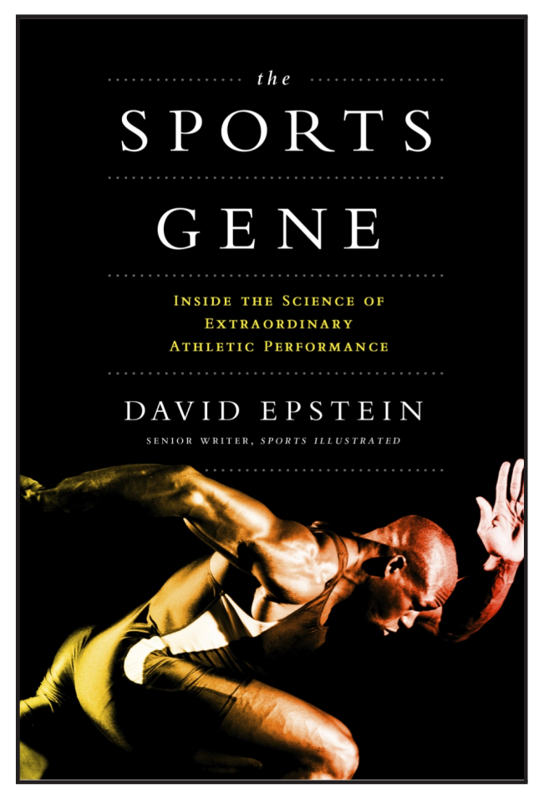


Classifying Olympic Athletes By Sport and Event

Matt Dickenson
mcd31@duke.edu

Motivation

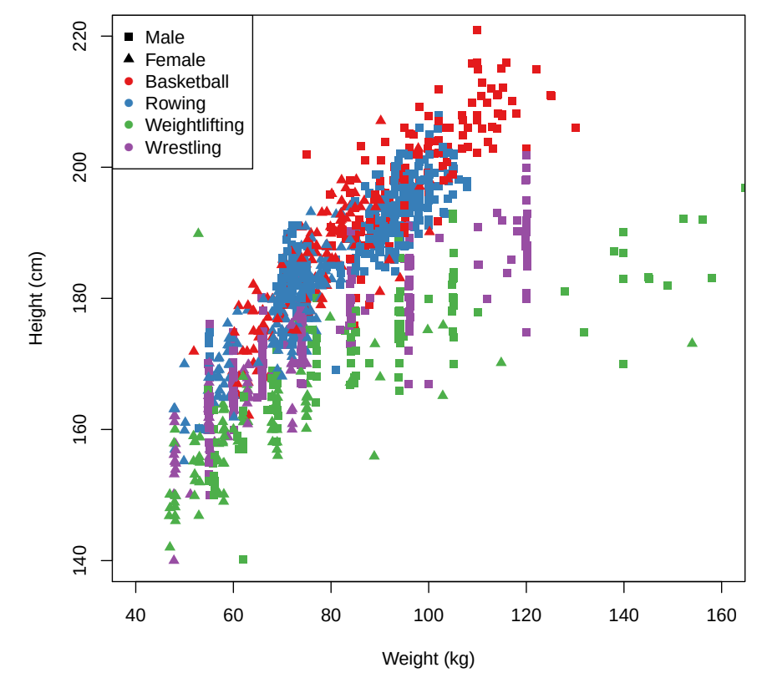
To what extent do biological traits determine sporting success? At the highest level of amateur sports–the Olympic games–we can notice differences in the physical characteristics of participating athletes across sports. Can these differences be exploited to classify individuals by sport or event given their physical attributes?

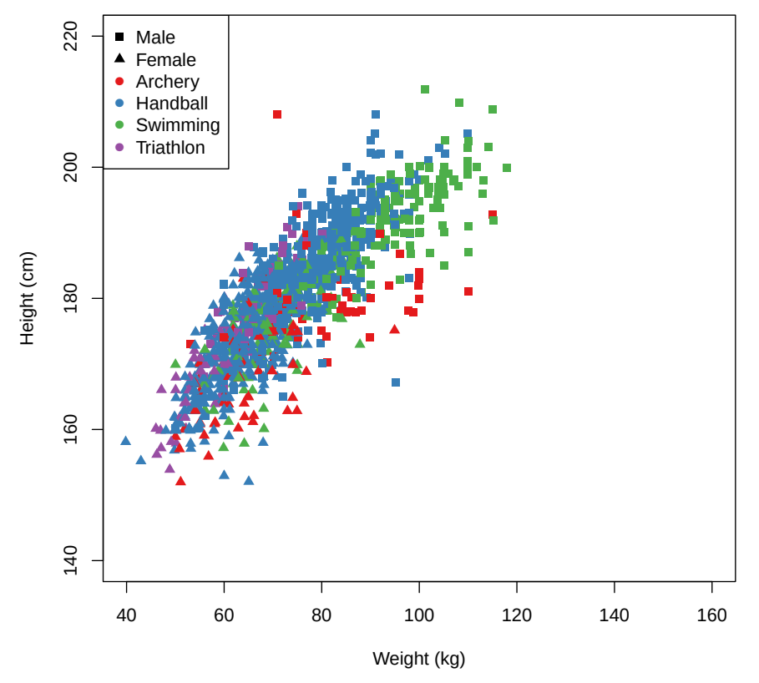
This project was inspired by a claim made by David Epstein, author of *The Sports Gene*. This claim is expressed in an interview with Russ Roberts:

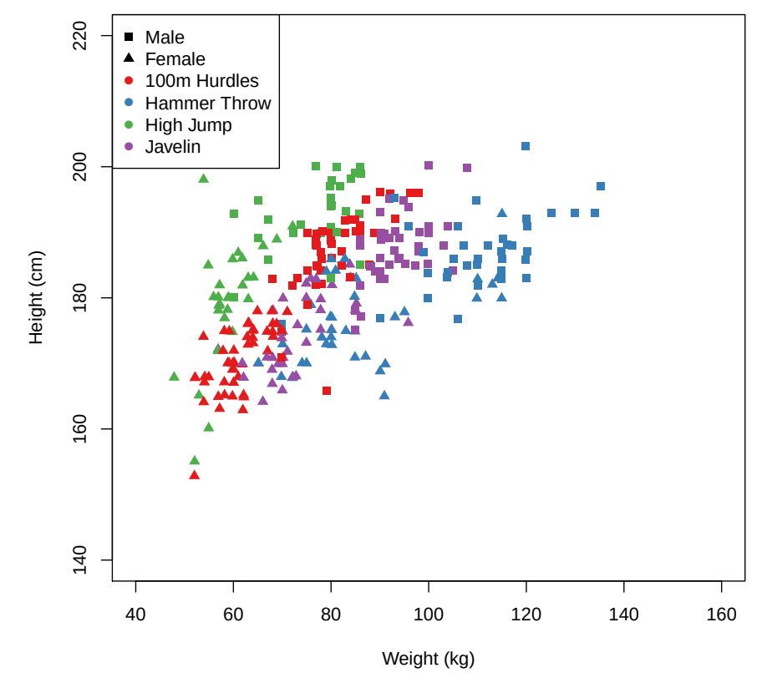
> **Roberts**: [You argue that] if you simply had the height and weight of an Olympic roster, you could do a pretty good job of guessing what their events are. Is that correct?
> **Epstein**: That's definitely correct. I don't think you would get every person accurately, but... *I think you would get the vast majority of them correctly.* And frankly, you could definitely do it easily if you had them charted on a height-and-weight graph, and I think you could do it for most positions in something like football as well.

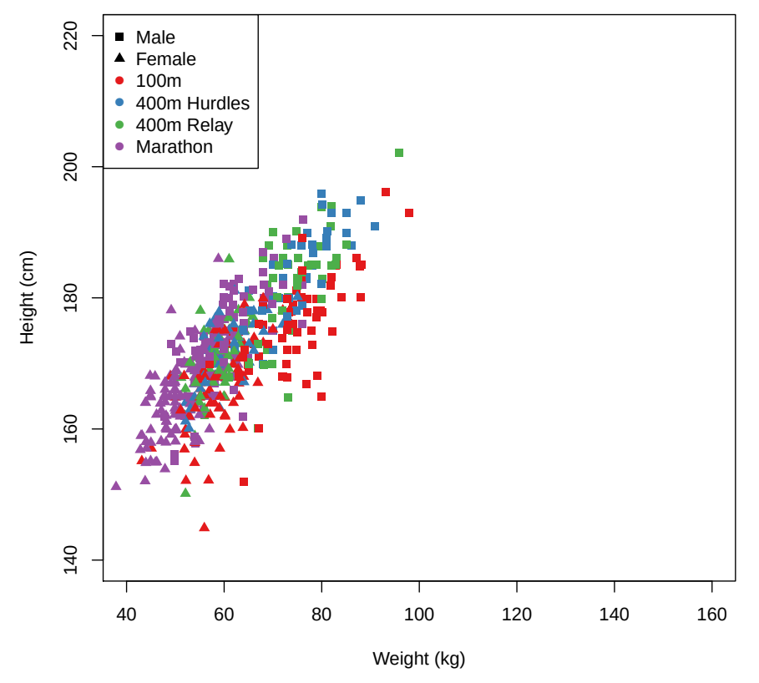
Data

Data was obtained for participants in the 2012 London Olympics from *The Guardian*. The original data consisted of 10,383 participants, which was reduced to 8,856 observations after data processing. Of these, 6,956 participants were split into training (n=3,520) and test (n=3,436) sets for classification by sports (k = 27). The remaining 1,900 Athletics participants were split into training (n=907) and test (n=993) sets for classification by event (k = 48). Participants' height, weight, age, and sex were used as features. Some sports and events exhibit relatively well-clustered features, whereas others are less defined.





