University of Sheffield

Software Repository Mining

Software Reengineering

(COM3523 / COM6523)

The University of Sheffield

Software Repositories

Collection of source code, compiled artifacts and software metadata

- Typically used for project collaboration and version control

Source Code repositories

- Github, Gitlab, IBM RTC

Not only source code

- Maven Central (Java), PyPi (Python), NPM (Node JS)

Software Reengineering

What has this got to do with reengineering?

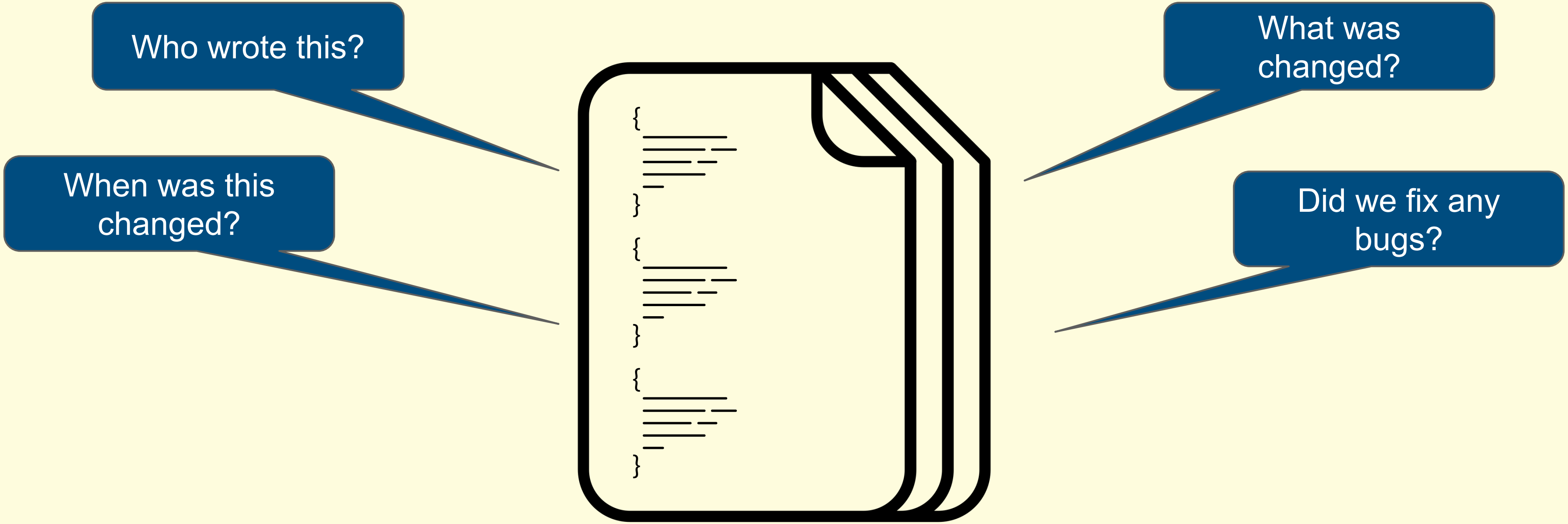
Key Takeaways

What is software repository metadata?

How can we obtain metadata?

How can we use metadata?

Repository Metadata



Who wrote this?

When was this changed?

What was changed?

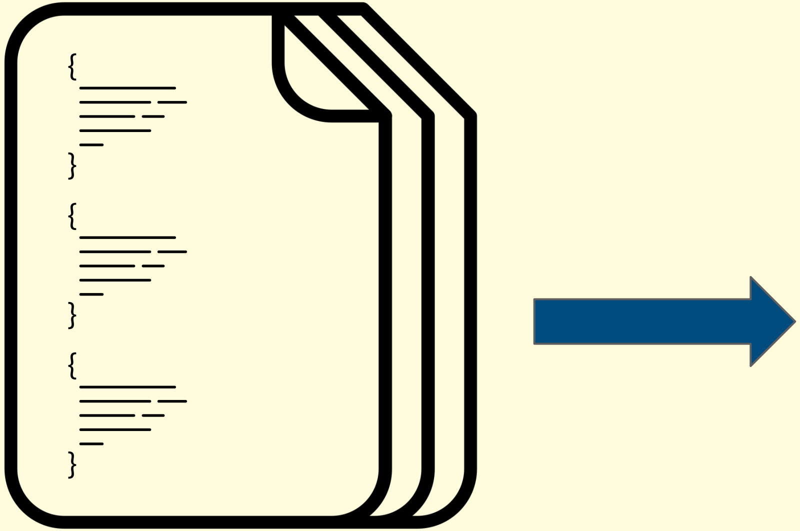
Did we fix any bugs?

