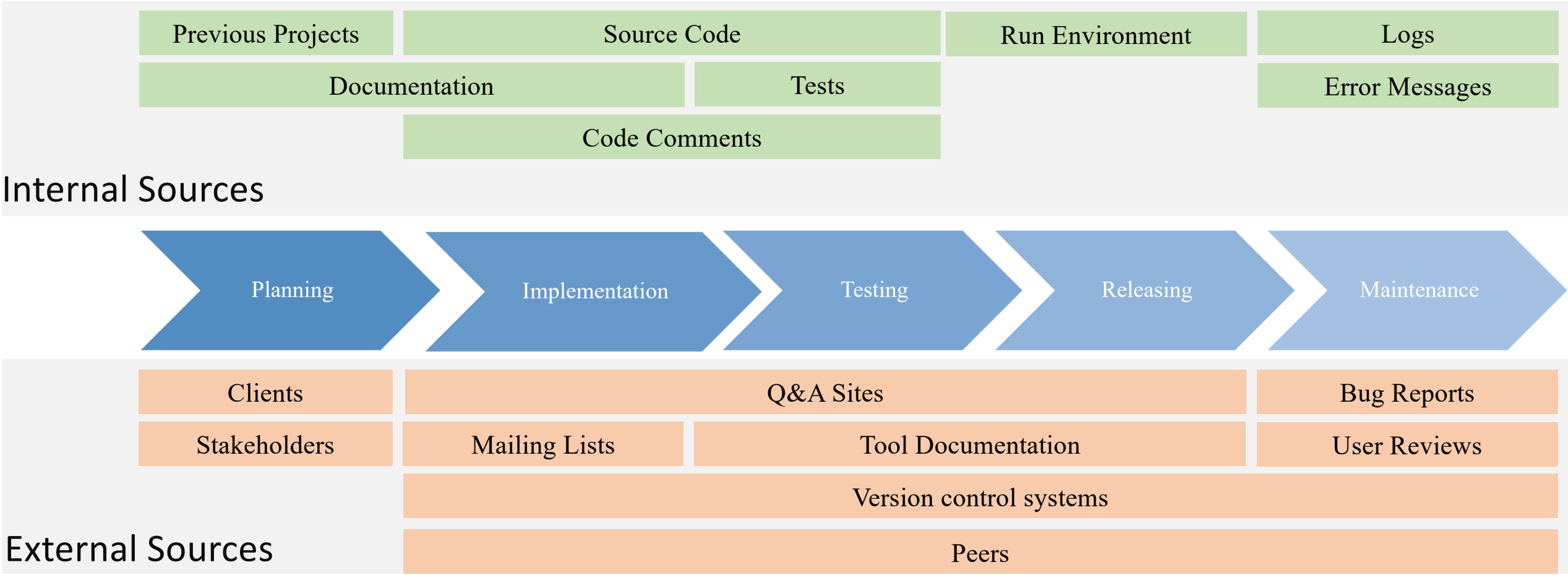


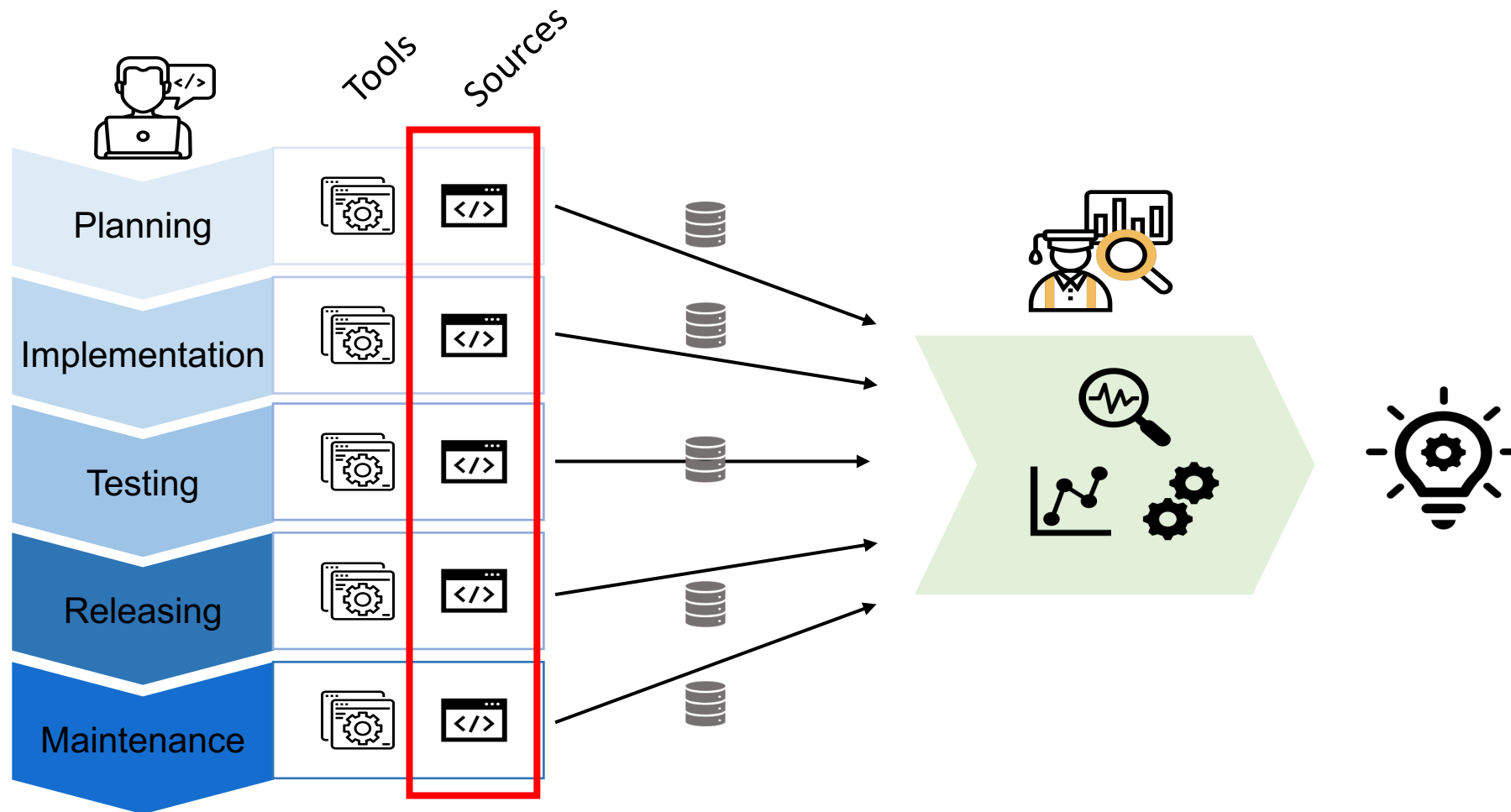
What do Developers Discuss about Code Comment Conventions on Social Media?

Pooja Rani, Mathias Birrer, Sebastiano Panichella
Mohammad Ghafari, Oscar Nierstrasz

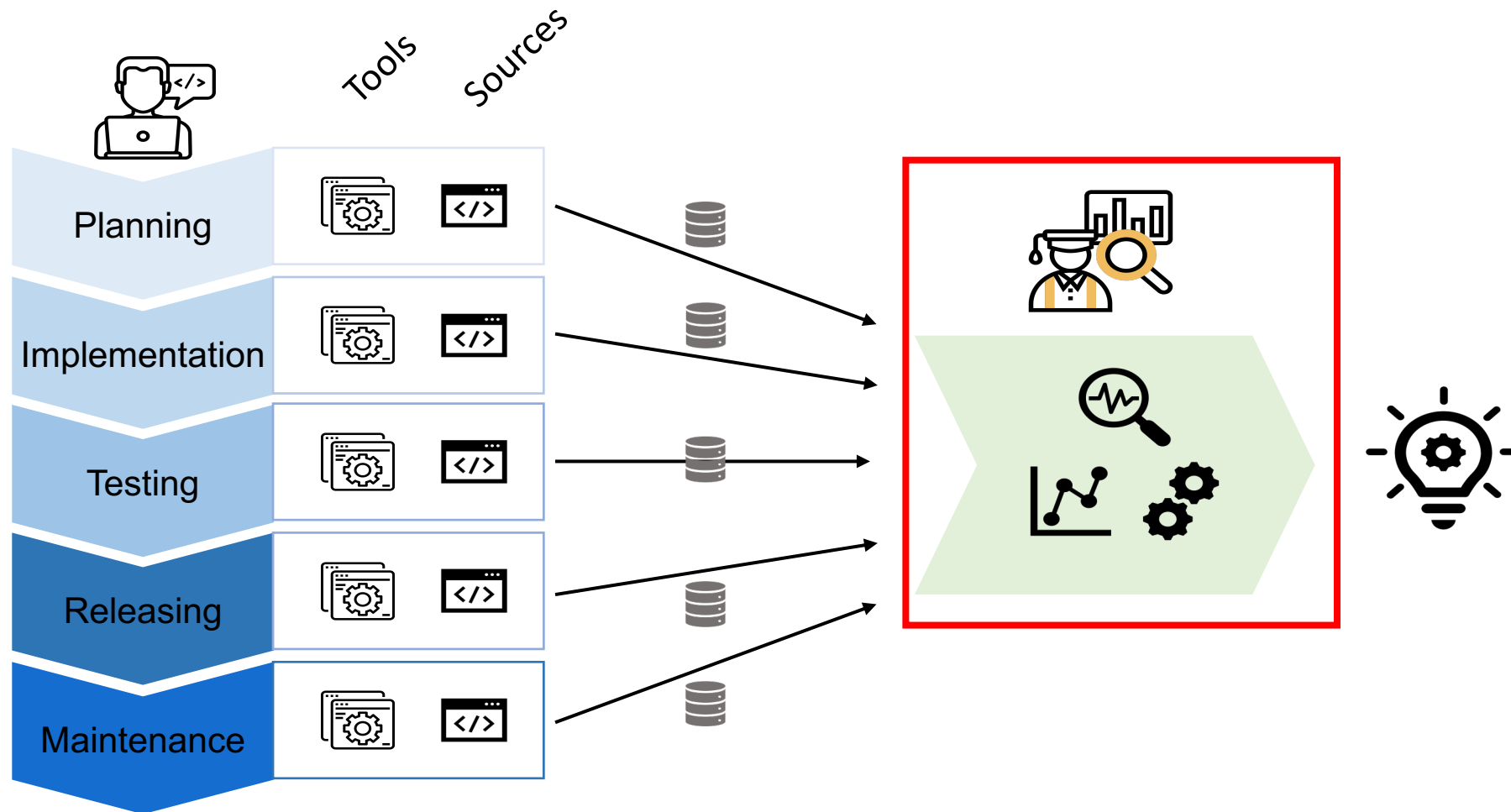
Throughout development phases developers typically consult internal or external sources of information depending on the phase and task to perform to understand the software structure and its components.

