Welcome to General Assembly

WiFi GA Guest
Password yellowpencil

GENERAL ASSEMBLY

DATA SCIENCE

Lesson 2, Week 1

Elements of Data Science

What tasks do data scientists routinely perform, and what tools do they use?

Learning Objectives

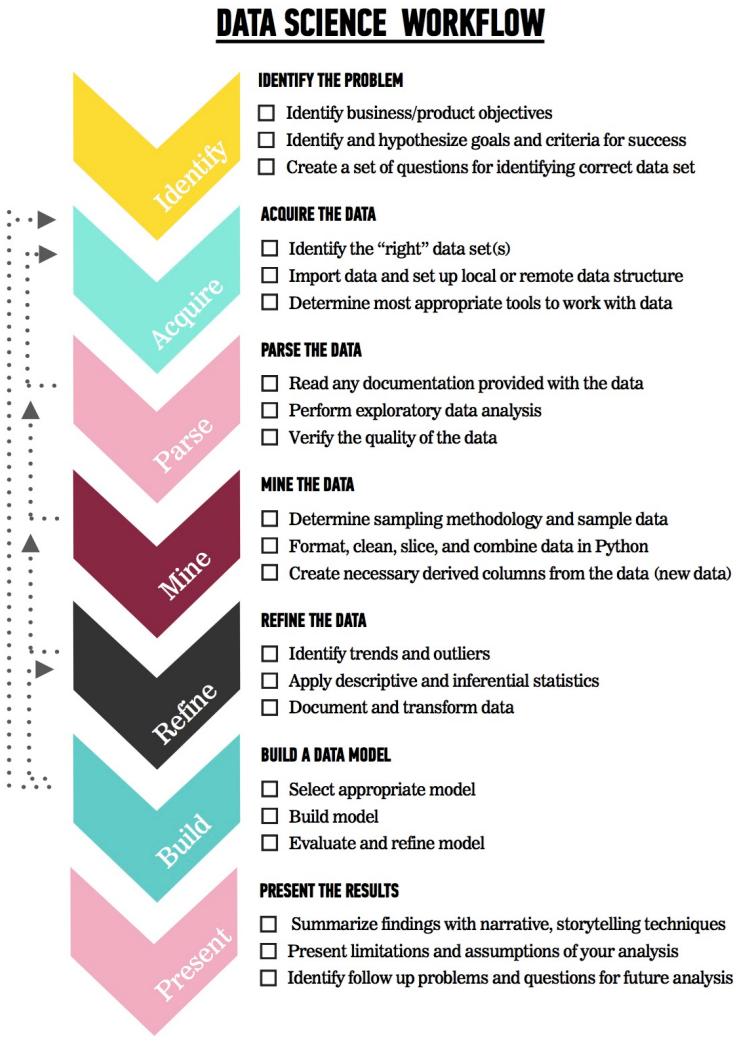
By the end of this lesson you should be able to ...

- Describe what data scientists create and how we get started on a problem
- Use the SMART scheme for formulating the objectives of a data science problem
- Describe the forms that data can take and where it may come from
- Use an iPython Notebook to manipulate a dataset
- Use Git to keep your local repo in sync with the course repo

RECAP:-
Data Science Workflow

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



Quick Quiz (5-mins)

1. Name one step in the data science workflow?

3. Which version of python are we using in this course?

5. What is the name of the online code storage platform we are using?

7. How do you run a code cell in jupyter notebook?

9. Is age a continuous or categorical variable?

Action	Unix Command		Action	Unix Command
Printing the working directory:				
Changing directory:				
Creating a new directory:				
Show contents of files in directory:				
Copy files				

Quick Quiz Answers

1. **Name one step in the data science workflow?**
(Identify the problem, Acquire the data, Parse the data, Mine the data, Refine the data, Build a data model, Present the results)

3. **Which version of python are we using in this course?** (3.5, 3.6)

5. **What is the name of the online code storage platform we are using?** (github)

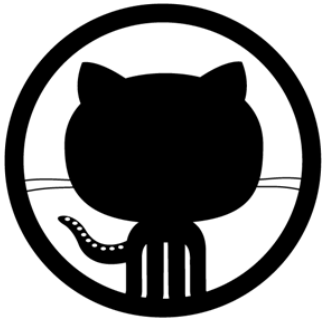
7. **How do you run a code cell in jupyter notebook?** (control+enter, play)

9. **Is age a continuous or categorical variable?** (Continuous)

Action	Unix Command		Action	Unix Command
Printing the working directory:	pwd		Move (rename) files:	mv
Changing directory:	cd		Go to “Home” directory:	cd ~
Creating a new directory:	mkdir			
Show contents of files in directory:	ls			
Copy files	cp			

Let’s get started...

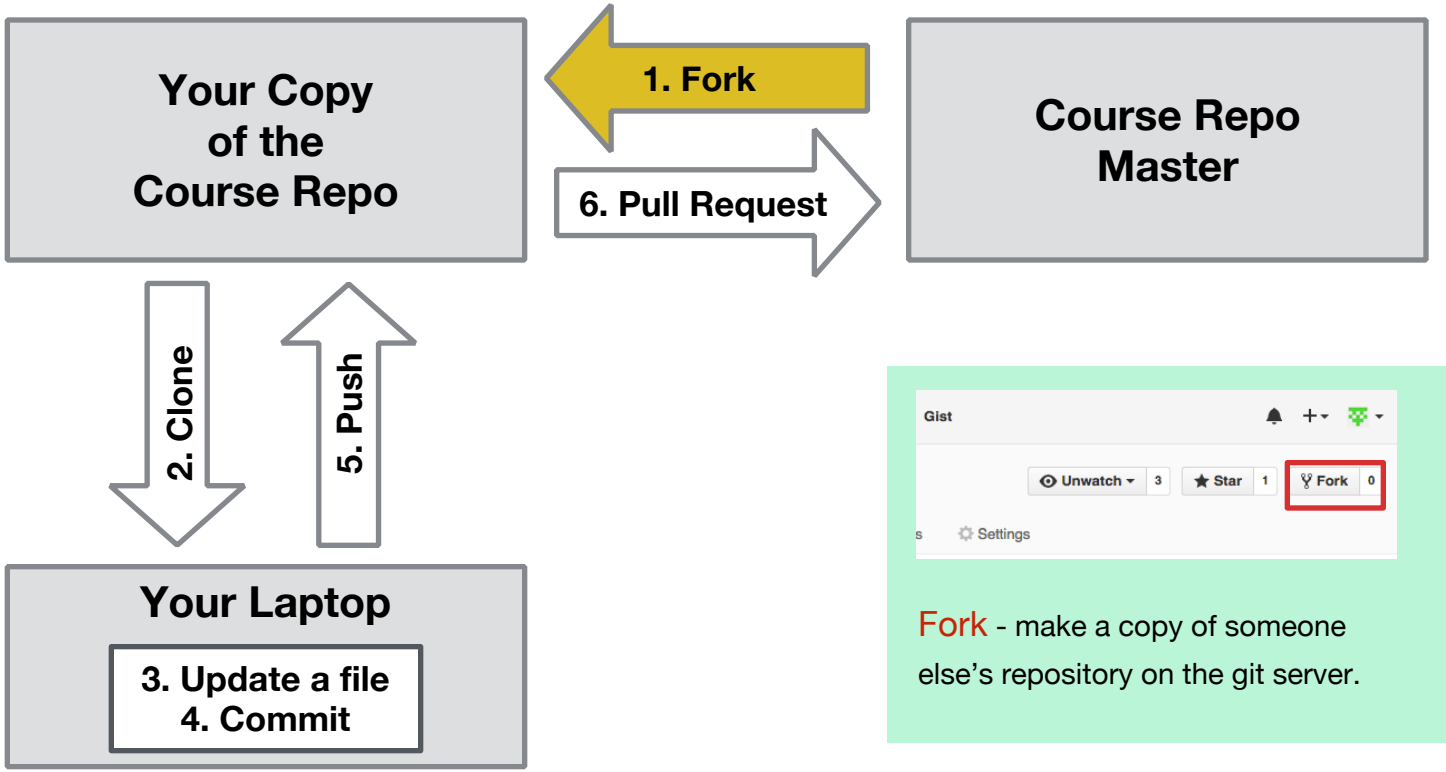
Review:
Git and GitHub



Git and GitHub

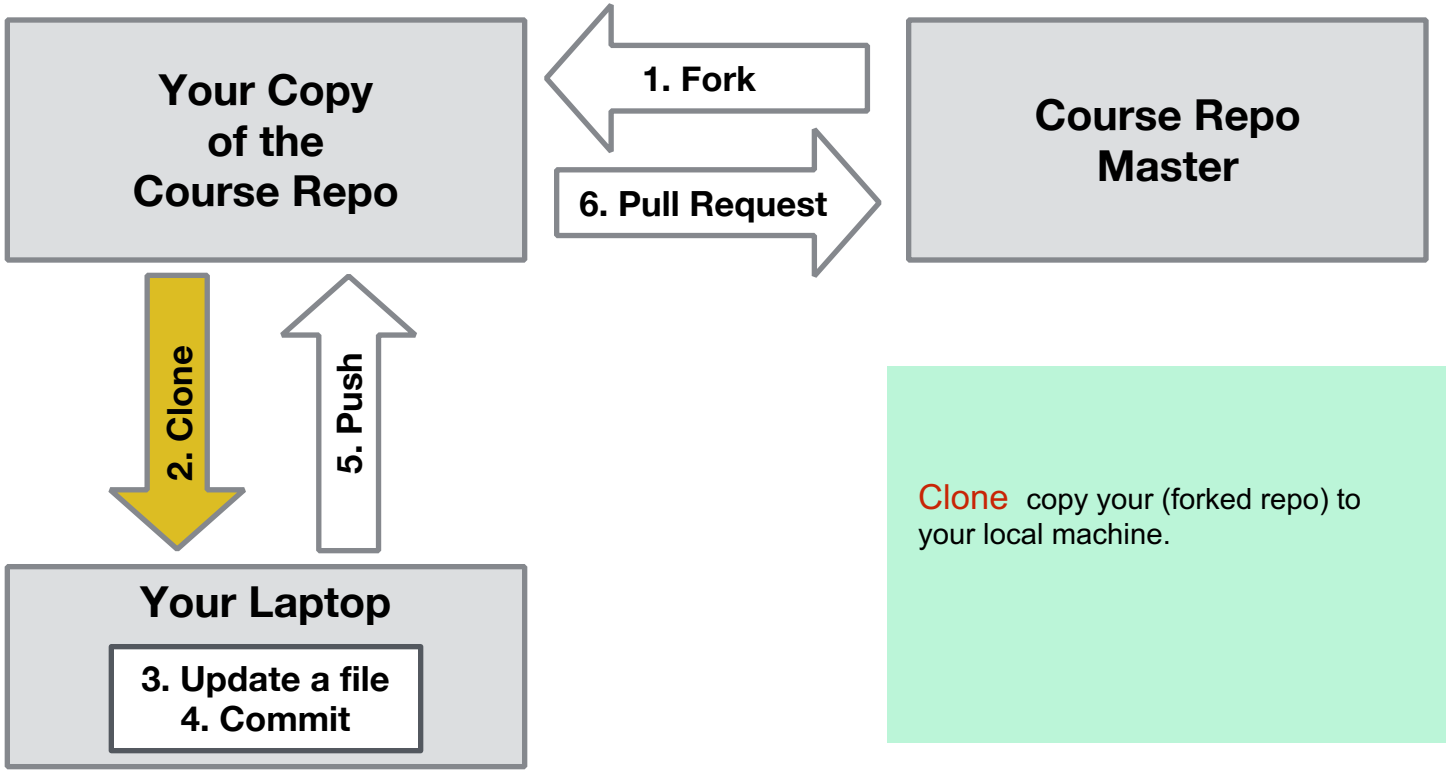
- Track files and file changes in a repository (“repo”)
  - version control
  - branching → divergent development
  - backup
  - collaboration (closed team or open source)
  - delivery
- Most widely used version control system, largest code host
- Runs from command line or a GUI (GitHub Desktop)
- Repos are online

