DSCI353-353m-453: Class 01a Intro Class

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

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

19 January, 2023

Contents

1.1.1.1 Class Readings, Assignments, Textbooks Syllabus Topics

1.1.1.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today:
  - For next class: ISRL1,2 (R4DS)
- Laboratory Exercises:
  - LE0 : Do this as a refresher
  - LE1 : Given out next Tuesday Jan. 24th
  - LE2 : Is Due Thursday Feb. 2nd
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesdays @ 4:00 PM to 5:00 PM
  - Saturdays @ 3:00 PM to 4:00 PM
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - DSCI 453 Students Biweekly Updates Due
    - Update #1 is Due **Friday Jan. 27th**
  - DSCI 453 Students
    - Next Report Out #1 is Due **Friday Feb. 17th**
  - All DSCI 353/353M/453, E1453/2453 Students:
    - Peer Grading of Report Out #1 is Due **Thursday March 2nd**
  - Exams
    - MidTerm: **Thursday March 9th**, in class or remote, 11:30 - 12:45 PM
    - Final: **Thursday May 4th**, 2023, 12:00PM - 3:00PM, Nord 356 or remote

1.1.1.1.2 Textbooks -Text Books for DSCI353/353M/453

- R4DS: Wickham: R for Data Science
- ISLR: Intro to Statistical Learning with R, 2nd Ed.
- DLwR: Deep Learning with R, Chollet, Allaire,
- DLGB: Deep Learning, Goodfellow, Bengio, Courville
- Magazine Articles about Deep Learning
  - DL1 to DL12 are "Deep Learning" articles in 3-readings/2-articles/
- Books from DSCI351/351M/451
  - Peng: R Programming for Data Science
  - Peng: Exploratory Data Analysis with R
  - Open Intro Stats, v4
  - R4DS: Wickham: R for Data Science

1.1.1.1.3 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet
  - **Tidyverse For Beginners Cheatsheet**
    - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder

  - **Data Wrangling with dplyr and tidyr Cheatsheet**

  Tidyverse Functions & Conventions

  - The pipe operator `%>%`
  - Use `dplyr::filter()` to subset data row-wise.
  - Use `dplyr::arrange()` to sort the observations in a data frame
  - Use `dplyr::mutate()` to update or create new columns of a data frame
  - Use `dplyr::summarize()` to turn many observations into a single data point
  - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
  - Use `dplyr::select()` to choose variables from a tibble,
    - keeps only variables you mention
  - Use `dplyr::rename()` keeps all the variables and renames variables
    - rename(iris, petal_length = Petal.Length)
  - These can be combined using `dplyr::group_by()`
    - which lets you perform operations "by group".
  - The `%in%` matches conditions provided by a vector using the c() function
  - The **forcats** package has tidyverse functions
    - for factors (categorical variables)
  - The **readr** package has tidyverse functions
    - to read_…, melt_… col_…, parse_… data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

