THE GRADUATE INSTITUTE GENEVA

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Interdisciplinary Programmes

Academic year 2019-2020

Big Data Analysis

MINT078 - Spring - 3 ECTS
24 April & 8 May 2020

Course Description

This block course provides a basic introduction to big data and corresponding quantitative research methods. The objective of the course is to familiarize students with big data analysis as a tool for addressing substantive problems. The course begins with a basic introduction to big data and discusses what the analysis of these data entails, as well as associated technical, conceptual and ethical challenges. Strength and limitations of big data research are discussed in depth using real-world examples. Students then engage in case study exercises in which small groups of students develop and present a big data concept for a specific real-world case. These exercises are designed to familiarize students with the format of big data and to gain a first, hands-on experience with potential applications for large, complex data in policy-relevant settings. The block course is designed as a primer for anyone interested in attaining a basic understanding of what big data analysis entails and does not entail technical training for scripting etc. There are no prerequisite requirements for this course.

Note that the Big Data Analysis course is offered in two separate sessions: April 24 (course MINT078-3) and May 8 (course MINT078-4). Please ensure that you are attending the session you are registered for!

IMPORTANT: This course has been restructured for an online-only teaching setting. Please see the revised course schedule below!

PROFESSOR

Karsten Donnay
karsten.donnay@graduateinstitute.ch
www.karstendonnay.net

Syllabus

Course Requirements

Requirement 1: (Virtual) Attendance in all required parts of the workshop is mandatory and students are expected to actively engage with the recommended readings and/or online resources in preparation for the course.

Requirement 2: Students will be required to complete case study exercises in small groups. Evaluation will be based on the written project report of each group. You are expected to actively work with the members of your group through suitable online collaboration channels (Google Docs, Google Hangout etc.).

Course Evaluation

Performance in the course depends both on active participation in the required (online) sessions of the course and performance in the case study exercises. Evaluation will be based on:

1. Active participation and contribution to the course 20%
2. Performance in case study exercises 80%

Course Material

The following are recommended for anyone interested in background readings on big data written for scientific and general audiences. Recommended scientific readings and/or online resources for individual sessions are provided with stable links in the course schedule below.

- Matthew J. Salganik. (2017). Bit by Bit: Social Research in the Digital Age. Princeton University Press.
- Cathy O'Neil. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Penguin Books.
- Rob Kitchin. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE Publications.

The first book is written for social scientists interested in conducting big data analysis and a useful guide for everybody interested in data science. The second book focuses primarily on possible downsides of algorithms and big data analysis in various domains. And the third book both provides an overview of big data, open data and data infrastructures and associated concepts as well as a discussion of potential shortcoming and (unintended) consequences of this paradigm shift for science and society.

Overview of the Course

The first part focuses on providing a theoretical and practical introduction to big data, its analysis and associated challenges. This part entails an interactive session through the video conferencing tool *Google Meet* as well as four pre-recorded in-depth sessions. In the second part, students then apply this knowledge in the context of a case study and prepare a written case study report. The course provides a conceptual overview of technical aspects of big data analysis but students are not required to complete practical programming exercises and no prior knowledge of scripting etc. is assumed.

Course Website

Please refer to the course website on Moodle for the most up-to-date information on the class. The lecture slides, pre-recorded lectures, case study materials etc. will all be made available through the website. We will also use its forum for course-related communication. Please use the link below or search for "MINT078-3" on Moodle:

https://moodle.graduateinstitute.ch/course/view.php?id=1679

Course Schedule with Recommended Readings and Online Resources

Part 1: Fundamentals of Big Data Analysis

Overview Session: Fundamentals of Big Data Analysis
Friday, Apr. 24 / May 8, 15:00-16:30
This session takes place via the video conferencing platform Google Meet

This interactive session provides an overview of the topic of big data analysis including a general introduction, background on handling and processing of big data, methodological challenges and problems as well as use cases. You will be required to log onto the video conferencing system (see link above) and attend the session online; the slides will be provided he

In-depth Sessions

For each of the topics covered in the overview session four in-depth sessions will be provided as pre-recorded lectures and can be accessed whenever it is most convenient for you. Readings for each of these sessions are provided below. The links to the recordings will be provided through the course website.

