Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center

Statistical Methods for High Dimensional Data

Kwang-Youn Kim, PhD

Assoc. Professor

Dept of Preventive Medicine

kykim@northwestern.edu

All lectures will be held from 1 to 2 pm in Baldwin Auditorium, Robert H. Lurie Medical Research Center, 303 E. Superior St.

BCC: Biostatistics Collaboration Center

Who We Are

Leah J. Welty, PhD
Assoc. Professor
BCC Director

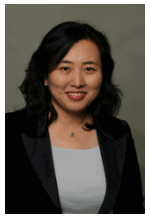
Lauren Balmert, PhD
Asst. Professor

Jody D. Ciolino, PhD
Asst. Professor

Kwang-Youn A. Kim, PhD
Assoc. Professor

Masha Kocherginsky, PhD
Assoc. Professor
QDSC Director

Mary J. Kwasny, ScD
Professor



Julia Lee, PhD, MPH
Assoc. Professor

David Aaby, MS
Biostatistician

Elizabeth Gray, MS
Senior Statistical Analyst

Liqi Chen, MS
Stat. Analyst

Nina Srdanovic, MS
Stat. Analyst

Chen (Jen) Yeh, MS
Stat. Analyst

Kate Zumpf, MS
Stat. Analyst

Ashley Knudson
Business Coordinator

Biostatistics Collaboration Center (BCC)

Mission: to support investigators in the conduct of high-quality, innovative health-related research by providing expertise in biostatistics, statistical programming, and data management.

How do we accomplish this?

1. Every Northwestern investigator is provided a **FREE** initial consultation of 1-2 hours, subsidized by **FSM Office for Research**. Thereafter:
   a) Grants
   b) Re-charge (Hourly) Rates
   c) Subscriptions/partnerships

2. Most grant writing (e.g. developing analysis plans, power/sample size calculations) is also supported by FSM at **no cost to the investigator**, with the goal of establishing successful collaborations.

BCC: Biostatistics Collaboration Center

What We Do

- Many areas of expertise, including:
  - Bayesian Methods
  - Big Data
  - Bioinformatics
  - Causal Inference
  - Clinical Trials
  - Database Design
  - Genomics
  - Longitudinal Data
  - Missing Data
  - Reproducibility
  - Survival Analysis

Many types of software, including:

BCC: Biostatistics Collaboration Center

Shared Statistical Resources

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Stanley Manne Children's Research Institute | Ann & Robert H. Lurie Children's Hospital of Chicago

Biostatistics Collaboration Center (BCC)

- Supports **non-cancer** research at NU
- Provides investigators an initial 1-2 hour consultation subsidized by the FSM Office of Research
- Grant, Hourly, Subscription

Shirley Ryan AbilityLab

Quantitative Data Sciences Core (QDSC)

- Supports all **cancer-related** research at NU
- Provides free support to all Cancer Center members subsidized by RHLCCC
- Grant

Biostatistics Research Core (BRC)

- Supports **Lurie Children's Hospital affiliates**
- Provides investigators statistical support subsidized by the **Stanley Manne Research Institute at Lurie Children's**
- Hourly

BCC: Biostatistics Collaboration Center

Shared Resources Contact Info

- Biostatistics Collaboration Center (BCC)
  - Website: http://www.feinberg.northwestern.edu/sites/bcc/index.html
  - Email: bcc@northwestern.edu
  - Phone: 312.503.2288

- Quantitative Data Sciences Core (QDSC)
  - Website: http://cancer.northwestern.edu/research/shared_resources/quantitative_data_sciences/index.cfm
  - Email: qdsc_rhlccc@northwestern.edu
  - Phone: 312.503.2288

- Biostatistics Research Core (BRC)
  - Website: https://www.luriechildrens.org/en-us/research/facilities/Pages/biostatistics.aspx
  - Email: mereed@luriechildrens.org
  - Phone: 773.755.6328

Statistical Methods High Dimensional Data

What is High Dimensional Data?

- Most data can be represented in two dimensions
  - Variables in the columns
  - Observations (or samples) in the rows
- High dimensional data is a case where p >> n

Variables

"wide" format

Observations

Sample	Var 1	Var 2	Var 3	Var 4	…	Var p
1	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1p}
2	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2p}
…						
n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	…	x_{1p}

Note: "Sample" in the first row is also a variable name

Why should we be concerned about high dimensionality

- Isn't high dimension a good thing? More data is better, right?
  - More data refers to samples, not variables
    - Large *p* is not the same as large *n*
    - In fact *p* >> *n* is called the curse of dimensionality problem
  - Most statistical theorems are based on large *n* (not large *p*)
  - High dimensionality gives rise to overfitting and high variance
  - Independent samples give you the maximum information
    - Imagine if measuring heights of 200 pairs of identical twins
  - p >> N often occurs in genomics such as measuring gene expression in the whole blood

Mild Introduction to Statistics

- Learn about the methods to handle high dimensional data
  - Dimension reduction techniques
  - Learn machine learning techniques: unsupervised, supervised

Terminology

- **Sample**: An object we have data for (e.g. a study participant)
- **Feature**: A variable measured in our sample (e.g. gene expression for gene A)
- **Class**: A characteristic of the sample that is not a feature (e.g. x_{ij} = 0 or 1; death status)
- **Machine learning**: A broad category of techniques devoted to pattern recognition

Sample	Var 1	Var 2	Var 3	Var 4	…	Var p
1	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1p}
2	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2p}
…						
n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	…	x_{1p}

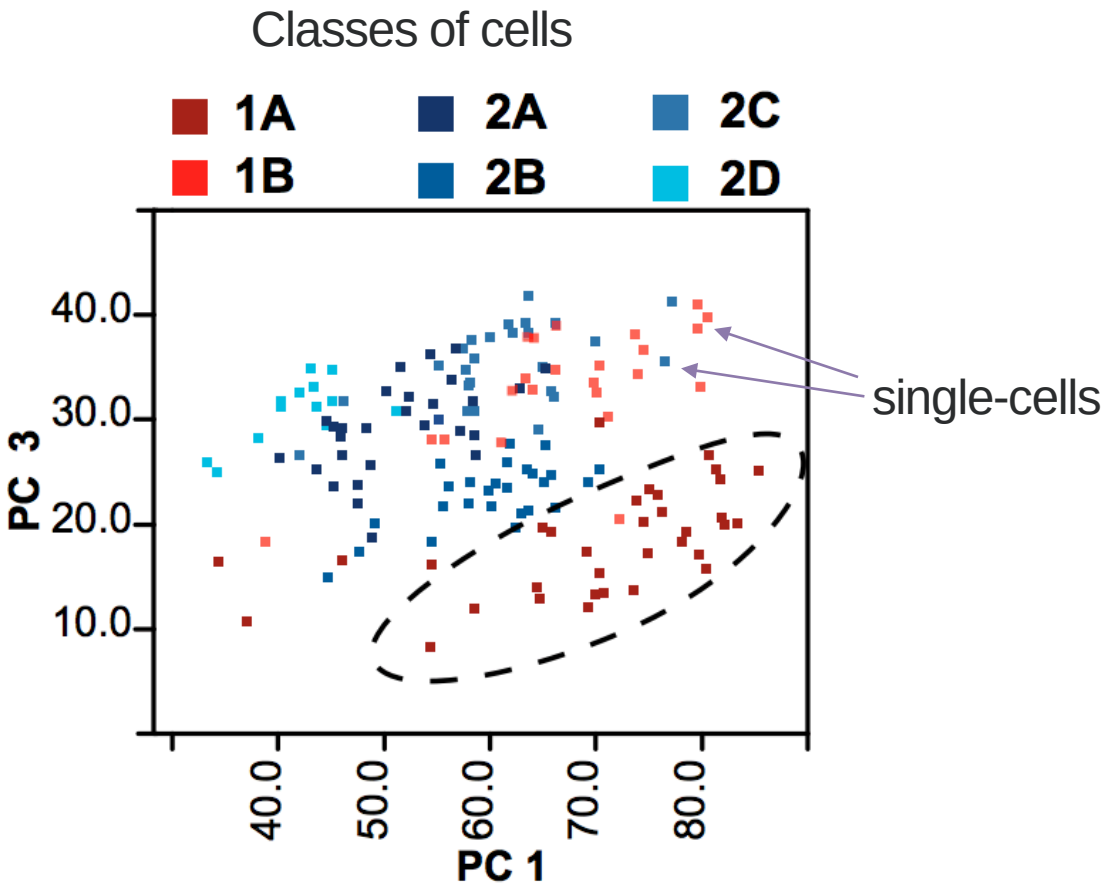
Northwestern Medicine®
Feinberg School of Medicine

Dimension Reduction Methods

Dimension Reduction Methods

- Principal Components Analysis
- Shrinkage Methods
  - Ridge regression
  - LASSO
  - Elastic Net

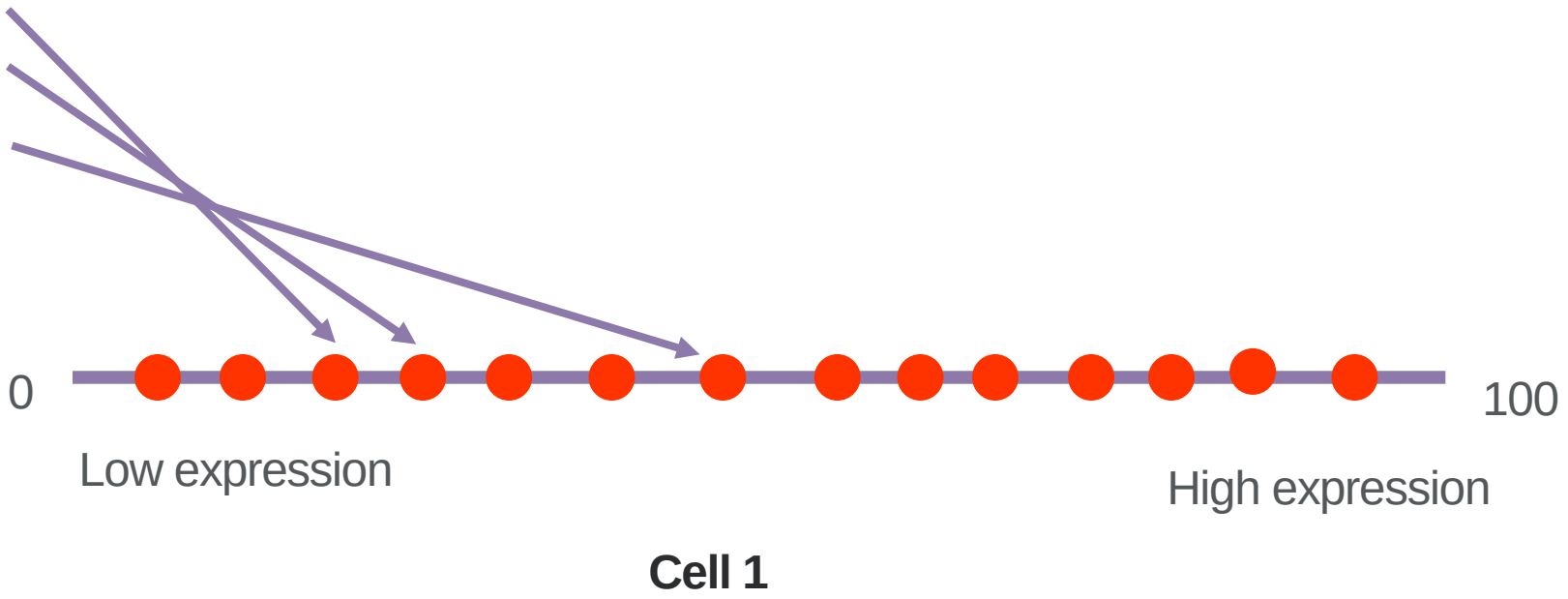
Dimension Reduction Methods: Principal Components Analysis



Poulin et al. 2014

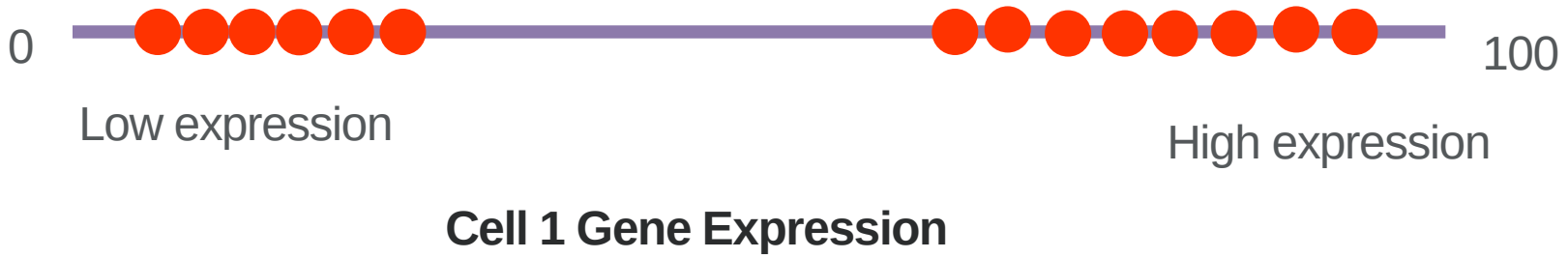
Let’s start with only 1 dimension

- Gene Cell 1
- A 2.4
- B 3.2
- C 20



Let’s start with only 1 dimension

- We are plotting multiple genes from a single cell only.



Now onto 2D

Gene	Cell 1	Cell 2
A	2.4	2.6
B	3.2	4.2
C	10.0	11.9

Now onto muti-dimensions

- 1 cell → 1D graph
- 2 cells → 2D graph
- 3 cells → 3D graph
- 4 cells → 4D graph
- .
- .
- .
- N cells → N-dimension graph

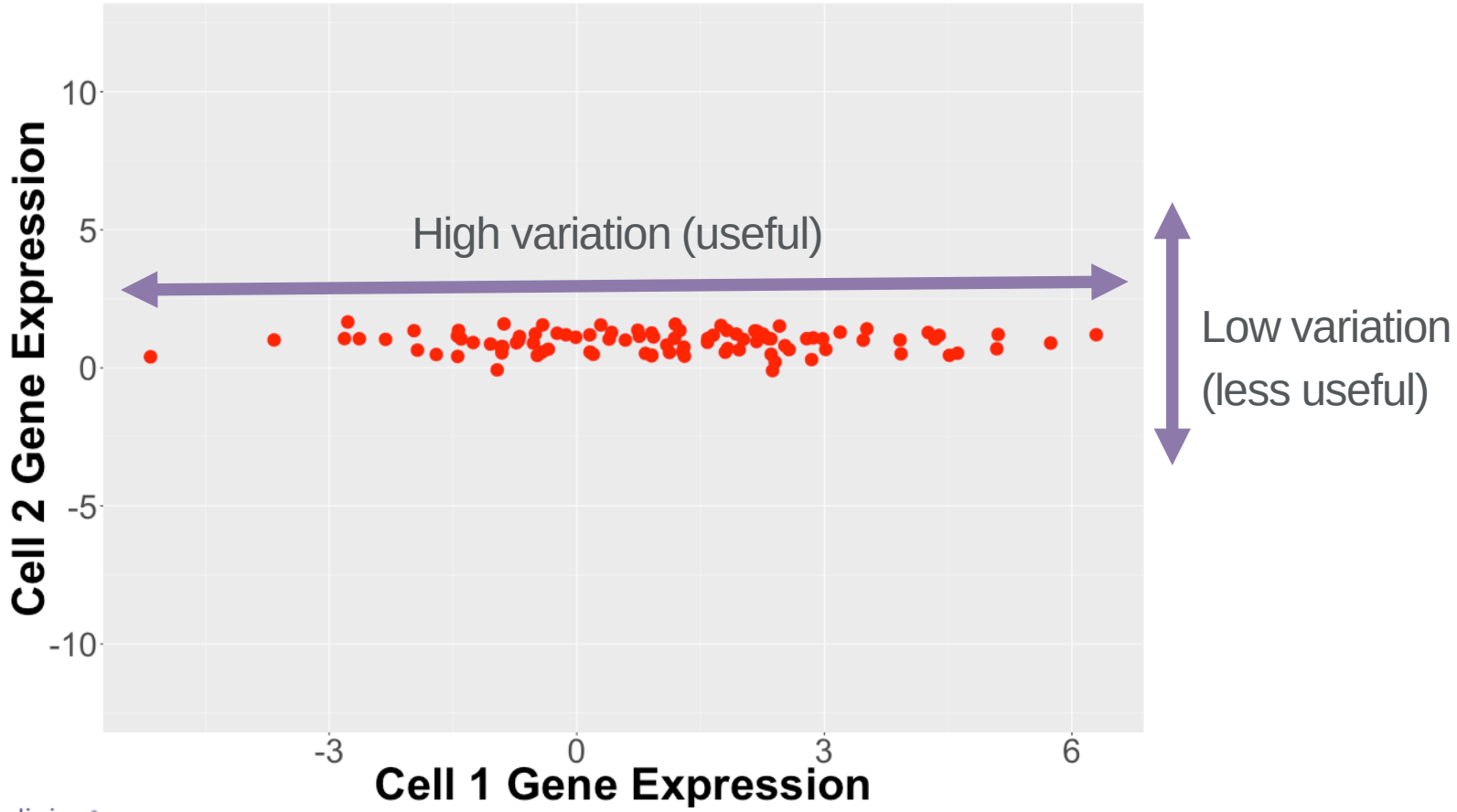
- How can we draw N-dimensional graph? You CAN'T!!

