FALL 2018

TEXAS
The University of Texas at Austin

APPLIED ENGINEERING DATA ANALYSIS, OPTIMIZATION AND VISUALIZATION

Class overview and introduction to data

JOSHUA RHODES, PHD
Research Fellow/Adjunct Professor, The University of Texas at Austin

FALL 2018

TEXAS
The University of Texas at Austin

TOO BIG TO EXCEL

APPLIED ENGINEERING DATA ANALYSIS, OPTIMIZATION AND VISUALIZATION

Class overview and introduction to data

JOSHUA RHODES, PHD
Research Fellow/Adjunct Professor, The University of Texas at Austin

Observance of University policies

- Standard University policies relating to accommodation for students with disabilities and to scholastic dishonesty will be followed in this course. Information regarding these policies may be found in the General Information Bulletin.
- The University of Texas at Austin provides upon request appropriate academic adjustments for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TDD or the College of Engineering Director of Students with Disabilities at 471-4321.

A note about the waitlist

- There are 10+ people on the waitlist
- Thus, I will institute an attendance policy
  - TBD, but not showing up hurts your grade
- If you are not going to show up for lectures, please drop the class

Class computer requirements

- This class will have in-class live demos as well as exercises
- You will need to bring a laptop to every class
- Windows or Mac is acceptable
- https://www.me.utexas.edu/laptopreq/meter/laptopreq

This class will seek to provide you with a framework to work better with data

- Intro to data
- How to deal with ”big data”
- Leveraging UT’s computing and data resources
- Using collaborative working environments
- Analyzing data
- Using HPC resources to analyze data
- Visualize that data
- Work with data with a spatial dimension

Assignments + Tests (TENTATIVE)

- Likely there will be 2-3 minor coding exercises and 1 larger one
- Open book/internet midterm (1)
  - You should still study
- Final project that utilizes most of what we learn

Term projects

- Please set up a meeting with me soon to go over what you might like to do with your own research data
- These meetings will be used to craft your class project
  - Hope is that this class will not be an extra burden, but will help accelerate your research
- Projects will also have a collaborative aspect, so we will be working in teams (on Github) on some

Current schedule of topics looks like this, very tentative:

DOY	Date	Class	Lecture
Wednesday	29-Aug	1	Intro to class
Monday	3-Sep	NO CLASS	Labor Day - No Class
Wednesday	5-Sep	G1	TACC guest lecture
Monday	10-Sep	2	Intro to Terminal/TACC
Wednesday	12-Sep	3	Intro to R
Monday	17-Sep	G2	UT library data guest lecture
Wednesday	19-Sep	4	Cleaning data
Monday	24-Sep	5	Data with R 1
Wednesday	26-Sep	6	Data with R 2
Monday	4-Oct	7	GIS
Wednesday	3-Oct	8	GIS
Monday	8-Oct	9	Using TACC
Wednesday	10-Oct	10	PostgreSQL
Monday	15-Oct	11	PostgreSQL2
Wednesday	17-Oct	12	Machine learning + Regression
Monday	22-Oct	13	Regression
Wednesday	24-Oct	14	Neural Networks
Monday	29-Oct	15	Neural Networks
Wednesday	31-Oct	16	Clustering
Monday	5-Nov	17	Clustering
Wednesday	7-Nov	18	Optimization
Monday	12-Nov	19	Optimization
Wednesday	14-Nov	20	GitHub
Monday	19-Nov	21	GitHub
Wednesday	21-Nov	NO CLASS	Thanksgiving, no class
Monday	26-Nov	22	Midterm
Wednesday	28-Nov	23	Running applications faster, parellel R
Monday	3-Dec	24	Data Vis
Wednesday	5-Dec	25	Data Vis
Monday	10-Dec	26	Data Vis

This is a lot to cover

- Wide and shallow
- #ExperientialLearning
- Get over actuation energy
- Other classes will be deeper dives

This class has no prerequisites, but we will be writing code

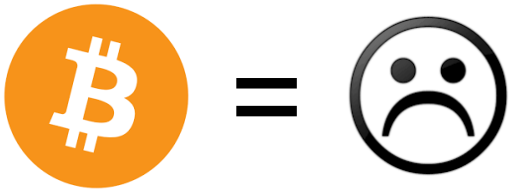
- I am expecting some familiarity, or at least a willingness to self-learn
- I hope to show you how to be a “scrappy programmer”

We will be using the Texas Advanced Computing Center's resources

- You need to set up an account if you haven't already
  - https://portal.tacc.utexas.edu/
- Send me your TACC username so that I can add you to our class allocation
- Do it today

Use TACC resources for class use only

- If you want to use TACC resources for research (and you should!), ask your advisor to set up an allocation
- Do not abuse TACC resources
  - If you mine cryptocurrency for profit, you will fail this class



Test your TACC account, requires 2-factor authentication

```
Connection to maverick.tacc.utexas.edu closed.
[jdr2823@eins-a16551:~$ ssh joshdr@maverick.tacc.utexas.edu
To access the system:

1) If not using ssh-keys, please enter your TACC password at the password prompt
2) At the TACC Token prompt, enter your 6-digit code followed by <return>.


[Password:
TACC Token Code:
```

Set up a GitHub account

- Set up an account with your @utexas.edu address
  - Students can set up “private” repositories (repos) for free
  - Might want to for keeping your ”research-grade” software to yourself at this time
  - IF YOU DON’T DO THIS EVERYTHING DEFAULTS TO OPENSOURCE
  - Do not put any sensitive information in repos
- Run through this GitHub tutorial
  - https://github.com/joshdr83/No-Nonsense-Github-Project
- We will be using GitHub for some projects in this class

