Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization

Hannah Kim[1], Jaegul Choo[2], Jingu Kim[3], Chandan K. Reddy[4], Haesun Park[1]

Georgia Tech[1], Korea University[2], Netflix Inc.[3], Wayne State University[4]

hannahkim@gatech.edu

August 11th 2015 SIGKDD

Outline

- Motivation
- Topic Modeling via NMF
- Experiments
  - Quantitative Evaluation
  - Case Study
- Conclusion

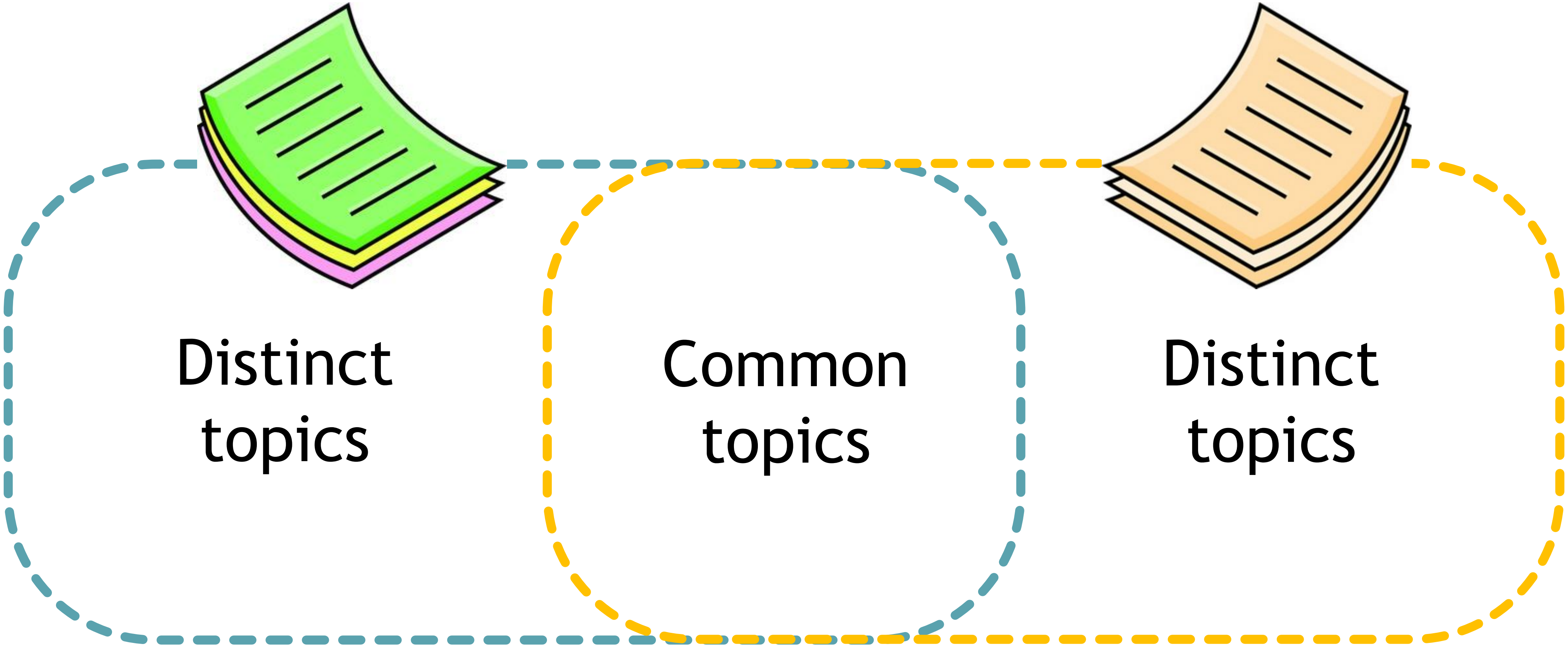
Motivation

- Understanding large-scale document collections is important
- In many real world applications,
  we often need to **compare** and **contrast** document sets

- We may want to analyze w.r.t. additional information
  - author information (e.g., gender, age, and location)
  - network information (e.g., co-authorship and citation)
  - publishing information (e.g., year, publisher, and venue)

Example (1)

