CSC 390
Topics in Artificial Intelligence

"Unsupervised Machine Learning"

Fall 2016

Prof. Sara Mathieson
Smith College

Outline: 10/27

- Today:
  - Recap Homework 4
  - Hidden Markov Models (HMMs)
  - Viterbi Algorithm
  - Lab 4

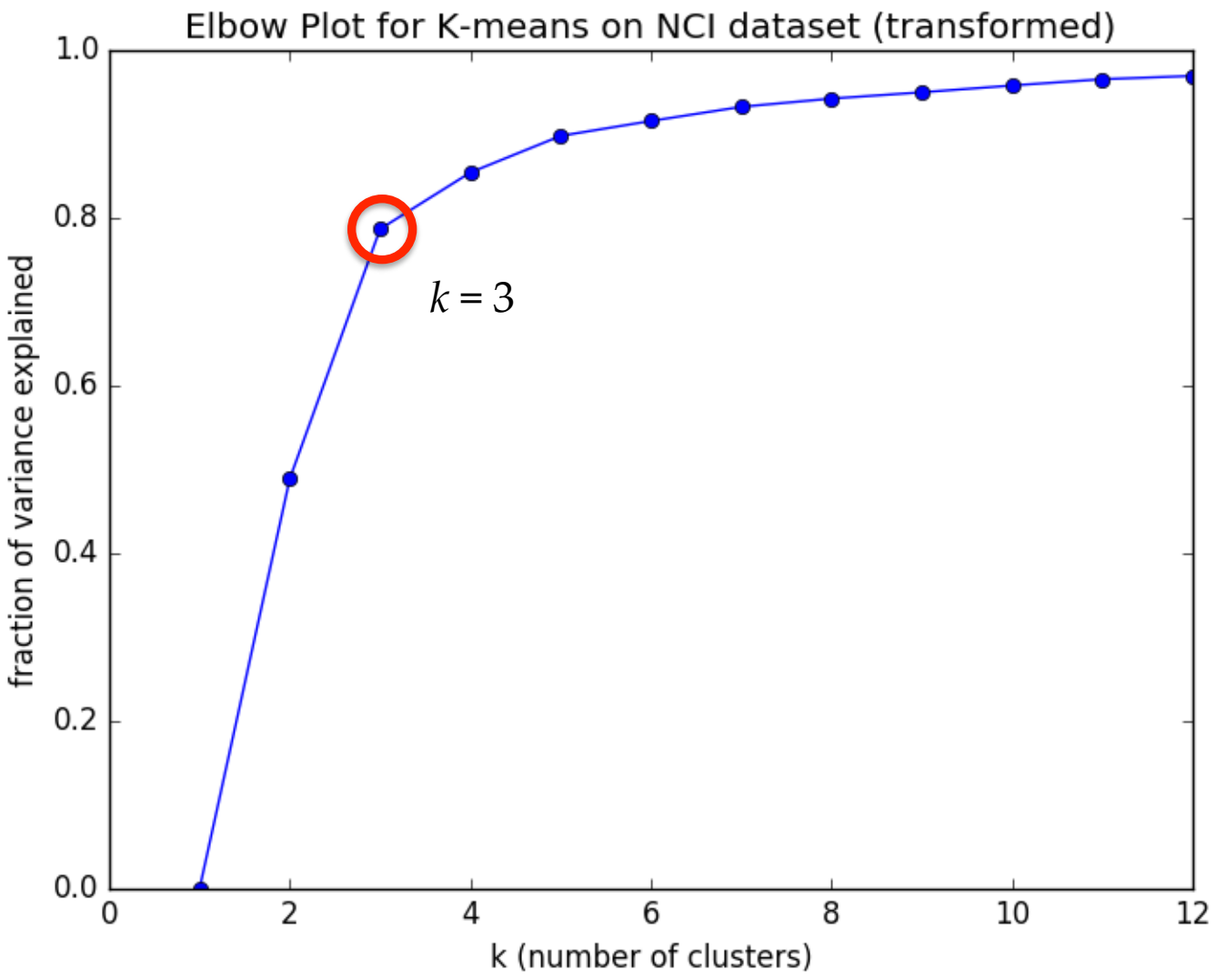
- Office Hours today: 4-5pm, Ford 355

- Next time: start mid-semester presentations
  - Email me your paper!

