CLUSTERING

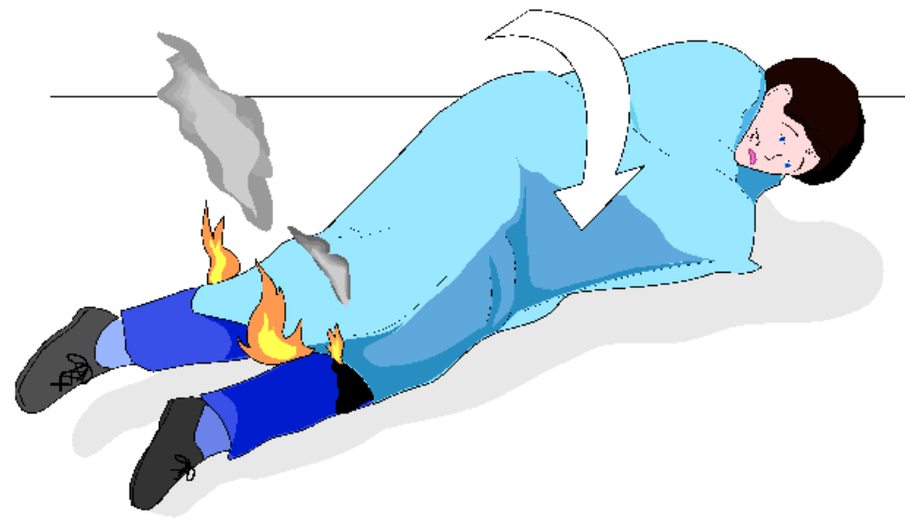
12.2.2020

END-OF-SEMESTER PLAN

* on Friday we'll talk about **artificial neural networks**

* on Monday we'll **review for the final**

  * next Tuesday's office hours will also be available for **final review**



RECAP: PCA

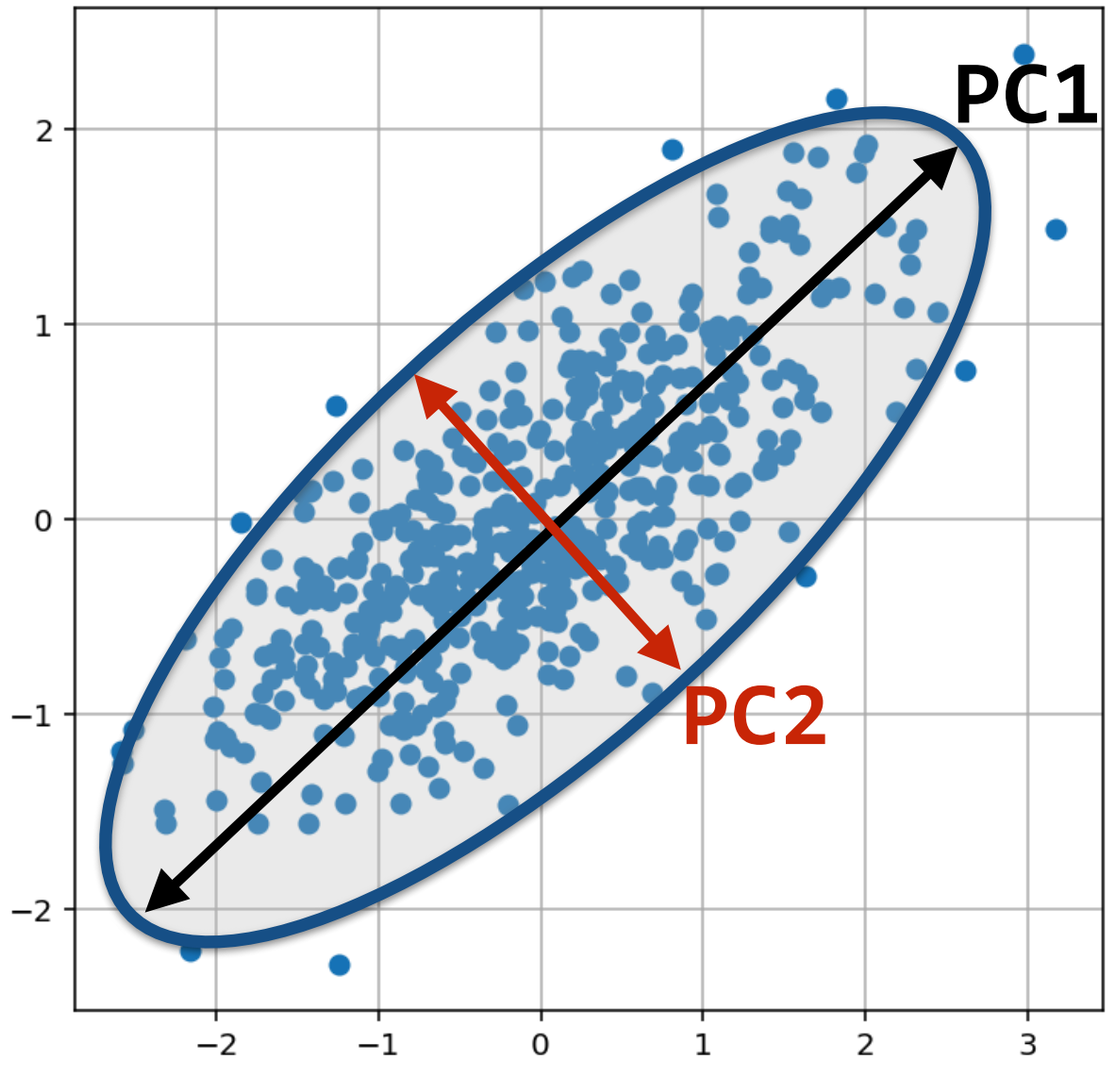
* **Principal Components Analysis** is an *unsupervised* method for finding structure in datasets