- E.g., Male- vs. female-authored documents



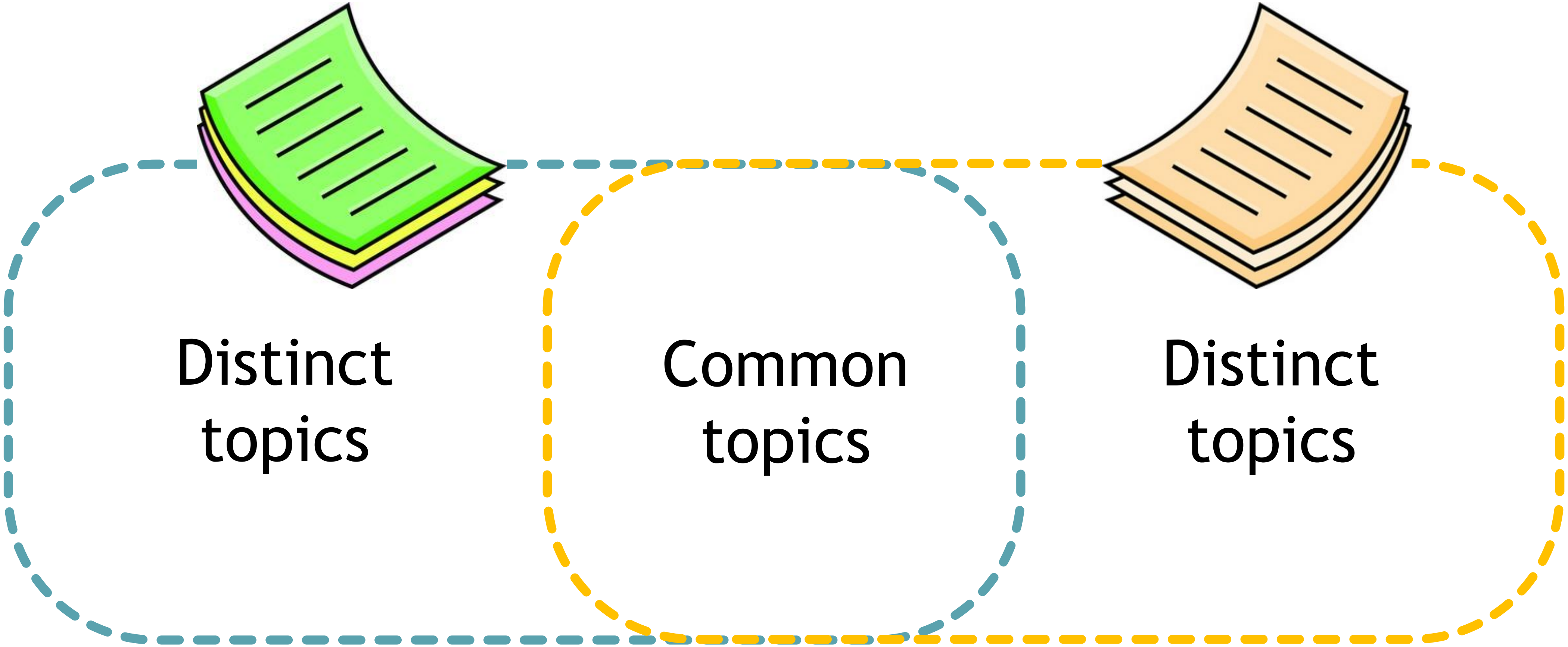
Distinct topics

Common topics

Distinct topics

Example (2)

- E.g., Old documents vs. new documents



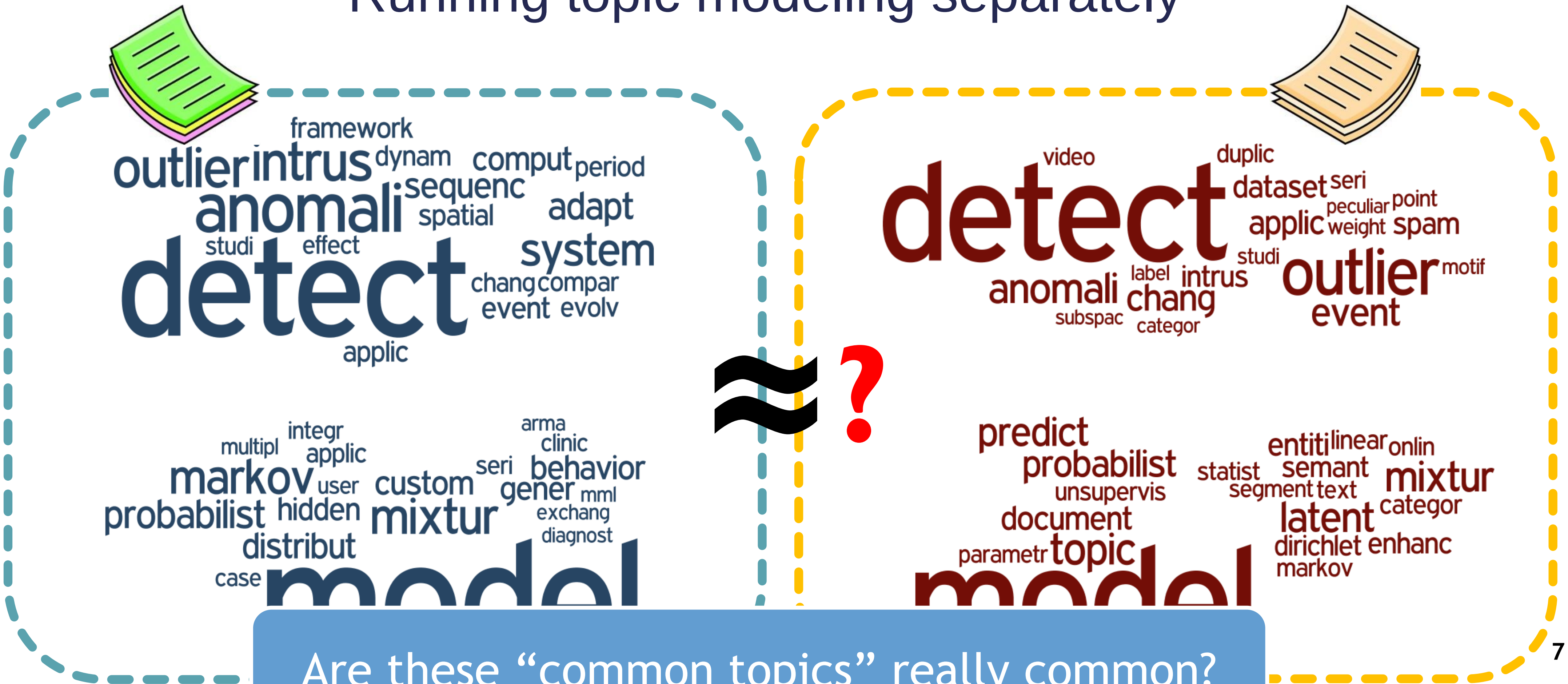
Motivation

- However, standard topic modeling cannot fully satisfy the needs to compare and contrast document sets

- Independently running standard NMF algorithms on different document sets **does not clearly reveal** their common and discriminative topics

