# COS424: Interacting with Data (Spring 2014)

DAVID M. BLEI

April 7, 2014

## COURSE DESCRIPTION

Problems about data abound. Here are some examples:

- Netflix collects ratings about movies from millions of its users. From these ratings, how can they predict which movies a user will like?
- JSTOR scans and runs OCR software on millions of articles. Scholars want to browse and explore their collection. How should JSTOR organize it?
- A biologist has collected hundreds of thousands of measurements about the genotypes and traits of a large population. She would like to run a new experiment. Can she make a hypothesis about which genes are associated with which traits?
- Google sends and receives hundreds of millions of email messages each day. Are some of them spam? Which advertisements should they show next to each one?

Data analysis is central to many modern problems in science, industry and culture. Scientists and engineers have to be fluent in thinking about how to solve modern data analysis problems. This class puts you on the path towards that fluency.

In this course, we will learn about a suite of tools in modern data analysis: when to use them, the kinds of assumptions they make about data, their capabilities, and their limitations. More importantly, we will learn about the language for and process of solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it to data, and understand the meaning of the result.

## Prerequisites

The prerequisite knowledge is calculus, linear algebra, computer programming, and some exposure to probability and/or statistics.

## Lecture

Tuesdays and Thursdays, 1:30PM-2:50PM
Location: Friend 101

## Recitation Office Hour

There will be a weekly recitation office hour, which is optional to attend. The time and location is TBD.

## Course Staff

- Dr. David Blei (Professor)
- Dr. Xiaoyan Li (Lecturer)
- Allison Chaney (TA)
- Rajesh Ranganath (TA)
- Sachin Ravi (TA)
- Pingmei Xu (TA)

## PROGRAMMING

We will use R, which is a powerful open-source platform for statistical computing and visualization. We will hold a special session about learning R in the beginning of the semester.

You can download R for many platforms at www.r-project.org. We also recommend RStudio (www.rstudio.com), which is an excellent environment within which to develop and use R.

To get started consider *Introductory Statistics with R* by Peter Daalgard. It is available as a PDF from the Princeton Library. There are many tutorials on the web.

Optionally, if you prefer, you can use Python in combination with scientific computing packages like SciPy and NumPy.

## WRITTEN WORK

There are three kinds of written work for the class: Reading responses, homework, and the final project.

### Reading Responses

- Every Thursday, you will hand in a reading response about the week's reading. It should simply be a few paragraphs about what you thought of the reading, or what thoughts the reading brought up. The maximum length is one page; there is no minimum length. There's no need to polish these responses.

- The reading responses are not graded, but are noted each week. Further we will read 3 of each student's responses, chosen at random throughout the semester.

- We will not accept late responses. If you miss class on Tuesday, you will not be able to hand in the response.

### Assignments

- There will be five assignments, due every two weeks for ten weeks. They are due at the beginning of the Thursday lecture.

- We require that you do the assignments in pairs. Each pair hands in one assignment.

- Each assignment contains written questions and an open-ended question. Each part is worth 50% of the assignment grade.

- The open-ended question will ask you to find and analyze some data using one of the ideas that we studied in class. You will be asked to report on what you did and what you learned about the data from the analysis. You can use open-source software or write your own software. You can use public data or analyze data that you have private access to.

- The write-up is one page long. You can use an additional page for plots. It is graded on correctness, clarity, creativity, and difficulty. Once you have a final project in mind, we encourage you to start getting involved in the data via the open-ended question.

### Final Project

- The centerpiece of the course is a final project. We will give details later.
- We require that you do final projects in groups of four. Each group hands in one project.
- Each group writes a project proposal.
- Each group writes a five-page report (with an additional three pages for figures and plots) and participates in a class-wide poster session.

### Formatting and Page Limits

- All work should be typed. It should be single spaced with 12-point fonts and one-inch margins.
- We prefer you to prepare your work with LaTeX. But we accept any work that adheres to the formatting.
- We will not read beyond the specified page limit. We will penalized work that goes over the limit. We will *not* penalize or reward work that goes under the limit.

