January 30th, 2013 – Meeting Notes

Supervised Play Recognition (Celtics Data Only)

We start with data from 12 Celtics home games (Nov 7th – Dec 12), in these games we have "play calls" for over 1000 plays, totaling 77 unique play ids. We considered a subset of these plays for which we had at least 25 instances. In addition we consider only plays with ParentPlay=HalfCourt (could also be Transition, or SOB). This resulted in the following subset:

Table 1: Play Calls Considered

PlayID	PlayName	Count
26	Elbow	72
42	Floppy	58
45	13	39
51	Drop	41
92	Invert	38
98	Delay	47
164	Custom ATO	56

We begin by considering Elbow vs. Floppy. After some initial preprocessing (to remove possessions with stoppage) we transform position data for each possession (with either Elbow or Floppy label) into a feature vector consisting of 24 features. We consider three sets of 8 features stacked in such a way that temporal order is maintained. Each of the 8 features corresponds to a segment of the court, and we simply count the number of times a player is in that segment during the possession. Each possession is broken into three segments: beginning, middle and end. Features are extracted for each segment and then stacked.

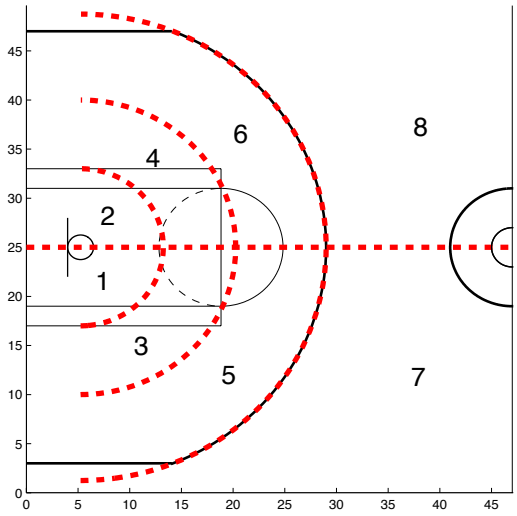

Figure 1: We divide the court into 8 "zones", and use this discretization and position data to calculate feature vectors for each possession.

This resulted in excellent separation for the two plays we initially considered:

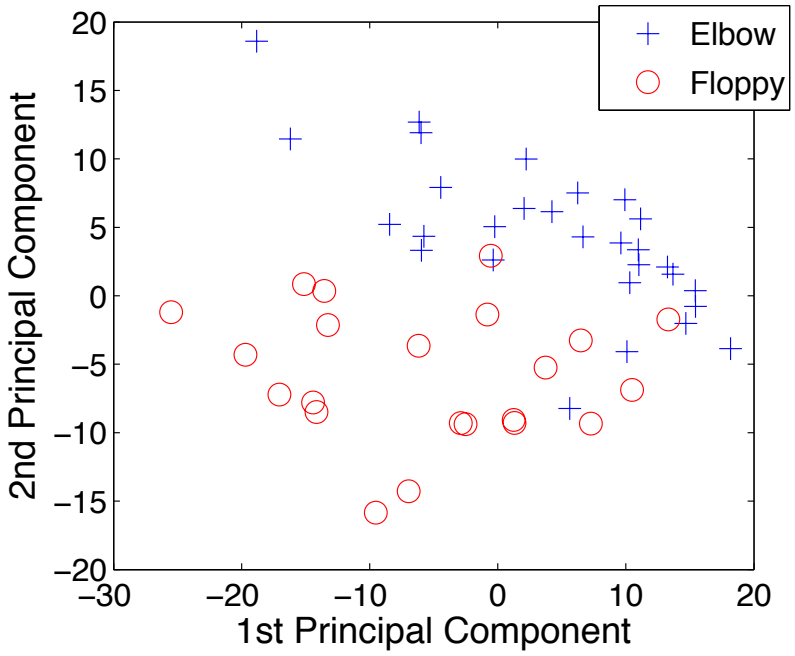

Figure 2: For visualization purposes only, we plot the play data for Elbow and Floppy in the two first principal components.

