DSCI351-351m-451: Class 01a, ODS Tool Chain (CWRU, Pitt, UCF)

Profs: R. H. French, Pawan Tripath, P. Leu, M. Li, K. Davis

TAs: Redad Medhi, Olatunde Akanbi

r format(Sys.time(), '%d %B, %Y')

{r setup, include = FALSE} knitr::opts\_chunk$set( cache = FALSE, # if TRUE knitr will cache results to reuse in future knits fig.width = 5, # the width for plots created by code chunk fig.height = 4, # the height for plots created by code chunk fig.align = 'left', # how to align graphics. 'left', 'right', 'center' dpi = 300, dev = 'png', # Makes each fig a png, and avoids plotting every data point # eval = FALSE, # if FALSE, then the R code chunks are not evaluated # results = 'asis', # knitr passes through results without reformatting echo = TRUE, # if FALSE knitr won't display code in chunk above it's results message = TRUE, # if FALSE knitr won't display messages generated by code strip.white = TRUE, # if FALSE knitr won't remove white spaces at beg or end of code chunk warning = TRUE, # if FALSE knitr won't display warning messages in the doc error = TRUE) # report errors # options(tinytex.verbose = TRUE)

#### Class Readings, Assignments, Syllabus Topics

##### Reading, Lab Exercises, SemProjects

* Readings:
  + For today:
  + For next class: Peng R Programming, pages 4-33
* Laboratory Exercises:
  + LE1 : Given out Thursday
  + LE1 : Is Due Tuesday Sept. 10th
* Office Hours: (Class Canvas Calendar for Zoom Link)
  + Mondays @ 5:00 PM to 6:00 PM, Redad Medhi
  + Wednesday @ 3:00 PM to 5:00 PM, Olatunde Akanbi
  + **Office Hours are on Zoom, and recorded**
* Semester Projects
  + DSCI 451 Students Biweekly Update Due
  + DSCI 451 Students
    - Next **Report Out #4 (Full Report) is Due December 12th**
  + All DSCI 351/351M/451 Students:
    - **Peer Grading of Report Out # is Due**
  + SemProject Office Hours: (Class Canvas Calendar for Zoom Link)
    - Prof. Pawan Tripathi guides the SemProjs for DSCI451 students
      * @Abhishek Daundkar aad157@case.edu
    - SemProject Office Hours. (twice a week during the first month)
      * Tuesdays @ 3:00 PM
      * Thursdays @ 2:00 PM
* Exams
  + Final: Monday December 16, 2024, 12:00PM - 3:00PM, Nord 356 or remote

##### 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/

##### Syllabus

DSCI351-351M-451 Syllabus

DSCI351-351M-451 Syllabus

##### Prof. Pawan Tripath will present in class 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**
  + The SemProj TA is Abhishek Daundakar
    - White 540, aad157@case.edu

#### 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

#### Where we are at present in Class

getting started

getting started

And getting familiar with data science

getting started

getting started

* 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)
* 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 24f-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 ( Redad Medhi, Hein Htet Aung, Nicole Lipa )

* 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

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

ODS Infrastructure

ODS Infrastructure

* 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 4.4.1
  + You can also get an XFCE graphical desktop on Markov

CWRU HPC provides Markov

* [CWRU’s HPC (High Peformance Computing) Markov Cluster](https://sites.google.com/a/case.edu/hpcc/)
  + 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](https://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 xfce graphical desktop session
    - for simple file operations or a browser

Markov Cluster

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)](https://en.wikipedia.org/wiki/Tensor_processing_unit)

#### 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 XFCE Desktop (rxf131)
    - make a Git folder under /home/CaseID/
    - Login to DSCI Slack in your firefox browser on XFCE desktop
  1. 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
  1. 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
  1. Git Fork the Class “Prof” Repo
  + In your Bitbucket Account
  1. Git Clone your Fork of the Class Repo
  2. 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.
  1. Setup Bitbucket account
  2. Setup [DSCI Slack Account](http://cwru-dsci.slack.com)
  3. Setup StackExchange account

##### 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](http://cwru-dsci.slack.com)
  + This is [an invite link to CWRU DSCI Slack](https://join.slack.com/t/cwru-dsci/shared_invite/zt-gy4t8444-WZifMnxeReZRw8q~uJUxpQ)
* For you cloud Git server
  + [Bitbucket.org](http://bitbucket.org)
* A [Stack Exchange account](https://stackexchange.com/)

#### Your Open Data Science Tool Chain

ODS Infrastructure

ODS Infrastructure

##### 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](https://creativecommons.org/)
  + Or other [Open Source Licenses](https://en.wikipedia.org/wiki/Comparison_of_free_and_open-source_software_licenses)
  + Such as the [Gnu Public License: GPL](https://en.wikipedia.org/wiki/GNU_General_Public_License)
  + Or one of my favorites, [the Apache License](https://en.wikipedia.org/wiki/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

##### Online Git Server Communities

* We use [BitBucket Account](http://bitbucket.org)
  + In class, for our class code repositories
  + These are private repositories
* You’ll probably also want a [GitHub](http://github.com/) account.
  + Many Rprojects are there, and
  + you can fork their repo’s as inspect the code very easily.

ODS Infrastructure

ODS Infrastructure

##### Slack, another component of Agile Sofware Development

* [Slack.com](https://slack.com/)
  + We have a [CWRU DSCI Slack room](http://cwru-dsci.slack.com)
  + 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

#### 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

##### 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 Stack Overflow (SO)

###### [Stack Exchange’s Tour of Stack Overflow](http://stackoverflow.com/tour)

###### Specific Stack Exchange websites

* for [SX Data Science](https://datascience.stackexchange.com/)
* for [SX Statistics on Cross Validated](http://stats.stackexchange.com/)
* for [SX Open Data](http://opendata.stackexchange.com/)

##### Efficiently browse you SX sites

* Google (but more random)
* [The Stack Exchange apps](http://stackapps.com/)
* Using an [RSS Feed reader such as Feedly](http://feedly.com/i/latest) is a good way

##### An Example, Emeline Liu

* [emelineliu.com](http://emelineliu.com/)
  + This website, which runs off of [Github Pages](https://pages.github.com/) and [Jekyll](https://jekyllrb.com/), is my latest project.
  + Right now, I’m using [Poole](http://getpoole.com/) as a foundation for my website/blog.

#### Links

* http://www.r-project.org
* Rory Winston, for the [Learning R Intro](http://www.theresearchkitchen.com/archives/1017)
* StackExchange <http://stackexchange.com/sites>
* Slack <http://slack.com>
* [CWRU-DSCI Slack](http://cwru-dsci.slack.com)
* [emelineliu.com](http://emelineliu.com/)
* [Github Pages](https://pages.github.com/)
* [Kaggle.com](https://kaggle.com)
* [Colaboratory](https://colab.research.google.com/notebooks/welcome.ipynb)