

Artifact Evaluation and Competitions

Lightning Talk - Research Software Engineering Panel @ SE24

Ben Hermann - 29.02.2024

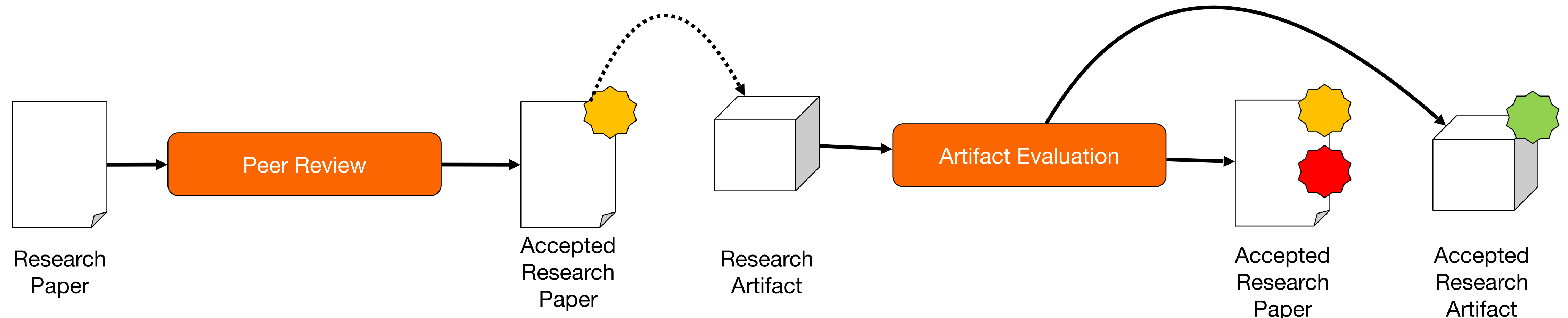
@benhermann.eu

@sse.cs.tu-dortmund.de

Artifact Evaluation

Quick Overview

- In Computer Science there has been a peer review process established for research artifacts: **Artifact Evaluation**
- It was piloted in 2011 at the ESEC/FSE conference
- Since then, it spread to various other conferences in the field and in other fields of computer science



