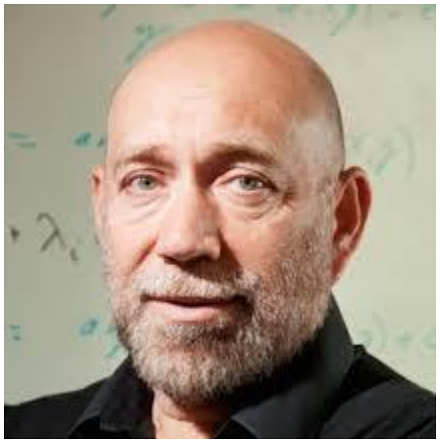
Text Classification without Supervision: Incorporating World Knowledge and Domain Adaptation

Yangqiu Song
Lane Department of CSEE
West Virginia University

Collaborators

Dan Roth | Haixun Wang | Shusen Wang | Weizhu Chen



Text Categorization

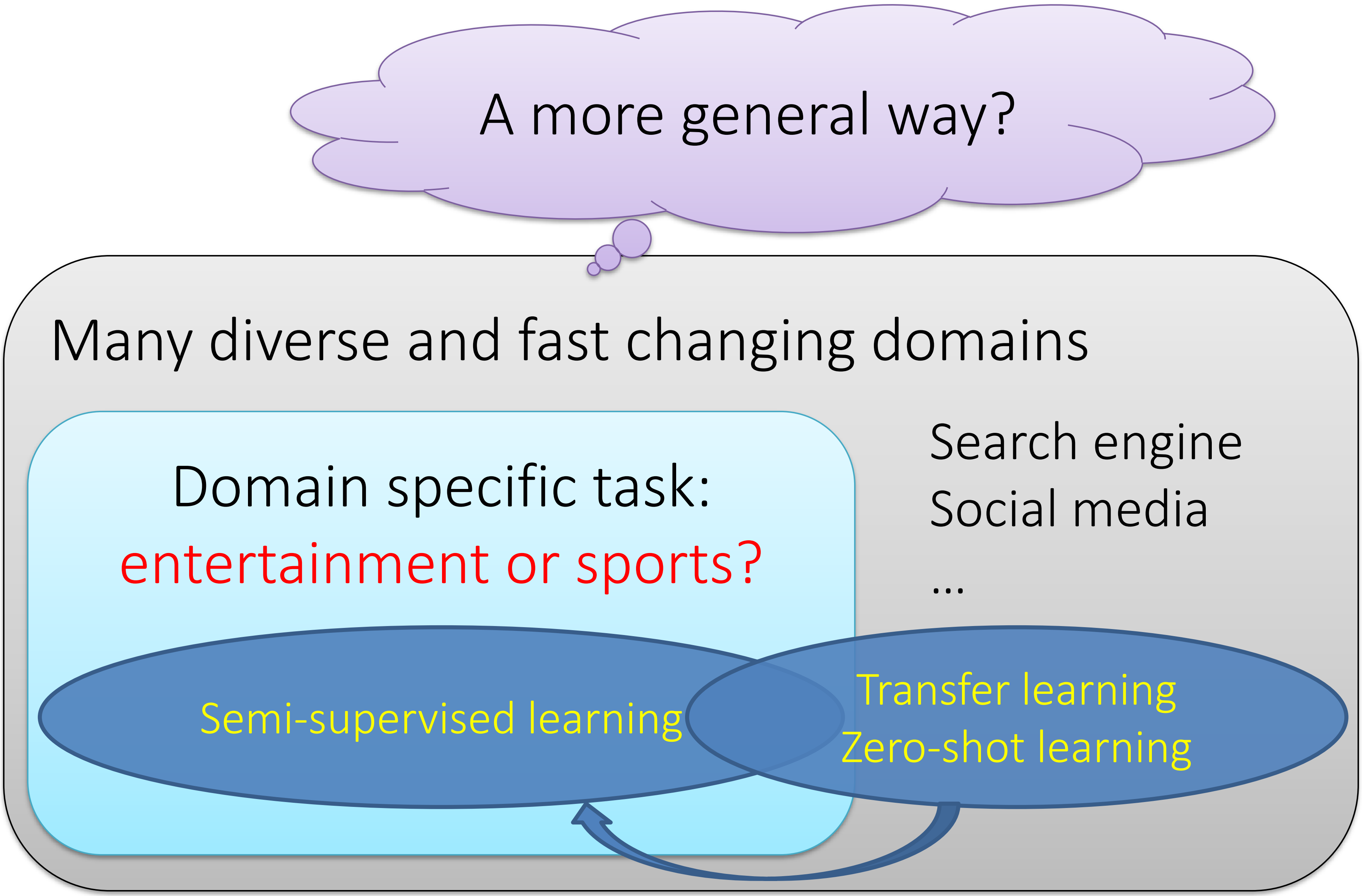
- Traditional machine learning approach:

Label data → Train a classifier → Make prediction

Challenges

- Domain expert annotation
  - Large scale problems
- Diverse domains and tasks
  - Topics
  - Languages
  - …
- Short and noisy texts
  - Tweets,
  - Queries,
  - …

Reduce Labeling Efforts



Our Solution

- Knowledge enabled learning
  - Millions of entities and concepts
  - Billions of relationships

- Labels carry a lot of information!
  - Traditional models treat labels as “numbers or IDs”

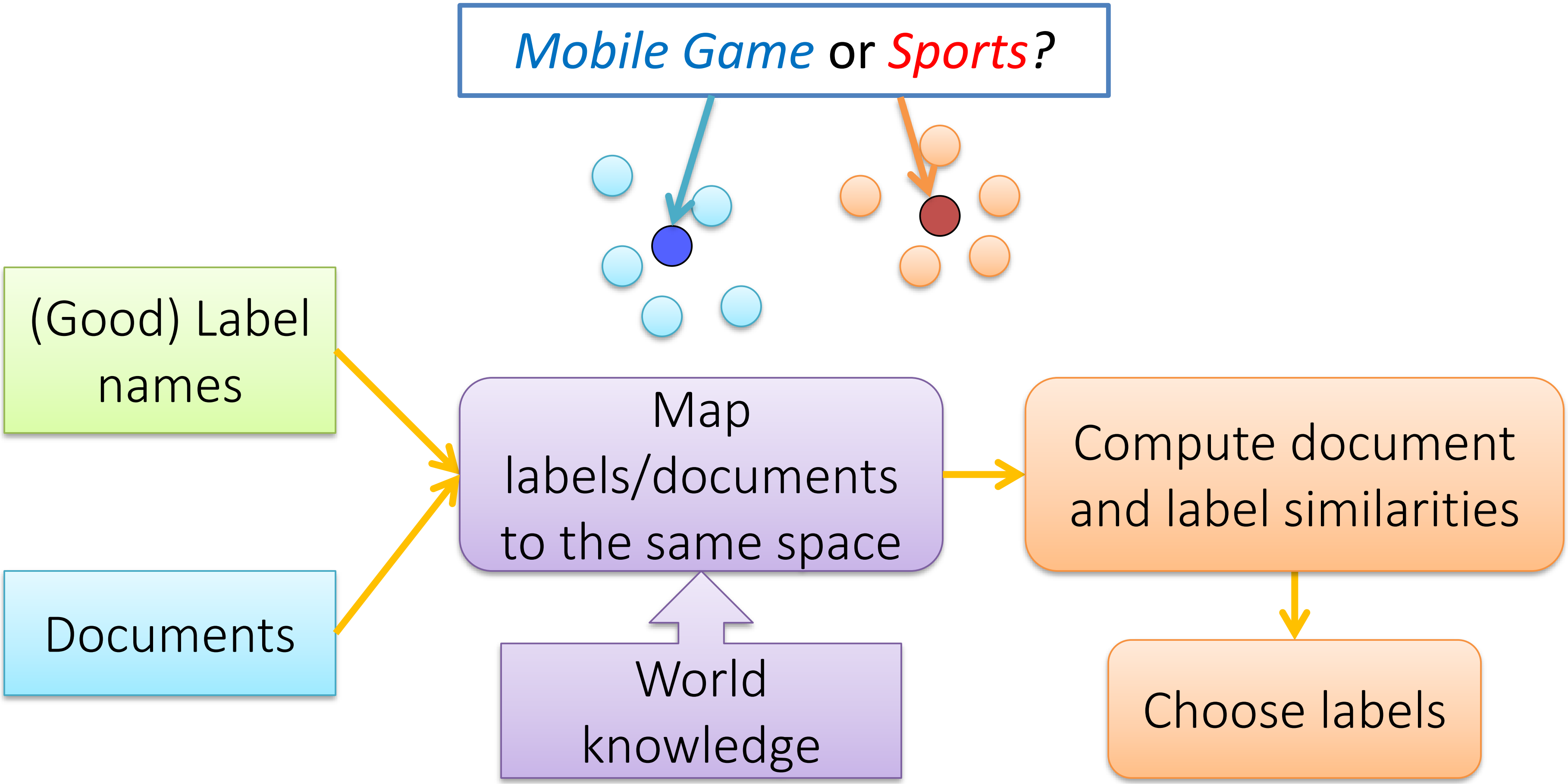
Example:
Knowledge Enabled Text Classification

Dong Nguyen announced that he would be removing his hit game Flappy Bird from both the iOS and Android app stores, saying that the success of the game is something he never wanted. Some fans of the game took it personally, replying that they would either kill Nguyen or kill themselves if he followed through with his decision.

Pick a label:

Mobile Game *or* *Sports*

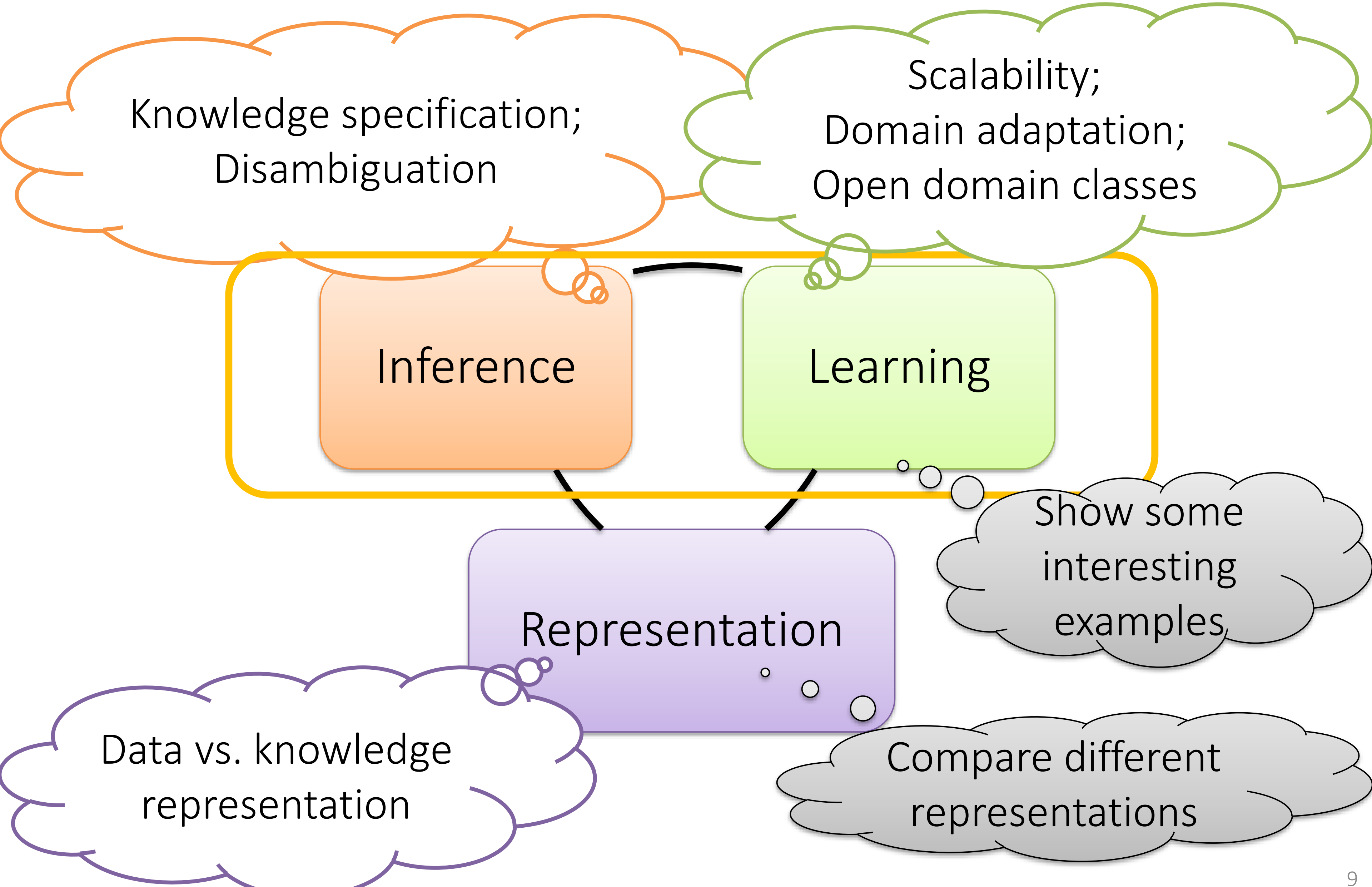
Dataless Text Categorization: Classification on the Fly



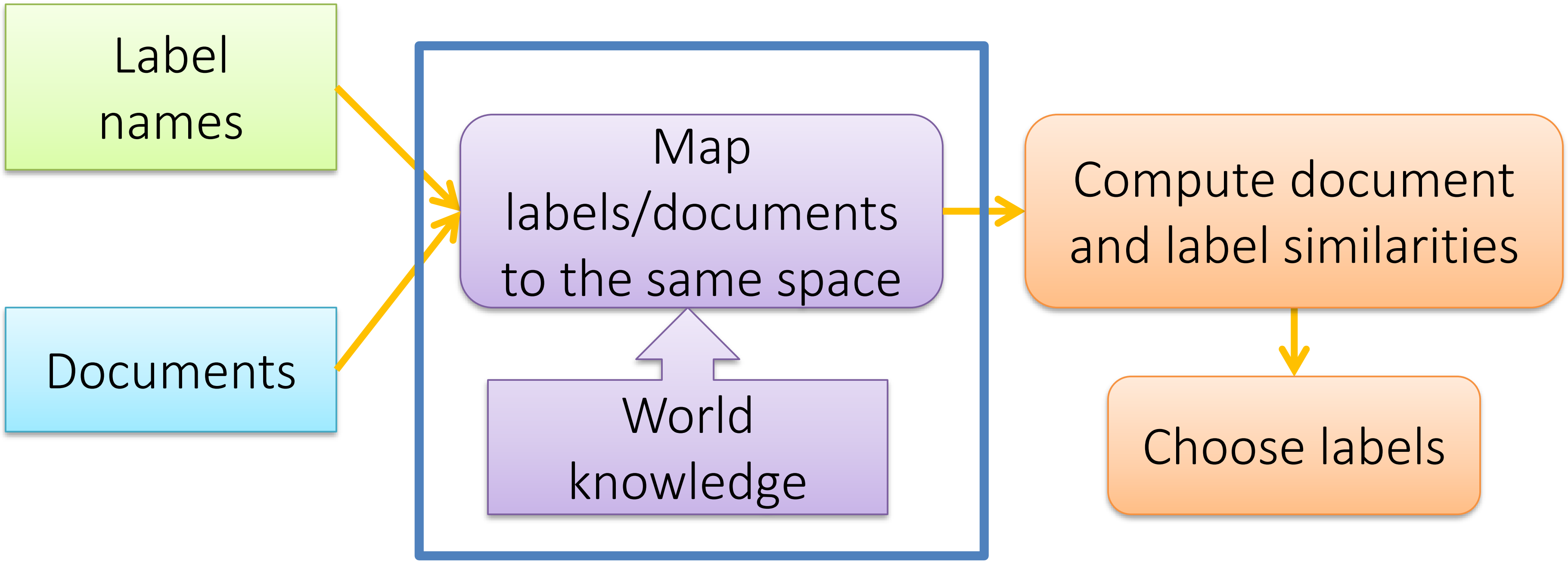
M.-W. Chang, L.-A. Ratinov, D. Roth, V. Srikumar: Importance of Semantic Representation: Dataless Classification. AAAI 2008.

