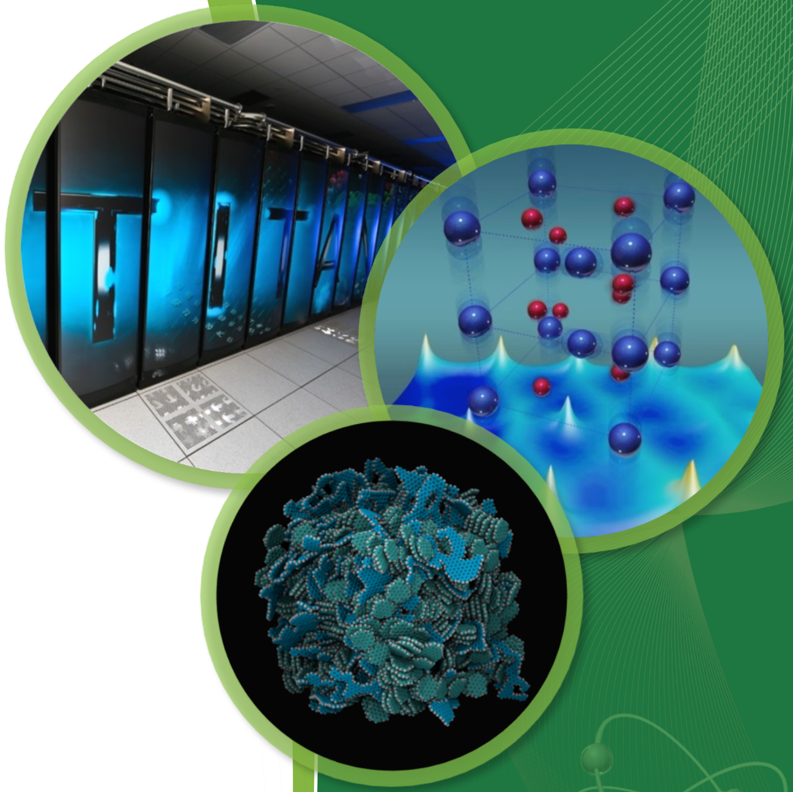
Active Learning Approach to Record Linking in Large Geodatasets

Alexandre Sorokine
Jason Kaufman
Robert Stewart
Jessie Piburn

Oak Ridge National Laboratory



November 2020

AUTOCARTO 2020
Virtual Meeting

ORNL is managed by UT-Battelle
for the US Department of Energy

Introduction

Integration of diverse datasets

- Very common task in geodata domain
- Technical issues
  - file formats
  - data transfer
  - projections
  - etc.
- Most technical issues have been solved

Present challenges

- Variety
  - Semantic diversity: PoIs, historic maps, OSM, traditional map products
- Volume
  - Dozens millions of features in a dataset is a new norm
- Automation is needed to make data integration feasible

Problem Statement

- Detect records that refer to the same real-world entity
  – Also this is known as conflation, data matching, record linking, entity resolution or alignment
- Goals of matching
  – Creation of a new datasets that incorporates original data in part or as a whole
  – Cross-verification of the datasets
  – Filling the gaps
  – Updating with newly acquired records
  – Establishing sameness or other types of relations among the features

