Elements of Data Science: A First Course
Fall 2017

Time: TBD

Instructors: Bryan Gibson (Section 001, Python) | Harry Wang (Section 002, R)

Textbook:

- Section 001: Python Data Science Handbook, Jake VanderPas
  (Free at: https://github.com/jakevdp/PythonDataScienceHandbook)
- Section 002*: Practical Data Science with R*, Nina Zumel and John Mount, Manning (2014), (ISBN-13: 978-1617291562) (PDSR)

Prerequisite(s):

- Introductory programming class as well as basic familiarity with R or Python 3.
- Linear algebra; concepts such as vectors and matrices as well as basic data structures such as arrays, hashes, trees, etc.

Course Description

This course is designed as an introduction to elements that constitutes the skill set of a data scientist. The course will focus on the utility of these elements in common tasks of a data scientist, rather than their theoretical formulation and properties. The course provides a foundation of basic theory and methodology with applied examples to analyze large engineering, business, and social data for data science problems. Hands-on experiments with R or Python will be emphasized. The programming language utilized will depend on the course section, which provides the opportunity to focus on a specific programming language while covering the same skills. Section 001 will focus on Python, while section 002 will focus on R.

Topics include:

- Data Cleaning, Exploration and Visualization
- Classification, Regression and Clustering
- Dimensionality Reduction
- Model Evaluation and Model Selection
- Feature Engineering and Feature Selection
- Statistical Analysis and Hypothesis Testing
- Natural Language Processing and Topic Modeling
- Data processing and delivery using ETL and APIs
- Time Series Analysis
- Recommendation Engines
- Image Recognition

Assignments and Grading

Participation (in class communication, discussion, and pop-up quiz)	**10%**
Homework Assignments (Four, equally weighted at 10% each)	**40%**
Midterm Exam	**25%**
Final Exam	**25%**

TOTAL	100%

Quality of Performance	Letter Grade	Range %	GPA/ Quality Pts.
Excellent - work is of exceptional quality	A+	98 - 100	4.33
	A	93 – 97.9	4.0
	A-	90 - 92.9	3.67
Good - work is above average	B+	87 - 89.9	3.33
Satisfactory	B	83 - 86.9	3.0
Below Average	B-	80 - 82.9	2.67
Poor	C+	77 - 79.9	2.33
	C	73 - 76.9	2.0
	C-	70 - 72.9	1.67
	D	65-69.9	1.0
	D-	60-64.9	0.67
Failure	F	< 60	0.0

Weekly Outline by Section

Section 001

Unit	Topic	Readings	”To Do”
Week #1	Introduction to Data Science problems and tools		
Week #2	Data Processing and Delivery: ETL and API		
Week #3	Data Exploration and Visualization		
Week #4	Data Cleaning and Management		
Week #5	Classification and Regression		
Week #6	Feature Engineering and Application: Natural Language Processing		
Week #7	Midterm Exam		
Week #8	Feature Selection, Model Evaluation and Selection		
Week #9	Project Reporting and Application: Time Series		

Unit	Topic	Readings	”To Do”
Week #10	Dimensionality Reduction and Application: Image Recognition		
Week #11	Statistical Modeling and Hypothesis Testing		
Week #12	Clustering and Topic Modeling		
Week #13	Application: Recommendation Engine		
Week #14	Review and summary, final exam		

Section 002

Unit	Topic	Readings	”To Do”
Week #1 Sep 5	An introduction to data science; Data science problems and tools; Lab R	PDSR: Chapter 1 PDSR: Appendix A	
Week #2 Sep 11	Data processing and delivery using ETL and APIs	PDSR: Chapter 2	HW 1 is due
Week #3 Sep 18	Data exploration and visualization	PDSR: Chapter 3	
Week #4 Sep 25	Data management, data cleaning, and sampling	PDSR: Chapter 4	HW 2 is due
Week #5 Oct 2	Statistical modeling and hypothesis testing	PDSR: Appendix B	
Week #6 Oct 9	Linear and logistic regression models	PDSR: Chapter 7	HW 3 is due
Week #7 Oct 16	Midterm Exam		Midterm
Week #8 Oct 23	Application: data cleaning, visualization, and predictive modeling using regression	Selected research paper and reading	
Week #9 Oct 30	Model evaluation and validation	PDSR: Chapter 5-6	
Week #10 Nov 6	Dimension reduction and application	Selected research paper and reading	HW 4 is due
Week #11 Nov 13	Unsupervised Methods: cluster analysis and association rules	PDSR: Chapter 8	
Week #12 Nov 20	Application: recommendation engine	Selected research paper and reading	HW 5 is due

Unit	Topic	Readings	”To Do”
Week #13 Nov 27	Documentation, model comparison, and deployment	PDSR: Chapter 10-11 PDSR: Appendix C	
Week #14 Dec 4	Review and summary, final exam		Final

Response Policy
You will usually get a response within 24 hours. Expect slower response over the weekend, or when out of town. If you have a question about an assignment, you are advised to email us several days before it is due; if your email arrives within 24 hours of the due date, you may not get a timely response.

Use of Technology in the Course: CourseWorks Online Course System
Class announcements will be made in CourseWorks. You are expected to check CourseWorks course page regularly. A copy of the most recently updated syllabus will be on CourseWorks. Occasionally, there will be other course related handouts posted in CourseWorks. Lecture slides will be posted on Courseworks, however, blackboard lectures will not be posted on Courseworks.

Students with learning disabilities, test or math/statistics anxiety
If you have a learning disability, contact the instructor as early as possible. If you have any extenuating circumstances at any time during the course, contact the instructor as early as possible. Before bringing the disability forms for signature, make sure all information is filled in INK, such as: Your name/Uni/Phone Number, Course Name/Course Number/Section Number/, instructors name, TA Name/e-mail, exam dates/start times.

Academic Dishonesty
Cheating in any form is unacceptable. Standard school policies will be enforced in the case any student is caught cheating. In addition, if you get caught cheating during an exam, you get a score of zero from that exam and are strongly encouraged to withdraw from the course. You are encouraged to check The Columbia University Undergraduate Guide to Academic Integrity at https://www.college.columbia.edu/academics/academicintegrity