CSE P 590

Building Data Analysis Pipelines

Fall 2024

Advanced statistical modeling

Today

- Homework 1: big picture
    - A first end-to-end data analysis
    - Domain and data set
    - Modeling and statistical methods
- Live demo: Data modeling
- Homework 1: brainstorming

Homework 1: big picture

What is Defects4J?

What is APR?

What is the data set?

What is Defects4J?

Database of **E**xisting **F**aults to **E**nable **C**ontrolled **T**esting **S**tudies **For** **J**ava programs

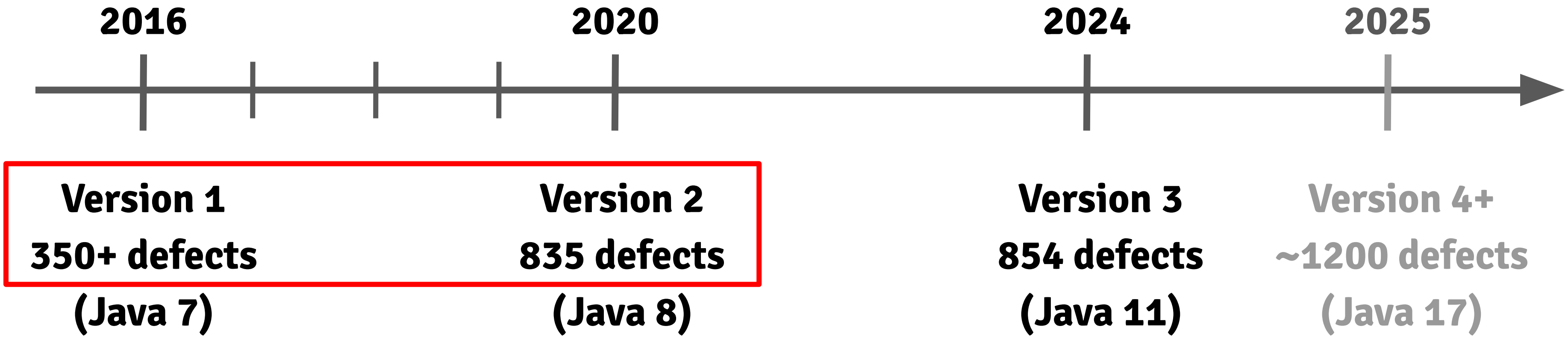
1. **Database 854 defects (17 software systems)**
   - Linked to issues in an issue tracker
   - Reproducible with known triggering test(s)
   - Isolated defects (excl. irrelevant changes)

Suitable for benchmarking testing/debugging approaches.

2. **Supporting infrastructure**
   - Uniform interface to checkout, compile, and analyze defects
   - Support for large-scale experimentation
   - Defect-mining infrastructure plus guidelines and validation

Defects4J over time

Defects4J -- version 3.0.0



Version 1
350+ defects
(Java 7)

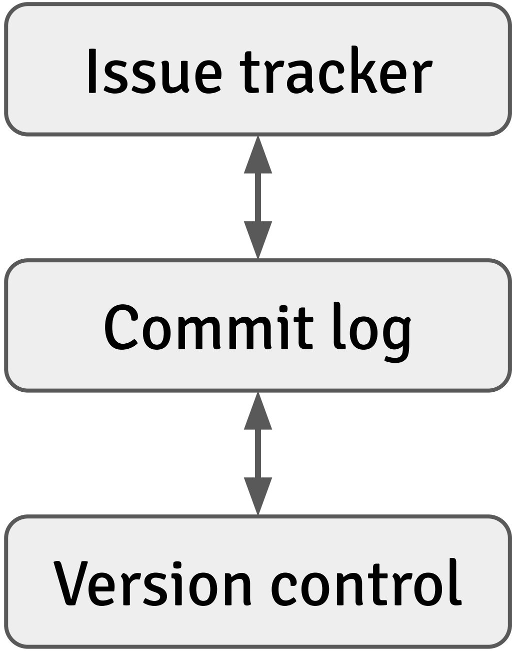
Version 2
835 defects
(Java 8)

Version 3
854 defects
(Java 11)

Version 4+
~1200 defects
(Java 17)

Key focus of HW1: Differences between these versions.

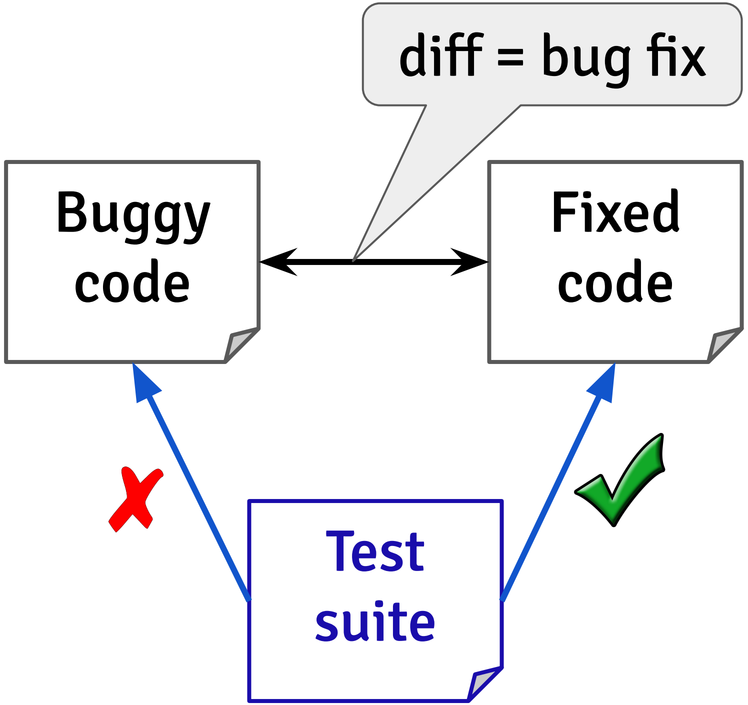
Building Defects4J: how hard can it be?