Using Git



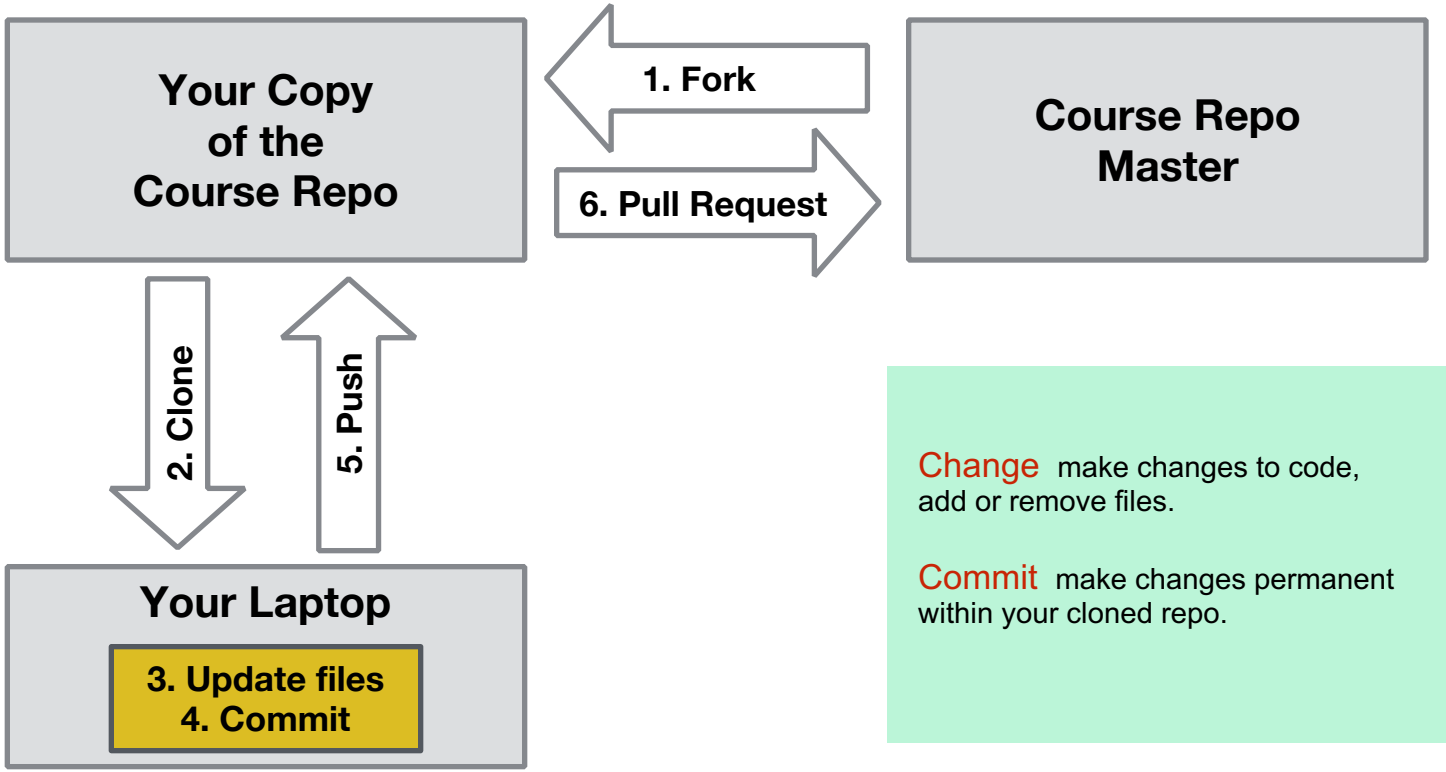
Fork - make a copy of someone else's repository on the git server.

Using Git



Clone copy your (forked repo) to your local machine.

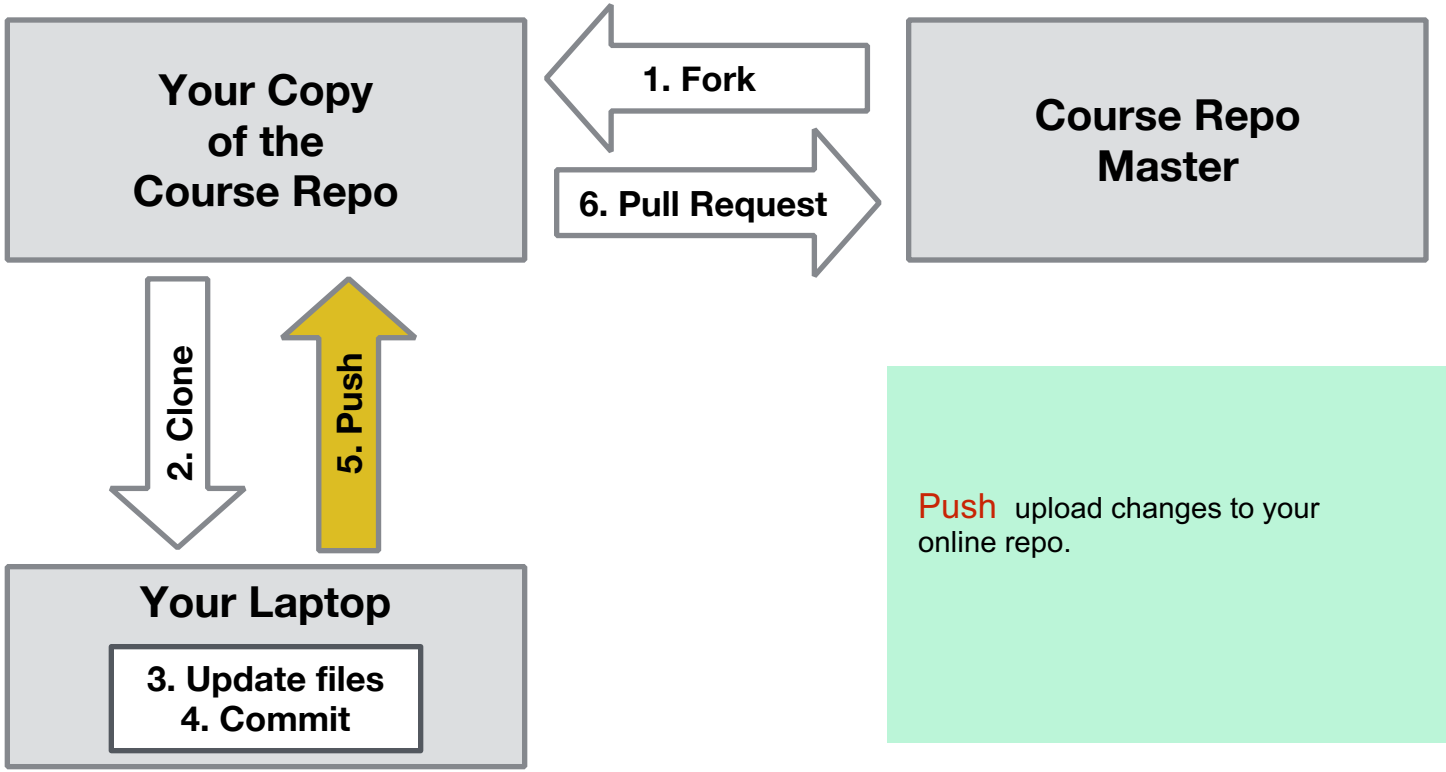
Using Git



Change make changes to code, add or remove files.

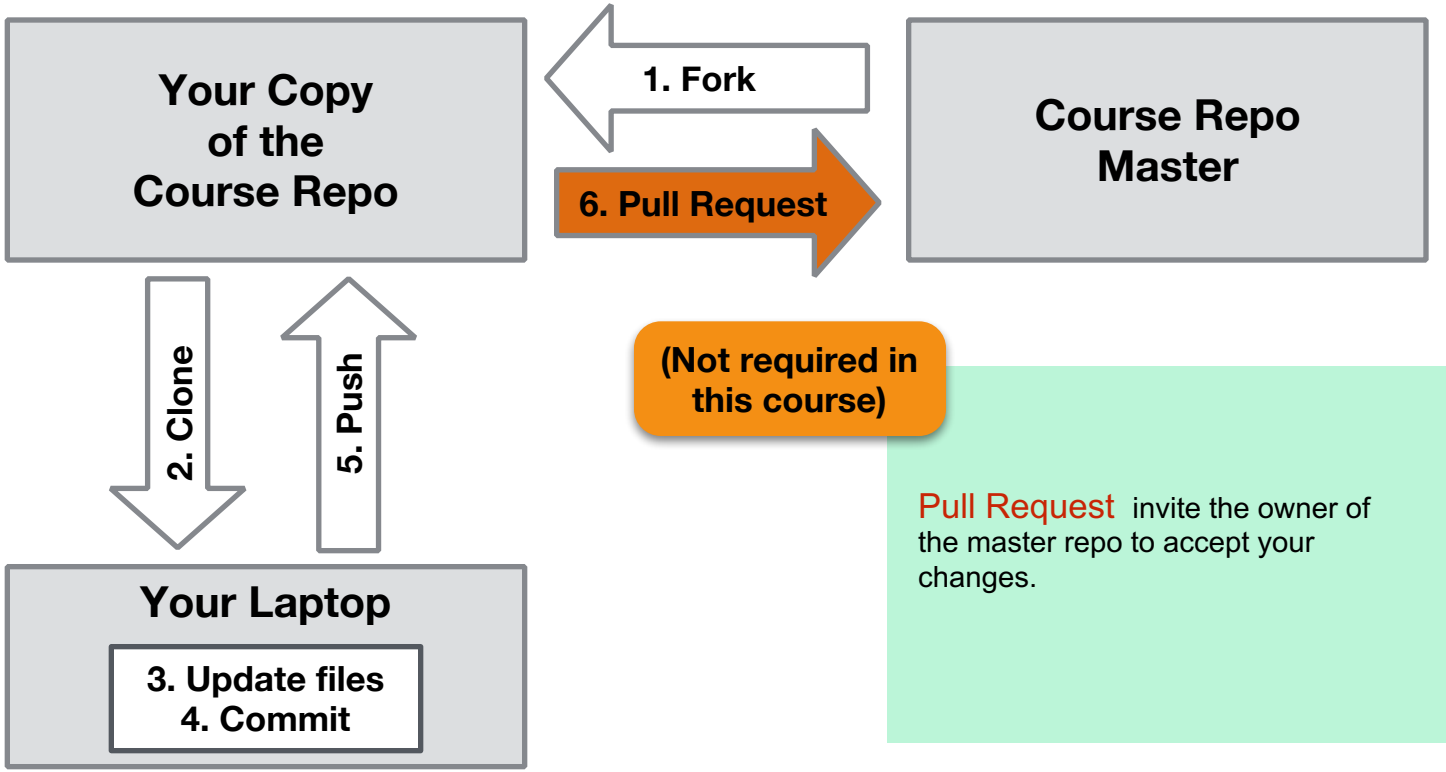
Commit make changes permanent within your cloned repo.

