

# **DATA SCIENCE**

**CSC 405/605**



THE UNIVERSITY *of* NORTH CAROLINA  
**GREENSBORO**

# VERSION CONTROL



# GIT

## What is Git?

- Distributed Version / Revision Control System
- Large projects – Multiple members/coders
- Software development
- Decentralized, Fast, Flexible

## Git Servers

- GitHub (<https://education.github.com/pack>)
  - *We are going to be using GitHub – Get an account*
- BitBucket (<https://www.atlassian.com/software/views/bitbucket-academic-license.jsp>)
- ....

## Git Tools

- Command line git (Learn this)
- Tower - <https://www.git-tower.com/mac>
- Github (Windows, Mac) - <https://desktop.github.com/>
- Gitbox, SourceTree



# GIT SETUP

## Windows

- <https://gitforwindows.org/>

## Mac

- <http://git-scm.com/download/mac>

## Linux

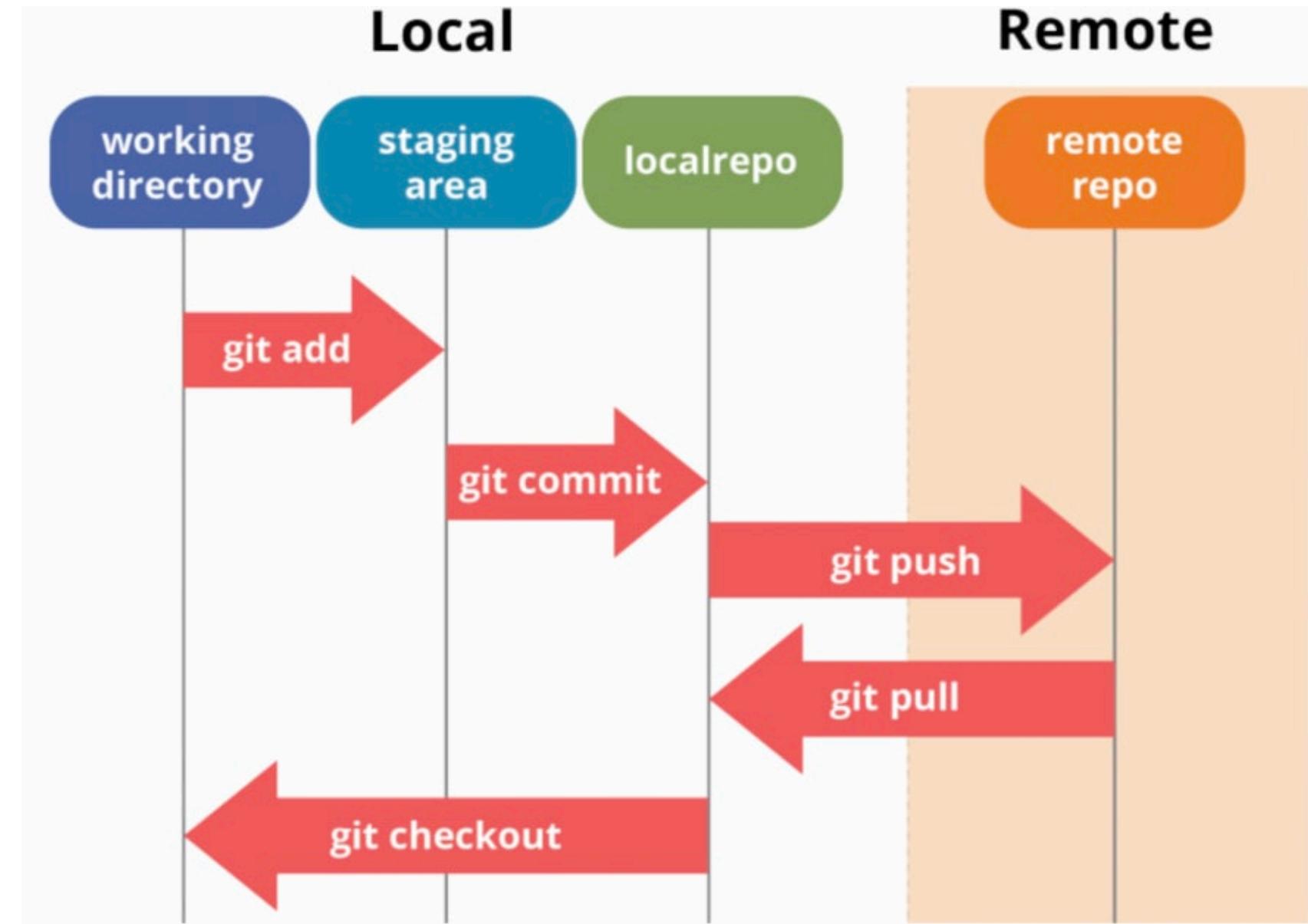
- <http://git-scm.com/book/en/v2/Getting-Started-Installing-Git>

