Introduction to Python

for applications to biomedical industries

BME 6303 | CRN 19454 | 3 credits

Asynchronous Learning | Lectures & Assignments Available Online

Office Hours, Mondays, 2:30 pm Central

QutubLab.org/python

Instructor: Dr. Amina Ann Qutub

Amina.Qutub@utsa.edu

Additional Assistants:

Byron Long (Byron.Long@utsa.edu)

Erin Pollet (Erin.Pollet@utsa.edu)

Jenny Brethen (Jennifer.Brethen@utsa.edu)

Office Hours, Mondays, 2:30 pm Central or by appointment

Accessing course materials

NOTE: Links to the course site are also available through UTSA's Blackboard. Three coding challenges and one final report are submitted through UTSA's Blackboard. **Exception:** Quantu Project volunteers.

Main Course Site:

QutubLab.org/python

Materials (Modules, Videos, Reading):

QutubLab.org/pythonmaterials

Glossaries & Cheat Sheets

QutubLab.org/pythonglossary

Topics included: overall Python syntax, file input/output, data handling

What are included in modules?

16 Weekly Modules:

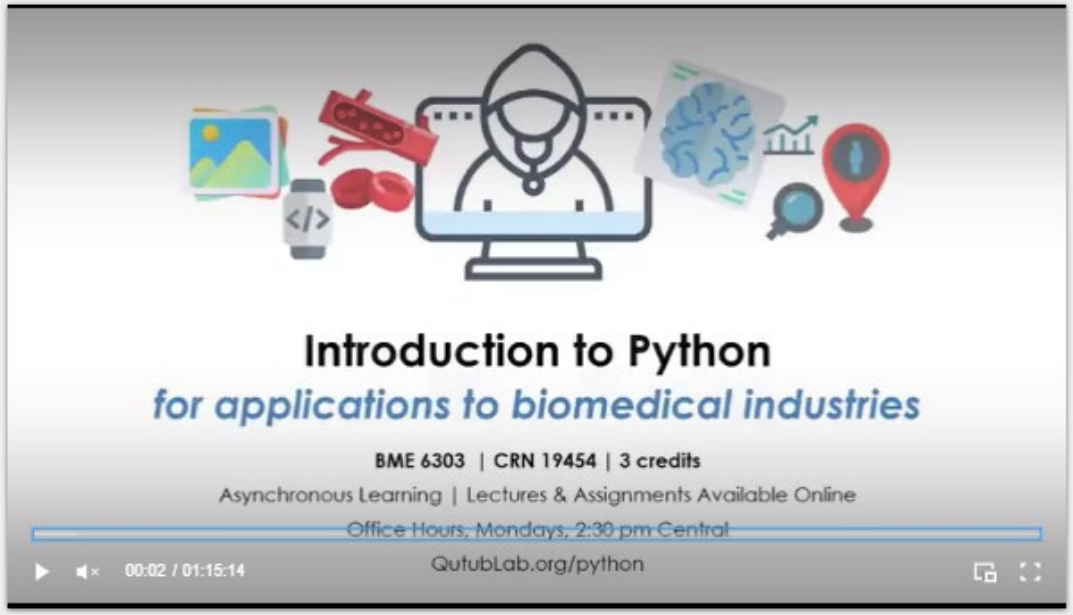
QutubLab.org/pythonmaterials

READ: Suggested reading

WATCH: Videos (not every week), tutorials

DO: Downloads, programming, coding challenges

Module 1
Welcome & Introduction



Weeks 1-2

Read
Beginner's Guide to Python

Watch
Introductory Video for Module 1 (above)

Do
1. Download & install Python
2. Download & install PyCharm (or other editor)
3. Bookmark W3Schools Python Intro
4. Sign up for zyBooks and subscribe to the Python book (Code: **UTSABME6303Fall2020**)
5. Try the first code "3 ways" as presented in the video lecture (@ ~28:20 into the video)

Module 5:
Applications: Omics & Biosensor Data Processing

Read

1. New to biomedicine? Read about "Omics" (Wikipedia Entry)
2. A Primer on Data Analytics in Functional Genomics: How to Move from Data to Insight?
3. 10 Clustering Algorithms with Python
4. Read about the data and challenges for the 3 class project options:
   - Activity data from fitness wearables (data download)
   - Proteomics & clinical data from cancer patients (data download)
   - Facebook data on COVID symptoms

Watch

This Live Class or Video for Module **5**

Module 5:
Applications: Omics & Biosensor Data Processing

Do

1) Introduce yourself to your project teammates (assigned via email on 9/21-9/22)
2) Decide on class project topic with your teammates: (1) identify trends in cardiovascular health from fitness watch & smartphone data; (2) predict leukemia patients' response to chemotherapy from clinical and proteomics data or (3) develop an early detection method from COVID Facebook symptoms data
3) Read about Biopython and download Biopython if your team plans to use it for your course project

Today's module

Specific Objective II: To gain familiarity with the methods, open-source programs, and other tools available for programming using Python.
Specific Objective III: To gain the programming skills needed to apply Python code to interpret large, complex, multimodal data (images, videos, protein-DNA interaction data, etc.) and be knowledgeable of the ways to optimize code.

Module 5: *Python Applications: Omics & Biosensor Data Processing*

QutubLab.org/pythonmaterials

Structure of today's module

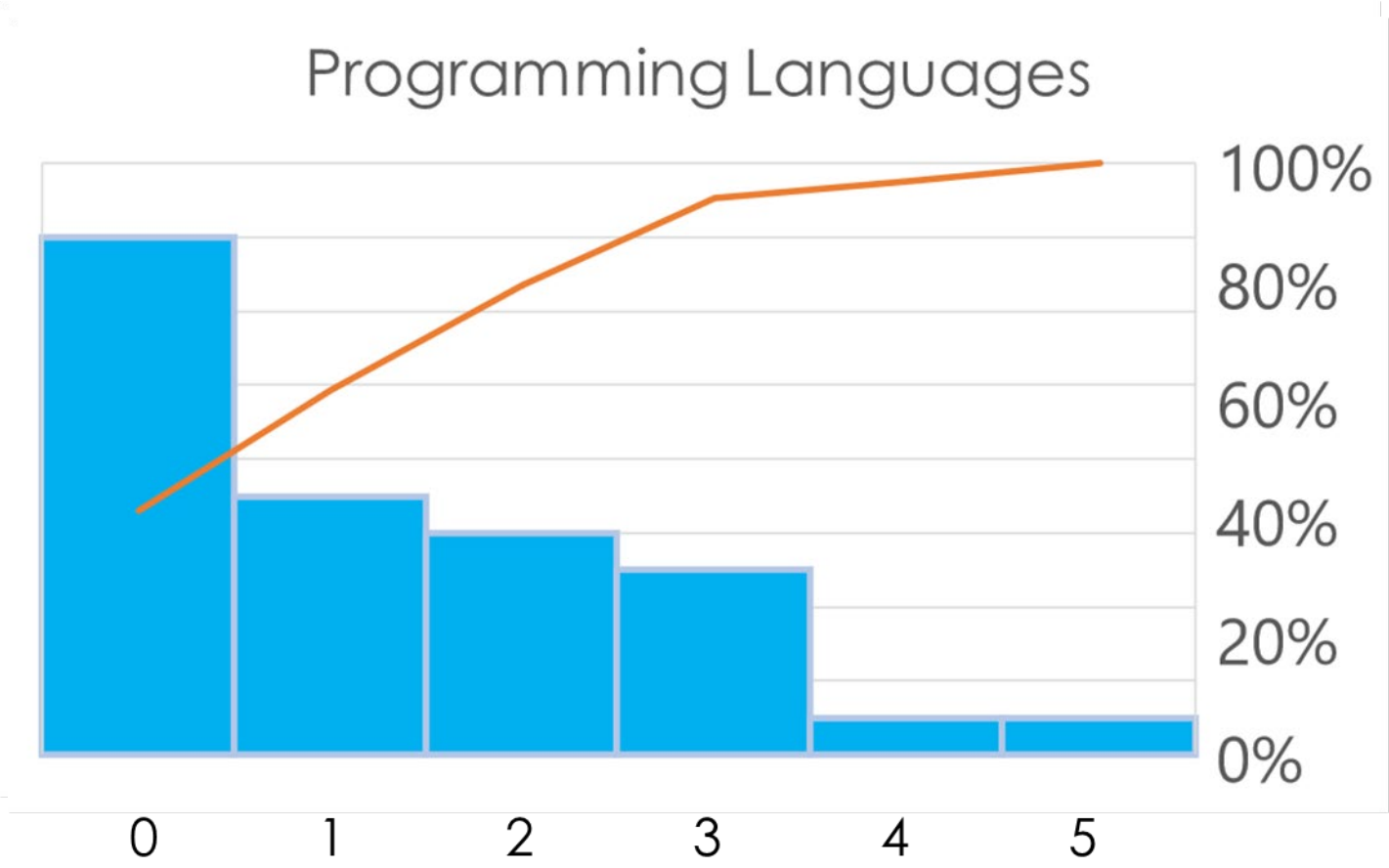
Introduction to Classmates & Class Project	~5-10 min
Class Project Options & Data Types	~25-30 min
Introduction to Clustering High Dimensional Data	~25-30 min

Next modules:

- Introduction to Python Open Source Packages & Importing Python Packages – 15-20 min
- Python Functions

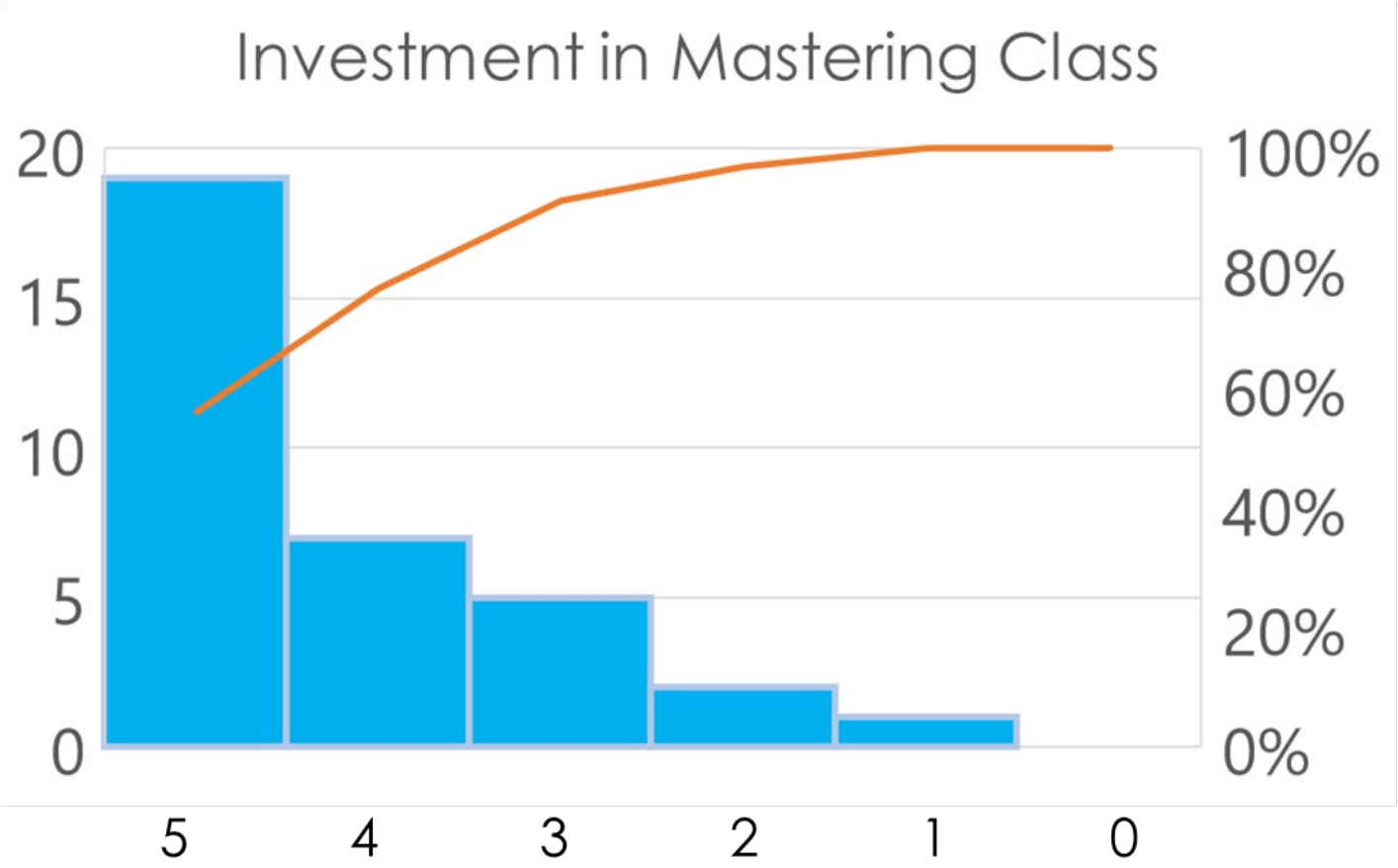
An Intro to Classmates

Medical students, engineers, retirees, teachers, artists, nurses
4 of 34 survey responders know cluster analysis



An Intro to Classmates

Medical students, engineers, retirees, teachers, artists, nurses
4 of 34 survey responders know cluster analysis



