Joint LMU-TUM Statistics Colloquium:

Data Thinning and beyond

Daniela Witten

(University of Washington)

06/05/2026, 4:15 pm

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 631 1190 7291; Password: StatsCol)

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 two strategies inspired by data thinning that enable valid post-selection inference without such assumptions. One strategy considers thinning summary statistics of the data, rather than the data itself, in order to take advantage of asymptotic properties of the summary statistics. The second strategy involves generating training and test sets that are not independent, and then orthogonalizing the latter with respect to the former in order to conduct valid inference.

About the Speaker:
Daniela Witten is a professor of Statistics and Biostatistics at University of Washington, and the Dorothy Gilford Endowed Chair in Mathematical Statistics. She develops statistical machine learning methods for high-dimensional data, with a focus on unsupervised learning. She has received a number of awards for her research in statistical machine learning: most notably the Spiegelman Award from the American Public Health Association for a (bio)statistician under age 40, and the Presidents' Award from the Committee of Presidents of Statistical Societies for a statistician under age 41. Daniela is a co-author of the textbook *Introduction to Statistical Learning*, and since 2023 serves as Joint Editor of Journal of the Royal Statistical Society, Series B.