1.1.1.1.4 Syllabus

1.1.1.2 The DSCI courses and class sections

1.1.1.2.1 In these Applied Data Science (DSCI) classes

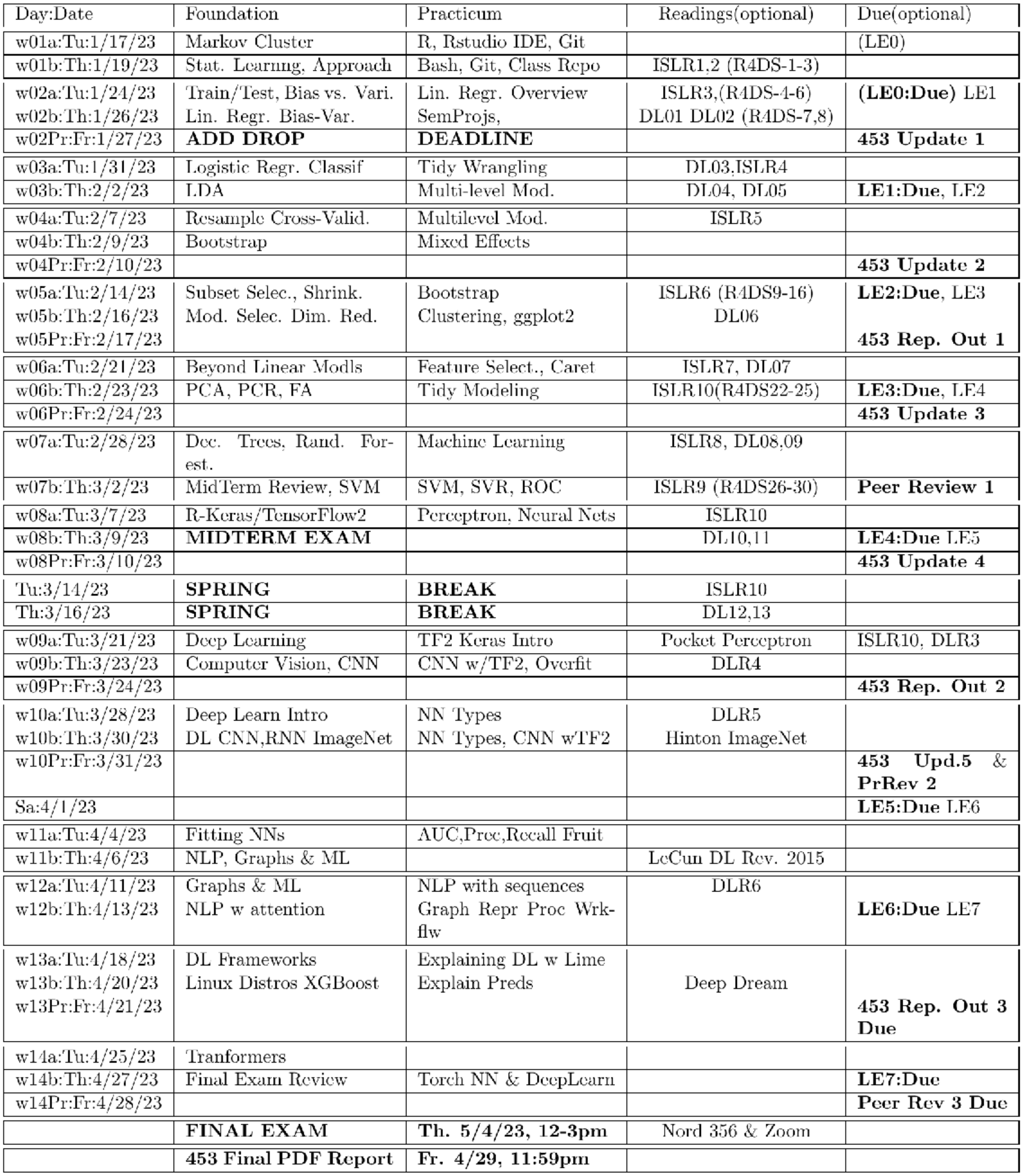
- We focus on teaching all necessary data science skills
  - Including coding in R
  - Use of Rmarkdown for data analysis reports and presentations
  - Git for code versioning and collaboration
  - Linear and non-linear regression and classification
  - Beyond linear modeling, including Support Vector Machines, Random Forest
  - Machine Learning, including Neural Networks, non-parametric regression
  - Deep Learning, including Keras/TensorFlow running on GPUs

1.1.1.2.2 The course sections

- DSCI35x (x = 1,3,2)
  - Is undergraduate class for "general" applied data science
- DSCI35xM (x=1,2,3) focuses on materials science systems
- DSCI45x (x=1,2,3)
  - Is a graduate level class
  - With the same class material and DSCI35x
  - Additionally the students do a 40 point Semester Data Analysis Project

