

# 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.

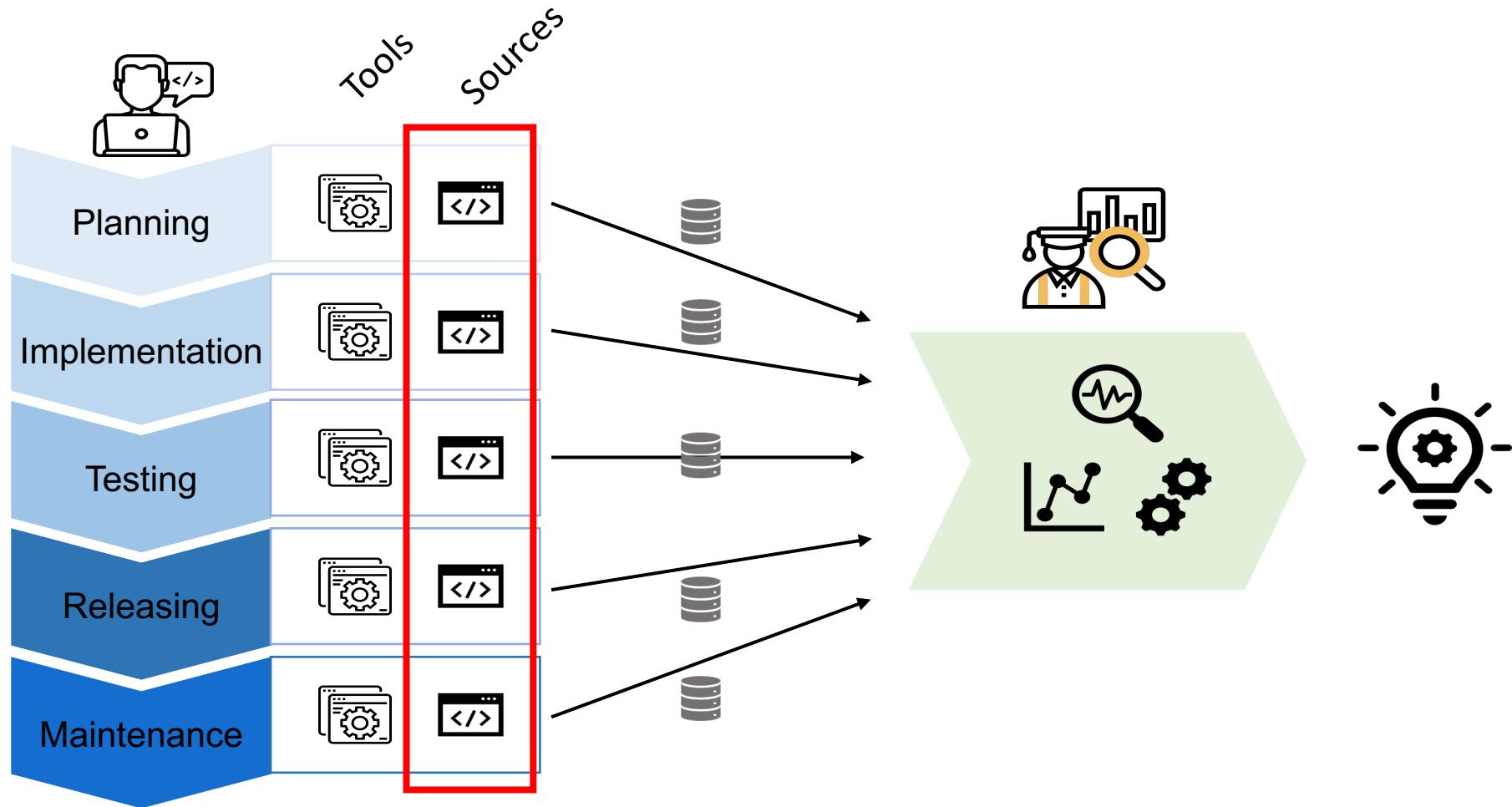
# Developers' Information Needs



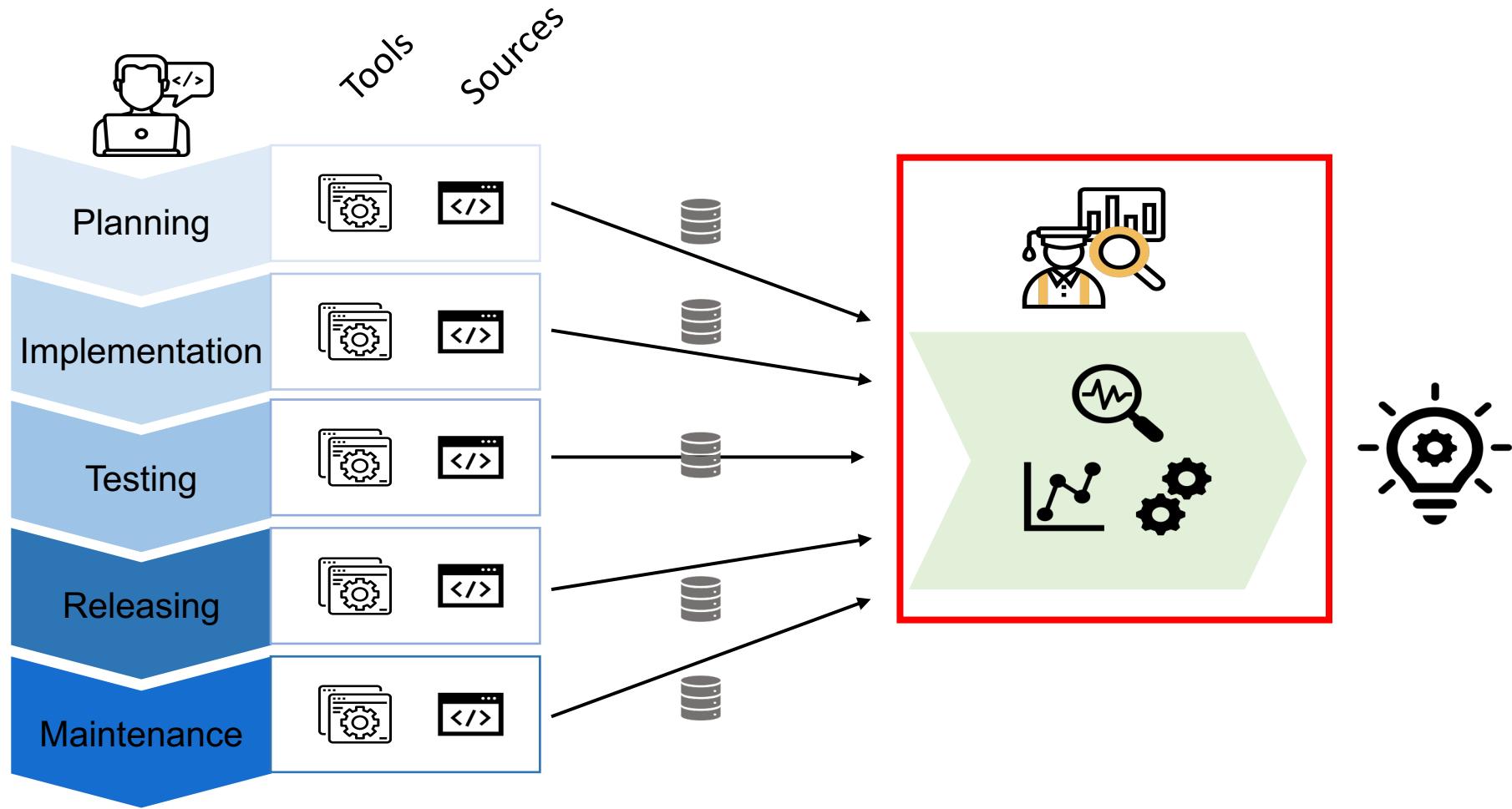
## Internal Sources



## External Sources

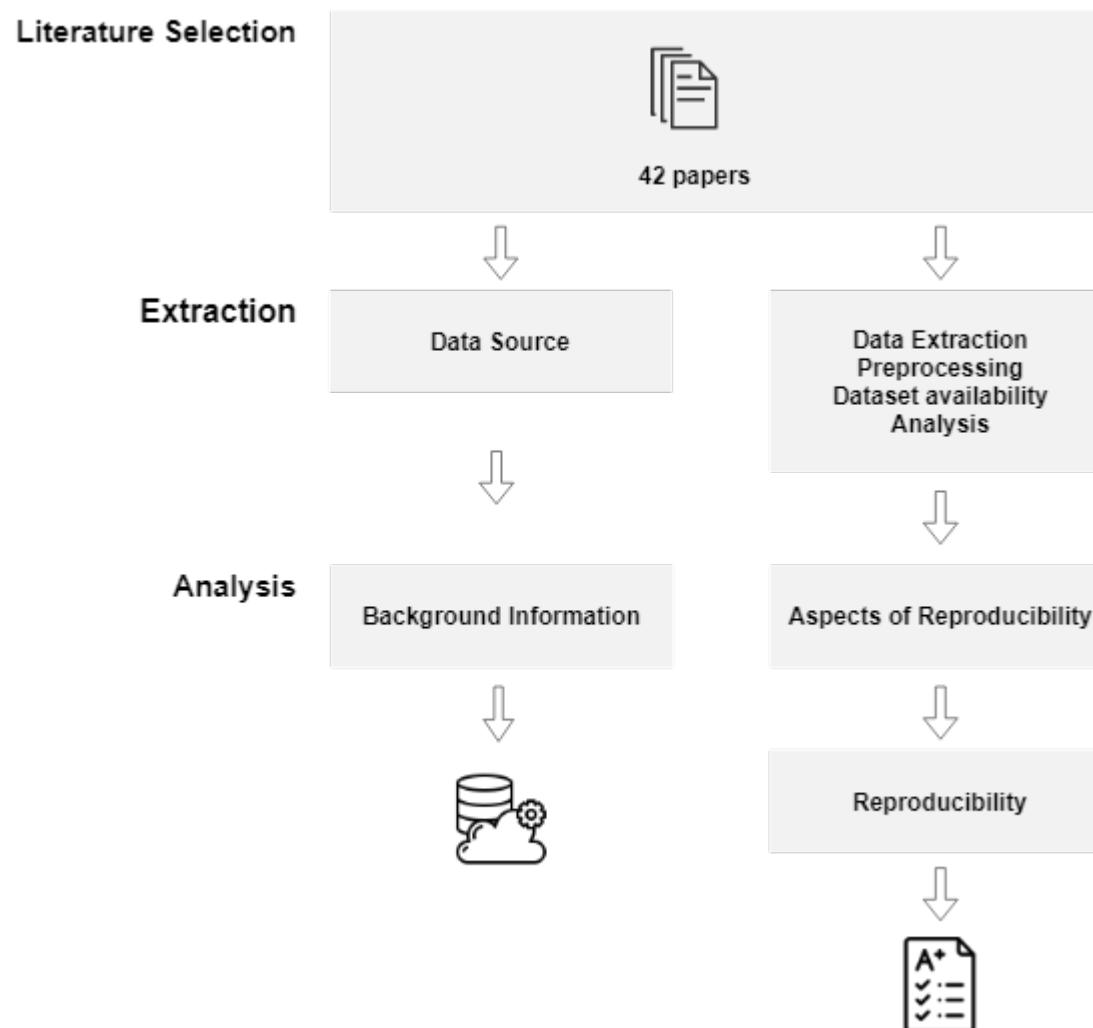


**Background Study:** Which data sources are typically analyzed by researchers to understand developers' information needs? What challenges do researchers face in conducting these studies?

