



NYC DATA SCIENCE
ACADEMY

Introduction to Git & Github

Data Science Bootcamp

Outline

- ❖ **Set up Git and GitHub**

- ❖ **Introduction to Git**

- **Creating a Git Repository**
- **Manipulating files**

- ❖ **Introduction to GitHub**

- **Lightning Tour of Github**
- **Create a Remote Repository**

Pre-requisites

- ❖ Patience and willingness to keep learning
- ❖ Have a good text editor for editing plain text files. Good examples are:
 - Notepad++
 - TextWrangler
 - Sublime
 - Emacs
 - Vim
 - LightTable
- ❖ Syntax highlighting and side-by-side editing are also useful features.

Pre-requisites for Windows

- ❖ Notepad++: <https://notepad-plus-plus.org/download/>
- ❖ The default is plain UTF-8 encoding, without a byte-order mark (BOM).

The screenshot shows the Notepad++ interface with two tabs open: "index.html" and "variant-duo.css".

index.html:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd>
<html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
<head>
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
<meta name="description" content="Your description goes here" />
<meta name="keywords" content="your,keywords,goes,here" />
<meta name="author" content="Your Name" />
<link rel="stylesheet" type="text/css" href="variant-duo.css" title="Variant Duo" media="screen,projection" />
<title>Variant Duo v1.0</title>
</head>
<body>
<div id="wrap">
<h1><a href="index.html">Variant Duo</a></h1>
<p class="slogan">Dual-column content template</p>
<div id="menu">
<strong>menulinks</strong>
<a class="menulink" href="index.html">Home</a><span class="hide">|</span>
<a class="menulink active" href="index.html">Page 2</a><span class="hide">|</span>
<a class="menulink" href="index.html">Page 3</a><span class="hide">|</span>
<a class="menulink" href="index.html">Page 4</a><span class="hide">|</span>
<a class="menulink" href="index.html">Page 5</a>
</p>
</div>

<div id="content">
<div class="left">
<h2>Introducing Variant Duo</h2>
<p>If you are looking for a really simple website template with a basic dual-column layout that is easy to get started with, then Variant Duo may be a good starting point. This template is completely free and may be used without any limitations or obligations. I kindly ask you to leave the design credit link in

```

variant-duo.css:

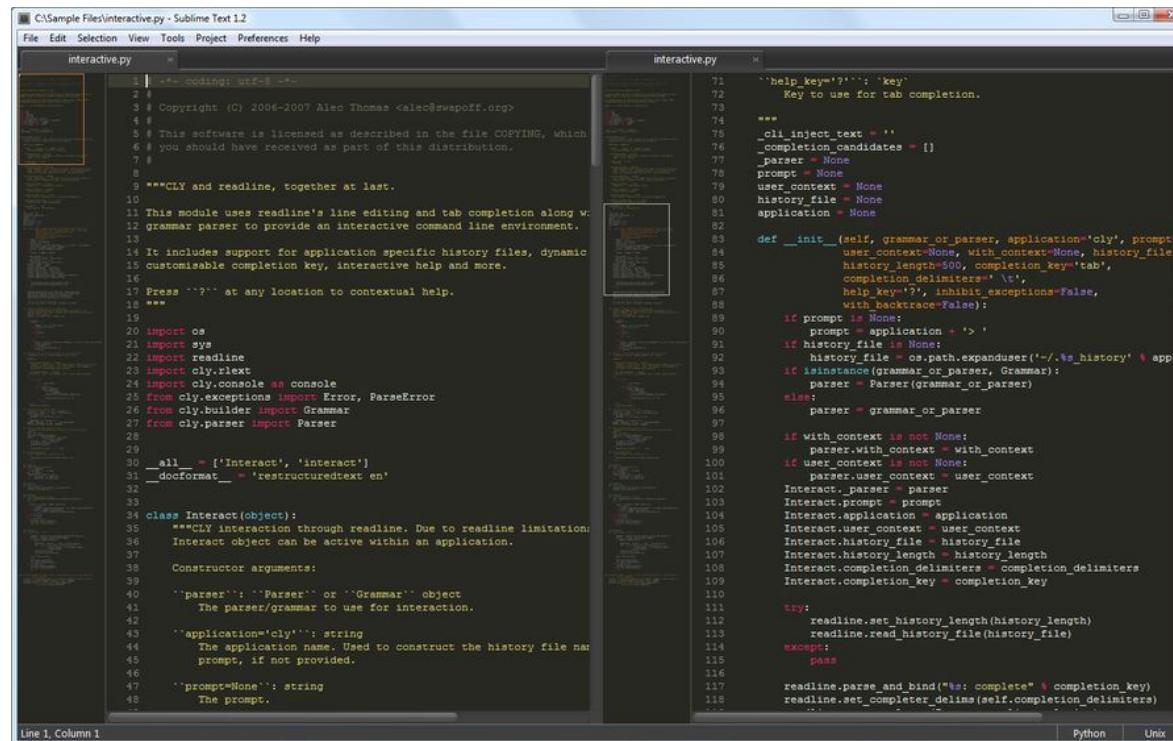
```
Original design Variant Duo (v1.0 - Oct 08, 2010) - A free xhtml/css website template by Andreas Viklund.
For more information, see http://andreasviklund.com/templates/variant-duo/
/*
Main containers */
body {padding:0; margin:0; font:83% tahoma,verdana,sans-serif; background-color:#e6e4e1; color:#333;
#wrap {width:980px; text-align:left; margin:0 auto;}
#menu {text-align:center; margin-top:40px;}
#content {background:#fff; text-align:left; padding:20px 20px 5px; margin:15px 0 15px 0;}
```

The status bar at the bottom of the Notepad++ window displays the following information:

- Cascade Style Sheets File
- 2078 chars 2117 bytes 40 lines
- Ln:1 Col:1 Sel:0 (0 bytes) in 0 ranges
- UNIX ANSI as UTF-8 INS

Pre-requisites for Windows

- ❖ Sublime Text: <http://www.sublimetext.com/3>
- ❖ Heavy-weight IDE: RStudio, Eclipse, Visual Studio, PyCharm



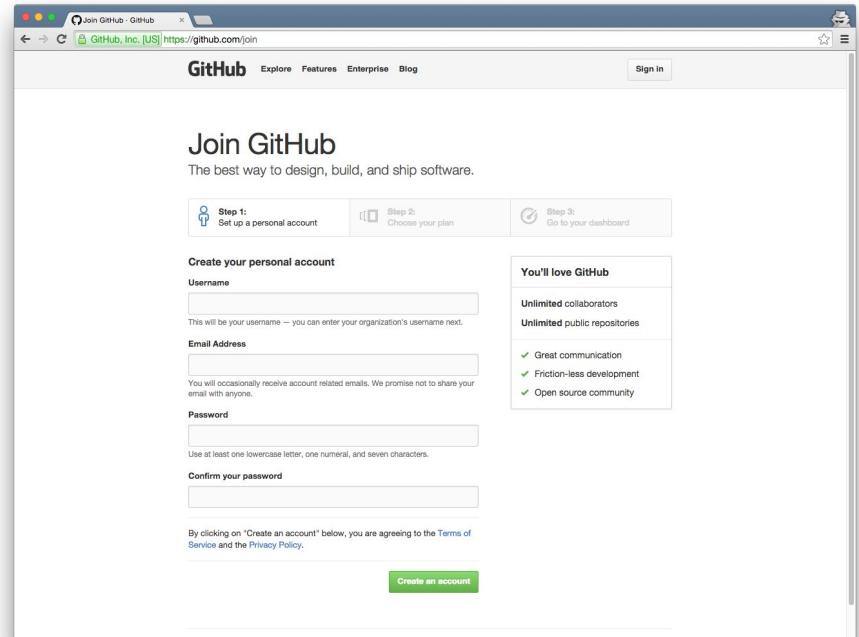
```
 1 #!/usr/bin/python
 2 #
 3 # Copyright (C) 2006-2007 Alec Thomas <alec@swipoff.org>
 4 #
 5 # This software is licensed as described in the file COPYING, which
 6 # you should have received as part of this distribution.
 7 #
 8 #
 9 """CLY and readline, together at last.
10
11 This module uses readline's line editing and tab completion along with
12 a grammar parser to provide an interactive command line environment.
13
14 It includes support for application specific history files, dynamic
15 customizable completion key, interactive help and more.
16
17 Press `??` at any location to contextual help.
18 """
19
20 import os
21 import sys
22 import readline
23 import clyrixet
24 import cly.console as console
25 from cly.exceptions import Error, ParseError
26 from cly.builder import Grammar
27 from cly.parser import Parser
28
29
30 __all__ = ['Interact', 'interact']
31 __docformat__ = 'restructuredtext en'
32
33
34 class Interact(object):
35     """CLY interaction through readline. Due to readline limitation,
36     Interact object can be active within an application.
37
38     Constructor arguments:
39
40     'parser': 'Parser' or 'Grammar' object
41         The parser/grammar to use for interaction.
42
43     'application='cly)': string
44         The application name. Used to construct the history file name
45         prompt, if not provided.
46
47     'prompt=None': string
48         The prompt.
49
50     'help_key='?': string
51         Key to use for tab completion.
52
53     'cli_inject_text': string
54     'completion_candidates': []
55     '_parser': None
56     'prompt': None
57     'user_context': None
58     'history_file': None
59     'application': None
60
61     def __init__(self, grammar_or_parser, application='cly', prompt=None,
62                  user_context=None, history_file=None,
63                  history_length=500, completion_key='tab',
64                  completion_delimiters=' \t',
65                  help_key='?', inhibit_exceptions=False,
66                  with_backtrace=False):
67         if prompt is None:
68             prompt = application + '> '
69         if history_file is None:
70             history_file = os.path.expanduser('~/.%s_history' % application)
71         if isinstance(grammar_or_parser, Grammar):
72             parser = Parser(grammar_or_parser)
73         else:
74             parser = grammar_or_parser
75
76         if with_context is not None:
77             parser.with_context = with_context
78         if user_context is not None:
79             parser.user_context = user_context
80         Interact.parser = parser
81         Interact.prompt = prompt
82         Interact.application = application
83         Interact.user_context = user_context
84         Interact.history_file = history_file
85         Interact.history_length = history_length
86         Interact.completion_delimiters = completion_delimiters
87         Interact.completion_key = completion_key
88
89     try:
90         readline.set_history_length(history_length)
91         readline.read_history_file(history_file)
92     except:
93         pass
94
95     readline.parse_and_bind("%s: complete \"%s\" completion.key")
96     readline.set_completer_delims(self.completion_delimiters)
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
```

Pre-requisites for Mac

- ❖ Sublime Text: <http://www.sublimetext.com/3>
- ❖ TextMate: <http://macromates.com/download>
- ❖ LightTable: <http://lighttable.com/>
- ❖ MacVim: <https://github.com/macvim-dev/macvim>
- ❖ Many other options ...

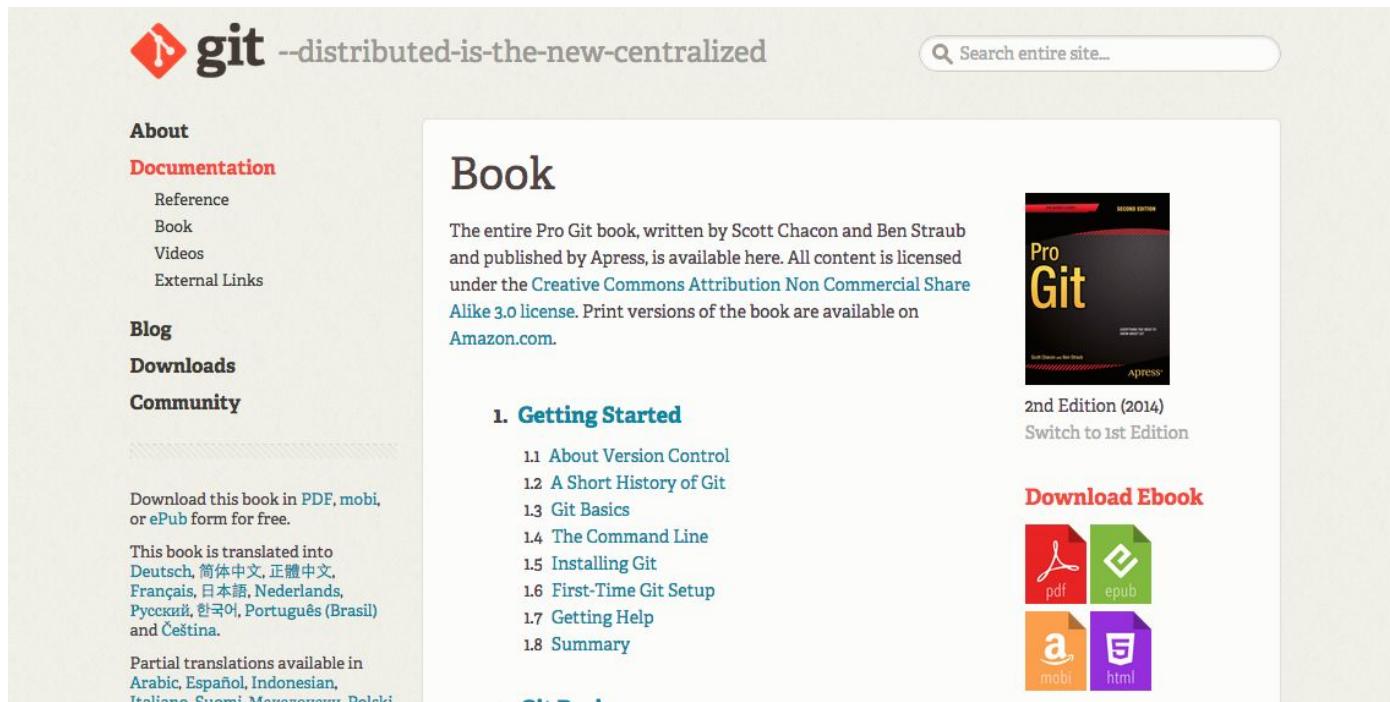
Pre-requisites

- ❖ Create a personal GitHub account.
- ❖ Choose the free option: get unlimited open repositories, but no private repositories.



Pre-requisites

- ❖ Bookmark the free Pro Git book <http://git-scm.com/book/en/v2/>



The screenshot shows the homepage of the Pro Git book website. At the top, there's a navigation bar with links for "About", "Documentation", "Blog", "Downloads", and "Community". Below the navigation, there's a section for downloading the book in PDF, mobi, or ePub format. A note indicates that the book is translated into multiple languages. The main content area is titled "Book" and contains a brief description of the book, its authors, and its availability under a Creative Commons license. It also mentions that print versions are available on Amazon.com. To the right of the description is an image of the book cover, which is black with the title "Pro Git" in large white letters. Below the book image, it says "2nd Edition (2014)" and "Switch to 1st Edition". Further down, there's a section titled "Download Ebook" with icons for PDF, EPUB, MOBI, and HTML formats.

git --distributed-is-the-new-centralized

Search entire site...

About

Documentation

Reference
Book
Videos
External Links

Blog

Downloads

Community

Download this book in PDF, mobi, or ePub form for free.

This book is translated into Deutsch, 简体中文, 正體中文, Français, 日本語, Nederlands, Русский, 한국어, Português (Brasil) and Čeština.

Partial translations available in Arabic, Español, Indonesian, Italiano, Svenska, Magyar, and Dansk

Book

The entire Pro Git book, written by Scott Chacon and Ben Straub and published by Apress, is available here. All content is licensed under the [Creative Commons Attribution Non Commercial Share Alike 3.0 license](#). Print versions of the book are available on [Amazon.com](#).

1. Getting Started

- 1.1 About Version Control
- 1.2 A Short History of Git
- 1.3 Git Basics
- 1.4 The Command Line
- 1.5 Installing Git
- 1.6 First-Time Git Setup
- 1.7 Getting Help
- 1.8 Summary

2nd Edition (2014)
Switch to 1st Edition

Download Ebook

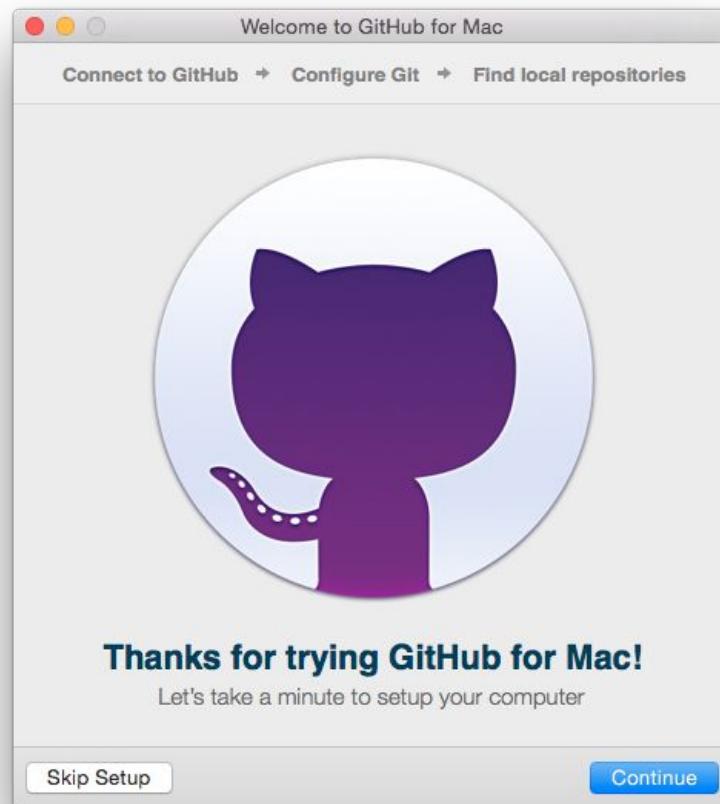
pdf epub
mobi html

Download and Install Github GUI For Your OS (Optional)

- ❖ Easy method: Download and install GitHub's application. May not work for all versions of your OS. If this doesn't work, it's ok, let an instructor know.
- ❖ Windows: <https://windows.github.com/>
- ❖ Mac: <https://mac.github.com/>

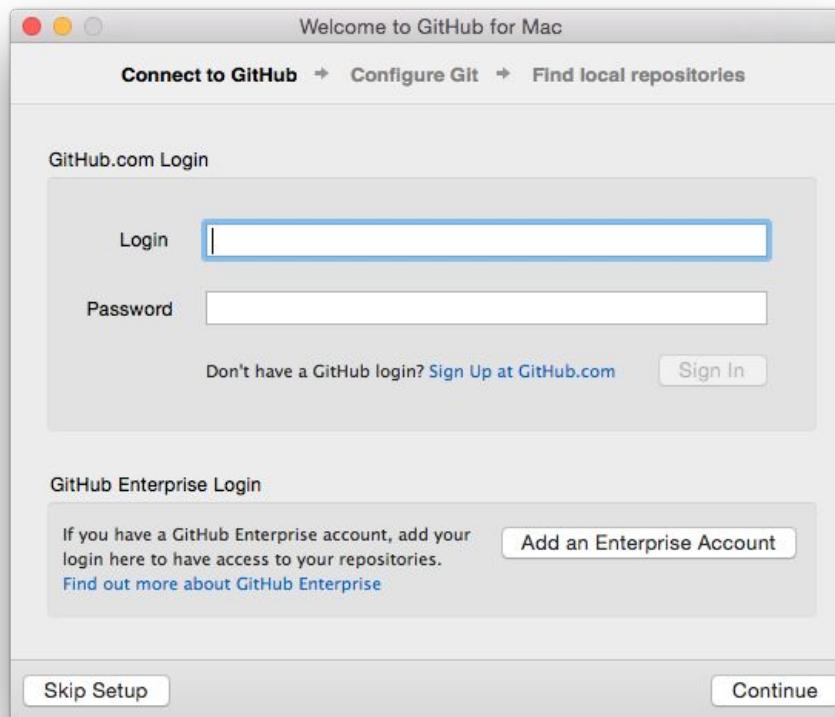
Download and Install Github GUI For Your OS (Optional)

- ❖ After you install the GUI, find the application in the start menu or Applications, and open it.



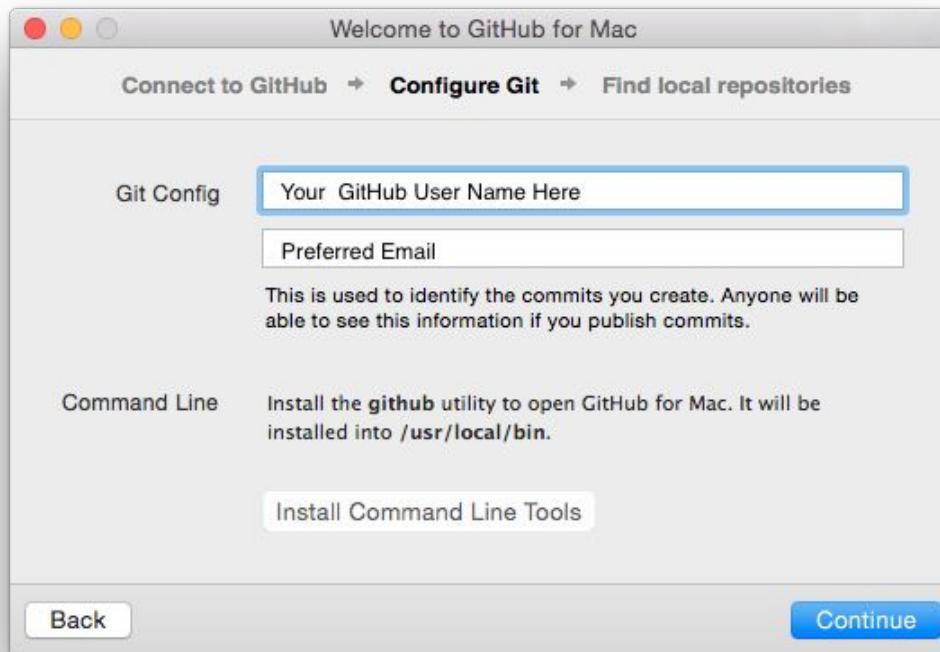
Download and Install Github GUI For Your OS (Optional)

- ❖ Enter your GitHub credentials through the setup wizard. This step should automatically configure your SSH keys for identification.



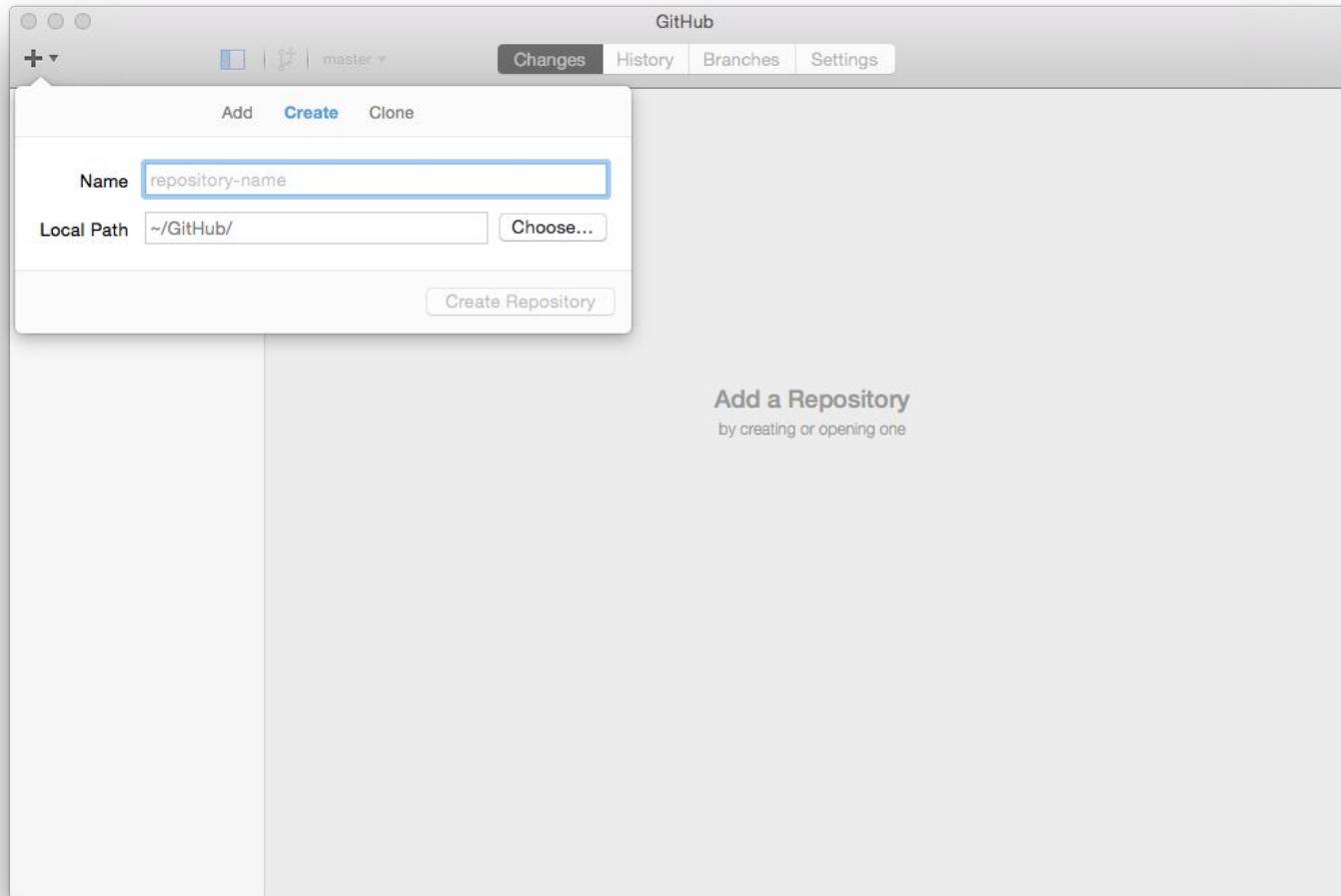
Download and Install Github GUI For Your OS (Optional)

- ❖ Enter the github username you created, and the preferred email you used as well. Your commits will be identified by these details, so make sure it's okay to share.



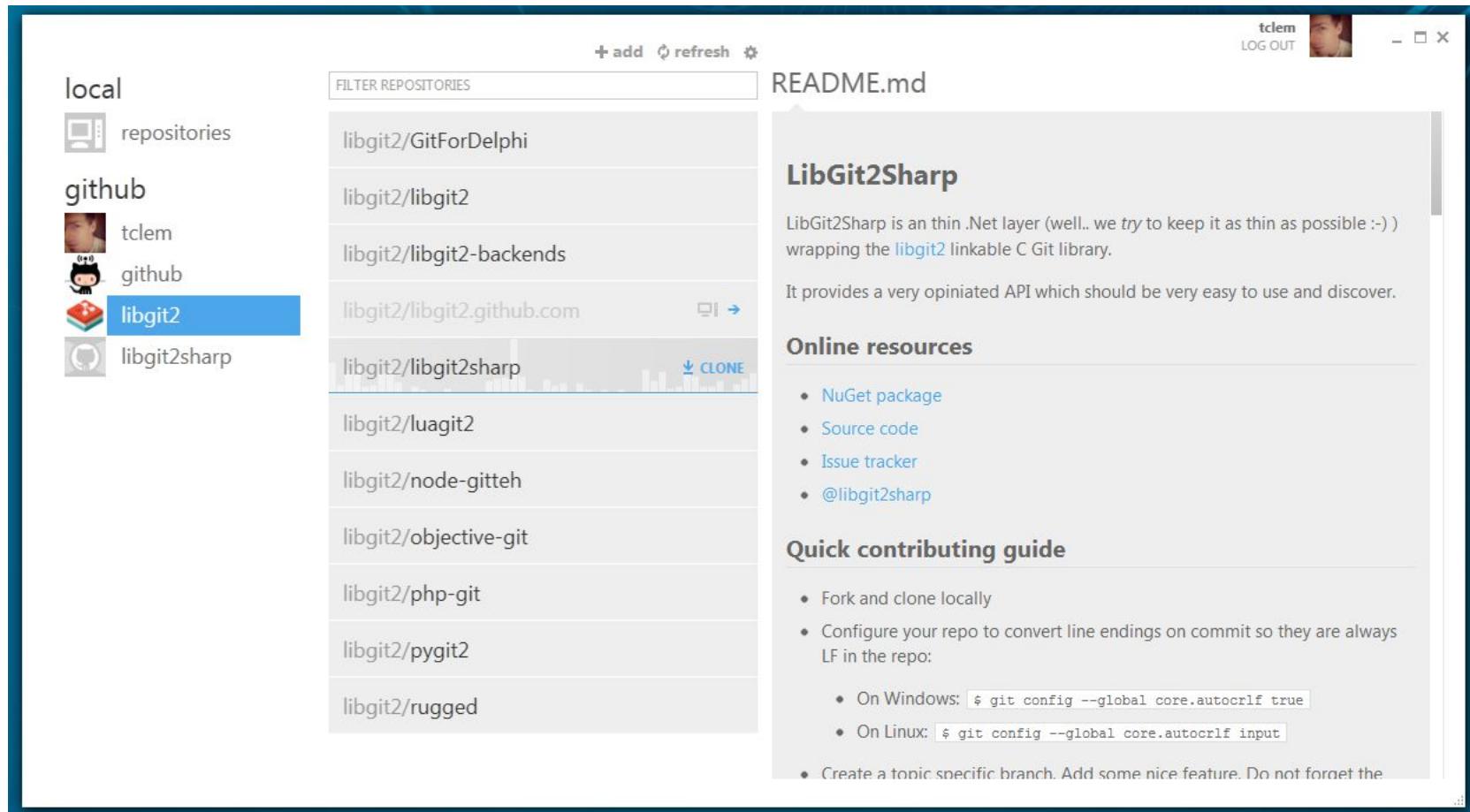
Download and Install Github GUI For Your OS (Optional)

- ❖ Example Application for Mac



Download and Install Github GUI For Your OS (Optional)

❖ Example Application for Windows



Command-Line For Windows

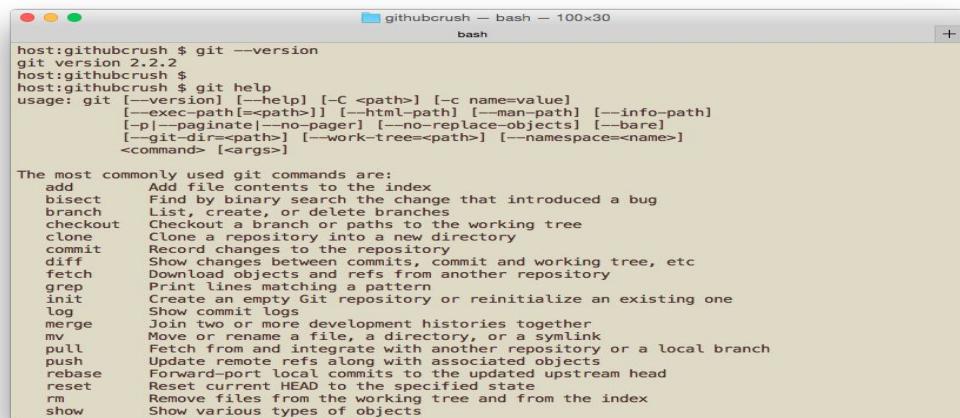
- ❖ Download the package msysgit from <https://git-for-windows.github.io/> and install it. This wizard will take you through the steps to install it.
- ❖ Assuming the default settings, open your 'My Documents' folder to find the root folder of the command-line environment that msysgit created.
- ❖ On your computer, find Git in the start menu, and open GitBash. If you find something like the following window, the installation is successful!



The screenshot shows a terminal window titled "MINGW32:/c/Users/alisa". The title bar also includes the path "MINGW32:/c/Users/alisa". The window contains the following text:
Welcome to Git (version 1.9.4-preview20140815)
Run 'git help git' to display the help index.
Run 'git help <command>' to display help for specific commands.
alisa@WWW ~ (master)
\$

Command-Line for Mac

- ❖ Three alternative install methods:
 - Use the pre-packaged version of git provided by OS X
 - Directly install it from a package <http://git-scm.com/download/mac>
 - Install the Homebrew package manager <http://brew.sh/> and then open your terminal type `$ brew install git`
- ❖ Within Applications, find the Utilities folder, and open the Terminal application. If you can successfully type `git help` at the terminal, then installation was successful!



