**CS109 Data Science Syllabus**

**CS109 Data Science**

Welcome to CS109! The course is also listed as STAT121 and AC209, and offered through the Harvard University Extension School as distance education course CSCI E-109. All lectures and labs will be recorded and the videos will be archived and streamed live during meeting times. The requirements for these four labelings of the course are the same, except that for students registered for AC209, since they will be receiving graduate-level credit, homeworks and the final project will be held to a higher standard and there may be additional readings.

**Instructors:**

* Rafael Irizarry, Biostatistics
* Verena Kaynig-Fittkau, Computer Science

**What is this class about?**

This class is about learning from data, in order to gain useful predictions and insights. Separating signal from noise presents many computational and inferential challenges, which we approach from a perspective at the interface of computer science and statistics. Through real-world examples of wide interest, we introduce methods for five key facets of an investigation:

* *data munging/scraping/sampling/cleaning* in order to get an informative, manageable data set
* *data storage and management* in order to be able to access data - especially big data - quickly and reliably during - subsequent analysis
* *exploratory data analysis* to generate hypotheses and intuition about the data
* *prediction* based on statistical tools such as regression, classification, and clustering
* *communication* of results through visualization, stories, and interpretable summaries

**Why take this class?**

Hal Varian, Chief Economist at Google, said that:

*“The sexy job in the next ten years will be statisticians The ability to take datato be able to understand it, to process it, to extract value from it, to visualize it, to communicate it thats going to be a hugely important skill.”*

More and more applications in industry, academia, and everyday life are or should be based on careful analysis of data. For example, consider the book *Moneyball* (for sports) and the work of Nate Silver (for elections as well as sports). More and more data sets are becoming available, to the point where some companies have described themselves as “drowning in data”. This presents many opportunities, but the right tools from *both* computer science and statistics are needed so that you can learn from the data *without* drowning.

**Expected Learning Outcomes**

After successful completion of this course, you will be able to…

* Use Python and other tools to scrape, clean, and process data
* Use data management techniques to store data locally and in cloud infrastructures
* Use statistical methods and visualization to quickly explore data
* Apply statistics and computational analysis to make predictions based on data
* Apply basic computer science concepts such as modularity, abstraction, and encapsulation to data analysis problems
* Implement data-intensive computations on cluster and cloud infrastructures using MapReduce
* Effectively communicate the outcome of data analysis using descriptive statistics and visualizations

**Who should take this class?**

The prerequisite for this class is programming knowledge at the level of CS 50 (or above), and statistics knowledge at the level of Stat 100 (or above). Both undergraduates and graduate students are welcome to take the course.

**What is the structure of this class?**

The class will consist of both labs and lectures including class puzzles and interactive in class examples.

**Course Logistics**

**Prerequisites**

Programming knowledge at the level of CS 50 or above, and statistics knowledge at the level of Stat 100 or above (Stat 110 recommended). Extension school students are required to have taken CSCI E-26 and STAT E-150 or above. Exceptions with permission of the instructors.

**Required Textbook**

None. Instead, we have a list of recommended readings on the web site.

**Online Discussion Forum**

We’ll be using [Piazza](http://cs109.github.io/2014/pages/www.piazza.com) as our online forum. You can register [here](https://piazza.com/harvard/fall2014/cs109) for the Fall 2014 course. Piazza is your main venue to ask questions, discuss problems, and help each other out. Piazza is a question-and-answer system designed to streamline class discussion outside of the classroom. It should always be your first recourse for seeking answers to your questions about the course, lecture or reading material, or the assignments. Piazza supports LaTeX code formatting, embedding of images, and attaching of files. We will also use Piazza for all announcements, so it is important that you are signed up.

**Online Videos**

All [lectures will be posted online](http://cm.dce.harvard.edu/2015/01/14328/publicationListing.shtml) and should be available 24 hours after meeting time.

**Office Hours**

The staff will hold weekly office hours, either in person or via Skype for distance education students. Office hour times and locations will be listed on Piazza. Office hours provide you with an opportunity to review and discuss course materials as well as provide further guidance for your homework in a more intimate environment, with only your teaching fellow and maybe a handful of classmates present. Online students can make special arrangements directly with their assigned Teaching Fellows to meet on Skype.

**Course Components**

**Lectures**

The class meets twice a week for lectures and joint class activities. The class activities are designed to help you master the relevant materials, to work on your homework in groups, and to get you started on your project. The weekly schedule of lectures is posted on the course web site.

**Labs**

Lectures are supplemented by labs led by the teaching fellows or guest lecturers. The labs are meant to supplement material from lectures with examples and discuss programming environments (e.g., iPython).

**Project**

Towards the middle of the course you will start to work on a data science project. The goal of the project is to go through the complete data science process to answer questions you have about some topic of your own choosing. You will acquire the data, design your visualizations, run statistical analysis, and communicate the results. Part of this project will be assessed midway through the course and then the rest of it will be assessed at the end of the course.

You will work closely with other classmates in a **3-4 person project team**. You can come up with your own teams and use Piazza to find prospective team members. If you cant find a partner we will team you up randomly. We recognize that individual schedules, different time zones, preferences, and other constraints might limit your ability to work in a team. If this the case, ask us for permission to work alone.

**Homework**

The homework is going to provide an opportunity to learn data science skills and to test your understanding of the material. See the homework as an opportunity to learn, and not to “earn points”. The homework will also be graded to reflect this objective.

**Reading Assignments**

The course schedule includes required readings. The goal of the reading assignments is to prepare for class, to familiarize yourself with new terminology and definitions, and to determine which part of the subject needs more attention. The homework assignments may contain questions about the mandatory readings. When answering those please be brief and to the point!

