



Software Developers' Work Habits and Expertise

Empirical Studies on Sketching, Code Plagiarism, and Expertise Development

Sebastian Baltes

 @s_baltes



THE UNIVERSITY
of ADELAIDE

 [empirical-software.engineering](https://github.com/empirical-software-engineering)

 **Universität** Trier

Ein bisschen Geographie...





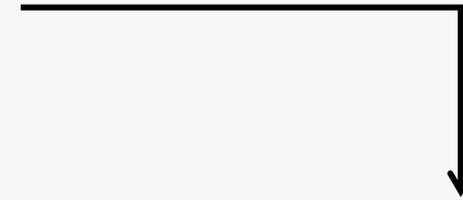
Evidenz-basierte Praxis durch **Praxis-basierte Evidenz**

Meine Dissertation

Beobachten
Beschreiben
Erklären



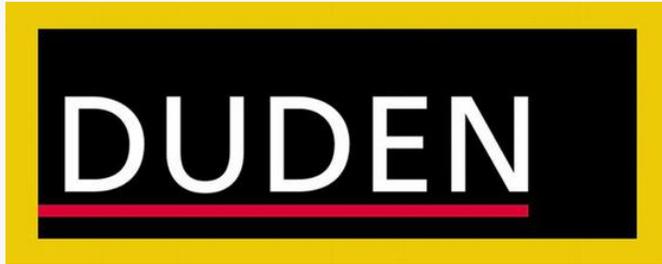
Arbeitsgewohnheiten
von
Softwareentwicklern



- Anforderungen an Entwicklungswerkzeuge
- Verbesserungen am Entwicklungsprozess
- Wissenschaftskommunikation!



Arbeitsgewohnheiten?