```
githubcrush $ git --version
git version 2.2.2
githubcrush $
githubcrush $ git help
usage: git [--version] [--help] [-C <path>] [-c name=value]
           [-exec-path[=<path>]] [-html-path] [--man-path] [--info-path]
           [-p|--paginate[--no-pager]] [--no-replace-objects] [--bare]
           [--git-dir=<path>] [--work-tree=<path>] [--namespace=<name>]
           <command> [<args>]

The most commonly used git commands are:
add      Add file contents to the index
bisect   Find by binary search the change that introduced a bug
branch   List, create, or delete branches
checkout  Checkout a branch or paths to the working tree
clone    Clone a repository into a new directory
commit   Record changes to the repository
diff     Show changes between commits, commit and working tree, etc
fetch   Download objects and refs from another repository
grep    Print lines matching a pattern
init    Create an empty Git repository or reinitialize an existing one
log     Show commit logs
merge   Join two or more development histories together
mv      Move or rename a file, a directory, or a symlink
pull   Fetch files and update local branches, a repository or a local branch
push    Update remote refs along with associated objects
rebase  Forward-port local commits to the updated upstream head
reset   Reset current HEAD to the specified state
rm      Remove files from the working tree and from the index
show   Show various types of objects
```

View All Configuration Information

- ❖ Start the terminal program for your OS. Type in this command to list all the known global Git configuration. If you don't see a **user.name** or **user.email** setting in your global config, then it has to be manually configured.

