Intro to Github and Git

Sasan Bahadaran

May 9, 2017

COMMERCE
DATA SERVICE

COMMERCE
DATA SERVICE

Commerce Data Academy

- A data education initiative of the Commerce Data Service.
- Launched by CDS to offer data science, data engineering, and web development training to employees of the US Department of Commerce.
- Course schedule and materials (e.g. slides, code, papers) produced for the Commerce Data Academy on Github.
- Questions? Feel free to write us at Data Academy (dataacademy@doc.gov).

Goals

Our goals for the class

- Explain and make the case for version control.
- Collaboration in coding/software engineering.
- Illustrate what Git software is and what it can do.
- Differentiate Git (the software) and Github (the website).
- Describe how we integrate Git and Github into our project workflows.

Goals

Your goals for the class

- Understand what version control is and why should you use it for your projects.
- Start using Git on the command line.
- Experiment with pushing repos to Github.
- Practice working with a team using Waffle.io.

Prerequisites

1. Create your own Github account
2. Create your own Waffle.io account
3. Download/install Git
4. Download/install Anaconda's Python distribution
5. Verify your access to Terminal (Mac) or Powershell (Windows)

Any challenges? Questions?

Open Sources Installations

- We use open source and free software, so they should have a minimal impact on your IT department!
- DOC has provided guidance that states that states that Github and all the tools that we are teaching are permissible under policy.
- However, it is up to the CIO of each bureau to accept this guidance policy or not.
- DOC has a formalized Github policy: https://github.com/CommerceGov/Policies-and-Guidance/blob/master/GithubGuidanceforDepartmentofCommerce.md

Review

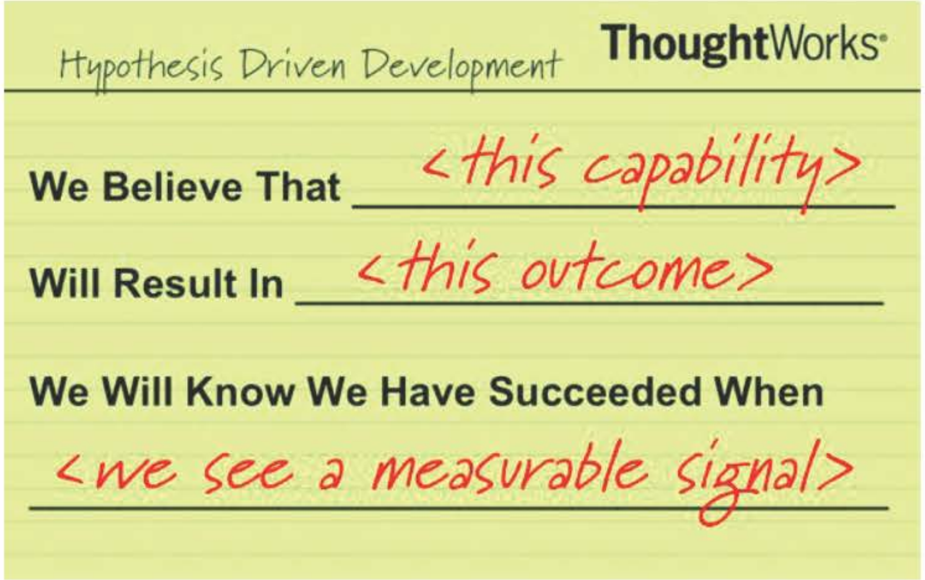
What is data science?

“Data science is the practice of **transforming raw data into insights, products, and applications** to **empower data-driven decision making**. It combines proven, time-tested methods from fields including statistics, natural sciences, computer science, operations research, and design in ways that are particularly well-suited to the data age. These methods, which range from data mining and visualization to predictive modeling, can scale from small to large datasets and can handle structured data as well as unstructured data like text and images.”

Jeff Chen, Chief Data Scientist
U.S. Department of Commerce

How is data science different from data analytics?

What is hypothesis-driven development?



Hypothesis Driven Development

ThoughtWorks®

We Believe That <this capability>

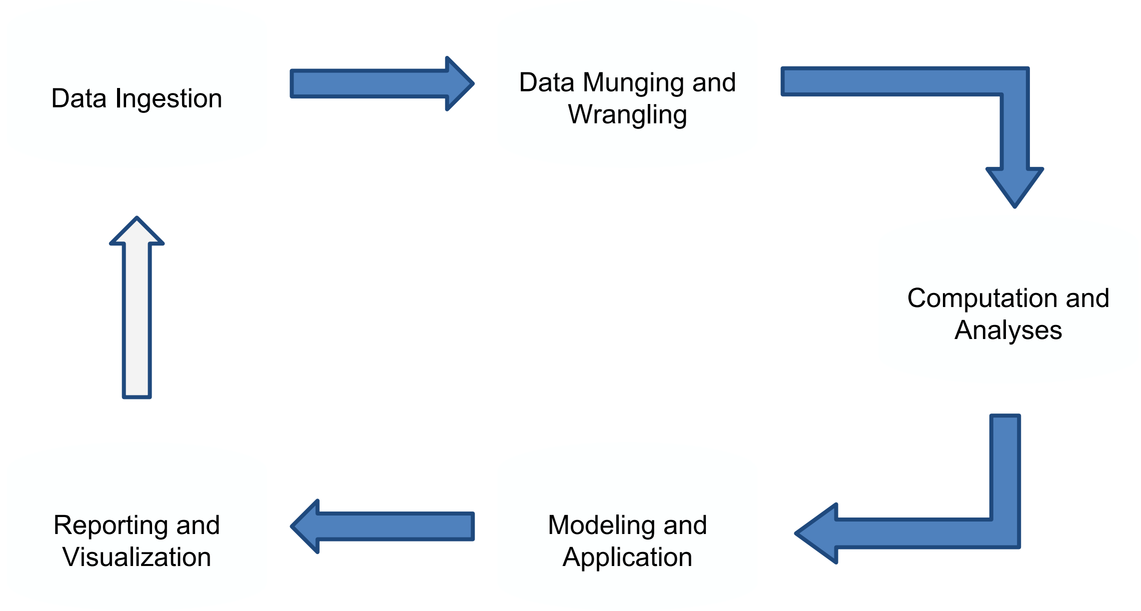
Will Result In <this outcome>

We Will Know We Have Succeeded When

<we see a measurable signal>

What tools do data scientists use?

What is the data science pipeline?



Data Ingestion

Data Munging and Wrangling

Computation and Analyses

Modeling and Application

Reporting and Visualization

What is a data product?

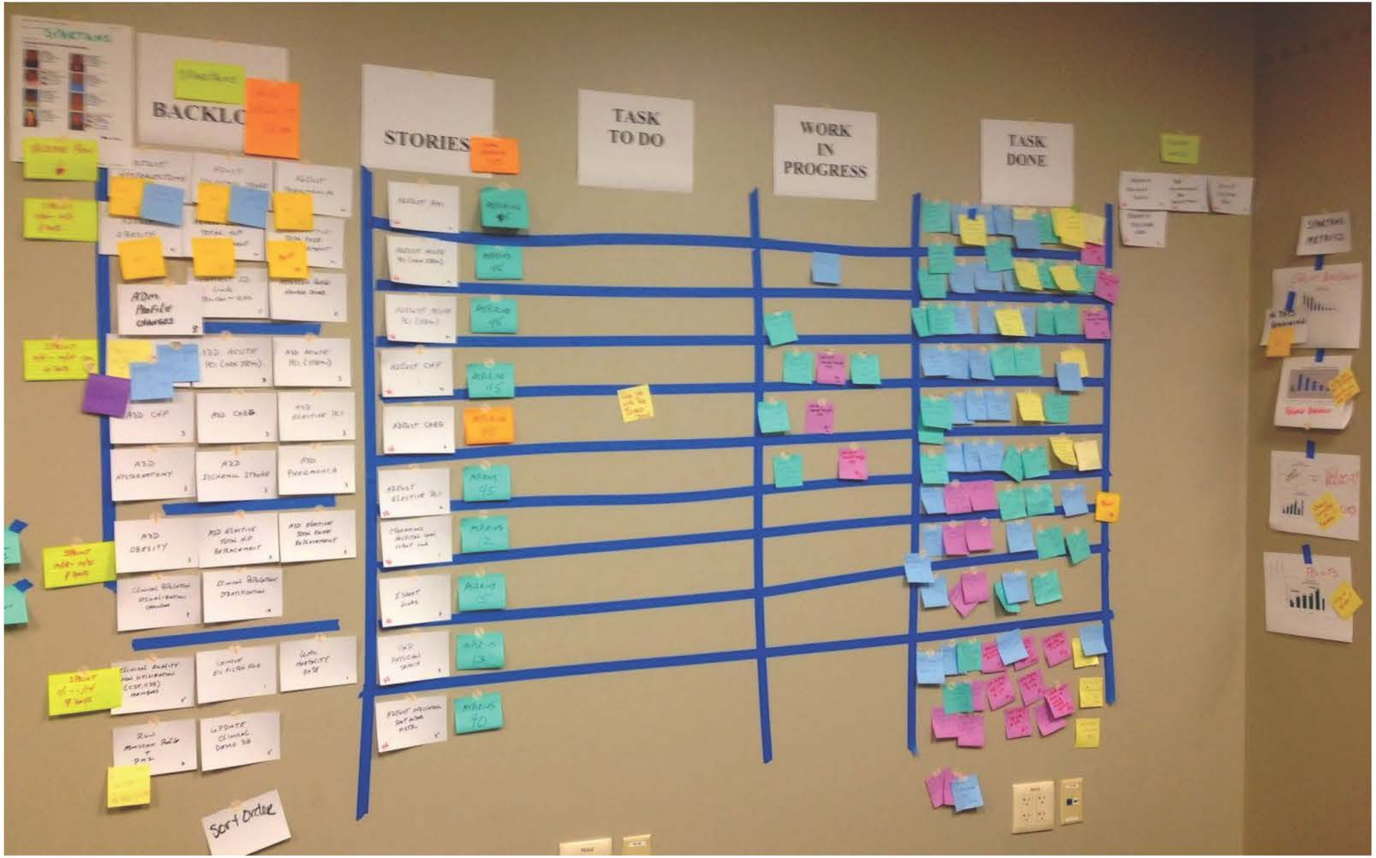
How are data products different from analytical insights?

