What Users Ask a Search Engine: Analyzing One Billion Question Queries

Michael Völske[1] Pavel Braslavski[2,3] Matthias Hagen[1]

Galina Lezina[2,4] Benno Stein[1]

[1]Bauhaus-Universität Weimar
<firstname>.<lastname>@uni-weimar.de
www.webis.de

[2]Ural Federal University
[3]pbras@yandex.ru
[4]galina.lezina@gmail.com

Question Queries

Question Queries

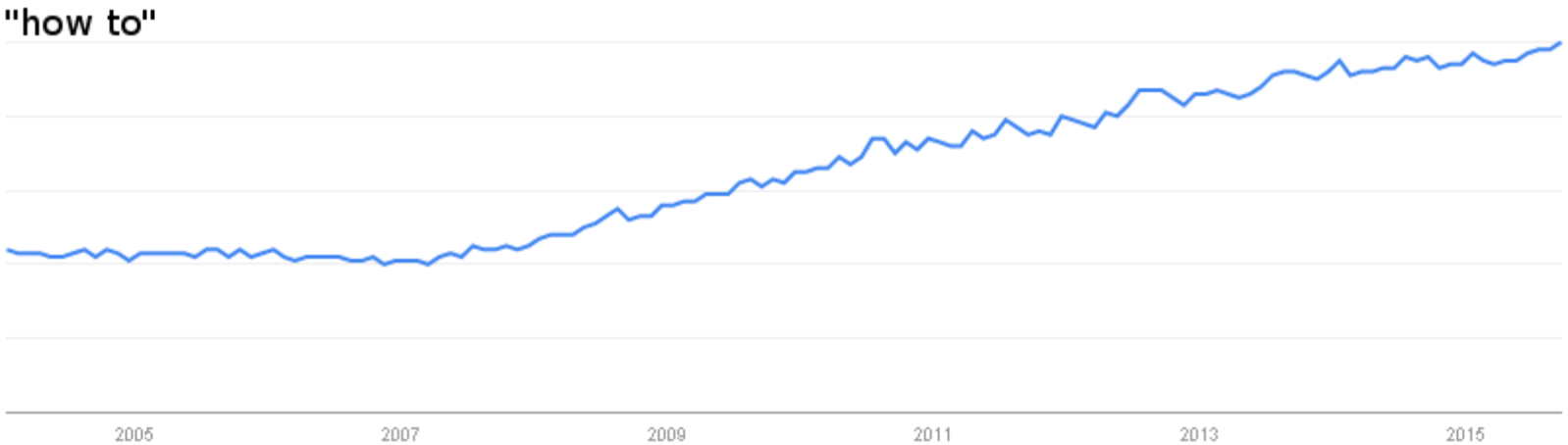
Question Queries

lose weight

how much should I exercise to lose ten pounds

Question Queries

Relevance



[Google Trends, 2015]

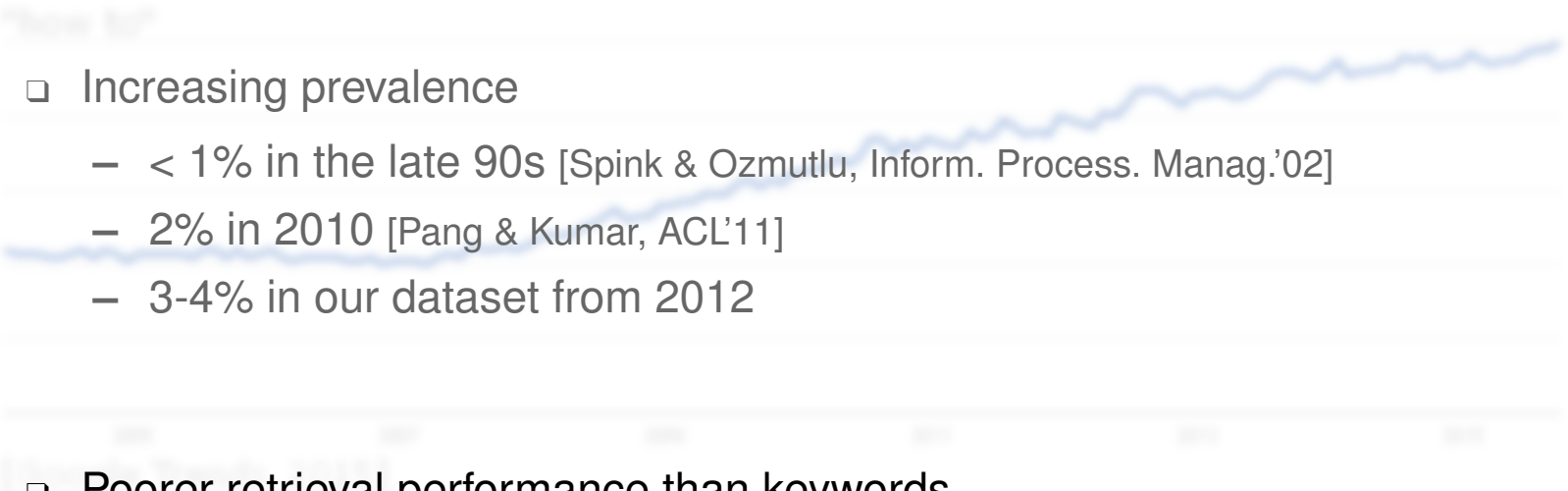
Question Queries

Relevance

- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.'02]
  - 2% in 2010 [Pang & Kumar, ACL'11]
  - 3-4% in our dataset from 2012

Question Queries

Relevance



- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.'02]
  - 2% in 2010 [Pang & Kumar, ACL'11]
  - 3-4% in our dataset from 2012

- Poorer retrieval performance than keywords
  [Bendersky & Croft, WSCD'09] [Aula et al., CHI'10]

Question Queries

Relevance

- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.'02]
  - 2% in 2010 [Pang & Kumar, ACL'11]
  - 3-4% in our dataset from 2012

- Poorer retrieval performance than keywords
  [Bendersky & Croft, WSCD'09] [Aula et al., CHI'10]

- Topical query classification benefits
  - General search [Bailey et al., ACM TWEB'10]
  - Query disambiguation [Li et al., SIGIR'08]
  - Search advertising [Broder et al., SIGIR'07]

What Users Ask a Search Engine

. . . About this Talk

- Large dataset of ~1 billion question queries from Yandex
- Question query classification using CQA data as training set
- Three classification pipelines: Retrieval, BoW, Topic models
- Insights into asker behavior

What Users Ask a Search Engine

Our Approach

- **Classification task:** given unlabeled question query, predict category
  - Click information not helpful: QQ are rare

What Users Ask a Search Engine

Our Approach

- **Classification task:** given unlabeled question query, predict category
  - Click information not helpful: QQ are rare

- Community question answering (CQA) data as training set
  - CQA users manually select appropriate category for their question

What Users Ask a Search Engine

Our Approach

- **Classification task:** given unlabeled question query, predict category
  - Click information not helpful: QQ are rare
- Community question answering (CQA) data as training set
  - CQA users manually select appropriate category for their question
- Train a classifier that correctly categorizes CQA, then transfer to QQ

Datasets

Overview

Dataset	Queries	Labels
Question Queries (yandex.ru)	1 980 million	unlabeled

Datasets

Overview

Dataset	Queries	Labels
Question Queries (yandex.ru)	1 980 million	unlabeled
- after cleaning	900 million	unlabeled

Cleaning to remove:

- Spam & bots
- Repeated submissions

Datasets

Overview

Dataset	Queries	Labels
Question Queries (yandex.ru)	1 980 million	unlabeled
- after cleaning	900 million	unlabeled
CQA Questions (Otvety@Mail.ru)	11 million	hierarchical (189)

Cleaning to remove:

- Spam & bots
- Repeated submissions

Datasets

Overview

Dataset	Queries	Labels
Question Queries (yandex.ru)	1 980 million	unlabeled
- after cleaning	900 million	unlabeled
CQA Questions (Otvety@Mail.ru)	11 million	hierarchical (189)
- after cleaning	6 million	flat (14)

Cleaning to remove:

- Spam & bots
- Repeated submissions
- Mis-categorized CQA questions

Datasets

Train and Test Set

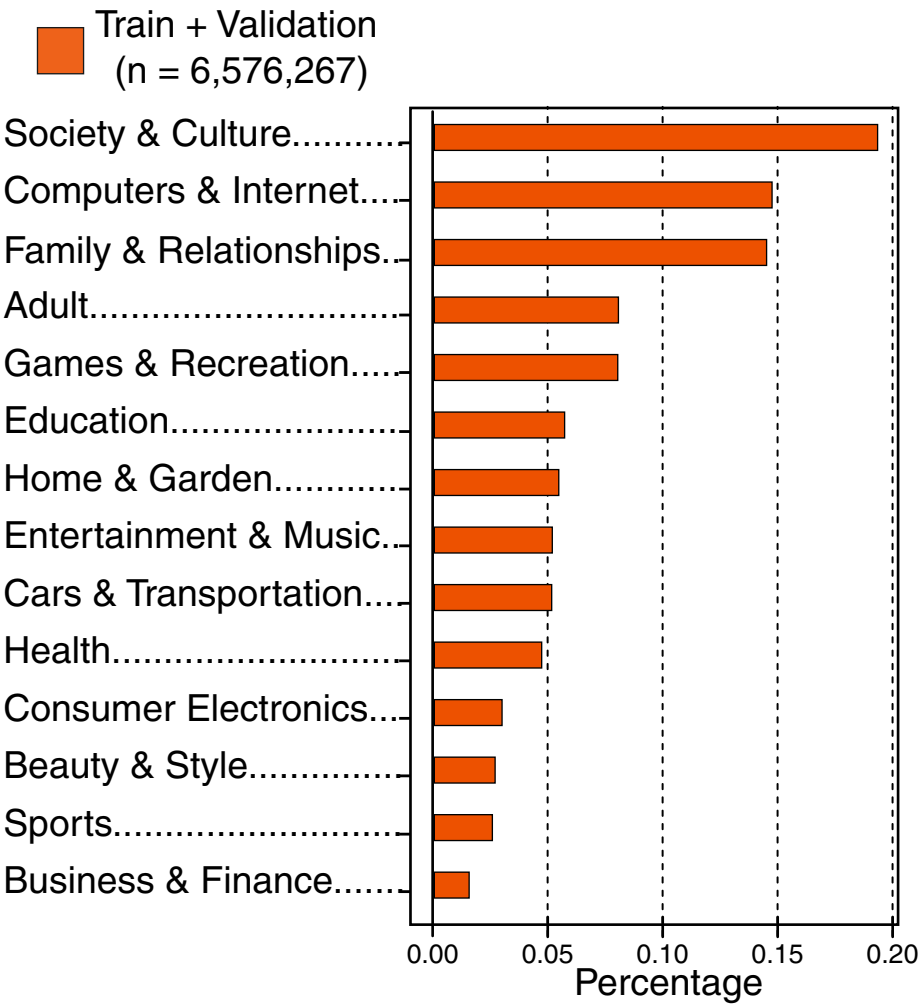
- CQA data
  - 14 classes derived from CQA taxonomy

