

# Mining Software Repositories

John Businge  
Henrique Rocha

# References



## **The Road Ahead for Mining Software Repositories**

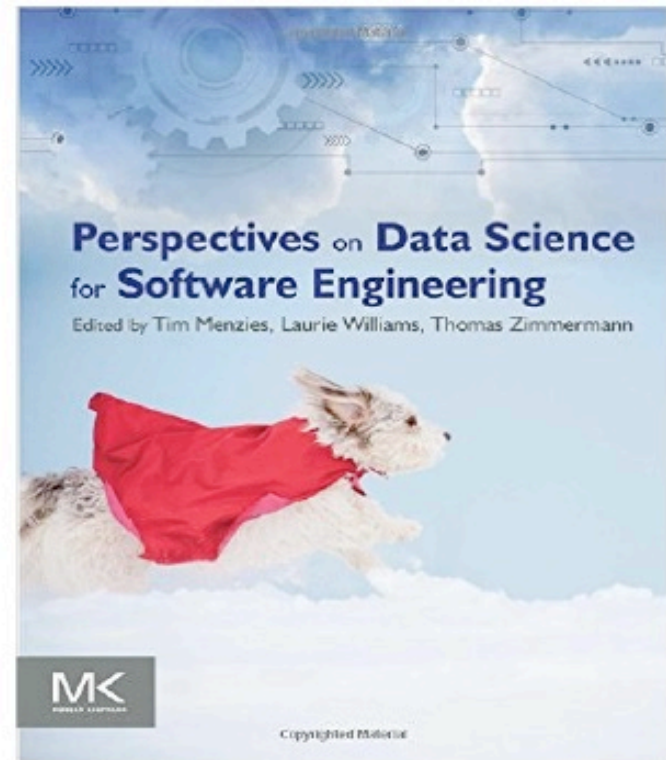
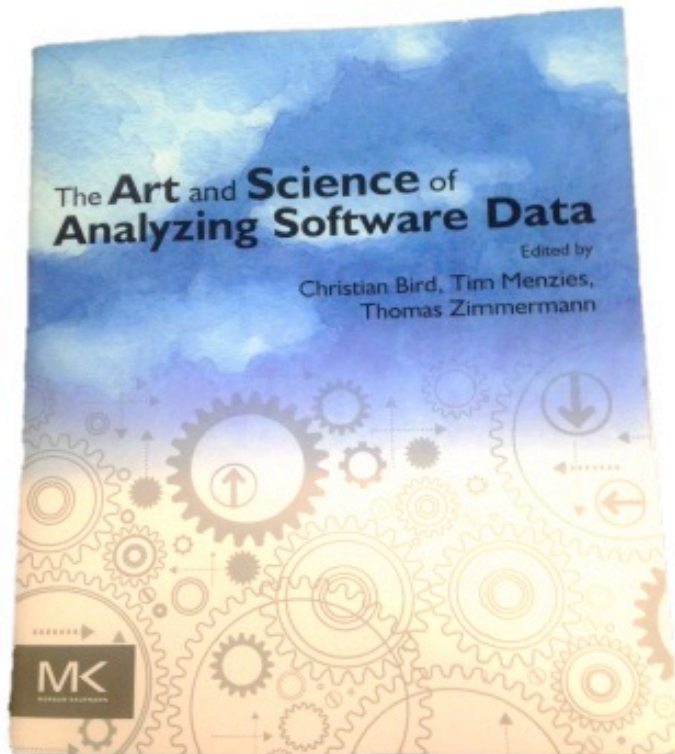
Ahmed E. Hassan  
Software Analysis and Intelligence Lab (SAIL)  
School of Computing, Queen's University, Canada  
ahmed@cs.queensu.ca

## **Software Intelligence: The Future of Mining Software Engineering Data**

Ahmed E. Hassan  
School of Computing  
Queen's University  
Kingston, ON, Canada  
ahmed@cs.queensu.ca

Tao Xie  
Department of Computer Science  
North Carolina State University  
Raleigh, NC, USA  
xie@csc.ncsu.edu

# More References



# Acknowledgement



**Ahmed E. Hassan**

Queen's University

[www.cs.queensu.ca/~ahmed](http://www.cs.queensu.ca/~ahmed)

[ahmed@cs.queensu.ca](mailto:ahmed@cs.queensu.ca)

**Tao Xie**

North Carolina State University

[www.csc.ncsu.edu/faculty/xie](http://www.csc.ncsu.edu/faculty/xie)

[xie@csc.ncsu.edu](mailto:xie@csc.ncsu.edu)

**Bram Adams**

Polytechnique Montréal, Canada

<http://mcis.polymtl.ca/index.html>

[lab.mcis@gmail.com](mailto:lab.mcis@gmail.com)

With updates by **John Businge** from University of Antwerp, Belgium

# Lecture Goals



- **Learn about:**

- Classic and notable research and researchers in mining SE data
- Data mining and data processing techniques and how to apply them to SE data
- Risks in using SE data due to e.g., noise

- **After the lecture, you should be able to:**

- Retrieve SE data
- Prepare SE data for mining
- Mine interesting information from SE data

