Learning from Microarray Data

Using gene expression data to classify cancer types

Final Presentation

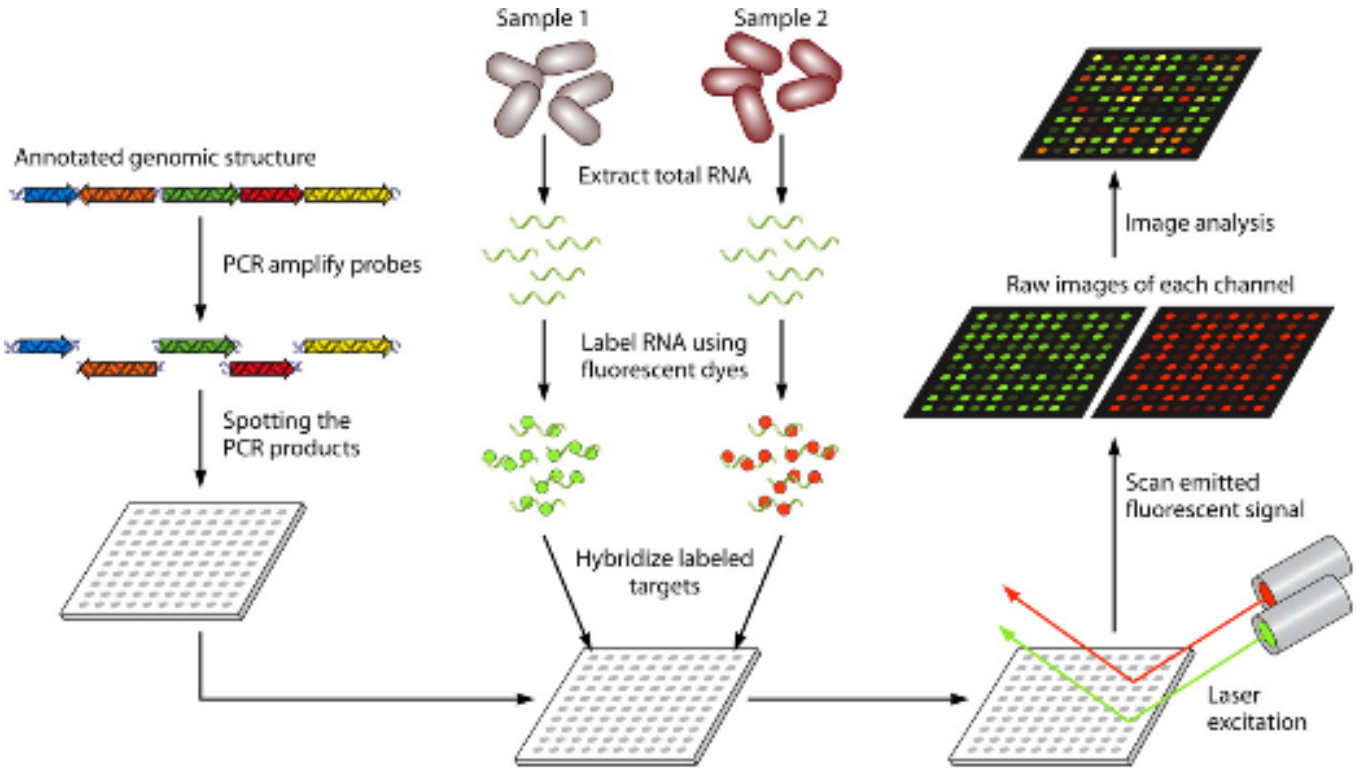
EAP 2019

Arno Veletanlic
Carlos Perez-Guerra

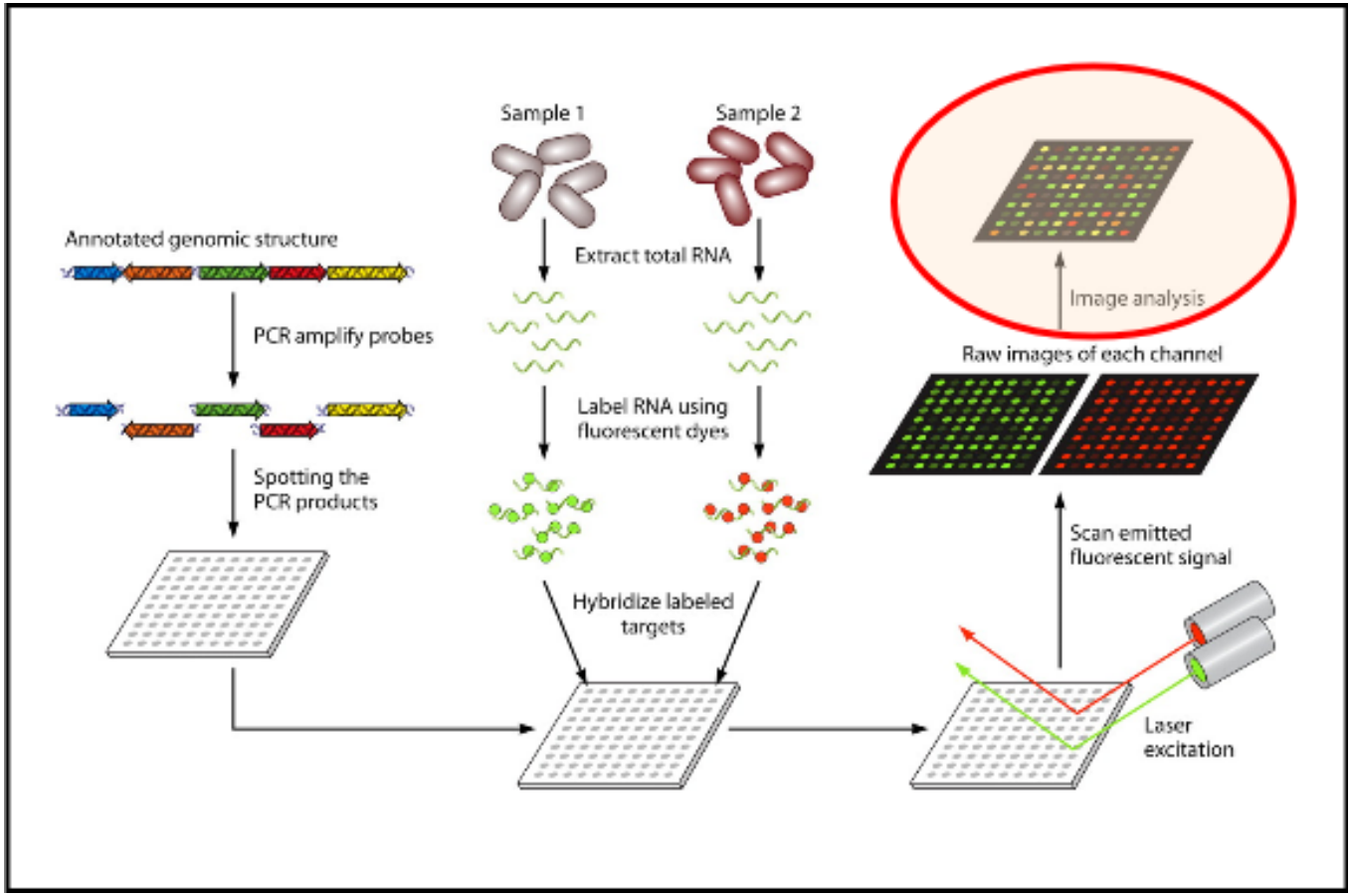
Shanghai Jiaotong University

July 28, 2019

Microarray Genechip Affymetrix



Our Dataset


Sample 1
Sample 2
Annotated genomic structure
Extract total RNA
PCR amplify probes
Label RNA using fluorescent dyes
Spotting the PCR products
Hybridize labeled targets
Image analysis
Raw images of each channel
Scan emitted fluorescent signal
Laser excitation

First Impressions

Shape of the data

- **Variety of labels:** 130 or more diseases recorded
- **Variety of features:** 22,000 or more measurements for each sample
- **Number of samples:** Around 6000 (sufficient)