## Guide

- <http://rogerdudler.github.io/git-guide/>
- [http://slides.com/moji3000/git\\_presentation/fullscreen#/](http://slides.com/moji3000/git_presentation/fullscreen#/)



# GIT SETUP



# GIT SETUP

## Git Config

- Add your user name and email

```
Somyas-MBP:test_proj Som$ git config --global user.name "Somya Mohanty"  
Somyas-MBP:test_proj Som$ git config --global user.email "mohanty.somya@gmail.com"
```

```
Somyas-MBP:test_proj Som$ git config --list  
user.name=Somya Mohanty  
user.email=mohanty.somya@gmail.com  
filter.media.clean=git-media-clean %f  
filter.media.smudge=git-media-smudge %f  
filter.lfs.clean=git lfs clean %f  
filter.lfs.smudge=git lfs smudge %f  
filter.lfs.required=true  
core.repositoryformatversion=0  
core.filemode=true  
core.bare=false  
core.logallrefupdates=true  
core.ignorecase=true  
core.precomposeunicode=true
```



# GIT

## Git Init

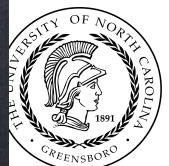
- Create a new repository
  - Create a new directory
  - Open it
  - Execute git init

```
Somyas-MBP:test_proj Som$ git init
Initialized empty Git repository in /Users/Som/test_proj/.git/
```

## Git Clone

- Create a working copy from remote server

```
Somyas-MBP:test_proj Som$ git clone https://github.com/idl/Twitter_sentiment
Cloning into 'Twitter_sentiment'...
Username for 'https://github.com': somyamohanty
Password for 'https://somyamohanty@github.com':
remote: Counting objects: 73, done.
remote: Total 73 (delta 0), reused 0 (delta 0), pack-reused 73
Unpacking objects: 100% (73/73), done.
Checking connectivity... done.
Somyas-MBP:test_proj Som$ ls
Twitter_sentiment
```



# GIT

## Git Status

- Status of the project

```
Somyas-MBP:test_proj Som$ git status
On branch master

Initial commit

nothing to commit (create/copy files and use "git add" to track)
```