Using Git

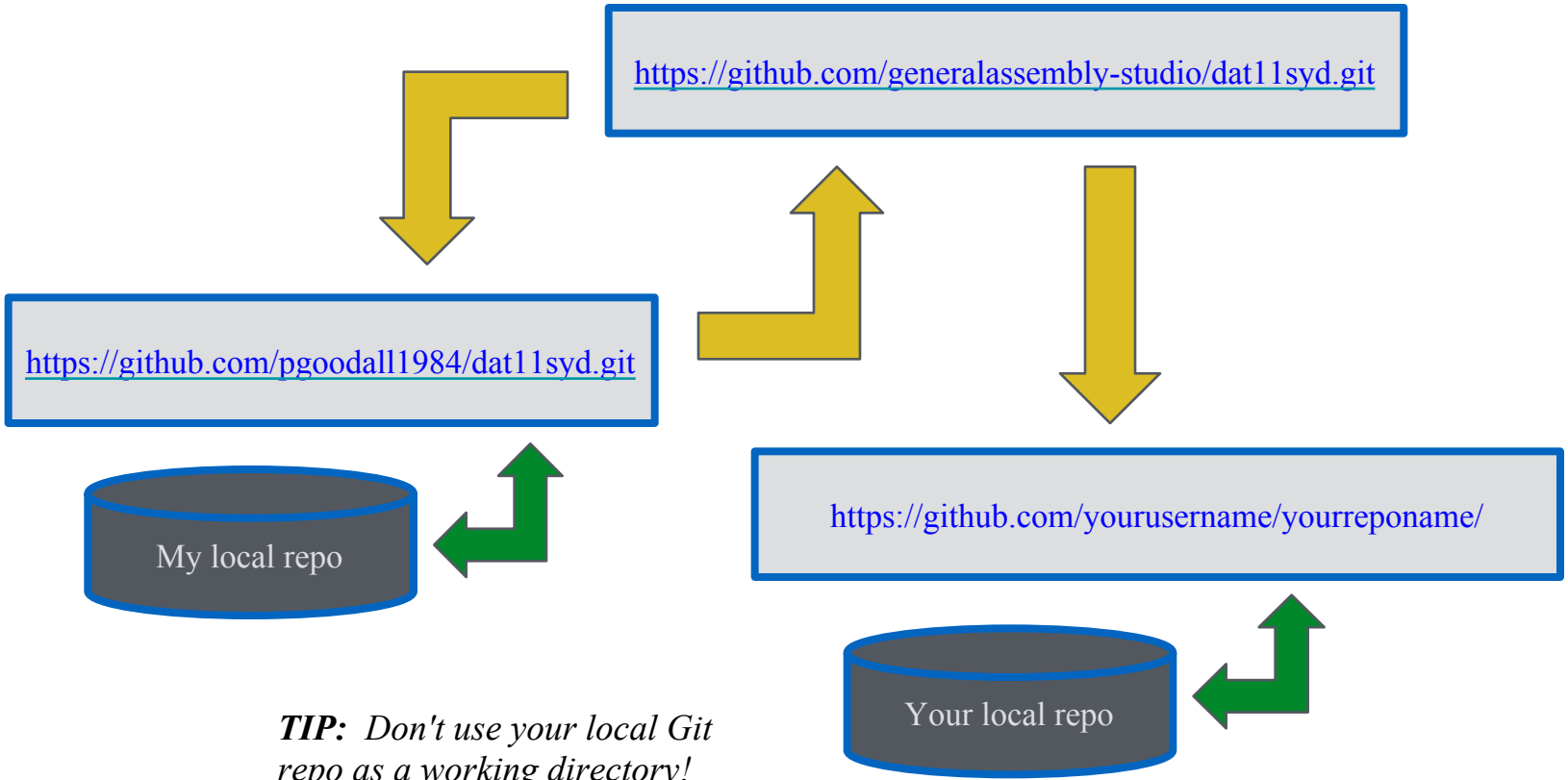


Push upload changes to your online repo.

Using Git

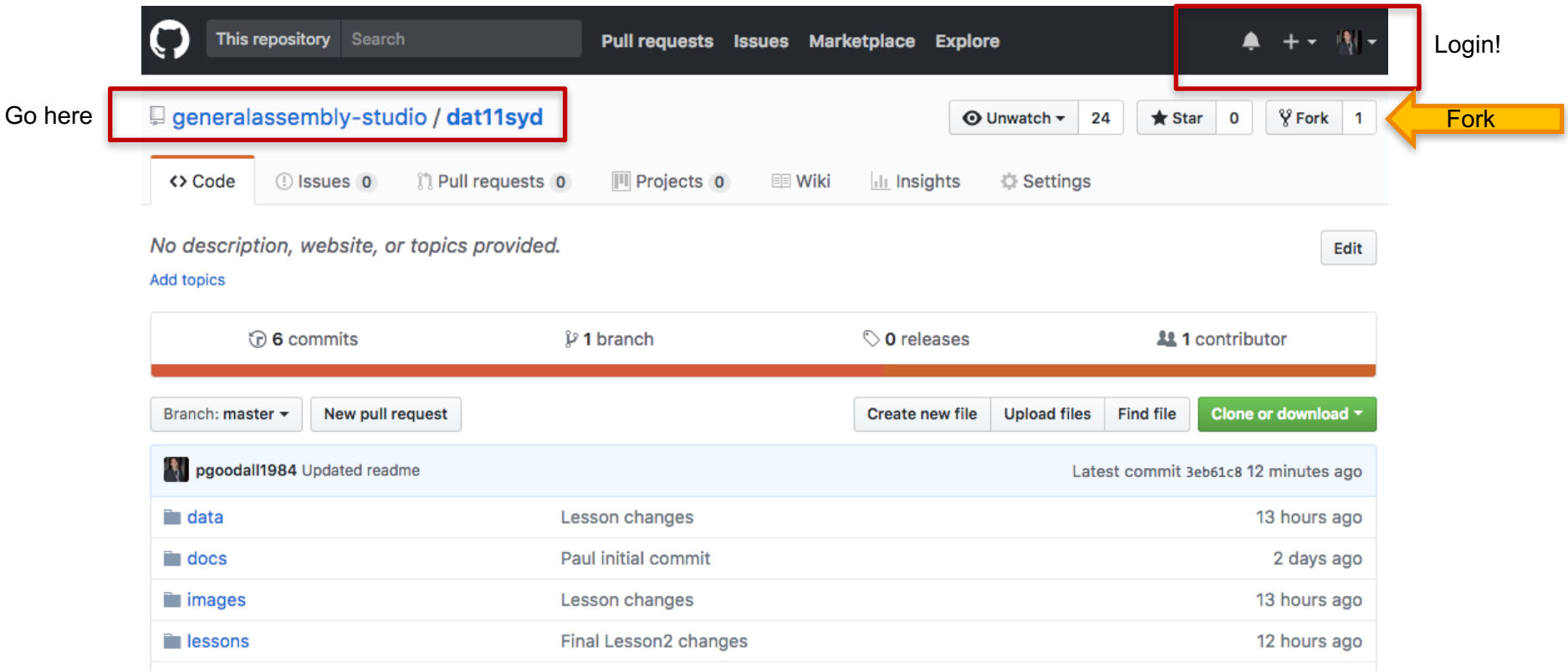

Your Copy of the Course Repo
1. Fork
Course Repo Master
6. Pull Request
2. Clone
5. Push
(Not required in this course)
Pull Request invite the owner of the master repo to accept your changes.
Your Laptop
3. Update files
4. Commit

Course Repos ('Fork & Pull' Model)



TIP: *Don't use your local Git repo as a working directory!*

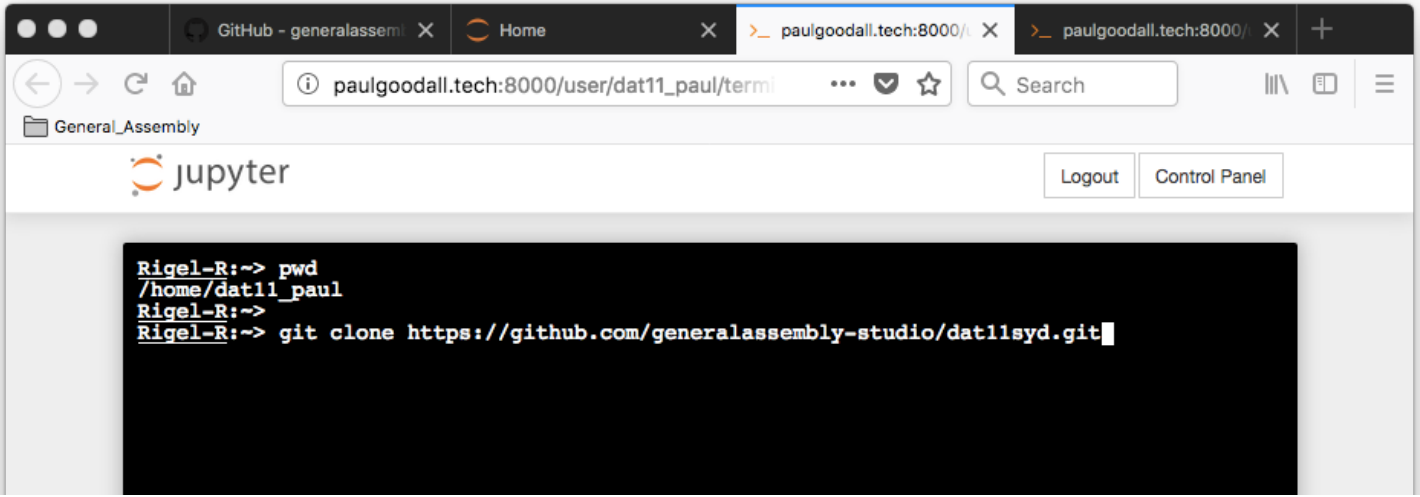
Course Repos ('Fork & Pull' Model)



Let’s get started...

Using the course Repo:

Navigate to the class Python Jupyter Hub:

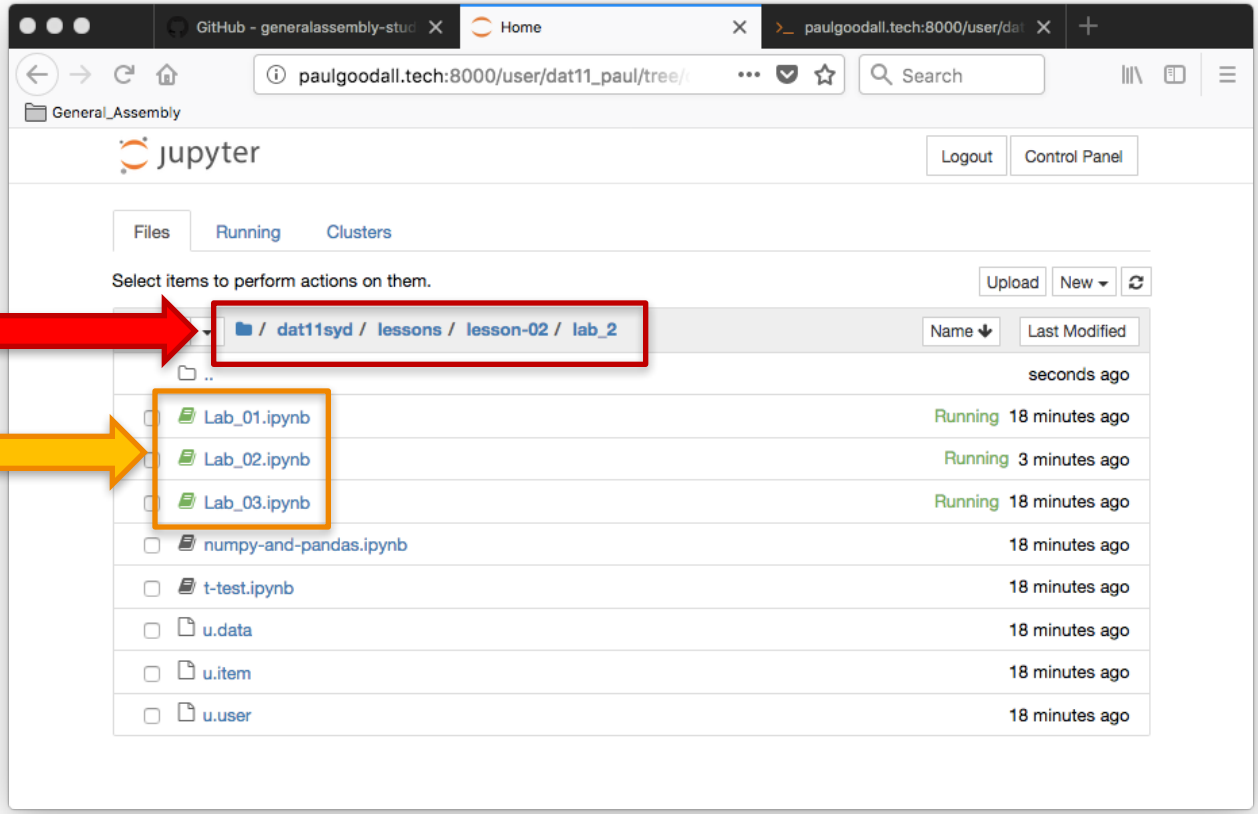


Python Exercises

- Navigate to the class Python Jupyter Hub: http://paulgoodall.tech:8000

Navigate to this directory

Do Lab1- Go! ☺



Jupyter Tips & Tricks

- special commands
  - ![cmd]
    - runs bash command [cmd]

