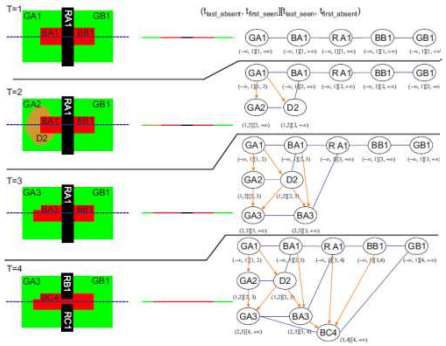
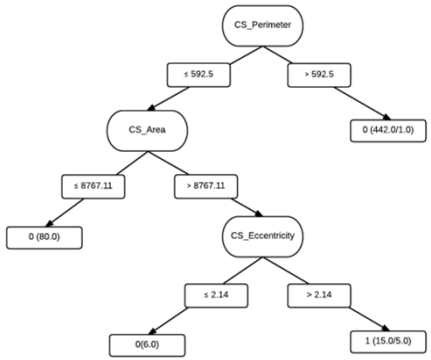
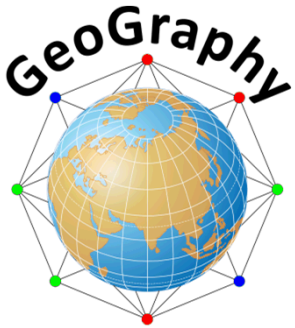
SAND2015-6160PE

Exceptional service in the national interest

Sandia National Laboratories







Data Mining-based Search Template Generation

Ryan Cooper
01464

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

GeoGraphy Background

1. Georegister data to have spatial location as the common semantic space.
2. Take the best information provided by each data source and build semantic hierarchies.

