## PCA: Principal Component Analysis

* PCA: principal component analysis
  + PCA projects the features onto the principal components
  + The motivation is to reduce the features dimensionality while only losing a small amount of information
    - Principal components are new low dimensional axis (or surfaces) closest to the observations
    - The first principal component (PC) follows a “best fit” through the data points
* What does PCA aim to do?
  + PCA takes a dataset with a lot of dimensions (i.e. lots of cells) and flattens it to 2 or 3 dimensions so we can look at it
    - It tried to find a meaningful way to flatten the data by focusing on the things that are different between cells
  + For example
    - These two “new” axes that describe the variation in the data are “Principal Components” (PCs)
      * PC1 (the first principal component) is the axis that spans the most variation
      * PC2 is the axis that spans the second most variation
    - Some genes have more influence on PC1 than others
      * Genes with little influence on PC1 get values close to zero, and genes with more influence get numbers further from zero
  + Using the two Principle Components to plot cells
    - Combining the read counts for all genes in a cell to get a single value
      * Cell 1 PC1 score = (read count \* influence) + … for all genes
* **BONUS: predictive modeling with PCA components**
  + We can use our PCA and clustering models to predict the potential malignancy of new samples
    - ## Predicting Malignancy of new samples
    - new <- read.csv(“https://tinyurl.com/new-samples-CSV”)
    - npc <- predict(wisc.pr, newdata = new)
    - plot(wisc.pr$x[,1:2], col = (diagnosis+1))
    - points(npc[,1], npc[,2], col = “blue”, pch=16)