Y. Song, D. Roth: On dataless hierarchical text classification. (AAAI). 2014.

Challenges of Using Knowledge

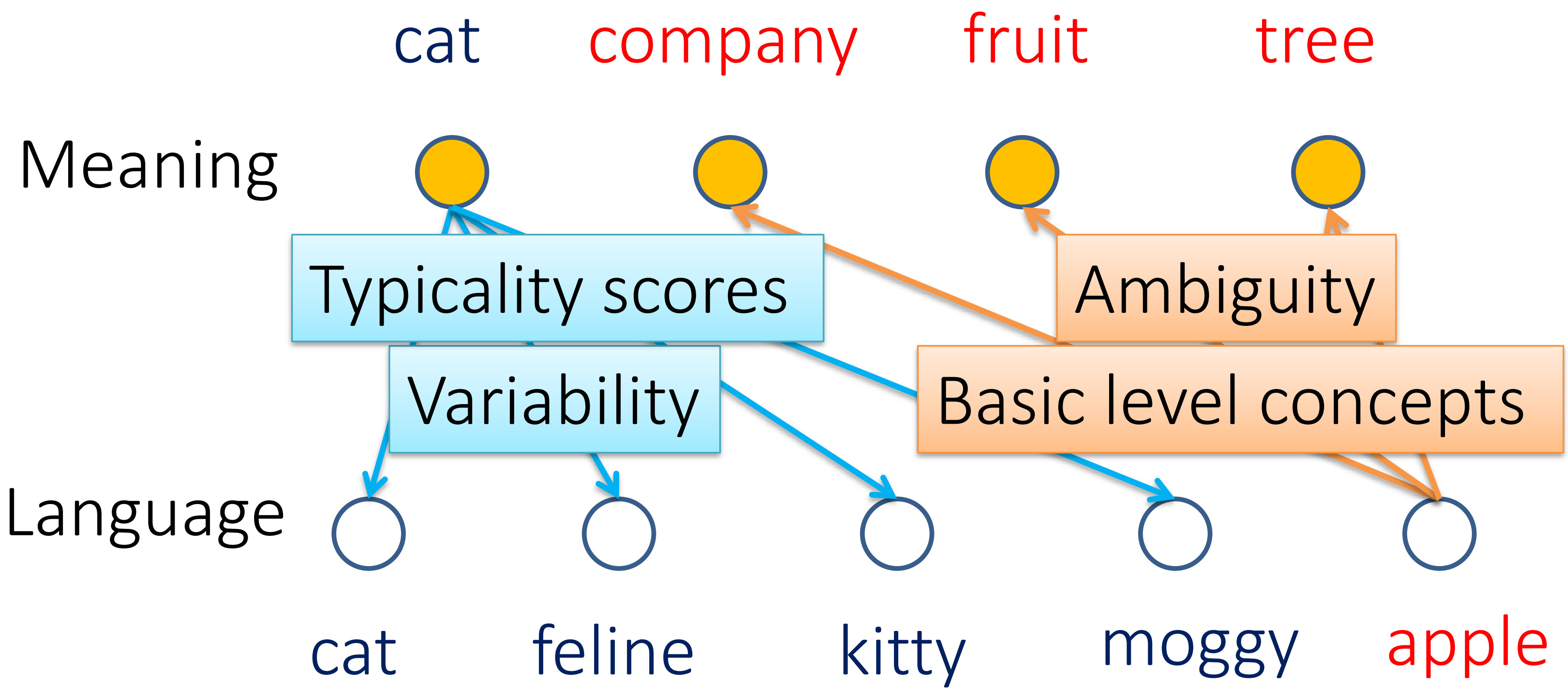


Outline of the Talk
Dataless Text Classification: Classify Documents on the Fly



Difficulty of Text Representation

- Polysemy and Synonym



Rosch, E. et al. Basic objects in natural categories. Cognitive Psychology. 1976.
Rosch, E. Principles of categorization. In Rosch, E., and Lloyd, B., eds., Cognition and Categorization. 1978.

Typicality of Entities

bird

Basic Level Concepts

What do we usually call it?

pug pet $\boxed{\text{dog}}$ mammal animal

pug

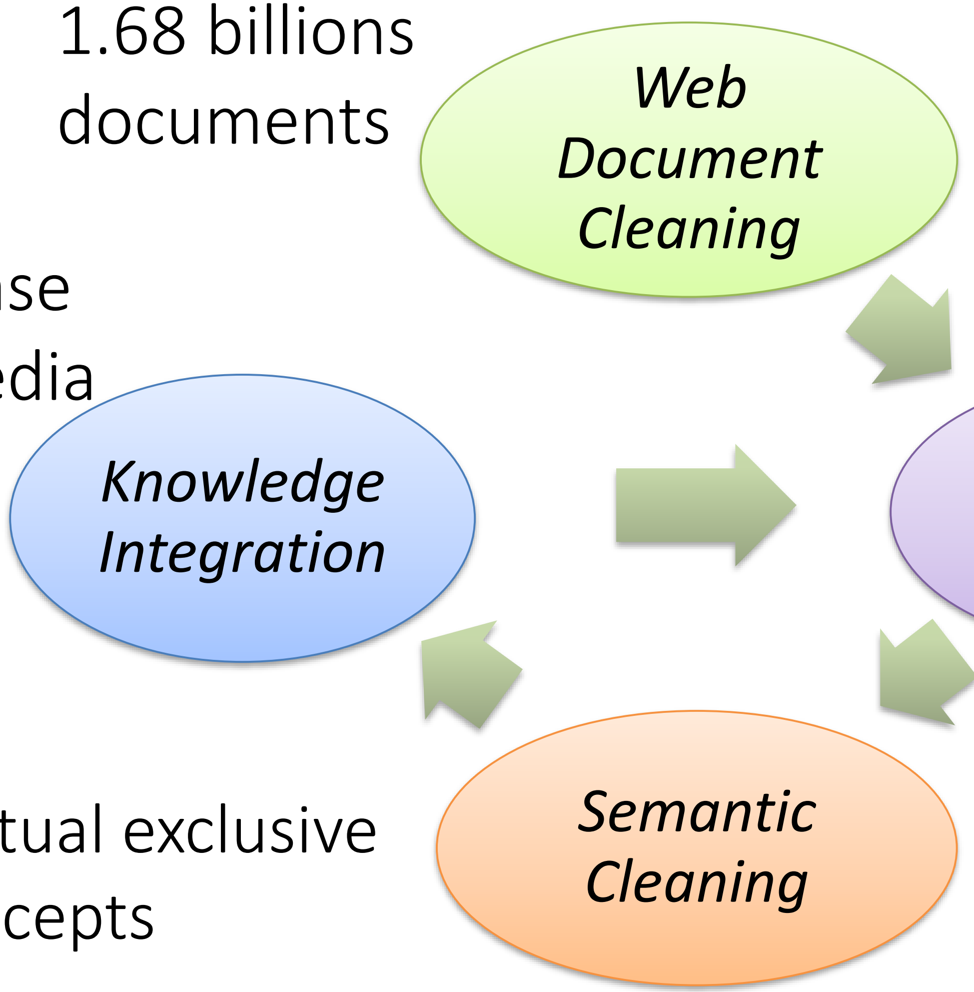
bulldog

We use the right level of concepts to describe things!

Probase: A Probabilistic Knowledge Base

1.68 billions documents

Freebase
Wikipedia



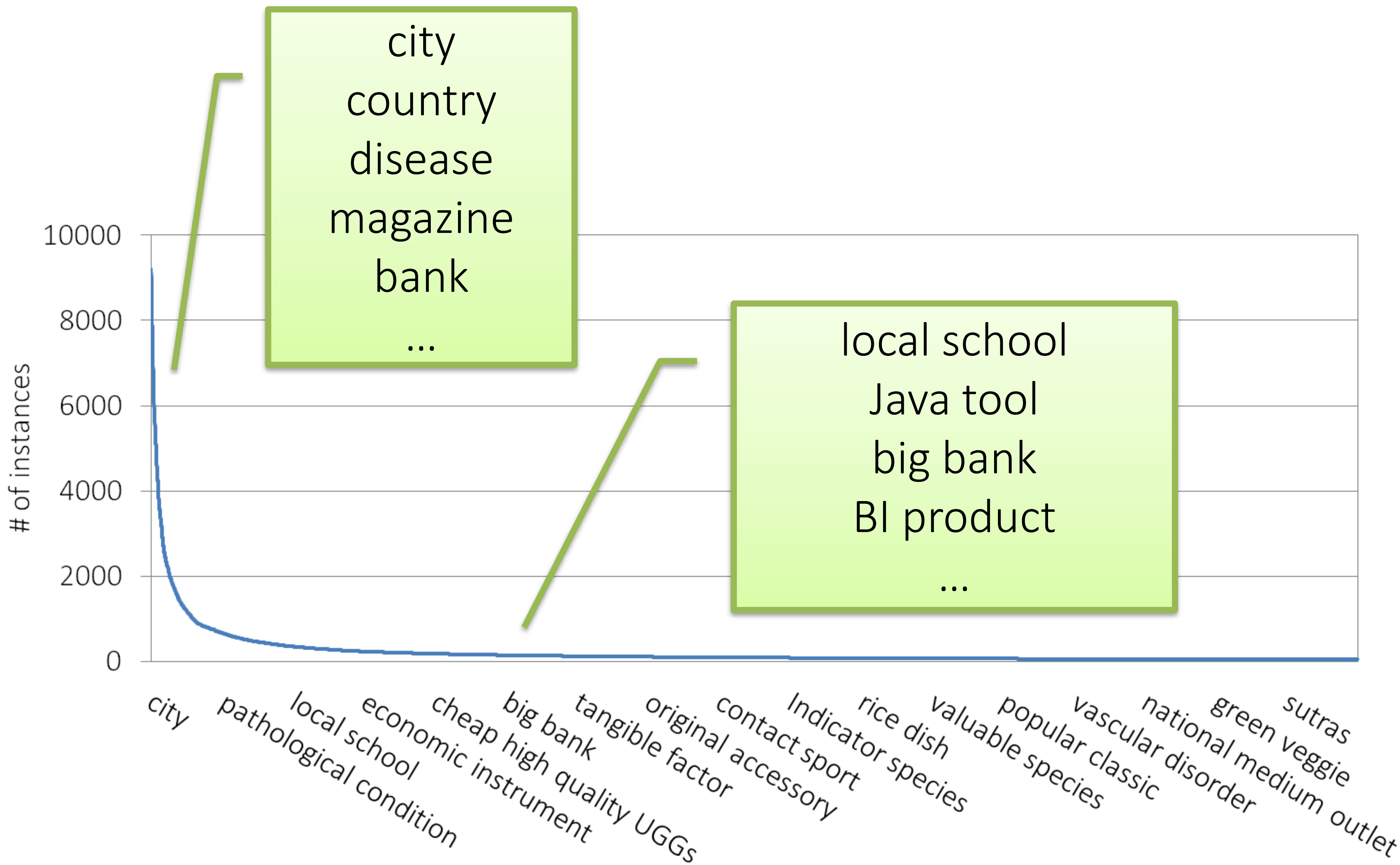
Hearst patterns
"Animals *such as* dogs and cats."

Mutual exclusive concepts

concept	entity	frequency
animal	dog	2111
animal	cat	1613
animal	horse	1288
animal	bird	1190
animal	rabbit	1179
animal	deer	1123
animal	cow	921
animal	sheep	909
animal	goat	897

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. Int. Conf. on Comp. Ling. (COLING).1992.
W. Wu, et al. Probase: A probabilistic taxonomy for text understanding. In ACM SIG on Management of Data (SIGMOD). 2012. (**Data released http://probase.msra.cn**)

Concept Distribution

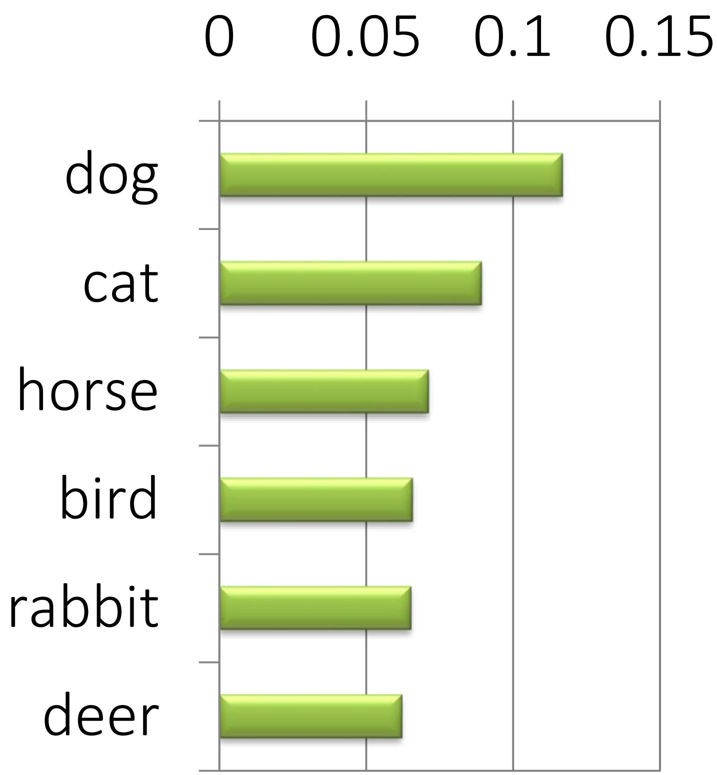


Distribution of Concepts

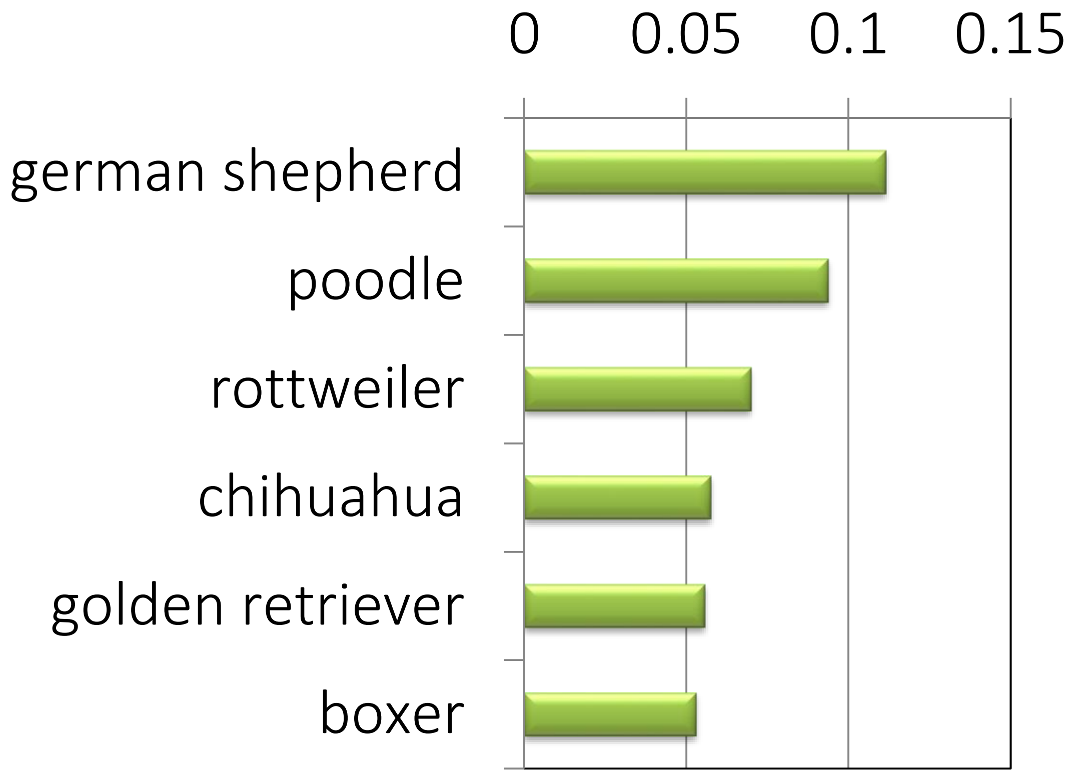
Typicality

$$P(\text{entity} \mid \text{concept}) = \frac{n(\text{entity}, \text{concept})}{n(\text{concept})}$$

