SAND2014-1829C

# PANTHER DA Uncertainty

# Approximate Pattern Matching under Uncertainty in Geospatial Semantic Graphs

Randy C. Brost
Cynthia A. Phillips
David G. Robinson
David J. Stracuzzi
Diane Myung-kyung Suh
(Sandia National Laboratories)

Mark Kaiser
Daniel Nordman
(Iowa State U.)
Alyson Wilson
(North Carolina State University)

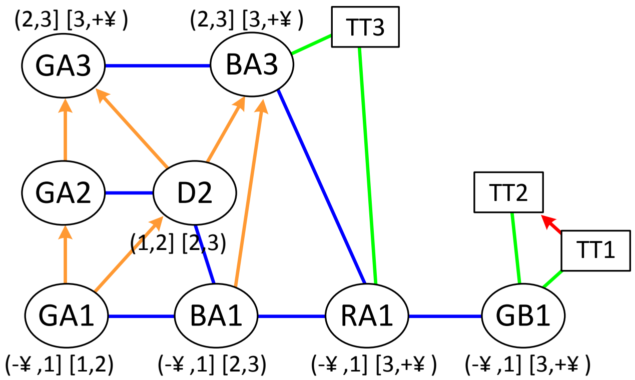
March 7, 2014

Sandia National Laboratories

*Exceptional service in the national interest*

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

# High-Level View, Overall Problem



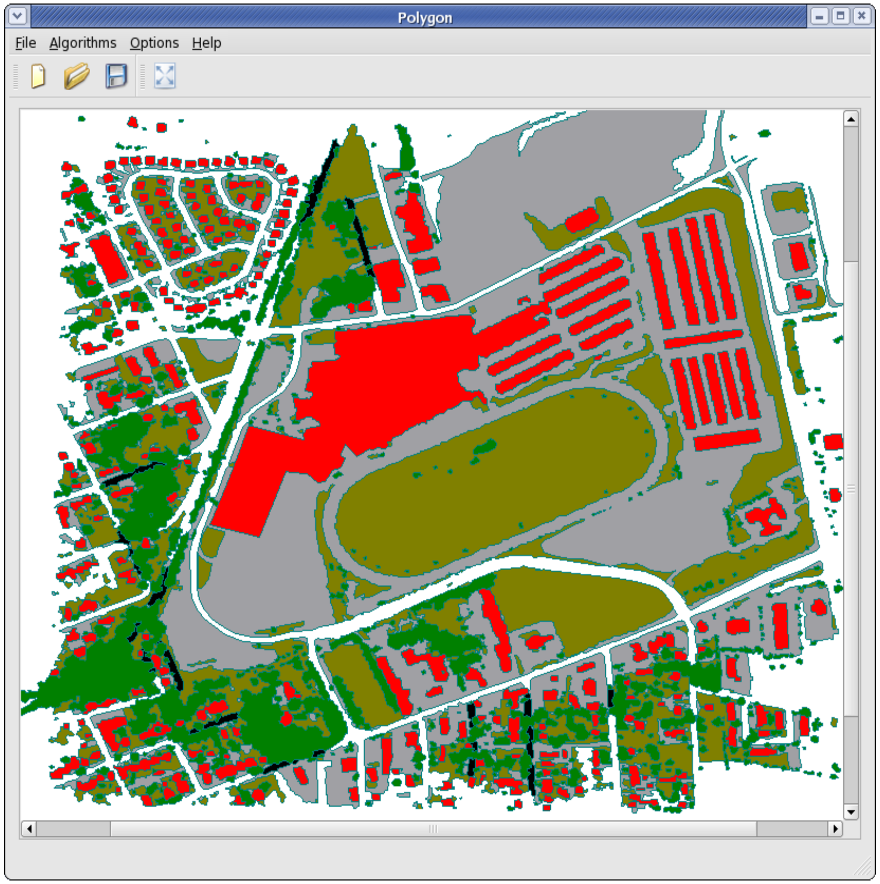
Human Analytics

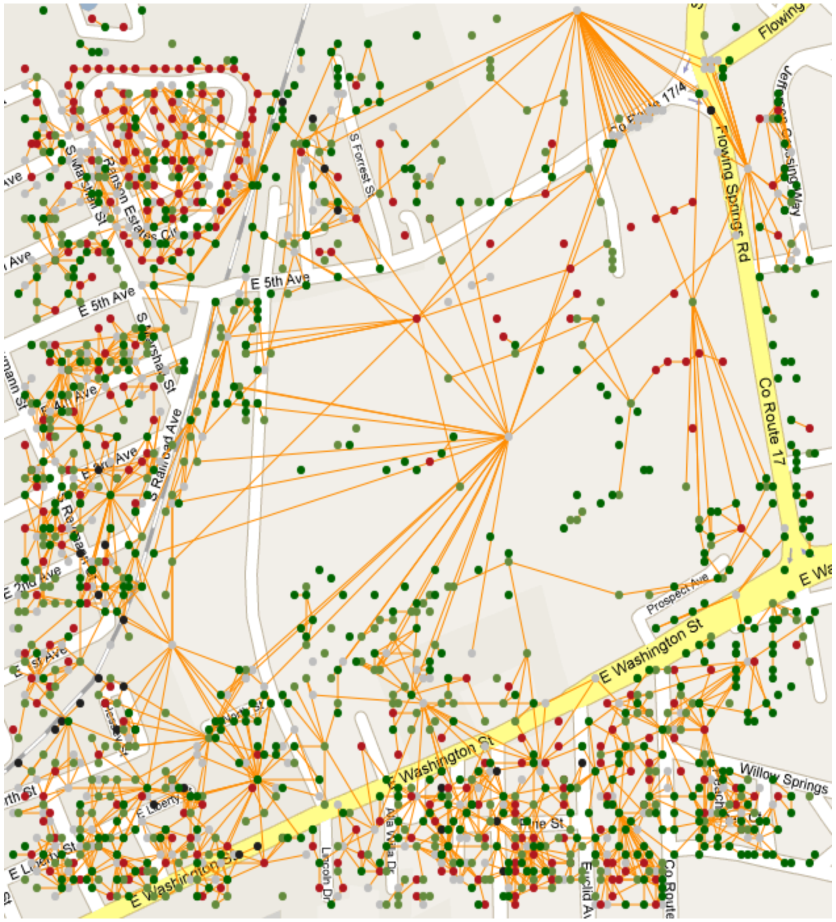
Remote Sensing → Sensor Exploitation → Spatiotemporal Graph →

