# Developers' Information Needs on Collaborative Platforms

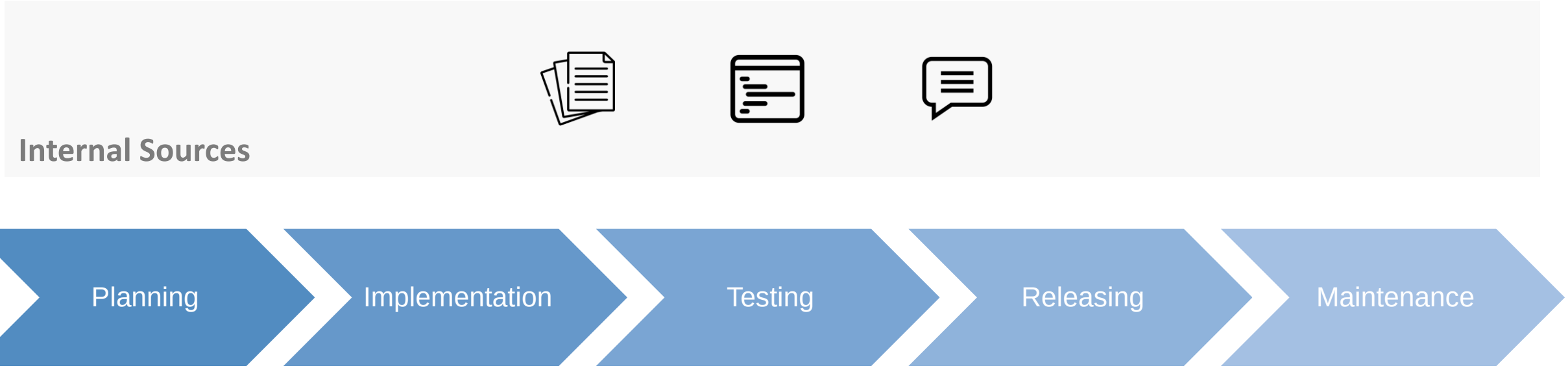
Master Thesis of Mathias Birrer
Supervised by Pooja Rani

Final Presentation

UNIVERSITÄT BERN

MASTER IN COMPUTER SCIENCE

# Developers' Information Needs



# External Sources of Information

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

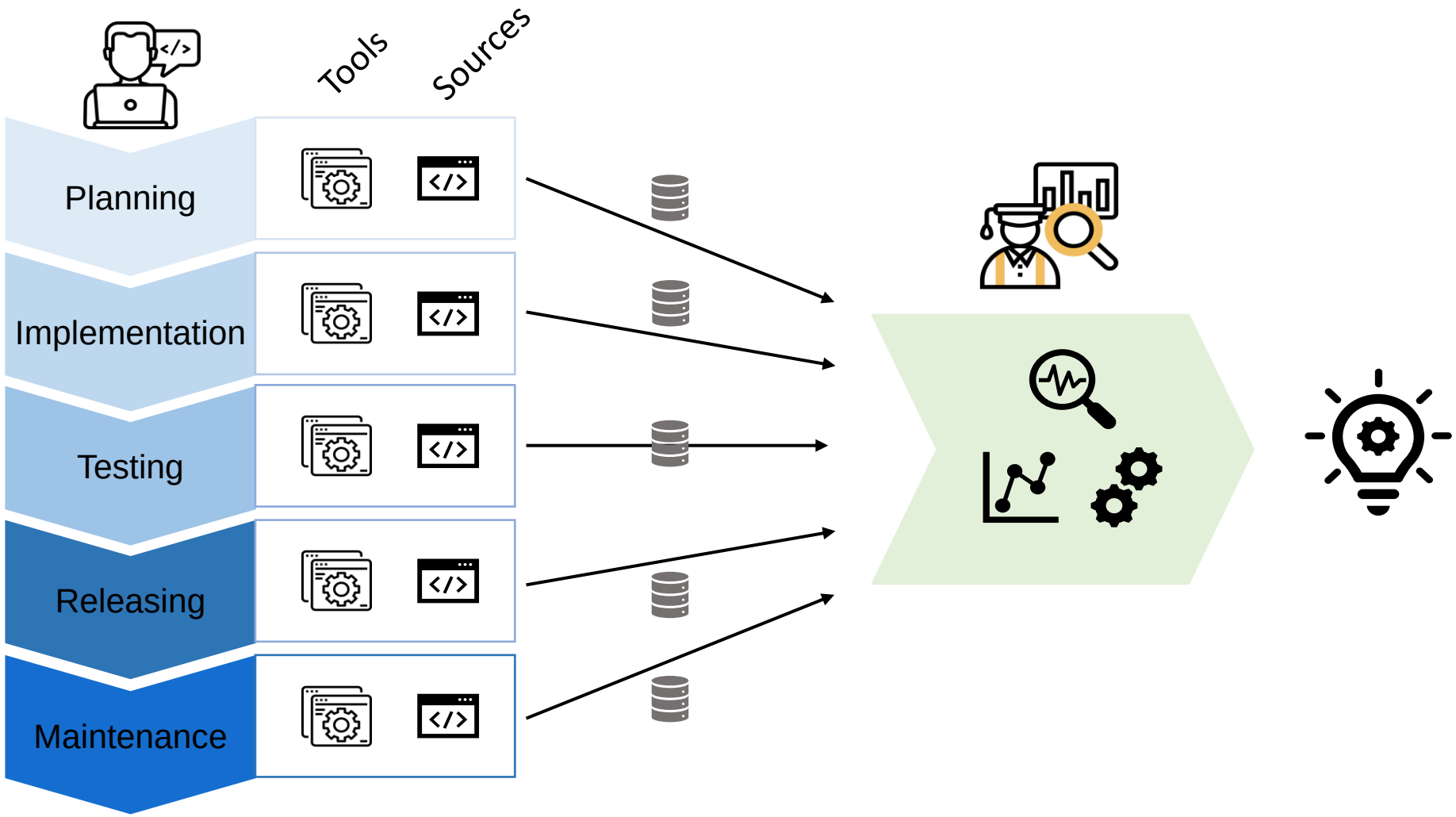
- Unstructured data
- Fragmented developer workflow
- Diversity

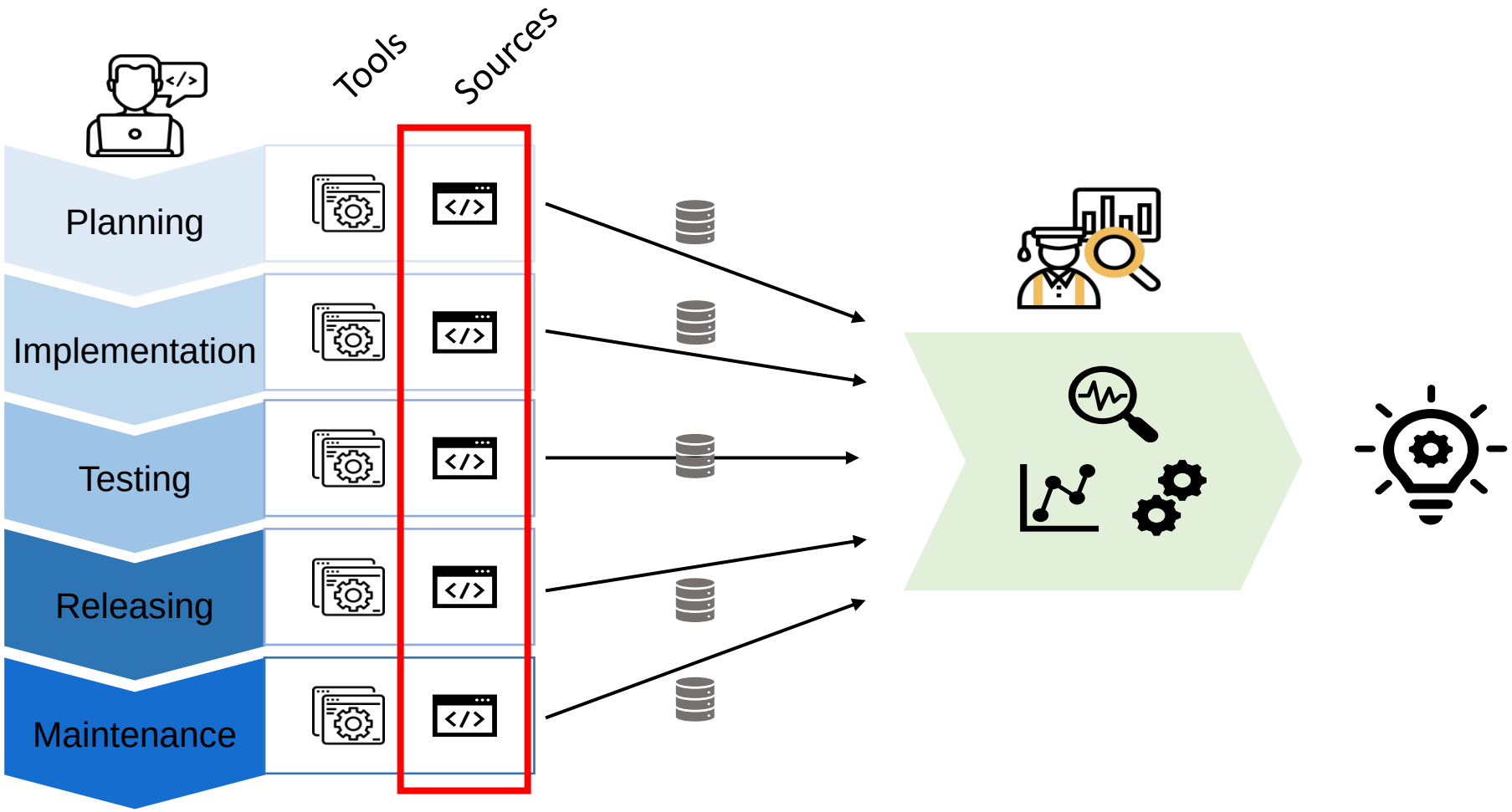
→ Researchers started investigating these sources

