

# DATA SCIENTIST'S TOOLBOX OVERVIEW, BASIC R

---

**Lecturer : Professor Pao-Ann Hsiung**

Teaching Assistant: Hendrik

Embedded Systems Laboratory  
National Chung Cheng University  
Chiayi, Taiwan-62102

# Learning Objective

- Install the R and RStudio software packages
- Git and GitHub
- Download and install the swirl package for R
- Execute basic arithmetic operations
- Get to know the sense of R

# Outline

- Data Scientist's Toolbox
  - Introduction to basic tools
  - R&Rstudio
  - R tools
  - Git&Github
  - Getting Help
- Basic R
  - Data input
  - First program
  - R packages
- Demo & Assignment

# R & RStudio

- R distribution available on <https://cran.r-project.org/>

The image shows a screenshot of the RGui (64-bit) application window and the Comprehensive R Archive Network (CRAN) website. The RGui window is on the left, displaying the R console with the following text:

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

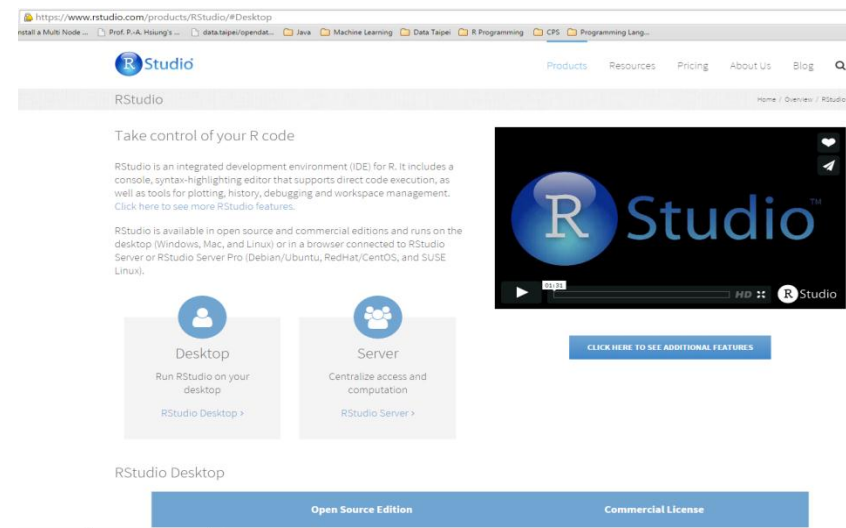
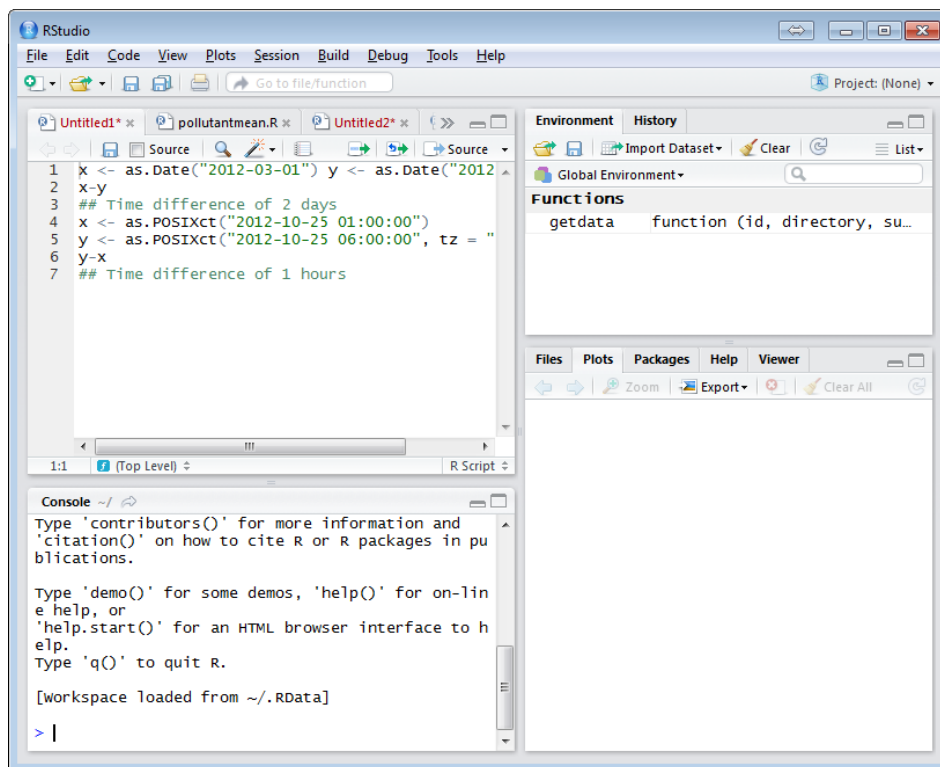
[Previously saved workspace restored]

> |
```

The CRAN website is on the right, showing the "Download and Install R" section. It provides links for downloading R for Linux, Mac OS X, and Windows. It also includes information about the latest release (2015-06-14, Fire Safety) and sources of R alpha and beta releases. The website also mentions that R is part of many Linux distributions and that Windows and Mac users should check their Linux package management system. It provides links for downloading R for Linux, Mac OS X, and Windows, and for downloading R for Windows. It also includes information about the latest release (2015-06-14, Fire Safety) and sources of R alpha and beta releases. The website also mentions that R is part of many Linux distributions and that Windows and Mac users should check their Linux package management system. It provides links for downloading R for Linux, Mac OS X, and Windows, and for downloading R for Windows. It also includes information about the latest release (2015-06-14, Fire Safety) and sources of R alpha and beta releases.

# R & Rstudio

- **Rstudio**: IDE for R includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management
- Available on <https://www.rstudio.com/products/rstudio/>



# R Tools

- A collection of tools for building R packages in windows
- Available at <https://cran.r-project.org/bin/windows/Rtools/>
- For detail refer to that website

\* Not necessary for linux or mac

# Packages

- See available packages on CRAN

`available.packages()`

- Install

`install.packages("namesPackages")`

- Load

`library("namesPackages")`

The screenshot shows the RStudio environment. The main editor window displays a script with the following code:

```
1 install.packages("swirl")
2 #single call to install
3 install.packages(c("slidify", "ggplot2", "devt
4 #place the name of the packages in a chara
5 source("http://bioconductor.org/biocLite.R"
6 #to get basic installer (multiple packages)
7 biocLite()
8 library(ggplot2)
9 #load packages
10 search()
11 #list what packages in the workspace
```

The console window at the bottom shows the output of the first command:

```
[1] 22 33 44 55
> install.packages("swirl")
warning in install.packages :
  downloaded length 227 != reported length 227
Installing package into 'C:/Users/Infinity/Docume
nts/R/win-library/3.1'
(as 'lib' is unspecified)
```

# Git & GitHub



You don't need GitHub to use Git

Git = Local (On your computer);  
Github=Remote (on the Web)

## Vs



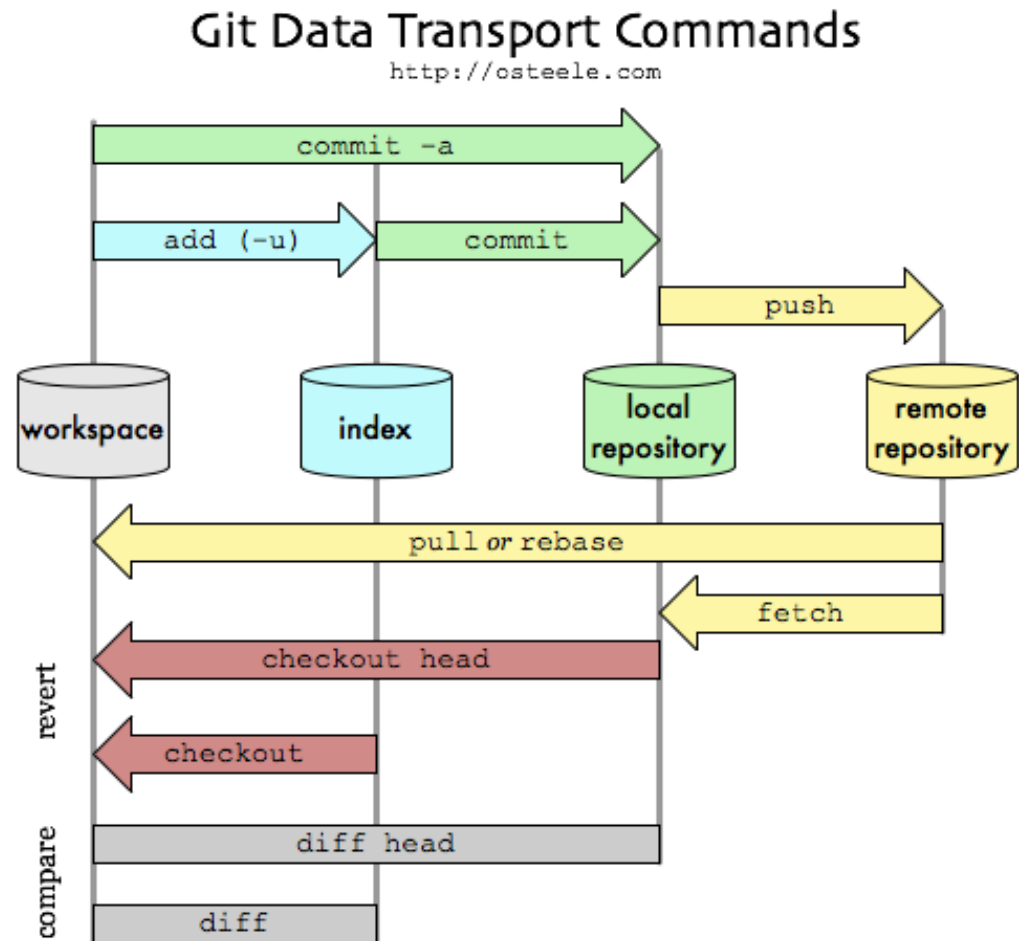
- GitHub Allow you to:
  1. Share your Repositories with others
  2. Access other user's repositories
  3. Store remote copies of your repositories (on GitHub's Server) in case something happens to your local copies (on your computer)

Go to the following website to download <https://git-scm.com/downloads>



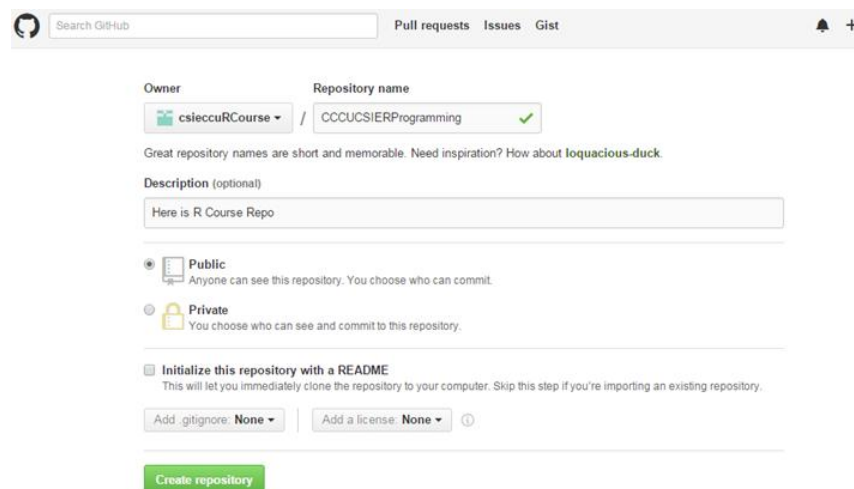
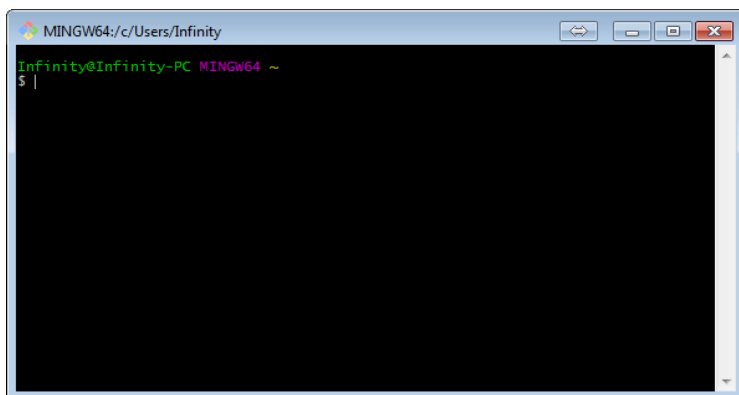
# Git Basic

- **init**
- **clone**
- **add**
- **status**
- **commit**
- **log**
- **diff**
- **checkout**



# Getting started: Git and Github

- Create a your own github and make a new repository named it *Lab1\_studentID*
- Open your git bash, download git here



# Creating GitHub Repo

- Two Methods
  1. Start Repository from scratch
  2. Fork another user's repository
- Repo From Scratch
  - Go to your profile page(<https://github.com/usernameHere>) and click on “Create a new repo” in the righthand corner of the page
  - Or go directly to <https://github.com/new>
  - Select public(Private repos require a paid or educational account)
  - Check the box next to initalize this repository with a README

# Creating a Local Copy

- Now you need to create a copy of this repo on your computer
- Open git bash
- Create a directory on your computer where you will store your copy of the repo

