Boise Data Science

organizer: ***Randall Shane, PhD***
month year: ***January 2017***
topic: **GitHub for DS applications**

…DISCLAIMERS…

(1) Code written in this presentation is syntactically expressed using either R or Python. Both are open source programming languages with statistical libraries available.

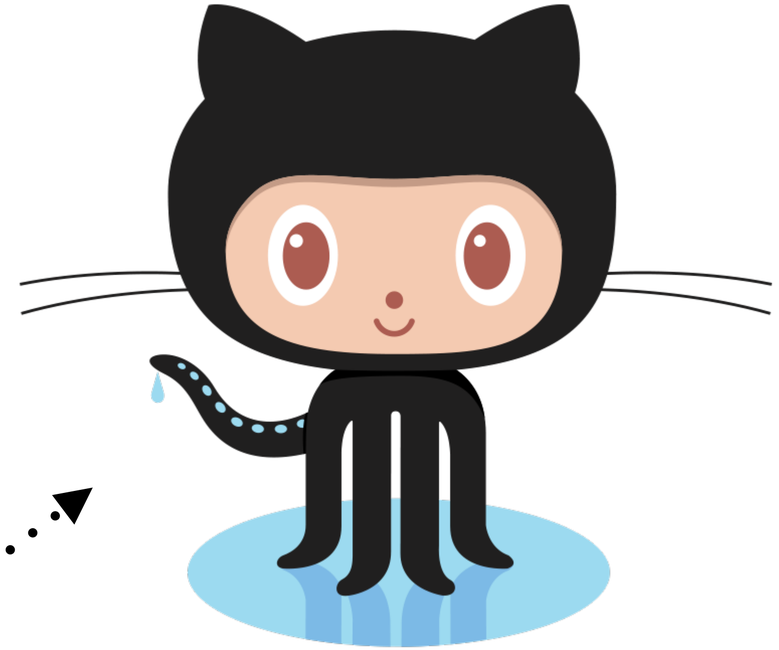
(2) Open source products including data storage technologies, databases, languages and operating systems are used exclusively for this group.

(3) Images and some text has been borrowed from the inter webs. Apologies if I did not credit. Thanks for contributing to making the world a smarter place!!

Boise Data Science Meetup

Using Github for Data Science Applications

- What is GitHub?
- How does it work?
- Why would I use this?
- What commands do I need to know?
- Demonstrate a common workflow



What is that?
(it's an octocat)

Boise Data Science Meetup

What is GitHub?

GitHub is a web-based Git repository hosting service. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features.

GitHub - Wikipedia
https://en.wikipedia.org/wiki/**GitHub**

GitHub provides source control for code bases. There are options for public and private repositories as well as multi-member participation.

Boise Data Science Meetup

Caveats…

-There are other commands and ways to use git.
-This is how I do it.

Git References:

https://help.github.com
https://git-scm.com/book/en/v2
https://git-scm.com/docs/gittutorial

Boise Data Science Meetup

How does GitHub work?

Setup:

Step 1: Sign Up for an account
Accounts are free, paid accounts possess private repositories

Step 2: Install Resources
https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

Step 3: Setup Access
Common method includes ssh keys but token based auth is available

Boise Data Science Meetup

How does GitHub work?

Use:

Step 1: Create a Repository or be granted access to an existing one
There are public and private repositories.

Step 2: Repositories have branches
'master' *is the default main branch for repos.*
Generally, you don't work in master.
Work is 'committed' to branches.
Branches are merged via 'pull request'.

Boise Data Science Meetup

Why would I use GitHub?

Main Reasons:

- Source control - code old and new, history and ‘blame’
- Multiple people contributing to single code base

http://www.howtogeek.com/180167/htg-explains-what-is-github-and-what-do-geeks-use-it-for/

Boise Data Science Meetup

…& now, the meat of our topic!…

Boise Data Science Meetup

What commands do I need to know?