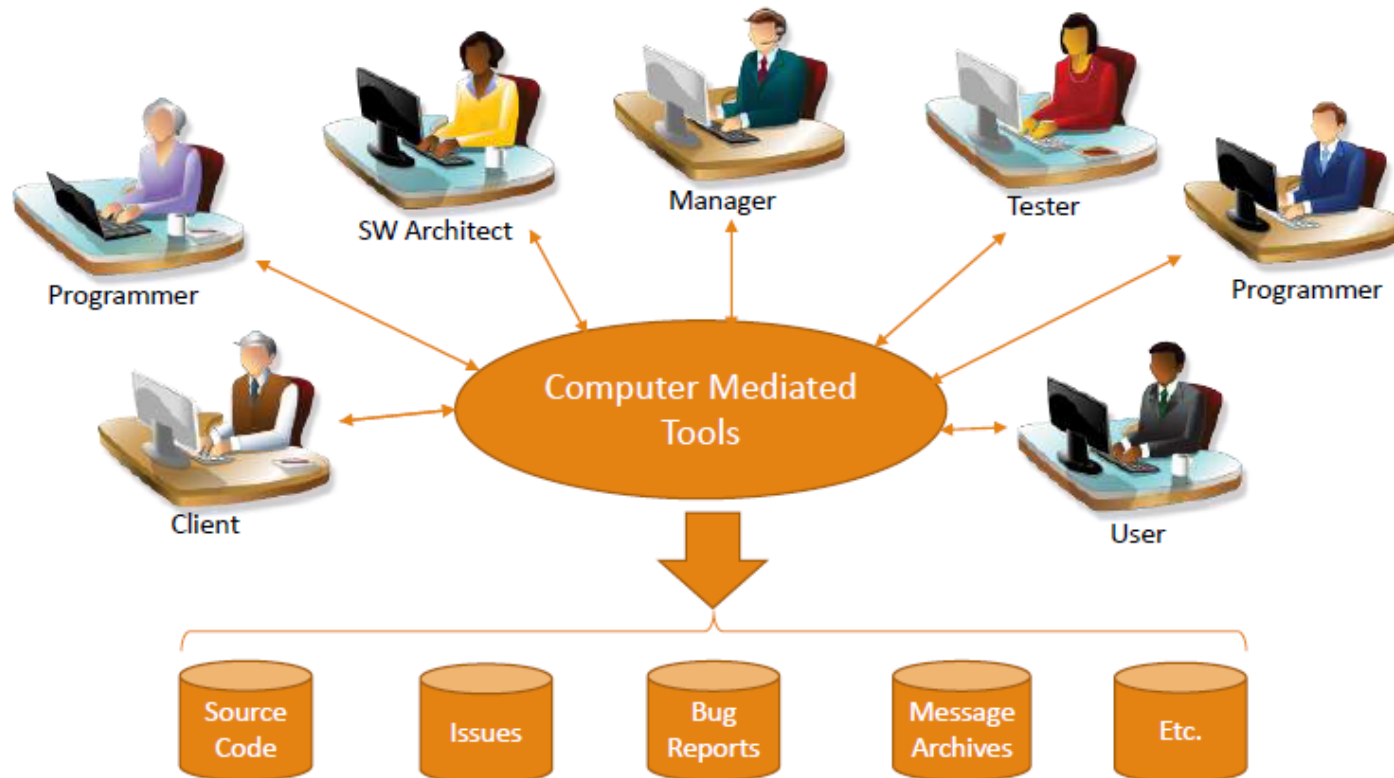
# Why mine SE data?

- **SE data can be used to:**

- Gain empirically-based understanding of software development
- Predict, plan, and understand various aspects of a project
- Support future development and project management activities



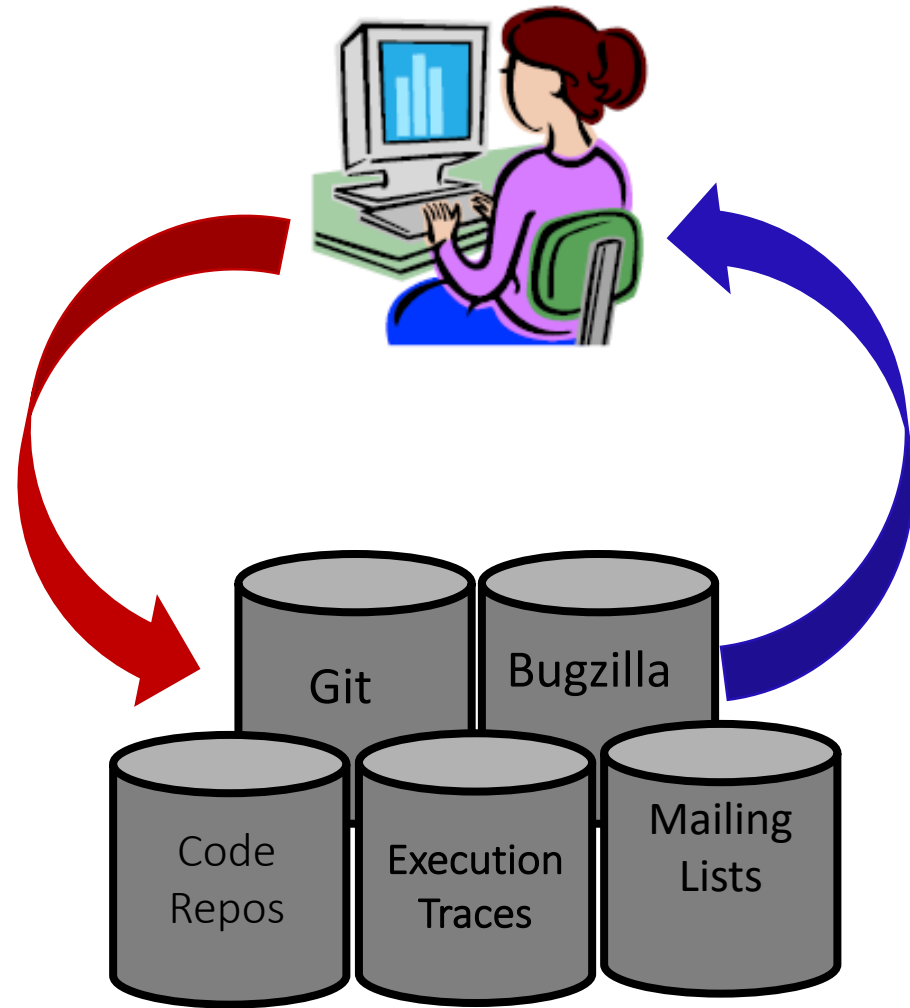
# How is SE Data generated?