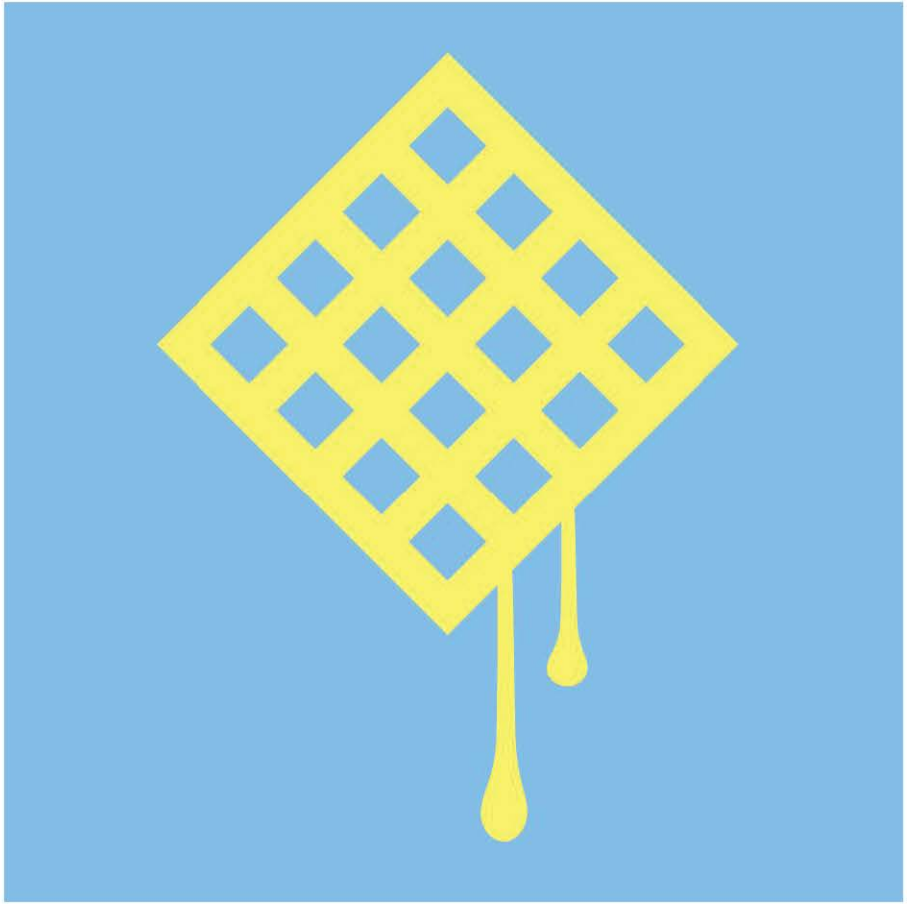
Data products are self-adapting, broadly applicable economic engines that derive their value from data and generate more data by influencing human behavior or by making inferences or predictions upon new data.

Benjamin Bengfort

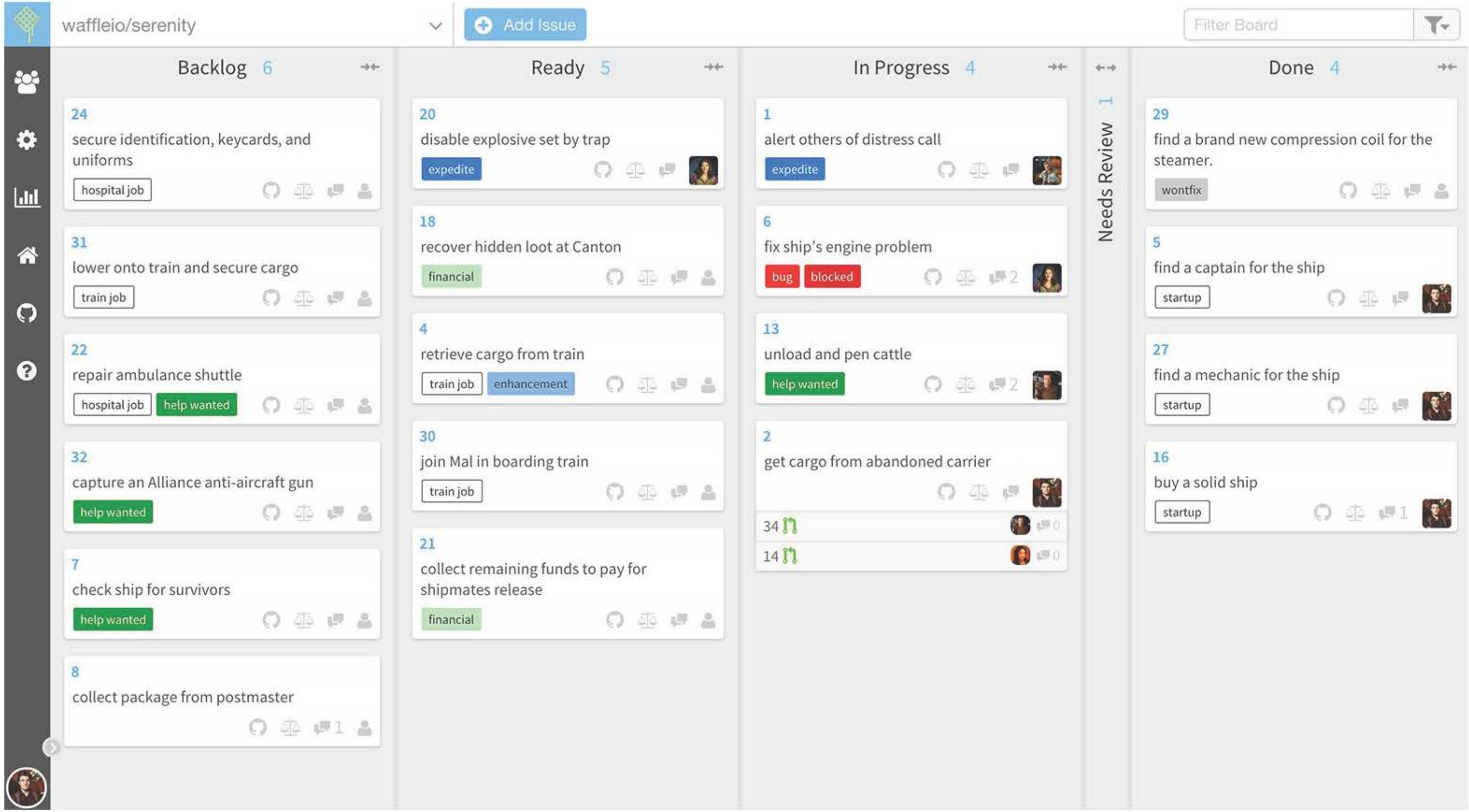
What is software engineering?

What does collaboration look like in a data group?


BACKLOG
STORIES
TASK
TO DO
WORK
IN
PROGRESS
TASK
DONE
Sort Order



COMMERCE
DATA SERVICE



Backlog 6

24
secure identification, keycards, and uniforms
hospital job

31
lower onto train and secure cargo
train job

22
repair ambulance shuttle
hospital job help wanted

32
capture an Alliance anti-aircraft gun
help wanted

7
check ship for survivors
help wanted

8
collect package from postmaster
1

Ready 5

20
disable explosive set by trap
expedite

18
recover hidden loot at Canton
financial

4
retrieve cargo from train
train job enhancement

30
join Mal in boarding train
train job

21
collect remaining funds to pay for shipmates release
financial

In Progress 4

1
alert others of distress call
expedite

6
fix ship's engine problem
bug blocked 2

13
unload and pen cattle
help wanted 2

2
get cargo from abandoned carrier
34 0
14 0

Needs Review 1

Done 4

29
find a brand new compression coil for the steamer.
wontfix

5
find a captain for the ship
startup

27
find a mechanic for the ship
startup

16
buy a solid ship
startup 1

Version Control

Examples?

COMMERCE
DATA SERVICE

Google Drive

SharePoint

Dropbox

Wikis

Bitbucket

TortoiseSVN

What is version control?

Other names?

What problems does this solve?

What are the benefits?

What are some common features?

Definition:

The management of changes to electronic documents and, in particular, computer programs.

"In computer software engineering, revision control is any kind of practice that tracks and provides control over changes to source code."

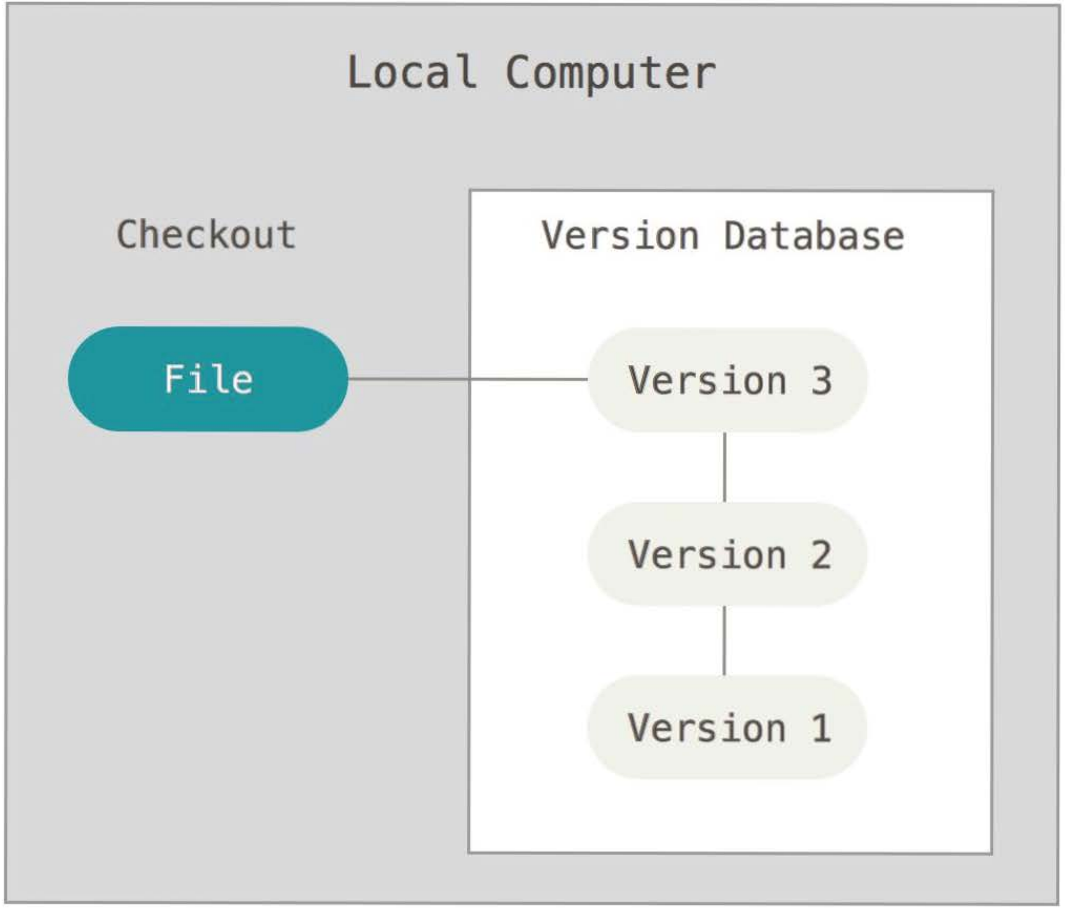
Wikipedia knows everything

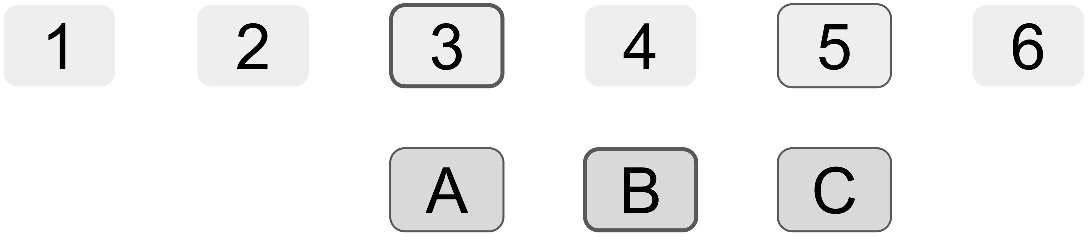
Tell us about a time when you could have used some version control...

