Sample Size
vs
Bias in Defect Prediction

Research Questions

Do different sources of bias have different impacts on prediction performance?

Considering bias, pollution, and size, which aspect of missing links affects prediction models the most?

Experimental Group

- Select programs from the bug tracking software JIRA
- Very high linking rates

Project	Description	Releases	Avg Files	Avg SLOC	Link Rate
CXF	Services Framework	July 2007–April 2012, 6 releases	4038.33	358846.67	0.77
Camel	Enterprise Integration Framework	January 2008–March 2012, 8 releases	4600.38	241668.12	0.84
Derby	Relational Database	February 2006–October 2011, 7 releases	2497.29	530633.00	0.85
Felix	OSGi R4 Implementation	August 2007–November 2011, 9 releases	2740.56	249886.22	0.85
HBase	Distributed Scalable Data Store	Jun 2007–May 2012, 8 releases	934.75	187953.38	0.85
HadoopC	Common libraries for Hadoop	Jun 2007–September 2009, 6 releases	1047.17	142257.33	0.79
Hive	Data Warehouse System for Hadoop	October 2008–May 2012, 7 releases	966.29	152079.86	0.75
Lucene	Text Search Engine Library	October 2005–November 2009, 7 releases	990.86	122527.00	0.85
OpenEJB	Enterprise Java Beans	August 2007–April 2012, 7 releases	2895.43	225018.43	0.86
OpenJPA	Java Persistence Framework	January 2007–February 2012, 8 releases	3181.50	321033.50	0.92
Qpid	Enterprise Messaging system	November 2008–May 2012, 7 releases	1724.00	198311.86	0.73
Wicket	Web Application Framework	November 2008–May 2012, 5 releases	2295.20	152565.40	0.77

Bias Types Evaluated

Experience- how experienced the defect-fixer was(percent of all commit made by that fixer)

Severity- The importance of this bug

Proximity- Number of days between the day of the bug-fix and next release (how close to a deadline was this bug fixed)

Latency- Number of days between the day of reporting and fixing the bug. (how long it took to fix the bug)

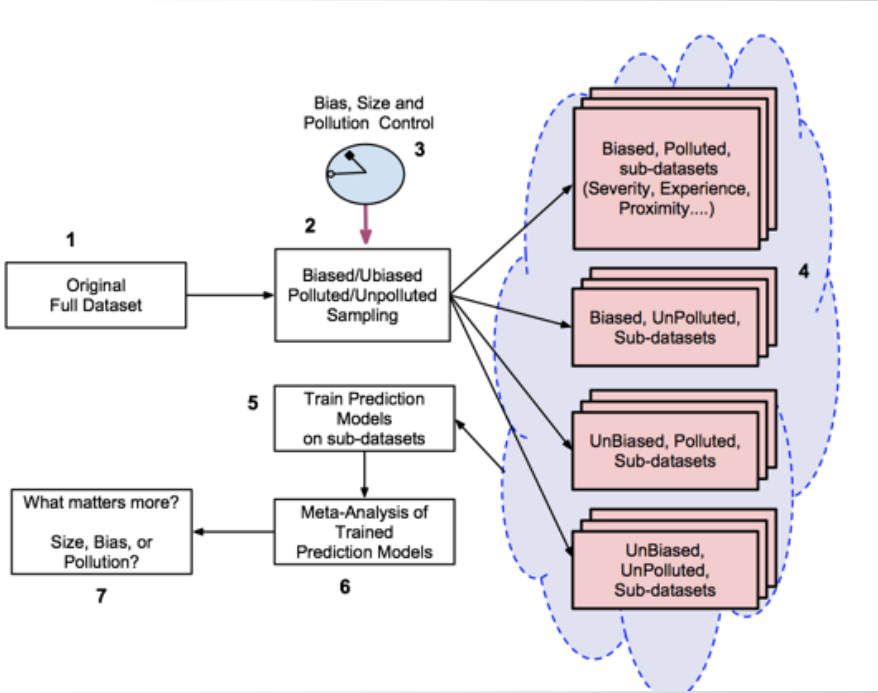
Cardinality- Number of files in the bug-fixing commit.