Ranked list of matches
+
Match quality or probability
+
Confidence intervals

Query:
Graph topology
Node properties
Edge properties

# Geospatial Semantic Graphs

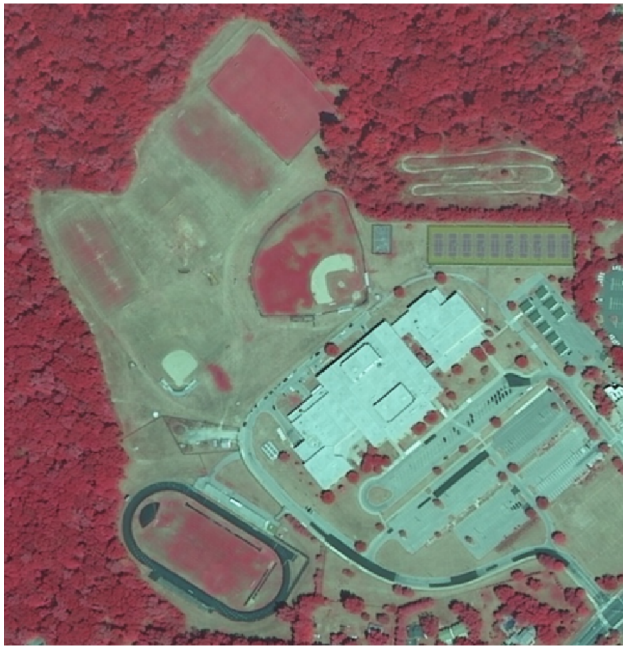


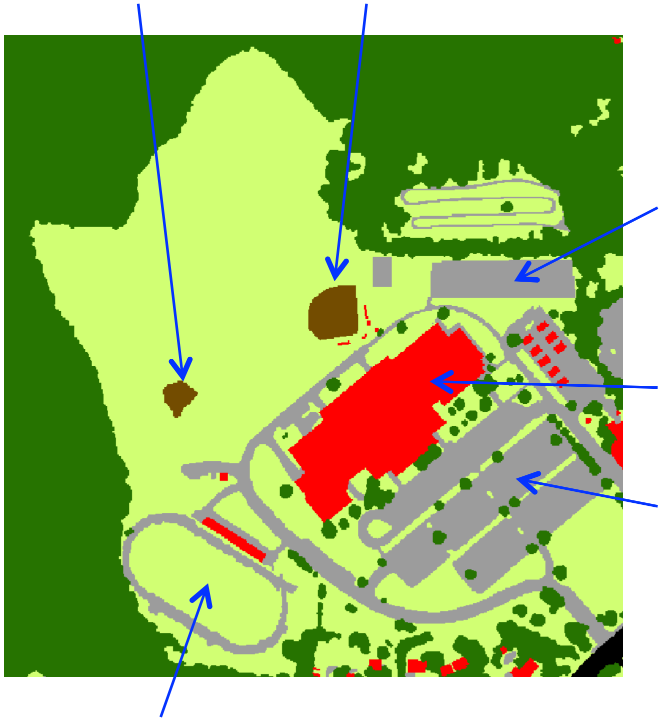


- Left: ground cover representation from an image
- Right: derived graph

# Example: Find US High School

- Ann Arundel County, MD





Baseball Field #2
d = 105 m

Baseball Field
d = 36 m

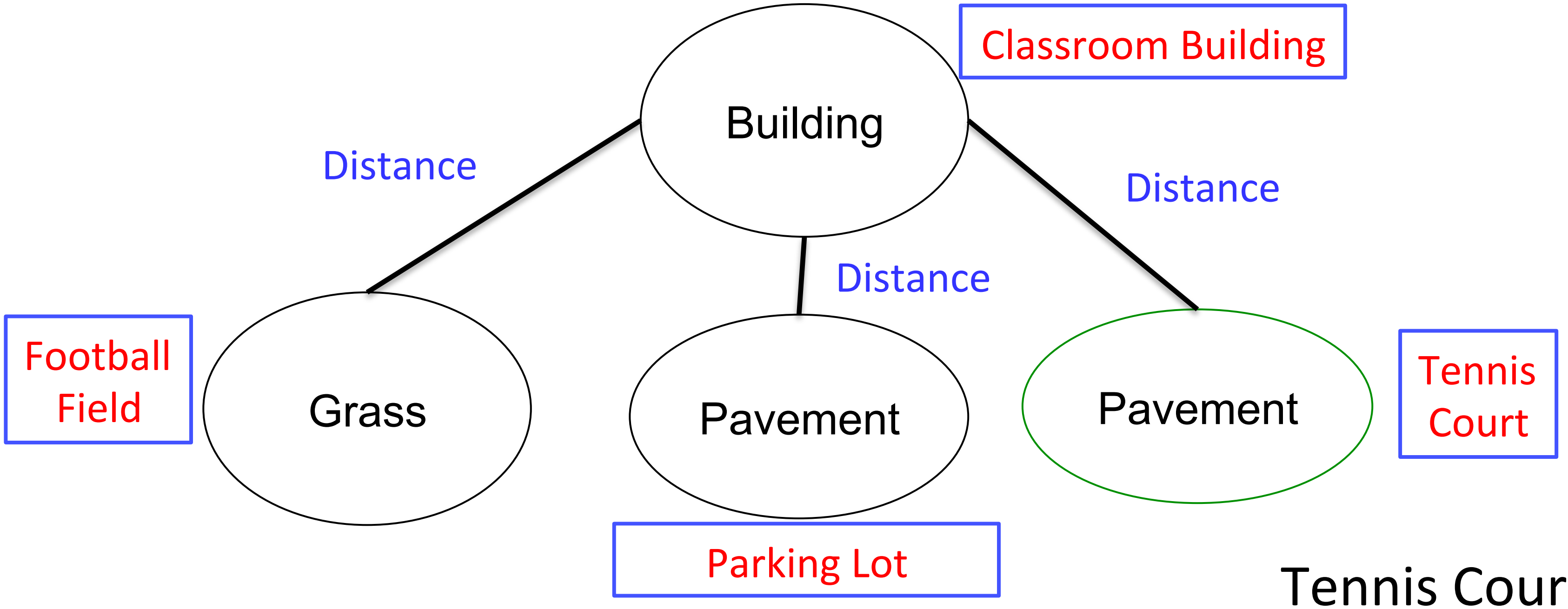
Tennis Courts
A = 4,940 m$^2$
w = 32 m
d = 24 m

Classroom Building
A = 16,040 m$^2$

Parking Lot
A = 36,450 m$^2$
d = 0 m

Football Field
A = 10,160 m$^2$
aspect = 2.37
d = 78 m

# Example Template: High School

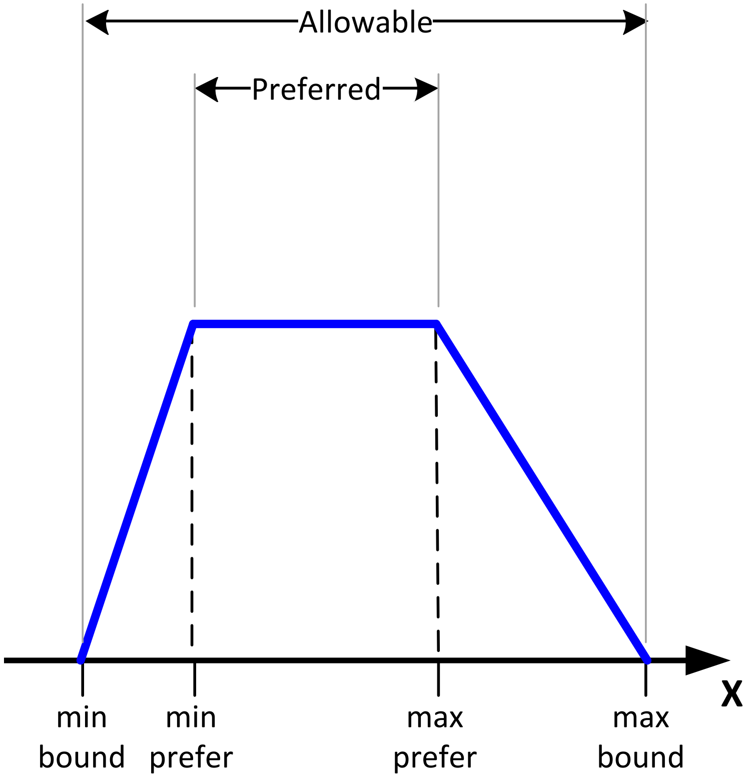


Tennis Court optional

- All nodes have: Land Cover, size.
- Football field has aspect ratio.
- Tennis court has width (assume one line of courts).

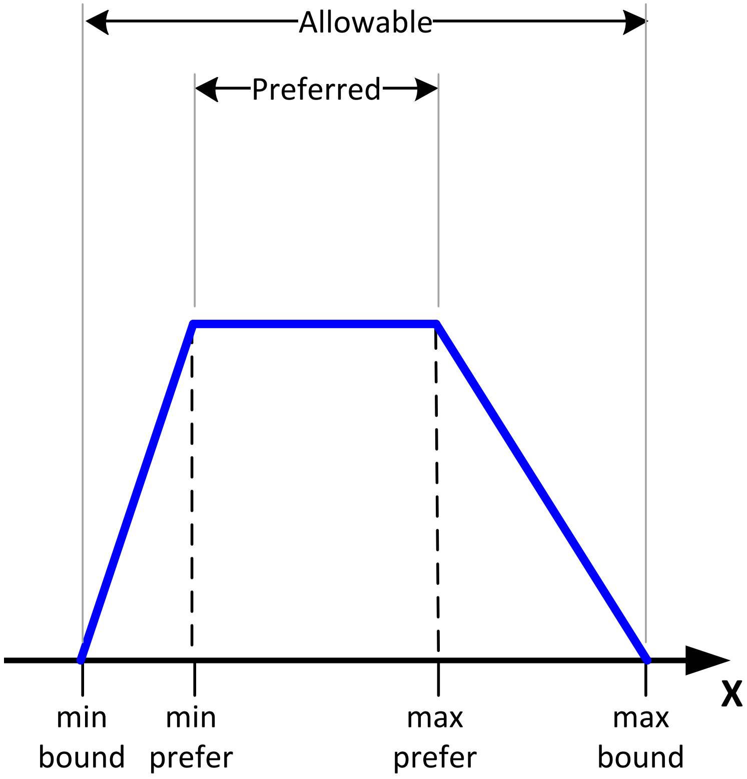
# Numerical Tolerances

- Classroom Building
  - Size : [5000,  6000, 25000, 30000]
- Parking Lot
  - Size: [9000, 10000, 1000000000, 1000001000]
  - Distance to Classroom building : [ 0, 0.1, 100, 101]
- Football Field
  - Size: [8000, 8200, 10500, 10700]
  - Aspect Ratio: [1.7, 1.8, 3.2,3.3]
  - Distance : [0, 1, 370, 600]
- Tennis Court
  - Size: [3700, 3800, 5500, 5600]
  - Width: [20, 30, 40, 50]
  - Distance : [0, 0.1, 200, 201]



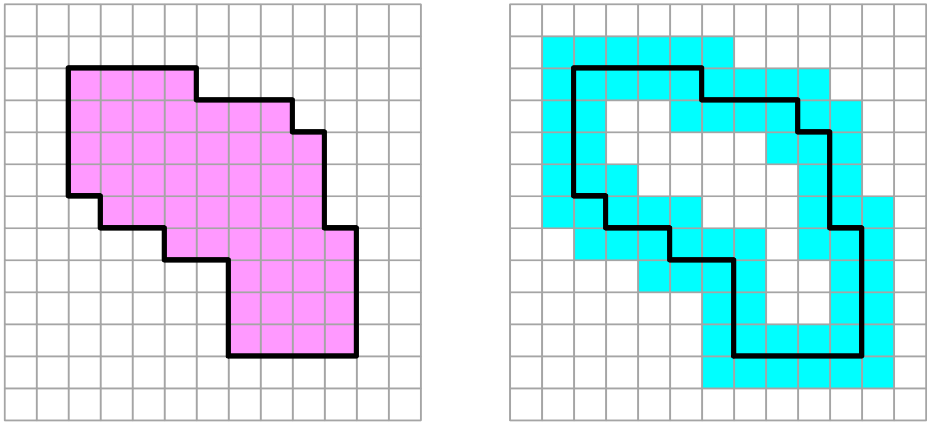
# Numerical Tolerances

- Allows for variation among high schools
- Appropriate for any numerical value
- Discretization



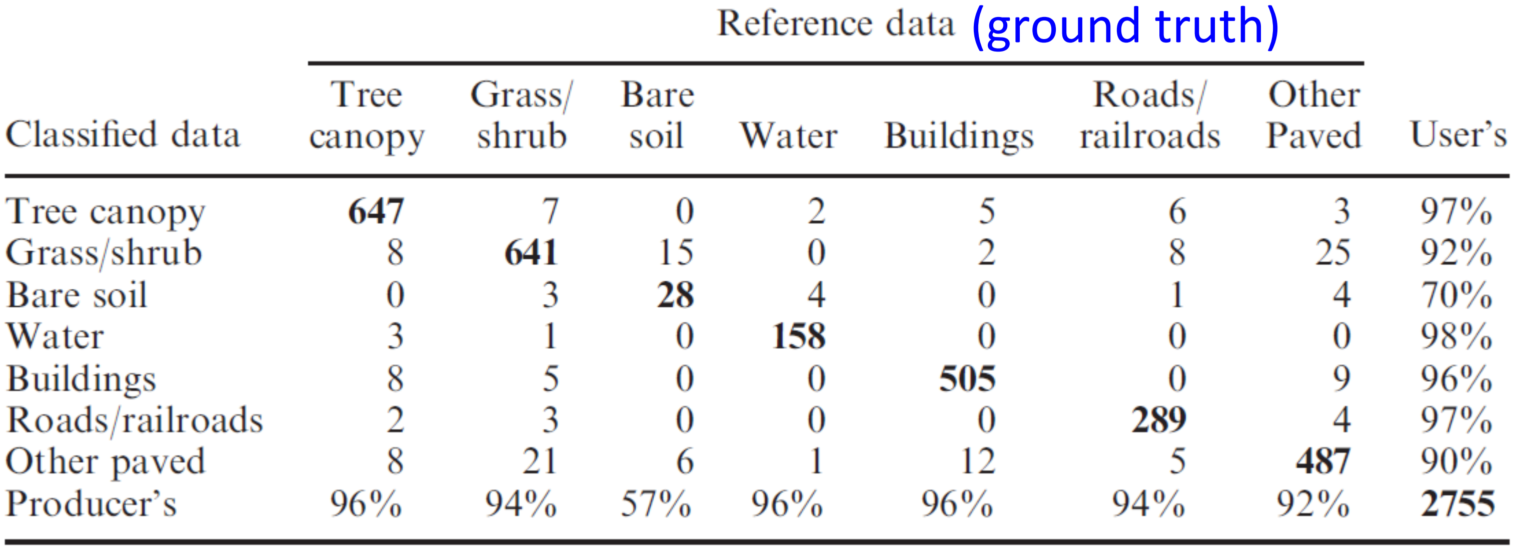
# Sources of Uncertainty

- (Physics) limitations of sensors
- Algorithms
- Example: Boundary uncertainty
  - Leads to uncertainty in area, aspect ratio, distances between objects



# Landcover Uncertainty

Error matrix for a similar Philadelphia data set:



Reference data (ground truth)

Algorithm output

| Classified data | Tree canopy | Grass/ shrub | Bare soil | Water | Buildings | Roads/ railroads | Other Paved | User's |
|---|---|---|---|---|---|---|---|---|
| Tree canopy | **647** | 7 | 0 | 2 | 5 | 6 | 3 | 97% |
| Grass/shrub | 8 | **641** | 15 | 0 | 2 | 8 | 25 | 92% |
| Bare soil | 0 | 3 | **28** | 4 | 0 | 1 | 4 | 70% |
| Water | 3 | 1 | 0 | **158** | 0 | 0 | 0 | 98% |
| Buildings | 8 | 5 | 0 | 0 | **505** | 0 | 9 | 96% |
| Roads/railroads | 2 | 3 | 0 | 0 | 0 | **289** | 4 | 97% |
| Other paved | 8 | 21 | 6 | 1 | 12 | 5 | **487** | 90% |
| Producer's | 96% | 94% | 57% | 96% | 96% | 94% | 92% | **2755** |

From O'Neil-Dunne, et al, "An Object-Based System for LiDAR Data Fusion and Feature Extraction," Geocarto International, 2012.

