HOLT SKINNER

K MEANS, FUZZY C-MEANS, POSSIBILISTIC C-MEANS

CLUSTERING

WHAT IS CLUSTERING?

- Unsupervised Learning
- Grouping a set of objects such that objects in the same group (called a **cluster**) are more similar to each other than to those in other groups

HOW DOES IT WORK?

- Simplest Algorithm: K-Means

1. Pick K Random Initial Cluster Centers (You pick K)
2. Find out which cluster each point belongs to (Distance)
3. Update Cluster Centers (Take the Mean Value of all points in a cluster)
4. Go back to Step 2 & Repeat Until convergence

IMPROVEMENT – FUZZY C MEANS

- What if each point could partially belong to multiple clusters???
- Same Basic Algorithm

$$u_{ik} = \frac{\left(1/d(x_k, v_i)\right)^{2/(m-1)}}{\sum_{j=1}^{C} \left(1/d(x_k, v_j)\right)^{2/(m-1)}} \qquad v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}$$

SOURCE: KELLER - FUNDAMENTALS OF COMPUTATIONAL INTELLIGENCE

FURTHER IMPROVEMENT – POSSIBILISTIC C MEANS

- What if your data is noisy or full of outliers?
- Some points shouldn't belong in any of the clusters.
- Algorithm created by MU Professor Jim Keller

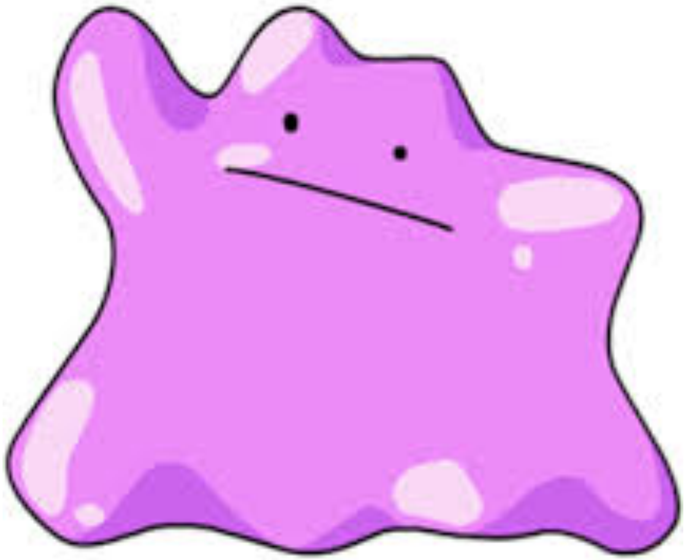
$$u_{ik} = \frac{1}{1 + \left(d^2(x_k, v_i)/\eta_i\right)^{1/(m-1)}}$$

