Module 4

Lecture 4

Introduction to Machine Learning: Clustering & Supervised Learning

Sarah Groves, Ph.D.

Shayn Peirce-Cottler, Ph.D.



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 SUNDAY APRIL 19
midnight:
GitHub Check-in #2

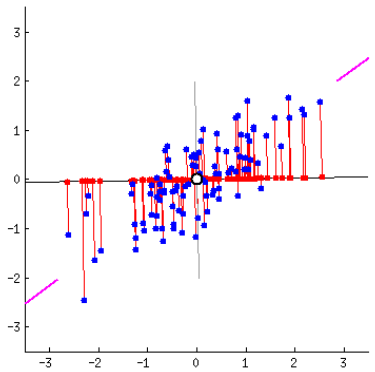
Due SUNDAY APRIL 26
midnight:
GitHub Check-in #3

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

Recap from last class

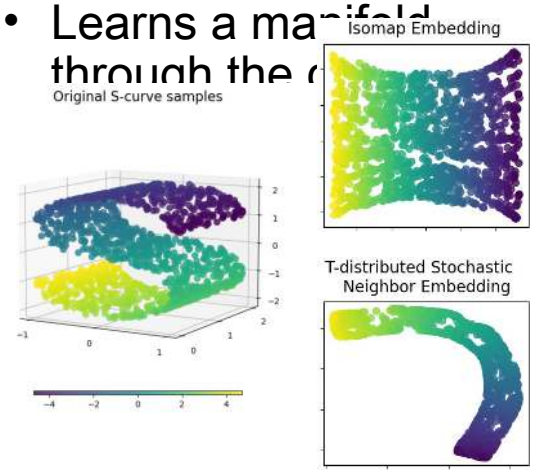
PCA

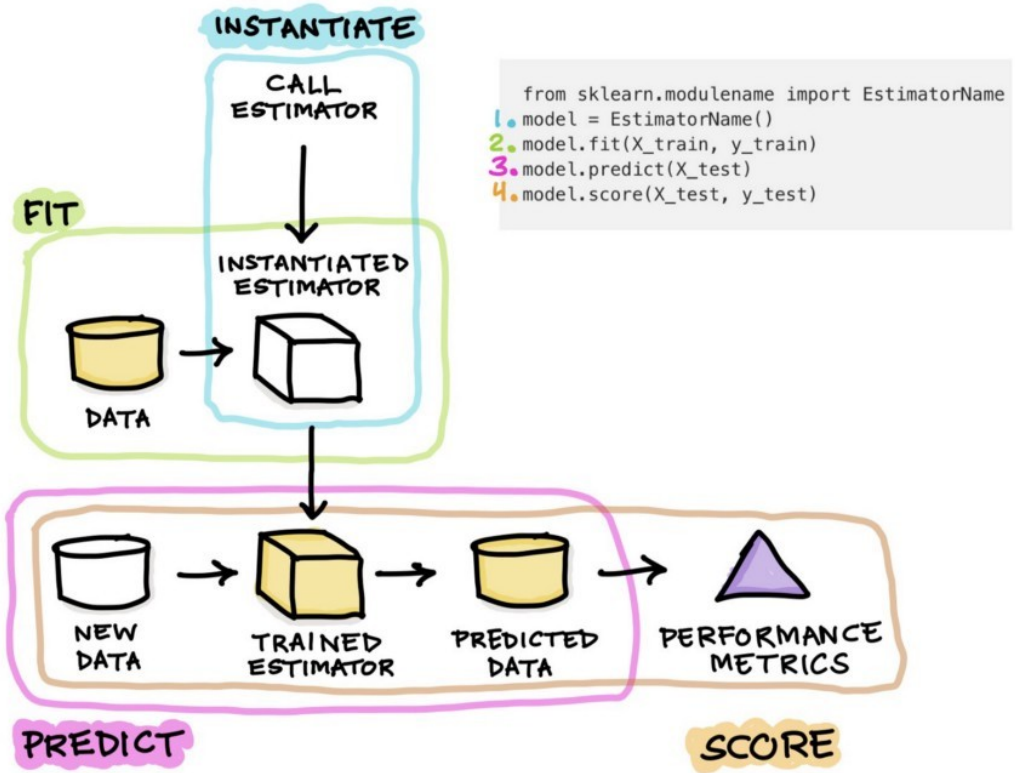
- Find axes of highest variance(pca.fit()) and rotate the datato view it from the new axes (pca.transform())
- Each new axis is meaningful and tells you something about the underlying patterns in the features



Manifold learning (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
- Learns a ma… through the …