```
$ git config --list --global  
...  
user.name='your username'  
user.email=your email  
...
```

Configuration for Git

- ❖ View the version information and verify that the installation was successful. E.g. on Windows:

```
$ git --version  
git version 1.9.4.msysgit.1
```

- ❖ Set your user information in the command window/terminal

```
$ git config --global user.name "username"  
$ git config --global user.email "your email"
```

- ❖ Note : The parameter --global in the git config command indicates that all repositories on this host will use this configuration. For setting a different configuration on an individual repository, use the --local parameter.

View Specific Configuration Information

- ❖ If you want to view a particular setting, like user.email, specify that parameter after the git config command.

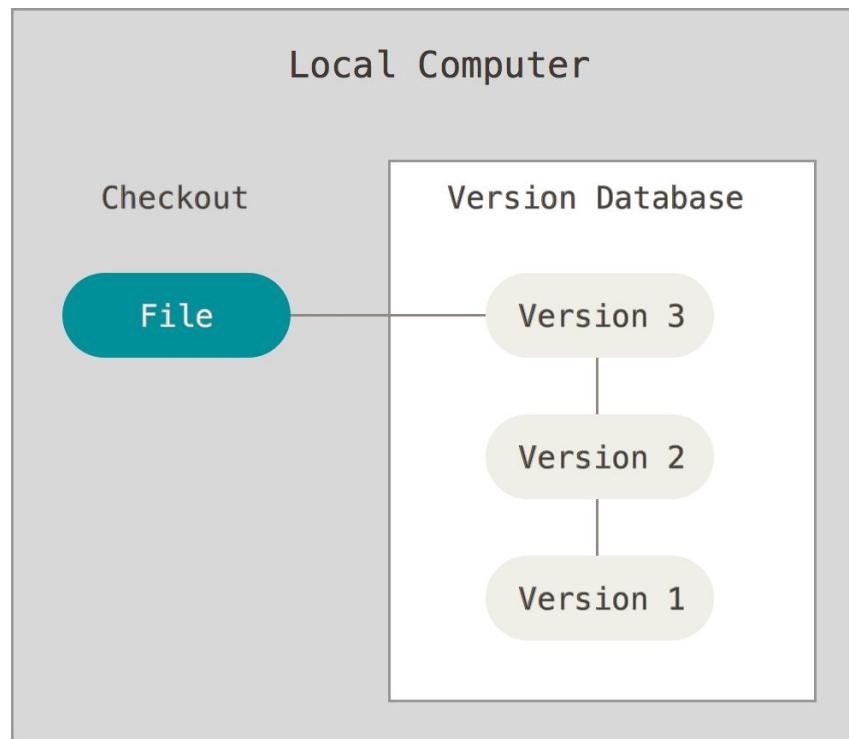
```
$ git config user.email  
my.email@domain.com
```

Outline

- ❖ Set up Git and GitHub
- ❖ Introduction to Git
 - Creating a Git Repository
 - Manipulating files
- ❖ Introduction to GitHub
 - Lightning Tour of Github
 - Create a Remote Repository

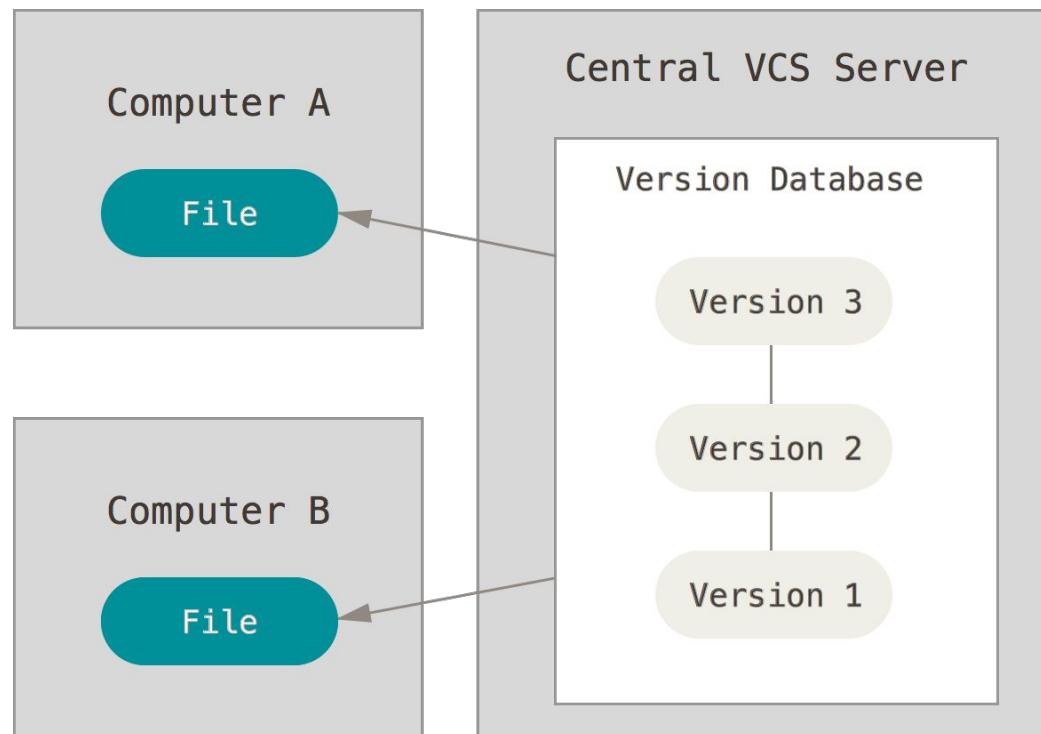
What is Version Control?

- ❖ Version control allows us to capture and refer back to a known state of a file across a series of changes.
- ❖ E.g. Mac OS X File System, MS Office App 'Track Changes' feature



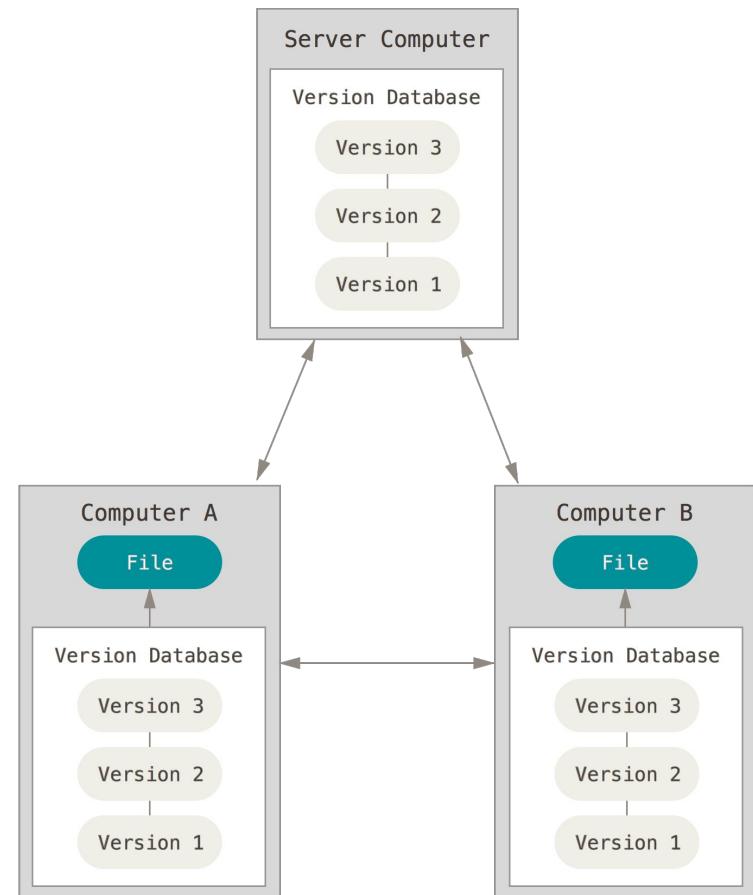
What is Centralized Version Control?

- ❖ Sharing changes between hosts is done through a central authority (the source of truth).
- ❖ E.g. standard Google spreadsheet



What is Distributed Version Control?

- ❖ Sharing changes between hosts is done using a peer-to-peer model (no real source of truth).
- ❖ E.g. Git, offline Google spreadsheet



Introduction to Git

- ❖ Distributed version control system, great for collaboration over source code
- ❖ Linus Torvalds developed Git in 2005 to help manage the Linux kernel development across thousands of volunteers
- ❖ Fast and robust performance
- ❖ Wide community adoption
- ❖ Allows massive, parallel branch development
- ❖ Has the ability to efficiently manage large scale projects

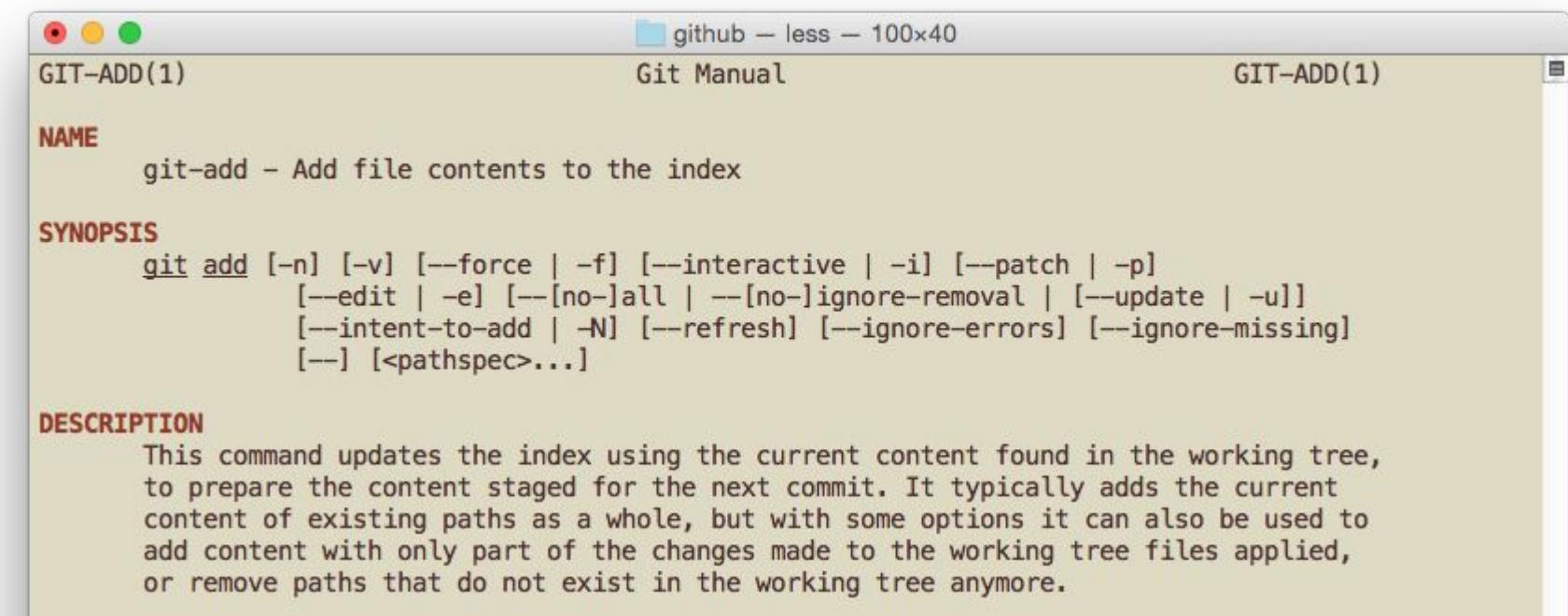
Introduction to Git

- ❖ If you're coming from another Version Control System (VCS), note that Git has some differences:
 - Doesn't really work like CVS, Perforce, i.e. centralized version control.
 - Instead of encoding deltas between one version and the next, Git works with files as a whole.
 - Treats the history as a series of snapshots, which are then efficiently stored in the local database.
 - Many commands in Git work on primarily the local data. No need to talk to a remote server, unless you want to exchange history, branches, and/or tags.

Help with Git Commands

- ❖ For more information about Git's many commands, consult the help command.

```
$ git add --help  
$ git help add
```



The screenshot shows a terminal window titled "github — less — 100x40". The window contains the output of the "git help add" command. The output is organized into sections: NAME, SYNOPSIS, and DESCRIPTION. The NAME section contains "git-add - Add file contents to the index". The SYNOPSIS section contains the command line options for "git add". The DESCRIPTION section contains a detailed explanation of the command's purpose and how it updates the index.

```
GIT-ADD(1)                               Git Manual                               GIT-ADD(1)

NAME
    git-add - Add file contents to the index

SYNOPSIS
    git add [-n] [-v] [--force | -f] [--interactive | -i] [--patch | -p]
                [--edit | -e] [--[no-]all | --[no-]ignore-removal | [--update | -u]]
                [--intent-to-add | -N] [--refresh] [--ignore-errors] [--ignore-missing]
                [--] <pathspec>...

DESCRIPTION
    This command updates the index using the current content found in the working tree,
    to prepare the content staged for the next commit. It typically adds the current
    content of existing paths as a whole, but with some options it can also be used to
    add content with only part of the changes made to the working tree files applied,
    or remove paths that do not exist in the working tree anymore.
```

Help on the Web

- ❖ Many online resources, but here is one: <http://git-scm.com/docs/git-add>

The screenshot shows a web browser window with the URL <http://git-scm.com/docs/git-add> in the address bar. The page is titled "Git - git-add Documentation". The main content area displays the "git-add" command documentation. It includes sections for NAME, SYNOPSIS, DESCRIPTION, OPTIONS, and a detailed description of the command's behavior. The SYNOPSIS section shows the command line syntax:

```
'git add' [-n] [-v] [--force | -f] [--interactive | -i] [--patch | -p]
[--edit | -e] [--no-all | --[no-]ignore-removal | --update | -u
--intent-to-add | -N] [--refresh] [--ignore-errors] [--ignore-mis
[--] [<pathspec>...]
```

The DESCRIPTION section explains that the command updates the index using the current content found in the working tree to prepare the content staged for the next commit. It typically adds the current content of existing paths as a whole, but with some options it can also be used to add content with only part of the changes made to the working tree files applied, or remove paths that do not exist in the working tree anymore. The OPTIONS section details various flags like -n (dry-run), -v (verbose), and --patch (add individual files).

Outline

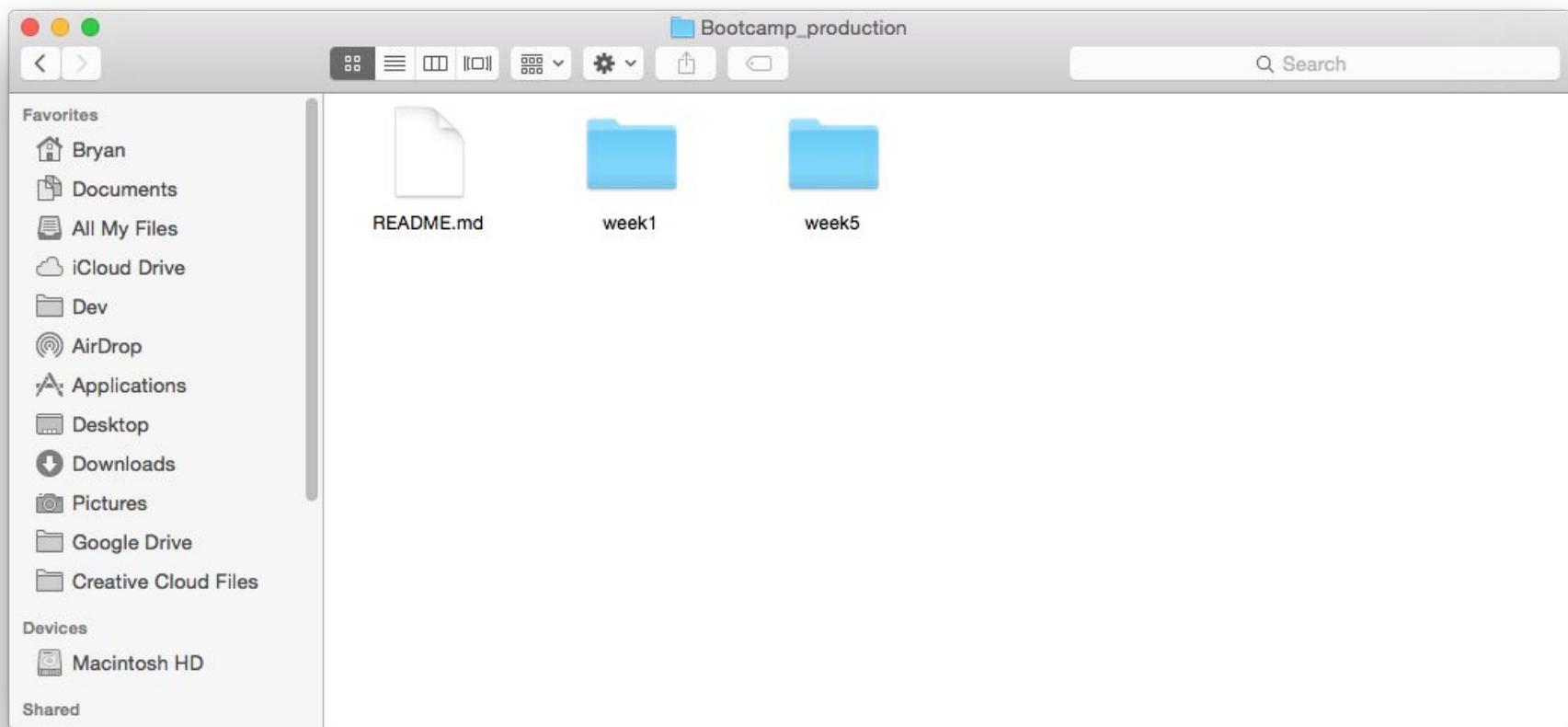
- ❖ Set up Git and GitHub
- ❖ Introduction to Git
 - Creating a Git Repository
 - Manipulating files
- ❖ Introduction to GitHub
 - Lightning Tour of Github
 - Create a Remote Repository

What is a Git Repository?

- ❖ View the repository as a controlled working area (a directory) with a version database.
 - Some or all of the files in this directory can be managed by Git:
 - add and commit files to the repository
 - move or delete files
 - view updated files
 - view history
 - cancel local changes
 - Git can even follow renamings (or moving) of files!

What is a Repository?

- ❖ A repository is a directory that contains other files and subdirectories on some volume.



Create a Repository

- ❖ There are two ways to create a repository:
 - Create a local repository directly:
 - Create a repository locally using the `git init` command.
 - Later, add a remote repository reference to the configuration, and push data to that remote.
 - Clone a remote repository to your local machine.

Create a Local Repository

- ❖ **Step 1:** Choose a working directory and initialize an empty directory
 - Type the following commands one at a time. Everything after the double forward slash (//) are just comments.