- Animal



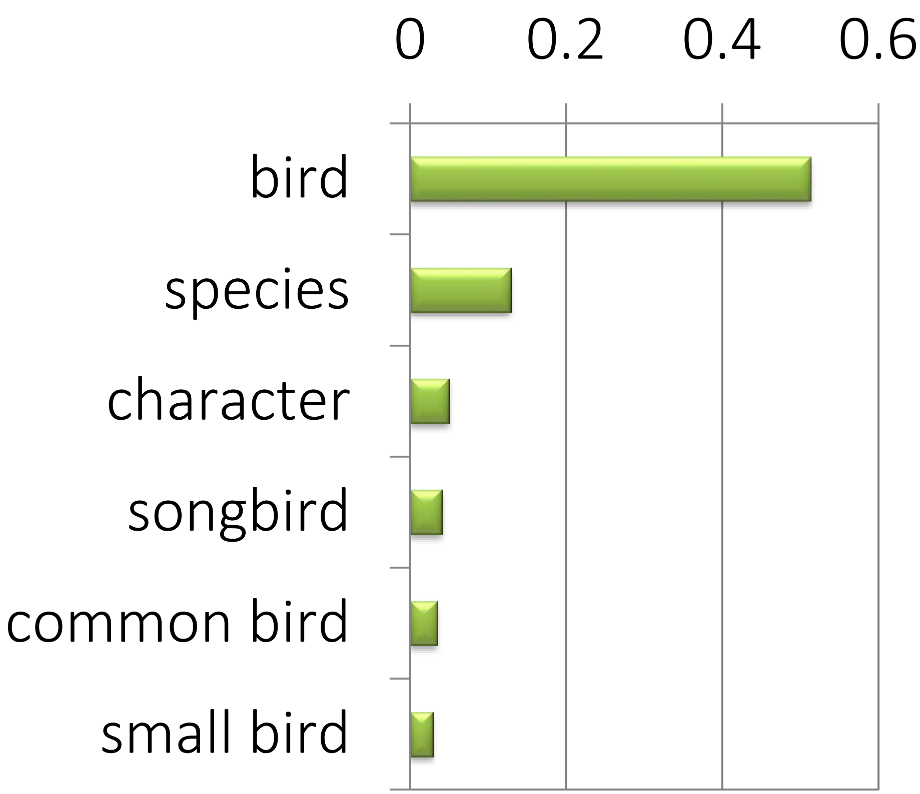
- Dog



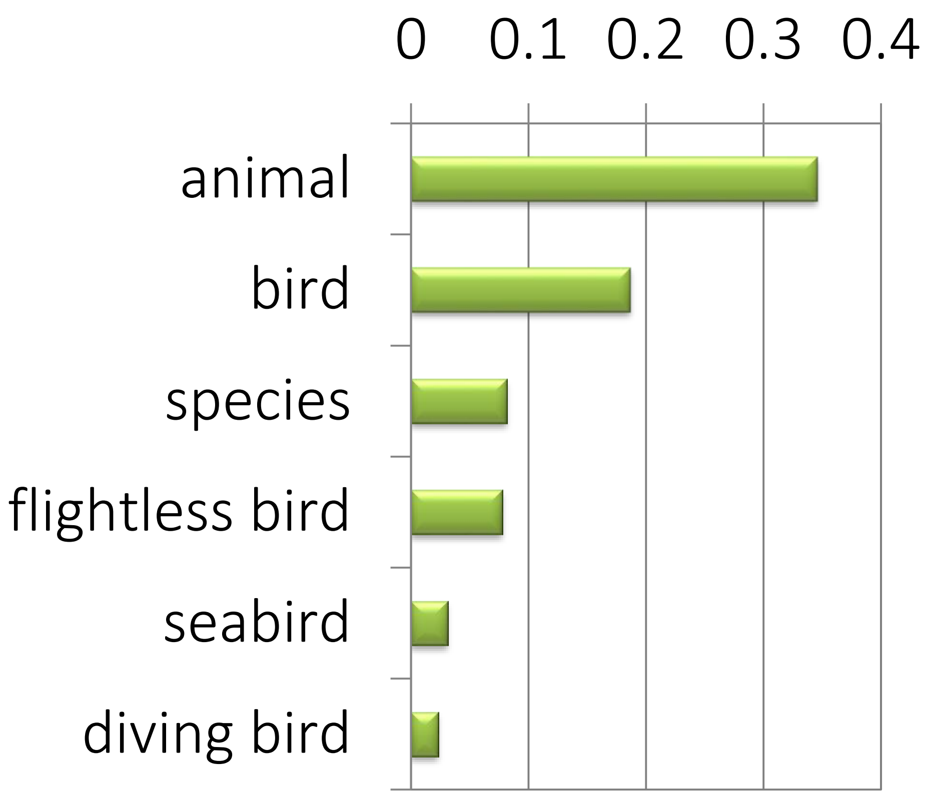
Basic Level Concepts

$$P(\text{concept} \mid \text{entity}) = \frac{n(\text{entity}, \text{concept})}{n(\text{entity})}$$

- Robin



- Penguin



Concepts of Multiple Entities

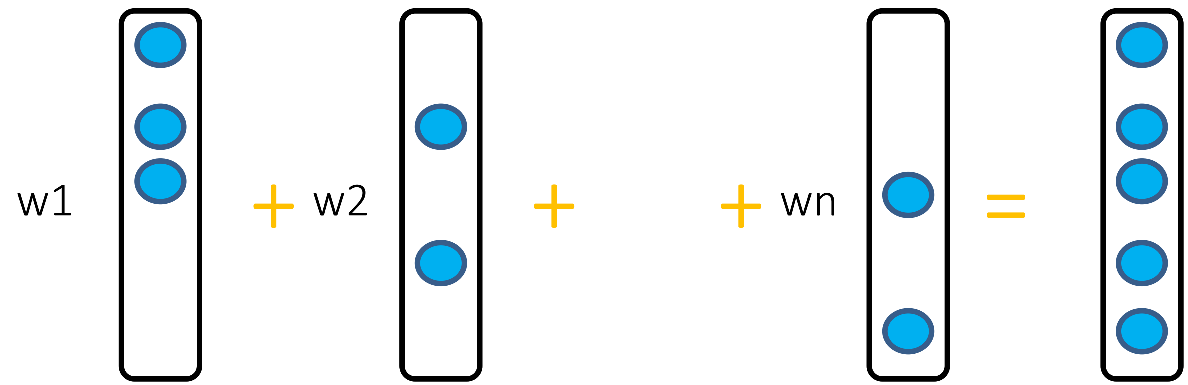
Obama’s real-estate policy

president, politician

investment, property, asset, plan

president, politician, investment, property, asset, plan

Explicit Semantic Analysis (ESA)



E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. J. of Art. Intell. Res. (JAIR). 2009.

Multiple Related Entities

apple

adobe

software company, brand, fruit

brand, software company

software company, brand, fruit
software company, brand

Intersection instead of union!

Probabilistic Conceptualization

P(concept | related entities)

P(fruit | adobe, apple) = 0

P(adobe | fruit) = 0

$$P(c_k \mid E) = \frac{P(E \mid c_k)P(c_k)}{P(E)} \propto P(c_k)\prod_{i=1}^{M} P(e_i \mid c_k)$$

$$E = \{e_i \mid i = 1, \ldots, M\}$$

$$P(e_i \mid c_k) = \frac{P(e_i, c_k)}{P(c_k)}$$

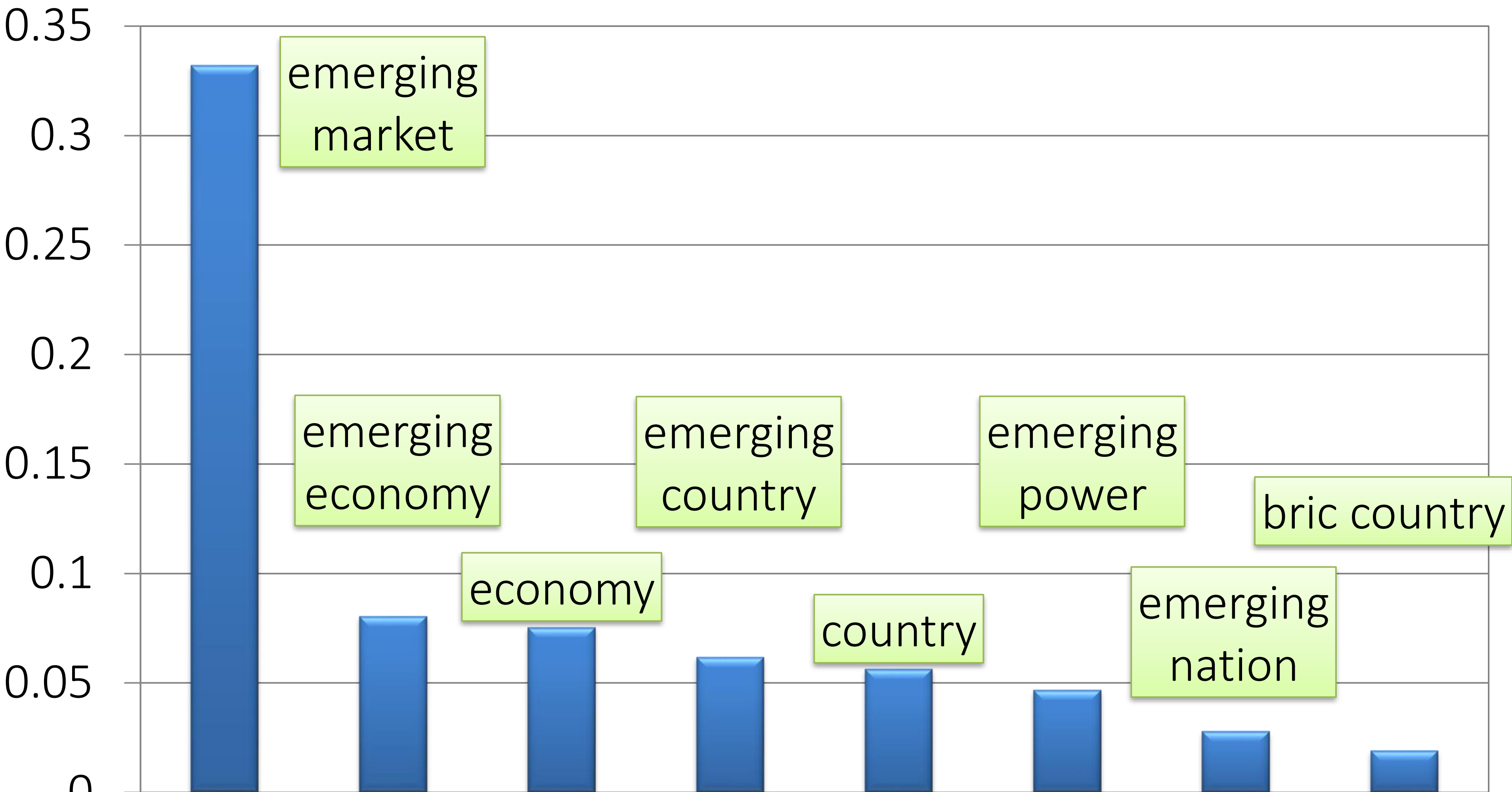
$$P(c_k)$$

Basic Level Concept

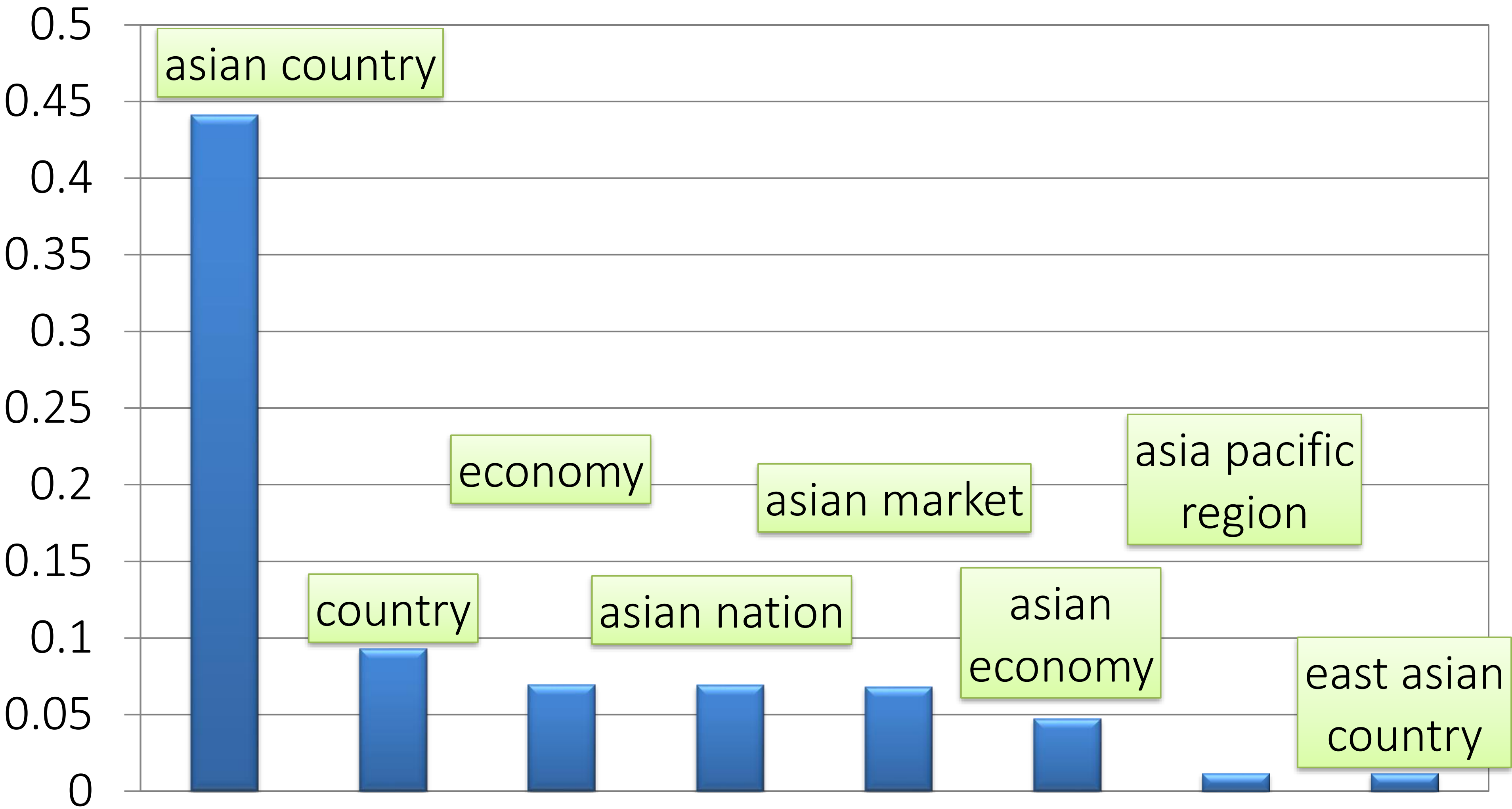
Typicality

Song et al., Int. Joint Conf. on Artif. Intell. (IJCAI). 2011.