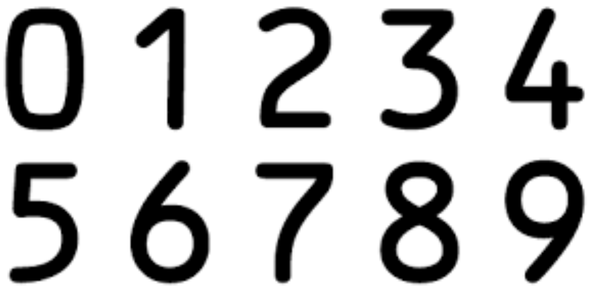
SOURCE: KELLER - FUNDAMENTALS OF COMPUTATIONAL INTELLIGENCE

PROJECT IDEA

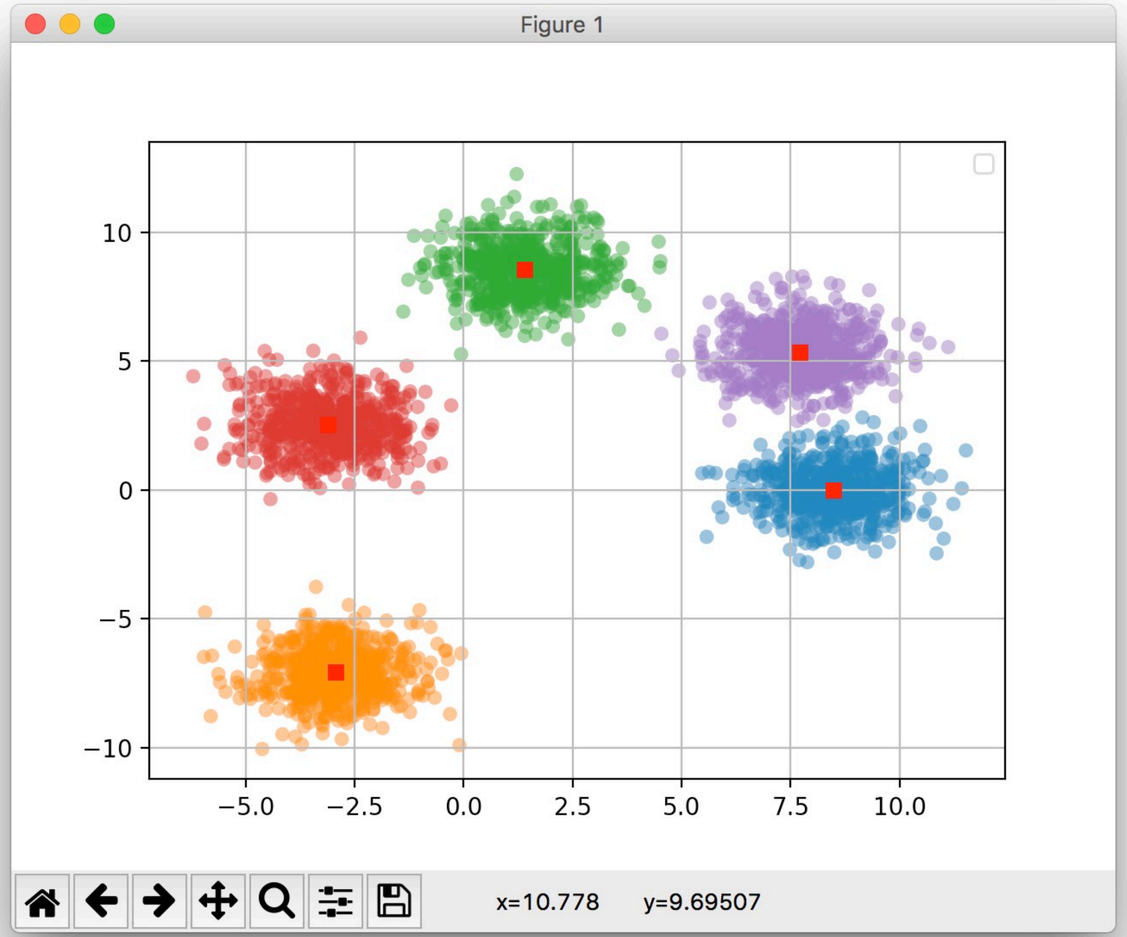
- The Possibilistic C Means Algorithm provides a great deal of promise, but there's no open source library for it.
- Capstone project uses a large data set, but existing library for FCM is too slow.
- Solution: Reinvent the Wheel!
  - Implement K Means, Fuzzy C Means and Possibilistic C Means
  - Python, Numpy, & MatPlotLib
- Verify clusters using distances from cluster centers compared to actual means of classes (For test data)