# GIT

## Git Status

- Uncommitted

```
Somyas-MBP:test_proj Som$ git status
On branch master
Initial commit

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

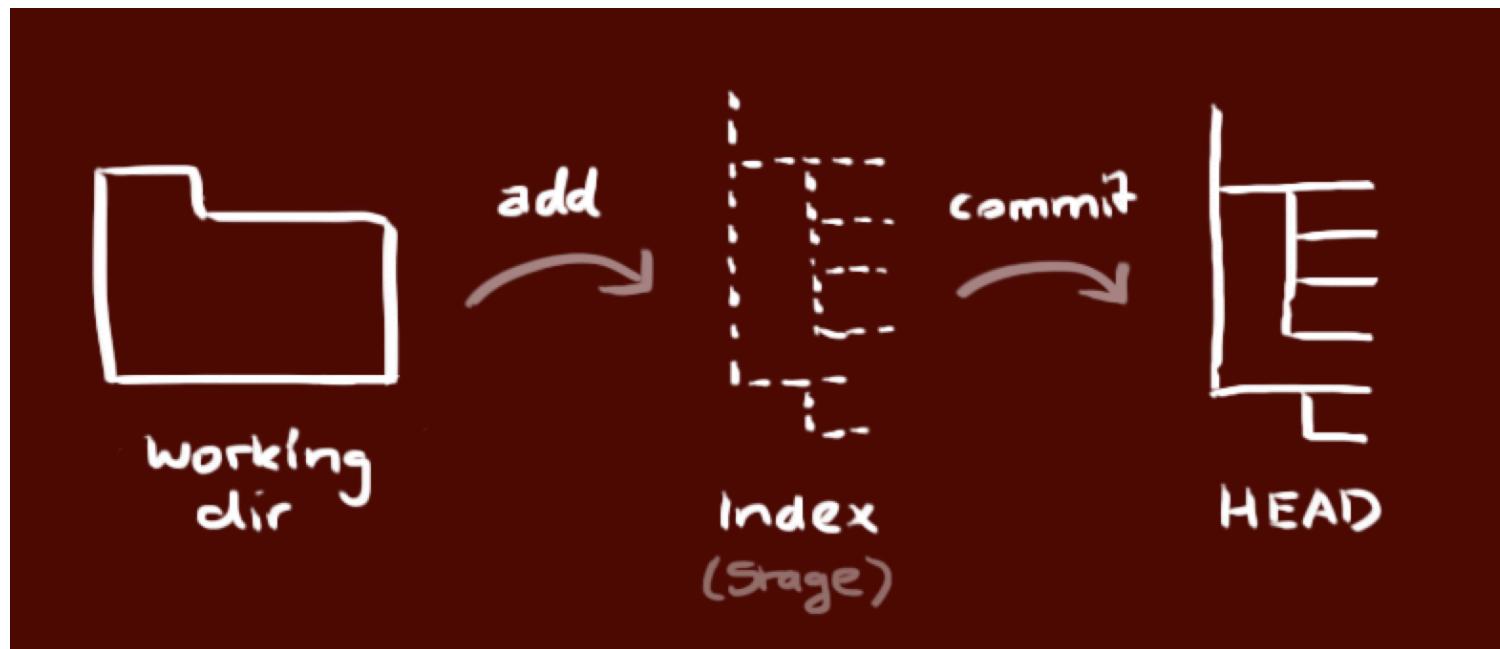
    Twitter_sentiment/
nothing added to commit but untracked files present (use "git add" to track)
```



# GIT

## Git Workflow

- Working Directory - Actual Files
- Index – Staging Area
- Head – Commit Pointer



# GIT

## Git Add

- Change a file

```
Somyas-MBP:Twitter_sentiment Som$ git add README.md
Somyas-MBP:Twitter_sentiment Som$ git status
On branch master
Your branch is up-to-date with 'origin/master'.
Changes to be committed:
  (use "git reset HEAD <file>..." to unstage)

    modified:   README.md
```

## Git Commit

- Commit the changes made to a file

```
Somyas-MBP:Twitter_sentiment Som$ git commit -m 'testing commit'
[master 68d54c6] testing commit
 1 file changed, 1 insertion(+)
```



# GIT

## Git Push

- Push head to include the new commit

```
Somyas-MBP:Twitter_sentiment Som$ git push
warning: push.default is unset; its implicit value has changed in
Git 2.0 from 'matching' to 'simple'. To squelch this message
and maintain the traditional behavior, use:

  git config --global push.default matching

To squelch this message and adopt the new behavior now, use:

  git config --global push.default simple

When push.default is set to 'matching', git will push local branches
to the remote branches that already exist with the same name.

Since Git 2.0, Git defaults to the more conservative 'simple'
behavior, which only pushes the current branch to the corresponding
remote branch that 'git pull' uses to update the current branch.

See 'git help config' and search for 'push.default' for further information.
(the 'simple' mode was introduced in Git 1.7.11. Use the similar mode
'current' instead of 'simple' if you sometimes use older versions of Git)

Counting objects: 3, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 292 bytes | 0 bytes/s, done.
Total 3 (delta 2), reused 0 (delta 0)
To https://github.com/idl/Twitter_sentiment
  49a31cb..68d54c6  master -> master
```



# GIT

## Git Show

```
Somyas-MBP:Twitter_sentiment Som$ git show  
commit 68d54c60620befb11235ded6d462d6c09cbfde2b  
Author: Somya Mohanty <mohanty.somya@gmail.com>  
Date: Fri Jul 17 12:42:57 2015 -0500
```

testing commit

```
diff --git a/README.md b/README.md  
index 028a6d0..8239c85 100755  
--- a/README.md  
+++ b/README.md  
@@ -3,4 +3,5 @@ Twitter Sentiment
```

