Building Curriculum and Infrastructure for Quantitative Social Science

Kosuke Imai

Department of Politics
Princeton University

Talk at the School of Political Science and Economics
Waseda University
June 30, 2015

Quantitative Social Science Today

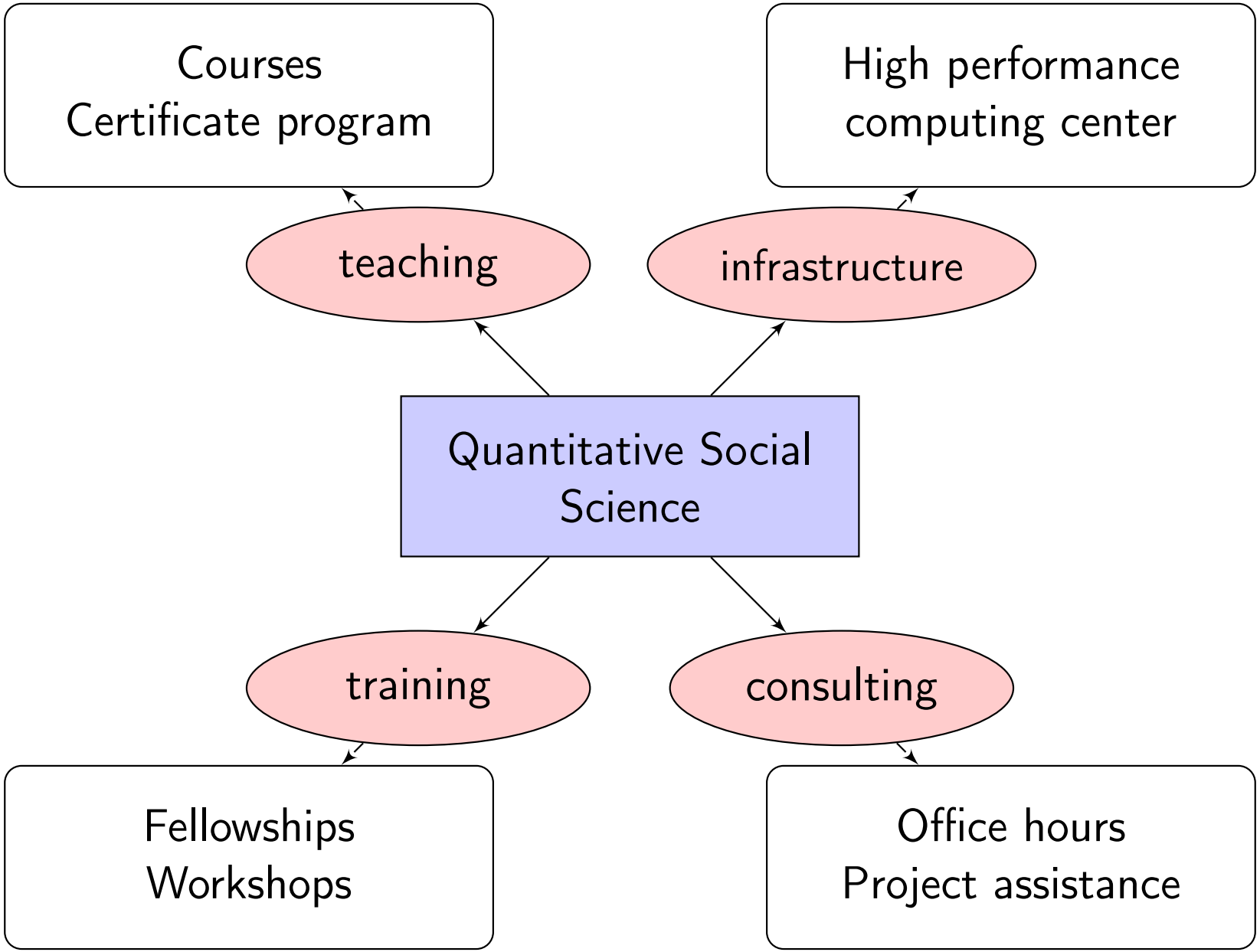
- Internet and computing revolution over the last two decades
- Transformed social science research and education
- Data, Data, and Data!
- Past: government data, national survey data
- Today: more of old types of data and lots of new data
  - Randomized experiments and surveys conducted by researchers
  - Administration records: voter files, contributions, lobbying, ...
  - Economic data: trade, company information, finance, ...
  - Military data: casualty, insurgent attacks, ...
  - Social media data: websites, blogs, tweets, cell phones, ...
  - GIS data: satellite, climate, natural resource discoveries, ...
  - Text, images, sounds: news, speeches, bills, commercials, ...

