# Leveraging AI for the classification of SE artifacts

## - industrial cases working with defect reports


**Work by: Muhammad Laiq, Nauman bin Ali, Emelie Engström and Jürgen Börstler**

nauman.ali@bth.se


# NAUMAN BIN ALI

- Senior lecturer (Docent) at Blekinge Institute of Technology, (BTH) Karlskrona
- Moved to Vallentuna (30 kms north)
- Industry collaborators:

SOFTHOUSE
SONY
ALSTOM
qvantel

AXIS COMMUNICATIONS
ERICSSON
TEST SCOUTS
INGKA™

- **Research interests**
  - Test case and suite quality
  - Test case selection and prioritization
  - Test flakiness
  - Visual and data analytics for decision support
  - Levering test results for decisions about:
    - test scoping
    - release readiness
  - Anomaly detection in cloud based products and services

BLEKINGE TEKNISKA HÖGSKOLA • BTH •

# Defect report

- bug report,
- trouble report,
- issue report,
- failure report,
- error report,
- incident report,
- problem log

Description of a mismatch between the experienced and expected behavior of a system

# Not all defect reports are equal

It's Not a Bug, It's a Feature

Type of solution required:

- Code correction – to fix an error in the implementation
- Documentation update – to fix an error/clarification in the description of features
- Training and education – to improve awareness of the systems
- New feature development – requires a decision where the functionality should be added to the system

Who should be involved in the investigation and resolution of the bug report, and when?

Herzig, Kim, Sascha Just, and Andreas Zeller. "It's not a bug, it's a feature: how misclassification impacts bug prediction." *2013 35th international conference on software engineering (ICSE)*. IEEE, 2013.

# Defect report classification

Why categorize:

| Task | Classification |
| --- | --- |
| Should developers be involved? | Defect classification based on the type of resolution required.<br><br>A solution may require a code change, customer or support personnel education, an update to the documentation, or requesting a new feature. |

**Table 6** Lead time (mean calendar days from submission to resolution) and median priority (A: High, B: Medium, C: Low) for valid (V) and invalid (I) bug reports

| Product | Lead time (V) | Lead time (I) | Priority (V) | Priority (I) |
| --- | --- | --- | --- | --- |
| P1 | ∼15 | ∼18 | C | B |
| P2 | ∼19 | ∼16 | B | B |
| P3 | ∼15 | ∼13 | B | B |

# Not straight forward

- There are several review steps to filter and sort the defect reports manually.
- The sheer volume of defect reports in a large-scale product makes it difficult, time-consuming, and error-prone to identify this early on correctly.

| Case | Product size | Code base age | People | Bug reports | Not requiring a code correction |
|---|---|---|---|---|---|
| P1 | Large | 10 + years | 100 | 3300 | 13% |
| P2 | Medium | 10 + years | 100 | 3400 | 16% |
| P3 | Small | 8 + years | 70 | 1000 | 18% |
| 5 OSS projects | Small to medium | mix | | 5591 | 42.6% |

# Motivation for the use of AI-Based solutions

**Classification:** Supporting defect report *routing* by automatic and early assessment of the solution required for the reported defect.