Team Projects

Scoring Criteria

1. Organization
2. Clarity
3. Methodology & its Justification
4. Conclusions
5. Impact
6. Creativity

Peer weighting scales grade

3 Team Project Options

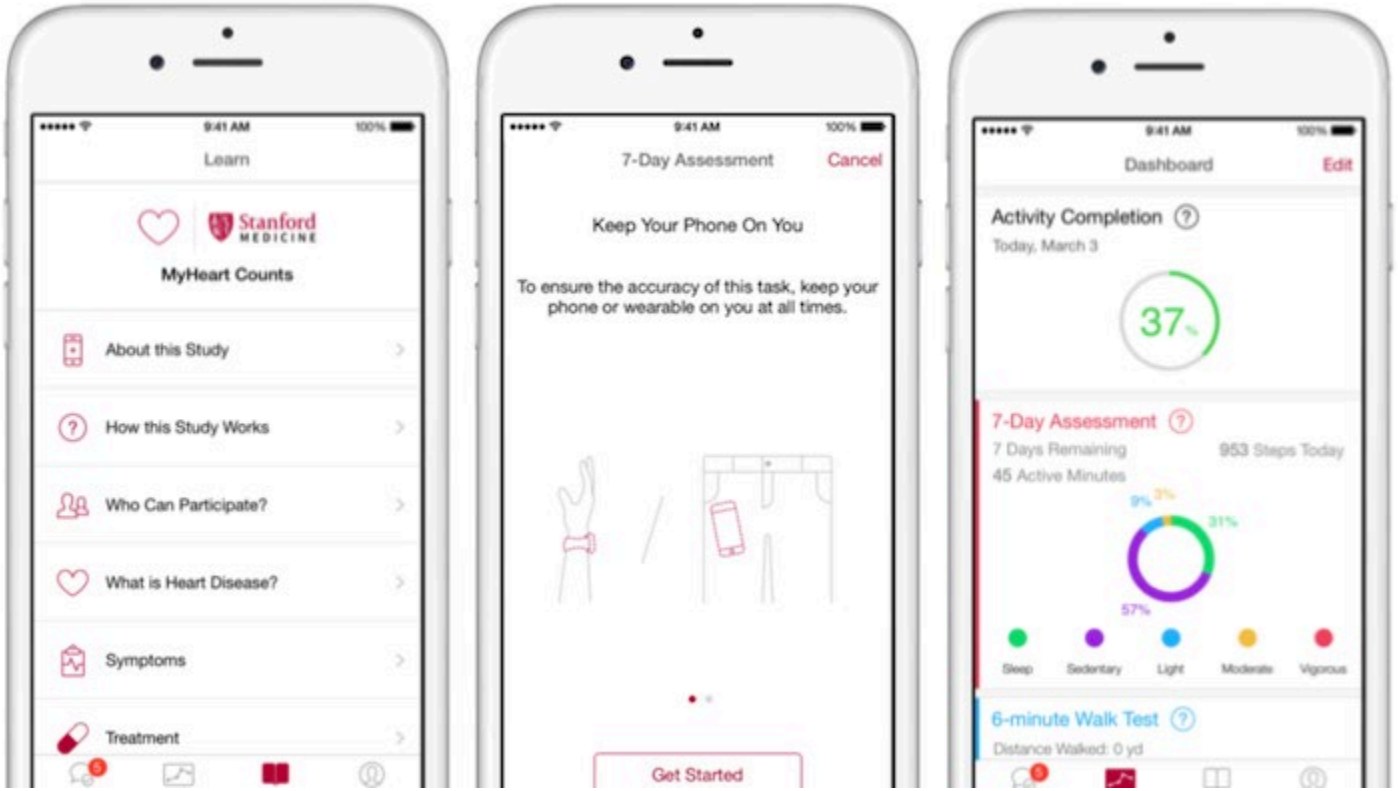
1. Identify trends in cardiovascular health from fitness watch & smartphone data

2. Predict leukemia patients' response to chemotherapy from clinical and proteomics data

3. Develop an early detection method from COVID Facebook symptoms data

Project Option 1

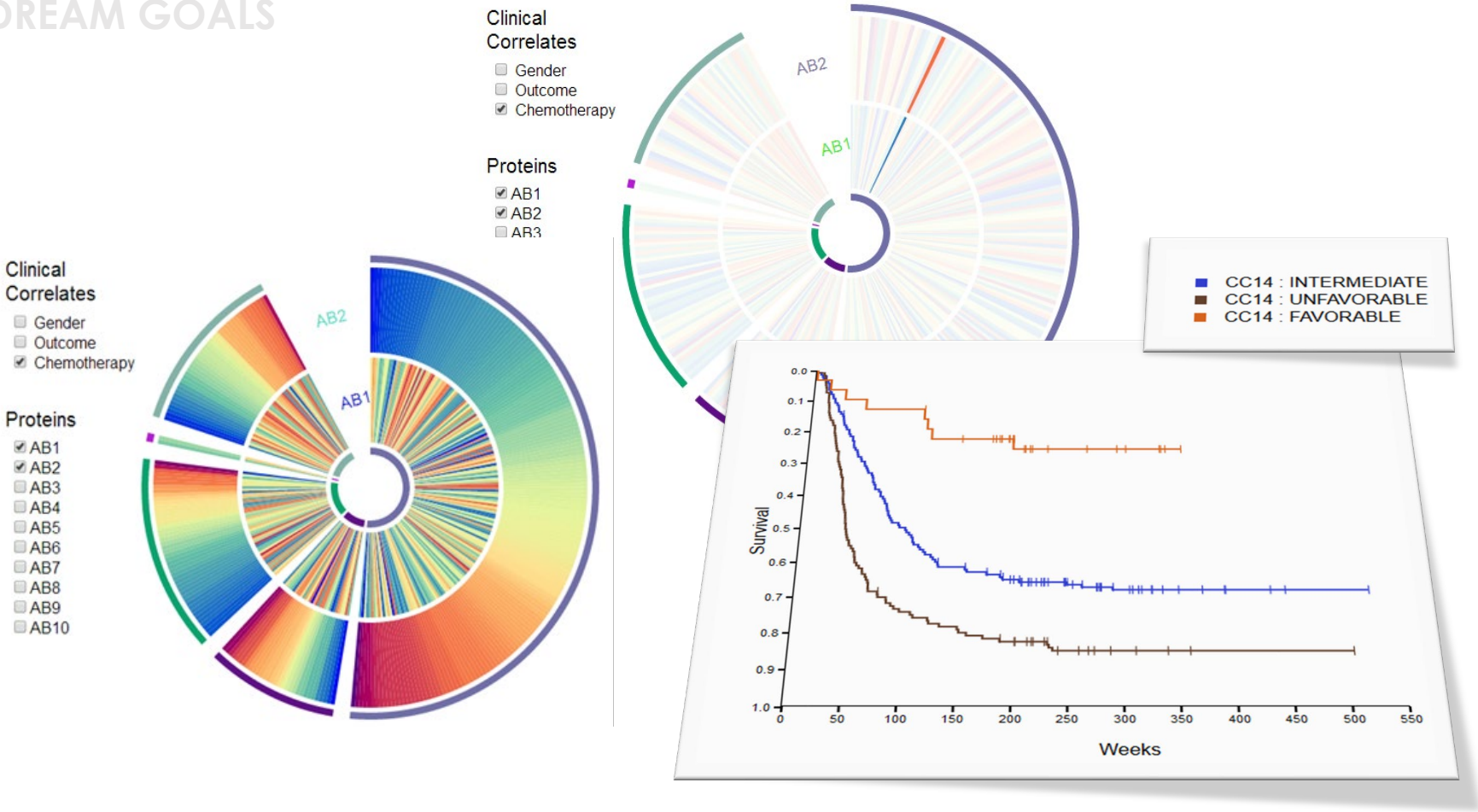
Predict cardiovascular health from 7-day assessment data



Project Option 2

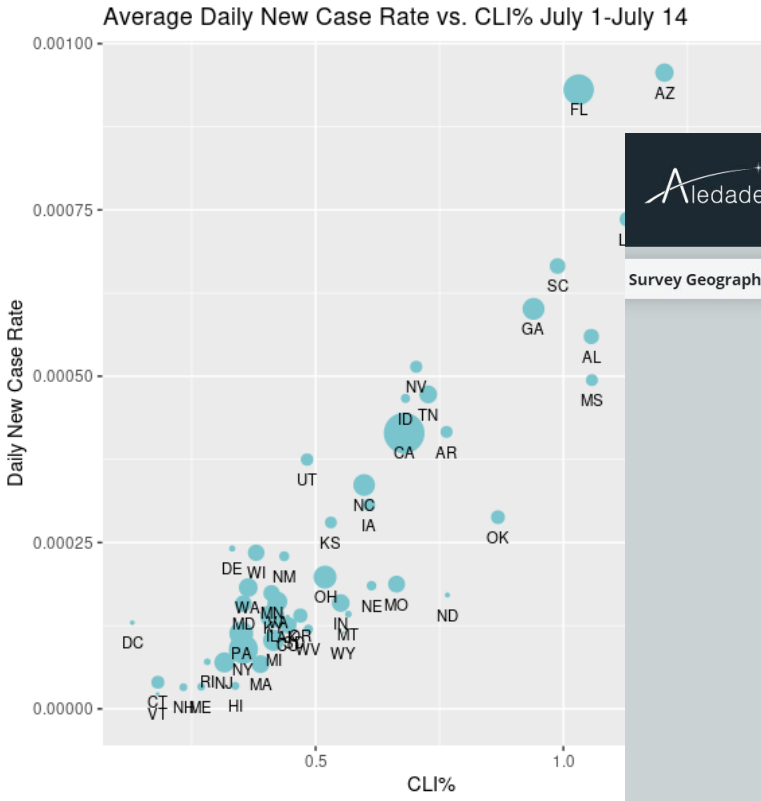
Predict which AML patients will have Complete Remission or will be Primary Resistant

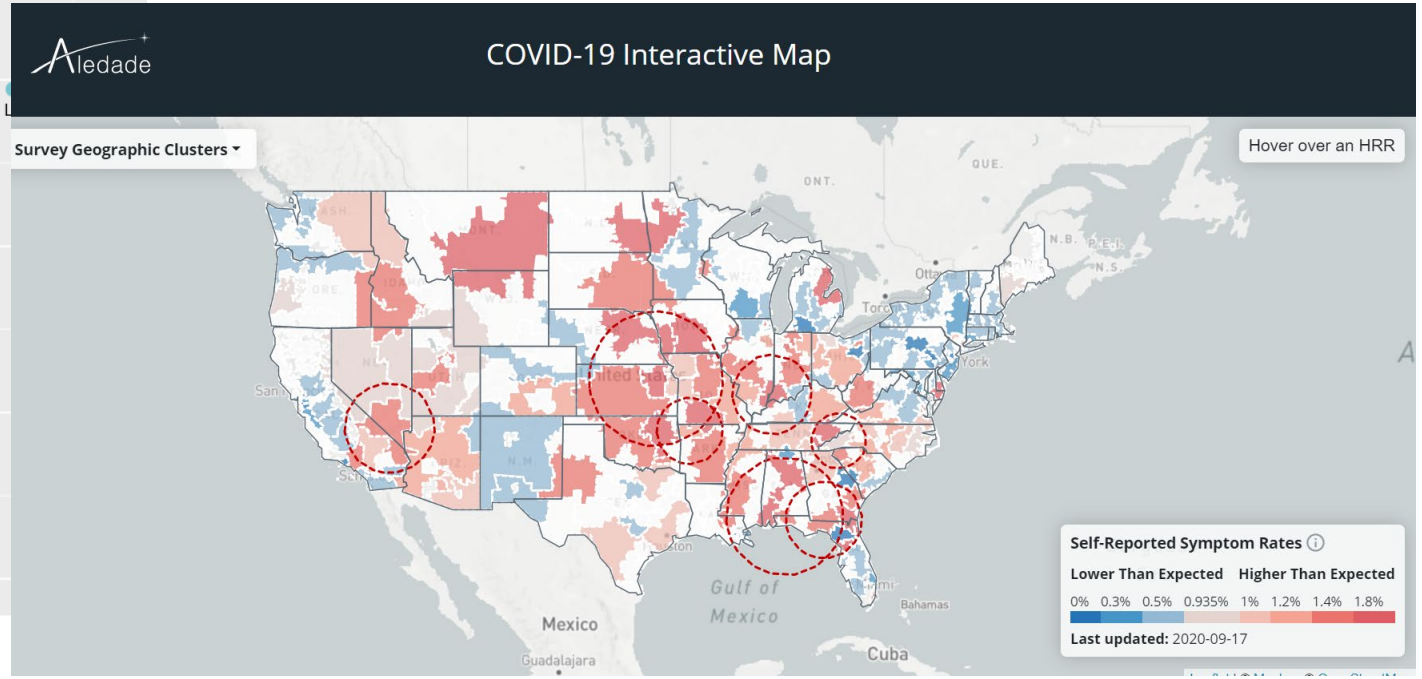
DREAM GOALS



Project Option 3

Predict infection and recovery rate from COVID19 from Facebook data





Acute Myeloid Leukemia Data

Clinical & Proteomics Dataset

Pages

Acute Myeloid Leukemia Outcome Prediction Challenge

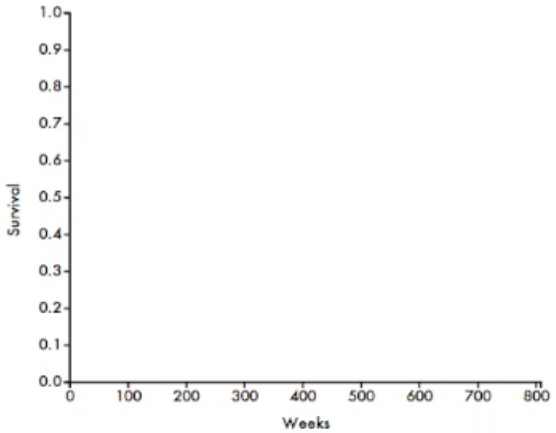
<<

Data Selection

Proteins	Clincal Correlates	Groups	Patients
ACTB	Age.at.Dx	SEX	2028430
AIFM1	AHD	PRIOR.MAL	2079222
AKT1	CR.duration	PRIOR.CHEMO	2481128
AKT1_2_3.pS473	WBC	PRIOR.XRT	2503784
AKT1_2_3.pT308	ABS BLST	Infection	2617320
ARC	BM.BLAST	cyto.cat	2659168
ASH2L	BM.MONOCYTES	ITD	2690050
ASNS	BM.PROM	D835	2709522
ATF3	PB.BLAST	Ras.Stat	2720828
ATG7	PB.MONO	Chemo.Simplest	2773346

Protein Heatmap

Survival Curve



Clincial Correlate Heatmap

Acute Myeloid Leukemia Data

Clinical & Proteomics Dataset

290 Patients

- Subset of 511 patients seen at MDACC, treated with ARA-C therapies
- 2/3rd given for training; 1/3rd saved for scoring

