Northwestern Medicine®
Feinberg School of Medicine

Finding Signals in Big Data

Kwang-Youn A. Kim, PhD
Assistant Professor, Department of Preventive Medicine
Biostatistics Collaboration Center
kykim@northwestern.edu

BCC: Biostatistics Collaboration Center

Who We Are

•Leah J. Welty, PhD
•Assoc. Professor
•BCC Director

•Joan S. Chmiel, PhD
•Professor

•Jody D. Ciolino, PhD
•Asst. Professor

•Kwang-Youn A. Kim, PhD
•Asst. Professor

•Masha Kocherginsky, PhD
•Assoc. Professor

•Mary J. Kwasny, ScD
•Assoc. Professor

•Julia Lee, PhD, MPH
•Assoc. Professor

•Alfred W. Rademaker, PhD
•Professor

•Hannah L. Palac, MS
•Senior Stat. Analyst

•Gerald W. Rouleau, MS
•Stat. Analyst

•Amy Yang, MS
•Senior Stat. Analyst

•Not Pictured:
•1. David A. Aaby, MS
•Senior Stat. Analyst

•2. Tameka L. Brannon
•Financial | Research Administrator

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

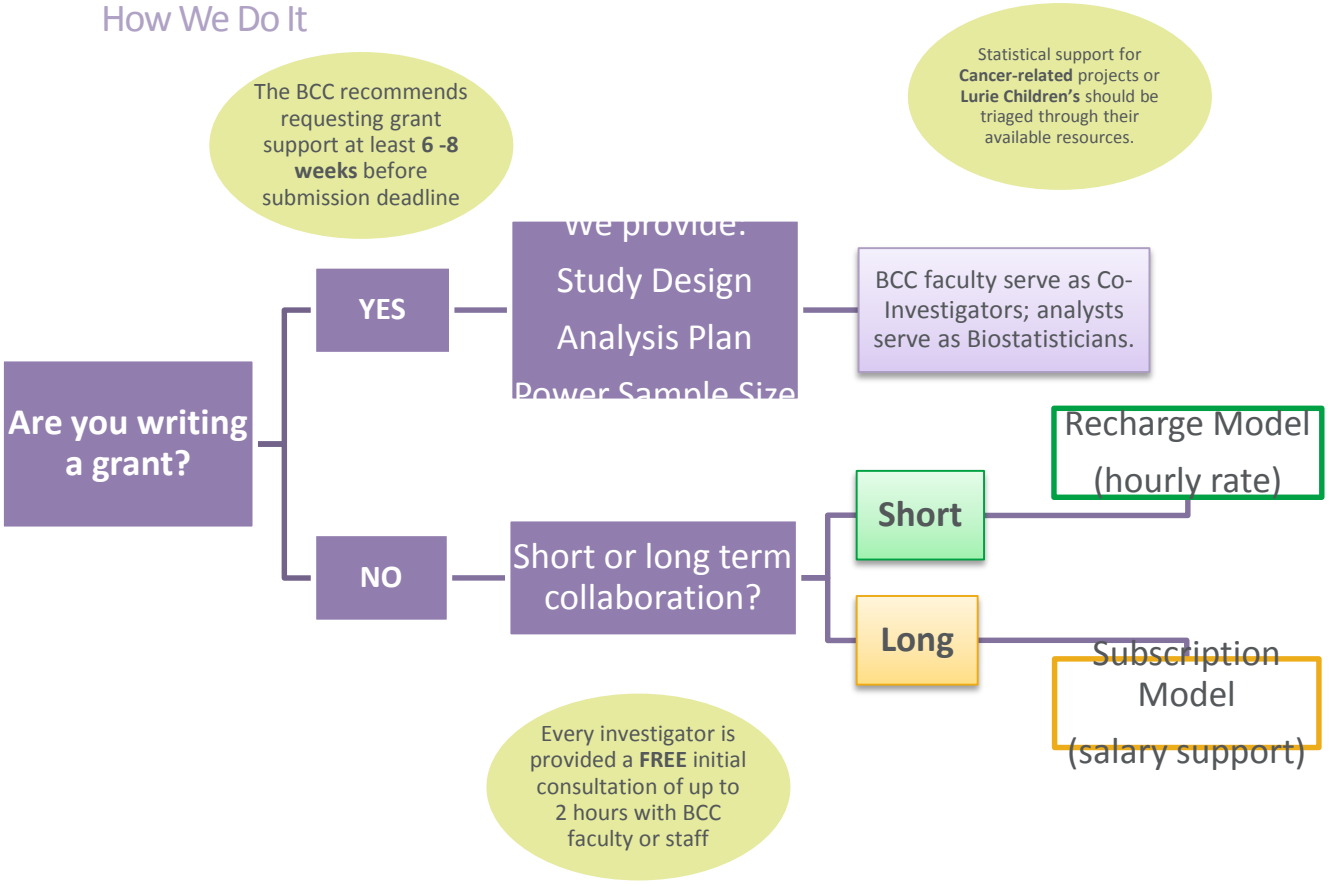
BCC: Biostatistics Collaboration Center

What We Do

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

BCC: Biostatistics Collaboration Center

How We Do It



BCC: Biostatistics Collaboration Center

How can you contact us?

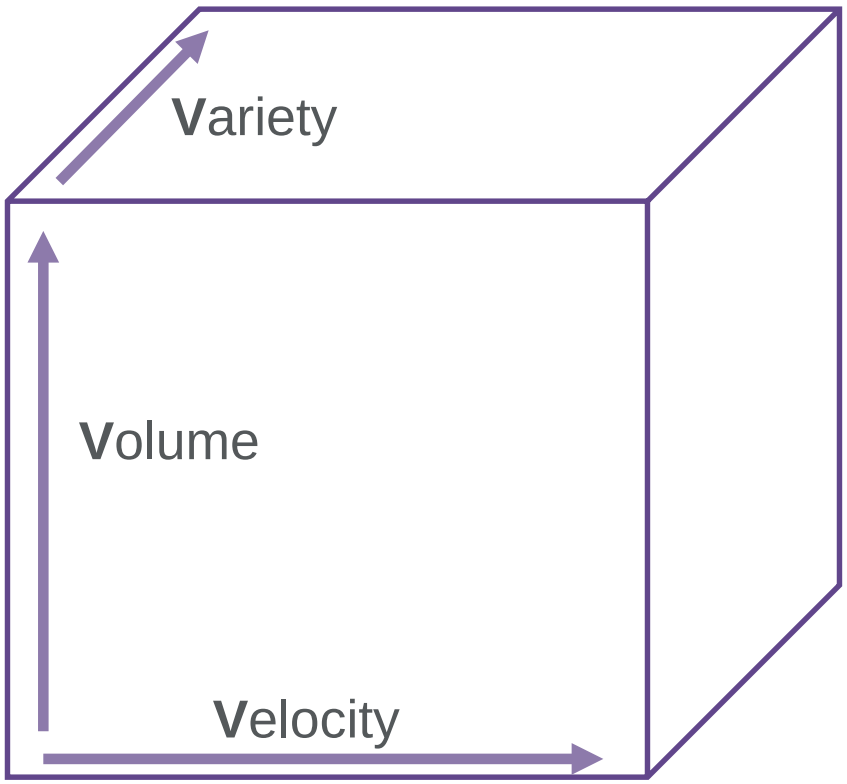
- Request an Appointment
  - http://www.feinberg.northwestern.edu/sites/bcc/contact-us/request-form.html
- General Inquiries
  - bcc@northwestern.edu
  - 312.503.2288
- Visit Our Website
  - http://www.feinberg.northwestern.edu/sites/bcc/index.html

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

Topic for Today

Statistical Methods in Medical Research Involving Big Data

What is Big Data?



- New paradigm and an ecosystem that transforms case-based studies to large-scale, data-driven research.
- "-omics" sequencing datasets: genomics, proteomics, metabolomics, phenomics
- Unstrucutred datasets: Notes from EHRs, medical images, sensor data
- Social media data

Biomedical Big Data is…

- More than just very large data or a large number of data sources. Big Data refers to the complexity, challenges, and new opportunities presented by the combined analysis of data. In biomedical research, these data sources include the diverse, complex, disorganized, massive, and multimodal data being generated by researchers, hospitals, and mobile devices around the world.
- Diverse and complex. It includes imaging, phenotypic, molecular, exposure, health, behavioral, and many other types of data. These data could be used to discover new drugs or to determine the genetic and environmental causes of human disease.
- Faces many challenges. The unwieldy amount of information, lack of organization and access to data and tools, and insufficient training in data science methods make it difficult for Big Data's full power to be harnessed.
- Provides spectacular opportunities. Big Data methods allow researchers to maximize the potential of existing data and enable new directions for research. Biomedical Big Data can increase accuracy and supports the development of precision methods for healthcare.

Source: https://datascience.nih.gov/bd2k/about/what

Precision Medicine

- Precision Medicine Initiative (PMI)
- Engage a group of >1 million participants VOLUME
- Share biological samples, genetic data and diet/lifestyle information, all linked to their electronic health records VARIETY
- The PMI Cohort Program will be a participant-engaged, data-driven enterprise supporting research at the intersection of human biology, behavior, genetics, environment, data science and computation, and much more to produce new knowledge with the goal of developing more effective ways to prolong health and treat disease.

