DSCI351-351m-451: Class 01a, (CWRU, Pitt, UCF, UTRGV)

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

01 September, 2022

Contents

1.2.1.1 Class Readings, Assignments, Syllabus Topics . . . 1
1.2.1.1.1 Reading, Lab Exercises, SemProjects . . . 1
1.2.1.1.2 Textbooks . . . 2
1.2.1.1.3 Syllabus . . . 2
1.2.1.1.4 Prof. Laura Bruckman will present in class today, Thursday, on SemProjs . . . 2
1.2.1.2 The Lab Exercises (LEs) . . . 2
1.2.1.3 Where we are at present in Class . . . 2
1.2.1.4 Markov HPC and Open Data Science (ODS) Compute Engines . . . 4
1.2.1.5 What we need to do now . . . 6
1.2.1.5.1 So go make accounts, using your case.edu email address . . . 7
1.2.1.6 Your Open Data Science Tool Chain . . . 7
1.2.1.6.1 Its all about a Data Science Tool Chain . . . 7
1.2.1.6.2 Online Git Server Communities . . . 7
1.2.1.6.3 Kaggle Account . . . 7
1.2.1.6.4 Slack, another component of Agile Sofware Development . . . 8
1.2.1.7 Your Online Data Science Portfolio . . . 8
1.2.1.7.1 Twitter used for Data Science . . . 8
1.2.1.7.2 Sign up for a Stack Exchange Account . . . 9
1.2.1.7.3 Efficiently browse you SX sites . . . 9
1.2.1.7.4 An Example, Emeline Liu . . . 9
1.2.1.8 Links . . . 9

1.2.1.1 Class Readings, Assignments, Syllabus Topics

1.2.1.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today:
  - For next class: Peng R Programming (PRP), p 4-33
- Laboratory Exercises:
  - LE0 : An intro to R exercise, that counts as 0 points
  - LE1 : Given out in Thursday W01b
    - **LE1 due Tuesday Sept. 13th**
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Mondays @ 4:00 PM to 5:00 PM, Will Oltjen
  - Saturday @ 3:00 PM to 4:00 PM, Kristen Hernandez
  - **Office Hours are on Zoom, and recorded**
- Exams

  - MidTerm: Tuesday October 18th, in class or remote, 11:30 - 12:45 PM
  - Final: Monday 12/19/2022, 12:00PM - 3:00PM, Nord 356 or remote

1.2.1.1.2 Textbooks

- Introduction to R and Data Science
  - For R, Coding, Inferential Statistics
    - Peng: R Programming for Data Science
    - Peng: Exploratory Data Analysis with R

Textbooks for this class

- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4
- R4DS = Wickham, Grolemund: R for Data Science

Textbooks for DSCI353/353M/453, And in your Repo now

- ISLR = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R 2nd Ed.
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R

Magazine Articles about Deep Learning

- DL1 to DL13 are "Deep Learning" articles in 3-readings/2-articles/

1.2.1.1.3 Syllabus

1.2.1.1.4 Prof. Laura Bruckman will present in class today, Thursday, on SemProjs

- To give more information on the Semester Projects for DSCI453 students
  - This includes 3 Reports Outs by 453 Students
  - That **all students will view and do peer grading of**

1.2.1.2 The Lab Exercises (LEs)

- Each LE is worth
  - LE1,2 are 7 points
  - LE3-7 are 10 points
    - (except LE0 = 0 points)

So 64 points are in the Lab Exercises

- So these are important and critical to learning
- You will need to start on the early
  - This is why you are given two weeks to do them
- You turn in both the .Rmd and the .pdf file
  - We grade on the .pdf file in Canvas
- We expect good code styling
  - That matches the Google/Rstudio R Style Guide
  - Since this aides collaboration

1.2.1.3 Where we are at present in Class

- So as of today,

We need to make all elements for the ODS tools chain working for you

- You have logged into your CaseID email at http://webmail.case.edu
  - And have setup Duo for Two Factor Authentication (2FA)