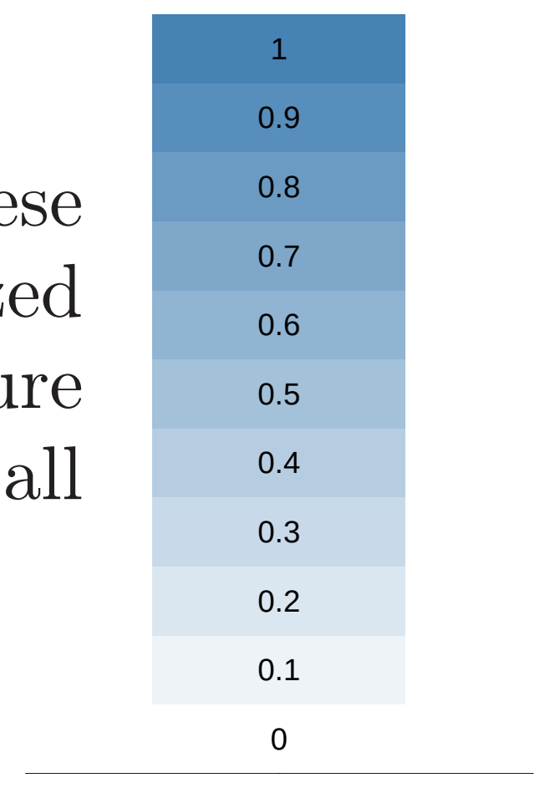




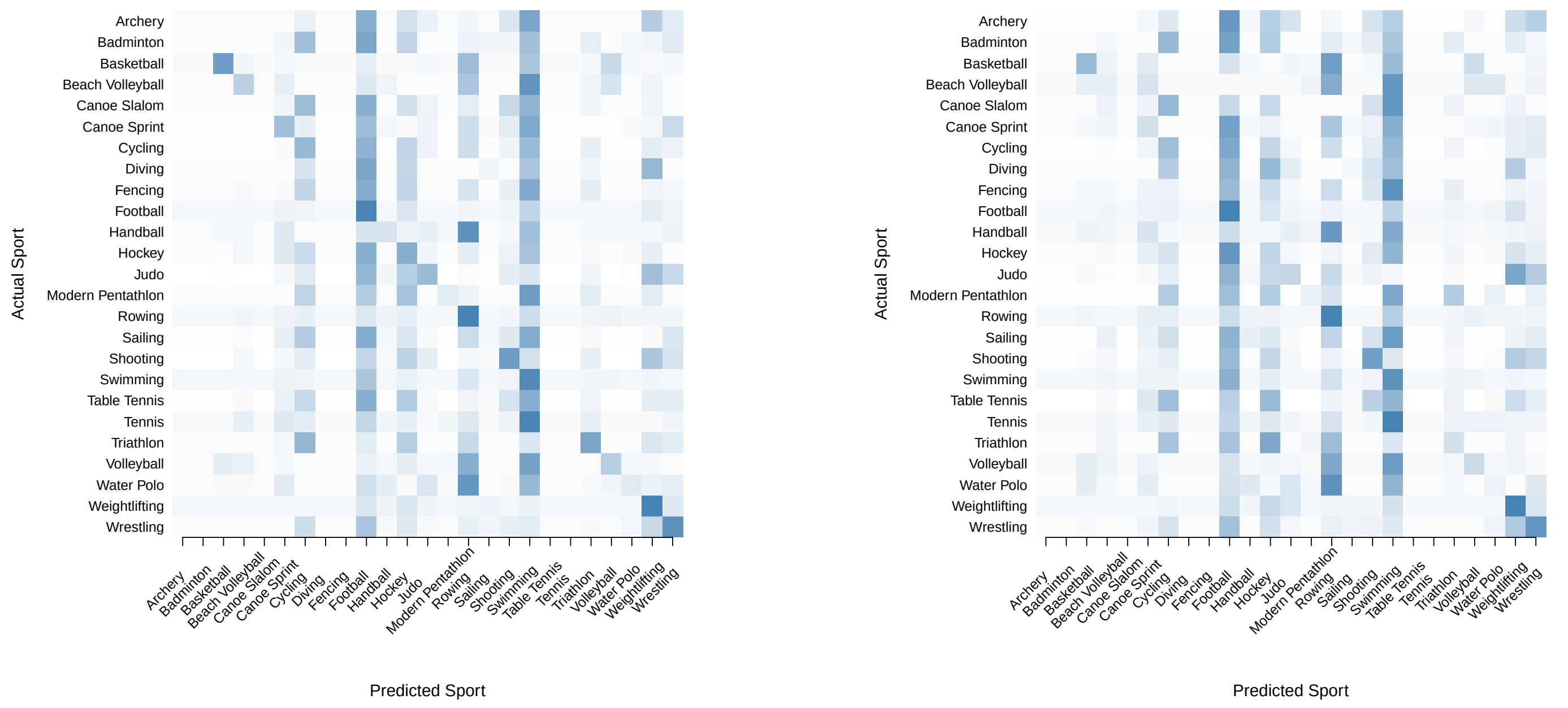
Methodology

Several machine learning methods were applied to this classification problem. Conditional inference trees were formed by recursive binary partitioning. Evolutionary trees seek a globally optimal tree by minimizing the misclassification rate. Breiman's Random Forest algorithm was used with 500 trees. Single-hidden-layer neural networks were constructed with 30 units in the hidden layer for sport classification and 50 for event classification.

