**Introduction**

**Motivation**
- Big spatial data analytics often needs to geocode massive address datasets
- Privacy is a common concern for many applications, such as public health studies
- Cost effective geocoding is preferred by researchers or small businesses

**Limitations of Current Geocoding Systems**
- Open source ones are highly limited on accuracy
- Commercial geocoders are pricey for large address datasets
- Most are Web based and require sharing of private data over the Web

**Objective**
- Our aim is to develop an effective open source geocoder with an intelligent integrative geocoding model based on open data sources and machine learning, which achieves high accuracy, scalability, and privacy protection.

---

**Traditional Geocoding vs Integrative Intelligent Geocoding**

**System Overview**
- Integrate multiple open data sources
- Use machine learning to select best result intelligently from multiple candidates

**Parsing, Cleaning, Searching**
- Normalize the data after tokenizing: case sensitivity, abbreviations, house number formats, …
- Search input address as a query in database(s)

**Customized Query Generation**
- Address tokens are not always present in a data source, e.g., city name is missing in TIGER
- Generate queries based on sources characteristics to minimize error
- Generate customized queries for relaxation and resending to database

**Intelligent Answer Selection**
- Each query returns the best answer based on text similarity in a data source
- Each data source could be more accurate in one specific region or one specific input address
- Use machine learning with consideration of spatial proximity and implicit factors such as population density
  - Auto correction of building number separators, E.g. “74-58” mistyped as “7458”

**Gradient Tree Boost**
- Decision trees are learnt from training set
- Distance vector of input address to each candidate is calculated
- Decision trees specify which candidate is probabilistically nearer to input address

---

**Software Performance & Demonstration**

**Software**
- Web-based API for single and batch address geocoding
- The centroid of city/zip code/county also provided in results
- OS independent: MacOS, Windows and Linux
- Competitive with well-known geocoding systems
- Fast, efficient and highly scalable
- Open source

http://bmidb.cs.stonybrook.edu/easergeocoder/

---

**Acknowledgement**

This research is supported in part by grants from National Science Foundation ACI 1443054 and IIS 1350885.