Issue tracker

Commit log

Version control



diff = bug fix

Buggy code

Fixed code

Test suite

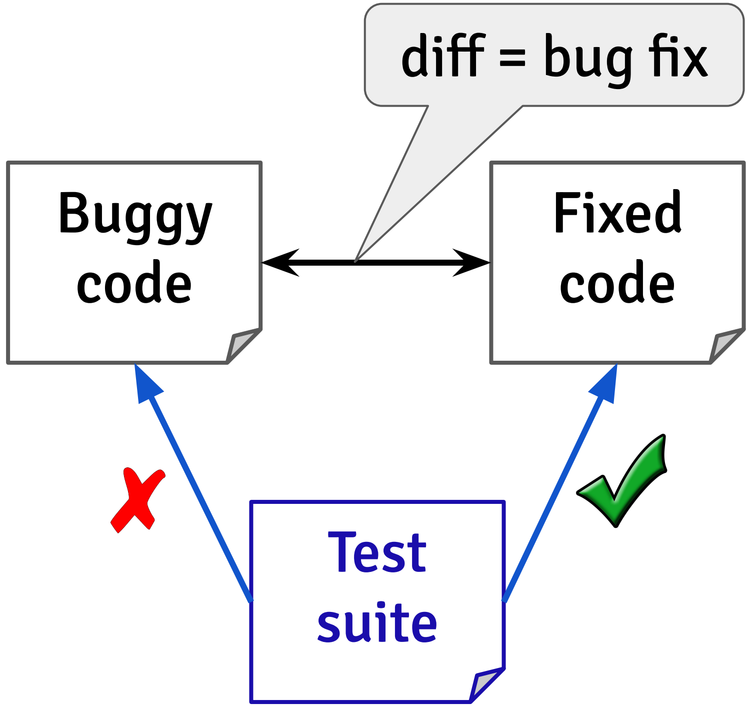
✗

✓

Building Defects4J: how hard can it be?

Real-world programs

- Complex build systems
- Build dependencies
- Broken and flaky tests
- Non-atomic commits



Automated defect **mining is easy**, but **curation is hard!**

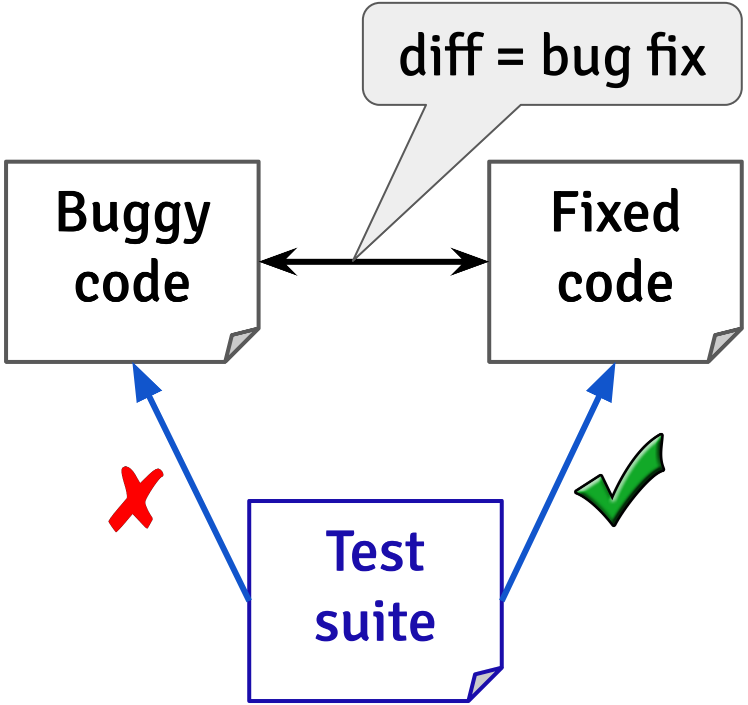
Building Defects4J: benchmark curation

Curation

- **Defect isolation**: separate bug fix from features/refactorings
- **Clean test suite**: remove broken and flaky tests

Usability and experimental control

- **Improve precision** of bug (fix) location and complexity
- **Reduce false-positives** (triggering tests)