Day:Date	**Foundation**	**Practicum**	**Reading**	**Due**
w01a:Tu:8/30/22	ODS Tool Chain	R, Rstudio, Git		
w01b:Th:9/1/22	Setup ODS Tool Chain	Bash, Git, Slack, Agile	PRP4-33	LE1
w02a:Tu:9/6/22	What is Data Science	OIS:Intro2R, Git	PRP35-64	
w02b:Th:9/8/22	Summarizing Data	Intro2R	OIS1,2	
w02Pr:Fr:9/9/22			PRP65-93	**451 Update1**
w03a:Tu:9/13/22	Summarizing Data	Git, Rmds, Loops,	PRP94-116	LE2 **LE1 Due**
w03b:Th:9/15/22	Rand. Var. Normal Dist.	Data Analytic Style	OIS4	
w04a:Tu:9/20/22	Tidy Check Explore	Tidy GapMinder	EDA1-31	
w04b:Th:9/22/22	Inference, DSCI Process	Other Distrib. 7 ways	R4DS1-3	LE3 **LE2 Due**
w04Pr:Fr:9/23/22			EDA32-58	**451 Update2**
w05a:Tu:9/27/22	OIS4 Rand. Var.	EDA of PET Degr.	OIS5	
w05b:Th:9/29/22	OIS5 Found. of Infer.	Multivar Corr. Plot	R4DS4-6	
w05Pr:Fr:9/30/22				**451 RepOut1**
w06a:Tu:10/4/22	Pred., Algorithm, Model	Anscombe's Quartets	R4DS7-8	
w06b:Th:10/6/22	EDA stats, vis	Summ. Stats & Vis.	R4DS9-16	LE4 **LE3 Due**
w06Pr:Fr:10/7/22	Corr. Coeff. Pairs Plots			**451 Update3**
w07a:Tu:10/11/22	Confidence Intervals	Penguins	OIS6.1-2	**PeerRv1 Due**
w07b:Th:10/13/22	Midterm Rev.	Hypo.Test, Sampl. Dist.		
w08a:Tu:10/18/22	**MIDTERM**	**EXAM**		
w08b:Th:10/20/22	Programming & Coding	Coding Expect.		**LE4 Due**
w08Pr:Fr:10/21/22				**451 Update4**
Tu:10/24,25	**CWRU**	**FALL BREAK**	R4DS17-21	
w09b:Th:10/27/22	Cat. Inf. 1 & 2 propor.	Indep. Test,2-way tables	OIS6.3-4	LE5
w09Pr:Fr:10/28/22				**451 RepOut2**
w10a:Tu:11/1/22	Goodness of Fit, χ^2 test	t-tests 1&2 means	OIS7.1-4	
w10b:Th:11/3/22	Num. Infer, Cont. Tables	Stat. Power		
w10Pr:Fr:11/4/22				**451 Update5**
w11a:Tu:11/8/22	Sample & Effect Size	Stat. Power GGmap	OIS8	**PeerRv2 Due**
w11b:Th:11/10/22	Inf. 4 Regr, Test & Train	Curse of Dimen.	ISLR1,2.1,2	LE6 **LE5 Due**
w12a:Tu:11/15/22	Lin. Regr. Part 1	Residuals	OIS9	
w12b:Th:11/17/22	Lin. Regr. Part 2	Regr. Diagnostics		
w12Pr:Fr:11/18/22				**451 Update6**
w13a:Tu:11/22/22	Mult. Lin. Regr.	Var. & Mod. Selec.,	ISLR3.1	LE7 **LE6 due**
w13b:Th:11/24/22	Log. Regr.	GIS Trends	ISLR3.2	
w13Pr:Fr:11/25/22				**451 RepOut3**
w14a:Tu:11/23/22	Classificat., Sup. Lrning	Caret, Broom 4 modeling	ISLR4.1-3	
Th,Fr:11/24,25	**THANKSGIVIING**	**Vacation**		
w15a:Tu:11/29/22		Clustering		**PeerRv3 Due**
w15b:Th:12/1/22	Big Data Analytics	Dist. Comp., Hadoop		
w15SPr:Fr:12/2/22		Read Article by	Mirletz,2015	
w16a:Tu:12/6/22	Final Exam Review			
w15b:Th:12/8/22				**LE7 due**
Friday 12/12	**SemProj**	**Final Report**		**SemProj4 due**
Monday 12/19	**FINAL EXAM**	**12:00-3:00pm**	Nord 356	or remote

