POL 245: Visualizing Data

Summer 2015

James Lo, Will Lowe (Instructors)
Winston Chou, Elisha Cohen (Preceptors)
Alex Tarr (QuantLab Coordinator)
Kosuke Imai (Course Head)

Department of Politics, Princeton University

In this course, we consider ways to illustrate compelling stories hidden in a blizzard of data. Equal parts art, programming, and statistical reasoning, data visualization is a critical tool for anyone doing analysis. In recent years, data analysis skills have become essential for those pursuing careers in policy advocacy and evaluation, business consulting and management, or academic research in the fields of education, health, medicine, and social science. This course introduces students to the powerful R programming language and the basics of creating data-analytic graphics in R. From there, we use real datasets to explore topics ranging from network data (like social interactions on Facebook or trade between counties) to geographical data (like county-level election returns in the US or the spatial distribution of insurgent attacks in Afghanistan). No prior background in statistics or programming is required or expected.

Contact Information

Name	James Lo	William Lowe	Kosuke Imai
Office	Corwin 029	TBA	Corwin 036
Email	`jameslo@princeton.edu`	`will.lowe@uni-mannheim.de`	`kimai@princeton.edu`

Name	Winston Chou	Elisha Cohen	Alex Tarr
Office	Corwin 127	Corwin 127	TBA
Email	`wchou@princeton.edu`	`eacohen1@princeton.edu`	`atarr@princeton.edu`

During our office hours we may be in either the office that is listed above or in Corwin 023 (just across the hall) which has space for more students. If these office hours do not fit your schedule, you should feel free to contact us directly via email. Also, do not forget about Piazza and QuantLab (see below) where you can ask questions and receive answers back immediately.

Logistics

The schedule during the first week deviates from this, details are below in the Course Outline section.

Lectures. Monday and Wednesday, 1:30pm–2:30pm, Sherrerd Hall 101. To make lectures interactive, lecture slides will be posted on Blackboard immediately *after* the lecture. However, students are expected to take notes during the lecture.

Precepts.   Tuesday and Thursday, 1:30pm–2:50pm, Frist Campus Center 307, 309, and 329. We ask you to bring your personal laptop to precepts.

QuantLabs.   Monday, Tuesday, and Thursday, 3:00pm - 4:30pm, Frist Campus Center 307, 309, and 329. (following immediately after the lecture on Mondays and the precepts on Tuesdays and Thursdays) in the same room as your precepts. You will be working with tutors on review questions, practice exercises, and problem sets. Bring your own laptop to the QuantLabs.

Guest Lectures.   Friday, 10:30am–11:50am, Wallace Hall 300. These sessions occur during the second week of FSI through the final week. They involve guest speakers from various industries where data visualization is used.

Lunch with Guest Speaker.   Friday, 12:00pm–1:30pm, Prospect House. *The lunch with the last speaker will be held at Mediterra, a downtown Princeton restaurant.* Students will sign up to have lunch with one of the five guest speakers at the beginning of the course. During the selected week, students and the course team will meet with the guest speaker during a casual, catered lunch.

Course Requirements

- **Class participation (15%):** Students should actively participate in all aspects of the course. Class participation will be judged based on questions asked/answered during the lectures, the precepts, and on the online discussion board. Each portion is equally weighted.

- **Review Questions (15%):** During the QuantLab, students will work on the assigned portion of the textbook and electronically submit a small set of *Review Questions*. The answers to *Review Questions* will be graded pass fail. Details on these assignments are announced at the QuantLab. *This is an individual assessment with limited collaboration.*

- **Problem sets (50%):** Each week will end with the posting of a problem set. These assignments will be posted on Thursday at the end of QuantLab via Blackboard. Hard copies of your problem sets must be turned in at the beginning of the Tuesday precept. Electronic submission of your computer code via Blackboard must also be done by then. Each problem set will be equally weighted. *This is an individual assessment with no collaboration.*

- **Final Project (20%):** This is a group data analysis project. Students will be assigned to groups. Analyzing a data set of their choice, students will write a report of no more than 1,000 words summarizing a compelling relationship or story they identified in the data. No more than 3 figures/tables can be used. Details regarding the final project will be announced later in the course. *This is a group assessment with collaboration allowed only within the assigned groups.*

Collaboration Policy

The assignments in this course are designated as individual or group assessments. The degree of permissible collaboration depends on the kind of assignment:

- **Review Questions.** Students are encouraged to interact with each other, the instruction team, and QuantLab tutors in discussing their approaches and solutions. This includes conceptual discussion and actual computer code. *However, for all other assignments, this degree of collaboration is not appropriate!*

- **Problem Sets.** No collaboration is allowed. Students may ask clarifying questions regarding problem set and midterm questions to the instruction team through Piazza. This allows all students to benefit from clarifications equally. Clarifying questions about the problem sets may not be asked of QuantLab tutors, however.

- **Final Project.** Students may fully collaborate within their assigned groups, and may discuss their group's work with other students, the instruction team, and QuantLab tutors.

