Worse than

Issues In Sampling Software Developers

Sebastian Baltes
University of Trier, Germany

@s_baltes
research@sbaltes.com

10th International Symposium on
Empirical Software Engineering and Measurement
September, 8-9 – Ciudad Real, Spain

Motivation

- **Reaching out** to professional software developers is crucial part of empirical software engineering research

- **Survey research** is important method to investigate state of practice

- When sampling developers for surveys, several **practical and ethical issues** arise

Content of this talk:

1. The problem of **convenience samples**
2. **Own experience** with different sampling strategies
3. **Ethical** implications of these strategies
4. Assessment strategy for **external validity**

Sampling: Ideal Scenario

Theoretical Population

Who do you want to generalize to?

Study Population

What populations can you get access to?

Sampling Frame

Random Selection

n … get access to them?

Sample

Who is in your study?

Respondents

Who participated?

Sampling: Common Scenario

Main problem: Availability of suitable sampling frames, reachability of participants.

➔ Reliance on available subjects:
convenience sampling, snowball sampling

➔ Likely leads to **biased samples**:

- Self-selection bias
- Researchers contact people from their own social and cultural group
- **Limited generalizability**

Strategies:

- (Try to) select **broad cross-section of the target population**
- **Clear description** of sampling approach and participants
- Take care not to overgeneralize
- Alert readers to the **limitations**

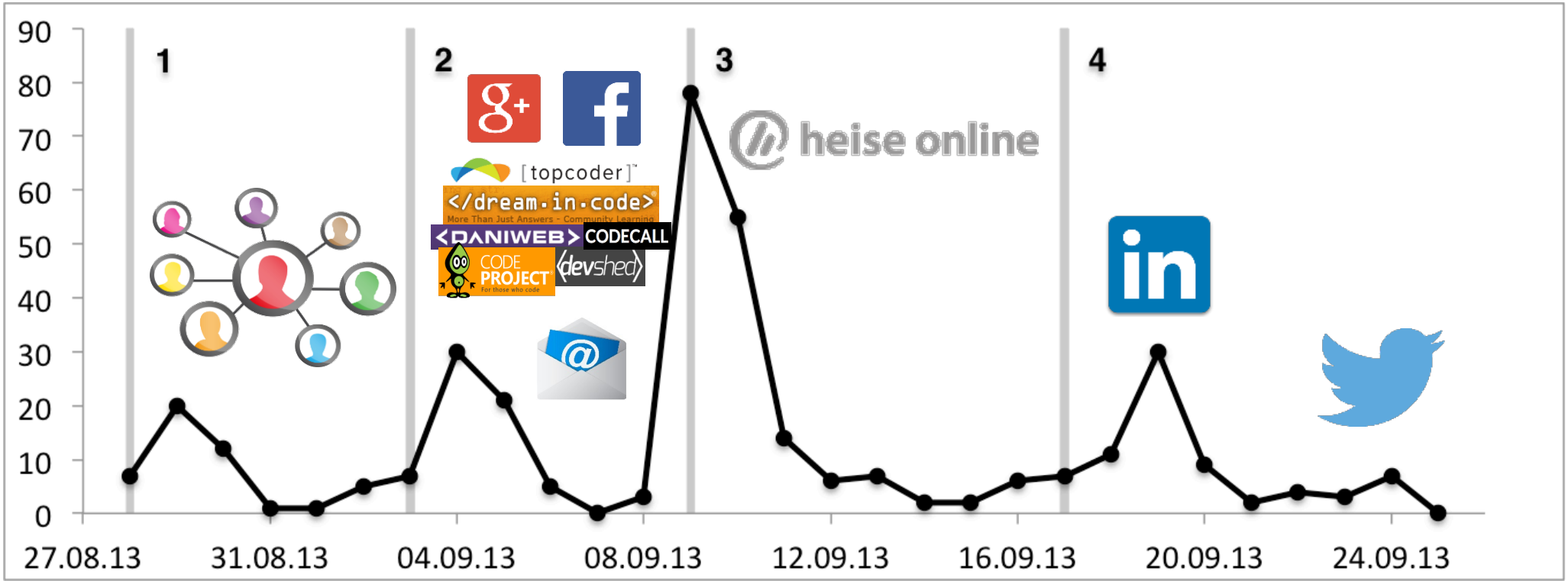
Sampling Strategies

Our Experience with Sampling Strategies

- **Survey** on the usage of sketches and diagrams in software development with **394 participants**
- **Four** recruitment phases

FSE 2014

Sebastian Baltes and Stephan Diehl:
Sketches and Diagrams in Practice



Our Experience: Summary

- **Personal network:**
  - Not very effective
  - May dependent on quality and quantity of network
  - Better suited for other study designs (interviews, controlled experiments)