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

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

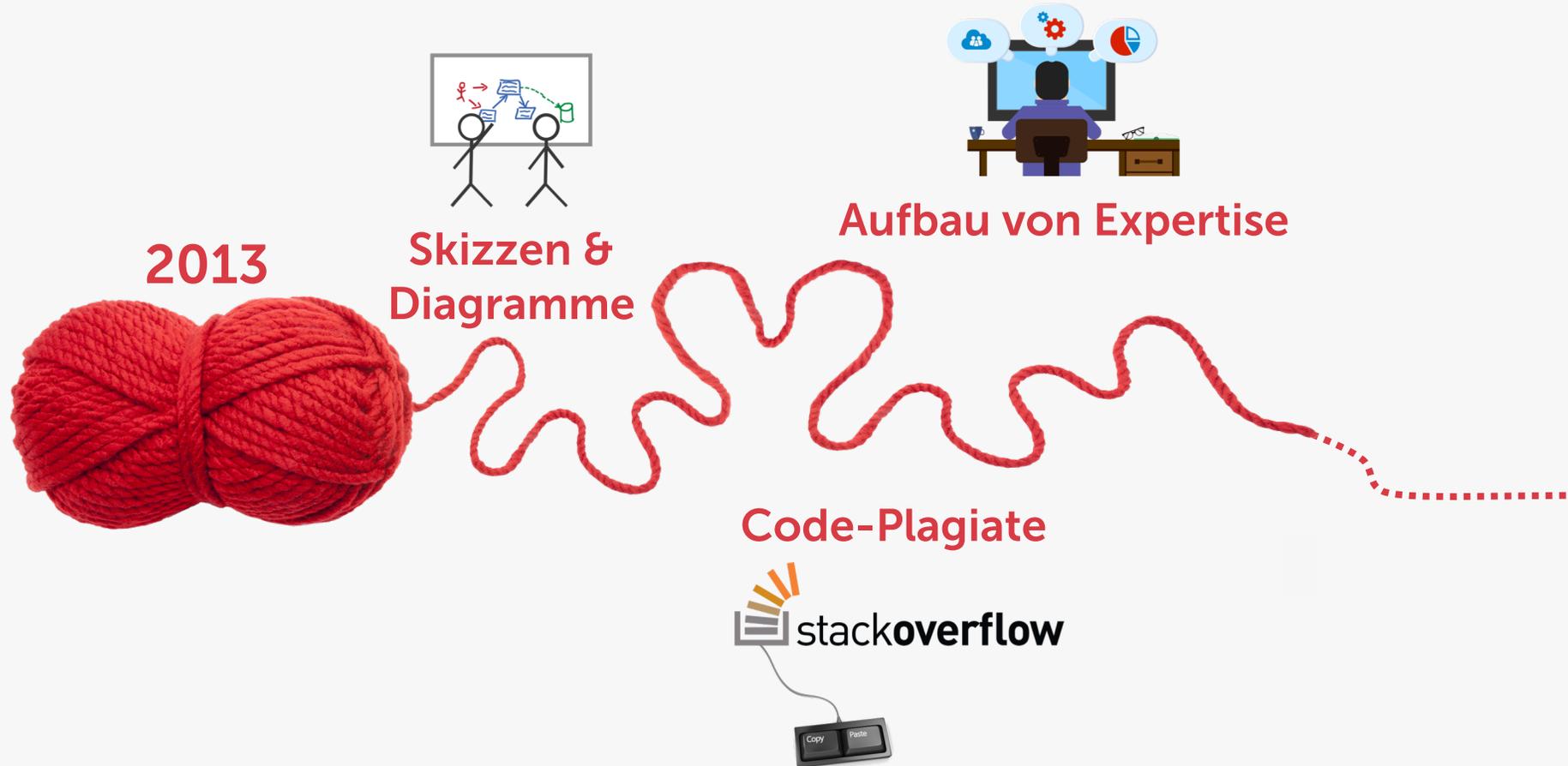
Private Gewohnheiten



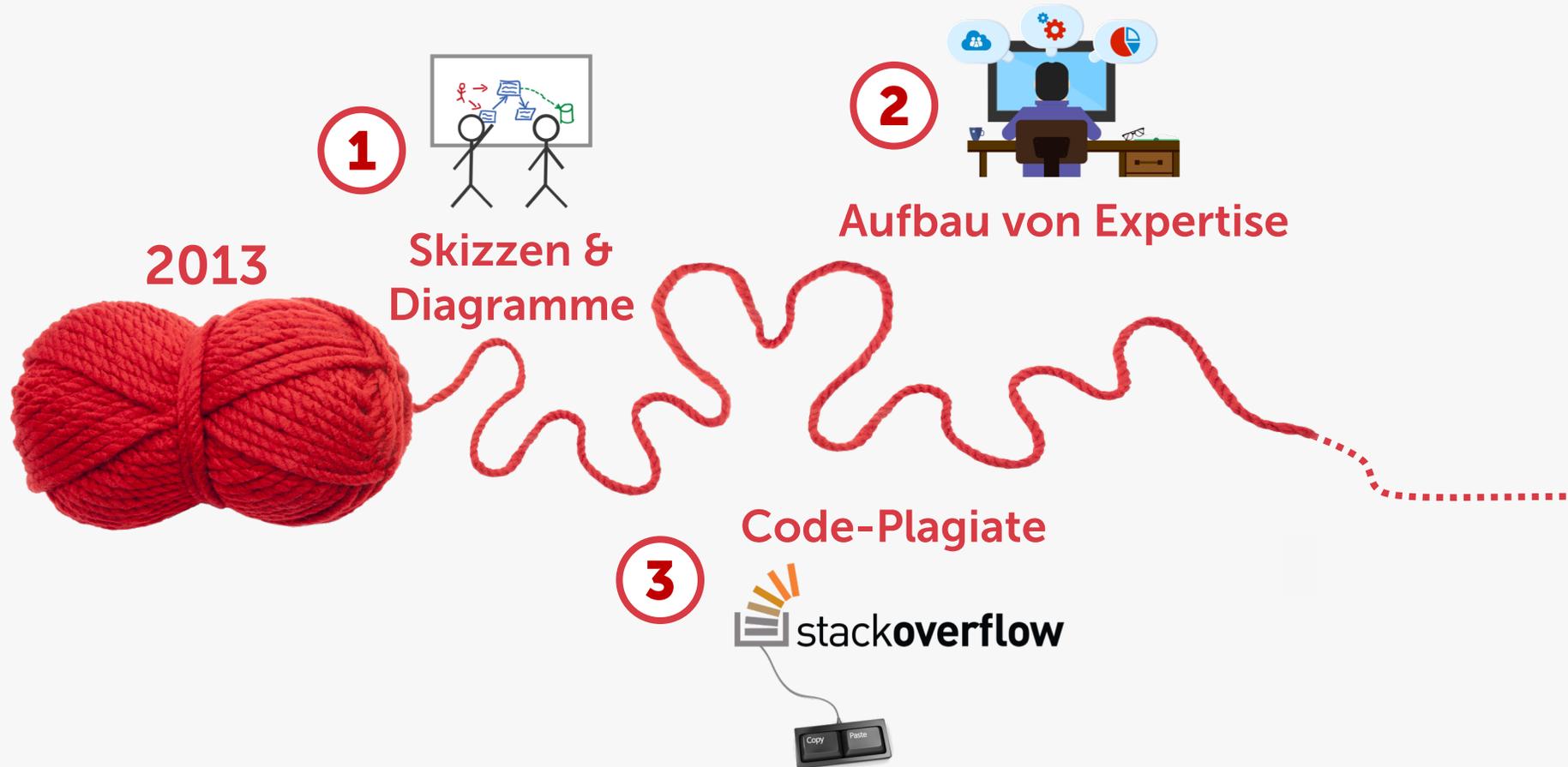
Arbeitsgewohnheiten



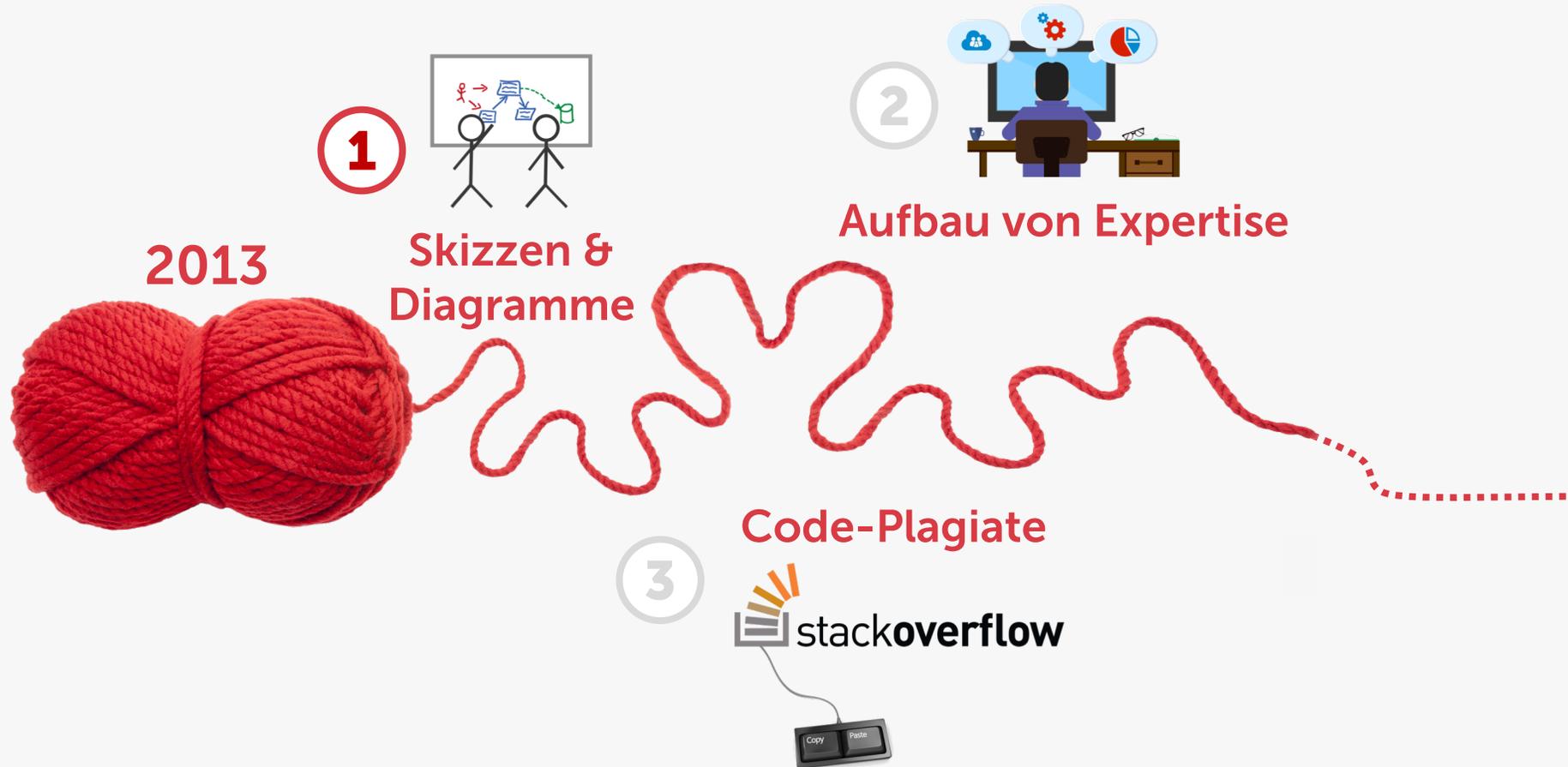
Behandelte Themenbereiche

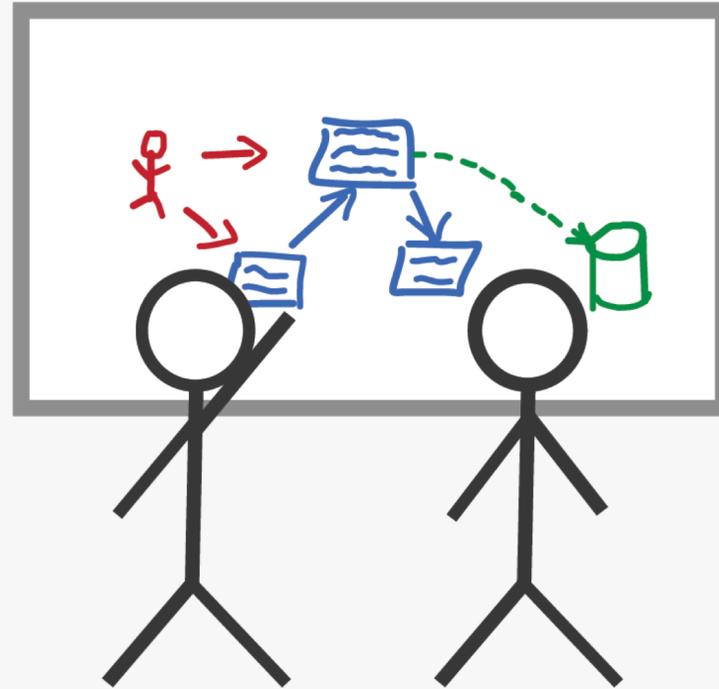


Struktur dieses Vortrags

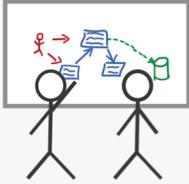


Struktur dieses Vortrags





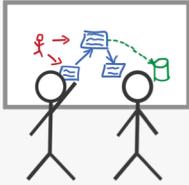
Skizzen & Diagramme



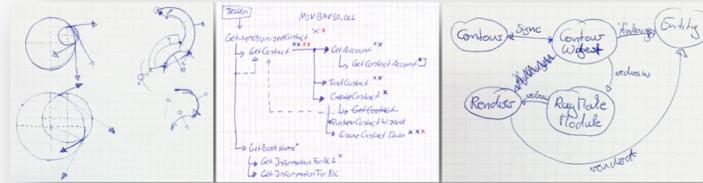
Wie und **warum** verwenden Softwareentwickler/-innen **Skizzen und Diagramme**?

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

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



• Feldstudie



• Onlineumfrage

Questionnaire on the Use of Sketches and Diagrams in Software Development

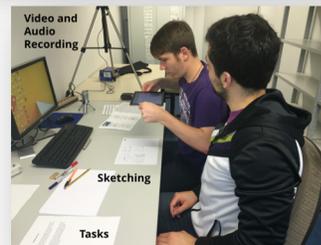
The goal of this short survey (5-10 minutes) is to evaluate how software developers and other people involved in the development of software use sketches and diagrams in their daily work. By the terms sketches and diagrams, we mean visual artifacts created by hand (e.g. hand-drawn sketches on scrap paper or whiteboards) as well as the ones created using computer programs (e.g. automatically generated UML diagrams).