Developers' Information Needs



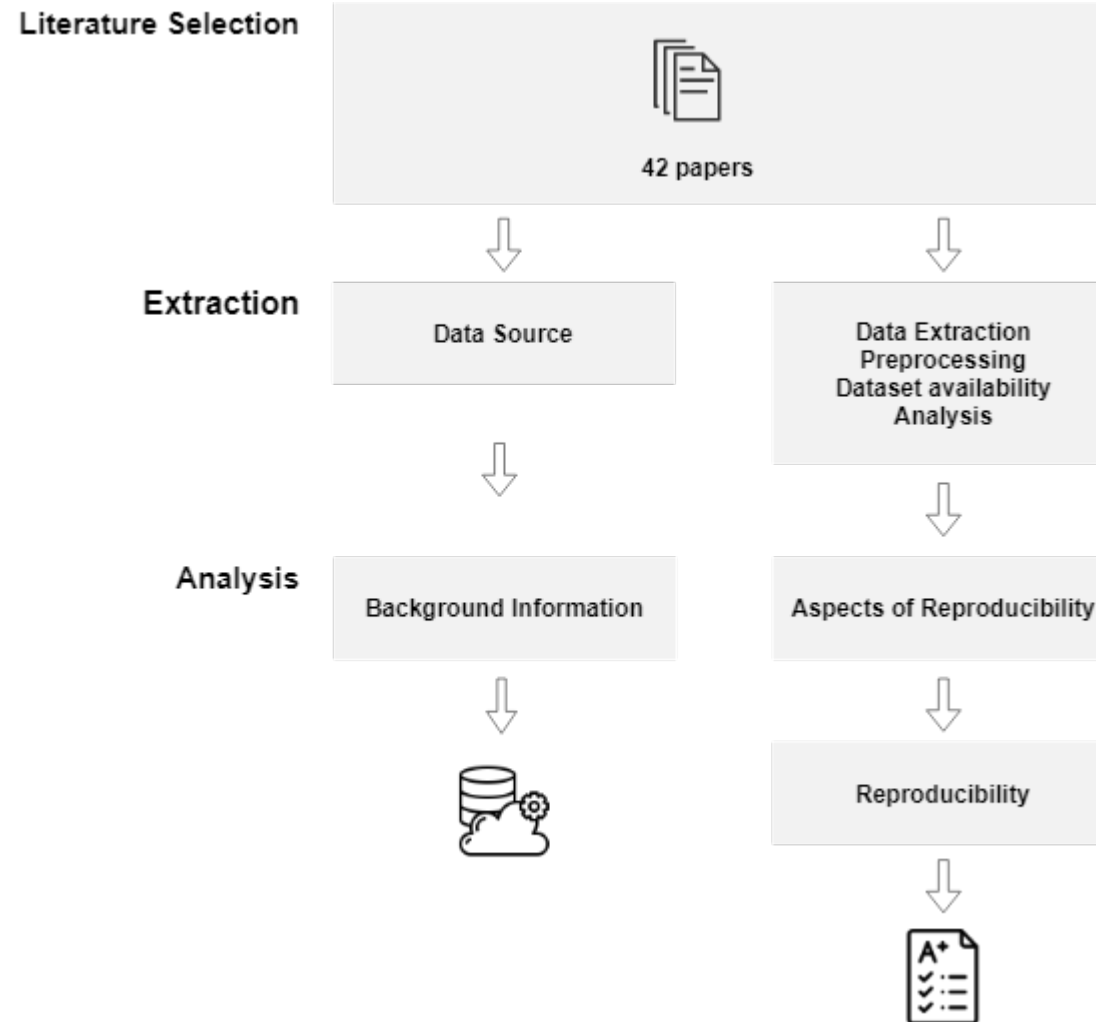


Background Study: Which data sources are typically analyzed by researchers to understand developers' information needs? What challenges do researchers face in conducting these studies?

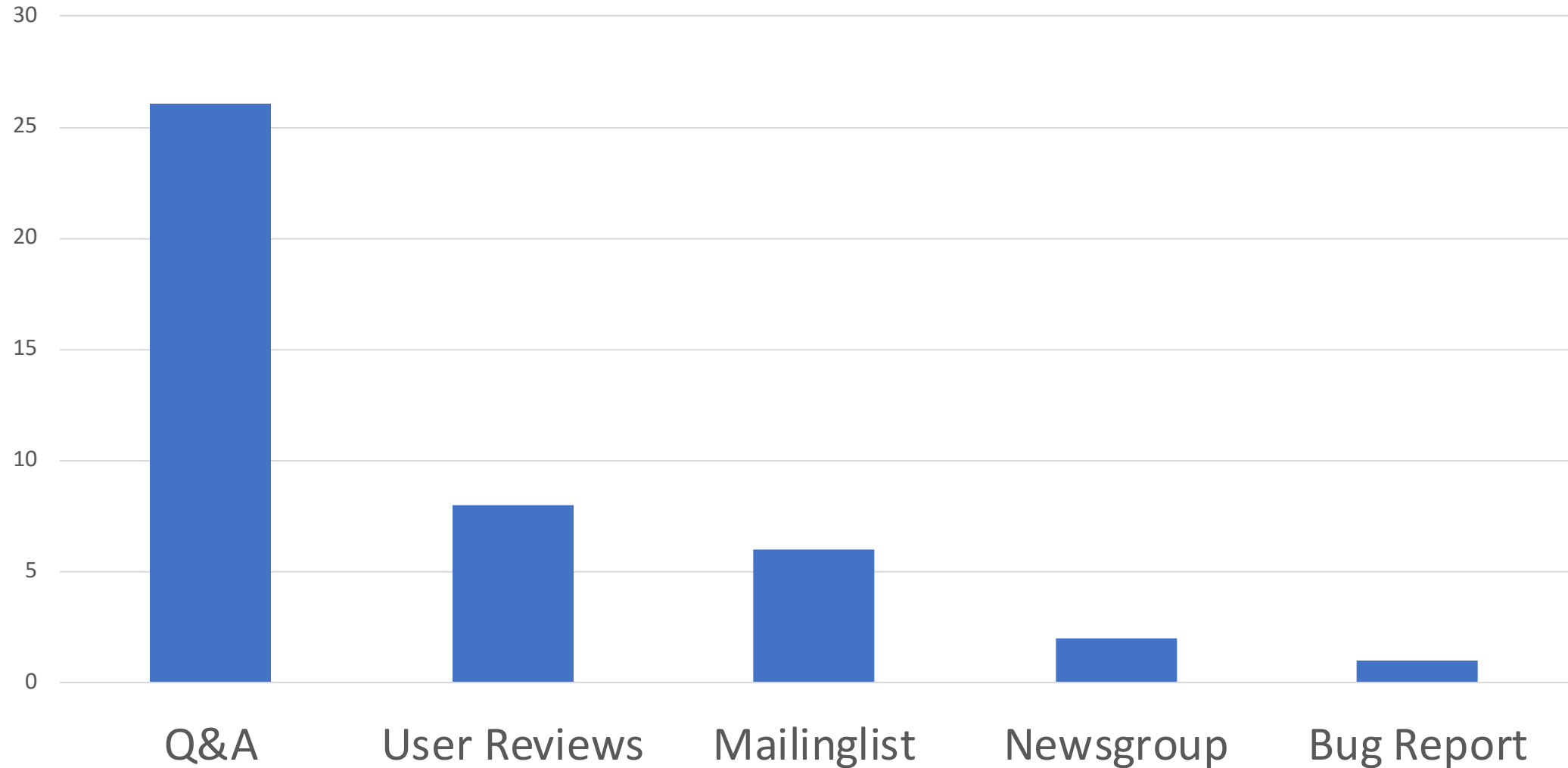


Background Study: What is the impact of such studies in *reproducibility*?

Methodology for Background Study



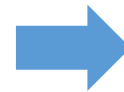
Results: Data sources analyzed by researchers



Benefits of External Sources

Sources: *Mailing Lists, Q&A Sites, Bug Trackers, News Sites, ...*

- Present recurrent questions of developers
- Easy access to their data
- Experts answer developer questions
- Contains years of data and thus present an evolution aspect
- To analyze developer and user feedback

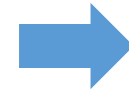


Researchers started analyzing these sources

Challenges in External Sources

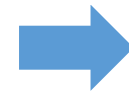
Sources: *Mailing Lists, Q&A Sites, Bug Trackers, News Sites, ...*

- Contains unstructured data
- Selecting relevant data
- Cleaning data for the analysis



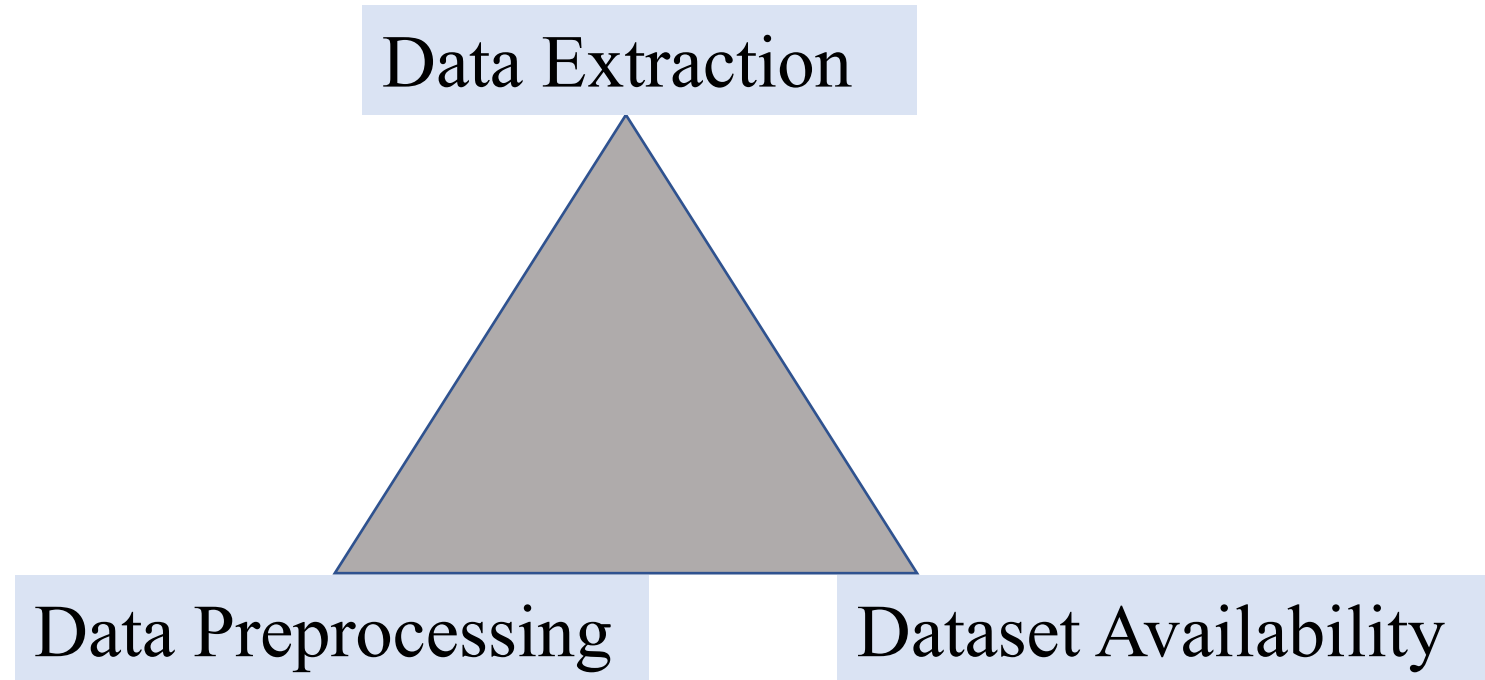
Researchers started investigating these sources

- To compare different sources / communities
- To uncover their evolution
- To reuse datasets

