Pre-computing Search Features for Fast and Accurate Query Classification

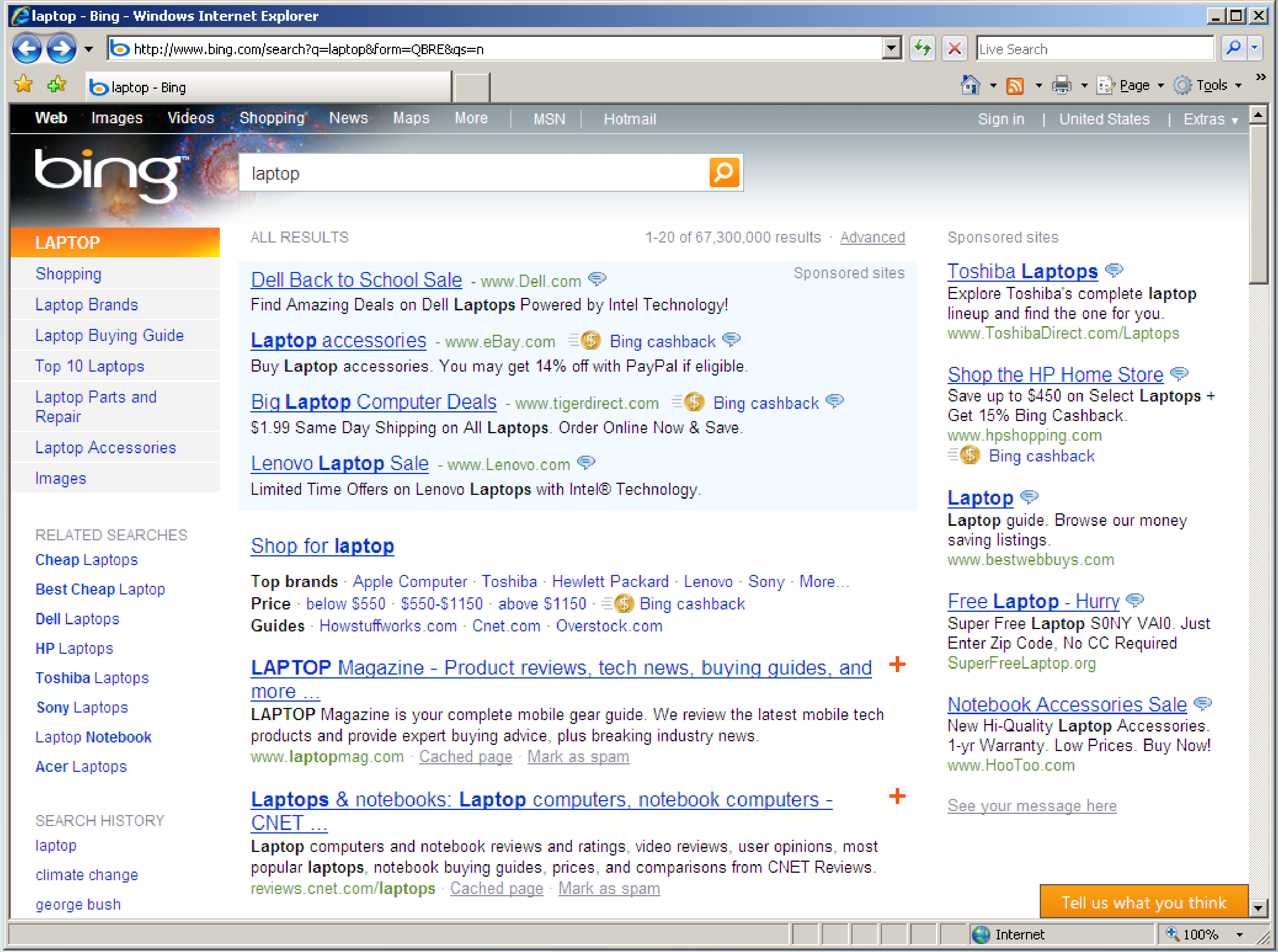
Arnd Christian König

Venkatesh Ganti

Xiao Li

Microsoft Research

Query Classification



LAPTOP
- Shopping
- Laptop Brands
- Laptop Buying Guide
- Top 10 Laptops
- Laptop Parts and Repair
- Laptop Accessories
- Images