The information collected in this survey is handled confidentially; only anonymized results will be published. If you like to, you can send us some of your sketches and diagrams. For further information, please read the instructions shown after submitting the questionnaire.

To answer the following questions, please think of the last sketch or diagram that you created for your professional work and that is related to a software project.

| | |
|---|-------------------------------------|
| 1. When did you create your last sketch/diagram? | Please choose: <input type="text"/> |
| 2. Has the sketch/diagram been revised after its initial creation? | Please choose: <input type="text"/> |
| 3. How much effective work time went into the creation and revision of the sketch/diagram up to now (if several persons were involved, add up the individual work times of all contributors)? | Please choose: <input type="text"/> |

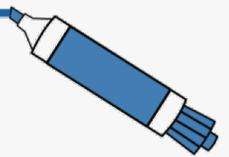
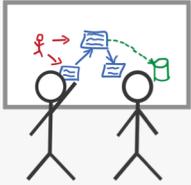
• Laborexperiment



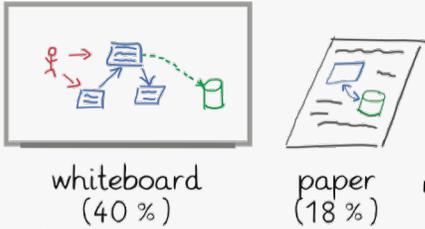
• Formative Evaluierungen

```
/**  
 * Creates an two-dimensional image of pixels  
 * creating a ray for each pixel and tracing  
 * @sketchlink Found sketches:  
 */  
void renderImage (2013-12-02 SDRaytracer Overview  
2013-11-05 Pixel Raster  
System.out.println("Render Image: " + fov
```