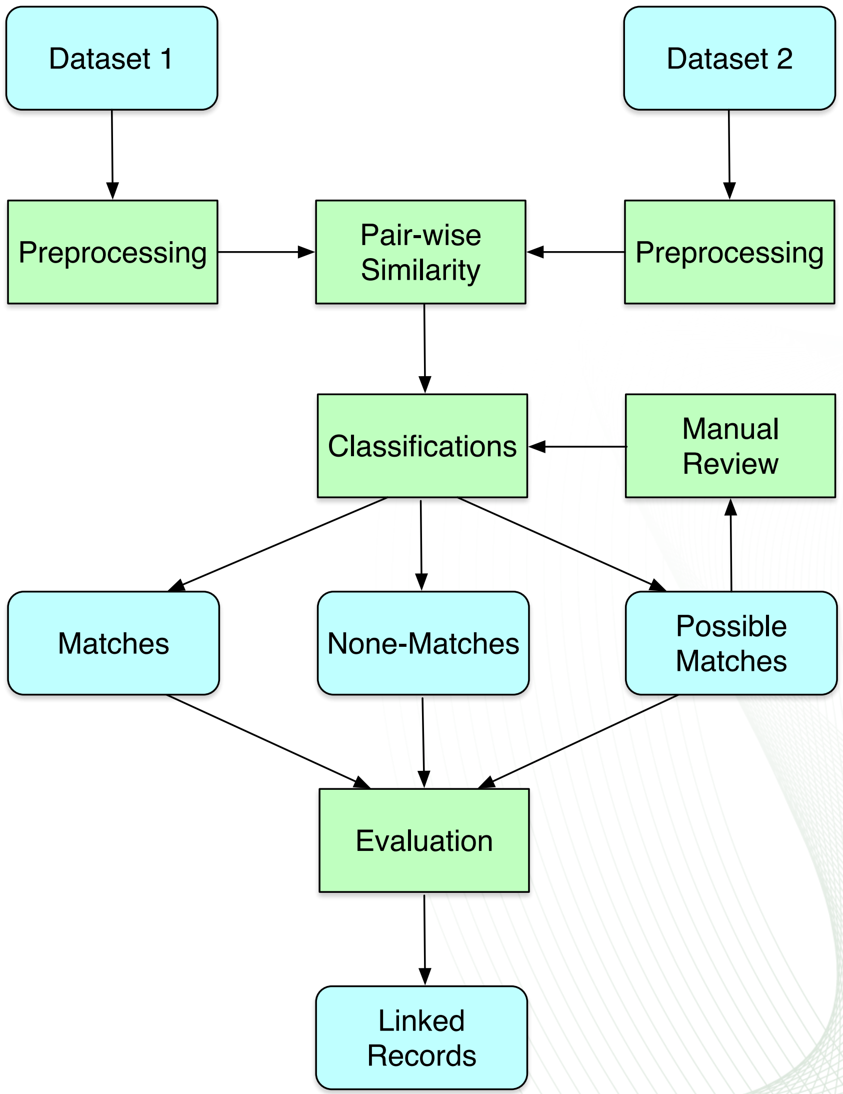
Earlier Work

- Conflation outside of geodata domain
  - problem formulated as early as 1960s
  - Medical records
  - Census data
  - Bibliographies, product catalogues, inventories, ...
- Geodata conflation: the term used since ca. 1985 at AutoCarto
  - Early work: geometric alignment of features
  - Present interest: VGI
    - NGA Hootenanny: https://github.com/ngageoint/hootenanny
  - Methods
    - Machine Learning – reduce hardcoded matching rules

OAK RIDGE National Laboratory

Record Linking Workflow

- Preprocessing
  - conversion to common format or API
- Pairwise similarity
- Classification of pairs
  - matches
  - possible matches
  - none-matches
- Evaluated for correctness
  - some matches may be reconsidered



Challenges Matching Medical and Census Records

- An entity having multiple records in different or in the same datasets
- Records often entered lack a common identifier or identifiers are wrong
  - *e.g.*, SSN should never be trusted
- Matching is achieved by
  - comparing salient attributes
  - discounting data entry errors
  - controlling spelling variations
  - handling missing values
  - detecting special circumstances like change of name or gender.