RELATED SEARCHES
- Cheap Laptops
- Best Cheap Laptop
- Dell Laptops
- HP Laptops
- Toshiba Laptops
- Sony Laptops
- Laptop Notebook
- Acer Laptops

SEARCH HISTORY
- laptop
- climate change
- george bush

ALL RESULTS 1-20 of 67,300,000 results · Advanced

Sponsored sites

Dell Back to School Sale - www.Dell.com
Find Amazing Deals on Dell **Laptops** Powered by Intel Technology!

Laptop accessories - www.eBay.com Bing cashback
Buy **Laptop** accessories. You may get 14% off with PayPal if eligible.

Big **Laptop** Computer Deals - www.tigerdirect.com Bing cashback
$1.99 Same Day Shipping on All **Laptops**. Order Online Now & Save.

Lenovo **Laptop** Sale - www.Lenovo.com
Limited Time Offers on Lenovo **Laptops** with Intel® Technology.

Shop for **laptop**

Top brands · Apple Computer · Toshiba · Hewlett Packard · Lenovo · Sony · More...
Price · below $550 · $550-$1150 · above $1150 · Bing cashback
Guides · Howstuffworks.com · Cnet.com · Overstock.com

LAPTOP Magazine - Product reviews, tech news, buying guides, and more ... +
LAPTOP Magazine is your complete mobile gear guide. We review the latest mobile tech products and provide expert buying advice, plus breaking industry news.
www.**laptop**mag.com · Cached page · Mark as spam

Laptops & notebooks: **Laptop** computers, notebook computers - CNET ... +
Laptop computers and notebook reviews and ratings, video reviews, user opinions, most popular **laptops**, notebook buying guides, prices, and comparisons from CNET Reviews.
reviews.cnet.com/**laptops** · Cached page · Mark as spam

Sponsored sites

Toshiba **Laptops**
Explore Toshiba's complete **laptop** lineup and find the one for you.
www.ToshibaDirect.com/Laptops

Shop the HP Home Store
Save up to $450 on Select **Laptops** + Get 15% Bing Cashback.
www.hpshopping.com
Bing cashback