Local Version Control Systems

COMMERCE DATA SERVICE

Version Control:
A Visualization

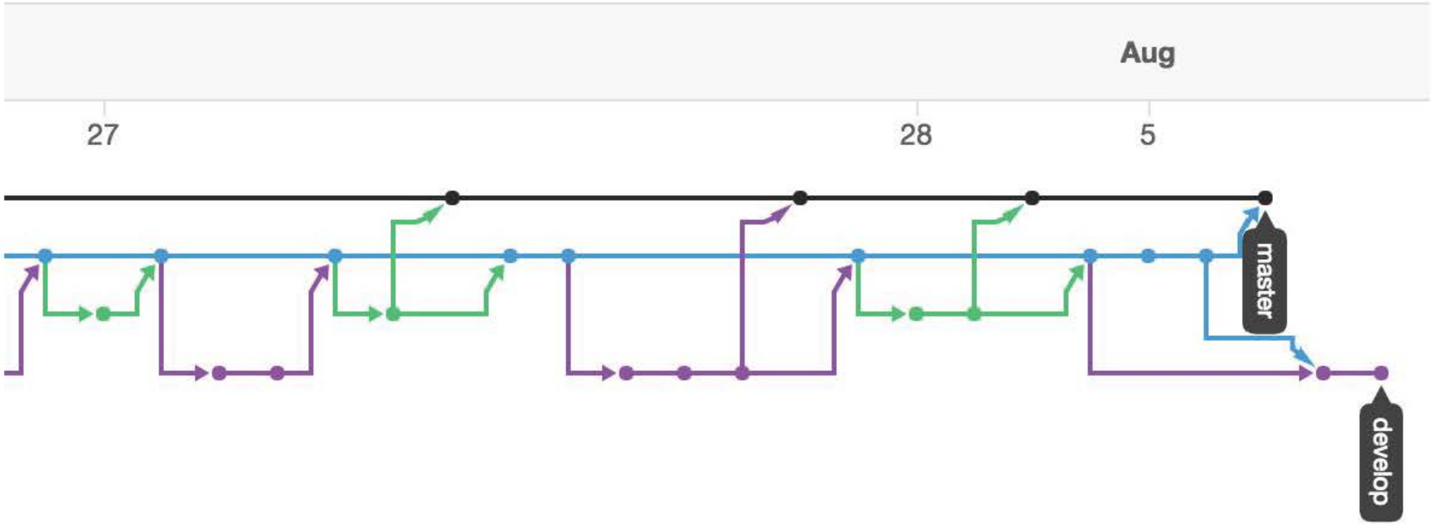

Local Computer
Checkout
File
Version Database
Version 3
Version 2
Version 1



1 2 3 4 5 6

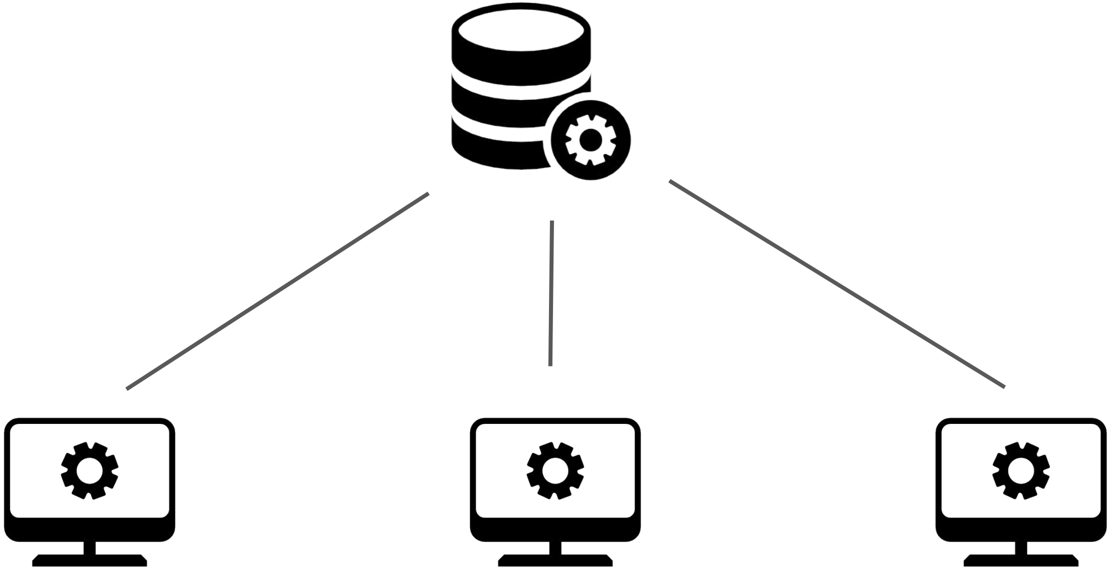
A B C

Branches and revisions through time - example scenario



Branches and revisions through time - actual workflow

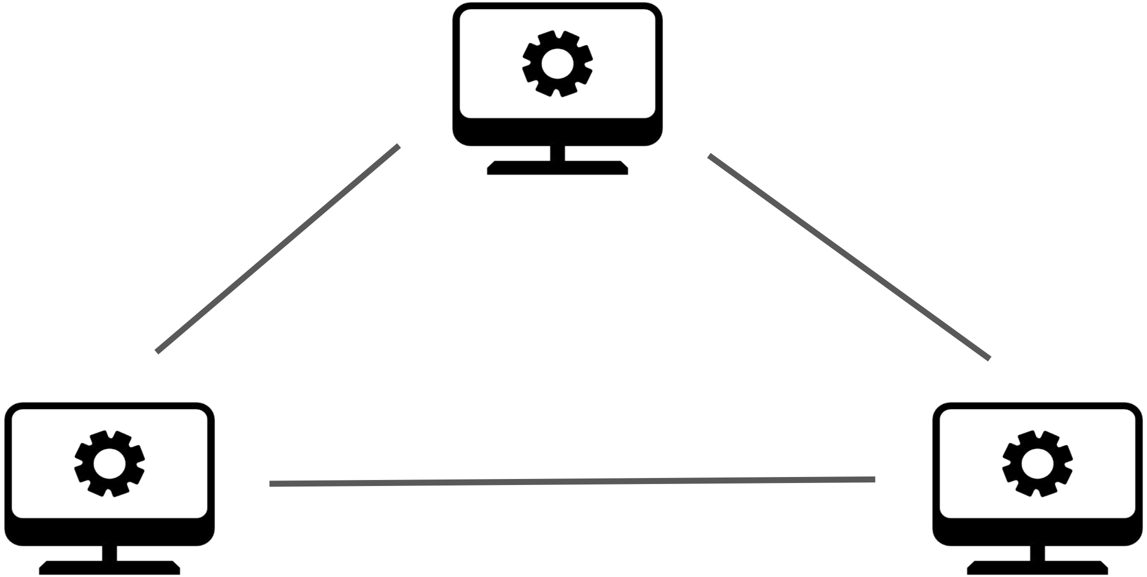
Distributed vs. Centralized



What are the benefits?

What are the weaknesses?

Centralized



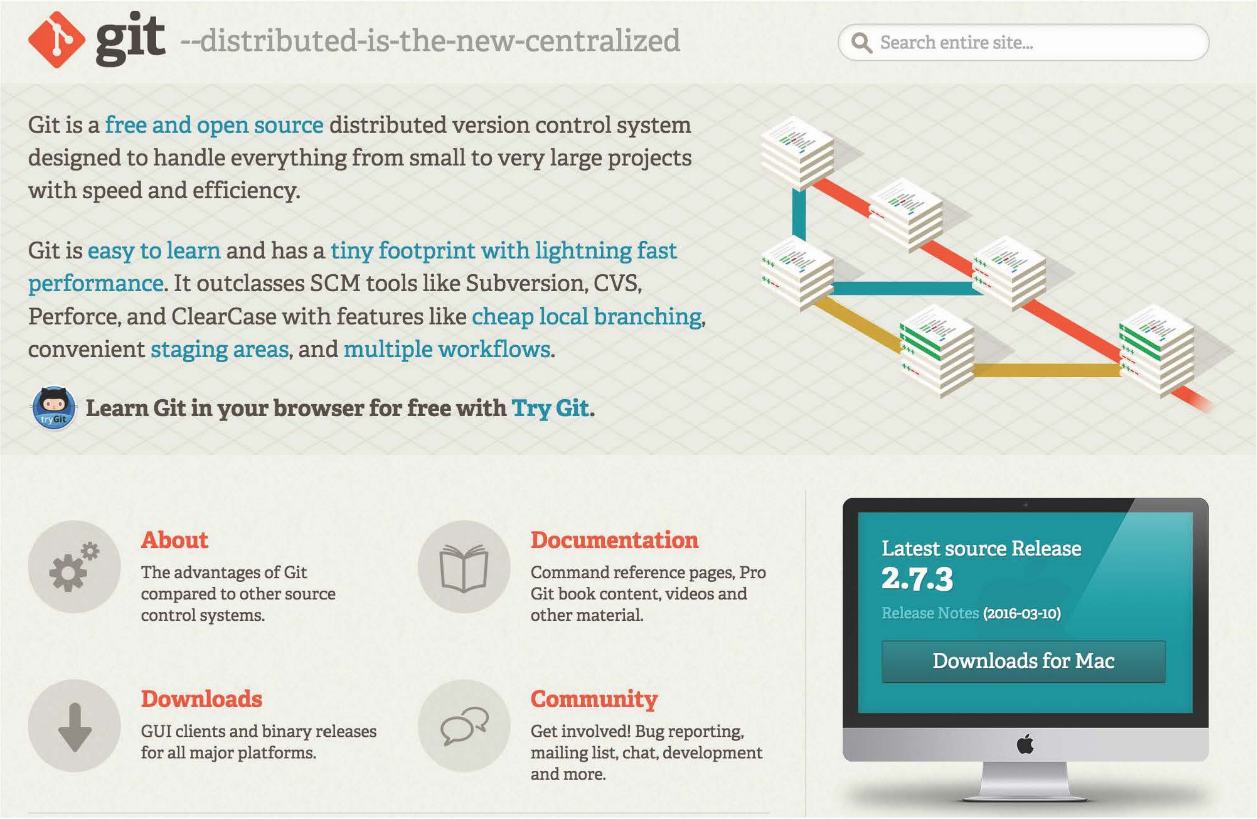
What are the benefits?

What are the weaknesses?

Decentralized

Git

Installing Git



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

Search entire site...

Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

Git is easy to learn and has a tiny footprint with lightning fast performance. It outclasses SCM tools like Subversion, CVS, Perforce, and ClearCase with features like cheap local branching, convenient staging areas, and multiple workflows.

Learn Git in your browser for free with Try Git.

About
The advantages of Git compared to other source control systems.

Documentation
Command reference pages, Pro Git book content, videos and other material.