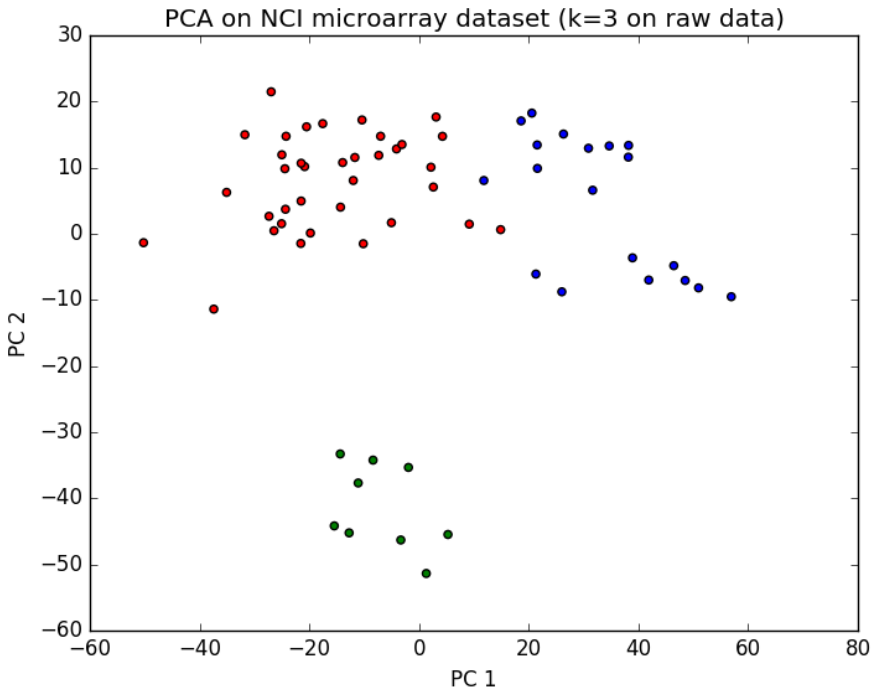
Followups

- Extra materials on Deep Learning and HMMs on Piazza
- Homework 6: we will use a Python package for HMMS
- From notecards:
  - More on PCA, autoencoders, and stats/probability

