# Breast Cancer Detection

*John Clements, Rong Huang, Shantel Ward*

## Keywords:

Classification, Principal Components Analysis (PCA), MANOVA, Breast Cancer

## Introduction:

We are interested in identifying the ways in which malignant and benign tumors differ in their characteristics, if an informative lower dimensional structure of the variables exists, and using those differences for the purpose of building a classifier. We will judge classifiers on accuracy, but also discuss the relative costs of False Positives versus False Negatives.

Scientific Questions:

1. Are there statistically significant differences between mean vectors for benign tumors and malignant tumors?
2. Can we use PCA to reduce the dimensionality of the data and to identify/summarize crucial variables?
3. Which classification model yields the highest accuracy for predicting if a tumor is benign or malignant?
4. Are there any outliers in the multivariate data set?

## Methods:

The data was collected from 699 patients of Dr. William H. Wolberg at the University of Wisconsin between January 1989 and November 1991. Measurements were derived from Fine Needle Aspirations (FNA) of human breast masses and analyses were performed on the masses. Each observation is described by nine features (Clump Thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitosis). Each of these nine features of the fine needle aspirates was graded 1 to 10 at the time of sample collection, with 1 being the closest to benign and 10 the closest to malignant.

We intend to use one-way MANOVA analysis (or Hotelling's T2 test for two independent samples) and methods of analyzing pairwise differences to determine the characteristics distinguishing malignant from benign masses, if differences exist. We will also attempt to find a low-dimensional representation of the data using Principal Components Analysis. Finally, we will use cross-validation to compare multiple classification algorithms to decide which one to test on a completely withheld subset of the data.

## Expected Outcome:

We hope to determine what features are significantly different between malignant and benign breast masses, if there is a lower-dimensional representation of the features, and if we can find a classifier that outperforms the no-information rate. We will also analyze the tradeoff between False Negatives and False Positives in the context of diagnosing breast cancer.