- ‘magics’
  - %[cmd] "line magics"
    - [cmd]'s args come from same line
  - %%[cmd] "cellmagics"
    - [cmd]'s args come from entire cell
  - %lsmagic
    - lists all magics
  - %pwd
    - shows working dir
  - %ls -l
    - shows dir contents

%matplotlib inline

causes matplotlib plots to display within notebook

%%HTML

renders entire cell as HTML

use for embedding images, links, video, etc.

%%timeit

times execution of code cell

Now, back to Data Science ...

What does a data scientist do?

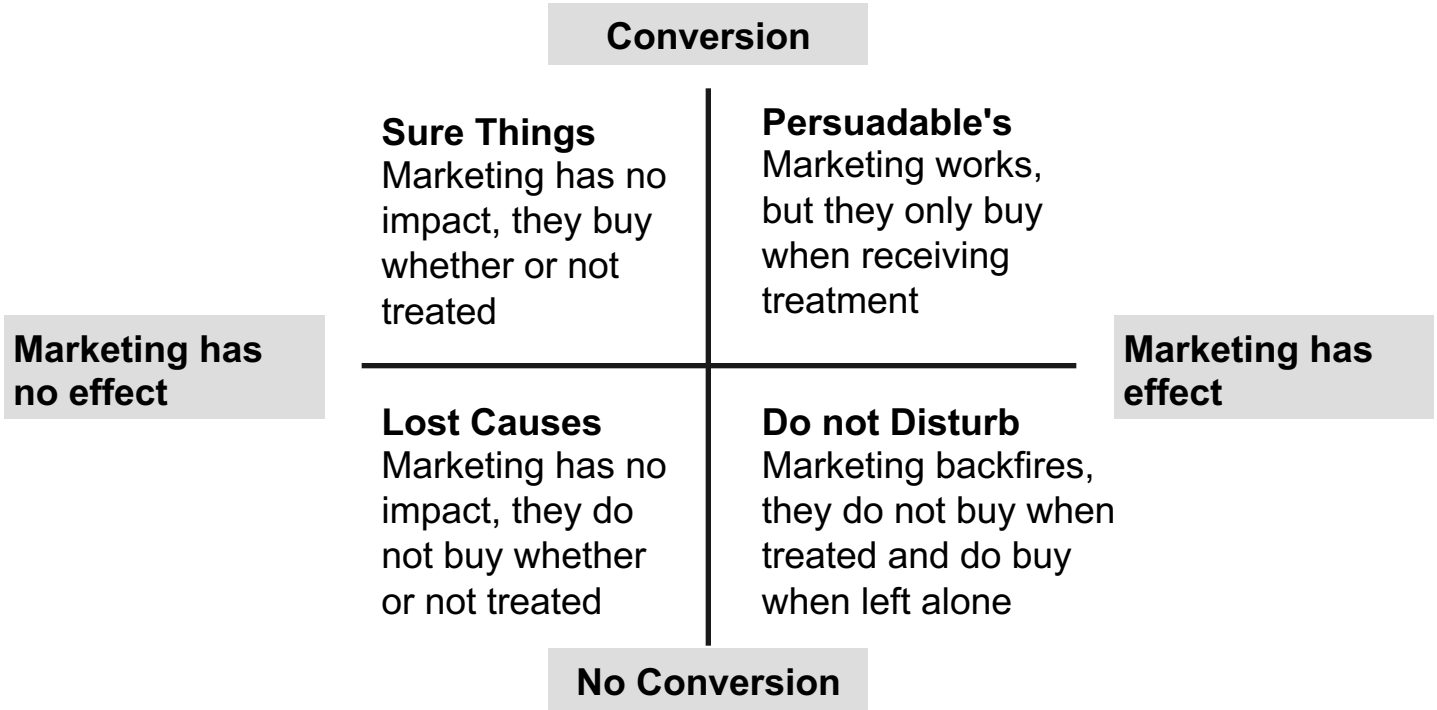
- Michael E Driscoll:
  - "Data scientists are better statisticians than most programmers and better programmers than most statisticians."
- DJ Patil:
  - "Those who use both data and science to create something new."
- Allen Nugent:
  - “Business analysis on steroids, tackling hard challenges with big rewards, using cutting-edge computing and statistical techniques, and guided by scientific discipline.”
  - “Create clever, useful things that exploit data invisibly.”
  - “Turn data into information, and information into insight.”

➢actionable information

➢data products

Examples of Actionable Information

- "Customers who do not place an order for 2 months and who are not subsequently contacted our Sales & Marketing are likely not to place any further orders (74% probability)."



Examples of Actionable Information

- "The new **stocking policy** would increase branch inventory by $1.2M ± $0.5M, averaged over the 3 branches I modelled, with no significant increase in availability."
- "The trial changes we made to the website were accompanied by a **500% increase in visit duration** (from 32 sec to over 3 minutes) and an **increase in average click-through depth** (from 1.8 pages to 4.7 pages). The catalog pages received 236% more hits, and sales revenue increased by 28%."

Data Science Workflow
Step 1:
Identify the Problem

Every Solution Begins with a Question

Whether we are seeking a one-off answer to a business question, a reusable information device, or a clever data product, we **begin with a set of questions that frame an analysis**.

By having a high quality question/aim you **set yourself up for a successful analysis**.

You also establish the basis for making your analysis reproducible.

A clearly articulated research question not only **helps other data scientists learn** from, and reproduce your work, but helps them expand on your work in the future.

"A problem well stated is half solved."
— Charles Kettering

"Judge a man by his questions rather than by his answers."
— Voltaire

“

I shall not today attempt further to define … [pornography]. But I know it when I see it.

United States Supreme Court Justice Potter Stewart, 1964

What is your question?

- A business challenge may be vague
  - “How can we grow our online market share?”
- Data science questions must be more focused
  - “Is our website engaging?”
  - “Are we presenting our products effectively to website visitors?”
  - “Are our prices competitive?”
  - “Is this market niche saturated?”

➢ *Even these examples are a bit vague, but we could easily break them down into quantitative questions*

What is your quest?
What is your favourite colour?
What is the average airspeed of an unladen sparrow?

Applying the Data Science Workflow:
What makes a good question?

Specific	The dataset and key variables are clearly defined.
Measurable	The type of analysis and major assumptions are articulated.
Attainable	The available data are amenable to the question and unlikely to be biased.
Reproducible	The analysis can be repeated by another person or at another time.
Time-bound	The time period and population to which the analysis pertains is clearly stated.

Knowledge Check

Does this question follow the SMART framework?

"Is there an association between number of passengers with carry-on luggage and delayed take-off time?"

Knowledge Check

How about the revised question?

"Is there an association between number of passengers with carry-on luggage and delayed take-off time?"

"Is there an association between the number of passengers (on JetBlue, Delta, and United domestic flights) with carry-on luggage and delayed take-off time in the data from flightstats.com between January 2015 and December 2015."

Let's get practical

- Work in groups of 3-4
- Dataset: *Survivors of the Titanic*
  - https://www.kaggle.com/c/titanic/data
  - dataset
  - data dictionary
- Formulate a question you could ask of this dataset, abiding by the SMART framework
- Elect a spokesman to present your proposal

Variable	Definition	Key
Survived	survival	0 = No, 1 = Yes
Pclass	ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	name	
Sex	sex	
Age	age in years	
Sibsp	# of siblings, spouses aboard	
Parch	# of parents, children aboard	
Ticket	ticket number	
Fare	passenger fare	
Cabin	cabin number	
Embarked	port of embarkation	

Example Answers:

"Using data from April 15, 1912, taken from the Titanic disaster, we will determine the association of gender and age (in years) with survival."

"Using data from April 15, 1912, taken from the Titanic disaster, we will test the hypothesis that women and children were preferentially given places in lifeboats by analysing relative survival rates of men, women, and children."

Questions?

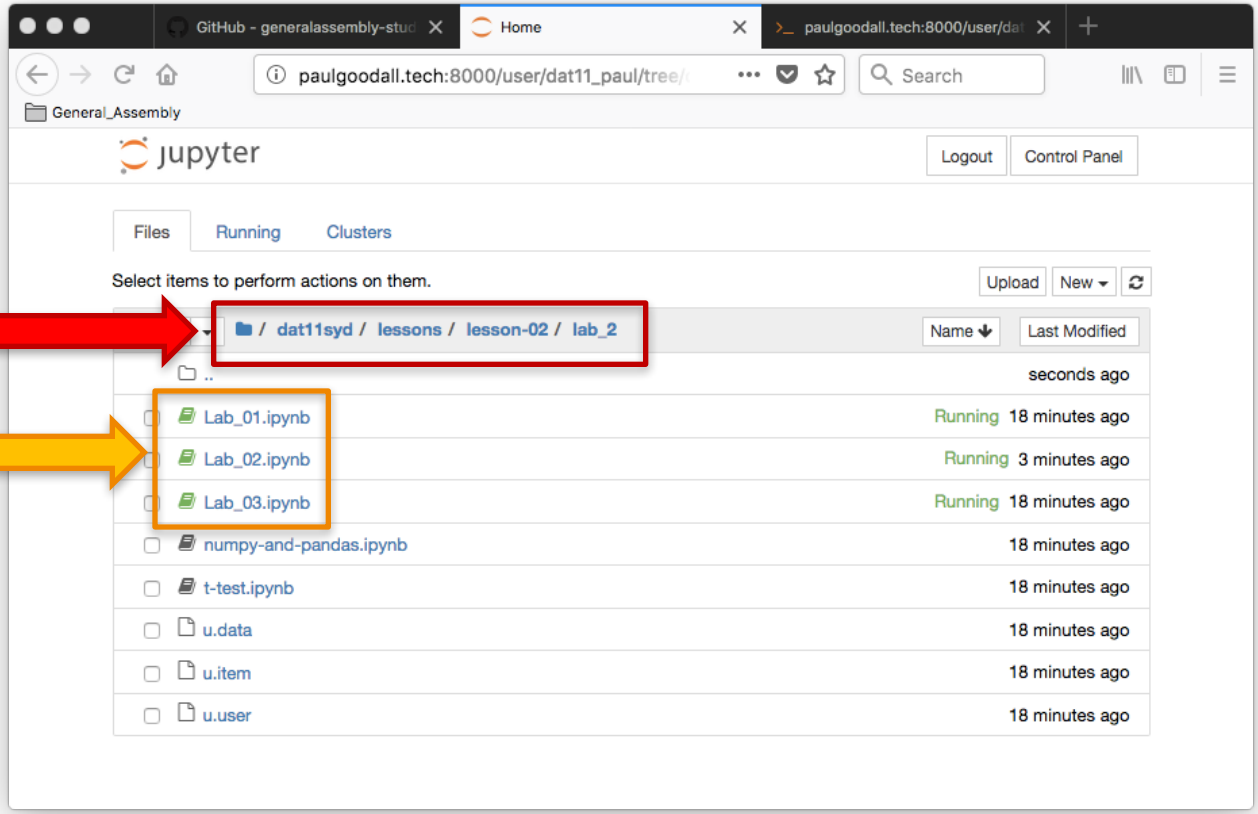
Let’s get started...

