# Group 3: Pre-requisites for Internet Data Analysis and Fundamentals of Research Methods

## （Advanced Level）

Main Instructor: Bernie Hogan Director, MSc in Social Data Science
Senior Research Fellow, Oxford Internet Institute, University of Oxford
Teaching assistant: Linda Li, the Dphil student
the teaching assistant of Dr Bernie Hogan

## Software installation

For research in data science we will expect the following set up:

1. Python 3.7 or 3.8 installed as a part of the Anaconda Python distribution of Data Science, or equivalent.
2. A Github account
3. Installation of GitHub desktop or familiarity with the GitHub command line.

We will expect that the class members will have familiarity with Python and basic statistics. A primer sheet will be sent to students shortly with the expected skills with resources concerning these programming skills below

## Programming

- Introductory Python.
    - Students should know the following data structures:
        - String, list, dictionary, set, tuple
    - Students should know the following abstractions:
        - Lists and list comprehensions
        - Functions
            - With arguments including default, list, and keyword arguments
        - Classes
            - What are they, how to create one, what does it mean to extend a class?
    - Libraries:
        - We will be using the following libraries. You do not need to be familiar with all of them, but you should be able to import them. They come standard with Anaconda if you are using this to run Jupyter:
            - Pandas, beautifulsoup (bs4), os, json, network, nltk, scipy.stats, numpy
            - We will be covering each of these in class with examples.

## Math and statistics

Although the math and stats will be lightweight for this course, there will be some. And we will expect rudimentary statistical knowledge:

- Math:
    - What is a vector, what is a matrix, how do you transpose a matrix?
- Statistics:
    - What are measures of central tendency? (mean, median, mode).
    - What is a correlation?
    - What is a p-value?
    - We will be doing basic tests using one-way anovas, correlations, and t-tests.

## Timetable (First 2-hour teaching session plus 90 mins' TA session)

TA (teaching assistant) session means 90 minutes after the first 2-hour teaching session given by the instructor Bernie, for the exercises and explanation given by teaching assistant Linda Li. And you could ask questions to the teaching assistant during the TA sessions.

### Week 1. Data Wrangling and Cleaning in Python

| Week 1 | Internet Data Analysis | | TA Sessions | |
|---|---|---|---|---|
| Feb 1 MON | 16:00-18:00 BJS | 8:00-10:00 GMT | No Sessions | |
| Feb 2 TUE | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30GMT |
| Feb 3 WED | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30GMT |
| Feb 4 THU | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30GMT |
| Feb 5 FRI | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30GMT |
| Feb 6 SAT (Optional) | 16:00-18:00 BJS | 8:00-10:00 GMT | Two sharing sessions by Oxford PhD students on My Life in Oxford (University application experience; the life and my research) | |

### Week 2. Internet Research and Intro to Text Mining in English

This week will focus on developing research questions with data and introduce some basic text mining and classification

| Week 2 | Internet Data Analysis | | TA Sessions | |
|---|---|---|---|---|
| Feb 8 MON | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30GMT |
| Feb 10 WED | 16:00-17:30 BSJ | 8:00-11:00 GMT | Public Lecture 1: Big Data and AI by Prof. Pietro Lio, Computer Science and Technology, University of Cambridge | |
| | 17:30-19:00 BSJ | 11:00-12:30 GMT | Public Lecture 2: The Future of AI and Machine Learning by Prof. Alex Rogers, Computer Science, University of Oxford | |

### Week 3. Network Analysis

This week will focus on network analysis as a paradigm and approaches to the collection and processing of online data

| Week 3 | Internet Data Analysis | | TA Sessions | |
|---|---|---|---|---|
| Feb 15MON | 16:00-18:00 BJS | 8:00-10:00 GMT | 20:30-22:00 BSJ | 12:30-14:00 GMT |
| Feb 16 TUE | 16:00-18:00 BJS | 8:00-10:00 GMT | 21:00-22:30 BSJ | 13:00-14:30 GMT |
| Feb 17 WED | 16:00-18:00 BJS | 8:00-10:00 GMT | 20:30-21:00 BSJ | 12:30 - 13:00 GMT |
| | 21:00-22:30 BJS | 13:00-14:30 GMT | Public Lecture 3: Me, My Spouse and the Internet: Meeting, Dating and Marriage in the Digital Age<br>Dr. Bernie Hogan, Senior Research Fellow, Oxford Internet Institute, University of Oxford | |
| Feb 18 THU | 16:00-18:00 BJS | 8:00-10:00 GMT | No Sessions | |
| Feb 19 FRI | 13:00-14:30 BJS | 21:00-22:30 GMT | Public Lecture 4: The secrets of creativity | |

| | | | |
|---|---|---|---|
| | | | from big data: How do Jazz musicians, video game developers, and software programmers succeed by choosing the right networks?<br>Prof. Balazs Vedres, Oxford Internet Institute, University of Oxford |

## Week 4. Research questions, visualisations, analytical claims

This week will focus on visualisation, network analysis as a paradigm, and how to answer questions analytically and descriptively

| Week 4 | Internet Data Analysis | | TA Sessions | |
|---|---|---|---|---|
| Feb 22 MON | 16:00-18:00 BJS | 8:00-10:00 GMT | 19:00-21:00 BSJ | 11:00-12:30 GMT |
| Feb 23 TUE | 16:00-18:00 BJS | 8:00-10:00 GMT | 19:00-21:00 BSJ | 13:00-14:30 GMT |