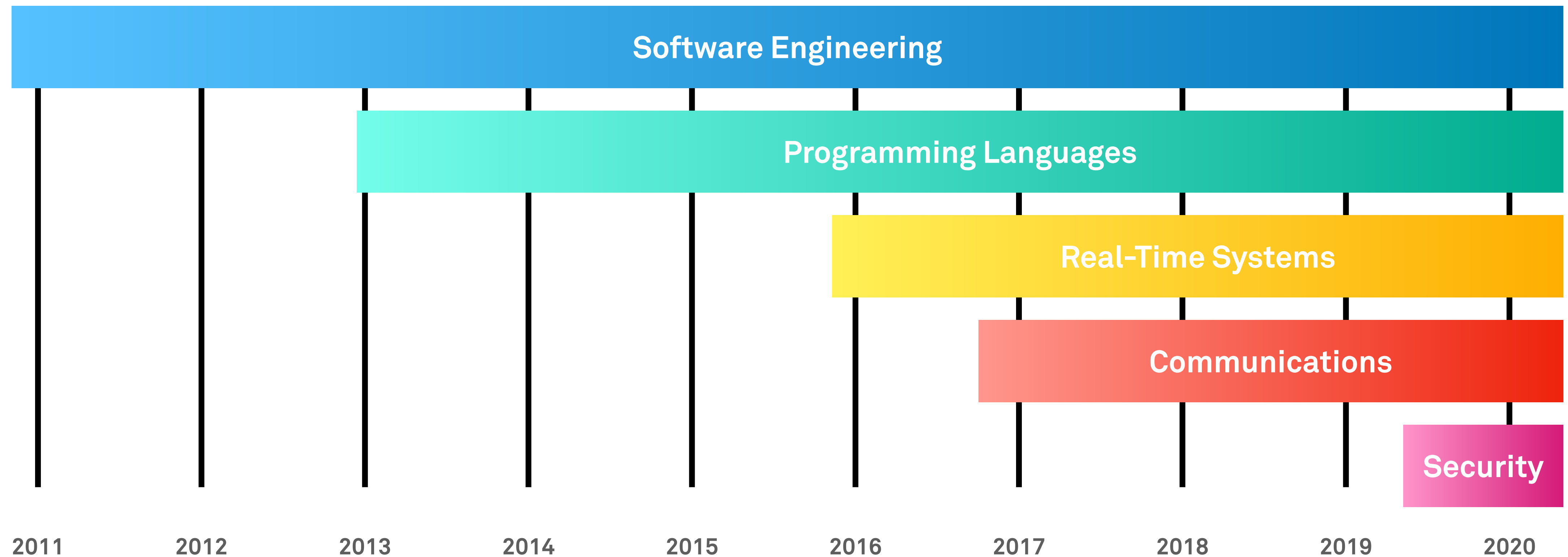
Research Artifact

... a self-contained work result with a context-specific purpose.

Daniel Méndez Fernández, Wolfgang Böhm, Andreas Vogelsang, Jakob Mund, Manfred Broy, Marco Kuhrmann, and Thorsten Weyer. 2019. Artefacts in software engineering: a fundamental positioning. *Software & Systems Modeling* 18, 5 (2019), 2777–2786