* (This is different from regression & classification, which are examples of *supervised learning*. They learn a function f(X)=y. Here we only have X!)

RECAP: PCA

* PCA “fits an ellipse” to a cloud of datapoints

* The axes of the ellipse are the “principal components”

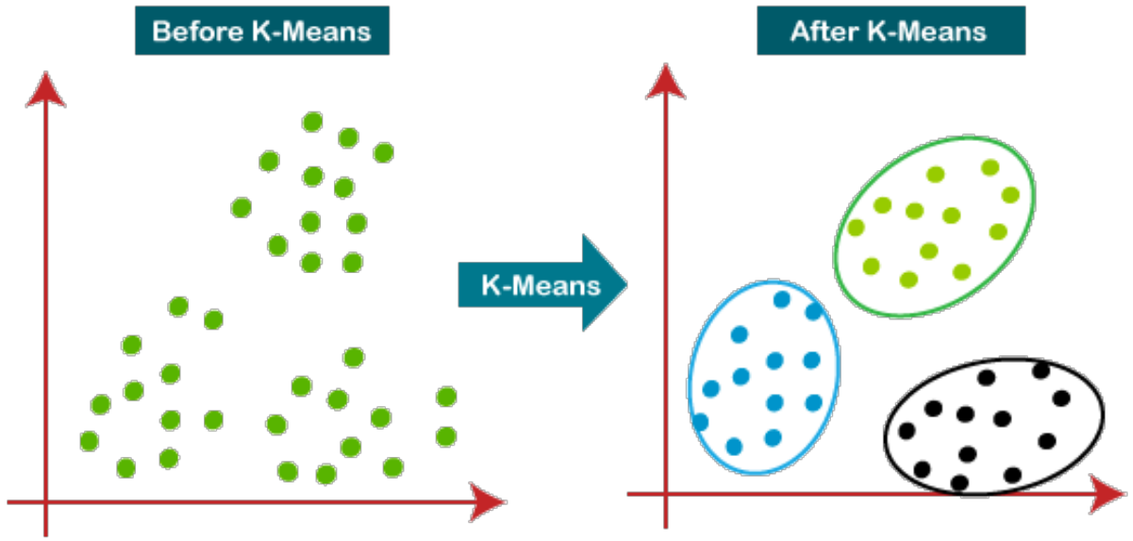


CLUSTERING

* Another commonly-used type of unsupervised learning is **clustering**

* **Clustering** involves assigning your data points to groups (aka clusters)