### Various Policies

- We accept on-time work during the beginning of class. After class, you must use late days (see below). We accept late work during an office hour.
- Each student is allowed five late days to be used for the assignments. Late days cannot be used for the reading responses or the final project.
- There are no partial late days.
- Beyond your late days, late work is not permitted.
- Any cheating will result in an F for the course.

## COURSE GRADES

Your final grade is based on the following:

- Final project (50%)

- Assignments (40%)
- Reading responses (5%)
- Participation in class and on Piazza (5%)

If you fail any written component of the course (e.g., do not hand in any reading responses), you may receive a D or F.

## Disputing a grade

If you think we made a mistake on your grade for an assignment, see a TA during his or her office hour. You must see a TA within one week of the return of the graded assignments.

If you think we made a mistake on your final course grade, first check with Dr. Li that there was not a numerical error.

If you still think that we made a mistake, send Prof. Blei the following email: "I think there was a mistake with my grade. Please re-evaluate my body of work for COS424." Prof. Blei will re-evaluate all of your work for the semester. Your grade will go up, stay the same, or go down.

## PIAZZA

We will use Piazza to host all communication. Sign up for the Piazza site at piazza.com/class#spring2014/cos424.

Here are some uses for Piazza:

- Ask questions about the course.
- Answer questions about the course.
- Point fellow students to related material on the internet.
- Give advice about course work.
- Share resources, open-source code, public data sets, etc.
- Receive important announcements from the instructors.
- Privately communicate with the instructors.

Please use Piazza to privately communicate with the instructors. Only use email in exceptional situations.

## SYLLABUS AND READINGS

Most readings come from

- Barber, D. *Bayesian Reasoning and Machine Learning.* Cambrdige University Press, 2012 (BRML).
- Bishop, C. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006. (PRML)
- Murphy, K. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2013. (MLAPA)
- Hastie, T., Tibshirani, R. and Freedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Edition, Springer, 2009. (ESL)

Readings will be posted on the course website. (BRML and ESL are also available for free online from the authors.)

| # | Date | | | Subject | Reading |
|---|---|---|---|---|---|
| 01 | T | 04 | Feb | Introduction | PRML Ch 1.2.1–1.2.4, 2.1–2.3 |
| 02 | R | 06 | Feb | Probability theory review | |
| 03 | T | 11 | Feb | Statistical models | BRML Ch 10 |
| 04 | R | 13 | Feb | Classification with probabilistic models | |
| 05 | T | 18 | Feb | Graphical models | PRML Ch 8.1–8.2; 9.1–9.2 |
| 06 | R | 20 | Feb | Graphical models | |
| 07 | T | 25 | Feb | EyeWire, a game to map the brain [Prof. Sebastian Seung] | |
| 08 | R | 27 | Feb | Computer Vision [Prof. Jinaxiong Xiao] | |
| 09 | T | 04 | Mar | Mixture models and expectation maximization | MLAPP Ch 11.1–11.4 |
| 10 | R | 06 | Mar | Expectation maximization | |
| 11 | T | 11 | Mar | Sequence models and hidden Markov models | PRML Ch 13.1–13.2 |
| 12 | R | 13 | Mar | Hidden Markov models II | |
| — | T | 18 | Mar | *Spring break* | |
| — | R | 20 | Mar | *Spring break* | |
| 13 | T | 25 | Mar | Hidden Markov models III | ESL 3.1, 3.2 |
| 14 | R | 27 | Mar | Linear regression | |
| 15 | T | 01 | Apr | Recommendation Systems [Dr. Laurent Charlin] | ESL 3.3, 3.4, 3.6 |
| 16 | R | 03 | Apr | Regularized linear regression | |
| 17 | T | 08 | Apr | Logistic regression | MLAPP 1.4.6, 8.1–8.3.3 |
| 18 | R | 10 | Apr | Exponential families and generalized linear models | |
| 19 | T | 15 | Apr | Scalable machine learning I | |
| 20 | R | 17 | Apr | Scalable machine learning II | |
| 21 | T | 22 | Apr | Gaussian dimension reduction | PRML 12.1, 12.2, PRML 2.4, Blei (2012) |
| 22 | R | 24 | Apr | Non-negative matrix factorization | |
| 23 | T | 29 | May | Probabilistic topic models | Blei (2012) |
| 24 | R | 01 | May | Summary and discussion | |