* See R. Congalton and K. Green. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, 2nd Edition.* CRC Press/Taylor and Francis, 2009.

# Evaluating Candidates

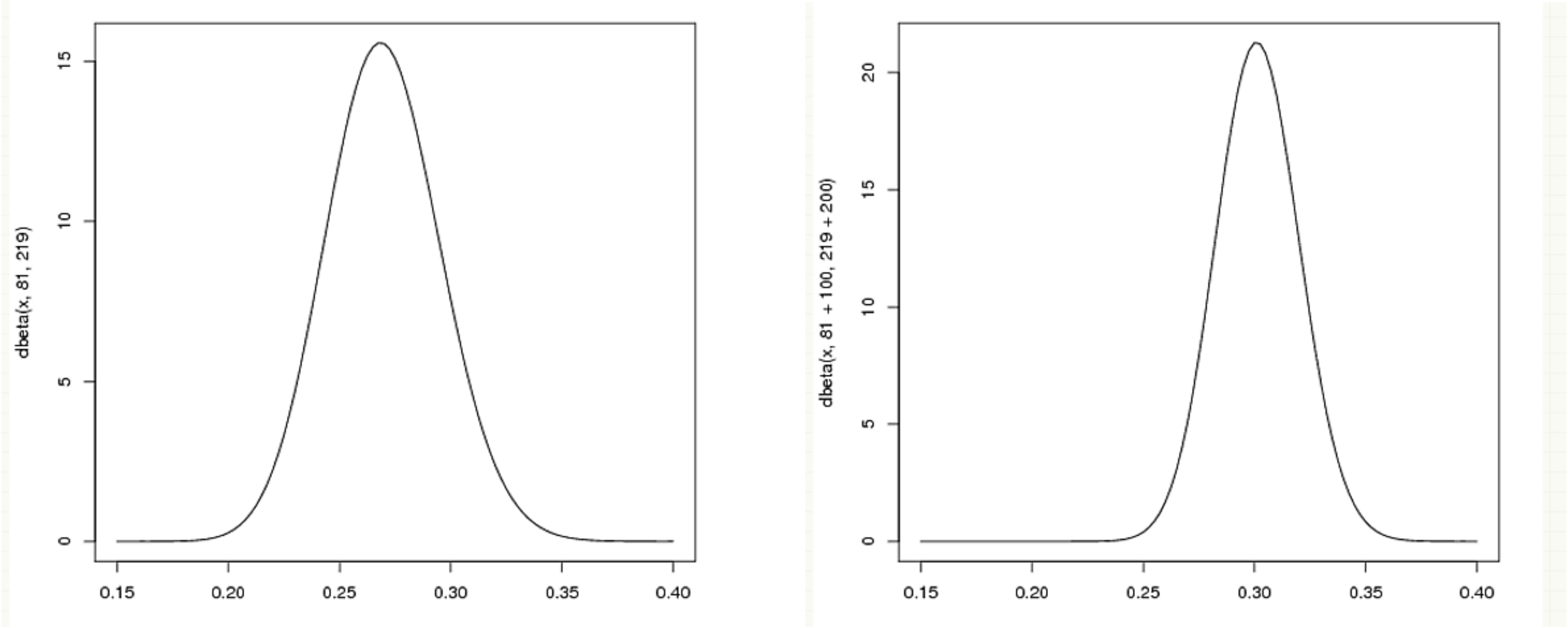
- Goal: Given set of candidates, approximately evaluate, rank
- Methods that give probability of a match (e.g. high school)
  1. Elicitation, regression to a beta distribution
  2. Bayes Network
  3. Hierarchical Naïve Bayes
- Methods that score match/distance to the template
  4. Our quality score metric
  5. Earth Mover's Distance
- Score does not comment on semantic meaning, only the quality of match to what the analyst asked for.

# 1. Elicitation-based method

- Assume training data is rare: “unicorn farm.”
- Elicit information from experts one attribute at a time.
  - Order attributes by importance.
  - Elicit trapezoid values: preferred and allowable ranges.
  - Estimate $p_1$, probability of HS if attribute one is in preferred range.
  - Estimate $p_i$, additional probability of a HS if all of the first $i$-1 attributes in preferred range.
  - If $\sum_{i=1}^{n} p_i < 1$ , remaining probability represents other hypotheses
- This elicitation is difficult for the “unicorn farm” expert

# Beta Distribution

- Beta(α,β) has mean α/(α+β). Larger values of α,β tighter.



David Robinson, from stats.stackexchange.com

- Distribution of probabilities
- Represent prior knowledge. Posterior knowledge still beta.

# Compute a beta distribution

- Assume the mean at each step is linear function of attributes:

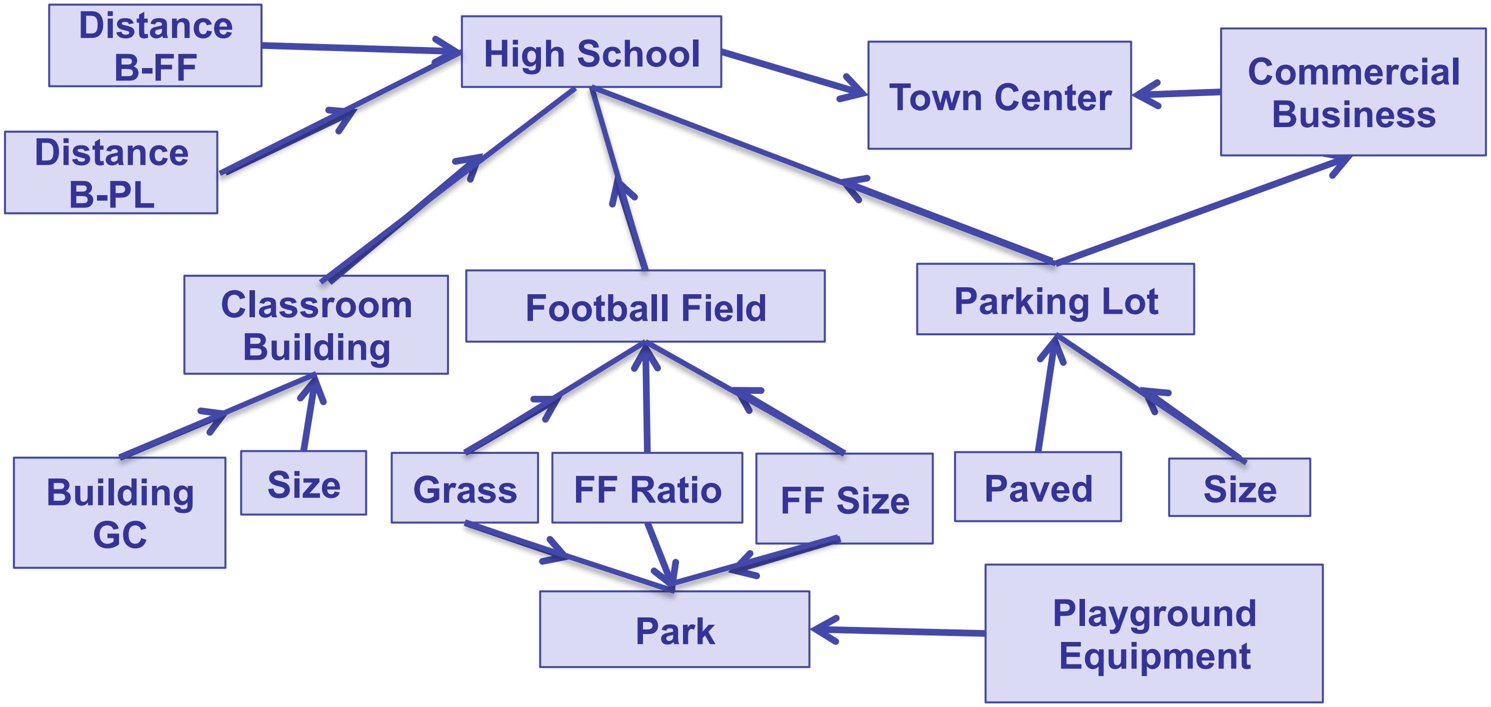
$$\sum_{j=1}^{i}(\gamma_{j,0}+\gamma_{j,1}z_j)$$

where $\gamma$ are regression values, $z_j$ represent deviation from preferred value range for the candidate match.

- Infer $\gamma_{j0}$ from preferred values where $z_j$ =0.
- Infer $\gamma_{j1}$ from the instances.
- For confidence intervals: confidence from experts gives a measure of variance. Can compute weight around mean.
- Order matters: Don't get full credit if the earlier values upon which $p_i$ is conditioned are not perfect.

# 2. Bayes Nets

- Directed Acyclic Graph
  - Nodes are variables.
  - Edges represent conditional dependence.
  - Non-edges represent conditional independence.