- Compare sources / communities
- Uncover evolution
- Reuse datasets

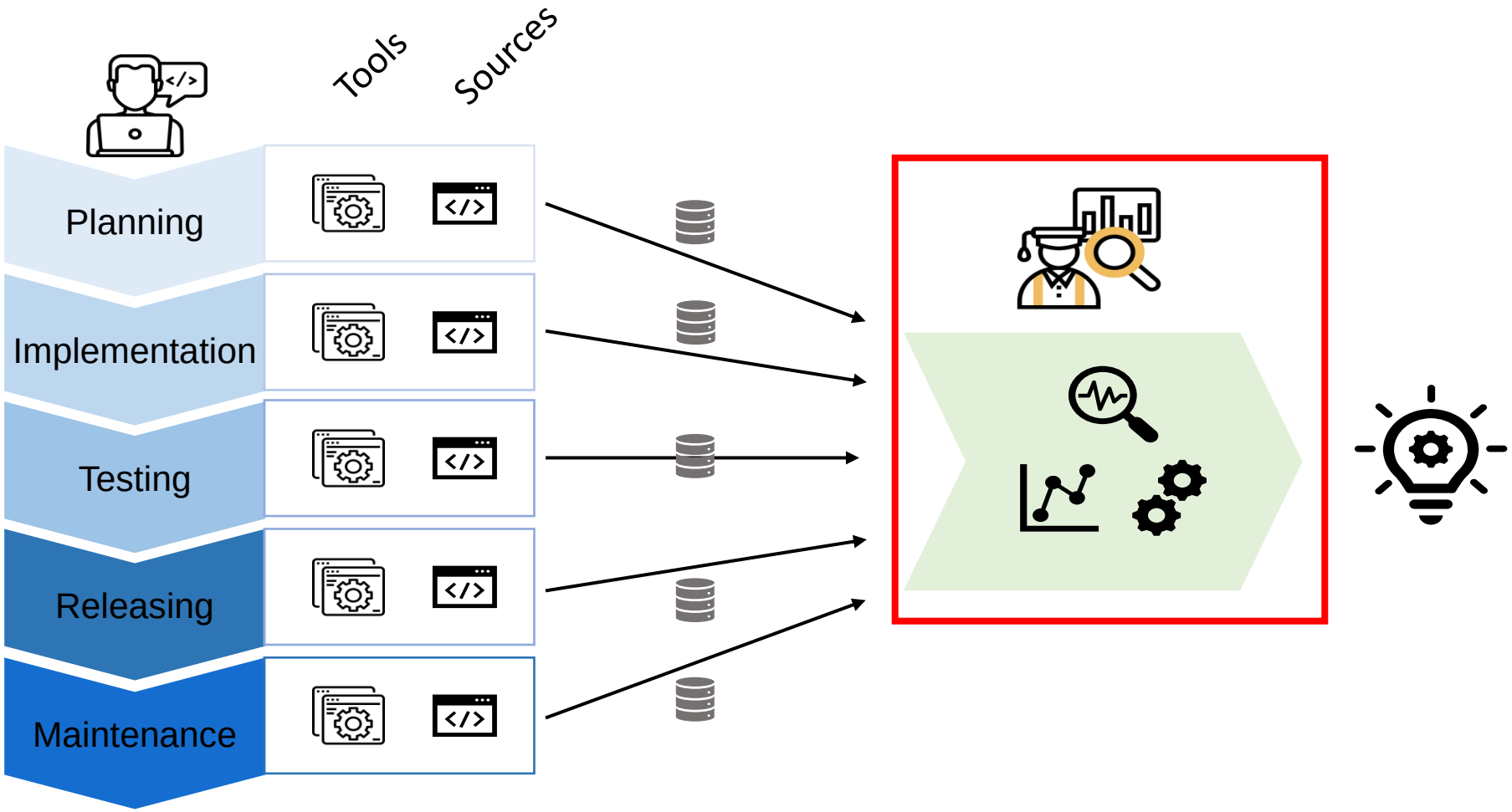
→ **Reproducibility** is highly important