Challenges

1. Data availability is NOT the problem
   - Need for new algorithms, models, and methods
   - Need for new computing technologies
     - parallel computing
     - natural language, image, and video processing
     - database management
     - data visualization

2. Everyone, not just statisticians and methodologists, are analyzing data
   - Need for methodological training at all levels
   - Undergrads, not just graduate students
     - Curriculum development
     - Independent research
     - Job market in virtually all areas

Overview of Quantitative Social Science at Princeton



Undergraduate Teaching: Let's Look at the Data First

- Introductory statistics courses are unpopular
  - Non-politics introductory statistics courses in social sciences
  - 5 year average: 2008/09–2013/14
  - ECO 302, PSY 251, SOC 301, WWS 200, WWS 332

	Lectures	Assignments	Readings	Precepts	Overall
Statistics	3.2	3.3	3.1	3.6	3.1
All PU courses	3.8	3.7	3.7	4.0	3.9

- Politics introductory statistics courses:

	Lectures	Assignments	Readings	Precepts	Overall
POL 245	4.4	3.9	3.5	3.9	4.3
POL 345	4.0	3.8	3.7	4.2	4.1

What Students are Saying about the Course

- A happy student ☺:
  "It is an immensely difficult course. However, I am happy to have learned a valuable skill."

- An unhappy student ☹:
  "I should have dropped this course in the first week and signed up for something easier – that's my fault."

- What's the secret of success?

Why Teaching Introductory Statistics Courses is Hard

1. Students are not interested in statistics:

	Professor	Distribution Requirement	Departmental	Certificate Program	General Interest
Statistics	0%	20%	71%	3%	6%
All PU courses	6%	12%	32%	7%	42%

"Professor Imai tried hard to make statistics interesting. But, statistics is boring."

2. Students have weak mathematical and programming background

"as a person not naturally inclined towards statistics and probability, I don't feel at all qualified to pass judgement on how the course might have been improved."

The Problems of Current Teaching

- Standard introductory statistics course:
  1. Descriptive statistics:
     - univariate: histogram, mean, median, standard deviation, ...
     - bivariate : scatterplot, correlation, regression, ...
  2. Probability:
     binomial, normal, law of large numbers, central limit theorem, ...
  3. Statistical inference:
     hypothesis tests, standard errors, confidence intervals, ...
- The Problem for social science students: Boring!!
- Other problems:
  - only teaches statistical concepts but not actual data analysis
  - often does not teach basic statistical programming
- How can we make these materials more fun and relevant?

Development of New Curriculum (and Textbook)

- Key ideas:
  1. Teach data analysis *before* probability and statistics
  2. Emphasize applications rather than methods themselves
  3. Teach statistical programming as well as concepts

- Overview:
  - Teach applications, concepts, and programming altogether
  - Use of **R** – code chunks embedded in the book
  - Examples and exercises drawn from actual research
  - Chapters
    1. Causality
    2. Measurement
    3. Prediction
    4. Discovery
    5. Probability
    6. Uncertainty