Sketches and Diagrams in Practice



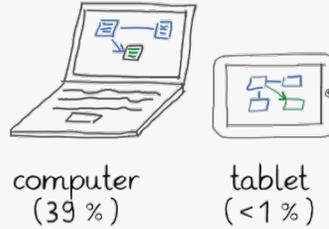
Revision



Media

transitions between media are common

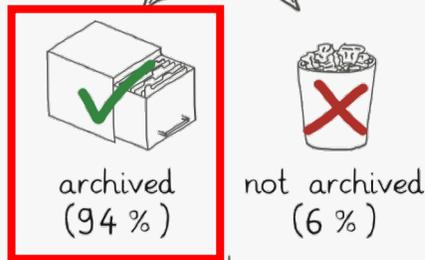
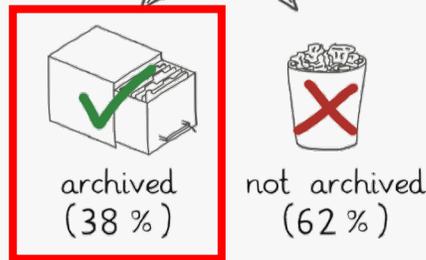
Revision



analog (58%)

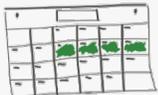
digital (40%)

Archiving

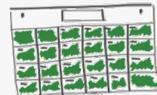


Lifespan

several work days



several months



Purpose

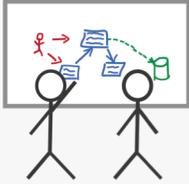
- Designing (75%)
- Explaining (60%)
- Understanding (56%)
- Analyzing Requirements (45%)

Relation to Source Code



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

Skizzen & Diagramme



SketchLink

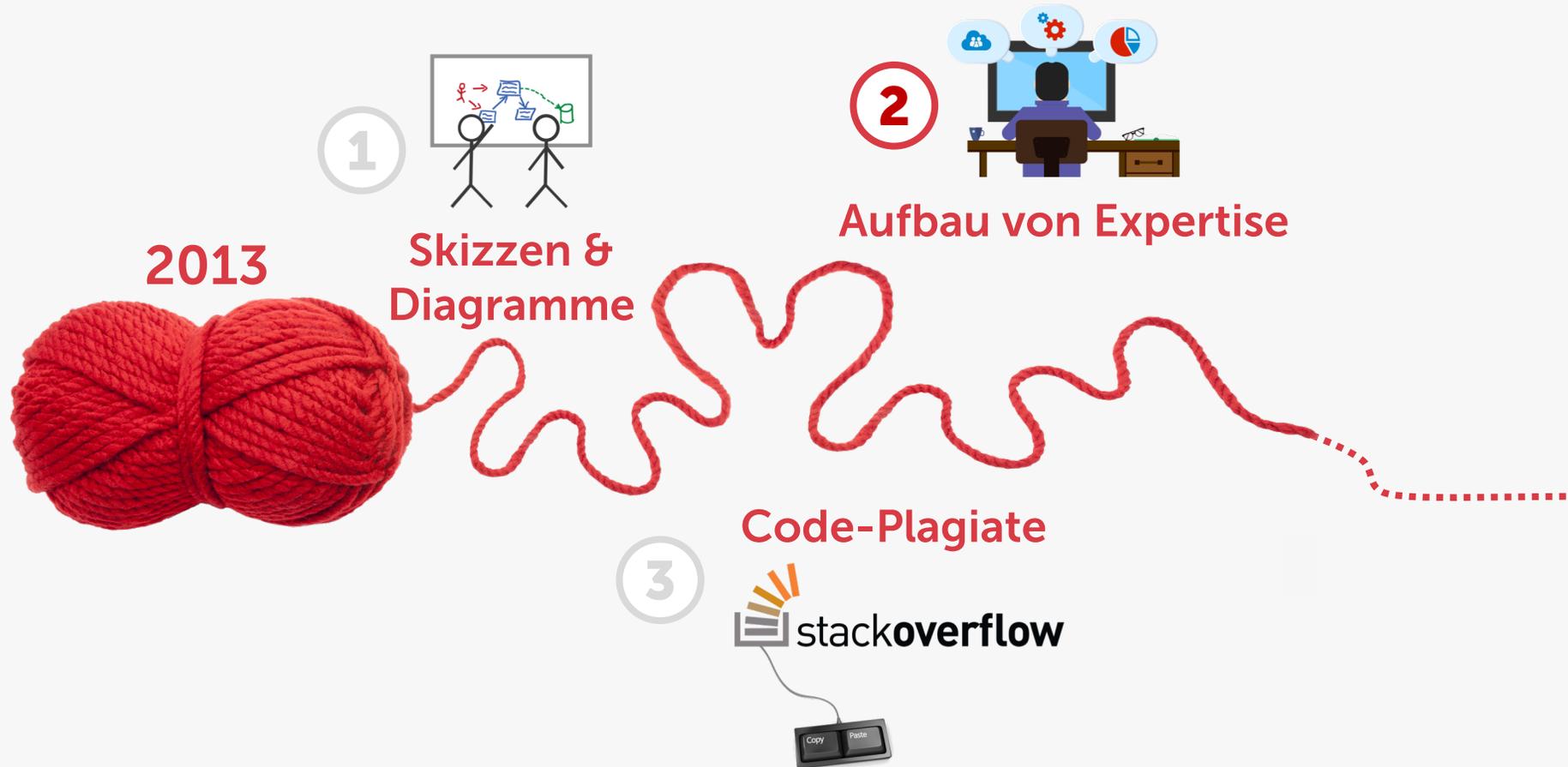
```
100 Ray eye_ray = new Ray();
101
102 /**
103  * Creates a two-dimensional image of pixels and computes the pixel colors by
104  * creating a ray for each pixel and tracing this ray through the scene.
105  * @sketchlink
106  */
107 void renderImage() {
108     System.out.println("Render Image: " + fov);
109     double tan_fovx = Math.tan(fov);
110     double tan_fovy = Math.tan(fov);
111     for (int i = 0; i < width; i++)
112         for (int j = 0; j < height; j++) {
113             image[i][j] = new RGB(0, 0, 0);
114             for (int k = 0; k < rayPerPixel; k++) {
115                 double di = i + (Math.random() / 2 - 0.25);
116                 double dj = j + (Math.random() / 2 - 0.25);
117                 Ray ray = new Ray(0, 0, 0, 1, 0, 0, 1);
118                 ray.setStart(startX, startY, startZ); // ro
119                 ray.setDir((float) (((0.5 + di) * tan_fovx * 2.0) / width - tan_fovx),
120                         (float) (((0.5 + dj) * tan_fovy * 2.0) / height - tan_fovy), (float) 1f);
121                 eye_ray.normalize();
122                 image[i][j] = addColors(image[i][j], rayTrace(eye_ray, 0), 1.0f / rayPerPixel);
123             }
124         }
125     }
126 }
127 }
128 }
129 }
```