Using leave-one-out cross-validation we achieve an accuracy of 94% for predicting Floppy vs. Elbow. The SVM learns the weight of each feature (features with low weights omitted from table:

Table 2: Feature weights with the greatest absolute value

Feature ID	Weight
24	0.10
13	0.06
5	0.05
14	0.05
23	0.05
10	-0.06
19	-0.07
18	-0.08
20	-0.15

We repeated this analysis for the remaining pairs of plays (7 plays, 21 total pairs).

Table 3: Binary classification accuracy for all pairwise combinations

ID	Classifier	LOOCV Accuracy
'26-45'	Elbow vs. 13	0.89
'26-51'	Elbow vs. Drop	0.70
'26-92'	Elbow vs. Invert	0.61
'26-98'	Elbow vs. Delay	0.73
'26-164'	Elbow vs. Custom ATO	0.55
'42-45'	Floppy vs. 13	0.78
'42-51'	Floppy vs. Drop	0.80
'42-92'	Floppy vs. Invert	0.78
'42-98'	Floppy vs. Delay	0.91
'42-164'	Floppy vs. Custom ATO	0.89
'45-51'	13 vs. Drop	0.70
'45-92'	13 vs. Invert	0.68
'45-98'	13 vs. Delay	0.88
'45-164'	13 vs. Custom ATO	0.61
'51-92'	Drop vs. Invert	0.62
'51-98'	Drop vs. Delay	0.52
'51-164'	Drop vs. Custom ATO	0.33
'92-98'	Invert vs. Delay	0.68
'92-164'	Invert vs. Custom ATO	0.63
'98-164'	Delay vs. Custom ATO	0.63

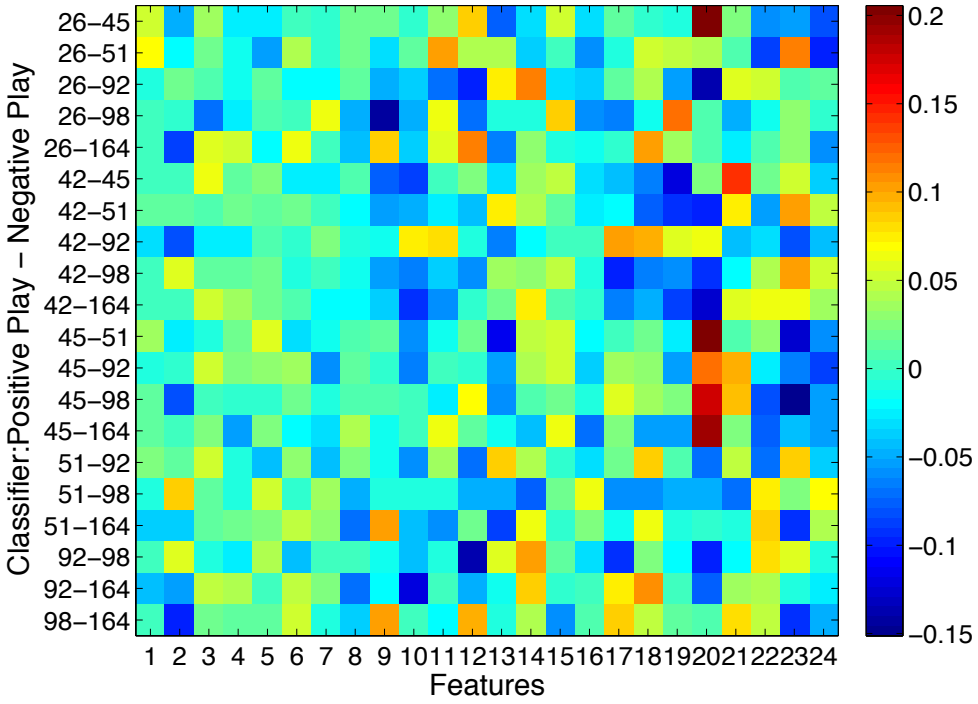

Figure 3: Distribution of feature weights across classifiers.

