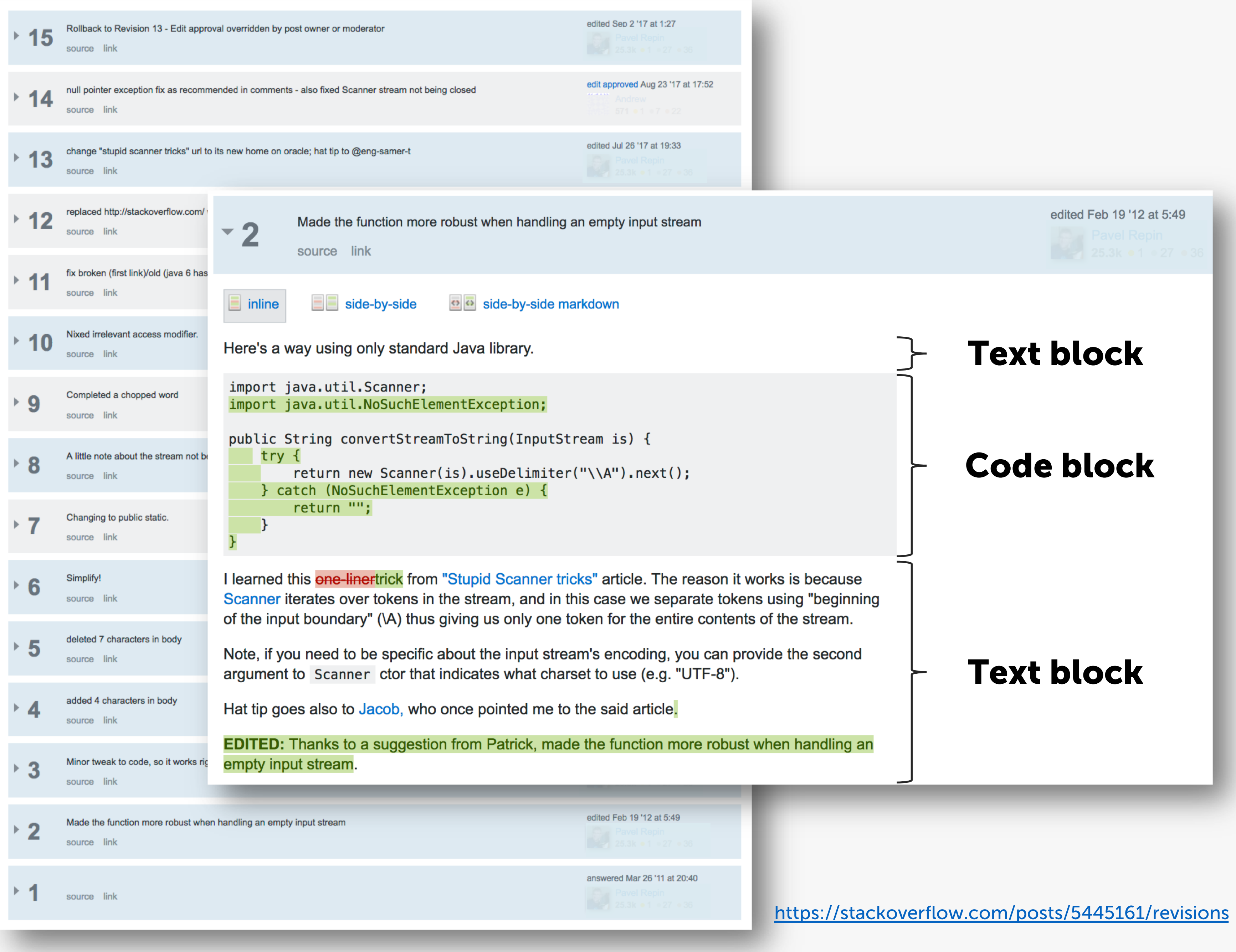
- **Evolution of code on SO** differs from regular software projects:
  - **Short** code snippets (12 LOC on average)
  - **No bug tracking** system (just comments and new answers)
  - **No versioning** for individual snippets (just whole posts)



SO Revisions

Problems:

- Version history is only available on the level of whole posts, thus **individual code snippets hard to trace**
- **Comments and edits** are not linked
- Unclear how **external sources** are related



15 Rollback to Revision 13 - Edit approval overridden by post owner or moderator — edited Sep 2 '17 at 1:27
source link

14 null pointer exception fix as recommended in comments - also fixed Scanner stream not being closed — edit approved Aug 23 '17 at 17:52
source link

13 change "stupid scanner tricks" url to its new home on oracle; hat tip to @eng-samer-t — edited Jul 26 '17 at 19:33
source link

12 replaced http://stackoverflow.com/
source link

11 fix broken (first link)/old (java 6 has
source link

10 Nixed irrelevant access modifier.
source link

9 Completed a chopped word
source link

8 A little note about the stream not b
source link

7 Changing to public static.
source link

6 Simplify!
source link

5 deleted 7 characters in body
source link

4 added 4 characters in body
source link

3 Minor tweak to code, so it works ri
source link

2 Made the function more robust when handling an empty input stream — edited Feb 19 '12 at 5:49
source link

1 — answered Mar 26 '11 at 20:40
source link

2 Made the function more robust when handling an empty input stream — edited Feb 19 '12 at 5:49
source link

inline | side-by-side | side-by-side markdown

Here's a way using only standard Java library.

Text block

```
import java.util.Scanner;
import java.util.NoSuchElementException;

public String convertStreamToString(InputStream is) {
    try {
        return new Scanner(is).useDelimiter("\\A").next();
    } catch (NoSuchElementException e) {
        return "";
    }
}
```

Code block

I learned this ~~one-liner~~trick from "Stupid Scanner tricks" article. The reason it works is because Scanner iterates over tokens in the stream, and in this case we separate tokens using "beginning of the input boundary" (\A) thus giving us only one token for the entire contents of the stream.

Note, if you need to be specific about the input stream's encoding, you can provide the second argument to `Scanner` ctor that indicates what charset to use (e.g. "UTF-8").

Hat tip goes also to Jacob, who once pointed me to the said article.

EDITED: Thanks to a suggestion from Patrick, made the function more robust when handling an empty input stream.

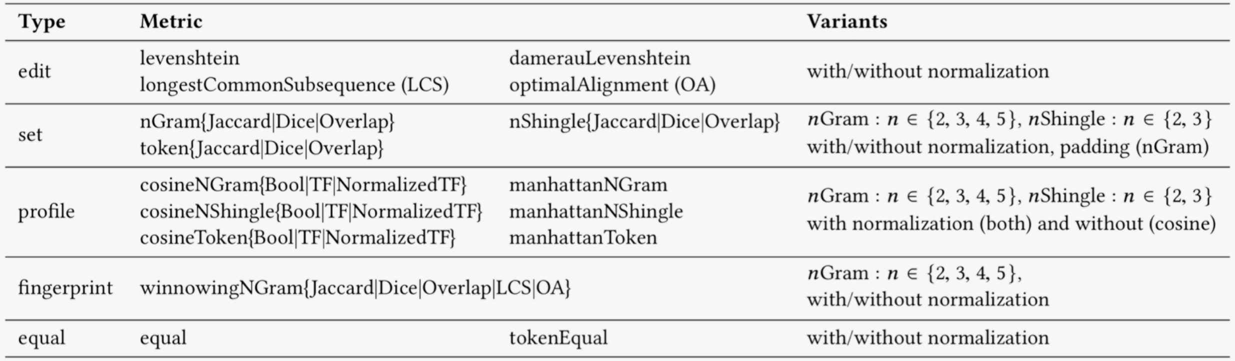
Text block

https://stackoverflow.com/posts/5445161/revisions

SOTorrent

Fork me on GitHub

- Among other features, the dataset provides the **version history** of Stack Overflow content on the **level of individual text or code blocks**
- Extraction of post blocks and mapping to their predecessors was required, involving an extensive evaluation of similarity metrics



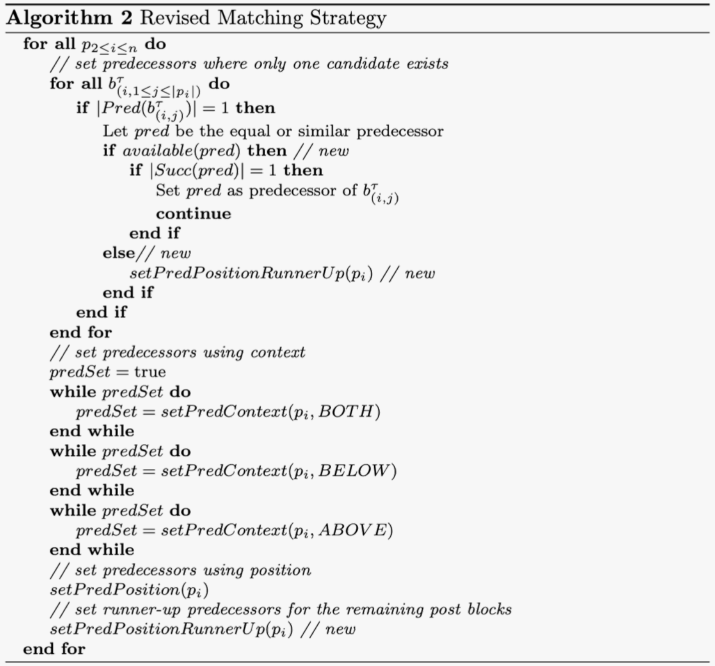
Type	Metric		Variants
edit	levenshtein longestCommonSubsequence (LCS)	damerauLevenshtein optimalAlignment (OA)	with/without normalization

set	nGram{Jaccard|Dice|Overlap}
	token{Jaccard|Dice|Overlap}
		nShingle{Jaccard|Dice|Overlap}	$nGram : n \in \{2, 3, 4, 5\}$, $nShingle : n \in \{2, 3\}$
			with/without normalization, padding (nGram)

profile	cosineNGram{Bool|TF|NormalizedTF}
	cosineNShingle{Bool|TF|NormalizedTF}
	cosineToken{Bool|TF|NormalizedTF}
		manhattanNGram
		manhattanNShingle
		manhattanToken
			$nGram : n \in \{2, 3, 4, 5\}$, $nShingle : n \in \{2, 3\}$
			with normalization (both) and without (cosine)

fingerprint	winnowingNGram{Jaccard|Dice|Overlap|LCS|OA}		$nGram : n \in \{2, 3, 4, 5\}$, with/without normalization
equal	equal	tokenEqual	with/without normalization

