Week 7 Video 3

Advanced Clustering Algorithms

Today…

- Multiple advanced algorithms for clustering

Gaussian Mixture Models

- Often called EM-based clustering

- Kind of a misnomer in my opinion
  - What distinguishes this algorithm is the kind of clusters it finds
  - Other patterns can be fit using the Expectation Maximization algorithm

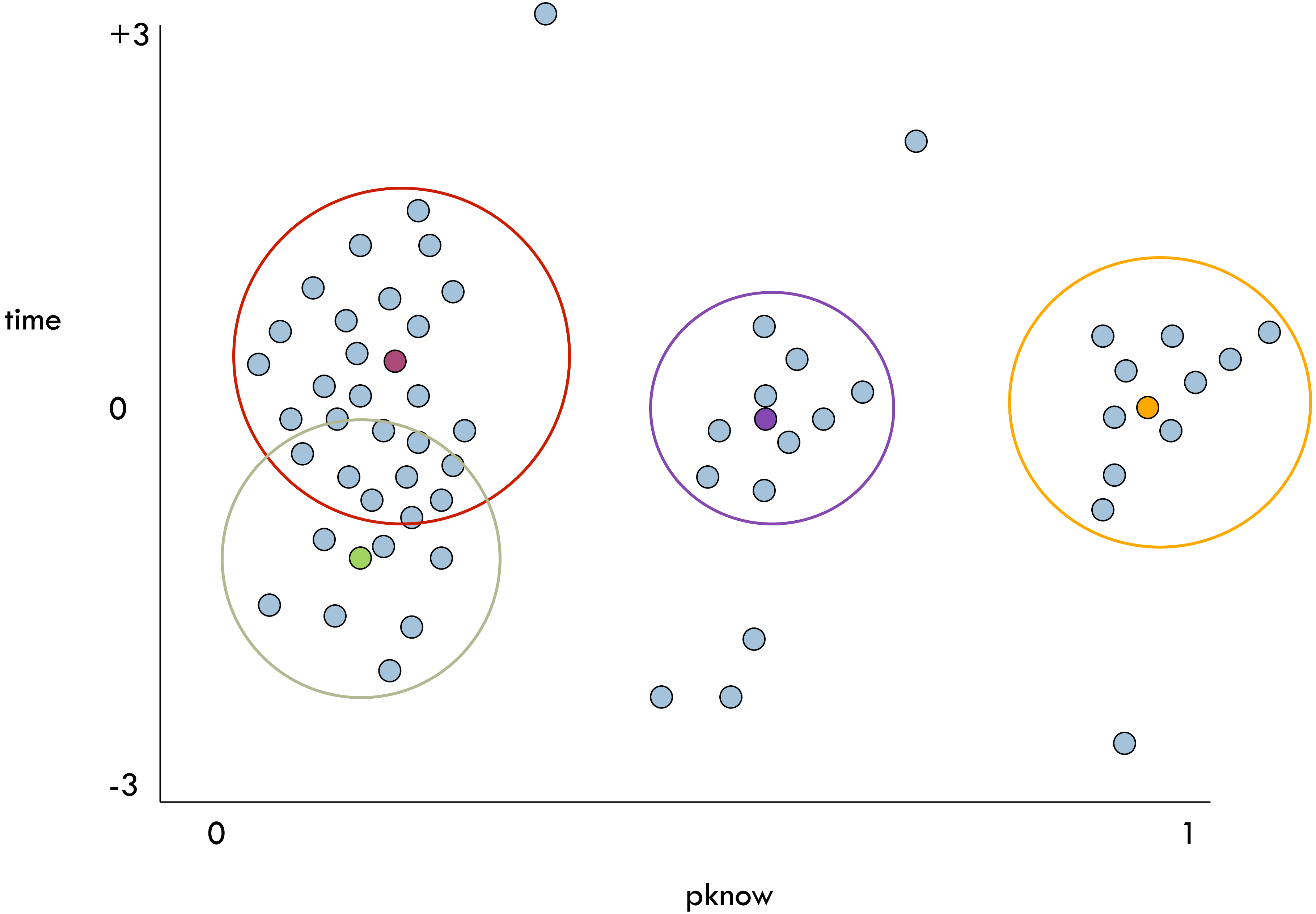
- I'll use the terminology Andrew Moore uses, but note that it's called EM in RapidMiner and most other tools