First Impressions

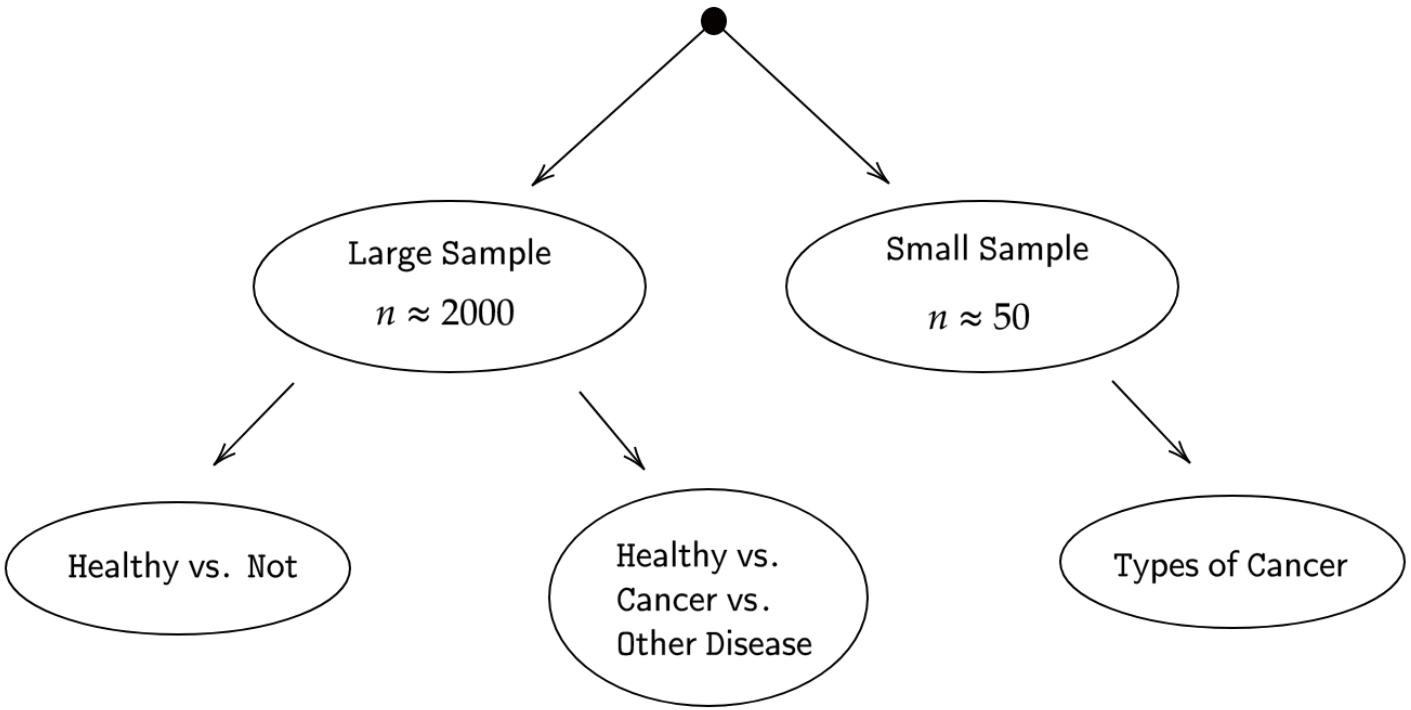
Shape of the data

- **Variety of labels:** 130 or more diseases recorded
- **Variety of features:** 22,000 or more measurements for each sample
- **Number of samples:** Around 6000 (sufficient)

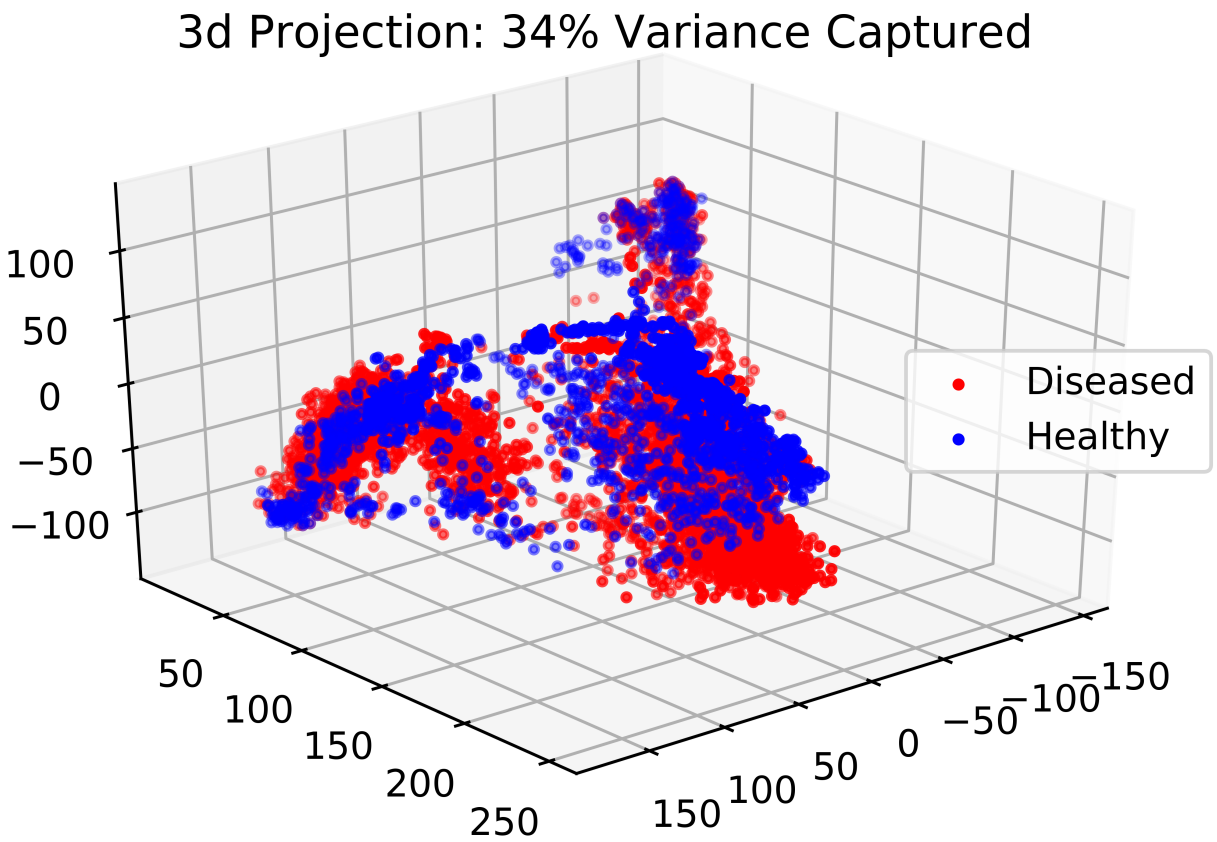
Challenges

- **Variety of labels:** too much heterogeneity
- **Variety of features:** high computational cost!
- **Number of samples:** for many illnesses, small sample number (bad)