Sentiment Analysis of Twitter Messages

+Test

```
Somyas-MBP:Twitter_sentiment Som$ git status  
On branch master  
Your branch is up-to-date with 'origin/master'.  
nothing to commit, working directory clean
```



# GIT

Commits on Jul 17, 2015



## testing commit

somyamohanty authored 44 seconds ago



68d54c6



## testing commit

master



somyamohanty authored a minute ago

1 parent 49a31cb commit 68d54c60620befb11235ded6d462d6c09cbfde2b

Browse files



Showing 1 changed file with 1 addition and 0 deletions.

Unified

Split

1 1 README.md



View



@@ -3,4 +3,5 @@ Twitter Sentiment

3 3

4 4

Sentiment Analysis of Twitter Messages

5 5

+Test

6 7



# GIT

## Git Pull

- Get new updates

```
Somyas-MBP:Twitter_sentiment Som$ git pull
remote: Counting objects: 3, done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (3/3), done.
From https://github.com/ndl/Twitter_sentiment
  68d54c6..97ef6a1 master      -> origin/master
Updating 68d54c6..97ef6a1
Fast-forward
 README.md | 1 +
  1 file changed, 1 insertion(+)
Somyas-MBP:Twitter_sentiment Som$ git show
commit 97ef6a176552a818eb5b9437f523f02d350c5e6f
Author: Somya Mohanty <mohanty.somya@gmail.com>
Date:   Fri Jul 17 13:16:12 2015 -0500
```

Testing again

```
diff --git a/README.md b/README.md
index 8239c85..c6ac490 100755
--- a/README.md
+++ b/README.md
@@ -5,3 +5,4 @@ Sentiment Analysis of Twitter Messages
```

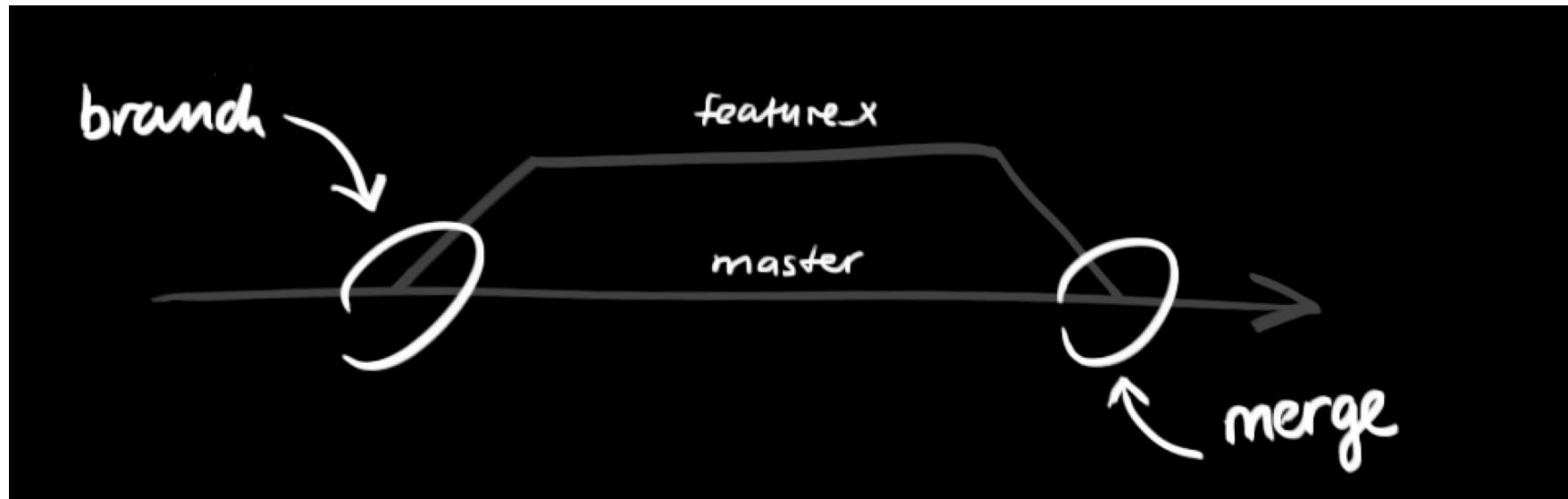
Test

+Test agin



# GIT

## Git Branch



# GIT

## Git Branch

```
Somyas-MBP:Twitter_sentiment Som$ git checkout -b 'feature_x'  
Switched to a new branch 'feature_x'
```

```
Somyas-MBP:Twitter_sentiment Som$ git branch  
* feature_x  
  master
```

