Module 3

Lecture 4

Introduction to Machine Learning: Unsupervised Learning

Sarah Groves, Ph.D.

Shayn Peirce-Cottler, Ph.D.



Learning Objectives for Module 3: Cancer

- *I can answer the question "what is cancer?" and explain what the **hallmarks** of cancer are.*
- *I can use **pandas and scikit-learn** to work with **RNA-seq data**, apply a machine learning method to a dataset, and validate the model.*
- *I understand the difference between **supervised** and **unsupervised** learning methods and know when to use them.*
- *I understand how machine learning methods use **optimization** of a cost function to determine the best answer to my research question.*
- *I know how to **avoid overfitting** and determine if my model is "good enough" using quantitative metrics.*

Where we are so far…

Week 1

~~**STEP 1: Learn** about cancer (and a specific cancer type if desired!)~~

~~**STEP 2: Explore the data-set** (you will be given one)~~

STEP 3: Ask a question related to the data set that you are curious about

Week 2

STEP 4: Visualize the data with **dimensionality reduction** (unsupervised)

STEP 5: Implement clustering, **regression**, or **classification** using Python/scikit-learn to answer the question (supervised)

Week 3

STEP 6: Verify and validate the answer using **ML metrics**

STEP 7: Draw conclusions and **reflect on the ethical implications**

Week 4

STEP 8: Communicate your question's background, methods, conclusions, limitations, implications, and future work

~~Due Thursday April 9 by class: Cancer Hallmark ABM simulations activity~~

~~Due Friday April 10 midnight: GitHub Check-in #1~~

Due Tuesday April 14 by class:
Video comprehension quiz

Due Friday April 17 midnight:
GitHub Check-in #2

Due Friday April 24 midnight:
GitHub Check-in #3

Due Tuesday May 5 midnight:
Final Project
Presentations: May 5 2pm (Final exam block)

Let’s pick genes specific to our hallmark of interest

- Pull my upstream remote branch
- Open EDA.py and in_class_DR_clustering.ipynb
  - How many total genes are in the dataset?
  - How do we visualize this?
- To get a comprehensive picture of the hallmark activity, use a “gene set”: genes with coordinated expression related to a hallmark
  - https://cancerhallmarks.com/download

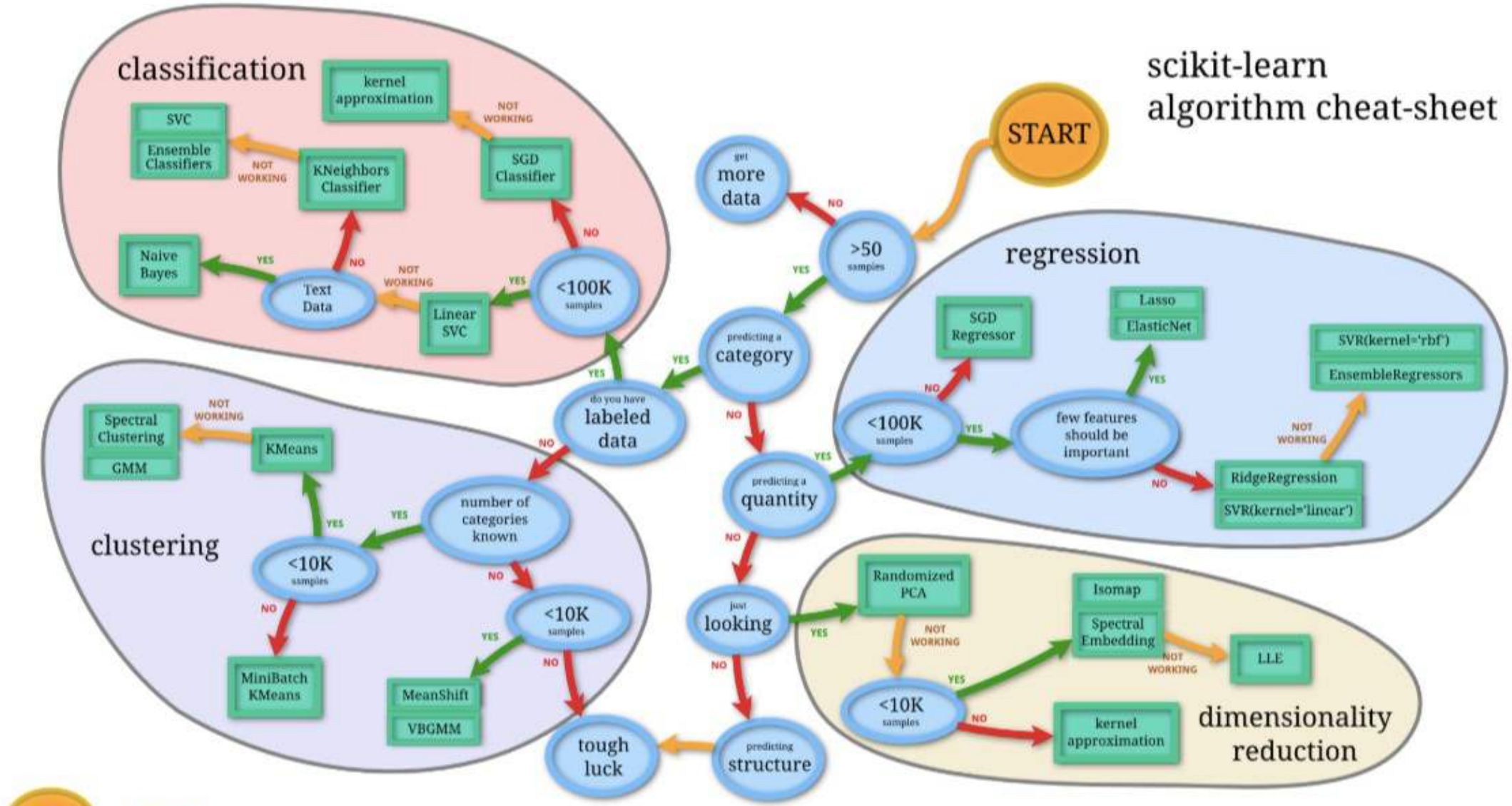
Using sci-kit learn for machine learning