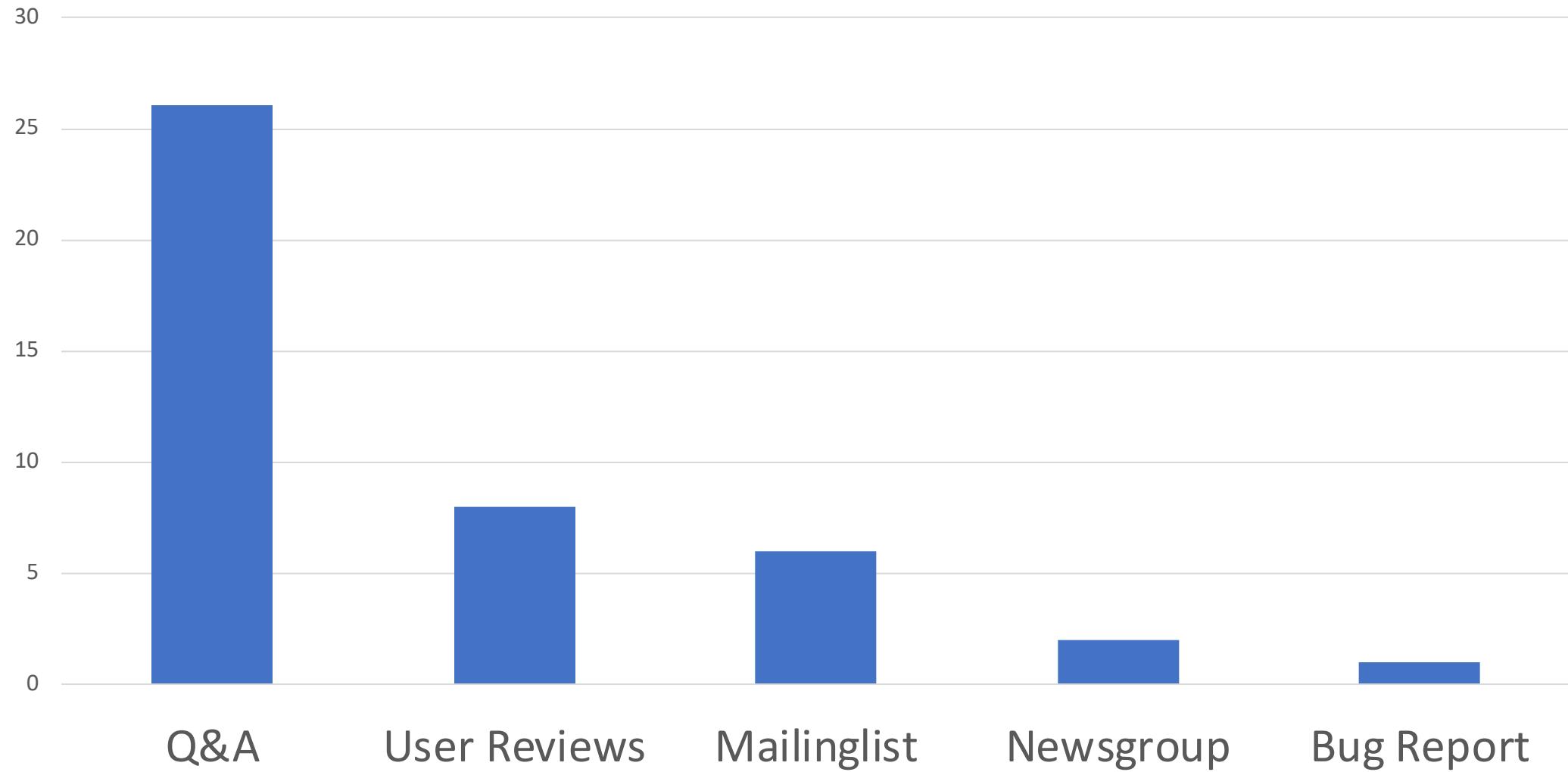


**Background Study:** What is the impact of such studies in *reproducibility*?

# 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

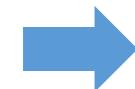


Researchers started analyzing  
these sources

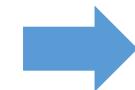
# Challenges in External Sources

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

- Contains unstructured data
  - Selecting relevant data
  - Cleaning data for the analysis
- 
- To compare different sources / communities
  - To uncover their evolution
  - To reuse datasets