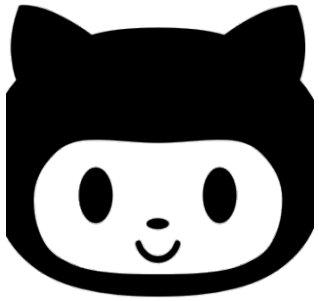
- **Online networks and communities:**
  - Not very effective
  - Mostly positive feedback
  - Some criticism in IRC channels

- **Directly contacting companies:**
  - Difficult to cross company borders without a *gatekeeper*

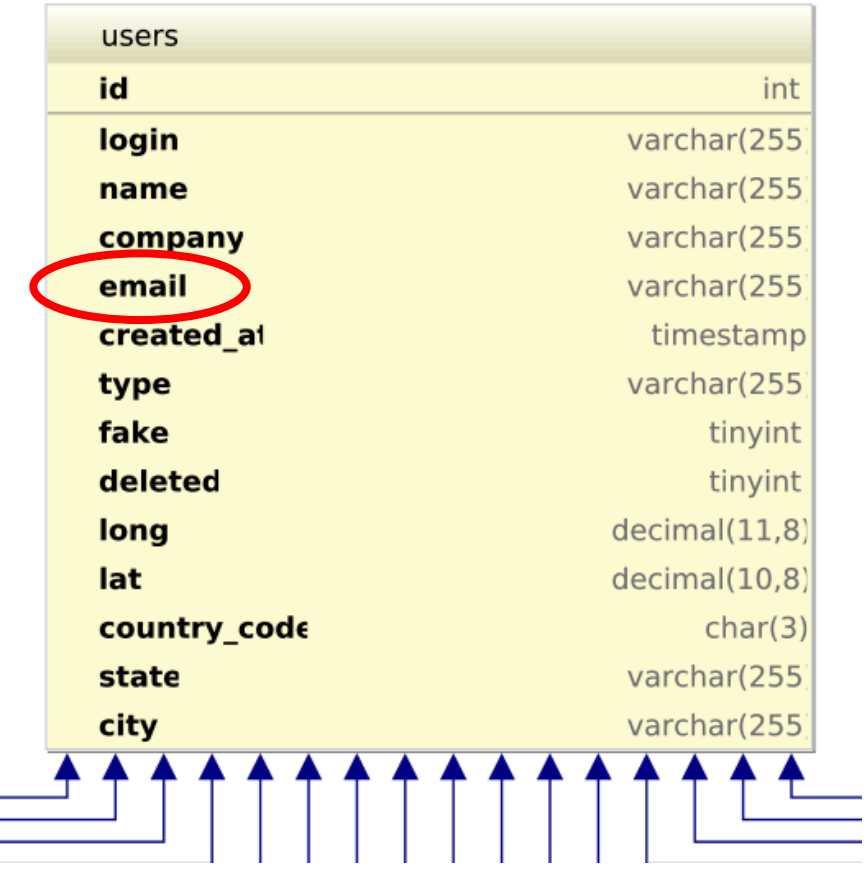
- **Public media:**
  - Most effective and efficient strategy (about 40% of responses)
  - Again gatekeeper in editorial team helpful

- **“Testimonials” (Twitter):**
  - Rather efficient
  - Again problem of biased sample

Alternative: Sampling Using GHTorrent

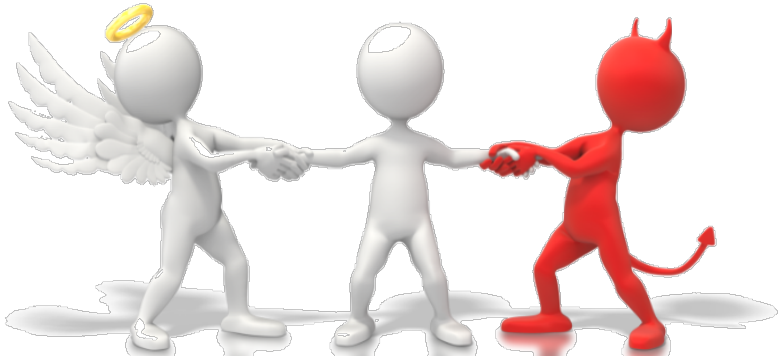


GHTorrent



- GHTorrent:
  - Project collecting **data about public GitHub projects**
  - Available online and as data dump
- Possibility to filter users according to their activity on GitHub
- **Random sampling**
- Email addresses removed in March 2016 after heated discussion on GitHub
- Alternative: Collect email addresses from user profiles or commits
- Convenient, but raises **ethical issues**