Society & Culture
Computers & Internet
Family & Relationships
Adult
Games & Recreation
Education
Home & Garden
Entertainment & Music
Cars & Transportation
Health
Consumer Electronics
Beauty & Style
Sports
Business & Finance

Datasets

Train and Test Set

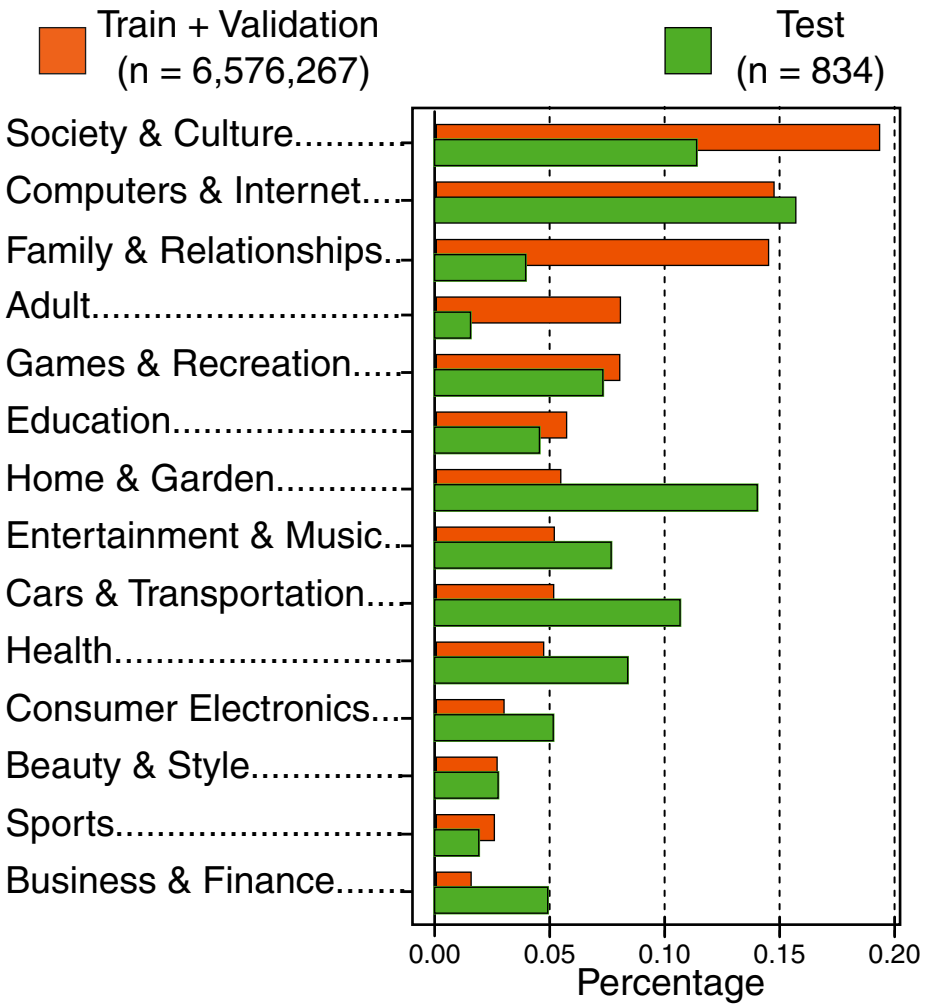
- CQA data
  - 14 classes derived from CQA taxonomy
  - Training/validation set: 70/30 split



Datasets

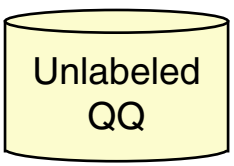
Train and Test Set

- CQA data
  - 14 classes derived from CQA taxonomy
  - Training/validation set: 70/30 split
- Question queries
  - Test set: 1000 instances hand-labeled
  - 834 with majority agreement