```
$ mkdir ~/test-repo
```

```
$ cd ~/test-repo
```

Initialize a local git repo

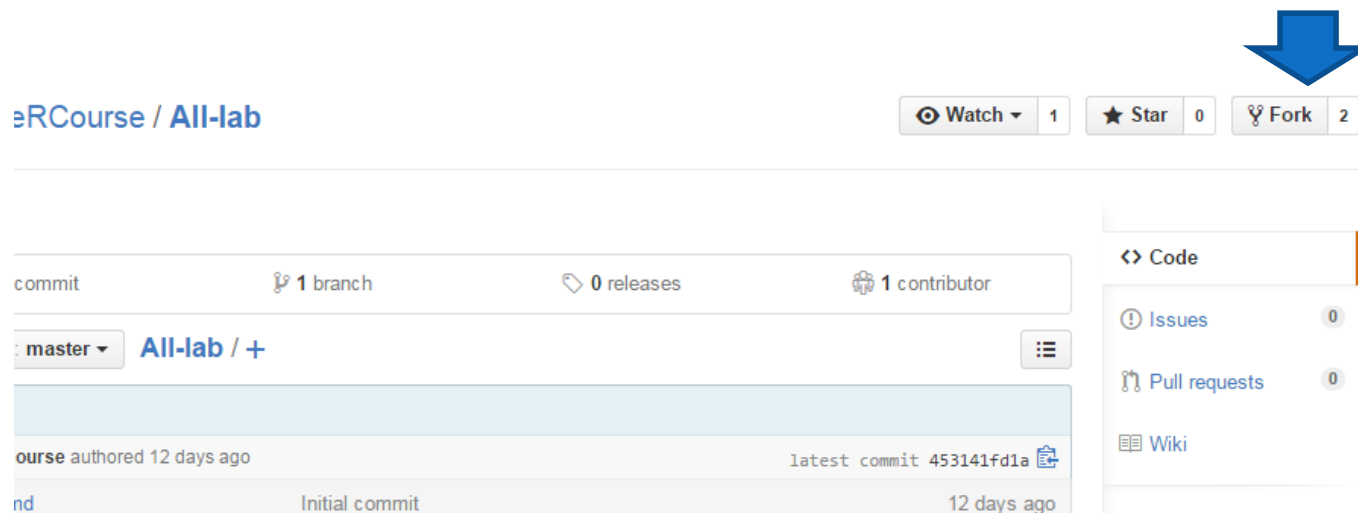
```
$ git init
```

Initialize Point your local repository at the remote repository you just created on the GitHub server