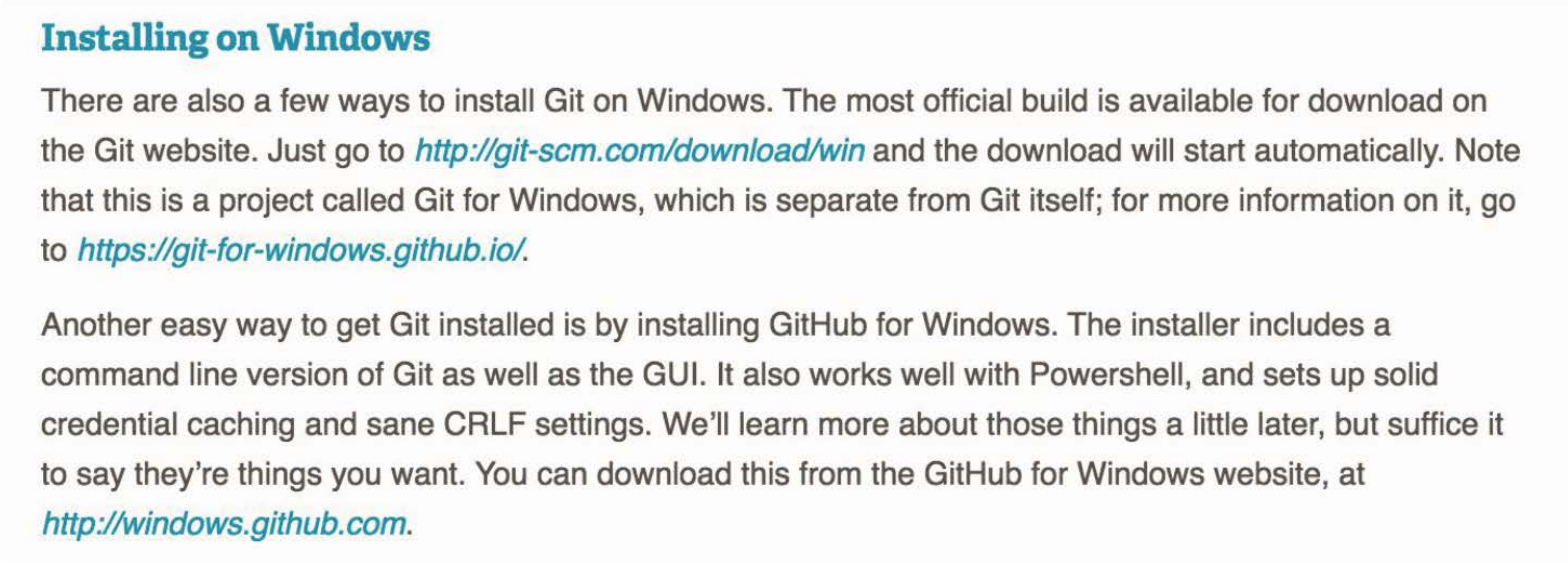
Downloads
GUI clients and binary releases for all major platforms.

Community
Get involved! Bug reporting, mailing list, chat, development and more.

Latest source Release
2.7.3
Release Notes (2016-03-10)

Downloads for Mac

Installing Git



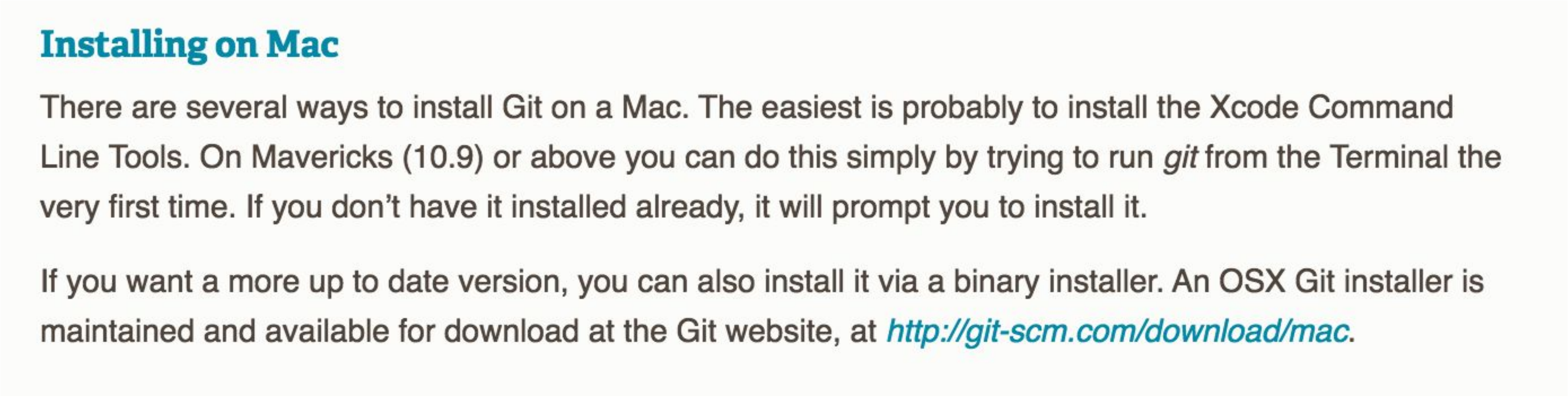
Installing on Windows

There are also a few ways to install Git on Windows. The most official build is available for download on the Git website. Just go to *http://git-scm.com/download/win* and the download will start automatically. Note that this is a project called Git for Windows, which is separate from Git itself; for more information on it, go to *https://git-for-windows.github.io/*.

Another easy way to get Git installed is by installing GitHub for Windows. The installer includes a command line version of Git as well as the GUI. It also works well with Powershell, and sets up solid credential caching and sane CRLF settings. We'll learn more about those things a little later, but suffice it to say they're things you want. You can download this from the GitHub for Windows website, at *http://windows.github.com*.

http://git-for-windows.github.io/

Installing Git



Installing on Mac

There are several ways to install Git on a Mac. The easiest is probably to install the Xcode Command Line Tools. On Mavericks (10.9) or above you can do this simply by trying to run *git* from the Terminal the very first time. If you don't have it installed already, it will prompt you to install it.

If you want a more up to date version, you can also install it via a binary installer. An OSX Git installer is maintained and available for download at the Git website, at *http://git-scm.com/download/mac*.

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

Git - History Lesson

- Originally conceived/created by Linus Torvalds (after a fight with BitKeeper)
- Distributed Version Control
- Open Source
- Initial release: 7 April 2005
- All metadata is stored in the .git directory

- Speed
- Simple design
- Strong support for non-linear development (thousands of parallel branches)
- Fully distributed
- Able to handle large projects like the Linux kernel efficiently (speed and data size)

Git - Advantages

Object Database

where git stores metadata about each commit

Index / Staging Area

file snapshots to be included in next commit

Working Directory

the “physical” files on a computer

Git - “Places”

Committed

data is safely stored in your local object database

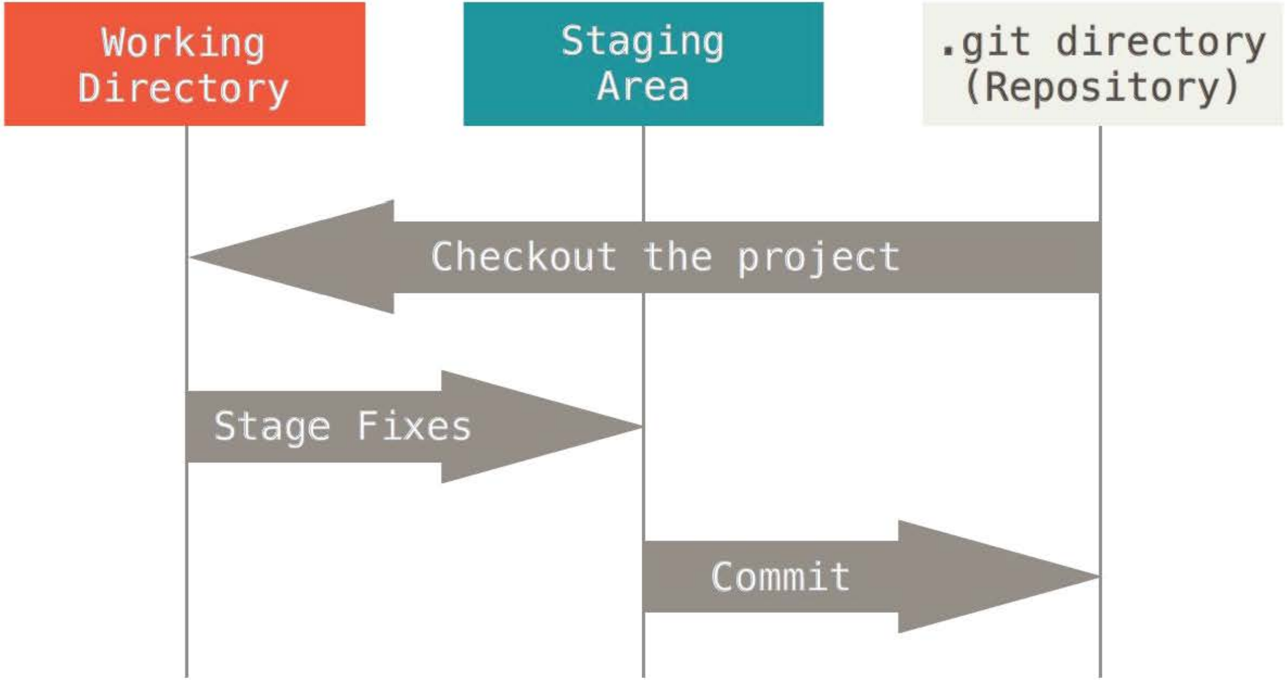
Staged

marked such that the current state of the modified file will be included in the next commit

Modified

changed but not staged or committed

Git - “Stages”



Git - Areas/places

COMMERCE
DATA SERVICE

Git Commands

git init

create a new git repository to manage the current folder

git clone <repository address>

downloads an existing git repository for the first time

git add <file path>

marks individual/modified files to be added to the index/staging area for next commit

git commit -m <message>

takes metadata/changes from staging and adds to the object database

Git - Basic Commands

git fetch <server> <branch>

updates your object database but does not change the working directory

git merge <source branch>

applies the commits from source branch to the current working directory (which is the manifestation of another branch)