Current and historical artifacts and interactions are registered in software repositories

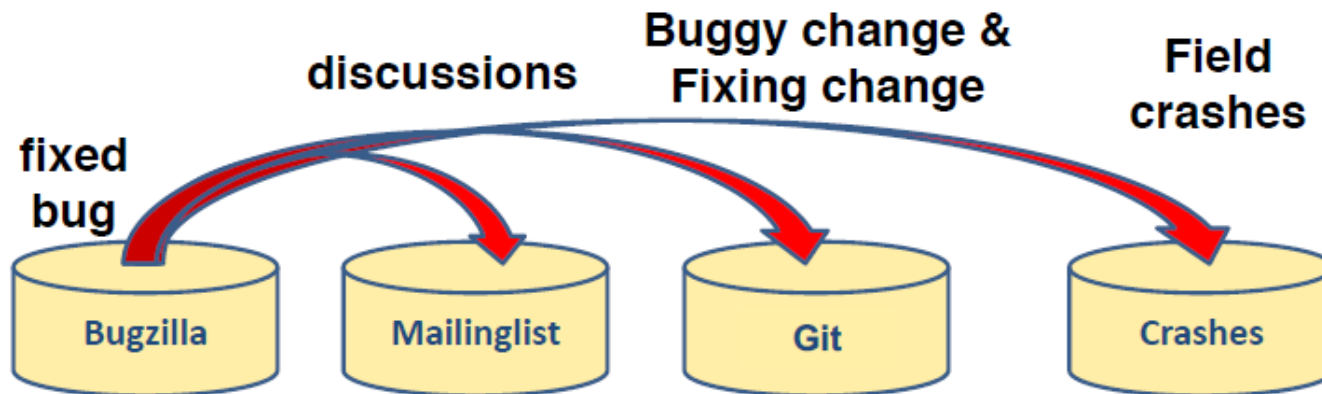
# What is MSR?

- Transforming static record keeping SE into active data
- Making SE data actionable by uncovering patterns and trends





# MSR researchers analyze and cross-link repositories



New bug report  
Estimate fix effort  
Suggest experts and fix!

# Study Outline



- **Part I:** What can we learn from SE data?
  - A sample of notable findings for different SE data types
- **Part II:** How can we mine SE data?
  - Understand the structure of SE data

# MSR studies – Bugs – Part I

## Using imports to predict Bugs

71% of files that import **compiler** packages,  
had to be fixed later on.

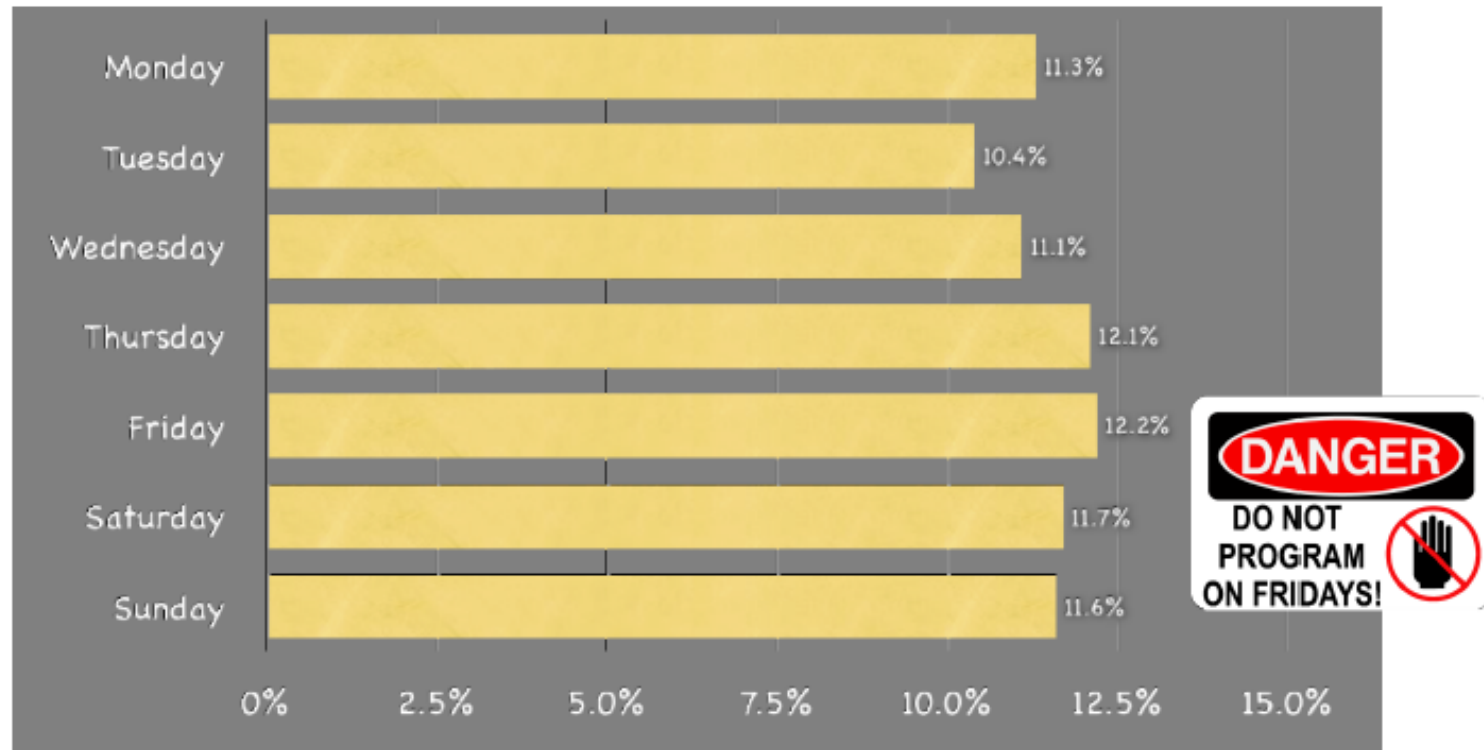
```
import org.eclipse.jdt.internal.compiler.lookup.*;
import org.eclipse.jdt.internal.compiler.*;
import org.eclipse.jdt.internal.compiler.ast.*;
import org.eclipse.jdt.internal.compiler.util.*;
...
import org.eclipse.pde.core.*;
import org.eclipse.jface.wizard.*;
import org.eclipse.ui.*;
```