```
$ git remote add origin git@github.com:yourUserName/test-repo.git
```

# Fork a Another User's Repository

- The second method of creating a repository is to make a copy of someone else's
- Begin by navigating to the desired repository on GitHub website and click the “Fork” button

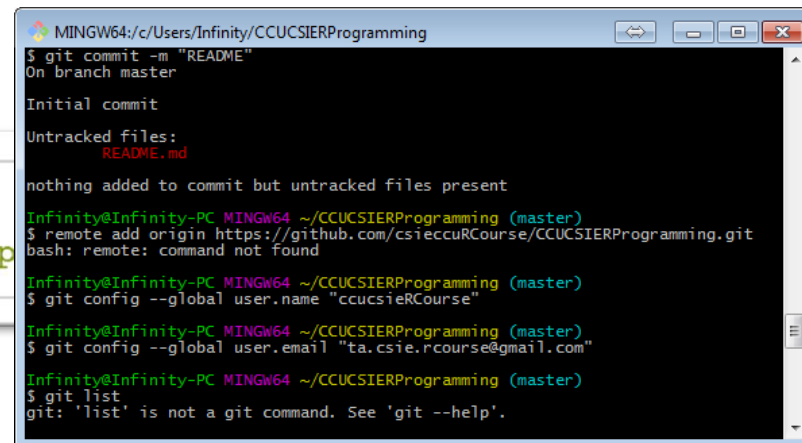


<https://help.github.com/articles/fork-a-repo/>

# Configure username and email

- Each commit to a Git repository will be tagged with the username of the person who made the commit
- To set your username and email, Enter the following commands in Git Bash

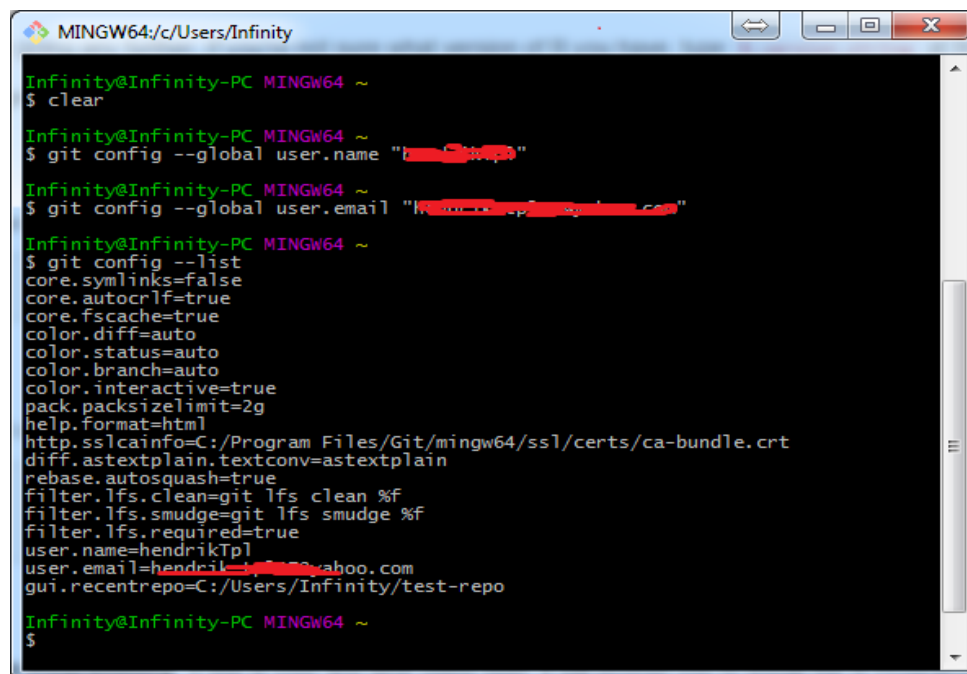
```
$ git config --global user.name "Your Name Here"  
$ git config --global user.email "your_email@example.com"
```



```
MINGW64:/c/Users/Infinity/CCUCSIErProgramming  
$ git commit -m "README"  
On branch master  
Initial commit  
Untracked files:  
  README.md  
nothing added to commit but untracked files present  
Infinity@Infinity-PC MINGW64 ~/CCUCSIErProgramming (master)  
$ remote add origin https://github.com/csieccuCourse/CCUCSIErProgramming.git  
bash: remote: command not found  
Infinity@Infinity-PC MINGW64 ~/CCUCSIErProgramming (master)  
$ git config --global user.name "ccucsieRCourse"  
Infinity@Infinity-PC MINGW64 ~/CCUCSIErProgramming (master)  
$ git config --global user.email "ta.csie.rcourse@gmail.com"  
Infinity@Infinity-PC MINGW64 ~/CCUCSIErProgramming (master)  
$ git list  
git: 'list' is not a git command. See 'git --help'.
```

# Cont.

- To confirm your changes, type the following command
  - `$ git config --list`



```
MINGW64/c/Users/Infinity

Infinity@Infinity-PC MINGW64 ~
$ clear

Infinity@Infinity-PC MINGW64 ~
$ git config --global user.name "h[REDACTED]"

Infinity@Infinity-PC MINGW64 ~
$ git config --global user.email "[REDACTED]@yahoo.com"

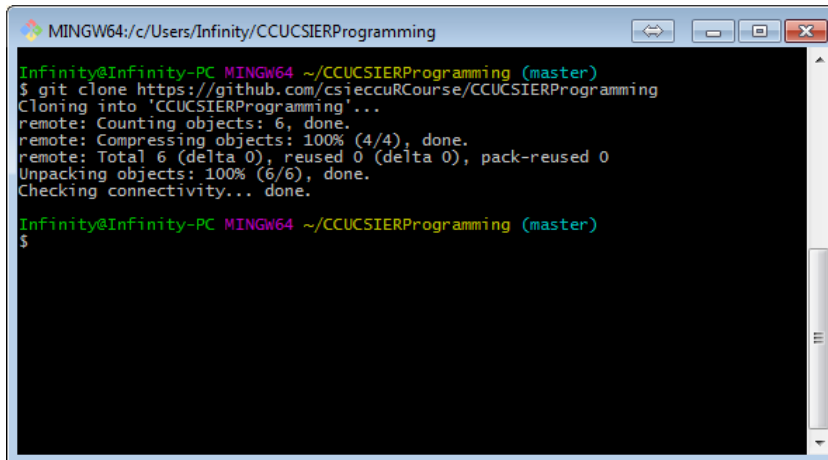
Infinity@Infinity-PC MINGW64 ~
$ git config --list
core.symlinks=false
core.autocrlf=true
core.fscache=true
color.diff=auto
color.status=auto
color.branch=auto
color.interactive=true
pack.packsizelimit=2g
help.format=html
http.sslcainfo=C:/Program Files/Git/mingw64/ssl/certs/ca-bundle.crt
diff.astextplain.textconv=astextplain
rebase.autosquash=true
filter.lfs.clean=git lfs clean %f
filter.lfs.smudge=git lfs smudge %f
filter.lfs.required=true
user.name=hendrikTp1
user.email=hendrik[REDACTED]@yahoo.com
gui.recentrepo=C:/Users/Infinity/test-repo

Infinity@Infinity-PC MINGW64 ~
$
```

# Clone

- To make a local repository on your computer
- to get a copy of an existing Git repository on your GitHub Account

```
git clone https://github.com/yourUserName/repoName
```



```
MINGW64:/c/Users/Infinity/CCUCSIERProgramming
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ git clone https://github.com/csieccuRCourse/CCUCSIERProgramming
Cloning into 'CCUCSIERProgramming'...
remote: Counting objects: 6, done.
remote: Compressing objects: 100% (4/4), done.
remote: Total 6 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (6/6), done.
Checking connectivity... done.

Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$
```

This will clone repo into your current directory



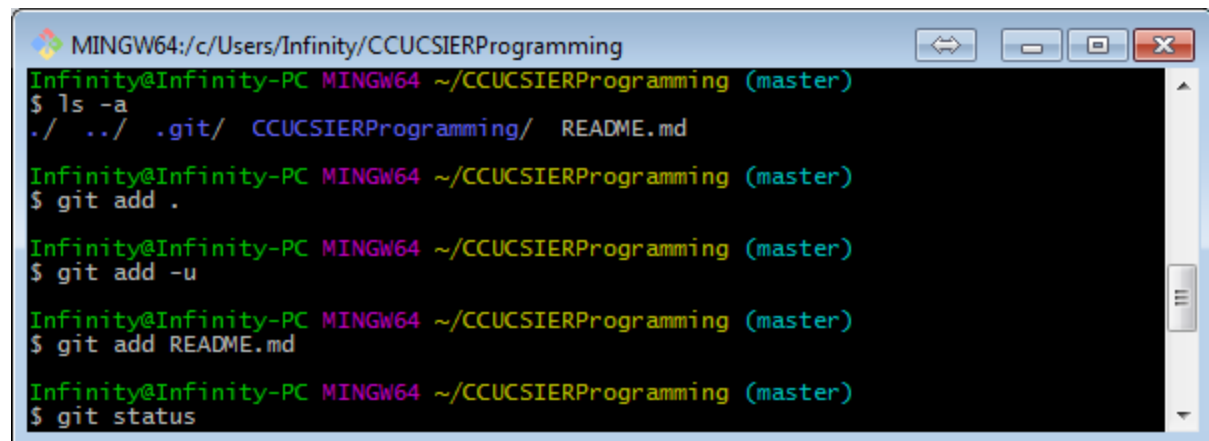
# Add

- Suppose you add new files to a local repository under version control
- You need to let git know that they need to be tracked
- Should do this before committing

**git add .** adds all news file

**git add -u** updates tracking for files that changed names or were deleted

**git add -A** both



```
MINGW64:/c/Users/Infinity/CCUCSIERProgramming
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ ls -a
./ ../ .git/ CCUCSIERProgramming/ README.md

Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ git add .

Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ git add -u

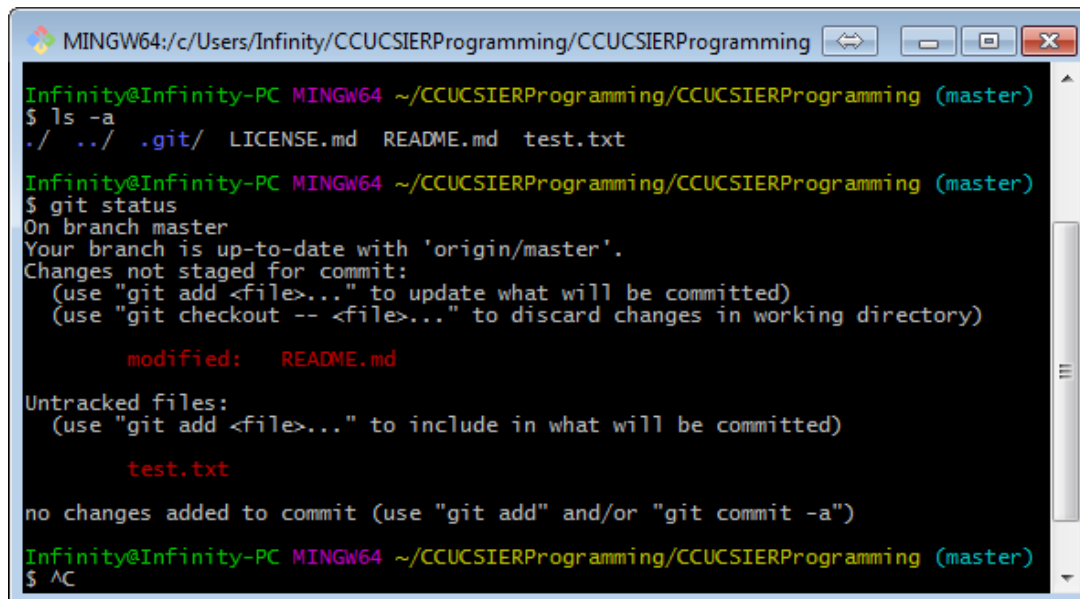
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ git add README.md

Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming (master)
$ git status
```

# Status

- As you saw in the git add section, in order to see what the status of your staging area is compared to the code in your working directory

git status  
git status -s

A screenshot of a terminal window titled 'MINGW64:/c/Users/Infinity/CCUCSIERProgramming/CCUCSIERProgramming'. The terminal shows the following commands and output:

```
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$ ls -a
./ ../ .git/ LICENSE.md README.md test.txt

Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$ git status
On branch master
Your branch is up-to-date with 'origin/master'.
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:   README.md

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

        test.txt

no changes added to commit (use "git add" and/or "git commit -a")
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$ AC
```

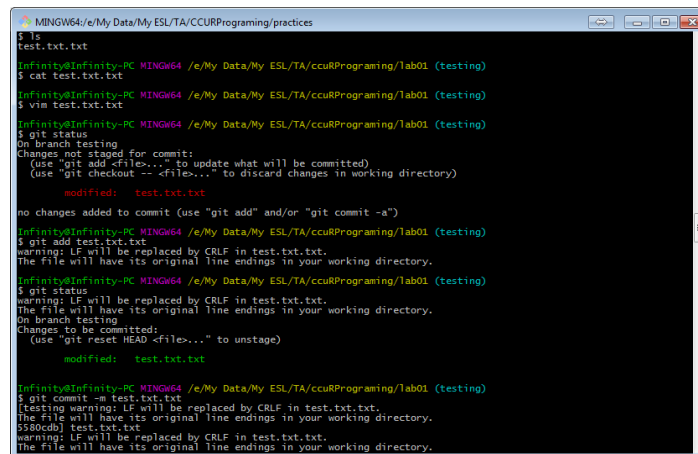
# Commit

- Commit the files that you've staged in your local repository

