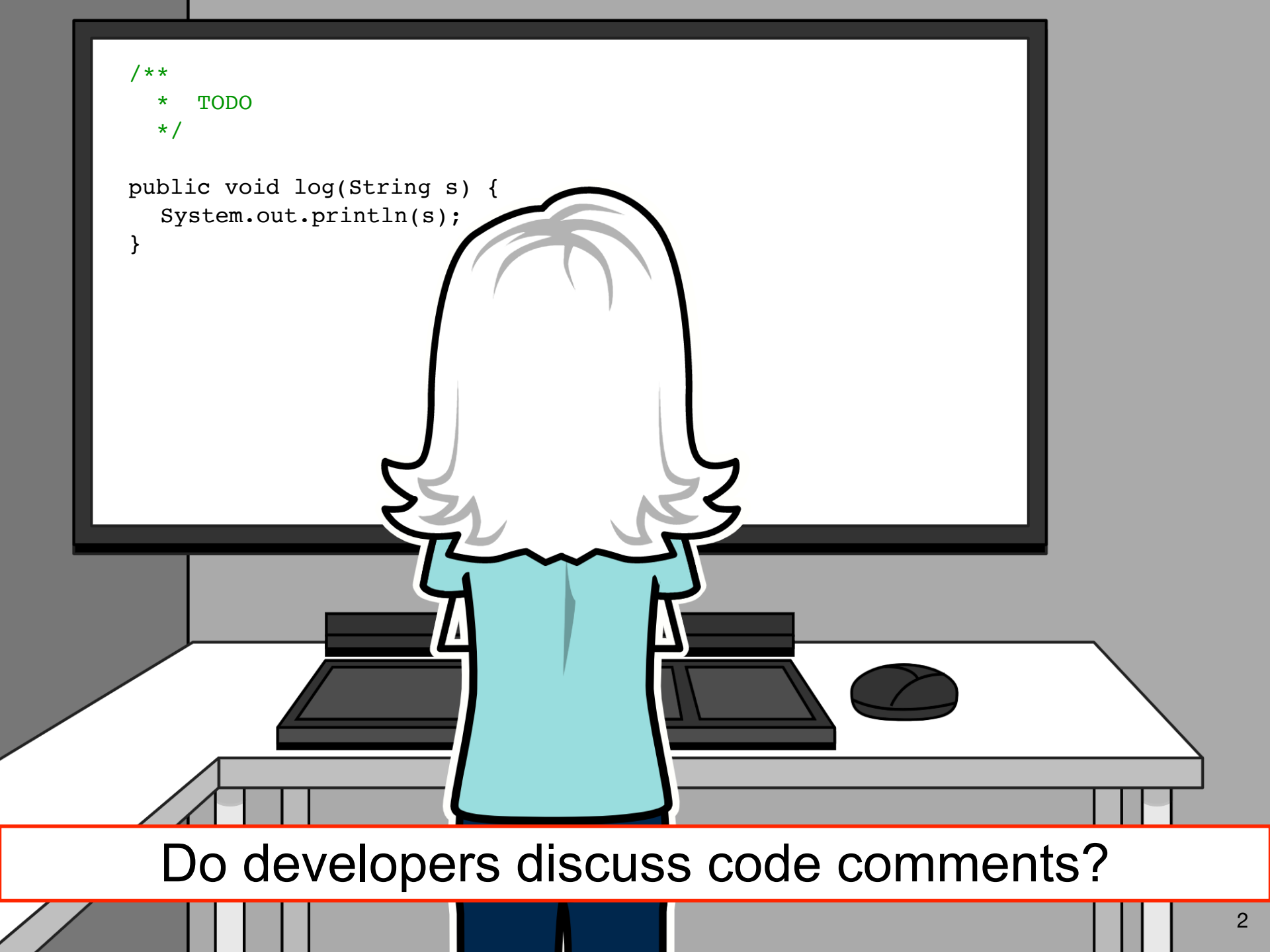
UNIVERSITÄT BERN

# Makar: A Framework for Multi-source Studies Based on Unstructured Data

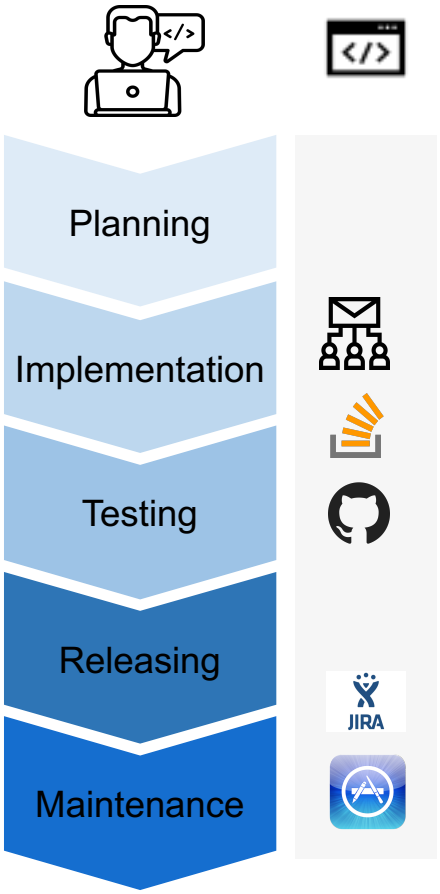
Mathias Birrer, Pooja Rani, Sebastiano Panichella, Oscar Nierstrasz