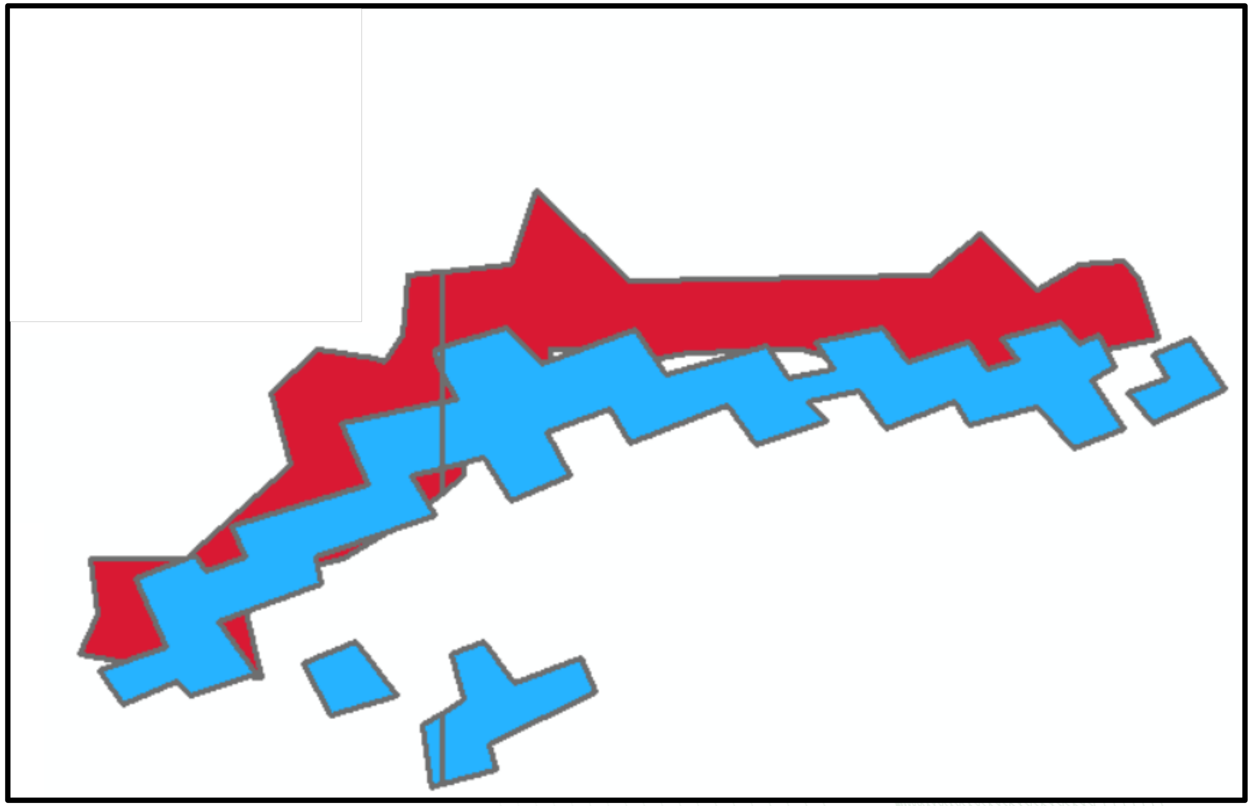
Semantics of Matching Geographic Features

- Locational information
- Generalization and scale
- Geographic categories
- Temporality: updates and change
- Relations among the objects
- Geophysical fields

What does it mean to be the same in the geographic space?

Locational Information

- Reduces number of potential matches
  - Safe to assume that nearby or overlapping features are at least related or the same real-world object
- Positional accuracy
  - multiple match candidates may fall within error bounds
  - significant problem in VGI
  - lack of attribute-level matching significantly reduces confidence
  - mixing up with neighbors



