Prof Sue Wilson, UNSW

Title: Analysis of Very Large Data Sets containing Different Types of Variables

Abstract: Major challenges arising from today's "data deluge" include how to handle the commonly occurring situation of different types of variables being simultaneously measured, as well as how to address the accompanying flood of questions. Based on information theory, a mutual information (MI) measure of association that is valid and estimable between all basic types of variables has been proposed. It has the advantage of being able to identify non-linear as well as linear relationships. Based on this MI measure, a novel exploratory approach to finding associations in large data sets has been developed. These associations can be used as a basis for finding clusters and networks, say, in large data sets in which continuous and categorical variables have been collected. The application of this approach is very general. Here it will be illustrated on genomic data that includes continuous variables (gene expression values), categorical variables (genotypes) and clinical variables. As well, comparisons will be made with the recently proposed measure, maximal information coefficient (MIC: Science 334:1518-1524). This is joint research with PhD student Chris Pardy.