* Where points within a group are more similar than points in different groups



CLUSTERING

* Clustering is kind of like PCA (and other factor analysis methods), except that its output is **discrete** (i.e. 1, 2, 3) rather than continuous (i.e. -0.2, 0.7, 3.5)

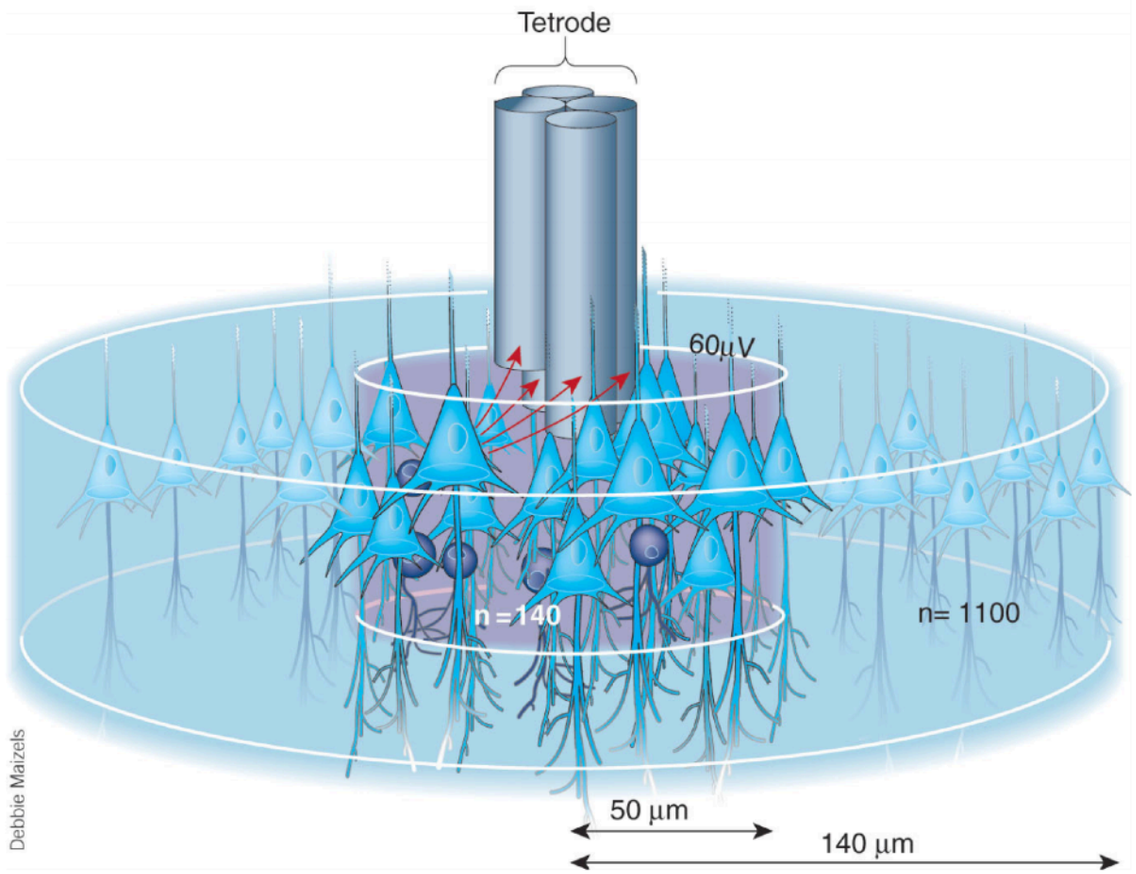
CLUSTERING

	Supervised (given X & y, learn X→y)	Unsupervised (given X, learn its structure)
Continuous (output is real numbers)	Regression	PCA
Discrete (output is discrete classes)	Classification	**Clustering**

APPLICATION: SPIKE SORTING

* Spikes from several neurons are recorded using a multi-electrode probe (e.g. a tetrode)