Gaussian Mixture Models

- A centroid ***and*** a radius
- Fit with the same approach as k-means
  (some subtleties on process for selecting radius)

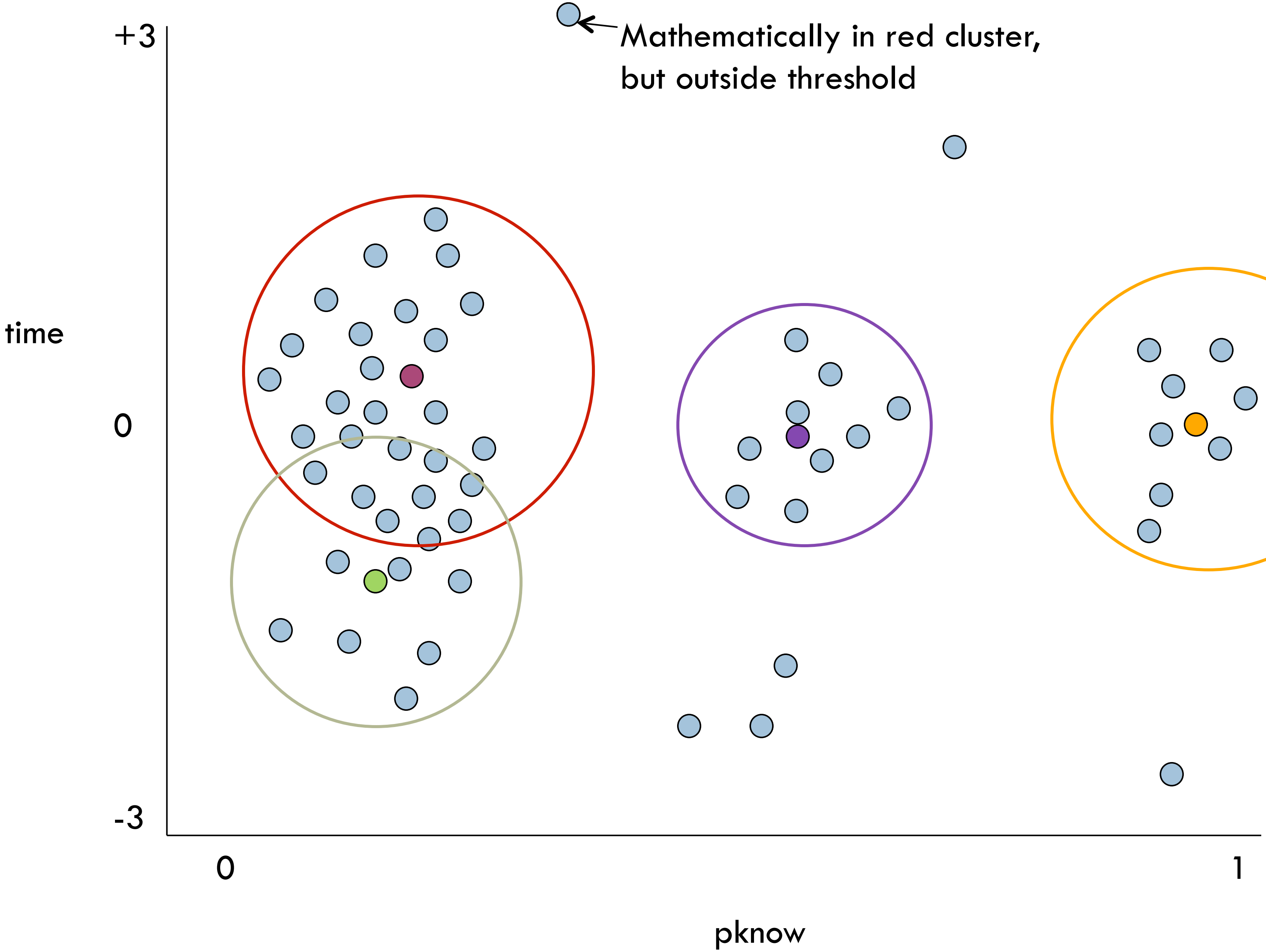
Gaussian Mixture Models

- Can do fun things like
  - Overlapping clusters
  - Explicitly treating points as outliers


+3
time
0
-3
0
1
pknow

Nifty Subtlety

- GMM still assigns every point to a cluster, but has a threshold on what's really considered "in the cluster"
- Used during model calculation