And we have University of Pittsburgh (Pitt) students

- Taken E1453 and E2453
- With Prof. Paul Leu and TA Mingxuan Li



Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23	Markov Cluster	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/19/23	Stat. Learnng, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Train/Test, Bias vs. Vari.	Lin. Regr. Overview	ISLR3,(R4DS-4-6)	**(LE0:Due)** LE1
w02b:Th:1/26/23	Lin. Regr. Bias-Var.	SemProjs,	DL01 DL02 (R4DS-7,8)	
w02Pr:Fr:1/27/23	**ADD DROP**	**DEADLINE**		**453 Update 1**
w03a:Tu:1/31/23	Logistic Regr. Classif	Tidy Wrangling	DL03,ISLR4	
w03b:Th:2/2/23	LDA	Multi-level Mod.	DL04, DL05	**LE1:Due**, LE2
w04a:Tu:2/7/23	Resample Cross-Valid.	Multilevel Mod.	ISLR5	
w04b:Th:2/9/23	Bootstrap	Mixed Effects		
w04Pr:Fr:2/10/23				**453 Update 2**
w05a:Tu:2/14/23	Subset Selec., Shrink.	Bootstrap	ISLR6 (R4DS9-16)	**LE2:Due**, LE3
w05b:Th:2/16/23	Mod. Selec. Dim. Red.	Clustering, ggplot2	DL06	
w05Pr:Fr:2/17/23				**453 Rep. Out 1**
w06a:Tu:2/21/23	Beyond Linear Modls	Feature Select., Caret	ISLR7, DL07	
w06b:Th:2/23/23	PCA, PCR, FA	Tidy Modeling	ISLR10(R4DS22-25)	**LE3:Due**, LE4
w06Pr:Fr:2/24/23				**453 Update 3**
w07a:Tu:2/28/23	Dec. Trees, Rand. Forest.	Machine Learning	ISLR8, DL08,09	
w07b:Th:3/2/23	MidTerm Review, SVM	SVM, SVR, ROC	ISLR9 (R4DS26-30)	**Peer Review 1**
w08a:Tu:3/7/23	R-Keras/TensorFlow2	Perceptron, Neural Nets	ISLR10	
w08b:Th:3/9/23	**MIDTERM EXAM**		DL10,11	**LE4:Due** LE5
w08Pr:Fr:3/10/23				**453 Update 4**
Tu:3/14/23	**SPRING**	**BREAK**	ISLR10	
Th:3/16/23	**SPRING**	**BREAK**	DL12,13	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro	Pocket Perceptron	ISLR10, DLR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLR4	
w09Pr:Fr:3/24/23				**453 Rep. Out 2**
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLR5	
w10b:Th:3/30/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2	Hinton ImageNet	
w10Pr:Fr:3/31/23				**453 Upd.5 & PrRev 2**
Sa:4/1/23				**LE5:Due** LE6
w11a:Tu:4/4/23	Fitting NNs	AUC,Prec,Recall Fruit		
w11b:Th:4/6/23	NLP, Graphs & ML		LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrkflw		**LE6:Due** LE7
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				**453 Rep. Out 3 Due**
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		**LE7:Due**
w14Pr:Fr:4/28/23				**Peer Rev 3 Due**
	FINAL EXAM	**Th. 5/4/23, 12-3pm**	Nord 356 & Zoom	
	453 Final PDF Report	**Fr. 4/29, 11:59pm**		

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y, DLGBx.y refers to chapters and sections assigned as reading in our textbooks. DLx are deep learning articles.

Figure 1: DSCI351-351M-451 Syllabus

1.1.1.2.3 The specific courses

- DSCI351, 351M, 451
  - Is an introduction to Exploratory Data Science
- DSCI353, 353M, 453
  - Focuses on Modeling, Prediction and Machine Learning
- DSCI 352, 352M, 452
  - Is a Semester long Data Science Project Class
  - Providing a data analysis for inclusion
  - In your Data Science Portfolio
- DSCI 354, 354M, 454
  - Is on Data Visualization and Analytics
  - Alternative Level 5 course for the ADS UG Minor

1.1.1.2.4 DSCI45x Graduate level courses

- For graduate students,
  - DSCI451 is not listed as a suggested prerequisite
- Therefore some DSCI453 grad. students
  - Do not have familiarity with Open Data Science, R, Git etc.
- For these "New to R" students
  - The initial weeks in class have optional content
  - To get people familiar with Open Data Science

1.1.1.2.5 Semester Data Science Projects

- Are done in DSCI352, 352M by students who have completed both DSCI351,3
- And by graduate students in DSCI 451, 453 and 452

For DSCI45x students, their Semester Project is developed in the DSCI352 course

- With Prof. Laura Bruckman
- During team meetings during Friday Community Hour
  - 12:45 to 1:45 in Olin 303
- And during class office hours
  - Monday/Wednesday 4pm to 5pm in White 540
- There are weekly SemProj updates due each week on progress
- And 3 SemProj Presentations in DSCI35x class

