ARCL0160: Archaeological Data Science

2023-2024

15 credits
Thursdays 11-2
Term I

Moodle Password: IoA2324
Turnitin: See Moodle site
Co-ordinator: Mark Altaweel
Co-ordinator's Email: m.altaweel@ucl.ac.uk
Room 103 Tel: 020 7679 74607 (Internal: 24607)

IMPORTANT INFORMATION REGARDING ASSESSMENTS:

The **coursework coversheet** is available on the course Moodle pages and here: https://www.ucl.ac.uk/archaeology/current-students under "Policies, Forms and Guidelines".

Please enter **your five-digit candidate code on the coversheet and *in the subject line*** when you upload your work in Moodle.

Please use **your five-digit candidate code as the name of the file** you submit.

Please refer to the IoA Student Handbook and IoA Study Skills Guide:
https://www.ucl.ac.uk/archaeology/current-students/ioa-student-handbook
https://www.ucl.ac.uk/archaeology/current-students/ioa-study-skills-guide
for instructions on coursework submission, IoA referencing guidelines and marking criteria, as well as UCL policies on penalties for late submission, over-length work and academic misconduct.

1. MODULE OVERVIEW

Module description

This course introduces principles of information and data science and its application to archaeology. The areas covered include basic principles of computer languages and scripts, database management and design, web science, document scraping, natural language processing, and artificial intelligence. All of these will also reinforce computational methodologies while also addressing new, developing ways in which archaeological problems are beginning to be investigated. The course is intended to provide a basic understanding in core principles and practical application on how emerging technologies and methods can be applied. The course is intended to give students skills that allow them to apply computational methods for their own research and professional interests. Lectures will cover how these techniques have been used in archaeology and other fields, including benefits and limitations. The course will also utilize hands on training and practicals to reinforce methods and techniques taught, with code provided. The use of open source and free tools will be emphasized as part of training.

Module Aims

The aim is to introduce students to key concepts in data/information science as it applies to archaeology. With increased use of technologies and increasing availability of data and data repositories, along with the fact data science is now transforming many fields, archaeologists need techniques and methods to understand how to promote their work in a modern format to other researchers and the public, while also utilizing information to make new and deeper understanding into archaeological problems. This includes being aware of current technical tools that are available as well as having abilities to customize tools to solve problems of interest.

<u>Objectives</u>

On successful completion of this course a student should:

- Understanding basic software and computer science applications for archaeology/heritage
- Recognise how to configure and create tools for research needs (rather than tools dictating what you can do)
- Be able to apply tools to archaeological/heritage problems
- Be familiar with different computing methods that can be created for various research problems

Learning Outcomes

On successful completion of the course students should be able to demonstrate/have developed:

- a basic understanding of fundamentals in data management, data science, scripts and software creation, machine learning techniques, web/document scraping, databases, and other computational skills

- be able to create a basic design and practical implementation of software and technical approaches to archaeological and/or heritage problems

The assignments provide assessment of the core skills taught. The research project allows students to think creatively and apply skills to a research problem specific to their interests or of relevance. Although the final project is left open for the student to decide what to apply it on, the course instructor is available to provide ideas and tasks that could applied to the final project.

Methods of Assessment

This course is assessed by two short assignments, (60% of the total mark), with the first assignment used as practice in code development and training. There is one research project (40%).

This course is assessed by means of:

(a) *The two* pieces of coursework that will be posted on Moodle are marked, with each being 500 words, which each contribute 3*0*% to the final grade for the course. These assignments will assess your ability to 1) understand the task given, 2) develop appropriate methods to solve the assigned problem, 3) provide solutions to common data science problems, 4) give good justification to your applied methods, and 5) determine how well your work is portable and works on different computers.

(b) a research project of 3000 words. The criteria is the same as the assignments; however, the main addition is that you must be able to demonstrate that your project can address a useful archaeological problem. You are encouraged to apply methods learned in class, but other approaches could be applied. The specific task you address is up to you but it should involved an archaeological/heritage related problem.