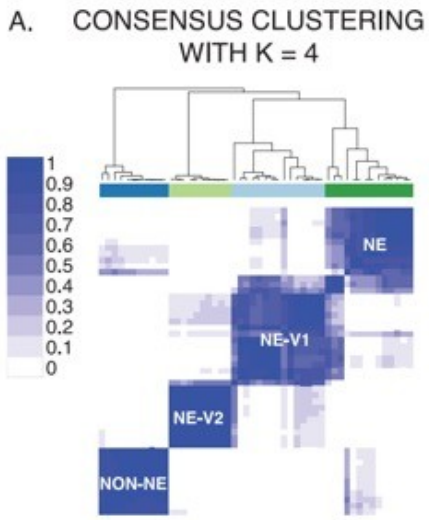


Implementing PCA with sklearn

- 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

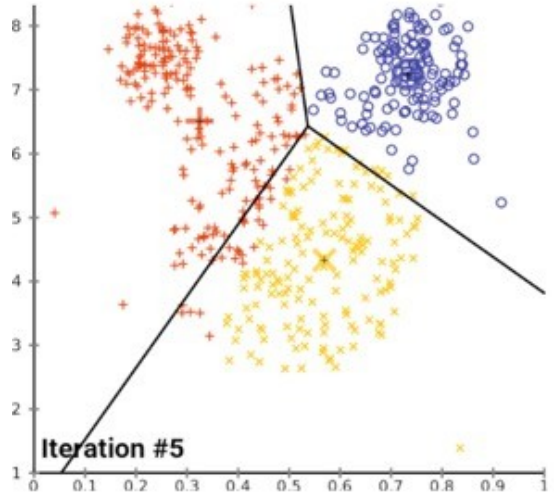
David J. Wooten[1,2☯], Sarah M. Groves[2☯], Darren R. Tyson[2], Qi Liu[3], Jing S. Lim[4], Réka Albert[1], Carlos F. Lopez[2], Julien Sage[4], Vito Quaranta[2*]

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

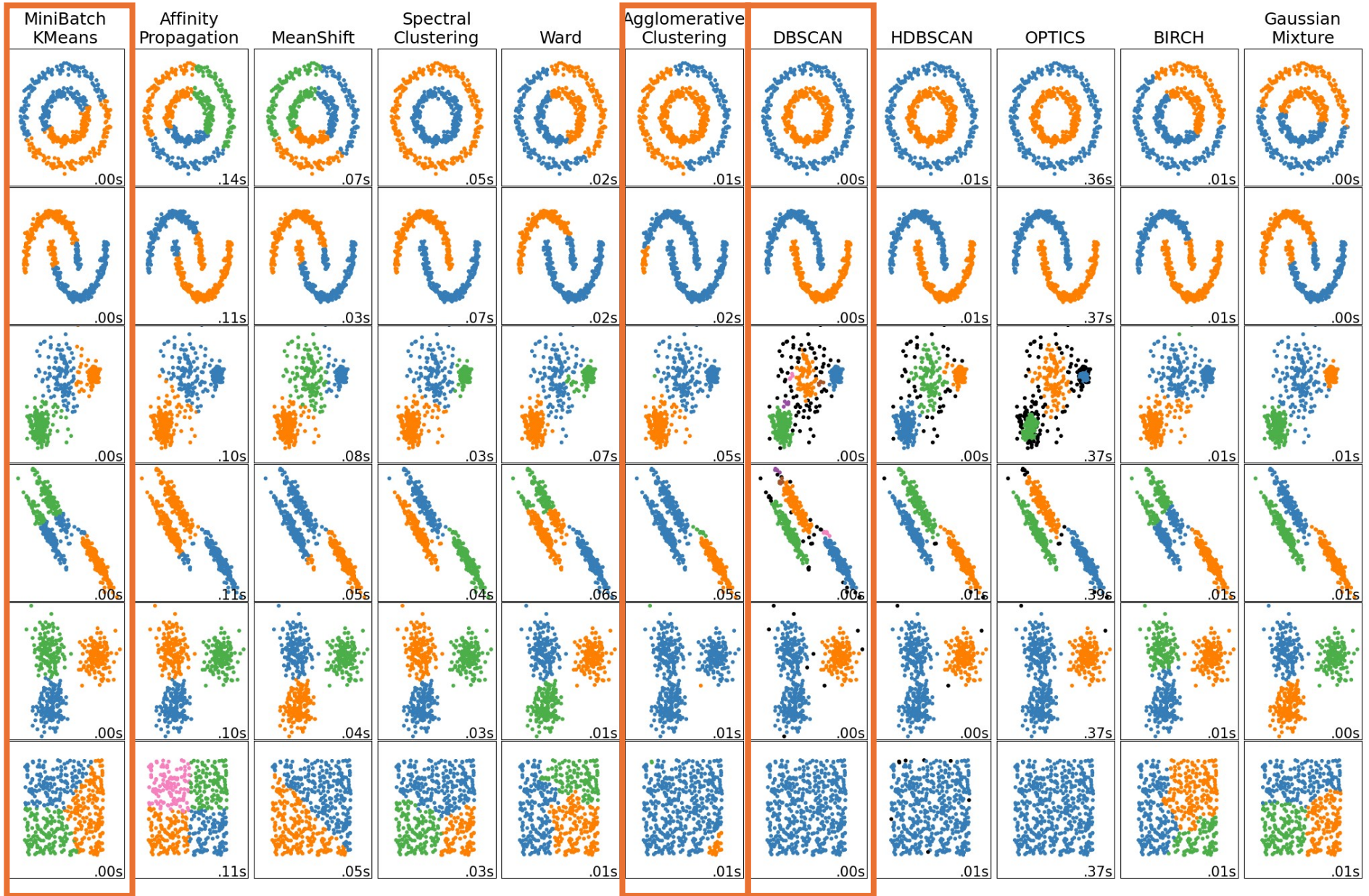
Cost function: minimize the mean distance (norm) between samples and centroids