[Schröter et al. 06]

14% of all files that import **ui** packages, had  
to be fixed later on.

# MSR studies - Bugs

Do not program on Friday ;-)

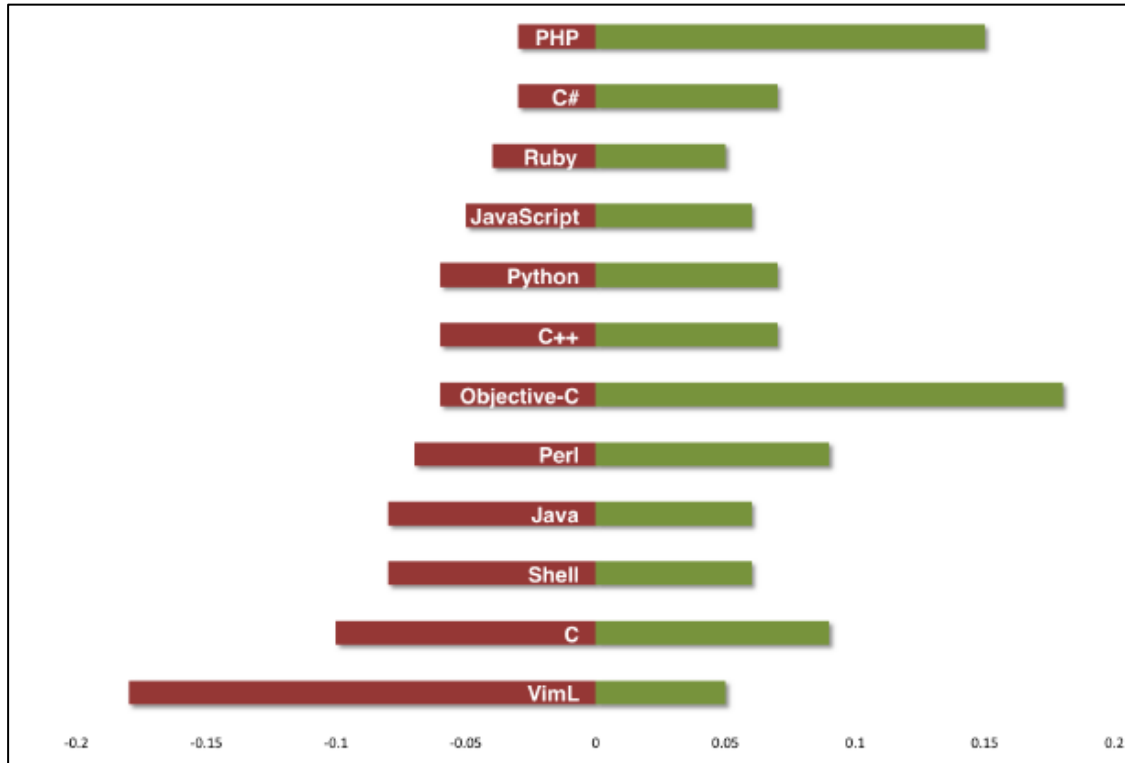


Percentage of bug-introducing changes for eclipse

[Zimmermann et al. 05]

# MSR studies – Sentiment Analysis

Anger vs. Joy



How they stack up?

- **PHP**, **Object-C**, and **C#** are net positive
- **Java**, **Shell**, and **C** are fairly even while **VimL** is just bad news.