In-depth Session 1: Introduction – What is Big Data?

Dutcher, Jenna. (2014). What is Big Data? *UC Berkeley Data Science Blog.*
Press, Gil. (2014). 12 Big Data Definitions: What's Yours? *Forbes Blog*.
Manovich, Lev. (2012). Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities*, edited by Matthew K. Gold. The University of Minnesota Press.
Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. (2009). Computational Social Science. *Science* 323(5915): 721-723.

In-depth Session 2: Handling and Processing Big Data

Atz, Ulrich. (2013). 11 Tips on How to Handle Big Data in R. *Open Data Institute Blog*.
Lockwood, Glenn. (2014). Conceptual Overview of Map-Reduce and Hadoop. *Blog Post.*
Jacobs, Bill. (2015). Using Hadoop with R: It Depends. *Blog Post.*
Penchikala, Srini. (2015). Big Data Processing in Apache Spark – Part 1: Introduction. *InfoQ Article.*
Venkataraman, Shivaram. (2015). Announcing SparkR: R on Spark. *Databricks Blog Post.* (*R* package)

In-depth Session 3: Methodological Challenges and Problems

Bollier, David (2010). The Promise and Peril of Big Data. *The Aspen Institute Communications and Society Program*.
Cate, Fred H. (2014). The Big Data Debate. *Science* 346(6211): 818-818.
Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176): 1203-1205.
Lazer, David. (2015). The Rise of the Social Algorithm. *Science* 348(6239): 1090-1091.

In-depth Session 4: Example Applications

Viktoria Spaiser, Thomas Chadefaux, Karsten Donnay, Fabian Russmann, and Dirk Helbing. (2017). Communication Power Struggles on Social Media: A Case Study of the 2011-12 Russian Protests. *Journal of Information Technology & Politics* 14(2): 132-153.
Karsten Donnay. (2017). Big Data for Monitoring Political Instability. *International Development Policy* 8.1 (Online).
Pablo Barberá and Thomas Zeitzoff. (2018). The New Public Address System: Why Do World Leaders Adopt Social Media? *International Studies Quarterly* 62(1): 121-130.

Part 2: Big Data Analysis in Practice

Case Studies Coordination Session
Friday, Apr. 24 / May 8, 16:30-17:30
This session takes place via the video conferencing platform Google Meet

The successful completion of this class entails working on one of the three case studies listed below that allow you to apply the concepts of big data analysis to practical questions. The case studies are:

- Case Study 1: Using Big Data for Malnutrition Early Warning
- Case Study 2: Using Big Data for Urban Development
- Case Study 3: Using Big Data to Analyze Aid Flows

In this online-only setting, the case study work will have to be completed remotely. Case study reports are due at the end of next week, on **Sunday May 3 / May 17, 2020 (midnight)**. You may work on the case studies in groups of max. 2-3 people. Details on the coordination of groups and topics will be given in this session.

Case Study Check-in Session (non-obligatory)
Wednesday, Apr. 29 / May 13, 16:30-17:30
This session takes place via the video conferencing platform Google Meet

This is a non-obligatory check-in session in case you and your group have any questions about your case. You can also send an email any time to karsten.donnay@graduateinstitute.ch and I will try to get back to you as soon as possible.

Case Study Materials and Instructions

For each of the three case studies the case description and the specific tasks will be made available through the course website on Moodle. When working on the case studies you may use any information you find as long as it is properly sourced (e.g. using proper citations). The case studies will be made available only at the beginning of the coordination session on Friday to guarantee that each group has the same time to work on them.