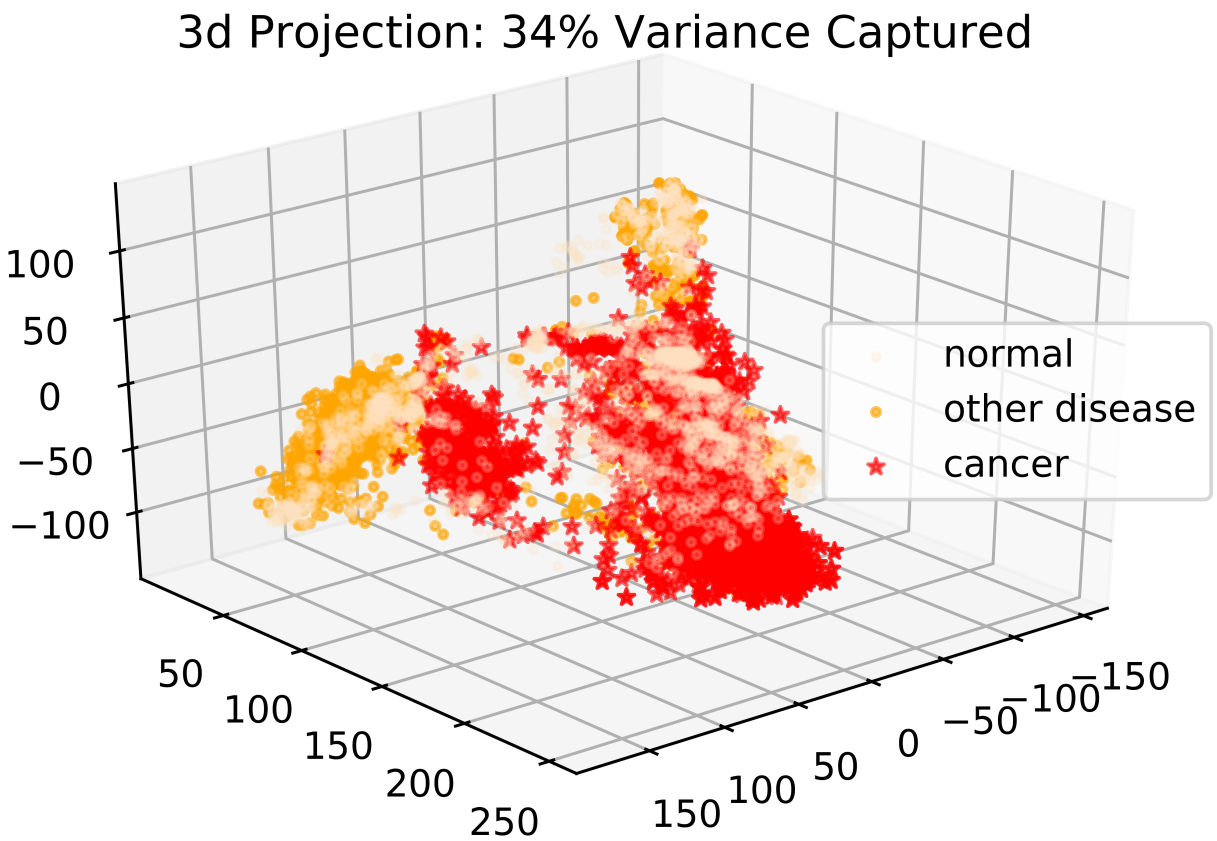
Exploring the Data



3D Representation: PCA



3D Representation: PCA



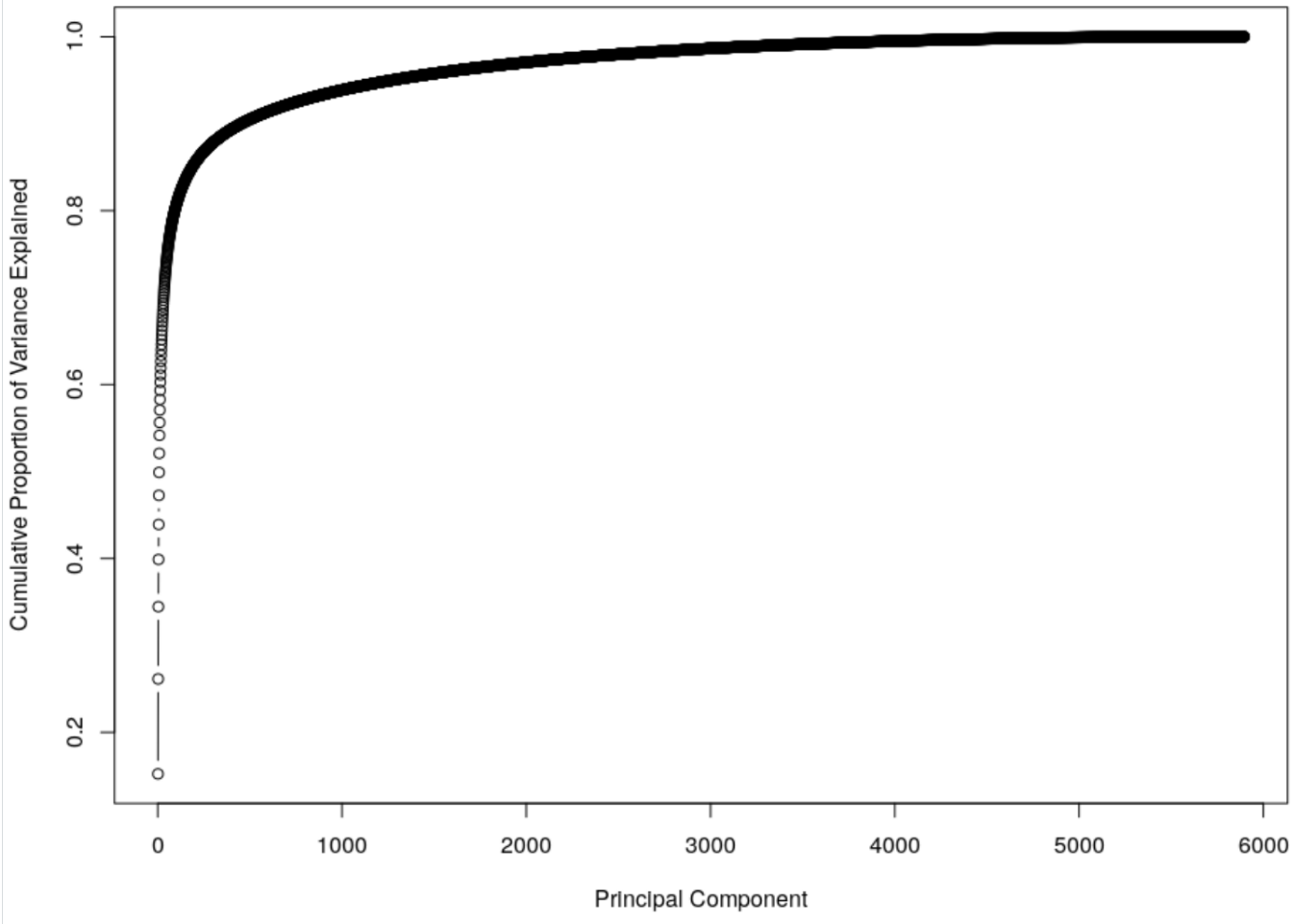
PCA is known to be unstable

Theorem: Asymptotic Instability of PCA.

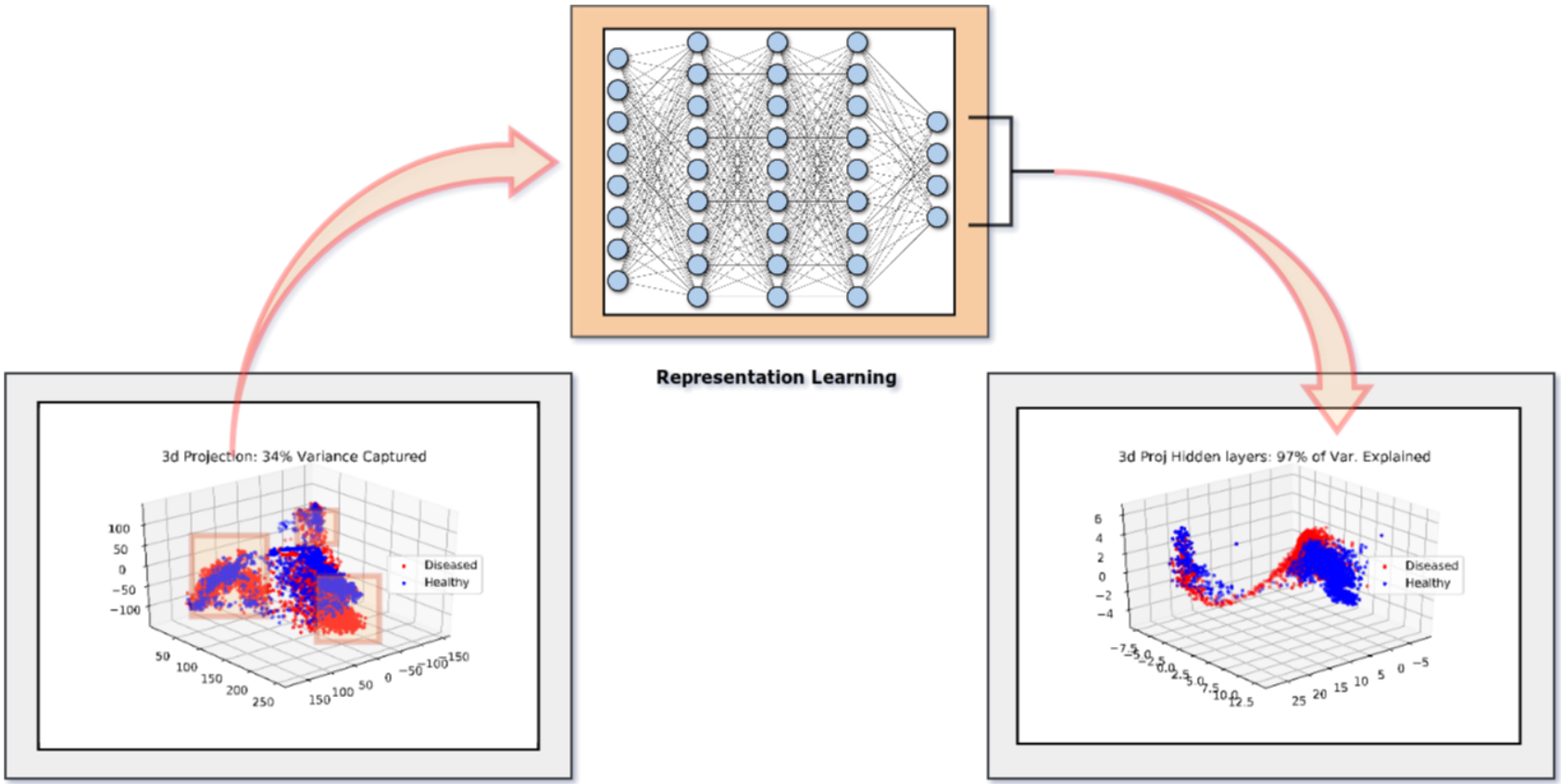
If (# features) $\gg$ (# samples), the sample's PCA doesn't approximate the population's PCA well.