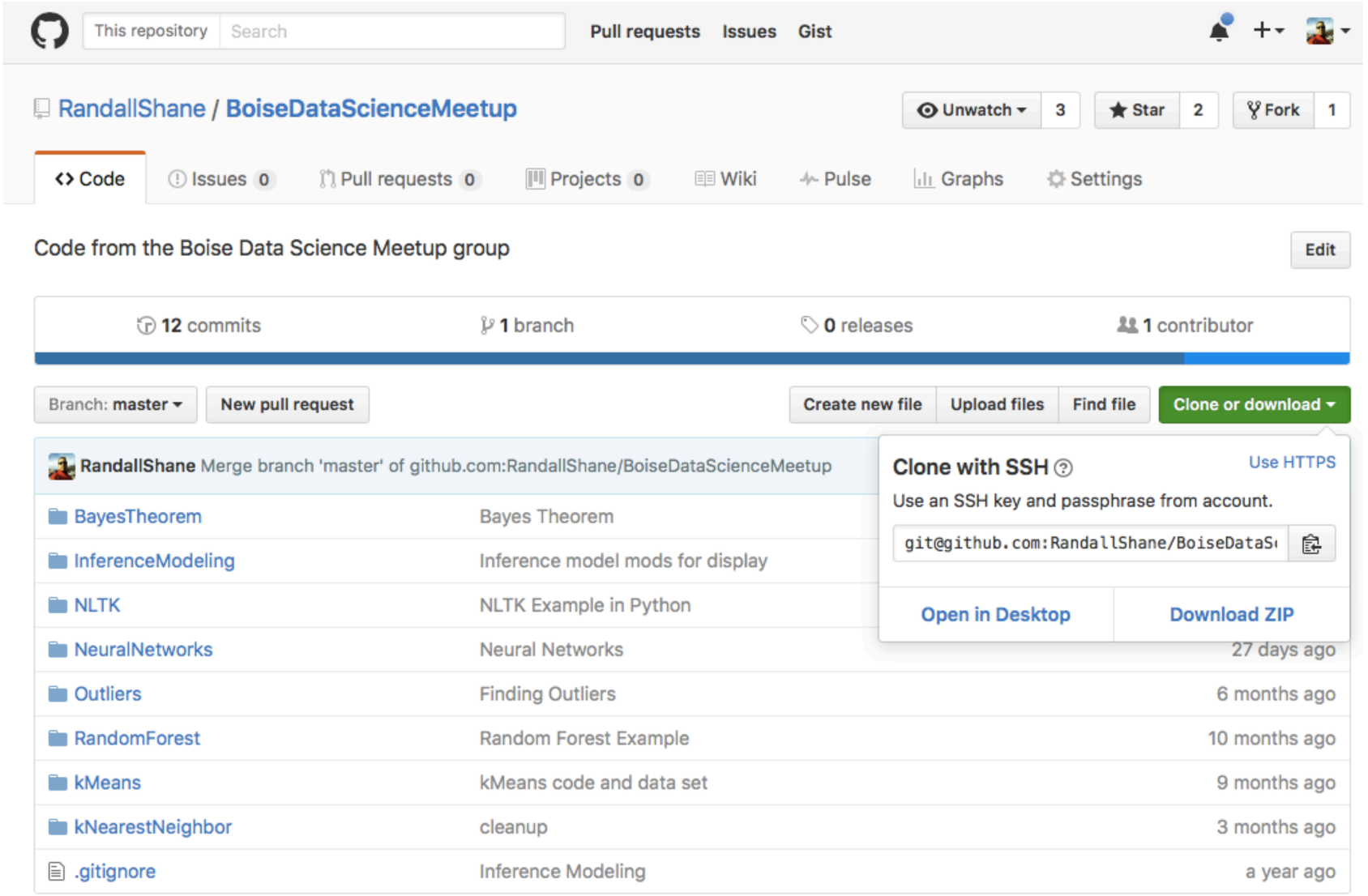
Setup:

1) local git config
   http://git-scm.com/docs/gittutorial
      $ git config --global user.name "Your Name Comes Here"
      $ git config --global user.email you@yourdomain.example.com

2) Initialize local directory
   git init

3) Clone repository local
   git clone git@github.com*:RandallShane/BoiseDataScienceMeetup.git*

Boise Data Science Meetup



NOTE

Boise Data Science Meetup

Demonstrate a common workflow - COMMANDS

1) Branches
   create branch: git checkout -b <branch>
   list branch: git <branch>
   checkout branch: git checkout <branch>
   checkout remote branch: git fetch, then git checkout <branch>
   delete local branch: git branch -d <branch>
   delete remote branch: git branch -D <branch>

2) Code Commits:
   commit only: git commit -a
   commit w/message: git commit -am 'Commit message'
   commit new branch git push -u origin <branch>
   merge branch and master: git merge <branch>

