The credit for creating these slides belongs to Fall 2014 CS 521/621 students. Student names have been removed per FERPA regulations.
Empirical Software Engineering

- Empirical Software Engineering- field of research that emphasizes empirical methodology
  - Experiments, case studies, statistical analyses, and surveys
- In order to create useful information, we need data to reuse and share
Bug-fix datasets and links

- Detailed data on bugs is **crucial** in empirical software engineering
  - Data is collected in **bug-fix datasets**
    - Contain **links** to the commit bug report info and the bug location in the source code
- This data is used to create software defect prediction models
  - Gives us better understanding to what affects software quality
Biases in bug-fix datasets

● Concerns with bias bug-fix datasets
  ○ Sample bias lead to underperforming prediction models
  ○ Validity of studies based on bias bug-fix datasets are suspect
Size as a factor

- But a factor not examined carefully is size
  - Bias bug-fix datasets could be large or small
  - Smaller datasets provide less reliable basis for estimating models → underperforming prediction models

- What affects performance more? Bias or size?
  - Can a large dataset be useful even if it is bias?
Related Work and Theory

Past research determined defect-fixing changes by linking commit comments with keywords (e.g. “fix”, “bug”) to code change

- Due to *human error*, might lead to faults within automated process and resulting in missing links
  - **Missing links** - Defect fixing changes that automated processes fail to recover
    - Bias and pollution are causes of missing links
Key Principle Idea

What affects performance more, bias, or size?
Contributions

- Examined effects of bias in defect prediction models from datasets known to be high quality (dataset free of bias)

- Relative effects of different types of bias on defect prediction using meta-models

- Used meta-models to study relative effect of bias and size

- Effects of pollution
Research Questions

Research Question 1: Do varying sources of bias have varying impact on prediction performance?

Research Question 2: Considering Bias, Pollution, and Size, which aspect of missing links affects prediction models the most?
Experimental Methodology

The given dataset is highly linked.
Experimental Methodology

The datasets are selectively sampled by bias, size, and pollution.
Experimental Methodology
Biases and Pollution

**BIAS**

**Experience** - Defect-fixer experience

**Severity** - Severity of the fixed defects

**Proximity** - Proximity of defect-fixing commit to the release deadline

**Latency** - Amount of time it took to fix a defect

**Cardinality** - Size of commits

Ex: only experienced developers are annotating their changes. Automated tool will only identify defect-fixing changes done by experienced developers.

**Pollution**

Bug-fix datasets containing defective files, indeed, that are labeled as defect-free.
Experimental Methodology

The datasets are selectively sampled by bias, size, and pollution.
Experimental Methodology

Create a cloud of sub-datasets through the samplings that had been controlled by size, bias, and pollution.
Experimental Methodology

Each sub-datasets are used to estimate prediction models.
Experimental Methodology

Each prediction model gets subjected to multiple regression meta-modeling.
Experimental Methodology

By discerning meta-analysis, we get useful info on the effects of size, bias, and pollution.
R1: Do varying sources of bias have varying impact on prediction performance?

Figure 2: Median of Performance for different bias sources. \( E \) for EXPERIENCE; \( C \) for CARDINALITY; \( S \) for SEVERITY; \( L \) for LATENCY; \( P \) for PROXIMITY.
R1: Do varying sources of bias have varying impact on prediction performance?

Experience
Cardinality
Severity
Latency
Proximity
Result

R1: Do varying sources of bias have varying impact on prediction performance?

Figure 2: Median of Performance for different bias sources. $E$ for EXPERIENCE; $C$ for CARDINALITY; $S$ for SEVERITY; $L$ for LATENCY; $P$ for PROXIMITY
R1: Do varying sources of bias have varying impact on prediction performance?

Result

No change!

Figure 2: Median of Performance for different bias sources. E for Experience; C for Cardinality; S for Severity; L for Latency; P for Proximity
R1: Do varying sources of bias have varying impact on prediction performance?

examined the stability of prediction performance using the variance of the performance measures per release.

- Indicates that variance is sensitive when using difference biases.

But...
R2: Considering Bias, Pollution, and Size, which aspect of missing links affects prediction models the most?

When considering AUC, SIZE is much more important than BIAS polarity. Focusing effort on collecting more samples may mitigate much of the impact of BIAS polarity.

For the IR performance measures AUC and F50, the effect of SIZE strongly dominates that of BIAS. For the AUCEC performance measure, SIZE has as much of an influence as BIAS polarity.

For F50 performance, POLLUTION plays a more damaging role than BIAS polarity. SIZE still has the most impact.
When considering AUC, SIZE is much more important than BIAS polarity. Focusing effort on collecting more samples may mitigate much of the impact of BIAS polarity.

L: bias polarity to low values
M: bias polarity to higher values
P: stands for pollution
S: stands for size
R2: Considering Bias, Pollution, and Size, which aspect of missing links affects prediction models the most?

For the IR performance measures AUC and F$_{50}$, the effect of SIZE strongly dominates that of BIAS. For the AUCEC performance measure, SIZE has as much of an influence as BIAS polarity.

For F$_{50}$ performance, POLLUTION plays a more damaging role than BIAS polarity. SIZE still has the most impact.
R2: Considering Bias, Pollution, and Size, which aspect of missing links affects prediction models the most?
Discussion Questions

1. Why might smaller sized datasets lead to having underperforming prediction models?
2. Why might an automated process fail to recover defect-fixing changes? (aka missing links)
3. What might be an example of an experience bias?
Discussion Questions

4. What other sources of biases can you think of other than the ones we discussed?
5. If they had used lower quality datasets, what might happen.
6. What are the advantages of having a larger sized dataset as compared to having an unbiased dataset.
References

http://dl.acm.org/citation.cfm?id=2491411.2491418

Thank You.