Data mining papers published in 2000-2005 vs. 2006-2008

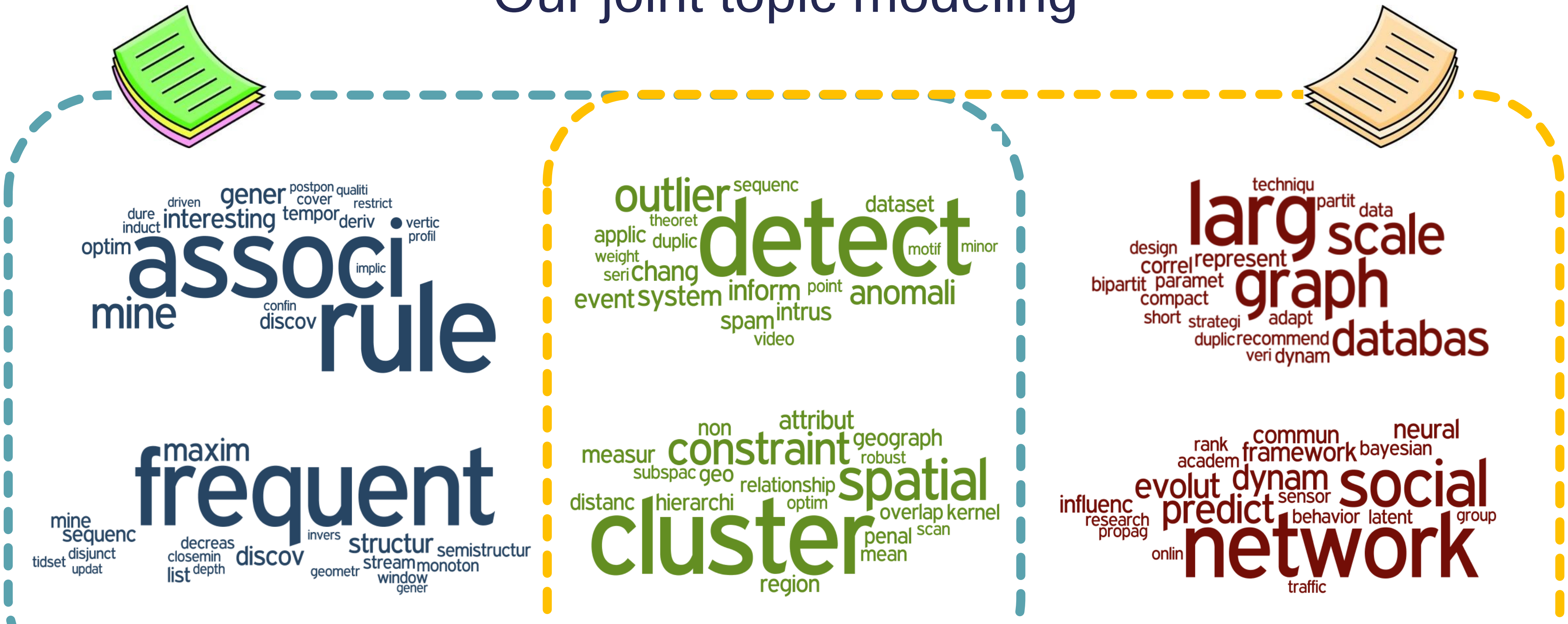
Running topic modeling separately



Are these “common topics” really common?

Data mining papers published in 2000-2005 vs. 2006-2008

Our joint topic modeling

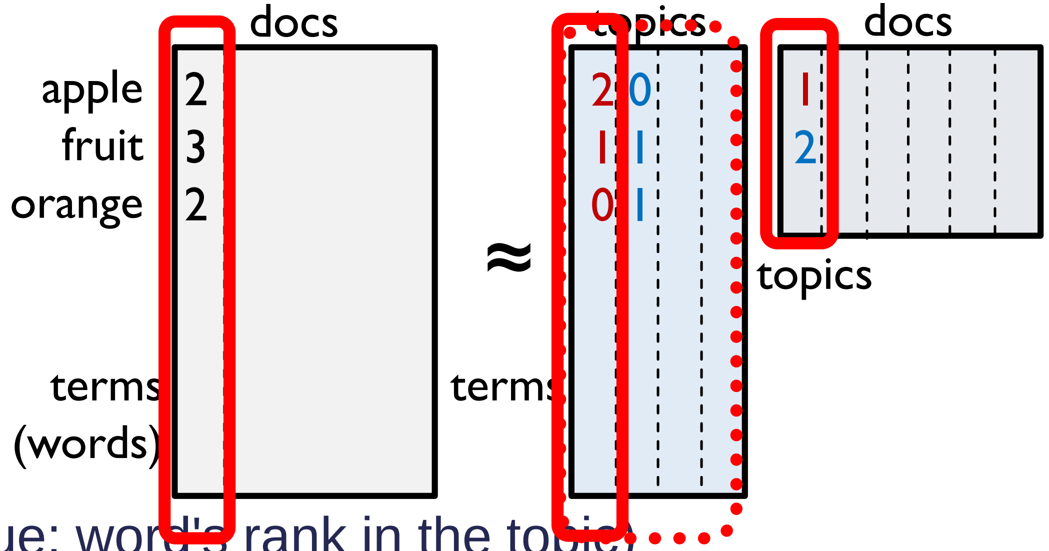

assoc rule mine optim interesting gener driven dure induct tempor deriv postpon cover qualiti restrict vertic profil implic confin discov
frequent maxim mine sequenc tidset disjunct updat decreas closemin list depth discov invers structur semistructur stream monoton geometr window gener
outlier detect sequenc theoret applic duplic weight seri chang dataset motif minor event system inform point anomali spam intrus video
cluster constraint non attribut geograph robust measur subspac geo relationship spatial distanc hierarchi optim overlap kernel penal scan mean region
larg scale techniqu partit data design correl represent bipartit paramet compact graph short strategi adapt duplic recommend databas veri dynam
network social commun neural rank framework bayesian academ evolut dynam sensor influenc predict behavior latent group research propag onlin traffic

Our joint NMF model identifies common topics and discriminative topics

Nonnegative Matrix Factorization (NMF) for Topic Modeling

- $X \approx WHT$

  term-document matrix(X)
  → term-topic matrix(W),
  topic-document matrix(H^T)