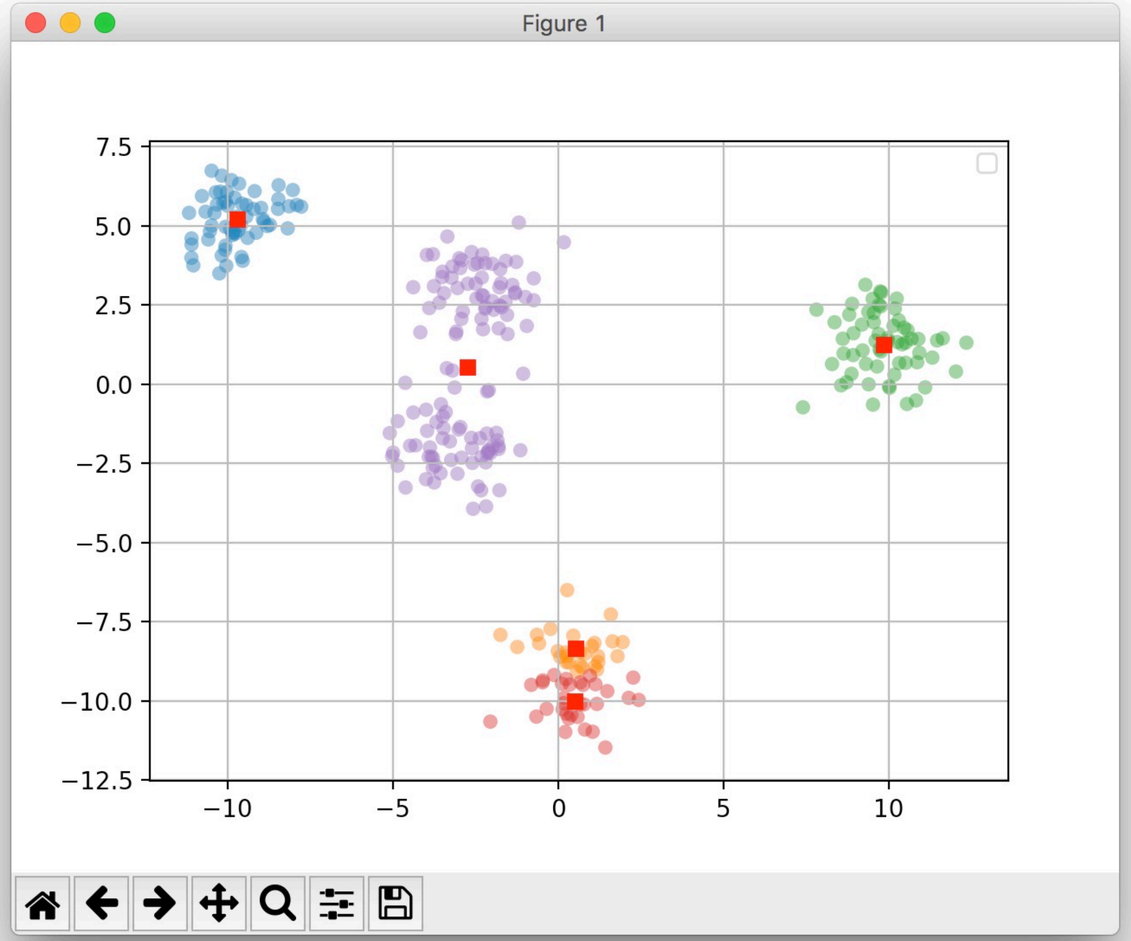
TESTING DATA SETS