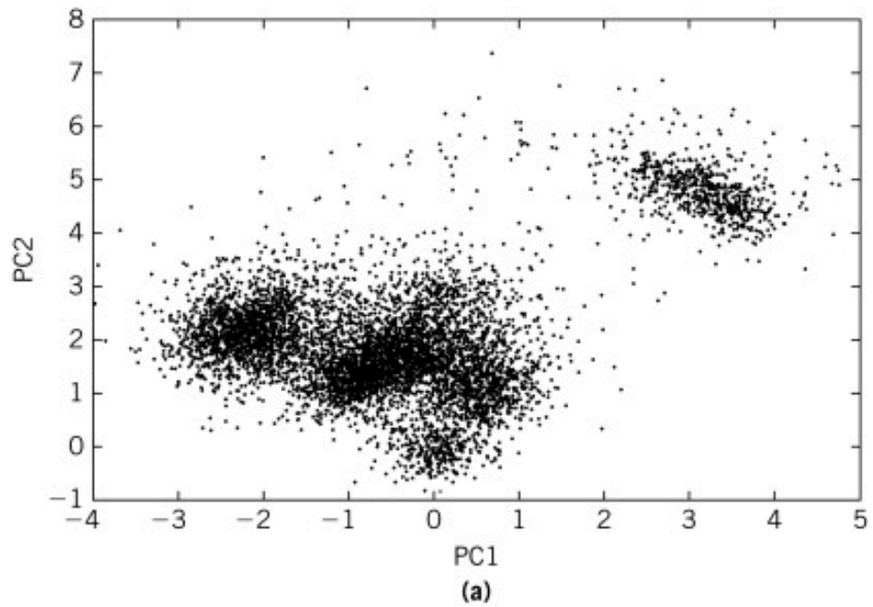
* How do we figure out which spikes came from the same neuron?

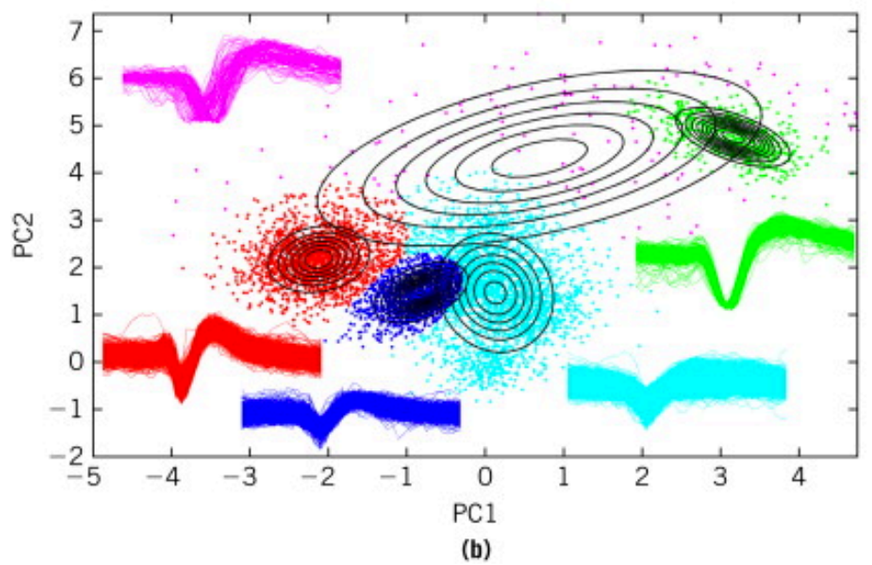


APPLICATION: SPIKE SORTING

* Step 1: Use PCA on the spike waveforms to represent the spikes in a lower-dimensional space

