Data Analytics & Society

Understanding Data and their Environment

Course Information Handbook

2020

Contents

Contents

UNDERSTANDING DATA AND THEIR ENVIRONMENT 1

COURSE OUTLINE 3

Session 1: Introduction: Information about data 5

Session 2: Data Provenance ..... Error! Bookmark not defined.

Sessions 3, 4, and 5: Data Pre-processing 6

Sessions 6: Formative Data Analytics Exercise 6

Weeks 7-9: Anonymisation: Data Protection, Data Ethics and Statistical Disclosure 7

Session 10: Formative Data Analytics Exercise part II; Presentations and Introduction. 8

OTHER INFORMATION 8

Assessment 8

Support 9

READING LIST 10

Course Outline

This module aims to introduce you to a set of interrelated ideas about and approaches to understanding data and the environment in which those data sit. It is a combination of technical and non-technical topics all related to critical externalities to the data analytics process.

Day	Session times (GMT with daylight savings)	TOPIC	T&L	Lead
Monday	9:15-12:30	Information about data	Exercises (group) webinars (group and whole class),	Nuno & Mark
	13:15-16:30	Data Provenance	Exercises (group), presentaions (group), webinars (group and whole class)	Nuno & Mark
Tuesday	9:30-12:30	Data Pre-Processing	Live lectures, Exercises (individual), Q&A webinar (whole class)	Yu-wang
	13:30- 16:30			
Wednesday	9:30-12:30			
	13:30- 16:30	Formative Data Analytics Exercise 1	Exercises (group), Q&A webinars.	Mark & Yu-Wang
Thursday	9:30-12:30	Anonymisation: Data Protection, Data Ethics and Statistical Disclosure	Exercises (group) Q&A webinar (whole class)	Mark
	13:30- 16:30			
Friday	9:30-12:30			
	13:30- 16:30	Formative Data Analytics Exercise 2	Exercises (group), Group presentations. Q&A about asssessment (whole class)	Mark & Yu-Wang

The primary aim of the module is to demonstrate that data science cannot be carried out in a vacuum that a whole range of extrinsic considerations

affect our ability to carry out the research that we wish to carry out. However, appropriate management of these externalities can lead to higher quality as well more responsible research.

We aim to enable you to develop (i) a basic understanding of the technical processes of data provenance, anonymisation, disclosure control, data linkage, (ii) an awareness of the issues around the use of data for analysis (iii) fundamental skills in data husbandry.

On successful completion of this unit you should be able to:

- Understand the ethical issues surrounding the use of data in research.
- Understand the concepts and technical vocabulary of anonymisation and statistical disclosure.
- Pre-process data to optimise its analytical value.
- Demonstrate a basic understanding of data provenance.
- Make informed decisions about linkage/integration of data and carry out a basic data linkage.
- Conduct a basic anonymisation process with a dataset.
- Identify an appropriate collection of data sources for a project and to identify the issues in using those data sources.

The course is divided into ten sessions as detailed below.

What you need to do (General information)

Webinars/Live lectures

For the webinar/live lecture slots - unless it says differently in the session description below - we will be using blackboard collaborate ultra. To find the link for the particular session:

1. Go to the Understanding Data and Environment Blackboard site
2. Click on course content on the left hand menu
3. Click on the folder for the current session.
4. Click on the webinar links item
5. Click on the link for the particular webinar

Should Blackboard be down

Then go to one of the following links depending on who is (first) named tutor doe the session in the table on the previous page:

Nuno: https://zoom.us/j/436644021
Mark: https://global.gotomeeting.com/join/147813469

Yu-Wang: https://zoom.us/j/4673375977

Groupwork

In the group work sessions, you will generally need to produce a small piece of work collectively. You will need to agree amongst yourselves how you are going to do that. To state what may be obvious to you, each group will need:

1. An agreed communication channel to meet and discuss the task (Whatsapp, Skype, Zoom etc).
2. A mechanism for sharing/producing group outputs (at various points these include spreadsheets, text documents, diagrams and presentations).

Session by session

Sessions 1 & 2: Information about data

Tutor: *Nuno Pinto, Mark Elliot*
Summary: *These sessions are about how we conceptualise and understand data. Important topics metadata and paradata; data provenance and the data generating process; issues about data quality and the impact on inference; accessing and finding the data you need. You will do exercises using the data that you have created in advance of the module.*

What you need to do:

Session 1

In advance: Watch the session 1 podcast and exercise (metadata) brief
9:15 – 9:30: Attend webinar Q&A (Whole class)
9:30 – 10:45: Work on the metadata exercise (in groups)
@ 10:45 or 11:00 Attend 15 minute group webinars (in groups)

10:45 Leeds with Mark
https://global.gotomeeting.com/join/147813469

10:45 Manchester with Nuno
https://zoom.us/j/436644021

11:00 Liverpool with Mark
https://global.gotomeeting.com/join/147813469

11:00 Sheffield with Nuno
https://zoom.us/j/436644021

11:00-12:30 Work to complete the metadata exercise (in groups)
@12:30 post solution to Dropbox

Session 2
In advance: Watch the session 2 podcast and exercise (indicator) and exercise (provenance) briefs
1:15 – 1:30: Attend webinar Q&A (Whole class)
1:30 – 2:30: Work on metadata exercise 2 (in groups)
2:30 – 3:10: Group presentations 1-2 slides (whole class)
3:10 – 4:10: Work on provenance exercise (in groups)
4:10 – 4:30: Closing Q&A (whole class)

After session, upload your final provenance exercise group to blackboard by 9AM on Tuesday. (Mark will provide feedback).