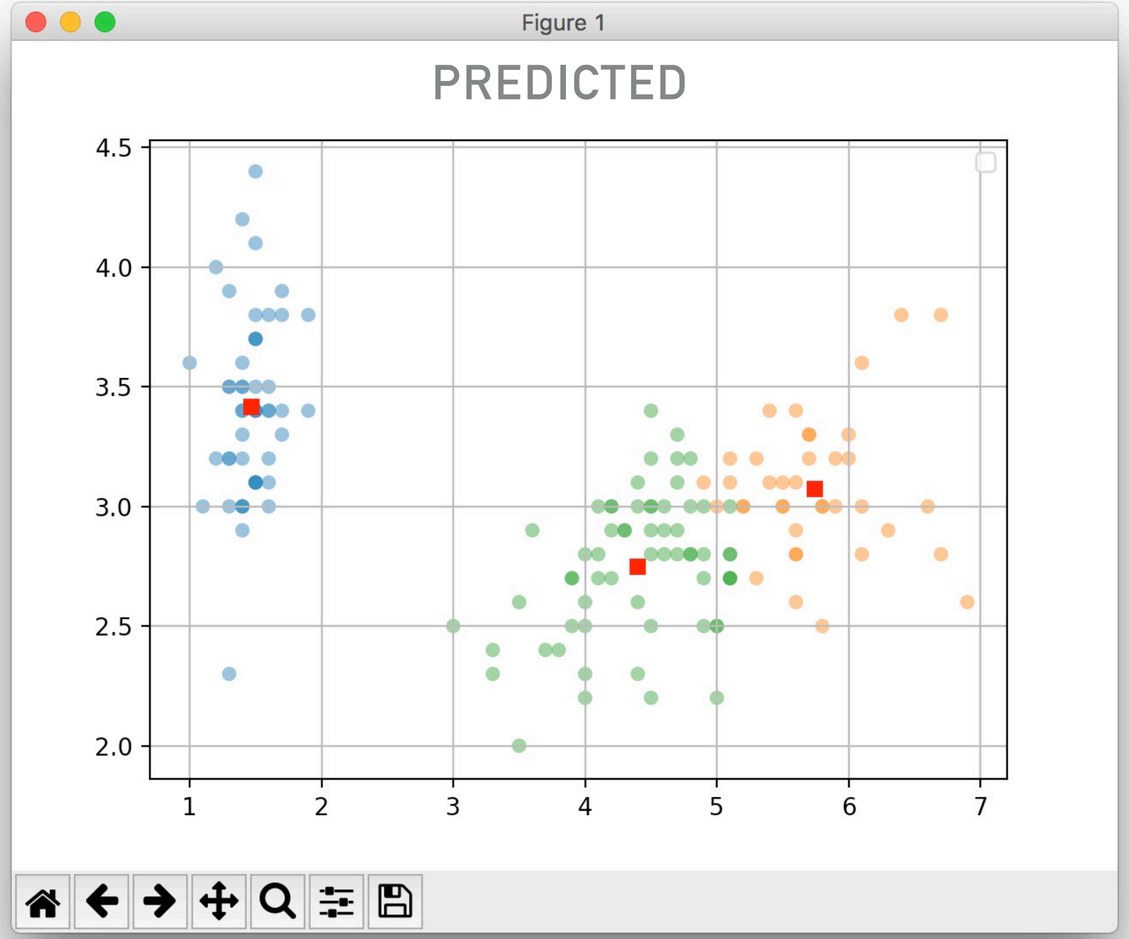


