# Plan

- ssh + running on labs computers
- k-means clustering
  - ↳ notebook sklearn

Logistical stuff

grading details

data available on labs

## using 'screen'

- keeps programs running in background (nohup)
- detach

  ```
  C-a d
  ```

- reattach

  ```
  screen -D -R
  ```

- kill

  ```
  C-a k
  ```

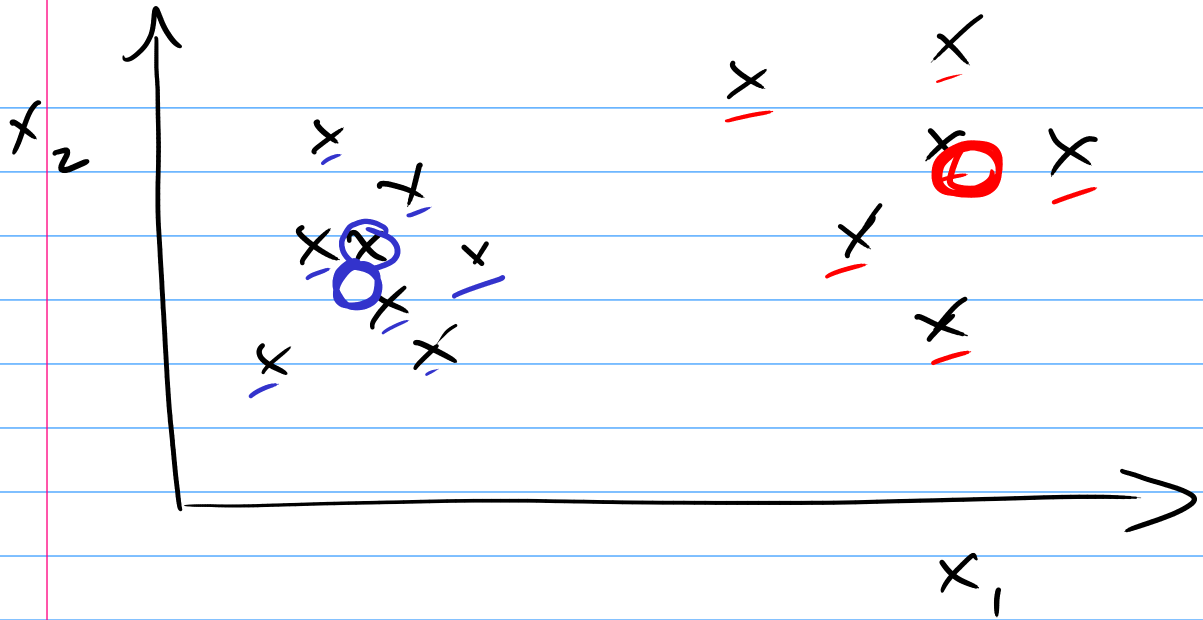
# K-means (not k-NN)

**Clustering** method: have a dataset

$$X, \{\vec{x}_i \in \mathbb{R}^d\}_{i=1}^n$$

want to divide the vectors into sets, produce a label $y_i \in \{1, \dots, k\}$ one of k diff. clusters.

Diff from classification: no y labels to start with



$(d=2)$
$k=2$

○ = center 1
○ = center 2

Initialize w/ K different centers

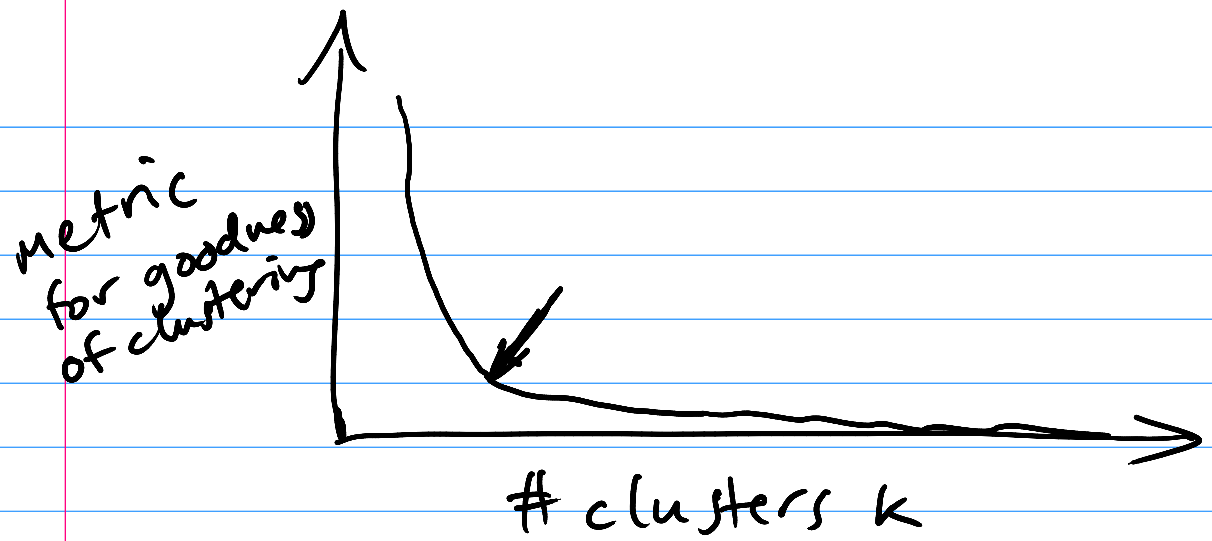
Repeat until convergence:

1) Label pts w/ by nearest center

2) Calculate the mean of all points in a given label and use that as new center

Details: initialization

Similar to Gaussian mixture model ← better in general



$$\text{metric} = \sum_{i=1}^{n} (\text{distance of pt } i \text{ to cluster center})$$

often called the "distortion"

= error in distances from replacing each data pt with cluster center

# Project :

use linear model as baseline

Ridge , RidgeCV

Logistic , LogisticCV

SVM

np.loadtxt