# NHL Deployment Clusters

## The Motivation

**What on earth do you mean by “Top 4 Defenceman?”,** I said to myself for the hundredth time while watching Singing Season this past July 1st. In recent years I’ve become increasingly frustrated by the lack of nuance in public discussions about archetypes a hockey player may fall into.

Phrases such as “Middle Six Winger” and “Bottom Pair Dman” are so ingrained in the vernacular of hockey culture, yet rarely are they expanded upon. Does this middle six winger play on the power play and does their coach trust them to log heavy minutes while trailing by a goal in the third period. Someone like Max Domi perhaps. Or does this player thrive on the penalty kill, as well as being trusted to bear the burdens of many defensive zone faceoffs, a Calle Jarnkrok like player.

These two player archetypes perform vastly different roles, and there’s very little in the public sphere of hockey data that shows their differences in a digestible and easy to understand way. This lack of a measure to convey such nuanced data is understandable however. Many facets come into play when considering player deployment, and data larger than 2 dimensions does not lend itself well to plotting or charts.

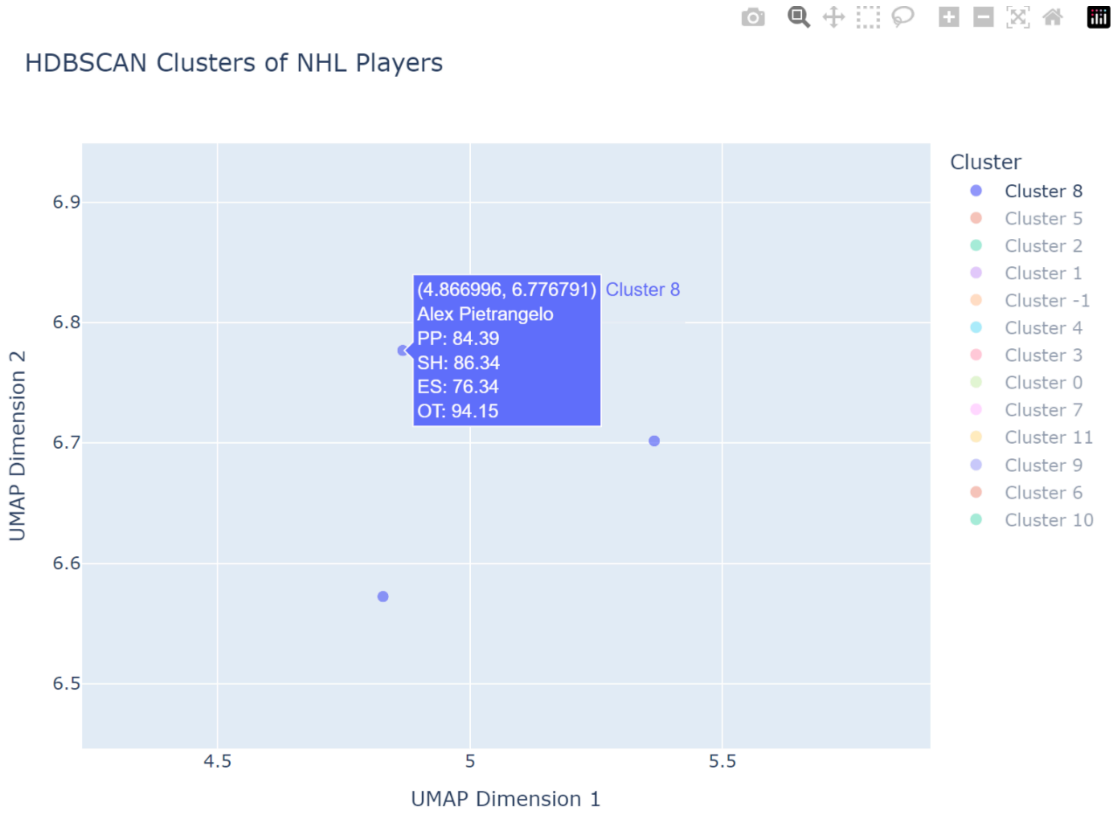
## The Solution

Much of my time spent at my day job as a data scientist specializing in Natural Language Processing has me working with text embeddings, massive vectors with hundreds of dimensions required to represent the semantic meaning of human readable text. To extract value from these embeddings, I frequently turn to state-of-the-art algorithms UMAP and HDBSCAN. Essentially UMAP reduces the dimensionality of massive vectors into human understandable space, while HDBSCAN is used to create hierarchical clusters of data points that have common traits. Together what they leave you with is a 2D scatter plot of clusters that can be grouped together based on an enormous number of variables. A perfect solution for factoring in all sorts of variables that go into NHL player deployment, yet still delivering a digestible solution to the end consumer.