Illinois Precision Medicine Consortium

DESCRIPTION | DETAILS | RESULTS | HISTORY | SUBPROJECTS | SIMILAR PROJECTS | NEARBY PROJECTS BETA | LINKS | NEWS AND MORE

Project Number: 1UG3OD023189-01
Title: ILLINOIS PRECISION MEDICINE CONSORTIUM (IPMC)

Contact PI / Project Leader: GREENLAND, PHILIP
Awardee Organization: NORTHWESTERN UNIVERSITY AT CHICAGO

Contact PI / Project Leader Information:

Name: GREENLAND, PHILIP
Email: Click to view Contact PI / Project Leader email address
Title: PROFESSOR AND CHAIR

Program Official Information:

Name: REIS, JARED P
Email: Click to view PO email address

Other PI Information:

AHSAN, HABIBUL
WINN, ROBERT ANDREW

Profile Exists | No Profile

Organization:

Name: NORTHWESTERN UNIVERSITY AT CHICAGO
City: CHICAGO **Country:** UNITED STATES (US)

Department/ Organization Type:

PUBLIC HEALTH & PREV MEDICINE
SCHOOLS OF MEDICINE

Congressional District:

State Code: IL
District: 07

Other Information:

FOA: RFA-PM-16-002
Study Section: Special Emphasis Panel [ZRG1-PSE-N (51)R]
Fiscal Year: 2016 **Award Notice Date:** 6-JUL-2016

DUNS Number: 005436803
Project Start Date: 6-JUL-2016
Budget Start Date: 6-JUL-2016

CFDA Code: 310
Project End Date: 30-JUN-2021
Budget End Date: 30-JUN-2017

Administering Institutes or Centers:

OFFICE OF THE DIRECTOR, NATIONAL INSTITUTES OF HEALTH

Project Funding Information for 2016:

Total Funding: $5,301,304 **Direct Costs:** $4,739,859 **Indirect Costs:** $561,445

Year	Funding IC	FY Total Cost by IC
2016	OFFICE OF THE DIRECTOR, NATIONAL INSTITUTES OF HEALTH	$5,301,304

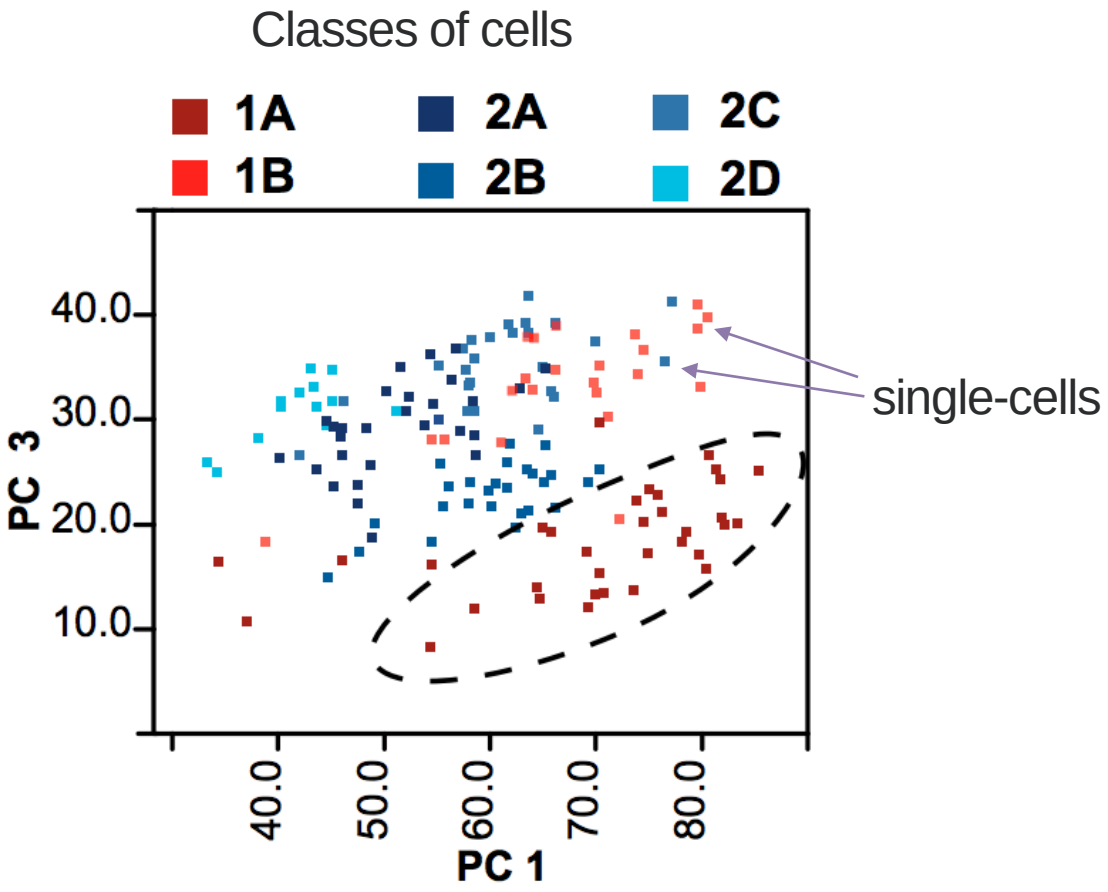
Mild Introduction to Statistics

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

Terminology

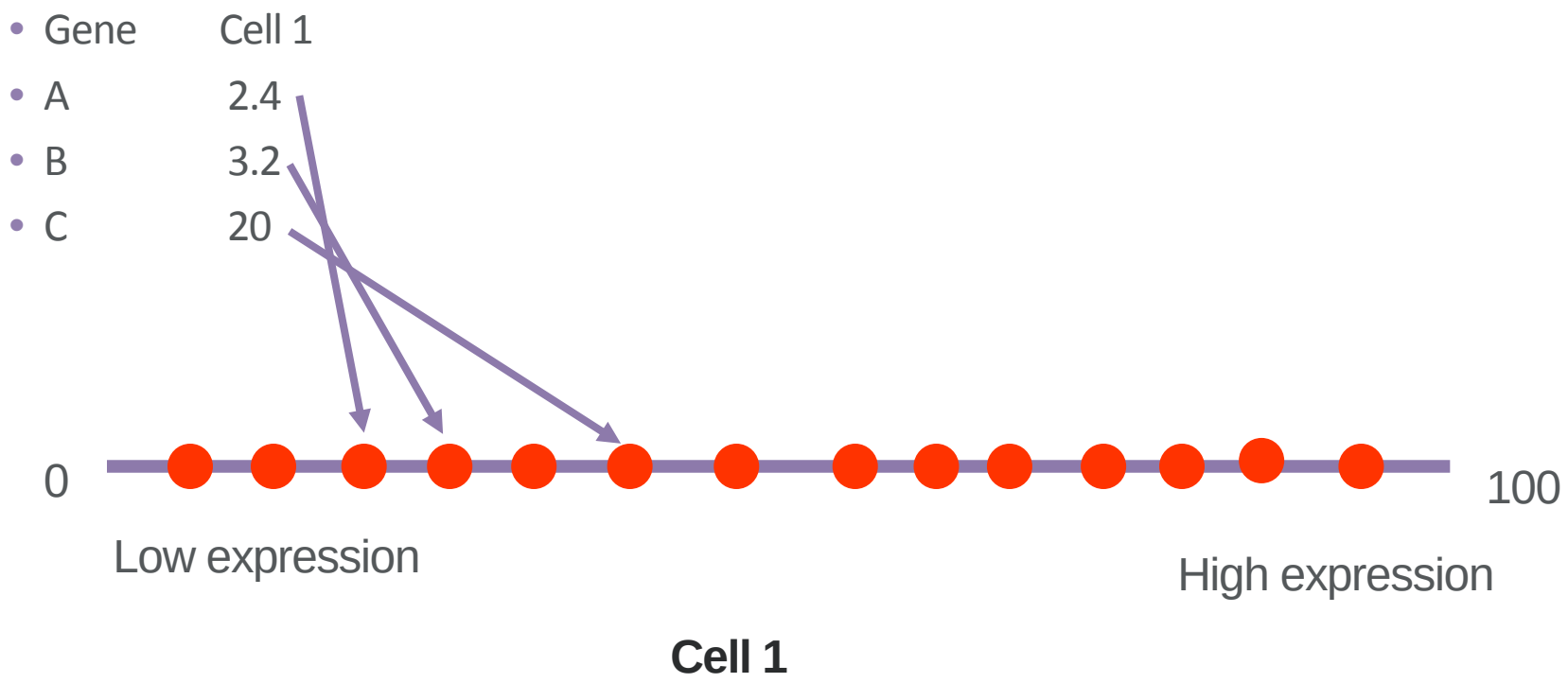
- **Sample**: An object we have data for (e.g. a study participant)
- **Feature**: A variable measured in our sample (e.g. gene expression for gene A)
- **Class**: A characteristic of the sample that is not a feature (e.g. death status)
- **Machine learning**: A broad category of techniques devoted to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

