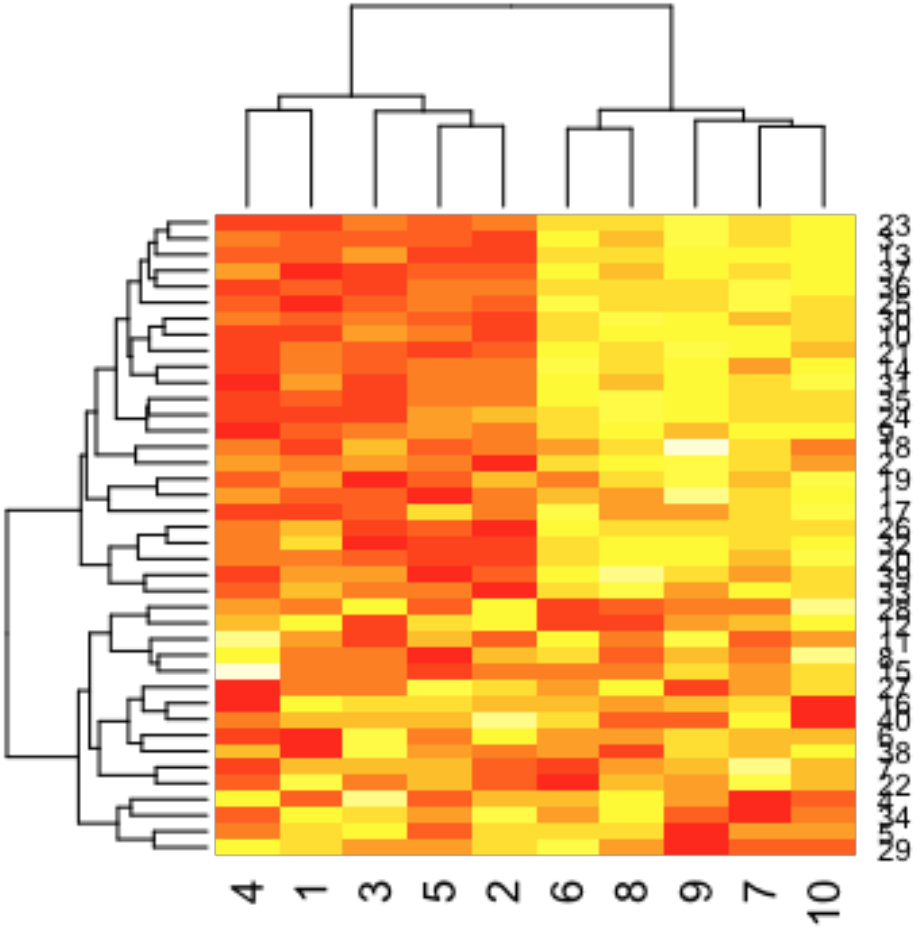
# Dimension reduction

Jeff Leek

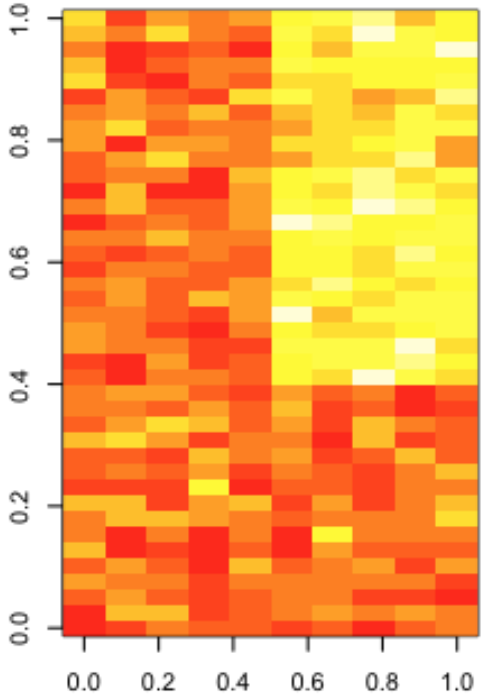
@jtleek

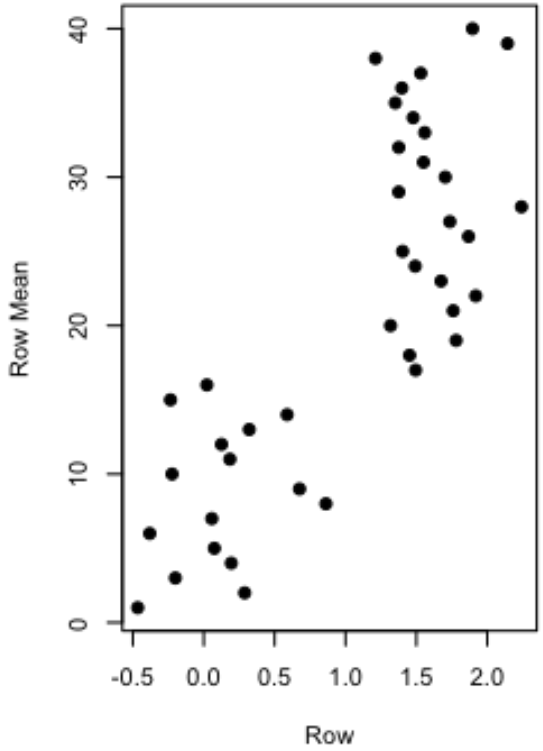
www.jtleek.com

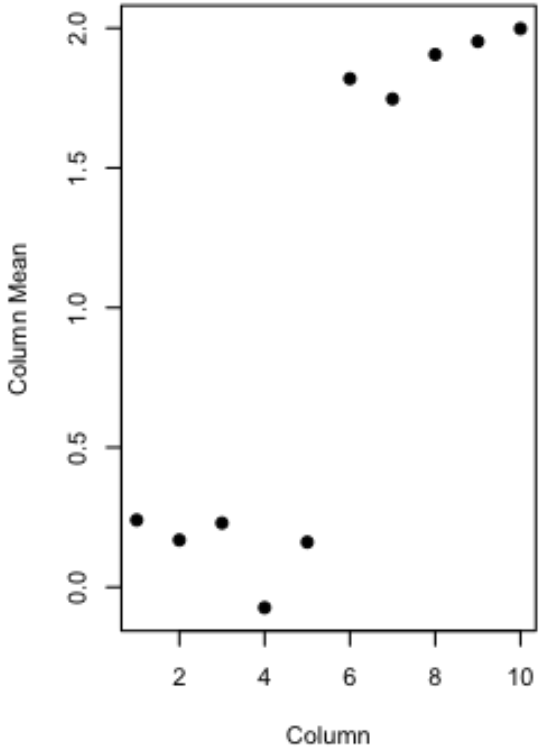
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction



https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction



# PCA and SVD

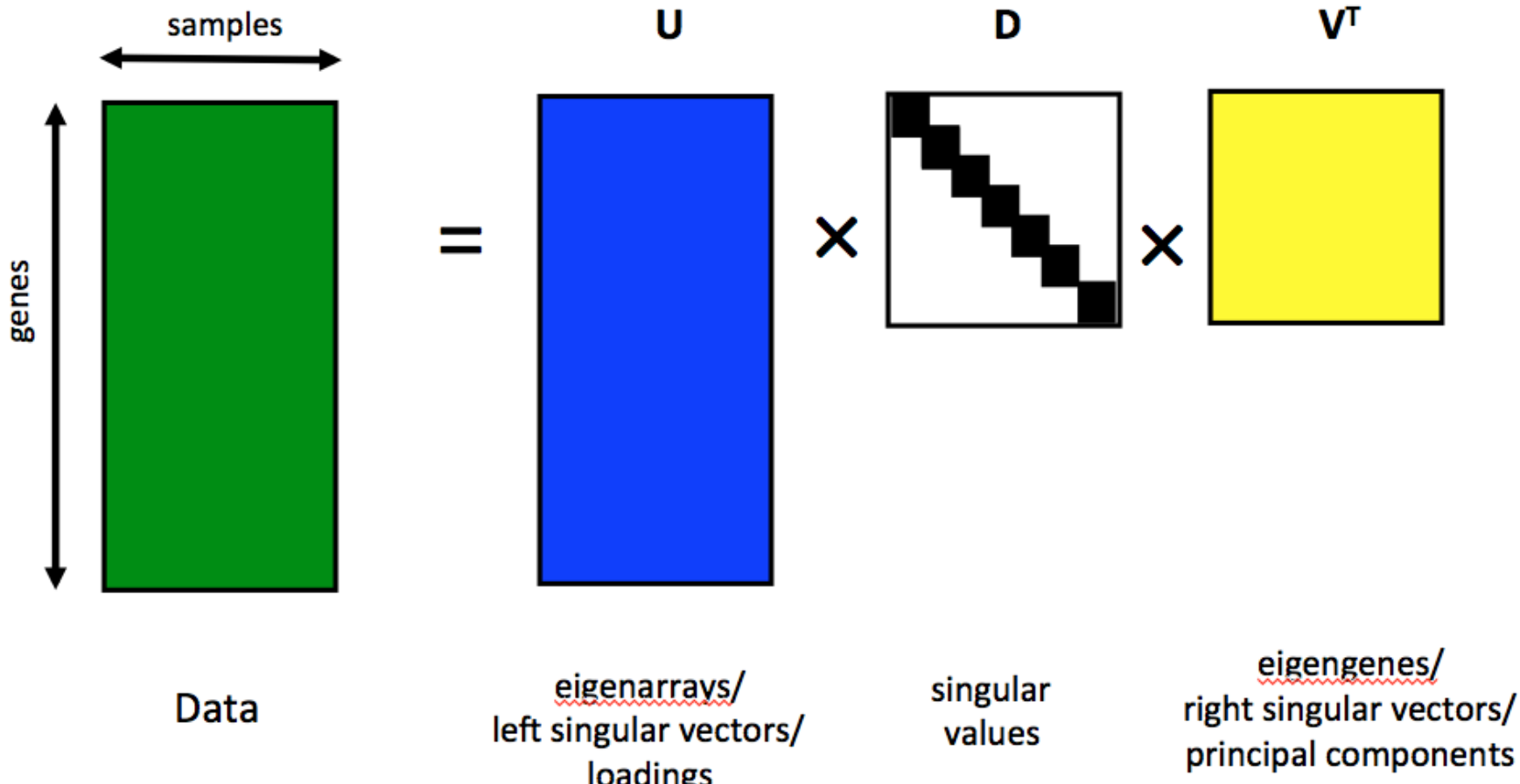
- PCA & SVD have different math goals
- SVD can be used to estimate PCs
- First proposed in genomics by Alter et al. 2000 PNAS

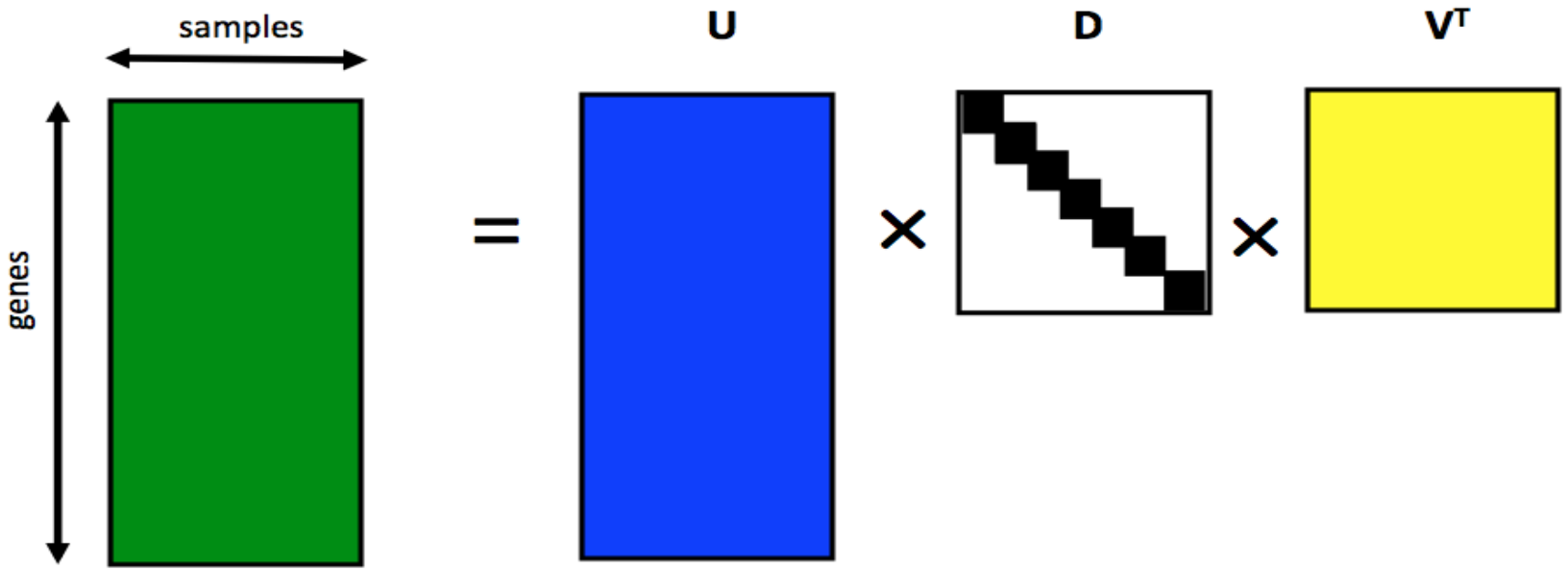
# Related Problems

You have multivariate matrix of data **X**

- Find a new set of multivariate variables that are uncorrelated and explain as much variance across rows as possible.
- Find the best matrix created with fewer variables (lower rank) that explains the original data.