# Researchers workflow

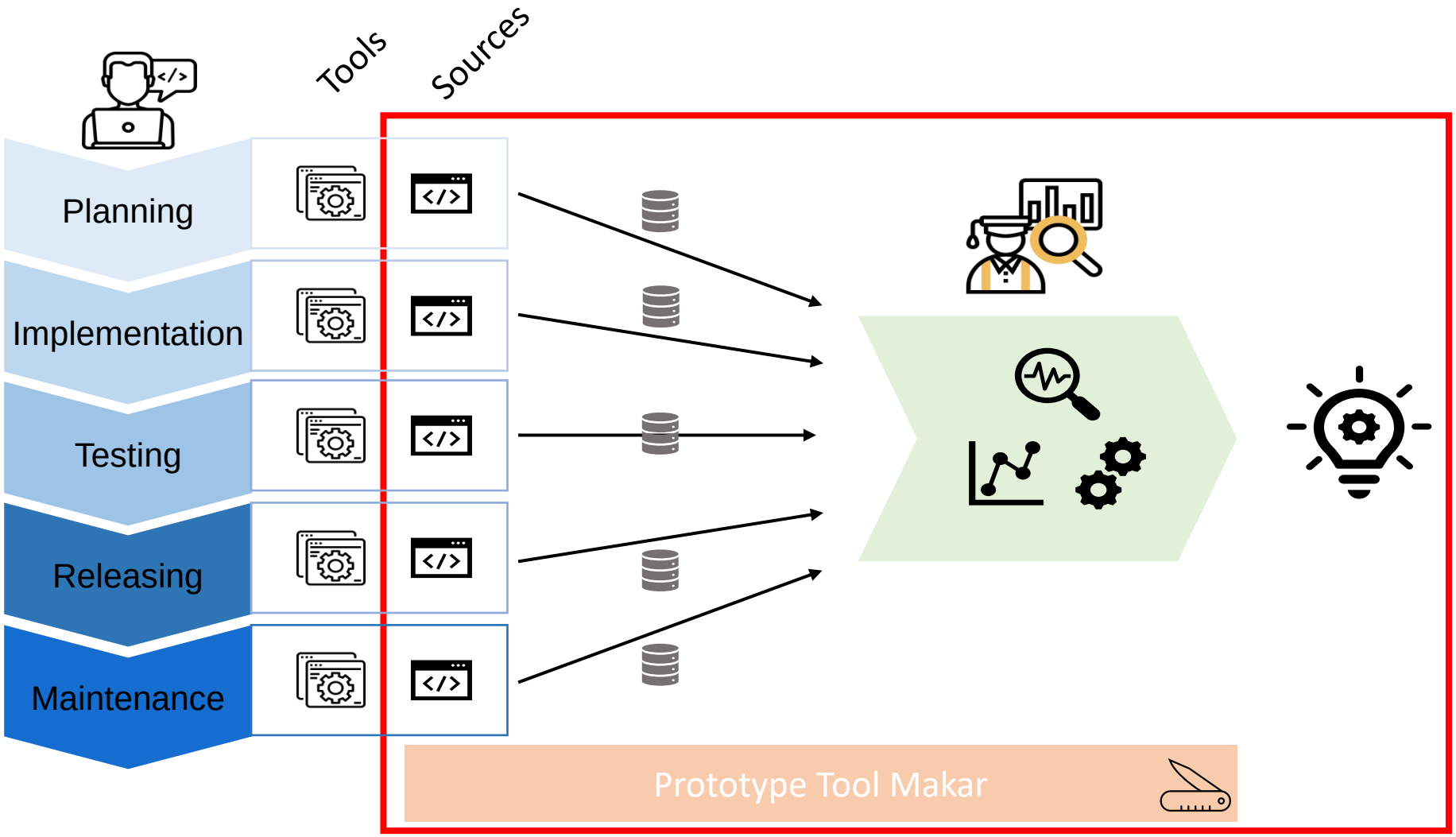





**Background**: Which data sources are typically analyzed by researchers to understand developers' information needs?

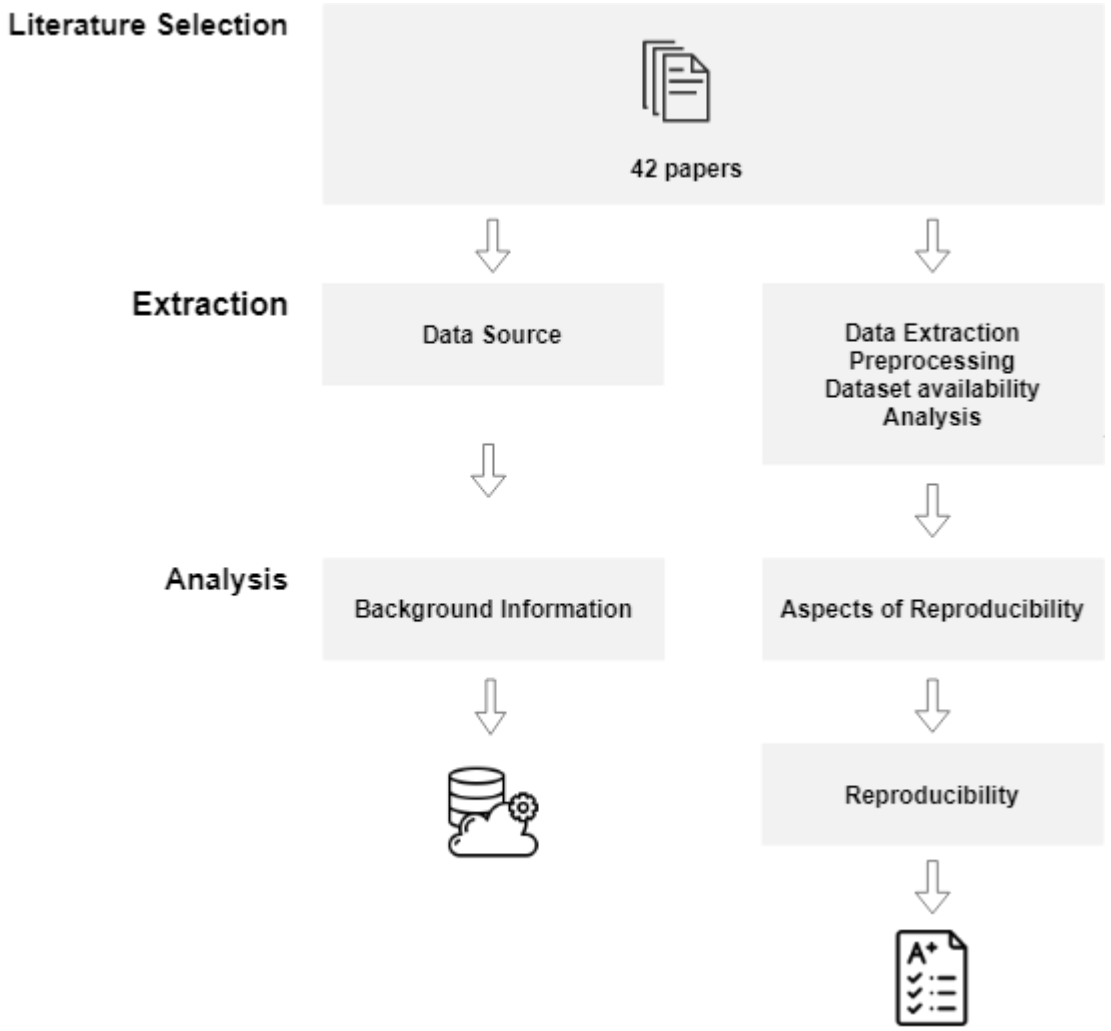



**RQ1**: How do researchers analyze developers' information needs on collaborative platforms?

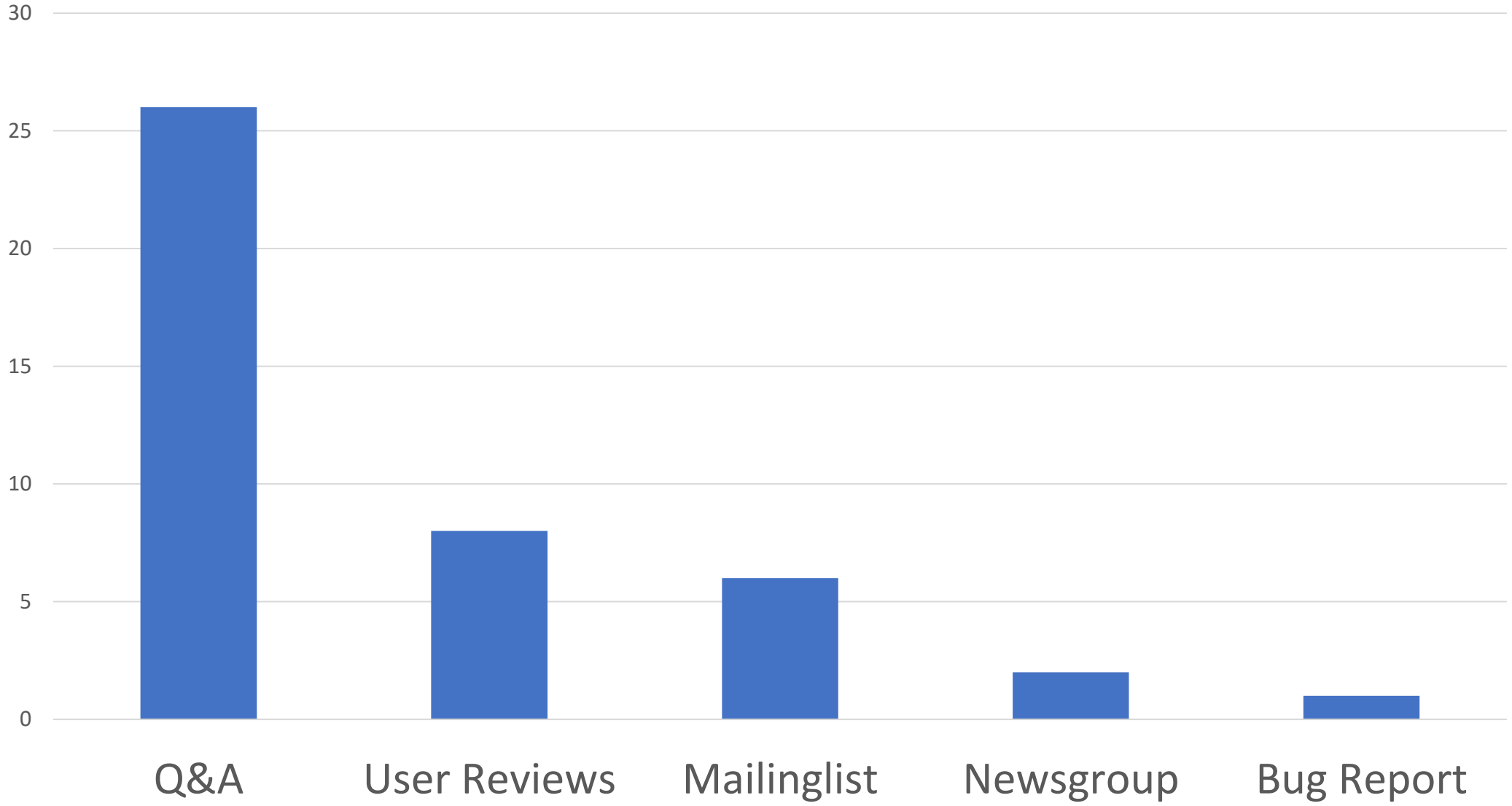