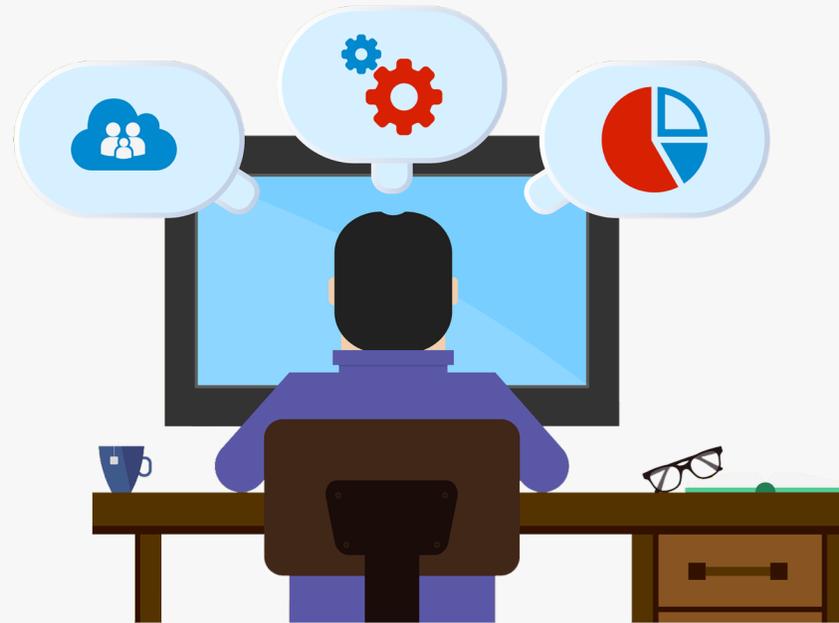
Found sketches:
2013-12-02 SDRaytracer Overview
2013-11-05 Pixel Raster

renderImage()

<http://sketchlink.sbaltes.com>

Struktur dieses Vortrags





Expertise in der Softwareentwicklung

Software Development Expertise?



Neue Features
implementieren

Algorithmen &
Datenstrukturen

Testen

Kommunikation

Debuggen



Software Development Expertise?



