

Characterizing Commits in Open-Source Software

Mívian Ferreira

Diego Gonçalves

Mariza Bigonha

Kecia Ferreira

Motivação

**Dados do GitHub têm sido amplamente utilizados
por pesquisadores**

Motivação

Dados do GitHub têm sido amplamente utilizados por pesquisadores

COMMITTS

Motivação

Dados do GitHub têm sido amplamente utilizados por pesquisadores

COMMITTS

Pouco se sabe sobre as características desses dados

Motivação

Por que devemos conhecer as características desses dados?

```
graph TD; A[Por que devemos conhecer as características desses dados?] --- B[Evitar a introdução de viés]; A --- C[Mitigar ameaças à validade]
```

Evitar a introdução de viés

Mitigar ameaças à validade

Objetivo

Caracterizar commits em relação:

- Número de arquivos modificados (RQ1)
- Número de arquivos de código-fonte modificados (RQ2)
- Número de arquivos modificados de acordo com a atividade relacionada ao commit (RQ3)
- Intervalo de tempo em que commits são realizados pelos desenvolvedores (RQ4)

Como fizemos?

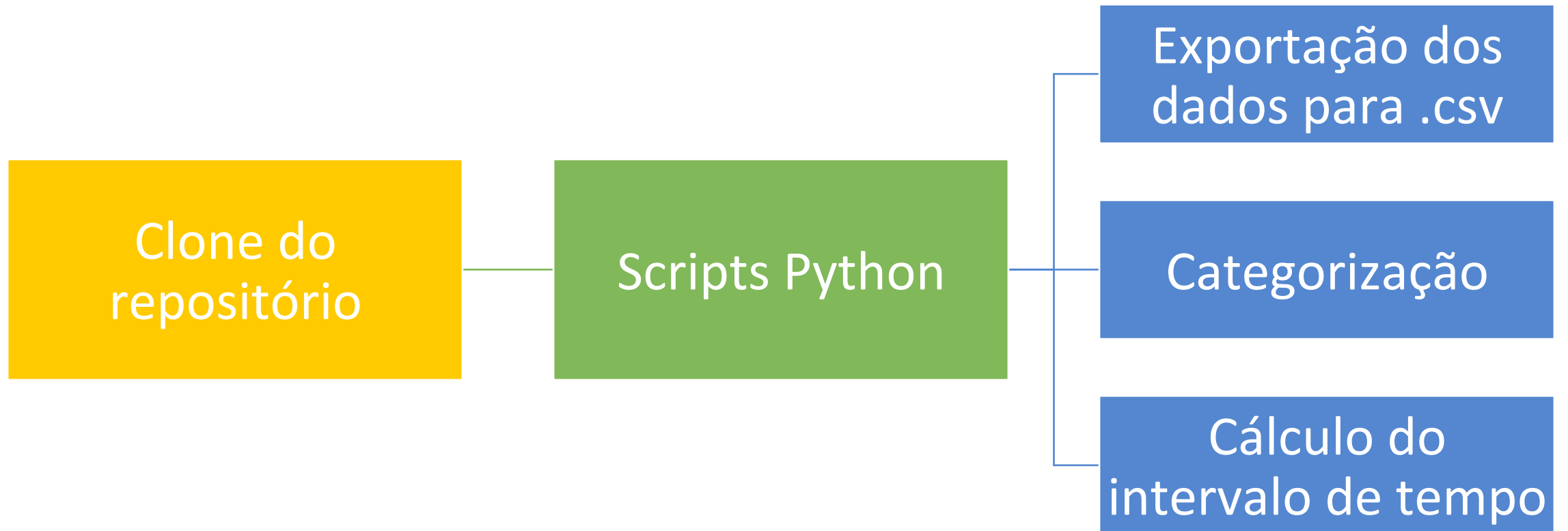
Dataset

- 24 projetos Java do GitHub
 - Projetos mais populares
 - Sistemas maduros – 3 a 11 anos
 - **≈ 1 milhão de commits analisados**