Ethical Considerations



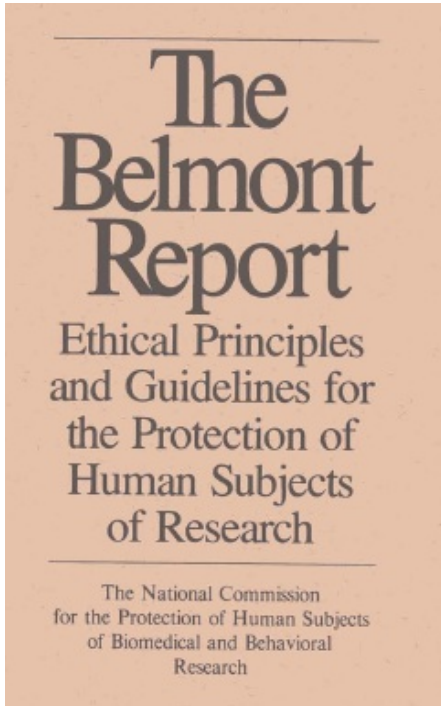
Ethical Considerations

- **Ethics:** "Rules of behavior based on ideas about what is morally good and bad" [Merriam-Webster]
- Legal aspects out of scope for this talk

*"I get emails like this **every week**. You might not realize this but it's majorly annoying and I consider this problem now **worse than spam**, since Google at least filters out spam for me. [...] **[Y]ou send one, I get one per week - or more.** I was playing along for the first 30 or so, and by now (after several hundred emails) **I'm quite annoyed**."*

- Sending mails to users on a large scale causes **costs**, even if they don't participate
- Active users get annoyed and do not answer → **selection bias**

Ethical Considerations: Resources



- USA: *Belmont Report* (1979) and subsequent legislation of *Common Rule* (1981)

- Three guiding ethical principles:
  - **Respect** for research participants
    - Must enter research voluntarily and with adequate information (informed consent)
  - **Beneficence**
    - (1) Do not harm
    - (2) Maximize possible benefits and minimize possible harms
  - **Justice** in participant selection
    - Fairly distribute benefits and burdens of research

Ethical Considerations: Belmont vs. GHTorrent

Sampling using GHTorrent:

- Users may change their behavior due to “survey spam” (e.g., remove email address from profile)

Beneficence?

- Active users are likely to get contacted more often

Justice in partici-pant selection?

- Frequently contacted users may refuse to answer → biased samples

Beneficence?

Ethical Considerations: Resources

CASRO code of ethics has section about "internet research"

Criteria for email recruitment:

(1) substantive **pre-existing relationship**
(2) based on relationship "reasonable expectation" to be contacted
(3) not opted out
(4) **no recruitment via unsolicited emails**

Problematic strategies: Contacting companies and using GHTorrent

- No substantive pre-existing relationship
- Unsolicited emails
- GitHub users did not share email to be contacted for research

Assessing External Validity

Assessing External Validity

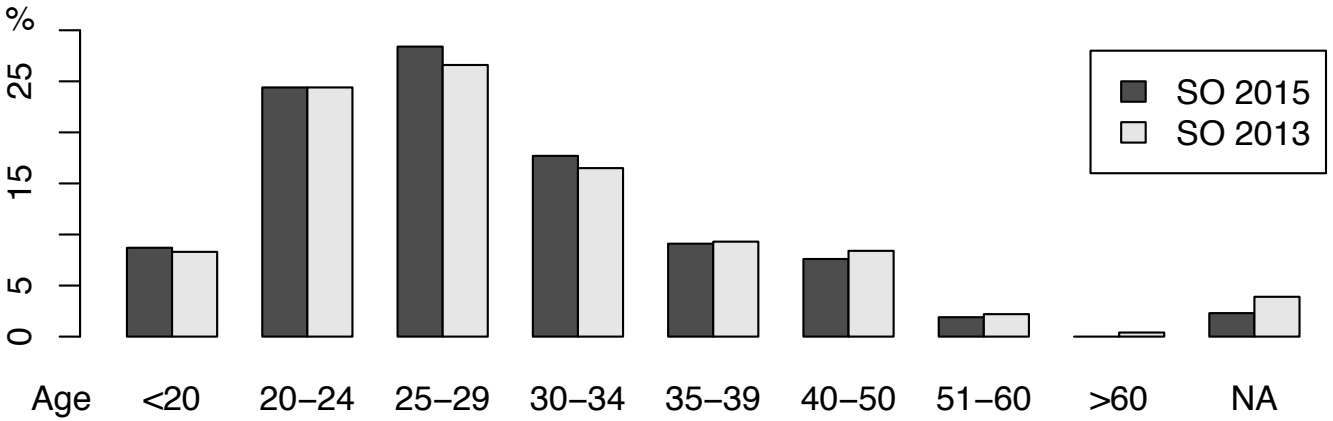
What do we know about the target population of software developers?