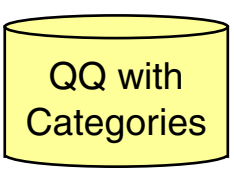


Question Query Classification Pipelines

Question Query Classification Pipelines

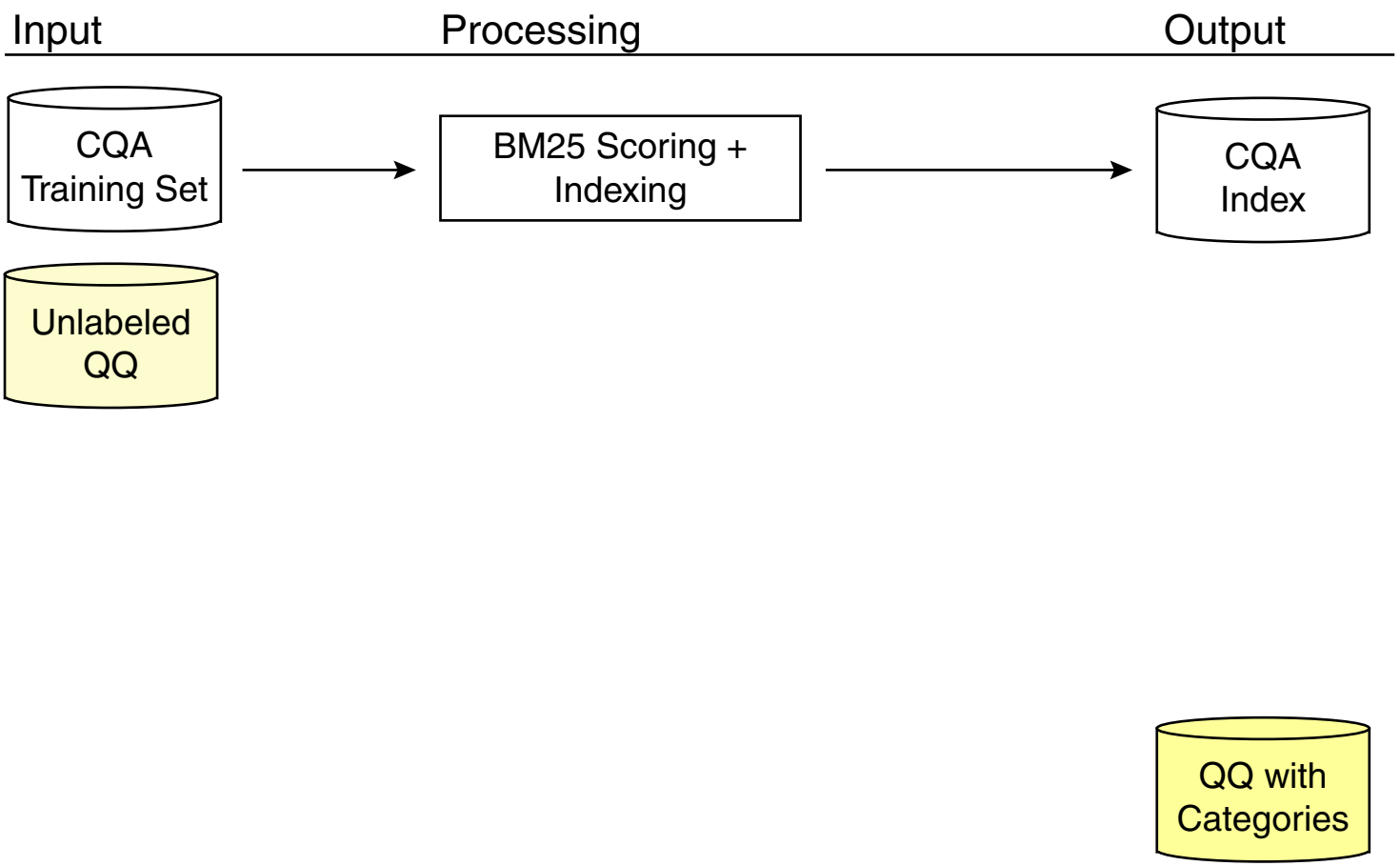


- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models



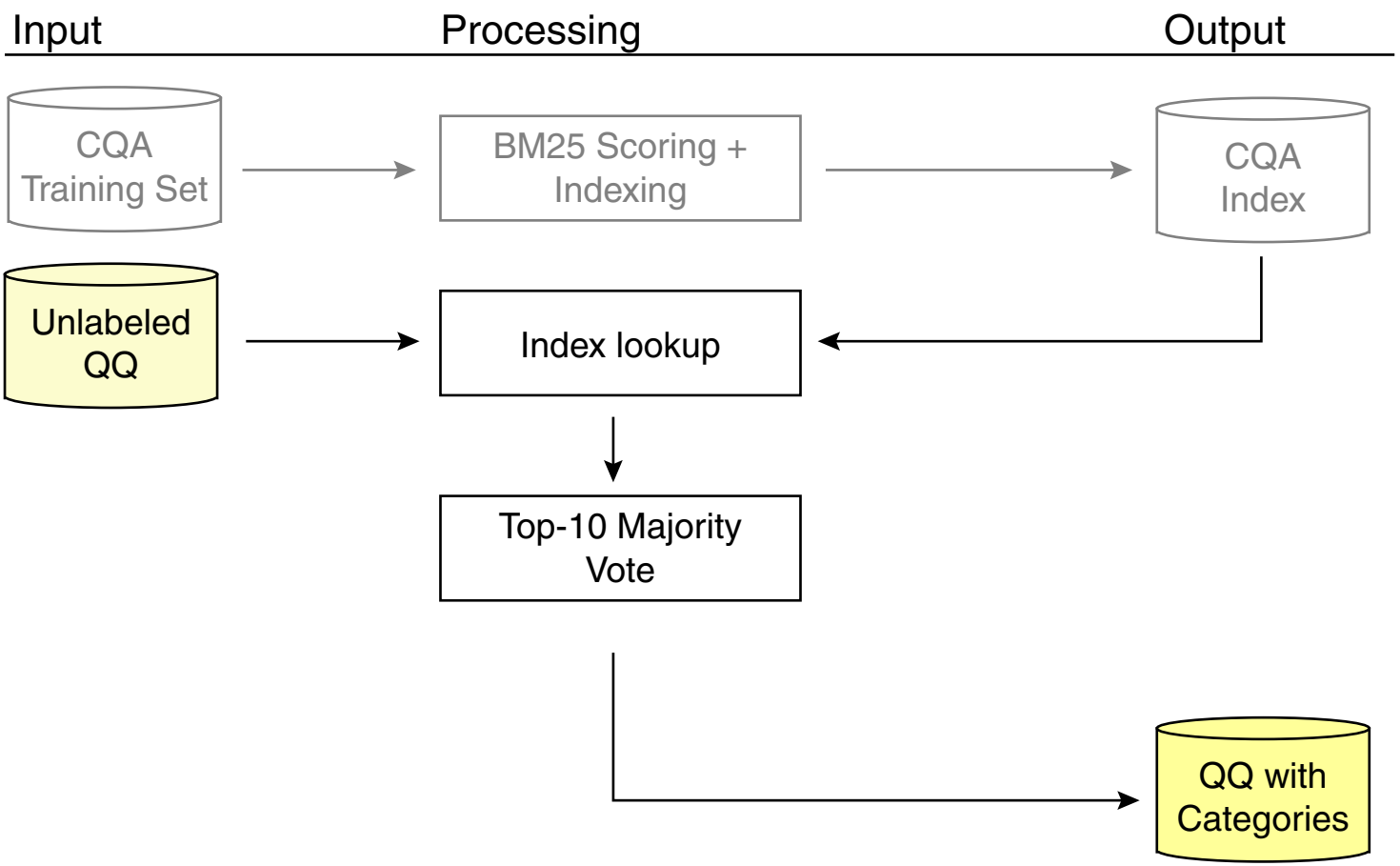
Question Query Classification Pipelines

Pipeline 1: CQA Retrieval

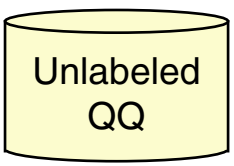


Question Query Classification Pipelines

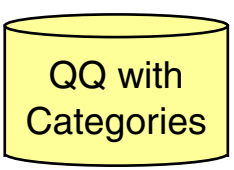
Pipeline 1: CQA Retrieval



Question Query Classification Pipelines

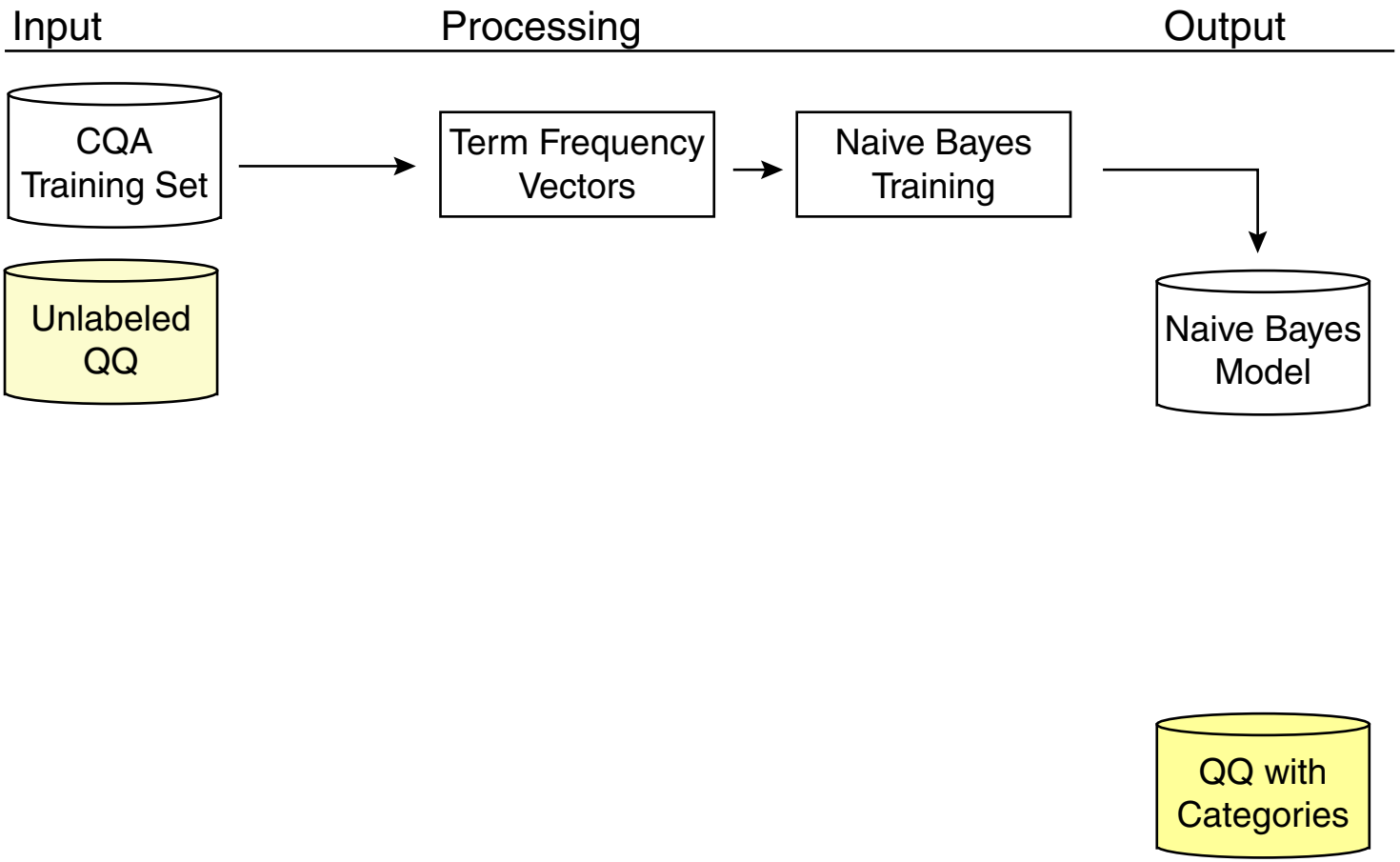


- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models