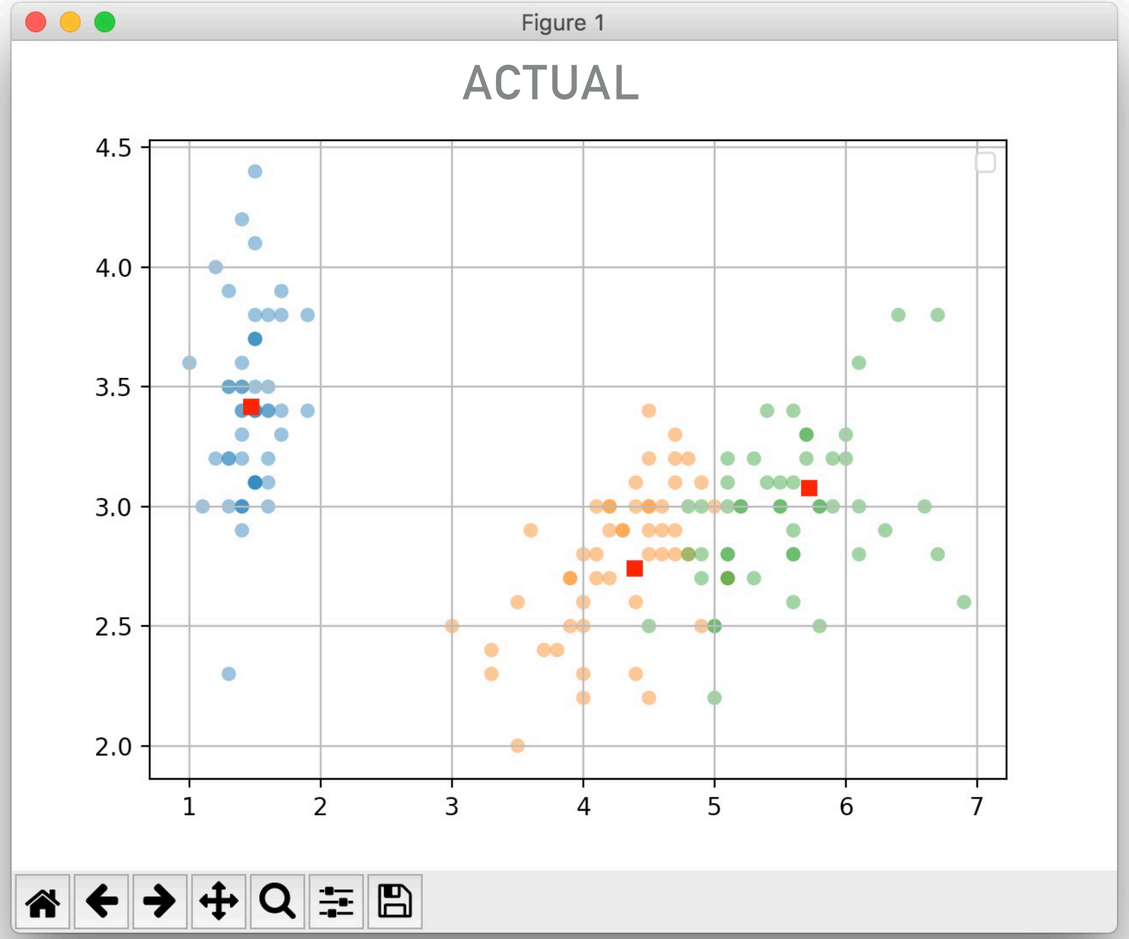
RESULTS – K MEANS (GAUSSIAN BLOBS)