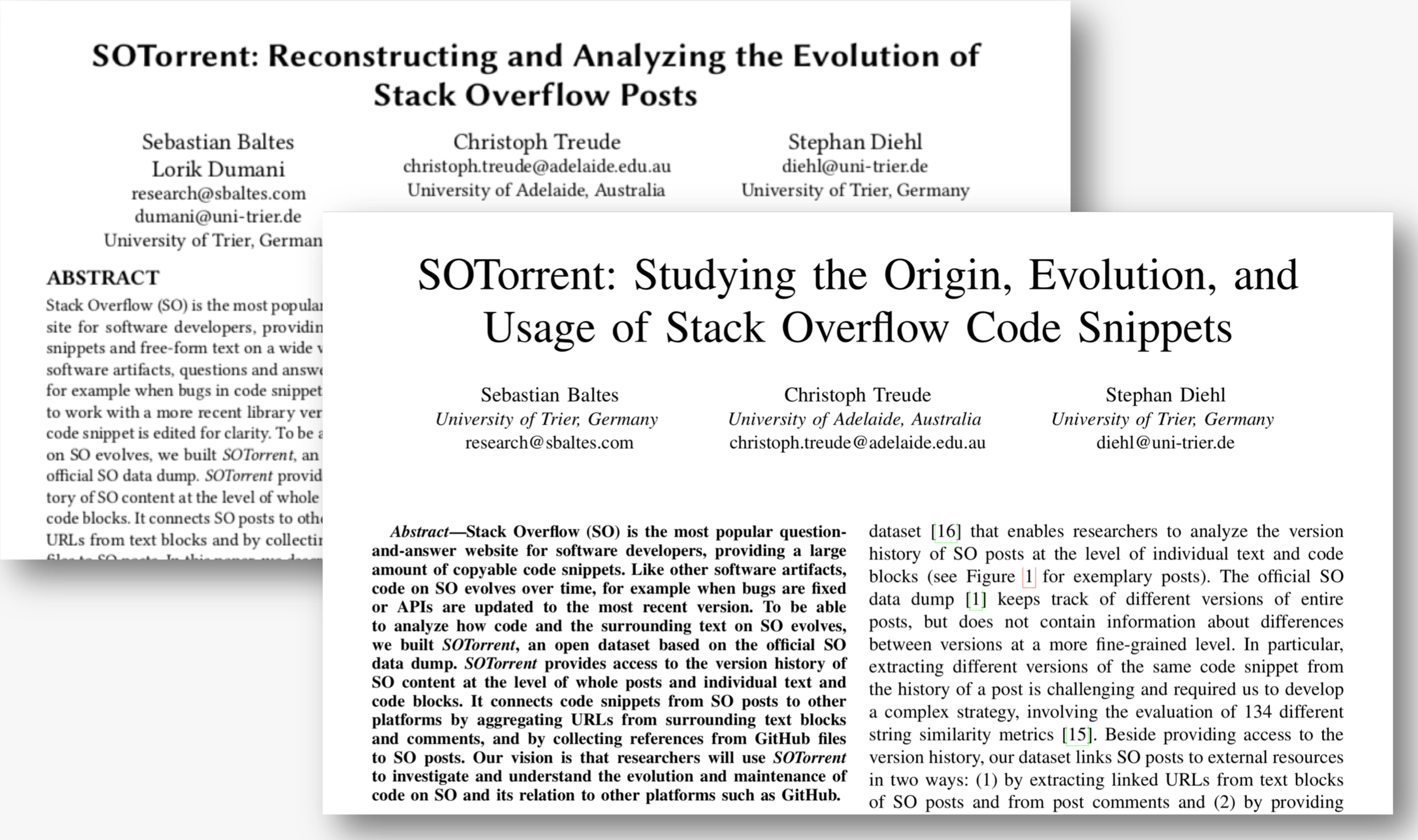
https://github.com/sotorrent/string-similarity



Algorithm 2 Revised Matching Strategy

```
for all $p_{2 \le i \le n}$ do
    // set predecessors where only one candidate exists
    for all $b^{\tau}_{(i,1 \le j \le |p_i|)}$ do
        if $|Pred(b^{\tau}_{(i,j)})| = 1$ then
            Let pred be the equal or similar predecessor
            if available(pred) then // new
                if $|Succ(pred)| = 1$ then
                    Set pred as predecessor of $b^{\tau}_{(i,j)}$
                    continue
                end if
            else// new
                setPredPositionRunnerUp($p_i$) // new
            end if
        end if
    end for
    // set predecessors using context
    predSet = true
    while predSet do
        predSet = setPredContext($p_i$, BOTH)
    end while
    while predSet do
        predSet = setPredContext($p_i$, BELOW)
    end while
    while predSet do
        predSet = setPredContext($p_i$, ABOVE)
    end while
    // set predecessors using position
    setPredPosition($p_i$)
    // set runner-up predecessors for the remaining post blocks
    setPredPositionRunnerUp($p_i$) // new
end for
```

https://github.com/sotorrent/posthistory-extractor



SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts

Sebastian Baltes
Lorik Dumani
research@sbaltes.com
dumani@uni-trier.de
University of Trier, German

Christoph Treude
christoph.treude@adelaide.edu.au
University of Adelaide, Australia

Stephan Diehl
diehl@uni-trier.de
University of Trier, Germany

ABSTRACT

Stack Overflow (SO) is the most popular
site for software developers, providin
snippets and free-form text on a wide v
software artifacts, questions and answ
for example when bugs in code snippet
to work with a more recent library ver
code snippet is edited for clarity. To be a
on SO evolves, we built *SOTorrent*, an
official SO data dump. *SOTorrent* provid
tory of SO content at the level of whole
code blocks. It connects SO posts to oth
URLs from text blocks and by collectin

SOTorrent: Studying the Origin, Evolution, and Usage of Stack Overflow Code Snippets

Sebastian Baltes
University of Trier, Germany
research@sbaltes.com

Christoph Treude
University of Adelaide, Australia
christoph.treude@adelaide.edu.au

Stephan Diehl
University of Trier, Germany
diehl@uni-trier.de

***Abstract*—Stack Overflow (SO) is the most popular question-and-answer website for software developers, providing a large amount of copyable code snippets. Like other software artifacts, code on SO evolves over time, for example when bugs are fixed or APIs are updated to the most recent version. To be able to analyze how code and the surrounding text on SO evolves, we built *SOTorrent*, an open dataset based on the official SO data dump. *SOTorrent* provides access to the version history of SO content at the level of whole posts and individual text and code blocks. It connects code snippets from SO posts to other platforms by aggregating URLs from surrounding text blocks and comments, and by collecting references from GitHub files to SO posts. Our vision is that researchers will use *SOTorrent* to investigate and understand the evolution and maintenance of code on SO and its relation to other platforms such as GitHub.**

dataset [16] that enables researchers to analyze the version history of SO posts at the level of individual text and code blocks (see Figure 1 for exemplary posts). The official SO data dump [1] keeps track of different versions of entire posts, but does not contain information about differences between versions at a more fine-grained level. In particular, extracting different versions of the same code snippet from the history of a post is challenging and required us to develop a complex strategy, involving the evaluation of 134 different string similarity metrics [15]. Beside providing access to the version history, our dataset links SO posts to external resources in two ways: (1) by extracting linked URLs from text blocks of SO posts and from post comments and (2) by providing

MSR 2018/19

sotorrent.org

Open Data

Dataset available on Zenodo and BigQuery

Cloud Icon CC BY 3.0 smashicons on flaticon.com

stackoverflow
Copy
Paste

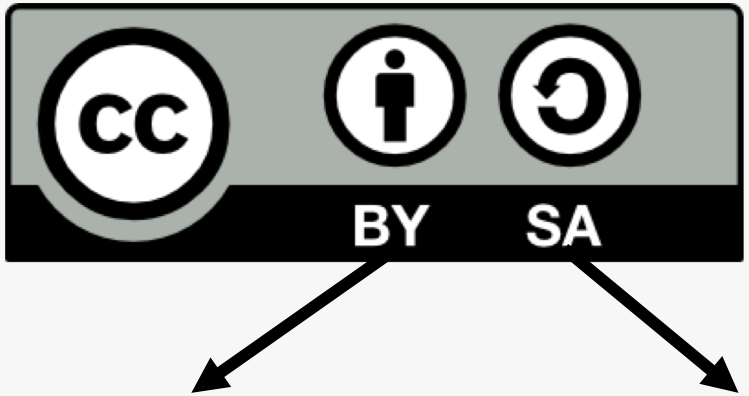
Question for the Audience I

Who admits regularly copying non-trivial code snippets from Stack Overflow?

Question for the Audience II

Who knew that all content on Stack Overflow is licensed under CC BY-SA?

*"You must give **appropriate credit** [...] and indicate if changes were made."*



Attribution

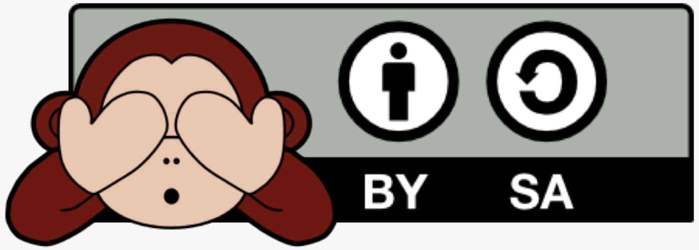
Share-alike

*“If you [...] **build upon** the material, you must **distribute your contributions** under the same license as the original.”*

Results from our Online Surveys

- **46%** of the participants admitted copying code from Stack Overflow **without attribution**
- **75%** did **not know** that content on SO is licensed under **CC BY-SA**
- **67%** did **not know** that **attribution is required**

→ **Lack of awareness**

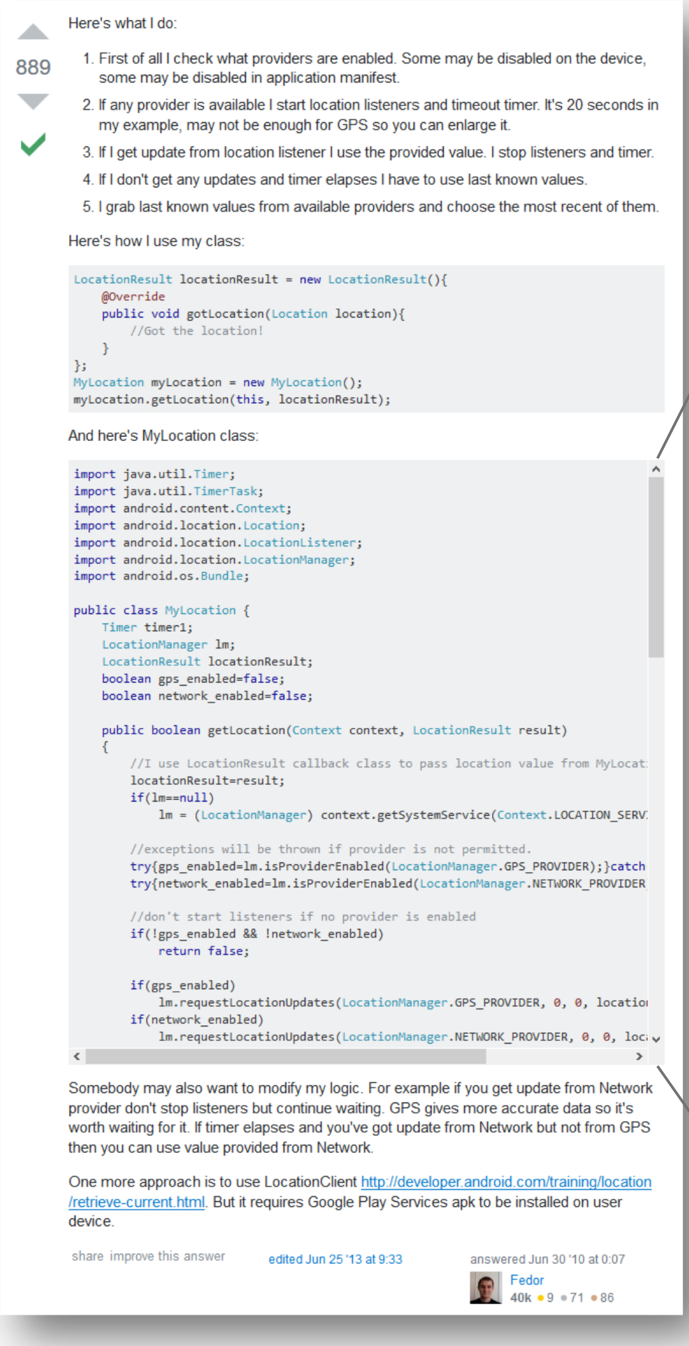


Background

http://theconversation.com/why-universities-cant-be-expected-to-police-copyright-infringement-82677

"Well, but these snippets are rather trivial and not protected by copyright."

- Not all code snippets on Stack Overflow are copyrightable
- "A snippet that is more than one or two lines of standard function calls would typically be creative enough for copyright" [Engelfriet 2016]
- But no "international standard for originality" [Creative Commons 2017b]



Here's what I do:

1. First of all I check what providers are enabled. Some may be disabled on the device, some may be disabled in application manifest.
2. If any provider is available I start location listeners and timeout timer. It's 20 seconds in my example, may not be enough for GPS so you can enlarge it.
3. If I get update from location listener I use the provided value. I stop listeners and timer.
4. If I don't get any updates and timer elapses I have to use last known values.
5. I grab last known values from available providers and choose the most recent of them.