**RQ2 (Case Study)**: What are developers' questions about code comment conventions?

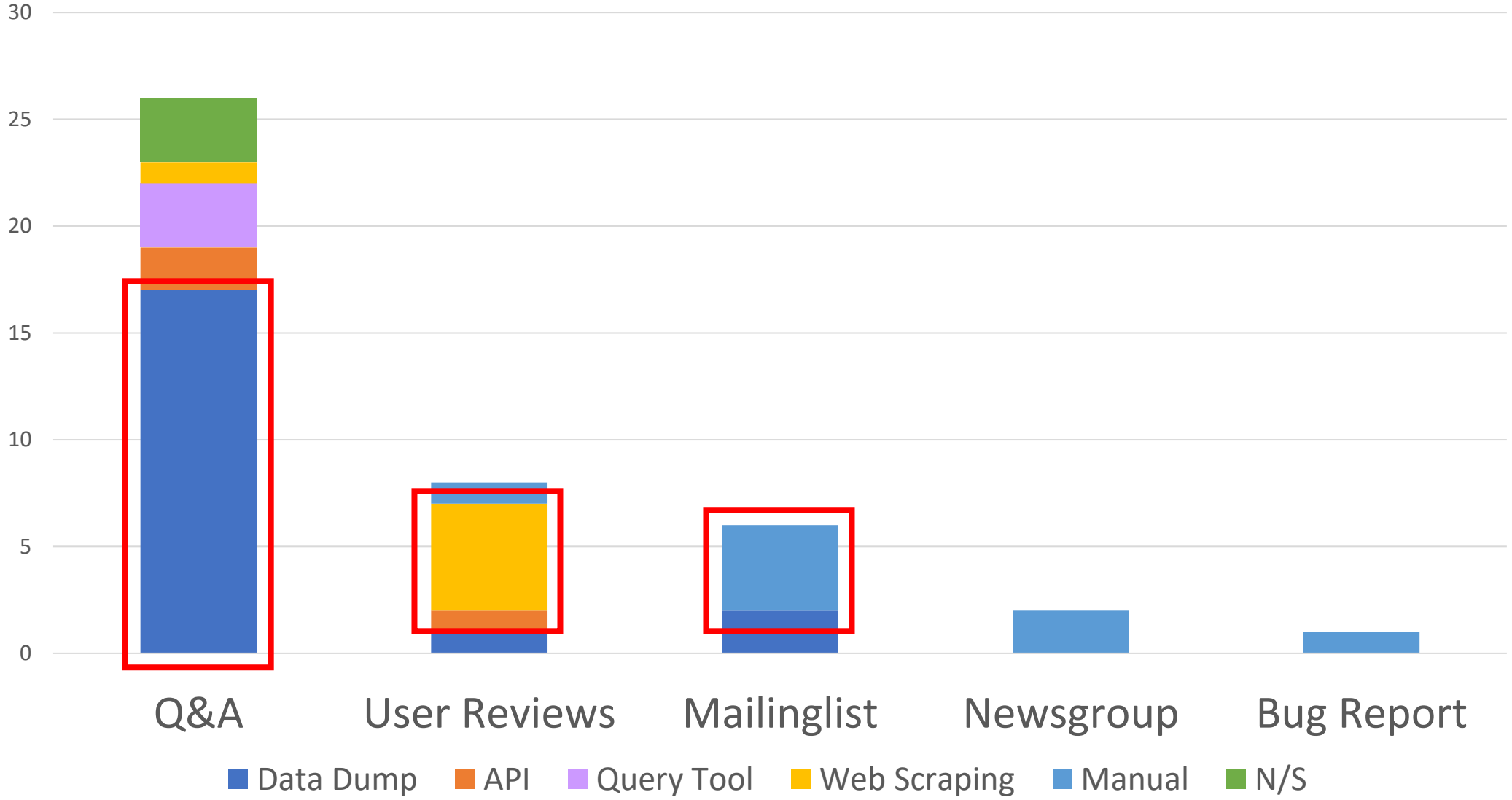
# Methodology RQ1



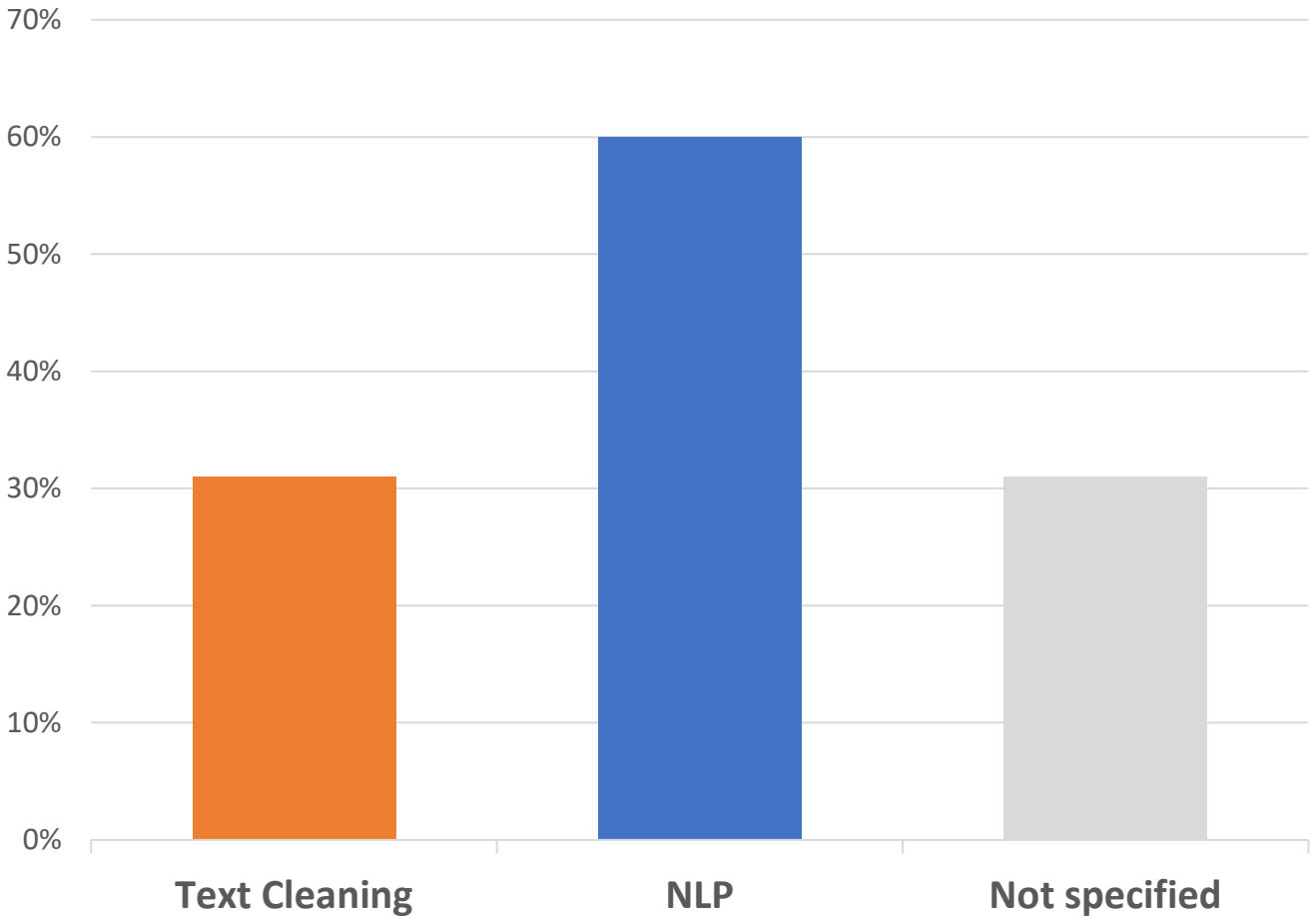
# Results RQ1 – Source Categories



# Source Categories with Extraction Method



# Data Preprocessing

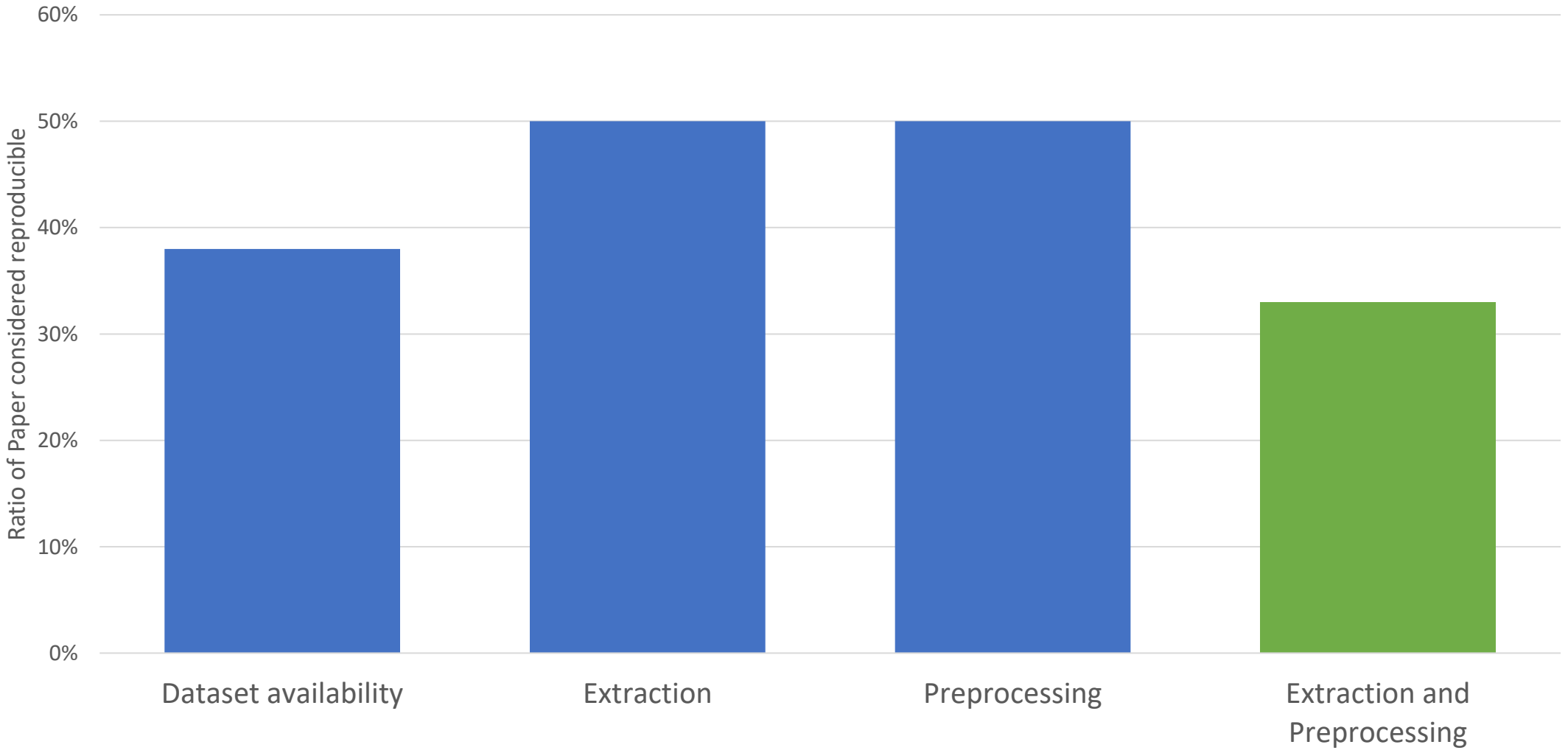



- Remove **source code**
- Remove **HTML tags**
- Remove **Punctuation**
- Remove **Non-Alpha-Numeric**