Example Data Sources

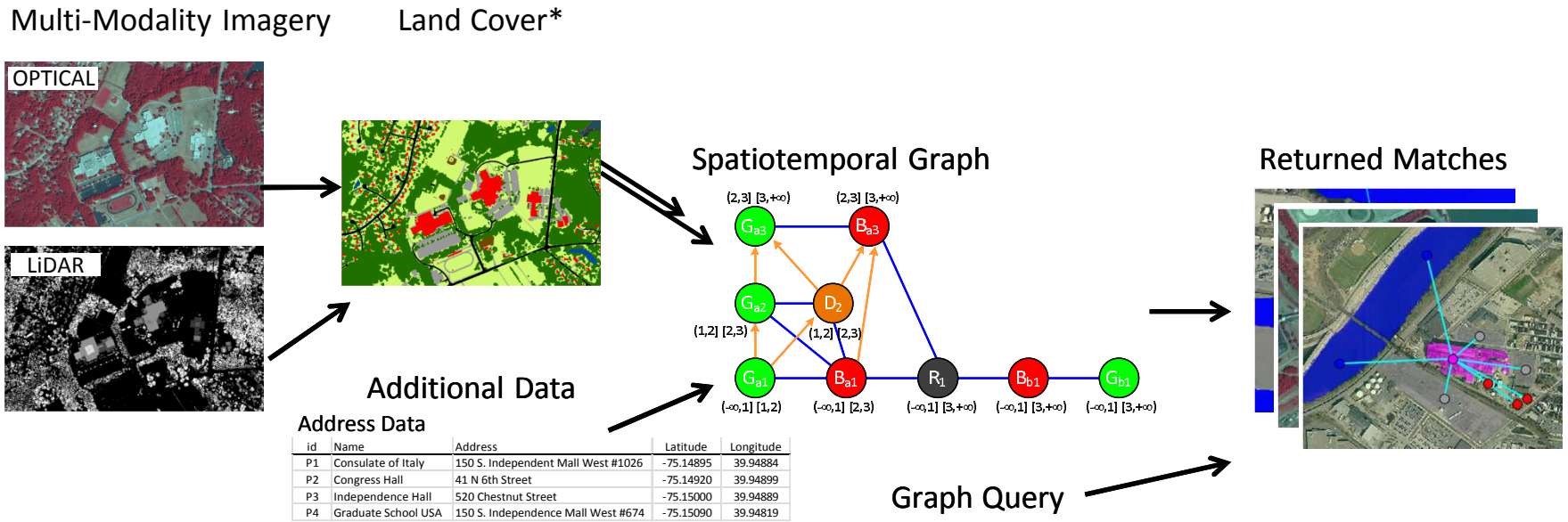
RGB+IR Optical Image

LiDAR Height Map

GIS Road Polygons

Land Cover Map

Location Database



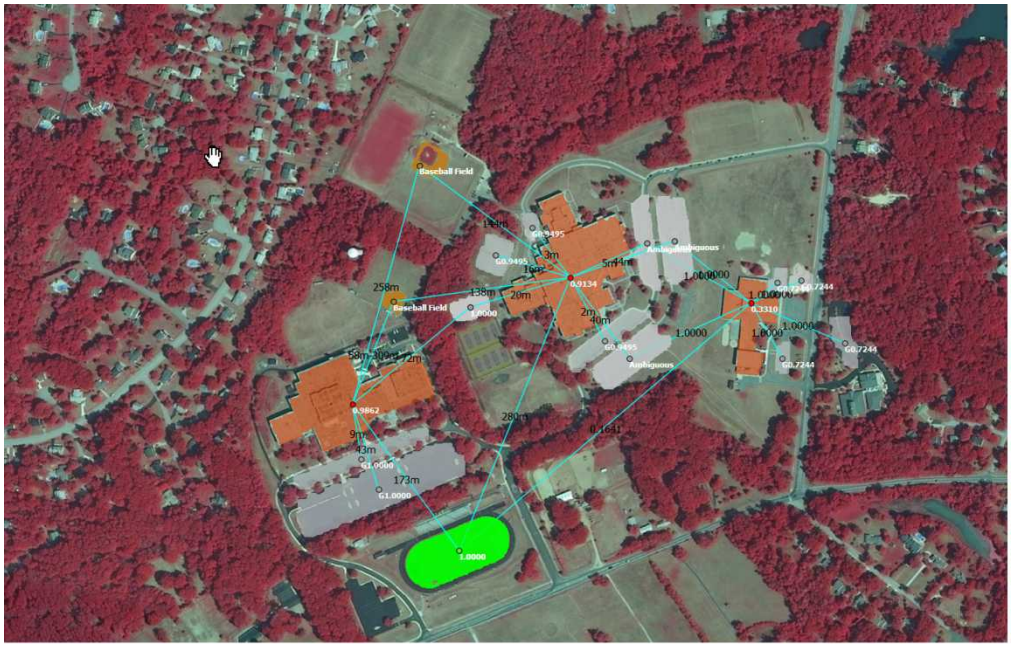
id	Name	Address	Latitude	Longitude
P1	Consulate of Italy	150 S. Independent Mall West #1026	-75.14895	39.94884
P2	Congress Hall	41 N 6th Street	-75.14920	39.94899
P3	Independence Hall	520 Chestnut Street	-75.15000	39.94889
P4	Graduate School USA	150 S. Independence Mall West #674	-75.15090	39.94819

* O'Neil-Dunne, et al, An object-based system for LiDAR data fusion and feature extraction, Geocarto (28), pp. 227–242, 2012.

Motivation

- I was presented with an opportunity to optimize the High School search for GeoSearch to reduce false positive results, and record my methodology for use on future optimization projects.

- This specific problem would require steps to be developed and processes to be created in order to differentiate the true positives from false positives.



Motivation

Sandia National Laboratories

- This specific search was designed with the implication of finding high schools, and obviously not all high school are identical, and this creates a variance in the true positives.
- This process of optimization is intended to define that variance, and be able to select high schools from a GeoSearch solely based on criteria.
- Originally, the High School search was able to identify 67 possible High Schools in Anne Arundel County, of which, 12 were true positives and 55 were false positives.

Process

Sandia National Laboratories

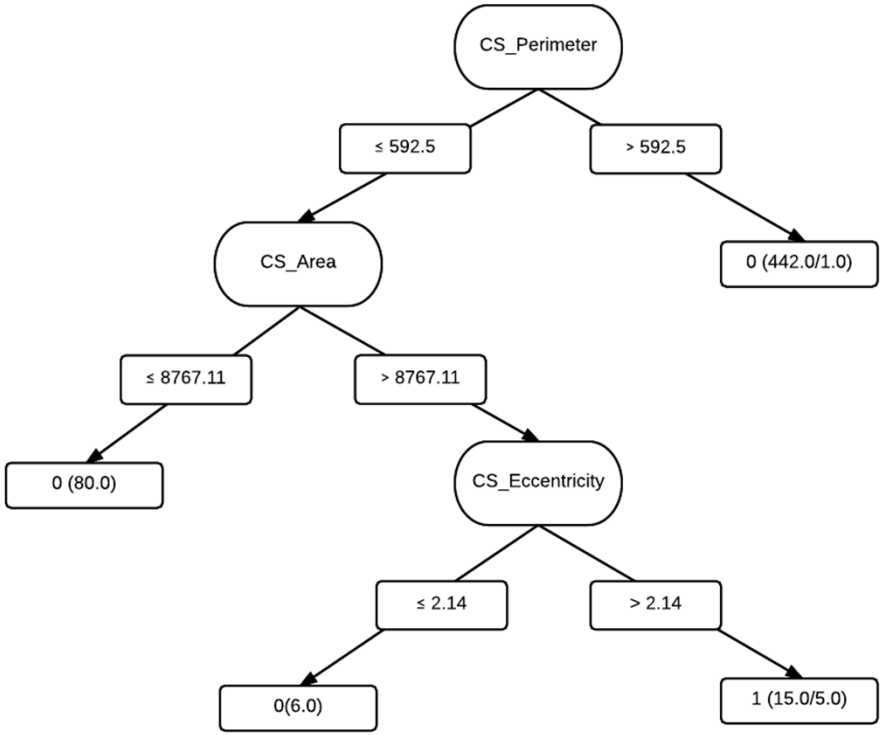
- The search was comprised of: a classroom building, football field, tennis court, parking lot, baseball field, and their relativity to each other, which I had to individually conduct a classification on each step in the search in order to refine the parameters.

