Software Developers' Work Habits and Expertise

Empirical Studies on Sketching, Code Plagiarism, and Expertise Development

Sebastian Baltes

@s_baltes

THE UNIVERSITY of ADELAIDE

Universität Trier

empirical-software.engineering

Ein bisschen Geographie…


Evidenz-basierte Praxis durch Praxis-basierte Evidenz

Meine Dissertation

Beobachten
Beschreiben
Erklären

Arbeitsgewohnheiten von Softwareentwicklern

- Anforderungen an Entwicklungswerkzeuge
- Verbesserungen am Entwicklungsprozess
- Wissenschafts-kommunikation!

Arbeitsgewohnheiten?

DUDEN

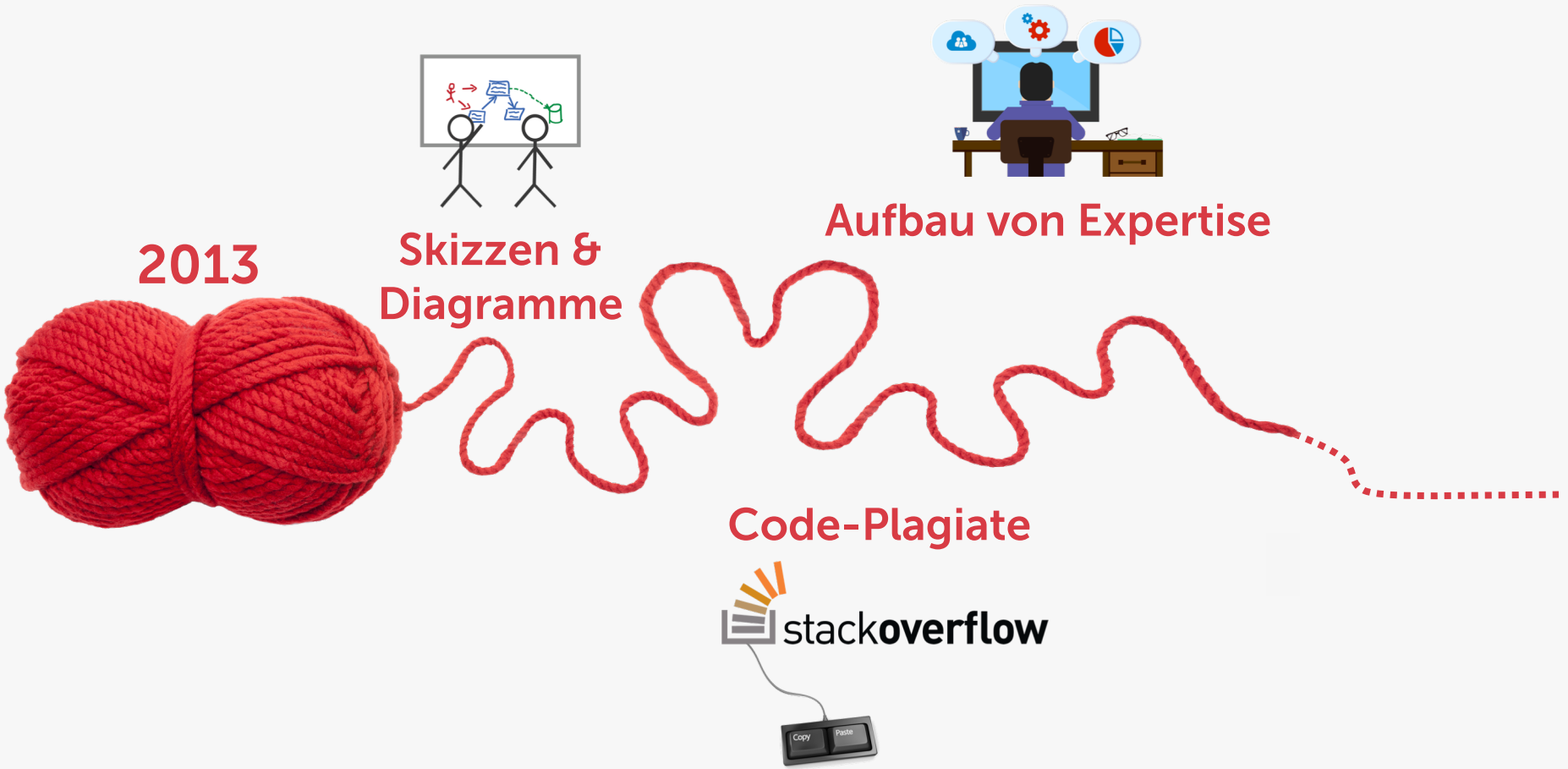
Gewohnheit: „durch häufige und stete Wiederholung **selbstverständlich** gewordene **Handlung**“