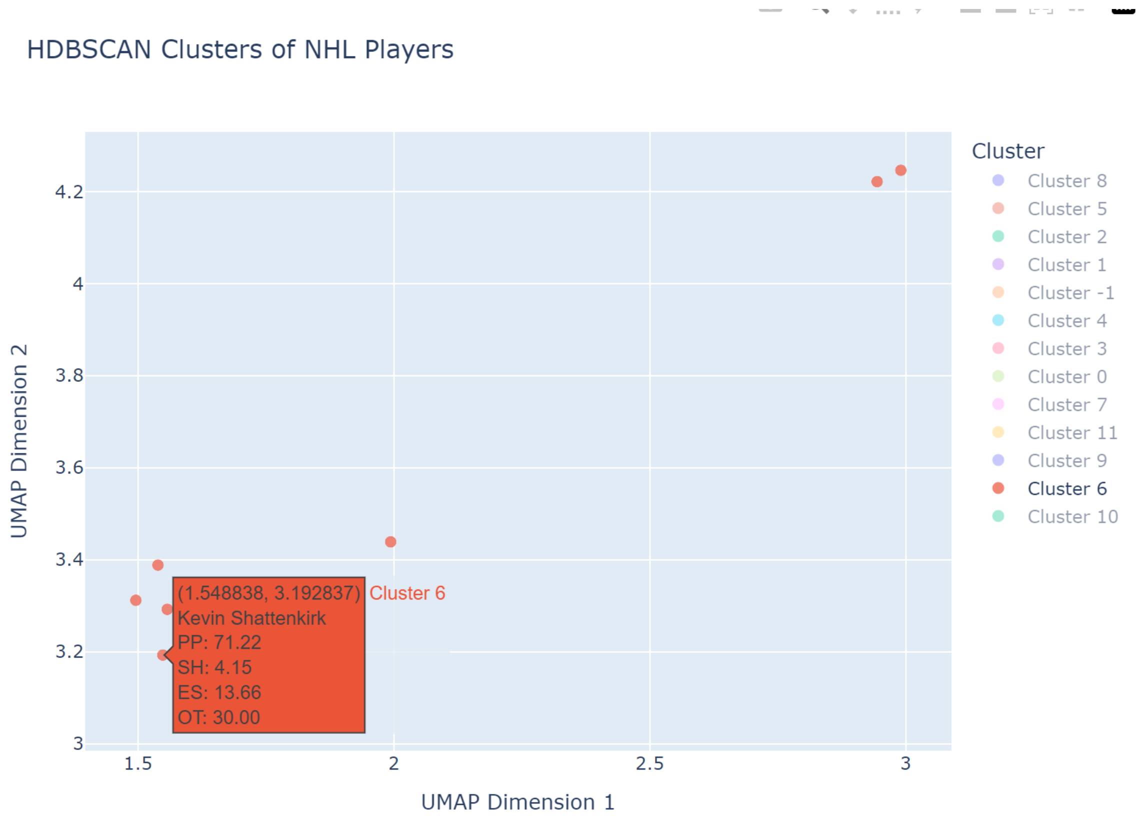
## In Practice

In the plots below, I used time on ice data from the 2023-24 season of players who played a minimum of 40 NHL games. I generated 2 plots, one for forwards and defencemen. Clusters were generated based on 4 deployment percentiles, even strength, power play, short handed and overtime. The idea was that by gaining a deeper understanding of how a coach deploys their players, we can better know a players role on a team, and the player archetype they fall into.

Starting off with the defence, let's take a look at cluster 8, one most hockey fans would easily spot on their favourite team, the work horse defencemen. This player is often found on the ice in all game situations, munching minutes as a steady hand in their coaches arsenal. It should come as no surprise to see players such as Alex Pietrangelo, Mike Matheson and Cam Fowler showing up here as each of these players clear the 75th percentile in each deployment type.



In cluster 6, we see a totally different type of player, the power play specialist. Here we have players that don’t necessarily see a ton of ice time, but when they do, it's often with the man advantage. The majority of these players are in the 60th percentile of power play usage or higher by defenceman, but rank significantly lower in other situations.



## The Implications

While these simple TOI metrics that I’ve used are by no means earth shattering, they do speak to the potential of utilizing HDBSCAN and UMAP to generate bins of players when even more nuanced data is added as dimensions to our player vectors. My goal here wasn’t to go into super deep detail on how I could use advanced quality of competition metrics in my vector. It was to show that with limited data, these algorithms were able to generate reasonably smart bins of player archetypes.

For example, one could utilize statistics such as ice time when a team is down or up a goal or defensive/offensive zone start percentages. Going even further,  one could even use their own expected goals for and against metrics so that these algorithms not only grouped players by their deployment, but also their success in said deployment.

## Conclusions

With limited data, hdbscan and umap effectively grouped players into bins in an accessible and digestible way. These clusters are well beyond what most hockey media classifies players into, and I am extremely excited to continue this work with more advanced metrics added into the pot going forward. I see these algorithms as being a great way for front offices to assess a player archetype they are hoping to acquire, as well as validate how a player succeeds in that role currently.