Research at Scale

The Mondego Group @ UC Irvine & ISR

Crista Lopes

Pedro Martins, Assistant Project Scientist

Rohan Achar, Graduate Student

Di Yang, Graduate Student

Vainhav Saini, Graduate Student

Eugenia Gabrielova, Graduate Student

Wen Shen, Graduate Student

Farima Farmahinifarahani, Graduate Student

ISR Research Forum, June 2017

A Couple of Projects

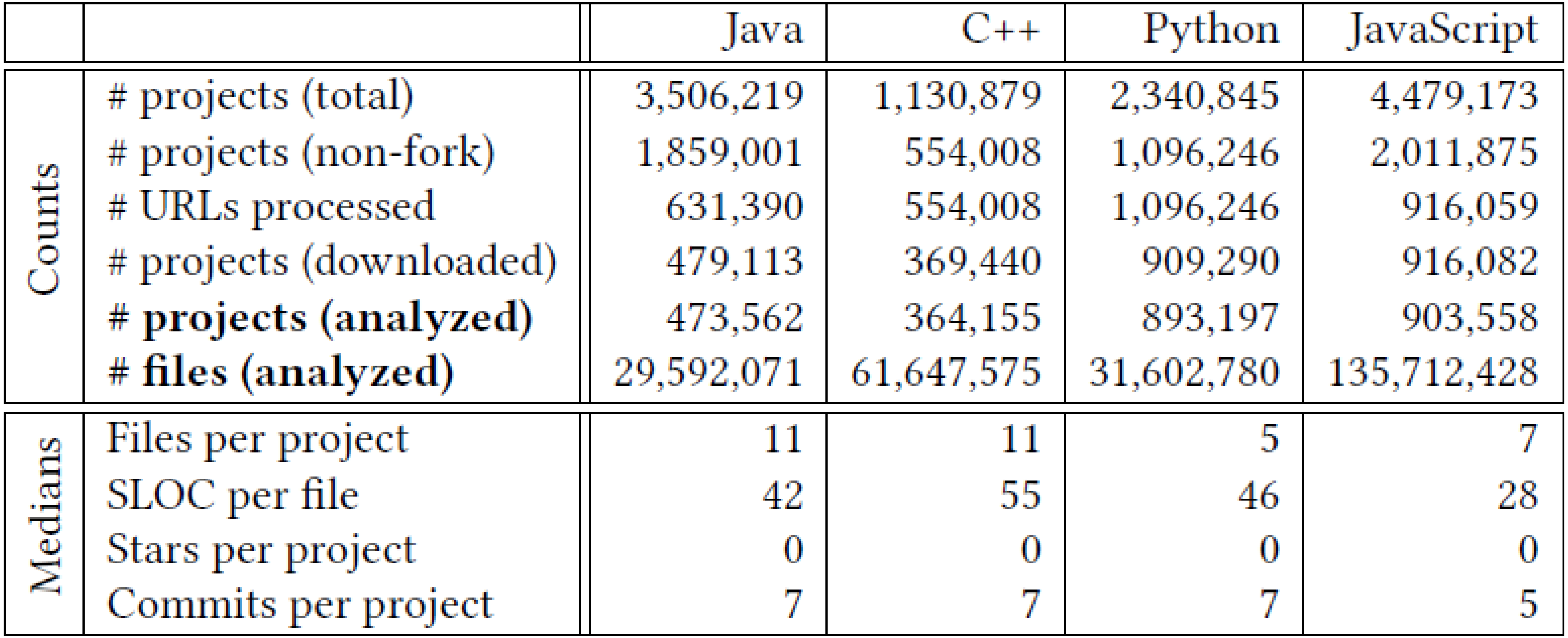
- Code cloning in Java, C++, Python, JavaScript (Collaboration with Jan Vitek, NEU)
- Sourcerer’s Java Build Framework
- [Is there gold in Stack Overflow data?]
  - Ask me offline!