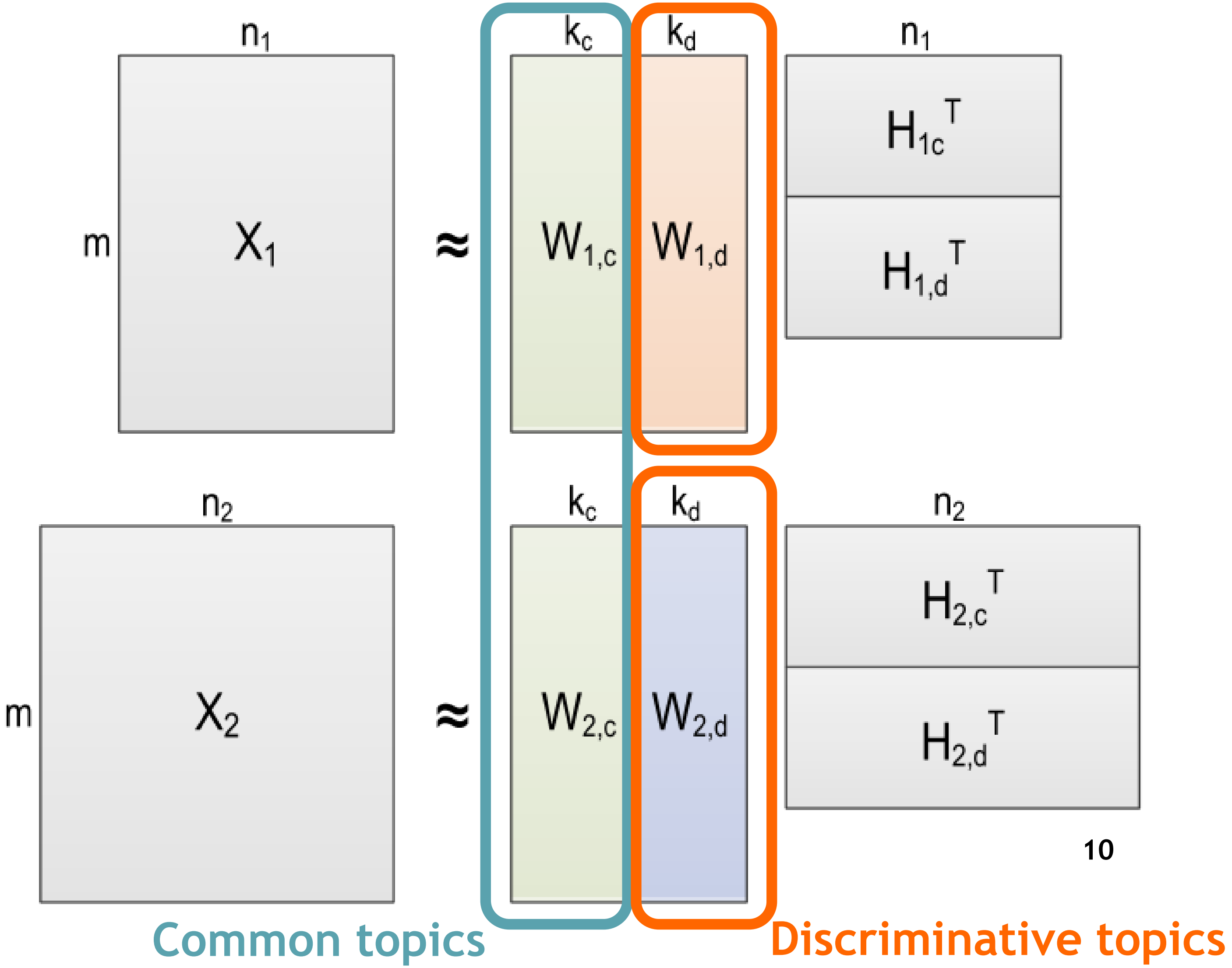
- Each topic,
  a nonnegative vector of words (value: word's rank in the topic)
- Each document, a linear combination of topic vectors
- Algorithm
  - Initialize W, H
  - Update W, H to optimize $\min_{W,H\geq 0} \|X - WH^T\|_F^2$

Our Joint NMF-based Model

- **GOAL**: Given two datasets, find common topics and discriminative topics from each dataset
- Formula

$$X_1 \approx W_1 H_1^T$$
$$X_2 \approx W_2 H_2^T,$$

where $W_{1,c} \cong W_{2,c}$
and $W_{1,d} \neq W_{2,d}$



Our Batch Processing Approach

- Optimize

$$\min_{W_1,H_1,W_2,H_2\geq 0} \frac{1}{n_1}\left\|X_1 - W_1 H_1^T\right\|_F^2 + \frac{1}{n_2}\left\|X_2 - W_2 H_2^T\right\|_F^2 \underbrace{+\alpha\left\|W_{1,c} - W_{2,c}\right\|_F^2}_{\text{Commonality penalty term}} + \underbrace{\beta\left\|W_{1,d}^T W_{2,d}\right\|_{1,1}}_{\text{Distinctiveness penalty term}}$$

- Block-coordinate descent framework:
  - Solve the objective function for a column while fixing the other column vectors of W_1, W_2, H_1, H_2