## Git push branch

```
Somyas-MBP:Twitter_sentiment Som$ git push origin 'feature_x'  
Total 0 (delta 0), reused 0 (delta 0)  
To https://github.com/idl/Twitter_sentiment  
 * [new branch]      feature_x -> feature_x
```



# GIT

## Git Checkout

```
Somyas-MBP:Twitter_sentiment Som$ git checkout master
Switched to branch 'master'
Your branch is up-to-date with 'origin/master'.
```

## Git merge branch

```
Somyas-MBP:Twitter_sentiment Som$ git merge feature_x
Updating 97ef6a1..77bbeea
Fast-forward
 README.md | 2 ++
 1 file changed, 2 insertions(+)
Somyas-MBP:Twitter_sentiment Som$ git show
commit 77bbeea597407a2f9e3093ce0cd02b7695aaea8b
Author: Somya Mohanty <mohanty.somya@gmail.com>
Date:   Fri Jul 17 13:27:10 2015 -0500

        testing merging

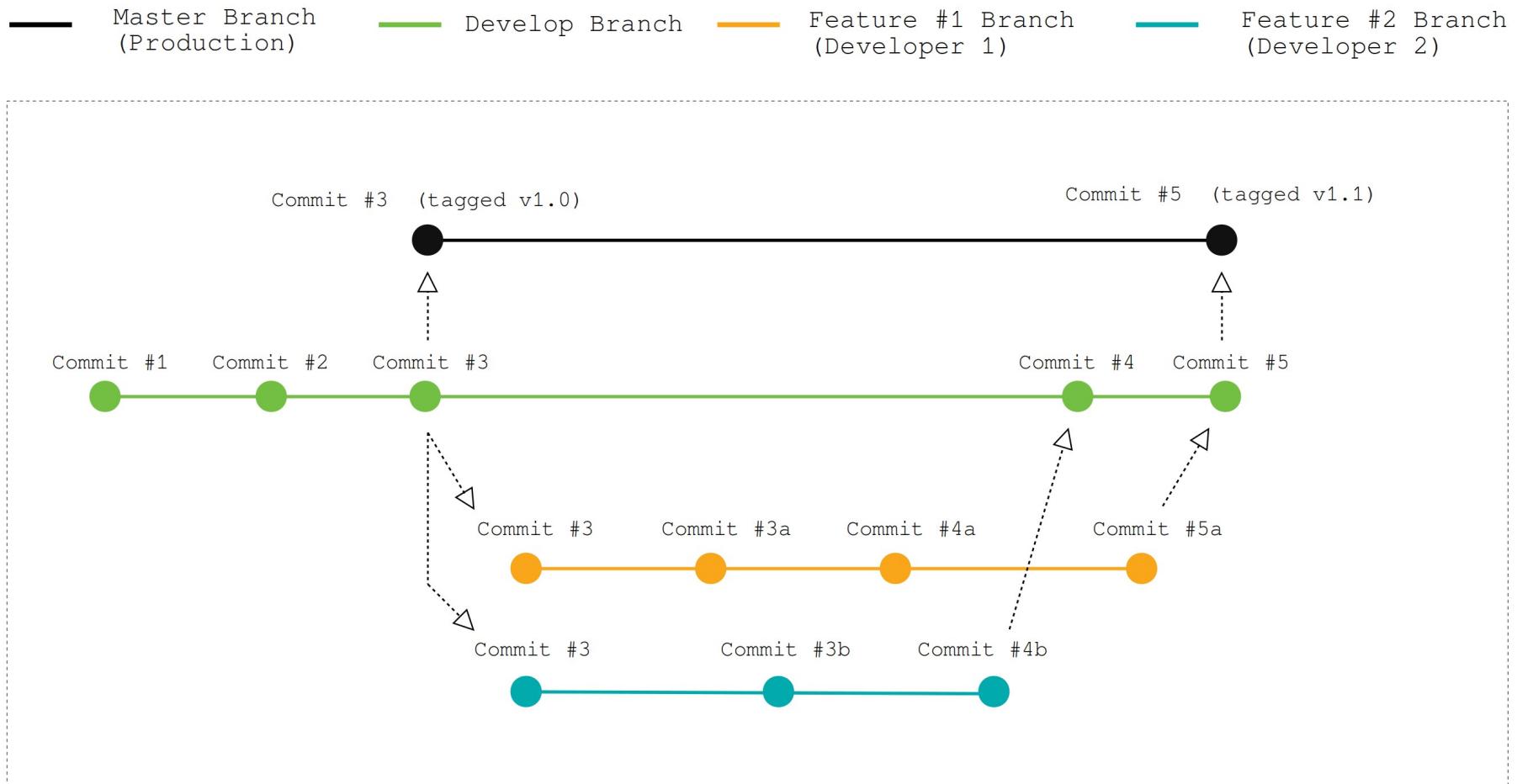
diff --git a/README.md b/README.md
index c6ac490..9f81ac5 100755
--- a/README.md
+++ b/README.md
@@ -6,3 +6,5 @@ Sentiment Analysis of Twitter Messages
 Test

 Test agin
+
+Testing merging branch
```



