# Statistics W4240: Data Mining
# Columbia University
# Spring, 2014

Version: January 30, 2014. The syllabus is subject to change, so look for the version with the most recent date.

## Course Description

Massive data collection and storage capacities have led to new statistical questions:

- Amazon collects purchase histories and item ratings from millions of its users. How can it use these to predict which items users are likely to purchase and like?
- Yahoo news acts as a clearinghouse for news stories and collects user click-through data on those stories. How should it organize the stories based on the click-through data and the text of each story?
- Advances in molecular biology have allowed scientists to gather massive amounts of genomic data. How can this be used to predict gene interactions?
- Large medical labs can receive thousands of tissue and cell samples per day. How can they automatically screen cancerous specimens from non-cancerous ones, preferably with a higher accuracy than doctors?
- Facebook gets millions of photographs annotated by its users. How can it use this data to automatically detect who is in newly uploaded photos?

Many new problems in science, industry, arts and entertainment require traditional and non-traditional forms of data analysis. In this course, you will learn how to use a set of methods for modern data mining: how to use each method, the assumptions, computational costs, how to implement it, and when *not* to use it. Most importantly, you will learn how to think about and model data analysis problems.

## Administrative

**Lecture**

(Note that seating will be very tight, so it is critical that you go to your assigned section)

- Section 1 (Hannah):
  Tuesday and Thursday, 6:10PM–7:25PM
  Location: 614 Schermerhorn
- Section 2 (Cunningham):
  Monday and Wednesday, 6:10PM–7:25PM
  Location: 329 Pupin Laboratories

**Instructors**

- Lauren Hannah
  Office: Department of Statistics, 1009 School of Social Work, 1255 Amsterdam
  Email: lah2178@columbia.edu

- John Cunningham
  Office: Department of Statistics, 1026 School of Social Work, 1255 Amsterdam
  Email: jpc2181@columbia.edu

**Teaching Assistants**

- Phyllis Wan
  Office Hours: Mondays 2:40pm - 4:40pm
  Office: Department of Statistics, 10th Floor School of Social Work, 1255 Amsterdam
  Email: pw2348@columbia.edu

- Feihan Lu
  Office Hours: Tuesdays 3:30pm - 5:30pm
  Office: Department of Statistics, 10th Floor School of Social Work, 1255 Amsterdam
  Email: fl2238@columbia.edu

**Virtual Office Hours**

Owing to student schedules and the subsequent challenges of finding mutually suitable office hours, we will use a virtual platform. Piazza is a highly regarded forum for students to discuss class questions, homework problems, and more. Discussing problems is encouraged, but full solutions should not be posted (see section on academic integrity). The tool can be found at: `https://piazza.com/class/hq76cfqlygz6no?cid=6`. Many Columbia classes find this to be a much quicker and more effective tool than traditional office hours, and we encourage students to use it both to ask questions and to improve their own understanding by posting answers and comments.

# Prerequisites

A previous course in statistics, elementary probability, multivariate calculus, linear algebra and ability to do moderate coding in R. A quiz will be given during the first class to allow students to self-assess their preparedness for this course.

# Grading and Academic Integrity

We take the honor code very seriously; students caught cheating or otherwise in violation will face disciplinary action. Please note the Barnard honor code text:

> "We... resolve to uphold the honor of the College by refraining from every form of dishonesty in our academic life. We consider it dishonest to ask for, give, or receive help in examinations or quizzes, to use any papers or books not authorized by the instructor in examinations, or to present oral work or written work which is not entirely our own,

unless otherwise approved by the instructor.... We pledge to do all that is in our power to create a spirit of honesty and honor for its own sake."
`http://barnard.edu/node/2875`
`https://www.college.columbia.edu/academics/academicintegrity`

Your grade will be determined by three different components:

- **Homework (30%).** Homework will contain both written and `R` data analysis elements. This is due online by the beginning of class on the due date.
- **Midterm Exam (30%).** This will be given in class during midterm week. You will be permitted use one handwritten page, front and back, of notes.
- **Final Exam (40%).** This will be given in class during the finals period. You will be permitted use one handwritten page, front and back, of notes.

Failure to complete any of these components may result in a D or F. Grades may be adjusted upward or downward to reflect trajectory.

**Late Work and Regrading Policy:** No late work or requests for regrades are accepted.

**Homework:** Students are encouraged to work together, but homework write-ups must be done individually and must be entirely the author's own work. Homework is due at the **beginning** of the class for which it is due. **Late homework will not be accepted under any circumstances.** To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), class (STAT W4240), and section number. All homework must be turned in online through Courseworks in two parts: 1) The written part of submitted homework must be in PDF format, have a .pdf extension (lowercase!), and be less than 4MB; and 2) the code portion of submitted homework must be in R and have a .R extension (uppercase!). Homeworks not adhering to these requirements will receive no credit.

# Syllabus and Readings

Readings come from the following texts:

- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning* Springer, 2014.
- Torgo, L. *Data Mining with R.* CRC Press, 2011.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition.* Springer, 2009.

The readings will cover more than we will in class.

Other useful books:

- Adler, J. *R in a Nutshell: A Desktop Quick Reference.* O'Reilly Media, 2010.
- Bishop, C. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006.
- Witten, I. H., Frank, E. and Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufman, 2011.
- Wu, X. and Kumar, V., eds. *Top Ten Algorithms in Data Mining.* CRC Press, 2009.