https://www.duden.de/rechtschreibung/Gewohnheit

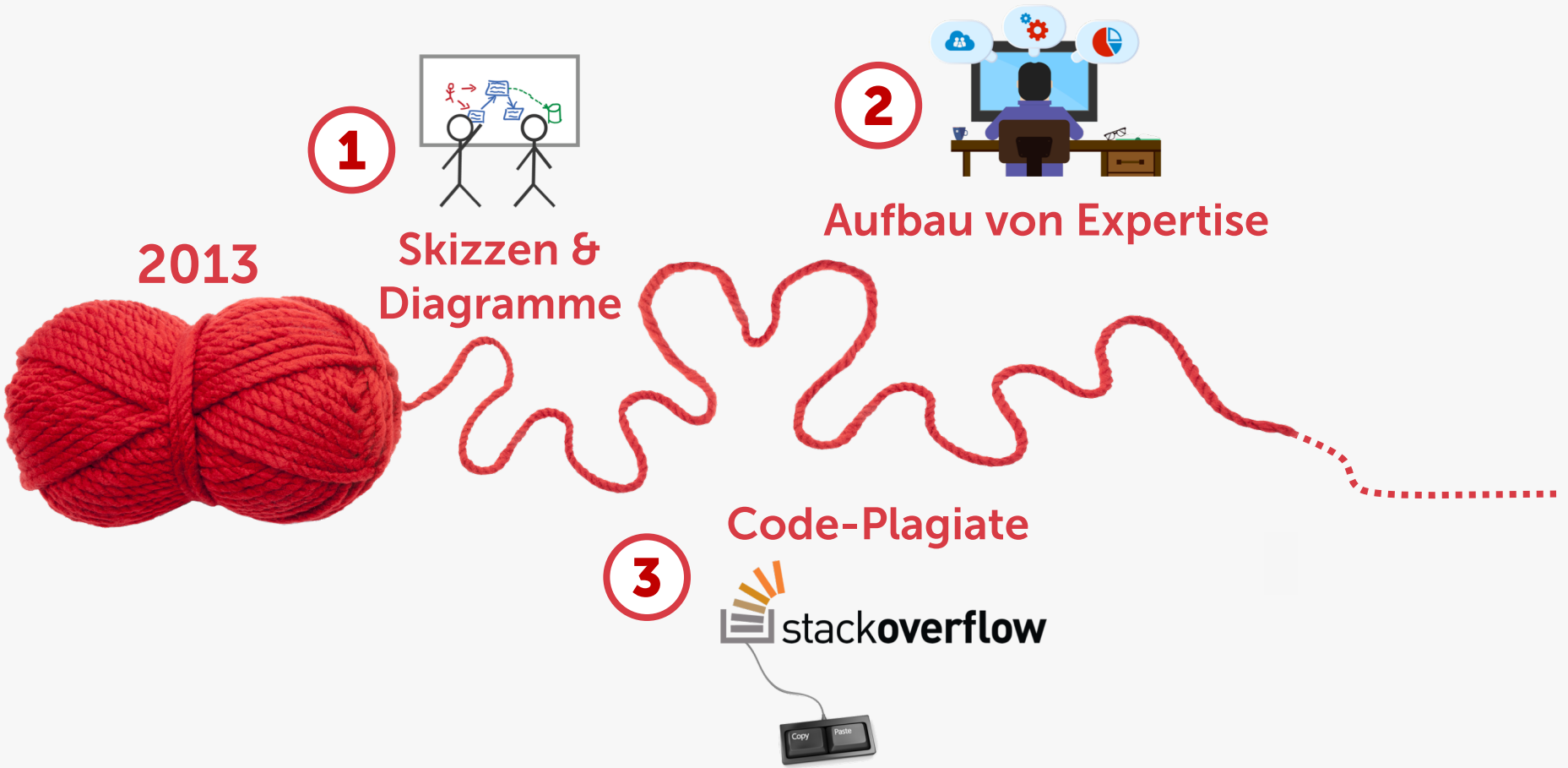
Private Gewohnheiten

Arbeitsgewohnheiten

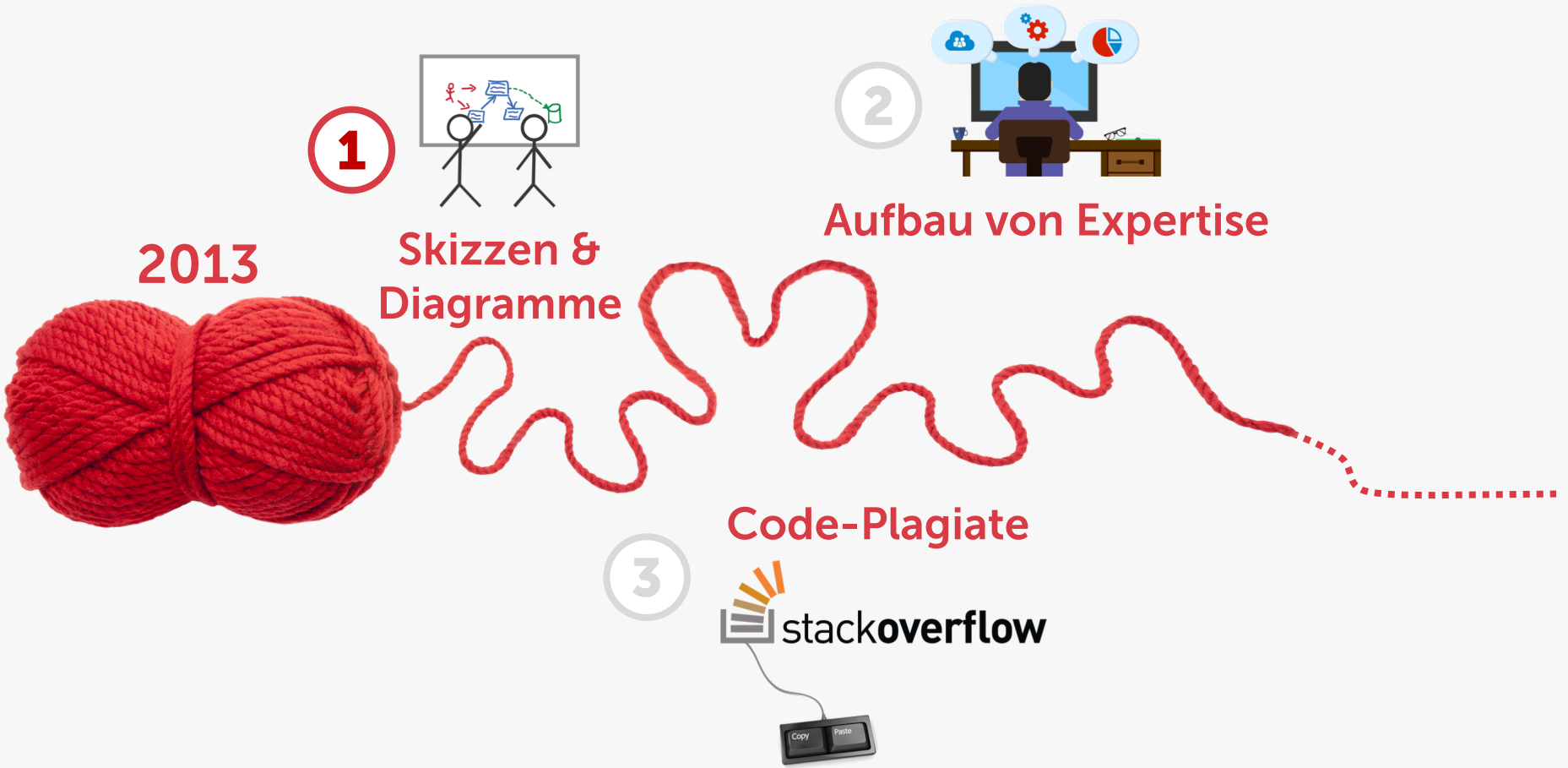
Behandelte Themenbereiche

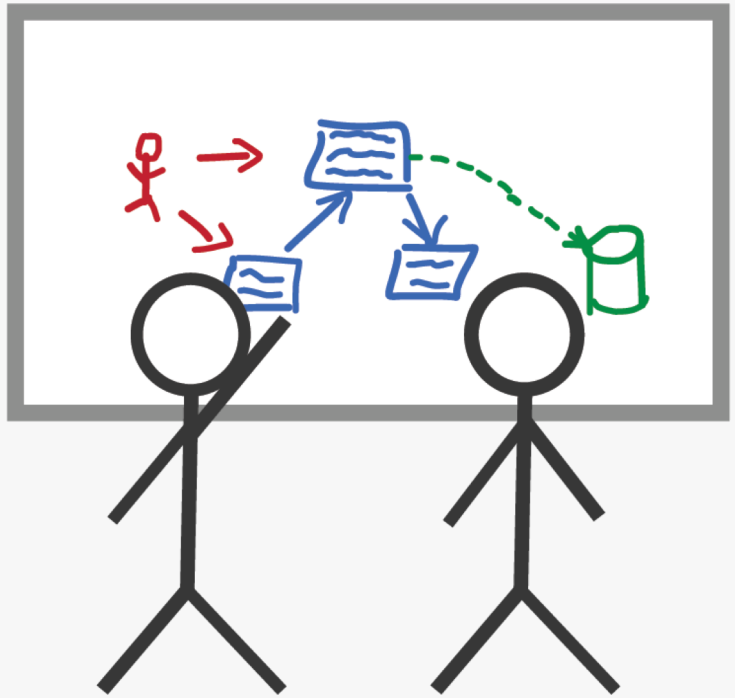


Struktur dieses Vortrags



Struktur dieses Vortrags





Skizzen & Diagramme

Skizzen & Diagramme

Forschungsfragen

Wie und **warum** verwenden Softwareentwickler/-innen **Skizzen und Diagramme**?

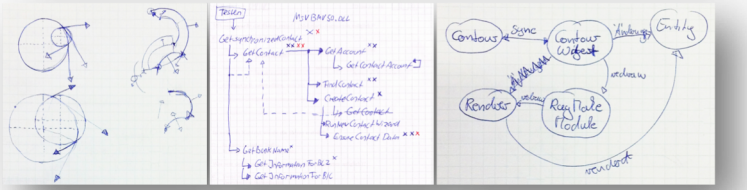
Wie ist deren **Beziehung zum Quellcode**, den sie dokumentieren?

Wie können wir Entwickler/-innen besser mit **Werkzeugen unterstützen**?

