STATS 769
Overview

Paul Murrell

The University of Auckland

July 18, 2019

Overview

- This course covers topics related to the analysis of large and/or complex data sets.
- We will cover a range of data technologies for accessing and processing data.

In my view, we do need a term that covers that full life cycle - not building an IT platform and network for databases perhaps, but yes, building the occasional database given the infrastructure is in place; managing large amounts of data and knowing enough about IT at least to know where and how to get help; reshaping and wrangling many many datasets and knowing the tricks of the trade to combine them in meaningful ways; knowing how to manage complex, large, conflicting systems of classifications; fitting statistical models and performing inference; and efficiently presenting results in high quality presentations and graphics using tools like LaTeX, JavaScript, Shiny, etc.

Peter Ellis
MANAGER SECTOR PERFORMANCE
Ministry of Business, Innovation & Employment

Data technology topics

- Working in Linux
  - Surviving the shell: a CLI for the operating system.
  - Prospering in the shell: power and automation.
- Coping with data that is not already rectangular.
  - Data formats (JSON and XML).
  - Data from the Web.
- Coping with data that is too large (for RAM or your hard drive).
  - Make the machine bigger.
  - Make the data smaller.
  - Use different software.
- Making your code run faster.
  - Writing fast (and slow) R code.
  - Making use of multiple cores.
  - Making use of remote machines.

2019-07-22 week 1 Paul	2019-07-23 week 1 Paul	2019-07-24 week 1 overview Paul	2019-07-25 week 1 data science workflow Paul	2019-07-26 week 1 Paul
2019-07-29 week 2 Paul	2019-07-30 week 2 Paul	2019-07-31 week 2 data tech review Paul	2019-08-01 week 2 data tech review Paul	2019-08-02 week 2 Lab 1 Paul
2019-08-05 week 3 Paul	2019-08-06 week 3 Paul	2019-08-07 week 3 surviving linux Paul	2019-08-08 week 3 surviving linux Paul	2019-08-09 week 3 Lab 2 Paul
2019-08-12 week 4 Paul	2019-08-13 week 4 Paul	2019-08-14 week 4 prospering in linux Paul	2019-08-15 week 4 prospering in linux Paul	2019-08-16 week 4 Lab 3 Paul
2019-08-19 week 5 Paul	2019-08-20 week 5 Paul	2019-08-21 week 5 data formats Paul	2019-08-22 week 5 data formats Paul	2019-08-23 week 5 Lab 4 Paul
2019-08-26 week 6 Paul	2019-08-27 week 6 Paul	2019-08-28 week 6 web scraping Paul	2019-08-29 week 6 TERM TEST Paul	2019-08-30 week 6 Lab 5 Paul
2019-09-02	2019-09-03	2019-09-04	2019-09-05	2019-09-06
2019-09-09	2019-09-10	2019-09-11	2019-09-12	2019-09-13
2019-09-16 week 7 Paul	2019-09-17 week 7 Paul	2019-09-18 week 7 large data problems Paul	2019-09-19 week 7 large data problems Paul	2019-09-20 week 7 Lab 6 Paul
2019-09-23 week 8 Paul	2019-09-24 week 8 Paul	2019-09-25 week 8 large data solutions Paul	2019-09-26 week 8 large data solutions Paul	2019-09-27 week 8 Lab 7 Paul
2019-09-30 week 9 Paul	2019-10-01 week 9 Paul	2019-10-02 week 9 code efficiency Paul	2019-10-03 week 9 code efficiency Paul	2019-10-04 week 9 Lab 8 Paul
2019-10-07 week 10 Paul	2019-10-08 week 10 Paul	2019-10-09 week 10 parallel code Paul	2019-10-10 week 10 parallel code Paul	2019-10-11 week 10 Lab 9 Paul
2019-10-14 week 11 Paul	2019-10-15 week 11 Paul	2019-10-16 week 11 HPC Paul	2019-10-17 week 11 HPC Paul	2019-10-18 week 11 Lab 10 Paul
2019-10-21 week 12 Paul	2019-10-22 week 12 Paul	2019-10-23 week 12 other systems Paul	2019-10-24 week 12 other systems Paul	2019-10-25 week 12 Paul

Course structure

- Lecture Wednesday 2-3 303.101
- Lecture Thursday 2-3 303.101
- Lab Thursday 4-6 302.190
- Lab Friday 12-2 303S.175

Assessment

- There are **10 labs**.
- There is a **TERM TEST**.
- There is an **EXAM**.

- Each Lab will be worth 3% of your mark (for a total of 30%).
- The Term Test will be worth 20% (with **plussage**).
- There will be a (closed-book, no calculator) written exam worth 50%.
  You must pass the exam in order to pass the course.

Assessment

- We will cover one topic each week.
- Handouts for each lab will be made available at the start of the week.
- Lab attendance is not compulsory, but is the best place to receive help and feedback.
- The lab submission is **online** and will be due on **the following Monday** (consult Canvas for exact submission date-times).
- There is no lab in the first week and no lab in the last week.

Resources

- "An Introduction to Statistical Learning"
  Gareth James,Daniela Witten, Trevor Hastie and Robert Tibshirani
  `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`
- "Introduction to Data Technologies" (assumed knowledge)
  Paul Murrell
  `https://www.stat.auckland.ac.nz/~paul/ItDT/`
- "Advanced R"
  Hadley Wickham
  `http://adv-r.had.co.nz/`
- "XML and Web Technologies for Data Sciences with R"
  Deborah Nolan and Duncan Temple Lang
  `http://link.springer.com/book/10.1007%2F978-1-4614-7900-0`

Class Rep

- Volunteer(s) required !