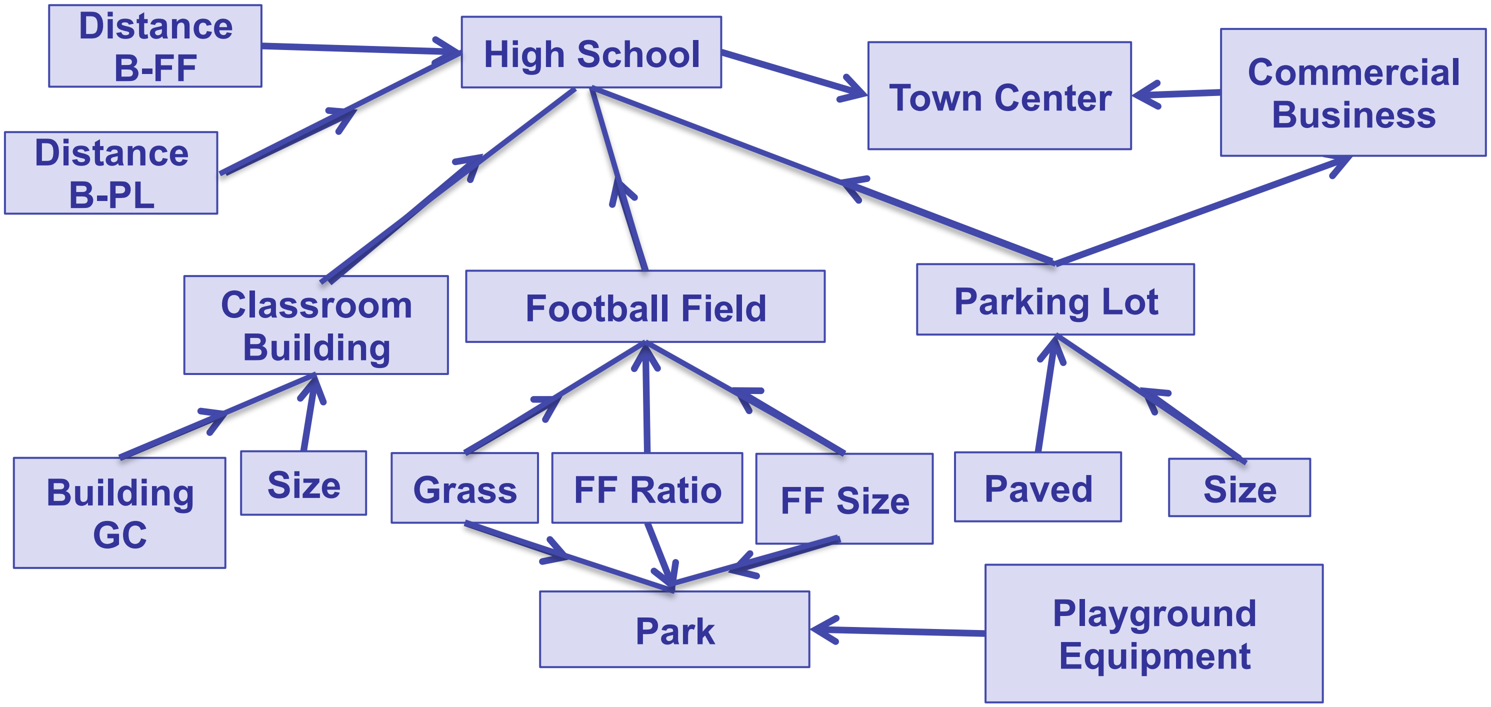


# Bayes Nets

- Requires more data than we are ever likely to have.

| Conditional Prob of HS | Football Field | Classroom Building | Parking Lot | Distance (BP) | Distance (BF) |
|---|---|---|---|---|---|
| .95 | 1 | 1 | 1 | 1 | 1 |
| .20 | 1 | 0 | 1 | 0 | 0 |
| .30 | 0 | 1 | 1 | 1 | 0 |
| .70 | 1 | 1 | 0 | 0 | 1 |

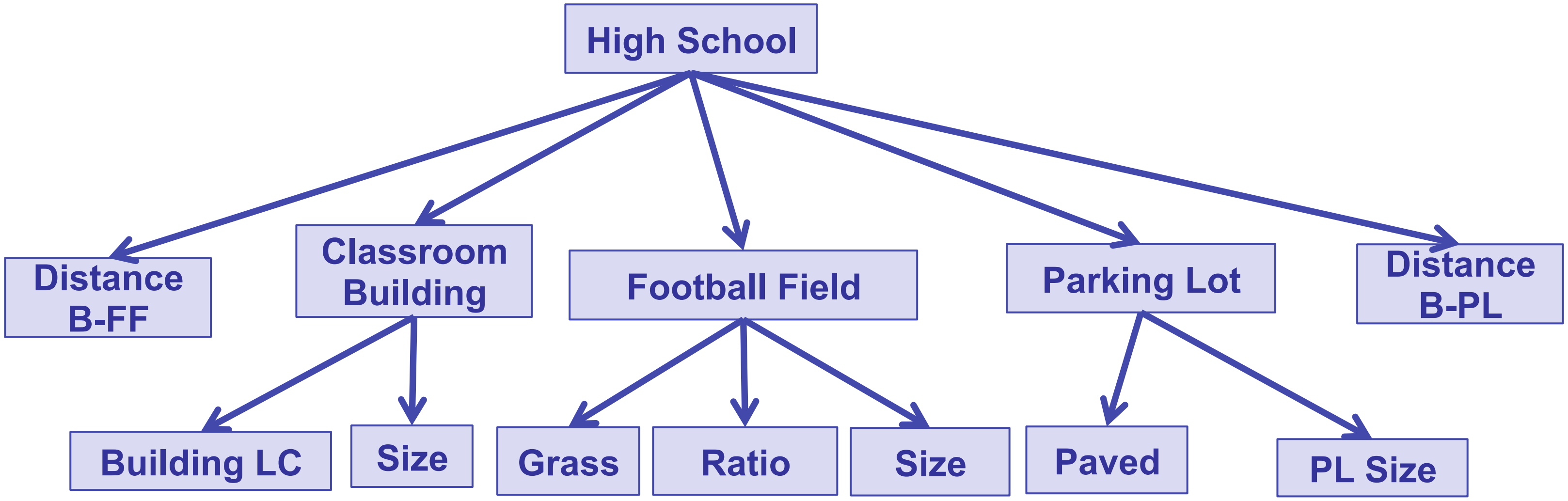
Etc…



# 3. Hierarchical Naïve Bayes

- We create ensembles from features and other ensembles
- Assume attribute values are conditionally independent given the object class.

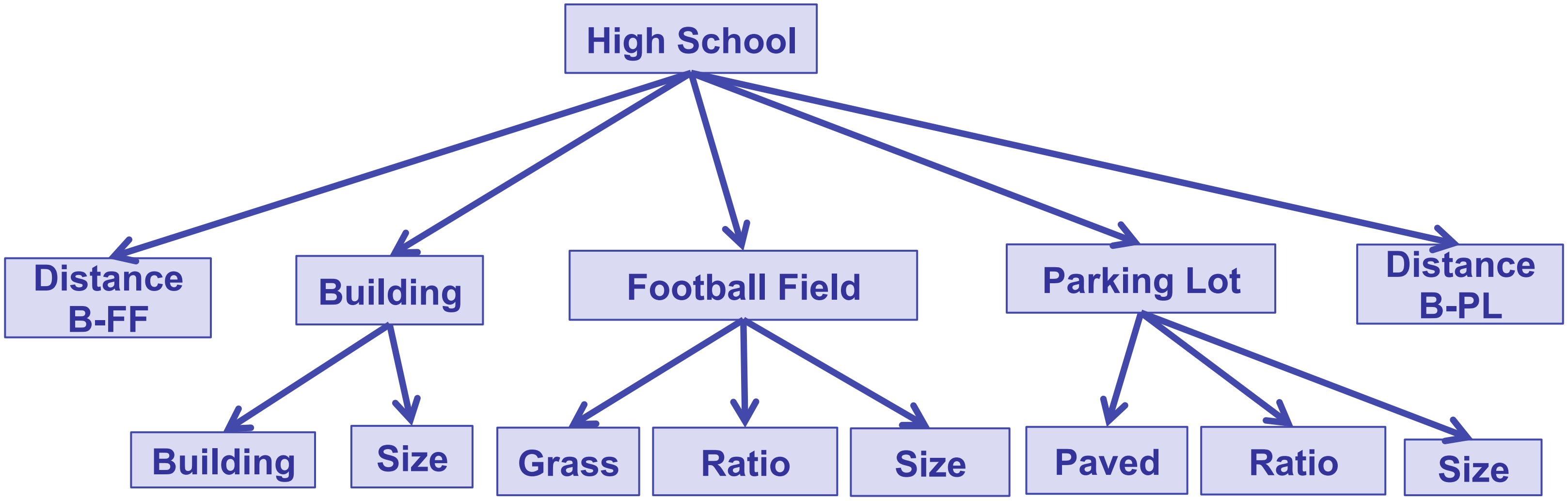
$$P(H|F,B,L,d_1,d_2) = \frac{P(H)P(F|HS)P(B|HS)P(L|HS)P(d_1|HS)P(d_2|HS)}{P(F,B,L,d_1,d_2)}$$



# Hierarchical Naïve Bayes

- Requires training data.
- Challenge defining what to count/how to estimate conditional probabilities.

$$P(H|F,B,L,d_1,d_2) = \frac{P(H)P(F|HS)P(B|HS)P(L|HS)P(d_1|HS)P(d_2|HS)}{P(F,B,L,d_1,d_2)}$$



# 4. Scoring Match Quality

- Scoring individual pieces $q_i$
  - Ground cover: score using confusion matrix
  - Area, distance, aspect ratio:
    - Compute a range of values based on sensor limitations
    - Score is percentage of the range within preferred/acceptable range
  - Other options for more complex attributes

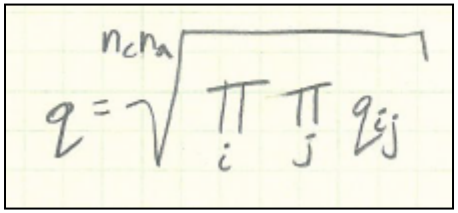
- How to combine the individual scores?

# Desired Score Properties

1. Intuitive magnitude ⇒ range (0,1]; good scores e.g. ≈0.9.
2. Increasing number of attributes, components ⇒ stable score.
3. Low-probability required components ⇒ very low score.
4. Low-probability optional components ⇒ stable score.
5. Optional components present ⇒ increase score.
6. Smooth:
   - Smooth degradation with noise, uncertainty.
   - Avoid binary decisions causing discontinuous behavior.
7. Monotonic response to component scores.
8. Not sensitive to order of attributes or components.
9. Differentiate required vs. optional components.
10. Adjustable contribution of required vs. optional components.
11. Adjustable weights of individual components.
12. Understand relationship between attribute and component contributions.
    - Avoid instabilities.

# Geometric Mean Approach

Normalizing joint probability product:



$$q = \sqrt[n_c n_a]{\prod_i \prod_j q_{ij}}$$

where:

| | |
|---|---|
| $q$ | Overall quality score. |
| $n_c$ | Number of template components. |
| $n_a$ | Number of attributes per component. |
| $i$ | Component index. |
| $j$ | Attribute index. |
| $q_{ij}$ | Quality of attribute j for component i. |

- Intuitive: Good scores remain near 1.0. → Supports goal #1.
- Provides stable aggregate score, regardless of n. → Supports goal #2.
- Low-probability components drive a small score. → Supports goal #3.
- Smooth, monotonic. → Supports goals #6 and #7.
- Insensitive to order. → Supports goal #8.
- Understandable attribute/component relationship. → Supports goal #12.