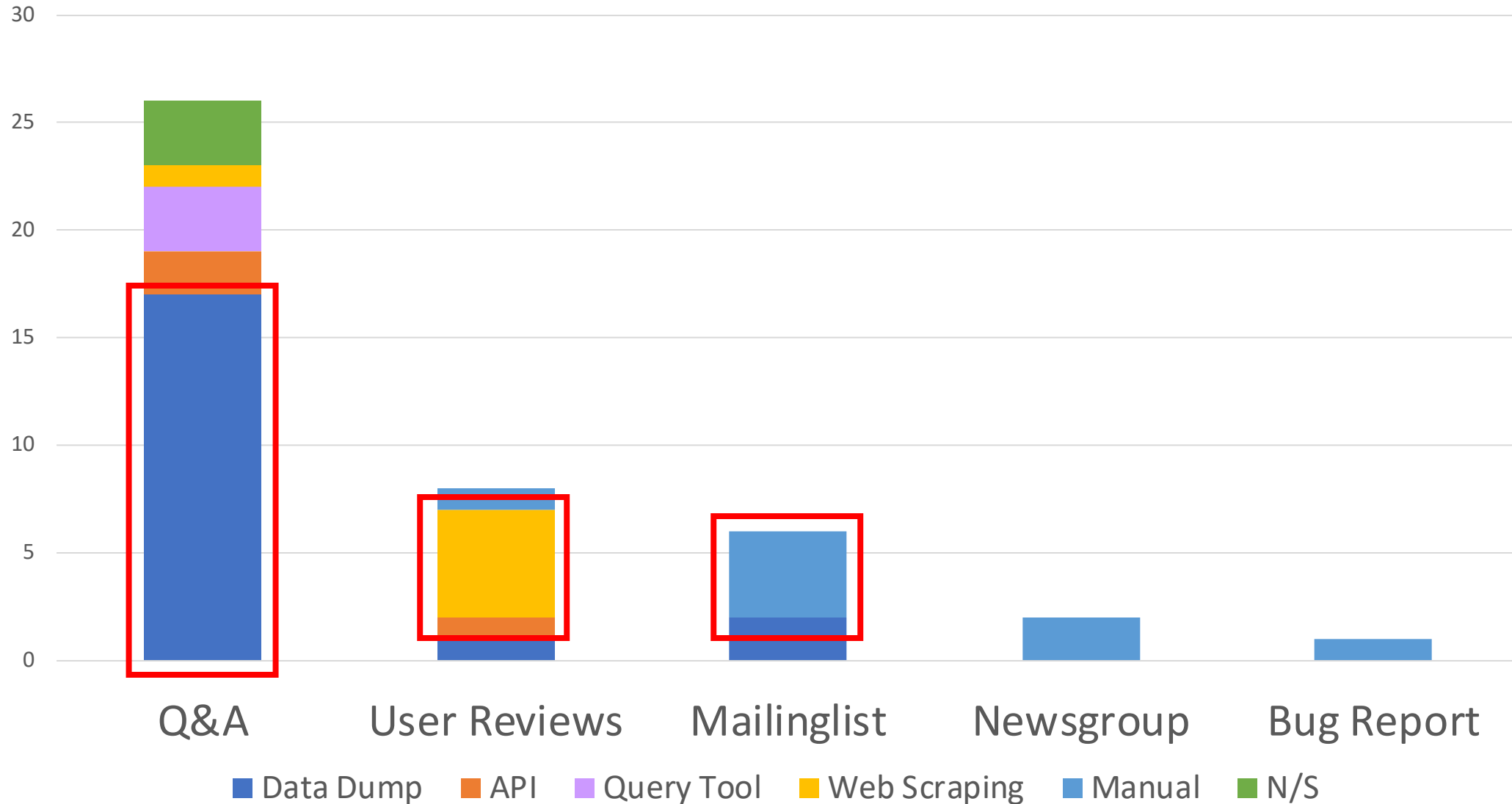


Reproducibility is highly important in such studies

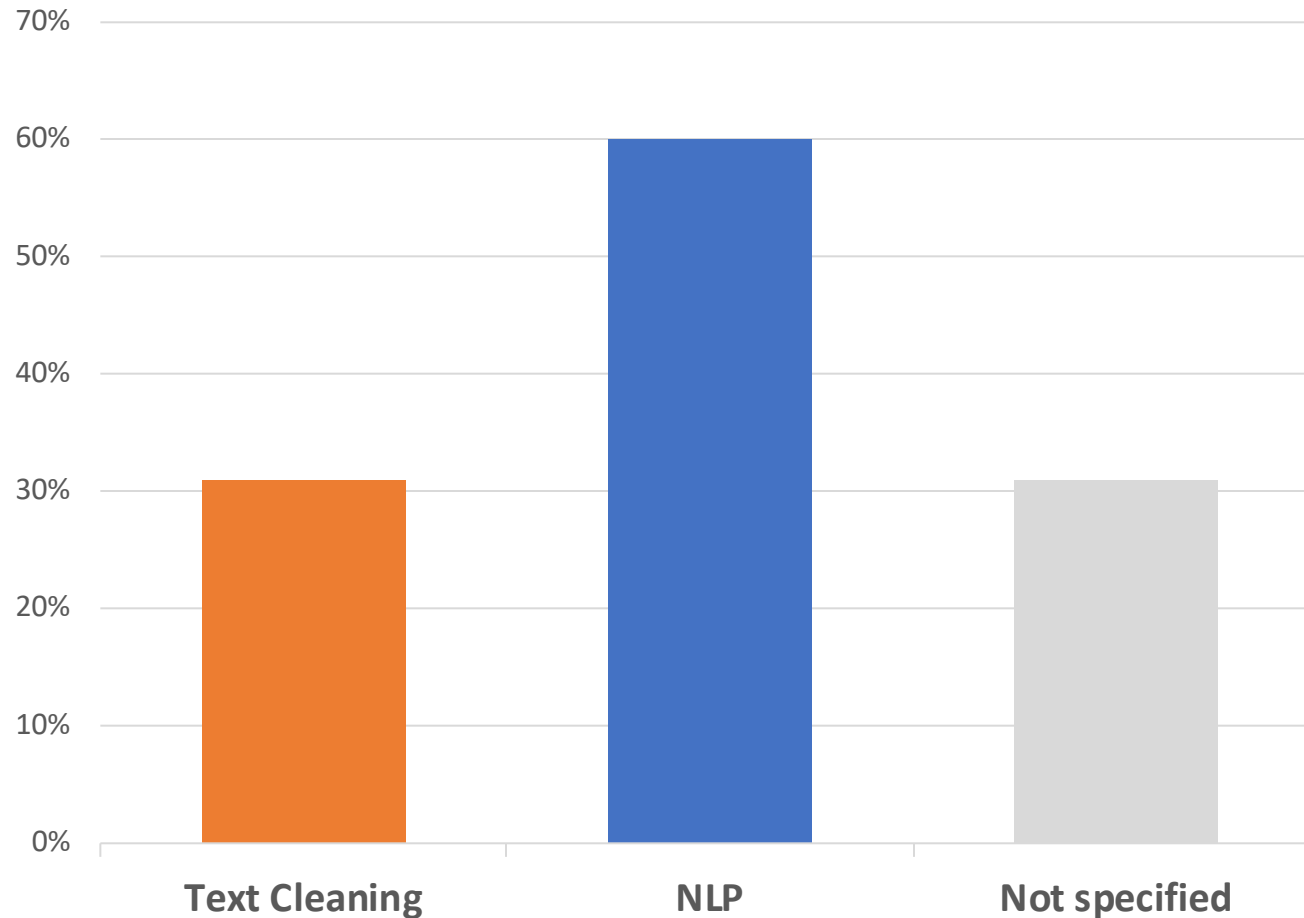
Reproducibility Aspects



Data Extraction: How data is extracted from the sources?



Data Preprocessing: Type of Preprocessing performed



Text cleaning

- Remove **source code**
- Remove **HTML tags**
- Remove **Punctuation**
- Remove **Non-Alpha-Numeric**

NLP cleaning

- Remove **Stop words**
- Apply **Word stemming**
- Apply **Lemmatization**
- Filter **Non-English**
- **Case unification**

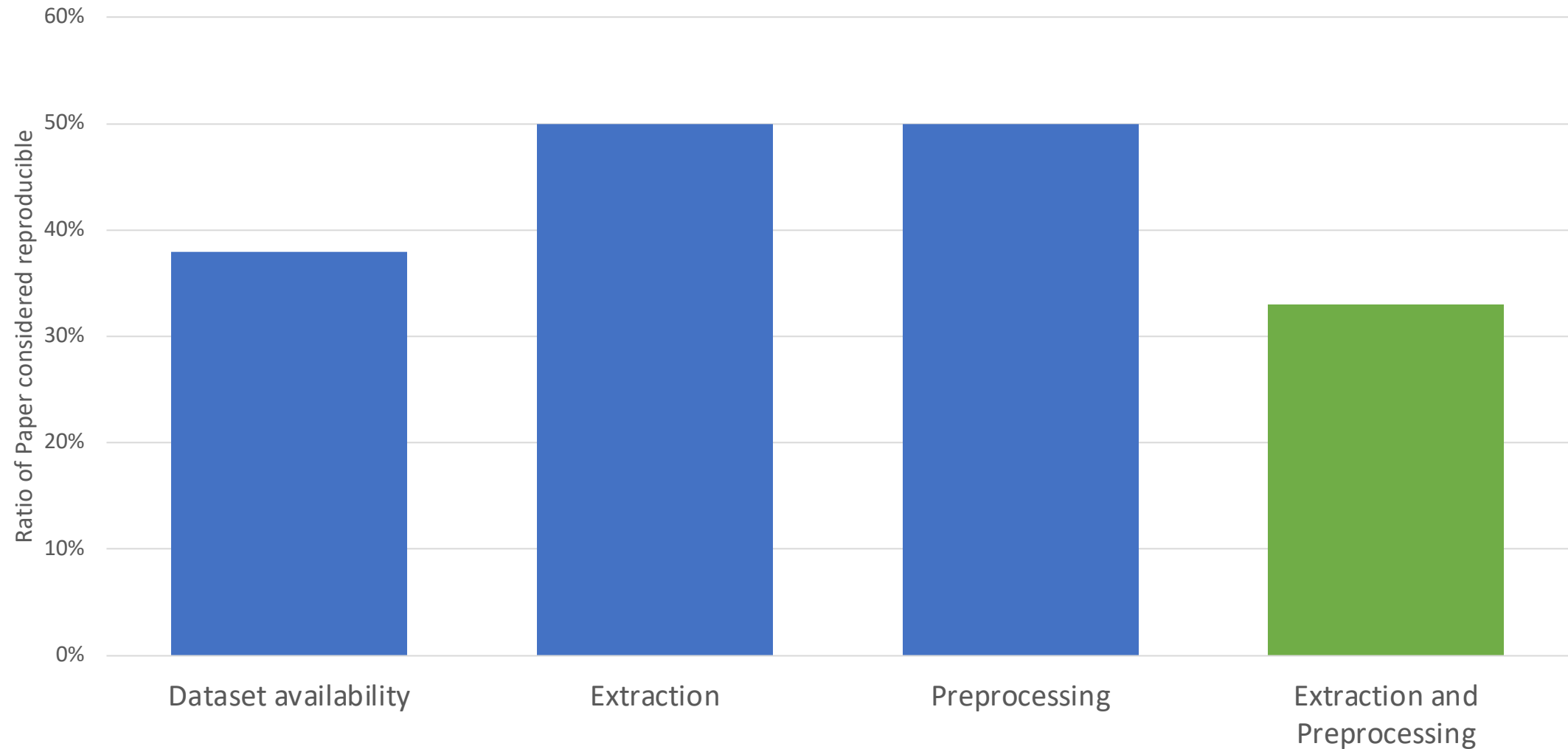
Text cleaning is highly dependent of the data source and extraction method, e.g Stack Overflow questions have HTML tags which need to be cleaned. NLP preprocessing is dependent on the analysis to be performed with the data.

