5 GitHub Tips on Working with (Sensitive) Data

by W. Jenny Shi, Peter DeWitt, Matthew Mulvahill

September 5, 2018

**Disclaimer**: This cheat sheet summarizes five cautionary points when using [GitHub](https://github.com/) to work with (sensitive) data. The list is not comprehensive.

## 1. Do NOT include [personally identifiable information (PII)](http://privacy.med.miami.edu/faq/privacy-faqs/what-is-personally-identifiable-information-pii)!

Unless the research analysis requires [PII](http://privacy.med.miami.edu/faq/privacy-faqs/what-is-personally-identifiable-information-pii), you should only be working with de-identified data. If you are given any PII, please inform the investigators and permanently delete the data right away.

## 2. GitHub is NOT HIPAA-compliant!

If you are working with materials protected by the Health Insurance Portability and Accountability Act (HIPAA), DO NOT EVER place it in a GitHub repository. You should also avoid storing on iCloud and emailing the materials. Most of cloud storage services, such as Amazon S3 and Dropbox, are NOT HIPAA-compliant out-of-the-box.

## 3. GitHub allows public and private repositories.

Anyone can get a free GitHub account and publish public repositories. Public repositories are viewable and accessible by everyone visiting the web. GitHub users can pay a small fee, and obtain private repositories that are not accessible by general public. Even if you have a paid account, you should NOT upload sensitive information. You can change a repository’s visibility, but switching between public and private does NOT change the commit history.

## 4. Data files can be excluded from commits with .gitignore.

As the name states, [.gitignore](https://git-scm.com/docs/gitignore) tells git which files it should ignore. If there is a data set to be included in a subdirectory of the repository, upon adding the file to the repository add the file name to the .gitignore file. It is recommended, at the very least, that an md5sum or SHA of the file is versioned and checked during builds. Another option is to keep data in a separate git repository stored locally, with no ‘remote’.

## 5. Deleted files may still be recoverable by others.

You can [delete any file](https://help.github.com/articles/deleting-files/) within your repositories on GitHub. However, since Git is a version control system, unless you take additional actions, the file is NOT completely removed, and it can be recovered from the commit history. To COMPLETELY remove a file from a repository, follow the steps [here](https://help.github.com/articles/removing-sensitive-data-from-a-repository/). If you have pushed sensitive files to GitHub, additional mitigation will be necessary. This may include reviewing responsibilities under HIPAA and institutional rules, requesting GitHub delete internal cache, and assessing the extent of unauthorized access by requesting GitHub compare IP address access records with a list of potentially valid institutional IP addresses.