# Extending to Include Weighting

Attribute equation:

$$q_i = \sqrt[n_{ia}]{\prod_j q_{ij}}$$

where:

| | |
|---|---|
| $q_i$ | Quality score for component i. |
| $n_{ia}$ | Number of attributes for component i. |
| j | Attribute index. |
| $q_{ij}$ | Quality of attribute j for component i.<br>$q_{ij} \in (0, 1]$; note that $q_{ij} \neq 0$. |

Component equation:

$$q = \sqrt[\Sigma k_i]{\prod_i q_i^{k_i}}$$

where:

| | |
|---|---|
| q | Overall quality score. |
| $k_i$ | Weighting exponent for component i. |
| i | Component index. |
| $q_i$ | Quality score for component i .<br>$q_i \in (0, 1]$ follows from constraint on $q_{ij}$. |

The vector of weights $[k_1 \;\; k_2 \;\; k_3 \;\; \ldots \;\; k_{nk}]$ can be arbitrary.

# What If All k Values are Equal?

If all $k_i = 1$:

$$q = \left[ \prod_i q_i \right]^{\frac{1}{n}}$$

Quality q reduces to simply the geometric mean.

Compare to arithmetic mean:

$$q_{arith} = \left[ \sum_i q_i \right] \cdot \frac{1}{n}$$

# Partitioning Required and Optional

General equation:

$$q = \left[\prod_i q_i^{k_i}\right]^{\frac{1}{\Sigma k_i}}$$

- $k_i$ values are an arbitrary vector.

Reducing free parameters:

- Partition components into "required" and "optional" subsets.
- For each required component, select $k_i = k$.
- For each optional component, select $k_i = 1$.
- Resulting quality score:

$$q = \left\{\left[\prod_{req} q_i^{k}\right]\left[\prod_{opt} q_i\right]\right\}^{\frac{1}{n_{req}\cdot k + n_{opt}}}$$

- This allows quality score to have a different response to required vs. optional components, controlled by a single adjustment parameter k.

# Calculating k

Method:

- Choose small “zero” quality score $q_{zero}$ corresponding to clear absence of a required feature (for example, select $q_{zero} \equiv 0.0001$).* Assure all $q_i \geq q_{zero}$.

- Choose desired quality score $q_{pr}$ desired when all required components have perfect quality (for example, select $\boxed{q_{pr} \equiv 0.75}$). ← Free adjustment parameter.

- Compute k:

$$k = \left(\frac{n_{opt}}{n_{req}}\right)\left[\left(\frac{\log(q_{none})}{\log(q_{pr})}\right) - 1\right]$$

where:

| | |
|---|---|
| $n_{req}$ | Number of required components. |
| $n_{opt}$ | Number of optional components. |

* The special value $q_{zero} > 0$ is required:
  (a) to prevent a missing optional component from driving the overall score to zero, and
  (b) to prevent $\log(q_{zero})$ from blowing up when computing k.

If the same $q_{zero}$ value is used when computing k and setting minimum quality scores, then the overall q value is not sensitive to the choice of $q_{zero}$, when all required components are present.

# Advantages

Advantages of partitioned geometric mean:

- Intuitive: Good scores remain near 1.0. → Supports goal #1.
- Provides stable aggregate score, regardless of n. → Supports goal #2.
- Low-probability required components drive a small score. → Supports goal #3.
- Low-probability optional components leave score stable. → Supports goal #4.
- High-probability optional components increase score. → Supports goal #5.
- Smooth, monotonic. → Supports goals #6 and #7.
- Insensitive to order. → Supports goal #8.
- Adjustable required/optional differentiation. → Supports goals #9 and #10.
- Individual component weights possible. → Supports goal #11.
- Understandable attribute/component relationship. → Supports goal #12.

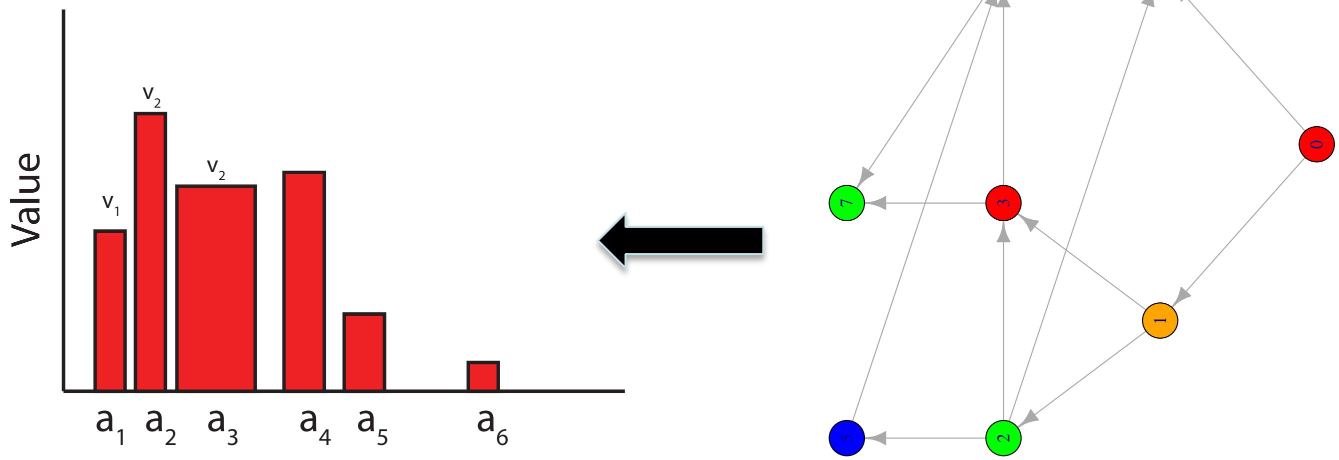
# Confidence intervals for Scores

Bootstrapping:

- Generate many random tuples of values from the initial data based on uncertainty
  - Draw from the confusion matrix for ground cover
  - Draw from intervals for numerical values
- Compute score for each instance
- Compute desired percentile (e.g. 5% to 95%)

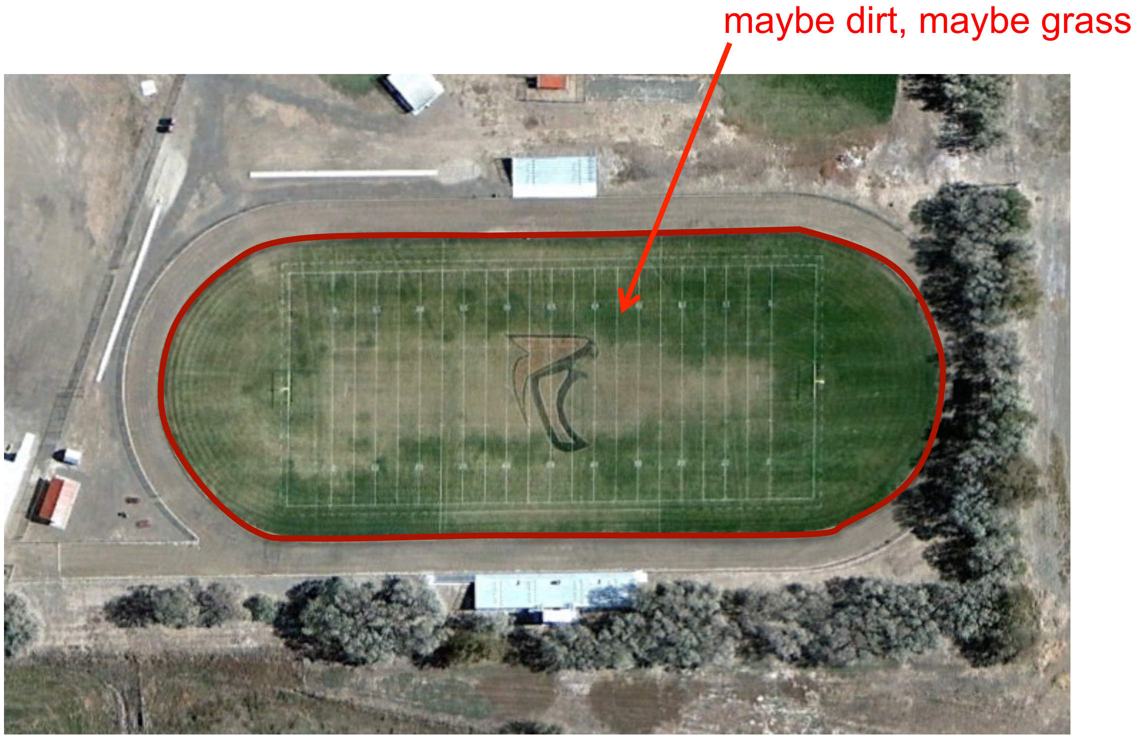
# 5. Bertillonage

- Compute a distance/similarity between attributed graphs
  - In general does not require knowledge of a matching between nodes
  - In general does not require graphs to be the same size
- Create a signature (distribution)
  - Depends on the application
  - For large graphs (10,000 nodes), local topological measures from each node does very well



# Ground Cover Revisited

- Consider ground cover as a distribution
  - Ground cover program finds a boundary and counts pixel/patch types



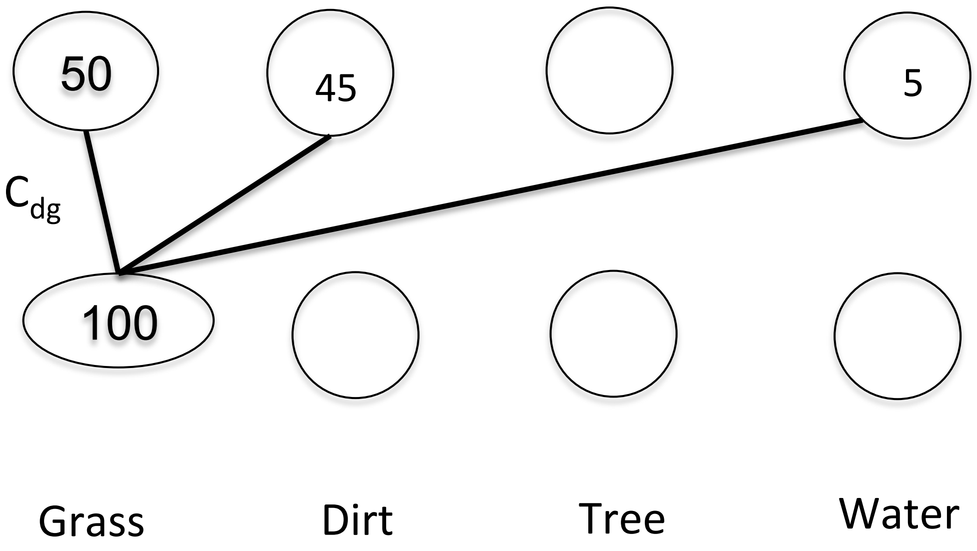
A football field near Capulin, Colorado.

*Image from Google Earth.*

# Earth Mover's Distance

- For computing distances between distributions.
  - Also has meaning for unnormalized distributions (e.g. compare size).
  - A true metric if the ground distance is a metric. (Mallow's)
  - Any node can move weight to any other node.