Common pre-processing methodology emerged from the studies



NB: This common workflow emerged mainly from the studies using Stack Overflow as data source

Dataset Availability



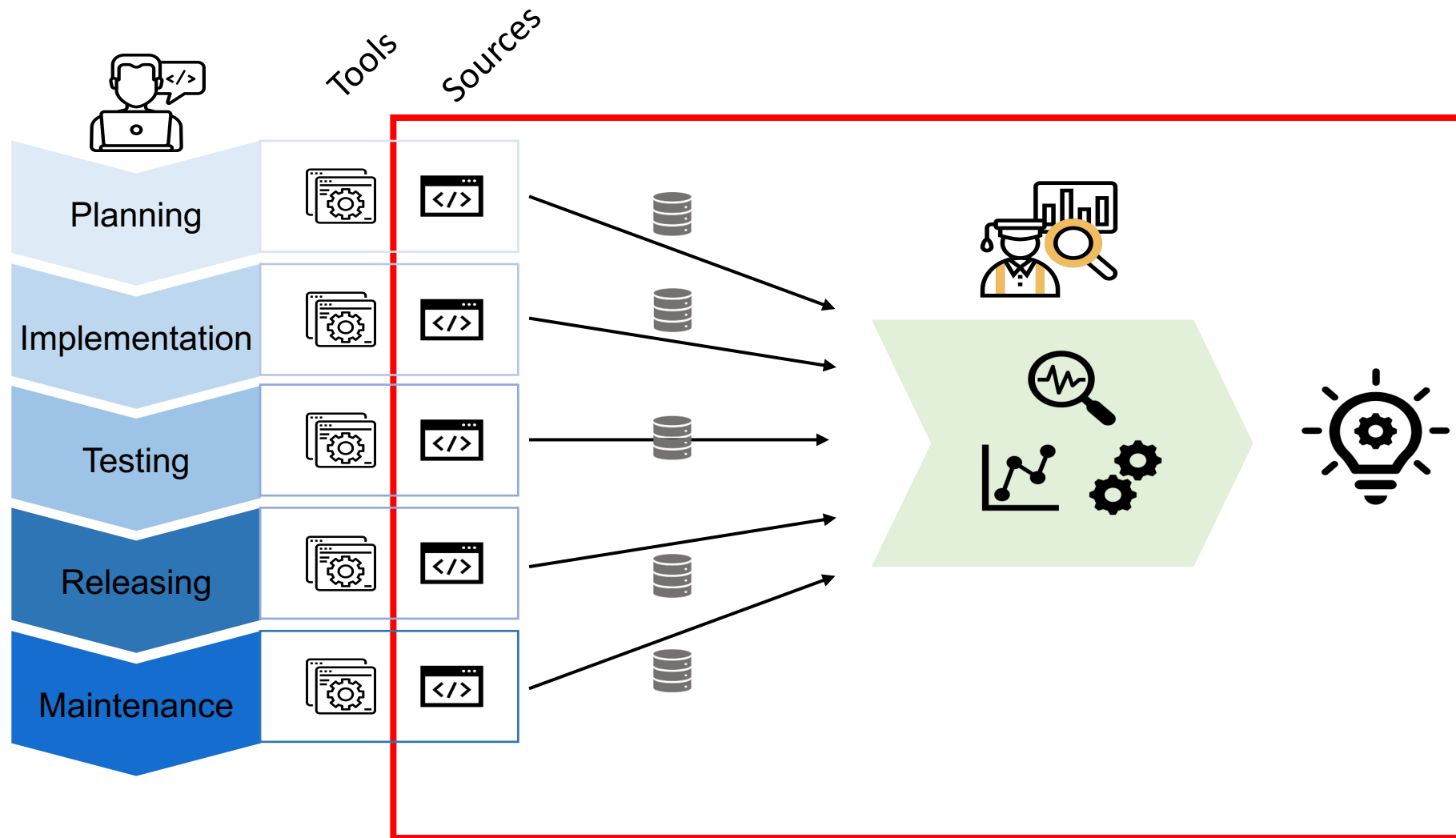
Despite low requirement of dataset availability, only 38% of the analysed studies publish a dataset for replication purposes

Study Goal

What Developers discuss about “*Code Comment Conventions*” on Social Media?

Code Comment Conventions

- Code comments are trustworthy form of documentation
- Basis for documentation tools
- Style & Syntax is not enforced by compiler
- Different conventions for Languages, Companies, Projects, Developers
- **Confusion amongst developers**



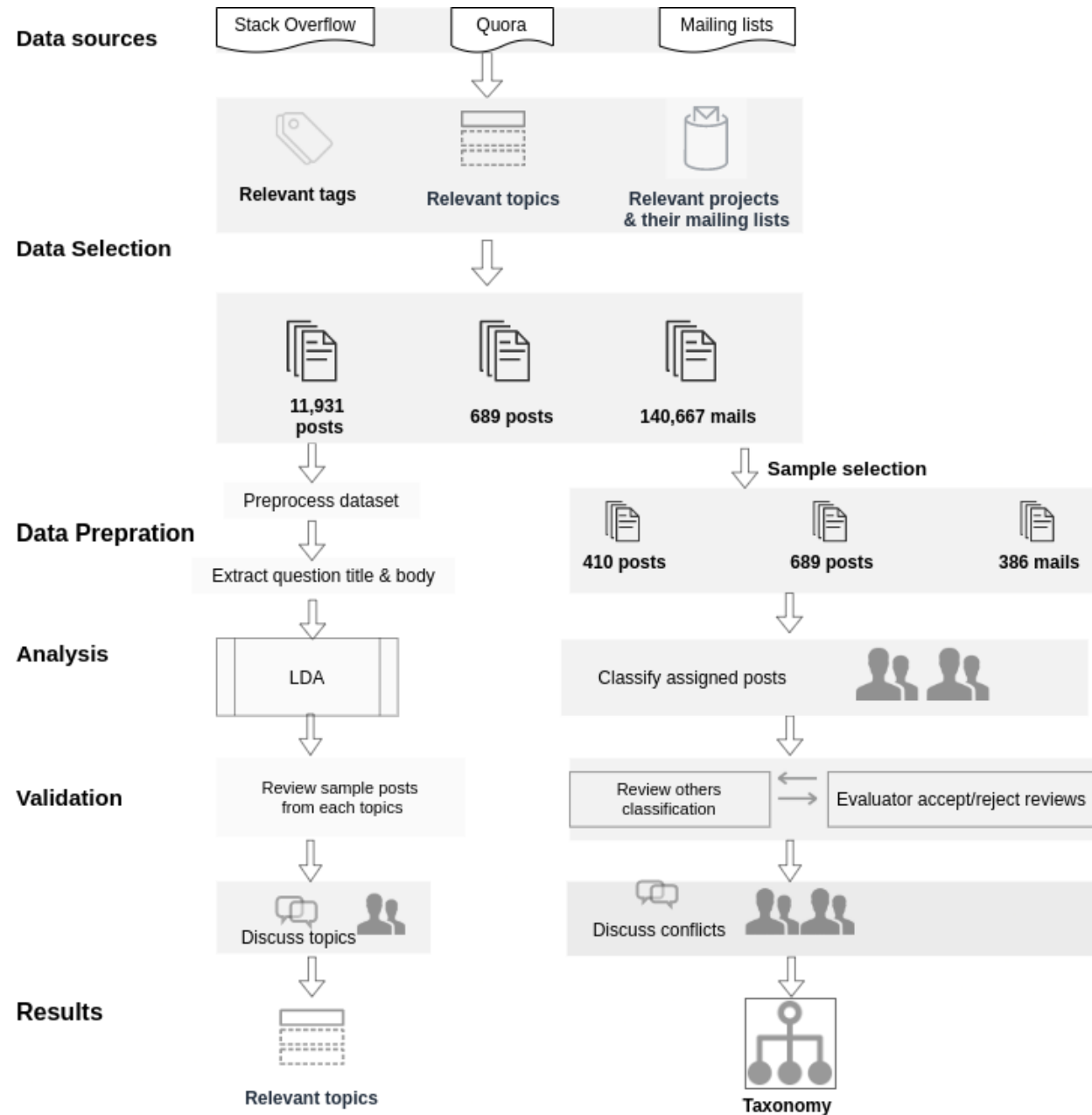
Main Study: What Developers discuss about “*Code Comment Conventions*” on Social Media?

Research Questions

RQ1: What topics are discussed by developers about commenting conventions?

RQ2: What types of questions and problems developers discuss on various platforms?

Methodology



Results: Mailing Lists

- We investigated the replication package of the previous study in addition to manual analysis of statistical significant sample set.
- Despite previous study on documentation issues
- No relevant data found concerning code commenting conventions

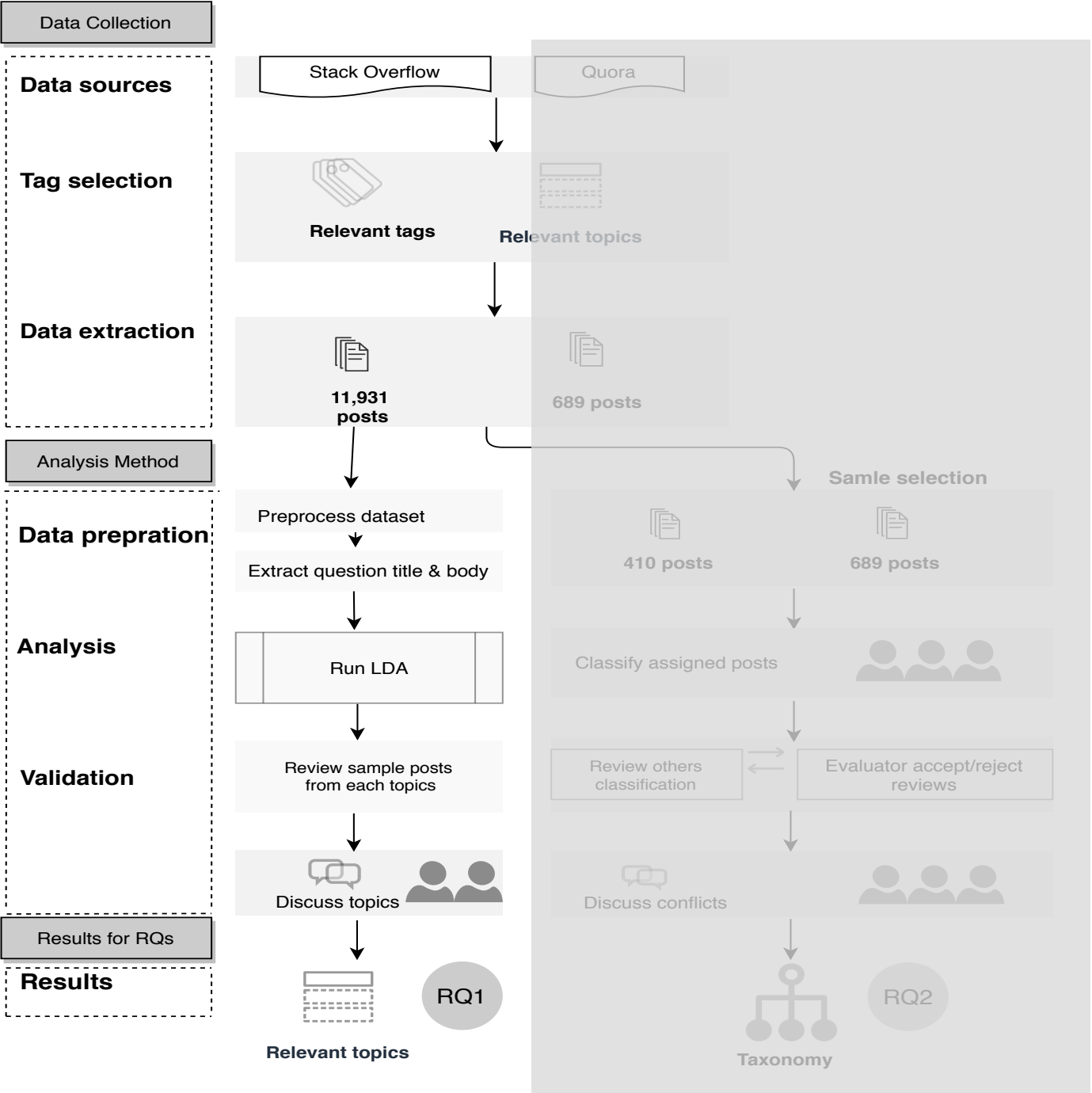
New Methodology



Research Questions

RQ1: What topics are discussed by developers about commenting conventions?