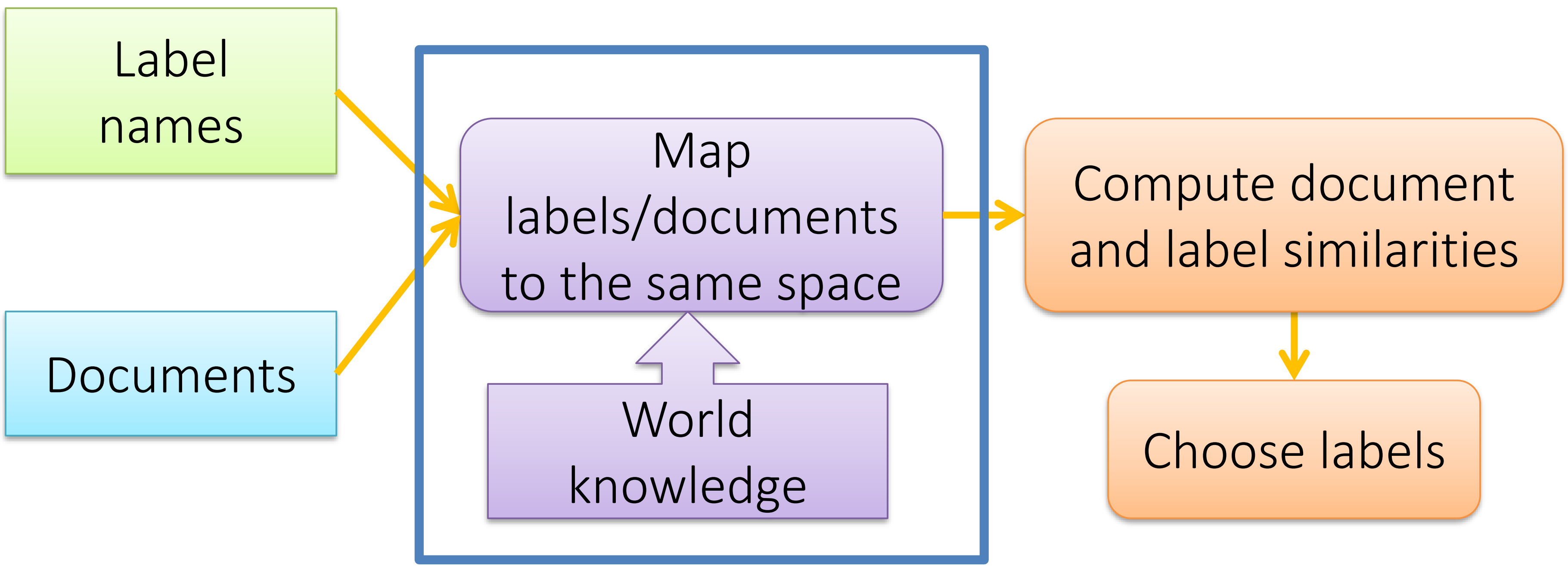
Given “China, India, Russia, Brazil”



Given “China, India, Japan, Singapore”



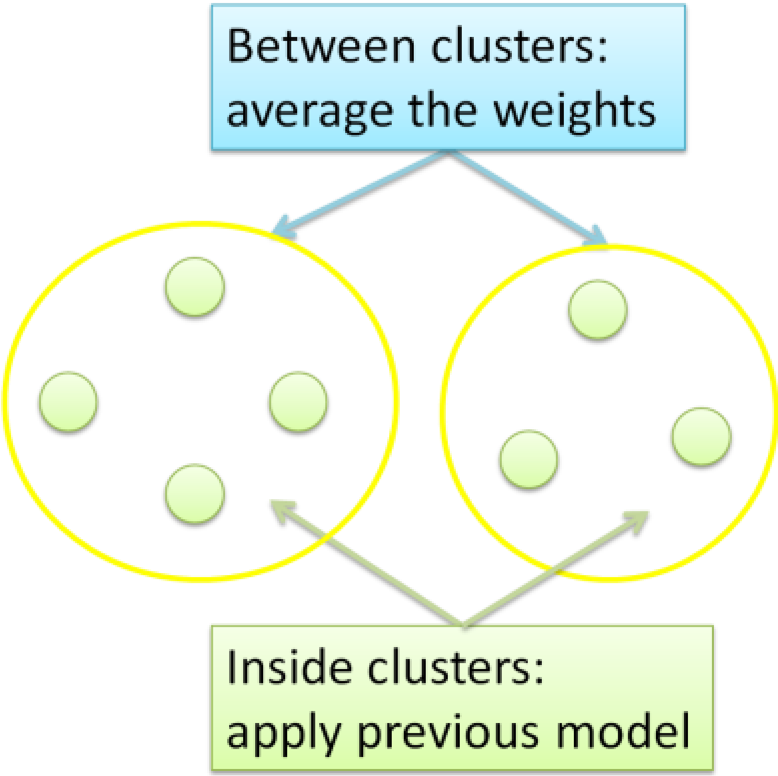
Outline of the Talk
Dataless Text Classification: Classify Documents on the Fly



Generic Short Text Conceptualization

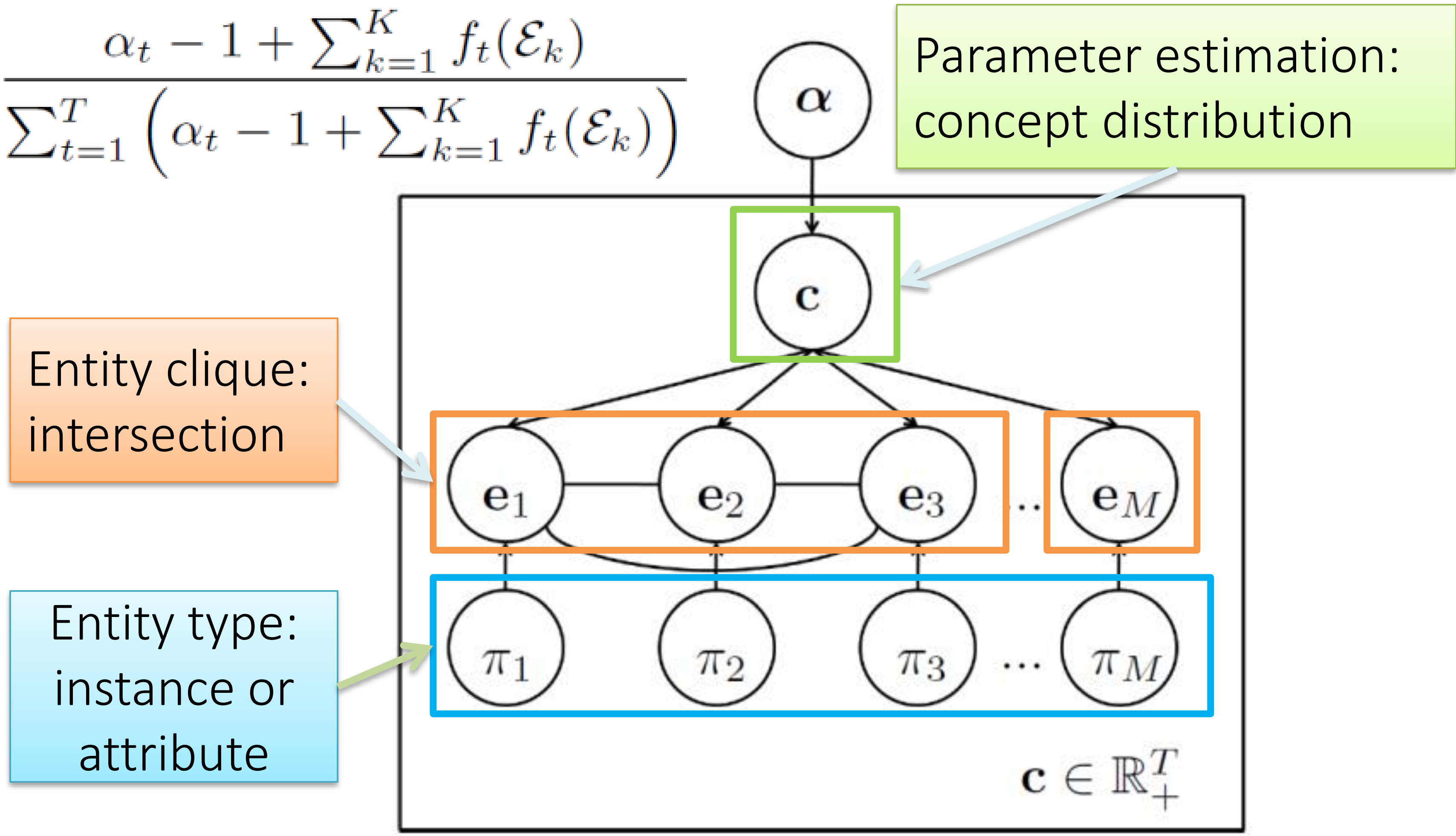
P(concept | short text)

1. Grounding to knowledge base
2. Clustering entities
3. Inside clusters: intersection
4. Between clusters: union



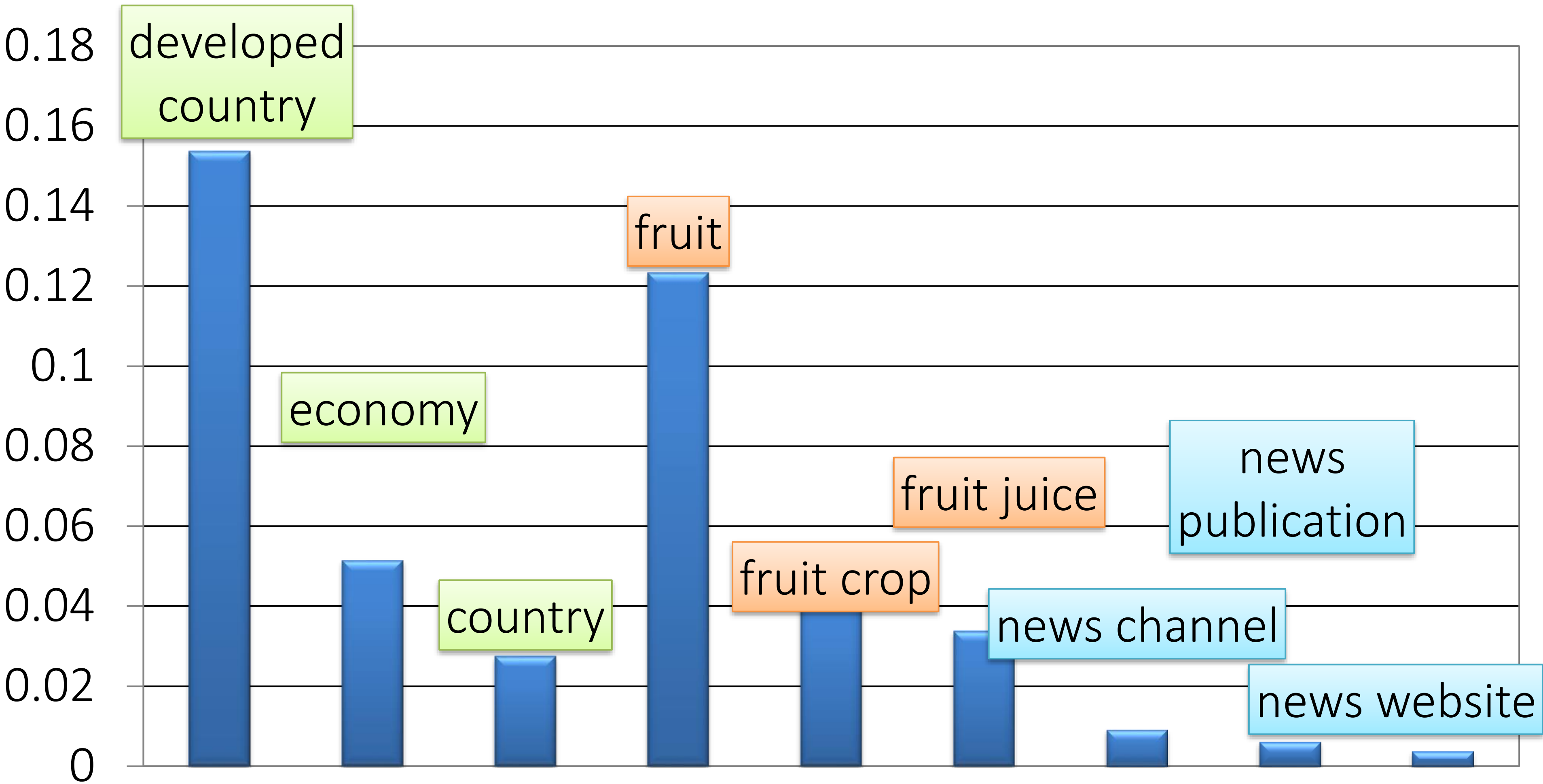
Markov Random Field Model

$$c_t^{\mathrm{opt}} = \frac{\alpha_t - 1 + \sum_{k=1}^{K} f_t(\mathcal{E}_k)}{\sum_{t=1}^{T} \left(\alpha_t - 1 + \sum_{k=1}^{K} f_t(\mathcal{E}_k) \right)}$$

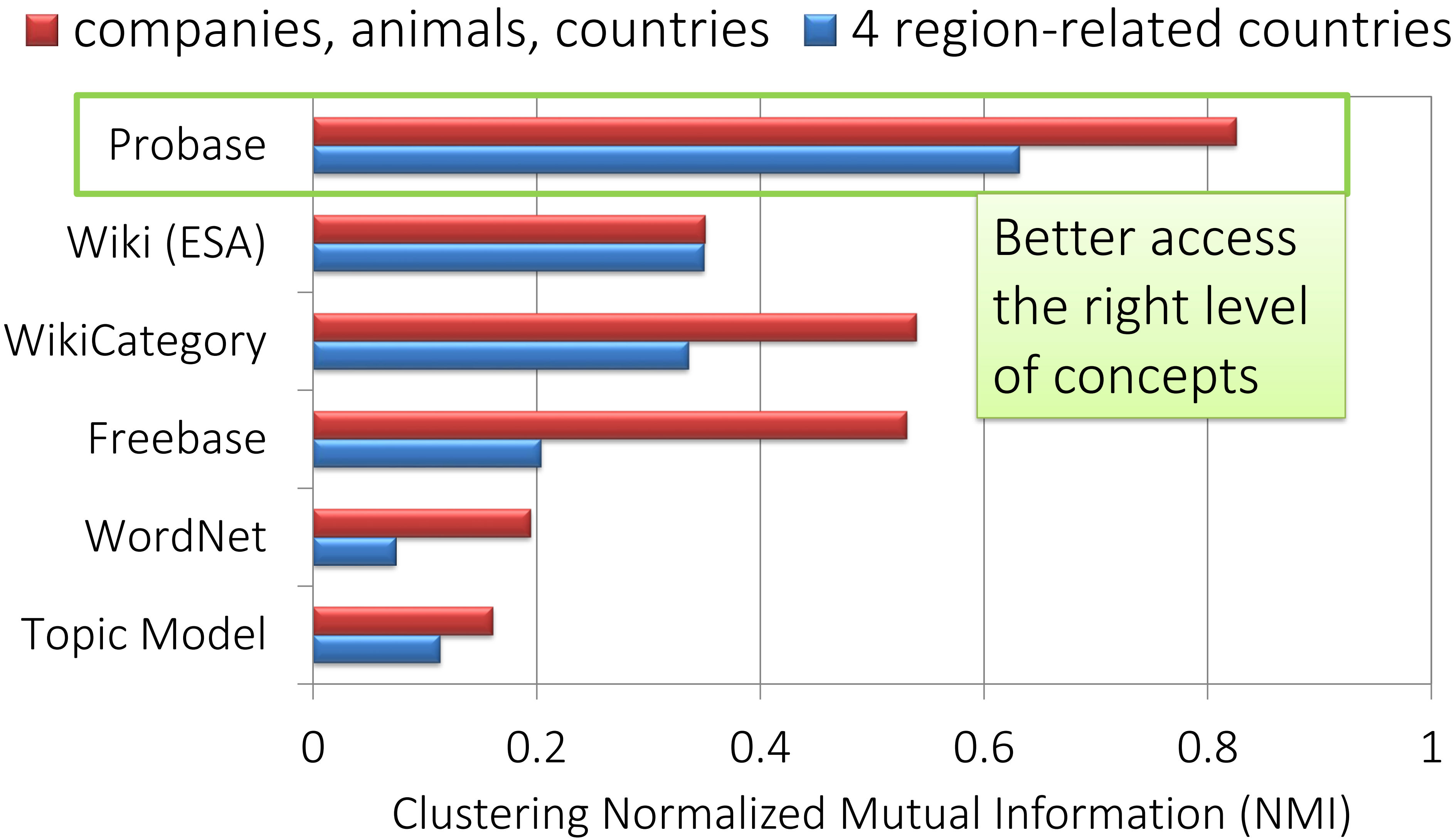


$$P_\Phi(\boldsymbol{\alpha}, \mathbf{c}, \{\mathbf{e}_m\}_{m=1}^{M}, \{\pi_m\}_{m=1}^{M}) = \frac{1}{Z} \phi(\boldsymbol{\alpha}, \mathbf{c}) \prod_{k=1}^{K} \phi(\mathcal{E}_k, \mathbf{c})$$

Given “U.S.”, “Japan”, “U.K.”; “apple”, “pear”; “BBC”, ”New York Time”