Unsupervised Play Recognition

In this analysis we start with all of the data (STATS has given us) for the 2011-2012 season. We consider all possessions in which there were no stoppages between the time the ball crosses half-court to the time the first field goal of the possession is released. In this preliminary analysis we consider only the position of the ball during this time interval. We consider all instances lasting [8,10] seconds in duration.

We resample the ball data in space instead of time, every 0.5 feet. This essentially is a way of reweighting the path, otherwise when calculating distance between paths parts where the ball is stationary would get greater weight.

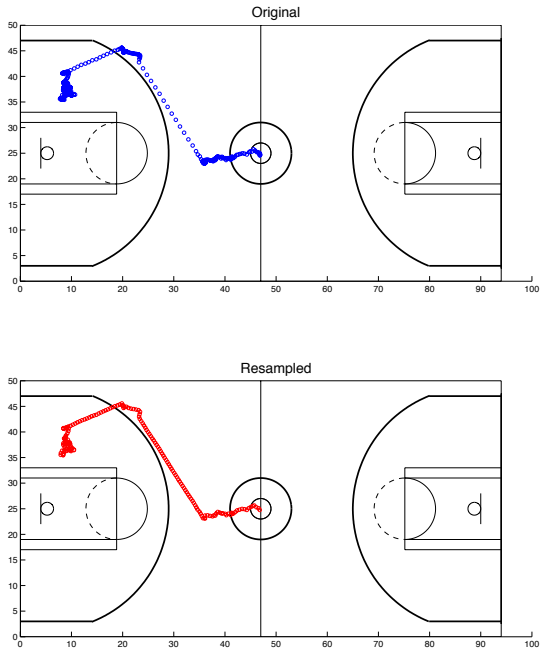

Figure 4: Resampling the ball position in space instead of time.

Next we consider the subset of ball paths of approximately the same length (100,120) feet to facilitate comparison. After this preprocessing we measure the distance between two possessions as Euclidean distance between ball paths (linearly interpolated to all be the same length).

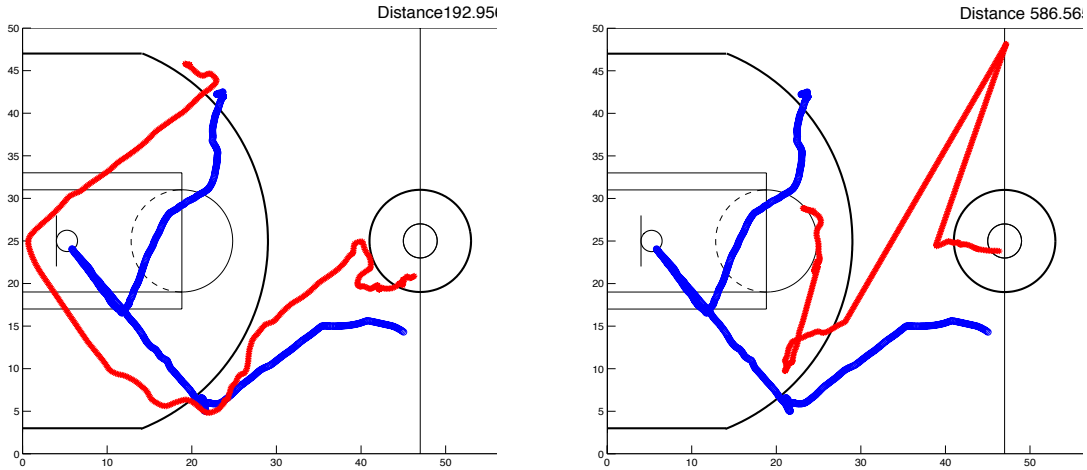

Figure 5: Developing a distance metric - Measuring the difference between pairs of ball trajectories.

Given this distance metric, we can cluster the data (agglomerative hierarchical clustering with average linkage criteria).

