NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

PRACTICAL DATA SCIENCE

INFO-GB.3359.10 Fall 2013

SYLLABUS

Professors	**Josh Attenberg**
Office; Hours	Wednesdays 2-3, KMC 8-171 & By appointment
Email	jattenbe@stern.nyu.edu **Emails should have subject tag: subject: [PDS] …** ← **note!**
Course Webpage	http://practical-data-science.com, http://pds-2013.blogspot.com
Classroom	KMC 5-140
Meeting time	Wednesdays, 6pm-9pm
First/Last Class	Sept 25, 2013/December 18, 2013
Final Exam	None
Course Assistants CA Office Hour	Kumar Bharath Prabhu (kumar.prabhu@nyu.edu) Mondays 2-4, KMC 7-100

1. Course Overview

Data is the new oil. Data is a new class of economic asset. Those were the conclusions of the reports issued by the World Economic Forum at Davos in January 2011 and January 2012. Research published in 2011 by MIT economists shows that companies adopting "data-driven decision-making" achieved significant productivity gains over other firms. In industry, the hottest job these days is the Data Scientist. Data scientists combine technical and statistical skills, analytical thinking, and business acumen. One of the complaints about the data scientists trained in computer science departments is that they're "just technical", understanding algorithms well, but lacking important skills in problem formulation, evaluation, and analysis generally. On the other hand, those trained in math and statistics departments, in addition to those trained in business schools tend to have underdeveloped technical skills. This course will cover all of these aspects of being a data scientist.

This class is an introduction to the practice of data science. The student will leave the class with a broad set of practical data analytic skills based on building real analytic applications on real data. These skills include accessing and transferring data, applying various analytical frameworks, applying methods from machine learning and data mining, conducting large-scale rigorous evaluations with business goals in mind, and the understanding, visualization, and presentation of results. The student will have experience processing "big data," the latest buzz concept in a field awash with buzz. Specifically, the student will be able to analyze data that are too big to fit in the computer's memory, and therefore thwart many standard analytical tools. The student will have experience with unstructured data, for example processing text for applications such as "sentiment analysis" of user-generated content on the web.

Students will program throughout the course and are expected to have some programming experience coming in, or a keen desire to learn on the fly. The emphasis of the course will be on rigor and practical usefulness. This is not a replacement for a class on machine learning, database, data mining theory or algorithms.

2. Focus and interaction

The course will explain through lectures, in-class exercises, and real-world cases the fundamental concepts underlying the practice of data science. The emphasis primarily is on being able practically to manage, mine, and visualize data. The course is not a "breadth" course, in that we will not try to be comprehensive across techniques. Nor is it a "depth" course, in that we will not dig deeply into some particular techniques. Instead, the course can be thought of as a practical course: we would like students to leave the class with the practical ability to do some data science, both in their careers and in other related courses.

Student participation is an essential part of the learning process. We expect you to be prepared for class discussions and in-class exercises by having satisfied yourself that you understand what we have done in the prior classes. You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class.

It is important for students to realize that practical data science involves spending time getting things to work. We will be there to help you, but if you tend to get frustrated easily with computers not quite "working right" or with figuring out technical details, then this class is not for you – and practical data science is not for you. We will expect you to be willing to work through the inevitable technical details. You should expect us to be responsive to requests for help, when you begin to get frustrated. Note that this is the first time through this class, and we may not have encountered a particular problem--for example, one that is specific to some specific operating system or system configuration. We expect you to be an active part of the class community, working to help your fellow students to deal with technical problems, and posting solutions you have found on the class discussion board so that others can benefit.

The course website will be used for distributing reading materials and assignments. Communications for the course will be over the Google group. Here, we will notify you of any new content on the web page, new assignments, and any late-breaking news. Additionally, the Google group can be used to ask any questions or to discuss any of the course material. Please feel free to make full use of the office hours of the course assistants and the professors. However, if you have the question, someone else may too and everyone may benefit from the answers being available from the Google group. Also, please try to answer your classmates' questions. In grading your class participation we will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class). Finally, the course blog will be used to give notification of new content published to the site, and will act as a repository for interesting and relevant information. Classes will be used to collect the homework assignments given throughout the class.

We will check our email at least once a day during the week (M-F). We get a tremendous amount of email, and cannot process it all daily. So: ***Your email will get priority if you include the special tag [PDS] in the email subject header as indicated in the information block on the first page of this syllabus***. We will sort/filter based on this tag in order to make sure to process class mail first. We cannot guarantee to be able to process your email promptly if you do not include the special tag, as we may not read it for a while.

In general, we will follow Stern default policies unless we state otherwise. We will assume that you have read them and agree to abide by them:

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

3. Tutorials and Readings

We know of no textbook that covers the breadth of practical, hands-on skills that are required for data science—so we will hand out tutorials and other materials. We have been working to create appropriate-level tutorials for the various hands-on tools we need in the class. Because the subject matter is so broad, we may have missed some things. It is your responsibility to alert us if this is the case. Please think carefully: where is the gap that needs to be filled. We will then work to fill it.

We also will assign readings that provide context and more breadth and depth on the fundamental principles of data science. These will complement our in-class focus on practical, hands-on data science issues. Readings will be taken from:

Learn Python the Hard Way, Shaw

Programming Collective Intelligence: Building Smart Web 2.0 Applications, Segaran

Thinking Stats, Downey

Data Science for Business: Fundamental principles of data mining and data analytic thinking, by Provost & Fawcett.

