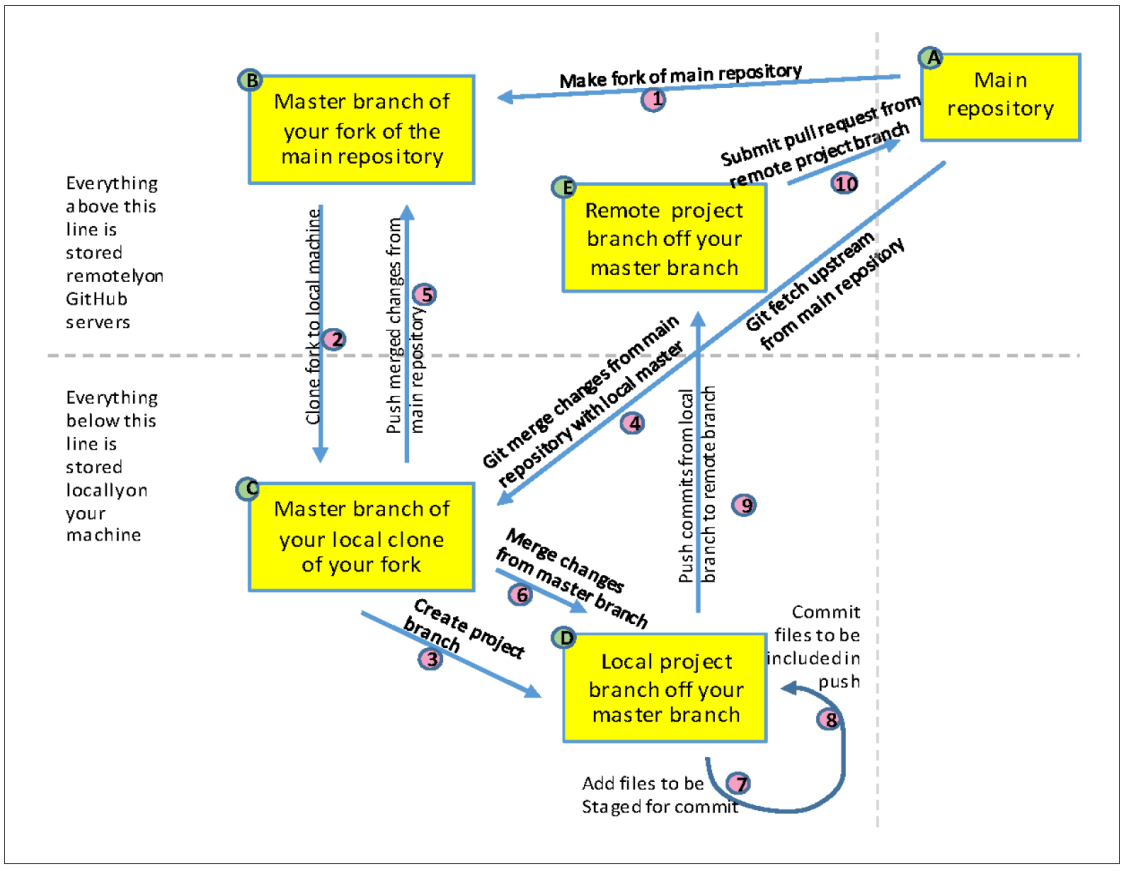
Name:_____

This quiz totals 20 points. The quiz is open-note, but not open-neighbor. Relax and try to do the best you can.

1. (4 points) Explain what a distributed version control system (DVCS) is.

> **Solution:** A DVCS is a system of tracking versions of files. In this system, collaborators each have a full copy of the repository's contents and GitHub (or another online service) tracks the contributions of each collaborator.

2. Referring to the following image, please answer the following questions.



(a) (4 points) Write down the command line syntax corresponding to arrow #4. (There are two commands — one for above the dotted line and one for below the dotted line)

Solution: The commands are:

```
git fetch upstream
```

and

```
git merge upstream/master -m "message"
```

(b) (4 points) Write down the command line syntax for arrow #2

Solution: The command is `git clone https://github.com/[fork_acct_username]/[owner_reponame].git`

(c) (4 points) Write down the command line syntax for arrow #5

Solution: The command is `git push origin master`

(d) (4 points) Explain how a pull request differs from a "push"

Solution: A pull request is when you submit suggested changes to an owner's repository. A push is when you send to GitHub a set of changes to a repository that you are the owner of.

Name:_____

This quiz totals 20 points. The quiz is open-note, but not open-neighbor. Relax and try to do the best you can.

1. (10 points) Explain the difference between a static webpage and a dynamic webpage. What is the implication of this difference as it relates to the workflow for automated web scraping?

   Solution: A dynamic webpage is one in which the HTML source does not contain the data of interest. For example, the website (e.g. Google Maps) requires the user to input an address, from which the data are revealed.

   In a static webpage, the data are nested inside the HTML source code and hence the webpage can be downloaded as a text file and parsed to obtain the data of interest.

   With a dynamic webpage, one needs to use a "`requests`" package to pass the environment variables corresponding to the exact data of interest, and then access the webpage in an automated fashion to get it to return the requested data.

