Chicago Booth 41201: Big Data — Syllabus for Winter 2016

Matt Taddy (mataddy@gmail.com)

COURSE SITE: faculty.chicagobooth.edu/matt.taddy/teaching

TAS: Sergiy Verstyuk (sv.uchicago@gmail.com)
Denise Lau (djlau@uchicago.edu)
Mohsen Mirtaher (mmirtaher@gmail.com)

ASSISTANT: Marshall Smith (marshall.smith@chicagobooth.edu)

OFFICE 340: by appointment please.

COURSE SUMMARY

Inference at Scale: : BUS 41201 is a course on Big Data. Students will learn how to explore and analyze large high-dimensional datasets, become adept at building powerful systems for prediction, and gain the understanding necessary for interpreting structure in such models.

This course includes the key concepts and tools that I see data scientists needing in business environments, and it is also designed to act as a primer for continued study. The course is heavily informed by my experience in large silicon valley firms and interacting with small startups, and I've focused on the topics that I see as most useful in those places. It is not specifically an introduction to computer science or machine learning, nor a class on high-dimensional econometrics and statistics; rather, like a good data scientist, the class borrows from multiple disciplines.

Techniques covered include an advanced overview of linear and logistic regression, model choice and false discovery rates, information criteria and cross validation, regularized regression and the lasso, bagging and the bootstrap, experiments and causal estimation, multinomial and binary regression, classification, latent variable models, principal component analysis, topic models, decision trees and random forests, text analysis and natural language processing.

We learn both basic underlying concepts and practical computational skills, including techniques for scalable analysis of distributed data. Heavy emphasis is placed on analysis of actual datasets, and on development of application specific methodology. Among other examples, we will consider consumer database mining, internet and social media tracking, asset pricing, network analysis, sports analytics, and text mining.

PREREQUISITES

BUS 41000 or 41100. Prerequisite material includes fundamentals of probability and the following: random variables (and functions thereof), normal and multinomial distributions, confidence/prediction intervals, hypothesis testing and sampling distributions. In particular, you should be comfortable with the basics of linear regression as covered in 41000 or 41100. It is your responsibility to ensure that you have a complete command of these concepts.

Waivers: Perhaps you have a mastery of statistics and regression, from work or previous classes. I do often grant access to the course to people with quantitative background, but the onus of preparation and self-assessment is entirely on the student. It is completely your responsibility to ensure

that you have the necessary background for the course, and to withdraw quickly otherwise. This is a difficult and fast-paced course, so it is easy to fall behind if there are gaps in your background.

In addition, this course does not cover basic statistics material that is assumed for other courses at Booth. Such classes may not accept BD as a replacement for 41000 or 41100 prerequisites.

All computing is conducted in R, a platform for statistical analysis. To make it possible to focus on data science concepts, I strongly encourage students to learn the basics of the language and software BEFORE starting the class. See the next section for detail and resources.

A pre-class example analysis (R script) has been posted to the class webpage. You should be able to run all of the steps in this example, and interpret the analysis. Questions are in the script comments.

COMPUTING

This class uses R, which is available for free via `www.r-project.org`. You can download and install the software following directions at `cran.us.r-project.org` (do this ASAP).

R is a widely used and hugely flexible analysis platform. It has a command line interface (you type commands to get what you want). Some students find the learning curve for such 'programming' to be very steep. I provide limited software instruction, in-class demonstration, and code to accompany lectures and assignments. We don't assume that you have used R in a previous class.

However, this is not a class on R. Like any language, R is only learned by doing. You should install R as soon as possible and familiarize yourself with basic operations. Ideally, you would start this course able to replicate any analysis from previous classes (e.g., 41000 or 41100) in R.

A great way to start learning is to buy a book and start working through tutorials. A good guide is Adler's *R in a Nutshell*. They have many tutorials to help you get up to speed. You can browse other options by searching 'R statistics' on Amazon.

If you are new to R (and even if not) you should complete a tutorial to familiarize yourself with the language. A great option is the **TryR code school** @ `http://tryr.codeschool.com`.