Dimension Reduction



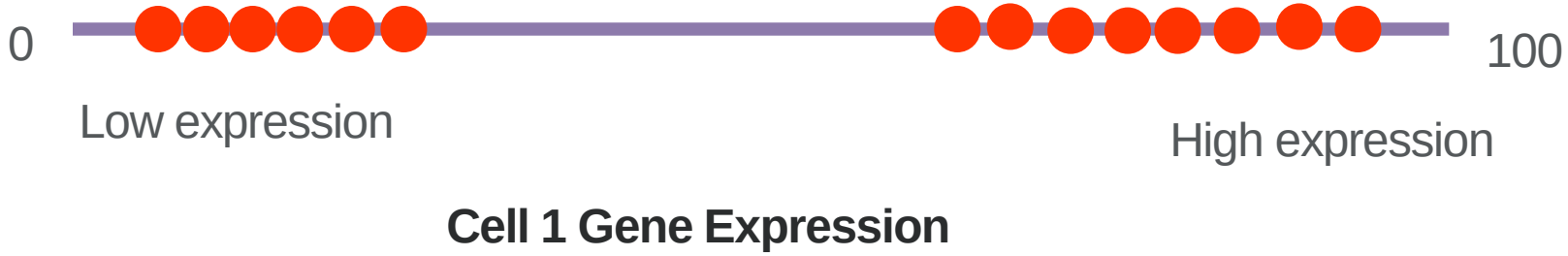
Poulin et al. 2014

Let’s start with only 1 dimension



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

In summary…

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

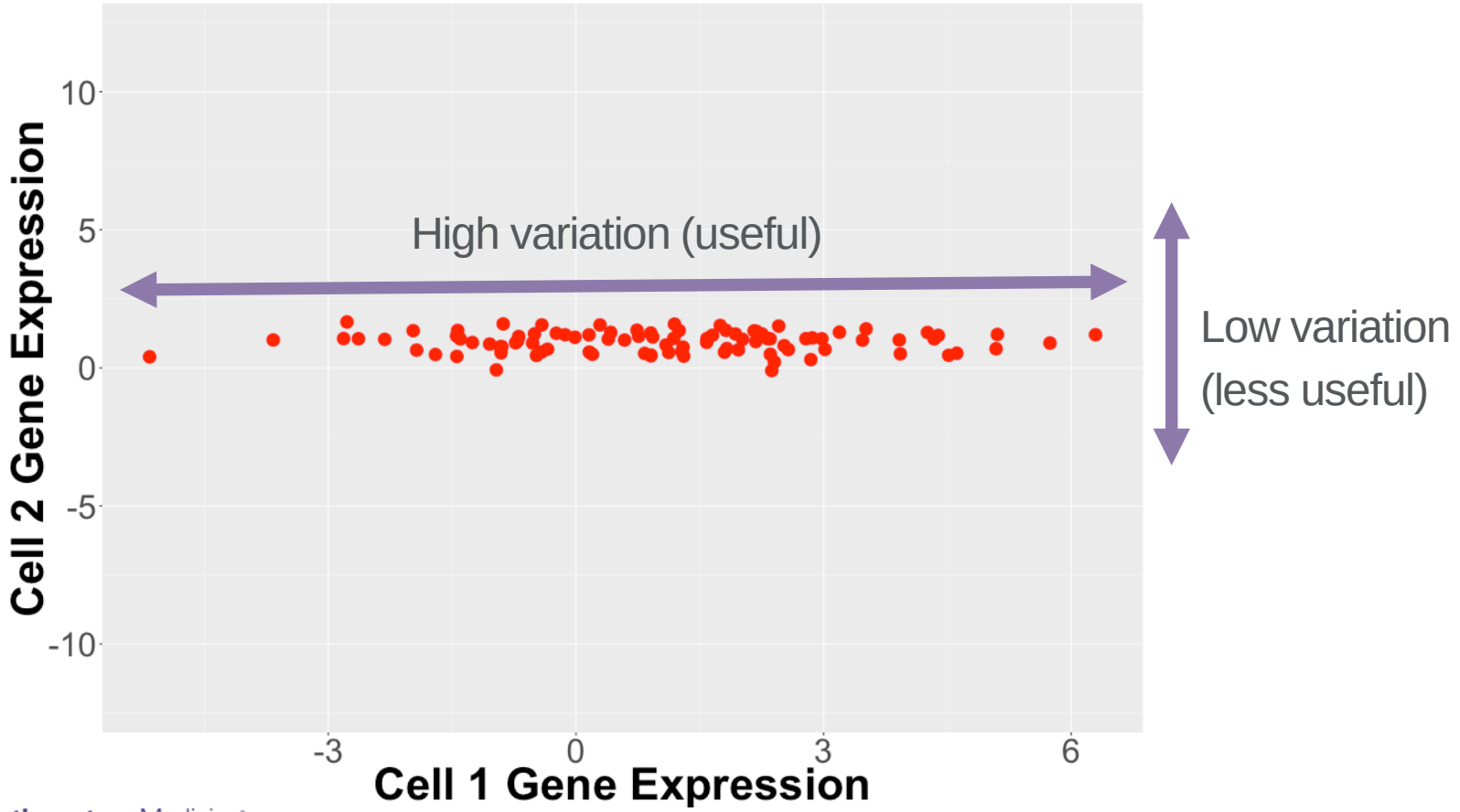
- How can we draw N-dimensional graph? You CAN'T!!

Not all dimensions are created equally

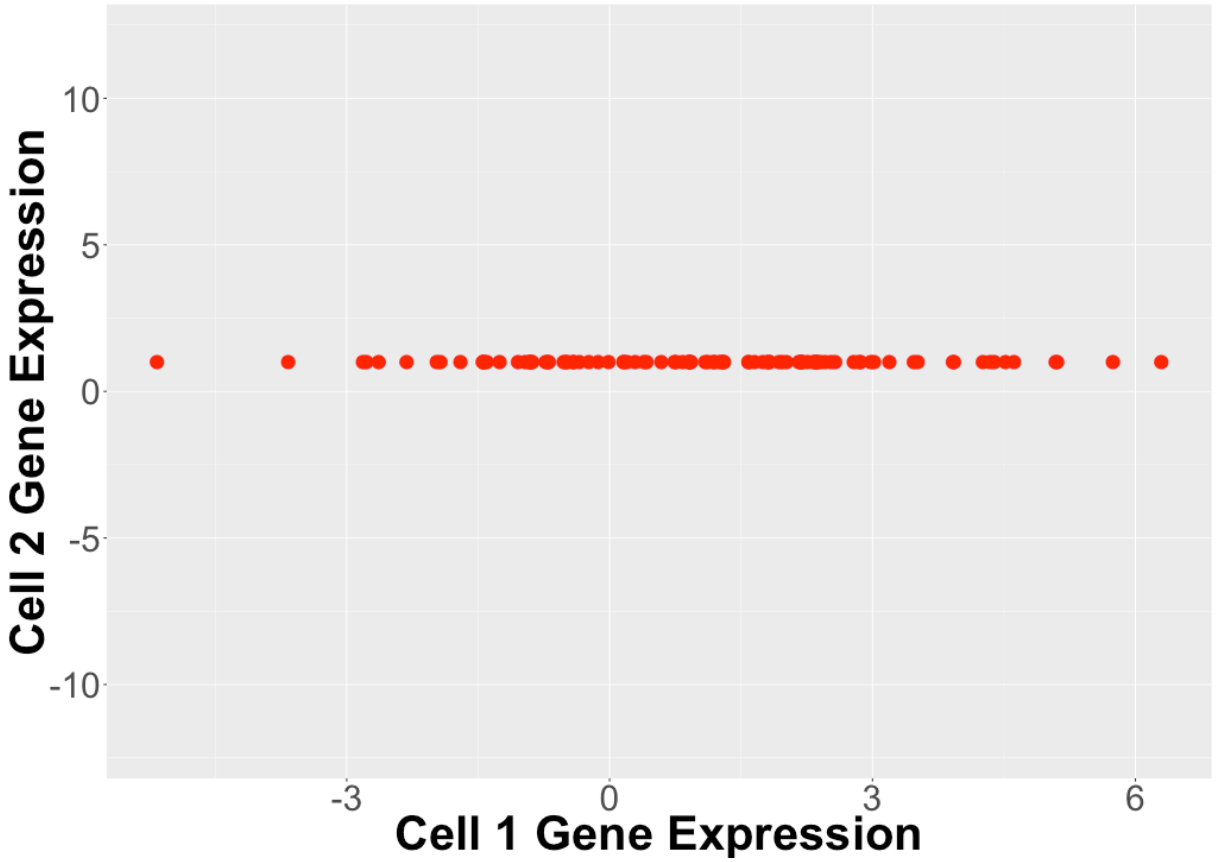
- Are all dimensions equally important?

Not all dimensions are created equal

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

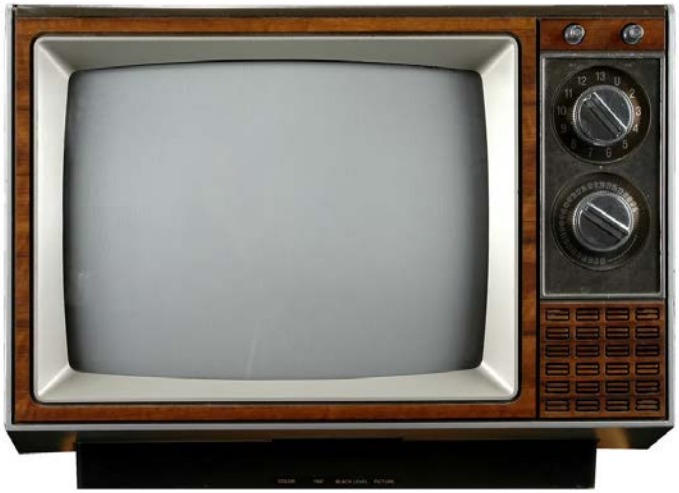


Dimension Reduction of Cell 2



Analogy

1 Dimension



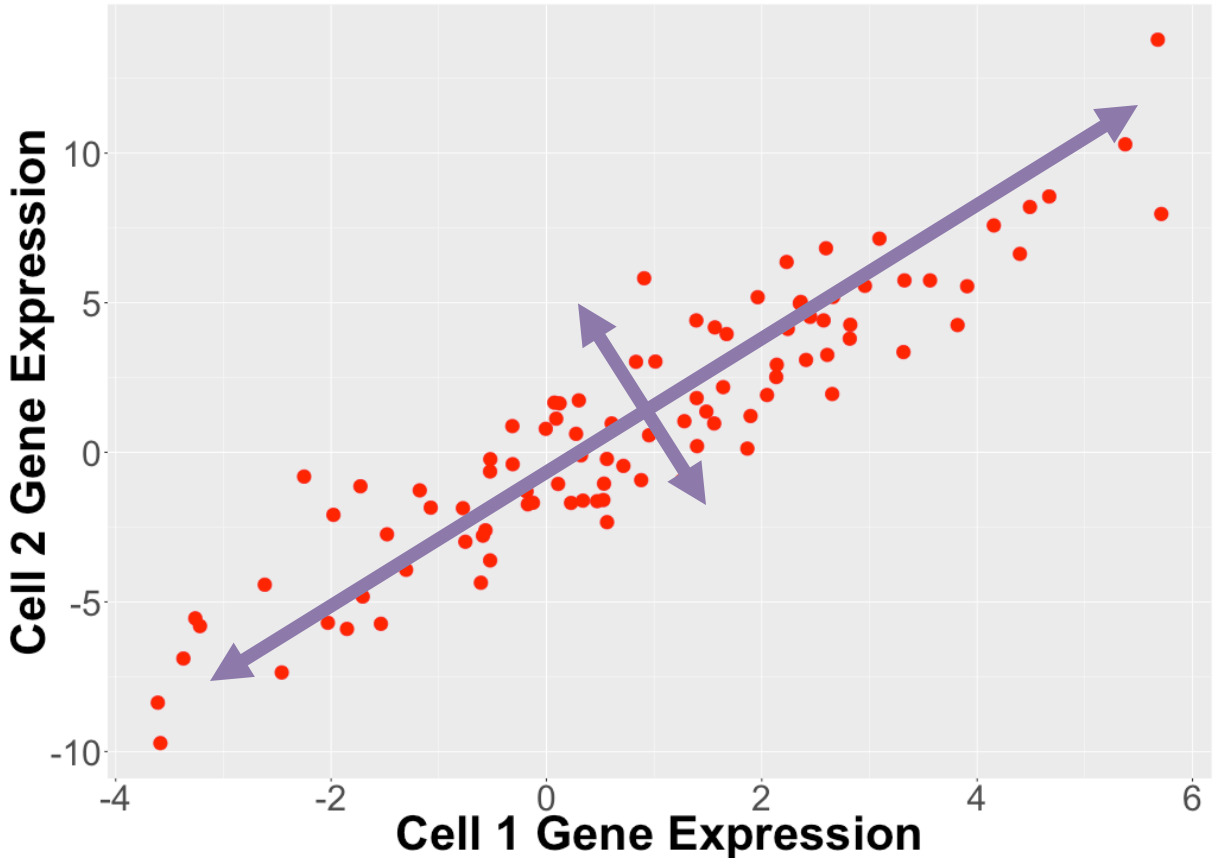
2 Dimensions

3 Dimensions

Principal Components Analysis (PCA)