University of Bern, Switzerland

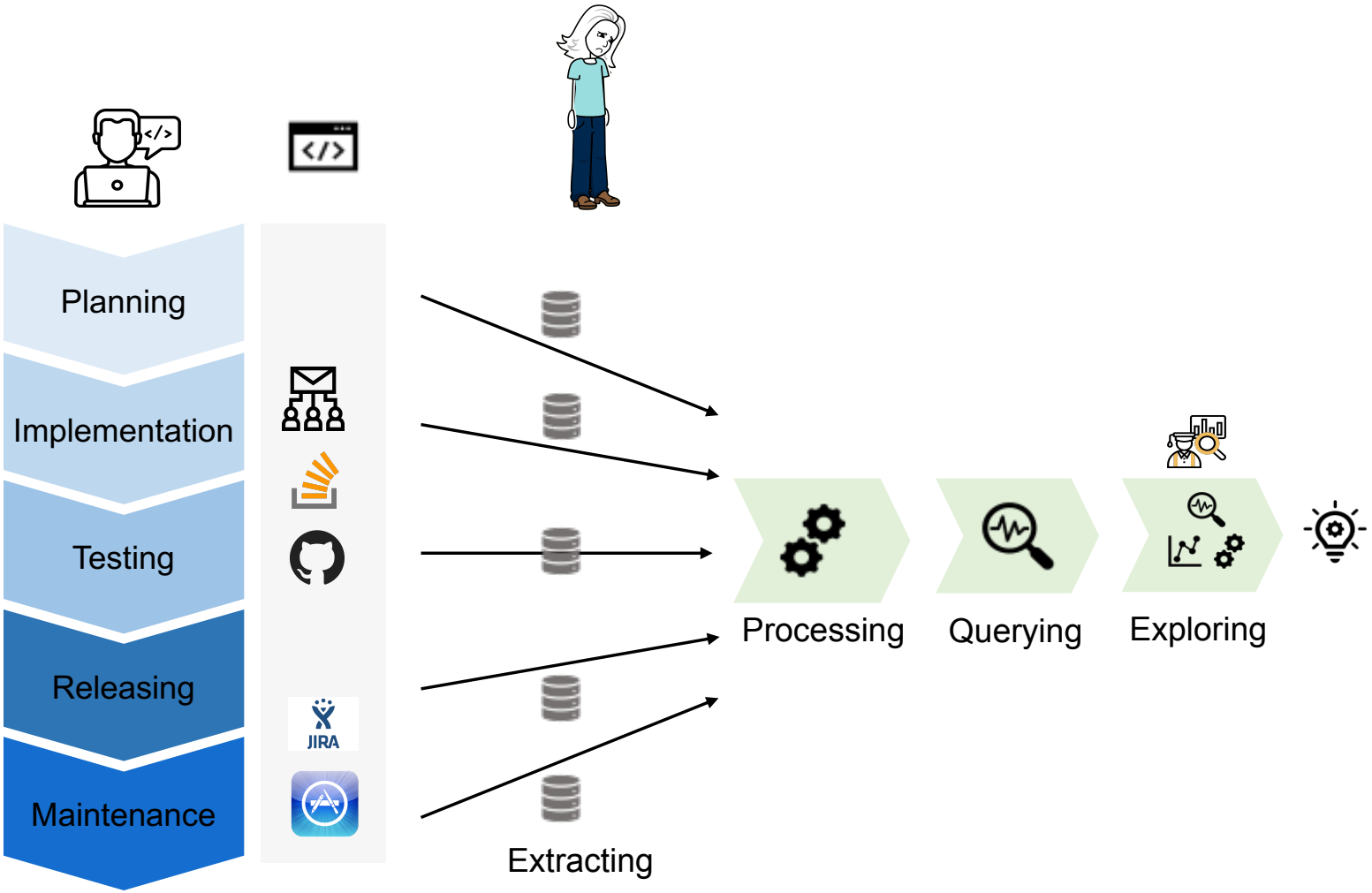



Do developers discuss code comments?