Additional resources

- Tutorials at `data.princeton.edu/R` are fantastic (and there are many others out there).
- I've linked to short reference cards online. Keep one of these handy (and add to it).
- youtube intros to R, e.g. the series from Google Developers linked to on our teaching page.
- Me and your classmates: work together, and chat on the discussion board.
- Rstudio is a free platform for both writing and running R, available at `www.rstudio.org`. Some students find it friendlier than basic R (especially in windows OS).
- The UC library has many e-books on R available. A search link is on my teaching page.
- Search: "do *X* in R". Try variations on *X* until you find an answer. Many answers will live on `stackoverflow.com`, which is a site where programmers compete to be helpful.

Links to anything else that looks useful will be posted in Piazza.

QUESTION & ANSWER

We're using Piazza for class discussion. Rather than emailing your questions, try to post on Piazza. If you have any problems, you can email me or `team@piazza.com`. Class page is at:
`https://piazza.com/chicagobooth/winter2016/41201/home`

Feel free to answer your classmate's questions; it's a huge help to us, and even if you're wrong everyone learns (we check the answers and clear up confusion). Also, we encourage you to take credit for your questions and answers (rather than posting anonymously)!

TEXT/NOTES

There is no required textbook: all materials will be available on the class website. The best preparation you can do before lectures is to go through class R code and work through examples.

A solid primer is *An Introduction to Statistical Learning*, by James, Witten, Hastie, and Tibshrani. However, it takes a very different approach from us and only partially overlaps on material.

A great advanced text is *Elements of Statistical Learning* by Hastie, Tibshrani, and Friedman, but it requires some mathematical sophistication and goes beyond the material we will be covering. The book is free at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Also good, but still advanced, are *Pattern Recognition & Machine Learning* by Bishop and (the encyclopedic) *Machine Learning* by Murphy. Both of these books introduce the material from a more computer engineering (or AI), rather than statistical, perspective.

EVALUATION

Grades will be determined by **homework (10%)**, a **take-home midterm exam (45%)**, and a **final project (45%). Late assignments, exams, or projects, will not be accepted.**

Homeworks are due weekly except for the 1st and 6th (midterm) classes.

Individual Midterm Take-Home Exam is assigned in week 5 and due in week 6.

Final Project will be due during exam week. There is no class that week.

Students pledge to adhere to *Booth Honor Code* standards on all work.

There will be 8 one-week homework assignments. Students are encouraged to form groups (max size 4) for homework and only one write-up per group need be turned in. Only the best 7 of 8 assignments will be counted towards your grade.

The individual midterm exam will be posted at the beginning of week 5, and is due the next week.

The final project can be done as group work, and students will develop their own Big Data application. Planning should begin soon after the midterm. Previous projects have included health risk prediction, social media analysis, text mining, and various machine learning contests.

All work will be submitted on-line; see detailed instructions on Piazza.

FAQs

Do you assign provisional grades?

As needed, but only after the midterm is graded.

I am going to miss a lecture. What should I do?

If there are seats available, you are free to attend an alternate section. If this is not possible, you will need to catch up by reading the lecture slides and talking to group mates and reviewing their notes. Try the homework, and then come to me if you have outstanding issues. I don't recommend making a habit of missing lectures: there is a lot of content that is not in the slides. Unfortunately, the Deans Office will not allow me to tape lectures.

Can we form groups across sections?

Yes. You need to hand-in all work before the first session attended by any member of the group.

Can I audit the course?

I discourage auditing. The course is very hands-on and project based; without that (and without group interaction), it's probably not worthwhile for either of us. If you really want to audit please introduce yourself before class. Space has also been an issue in the past.

Are there review sessions?

No. My experience in teaching this material is that review sessions just distract students from actually working on problems. The best way to learn is by doing (especially with programming), and we have such a wide breadth of backgrounds in the course that each student will get stuck on different issues. Instead, I go to great lengths to be accessible for specific questions about projects, homework, and class notes. This may be different from how you are treated in other classes, so make sure that you are working through examples and reaching out when you hit obstacles.