1.1.1.2.6 For the DSCI 453 students they have an EDA SemProj to do

- SemProjects:
  - SemProjects have a 5 progress update
    - due Friday's at 11:59 pm (5 updates)
  - Each update should be made in the report template
    - found in the Repo with each update filled out with the new things in the document
  - the update helps TA and professor grade and follow you project
  - The document should be filled in under each section and update throughout the semester until the final written report
  - SemProj Report Out #1 Class W5, (recorded 10 min presentation)
    - Peer Grading by All DSCI 353/353m/453 students due on syllabus
  - SemProj Report Out #2 in Class W9 (recorded 10 min presentation)
    - Peer Grading by All DSCI 353/353m/453 students due on syllabus
  - SemProj Report Out #3 in Class W13 (recorded 10 min presentation)
    - Peer Grading by All DSCI 353/353m/453 students due on syllabus
  - SemProj Report is full comprehensive written project

    - * (report template updated from each report)
- Assistance on SemProjects is done with DSCI353-353m-453 Class
  - SemProj's are taught by Prof. Laura Bruckman
  - SemProject office hours 9-10 am on Tuesdays

1.1.1.2.7 Care should be taken when choosing SemProj datasets.

- Report Out 1 focuses on
  - Explaining the 'why' of your research project
  - Describing your dataset
  - Presenting an analysis plan
  - Cleaning your data
- Report Out 2 focuses on:
  - EDA of your data
  - Visualizing your data
  - Further cleaning of your data
  - Reevaluation of your data analysis plan (Do you need more data?)
- Report Out 3:
  - More data visualization
  - Initial modeling
  - Conclusions about your data
  - Were you able to answer your why question?
  - What else would you need to do to get to understanding your data better?

1.1.1.3 Syllabus

1.1.1.4 Open Data Science (ODS) & HPC 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

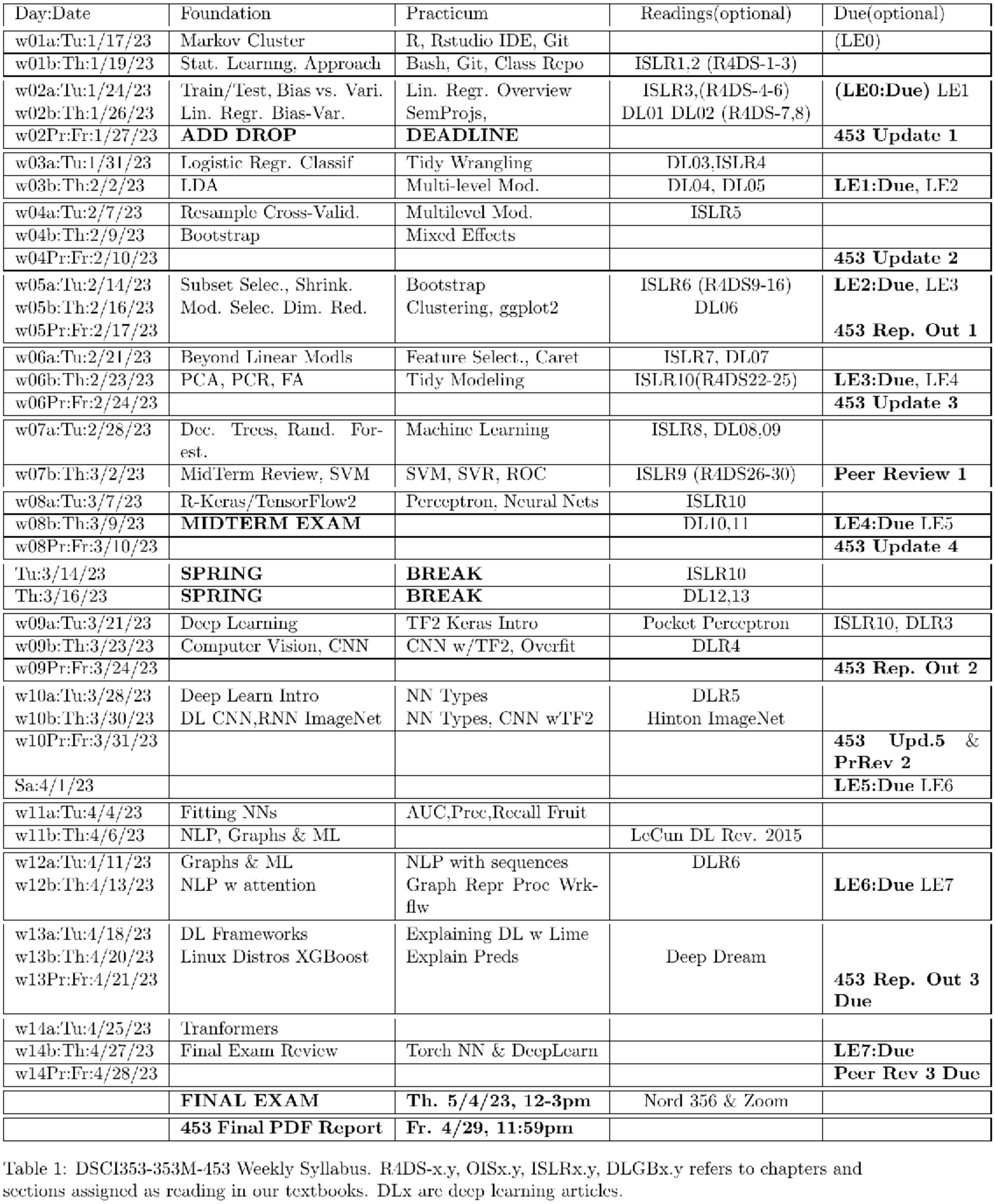
- In Class we'll use
  - Markov Data Science Cluster
    - A high performance computing cluster
  - or Open Data Science Desktops

