Often future business owners find themselves trying to figure out whether or not their businesses will be successful in certain domains. Whether a business is a food restaurant, grocery store, hardware store, medical office, or any other number of businesses, the general belief is that an establishment will do better if their services are suitable to and/or needed in the surrounding neighborhood. Taking this into account, businesses make decisions regarding whether they will invest in certain locations. If the calculation of anticipated business is under- or overestimated it can cause a business to fail and they suffer a large economic loss due to high start-up costs. Generating a method to estimate whether or not a proposed business is a good idea beforehand could be very useful to entrepreneurs.

Our solution to this problem was to build a Nearest Neighbor classifier, which we deemed useful based upon the notion that businesses tend to do better in certain locations which cater to the corresponding local demographic. With the Yelp dataset, we gained access to information about over 61,000 businesses in 10 countries. Our k-NN sought to compare a proposed business to the existing businesses in the dataset to determine the likely success, in terms of star ratings, of the proposed business. After initially looking at the data, it became clear that the most important attributes for the task were the latitude, longitude, and the categories section. We discovered very quickly that a lot of the data in the data set was either redundant or did not have an effect. For example, there were neighborhood, city, and state attributes that were redundant because we already had numerical values for the business' latitude and longitude. Aside from location information, it seemed that the strongest mark of the similarity of businesses lied in whether or not they were offering the same services (e.g. Restaurant, Medical Services, etc.). With these two things taken into account (i.e. location and business type), the other information provided about the businesses was redundant or unhelpful in determining a "Nearest Neighbor," or similar business. Our task, then, was to build a k-NN classifier which would push the less similar businesses (in location, business type, or both) far enough away that they did not have an influence on the classification while more similar businesses (in both location and business, concurrently) would be classified as a "Nearby" Neighbor and influence the overall classification.

As previously stated, the attributes used in our Nearest Neighbor were location, as determined by its longitude and latitude, and the businesses' classifications into categories (e.g. Food, Restaurant, Medical Services). Our classifier incorporated a Euclidean distance calculation based on location and coupled this result with a weighted calculation of the inverse of the Jaccard index, or Jaccard similarity coefficient, based on shared categories between businesses. While the Euclidean distance reflected how close or far a business and proposed business were located, the inverse of the Jaccard distance served to reflect the importance of similarity in the categories section of the businesses. In this way, a larger distance and/or a larger weighted inverse of the Jaccard index would represent dissimilarity of businesses, and push them away in our Nearest Neighbor system. Alternatively, a small Euclidean distance for location and/or a small weighted inverse Jaccard index was an indicator of similarity between businesses and, thus, businesses with these features were deemed "Nearby" Neighbors. In addition to location and business classification information, the classifier takes a value, *k,* for the k-Nearest Neighbor and then averages the stars of these k-Nearest Neighbors to suggest an expected rating value for the proposed business.

In order to test and train our data we first created a ZeroR classifier to benchmark our k-NN classifier's performance. The classifying attribute was the 'stars,' or the rating that the businesses instances received. The ZeroR classifier ignores all predictors and simply looks at the overall classification of a business, or the number of stars it received in its rating. Since this classification is numerical, it takes the mean number of stars and makes the prediction that any proposed business would garner this number of stars. Our ZeroR classifier's performance was measured by taking the difference between the mean rating and actual rating for each data point. The average of these differences, or the variance, was calculated to represent ZeroR's performance (0.7942). Trying a number of different *k* values and Jaccard index weighting values, we then tested our data. We used 0.03% cross-validation (while this is a small percentage it still represents a large number of businesses and was chosen due to the enormous size of the dataset and the time it took to perform trials) to test our k-NN classifier and compared the output classifications for the tested 0.03% to these businesses' actual star ratings. The average difference between our k-NN's classifications and the actual businesses' ratings (i.e. the variance) was calculated, and we compared this variance to the variance of the ZeroR classifier. All the trials that we ran were better at classifying than the ZeroR. However, as we altered the Jaccard index weighting system and the value *k*, there were considerable differences in how much more or less variant the classifier was from the actual data.

Based on the proposed rating estimation of our k-NN classifier, an entrepreneur can decide whether or not it would be beneficial for him or her to expand into the market and add their business. Our solution does not give a definitive answer but rather demonstrates to a potential entrepreneur an educated estimate of how their business may be received by the demographic.