System	Age	#Commits	#Stars
ballerina-lang/ballerina-platform	3	96,121	2,644
neo4j/neo4j	8	69,702	8,315
jdk/openjdk	2	62,947	6,553
elasticsearch/elasticsearch	10	57,414	52,228
camel/apache	11	50,138	3,489
graal/oracle	4	53,665	13,950
language-tool/language-tool-org	7	46,224	4,114
vespa/vespa-engine	4	46,403	3,363
lucene-solr/apache	4	34,703	3,863
rstudio/rstudio	9	34,292	3,423
alluxio/alluxio	7	31,587	4,805
hazelcast/hazelcast	8	30,936	4,033
jenkins/jenkinsci	9	31,136	16,463
sonarqube/sonar-source	9	30,480	5,272
beam/apache	4	30,519	4,362
spring-boot/spring-projects	8	30,671	51,678
bazel/bazelbuild	6	28,662	15,673
shardingsphere/apache	4	28,457	12,387
ignite/apache	5	27,401	3,518
selenium/seleniumhq	7	26,432	19,074
cassandra/apache	11	25,994	6,278
flink/apache	6	25,543	14,626
hadoop/apache	6	24,584	11,041
tomcat/apache	9	22,909	4,984

Table 1: Dataset systems sorted by number of commits.

Extração de Dados



Categorização

Baseado em Hattori e Lanza (2008)

Category	Keywords
Merge	merge, pull request
Corrective	bug, fix, correct, miss, proper, broken, corrupted, failure, fault, deprecate, throw/catch exception, crash, typo
Forward	implement, add, request, new, test, increase, expansion, include, initial, create, introduce, launch, define, determine, support, extend, set
Reengineering	parallelize, optimization, adjust, update, delete, remove, expunge, cut off, refactor, replace, modification, improve, is/are now, change, rename, eliminate, duplicate, obsolete, enhance, restructure, alter, rearrange, withdraw, conversion, revision, simplify, move, relocate, downgrade, exclude, reuse, revert, extract, reset, redefine, edit, read, revamp, decouple
Management	clear, license, release, structure, integration, copyright, documentation, manual, javadoc, migrate, review, polish, upgrade, style, standardization, TODO, migration, organization, normalize, configure, ensure, resolve conflict, bump, dump, comment, format code, do not use

Intervalo de Tempo entre Commits

Para cada commit:

- Identificação do autor
- Agrupamento dos commits por autor

- Cálculo do intervalo de tempo

- Dev A fez commits às 12h, 13h e 13h30
- 2 intervalos: 60 e 30 minutos

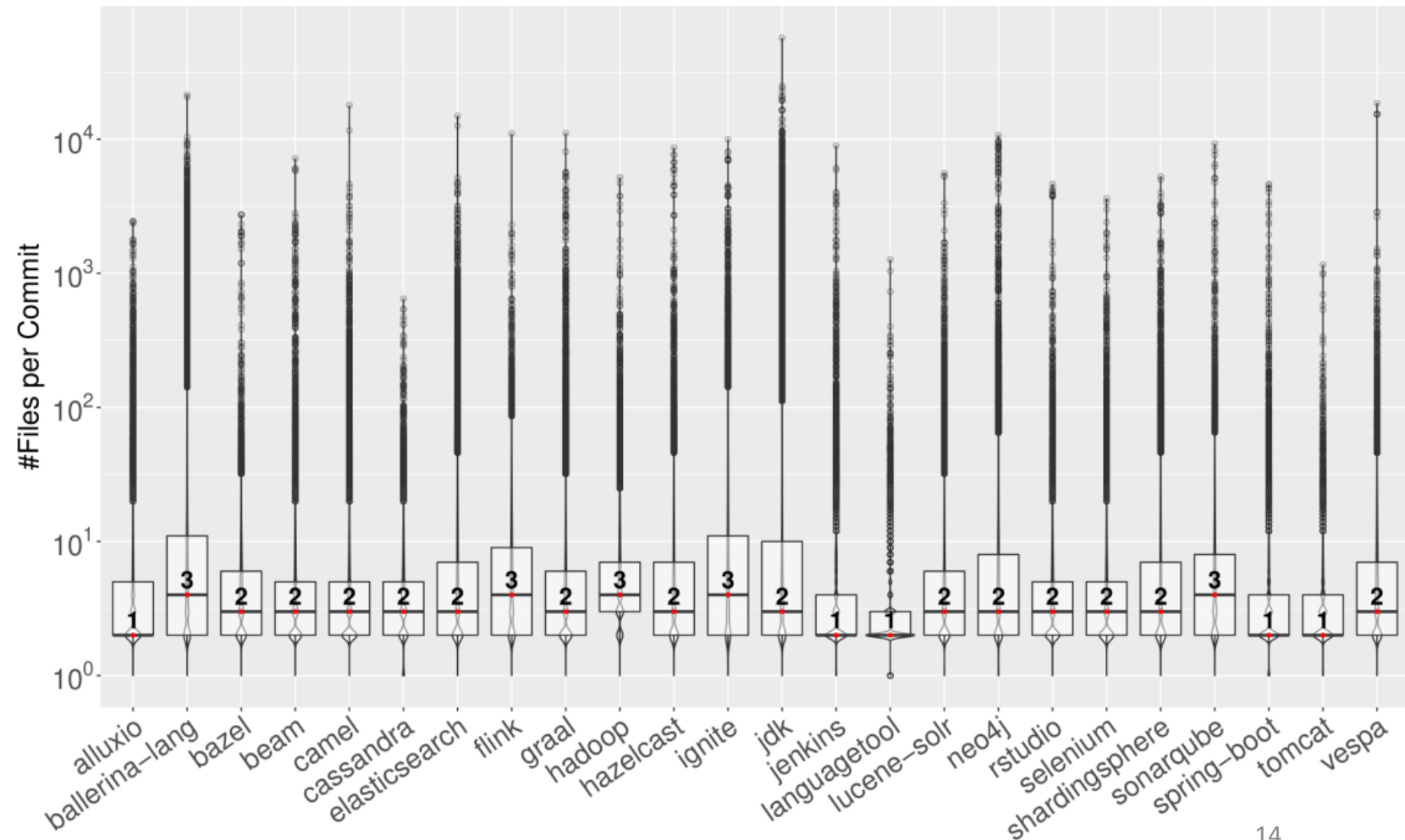
- Cálculo da média de tempo para cada desenvolvedor

Resultados

RQ1. Qual é o tamanho dos commits nos repositórios de software?

RQ1. Qual é o tamanho dos commits nos repositórios de software?

