EPID 600: Data Science for Biomedical Informatics (1 unit)
Fall 2015

Course Description: Data science refers broadly to using statistics and informatics techniques to gain insights from large datasets. Biomedical informatics refers to a range of disciplines that use computational approaches to analyze biomedical data to answer pre-specified questions as well as to discover novel hypotheses. In this course, we will use R and other freely available software to learn fundamental data science applied to a range of biomedical informatics topics, including those making use of health and genomic data. After completing this course, students will:

- Be able to retrieve and clean data, perform exploratory analyses, build models to answer scientific questions, and present visually appealing results to accompany data analyses.
- Be familiar with various biomedical data types and resources related to them.
- Know how to create reproducible and easily shareable results with R and github.

Course Director:
Blanca E Himes, PhD
Assistant Professor of Epidemiology
219 Blockley Hall
423 Guardian Dr
Philadelphia, PA 19104
(215) 573-3282
bhimes@upenn.edu

Guest Lecturers:
Casey Greene, PhD
John Holmes, PhD
Jason Moore, PhD
Michelle Ross, PhD
Lyle Ungar, PhD

TA:
Ryan Urbanowicz, PhD
ryanurb@upenn.edu

Time: Tuesday, Thursday 1-2:30pm
Location: BRB 251

Office Hours: By appointment.

Expectations: You are expected to attend all sessions of the course, read assigned chapters and articles prior to class (if/when assigned), participate during class sessions, and complete required exercises and the class project. **This course requires use of a laptop computer, which you must bring to fully participate in lectures and scheduled lab activities. You must be familiar with this laptop and able to install free programs onto it.**

Grading: The course is graded on a letter grade basis, according to the following proportions:

40% assignments (6 total)
40% biomedical data science project
20% participation in class and lab sessions

Format: This course meets twice weekly. The first half of each session is lecture-based, and the second half is spent working through computational exercises. Six assignments will be due throughout the semester. A final project requiring a substantial amount of work and creativity will be due at the end of the semester in lieu of a final exam. Students will be encouraged to work independently and seek help as needed.

Assignments: Due dates for the assignments: 9/4/15, 9/18/15, 10/2/15, 10/23/15, 11/9/15, 11/23/15.

Biomedical Data Science Project: The final project will answer a question selected by each student using publicly available biomedical data and some of the tools presented during the course. After students choose the topic to address on their own, each will identify three faculty/staff scientists/postdocs from different departments/fields to get feedback and help define a specific novel and interdisciplinary question. Students will work on these projects throughout the semester. Identification of topics is due 9/18/15 as part of course assignment 2. Project proposals with specific questions to be addressed are due 10/16/15 as part of course assignment 4. Final project reports are due 11/30/15. Grading will be based on three project components: (1) proposal that includes a novel interdisciplinary question and feedback provided by 3 diverse experts, (2) an R markdown document that describes the question, source of data, analysis, and results, and (3) a 10 minute oral presentation describing the work to classmates.

Textbooks: The following two textbooks will be provided to each student:

Lander JP, "R For Everyone: Advanced Analytics and Graphics" Addison-Wesley Professional. (2014)

Chang W, "R Graphics Cookbook" O'Reilly Media, Inc. (2013)

Prerequisites: Familiarity with basic statistical (e.g., EPID 526/7 or other first-year graduate level stats course) concepts is expected, as this course will not cover basic concepts in depth. A background in biology and computing would be helpful, but no formal requirements will be enforced.

Academic Honesty: All work submitted for credit is expected to be your own work. In the preparation of all papers and other written work, you should always take great care to distinguish your own ideas and knowledge from information derived from other sources. The term "sources" includes not only published primary and secondary material, but also information and opinions gained directly from other people. The responsibility for learning the proper forms of citation lies with you. You must acknowledge any collaboration and its extent in all submitted work. You are expected to follow Penn's standards of academic integrity as found in:
http://www.upenn.edu/academicintegrity/ai_codeofacademicintegrity.html

Students with Disabilities: University of Pennsylvania, provides reasonable accommodations to students with disabilities who have self-identified and been approved by the office of Student Disabilities Services (SDS). Please make an appointment to meet with me as soon as possible in order to discuss your accommodations and your needs. If you have not yet contacted SDS, and would like to request accommodations or have questions, you can make an appointment by calling (215) 573-9235.

The SDS office is located in the Weingarten Learning Resources Center at Stouffer Commons 3702 Spruce Street, Suite 300. All services are confidential.

Schedule

Session	Date	Lecture	Practicum
Data Science Fundamentals			
1	8/27/15	Intro to Data Science and R	R, Rstudio, github
2	9/1/15	Reproducible research	R markdown, knitr
3	9/3/15	R Programming Basics	Data structures, for loops
4	9/8/15	R Programming Basics II	Functions, scoping
5	9/10/15	Information Retrieval	Reading in data files, databases, web
6	9/15/15	Cleaning and Transforming Data	melt, dcast, dplyr
7	9/17/15	Descriptive statistics (Ross)	
8	9/22/15	Regression (Ross)	linear models, logistic regression
	9/24/15	No class - Papal visit	
9	9/29/15	Visualization (Moore)	ggplot2
10	10/1/15	Machine Learning I (Greene)	
11	10/6/15	Machine Learning II	SVM, RF
12	10/8/15	Model evaluation	ROC curves
Applications in Biomedical Informatics			
13	10/13/15	Overview of Biomedical Data	Public Biomedical Databases
14	10/15/15	Social media data (Ungar)	twitteR
15	10/20/15	Epidemiological data: BRFSS	maps
16	10/22/15	Genetic variation: GWAS	GenABEL
17	10/27/15	Genetic variation: Whole Genome/Exome	Galaxy
18	10/29/15	Health data: secondary data (Holmes)	FARS
19	11/3/15	Health data: claims (Holmes)	Relevant example
20	11/5/15	Gene expression: microarrays	Full workflow in R, heatmap
21	11/10/15	Gene expression: RNA-Seq	DeSEQ2, PCA
22	11/12/15	Functional annotation and enrichment analysis (Urbanowicz)	Examples for SNPs and gene expression from databases/R
23	11/17/15	Transcriptional regulation: methylation arrays and ChIP-Seq	minifi
24	11/19/15	Biological Networks	Bayesian networks, Cytoscape
25	11/24/15	High Performance Computing	
Final Project			
26	12/1/15	Project presentations - I	
27	12/3/15	Project presentations - II	
28	12/8/15	Project presentations - III	