School of Public Policy

PP434

Automated Data Visualisation for Policymaking

Course syllabus and readings

2023/24

(0.5 unit, AT)

Course teachers:

Richard Davies r.davies@lse.ac.uk

Automated Data Visualisation for Policymaking

Professor Richard Davies

Overview and motivation
This course explores ways of accessing large data sets to better understand the societies in which we live and ultimately to help guide policy decisions. The data we will encounter ranges from real-time measures of economic activity to micro data on local prices, to voting patterns and measures of pollution.

We will use methods from programming and economics to work on real-world problems. Students will learn the theory and policy history that lies behind data types, visualisation methods, data mapping and machine learning. With these tools in place, we will use APIs to access data programmatically, build scrapers and batch downloaders using Python. Cleaned and verified data are stored on GitHub, with students' work visualised using live and interactive web pages.

Topics include empirical strategy design, fetching and scraping data, data cleaning and storage, visualisation and interactivity. There is a focus on clear, replicable code that allows the automation of all these tasks in a policy setting. Students apply concepts of descriptive data analysis and may also use econometric techniques learned in parallel compulsory econometrics courses.

Learning outcomes
Students taking the course will understand the theory and history of data science as used by policymakers. In addition, they will be able to:

1. **Design**. To design an empirical strategy, linking a socioeconomic question to an on-line data sources so that trends can be tracked, and hypothesis tested.
2. **Automation**. To create a re-usable algorithm using code that will fetch (via an API) or scrape data from the web programmatically, building a novel database over time, and storing for analysis.
3. **Analysis.** To evaluate this data, in light of the empirical/policy question under investigation, interpreting descriptive statistics and simple statistical methods to test socio-economic questions.
4. **Visualisation**. To communicate the results of this analysis in a transparent, interactive and accessible manner.

Teaching details
The course will be taught using lectures, computer lab sessions and office hours. Each week consists of the following elements.

- **Preparation**. Asynchronous material. Students prepare for the coming week, watching videos, reading guides, and gathering data ideas.

- **Lecture and practical (2h).** A 1h lecture setting out the theory, history and principles of data analysis. After a short break, the skills and ideas we discussed are put to use in a 1h computer lab. Participants use their own laptops.

- **Project sessions (1h30, online).** A weekly compulsory slot for students to lean an additional skill, and then to troubleshoot problems, in a setting where course instructors will be on hand to advise. This session may also include guest speakers.

- **Office hours.** Bookable office hours with the teaching team.

Formative work (not graded**)**

Each week students have an opportunity to present workbooks and visualisations to their peers. This provides an opportunity for feedback from the teaching team, and an opportunity to iron out bugs and learn coding best practices. This work is not graded.

Grading Details (Summative Assessment)

The course is graded via the production of a professional-grade Data Science website. The website may consist of as many pages as students choose. The grades are given based on two pages: a portfolio, and a project. All graded work must be embedded in the website, hosted by GitHub pages. Students are given detailed lessons on how to do this. The deadline is 8th January 2025, at 4pm.

- **Portfolio (20%)**. This page demonstrates the tools that students learn in a practical setting, by using them to embed charts and diagrams of various types. There are 10 challenges, each of them demonstrating 1-2 skills and resulting in embedding 1-2 charts. The total score for this work is 20%, split evenly across each of these challenges.

  Portfolio skills include. *Building a web site. Embedding a live visualisation in a web site. Hosting data in the cloud. Editing and cleaning data. API-driven charts. Loops and APIs. Scrapers. Critical commentary on data. Advanced analytics. Interactivity.*

- **Project (80%)**. This page sets out the student's data science project. There are weekly on-line sessions in which students can discuss ideas with the teaching team. The project consists of between 5 and 8 charts, tables or visualisations. Students briefly discuss four topics: the aims, the data, analytical challenges, conclusions. Key marking criteria include: accessibility, empirical design, data approach, automation, interactivity, clarity of writing.

COURSE OVERVIEW

The foundation of the course is weekly lectures, practical sessions and project discussion sessions. These take participants from no coding experience at all to being able to design and maintain a live website that houses complex data visualisations.