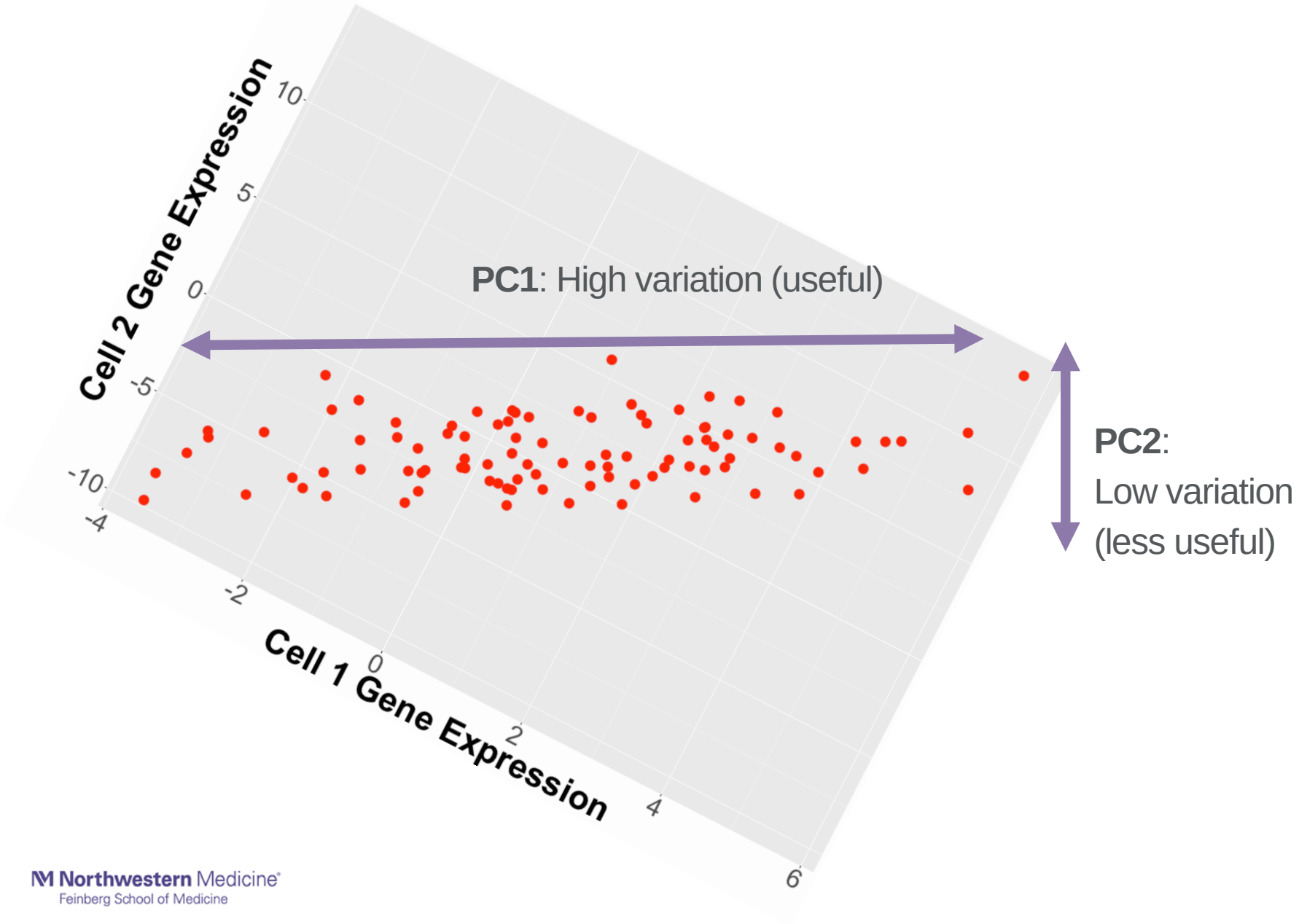
- Flattens the data without losing much information
- Goal is to find the important dimensions
- E.g. Using information of many cells, reduce it to a few dimensions that we can visualize

Now onto 2D



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

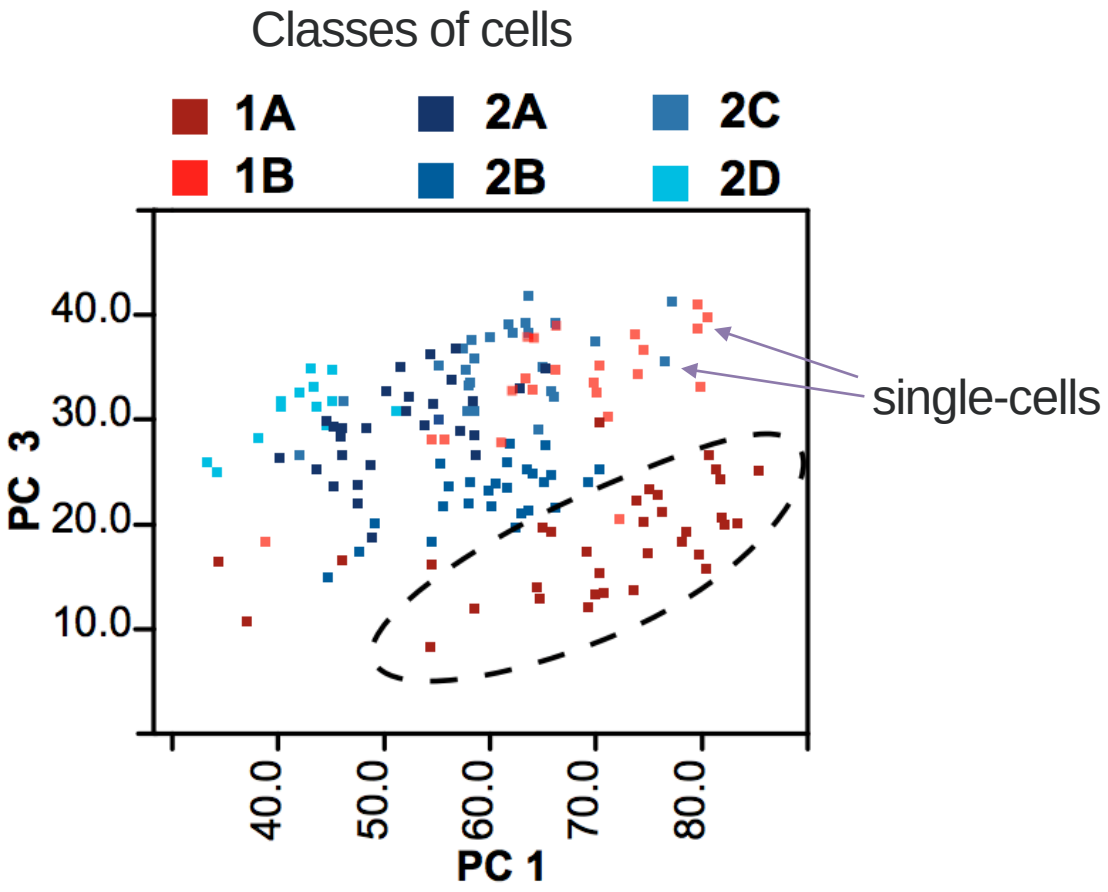
PCA Rotation



In summary

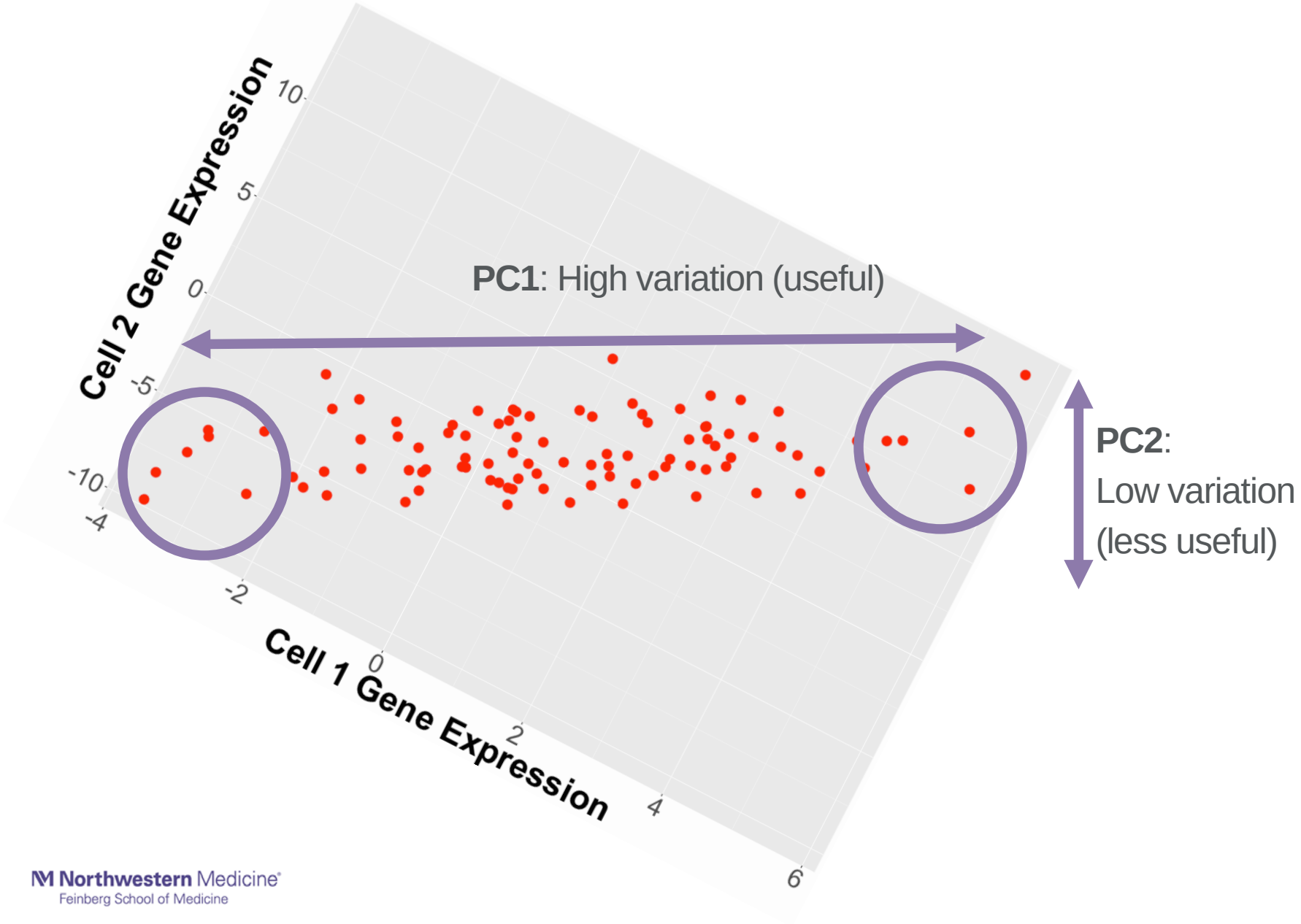
- If we have 2 cells
  - PC1 spans in the direction that captures the most variation
  - PC2 spans in the direction that captures the 2nd most variation
- If we have N cells
  - PC1 spans in the direction that captures the most variation
  - PC2 spans in the direction that captures the 2nd most variation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