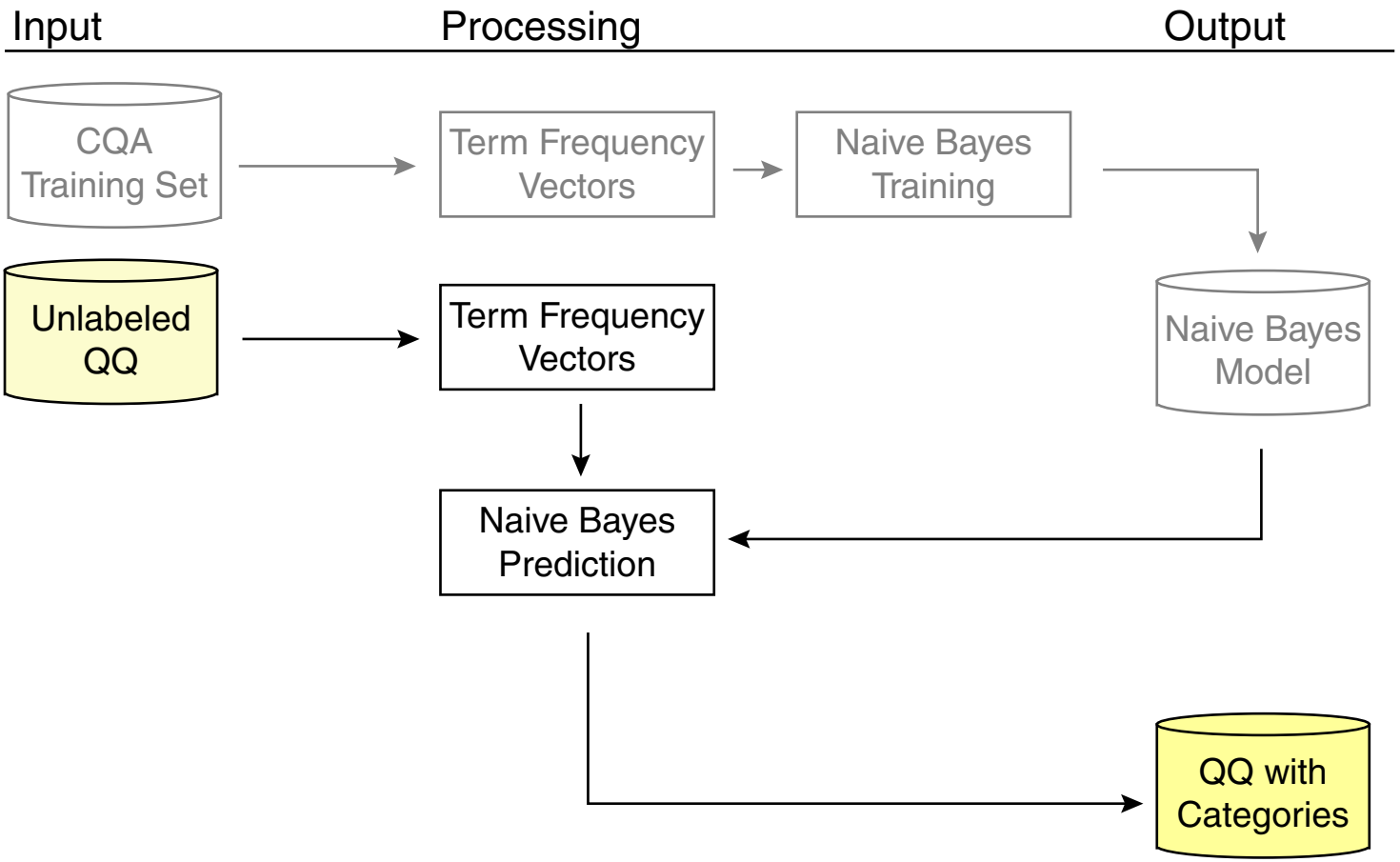
Question Query Classification Pipelines

Pipeline 2: Bag-of-Words Classifier

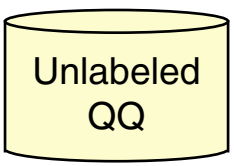


Question Query Classification Pipelines

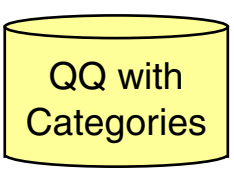
Pipeline 2: Bag-of-Words Classifier



Question Query Classification Pipelines

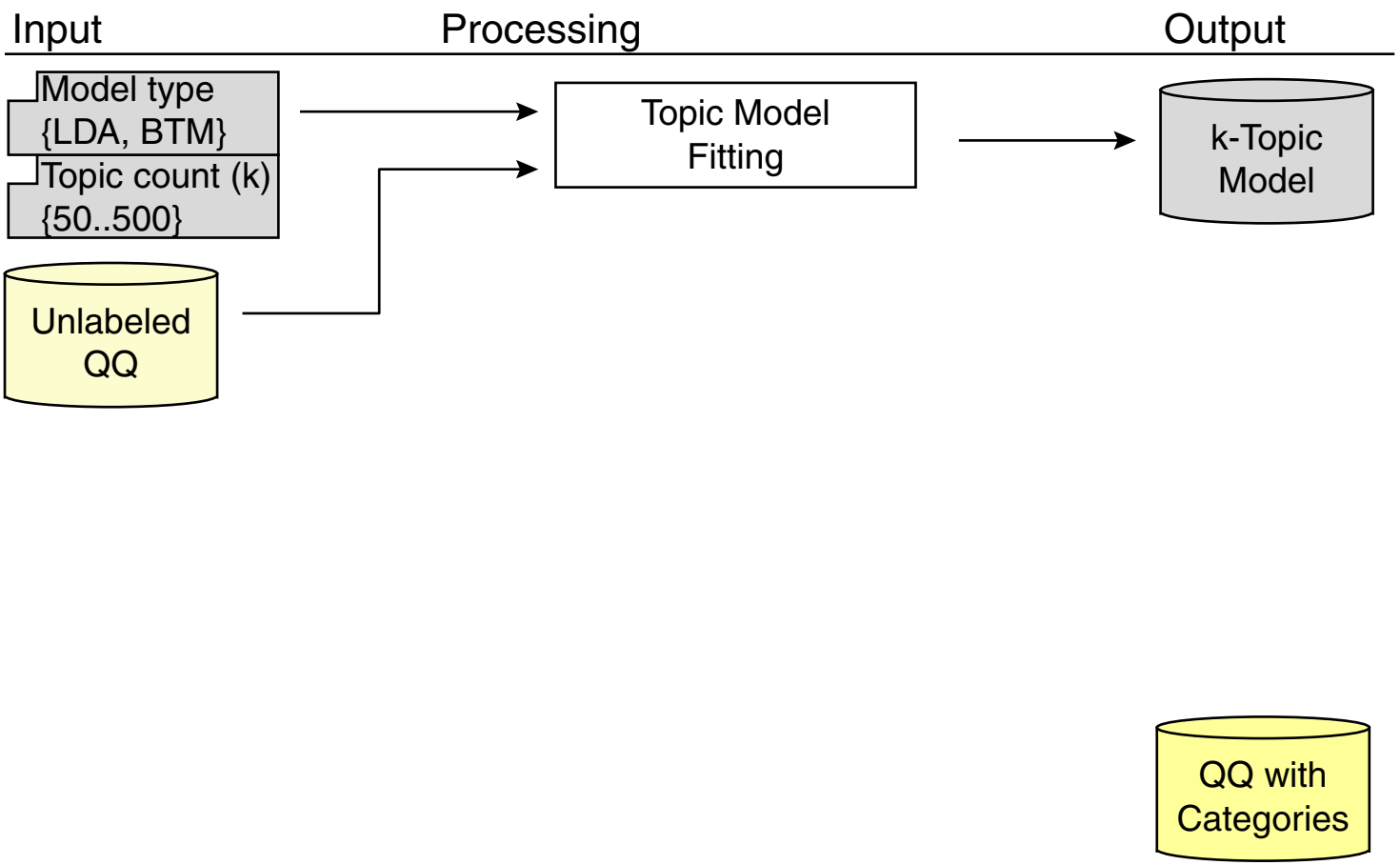


- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models



Question Query Classification Pipelines

Pipeline 3: Topic Models

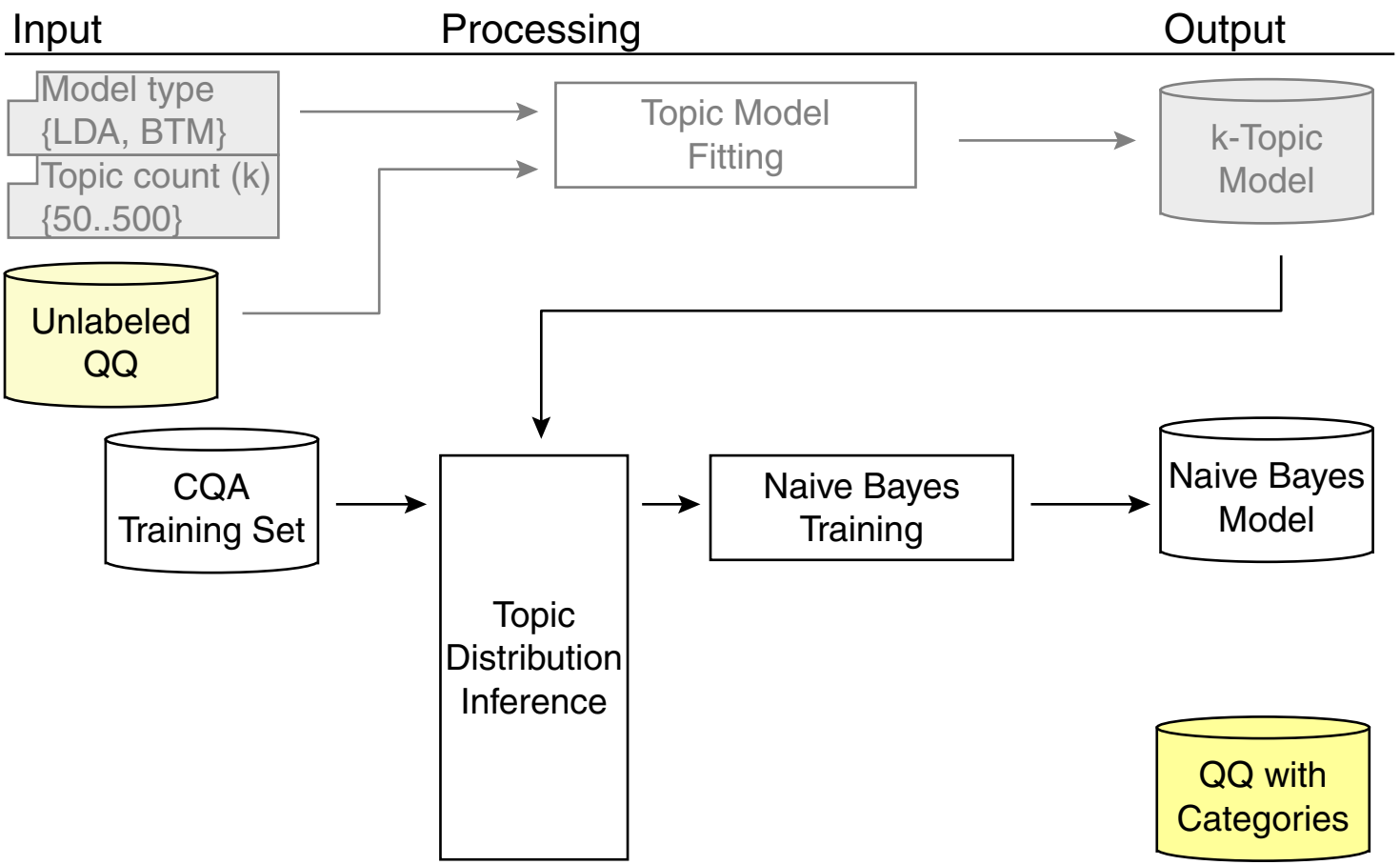


Latent Dirichlet Allocation: [Blei et al., JMLR'03]

Biterm Topic Model: [Yan et al., WWW'11]