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 PyCharm IDE

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

- Markov Data Science HPC Cluster
  - Log in to http://ondemand.case.edu
  - Using your CaseID and password



Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23	Markov Cluster	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/19/23	Stat. Learnng, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Train/Test, Bias vs. Vari.	Lin. Regr. Overview	ISLR3,(R4DS-4-6)	**(LE0:Due)** LE1
w02b:Th:1/26/23	Lin. Regr. Bias-Var.	SemProjs,	DL01 DL02 (R4DS-7,8)	
w02Pr:Fr:1/27/23	**ADD DROP**	**DEADLINE**		**453 Update 1**
w03a:Tu:1/31/23	Logistic Regr. Classif	Tidy Wrangling	DL03,ISLR4	
w03b:Th:2/2/23	LDA	Multi-level Mod.	DL04, DL05	**LE1:Due**, LE2
w04a:Tu:2/7/23	Resample Cross-Valid.	Multilevel Mod.	ISLR5	
w04b:Th:2/9/23	Bootstrap	Mixed Effects		
w04Pr:Fr:2/10/23				**453 Update 2**
w05a:Tu:2/14/23	Subset Selec., Shrink.	Bootstrap	ISLR6 (R4DS9-16)	**LE2:Due**, LE3
w05b:Th:2/16/23	Mod. Selec. Dim. Red.	Clustering, ggplot2	DL06	
w05Pr:Fr:2/17/23				**453 Rep. Out 1**
w06a:Tu:2/21/23	Beyond Linear Modls	Feature Select., Caret	ISLR7, DL07	
w06b:Th:2/23/23	PCA, PCR, FA	Tidy Modeling	ISLR10(R4DS22-25)	**LE3:Due**, LE4
w06Pr:Fr:2/24/23				**453 Update 3**
w07a:Tu:2/28/23	Dec. Trees, Rand. Forest.	Machine Learning	ISLR8, DL08,09	
w07b:Th:3/2/23	MidTerm Review, SVM	SVM, SVR, ROC	ISLR9 (R4DS26-30)	**Peer Review 1**
w08a:Tu:3/7/23	R-Keras/TensorFlow2	Perceptron, Neural Nets	ISLR10	
w08b:Th:3/9/23	**MIDTERM EXAM**		DL10,11	**LE4:Due** LE5
w08Pr:Fr:3/10/23				**453 Update 4**
Tu:3/14/23	**SPRING**	**BREAK**	ISLR10	
Th:3/16/23	**SPRING**	**BREAK**	DL12,13	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro	Pocket Perceptron	ISLR10, DLR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLR4	
w09Pr:Fr:3/24/23				**453 Rep. Out 2**
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLR5	
w10b:Th:3/30/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2	Hinton ImageNet	
w10Pr:Fr:3/31/23				**453 Upd.5 & PrRev 2**
Sa:4/1/23				**LE5:Due** LE6
w11a:Tu:4/4/23	Fitting NNs	AUC,Prec,Recall Fruit		
w11b:Th:4/6/23	NLP, Graphs & ML		LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrkflw		**LE6:Due** LE7
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				**453 Rep. Out 3 Due**
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		**LE7:Due**
w14Pr:Fr:4/28/23				**Peer Rev 3 Due**
	FINAL EXAM	**Th. 5/4/23, 12-3pm**	Nord 356 & Zoom	
	453 Final PDF Report	**Fr. 4/29, 11:59pm**		