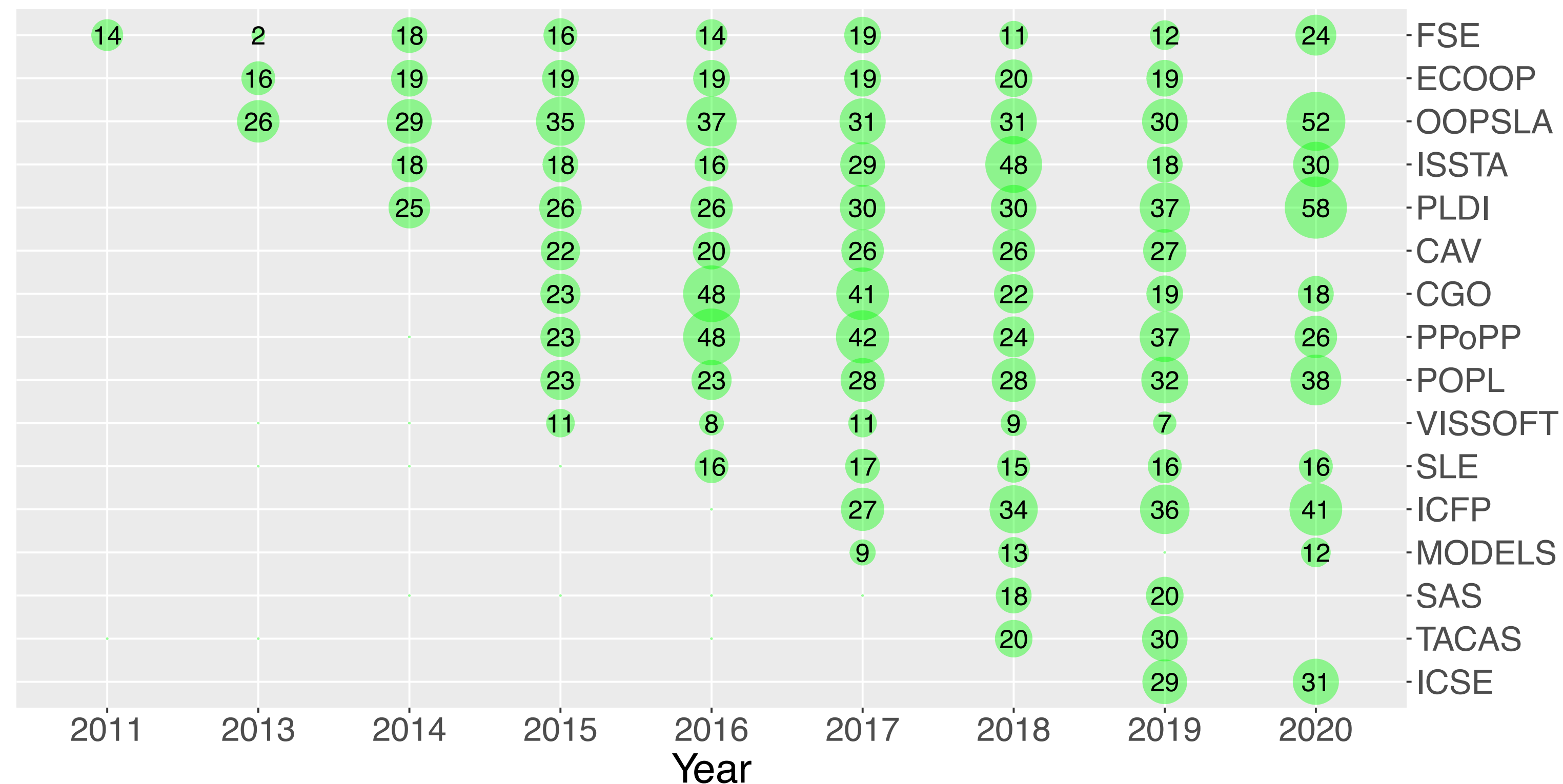
Adoption

A few examples (... and yes, there is more)



Community Effort

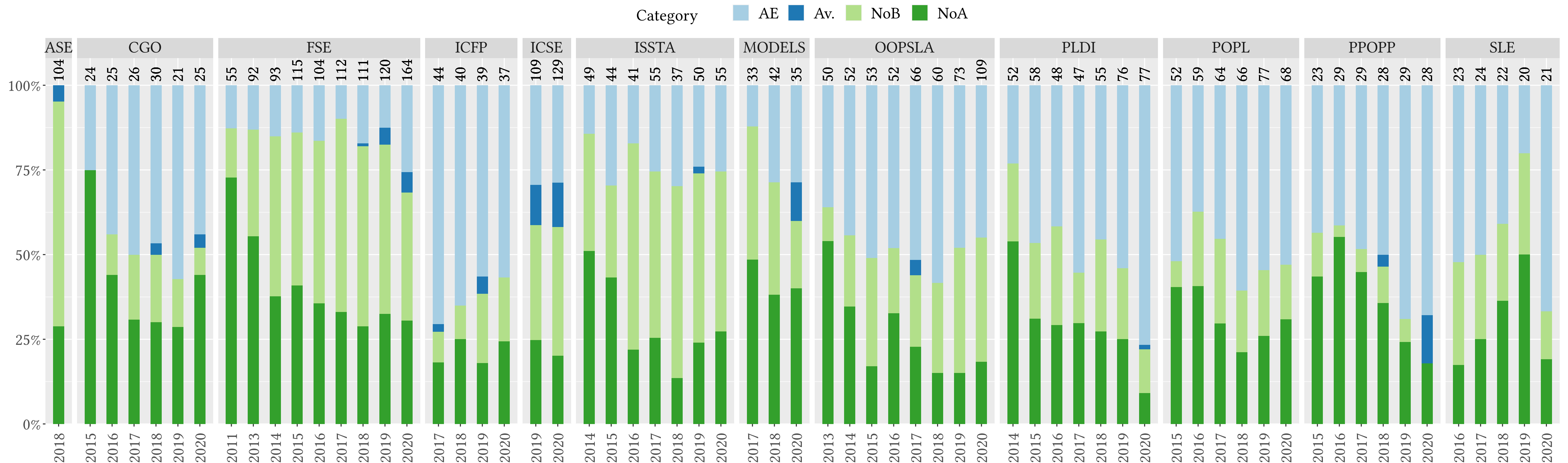
Over 1.000 artifact reviewers in SE and PL and counting



Ben Hermann, Stefan Winter, and Janet Siegmund. 2020. Community expectations for research artifacts and evaluation processes. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020). Association for Computing Machinery, New York, NY, USA, 469–480. <https://doi.org/10.1145/3368089.3409767>

Availability

... not the main issue anymore



Stefan Winter, Christopher S. Timperley, Ben Hermann, Jürgen Cito, Jonathan Bell, Michael Hilton, and Dirk Beyer. 2022. A retrospective study of one decade of artifact evaluations. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022). Association for Computing Machinery, New York, NY, USA, 145–156.

But what are the Quality Expectations?