RQ1- LDA Analysis



LDA Technical Details

- Stack overflow posts: 11, 931
- MALLET
- Topics $k = 14$
- Hyperparameters
 - $\alpha = 5$
 - $\beta = 0.01$

LDA topic modeling yielded the following 14 topics.
Expected topics like “Documentation Generation” or
“Comments Syntax” were successfully identified by LDA.

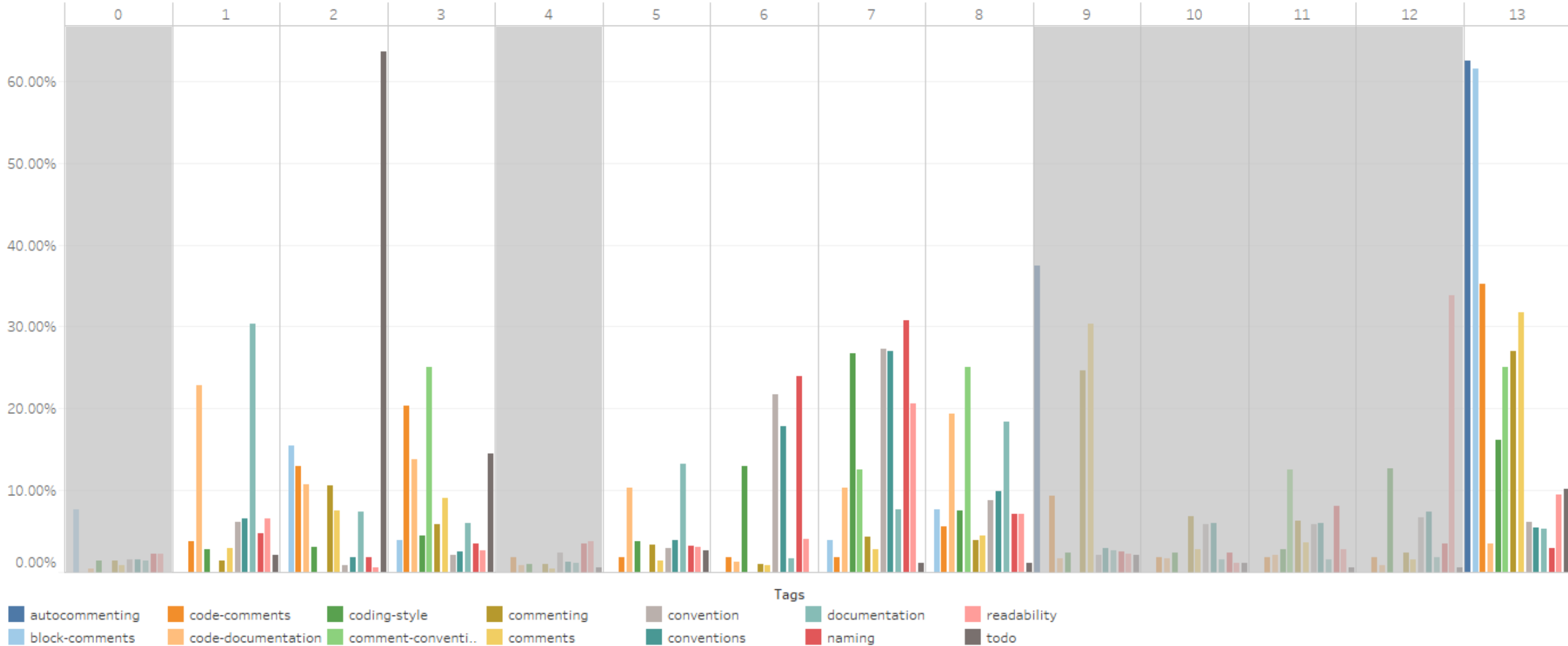
LDA Topics

#	Topic Name
0	Exceptions
1	Documentation Generation
2	IDE & Editors
3	Processing Code Comments
4	Testing
5	Project Documentation
6	Naming Conventions in Projects
7	Naming Code Entities
8	Comments Writing Strategies
9	Thread Comments in Websites
10	Development Framework for Thread Commenting
11	Database
12	Readability
13	Comments Syntax

LDA Topics: Irrelevant Topics

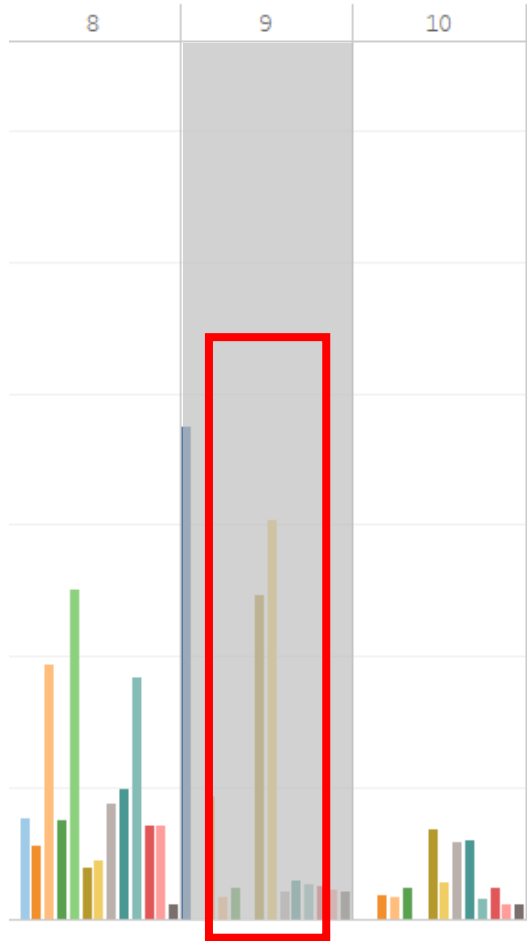
#	Topic Name
0	Exceptions
1	Documentation Generation
2	IDE & Editors
3	Processing Code Comments
4	Testing
5	Project Documentation
6	Naming Conventions in Projects
7	Naming Code Entities
8	Comments Writing Strategies
9	Thread Comments in Websites
10	Development Framework for Thread Commenting
11	Database
12	Readability
13	Comments Syntax

LDA Topics: Tag Distribution



Finding : The distribution of specific relevant tags tends to be concentrated highly in one of the topics identified by LDA, showing the specificity of the tags whereas some other relevant tags span multiple topics showing the generality of these tags. Tags spanning into several topics require manual intervention to confirm the results.

Problems with tags

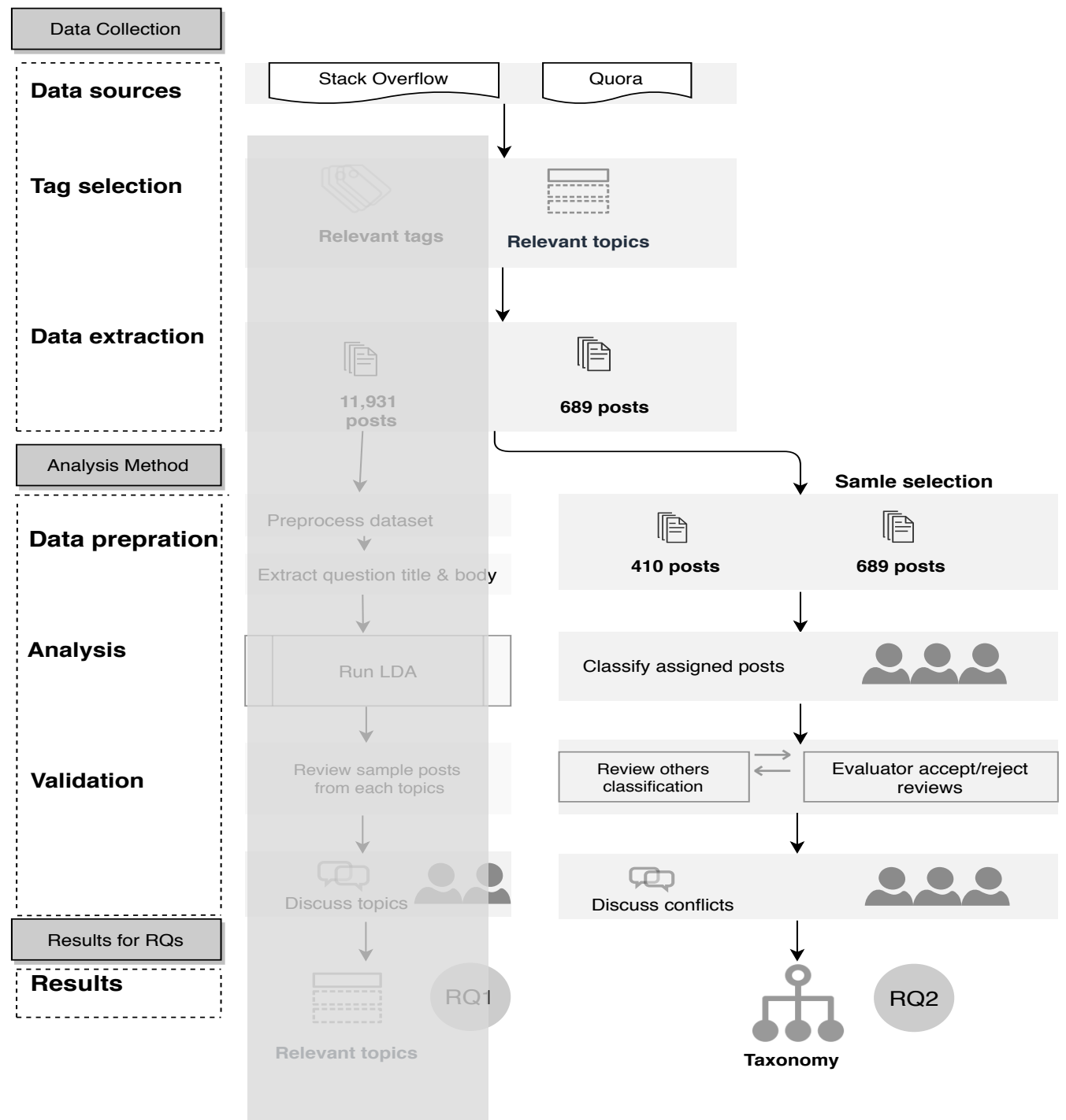


- Tags *comments* and *commenting*
- General and ambiguous
- Irrelevant despite large proportion of tags