**Understanding:** Supporting automatic detection of themes in defect reports classes to gain *insights*. For example,

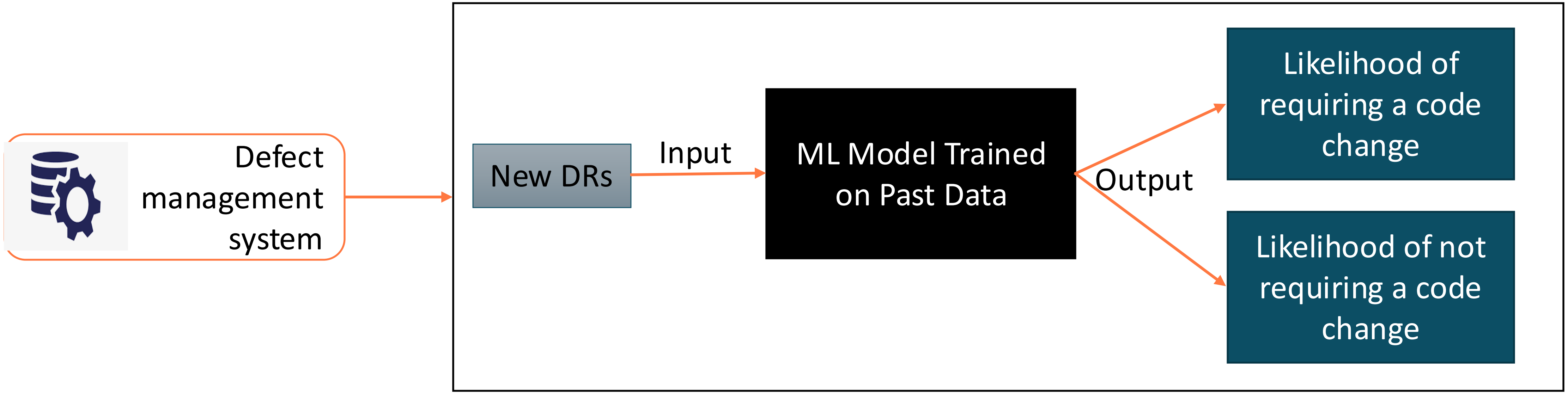
- Help identify common causes of bug reports, which practitioners can use to propose improvements.
- Help identify which features require better documentation.

# AI Methods for Classification of Defect Reports

| Attributes | NB | LR | SVM | SGD | KNN | RF | DT | AutoML-30 | AutoML-60 | AutoML-90 | RoBERTa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | ++ | ++ | ++ | ++ | − | + | + | +++ | +++ | +++ | +++ |
| Computational cost for training the models | + | + | + | + | + | + | + | − | − | − | −− |
| Generalizability | − | + | + | − | − − − | −− | − | − − − − − | ++ | ++ | + |
| Scalability | + | + | + | + | + | − | + | − | − | − | − |
| Learning curve | − | − | − | − | − | − | − | + | + | + | − |
| Explainability (global) | + | + | + | + | + | + | + | − | − | − | − |
| Explainability (local) | − | − | − | − | − | − | − | − | − | − | − |
| Maintenance of the model | − | − | − | − | − | − | − | + | + | + | + |
| Availability of open-source tools | + | + | + | + | | | | | | | |
| Technical support | + | + | + | + | | | | | | | |

| Attributes | NB | LR | SVM | SGD | KNN | RF | DT | AutoML-30 | AutoML-60 | AutoML-90 | RoBERTa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy (AUC) on three datasets | | | | | | | | | | | |
| D1 | 0.61 | 0.63 | 0.65 | 0.63 | 0.50 | 0.57 | 0.62 | 0.67 | **0.68** | **0.68** | **0.69** |
| D2 | 0.81 | 0.81 | 0.80 | 0.80 | 0.73 | 0.72 | 0.71 | 0.86 | **0.87** | **0.88** | **0.84** |
| D3 | 0.74 | 0.81 | 0.81 | 0.80 | 0.69 | 0.77 | 0.78 | ** | **0.85** | **0.86** | **0.85** |

# Classification approach



# From Prediction to Explanation

Title: ------------------------------------------------------------
Impact: ------------------------------------------------------------
------------------------------------------------------------
DESCRIPTION: ------------------------------------------------------------
------------------------------------------------------------

## Input

Valid Case: DR Id = 25

- **Submitter Experience** (Submitter validity rate = 0.95, Total DRs submitted in past = 20, DRs submitted in past 90 days =1)
- **DR Description** (Has testcase=False, Has attachment=True, other text features )
- **DR Priority** = Medium

## ML Model

## Output

Requires code correction:
Yes = 72%
No = 28%

0 **Base Value: 0.63** 1

Other features | Condition (text feature) | Total TRs submitted in past | Change (text feature) | Submitter validity rate

Sum = 0.72

name (text feature) | statement (text feature) | list (text feature) | Other features