$C_{dg}$ = cost to move from dirt to grass

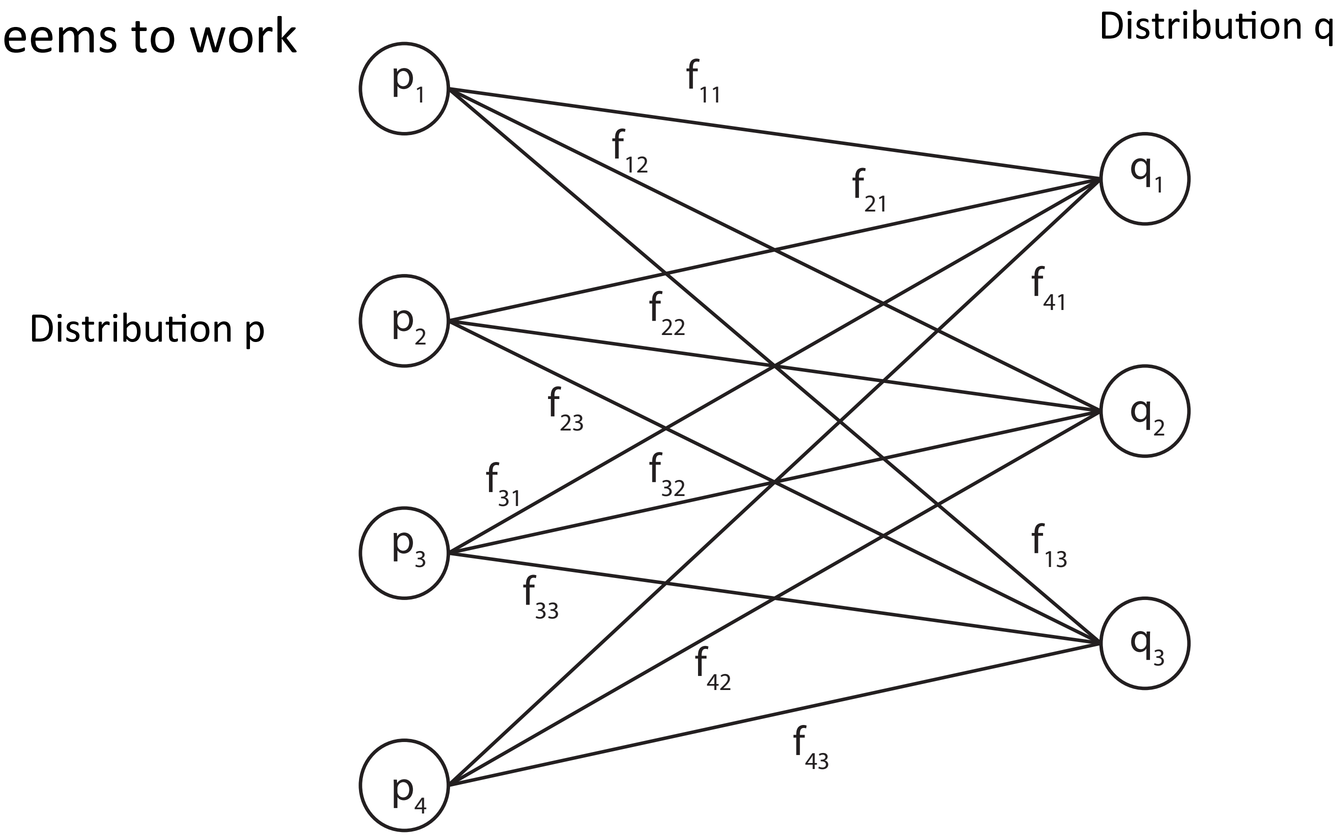


Candidate (%)

Target (%)

# EMD-based signatures

- Combine ground cover with area, distances, etc
- Challenge to normalize and combine
  - Simple seems to work

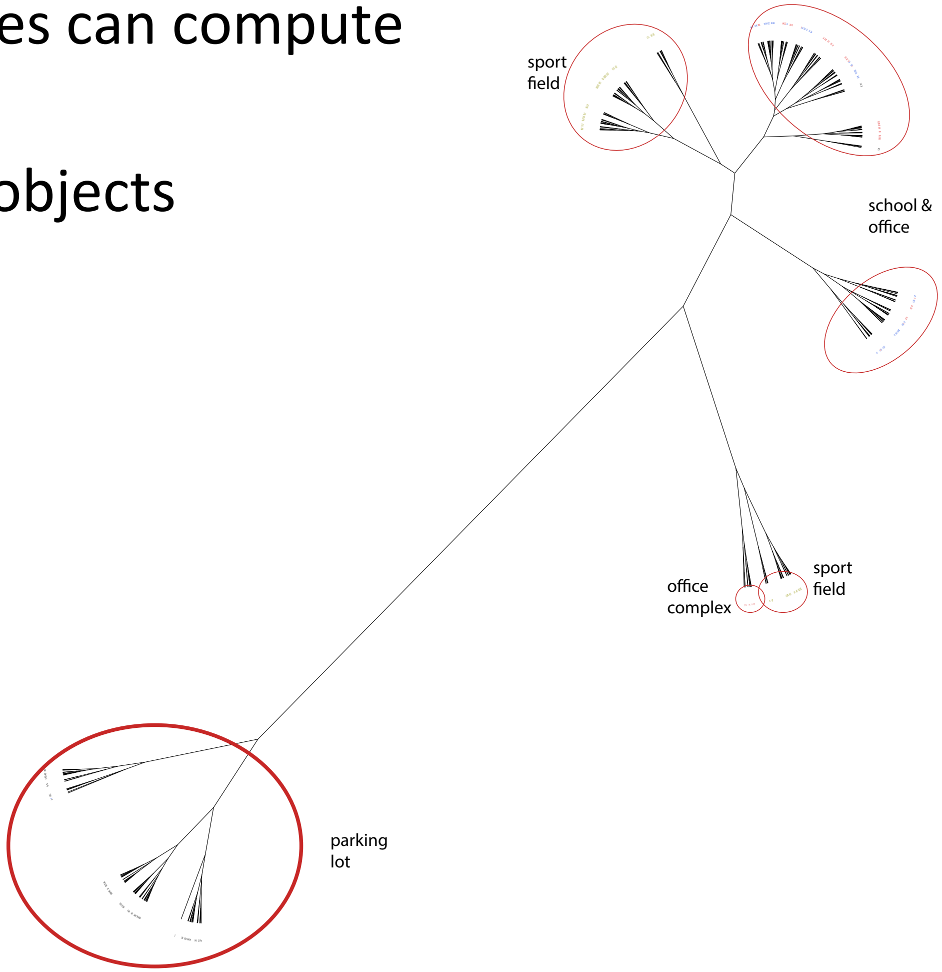


# What To Do with Distances

- Libraries of instances
  - Random instances based on uncertainty or tolerances
    - Confusion matrix, sensor measurement errors, etc
  - Instances we’ve seen before (e.g. interesting instances)
- Compare to library
  - What is this example most closely related to?
  - Are examples changing over time?
  - Have we seen something like this before?
- In high school example, comparison to template ranks candidates

# Dendrograms

- Given distances can compute dendrograms
- Clustering of objects



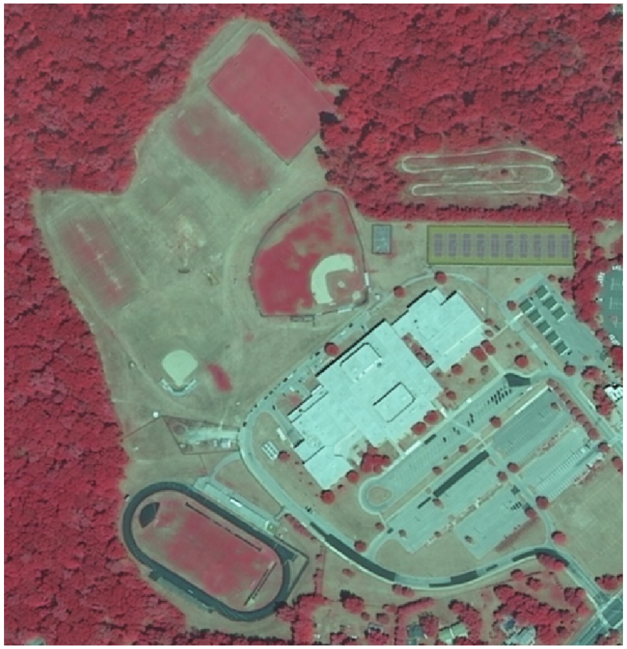
# Preliminary Experiments

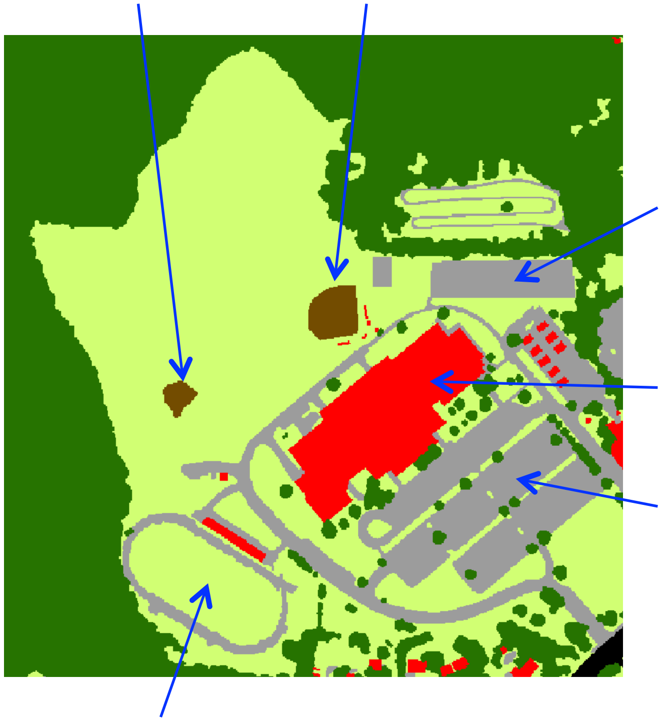
- Consider 5 high-school-(like) objects from Ann Arundel County
- Apply all but full Bayes Nets
- Many details still evolving

# Annapolis High School

Characteristics:

- Location (38.9746°, -76.5655°)
- Original NA-22 match: true
- Perfect – all components are present.