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y, DLGBx.y refers to chapters and sections assigned as reading in our textbooks. DLx are deep learning articles.

Figure 2: Modeling, Prediction and Machine Learning Syllabus

  - Launch the SDLE Rstudio Server-4.1.1
  - You can also get a KDE Desktop on Markov

1.1.1.4.2 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 KDE Desktop session

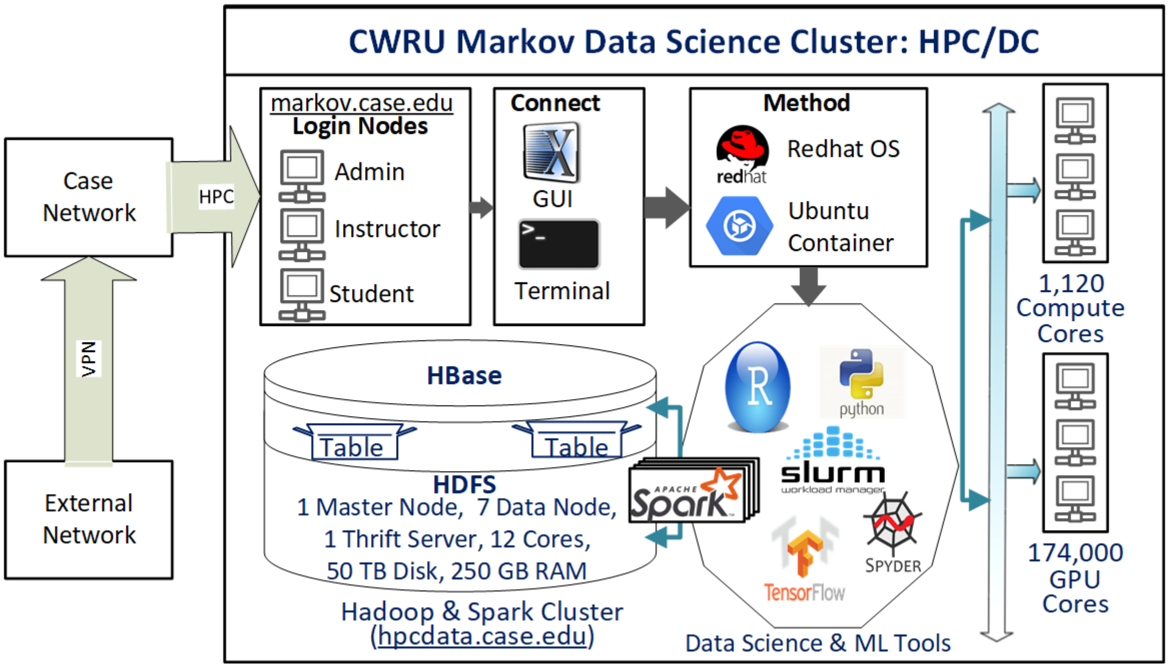


Figure 3: Markov Cluster

1.1.1.4.3 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.