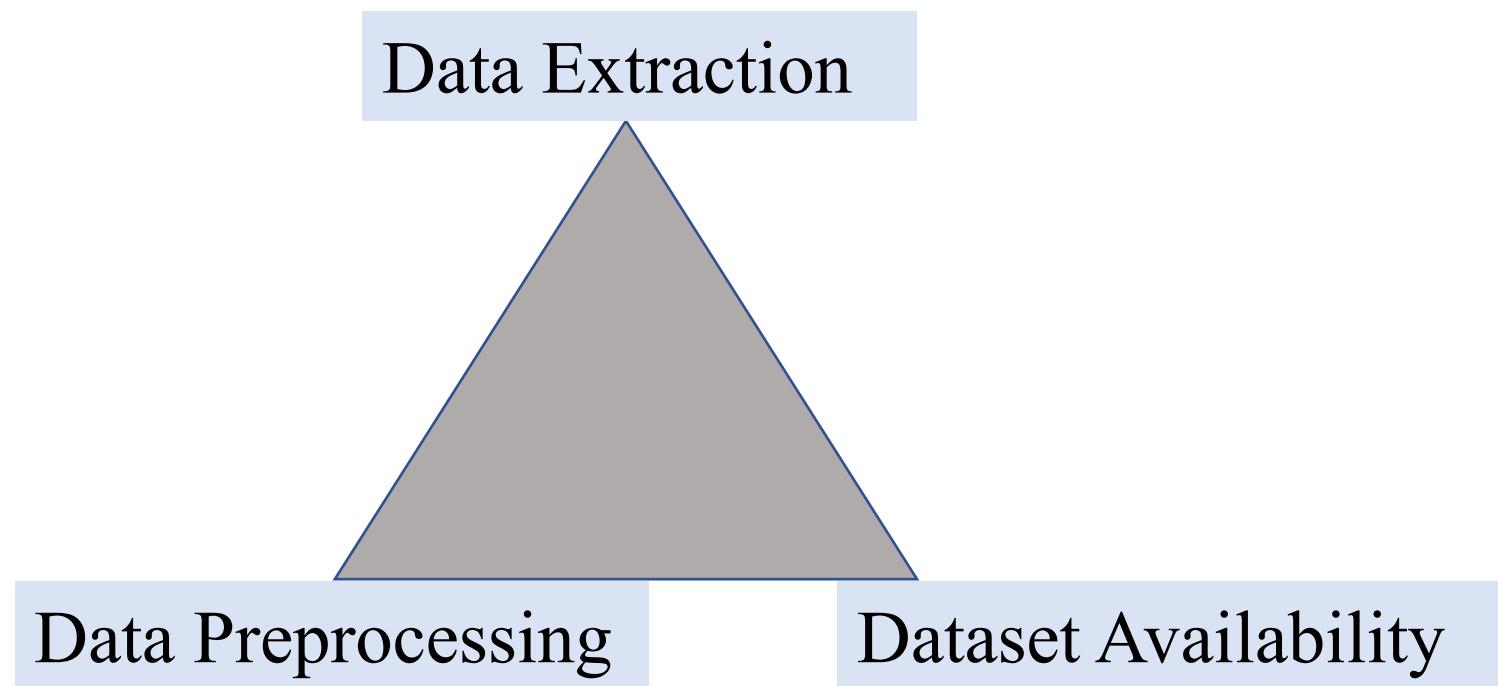


Researchers started investigating these sources

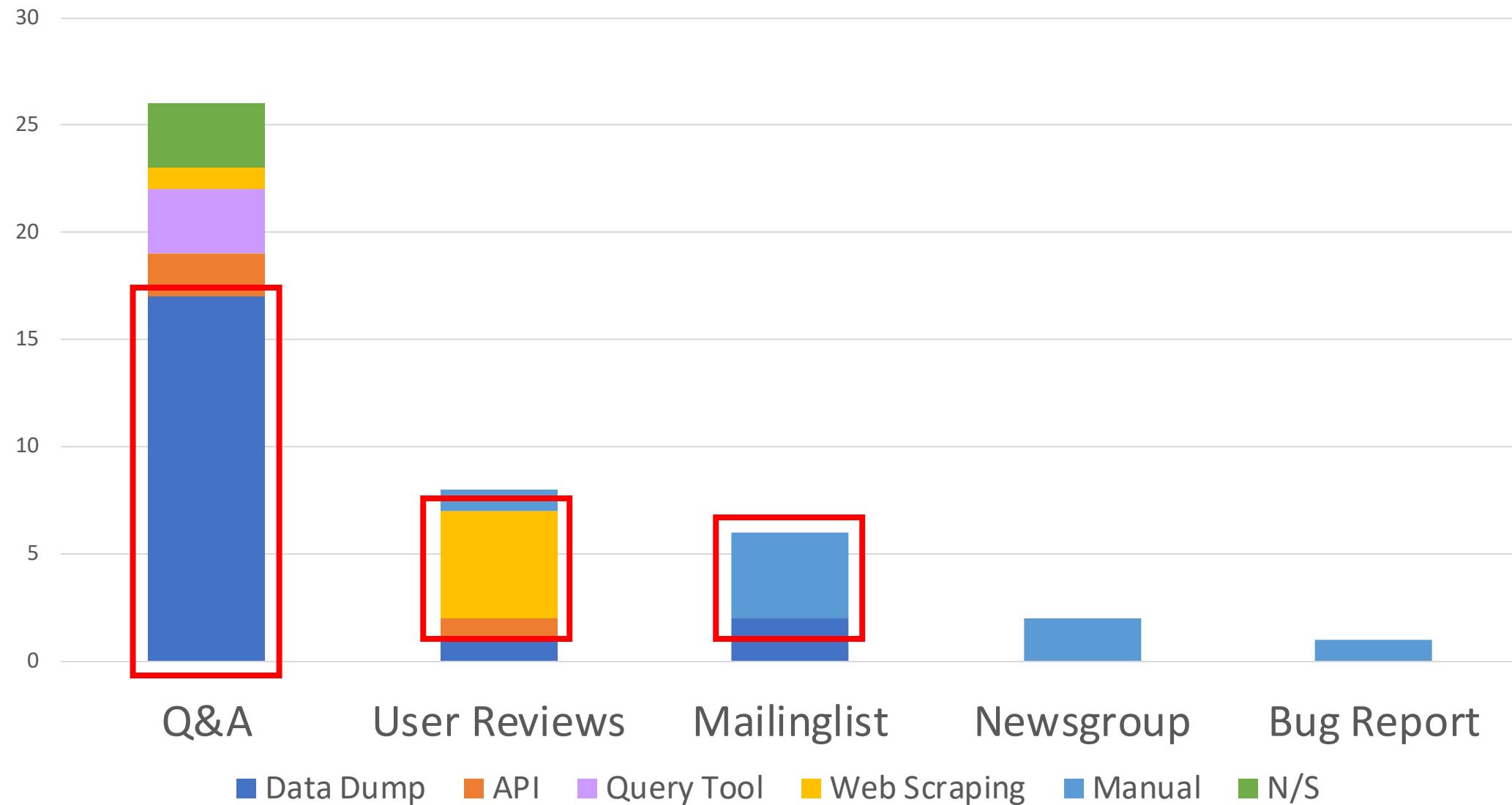


**Reproducibility** is highly important in such studies

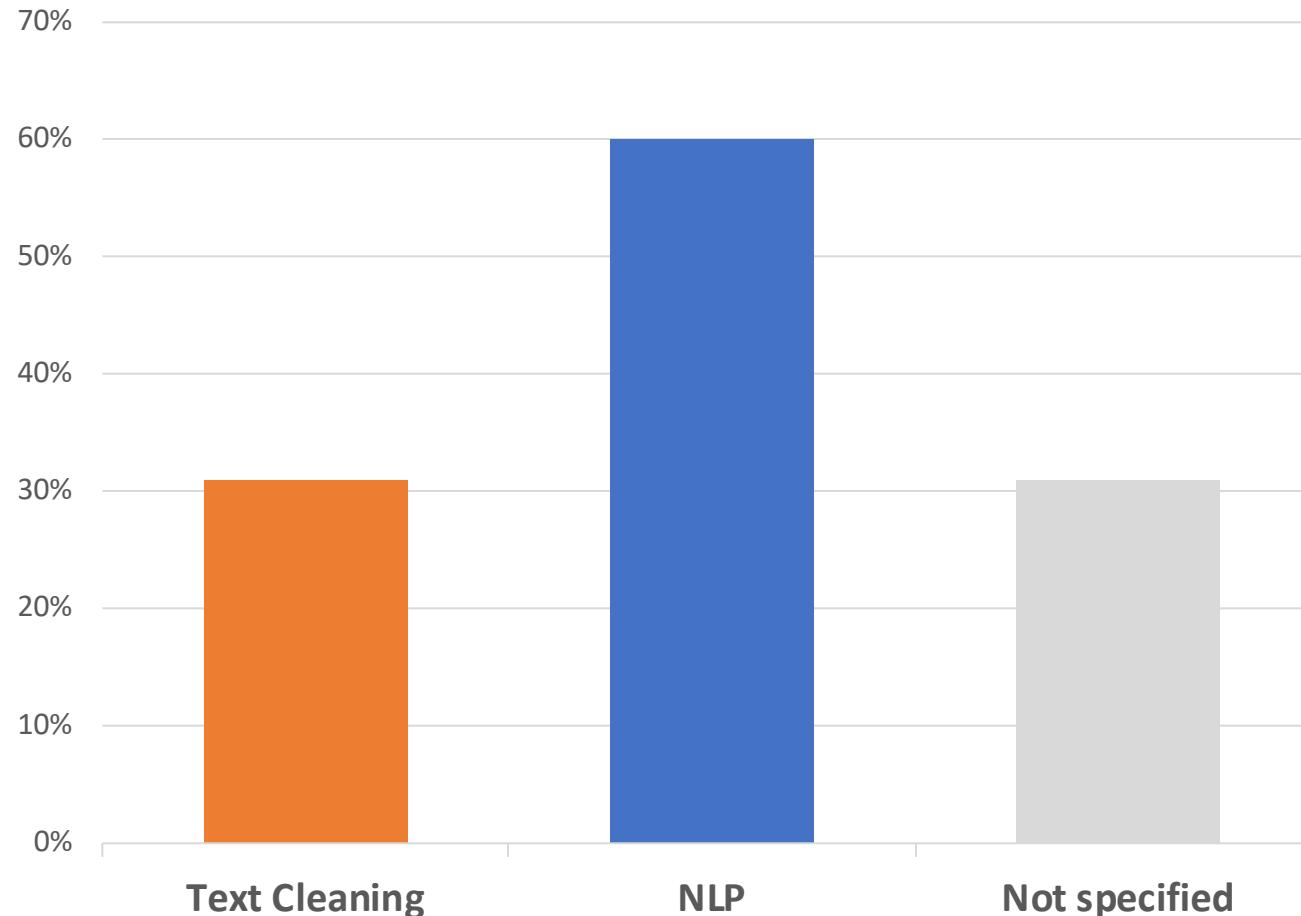
# Reproducibility Aspects



# 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**

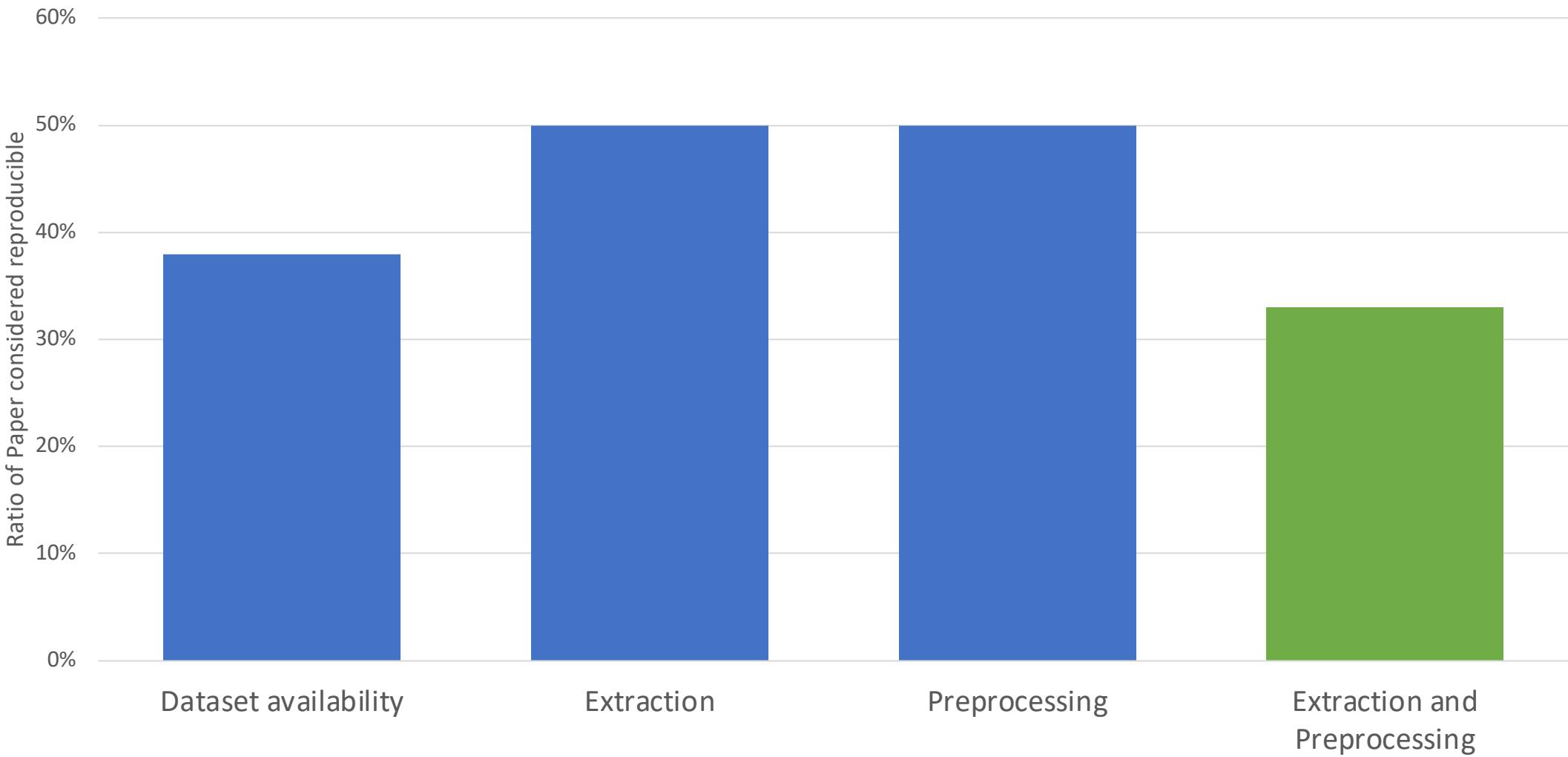
Text cleaning is highly dependent of the data source and extraction method, e.g Stack Overflow questions have HTML tags which need to be cleaned. NLP preprocessing is dependent on the analysis to be performed with the data.

