University of Essex

Department of Government

GV918-7-AU
Data for Social Data Science
2023-2024

Lecturer and Module Supervisor
Dr Akitaka Matsuo
Tel: 01206 873097
E-mail: a.matsuo@essex.ac.uk
Academic Support Hours:
Tuesday, 3-5 pm (Room TBA)

Module Administrator
govpgquery@essex.ac.uk

Module available for Study Abroad students: Yes ☐ No ☒

ASSESSMENT: This module is assessed by 100% coursework.

LISTEN AGAIN: ☒ **Listen again is available and should be fully utilised.**

INSTANT DEADLINE CHECKER - COURSEWORK

AU variant

Assignment Title	Due Date	Coursework Weighting	Feedback Due
Assignment 1	Week 5 (02-Nov)	30%	Week 8
Assignment 2	Week 8 (23-Nov)	35%	Week 11
Assignment 3	Week 16 (18-Jan)	35%	Week 19

ASSESSMENT

Each student is assessed by three coursework assignments, where we ask students to complete programming tasks in Python and to write up the interpretations of the outputs from the program. Students will get the assignment from the course github repository linked from course Moodle page and make the final submissions via FASER. All assignments are provided in the form of Jupyter Notebook. Students will conduct the computation using the notebook, and finished students will submit two files: One is the raw (executable) Jupyter notebook file with codes and write-ups about the findings from the computation; the other is an html file converted from the notebook. For the Jupyter notebook, students need to make sure that the notebook can provide the results without an error. If the notebook cannot be executed correctly, this will affect the mark. The converted html file has to be generated from a single step execution and should include all the write-ups. All assignments can be

completed in a computational environment on the cloud and students will be given an instruction for how to work in a cloud environment.

TOP READS

There are two textbooks for the module

1. VanderPlas Jake. Python *Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016 (available for download from the author's github page, https://jakevdp.github.io/PythonDataScienceHandbook/)
2. McKinney Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd Edition. O'Reilly Media; 2017.
3. James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert. *An Introduction to Statistical Learning: with Applications in R.* Springer; 2021 (available at https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf).

Other recommended textbooks are:

- Mitchell, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web, 2nd Edition.* O'Reilly; 2018.
- Teate, Renee M. *SQL for data scientists: a beginner's guide for building datasets for analysis,* John Wiley & Sons; 2021.

MODULE DESCRIPTION

This module introduces principles and applications of the electronic storage, structuring, manipulation, transformation, extraction, and dissemination of data. In the age of 'Big Data', the vast amount of data is generated in each day, and if equipped with a right set of skills, computational social scientists can obtain valuable insights only attainable through a data-driven approach. This module is aimed to provide an opportunity for learning such skills through programming in Python.

We focus on four key aspects of data management. The first is studying the various types of data, data shapes, and how to clean and transform them to fit for future data analysis. The next key component is the data acquisition. Most data nowadays are stored electronically on the Internet. We will learn what data are available online and how to obtain them through both scraping of websites and accessing APIs of online databases and social network services. The third key component of the module is to learn about the data storage solution, in particular about databases in both relational and non-relational forms. The module covers the fundamental concepts of database and how to create, populate, modify, and query relational databases. Lastly, this module uses a project-based learning approach, including group-based collaboration, essential ingredients of modern data science projects. We will learn various collaboration and management tools, such as the shared computational environment on the cloud and use of version control tools.

Advisory Note

Students are expected to have a basic understanding of statistical analysis with a successful completion of a module in introductory statistics. This includes experience in working with statistical computing software, such as R or Stata.

Students who do not have previous experience in statistical programming are strongly encouraged to consult with the module supervisor to be ready to take the course by acquiring necessary knowledge before the start of the term.