Plagiarism Policy

Violations of the above collaboration policy will be treated as instances of plagiarism. This course will follow a modified version of the guidelines used for computer science classes here at Princeton. *Please take this guideline seriously.* In the past, plagiarism cases typically result in one-year suspension from Princeton.

Programming necessitates that you reach your own understanding of the problem and discover a path to its solution. DO NOT, UNDER ANY CIRCUMSTANCES, COPY ANOTHER PERSON'S CODE. Incorporating someone else's code into your program in any form is a violation of academic regulations. Abetting plagiarism or unauthorized collaboration by sharing your code is also prohibited. Sharing code in digital form is an especially egregious violation: do not e-mail your code to anyone.

Novices often have the misconception that copying and mechanically transforming a program (by rearranging independent code, renaming variables, or similar operations) makes it something different. Actually, identifying plagiarized source code is easier than you might think. For example, there exists computer software that can detect plagiarism.

This policy supplements the University's academic regulations, making explicit what constitutes a violation for this course. Princeton Rights, Rules, Responsibilities handbook asserts:

> The only adequate defense for a student accused of an academic violation is that the work in question does not, in fact, constitute a violation. Neither the defense that the student was ignorant of the regulations concerning academic violations nor the defense that the student was under pressure at the time the violation was committed is considered an adequate defense.

If you have any questions about these matters, please consult a member of the instruction team.

Textbook

This course uses a draft manuscript of the following textbook.

> Imai, Kosuke. *Quantitative Social Science: An Introduction.* Under contract with Princeton University Press.

The textbook is made freely available to the students at Blackboard. Due to the copyright issues, this file should not be distributed to those who are not taking this class.

Statistical Software

In this course, we use the open-source statistical software **R** (http://www.r-project.org). **R** can be more powerful than other statistical software such as SPSS, STATA and SAS, but it can also be more difficult to learn. A variety of resources will be made available for POL 245 students in order to learn **R** as efficiently as possible. To help make using **R** easier, we'll be using **RStudio** (http://www.rstudio.com/)—a user-interface that simplifies many common operations.

Get Help

Many students will find the materials of this course to be challenging. As such, students must seek immediate help when struggling with the course. There are several ways, in which students can get in-person and online help.

In-person Help

- Office Hours: 2:00pm to 3:30pm on Fridays in Corwin 127. You will be able to ask the instruction team any questions you might have about the course materials. You may also e-mail to set up an appointment outside of the office hours.

- Problem Set Help Sessions: 7:00pm to 9:00pm on Sundays in Hargadon G001, G002, and G004 (located in Baker Hall, Whitman College) and QuantLab, 3:00pm to 4:30pm on Mondays. Tutors will not give you direct guidance on the actual problem set questions but will help you understand the concepts required for solving them.

- R Drop-in Office Hours: 3:30pm to 4:30pm on Fridays, Butler 028. Dima Gorenshteyn will be available to answer any questions about R programming.

- R Workshop: 4:30pm to 6:00pm on Fridays, Butler 028. Dima Gorenshteyn will run a short workshop that cover tricky R programming concepts introduced in each week. The details of a workshop will be announced each week.

On-line Help

In addition to office hours and individual appointments, we will be available online to answer any questions you may have about the course materials and the problem sets. We use the Piazza discussion forum that will be linked on Blackboard course page or accessible directly at `http://piazza.com`.

Before posting your question, please review previous posts to make sure that a similar question has not been answered. *In accordance with the collaboration policy described above, you should not directly post your code for a problem set.* You should frame your questions in general terms rather than trying to have us debug your code directly. You may subscribe to the Discussion Forum so that you receive your fellow students' questions and answers to those questions. You should also feel free to respond to questions that you can answer. Piazza also has a free smartphone application if you are interested.

Course Outline

Introduction: July 21 – July 22

During the first two days of the course, you will be introduced to R statistical programming environment through the use of **RStudio**.

Note, the first session on Tuesday, July 21 will be a lecture in Sherrerd Hall 101 and not a precept. The second session on Wednesday, July 22 will be a precept in the Frist Campus Center. The third session on Thursday, July 23 will be a lecture in Sherrerd Hall. And, for Friday, July 24 we will hold a precept in Frist Campus Center at 10:30 am. These changes affect only the first week of the course.

Lecture	July 21	Overview of the course
QuantLab	July 21	Work on Chapter 1; Submit Chapter 1 Review Questions 1 and 2
Precept	July 22	Application: Understanding World Population Dynamics

Causality: July 22 – July 28

We will learn how to infer causality from data. We learn the distinction between randomized experiments and observational studies. Our applications include the evaluation of strategies for increasing voter turnout and the effect of class size on educational achievement.

QuantLab 1	July 22	Work on Chapter 2 (2.1–2.4); Submit Chapter 2 Review Questions 1
Lecture 1	July 23	Experiments
QuantLab 2	July 23	Work on Chapter 2 (2.5–2.7); Submit Chapter 2 Review Questions 2
Precept 1	July 24	Application: Efficacy of Small-class Size in Primary Education
Lecture 2	July 27	Observational Studies
Precept 2	July 28	Application: Success of Leader Assassination as a Natural Experiment

Measurement: July 28 – August 4

We consider how to measure public opinion using sample surveys. We also learn about a measurement strategy regarding latent concepts like ideology. Our applications include surveys in Afghanistan and political polarization in US Congress.