Not all dimensions are created equal

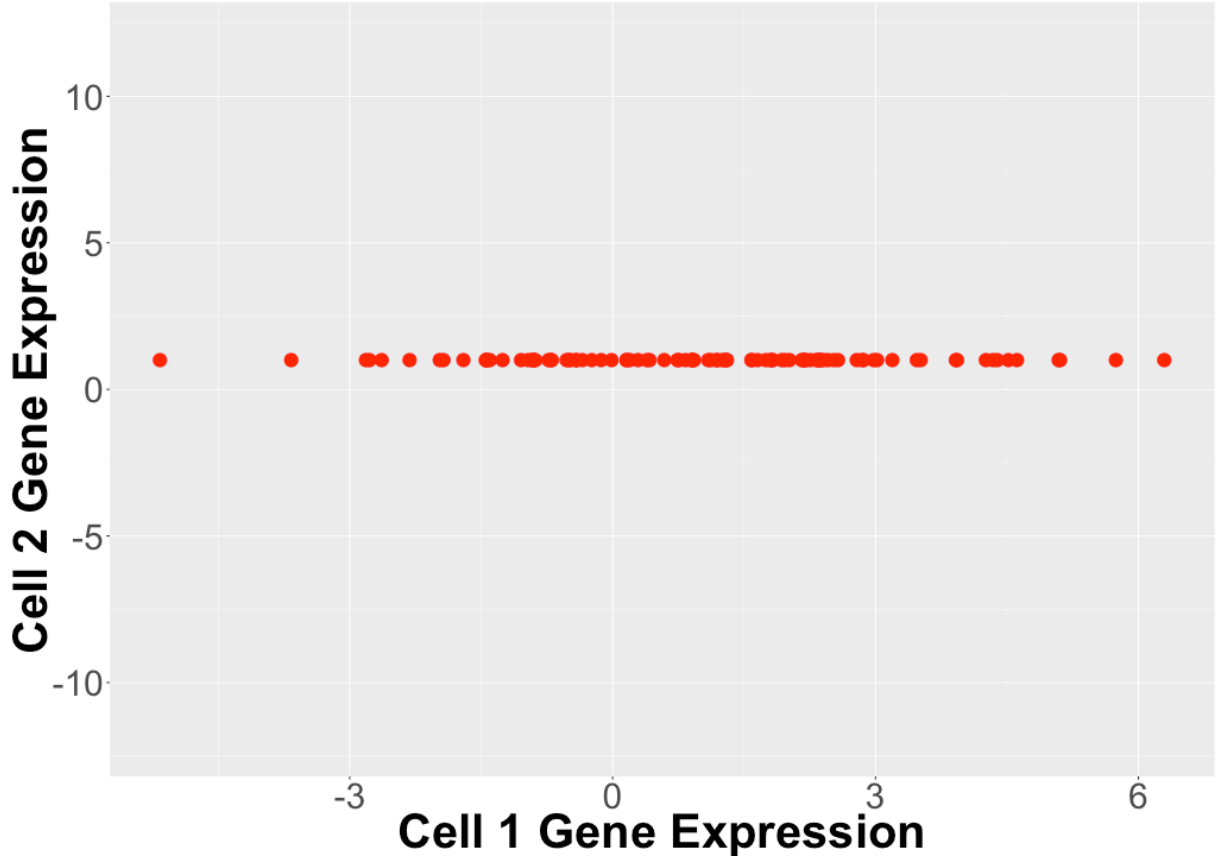
- Are all dimensions equally important?

Not all dimensions are created equal

- This is where dimension reduction comes into play.
- Are all of these dimensions (i.e. cells) equally important?

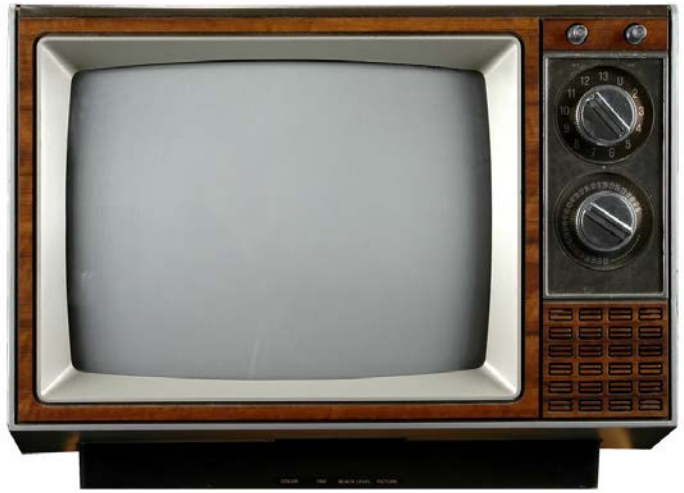


Dimension Reduction of Cell 2



Analogy

1 Dimension



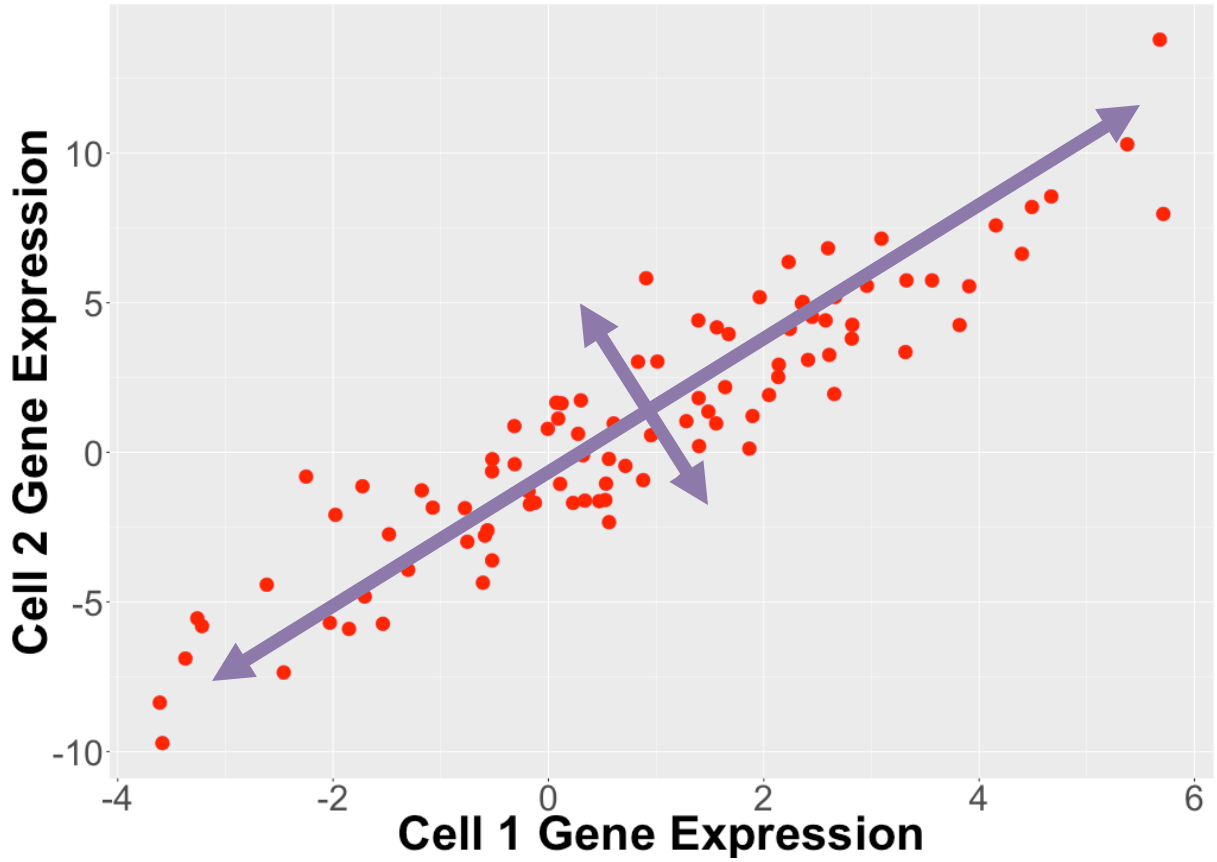
2 Dimensions

3 Dimensions

Principal Components Analysis (PCA)

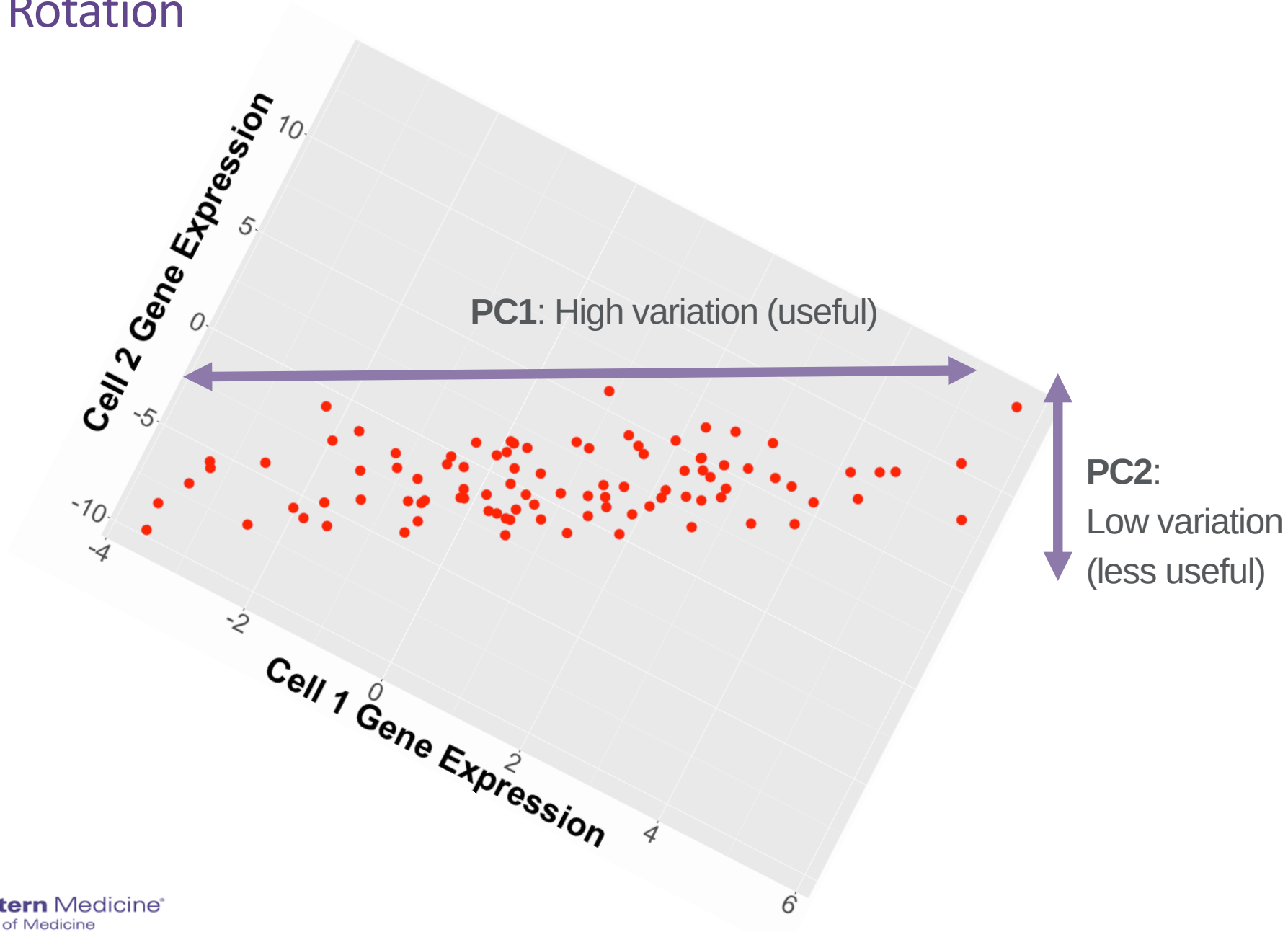
- Flattens the data without losing much information
- Goal is to find the important dimensions
- E.g. Using information of many cells, reduce it to a few dimensions that we can visualize

Now onto 2D



Gene	Cell 1	Cell 2
A	2.4	2.6
B	3.2	4.2
C	10.0	11.9

PCA Rotation

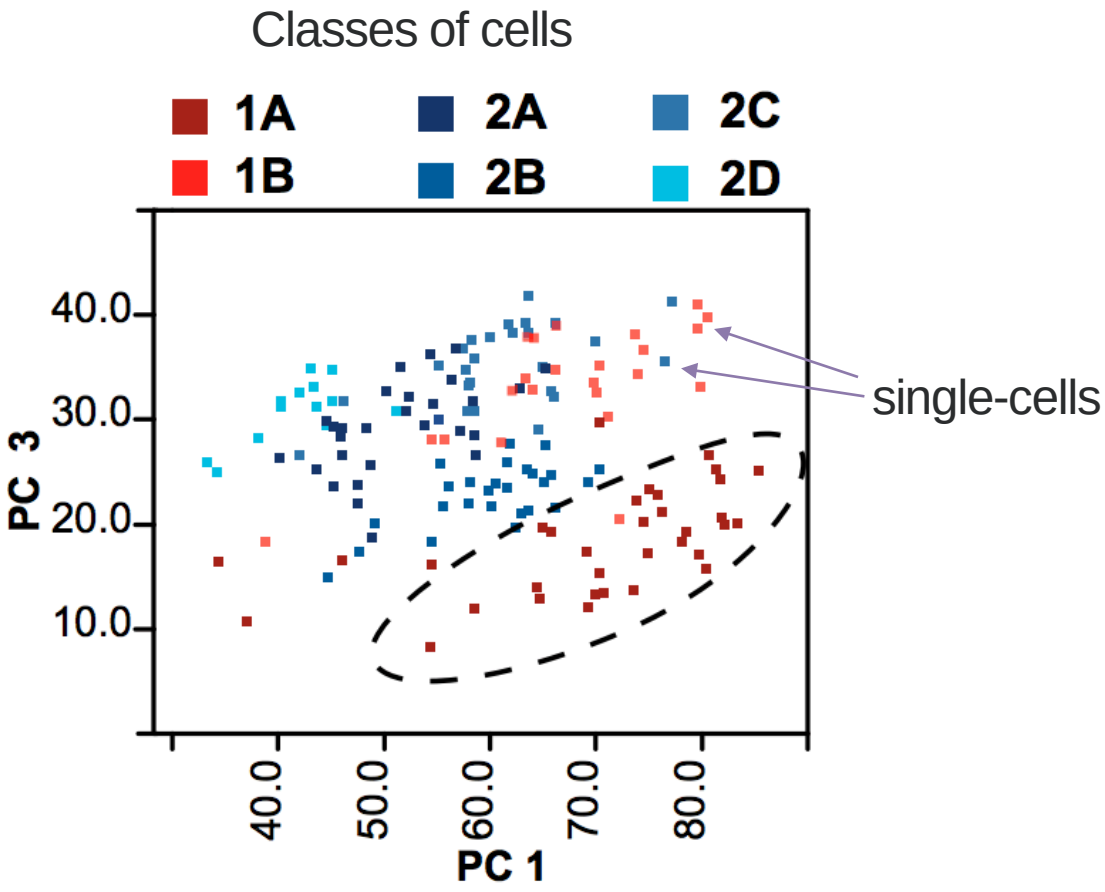


In summary

- If we have 2 cells
  - PC1 spans in the direction that captures the most variation
  - PC2 spans in the direction that captures the 2nd most variation

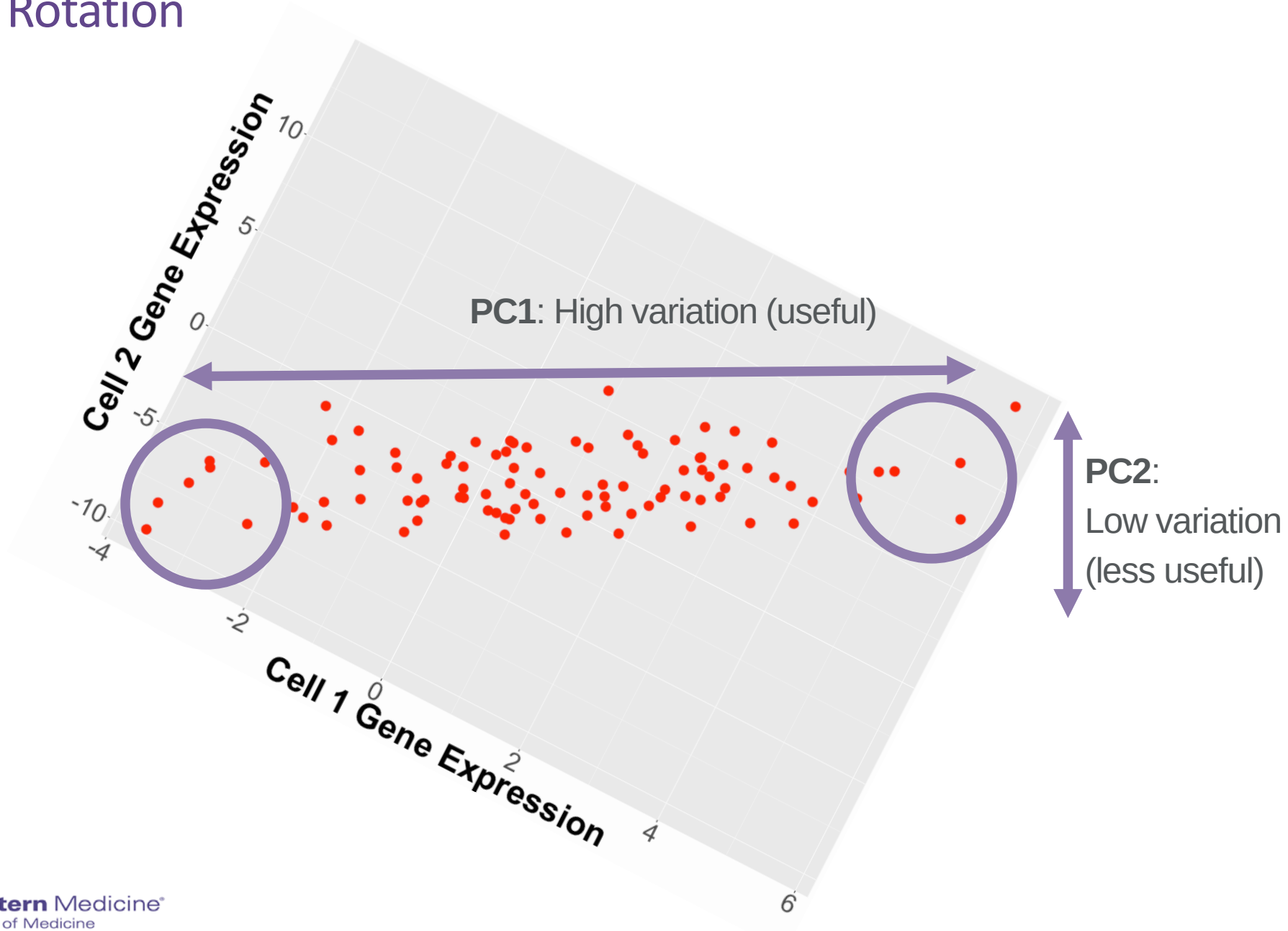
- If we have N cells
  - PC1 spans in the direction that captures the most variation
  - PC2 spans in the direction that captures the 2nd most variation
  - PC3 spans in the direction that captures the 3rd most variation
  - …
  - PCN spans in the direction that captures the least variation

Let’s revisit the PCA plot



Poulin et al. 2014

PCA Rotation



PCA Procedure

Each cell ends up with "some value" for each principal component.

Gene	Cell 1	Influence on PC1 (Loadings)
A	-2.1	Low (.1)
B	1.2	Low (.1)
C	12.4	High (10)
D	-5.3	Medium (.2)
E	1.2	Low (.2)
F	0.2	Low (.1)
…	…	…

Gene	Cell 2	Influence on PC1 (Loadings)
A	-0.2	Low (0.1)
B	1.7	Low (0.3)
C	3.4	medium
D	-2.3	medium
E	0.2	low
F	1.5	low
…	…	…

Cell 1: PC1 score = -2.1*0.1 + 1.2*0.1 + … = some value 1

Cell 1: PC2 score = similar idea with PC2 loadings for cell 1 = some value

…

Cell 1 *n*th PC score = similar strategy with the *n*th PC loadings for cell 1 = some value

PCA Procedure

Each cell ends up with "some value" for each principal component.