Note: all assignments are likely to include coding. Details will be discussed in class, as code should be submitted for evaluation.

Communications

- **Moodle is the main hub** for this course.
- Important information will be posted by staff in the **Announcements section of the Moodle page** and you will automatically receive an email notification for these.
- Please post any general queries relating to module content, assessments and administration **in the Moodle Forum for the course**. The forum will be checked regularly.
- For personal queries, please contact the co-ordinator by email.

Week-by-week summary

Week	**Date**	**Topic**	**Lecturers**
1	5/10/2023	Introduction Introduction to Programming	Mark Altaweel
2	12/10/2023	Using Python for Data Science	Mark Altaweel

3	19/10/2023	Programming Paradigms	Mark Altaweel
4	26/10/2023	Review Programming	Mark Altaweel
5	2/11/2023	Databases SQL	Mark Altaweel
6	READING WEEK		
7	16/11/2023	NoSQL	Mark Altaweel
8	23/11/2023	Web Science Web Scraping	Mark Altaweel
9	30/11/2023	PDF Scraping Natural Language Processing Topic Modelling	Mark Altaweel
10	7/12/2023	Methods in Artificial Intelligence Machine Learning	Mark Altaweel
11	14/12/2023	Computer Vision Deep Learning	Mark Altaweel

Lecturers (or other contributors)

All presentations are by Dr. Mark Altaweel

Weekly Module Plan

The course is taught through one hour lectures/seminars and one-two hours of hands on practical lessons during class. Each week will include presentations on concepts and presentation of examples that demonstrate concepts. In addition, students are expected to spend time outside of class to review and practice materials. Videos and examples of how to approach tasks given in class are given each week and these can be used to assist in understanding taught concepts.

Workload

Total work will constitute about 150 hours. This includes 10 hours for lecture, 20 for reading, 20 for additional work on practicals, and 90 hours for assignment preparation.

20 hours	*Staff-led teaching sessions (lectures, seminars, tutorials, discussion-board sessions)*
20 hours	*Self-guided session preparation and reading*
20 hours	*Practicing code and concepts learned through practicals*
90 hours	*Preparing assignments and work*

1. ASSESSMENT

Assessments will be further discussed in class. The first assessment will be marked but it will not count. For assignments 1-3, the marking will be based on your 1) coding, 2) ability to comment and explain what you did, 3) making what you did work, including providing the results that the code creates. These are used as the criteria to mark assignments 2-3. The final project will look at your code and code practice, but the focus will be on the essay and your ability to answer relevant archaeological/heritage questions with the analysis you apply.

Assignment 1: This will look at basic coding practices covered in class. The exact assignment is posted on Moodle. It will be marked but not count towards the final course mark. This will be used as an exercise and practice to prepare you for the marked work (assignments 2-3). It will not count but do use this to prepare for your other assignments.

Due date: 2/11/23

Assignment 2: This will look at database and data management techniques. The assignment will be posted on Moodle.

Due date: 23/11/23

Assignment 3: This assignment covers text and image scraping from websites along with the use of databases. The assignment will be posted on Moodle.

Due date: 21/12/23

Project: For the project, I want you to choose an archaeological/heritage topic and choose an appropriate set of methods covered in class. These methods should at least cover one or more topics, including: databases, web scraping, document scraping, natural language processing, machine learning/artificial intelligence, or related methods. Assessment will be based on your coding and ability to answer an archaeological/heritage question/research using appropriate citations, structure, and essay.

Due date: 19/1/24

If students are unclear about the nature of an assignment, they should discuss this with the Course Co-ordinator. Students are not permitted to re-write and re-submit essays in order to try to improve their marks. However, students may be permitted, in advance of the deadline for a given assignment, to submit for comment a brief outline of the assignment. The Course Co-ordinator is willing to discuss an outline of the student's approach to the assignment, provided this is planned suitably in advance of the submission date.

