DSCI353-353m-453: Class 02a-p Open Data Science Tool Chain

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

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

24 January, 2023

Contents

2.1.3.1 Class Readings, Assignments, Textbooks Syllabus Topics . . . 1
2.1.3.1.1 Reading, Lab Exercises, SemProjects . . . 1
2.1.3.1.2 Syllabus . . . 2
2.1.3.2 Current Status of Everyone in Class . . . 2
2.1.3.2.1 Prof. Laura Bruckman will present in class Today . . . 2
2.1.3.3 The Lab Exercises (LEs) . . . 4
2.1.3.4 Literate Programing: Donald Knuth . . . 4
2.1.3.4.1 Literature Programming, was another of his goals . . . 4
2.1.3.5 Agile Software Development . . . 5
2.1.3.6 Your Open Data Science Tool Chain . . . 5
2.1.3.6.1 Its all about a Data Science Tool Chain . . . 5
2.1.3.6.2 Twitter used for Data Science . . . 5
2.1.3.6.3 Sign up for a Stack Exchange Account . . . 6
2.1.3.6.4 Efficiently browse you SX sites . . . 6
2.1.3.6.5 Online Git Server Communities . . . 6
2.1.3.6.6 Slack, another component of [Agile Sofware Development] . . . 7
2.1.3.7 You Online Data Science Portfolio . . . 7
2.1.3.7.1 An Example, Emeline Liu . . . 7
2.1.3.8 Links . . . 7

2.1.3.1 Class Readings, Assignments, Textbooks Syllabus Topics

2.1.3.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today: ISLR3,(R4DS-4-6)
  - For next class: DL01 DL02 (R4DS-7,8)
- Laboratory Exercises:
  - LE0 : Due today
  - LE1 : Posted Today
    - LE1 is due Tuesday February 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 ** **
    - DSCI 453 Students
        - Next Report Out #1 is Due ** **
    - All DSCI 353/353M/453, E1453/2453 Students:
        - Peer Grading of Report Out #1 is Due ** **
- 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

2.1.3.1.2 Syllabus

2.1.3.2 Current Status of Everyone in Class

- So as of today, All the elements for the course should be working for you

If not, reach out to the TAs (@kristen hernandez and @will oltjen)

- Defining where you issue is
- And we'll fix it

You should all have the following Elements setup

- You have Forked the `23s-dsci353-453-prof` Bitbucket Repo - And changed the `prof` to your caseID - You can sync new changes in prof repo to your personal repo.
- And you have
    - Logged into http://ondemand.case.edu
    - And launched a "RStudio Server (rxf131)"
    - And made a Git folder in your /mnt/pan/courses/dsci353-453/caseID folder
    - And do the config commands for your name and email of your Git Server
    - And made the /mnt/pan/courses/dsci353-453/caseID folder
- You have joined the DSCI Slack
    - using your caseID email address
    - And joined the DSCI353-353m-453 Slack Channel
    - And for DSCI453 students, have joined the DSCI453 SemProj Channel

Lets check your "primary group" in HPC-Markov

- Launch the SDLE Diagnostics App

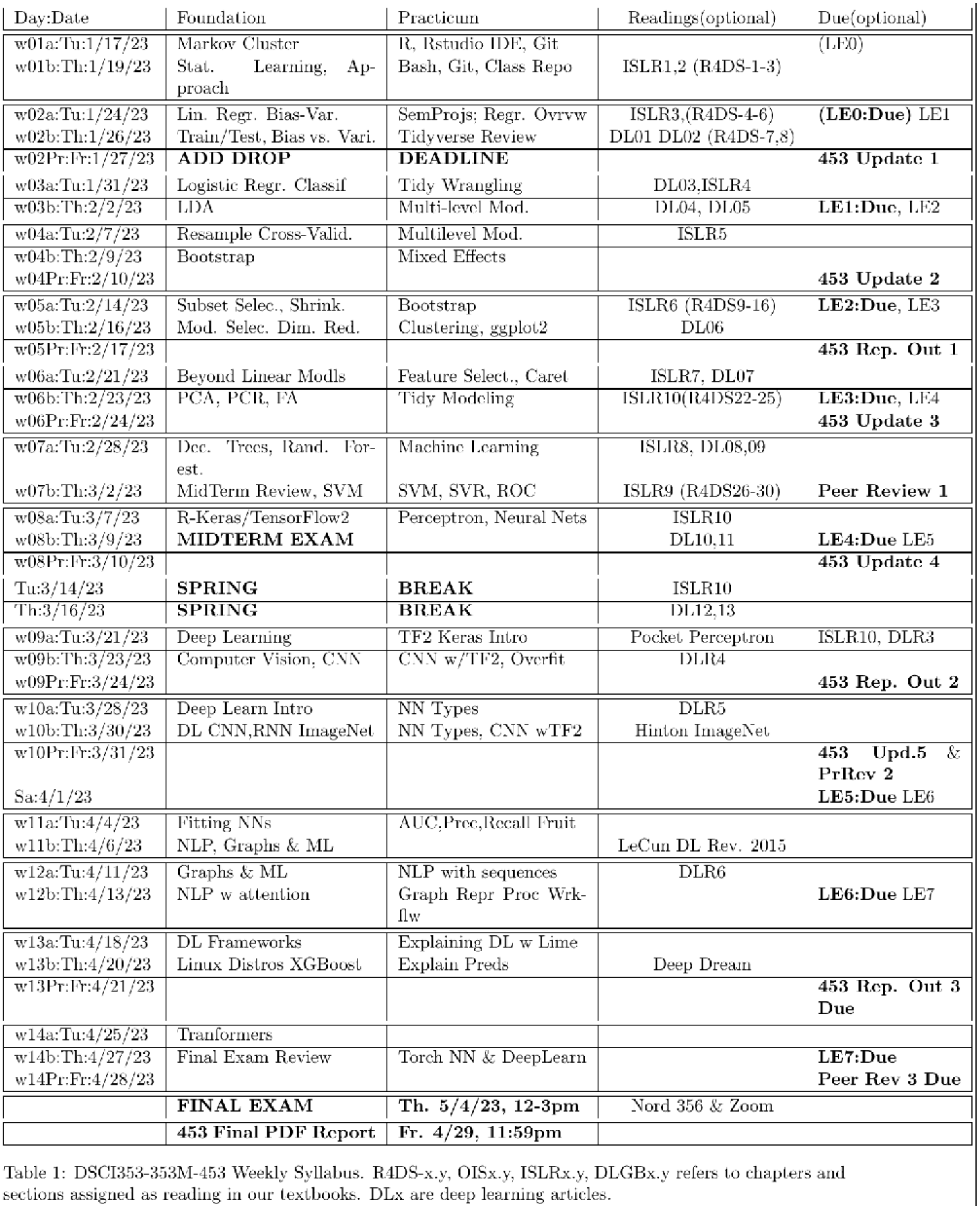
Confirm your /mnt/pan/courses/caseID folder