Code Duplication

Understanding Natural Code Duplication

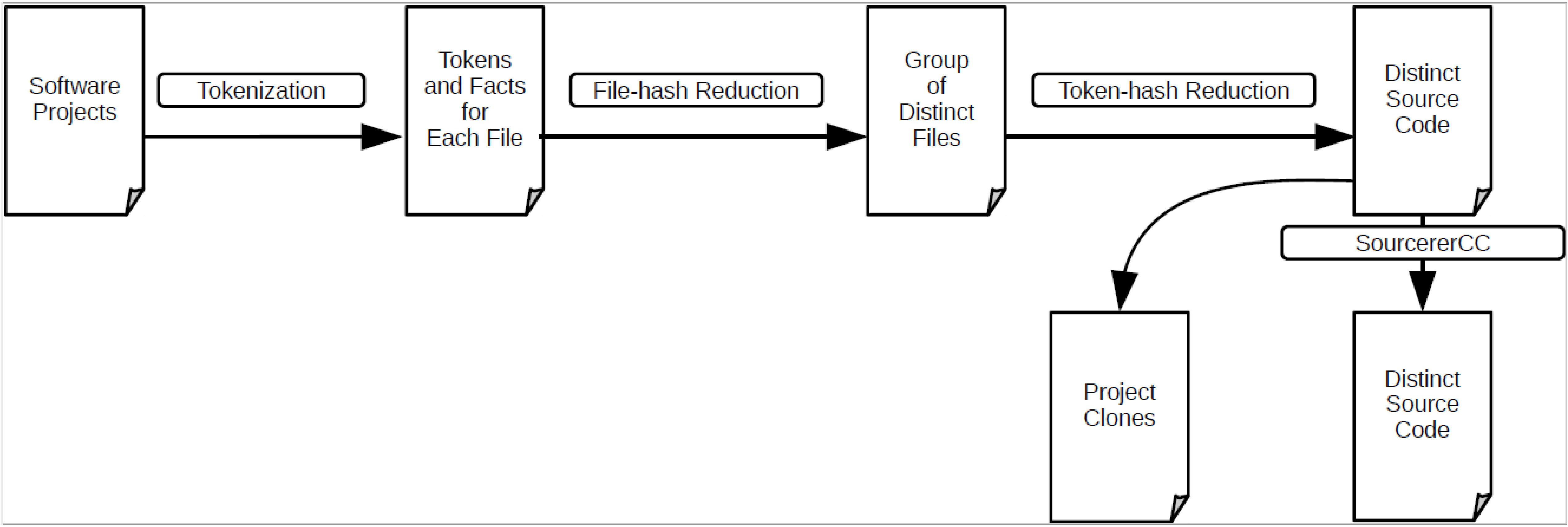
- Main objectives :
  - Measure it
  - Understand **what** is being cloned (qualitative analysis)
  - Understand main differences between different languages
  - Make duplication data available

Corpus

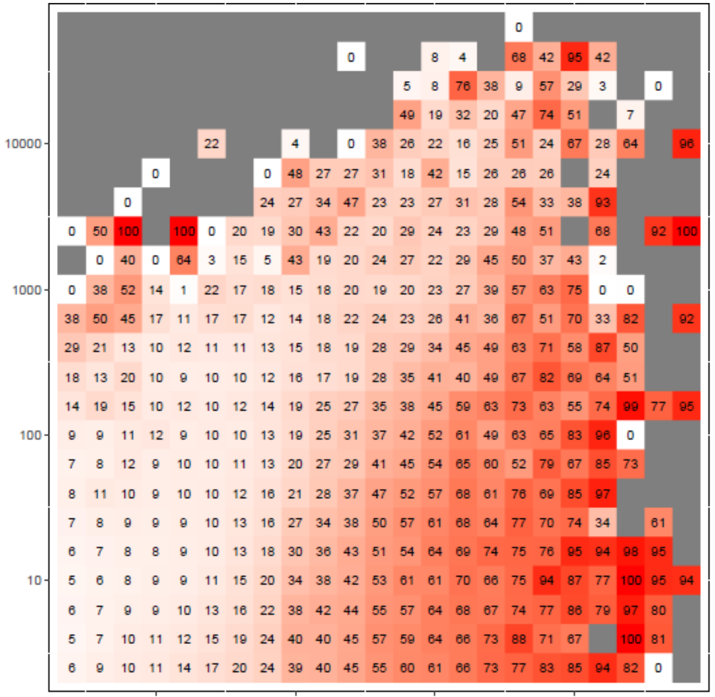


		Java	C++	Python	JavaScript
Counts	# projects (total)	3,506,219	1,130,879	2,340,845	4,479,173
	# projects (non-fork)	1,859,001	554,008	1,096,246	2,011,875
	# URLs processed	631,390	554,008	1,096,246	916,059
	# projects (downloaded)	479,113	369,440	909,290	916,082
	# projects (analyzed)	473,562	364,155	893,197	903,558
	# files (analyzed)	29,592,071	61,647,575	31,602,780	135,712,428
Medians	Files per project	11	11	5	7
	SLOC per file	42	55	46	28
	Stars per project	0	0	0	0
	Commits per project	7	7	7	5