Python Exercises

- Navigate to the class Python Jupyter Hub: http://paulgoodall.tech:8000

Navigate to this directory

Do Lab2- Go! ☺



Data Science Workflow
Step 2:
Acquire the Data

Data Sources

- databases (SQL, JSON, etc.)
- electronic files (CSV, TXT, XML, JSON)
- Web pages (HTML, XML, Google Analytics)
- surveys
- published datasets
- social media (TXT, JSON)
- documents (PDF, JPEG, etc.)
- measurements
  - experiments, QA
- devices & monitors (ASCII, binary)
  - sensors, meters, loggers
  - trackers, wearables

Data Temporality

Cross-sectional ('static')

- treated as a snapshot in time
- causality is simultaneous

Longitudinal ('time series') *Yearly vs Monthly*

- treated as a series of snapshots with a temporal or serial dependence

Dynamic ('streaming')

- continuously accumulated or refreshed

Dataset Characteristics

- size
- completeness
- bias
- accuracy, precision
- periodicity
- stationarity
- variance, heteroskedasticity
- correlated variables
  - causation or covariation
- correlated samples
  - time series / Markov series
  - contaminated or prejudiced sampling

Practical example

Suppose we have a dataset comprised of patients' responses to two different therapies:

- drug A (the old drug, or 'control' treatment)
- drug B (the new drug, or 'test' treatment).

The null hypothesis H_0 assumes that there is no significant difference between the two. In statistical terms, this means that the two distributions we get from the 'A' data and the 'B' data represent two sample sets from the *same* 'population'
(i.e. a common underlying distribution).

What does that prove?!

To prove that ‘B’ is better than ‘A’ we must, we need to make a convincing case that the true mean value of the ‘B’ data is different from that of ‘A’ (and in a favourable sense):

- we are trying to show that we can safely reject H_0

Thus, we must demonstrate that the probability that the ‘A’ and ’B’ distributions are samples of the same population *is less than some threshold* (usually $P < 0.05$).

- Note that we can’t really *prove* that ‘B’ is better than ‘A’, because there is always a finite chance that one or both of our sample sets is highly unrepresentative of its respective population.

How Much Data is Enough?

Statistical power:

The probability that we will correctly reject H_0

Below are some diagrams showing real differences (here, little vertical lines represent the [unknown] true means for the underlying populations).
For each pair, judge which one would be most easily detected by a study:

1. Size of difference: A B

Intuitively, a study would find it easier to detect the difference in 'A'.
So, the size of the true difference influences power

2. Distribution: C D

Intuitively, a study would find it easier to detect the difference in 'D',
so the variance of the distribution of results also influences power

3. Sample size: E F

The bars show the 95% confidence intervals. Intuitively, a study would find it easier to detect the difference in 'F'. For a given SD, the CI is affected by sample size [the formula for 95%CI = ±1.96 (SD / √n)]. Hence, sample size also influences power

Conclusion: The power of a study is influenced by the magnitude of the true difference, the SD of the population means, and the sample size.

www.med.uottawa.ca

Questions?

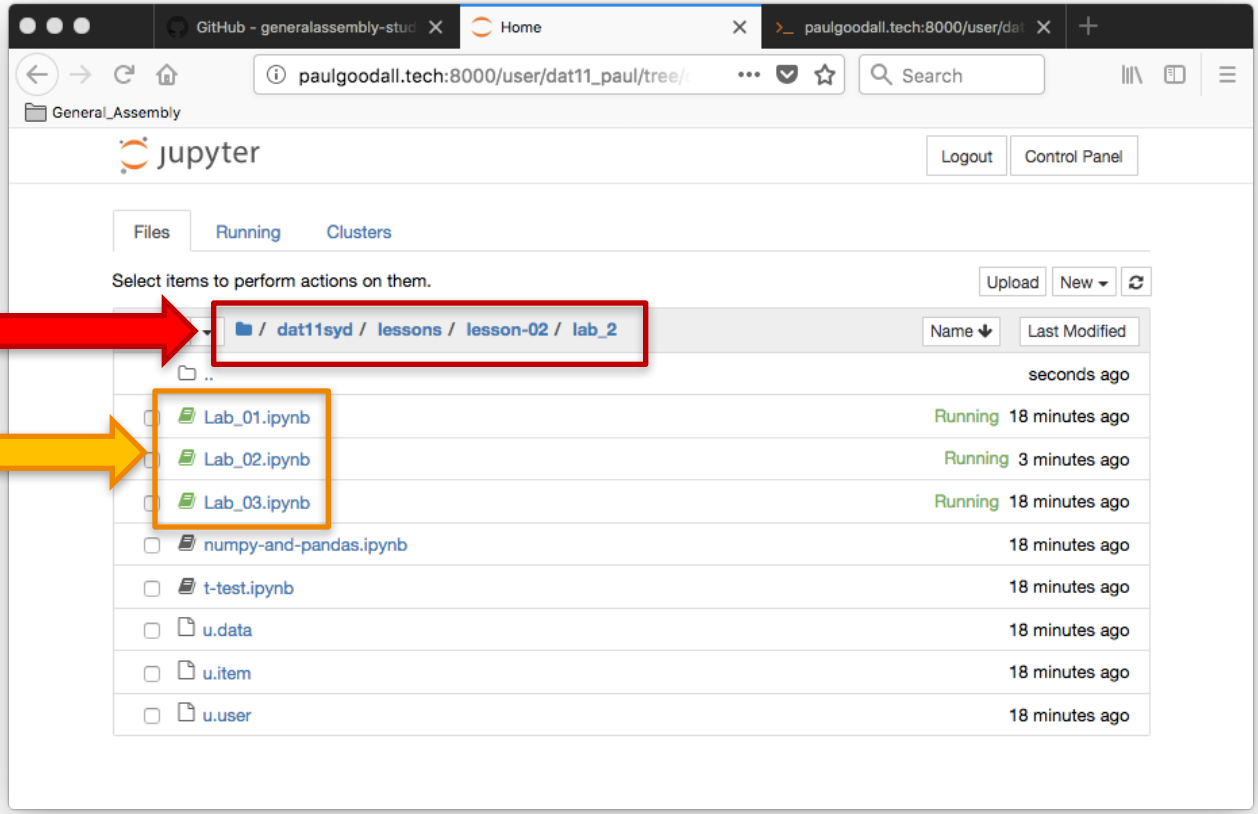
Let’s get started...

Python Exercises

- Navigate to the class Python Jupyter Hub: http://paulgoodall.tech:8000

Navigate to this directory

Do Lab3- Go! ☺



Data Science Workflow
Step 3:
Parse the Data

Data Types

Numeric	Text	Other
integer • signed, unsigned • 2, 4 B	character • 1 B (ASCII) • 2 B (unicode)	Boolean • true, false Binary • 2^n
floating-point (‘float’) • 4, 8 B • double = 2 x float	string • character array • 0-based *or* 1-based • null-terminated *or* length-encoded • usually immutable in OOP	unassigned • null • NA undefined • NA • +, − infinity
complex • 2 x double (real, imaginary)	document • key-value pairs (JSON strings)	BLOB • images, video • signals

Numeric Data Representation

Name	Symbol	Limits	Where
binary		base 2 range = 2	ubiquitous
octal	\123	base 8 range = 256	early digital computers
decimal		base 10	ubiquitous
12-bit		base 10 range = 4096	microcontrollers, analog-to-digital converters
hexadecimal	\x123, 123h	base 16 range = 65,536	modern digital computers

Data Types in Python and NumPy

Type	Python	Numpy	Usage
byte byte array	b‘any string’ bytearray()		• immutable • mutable
integer	int()	• 11 types	• signed, unsigned • 8, 16, 32, 64 bits, unlimited
floating-point	float()	• 3 types	• 16, 32, 64 bits
complex	complex()	• 2 types	• 64, 128 bits
unassigned	None		• object • myVar is not None
missing	nan	isnull(), notnull(), isnan()	• float, object

Data Forensics

Inputs

Features
Independent Variables
Predictors

A *predictor* is a *feature* that is useful in modelling the *response*. Specifically, its inclusion enables a *model* to account for more of the *variance* of the response.

Outputs

Outcomes
Dependent Variables
Responses

A *covariate* is a variable that is possibly predictive of the *response*. It could also represent an *interacting* variable.

A *confounding* variable is one which influences the response but has not been measured (i.e. it introduces bias).

Tidy Data

The end goal of the cleaning data process:

- each variable should be in one column
- each observation should comprise one row
- each type of observational unit should form one table
- key columns for linking multiple tables
- top row contains (sensible) variable names
- in general, save data as one file per table

Codd's 3rd normal form

➢search: “hadley wickham's tidy data paper”

Is this Tidy Data?

	treatment.A	treatment.B
patient.1	-	2
patient.2	16	11
patient.3	3	1

How about this?

Patient	Treatment	NumEvents
1	A	0
1	B	2
2	A	16
2	B	11
3	A	3
3	B	1

ASSUMPTION: "-" means zero

- could mean 'missing'!

Review

Review: Key Topics and Takeaways

- Data Science Workflow Step 1: _____
  - ?

Review: Key Topics and Takeaways

- Data Science Workflow Step 1: **Identify**
  - SMART

Review: Key Topics and Takeaways

- Data Science Workflow Step 2: _____
  - ?
  - ?
  - ?
  - ?

Review: Key Topics and Takeaways

- Data Science Workflow Step 2: **Acquire**
  - data sources
  - temporality
  - dataset characteristics
  - statistical significance & the null hypothesis

Review: Key Topics and Takeaways

- Data Science Workflow Step 3: _____
  - ?
  - ?
  - ?
  - ?

Review: Key Topics and Takeaways

- Data Science Workflow Step 3: **Parse**
  - data types
  - data forensics
  - tidy data
  - NumPy, Pandas

Q&A

Data Science Homework 1

DAT11 | Lesson 2 | Homework 1

Investigating Data

Instructions for the homework - time required: ~2 hours

- 1. Browse through the list of Data Science problems on Kaggle: https://www.kaggle.com/datasets
- 2. Choose 3 data sets that interest you,
- 3. Download the data and read it into Python Notebook on Jupyter Hub
- 4. Have a look at the datasets you have downloaded, and consider:

a. What is the data for?,
b. What are the data-types? (Numerics, Characteristics, etc?)
c. What alternative uses could you think of for the dataset?