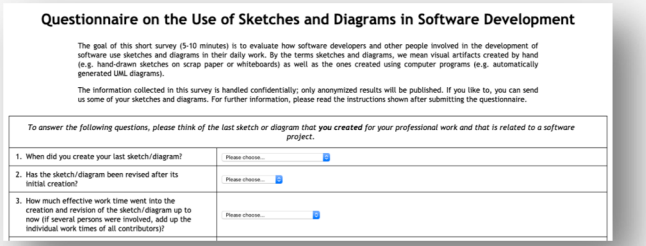
Skizzen & Diagramme

Forschungsmethoden

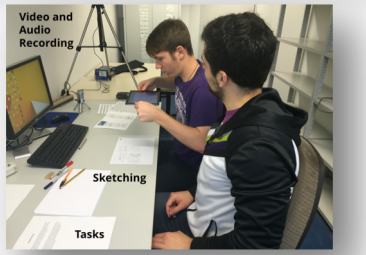
- Feldstudie



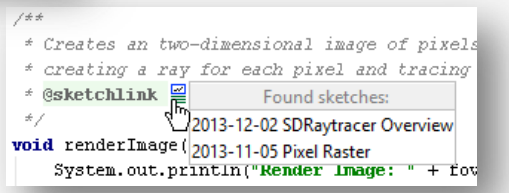
- Onlineumfrage



- Laborexperiment

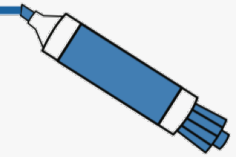


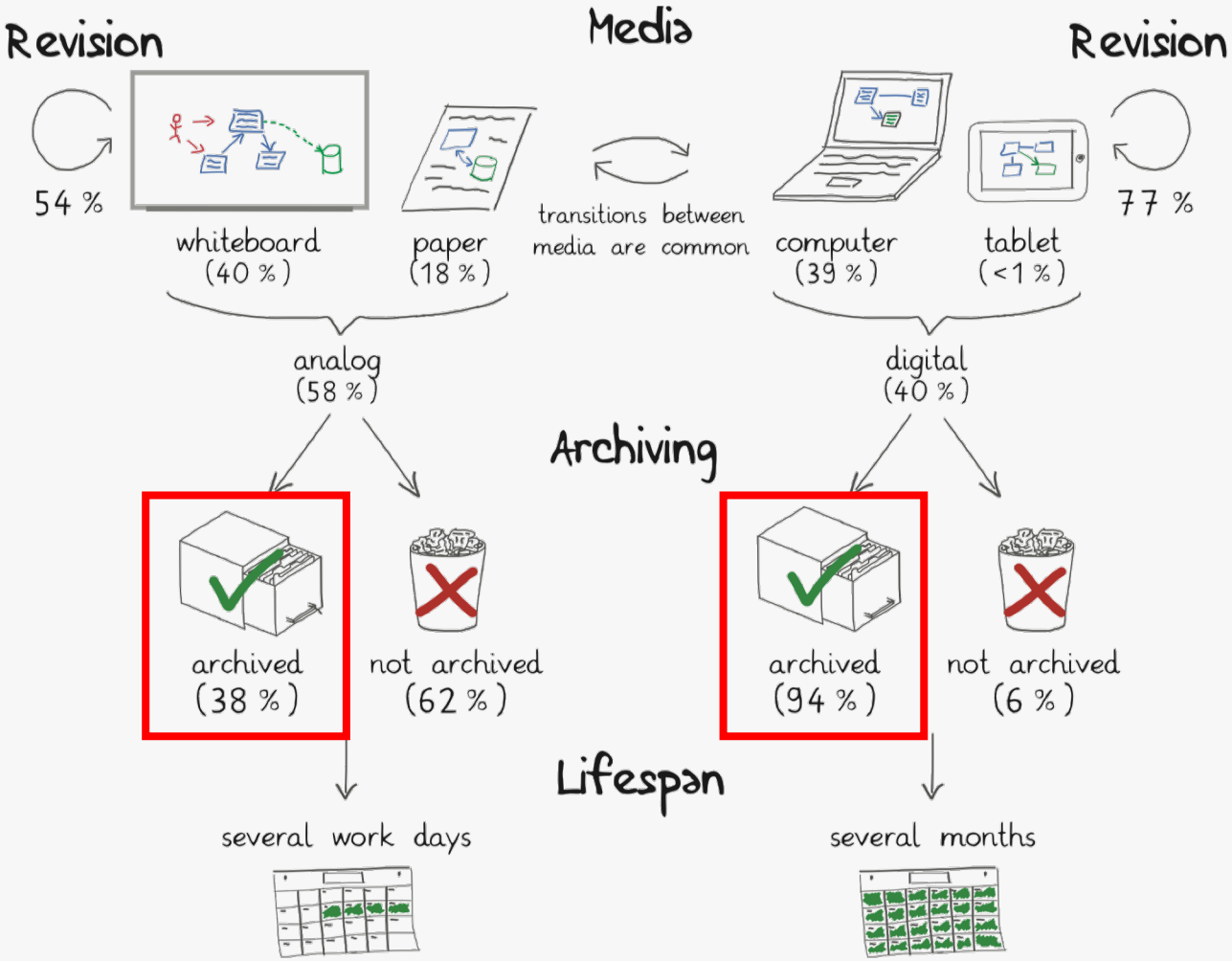
- Formative Evaluierungen



Skizzen & Diagramme

Sketches and Diagrams in Practice





Purpose

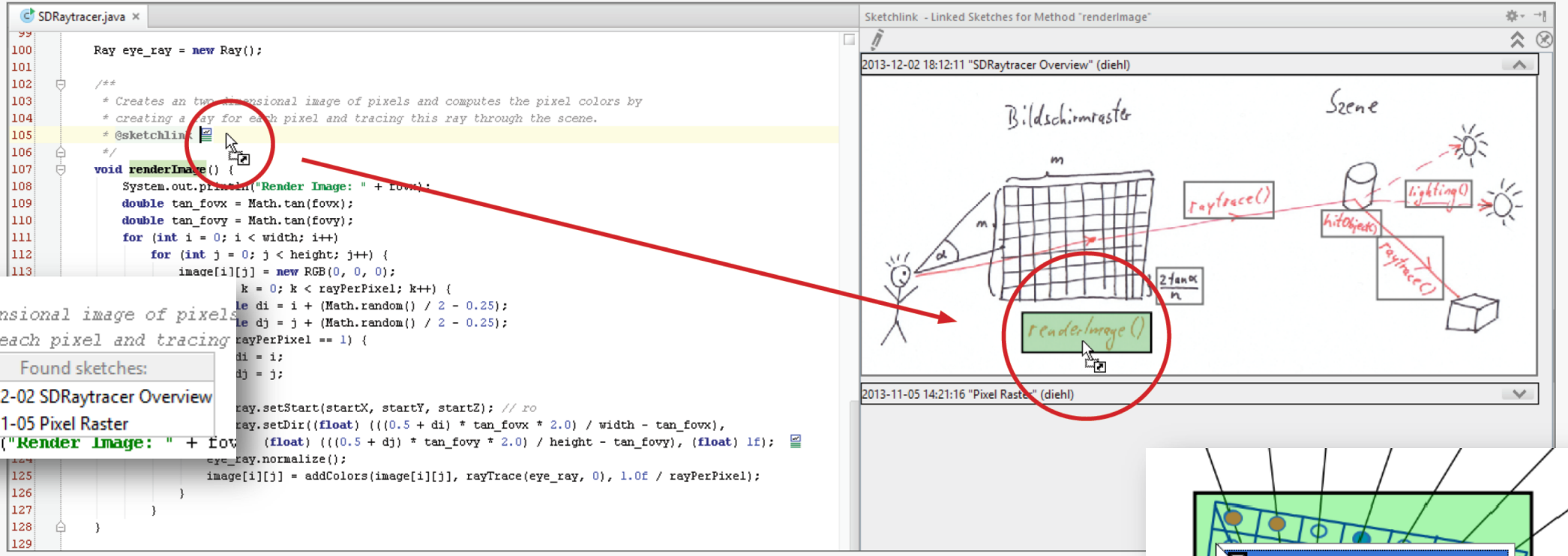
Designing (75 %) 1
Explaining (60 %) 2
Understanding (56 %) 3
Analyzing Requirements (45 %)