Question Query Classification Pipelines

Pipeline 3: Topic Models

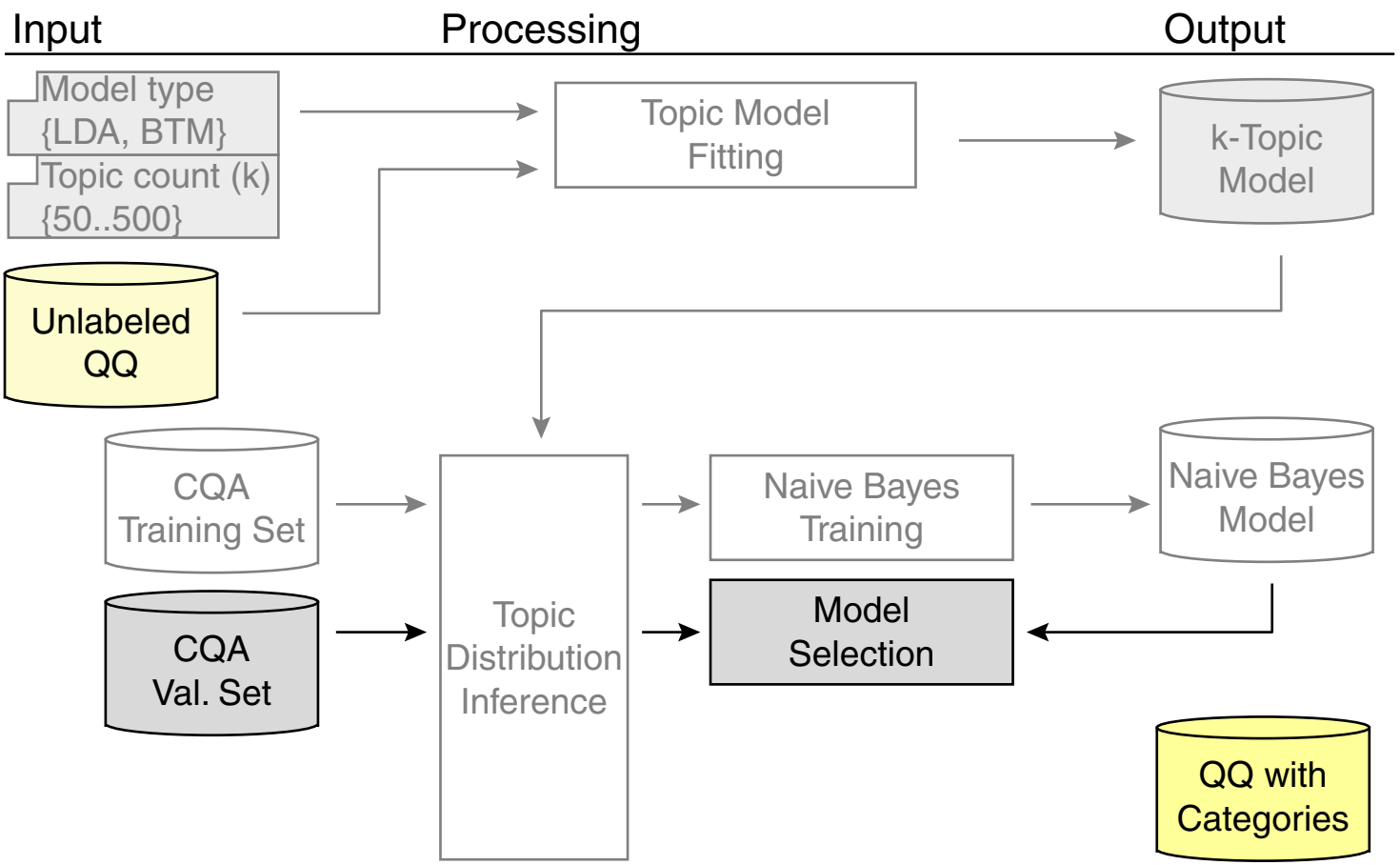


Latent Dirichlet Allocation: [Blei et al., JMLR'03]

Biterm Topic Model: [Yan et al., WWW'11]

Question Query Classification Pipelines

Pipeline 3: Topic Models

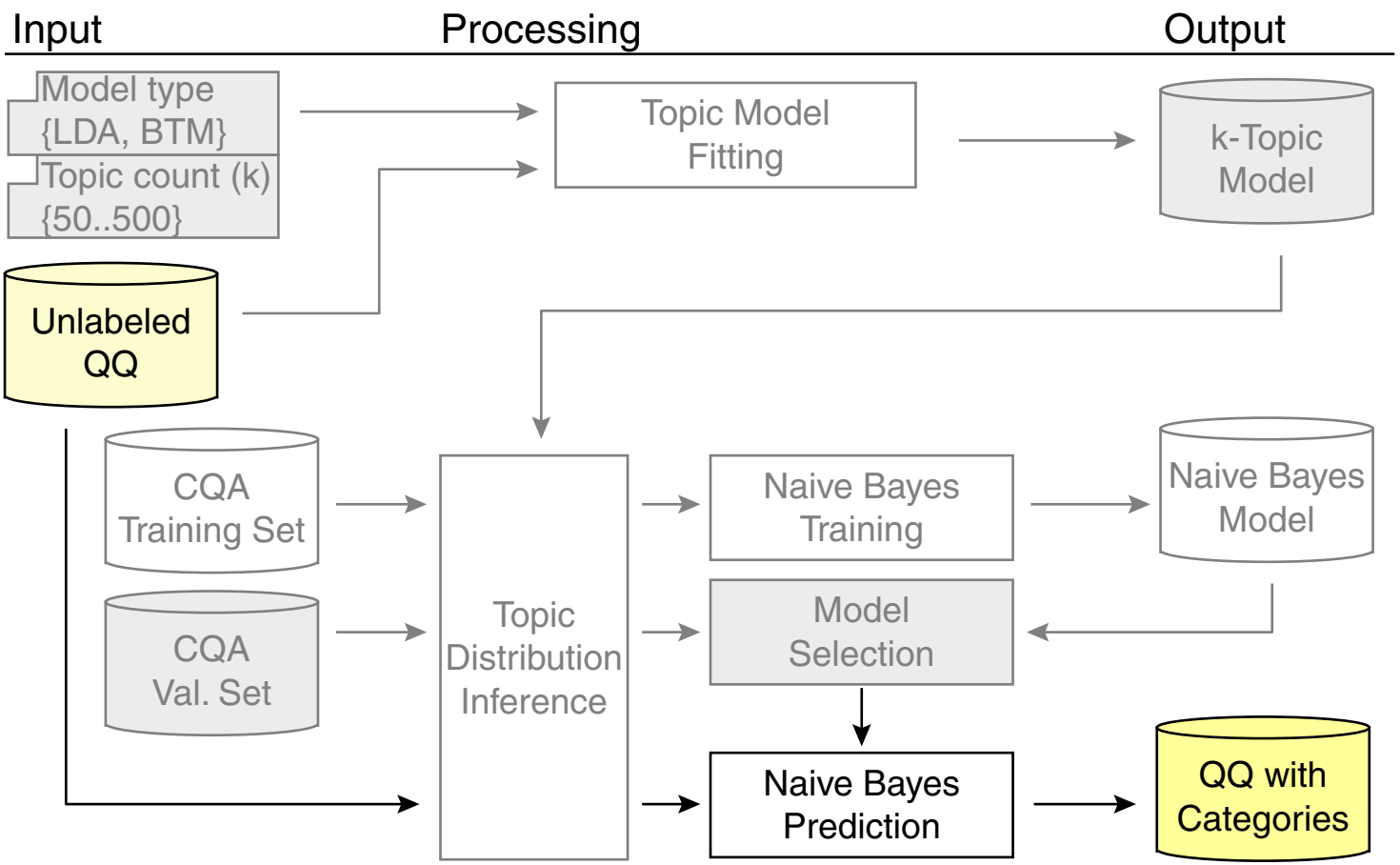


Latent Dirichlet Allocation: [Blei et al., JMLR'03]

Biterm Topic Model: [Yan et al., WWW'11]

Question Query Classification Pipelines

Pipeline 3: Topic Models



Latent Dirichlet Allocation: [Blei et al., JMLR'03]

Biterm Topic Model: [Yan et al., WWW'11]

Results

Classifier Performance

Test Set (n=834)			
Features	**Precision**	**Recall**	**F_1-Score**
CQA Retrieval			
6 million	0.67	0.66	0.66
Bag-of-Words			
137,032	0.61	0.70	0.65
LDA Topics			
500	0.40	0.39	0.40
Biterm Topics			
450	0.49	0.53	0.51

- Simple BoW classifier performs similarly to CQA retrieval
- Biterm topic model outperforms LDA

Results

Classifier Performance

Test Set (n=834)

Features	Precision	Recall	F_1-Score
CQA Retrieval			
6 million	0.67	0.66	0.66
Bag-of-Words			
137,032	0.61	0.70	0.65
LDA Topics			
500	0.40	0.39	0.40
Biterm Topics			
450	0.49	0.53	0.51

Model	Complexity	
	Training	Classification
CQA Retrieval	medium	high
Bag-of-Words	low	medium
Topic Models	high	low

- Simple BoW classifier performs similarly to CQA retrieval
- Biterm topic model outperforms LDA
- Topic models less accurate but faster at classification time

Results

Classifier Performance

Test Set (n=834)

