

The Best Versions of Ourselves

Christopher R. Genovese

Department of Statistics & Data Science

03 Sep 2024
Session #3

Plan

Discussion/Questions on TL1

Plan

Discussion/Questions on TL1

The Command Line

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Announcements

- Email subject: [750]
- Office Hours Wed 4:30-5:30, plus others.
- Documents (TL, slides, etc.) pushed to documents repo
- **Reading:**
 - Thursday:
 - Finish Thinking Languages, if not already
 - <https://clig.dev/>
 - Writing Commands on home page
 - Git from the Bottom Up by John Wiegley
 - <https://learngitbranching.js.org/>
- Try out the git-challenge exercise in the problem-bank repo
- **Homework:** swag due Thursday 5 Sep. Next HW coming soon.

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Why Use the Command Line?

Some disadvantages:

- It has a nontrivial learning curve.
- There are many commands and configuration options, many not particularly mnemonic.
- Many commands focus on processing text files.

But this approach has many compensating advantages:

- Operations may be easily combined to form new and useful operations.
- Very complex commands can be built from several simple pieces.
- Commands can be recorded for later recollection, reuse, modification, or automation.
- Commands can be parameterized for more general uses.
- Once mastered, working on the command line is **fast**.

Shells, Terminals, Files, Commands

- What is a shell?
- Terminals, ttys, and terminal emulators
- Files and directories

Principles

- ① Design for Composability
- ② Design for Humans
- ③ Make Consistent Interfaces
- ④ Effective feedback
- ⑤ Provide great help
- ⑥ Make your Interface Discoverable
- ⑦ Handle errors clearly
- ⑧ Decoration is wonderful (when appropriate)

Running Programs on the Command Line

Command structure

- Prompt
- Command name
- Flags and arguments
- Subcommands
- Pipes and redirection

Examples: `jq`, `git`, `curl`, `brew`, `ls`, ...

Input and Output

- Standard input, standard output, standard error
- Redirection
- Pipes and composability
- Text (line-based) and JSON

Program Interface Conventions

<https://clig.dev/> gives a nice overview, especially see Arguments section.

Always include:

- `--help`
- `--version`

Provide usage string with no arguments (if some required). `--help` can accept arguments for specialized documentation.

Where appropriate:

- `--quiet`
- `--verbose`
- `--dry-run`
- `--plain`
- `--no-input`
- `--no-color`
- `--json`

Other commonly used arguments:

- `--force`
- `--debug`
- `--output`

Program Interface Conventions

<https://clig.dev/> gives a nice overview, especially see Arguments section.

Other considerations:

- For file I/O, support '-' for stdin/stdout.
- For flags with optional arguments, have a special value for not supplied.
- Prompt for user input where appropriate but do not ever *require* a prompt.
- Confirm before executing a dangerous action (if possible).
- Check user input asap.

Handling Command-Line Arguments

- Multi-language: `docopt`
- Python: <https://docs.python.org/3/library/argparse.html> and `click`
- R: `argparse` and `optigrab`

Some very nice examples in other languages show some of the possibilities and are interesting to study.

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Some Benefits of Version Control

- Keep a complete record of changes (and when and by whom the changes were made). You can always revert back to a previous state and even continue forward development from a past state.
- Store a message with every change, so the rationale for changes is always recorded.
- Maintain multiple parallel branches of development, merging or discarding them as desired.
- View differences across time and identify the source of changes/problems across your history.
- Mark important snapshots of your code: "the version submitted to JASA" or "the version used for my conference presentation".
- Easily distribute your code or back it up, e.g., with a Git hosting service like GitHub or Bitbucket.

Git

Git is a popular, distributed version control system. It's underlying design is conceptually lean and elegant; it's practical complicated, like an overgrown garden (or a dumpster fire). High-Level Ideas:

- 1 Every object in git has a *unique identifier* (hash) that *derived from its content*.
- 2 A **commit** is an *immutable snapshot* of your project at specific time. Each commit is equipped with a pointer to its predecessor(s) (the state of HEAD at the time the commit is created).

