The University Of Sheffield.

Repository Analysis

Software Reengineering
(COM3523 / COM6523)

The University of Sheffield

Version Repositories

Software development can involve hundreds or thousands of developers.

- Often working asynchronously, from different parts of the globe.

Version repositories manage these changes.

Every clone of a repository includes entire history of code changes.

- **A valuable data-set for exploring the evolution of the software system.**
- **Often come with powerful command-line interfaces.**

Patches

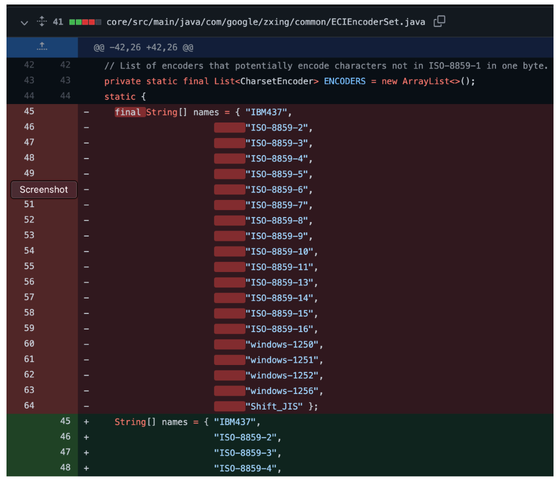
The contents of a commit in Git.

Each patch can affect one or more files.

A set of lines of code that are either added or deleted.

- A change to a line is achieved by deleting it, and adding the changed version.

Can include the creation of new files, or the removal of files.



```
41 core/src/main/java/com/google/zxing/common/ECIEncoderSet.java
@@ -42,26 +42,26 @@
    // List of encoders that potentially encode characters not in ISO-8859-1 in one byte.
    private static final List<CharsetEncoder> ENCODERS = new ArrayList<>();
    static {
-     final String[] names = { "IBM437",
-                              "ISO-8859-2",
-                              "ISO-8859-3",
-                              "ISO-8859-4",
-                              "ISO-8859-5",
-                              "ISO-8859-6",
-                              "ISO-8859-7",
-                              "ISO-8859-8",
-                              "ISO-8859-9",
-                              "ISO-8859-10",
-                              "ISO-8859-11",
-                              "ISO-8859-13",
-                              "ISO-8859-14",
-                              "ISO-8859-15",
-                              "ISO-8859-16",
-                              "windows-1250",
-                              "windows-1251",
-                              "windows-1252",
-                              "windows-1256",
-                              "Shift_JIS" };
+     String[] names = { "IBM437",
+                        "ISO-8859-2",
+                        "ISO-8859-3",
+                        "ISO-8859-4",
```

Useful information about the system

Which files do developers work on most frequently?

- Tells us which areas are particular important, or problematic.

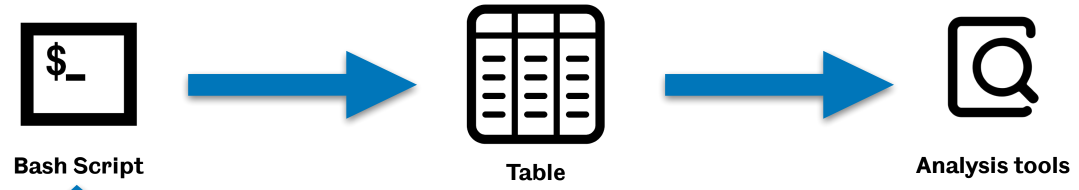
Which files were most associated with bug fixes?

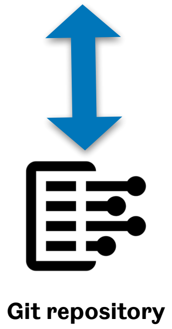
- Which areas of the system are weak, perhaps need some re-design?

Which files are most frequently changed at the same time?

- Which areas are probably related to each other?

Process





Get the “raw” data into an accessible form.