# 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



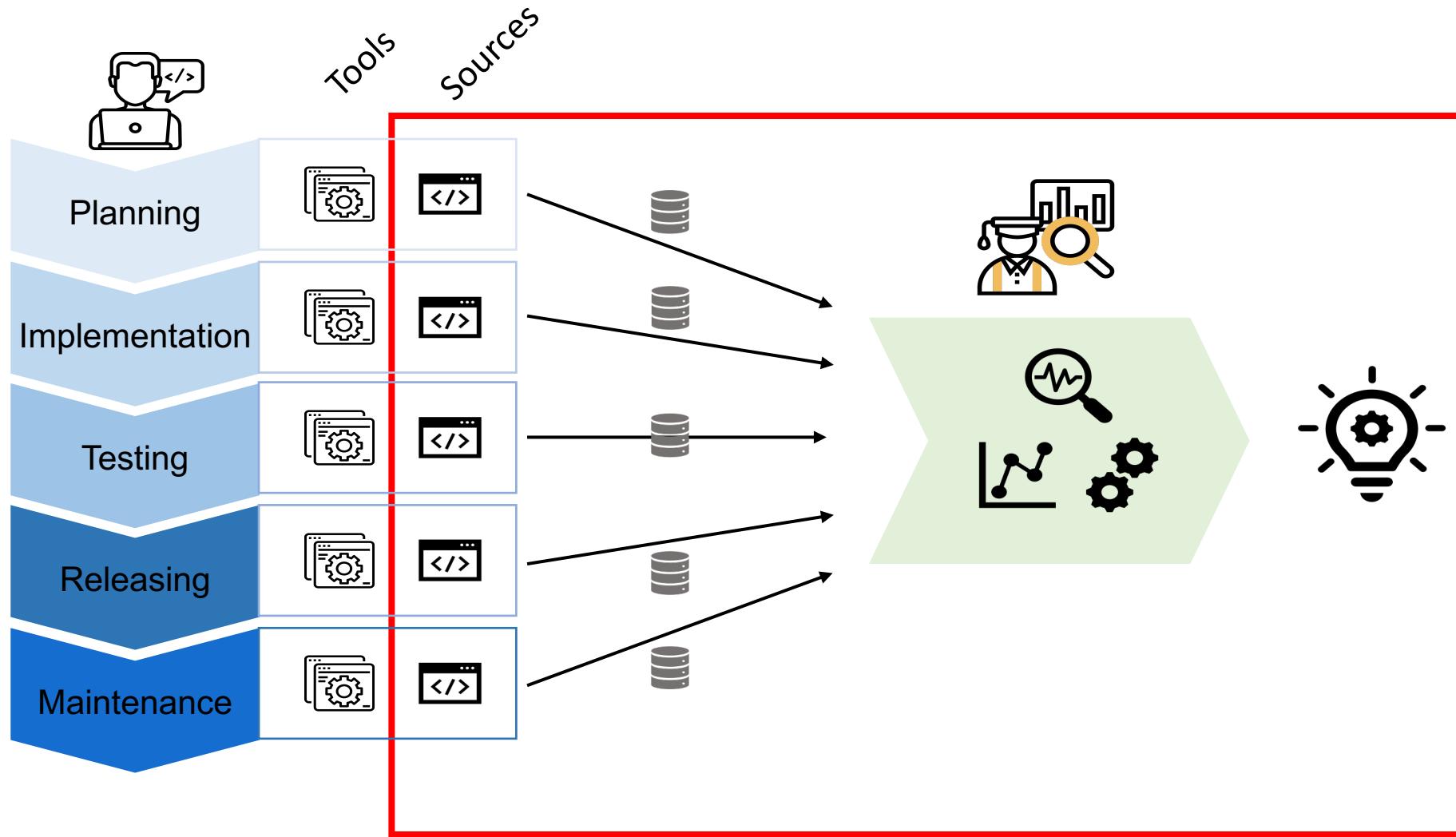
Despite low requirement of dataset availability, only 38% of the analysed studies publish a dataset for replication purposes

# 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**



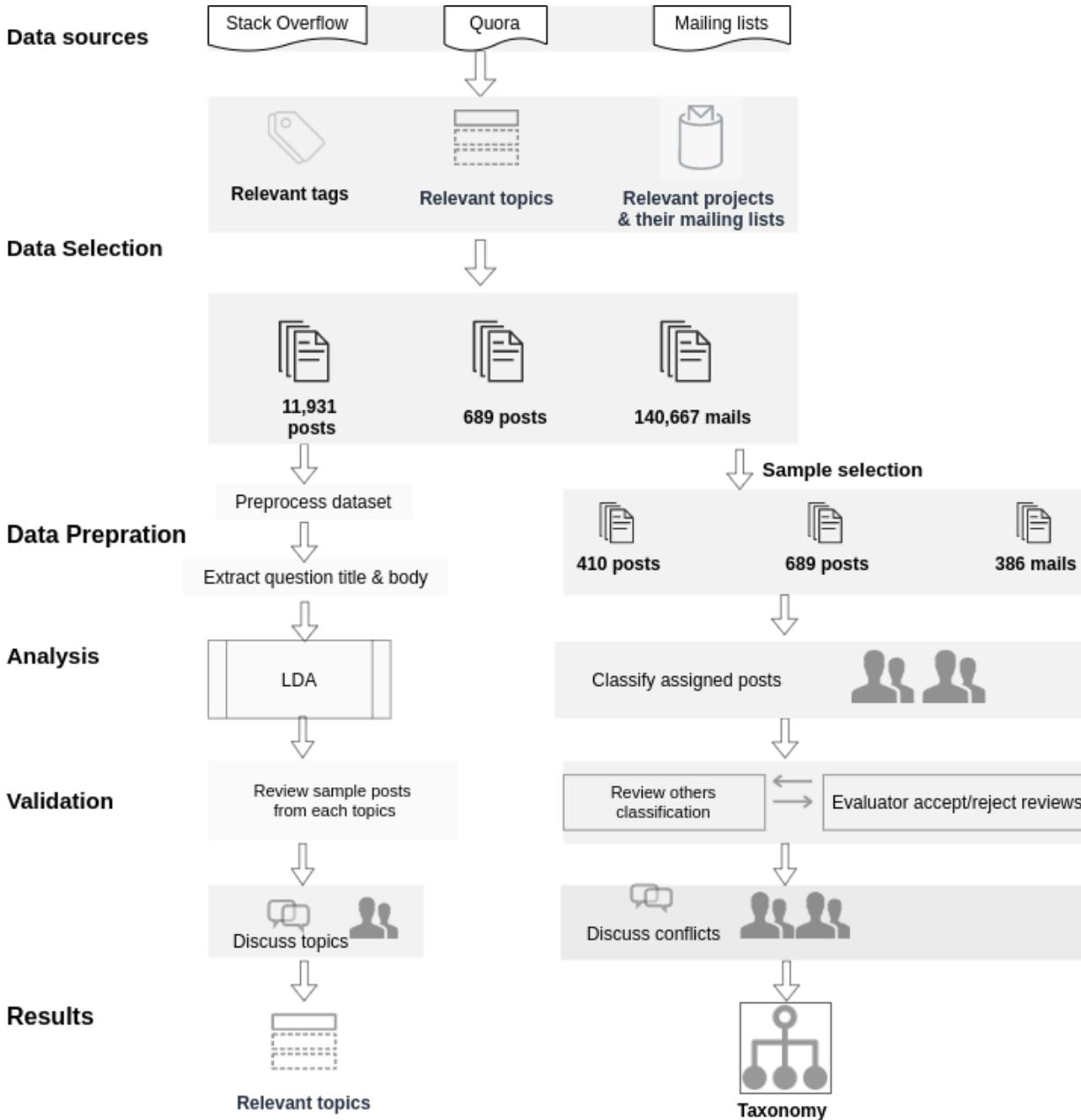
Main Study: What Developers discuss about “*Code Comment Conventions*” on Social Media?