Benchmark curation: design considerations

Internal validity

Experimental control

External validity

Realism

← **Benchmarks** **Real deployment** →

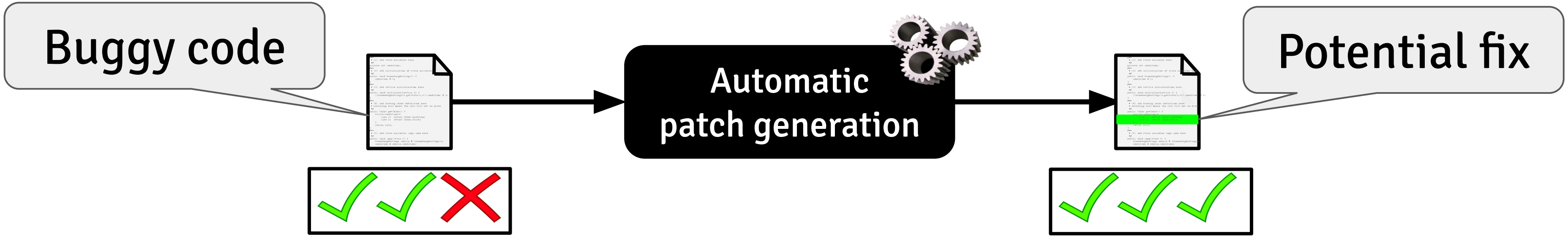
What is Defects4J?

What is APR?

What is the data set?

APR: Automated Program Repair

Goal: patch software bugs automatically



Generate-and-validate Approaches:

- Fault localization
- Mutation + fitness evaluation
- Patch validation (test executions)

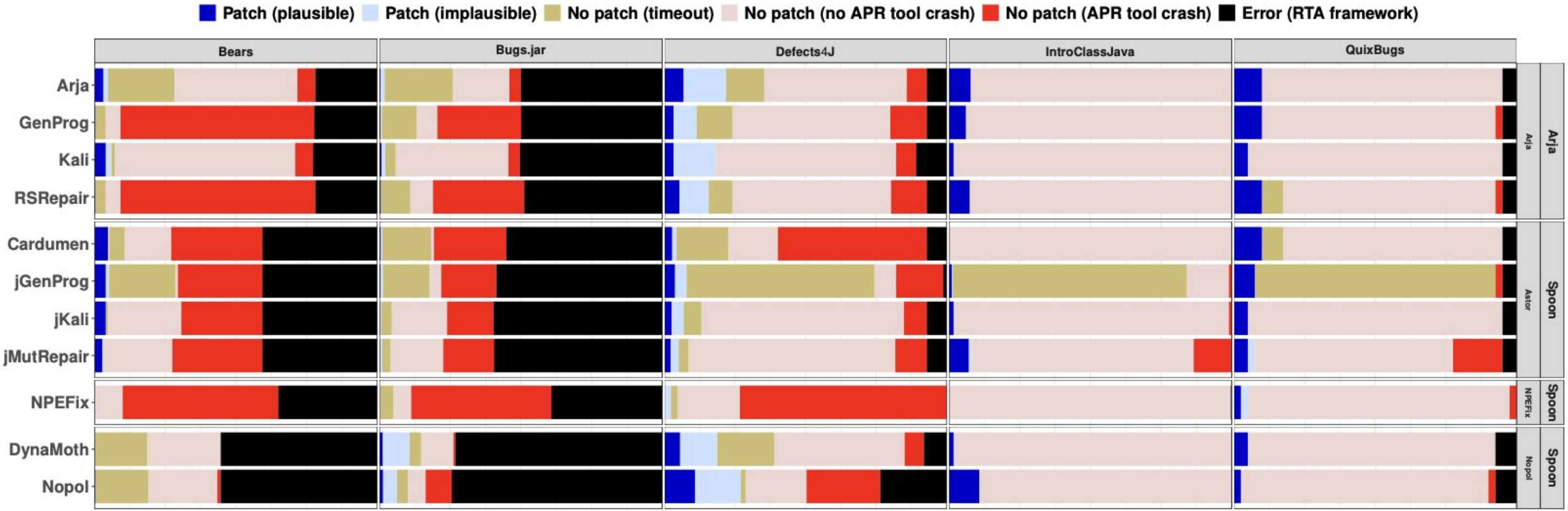
Many different approaches and evaluations (10+ years of research)

What is Defects4J?

What is APR?