3) Repository:
   push branch to repo: git push
   pull repo local: git pull
   rebase branch to master git rebase origin/master

Boise Data Science Meetup

Demonstrate a common workflow

-Know your status: git status
git branch

-Start with master branch on repo
--make local copy: git pull
--create branch to work from: git branch -b <branch>
--switch to branch to work: git checkout <branch>
--if master changes, rebase branch to master: git rebase origin/master

-Commit work back to repo
--add any new files: git add .
--commit with message: git commit -am '<message>'
--first branch push: git push -u origin <branch>
--subsequent push(es): git push

Boise Data Science Meetup

Demonstrate a common workflow

Pull Requests (merging):

-Code Review
--On github, go to branch and make a pull request
--Pull requests are peer reviewed and then merged
--Once merged, the branch can be deleted from the repo

Boise Data Science Meetup

Boise Data Science Meetup

```
Mashed:BoiseDataScienceMeetup randallshane$ git pull
Already up-to-date.
Mashed:BoiseDataScienceMeetup randallshane$ git checkout -b fix
Switched to a new branch 'fix'
Mashed:BoiseDataScienceMeetup randallshane$ echo 'make changes in code here'
make changes in code here
Mashed:BoiseDataScienceMeetup randallshane$ git status
On branch fix
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:   RandomForest/randomforest.py

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

        Ratings/

no changes added to commit (use "git add" and/or "git commit -a")
Mashed:BoiseDataScienceMeetup randallshane$ git add .
```

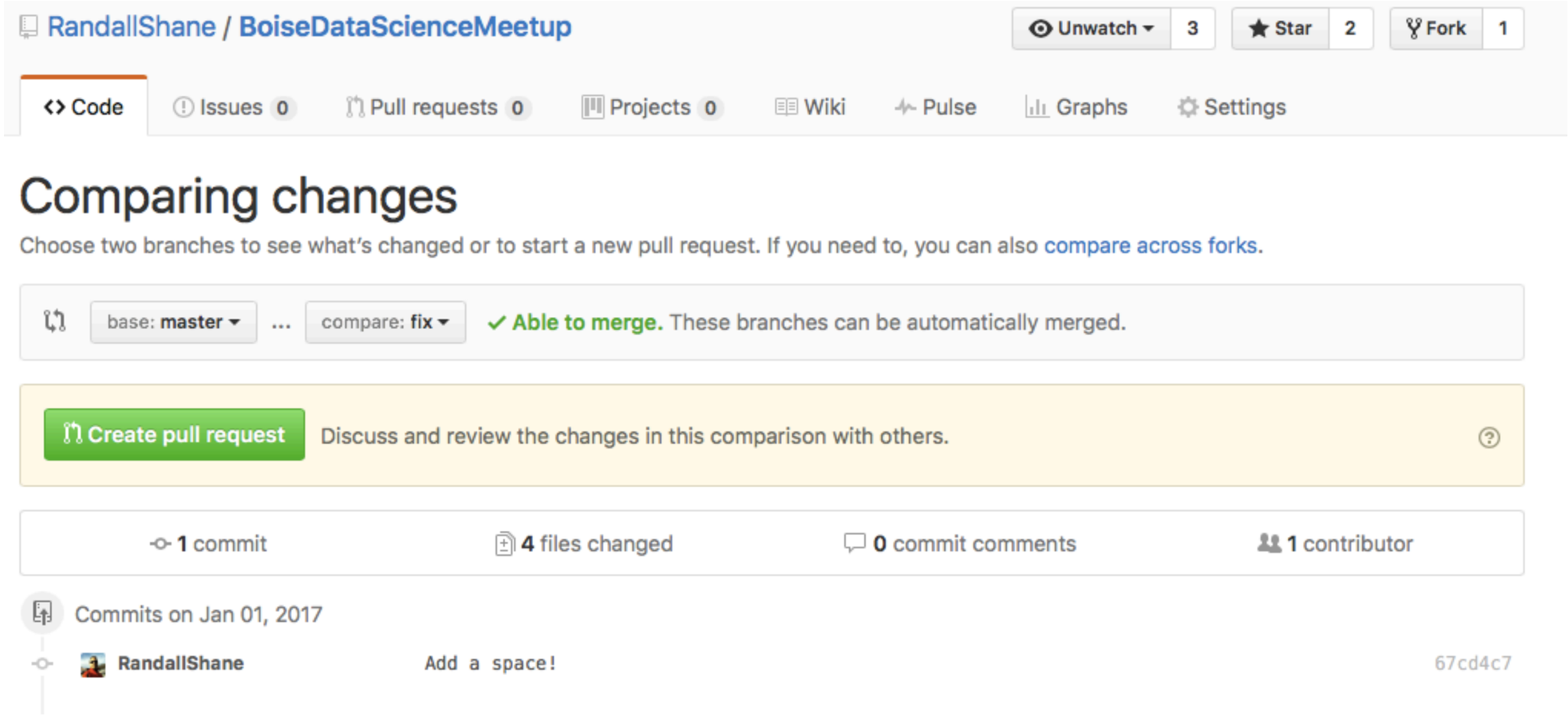
Boise Data Science Meetup

```
Mashed:BoiseDataScienceMeetup randallshane$ git commit -am 'Add a space!'
[fix 67cd4c7] Add a space!
 4 files changed, 1428653 insertions(+)
 create mode 100644 Ratings/BX-Book-Ratings.csv
 create mode 100644 Ratings/BX-Users.csv
 create mode 100644 Ratings/ratings.py
Mashed:BoiseDataScienceMeetup randallshane$ git push
fatal: The current branch fix has no upstream branch.
To push the current branch and set the remote as upstream, use

    git push --set-upstream origin fix

Mashed:BoiseDataScienceMeetup randallshane$ git push --set-upstream origin fix
Counting objects: 8, done.
Delta compression using up to 4 threads.
Compressing objects: 100% (8/8), done.
Writing objects: 100% (8/8), 9.49 MiB | 638.00 KiB/s, done.
Total 8 (delta 2), reused 0 (delta 0)
remote: Resolving deltas: 100% (2/2), completed with 2 local objects.
To github.com:RandallShane/BoiseDataScienceMeetup.git
 * [new branch]      fix -> fix
Branch fix set up to track remote branch fix from origin.
Mashed:BoiseDataScienceMeetup randallshane$
```