Mathematically in red cluster, but outside threshold

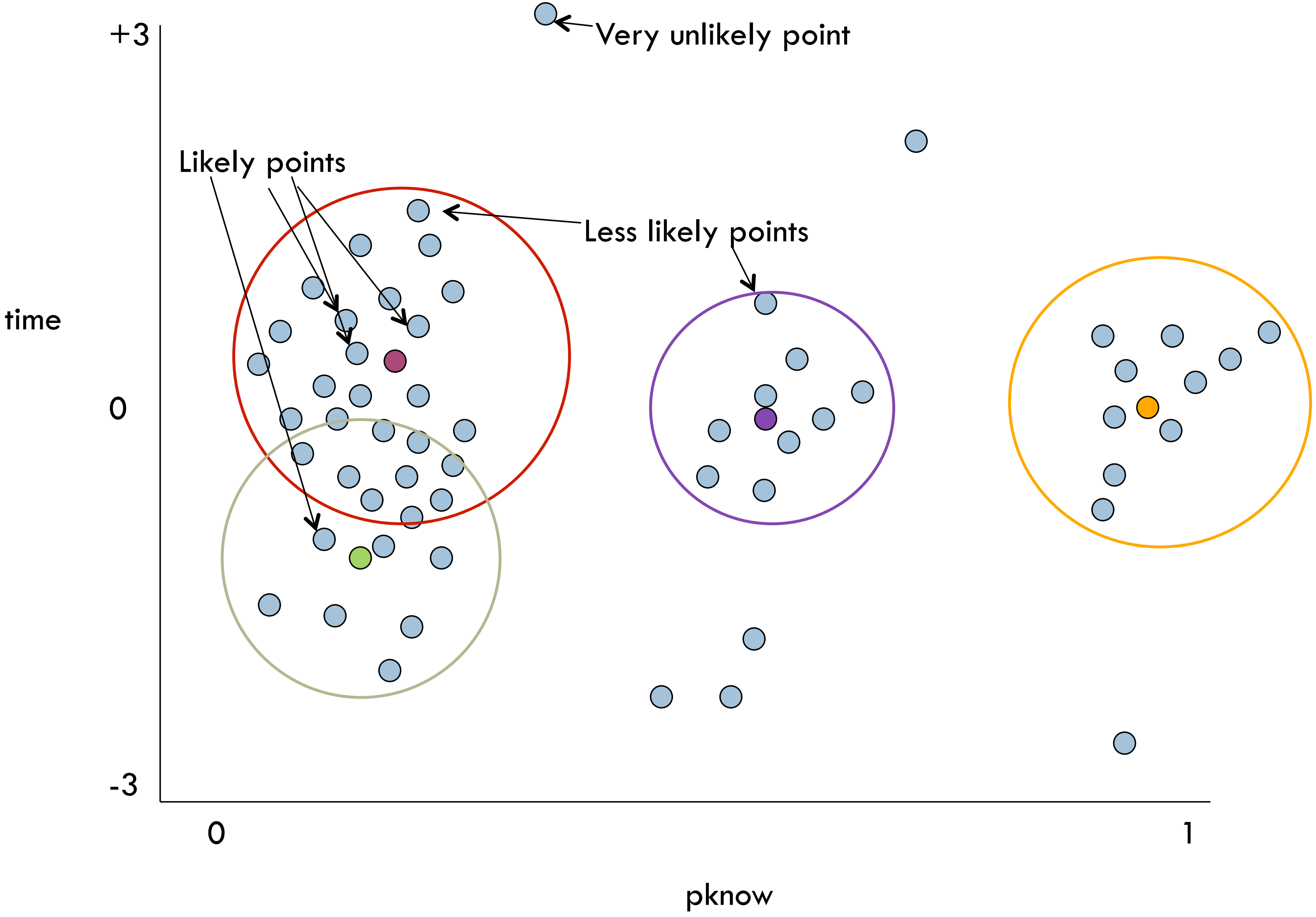
Assessment

- Can assess with same approaches as before
  - Distortion
  - BiC
- Plus

Likelihood

- (more commonly, log likelihood)
- The probability of the data occurring, given the model
- Assesses each point's probability, given the set of clusters, adds it all together