Strategy for dealing with convenience samples:

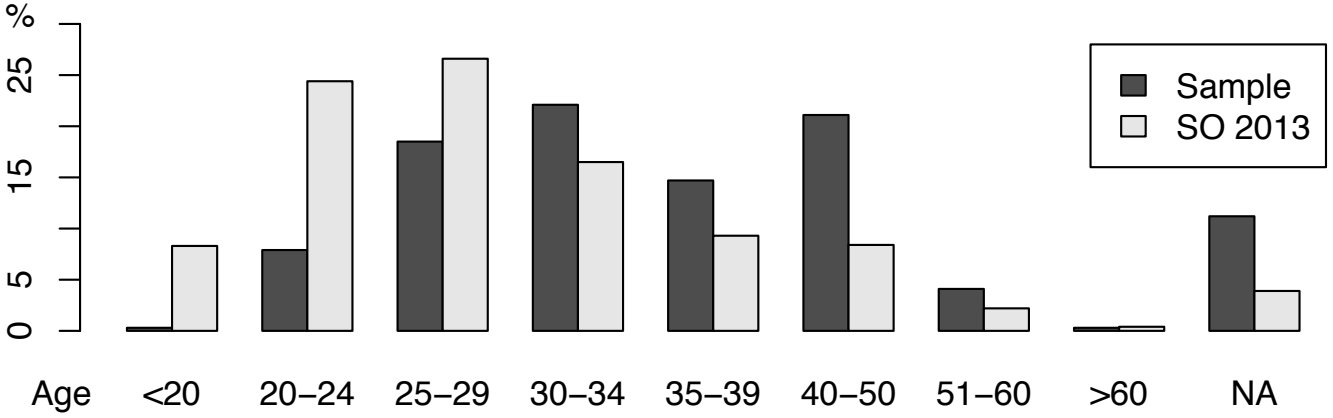
> **"Carefully select broad cross-section of the target population"**

- No structured and systematic database with demographics of software developers available
- Yearly Stack Overflow developer survey (2010-2016)

Assessing External Validity



- No major differences between 2013 (n=7,644) and 2015 (n=26,086) data set



- Our sample **biased towards older and more experienced developers**
- More participants refused to provide their age (5.6% vs. 1.8%)
- Fewer female respondents (2.8% vs. 4.8%)

Conclusion

Conclusion

- **Gatekeepers** are important to cross company borders

- "Testimonials" on Twitter and an article on a IT news website worked best for us

- Using GHTorrent for sampling is compelling, but raises ethical issues
  - We should **discuss ethical implications** of our research at workshops and conferences (see, e.g., CHI and CSCW).
  - Survey with SE researchers about their notion of ethics

- To assess external validity of a survey, a **collection of developer demographics** is needed
  - Systematic literature review (e.g., age, experience, typical sample sizes)
  - Contacting authors of surveys conducted over the past years

Questions?

@s_baltes
research@sbaltes.com

Statement from GitHub

"Our position is that we allow our users to provide their email addresses to the public or make them private. ***If they do make their email addresses public, the public may contact them, including researchers.*** *If they do make their email addresses private, we don't share that information with third parties or allow third parties to access it."*

"We ***offer documentation for keeping email addresses private****. It should be the case that when a user sets his or her email address private in https://github.com/settings/emails, the email address is consistently private for commits and other git functions. However,* ***some users don't remember to also set their emails in Git****, so when they're working on the command line, their email addresses are exposed in commits. The problem there is that* ***we can't control how Git works****; we can only control how GitHub works, and provide as much warning and documentation as possible for integrating with Git."*

New GitHub Privacy Statement

Public Information on GitHub

"Much of GitHub is public-facing. If your content is public-facing, third parties may access and use it in compliance with our Terms of Service. We do not sell that content; it is yours. However, ***we do allow third parties, such as research organizations or archives, to compile public-facing GitHub information****. Your Personal Information, associated with your content, may be gathered by third parties in these compilations of GitHub data. If you do not want your Personal Information to appear in third parties' compilations of GitHub data,* ***please do not make your Personal Information publicly available and be sure to configure your email address to be private in in your user profile****. If you would like to compile GitHub data, you may only use any public-facing Personal Information you gather for the purpose for which our user has authorized it. For example, where a GitHub user has made an email address public-facing for the purpose of identification and attribution, do not use that email address for commercial advertising. We expect you to reasonably secure any Personal Information you have gathered from GitHub, and to* ***respond promptly to complaints, removal requests, and "do not contact" requests from GitHub or GitHub users****."*

Effective date: **August 29, 2016**