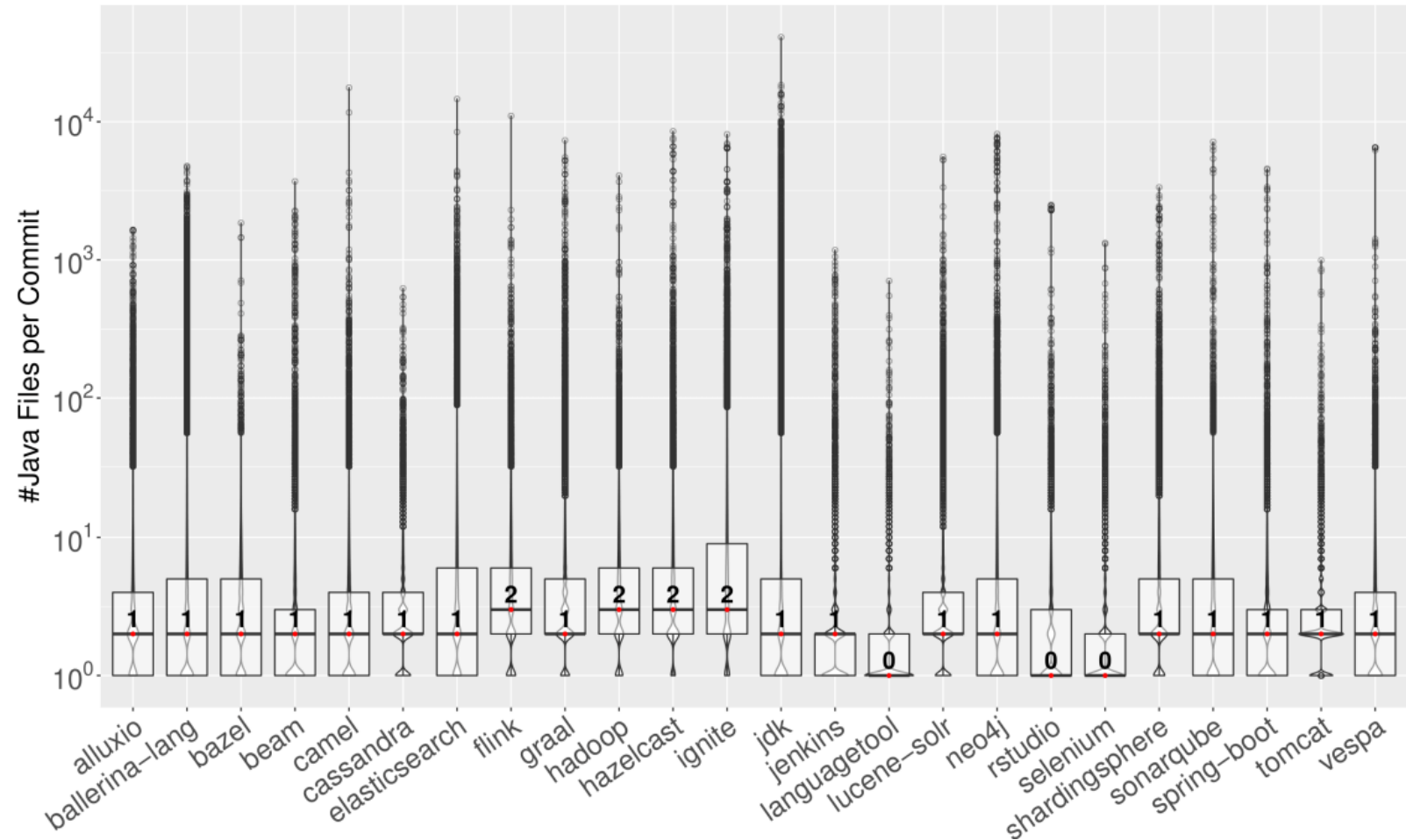
- Dados têm distribuição de cauda pesada
- Alguns commits modificam muitos arquivos de uma vez
- De modo geral, #arquivos varia de 1 a 10



RQ2. Qual é o tamanho dos commits envolvendo apenas arquivos Java?

RQ2. Qual é o tamanho dos commits envolvendo apenas arquivos Java?