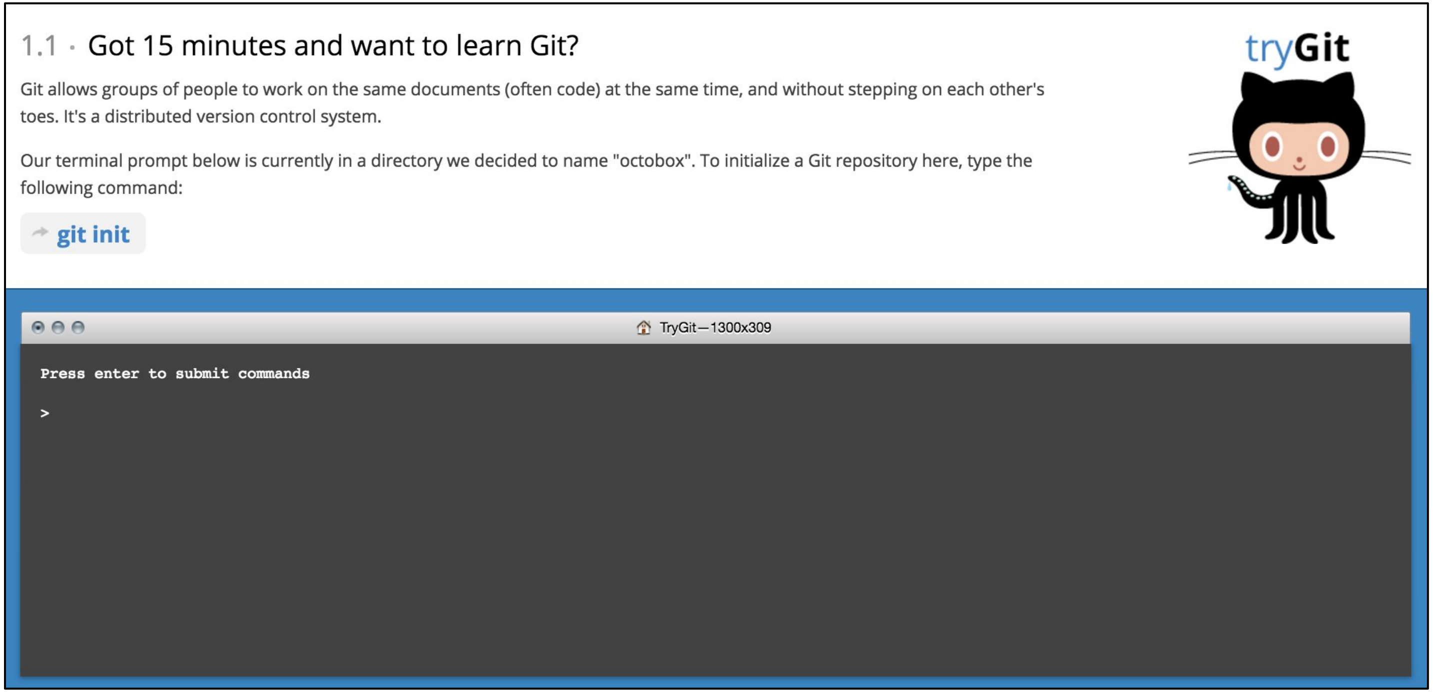
git pull <server> <branch>

performs a fetch and then merges those changes into your working directory

git push <server> <branch>

sends your latest branch commits to the remote server

Git - Basic Commands



1.1 · Got 15 minutes and want to learn Git?

Git allows groups of people to work on the same documents (often code) at the same time, and without stepping on each other's toes. It's a distributed version control system.

Our terminal prompt below is currently in a directory we decided to name "octobox". To initialize a Git repository here, type the following command:

```
git init
```

```
Press enter to submit commands

>
```

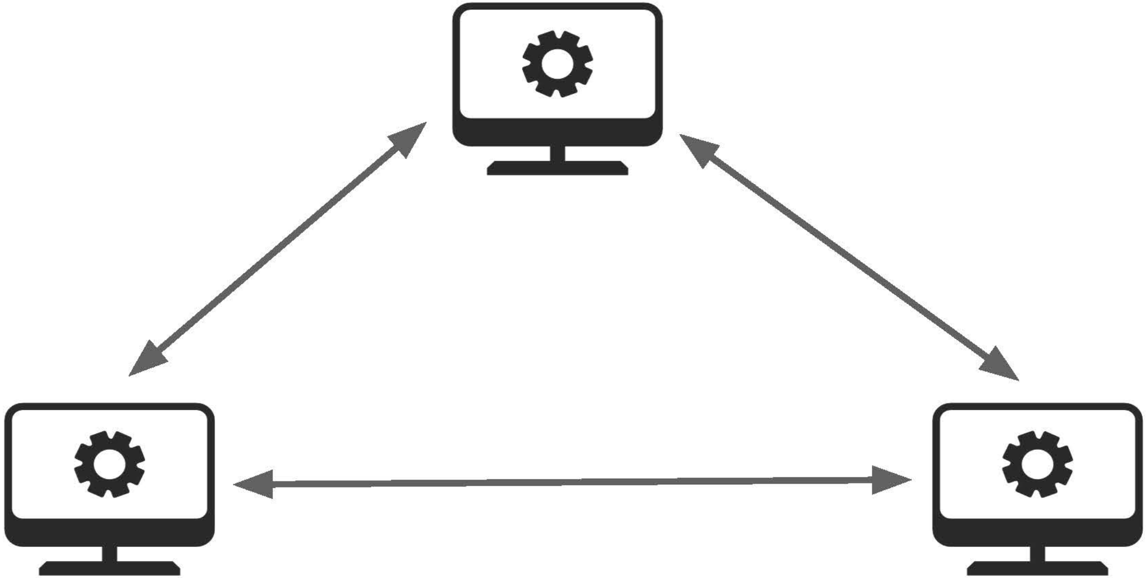
Git Challenge (20 minutes)

https://try.github.io/levels/1/challenges/1

Github

- A remote git repository
- A website
  - provides secure access
  - provides repository metadata & reports
  - provides tools for development teams
- Launched: April 10, 2008
- ~10 million users in 2015

Github



Non-local git repositories are called “remotes”

Object Database

where git stores metadata about each commit

Index / Staging Area

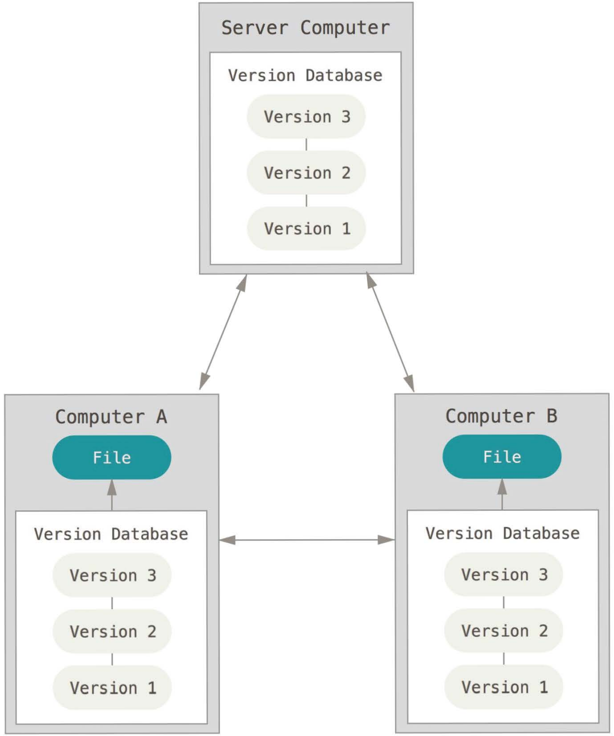
file snapshots to be included in next commit

Working Directory

the “physical” files on a computer

Git - “Places”

COMMERCE DATA SERVICE



Github: A Distributed Version Control example

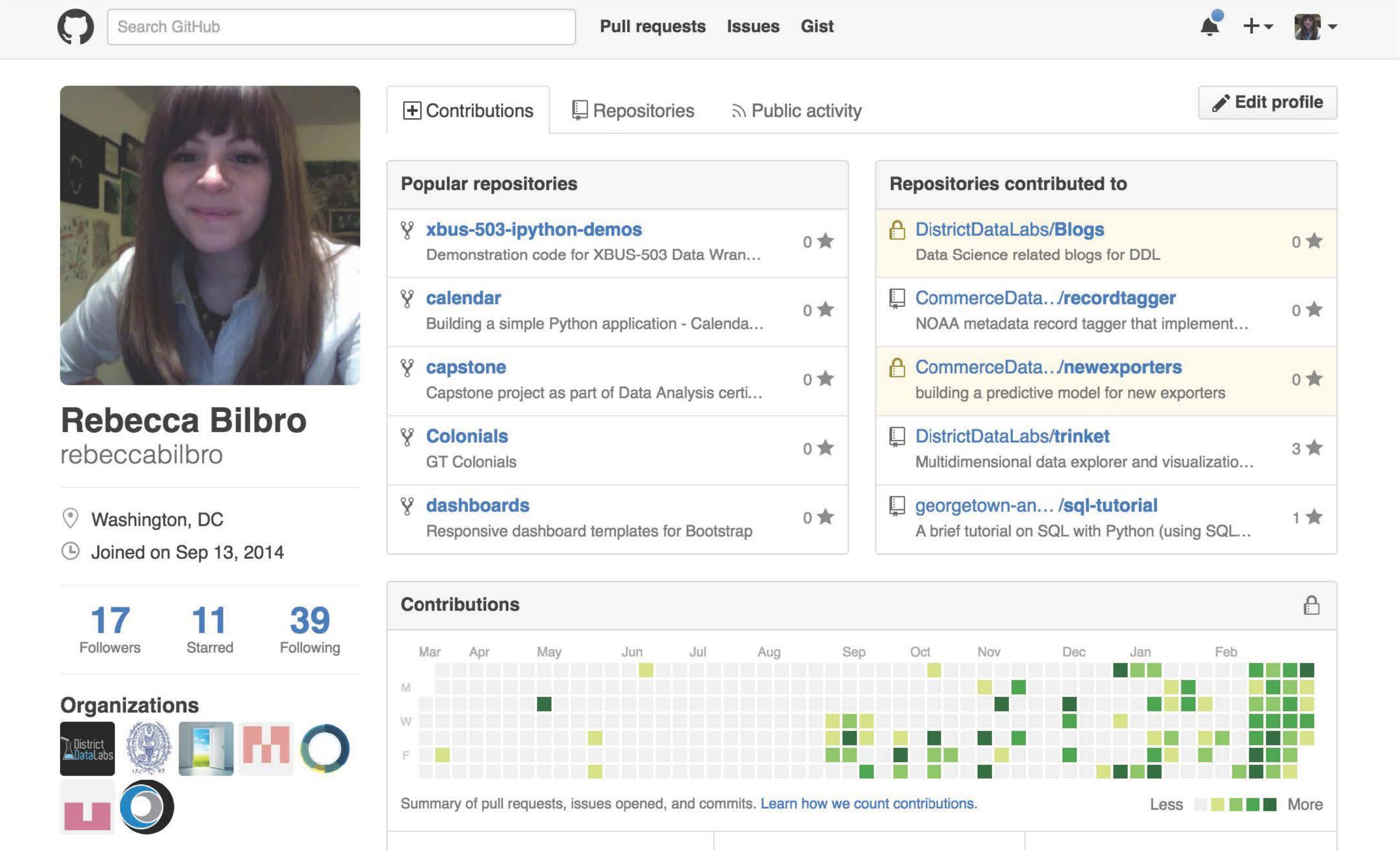
- The “origin” remote is automatically created when you clone
- It is the default remote to use for pushing and pulling
- There is nothing special about “origin” it is just a default name

Git - “Origin”

COMMERCE DATA SERVICE

User Account

COMMERCE
DATA SERVICE



Rebecca Bilbro
rebeccabilbro

Washington, DC
Joined on Sep 13, 2014

17 Followers
11 Starred
39 Following

Organizations

Popular repositories

xbus-503-ipython-demos Demonstration code for XBUS-503 Data Wran...	0 ★
calendar Building a simple Python application - Calenda...	0 ★
capstone Capstone project as part of Data Analysis certi...	0 ★
Colonials GT Colonials	0 ★
dashboards Responsive dashboard templates for Bootstrap	0 ★

