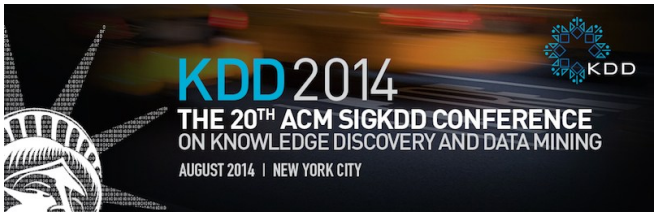
# Mining Topics in Documents Standing on the Shoulders of Big Data

Zhiyuan (Brett) Chen and Bing Liu




UIC UNIVERSITY OF ILLINOIS AT CHICAGO
Department of Computer Science
COLLEGE OF ENGINEERING

1st International Workshop on Big Data Discovery & Curation

# Topic Models

Widely used in many applications

Most of them are unsupervised

However, topic models

- Require a large amount of docs
- Generate incoherent topics

# Example Task

Finding product features from reviews

Most products do not even have 100 reviews.

# Example Topics of LDA

LDA topics with 100 reviews

Poor performance.

| Topic A | Topic B |
|---|---|
| price | sleeve |
| bag | hour |
| battery | design |
| file | simple |
| screen | video |
| dollar | mode |
| headphone | mouse |

Can we improve modeling using Big Data?

# Human Learning

A person 

sees a new situation 

uses previous experience (Years of Experience)

# Model Learning

A model 


sees a new domain 

uses data of many previous domains (Big Data)

# Motivation

Learn as humans do, Lifelong Learning

Retain the results learned in the past

Use them to help learning in the future

# Proposed Model Flow

Retain the topics from previous domains

Learn the knowledge from these topics

Apply the knowledge to a new domain

What's the knowledge representation?

How does a  gain knowledge?

Should / Should not

# Knowledge Representation

Should => Must-Links

e.g., {battery, life}

Should not => Cannot-Links

e.g., {battery, beautiful}

# Proposed Model Flow

**Algorithm 1** AMC($D^t$, $S$, $M$)

1: $A^t \leftarrow$ GibbsSampling($D^t$, $N$, $M$, $\emptyset$); // $\emptyset$: no cannot-links.
2: **for** $r = 1$ **to** $R$ **do**
3: $\quad C \leftarrow C \cup$ MineCannotLinks($S$, $A^t$);
4: $\quad A^t \leftarrow$ GibbsSampling($D^t$, $N$, $M$, $C$);
5: **end for**
6: $S \leftarrow$ Incorporate($A^t$, $S$);
7: $M \leftarrow$ MiningMustLinks($S$);

# Proposed Model Flow

**Algorithm 1** AMC($D^t$, $S$, $M$)

1: $A^t \leftarrow$ GibbsSampling($D^t$, $N$, $M$, $\emptyset$); // $\emptyset$: no cannot-links.
2: **for** $r = 1$ **to** $R$ **do**
3: $C \leftarrow C \cup$ MineCannotLinks($S$, $A^t$);
4: $A^t \leftarrow$ GibbsSampling($D^t$, $N$, $M$, $C$);
5: **end for**
6: $S \leftarrow$ Incorporate($A^t$, $S$);
7: $M \leftarrow$ MiningMustLinks($S$);

# Knowledge Extraction

Motivation: a person learns knowledge when it happens repetitively.

A piece of knowledge is reliable if it appears frequently.

# Frequent Itemset Mining (FIM)

Issue of single minimum support threshold

Multiple minimum supports frequent itemset mining (Liu et al., KDD 1999)

Directly applied to extract Must-Links

# Extracting Cannot-Links

O(V^2) Cannot-links in total

A domain has a small set of vocabulary

Only for those top topical words

# Related Work about Cannot-Links

Only two topic models were proposed to deal with cannot-type knowledge:

DF-LDA (Andrzejewski et al., ICML 2009)

MC-LDA (Chen et al., EMNLP 2013)

However, both of them assume the knowledge to be correct.

