[ dɪsiː aɪtiː ]
Daimler Center for
Automotive Information Technology Innovations

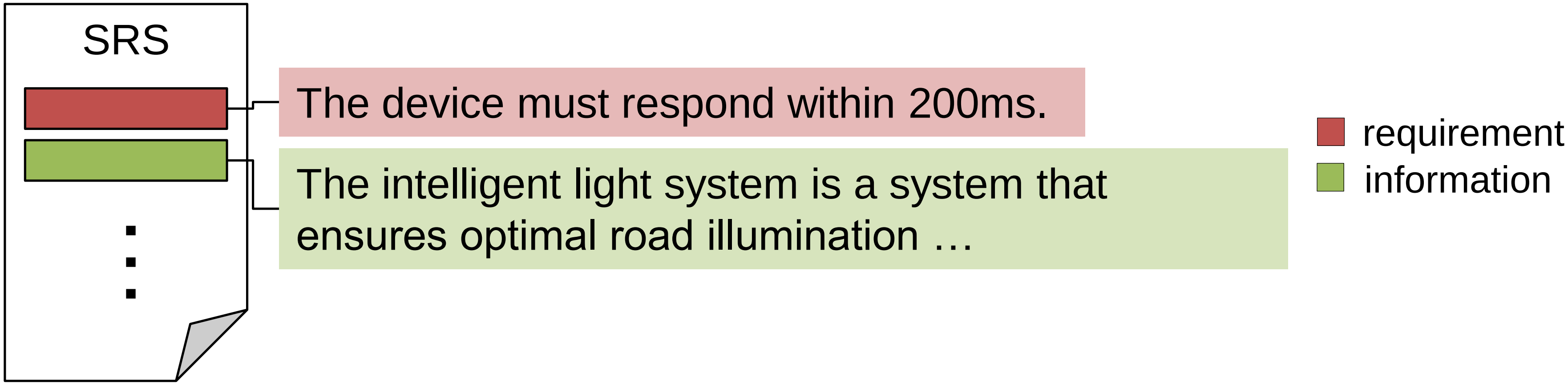
# Using Tools to Assist Identification of Non-Requirements in Requirements Specifications – A Controlled Experiment

REFSQ’18, Utrecht, The Netherlands

Jonas Paul Winkler, Andreas Vogelsang
DCAITI, Technische Universität Berlin

March 20, 2018

# Background – Requirements vs Information



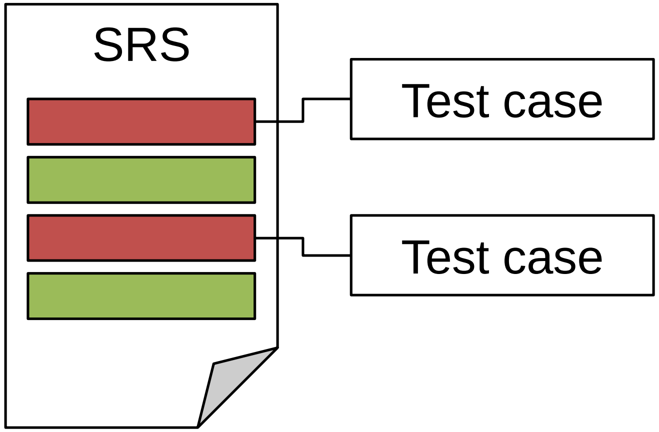
The device must respond within 200ms.

The intelligent light system is a system that ensures optimal road illumination …

