Digital Social Research

Academic Year 2013-2014 Hilary Term
Day and Time: TBC
Location: Meeting Room, 1 St Giles

Course Providers
Dr Bernie Hogan, Oxford Internet Institute, bernie.hogan@oii.ox.ac.uk
Dr Jonathan Bright, Oxford Internet Institute, jonathan.bright@oii.ox.ac.uk

Background
Digital data (and the now fashionable “Big Data”) provide new opportunities for creative and compelling social science. As great as these opportunities are, they also provide new challenges for standard social science research designs. Issues including data management, access, and interactivity require technical skills not typically found in the social science curricula. This course is designed to be a simultaneous introduction to digital social research by examining examples of current research in this area as well as an introduction to the skill set necessary to begin conducting such research.

The Internet is a ready-made source of usable and interesting data. Thus, unlike the Research Methods course, which focuses on reactive data via interviews, surveys, experiments, etc...we focus here on non-reactive data. In particular we make use of Twitter, Wikipedia, email archives, Facebook and Reddit, but the techniques are amenable to most social data providers such as YouTube, Spotify, Last.FM, Kiva and Tumblr. We will explore four basic dimensions of the structure of social life: geography, temporality, relational structure, and topicality. We demonstrate robust and ethical approaches to the analysis of each using data readily found online.

Key Themes

- Where can digital social research provide novel insights into perennial social science questions, as well as new ones created by interactions via the internet and web?
- How do digital contexts shape online behaviors and how to these contexts merely express existing cultural values and psychological traits?
- What are the ethical and legal issues involved in collecting non-reactive data on the internet?
- How can a data collection effort be robust, scalable and statistically meaningful?

Course Objectives

This course will familiarize students with a variety of methods for capturing online data in a host of formats (.json, .xml, .graphml, .csv) and techniques for answering research questions based on data from the web.
In doing so, this course will situate the use of data processing techniques within the larger paradigm of ‘computational social science’. This paradigm suggests a newfound capacity for research insights and predictive analytics that are available from naturalistic trace data and field experiments rather than representative surveys and in-depth interviews. In doing so, it will address the relationship between micro behaviours and macro outcomes on the web, and speak to the wider social science debates about when and how digital data make a difference to social processes. Each week will discuss both practical examples of digital social research but also substantive issues and research topics that are informed by these practical examples.

Learning Outcomes

Upon successful completion of this course students should:

- Have a familiarity with the breadth of digital social research, and distinguish digital social research from conventional survey and interview-based analysis.
- Understand how to use computational tools to answer social science research questions
- Be able to transfer data from machine-readable formats to formats amenable to statistical analysis.
- Be able to engage in collaborative research exercises, including developing tools that can be reused in different contexts.
- Assess the practicality of designing research around streaming / live data.
- Critically appraise contemporary claims about the power of Big Data.

Teaching Arrangements

This course will consist of weekly lectures during Hilary term. Each class is divided into two parts: a shorter, lecture style substantive session where key themes and concepts for the week are explored, and a longer ‘lab’ session where students will be assigned a number of practical programming challenges. Where possible both instructors will be on hand. Code samples will be distributed on Plato before class and students will be expected to submit assignments to a shared repository.

Pre-requisites

This course requires the use of the programming language Python. Students should have a working familiarity with this language, even if it is rudimentary. If a student has no experience with coding, we require one of two routes: first, completion of the three-week, three hour python tutorial held in the department in the first term. An alternative is the successful completion of Codecademy’s online Python beginner track. Students wishing to be exempted from a training course will be asked to submit a code sample in week 4 of Michaelmas term or speak to one of the instructors.

Assessment

Students will be assessed through a final essay that is no longer than 5000 words plus references. An appendix to this essay may contain 500 lines of code or less and detail the technique that was used for data collection in fully-commented executable code. However, the essay will be graded on the basis of the main text, and should consist of a proper literature review, a research question and an original data analysis contribution.