[Doll and Grigorik, 2012]

# MSR studies – Sentiment Analysis



10/23/12 3:56 AM

Disable 'showmatch' option Matching parens  
are highlighted even without this option;  
what it does is jump the cursor to the  
matching paren which is [REDACTED] insane.



10/23/12 3:28 AM

styles everywhere, tutorial for first user  
login, fixing some css [REDACTED]



10/23/12 2:57 AM

[REDACTED] again



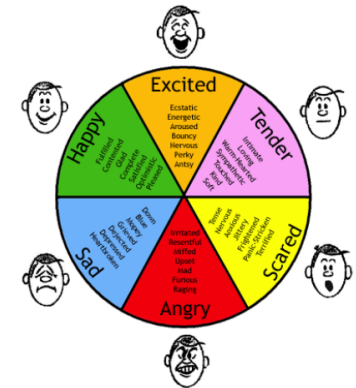
10/23/12 2:30 AM

more [REDACTED]



10/23/12 2:29 AM

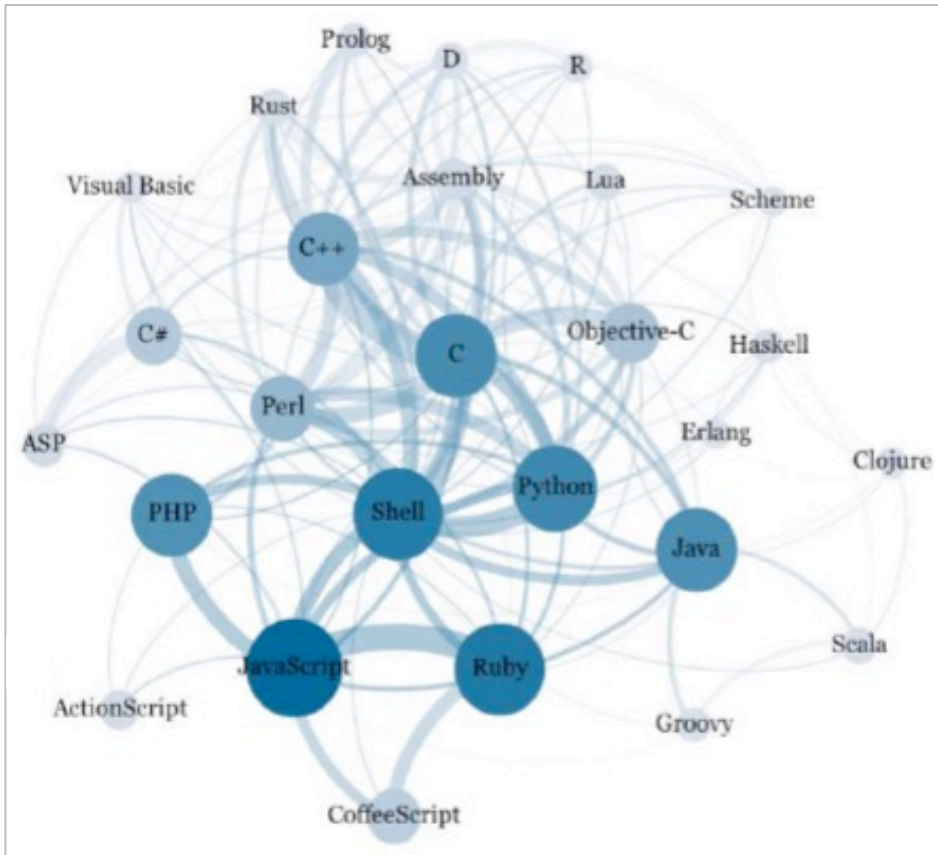
Security [REDACTED] worked out



<http://www.commitlogsfromlastnight.com/>

# MSR studies – Programming languages

## Programming language relations



A **Ruby** programmers is **very likely to know Javascript**, while a **Perl** programmer is not

**Java** is a popular programming language but stands primarily alone

<https://github.com/mjwillson/ProgLangVisualise>

# MSR studies – Changes by programmers

## Programming language relations

A) The user inserts a new preference into the field fKeys[]

B) ROSE suggests locations for further changes, e.g. the function initDefaults()

```
public final OverlayPreferenceStore OverlayKey[] fKeys = new OverlayPreferenceStore OverlayKey[] {
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, OPEN_STRUCTURE.COM,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, SYNCHRONIZE_SCROLL,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, SHOW_PSEUDO_CONFLI,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, INITIALLY_SHOW_ANC,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, SHOW_MORE_INFO),
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, IGNORE_WHITESPACE),
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, PREF_SAVE_ALL_EDIT,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, NEW_PREFERENCE),
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.STRING, AbstractTextEditor,
    new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, AbstractTextEditor,
    //new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.SINGLELINE, USE_SPLINES),
    //new OverlayPreferenceStore OverlayKey(OverlayPreferenceStore.BOOLEAN, USE_SINGLE_LINE),
};

public static void initDefaults(IPreferenceStore store) {
    store.setDefault(OPEN_STRUCTURE.COMPAR, true);
    store.setDefault(SYNCHRONIZE_SCROLLING, true);
    store.setDefault(SHOW_PSEUDO_CONFLICTS, false);
    store.setDefault(INITIALLY_SHOW_ANCHOR_PANE, false);
    store.setDefault(SHOW_MORE_INFO, false);
    store.setDefault(IGNORE_WHITESPACE, false);
    store.setDefault(PREF_SAVE_ALL_EDITORS, false);
    //store.setDefault(USE_SPLINES, false);
    store.setDefault(USE_SINGLE_LINE, true);
}
```