Predictor Metrics

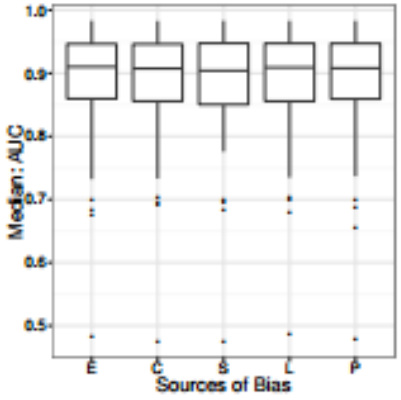
Short Name	Description
COMM	Commit Count
ADEV	Active Dev Count
DDEV	Distinct Dev Count
ADD	Normalized Lines Added
DEL	Normalized Lines Deleted
OWN	Owner's Contributed Lines
MINOR	Minor Contributor Count
NADEV	Neighbor's Active Dev Count
NDDEV	Neighbor's Distinct Dev Count
NCOMM	Neighbor's Commit Count
OEXP	Owner's Experience
EXP	All Committer's Experience

The Experimental Procedure

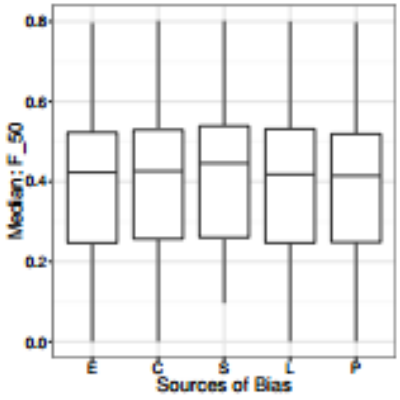
- For each type of bias tested, sub-datasets are sampled from the original dataset with that bias
- Each biased dataset is used to train a prediction model, and the models are evaluated
- Conclusions drawn from the results of these prediction models