Check out the User Guide on sklearn website!

conda activate BME2315

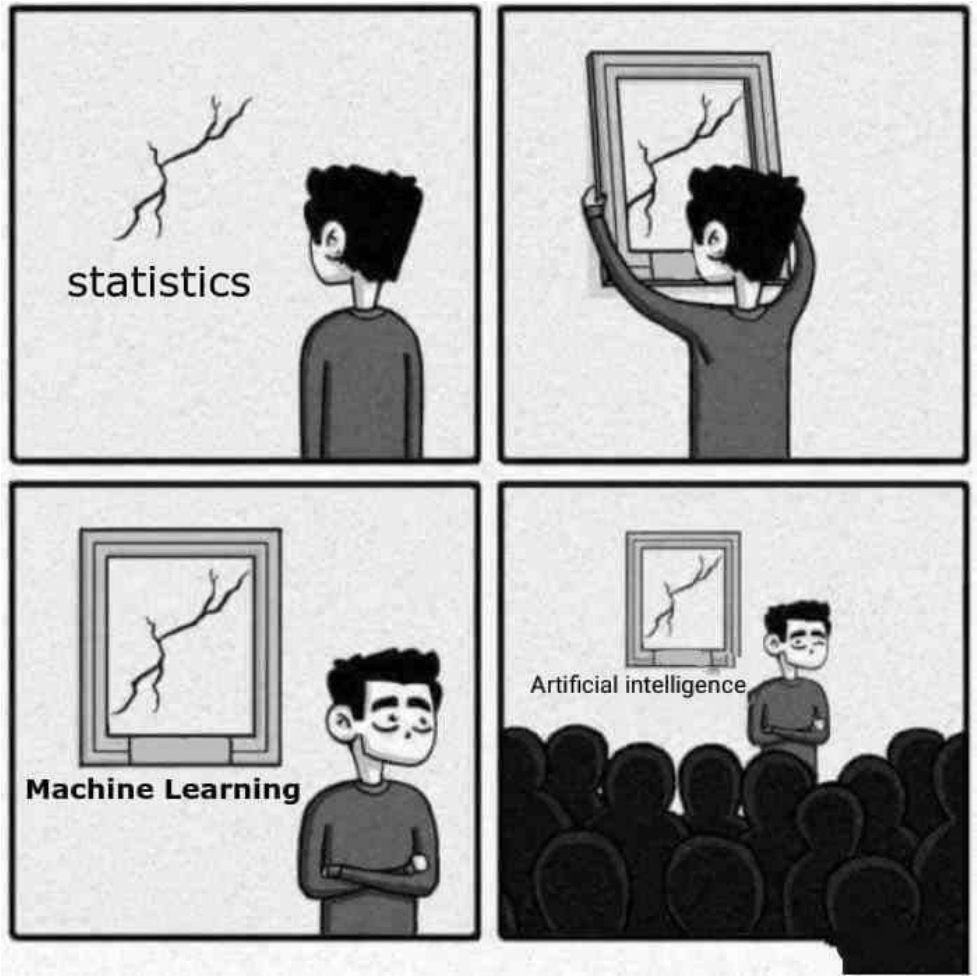
pip install scikit-learn

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
Lasso
ElasticNet
<100K samples
few features should be important
SVR(kernel='rbf')
EnsembleRegressors
RidgeRegression
SVR(kernel='linear')
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn
Screenshot