47 Clinical Correlates

- Includes mutation status and cytogenetic categories

Outcomes

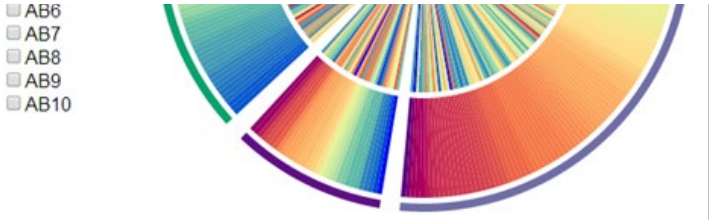
- Values for CR, PR, Remission Duration and Overall Survival up to 3 years on all patients

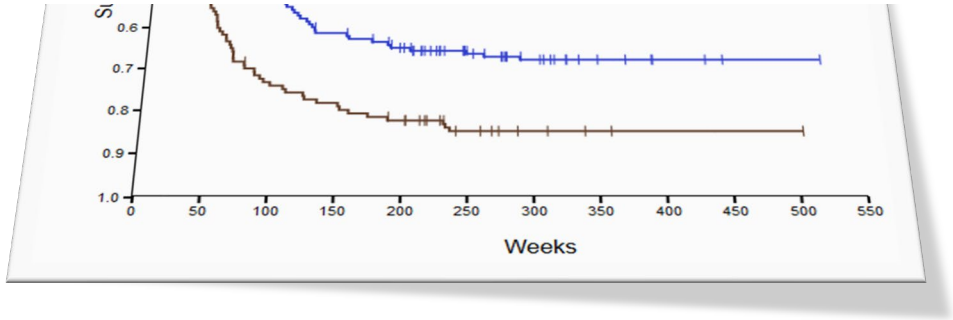
231 Protein Expression Levels

AML Outcome Prediction

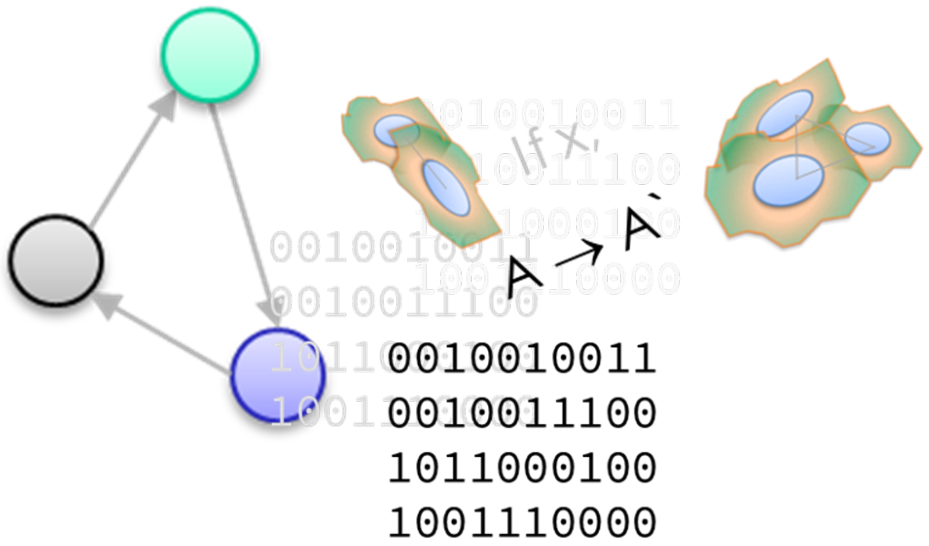
DREAM GOALS

Determine the best model to predict which AML patients will have Complete Remission or will be Primary Resistant.





Clustering Introduction



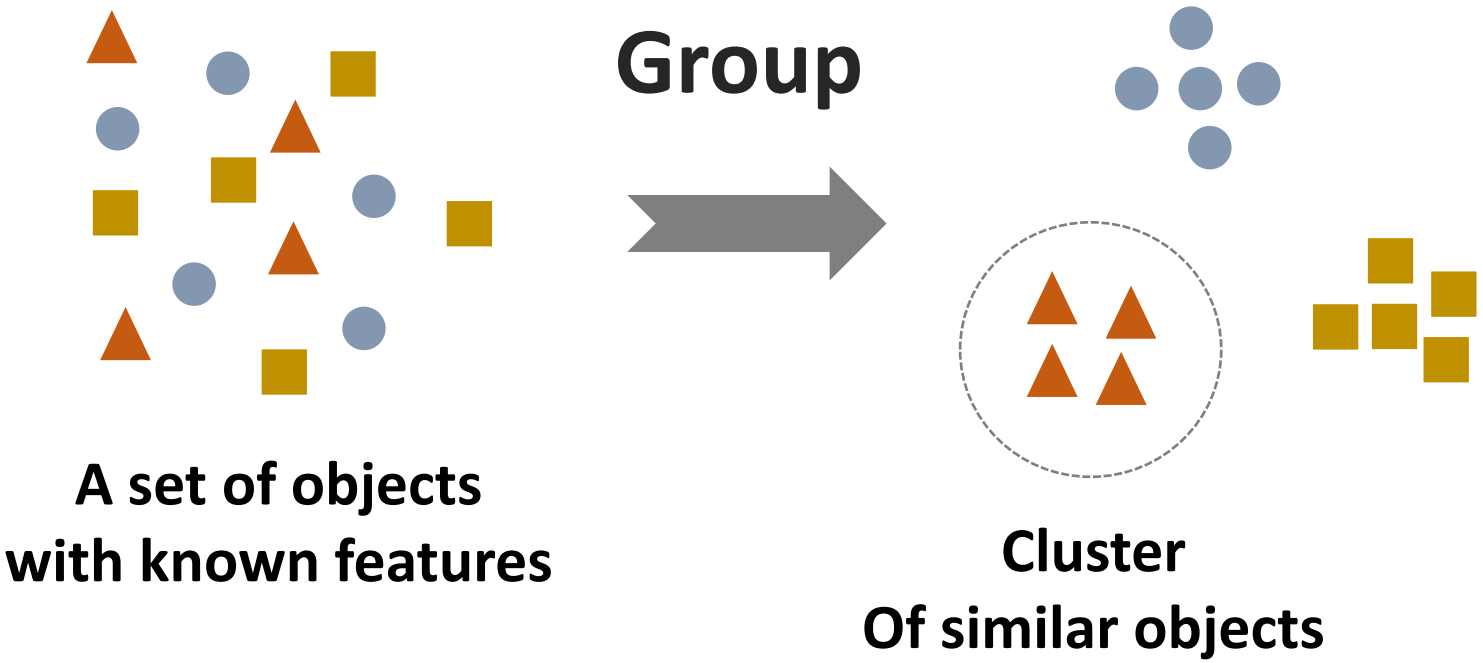
Clustering Intro. Slide Credits:

Dr. Chenyue Wendy Hu, Qutub Lab Alumna, Data Scientist Uber Technologies

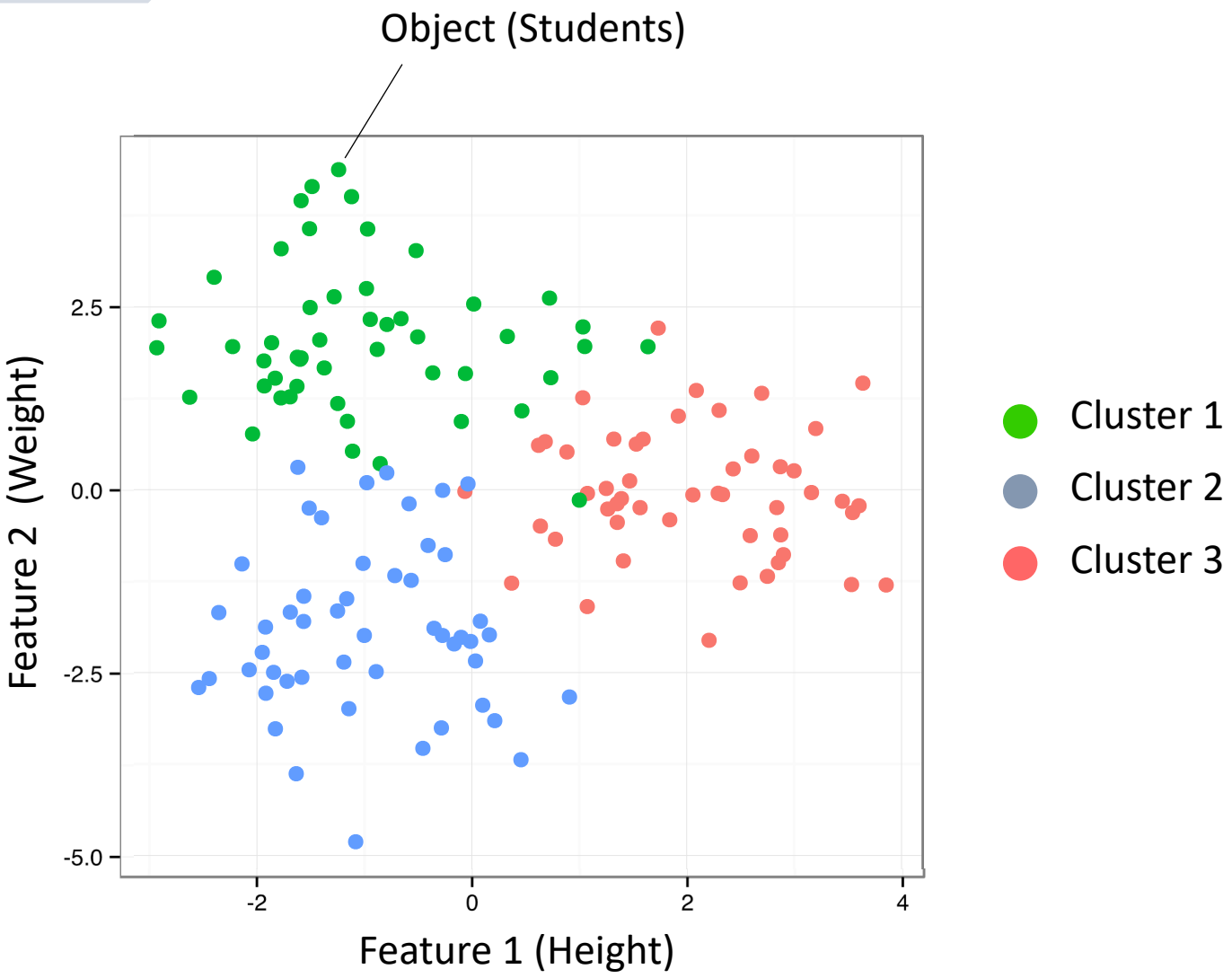
Outline

- What is cluster analysis?
  - concept, examples, application scenarios
- How does cluster analysis work?
  - Similarity (Euclidean distance)
  - K-means Clustering
  - Hierarchical Clustering
- How to pick the cluster number?
  - Elbow Method
  - Silhouette
- Best practices & review

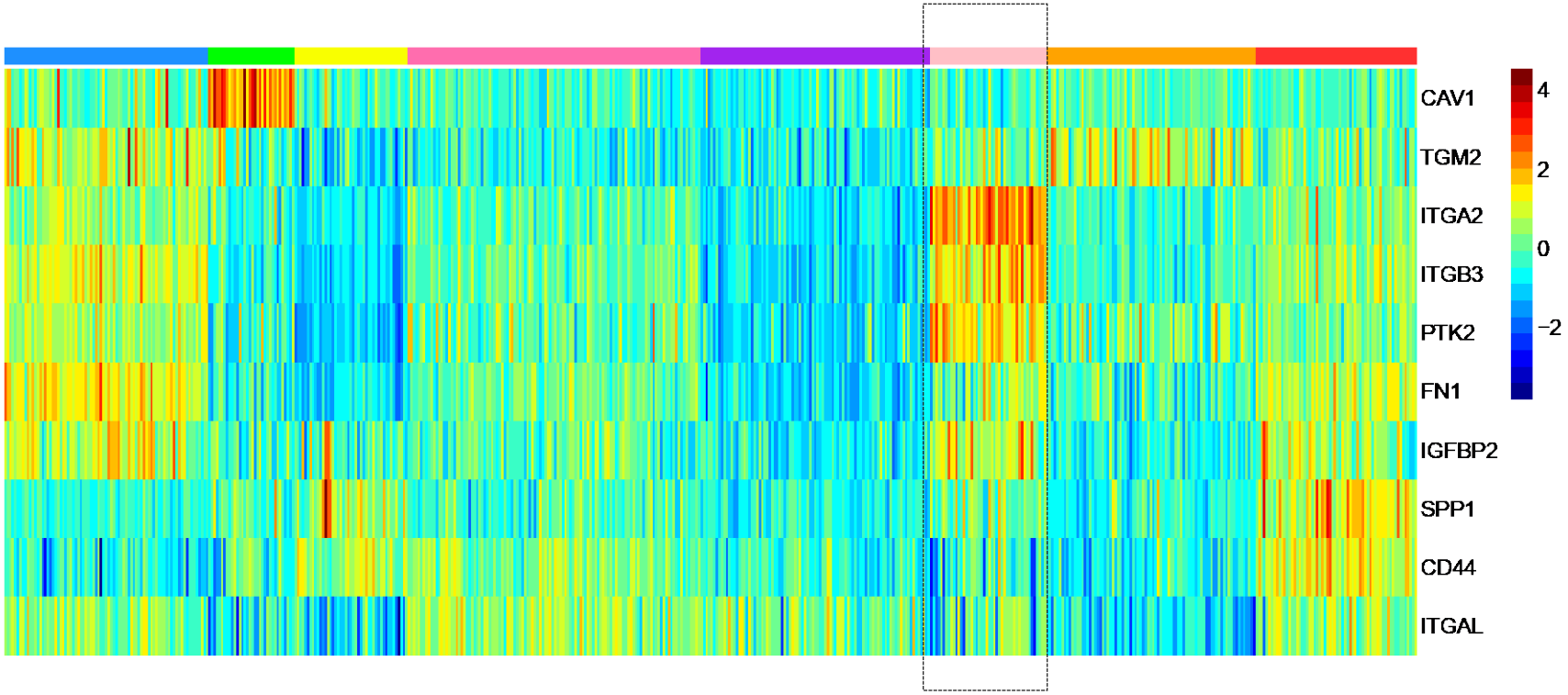
Concept


Group
A set of objects
with known features
Cluster
Of similar objects

Example 1


Object (Students)
2.5
0.0
-2.5
-5.0
-2
0
2
4
Feature 2 (Weight)
Feature 1 (Height)
Cluster 1
Cluster 2
Cluster 3