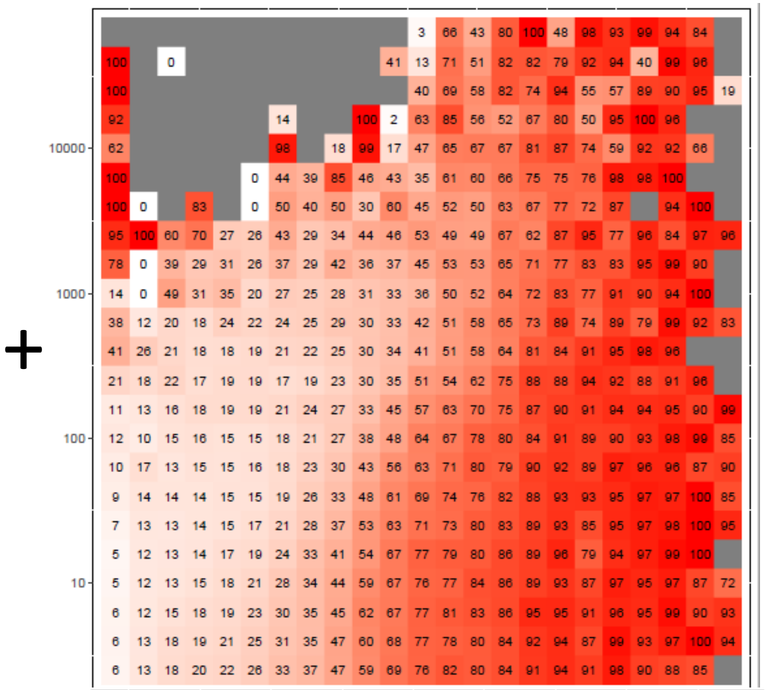
Data Processing Pipeline



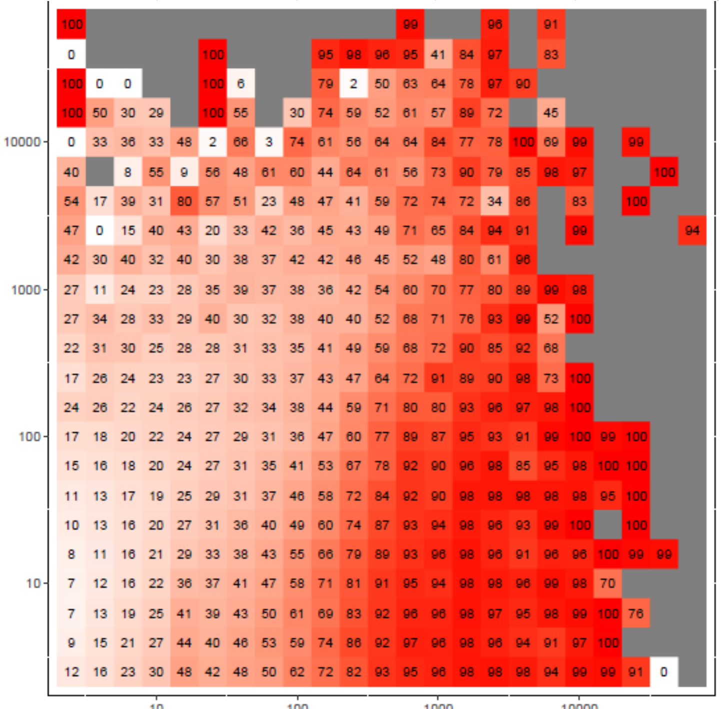
Duplication vs. # files, # commits



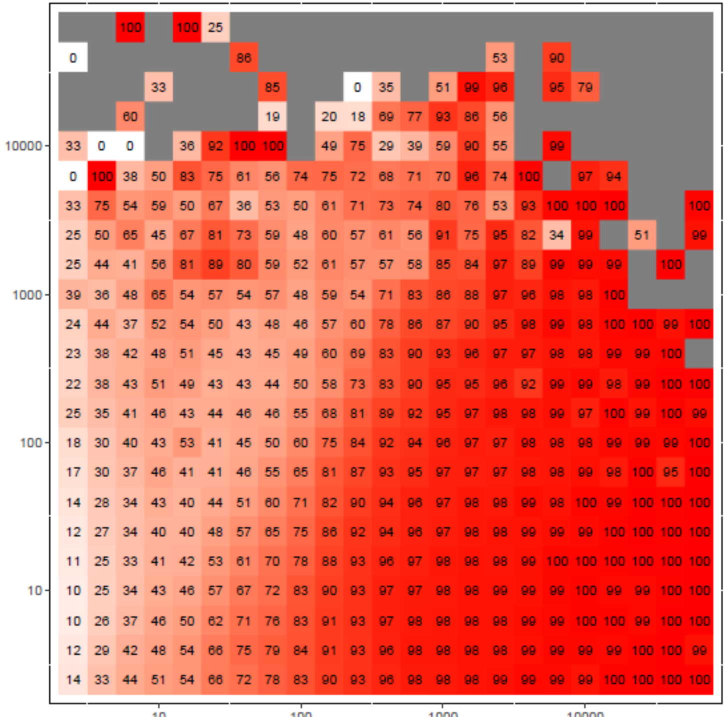
Java



C++

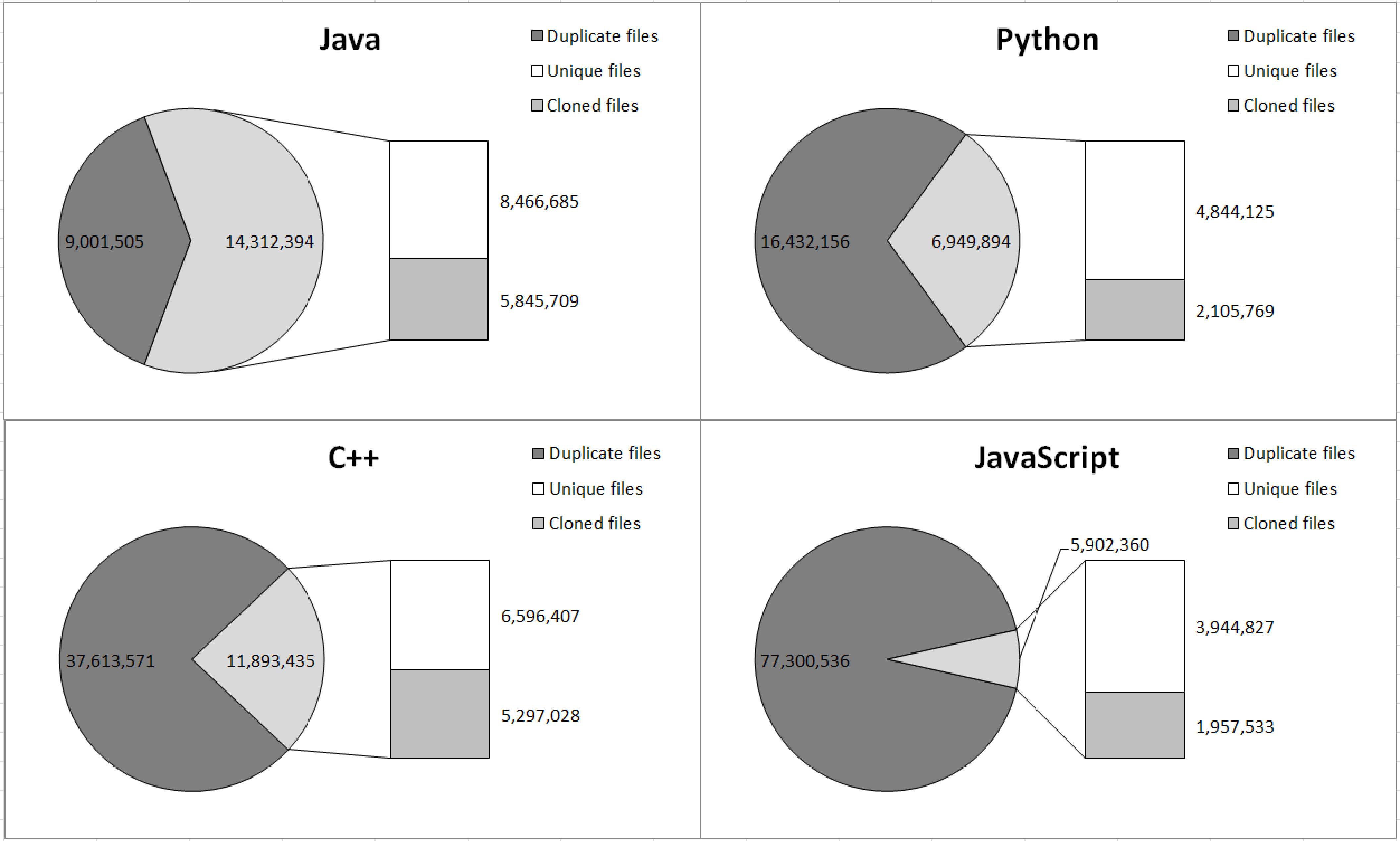


Python