For instance…

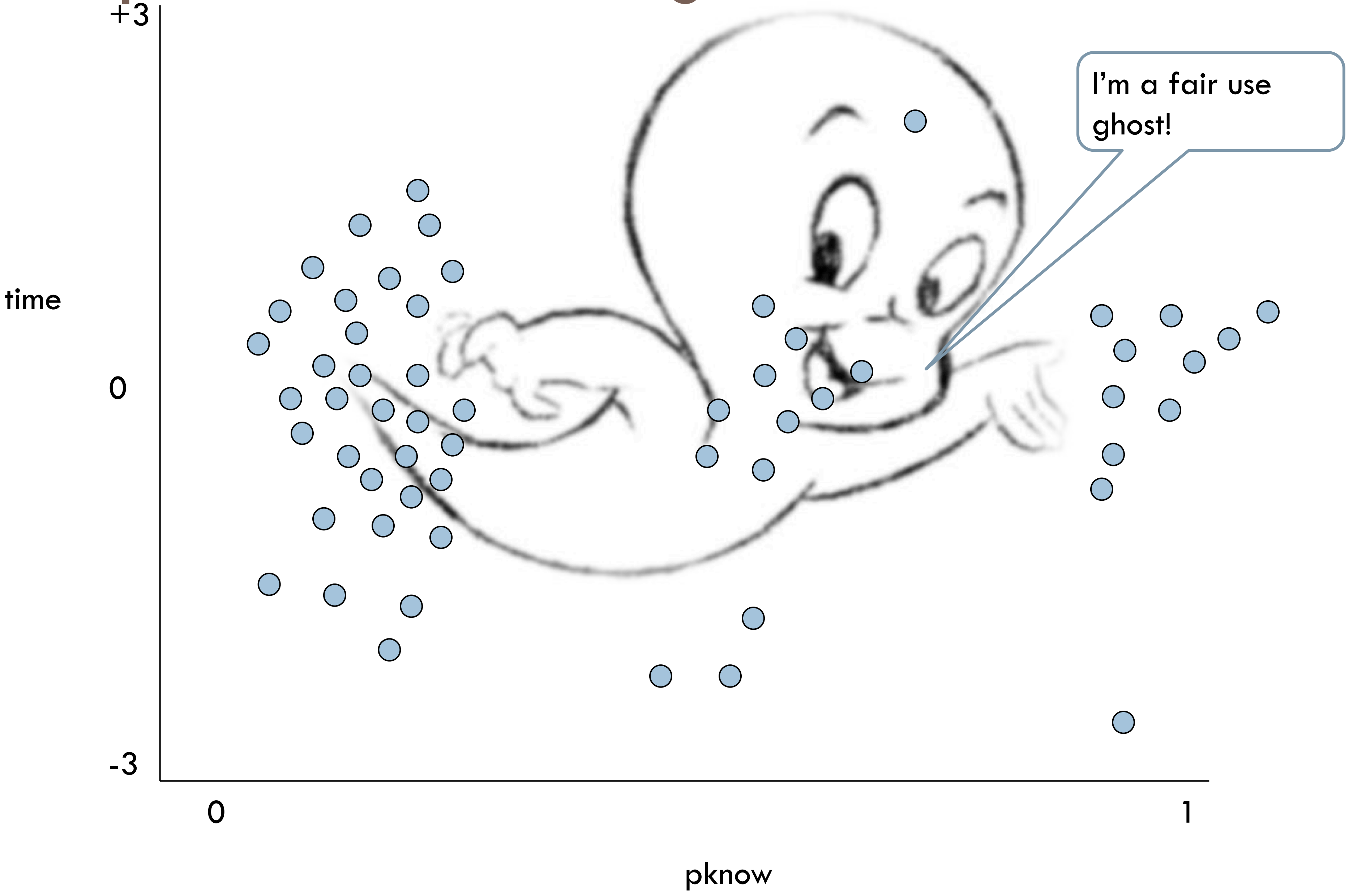


Disadvantages of GMMs

- Much slower to create than k-means
- Can be overkill for many problems

Spectral Clustering

Spectral Clustering

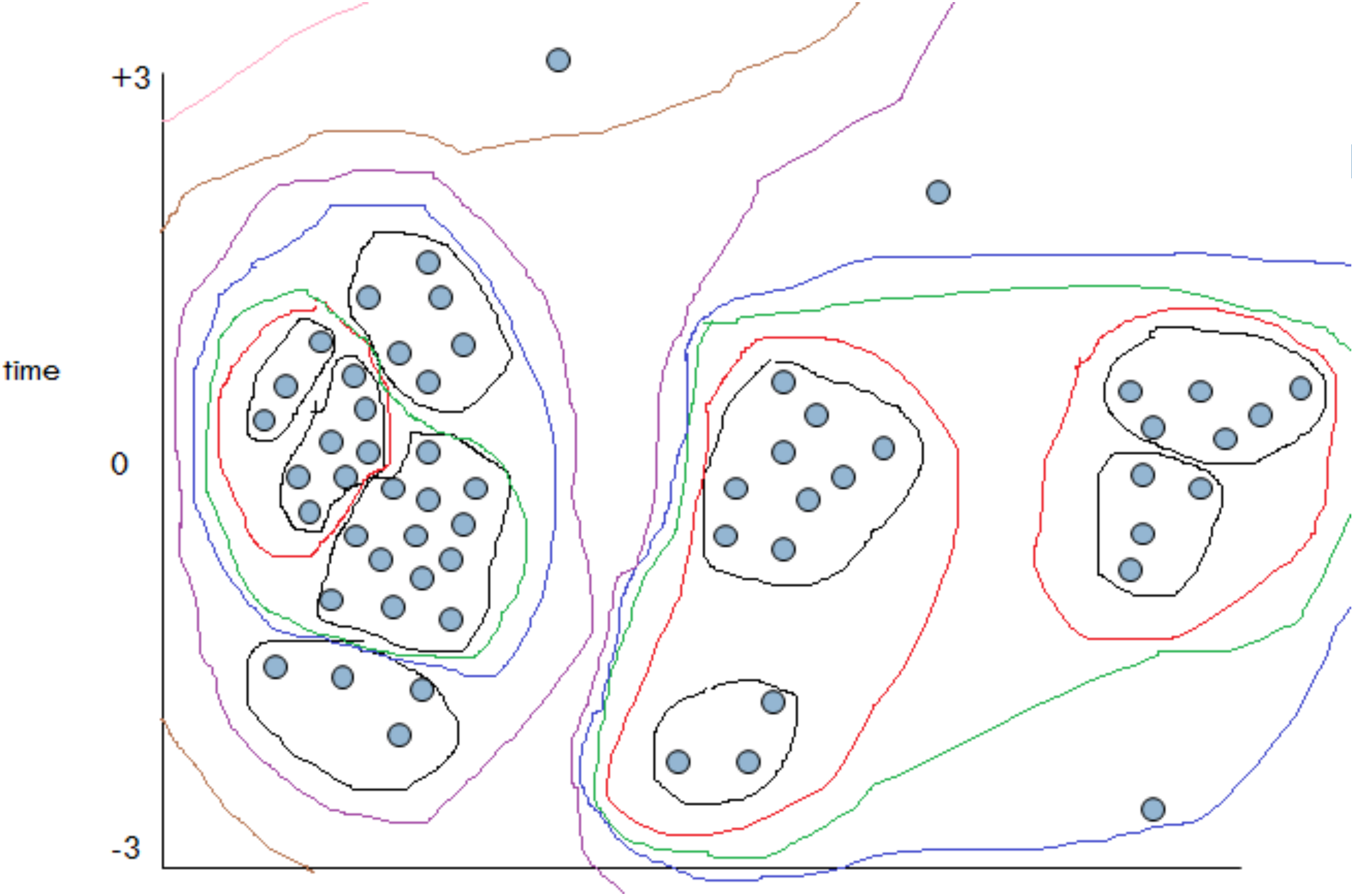


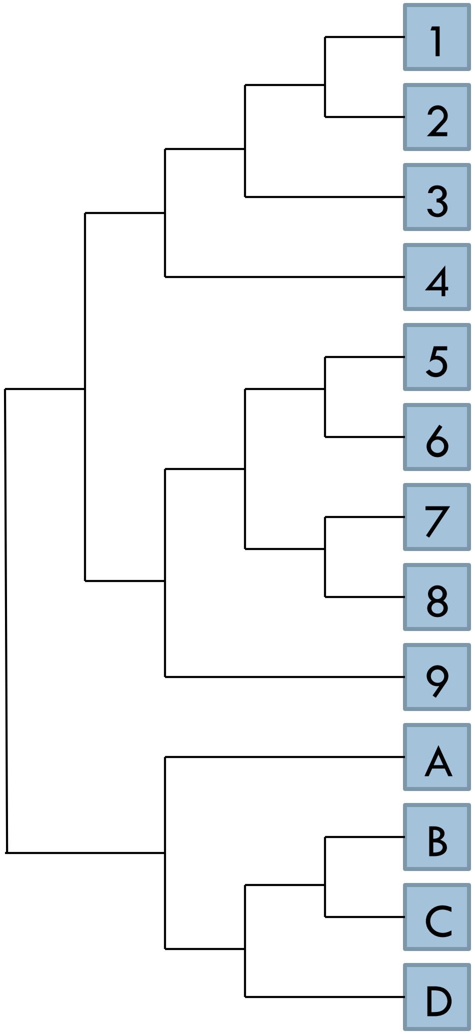
Spectral Clustering

- Conducts dimensionality reduction and then clustering
  - Like support vector machines
  - Mathematically equivalent to K-means clustering on a non-linear dimension-reduced space

Hierarchical Clustering

- Clusters can contain sub-clusters


+3
time
0
-3


1
2
3
4
5
6
7
8
9
A
B
C
D

Hierarchical Agglommerative Clustering (HAC)

- Each data point starts as its own cluster
- Two clusters are combined if the resulting fit is better
- Continue until no more clusters can be combined

Many types of clustering

- Which one you choose depends on what the data looks like
- And what kind of patterns you want to find

Next lecture

- Clustering – Some examples