Lectures and labs

Week	*Lecture*	*Lab*
1	**Introduction + building blocks** • Motivation. What is good Data Science? Where does it sit? Principles. • History and principles of 3 main languages: HTML, CSS and JS. • Data types (inc JSON).	**Building blocks** • Building your first interactive web site. • GitHub pages.
2	**Visualisation theory and principles I** • The good the bad and the ugly. Charts and misinformation. • History. From Playfair to today. • Good charting guide. What research tells us about visual communication.	**Visualisation 1**. • Making two live and interactive charts. • Introduction to charting libraries
3	**Visualisation theory and principles II** • The Semiology of Graphics (Bertin) • Makinley design criteria. • The Grammar of Graphics and cutting-edge chart packages	**Accessing Data 1**. APIs • Direct use in JavaScript. • Python access. Inc: how to export CSVs / JSON.
4	**Programming for Data Science** • Computational efficiency. • Control Flow principles. • Conditionality statements. Loops.	**Accessing Data 2.** Scrapers and loops. • Scraper examples: Google Finance, Sainsburys, FT, Olympics • Using loops with APIs and scrapers.
5	**Maps, data, and policy** • Cartography theory. Projections. • Choropleth maps: policy history. • Modern mapping.	**Making maps** • Geo-JSON data – base maps • Linking policy data to maps.
6. READING WEEK – NO CLASSES OR LECTURES – STUDENTS TO WORK ON PROJECTS		
7	**Advanced data: storage, reshaping** • Re-shaping • Transforms • Modern databases	**Analysis: cleaning and re-shaping** • Wrangling. • Practical examples of this.
8	**Advanced analysis** • Deeper numbers: using descriptive statistics to your advantage. • Establishing relationships. Correlation, causality, regression.	**Analysis 2**. • Generating and plotting descriptive statistics. • Plotting data to demonstrate causation
9	**Machine Learning I** • Machine learning. Theory and history • Supervised learning.	**ML 1** • Supervised • Unsupervised • Visualizing ML.
10	**Machine Learning II** • Unsupervised learning • Chat-GPT in policy settings	**Visualisation 2**. • Interactive visualisations. • Drop-downs, sliders.

READINGS

There is no single textbook that adequately covers this fast-moving area. In terms of classic texts, if there is one book to buy, I recommend Tufte's *Visual Display of Quantitative Information*.

Week	Reading
1	Mattmann, C.; A vision for data science. *Nature* 493, 473–475 (2013). https://doi.org/10.1038/493473a Ferguson, A.; A History of Computer Programming Languages, *Brown University*, 2000, https://cs.brown.edu/~adf/programming_languages.html -and an update of the original by *Hewlett-Packard*, 2018, https://www.hp.com/us-en/shop/tech-takes/computer-history-programming-languages
2	Friendly, M., Wainer, H., 2021; *A History of Data Visualization and Graphic Communication*, Chapter 5, pp. 95-120, Harvard University Press Tufte, E., 2007; *The Visual Display of Quantitative Information*, 2nd ed., Chapter 1 Graphical Excellence and Chapter 5: Chartjunk, Graphics Press LLC Davies, R.; *Heroes and Heroines*, [website]. URL https://www.playfairprize.com/william-playfair Playfair, W.; *The Commercial and Political Atlas*, 1785 Norman, J.; *History of Information* [website], 2021 https://historyofinformation.com/
3	Bertin, J.; *Semiologie Graphique*, 1967 Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. "A tour through the visualization zoo." Communications of the ACM 53.6 (2010): 59-67. Heer, Jeffrey, and Michael Bostock. "Declarative language design for interactive visualization." IEEE transactions on visualization and computer graphics 16.6 (2010): 1149-1156. Satyanarayan, A., Moritz, D., Wongsuphasawat, K., Heer, J., 2017; Vega-Lite: A Grammar of Interactive Graphics, IEEE Transactions on Visualization and Computer Graphics, 23(1), pp. 341-350, January 2017. URL https://doi.org/10.1109/TVCG.2016.2599030 Tufte, Edward, and P. Graves-Morris. "The visual display of quantitative information.; 1983." Diagrammatik-Reader. Grundlegende Texte aus Theorie und Geschichte. Berlin: De Gruyter (2014): 219-230. Mackinlay, Jock. "Automating the design of graphical presentations of relational information." Acm Transactions On Graphics (Tog) 5.2 (1986): 110-141. Stevens, Stanley Smith. "Adaptation-level vs. the relativity of judgment." The American Journal of Psychology 71.4 (1958): 633-646. Stevens, Stanley Smith. "Measurement, statistics, and the schemapiric view." Science 161.3844 (1968): 849-856.
4	Ederer, F., Goldsmith-Pinkham, P., Jensen, K., 2023; Anonymity and Identity Online, October 2023. URL https://florianederer.github.io/ejmr.pdf Ederer, F., Goldsmith-Pinkham, P., Jensen, K., 2023; Anonymity and Identity Online, NBER SI presentation, July 2023. URL https://www.youtube.com/watch?v=JwmZuxN0LDI