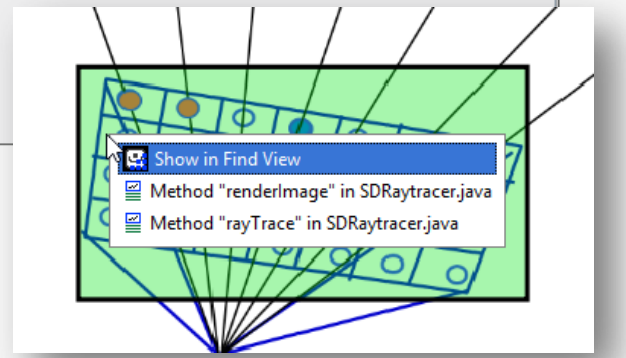
Relation to Source Code

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

Skizzen & Diagramme

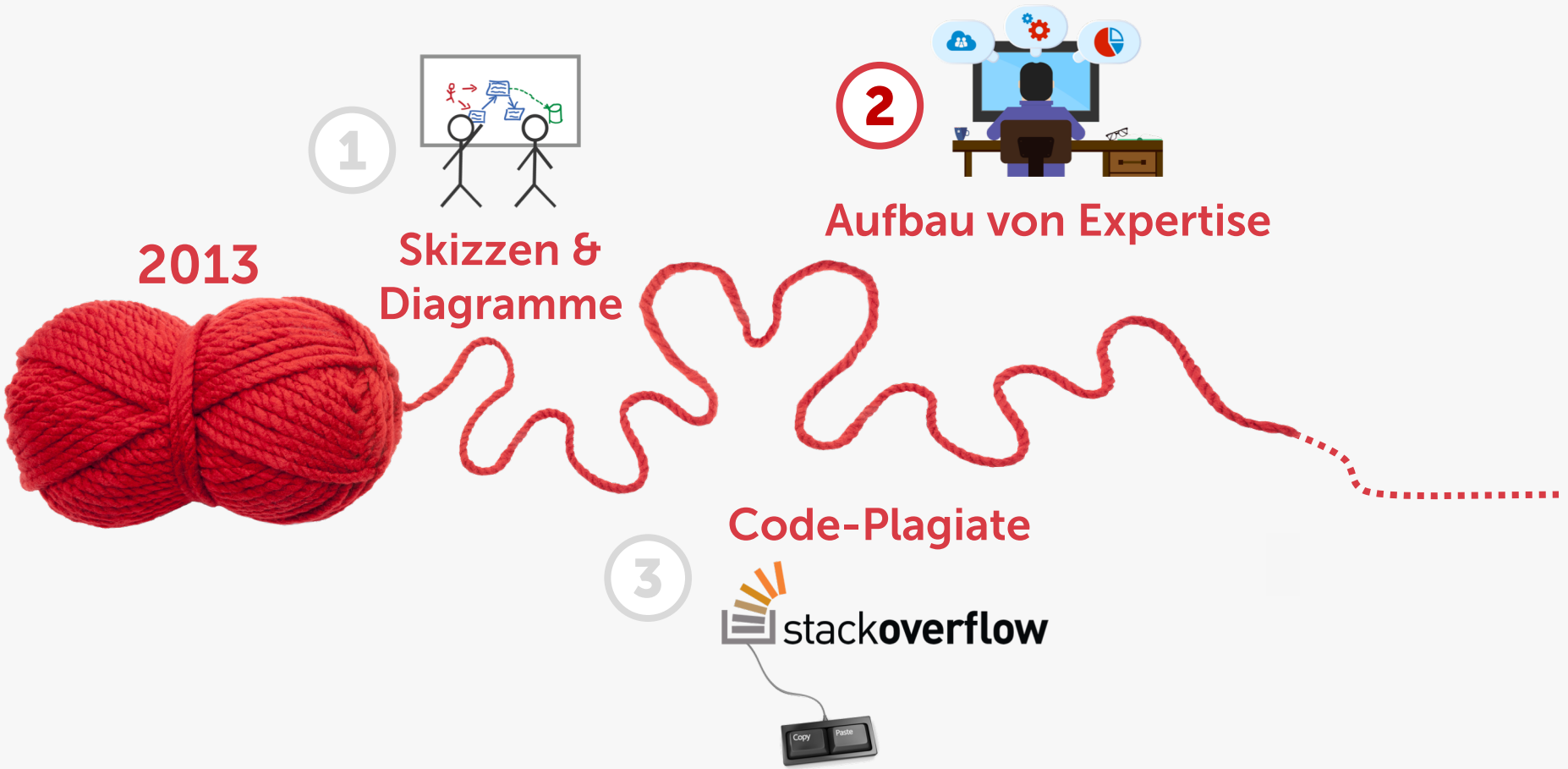
SketchLink





http://sketchlink.sbaltes.com

Struktur dieses Vortrags



Expertise in der Softwareentwicklung

Expertise

Software Development Expertise?

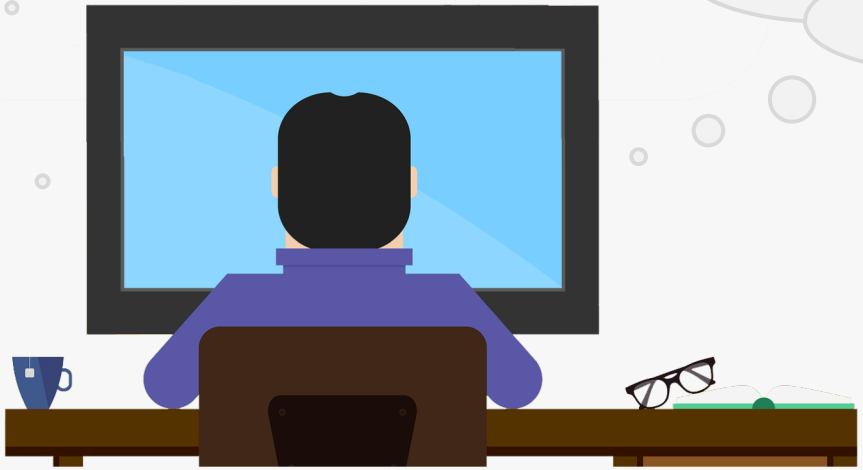
Neue Features implementieren

Algorithmen & Datenstrukturen

Testen

Kommunikation

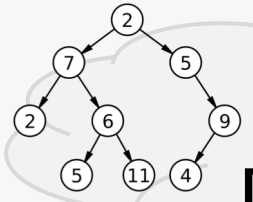
Debuggen



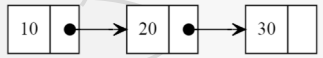
Expertise

Software Development Expertise?