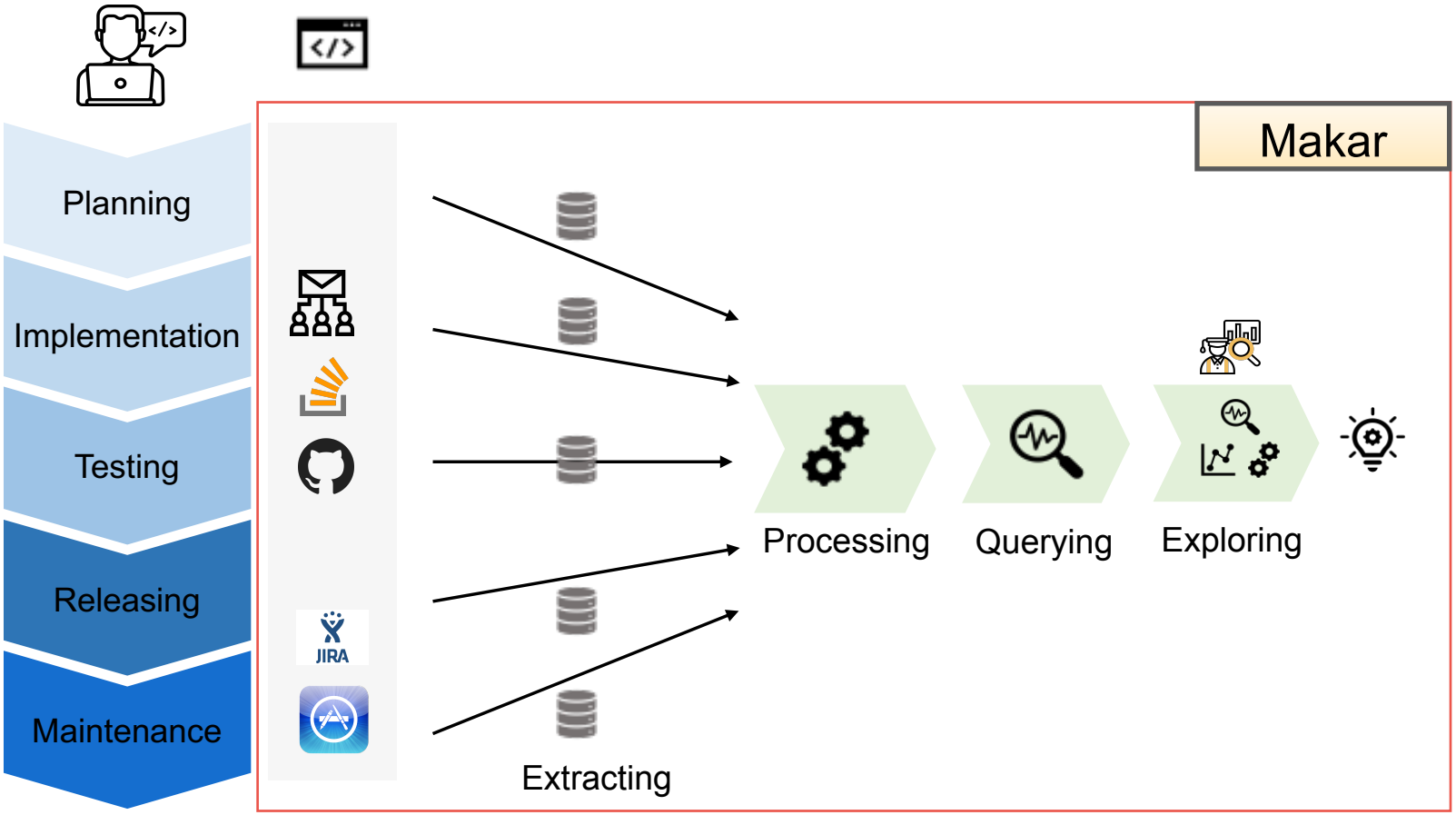
# Developers use various discussion sources



# Challenges


Planning
Implementation
Testing
Releasing
Maintenance
JIRA
Extracting
Processing
Querying
Exploring

# Makar: A tool for Multi-source Studies




https://github.com/maethub/makar

# Features

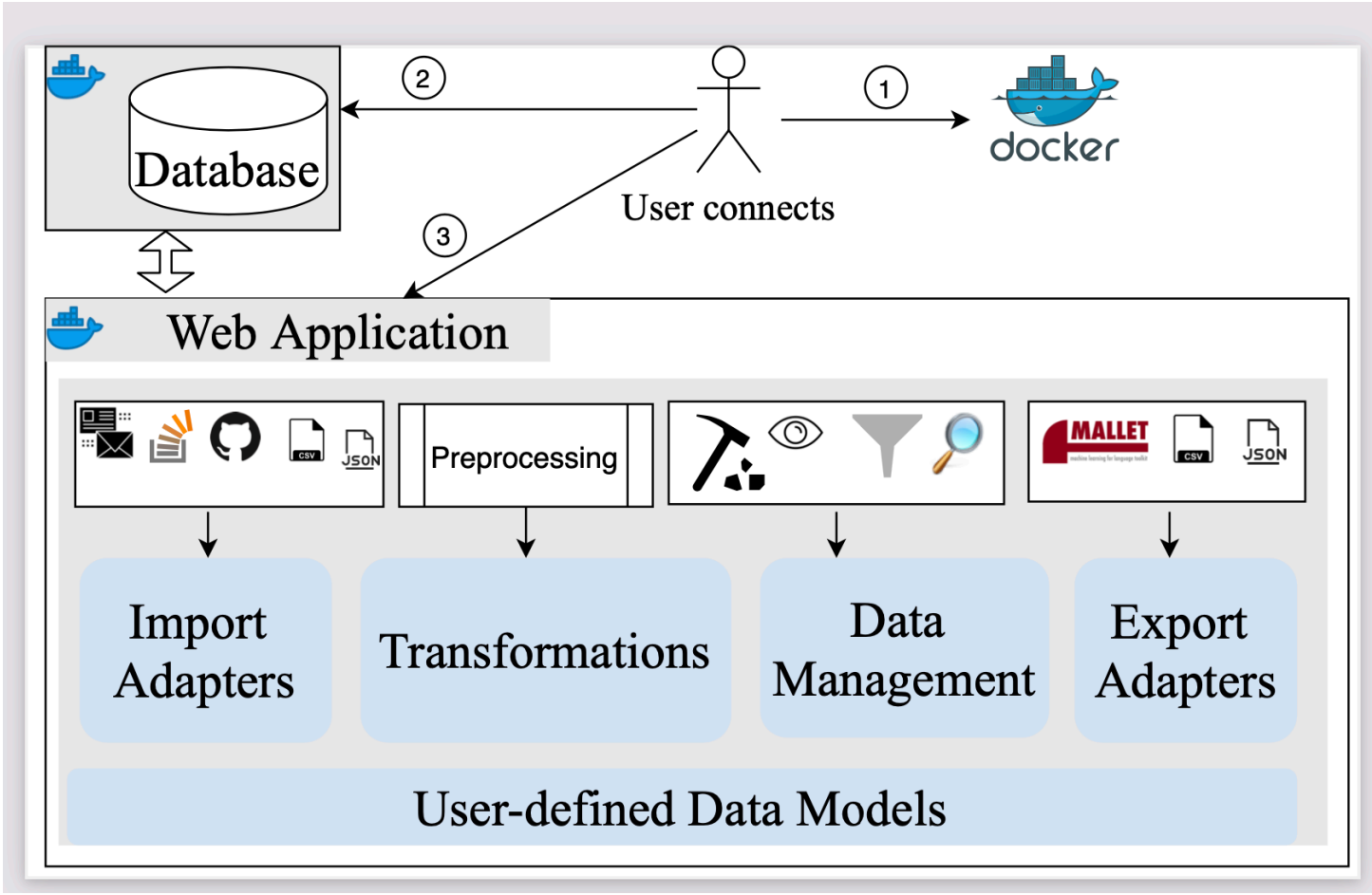
Extract data from different sources
e.g., Stack Overflow, Github, Mailing Lists

