Seminar on Combinatorial Computing
February 13, 2002, Wednesday, 6:30 p.m.
Room 6417, Graduate Center
365 Fifth Avenue, New York

# An algorithm for clustering high-dimensional data

Sanjoy Dasgupta
AT&T Laboratories

**Abstract**

Clustering – finding groups in data – lies at the heart of some of the most exciting new developments in computer science, such as computational genomics and web document retrieval. These emerging areas require algorithms for rapidly clustering large data sets in very high dimension.

One of the most common ways to cluster data is to model it as a "mixture of Gaussians" (the standard statistical model of clustered data). Despite its importance, and despite over a century of close study, the only algorithms known for this clustering task have been heuristics with weak performance guarantees.

I shall describe the first efficient, provably correct algorithm to learn a mixture of Gaussians from arbitrarily high-dimensional data.

The algorithm depends crucially upon an understanding of the geometric structure of high-dimensional clusters, and upon a particular projection technique which reduces the dimension of the data while preserving important aspects of this structure.

For further information contact Janos Pach at pachcims.nyu.edu, or visit our website: http://cs.gc.cuny.edu/ pach/