Neue Features implementieren



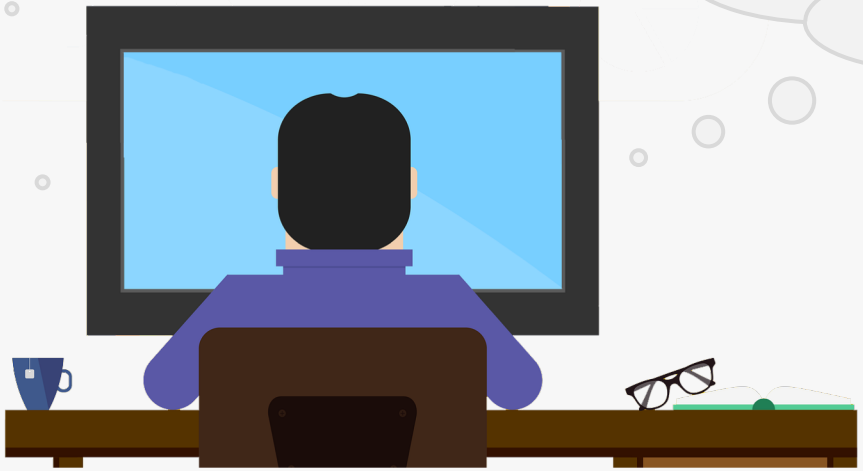
Algorithmen & Datenstrukturen



JUnit 5 Testen jbehave

Kommunikation

Debuggen



Expertise

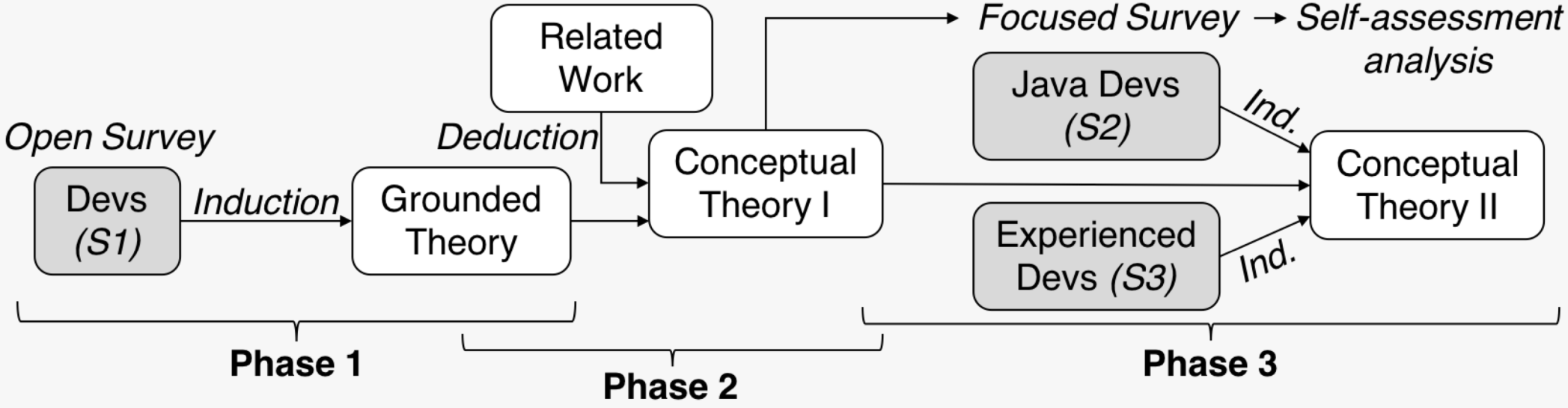
Forschungsfragen

Wie könnte ein **Modell** aussehen, dass wichtige Aspekte im Zusammenhang mit **Expertise in der Softwareentwicklung** strukturiert?

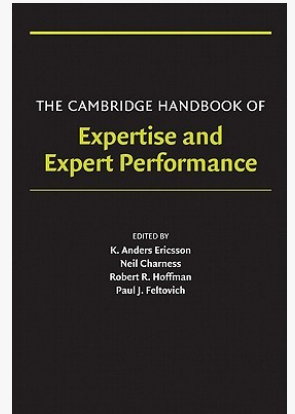
Welche **Faktoren** beeinflussen den Aufbau von Expertise **über die Zeit**?

Expertise

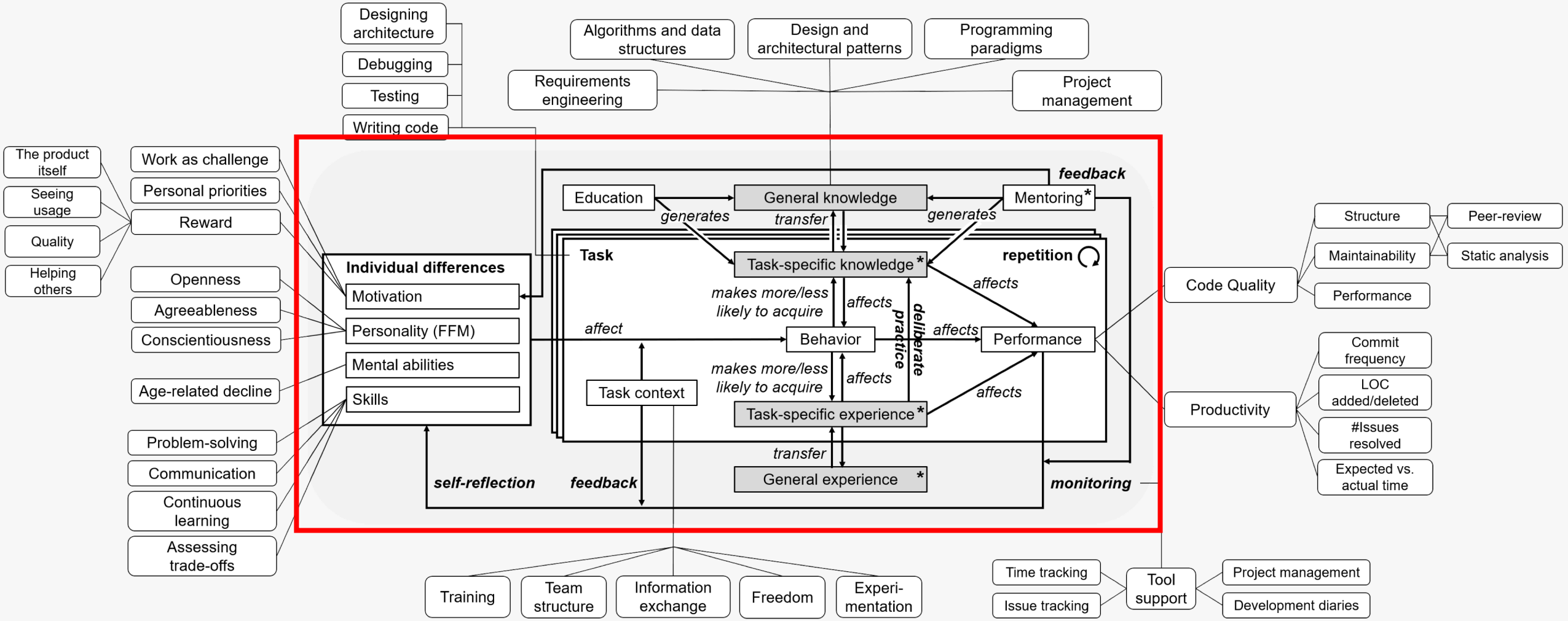
Forschungsmethode: Iterative Theoriebildung



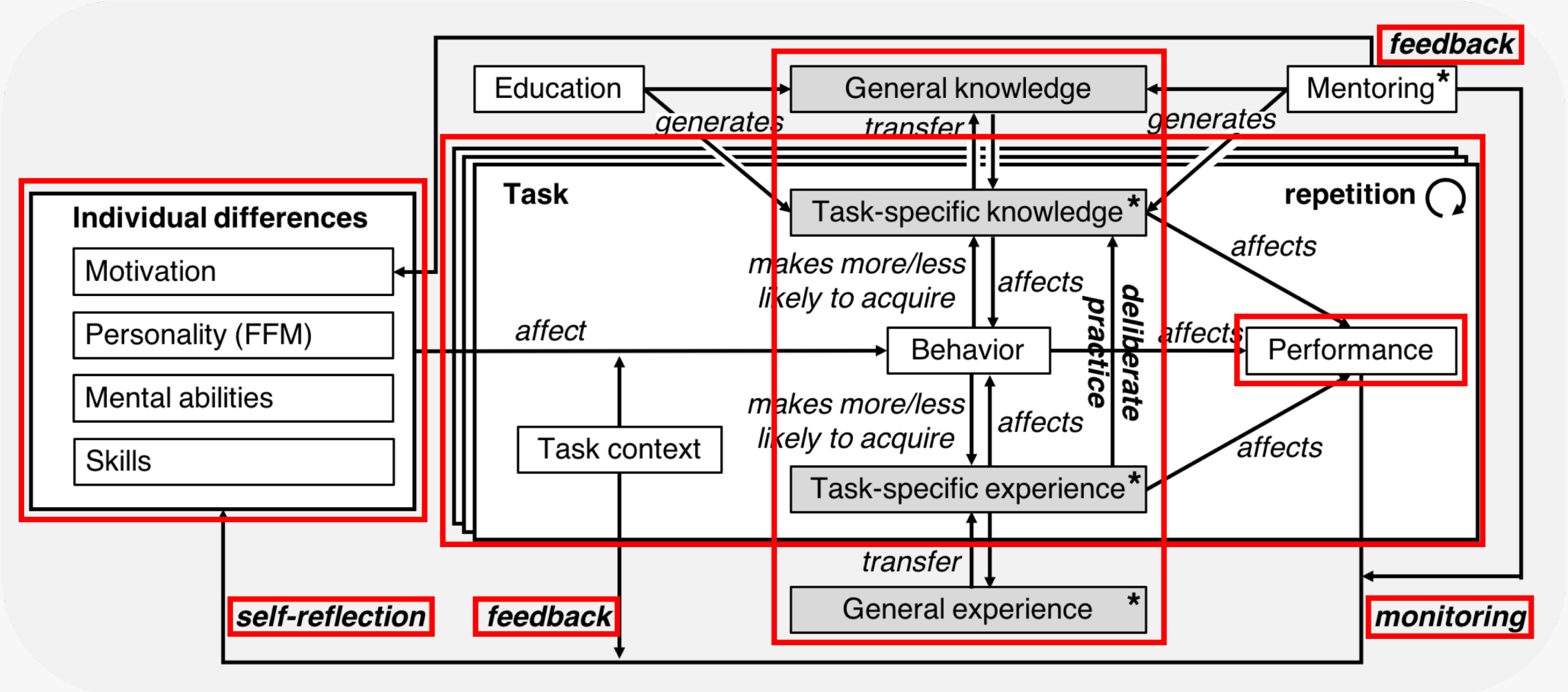
- **Induktiver Teil:** 335 Softwareentwickler
- **Deduktiver Teil:** *"Cambridge Handbook of Expertise and Expert Performance"* und andere verwandte Arbeiten



