CODE-SHARING DETECTION

Michael Monschau, Andreas Deisenau, Johannes Klöckner, Kevin Klein, Thomas Schmorleiz, ESE13 @ Uni Koblenz

RESEARCH QUESTION

How is code shared among Python contributions in the 101companies corpus?

HYPOTHESIS

HYPOTHESIS

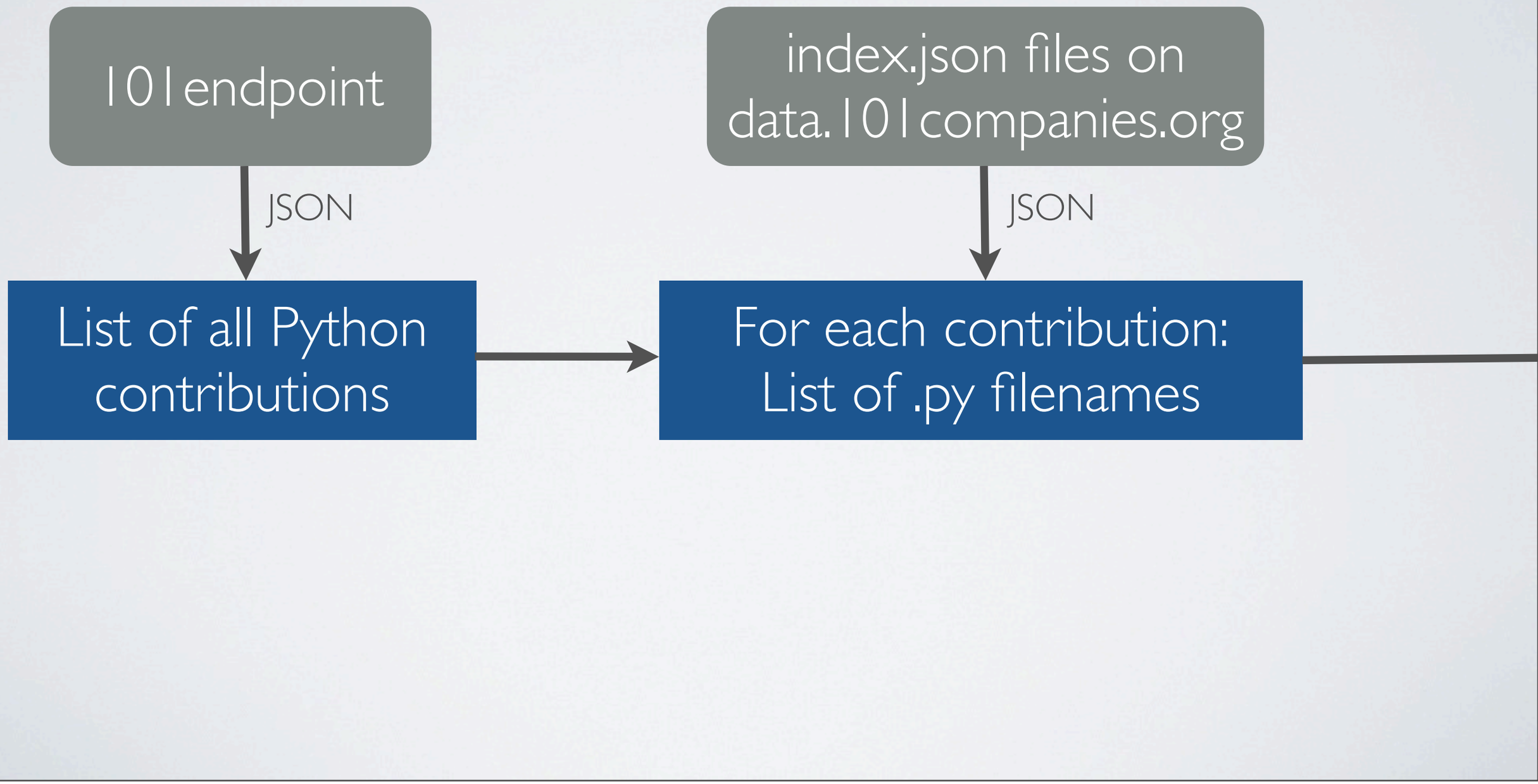
1. There are no attempts to modularize code between different contributions

2. All contributions were either partial clones, or were cloned from another contribution

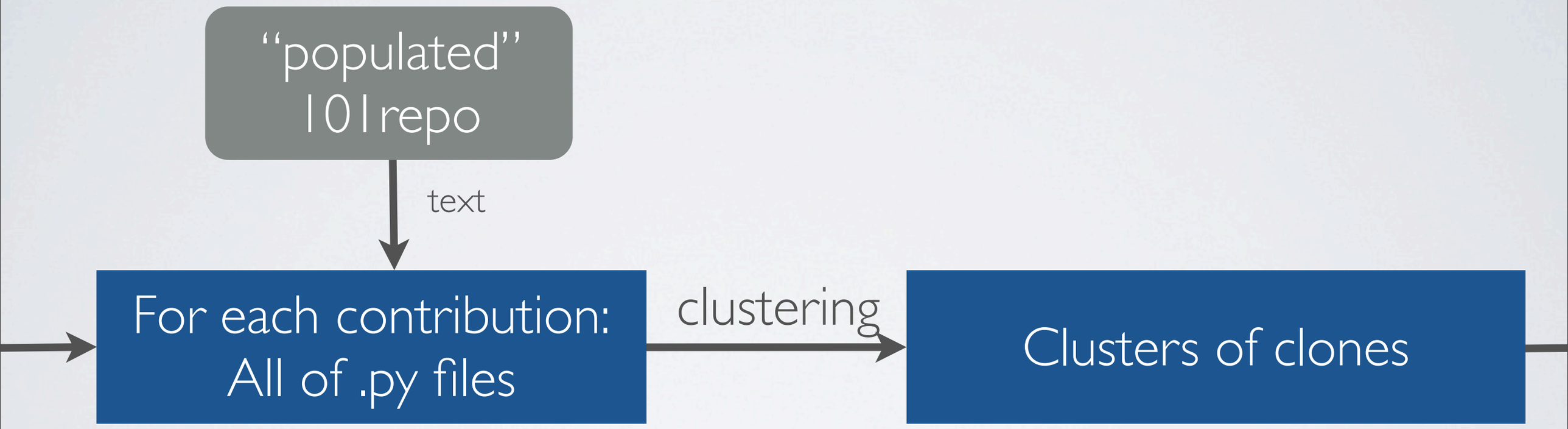
3. There is no single file that can be reused as-is by multiple contributions