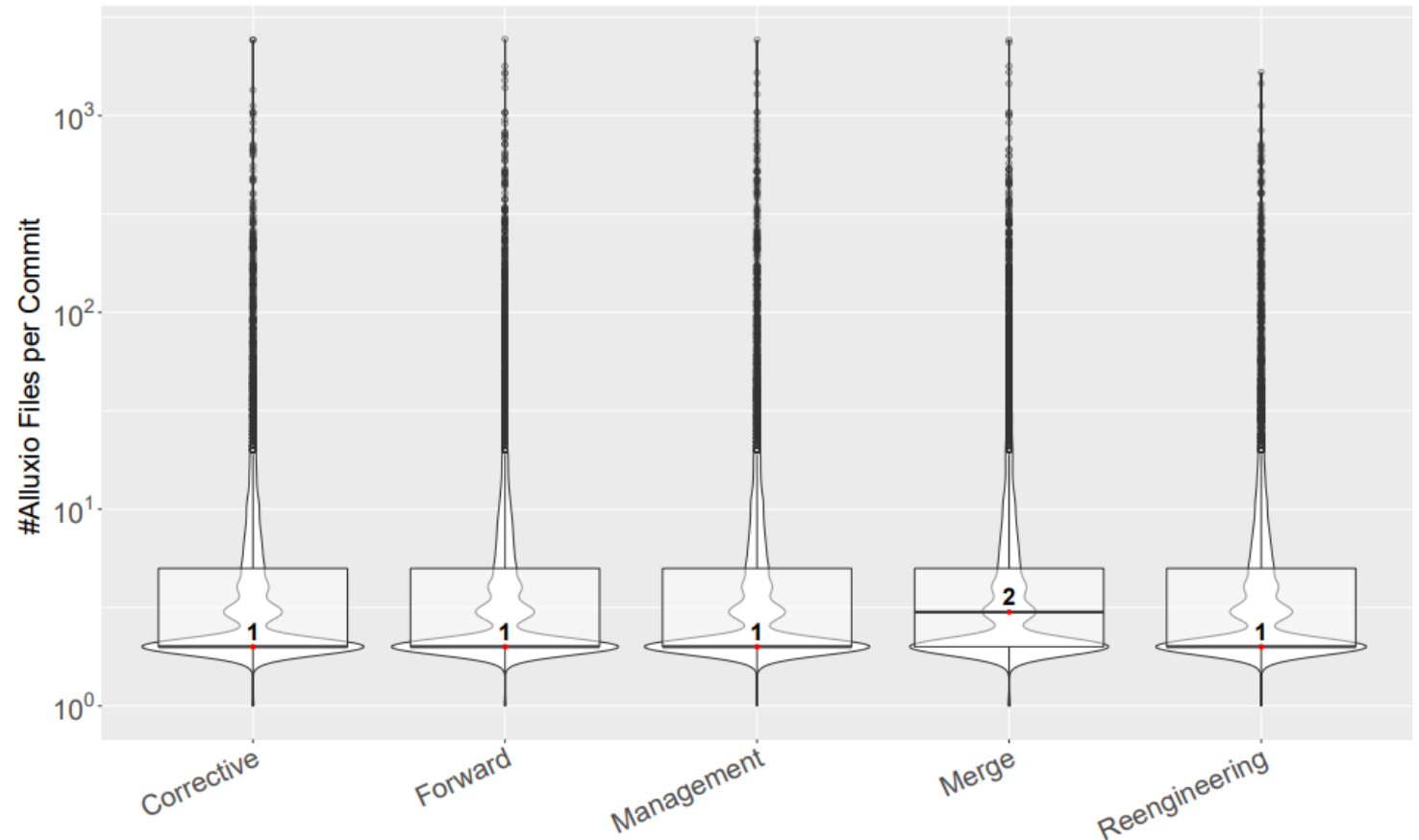
- Arquivos .java também apresentam distribuição de cauda pesada
- Alguns commits modificam muitos arquivos de uma vez
- De modo geral, #arquivos .java varia de 1 a 4



RQ3. Qual é o tamanho dos commits de acordo com as categorias às quais eles pertencem?

RQ3. Qual é o tamanho dos commits de acordo com as categorias às quais eles pertencem?