Here's how I use my class:

```
LocationResult locationResult = new LocationResult(){
    @Override
    public void gotLocation(Location location){
        //Got the location!
    }
};
MyLocation myLocation = new MyLocation();
myLocation.getLocation(this, locationResult);
```

And here's MyLocation class:

```
import java.util.Timer;
import java.util.TimerTask;
import android.content.Context;
import android.location.Location;
import android.location.LocationListener;
import android.location.LocationManager;
import android.os.Bundle;

public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled=false;
    boolean network_enabled=false;

    public boolean getLocation(Context context, LocationResult result)
    {
        //I use LocationResult callback class to pass location value from MyLocat
        locationResult=result;
        if(lm==null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERV

        //exceptions will be thrown if provider is not permitted.
        try{gps_enabled=lm.isProviderEnabled(LocationManager.GPS_PROVIDER);}catch
        try{network_enabled=lm.isProviderEnabled(LocationManager.NETWORK_PROVIDER

        //don't start listeners if no provider is enabled
        if(!gps_enabled && !network_enabled)
            return false;

        if(gps_enabled)
            lm.requestLocationUpdates(LocationManager.GPS_PROVIDER, 0, 0, locatio
        if(network_enabled)
            lm.requestLocationUpdates(LocationManager.NETWORK_PROVIDER, 0, 0, loc
```

Somebody may also want to modify my logic. For example if you get update from Network provider don't stop listeners but continue waiting. GPS gives more accurate data so it's worth waiting for it. If timer elapses and you've got update from Network but not from GPS then you can use value provided from Network.

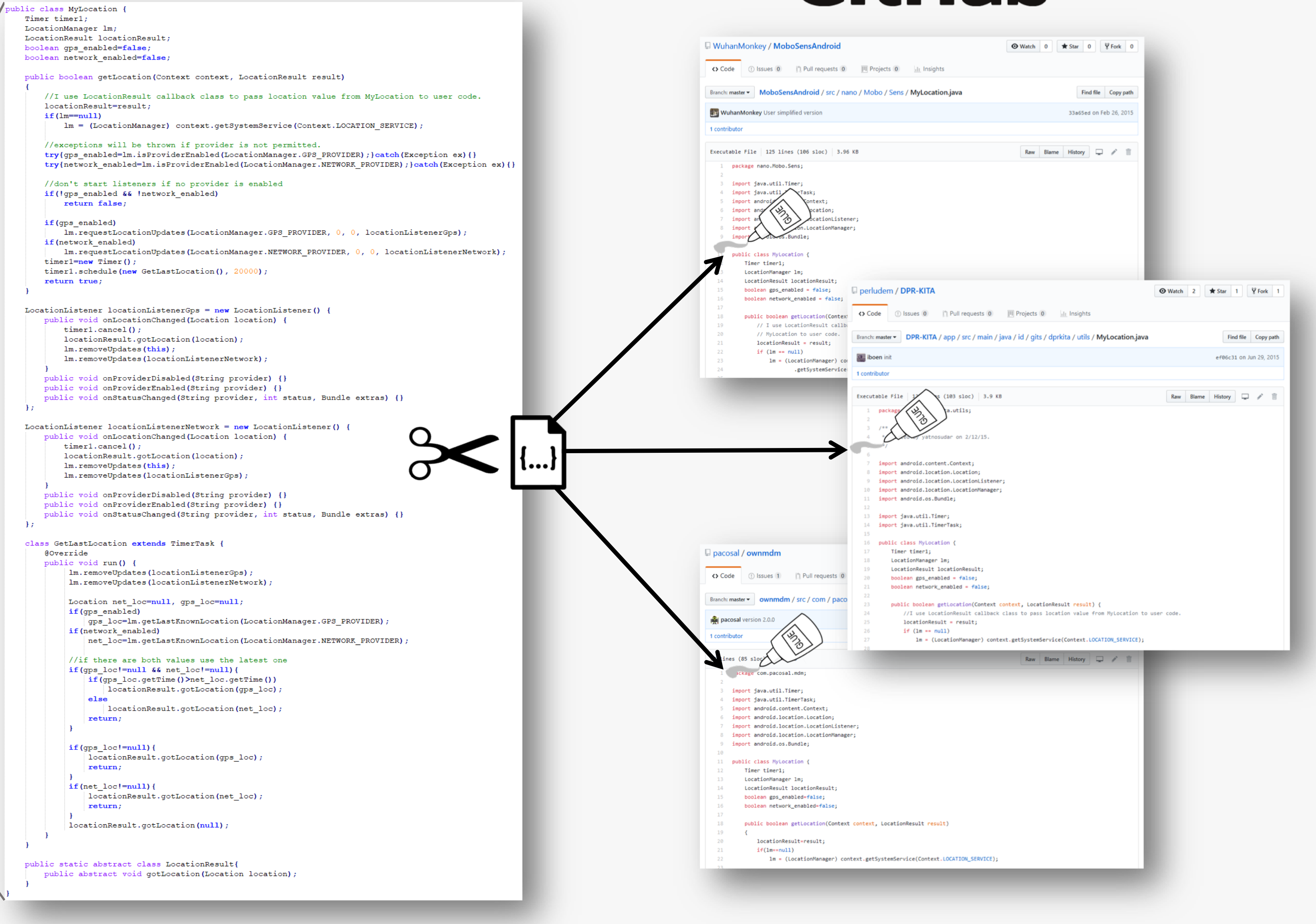
One more approach is to use LocationClient http://developer.android.com/training/location/retrieve-current.html. But it requires Google Play Services apk to be installed on user device.

share improve this answer   edited Jun 25 '13 at 9:33   answered Jun 30 '10 at 0:07

Fedor
40k ● 9 ● 71 ● 86

stackoverflow

GitHub



```
public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled=false;
    boolean network_enabled=false;

    public boolean getLocation(Context context, LocationResult result)
    {
        //I use LocationResult callback class to pass location value from MyLocation to user code.
        locationResult=result;
        if(lm==null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERVICE);

        //exceptions will be thrown if provider is not permitted.
        try{gps_enabled=lm.isProviderEnabled(LocationManager.GPS_PROVIDER);}catch(Exception ex){}
        try{network_enabled=lm.isProviderEnabled(LocationManager.NETWORK_PROVIDER);}catch(Exception ex){}

        //don't start listeners if no provider is enabled
        if(!gps_enabled && !network_enabled)
            return false;

        if(gps_enabled)
            lm.requestLocationUpdates(LocationManager.GPS_PROVIDER, 0, 0, locationListenerGps);
        if(network_enabled)
            lm.requestLocationUpdates(LocationManager.NETWORK_PROVIDER, 0, 0, locationListenerNetwork);
        timer1=new Timer();
        timer1.schedule(new GetLastLocation(), 20000);
        return true;
    }

    LocationListener locationListenerGps = new LocationListener() {
        public void onLocationChanged(Location location) {
            timer1.cancel();
            locationResult.gotLocation(location);
            lm.removeUpdates(this);
            lm.removeUpdates(locationListenerNetwork);
        }
        public void onProviderDisabled(String provider) {}
        public void onProviderEnabled(String provider) {}
        public void onStatusChanged(String provider, int status, Bundle extras) {}
    };

    LocationListener locationListenerNetwork = new LocationListener() {
        public void onLocationChanged(Location location) {
            timer1.cancel();
            locationResult.gotLocation(location);
            lm.removeUpdates(this);
            lm.removeUpdates(locationListenerGps);
        }
        public void onProviderDisabled(String provider) {}
        public void onProviderEnabled(String provider) {}
        public void onStatusChanged(String provider, int status, Bundle extras) {}
    };

    class GetLastLocation extends TimerTask {
        @Override
        public void run() {
             lm.removeUpdates(locationListenerGps);
             lm.removeUpdates(locationListenerNetwork);

             Location net_loc=null, gps_loc=null;
             if(gps_enabled)
                 gps_loc=lm.getLastKnownLocation(LocationManager.GPS_PROVIDER);
             if(network_enabled)
                 net_loc=lm.getLastKnownLocation(LocationManager.NETWORK_PROVIDER);

             //if there are both values use the latest one
             if(gps_loc!=null && net_loc!=null){
                 if(gps_loc.getTime()>net_loc.getTime())
                     locationResult.gotLocation(gps_loc);
                 else
                     locationResult.gotLocation(net_loc);
                 return;
             }

             if(gps_loc!=null){
                 locationResult.gotLocation(gps_loc);
                 return;
             }
             if(net_loc!=null){
                 locationResult.gotLocation(net_loc);
                 return;
             }
             locationResult.gotLocation(null);
        }
    }

    public static abstract class LocationResult{
        public abstract void gotLocation(Location location);
    }
}
```

WuhanMonkey / MoboSensAndroid

MoboSensAndroid / src / nano / Mobo / Sens / MyLocation.java

WuhanMonkey User simplified version   33a65ed on Feb 26, 2015

1 contributor

Executable File | 125 lines (106 sloc) | 3.96 KB

```
package nano.Mobo.Sens;

import java.util.Timer;
import java.util.  rTask;
import androi      Context;
import and         ocation;
import an          ocationListener;
import             on.LocationManager;
import   oid.os.Bundle;

public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled = false;
    boolean network_enabled = false;

    public boolean getLocation(Contex
        // I use LocationResult callb
        // MyLocation to user code.
        locationResult = result;
        if (lm == null)
            lm = (LocationManager) co
                    .getSystemService
```

perludem / DPR-KITA

DPR-KITA / app / src / main / java / id / gits / dprkita / utils / MyLocation.java

iboen init   ef06c31 on Jun 29, 2015

1 contributor

Executable File | 12   s (103 sloc) | 3.9 KB

```
package         a.utils;

/**
 *    ed by yatnosudar on 2/12/15.
 */

import android.content.Context;
import android.location.Location;
import android.location.LocationListener;
import android.location.LocationManager;
import android.os.Bundle;

import java.util.Timer;
import java.util.TimerTask;

public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled = false;
    boolean network_enabled = false;

    public boolean getLocation(Context context, LocationResult result) {
        //I use LocationResult callback class to pass location value from MyLocation to user code.
        locationResult = result;
        if (lm == null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERVICE);
```

pacosal / ownmdm

ownmdm / src / com / paco

pacosal version 2.0.0

1 contributor

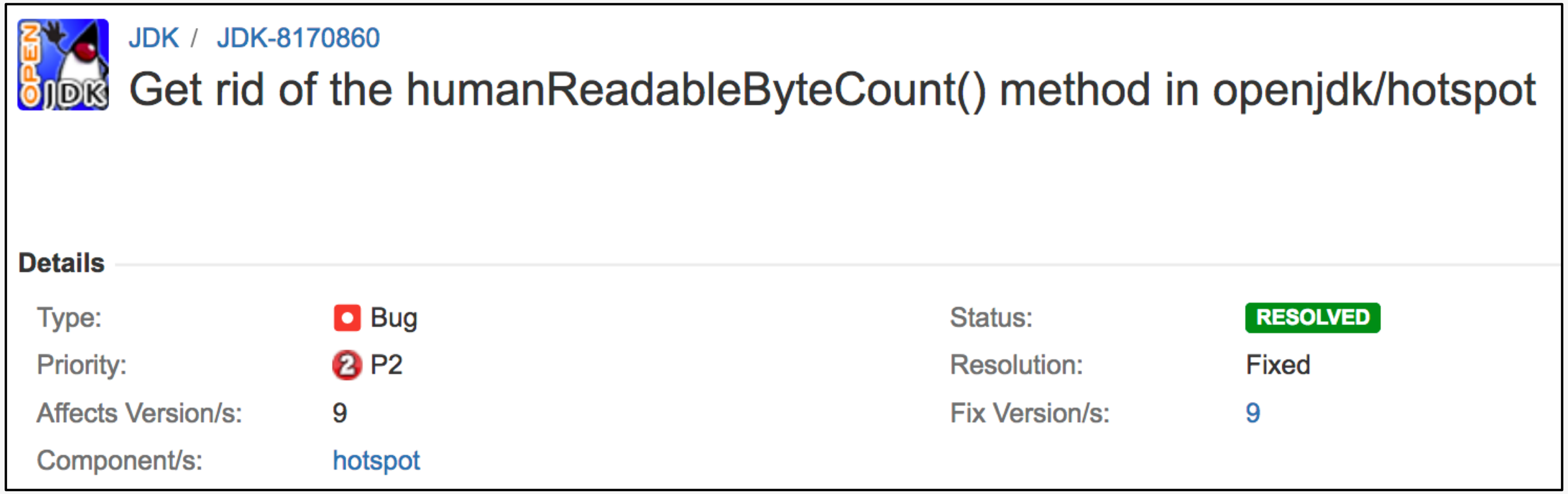
ines (85 sloc

```
 ckage com.pacosal.mdm;

import java.util.Timer;
import java.util.TimerTask;
import android.content.Context;
import android.location.Location;
import android.location.LocationListener;
import android.location.LocationManager;
import android.os.Bundle;

public class MyLocation {
    Timer timer1;
    LocationManager lm;
    LocationResult locationResult;
    boolean gps_enabled=false;
    boolean network_enabled=false;

    public boolean getLocation(Context context, LocationResult result)
    {
        locationResult=result;
        if(lm==null)
            lm = (LocationManager) context.getSystemService(Context.LOCATION_SERVICE);
```