The first goal is statistical and the second goal is data compression.


samples
genes
=
U
D
V^T
×
×
Data
eigenarrays/
left singular vectors/
loadings
singular
values
eigengenes/
right singular vectors/
principal components




Columns of $V^T$/rows of U are orthogonal and calculated one at a time

Columns of $V^T$ describe patterns across genes

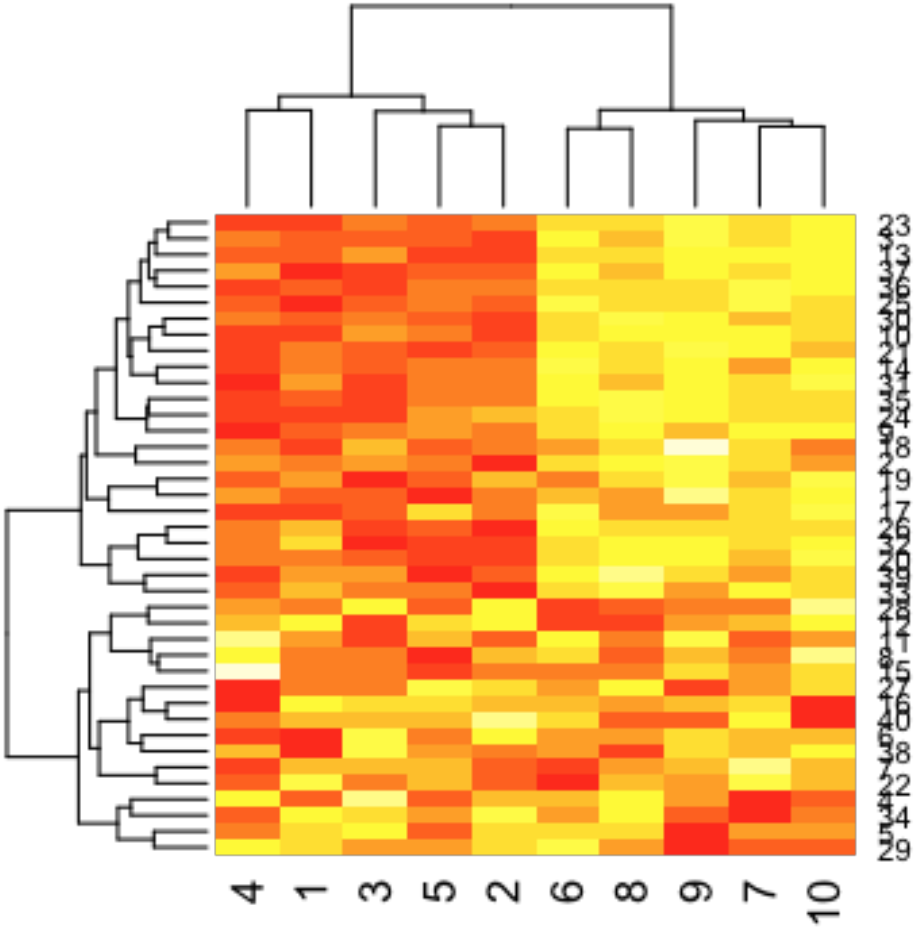
Columns of U describe patterns across arrays

$d_i^2 / \sum_{i=1}^{n} d_i^2$ is the percent of variation explained by the ith column of V

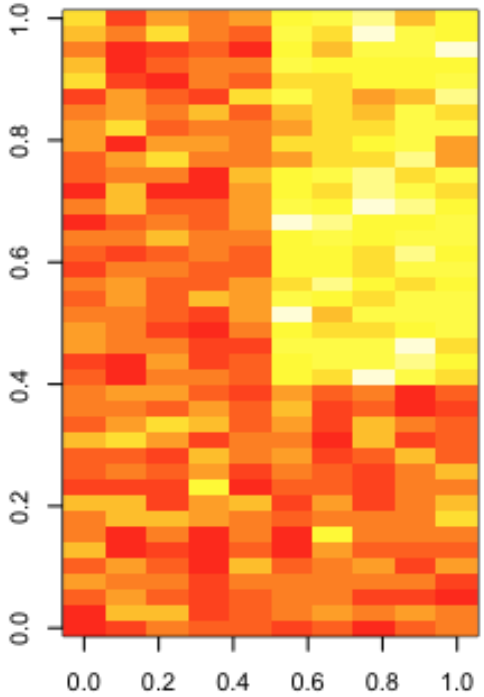
# Singular vectors/principal components

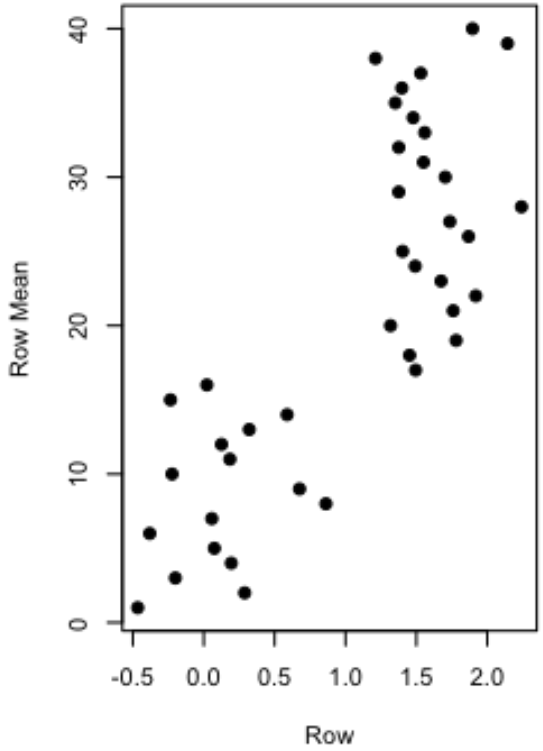
## Method to identify patterns in the data

