IM939-15 Data Science Across Disciplines: Principles, Practice and Critique

23/24

Department
Centre for Interdisciplinary Methodologies
Level
Taught Postgraduate Level
Module leader
Cagatay Turkay
Credit value
15
Module duration
10 weeks
Assessment
100% coursework
Study location
University of Warwick main campus, Coventry

Description

Introductory description

This module introduces students to the fundamental techniques, concepts and contemporary discussions across the broad field of data science. With data and data related artefacts becoming ubiquitous in all aspects of social life, data science gains access to new sources of data, is taken up across an expanding range of research fields and disciplines, and increasingly engages with societal challenges. The module provides an advanced introduction to the theoretical and scientific frameworks of data science, and to the fundamental techniques for working with data using appropriate procedures, algorithms and visualisation. Students learn how to critically approach data and data-driven artefacts, and engage with and critically reflect on contemporary discussions around the practice of data science, its compatibility with different analytics frameworks and disciplinary, and its relation to on-going digital transformations of society. As well as lectures discussing the theoretical, scientific and ethical frameworks of data science, the module features coding labs and workshops that expose students to the practice of working effectively with data, algorithms, and analytical techniques, as well as providing a platform for reflective and critical discussions on data science practices, resulting data artefacts and how they can be interpreted, actioned and influence society.

Module web page

Module aims

In this module, students gain both formal knowledge and practical experience of the theoretical, scientific and ethical frameworks underpinning data science and critically reflect on the scope and impact of these frameworks. Lectures will provide a grounded understanding of the theoretical and scientific frameworks underpinning data science. In workshops, students gain experience of the fundamentals of the practice of data science, and through seminars they will be exposed to academic debates in data studies and related fields about the changing role of data science in society as seen in, for instance, the increasing use of data artefacts in policy and decision making in governmental bodies and businesses, how scientific discoveries are made and communicated, or how (in)equalities and power (im)balances are surfacing in uses of data. The module aims to build the required skills to apply data science techniques and algorithms within and across analytics frameworks developed in different disciplines. The module aims to cultivate a holistic data science practice which reviews the whole data science process critically and inquisitively, and handles problems through a user-centred thinking. This practice also embraces critical reflection about the data, algorithms, and data artefacts, as well as the ethical, societal, and cultural implications of data science broadly conceived.

Outline syllabus

This is an indicative module outline only to give an indication of the sort of topics that may be covered. Actual sessions held may differ.

Session-01: INTRODUCTION, HISTORICAL PERSPECTIVES & BASIC CONCEPTS
This week discusses data science as a field that cuts across disciplines and provides a historical perspective on the subject. We discuss the terms Data Science and Data Scientists, reflect on examples of Data Science projects, and discuss the research process at a methodological level. We will also use the examples as probes to think broadly on the potential influence of data-intensive scientific approaches on knowledge, industry and the wider society.

The practical lab session help students get acquainted with the analytical platform that will be used throughout the term and provides a first experience working with data sets within a data science approach.

Session-02: THINKING DATA: THEORETICAL AND PRACTICAL CONCERNS
This week explores the cultural, ethical, and critical challenges posed by data artefacts and data-intensive scientific processes. Engaging with Critical Data Studies, we discuss issues around data capture, curation, data quality, inclusion/exclusion and representativeness. The session also discusses the different kinds of data that one can encounter across disciplines, the underlying characteristics of data and how we can analytically and practically approach data quality issues and the challenge of identifying and curating appropriate data sets.

The practical lab session walks students through the earlier stages of the data science process. We start by looking at different types of data suitable for analysis within a data science framework and move on to how to wrangle the data to make it available for further use.

Session-03: ABSTRACTIONS & MODELS

This week discusses ways of abstracting data. We start by visiting statistics as a means of representing data and its inherent characteristics. The session moves on to discuss the notion of a “model” and visit the different schools of thought within model-ing, as well as a tour of fundamental statistical models that help abstract data and its inherent relations.

The practical part explores processing data and data transformations, summarizing data through descriptive statistics, the case of outliers and a brief overview of robust statistics, as well as investigating relations within different aspects of the data and explore concepts such as correlation, regression, and their relevance within different disciplinary frameworks.

Session-04: STRUCTURES AND SPACES
This week explores the notion of structures and how data science can enable the extraction of “hidden” underlying groups – clusters -- and hierarchical structures from data. We discuss the different techniques to surface and generate artificial boundaries and how the resulting artefacts can be interpreted. This session then investigates how artificial and abstract spaces can be constructed through different “projection” techniques, and how these spaces help us navigate data that are high-dimensional in nature and apply analytic frameworks to them.