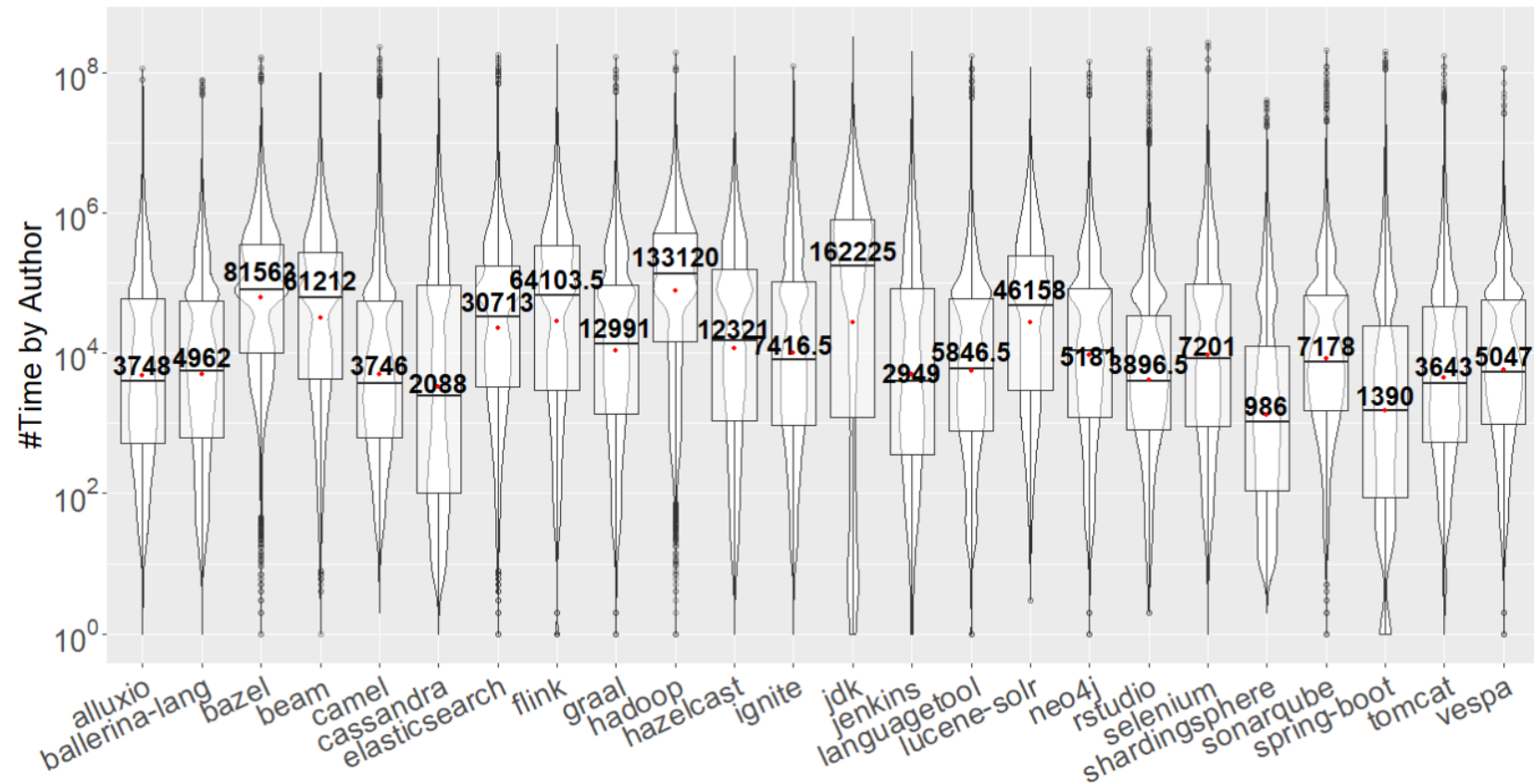
- Não existe diferença significativa entre o número de arquivos por categoria
- Todas as categorias apresentam valores entre 1 e 3 arquivos, em geral



RQ4. Qual é o intervalo de tempo em que um desenvolvedor realiza um commit?

RQ4. Qual é o intervalo de tempo em que um desenvolvedor realiza um commit?

- Desenvolvedores realizam commits em média a cada 8 horas



Conclusões

Conclusões

- Compreender as características do *dataset* é fator crítico no desenvolvimento de uma pesquisa
- O tamanho do commit importa
 - Um único commit pode conter centenas de arquivos e isso não ocorre apenas para Merge e Management
- Não podemos aplicar métodos estatísticas de distribuição Normal a esses dados
- Resultado relevante para trabalhos sobre *co-change*, autoria de código, análise de impacto de modificação, entre outros.

Trabalhos Futuros

- Análise de commits em projetos de outras linguagens de programação
- Relação entre commits e outros dados disponíveis no GitHub como issues
- Uso de técnicas como processamento de linguagem natural na categorização dos commits

Obrigada pela atenção!

Dúvidas?

Sinta-se à vontade para entrar em contato conosco:

mivianmarques@gmail.com

disantosg18@gmail.com

mariza@dcc.ufmg.br

kecia@cefetmg.br