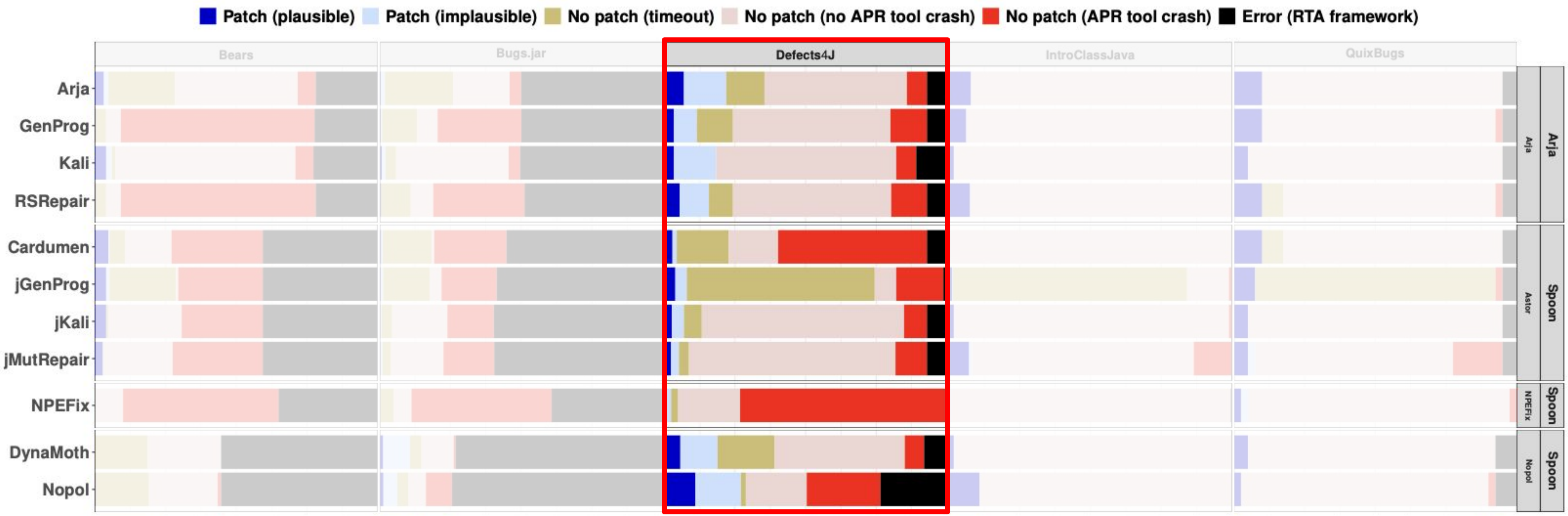
What is the data set?

What do APR evaluations look like?



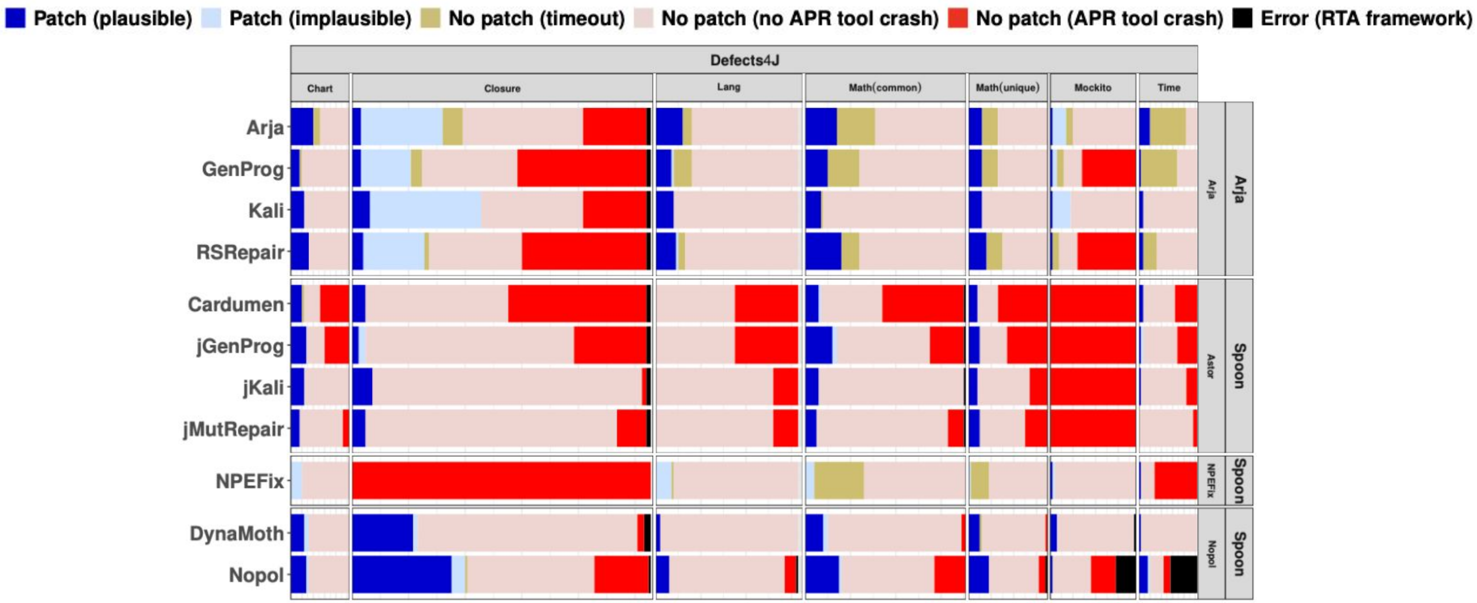
Data: Mapping of *Tool x Bug* to *Outcome*

What do APR evaluations look like?