	Davis, D. R., Dingel, J. I., Monras, J., Morales, E., 2019; How Segregated Is Urban Consumption?, June 2019. URL http://www.jdingel.com/research/DavisDingelMonrasMorales.pdf Edelman, B., Luca, M., Svirsky, D., 2017; Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment, American Economic Journal: Applied Economics, 9(2), pp. 1-22. URL https://www.aeaweb.org/articles?id=10.1257/app.20160213
5	Lapon, L.; Ooms, K.; De Maeyer, P. The Influence of Map Projections on People's Global-Scale Cognitive Map: A Worldwide Study. *ISPRS Int. J. Geo-Inf.* 2020, *9*, 196. https://doi.org/10.3390/ijgi9040196 Krygier, John B. "Cartography as an art and a science?." The Cartographic Journal 32.1 (1995): 3-10. Cosgrove, Denis. "Maps, mapping, modernity: Art and cartography in the twentieth century." Imago Mundi 57.1 (2005): 35-54. Vaughan, Laura. "Charles Booth and the Mapping of Poverty." In *Mapping Society: The Spatial Dimensions of Social Cartography*, 61–92. UCL Press, 2018. https://doi.org/10.2307/j.ctv550dcj.8. Kraak, Menno-Jan, and Ferjan Ormeling. Cartography: visualization of geospatial data. CRC Press, 2020.
6. READING WEEK – NO CLASSES OR LECTURES – STUDENTS TO WORK ON PROJECTS	
7	Wickham, Hadley. "Tidy data." Journal of statistical software 59 (2014): 1-23. Mckinney, Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Al Sweigart. Automate The Boring Stuff. (Textbook, available free online here) Lohr, S., 2014; For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights, The New York Times. URL https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
8	Cunningham, S.; Causal Inference: The Mixtape, Introduction, Section 1.1 What Is Causal Inference, [book]. Available at https://mixtape.scunning.com/01-introduction
9	VanderPlas, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016. Andreas C. Müller, Sarah Guido. Introduction to Machine Learning with Python. O'Reilly. Athey, S., 2020; The Impact of Machine Learning on Economics, in Agrawal, A., Gans, J., Goldfarb, A. (eds), The Economics of Artificial Intelligence: An Agenda, Chicago, IL, 2019; online edn, Chicago Scholarship Online, 23 Jan. 2020. DOI: https://doi.org/10.7208/chicago/9780226613475.003.0021 Davis, J. M. V., Heller, S. B., 2017; Using Causal Forests to Predict Treatment Heterogeneity - An Application to Summer Jobs, American Economic Review. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20171000
10	Faria e Castro, M., Leibovici, F., 2024; Artificial Intelligence and Inflation Forecasts, Federal Reserve Bank of St. Louis Working Paper 2023-015. URL https://doi.org/10.20955/wp.2023.015 Kantayya, S., director, 2020; Coded Bias, [documentary], January 2020, available on Netflix