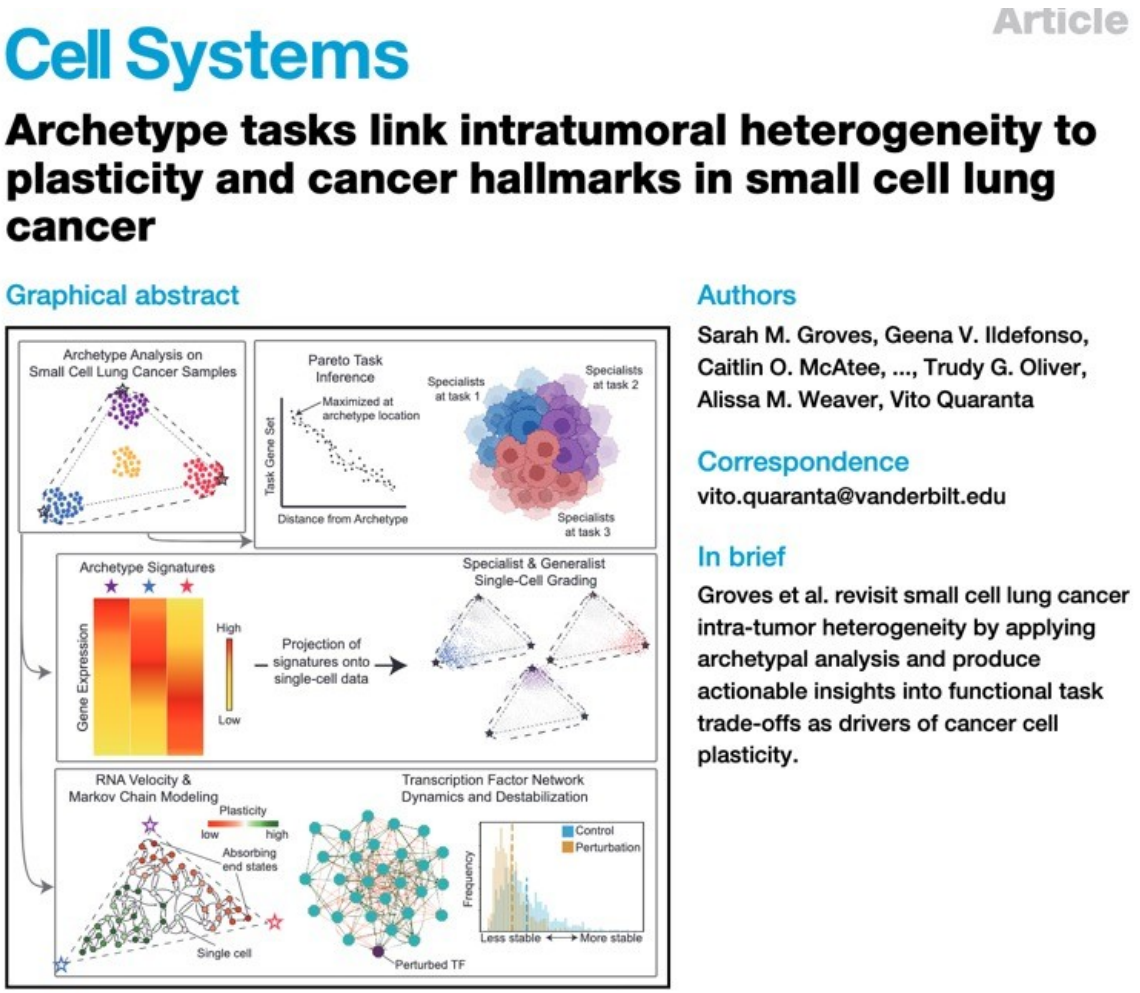
Unsupervised learning methods

- **I have “unlabeled” data**
  - No labels or “right answers” (or the labels given are irrelevant to my question)
  - I want to find patterns or structure in the data
- Dimensionality reduction
  - Preserve information while reducing features
  - PCA
  - Manifold learning (UMAP, tSNE)
- Clustering
  - K-means
  - Hierarchical clustering
  - DBSCAN
- Anomaly detection
  - Finding unusual data points



Dimensionality reduction and manifold learning

- Problem: our data has thousands of features (genes)
  - Even after subsetting to a hallmark,
  - We need a way to look at the “most important” dimensions in 2D or 3D
- “What gene shows the most variable expression across breast cancer?” (PCA)
- “Is there a pattern of continuous cell identity (i.e. plasticity) connecting small cell lung cancer subtypes?” (manifold learning)



Article

Cell Systems

Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancer

Graphical abstract

Authors

Sarah M. Groves, Geena V. Ildefonso, Caitlin O. McAtee, ..., Trudy G. Oliver, Alissa M. Weaver, Vito Quaranta

Correspondence

vito.quaranta@vanderbilt.edu

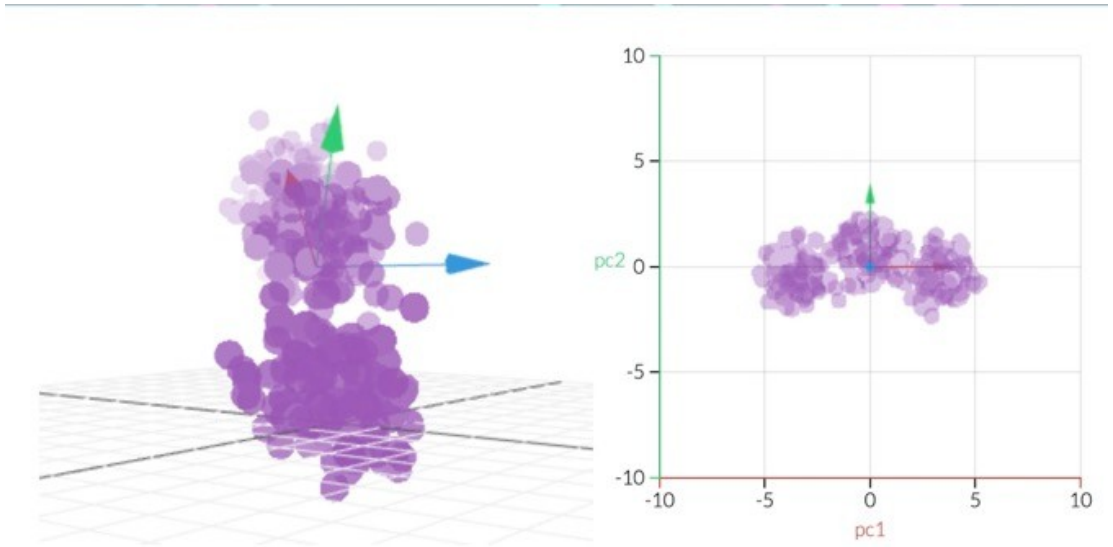
In brief

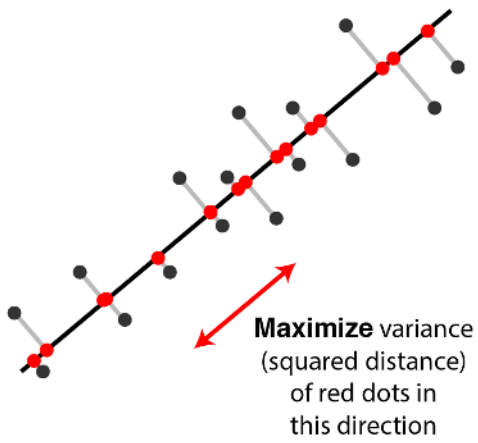
Groves et al. revisit small cell lung cancer intra-tumor heterogeneity by applying archetypal analysis and produce actionable insights into functional task trade-offs as drivers of cancer cell plasticity.