Gene	Cell 1	Influence on PC1 (Loadings)	Gene	Cell 2	Influence on PC1 (Loadings)
A	-2.1	Low (.1)	A	-0.2	Low (0.1)
B	1.2	Low (.1)	B	1.7	Low (0.3)
C	12.4	High (10)	C	3.4	medium
D	-5.3	Medium (-2)	D	-2.3	medium
E	1.2	Low (.2)	E	0.2	low
F	0.2	Low (.1)	F	1.5	low
…	…	…	…	…	…

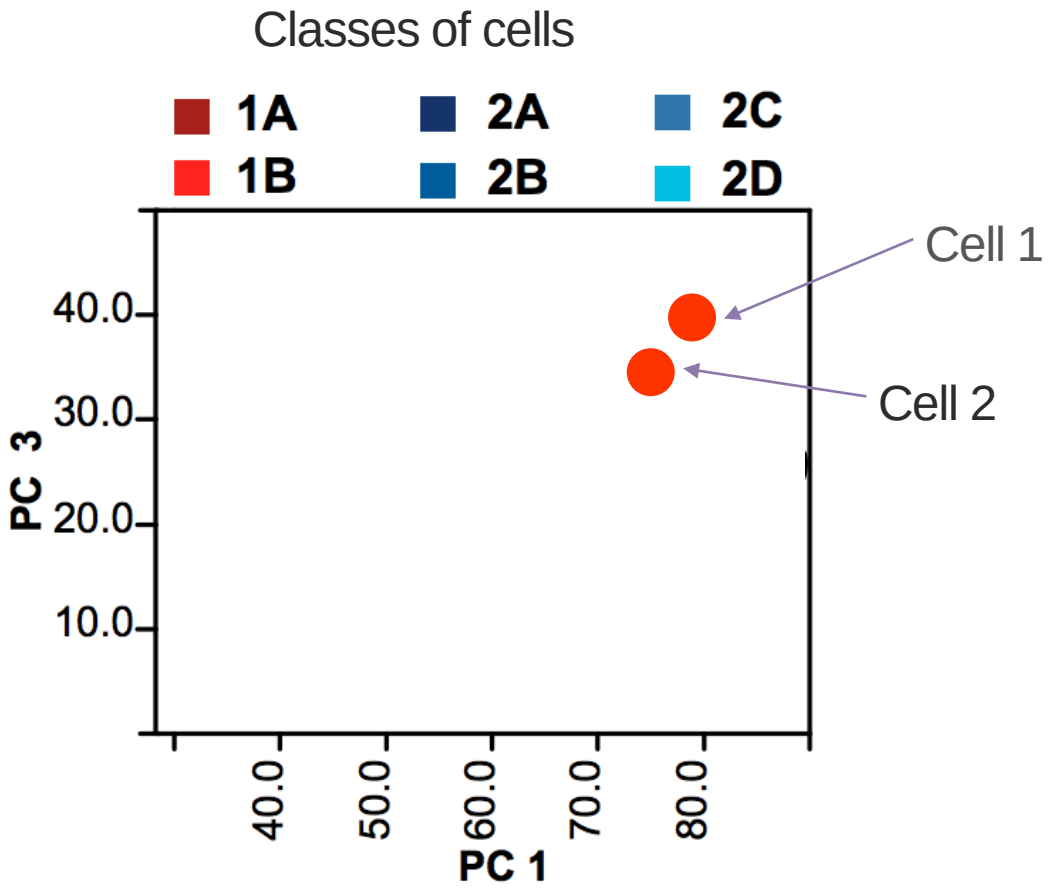
Cell 2 PC1 score = -0.2*0.1 + 1.7*0.3 + … = some value 2

Cell 2 PC2 score = similar strategy with PC2 loadings for cell 2 = some value

…

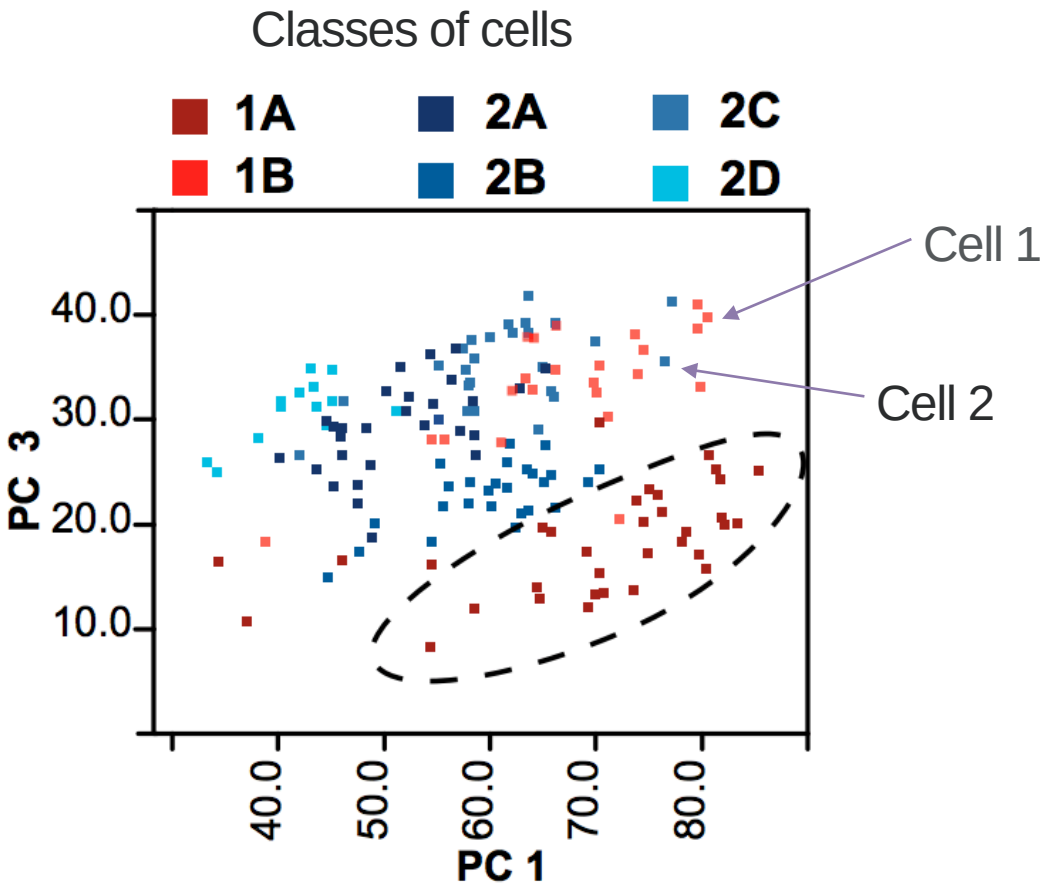
Cell 2 PCn score = similar strategy with the Nth PC loadings for cell 2 = some value

Revisit PCA Plot



Poulin et al. 2014

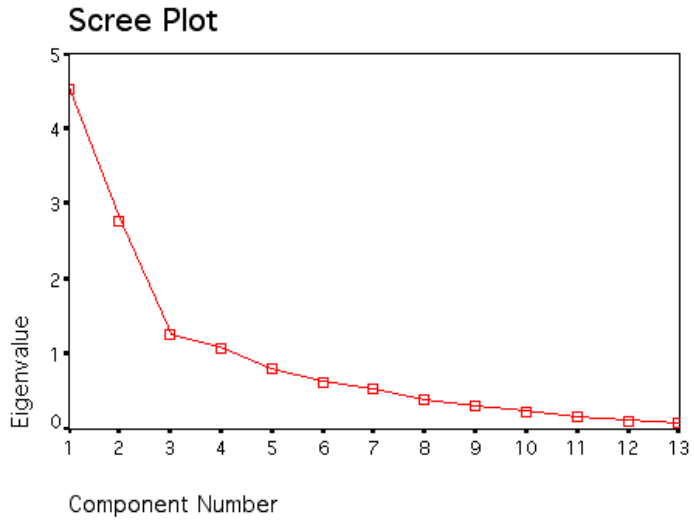
Revisit PCA Plot



Poulin et al. 2014

In summary

- PCA is a way to reduce the dimension into the most influential principal components
- Genes with high “impact score" (loadings) in a principal component are more influential
- Scree plot shows the variation accounted for by each principal component

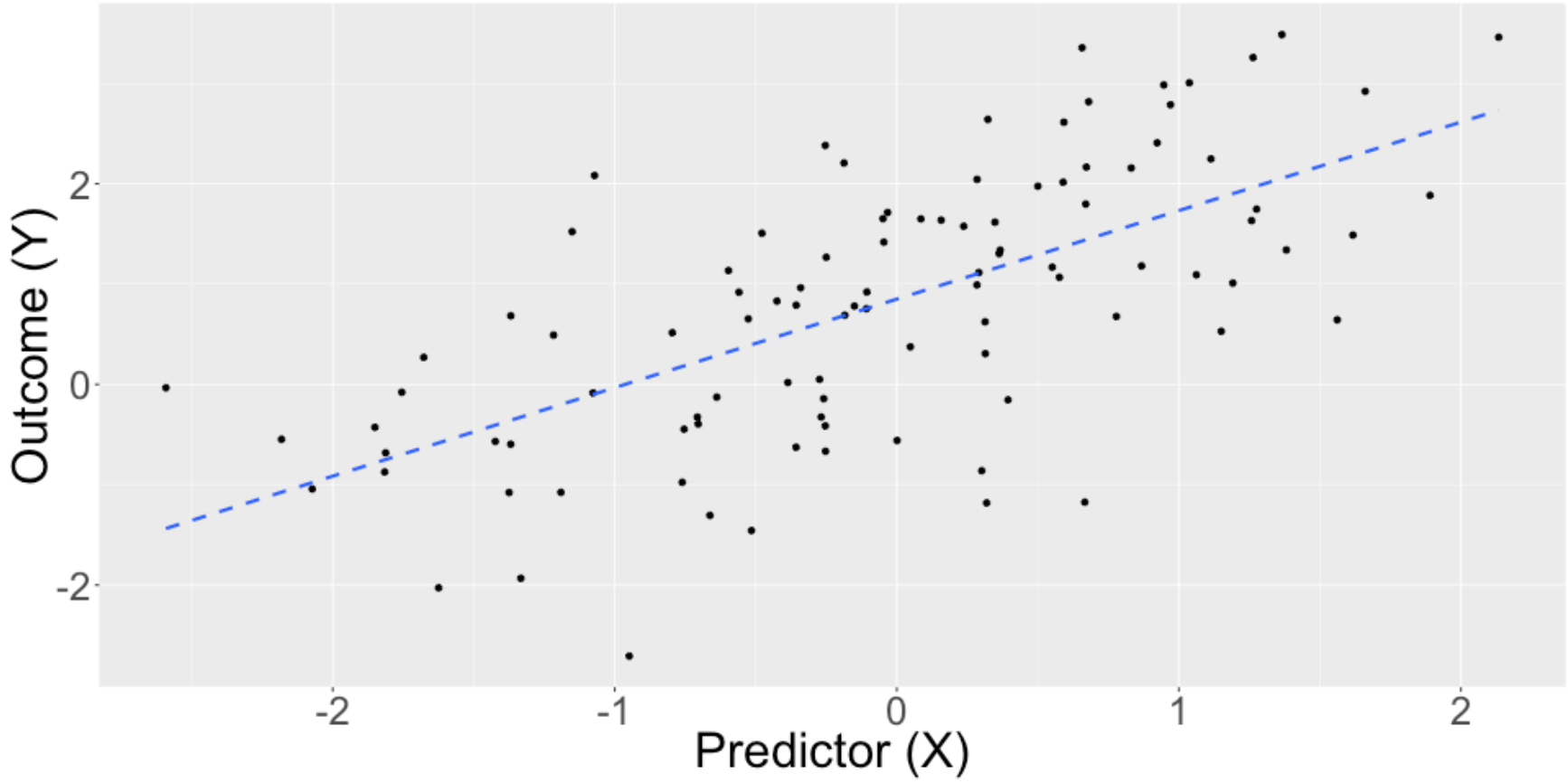


Northwestern Medicine®
Feinberg School of Medicine

Shrinkage Methods

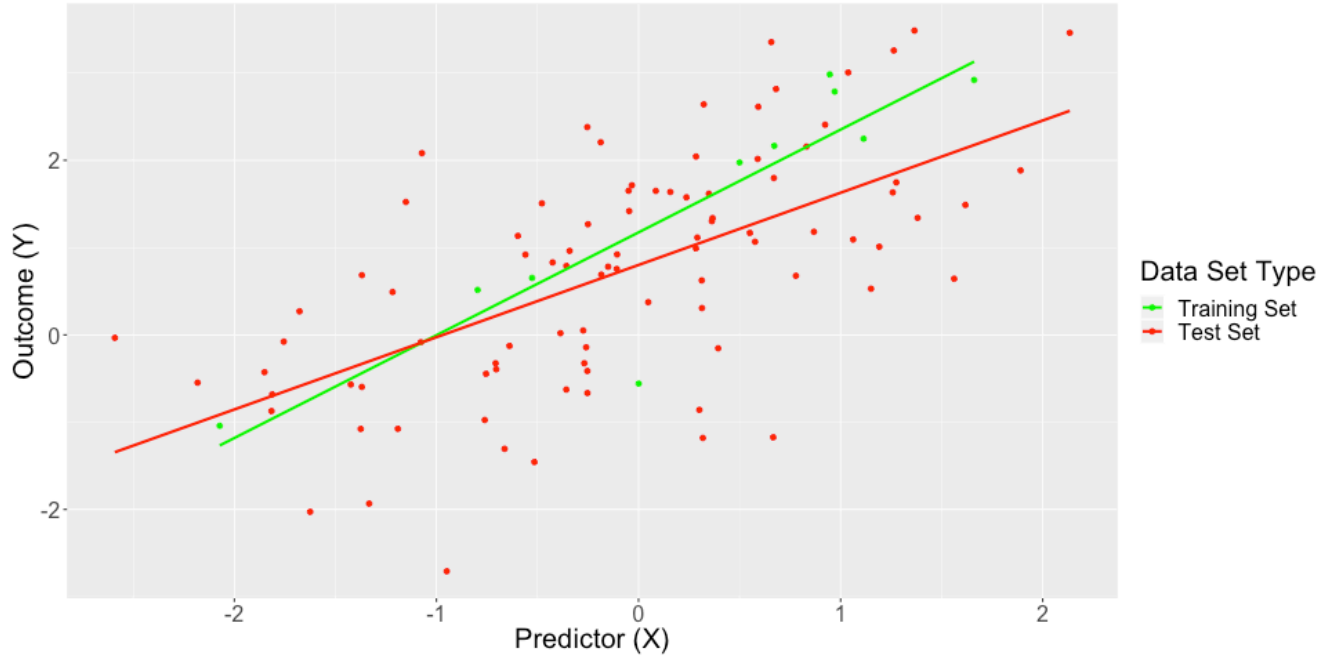
Regression Models

- Think about the regression models, where you identify the association between two variables: X and Y



Ridge Regression

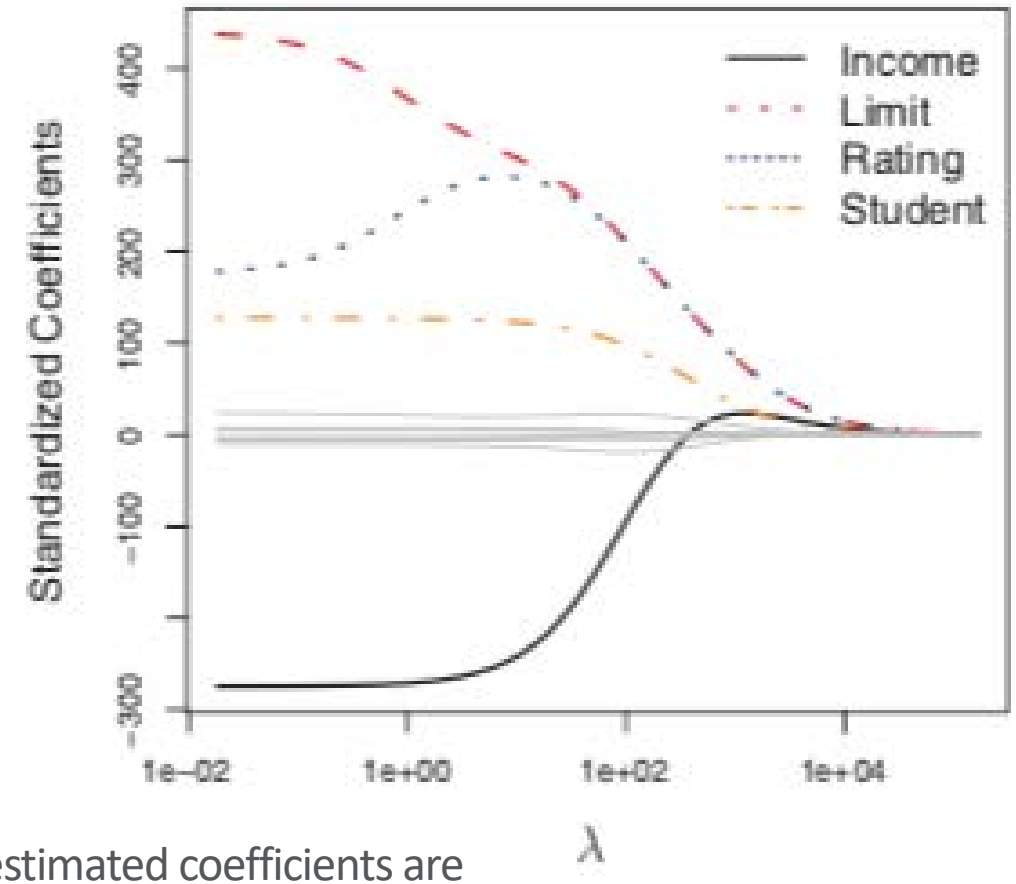
- Purpose of the model is to reduce model variance by increasing a little bias (variance-bias tradeoff)
- This is done by introducing a penalty term (or tuning parameter)
- The result of fitting a ridge regression is that the model is less sensitive to the training dataset and reduces overfitting



tuning parameter

$$\sum_{i=1}\left(y_i-\beta_0-\sum_{j=1}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

The effect of Tuning Parameter in Ridge Regression

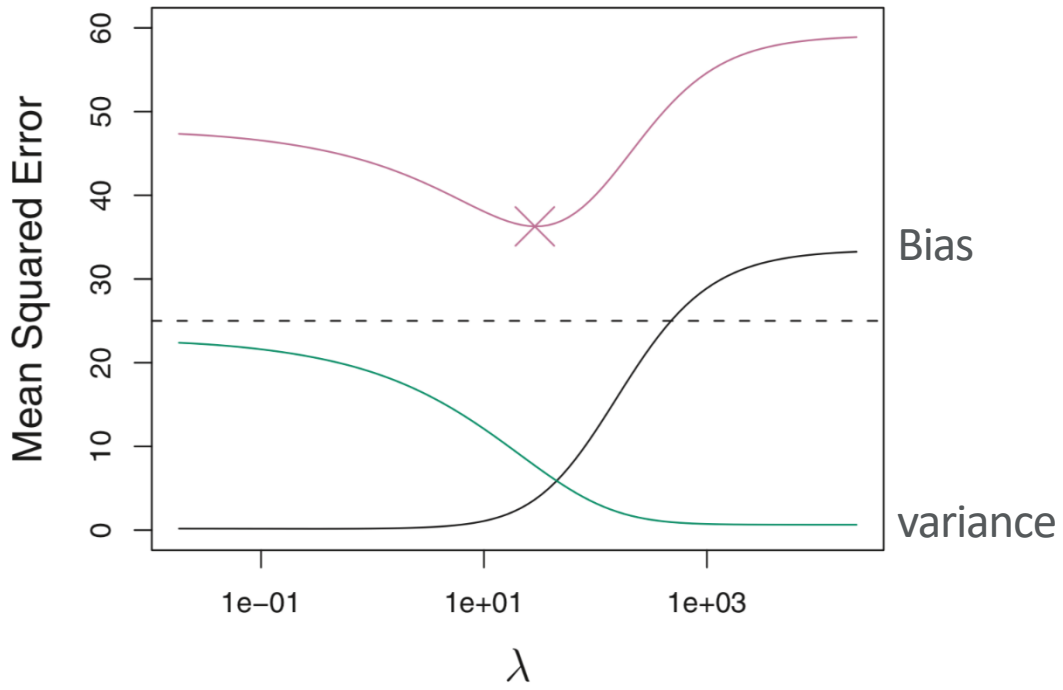


Note: the estimated coefficients are standardized

Source: An Introduction to Statistical Learning (2017)

Why does Ridge Regression Improve over OLS?