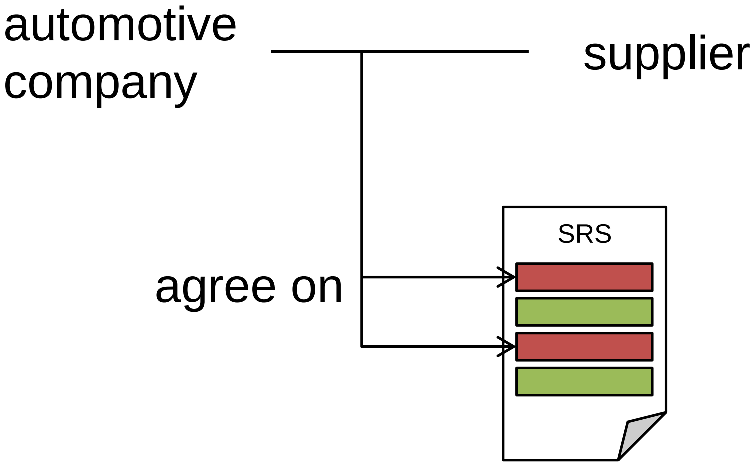
🟥 requirement
🟩 information

**Why is this important?**

1) Test case creation



2) Document change management

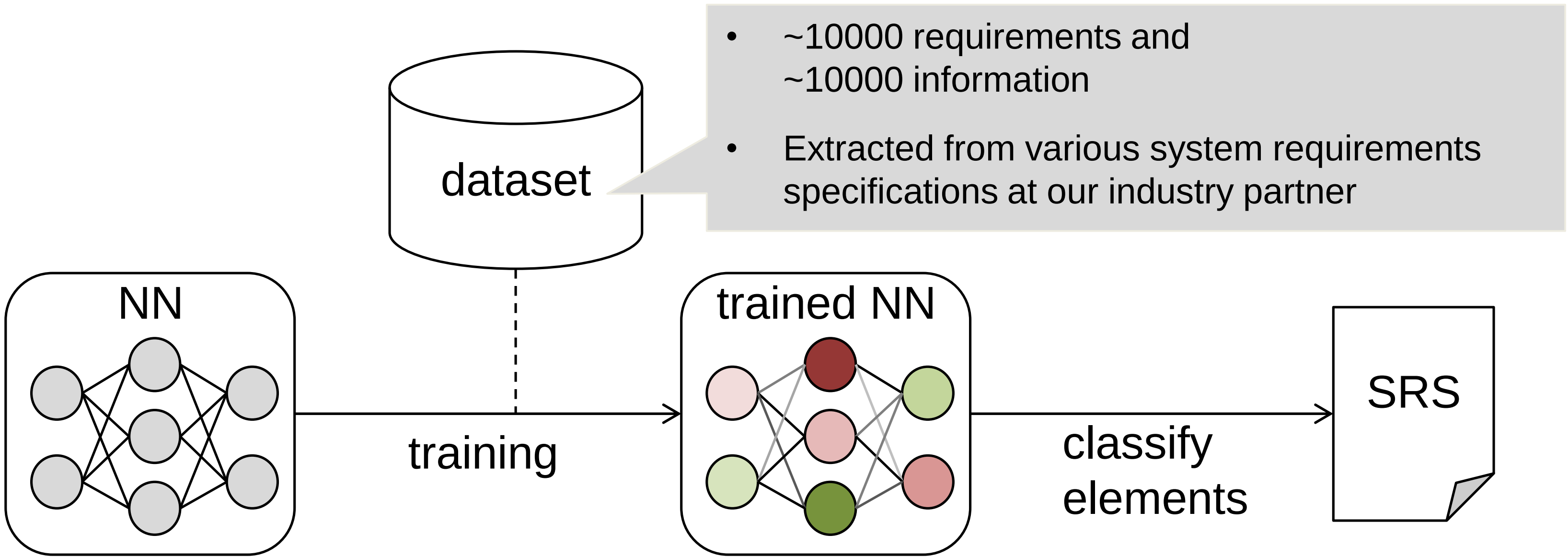


# Background – Classifying Requirements

- Explicit labelling of requirements specification content elements at our industry partner („object type“)
- Quality reviews: requirement documents are manually inspected for defects
  - Common quality criteria: correct, unambiguous, complete, verifiable…
  - Also: correct labelling regarding object type
- Manual labelling is time-consuming and error-prone

**Our goal:**
Assist requirements engineers in verifying correct labelling of requirements and non-requirements

# Background – Automatic Classification



- ~10000 requirements and ~10000 information
- Extracted from various system requirements specifications at our industry partner

- We did: Integration into a tool that issues warnings on incorrectly labelled items (“defects”)

**Main question:** Does using such a tool provide benefits?

Winkler, Jonas P.; Vogelsang, Andreas (2016): Automatic Classification of Requirements Based on Convolutional Neural Networks. In : 3rd IEEE International Workshop on Artificial Intelligence for Requirements Engineering (AIRE). Beijing.