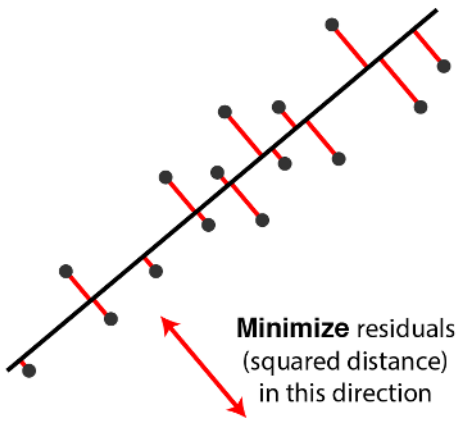
WE ARE ALL GOING TO USE DIMENSIONALITY REDUCTION TO VISUALIZE OUR DATA!

Matrix Factorization/Decomposition: PCA

- Find the “directions” in the dataset that give the most information (have the most variance)
  - Done by “decomposing” the matrix of data (genes/features x samples) into new features (PC1, PC2, etc)
  - Each PC is a combination of the old features
  - Each PC is orthogonal (90deg) from others
  - Maximizing variance ~ **minimizing the distance of the projection in a least-squares sense**
- Very useful for visualizing high-dimensional data
  - Can plot things like cluster labels on top of the PCA for communicating clustering results visually