The practical lab explores the use of clustering techniques, compares alternatives, and discusses interpretability issues, and we also review how we can deal with data sets that consists of several variables.

Session-05: MULTI-MODEL THINKING AND RIGOUR IN DATA SCIENCE
This week we focus on multi-model approaches as a way of thinking and how critical, pluralistic thinking can improve our understanding of the underlying phenomena implicit in data. We also discuss how to adopt a comprehensive approach to the data science process, and investigate indicators of rigor in data science.

The practical session involves combining perspectives derived from different computational models, as well as considering how diverse theoretical frameworks can help us approach phenomena of interest in different ways.

Session-06: RECOGNISING AND AVOIDING TRAPS
This week we discuss how we can be aware of various methodological and ethical traps and pitfalls that one can encounter during the data science process. We will discuss causality and when and to what extent it can be expected and observed, we will touch upon statistical traps such as Base-rate fallacy / prosecutor's fallacy, Simpson's paradox, as well as how visualisations can deceive and how to avoid representational traps.

The practical session involves hands-on examples where such traps are encountered and responded.

Session-07: DESIGN THINKING IN DATA SCIENCE
This week explores the question “Can we approach data science as a design problem?” and discusses how one can embrace a user-centred approach to design appropriate data science processes. We will explore how design thinking can inform and influence how we approach underlying problems, understand the interests and investments of different disciplines and expert stakeholders and how this can inform the design of analytic frameworks and disciplines accordingly.

The practical session involves the hands-on exploration of the topic through an applied example

where students respond to a design brief through some of the design techniques discussed.

Session-08: DATA SCIENCE & SOCIETY
We will engage with academic and practices discourse on the social, cultural and ethical aspects of data science, and discuss around how one can responsibly carry out data science research on social phenomena, whether data science can be a transformative power in society, and what ethical and social frameworks can help us to critically approach data science practices and its effects on society, and what are ethical practices for data scientists.

Session-09: DATA SCIENCE WORKSHOP - 1
This week will involve hands-on practical where we go through the data science process over applied examples. During the workshop, we will explore concepts such as reproducibility, openness and transparency and how best to communicate the analytical and reasoning process as well as critically reflecting on the various data artefacts produced.

Session-10: DATA SCIENCE WORKSHOP - 2
This week will involve hands-on practical workshops where we go through the data science process using applied examples. During the workshop, we will also explore concepts such as narrative and visual storytelling, as well as reflect on the design process for our analysis and artefacts.

Learning outcomes

By the end of the module, students should be able to:

- Demonstrate an understanding of the workings and the practicalities of the data science process
- Demonstrate an in-depth understanding of the theoretical underpinnings, scientific and ethical frameworks of data science as applied across disciplines
- Demonstrate a critical understanding of the role that data and data intensive practices play in research, industry and the wider society
- Apply and evaluate data science techniques and tools for particular scenarios and argue their suitability
- Demonstrate an ability to critique any resulting data artefacts, such as data-informed decisions to data-driven models, including from a user-centred perspective
- Develop and demonstrate an understanding of the societal, ethical, and cultural implications of advances in and applications of data science

Indicative reading list

- **Data science as a scientific practice** : Dhar, Vasant. "Data science and prediction." Communications of the ACM, 56.12 (2013): 64-73.
- **On Google's influenza epidemic application**: Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." Nature (2008)
- Interviews with analysts on challenges in data analysis: Kandel, Sean, et al. "Enterprise data analysis and visualization: An interview study." *Visualization and Computer Graphics, IEEE Transactions on* 18.12 (2012): 2917-2926.
- On data wrangling: Kandel, Sean, et al. "Research directions in data wrangling:

Visualizations and transformations for usable and credible data." *Information Visualization* 10.4 (2011): 271-288.

