

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, March 13, 2007
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

David G. Stork
Chief Scientist, Ricoh Innovations
Visiting Lecturer, Department of Statistics, Stanford University

Toward a statistical theory of data acquisition

There are deep theoretical justifications and compelling experimental verification that there is no “best” general method for statistical pattern classification and, further, that classifiers perform better the larger their training sets. Even simple unbiased classifiers, when trained with sufficiently large training sets, can outperform more sophisticated classifiers. These facts imply that the most promising avenues for research in pattern classification are no longer in developing refinements to general classification methods themselves, but rather in developing novel, efficient, and accurate methods for collecting, labelling and “truthing” large data sets for training simple, scalable classifiers.

This talk will review the foundations of a statistical theory of data acquisition, including problems such as acquiring data under cost constraints, estimating the accuracy and reliability of contributors, organizing the self-policing among data contributors, and identifying “malicious” contributors. It will describe these and other challenges and opportunities associated with novel methods of data acquisition, such as the Open Mind Initiative, in which non-experts openly contribute data over the internet. This talk will explore the relationship of this nascent theory of data acquisition to polling theory, experimental design, and interactive learning, and it will conclude by describing a number of open research problems.

Joint work with Chuck Lam

David G. Stork is Chief Scientist of Ricoh Innovations and Visting Lecturer in Statistics at Stanford University, where he will teach **Stat 328**, “Statistical theory of data acquisition,” this spring quarter. He has held academic posts in eight different disciplines, served on five editorial boards, and holds 35 patents. His roughly 120 scholarly publications are in theoretical mechanics, human visual perception, pattern classification, machine learning, computer lipreading, theory of concurrency, optical design, and image processing. Most recently he has pioneered and lectured widely on the use of computer vision methods for the analysis of Renaissance and Baroque master paintings. His five books include **Pattern Classification** (2nd ed.) with R. Duda and P. Hart, and **HAL’s Legacy**, the companion to his PBS television documentary about the famous computer in *2001: A Space Odyssey*.