1.1.1.4.4 And you can also use Google's Kaggle.com

- Here one can run R or Python

- Using Jupyter Notebooks Interface
- Has Free GPUs

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)

1.1.1.5 Operating Systems: Windows, OSX and Linux

- Command Line Environments
  - Linux: Bash on Linux, or Git Bash on Windows
  - Mac OSX: Bash in Terminal
  - Windows: Command.com Terminal
  - In R: R Console, or Console in RStudio

Item	Linux OS	X Mac Wi	ndows
folder demarcation	/	/	"\" don't use
directory listing	ls	ls	dir
present work. dir	pwd	pwd	
change directory	cd	cd	cd
drives	root	root	drive letters
NO SPACES in	filenames	spaces	don't work

1.1.1.5.1 Basic/Universal Rules

- No Spaces in Filenames
- Only 1 period in a filename, before file extension
- No other periods
- Only Letters, Underscore (_), and Dashes (-) in Filenames
- In code scripts, use forward slash in all file paths and directories
- You can use CamelBack or snake_case in variable or file names
  - To make code easier to read.
- Code Style is Rstudio or Google R style
- No use of = for Assignments
- Only use <- as the Assignment Operator in R
  - Rstudio Cheat Sheet says <- is "Alt -" in R code

1.1.1.6 Quick Introduction to R/Rstudio/Git

- R is the statistical programming language

Rstudio is the Integrated Development Environment (IDE)

Git is the distributed content versioning system

1.1.1.7 What we need to do this week

- 1. Setup our Markov and Open Data Science (ODS) Computers
  - For Markov Data Science Cluster
    - login to http://ondemand.case.edu

        - Launch the SDLE Rstudio Server-4.1.1
    - For the ODS Desktop
        - Rstudio
        - Drag icons of R, Rstudio, Git Bash, Spyder, Jupyter Notebook, DSCI Slack
            - to desktop

2. 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

3. Setup Bitbucket account
4. Setup DSCI Slack Account
5. Setup StackExchange account
6. Git Fork the Class "Prof" Repo

- In your Bitbucket Account

8. Git Clone your Fork of the Class Repo

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

- Most students have already been invited
    - Pitt students have been issued CaseIDs
        - That you will use for logging in to
        - 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

1.1.1.8 Your Open Data Science Tool Chain

1.1.1.8.1 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

1.1.1.8.2 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.

1.1.1.9 Things you need to do

1.1.1.9.1 Online accounts

- Sign up for our Class Slack with your personal or case.edu email
- Sign up for a bitbucket.org account
  - with your case.edu address

- Sign up for a twitter account,
  - then follow @frenchrh, @hadleywickham, @dataandme, @JennyBryan
  - @minebocek, @juliasilge, @rdpeng, @jtleek, @robjhyndman, @daniela_witten
  - and others as you want, such as
  - @fchollet, @TensorFlow, @ylecun, @GoogleAI, @egorzakharovdl
- Sign up for a stack overflow account on stack exchange

1.1.1.9.2 Lab Exercises are submitted and graded on Canvas

- Assignment turn in pages will be posted when LE are given out.

1.1.1.9.3 Your Class Git Repo

- My "Professor" Repo is 20s-dsci353-353m-453-prof
  - On bitbucket, you will fork this repo to your own account
  - Each day prior to class, update your fork from my prof. repo

1.1.1.10 Intro to some R: Data Types

- Primitives (numeric, integer, character, logical, factor)
- Data Frames
- Lists
- Tables
- Arrays
- Environments
- Others (functions, closures, promises..)

```
x <- 1
class(x)
## [1] "numeric"
y <- "Hello World"
class(y)
## [1] "character"
```