Figure 1: DSCI351-351M-451 Syllabus

- You have joined the DSCI Slack
  - At https://cwru-dsci.slack.com
  - Using your CaseID@case.edu email
- You setup a bitbucket.org account
  - using your CaseID email account
  - And have setup your Bitbucket “App Password”
- You have “forked” the 22f-dsci351-451-prof “prof” repo
  - And have change “prof” to your caseID
  - And made your fork in the CWRU-DSCI team
- You have configured your git server
  - on both Markov, in your /home/CaseID/Git folder
    - and on ODS Desktop, in your H:/Git folder
    - and on your personal notebook computer, in a Git folder you make
  - And these configurations define your name and email
    - `git config --global user.name "[name]"`
    - `git config --global user.email "[email address]"`
- Then you want to clone your personal course repo to 3 places
  - Markov/OnDemand: git clone… to /home/CaseID/Git/

  - ODS Desktop/MyApps: git clone… to H:/Git/

  - On your own computer to Git folder (to enable easy reading pdf)

If not, reach out to the TAs (Will Oltjen, Krisen Hernandez, Mingxuan Li)

- Using the http://cwru-dsci.slack.com
  - Which you can join directly using your CaseID@case.edu email address
- Defining where you issue is
- And we’ll fix it

1.2.1.4 Markov HPC and Open Data Science (ODS) Compute Engines

- You can do data analysis on your notebook computer

  - You can setup your own notebook
    - For data science using R or Python
    - Full instructions are in the class syllabus
      - Section 11
    - For Linux, Mac’s or Windows Operating Systems
    - But Many times you’ll need more compute power than your notebook
      - Such as GPUs (Graphics Processing Units) to accelerate computations

But its useful to learn about a variety of Compute Resources

- In Class we’ll use
  - Markov Data Science Cluster
    - A high performance computing cluster
    - via http://ondemand.case.edu
  - or Open Data Science Desktops
    - A Win10 cloud desktop
    - via http://myapps.case.edu These are all configured the same
- Independent of the Operating System
- They have R with Rstudio IDE (Integrated Development Environment)
- Git for code versioning
- LaTeX for publication quality report generation
- And also Python3 with VS Codium or PyCharm IDE

The two cloud computing systems: Markov HPC Cluster & ODS Win10 Desktop

- Markov Data Science HPC Compute Cluster, via OnDemand
  - Log in to http://ondemand.case.edu
  - Using your CaseID and password
  - Launch the Rstudio Server (rxf131)
    - Which runs R version 4.2.1
  - You can also get an LXDE graphical desktop on Markov

CWRU HPC provides Markov

- CWRU's HPC (High Peformance Computing) Markov Cluster
  - This runs RedHat Linux version 7
  - Has 4400 CPU cores
  - Has 100,000 GPU cores
  - Up to a terabyte of Ram
- And has a new Data Science Cluster, named Markov.case.edu
  - With a Hadoop Cluster for distributed computing
  - And dedicated GPUs
- You'll get accounts on CWRU HPC
- And use http://ondemand.case.edu
  - To login to Markov and get a Rstudio Server (rxf131) session
  - Or a LXDE graphical desktop session
    - for simple file operations or a browser