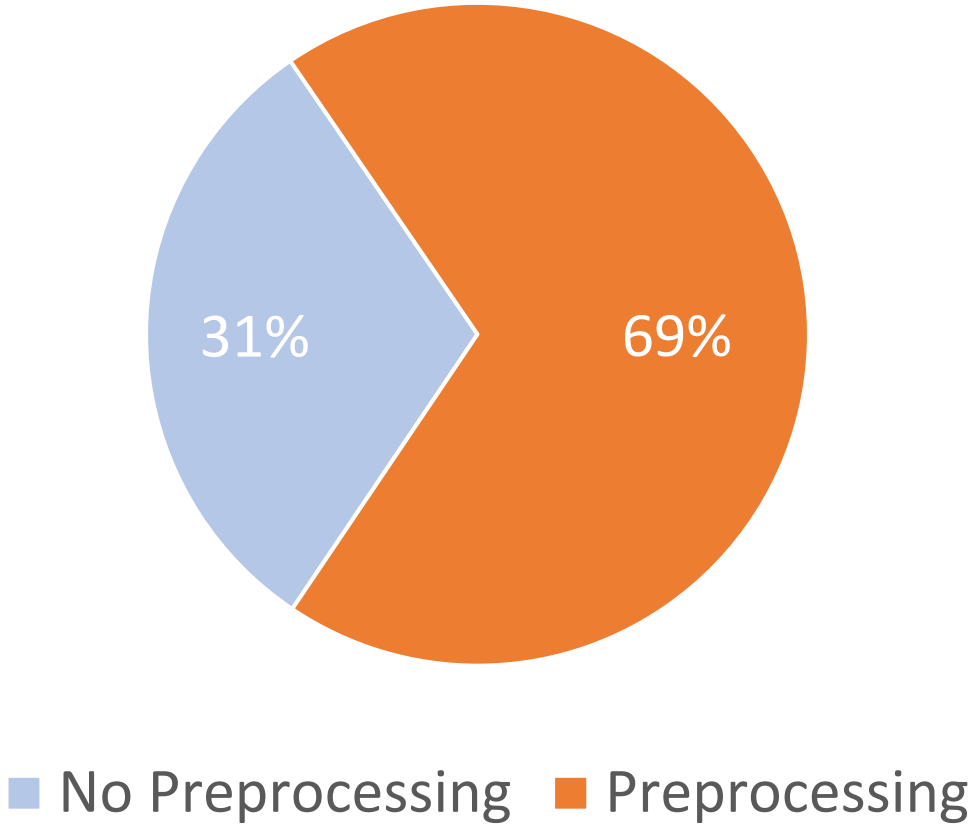
- Remove **Stop words**
- Apply **Word stemming**
- Apply **Lemmatization**
- Filter **Non-English**
- **Case unification**

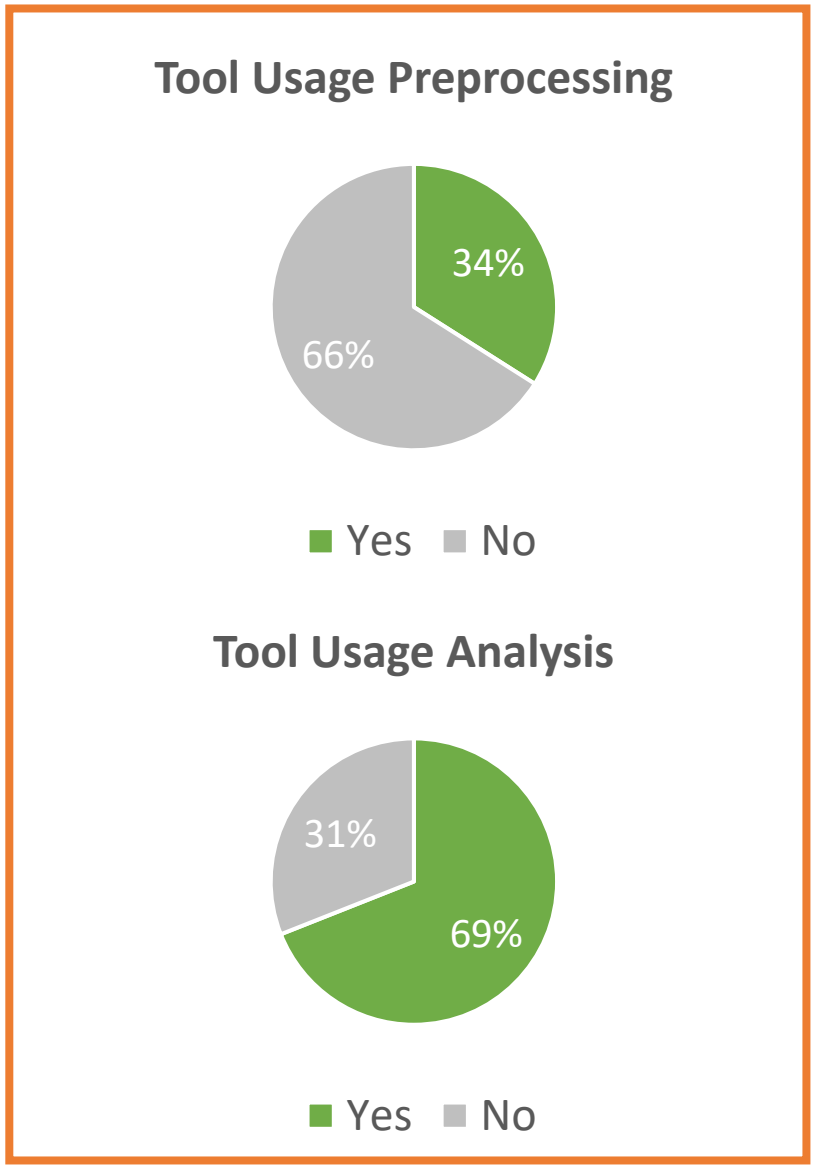
# Common methodology

# Reproducibility Aspects



# Tool Usage for Preprocessing



# Tools for Preprocessing

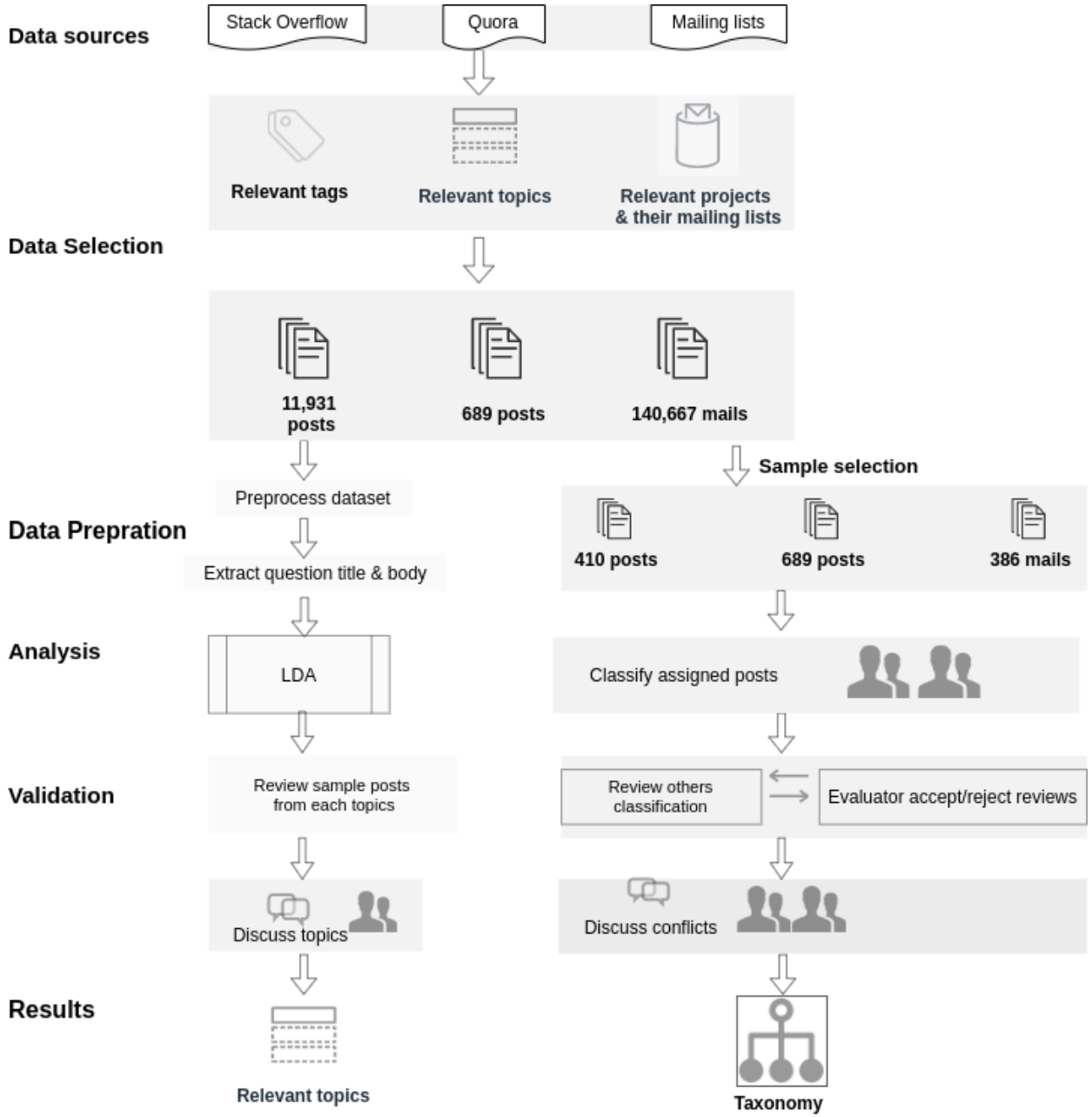
- Common Tools
  *Porter Stemmer, Stanford Parser, Python NLTK, Apachen OpenNLP*
- Tools for NLP
- **No tool to manage or automate preprocessing workflow**

# RQ2 – Case Study

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

# Code Comment Conventions

- Trustworthy form of documentation
- Basis for documentation tools
- Style & Syntax cannot be enforced
- Conventions for Languages, Companies, Projects, Developers
- **Confusion amongst developers**