- 5. Derive some simple, interesting facts from each dataset to report back to the class (plot a trend, or determine a relationship)

For the next class (Monday 26th Feb), please prepare:

```
* 1 single slide about your investigations - plan to keep your presentation to <= 5mins *
```

Additional Resources

- Statistics tutorials:
  - http://stattrek.com/tutorials/statistics-tutorial.aspx
- NumPy reference:
  - https://docs.scipy.org/doc/numpy-1.13.0/reference/
- GitHub:
  - https://guides.github.com/introduction/flow/

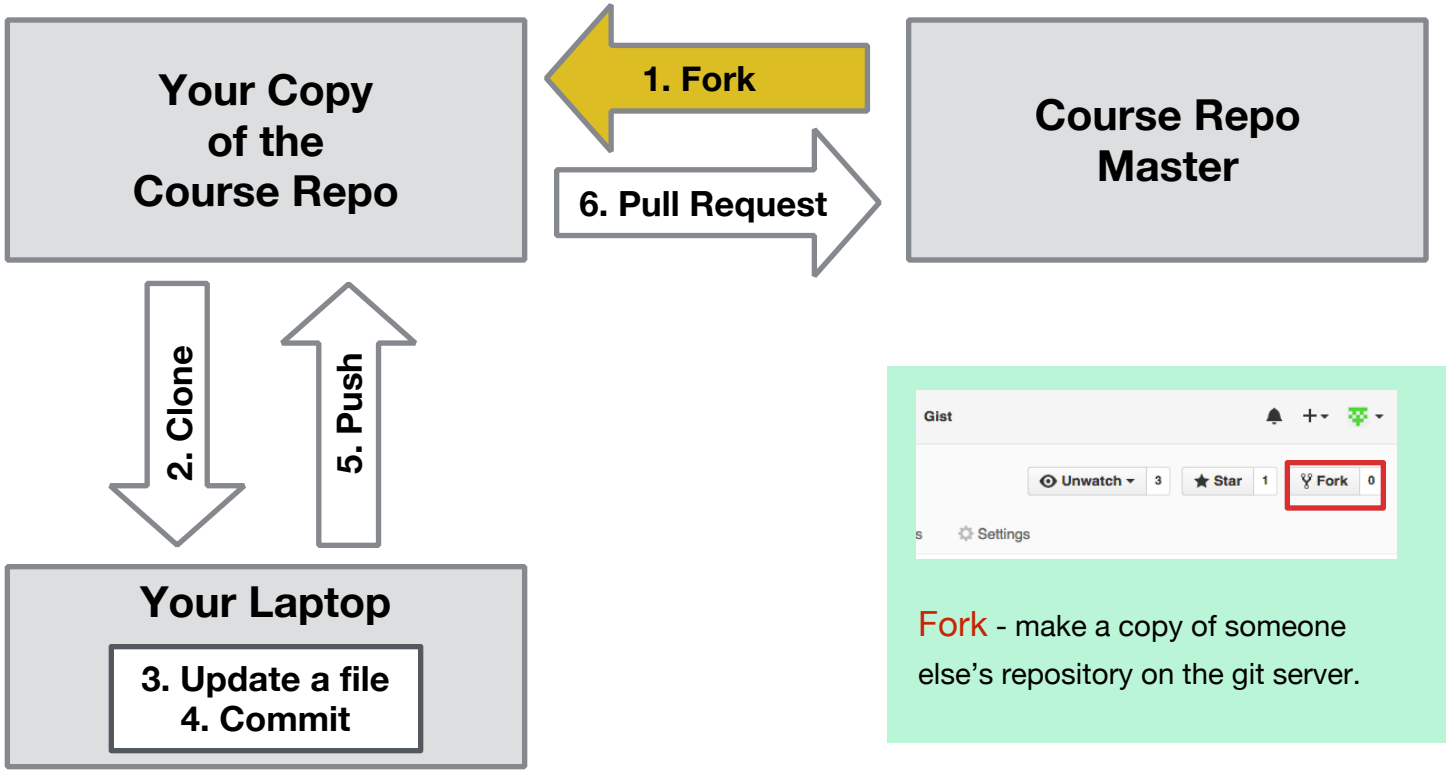
Please don’t forget the survey!

Exit Tickets:

https://docs.google.com/forms/d/e/1FAIpQLSdaBdMebhoXQpac2edDbIkHD-78HrTpBk3VKGH6lcDMVorSIQ/viewform

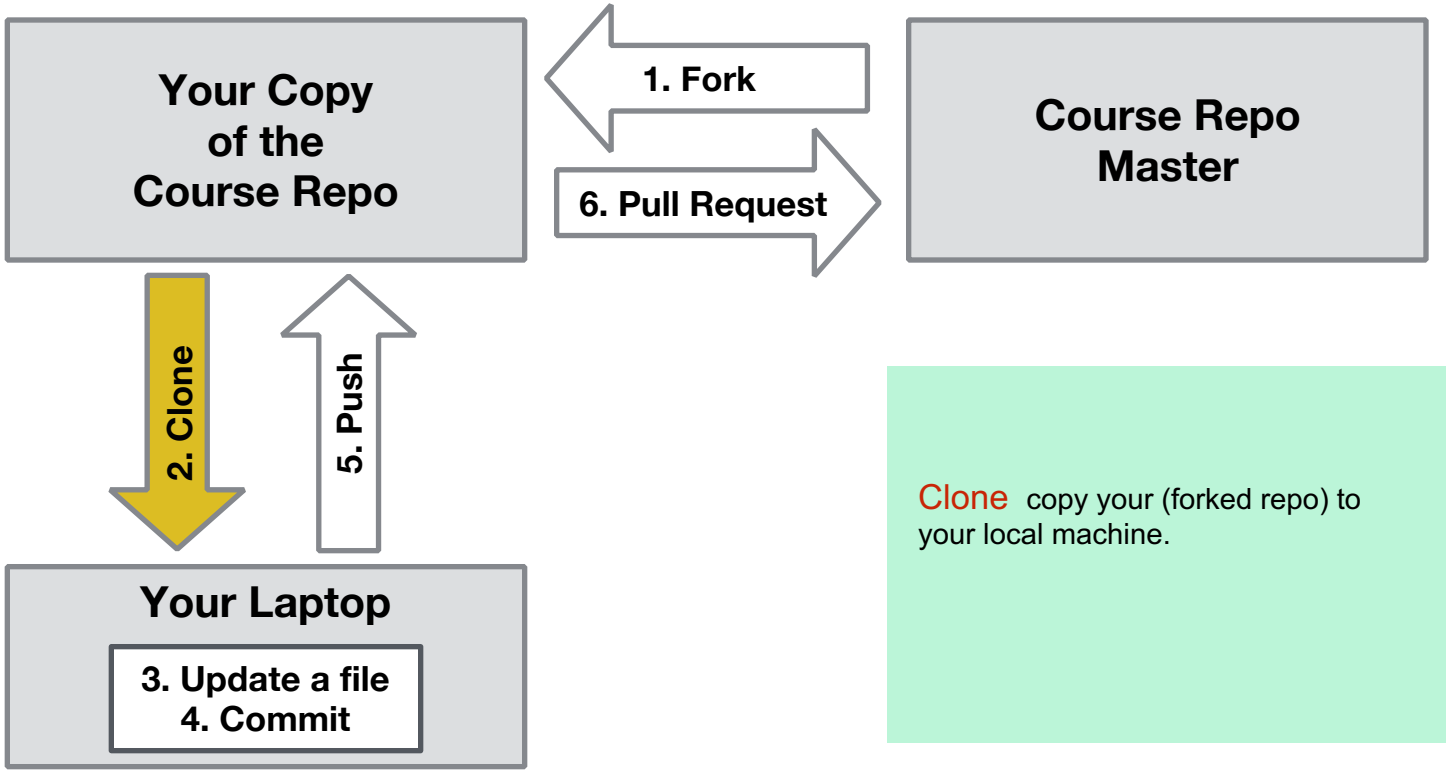
Thank You!

Using Git



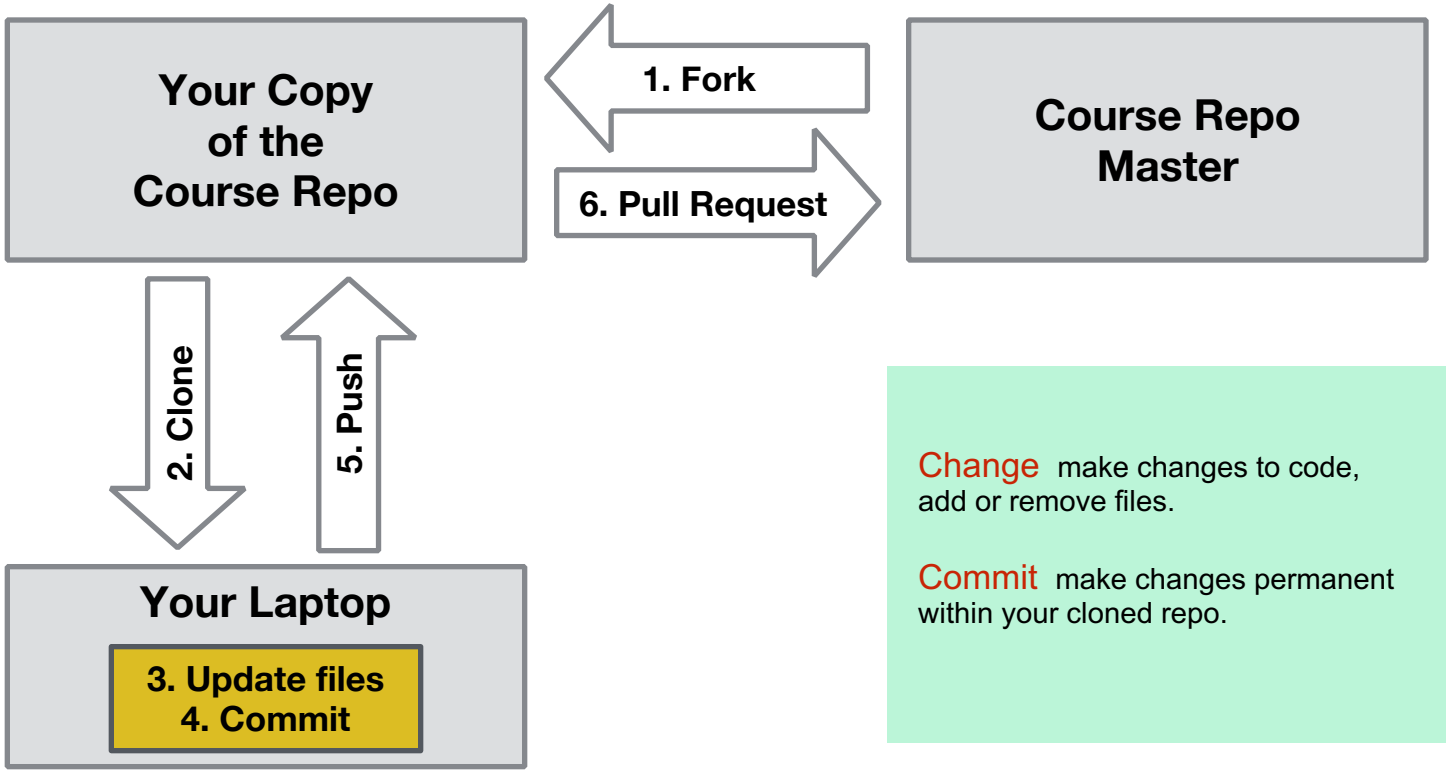
Fork - make a copy of someone else's repository on the git server.

Using Git



Clone copy your (forked repo) to your local machine.