Table 1: Section 001

| # | Class | Subject | Reading | Other |
|---|---|---|---|---|
| 01 | T Jan 21 | (cancelled) | | |
| 02 | Th Jan 23 | Intro | | |
| 03 | T Jan 28 | Intro to R | James Ch. 2.3 | |
| 04 | Th Jan 30 | Probability Models | Torgo Ch. 1 | |
| 05 | T Feb 4 | Dimension Reduction I | James Ch. 10.1, 10.2, 10.4 | **HW 1 Due** |
| 06 | Th Feb 6 | Dimension Reduction II | | |
| 07 | T Feb 11 | Supervised Learning I | James Ch. 2.1 | |
| 08 | Th Feb 13 | Supervised Learning II | James Ch. 2.2 | |
| 09 | T Feb 18 | Linear Models I | James Ch. 3.1-3.3.1 | **HW 2 Due** |
| 10 | Th Feb 20 | Linear Models II | James Ch. 3.3.2-3.6 | |
| 11 | T Feb 25 | Logistic Regression | James Ch. 4.1-4.3 | |
| 12 | Th Feb 27 | Linear Discriminant Analysis | James Ch. 4.4-4.6 | |
| 13 | T Mar 4 | Naive Bayes | Hastie Ch. 6.6 | **HW 3 Due** |
| 14 | Th Mar 6 | Cross Validation | James Ch. 5.1 | |
| 15 | T Mar 11 | **Midterm** | | |
| 16 | Th Mar 13 | The Bootstrap | James Ch. 5.2-5.3 | |
| | *T Mar 18* | *No Class* | *Spring Break* | |
| | *Th Mar 20* | *No Class* | *Spring Break* | |
| 17 | T Mar 25 | Subset Selection | James Ch. 6.1, 6.5 | |
| 18 | Th Mar 27 | Shrinkage | James Ch. 6.2-6.4, 6.6 | |
| 19 | T Apr 1 | Trees I | James Ch. 8.1 | **HW 4 Due** |
| 20 | Th Apr 3 | Trees II | | |
| 21 | T Apr 8 | Boosting | Hastie 10.1-10.9 | |
| 22 | Th Apr 10 | Bagging and Random Forests | James Ch. 8.2-8.3 | |
| 23 | T Apr 15 | SVMs I | James 9.1-9.2 | **HW 5 Due** |
| 24 | Th Apr 17 | SVMs II | James Ch. 9.3-9.6 | |
| 25 | T Apr 22 | SVMs III | | |
| 26 | Th Apr 24 | Clustering I | James Ch. 10.3, 10.5 | |
| 27 | T Apr 27 | Clustering II | | |
| 28 | Th May 1 | A Priori | Hastie Ch. 14.1-14.2 | **HW 6 Due** |

Table 2: Section 002

| # | Class | Subject | Reading | Other |
|---|---|---|---|---|
| 01 | W Jan 22 | Intro | | |
| 02 | M Jan 27 | Intro to R | James Ch. 2.3 | |
| 03 | W Jan 29 | Probability Models | Torgo Ch. 1 | |
| 04 | M Feb 3 | Data Cleaning | | |
| 05 | W Feb 5 | Dimension Reduction I | James Ch. 10.1, 10.2, 10.4 | **HW 1 Due** |
| 06 | M Feb 10 | Dimension Reduction II | | |
| 07 | W Feb 12 | Supervised Learning I | James Ch. 2.1 | |
| 08 | M Feb 17 | Supervised Learning II | James Ch. 2.2 | |
| 09 | W Feb 19 | Linear Models I | James Ch. 3.1-3.3.1 | **HW 2 Due** |
| 10 | M Feb 24 | Linear Models II | James Ch. 3.3.2-3.6 | |
| 11 | W Feb 26 | Logistic Regression | James Ch. 4.1-4.3 | |
| 12 | M Mar 3 | Linear Discriminant Analysis | James Ch. 4.4-4.6 | |
| 13 | W Mar 5 | Naive Bayes | Hastie Ch. 6.6 | **HW 3 Due** |
| 14 | M Mar 10 | Cross Validation | James Ch. 5.1 | |
| 15 | W Mar 12 | **Midterm** | | |
| | *M Mar 17* | *No Class* | *Spring Break* | |
| | *W Mar 19* | *No Class* | *Spring Break* | |
| 16 | M Mar 24 | The Bootstrap | James Ch. 5.2-5.3 | |
| 17 | W Mar 26 | Subset Selection | James Ch. 6.1, 6.5 | |
| 18 | M Mar 31 | Shrinkage | James Ch. 6.2-6.4, 6.6 | |
| 19 | W Apr 2 | Trees I | James Ch. 8.1 | **HW 4 Due** |
| 20 | M Apr 7 | Trees II | | |
| 21 | W Apr 9 | Boosting | Hastie 10.1-10.9 | |
| 22 | M Apr 14 | Bagging and Random Forests | James Ch. 8.2-8.3 | |
| 23 | W Apr 16 | SVMs I Text | James 9.1-9.2 | **HW 5 Due** |
| 24 | M Apr 21 | SVMs II | James Ch. 9.3-9.6 | |
| 25 | W Apr 23 | SVMs III | | |
| 26 | M Apr 28 | Clustering I | James Ch. 10.3, 10.5 | |
| 27 | W Apr 30 | Clustering II | | |
| 28 | M May 5 | A Priori | Hastie Ch. 14.1-14.2 | **HW 6 Due** |