Detailed Contents

Applications	Concepts	Programming
Introduction		
demographic statistics (birth and death rates); voting in Weimar Republic	sample average; aggregation bias	arithmetic operations; functions; loading and saving data
Causality		
racial discrimination; social pressure and turnout; conditional cash transfer and incumbency support; minimum wage and unemployment;	potential outcomes; randomized experiments; instrumental variables; confounding; observational studies; difference-in-differences; standard deviation; quantiles	logical values and operators; relational operators; subsetting; conditional statements

Applications	Concepts	Programming
Measurement		
public opinion survey in Afghanistan; list experiment; academic integrity and changing minds on gay marriage; ideal points and political polarization	random sampling; non-response bias; distribution; density; correlation; clustering; k-means algorithm	handling missing data; basic graphics (barplot, boxplot, histogram, scatterplot, qq-plot)
Prediction		
pre-election polling; betting markets and election forecasting; facial appearance and election outcomes; women as policy makers; return to political office in Britain	prediction and classification error; linear models; least squares; regression to the mean; residuals; R^2; regression and causality; regression discontinuity design	loop; merging data sets

Applications	Concepts	Programming
Discovery		
Federalist papers and authorship prediction; Marriage network in Renaissance Florence; Twitter following among politicians; John Snow and Cholera; Spatial patterns of US presidential elections; Expansion of Walmart	Basic text analysis (document-term matrix, tf-idf, topic discovery); Basic network analysis (undirected and directed networks, weighted and unweighted networks, centrality measures, PageRank algorithm, community detection)	list and matrix; use of packages (text analysis, network analysis, spatial analysis); visualization using various packages; animation

- Two more chapters – Probability and Uncertainty – to be written
- Full manuscript available in September

Motivating Students

1. Statistics as a necessary tool for modern social science research
   - Junior papers and senior thesis
   - Reanalysis of data from published research
2. Statistics as a useful skill for post-graduate career
   - Guest speakers from industry
   - Stories from course alumni/alumnus in various industries

Recent Emails from Alumni

- Senoir thesis research:
  "I took your POL 345 course last semester, and I think I'm going to need to use statistical analysis skills in my research this semester. I'm currently studying abroad in Beijing, and working on a research project to assess which methods can best improve the quality of education and access to education for migrant children in Beijing..."

- Graduate school:
  "When I first decided to take Professor Imai's statistics course as a senior majoring in Politics at Princeton, I was excited at the idea of gaining new skill sets but wasn't entirely sure why statistics and learning R would be useful for me. In retrospect, I am very grateful for having taken the course, which helped me become fully ready for my present graduate studies in political science..."

- Finance:
  "It was a pleasure, and quite a coincidence, running into on Wall St. the other week. I have actually been meaning to send you an e-mail to thank you for what Pol 345 has done for me. I recently returned from London for two months of training with DB, and spent a large majority of the time discussing options pricing models, which is highly contingent on statistical assumptions..."

- Small-town newspaper:
  "I graduated from Princeton last June, and took POL 345 a year ago. I was a history major, and my job now has little to do with statistics – I'm a sports reporter for a small-town newspaper. But I did find a way to employ **R** quite usefully..."

Helping Students Learn Efficiently

1. Short but frequent assignments
   - Pre-precept assignments
   - Problem sets, quizzes
2. Hands-on instruction in computer labs
   - Detailed handouts
   - Practice exercises
3. Assistance outside of the classroom
   - Extra office hours, Peer tutoring
   - Piazza online discussion board

Freshman Scholars Institute as a Testing Case

- 6-week long summer school for a subset of incoming freshman
- an incubator for new courses
- Enrollment: 30 – 40 students
  - come from "disadvantaged" background
  - first generation college students
  - lack mathematical and computing background
- Goals:
  - transition them from high school to college
  - get them used to Princeton before the semester starts
  - offer head start by earning early PU course credits
- Similar programs at other schools: http://nyti.ms/1gjJOoU
- A hard test for our teaching strategies