Formative Assessment

This is a hands-on course, and as such, there will be a variety of weekly formative assessments, each relating to a pedagogical goal for the course and the research theme for that week. The reading load for this class has been deliberately made lighter than in other courses to allow time for the completion of these exercises. Coding requires practice.
Submission of Assignments

All coursework should be submitted in person to the Examinations School by the stated deadline. All coursework should be put in an envelope and must be addressed to ‘The Chairman of Examiners for the MSc in Social Science of the Internet C/o The Clerk of Examination Schools, High Street. Students should also ensure they add the OII coversheet at the top of the coursework and that two copies of the coursework are submitted. Please note that all work must be single sided. An electronic copy will also need to be submitted to the department. Please note that all coursework will be marked anonymously and therefore only your candidate number is required on the coversheet.

Please note that work submitted after the deadline will be processed in the standard manner and, in addition, the late submission will be reported to the Proctors’ Office. If a student is concerned that they will not meet the deadline they must contact their college office or examinations school for advice. For further information on submission of assessments to the examinations school please refer to http://www.admin.ox.ac.uk/schools/oxonly/submissions/index.shtml.

For details on the regulations for late and non-submissions please refer to the Proctors website at http://www.admin.ox.ac.uk/proctors/info/pam/section9.shtml.

Any student failing this assessment will need to follow the rules set out in the OII Examing Conventions regarding re-submitting failed work.

Topics

1. Computational Social Science and the language of data
2. Research designs and authentication
3. Data cleaning and geographic mash-ups
4. Webometrics and network visualizations
5. User generated content and topic classification
6. The logic of participation and temporality
7. Consolidating research and ethics through Facebook
8. Reflections on Big Data and the “End of Theory”

General Readings

The general readings are structured in essential and optional. If a reading is essential, it is because we use multiple chapters from the book and consider it a key reference for this course. The optional readings are resources that we return to on a regular basis, but will not help directly shape the course. We expect students to have purchased a (paper or digital) copy of the essential readings.


Week 1. Computational Social Science and the language of data

Substantive: In this week we introduce the concept of digital social research as a methodological frontier meant to consolidate a variety of research strands within social, computer and information sciences. We will explore the growing importance of this (sub)field for both academic research and the private sector, and highlight briefly some banner examples of ‘DSR’ research.

Practical: This week begins with a discussion of the ways in which data can be accessed through the Internet, via APIs. We cover the following concepts: json, xml, rss, csv and serialization. We will be using Reddit’s json feature to explore online data access.


*Shadbolt, Nigel Berners-Lee, Tim


Welser, Howard T. Smith, Marc Fisher, Danyel Gleave, Eric

“Distilling Digital Traces: Computational Social Science Approaches” SHoORM. Pp. 116-140.

Dutton, William H.


Week 2. Research design and authentication

Substantive: This class will explore the overall design of digital social research projects, and highlight a number of key factors to consider such as sample size and sampling window. Many DSR projects require data to be collected as it is created, which introduces a number of practical problems to the research process.

Practical: We will explore the mechanics of streaming data through the Twitter API, looking at how to authenticate requests, how to avoid rate limiting, and how to effectively store large amounts of data.


Chapter 1: Mining Twitter, pp. 3-21

Bruns, Axel Stiegilz, Stefan


Sloan, L., et al.

“Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter”, Sociological Research Online. 18(3): 7.
Week 3. Data cleaning and geographic mash-ups

**Substantive:** This class will start to explore some of the exciting possibilities created by harvesting large amounts of data. We

**Practical:** We will explore how the twitter data collected in the previous week can be processed, how useful information can be extracted, and how some of this information can be displayed on a map.

* Martin, David
  Cockings, Samantha
  Leung, Samuel


* Russell, Matthew A.

  - Chapter 8: Microformats, pp. 315-324

Graham, Mark
Hale, Scott
Gaffney, Devin

  [http://www.geospace.co.uk/files/icwsm_paper2.pdf](http://www.geospace.co.uk/files/icwsm_paper2.pdf)

Monmonier, Mark
De Blij, H.J.


Leetaru, Kalev et al.

  "Mapping the global twitter heartbeat: The geography of Twitter". 2013. First Monday. 18(5-6).

Kitchin, Rob
Dodge, Martin

  - Chapter 4: *The Transduction of Space*, pp. 65-80
Week 4. Webometrics and network visualizations

**Substantive:** This class will move away from APIs to look at slightly more unstructured forms of data capture from websites. We will explore how to ‘spider’ across webpages, hopping from one link to the next, and how the resulting information can be used to tell us something about the network structure of what we are accessing.