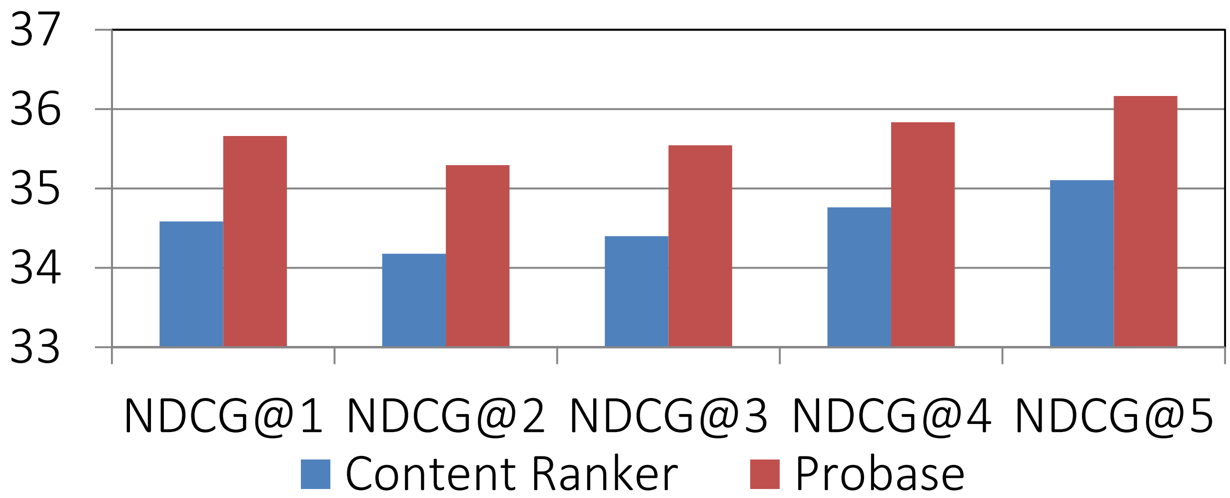
Tweet Clustering

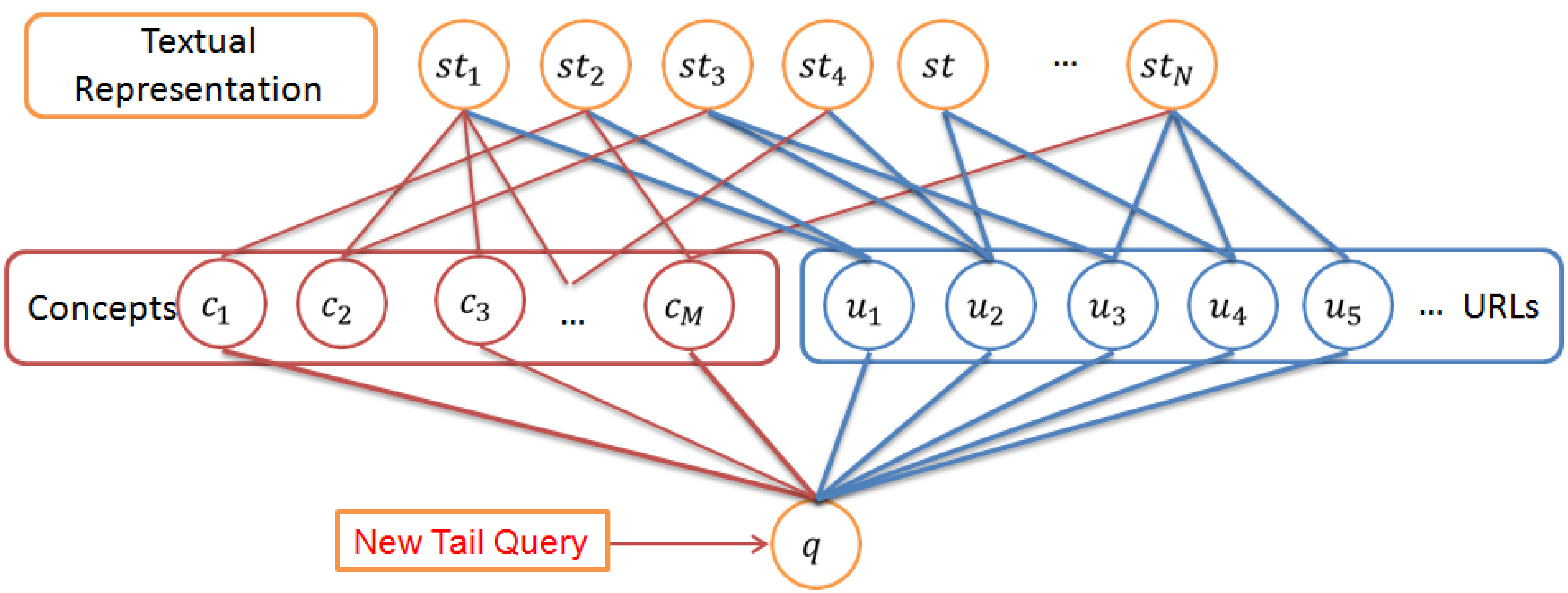


Song et al., Int. Joint Conf. on Artif. Intell. (IJCAI). 2011.

Web Search Relevance

- Evaluation data:
  - 300K Web queries
  - 19M query-URL pairs
- Historical data:
  - 8M URLs
  - 8B query-URL clicks





Song et al., Int. Conf. on Inf. and Knowl. Man. (CIKM). 2014.

Domain Adaptation

- World knowledge bases
  - General purpose
  - Information bias
- Domain dependent tasks
  - E.g., classification/clustering of entertainment vs. sports
  - Knowledge about science/technology is useless

Domain Adaptation for Corpus

$$\alpha_t^{\text{new}} \leftarrow \frac{\alpha_t \boxed{\sum_{n=1}^{N} \left(\Psi\left(\alpha_t + \sum_{k=1}^{K^{(n)}} f_t(\mathcal{E}_k^{(n)})\right) - \Psi(\alpha_t) \right)}}{\sum_{n=1}^{N} \left(\Psi\left(\sum_{t=1}^{T} (\alpha_t + \sum_{k=1}^{K^{(n)}} f_t(\mathcal{E}_k^{(n)})) \right) - \Psi(\sum_{t=1}^{T} \alpha_t) \right)}$$

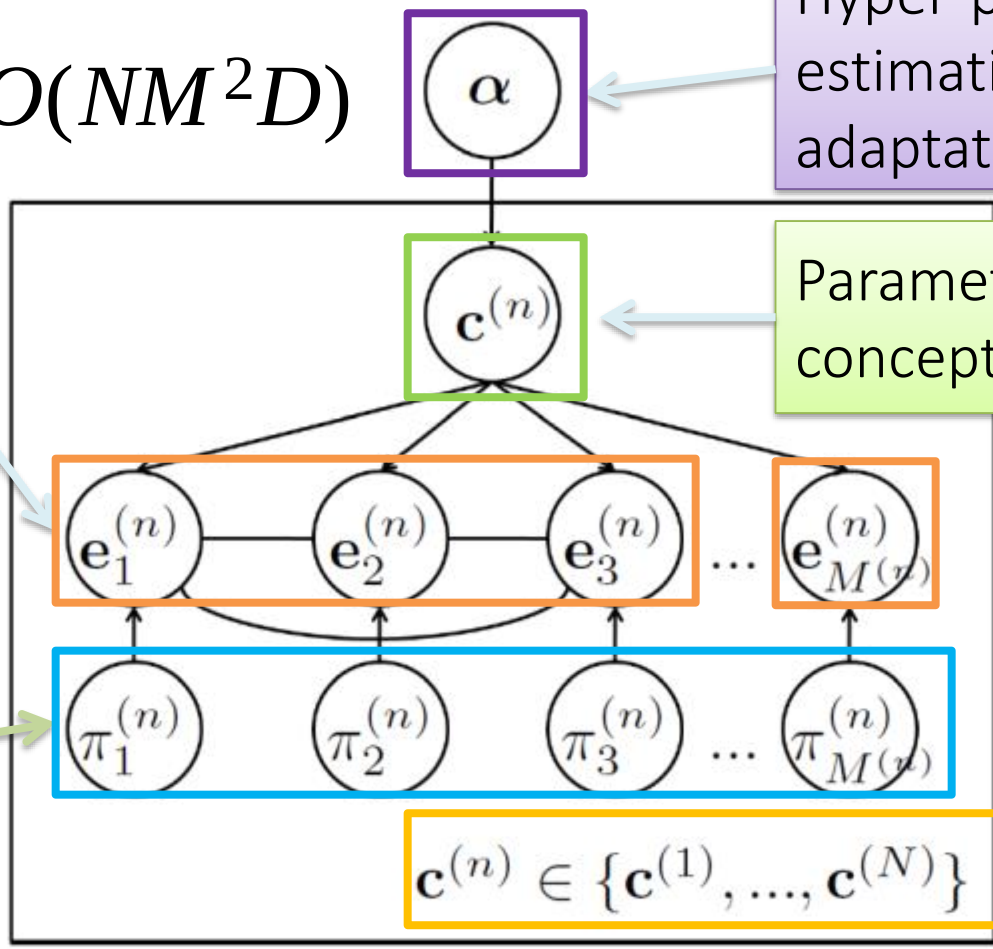
Complexity: $O(NM^2D)$

Hyper-parameter estimation: domain adaptation

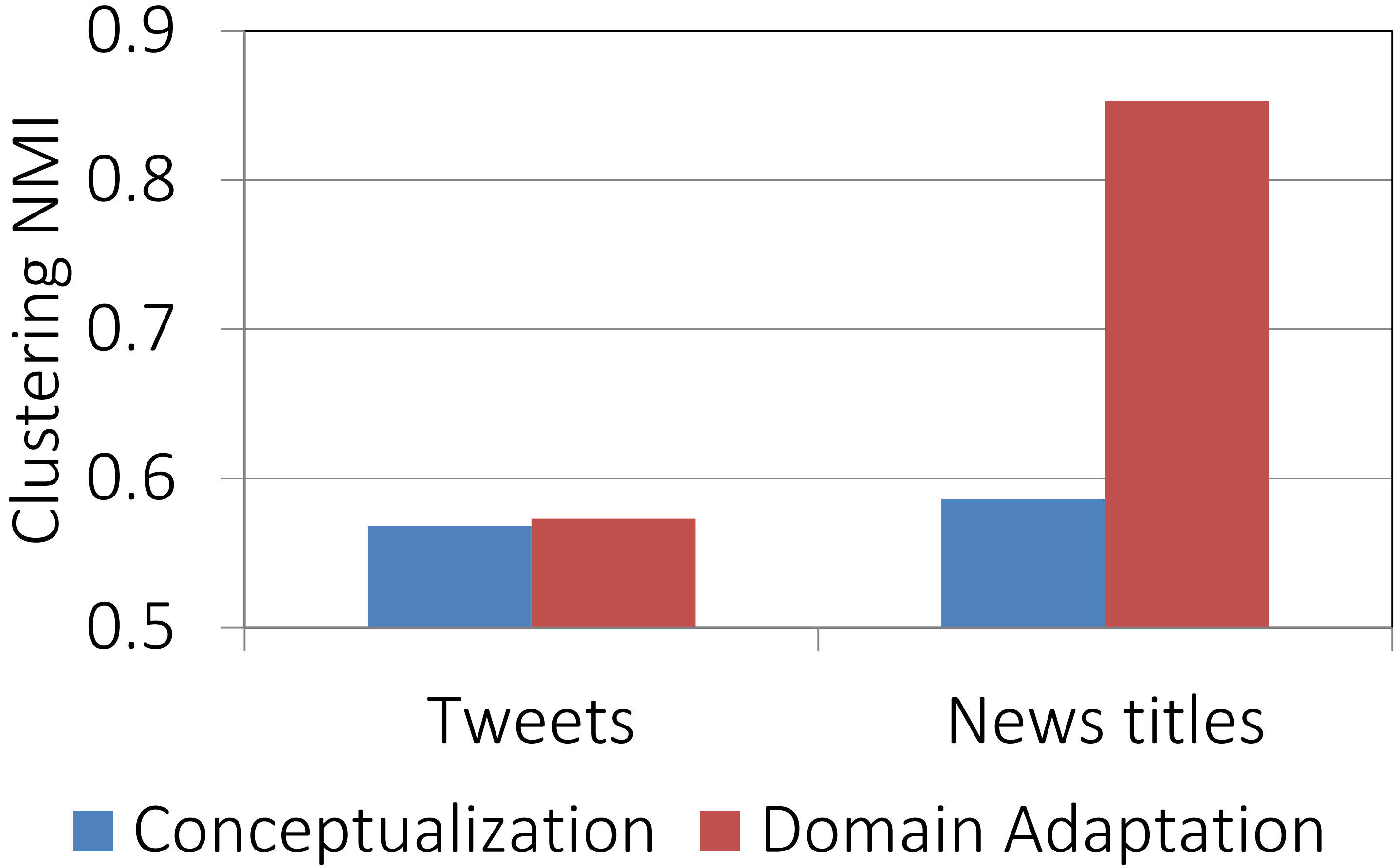
Parameter estimation: concept distribution

Entity clique: intersection

Entity type: instance or attribute



Domain Adaptation Results



Song et al., Int. Joint Conf. on Artif. Intell. (IJCAI). 2015.

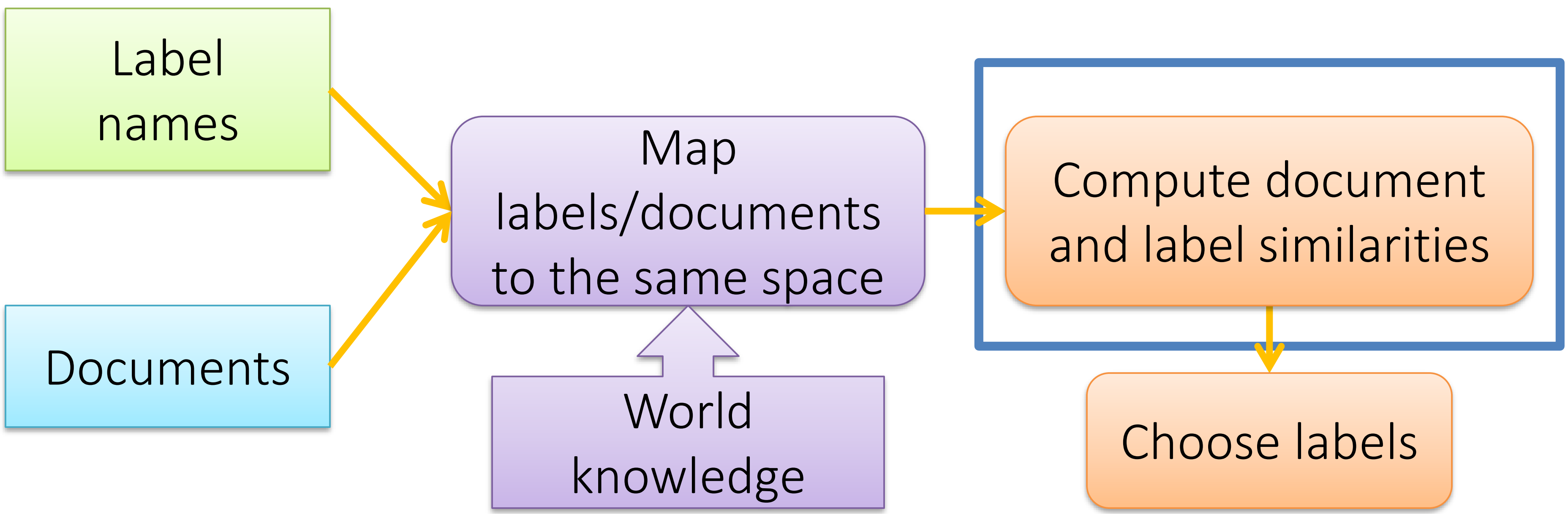
Similarity and Relatedness

- Similarity
  - a specific type of relatedness
  - synonyms, hyponyms/hypernyms, and siblings are highly similar
    - doctor vs. surgeon, bike vs. bicycle

- Relatedness
  - topically related or based on any other semantic relation
    - heart vs. surgeon, tire vs. car

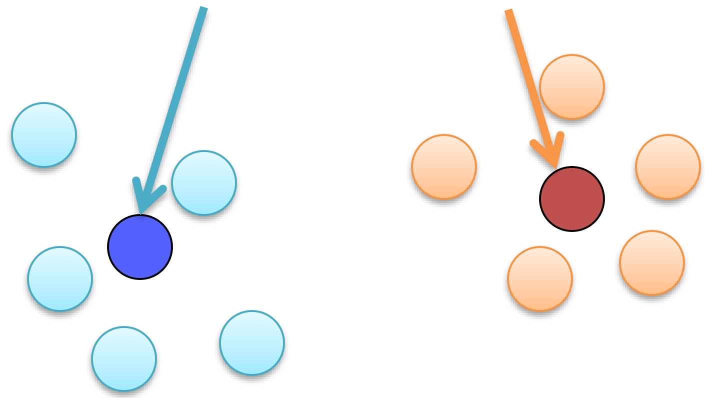
- In the following, we focus on Wikipedia!
  - The methodologies apply
    - Entity relatedness
    - Domain adaptation