# GIT

## A Simple, Effective Git Workflow



# GIT

## Git directory structure for class project

- **./src** – Source code / IPython Notebooks
  - **./util** – Library files
- **./data** – Sample Data (Public Data Only)
- **./doc** - Project Documentation
- **./Readme.MD** – Project Introduction
- **./requirement.txt** – Libraries/Tools/Module use



# PACKAGE MANAGEMENT



# PIP PYTHON

## What is Pip?

- **Package manager to install python libraries**
  - pip install <some-package-name>
- **How to install pip?**
  - Linux (ubuntu)
    - sudo apt-get install python-pip
  - Mac
    - sudo easy\_install pip
  - Windows
    - <https://github.com/BurntSushi/nfldb/wiki/Python-&-pip-Windows-installation>
    - Download get-pip.py from <https://bootstrap.pypa.io/get-pip.py>
    - Pip install
      - python get-pip.py --no-index --find-links=/local/copies
      - python get-pip.py -user
- **Upgrade pip**
  - pip install -U pip



# VIRTUAL ENVIRONMENT

## What is Virtual Environment?

- **Dependencies relevant to project**
  - Local to projects
  - Easy to do pip-freeze and get the requirements
- **Solves “Project X depends on version 1.x but, Project Y needs 4.x”**
- **Helps in re-producible projects**
- **Steps**
  - Install virtualenv and virtualenvwrapper (makes it much easier to work with virtualenv)
    - Linux or Mac
      - `pip install virtualenv`
      - `pip install virtualenvwrapper`
    - Windows
      - `pip install virtualenvwrapper-win`



# VIRTUAL ENVIRONMENT

- **Steps**
  - Set your home environment
    - `export WORKON_HOME=~/Envs`
    - `source /usr/local/bin/virtualenvwrapper.sh`
  - Create a project directory
    - `mkdir assignment_1`
    - `cd assignment_1`
  - Create virtual environments
    - `mkvirtualenv assignment_1`
    - `workon assignment_1`
  - List all environments
    - `lsvirtualenv`
  - List all packages
    - `lssitepackages`
- <http://docs.python-guide.org/en/latest/dev/virtualenvs/>



# VIRTUAL ENVIRONMENT

- **Get the requirements of project**
  - Inside the project directory
    - pip freeze
    - pip freeze > requirements.txt



# LOCAL SETUP

- **Install build dependencies**
  - Ubuntu
    - sudo apt-get install python-numpy python-scipy python-matplotlib ipython ipython-notebook python-pandas python-sympy python-nose
    - sudo apt-get install libfreetype6-dev libpng-dev
    - pip install pandas, numpy, scipy, matplotlib
  - Mac
    - Use pip to install pandas, numpy, scipy, matplotlib
- **Install gcc**
  - Ubuntu
    - sudo apt-get install gcc
  - Mac
    - brew install gcc
- **Install g++**
  - Ubuntu
    - sudo apt-get install g++