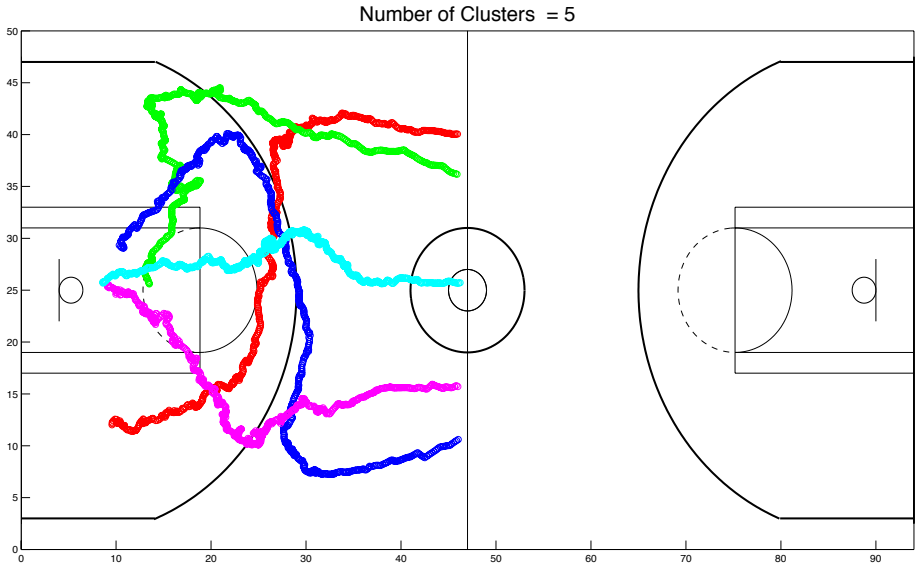

Figure 6: Representative ball trajectory of each cluster, when number of clusters is set to 5. Note the symmetry.

In addition to considering the position of the ball, we also consider the outcome of the possession. We assign +1 to the possession if the field goal is made or if the field goal is missed but immediately followed by a foul and free throws, otherwise we assign 0. We can then cluster the data, and assign a mean outcome to each cluster. We also note which team is in on offense and can look at the distribution across clusters for each team.

These are very preliminary results. We are considering only the path of the ball, whereas the positioning of the players will have more information (and more variability). Going forward we can incorporate these additional features in the distance metric to generate more interesting clusters. But this initial analysis suggests this unsupervised direction is promising for efficiently analyzing a large number of unlabeled possession data.