Objectives and transferable skills

This module aims to provide the following knowledge and comprehension on the basic of modern data science, through lectures and hands-on coding classes:

1. An overview of the lifecycle of the data in social data science, from data acquisition, pre-processing, storing to analysis
2. Knowledge of collaborative working space such as shared computing environments and version control systems
3. A general review of cloud computing
4. Basic principles of machine learning

By the end of the module, students will be:

1. Able to work with data sets using Python programming language and to summarise and visualise the data
2. Able to work with colleagues securely and effectively using online collaborative working space
3. Familiar to how to set up the cloud computing environment and able to know when to go on the cloud.
4. Capable of implementing online data collection projects for their research and managing/handling large data sets
5. Equipped with the understanding the fundamentals of machine learning, essential to the next steps of their data science learning.

MODULE STRUCTURE AND TEACHING

This module will be delivered over 4 hours per week – a 2-hour lecture and a 2-hour class.

What we expect of you during lecture and classes:

- To attend all lectures and classes after having done the required reading
- To pay attention and take notes as necessary.
- To think about the readings and lectures notes before the class and be ready to discuss them: try to identify the key assumptions in the texts; map the structure of the argument; underline the conclusions. Highlight to yourself points you don't understand. (If you don't understand it, there's great likelihood others have not understood it either, so don't be shy to ask.) Ask yourself whether you agree with the text, whether you can identify weaknesses or gaps in the argument, and what could someone who disagrees with it argue against it.
- To offer your participation as required (answering questions, asking questions etc.). Learning about and discussing these texts is a communal endeavour and it is a matter of good citizenship to contribute. Further, part of what we want you to achieve, and what we mark you for, is clear and confident oral presentation. You are expected to answer questions, raise new points, and

contribute to the progression of discussion in class.

How to submit your essay using FASER

You will be able to access the online submission system via your myEssex portal or via https://FASER.essex.ac.uk. FASER allows you to store your work-in-progress. This facility provides you with an ideal place to keep partially completed copies of your work and ensures that no work, even drafts, is lost. If you have problems uploading your coursework, you should contact *ltt@essex.ac.uk*. You may find it helpful to look at the FASER guide https://faser.essex.ac.uk/Student/Help. If you have any questions about FASER, please contact your administrator or refer to the handbook.

Under NO circumstances is your coursework to be emailed to the administrators or the lecturer. This will NOT be counted as a submission.

Coursework deadline policy for students

There is a single policy at the University of Essex for the late submission of coursework. Essays must be uploaded before 09.45 on the day of the deadline.

All coursework submitted after the deadline will receive a mark of zero. The mark of zero shall stand unless the student submits satisfactory evidence of extenuating circumstances that indicate that the student was unable to submit the work prior to the deadline. For further information on late submission of coursework and extenuating circumstances procedures please refer to http://www.essex.ac.uk/students/exams-and-coursework/ext-circ.aspx.

Essay feedback will be given via FASER.

ALL submissions should be provided with a coversheet (Available from Moodle).

Plagiarism

Plagiarism is a very serious academic offence and whether done wittingly or unwittingly it is your responsibility. **Ignorance is no excuse**! The result of plagiarism could mean receiving a mark of zero for the piece of coursework. In some cases, the rules of assessment are such that a mark of zero for a single piece of coursework could mean that you will fail your degree. If it is a very serious case, you could be required to withdraw from the University. It is important that you understand right from the start of your studies what good academic practice is and adhere to it throughout your studies.

All work submitted to the Department will be run through plagiarism detection software and lecturers are very good at spotting work that is not your own. **Plagiarism gets you nowhere; DON'T DO IT!**

Following the guidance on referencing correctly will help you avoid plagiarism.

Please familiarise yourself with the University's policy on academic offences: http://www.essex.ac.uk/about/governance/policies/academic-offences.aspx.

Extenuating circumstances for late submission of coursework

The university has guidelines on what is acceptable as extenuating circumstances for later submission of coursework. If you need to make a claim, you should upload your coursework to FASER and submit a late submission of coursework form which can be found here: http://www.essex.ac.uk/students/exams-and-coursework/late-submission.aspx. This must be done within seven days of the deadline. FASER closes for all deadlines after seven days. The Late Submissions committee will decide whether your work should be marked and you will be notified of the outcome.

If you experience significant longer-term extenuating circumstances that prevent you from submitting your work either by the deadline or within seven days of the deadline, you should submit an Extenuating Circumstances Form for the Board of Examiners to consider at the end of the year http://www.essex.ac.uk/students/exams-and-coursework/ext-circ.aspx.