Baseball Field #2
d = 105 m

Baseball Field
d = 36 m

Tennis Courts
A = 4,940 m$^2$
w = 32 m
d = 24 m

Classroom Building
A = 16,040 m$^2$

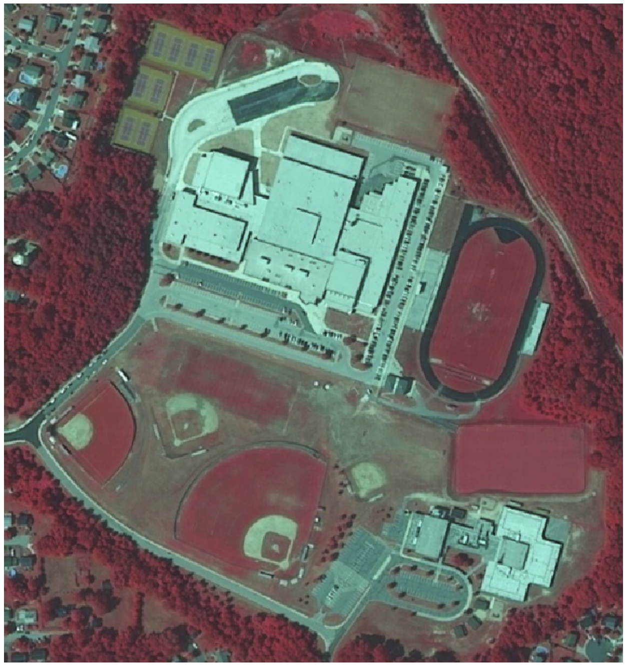
Parking Lot
A = 36,450 m$^2$
d = 0 m

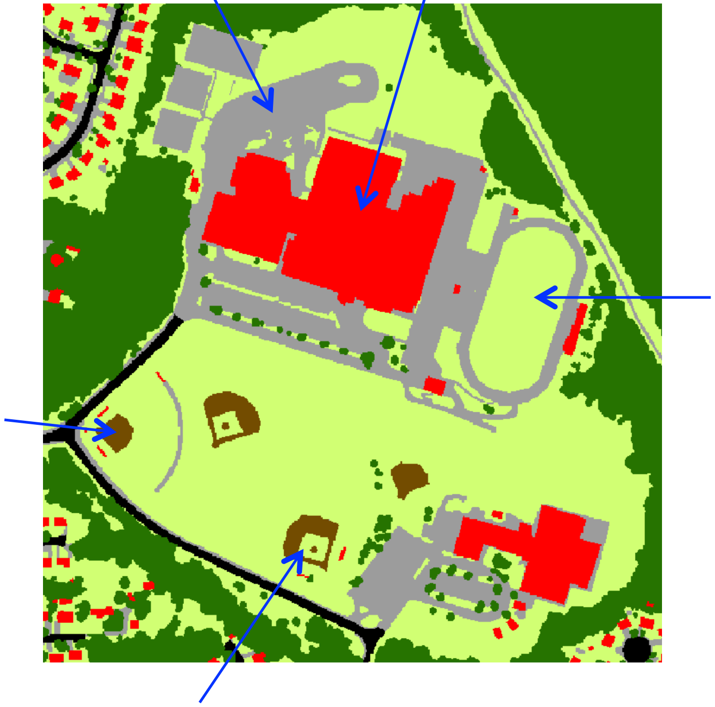
Football Field
A = 10,160 m$^2$
aspect = 2.37
d = 78 m

# North County High School

<u>Characteristics:</u>

- Location (39.1930 °, -76.6370°)
- Original NA-22 match: true
- No tennis courts.
  (Broken into too-small segments)





Parking Lot
A = 29,050 m$^2$
d = 0 m

Classroom Building
A = 22,550 m$^2$

Football Field
A = 9,950 m$^2$
aspect = 2.26
d = 59 m

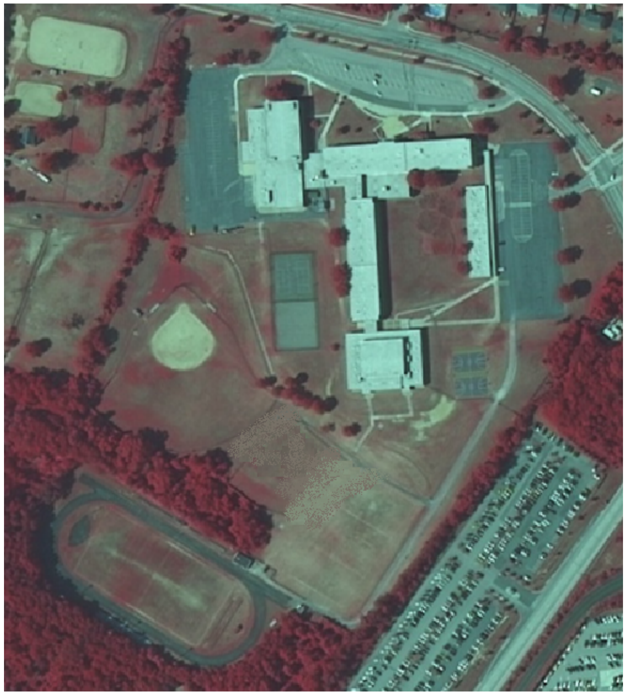
Baseball Field #2
d = 178 m

Baseball Field
d = 197 m

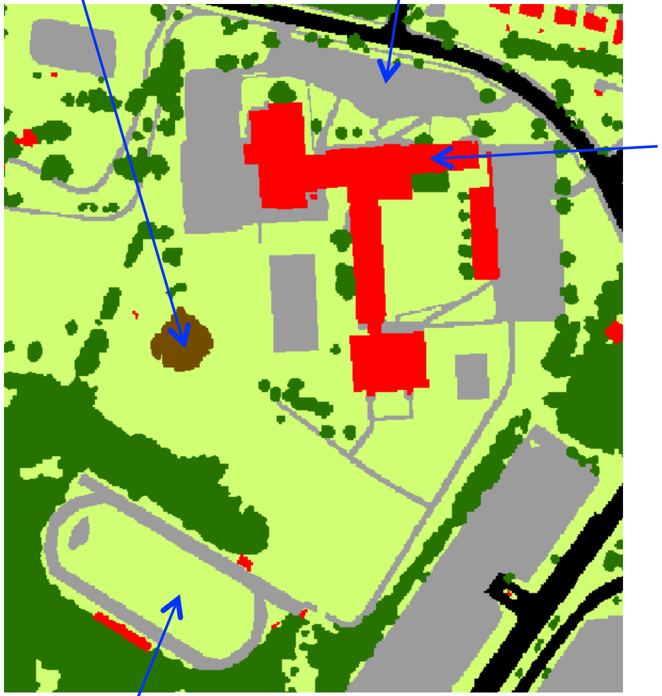
# Lindale Middle School

Characteristics:

- Location (39.1965°, -76.6620°)
- Original NA-22 match: true
- No tennis courts.
- No baseball field #2.



Note: Manually edited to support example.



Baseball Field
d = 95 m

Parking Lot
A = 13,730 m$^2$
d = 0 m

Classroom Building
A = 12,840 m$^2$

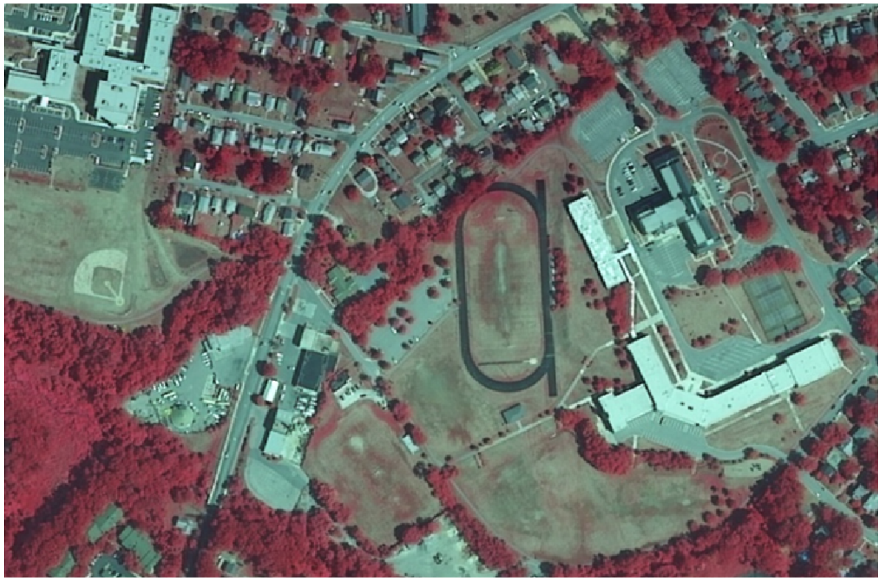
Football Field
A = 8,455 m$^2$
aspect = 2.84
d = 253 m

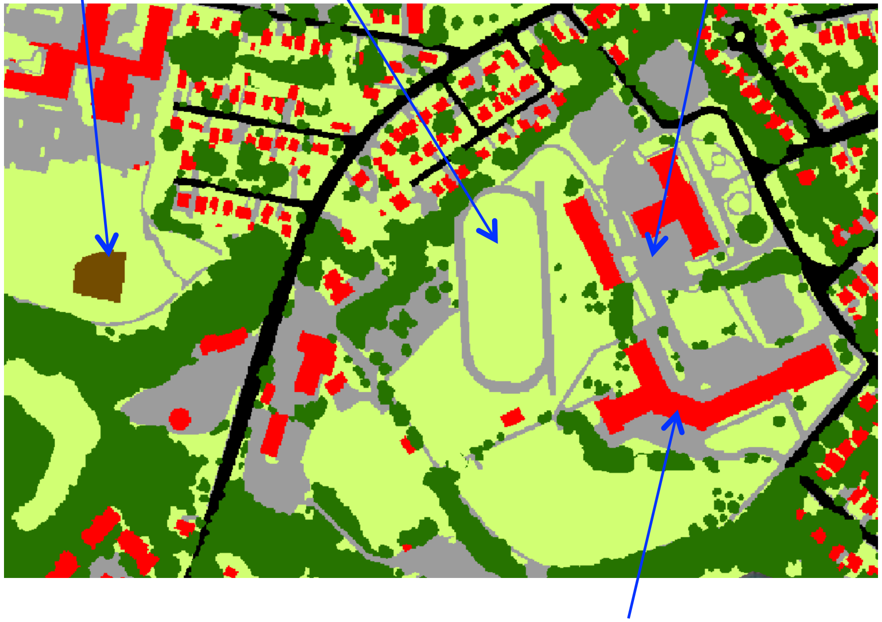
Note: Manually edited to support example.

# Wiley Bates Middle School

Characteristics:

- Location (38.9723°, -76.5053°)
- Original NA-22 match: MARGINAL
- No tennis courts.
- No baseball field #2.
- Building area marginal.
- Baseball field distance marginal.