# Research Questions

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

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

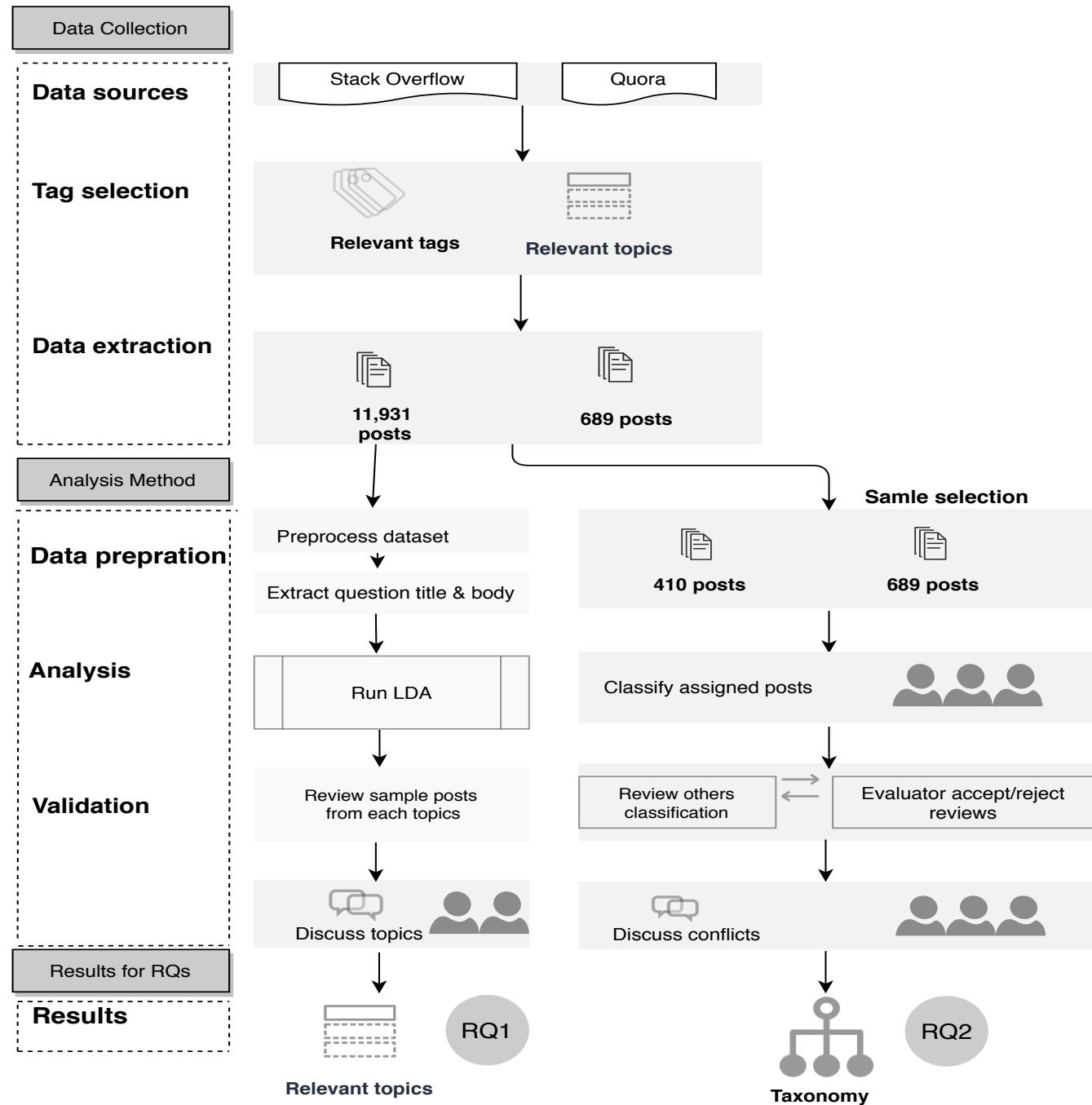
# Methodology



# 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

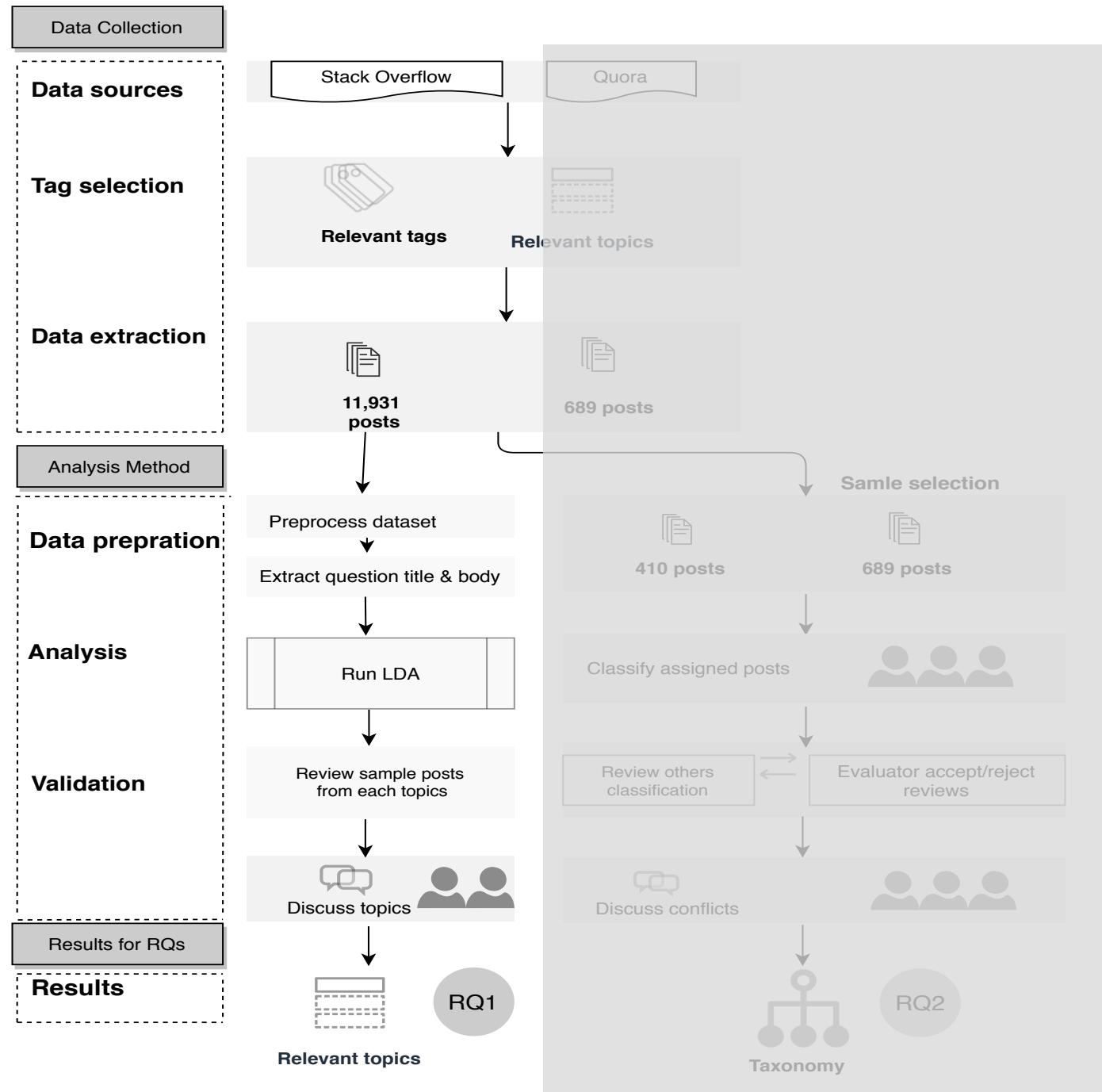
# New Methodology



# 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.

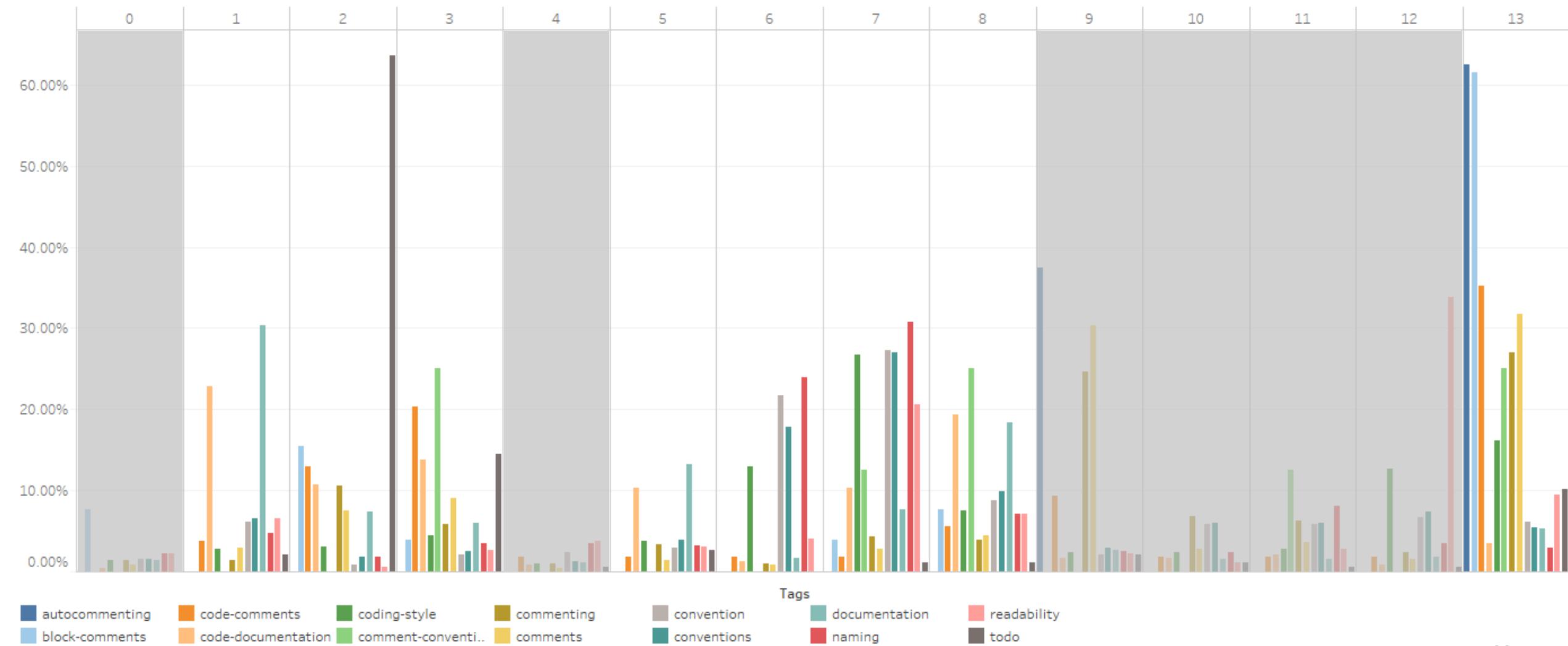
# LDA Topics

#	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

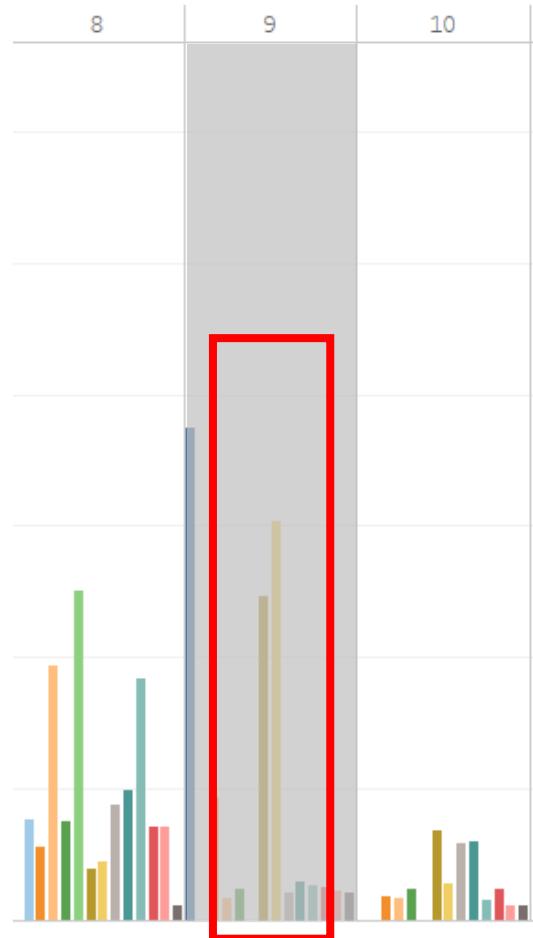
#	Topic Name
0	<b>Exceptions</b>
