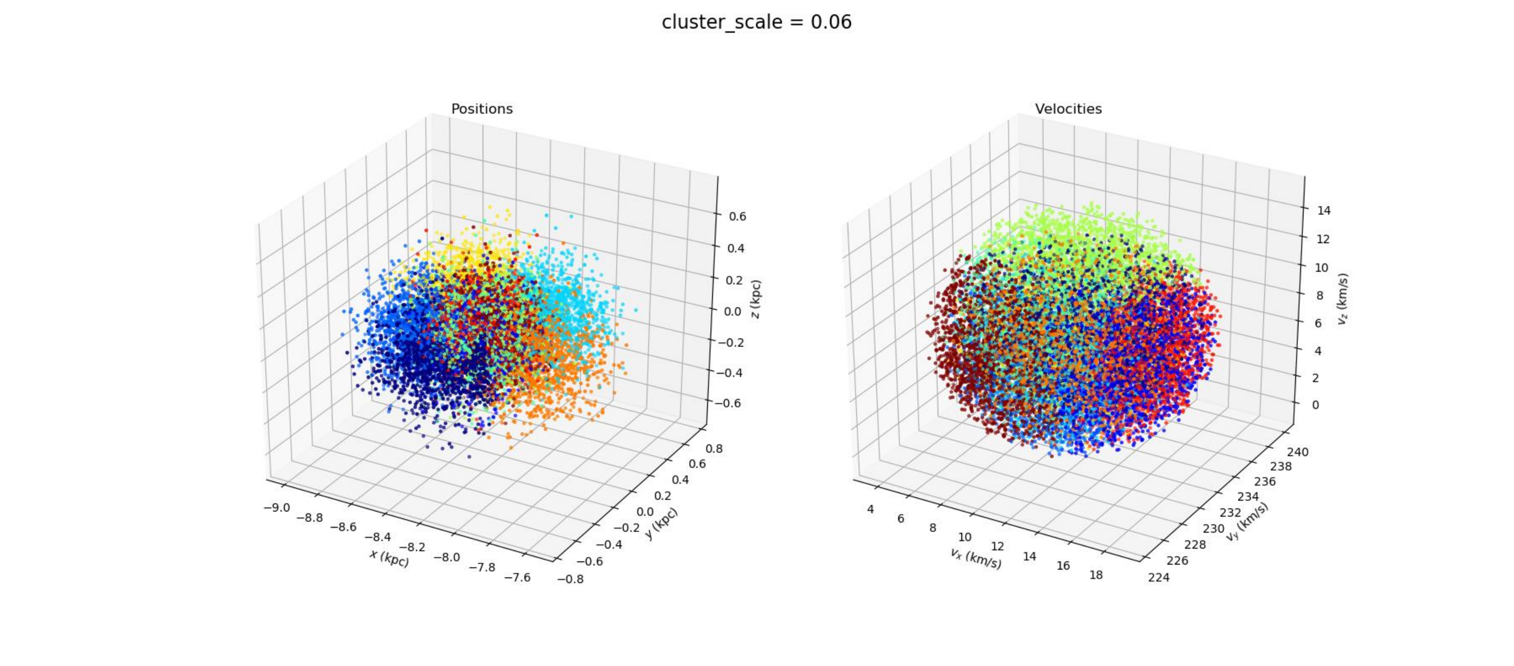
KMeans Clustering in 6 Dimensions to sample phase space

By: Michael Poon, Mathew Bub

June 2018

‘clustering technique to reduce 6D dataset’



By: Michael Poon, Mathew Bub

June 2018

Overview: Gaia DR2 RV (6D) -> main program -> 3D or 4D subspace?