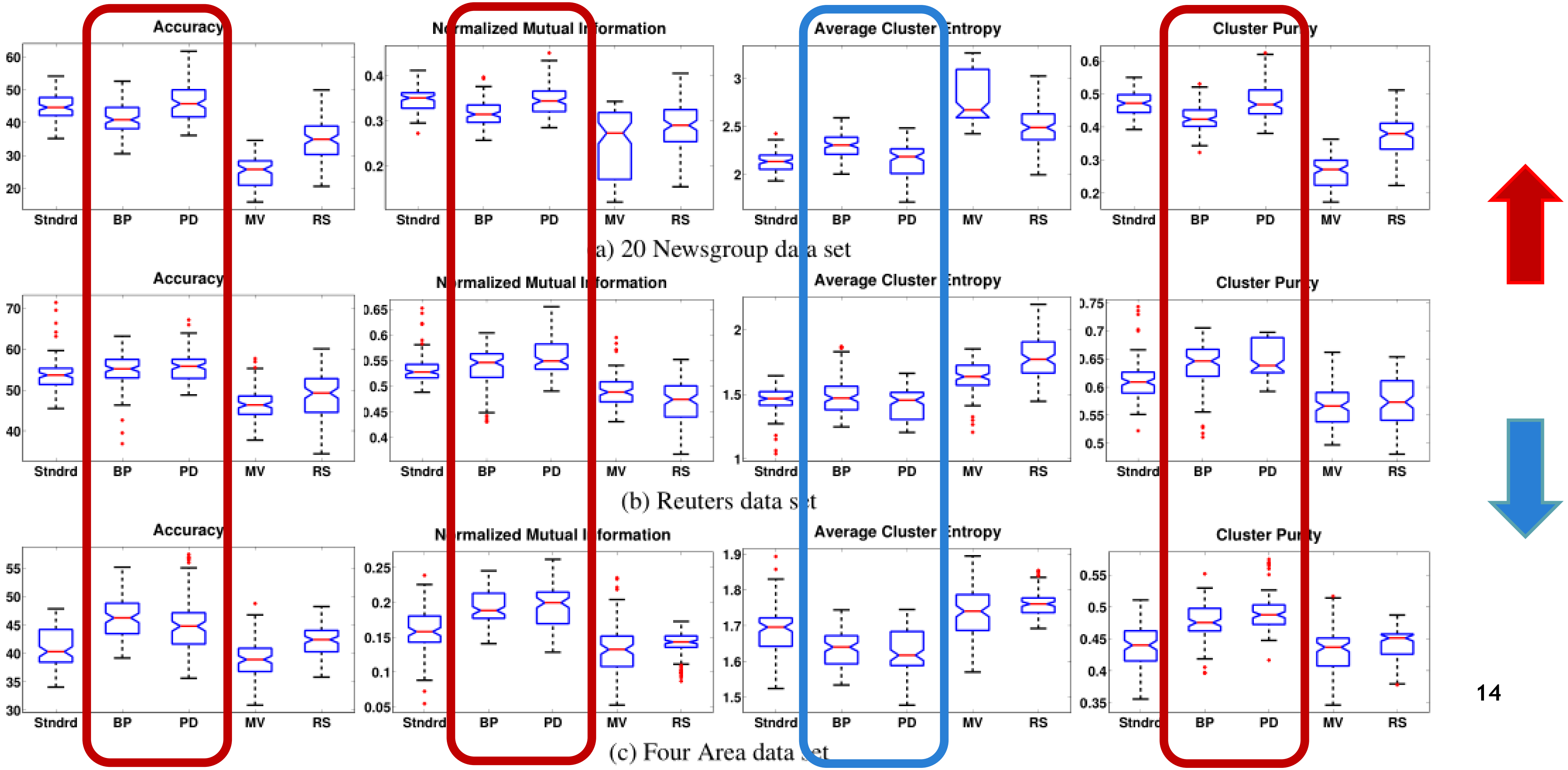
Our Pseudo-deflation Approach

- In practice, to understand topics, people check only a small number of the most representative, thus meaningful keywords.
- Our pseudo-deflation approach considers only the top keywords in each topic.
- However, considering only the top keywords presents a challenge – the objective function could change every iteration.
- To solve this, our pseudo-deflation approach discovers discriminative topics one by one, in a manner similar to a rank-deflation procedure.
- Please see our paper for detailed algorithm (Section 3.4)

Quantitative Evaluation – Clustering

- Assumption: by jointly performing clustering on multiple data sets and allowing both common and discriminative topics, our method would show better clustering performance
- Compared methods:
  - Standard NMF
  - **Our batch processing method (BS)**
  - **Our pseudo-deflation method (PD)**
  - Multiview NMF (MV) by Liu *et al.* SDM '13
  - Regularized shared subspace NMF (RS) by Gupta *et al.* DMKD '13
- Performance measures: accuracy, normalized mutual information, average cluster entropy, and cluster purity