- Data mining is an analytical process designed to explore data.

- Through data mining, I was able to conduct a classification algorithm called C4.5 tree to help identify factors leading to false positives, and restrict the search parameters.

Classification Algorithm

- I primarily used the C4.5 decision tree*, which is a classification algorithm that makes predictions to decide a target value based on various attribute values of an available data set.



* Quinlan, J. Ross. *C4. 5: programs for machine learning*. Elsevier, 2014.

Results

Sandia National Laboratories

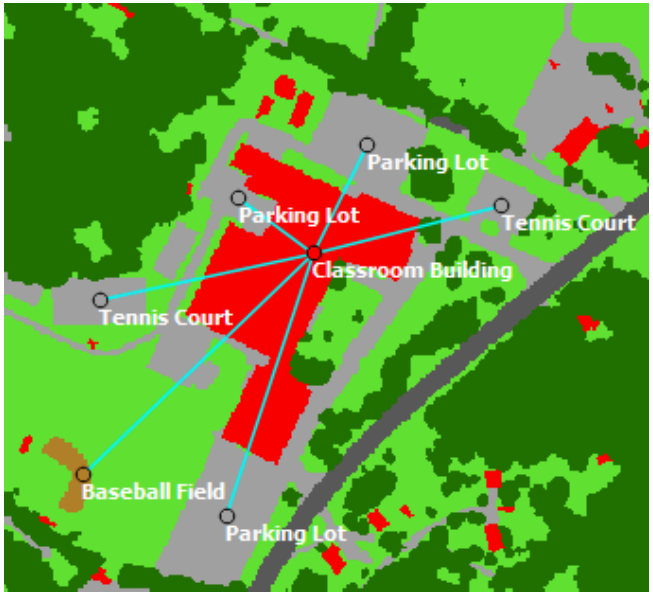
- After concluding the final optimization, I compiled all of the redefined search criteria and ran the search again.

- The redefined criteria resulted:
  - 27 possible High Schools with an optional football field
    - A reduction from 55 false positives to 15 false positives, or a 72.73% reduction in overall false positives.
    - Of which, 13 out of 15 false positives were other types of schools (i.e. Elementary School, Middle School or Private High School)
  - 16 possible High Schools with a required football field
    - On a broad spectrum, not all High Schools have a football field, however, all true positives in this example did.
  - No false negatives!

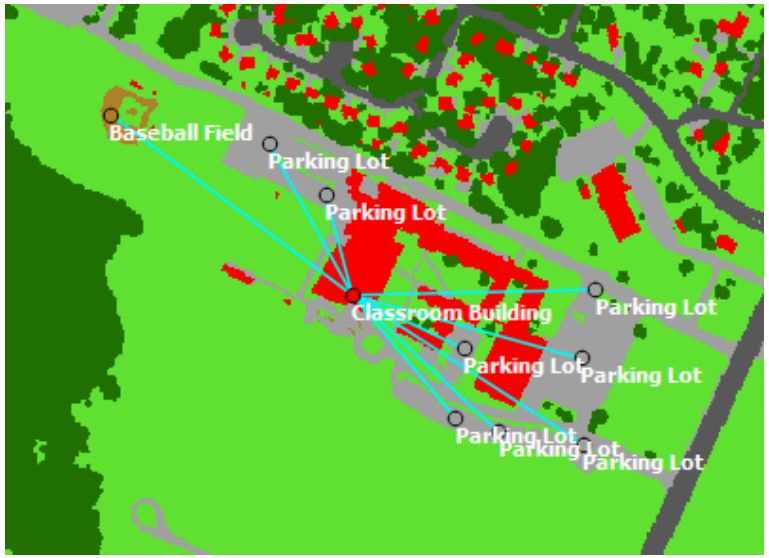
Analysis

- Attributes of the final 27 results were consistent with each other, in that they were schools, with the exception of 2 nodes.

- With the available data, it is almost impossible to differentiate between a middle school or a high school or between a public high school and a private school ; their features are so similar.



High School



Middle School

Takeaways

Sandia National Laboratories

- Experience with large amounts of data and database manipulation.
- Expanded knowledge of SQL (Structured Query Language).
- Practice of applications in Data Mining.
  - Specifically the J48 classification tree.
- Applications of C++ programming.
- Version Control: Git
- SCRUM Meetings
- Deadline management and workplace experience.