Due to constraints from: Conserved (1) Energy, (2) Ang. Momentum, (3?) *Mystery*

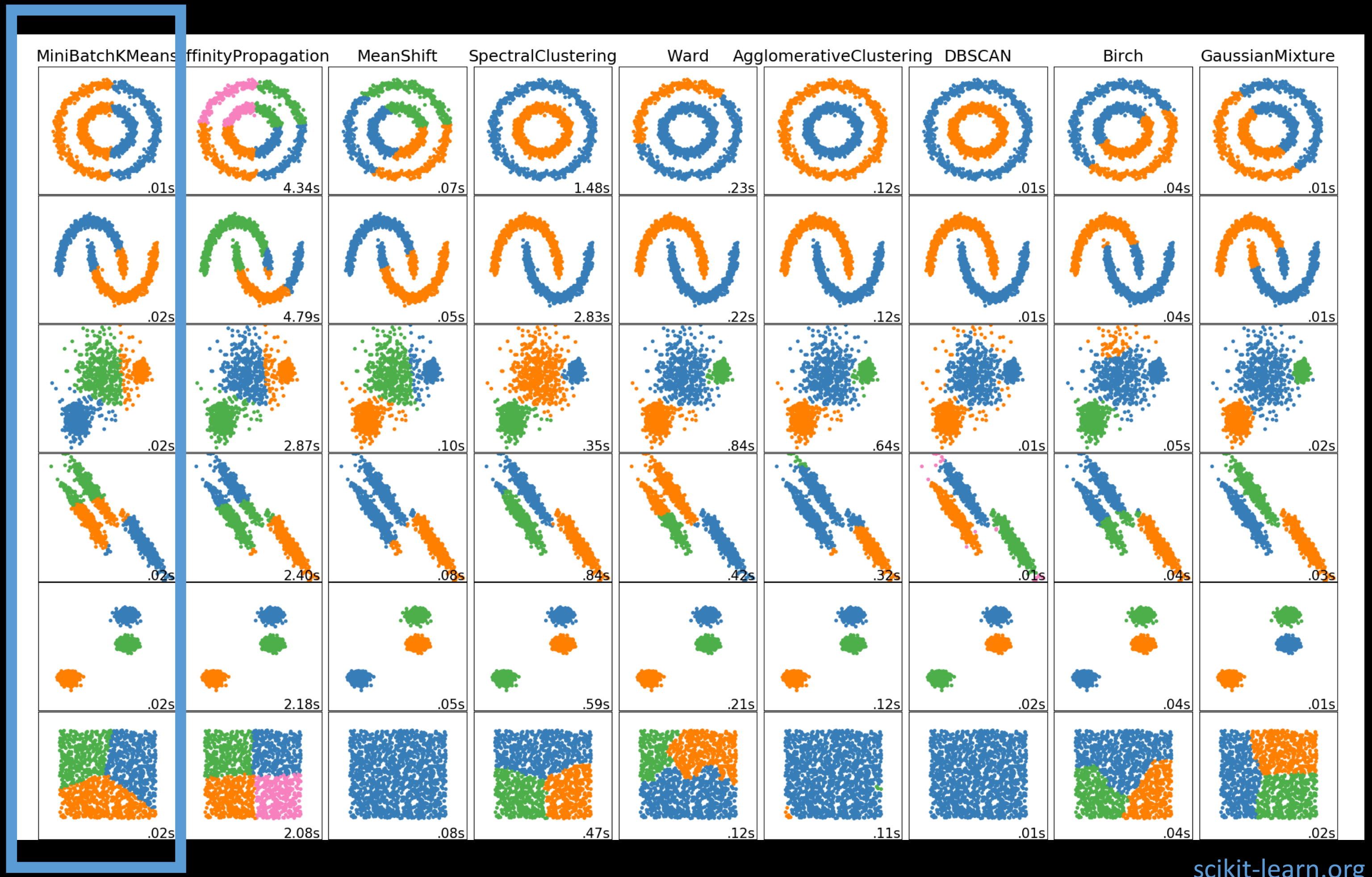
Problem: main program bottleneck, reduce dataset to a reasonable sample

GOAL:

Want a reduction: 7,224,631 Stars -> quality cuts -> KMeans Clustering -> 100,000? Cluster Centers

What is KMeans Clustering?

