Abstracts from Statistical Modelling and Analysis of Big Data workshop 2015

Dr Simon Angus, Monash University

Drinking from the fire-hydrant: global online/offline internet activity, four times an hour.

Abstract:
In this talk, I will share our group's work so far on handling and analysing global Internet Protocol (IP) address activity (online/offline) at granular (15min) temporal resolution over a 7 year period. So far, we have successfully linked an ip-activity database (raw 150TB) to a commercial geo-location database (500GB), providing opportunities to build unique datasets at any internet-connected location, such as a country, state, LGA, or city. Presently, we are focussing on >100,000 population cities as spatial units of analysis, given the increasing interest in cities as loci of economic activity. The work has drawn extensively on high-performance, distributed, computing assets available to Australian researchers, and has presented numerous data-processing and data-science challenges. Methodologically, stream-processing map-reduce tools together with wavelet, clustering and geo-spatial tools have been prominent. Preliminary results at a single-city, and multi-city level will be presented, hinting at the breadth of social science opportunities such a dataset affords.

**********

Prof John Geweke, University of Technology Sydney

A Hierarchical Forecasting Engine for Massive Longitudinal Data

Abstract:
The capacity to store, retrieve and manipulate large volumes of data has grown dramatically in recent years and will continue to do so in the foreseeable future. These innovations bear on all established agendas in forecasting and define new ones. Responding to these developments, this paper develops a large hierarchical tree structure for the modeling and forecasting of longitudinal discrete data that is applicable to data having millions of cross-section dimensions and thousands of time dimensions. It caters to circumstances in which models for different parts of the tree are developed by subject matter experts, a situation arising both in the academic world as well as large business establishments and government agencies. The structure ensures that models and forecasts are logically consistent, despite the decentralization, and permits the generation of forecasts individual tailored to selected cross-sectional and time dimensions in real time.

**********

Prof Rob Hyndman, Monash University

Visualizing and forecasting big time series data.

Abstract:
Many organizations are collecting enormous quantities of time series data. For example, a manufacturing company can disaggregate total demand for their products by country of sale, retail outlet, product type, package size, and so on. As a result, there can be millions of individual time series to forecast at the most disaggregated level, plus additional series to forecast at higher levels of aggregation.

The first problem with handling such large numbers of time series is how to produce useful graphics to uncover structures and relationships between series. Data visualization provides an essential tool for exploring, studying and understanding structures and patterns, but the sheer quantity of data challenges the current methodology. I will demonstrate some data visualizations tools that help in exploring big time series data.

The second problem is how to forecast large quantities of time series data, while respecting the various aggregation constraints that often apply. This is known as forecast reconciliation. I will show that the optimal reconciliation method involves fitting an ill-conditioned linear regression model where the design matrix has one column for each of the series at the most disaggregated level. For problems involving huge numbers of series, the model is impossible to estimate using standard regression algorithms. I will also discuss some fast algorithms for implementing this model that make it practicable for implementing in business contexts.
**********

Dr Steve Scott, Google

Bayes and Big Data: The Consensus Monte Carlo Algorithm

Abstract:
A useful definition of ``big data'' is data that is too big to comfortably process on a single machine, either because of processor, memory, or disk bottlenecks. Graphics processing units can alleviate the processor bottleneck, but memory or disk bottlenecks can only be eliminated by splitting data across multiple machines. Communication between large numbers of machines is expensive (regardless of the amount of data being communicated), so there is a need for algorithms that perform distributed approximate Bayesian analyses with minimal communication. Consensus Monte Carlo operates by running a separate Monte Carlo algorithm on each machine, and then averaging individual Monte Carlo draws across machines. Depending on the model, the resulting draws can be nearly indistinguishable from the draws that would have been obtained by running a single machine algorithm for a very long time. Examples of consensus Monte Carlo are shown for simple models where single-machine solutions are available, for large single-layer hierarchical models, and for Bayesian additive regression trees (BART).