The resulting chains of commits creates a “history” for the project.

- 3 A **branch** is a mutable *named pointer* to commit (and thus its linked history).

Branches are not special, except when a commit is made “on a branch,” the pointer is advanced to the new commit.

- 4 A **tag** is an *immutable* named pointer to a commit with its own description.

To use git effectively, it is surprisingly helpful to understand the underlying concepts.

Key Concept: The Three Trees

A git repository is a collection of commits and tags, along with a set of branches that point to the tips of the history along one or more several lines of development.

In the background, git manipulates four different types of objects: blobs, trees, commits, and tags. All the objects are stored under `.git/objects` in your project.

These objects are arranged into high-level “trees” that describe the state of your project at a given moment. There are three high-level trees that structure the git workflow.

- ① HEAD – the latest commit on the current branch
- ② Staging Area (Index) – a tree forming a *proposed next commit*, a set of changes from HEAD that are currently registered to be added to the next commit.
- ③ Working tree – the current (possibly modified) file tree for your project.

When you edit your code or add or remove files, you are changing the working tree. You then [stage](#) selected changes into the Index, building up a new commit. Eventually, you commit those staged changes, which creates a new commit with HEAD as its predecessor and advances HEAD to point to it.

Key Concept: The Three Trees (cont'd)

Here's a stylized view of HEAD in a fictitious repository.

[MyProject]		100000 MyProject	tree [100001, 100002, 100003, 100006, 100007, 100010, 100014]
README		100001 README	blob ad6b32ed752dac56d8dc7559308195ec
core.py		100002 core.py	blob e25e6b0e7c7bf92d4e861734436c2fd2
[data]		100003 data	tree [100004, 100005]
maps.gis		100004 map.gis	blob 239ccf3e9b7c400c09946af1581dd711
config.json		100005 config.json	blob d383ccd77254e01ea22a55313e94ad14
gis-util.py		100006 gis-util.py	blob abb6712b4ca6d1ba7a2ee6f4a1642c8e
[tests]		100007 tests	tree [100008, 100009]
test-gis.py		100008 test-gis.py	blob e488f57b8c424702898b776d195f2f1c
test-gfx.py		100009 test-gfx.py	blob 731ea463f7b3c9632b32ce57d9805375
[doc]		100010 doc	tree [100011]
[generated]		100011 generated	tree [100012, 100013]
OAF323DB1		100012 OAF323DB1	blob 726232cd6702e61932fbf3ea44a03c6f
BFEEC9978		100013 BFEEC9978	blob d792cf03a1af84fd546054caee49613c
doc.md		100014 doc.md	blob cee77a23ad50c9e419bf77e835f36aed
...		100015

On the left is high-level file-system view of the tree; on the right a more packed encoding like what git uses. (Ids in trees are shorthand for hashes.)

Key Concept: The Three Trees (cont'd)

This gives an associative map of hashes to contents. We do not need file names for this to work; names become *metadata*.

This tree also has a hash, so we end up with a snapshot, a single commit:

```
100000  
  HEAD, main
```

HEAD and main are named pointers to this commit – branches.

Key Concept: The Three Trees (cont'd)

Now we make a change to one file README. What happens?

100000	MyProject	tree	[100001, 100002, 100003, 100006, 100007, 100010, 100014]
100001	README	blob	ad6b32ed752dac56d8dc7559308195ec
100002	core.py	blob	e25e6b0e7c7bf92d4e861734436c2fd2
100003	data	tree	[100004, 100005]
100004	map.gis	blob	239ccf3e9b7c400c09946af1581dd711
100005	config.json	blob	d383ccd77254e01ea22a55313e94ad14
100006	gis-util.py	blob	abb6712b4ca6d1ba7a2ee6f4a1642c8e
100007	tests	tree	[100008, 100009]
100008	test-gis.py	blob	e488f57b8c424702898b776d195f2f1c
100009	test-gfx.py	blob	731ea463f7b3c9632b32ce57d9805375
100010	doc	tree	[100011]
100011	generated	tree	[100012, 100013]
100012	OAF323DB1	blob	726232cd6702e61932fbf3ea44a03c6f
100013	BFEEC9978	blob	d792cf03a1af84fd546054caee49613c
100014	doc.md	blob	cee77a23ad50c9e419bf77e835f36aed
100015
101033	README'	blob	a1d09363bc48b91ecf2b2bae8a0a533b
101034	MyProject'	tree	[100133, 100002, 100003, 100006, 100007, 100010, 100014]

