# CALIFORNIA SCIENCE & ENGINEERING FAIR
# 2019 PROJECT SUMMARY

| | |
|---|---|
| **Name(s)**<br>**Jian Park** | **Project Number**<br>**S1411** |

**Project Title**

## DT-DBSCAN: Density-Based Spatial Clustering in Linear Expected Time Using Delaunay Triangulation

### Abstract

**Objectives**

DBSCAN is an unsupervised learning algorithm that connects locally dense regions of data to form data clusters. This algorithm is frequently used for medical, engineering, and financial applications, and it is also used to train machine learning algorithms. However, the fastest iteration of DBSCAN runs in quasi-linear time, which makes the algorithm unusable for larger sets of data. Therefore, the goal of this project was to develop a DBSCAN algorithm that has a linear expected run-time.

**Methods**

To determine locally dense regions, current DBSCAN iterations use R - trees and hashing methods to check for points within a locality, which takes O(log n) time for each point. DT-DBSCAN, or Delaunay triangulated DBSCAN, conducts the locality check by tree-searching along the edges of the Delaunay triangulation of the initial data set. Given that the maximum density of a cluster is bounded by a finite constant, I prove that this process takes provably constant time for each point. After classifying locally dense points, DT-DBSCAN connects locally dense points to form clusters in linear time by exploiting the linearity of the edges of a planar graph. After mathematically proving the run-time of DT-DBSCAN, an experiment was conducted to verify this claim. Procedurally generated data-sets were created with varying sizes, and the run-times between DT-DBSCAN and conventional DBSCAN algorithms were compared.

**Results**

After testing a total of 3,750 datasets with varying sizes, the run-time test showed that DT-DBSCAN heavily outperformed DBSCAN for data sets with more than 15,000 points. In addition, DT-DBSCAN displayed a linear growth in run-time as the data size increased, which affirmed the mathematical proofs on the run-time.

**Conclusions**

The proofs and the experiment both demonstrated the advantages of DT-DBSCAN over conventional DBSCAN iterations. DT-DBSCAN generates the exact same cluster results as DBSCAN, but has a linear expected run-time. This provides a unique solution for practically computing density-based clusters from larger sets of data.

**Summary Statement**

I developed DT-DBSCAN, an algorithm that classifies density-based data clusters faster than currently existing methods.

**Help Received**

I developed the run-time proofs and the DT-DBSCAN algorithm independently. My math teacher reviewed my project.