Support mapping and processing the data

Explore and perform ad-hoc searches

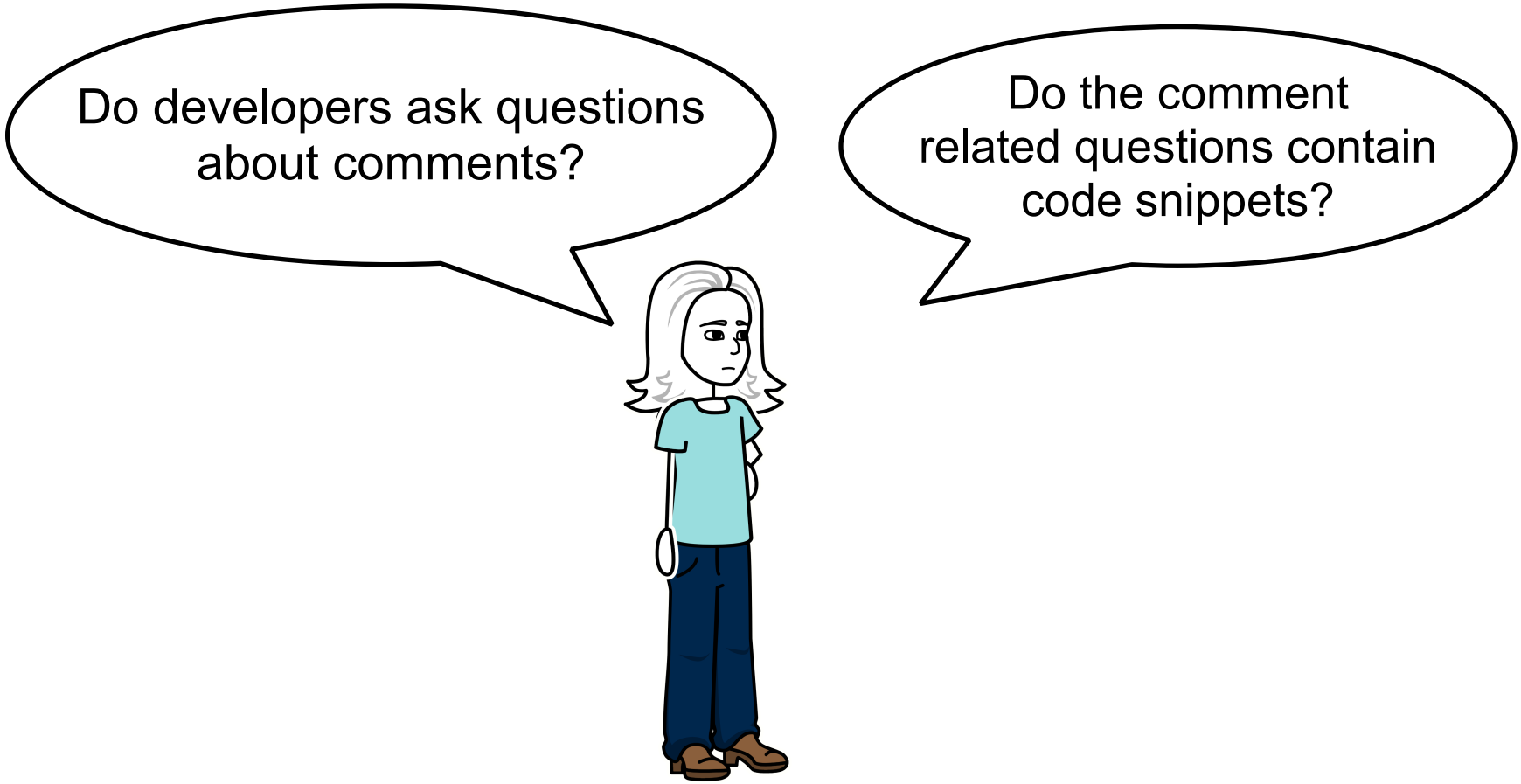
Extending the dataset easily

# Makar Architecture

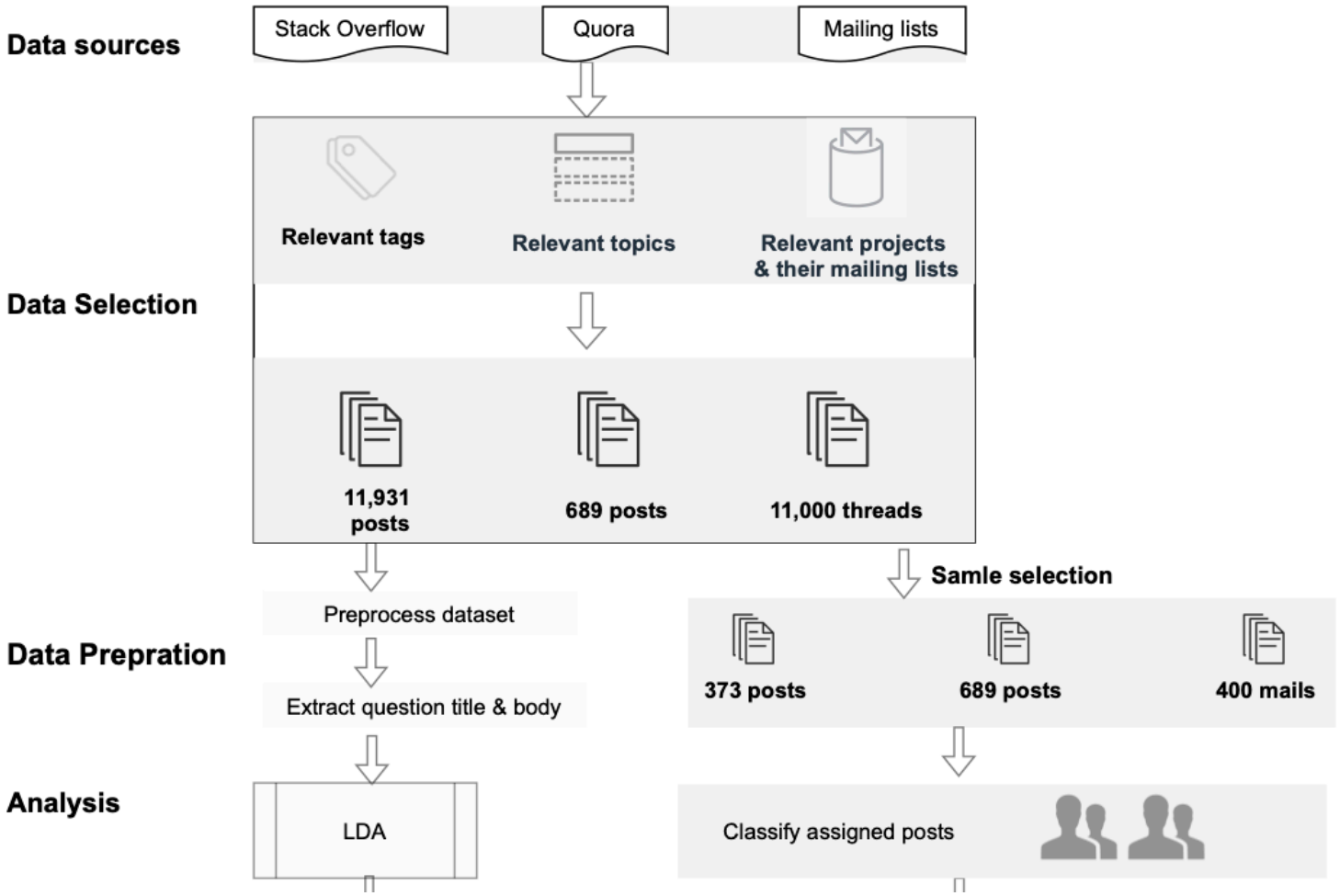



https://github.com/maethub/makar

# Demo


Do developers ask questions about comments?
Do the comment related questions contain code snippets?

# Case Study


Data sources
Stack Overflow
Quora
Mailing lists
Data Selection
Relevant tags
Relevant topics
Relevant projects & their mailing lists
11,931 posts
689 posts
11,000 threads
Samle selection
Data Prepration
Preprocess dataset
Extract question title & body
373 posts
689 posts
400 mails
Analysis
LDA
Classify assigned posts