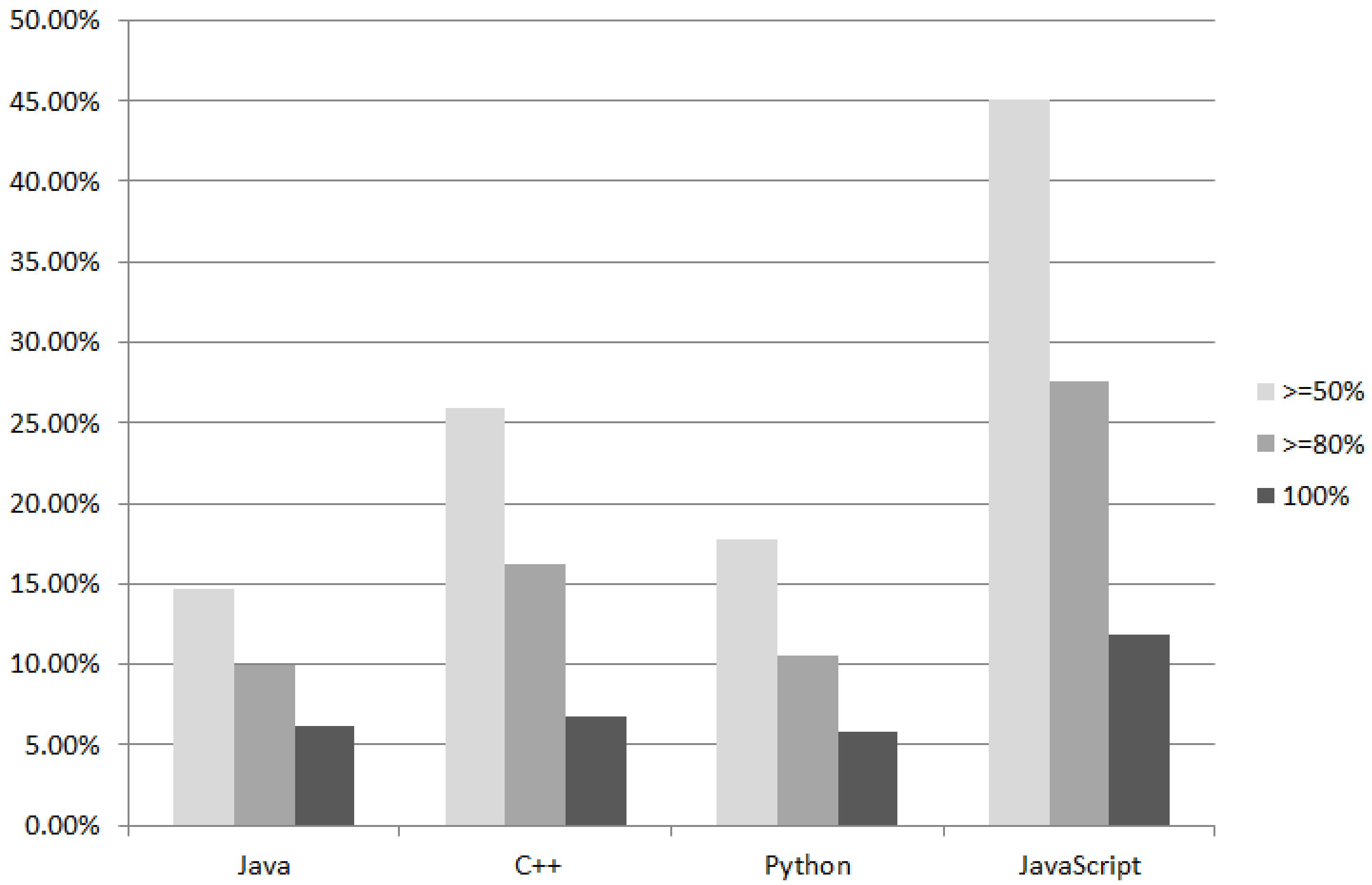


JS

Types and Amounts of Duplication



Project-Level Duplication



Current Work

- DejàVu: a Web service that returns all duplicates of a given file in GitHub
- Performance improvements to clone detection