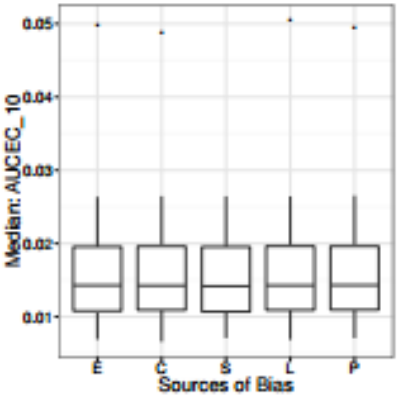
Effects of Bias



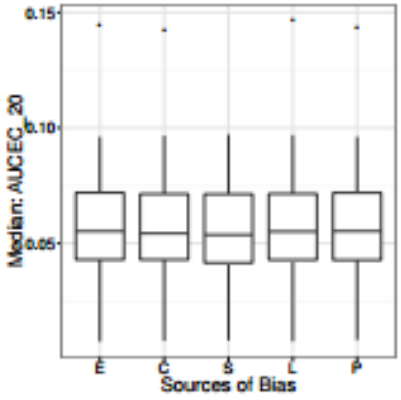
(a) Median AUC



(b) Median F_{50}

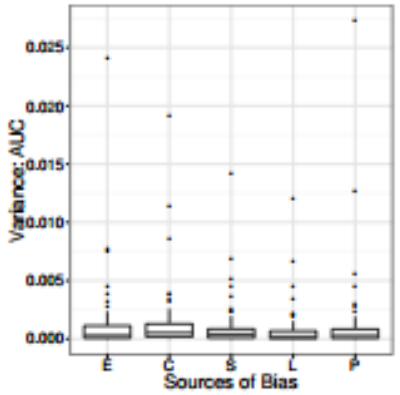


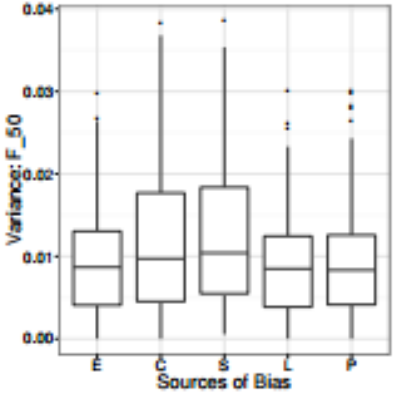
(c) Median AUCEC_{10}



(d) Median AUCEC_{20}

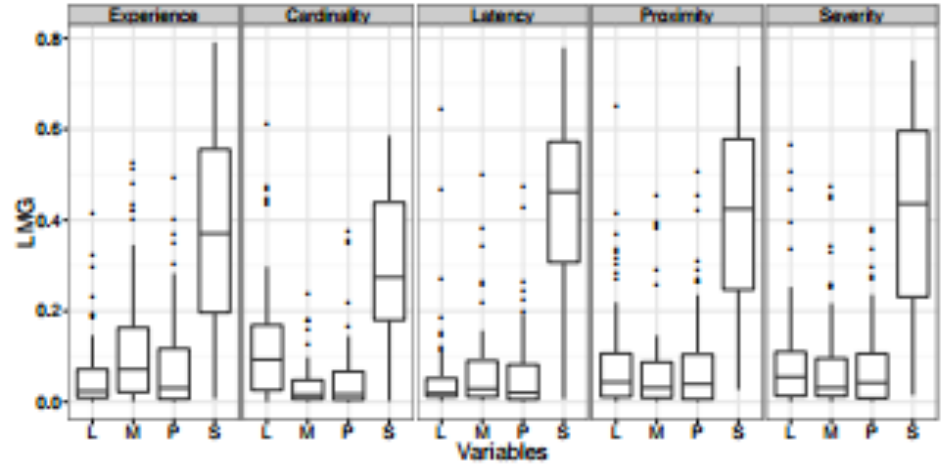
Figure 2: *Median of Performance for different bias sources. E for* EXPERIENCE*; C for* CARDINALITY*; S for* SEVERITY*; L for* LATENCY*; P for* PROXIMITY