SCHEDULE OF TOPICS AND READINGS

Syllabus

Week	Topic
Week 02	Data in social science
Week 03	Data manipulation
Week 04	Data visualisation
Week 05	Cloud computing
Week 06	Using the data from the Internet
Week 07	Working with APIs
Week 08	Working with databases I
Week 19	Working with databases II
Week 10	Introductory machine learning I
Week 11	Introductory machine learning II

List of courses in which the module is offered / to be offered, and module status - core, compulsory, optional)

MSc Social Data Science, Core

READING AND LAB SCHEDULE

Week 02: Data in Social Science

Required readings:

- Grimmer, Justin. We are all social scientists now: How big data, machine learning, and causal inference work together. PS - Polit Sci Polit. 2015;48(1):80-83.
- Counts S, De Choudhury, Munmun, Diesner Jana, et al. Computational social science. 2009;323(February):105-108. doi:10.1145/2556420.2556849
- VanderPlas, Jake. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, Inc.; 2016, Chapter 1.
- McKinney, Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2nd Edition. O'Reilly Media; 2017, Chapter 2, 3.
- Johari, Aayushi, *How To Use GitHub – Developers Collaboration Using GitHub* (https://www.edureka.co/blog/how-to-use-github/)

Further resource:

- https://www.linkedin.com/learning/python-essential-training-2018/welcome?u=51088249 (if you don't have prior experience in Python. Especially from Section 1 to Section 7).
- GitHub Guides (https://guides.github.com), especially: "Understanding the GitHub Flow", "Hello World", and "Getting Started with GitHub Pages".

Lab:

- Introduction to Google Colab, github
- Introduction to Jupyter Notebook
- Introduction to Python (Numpy and Pandas)
-

Week 03: Data Manipulation

Required readings:

- VanderPlas Jake. Python *Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016, Chapter 2-3

Recommended readings:

- McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd Edition. O'Reilly Media; 2017, Chapter 3-7.

Lab:

- Working with data using Pandas
    - Data manipulation
    - File input/ouput
    - Reshaping data

Week 04: Data Visualisation

Required readings:

- VanderPlas J. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016, Chapter 4
- Hughes, A. Visualizing inequality: How graphical emphasis shapes public opinion. *Research and Politics*. 2015.

Further readings:

- Tufte, E. *The visual display of quantitative information*. 2002. (https://www.edwardtufte.com/tufte/books_vdqi).
- McKinney W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd Edition. O'Reilly Media; 2017, Chapter 10

Lab:

- Data visualisation using Matplotlib and Seaborn

Week 05: Cloud Computing

Required readings:

- Rajaraman, V. 2014. "Cloud Computing." Resonance 19(3): 242–58. (https://www.ias.ac.in/article/fulltext/reso/019/03/0242-0258)
- AWS: *What is cloud computing*. (https://aws.amazon.com/what-is-cloud-computing/)
- Azure: *Developer guide*. (https://docs.microsoft.com/en-us/azure/guides/developer/azure-developer-guide)

Recommended readings:

- Puparelia, Nayan. 2016. "Cloud Computing." MIT Press. Ch. 1-3.
- Botta, Alessio, Walter De Donato, Valerio Persico, and Antonio Pescapé. 2016. "Integration of Cloud Computing and Internet of Things: A Survey." Future Generation Computer Systems 56: 684–700. (http://iranarze.ir/wp-content/uploads/2017/03/6229-English-IranArze.pdf)

Lab:

- Introduction to cloud computing using AWS

Week 06: Data from the Internet

Required readings:

- Mitchell, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web, 2nd Edition.* O'Reilly; 2018. Chapter 1-3, 5.

Further readings:

- Mitchell, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web, 2nd Edition.* O'Reilly; 2018. Chapter 17, 18.
- Silvey, Megan, Build Three Real-World Python Applications, Linkedin Learning (https://www.linkedin.com/learning/build-three-real-world-python-applications/scheduling-placeholder-movie), Chapter 1: Scraping Wisdom Pet Medicine Website.