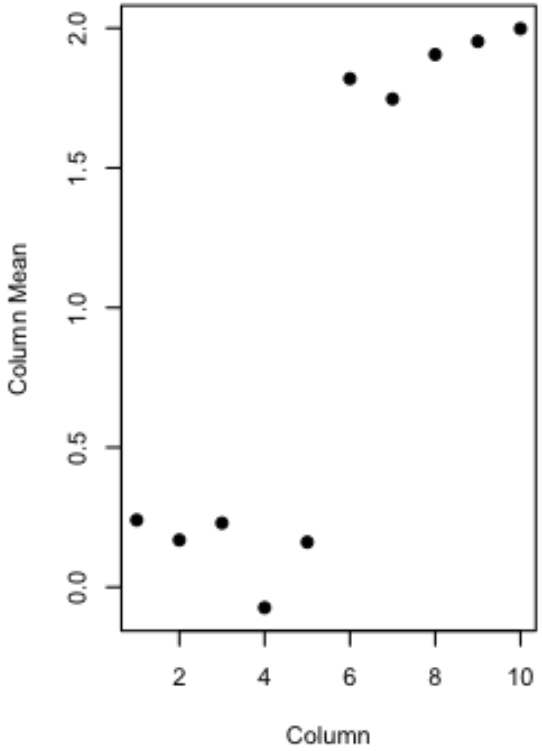
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction



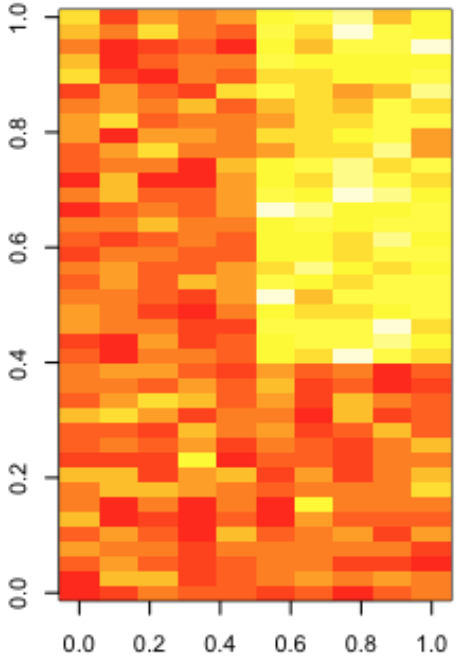
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction

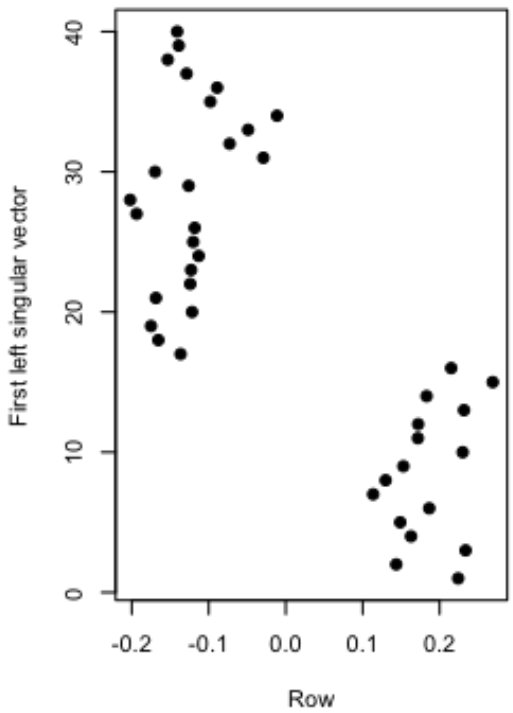


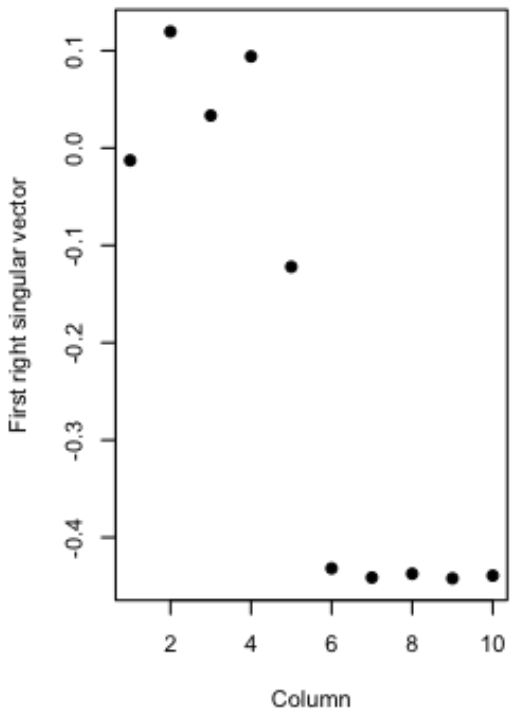



https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction



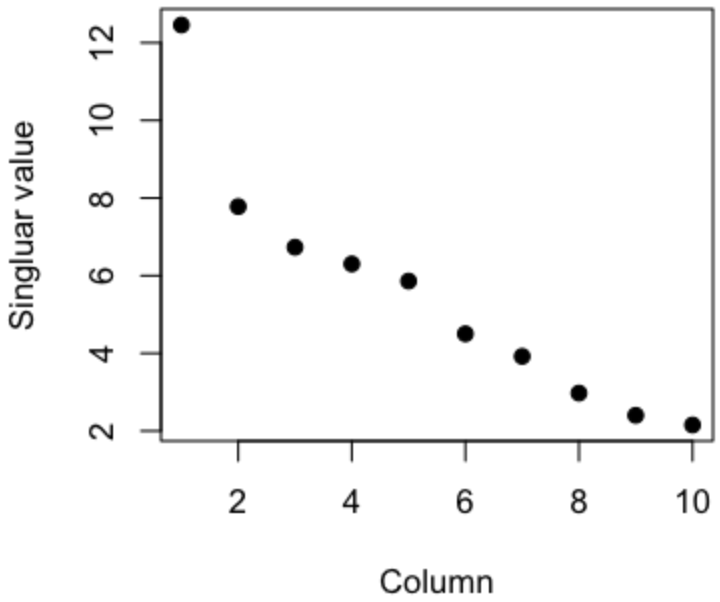
# Singular values

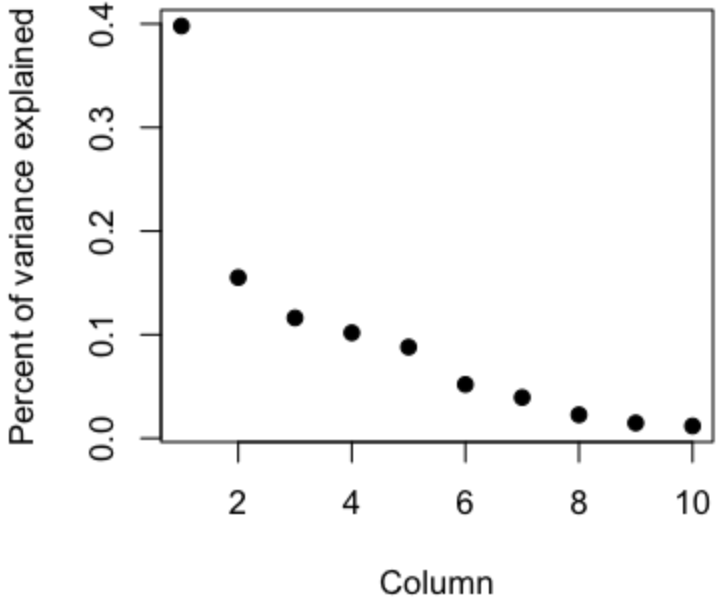
D is a diagonal matrix

$d_{ii}$ = ith singular value

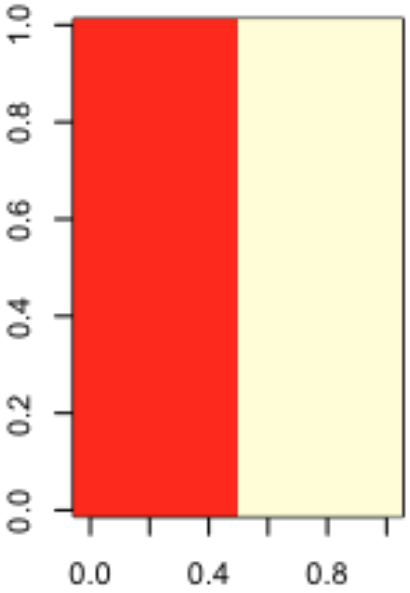
$d_{ii}^2 / \sum d_{jj}^2$ = percent variance explained by ith singular vectors

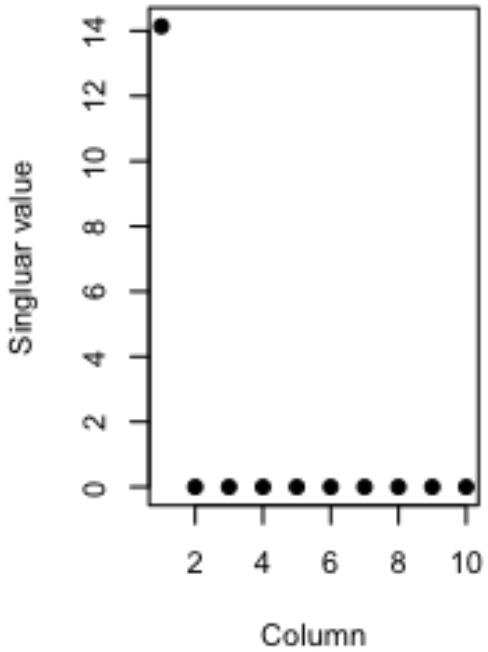
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction

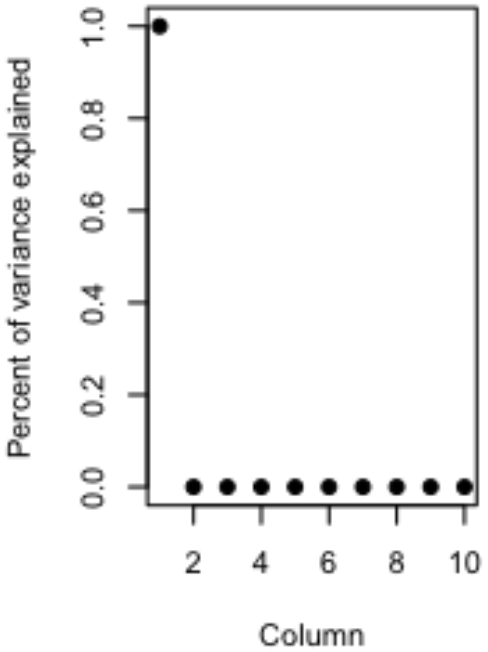


https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction



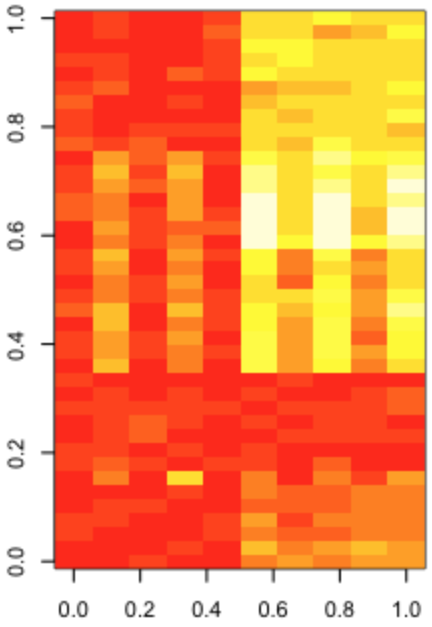



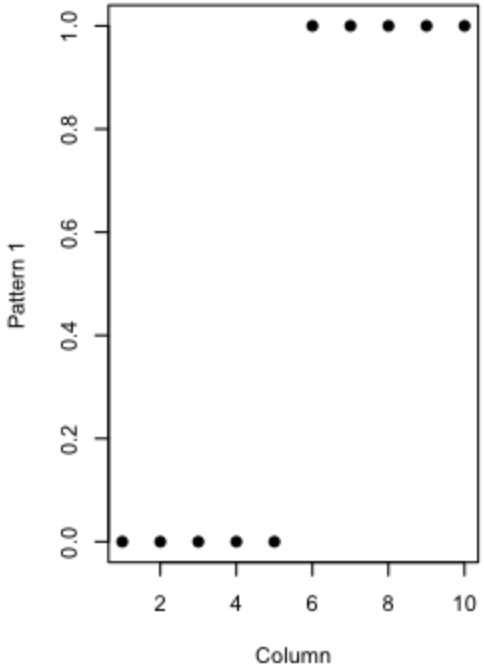



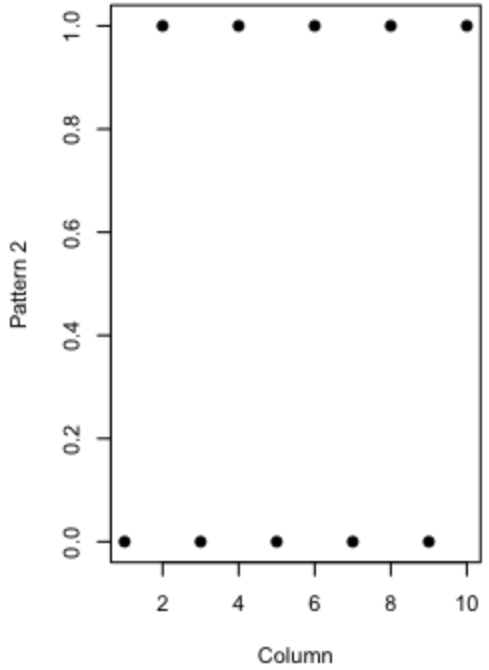
# More than one pattern

- Patterns are orthogonal
- One PC/SV may not equal one “variable”