Committed Change

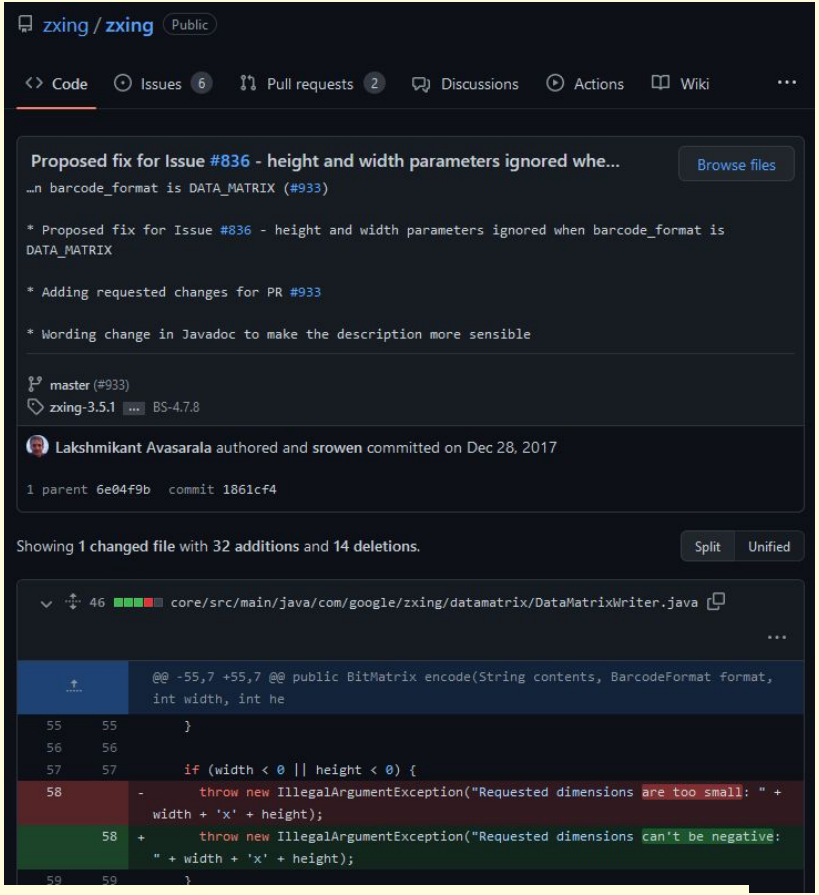
What are some examples of source-code metadata?

ⓘ Start presenting to display the poll results on this slide.

Repository Metadata



Committed Change



Proposed fix for Issue #836 - height and width parameters ignored whe…

```
…n barcode_format is DATA_MATRIX (#933)

* Proposed fix for Issue #836 - height and width parameters ignored when barcode_format is
DATA_MATRIX

* Adding requested changes for PR #933

* Wording change in Javadoc to make the description more sensible
```

master (#933)

zxing-3.5.1 … BS-4.7.8

Lakshmikant Avasarala authored and srowen committed on Dec 28, 2017

1 parent 6e04f9b commit 1861cf4

Showing 1 changed file with 32 additions and 14 deletions.

46 core/src/main/java/com/google/zxing/datamatrix/DataMatrixWriter.java

```
@@ -55,7 +55,7 @@ public BitMatrix encode(String contents, BarcodeFormat format, int width, int he
        }

        if (width < 0 || height < 0) {
-         throw new IllegalArgumentException("Requested dimensions are too small: " + width + 'x' + height);
+         throw new IllegalArgumentException("Requested dimensions can't be negative: " + width + 'x' + height);
        }
```

Repository Metadata Metrics

Previously explored codebase metrics

- Lines of code, cyclomatic complexity, object coupling

Repository analysis produce ‘Process Metrics’

- Number of commits, lines added, active developers
- Allows for active/legacy projects to be identified
- Identification of problem areas in the codebase

Repository Mining using Bash

git log **output the logs for a git repository**

```
git log --no-merges --pretty=format:"%h"
```

Ignore merge commits

Formatting

git show 1861cf4 **will output the metadata for a given commit hash**

```
git show -s --format='%an' 1861cf4
```

Only output author metadata

Repository Mining using Bash

Git bash demo

Using Mined Metadata

We found each commit for a project

- Each commit was mined for metadata
- Generated a large csv of each atomic change

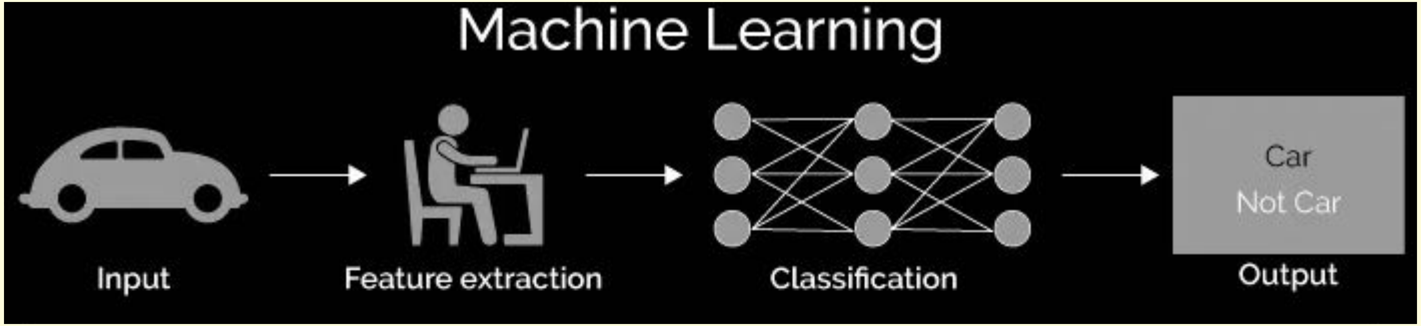
Explored the use of Pivot Tables to combine data

- We are able to identify files of interest through number of commits
- Also able to find active members of the repository

Many more complex tools to use outside of bash…

Cutting Edge Uses for Mined Metadata

Now we have the data, what do we do with it?



Recent work into fault prediction software

- How should a company best divide its programmer budget?

Unfinished work classification

- Did employee X leave before finishing their pull request?

Key Takeaways

What is software repository metadata?

Information about the software program outside of the code itself

How can we obtain metadata?

We explored the use of bash git commands to obtain metadata

How can we use metadata?

Combining data to make the sheer quantity of data

Covered some cutting edge techniques

slido

Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.