```
$ pwd                  // show current directory  
/c/Users/dev/githubcrush  
$ mkdir datascience    // create empty directory  
$ cd datascience      // enter the directory  
$ git init             // initial new repository, an empty repository  
Initialized empty Git repository in  
c:/Users/dev/githubcrush/datascience/.git/
```

- ❖ Note: The git init command is used to change a local directory into a repository managed by Git.

What is a Repository?

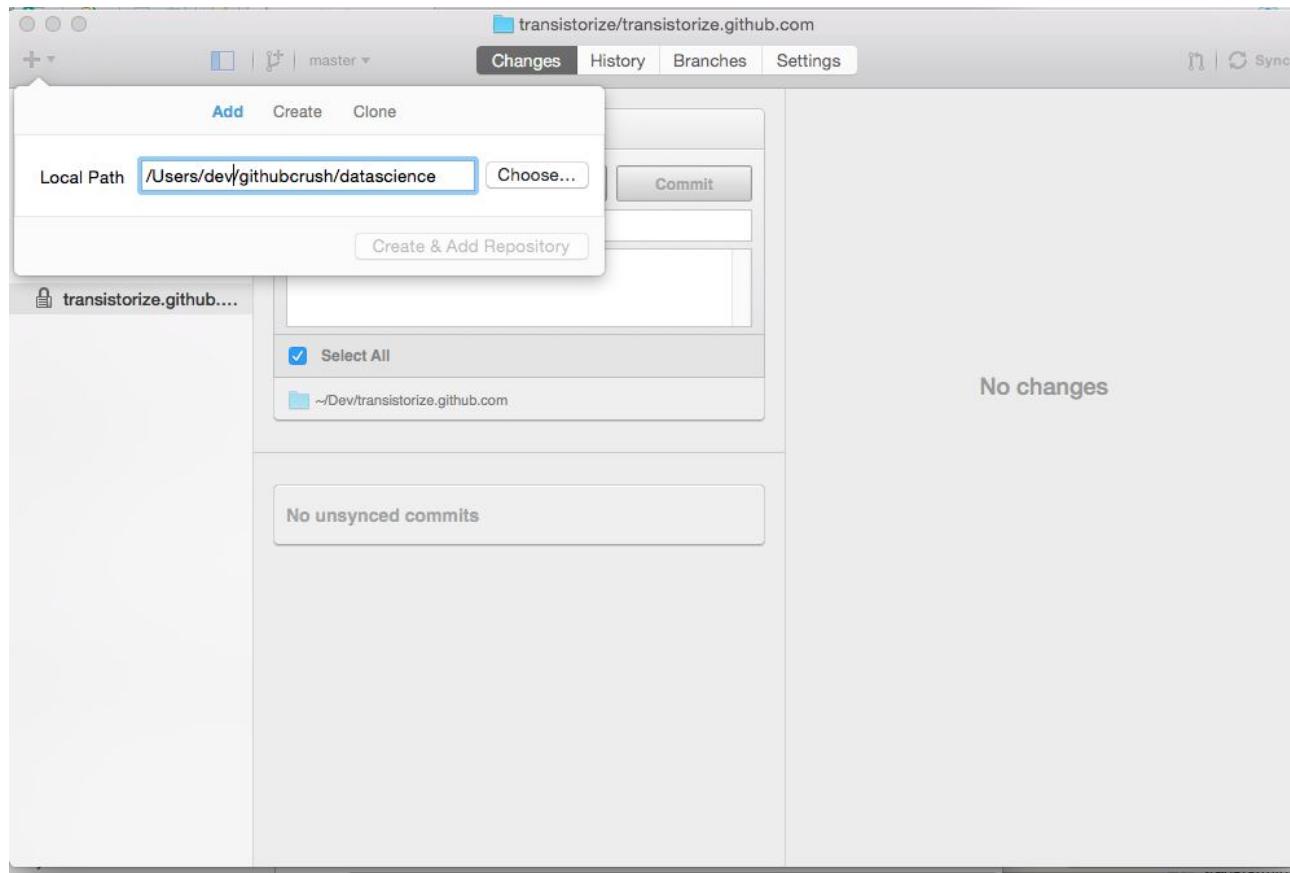
- ❖ After initializing the directory as a git repository, a hidden .git database is created inside that directory.



```
Bryans-iMac:Bootcamp_production Bryan$ ls -lah
total 8
drwxr-xr-x  6 Bryan  staff  204B Jan 20 12:28 .
drwxr-xr-x 20 Bryan  staff  680B Jan 20 12:24 ..
drwxr-xr-x 13 Bryan  staff  442B Jan 20 14:46 .git
-rw-r--r--  1 Bryan  staff  2.7K Jan 20 12:28 README.md
drwxr-xr-x  3 Bryan  staff  102B Jan 20 12:28 week1
drwxr-xr-x  7 Bryan  staff  238B Jan 20 12:28 week5
Bryans-iMac:Bootcamp_production Bryan$
```

Create a Local Repository

- ❖ Optionally, add the repository to the GitHub UI.



Add Content to the Local Repository

- ❖ **Step 2:** Create a new text file using your text editor, create a file named data.txt within the datascience directory you created before. Add the following content to this file:

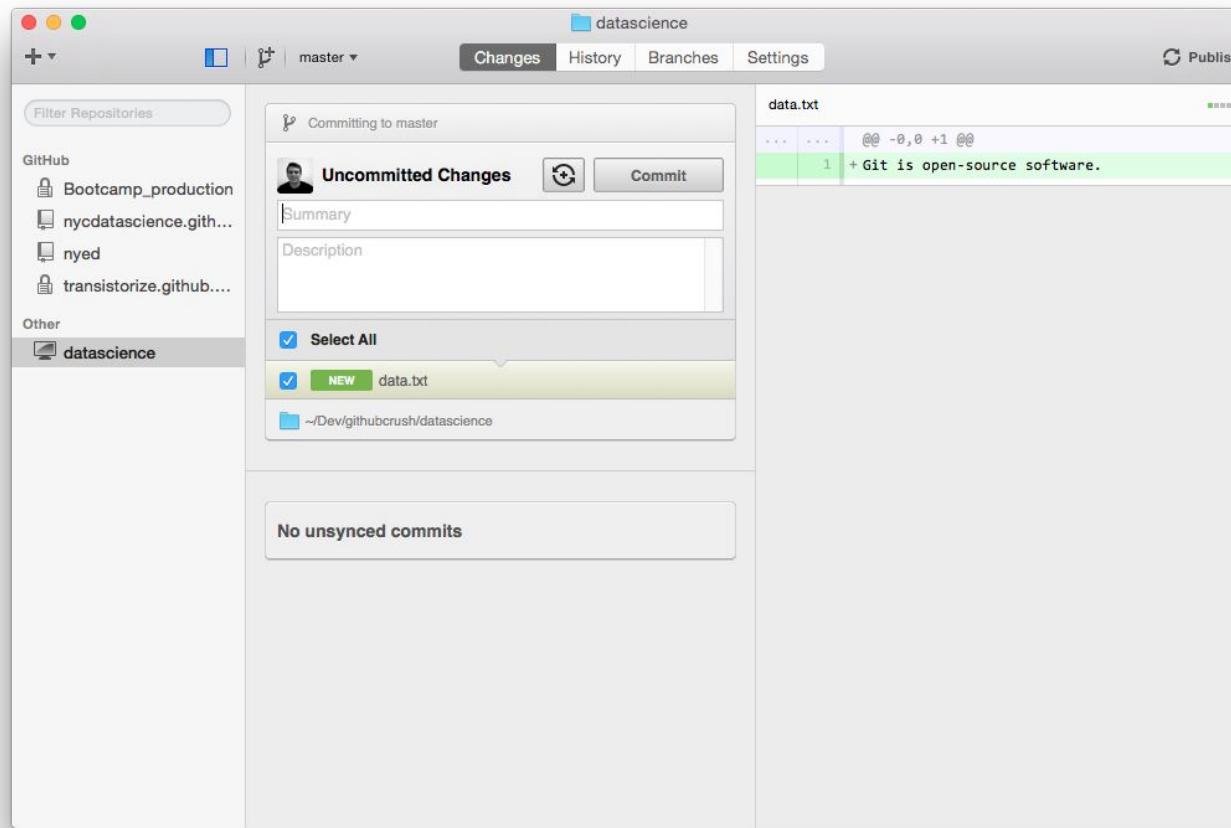
```
Git is open-source software.
```

- ❖ Note: on Windows, make sure Notepad++ is configured with the settings: UTF-8 without BOM.
- ❖ FYI: On Linux and Mac, you can create a file via the command line.

```
$ touch data.txt          // creates an empty file
$ echo "Git is open-source software." > data.txt
$ ls -a                  // lists files in the directory
.   ..      .git           data.txt
```

Status of the Local Repository

- ❖ Verify in the UI that a new file has been created.

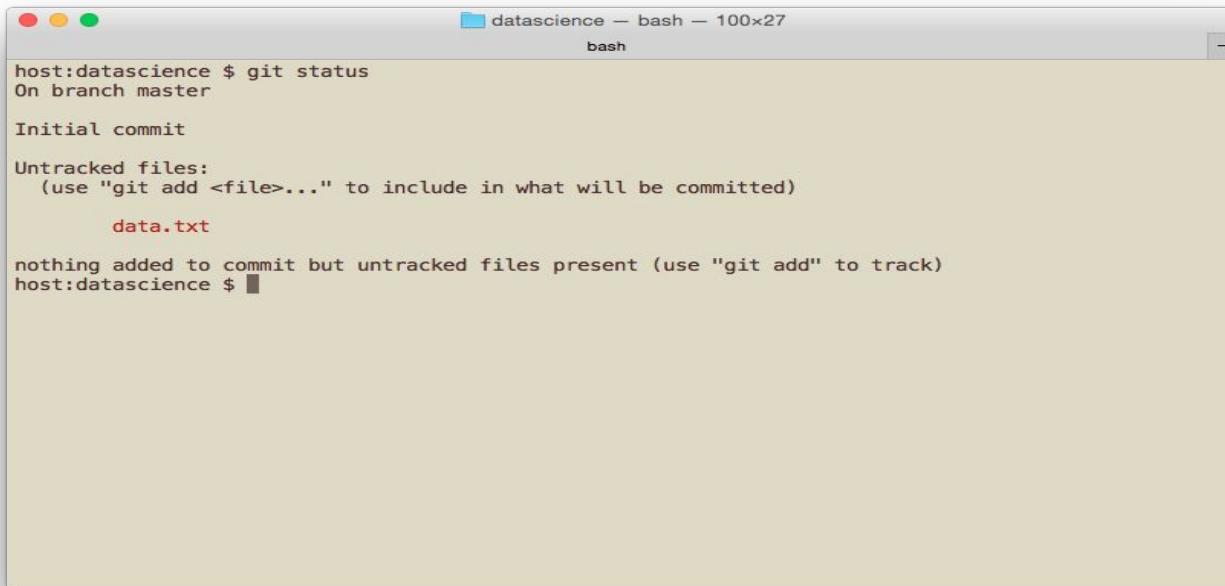


Status of the Local Repository

- ❖ The status command is used to show what changes git has detected in the local working directory.

```
$ git status          //examine git's point of view
```

- ❖ What's going on here? What does it mean to be untracked?



A screenshot of a terminal window titled "data-science — bash — 100x27". The window shows the following text output:

```
host:data-science $ git status
On branch master

Initial commit

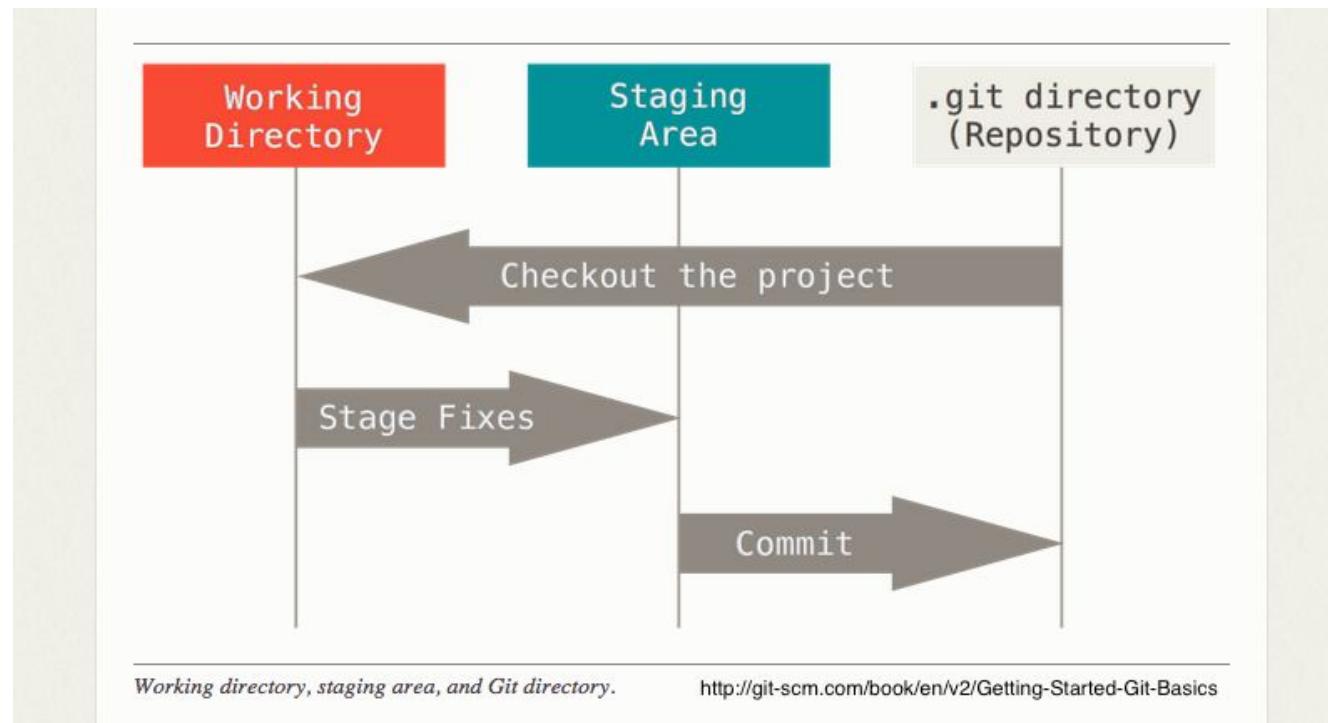
Untracked files:
  (use "git add <file>..." to include in what will be committed)

    data.txt

nothing added to commit but untracked files present (use "git add" to track)
host:data-science $
```

Git Model: Working Directory

- ❖ Untracked or modified items have to be added to the index file in the .git repository, a.k.a. the staging area



- ❖ Note: Later, we will learn how to go straight from creating or modifying a file to commit it.

Add a file to the staging area

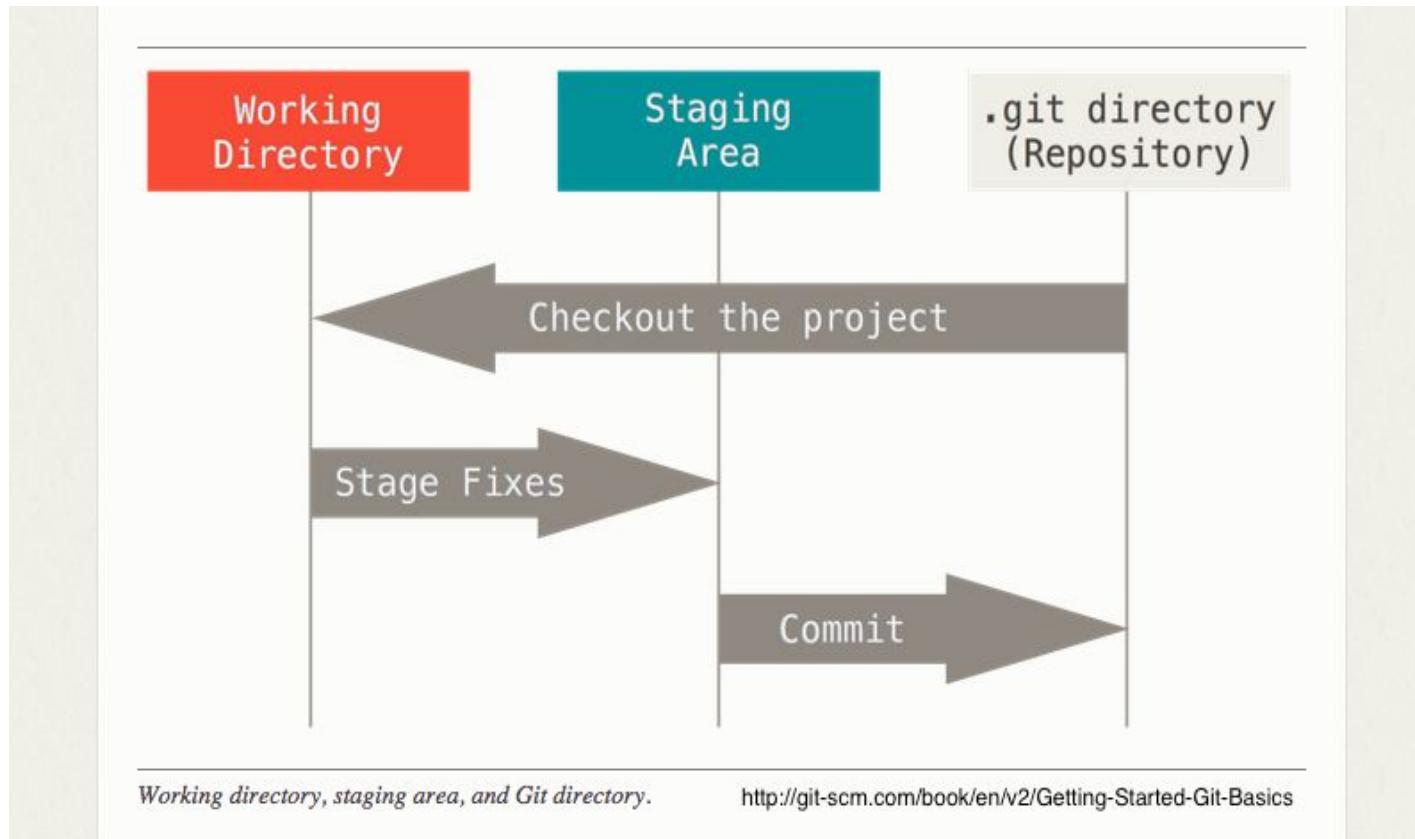
- ❖ **Step 3:** Once the file is created, we must add it to the staging area and then commit the file to the repository.