**Practical:** We will explore web crawling in Wikipedia, looking at how chains of articles can be downloaded from original source pages and how ‘regular expressions’ can be used to facilitate this task. We will explore different structures of linking in different language versions of Wikipedia, and also show how to deal with different text ‘encoding’ schemes.

- Chapter 5: Part 1, pp. 175-184

- Chapter 4: Hyperlink Networks, pp. 78-94.

Adamic, Lada
Glance, Natalie

“Divided they blog”. WWW ‘05. ACM Press.

Hargittai, Eszter
Gallo, Jason
Kane, Matthew


Kamps, Jason
Koolen, Marijn


Hecht, Brent
Gergle, Darren


Hogan, Bernie

*The Sage Handbook of Online Research Methods.* 2008. SAGE.
- Chapter 8: *Analysing Social Networks via the Internet,* p. 141-160

Week 5. User generated content and topic classification

**Substantive:** Digital research is about more than just automating data collection. Techniques are also developing which allow automatic coding of what has been collected, allowing us to say something about the topic of a piece of text, its writing style, or the sentiment it was written in. This session discusses the latest research in this area comparing two styles: Machine Learning and Lexicographic approaches.
**Practical:** We will develop a simple (and yet still industry standard) naive bayesian topic classifier for some of the material collected from Wikipedia in week 4, using Python's natural language toolkit.

*Russell, Matthew A.  
- Chapter 5: part 2, pp. 185-218

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 6: Document Filtering, pp. 117-141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindman, Matthew</td>
<td>Political Traffic and the Politics of Search</td>
<td></td>
</tr>
<tr>
<td>Askham, Harry</td>
<td>Crowcroft, Jon</td>
<td></td>
</tr>
</tbody>
</table>

### Week 6. The logic of participation and temporality

**Substantive:** The web lives and breathes with the rhythms of human activity. Sites that are active during the day are ignored at night, the weekend varies from the weekday and some sites fade from relevance over time. In this week we address the issue of examining the web and online activity as a temporal process, and consider it in relation to the sociology of time.

**Practical:** We will explore the capture and visualization of temporal patterns in blog and forum commenting, publication and discussion, and look at the use of RSS feeds as a method for ‘streaming’ data where a streaming API does not exist.

*Russell, Matthew A.  

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
</table>

• Chapter 1: Temporal Regularity, pp. 1-30

Fisher, Danyel  

Flaherty, Michael  

Yasseri, Taha  
Sumi, Robert  

---

**Week 7. Consolidating research and ethics through Facebook**

**Substantive:** This week places a focus on Facebook, a major player in the developing field of digital social research. We will discuss existing Facebook research and some of the potential opened up by their extraordinary collection of data; and we’ll also tackle some tricky ethical and legal questions which researchers need to grapple with when using such data. While we do not cover work by the Facebook Data Team, as we are exploring research capable by outside academics, we still recommend perusing their blog: [https://www.facebook.com/data](https://www.facebook.com/data)

**Practical:** This week will represent somewhat of a consolidation of previous techniques from data access and authentication to geographical and network mapping, topic classification and temporal rhythms. All of these topics will be applied to the rich (and still developing) Facebook Graph API.

• Chapter 2: Mining Facebook, pp. 41-80

Zimmer, Michael  "But the data is already public": On the ethics of research in Facebook.” 2010. Ethics and Information Technology. 12(4):313-325.

Lewisa, Kevin  
Gonzales, Marco  

Week 8. Reflections on Big Data and the “End of Theory”

**Substantive:** This last week is purely substantive, and is focused on the changing nature of social science which is implied by the rise of digital social research. Pundits and scholars alike have pondered whether “Big Data” can eliminate or reduce key social science questions. Is there room left for new theories? Is big data analysis merely “tuning the machine” or can it have transformative potentials? What are we losing and who are we excluding when we gain so much data?

*Savage, Mike*  
Burrows, Roger  

*boyd, danah,*  
Crawford, Kate  

*Gillespie, Tarleton*  
http://culturedigitally.org/2012/11/the-relevance-of-algorithms/

*Anderson, Chris*  
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

*Crompton, Rosemary*  

*Webber, Richard*  
Savage, Mike
Burrows, Roger