* Step 2: Use clustering to group the spikes together into putative single units





APPLICATION: FINDING CELL TYPES

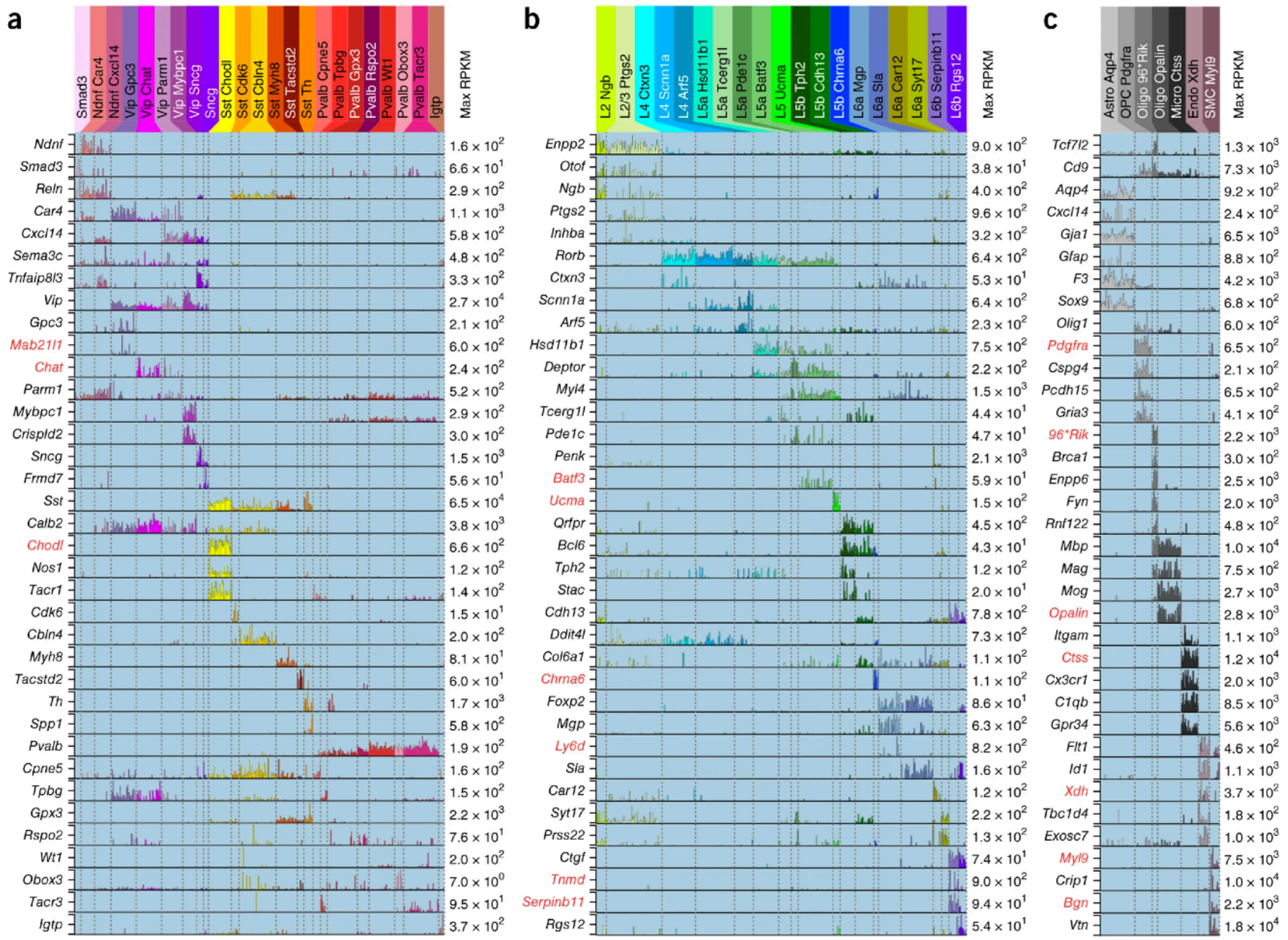
* How many different types of neurons are there? What does each type do? How can we distinguish them?

* The Allen Institute has been trying to answer this question using single-cell transcriptomics (aka RNA-seq)

  * For each tested neuron they measure how often each of thousands of genes are transcribed in that neuron

  * Then they use clustering to group the neurons into discrete types

APPLICATION: FINDING CELL TYPES



https://www.nature.com/articles/nn.4216

CLUSTERING

* How can we do clustering? How does it work?

* Let’s talk about one clustering method: **k-means**

  * k-means is pretty simple!