- Bias-variance tradeoff
- With increase in lambda, there is less flexibility in estimating the parameters, i.e. bias increases



Source: An Introduction to Statistical Learning (2017)

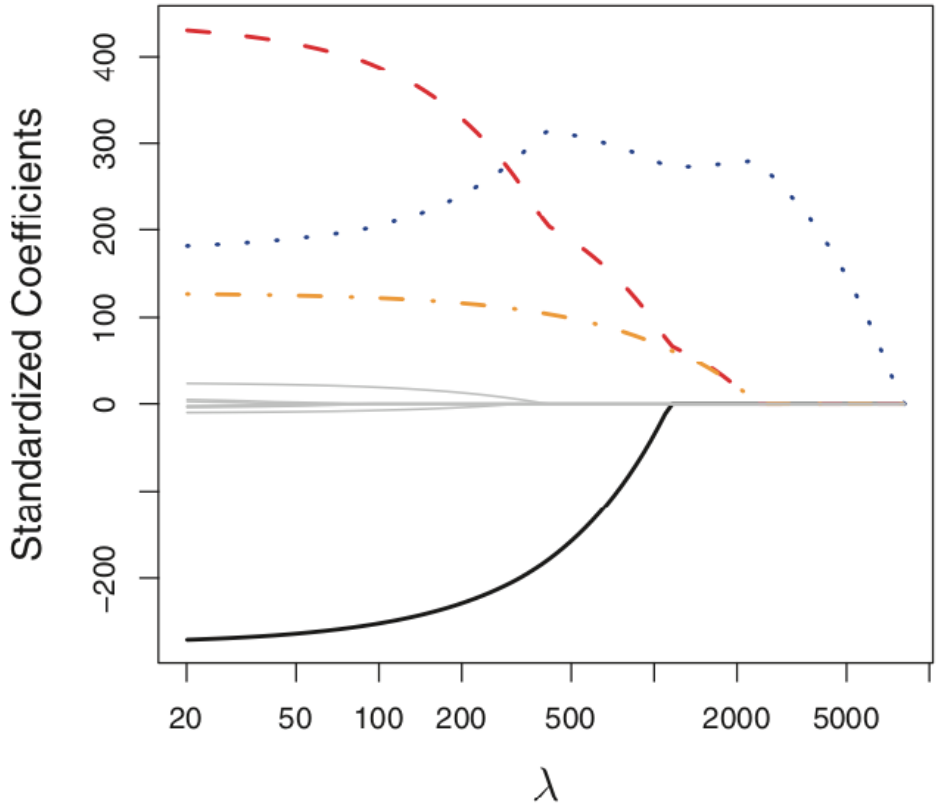
Least Absolute Shrinkage and Selection Operator (LASSO)

- Similar to Ridge regression, LASSO also uses a tuning parameter
- Instead of using the squared term, LASSO uses absolute value
- LASSO can decrease the slope to zero, not just asymptotically zero
  - This subtle difference has a huge implication

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

- Ridge regression works better when most variables are important
- LASSO is more useful when many of the variables are useless

The effect of Tuning Parameter in LASSO



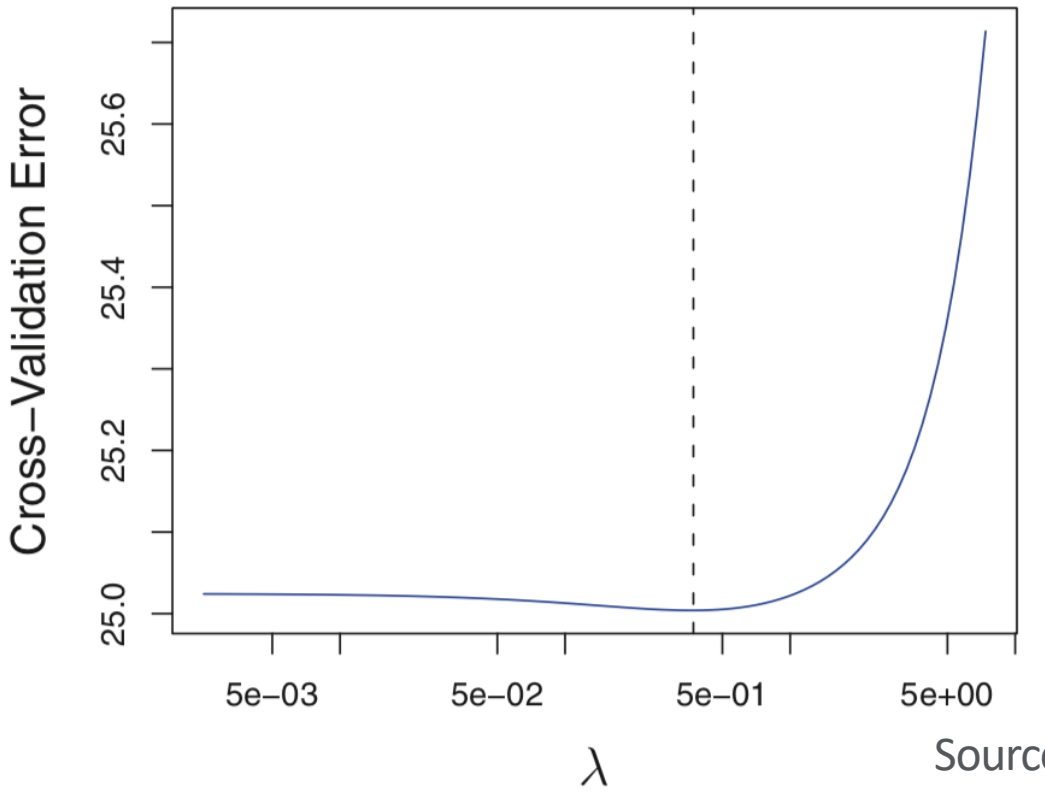
Notice that the parameters shrink to an absolute zero

Note: the estimated coefficients are standardized

Source: An Introduction to Statistical Learning (2017)

Selecting Tuning Parameter

- Using cross-validation, we choose a grid of lambda values and compute the cross-validation error
- Then we could decide the optimal value of lambda



Source: An Introduction to Statistical Learning (2017)

Elastic Net

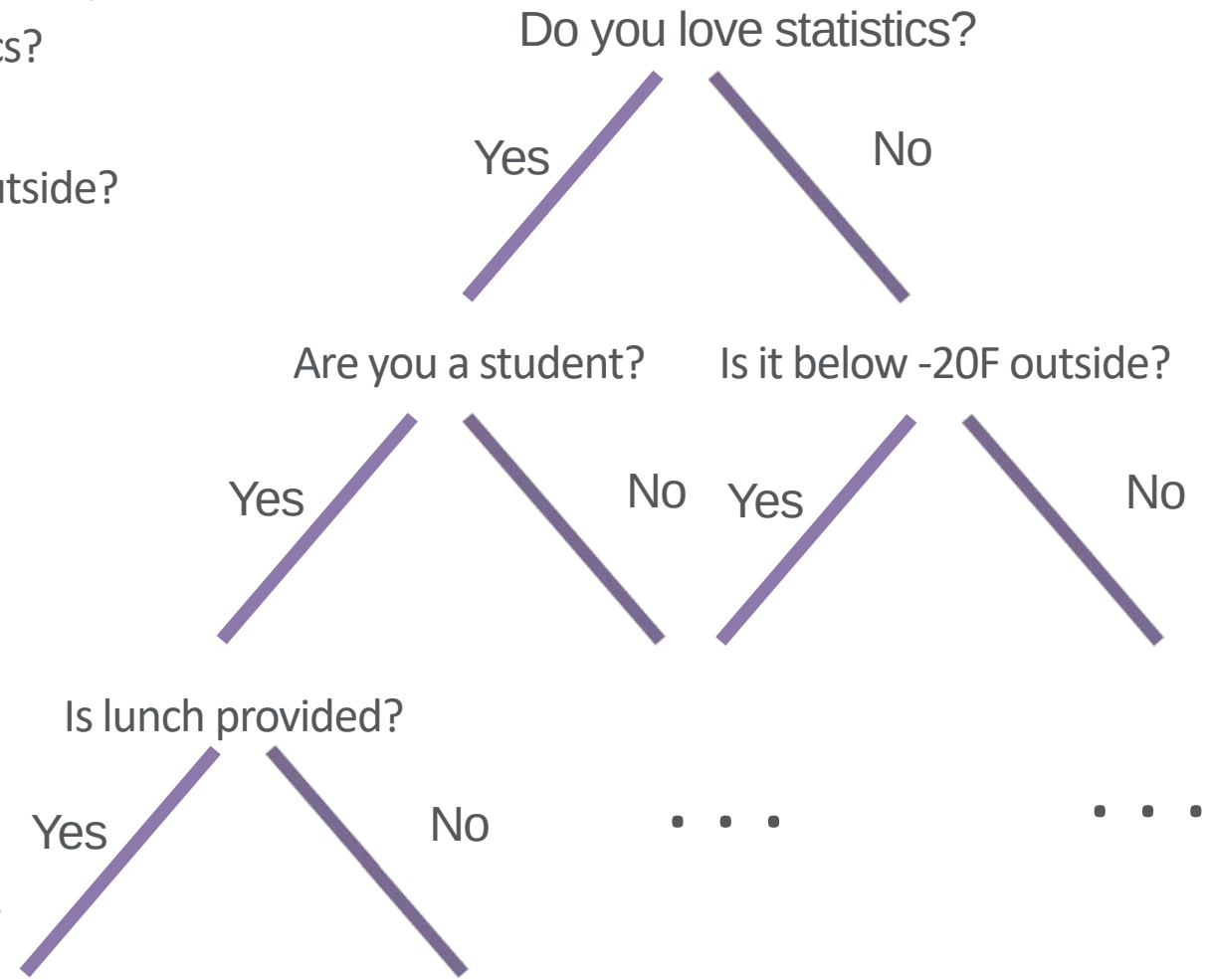
- When faced with choosing between ridge regression or LASSO, you could use Elastic Net
- Elastic net combines the penalty terms from the ridge regression and LASSO
- Cross-validation to find the optimal tuning parameters for Ridge and LASSO
- Elastic-net groups and shrinks the parameters associated with the correlated

Northwestern Medicine®
Feinberg School of Medicine

Machine Learning

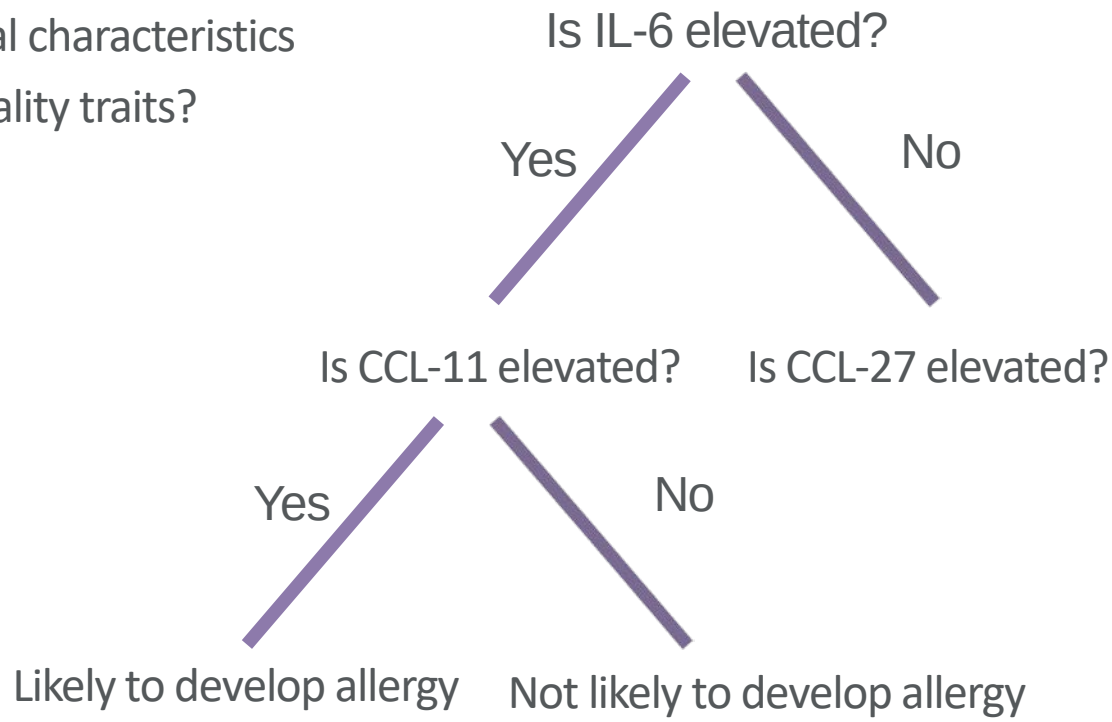
A thought experiment

- Among the people who responded that they will attend the Statistically Speaking Lecture Series, who are most likely to actually attend?
- Do you love statistics?
- Are you a student?
- Is it -20F degrees outside?
- Is lunch provided?



Examples of Machine Learning

- A big goal of machine learning is to make prediction
- To assess the model performance, we use training set to fit the model and test set to test our prediction model
- For example, can a set of biomarkers predict if a child will develop allergy in 5 years?
- Machine learning can also be used to identify patterns
- For example, based on the behavioral characteristics
  can we group people into personality traits?



Classes of Machine Learning

- Unsupervised learning (no class assignment)
  - Cluster analysis
- Supervised learning (class assignment provided)
  - kNN classification: `clusterCons`
  - Nearest shrunken centroids: `class`
  - Elastic nets: `glmnet`
  - Classification and regression trees: `rpart`
  - Random forests: `randomForest`

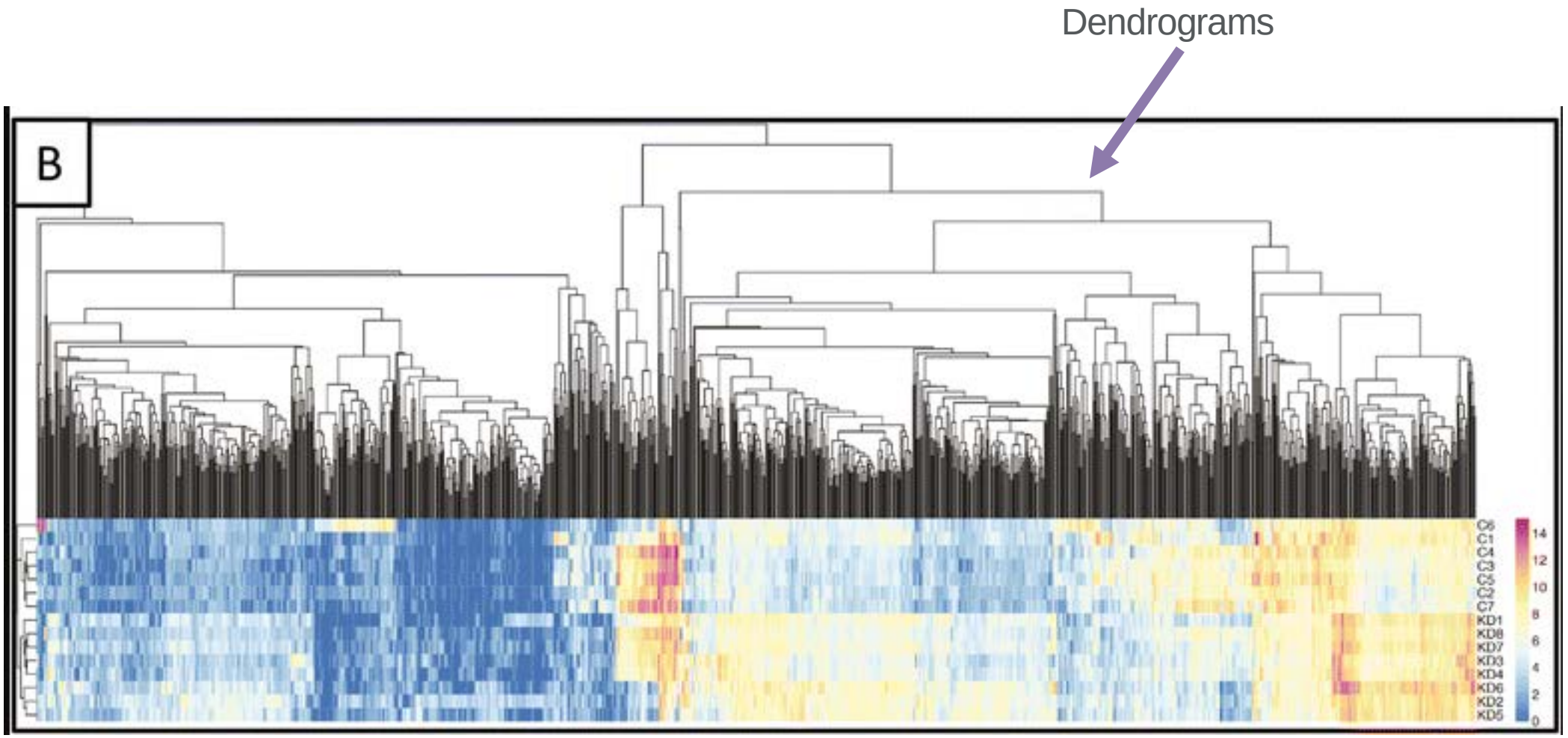
Machine Learning: **Unsupervised Learning**

