CWRU DSCI351-451: Big Data Analytics

Roger H. French, JiQi Liu

25 November, 2018

Contents

14.1.2.1 Reading, Homeworks, Projects, SemProjects

- Homework:
  - All Done
- Readings:
  -
- Projects: We will have four 2 week EDA projects
  - Project 4, Samsung Sensor Machine Learning
  - Due Thursday Dec. 6th
- 451 SemProjects:
  - Turn in short summary of SemProj expected outcomes
  - Report Outs 3 next week 15a 15b

14.1.2.2 Syllabus

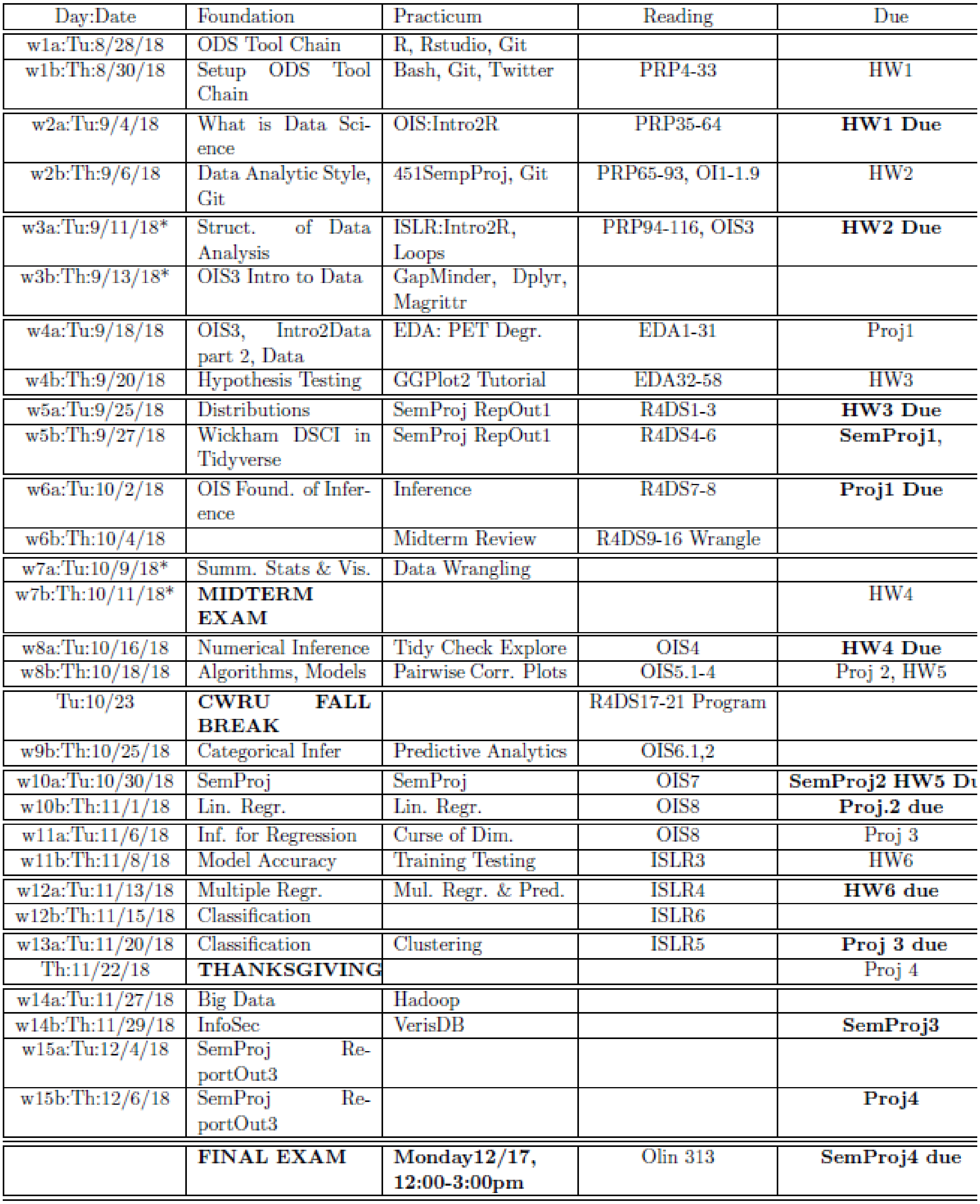
14.1.2.3 Hadoop and Big-Data Analytics

14.1.2.4 3 Seminal Papers from Google

14.1.2.4.1 Google File System

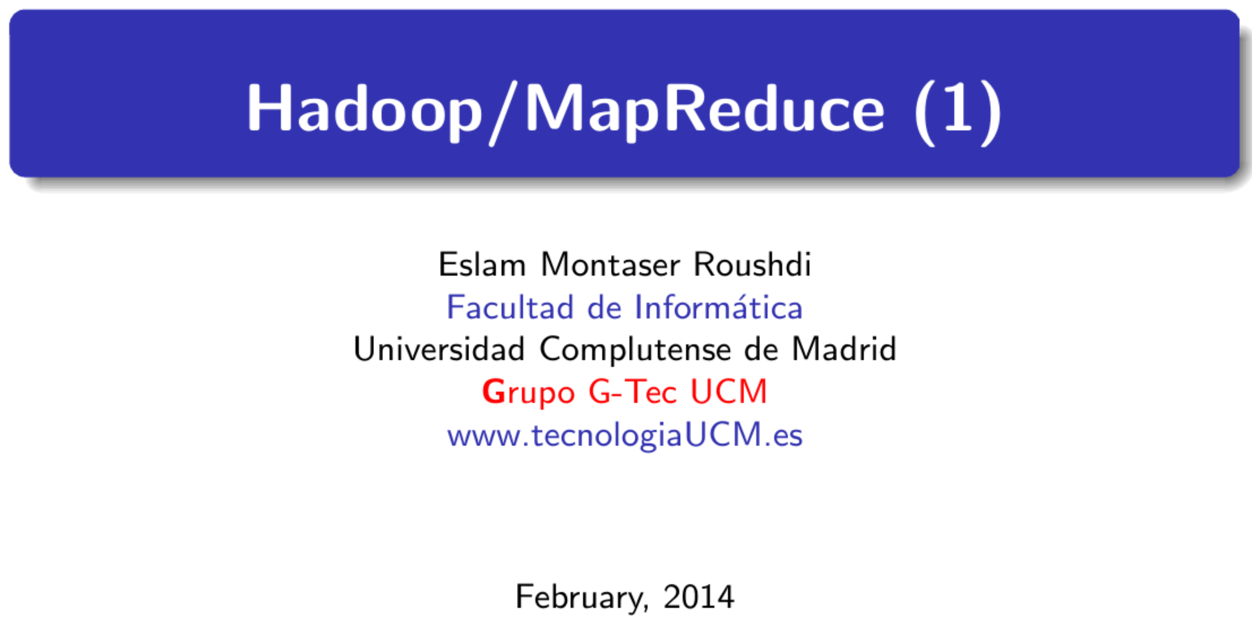
Copies of these papers are in your readings folder of your Repo.

- Ghemawat, S., Gobioff, H., Leung, S.-T., 2003. The Google file system. ACM SIGOPS Operating Systems Review 37, 29–43. doi:10.1145/1165389.945450
- Google File System



Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI351/451 Syllabus



Hadoop/MapReduce (1)

Eslam Montaser Roushdi
Facultad de Informática
Universidad Complutense de Madrid
Grupo G-Tec UCM
www.tecnologiaUCM.es

February, 2014

Figure 2: Hadoop/MapReduce

14.1.2.4.2 MapReduce

- Dean, J., Ghemawat, S., 2004. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM 51, 107–113. doi:10.1145/1327452.1327492
- Google File System

14.1.2.4.3 BigTable

- Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E., 2006. Bigtable: A Distributed Storage System for Structured Data. ACM Transactions on Computer Systems (TOCS) 26, 1–26. doi:10.1145/1365815.1365816
- BigTable

