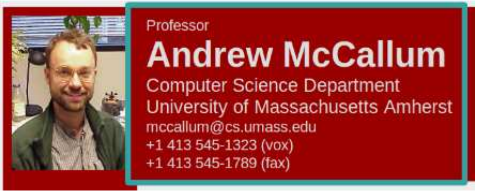
Paper 201: Extracting Researcher Metadata with Labeled Features

- Researcher metadata (*employment position*, *university affiliation*, *contact info*) is used in disambiguation, expert search and profile extraction [1,2]
- Researcher homepages are a good source of researcher metadata! For example:



Professor

Andrew McCallum

Computer Science Department
University of Massachusetts Amherst
mccallum@cs.umass.edu
+1 413 545-1323 (vox)
+1 413 545-1789 (fax)

- How can we reduce the amount of labeled data needed to learn accurate metadata extractors? Use Feature Labeling! [3]

1. Tang, et al. "ArnetMiner: extraction and mining of academic social networks" in KDD 2008
2. Balog, et al. "Broad expertise retrieval in sparse data environments" in SIGIR 2007
3. Druck, et al. "Learning from labeled features using generalized expectation criteria" in SIGIR 2008

Paper 201: Extracting Researcher Metadata with Labeled Features

- Labeled Feature = (feature, label-distribution)
  - (‘department’: AFFL=0.9, O=0.1)
  - (‘research’: POS=0.5, AFFL=0.3, O=0.2)
- We extract labeled features using field-specific terms and proximity between metadata fields on a homepage
  - Significant improvements in the tagging performance!
  - Considerable reduction in the number of training examples!

Field	**Precision**		**Recall**		**F1**	
	Basic	Best	Basic	Best	Basic	Best
AFFL	0.6670	0.4571	0.4302	0.7095	0.5219	**0.5554** (+6.4%)
EMAIL	0.9178	0.8889	0.8136	0.8693	0.8624	0.8788 (+1.9%)
FAX	0.9543	0.9501	0.9295	0.9406	0.9417	0.9453
PHN	0.9370	0.9310	0.8899	0.9296	0.9128	0.9303 (+1.9%)
POS	0.8048	0.7470	0.5995	0.6835	0.6870	**0.7138** (+3.9%)
UNIV	0.7203	0.6596	0.5827	0.7336	0.6432	**0.6940** (+7.9%)