Dataless Text Classification: Classify Documents on the Fly



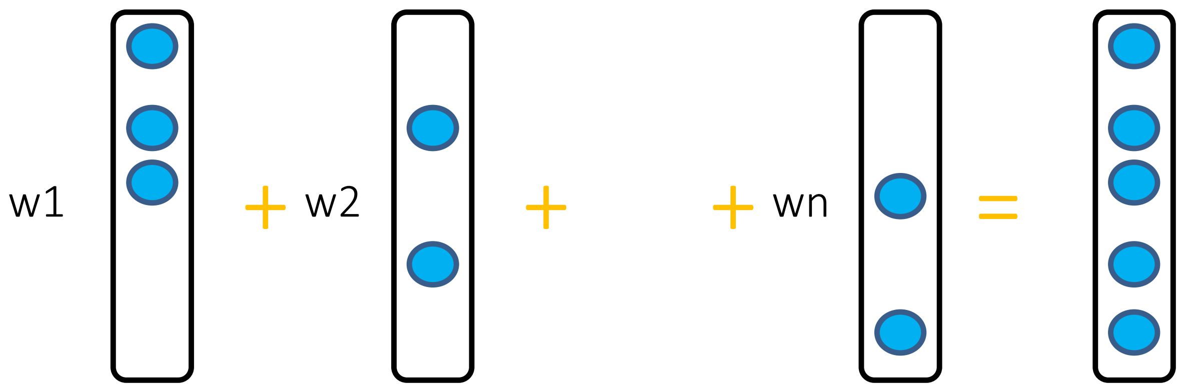
Classification in the Same Semantic Space

Mobile Game or *Sports*?



$$l = \arg\min_{l=l_i} Dist(\phi(x), \phi(l_i))$$

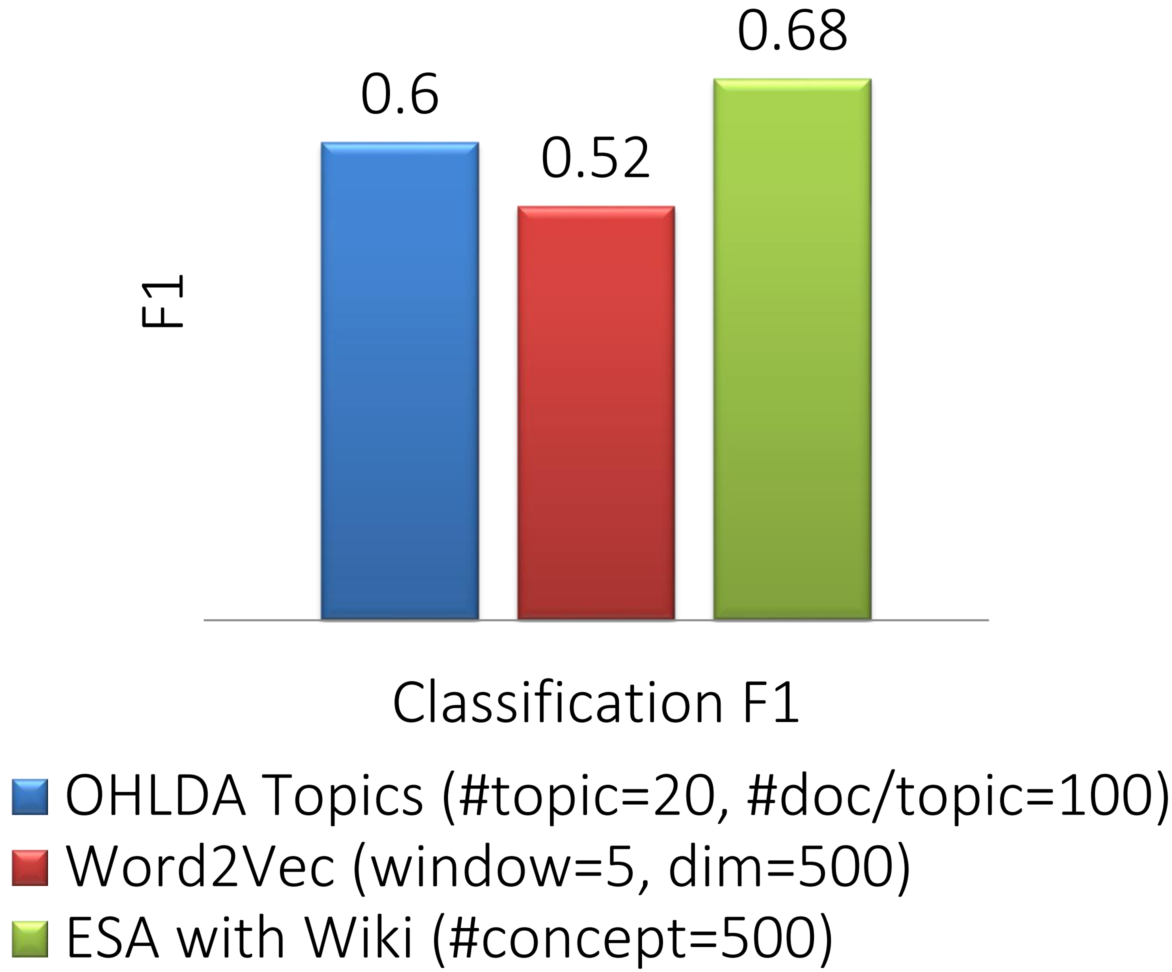
Explicit Semantic Analysis (ESA)



E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. J. of Art. Intell. Res. (JAIR). 2009.

Classification of 20 Newsgroups Documents: Cosine Similarity

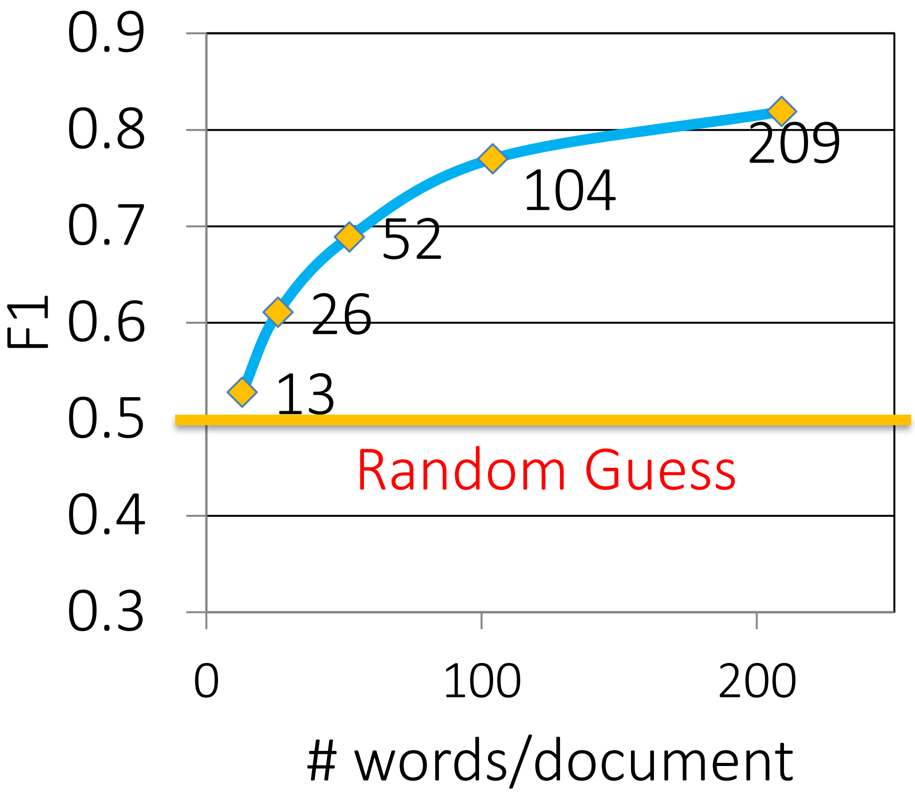
- 20 newsgroups
  - L1: 6 classes
  - L2: 20 classes
- OHLDA:
  - Same hierarchy
- Word2vec
  - Trained on wiki
  - Skipgram



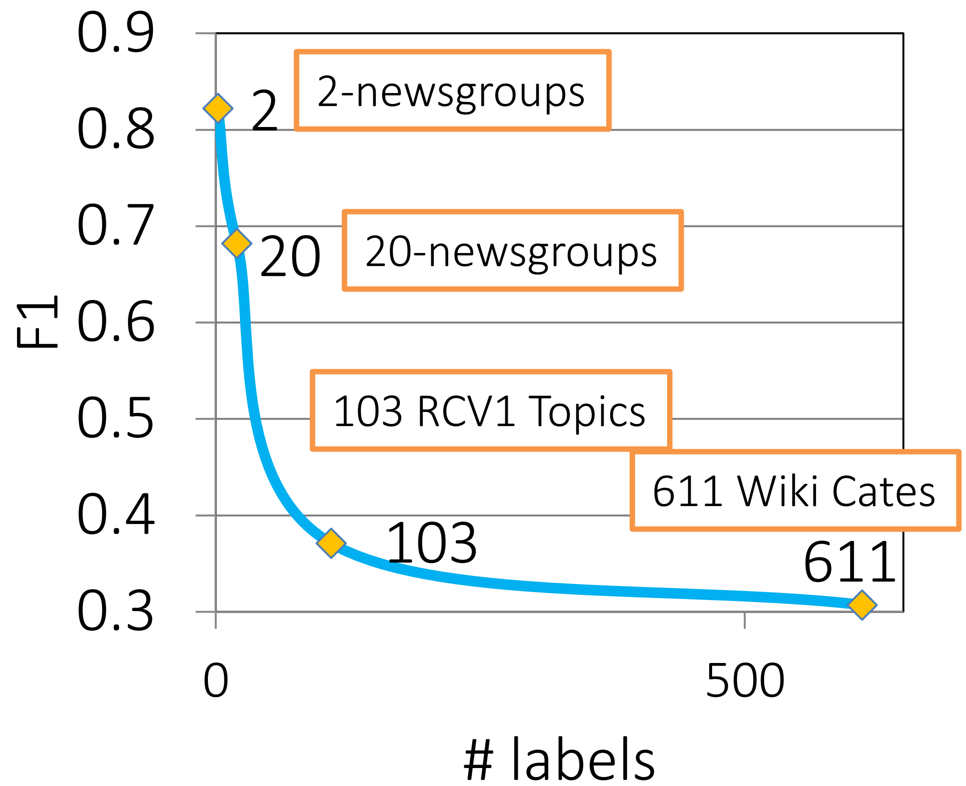
V.Ha-Thuc, and J.-M. Renders, Large-scale hierarchical text classification without labelled data. In WSDM 2011.
Blei et al., Latent Dirichlet Allocation. J. of Mach. Learn. Res. (JMLR). 2003.
Mikolov et al. Efficient Estimation of Word Representations in Vector Space. NIPS. 2013.

Two Factors in Dataless Classification

- Length of document
- Number of labels



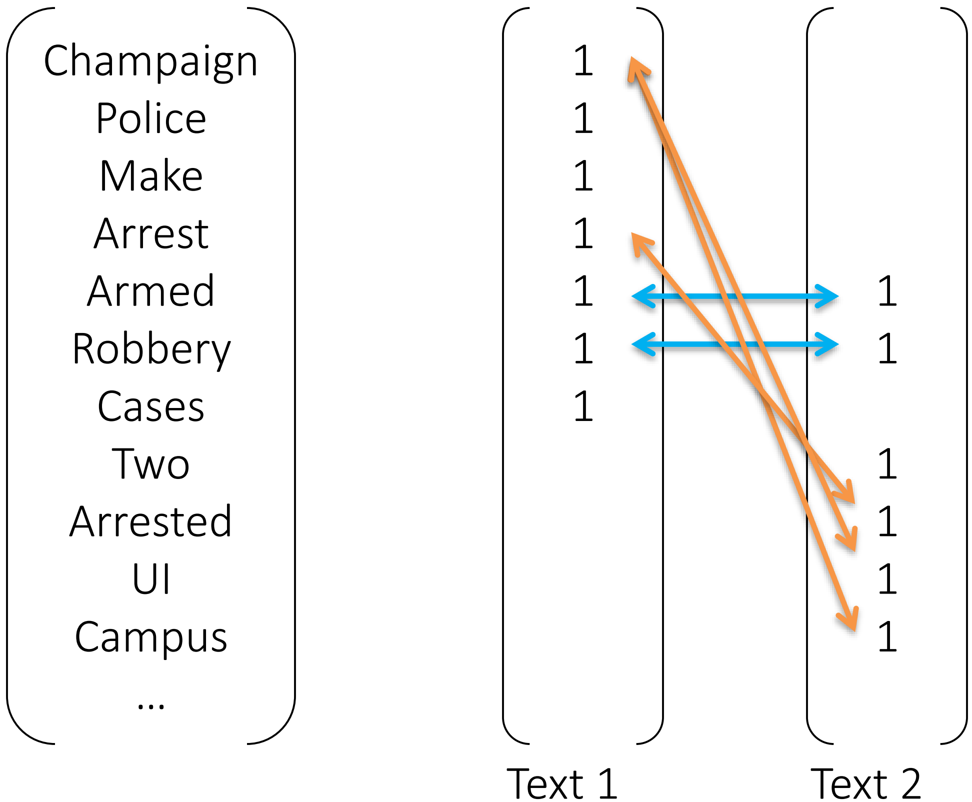
Balanced binary classification



Multi-class classification

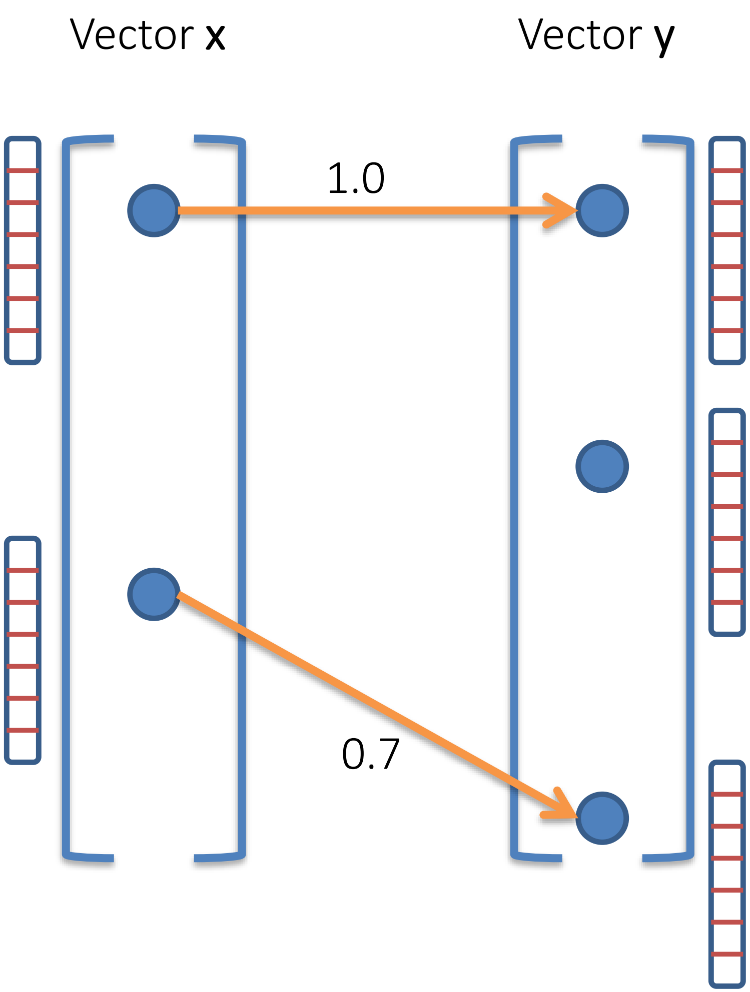
Similarity

- Cosine



$$\cos(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\delta(a_i - b_j)x_{a_i}y_{b_j}}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$$

Representation Densification



Cosine

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\delta(a_i - b_j)x_{a_i}y_{b_j}}{||\mathbf{x}||\cdot||\mathbf{y}||}$$

Average

$$S_A(\mathbf{x},\mathbf{y}) = \frac{1}{n_x||\mathbf{x}||\cdot n_y||\mathbf{y}||}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} x_{a_i}y_{b_j}\phi(a_i,b_j)$$