Example 2



Cluster

Application

Application scenarios

- You have multiple features (> 1)
  - matrix or data table
- You believe the population is heterogeneous
- You don’t know the true grouping of objects

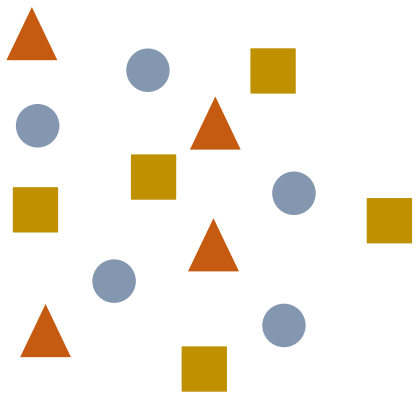
Application goals

✓ End-step: Define classes or categories

✓ Pre-step: Segment samples for subsequent analyses

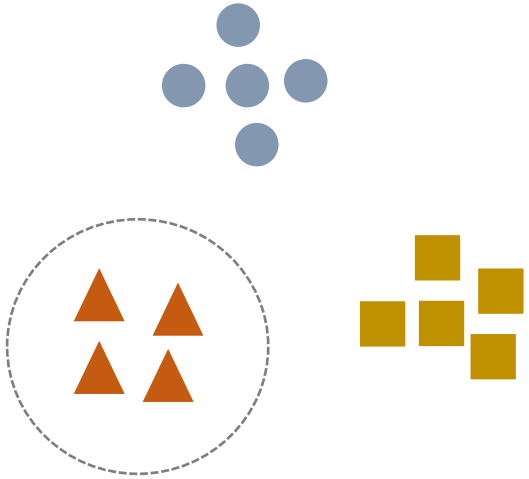
Concept

Clustering Algorithm



A set of objects with known features

Group

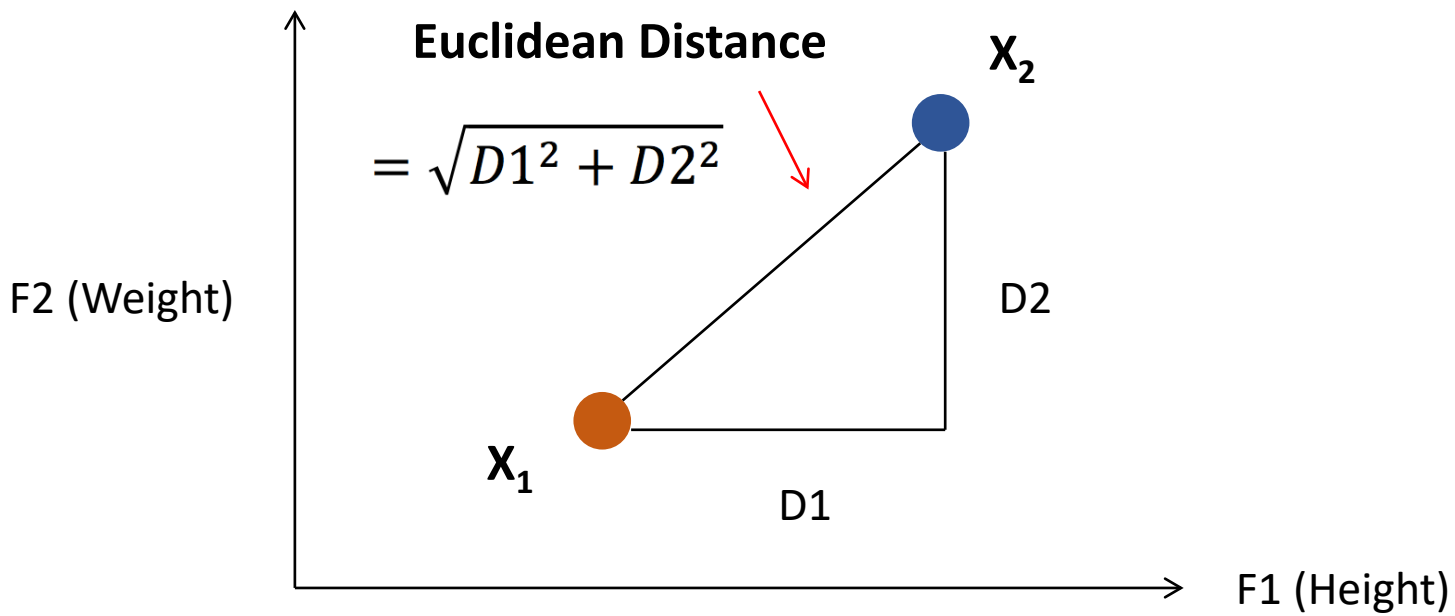


Cluster Of similar objects

Defining similarity / distance

Similarity

Distance (Dissimilarity)



Generalized Euclidean Distance
Between X_i and X_j
In n dimensional space

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(D_{ik} - D_{jk}\right)^2}$$

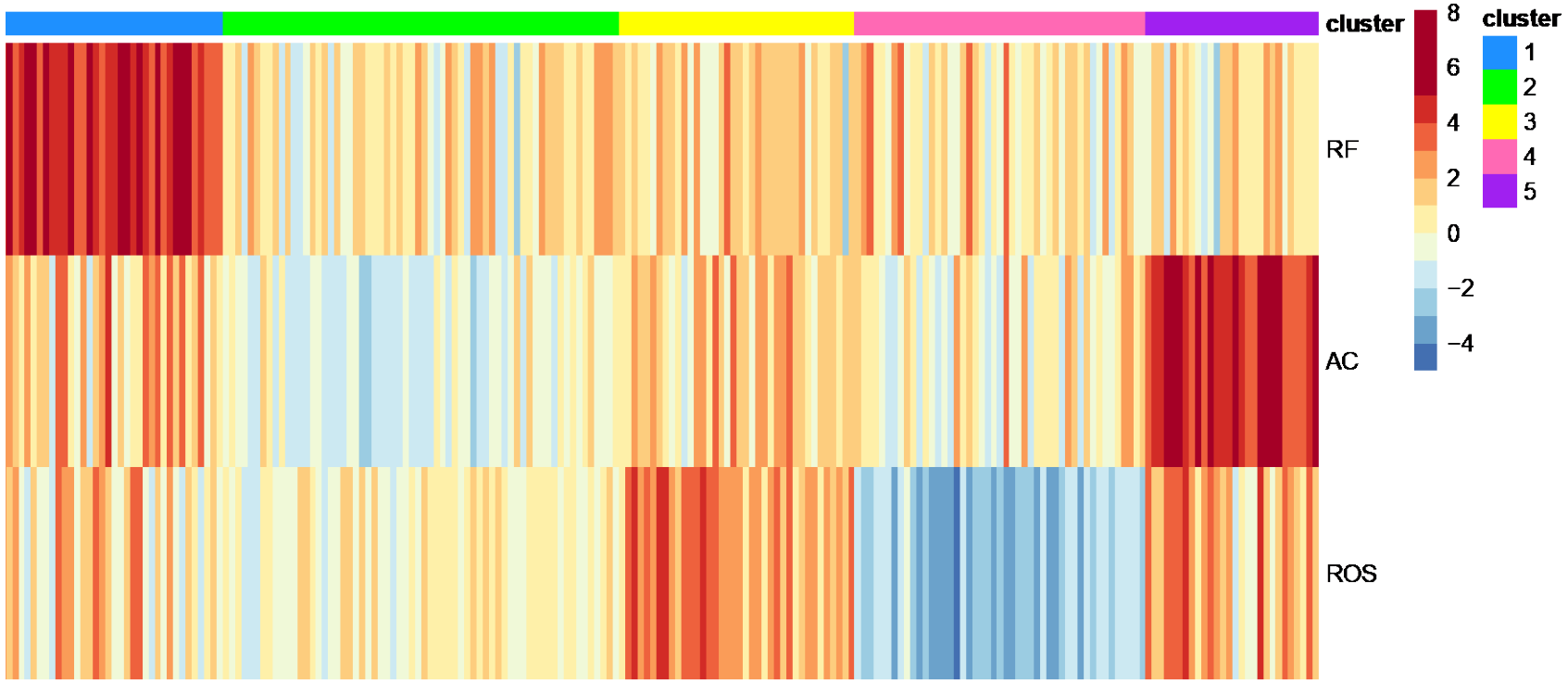
Algorithms

Common Clustering Algorithms

- **Hierarchical Clustering**
- **K-means**
- Affinity Propagation
- Spectral Clustering
- DBSCAN

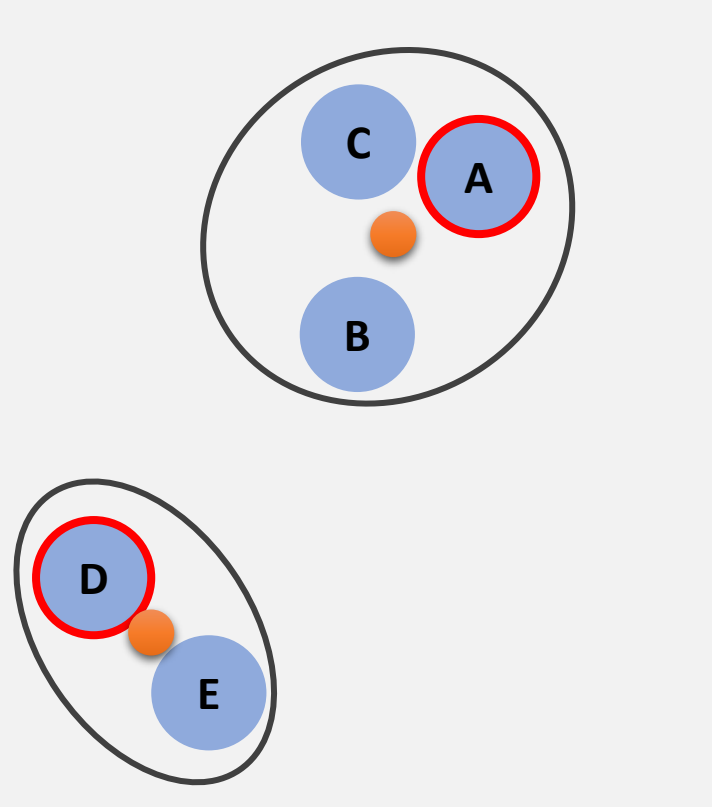
K-means

No dendrogram



K-means

- Start with K initial random cluster centers
- Repeat the following steps:
  - Assign objects to their closest cluster centers
  - Recompute cluster centers
- End when the solution is optimized (nothing to do)



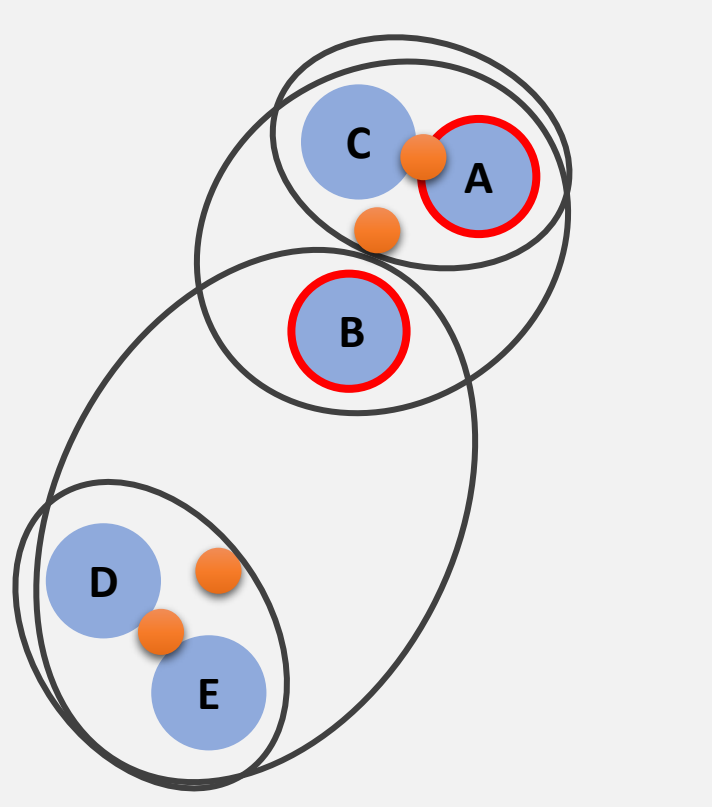
Initialization: A, D

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

K-means

- Start with K initial random cluster centers
- Repeat the following steps:
  - Assign objects to their closest cluster centers
  - Recompute cluster centers
- End when the solution is optimized (nothing to do)