# Knowledge Verification

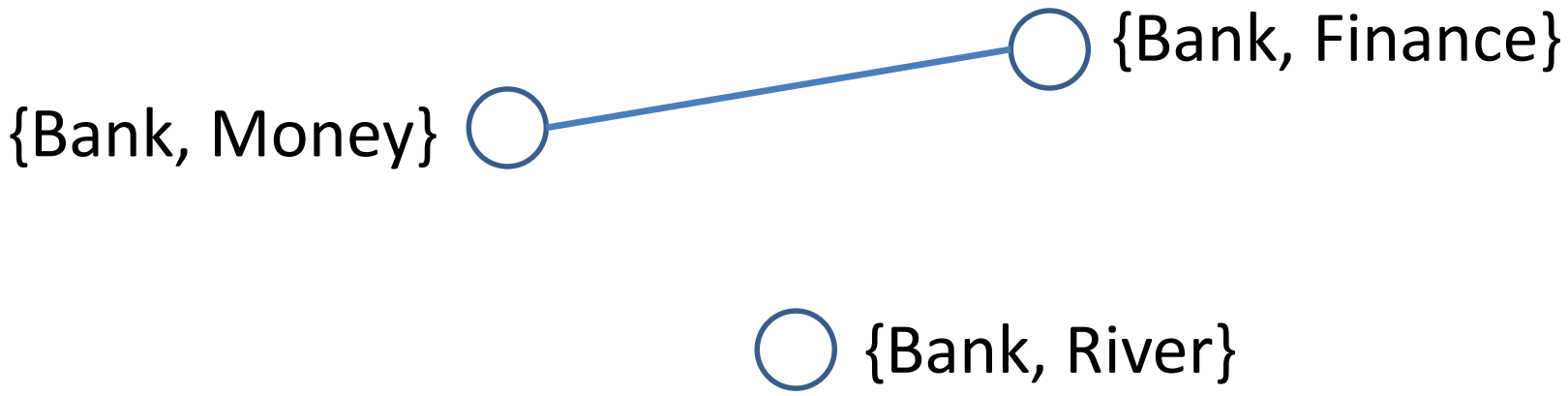
Motivation: a person's knowledge may not be applicable to a particular domain.

The knowledge needs to be verified towards a particular domain.

# Must-Link Graph

Vertex: must-link

Edge: must-links have original topic overlapping



# Pointwise Mutual Information

Estimate the correctness of a must-link

A positive PMI value implies semantic correlation

Will be used in the Gibbs sampling

# Cannot-Links Verification

Most words do not co-occur with most other words

Low co-occurrence does not mean negative sematic correlation

# Proposed Gibbs Sampler

M-GPU (*multi-generalized Pólya urn*) model

Must-links: increase the probability of both words of a must-link

Cannot-links: decrease the probability of one of words of a cannot-link

# Example

See word **speed** under topic 0:

Increase prob of seeing **fast** under topic 0 given must-link: {speed, fast}

Decrease prob of seeing **beauty** under topic 0 given cannot-link: {speed, beauty}

# M-GPU

Sample a must-link of word *w*

$$P(m_i = m|k) \propto P(w_1|k) \times P(w_2|k)$$

Construct a set of must-link {*m'*} given must-link graph

# M-GPU

Increase prob by putting must-link words into the sampled topic:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{w}, \alpha, \beta, \lambda)$$

$$\propto \frac{n_{d,k}^{-i} + \alpha}{\sum_{k'=1}^{K}(n_{d,k'}^{-i} + \alpha)}$$

$$\times \frac{\sum_{\{w',w_i\}\in\{m'\}} \lambda_{w',w_i} \times n_{k,w'}^{-i} + \beta}{\sum_{v=1}^{V}(\sum_{\{w',v\}\in\{m'_v\}} \lambda_{w',v} \times n_{k,w'}^{-i} + \beta)}$$