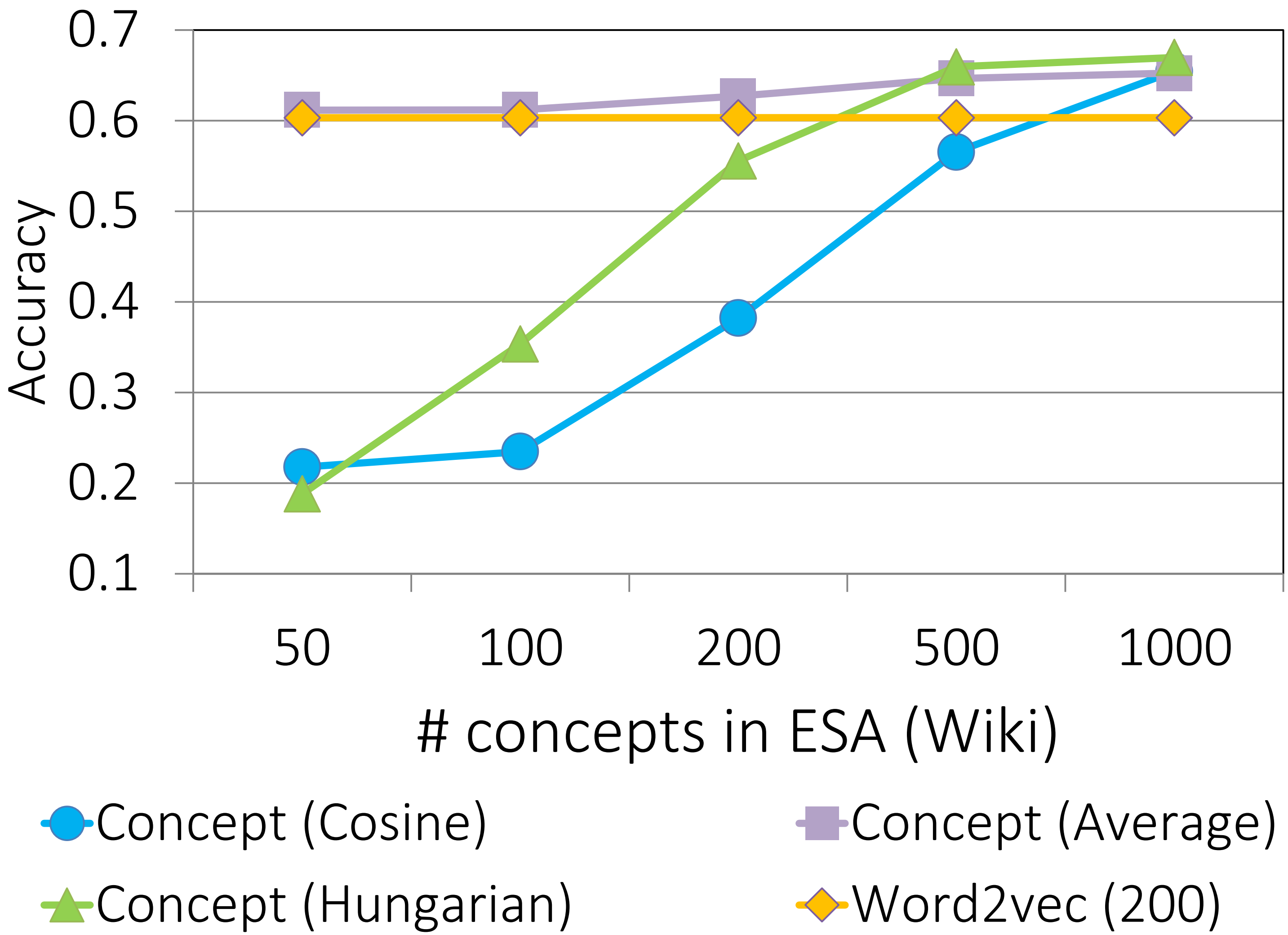
Max matching

$$S_M(\mathbf{x},\mathbf{y}) = \frac{1}{||\mathbf{x}||\cdot||\mathbf{y}||}\sum_{i=1}^{n_x} x_{a_i}y_{b_j}\max_j \phi(a_i,b_j)$$

Hungarian matching

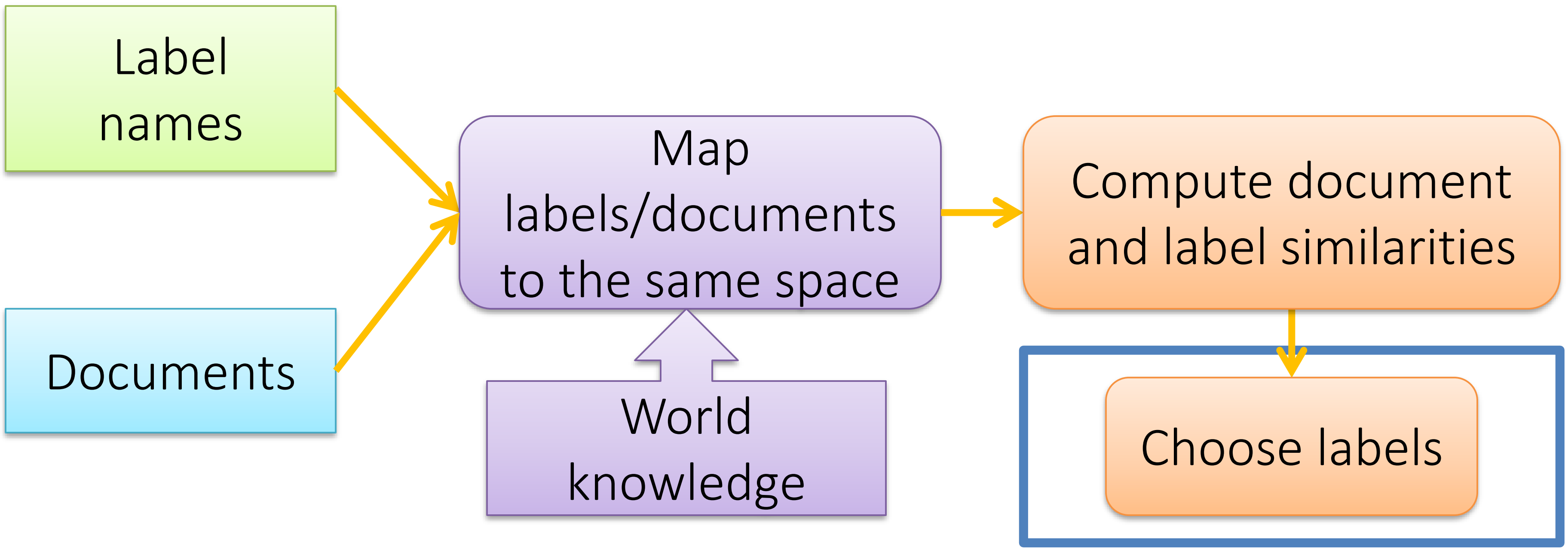
$$S_H(\mathbf{x},\mathbf{y}) = \frac{1}{||\mathbf{x}||\cdot||\mathbf{y}||}\sum_{i=1}^{n_x} x_{a_i}y_{h(a_i)}\phi(a_i,h(a_i))$$

rec.autos vs. sci.electronics (1/16 document: 13 words per text)

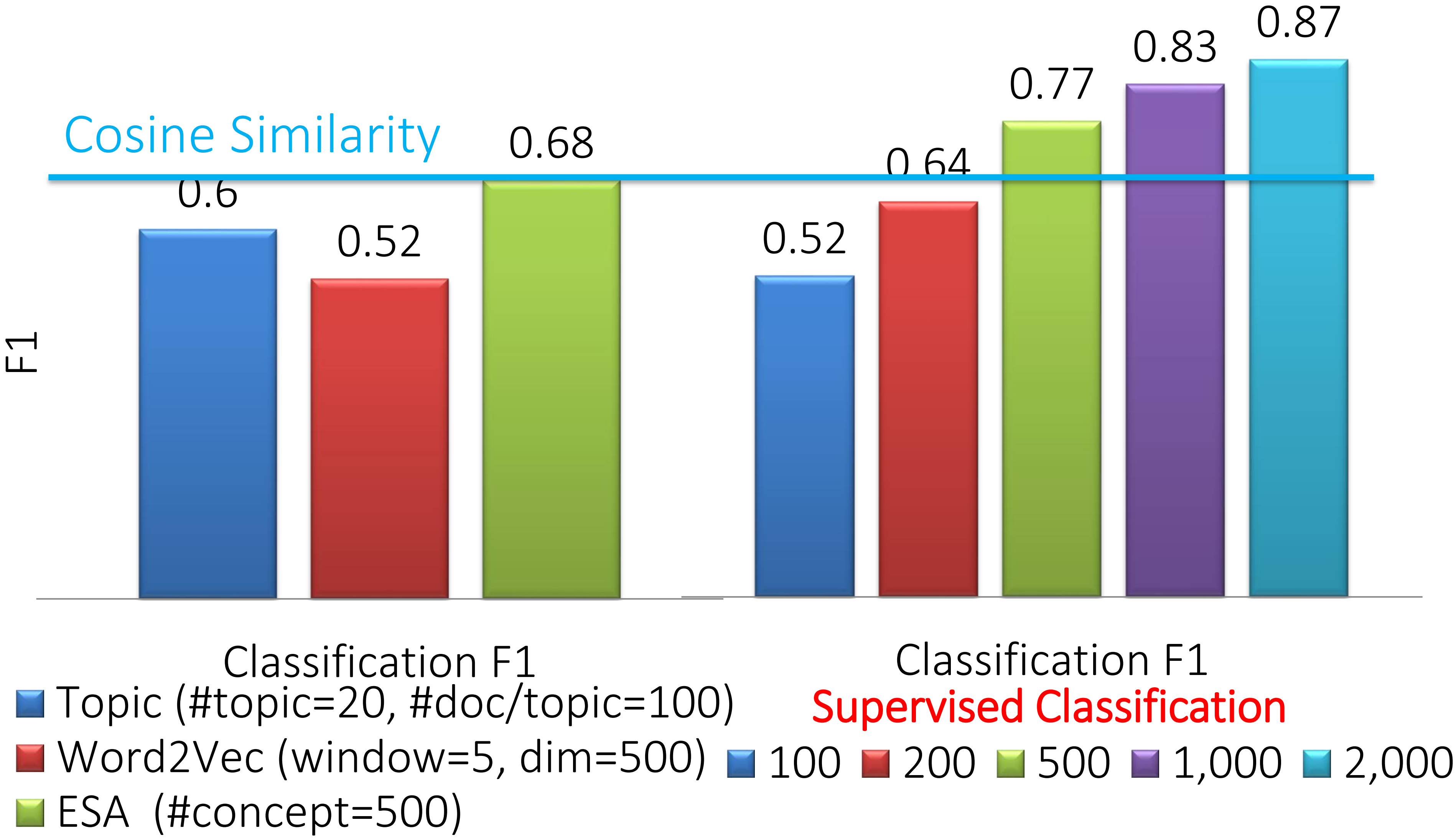


Song and Roth. North Amer. Chap. Assoc. Comp. Ling. (NAACL). 2015.

Dataless Text Classification: Classify Documents on the Fly



Classification of 20 Newsgroups Documents



Blei et al., Latent Dirichlet Allocation. J. of Mach. Learn. Res. (JMLR). 2003.
Mikolov et al. Efficient Estimation of Word Representations in Vector Space. Adv. Neur. Info. Proc. Sys. (NIPS). 2013.