Data sources
Stack Overflow
Quora
Mailing lists
Relevant tags
Relevant topics
Relevant projects & their mailing lists
Data Selection
11,931 posts
689 posts
140,667 mails
Sample selection
Data Prepration
Preprocess dataset
Extract question title & body
410 posts
689 posts
386 mails
Analysis
LDA
Classify assigned posts
Validation
Review sample posts from each topics
Review others classification
Evaluator accept/reject reviews
Discuss topics
Discuss conflicts
Results
Relevant topics
Taxonomy

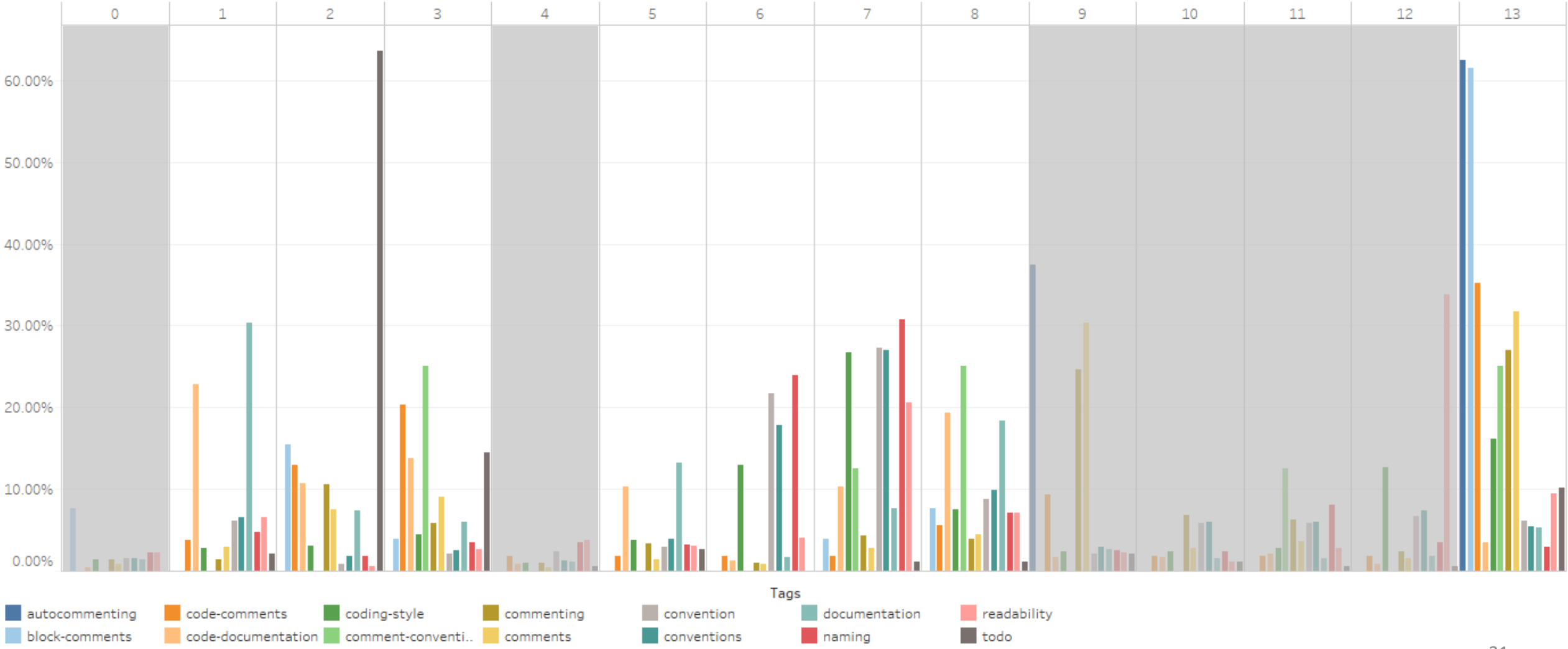
# LDA Topics

| # | Topic Name |
|---|---|
| 0 | **Exceptions** |
| 1 | **Documentation Generation** |
| 2 | **IDE & Editors** |
| 3 | **Processing Code Comments** |
| 4 | **Testing & Naming Conventions** |
| 5 | **Project Documentation** |
| 6 | **Project Naming Conventions** |
| 7 | **Code Entities Naming Conventions** |
| 8 | **Comments Writing Strategies** |
| 9 | **Comment Functionality Websites** |
| 10 | **Comment Framework** |
| 11 | **Database** |
| 12 | **Code Conventions** |
| 13 | **Comments Syntax** |

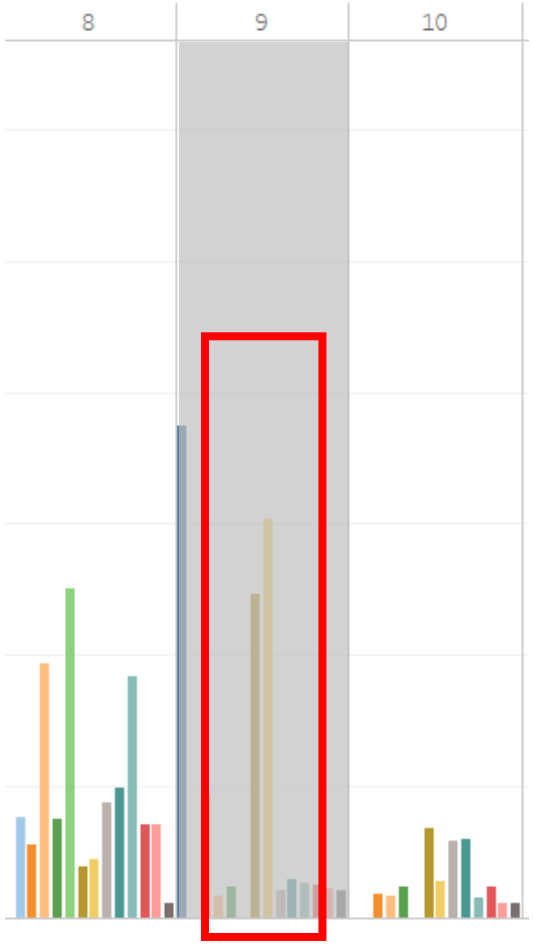
# LDA Topics

| # | Topic Name |
|---|---|
| 0 | **Exceptions** |
| 1 | **Documentation Generation** |
| 2 | **IDE & Editors** |
| 3 | **Processing Code Comments** |
| 4 | **Testing & Naming Conventions** |
| 5 | **Project Documentation** |
| 6 | **Project Naming Conventions** |
| 7 | **Code Entities Naming Conventions** |
| 8 | **Comments Writing Strategies** |
| 9 | **Comment Functionality Websites** |
| 10 | **Comment Framework** |
| 11 | **Database** |
| 12 | **Code Conventions** |
| 13 | **Comments Syntax** |

# LDA Topics– Tag Distribution



# Problems with tags

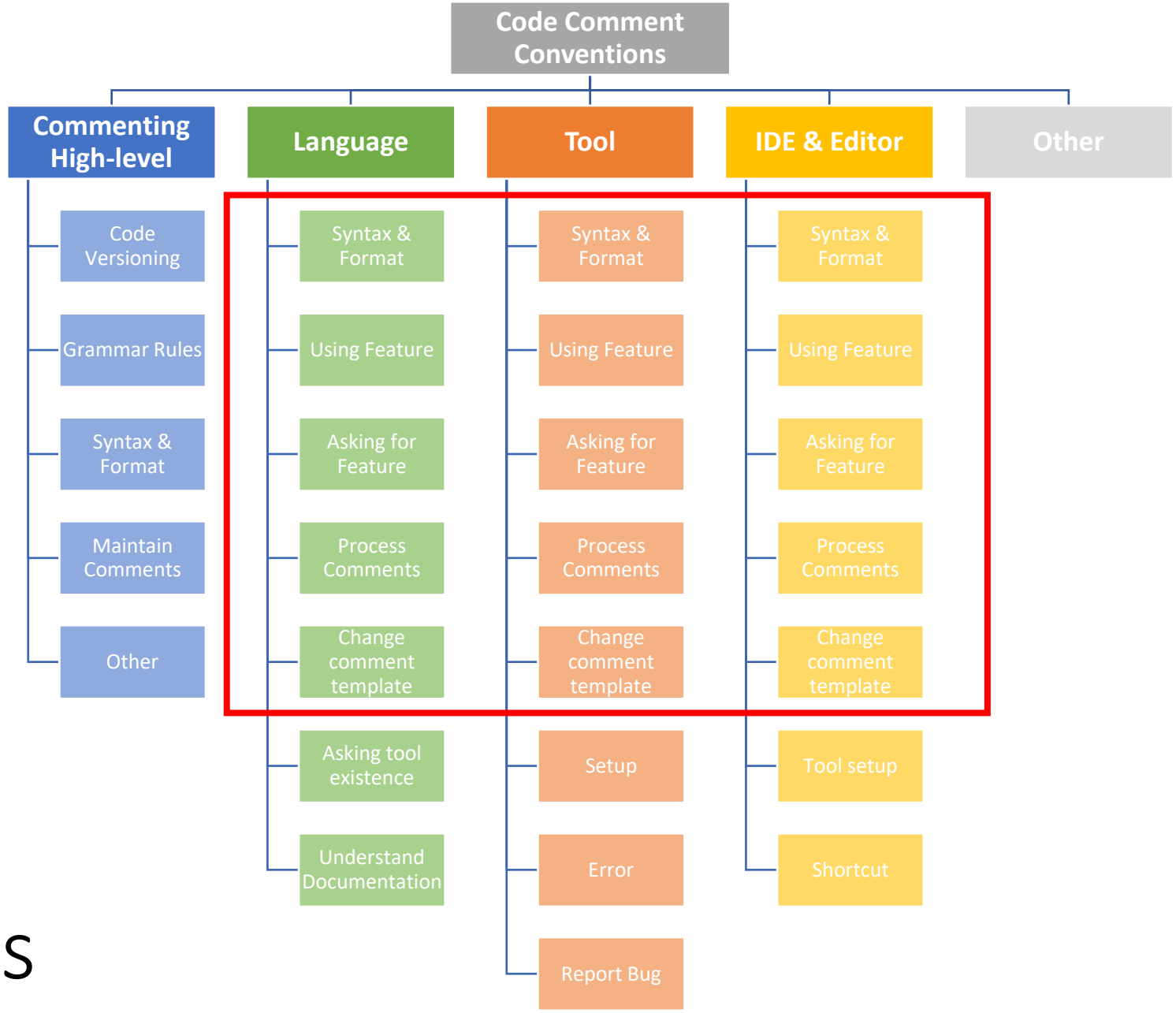



- Tags *comments* and *commenting*
- General and ambiguous
- Irrelevant despite large proportion of tags