4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 15%
2. Take Home Assessments: 35%
3. Term Project: 30%
4. Participation & Class Contribution: 20%

At NYU Stern we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. We expect students generally to receive As and Bs in this class, with only those who are not committed to the class earning C's or lower. Of course, the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

Homework Assignments
The homework assignments are listed (by due date) in the class schedule below. You can interact with your classmates to understand how to complete the homeworks. You must complete the homeworks on your own. What does this mean practically? You cannot copy code from another student. However, you can ask another student to help you write the code for something similar to the assignment, and then use that as a template to complete your assignment. *This applies only to the "homeworks" and not to the "take home assessments", described below;* ***the take home assessments you must do completely on your own.***

Completed assignments must be handed on blackboard at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated. Answers to homework that go beyond program output should be well thought out and communicated precisely, avoiding sloppy language, poor diagrams, and irrelevant discussion.

Generally the Teaching Assistants should be the first point of contact for questions about and issues with the homeworks.

"Take Home Assessments"
We will not have traditional exams in this class. The main homeworks will be lightly graded, as they are designed for you to develop your skills. There will be 2 take-home "assessments" at approximately 1/3 and 2/3 through the class (see schedule). These will assess some subset of the skills you have developed to that point in the class, and will be more rigorously graded. The idea is that you will have had the opportunity to test yourself on the skills and talk to the instructors/TAs about them before being rigorously assessed. We will discuss them more in class when the time comes.

Worth repeating: except as explicitly stated otherwise, you are expected to complete the take home assessments on your own—without interacting with others.

Term Project
The term project will be the other main assessment of the development of your skills. A term project report will be prepared by student teams. The report will include an appendix that details the contributions of the team members. Student teams will comprise 3 students. *We will choose your teams and inform you by the second class. Teams will be chosen to have a mix of skill levels.* Teams are encouraged to interact with the instructors and TAs electronically or face-to-face in developing their projects and project reports. You will submit a proposal for your project about halfway through the course. We will discuss the project requirements in class.

Late Assignments
As stated above, assignments are to be submitted on Blackboard at least *one hour prior* to the start of the class on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.

Participation/Contribution/Attendance/Punctuality
Please see Section 2.

Re-grading
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or assessment, please write a formal memo to us describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., a photocopy of class materials). We will make a decision and get back to you as soon as we can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

***FOR STUDENTS WITH DISABILITIES*:** If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

Please read the policies for Stern courses

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

Please keep in mind the Stern Honor Code

http://www.stern.nyu.edu/mba/studact/mjc/hc.html

Class Schedule

Class Number	Date	**Topics** (tentative)	Readings	Deliverables
1	Sept 25	**Introduction** Data science practice, Class objectives, Detailed example, Boring syllabus details, getting started lab		**Homework 0**
2	Oct 2	**Data Processing & Programming** Unix Command Line Tools. Data structures, program command, elements of python. Examples from "Learn Python the Hard Way"	Learn Python the Hard Way, Linux Notes	**HW#1 Due**
3	Oct 9	**Data Storage and Representation** More on programming python, class objects and methods. XML, JSON, YAML. Structured data: information architecture, ER representations Reading/writing/transforming data.		**HW#2 Due**
4	Oct 16	**Predictive models I.** Data representation for predictive modeling, models, data-driven-model applications. Evaluating models, metrics for model quality. *Example: Evaluate several predictive models on "test" data, compute metrics, visualize results, compare methods. How to decide which model is best?* **(Guest Speaker: Foster Provost, Stern)**		**HW#3 Due**

Class Number	Date	Topics	Readings	Deliverables
5	Oct 23	**Predictive models II**. Learning models from data. Training. How does that work for selected models? Overfitting, holdout evaluation, cross-validation, overfitting avoidance. *Example: build predictive models from data; then apply evaluation methods from PM I, plus overfitting analysis.* **(Guest Speaker: Foster Provost, Stern)**		
6	Oct 30	**ML Applications** Practical issues building predictive models for financial applications. Building a solution for a Kaggle competition. **(Guest Speakers: Vasant, Puneet and Eli)**		
7	Nov 6	**Databases and Text** Relational databases, table structure. SQL. Indexes. Text and Regular Expressions *Example: get data from a non-trivial external database, process locally to derive insight.*		**Take home assessment #1 due on Friday**
8	Nov 13	**Predictive Models III.** Scikit.learn- building machine learning models in python. Data transformation, learning from text data. *Example: learning from real data, example deploying model.*		**Get projects in order**
9	Nov 20	**Statistics, Systems and Experimentation.** Experimental design, evaluation, statistical issues. Evaluating live *Example: how do deploy a data-driven system to real users in a controlled experiment, collect and evaluate it's effects* **(Guest Speaker: Nell Thomas, Etsy)**		**HW#5 Due**
10	Dec 4	**Data Visualization** **(Guest Speaker: Kristen Sosulski, Stern)**		**Take home assessment #2 due**
11	Dec 11	**Big Data & APIs** Distributed file systems, Map/reduce, Hadoop (what's that? when is it useful?). Related big data technologies/platforms: Pig, HBase Programmatic access to get (and post) data. *Example: get data from web source, process locally, visualize using visualization API.*		
12	Dec 18	**Wrap up**		**Project due**