Repositories contributed to

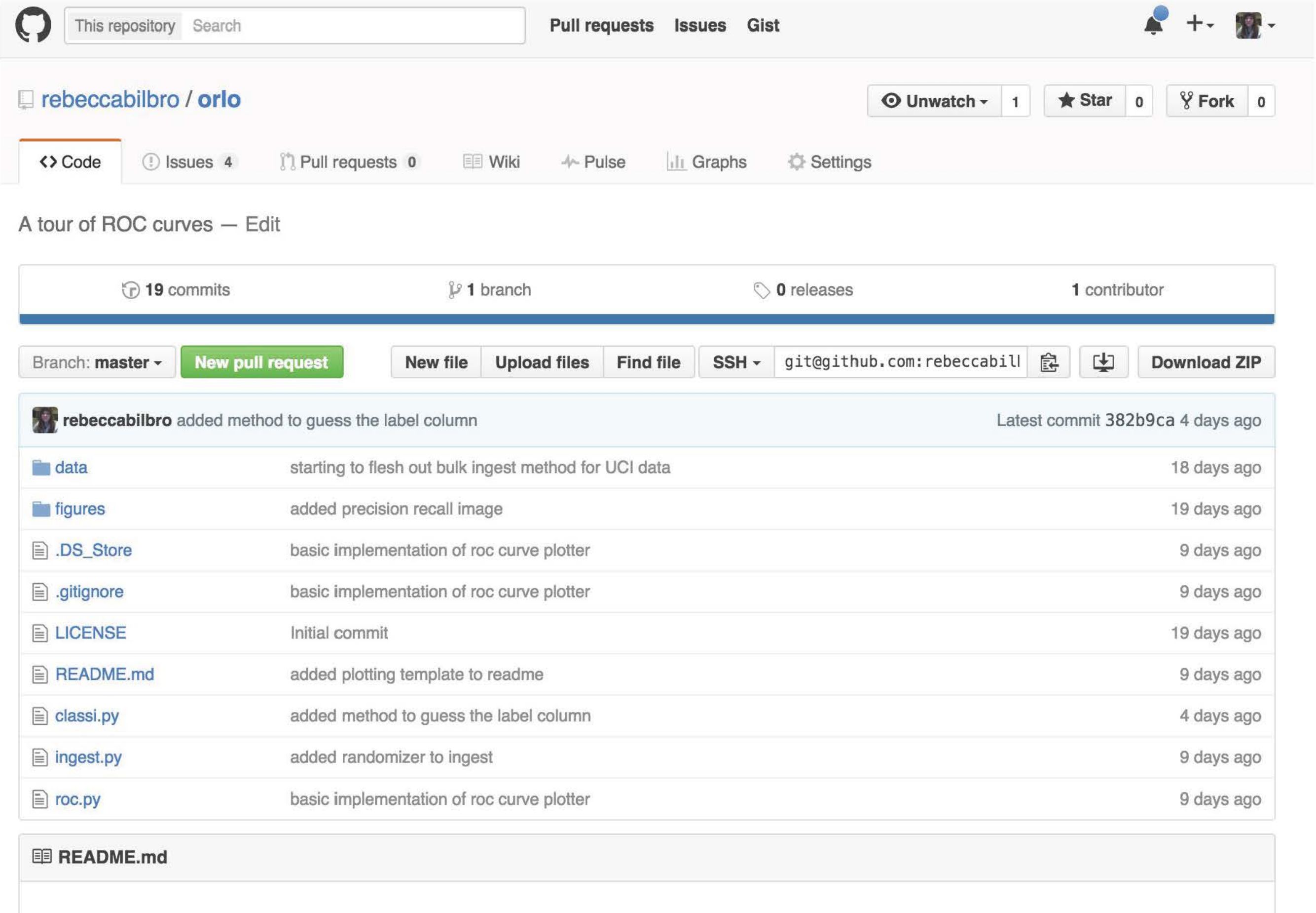
DistrictDataLabs/**Blogs** Data Science related blogs for DDL	0 ★
CommerceData.../**recordtagger** NOAA metadata record tagger that implement...	0 ★
CommerceData.../**newexporters** building a predictive model for new exporters	0 ★
DistrictDataLabs/**trinket** Multidimensional data explorer and visualizatio...	3 ★
georgetown-an... /**sql-tutorial** A brief tutorial on SQL with Python (using SQL...	1 ★

Contributions

Summary of pull requests, issues opened, and commits. Learn how we count contributions.

Less ⬜🟩🟩🟩🟩 More

Repo



rebeccabilbro / orlo

A tour of ROC curves — Edit

19 commits | 1 branch | 0 releases | 1 contributor

git@github.com:rebeccabill

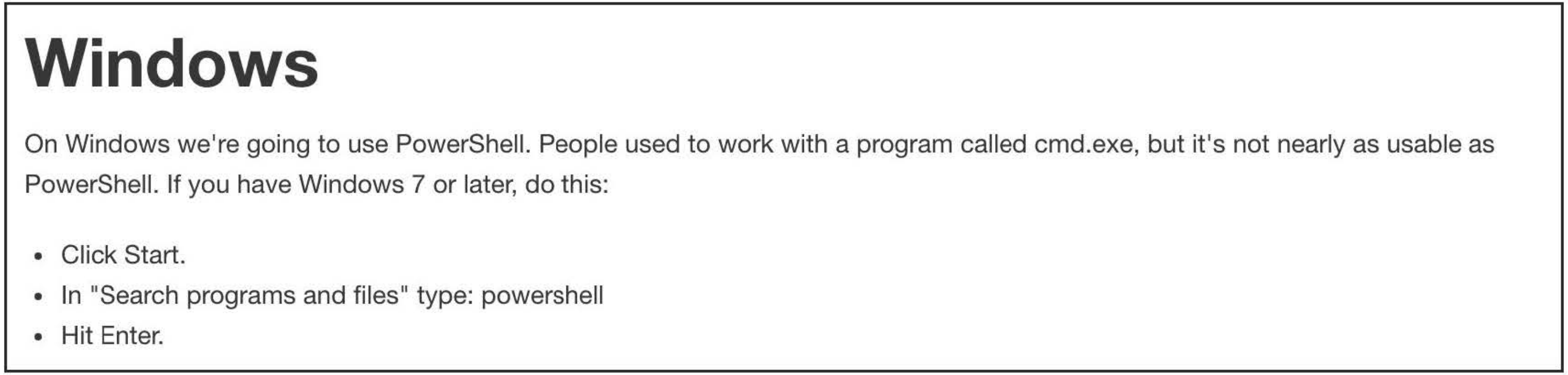
rebeccabilbro added method to guess the label column — Latest commit 382b9ca 4 days ago

data	starting to flesh out bulk ingest method for UCI data	18 days ago
figures	added precision recall image	19 days ago
.DS_Store	basic implementation of roc curve plotter	9 days ago
.gitignore	basic implementation of roc curve plotter	9 days ago
LICENSE	Initial commit	19 days ago
README.md	added plotting template to readme	9 days ago
classi.py	added method to guess the label column	4 days ago
ingest.py	added randomizer to ingest	9 days ago
roc.py	basic implementation of roc curve plotter	9 days ago

README.md

Command Line

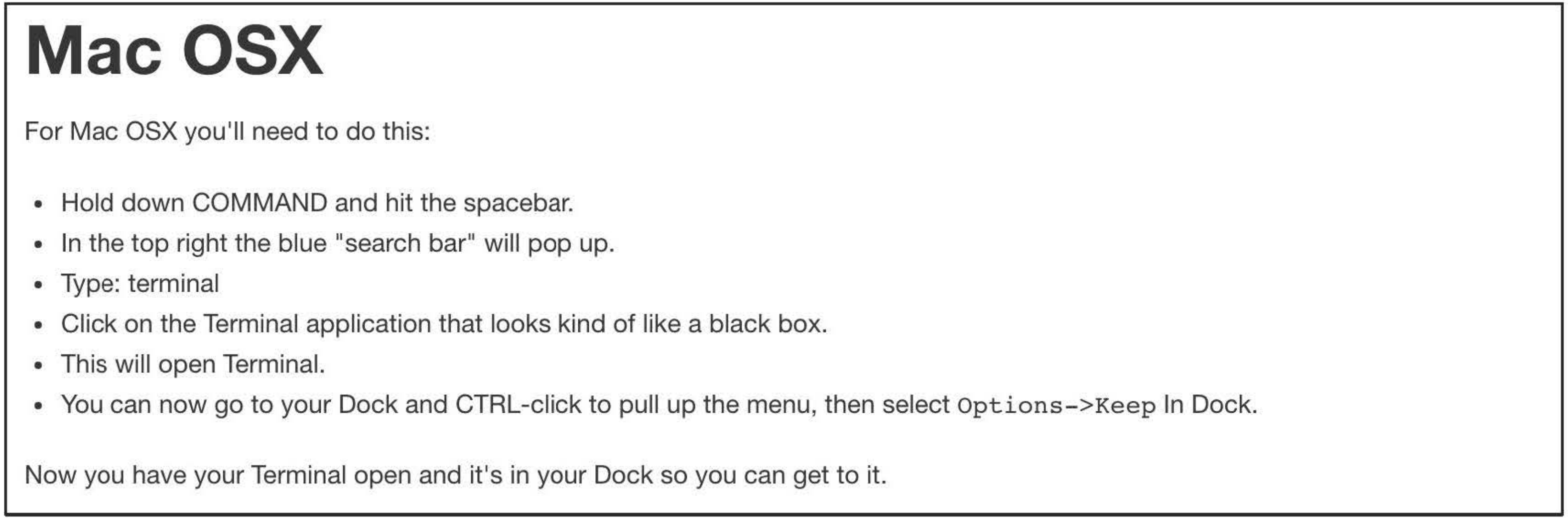
Shifting to the command line...



Windows

On Windows we're going to use PowerShell. People used to work with a program called cmd.exe, but it's not nearly as usable as PowerShell. If you have Windows 7 or later, do this:

- Click Start.
- In "Search programs and files" type: powershell
- Hit Enter.



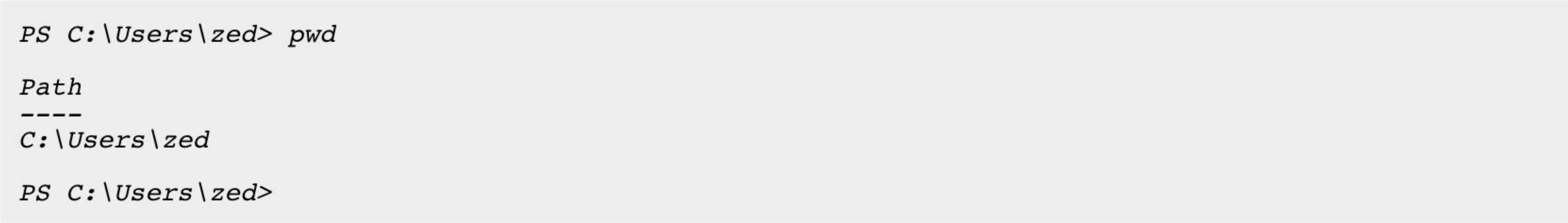
Mac OSX

For Mac OSX you'll need to do this:

- Hold down COMMAND and hit the spacebar.
- In the top right the blue "search bar" will pop up.
- Type: terminal
- Click on the Terminal application that looks kind of like a black box.
- This will open Terminal.
- You can now go to your Dock and CTRL-click to pull up the menu, then select `Options->Keep` In Dock.