1	<b>Documentation Generation</b>
2	<b>IDE &amp; Editors</b>
3	<b>Processing Code Comments</b>
4	<b>Testing</b>
5	<b>Project Documentation</b>
6	<b>Naming Conventions in Projects</b>
7	<b>Naming Code Entities</b>
8	<b>Comments Writing Strategies</b>
9	<b>Thread Comments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

# 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

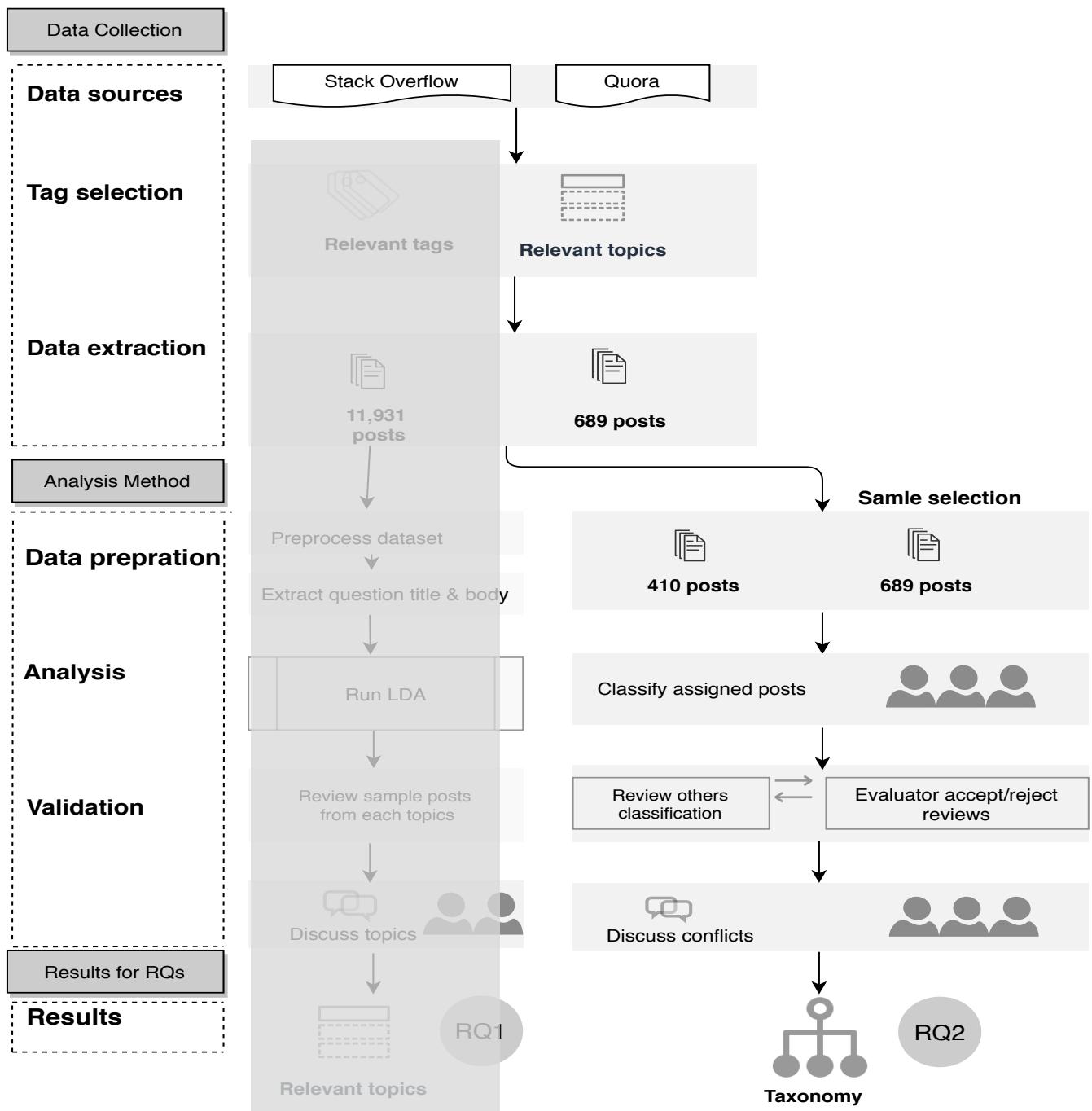


- Tags *comments* and *commenting*
- General and ambiguous
- Irrelevant despite large proportion of tags