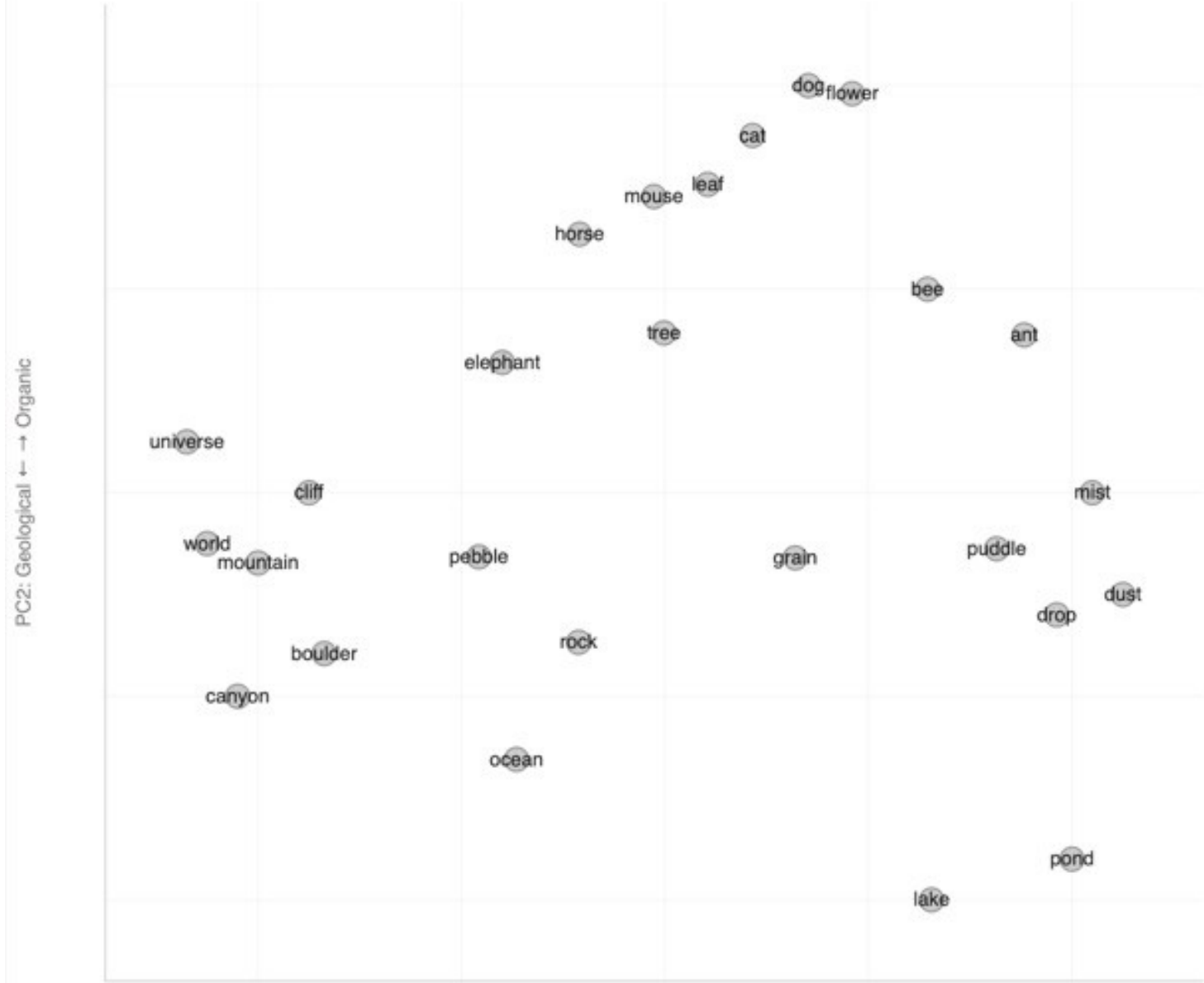




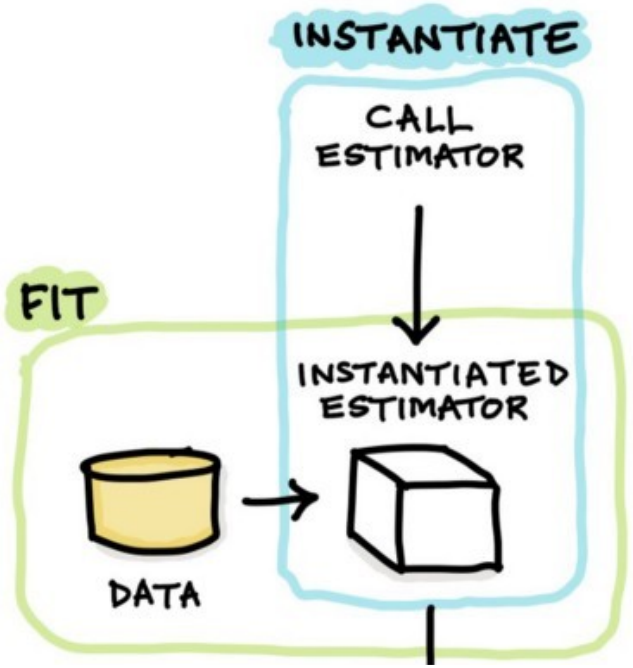


https://setosa.io/ev/principal-component-analysis/

https://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/


dog
flower
cat
leaf
mouse
horse
bee
tree
ant
elephant
universe
cliff
mist
world
mountain
pebble
grain
puddle
dust
drop
rock
boulder
canyon
ocean
pond
lake
PC2: Geological ← → Organic
PC1: Big ← → Small

How to use ANY sklearn model/estimator




```
from sklearn.modulename import EstimatorName
model = EstimatorName()
model.fit(X_train, y_train)
model.predict(X_test)
model.score(X_test, y_test)
```

- X_train = features you want to train the model on
- y_train = if you have labeled data, this is the “target”
- X_test & Y_test = new dataset

How to use PCA from the sklearn.decomposition module

```
X = data # features X samples

#instantiate model
pca = PCA(n_components=2)

#fit the PCA to data AND transform that same data
X_pca = pca.fit_transform(X)

#X_pca = transformed dataset
#pca.Attribute = info about model
```

Understanding sci-kit learn via the API

```
conda activate BME2315
pip install scikit-learn
```



https://www.menti.com/al29szz58jmf

Manifold learning: UMAP

- High-dimensional data often lies on a lower-dimensional, *non-linear* “manifold” (like a curved surface)
- “Fuzzy graph” where each data point is connected to its nearest neighbors
  - Preserves high-dimensional structure
- Optimize the layout of the graph in 2D