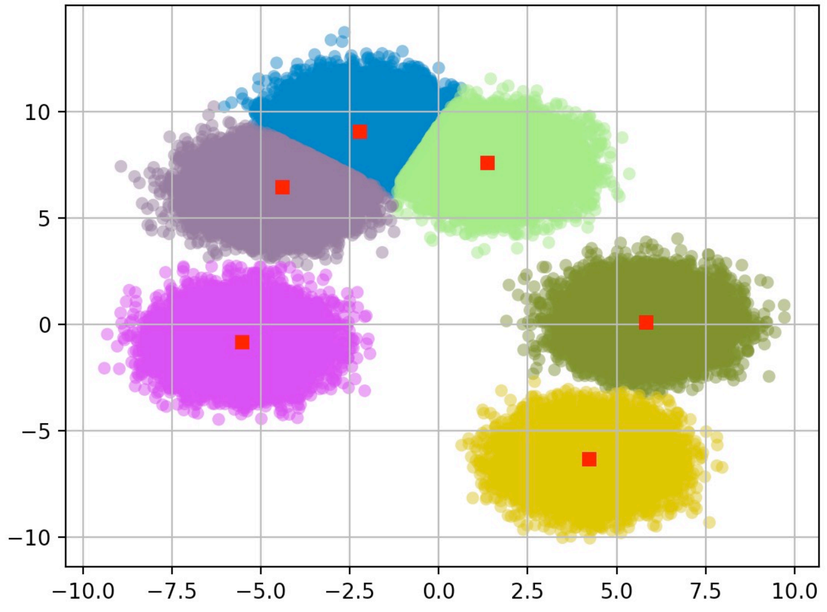


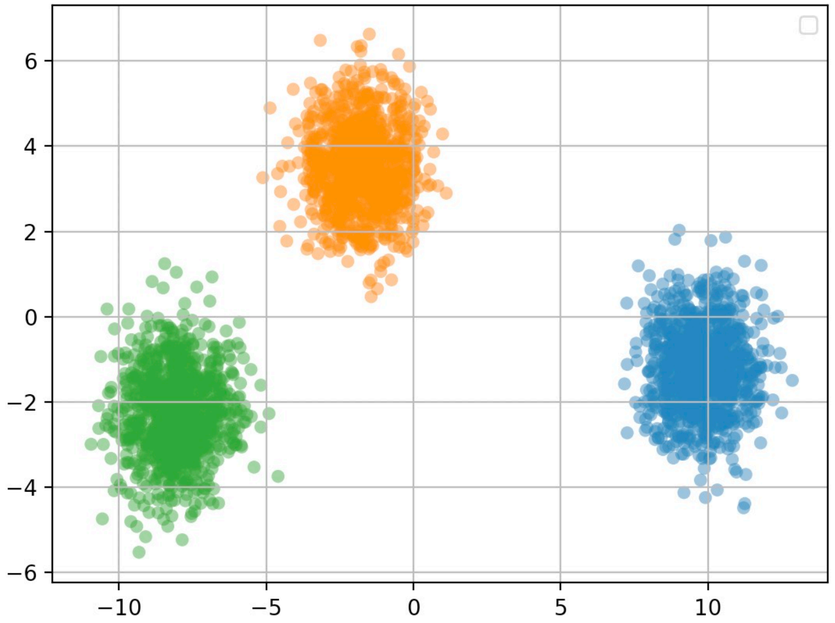
RESULTS – K MEANS (IRIS DATA)

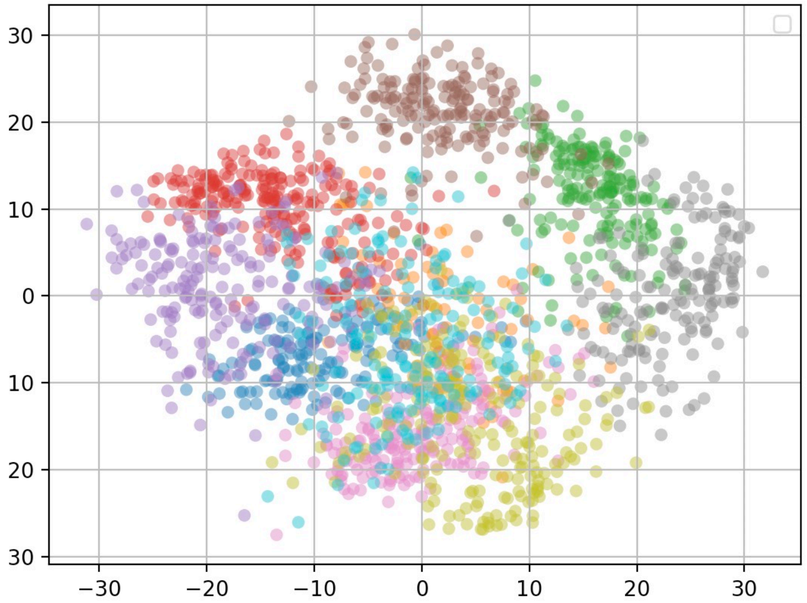




RESULTS – FUZZY C MEANS (GAUSSIAN BLOBS & DIGITS)



