Populating Entity Name Systems for Big Data Integration
Mayank Kejriwal¹ and Daniel P. Miranker (advisor)¹

¹{kejriwal,miranker}@cs.utexas.edu / The University of Texas at Austin

Heterogeneous Big Data Entity Resolution (ER)
50-year-old AI and database problem: How do we automatically resolve logically equivalent entities that occur in several syntactically different forms?

Now a Big Data problem!

Research Question and Thesis Statement
Can current state-of-the-art ER solutions address needs of Big Data integration?

Innovations
Automation: Generate a training set automatically using inexpensive heuristics, and make learning algorithms robust to potential noise.

Heterogeneity: Serialize RDF tabularly.

Quality:
- Use phonetic features
- Bootstrap aggregating
- Accommodate domain expertise if available

Full Scalable System: Implement key steps in MapReduce; impose tight theoretical bounds on serial module.

Data Integration
The Emerald architecture

In order to meet instance matching needs...

Entity Name System: a thesaurus for entities

Standard 2-step ER workflow
State-of-the-art has not managed to make the setup fully unsupervised yet.

Evaluation Plan

Preliminary Experimental Results
Automation: Noisy duplicates generation

Quality: Feature Selection techniques to compensate for noise

Competitive with supervised baseline by Bilenko, Kamath and Mooney (2006)

Open Questions

Acknowledgements
We would like to thank the National Science Foundation (NSF) for supporting the research and this trip.