OAK RIDGE National Laboratory

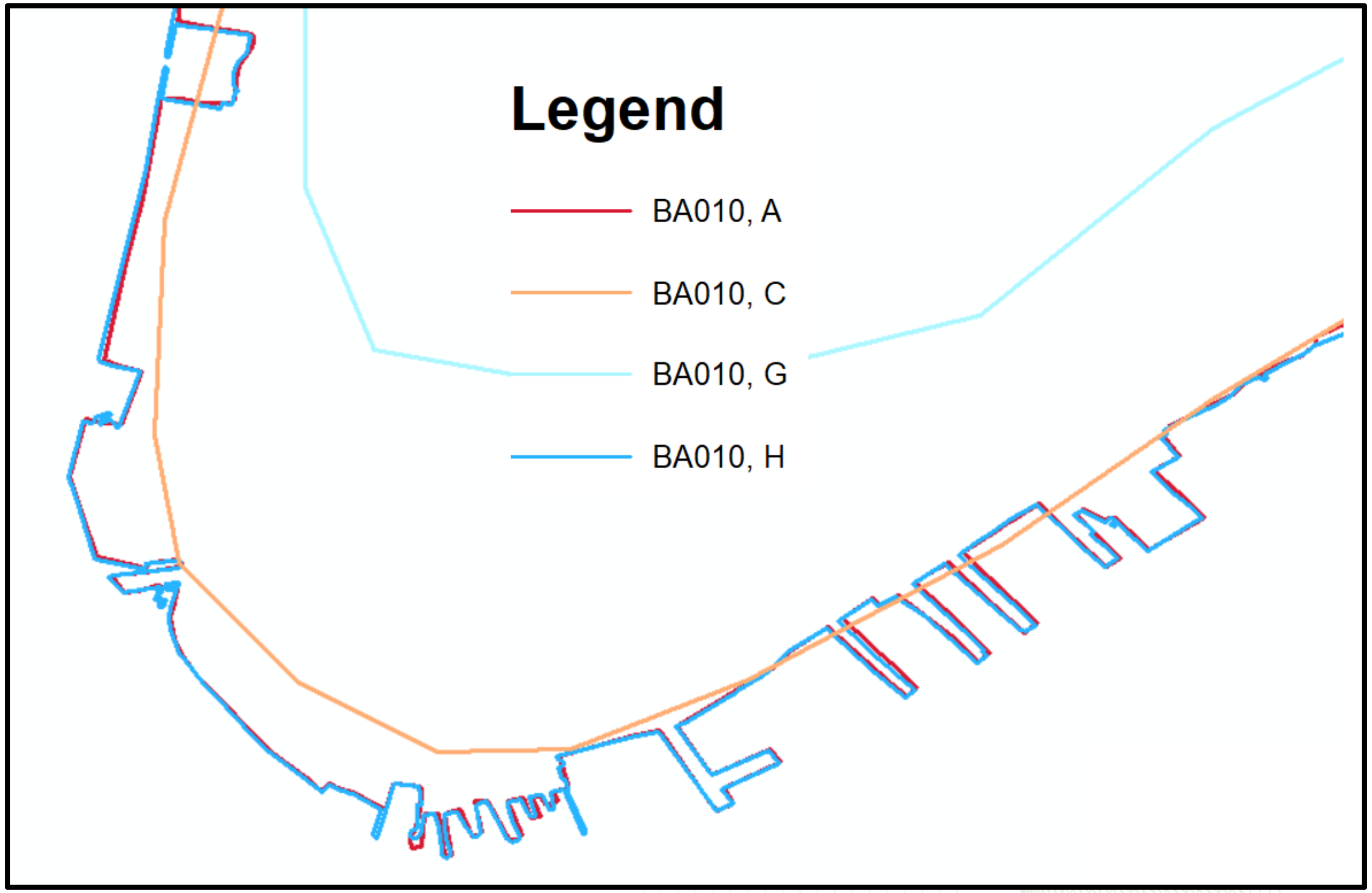
Geographic Categories and Feature Definitions

- Assumption: matched records should describe real-world objects of the same feature class
  - No such problem in medical and census records
- Same category objects occupying the same space
  - Administrative unit vs. municipality with the same name
- Compatibility of feature definitions
  - Convenience store and a gas station
- Problem of the subcategory “other”

OAK RIDGE National Laboratory

Generalization

- Matching across scales
- Link multiple records with different geometric representations
- Different positional accuracy at different scales