# M-GPU

Increase prob by putting must-link words into the sampled topic:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{w}, \alpha, \beta, \lambda)$$

$$\propto \frac{n_{d,k}^{-i} + \alpha}{\sum_{k'=1}^{K}(n_{d,k'}^{-i} + \alpha)}$$

$$\times \frac{\sum_{\{w',w_i\}\in\{m'\}} \lambda_{w',w_i} \times n_{k,w'}^{-i} + \beta}{\sum_{v=1}^{V}(\sum_{\{w',v\}\in\{m'_v\}} \lambda_{w',v} \times n_{k,w'}^{-i} + \beta)}$$

# M-GPU

Increase prob by putting must-link words into the sampled topic:

$$p(z_i = k|\boldsymbol{z}^{-i}, \boldsymbol{w}, \alpha, \beta, \lambda)$$

$$\propto \frac{n_{d,k}^{-i} + \alpha}{\sum_{k'=1}^{K}(n_{d,k'}^{-i} + \alpha)}$$

$$\times \frac{\sum_{\{w',w_i\}\in\{m'\}} \lambda_{w',w_i} \times n_{k,w'}^{-i} + \beta}{\sum_{v=1}^{V}(\sum_{\{w',v\}\in\{m'_v\}} \lambda_{w',v} \times n_{k,w'}^{-i} + \beta)}$$

# M-GPU

Decrease prob by transferring cannot-link word into other topic with higher word prob:

$$P(z_{q_c} = k | \boldsymbol{z}^{-q_c}, \boldsymbol{w}, \alpha, \beta, \lambda, q = q_c)$$

$$\propto \boldsymbol{I}_{[0, p(w_c|k)]}(P(w_c|z_c))$$

$$\times \frac{n_{d,k}^{-q_c} + \alpha}{\sum_{k'=1}^{K}(n_{d,k'}^{-q_c} + \alpha)}$$

$$\times \frac{\sum_{\{w', w_i\} \in \{m'_c\}} \lambda_{w', w_i} \times n_{k,w'}^{-q_c} + \beta}{\sum_{v=1}^{V}(\sum_{\{w', v\} \in \{m'_v\}} \lambda_{w', v} \times n_{k,w'}^{-q_c} + \beta)}$$

# M-GPU

Decrease prob by transferring cannot-link word into other topic with higher word prob:

$$P(z_{q_c} = k | \boldsymbol{z}^{-q_c}, \boldsymbol{w}, \alpha, \beta, \lambda, q = q_c)$$

$$\propto \boldsymbol{I}_{[0, p(w_c|k)]}(P(w_c|z_c))$$

$$\times \frac{n_{d,k}^{-q_c} + \alpha}{\sum_{k'=1}^{K}(n_{d,k'}^{-q_c} + \alpha)}$$

$$\times \frac{\sum_{\{w', w_i\} \in \{m'_c\}} \lambda_{w', w_i} \times n_{k,w'}^{-q_c} + \beta}{\sum_{v=1}^{V}(\sum_{\{w', v\} \in \{m'_v\}} \lambda_{w', v} \times n_{k,w'}^{-q_c} + \beta)}$$

# M-GPU

Note that we do not increase the number of topics as MC-LDA did.

Rational: cannot-links may not be correct, e.g., {battery, life}.