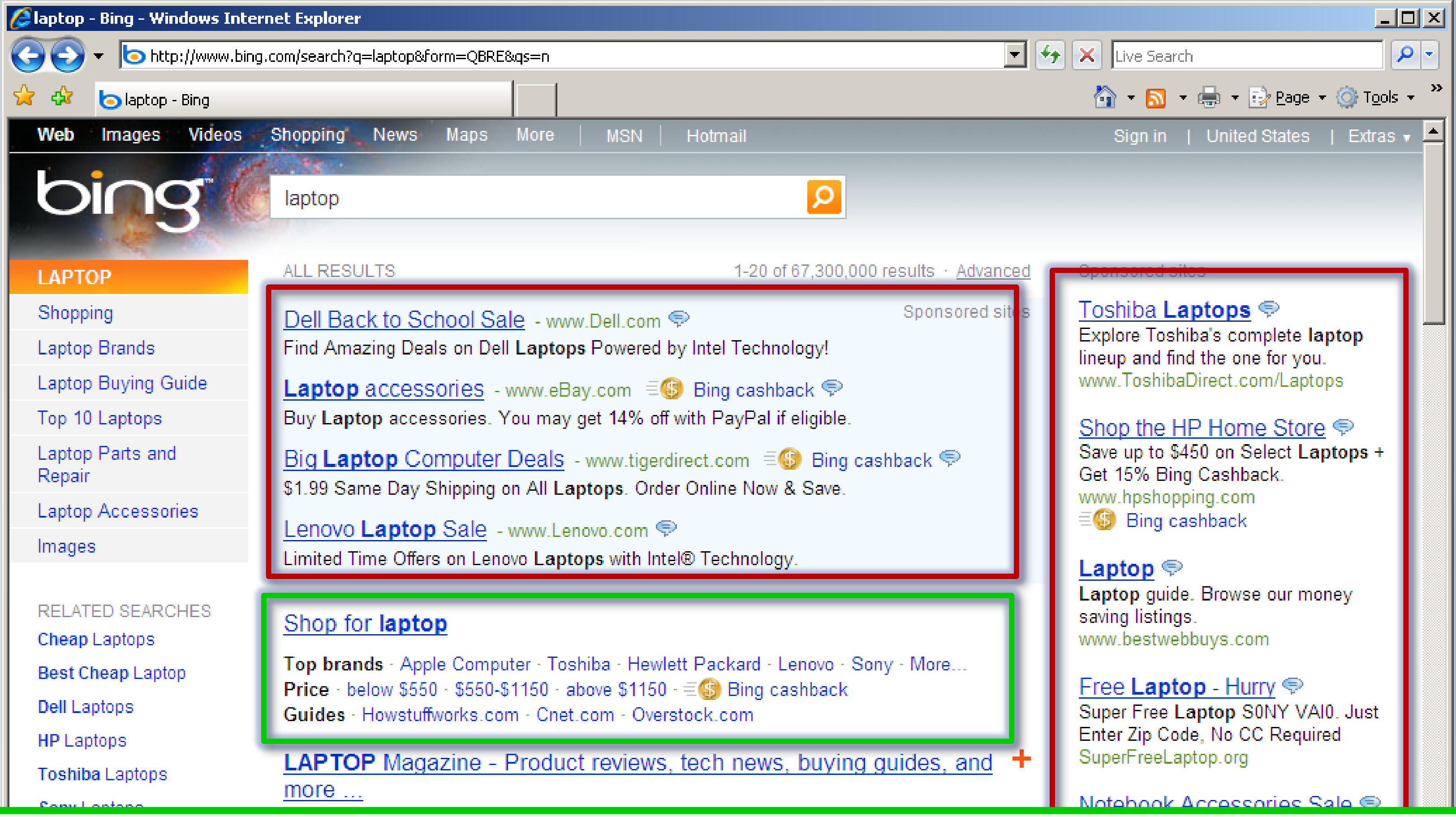
Laptop
Laptop guide. Browse our money saving listings.
www.bestwebbuys.com

Free **Laptop** - Hurry
Super Free **Laptop** SONY VAIO. Just Enter Zip Code, No CC Required
SuperFreeLaptop.org

Notebook Accessories Sale
New Hi-Quality **Laptop** Accessories. 1-yr Warranty. Low Prices. Buy Now!
www.HooToo.com

See your message here

Query Classification



laptop - Bing - Windows Internet Explorer

http://www.bing.com/search?q=laptop&form=QBRE&qs=n

laptop

LAPTOP

- Shopping
- Laptop Brands
- Laptop Buying Guide
- Top 10 Laptops
- Laptop Parts and Repair
- Laptop Accessories
- Images

RELATED SEARCHES

- Cheap Laptops
- Best Cheap Laptop
- Dell Laptops
- HP Laptops
- Toshiba Laptops
- Sony Laptops

ALL RESULTS 1-20 of 67,300,000 results · Advanced

Sponsored sites

Dell Back to School Sale - www.Dell.com
Find Amazing Deals on Dell **Laptops** Powered by Intel Technology!

Laptop accessories - www.eBay.com Bing cashback
Buy **Laptop** accessories. You may get 14% off with PayPal if eligible.

Big **Laptop** Computer Deals - www.tigerdirect.com Bing cashback
$1.99 Same Day Shipping on All **Laptops**. Order Online Now & Save.

Lenovo **Laptop** Sale - www.Lenovo.com
Limited Time Offers on Lenovo **Laptops** with Intel® Technology.

Shop for **laptop**

Top brands · Apple Computer · Toshiba · Hewlett Packard · Lenovo · Sony · More...
Price · below $550 · $550-$1150 · above $1150 · Bing cashback
Guides · Howstuffworks.com · Cnet.com · Overstock.com

LAPTOP Magazine - Product reviews, tech news, buying guides, and more ...