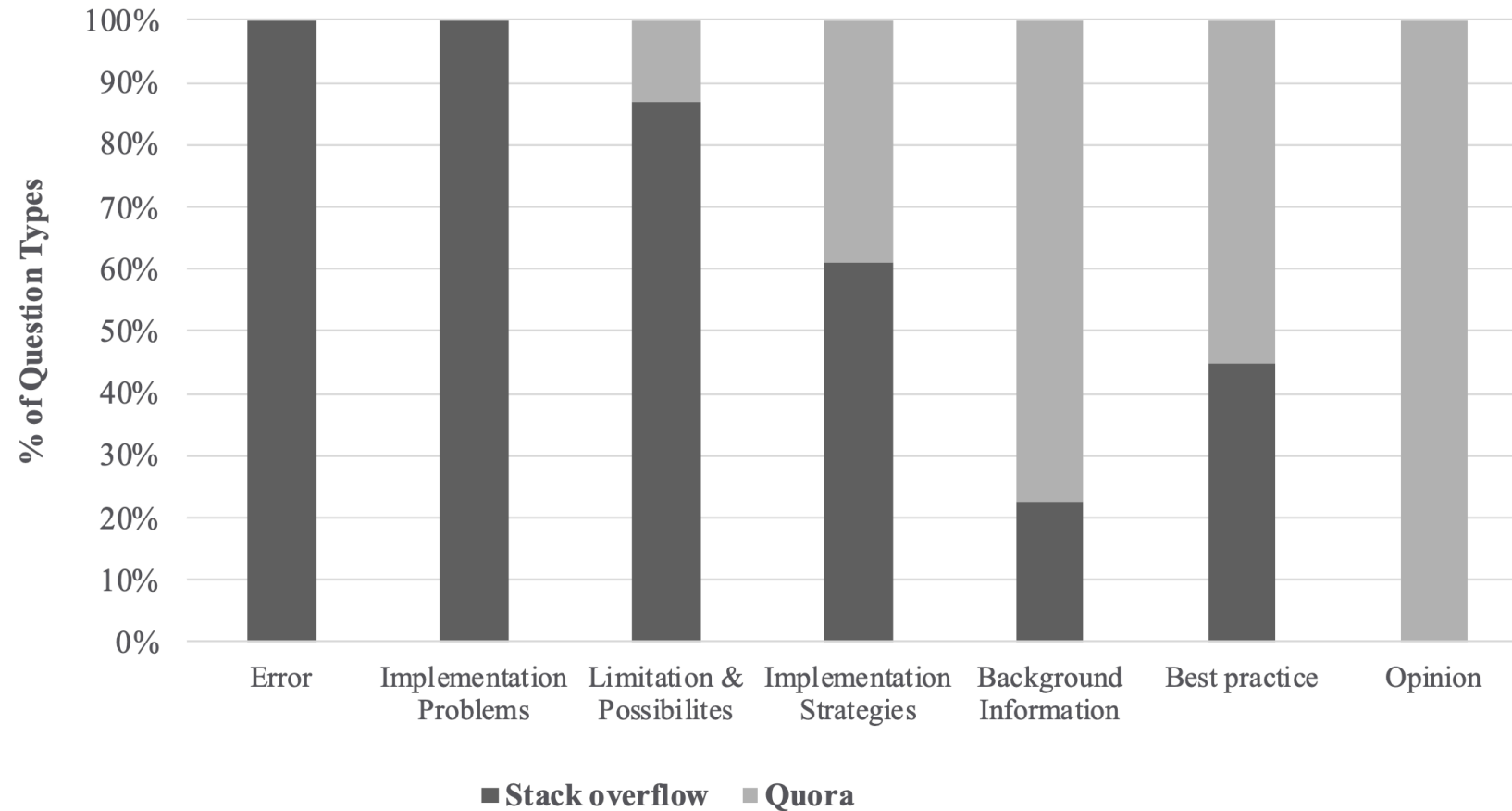
Research Questions

RQ2: What types of questions and problems developers discuss on various platforms?

RQ2: Manual Analysis

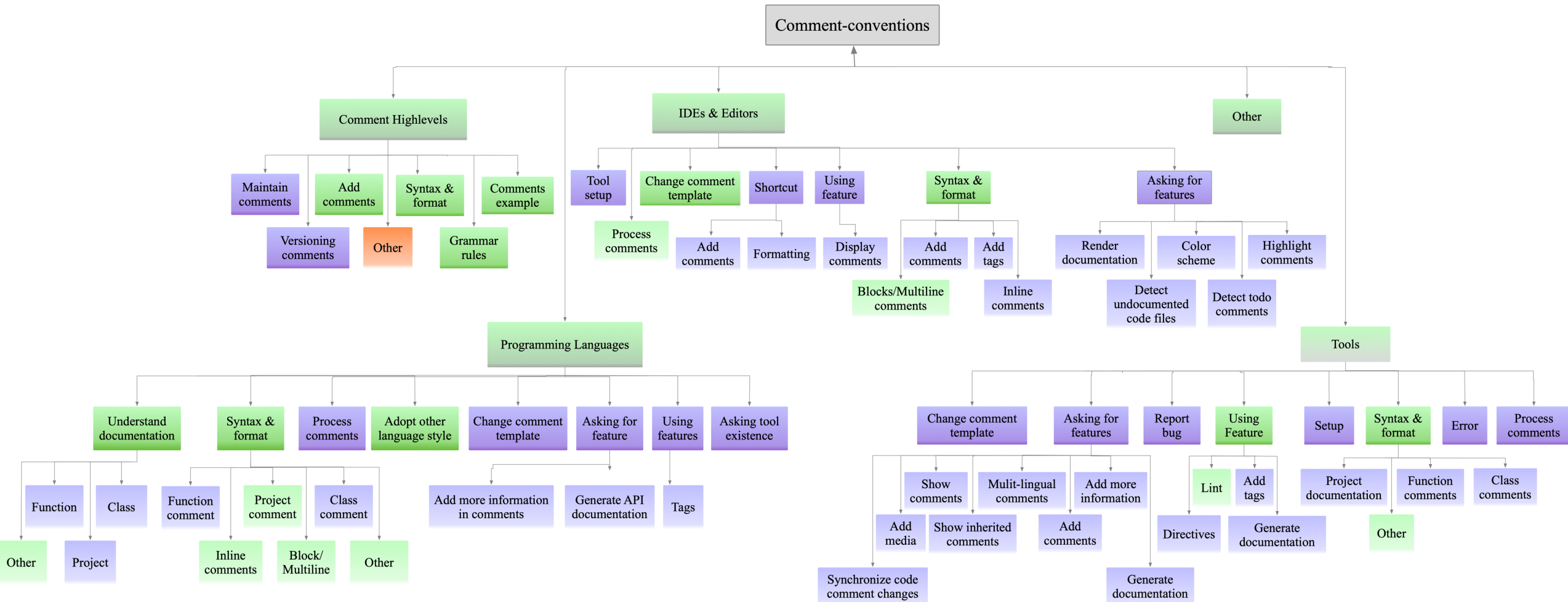


Question Type on Quora vs. Stack Overflow



Finding: Different kinds of questions are prevalent on Stack Overflow and Quora. Developers ask questions about implementation strategies and implementation problem more on Stack Overflow compared to Quora. On the other hand, Quora also observed questions about commenting on best practices and background information apart from opinion-based questions.

Taxonomy: Which types of information developers seek?



Taxonomy: Most discussed categories

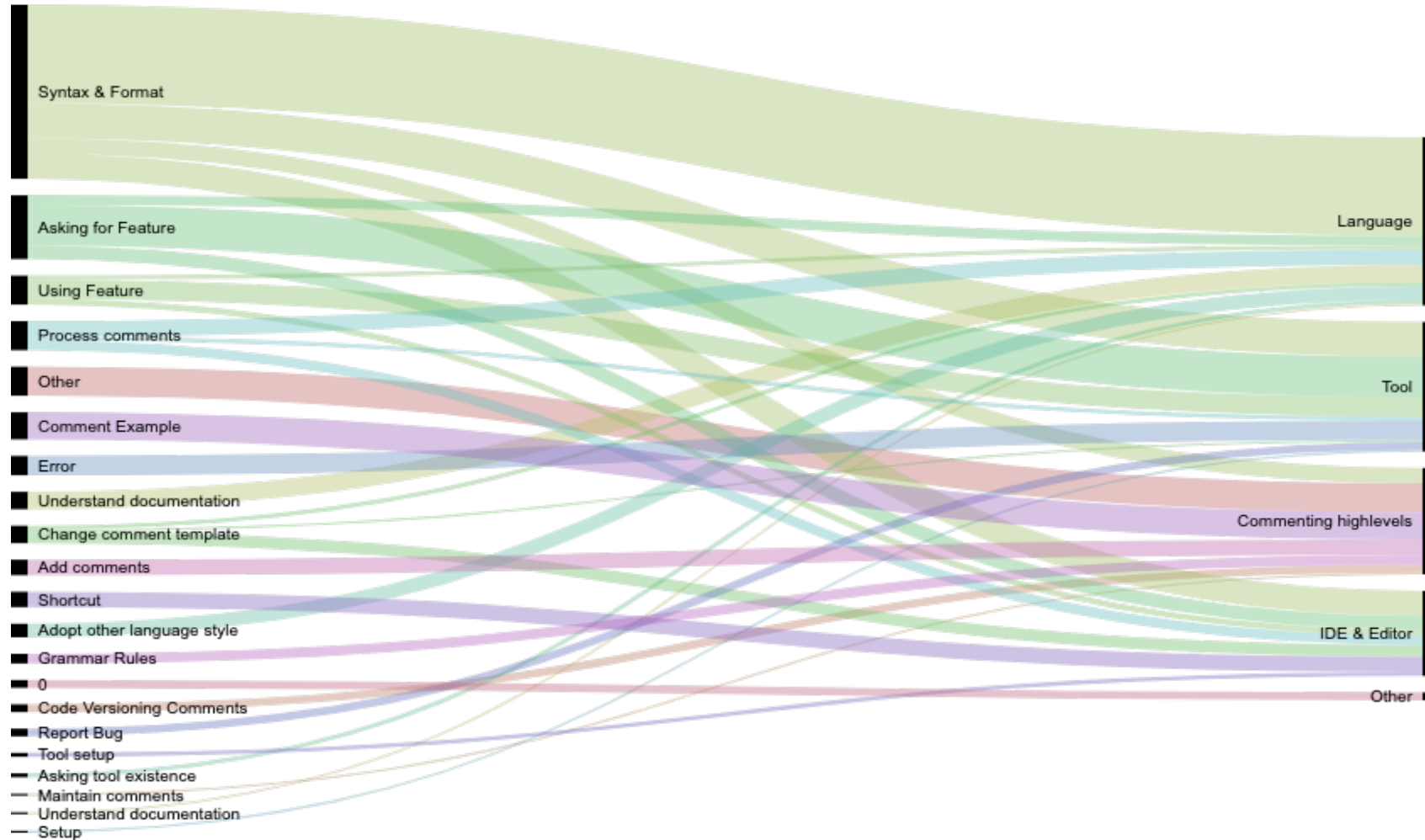
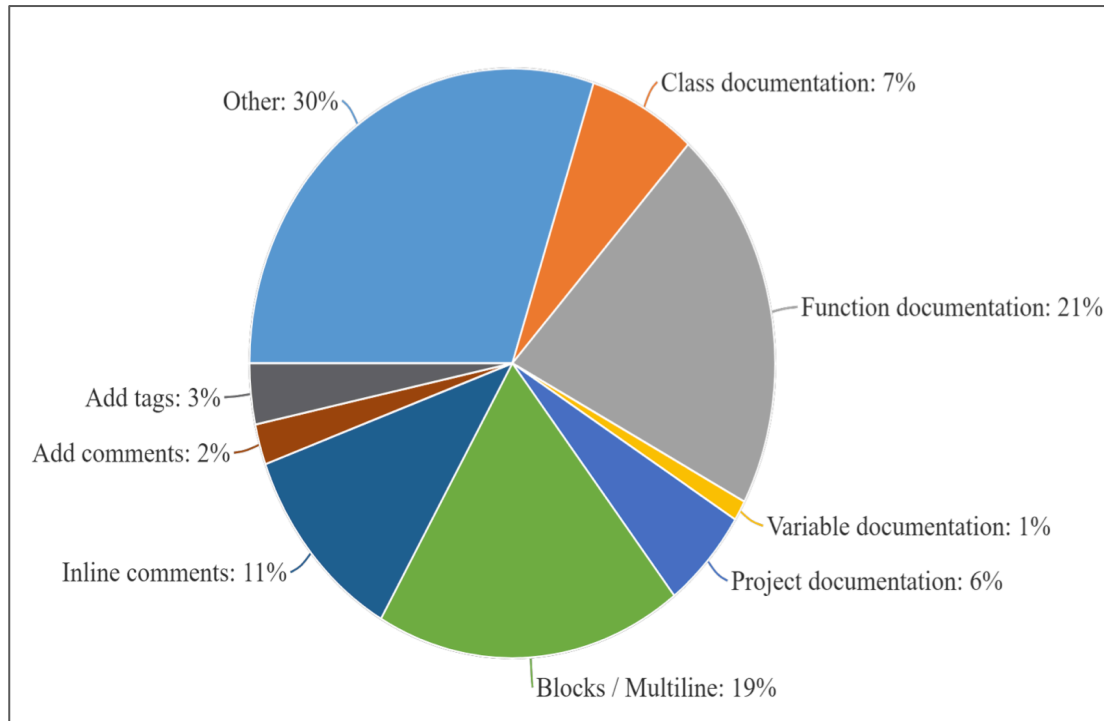