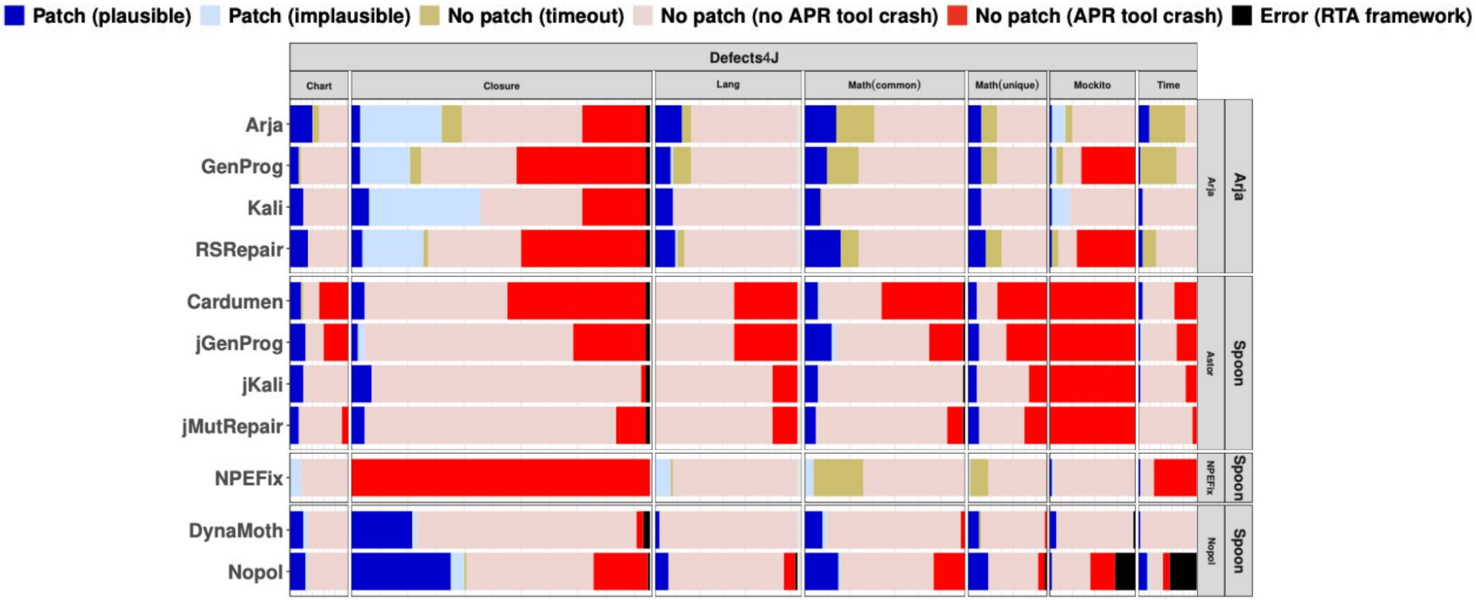
Let’s drill deeper: benchmark composition

What do APR evaluations look like?



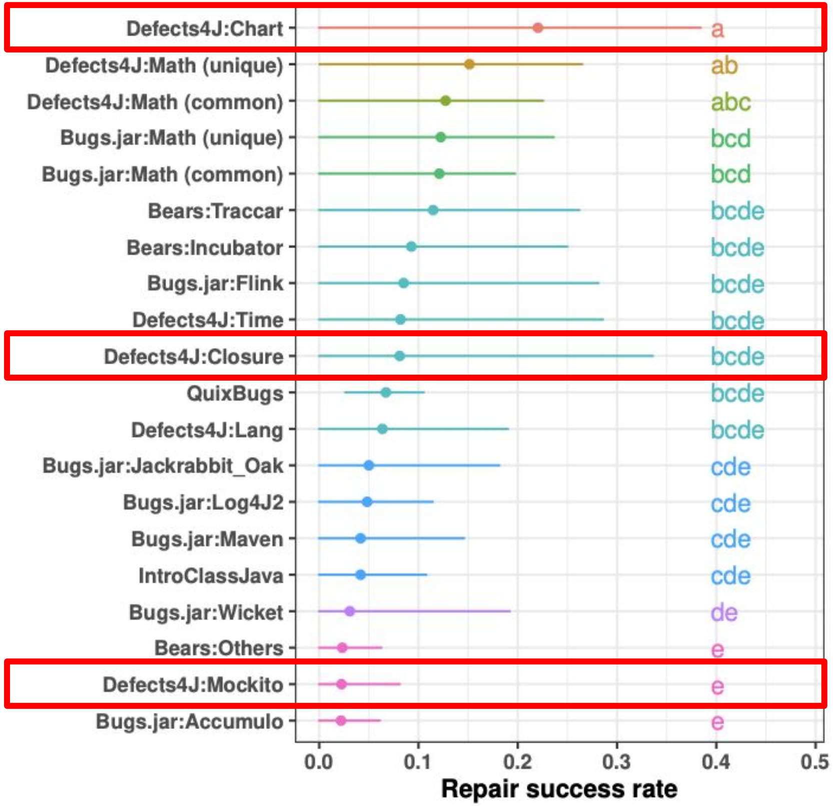
Data: *Mapping of Tool x Bug* to *Outcome – grouped by Project*