```
$ git add data.txt      // add file to staging
$ git status            // view status of current working
directory
On branch master        // name of branch currently checked out
Initial commit
Changes to be committed:
  (use "git rm --cached <file>..." to unstage)
    new file:   data.txt
```

- ❖ Note: The command `git status` tells us a new file was added and reminds us to commit it. Also, we can cancel the tracked status of the file by using the command `git rm --cached`

Git Model: Staging Area (Index)

- The staging area (index) tracks all the modified file content.



Commit Content to the Repository

- ❖ **Step 4:** Commit the file to the local repository

```
$ git commit -m "Add new data."  
[master (root-commit) 2a6487f] Add new data to the project.  
 1 file changed, 1 insertion(+)  
   create mode 100644 data.txt  
$ git status  
On branch master  
nothing to commit, working directory clean
```

- ❖ Note : After the parameter -m in git commit, provide a short, clear description of this change. It's makes it easier to find and understand changes in the history log. Also, we know the commit ID of the file on the master branch begins with 2a6487f. The description of this file submitted is 'Add new data to the project'.

Commit History

- ❖ View the commit log to see the recent history

```
$ git log --graph --decorate  
* commit 2a6487f3bd883297451f3c6412b81555ef45942f  
  Author: user.name <my.email@domain.com>
```

Add new data.

- ❖ Note: Try `$ git log --oneline` You can mix flags to find your output preference. See `git help log` for more details.

Exercise 1

- ❖ Create a new directory that is a sibling to the datascience directory. Call it exercise_1. Initialize that directory to be another git repository. If it helps, add that second repository to the UI tool.
- ❖ Create a new empty file in the exercise_1 directory. Add and commit this new file to your second repository.
- ❖ What branch did you commit this change to?
- ❖ Change directories back to the datascience directory. What is the status of the datascience repository?

Change Content in the Repository

- ❖ **Step 1:** Create multiple files using a text editor.
 - Modify data.txt file, and append this line to the file: **R is open-source software.**
 - Create the new file www.txt to have the string 'www'
 - Create the new file test.txt to have the string 'test'
 - Run `$ git status` Are all the files treated the same?

Change Content in the Repository

- ❖ **Step 2:** Try adding multiple files simultaneously to the staging area
- ❖ Using the --all parameter, add all files to the repository simultaneously.

```
$ git add --all
```

- ❖ Run `$ git status` What do you find?

```
$ git status
On branch master
Changes to be committed:
  (use "git reset HEAD <file>..." to unstage)

modified:   data.txt
new file:   test.txt
new file:   www.txt
```

Reset a File from Staging

- ❖ **Step 3:** Reset the files you just added, out of the staging area.

```
$ git reset HEAD data.txt test.txt www.txt
```

```
$ git status
```

- ❖ Note: In this case, this command does not modify the actual files. For now, ignore the mention of the HEAD reference.

Reset a File from Staging

- ❖ **Step 4:** Re-add all the files you just reset, back into the staging area.

```
$ git add data.txt test.txt www.txt
```

```
$ git status
```

Change Content in the Repository

- ❖ **Step 5:** Commit all the files as one change. Pick an appropriate message.

```
$ git commit -m "..."
```

- ❖ When nothing is modified or untracked, it is said you have a clean working directory.

```
$ git status  
On branch master  
nothing to commit, working directory clean
```

Change Content in the Repository

- ❖ Without looking at the slides, what are the steps to creating or modifying content within a repository?
- ❖ Could I add an image to the datascience repository?
- ❖ When you touch a file, does git consider that a modification? How can you tell?

Commit a Modified File in One Command

- ❖ 'Skipping' the add-to-staging part
- ❖ Adding the -a parameter to git commit will commit all known files to the repository simultaneously. A file is known if it has been added and committed to the git index before. Still, be careful when using this combination.
- ❖ Modify test.txt to read as follows:

```
test! test! test!
```

- ❖ Then execute:

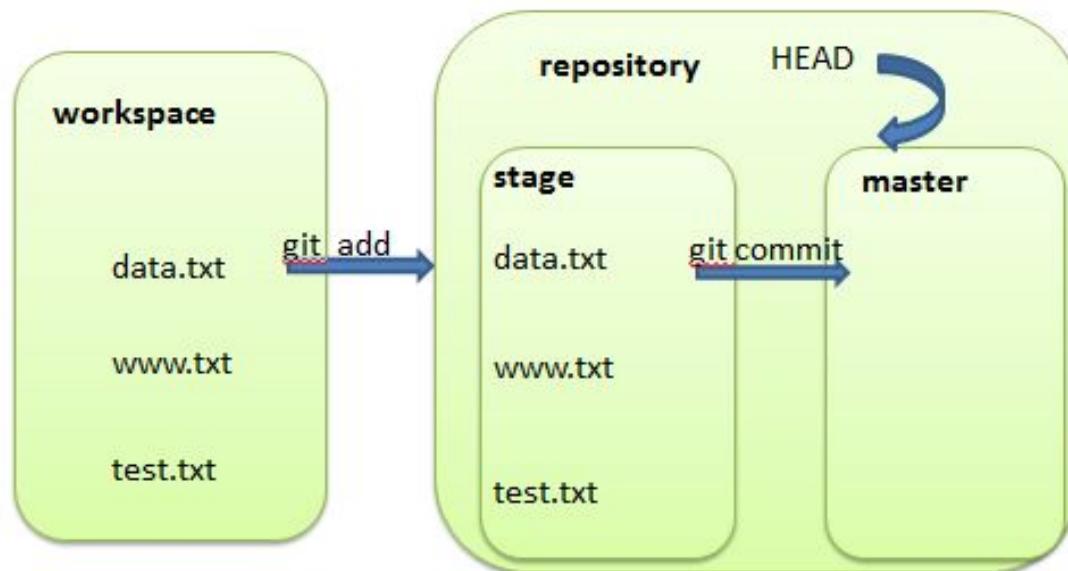
```
$ git commit -a -m "Commit more tests"  
[master ed1d0f8] Commit more tests  
 1 file changed, 1 insertion(+), 1 deletion(-)
```

Quick Review

- ❖ In order to better understand the purpose of Git commands, keep the following model in mind:
 - **Branch:** A stream of file snapshots, like a tree branch. Only one branch is active at a time in a repository, often defaults to the master branch.
 - **Workspace:** What we see in the directory, the datascience folder created before is the working area.
 - **Staging:** Used to prepare a commit of one or more files, and you can review the change list.
 - **Repository:** There is a hidden directory called .git in the workspace, that is the 'real' repository. Keeps track of snapshots, history, the index, the HEAD pointer (references), branches, and more.

Quick Review - Workflow

- ❖ The workflow figure of adding and committing files to repository.
 - Firstly, git add adds files in the workspace to staging.
 - Secondly, git commit commits files in staging to the current branch.



Outline

- ❖ **Set up Git and GitHub**
- ❖ **Introduction to Git**
 - **Creating a Git Repository**
 - **Manipulating files**
- ❖ **Introduction to GitHub**
 - **Lightning Tour of Github**
 - **Create a Remote Repository**

Change a Local Repository - Delete a file

- ❖ We take the test.txt file, committed under datascience, for example.
 - the delete command is rm [file name]

```
$ rm test.txt

$ ls -a
.  ..  .git  data.txt  www.txt
```

Change a Local Repository - Delete a file

- ❖ When you delete the test.txt file, view the current status of the workspace.

```
$ git status
On branch master
Changes not staged for commit:
  (use "git add/rm <file>..." to update what will be committed)
    (use "git checkout -- <file>..." to discard changes in working
     directory)

          deleted:    test.txt

no changes added to commit (use "git add" and/or "git commit
-a")
```

Change a Local Repository - Delete a file

- ❖ See the difference between the current workspace and the latest version of the repository.

```
$ git diff HEAD -- test.txt
diff --git a/test.txt b/test.txt
deleted file mode 100644
index e69de29..0000000
```

Change a Local Repository - Delete a file

- ❖ Command rm test.txt just deleted the file in the workspace, but did not remove it from the repository, so:
 - If you want to restore files to the workspace, use the command:

```
$ git checkout -- test.txt
```

- ❖ Note : Command git checkout -- [filename] can restore the file to the workspace, so you can undo accidental, non-commited changes. The --, in this case, is a generic way to refer to the current branch.

Change a Local Repository - Delete a file

- ❖ If you really want to remove the file from the repository, use the git rm command:

```
$ git rm test.txt  
rm 'test.txt'  
  
$ git commit -m "delete test.txt"  
[master b81549a] delete test.txt  
 1 file changed, 0 insertions(+), 0 deletions(-)  
 delete mode 100644 test.txt
```

- ❖ Note: When you delete a file, commit the updated status of the file. When you do this, you are removing the file from future snapshots, but the previous versions of the file are still in the history where it can be retrieved.

Change a Local Repository - View Changes

- ❖ We completely overwrite the contents of data.txt, by changing its contents to:

```
R is software.
```

```
R is free software.
```

- ❖ Before adding file to the staging area, view its status:

```
$ git status
```

```
On branch master
```

```
Changes not staged for commit:
```

```
  (use "git add <file>..." to update what will be committed)
```

```
  (use "git checkout -- <file>..." to discard changes in working  
  directory)
```

```
    modified:   data.txt
```

```
no changes added to commit (use "git add" and/or "git commit -a")
```

Change a Local Repository - View Changes

- ❖ After modifying the file within the workspace, view the status of the file.
 - if the file is unstaged, git diff views the difference between the current workspace and last known snapshot:

```
$ git diff -- data.txt
diff --git a/data.txt b/data.txt
index 15b81a8..a5c20b1 100644
--- a/data.txt
+++ b/data.txt
@@ -1,2 +1,2 @@
-Git is open-source software.
-R is open-source software.
+R is software.
+R is free software.
```

Change a Local Repository - View Changes

- ❖ After modifying the file within the workspace, view the status of the file.
 - if the file is staged already, have to add the --cached parameter:

```
$ git add data.txt

$ git diff data.txt

$ git diff --cached data.txt
diff --git a/data.txt b/data.txt
index 15b81a8..a5c20b1 100644
--- a/data.txt
+++ b/data.txt
@@ -1,2 +1,2 @@
-Git is open-source software.
-R is open-source software.
+R is software.
+R is free software.
```

Change a Local Repository - View history

- ❖ Firstly, we add the modified file data.txt and commit it to the repository.

```
$ git add data.txt  
$ git commit -m "Add free to data"
```

- ❖ Note: Under Windows, sometimes adding a file to staging outputs the following warning:

```
$ git add data.txt  
warning: LF will be replaced by CRLF in data.txt.  
The file will have its original line endings in your working  
directory.
```

- ❖ It doesn't matter, just ignore it. You can enforce this behavior by issuing the command:

```
$ git config core.autocrlf true
```

Change a Local Repository - View history

- ❖ If we don't want too much information, output can be controlled by the parameters --oneline, or to see the full SHA-1 commit IDs, use --pretty=oneline.

```
$ git log --pretty=oneline  
96f01db1d6e026b7284a7d1267c1adb70b20aa42 Add free to data  
063624c7165dc5a6c6f5949ab7cc82a7d8a28036 delete test.txt  
ed1d0f8ad270606bda51290a1bfed37f40f15469 Commit more tests  
4d8db1c6129c836e48c36d265bcb938bc07abf21 Create three more  
files.  
a638a8aab1a396c6cbd8ab425e8a8976af498dfb Add new data.
```

- ❖ Note: Output information may not look exactly as above, the important part is that only the commit IDs and description are printed.

Change a Local Repository - Rename a file

- ❖ Git can track file renames and also handle moving ('mv') files and directories about.

```
$ git mv www.txt www2.txt // make sure www.txt was committed  
$ git status  
On branch master  
Changes to be committed:  
(use "git reset HEAD <file>..." to unstage)  
  
renamed:    www.txt -> www2.txt  
...
```

- ❖ Note: Note how the rename automatically added itself to the staging area. Now commit it.

Hide files from Git

- ❖ Sometimes, we want to tell Git to ignore certain files. These could be generated files, files with secret keys, or special IDE files that should not be committed. The `.gitignore` exists to hide these kind of files and prevent mistakes.