Initialization: A, D

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

Initialization: A, B

Cluster: {A, C}, {B, D, E}

Cluster: {A, B, C}, {D, E}

Cluster: {A, B, C}, {D, E}

K-means

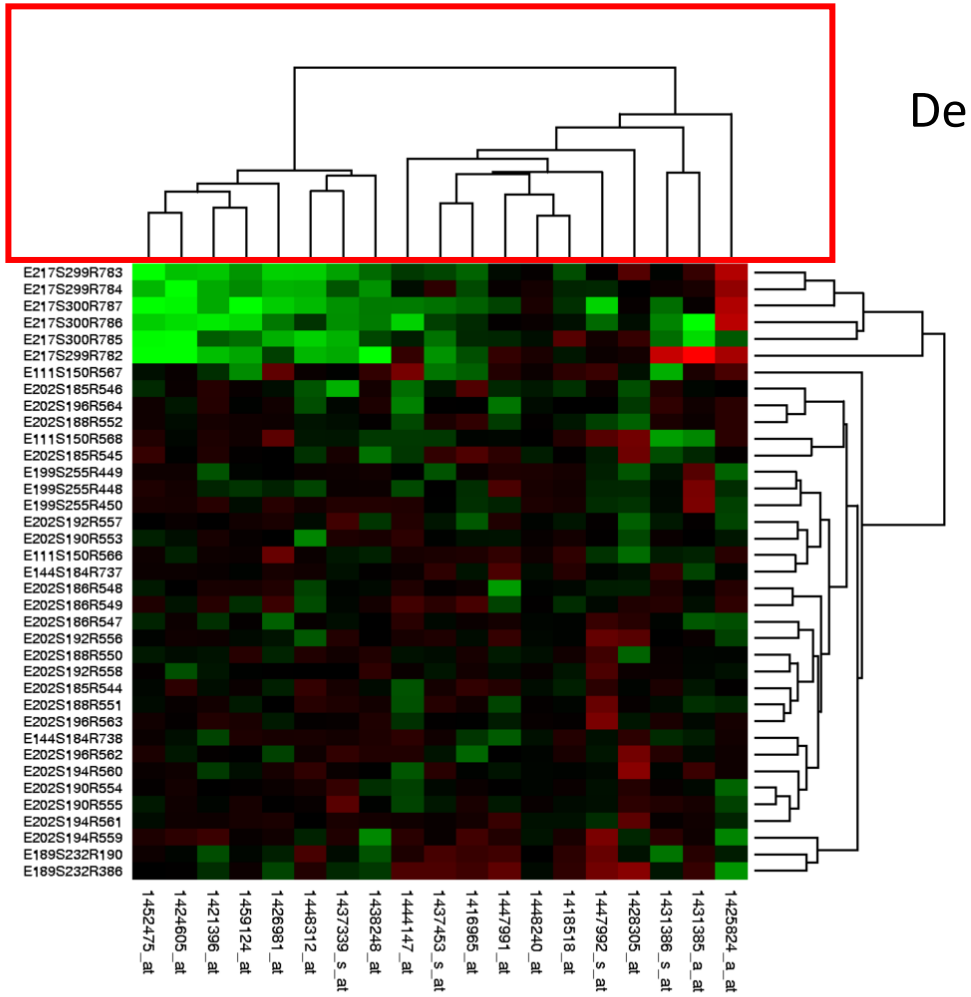
Summary

- Need to provide K (number of clusters)
- Iterative assignment and center update
- Optimize within-cluster distances
- Converge to local optimum

- Not necessarily the global optimum
- Different random initiations -> different solutions
- Make sure to run the algorithm multiple times (>=100)

Slower than Hierarchical, but better clustering quality

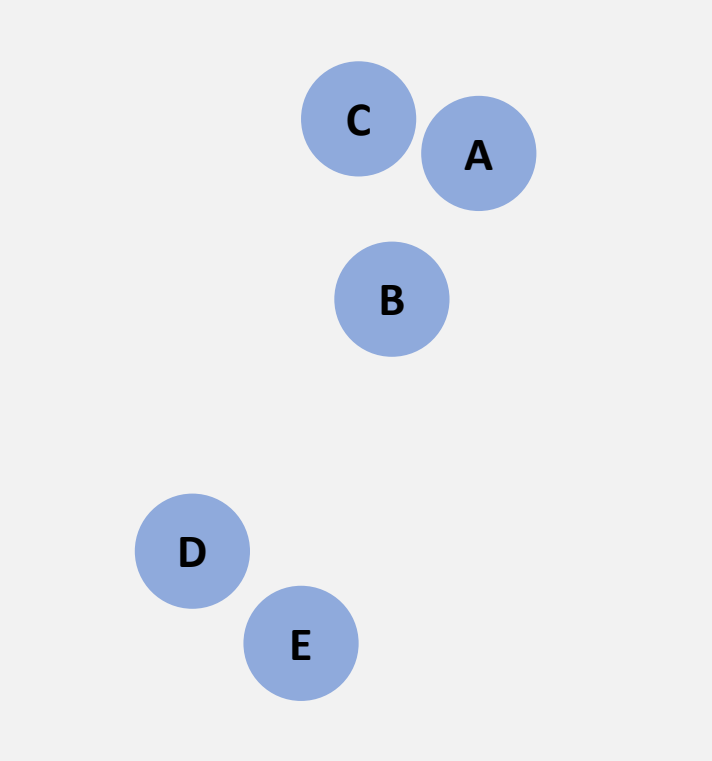
Hierarchical

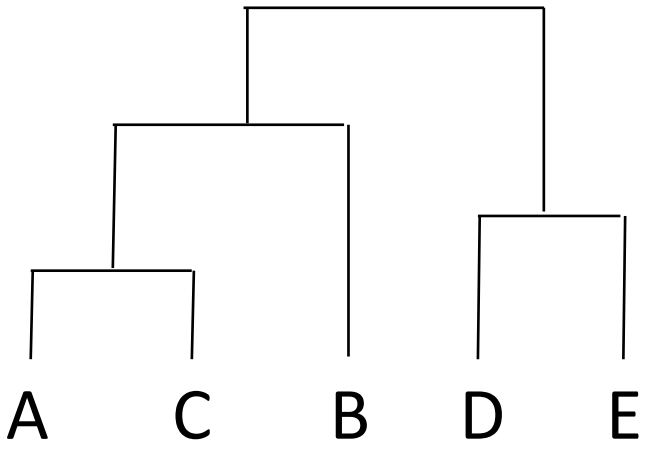


Dendrogram

Hierarchical

- Start with each object as a cluster
- Repeatedly connect the closest two clusters
- End when there is only one cluster left

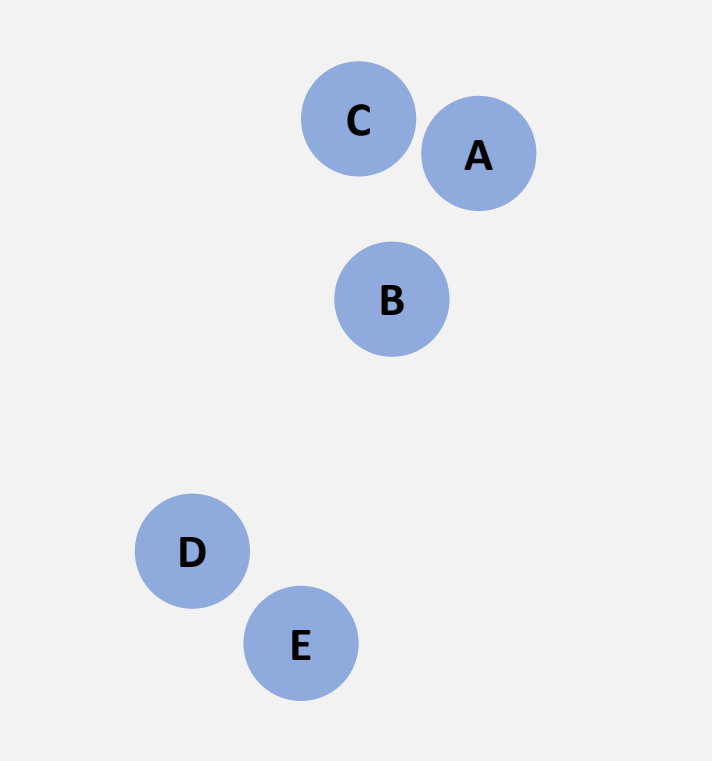


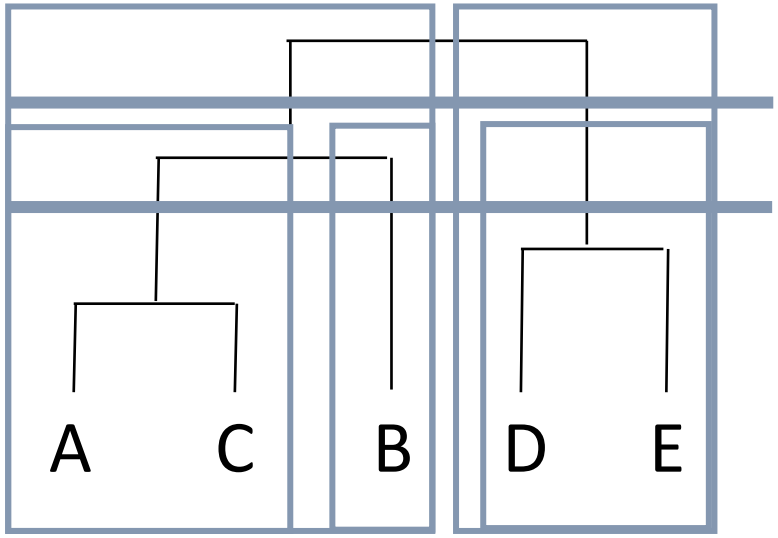


Minimum **Linkage**: the smallest distance between elements of each cluster

Hierarchical

Cut tree to obtain a clustering solution





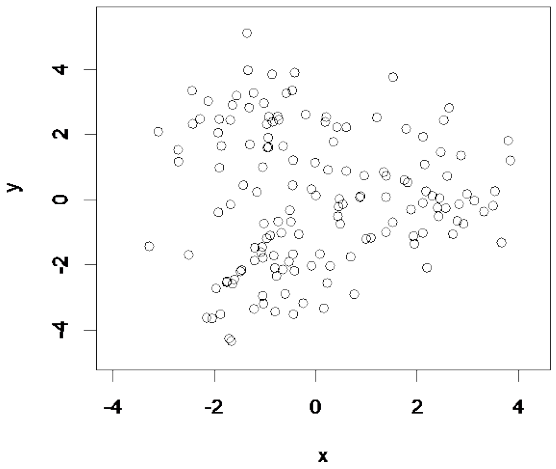
Hierarchical

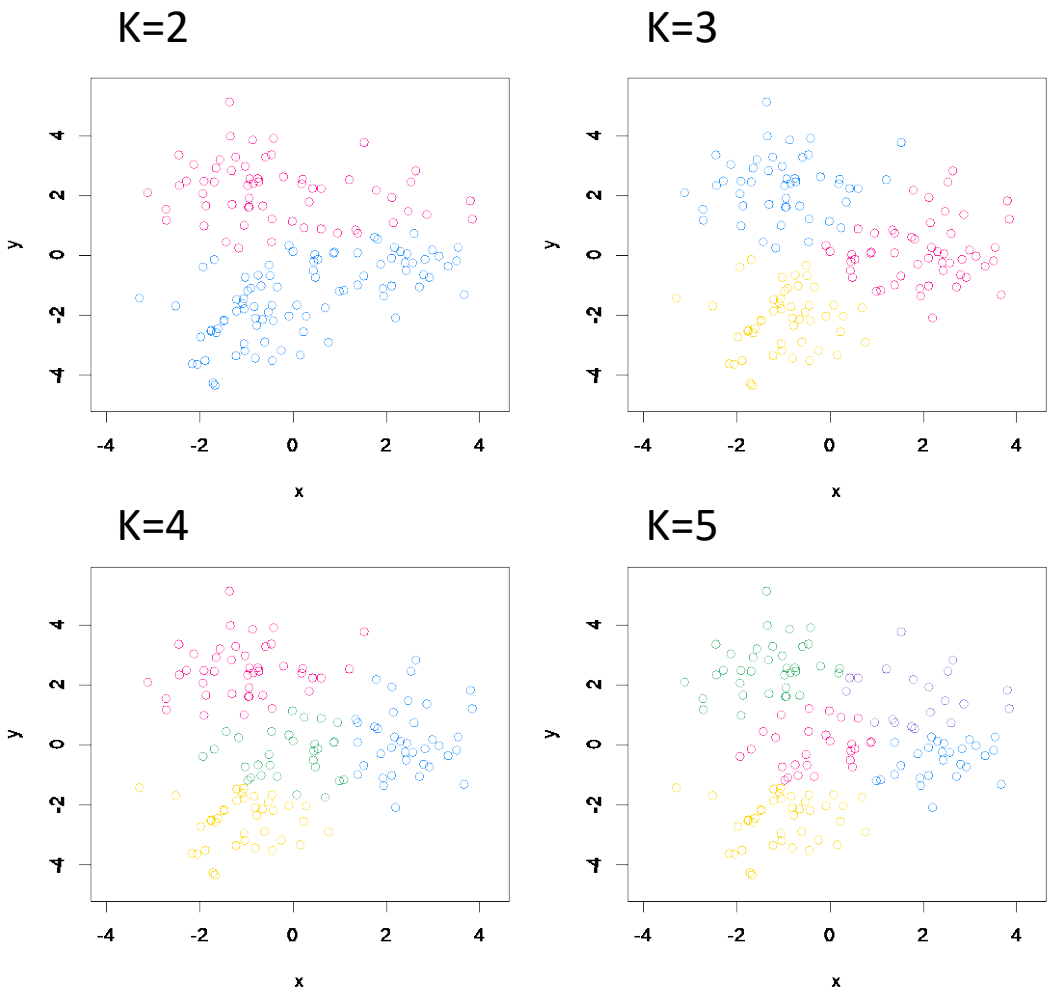
Summary

- Dendrogram (tree structure)
- Bottom-up or agglomerative approach
- Linkage Method
- Cut-tree

- Linkage Method Choice: Maximum (Complete), Minimum (Single), Average, Ward
- Pro: Good visualization, Fast Computing
- Con: Biased towards outliers

