# Usage and Attribution of Code Snippets in GitHub Projects

## License of Stack Overflow Content

| Attribution: Developers using code snippets from Stack Overflow must attribute author and origin of the snippet. |
| Share Alike: Derived work must be distributed under the same license. |

---

## Research Questions

### RQ1

**Method:** We searched for references to Stack Overflow content in source code files using the BigQuery GitHub data set and a regular expression. Then, we quantitatively and qualitatively analyzed the found references and referenced code snippets.

**Results:**
- On average three times more references to whole threads than to specific answers.
- One out of 357 files (0.28%) and one out of 32 repositories (3.15%) contained a reference to Stack Overflow.
- JavaScript, Python, and R contained more references than other languages.

### RQ2

**Method:** We further analyzed the data collected for RQ1.

**Results:**
- Frequently referenced questions and answers have a significantly higher view count and score.
- However, the dispersion of values is relatively high.

### RQ3

**Method:** (1) We build regular expressions matching the ten most frequently referenced Java code snippets and used BigQuery to match and analyze source code files. (2) We calibrated and used a code clone detector to find exact clones of Java snippets in a random sample of active GitHub Java projects.

**Results:**
1. For the code snippets from the ten most frequently referenced Java answers on Stack Overflow, we found that at most 27% of their usages were attributed (in all GitHub Java projects).
2. Using the code clone detector CPD, we found that in a random sample of Java projects (n=2,313), 207 repositories (9%) contained a copy of a snippet from our set of snippets (n=396). Only 19% of the matched files contained a reference to Stack Overflow.

---

*Usage without attribution leads to legal and maintenance issues.*

# Winter School on Big Software on the Run

October 23-28, 2016, Ede/Netherlands