```
$ echo ".DS_Store" > .gitignore          // Ignore common OSX file.  
$ echo "Thumbs.db" > .gitignore          // Ignore common Windows  
file.  
$ git add .gitignore; git commit -m "add gitignore file"  
$ git status
```

- ❖ Note: `.gitignore` files support a rich set of matching patterns, to match multiple files at any folder depth.

Wrapping Up Part 1

- ❖ Covered primary local commands
 - help, config, log, diff, status
 - init, add, commit, rm, mv
- ❖ practice
 - Read the freely available GitSCM book <http://git-scm.com/book>, chapters 1 & 2

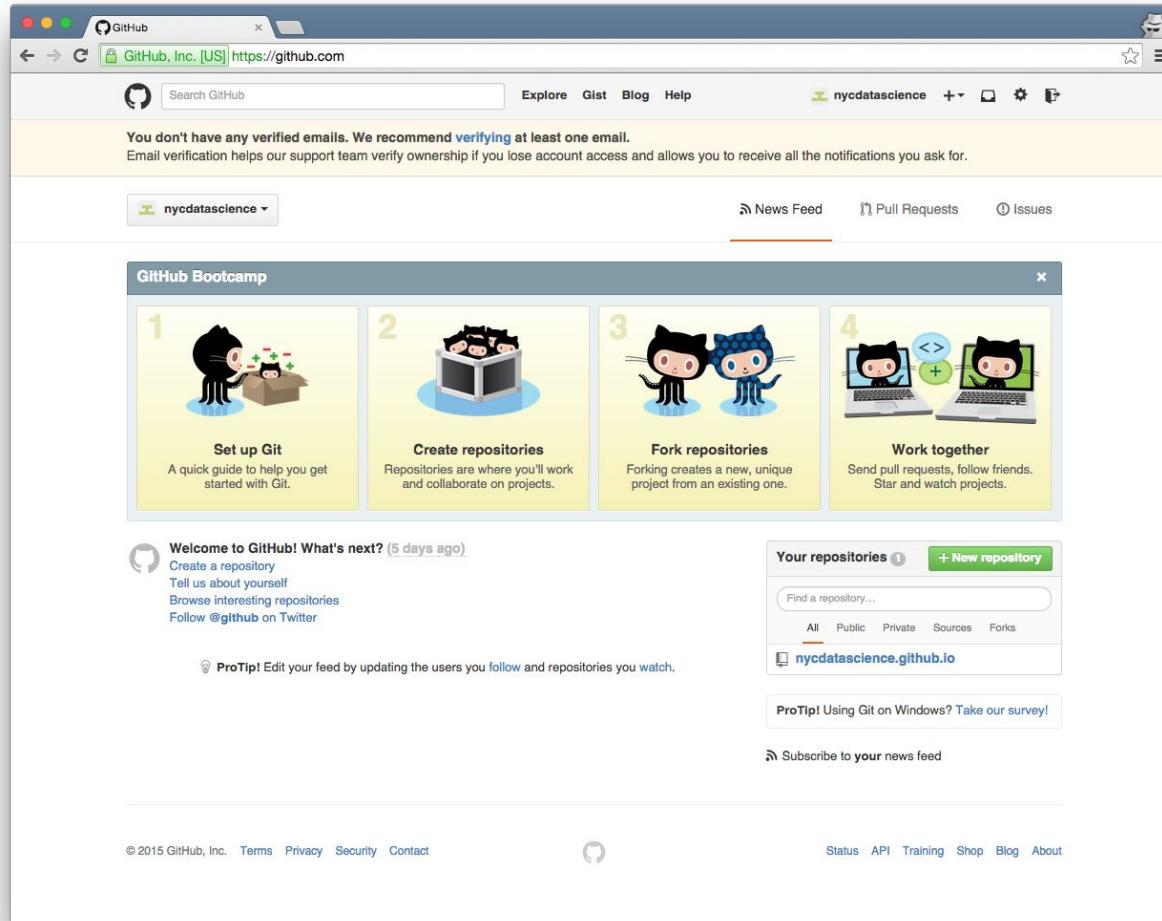
Outline

- ❖ **Set up Git and GitHub**
- ❖ **Introduction to Git**
 - **Creating a Git Repository**
 - **Manipulating files**
- ❖ **Introduction to GitHub**
 - **Lightning Tour of Github**
 - **Create a Remote Repository**

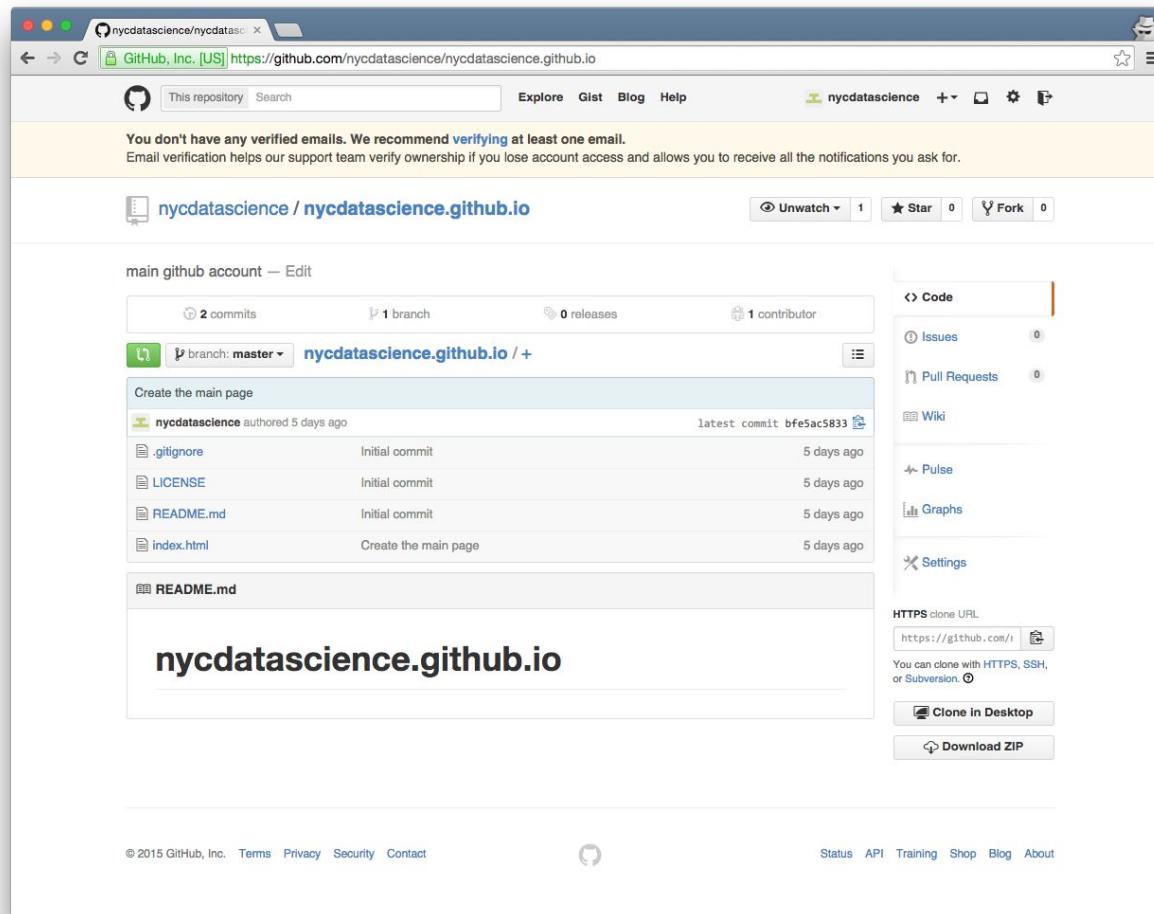
SSH Keys

- ❖ If sticking with the command line, you need to manually configure SSH keys.
- ❖ Please follow the link to generate an SSH key from your local computer:
 - <https://help.github.com/articles/generating-an-ssh-key/>
- ❖ To add your SSH key to your github account:
 - <https://help.github.com/articles/adding-a-new-ssh-key-to-your-github-account/>

Quick Feature Tour - Main Page



Quick Feature Tour - Personal GitHub Page



Quick Feature Tour - Organization Page

The screenshot shows a GitHub organization page for "The Apache Software Foundation". The page features a header with the Apache logo and navigation links for Explore, Gist, Blog, and Help. A sidebar on the left lists repositories: "infrastructure-puppet", "incubator-nifi", "ambari", and "pdfbox". Each repository entry includes a thumbnail, the name, a star icon, a commit count, and a last updated timestamp. To the right of the repositories is a "People" section displaying a grid of 156 profile pictures. At the bottom of the page is a footer with the URL "https://github.com/apache".

Quick Feature Tour - Forking Projects

The screenshot shows the GitHub repository page for `apache/hadoop`. The repository is described as a "Mirror of Apache Hadoop". Key statistics displayed include 9,837 commits, 131 branches, 199 releases, and 40 contributors. The main commit list is filtered to show changes for the `trunk` branch. A purple box highlights the top navigation bar, specifically the "Watch" (118), "Star" (328), and "Fork" (399) buttons.

Commit	Description	Time Ago
HADOOP-11473. test-patch says "-1 overall" even when all checks are +1	17 days ago	
HADOOP-10530 Make hadoop build on Java7+ only (stevel)	2 months ago	
HADOOP-11412 POMs mention "The Apache Software License" rather than "..."	a month ago	
HADOOP-9907. Webapp http://hostname:port/metrics link is not working....	2 hours ago	
HADOOP-11268. Remove no longer supported activation properties for pa...	3 months ago	
HDFS-7603. The background replication queue initialization may not le...	an hour ago	
HADOOP-9907. Webapp http://hostname:port/metrics link is not working....	2 hours ago	
HADOOP-11419 improve hadoop-maven-plugins. (Hervé Boute my via stevel)	4 days ago	
HADOOP-11412 POMs mention "The Apache Software License" rather than "..."	a month ago	
Merge from trunk to branch	6 months ago	
HADOOP-10574. Bump the maven plugin versions too -moving the numbers ...	3 hours ago	
HDFS-7566. Remove obsolete entries from hdfs-default.xml (Ray Chiang ...)	2 days ago	
HADOOP-9907. Webapp http://hostname:port/metrics link is not working....	2 hours ago	
HADOOP-10040. svn propset to native line endings on Windows files. Co...	a year ago	
HADOOP-10714. AmazonS3Client.deleteObjects() need to be limited to 10...	3 months ago	
BUILDING.txt	HADOOP-11428. Remove obsolete reference to Cygwin in BUILDING.txt. Co...	
LICENSE.txt	HADOOP-11184. Update Hadoop's lz4 to r123 (cmccabe)	

Quick Feature Tour - Forking Projects

The screenshot shows a GitHub fork page for the 'hadoop' repository. The repository is owned by 'nycdatascience' and is a mirror of the Apache Hadoop project. The main statistics are 9,837 commits, 131 branches, 199 releases, and 40 contributors. The current branch is 'trunk'. A list of recent commits is displayed, including:

- HDFS-7603. The background replication queue initialization may not be... (Kihwal Lee, 1 hour ago)
- HADOOP-11473. test-patch says "-1 overall" even when all checks are +1 (dev-support, 17 days ago)
- HADOOP-10530 Make hadoop build on Java7+ only (stevel, 2 months ago)
- HADOOP-11412 POMs mention "The Apache Software License" rather than "... (hadoop-client, a month ago)
- HADOOP-9907. Webapp http://hostname:port/metrics link is not working.... (hadoop-common-project, 2 hours ago)
- HADOOP-11266. Remove no longer supported activation properties for pa... (hadoop-dist, 3 months ago)
- HDFS-7603. The background replication queue initialization may not be... (hadoop-hdfs-project, an hour ago)
- HADOOP-9907. Webapp http://hostname:port/metrics link is not working.... (hadoop-mapreduce-project, 2 hours ago)
- HADOOP-11419 improve hadoop-maven-plugins. (Hervé Bouteemy via stevel, 4 days ago)
- HADOOP-11412 POMs mention "The Apache Software License" rather than "... (hadoop-minicluster, a month ago)
- Merge from trunk to branch (hadoop-project-dist, 6 months ago)
- HADOOP-10574. Bump the maven plugin versions too -moving the numbers ... (hadoop-project, 3 hours ago)
- HDFS-7566. Remove obsolete entries from hdfs-default.xml (Ray Chiang ..., 2 days ago)
- HADOOP-9907. Webapp http://hostname:port/metrics link is not working.... (hadoop-tools, 2 hours ago)
- HADOOP-11412 POMs mention "The Apache Software License" rather than "... (hadoop-yarn-project, a year ago)
- .gitattributes (HADOOP-10040. svn propset to native line endings on Windows files. Co..., 3 months ago)
- .gitignore (HADOOP-10714. AmazonS3Client.deleteObjects() need to be limited to 10..., a month ago)

On the right side, there are links for 'Code', 'Pull Requests', 'Pulse', 'Graphs', 'Settings', and download options ('Clone In Desktop', 'Download ZIP').

Quick Feature Tour - Pull Requests

The screenshot shows the GitHub interface for the `apache/hadoop` repository. The title bar indicates the current view is "Pull Requests". The main header includes the repository name "apache / hadoop", a star count of 328, and a fork count of 399. Below the header, there are tabs for "Pull requests" (which is selected), "Labels", and "Milestones". A search bar with the query "is:pr is:open" is present, along with a "New pull request" button. The main content area displays a list of 8 open pull requests:

- #13 Merging from apache trunk (opened on Dec 22, 2014 by vinayrpert)
- #12 Hdfs ec (opened on Dec 3, 2014 by divyam)
- #8 Branch 2 (opened on Nov 6, 2014 by peiyuefeng)
- #7 YARN-1964 Launching containers from docker (opened on Oct 6, 2014 by ashahab-alticas)
- #6 YARN-1964 Launching containers from docker (opened on Sep 28, 2014 by ashahab-alticas)
- #4 Update TaskInputOutputContext.java javadoc (opened on Sep 25, 2014 by sebastiancadena)
- #3 [HADOOP-10724] better interoperation with `sort -h` (opened on Sep 23, 2014 by sam-s)
- #1 MAPREDUCE-6096.SummarizedJob Class Improvement (opened on Sep 18, 2014 by piaoyu)

A "ProTip!" message at the bottom left says: "Adding no:label will show everything without a label." The footer contains links for "Status", "API", "Training", "Shop", "Blog", and "About".

Quick Feature Tour - Example Branching Structure

The screenshot shows a GitHub page for the Apache Cassandra repository. The URL is <https://github.com/apache/cassandra/blob/trunk/README.asci>. The page displays the README file content, which includes a sidebar for switching branches and tags. The 'trunk' branch is currently selected. The content of the README file discusses partitioning and row store, and provides links to requirements and getting started guides.

Switch branches/tags

Filter branches/tags

Branches Tags

- cassandra-1.0
- cassandra-1.1
- cassandra-1.2
- cassandra-2.0
- cassandra-2.1
- trunk**

Rows are organized into tables with a required primary key.

Partitioning means that Cassandra can distribute your data across multiple machines in an application-transparent matter. Cassandra will automatically repartition as machines are added and removed from the cluster.

Row store means that like relational databases, Cassandra organizes data by rows and columns. The Cassandra Query Language (CQL) is a close relative of SQL.

For more information, see the [Apache Cassandra web site](#).

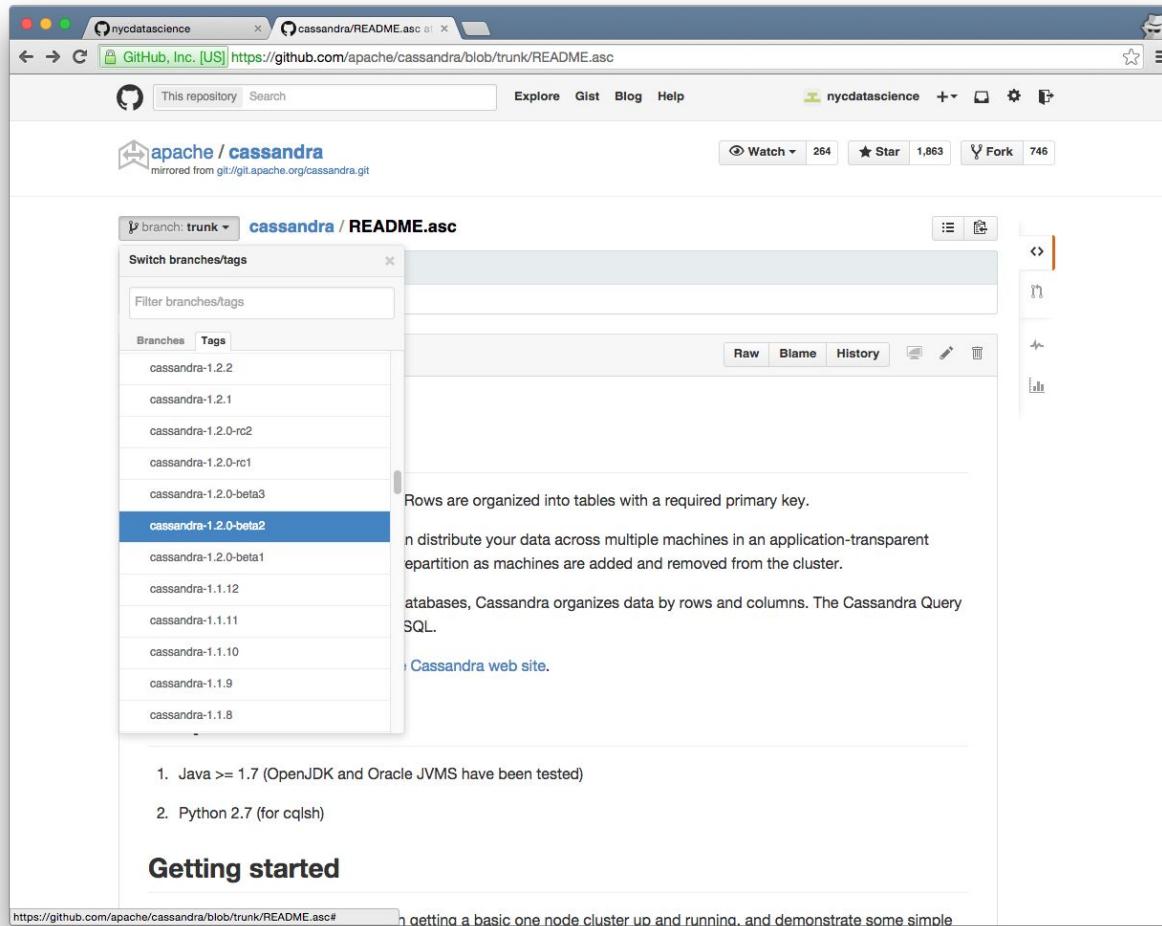
Requirements

1. Java >= 1.7 (OpenJDK and Oracle JVMS have been tested)
2. Python 2.7 (for cqlsh)

Getting started

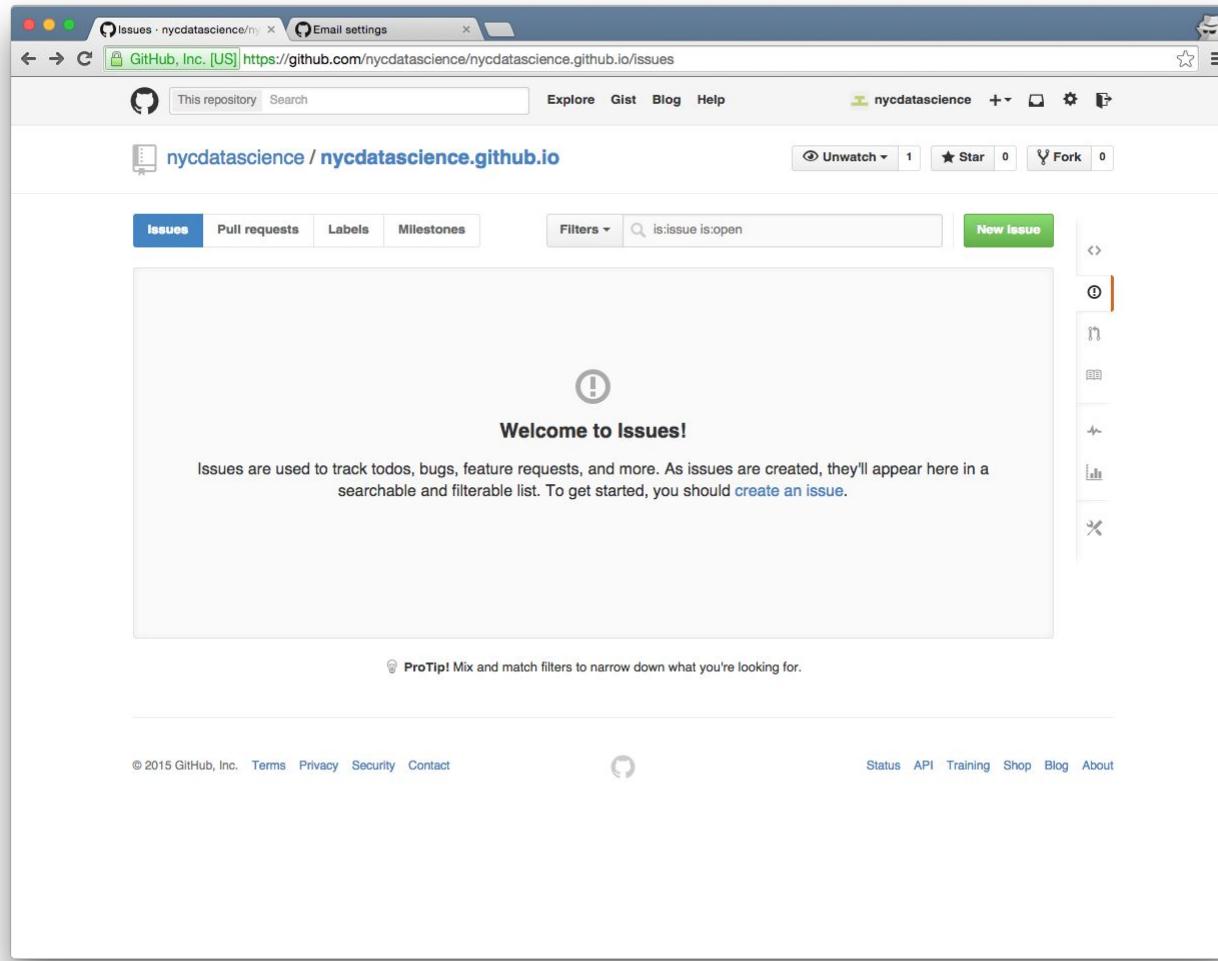
This short guide will walk you through getting a basic one node cluster up and running, and demonstrate some simple

Quick Feature Tour - Example Tagging Structure



The screenshot shows a GitHub repository page for 'apache / cassandra'. The URL is <https://github.com/apache/cassandra/blob/trunk/README.asc>. The page displays the contents of the README.asc file. A dropdown menu is open over the word 'cassandra' in the first line of the file, showing a list of branches and tags. The 'cassandra-1.2.0-beta2' tag is selected and highlighted in blue. The dropdown menu has a 'Filter branches/tags' input field and tabs for 'Branches' and 'Tags'. The main content area shows the file's text, which includes information about primary keys, data distribution, and the Cassandra Query Language (CQL). Below the file content, there is a 'Getting started' section with two bullet points: '1. Java >= 1.7 (OpenJDK and Oracle JVMs have been tested)' and '2. Python 2.7 (for cqlsh)'. At the bottom of the page, there is a navigation bar with links for 'Explore', 'Gist', 'Blog', and 'Help', and a sidebar with user profile information.

Quick Feature Tour - Issue Tracking



Quick Feature Tour - Issue Tracking

The screenshot shows the GitHub issue creation interface for the repository `nycdatascience/nycdatascience.github.io`. The page title is "New Issue · nycdatascience". The top navigation bar includes links for Explore, Gist, Blog, Help, and the repository name `nycdatascience`. A message at the top encourages email verification. The main content area shows an issue titled "Missing student pages for G001" with the following details:

- Assignee: No one is assigned
- Milestone: No milestone
- Description:

Missing Student Profiles

This issue will be solved in class as part of the tutorial.
- Labels: enhancement (selected), bug, duplicate, help wanted, invalid, question, wontfix

At the bottom of the form is a "Submit new issue" button. The footer of the page includes links for Terms, Privacy, Security, Contact, Status, API, Training, Shop, Blog, and About, along with a GitHub icon. The URL in the address bar is `https://github.com/nycdatascience/nycdatascience.github.io/milestones`.

Quick Feature Tour - Issue Tracking

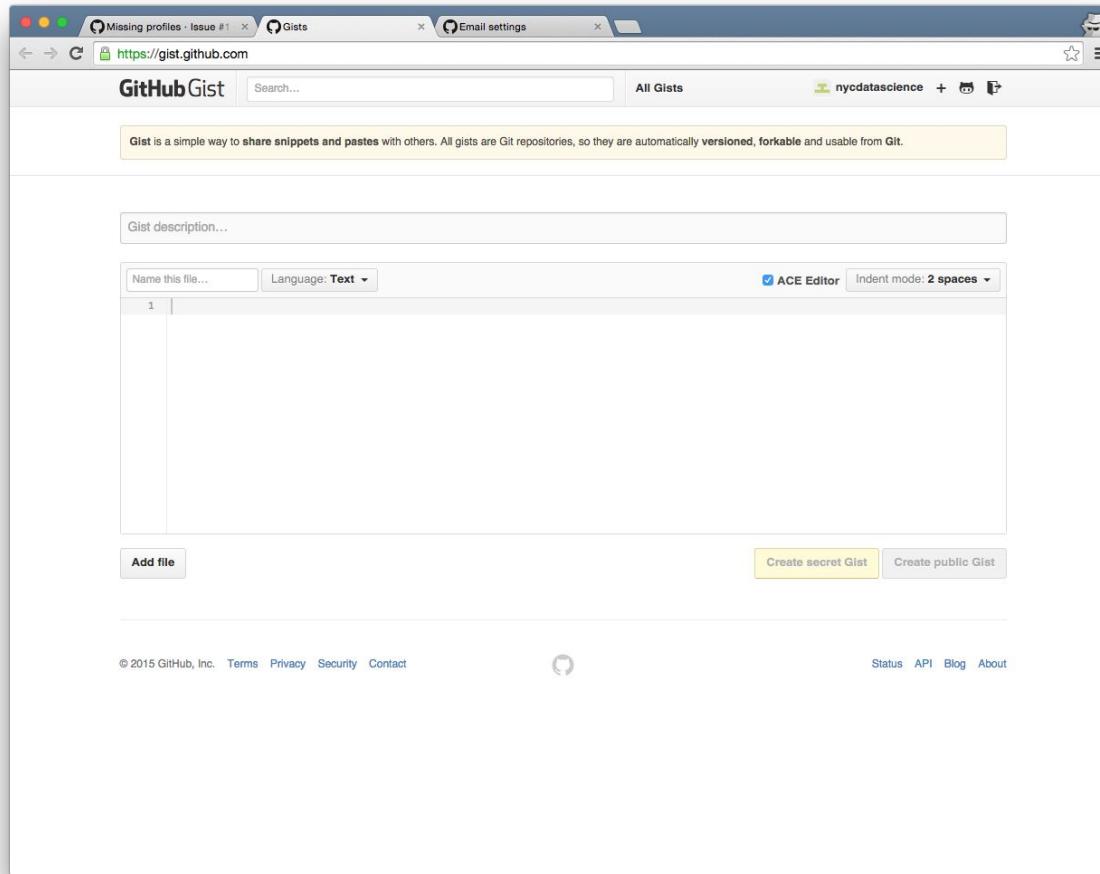
The screenshot shows a GitHub issue page for a repository named `nycdatascience / nycdatascience.github.io`. The issue is titled "Missing profiles #1". A green button indicates it was opened by `nycdatascience` a minute ago with 0 comments. The main content of the issue is a comment from `nycdatascience` stating "Missing Student Profiles" and noting that the issue will be solved in class as part of the tutorial. This comment was added a minute ago. The issue has been labeled "enhancement" and assigned to the "M1" milestone. The sidebar on the right provides options to edit the issue, change labels, set milestones, assignees, and notifications. It also shows that there is 1 participant and a link to unsubscribe from notifications. At the bottom, there is a comment input field with "Write" and "Preview" tabs, a note about Markdown support, and buttons for "Close issue" and "Comment".

Quick Feature Tour - Wiki Pages

The screenshot shows a GitHub repository page for `thinkaurelius/titan`. The main content area displays the `Home` page of the wiki. It features a large image of the Greek titan Atlas holding up the celestial spheres, with the word "TITAN" written in large, bold, capital letters below it. A text block describes Titan as a distributed graph database optimized for storing and querying graphs over a cluster of machines, mentioning its compatibility with various database technologies like Apache Cassandra, Apache HBase, and Oracle BerkeleyDB. Below the text is a red "DOWNLOAD" button with a small icon. On the right side, there is a sidebar titled "Pages" containing a list of wiki pages such as Acknowledgements, Advanced Blueprints, Advanced Indexing, Beginner's Guide, and many others. At the bottom of the sidebar, there is a link to "Show 35 more pages...". The top of the page includes the GitHub header with navigation links like Explore, Gist, Blog, Help, and the user's profile "nycdatascience". The URL in the address bar is <https://github.com/thinkaurelius/titan/wiki>.

Quick Feature Tour - Gist

- ❖ Gist: Sharing quick snippets of code - <https://gist.github.com/>



Disk Quotas

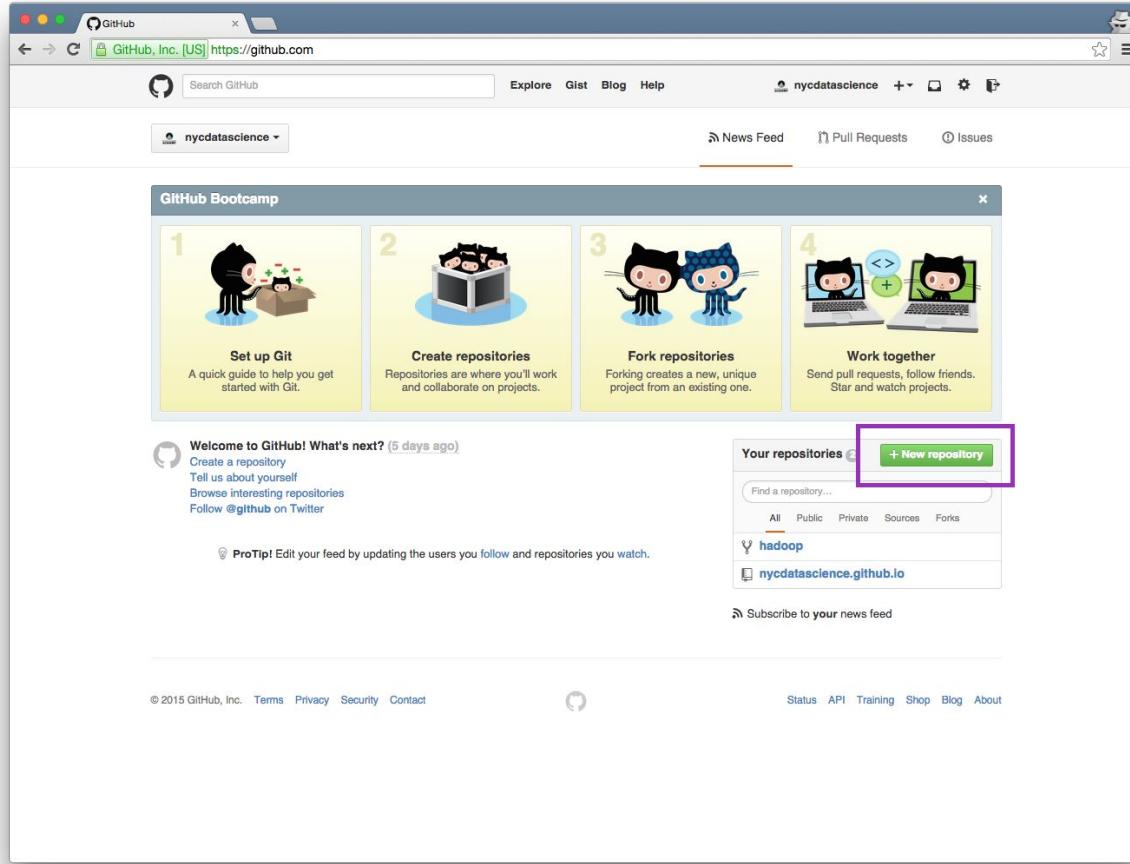
- ❖ How much information can you store on GitHub?
- ❖ Try to keep large datasets out of Git, and try to keep repositories below 100 MB. Instead, link to where a person can find a secure copy. See this help article.
- ❖ Also, for really large data sets, consider using BitTorrent, GoogleDrive, or DropBox with SHA-256 checksums for security.

Outline

- ❖ **Set up Git and GitHub**
- ❖ **Introduction to Git**
 - **Creating a Git Repository**
 - **Manipulating files**
- ❖ **Introduction to GitHub**
 - **Lightning Tour of Github**
 - **Create a Remote Repository**

Create a Remote Repository

- ❖ Look for the **+ New repository** button.



Create a Remote Repository

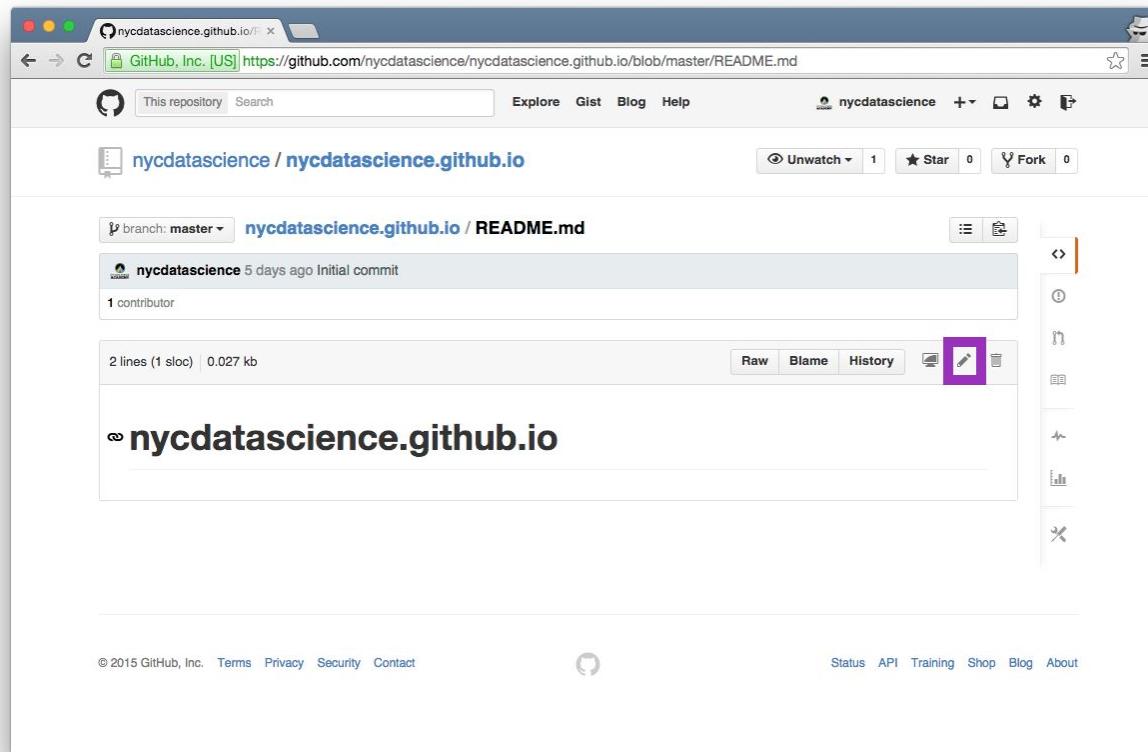
- ❖ For the repository name, create the domain name for your personal page, using your personal github username, in the following format
your_username.github.io
- ❖ username: nycdatascience

