

STANFORD UNIVERSITY  
DEPARTMENT OF STATISTICS  
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, October 10, 2000  
Sequoia Hall Rm. 200  
(Cookies at 3:45 in 1st Floor Lounge)

*David L. Donoho*  
*Department of Statistics*  
*Stanford University*

**High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality**

This is a replay of a talk I gave at the American Math Society's meeting this summer on "Mathematical Challenges of the 21st Century". I tried to identify a large trend, lasting decades, in which mathematics (and statistics) could play a role. Here is a brief extract of what I came up with ...

The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. Hyperspectral Imagery, Internet Portals, Financial tick-by-tick data, and DNA Microarrays are just a few of the better-known sources, feeding data in torrential streams into scientific and business databases worldwide.

"In traditional statistical data analysis, we think of observations of instances of particular phenomena (e.g. instance  $\leftrightarrow$  human being), these observations being a vector of values we measured on several *variables* (e.g. blood pressure, weight, height, ...). In traditional statistical methodology, we assumed many observations and a few, well-chosen variables. The trend today is towards more observations but even more so, to radically larger numbers of variables – voracious, automatic, systematic collection of hyper-informative detail about each observed instance. We are seeing examples where the observations gathered on individual instances are curves, or spectra, or images, or even movies, so that a single observation has dimensions in the thousands or billions, while there are only tens or hundreds of instances available for study. Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector. We can say with complete confidence that in the coming century, high-dimensional data analysis

will be a very significant activity, and completely new methods of high-dimensional data analysis will be developed; we just don't know what they are yet.

“Mathematicians are ideally prepared for appreciating the abstract issues involved in finding patterns in such high-dimensional data. Two of the most influential principles in the coming century will be principles originally discovered and cultivated by mathematicians: the blessings of dimensionality and the curse of dimensionality.

...(Elaboration)

There is a large body of interesting work going on in the mathematical sciences, both to attack the curse of dimensionality in specific ways, and to extend the benefits of dimensionality. I will mention work in high-dimensional approximation theory, in probability theory, and in mathematical statistics. I expect to see in the coming decades many further mathematical elaborations to our inventory of Blessings and Curses, and I expect such contributions to have a broad impact on society's ability to extract meaning from the massive datasets it has decided to compile.”

You can see more at

<http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html>