- Is only accessible to yourself
- If others can enter your /mnt/pan/courses/dsci353-453/caseID folder
    - Then in an LXDE desktop
    - Open the file manager
    - navigate to /mnt/pan/courses/dsci353-453/caseID folder
    - right click on your folder
    - go to permissions tab
    - and change all 3 access choices to be `only owner`

2.1.3.2.1 Prof. Laura Bruckman will present in class Today

- 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**



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. Learning, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Lin. Regr. Bias-Var.	SemProjs; Regr. Ovrvw	ISLR3,(R4DS-4-6)	**(LE0:Due)** LE1
w02b:Th:1/26/23	Train/Test, Bias vs. Vari.	Tidyverse Review	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

2.1.3.3 The Lab Exercises (LEs)

- Each LE is worth 9 points (except LE0 = 0 points)

So 63 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

The Deep learning, TensorFlow, GPU problems

- Are after the midterm break
- And these problems can be quite challenging
- So start on the LEs early
  - And ask questions in the DSCI Slack Channel

LE1 is posted today.

2.1.3.4 Literate Programing: Donald Knuth

- Donald Knuth

  - Bachelors and Masters degrees from CWRU
  - PhD from CalTech
  - CS Professor at Stanford

Did a great many things in Computer Science

- TAOCP: The Art of Computer Programming
  - Started in 1962, and not yet finished
  - Currently 7 volumes
- He also develeded TeX, the precursor to LaTeX

2.1.3.4.1 Literature Programming, was another of his goals

- Literate programming is a programming paradigm introduced by Donald Knuth

  - in which a program is given as an explanation of the program logic
    - in a natural language, such as English,
  - interspersed with snippets of macros and traditional source code,
    - from which a compilable source code can be generated.

The literate programming paradigm, as conceived by Knuth,

- represents a move away from writing programs
  - in the manner and order imposed by the computer,
  - and instead enables programmers to develop programs in the order
  - demanded by the logic and flow of their thoughts.
- Literate programs are written as an uninterrupted exposition of logic
  - in an ordinary human language, much like the text of an essay,
  - in which macros are included to hide abstractions and traditional source code.
- Literate programming (LP) tools are used
  - to obtain two representations from a literate source file:
  - one suitable for further compilation or execution by a computer, the “tangled” code,

  - and another for viewing as formatted documentation, which is said to be "woven" from the literate source.
- While the first generation of literate programming tools
  - were computer language-specific,
  - the later ones are language-agnostic
  - and exist above the programming languages.

Now a days one can integrate R and Python code in a common shared environment,

- as can be done with Rstudio v1.2 and the reticulate package.
- We use this in our data analytics in the SDLE Research Center at CWRU.

2.1.3.5 Agile Software Development

- In early 2000's the way software is developed changed Radically
  - With the Agile Manifesto
  - And the Agile Software Development Principles
  - Overview of Agile Software Development
  - Agile Development Philosophy

Agile is enabled by Literate Programming

- And Relies on an Open Tool Chain

2.1.3.6 Your Open Data Science Tool Chain

2.1.3.6.1 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

Pilot your DSCI studies using available data

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

Use Git repositories

- For version control
- For Collaboration
- For Open Science sharing

2.1.3.6.2 Twitter used for Data Science

- As part of setting up our Data Science Tool Chain
  - Sign up 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

- @daniela_witten Daniela Witten one of the ISLR authors
- @simplystats
- @Rbloggers
- @JennyBryan
- @hspter Hilary Parker
- @NSSDeviations
- @rstudio
- @rstudiotips
- @R_Programming
- @CRANberriesFeed
- @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
-

2.1.3.6.3 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

2.1.3.6.4 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

2.1.3.6.5 Online Git Server Communities

- After your BitBucket Account
- You'll probably want a GitHub account,.
- Many R Projects are there, and
- you can fork their repo's to inspect the code very easily.

2.1.3.6.6 Slack, another component of [Agile Sofware Development]

- cwru-dsci.slack.com
  - an online collaboration tool
- Its an intrinsic part of agile software development
  - There is slack app for phones
  - And client for computers, its on VDI.

2.1.3.7 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
- Need to be conscious of data use terms and agreements
  - Funded research at CWRU falls under IP agreements
  - So when you consider licenses you want to use
  - They must be consistent with the IP terms that came
  - With datasets and codes

2.1.3.7.1 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.

2.1.3.8 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
- emelineliu.com