Now you have your Terminal open and it's in your Dock so you can get to it.

Windows Powershell

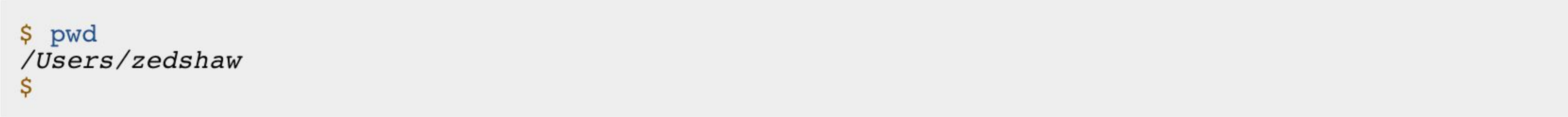


```
PS C:\Users\zed> pwd

Path
----
C:\Users\zed

PS C:\Users\zed>
```

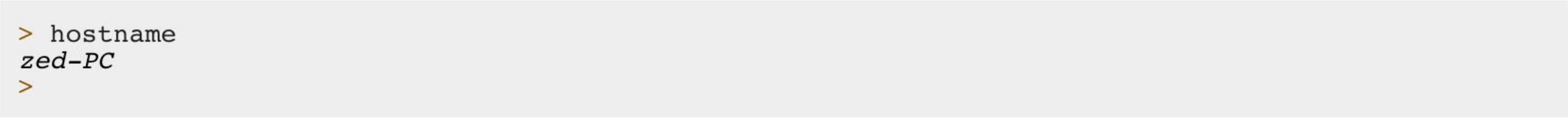
Mac OSX Terminal



```
$ pwd
/Users/zedshaw
$
```

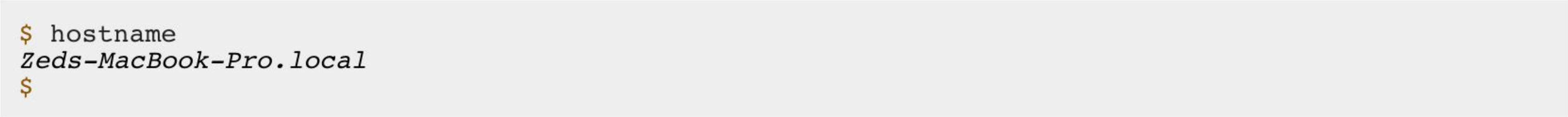
Where am I?

Windows Powershell



```
> hostname
zed-PC
>
```

Mac OSX Terminal



```
$ hostname
Zeds-MacBook-Pro.local
$
```

What’s my name?

Windows Powershell

```
> mkdir temp
> mkdir temp/stuff
> mkdir temp/stuff/things
> mkdir temp/stuff/things/frank/joe/alex/john
>
```

Mac OSX Terminal

```
$ mkdir temp
$ mkdir temp/stuff
$ mkdir temp/stuff/things
$ mkdir -p temp/stuff/things/frank/joe/alex/john
$
```

Make a directory

Windows Powershell

```
> cd temp
> pwd
>
```

Mac OSX Terminal

```
$ cd temp
$ pwd
$
```

Change between directories

Windows Powershell

```
> dir
>
```

Mac OSX Terminal

```
$ ls
$
```

List files and directories

Windows Powershell

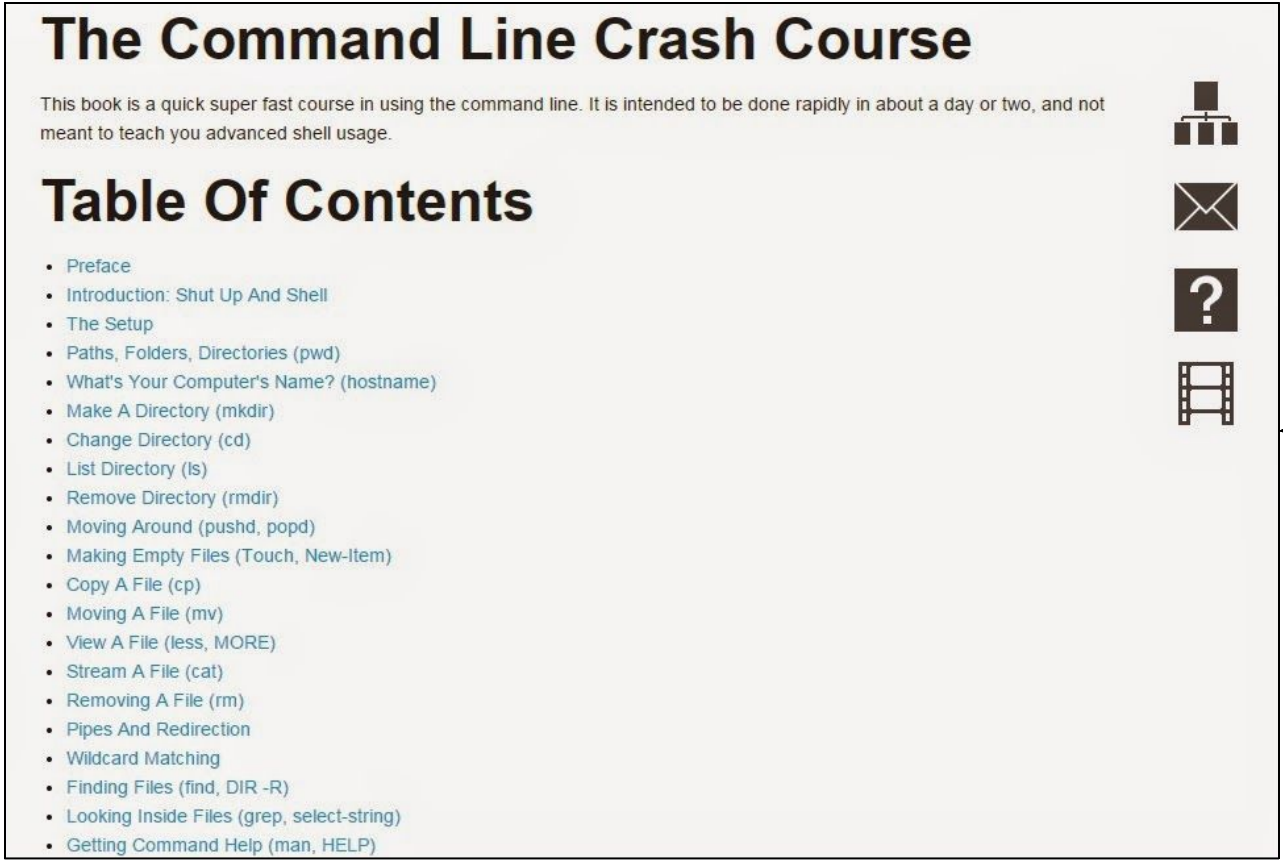
```
> cd temp
> New-Item iamcool.txt -type file
> dir
>
```

Mac OSX Terminal

```
$ cd temp
$ touch iamcool.txt
$ ls
$
```

Make an empty file

COMMERCE DATA SERVICE



The Command Line Crash Course

This book is a quick super fast course in using the command line. It is intended to be done rapidly in about a day or two, and not meant to teach you advanced shell usage.

Table Of Contents

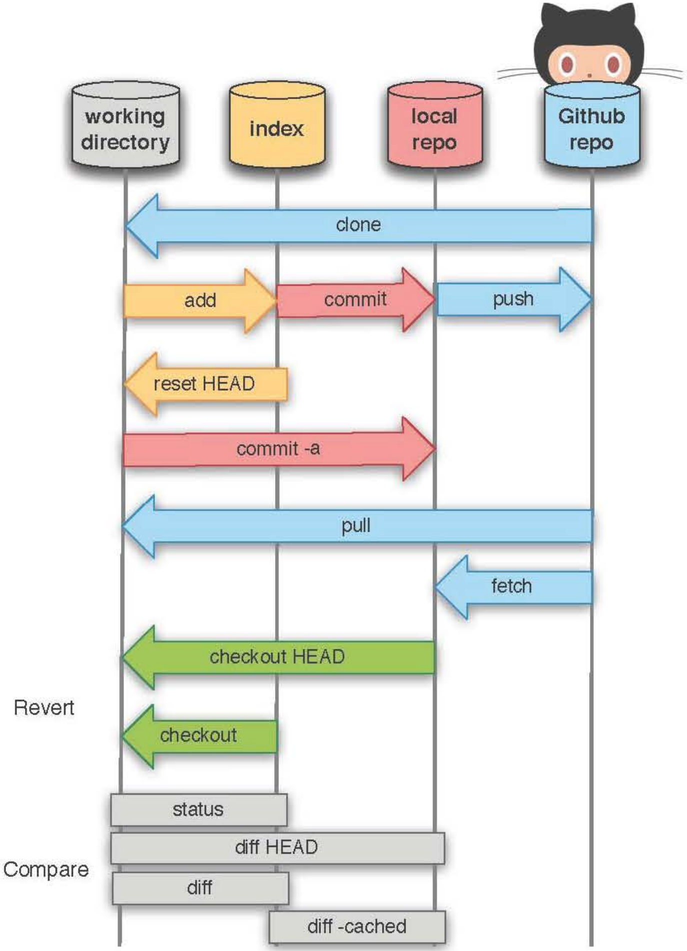
Zed Shaw's book

Let’s use what we’ve learned!

