# Drawbacks and Alternatives

KNN has several drawbacks for predicting real time location. The model is not easily transferable without the training data, it classifies too many categories, its speed decreases with more observations, and it could have difficulty with outliers.

KNN predictions are made by comparing a point to its closest neighbors then determining its location by either voting or an average of its neighbors’ locations. This means you must have other data points to make a prediction. If the training set is large, it would need to be copied to every location where the model would need to be run. If it is running on a single server somewhere that might not be an issue, but if it is running on user’s cell phones that might make the transfer time consuming and the storage cost to the user might be prohibitive.

Large training set sizes could also be an issue when it comes time to compute the prediction. The more observations in the training set, the more need to be searched through to find the nearest neighbors, increasing prediction time. The cost of this would be dependent on the use case. If it is merely observational, perhaps a few seconds difference might not matter. But it a decision needs to be made quickly based on location, then it could be a significant problem. For instance, if a robotic currier needed to decide where to turn based on its location, or trying to intercept an item moving through a storage facility.

A classification approach uses a seperate class for every x-y pair, and is probably not the best since it doesn't help the model realize 0-1 and 0-2 are close to one another. We think it would be difficult to get results with this approach.

Finally, there could be issues with outliers or irregular spaces. A new data point far outside the space would be placed by the average of its three closest neighbors, which could be far away from its actual location. I think that would manifest itself as a “ghost”, where the system would say something was present when the signal was coming from outside the space. Depending on the sensors used, there might be some maximum distance the sensors can transmit which might mitigate this issue. Another issue might be when the nearest neighbors might place an object outside the space.

A regression or random forest approach might solve the first two issues. It would also be more natural to represent the targets as continuous rather than levels of a categorical variable. These models are more portable in that they do not require the training data to make predictions and faster because they don’t need to search the training set for nearest neighbors. To address the outlier issue, we could limit our predictions based on the boundaries of the space. Any predictions outside the space are move to the nearest in-space location. We could potentially deal with outliers in preprocessing, perhaps looking for signal strengths that don’t make sense for the space.

# Conclusion

We built six models using KNN and multi-target regression methods, with a closer look at two mac addresses that appear to be in the same location. Each method tested three models, two that used each co-located mac address separately with the other 5 mac addresses, and one with all 7 mac addresses. The multi-target regression looks to address some of the issues with the KNN method.

The KNN model performed best with 3 neighbors and keeping all 7 mac addresses. We achieved 52.96% accuracy, however the worst model was not far behind at 49.87%. This method has drawbacks including being difficult to transport, modeling continuous variables as categorical, long run times and difficulty with outliers.

The multi-target regression represents the target variables as continuous, which seems more natural for this coordinate system. We used an average Euclidean distance metric to assess performance, and we found that keeping just the mac address "00:0f:a3:39:e1:c0" performed the best, rather than keeping both it and its co-located partner. This method has its own issues. First, most of the predictions fall along the center of the space, so it did not capture the non-linearity in the data well. Second, predictions can occur outside the space, so in practice we would need rules to confine predictions to the floor plan.

We would recommend additional feature creation and testing other algorithms to see if we can better capture the non-linearity of the data and improve our predictions.

## References

(http://rdatasciencecases.org/) Case Studies in Data Science with R by Nolan and Lang

(https://www.jstatsoft.org/article/view/v018i02) The pls Package: Principal Component and Partial Least Squares Regression in R, Mevik, Journal of Statistical Software