**Computer Science 109a/Applied Computation 209a/Statistics 121a Data Science 1: Introduction to Data Science**

**Pavlos Protopapas, Kevin Rader, Weiwei Pan**

**Fall Term**   
**Class Time**:

**Mon** and **Wed** 1‐2:30 pm in Harvard's Northwest Building (NW), Room B-103

**Labs**: **Thur** 4:00-5:30 pm and **Fri** 10:00-11:30 am in NW Basement Lobby (content is identical, students may only attend one)

**TF Office Hours**  
We have open Office Hours where TFs are present to help you. You do not need to sign up. Just show up.

**Mondays** and **Thursdays** 7:00pm-8:30pm room in the Red couch area in NW basement  
**Tuesdays** 4:00-5:30pm in the Red couch area in NW basement.  
**Tuesdays** 6:00-8:00pm in Maxwell Dworkin lobby outside MD 119 - these are joint Office Hours for all CS courses, you are welcome to attend those too.

**Instructor Office Hours**

**Pavlos**: Thursday 1:00-2:00 pm, Northwest Building B-155  
**Kevin**:   Tuesday 1:00-2:00 pm, Science Center 614  
**Weiwei**: Thursday 5:30-7:00pm, Northwest Building B-164

See Content in [Course Calendar](https://canvas.harvard.edu/courses/12656/pages/course-calendar" \o "Course Calendar" \t ")

**Description**: Data Science 1 is the first half of a one‐year introduction to data science. The course will focus on the analysis of messy, real life data to perform predictions using statistical and machine learning methods. Material covered will integrate the five key facets of an investigation using data: (1) data collection ‐ data wrangling, cleaning, and sampling to get a suitable data set; (2) data management ‐ accessing data quickly and reliably; (3) exploratory data analysis – generating hypotheses and building intuition; (4) prediction or statistical learning; and (5) communication – summarizing results through visualization, stories, and interpretable summaries.

**Course Notes**: Only one of CS 109a, AC 209a, or Stat 121a can be taken for credit. Students who have previously taken CS 109, AC 209, or Stat 121 cannot take CS 109a, AC 209a, or Stat 121a for credit.

**Recommended Prep**: Programming knowledge at the level of CS 50 or above, and statistics knowledge at the level of Stat 100 or above (Stat 110 recommended).

Lectures will be videotaped and posted online about 24 hours later.

**Textbook**: An Introduction to Statistical Learning by James, Witten,Hastie, Tibshirani  
***HOLLIS***: [http://link.springer.com.ezp-prod1.hul.harvard.edu/book/10.1007%2F978-1-4614-7138-7](http://link.springer.com.ezp-prod1.hul.harvard.edu/book/10.1007%2F978-1-4614-7138-7" \o "" \t "blank)  
***Amazon****:* [https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370 (Links to an external site.)](https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370" \o "" \t "_blank)  
***Free electronic version***: [http://www-bcf.usc.edu/~gareth/ISL/ (Links to an external site.)](http://www-bcf.usc.edu/~gareth/ISL/" \t "_blank)[(Links to an external site.)](http://www-bcf.usc.edu/~gareth/ISL/" \o "" \t "_blank)

**Labs:**Optional (but **strongly** suggested) labs will be held Thursday afternoon and Friday morning (they are identical) throughout the course.  Labs will be very hands-on and will go over practice problems similar to the homework problems and review difficult material.

**Help for the Course:**There are two dedicated email help-lines for the course. Send all questions regarding grading to: [cs109A+grading@gmail.com](mailto:cs109A+grading@gmail.com). All questions regarding course materials and requests for extra assistance should be sent to: [cs109A+help@gmail.com](mailto:cs109A+help@gmail.com). Submitting your questions to the correct help-line will facilitate speedy responses.

**Software:**We will be using Python 2.x.

Please download and install the [Anaconda platform (Links to an external site.)](https://docs.continuum.io/anaconda/" \o "" \t "_blank), which includes IPython as well all packages that will be required for the course.