-unsupervised (no “correct” solution) machine learning technique



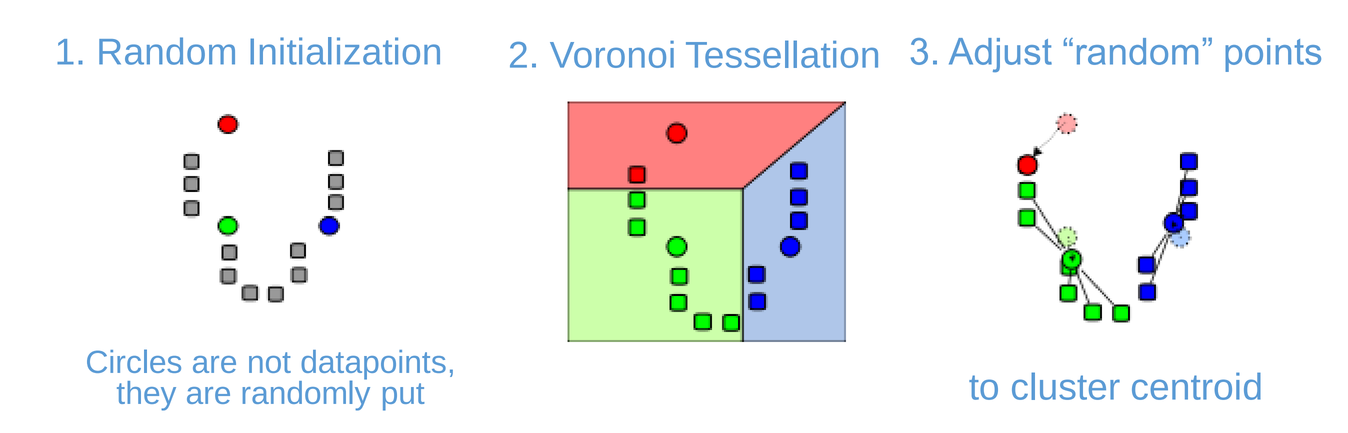
scikit-learn.org

Clusters how we want it to, and runs (relatively) fast:
Linear in Big O: O(kN), k - #iterations, N - #datapoints

How Does KMeans Work?

Voronoi Tessellation:

KMeans Clustering:



1. Random Initialization

2. Voronoi Tessellation

3. Adjust “random” points

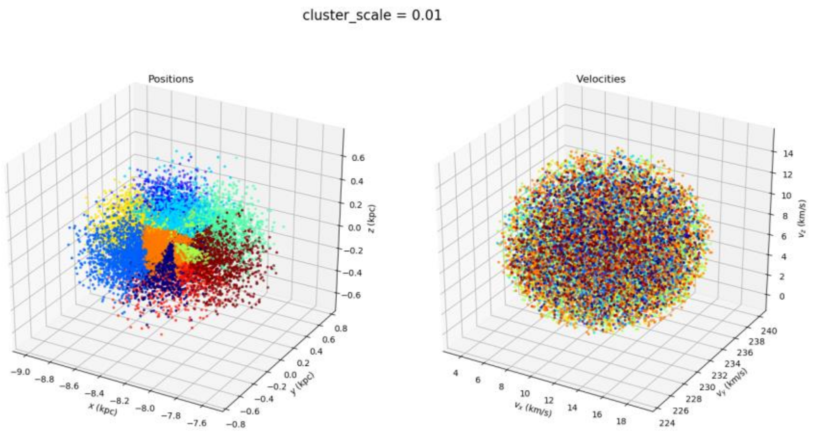
Circles are not datapoints, they are randomly put

to cluster centroid

wikipedia.org/wiki/K-means_clustering

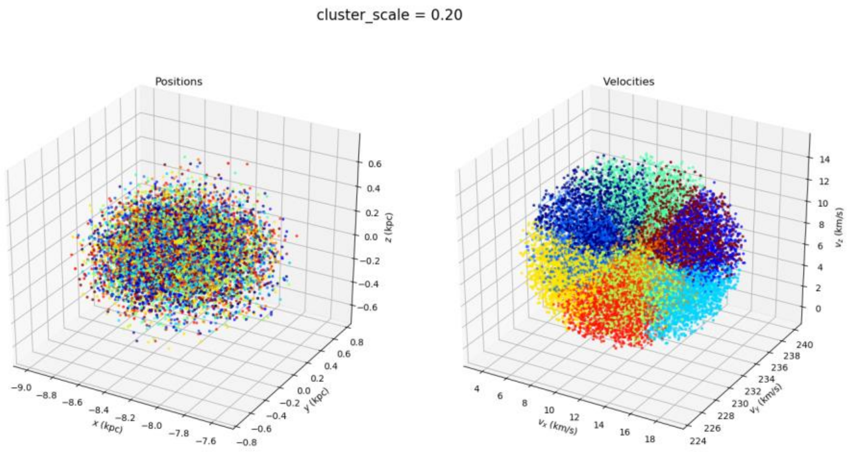
Preliminary Results:

#stars: 15,000


cluster_scale = 0.01

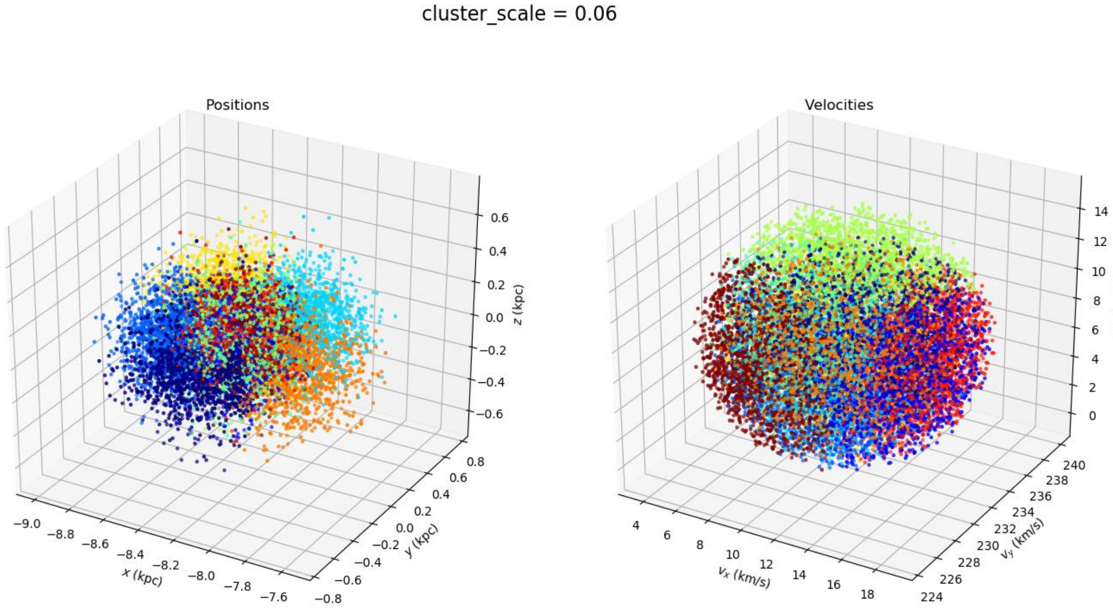
Positions

Velocities


cluster_scale = 0.20

Positions

Velocities


cluster_scale = 0.06

Positions

Velocities

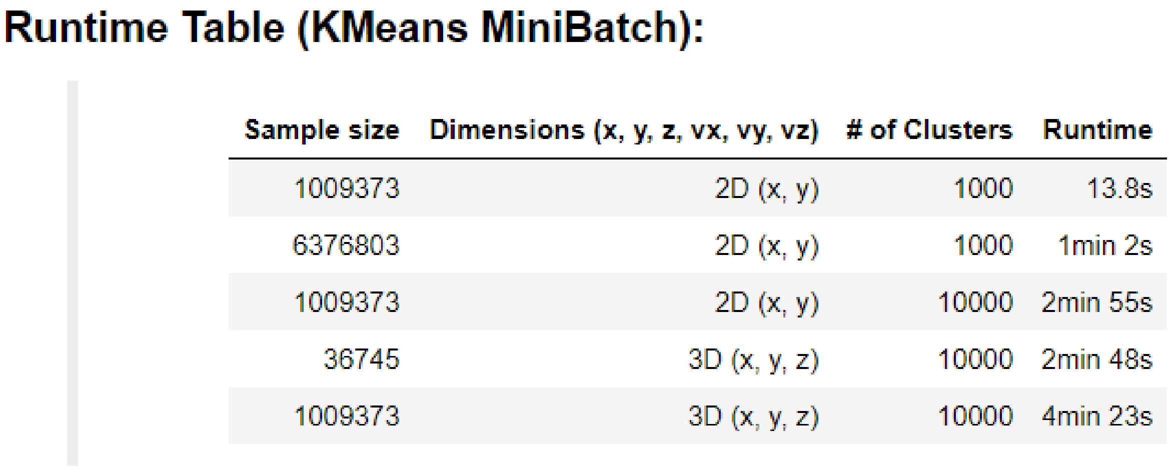
Discuss:

Reason for cluster_scale (Normalization)

Consequences to cluster_scale

Alternatives: Standard Dev. / Interq. Range

Next Steps:



Runtime Table (KMeans MiniBatch):

Sample size	Dimensions (x, y, z, vx, vy, vz)	# of Clusters	Runtime
1009373	2D (x, y)	1000	13.8s
6376803	2D (x, y)	1000	1min 2s
1009373	2D (x, y)	10000	2min 55s
36745	3D (x, y, z)	10000	2min 48s
1009373	3D (x, y, z)	10000	4min 23s

What sample size should we start with? Or use all?

How many clusters should we make?

KMeans vs. KMeans MiniBatch?