POIR 613: Computational Social Science

Pablo Barberá

University of Southern California
pablobarbera.com

Course website:
pablobarbera.com/POIR613/

→

Shift in communication patterns

How can we analyze *digital trace data* to answer Political Science questions?

POIR 613

Goals

- Read and evaluate research applying computational methods to political science problems
- Learn how to collect and manipulate digital trace data
- Develop skills necessary to analyze large and heterogeneous quantitative datasets

Outline (see detailed scheduled here)

- Weeks 1-3: Introduction. Ethics. Experiments
- Weeks 4-9: Text as data methods
- Weeks 10-12: Social network analysis
- Week 13: “Big Data”

Hello!

About me

- Pablo (he/his/him)
- Associate Professor at USC POIR; Research scientist in Core Data Science team at Meta
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU (2015–16), Assistant Professor at LSE Methodology (2017–18)
- My research:
  - Social media and politics, comparative electoral behavior
  - Computational methods: text as data, social network analysis, large-scale data manipulation
- Contact:
  - `pbarbera@usc.edu`
  - `www.pablobarbera.com`
  - Office hours: Mondays 4-5pm, Fridays 9-10am (Zoom)

Your turn!

1. Name? Pronouns?
2. Department, year?
3. Research interests?
4. Previous experience with R?
5. Why are you interested in this course?

The plan for today

- Introductions
- Logistics
- What is CSS?
- R and RStudio
- GitHub and version control

Course philosophy

How to learn the techniques in this course?

- Lecture approach: not ideal for learning computational social science methods
- You can only learn by doing:
  - → Reading and criticizing research
  - → Applying methods to social science problems
- Structure of each session:
  1. Introduction to the topic (30 minutes)
  2. Discussion of research (50 minutes)
  3. Guided coding session (30-40 minutes)
  4. Coding challenges (30 minutes)
- You will continue working on the coding challenges after class and submit before beginning of next class

Course website

POIR 613 | Overview | Syllabus | Project | Code | Resources

- POIR 613
- Instructor
- Schedule
- Prerequisites
- Course structure
- Course requirements and grading
- Software
- License and credit
- Feedback

POIR 613

Computational Social Science

University of Southern California, Fall 2021

Citizens across the globe spend an increasing proportion of their daily lives online. Their activities leave behind granular, time-stamped footprints of human behavior and personal interactions that represent a new and exciting source of data to study standing questions about political and social behavior. At the same time, the volume and heterogeneity of digital data present unprecedented methodological challenges. The goal of this course is to introduce students to new computational social science methods and tools required to explore and harness the potential of digital trace data using the R programming language.

The course will follow a “learning-by-doing” approach and will place emphasis on gaining experience in analyzing data using R. Students are expected to do the required readings and coding exercises for each week. The lectures will build upon the content of the readings with a series of data challenges that will introduce new statistical and programming concepts, which will then be applied to the analysis of data from published research papers or common tasks in computational social science. Most of the applications will be related to Political Science and International Relations questions, but the course should be of interest to social science students more generally.

pablobarbera.com/POIR613

Evaluation

- **Class participation**: 10%
  - Do all “readings for discussion” (required)
  - If unfamiliar with topic, also background reading
- **Referee reports and presentations**: 20%
  - TWO peer reviews (800-1000 words) of readings for discussion, due 8pm day before the class via email
  - 12-minute presentation in class
- **Coding challenges**: 20%
  - Not graded but submission (.Rmd + html/pdf files) of at least FIVE is required before next class
- **Research project**: 50%
  - Original research paper (8,000 words) that employs computational methods in political science. Individual or group project (up to 3 people)

Research project

Goal: demonstrate ability to conduct research that applies computational methods to political science questions.

Key deadlines:

10/14 Project idea (one page)

11/04 Descriptive statistics (5 pages)

12/02 First full draft (10-15 pages)

12/15 Final paper due

See course website for more information.

Learning in this (endless) age of covid

During these uncertain times...

- If you feel sick, feel free to miss sessions. Please send me a quick email to let me know.
- Same for problem set / project deadlines.
- I'll try to share lecture recordings if possible
- Feel free to come to office hours to catch up
- Don't worry about grades – you're here to learn

