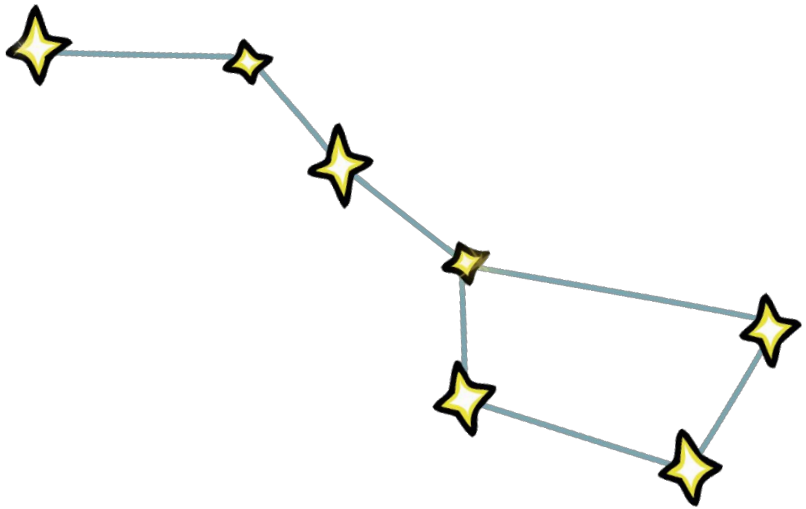
# Above and Beyond: Clustering Stars

Erica Lee, Jaclyn Nguyen, Kimberly Siegler

# Motivation

Have you ever wondered how constellations are defined?



# Approaches

Supervised: Two-Step Process

- Logistic Regression
  - Binary
  - Multi-Class
- K-Nearest Neighbors

Unsupervised

- k-means++
- Hierarchical clustering
- Affinity Propagation
- Mean Shift

# Data

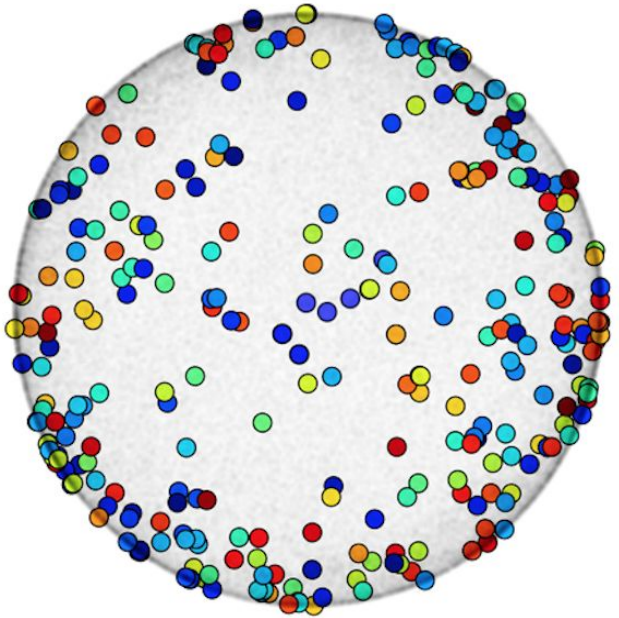
- Tycho-2
  - 2.5 million stars
  - 37 features
    - Galactic Latitude/Longitude
    - Right Ascension/Declination
    - Luminosity
- AstroNexus
  - 120,000 stars
  - Ground truth constellation labels

Python Packages: numpy, pandas, astropy, matplotlib

# Supervised Approach

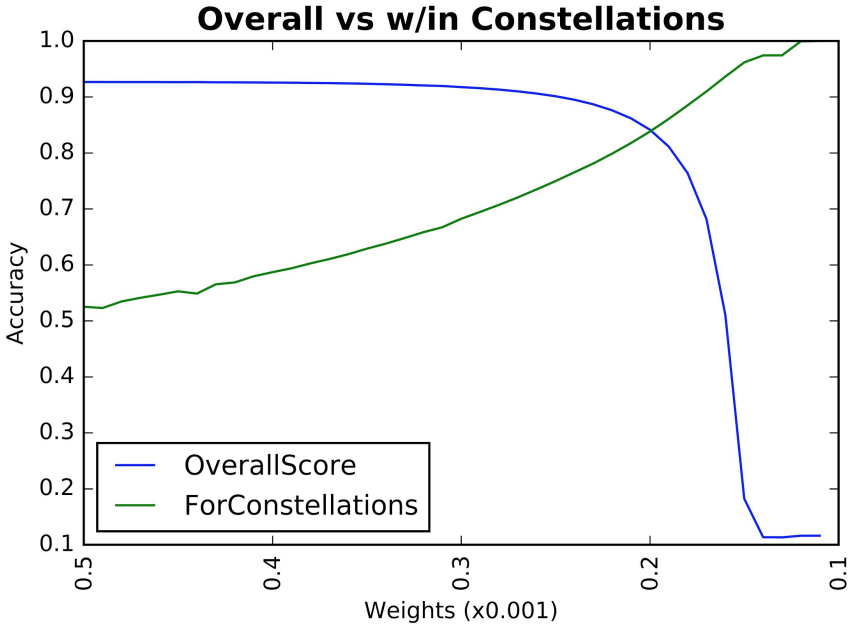
- Two Step Classification Process
  - Logistic Regression {0,1}
  - K-NN {Multi-class}