Symbol	File	Support	Confidence
initDefaults(IPreferenceStore store)	ComparePreferencePage.java	8	1.0
org.eclipse.compare.plugin.properties	plugin.properties	7	0.875
org.eclipse.compare.buildnotes.compare.html	buildnotes_compare.html	6	0.75
TextMergeViewer(Composite parent, int style, CompareConfiguration configuration)	TextMergeViewer.java	6	0.75
propertyChanged(PropertyChangeEvent event)	TextMergeViewer.java	6	0.75
createGeneralPage(Composite parent)	ComparePreferencePage.java	5	0.625
createTextComparePage(Composite parent)	ComparePreferencePage.java	5	0.625
handleDispose(DisposeEvent event)	TextMergeViewer.java	4	0.5

After the programmer has made some changes to the source (above), **ROSE** suggests locations (below) where, in similar transactions in the past, further changes were made

[Zimmermann, 2005]

Mining Version Histories to Guide Software Changes



# How can we mine SE Data – Part II

## Repositories of Repositories



January 2020:  
100 Million repositories  
40 Million Users



January 2020:  
430K repositories  
3.7 Million Users



April 2019  
28 Million repositories  
10 Million Users



April 2019  
28K projects



# How can we mine SE Data



# How can we mine SE Data



Search entire site...

Git is a **free and open source** distributed version control system designed to handle everything from small to very large projects with speed and efficiency.


Git is **easy to learn** and has a **tiny footprint with lightning fast performance**. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like **cheap local branching**, convenient **staging areas**, and **multiple workflows**.



Easiest to obtain local copy (distributed version control!), and has distinction between authors and committers, but ... branches can be pain to analyze



# How can we mine SE Data



Explore Gist Blog Help


bramadams + -

bramadams

News Feed Pull Requests Issues


### GitHub Bootcamp

1




**Set up Git**  
A quick guide to help you get started with Git.

2




**Create repositories**  
Repositories are where you'll work and collaborate on projects.

3




**Fork repositories**  
Forking creates a new, unique project from an existing one.

4



**Work together**  
Send pull requests, follow friends.  
Star and watch projects.

Access to thousands of  
Git-based projects via  
GitHub





**Better Word Highlighting in Diffs**

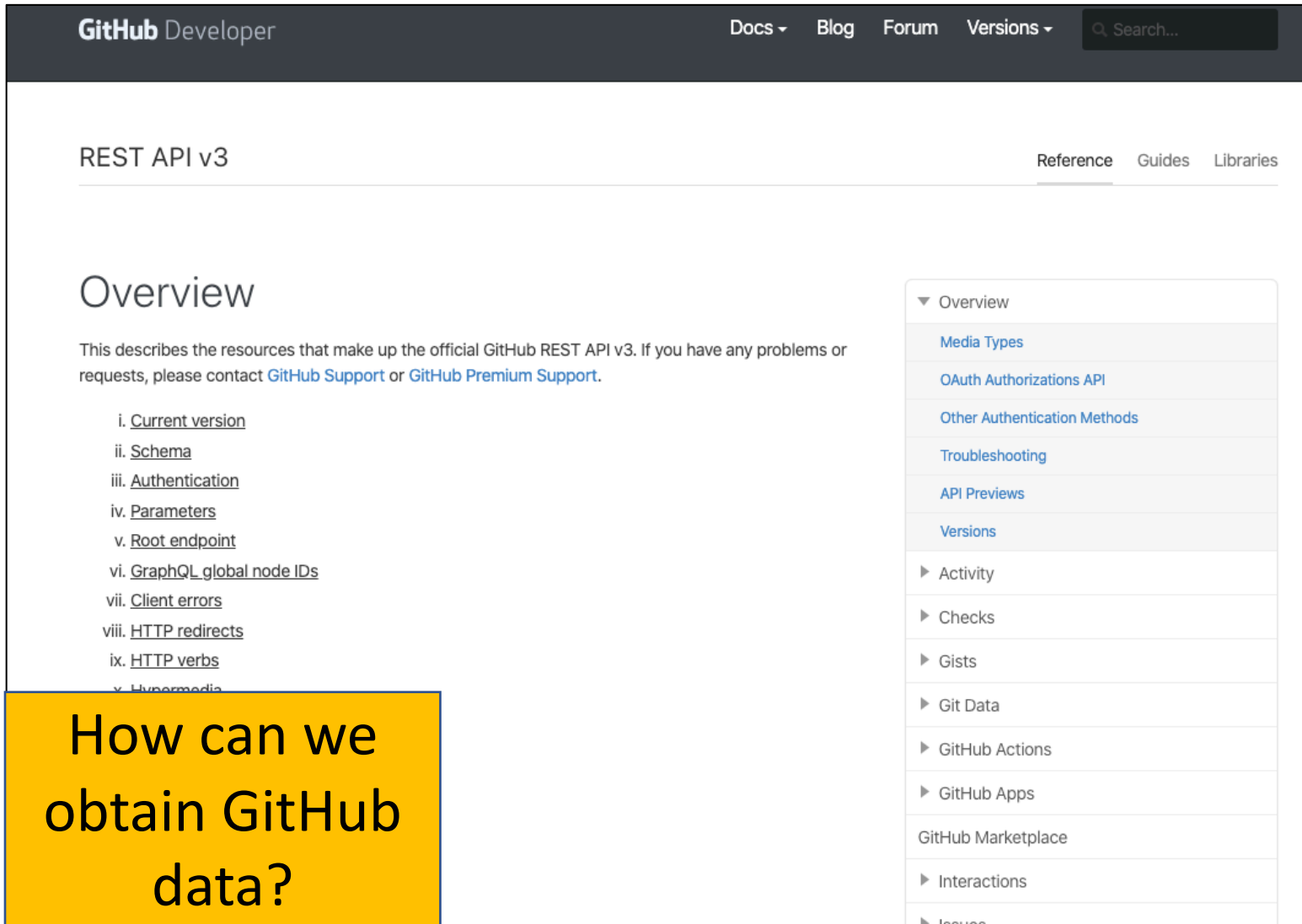
Commits, compare views, and pull requests now highlight individual changed words.