https://github.com/pacosal/ownmdm/blob/master/src/com/pacosal/mdm/MyLocation.java

https://stackoverflow.com/a/3145655

Stack Overflow Code in the OpenJDK

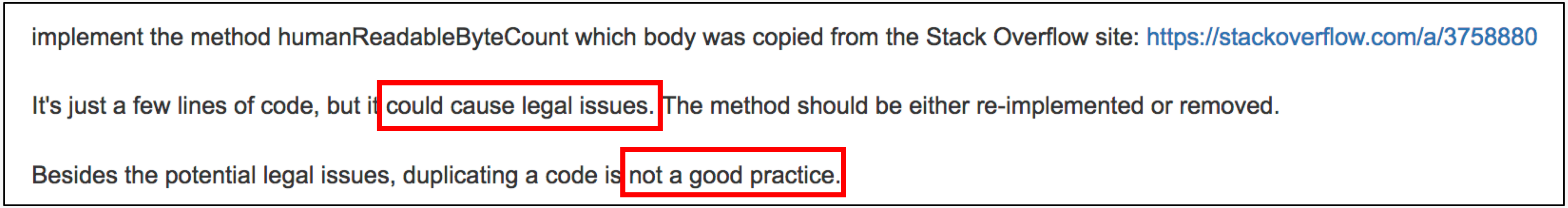


JDK / JDK-8170860

Get rid of the humanReadableByteCount() method in openjdk/hotspot

Details

Type:	Bug	Status:	RESOLVED
Priority:	P2	Resolution:	Fixed
Affects Version/s:	9	Fix Version/s:	9
Component/s:	hotspot		



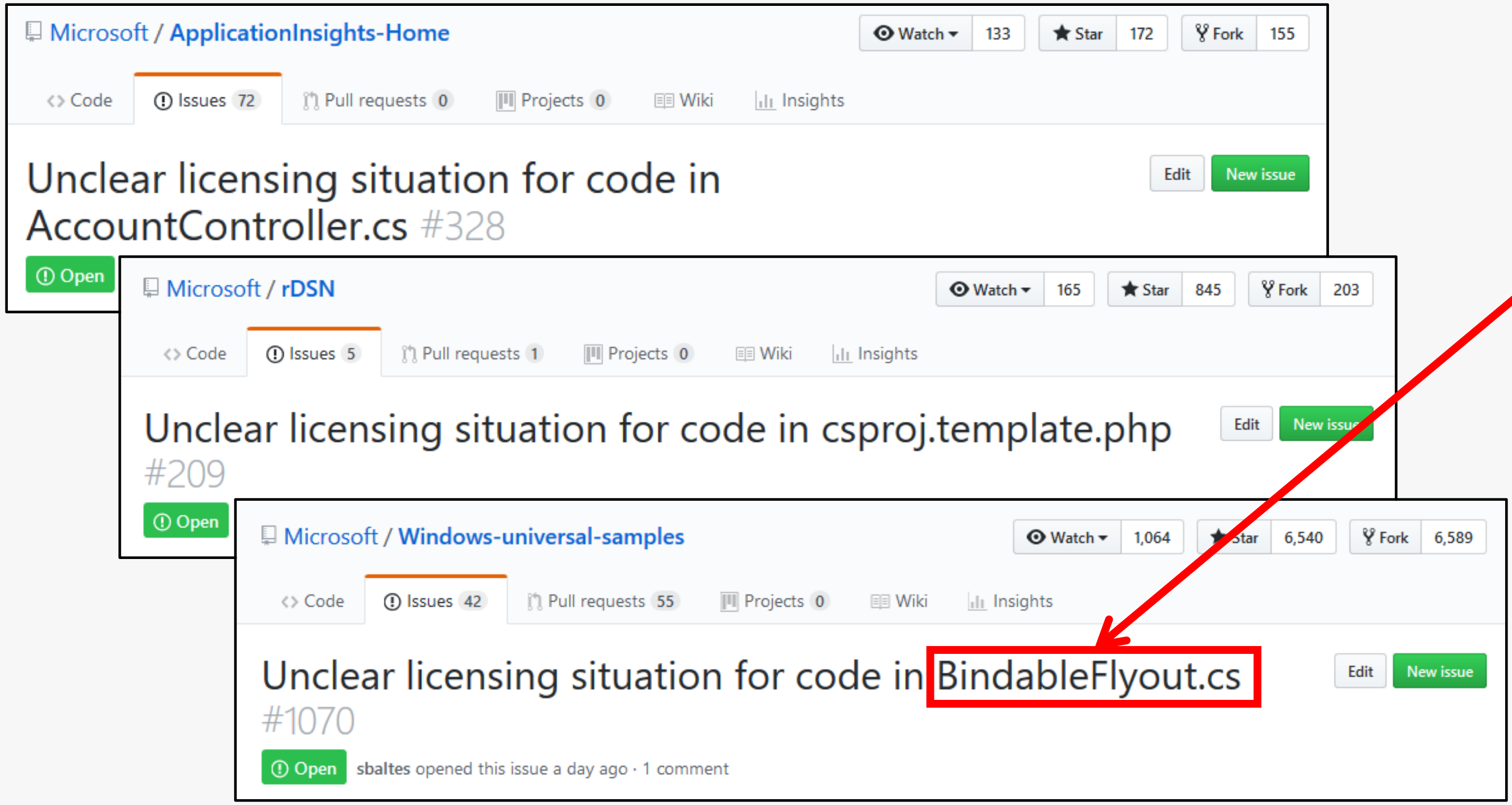
implement the method humanReadableByteCount which body was copied from the Stack Overflow site: https://stackoverflow.com/a/3758880

It's just a few lines of code, but it could cause legal issues. The method should be either re-implemented or removed.

Besides the potential legal issues, duplicating a code is not a good practice.

https://bugs.openjdk.java.net/browse/JDK-8170860

… and in Microsoft GitHub Repos



Microsoft / ApplicationInsights-Home

Unclear licensing situation for code in AccountController.cs #328

ⓘ Open

Microsoft / rDSN

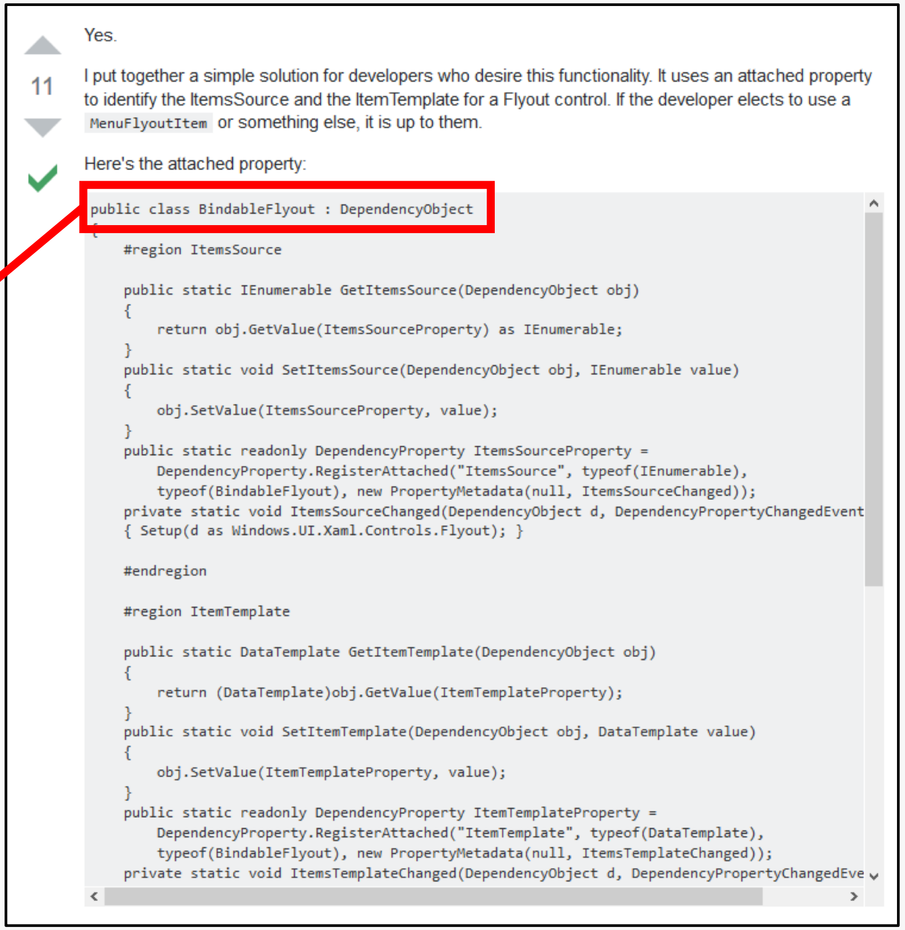
Unclear licensing situation for code in csproj.template.php #209

ⓘ Open

Microsoft / Windows-universal-samples

Unclear licensing situation for code in BindableFlyout.cs #1070

ⓘ Open sbaltes opened this issue a day ago · 1 comment



Yes.

I put together a simple solution for developers who desire this functionality. It uses an attached property to identify the ItemsSource and the ItemTemplate for a Flyout control. If the developer elects to use a `MenuFlyoutItem` or something else, it is up to them.

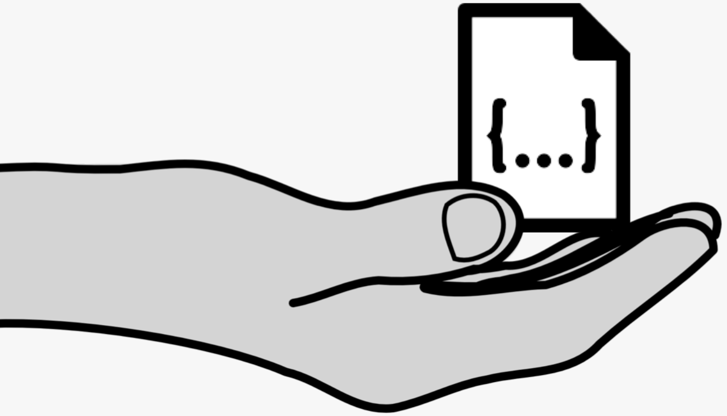
Here's the attached property:

```
public class BindableFlyout : DependencyObject
{
    #region ItemsSource

    public static IEnumerable GetItemsSource(DependencyObject obj)
    {
        return obj.GetValue(ItemsSourceProperty) as IEnumerable;
    }
    public static void SetItemsSource(DependencyObject obj, IEnumerable value)
    {
        obj.SetValue(ItemsSourceProperty, value);
    }
    public static readonly DependencyProperty ItemsSourceProperty =
        DependencyProperty.RegisterAttached("ItemsSource", typeof(IEnumerable),
        typeof(BindableFlyout), new PropertyMetadata(null, ItemsSourceChanged));
    private static void ItemsSourceChanged(DependencyObject d, DependencyPropertyChangedEvent
    { Setup(d as Windows.UI.Xaml.Controls.Flyout); }

    #endregion

    #region ItemTemplate

    public static DataTemplate GetItemTemplate(DependencyObject obj)
    {
        return (DataTemplate)obj.GetValue(ItemTemplateProperty);
    }
    public static void SetItemTemplate(DependencyObject obj, DataTemplate value)
    {
        obj.SetValue(ItemTemplateProperty, value);
    }
    public static readonly DependencyProperty ItemTemplateProperty =
        DependencyProperty.RegisterAttached("ItemTemplate", typeof(DataTemplate),
        typeof(BindableFlyout), new PropertyMetadata(null, ItemsTemplateChanged));
    private static void ItemsTemplateChanged(DependencyObject d, DependencyPropertyChangedEve
```

Implications of Stack Overflow's License

Permissive Licenses

- Permit using the licensed source code in proprietary software **without publishing changes** or the **derived work**
- *Examples:* MIT, Apache, and BSD license families



Copyleft Licenses

- Requires either modifications to the licensed content or the complete derived work to be **published under the same or a compatible license** (share-alike)
- *Examples (weak copyleft):* Mozilla/Eclipse Public Licenses
- *Examples (viral copyleft):* GNU General Public Licenses, Creative Commons Share-Alike Licenses (e.g., **CC BY-SA**)

Enforceability of Copyleft Licenses

- Courts in the US and Europe ruled that open source licenses are **enforceable contracts**
- Authors are able to **sue** when terms such as the share-alike requirement are violated:
  - **Interdict distribution** of derived work
  - **Claim monetary damages**
- USA: DMCA takedown notices for allegedly infringed copyright
  - Example: https://github.com/github/dmca
- Risk in mergers and acquisitions of companies
  - Example: FSF vs. Cisco lawsuit

Research Question

Question:

How **frequently** is code from Stack Overflow posts used in public GitHub projects **without** the required **attribution**?