Quantitative Evaluation



(a) 20 Newsgroup data set

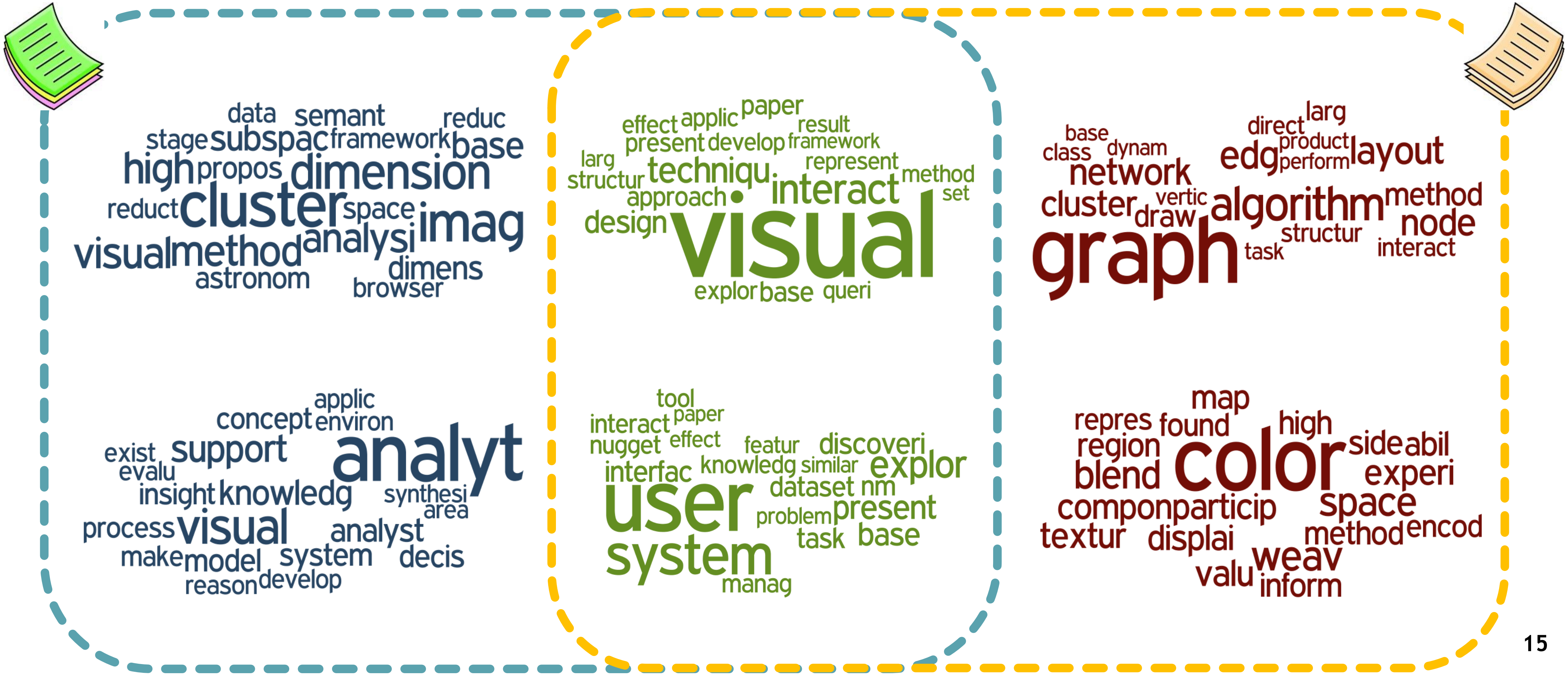
(b) Reuters data set

(c) Four Area data set

Case Study (1) – VAST vs. InfoVis Conferences

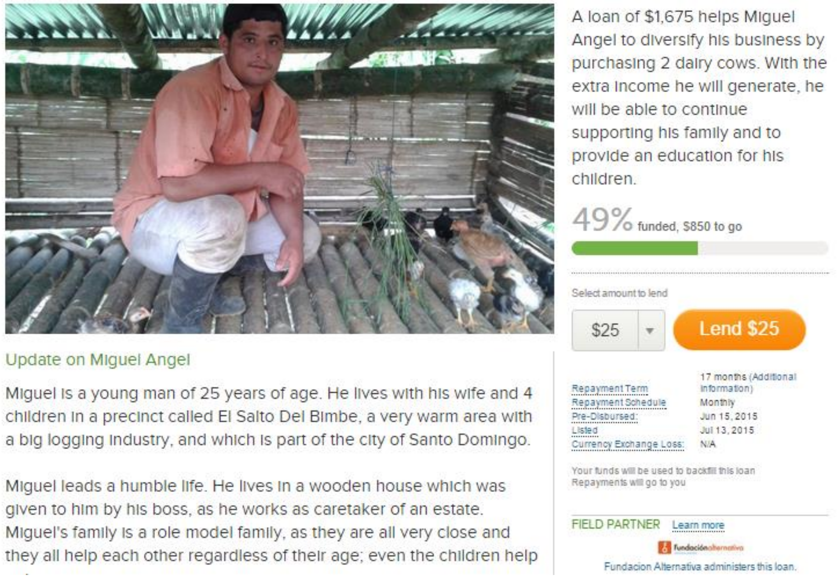
Visual Analytics Science and Technology (VAST)

Information Visualization (InfoVis)



Case Study (2) – Loan Description in Micro-finance

- KIVA.org is a nonprofit crowd-funding website where people in developing countries post loan requests
- Lenders can make a loan individually or as a team
- By analyzing loan description data, our method can help to characterize and promote lending activities



A loan of $1,675 helps Miguel Angel to diversify his business by purchasing 2 dairy cows. With the extra income he will generate, he will be able to continue supporting his family and to provide an education for his children.

49% funded, $850 to go

Select amount to lend

Update on Miguel Angel

Miguel is a young man of 25 years of age. He lives with his wife and 4 children in a precinct called El Salto Del Bimbe, a very warm area with a big logging industry, and which is part of the city of Santo Domingo.

Miguel leads a humble life. He lives in a wooden house which was given to him by his boss, as he works as caretaker of an estate. Miguel's family is a role model family, as they are all very close and they all help each other regardless of their age; even the children help

Repayment Term	17 months (Additional information)
Repayment Schedule	Monthly
Pre-Disbursed:	Jun 15, 2015
Listed	Jul 13, 2015
Currency Exchange Loss:	N/A