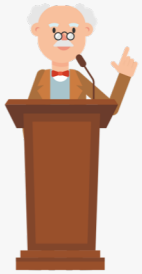
Konzeptionelle Theorie



Konzeptionelle Theorie



Kurzzusammenfassung



Forscher können...

- Mit Hilfe unseres Modells **neue Studien entwerfen**, um bestehende Wissenslücken zu schließen
- Unseren iterativen Ansatz zur **Theoriebildung** adaptieren

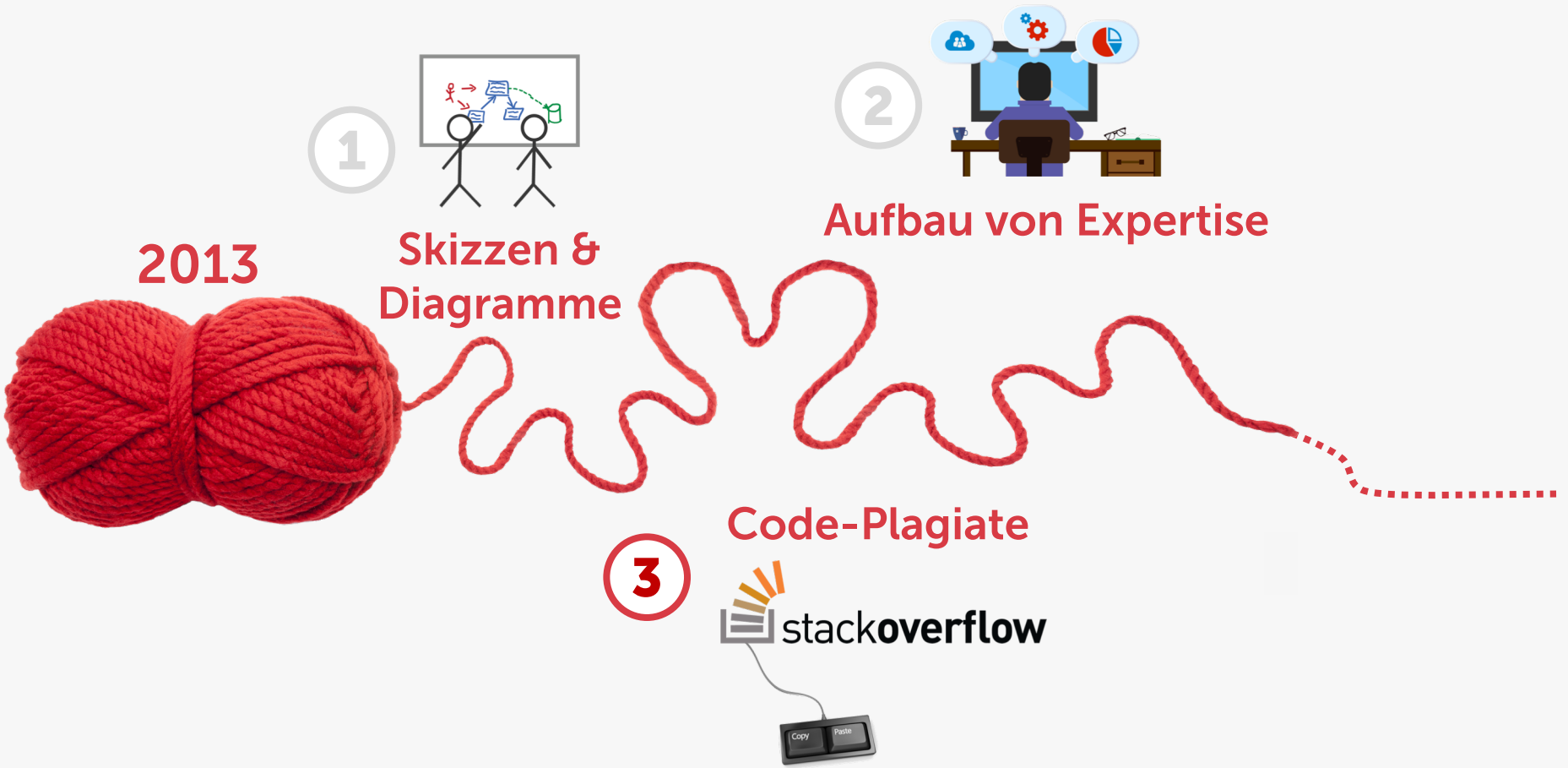
Entwickler können...

- Lernen, was andere Entwickler von **Experten/Mentoren** erwarten
- Erfahren, welche **Lernmodelle** beim Aufbau von Expertise helfen können

Arbeitgeber können...

- Erfahren, was Entwickler **demotiviert**
- Faktoren identifizieren, die eine **Arbeitsumgebung** ausmachen, die die Weiterbildung der Entwickler fördert

Struktur dieses Vortrags



Code-Plagiate

Code-Plagiate

GitHub

stackoverflow

Code-Plagiate

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

Frage

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

Antwort

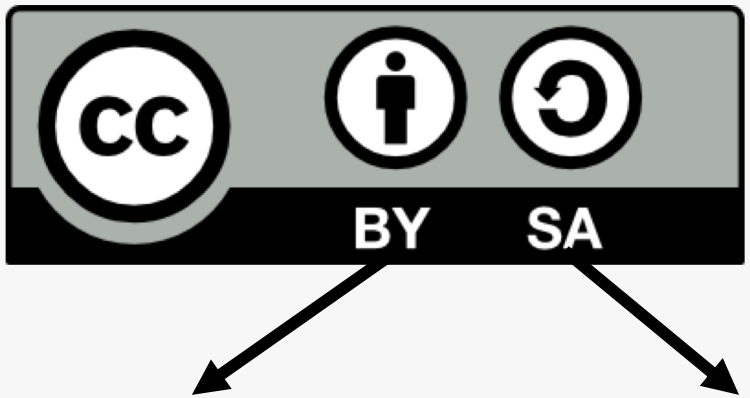
https://stackoverflow.com/a/5445161


stackoverflow
Copy
Paste

Lizensierung von Inhalten

Alle Inhalte auf Stack Overflow sind CC-BY-SA-lizensiert

*"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."*

Forschungsfragen

Wie häufig wird Code vom Stack-Overflow-Beiträgen in öffentliche Softwareprojekte auf GitHub **ohne** die erforderliche **Zuschreibung** kopiert?

Sind sich **Softwareentwickler** der Lizenzsituation und deren Implikationen **bewusst**?

Code-Plagiate

Forschungsmethoden

- Recherche der rechtlichen Situation §
- Data Mining, Code-Clone-Detektoren Google Cloud PMD DON'T SHOOT THE MESSENGER
- Kontaktierung von Entwicklern mit potentiell problematischem Code in deren Projekten

Code-Plagiate

Urheberrecht

"Diese Code-Fragemente sind trivial und nicht vom Urheberrecht geschützt."