QuantLab 1	July 28	Work on Chapter 3 (3.1–3.4); Submit Chapter 3 Review Questions 1
Lecture 1	July 29	Survey Sampling
Precept 1	July 29	Application: Political Efficacy in China and Mexico
QuantLab 2	July 30	Work on Chapter 3 (3.5–3.8); Submit Chapter 3 Review Questions 2
Lecture 2	August 3	Measurement and Clustering
Precept 2	August 4	Application: Voting in the United Nations General Assembly

Prediction: August 4 – August 13

We learn about prediction starting with the application of US presidential election forecasting. Students will be introduced to linear regression and how it is related to causality.

QuantLab 1	August 4	Work on Chapter 4 (4.1); Submit Chapter 4 Review Questions 1
Lecture 1	August 5	Prediction and Loop
Precept 1	August 6	Application: Prediction Based on Betting Markets
QuantLab 2	August 7	Work on Chapter 4 (4.2); Submit Chapter 4 Review Questions 2
Lecture 2	August 10	Regression
Precept 2	August 11	Application: Election and Conditional Cash Transfer Program in Mexico
QuantLab 11	August 11	Work on Chapter 4 (4.3–4.4); Submit Review Questions 3
Lecture 8	August 12	Regression and Causation
Precept 8	August 13	Application: Government Transfer and Poverty Reduction in Brazil

Discovery: August 13 – August 25

We cover how to analyze three different types of data; textual data, network data, and spatial data. Our applications include the prediction of disputed authorship of The Federalist Papers, the marriage network in Renaissance Florence, and the expansion of Wal-mart.

QuantLab 1	August 13	Work on Chapter 5 (5.1); Submit Chapter 5 Review Questions 1
Lecture 1	August 17	Textual Data
Precept 1	August 18	Application: Predicting Blog Tone
QuantLab 2	August 18	Work on Chapter 5 (5.2); Submit Review Questions 2
Lecture 2	August 19	Network Data
Precept 2	August 20	Application: International Trade Network
QuantLab 3	August 20	Work on Chapter 5 (5.3); Submit Review Questions 3
Lecture 3	August 24	Spatial Data
Precept 3	August 25	Application: Spatial Mapping US Election Results Over Time

Final Project: August 25 – August 30

Students will work on the final project of the course in a small group with the help from the instruction team and QuantLab tutors.

QuantLab	August 25	Final Project Work Session
Lecture	August 26	Wrapping Up
Precept	August 27	Final Project Work Session
QuantLab	August 27	Final Project Work Session

POL 245 Schedule

Due to the condensed nature of the course, the semester and weekly schedules (including assignments) can seem very full. To help with this, we have a Google Calendar with events, due dates, and sessions. The URL for this calendar is `http://goo.gl/D3pKXL`. We also outline a typical week of assignments in POL 245 below:

Sun	Mon	Tue	Wed	Thur	Fri	Sat
7:00pm–9:00pm Problem Set Help Session	**1:30pm–2:30pm** Lecture **3:00pm–4:30pm** Problem Set Help Session	**1:30pm** Problem Set deadline **1:30pm–2:50pm** Precept **3:00pm–4:30pm** QuantLab	**1:30pm–2:30pm** Lecture	**1:30pm–2:50pm** Precept **3:00pm–4:30pm** QuantLab **4:30pm** New Problem Set posted	**10:30am–11:50am** Guest Speaker **2:00pm–3:30pm** Office hours **3:30pm–4:30pm** **R** Office hours **4:30pm–6:00pm** **R** Workshop	

Guest Speakers

For five of the six weeks during the course, we will have guest speakers during our Friday sessions. In addition to their presentation from 10:30am – 11:50am in Wallace Hall 300, they will be on campus for lunch with a small group of students. Students will sign up for lunch with a specific speaker at the beginning of the course.

FSI Week	Date	Guest	Affiliation
2	7/31	Rebecca Lai *Graphics Intern* VISUALIZING THE NEWS	New York Times
3	8/7	Aaron Strauss *Executive Director* HOW CAMPAIGNS USE ANALYTICS AND EXPERIMENTS TO INFLUENCE VOTERS	Analyst Institute
4	8/14	Dan Chapsky *Advertisement Data Scientist* TRUTH, BEAUTY AND SOCIAL DATA: USING OPEN DATA ONLINE TO PREDICT OFFLINE EVENTS	Facebook
5	8/21	Elizabeth Roodhouse *Social Scientist* BIG DATA AT GOOGLE & YOUTUBE: EXPLORING TRENDS IN THE ONLINE VIDEO LANDSCAPE	Google
6	8/28	Neil Paine *Senior Sportswriter* SPORTS AND DATA JOURNALISM IN THE POST-MONEYBALL ERA	FiveThirtyEight