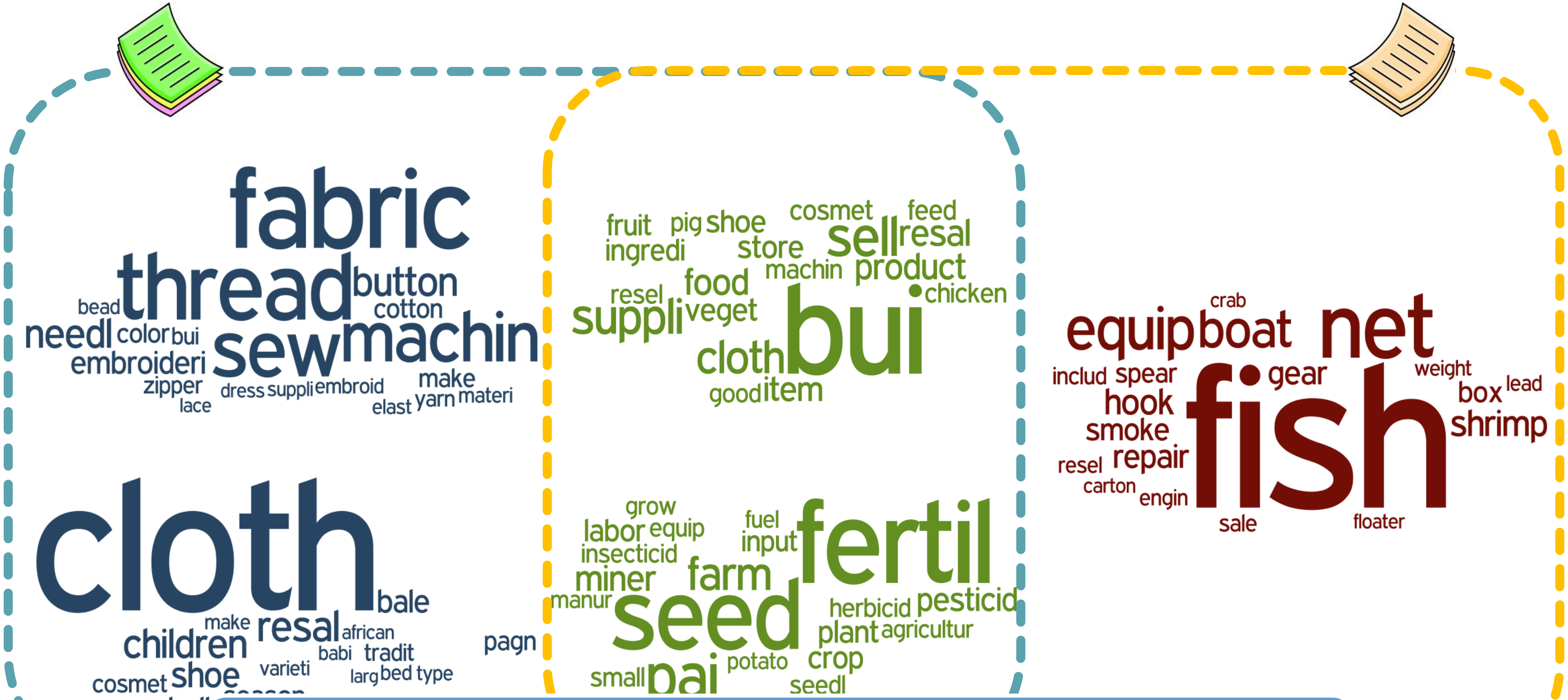
Your funds will be used to backfill this loan
Repayments will go to you

FIELD PARTNER Learn more

fundaciónalternativa

Fundacion Alternativa administers this loan.

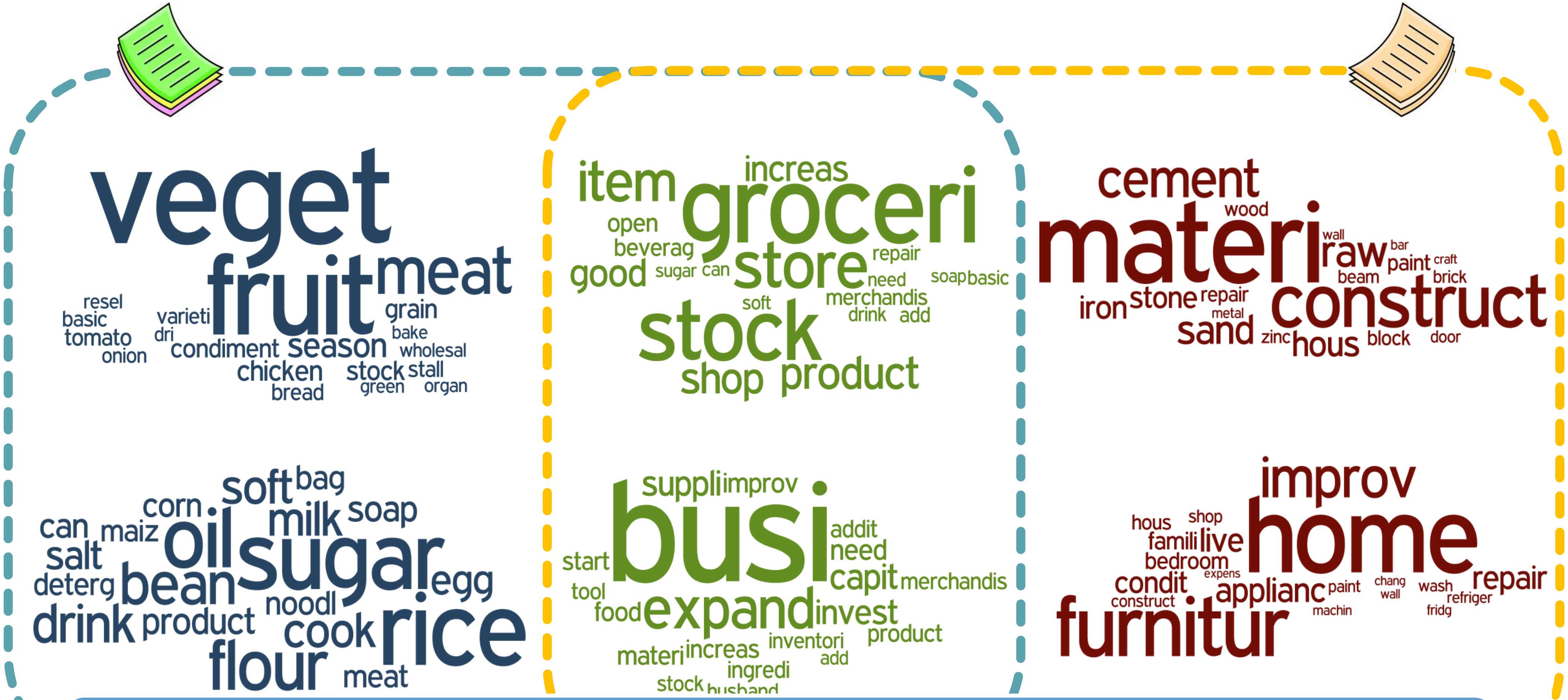
Teams ‘Etsy.com Handmade’ vs. ‘Guys holding fish’



“homophily”

People tend to fund loans similar to what they like.