Sponsored sites

Toshiba **Laptops**
Explore Toshiba's complete **laptop** lineup and find the one for you.
www.ToshibaDirect.com/Laptops

Shop the HP Home Store
Save up to $450 on Select **Laptops** + Get 15% Bing Cashback.
www.hpshopping.com
Bing cashback

Laptop
Laptop guide. Browse our money saving listings.
www.bestwebbuys.com

Free **Laptop** - Hurry
Super Free **Laptop** SONY VAIO. Just Enter Zip Code, No CC Required
SuperFreeLaptop.org

Notebook Accessories Sale

Query Classification important for

- Matching advertisements for a query
- Retrieval of additional non-web search results from verticals
- Load optimization for search verticals

Problem Statement

Query Classification:

- **Problem:** ~3 words in queries => little 'signal' for classification.
- Large vocabulary size => large, sparse feature space.
- Difficult to generalize across queries.

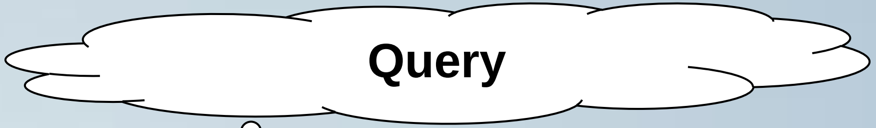
Post-Retrieval Features:

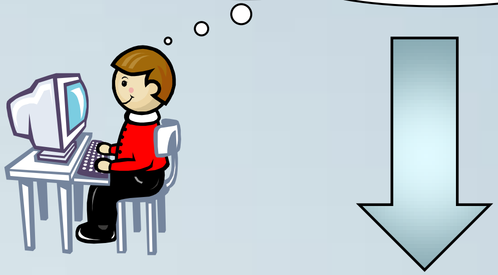
- Use search to obtain more context to derive features.

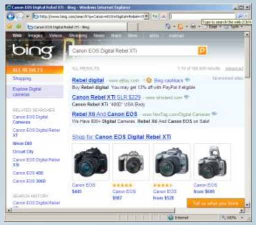
Post-Retrieval features

General approach:

- Issue the search query against a document corpus .
- Identify relevant sub-components of top results (e.g., titles, captions, key terms, etc.)
- Derive additional features from these components.






Canon EOS Digital Rebel XTi - Bing - Windows Internet Explorer

Canon EOS Digital Rebel XTi

ALL RESULTS

Rebel digital - www.eBay.com - Bing cashback

Buy Rebel digital. You may get 13% off with PayPal if eligible

Canon Rebel XTi SLR $229 - www.stroland.com

Canon Rebel XTi "400D" USA Body

Rebel Xti And Canon EOS - www.NexTag.com/Digital-Cameras

We Have 800+ Digital Cameras. Rebel Xti And Canon EOS on Sale!

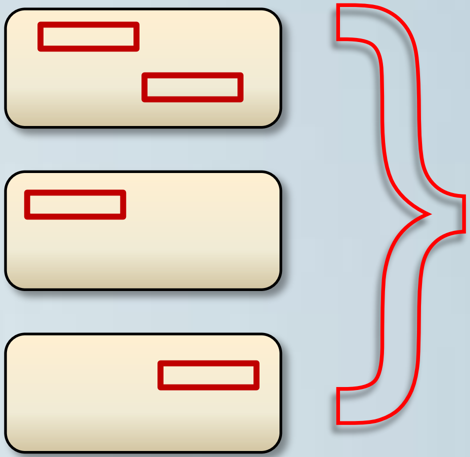
Shop for Canon EOS Digital Rebel XTi

Canon EOS $445

Canon EOS $987

Canon EOS from $528

Canon EOS from $680



$$F = \sum_{d \in \text{Result}} (\cdots)$$

Problem Statement

Query Classification:

- **Problem:** ~3 words in queries => little 'signal' for classification.
- Large vocabulary size => large, sparse feature space.
- Difficult to generalize across queries.