2. (10 points) You are using `git` from the command line and run into the following problem: You have made changes in your directory, but when you type `git push origin master`, nothing happens. (i.e. `git` returns with the message "`Everything up-to-date`"). More details below:

   When you type `git push origin master`, you see the following output:

   ```
   Username for 'https://github.com': [username]
   Password for 'https://[username]@github.com':
   Everything up-to-date
   ```

   When you type `git status`, you see the following output:

   ```
   # On branch master
   # Untracked files:
   # (use "git add <file>..." to include in what will be committed)
   #
   #     ./
   nothing added to commit but untracked files present
   (use "git add" to track)
   ```

   Explain what is going on and how to resolve this problem.

Solution: What is going on here? The problem is that the user needs to *add* the files to the staging area, then issue a commit, and then do the push. The steps are:

- `git add ./`
- `git commit -m "commit message"`
- `git push origin master`

Name:_____

This quiz totals 20 points. The quiz is open-note, but not open-neighbor. Relax and try to do the best you can.

1. (5 points) Explain the difference between a continuous and discrete variable. From a modeling standpoint, why do we care about this difference?

Solution: A continuous variable is one which can take on any value over a continuous range—typically $\mathbb{R}$ (the real number line) or $\mathbb{R}^+$ (the positive real numbers).

A discrete variable is one which can only take on a (typically small) finite set of values.

Within discrete variables, we have three groups:

1. binary (0/1)
2. categorical (0,1,...,K)
3. count (non-negative integers)

2. (5 points) Explain the difference between Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Which case will result in imputed values that are most plausible?

Solution: We have:

- **Missing Completely At Random**: missing values randomly happen to be missing
- **Missing At Random**: missing values are not MCAR, but are missing in such a way that we could reconstruct them using other observable information in our data set
- **Missing Not At Random**: missing values are missing in part because of some information that is not contained in the data set

MCAR is easiest to resolve, followed by MAR. *MNAR is essentially impossible to resolve unless you happen to know the process governing the missingness.*

3. (5 points) What is the most common use of the word "optimization" in Data Science? Why is optimization important?

> **Solution:** "Optimization" typically means "finding the minimum of an objective function." It's important because it is required to estimate any kind of statistical model!

4. (5 points) Explain what a closed-form solution is. What fraction of objective functions in data science have a closed-form solution?

> **Solution:** A closed-form solution is a solution to a problem that can be expressed analytically (i.e. with a written mathematical formula). Its opposite is a **numerical solution**, which requires a computer to iteratively solve.
>
> As luck would (not) have it, only a very small number of objective functions have a closed-form solution. (Basically, just OLS and a handful of others.)

Name:_____

This quiz totals 20 points. The quiz is open-note, but not open-neighbor. Relax and try to do the best you can.

1. (5 points) What is machine learning? How does it work? Give some examples of machine learning in the world today.

> **Solution:** Machine learning is the idea that we can get computers to "learn" for themselves without being explicitly programmed. The computer is "trained" according to some task and comes up with a function on its own that predicts well out of sample. It then uses this function to pass its knowledge on to other contexts of the same task.
>
> Popular examples of machine learning in the world today include voice assistants, Optical Character Recognition (OCR) systems, and gaming bots.

2. (5 points) What is the difference between L1 and L2 regularization?

> **Solution:** L1 regularization penalizes the objective ("cost") function by $\lambda \sum_k |\beta_k|$. L2 regularization penalizes by $\lambda \sum_k \beta_k^2$.

3. (5 points) What is the tradeoff that regularization attempts to solve? How is this related to overfitting?

> **Solution:** Regularization attempts to solve the bias-variance tradeoff. That is, an algorithm that is not flexible or complex enough will do a poor job of predicting both in-sample and out-of-sample. We say that this algorith has high bias or that this algorithm underfits.
>
> An algorithm that is too flexible or complex will do a great job of predicting in-sample but a poor job of predicting out-of-sample. We say that this algorithm has high variance or that this algorithm overfits.

4. (5 points) Machine learning is all about finding the best ____, and econometrics is all about finding the best ____.

   a. $\hat{y}$; $\hat{y}$

   b. $\hat{\beta}$; $\hat{y}$

   c. $\hat{y}$; $\hat{\beta}$

   d. $\hat{\beta}$; $\hat{\beta}$

Solution: The answer is (c). This is straight from the lecture notes.

Name:_____

This quiz totals 20 points. The quiz is open-note, but not open-neighbor. Relax and try to do the best you can.

1. (5 points) What is the point of doing cross validation, and why is CV important?

   Tying to find the optimal tradeoff between bias and variance

2. (5 points) Explain how to measure fit (i.e. prediction accuracy) in regression vs. classification models.

   Mean-squared error, root mean squared error, mean absolute squared errors

   F1, and confusion matrix (see solutions for more details)
   Do not use mean-squared errors for a classification problem

3. (5 points) Name the five tribes of machine learning and write down the master algorithm of each.

   Symbolists: tree models
   Connectionists: nnet
   Bayes : Support vector machines
   Others are on the five tribes section

4. (5 points) Explain difference between supervised and unsupervised machine learning.

   See answers