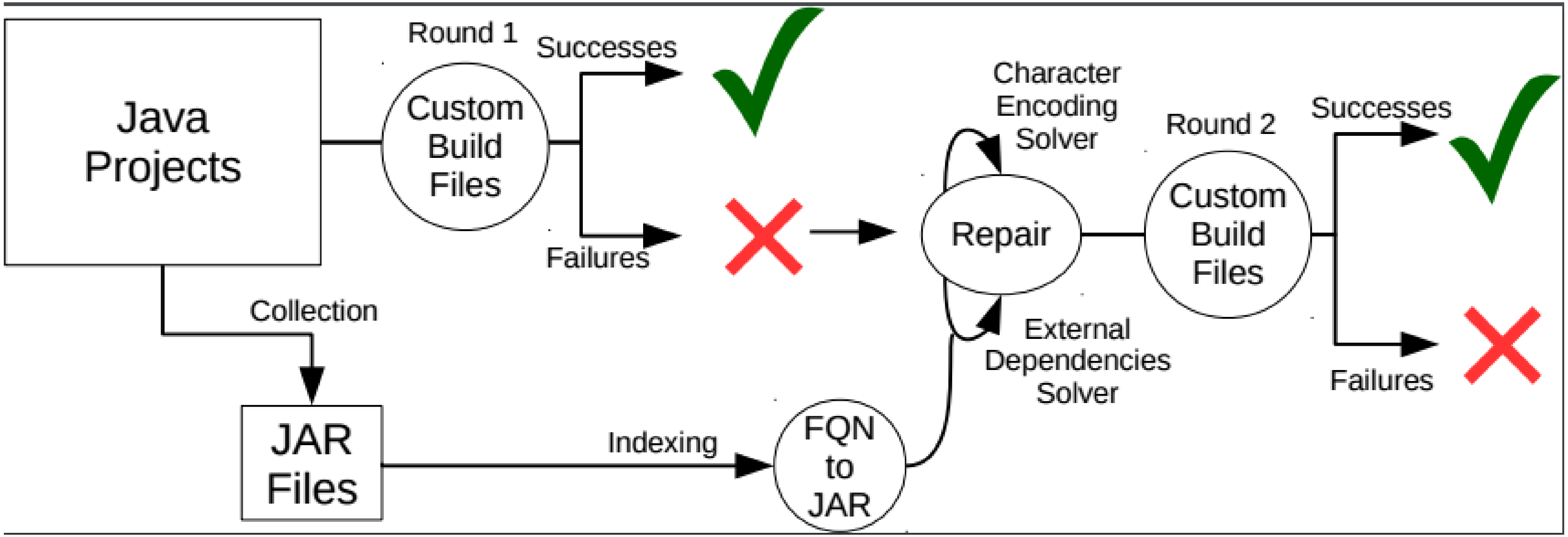
Sourcerer’s Java Build Framework

Goal

- Automatically build ALL of GitHub Java corpus

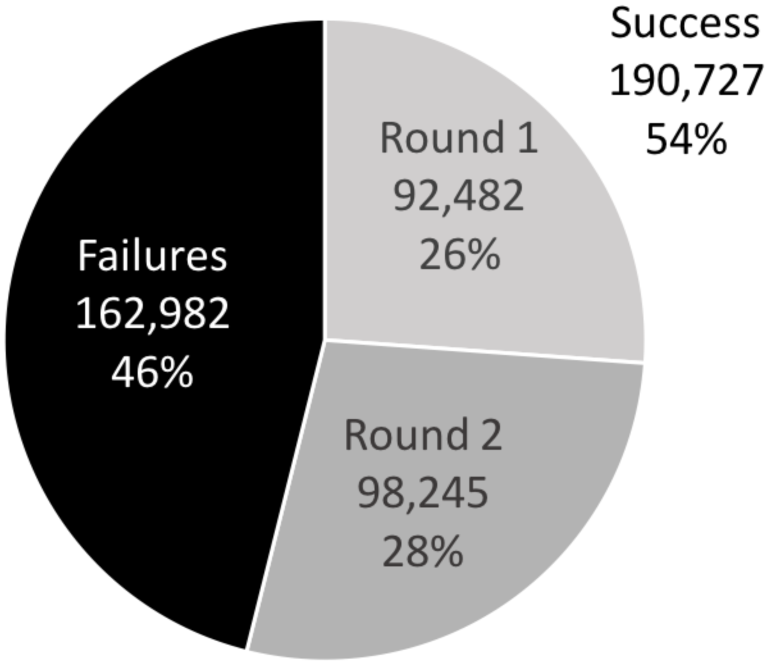