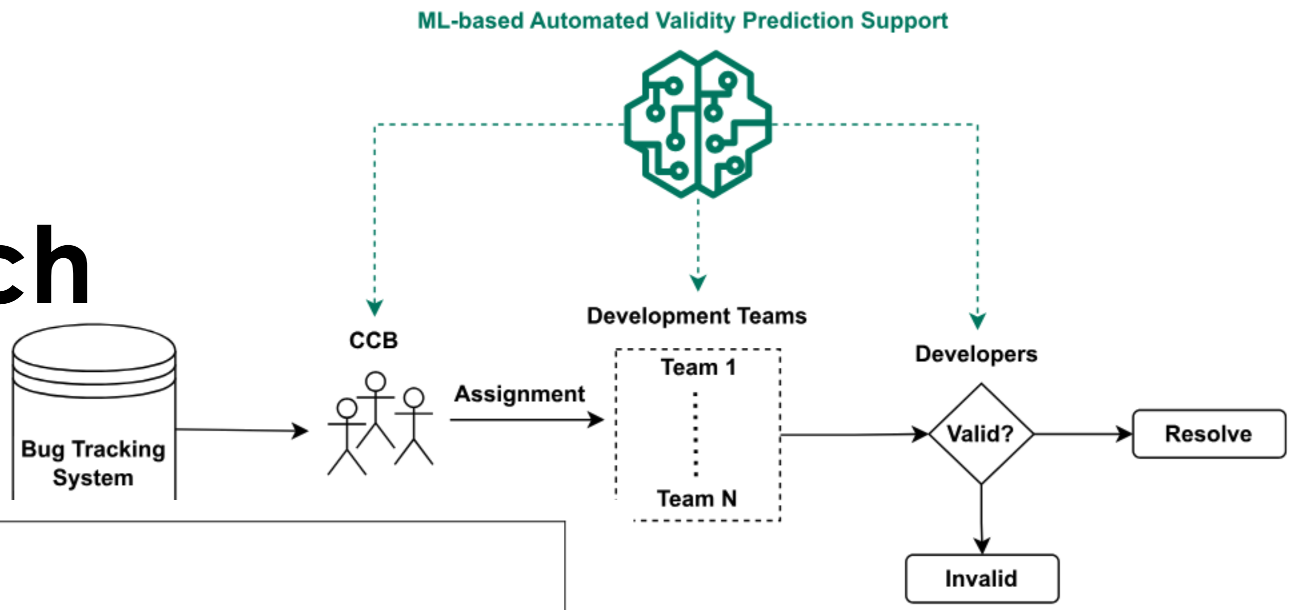
Sum = 0.28

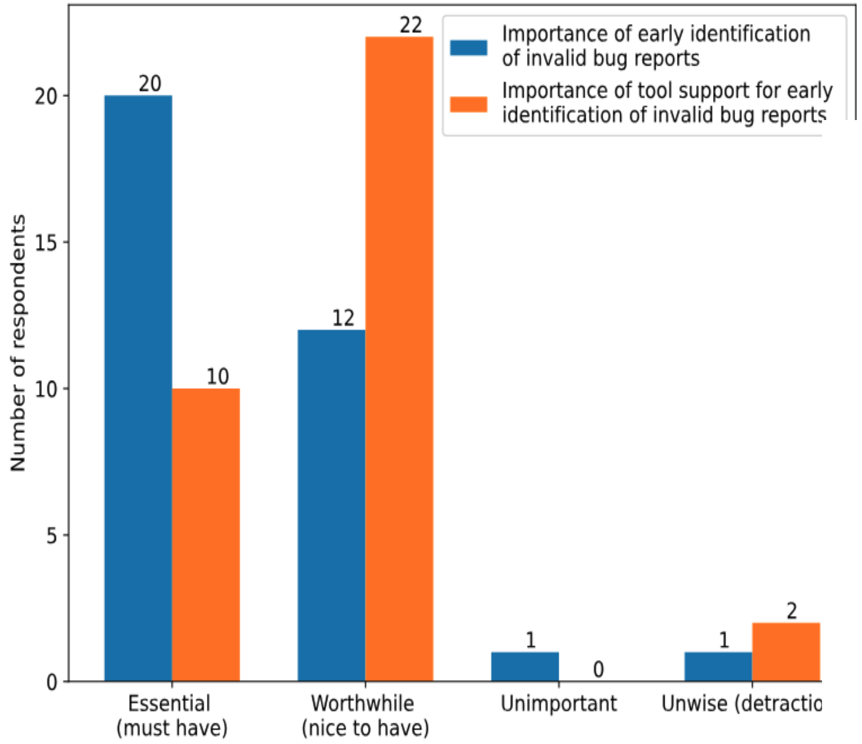
Major Contributor is Submitter Validity Rate

