Data thinning and beyond

A seminar by Daniela Witten

University of Washington

Thursday 22 Jan 2026 | 14:30-15:30

Room BENVENUTI

Department of Statistical Sciences

Contemporary data analysis pipelines often involve the use and reuse of data. For instance, a scientist may explore a dataset to select an interesting hypothesis, and then wish to test this hypothesis with the same data. From a statistical perspective, this double use of data is highly problematic: it induces dependence between the hypothesis generation and testing stages, which complicates inference. Failure to account for this dependence renders classical inference techniques invalid.

I will present "data thinning", a set of strategies for obtaining independent training and test sets so that the former can be used to select a hypothesis, and the latter to test it. Data thinning enables valid selective inference in settings for which no solutions were previously available. However, it is also restrictive, in the sense that it requires strong distributional assumptions. Therefore, I will also present strategies inspired by data thinning that enable valid post-selection inference without such assumptions.

Università degli Studi di Padova

Dipartimento di Scienze Statistiche