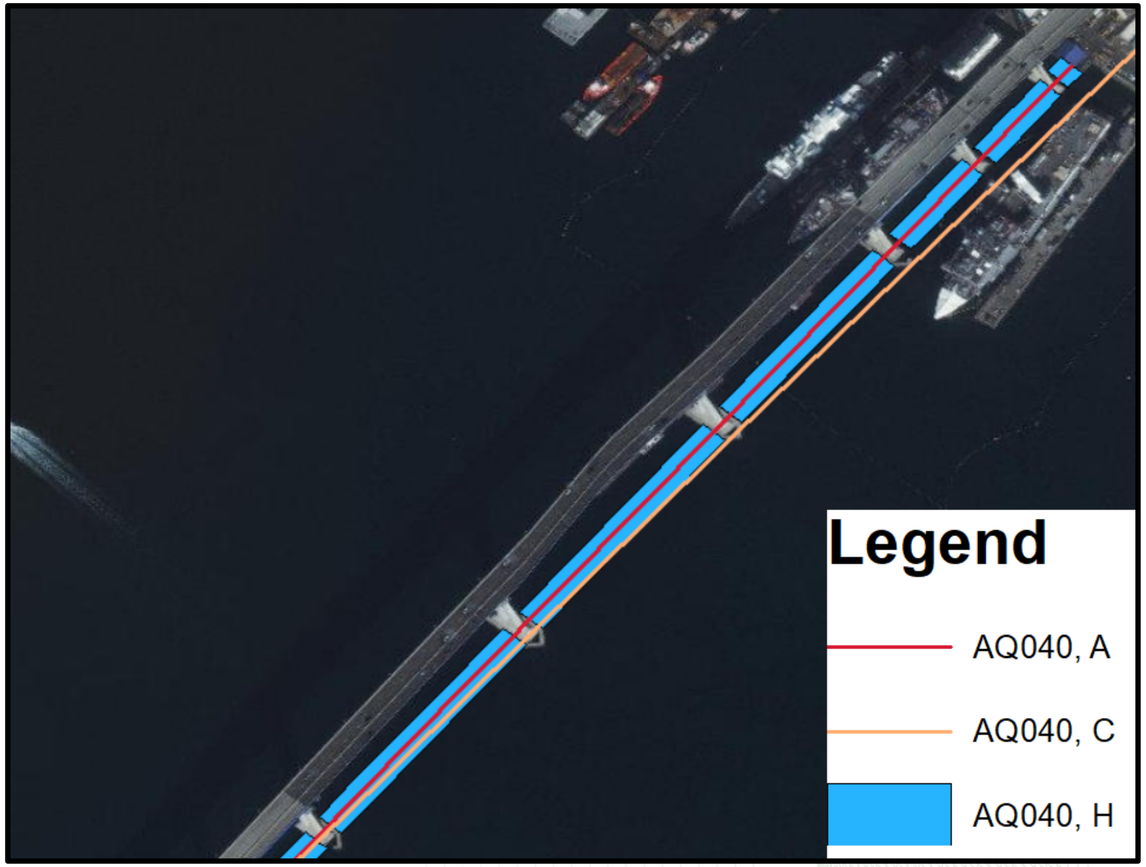


Temporality

- Very large range of temporal intervals
- Emerging, disappearing and changing objects vs. dataset updates
- Changing category
  - A province becomes an independent country
  - Lighthouse vs. museum
  - Restaurant replaced with barbershop
- Changing location
  - Settlement moved due to dam construction
  - Building physically moved
  - Islands merge

Object Relations

- Examples
  - Bridge and its pillars
  - Rock and a group of rocks
  - Museum and a restaurant
  - An arena and a gate
  - Building and main entrance
- Relations cannot be always expressed in the database schema



Case Study: Digital Nautical Chart by NGA

- Public domain data
  - https://dnc.nga.mil/
- More than 4 million features
- 4 scale levels
  - features at different scales are not linked to each other
- Significant temporal span of the data collection events
- Expectation of highly reliable results