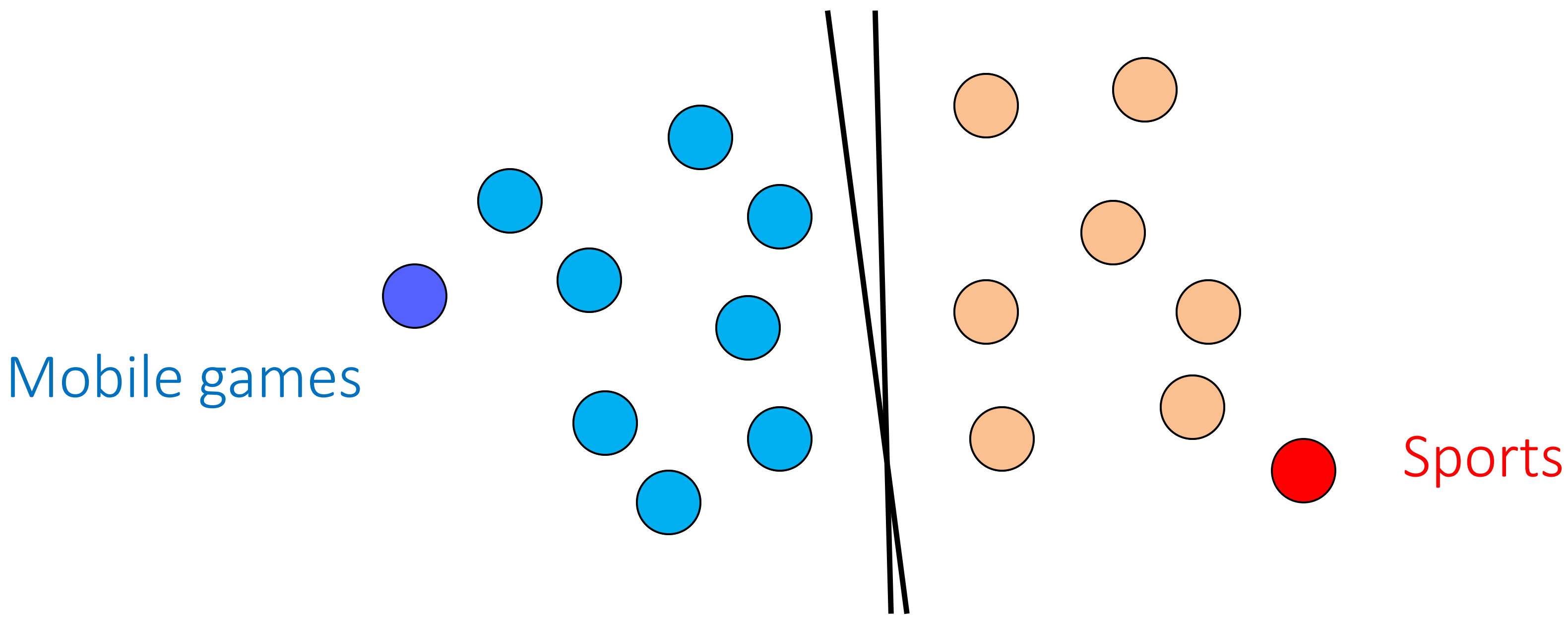
Bootstrapping with Unlabeled Data

Application of world knowledge of label meaning

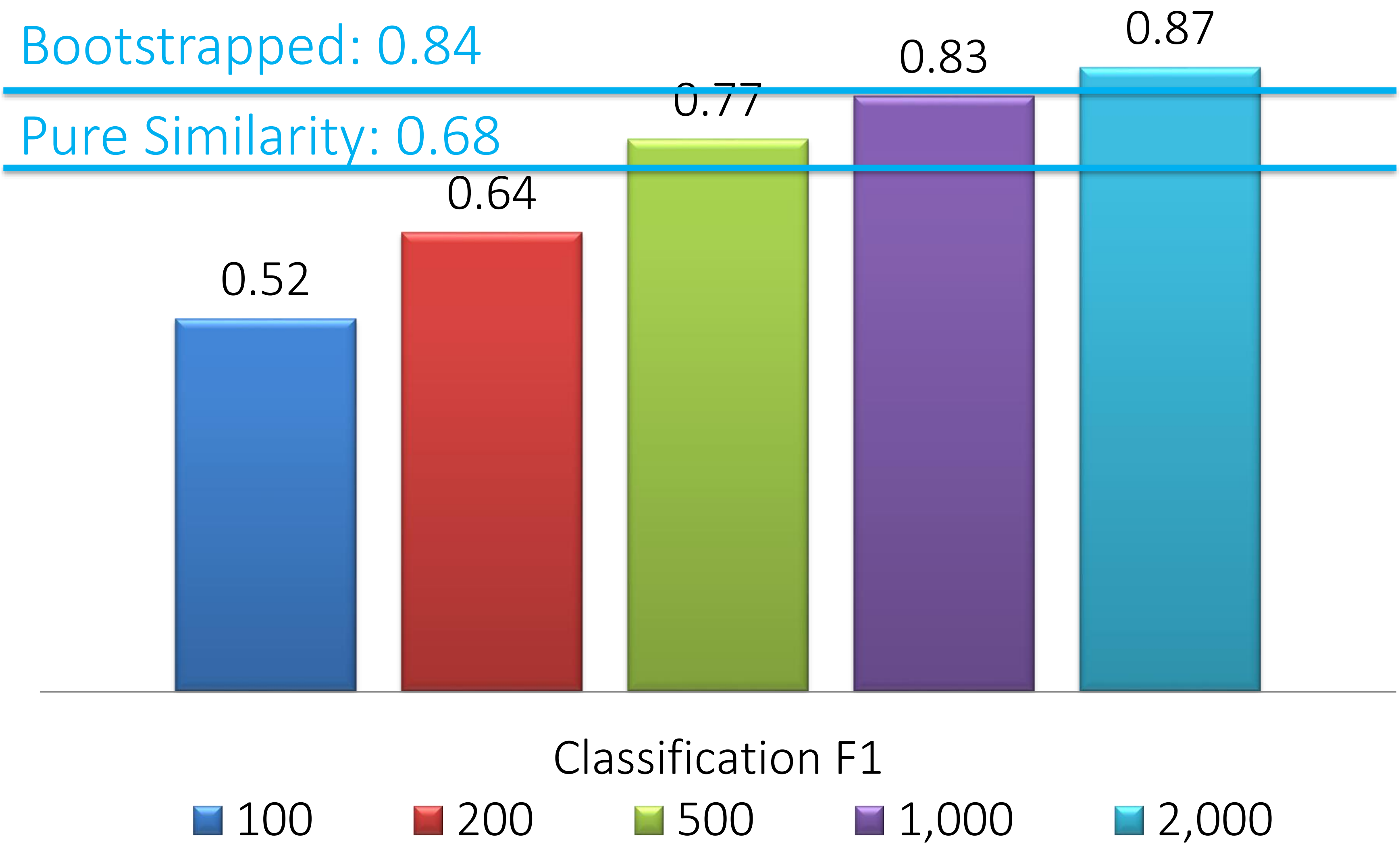
- Pure similarity based classifications

Domain adaptation

- Continue to label more data until no unlabeled document exists



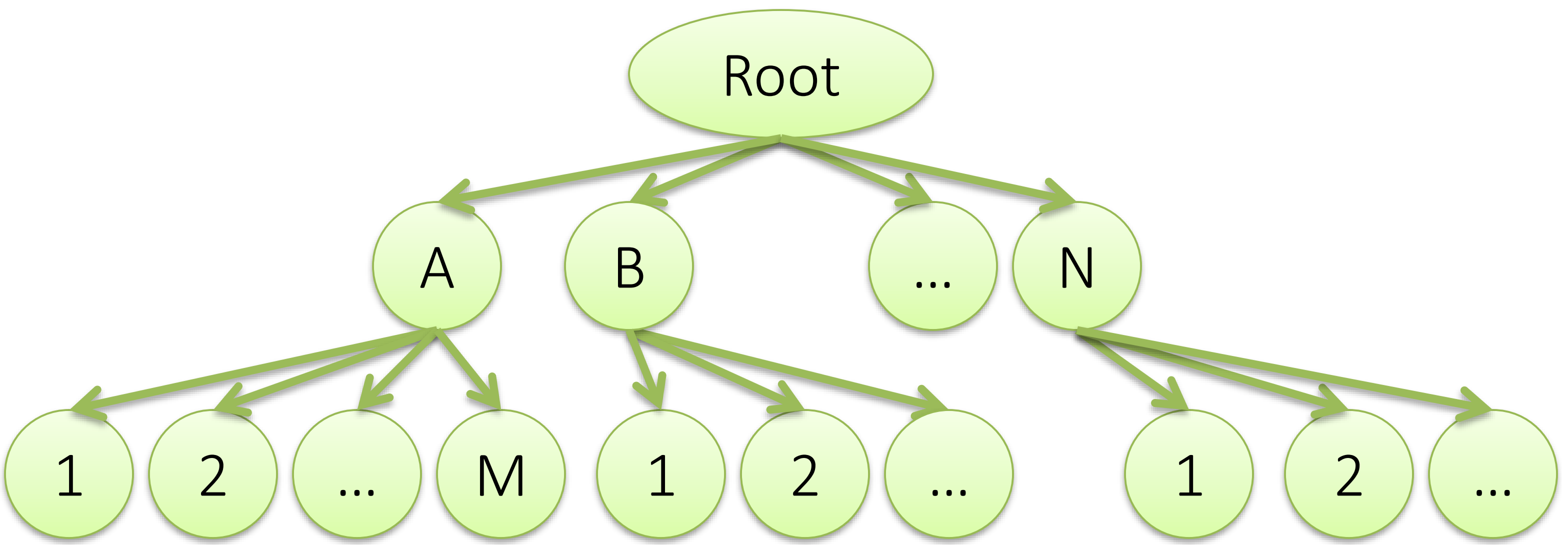
Classification of 20 Newsgroups Documents



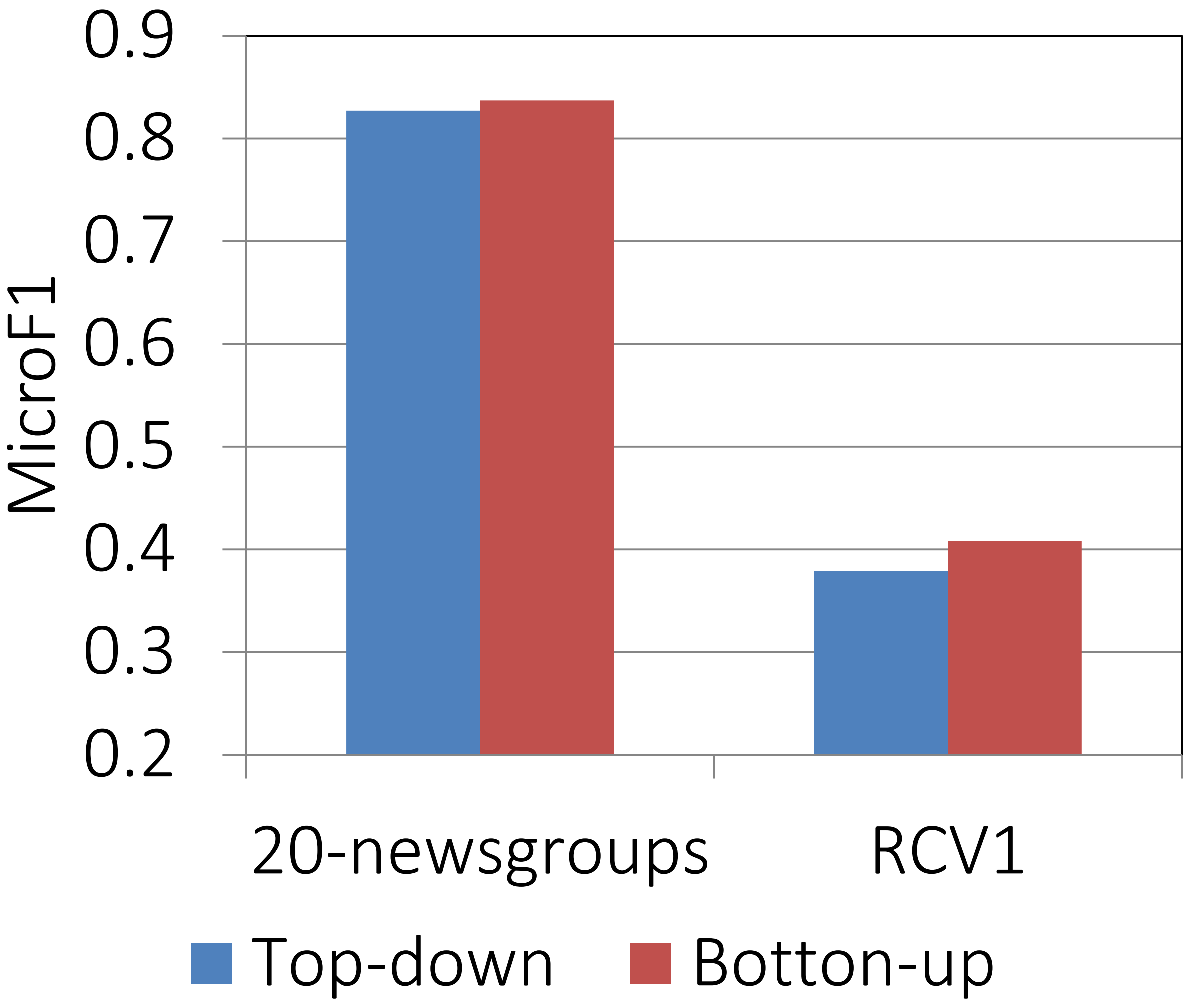
Song and Roth. Assoc. Adv. Artif. Intell. (AAAI). 2014

Hierarchical Classification: Considering Label Dependency

- Top-down classification
- Bottom-up classification (flat classification)



Top-down vs. Bottom-up



$$P = \frac{\Sigma_{t \in T} TP_t}{\Sigma_{t \in T} TP_t + FP_t}$$

$$R = \frac{\Sigma_{t \in T} TP_t}{\Sigma_{t \in T} TP_t + FN_t}$$

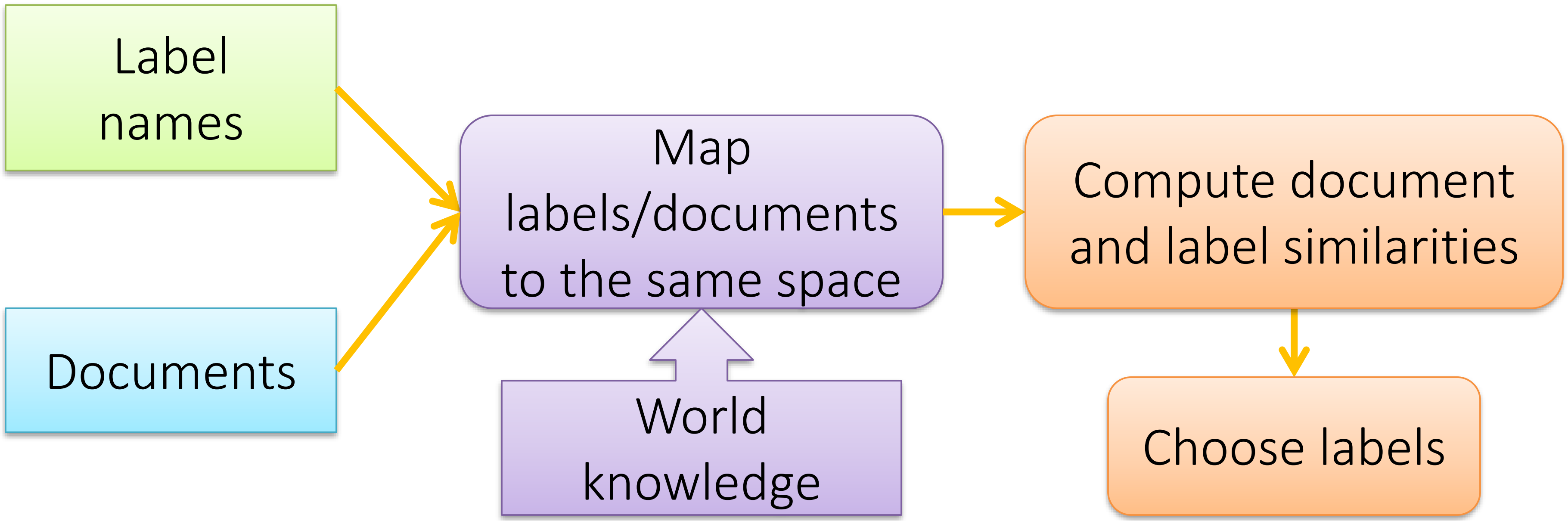
$$Micro\text{-}F_1 = \frac{2PR}{P + R}$$

RCV1

- 804,414 documents
- 82 categories in 4 levels
- 103 nodes in hierarchy
- 3.24 labels/document

Song and Roth. Assoc. Adv. Artif. Intell. (AAAI). 2014

Dataless Text Classification: Classify Documents on the Fly



	Labeled data in training	Unlabeled data in training	Label names in training	I.I.D. between training and testing
Supervised learning	Yes	No	No	Yes
Unsupervised learning	No	Yes	No	Yes
Semi-supervised learning	Yes	Yes	No	Yes
Transfer learning	Yes	Yes	No	No
Zero-shot learning	Yes	No	Yes	No
Dataless Classification (pure similarity)	No	No	Yes	No
Dataless Classification (bootstrapping)	No	Yes	Yes	Yes

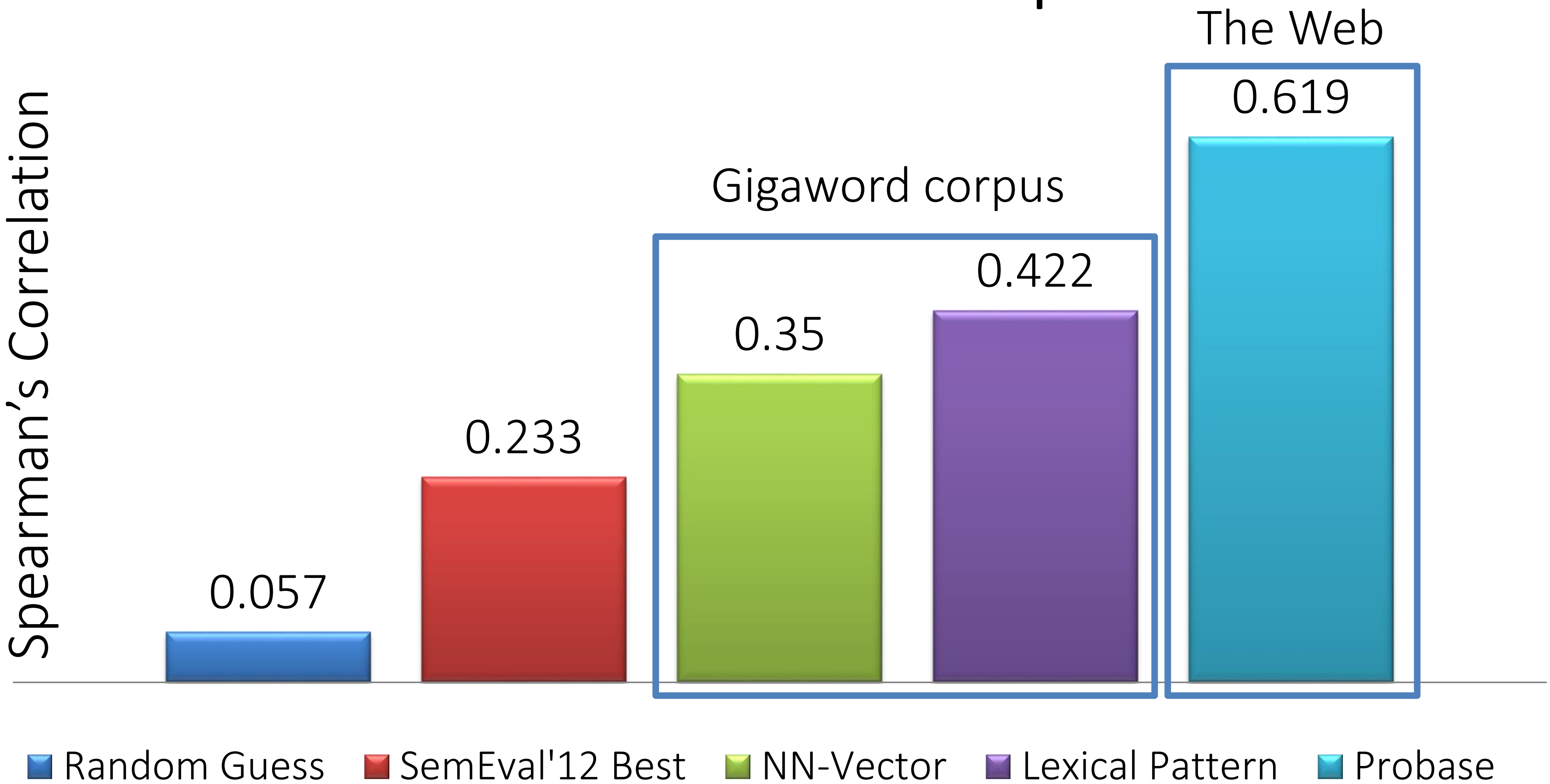
Conclusions

- Dataless classification
  - Reduce labeling work for thousands of documents
- Compared semantic representation using world knowledge
  - Probabilistic conceptualization (PC)
  - Explicit semantic analysis (ESA)
  - Word embedding (word2vec)
  - Topic model (LDA)
  - Combination of ESA and word2vec
- Unified PC and ESA
  - Markov random field model
- Domain adaptation
  - Hyper-parameter estimation
  - Boostrapping – refining the classifier

Advertisement:
Using knowledge as structured information instead of flat features!
Session 7B, DM835

Thank You! ☺

Correlation with Human Annotation of IS-A Relationships



Combining Heterogeneous Models for Measuring Relational Similarity. A. Zhila, W. Yih, C. Meek, G. Zweig & T. Mikolov. In NAACL-HLT-13.