Approach:

Triangulate an estimate for the attribution ratio using three different methods.

Method 1: Regular Expressions

Google Cloud GitHub

209m files in 4.1m projects

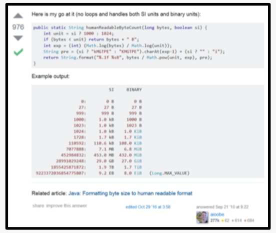
.java →

13m Java files in 336k projects

...stackoverflow\.com... →

10 most frequently referenced answers

Check external availability



Here is my go at it (no loops and handles both SI units and binary units):

976

```
public static String humanReadableByteCount(long bytes, boolean si) {
    int unit = si ? 1000 : 1024;
    if (bytes < unit) return bytes + " B";
    int exp = (int) (Math.log(bytes) / Math.log(unit));
    String pre = (si ? "kMGTPE" : "KMGTPE").charAt(exp-1) + (si ? "" : "i");
    return String.format("%.1f %sB", bytes / Math.pow(unit, exp), pre);
}
```

Example output:

```
                             SI     BINARY

                   0:       0 B       0 B
                  27:      27 B      27 B
                 999:     999 B     999 B
                1000:    1.0 kB    1000 B
                1023:    1.0 kB    1023 B
                1024:    1.0 kB   1.0 KiB
                1728:    1.7 kB   1.7 KiB
              110592:  110.6 kB 108.0 KiB
             7077888:    7.1 MB   6.8 MiB
           452984832:  453.0 MB 432.0 MiB
         28991029248:   29.0 GB  27.0 GiB
       1855425871872:    1.9 TB   1.7 TiB
 9223372036854775807:    9.2 EB   8.0 EiB   (Long.MAX_VALUE)
```

Related article: Java: Formatting byte size to human readable format

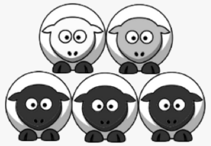
share improve this answer
edited Oct 29 '16 at 3:58
answered Sep 21 '10 at 9:22
aioobe
277k ● 62 ● 614 ● 684

Manually build regular expressions matching code snippets
(referenced usages as test cases)

```
((?i:String[\s]+\w+\([^\{]*long[^\{]+\)[\s]*\{[\s\S]+if
[\s]*\([^<]+<[^\)]+\)[\s\S]*return[^;]+\+[^;]*\"\ B\"
[\s\S]+int[\s][^\=]+\=[\s]*\([\s]*int[\s]*\)[\s]*\([\s]*
Math[\s]*\.[\s]*log[\s]*\([^\)]+\)[\s]*\/[\s]*Math[\s]*
\.[\s]*log[\s]*\([^\)]+\)[\s]*\)[\s\S]+return[^\}]+
String[\s]*\.[\s]*format[\s]*\([^\}]+\}))
```

←

4,198 files with matches



←

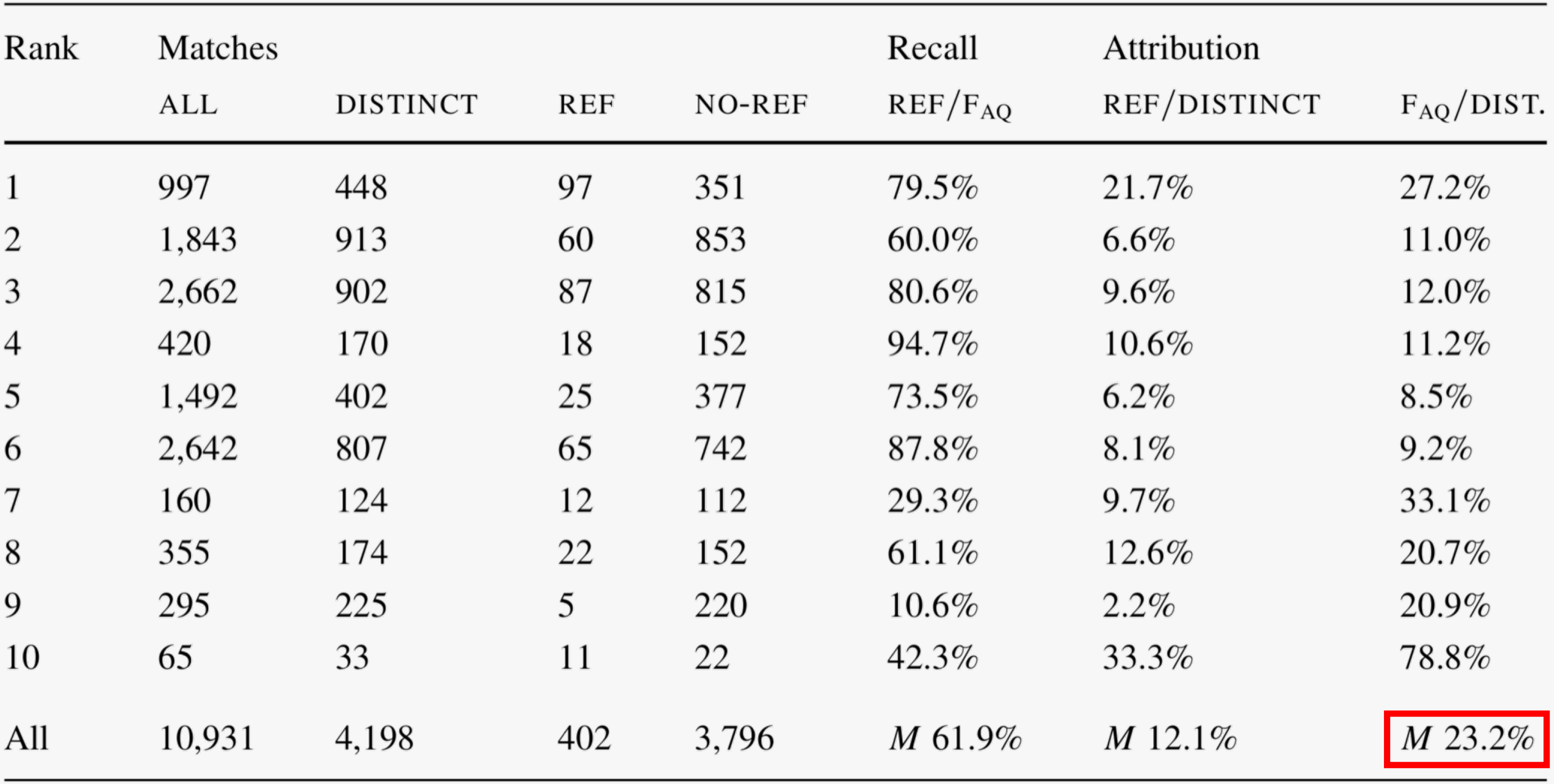
BY

Check if attributed
(URL to answer or corresponding question)

Check for false positives

https://iwsc2018.github.io/assets/img/sheep.png

Results



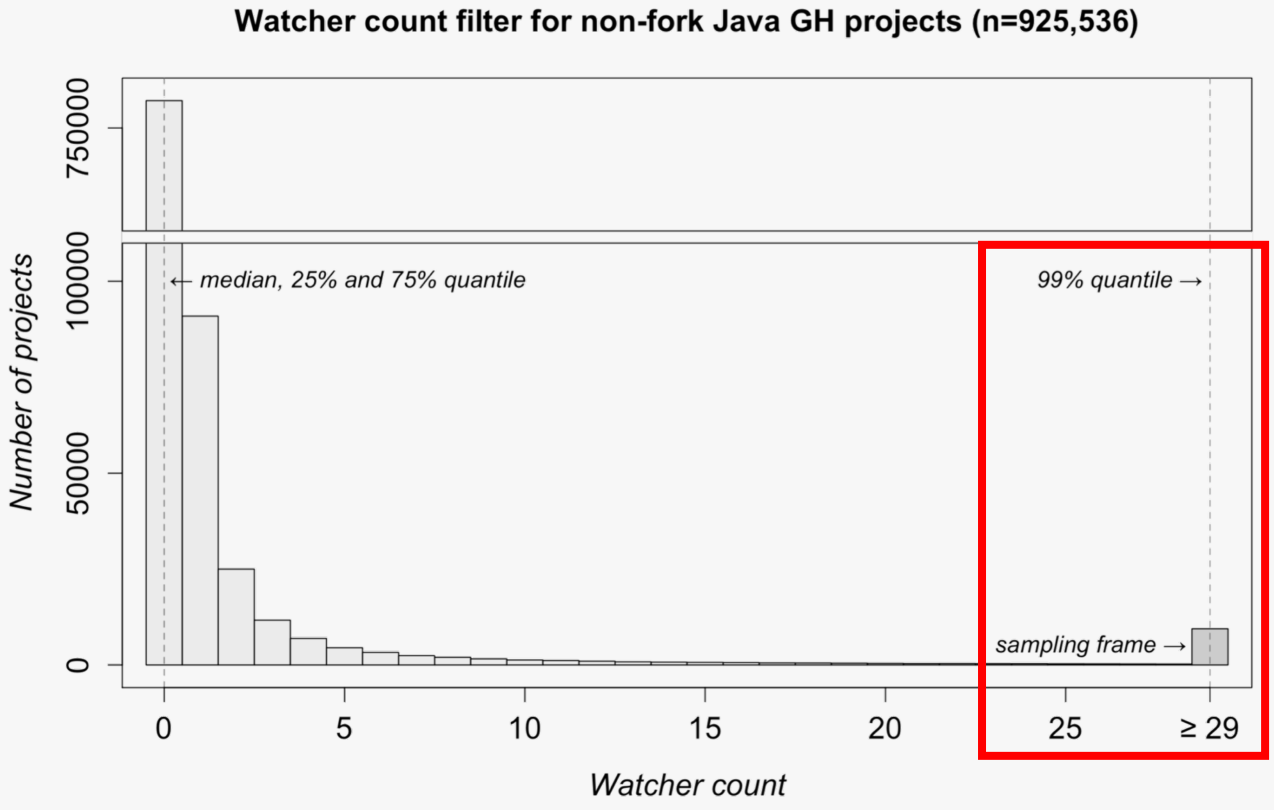
Rank	Matches				Recall	Attribution	
	ALL	DISTINCT	REF	NO-REF	REF/F_{AQ}	REF/DISTINCT	F_{AQ}/DIST.
1	997	448	97	351	79.5%	21.7%	27.2%
2	1,843	913	60	853	60.0%	6.6%	11.0%
3	2,662	902	87	815	80.6%	9.6%	12.0%
4	420	170	18	152	94.7%	10.6%	11.2%
5	1,492	402	25	377	73.5%	6.2%	8.5%
6	2,642	807	65	742	87.8%	8.1%	9.2%
7	160	124	12	112	29.3%	9.7%	33.1%
8	355	174	22	152	61.1%	12.6%	20.7%
9	295	225	5	220	10.6%	2.2%	20.9%
10	65	33	11	22	42.3%	33.3%	78.8%
All	10,931	4,198	402	3,796	*M* 61.9%	*M* 12.1%	*M* 23.2%