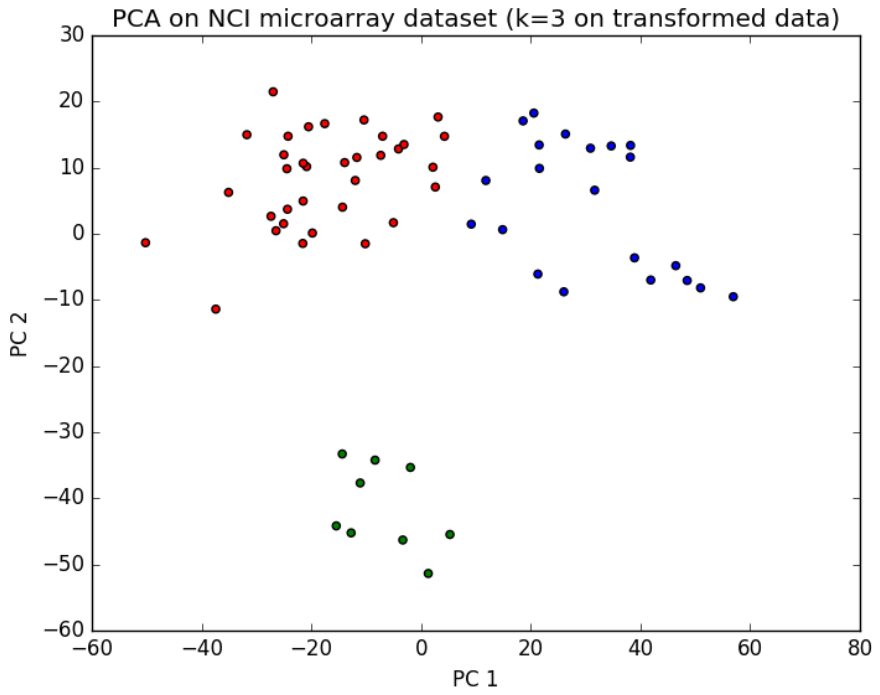
Homework 4: Elbow plot



Homework 4: k-means and PCA

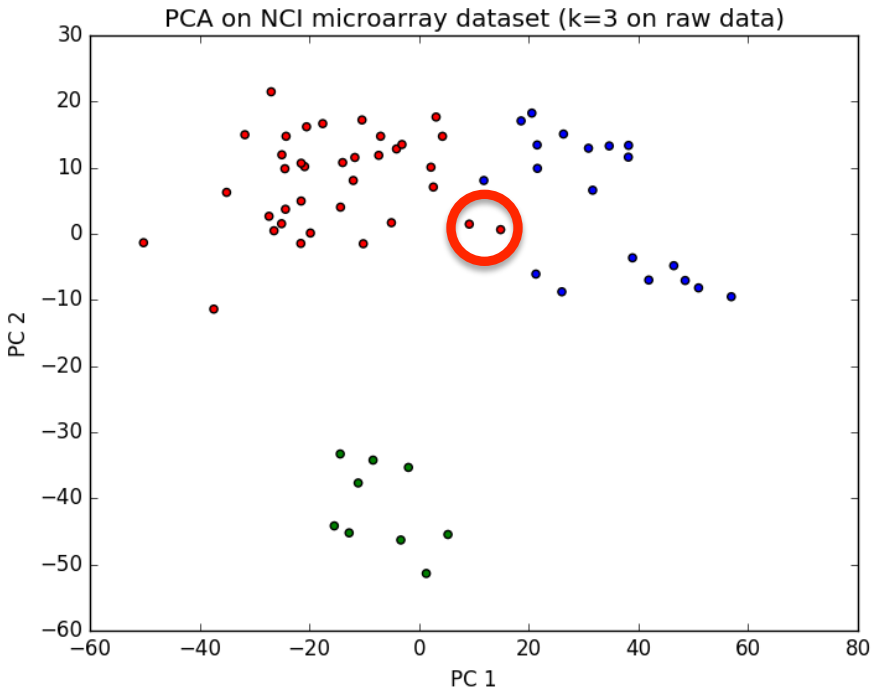


Raw Data

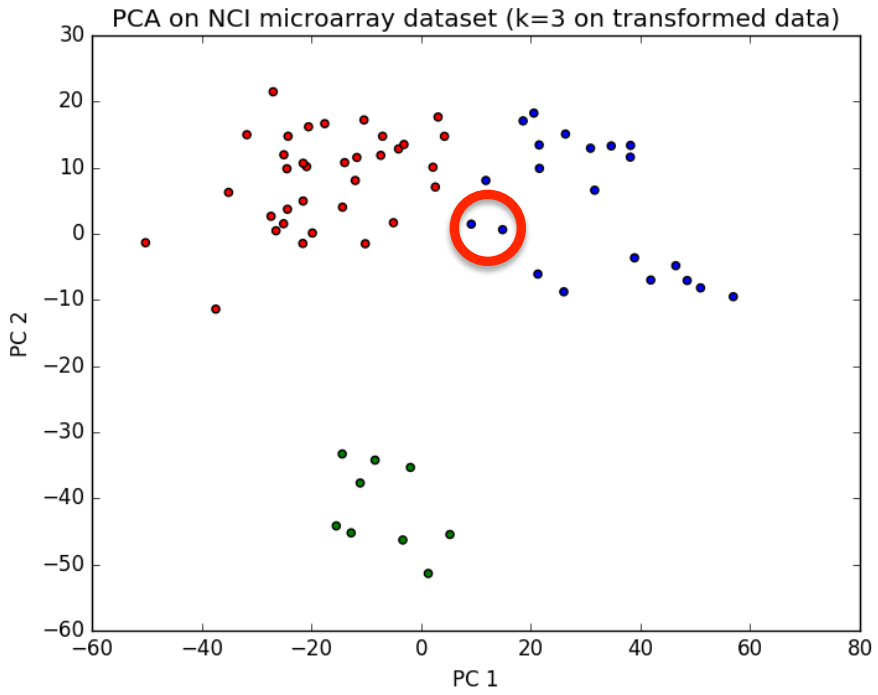


Transformed Data

Homework 4: k-means and PCA

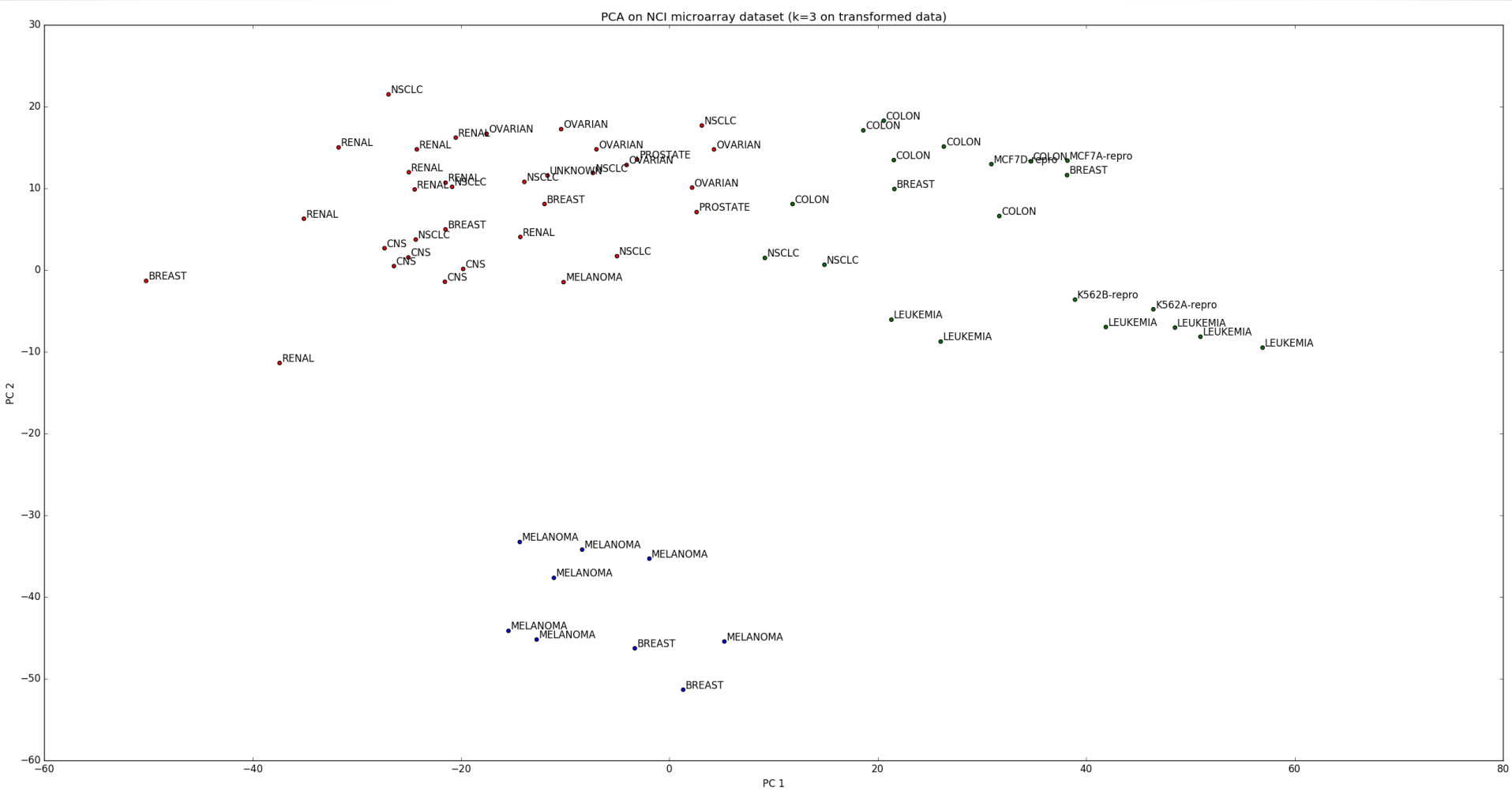


Raw Data

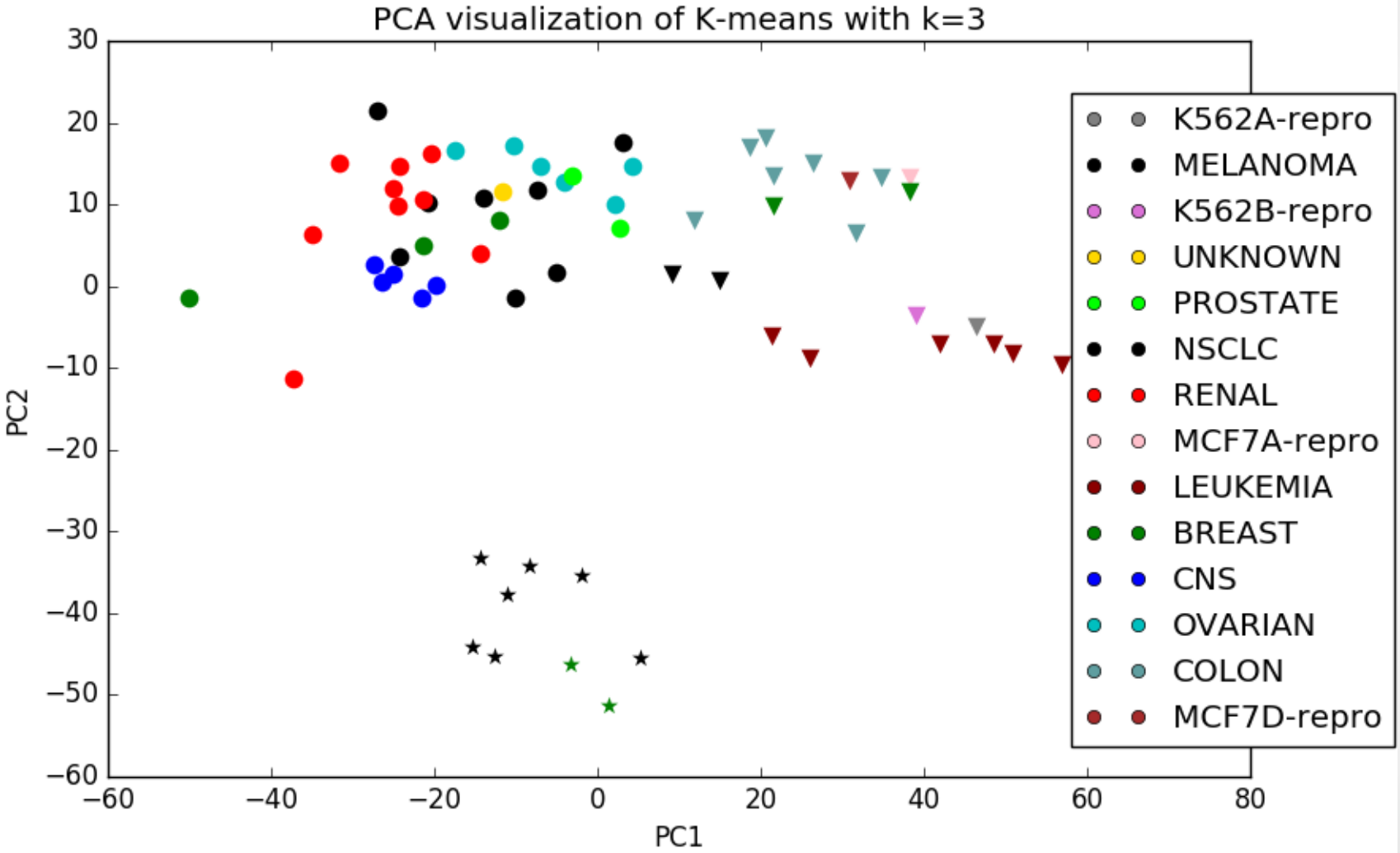


Transformed Data

Homework 4: visualization



Homework 4: visualization



Credit: Farida

Homework 4: observations

Some of our clusters seem to correspond to the labels of the data, but there are some surprises. The blue cluster contains almost all the melanoma samples, but also a few breast cancer samples. The leukemia samples are all in the far right of the green cluster, while the colon cancer samples are at the top, suggesting this cluster should perhaps be broken up. The red cluster also looks like it contains some sub-clusterings (renal, ovarian, CNS).

However, the tissue type is not the only thing that determines a gene expression profile (genetics being another important factor). In this particular real-world example, it might make more sense to focus on the groupings themselves (which are based on genes, which could respond to treatment) and not try to mold the clusterings to fit the labels.