- **Cluster Analysis**
  - Goal: Group similar data into groups
  - Groups are *a priori* undefined
  - Methods:
    - Hierarchical clustering
    - K-means clustering
    - Consensus clustering
    - Spectral clustering

Machine Learning: **Unsupervised Learning**

- **Cluster Analysis**
  - Goal: Group similar data into groups
  - Groups are *a priori* undefined
  - Methods:
    - Hierarchical clustering
    - K-means clustering
    - Consensus clustering
    - Spectral clustering

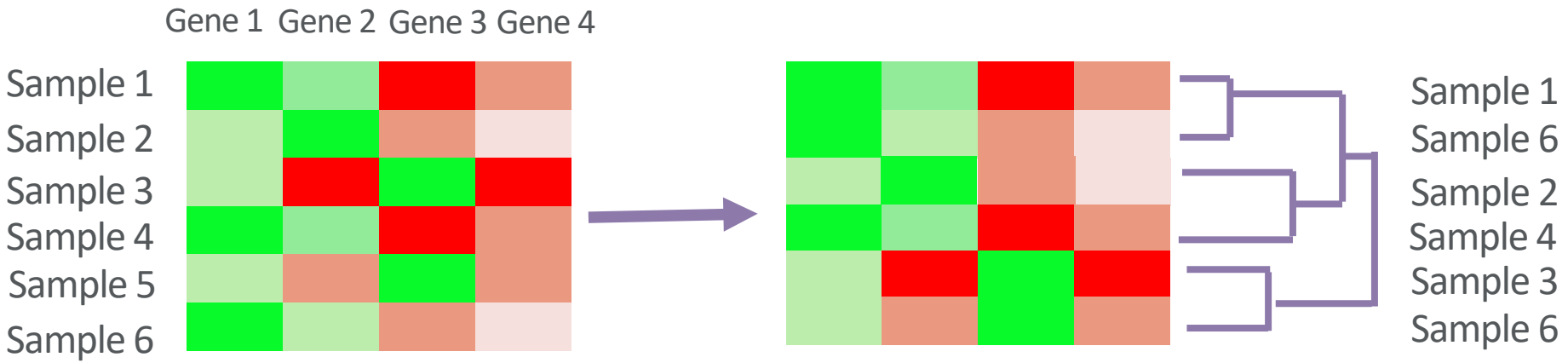
Hierarchical Cluster Example



Rowley et al. 2015

Hierarchical Clustering: Steps

- For each of the genes (row), find out which gene is most similar (we'll talk about how we determine similarity in the next few slides)
- Once we find which genes are most similar, then merge the two
- Repeat steps 1 and 2 until all genes are merged



How to measure distance

- We need to measure two metrics related to measuring distances
  - Linkage: how to define the distance, i.e. which points should we measure?
  - Distance metric: which metric of distance to use?

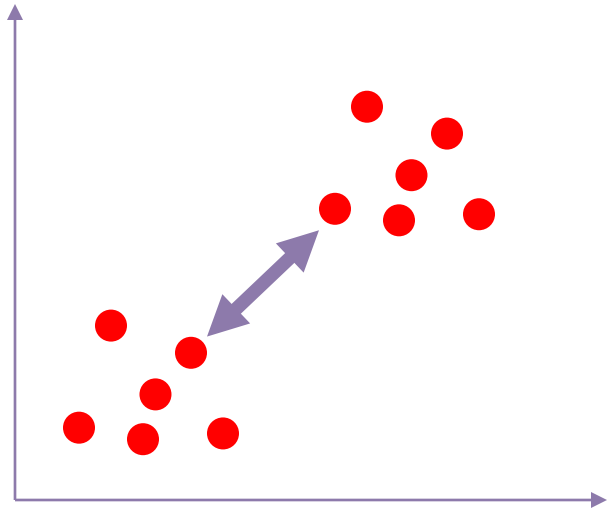
Distance Metric

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1} (a-b)}$ where S is the Covariance matrix

Source: Wikipedia

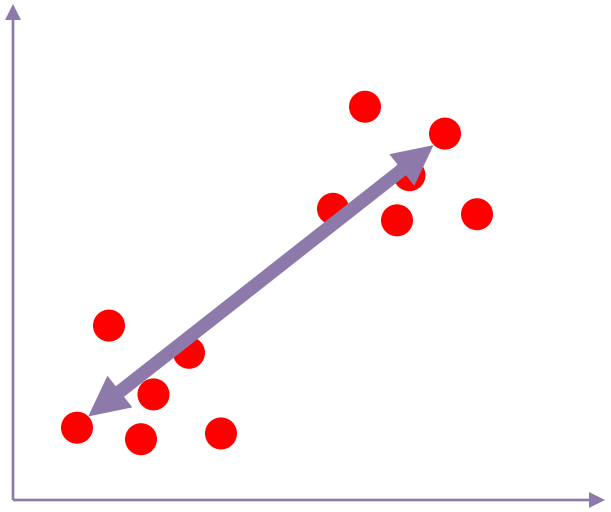
Linkage

- Nearest Neighbor (Single Linkage)



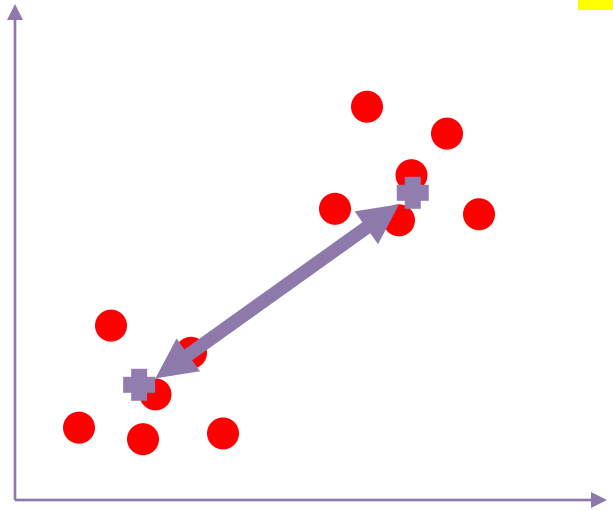
Linkage

- Furthest Neighbor (Complete Linkage)



Linkage

- Centroid



Be sure to check your software's default settings!!

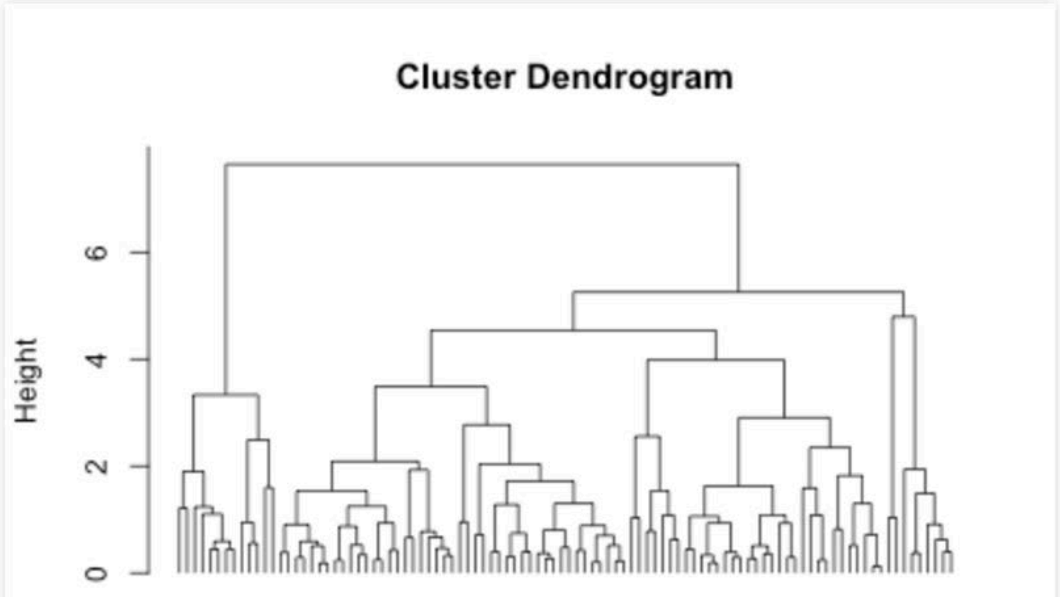
Pros and Cons of Hierarchical Clustering Analysis

- **Pros**
- Visually easy to inspect as a dendrogram
- Extremely popular to use in gene expression data

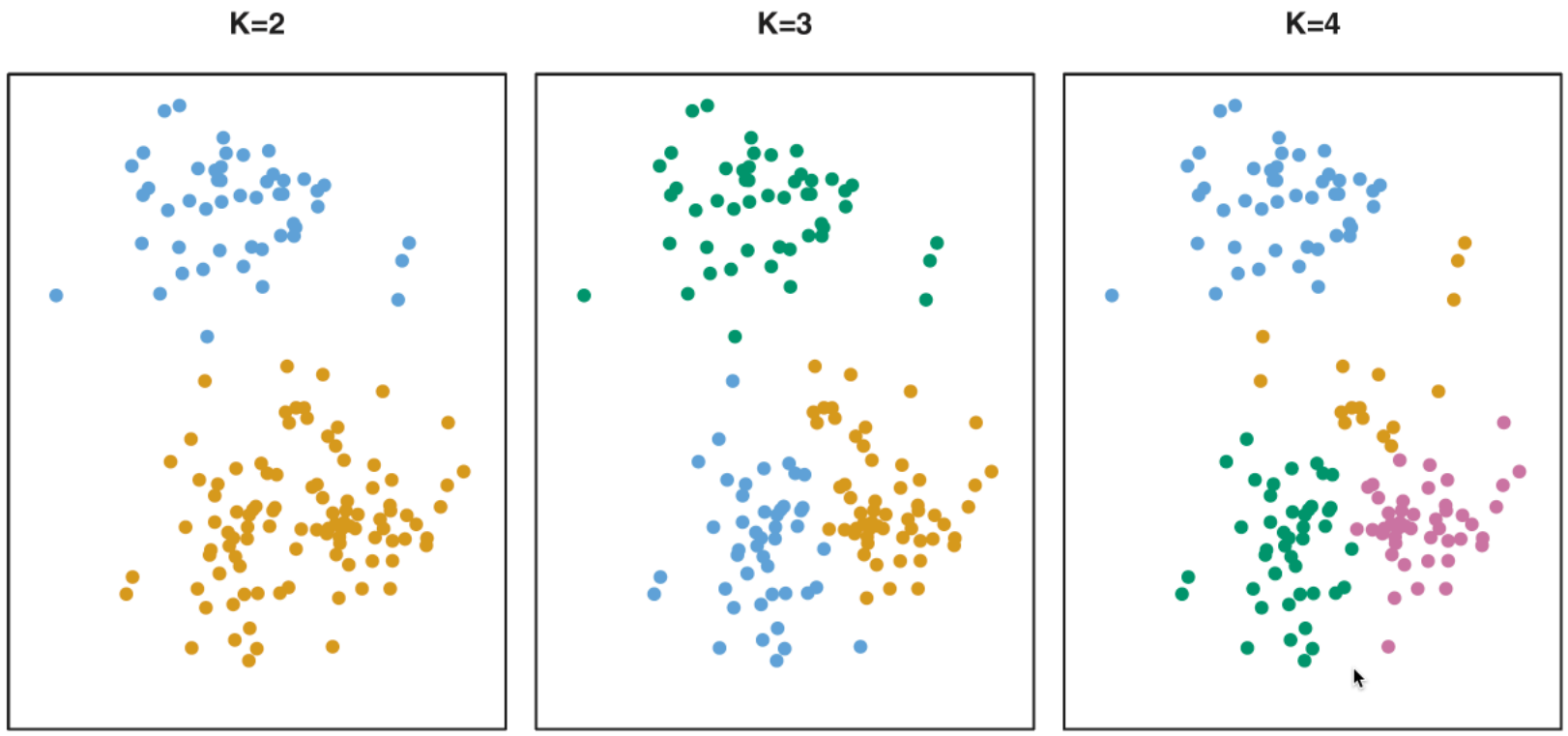
- **Cons**
- Sensitive to distance metric and linkage
- Hard to know if the hierarchical structure is real

Hierarchical Clustering with pure random noise

```
set.seed(100);
foo <- matrix(rnorm(300),nc=3) # PURE NOISE
plot(hclust(dist(foo)),hang=-1,xlab="",sub="",lab=F)
```



K-Means Clustering



Source: An Introduction to Statistical Learning (2017)

K-means Clustering

1. Select *k* items at random from the data set as the initial cluster centers
2. At each data point, calculate the distance from the data point to the center of the cluster and assign the data point to the cluster that was closest to the data point
3. Once all the data points are assigned, calculate the center of each cluster and use this as the new cluster center
4. Repeat steps 2 and 3 until none of the data points change cluster membership
5. Repeat steps 1 -- 4 multiple times then choose the clustering that produces the smallest within-cluster sum of squares

- Need to know how many clusters are present
- In R: built-in function **`kmeans`**

Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers

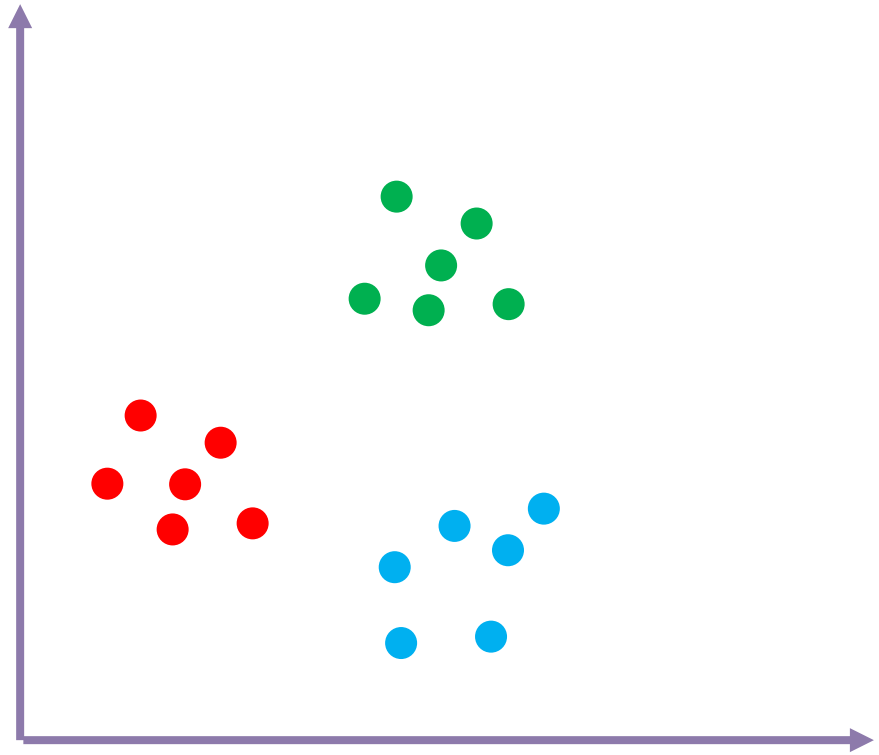


Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

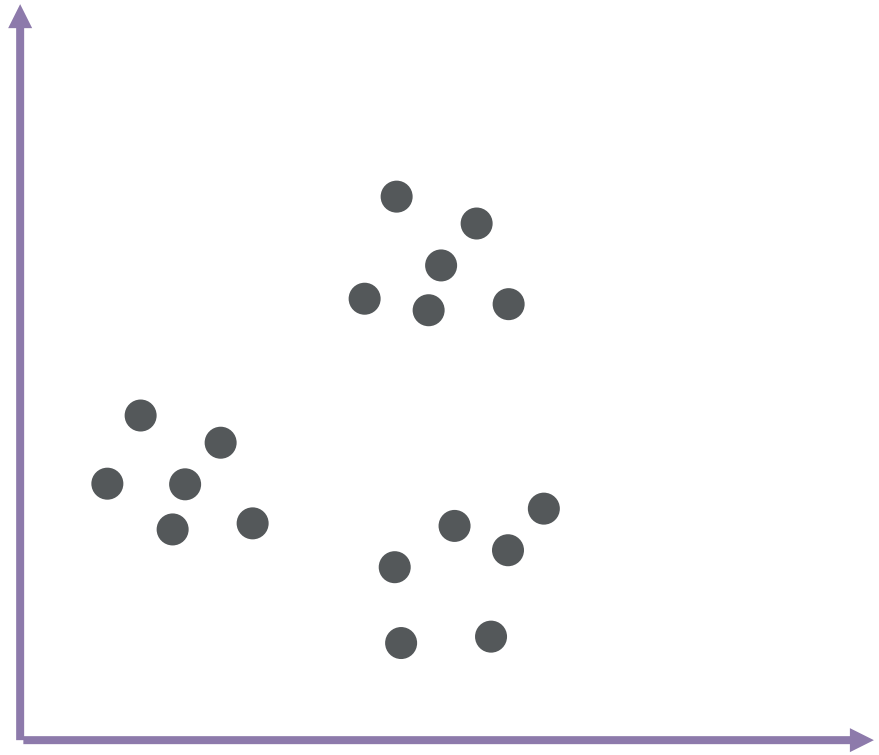


Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

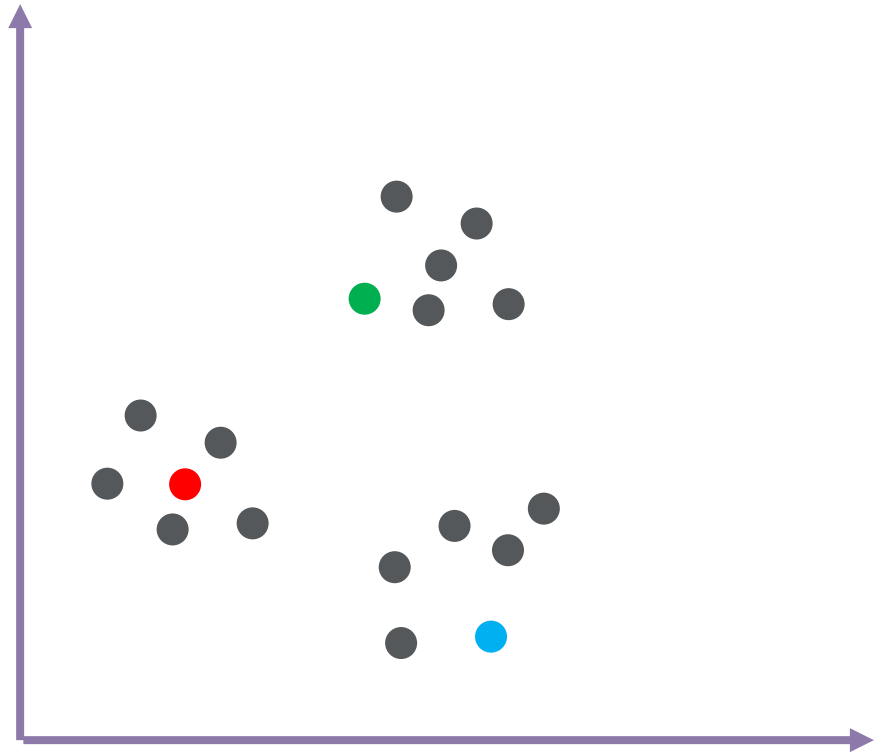


Illustration of K-means Clustering

- 2. At each data point, calculate the distance from the data point to the center of the cluster and assign the data point to the cluster that was closest to the data point