Sessions 3-5: Data Pre-processing

Tutor: *Yu-Wang Chen*
Summary: *This three session block covers the very important area of data pre-processing, an umbrella term covering a range data processing activities including description, summarization, visualization, cleaning, reduction, transformation, discretization, generalization and integration.*

What you need to do

Session 3
9:30-11:00 Join live teaching session on Blackboard Collaborate Ultra.
11:00-12:10 Work on a data prep task 1 (individual).
12:10-12:30 Attend a Q&A on the lecture and the task.

Session 4
13:30-15:00 Join live teaching session on Blackboard Collaborate Ultra.
15:00-16:10 Work on a data prep task 2 (individual).
16:10-16:30 Attend a Q&A on the lecture and the task.

Session 5
9:30-11:00 Join live teaching session on Blackboard Collaborate Ultra.
11:00-12:10 Work on a data prep task 3 (individual).
12:10-12:30 Attend a Q&A on the lecture and the task.

Session 6: Formative Data Analytics Exercise part I

Tutor: *Yu-Wang Chen, Mark Elliot*
Summary: *In this exercise, each group will be given some datasets and asked to produce a statistical analysis. This will provide some practice at the exercises required for the assessment.*

What you need to do

13:30 – 13:45 Attend webinar 1 for description of task and any immediate questions.
13:45 – 16:00 Work on the task (in groups).
16:00 – 16:30 Attend the webinar 2 for a Q&A on the task.

Sessions 7-9: Anonymisation: Data Protection, Data Ethics and Statistical Disclosure

Tutor: *Mark Elliot*
Summary: *These three sessions will cover the concepts of data protection, data ethics, and disclosure control. Each of these topics is huge in itself and each could easily consume a whole week's course. So here, you will be just getting a taster. In order to ensure that this is coherent the concepts will be presented and discussed through the overall lens of anonymisation.*

Anonymisation is a complex topic. We will endeavour to demystify it through the particular device of the Anonymisation Decision Making Framework, which ties together the statistical notion of risk with the legal notions of personal data and proportionality. We will, in particular demonstrate that data anonymisation should be considered as a ***process*** *that operates on the relationship between data and their environment rather than a state of data. Anonymisation is a data science problem in itself and one demonstrates aptly the need for interdisciplinary thinking.*

What you need to do

Session 7

In advance: Watch the session 7 podcast
9:30 – 10:00 webinar for a Q&A on the lecture
10:00 – 12:00 complete the data flow task (group)
12:00 – 12:30 attend the webinar for a Q&A on the task

Session 8

In advance: Watch the session 8 podcast

13:30 – 14:00 webinar for a Q&A on the lecture
14:00 – 16:00 Complete the data, legal and ethical sections of the ADF template (group)
16:00 – 16:30 Attend the webinar for a Q&A on the lecture and the task

Session 9

In advance: Watch the session 9 podcast
9:30 – 10:45 Join live teaching session on Blackboard Collaborate Ultra.
10:45 – 12:20 disclosure control sections of the template task (group)
12:20 – 12:45 attend the webinar for a Q&A on the task

Session 10: Formative Data Analytics Exercise part II; Presentations and Introduction.

Tutor: *Yu-Wang Chen, Mark Elliot*

Summary *In this session you will deliver a presentation on your groups solution for the formative exercise and we will talk through the assessment.*

What you need to do

13:30 – 15:30 Complete the task formative task and prepare a presentation (group)
15:30 – 16:30 Attend the webinar to present your solutions, receive feedback followed by a Q&A about the assessment.

Other Information

Assessment

Assessment will be in the form of an assessed essay of 2000 words and a project report of no more than 2500 words.

As part of the project report process, you will proceed at first in groups to develop a data pre-processing and analysis plan before writing individual reports. There will be a formative presentation session on the 29th April or 1st May.

You will be given more details about these in the relevant lecture.

Form	Length	Weighting	Deadline
Report	2500 words	60%	15th June
Essay	2000 words	40%	15th June

Support

The tutors are available during the office hours these are shown below:

Yu-Wang Chen	
Email	Yu-wang.Chen@manchester.ac.uk
Office hours	Monday 13:00 - 15:00
Link for office hours	TBC

Mark Elliot	
Email	Mark.elliot@manchester.ac.uk
Office hours	Wednesdays 9:00-10:00 Fridays 9:00-10:00
Link for office hours	https://global.gotomeeting.com/join/147813469

Nuno Pinto	
Email	nuno.pinto@manchester.ac.uk
Office hours	Thursdays 14.00-16.00
Link for office hours	https://zoom.us/j/575402348

Reading List

ARRINGTON, M. (2006) *AOL proudly releases massive amounts of user search data,* TechCrunch, available at: http://tinyurl.com/AOL-SEARCH-BREACH [accessed 30/5/2016].