4. For clones that result from copy-pasting there is no trivial way to utilize 101 fragments

DATA RETRIEVAL



DATA RETRIEVAL



COMPARING FILES

COMPARING FILES

- **Two metrics:**

1. diff based: Diffratio

2. fragment based: Ratio of matching classifiers

CLUSTERING

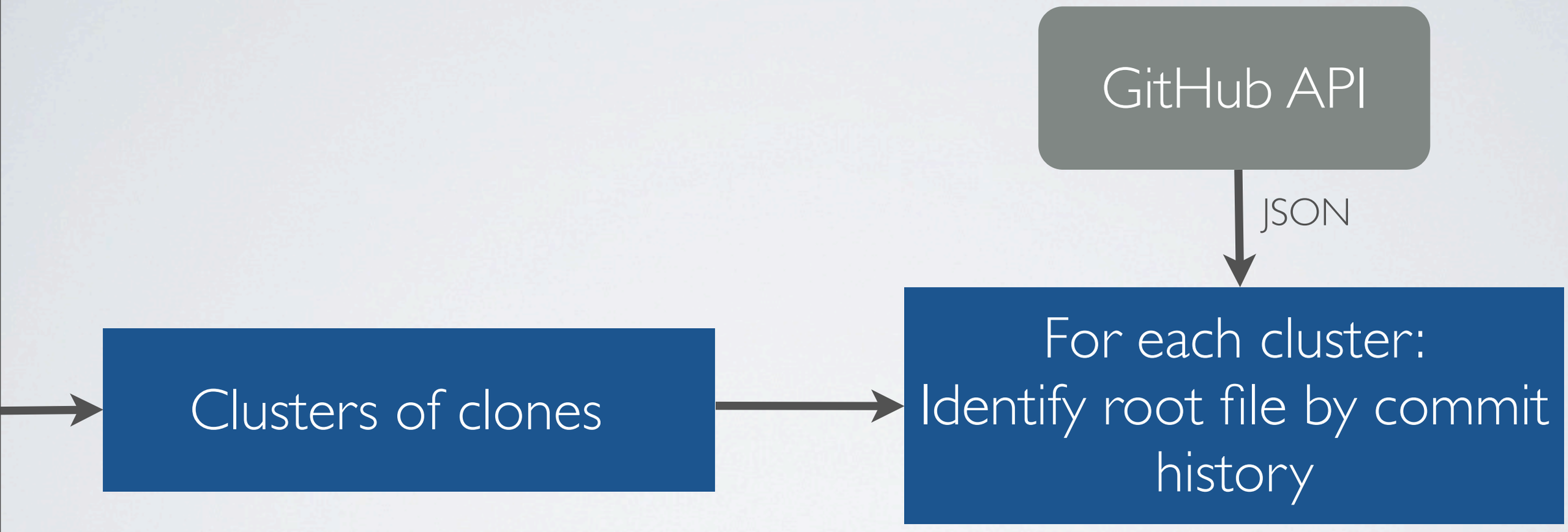
* for diff: 0.25, for fragments: 0.1

CLUSTERING

1. Iterate over all files, and for each file f:
    1. Create a new cluster for f
    2. Iterate over all exiting clusters
        1. Check if f fits in the cluster (i.e. if similarity ratio with all existing files is greater than threshold*)
2. Delete singleton clusters

* for diff: 0.25, for fragments: 0.1

DATA RETRIEVAL



DATA RESULTS: DIFF

DATA RESULTS: DIFF

- 5 clusters
- Only 2 / 7 contributions' files were not part of any clusters
- Some files were part of up to three clusters
- No entirely copied files

DATA RESULTS: FRAGMENTS

Problem: Slight changes breaks fragment based detection

DATA RESULTS: FRAGMENTS

- 1 cluster

Problem: Slight changes breaks fragment based detection

HYPOTHESIS

HYPOTHESIS

1. There are no attempts to modularize code between different contributions

2. All contributions were either partial clones, or were cloned from another contribution

3. There is no single file that can be reused as-is by multiple contributions

4. For clones that result from copy-pasting there is no trivial way to utilize 101 fragments

THREADS TO VALIDITY

- **Limited data volume:** Only 7 contributions
- **Low number of authors:** (4)
- **Noise for diff:** Renaming identifiers and formatting creates noise (Possible solution: normalize or compare AST)
- **Shallow comparison for fragments:** No recursive traversal strategy in place

OUTLOOK

- **Problem**: No files can be used as is for modularization.
- **Solutions**:
  - Refactoring: Factor out certain parts of files, shared across contributions
  - Provide base implementation (just data model) to be included in any new contribution