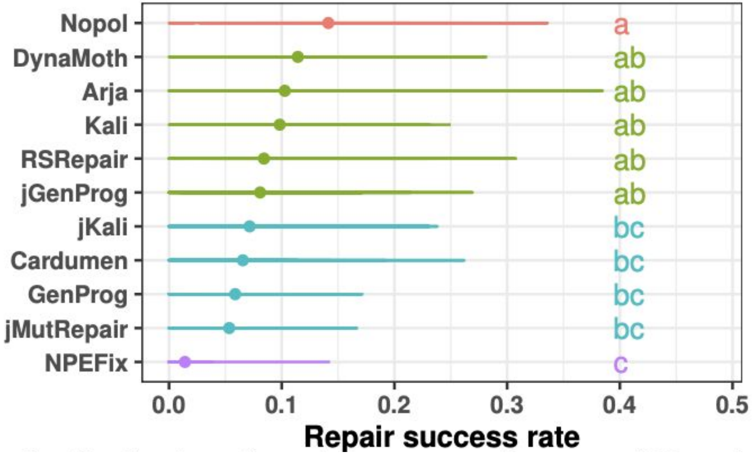
What do APR evaluations look like?



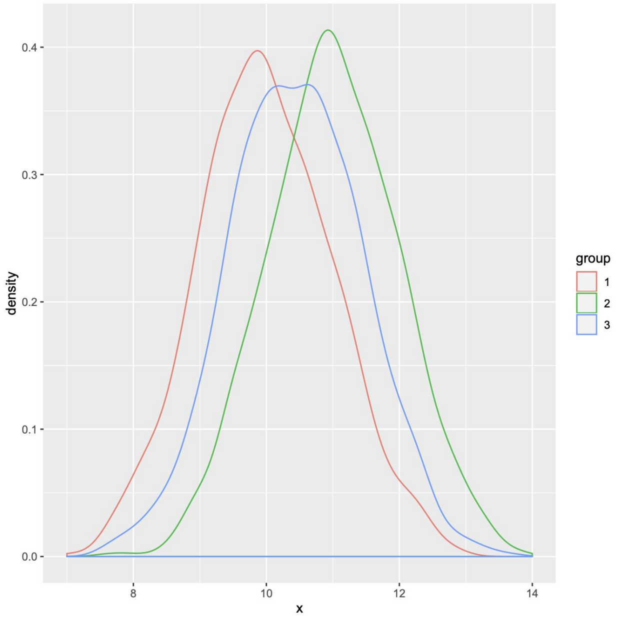
How would you (statistically) analyze the data?

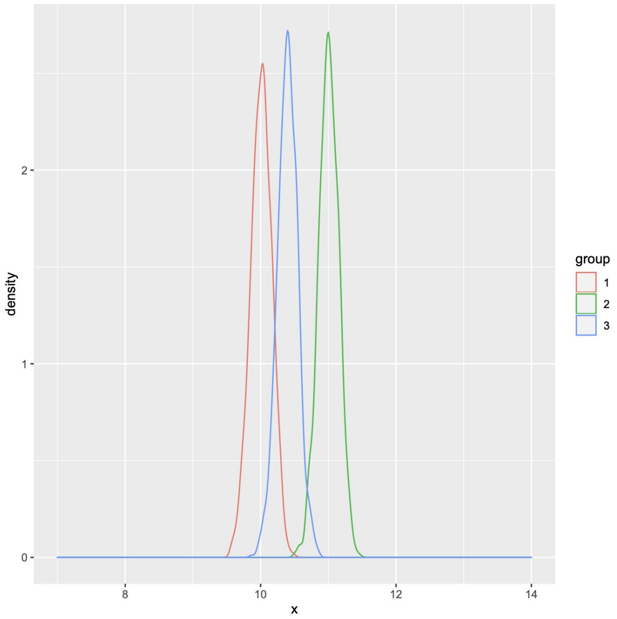
APR evaluation: one option (ANOVA and Tukey HSD)