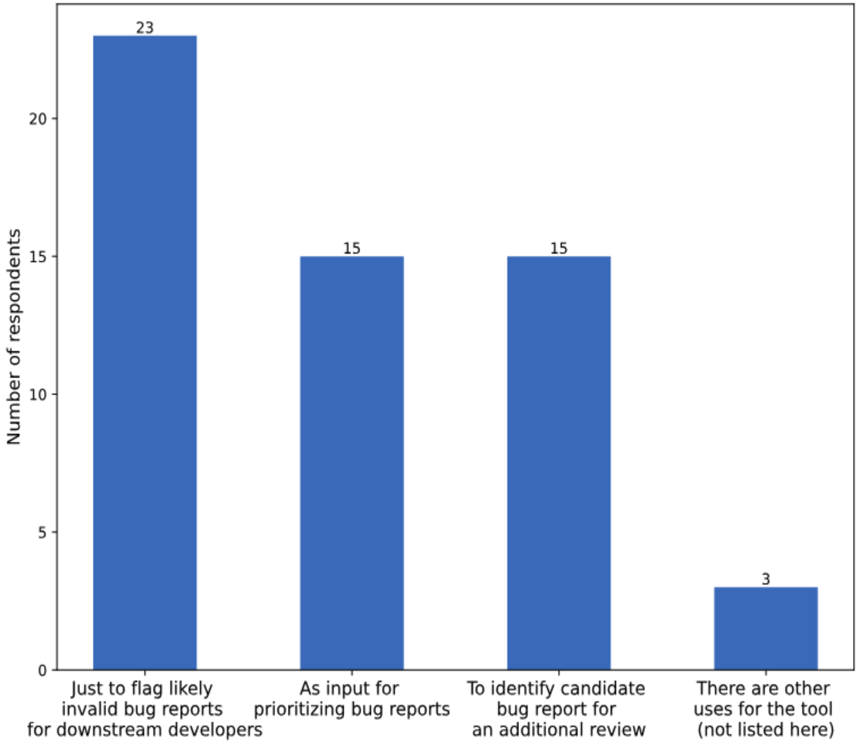
*Reasons why the model suggests this TR will likely be valid*