```
$ git commit -m 'First commit'
```

# Commits the tracked changes and prepares them to be pushed to a remote repository. To remove this commit and modify the file, use 'git reset --soft HEAD~1' and commit and add the file again.

- This only update your local repo , not the remote on github



```

MINGW64/e/My Data/My ESL/TA/ccurProgramming/practices
$ file
test.txt.txt
$ cat test.txt.txt
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ cat test.txt.txt
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ vim test.txt.txt
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ git status
On branch testing
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:   test.txt.txt

no changes added to commit (use "git add" and/or "git commit -a")
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ git add test.txt.txt
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ git status
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
On branch testing
Changes to be committed:
  (use "git reset HEAD <file>..." to unstage)

        modified:   test.txt.txt

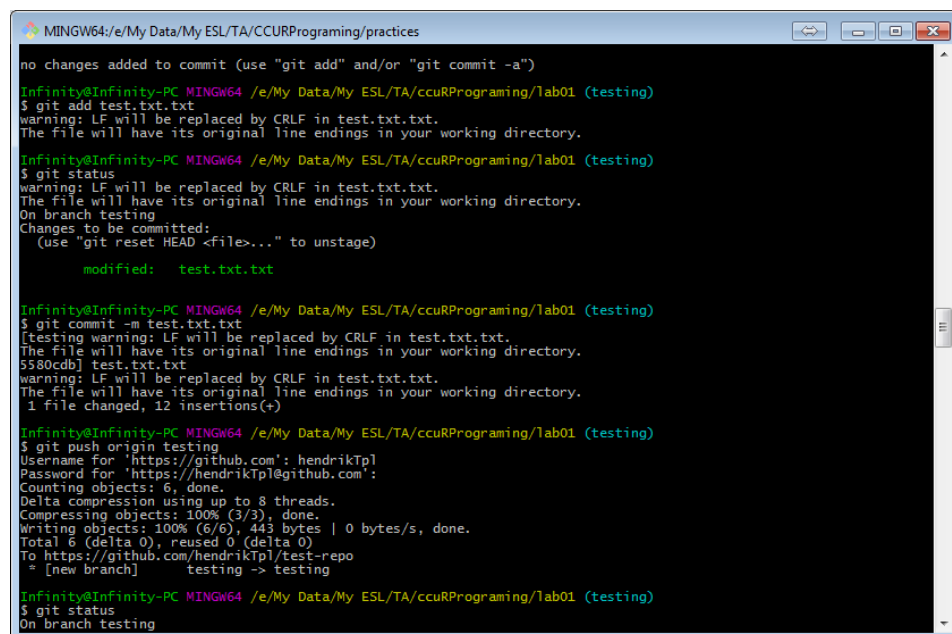
Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccurProgramming/lab01 (testing)
$ git commit -m test.txt.txt
[testing warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
5580c0db] test.txt.txt
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.

```

# Push

- Update local commits on the remote (Github)

```
$ git push origin master
# Pushes the changes in your local repository up to the
remote repository you specified as the origin
```



```
MINGW64:/e/My Data/My ESL/TA/CCURProgramming/practices
no changes added to commit (use "git add" and/or "git commit -a")

Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccuRProgramming/lab01 (testing)
$ git add test.txt.txt
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.

Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccuRProgramming/lab01 (testing)
$ git status
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
On branch testing
Changes to be committed:
  (use "git reset HEAD <file>..." to unstage)

        modified:   test.txt.txt

Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccuRProgramming/lab01 (testing)
$ git commit -m test.txt.txt
[testing warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
5580cdb] test.txt.txt
warning: LF will be replaced by CRLF in test.txt.txt.
The file will have its original line endings in your working directory.
1 file changed, 12 insertions(+)

Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccuRProgramming/lab01 (testing)
$ git push origin testing
Username for 'https://github.com': hendrikTp1
Password for 'https://hendrikTp1@github.com':
Counting objects: 6, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (3/3), done.
Writing objects: 100% (6/6), 443 bytes | 0 bytes/s, done.
Total 6 (delta 0), reused 0 (delta 0)
To https://github.com/hendrikTp1/test-repo
 * [new branch]   testing -> testing

Infinity@Infinity-PC MINGW64 /e/My Data/My ESL/TA/ccuRProgramming/lab01 (testing)
$ git status
On branch testing
```

# log & diff

- **git log** to find specific commits in your project history - by author, date,
- **git diff** to compare two different points in your history - generally to see how two branches differ or what has changed from one version of your software to another content or history.

The image shows two terminal windows side-by-side, both running in a MINGW64 environment at the path `c:/Users/Infinity/CCUCSIERProgramming/CCUCSIERProgramming`.

The left terminal window shows the output of `git diff` comparing the local version of `README.md` with the version in branch `b`. The output indicates that the local version has 1 line added and 5 lines changed compared to branch `b`.

```

MINGW64:/c/Users/Infinity/CCUCSIERProgramming/CCUCSIERProgramming
Update LICENSE.md
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming
$ AC
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming
$ git diff
diff --git a/README.md b/README.md
index 6f247c3..47500cd 100644
--- a/README.md
+++ b/README.md
@@ -1,3 +1,5 @@
 # CCUCSIERProgramming
+there is markdown
+
 CSIE R Programming Repo
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming
$ |
  
```

The right terminal window shows the output of `git log`, displaying the commit history. It shows two commits: one for creating `README.md` and another for updating `LICENSE.md`.

