Big Data Versus the Crowd
Looking for Relationships in All the Right Places
Ce Zhang, Feng Niu, Christopher Ré, and Jude Shavlik
Department of Computer Science, University of Wisconsin-Madison, USA

Motivating Application: DeepDive
DeepDive approaches relation extraction using human analysts and statistical signals from terabytes of data. It applies distant supervision to generate training examples from unlabeled text corpora.

Barack Obama brought his wife Michelle to Kenya three years later...

Contribution
Empirically study the factors that contribute to distant supervision quality and their relative impact.
Follow state-of-the-art approaches in each step, but study them in a new level of scalability (up to 100 million documents).

Takeaways
When developing a DS system, one should first expand the training corpus in order to improve recall and then worry about the precision of training examples.

Methodology: Follow State-of-the-art DS Scheme

Feature Extraction
- Mention extraction
  - Person, Location, Organization
  - Use Stanford CoreNLP
- Entity Linking
  - Link mentions to Freebase using exact string matching
- Linguistic pattern extraction
  - Dependency path between mentions
  - Word sequence between mentions

Distant Supervision
- Distant Supervision (DS)
  - Given Freebase as a knowledge base
  - find mention pairs supporting Freebase
- Human Feedback (HF)
  - Annotate mention pairs generated by distant supervision with Y/N:
    - 3 takers for each pair
    - with quality control

Impact of Big Data
- F1 is a log-linear function in the corpus size that we used for distant supervision.
  - The larger the corpus size, the higher the quality we can expect.
  - Large corpus increases the coverage of linguistic patterns in DS.

Impact of the Crowd
- F1 increases statistically significantly when we provide more human feedbacks.
  - But the slope is smaller than increasing the corpus size.
  - Human feedback has comparable precision as DS in our current protocol.

Future Work
- Study more sophisticated models (than LR) and distant supervision scheme.
- Explore other (more effective) usage of human feedbacks.

This work is generously supported by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of any of the above sponsors including DARPA, AFRL or the US government.

Big Data vs. the Crowd
Distant Supervision (DS)* automatically generates labeled training data from a text corpus. However, some labels are not accurate.

To Combat Inaccurate Labels in DS
- Use broad coverage and redundancy in the large corpus.
- Ask humans in the crowd to provide feedbacks.

Experiment Setup
- To Study our Hypotheses:
  - Subsample the document corpus to get different DS training set.
  - Subsample the human annotations to get different HF training set.
- Machine Learning Model
  - Train a logistic regression model using training data obtained from DS and HF.
- Data sets (another data set reported in paper)
  - TAC-KBP: 1.8M news articles
  - ClueWeb: 500M Web pages