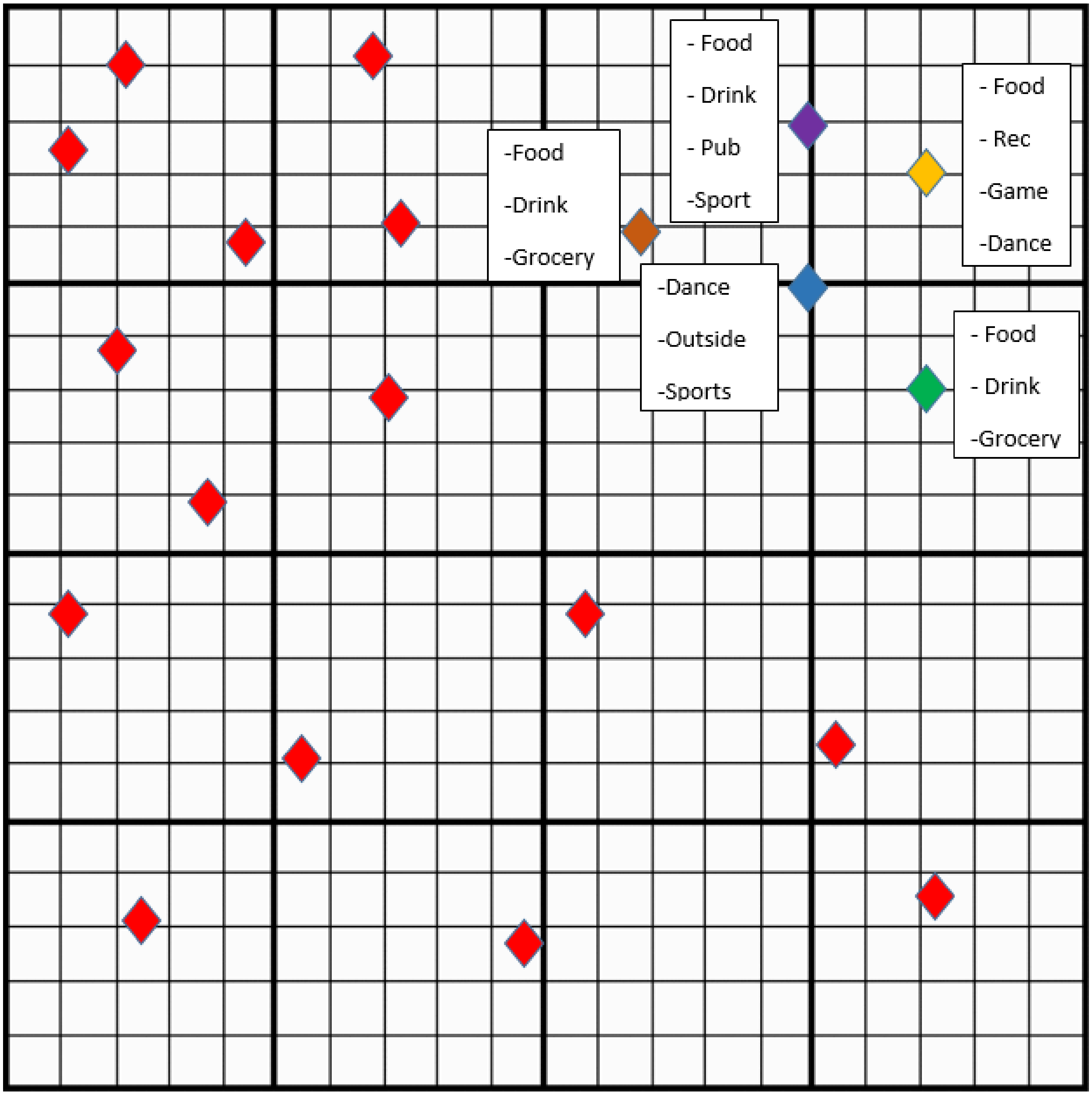

Figure 1

- The red diamonds indicate businesses that do not have any categories and have been placed in the 3D space and are only affected by latitude and longitude.
- The green diamond represents the placement of the proposed business based solely on latitude and longitude and has its categories listed next to it as well.
- The remaining diamonds are neighbors of the proposed business and have categories listed next to them. However they are currently only placed based on latitude and longitude