Lab:

- Web-scraping from the UK government websites

Week 07: Working with APIs

Required readings:

- Mitchell Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web, 2nd Edition.* O'Reilly; 2018. Chapter 11, 12.
- Ruths, Derek, Pfeffer Jurgen. Social media for large studies of behavior. *Science*. 2014;346(6213):1063-1064.

Further readings:

- Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman. "Big data, digital media, and computational social science: Possibilities and perils." *The ANNALS of the American Academy of Political and Social Science* 659, no. 1 (2015): 6-13.
- Ali, Imran, Maria Balta, and Thanos Papadopoulos. "Social media platforms and social enterprise: Bibliometric analysis and systematic review." *International Journal of Information Management* 69 (2023): 102510.
- Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. "Fake news on Twitter during the 2016 US presidential election." *Science* 363, no. 6425 (2019): 374-378.
-

Lab:

- Working with APIs

Week 08: Working with Databases I

Required readings:

- Teate, Renee M. *SQL for data scientists: a beginner's guide for building datasets for analysis,* John Wiley & Sons; 2021. Chapter 2-4.

Further readings:

- Beaulieu A. *Learning SQL: Master SQL Fundamentals.* O'Reilly; 2009. Chapter 1-4.

Lab:

- Creating and managing SQLite databases
- Running simple queries

Week 09: Working with Databases II

Required readings:

- Teate, Renee M. *SQL for data scientists: a beginner's guide for building datasets for analysis,* John Wiley & Sons; 2021. Chapter 6-8.
- Boicea A, Radulescu F, Agapin LI. *MongoDB vs Oracle - Database comparison.* Proc - 3rd Int Conf Emerg Intell Data Web Technol EIDWT 2012. 2012:330-335. doi:10.1109/EIDWT.2012.32
- Parker Z, Poe S, Vrbsky S V. Comparing NoSQL MongoDB to an SQL DB. Proc Annu Southeast Conf. 2013. doi:10.1145/2498328.2500047

Further Materials:

- Beaulieu A. *Learning SQL: Master SQL Fundamentals.* O'Reilly; 2009. Chapter 5, 8, 9.
- MongoDB Basics (https://www.edx.org/course/mongodb-basics)
- Analyzing Big Data in less time with Google BigQuery (https://www.youtube.com/watch?v=qqbYrQGSibQ)

Lab:

- Advanced SQL syntax practice
- No SQL database demo

Week 10: Introductory Machine Learning I

Required reading:

- Molina, Mario, and Filiz Garip. "Machine learning for sociology." *Annual Review of Sociology*, 2019, 45(1), 27–45.
- Grimmer, Justin, Roberts, Margaret E., and Stewart, Brandon M. Machine Learning for Social Science: An Agnostic Approach. Annual Review of Political Science, 2021, 24(1), 395–419.
- VanderPlas J. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016, Chapter 5. (From *What is machine learning* to *Feature engineering*)
- James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.* Springer; 2013. Chapter 1, 2.1-2.2, 3.1-3.3, 6.1-6.2.

Further reading:

- VanderPlas J. Python Data Science Handbook: *Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016, Chapter 5. (From *In-depth: Naïve Bayes Classification* to *Principal Component Analysis*)

Lab:

- Introduction to Scikit-learn
- Regression problem

Week 11: Introductory Machine Learning II

Required reading:

- Cranmer, Skyler J., & Desmarais, Bruce A. What Can We Learn from Predictive Modeling?. *Political Analysis*, 2017, 25(2), 145-166.
- Grimmer, Justin, Roberts, Margaret E., and Stewart, Brandon M. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 2021, 24(1), 395–419.
- VanderPlas J. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc.; 2016, Chapter 5. (*Naïve Bayes Classification Support Vector Machine, Decision Trees and Random Forests*)
- James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.* Springer; 2013. Chapter 4.1-4.4, 5.1, 8.1-8.2.

Further reading:

- Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts et al. "Integrating explanation and prediction in computational social science." *Nature* 595, 2021, 7866: 181-188.

Lab:

- Working with classification problem using scikit-learn