**Accommodations for students with disabilities**:   
Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the [Accessible Education Office](http://www.aeo.fas.harvard.edu/) (AEO) and speak with Kevin by the end of the third week of the term: Friday, September 18.  Failure to do so may result in us being unable to respond in a timely manner. All discussions will remain confidential.

**Collaboration:**You are encouraged to discuss homework with other students (and with the instructor and TAs, of course), but you must write your final answers yourself, in your own words.  Solutions prepared “in committee” or by copying or paraphrasing someone else’s work are not acceptable; your handed-in assignment must represent your own thoughts.  All computer output you submit must come from work that you have done yourself. **Please indicate on your problem sets the names of the students with whom you worked.**All exams (midterm and final) are individual work.

**Grading Guidelines:**Your final score for the course will be computed using the following weights:

Homework 40%

Quizzes/Readings  10%

Midterm I  15%

Midterm II  15%

Project  20% .

**Total  100%**

**Homework:**There will be 8 homework assignments (they are essentially weekly). No HW scores will be “dropped.”  You are allowed**up to 4 days of late HW submissions, maximum of 2 days on any single assignment**, no questions asked (solutions will be posted then).  Any other late homework submissions will not be accepted without a written note from UHS or your resident dean’s office.

**Project:**TBD

**Midterms:**There will be two midterms: Thur, October 13, and Mon, November 21. The midterms are open book.

**Quizzes:**Quizzes will be based on short reading assignments given before class. They will be given out at the beginning of the class (you will be given an access code in order to do them)