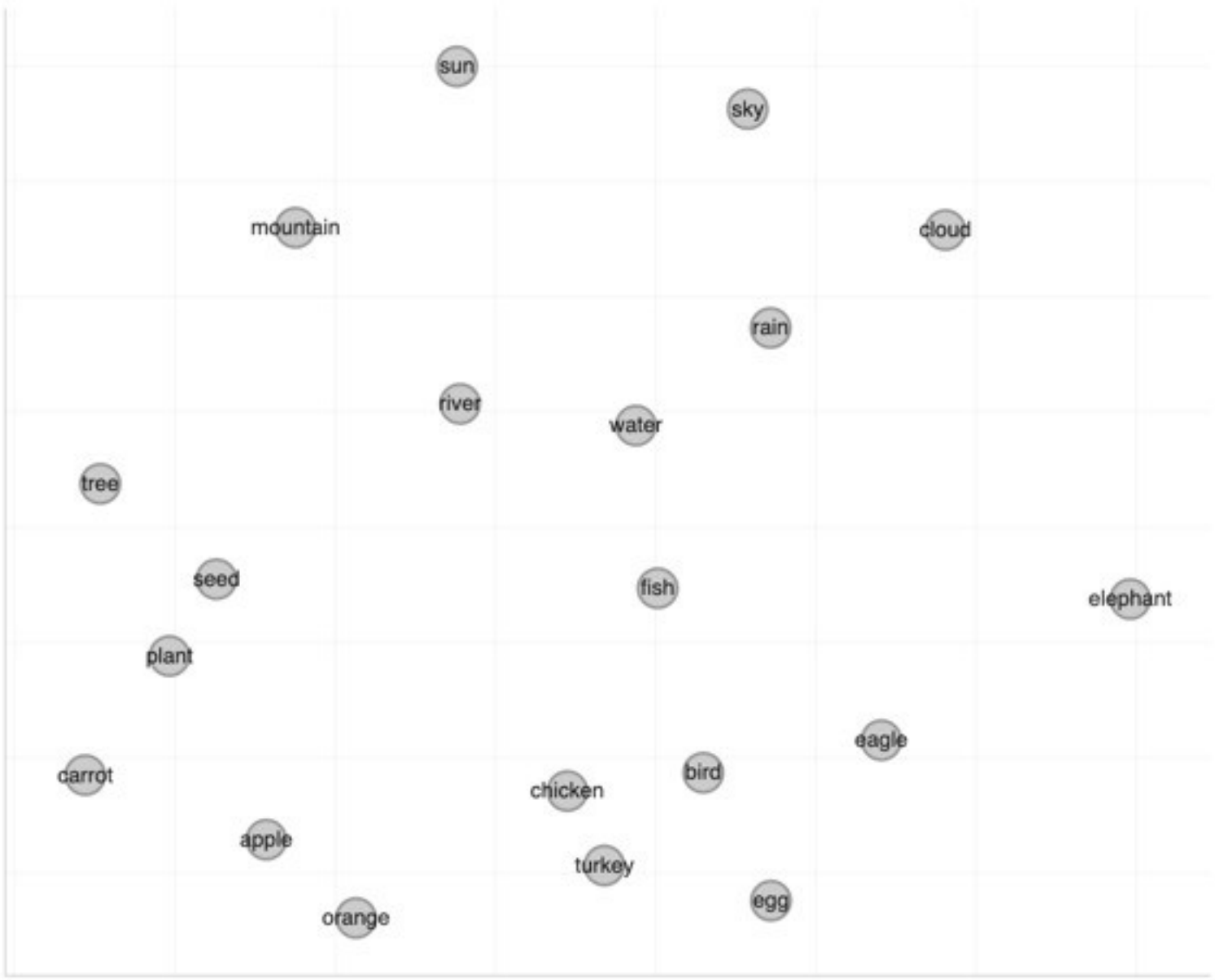
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