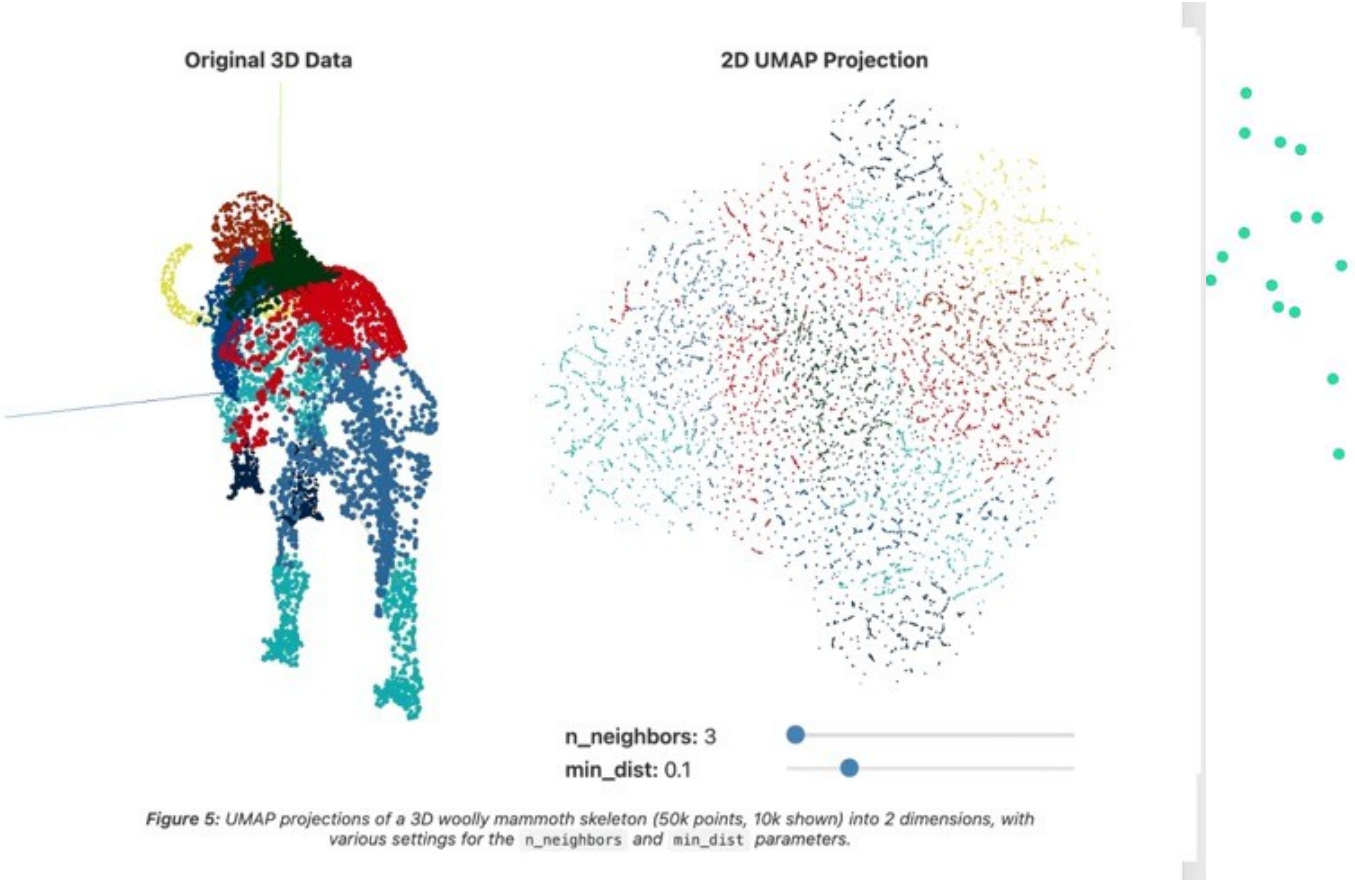


Figure 5: UMAP projections of a 3D woolly mammoth skeleton (50k points, 10k shown) into 2 dimensions, with various settings for the `n_neighbors` and `min_dist` parameters.

https://biostatsquid.com/umap-simply-explained/

https://pair-code.github.io/understanding-umap/

Where we are so far…

Week 1

~~**STEP 1: Learn** about cancer (and a specific cancer type if desired!)~~

~~**STEP 2: Explore the data-set** (you will be given one)~~

Week 2

STEP 3: Ask a question related to the data set that you are curious about

STEP 4: Visualize the data with **dimensionality reduction** (unsupervised)

Week 3

STEP 5: Implement clustering, **regression**, or **classification** using Python/scikit-learn to answer the question (supervised)

STEP 6: Verify and validate the answer using **ML metrics**

Week 4

STEP 7: Draw conclusions and **reflect on the ethical implications**

STEP 8: Communicate your question's background, methods, conclusions, limitations, implications, and future work

~~Due Thursday April 9 by class: Cancer Hallmark ABM simulations activity~~

~~Due Friday April 10 midnight: GitHub Check-in #1~~
~~Due Tuesday April 14 by class: Video comprehension quiz~~

Due SUNDAY APRIL 19 midnight:
GitHub Check-in #2

Due SUNDAY APRIL 26 midnight:
GitHub Check-in #3

Due Tuesday May 5 midnight:
Final Project
Presentations: May 5 2pm (Final exam block)

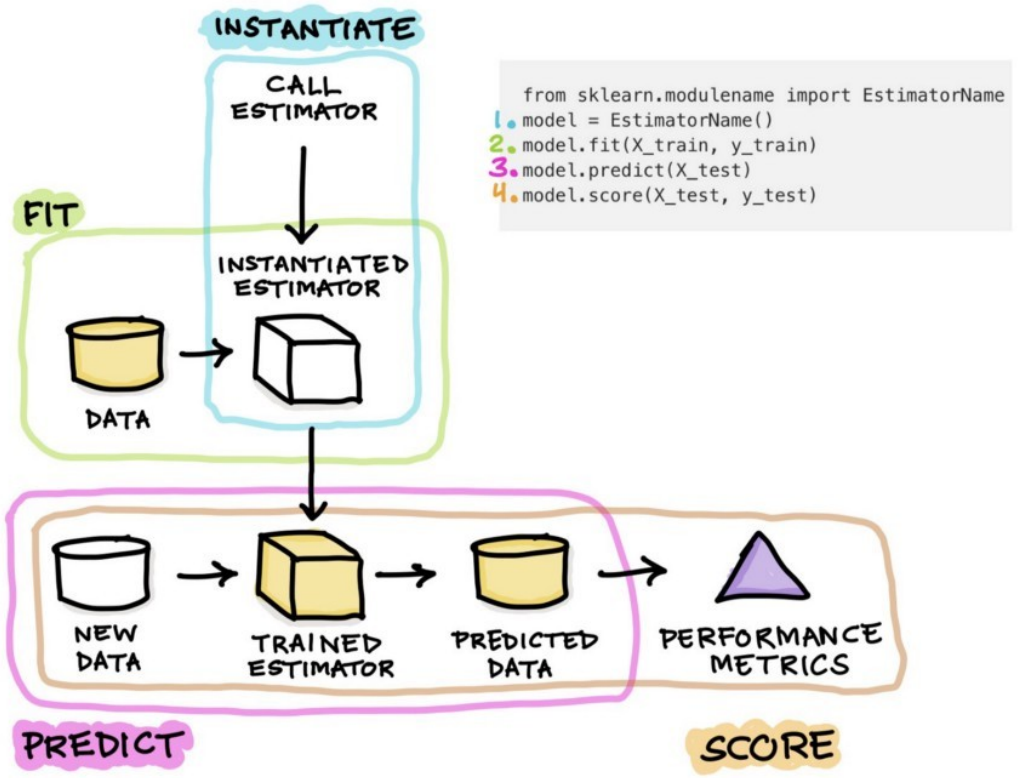
Recap from last class

PCA

- Find axes of highest variance(pca.fit()) and rotate the datato view it from the new axes (pca.transform())
- Each new axis is meaningful and tells you something about the underlying patterns in the features

UMAP

- Perform a *nonlinear* transformation by:
  - building a “fuzzy graph” that connects the neighbors in high-dimensional space
  - Figure out a good way to “smoosh” this graph into two dimensions that preserves the high-dimensional structure
- Visualizes a manifold through the data