Dimension Reduction



Poulin et al. 2014

PCA Rotation



PCA Procedure

Gene	Cell 1	Influence on PC1 (Loadings)
A	-2.1	Low (.1)
B	1.2	Low (.2)
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
…	…	…

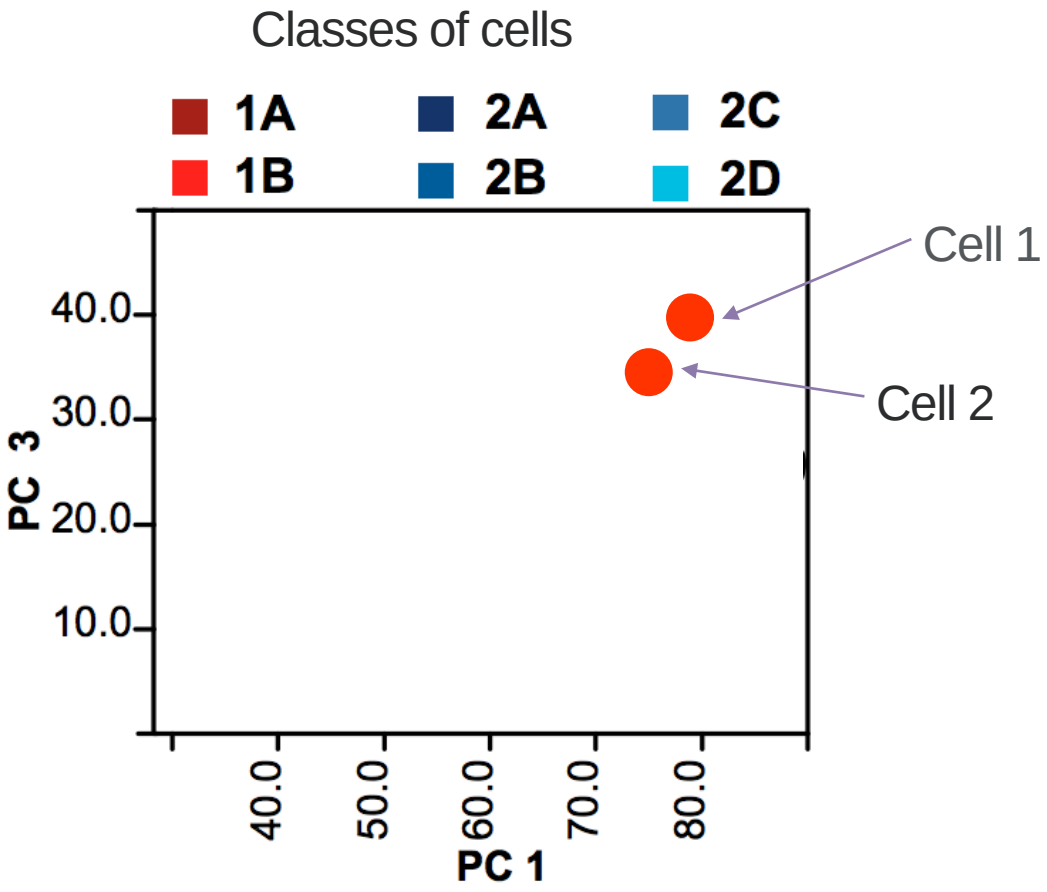
Cell 1 PC1 score = -2.1*0.1 + 1.2*0.2 + … = some value 1

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

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

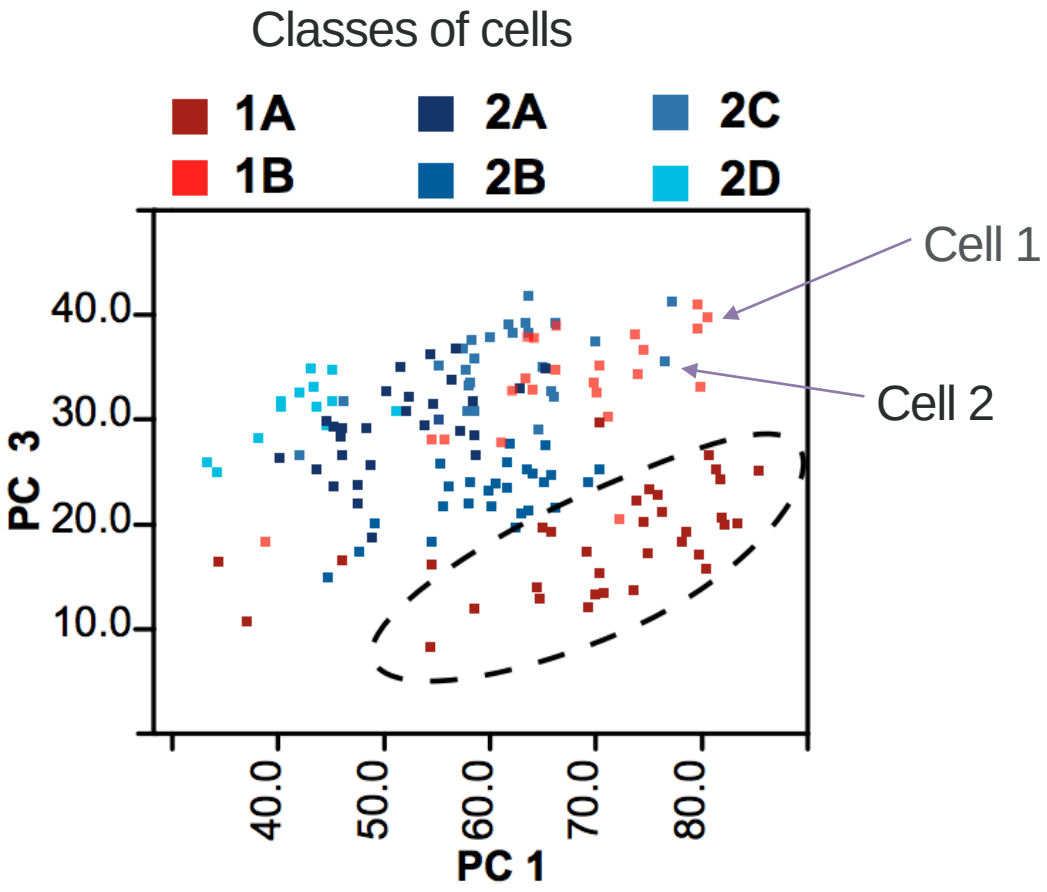
Cell 2 PC2 score = similar strategy with PC2 loadings for cell 2 = some value