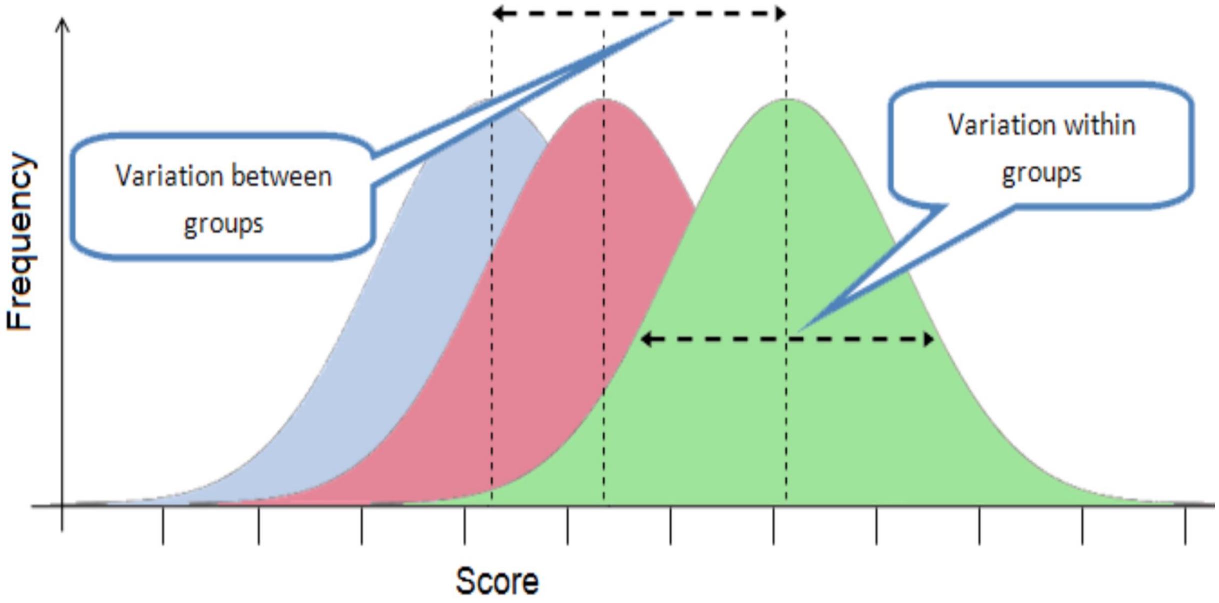
ANOVA: Motivation





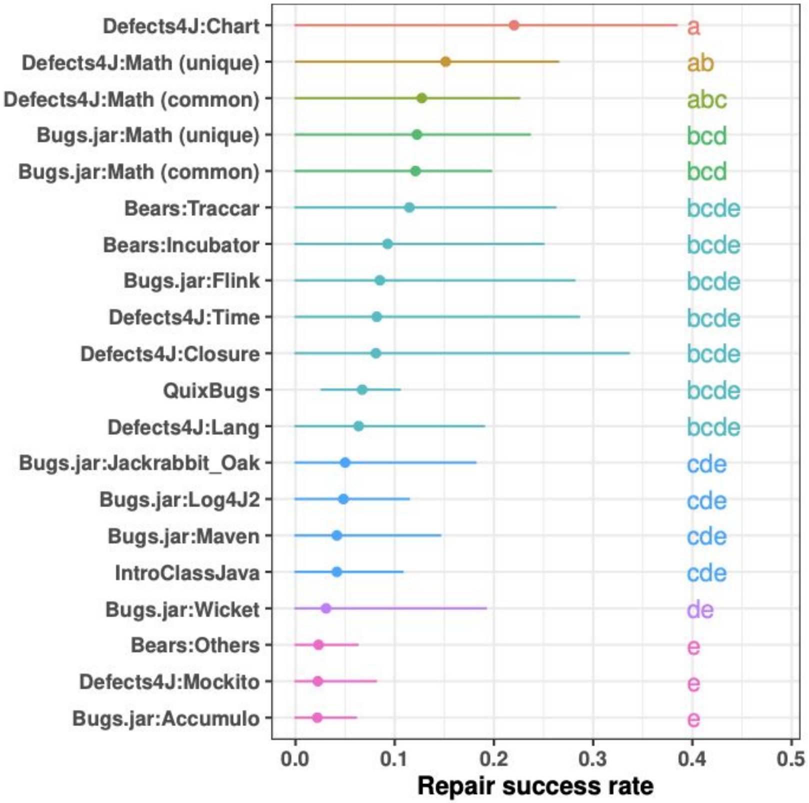
Are the group means significantly different?
(Do all 3 group samples come from the same population?)

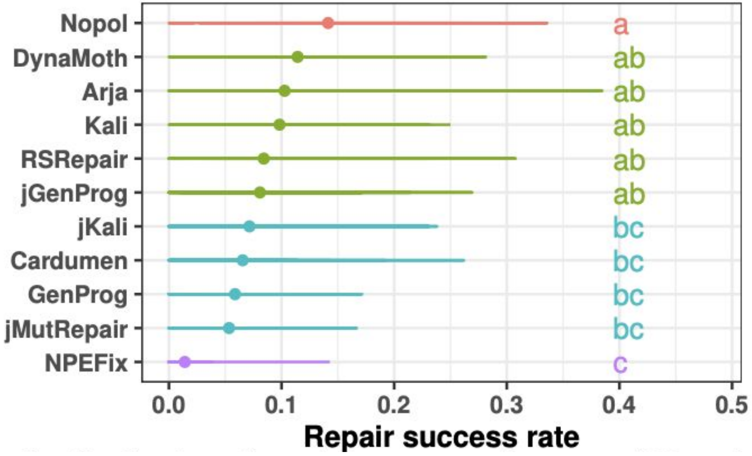
ANOVA: ANalysis Of VAriance



ANOVA: Is there a significant difference between some groups?
Post-hoc: What groups are significantly different from one another?