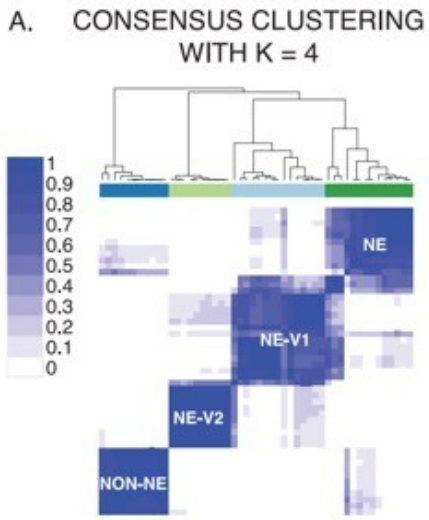


Implementing PCA with sklearn

- in_class_DR_clustering.ipynb

K-means clustering

- Used when we want to find groups of commonality in the data
- “Are there subgroups of small cell lung cancer based on the expression of key transcription factors?”
- We need to tell the algorithm how many clusters to look for



PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers

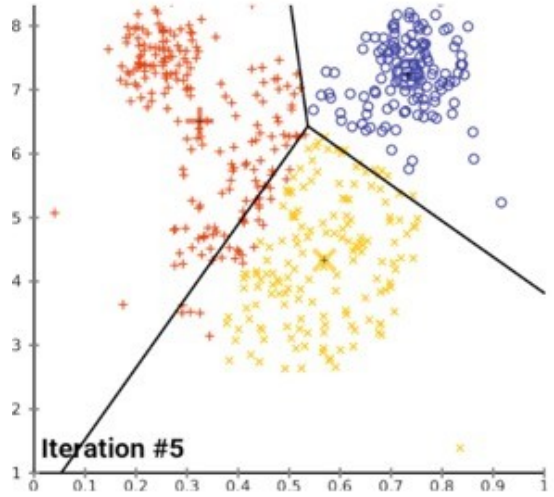
David J. Wooten[1,2☯], Sarah M. Groves[2☯], Darren R. Tyson[2], Qi Liu[3], Jing S. Lim[4], Réka Albert[1], Carlos F. Lopez[2], Julien Sage[4], Vito Quaranta[2*]

1 Department of Physics, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **3** Departments of Biomedical Informatics and Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **4** Departments of Pediatrics and Genetics, Stanford University, Stanford, California, United States of America

☯ These authors contributed equally to this work.
* vito.quaranta@vanderbilt.edu

1. Randomly place “centroids”
2. Assign each sample to the closest centroid
3. Move the centroids so they are actually in the middle of all the points assigned to them

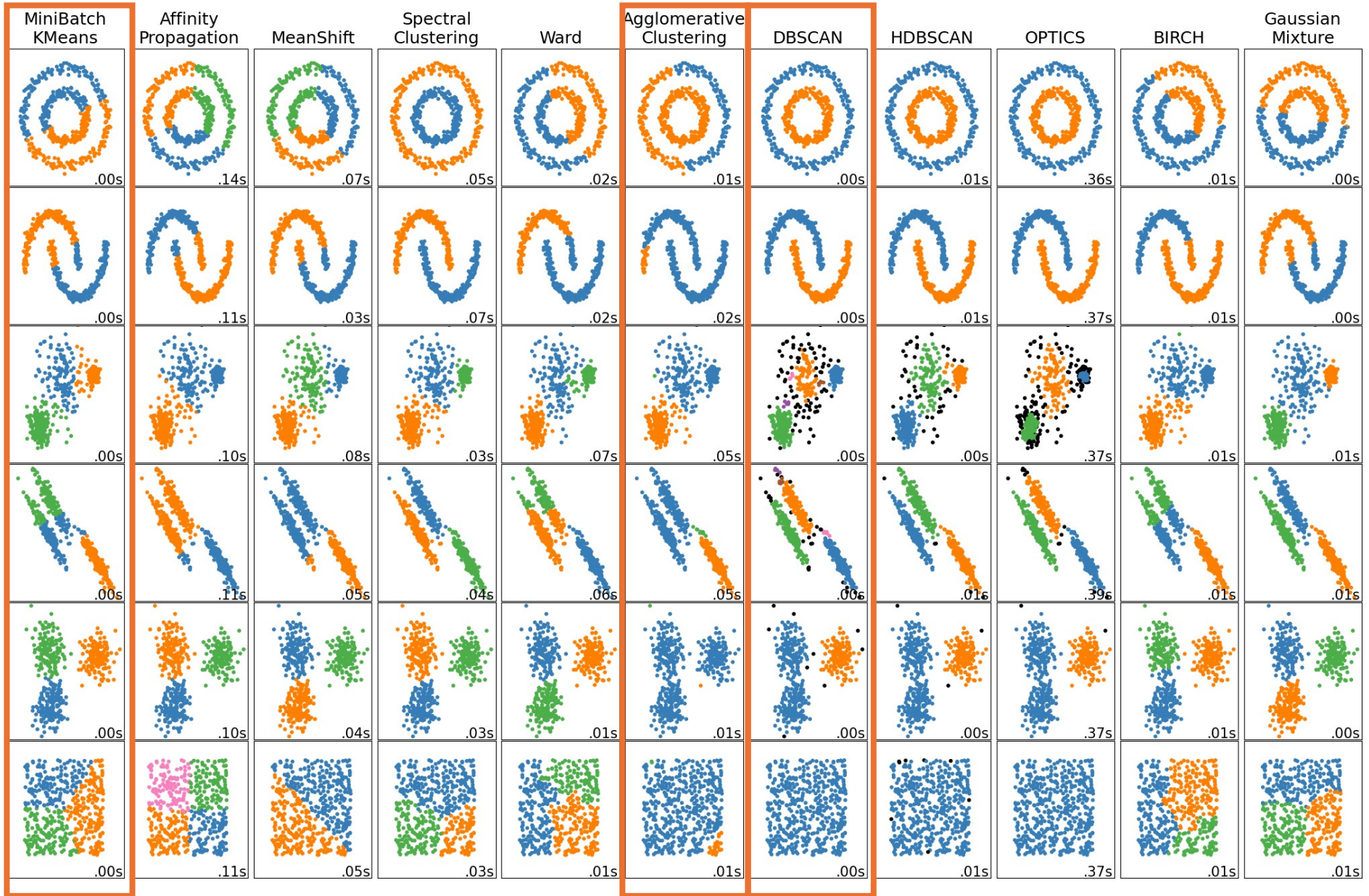
Cost function: minimize the mean distance (norm) between samples and centroids