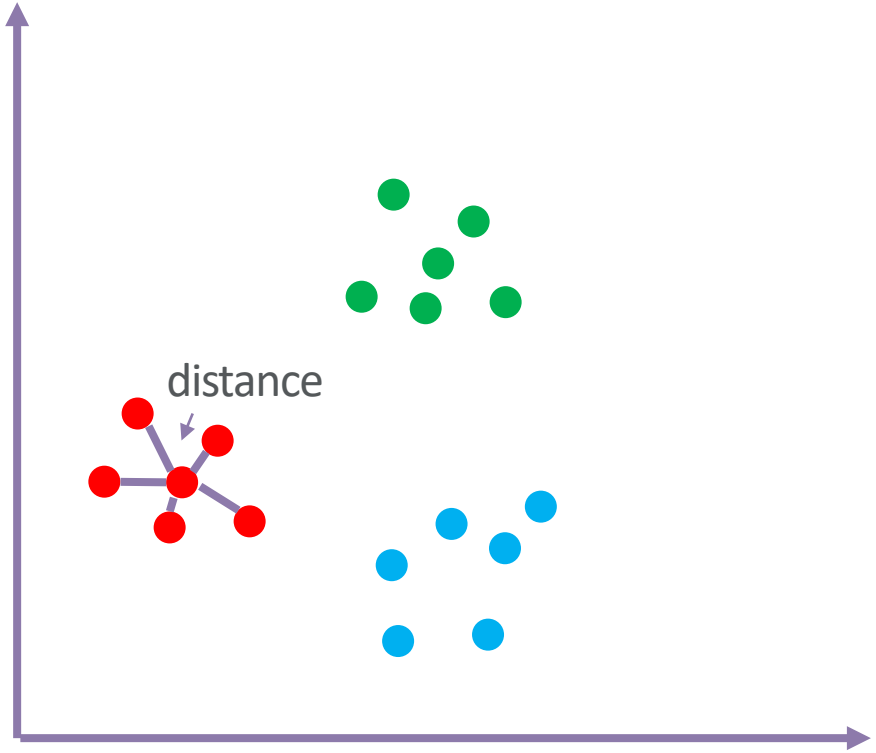


Illustration of K-means Clustering

- 2. At each data point, calculate the distance from the data point to the center of the cluster and assign the data point to the cluster that was closest to the data point

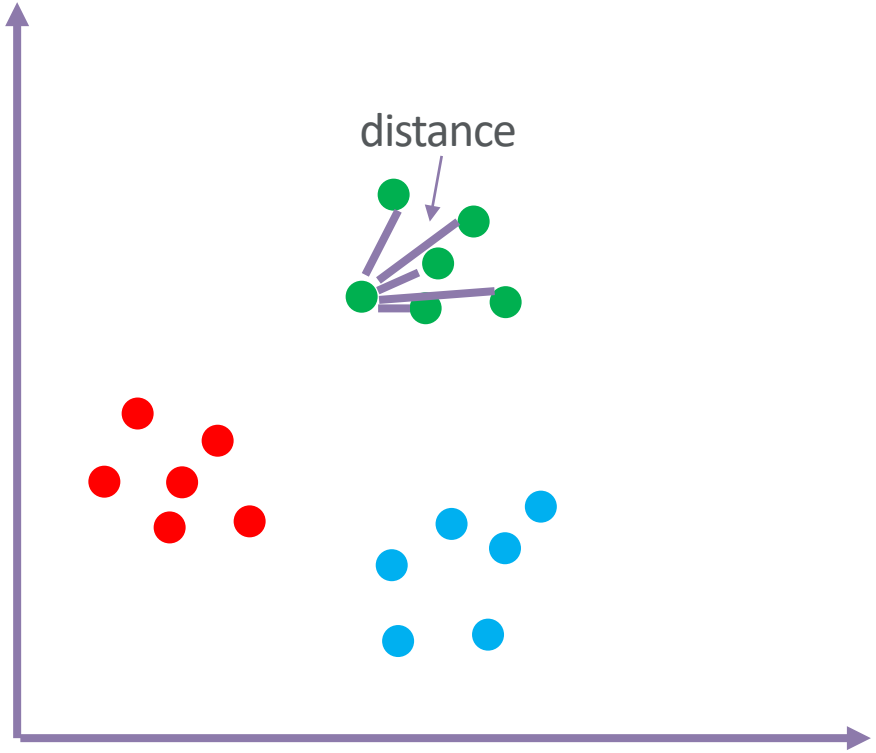


Illustration of K-means Clustering

- 2. At each data point, calculate the distance from the data point to the center of the cluster and assign the data point to the cluster that was closest to the data point

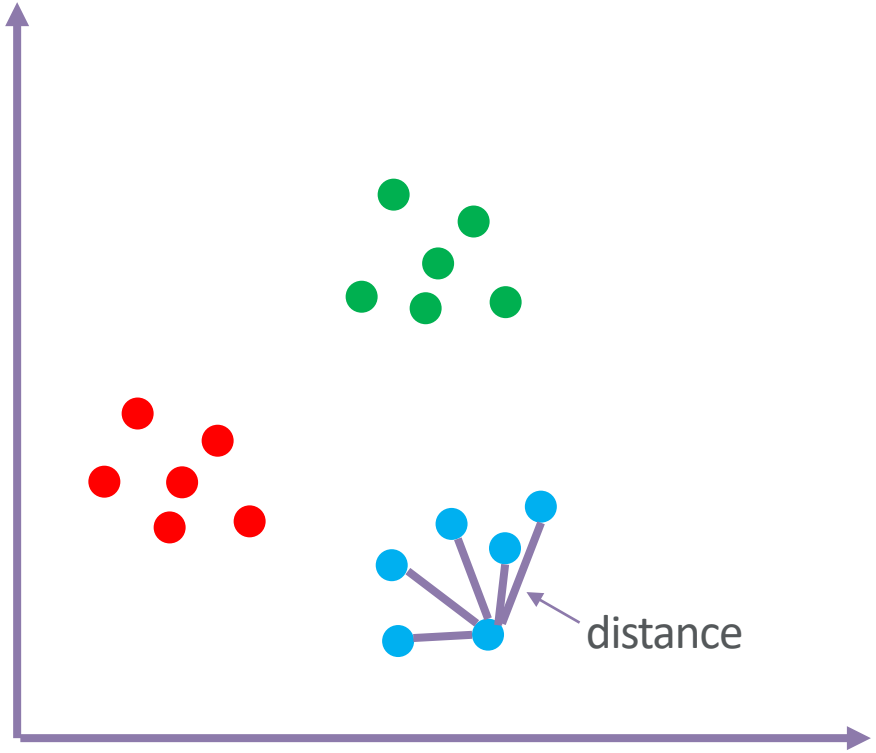


Illustration of K-means Clustering

- 3. Once all the data points are assigned, calculate the center of each cluster and use this as the new cluster center

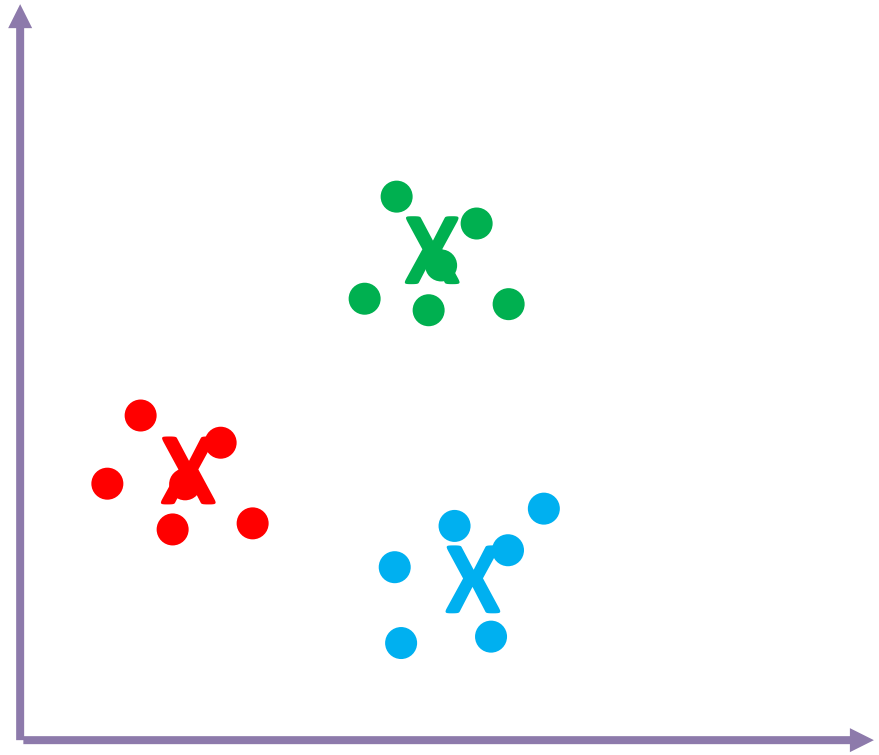
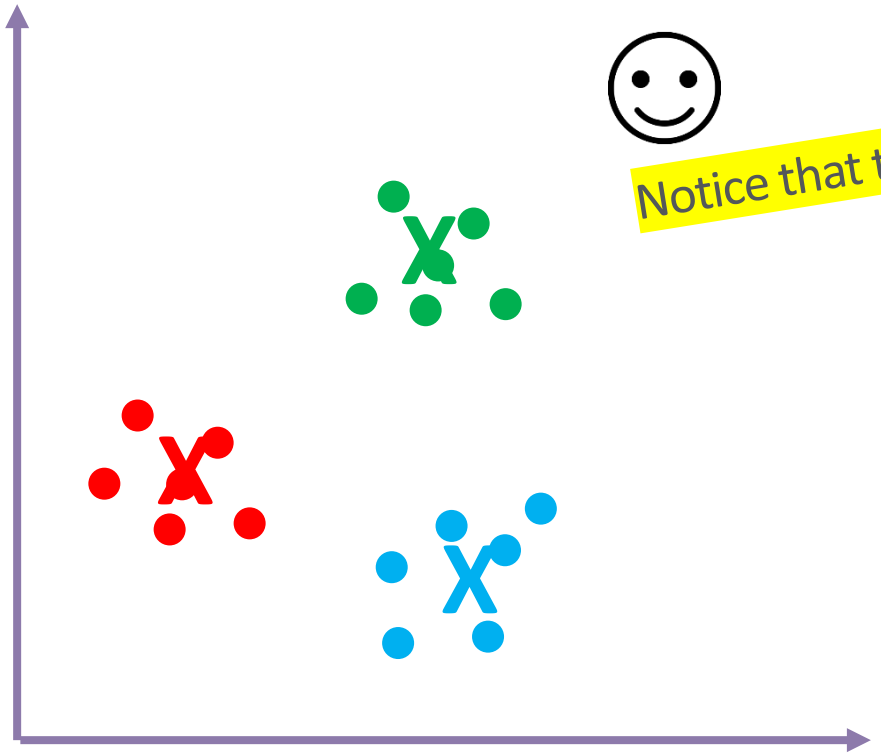


Illustration of K-means Clustering

- 3. Once all the data points are assigned, calculate the center of each cluster and use this as the new cluster center



Notice that the cluster membership is optimized!!

Illustration of K-means Clustering with different initial points

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

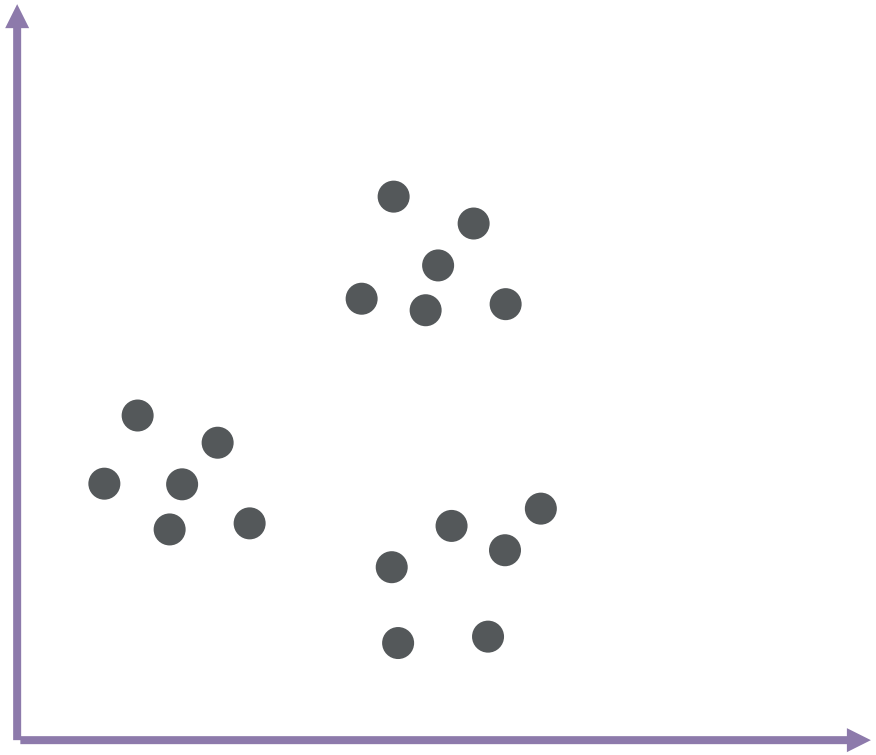


Illustration of K-means Clustering with different initial points

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

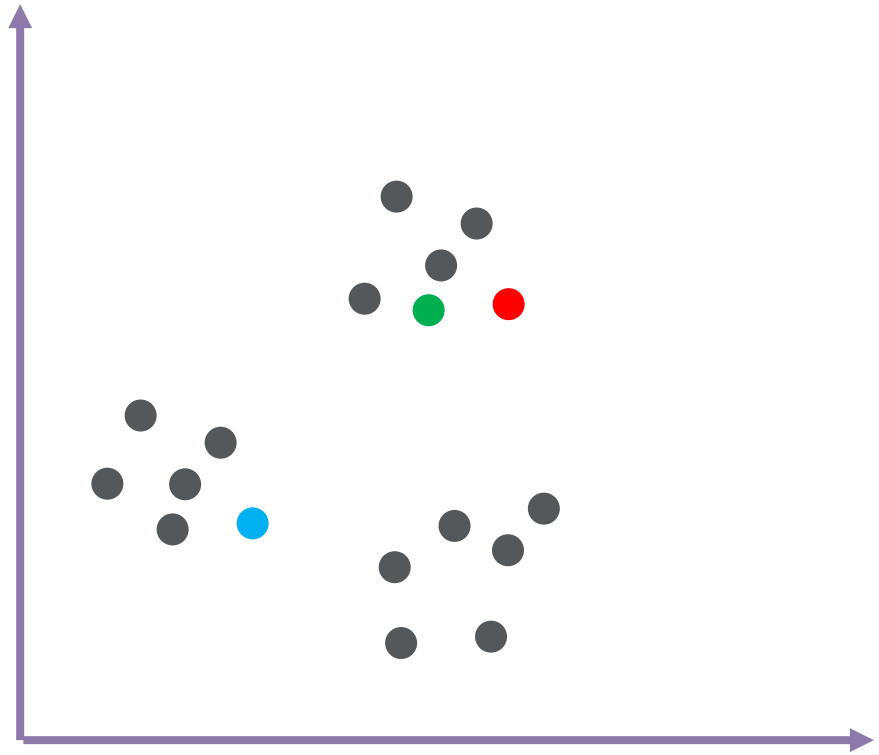


Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

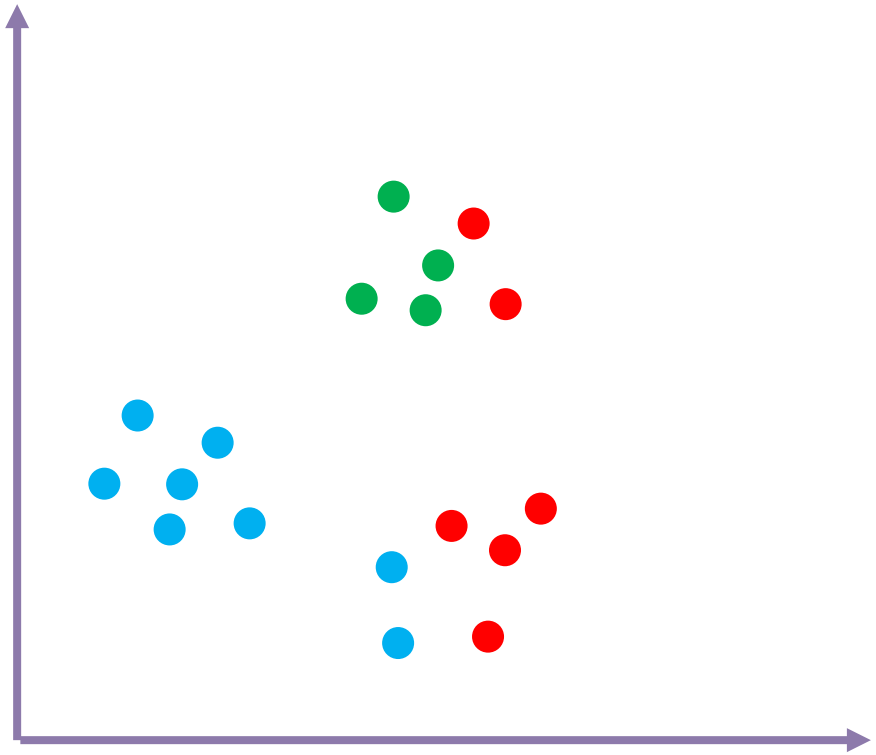


Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

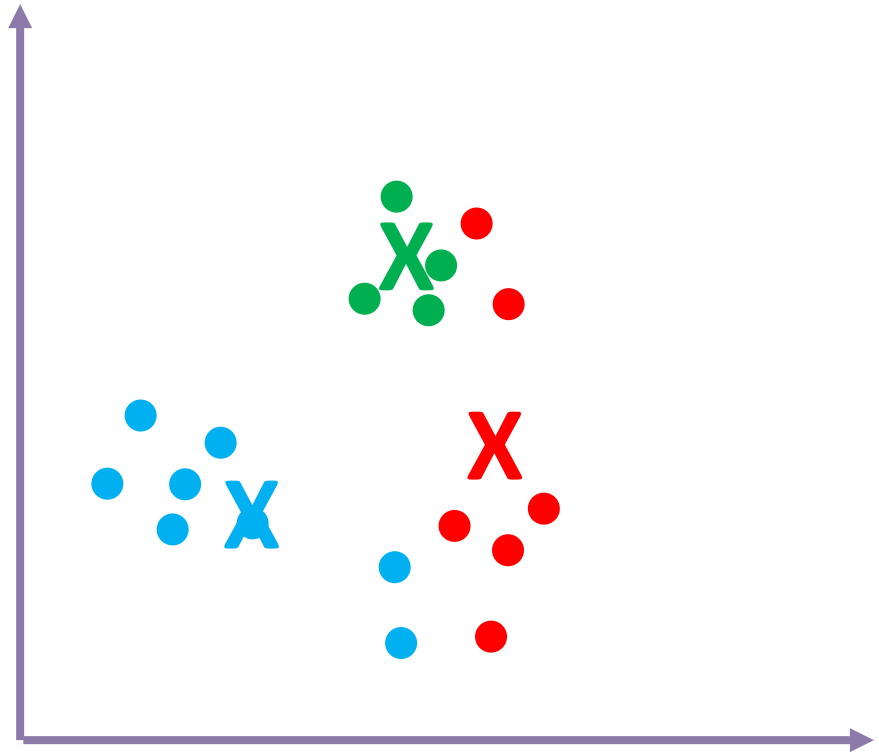


Illustration of K-means Clustering

- 1. Select *k* items at random from the data set as the initial cluster centers (let's set k=3)