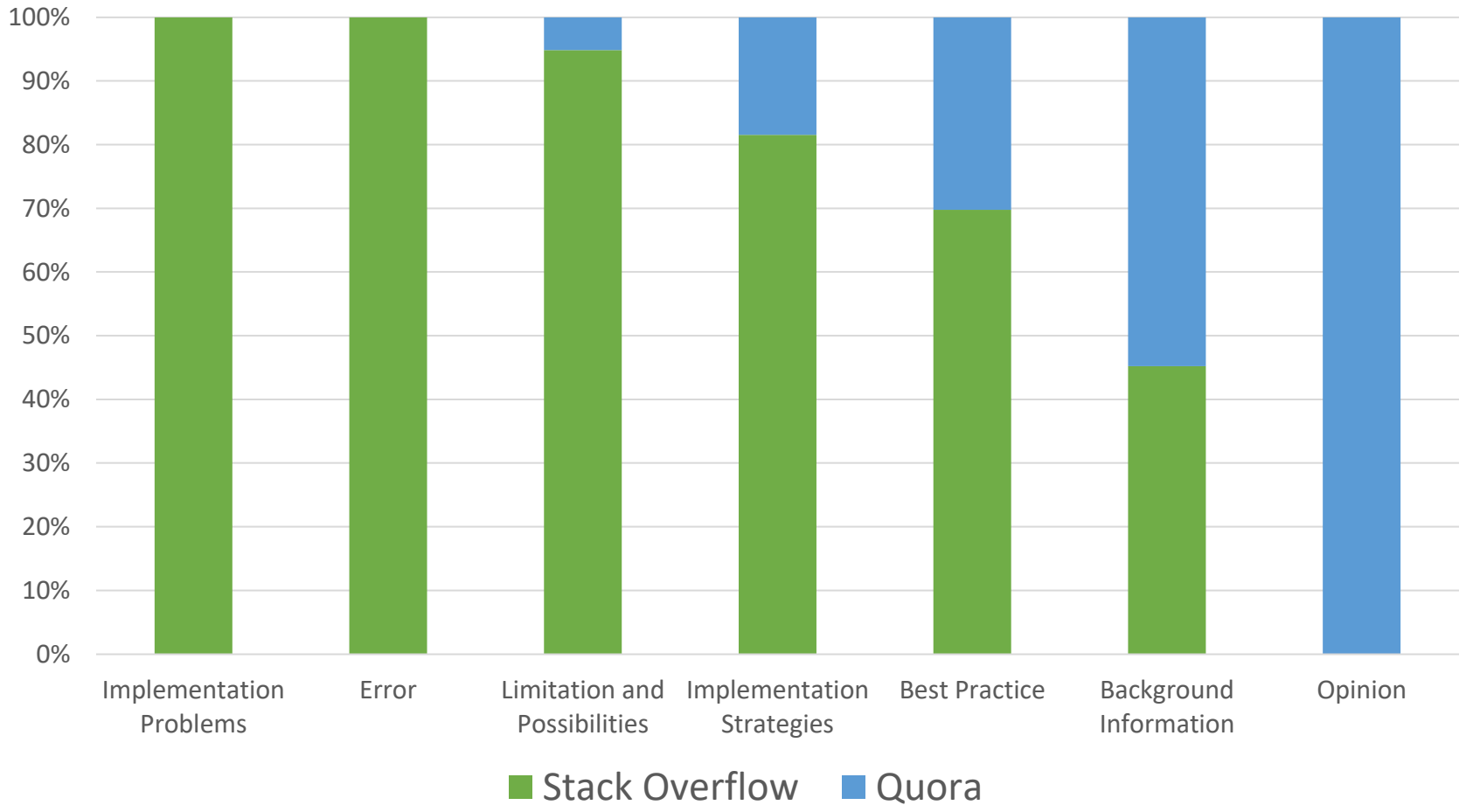
# LDA Technical Details

- MALLET
- Topics $k = 14$
- Hyperparameters
  - $\alpha = 5$
  - $\beta = 0.01$

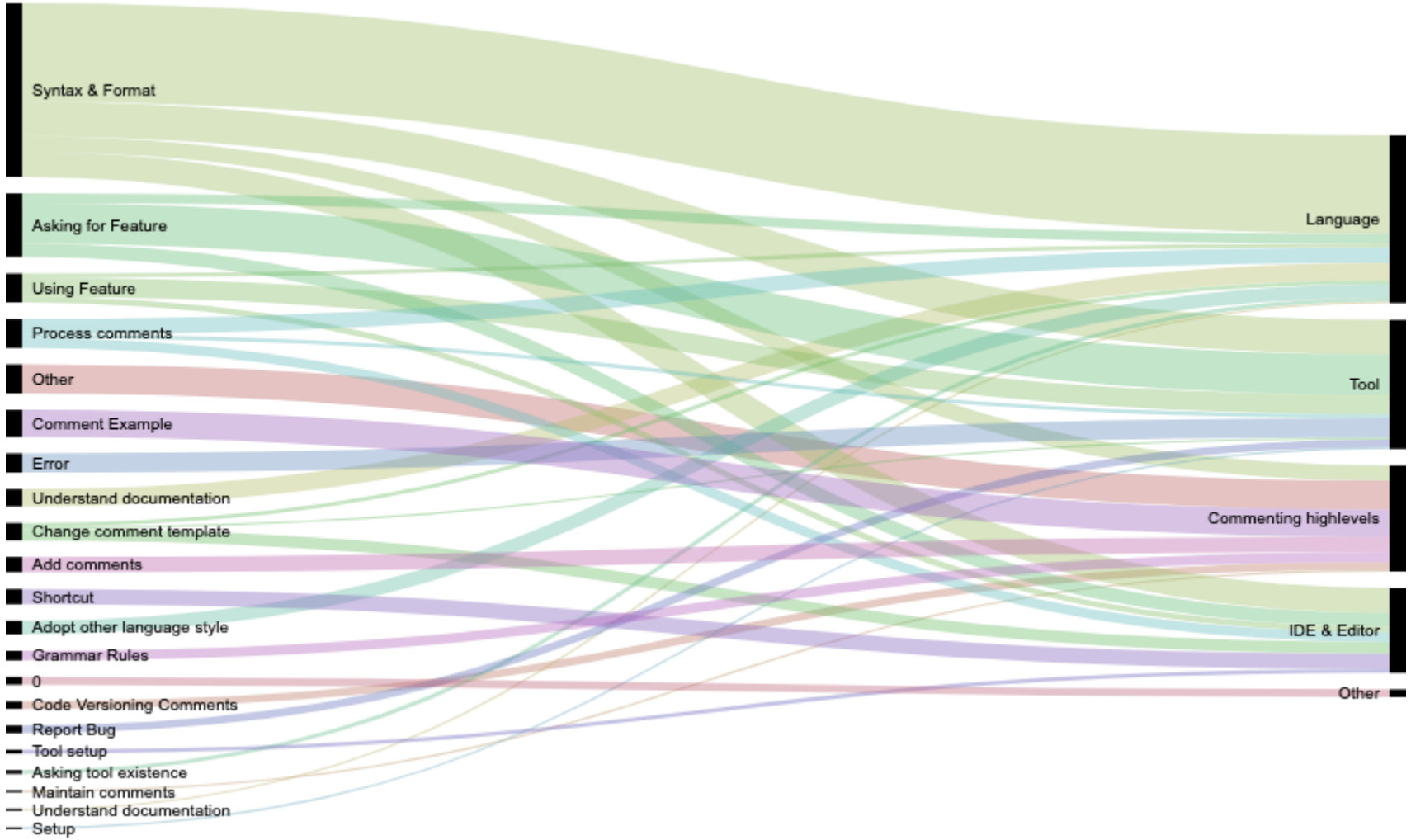
# Manual Analysis



# Question Type on Quora vs. Stack Overflow



# Taxonomy – Most discussed features



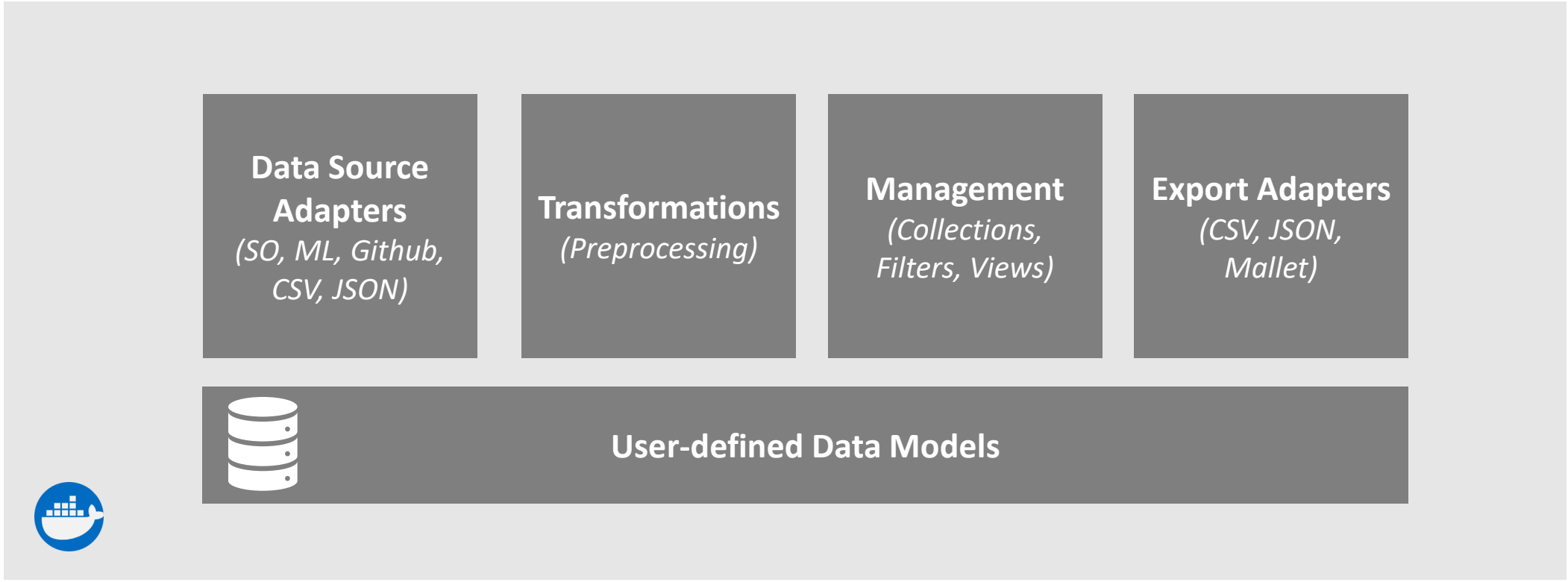
# What happened with Mailing Lists?

- No relevant data found regarding commenting practices
- Despite previous study on documentation issues

# 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*

# Makar – Data Management Tool



# Makar - Functionality

- User-Defined Data Models
- Import Adapters (CSV, JSON, Stack Overflow, Github)
- Transformations for Preprocessing
  - Common preprocessing steps built-in
  - Extensible
- Data Management
  - Collections
  - Search Interface
- Export Adapters (CSV, JSON, Mallet)

# Main Contributions

- Characteristics and challenges of external data sources
- Analysis of reproducibility aspects
- Multi-source study
- Developer questions about *Code Comment Conventions*
- Development of prototype tool *Makar*

# Future Work

**Research on Developers'Information Needs**

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

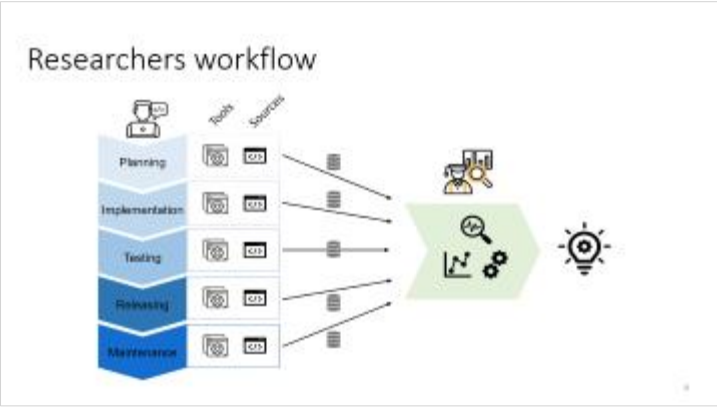
**Code Comment Conventions**

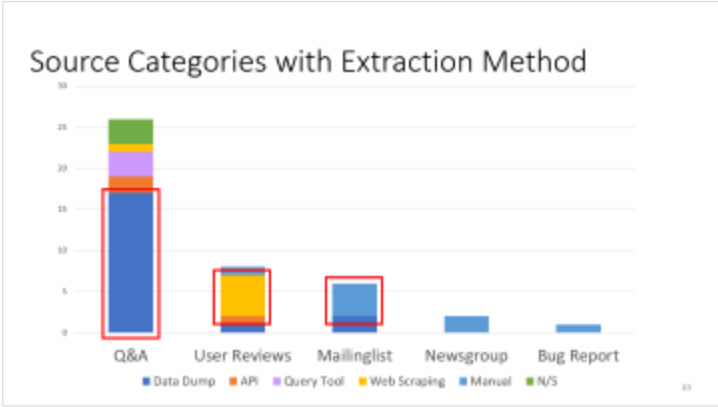