```
repository name = nycdatascience.github.io
```

- ❖ Click on the checkbox Initialize this repository with a README. The license choice is up to you.
- ❖ Click the Create repository button.

Edit From the Site

- ❖ Once the new repository has been created, click on your repository's README.md file. For example, this was the NYC Data Science README:



Edit From the Site

- ❖ Markdown <https://help.github.com/articles/github-flavored-markdown/>
- ❖ Change the contents of the README to have a good description, optionally using markdown like in the following:

```
# my_user_name.github.io
```

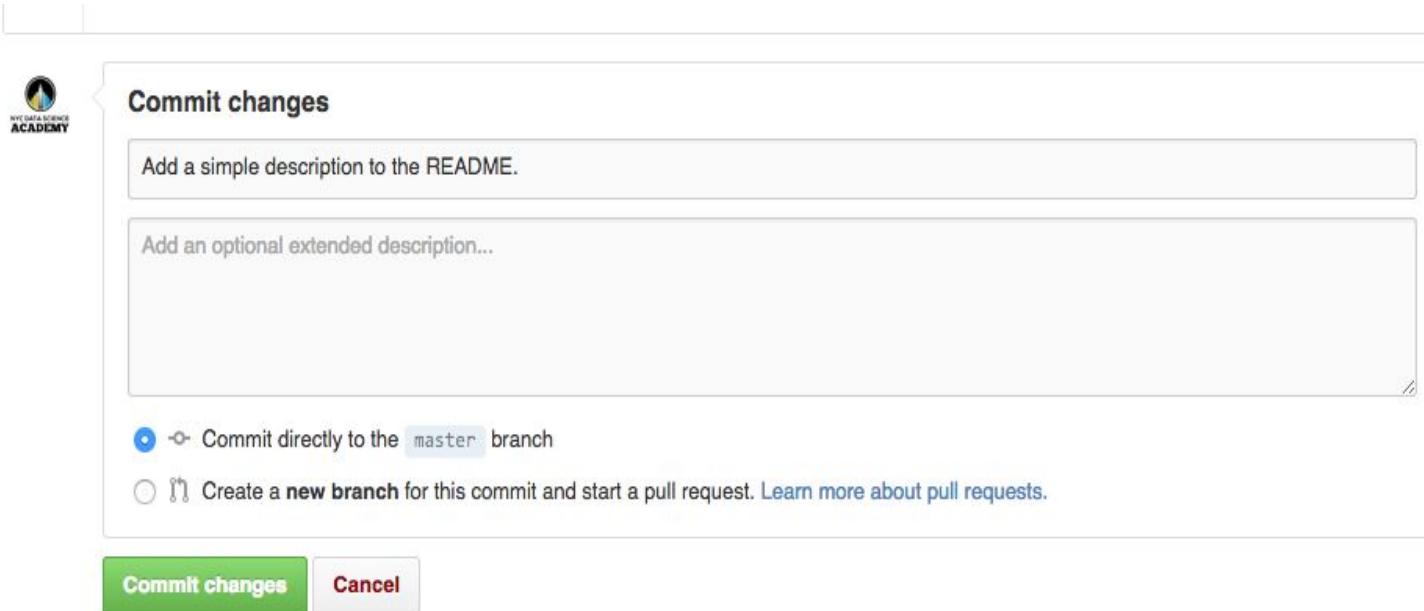
```
Copyright @ My Name
```