Baseball Field
d = 411 m

Football Field
A = 9,147 m$^2$
aspect = 2.56
d = 60 m

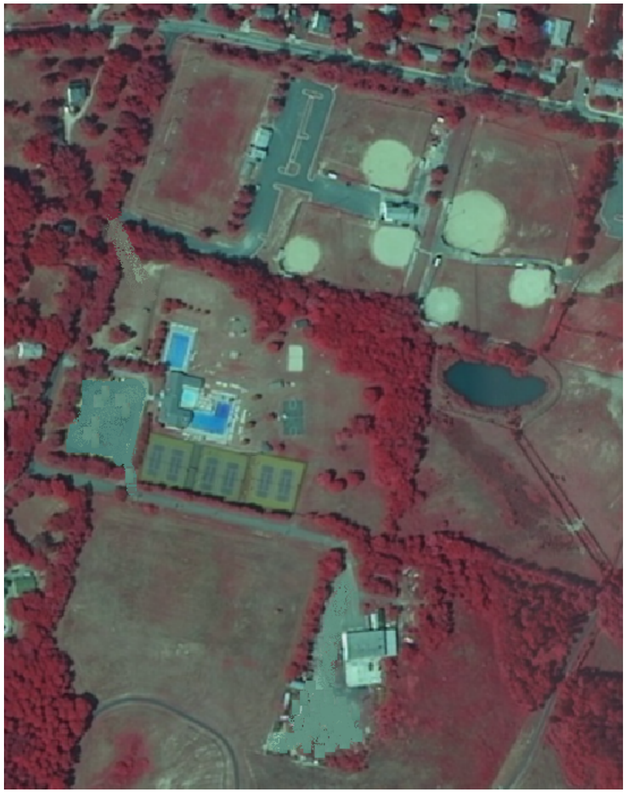
Parking Lot
A = 14,120 m$^2$
d = 0 m

Classroom Building
A = 5,700 m$^2$

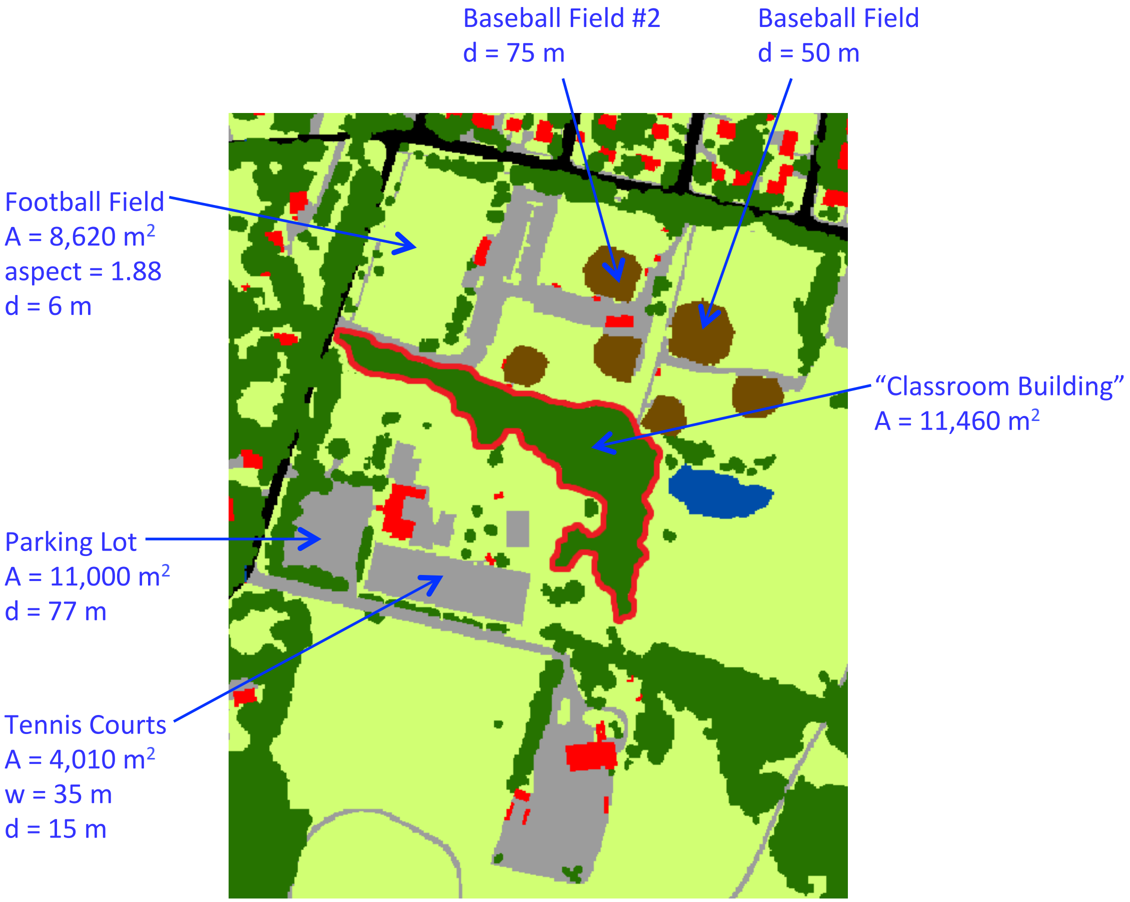
# Andover Park

Characteristics:

- Location (39.1968°, -76.6680°)
- Original NA-22 match: false
- No building (tree misclassification).
- All other components present, including optional.



Note: Manually edited to support example.



Baseball Field #2
d = 75 m

Baseball Field
d = 50 m

Football Field
A = 8,620 m$^2$
aspect = 1.88
d = 6 m

"Classroom Building"
A = 11,460 m$^2$

Parking Lot
A = 11,000 m$^2$
d = 77 m

Tennis Courts
A = 4,010 m$^2$
w = 35 m
d = 15 m

Note: Manually edited to support example.

# ISU Method

| | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| **ISU Method (phi = 0.005)** | **0.9500000**<br>[0.9157908, 0.9756950] | **0.8726458**<br>[0.8230521, 0.9151563] | **0.8726458**<br>[0.8230521, 0.9151563] | **0.10941637**<br>[0.07002800, 0.1562255] | **0.9500000**<br>[0.9157908, 0.9756950] | **0.07212672**<br>[0.04053871, 0.11182950] |

# Graphical Bertillonage

- Rank
  1. North County High
  2. Wiley Bates Middle
  3. Annapolis High
  4. Lindale Middle
  5. Andover Park No Building
  6. Andover Park Tree as Building

# Naïve Bayes

- With Normalization

| | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| **Naïve Bayes** | **1.00E+00** | **1.00E+00** | **1.00E+00** | **1.00E+00** | **1.00E+00**<br>(Actual Eccentricity was outside of boundary) | **0.00E+00** |

- Without Normalization

| | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| **Naïve Bayes** | **2.89E-07** | **1.05E-16** | **1.05E-16** | **2.92E-35** | **8.05E-26**<br>Actual Eccentricity was outside of boundary) | **0.00E+00** |

# Match Quality

- Landcover Exponent = 1

| | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| Match Quality (q_zero= 0. 0001, q_pr = 0.85) | 0.9745259 | 0.8287246 | 0.828394 | 0.7813322 | 0.4934152 | 0.04799298 |

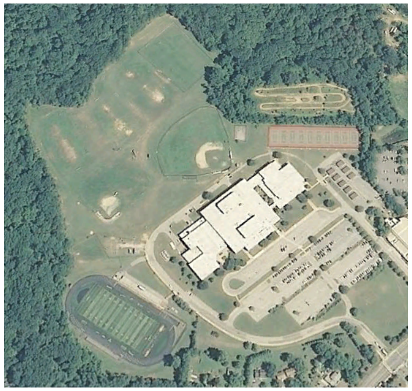
- Landcover Exponent = 100

| | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| Match Quality (q_zero= 0. 0001, q_pr = 0.85) | 0.926541 | 0.7889549 | 0.7889427 | 0.7880354 | 0.2416516 | 0.04602881 |

# Summary Table

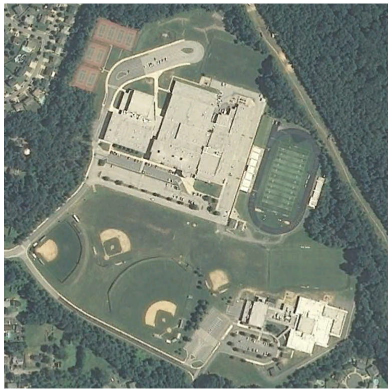
Quality score summary:

Annapolis High School



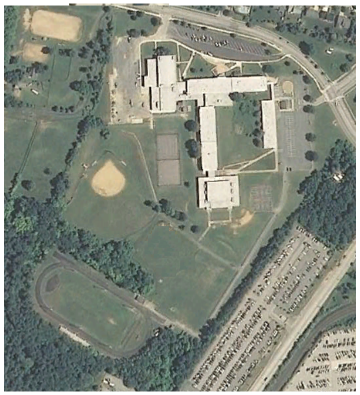
Ideal match.

North County High School



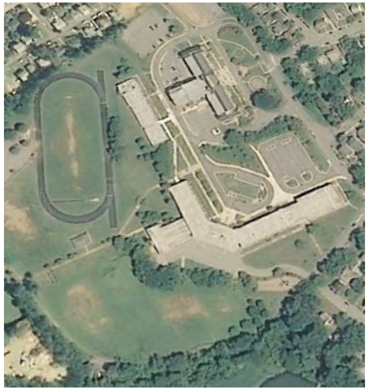
Missing one optional.

Lindale Middle School



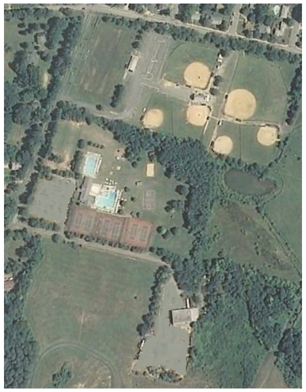
Missing two optional.

Wiley Bates Middle School



Missing two optional,
marginal parameters.

Andover Park



Missing one required,
other items perfect.

| $q_{pr} = 0.85$ | Annapolis High School | North County High School | Lindale Middle School | Wiley Bates Middle School | Andover Park (Trees as Building) | Andover Park (No Building) |
|---|---|---|---|---|---|---|
| Classroom Building | 0.98 | 0.98 | 0.98 | 0.82 | 0.12 | 0.00 |
| Parking Lot | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| Football Field | 0.98 | 0.98 | 0.96 | 0.98 | 0.95 | 0.95 |
| Baseball Field | 0.86 | 0.88 | 0.86 | 0.73 | 0.85 | 0.85 |
| Baseball Field #2 | 0.83 | 0.84 | 0.00 | 0.00 | 0.83 | 0.83 |
| Tennis Courts | 0.97 | 0.00 | 0.00 | 0.00 | 0.97 | 0.97 |
| **OVERALL QUALITY** | **0.94** | **0.88** | **0.80** | **0.74** | **0.56** ? | **0.10** |

# Our Current Method Choices

- Hierarchical Naïve Bayes when we have data
  - Primitive ensembles
- Match quality score when we do not

# More Challenges

- Dynamic/ephemeral components
- Human behavior
- Both
  - E.g, Tents on a mountainside. Cars in a parking lot.
  - Discontinuous snapshots may miss
- Elements whose existence reduces match likelihood
- Learning from negative examples