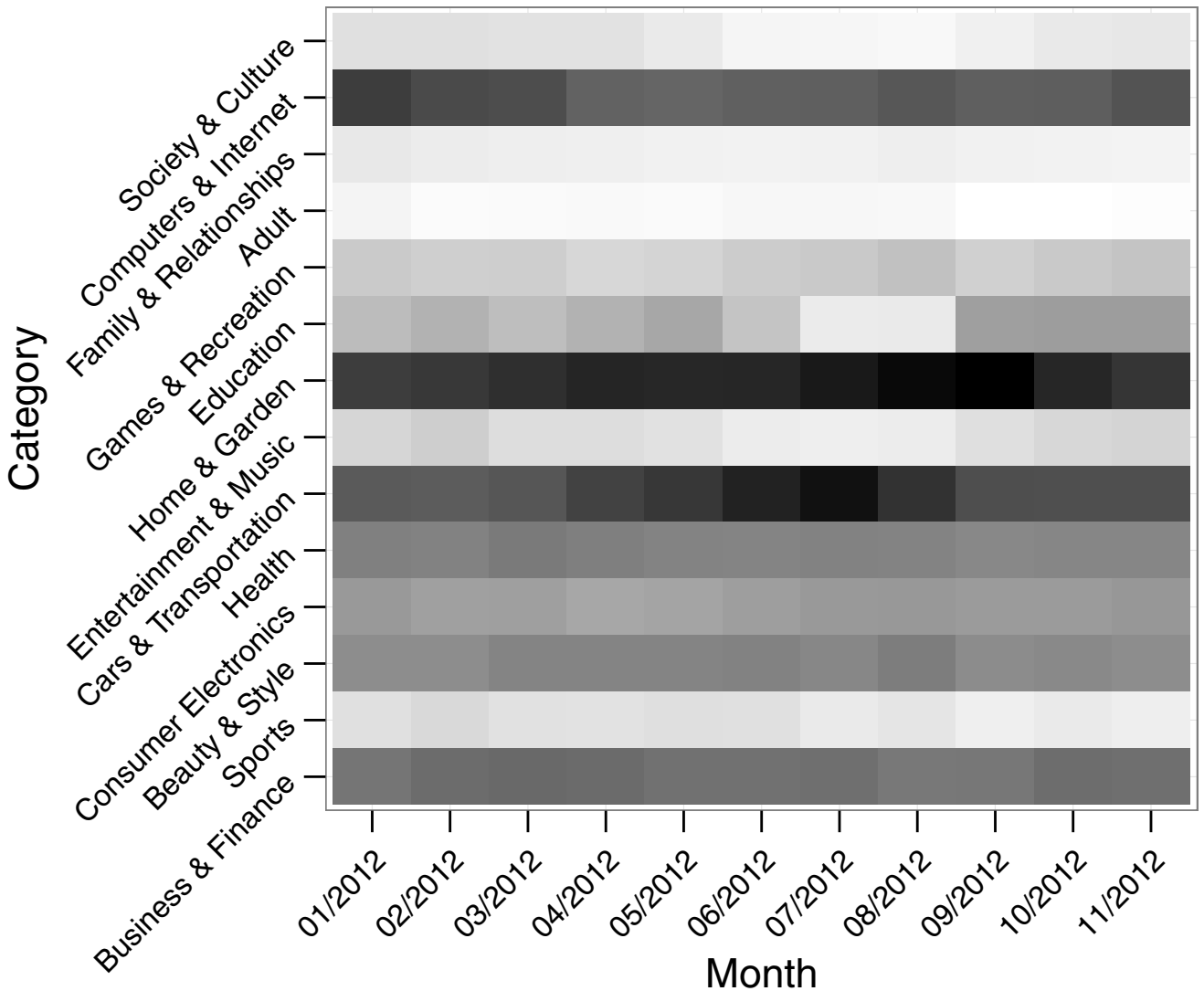
Features	Precision	Recall	F_1-Score
CQA Retrieval			
6 million	0.67	0.66	0.66
Bag-of-Words			
137,032	0.61	0.70	0.65
LDA Topics			
500	0.40	0.39	0.40
Biterm Topics			
450	0.49	0.53	0.51

Model	Complexity	
	Training	Classification
CQA Retrieval	medium	high
Bag-of-Words	low	medium
Topic Models	high	low

- Simple BoW classifier performs similarly to CQA retrieval
- Biterm topic model outperforms LDA
- Topic models less accurate but faster at classification time
- We use the Bag-of-Words classifier to analyze the question queries dataset

