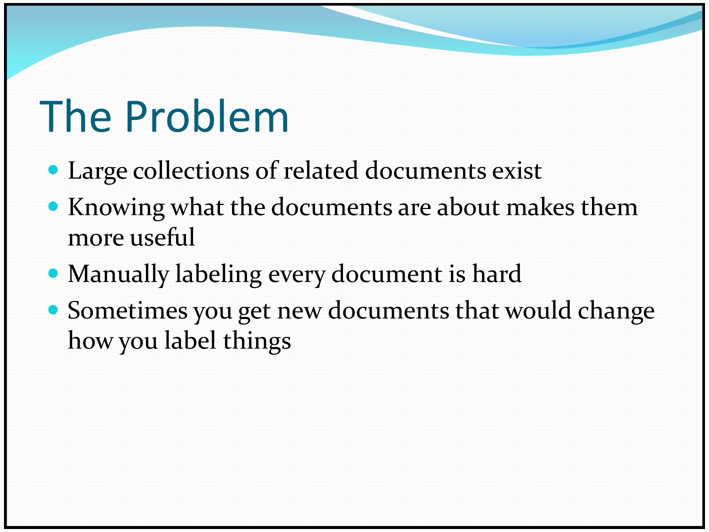
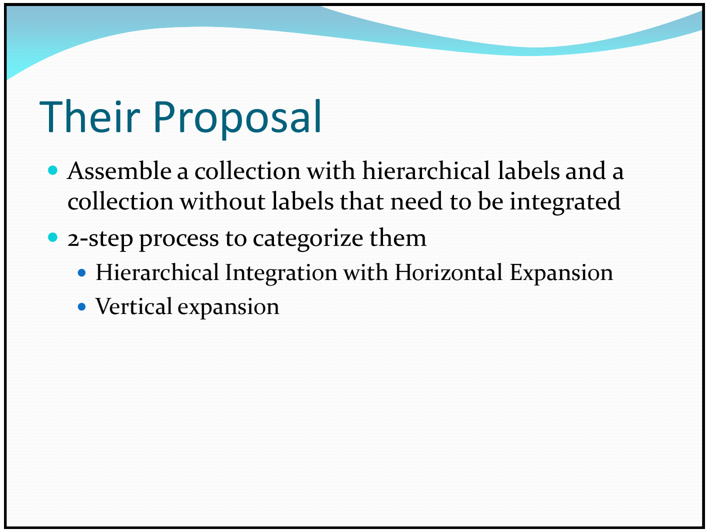
Hierarchical topic integration through semi-supervised hierarchical topic modeling

Xian-Ling Mao, Jing He, Hongfei Yan, and Xiaoming Li. 2012. Hierarchical topic integration through semi-supervised hierarchical topic modeling. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (CIKM '12). ACM, New York, NY, USA, 1612-1616. DOI=10.1145/2396761.2398483 http://doi.acm.org/10.1145/2396761.2398483



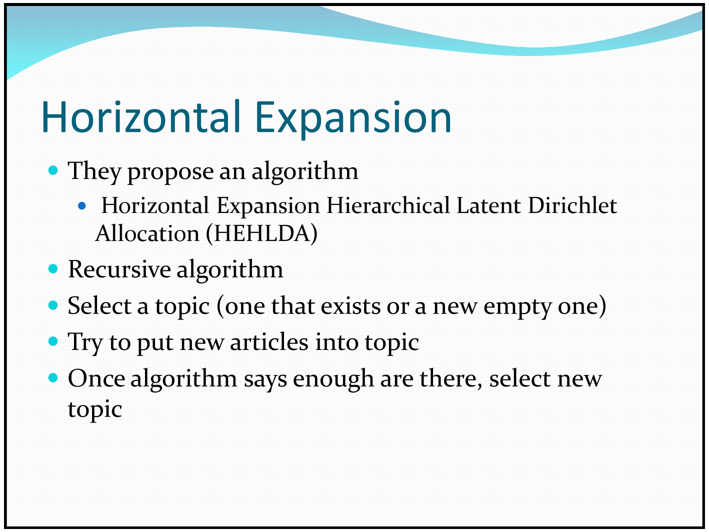
The Problem

- Large collections of related documents exist
- Knowing what the documents are about makes them more useful
- Manually labeling every document is hard
- Sometimes you get new documents that would change how you label things



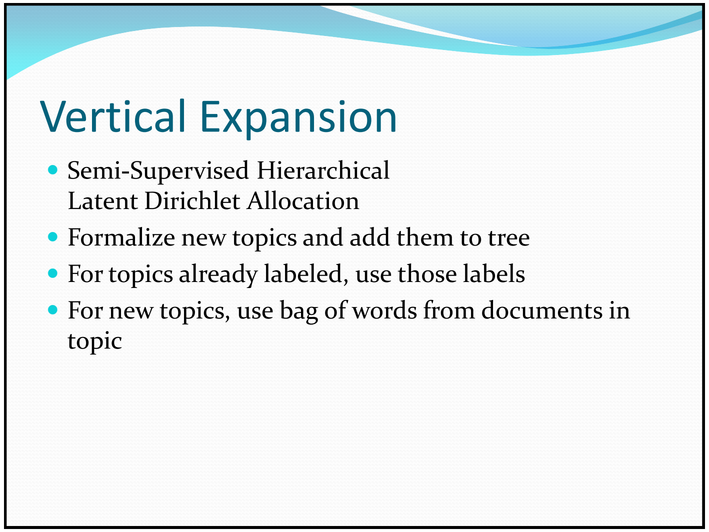
Their Proposal

- Assemble a collection with hierarchical labels and a collection without labels that need to be integrated
- 2-step process to categorize them
  - Hierarchical Integration with Horizontal Expansion
  - Vertical expansion