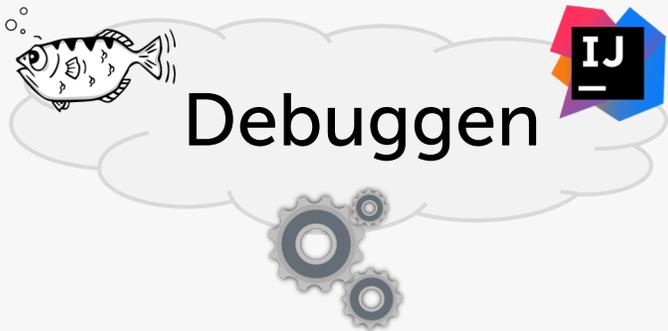
 Neue Features implementieren

Algorithmen & Datenstrukturen

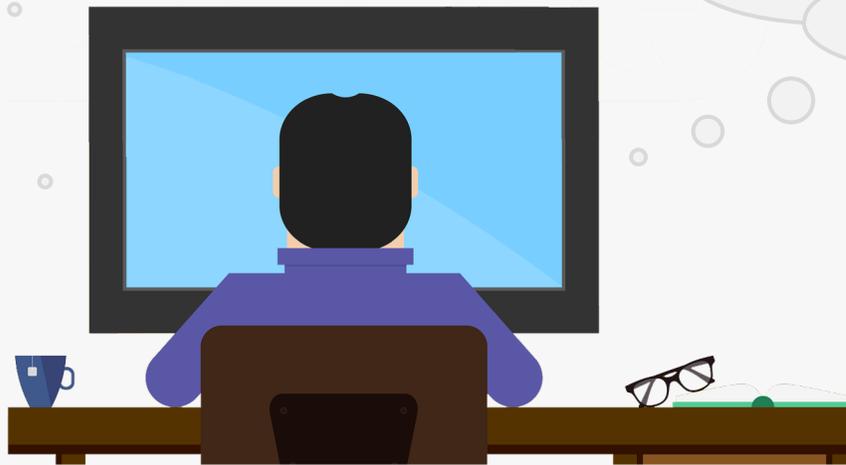


JUnit 5 Testen *jbehave*

Debuggen



Kommunikation





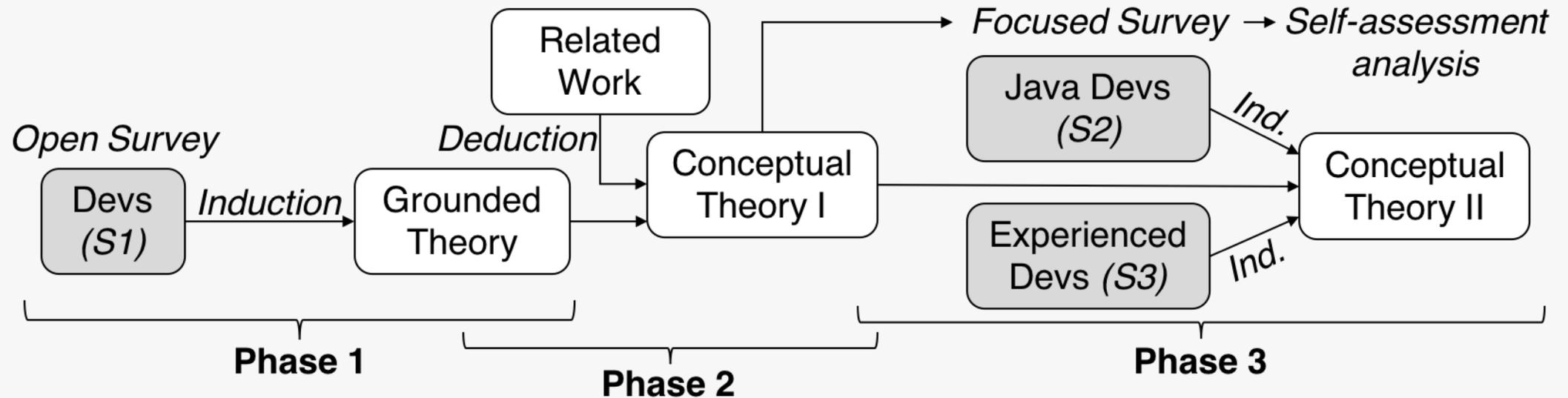
Forschungsfragen



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**?

Forschungsmethode: Iterative Theoriebildung

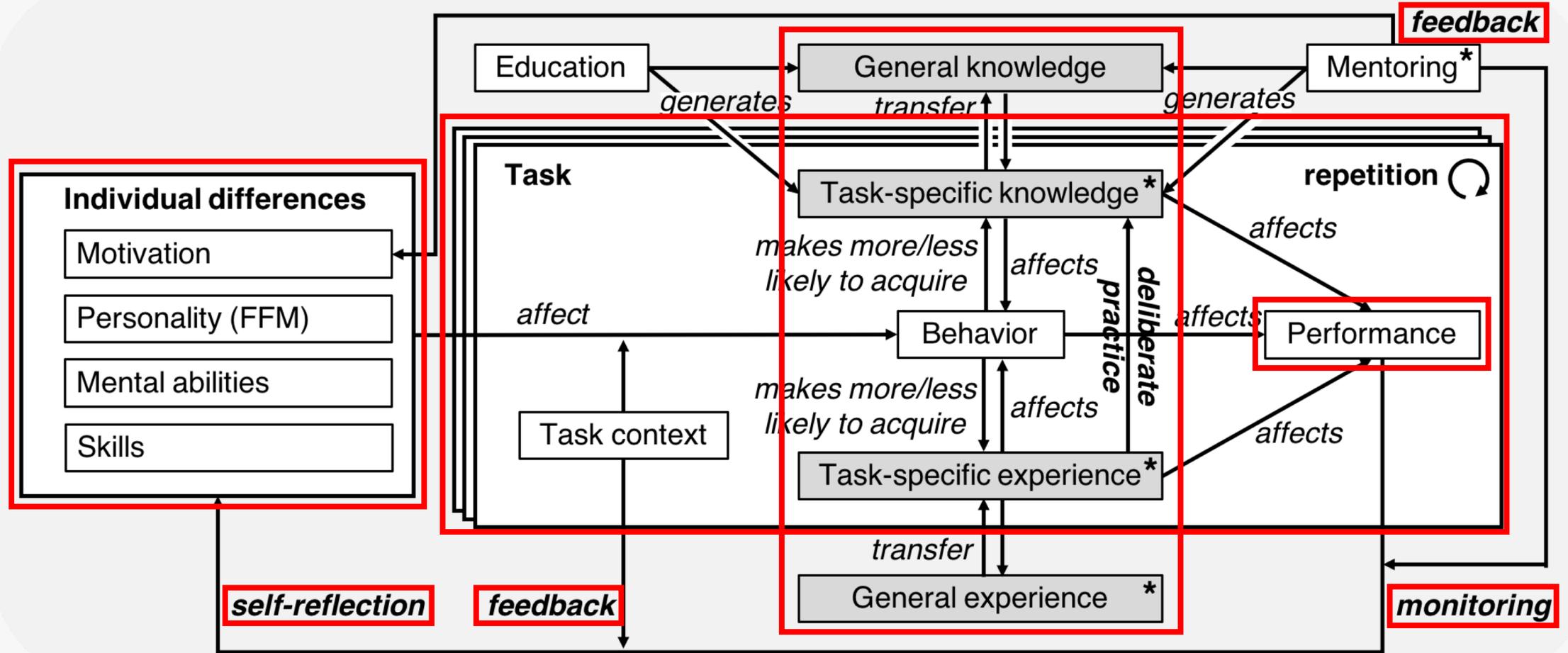


- **Induktiver Teil:** 335 Softwareentwickler
- **Deduktiver Teil:** "Cambridge Handbook of Expertise and Expert Performance" und andere verwandte Arbeiten

THE CAMBRIDGE HANDBOOK OF
Expertise and
Expert Performance

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

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...