# Evaluation

100 Domains (50 Electronics, 50 Non-Electronics), 1,000 review each

100 reviews for each test domain

Knowledge extracted from 1,000 reviews from other domains

# Model Comparison

AMC (AMC-M: must-links only)

LTM (Chen et al., 2014)

GK-LDA (Chen et al., 2013)

DF-LDA (Andrzejewski et al., 2009)

MC-LDA (Chen et al., 2013)

LDA (Blei et al., 2003)

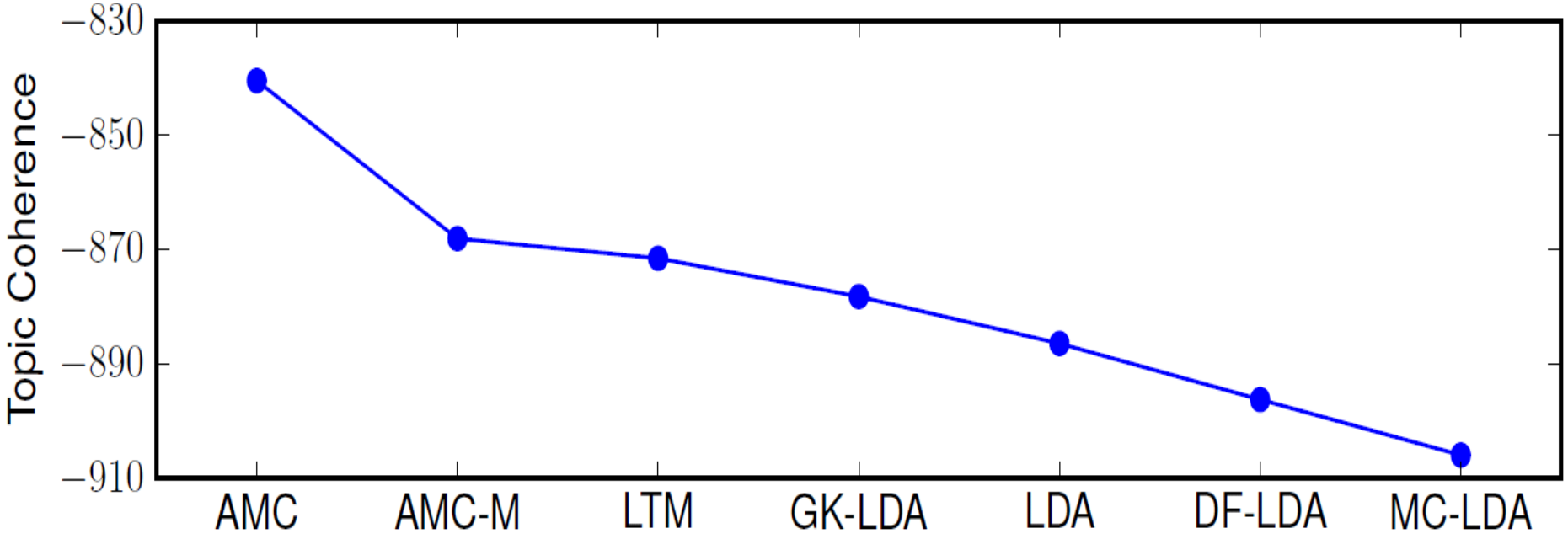
# Topic Coherence

Proposed by Mimno et al., EMNLP 2011