Complete Community Survey

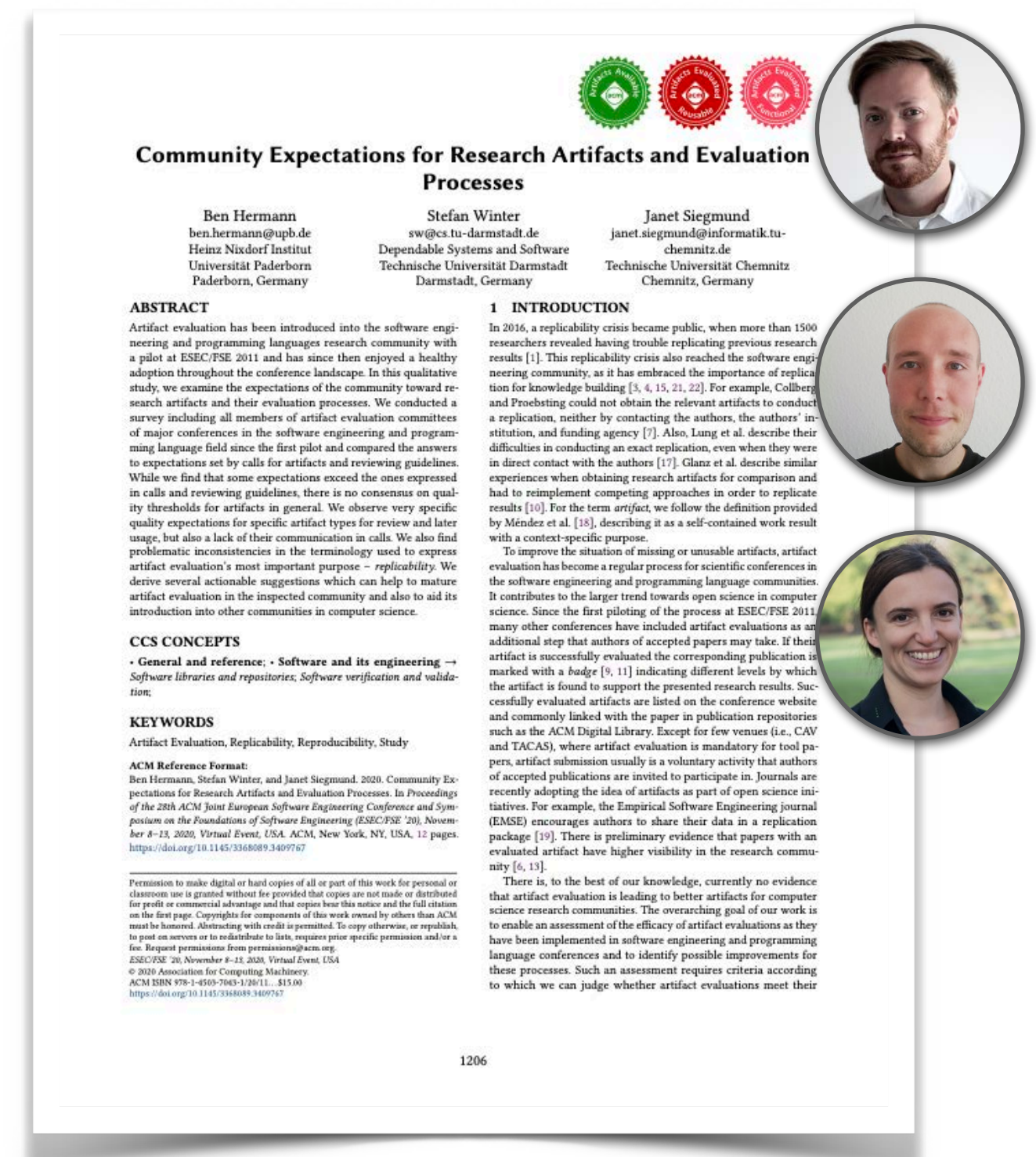
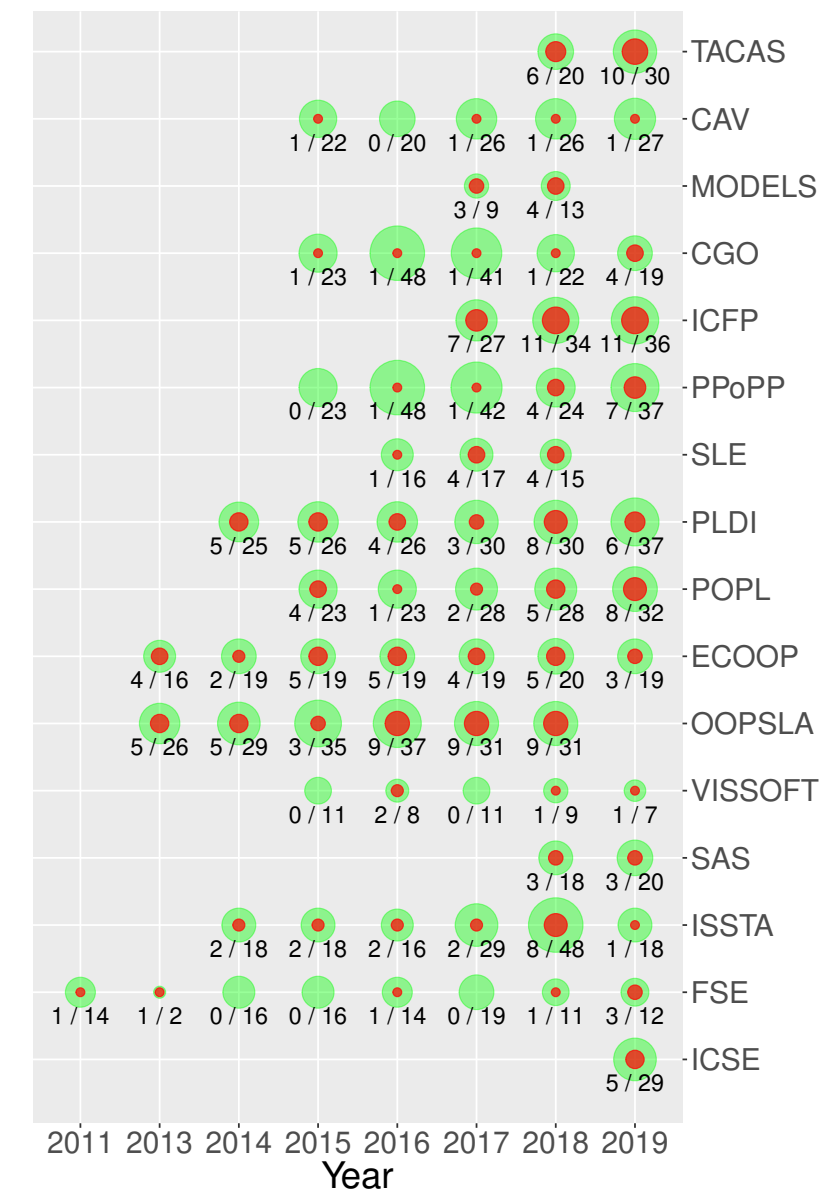
We asked all of the 1,034 reviewers and chairs involved in AE between 2011 and 2019 — 257 (~ 25%) replied.