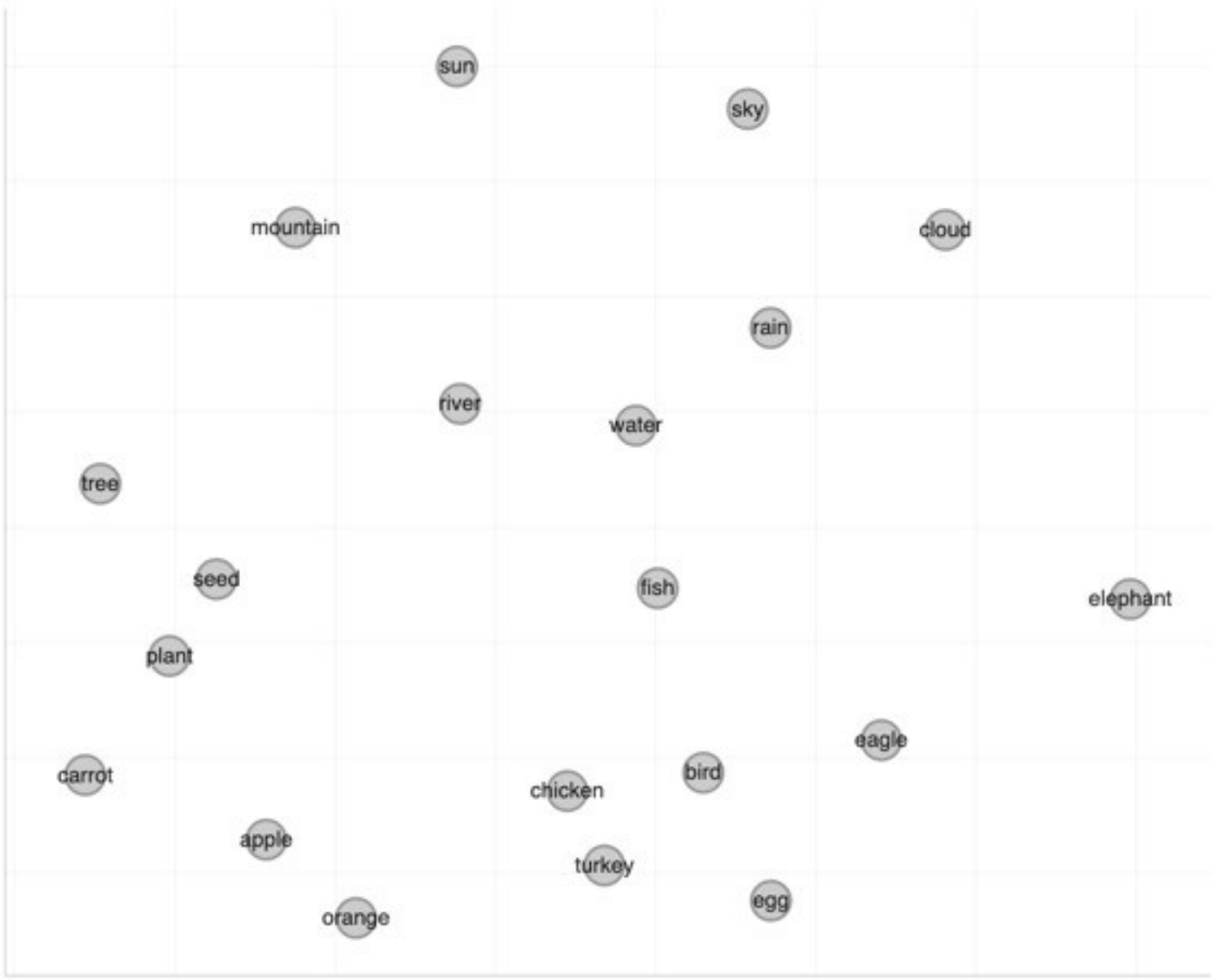
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Hierarchical clustering and DBSCAN

- HC: group most similar samples and then groups of groups
- DBSCAN: groups densely packed points while avoiding low-density regions



Hierarchical clustering and DBSCAN



Understanding sci-kit learn via the API

```
conda activate BME2315
pip install scikit-learn
```



https://www.menti.com/al29szz58jmf

In class work time on your projects

- Reduce the dataset to the cancer type(s) of interest and genes of interest related to your hallmark(s).
  - You should have >10 genes, but >100 is even better!
- Visualize your data with a PCA
  - Color the PCA by important features (at least two different views related to your project goal)
- Choose one of the following:
  - Visualize your data with a UMAP
  - Use clustering on your data and visualize the clusters on the PCA
- If you need to refine your question/project goal, do so!

2nd check-in: All the red should be completed!