```

MINGW64:/c/Users/Infinity/CCUCSIERProgramming/CCUCSIERProgramming
If no other git process is currently running, this probably means a
git process crashed in this repository earlier. Make sure no other git
process is running and remove the file manually to continue.
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$ git log
commit abe6117c91cbee196f8e4eedd93bfaef362c8035
Author: ccucsierCourse <ta.csie.rcourse@gmail.com>
Date: Mon Sep 7 02:27:33 2015 +0800

    Create README.md

commit f3fccbaa9ecfc172c6520aa7c3e62f6cc24008a8
Author: ccucsierCourse <ta.csie.rcourse@gmail.com>
Date: Mon Sep 7 02:26:49 2015 +0800

    Update LICENSE.md
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$ AC
Infinity@Infinity-PC MINGW64 ~/CCUCSIERProgramming/CCUCSIERProgramming (master)
$
  
```

# Branches

- Sometimes you are working on a project with a version being used by many people. You may not want to edit that version
- So you can create a branch with the command  
`git checkout -b branchname`
- To see what branch you are on type  
`git branch`
- To switch back to master branch type  
`git checkout master`

More about git, please kindly go to <https://git-scm.com/doc> or <http://gitref.org/>

# Getting Started

- Open R console
  - Commands entered will be executed immediately
- Demo some simple tasks
- Calculation
  - E.g.  $1+5*(2^3)/2$
- Statistical analysis
  - E.g. linear regression
- Graphing
  - E.g. boxplot

```

RGui (64-bit)
File History Resize Windows

R Console
Natural language support not running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> # this is a comment
> # commands entered will be executed immediately
> 3
[1] 3
> # a basic calculation
> 1+5*(2^3)/2
[1] 21
> #create a vector length 10
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
> #graping
> boxplot(1:10 , 1.5)
> |
  
```

# Entering Input & Evaluation

- `<-` symbol is the assignment operator      `VarNames <- c("variables")`
- `#` character indicates a comment

```
> x <- 1
> print(x)
[1] 1
> x
[1] 1
> msg <- "hello"
> x <- ## Incomplete expression
```

When a complete expression is entered at the prompt, it is evaluated and the result of the evaluated expression is returned

```
> x <- 5 ## nothing printed
> x ## auto-printing occurs
[1] 5
> print(x) ## explicit printing
[1] 5
```

The `[1]` indicates that `x` is a vector and `5` is the first element

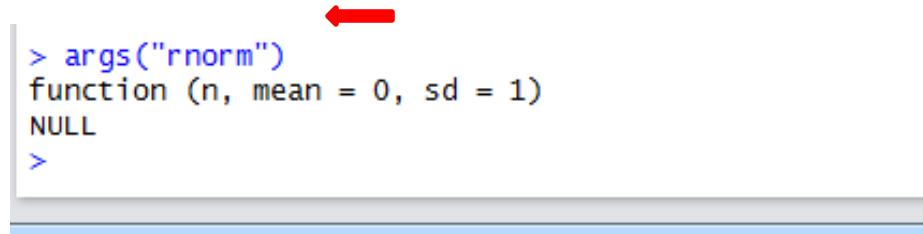
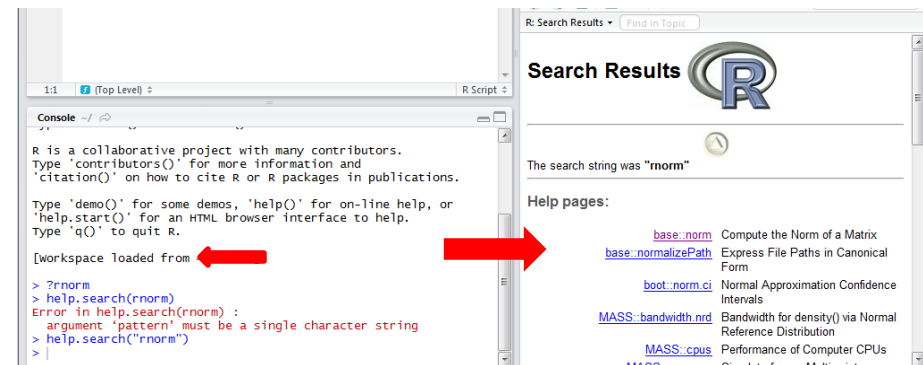
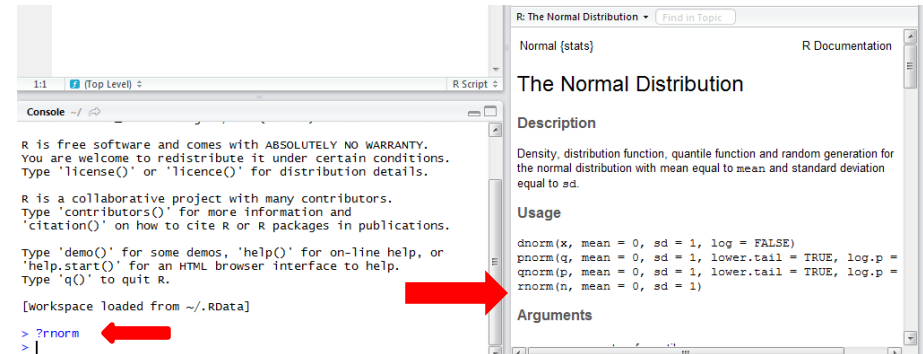


# First R

- `X <-1:10`      `#create x with 10 numbers`
  - `Print(x)`      `#print numbers in x`
  - `Y <-2*x-1`      `#create y based on x (also 10 num)`
  - `Print(y)`
  - `Mean(y)`      `#calculate average of numbers in y`
  - `Model <- lm(y~x)`      `#fit a linear regression model`  
                                 `#  $y=B_0+(B_1 \cdot X)$`
- `Print(Model)`      `# print the linear model`  
                                 `# $B_0 = -1$  ,  $B_1=2$`
- `Boxplot(x,y)`      `#plot distribution of x and y`

# Getting help

- Access help file
  - `?topic`
    - E.g. `?rnorm` for Normal Distribution
- Search help files
  - `help.search("topic")`
- Get arguments
  - `args("topic")`



# DEMO & ASSIGNMENT

---

# Part1 (setup your own environment)

## You are asked to setup your R & Git

- R
  - Install R
  - Install Rstudio
  - Open R & Rstudio and take a screenshot
- Github
  - Set up a Github account (you may use your own name or a pseudonym).
  - Create a repo called **Lab1\_studentID** (scratch)
  - Clone your github into local repo
  - Create a text file called HelloWorld.md
  - Add the line