K-MEANS CLUSTERING

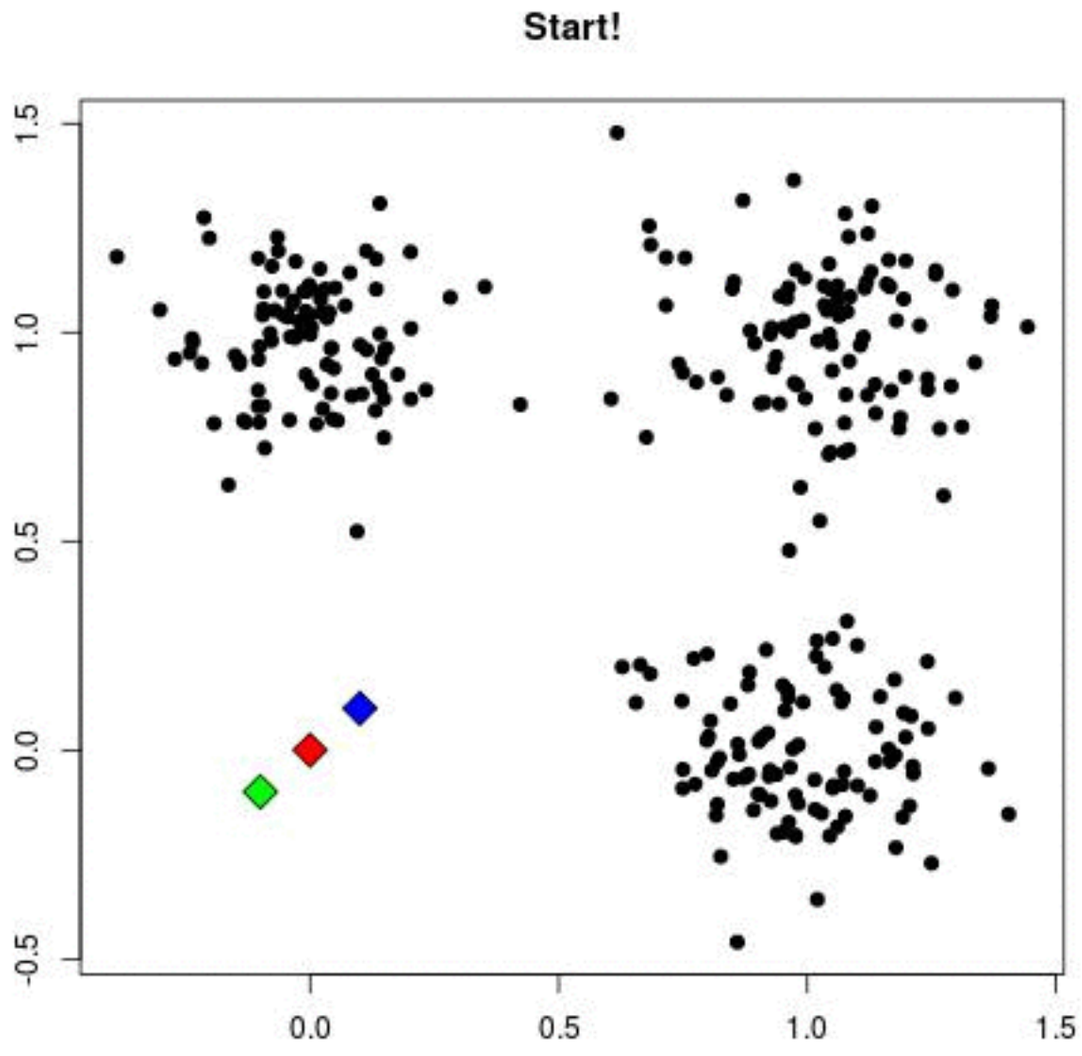
1. Randomly choose k datapoints to be the initial cluster “centroids”

2. Compute the distance from each datapoint to each centroid, then *assign* each datapoint to the closest one

3. Move each centroid to the mean location of its assigned datapoints

4. Go back to 2

K-MEANS CLUSTERING



https://mubaris.com/posts/kmeans-clustering/

K-MEANS CLUSTERING

* Note: this is not guaranteed to work!