Method 2: Code Clone Detector

- **Goal:** Use code clone detector to find clones of a sample of Stack Overflow snippets in a sample of GitHub projects
- *Why samples?*
  - Code clone detection is computationally expensive
- *Which snippets and projects to select?*
  - Random samples: Many **toy projects** on GitHub and many **irrelevant snippets** on Stack Overflow
  - Purposive sampling: Limited generalizability

GitHub Project Sample

- Focus on **popular** GitHub projects
- High precision in selecting “engineered” software projects [Munaiah et al. 2017]
- Greater (potential) impact of licensing issues



Sample size:
3,000 / 2,313

Stack Overflow Snippet Samples

- Non-trivial snippets retrieved from 100 most frequently referenced answers (n=111)

  $\Rightarrow \; S_{\text{top100}}$

- Non-trivial snippets retrieved from answers referenced in GitHub projects (n=137)

  $\Rightarrow \; S_{\text{gh}}$

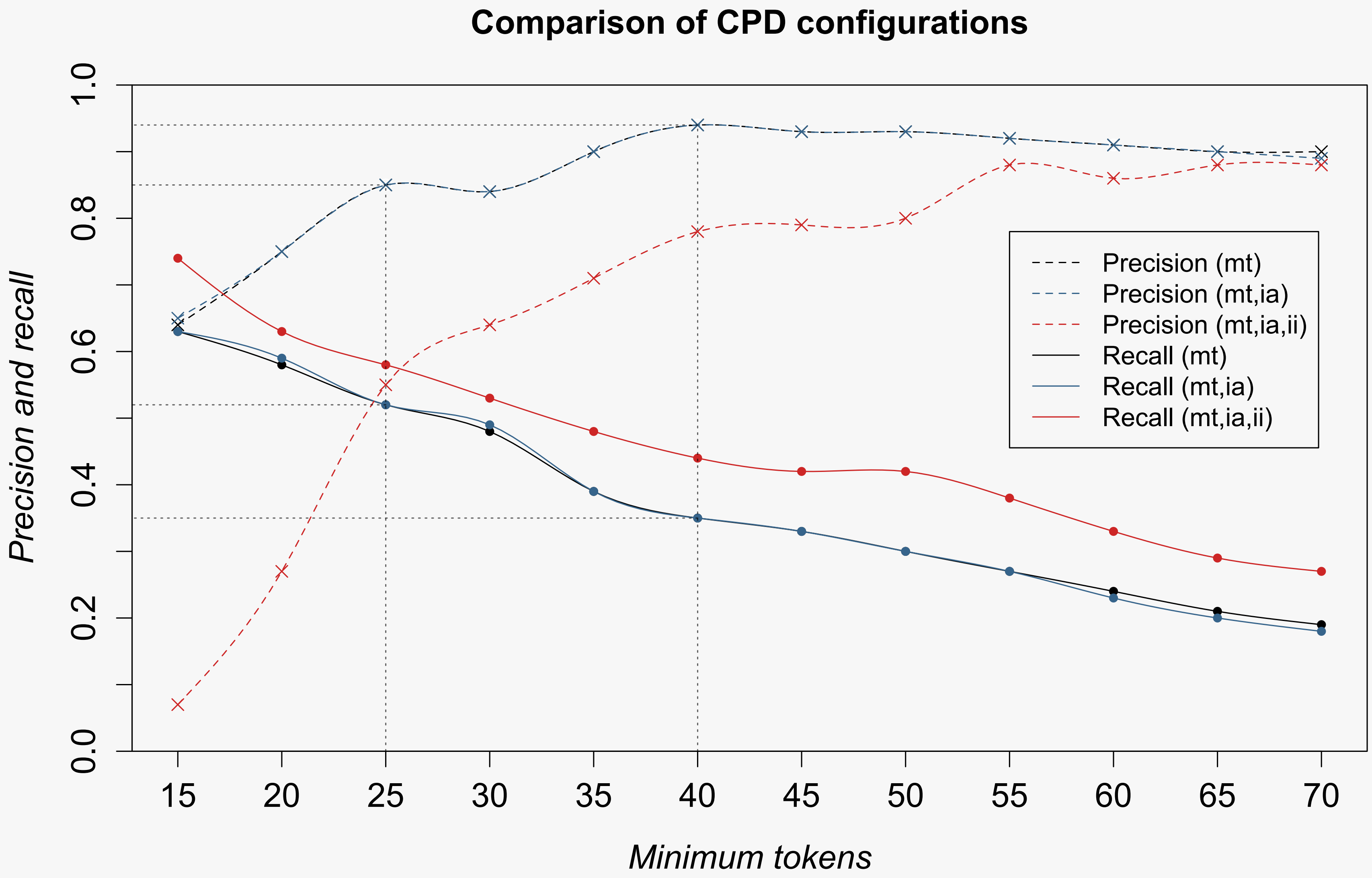
- *External sources:* Only three snippets available under a more permissive license than CC BY-SA

stackoverflow

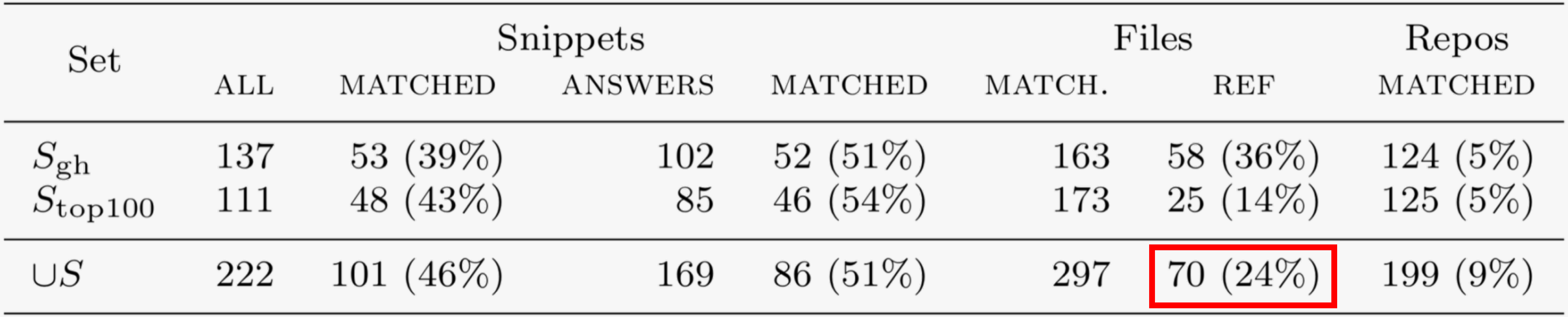
Code Clone Detector Calibration

DON'T SHOOT THE MESSENGER

https://pmd.github.io/



Results

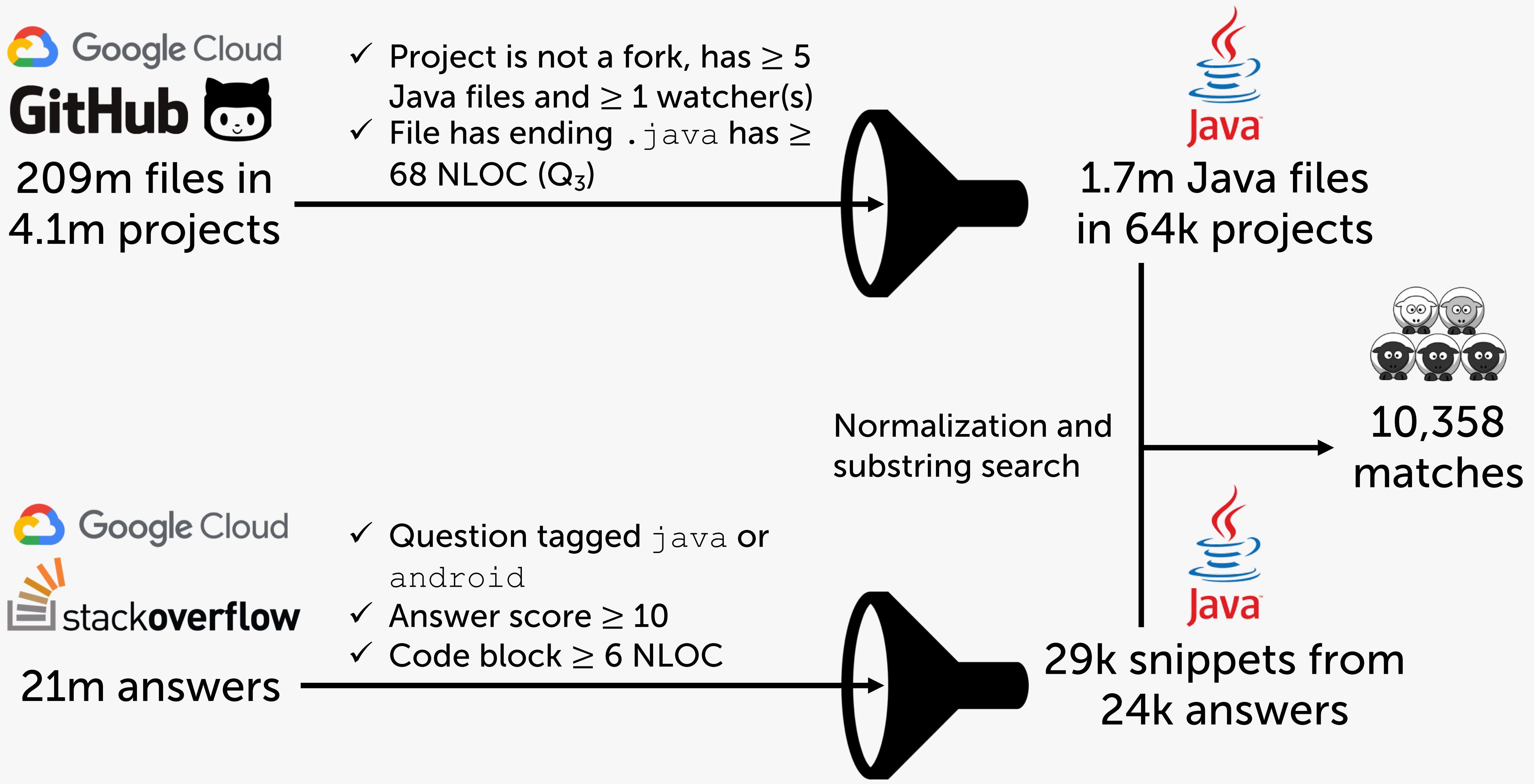


Set	Snippets				Files		Repos
	ALL	MATCHED	ANSWERS	MATCHED	MATCH.	REF	MATCHED
S_{gh}	137	53 (39%)	102	52 (51%)	163	58 (36%)	124 (5%)
S_{top100}	111	48 (43%)	85	46 (54%)	173	25 (14%)	125 (5%)
$\cup S$	222	101 (46%)	169	86 (51%)	297	70 (24%)	199 (9%)

Method 3: Exact Matches

- **Goal:** Address shortcomings of Method 1 and 2
  - Increase sample sizes
  - Exclude snippets available on external sources
  - Systematically exclude short snippets
- Select as many projects and snippets as possible and search for (almost) exact matches

Method 3: Exact Matches



Google Cloud
GitHub

209m files in
4.1m projects

- ✓ Project is not a fork, has ≥ 5 Java files and ≥ 1 watcher(s)
- ✓ File has ending `.java` has ≥ 68 NLOC (Q_3)