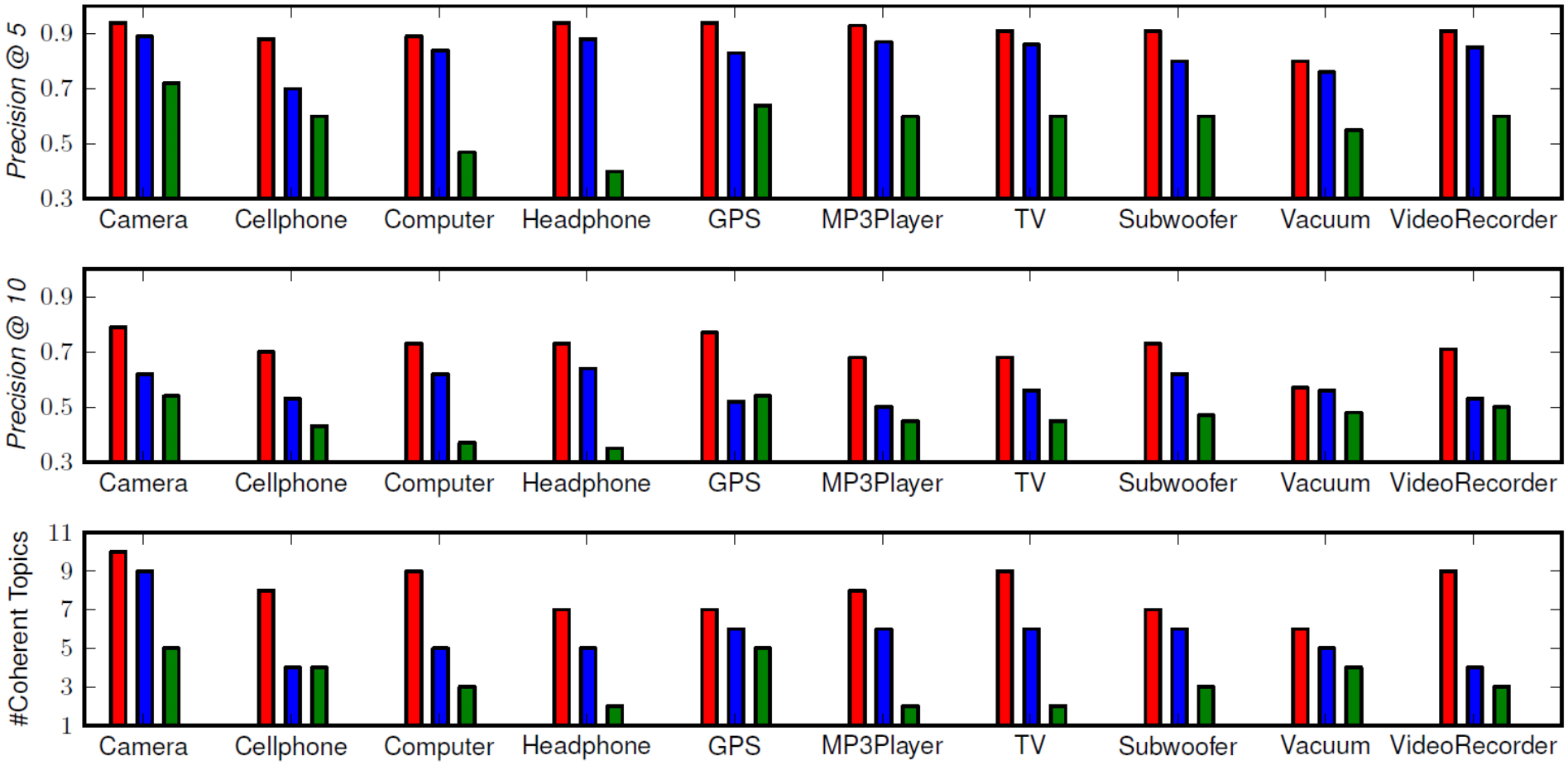
Higher score means more coherent topics

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

# Topic Coherence Results



# Human Evaluation Results

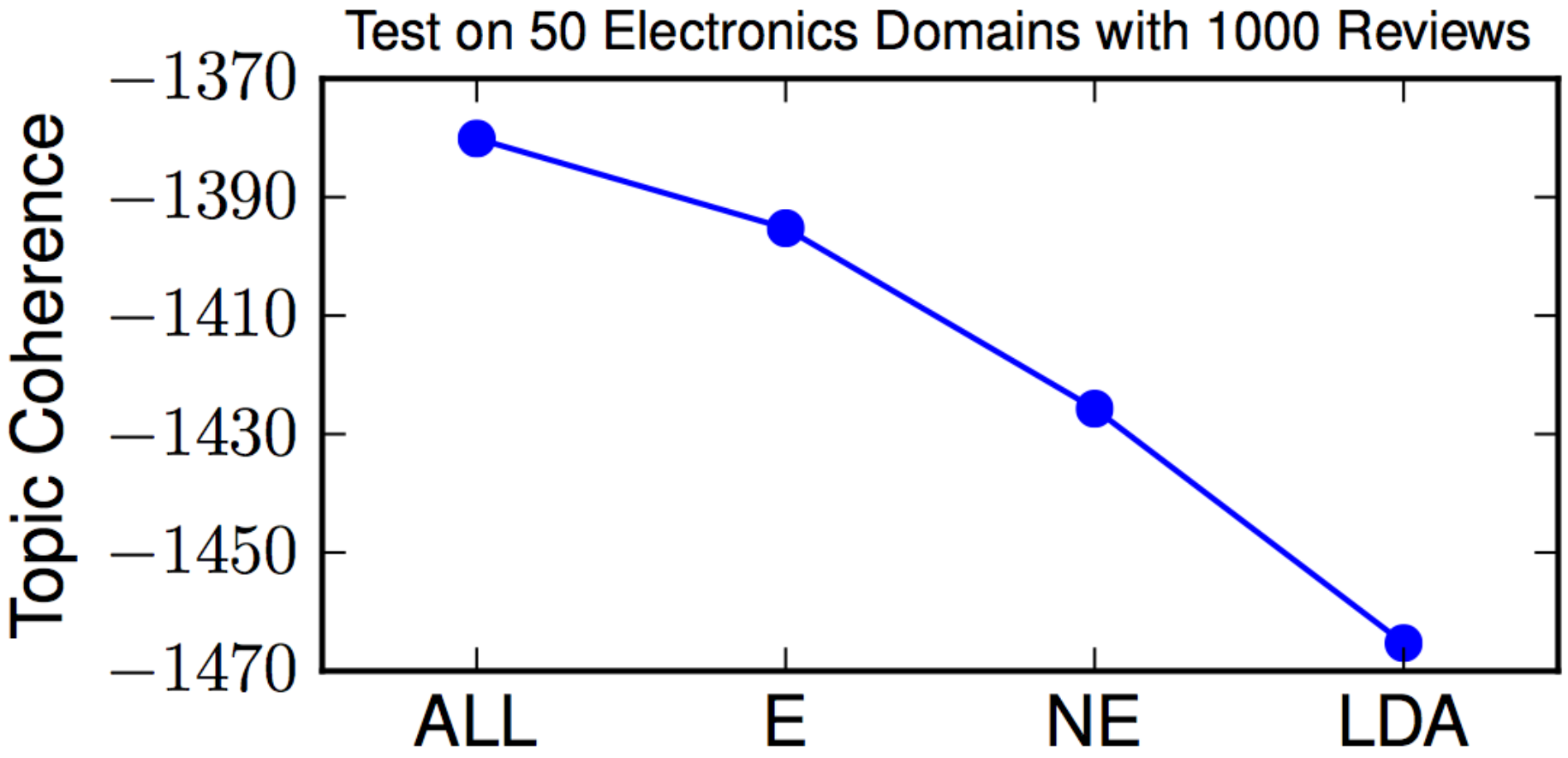



Red: AMC; Blue: LTM; Green: LDA

# Example Topics

| Price | | |
|---|---|---|
| **AMC** | **LTM** | **LDA** |
| money | *shot* | *image* |
| buy | money | price |
| price | *review* | *movie* |
| range | price | *stabilization* |
| cheap | cheap | *picture* |
| expensive | *camcorder* | *technical* |
| deal | *condition* | *photo* |
| *point* | *con* | *dslr* |
| *performance* | *sony* | *move* |
| *extra* | *trip* | *short* |

# Electronics vs. Non-Electronics



# Conclusions

Learn as humans do

Use big data to help small data

Knowledge extraction and verification

M-GPU model

# Future Work

Knowledge engineering: how to store/maintain the knowledge

Importance of domains, domain selection

# Q&A

Thank you!