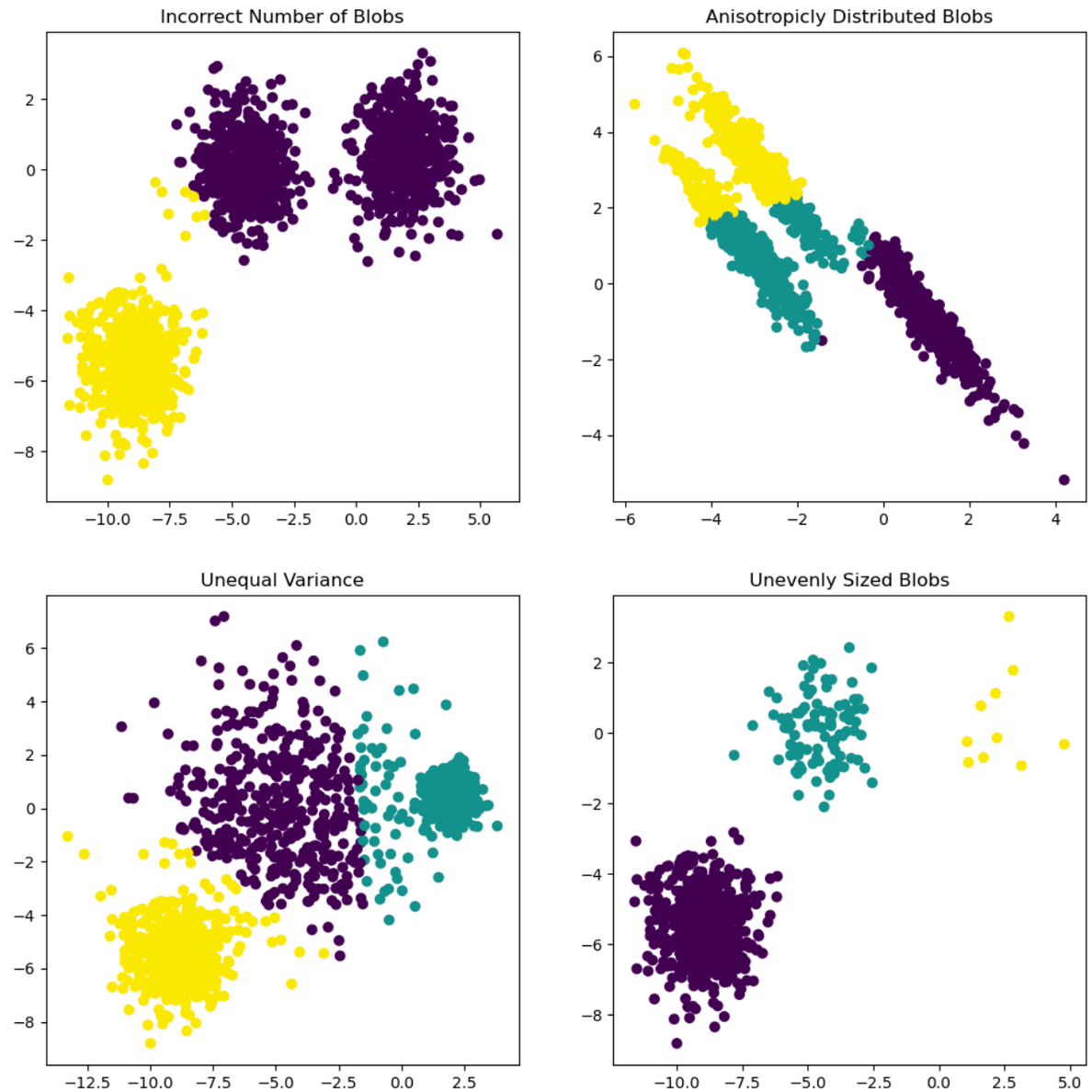
* There are lots of ways for k-means to mess up, and it’s very sensitive to initialization

* So if you want to use it, you should use clever implementations like **sklearn.cluster.KMeans**

K-MEANS CLUSTERING

* k-means also makes some assumptions:
  * you chose the correct number, k
  * all clusters are “round” (isotropic)
  * all clusters have equal variance