# Research Questions

1. Does the usage of our tool enable users to detect more defects?

2. Does the usage of our tool reduce the number of defects introduced by users?

3. Are users of our tool prone to ignoring actual defects because no warning was issued?

4. Are users of our tool faster in processing the documents?

5. Does our tool motivate users to rephrase requirements and information content elements?

# Experiment Design

- Two-by-two crossover study with students

- Students search and correct defects in a given SRS

- Control Group: Students without tool (manual review)
- Treatment Group: Students with tool (tool-assisted review)

| | Group 1 | Group 2 |
|---|---|---|
| Session 1 (SRS #1) | Manual | Tool-assisted |
| Session 2 (SRS #2) | Tool-Assisted | Manual |

- Compare the performance of students from both groups

# Experiment Materials

- Excerpts from actual work-in-progress SRS

| Document Name | Total Elements | Accuracy |
|---|---|---|
| Wiper Control | 115 | 82.6% |
| Window Lift | 261 | 75.8% |
| Hands Free Access | 147 | 85.0% |

- Size reduced to fit our experiment schedule
- Anonymized names as requested by our industry partner
- Determined true object type of all content elements

- Experiment was repeated after publishing
  - Presented in paper: Wiper Control, Window Lift
  - Performed after publishing: Wiper Control, Hands Free Access

# Evaluation Metrics & Hypotheses

- Defect Correction Rate:

$$DCR = \frac{Defects\ Corrected}{Defects\ Inspected}$$

- Defect Introduction Rate:

$$DIR = \frac{Defects\ Introduced}{Elements\ Inspected}$$

- Unwarned Defect Miss Rate:

$$UDMR = \frac{Unwarned\ Defects\ Missed}{Unwarned\ Defects\ Inspected}$$

- Time Per Element:

$$TPE = \frac{Total\ Time\ Spent}{Elements\ Inspected}$$

- Element Rephrase Rate:

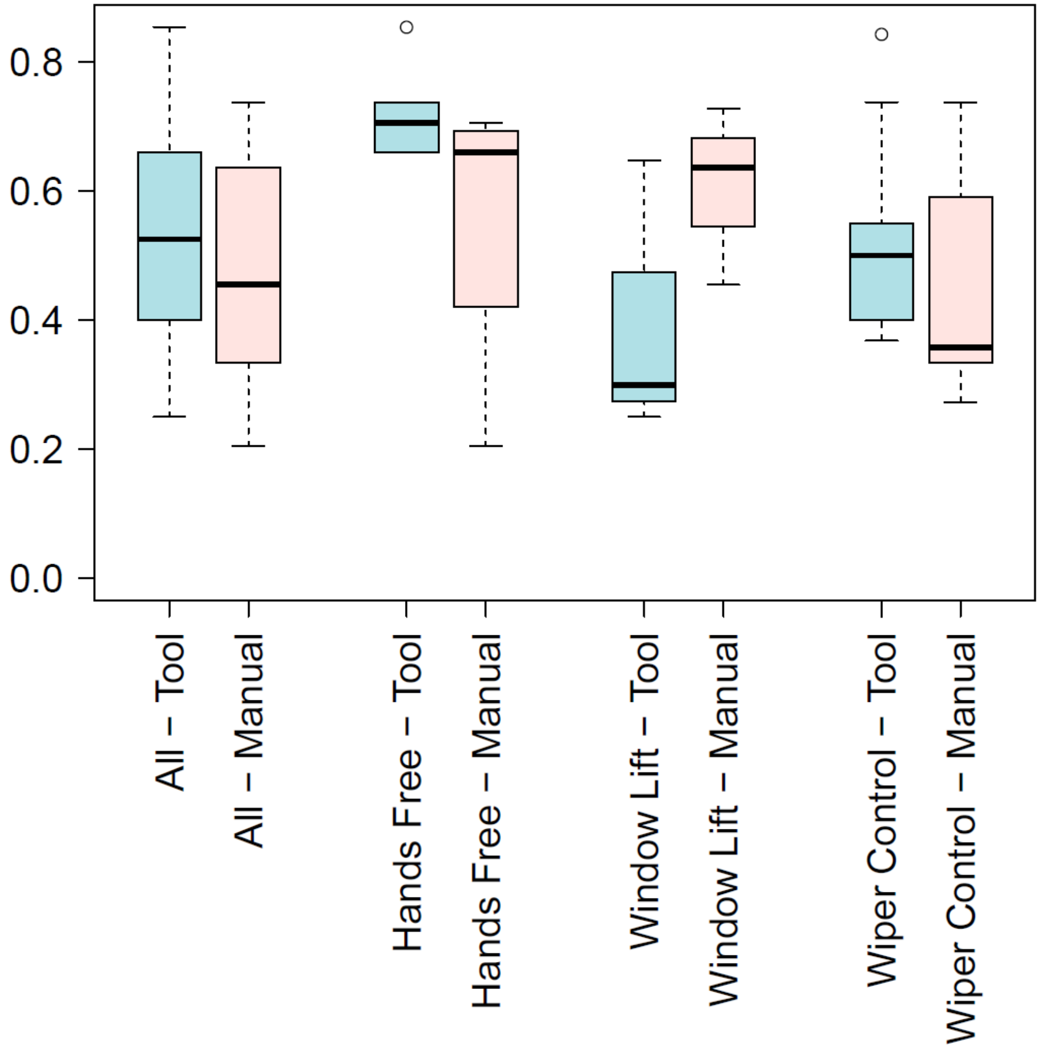
$$ERR = \frac{Elements\ Rephrased}{Elements\ Inspected}$$