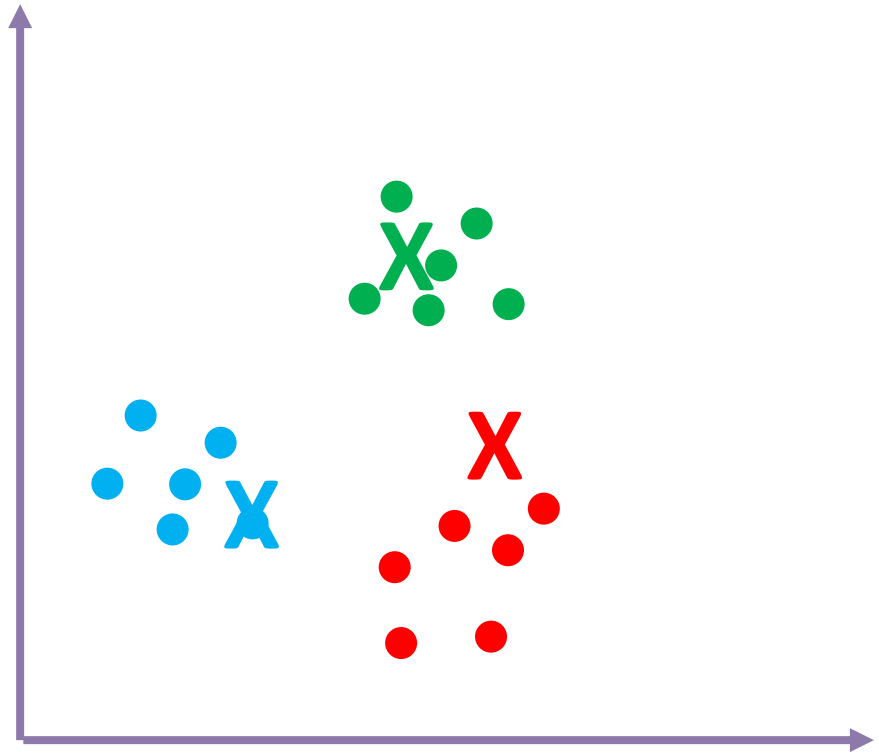
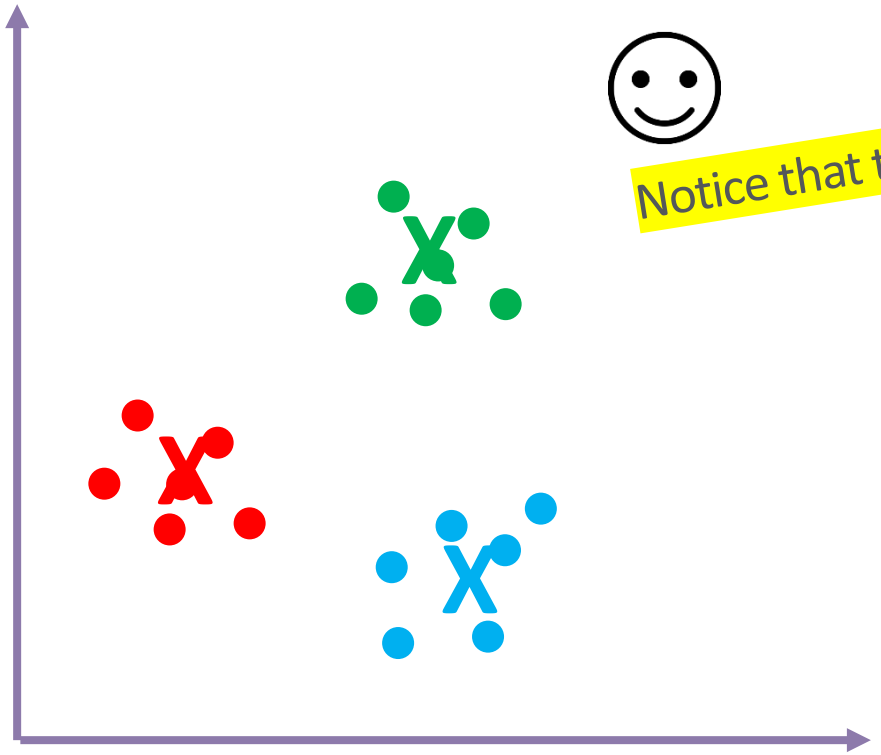


Illustration of K-means Clustering

- 3. Once all the data points are assigned, calculate the center of each cluster and use this as the new cluster center



Notice that the cluster membership is optimized!!

Illustration of K-means Clustering

- In reality, we do not know if there are three clusters
- We arbitrarily chose k = 3 based on the observation

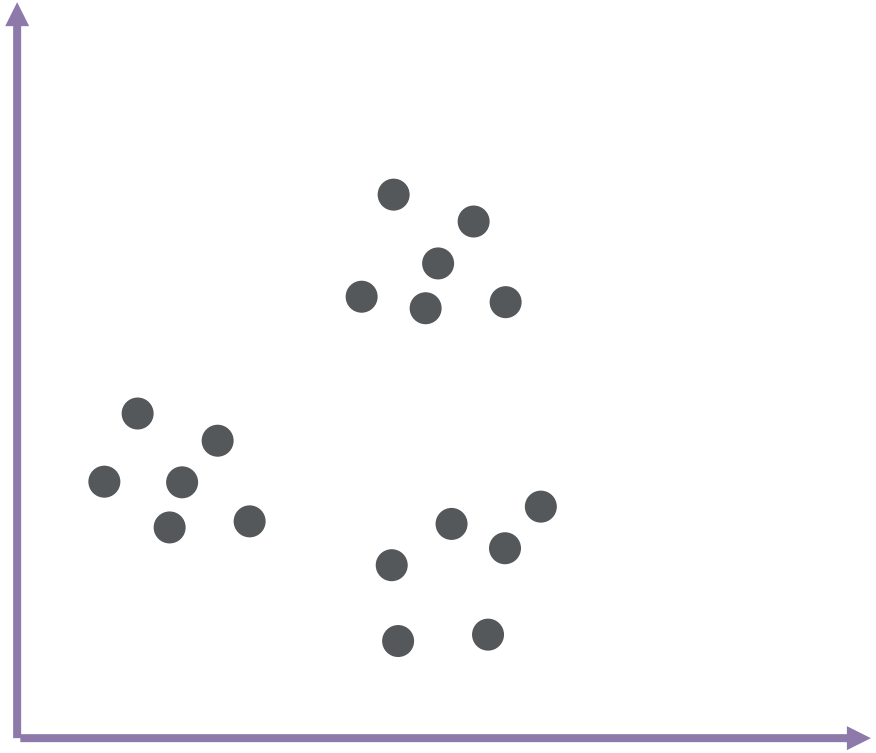
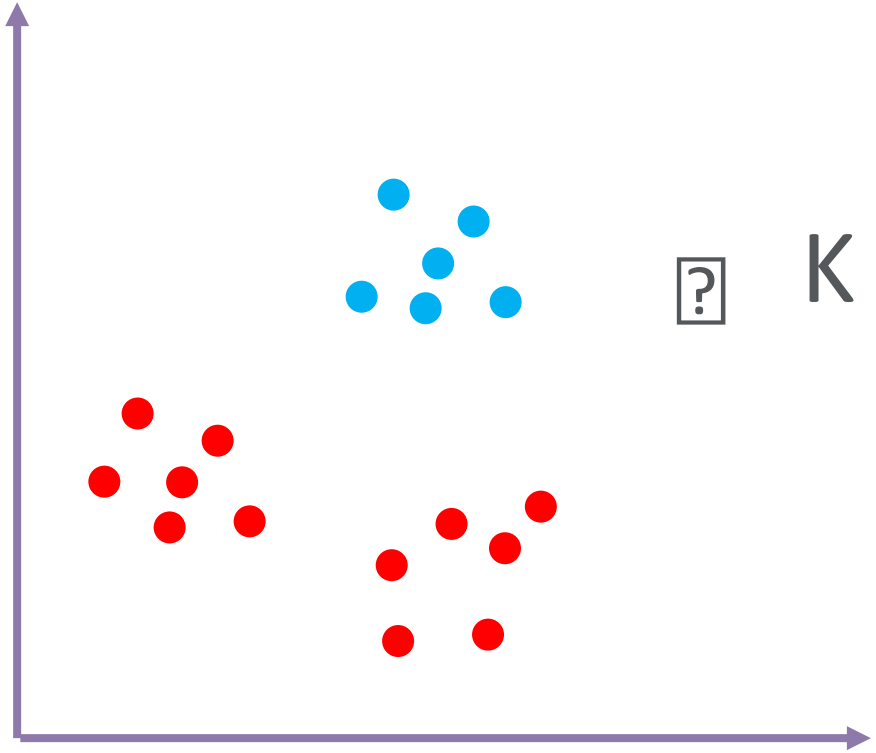


Illustration of K-means Clustering

- In reality, we do not know if there are three clusters
- We arbitrarily chose k = 3 based on the observation



? K may have been 2?

K-means clustering: How to choose K

- The most common method is to choose K by hand
- Use the elbow method

Clustering in R

- **stats** package (built-in)
  - hierarchical clustering (hclust, heatmap, cophenetic)
  - k-means (kmeans)
- **class** package
  - self-organizing maps (SOM)
- **mclust** package
  - EM / mixture models
- **clusterCons** package
  - consensus clustering

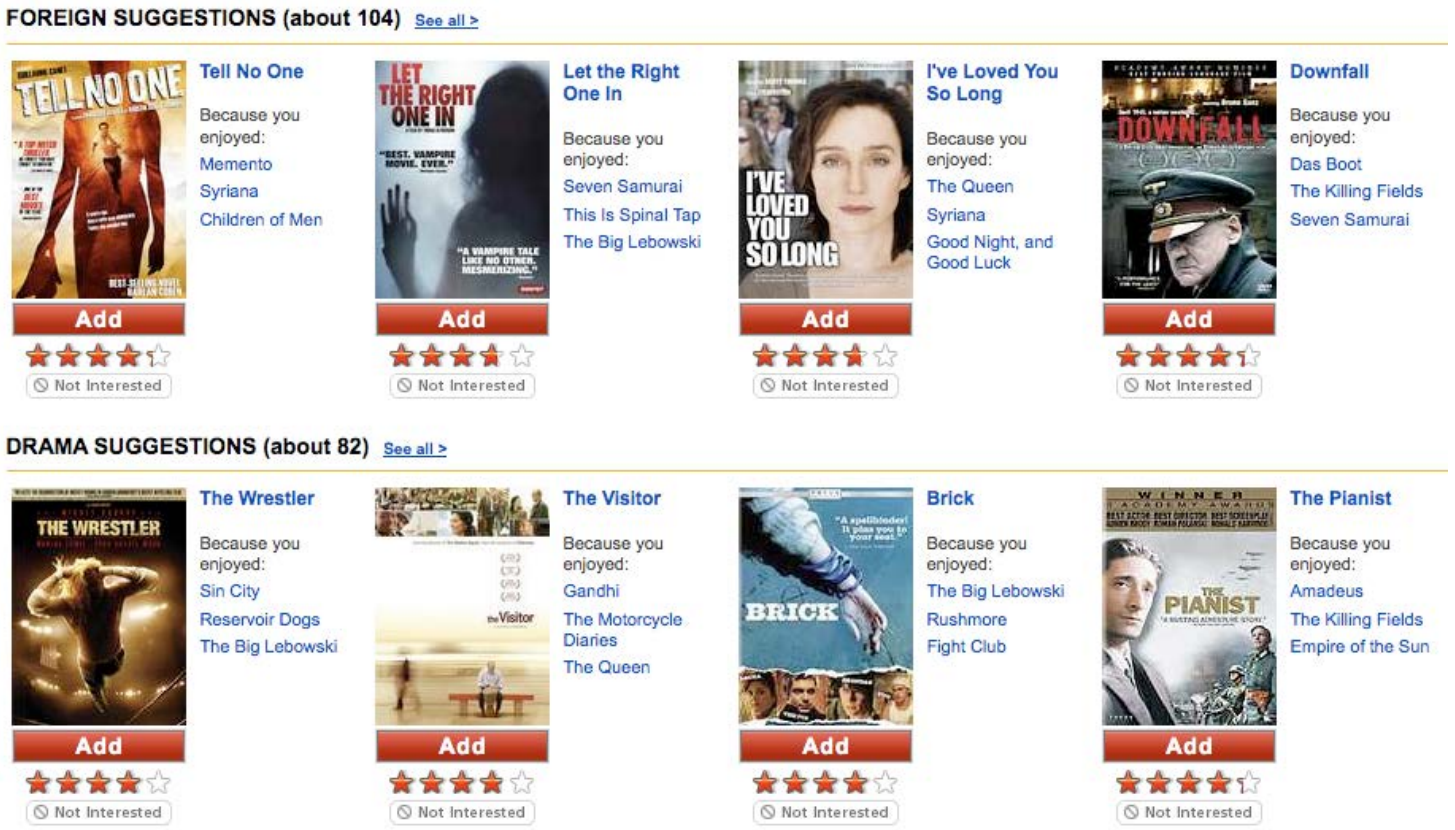
- **cluster** package
- AGglomerative NESting (agnes)
- DIvisive ANAlysis (diana)
- Fuzzy Analysis (fanny)
- Partitioning Around Medoids (pam)

Supervised Machine Learning

- **Goal**: Learn rules that can accurately **classify/predict** the sample characteristics from a sample's feature data