- Nicht alle Code-Fragmente auf Stack Overflow überschreiten die Schöpfungshöhe
- "A snippet that is more than one or two lines of standard function calls would typically be creative enough for copyright" [Engelfriet 2016]
- Kein "international standard for originality" [Creative Commons 2017b]

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

889

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

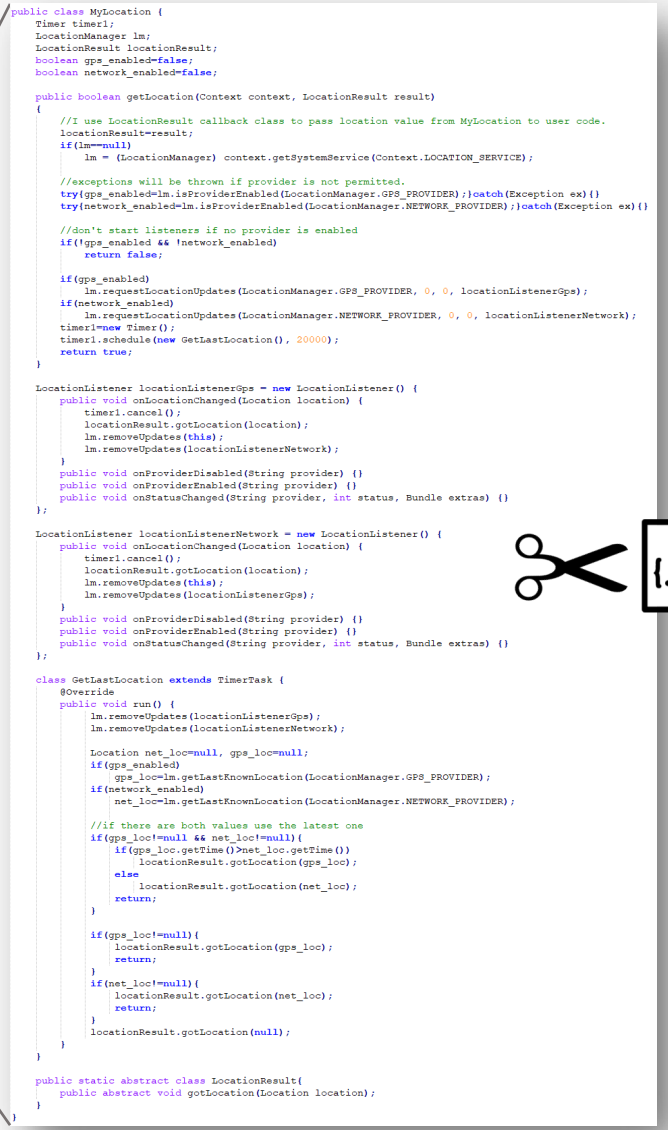
answered Jun 30 '10 at 0:07

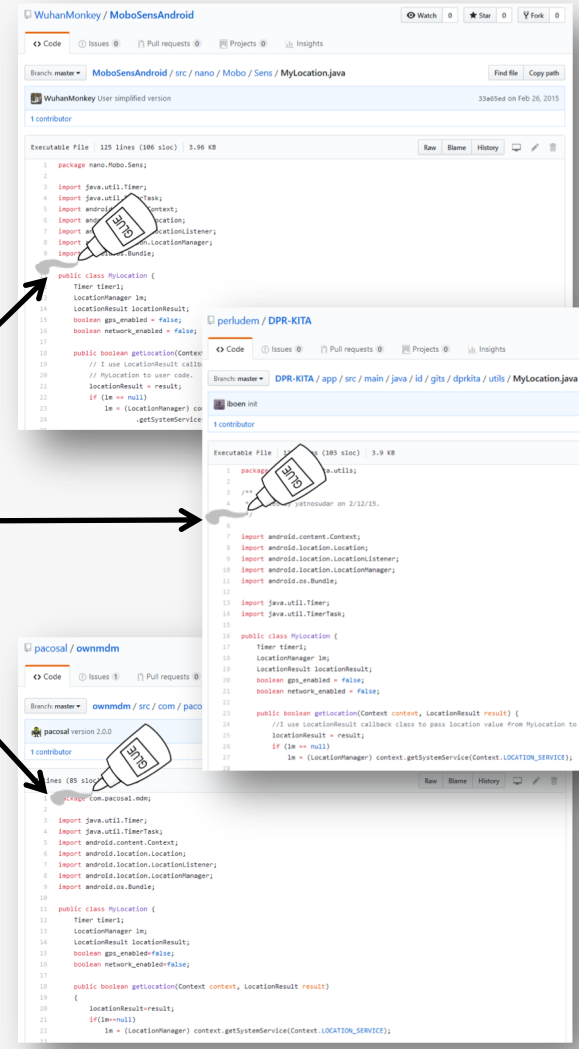
Fedor
40k ● 9 ● 71 ● 86

https://stackoverflow.com/a/3145655

stackoverflow

GitHub





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

Code-Plagiate

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.

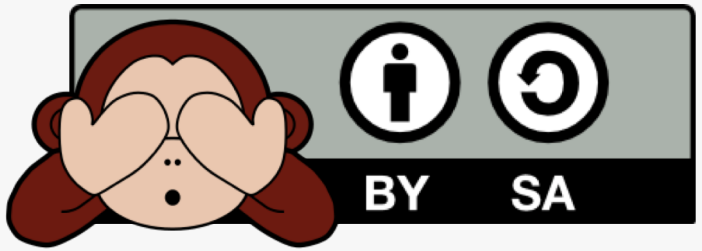
Besides the potential legal issues, duplicating a code is not a good practice.

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

Ergebnisse unsere Onlineumfrage

- **46%** der befragten Entwickler gaben zu, Code **ohne Zuschreibung** zu kopieren
- **75%** wussten nicht, dass der Code **CC BY-SA-lizensiert** ist.
- **67%** wussten nicht, dass eine **Zuschreibung erforderlich** ist

→ **Mangelndes Bewusstsein**



Code-Plagiate

Attribution

Anteil Kopien mit Zuschreibung:

- Methode 1 (Reguläre Ausdrücke): 23 %
- Methode 2 (Code-Clone-Detektor): 24 %
- Methode 3 (Exakte Kopien): 8 %

Konservative Abschätzung:

- **Anteil ≤ 25%**

Code-Plagiate

Share-alike

Nur **2%** aller analysierten Projekte mit Code-Fragmenten von Stack Overflow **wiesen auf die Quelle hin** und verwendeten eine **kompatible Lizenz** (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

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
software artifacts, questions and answe
for example when bugs in code snippet
to work with a more recent library ver
code snippet is edited for clarity. To be
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 collecti

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

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

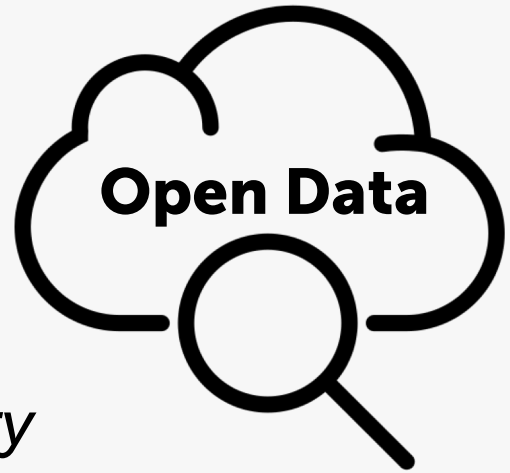
Christoph Treude
University of Adelaide, Australia
christoph.treude@adelaide.edu.au