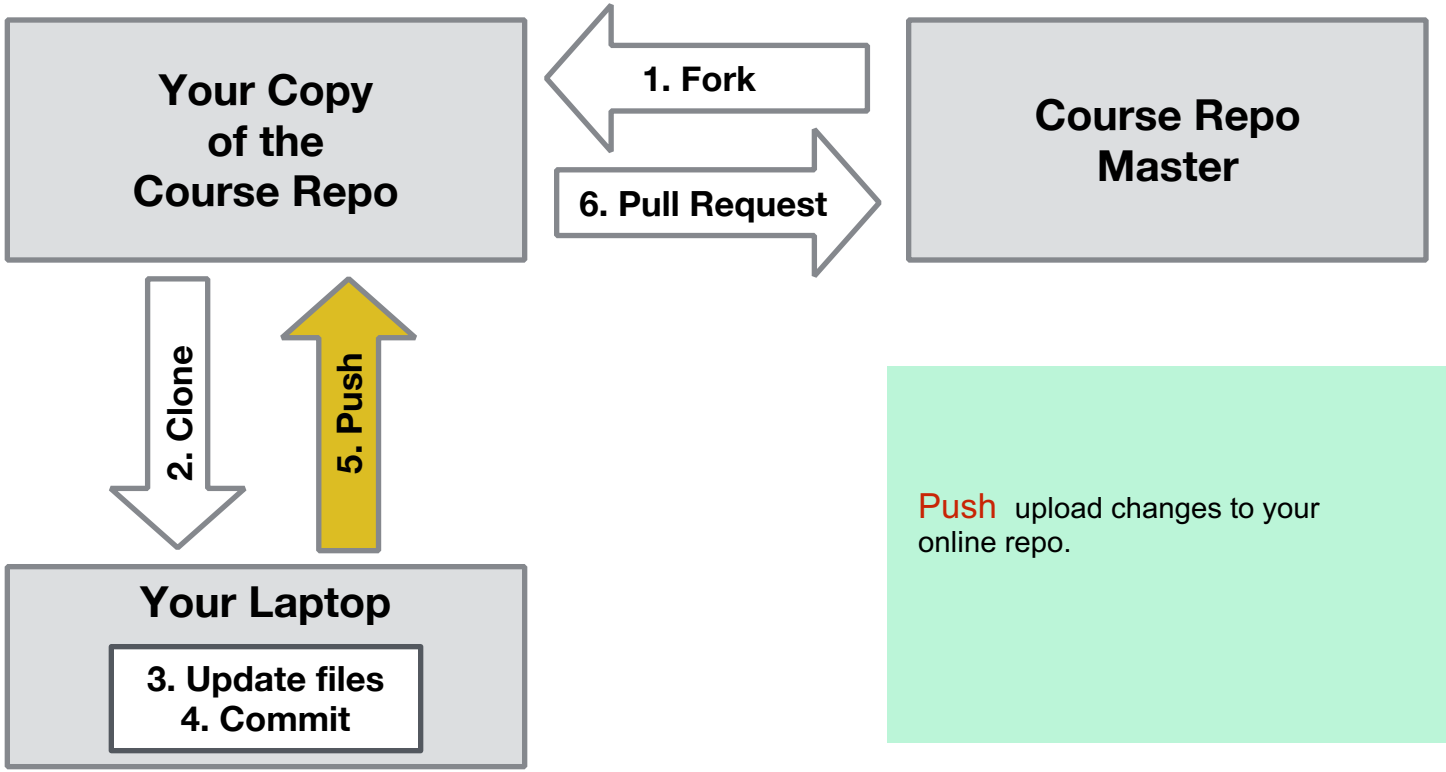
Using Git



Change make changes to code, add or remove files.

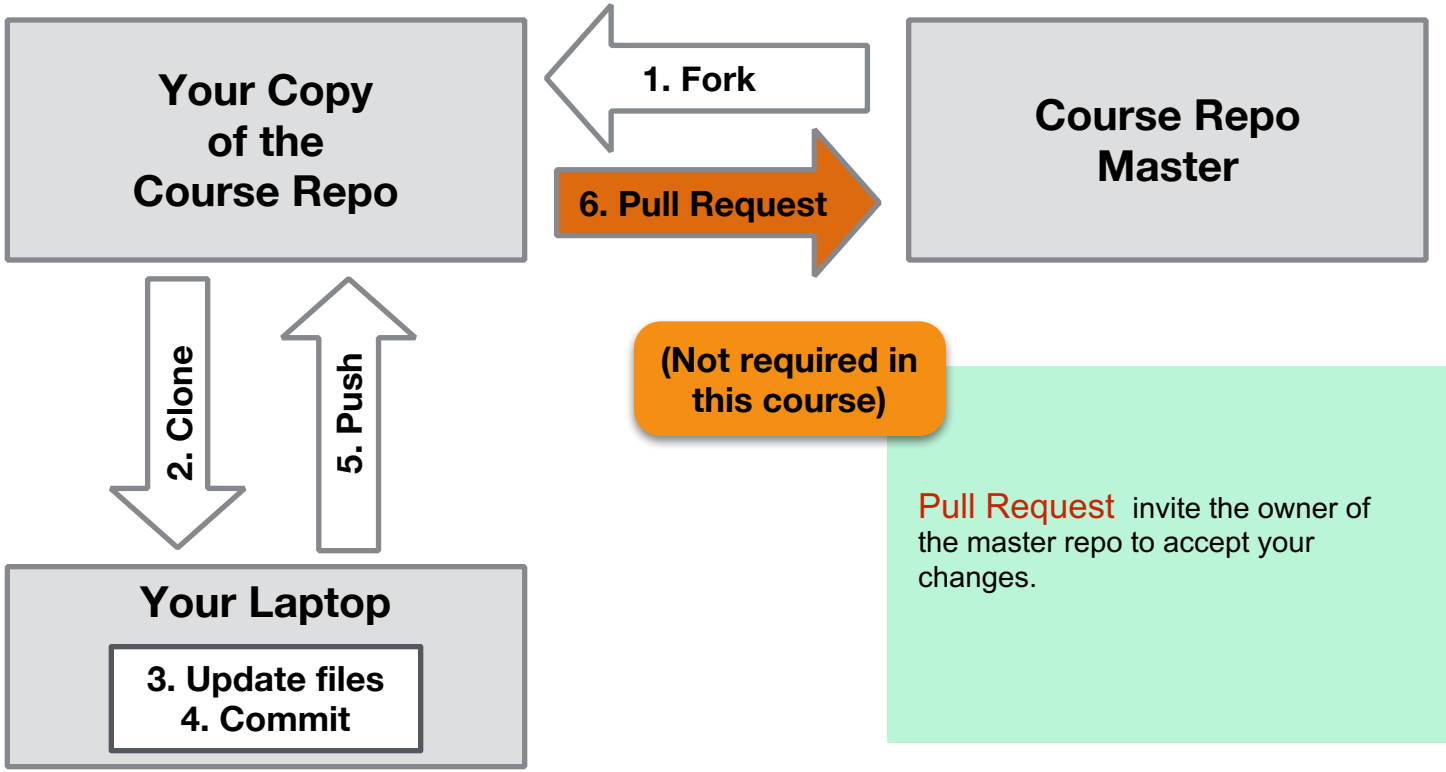
Commit make changes permanent within your cloned repo.

Using Git



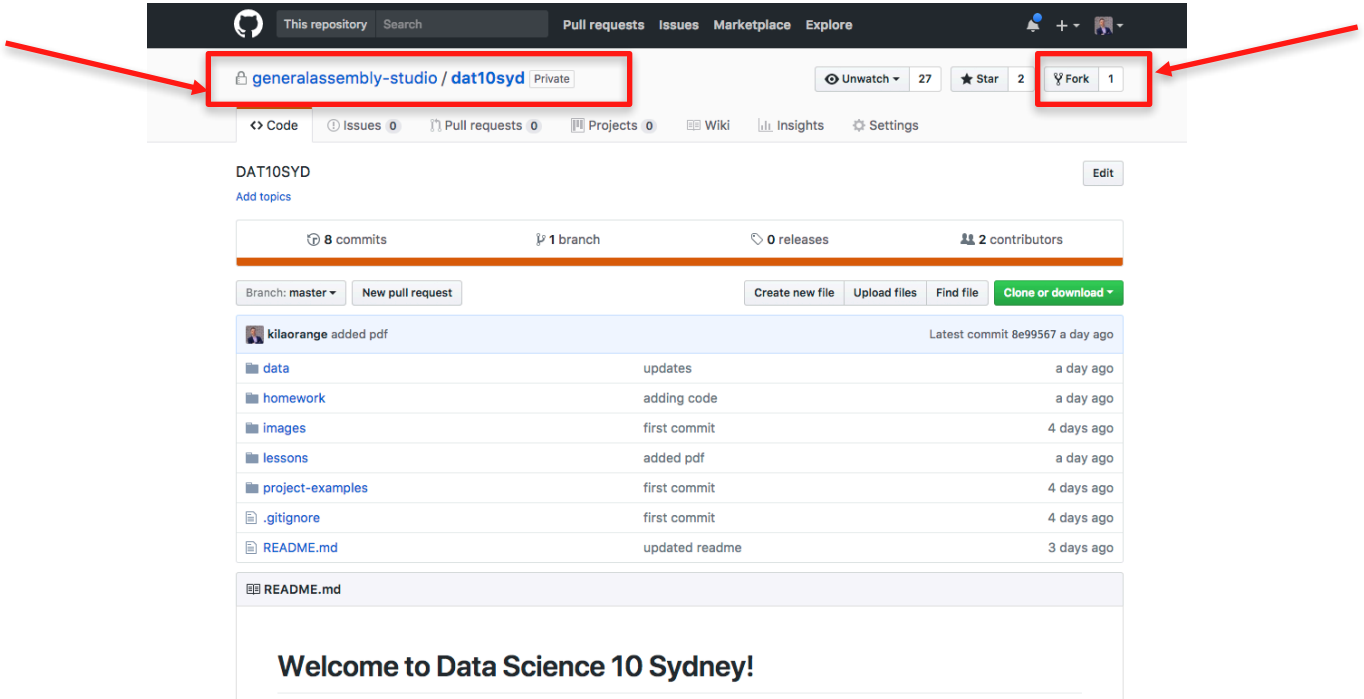
Push upload changes to your online repo.