Cluster Number





Cluster Number

Computational Methods

- Elbow Method
- Silhouette
- Gap Statistic
- Progeny Clustering*

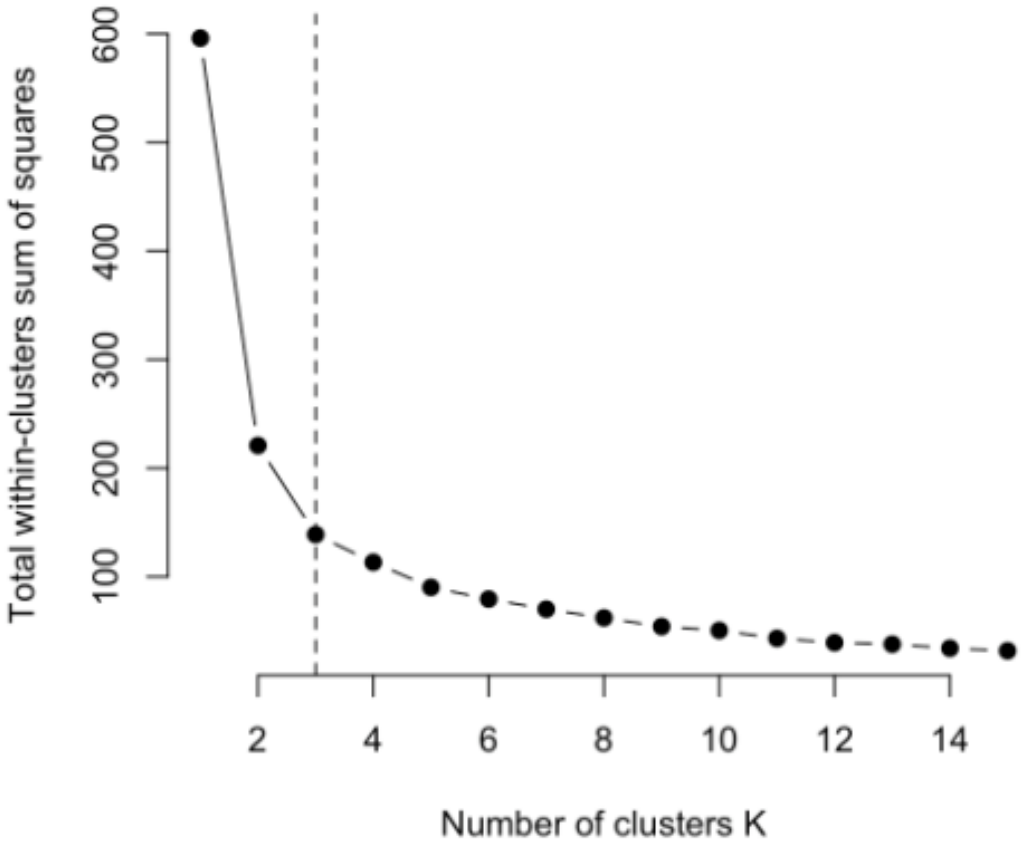
Procedure

1. Cluster data repeated with varying cluster numbers (K=2, …, 10)
2. Pick the number with best clustering quality

*Progeny Clustering (Hu et al.): https://www.nature.com/articles/srep12894

Cluster Number

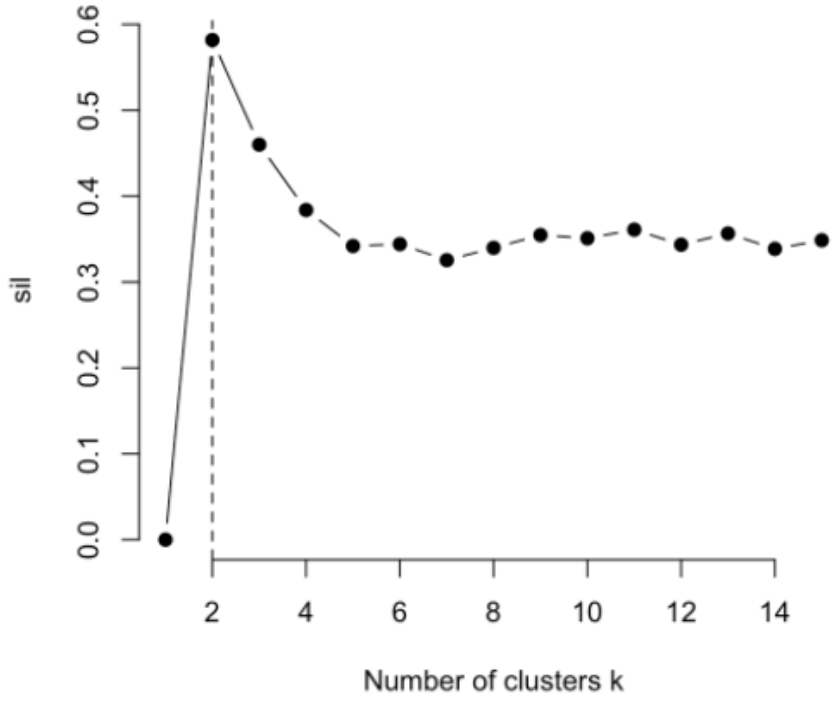
Elbow Method

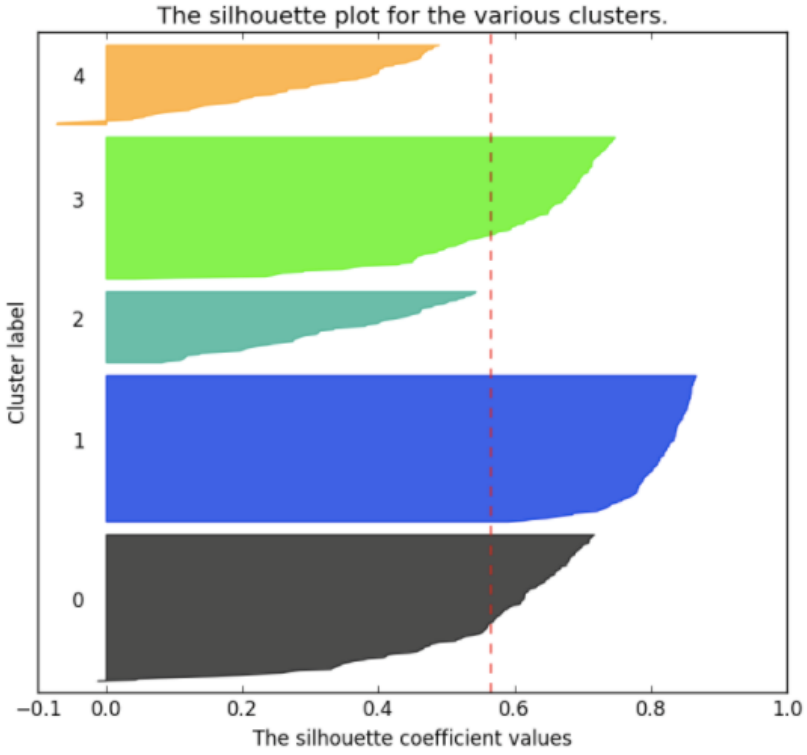


Cluster Number

Silhouette

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad -1 \leq s(i) \leq 1$$





Best Practice

- Data Normalization
  - Scaling -> Equal Variance
- Clustering Algorithm Choice
  - K-means with multiple runs
- Cluster Number Choice
  - Silhouette or your expert knowledge
- Visualization
  - Heatmaps with dendrograms or annotation bars
  - Scatter Plot with Principal Components (PCA)

Outline

- Concept
- Application
- Best Practices
- Circular Heatmap
- Tutorial

✓ what is a heatmap?

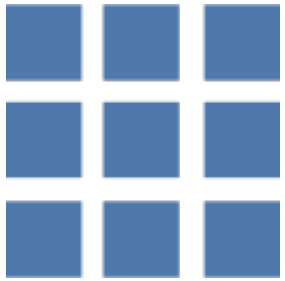
✓ when do I need a heatmap?

✓ how should I choose the color?

✓ what is a circular heatmap?

✓ I can draw heatmaps!

Concept



Map

Continuous: expression levels
Categorical: sex, mutation, condition

Matrix Values **Colors**

Application

Benefits

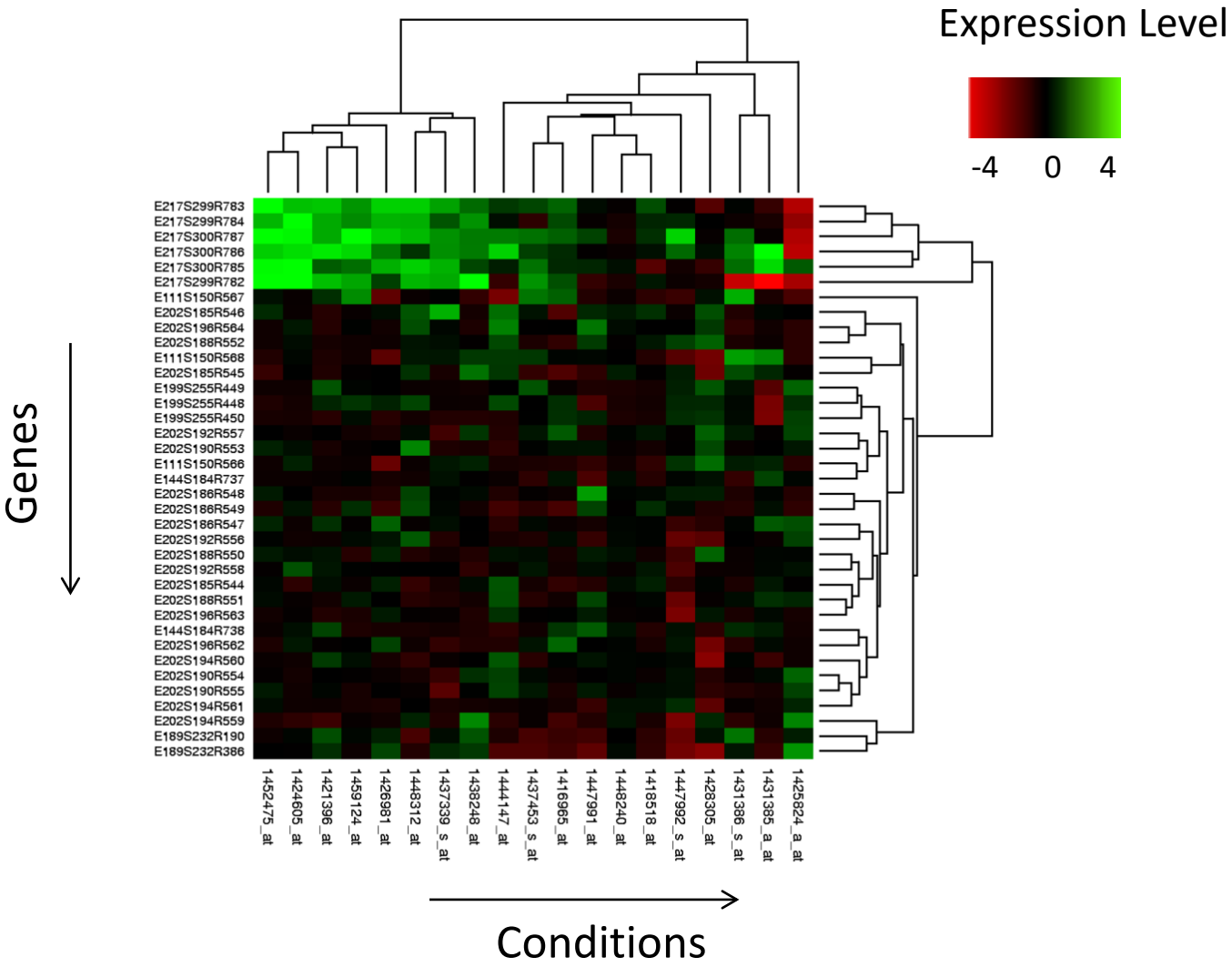
- Easy to interpret visually (color vs. text)
- Compact presentation of lots of information

Application Scenarios

- You have data of relatively large size
  - matrix or data table
- You want to see or show differences between data elements
  - accompanied by clustering