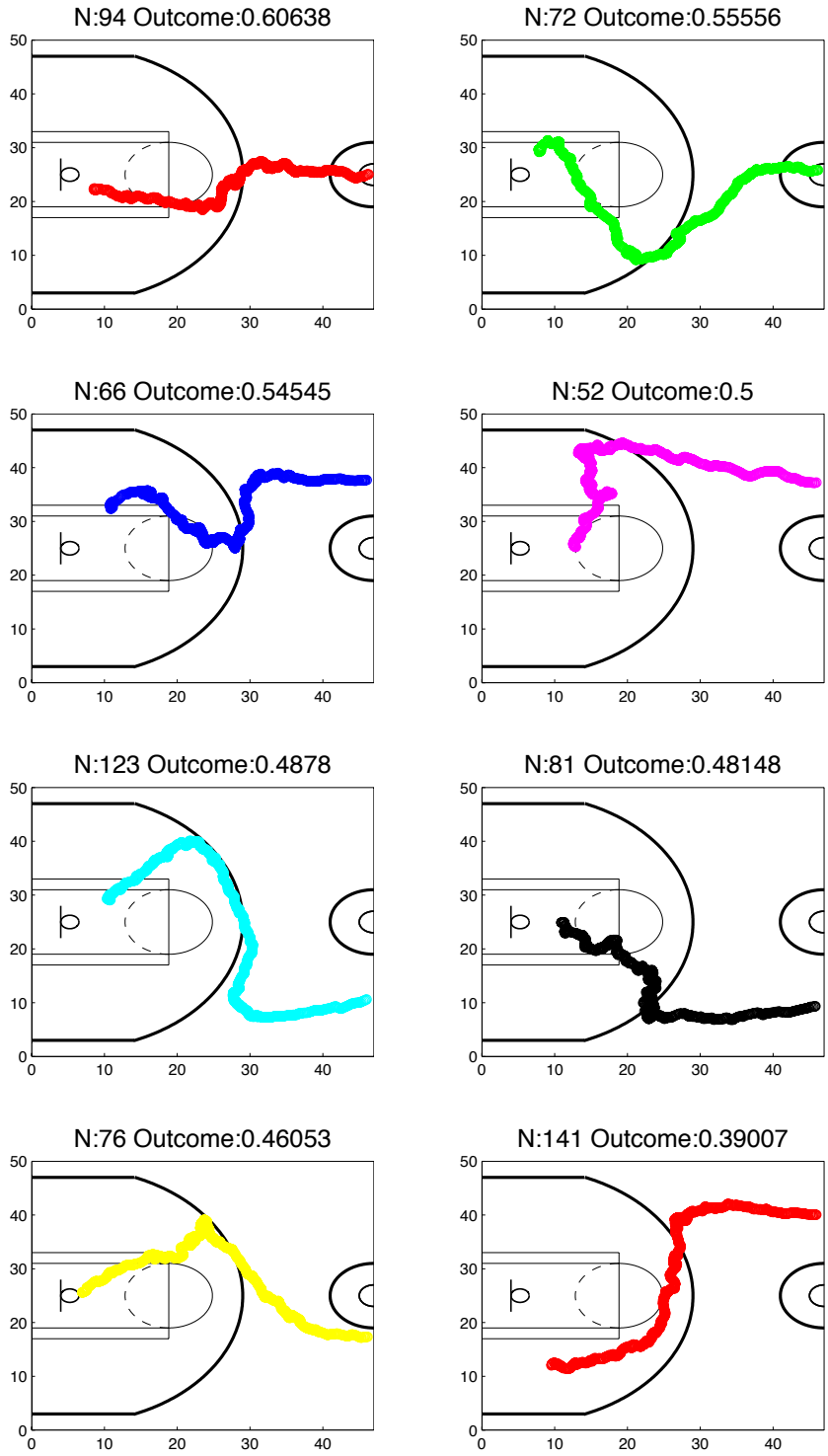

Figure 7: Representative trajectory when 10 clusters is chosen, each cluster is also associated with an average outcome (the closer to 1 the better). Note we only show clusters containing at least 10 instances.