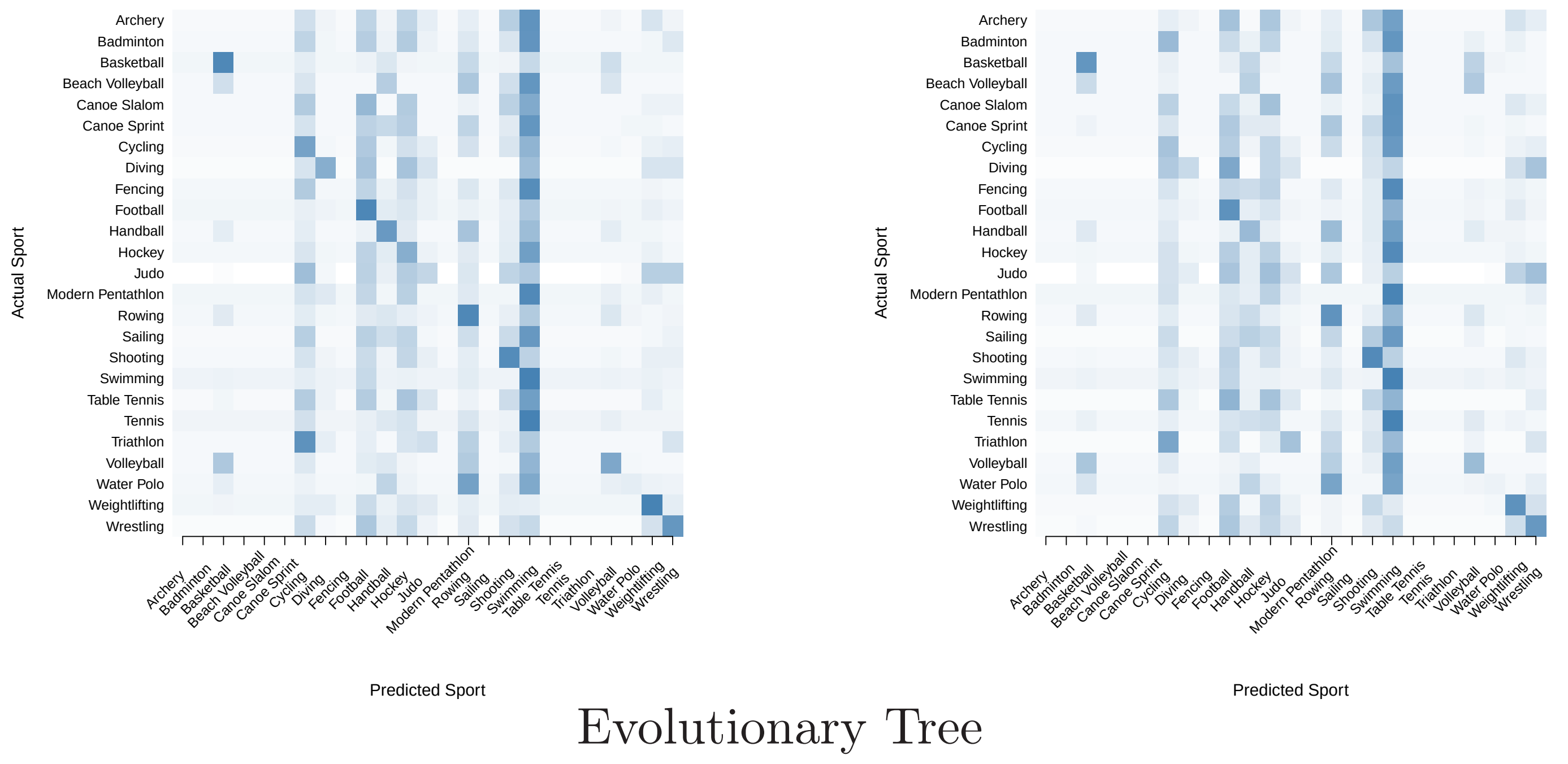
The columns at right present these results visually, with row-normalized observed frequencies. The figure at right serves as a legend for all heatmaps.