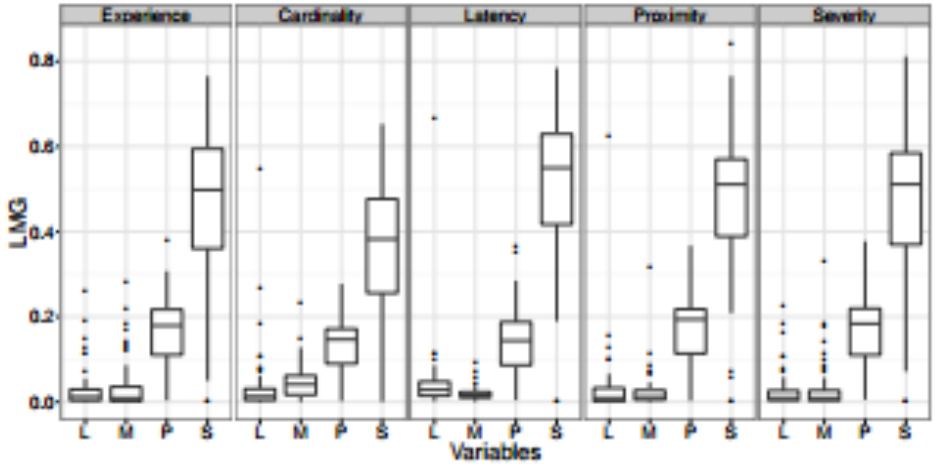


(a) AUC variance (b) F_{50} variance

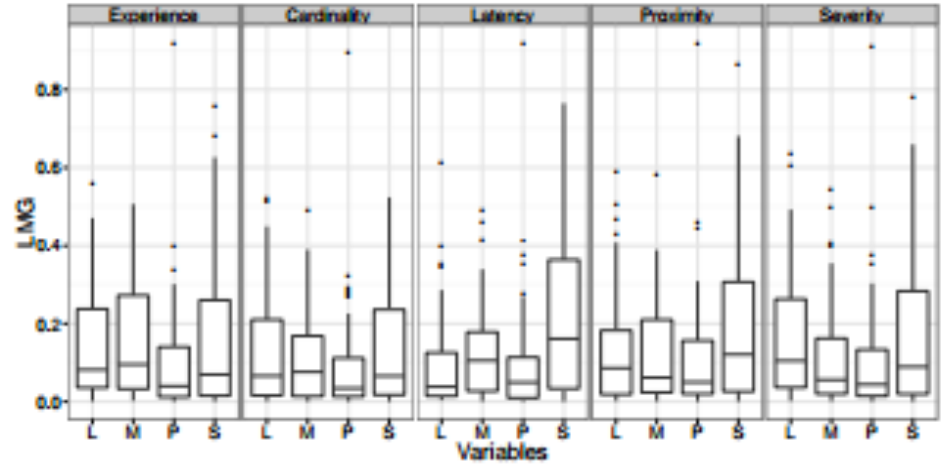
Figure 3: **Variance of Performance for different bias sources. E for Experience; C for Cardinality; S for Severity; L for Latency; P for Proximity.**



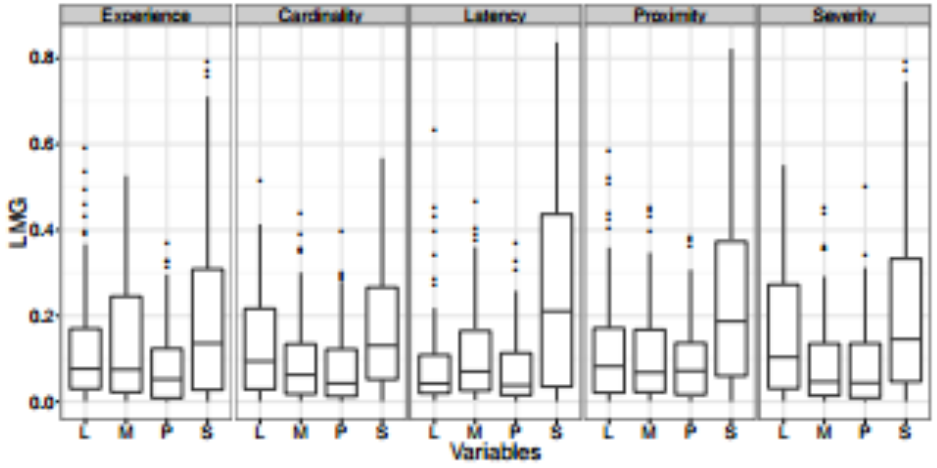
(a) Impact on AUC



(b) Impact on F_{50}



(c) Impact on $AUCEC_{10}$



(d) Impact on $AUCEC_{20}$

Analysis

Different sources of bias have very similar effects on performance; furthermore, the effect of varying rates of bias is also minimal on non-parametric measures of performance, for all sources of bias.

Conclusion

Size is much more important than bias polarity.

Focusing effort on collecting more samples may mitigate much of the impact of bias polarity.

Only for the F50 measure, pollution plays a bigger part than bias

Discussion Questions

1. What are some other sources of bias not identified in this study?

2. What would be the impact of using this study for defect prediction for tools like SemFix?

Discussion Questions

3. How do you know that the “high quality data set” is in fact sufficiently unbiased?

4. Why is it significant that dataset size is more important than bias?

Discussion Questions

5. How would different sources of bias impact prediction performance?

6. Which types of bias would it be more useful to know? What would you predict to be the most influential type of bias?

	Short Name	Description
+	COMM	Commit Count
	ADEV	Active Dev Count
	DDEV	Distinct Dev Count
	ADD	Normalized Lines Added
	DEL	Normalized Lines Deleted
+	OWN	Owner's Contributed Lines
	MINOR	Minor Contributor Count
	NADEV	Neighbor's Active Dev Count
	NDDEV	Neighbor's Distinct Dev Count
	NCOMM	Neighbor's Commit Count
+	OEXP	Owner's Experience
+	EXP	All Committer's Experience