[View 70 new broadcasts](#)

Repositories you contribute to

 <a href="#">inukshuk/jekyll-scholar</a>	152 ★
 <a href="#">smcintosh/moosetracks</a>	1 ★

# How can we mine GitHub Data



The screenshot shows the GitHub Developer REST API v3 Overview page. The page has a dark header with the GitHub Developer logo and navigation links for Docs, Blog, Forum, and Versions. A search bar is located on the right. The main content area is titled 'REST API v3' and includes tabs for Reference, Guides, and Libraries. The 'Overview' section is active, displaying a list of links for various API features. A yellow callout box is overlaid on the bottom left of the page, containing the text 'How can we obtain GitHub data?'.

GitHub Developer

Docs ▾ Blog Forum Versions ▾

## REST API v3

Reference Guides Libraries

### Overview

This describes the resources that make up the official GitHub REST API v3. If you have any problems or requests, please contact [GitHub Support](#) or [GitHub Premium Support](#).

- i. [Current version](#)
- ii. [Schema](#)
- iii. [Authentication](#)
- iv. [Parameters](#)
- v. [Root endpoint](#)
- vi. [GraphQL global node IDs](#)
- vii. [Client errors](#)
- viii. [HTTP redirects](#)
- ix. [HTTP verbs](#)
- x. [Hypertext](#)

- ▼ Overview
  - [Media Types](#)
  - [OAuth Authorizations API](#)
  - [Other Authentication Methods](#)
  - [Troubleshooting](#)
  - [API Previews](#)
  - [Versions](#)
- ▶ Activity
- ▶ Checks
- ▶ Gists
- ▶ Git Data
- ▶ GitHub Actions
- ▶ GitHub Apps
- GitHub Marketplace
- ▶ Interactions
- ▶ Issues

# How can we mine GitHub Data

GitHub Developer

Docs ▾ Blog Forum Versions ▾

## REST API v3

Reference Guides Libraries

### Overview

This describes the resources that make up the official GitHub REST API v3. If you have any problems or requests, please contact [GitHub Support](#) or [GitHub Premium Support](#).

- i. [Current version](#)
- ii. [Schema](#)
- iii. [Authentication](#)
- iv. [Parameters](#)
- v. [Root endpoint](#)
- vi. [GraphQL global node IDs](#)
- vii. [Client errors](#)
- viii. [HTTP redirects](#)
- ix. [HTTP verbs](#)
- x. [Hypertext](#)

▼ Overview

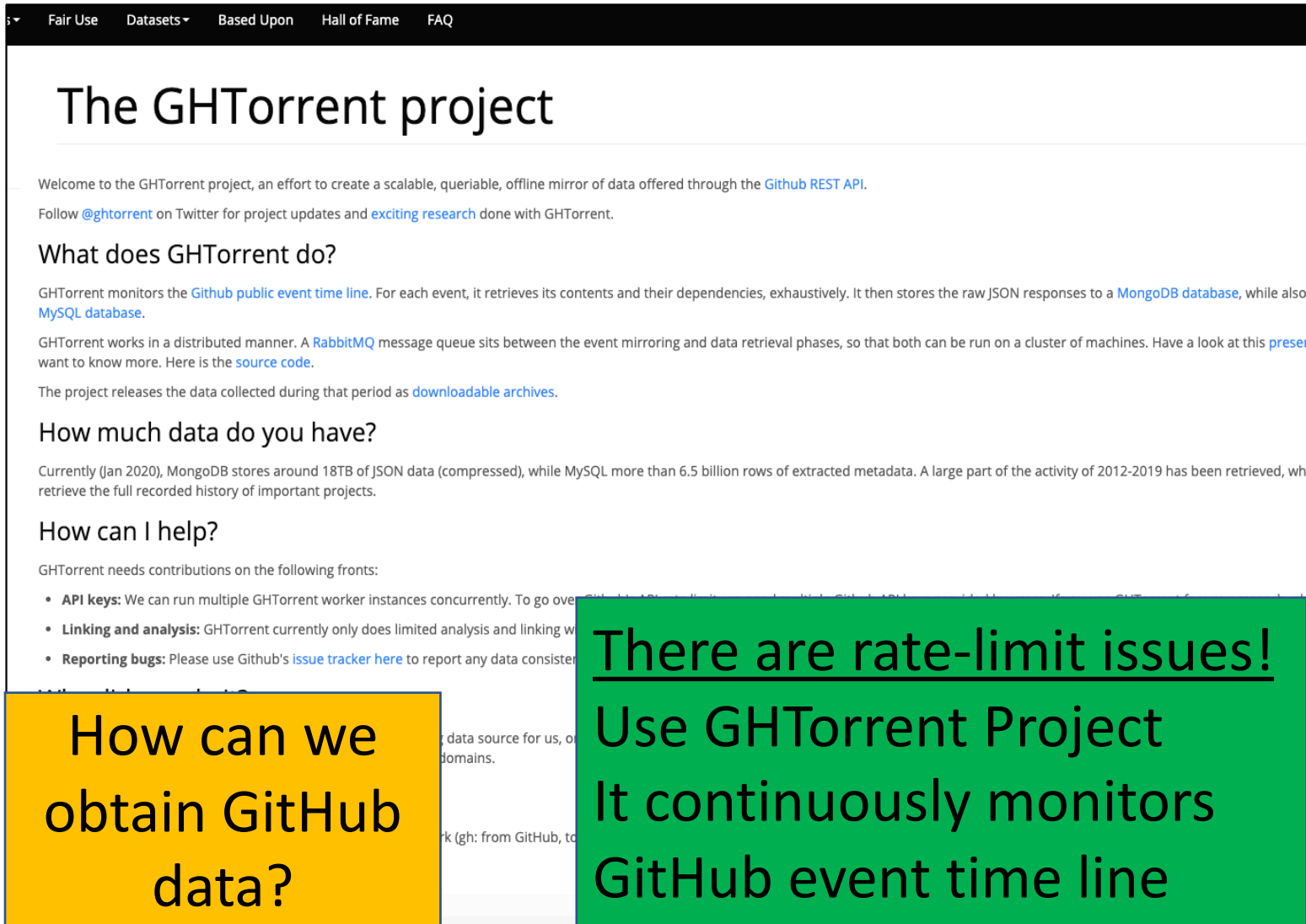
- [Media Types](#)
- [OAuth Authorizations API](#)
- [Other Authentication Methods](#)
- [Troubleshooting](#)
- [API Previews](#)
- [Versions](#)