Dimension Reduction



Poulin et al. 2014

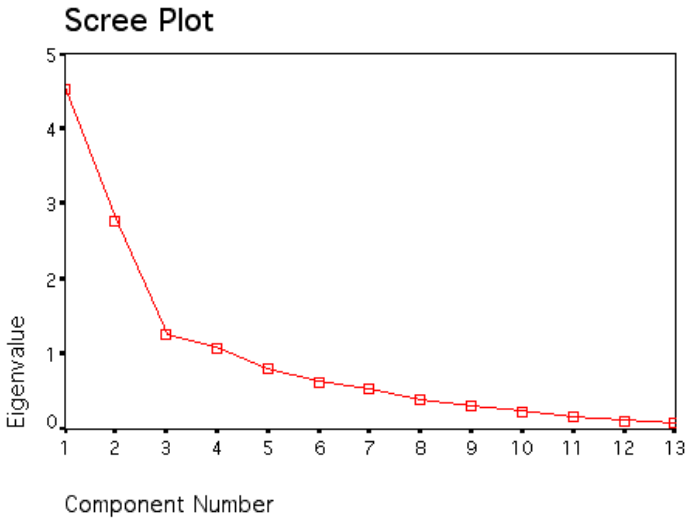
Dimension Reduction



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



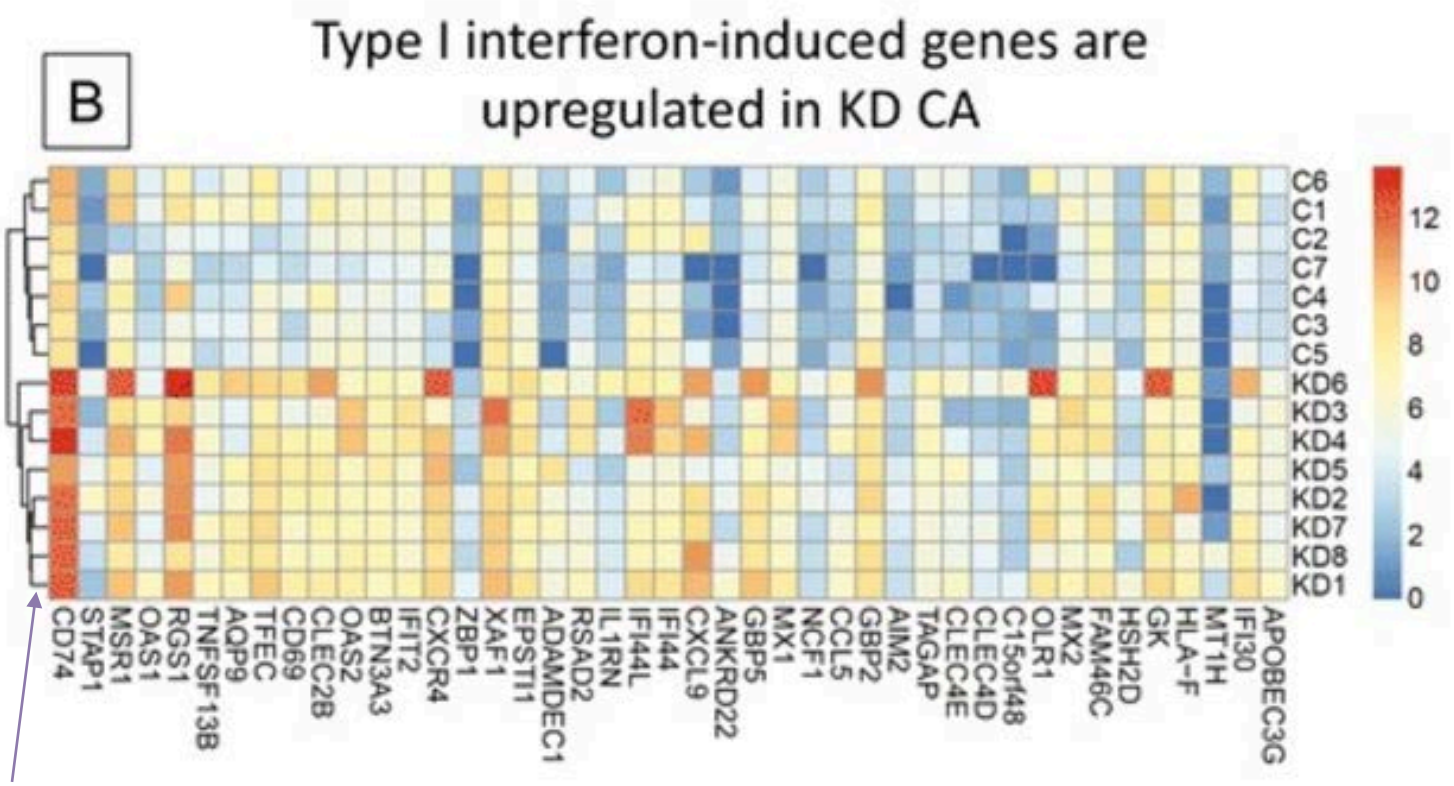
Machine Learning

- 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`

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

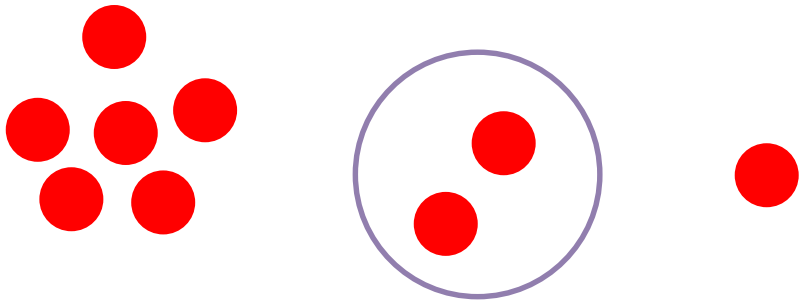


Distance of dissimilarity

Rowley et al. 2015

Agglomerative Hierarchical Clustering Explained

- Compute distances between all pairs of items
- Merge clusters according to the smallest distance between any pair of elements in the two clusters
- Continue until all clusters are merged



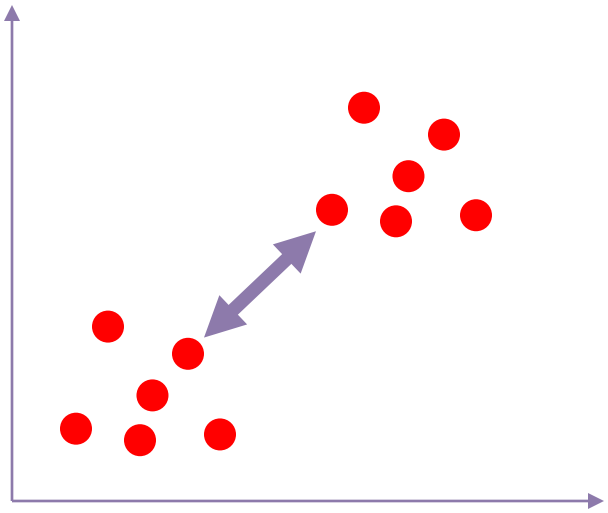
Distance Metric

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

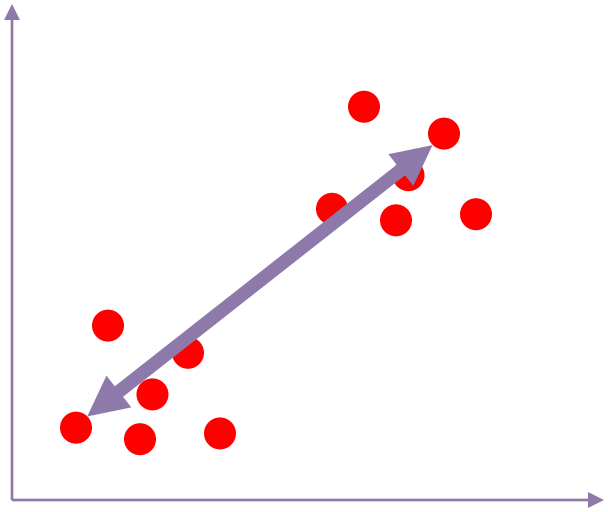
Linkage

- Nearest Neighbor (Single Linkage)



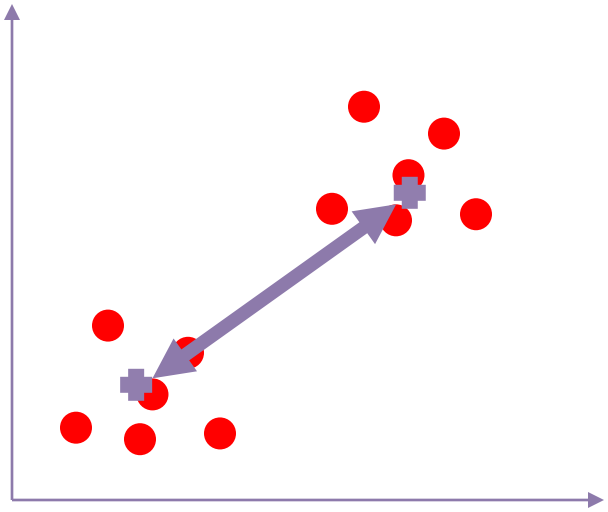
Linkage

- Furthest Neighbor (Complete Linkage)



Linkage

- Centroid



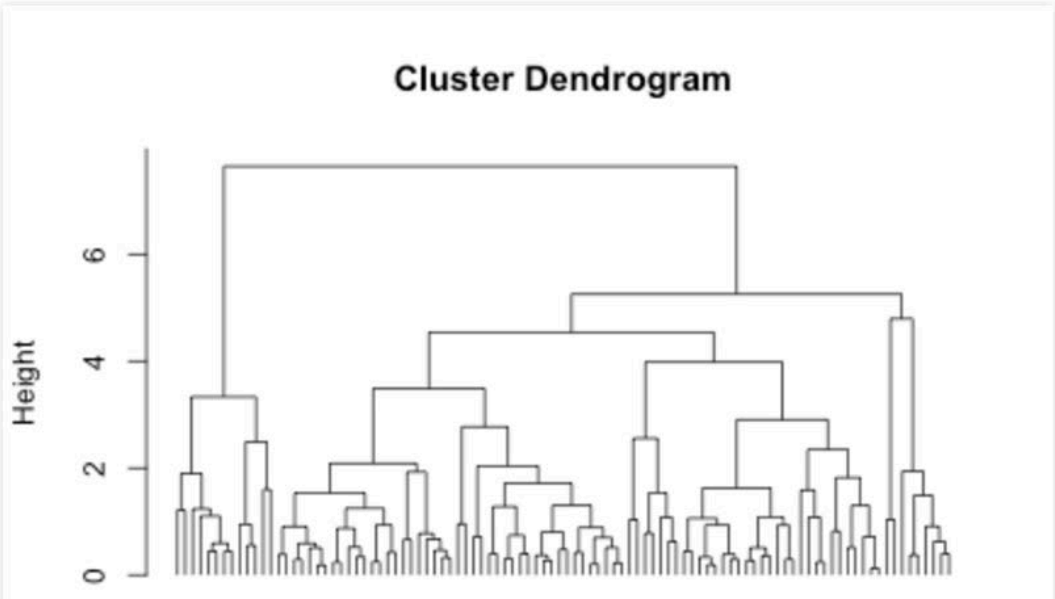
Pros and Cons of Hierarchical Clustering Analysis

- 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

Classification 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

1. Select *k* items at random from the data set as the initial cluster centers;
2. Cluster items based on the (Euclidean) proximity to the centers;
3. Set the new cluster centers to the centroid of the clusters from step (2);
4. Repeat (2) and (3) until the cluster assignment converges;
5. Perform (1)-(4) multiple times, choosing 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`**

Clustering in R

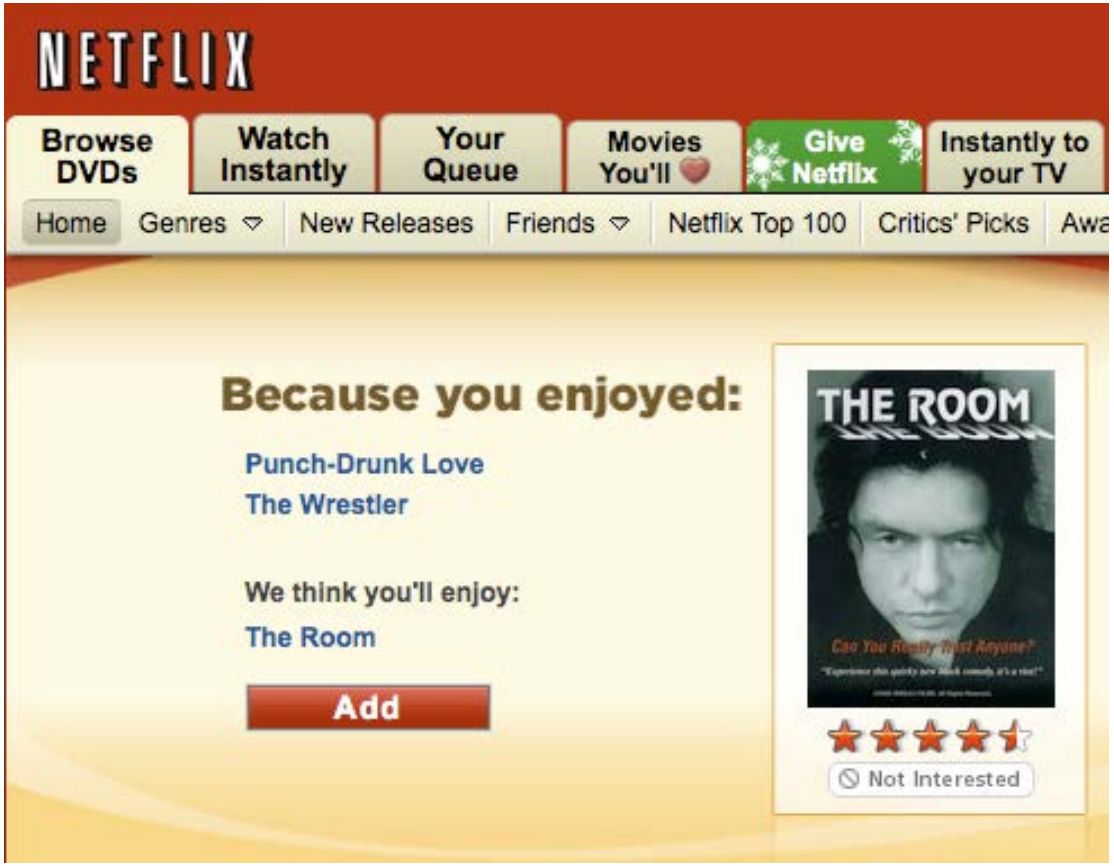
- **stats** package (built-in)
  - hierarchical clustering (hclust, heatmap, cophenetic)
  - k-means (kmeans)
- **class** package
  - self-organizing maps (SOM)
- **mclust** package
  - EM / mixture models
- **clusterCons** package
  - consensus clustering