14.1.2.5 Lets get introduced to the concepts

14.1.2.5.1 Hadoop/MapReduce

Hadoop/MapReduce

14.1.2.5.2 Intro Hadoop

Intro Hadoop

14.1.2.5.3 Python in a Big Data World

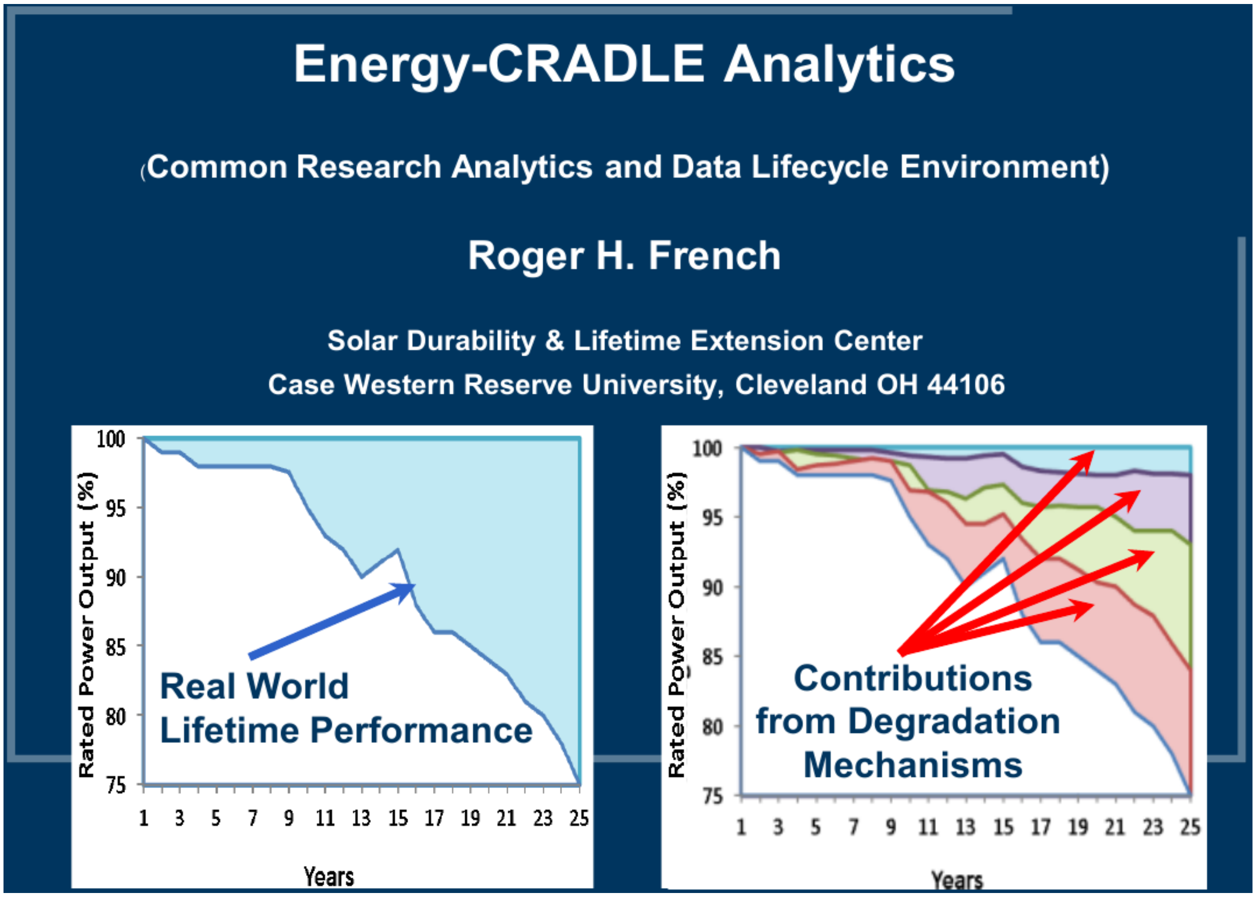
Python in a bid data world

I wonder what BigData, Hadoop and MapReduce is all about ☹

Figure 3: Intro Hadoop

{Python} in Big Data World

Figure 4: Python in a big data world



Energy-CRADLE Analytics

(Common Research Analytics and Data Lifecycle Environment)

Roger H. French

Solar Durability & Lifetime Extension Center
Case Western Reserve University, Cleveland OH 44106

Real World Lifetime Performance

Contributions from Degradation Mechanisms

Figure 5: Energy Cradle

14.1.2.5.4 Hadoop/Hbase: Energy-CRADLE for Energy Analytics

Energy Cradle

NoSQL Data Warehouse and Analytics Environment

14.1.2.5.5 SPARK for stream processing (In RAM)

Apache Spark Tutorials

14.1.2.6 Citations