

# CWRU DSCI351-451: Week01a-f

*Roger H. French*

*28 August, 2018*

## Contents

1.1.1.1	Reading, Homeworks, Projects, SemProjects . . . . .	1
1.1.1.2	Syllabus . . . . .	1
1.1.1.3	What we need to do this week . . . . .	1
1.1.1.4	Your Open Data Science Tool Chain . . . . .	3
1.1.1.4.1	Its all about a Data Science Tool Chain . . . . .	3
1.1.1.4.2	Online Git Server Communities . . . . .	3
1.1.1.4.3	Kaggle Account . . . . .	3
1.1.1.4.4	Twitter used for Data Science . . . . .	4
1.1.1.4.5	Sign up for a Stack Exchange Account . . . . .	4
1.1.1.4.6	Efficiently browse you SX sites . . . . .	5
1.1.1.4.7	Slack, another component of Agile Software Development . . . . .	5
1.1.1.5	You Online Data Science Portfolio . . . . .	5
1.1.1.5.1	An Example, Emeline Liu . . . . .	5
1.1.1.6	Links . . . . .	5

### 1.1.1.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - Peng R Programming pages 4 to 33
- Homeworks
  - HW1 given out thursday, due next tuesday
- Data Science Projects:
  -
- 451 SemProjects:
  -
- Friday Comm. Hour
  -

### 1.1.1.2 Syllabus

### 1.1.1.3 What we need to do this week

1. Setup VDI
  - Rstudio
  - Drag icons of R, Rstudio, Git Bash, Spyder, Jupyter Notebook, HipChat to desktop
2. Setup Git
  - make H:\Git folder
  - git config name and email
3. Setup Bitbucket account
4. Setup Kaggle account
5. Setup Twitter account
6. Setup StackExchange account

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	<b>HW1 Due</b>
w2b:Th:9/6/18	Data Analytic Style, Git	Teatime:Intro2R, For loops	PRP65-93	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis, SemProj	ISLR:Intro2R	PRP94-116	<b>HW2 Due</b>
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr	OI1-1.9,	
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	<b>HW3 Due</b>
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	<b>SemProj1,</b>
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	<b>Proj1 Due</b>
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	<b>MIDTERM EXAM</b>			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	<b>HW4 Due</b>
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	<b>CWRU FALL BREAK</b>		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	<b>SemProj2 HW5 Due</b>
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	<b>Proj.2 due</b>
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	<b>HW6 due</b>
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	<b>Proj 3 due</b>
Th:11/22/18	<b>THANKSGIVING</b>			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		<b>SemProj3</b>
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			<b>Proj4</b>
	<b>FINAL EXAM</b>	<b>Monday 12/17, 12:00-3:00pm</b>	Olin 313	<b>SemProj4 due</b>

Figure 1: DSCI351-351M-451 Syllabus

7. Setup Slack account
8. Git Clone
  - 18-sdle-teatime
    - for quick introduction to data science techniques and tools
  - For Class-Prof Repo
    - Clone your forked Class Repo

#### 1.1.1.4 Your Open Data Science Tool Chain

##### 1.1.1.4.1 Its all about a Data Science Tool Chain

- Use R and build on the communities foundation
- Use Rstudio as a comfy environment
- Share your Open Data and Open Source Code
- Produce Reproducible Science with Rmarkdown
  - Use [Creative Commons Licenses](#)
  - Or other [Open Source Licenses](#)
  - Such as the [Gnu Public License: GPL](#)
  - Or one of my favorites, [the Apache License](#)

Pilot your Data Science studies using available data

- Find available data sets
- Before starting the costly process of making data

Use Git repositories

- For Code Version Control
- For Collaboration
- For Open Science sharing

##### 1.1.1.4.2 Online Git Server Communities

- We use [BitBucket Account](#)
  - In class, for our class code repositories
  - These are private repositories
- You'll probably also want a [GitHub](#) account.
  - Many Rprojects are there, and
  - you can fork their repo's as inspect the code very easily.

##### 1.1.1.4.3 Kaggle Account

- [Kaggle](#) started as a data science competition site
- Its recently been bought by Google
  - And give free R and Python Notebooks
  - Including use of free GPUs
- It has a very good [Intro to R, Python, Machine Learning etc.](#)
  - First R Tutorial: [Getting staRted in R: First Steps](#)
  - 2nd [R Tutorial, Level 1, on Modeling](#)
  - 3rd [R Tutorial, Level 2, on tidyverse data manipulation](#)

#### 1.1.1.4.4 Twitter used for Data Science

As part of setting up our Data Science Tool Chain

- Signup for a Twitter account
- [Using Twitter in university research](#)
- [10 Commandments of Twitter for Academics](#)

Data Science People to follow on Twitter

- @hadleywickham
- @jtleek Jeff Leek JHU
- @rdpeng Roger Peng JHU
  
- @simplystats
- @Rbloggers
- @JennyBryan
- @hspter Hilary Parker
- @NSSDeviations
- @dataandme
- @rstudio
- @rstudiotips
- @R\_Programming
- @CRANberriesFeed
- @timoreilly
- @kaggle
- @SciPyTip
- @PyData
- @debian
- @ubuntu
- @GuardianData
- @UpshotNYT
- @EdwardTufte
- @ProjectJupyter
- @doctorow Cory Doctorow
- @gvanrossum Founder of Python
- @NateSilver538
- @cutting Founder of Hadoop
- @RProgLangRR
- @BitbucketStatus
- @CWRUITS\_STATUS
- @cshirky Clay Shirky

#### 1.1.1.4.5 Sign up for a Stack Exchange Account

Stack Exchange, Stack Overflow

- are a Q&A community focused on many topics.

Stack Overflow allows you to search by tag

- r and rmarkdown are useful tags for SO

[Stack Exchange's Tour of Stack Overflow](#)

Specific Stack Exchange websites

- for [SX Data Science](#)

- for [SX Statistics on Cross Validated](#)
- for [SX Open Data](#)

#### 1.1.1.4.6 Efficiently browse you SX sites

- Google (but more random)
- [The Stack Exchange apps](#)
- Using an [RSS Feed reader such as Feedly](#) is a good way

#### 1.1.1.4.7 Slack, another component of Agile Software Development

- [Slack.com](#)
  - We have a [CWRU DSCI Slack room](#)
  - There is Slack app for phones
  - And client for computers, its on vdi.
  - Slack client available for windows, mac and Linux
- an online collaboration tool

#### 1.1.1.5 You Online Data Science Portfolio

- Doing open, reproducible data science
- Lets you share a portfolio of codes and projects
- Cite it in your resume
- Build a community of supporters and collaborators

##### 1.1.1.5.1 An Example, Emeline Liu

- [emelineliu.com](#)
  - This website, which runs off of [Github Pages](#) and [Jekyll](#), is my latest project.
  - Right now, I'm using [Poole](#) as a foundation for my website/blog.

#### 1.1.1.6 Links

- <http://www.r-project.org>
- Rory Winston, for the [Learning R Intro](#)
- StackExchange <http://stackexchange.com/sites>
- Twitter <http://twitter.com>
- Slack <http://slack.com>
- [CWRU-DSCI Slack](#)
- [emelineliu.com](#)
- [Github Pages](#)
- [Jekyll](#)
- [Poole](#)