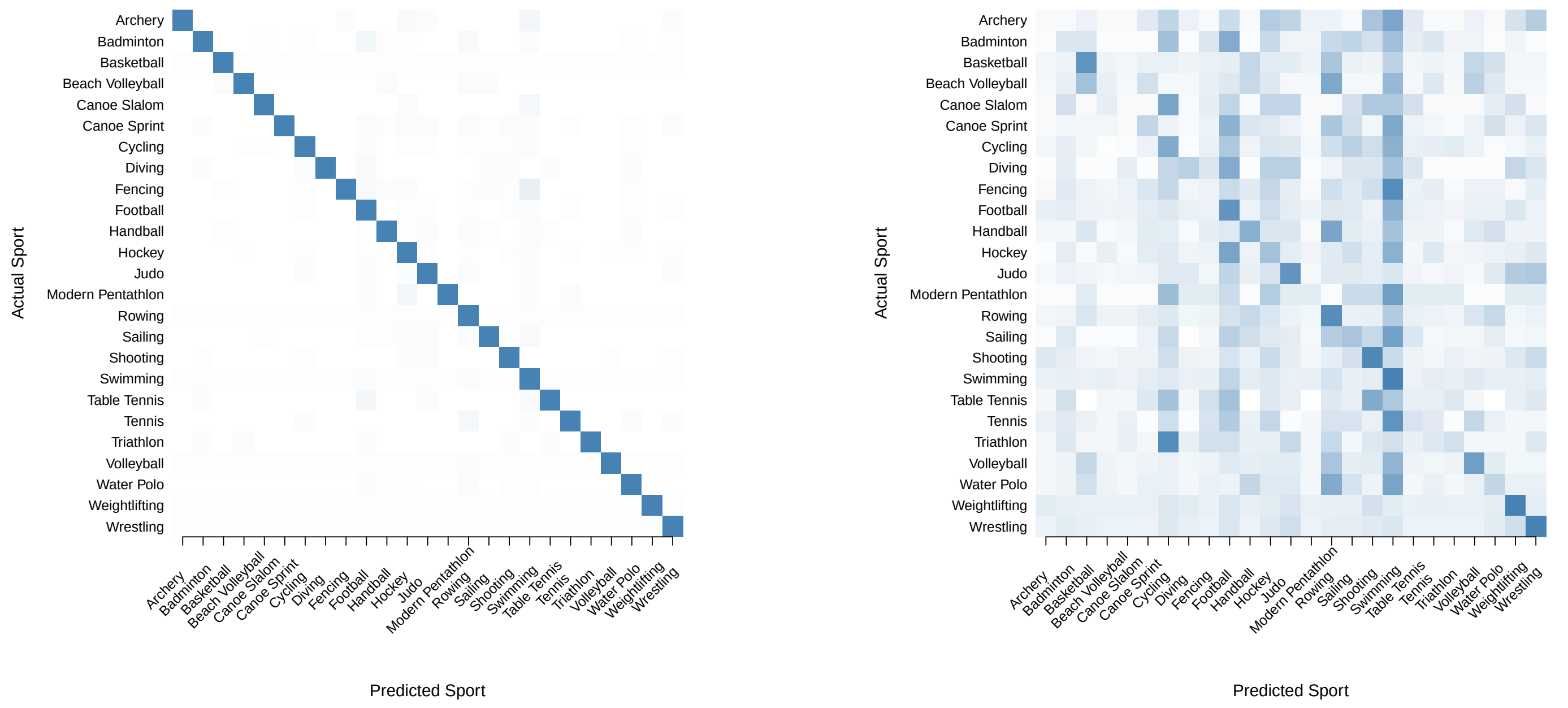
Classifying by Sport



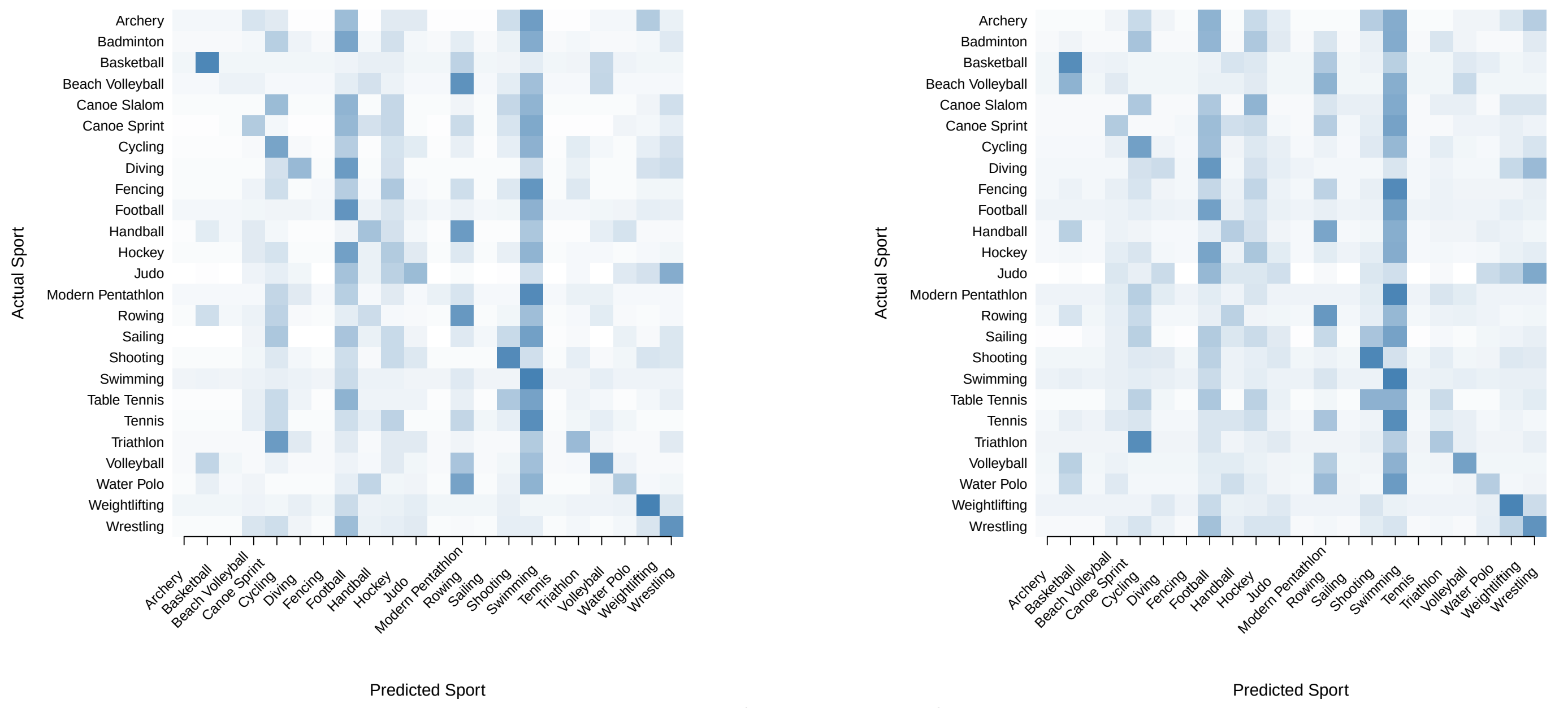
Conditional Inference Tree



Evolutionary Tree

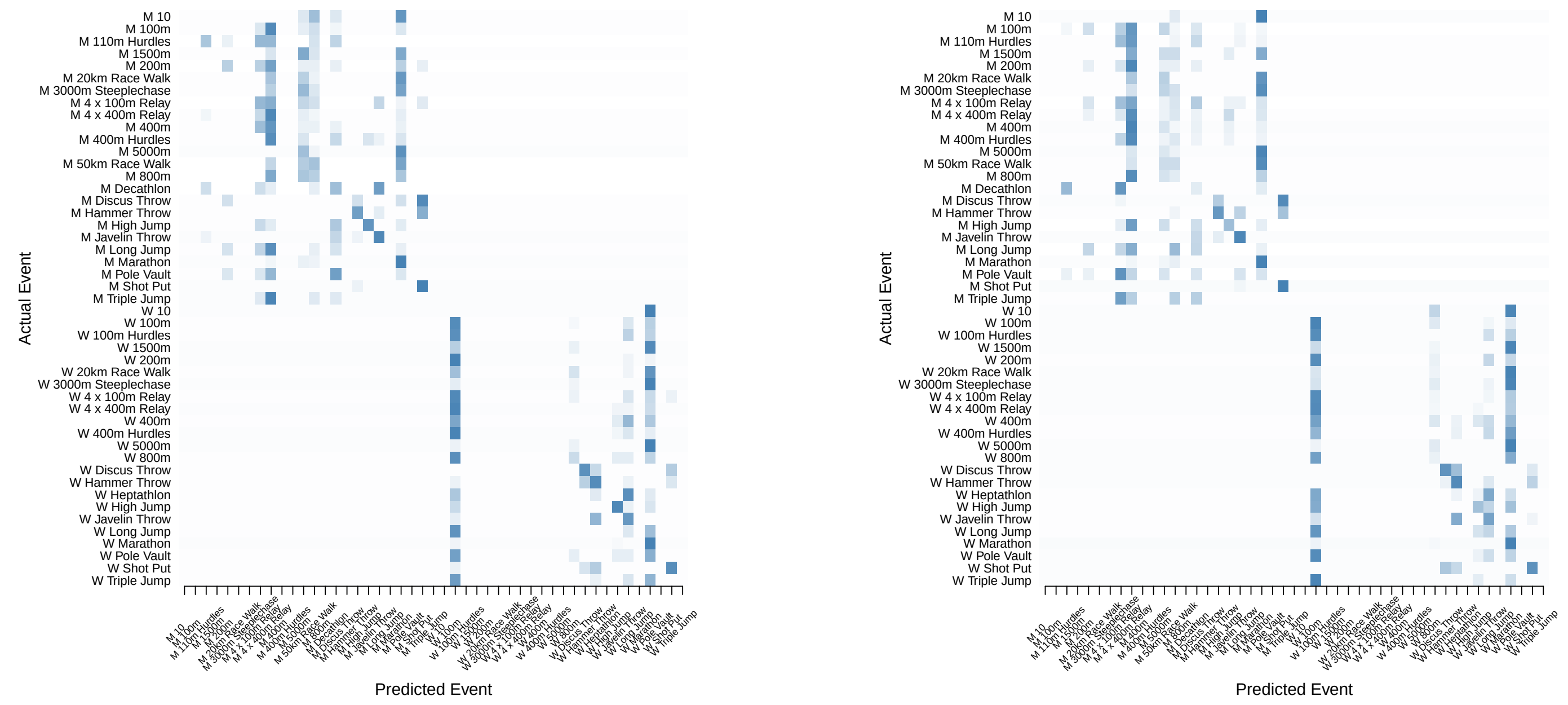


Random Forest

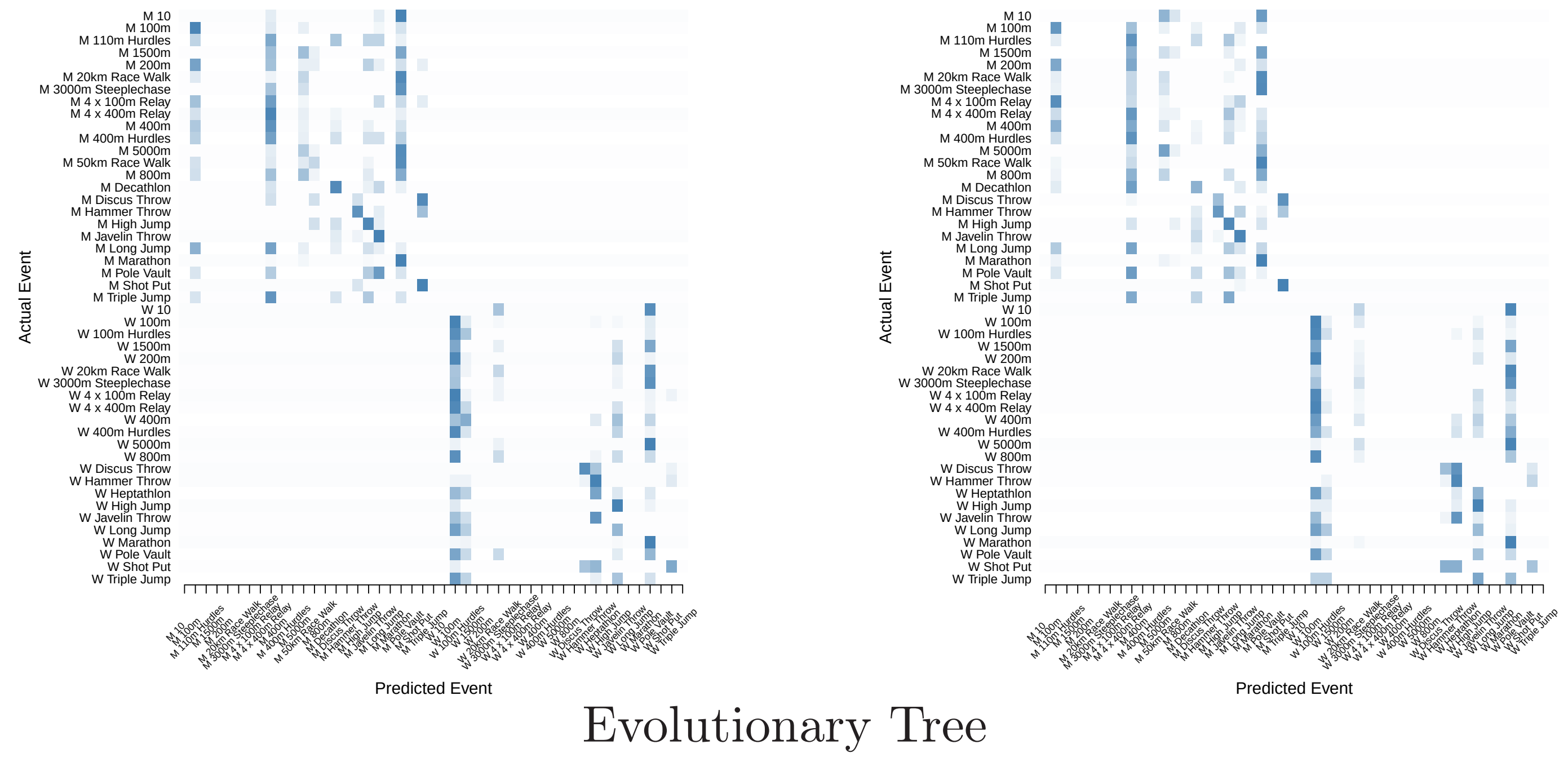


Neural Network

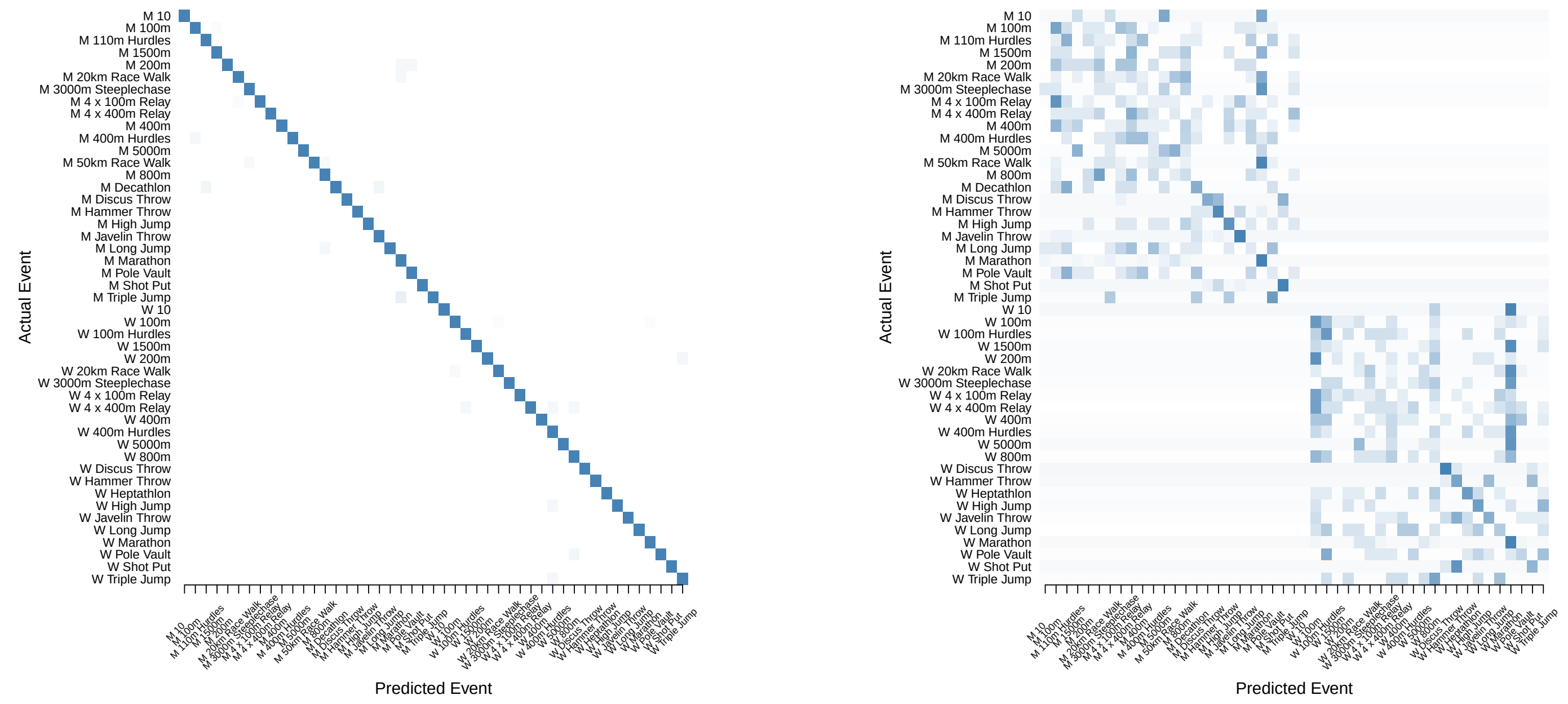
Classifying by Event



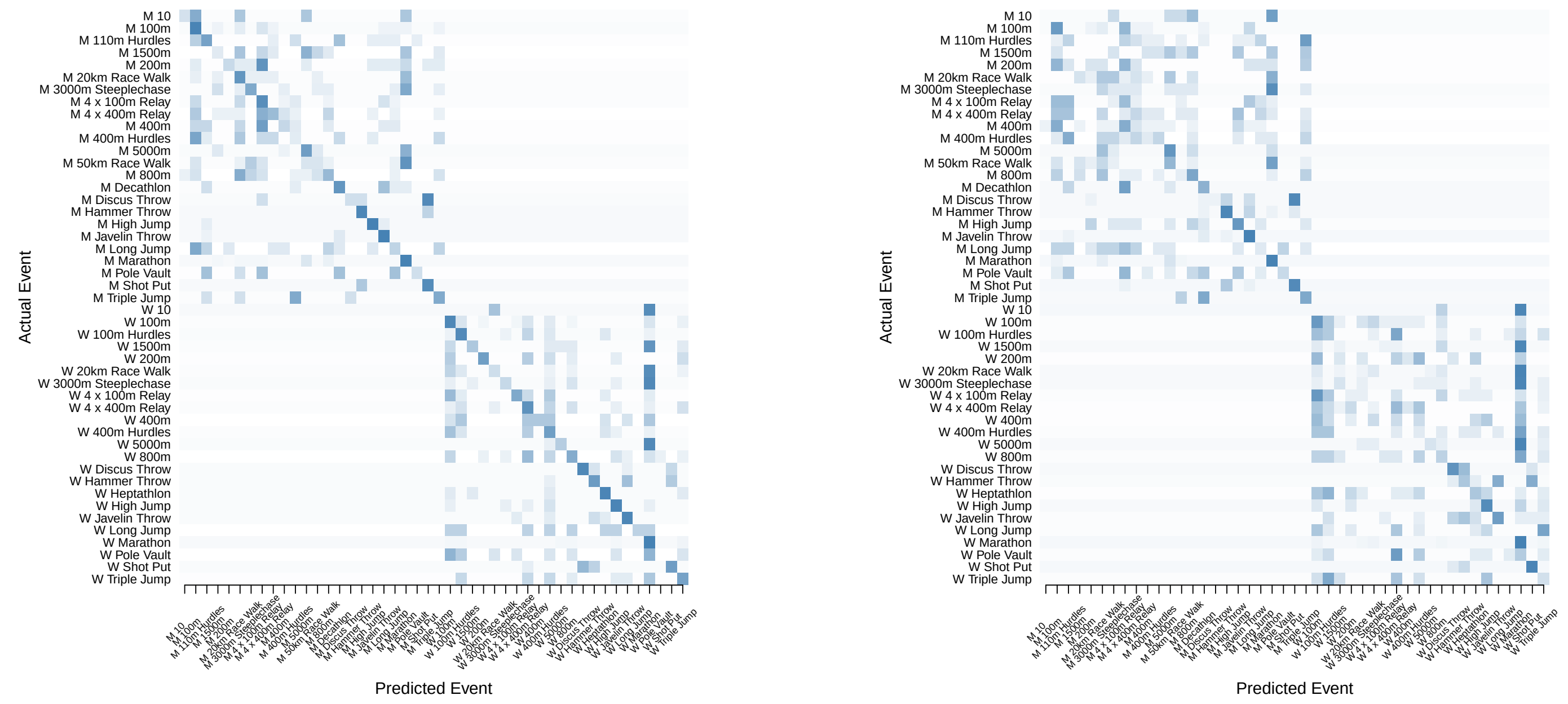
Conditional Inference Tree



Evolutionary Tree



Random Forest



Neural Network

Accuracy

Evolutionary trees and neural networks both exhibit higher accuracy for events than sports, and maintain acceptable out-of-sample accuracy. Conditional inference trees do slightly less well, for both training and test sets. Random forests tend to overfit the training data but still do well on the test set.

	Sports			Events		
	Train	Test	Ratio	Train	Test	Ratio
Conditional Inference Tree	.279	.219	.784	.277	.218	.787
Evolutionary Tree	.292	.236	.807	.303	.230	.757
Random Forest	.923	.244	.265	.976	.228	.233
Neural Network	.280	.265	.949	.397	.249	.623

Discussion

- Classifying athletes by sport can be achieved with moderate accuracy using only a few features
- Data with a large number of categories can be difficult to classify
- Traits of athletes in some sports and events exhibit noticeable clustering, while other categories are less distinct (multi-modal)
- Athletes in some sports and events have a well-defined body type, but Olympians exhibit a wide range of physical features