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

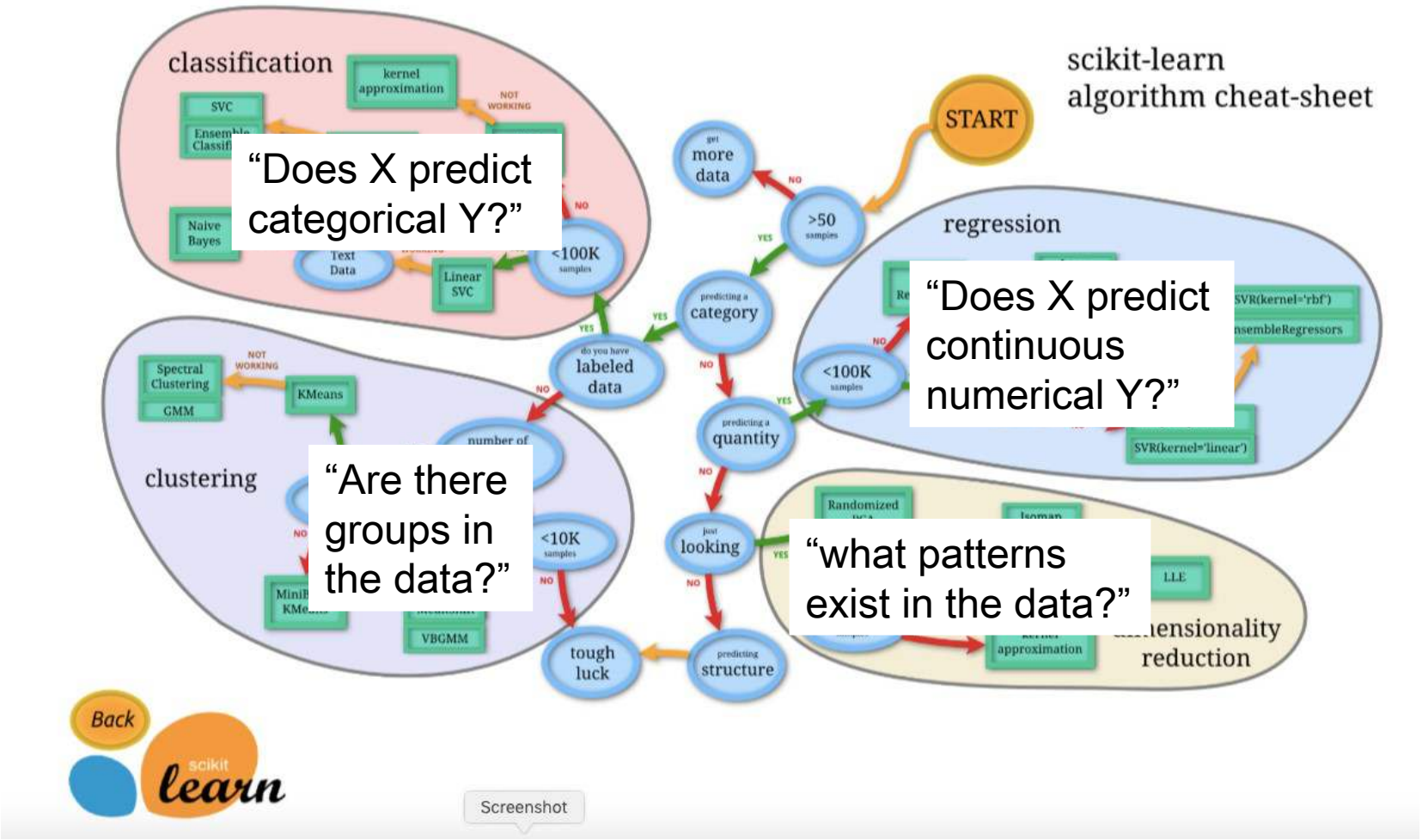
```
conda activate BME2315
pip install scikit-learn
```



https://www.menti.com/al29szz58jmf

Preview of supervised learning methods

- Data is “labeled,” and we are trying to predict the label (“target”)
- If target is a continuous numerical value → regression
- If target is categorical → classification



What category of methods will you use to answer your question?

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 at least 1 important feature related to your project goal:
    - Average expression of gene set
    - 1 key gene's expression
    - Metadata label
    - Cancer type (if using multiple cancer types)
- Choose one of the following:
  - Visualize your data with a UMAP (does this give you new/different information than the PCA?)
  - Use clustering on your data and visualize the clusters on the PCA (what do the clusters mean?)
- If you need to refine your question/project goal based on the patterns you are seeing in the data, do so!
  - You should now be able to state which method(s) you will use (clustering, classification, regression)

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