**Course Calendar:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **MONDAY** | **WEDNESDAY** | **LAB** | **NOTES** |
| **1** | **Week1** |  | 8/31/2016 | 9/1/2016 |  |
|  | Class |  | What is Data Science? Process. Why is it important? Who are we? Course overview. - PP/KR/W |  |  |
|  | Lab |  |  | Lab 0: TBD - Intro to Python (basic syntax, numpy, graphing libraries, basic program structure) (WP) |  |
|  | HW |  | HW0 release |  | Refreshing Statistical and Coding Knowledge and Introduce Python <- Not for Grade |
|  | Project |  |  |  |  |
| **2** | **Week 2** | 9/5/2016 | 9/7/2016 |  |  |
|  | Class | **LABOR DAY - no class** | Statistical Learning. Data Collection & Web Scraping Exploratory Data Analysis & Visualization (WP) |  |  |
|  | Lab |  |  | Lab 1: Intro to pandas, basic stats and visualization, scraping (WP) |  |
|  | HW |  | HW1 release |  | Scraping in Python and Plotting, Histograms, Statistical Intuition, Probability |
|  | Project |  |  |  |  |
| **3** | **Week 3** | 9/12/2016 | 9/14/2016 |  |  |
|  | Class | K-Nearest Neighbors. Data Cleanup (KR) | Probability Distributions & Inference Review Intro to Linear Regression I (PP) |  |  |
|  | Lab |  |  | Lab 2: comparing KNN with Linear regression (WP) |  |
|  | HW |  | HW1 Due - HW 2 Release |  | KNN, Pandas |
|  | Project |  |  |  |  |
| **4** | **Week 4** | 9/19/2016 | 9/21/2016 |  |  |
|  | Class | Introduction to Linear Regression II: Simple and Multiple (PP) | Cross-Validaion & Model Selection (PP) |  |  |
|  | Lab |  |  | Lab 3: Linear Regression/Model selection (Hari) |  |
|  | HW |  | HW2 Due - HW3 Release |  | Regression |
|  | Project |  |  |  |  |
| **5** | **Week 5** | 9/26/2016 | 09/28/2016 |  |  |
|  | Class | Lecture 6: Linear Model Regularization: Ridge & Lasso (PP) | Lecture 7: Linear Model Regularization: Ridge & Lasso (PP) |  |  |
|  | Lab |  |  | Lab 4: Variable selection, regularization, PCA (PP) |  |
|  | HW |  | HW3 Due - HW4 Release |  | Model Selection, LASSO, Ridge |
|  | Project |  | milestone 1 |  |  |
| **6** | **Week 6** | 10/3/2016 | 10/5/2016 |  |  |
|  | Class | Lecture 8: Principle Components & High Dimensionality (KR) | Lecture 9: Bayesian Thinking (WP) |  |  |
|  | Lab |  |  | Lab 5: Milestone lab (install, git, python environments and libraries) (WP) |  |
|  | HW |  | HW 4 Due - no new homework - midterm I |  |  |
|  | Project |  |  |  |  |
| **MID I** | **Midterm I - Week 7** | 10/10/2016 | 10/12/2016 | 10/13 | CURRENT WEEK |
|  | Class | **No Class – Columbus Day** | More on Exploratory Data Analysis & Visualizations (WP) Ethics? | **MIDTERM I** |  |
|  | Lab |  |  | No Lab/No OH |  |
|  | HW |  | no HW due | Midterm I Take Home Exam released 9:00am |  |
|  | Project |  |  |  |  |
| **8** | **Week 8** | 10/17/2016 | 10/19/2016 |  |  |
|  | Class | Intro to Logistic Regression (KR) | Logistic Regression Details (KR) |  |  |
|  | Lab |  |  | Lab 6: Logistic Regression (WP) |  |
|  | HW |  | HW5 release |  |  |
|  | Project |  | Milestone #2 Due |  |  |
| **9** | **Week 9** | 10/24/2016 | 10/26/2016 |  |  |
|  | Class | Bayesian Thinking/KNN revisited (KR) | LDA (KR) |  |  |
|  | Lab |  |  | Lab 7: LDA (WP) |  |
|  | HW |  | HW5 Due - HW6 release |  |  |
|  | Project |  |  |  |  |
| **10** | **Week 10** | 10/31/2016 | 11/2/2016 |  |  |
|  | Class | Regression Trees (PP) | Random Forests (PP) |  |  |
|  | Lab |  |  | Lab 8: Random Forests (WP) |  |
|  | HW |  | HW6 Due - HW7 release |  |  |
|  | Project |  | Milestone #3 Due |  |  |
| **11** | **Week 11** | 11/7/2016 | 11/9/2016 |  |  |
|  | Class | Support Vector Machines I (PP) | Support Vector Machines II (PP/Hari) |  |  |
|  | Lab |  |  | Lab 9: SVM (Hari) |  |
|  | HW |  | HW7 Due - HW8 release |  |  |
|  | Project |  |  |  |  |
| **12** | **Week 12** | 11/14/2016 | 11/16/2016 |  |  |
|  | Class | Experimental Design and testing (KR) | midterm review |  |  |
|  | Lab |  |  | Lab 9b: Project Milestone (all) |  |
|  | HW |  | HW 8 Due - no new homework - midterm II |  |  |
|  | Project |  | milestone 4 and 5 |  |  |
| **13** | **Midterm II - Week 13** | 11/21/2016 | 11/23/2016 |  |  |
|  | Class | **Midterm II Exam** | **No Class – Thanksgiving** |  |  |
|  | Lab |  |  | no lab |  |
|  | HW |  |  |  |  |
|  | Project |  |  |  |  |
| **14** | **Week 14** | 11/28/2016 | 11/30/2016 |  |  |
|  | Class | **More visualization, story telling, project related stuff (PP?)** | Wrap - Up |  |  |
|  | Lab |  |  |  |  |
|  |  |  |  |  |  |
|  | **Reading Period** | 12/5/2016 | 12/7/2016 | 12/8/2016 |  |
|  |  |  |  | Project Poster Presentations (1/3 of grade) |  |
|  |  |  |  |  |  |
|  | **Finals Week** | 12/12/2016 | 12/14/2016 |  |  |
|  |  |  | **Final Project Due (2/3 of grade)** |  |  |