Merge Conflict Workshop (20 minutes):

http://bit.ly/xbus501-workshop-git

COMMERCE DATA SERVICE

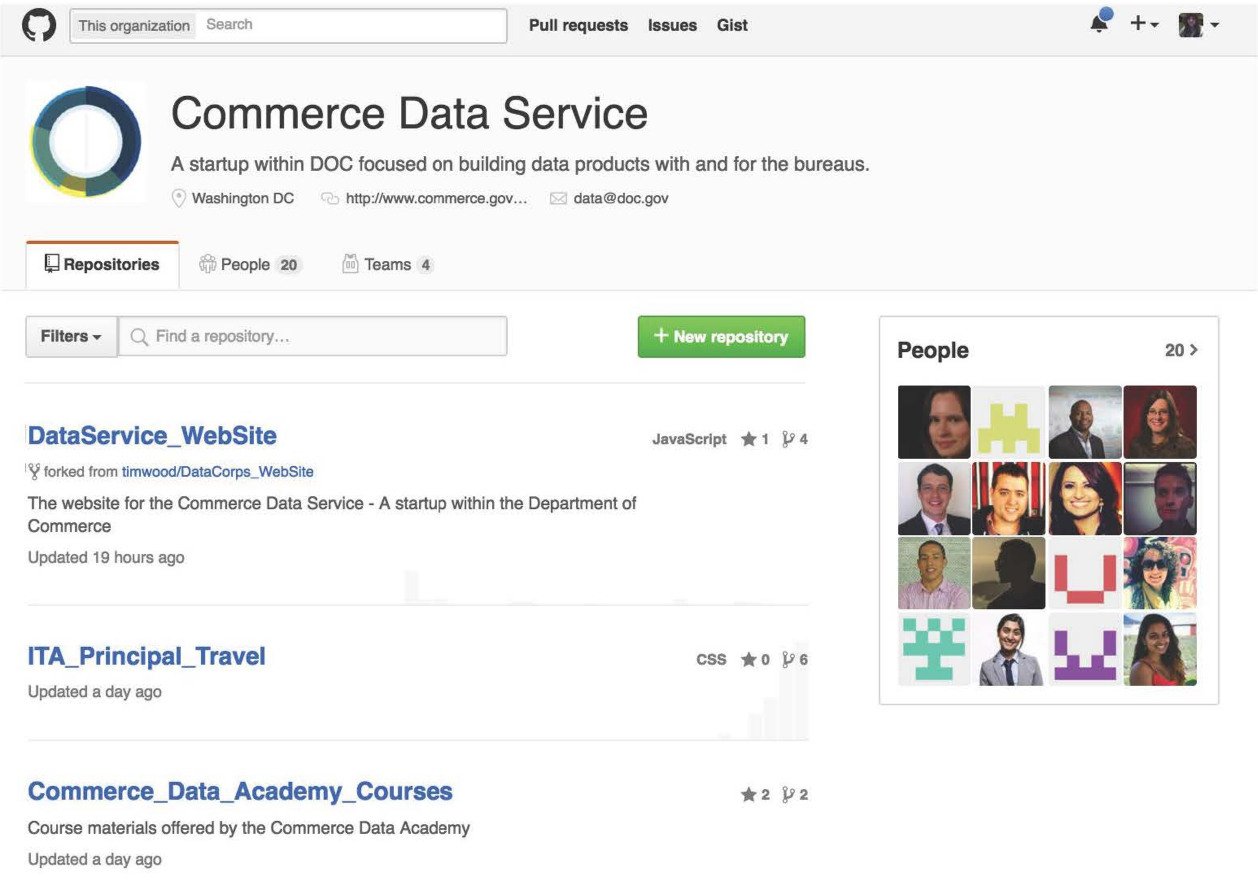


Teamwork

(makes the dream work!)

Organization

COMMERCE
DATA SERVICE



Commerce Data Service

A startup within DOC focused on building data products with and for the bureaus.

Washington DC http://www.commerce.gov… data@doc.gov

DataService_WebSite JavaScript ★ 1 4

forked from timwood/DataCorps_WebSite

The website for the Commerce Data Service - A startup within the Department of Commerce

Updated 19 hours ago

ITA_Principal_Travel CSS ★ 0 6

Updated a day ago

Commerce_Data_Academy_Courses ★ 2 2

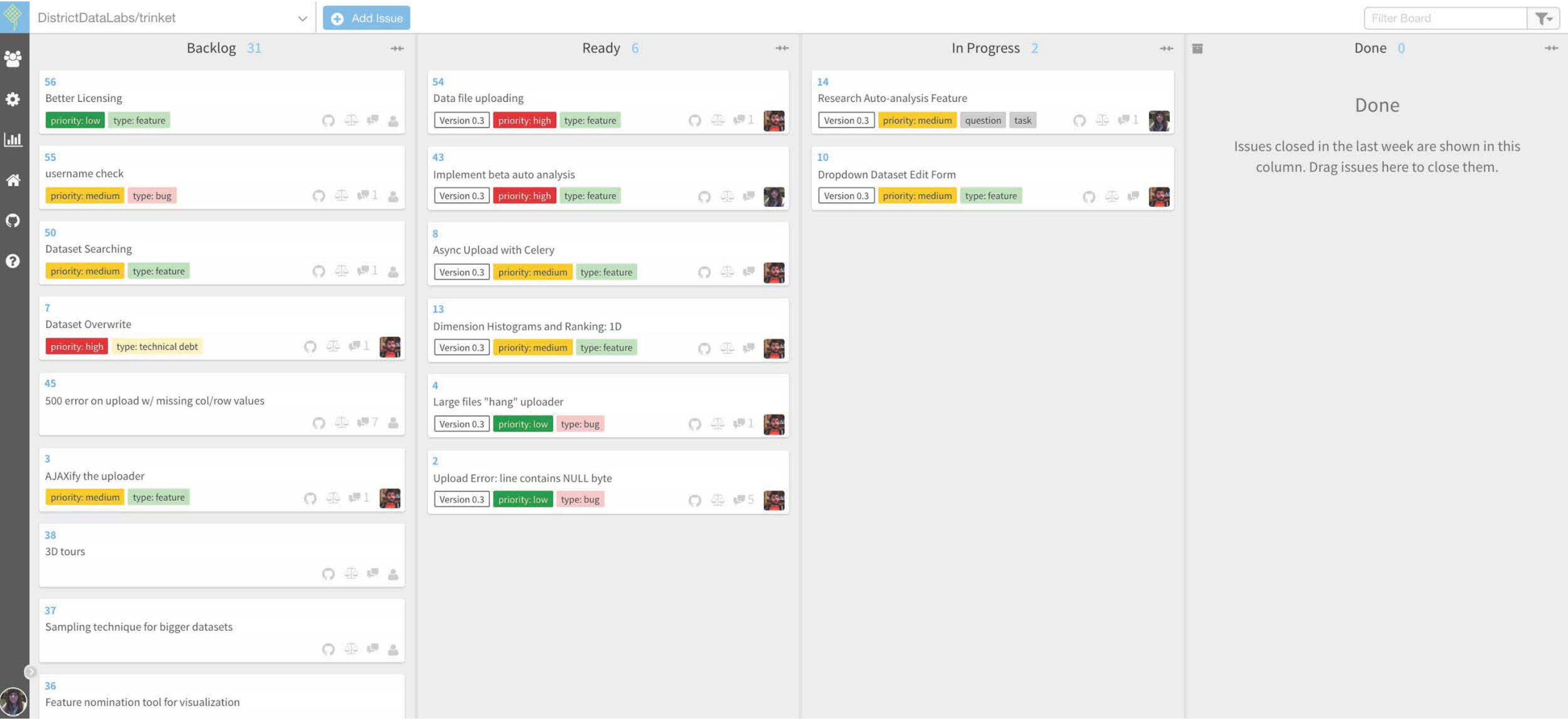
Course materials offered by the Commerce Data Academy

Updated a day ago

COMMERCE DATA SERVICE

Waffle

COMMERCE
DATA SERVICE



DistrictDataLabs/trinket

Add Issue

Filter Board

Backlog 31

56
Better Licensing
priority: low | type: feature

55
username check
priority: medium | type: bug

50
Dataset Searching
priority: medium | type: feature

7
Dataset Overwrite
priority: high | type: technical debt

45
500 error on upload w/ missing col/row values

3
AJAXify the uploader
priority: medium | type: feature

38
3D tours

37
Sampling technique for bigger datasets

36
Feature nomination tool for visualization

Ready 6

54
Data file uploading
Version 0.3 | priority: high | type: feature

43
Implement beta auto analysis
Version 0.3 | priority: high | type: feature

8
Async Upload with Celery
Version 0.3 | priority: medium | type: feature

13
Dimension Histograms and Ranking: 1D
Version 0.3 | priority: medium | type: feature

4
Large files "hang" uploader
Version 0.3 | priority: low | type: bug

2
Upload Error: line contains NULL byte
Version 0.3 | priority: low | type: bug

In Progress 2

14
Research Auto-analysis Feature
Version 0.3 | priority: medium | question | task

10
Dropdown Dataset Edit Form
Version 0.3 | priority: medium | type: feature

Done 0

Done

Issues closed in the last week are shown in this column. Drag issues here to close them.

Pair programming:
Make your own waffle!

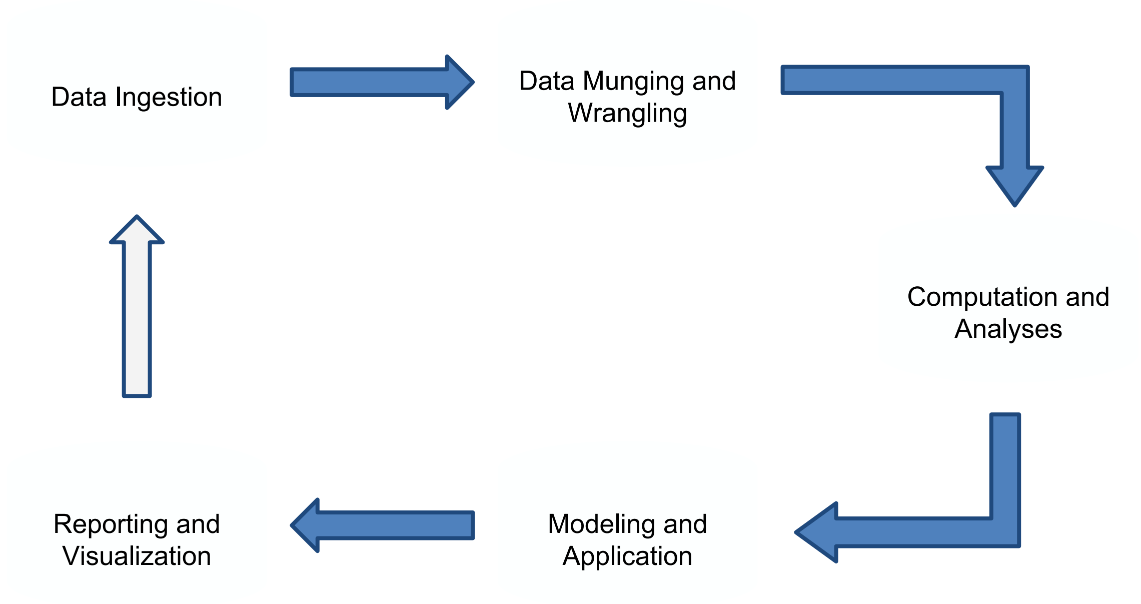
Communication: Commit Messages

git commit -m “try to be as helpful as possible”

(To your team and to future you)

Why?

Why do data scientists need version control?



Data Ingestion

Data Munging and Wrangling

Computation and Analyses

Modeling and Application

Reporting and Visualization

Where does version control fit into the data science pipeline?

Folder structure conventions on Github

README.md

COMMERCE DATA SERVICE

.gitignore

/fixtures

requirements.txt

Where to go from here?

Additional Tutorials

http://pcottle.github.io/learnGitBranching/

http://rogerdudler.github.io/git-guide/

http://www.tutorialspoint.com/git/

Resources

Git Desktop : https://desktop.github.com/

TortoiseGit: https://tortoisegit.org/

Git Cheat Sheet: https://training.github.com/kit/downloads/github-git-cheat-sheet.pdf

Getting Started: https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Basics: https://git-scm.com/book/en/v2/Git-Basics-Getting-a-Git-Repository

Branching: https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell

Github Setup: https://git-scm.com/book/en/v2/GitHub-Account-Setup-and-Configuration

Git Tools: https://git-scm.com/book/en/v2/Git-Tools-Revision-Selection

Git Commands: https://git-scm.com/book/en/v2/Git-Commands-Setup-and-Config