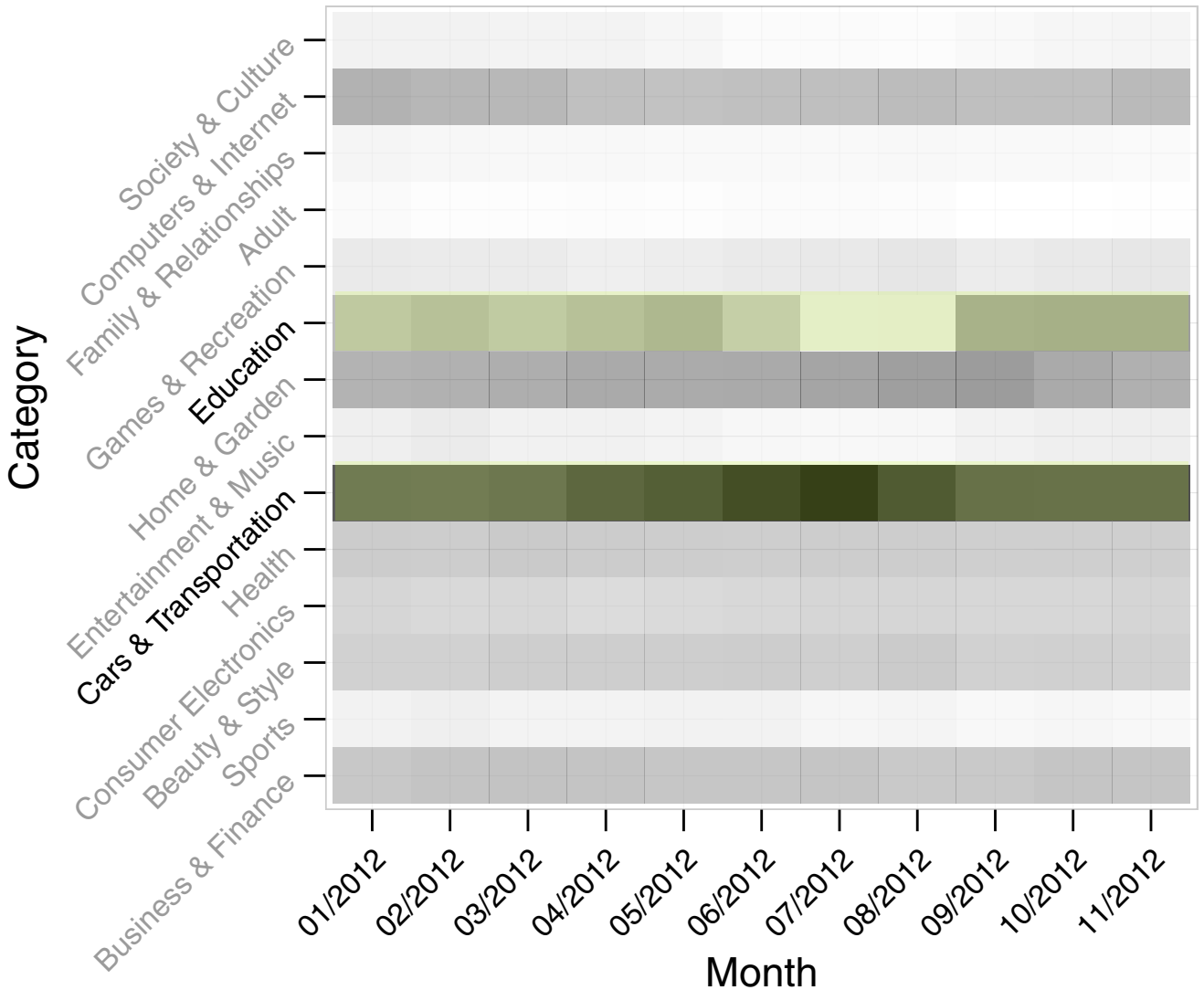
Results

Evolution of Categories over Time



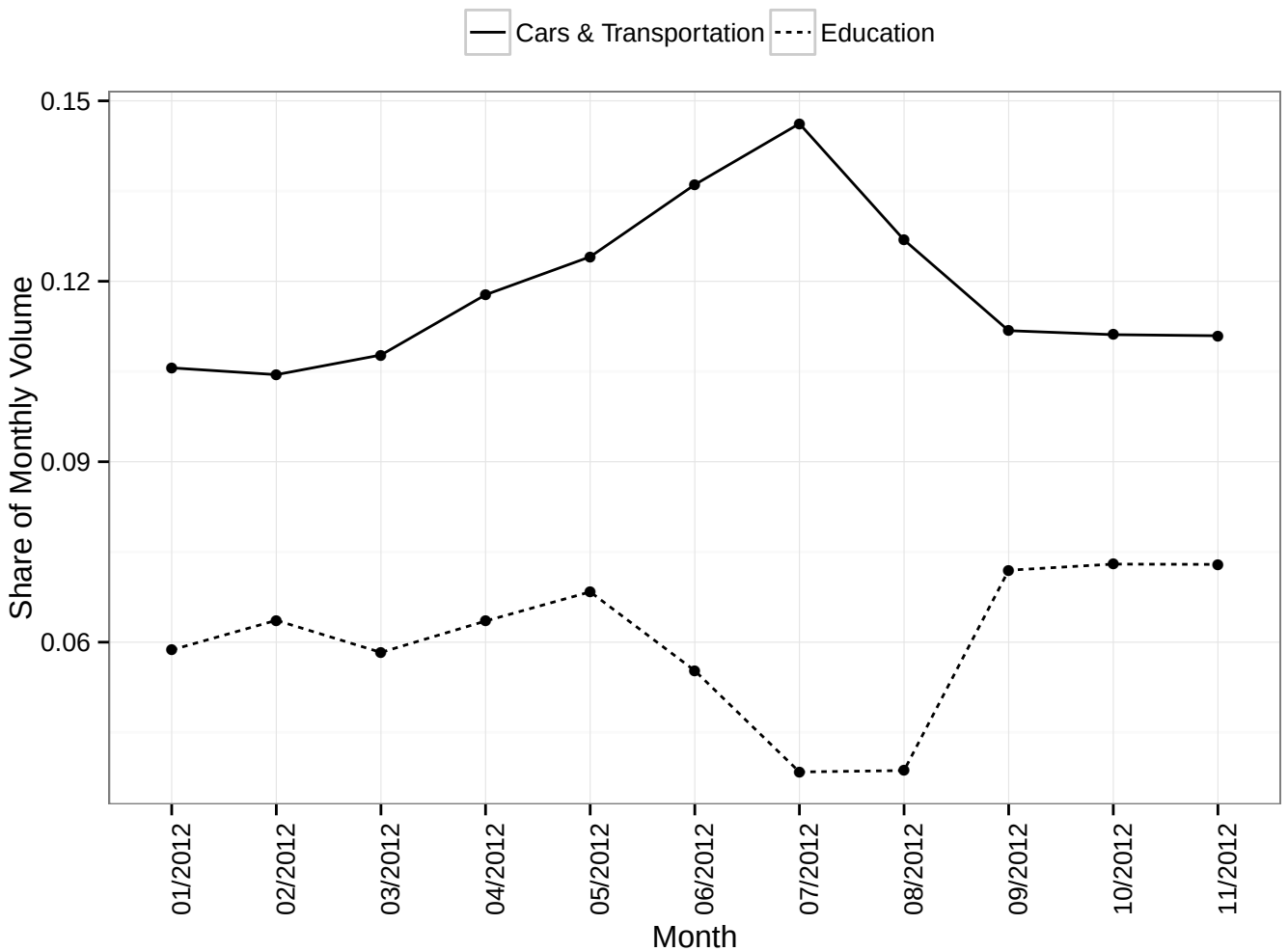
Results

Evolution of Categories over Time



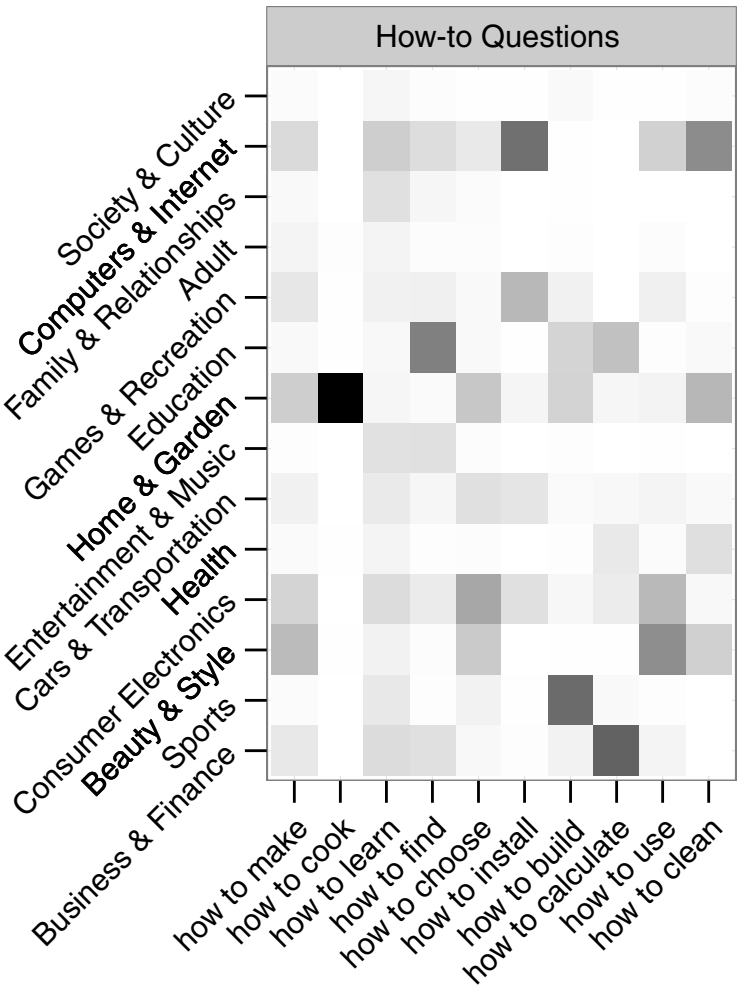
Results

Evolution of Categories Over Time: An Example



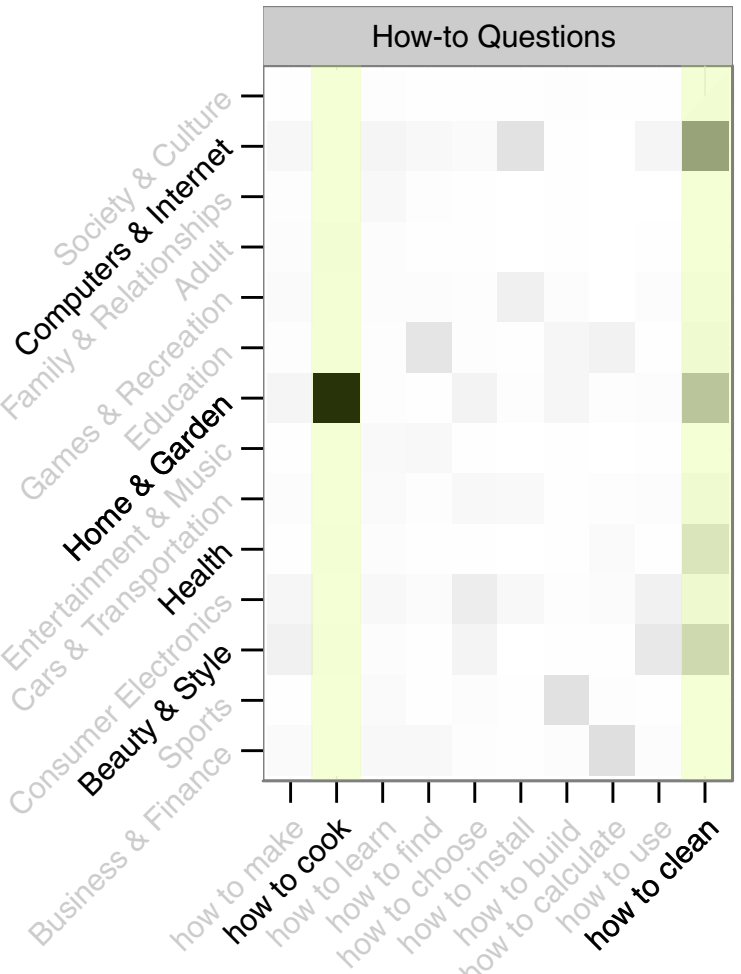
Results

Prefixes and Suffixes Across Categories



Results

Prefixes and Suffixes Across Categories



Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Asker behavior across categories

Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Asker behavior across categories

Future Work

1. Cross-language comparison
2. Deeper insights into asker behavior
3. More advanced classification schemes for short texts
4. Causes of increase in question query prevalence

Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Asker behavior across categories

Future Work

1. Cross-language comparison
2. Deeper insights into asker behavior
3. More advanced classification schemes for short texts
4. Causes of increase in question query prevalence

Thank you :-)

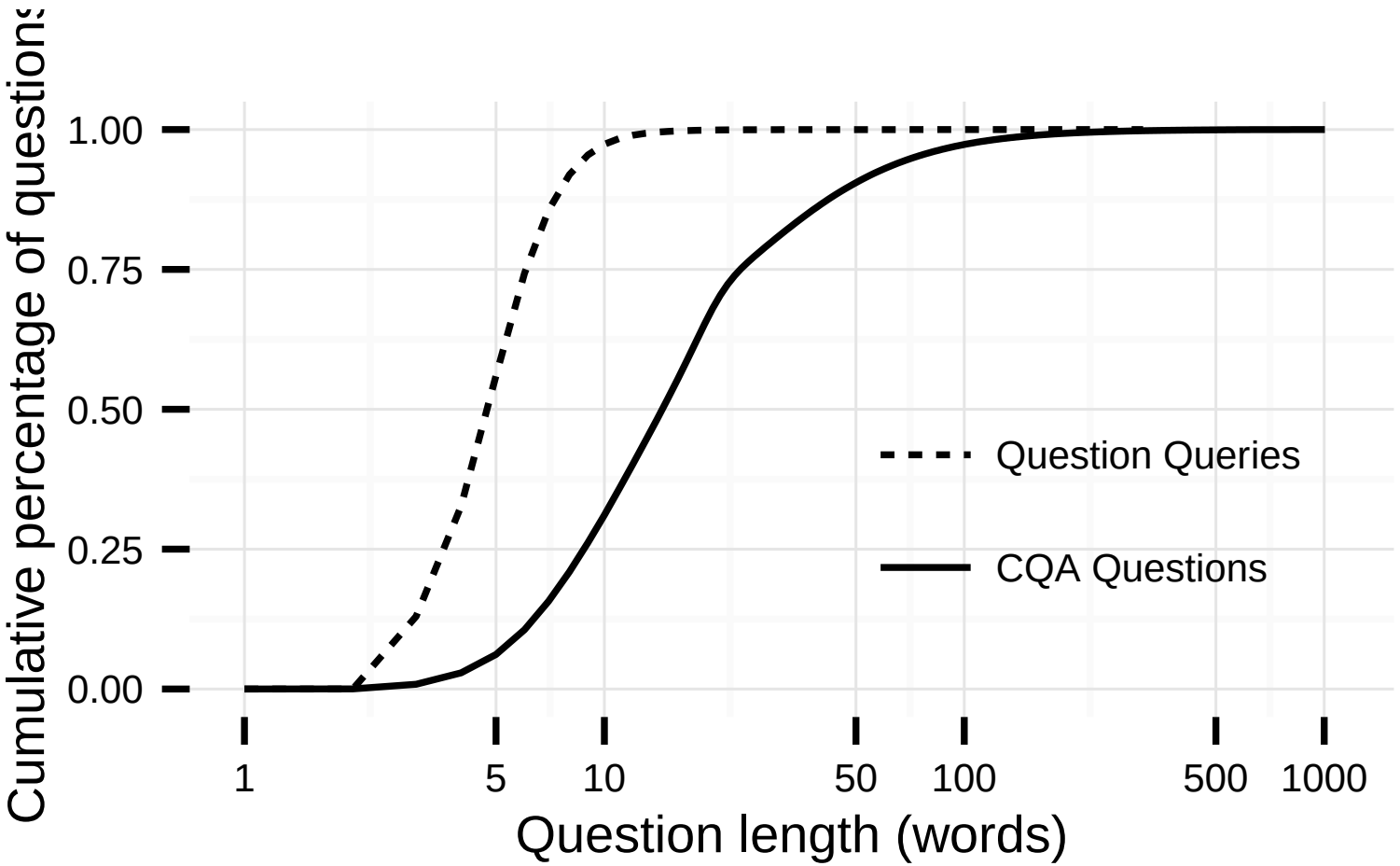
Acknowledgements

- Yandex (Alexey Gorodilov, Pavel Serdyukov, Alexander Sadovski)

- Mail.Ru (Andrey Oleynik)