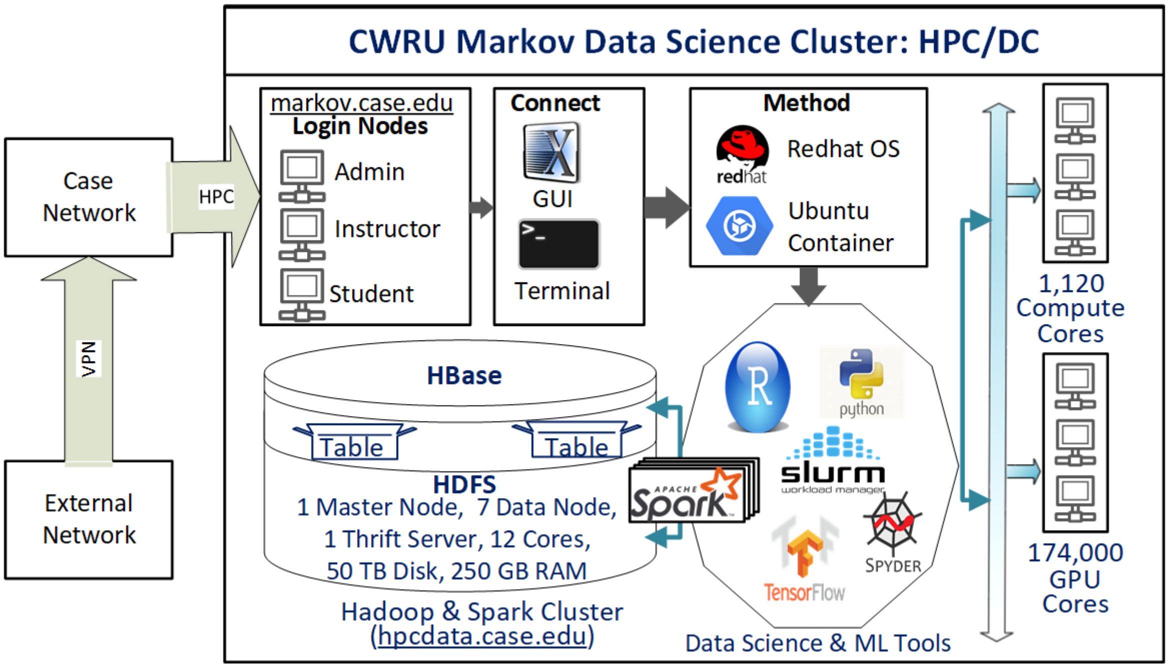

Figure 2: Markov Cluster

- You also have access to the ODS Win10 Desktops
  - These are cloud Windows computers
    - That you log into from a Browser
    - login to http://myapps.case.edu
    - With your CaseID and password
  - The ODS VDIs are Windows 10 computers

  - The ODS VDIs don't have GPUs

Not for class, but for your own data science projects.

And you can use Google's Collaboratory](https://colab.research.google.com/notebooks/welcome.ipynb)

- For Jupyter Notebooks
- Running Python3
- Doesn't support R language yet
- Free GPUs and TPUs (Tensor Processing Unit)

1.2.1.5 What we need to do now

- Setup our Markov and Open Data Science (ODS) Computers

  1. For Markov Data Science Cluster

  - login to `http://ondemand.case.edu` with your CaseID account
  - Launch the SDLE Rstudio Server (rxf131)
  - Check your "Library Paths"
    - in the R console
    - run `.libPaths()`
    - And the first directory MUST be
    - "/home/rxf131/ondemand/ubuntu2004/r4" "/usr/local/lib/R/site-library"
  - otherwise refer to the file in the root directory of your repo
    - named `FixRstudioServer-R-libPaths.txt`
    - and run the "source('/home/rxf131/ondemand/share/config/r-lib-path-fix.R')'
    - In the R console
    - then check your `.libPaths()` again
  - On Markov, launch `LXDE Desktop (rxf131)`
    - make a Git folder under `/home/CaseID/`

    - Login to DSCI Slack in your firefox browser on LXDE desktop

  2. For the ODS Desktop

  - login to `http://myapps.case.edu` with your CaseID account
  - Drag icons of to your desktop
    - for R, Rstudio, Git Bash, VScodium, PyCharm, Jupyter Notebook, Slack

  3. Setup Git

  - make /home/caseID/Git folder on Markov
    - git config your name and email of your git server
  - make H:\Git folder on ODS Desktop
    - git config your name and email of your git server

  4. Git Fork the Class "Prof" Repo

  - In your Bitbucket Account

  5. Git Clone your Fork of the Class Repo
  6. When in Rstudio (on Markov or ODS)

  - Its ESSENTIAL that you open the .Rproj file in the upper right corner
  - this tells Rstudio where your root directory of your project is.

  7. Setup Bitbucket account
  8. Setup DSCI Slack Account
  9. Setup StackExchange account

1.2.1.5.1 So go make accounts, using your case.edu email address

- Most students have already been invited
  - Pitt, UCF, UTRGV students have been issued CaseIDs
  - That you will use for logging in to
    - case email: at http://webmail.case.edu
    - Markov
    - ODS Desktop
    - DSCI Slack
    - CWRU Canvas
- Our DSCI Slack class channel
  - CWRU Data Science Slack
  - This is an invite link to CWRU DSCI Slack
- For you cloud Git server
  - Bitbucket.org
- A Stack Exchange account

1.2.1.6 Your Open Data Science Tool Chain

1.2.1.6.1 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

1.2.1.6.2 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.

1.2.1.6.3 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

1.2.1.6.4 Slack, another component of Agile Sofware 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

1.2.1.7 Your 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

1.2.1.7.1 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
- @robjhyndman
- @geoffreyhinton
- @ylecun
- @fchollet
- @TensorFlow
- @JeffDean
- @yudapearl
- @AndrewYNg

1.2.1.7.2 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

1.2.1.7.3 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

1.2.1.7.4 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.

1.2.1.8 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
- Kaggle.com

- Colaboratory