- Weights
  - Class-Weight for Unbalanced Data
  - Different Scale
    - x, y, z, Vmag

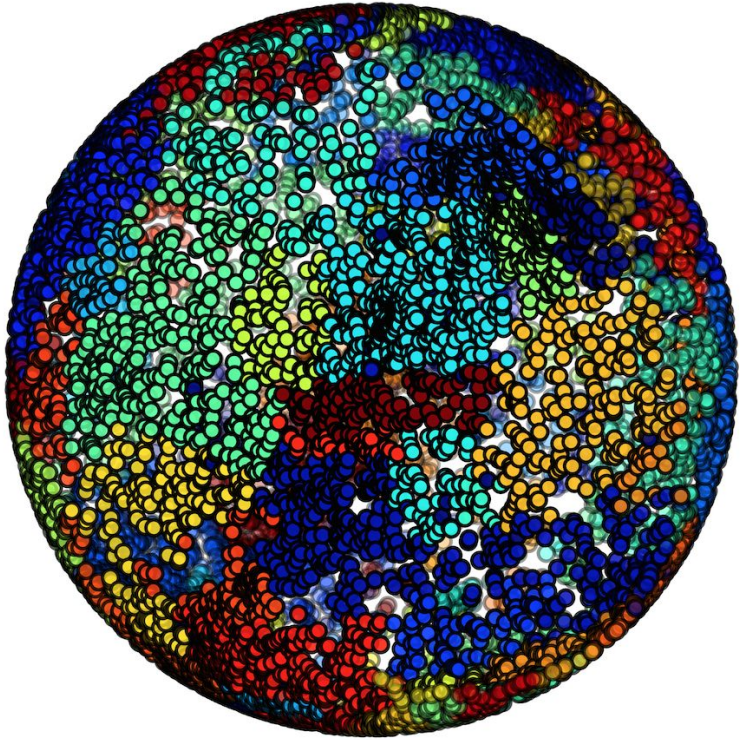


# Results - Supervised

- An ensemble method
  - Logistic Regression {0, 1}
    - Weighted-class
    - Accuracy : 84%
  - K-Nearest Neighbors
    - Weighted variable (V-mag)
    - k = 1 performed the best



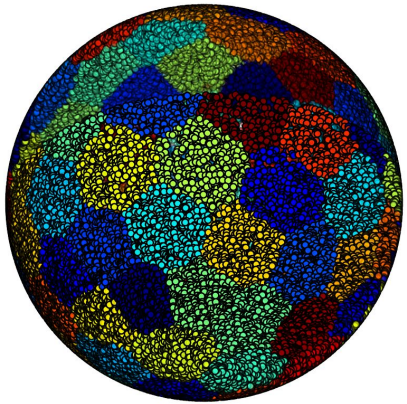
# Results - Unsupervised



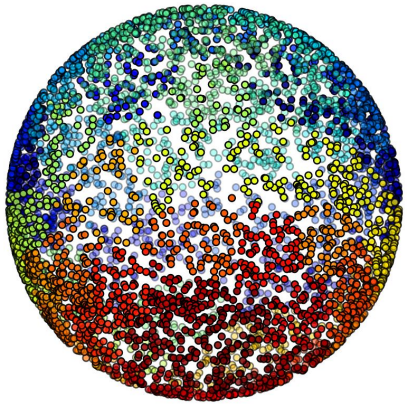
Current Constellation Membership

K-means++ Constellation Membership

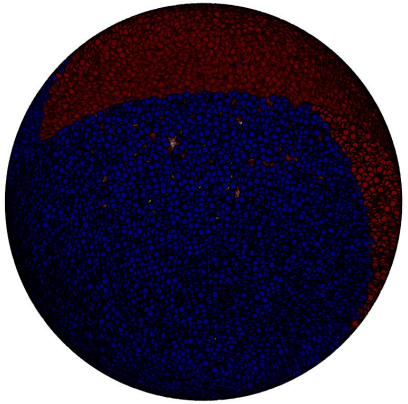
# Results - Unsupervised



K-means++



Affinity Propagation



Mean Shift

# Conclusion

- Best compact and separated clusters
  - k-means++
- Best clustered stars compared to current constellation labels
  - Affinity Propagation
- Best classifications
  - Ensemble method of weighted-class logistic regression + weighted -Knn