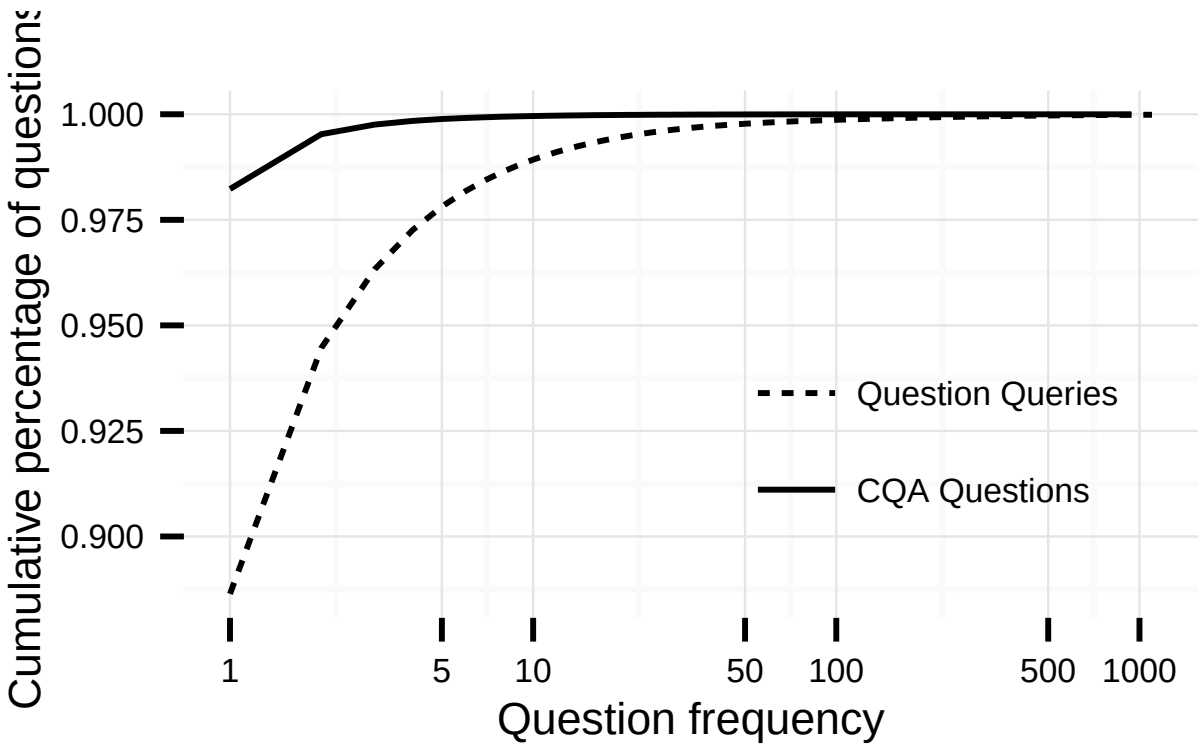
Appendices

Question Queries are Short



Appendices

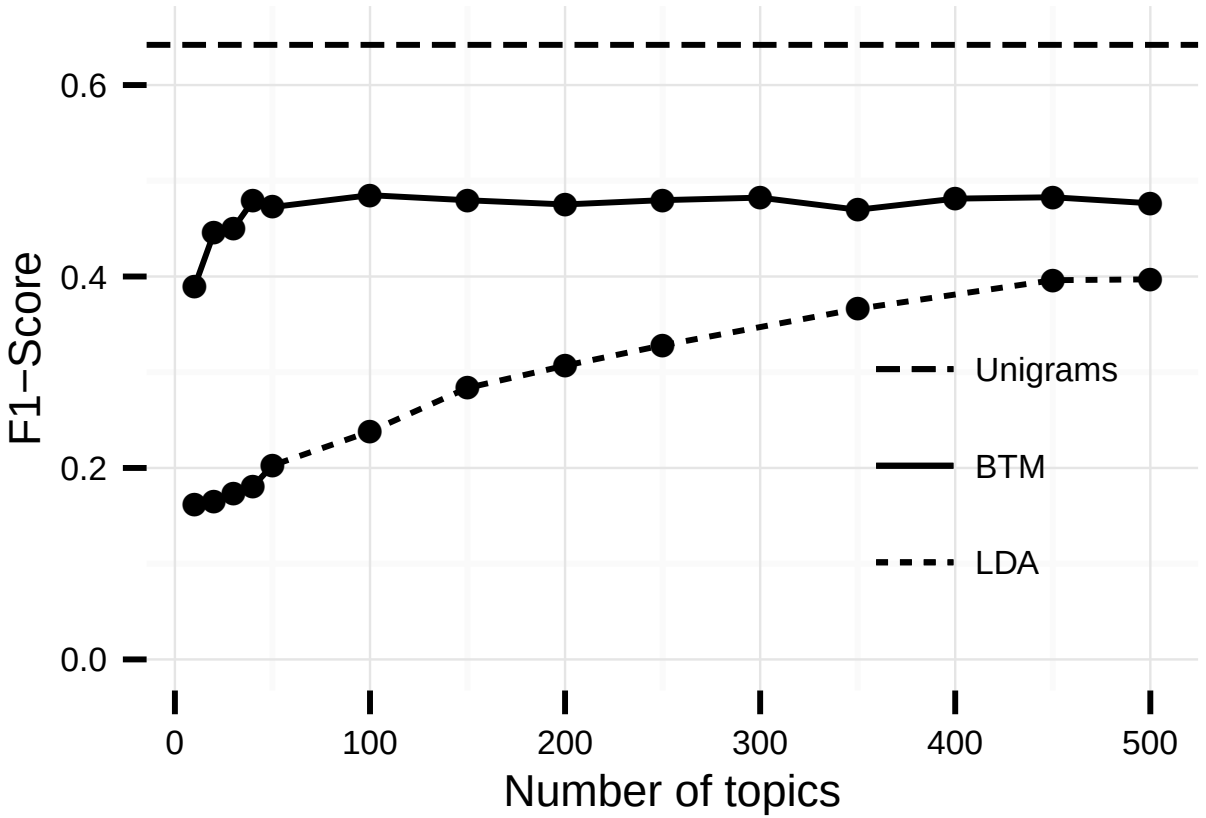
Question Queries are Unique



Appendices

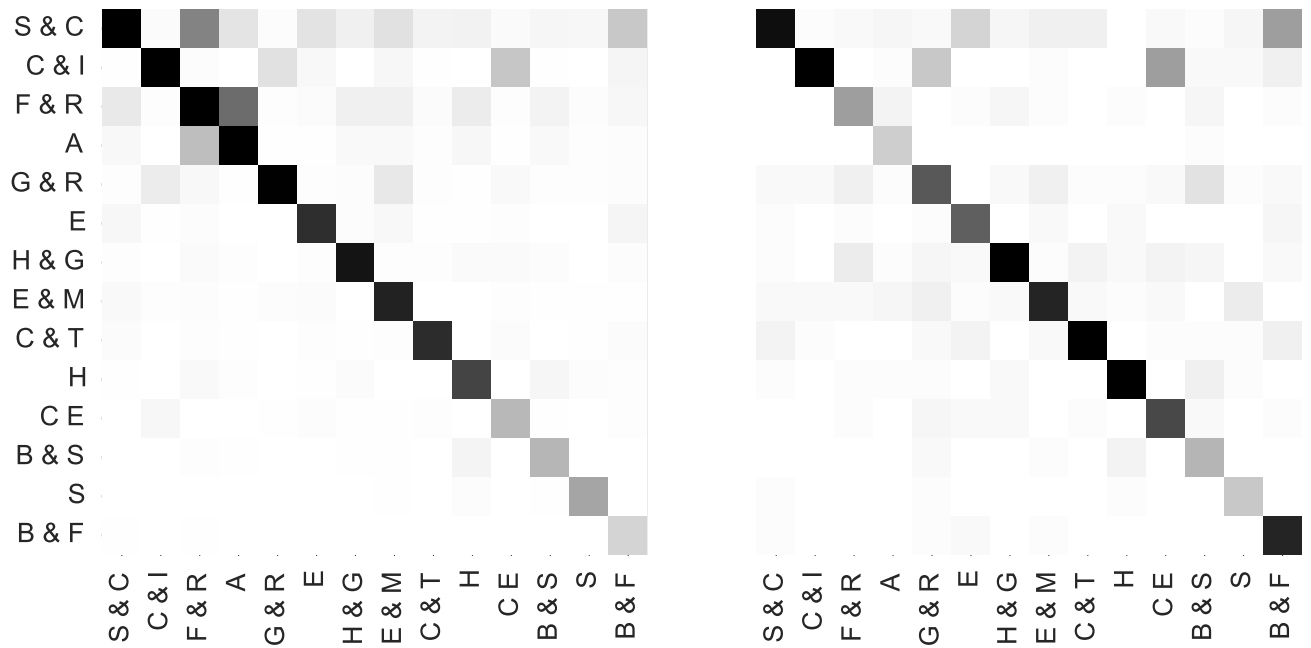
CQA Classification Performance

Validation Set (n ≈ 2 million)



Appendices

Confusion Matrix for Unigram Classifier



Left: CQA Validation set; Right: QQ test set

Appendices

References

Amanda Spink and H. Cenk Ozmutlu. Characteristics of question format web queries: An exploratory study. *Information processing & management*, 38 (4): 453–471, 2002.

Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of SIGIR 2007*, pages 231–238.

Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of CHI 2010*, pages 35–44.

Bo Pang and Ravi Kumar. Search in the lost sense of query: Question formulation in web search queries and its temporal changes. In *Proceedings of ACL 2011*, pages 135–140.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *Proceedings of the WSCD 2009 workshop*, pages 8–14.

Peter Bailey, Ryen W. White, Han Liu, and Giridhar Kumaran. Mining historic query trails to label long and rare search engine queries. *ACM Transactions on the Web*, 4(4):15, 2010.

Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *Proceedings of SIGIR 2008*, pages 339–346.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A Biterm Topic Model for Short Texts. In *Proceedings of WWW'13*, pages 1445–1456.