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

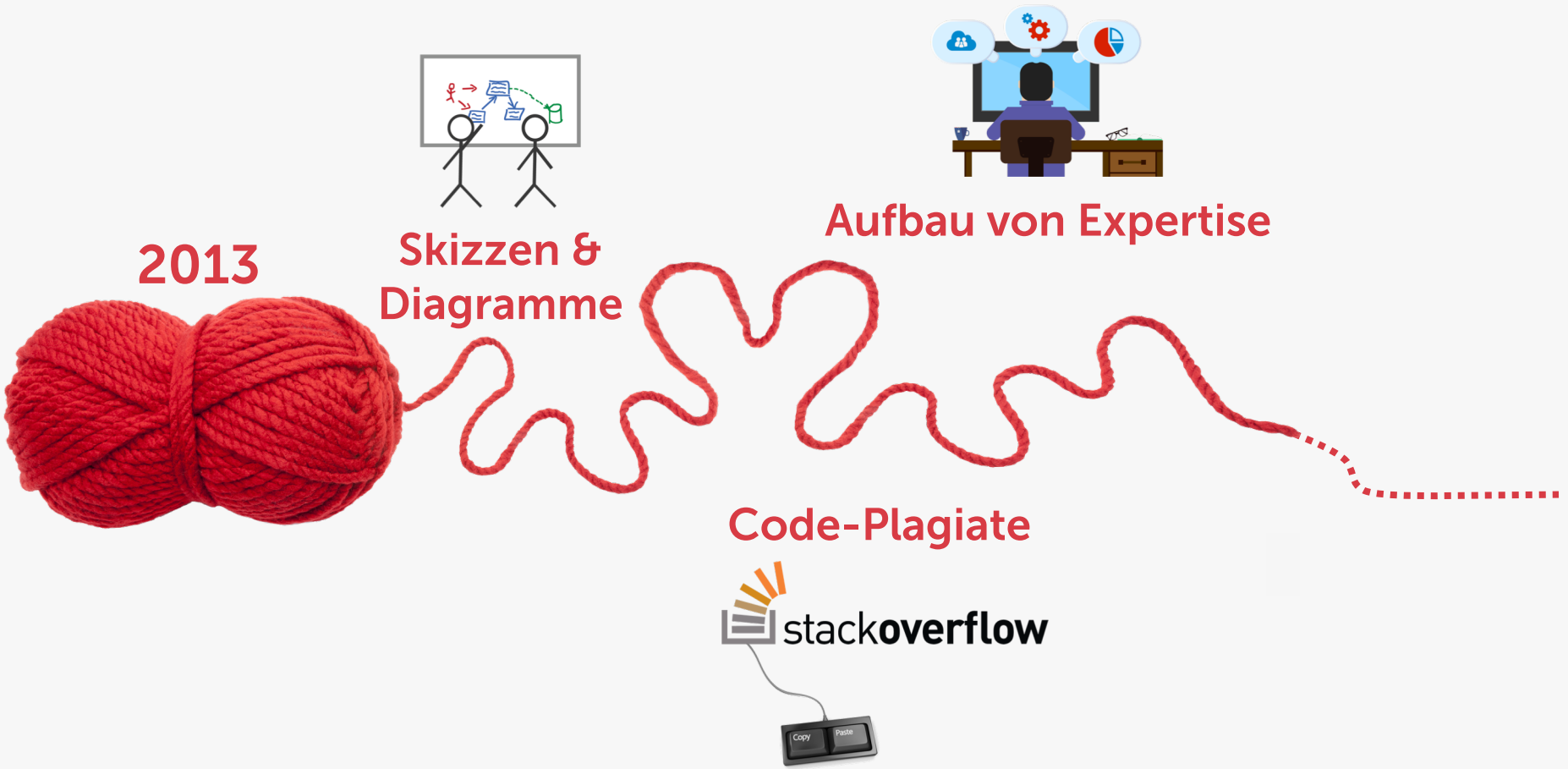
MSR 2018/19

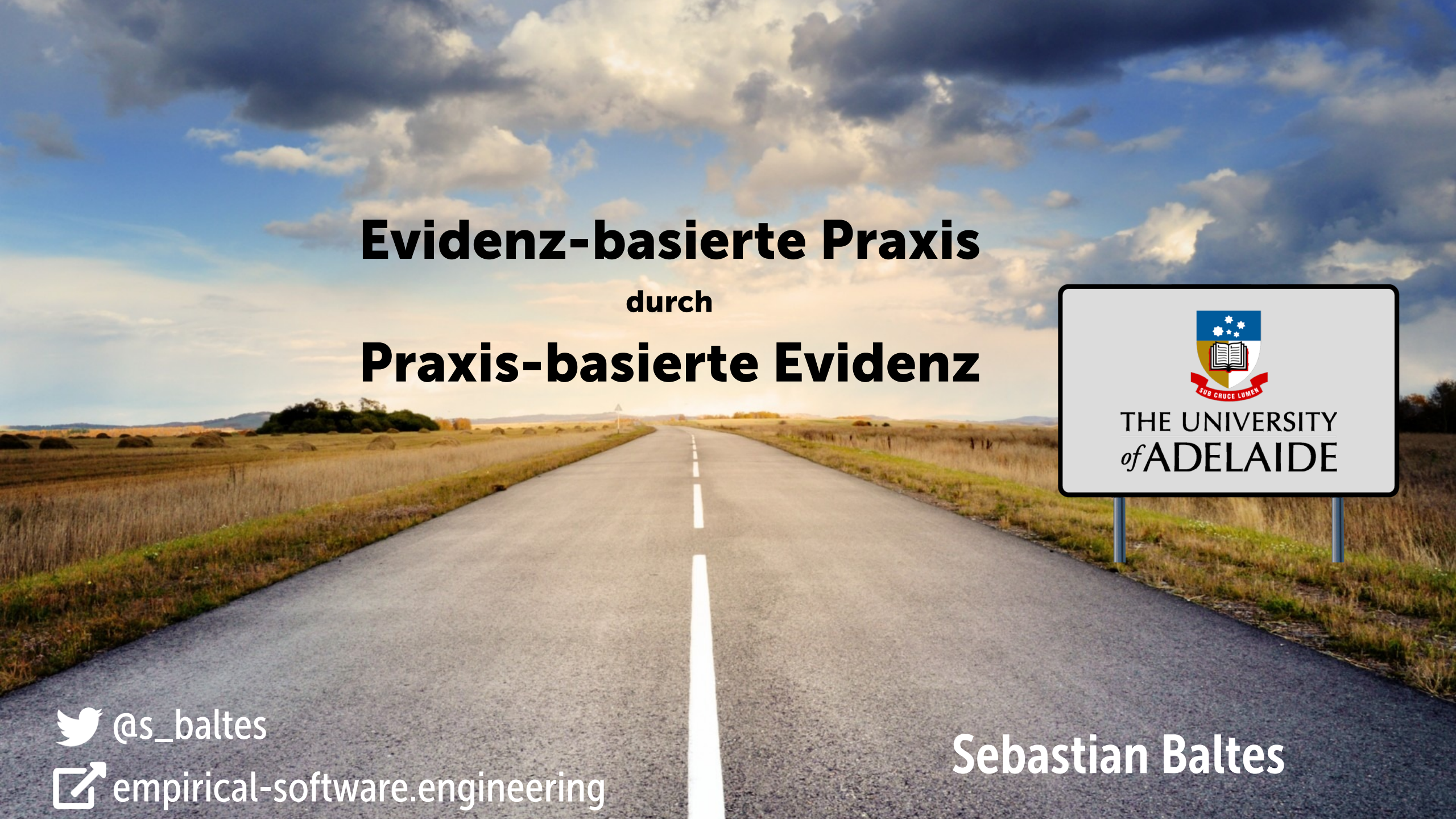


sotorrent.org

Dataset available on Zenodo and BigQuery

Behandelte Themenbereiche




Evidenz-basierte Praxis
durch
Praxis-basierte Evidenz
SUB CRUCE LUMEN
THE UNIVERSITY
of ADELAIDE
@s_baltes
empirical-software.engineering
Sebastian Baltes

Pandemic Programming

How COVID-19 affects software developers and how their organizations can help

Paul Ralph · Sebastian Baltes · Gianisa Adisaputri · Richard Torkar · Vladimir Kovalenko · Marcos Kalinowski · Nicole Novielli · Shin Yoo · Xavier Devroey · Xin Tan · Minghui Zhou · Burak Turhan · Rashina Hoda · Hideaki Hata · Gregorio Robles · Amin Milani Fard · Rana Alkadhi

http://covid19.sbaltes.com