Key Concept: The Three Trees (cont'd)

Now we make a change to one file README. What happens?

Now keys 101034 and 100000 point to two different versions in our history.
And do we have two snapshots in a sequence: two commits in a history.

```
100000          <--- 101034  
                      HEAD, main
```

Branches HEAD and main have automatically advanced to the tip of the history.

Key Concept: The Three Trees (cont'd)

We make another change, regenerating the files in doc.

```
00242 0AF323DB1' blob 1dcab182eb803c7f3e4459c69f20c5f0
00243 BFEEC9978' blob abb0acf24574303090c3455c71bbf94c
00244 generated tree [100242, 100243]
00250 doc' tree [100244]
02388 MyProject'' tree [100133, 100002, 100003, 100006,
                        100007, 100250, 100014]
```

This yields a chain of three commits that gives a history

```
100000 <--- 101034 <--- 102388
                        HEAD, main
```

Again branches HEAD and main have advanced to stay at the tip of the history.
(We can explicitly move them as well.)

Key Concept: The Three Trees (cont'd)

Thinking about it, we realize that we want to try something out starting at commit 101034. So we "move to that snapshot" and make a change: adding a new file packages inside the MyProject directory.

```
02521 packages      blob      e26fe3c3348c976e937b81423b561ebe
02388 MyProject''' tree      [100133, 100002, 100003, 100006, 100007,
                             100010, 100014, 102521]
```

We call this branch `experimental`, and our history now looks like

```
                experimental
            --- 102388
            |
            v
100000 <--- 101034 <--- 102388
                        HEAD, main
```

Key Concept: The Three Trees (cont'd)

Staging: Working Tree → Index

The **working tree** is the version of the files that we see in our editor. We can make what changes we like when we are working without changing the other trees.

When we have made changes that we think we want to save, we *first* add them to the **Index**. This is a tree that starts as a copy of the latest commit but is amended by changes **staged** from the working tree.

Key Concept: The Three Trees (cont'd)

Committing: Index → HEAD

When we have staged all the changes we like, we are ready to **commit**. We enter metadata like a *commit message* and the Index tree becomes the newest commit. It is added to the history with a pointer to the previous HEAD, and HEAD is moved to point to it.

Key Concept: The Three Trees (cont'd)

Checkout: HEAD → Working Tree

When we “move” in the history, we move the HEAD pointer to another commit and **checkout** that commit. The commit pointed to in the history is copied and becomes the files we see in the new working tree.

Key Concept: The Three Trees (cont'd)

That's git's main job. All the rest is details!

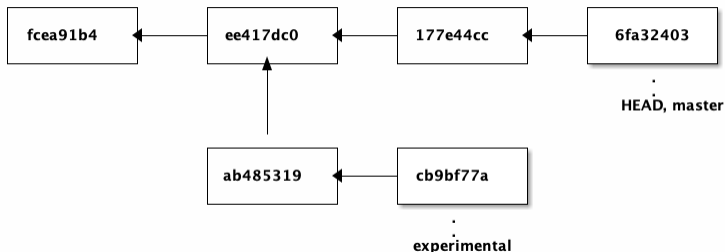
Branching and Merging

Creating a branch (like experimental above), creates a new named pointer that points to some specified commit. That commit is by definition considered the “tip” of a history.

Branching and Merging

Creating a branch (like experimental above), creates a new named pointer that points to some specified commit. That commit is by definition considered the “tip” of a history.