Specifically, the first few components will capture too much variance.

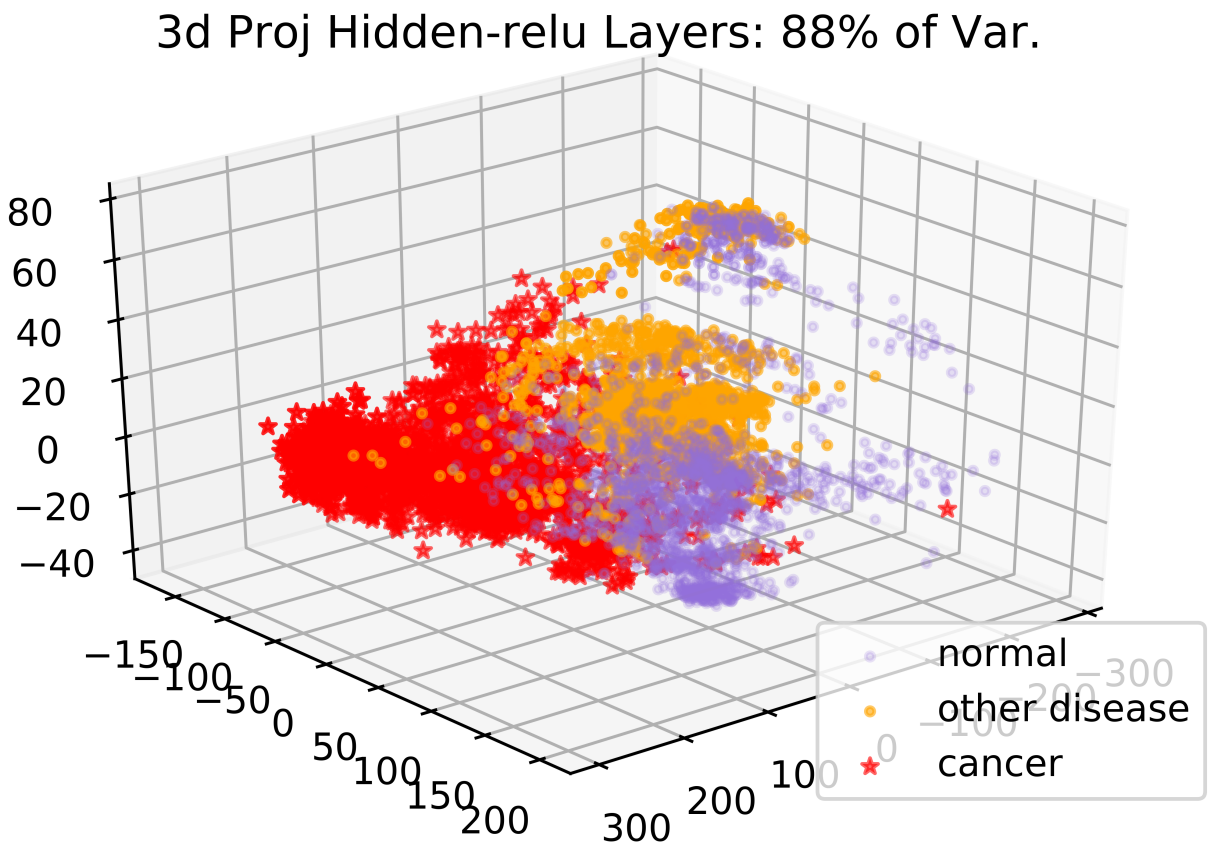
PCA is known to be unstable



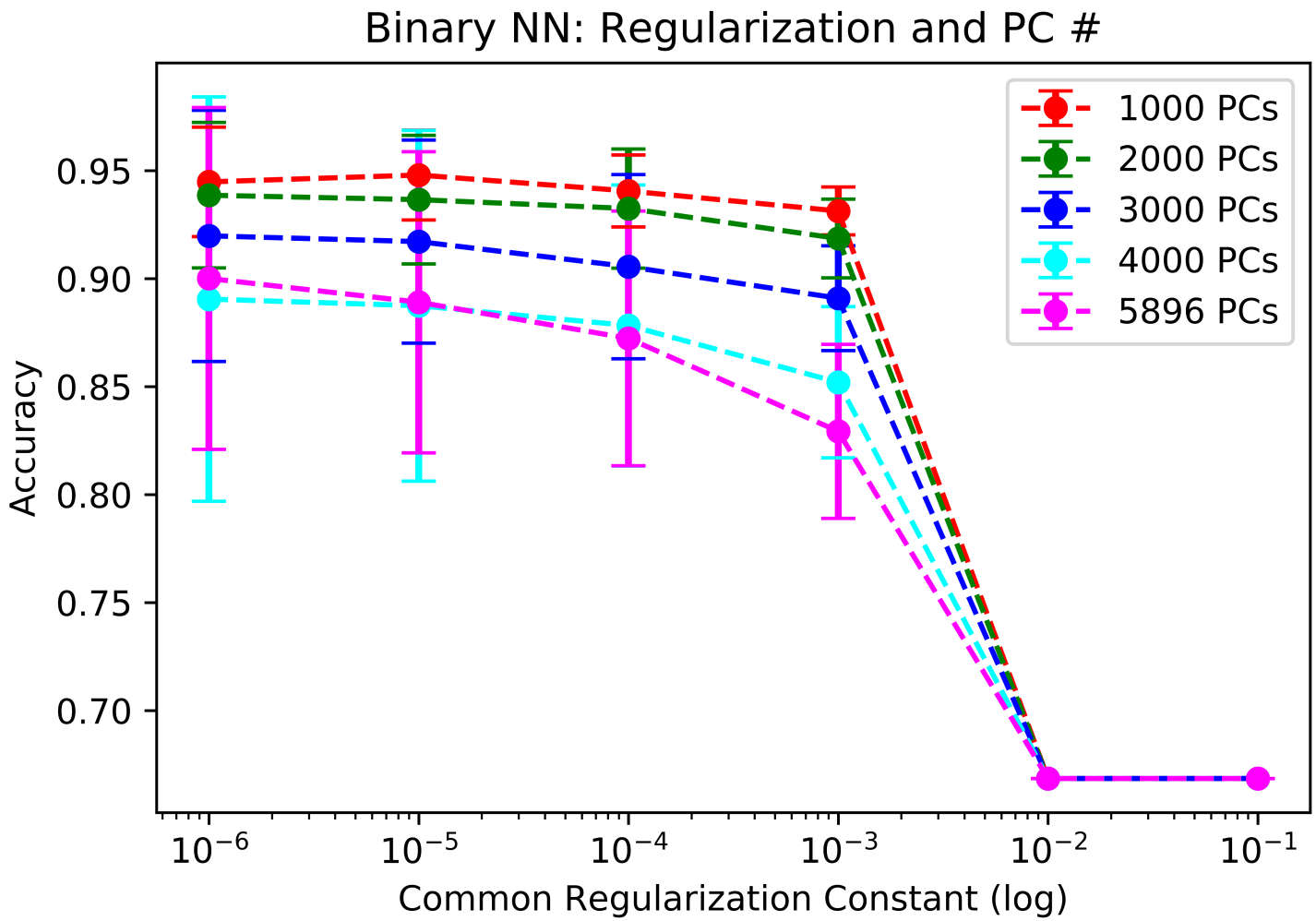
Learning better features



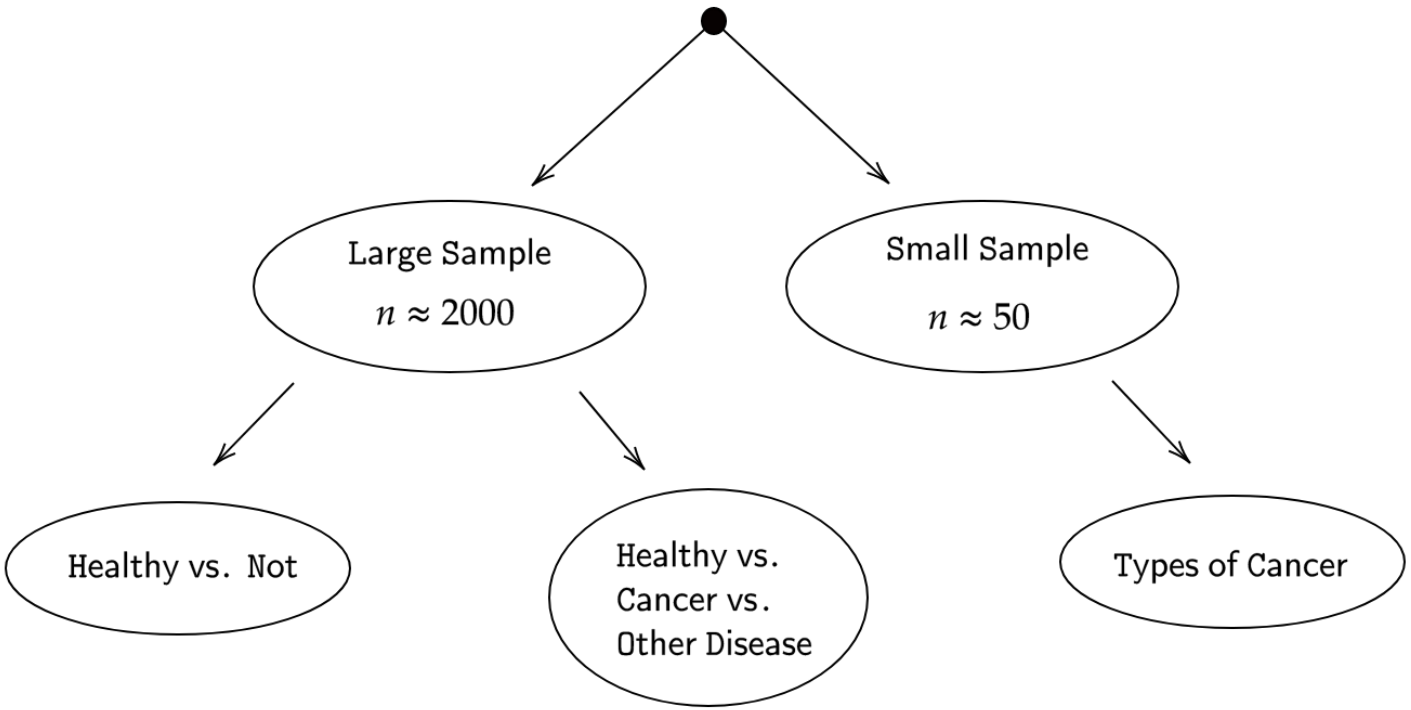
Learning better features



Performance



Exploring the Data



Why look at small samples?

Real world applications

- Initially we had a lot of valuable label information, such as whether the sample was taken from someone with a specific type of cancer.
- Use as **diagnostic tool**: often doctors need to understand what precise type of disease a patient suffers so they can prescribe the right treatment. While microarray analysis is expensive, getting screened for certain genes is not.