- A database or a table.

Separate your data extraction from your data analysis.

- Choose the best tool for the type of data and analysis you have.

git show

`git show` will extract *any* information you need about a commit in git.

Documentation available at: https://git-scm.com/docs/git-show

Can extract a single piece of data as follows:

```
git show -s —format=‘placeholder’ commit_hash_code
git show -s —format=‘%ci’
```

Shows the date as a Unix timestamp

Can also extract statistics for numbers of lines added / removed:

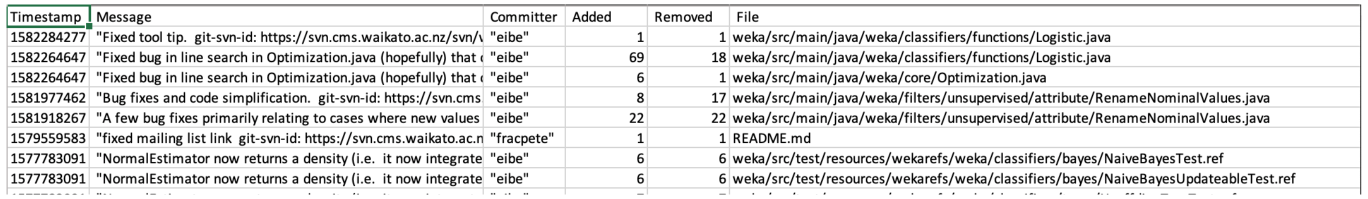
```
git show —numstat commit_hash_code
```

git show

Storage in a table

Attributes in the columns.

Each entry is a row.



Timestamp	Message	Committer	Added	Removed	File
1582284277	"Fixed tool tip. git-svn-id: https://svn.cms.waikato.ac.nz/svn/v	"eibe"	1	1	weka/src/main/java/weka/classifiers/functions/Logistic.java
1582264647	"Fixed bug in line search in Optimization.java (hopefully) that (	"eibe"	69	18	weka/src/main/java/weka/classifiers/functions/Logistic.java
1582264647	"Fixed bug in line search in Optimization.java (hopefully) that (	"eibe"	6	1	weka/src/main/java/weka/core/Optimization.java
1581977462	"Bug fixes and code simplification. git-svn-id: https://svn.cms.	"eibe"	8	17	weka/src/main/java/weka/filters/unsupervised/attribute/RenameNominalValues.java
1581918267	"A few bug fixes primarily relating to cases where new values	"eibe"	22	22	weka/src/main/java/weka/filters/unsupervised/attribute/RenameNominalValues.java
1579559583	"fixed mailing list link git-svn-id: https://svn.cms.waikato.ac.n	"fracpete"	1	1	README.md
1577783091	"NormalEstimator now returns a density (i.e. it now integrate	"eibe"	6	6	weka/src/test/resources/wekarefs/weka/classifiers/bayes/NaiveBayesTest.ref
1577783091	"NormalEstimator now returns a density (i.e. it now integrate	"eibe"	6	6	weka/src/test/resources/wekarefs/weka/classifiers/bayes/NaiveBayesUpdateableTest.ref

Summarising combinations of variables

Our “raw” CSV file is big.

- Every “atomic” change to a file has its own row.

Need to group and summarise changes to obtain useful summaries.

Lots of tools to do this - pick your favourite!

- Python - framworks such as Pandas can aggregate and summarise.
- R - Plyr, reshape2, etc.
- Excel - Pivot tables…

Key steps:

- (1) Select your “grouping” variables.
- (2) Select your “summary” operation to carry out on the grouped variables - to sum, to average, etc.

Key take-aways

Version repositories contain an extensive history of source code change.

- Can identify frequently changed files, active developers, co-changes, etc.

Can be particularly powerful when combined with other data sources.

- Information about file-sizes, speculative design documents, etc.

Often have powerful command-line interfaces.

- Relatively easy to mine with Bash scripts.