In a nutshell:

- Process is still neither mature nor stable
- Review criteria vary widely
- Communities struggle to position themselves between different goals: high availability vs. quality reuse vs. replication

Update on this: It got better since then

Community Expectations for Research Artifacts and Evaluation Processes, ESEC/FSE 2020



Boosting Your Bibliometrics?

Bad news: If there ever was an effect it's over now 😞

- We inspected 3650 papers from 64 conference instances
- These instances happened over the past 10 years
- We only find statistically significant citation count differences between short papers with "inofficial" artifacts and short papers without artifacts published at FSE 2014 and FSE 2019
- For all other papers we inspected there was no indication that papers with artifacts (badged or not) are better cited

Creating and publishing research artifacts does not generally have beneficial effects on citation counts.

A retrospective study of one decade of artifact evaluations , ESEC/FSE 2022

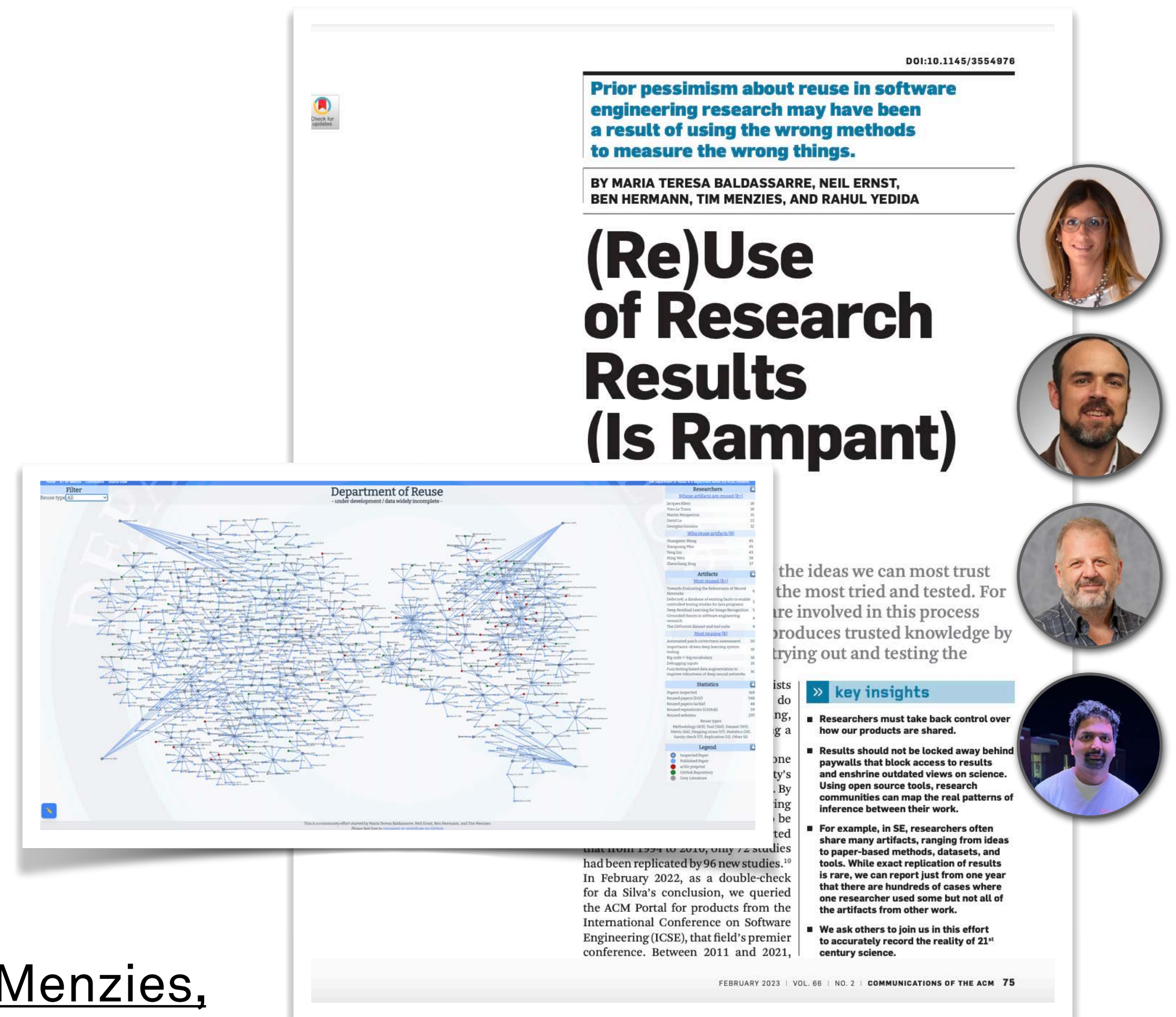


Reuse in Software Engineering Research

Crowdsourced Data Collection

- Formal replication studies are rare
- BUT we are VERY GOOD at reuse
- Prelim: 168 papers from 2020
- Website: Browse and Participate
- Adding a couple of more metrics to break h-index singularity

Maria Teresa Baldassarre, Neil Ernst, Ben Hermann, Tim Menzies, and Rahul Yedida. 2023. (Re)Use of Research Results (Is Rampant). Commun. ACM 66, 2 (February 2023), 75–81.



<https://www.reuse-dept.org/>

Competitions

- Some research communities run competitions
- Example: SV-COMP in the Software Verification Community
- Observations:
 - Competitions harmonize input- and output-formats and therefore advance comparability and interoperability
 - Inspire novel approaches and improvements, but require some maturity in the field
- But, competitions are very rare. **Why is this the case?**