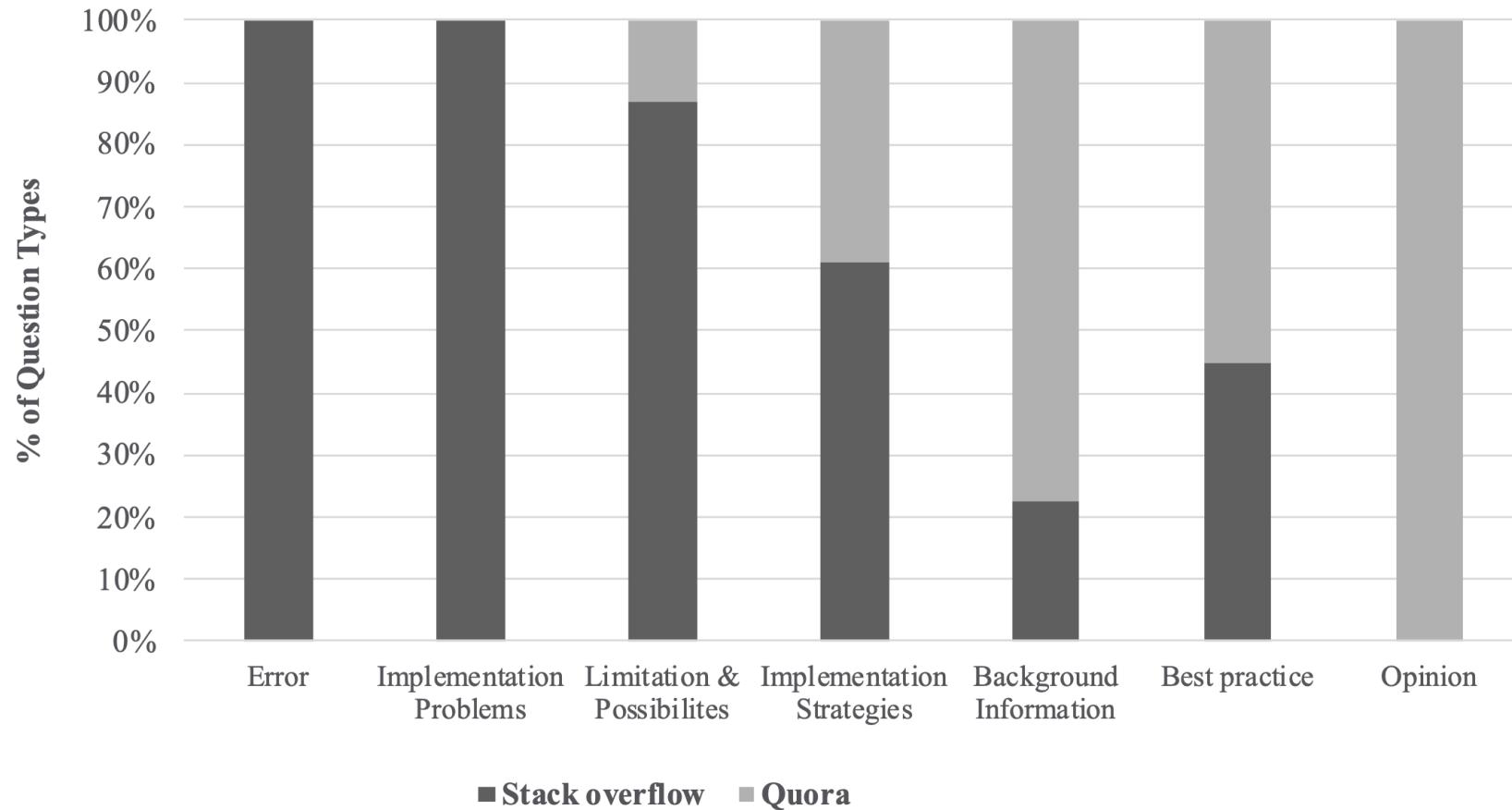
# Research Questions

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

# RQ2: Manual Analysis

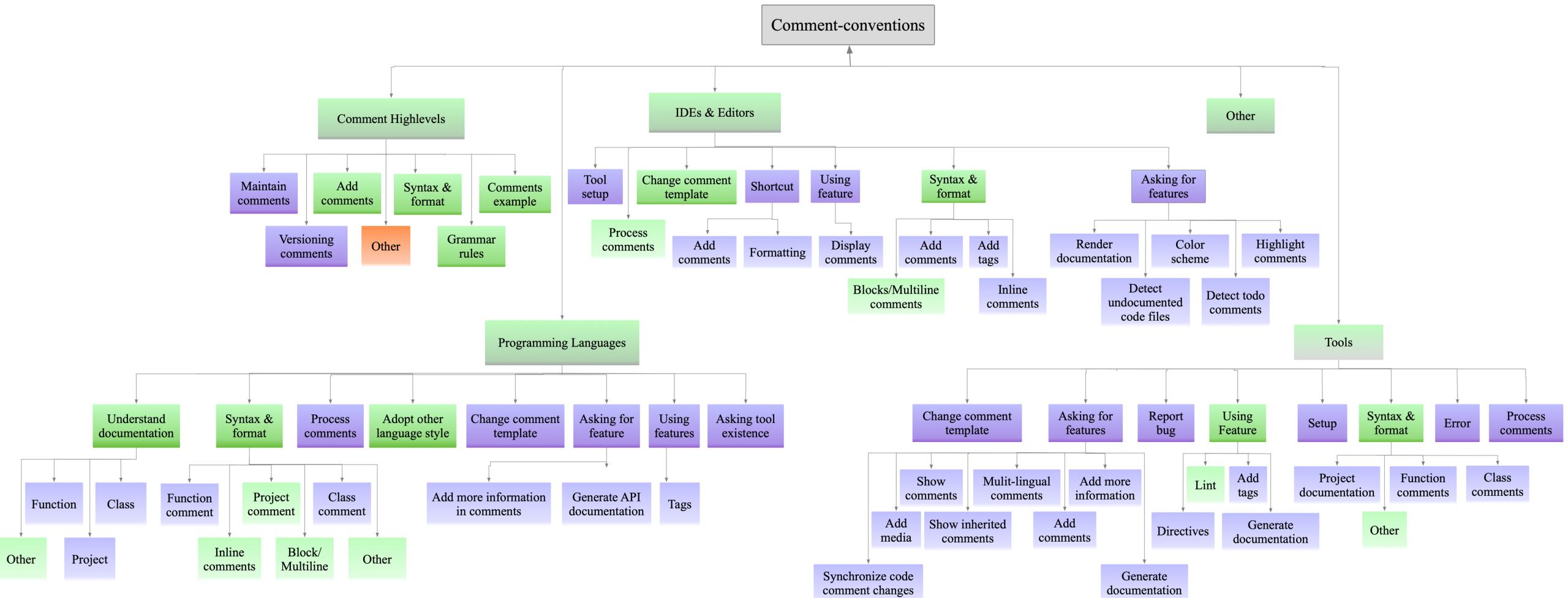


# Question Type on Quora vs. Stack Overflow



**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

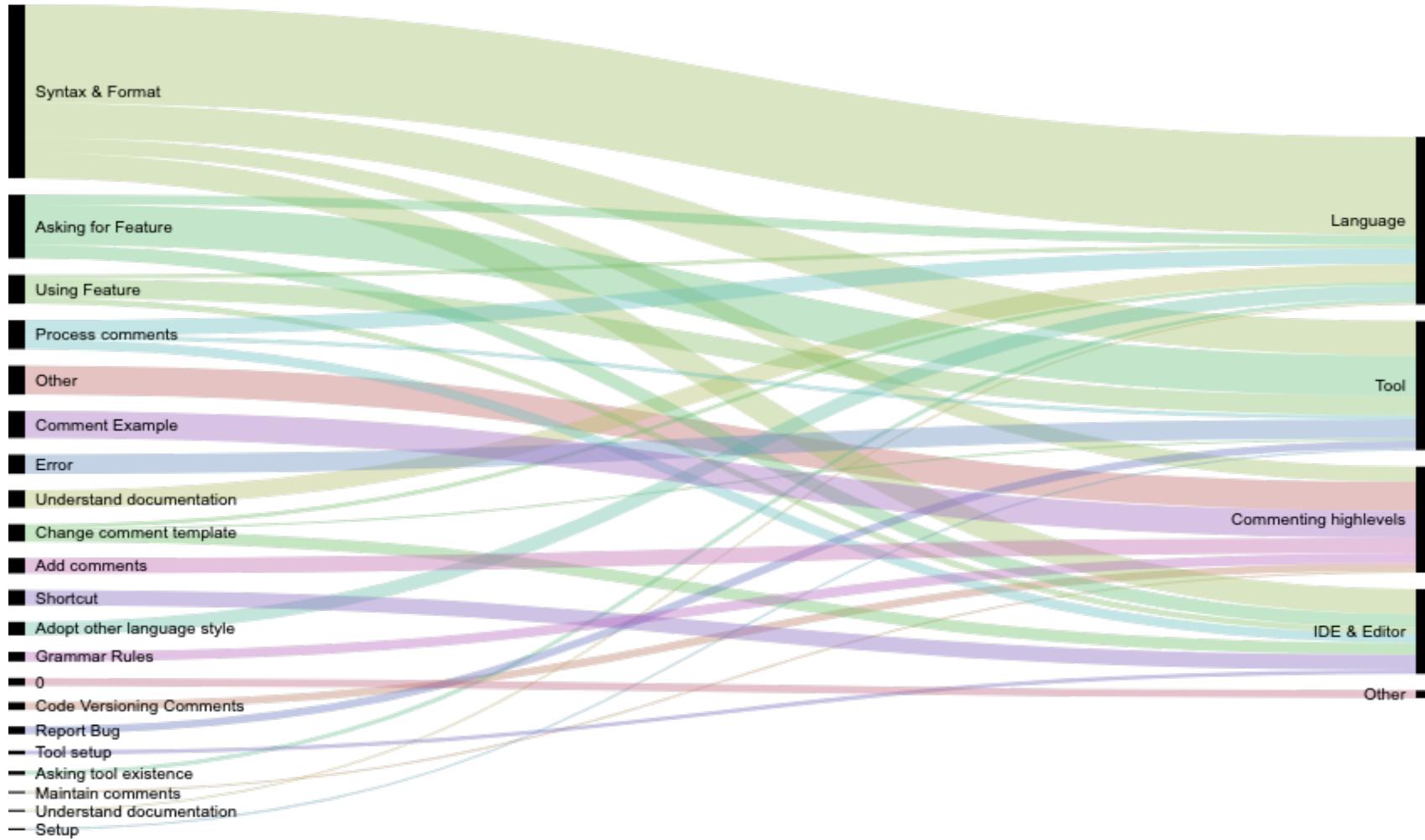
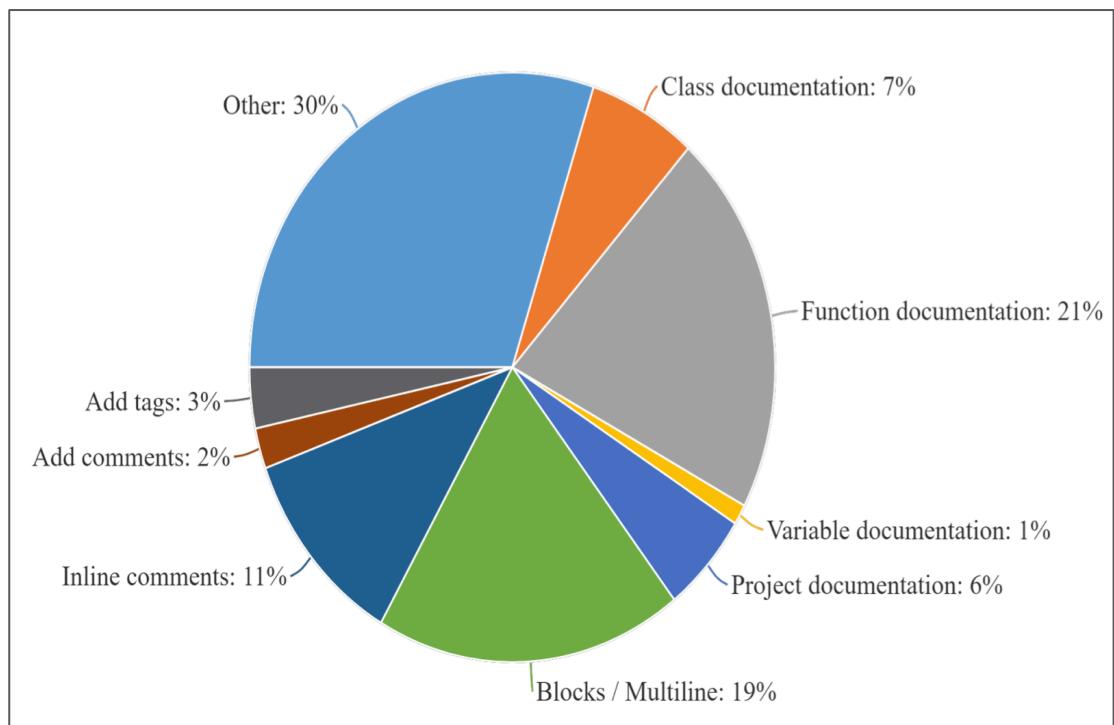


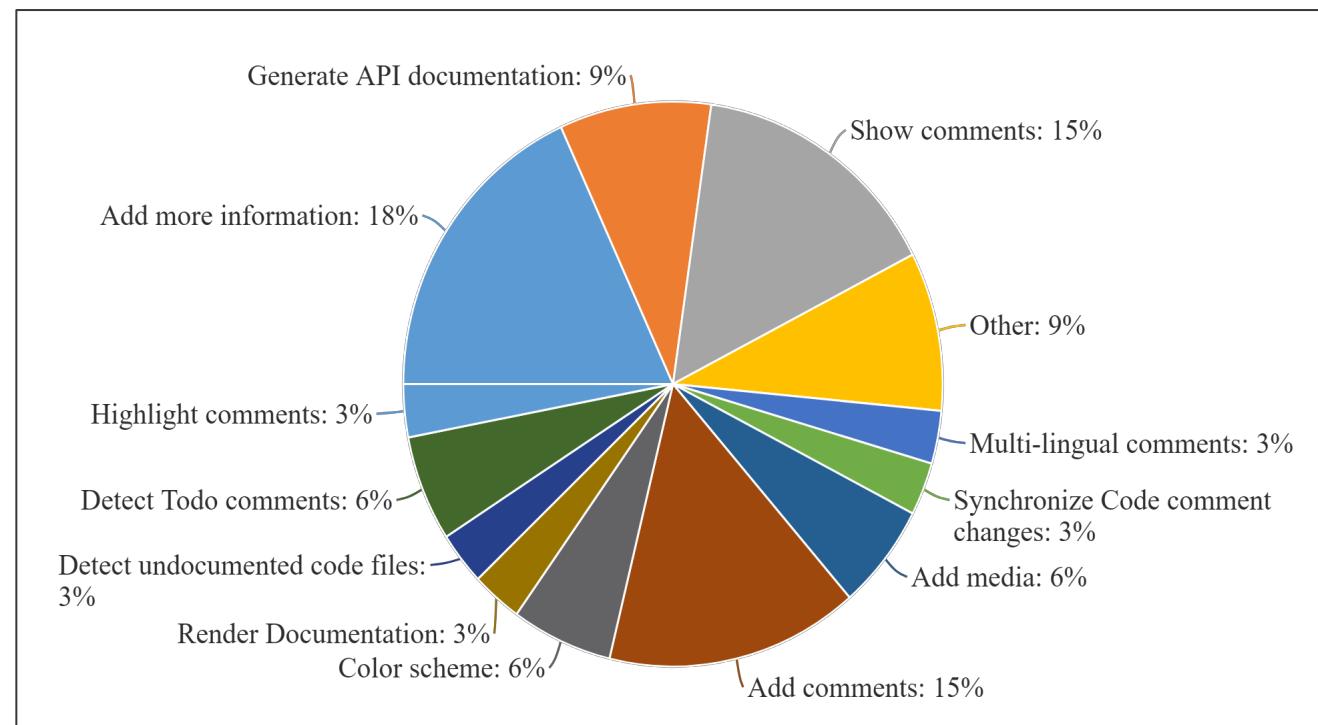
Fig: Second level Categories ordered according to their frequency