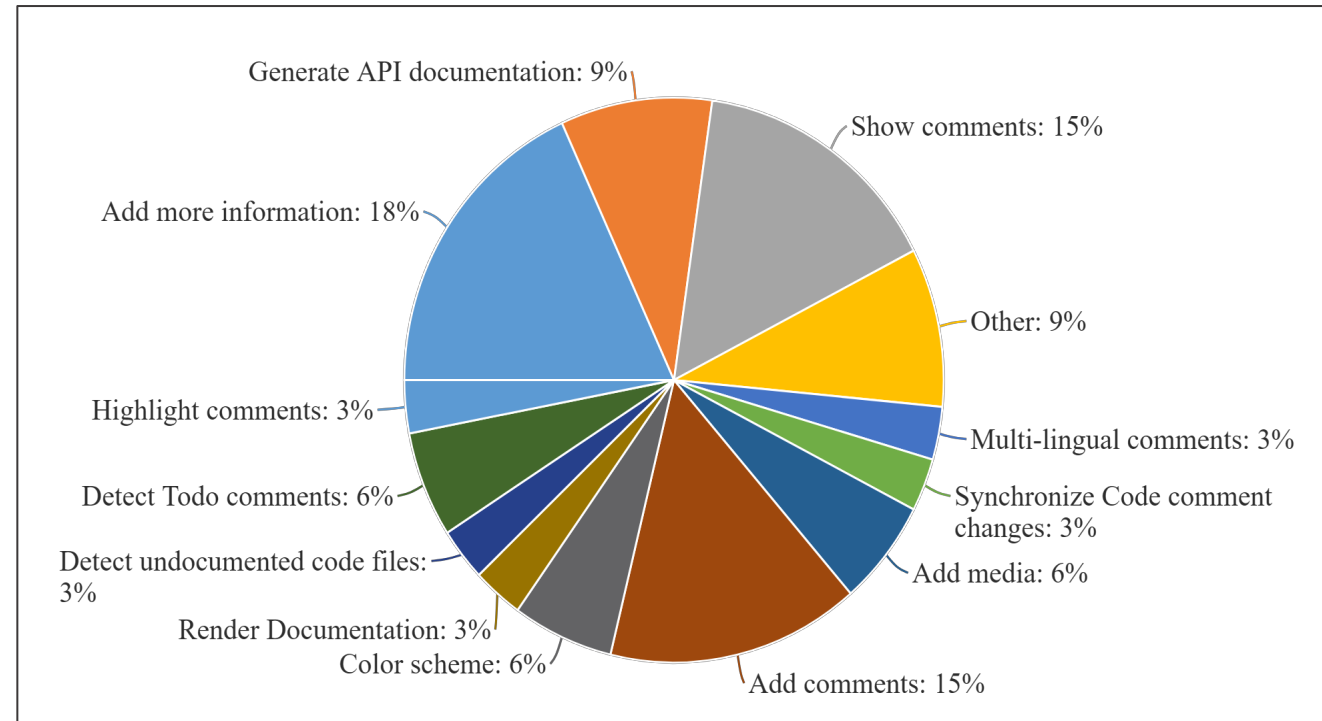
Fig: Second level Categories ordered according to their frequency

Finding: Developers often ask about the syntax used to write function (method) comments compared to other kinds of comments (class, variable, package). It shows the increasing efforts and trend towards API documentation. Another frequently asked feature is the conventions to add different kinds of information such as code examples, media, or custom tags in the code comments on Stack Overflow

Top two discussed categories



Syntax & Format discussions



Asking for Features discussions

Finding: Apart from syntax and features related discussions, developers ask questions about adopting commenting style from other programming languages, modifying comment templates, understanding code comments, and processing comments for various purposes

Code Comments Conventions – Challenges

- **Generality** and **ambiguity** of topic keywords
- Selection of **relevant tags**
- Selection of **relevant posts**
- **Conclusion:** Very hard to fully automate extraction and classification of “clean” dataset about *Code Comment Conventions*

Implication

Findings relevant for developers and researchers



Implication: Developers



Writing and checking syntax of comments



Organizing information in comments



Consistency in writing style of comments

Implication: Researchers



Need to survey code comment tools
(style checkers)

We present an initial picture of such a work.



Need to assess the relative
importance of comment conventions

We gathered conventions suggested by experts on Stack
overflow and Quora

We categorized the comment conventions provided by
experts.



Preventing duplication of comment
content

We present information need gaps identified by analyzing
developers questions.