# Result Overview

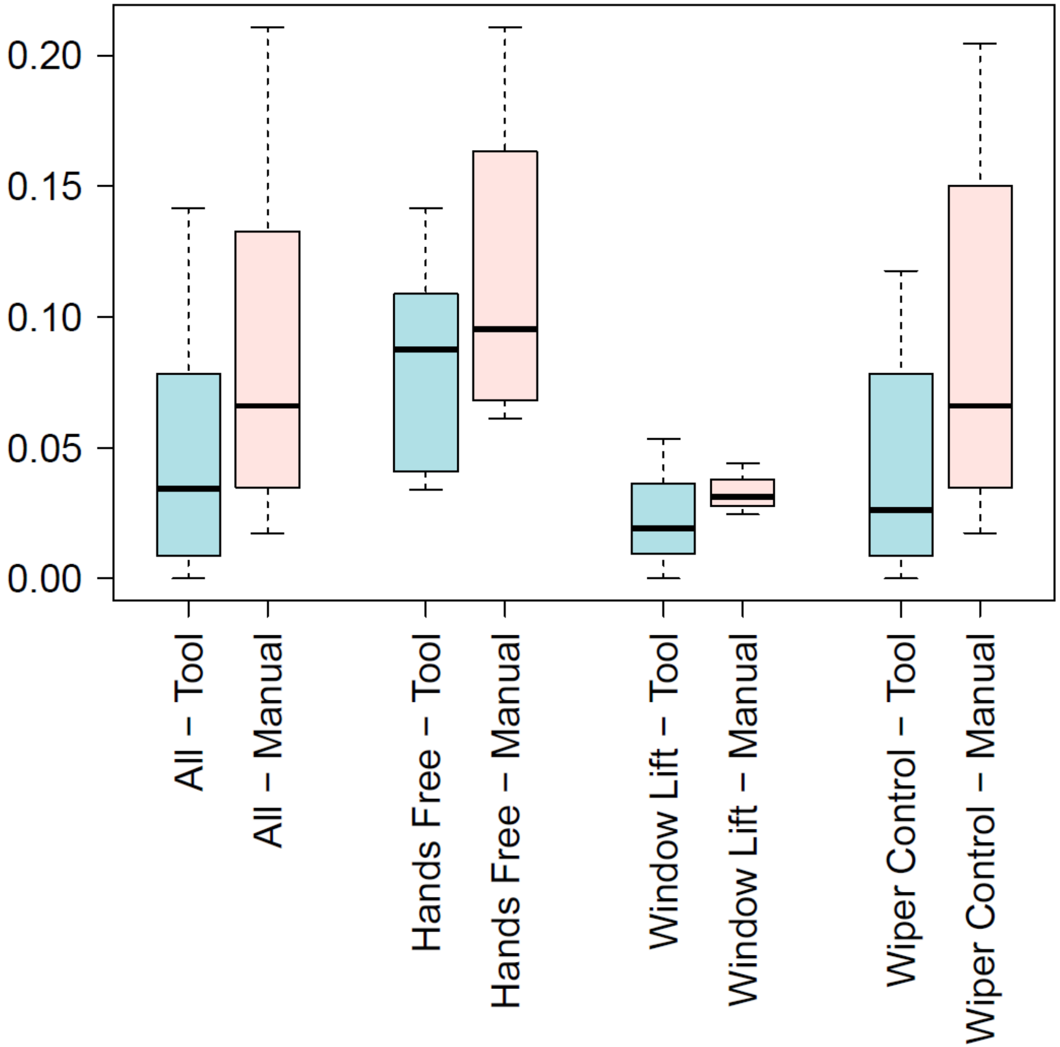
- Total number of students per experiment:
  - ~25 (experiment #1), ~20 (experiment #2)