- **Interpretability**: Researchers are interested in better understanding gene expression, which can be derived from the cDNA fragments. However transforming into PC space hinders this.

Small sample learning

Partial Least Squares

How do we learn from $n = 50, p = 22,000$ data?

Traditional ML Techniques fail

- Neural Networks will overfit.
- Random Forests will overfit.
- Difficult to get good parameter estimates.
- High degree of collinearity.

Statistical Learning Theory Techniques

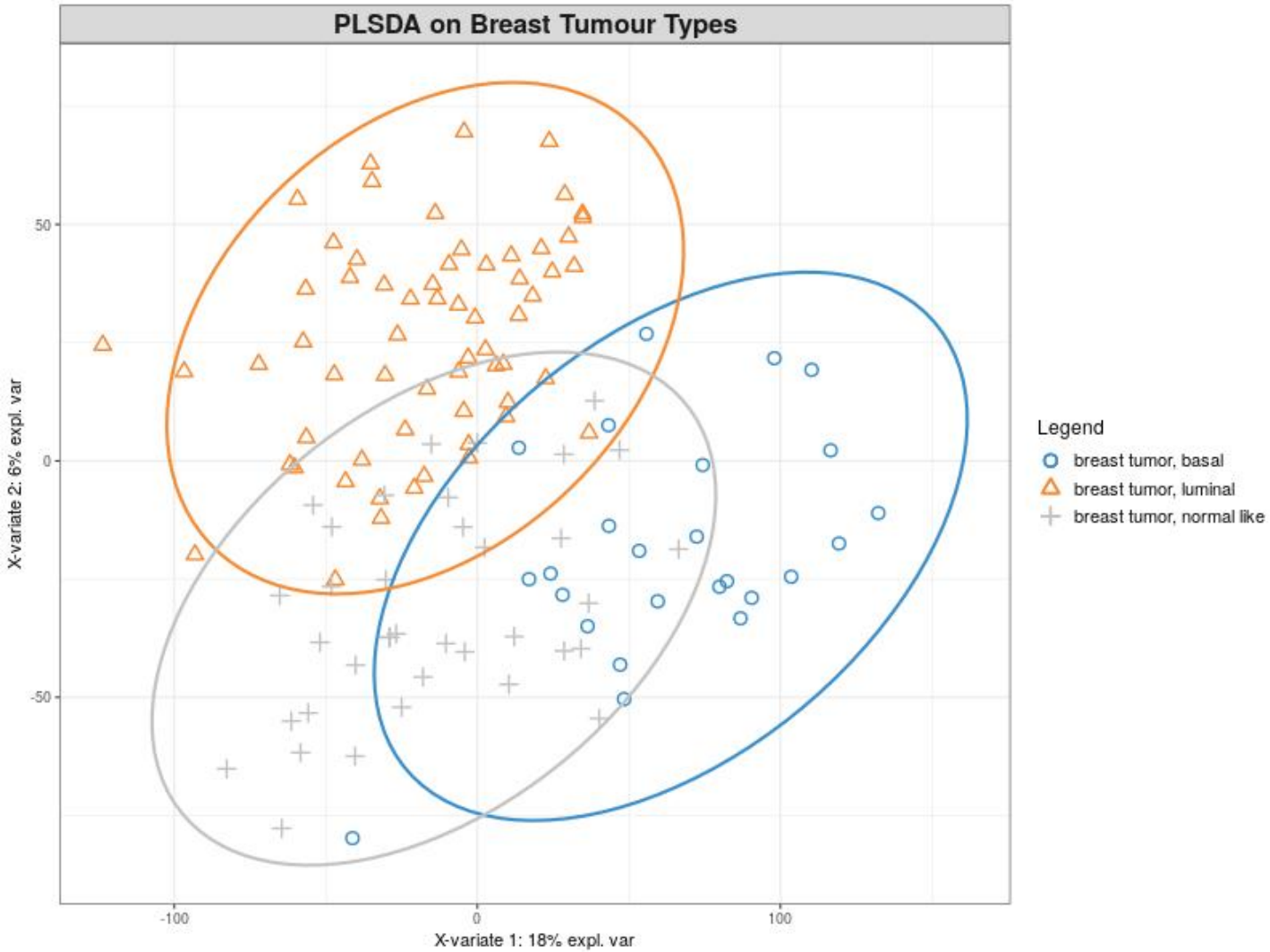
Partial Least Squares

PLS can be thought of as PCA's supervised sibling. Each variate is regressed on the labels, then the resulting vector is subtracted from the residuals of previous iterations. Algorithm is reminiscent of the Graham-Smith Process.