► Activity

**How can we obtain GitHub data?**

**There are rate-limit issues!**  
**You are only allowed only a limited number of GitHub requests per hour**

# How can we mine GitHub Data



The screenshot shows the GHTorrent project website. At the top is a navigation bar with links: Fair Use, Datasets, Based Upon, Hall of Fame, and FAQ. The main heading is 'The GHTorrent project'. Below it, a welcome message states: 'Welcome to the GHTorrent project, an effort to create a scalable, queriable, offline mirror of data offered through the [Github REST API](#). Follow [@ghtorrent](#) on Twitter for project updates and [exciting research](#) done with GHTorrent.'

### What does GHTorrent do?

GHTorrent monitors the [Github public event time line](#). For each event, it retrieves its contents and their dependencies, exhaustively. It then stores the raw JSON responses to a [MongoDB database](#), while also [MySQL database](#).

GHTorrent works in a distributed manner. A [RabbitMQ](#) message queue sits between the event mirroring and data retrieval phases, so that both can be run on a cluster of machines. Have a look at this [presentation](#) if you want to know more. Here is the [source code](#).

The project releases the data collected during that period as [downloadable archives](#).

### How much data do you have?

Currently (Jan 2020), MongoDB stores around 18TB of JSON data (compressed), while MySQL more than 6.5 billion rows of extracted metadata. A large part of the activity of 2012-2019 has been retrieved, while the rest is still being retrieved. We will eventually retrieve the full recorded history of important projects.

### How can I help?

GHTorrent needs contributions on the following fronts:

- **API keys:** We can run multiple GHTorrent worker instances concurrently. To go over GitHub's API rate limits, we need more API keys. If you have one, please [contact us](#).
- **Linking and analysis:** GHTorrent currently only does limited analysis and linking with other data sources. If you have ideas or code to improve this, please [contact us](#).
- **Reporting bugs:** Please use Github's [issue tracker here](#) to report any data consistency issues or bugs.

data source for us, or domains.

work (gh: from GitHub, to

How can we obtain GitHub data?

There are rate-limit issues!  
Use GHTorrent Project  
It continuously monitors GitHub event time line

# How can we mine GitHub Data

The screenshot shows the GitHub Developer REST API v3 Overview page. The page has a dark header with 'GitHub Developer' on the left and navigation links 'Docs', 'Blog', 'Forum', and 'Versions' on the right, along with a search bar. Below the header, the page title 'REST API v3' is followed by tabs for 'Reference', 'Guides', and 'Libraries'. The main content area is titled 'Overview' and contains a paragraph: 'This describes the resources that make up the official GitHub REST API v3. If you have any problems or requests, please contact [GitHub Support](#) or [GitHub Premium Support](#).' Below this is a list of links: i. [Current version](#), ii. [Schema](#), iii. [Authentication](#), iv. [Parameters](#), v. [Root endpoint](#), vi. [GraphQL global node IDs](#), vii. [Client errors](#), viii. [HTTP redirects](#), ix. [HTTP verbs](#), and x. [Hypertext](#). On the right side, there is a sidebar with a table of contents showing 'Overview' expanded, with links to 'Media Types', 'OAuth Authorizations API', 'Other Authentication Methods', 'Troubleshooting', 'API Previews', and 'Versions', followed by 'Activity'.

How can we obtain GitHub data?

There are rate-limit issues!  
GitHub allows one to use authentication where u get a token (5000 requests/hr)