ATOKAR (2014) *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, available at: http://tinyurl.com/NYC-TAXI-BREACH [accessed 30/5/2016].

BANNER, N., BURTON, P., ELLIOT, M. J. KNOPPERS B. M. and BANKS, J. (2017) 'Policies and strategies to facilitate secondary use of research data in the health sciences', *International Journal of Epidemiology* 46(6), 1729-1733

CHRISTEN, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media. http://users.cecs.anu.edu.au/~Peter.Christen/data-matching-book-2012.html

CNN MONEY (2010) *5 data breaches: From embarrassing to deadly*, available at: http://tinyurl.com/CNN-BREACHES/ [accessed: 30/5/2016].

DIBBEN, C., GOWANS, H. ELLIOT, M. J. AND LIGHTFOOT, D. (2015). 'The Data Linkage Environment' in *Methodological Developments in Data Linkage*; Harron, K. Goldstien H. and Dibben, C. (eds) 36-62 Wiley.

DUNCAN, G. T., ELLIOT, M., & SALAZAR-GONZÁLEZ, J. J. (2011). *Statistical Confidentiality.* Springer New York.

ELLIOT, M. J., MACKEY, E., O'HARA, K., & TUDOR, C. (2016). *The Anonymisation Decision-Making Framework*. UKAN publications; Manchester.

ELLIOT, M. J., RAAB, C., O'HARA, K., DIBBEN, C., GOWANS, H., MACKEY, E, O'KEEFE, C. PURDAM K. MCCULLAGH, K. (2018) 'Functional anonymisation: Personal data and the data environment', *Computer Law & Security Review* 34(2) April 2018 204-221

FAMILI, A., SHEN, W. M., WEBER, R., & SIMOUDIS, E. (1997). 'Data preprocessing and intelligent data analysis'. *Intelligent data analysis*, 1(1-4), 3-23.

GARCÍA S., LUENGO, J., HERRERA F. (2015). *Data preprocessing in data mining.* Springer

HAN, J., PEI, J. AND KAMBER, M. (2011). *Data mining: concepts and techniques.* Elsevier.

HAYNES, C. L., COOK, G. A., & JONES, M. A. (2007) Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease register; *Journal of Medical Ethics,* 33(5): 302–307, DOI:10.1136/jme.2006.016907.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. & DE WOLF, P. P. (2012) *Statistical Disclosure Control.* London: John Wiley & Sons.

LANE, J., STODDEN, V., BENDER, S. & NISSENBAUM, H. (Eds.) (2014) *Privacy, Big Data, and the Public Good.* Cambridge: Cambridge University Press.

MACKEY, E., & ELLIOT, M. (2013). Understanding the data environment. *XRDS: Crossroads* 20(1), 36-39.

MOREAU, L., & GROTH, P. (2013) *Provenance: An Introduction to PROV.* Available at https://tinyurl.com/PROV-BOOK [accessed 25/9/2019]

NISSENBAUM, H. (2004) Privacy as contextual integrity; *Washington Law Review*, 79 (119): 101-139, available at: http://tinyurl.com/j8xut58 [accessed 30/5/2016].

PURDAM K. AND ELLIOT, M. J. (2015). 'The Changing Social Data Landscape' in Halfpenny, P. and Procter, R. (eds.) *Innovation in Digital Research Methods.* 25-58 London: Sage.

RUBINSTEIN, I. S. (2013) Big data: the end of privacy or a new beginning?; *International Data Privacy Law,* 3(2): 74-87. Available at: http://tinyurl.com/q3wvd53 [accessed 28/5/16].

RUNKLER, T. A. (2012). *Data Analytics: Models and Algorithms for Intelligent Data Analysis,* Springer.

SMITH, D. & ELLIOT, M. J. (2014) 'A Graph-based Approach to Key Variable Mapping' Confidentiality and Privacy. 6(2), 81-115.

UK: INFORMATION COMMISSIONER'S OFFICE (2011) *Data sharing code of practice,* available at http://tinyurl.com/ICO-SHARE [accessed 25/5/2016].

WELLCOME TRUST (2016) *Public attitudes to commercial access to health data,* available at http://tinyurl.com/zl2es5b [accessed 19/6/2016]

UK: INFORMATION COMMISSIONER'S OFFICE (2012a) *Anonymisation:*

managing data protection risk code of practice, available at http://tinyurl.com/ICO-ANON [accessed 25/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2014a) *Data controllers and data processors: what the difference is and what the governance implications are,* available at http://tinyurl.com/ICO-CONT-PROC [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2014b) *Conducting privacy impact assessments code of practice,* available at http://tinyurl.com/ICO-PIA2 [accessed 30/5/2016].

UK: INFORMATION COMMISSIONER'S OFFICE (2016) *Guide to Data Protection version 2.4,* available at http://tinyurl.com/ICO-DPG2-4 [accessed 30/5/2019].