ANOVA and Tukey HSD



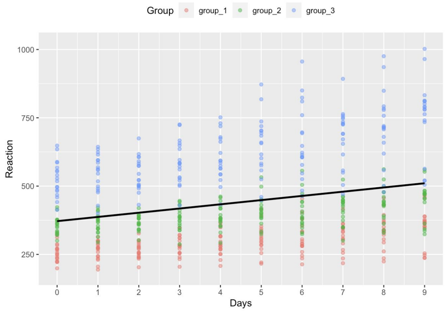


APR evaluation: an alternative (LM)

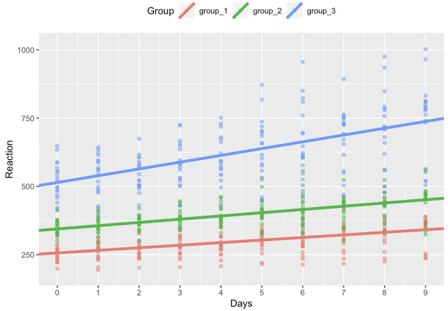
(Generalized) Linear Model

- Split the data set by groups.
- Model outcome as a function of variables of interest.

Entire data set



Split by groups



LM: Linear regression models

Assumptions

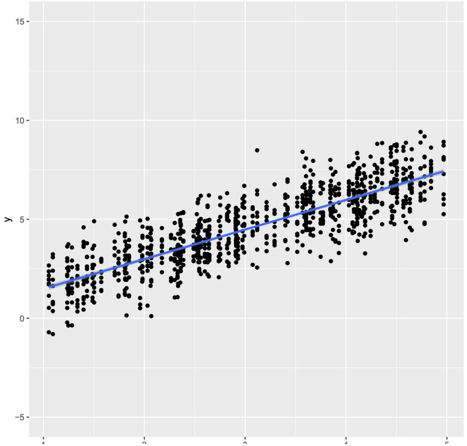
- Linearity
- Normality (residuals)
- Homoscedasticity (residuals)
- Independence (observations)
- Little to no multicollinearity (for inference).

LM: Linear regression models

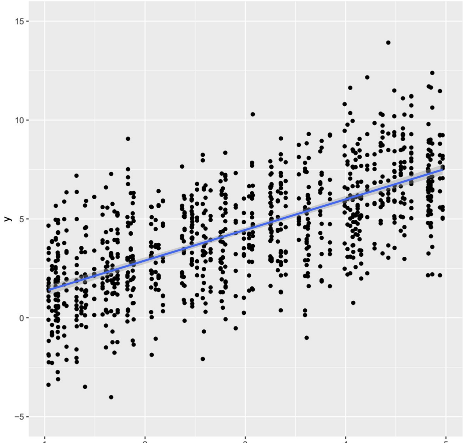
Interpretation of results

- Model fit: goodness of fit (R^2)
- Inference: significance of coefficients

R^2=0.7



R^2=0.4

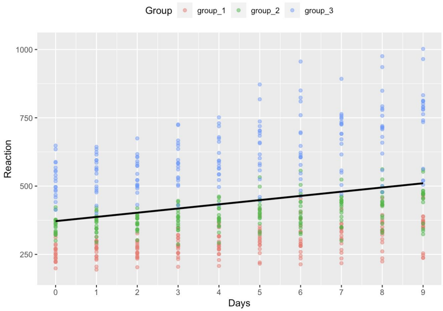


Which fitted linear model is “better”?

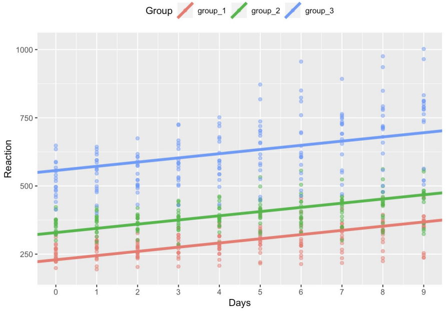
APR evaluation: another alternative (GLMM)

(Generalized) Linear Mixed Model

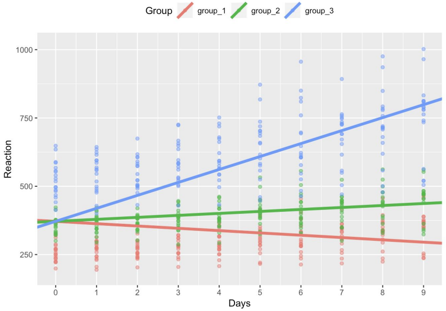
- Model fixed and random effects.
- Allow intercepts and/or slopes to vary.



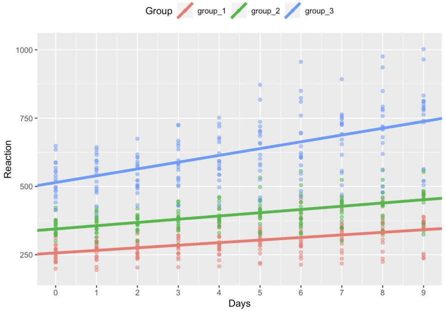
Fixed intercept, **fixed** slope



Random intercept, **fixed** slope



Fixed intercept, **Random** slope



Random intercept, **Random** slope

https://glennwilliams.me/r4psych/mixed-effects-models.html

Data modeling: live demo

Homework 1: brainstorming

HW1: An end-to-end data analysis

Goal

- Raise questions about terminology and concepts.
- Raise questions about the data set or data generation process.
- Raise questions about modeling challenges.

Set up

- Small groups (~6 students)
- Discuss and document open questions: https://tinyurl.com/abkwan7n