**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



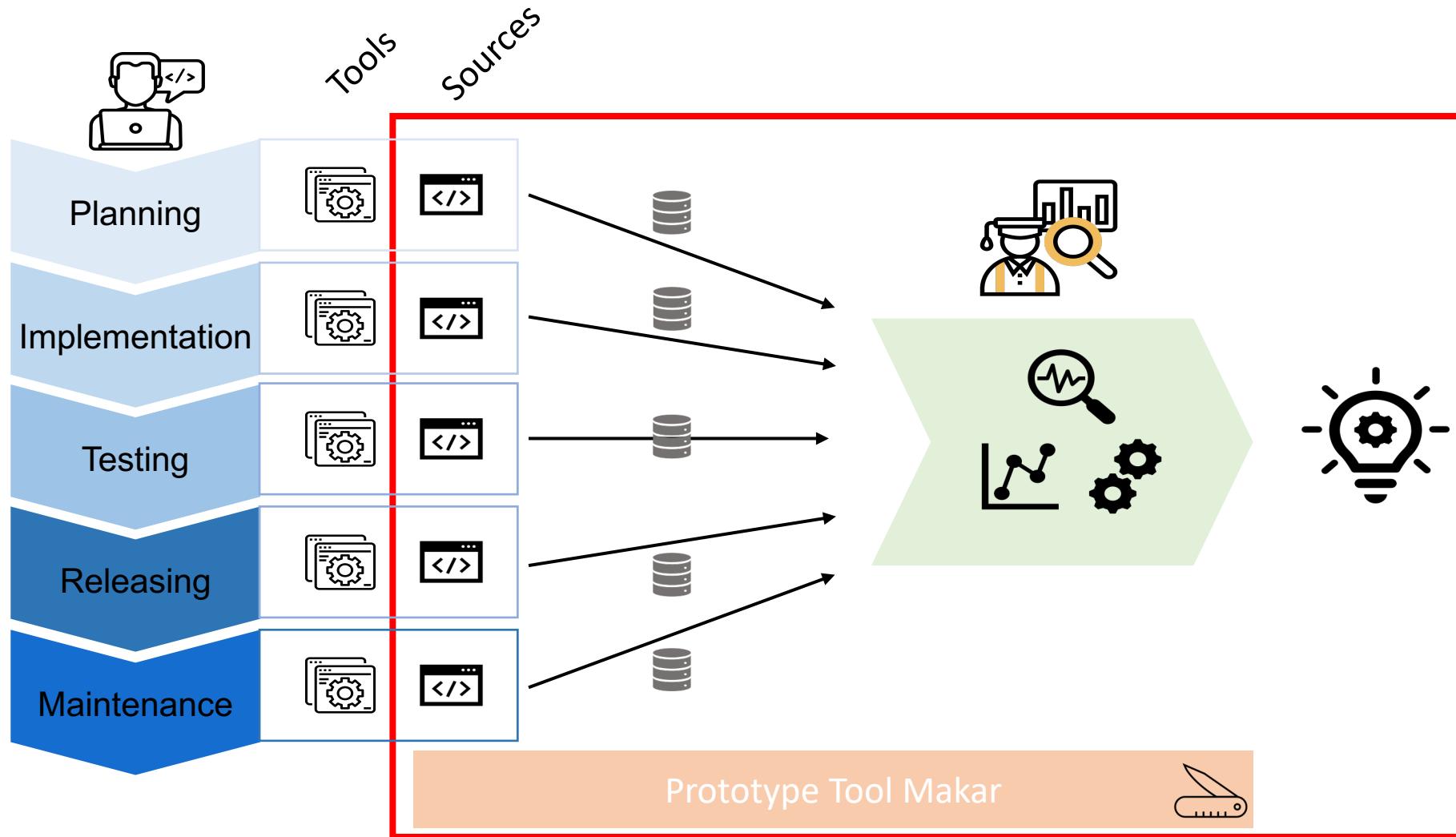
Tools to assess comment quality

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



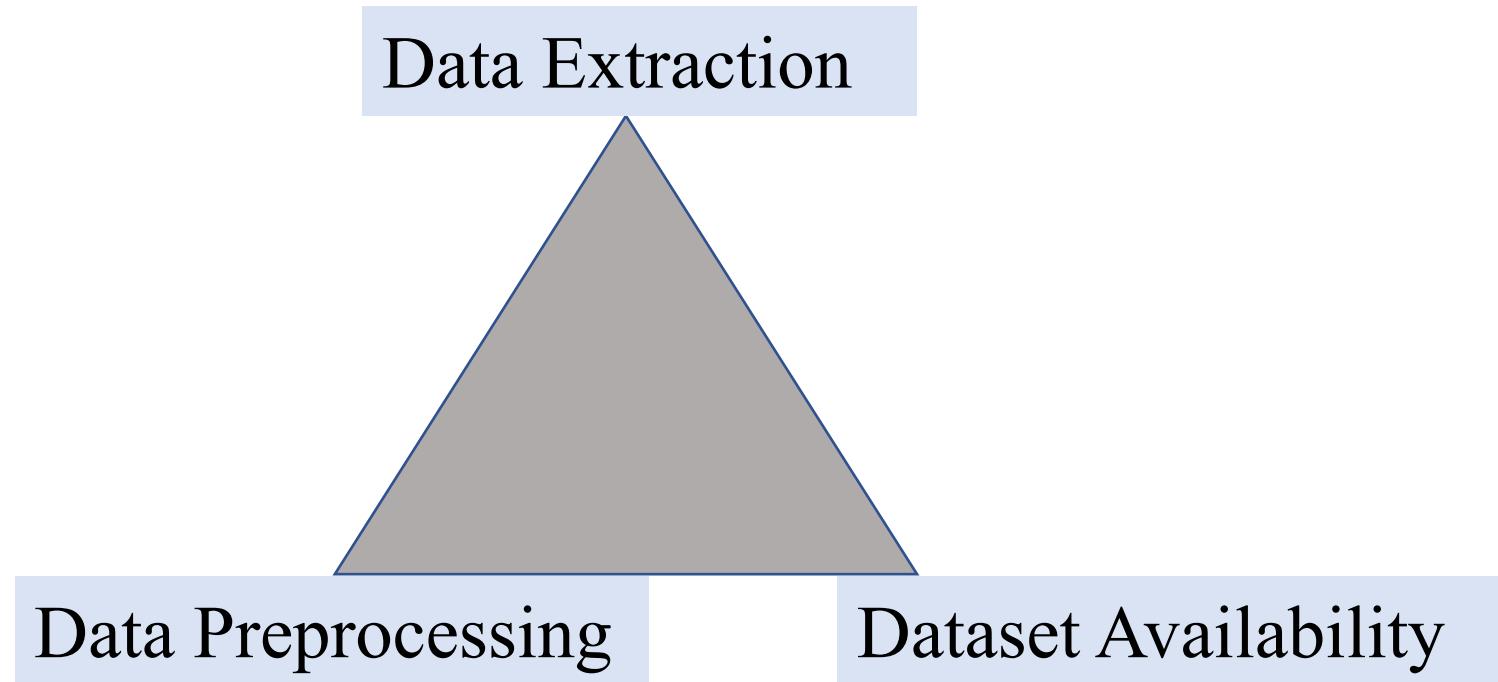
Ease of finding commenting guidelines

Developers face problems in locating conventions

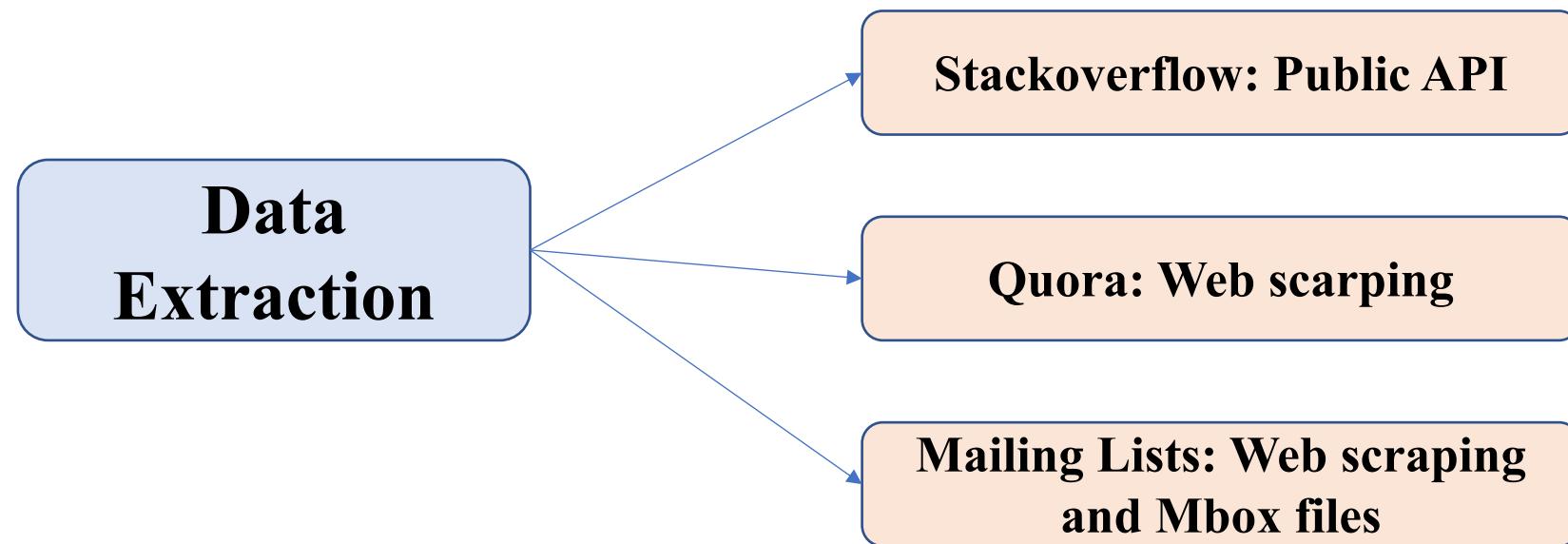


# Our Study Reproducibility

# Our Study Reproducibility: Parameters



# Our Study Reproducibility: Data Extraction

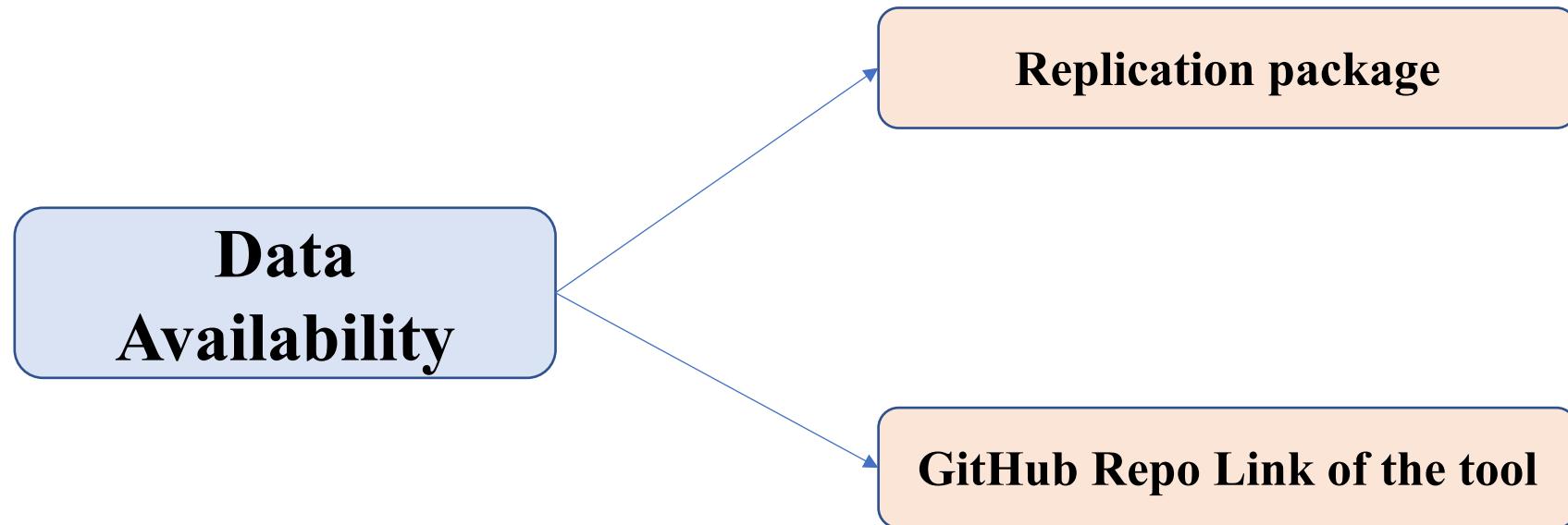


# Our Study Reproducibility: Data Pre-processing



Transformation	<code>extract_code</code>	<code>strip_html</code>	<code>string_replace</code>	<code>remove_stopwords</code>	<code>word_stemming</code>
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

### Developers' Information Needs

Internal Sources

External Sources

Planning → Implementation → Testing → Refining → Maintenance

### Dataset Availability

Category	Number of Papers
Research availability	~45
Extraction	~75
Processing	~75
Ingestion and Preprocessing	~20

### Study Goal

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

### New Methodology

### LDA Topics: Irrelevant Topics

- Topic 0: Team Roles
- Topic 1: Responsibilities
- Topic 2: Documentation Generation
- Topic 3: IDE & Editors
- Topic 4: Processing Code Conventions
- Topic 5: Coding
- Topic 6: Project Documentation
- Topic 7: Naming Conventions in Projects
- Topic 8: Naming Code Entities
- Topic 9: Continuous Delivery Pipeline
- Topic 10: Visual Editors in WebIDE
- Topic 11: Coding Framework for Visual Commenting
- Topic 12: Accessibility
- Topic 13: Continuous Spikes

### Taxonomy: Which types of information developers seek?

```
graph TD; CC[Code Comment Conventions] --> CS[Coding Style]; CC --> CP[Coding Practice]; CS --> JS[Java]; CS --> PY[Python]; CS --> C[C/C++]; CS --> G[Git]; CS --> CI[CI/CD]; CS --> T[Testing]; CP --> JCI[Java CI]; CP --> P[Python]; CP --> CCI[C/C++ CI]; CP --> GCI[Git CI]; CP --> CDT[Code Coverage Tools]; CP --> TCI[Testing CI]; CP --> JCI[Java CI]; CP --> PCI[Python CI]; CP --> CCI[C/C++ CI]; CP --> GCI[Git CI]; CP --> CDT[Code Coverage Tools]; CP --> TCI[Testing CI];
```

### Implication

Findings relevant for developers and researchers

### 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 information 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.