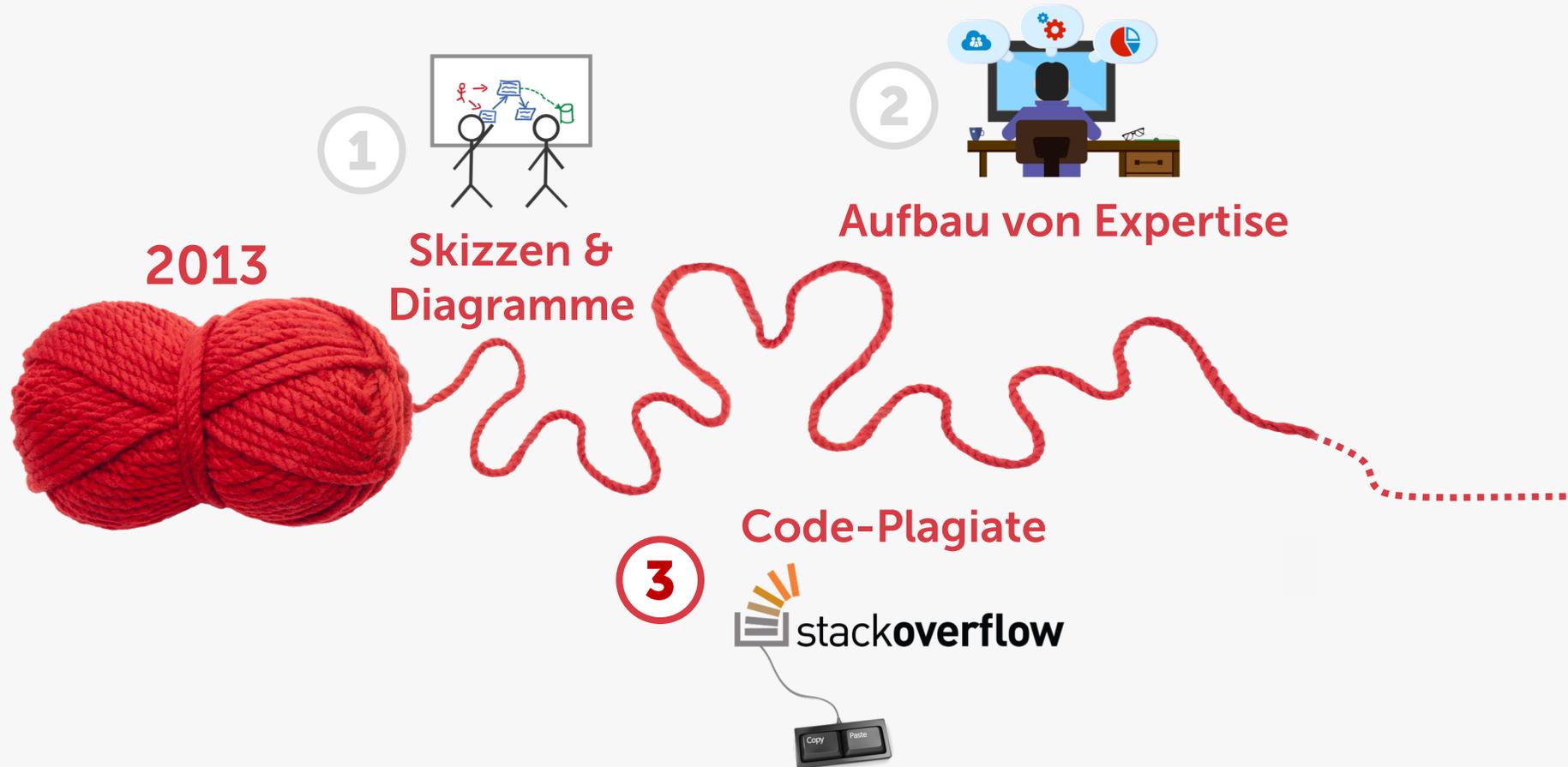
- Lernen, was andere Entwickler von **Experten/Mentoren** erwarten
- Erfahren, welche **Lernmodelle** beim Aufbau von Expertise helfen können



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





Read/convert an InputStream to a String

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

3101

▼ 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` ?