Some properties of PLSDA

- Stable in high dimensions.
- The projected feature space can be used to plot and interpret data.

Classifying Breast Cancer Types (PLSDA)

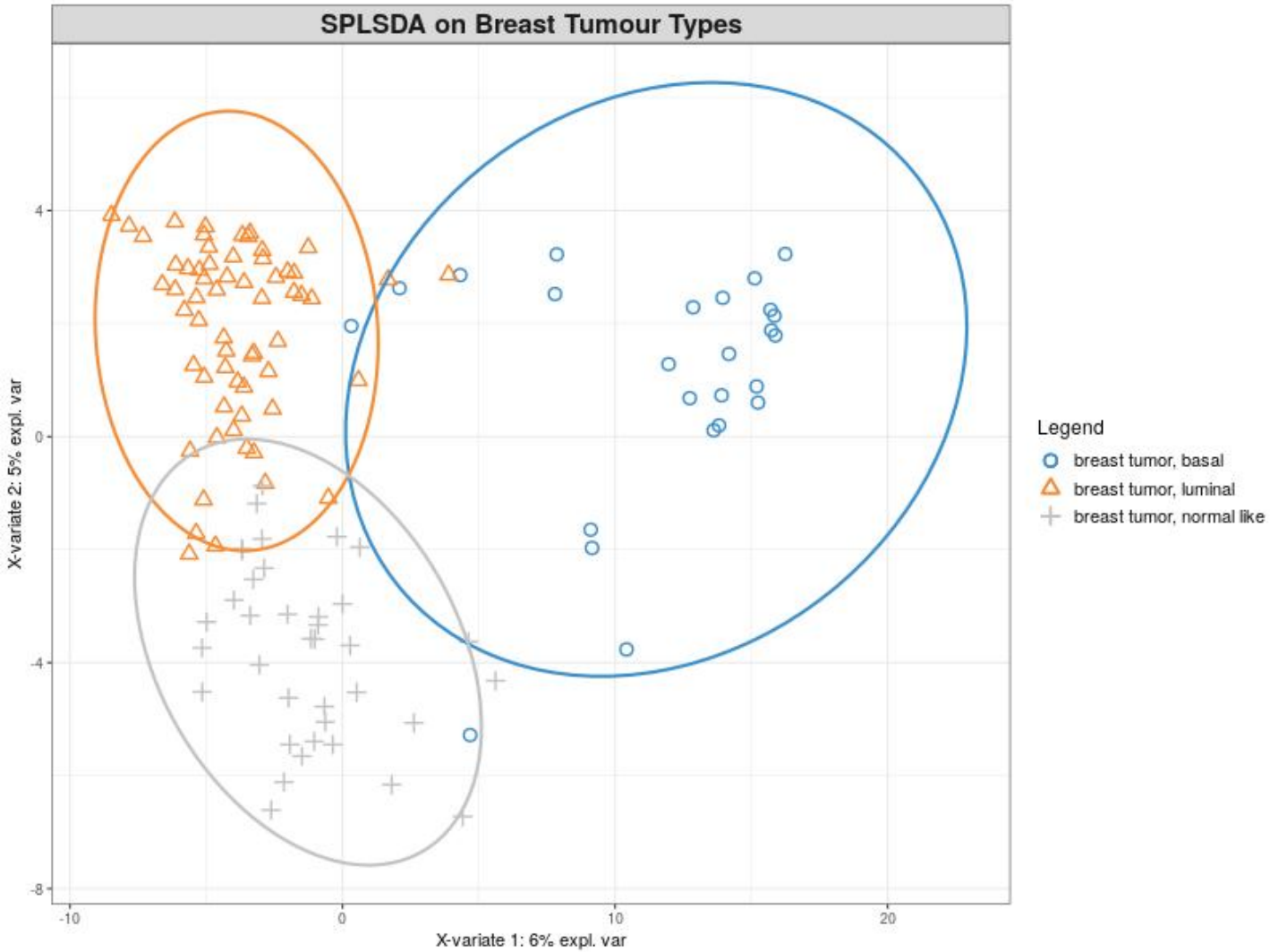


Interpretability of PLS

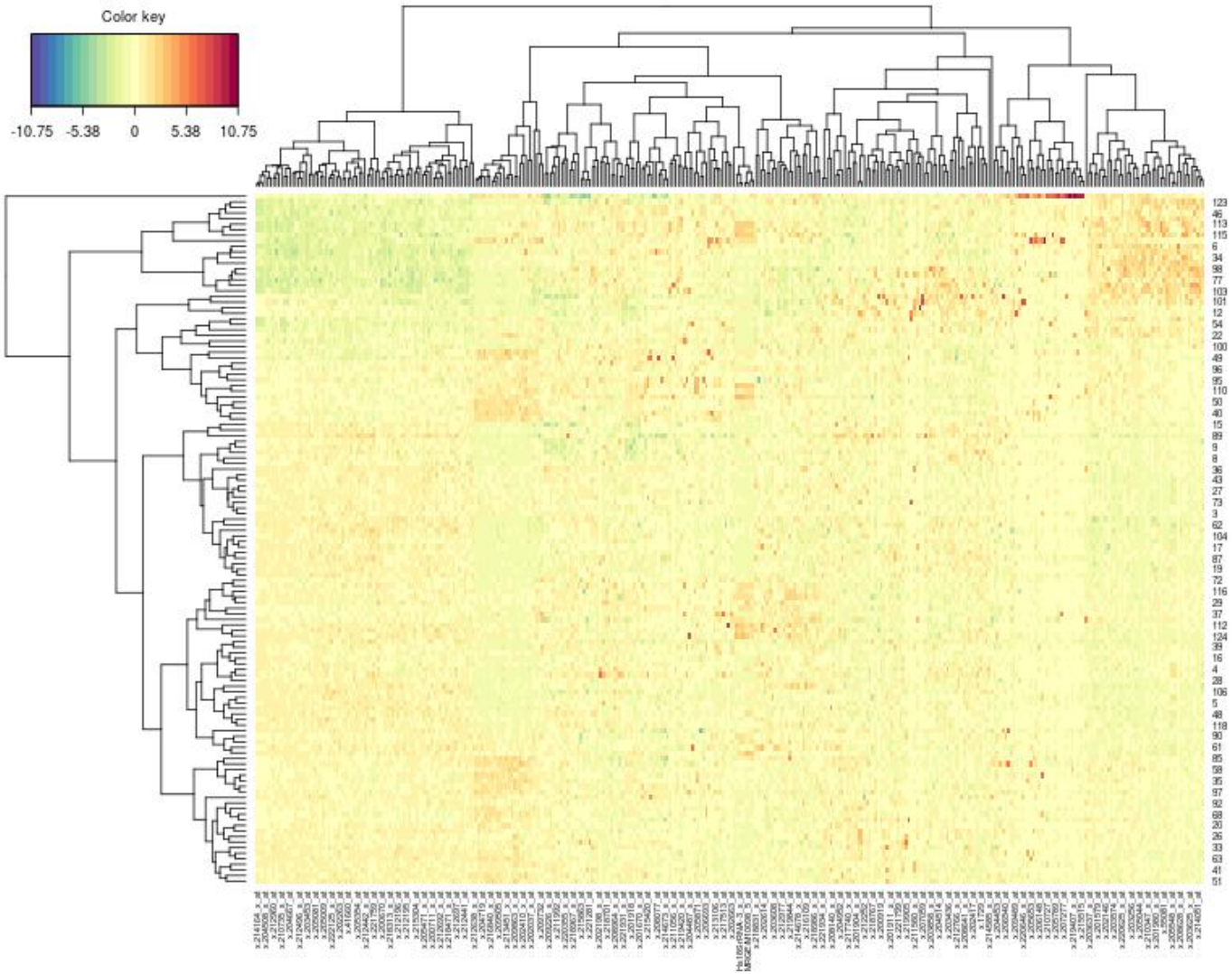
Problem with PLS: Interpretability

- PLS assigns a weight to every feature.
- However, we wish to know which clique of genes are responsible for the cell becoming cancer cells.
- Hence we need a sparsity constraint: we need the weight of most features to be 0.