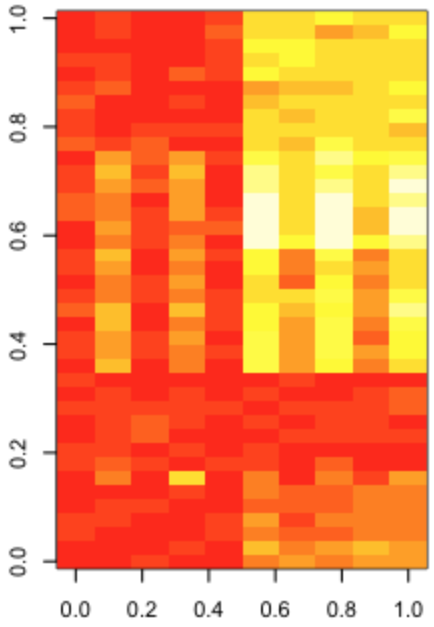
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction

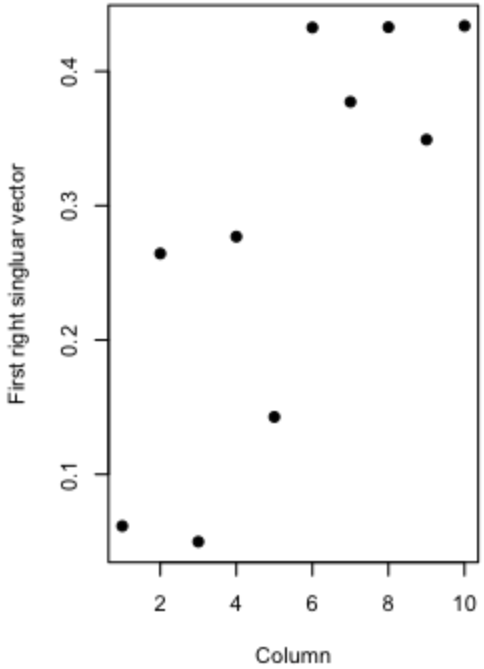


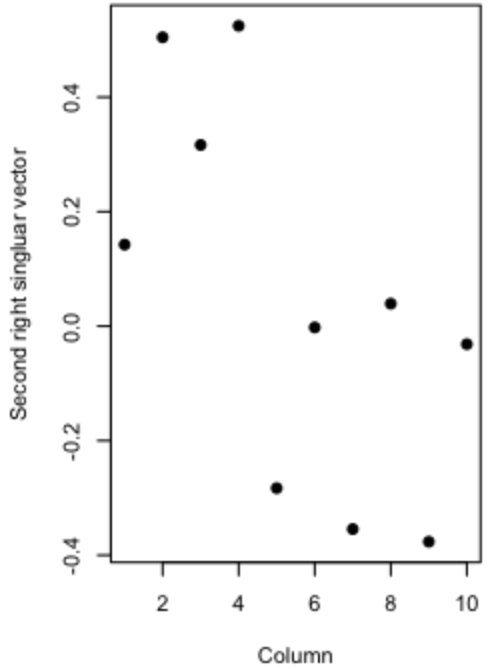



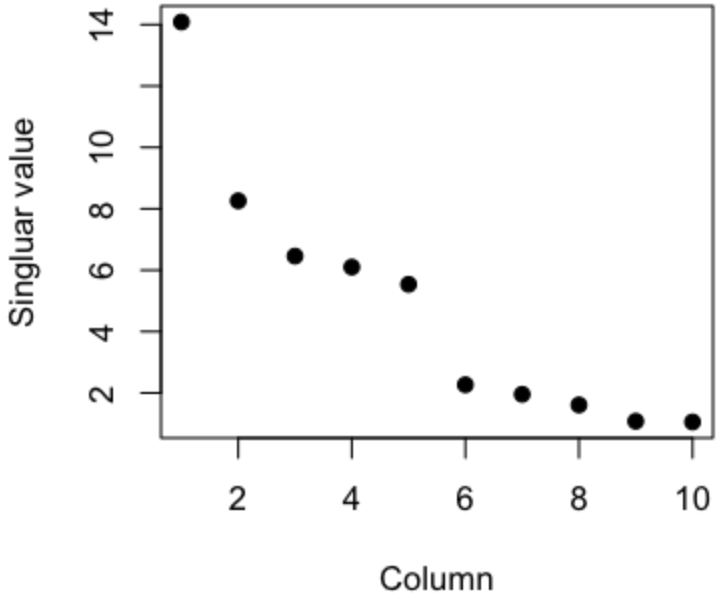
https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction

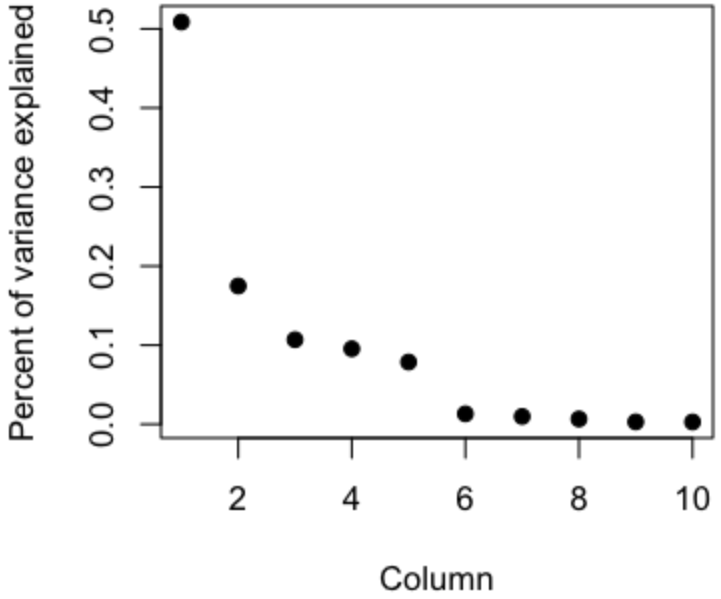


https://github.com/jtleek/dataanalysis/tree/master/week3/006dimensionReduction

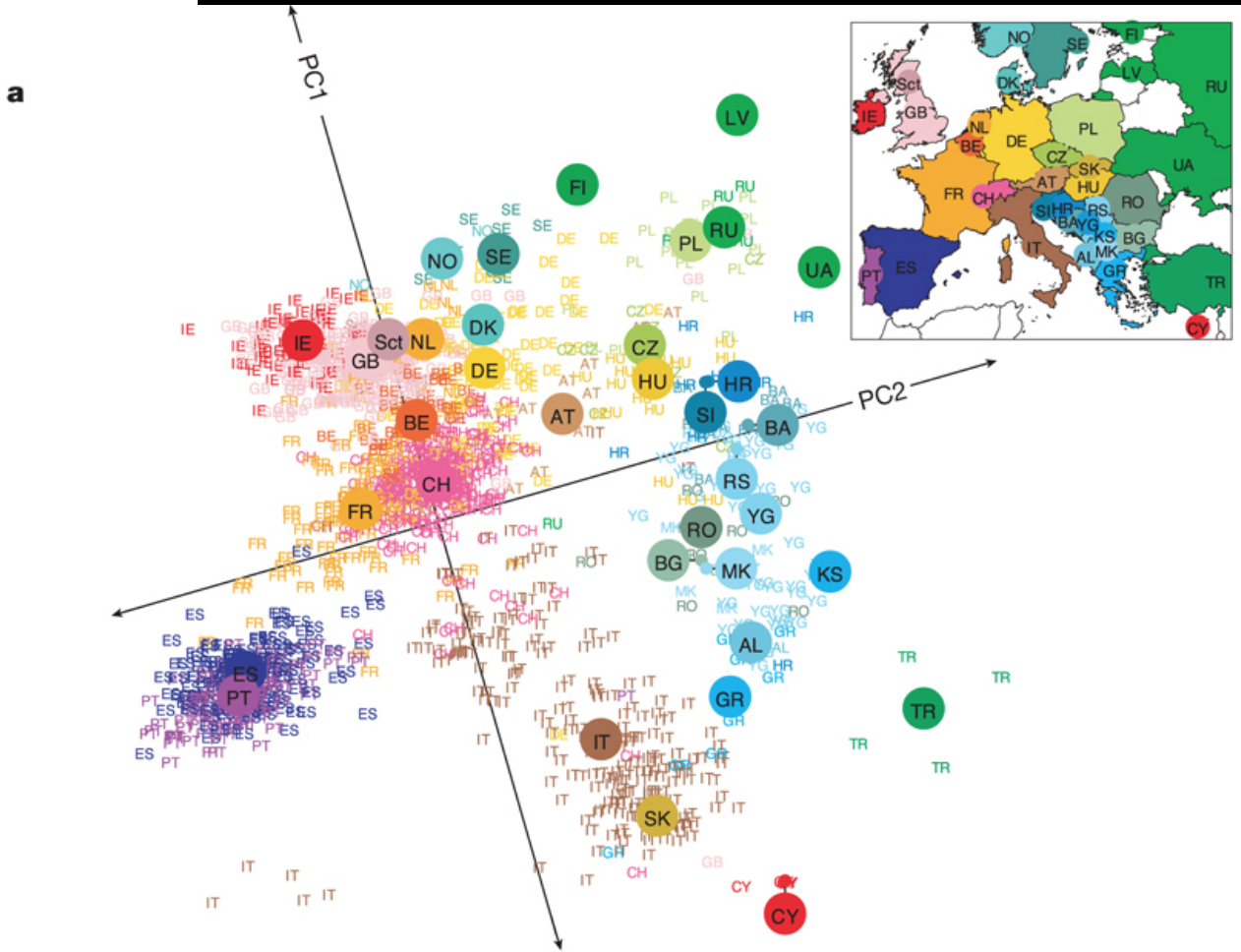


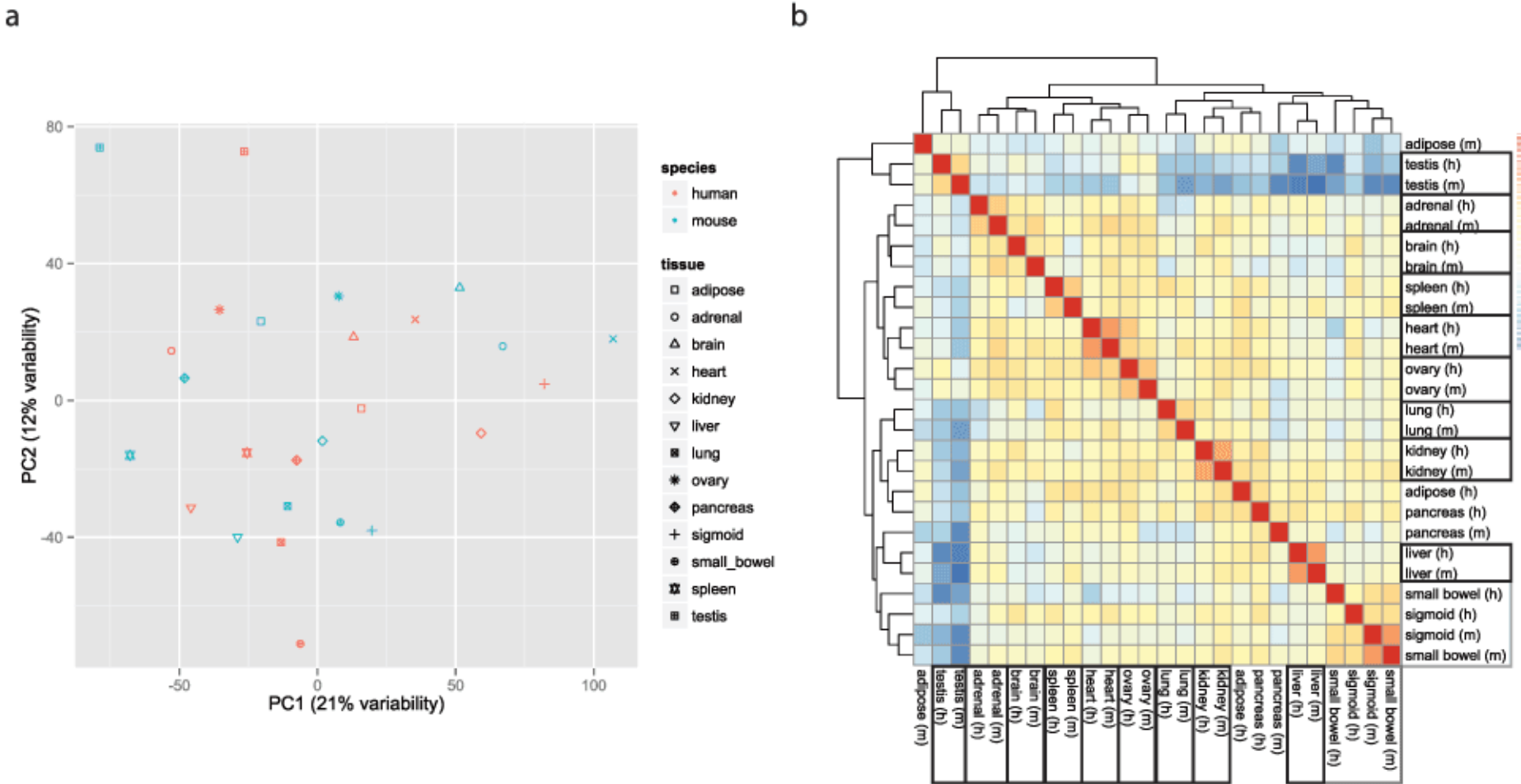



# How this is used

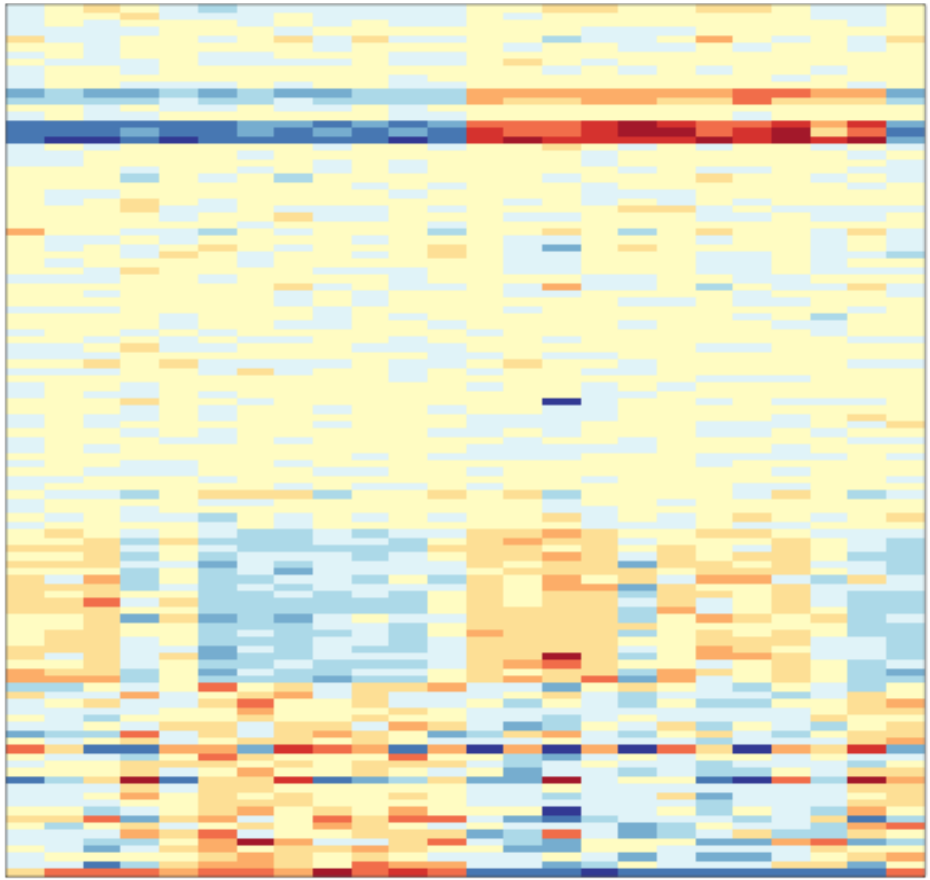
- Identify meaningful patterns
- Find batch effects

http://www.nature.com/nature/journal/v456/n7218/full/nature07331.html

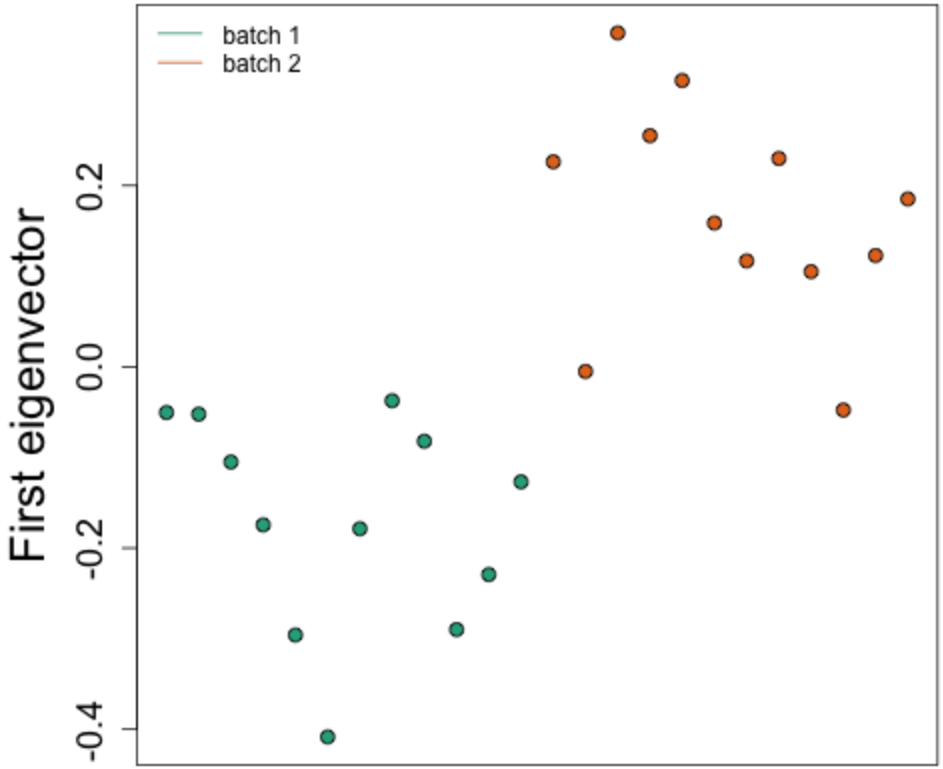



a
b
80
40
0
-40
-50
0
50
100
PC2 (12% variability)
PC1 (21% variability)
species
human
mouse
tissue
adipose
adrenal
brain
heart
kidney
liver
lung
ovary
pancreas
sigmoid
small_bowel
spleen
testis
adipose (m)
testis (h)
testis (m)
adrenal (h)
adrenal (m)
brain (h)
brain (m)
spleen (h)
spleen (m)
heart (h)
heart (m)
ovary (h)
ovary (m)
lung (h)
lung (m)
kidney (h)
kidney (m)
adipose (h)
pancreas (h)
pancreas (m)
liver (h)
liver (m)
small bowel (h)
sigmoid (h)
sigmoid (m)
small bowel (m)

http://genomicsclass.github.io/book/pages/svacombat.html



http://genomicsclass.github.io/book/pages/svacombat.html



# Notes and further resources

- Widely used for batch effects
  - http://www.nature.com/nrg/journal/v11/n10/full/nrg2825.html
- There are many more decompositions people use
  - multidimensional scaling, independent components analysis, non-negative matrix factorization
- More discussion in this course
  - https://www.edx.org/course/advanced-statistics-life-sciences-harvardx-ph525-3x