**Course Policies**

**Assessment Procedure**

Your final grade will be determined by the number of points you collect. You can collect various amounts of points for the different parts of the class:

* **Homework**: 65%, assessed on your individual submission.
* **Final Project Part I**: 10%, assessed on your individual submission.
* **Final Project Part II**: 25%, assessed on meeting the project criteria.
* **Best Projects**: We will elect the top three project submissions that will get extra points.

Homework, quizzes and project will be graded on a 10 point scale in increments of 1 using the following scale:

| **Points** | **Description** |
| --- | --- |
| 10 | Exceptional / no mistakes |
| 9 |  |
| 8 | Good / minor mistakes |
| 7 |  |
| 6 | OK / some mistakes |
| 5 |  |
| 4 | Not good / many mistakes |
| 3 |  |
| 2 | Very bad / major errors |
| 1 |  |
| 0 | Did not participate / did not hand in |

Scores >8 is equivalent to an A. Teaching Fellows will evaluate your work for mechanical correctness and holistically to assign an overall score of the assignment. In addition to the scores the Teaching Fellows will give written feedback for each problem.

**Project Group Peer Assessment**

In the professional world, three important features affect your productivity and success: your own effort, the effort of people you depend on, and the way you work together. For this reason we have chosen a team-based approach that values all three of those features. After the team-based project you will provide an assessment of the contributions of the members of your team, including yourself. Your teammates assessment of your contributions and the accuracy of your self-assessment will be considered as part of your overall course evaluation.

**Collaboration Policy**

You are welcome to discuss the course’s ideas, material, and homework with others in order to better understand it, but **the work you turn in must be your own** (or for the project, yours and your teammates). For example, you must write your own code, run your own data analyses, and communicate and explain the results in your own words and with your own visualizations. You may not submit the same or similar work to this course that you have submitted or will submit to another. Nor may you provide or make available solutions to homework to individuals who take or may take this course in the future. During the course of the semester, you will complete a number of questionnaires online. The purpose of these questionnaires is to evaluate how well this course works for you. Your answers will only be used to provide feedback on your learning and make adjustments to the course. They will not affect your grade in any way. Unless stated otherwise, you may neither look up any information, nor consult others during these questionnaires.

**Quoting Sources**

You must acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header). You can also acknowledge sources in a README.txt file if you used whole classes or libraries. Do not remove any original copyright notices and headers. However, you are encouraged to use libraries, unless explicitly stated otherwise!

You may use examples you find on the web as a starting point, provided its license allows you to re-use it. You must quote the source using proper citations (author, year, title, time accessed, URL) both in the source code and in any publicly visible material. You may not use existing complex combinations or large examples. For example, you may not use a ready to use multiple linked view visualization. You may use parts out of such examples.

**Missed Activities and Assignment Deadlines**

Projects and homework **must** be turned in on time, with the exception of late days for homeworks as stated below. It is important that everybody attends and proactively participates in class and online. We understand, however, that certain factors may occasionally interfere with your ability to participate or to hand in work on time. If that factor is an extenuating circumstance, we will ask you to provide documentation directly issued by the University, and we will try to work out an agreeable solution with you (and your teammates).

**Homework Deadlines and Late Days**

In the weeks when homework is due, it will be due on Thursdays at 11:59 pm, unless otherwise announced. Each student is given **six late days** for homework at the beginning of the semester. A late day extends the individual homework deadline by 24 hours without penalty. **No more than two late days may be used on any one assignment. Assignments handed in more than 48 hours after the original deadline will not be graded.** If you have already used all of your late days for the semester, we will deduct 2 points for assignments <24 hours late, and 4 points for assignments 24-48 hours late. We do not accept any homework under any circumstances more than 48 hours after the original deadline. Late days are intended to give you flexibility: you can use them for any reason no questions asked. You don’t get any bonus points for not using your late days. Also, you can only use late days for the individual homework deadlines all other deadlines (e.g., project milestones) are hard.

**Regrading Policy**

It is very important to us that all assignments are properly graded. If you believe there is an error in your assignment grading, please submit an explanation via email to us (the staff mailing list) **within 7 days of receiving the grade**. No regrade requests will be accepted orally, and no regrade requests will be accepted more than 7 days after you receive the grade for the assignment.

**Guest Lecture Attendance**

We are lucky to have some of the worlds best researchers take time out of their busy schedules to give guest lectures. We expect all non-distance students to attend these lectures in person and to engage the speakers with questions and comments. You must send an email to the staff at least one day before a guest lecture to be excused.

**Additional Information**

**Accessibility**

If you have a documented disability (physical or cognitive) that may impair your ability to complete assignments or otherwise participate in the course and satisfy course criteria, please meet with us at your earliest convenience to identify, discuss, and document any feasible instructional modifications or accommodations. You should also contact the Accessible Education Office to request an official letter outlining authorized accommodations.

**Credits**

Some of the material in this course is based on other classes. We have also heavily drawn on materials and examples found online and tried our best to give credit by linking to the original source. Please contact us if you find materials where the credit is missing or that you would rather have removed.

**User Notice for Copyrighted Materials on Course Websites**

This course website, and much of the text, images, graphics, audio and video clips, and other content of the site (collectively, the “Content”), are protected by copyright law. In some cases, the copyright is owned by third parties, and Harvard is making the third-party Content available to you under the fair use doctrine. Fair use permits only certain limited uses of the Content. You may use the website and its Content only for your personal, noncommercial educational and scholarly use. Some Content may be provided via streaming or other means that restrict copying; you may not circumvent those restrictions. If you wish to distribute or make any of the Content available to others, or to use any Content commercially, or to use any Content for any purpose other than your personal, noncommercial educational and scholarly use, you must obtain any required permission from the copyright holder. User notice courtesy of the Harvard University Office of General Counsel.