Speaker: Nicholas Chamandy, Google

Title: Estimating uncertainty in massive data streams

Abstract:

Massive data streams are becoming the norm in modern applied statistics, especially in the internet technology sector. Several properties of these structures make statistical analysis of the data therein challenging:

- Sheer volume of the data (large number of records)
- Inability to access multiple records simultaneously
- Lack of a meaningful sorting of the records
- Correlation between groups of records

While basic nearly-linear statistics can be computed exactly or approximately from such data, variance estimation and other “second-order” analysis is problematic. Unfortunately, massive sample sizes do not obviate the need for uncertainty calculations: modern data often have heavy tails, large coefficients of variation, tiny effect sizes, and generally exhibit bad behaviour.

In this talk, we describe in detail this “new frontier” in statistics, outline the computing infrastructure required, and motivate the need for modification of existing methods. We introduce two such modified procedures for basic uncertainty estimation, one derived from the bootstrap and the other from a form of subsampling. Their costs and properties are briefly discussed, and their use is demonstrated via a simulation study.

This is joint work with Omkar Muralidharan and Amir Najmi.