*Reasons why the model suggests this TR will likely be Invalid*

# Potential use of the classification approach



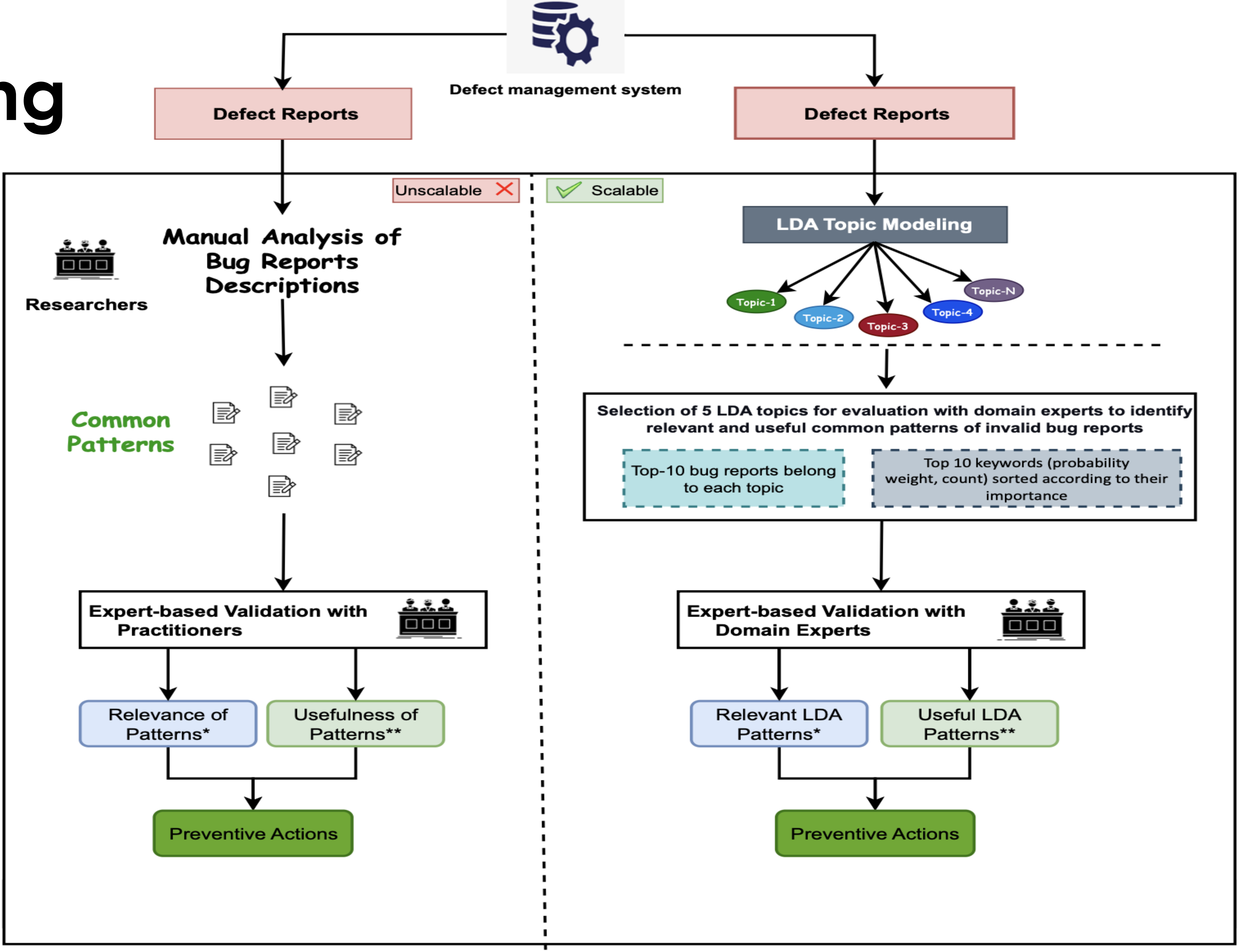




(b) Potential use cases of the tool support for identifying invalid bug reports

34 respondents

# Understanding and insights

From having to review
3500 DRs to 50 DRs



# Improvements suggested based on the analysis

- Features were identified where unclear documentation was leading to a number of defect reports.
- Additional review and onboarding for new and even experienced employees who are new to the product
- The use of the classification early in the process

# Some take aways

Reliable tool support in classification

- Helps to improve productivity
- Provides a cost-effective (several solutions like AutoML are reducing the steep learning curves)
- Help to generate context-aware improvements backed by data-driven insights.
- Works on other artifacts – pull requests,  reviewer feedback

# References

1. M. Laiq, N. bin Ali, J. Börstler, and E. Engström, "A data-driven approach for understanding invalid bug reports: An industrial case study," *Information and Software Technology*, vol. 164, p. 107305, 2023. Available: https://doi.org/10.1016/j.infsof.2023.107305.
2. U. Iftikhar, J. Börstler, and N. bin Ali, "On potential improvements in the analysis of the evolution of themes in code review comments," in the proceedings of the *49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2023, pp. 340-348. DOI: 10.1109/SEAA60479.2023.00059.
3. M. Laiq, N. bin Ali, J. Börstler, and E. Engström, "Industrial adoption of machine learning techniques for early identification of invalid bug reports," *Empirical Software Engineering*, vol. 29, no. 1, p. 130, 2024. Available: https://doi.org/10.1007/s10664-024-10502-3.

# Thanks!

Nauman bin Ali
nal@bth.se