Communication

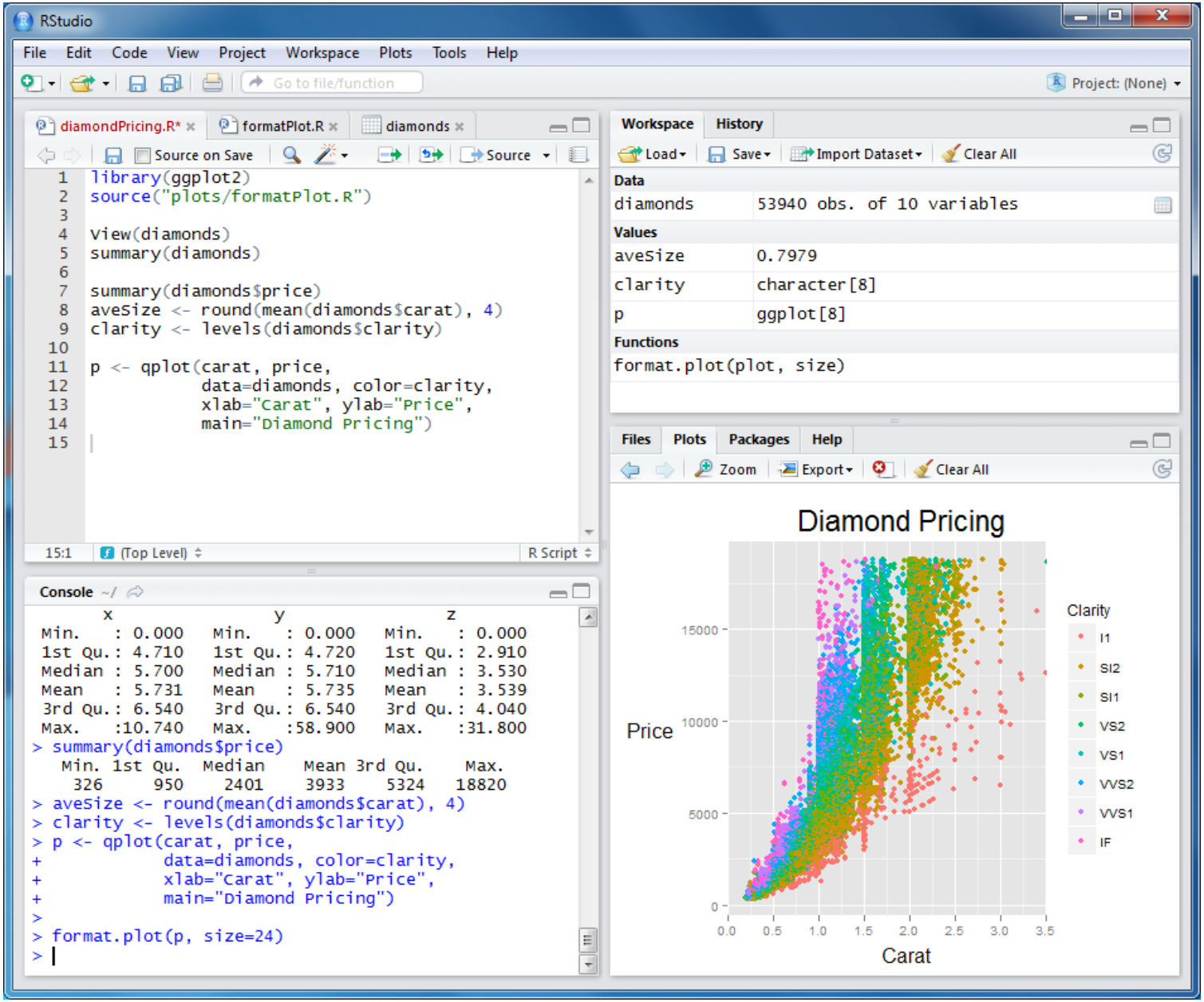
- We will use **Slack** for most course-related communication:
  - Announcements and reminders
  - Questions that you would regularly ask during class (to the whole group or to me directly) – anyone can reply!
- I would encourage you to install it on your laptop/phone - but make use of features to reduce notifications outside of working hours
- For important queries, feel free to email, but generally I am more responsive over chat than email.
- I generally answer chat messages within 2-3 hours; email within 24 hours during working days. For urgent questions, please tag me and I will prioritize.
- The fact that I may reply at odd times does **not** mean you should be working then!

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages* (over 20,000 and counting!)
- Powerful tool to conduct automated text analysis, social network analysis, and data visualization, with packages such as quanteda, igraph or ggplot2.
- Command-line interface and scripts favors reproducibility.
- Excellent documentation and online help resources.

R is also a full programming language; once you understand how to use it, you can learn other languages too.

RStudio


RStudio
File Edit Code View Project Workspace Plots Tools Help
Go to file/function
Project: (None)
diamondPricing.R*
formatPlot.R
diamonds
Source on Save
Source
library(ggplot2)
source("plots/formatPlot.R")
View(diamonds)
summary(diamonds)
summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)
p <- qplot(carat, price,
data=diamonds, color=clarity,
xlab="Carat", ylab="Price",
main="Diamond Pricing")
15:1
(Top Level)
R Script
Console ~/
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 950 2401 3933 5324 18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")
>
> format.plot(p, size=24)
>
Workspace History
Load Save Import Dataset Clear All
Data
diamonds 53940 obs. of 10 variables
Values
aveSize 0.7979
clarity character[8]
p ggplot[8]
Functions
format.plot(plot, size)
Files Plots Packages Help
Zoom Export Clear All
Diamond Pricing
Price
Carat
Clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

Computational Social Sciene: Opportunities and Challenges

Computational Social Science

"We have life in the network. We check our emails regularly, make mobile phone calls from almost any location ... make purchases with credit cards ... [and] maintain friendships through online social networks ... These transactions leave digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations and societies".

***Lazer** et al (2009) Science*

"Digital footprints collected from online communities and networks enable us to understand human behavior and social interactions in ways we could not do before".

***Golder and Macy** (2014) ARS*

Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

Big data and social science: challenges

1. Big data, big bias?
2. The end of theory?
3. Spam and bots
4. The privacy paradox
5. Generalizing from online to offline behavior
6. Ethical concerns

Computational **social** science

Challenge for social scientists: need for advanced technical training to collect, store, manipulate, and analyze massive quantities of semistructured data.

Discipline dominated by computer scientists who lack theoretical grounding necessary to know where to look.

Even if analysis of big data requires thoughtful measurement, careful research design, and creative deployment of statistical techniques (Grimmer, 2015).

New required skills for social scientists?

- Manipulating and storing large, unstructured datasets
- Webscraping and interacting with APIs
- Machine learning and topic modeling
- Social network analysis

For next week

1. Sign up for TWO peer reviews. Email with link will be sent tomorrow at 2pm.
2. Do readings for discussion: Vlasceanu and Amodio 2022; Hargittai 2018; and Kramer et al 2014 (and “Editorial Expression of Concern”).
   - Assuming they are selected to be discussed
3. New to CSS? Do background readings