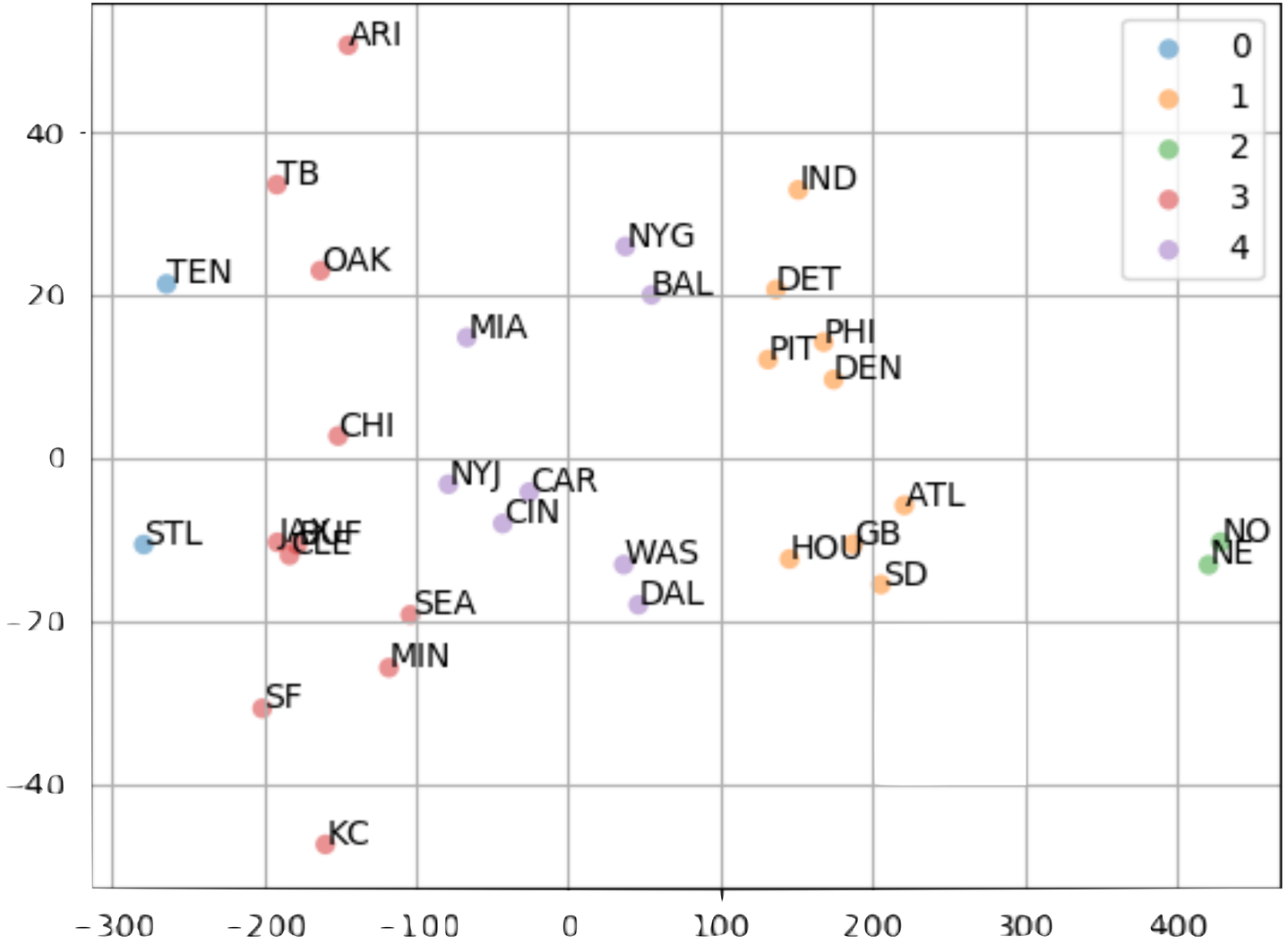




CAPSTONE APPLICATION – NFL TEAMS

- What if you could classify NFL Teams like Pokemon?
- Data Collected from Kaggle & Cleaned by Alex Hurt
- SciKit Fuzzy didn't work…

FUZZY C-MEANS & POSSIBILISTIC C MEANS (NFL DATA)



PROBLEMS

- Numpy Learning Curves...
  - Solution: Stack Overflow
- Initialization of Clusters makes a HUGE difference
- Fuzzifier (1.2 is best)
- Modularity of Code
  - Solution: Functional Programming!