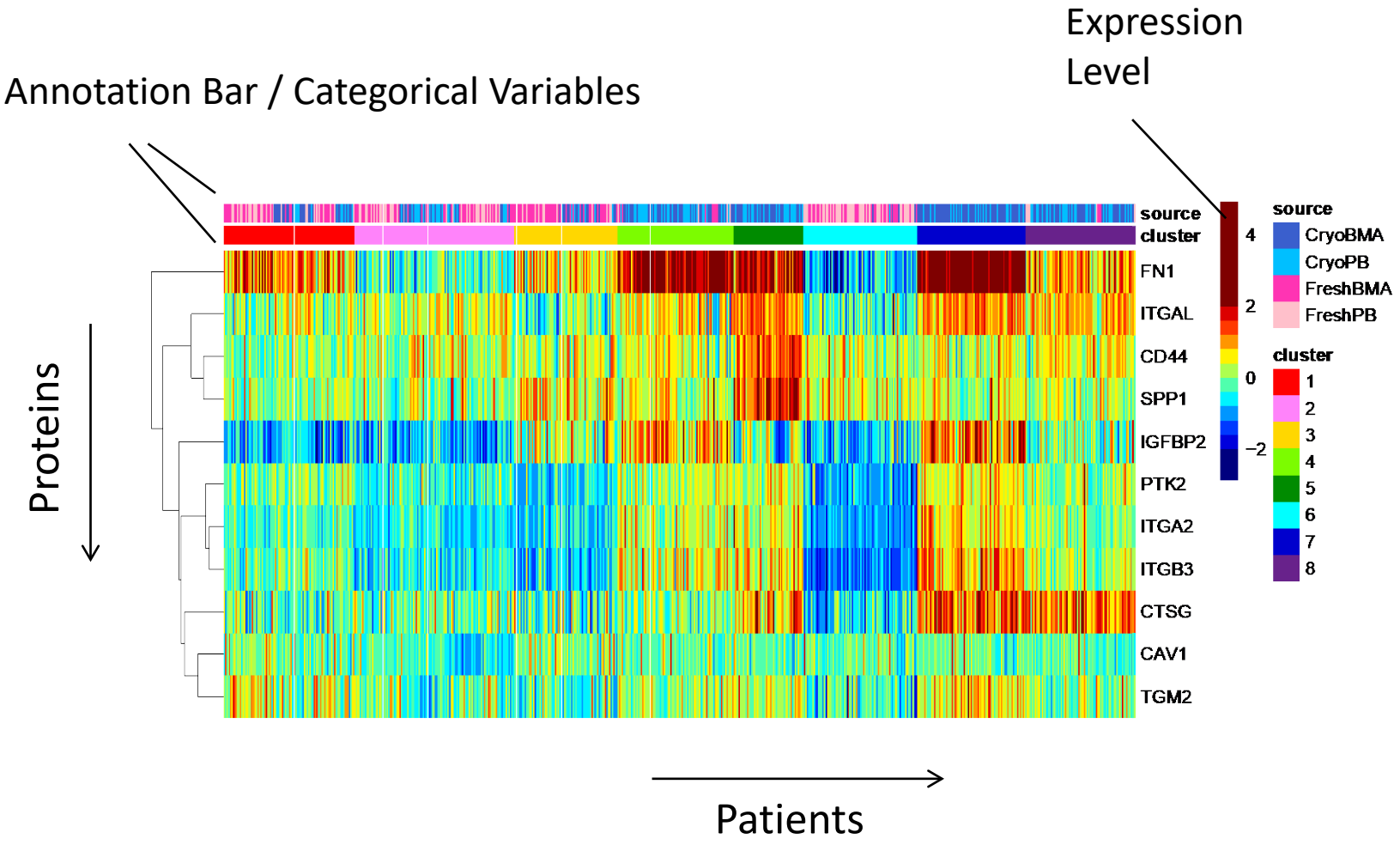
Examples

Microarray Data

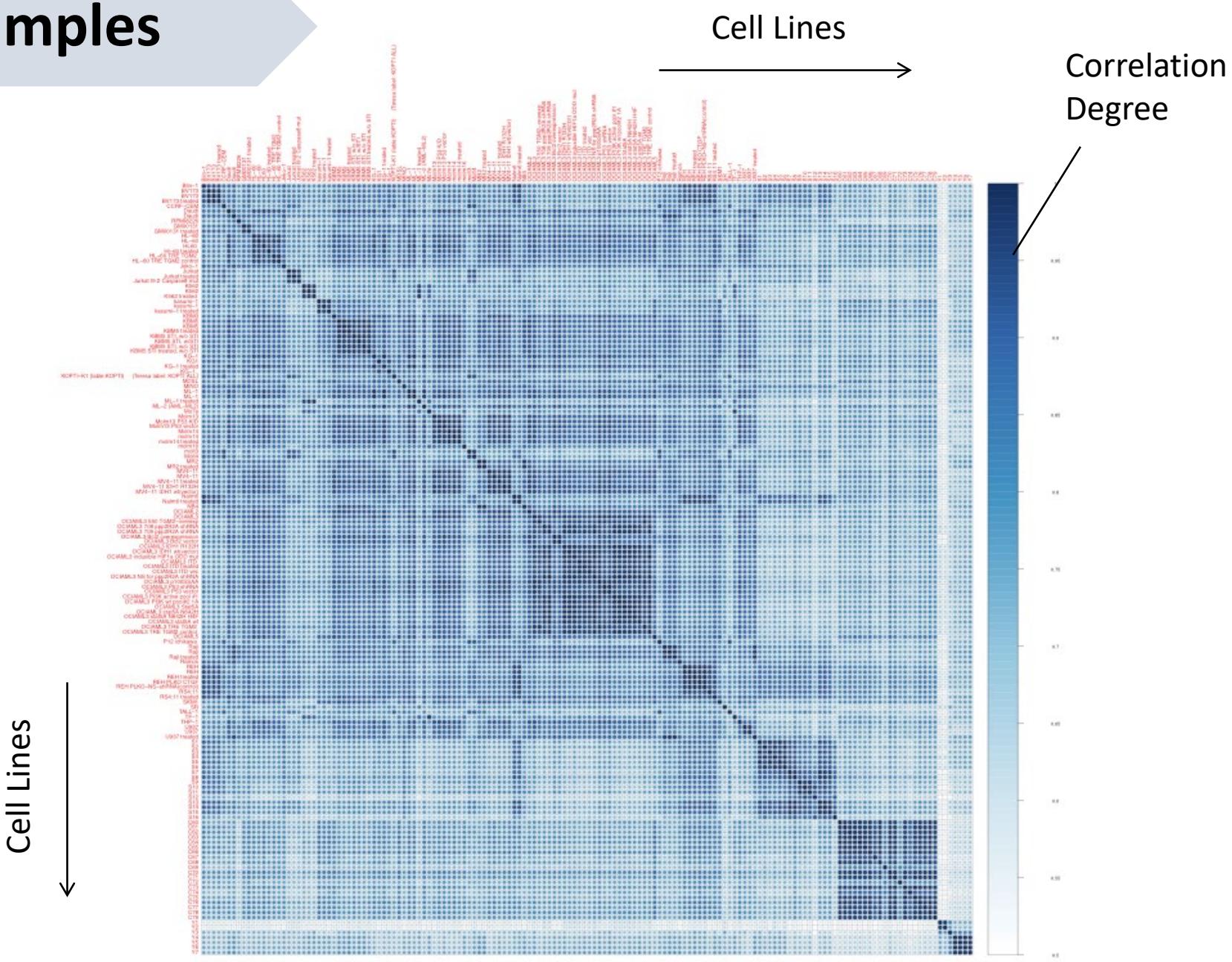


Examples

Protein Array Data



Examples



Best Practices

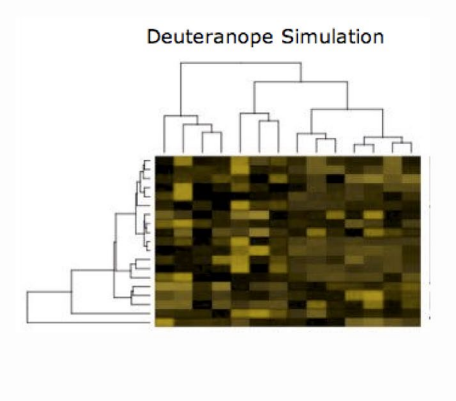
Color Choices

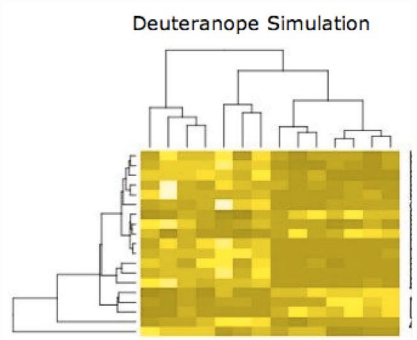
Traditional for microarray data

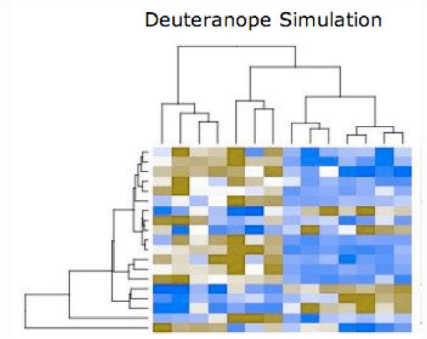
Color Blindness Consideration
vischeck.com

More shades and more perceivable detail;
Discouraged in some cases

Safe choice for print
But lack shades







Color Selection Criteria

- ✓ Display best contrast for data
- ✓ Color-blindness friendly

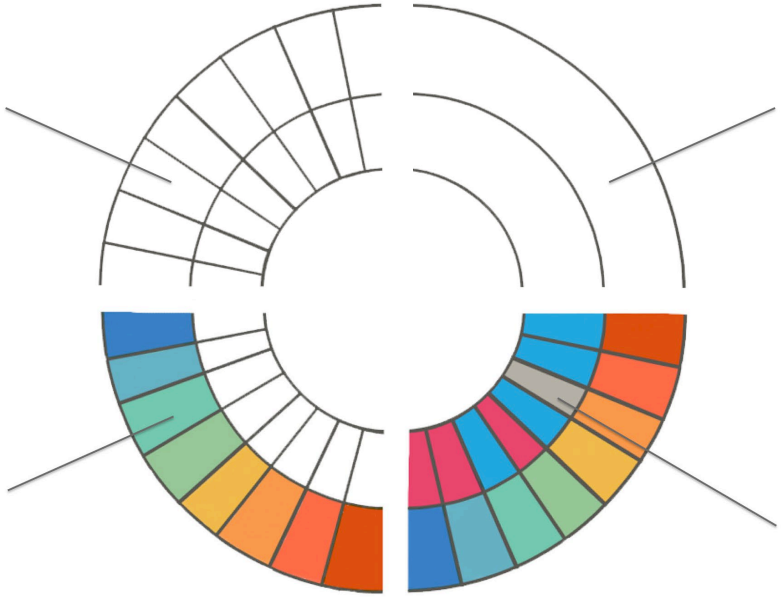
Circular Heatmap

Rows, Objects, Samples

e.g. employees, patients, students, companies, divisions, departments, products, time points, genes

Columns, Variables, Features

e.g. gender, age, disease type, test score, price, performance, material, revenue, downloads



Color Mapping for Continuous Variables

Numerical values from low to high are represented using a spectrum of colors. For example,

Color Mapping for Categorical Variables

Data categories are represented using distinct colors. For example,

Pro

- End-to-end comparison
- Center space ideal for drawing links (e.g. protein interactions)

Con

- Unbalanced focus on inner vs. outer rings

Tutorial

Biowheel (Academic Version)

- Free to use
- No programming required
- Interactive (for data exploration)
- Clustering

Data Preparation

- Data matrix in CSV or Excel format
- Apply appropriate normalization (e.g. z-score, min-max)

Clustering Options in Python Editor (of Many)

https://machinelearningmastery.com/clustering-algorithms-with-python/

For additional installation instructions specific to your platform, see:

- Installing scikit-learn

Next, let's confirm that the library is installed and you are using a modern version.

Run the following script to print the library version number.

```
# check scikit-learn version
import sklearn
print(sklearn.__version__)
```

Running the example, you should see the following version number or higher.

```
0.22.1
```

Clustering Dataset

Recap of today's module

Introduction to Classmates & Class Project

Class Project Options & Data Types

Introduction to Clustering High Dimensional Data

Next modules:

- Introduction to Python Open Source Packages & Importing Python Packages – 15-20 min
- Python Functions

Acute Myeloid Leukemia Outcome Prediction Challenge

RICE

K2I

HATRC Houston Area Translational Research Consortium RICE UNIVERSITY

Texas Medical Center® An Organization of Non-Profit Healthcare Providers

BIOHOUSTON BUILDING THE LIFE SCIENCE FUTURE

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

IBM Research

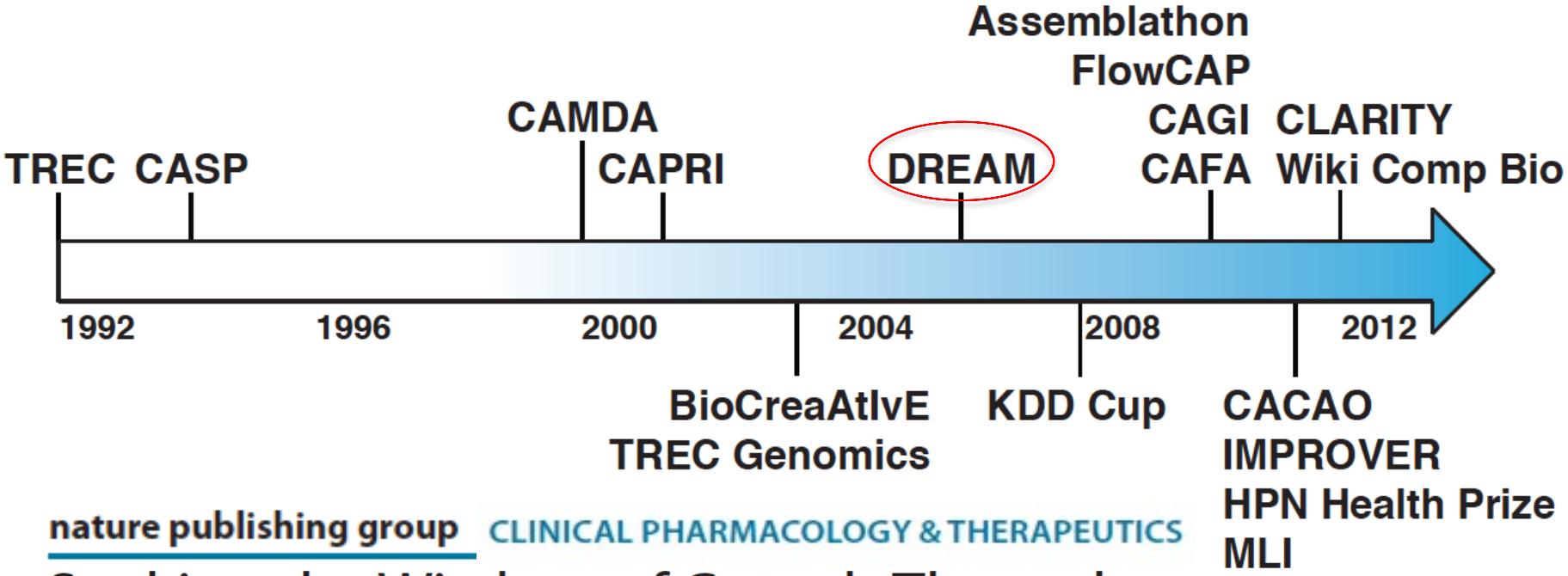
PLoS

DiBS

Sage BIONETWORKS

Challenges in Computational Bio

History of DREAM



nature publishing group CLINICAL PHARMACOLOGY & THERAPEUTICS

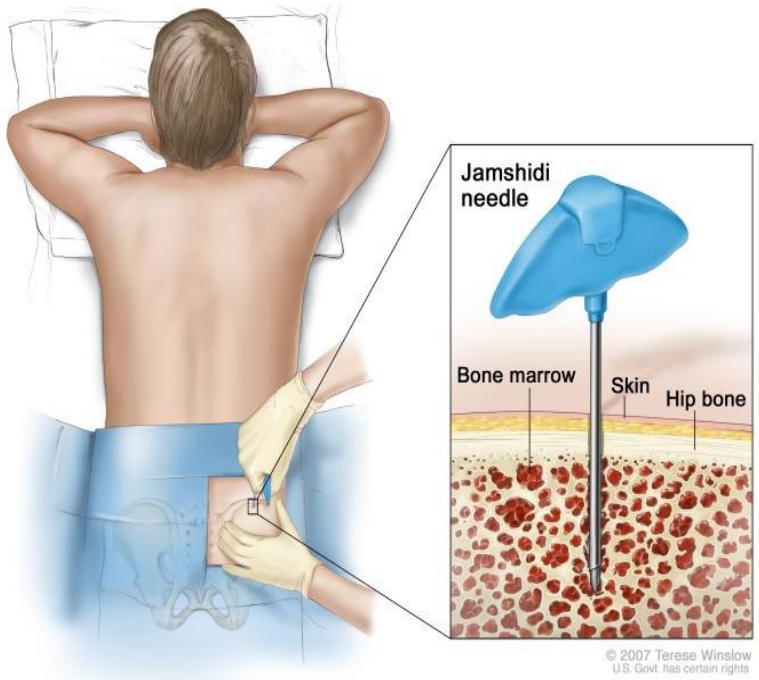
Seeking the Wisdom of Crowds Through Challenge-Based Competitions in Biomedical Research