929

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

java string io stream inputstream

share improve this question

edited May 19 '17 at 8:58

Pehlaj
4.8k ● 6 ● 25 ● 43

asked Nov 21 '08 at 16:47

Johnny Maelstrom
15.9k ● 5 ● 18 ● 17

Frage

<https://stackoverflow.com/q/309424>

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

2034

```
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

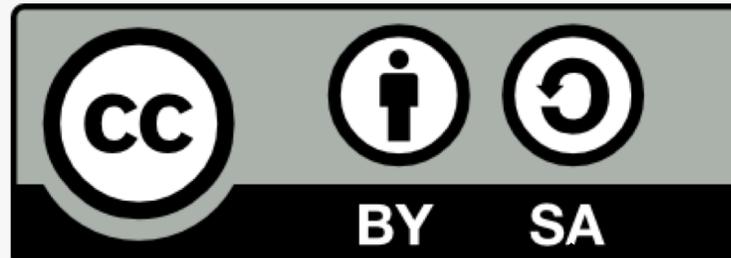
<https://stackoverflow.com/a/5445161>



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**?



Forschungsmethoden



- Recherche der rechtlichen Situation §

- Data Mining, Code-Clone-Detektoren



- Kontaktierung von Entwicklern mit potentiell problematischem Code in deren Projekten





Urheberrecht



“Diese Code-Fragmente 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]



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 getLocation(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



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

    public boolean getLocation(Context context, LocationResult result)
    {
        // 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.getLocation(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.getLocation(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) {}
    };

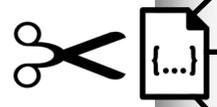
    class GetLastLocation extends TimerTask {
        @Override
        public void run() {
            lm.removeUpdates(locationListenerGps);
            lm.removeUpdates(locationListenerNetwork);

            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.getLocation(gps_loc);
                else
                    locationResult.getLocation(net_loc);
                return;
            }

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

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



The image shows three screenshots of GitHub repository pages. The top screenshot is for 'WuhanMonkey / MoboSensAndroid', the middle for 'perludem / DPR-KITA', and the bottom for 'pacosal / ownmdm'. Each screenshot shows the 'Code' tab with the 'MyLocation.java' file selected. A red circle highlights the class definition in each file, and a red arrow points from the central document icon to these highlights. The code in the screenshots matches the code shown in the Stackoverflow post.



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.**

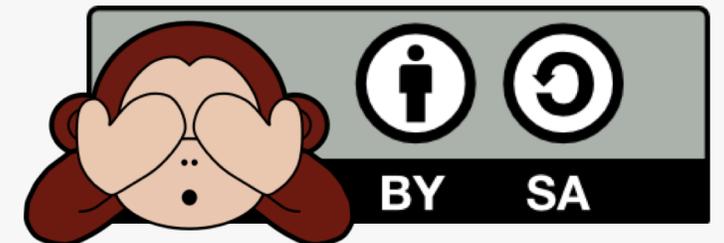
<https://bugs.openjdk.java.net/browse/JDK-8170860>



Ergebnisse unsere Onlineumfrage

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

→ **Mangelndes Bewusstsein**





Attribution



Anteil Kopien mit Zuschreibung:

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

Konservative Abschätzung:

- **Anteil \leq 25%**



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 (<i>n</i> = 2,962) | attributed (<i>n</i> = 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 (<i>n</i> = 144) | attributed (<i>n</i> = 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 (<i>n</i> = 1,169) | attributed (<i>n</i> = 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, Germany

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, providing snippets and free-form text on a wide range of software artifacts, questions and answers for example when bugs in code snippets to work with a more recent library version or a code snippet is edited for clarity. 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.

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

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

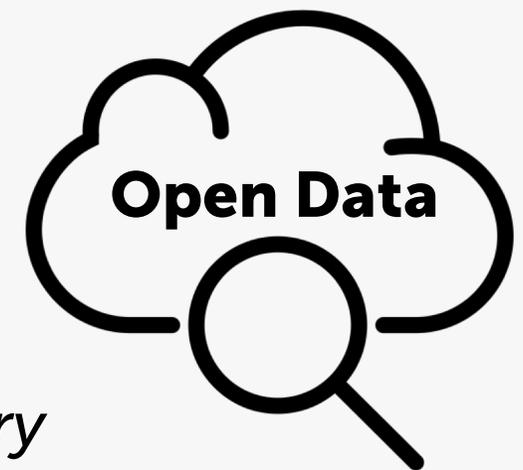
Stephan Diehl
University of Trier, Germany
diehl@uni-trier.de



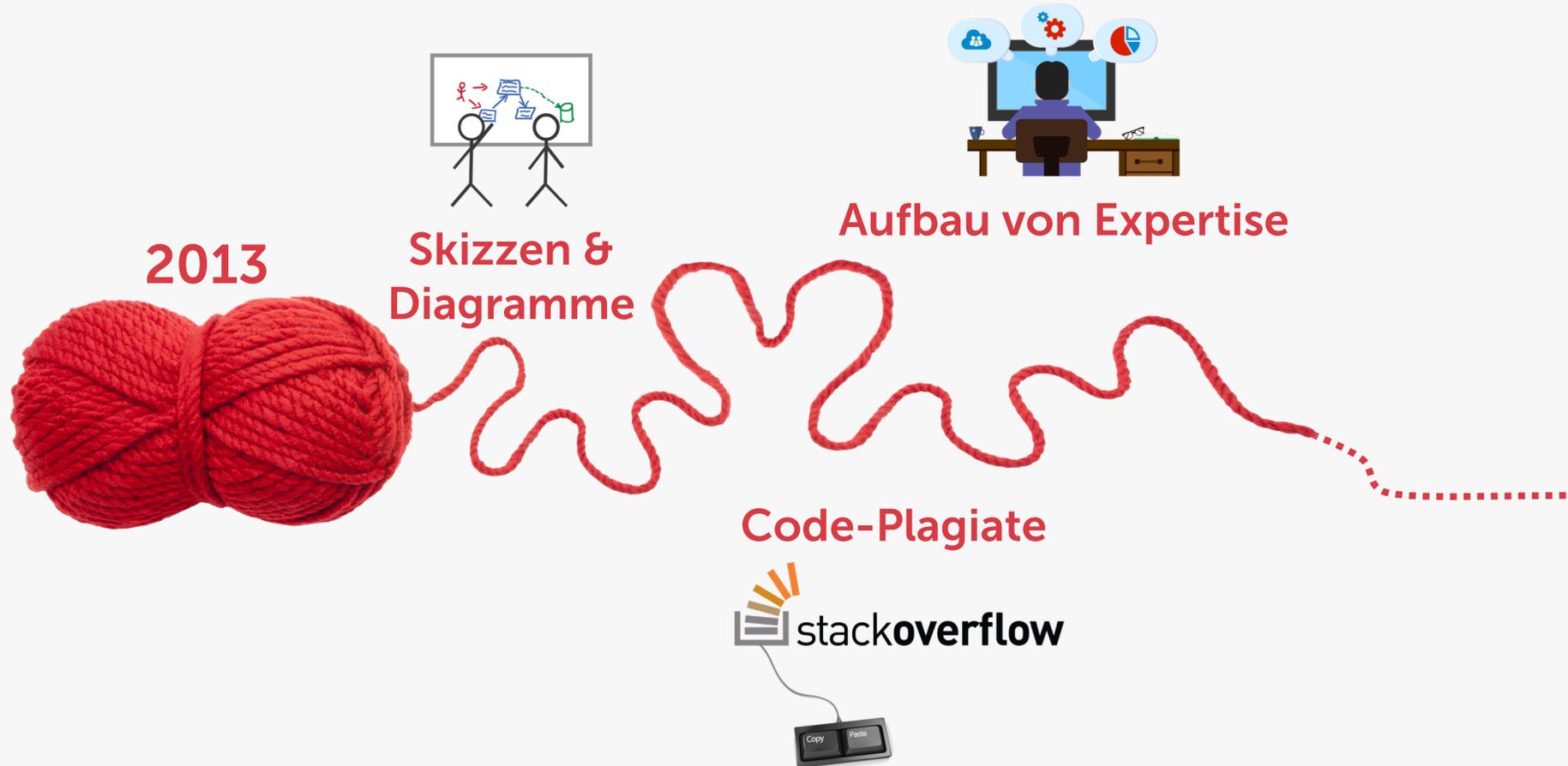
MSR 2018/19

sotorrent.org

Dataset available on Zenodo and BigQuery



Behandelte Themenbereiche



Evidenz-basierte Praxis durch Praxis-basierte Evidenz



 @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>