Teams ‘Thailand’ vs. ‘Greece’

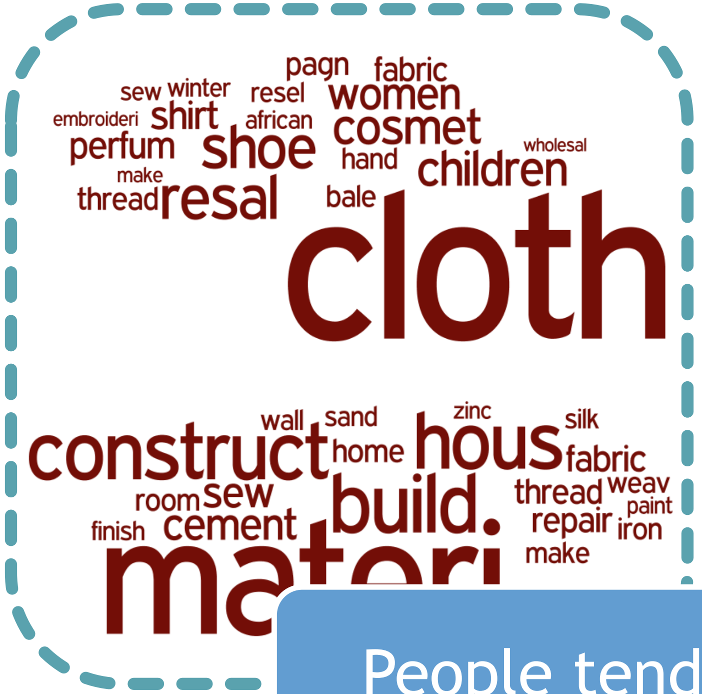


40% of Thai laborers work in agriculture (13% in Greece). Construction and Industrial Manufacturing are big in Greece.

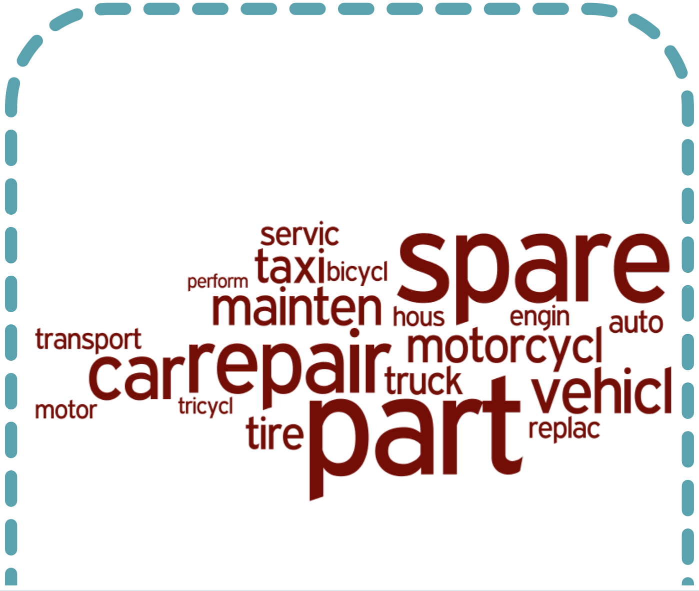
Lender Occupation

- Distinct topics of loans funded by a subset of lenders with the same occupation against the rest

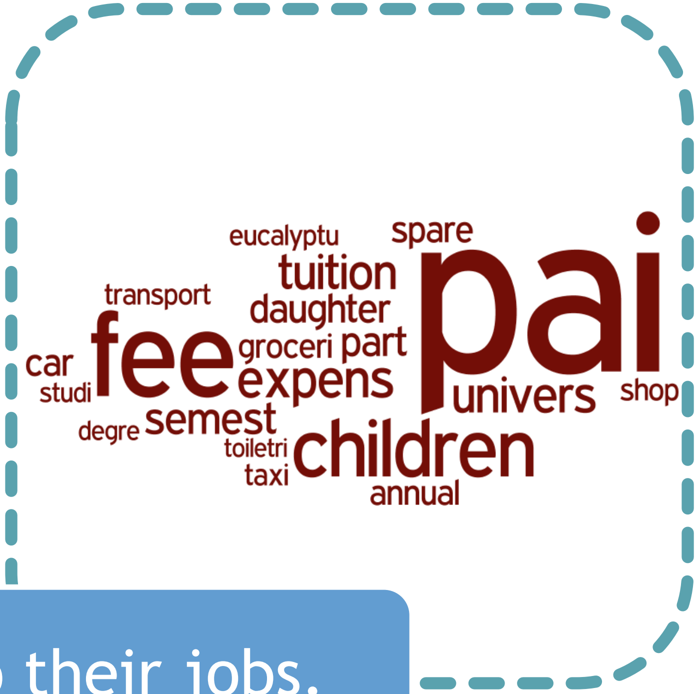
Art vs. non-art



Driver vs. non-driver



Teacher vs. non-teacher



People tend to fund loans related to their jobs.

Conclusion

- We presented a joint NMF-based topic model that identifies common and distinct topics between document sets
- We performed a detailed quantitative analysis as well as in-depth case studies
- We plan to
  - Build a real-time visual analytics system
  - Extend to compare multiple subsets
  - Apply block principal pivoting method

Thank you!

Hannah Kim
hannahkim@gatech.edu