Structure of the Course

- Module format for each week:
  1. Two 50 minute lectures
  2. Two 80 minute lab sessions
  3. One 80 minute guest lecture from industry, discussion over lunch (NYT, Facebook, Google, Political consulting firm, FiveThirtyEight)
  4. Three optional tutoring sessions with additional exercises

- Assignments:
  1. 5 problem sets with no collaboration
  2. 12 short non-graded pre-lab assignments
  3. Collaborative final project

After the Introductory Statistics Course

1. Second statistics course:
   - More regression modeling (e.g., logistic regression)
   - More causal inference (e.g., matching)
   - Simple panel data analysis (e.g., fixed effects regression)
2. Certificate Program in Statistics and Machine Learning
   - No double major at Princeton
   - New effort to coordinate SML teaching across departments
   - Five course requirement including Intro Stat and Intro ML
   - Independent research involving SML, poster presentation
   - More information at http://csml.princeton.edu
3. Results:
   - increasingly high enrollment
   - more senior theses with quantitative methods
   - more undergrads in the graduate methods sequence
   - more undergrads who apply (and get accepted) into PhD programs

The Graduate Methods Sequence

- Four course sequence:
  1. probability theory and mathematical statistics
  2. cross section data analysis
  3. panel data and survival analysis
  4. advanced topics (Bayesian statistics, Machine learning, etc.)
- Intermediate level: good use of calculus but no measure theory
- Some formal derivations but many applied data analysis exercises
- Methodologists take classes in computer science and statistics
- Goal: produce sophisticated applied empirical researchers
- Fast changing methodology: need to build a solid foundation
- Result: rapidly improved placements (Harvard, MIT, Yale, UCLA, etc.)
- More information at http://imai.princeton.edu/teaching/

Infrastructure

1. High performance computing center at Princeton
   - Collaboration across different departments
   - Originally built for natural science departments
   - Social sciences and humanities are now utilizing too
2. Computing clusters available to social scientists
   - Della (production cluster): 1536 cores, 4GB/8GB per core
   - Tukey (politics cluster): 384 cores, 3GB per core
   - Experimental "BigData" Hadoop cluster
3. Other resources and services
   - Visualization lab with a wall-size high-resolution display
   - GIS library: training and research support
   - Support for Subversion, GitHub, and Trac
   - Guest account services for non-Princeton collaborators
4. More information available at
   http://www.princeton.edu/researchcomputing/

Training

1. Statistical Programming Camp
   - Pre-requisite for graduate-level statistics courses
   - Taught by graduate students
   - Basic Data analysis and R programming
   - Our undergraduate introductory course taught in one week
2. Advanced Statistical Programming Camp
   - Taught by a research specialist
   - Integration of C++ and R
   - Parallel computing in R and C++
   - How to use cluster computing resources
3. Web-scraping and Text Processing Workshop
   - Taught by a research specialist
   - Python programming for scraping, text processing, twitter API etc.

- Other shorter workshops: LaTeX, Unix, Python, etc.
- More information at http://q-aps.princeton.edu

Consulting

- Consulting services run by research specialists and postdocs
- 2 hour session $\times$ 3 per week, free of charge
- Main clients:
  - Undergraduates: junior and senior independent research
  - Graduate students and faculty: research projects
- Consulting activities (2013–2014):
  - 430 sessions
  - 360 hours
  - 130 unique clients

- Breakdown by level:
  - Graduate students: 58%
  - Undergraduate students: 12%
  - Faculty: 17%

- Breakdown by department:
  - politics: 65%
  - other social sciences: 32%
  - natural sciences: 3%

Concluding Remarks

- Everyone – undergrads, grad students, faculty – analyzes data
- Need for better curriculum and infrastructure at all levels
  - Undergraduate sequence
  - Graduate sequence
- Importance of human resources: teaching, training, consulting
  - methodologist faculty
  - research specialists
  - postdocs, graduate and undergraduate fellows
- Challenge at Princeton: coordination across departments
- Creation of the Center for Statistics and Machine Learning
- All politics are local: things that work at one institution may not work at another institution