Horizontal Expansion

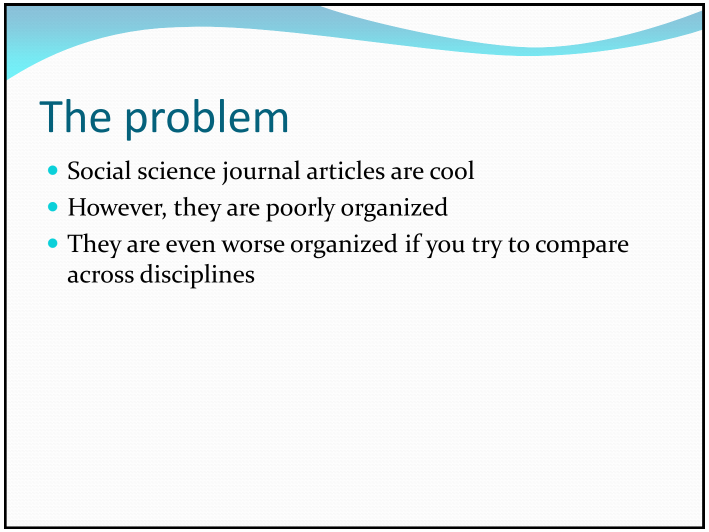
- They propose an algorithm
  - Horizontal Expansion Hierarchical Latent Dirichlet Allocation (HEHLDA)
- Recursive algorithm
- Select a topic (one that exists or a new empty one)
- Try to put new articles into topic
- Once algorithm says enough are there, select new topic



Vertical Expansion

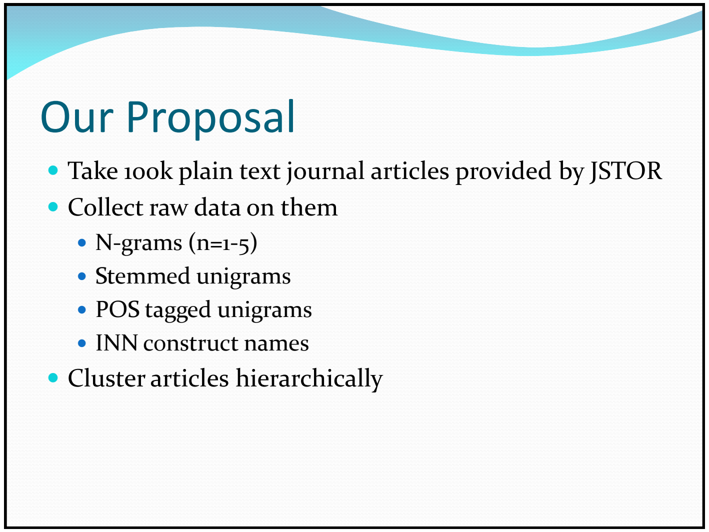
- Semi-Supervised Hierarchical Latent Dirichlet Allocation
- Formalize new topics and add them to tree
- For topics already labeled, use those labels
- For new topics, use bag of words from documents in topic

What we're doing



The problem

- Social science journal articles are cool
- However, they are poorly organized
- They are even worse organized if you try to compare across disciplines



Our Proposal

- Take 100k plain text journal articles provided by JSTOR
- Collect raw data on them
  - N-grams (n=1-5)
  - Stemmed unigrams
  - POS tagged unigrams
  - INN construct names
- Cluster articles hierarchically

What is JSTOR?

- JSTOR is a collection of articles form academic journals
- Contains around 5 million articles across disciplines
- We have just under 100k articles limited to social science

What is INN?

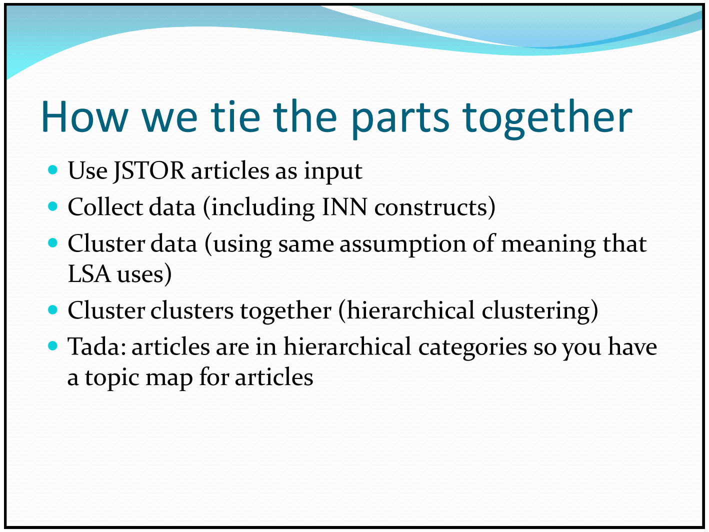
- Inter-Nomological Network
- Designed to be a “periodic table for social sciences”
- Connects constructs to articles and constructs to each other
- Constructs are connected even if they have different words

What is LSA?

- Latent Semantic Analysis
- Bag of words approach to finding how similar things are
- Based on assumption that words get meaning from frequent colocations
- Start by creating a semantic space, then using SVD to make it reasonably sized

Clustering Algorithms

- Three broad categories
  - Partition, density, grid-based
- Partition
  - Assumes clusters are all near a center point
- Density
  - Assumes clusters are a bunch of things close together separated from other clusters by empty space
- Grid-based
  - Designed for huge amounts of data with lots of dimensions



How we tie the parts together

- Use JSTOR articles as input
- Collect data (including INN constructs)
- Cluster data (using same assumption of meaning that LSA uses)
- Cluster clusters together (hierarchical clustering)
- Tada: articles are in hierarchical categories so you have a topic map for articles