- Today:
  - **54% non-Android**

SourcererJBF



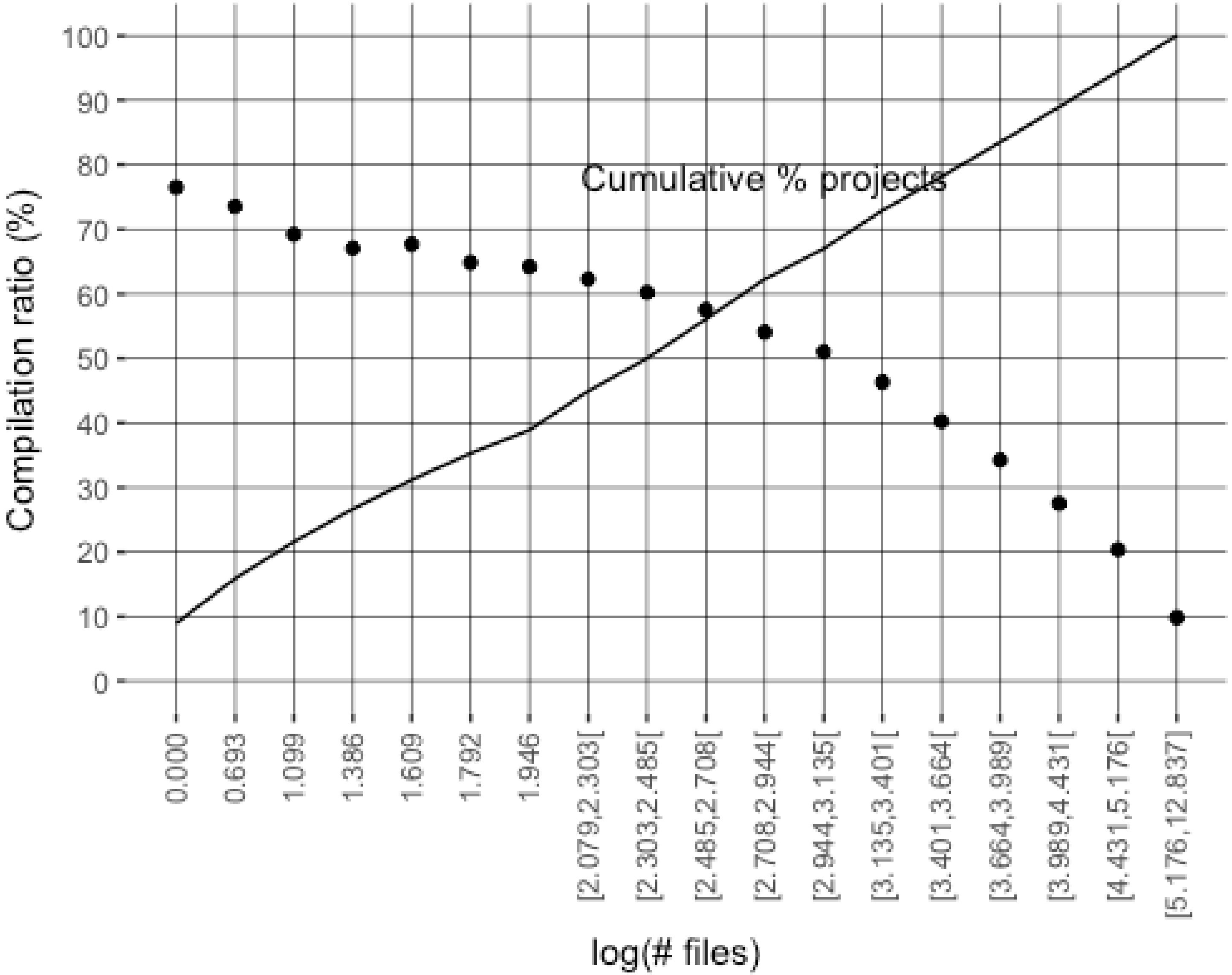
SourcererJBF Effectiveness

353,709 non-Android projects