Implication: Tool support



Need of automated style checkers

Some languages do not support automated style checkers



Hybrid style checkers

With the use of multiple programming languages in open source projects



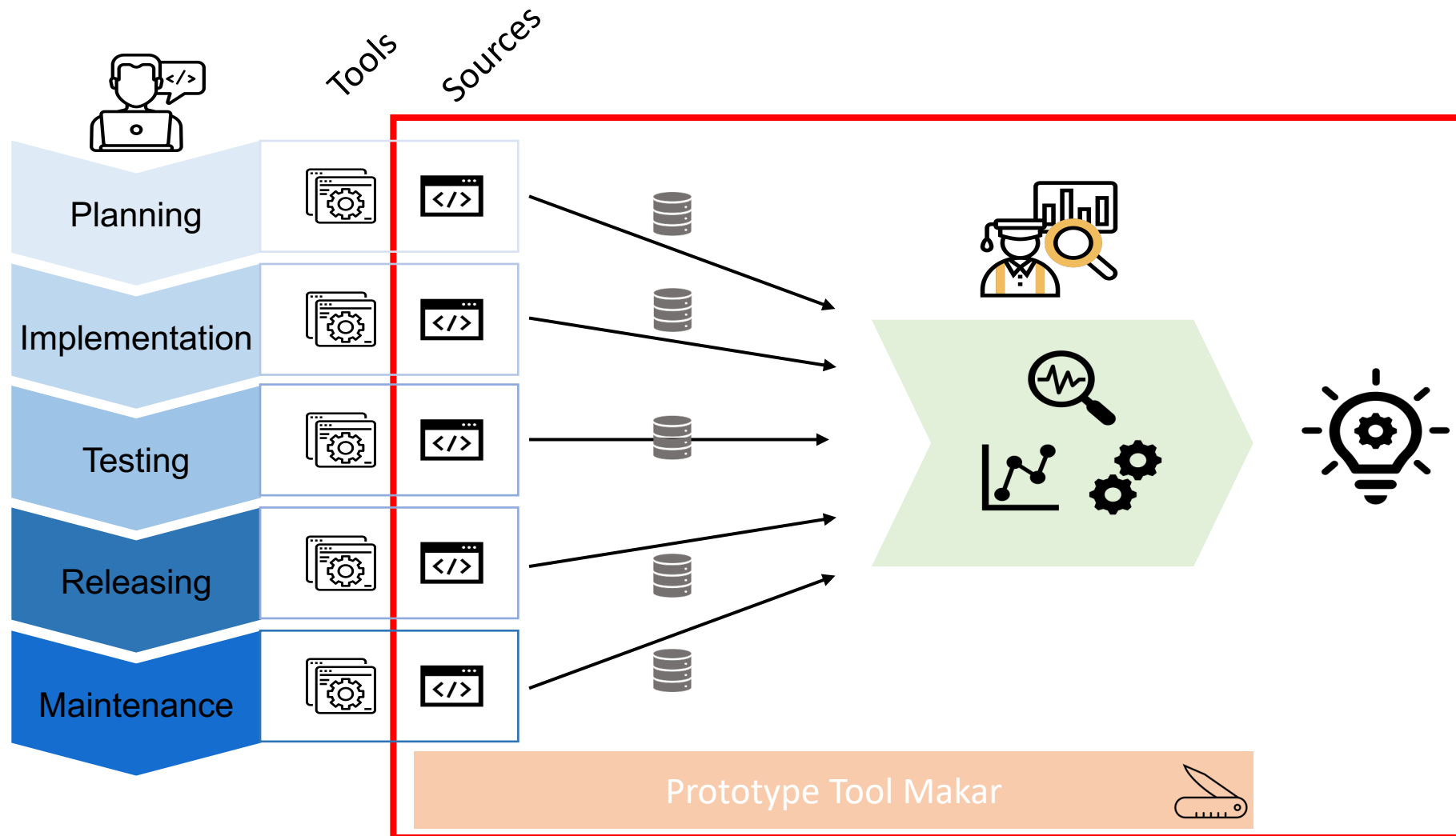
Tools to assess comment quality

Developers look for automated tools to assess the quality of their comments



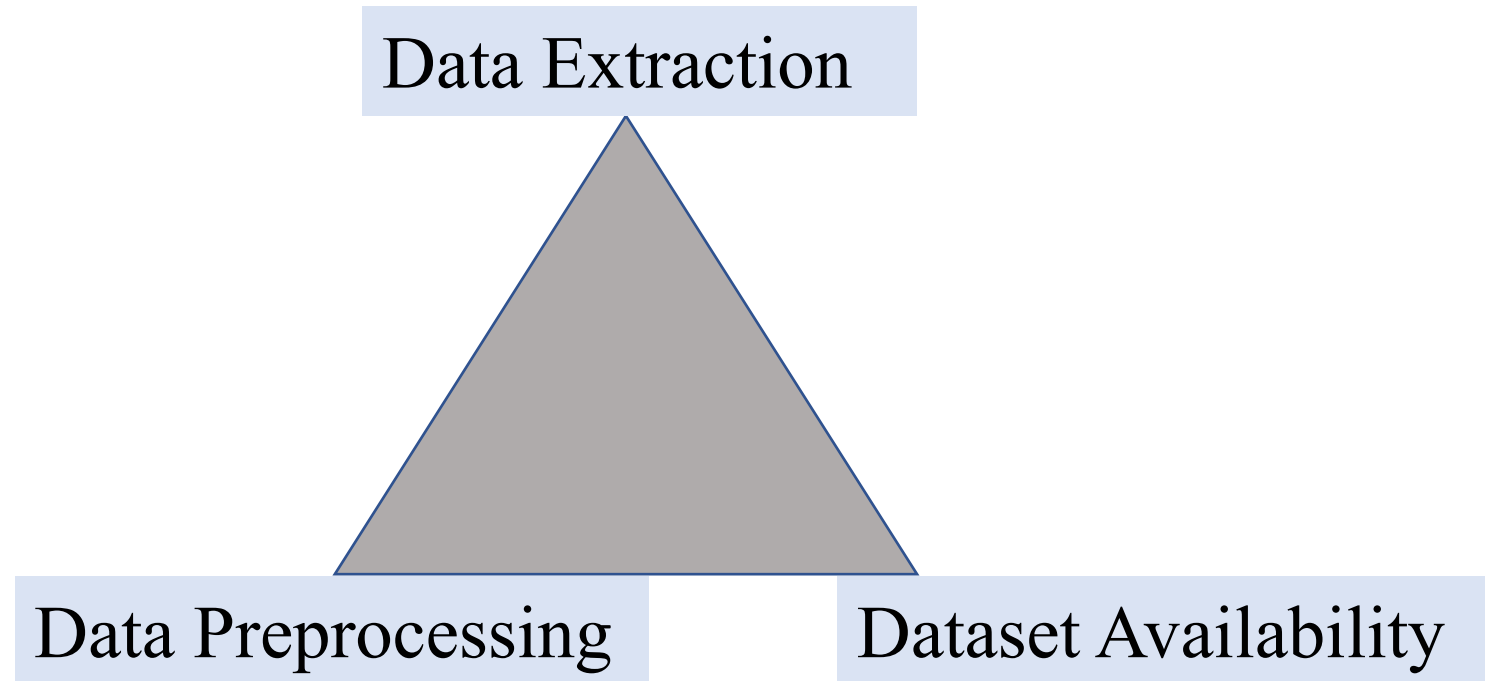
Ease of finding commenting guidelines

Developers face problems in locating conventions

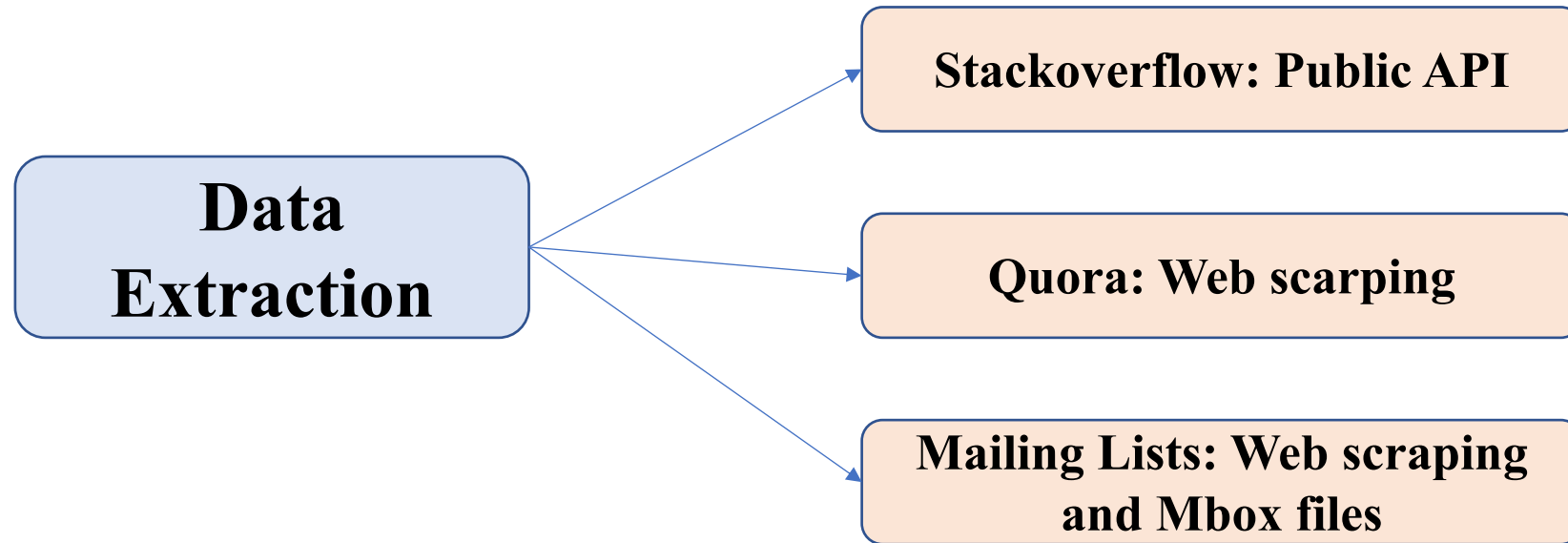


Our Study Reproducibility

Our Study Reproducibility: Parameters



Our Study Reproducibility: Data Extraction

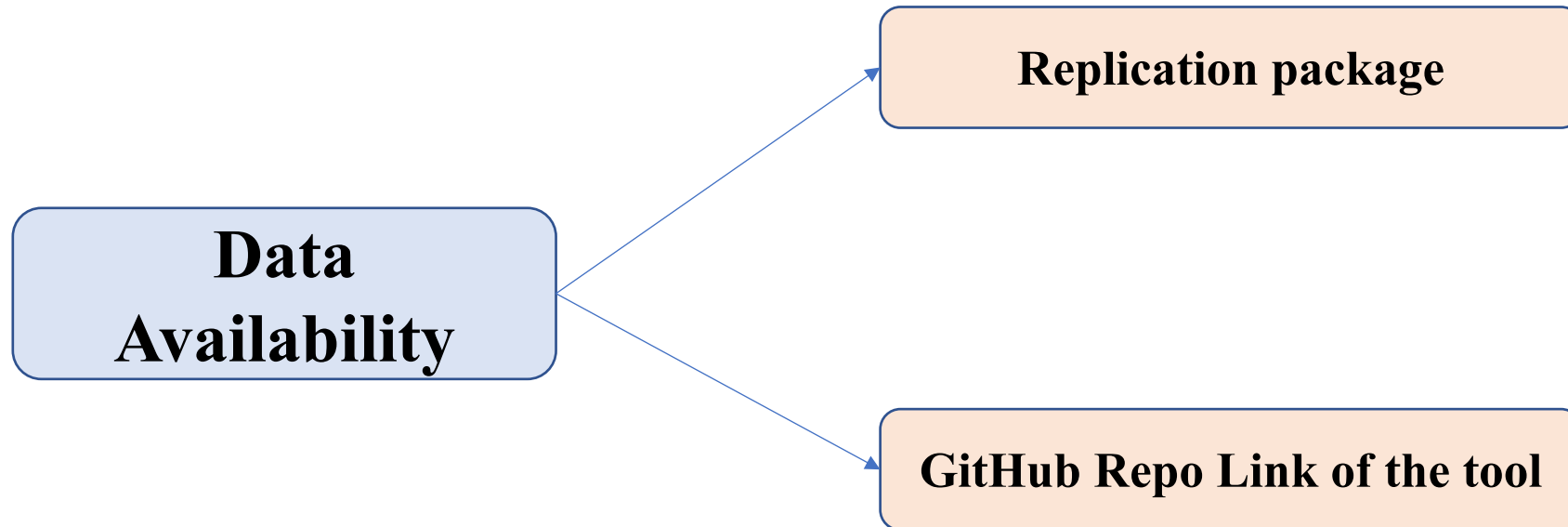


Our Study Reproducibility: Data Pre-processing



Transformation	<i>extract_code</i>	<i>strip_html</i>	<i>string_replace</i>	<i>remove_stopwords</i>	<i>word_stemming</i>
Attributes	- Question Body	- Question Body	- Question Body - Question Title	- Question Body - Question Title	- Question Body - Question Title

Our Study Reproducibility: Data Availability



Main Contributions

- An empirical investigation and analysis of comment convention related questions on different development social media
- An empirically validated taxonomy of code comment convention related questions
- An empirical and qualitative comparison of questions extracted from various sources
- A discussion of the challenges concerning the semi-automated extraction of relevant discussions from different sources
- A discussion about the potential gaps of available tools used to support code commenting practices
- A publicly available dataset including all validated data in the replication package

Future Work

Research on Developers' Information Needs

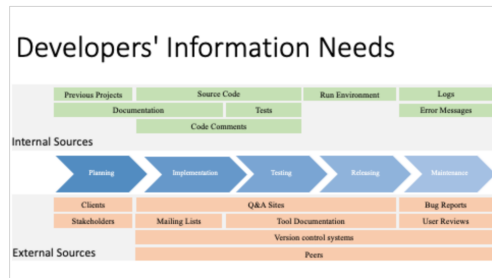
- Focus on multi-source studies
- Ease of research workflow

Code Comment Conventions

- Investigate more sources (e.g. GitHub, Jira)
- Gather comment conventions from various sources and assess their importance
- Survey developers to know which concerns are more important than others
- Gather comment conventions supported by style guidelines and by style checkers and compare them.



Summary



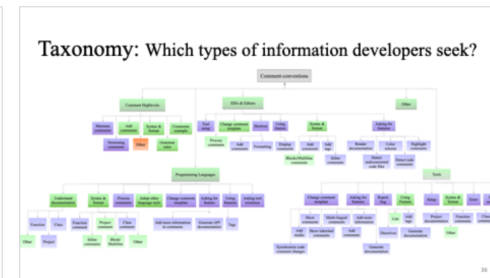
Study Goal

What Developers discuss about "Code Comment Conventions" on Social Media?



LDA Topics: Irrelevant Topics

Topic	Keywords
1. Topic 1	Keywords
2. Topic 2	Keywords
3. Topic 3	Keywords
4. Topic 4	Keywords
5. Topic 5	Keywords
6. Topic 6	Keywords
7. Topic 7	Keywords
8. Topic 8	Keywords
9. Topic 9	Keywords
10. Topic 10	Keywords
11. Topic 11	Keywords
12. Topic 12	Keywords
13. Topic 13	Keywords
14. Topic 14	Keywords
15. Topic 15	Keywords
16. Topic 16	Keywords
17. Topic 17	Keywords
18. Topic 18	Keywords
19. Topic 19	Keywords
20. Topic 20	Keywords



- ### Main Contributions
- An empirical investigation and analysis of comment-convention-related questions on different development social media
 - An empirically validated taxonomy of code comment-convention-related questions
 - An empirical and qualitative comparison of questions extracted from various sources
 - A discussion of the challenges concerning the semi-automated extraction of relevant discussions from different sources
 - A discussion about the potential gaps of available tools used to support code-commenting practices
 - A publicly available dataset including all validated data in the implication package

- ### Future Work
- Research on Developers' Information Needs**
- Focus on multi-source studies
 - Ease of research workflow
- Code Comment Conventions**
- Investigate more sources (e.g. GitHub, Jira)
 - Gather comment conventions from various sources and assess their importance
 - Survey developers to know which concerns are more important than others
 - Gather comment conventions supported by style guidelines and by style checkers and compare them