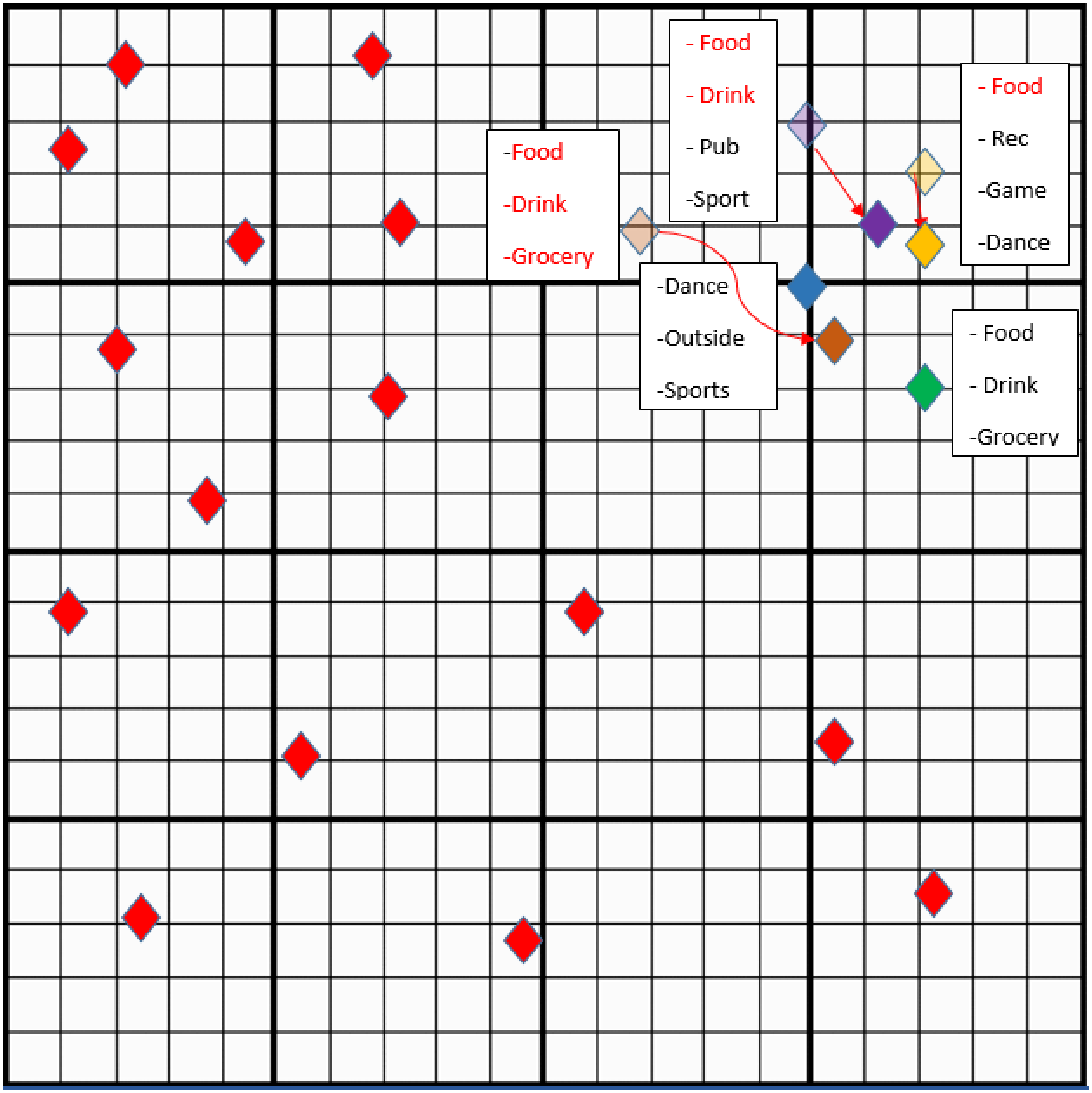

Figure 2

- This figure represents the concept of the Jaccard distance used in the nearest neighbor. The red components of the categories are ones found in both the proposed business and the existing businesses.
- The more similar the proposed business is to the actual business the closer that it gets pushed in the two dimensional space to the proposed business.
- The more transparent versions followed by arrows shows a generalization of where the business would be placed in the 2D space based on its similarity to the categories of the other business.
- Notice the fact that the blue diamond did not have any similar categories to the proposed business(green diamond) and the red diamonds were given no categories so they remain

in the same position in space after the weighting function is run. Its designed only to push similar businesses closer but it does not unrelated businesses further away.