Post-Retrieval Features:

- Use search to obtain more context to derive features.
- => significant improvements in classification accuracy.
- **Problem: Search Latency**
  - Even slight (100 ms) increases in latency decrease user satisfaction, increase in fraction of abandoned searches.

⇒ **Task:** Realize benefits of post-retrieval features at low overhead.

Our Approach

Query: Low-light snapshots

Classification Task: Product-intent
Tags: Entity Categories

Features based on the incidence of tags in the documents returned in response to a query.

$\Rightarrow$ Small feature space, features generalize across queries.
$\Rightarrow$ Less information to store, helping pre-computation.

Tag-Ratios

$q,t) :=$

$t \in result(q)$

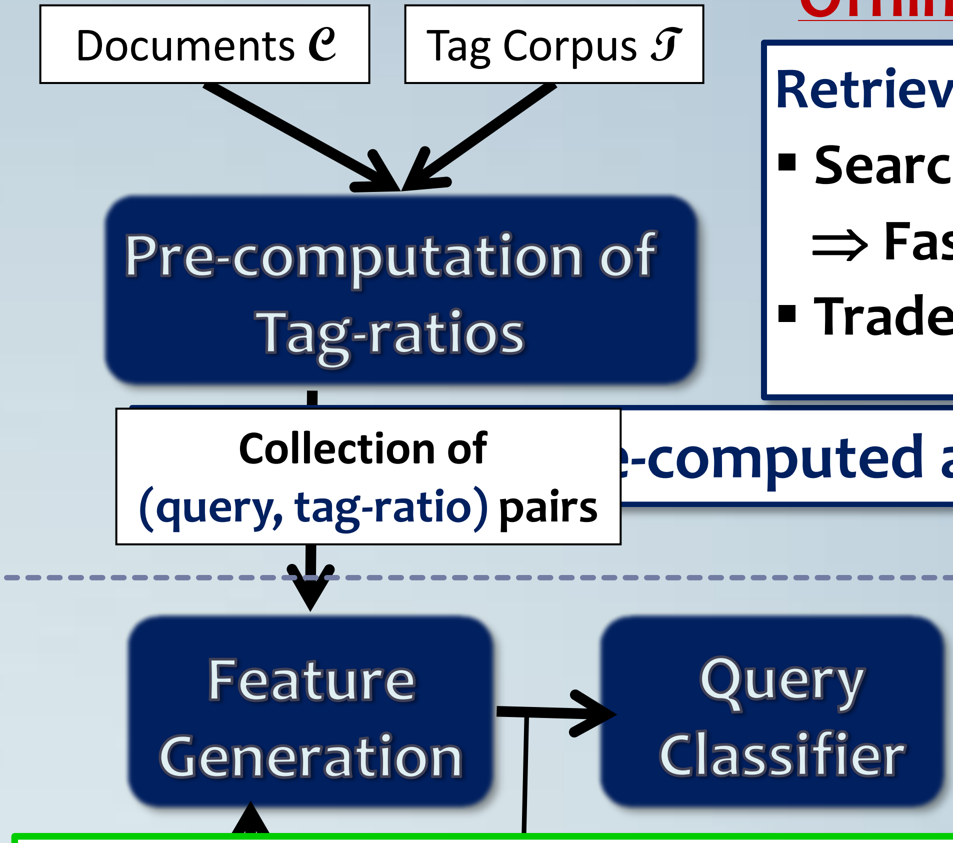
$esult(q)|$

Other examples:

Corpus: Sponsored Search Bids
- *Tags*: Advertiser-IDs
- Advertisers can be thought of as 'topics'

Corpus: Wikipedia
- *Tags*: Wikipedia-Category Tags

Our Approach (II)



Documents $\mathcal{C}$

Tag Corpus $\mathcal{T}$

Pre-computation of Tag-ratios

Collection of (query, tag-ratio) pairs

Feature Generation

Query Classifier

Offline

Retrieval Semantics: word-containment

- Search engine not involved in retrieval

  ⇒ Fast pre-computation of query sets

- Tradeoff: result relevance vs. result size

-computed and indexed in memory

Online

The rest of this talk:

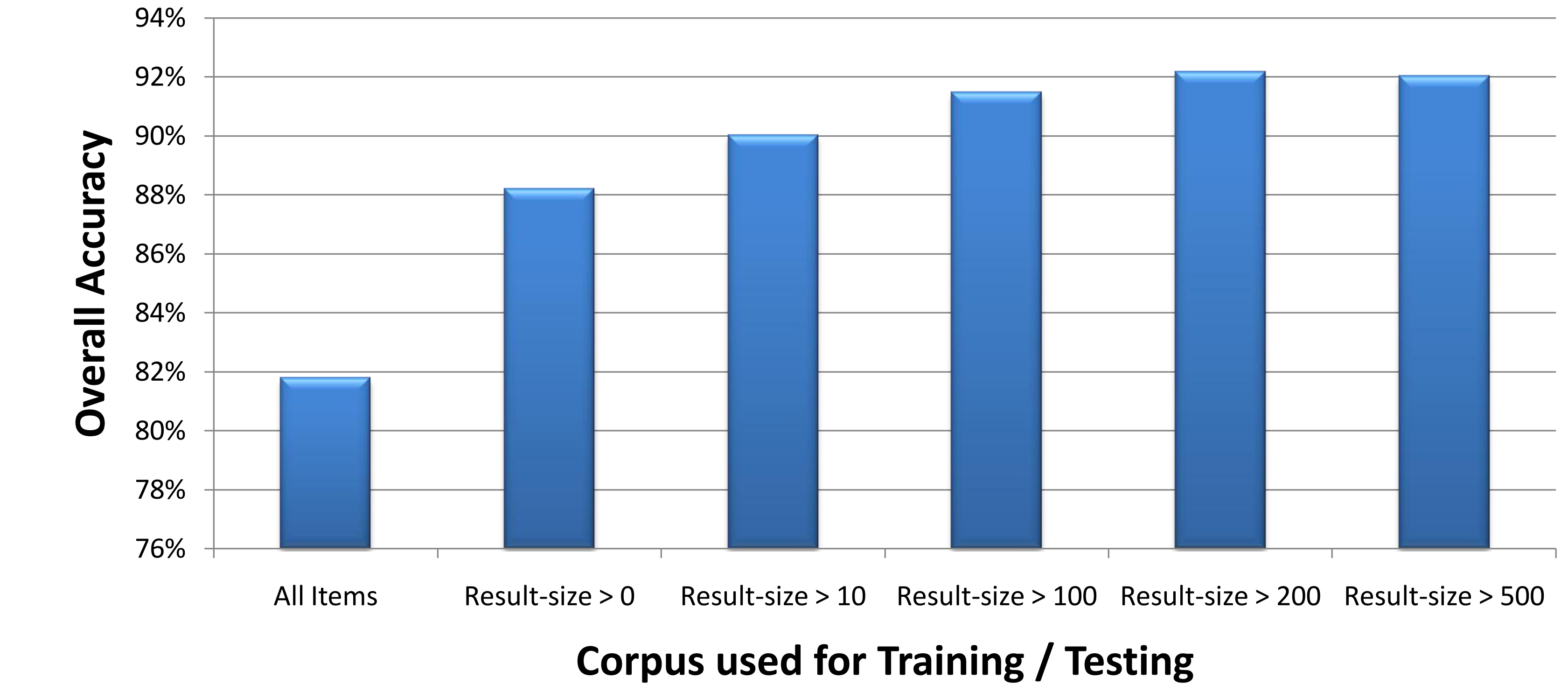
- How do we generate features from the ratios?
- Size-constraints: for which queries do we pre-compute ratios?
- How do we deal with query that is not pre-computed?