- Investigate more sources (e.g. Github, Jira)

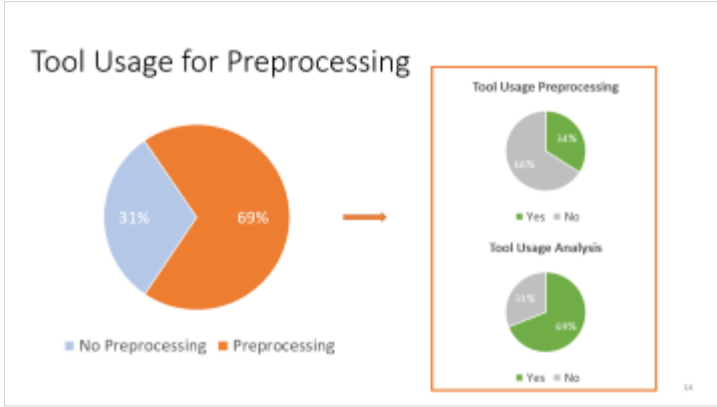
**Prototype Tool**

- Extend data source adapters
- Analysis and visualization components
- Evaluation against similar tools

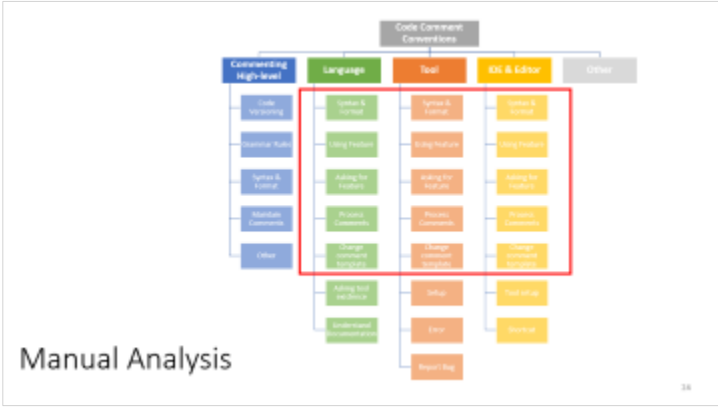
# Summary

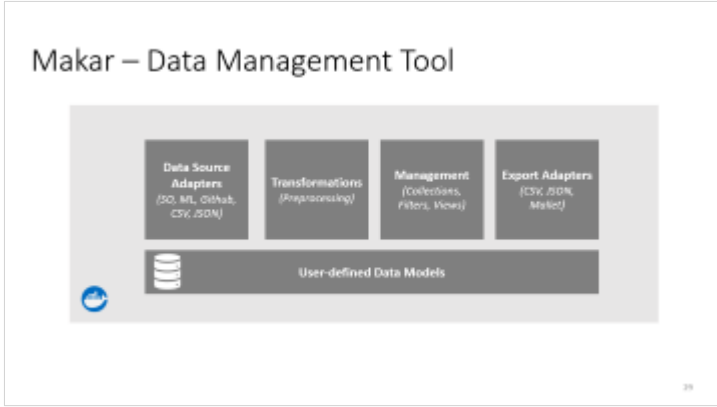

Researchers workflow
Tools
Sources
Planning
Implementation
Testing
Releasing
Maintenance


Source Categories with Extraction Method
Q&A
User Reviews
Mailinglist
Newsgroup
Bug Report
Data Dump
API
Query Tool
Web Scraping
Manual
N/S


Tool Usage for Preprocessing
31%
69%
No Preprocessing
Preprocessing
Tool Usage Preprocessing
14%
86%
Yes
No
Tool Usage Analysis
31%
69%
Yes
No


LDA Topics
# Topic Name
0 Exceptions
1 Documentation Generation
2 IDE & Editors
3 Processing Code Comments
4 Testing & Naming Conventions
5 Project Documentation
6 Project Naming Conventions
7 Code Entities Naming Conventions
8 Comments Writing Strategies
9 Comment Functionality Websites
10 Comment Framework
11 Database
12 Code Conventions
13 Comments Syntax


Code Comment Conventions
Commenting High-level
Language
Tool
IDE & Editor
Other
Manual Analysis


Makar – Data Management Tool
Data Source Adapters
Transformations
Management
Export Adapters
User-defined Data Models