Java

1.7m Java files
in 64k projects

Normalization and
substring search

10,358
matches

Google Cloud
stackoverflow

21m answers

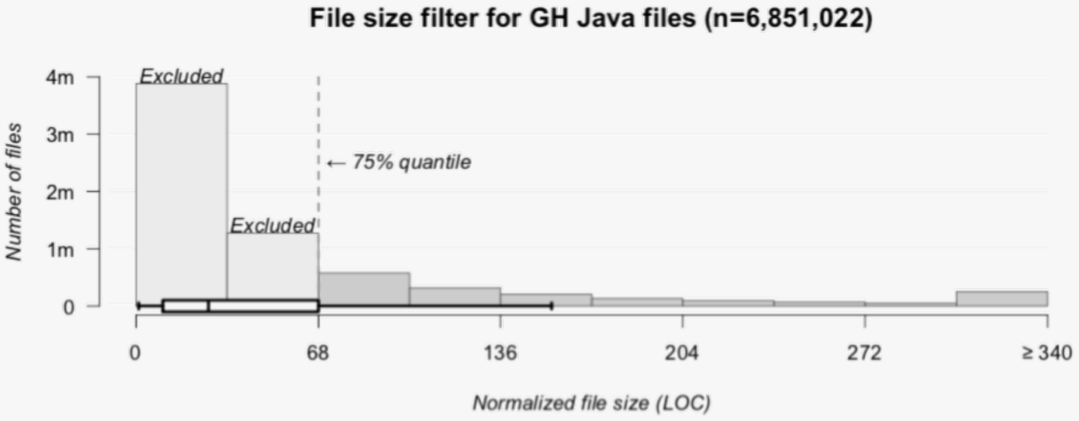
- ✓ Question tagged `java` or `android`
- ✓ Answer score ≥ 10
- ✓ Code block ≥ 6 NLOC

Java

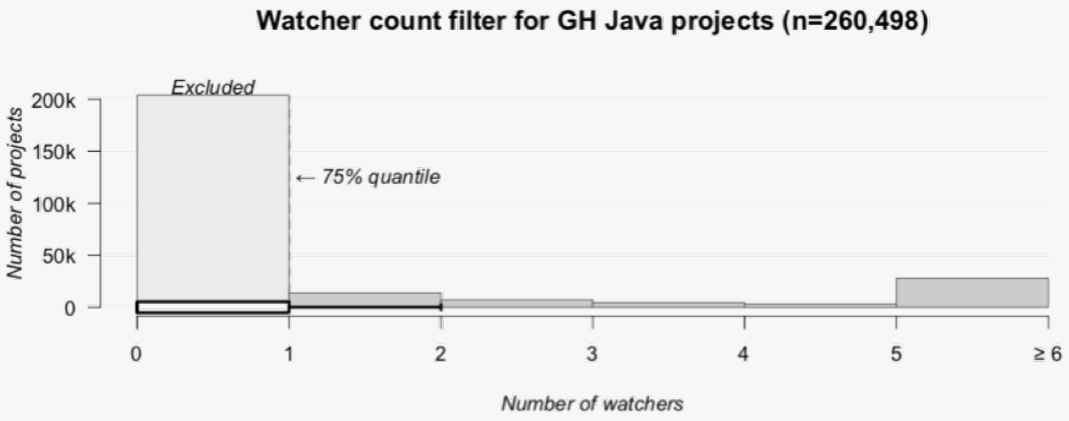
29k snippets from
24k answers

https://iwsc2018.github.io/assets/img/sheep.png

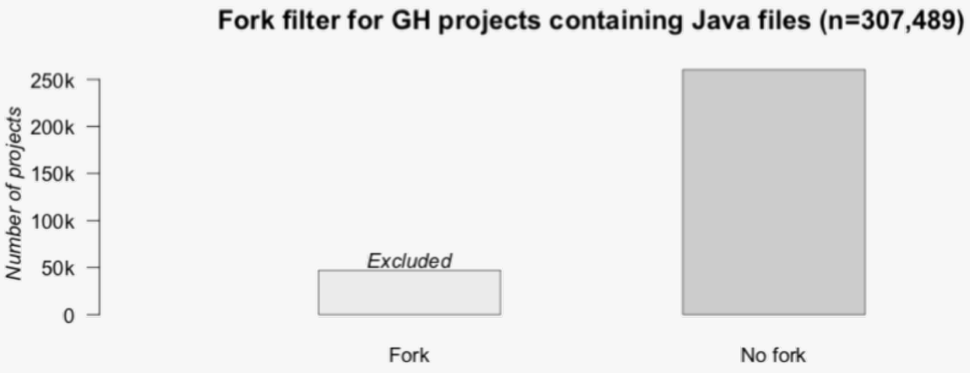
Details: Filtering of GitHub Projects



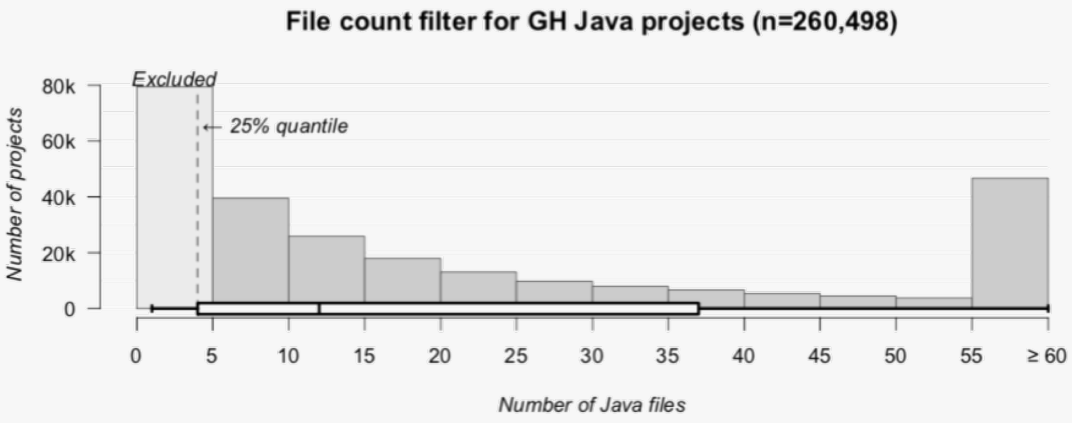
File size filter for GH Java files (n=6,851,022)



Watcher count filter for GH Java projects (n=260,498)

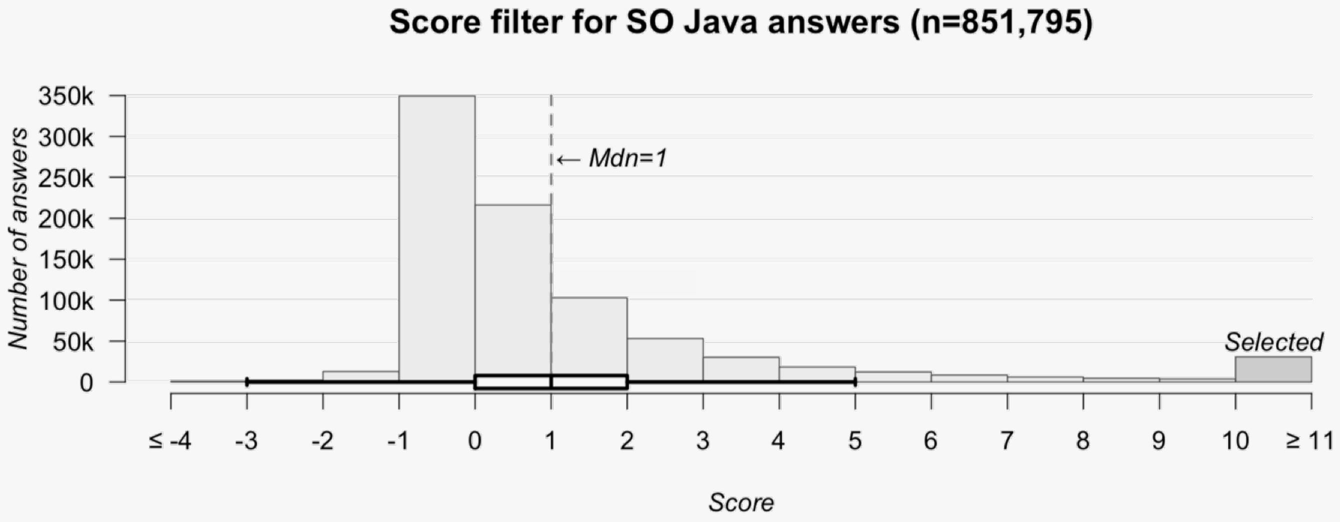


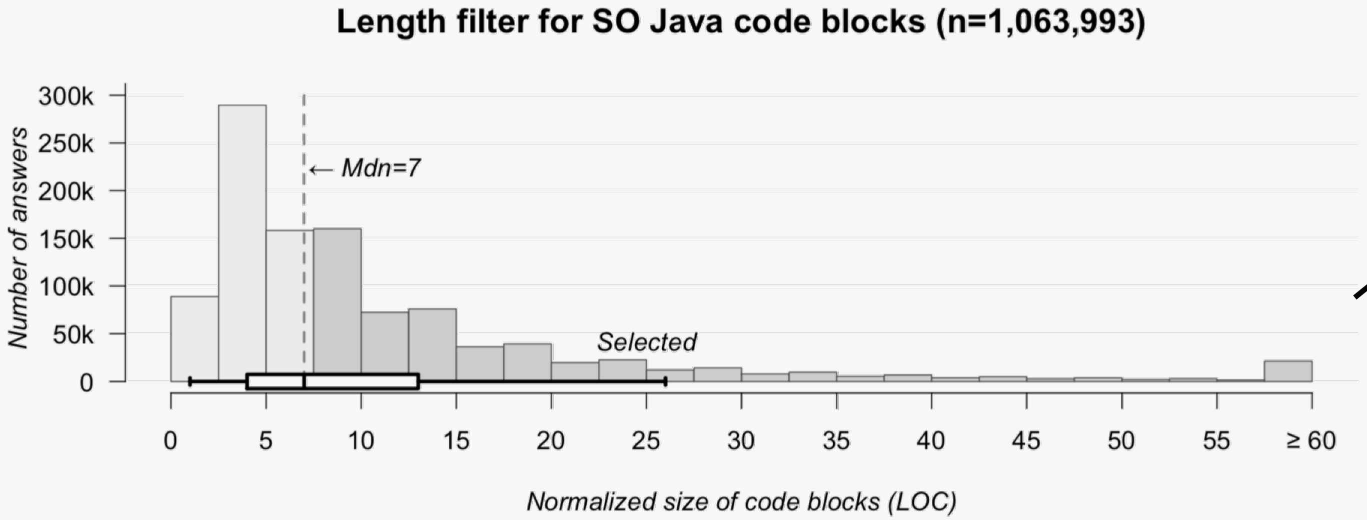
Fork filter for GH projects containing Java files (n=307,489)



File count filter for GH Java projects (n=260,498)

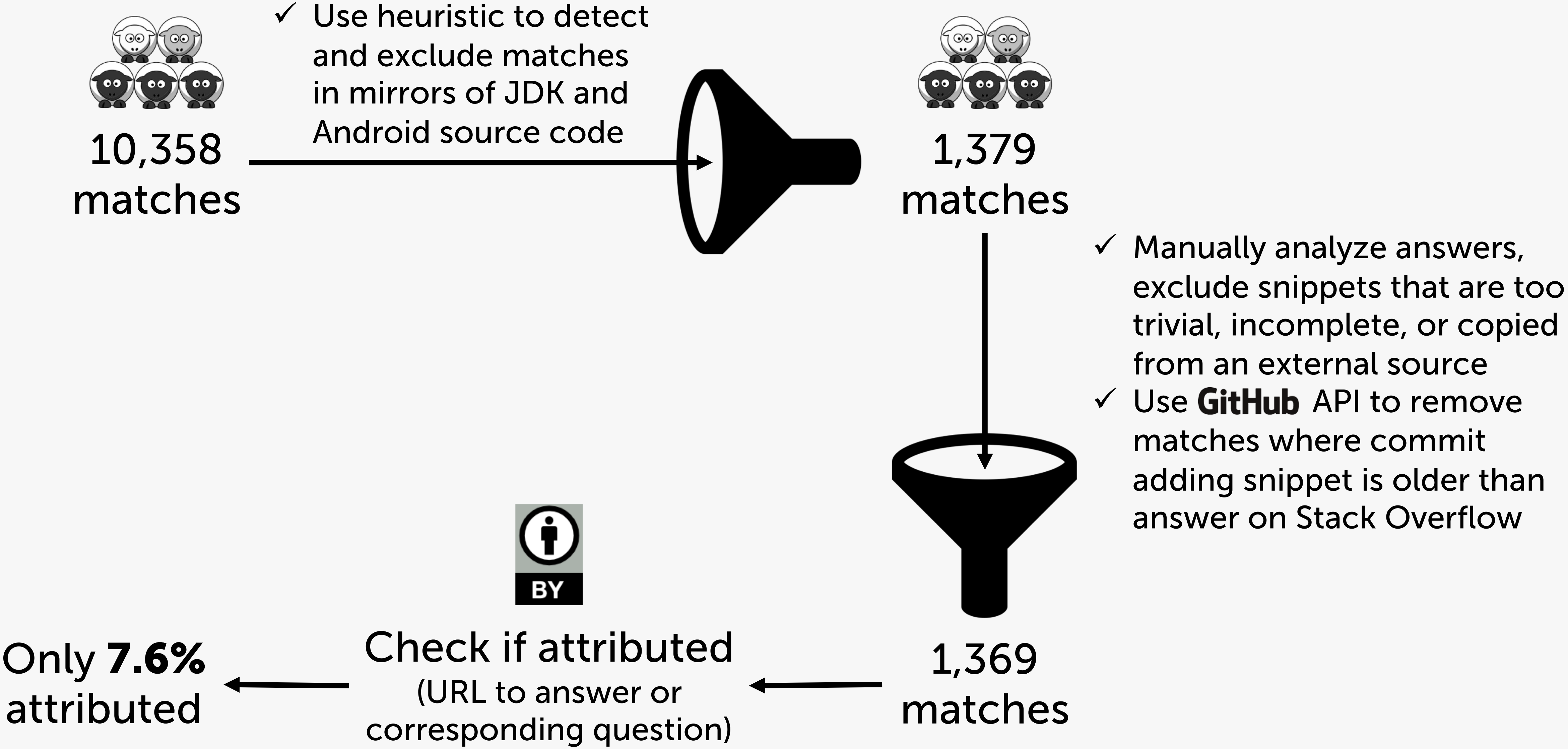
Details: Filtering of Stack Overflow Snippets