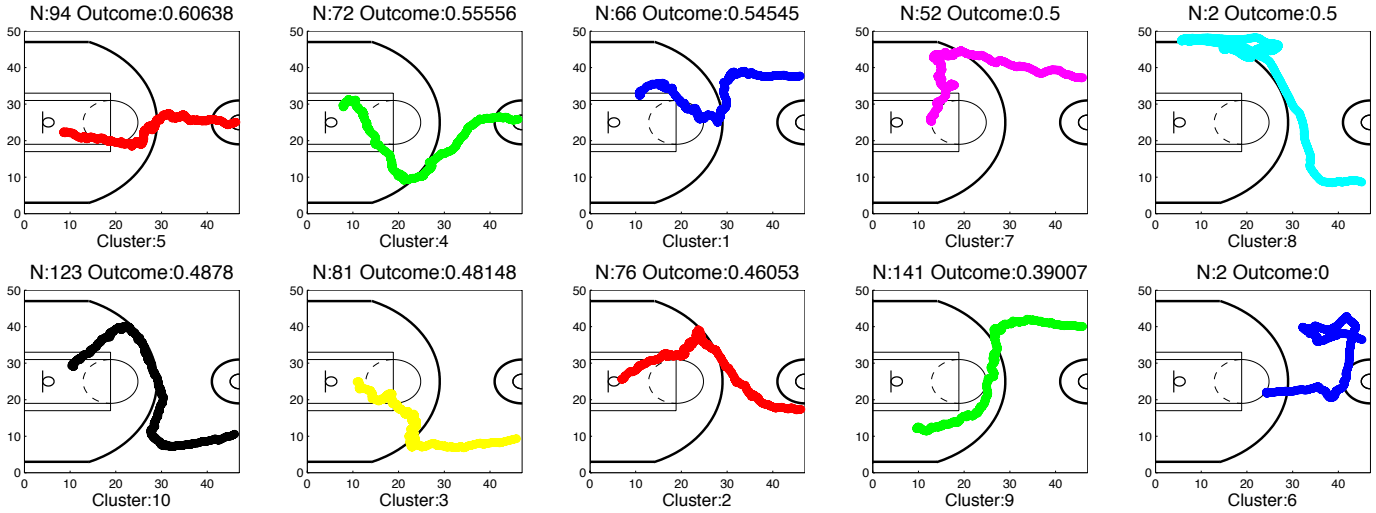

Figure 8: Representative trajectory of each cluster (nClusters=10).

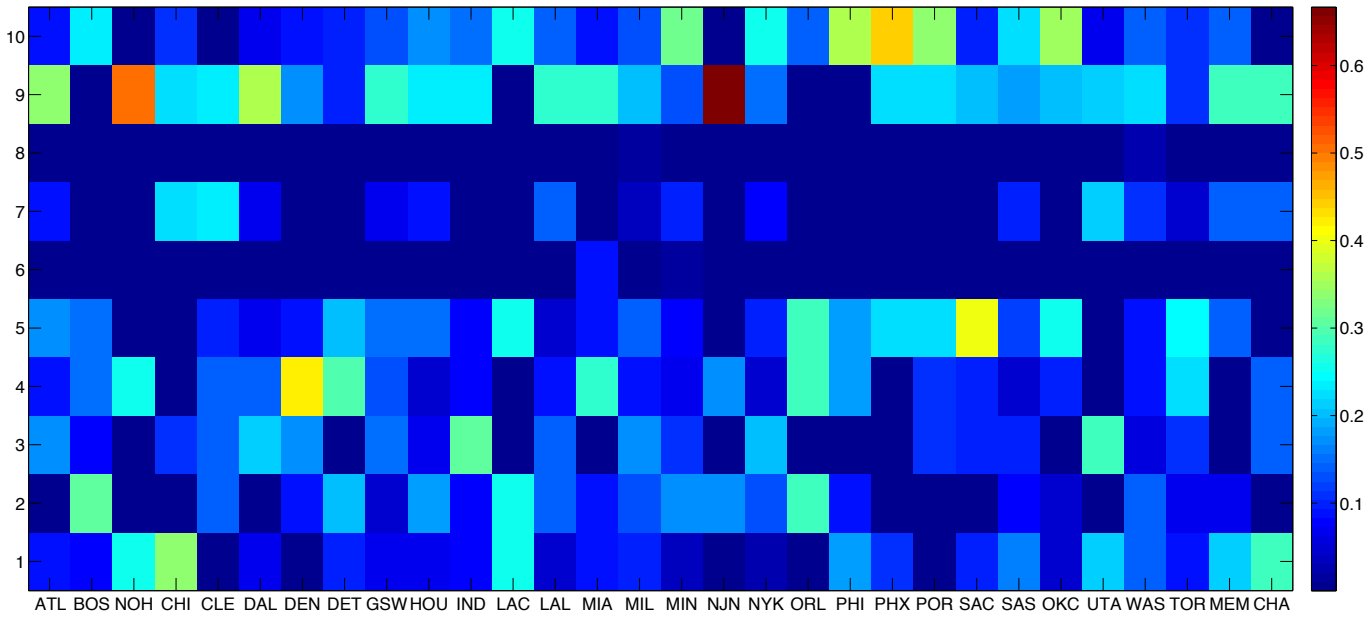

Figure 9: Distribution of possessions across clusters for each team.