2. RESOURCES AND PREPARATION FOR CLASS

Preparation for class

1) Do readings in advance of class. Readings are intended to introduce you to topics and allow you to understand concepts. We will not cover readings in class but use the readings to inform you about the needed background.

2) Have a look at the week's practical and see if you can get through it. You will have a chance to work on it during class but it helps to look at the practical in advance.

3) Practice coding. The best way to learn is practice. You can do this using the Eclipse and Python plugin used in class on your own computers. Everything used in class is free and open source so there is no restriction with what is available.

3. SYLLABUS

Weekly Outline

Week 1

Introduction
Introduction to Programming

Mark Altaweel

Summary: Basic concepts of modern computing and computer programming are given, including variables, methods, loops, and logical structures. The idea will also be to introduce students to the basics of Python as a language utilizing the concepts introduced. These skills will be foundational to skillsets taught in this and other courses where programming is central.

Readings

Concepts of Computing

Essential:

Liang, Y. Daniel. 2013. *Introduction to Programming Using Python*. Boston: Pearson. Chapters 1.6-6.
See: https://haseebsohail.files.wordpress.com/2013/02/introduction-to-programming-using-python-y-liang-pearson-2013-ww.pdf

Suggested:

Nash, Fahiemah, Angela Du Preez, and Conrad De Wet. 2009. *Computer Programming*.

Tucker, Allen B., Teofilo F. Gonzalez, and Jorge L. Diaz-Herrera, eds. 2014. *Computing Handbook*. Third edition. Boca Raton: CRC Press, Taylor & Francis Group. Topic 1.

Python Resources

https://docs.python.org/3/tutorial/
http://www.pythonlearn.com/ (Chapters 2-6)
See: https://www.python.org/about/gettingstarted/
See: https://developers.google.com/edu/python/
See: https://www.codecademy.com/learn/python

See also the Moodle site for additional resources for help on Python. Additionally, resources for learning will be discussed in class. Videos are provided for assistance from week to week in assisting with coding and other tasks discussed in class and presented.

Week 2

Using Python for Data Science

Mark Altaweel

Summary: Several languages have become popular for data science, with Python a useful one because it is flexible and extensible. We will continue to go over basic concepts in Python and how it can be applied for some basic data manipulation and use of information from different sources.

Readings

Essential:

Vanderplas, Jacob T. 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Sebastopol, CA: O'Reilly Media, Inc.

Boschetti, Alberto, and Luca Massaron. 2015. *Python Data Science Essentials: Become an Efficient Data Science Practitioner by Thoroughly Understanding the Key Concepts of Python*. Community Experience Distilled. Birmingham: Packt Publ.

Suggested:

https://www.learnpython.org/

Week 3

Programming Paradigms

Summary: There are different approaches to use in creating solutions for data science problems. This includes different computing paradigms, including object-oriented computing or procedural programming. We present some of these paradigms and how they can at times be used together.

Readings

Essential:

See: The Role of Programming Paradigms in the First Programming Courses: http://elib.mi.sanu.ac.rs/files/journals/tm/21/tm1122.pdf

https://scotch.io/bar-talk/s-o-l-i-d-the-first-five-principles-of-object-oriented-design

Suggested:

Object Oriented Programming

Farrell, Joyce. 2008. *Programming Logic and Design: Comprehensive*. 7th ed. Boston, Mass: Thomson Course Technology. Chapters 10-11.

Üçoluk, Göktürk, and Sinan Kalkan. 2012. *Introduction to Programming Concepts with Case Studies in Python*. Wien ; New York: Springer. Pp. 21-28.

Phillips, Dusty. 2010. *Python 3 Object Oriented Programming: Harness the Power of Python 3 Objects*. Community Experience Distilled. Birmingham: Packt Publ. Chapters 1-5

Other

Hofstedt, Petra. 2011. *Multiparadigm Constraint Programming Languages*. Cognitive Technologies. Heidelberg ; New York: Springer. Parts 3-5.