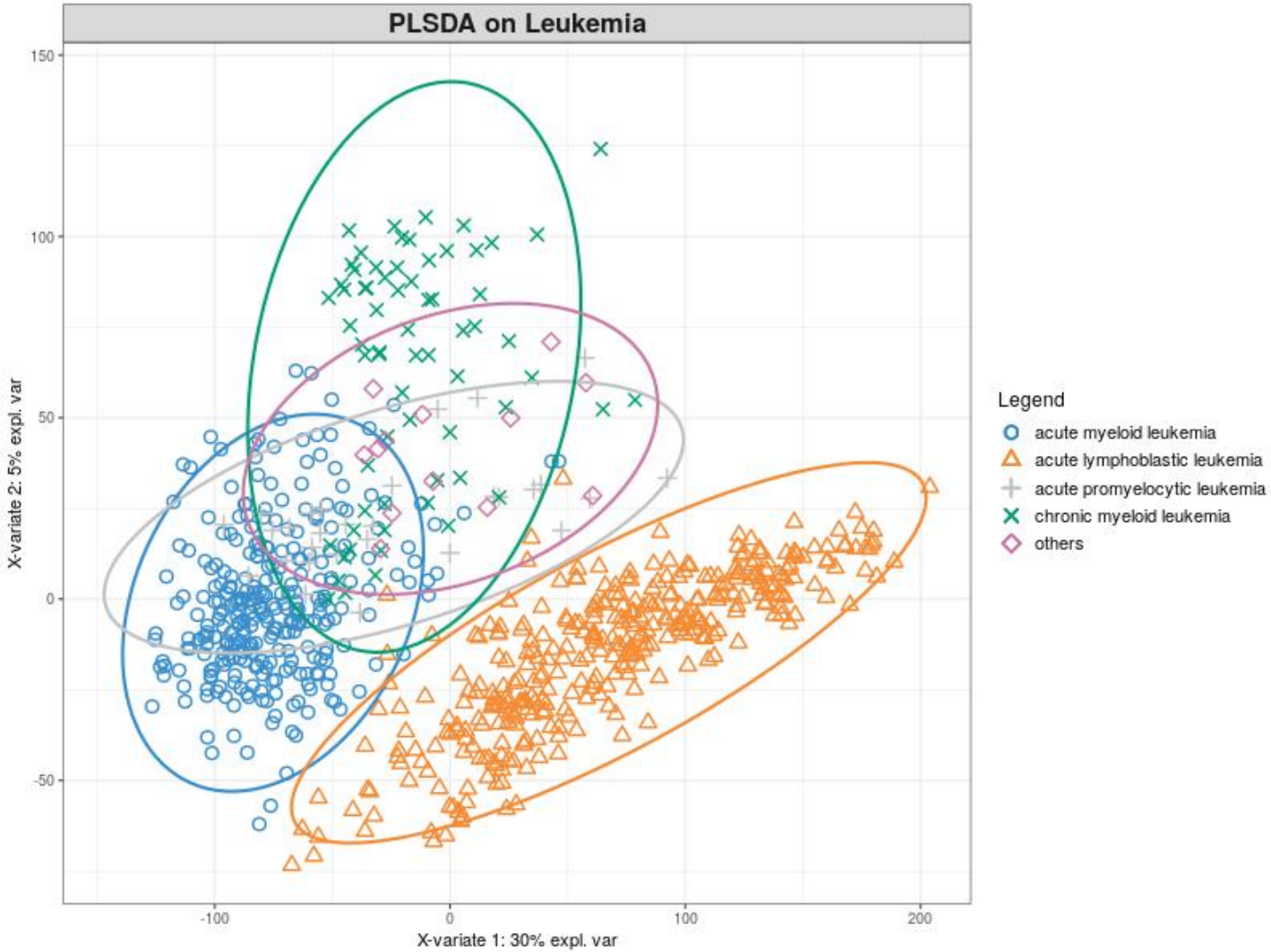
Classifying Breast Cancer Types (sPLSDA)



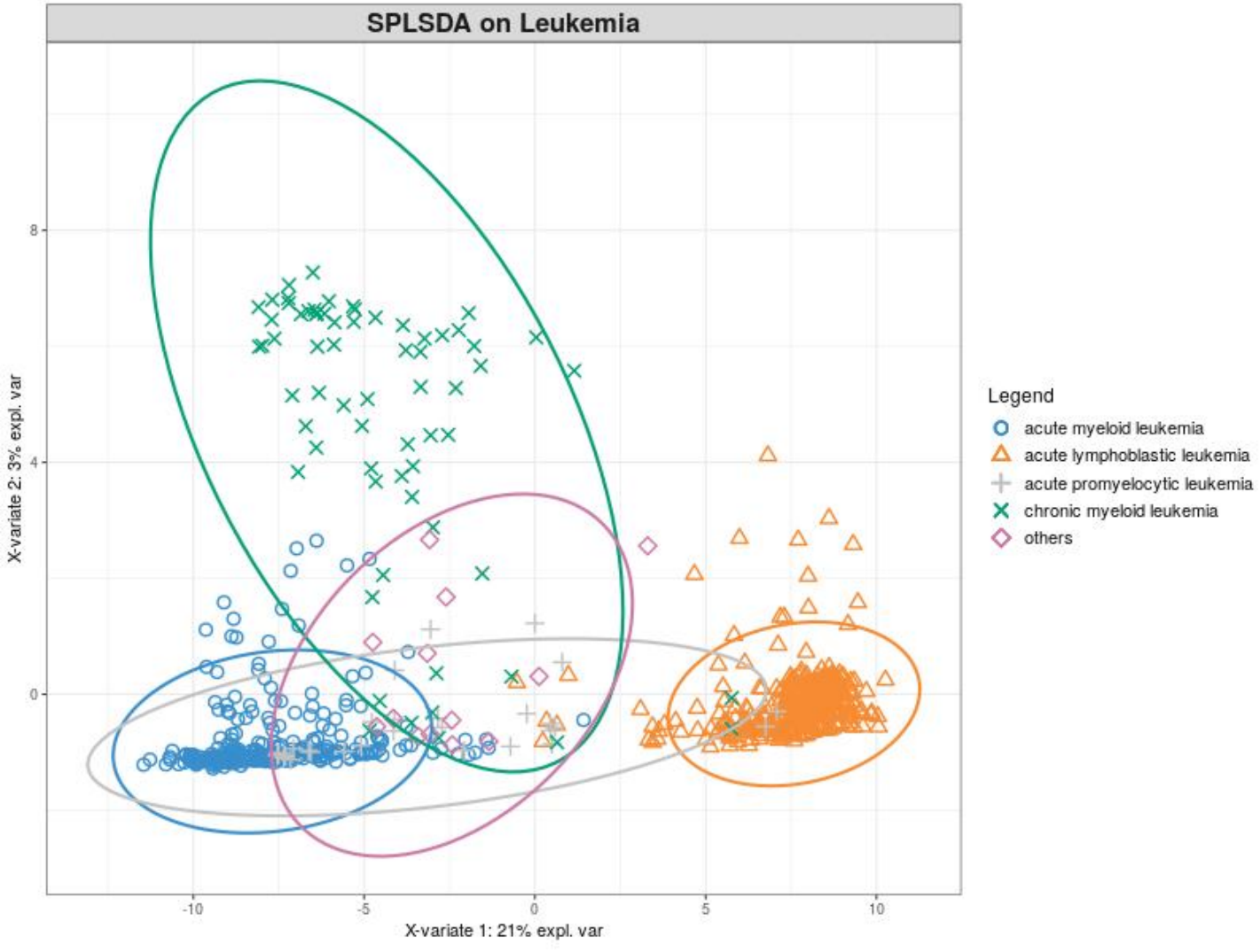
Classifying Breast Cancer Types (Clustered Image Map)



