THE GRADUATE INSTITUTE GENEVA

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

Interdisciplinary Programme
Academic Year 2016 - 2017

Big Data Analysis

MINT078 – Autumn 2016-17 - 3 ECTS
Workshop, October 21-22 (S5)

PROFESSOR

Karsten Donnay
karsten.donnay@graduateinstitute.ch
+41 22 908 59 42

ASSISTANT

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 research questions. 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. This includes practical exercises to familiarize students with the format of big data. It also provides a first hands-on experience in handling and analyzing large, complex data structures. The block course is designed as a primer for anyone interested in attaining a basic understanding of what big data analysis entails. There are no prerequisite requirements for this course.

Syllabus

Course Requirements

Requirement 1: Attendance in all parts of the workshop is required and students are expected to engage with the recommended readings and/or online resources in preparation for the course. It is essential that you come prepared and actively participate.

Requirement 2: Students will be required to complete case study exercises in small groups throughout the course. Evaluation will be based on (i) individual performance and participation throughout these exercises; (ii) a brief written case study report; and (iii) an oral presentation of results in the course. (ii) and (iii) are jointly prepared by each small case study group.

Course Evaluation

Performance in the course depends both on active participation 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 and looking for reference works with examples on recent trends in big data analysis. Recommended readings and/or online resources for individual sessions are provided with stable links in the course schedule below.

- Mayer-Schönberger, Viktor and Kenneth Cukier. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Mariner Books.
- McGovern, Tim (ed.). (2016). Big Data Now: 2015 Edition. O'Reilly Media.

The first is a popular science book that provides an excellent introduction to big data, predominant trends, promises and challenges written by two renown experts, one from the Oxford Internet Institute and the other from the Economist. Big Data Now is a compilation of blog posts that covers a wide range of topics. Its scope is beyond that of this class but it is a good resource to browse and read up on specific topics. It can be downloaded for free via the link provided (the previous 2014 edition is available here).

Overview of the Course

The first day focuses on providing a theoretical and practical introduction to big data, its analysis and associated challenges. During the second day, students then apply this knowledge in an in-depth case study and prepare a case study report and oral presentation. Any technical demonstrations throughout the course are done in *R* and their code will be made available.

Day 1: Fundamentals of Big Data Analysis

1. Introduction – What is Big Data?
2. Handling and Processing Big Data
3. Methodological Challenges and Problems
4. Example Application – Using Twitter to Analyze Political Discourse

Day 2: Big Data Analysis in Practice

5. Case Study Session 1
6. Case Study Session 2
7. Preparation of Case Study Report and Presentation
8. Case Study Presentation

Course Schedule with Recommended Readings and Online Resources

Day 1: Fundamentals of Big Data Analysis

Session 1: Introduction – What is Big Data?

Friday, October 21, 9:00-10:30

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.

Session 2: Handling and Processing Big Data

Friday, October 21, 10:45-12:15

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)

Session 3: Methodological Challenges and Problems

Friday, October 21, 14:00-15:30

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.
Ulfelder, Jay. (2015). The Myth of Comprehensive Data. *Blog Post.*
Weller, Nicholas, and Kenneth McCubbins. (2014). Raining on the Parade: Some Cautions Regarding the Global Database of Events, Language and Tone Dataset. *Blog Post*.

Session 4: Example Application – Using Twitter to Analyze Political Discourse

Friday, October 21, 15:45-17:15

Viktoria Spaiser, Thomas Chadefaux, Karsten Donnay, Fabian Russmann, and Dirk Helbing. (2014). Communication Power Struggles on Social Media: A Case Study of the 2011-12 Russian Protests *SSRN* doi:10.2139/ssrn.2528102.

Day 2: Big Data Analysis in Practice

Session 5: Case Study Session 1
Saturday, October 22, 9:00-10:30

Selection of case study topics and formation of small working groups. Students engage with the cases, read through background material provided in the session and work through an initial set of questions to deepen the understanding of the case. Sample applications and data is provided to help students familiarize themselves with the cases and available (big) data.

Session 6: Case Study Session 2
Saturday, October 22, 10:45-12:15

Groups are given a specific task relevant to the case in question and are expected to develop a corresponding big data concept using the knowledge gained in the course and the parameters set by the case study scenario. A set of questions that help guide through the scenarios will be provided.

Session 7: Preparation of Case Study Report and Presentation
Saturday, October 22, 14:00-15:30

Each group prepares a short 2-5 page report on their results and a 10 min oral presentation of their big data concept. There are no further requirements on the exact format of the report or the how the results are presented to the course (slides, flipchart etc.).

Session 8: Case Study Presentations
Saturday, October 22, 15:45-17:15

Presentation of big data concept to the group and discussion of results.