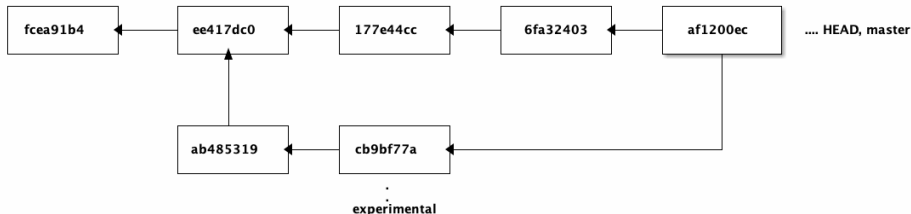
Switching to a branch moves HEAD to that branch, making it current. Commits at that point will advance both HEAD and the current branch.



Branching and Merging

Creating a branch (like experimental above), creates a new named pointer that points to some specified commit. That commit is by definition considered the “tip” of a history.

If we want to reconcile two branches of development, we **merge** them



We may need to address *conflicts* between the two trees!

Git Commands

Using git happens at several levels: porcelain, plumbing, and hosting services. Most common uses are at a high level, but the plumbing can be useful.

Common commands:

<code>git status</code>	Get current state of repo and three trees
<code>git log</code>	Look at repository history
<code>git diff</code>	See what has changed between commits
<code>git init</code>	Initialize a new repository
<code>git add</code>	Stage one or more files
<code>git branch</code>	Create and manage branches
<code>git switch</code>	Switch branches
<code>git checkout</code>	Checkout a commit
<code>git commit</code>	Commit the current index
<code>git merge</code>	Combine changes from multiple commits
<code>git remote</code>	Associate repo with a remote repo
<code>git push</code>	Send new commits to a remote copy of this repo
<code>git fetch/pull</code>	Receive new commits from a remote repo
<code>git worktree</code>	Work on multiple branches at the same time

Quick [demo](#)

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Homework Workflow

- 1 Use `new-homework` script to create and populate a branch for each assignment. Start in `main` branch when doing so.
- 2 Suggestion: use `git worktree` to work on an assignment branch. This prevents branch mixups and makes it easy to work on several assignments simultaneously.
- 3 Put an `exercises` directory in `~/s750` and outside your assignments directory. Use that as the target of your worktrees.
- 4 When you are done with an assignment/commit, push to your github repo.
- 5 When ready to (re)submit, create a Pull Request on Github, comparing the exercise branch to `clean-start`.

Questions?

Good Git Practices

- Commit often

Committing after individual, complete changes makes it much easier to track your progress – and move backward when needed.

Even provisional commits can be reasonable if so labeled.

- Write good commit messages

It should explain what you changed and *why* you changed it. The first line should be a short summary, easy to interpret from the log.

Label breaking changes, keep a log of bugs/issues and indicate the fix or change by id.

- Hooks can take regular actions at each commit, e.g., run tests.
- Use branches for trying things out, especially when collaborating.
- `git blame` is a great tool for finding changes.

Plan

Discussion/Questions on TL1

The Command Line

Version Control and Git

Aside: Homework Workflow

Activity: Migrit, Part 1/n

Migit

We are going to write a small git clone that handles many of the most common tasks. This will help us understand how git works and give us a practical case study for thinking about structuring, writing, and testing good code.

Design discussion, structure summary, snapshot, coding.

Migit

We are going to write a small git clone that handles many of the most common tasks. This will help us understand how git works and give us a practical case study for thinking about structuring, writing, and testing good code.

First steps:

- ① Separate executable and library
- ② Driver routine for the program

Design discussion, structure summary, snapshot, coding.

Migit

We are going to write a small git clone that handles many of the most common tasks. This will help us understand how git works and give us a practical case study for thinking about structuring, writing, and testing good code.

First steps:

- ❶ Separate executable and library
- ❷ Driver routine for the program

Next steps:

- ❶ The `init` command
- ❷ Create an abstraction for the Repository (at least in somewhat simplified form)

Design discussion, structure summary, snapshot, coding.

THE END