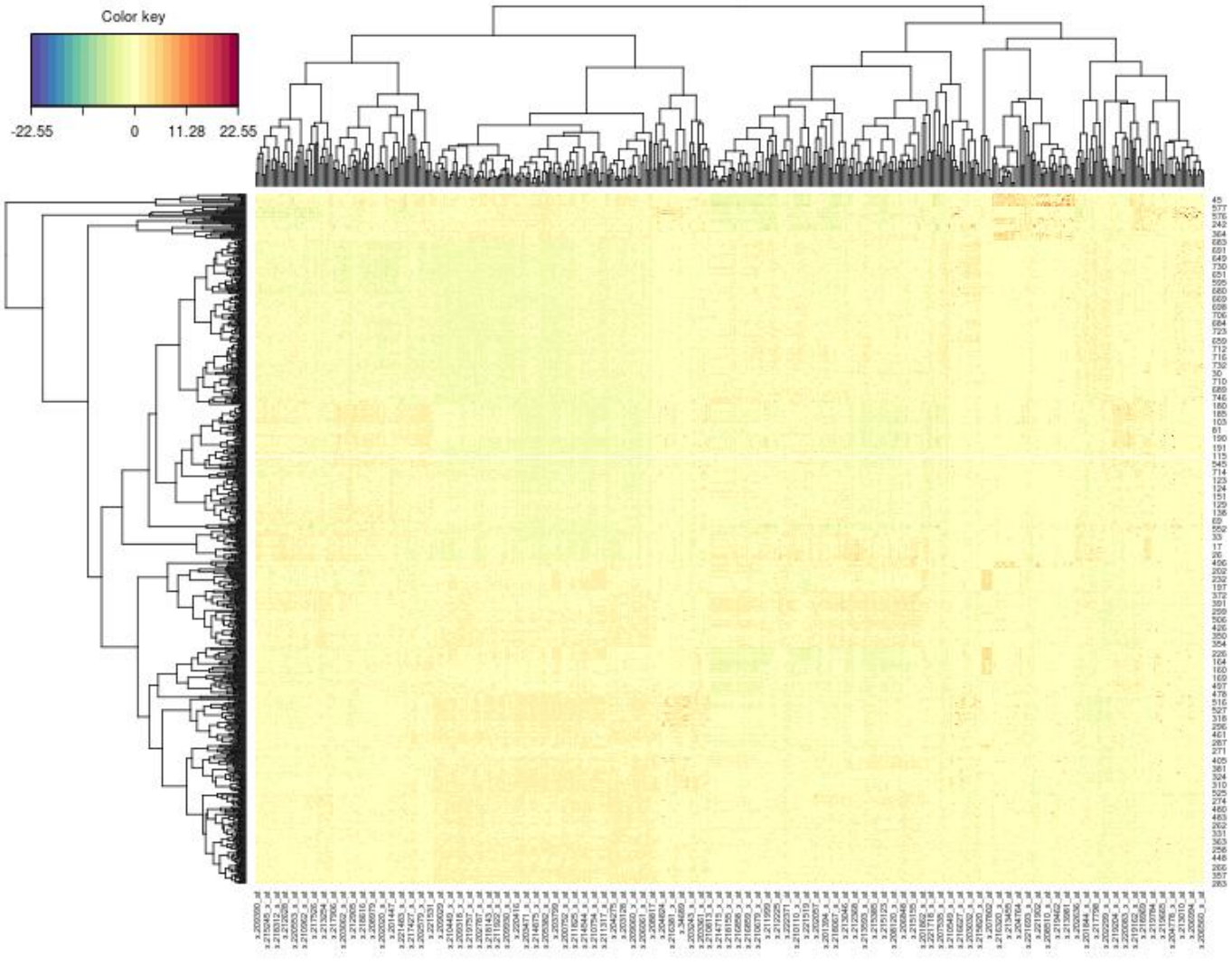
Classifying Leukemia Types (PLSDA)



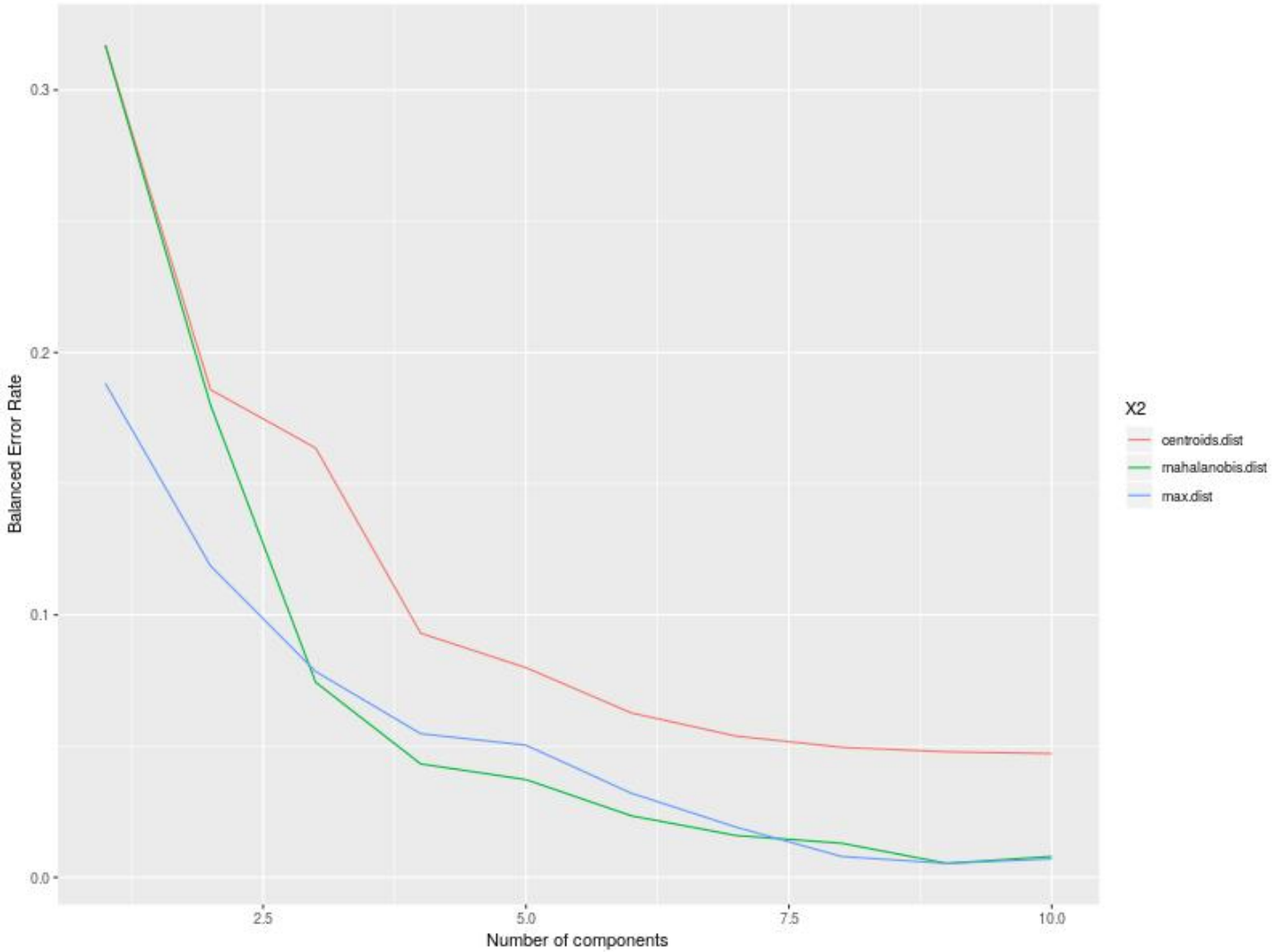
Classifying Leukemia Types (sPLSDA)



Classifying Leukemia Types (sPLSDA CIM)



Classifying Leukemia Types (PLSDA Performance)



Useful Links: Pictures

- **Picture 1:** ASM Journals , https://cmr.asm.org/content/22/4/611
- **Picture 2:** O'Reilly , https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html