Sarawagi Dedup with Active Learning

- Upshot:
  - 1) train on 100 rather than 5000 examples. (Out of 30K).
  - 2) Randomly tweaked committees of decision trees worked best. Decision trees are easy to understand.
- Machine Learning Basics
  - Vector space model
  - Classification. Supervised learning.
    - Binary classifiers
    - Decision trees
  - Clustering. Unsupervised learning.
- Representativeness
  - – e.g., density of node's neighborhood
  - – e.g., do sampling weighted by uncertainty. odds of picking an outlying item are low. Can mix sample of nodes with the top-k uncertainty nodes.
- This paper: active learning for training dedup classifiers
  - Entity resolution as a classification problem? Cross-product space.
    - Some efficiency issues to address
    - Vector space: 20-d. ngrams, overlapping words, edit distance, numbers off by 1, concatenated text,
  - What to train on? Hard to find good examples of false matches, even harder to find instructive examples of true matches.
- Ensemble ("Committee") Methods
  - – create n versions of a classifier (e.g. tweak parameters randomly)
  - – use the ensemble to decide which things are strongly classified, which are highly uncertain
- Active Learning: How to pick "instructive" examples? -- important for minimizing user
  - Uncertainty
  - Representativeness