* If these assumptions are false, then it can give bad results



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html

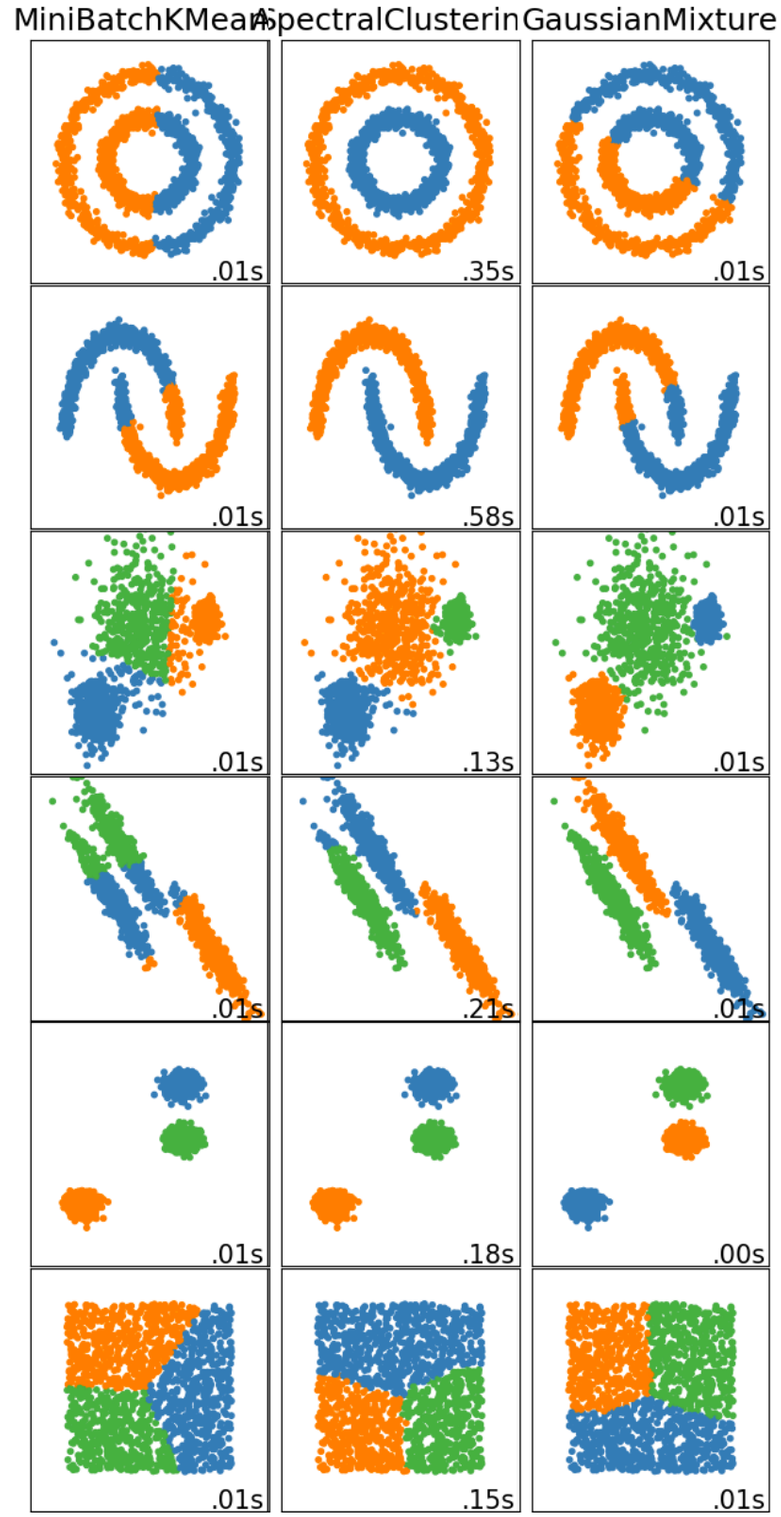
K-MEANS CLUSTERING

* but there are many other clustering methods that can solve these problems!

* Gaussian mixture models (GMMs)

* Spectral clustering

* Hierarchical clustering



THANK YOU!