# Case Study

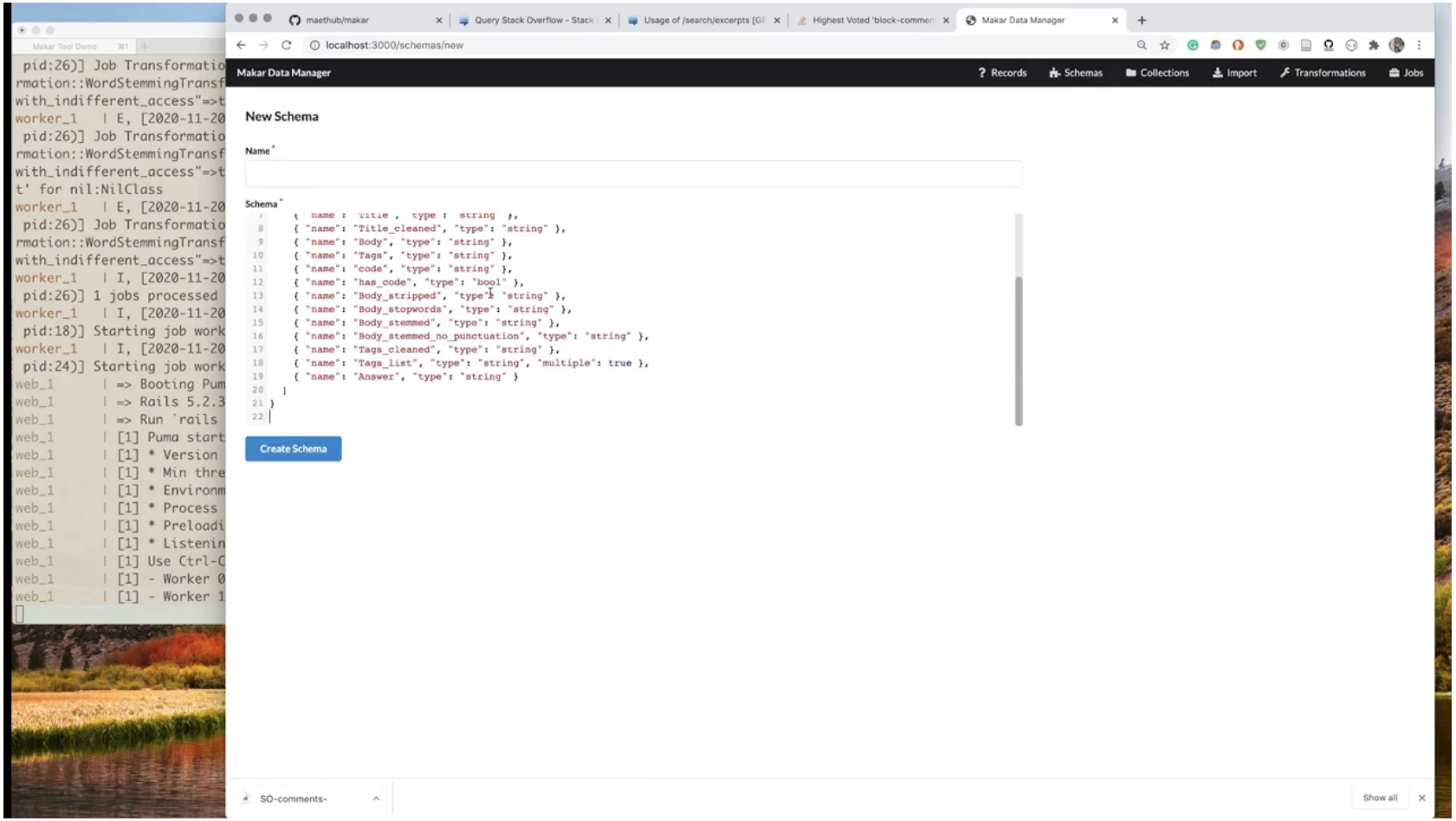
Import adapters: Stack overflow, CSV, Apache mailing list

Preprocess the data:


| Transformation | *extract_code* | *strip_html* | *string_replace* | *remove_stopwords* | *word_stemming* |
|---|---|---|---|---|---|
| **Attributes** | - Question \| Body | - Question \| Body | - Question \| Body<br>- Question \| Title | - Question \| Body<br>- Question \| Title | - Question \| Body<br>- Question \| Title |