Python

http://anandology.com/python-practice-book/object_oriented_programming.html
https://www.tutorialspoint.com/python/python_classes_objects.htm

Week 4

Review Programming

Mark Altaweel

Summary: This week we will review some of the concepts we have already covered in class. I will also introduces concepts in polyglot and multi-paradigm computing. The intent is to help create programmes with multiple languages (e.g., Python and R) or different paradigms (e.g., object-oriented, procedural programming, etc.). This will be the last week of programming training before we continue on to key data science concepts and paradigms.

Readings

Required:

Nowak, Robert M. 2014. "Polyglot Programming in Applications Used for Genetic Data Analysis." *BioMed Research International* 2014: 1–7. doi:10.1155/2014/253013.

Suggested:

Turnquist, Greg Lee. 2010. *Spring Python 1.1: Create Powerful and Versatile Spring Python Applications Using Pragmatic Libraries and Useful Abstractions*. Community Experience Distilled. Birmingham: Packt Publ. (Read Chapter 10 as an example, earlier chapters as needed).

Polyglot in Python with R

https://rpy2.github.io/doc/v3.0.x/html/overview.html

Polyglot in Python and java

http://www.jython.org/

Week 5

Databases
SQL

Mark Altaweel

Summary: This week introduces databases, their philosophies, with a focus on SQL-based relational structures.

Readings

Relational Databases

Garcia-Molina, H. et al. 2009. *Database Systems: The Complete Book.* http://infolab.stanford.edu/~ullman/fcdb/ch2.pdf (Chapter 2).

Kreibich, Jay A. 2010. *Using SQLite: Small. Fast. Reliable. Choose Any Three*. 1. ed. Beijing: O'Reilly.
Required:

Date, Chris J., and Hugh Darwen. 2007. *Databases, Types, and the Relational Model: The Third Manifesto*. 3. ed. Reading, Mass.: Addison-Wesley. Chapters 1-3, 5.

Data Mining

Cios, Krzysztof J., Witold Pedrycz, and Roman Świniarski. 1998. *Data Mining Methods for Knowledge Discovery*. The Kluwer International Series in Engineering and Computer Science, SECS 458. Boston: Kluwer Academic. Chapters 2-3.

Suggested:

Data Mining:

Cacciatore, Stefano ; Luchinat, Claudio ; Tenori, Leonardo. 2014. Knowledge discovery by accuracy maximization. *Proceedings of the National Academy of Sciences of the United States of America* 111(14): 5117-22.

Databases

https://www.mongodb.com/
https://www.mysql.com/
https://www.sqlite.org/

Week 6

NoSQL Databases

Mark Altaweel

Summary: This week continues from the previous week, with discussion and exercises on advanced data mining techniques utilized. The intent is to look at how statistical or machine learning methods can facilitate knowledge discovery. NoSQL databases are also covered.

Readings

NoSQL Databases

Dayley, Brad. 2015. *Sams Teach Yourself NoSQL with MongoDB in 24 Hours*. Indianapolis, Ind.: Sams.

Suggested:

Sadalage, Pramod J., and Martin Fowler. 2013. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Upper Saddle River, NJ: Addison-Wesley. Chapters 1-4, 7.

General

Hey, A. J. G. (Ed.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/M091000H.pdf. Pg. 111-116.

Data mining tools:

http://scikit-learn.org/stable/

Week 7

Introduction to Web Science
Web/Document Scraping

Mark Altaweel

Summary: This week we begin to look at web science today and how it impacts archaeology. The focus will be on web scraping methods for finding relevant documents and extracting key texts for further analysis from structured and unstructured data.

Introduction to Web Science

Required:

Ackland, Robert. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. Los Angeles: SAGE.

Web/Document Scraping

Mitchell, Ryan. 2015. *Web Scraping with Python: Collecting Data from the Modern Web*. First Edition. Sebastopol, CA: O'Reilly Media. Chapters 1-4, 6.

Web scraping tools:

http://scrapy.org/
https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Week 8

PDF Document Scraping and Natural Language Processing

Mark Altaweel

Summary: We continue with document scraping this week, where we now apply this to get information from unstructured pdf documents. The intent for this week is to allow you to potentially search and extract information from academic and other sources of information. We will introduce natural language processing (NLP) as an approach to understand unstructured data.