JC Costello[1] and G Stolovitzky[2]

Acute Myeloid Leukemia

A Heterogeneous Disease

AML = "A myeloid leukemia that is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells"



Acute Myeloid Leukemia

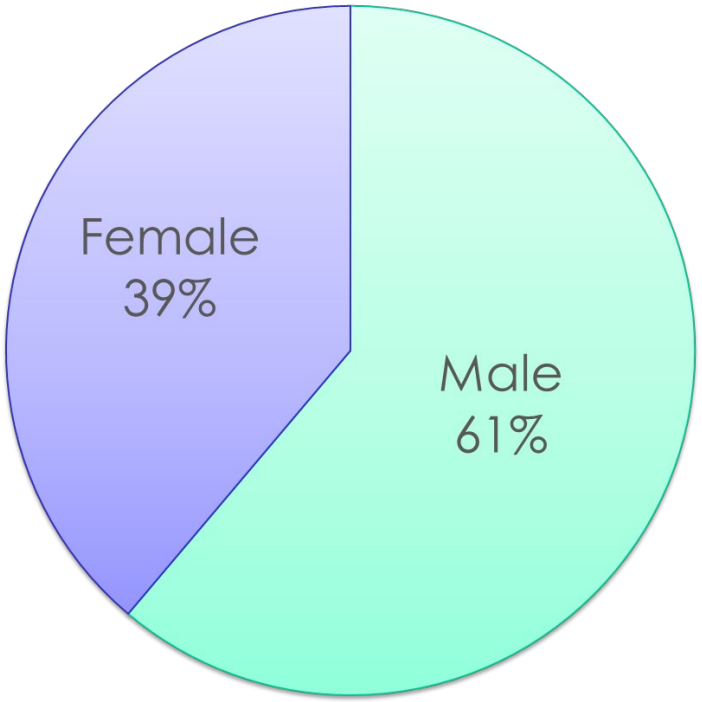
A Heterogeneous Disease

~ 18,600 new cases /year

~ 10,500 deaths (2013)

~ 25% cured

~ 65 median age



Definitions

Patient Disease Outcomes

Primary Resistant: Primary resistant to the given therapy. The patient fails to respond after one or two cycles of induction.

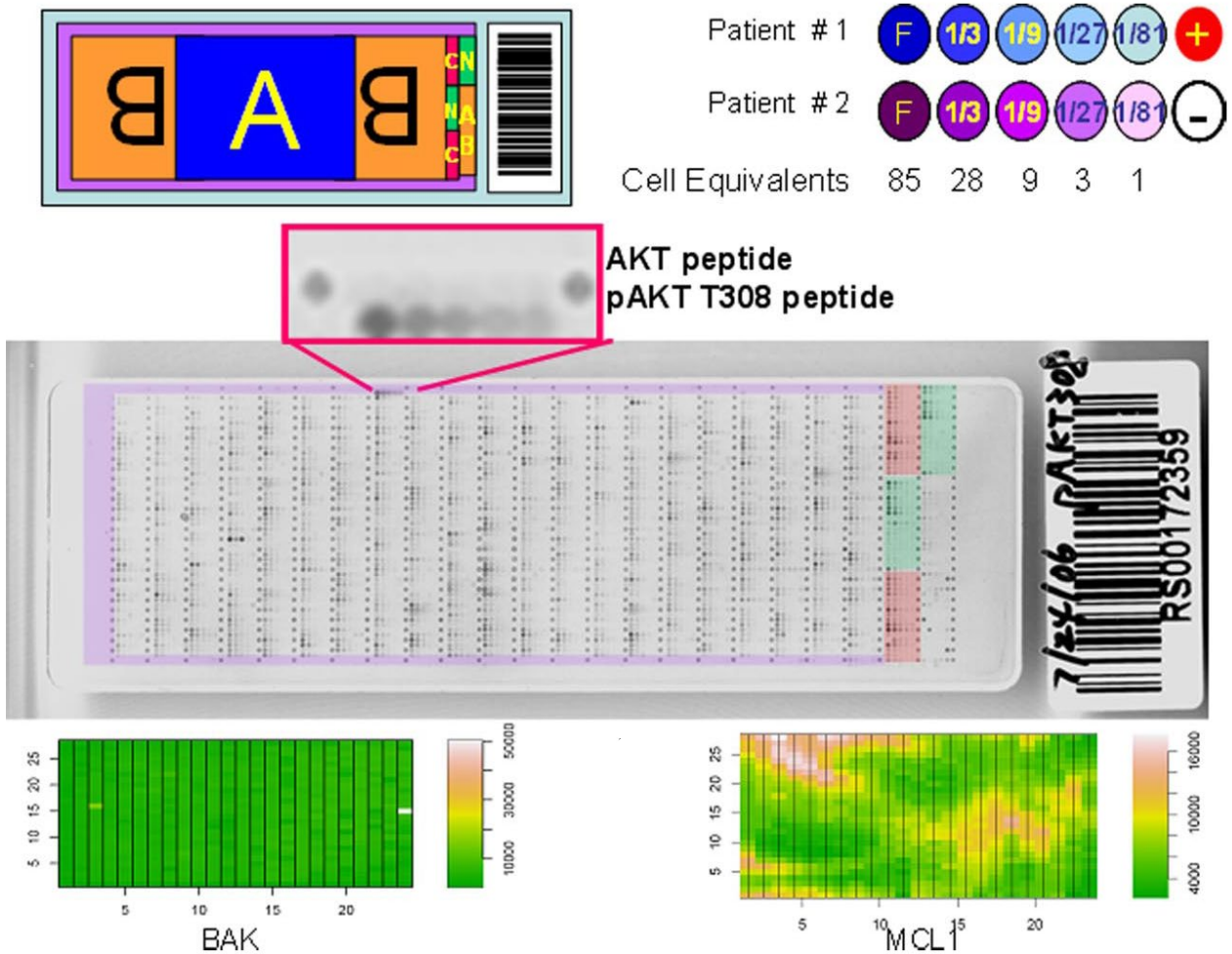
Complete Remission: No signs or symptoms of the disease. This includes normal peripheral blood cell count (absolute neutrophil count >1,000/mm^3 and platelet count >100,000/mm^3) and normocellular marrow with less than 5% blasts in the marrow.

Remission Duration: Length of remission (in weeks) for those patients who achieve Complete Remission. Remission ends if the patient relapses or dies.

Overall Survival: Length of life (in weeks) for each patient following initial diagnosis.

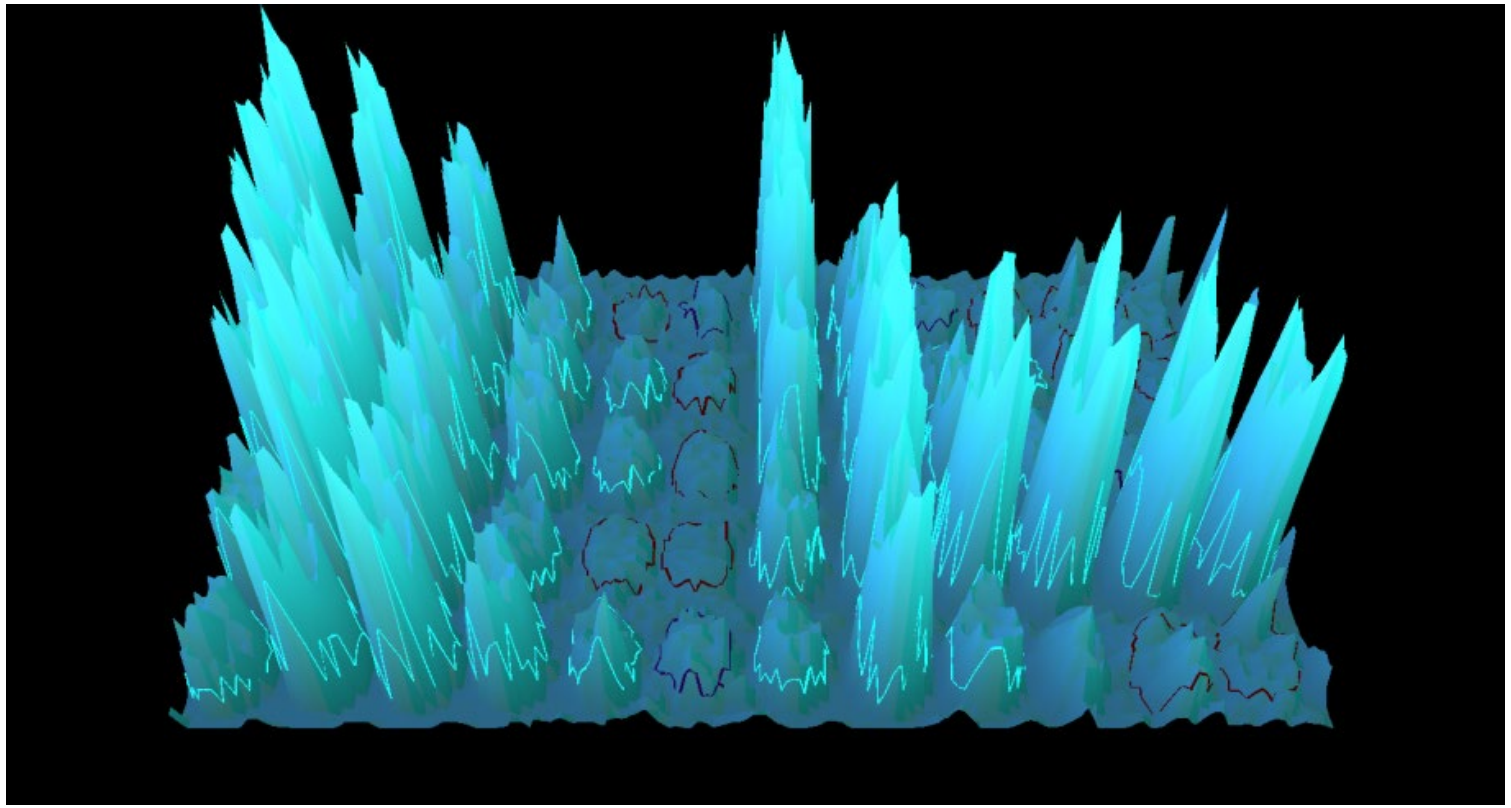
Definitions

Reverse Phase Protein Array



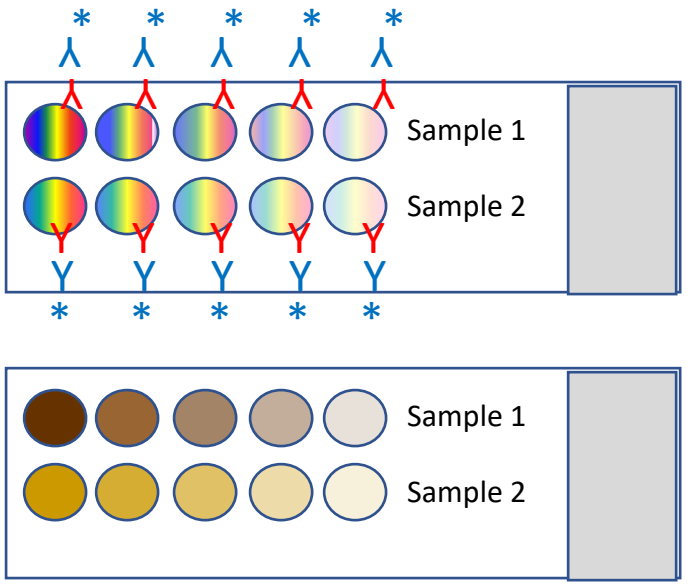
Definitions

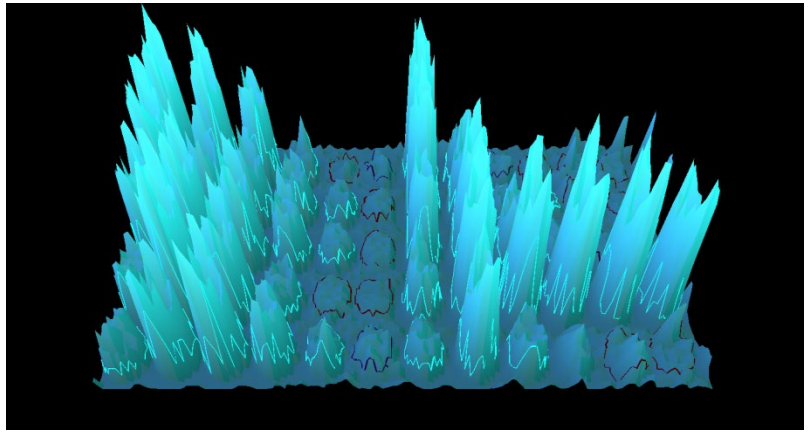
Reverse Phase Protein Array



Reverse Phase Protein Array (RPPA) Methodology

- Protein samples printed onto slide in 5 serial 1:2 dilutions (1152/slide)
- Slide probed with a validated antibody (Y)
- Slide probed with a secondary antibody with a detection signal (Y)
- DAB substrate give brown precipitate
- Slide scanned on desktop scanner





Tibes et al. (2006) Mol Cancer Therap
Kornblau et al. (2009) *Blood*