```
## Description
```

```
This will be the main portfolio page for the My Name. I am  
currently located in New York City.
```

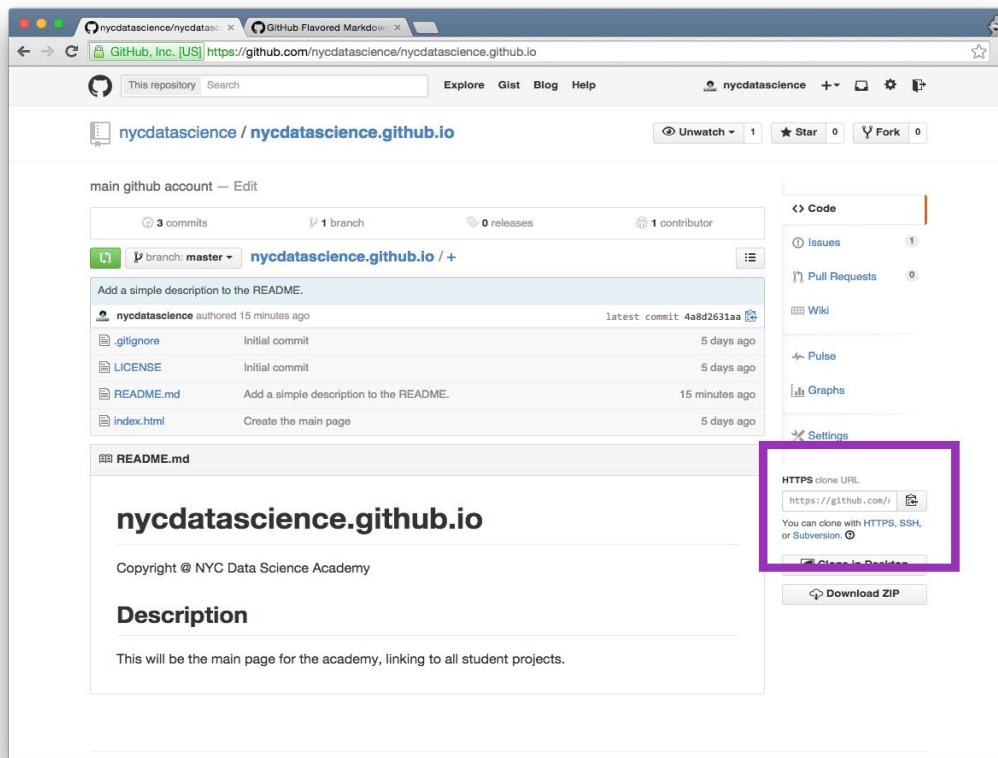
Edit From the Site

- ❖ Scroll down, add a commit message, and hit the **Commit changes** button. Leave the change to be made against the master branch.



Cloning

- ❖ Go back to the main page for the repository, and look for the remote HTTPS/SSH address of your repository:



Cloning

- ❖ We want to clone the repository using one of the following addresses, but your repository.
- ❖ HTTPS:

```
https://github.com/nycdatascience/nycdatascience.github.io.git
```

- ❖ Alternatives:

```
SSH: git@github.com:nycdatascience/nycdatascience.github.io.git
```

```
Subversion:
```

```
https://github.com/nycdatascience/nycdatascience.github.io
```

Cloning

- ❖ Copy the HTTPS address (the default), and open the local command line.

```
$ git clone  
https://github.com/<username>/<username>.github.io.git  
Cloning into 'nycdatascience.github.io'...  
remote: Counting objects: 11, done.  
remote: Compressing objects: 100% (9/9), done.  
remote: Total 11 (delta 2), reused 0 (delta 0)  
Unpacking objects: 100% (11/11), done.  
Checking connectivity... done.
```

New Local Repository

- ❖ Change directories into the folder that was cloned locally. Be sure to substitute your username in the command below.

```
$ cd <username>.github.io  
$ ls  
README.md
```

Remotes

- ❖ Repositories cloned from GitHub automatically have the 'remote' setting put into the local configuration.

```
$ git remote -v
origin
https://github.com/nycdatascience/nycdatascience.github.io.git
(fetch)
origin
https://github.com/nycdatascience/nycdatascience.github.io.git
(push)
```

- ❖ Note: Run a git status to see what the set branch is.

Remotes

- ❖ When the local repository and remote repository are both created independently, how can you associate them?
- ❖ Associate the local and remote database versions, by specifying a simple name instead of the full remote address. Use the command
 - \$ git remote add [simple name] [url]
- ❖ In the previous example, this was not necessary as the origin was set already.

```
// https protocol
$ git remote add origin
https://github.com/nycdatascience/nycdatascience.github.io.git
// or ssh protocol
$ git remote add origin
git@github.com:nycdatascience/nycdatascience.github.io.git
```

Pull

- ❖ Say you want check for the latest changes on the remote server. First, go back to GitHub, and modify the README.md file again. Then type one of the following.

```
$ git pull origin master
remote: Counting objects: 3, done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 3 (delta 2), reused 0 (delta 0)
Unpacking objects: 100% (3/3), done.
From https://github.com/nycdatascience/nycdatascience.github.io
  4a8d263..f527c76  master      -> origin/master
Updating 4a8d263..f527c76
Fast-forward
 README.md | 2 ++
 1 file changed, 1 insertion(+), 1 deletion(-)
```

Push

- ❖ Now we want to publish a local edit back to the remote GitHub repository. Create a new file in the local repository called `index.html`, with some sample content.

```
Hello, World!
```

Push

- ❖ Commit the new content to your local repository.

```
$ git add index.html  
$ git commit -m "Add the primary webpage for the portfolio."
```

Push

- ❖ Commit the new content to your remote personal repository's master branch.

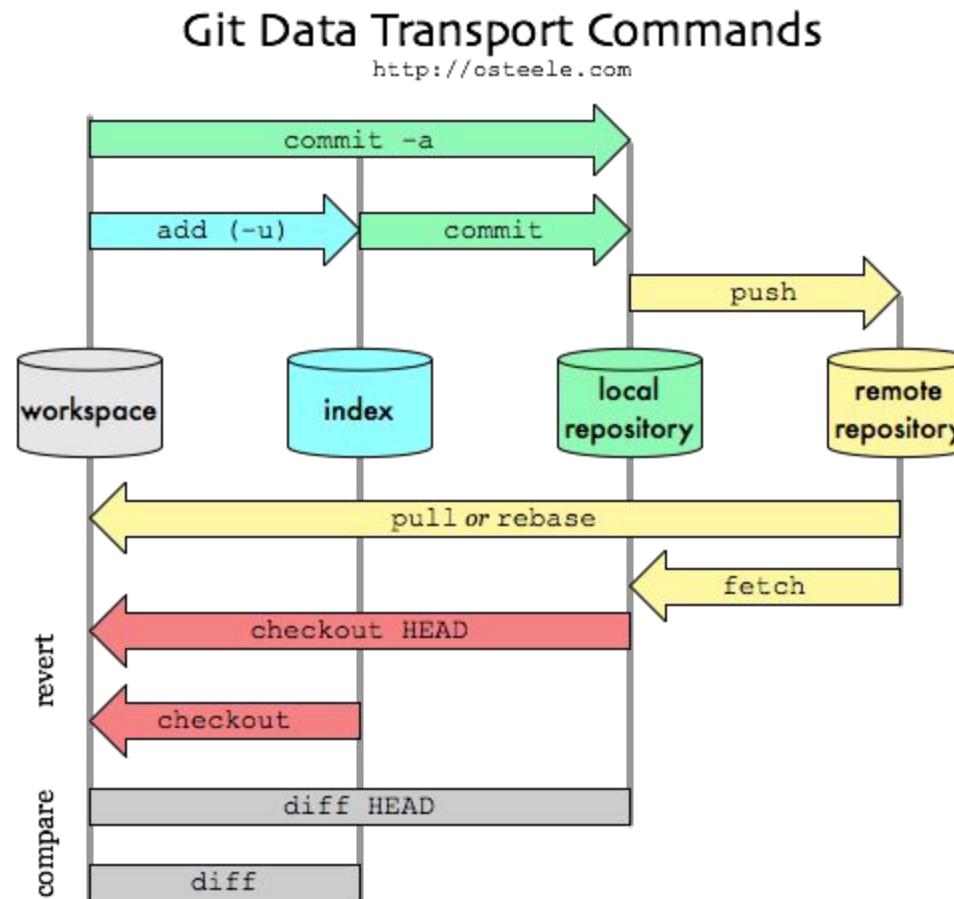
```
$ git push origin master
```

- ❖ If there are errors, then there is a problem with your setup. If not, wait a few seconds, and then refresh your GitHub page. Visit your actual portfolio page at <https://username.github.io/> where username is your Github username.

Wrapping Up Part 2

- ❖ Covered primary local commands
 - push, pull
- ❖ Practice!
 - Read the freely available GitSCM book <http://git-scm.com/book>, chapters 3, 5 - 7.

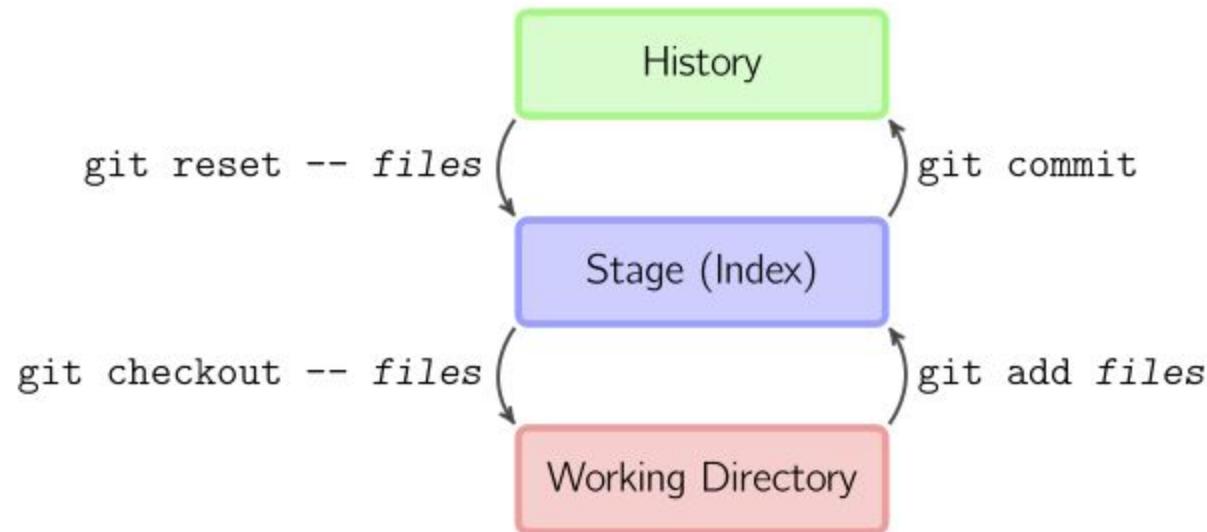
Workflow of Git/GitHub



Git cheatsheet: <http://ndpssoftware.com/git-cheatsheet.html>

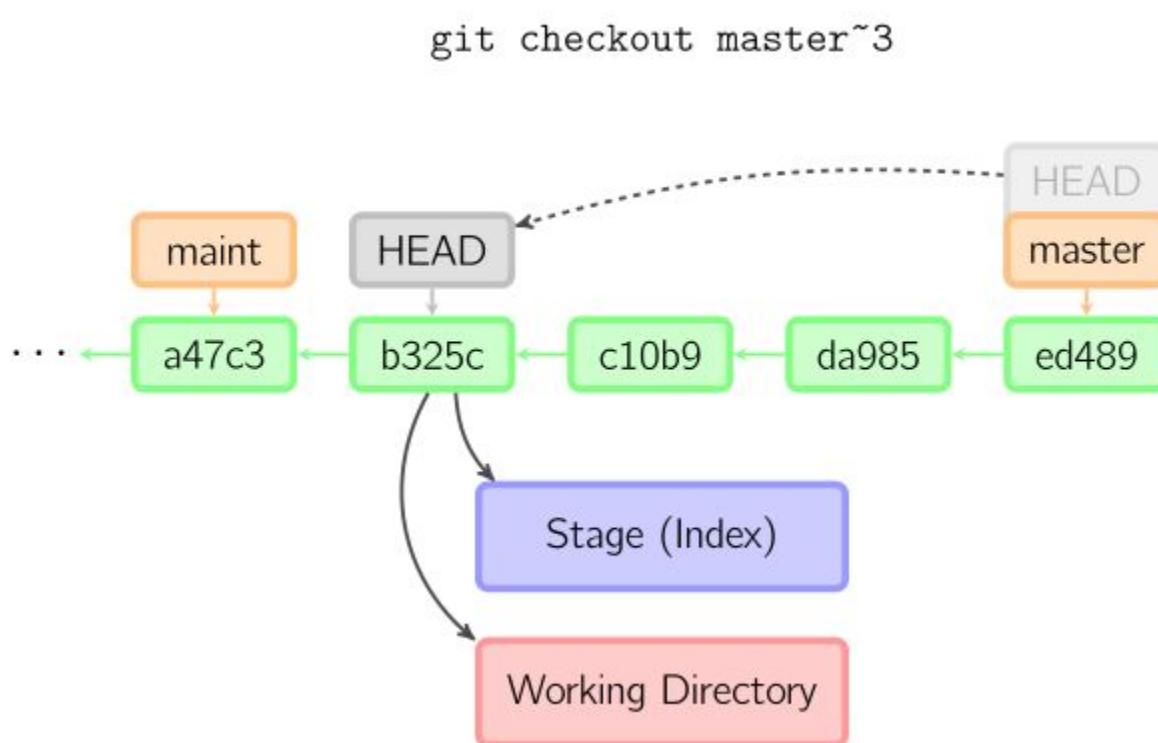
Workflow of Git/GitHub - Checkout & Reset

To reverse the changes made by *add* or *commit*:



Workflow of Git/GitHub - Checkout & Reset

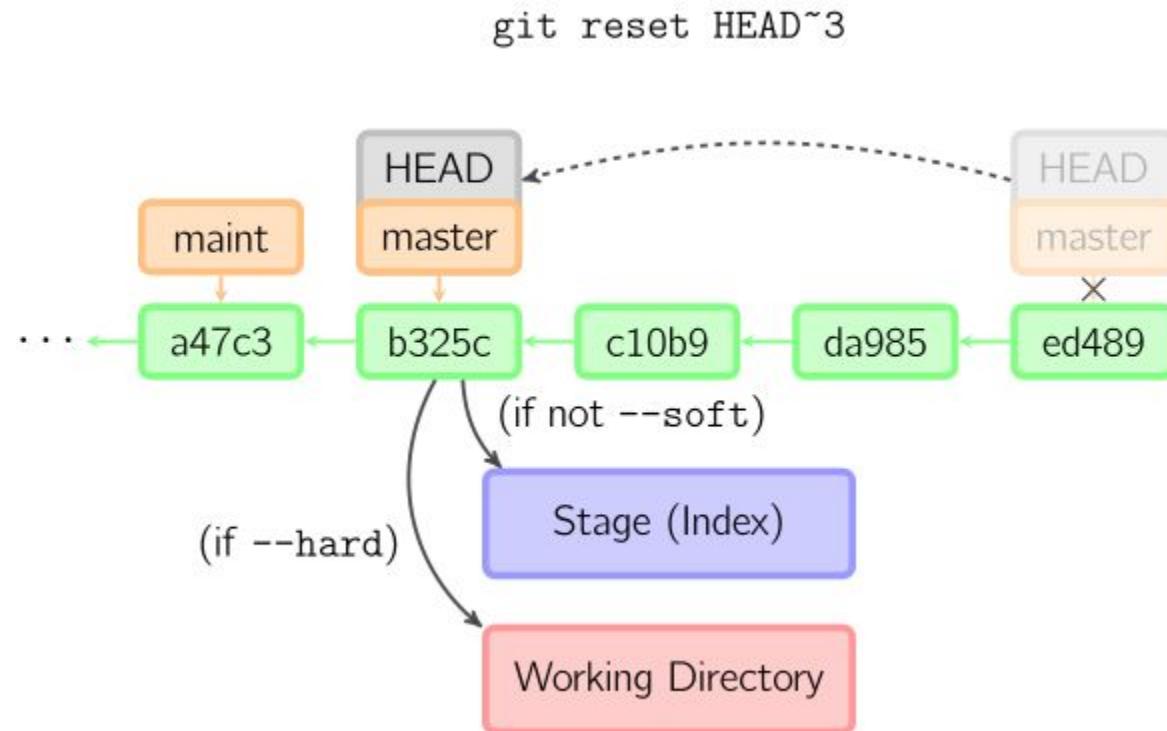
To switch branches or restore working tree files:



<https://git-scm.com/docs/git-checkout>

Workflow of Git/GitHub - Checkout & Reset

To reset current HEAD to the specified state:



<https://git-scm.com/docs/git-reset>