Required*:*

See reading from previous week.

Natural Language Processing:

Ghosh, Sohom, and Dwight Gunning. 2019. *Natural Language Processing Fundamentals*. http://proquest.safaribooksonline.com/?fpi=9781789954043.

Bonacchi, Chiara, Mark Altaweel, and Marta Krzyzanska. 2018. "The Heritage of Brexit: Roles of the Past in the Construction of Political Identities through Social Media." *Journal of Social Archaeology* 18 (2): 174–92. https://doi.org/10.1177/1469605318759713.

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers. 2016. "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism* 4 (1): 89–106. https://doi.org/10.1080/21670811.2015.1093271.

Suggested:

Altaweel, M. (2019). The Market for Heritage: Evidence from eBay using Natural Language Processing. Social Science Computer Review. 10.1177/0894439319871015.

Week 9
Methods in Artificial Intelligence
Artificial Intelligence

Mark Altaweel

Summary: This week we look more closely at artificial intelligence and machine learning methods more specifically. This includes using k-means clustering and related techniques for unsupervised and supervised classifications.

Required:

Altaweel M and Squitieri A. 2019. Finding a Relatively Flat Archaeological Site with Minimal Ceramics: A Case Study from Iraqi Kurdistan. *Journal of Field Archaeology* 44(8): 523–537. DOI: 10.1080/00934690.2019.1662269.

Conway, Drew, and John Myles White. 2012. *Machine Learning for Hackers*. 1st ed. Sebastopol, CA: O'Reilly Media. (All chapters are quite useful).

Wu, J. *2012. Advances in K-Means Clustering*. New York: Springer. Pg. 1-15.

Suggested:

Witten, I. H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. Morgan Kaufmann Series in Data Management Systems. Amsterdam ; Boston, MA: Morgan Kaufman. Pg. 131-142.

Altaweel, Mark, and Andrea Squitieri. 2020. "Quantifying Object Similarity: Applying Locality Sensitive Hashing for Comparing Material Culture." *Journal of Archaeological Science* 123 (November): 105257. https://doi.org/10.1016/j.jas.2020.105257.

Week 10

Computer Vision
Deep Learning

Mark Altaweel

Summary: This week we look at deep learning methods such as neural networks and layered techniques such as convultional neural networks (CNNs). These methods, more than any, attempt to mimic human learning and have been extensively used in so-called 'big data' problems. Additionally, we look at how computer vision in general can be used in archaeology.

Required:

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: The MIT Press. Chapter 6-9.

Gutherz G, Gordin S, Sáenz L, et al. (2023) Translating Akkadian to English with neural machine translation. *PNAS Nexus* Kearns M (ed.) 2(5): pgad096. DOI: 10.1093/pnasnexus/pgad096.

Winterbottom, T.; Leone, A.; Al Moubayed, N. A Deep Learning Approach to Fight Illicit Trafficking of Antiquities Using Artefact Instance Classification. *Sci Rep* 2022, *12*, 13468, doi:10.1038/s41598-022-15965-2.

Suggested:

UCL

Altaweel M, Khelifi A, Shana'ah MM. 2023. Monitoring Looting at Cultural Heritage Sites: Applying Deep Learning on Optical Unmanned Aerial Vehicles Data as a Solution. S*ocial Science Computer Review*. 08944393231188471. doi:10.1177/08944393231188471

Buduma, Nikhil, and Nicholas Locascio. 2017. *Fundamentals of Deep Learning: Designing next-Generation Machine Intelligence Algorithms*. First edition. Sebastopol, CA: O'Reilly Media.

Zhang, Liangpei, Lefei Zhang, and Bo Du. 2016. "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art." *IEEE Geoscience and Remote Sensing Magazine* 4 (2): 22–40. https://doi.org/10.1109/MGRS.2016.25407.