| Document | Manual group | | Tool-assisted group | |
|---|---|---|---|---|
| | # reviews | # elements | # reviews | # elements |
| Exp #1 (Wiper Control) | 7 | 506 | 7 | 749 |
| Exp #1 (Window Lift) | 4 | 772 | 3 | 435 |
| Exp #2 (Wiper Control) | 5 | 575 | 4 | 460 |
| Exp #2 (Hands Free) | 4 | 588 | 5 | 691 |
| **Total** | **20** | **2441** | **19** | **2335** |

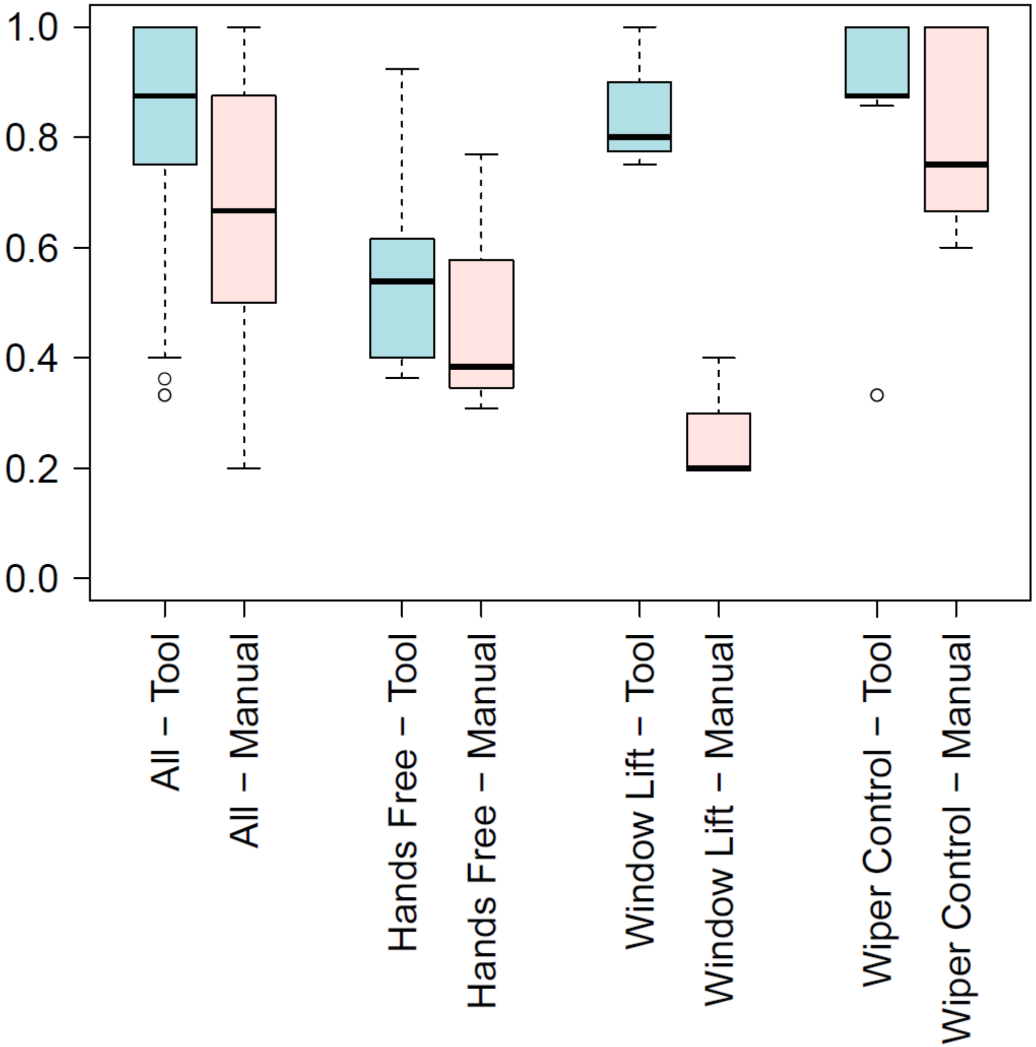
# Defect Correction Rate



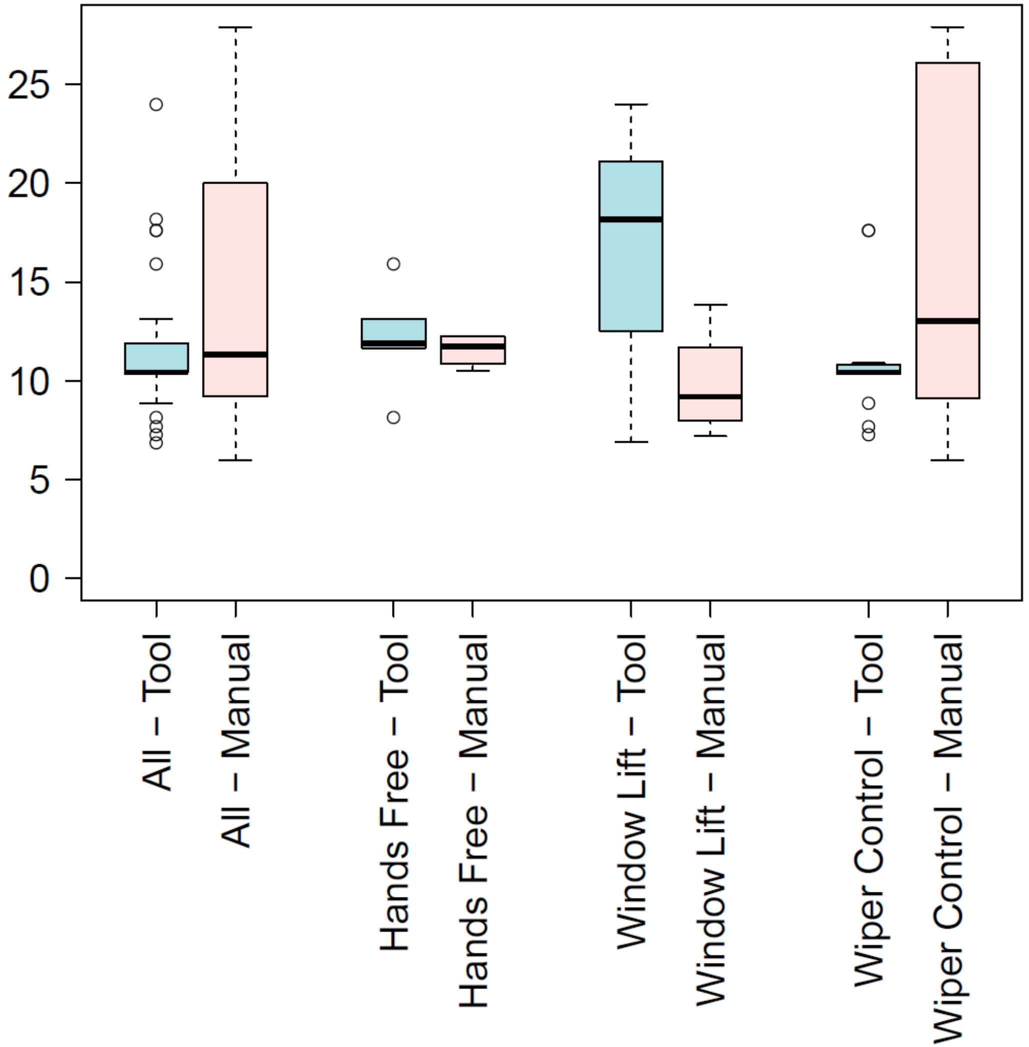
# Defect Introduction Rate



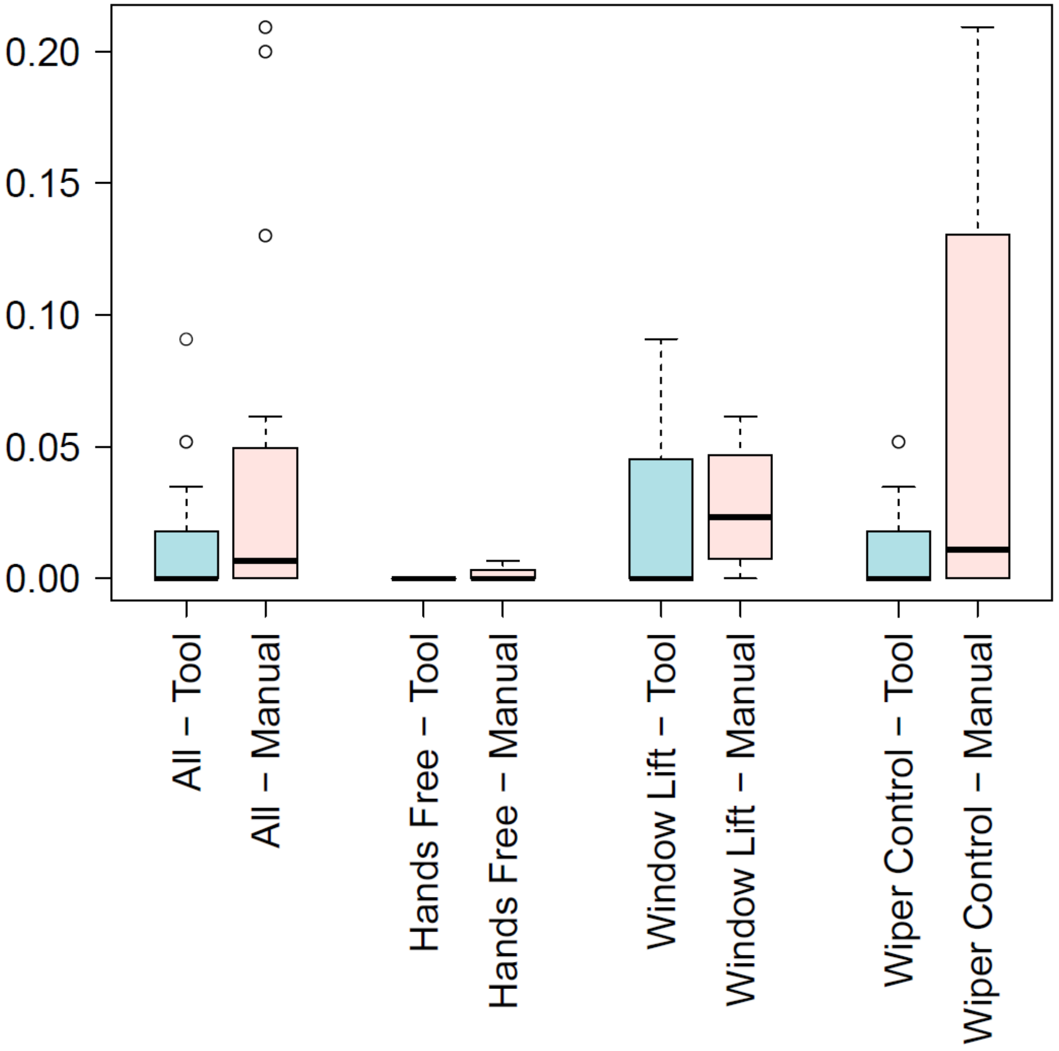
# Unwarned Defect Miss Rate



# Time Per Element



# Element Rephrase Rate



# Summary of Results

- **RQ1**: Users of our tool detect more defects, given that the accuracy is high enough.
- **RQ2**: Less defects are introduced when our tool is used.
- **RQ3**: Users are more likely to miss unwarned defects.
- **RQ4**: On our group of students, time did not improve significantly.
- **RQ5**: Students were not inclined to rephrase more elements when the tool was used.

# Threats to Validity

- Construct validity
  - Number of Participants
  - Definition of gold standard

- Internal validity
  - Maturation
  - Communication between groups
  - Time limit

- External validity
  - Students are no RE experts

# Summary & Future Work

- Tool support enables users to find more defects
- Repeated tool usage may also improve review time (maturation)
- Tool usefulness largely depends on classifier accuracy
- Future Work
  - Collect more data points
  - Repeat experiment with RE experts

**Thank you.**

jonas.winkler@tu-berlin.de ✉