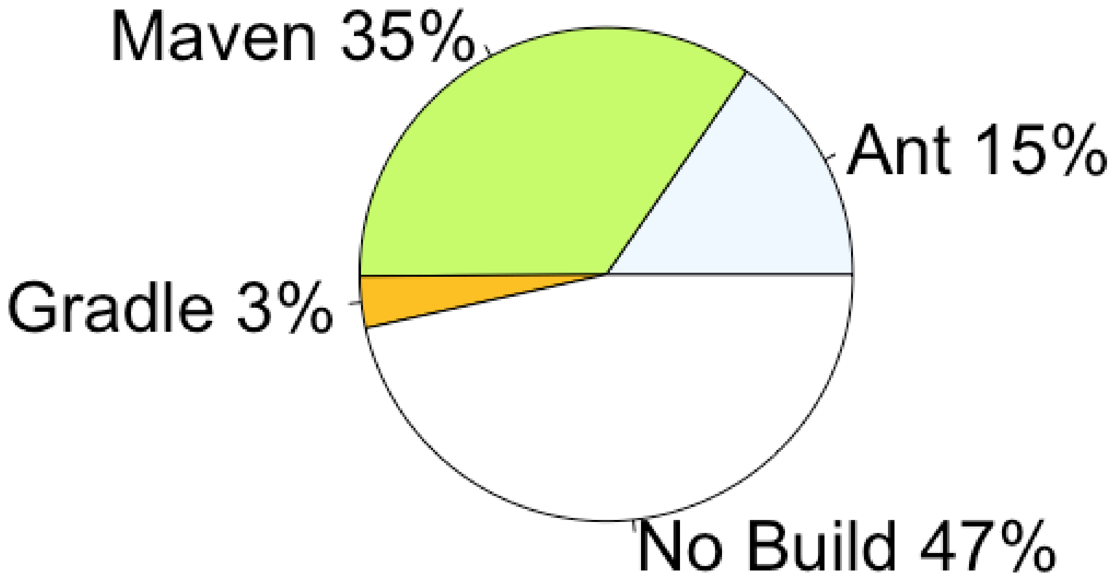
Success
54%

Correlation with Project Size?

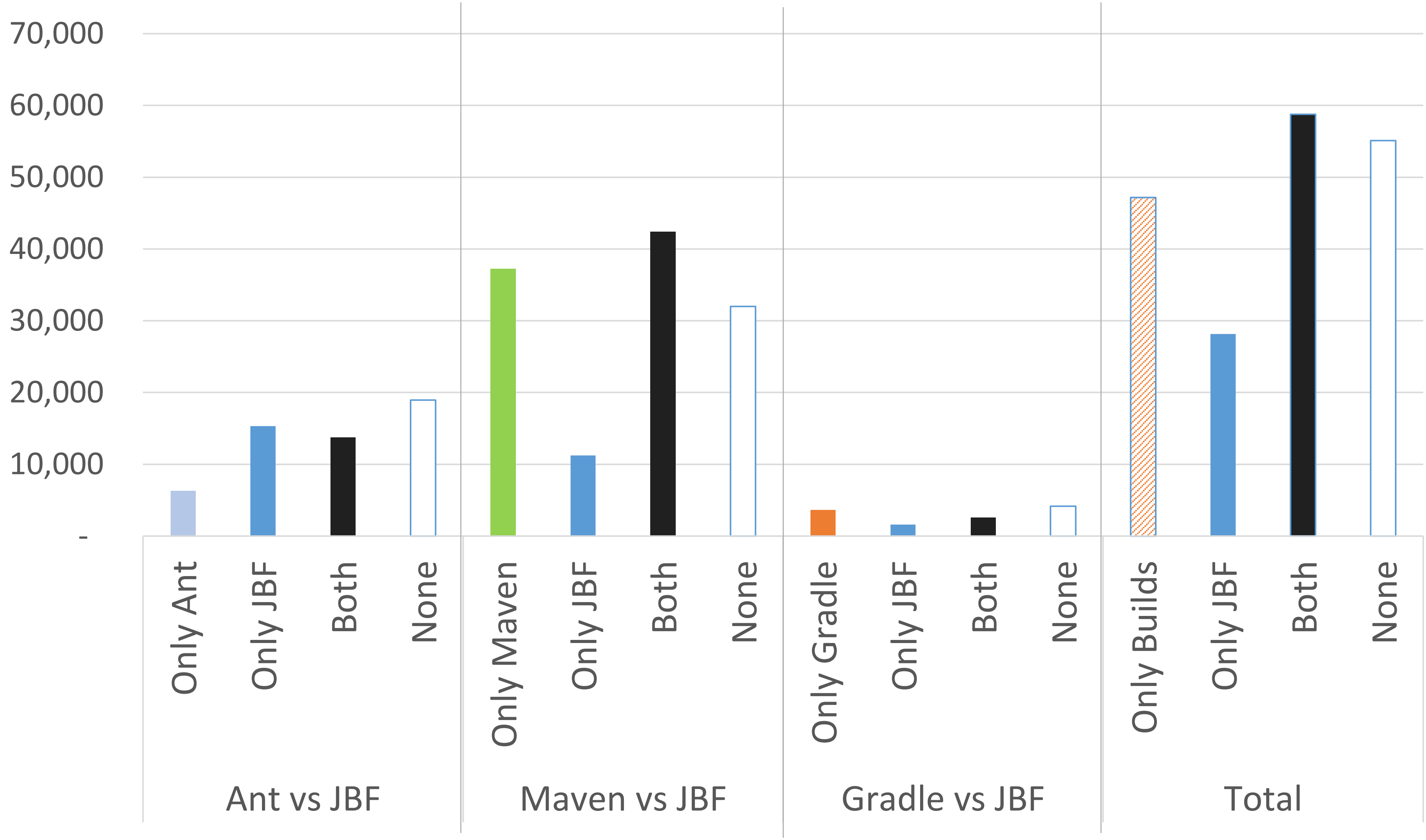


Could Own Build Scripts do Better?