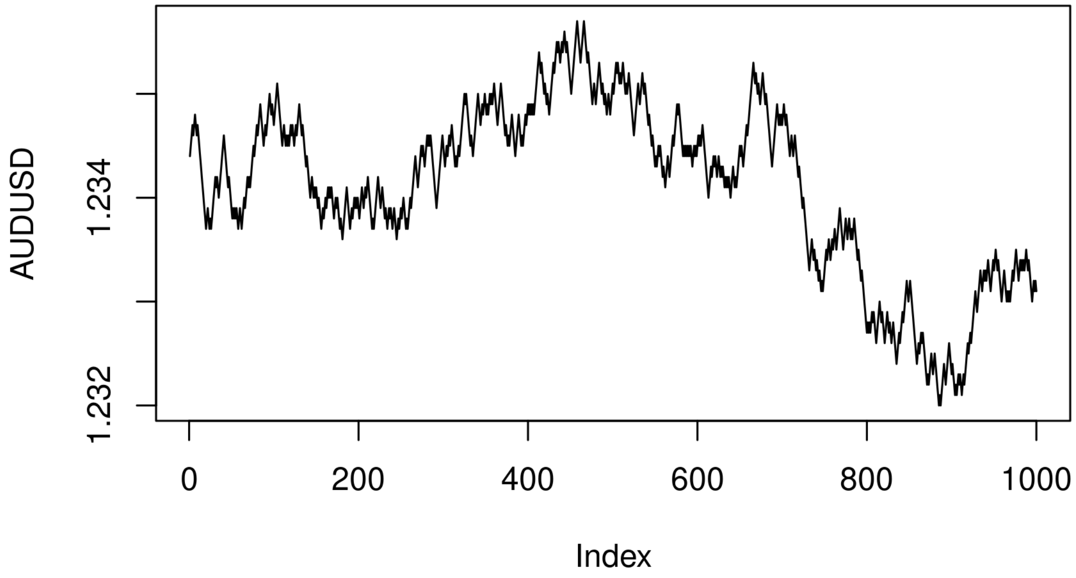
```
z <- TRUE
class(z)
## [1] "logical"
as.integer(z)
## [1] 1
```

1.1.1.10.1 Simple Types

```

randomWalk <- function(N)(cumsum(ifelse(rbinom(prob = 0.5, size = 1, N) == 0,-1,1)))
AUDUSD <- 1.2345 + randomWalk(1000)*.0001
plot(AUDUSD, type = 'l')
```

1.1.1.10.2 Example: Generating Random Data



1.1.1.11 Recommended R Libraries

- We're running R 4.1.2, named "Bird Hippie"

All our "Standard R Packages" are loaded in the Markov and the ODS Desktop

1.1.1.11.1 Basic useful packages (and many more than this)

- Rcpp - Convenient C++ interface
- zoo/xts - Time series libraries
- Matrix - Enhanced matrix library

1.1.1.11.2 Hadley Wickham Tidyverse packages

- This is the content of R for Data Science (R4DS) book.
  - Using Pipes "%>%" to replace loops
  - Makes syntax more compact and readable
  - Makes code faster
- Tidyverse Style Guide
  - Using tidy dataframes
- ggplot2 - Mini-DSL (domain specific language) for data visualization
- plyr/reshape - Data reshaping/manipulation

- dplyr
- data.table - Faster data.frame manipulation
- knitr - for markdown processing
- among others like purrr etc.

1.1.1.11.3 Statistical and Machine Learning

- e1071 Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier etc (142479 downloads)
- MASS tools for variable selection etc.
- rpart Recursive Partitioning and Regression Trees. (135390)
- igraph A collection of network analysis tools. (122930)
- nnet Feed-forward Neural Networks and Multinomial Log-Linear Models. (108298)
- randomForest Breiman and Cutler's random forests for classification and regression. (105375)
- caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. (87151)
- kernlab Kernel-based Machine Learning Lab. (62064)
- glmnet Lasso and elastic-net regularized generalized linear models. (56948)
- ROCR Visualizing the performance of scoring classifiers. (51323)
- gbm Generalized Boosted Regression Models. (44760)
- party A Laboratory for Recursive Partitioning. (43290)
- arules Mining Association Rules and Frequent Itemsets. (39654)
- tree Classification and regression trees. (27882)
- klaR Classification and visualization. (27828)
- RWeka R/Weka interface. (26973)
- ipred Improved Predictors. (22358)
- lars Least Angle Regression, Lasso and Forward Stagewise. (19691)
- earth Multivariate Adaptive Regression Spline Models. (15901)
- CORElearn Classification, regression, feature evaluation and ordinal evaluation. (13856)
- mboost Model-Based Boosting. (13078)

1.1.1.11.4 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
- @robjhyndman
- @geoffreyhinton
- @ylecun
- @fchollet
- @TensorFlow
- @JeffDean
- @yudapearl
- @AndrewYNg

1.1.1.12 Links

http://www.r-project.org

Rory Winston, for the Learning R intro http://www.theresearchkitchen.com/archives/1017

R for Data Science http://r4ds.had.co.nz/

- Or pull the R4DS repo from Bitbucket https://bitbucket.org/cwrudsci/r4ds
- Peng-Computing For Data Analysis Playlist