We will use GitHub later in the class, but perhaps sooner than I have scheduled

30 make_EIA_data_likert_chart_single.R — View

```
@@ -1,46 +1,66 @@
```

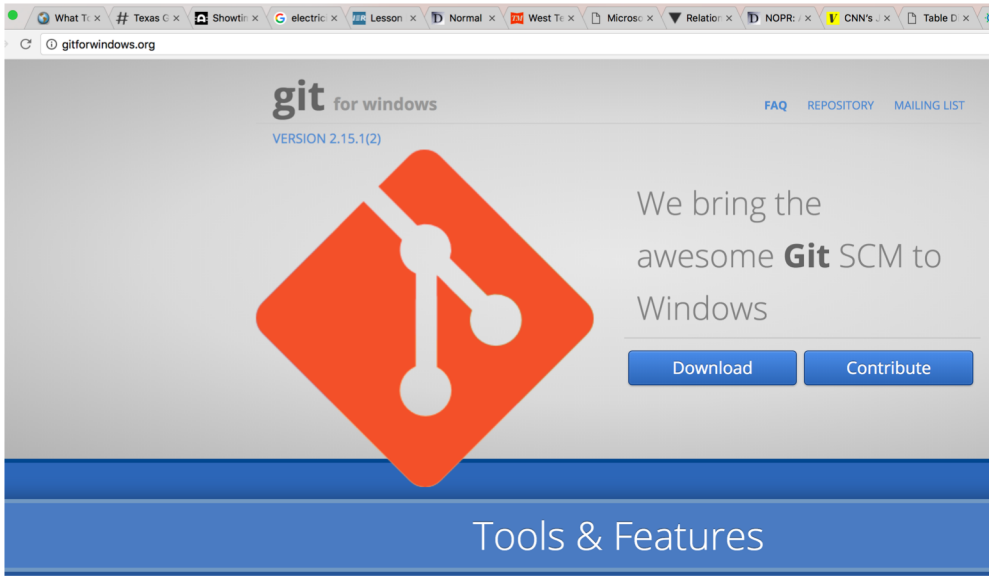
```
 make_EIA_data_likert_chart_single <- function(){
-
+#define the function "make_EIA_data_likert_chart_single"
+ make_EIA_data_likert_chart_single <- function(){
+  #this function takes no arguments. It reads in generator data from the csv file specified below and produces a likert chart showing the total amount of generation from different power plant technologies in the United States in 2016.
+
+  #tell R that it needs to use the "plotrix" and "animation" packages to implement this function
     library(plotrix)
     library(animation)

+  #read in the csv file and store it in a data.frame called "data". Be sure that the "merged_2016_f860_f923_data.csv" file is in your working directory.
     data <- read.csv('merged_2016_f860_f923_data.csv')

+  #the aggregate function will compile "data" into something more manageable, stored in a data.frame called "d2". As written, this command takes the sum of the "Net.Generation..Megawatthours" and "Nameplate.Capacity" columns aggregated by the power plant type in the "Technology" column.
     d2 <- aggregate(data[c('Net.Generation..Megawatthours.', 'Nameplate.Capacity')], by = list(data$Technology), FUN = sum)
+  #rename the column headers for d2 so they are easier to call in our function
     names(d2) <- c('tech', 'MWH_gen', 'MW_cap')

+  #convert MWh to TWh so we have something with more manageable significant figures. Note: 1 TWh (terra-watt-hour) equals 1-Million MWh (mega-watt-hours) and expresses an amount of energy.
     d2$TWH_gen <- d2$MWH_gen/1000000
+  #by default, the rows in d2 are sorted by alphabetical order according to the column we aggregated over (data$Technology), but we want to sort it by energy (TWh). "with(d2, order(-TWH_gen))" produces an array of indices that shows where each row ranks in terms of TWh. For example, "with(d2, order(-TWH_gen))[1]" returns a "10", meaning that the 10th row of d2 has the 1st highest ranking of TWh. "d2[##, ]" returns the data from all columns of d2 for the "##" row. So, "d2[with(d2, order(-TWH_gen)), ]" uses the array of TWh rankings to return d2 data, row by row, in order of the TWh rankings. "d2 <- " assigns that re-ordered data.frame back the object "d2", effectively deleting the old version of d2 and replacing it with one whose rows are sorted by TWh.
```

In this class we will use R and PostgreSQL and maybe some Python, so install:

- R: https://www.rstudio.com/
- pgAdmin 4: https://www.pgadmin.org/
- Python: https://www.python.org/

Windows users will need to install a terminal emulator

- git for windows will also be helpful for using Git/Github
- This will help you learn commands for TACC's systems



A very important aside on file names – VERY IMPORTANT!!!!

- If you send me (turn in a file named) this:
  - Assignment.doc, or hwk3.csv
  - I will NOT even open it, you get 0%
- Always name files so you (and I) know what they are
  - RHODES_HWK3.pdf OR
  - RHODES_midterm_dataset.csv
- This way, I know who it is from and what it is w/o opening it
- Also NO SPACES in filenames, use '_' instead

Why do you want to take this class?

By next class, you need:

- TACC account, tested!
- Github account
- RStudio installed on your computer
- PgAdmin4 installed
- git for windows (if you use PC)

Note: Next class will be a guest lecture at the TACC Visualization Lab on campus, POB 2.404a

- I will be out of town
- TA Philip White will be there
- https://www.tacc.utexas.edu/vislab