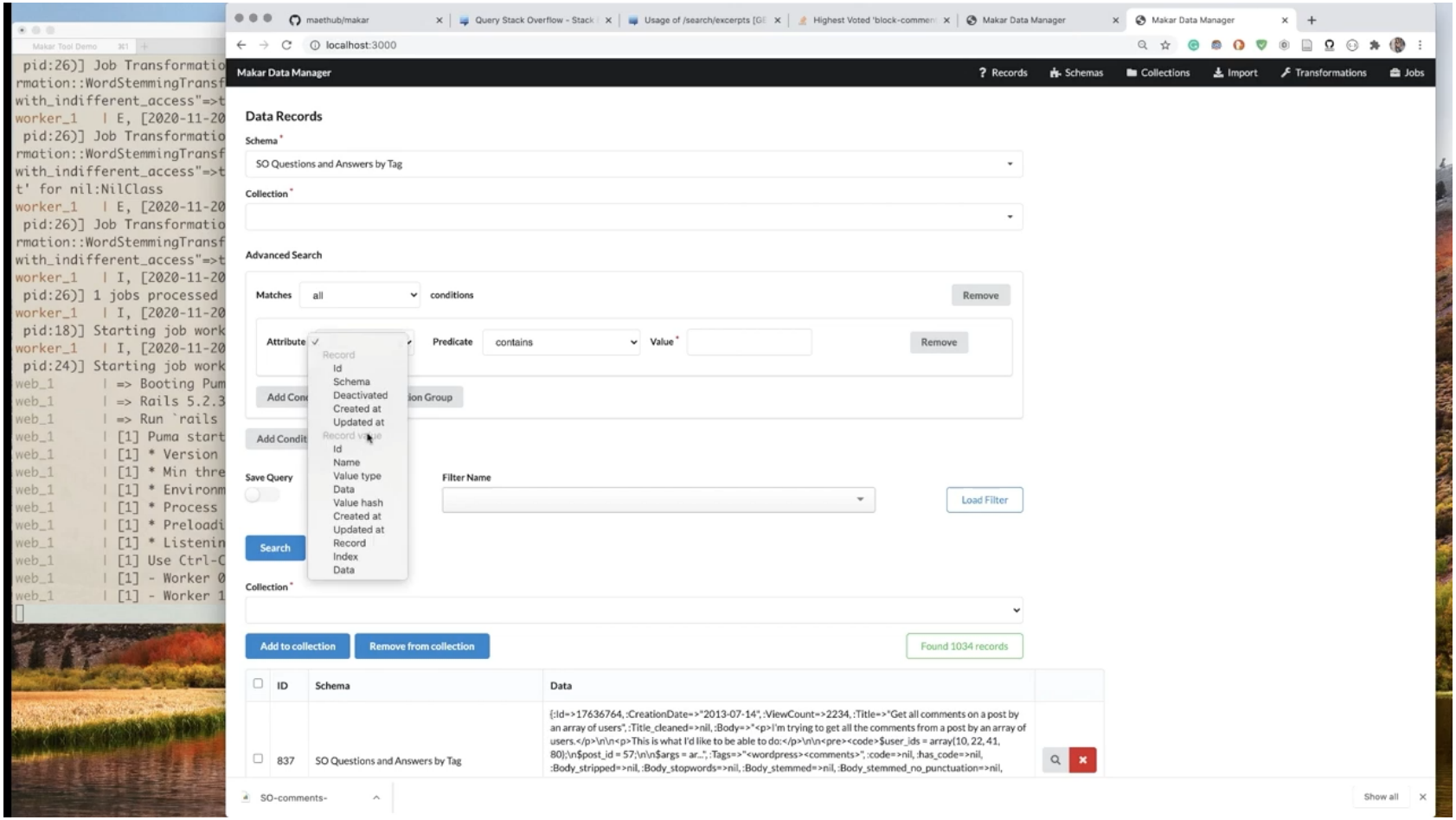
Export adapters: CSV adapters

# Features: Import



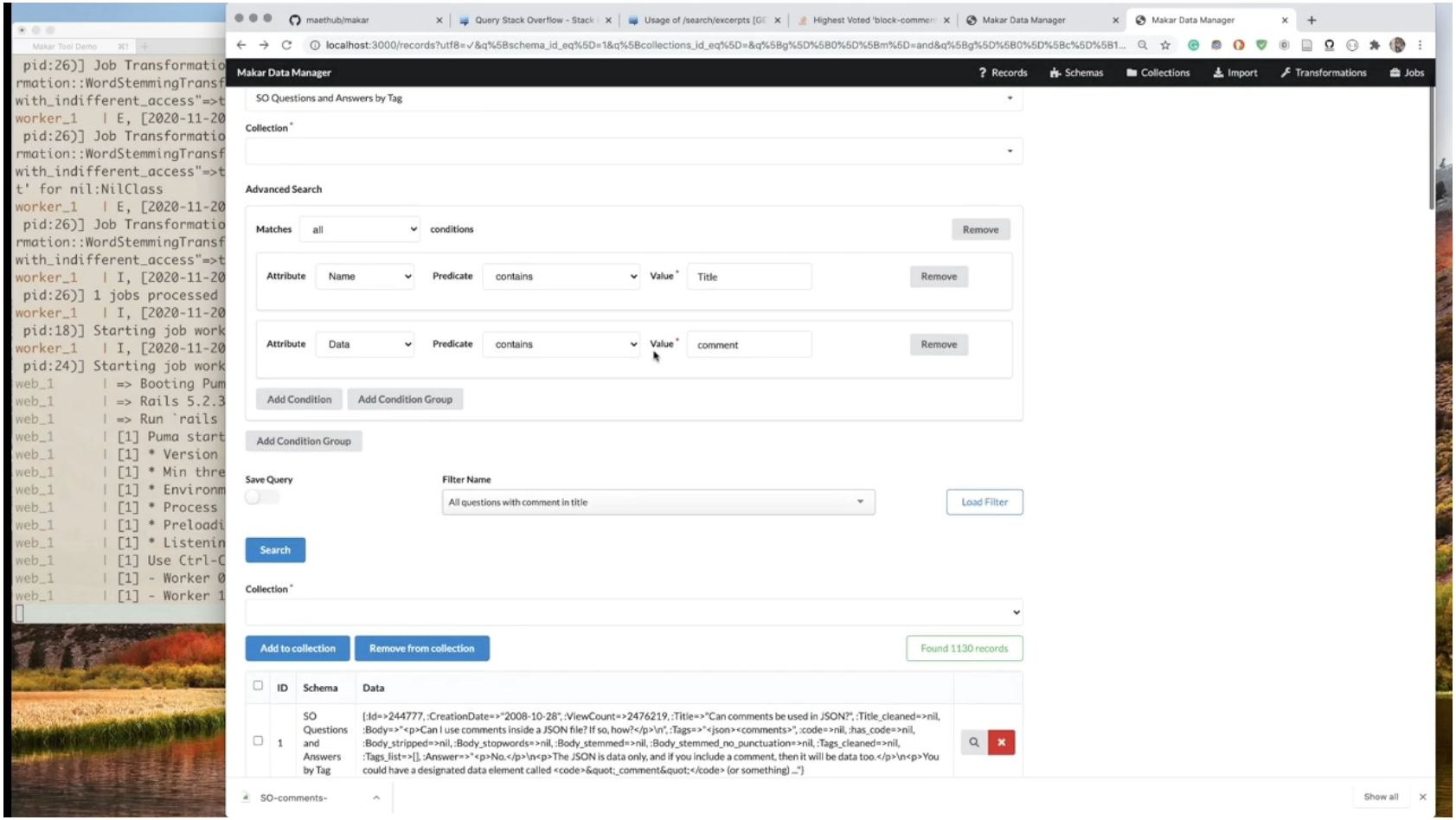

A complete demo is available at Youtube

# Features: Search



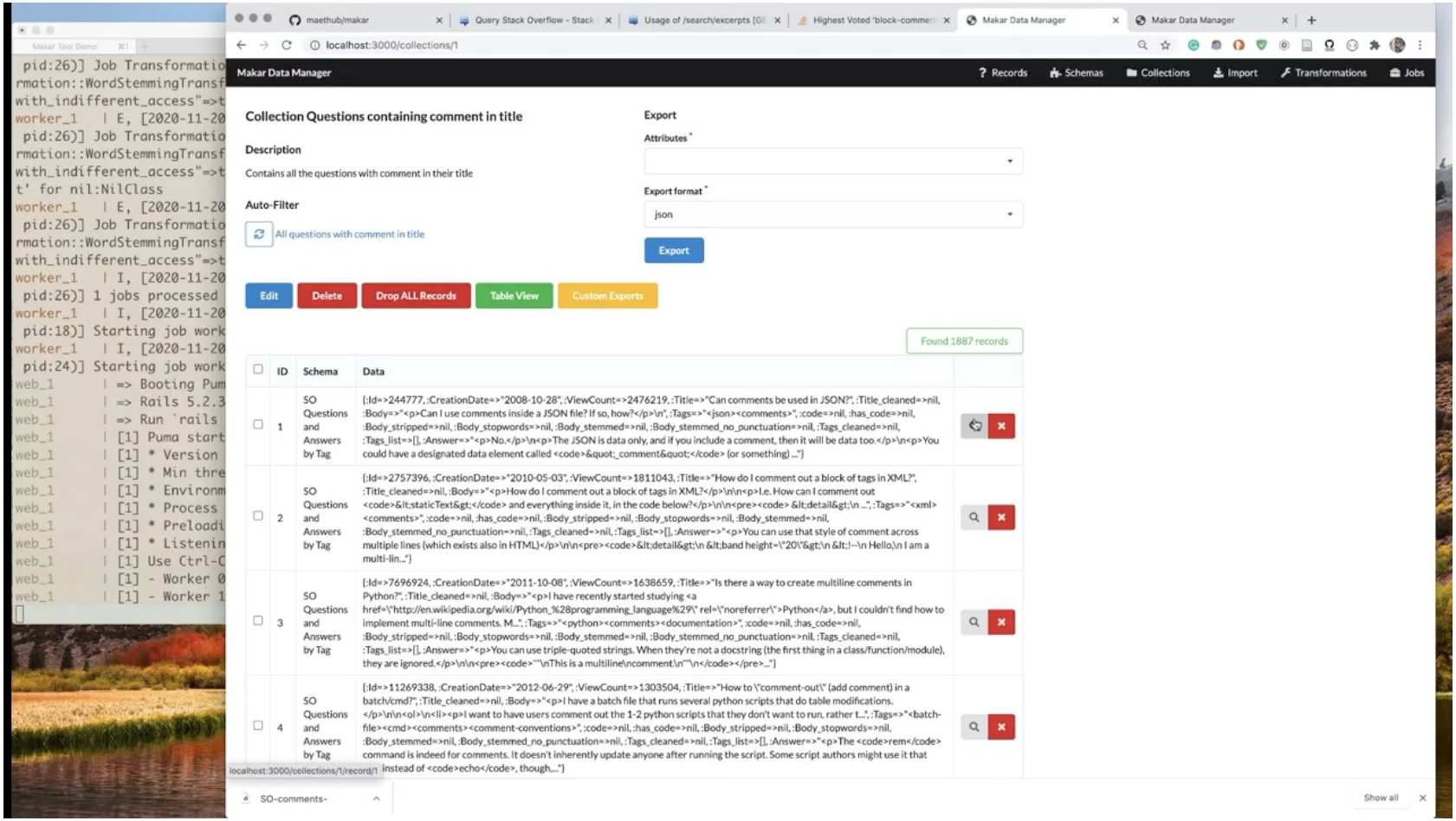

A complete demo is available at Youtube

# Features: Collection




A complete demo is available at Youtube

# Features: Transform




A complete demo is available at Youtube

# Tool Comparison

| Tool | Extract | Process | Manage |
|---|---|---|---|
| Octoparse | ✓ | ✗ | ✗ |
| Knime | Extension | ✓ | ✓ |
| Rapidminer | ✗ | Limited | Limited |
| ELKI | ✗ | ✓ | ✗ |
| Keel | ✗ | ✗ | ✗ |
| WEKA | ✗ | ✗ | ✗ |
| TrifactaWrangler | ✗ | ✓ | ✓ |
| Boa | Limited | ✓ | ✓ |
| OpenRefine | ✓ | ✓ | ✓ |
| Makar | ✓ | ✓ | ✓ |

# Future work

Extension of data source adapters

Building a UI of the study pipeline

Development of analysis and visualisation components

Facilitation of more multi-source studies

# Makar: A Framework for Multi-source Studies Based on Unstructured Data

Hosted on Github

https://github.com/maethub/makar

Demo at YouTube

https://youtu.be/Yqj1b4Bv-58

Replication Package at Zenodo

https://doi.org/10.5281/zenodo.4434822

Contact us

https://twitter.com/poojaruhal

http://scg.unibe.ch/staff/Pooja-Rani