# PYTHON



# PYTHON

## Why?

- **Wide array of libraries for most of the applications**
  - Numerical (NumPy, Pandas)
  - Scientific (SciPy)
  - Machine Learning (Sci-Kit, PyBrain, Gensim, NLTK)
  - Visualization (Matplotlib, Bokeh)
  - Database ORM (SQLAlchemy, PyMongo)
  - Big Data (PySpark, HadoopPy)
- **Easy to learn (in comparison to let say R)**
- **Multi-Processing**
- **Preferred language of choice for Data Scientists**



# IPYTHON

**IPython provides a rich architecture for interactive computing with:**

- Powerful interactive shells (terminal and Qt-based).
- A browser-based notebook with support for code, rich text, mathematical expressions, inline plots and other rich media.
- Support for interactive data visualization and use of GUI toolkits.
- Flexible, embeddable interpreters to load into your own projects.
- Easy to use, high performance tools for parallel computing.

<http://ipython.org/index.html>



# IPYTHON (JUPYTER) NOTEBOOK

- Runs code in a web-browser
- Stored in a json format
- Allows for code and text (MarkDown)
- Has Debugger
- Has Checkpoint
- Sharing / Co-editing is a lot easier
- Has access to all python libraries (locally imported and system installed)
- All your assignments are going to be in that. ☺



# IPYTHON (JUPYTER) NOTEBOOK

## Installation (Ubuntu and Mac)

- pip install ipython[notebook]
- Install all dependencies

## Alternative:

- Use Anaconda (Its pretty great for Data Science)
  - <https://conda.io/docs/user-guide/getting-started.html>
  - Install - <https://conda.io/docs/user-guide/install/index.html>
- Docker
  - <https://docs.docker.com/get-started/#prepare-your-docker-environment>
  - Created for class:  
[https://github.com/somyamohanty/datascience\\_docker](https://github.com/somyamohanty/datascience_docker)
    - All libraries are pre-installed

## Windows Users:

- <http://technivore.org/posts/2016/02/27/windows-jupyter-three-ways.html>



# IPYTHON (JUPYTER) NOTEBOOK

## Run

- Locally
  - `ipython notebook`
- Remote Server
  - `ipython notebook --no-browser --port=8888`
- Starts at home directory
- <http://localhost:8888/tree>
- `.ipynb`

## Ipython and Python Basics



# CURRENT LOCAL SETUP

- Current requirements (Installed IPython, NumPy, SciPy, Pandas)
  - backports.ssl-match-hostname==3.4.0.2
  - certifi==2015.4.28
  - decorator==4.0.2
  - funcsig==0.4
  - functools32==3.2.3.post2
  - ipykernel==4.0.3
  - ipython==4.0.0
  - ipython-genutils==0.1.0
  - Jinja2==2.8
  - jsonschema==2.5.1
  - jupyter-client==4.0.0
  - jupyter-core==4.0.4
  - MarkupSafe==0.23
  - matplotlib==1.4.3
  - mistune==0.7
  - mock==1.3.0
  - nbconvert==4.0.0
  - nbformat==4.0.0
  - nose==1.3.7
  - notebook==4.0.2
  - numpy==1.9.2
  - pandas==0.16.2
  - path.py==7.6.1
  - pbr==1.6.0
  - pexpect==3.3
  - pickleshare==0.5
  - ptyprocess==0.5
  - Pygments==2.0.2
  - pyparsing==2.0.3
  - python-dateutil==2.4.2
  - pytz==2015.4
  - pyzmq==14.7.0
  - scipy==0.16.0
  - simplegeneric==0.8.1
  - six==1.9.0
  - terminado==0.5
  - tornado==4.2.1
  - traitlets==4.0.0
  - wheel==0.24.0



# **RESOURCES**



# CLOUD RESOURCES

- **Microsoft Azure – UNCG**
  - <https://uncg-jupyterhub.eastus.cloudapp.azure.com/hub/login>
  - <https://kangaroo.uncg.edu>
  - Login with Spartan ID
    - It will show security exception, just confirm it and proceed.
    - It takes a few minutes to startup
- **Google CoLab**
  - <https://colab.research.google.com/notebooks/welcome.ipynb>
  - Links to google drive for data
- **Both have limitations on the computational resources**
  - Using your own laptop/desktop gives you more control.