189,220 out of 353,709 projects (53%)



Could Own Build Scripts do Better?



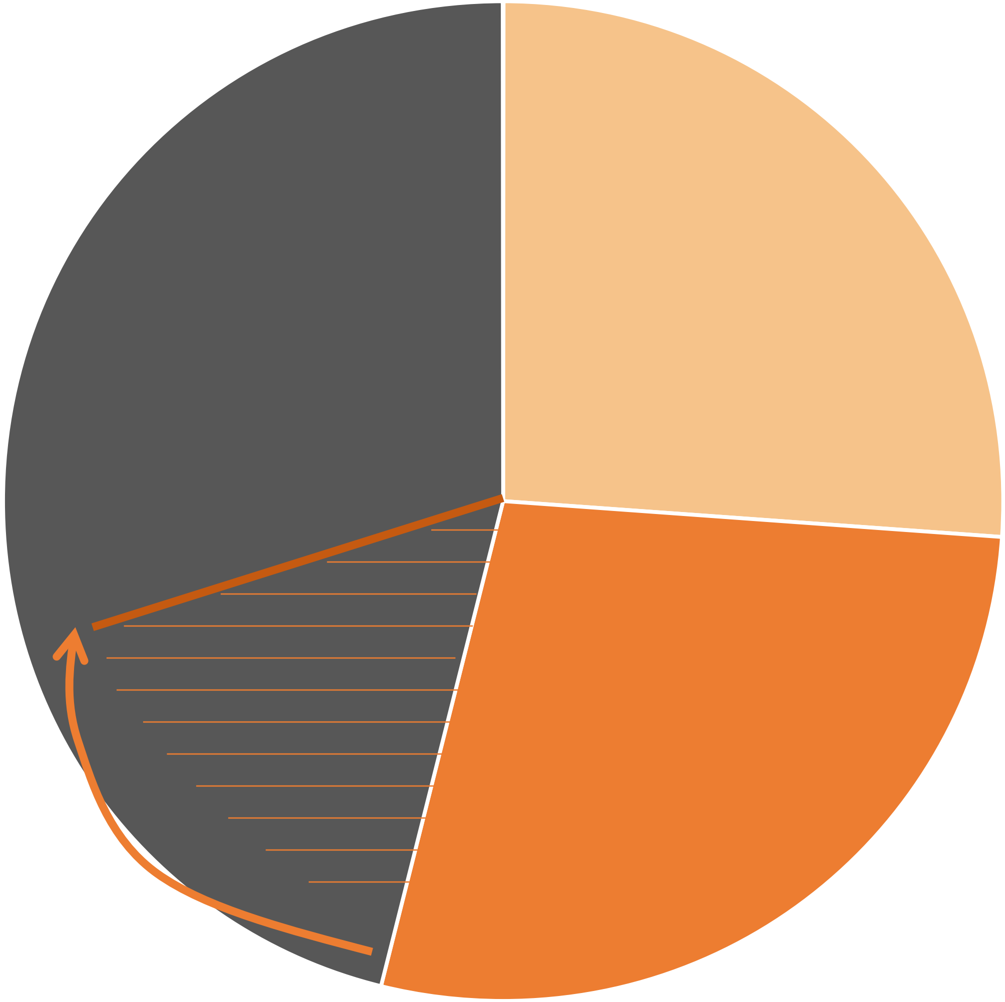
In **189,220** projects:
JBF: 86,926 (46%)
Own: 105,973 (56%)

In **353,709** projects:
JBF: 190,727 (**54%**) ←
Own: 105,973 (**30%**)

Problems with Own Builds

- Security and integrity of local build system
  - Crazy things happen!
- Unknown location of compiled code
  - Maybe jar’ed, may be moved into network, etc…
- Large variation of actions, not just compilation
  - “Success” means build script succeeded, not compilation succeeded
- Builds take much longer
  - JBF: 8 secs (median)
  - Own builds: 20 secs (median)

Improving SourcererJBF Effectiveness



Success now:
54%

↓

Success target:
67%

Doing Research with Big Data, the Bad

- Tera-byte sized datasets
  - Difficult to handle, share
- Requires $$ hardware
  - Currently: 112-core server, 512G RAM
- Processing can take weeks
  - Mistakes are expensive
- Scientific insights don’t necessarily need big data
  - Sampling

Doing Research with Big Data, the Good

- Useful applications require the whole data
- Scale presents new engineering challenges
  - Doctoral work worthy