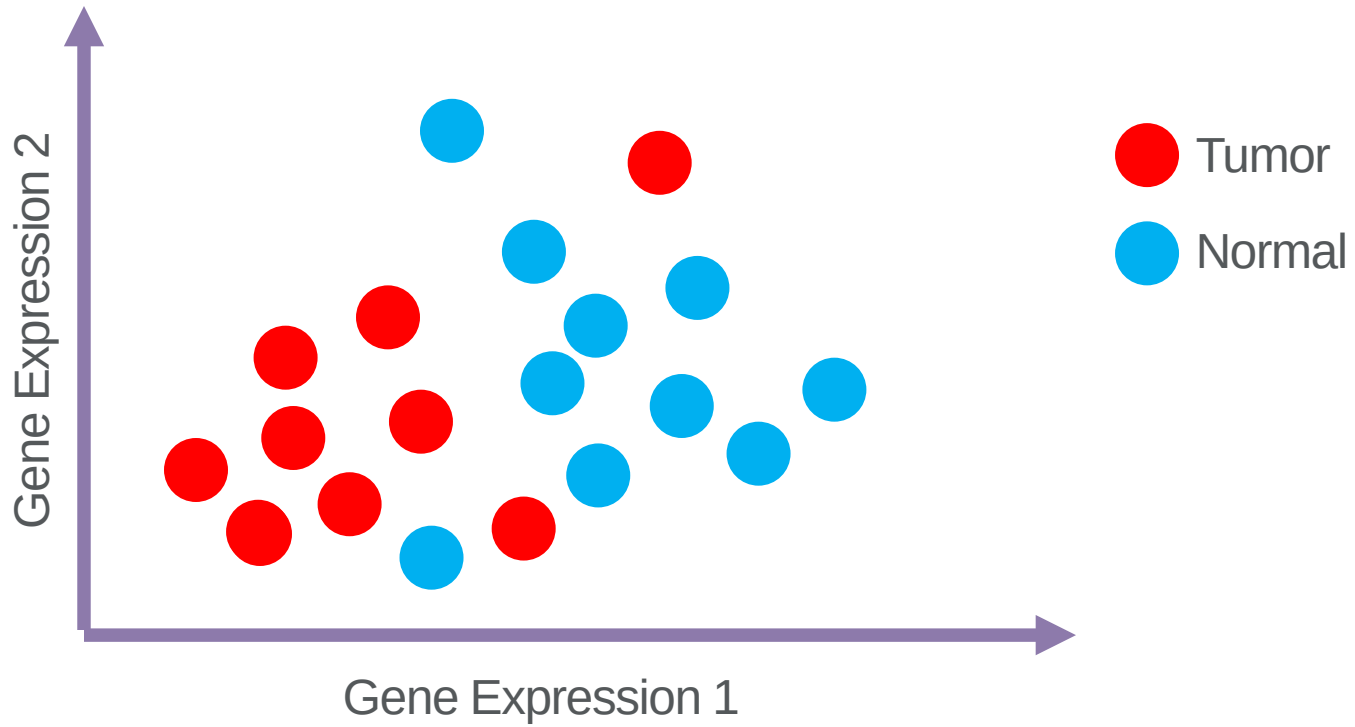
Netflix Recommendations

Because you enjoyed … you may enjoy…

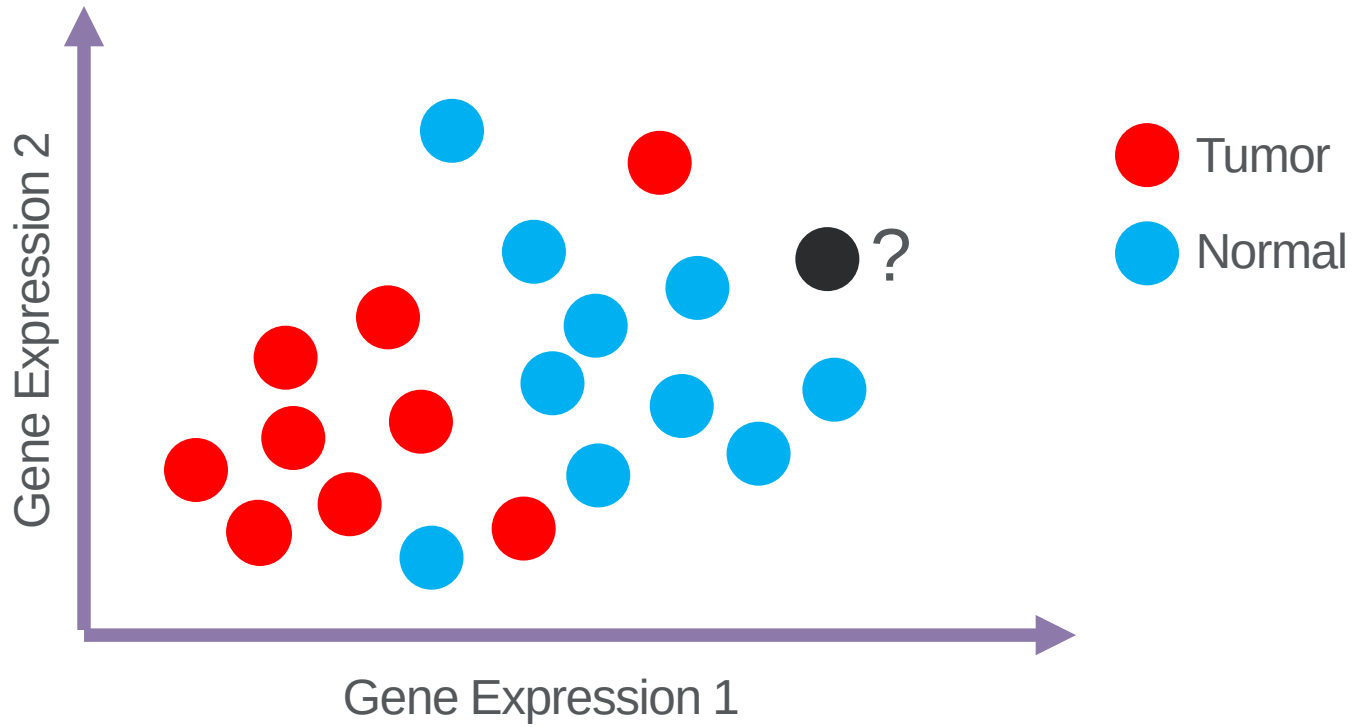


- Movie 1: romance
- Movie 2: thriller
- Movie 3: romance
- Movie 4: documentary
- Movie 5: documentary
- Movie 6: action

Supervised Learning Concept



Supervised Learning



Steps in Supervised Machine Learning

1. Pick a supervised learning algorithm
2. Select some training data
3. Train the machine
4. Test the accuracy of the machine with test data (not part of training data)

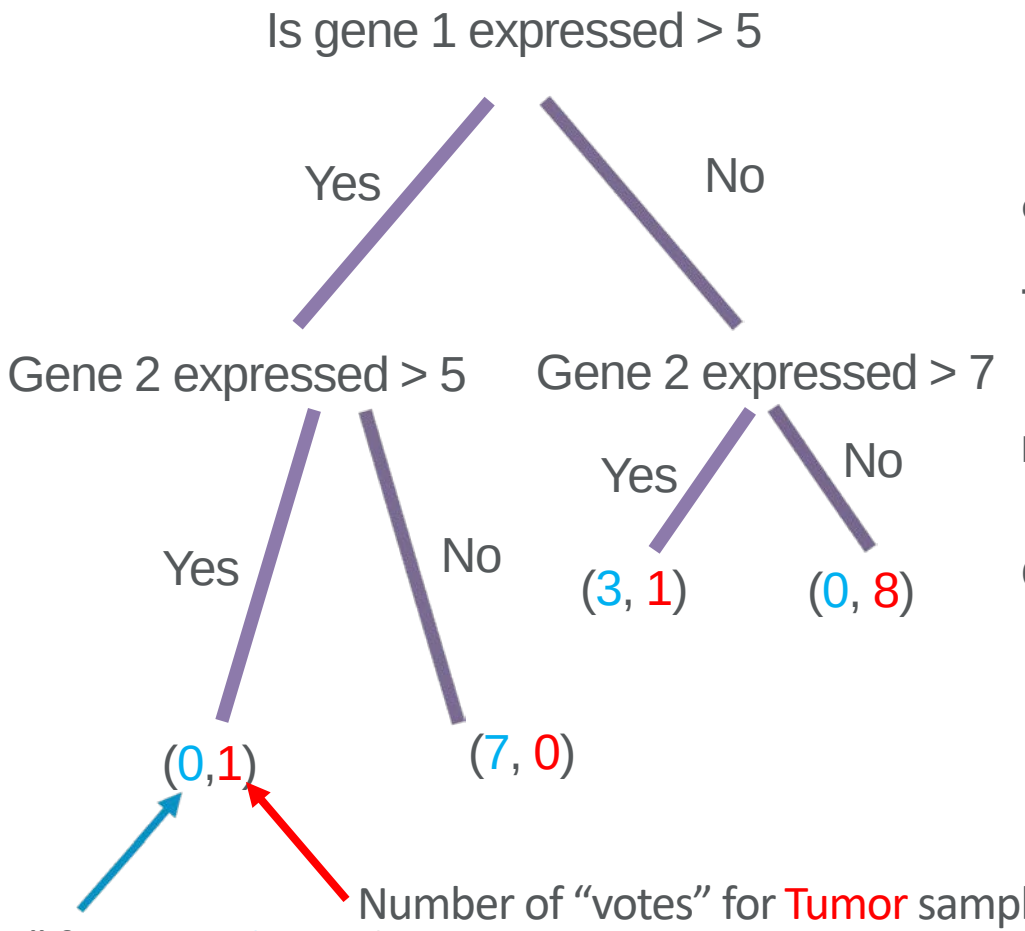
Assessing Accuracy: *K*-fold Cross-Validation

1. Break the samples into k blocks
2. Set one block aside for testing
3. Train on the other samples
4. Test on the samples in the testing block
5. Pick another one of the k blocks and repeat steps 2-4
6. Repeat step 5 until all blocks have been used for testing

Comment about Assessing Accuracy

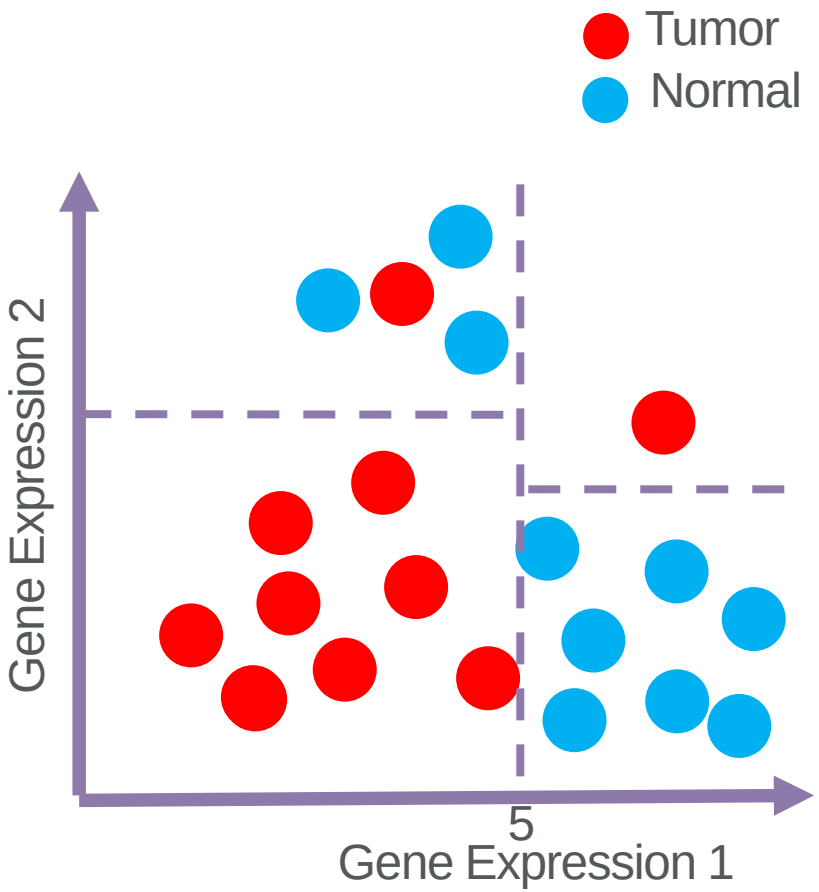
- The method is not a measure of generalizability
- It simply avoids “cheating"

Classification and Regression Tree (CART)



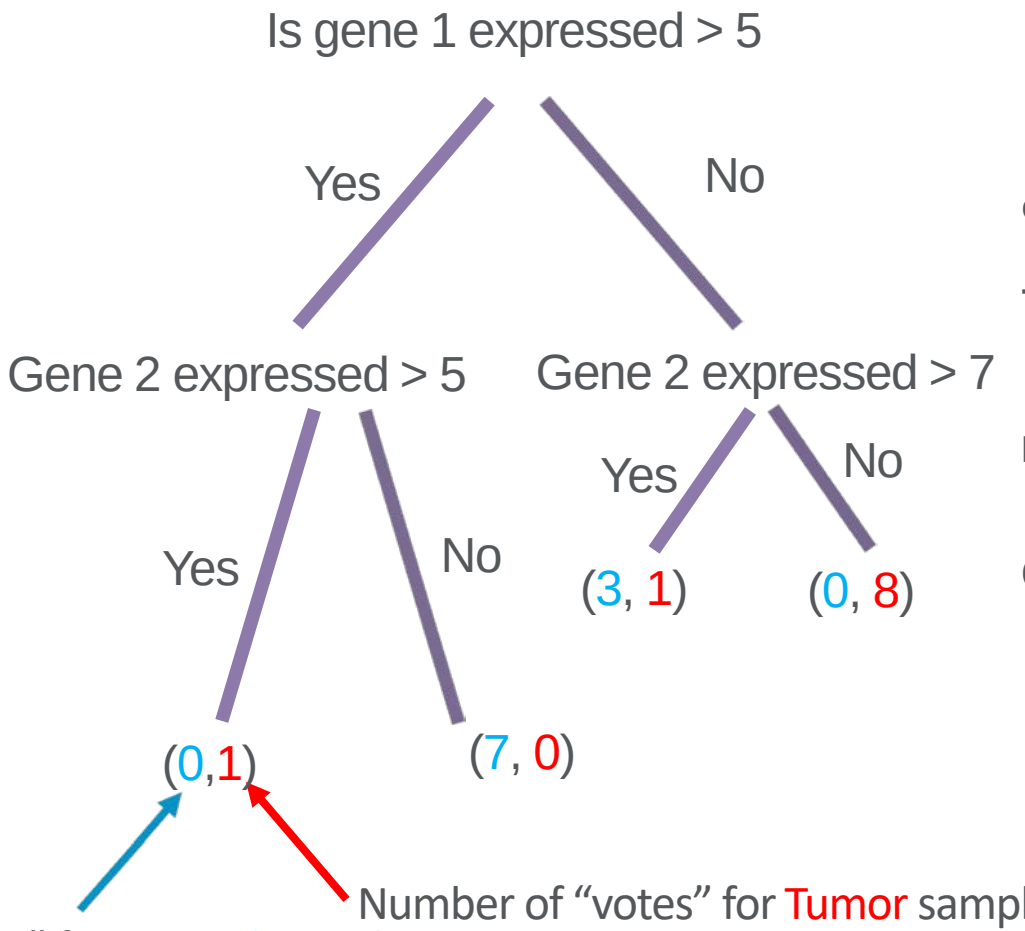
Number of "votes" for Tumor sample

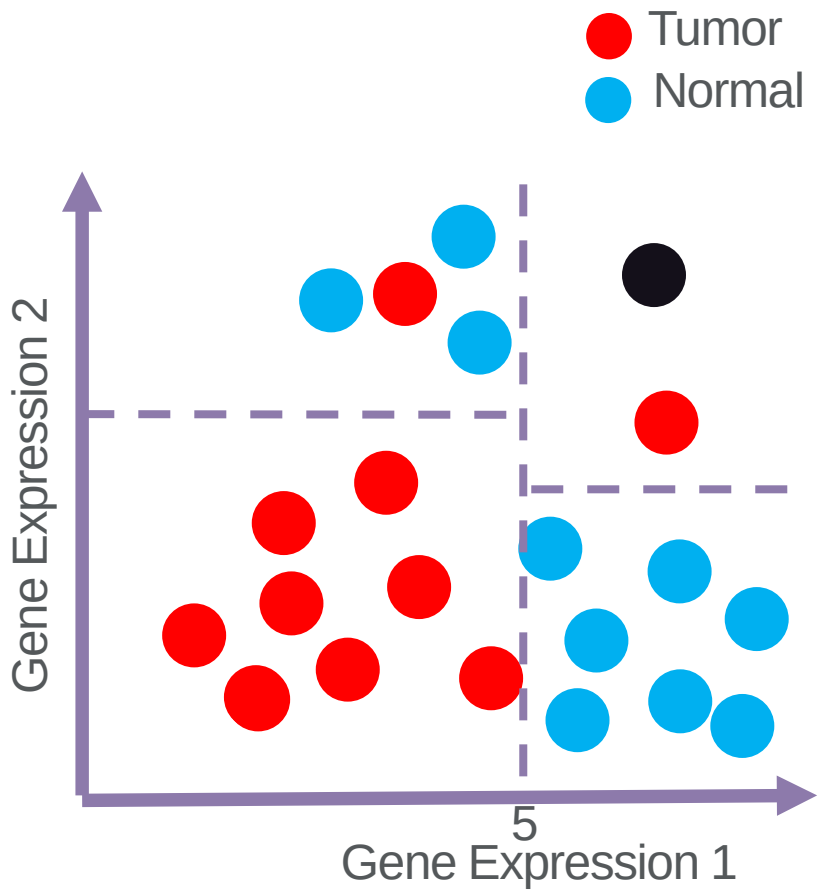
Number of "votes" for Normal sample



In R package **`rpart`**

Classification and Regression Tree (CART)





In R package **`rpart`**

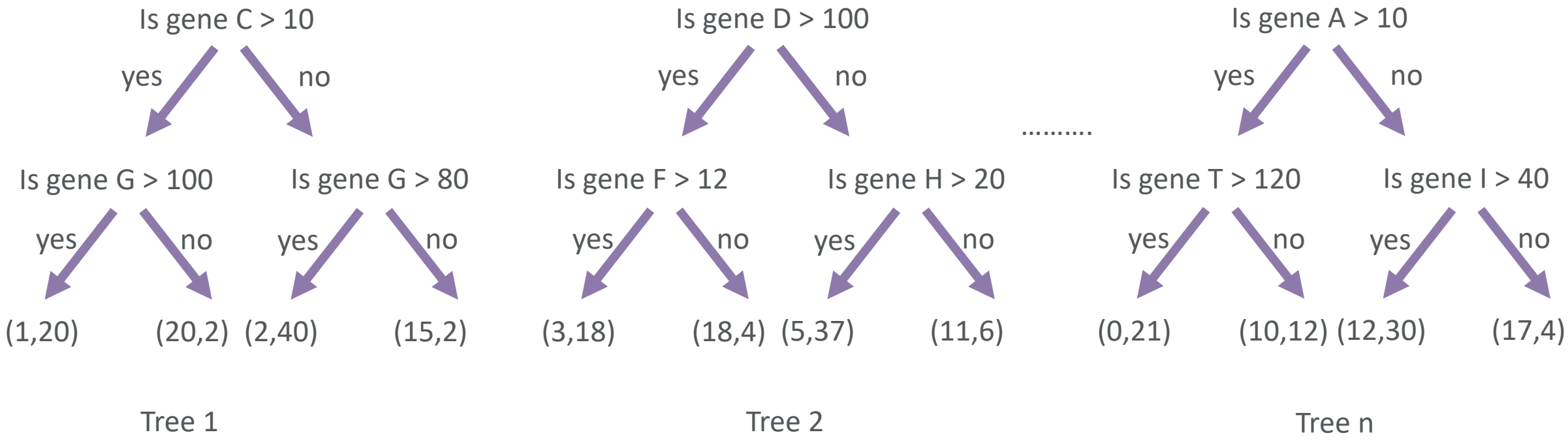
Random Forests Algorithm

- In random forests, we will construct many trees with bootstrap samples

1. For each tree, draw a random bootstrap sample of size N
2. Draw a random sample of m features. E.g. draw 10 features out of possible 1,000 features
3. Using the m features, split the node
4. Prediction of a new sample are the consensus of all the trees in the random forest

Many Trees

Supposed there are 1000 genes and 100 samples.
We will randomly sample 2 genes and 10 samples for each tree.



Recap

- Big data is complex and provide great opportunity
- Big data can be simplified using dimension reduction techniques
- Machine learning methods can be used for clustering and classification

BCC: Biostatistics Collaboration Center

Contact Us

- Request an Appointment
  - http://www.feinberg.northwestern.edu/sites/bcc/contact-us/request-form.html
- General Inquiries
  - bcc@northwestern.edu
  - 312.503.2288
- Visit Our Website
  - http://www.feinberg.northwestern.edu/sites/bcc/index.html

Biostatistics Collaboration Center |680 N. Lake Shore Drive, Suite 1400 |Chicago, IL 60611

Statistically Speaking: Upcoming Lectures

We hope to see you again!

RESCHEDULED

Monday, February 25

A Statistician's Guide to REDCap

Elizabeth Gray, MS, Statistical Analyst, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held from 1 to 2 pm in Baldwin Auditorium, Robert H. Lurie Medical Research Center, 303 E. Superior St.
http://www.feinberg.northwestern.edu/sites/bcc/education/lecture/2019.html

Your feedback is important to us! (And helps us plan future lectures).

Complete the evaluation survey to be entered in to a drawing to win 2 free hours of biostatistics consultation.