Boise Data Science Meetup



Boise Data Science Meetup

Add more commits by pushing to the **fix** branch on **RandallShane/BoiseDataScienceMeetup**.

This branch has no conflicts with the base branch
Merging can be performed automatically.

Merge pull request | You can also open this in GitHub Desktop or view command line instructions.

Write | Preview

Leave a comment

Attach files by dragging & dropping or selecting them.

Styling with Markdown is supported

Close pull request | Comment

ProTip! Add .patch or .diff to the end of URLs for Git's plaintext views.

Labels
None yet

Projects
None yet

Milestone
No milestone

Notifications
Unsubscribe
You're receiving notifications because you authored the thread.

1 participant

Lock conversation

Pull request successfully merged and closed
You're all set—the `fix` branch can be safely deleted.

Delete branch

Boise Data Science Meetup

```
Mashed:BoiseDataScienceMeetup randallshane$ git branch
* fix
  master
Mashed:BoiseDataScienceMeetup randallshane$ git checkout master
Switched to branch 'master'
Your branch is up-to-date with 'origin/master'.
Mashed:BoiseDataScienceMeetup randallshane$ git pull
remote: Counting objects: 1, done.
remote: Total 1 (delta 0), reused 1 (delta 0), pack-reused 0
Unpacking objects: 100% (1/1), done.
From github.com:RandallShane/BoiseDataScienceMeetup
   3365968..83b496d  master     -> origin/master
Updating 3365968..83b496d
Fast-forward
 RandomForest/randomforest.py |       1 +
 Ratings/BX-Book-Ratings.csv  | 1149781 ++++++++++++++++++++++++++++++++++++++++++++
 Ratings/BX-Users.csv         |  278860 +++++++++++
 Ratings/ratings.py           |      11 +
 4 files changed, 1428653 insertions(+)
 create mode 100644 Ratings/BX-Book-Ratings.csv
 create mode 100644 Ratings/BX-Users.csv
 create mode 100644 Ratings/ratings.py
Mashed:BoiseDataScienceMeetup randallshane$
```

Boise Data Science Meetup

https://github.com/RandalIShane/BoiseDataScienceMeetup

https://twitter.com/RandalIShanePhD