Using Git


Your Copy of the Course Repo
1. Fork
Course Repo Master
6. Pull Request
2. Clone
5. Push
(Not required in this course)
Pull Request invite the owner of the master repo to accept your changes.
Your Laptop
3. Update files
4. Commit

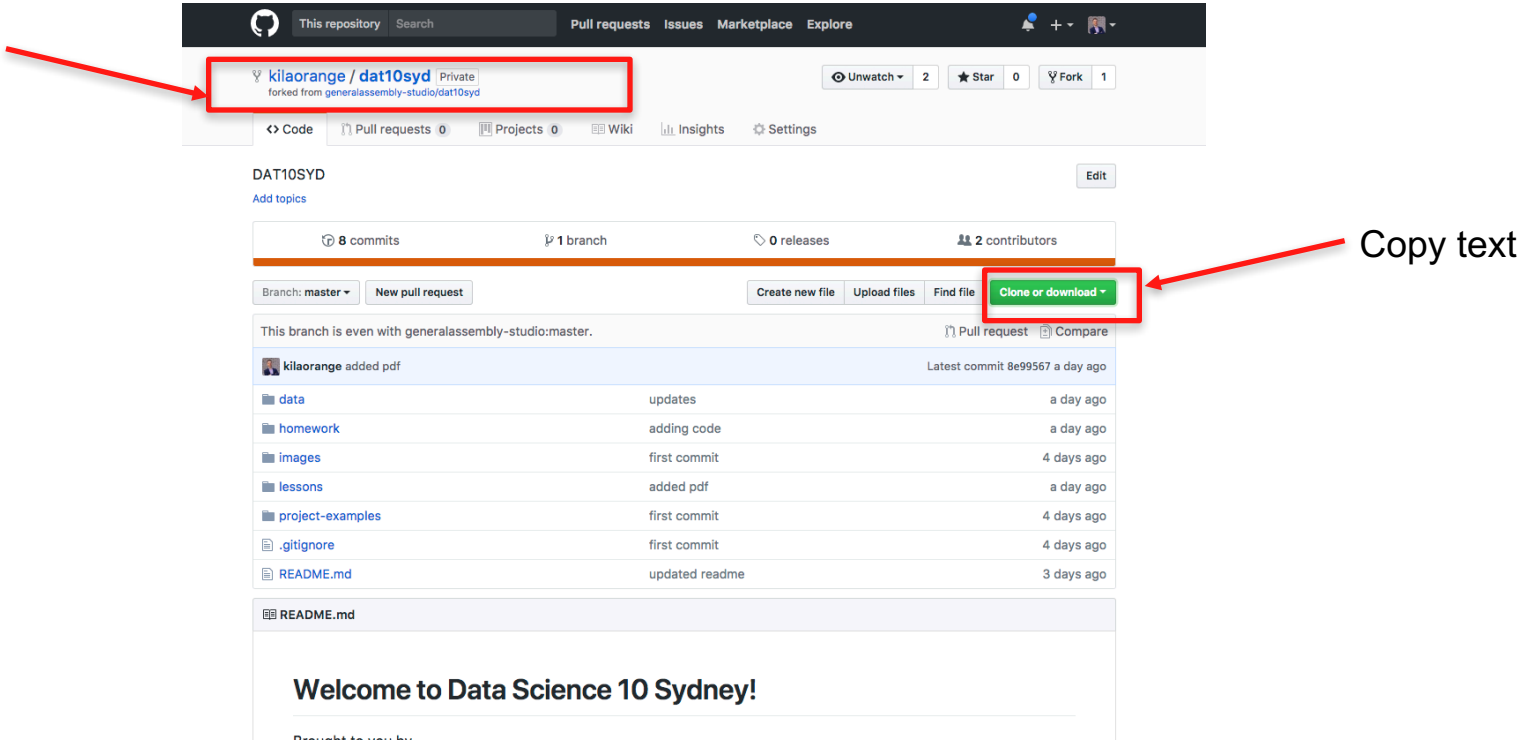
Preparing and Using Your Local Git

1. Improve on Tuesday night: delete Repo we cloned on Tuesday
2. Fork the GA Repo https://github.com/generalassembly-studio/dat10syd.git

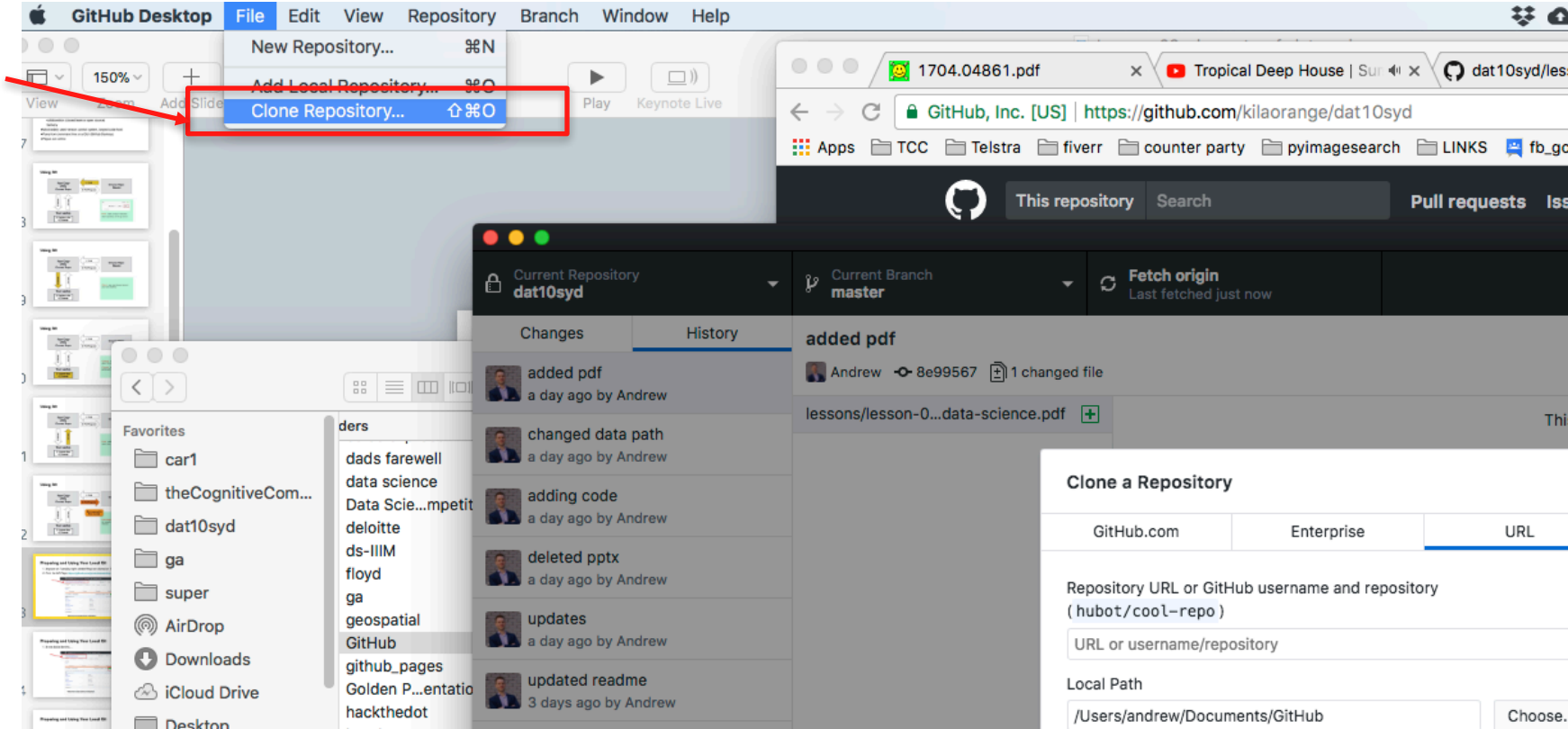


Preparing and Using Your Local Git

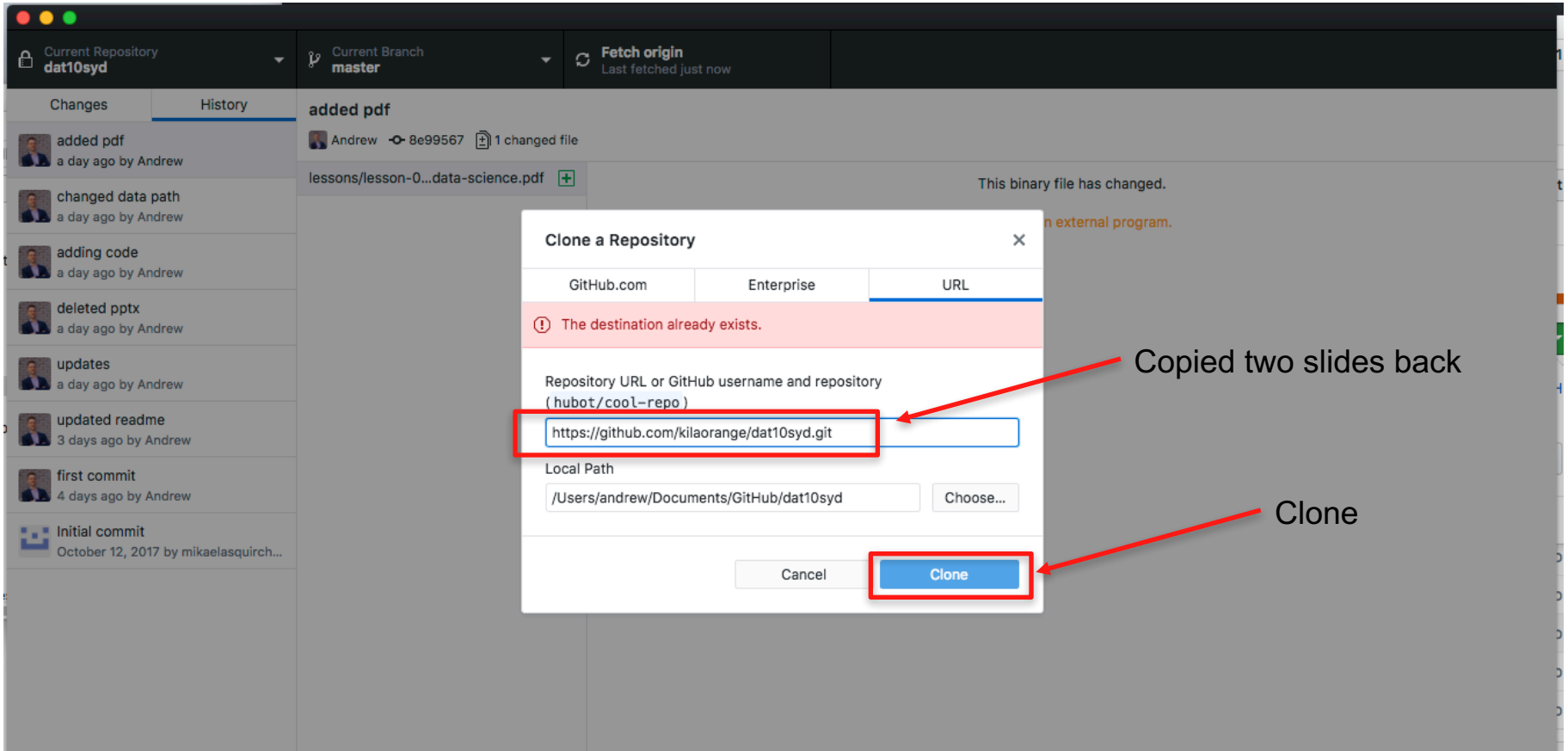
1. It now looks like this…



Preparing and Using Your Local Git



Preparing and Using Your Local Git



Preparing and Using Your Local Git

1. Fork the master repo to your GitHub account
2. Open a terminal (Mac, Linux) or git bash (Windows)
3. Change working directory to where you want the local repo

    $ cd yourpath

1. Clone

    $ git clone https://github.com/yourgithubname/yourgithubrepo

Preparing and Using Your Local Git

1. Fork the master repo to your GitHub account
2. Open a terminal (Mac, Linux) or git bash (Windows)
3. Change working directory to where you want the local repo

    $ cd yourpath

1. Clone

    $ git clone https://github.com/yourgithubname/yourgithubrepo

Preparing and Using Your Local Git:
Configure Automatic Updates (Optional)

5.Configure your local clone to point to the official course repository

```
        $ git remote -v
$ git remote add upstream course-repo.git
$ git remote -v                                                    # check
```

course-repo = ‘https://github.com/generalassembly-studio/dat9syd’

Download new material from official course repo (upstream) and merge it

1. Ensure you're in the master branch

   $ git checkout master

1. Grab the latest changes from the master

   $ git fetch upstream

1. Merge the master changes with your repo

   $ git merge upstream/master

WARNING: Be careful not to overwrite files you have already changed in your repo unless you want to replace them with the master versions!
(Consider renaming yours or doing a PULL REQUEST.)

Commit Changes to Your Local Git and Push to Your Master Repo

1. Commit

```
$ git status                                  # show changes
$ git add filename                            # stages one file
$ git add .                                   # stages all changed files
$ git commit -m your comments                 # commits file(s), with comments
$ git status                                  # check
```

1. Push

```
$ git push origin master                      # push request auto accepted
```

origin = your GitHub repo (forked from course repo)
master = course repo