- **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



Movie 1: romance

Movie 2: thriller

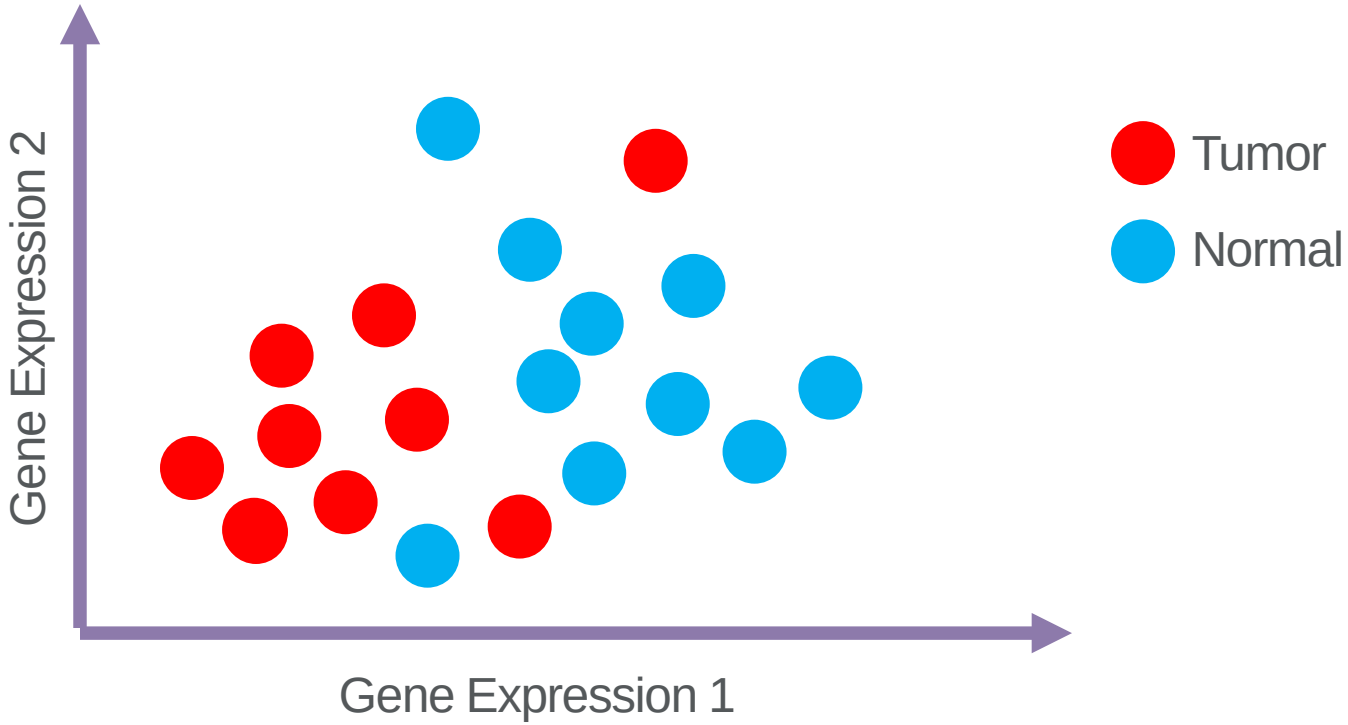
Movie 3: romance

Movie 4: documentary

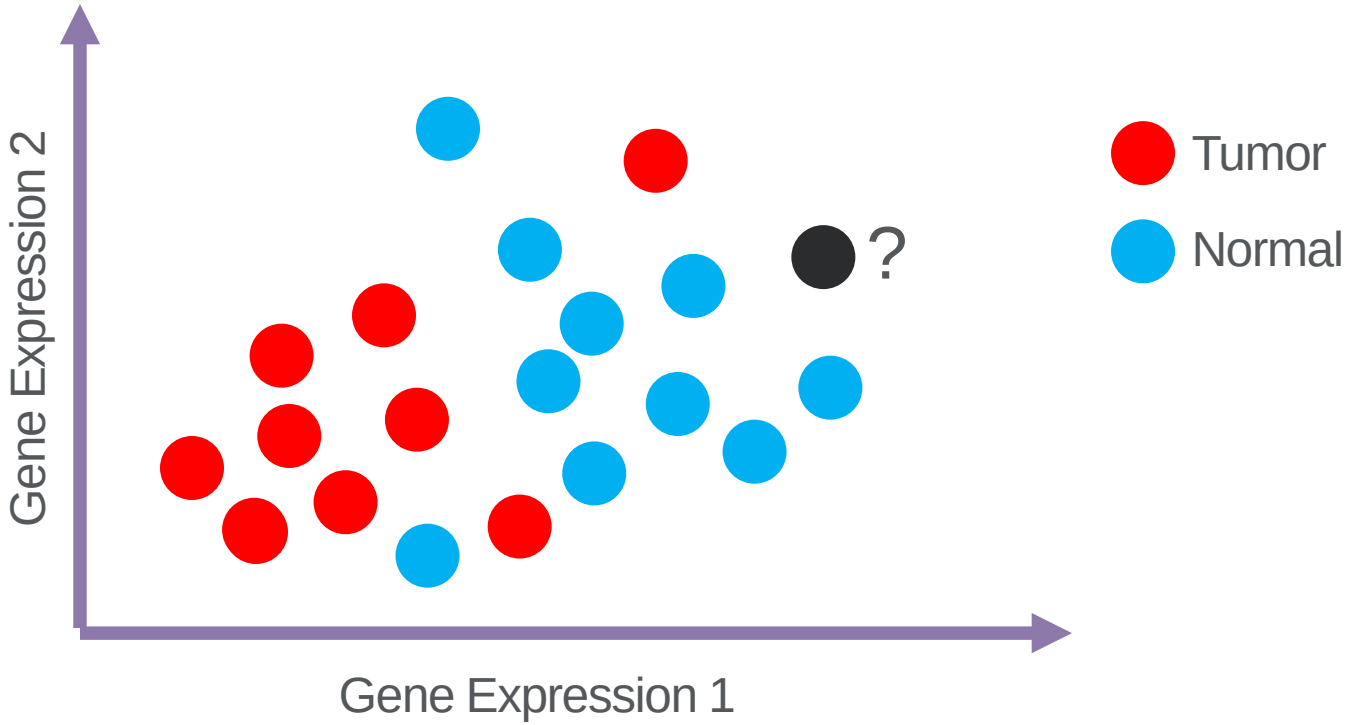
Movie 5: documentary

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)

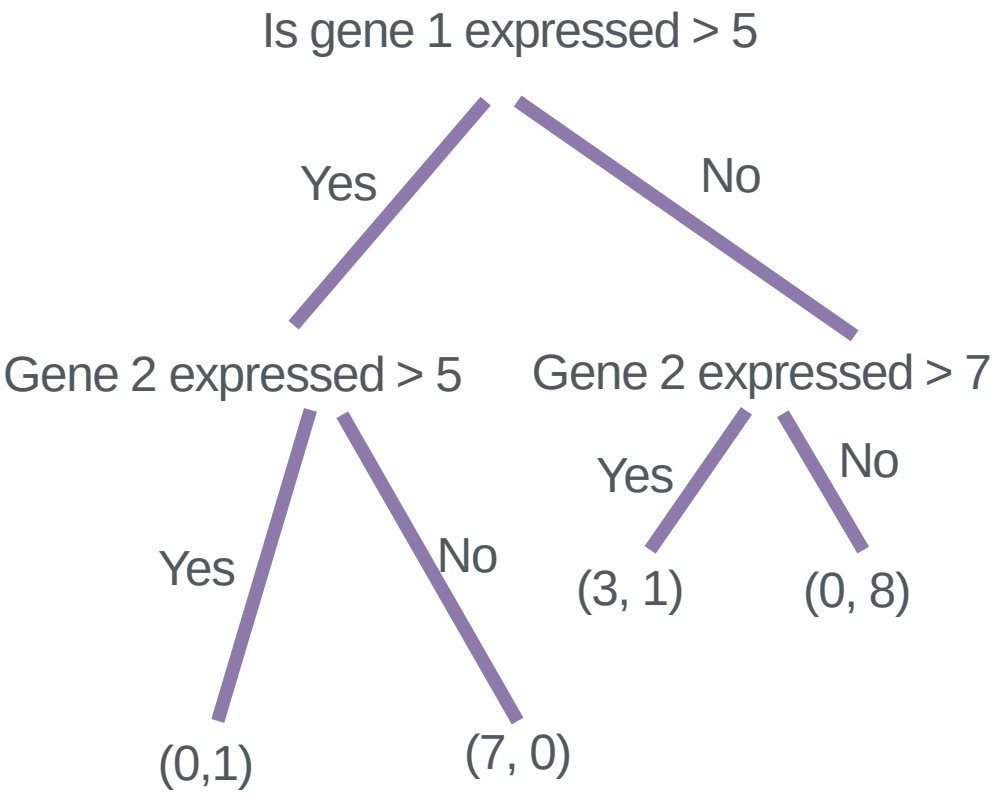
Assessing Accuracy: *K*-fold Cross-Validation