```
##This is a secondary heading
### This is a Tertiary heading
* first item list
* second item list
```
  - Commit every changes
  - Push the document to the **Lab1\_studentID** repo you created on Github
  - Fork the data sharing repository here <https://github.com/ccucsieRCourse/sharing>

# Part2 (My first R)

## **write your first R**

- Do some calculations (e.g.  $1*1+2*2+3*3+\dots+100*100$ )
- Learn some datasets (e.g. `?inspect`)
- Learn some function (e.g. `?seq`, `?mean`)
- Learn some package (e.g. `?stats`, `?swirl` )
- Try example in the help file

## Part 3 (Simple Data Analysis)

- Table morphometric measurements of eight birds.

Wingcrd	Tarsus	Head	Wt
59	22.3	31.2	9.5
55	19.7	30.4	13.8
53.5	20.8	30.6	14.8
55	20.3	30.3	15.2
52.5	20.8	30.3	15.5
57.5	21.5	30.8	15.6
53	20.6	32.5	15.6
55	21.5	NA	15.7

- Use the `c` function, create a variable `wingcrd` to hold the wing lengths for the 8 birds
  - Similarly, create `Tarsus` , `head` and `wt`
  - Compute basic statistics of the data (mean, median , min , max)

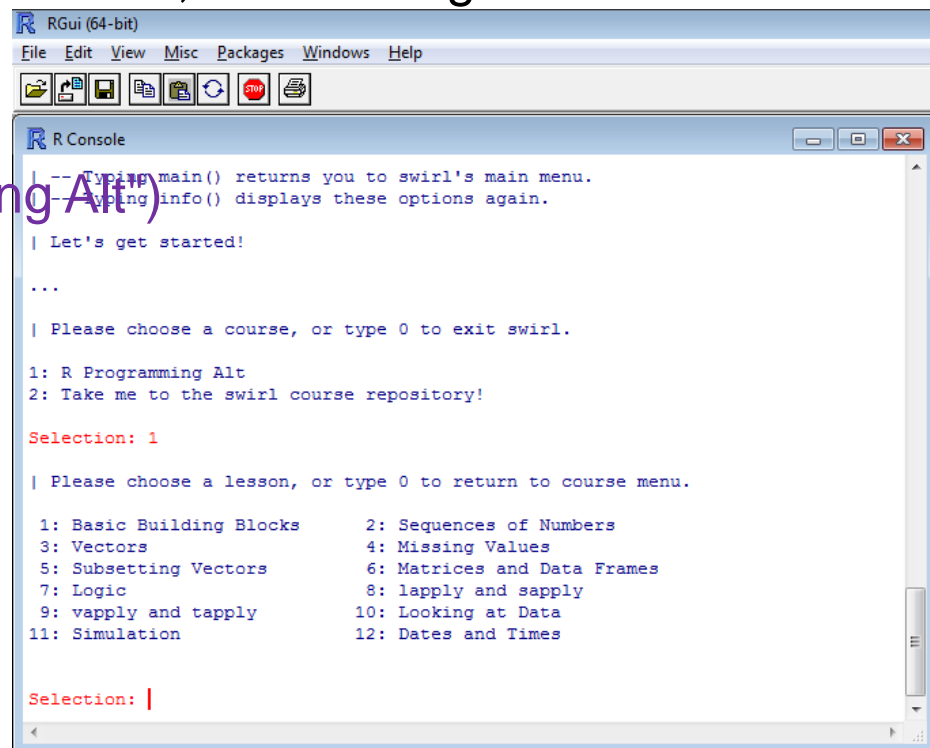
# Swirl exercise(home)

- 1) Make sure you have a recent version version of swirl:  
`install.packages("swirl")`
- 2) Enter the following from the R console, substituting the name of the course that you wish to install:

```
library(swirl)
```

```
install_from_swirl("R Programming Alt")
```

```
swirl()
```



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
| --Typing main() returns you to swirl's main menu.
| --Typing info() displays these options again.
| Let's get started!
...
| Please choose a course, or type 0 to exit swirl.
1: R Programming Alt
2: Take me to the swirl course repository!
Selection: 1
| Please choose a lesson, or type 0 to return to course menu.
1: Basic Building Blocks      2: Sequences of Numbers
3: Vectors                   4: Missing Values
5: Subsetting Vectors        6: Matrices and Data Frames
7: Logic                     8: lapply and sapply
9: vapply and tapply        10: Looking at Data
11: Simulation                12: Dates and Times
Selection: |
```

# Hand in

- To submit this lab
  - Use your own repo that you created on your Github
  - R scripts are called regards to part number (e.g. part2.R)
  - Zip **Lab1\_StudentID** that contain **R&Rstudio screenshot**, **part2.R** , **part3.R**
  - Commit each file you have modified
  - Submit the link to the forked repository on your Github account



# Hand in

- In order to record to assignment you still need to upload to
- The E-course System
  - <https://ecourse.ccu.edu.tw/>
- Upload **StudentID.zip** into “**Lab1**”

# GRADING POLICIES

---

# Grading Policies

- There will be 8 Labs =24%
  - Lab1 =2%, Lab2=2%, Lab3=3%, Lab4=3%, Lab5=3%, Lab6=3%, Lab7=4%, Lab8=4%
- Finish All Lab1 in order to get full credit (part1-part3)
  - In case you haven't finished by the end of class, you can still send it by reducing 20% from full score for each day..
  - e.g. Total score lab1 100, it will 80, 60, 40, ....

# References

- Roger, D. Peng.(2015). *R Programming for Data Science*. Victoria, British Columbia : Leanpub.
- Joseph, A. (2012). *R IN A NUTSHELL: A Desktop Quick Reference, 2<sup>nd</sup> Edition*. Sebastopol, CA : O'Reilly.
- Alain, F. Z. ,Elena N. L., Erik H.W.G. (2009). *A Beginner's Guide to R*. Newburgh,United Kingdom : Springer
- <https://www.r-project.org/>
- <http://gitref.org/>
- <https://git-scm.com/book/en/v2/Git-Basics-Getting-a-Git-Repository>
- <https://www.coursera.org/jhu>