Proxies for originality

Method 3: Filtering of Matches



10,358 matches

✓ Use heuristic to detect and exclude matches in mirrors of JDK and Android source code

1,379 matches

✓ Manually analyze answers, exclude snippets that are too trivial, incomplete, or copied from an external source

✓ Use GitHub API to remove matches where commit adding snippet is older than answer on Stack Overflow

1,369 matches

Check if attributed

(URL to answer or corresponding question)

Only **7.6%** attributed

Attribution

Attribution ratio:

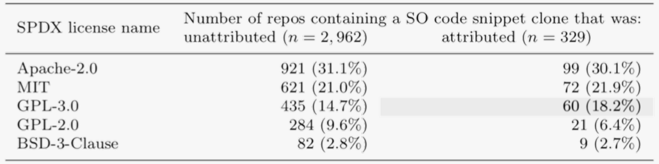
- Method 1 (regular expressions): 23 %
- Method 2 (code clone detector): 24 %
- Method 3 (exact matches): 8 %

Conservative estimate:

- **Attribution ratio** $\leq$ **25%**

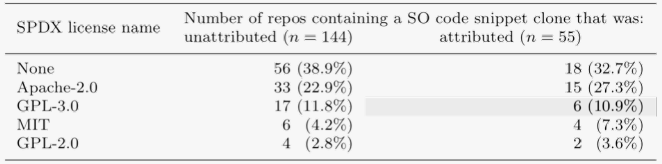
Share-alike

Only **2%** of all analyzed repositories (all methods) containing code from Stack Overflow **attributed** its source and used a **compatible license** (not CC BY-SA, but GPL 3.0).



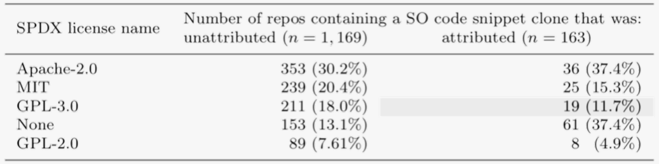
SPDX license name	Number of repos containing a SO code snippet clone that was: unattributed ($n = 2,962$)	attributed ($n = 329$)
Apache-2.0	921 (31.1%)	99 (30.1%)
MIT	621 (21.0%)	72 (21.9%)
GPL-3.0	435 (14.7%)	60 (18.2%)
GPL-2.0	284 (9.6%)	21 (6.4%)
BSD-3-Clause	82 (2.8%)	9 (2.7%)

Method 1



SPDX license name	Number of repos containing a SO code snippet clone that was: unattributed ($n = 144$)	attributed ($n = 55$)
None	56 (38.9%)	18 (32.7%)
Apache-2.0	33 (22.9%)	15 (27.3%)
GPL-3.0	17 (11.8%)	6 (10.9%)
MIT	6 (4.2%)	4 (7.3%)
GPL-2.0	4 (2.8%)	2 (3.6%)

Method 2



SPDX license name	Number of repos containing a SO code snippet clone that was: unattributed ($n = 1,169$)	attributed ($n = 163$)
Apache-2.0	353 (30.2%)	36 (37.4%)
MIT	239 (20.4%)	25 (15.3%)
GPL-3.0	211 (18.0%)	19 (11.7%)
None	153 (13.1%)	61 (37.4%)
GPL-2.0	89 (7.61%)	8 (4.9%)

Method 3

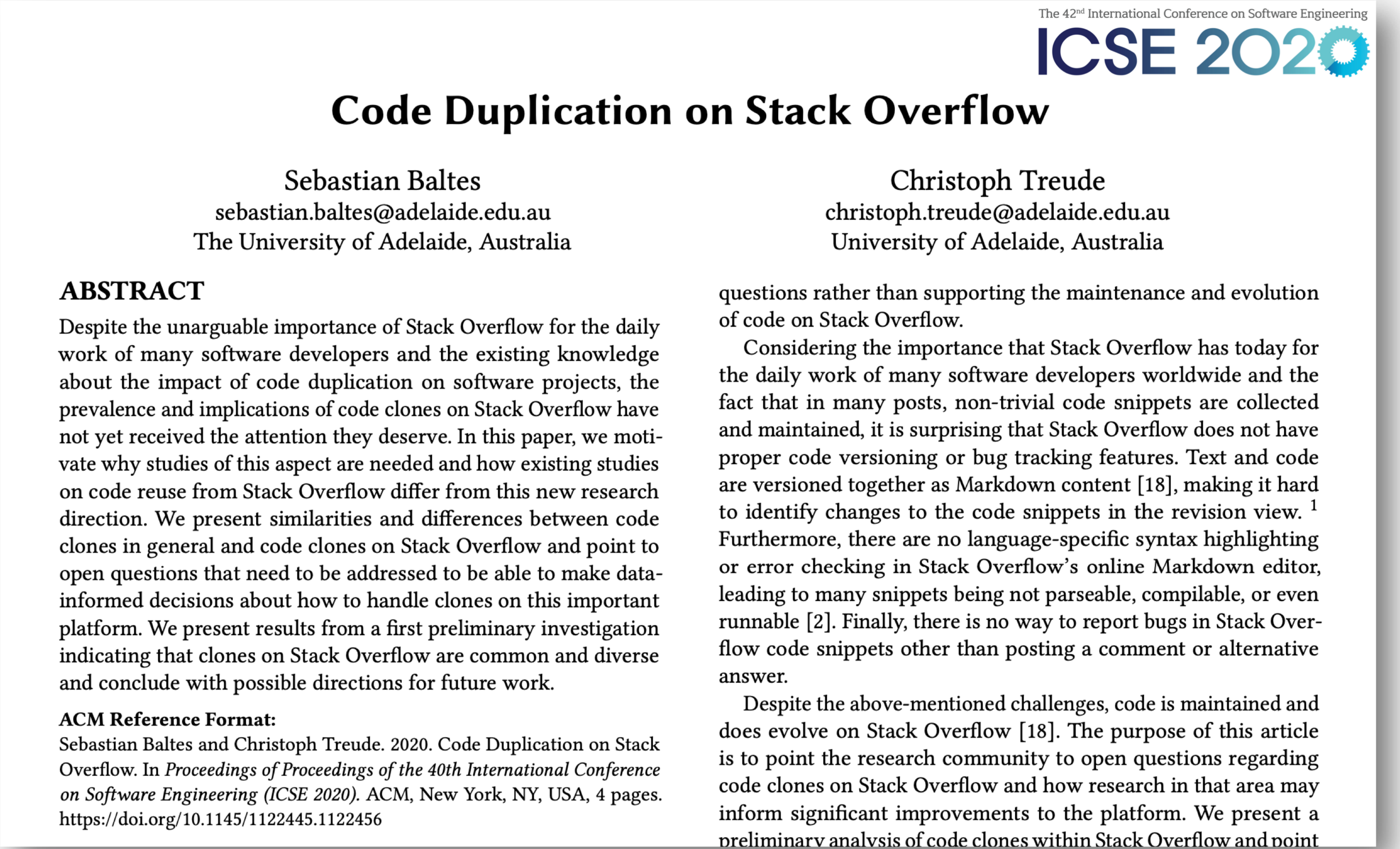
Reaching out to Developers

- **Contacted owners** of GitHub repositories containing copies of Stack Overflow snippets
- **75% not aware** of CC BY-SA licensing (see slide about online surveys)
- Many thankful responses

Future Work

- *Tool support*: Support **maintainability** of copied snippets by automatically adding links to sources, integration into CI tools
- *Education:* Help developers **understand complex licensing situations** (not only for complete libraries but also for individual snippets)
- *Study:* Analyze links to better understand Stack Overflow’s role in the **ecosystem** of documentation resources

Code Plagiarism



The 42nd International Conference on Software Engineering
ICSE 2020

Code Duplication on Stack Overflow

Sebastian Baltes
sebastian.baltes@adelaide.edu.au
The University of Adelaide, Australia

Christoph Treude
christoph.treude@adelaide.edu.au
University of Adelaide, Australia

ABSTRACT

Despite the unarguable importance of Stack Overflow for the daily work of many software developers and the existing knowledge about the impact of code duplication on software projects, the prevalence and implications of code clones on Stack Overflow have not yet received the attention they deserve. In this paper, we motivate why studies of this aspect are needed and how existing studies on code reuse from Stack Overflow differ from this new research direction. We present similarities and differences between code clones in general and code clones on Stack Overflow and point to open questions that need to be addressed to be able to make data-informed decisions about how to handle clones on this important platform. We present results from a first preliminary investigation indicating that clones on Stack Overflow are common and diverse and conclude with possible directions for future work.

ACM Reference Format:
Sebastian Baltes and Christoph Treude. 2020. Code Duplication on Stack Overflow. In *Proceedings of Proceedings of the 40th International Conference on Software Engineering (ICSE 2020).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

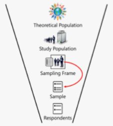
questions rather than supporting the maintenance and evolution of code on Stack Overflow.

Considering the importance that Stack Overflow has today for the daily work of many software developers worldwide and the fact that in many posts, non-trivial code snippets are collected and maintained, it is surprising that Stack Overflow does not have proper code versioning or bug tracking features. Text and code are versioned together as Markdown content [18], making it hard to identify changes to the code snippets in the revision view. [1] Furthermore, there are no language-specific syntax highlighting or error checking in Stack Overflow's online Markdown editor, leading to many snippets being not parseable, compilable, or even runnable [2]. Finally, there is no way to report bugs in Stack Overflow code snippets other than posting a comment or alternative answer.

Despite the above-mentioned challenges, code is maintained and does evolve on Stack Overflow [18]. The purpose of this article is to point the research community to open questions regarding code clones on Stack Overflow and how research in that area may inform significant improvements to the platform. We present a preliminary analysis of code clones within Stack Overflow and point

https://empirical-software.engineering/projects/snippets/

Studied Habits



Issues in Sampling Software Developers

Methodology

airbnb

Constructing Urban Tourism Space Digitally

Interdisciplinary Research

2013

Sketching

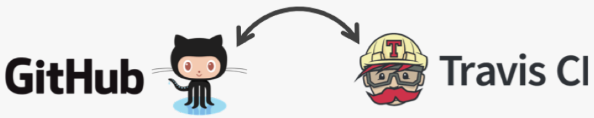
Expertise Development

Code Plagiarism

Regular Expressions

RegViz

Continuous Integration



Evidence-based Practice

through

Practice-based Evidence

THE UNIVERSITY *of* ADELAIDE

SUB CRUCE LUMEN

@s_baltes

empirical-software.engineering

Sebastian Baltes