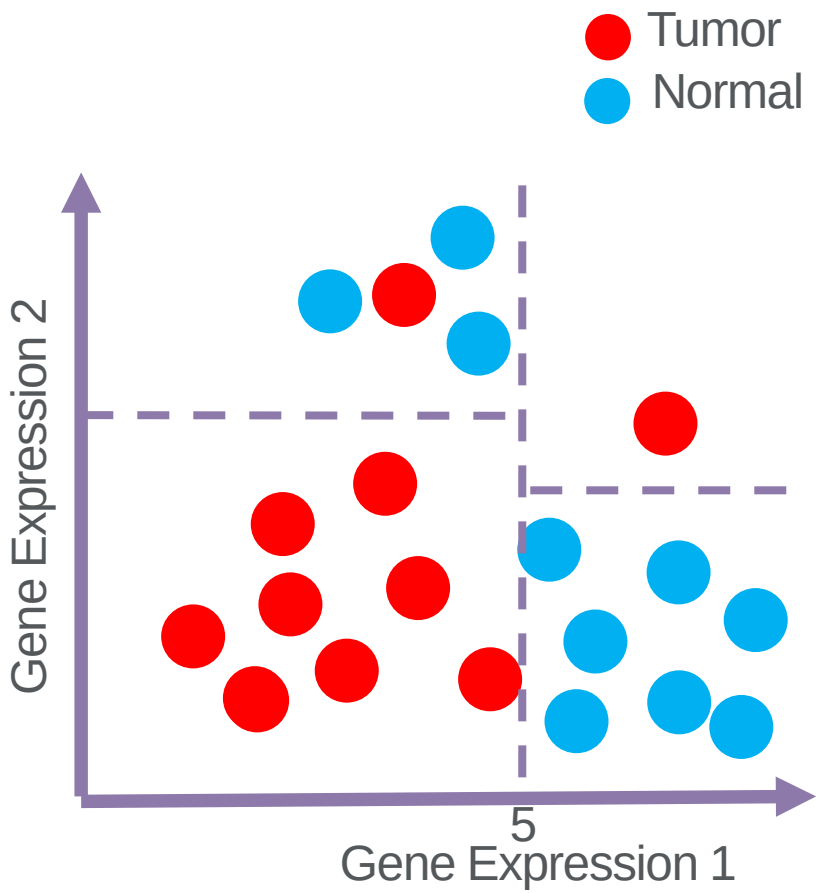
1. Break the samples into k blocks
2. Set one block aside for testing
3. Train on the other samples
4. Test on the samples in the testing block
5. Pick another one of the k blocks and repeat steps 2-4
6. Repeat step 5 until all blocks have been used for testing

Comment about Assessing Accuracy

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

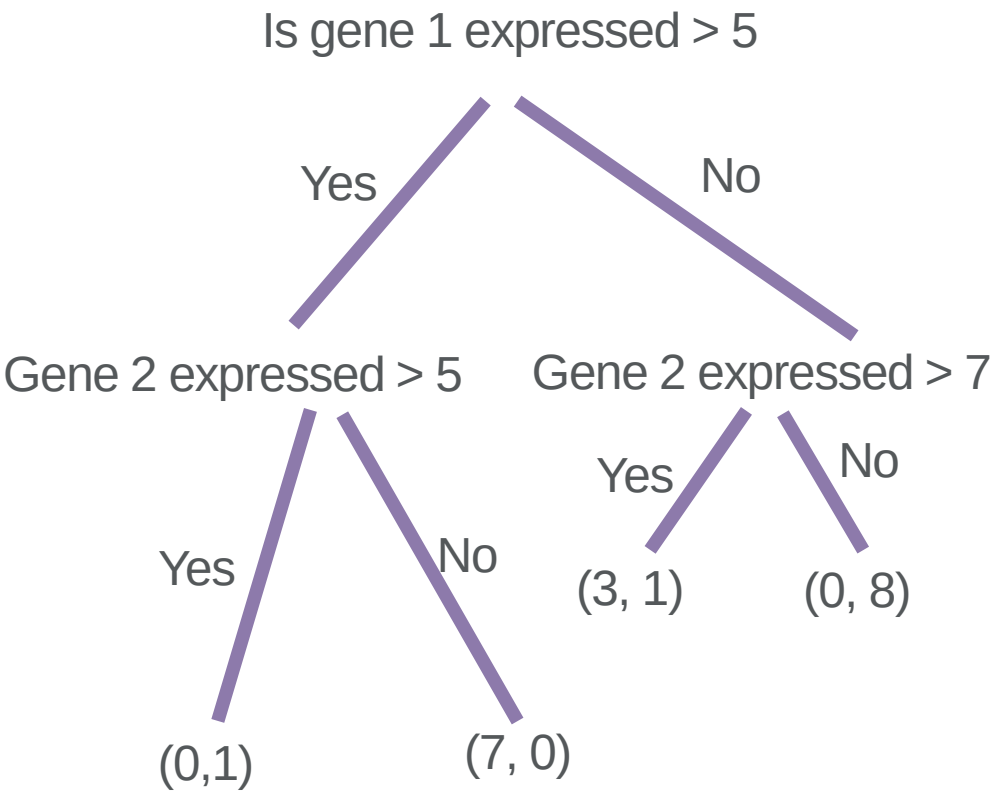
Classification and Regression Tree (CART)

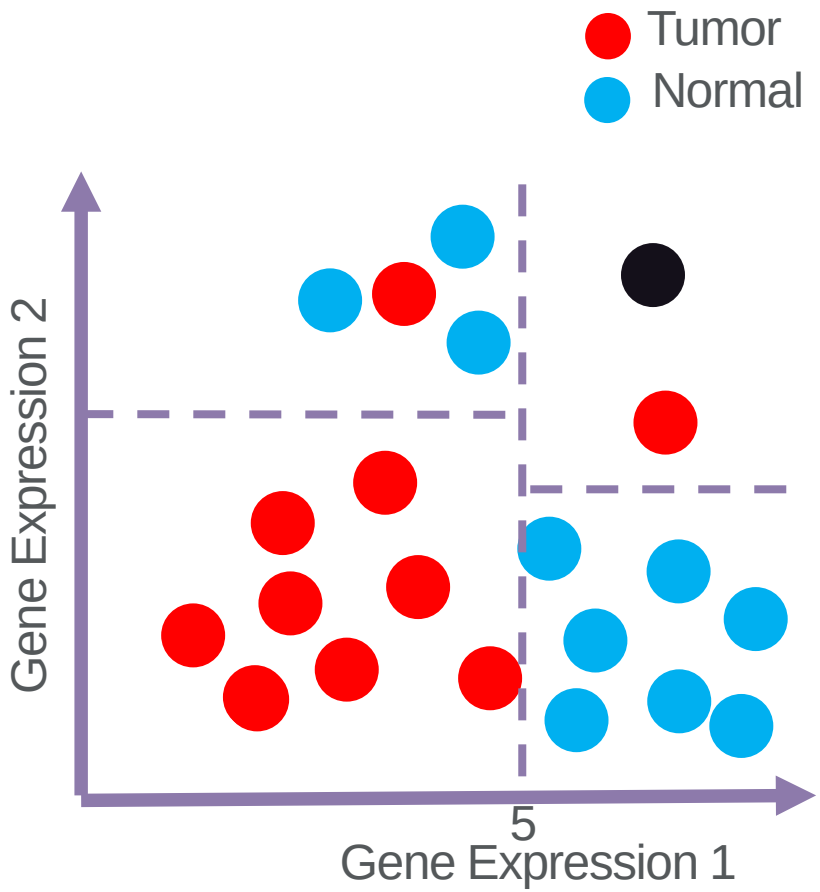




In R package **`rpart`**

Classification and Regression Tree (CART)





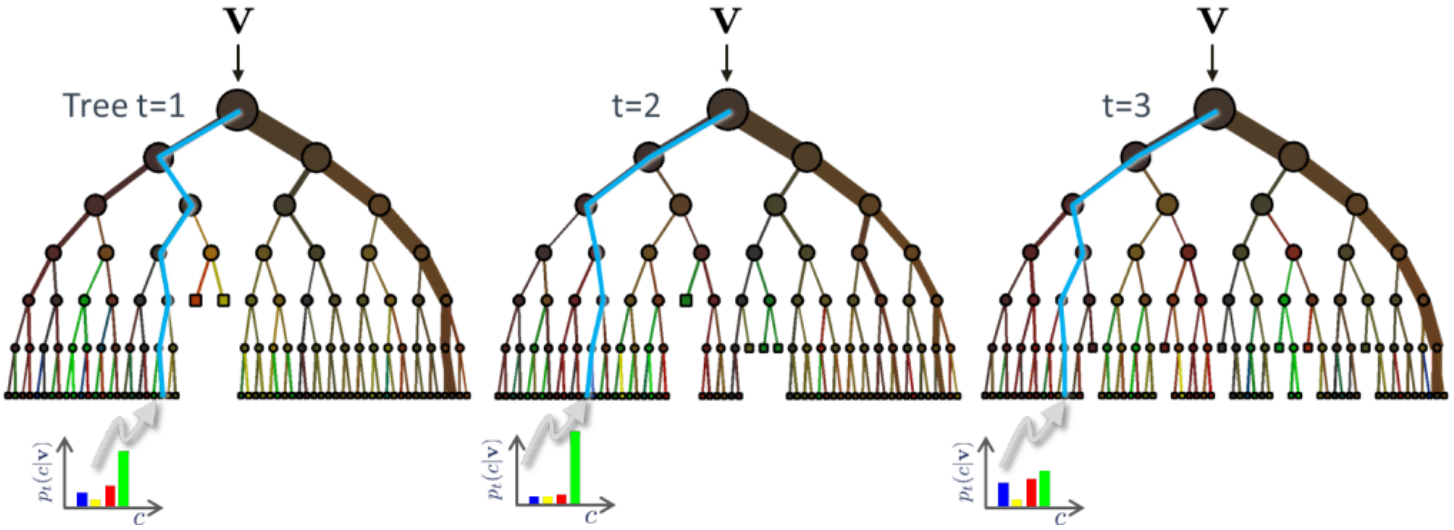
In R package **`rpart`**

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

Random Forests Illustration



Criminisi et al. 2011

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

Statistically Speaking …

What's next?

Tuesday, October 11	**Statistical Considerations for Sex Inclusion in Basic Science Research** Denise M. Scholtens, PhD, Associate Professor, Division of Biostatistics Associate Director, Department of Preventive Medicine
Friday, October 14	**The Impact of Other Factors: Confounding, Mediation, and Effect Modification** Amy Yang, MS, Sr. Statistical Analyst, Division of Biostatistics, Department of Preventive Medicine
Tuesday, October 18	**Statistical Power and Sample Size: What You Need and How Much** Mary Kwasny, ScD, Associate Professor, Division of Biostatistics, Department of Preventive Medicine
Friday, October 21	**Clinical Trials: Highlights from Design to Conduct** Masha Kocherginsky, PhD, Associate Professor, Division of Biostatistics, Department of Preventive Medicine
Tuesday, October 25	**Finding Signals in Big Data** Kwang-Youn A. Kim, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine
Friday, October 28	**Enhancing Rigor and Transparency in Research: Adopting Tools that Support Reproducible Research** Leah J. Welty, PhD, BCC Director, Associate Professor, Division of Biostatistics, Department of Preventive Medicine

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