- A good resource on data transformations and issues involving different types: Osborne, Jason. "Notes on the use of data transformations." *Practical Assessment, Research & Evaluation* 8.6 (2002): 1-8.
- Discussions on outliers in data analysis: Osborne, Jason W., and Amy Overbay. "The power of outliers (and why researchers should always check for them)." *Practical assessment, research & evaluation* 9.6 (2004): 1-12.
- Very good overview **on variable and feature selection** (have a look at the first 4 sections and can read the rest for more advanced methods): Guyon, Isabelle, and André Elisseeff. " An introduction to variable and feature selection." The Journal of Machine Learning Research 3 (2003): 1157-1182.
- On using PCA in biotechnology, gives an overview introduction: Ringnér, Markus (2008). " What is principal component analysis?". *Nature biotechnology (1087-0156)*, 26 (3), p. 303.
- A non-technical introduction to MDS : Jaworska, Natalia, and Angelina Chupetlovska-Anastasova. "A review of multidimensional scaling (MDS) and its utility in various psychological domains." *Tutorials in Quantitative Methods for Psychology* 5.1 (2009): 1-10.
- A very important read on Cross-validation methods: R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Intl. Jnt. Conf. AI
- A very good article on centrality on graphs: White, Douglas R., and Stephen P. Borgatti. " Betweenness centrality measures for directed graphs." *Social Networks* 16.4 (1994): 335-346.
- A very good article on interactive visualisation types: Heer, Jeffrey, and Ben Shneiderman. " Interactive dynamics for visual analysis." *Queue* 10.2 (2012): 30.
- Iliadis, A. and Russo, F., 2016. Critical data studies: An introduction. Big Data & Society, 3(2), p.2053951716674238.
- Ruckenstein, M. and Schüll, N.D., 2017. The datafication of health. Annual Review of Anthropology, 46, pp.261-278.
- Pink, S., Ruckenstein, M., Willim, R. and Duque, M., 2018. Broken data: Conceptualising data in an emerging world. Big Data & Society, 5(1), p.2053951717753228.

Research element

Students will need to conduct literature research into a subject area where they are approaching within a data science process. They will need to theoretically and empirically compare and evaluate multiple approaches, document the reasoning and report in a systematic and reflective manner. Also, students will need to engage with the theoretical and critical frameworks, evaluate and reflect on the results, process, and the implications in dialogue with these frameworks.

Interdisciplinary

Students will need to approach the subject of Data Science through an interdisciplinary lens in dialogue with theoretical, technical and critical advances primarily within the fields of Computer Science, Statistics and Social Sciences, as well as scholarly and professional literature that covers social, cultural, political and ethical aspects of data science. Students will further engage in interdisciplinary thinking through user-centred approaches to help them to systematically approach

the design of data-intensive approaches. In practical, hands-on sessions, students will not only be expected to engage with the technical methods in data science, but will also be asked to draw on critical thinking perspectives primarily from Data Studies and already interdisciplinary Data Science literature.

Subject specific skills

Data scientist skills such as:

- data wrangling and data processing
- data modelling and model interpretation
- data visualisation
- analytical programming
- interpreting computational artefacts

Transferable skills

- Critical thinking on the use of data and data-driven artefacts
- Analytical problem-solving skills
- Data-informed insight generation
- Presentation and communication skills

Study

Study time

Type	Required
Lectures	10 sessions of 1 hour (7%)
Supervised practical classes	10 sessions of 2 hours (13%)
Online learning (independent)	120 sessions of 1 hour (80%)
Total	150 hours

Private study description

No private study requirements defined for this module.

Costs

No further costs have been identified for this module.

Assessment

You do not need to pass all assessment components to pass the module.

Assessment group A

	Weighting	Study time
Critical Review	40%	
Critical Review of a Data Science project		
Final Project Analysis and Critique	60%	
A data-driven essay: An essay reporting on a data science project where students report on the data science process from initiation to evaluation to reflection while engaging with the relevant literature in the domain		

Feedback on assessment

Written feedback for all the assessments (report, poster & essay) as well as formative oral feedback for the poster.

Availability

Courses

This module is Optional for:

- Year 2 of TIMS-L990 Postgraduate Big Data and Digital Futures
- TIMA-L995 Postgraduate Taught Data Visualisation
  - Year 1 of L995 Data Visualisation
  - Year 2 of L995 Data Visualisation
- TIMA-L99A Postgraduate Taught Digital Media and Culture
  - Year 1 of L99A Digital Media and Culture
  - Year 2 of L99A Digital Media and Culture
- Year 1 of TIMA-L99D Postgraduate Taught Urban Analytics and Visualisation

This module is Core option list A for:

- Year 1 of TPSS-C803 Postgraduate Taught Behavioural and Data Science

This module is Core option list C for:

- Year 1 of TPSS-C803 Postgraduate Taught Behavioural and Data Science

This module is Option list A for:

- Year 1 of TIMS-L990 Postgraduate Big Data and Digital Futures