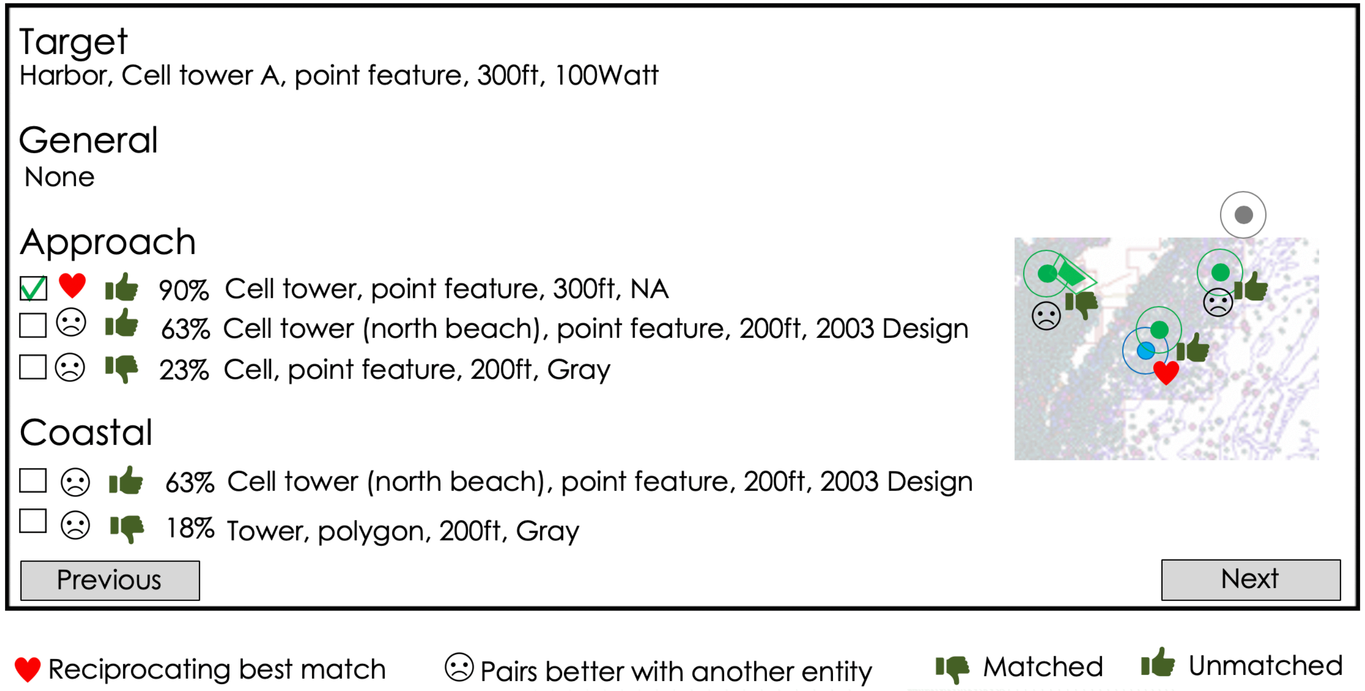
OAK RIDGE
National Laboratory

Approach

- Goal: highly automated process
  – close to 100% reliability required
- Recommender system with active machine learning learning
  – Each match must be approved by an analyst
  – Analyst feedback is fed back to ML to improve further recommendation
- Steps
  – Preprocessing: all feature loaded into a single table
  – Classification based on minimal distance and a feature class
    - Matches: within predefine accuracy with exact attribute match
    - None-matches: if distance exceeds predefined threshold
    - The rest are possible matches
  – Possible matches are handled by the recommender system

Recommender System

- **Recommender Systems** are tools that support user decision making by suggesting items that they are interested in
- **Active Learning (AL)** incorporates a user's response to its recommendations and re-trains the model to improve recommendations over time
- **Goal** is to provide initially useful and continuously improved recommendations



Similarity Vector

$$S_{i,j} = [d_1, d_2, \ldots, a_1, a_2, \ldots]$$

- Geographic proximity
  - minimal Euclidean distance
  - Hausdorf and Fréchet distances
  - percentage of the buffered overlap
- Attribute similarity
  - physical measurements: normalized difference
  - categorical values: exact match/not
  - entity names: Levenshtein distance
  - sets of attributes: Jaccard coefficient

Similarity Score

$$Score = [d_{1,} d_2 \dots, a_1, a_2, \dots] \cdot \begin{bmatrix} w_{d1}^0 \\ w_{d2}^0 \\ \dots \\ w_{a1}^0 \\ w_{a2}^0 \\ \dots \end{bmatrix}$$

- Weights are adjusted after each recommendation using Hierarchical Bayesian Logistic Regression

Summary

- Summary of the challenges for feature matching in diverse geodatasets
- Outline for a recommender-based active learning record matching system
- Potential improvements: adding more dimensions to the similarity vector
  - Neighbourhood measures
  - Text similarity between description categories

Questions?

Copyright and Disclaimer