Creating Features from Tag Ratios

Features = Ratios?

$$F = \left[ratio(q, t_1), \cdots, ratio(q, t_k) \right]$$

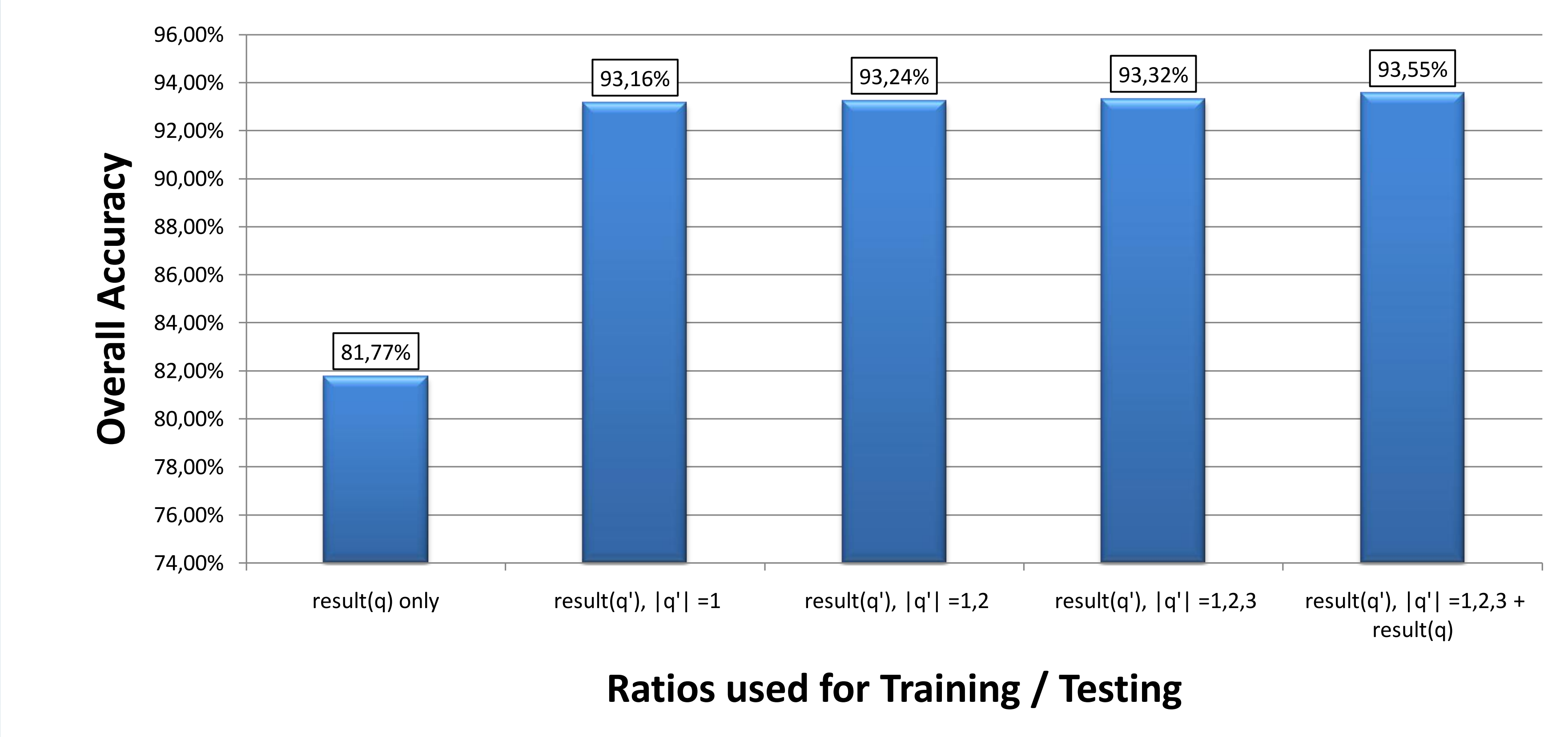
- **Problem I: Small result sizes**



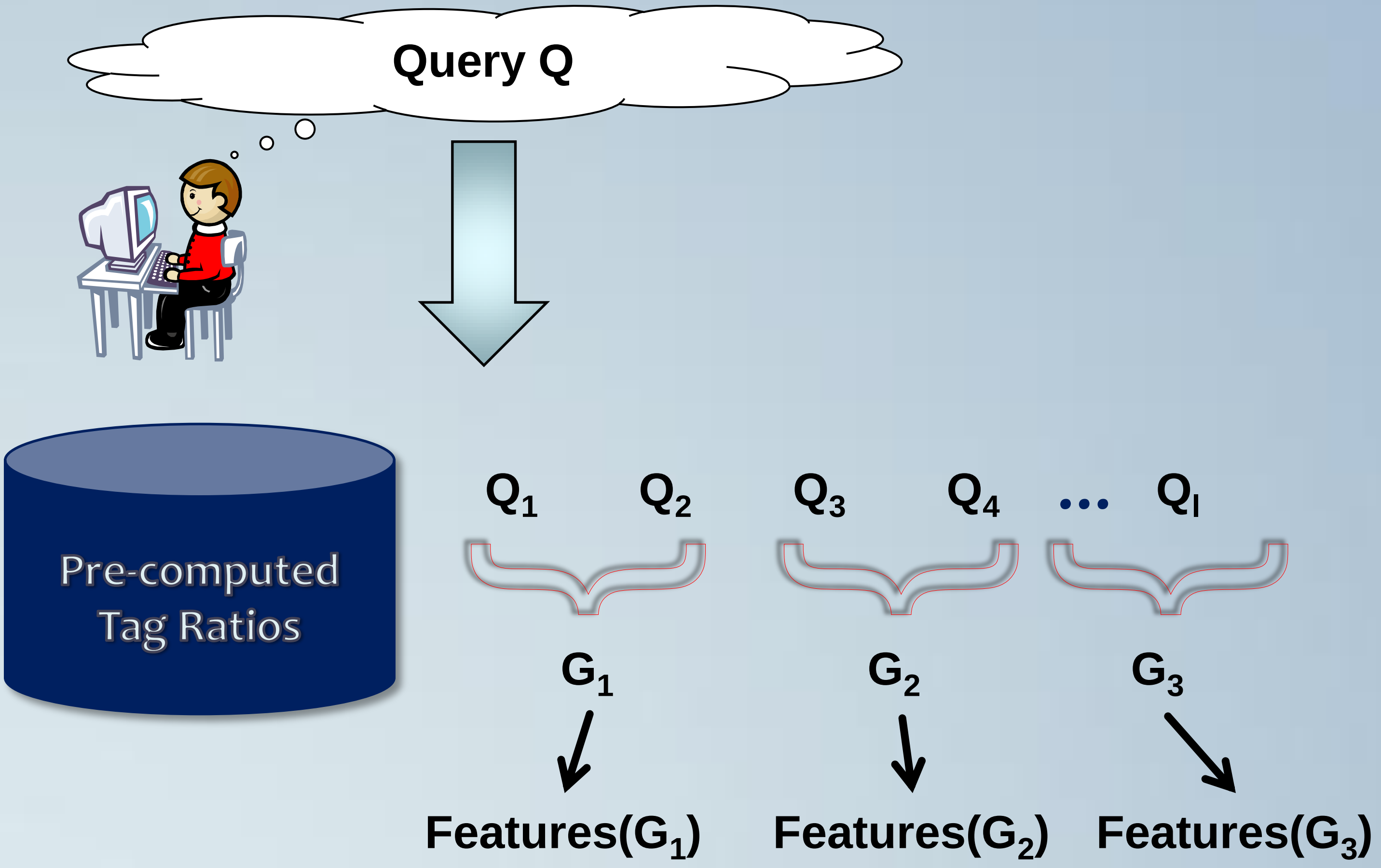
Creating Features from Tag Ratios

Features = Ratios?

$$F = \left[ratio(q,t_1), \cdots, ratio(q,t_k) \right]$$



Creating Features from Tag Ratios



Query Q

Pre-computed Tag Ratios

Q_1 Q_2 Q_3 Q_4 … Q_l

G_1 G_2 G_3

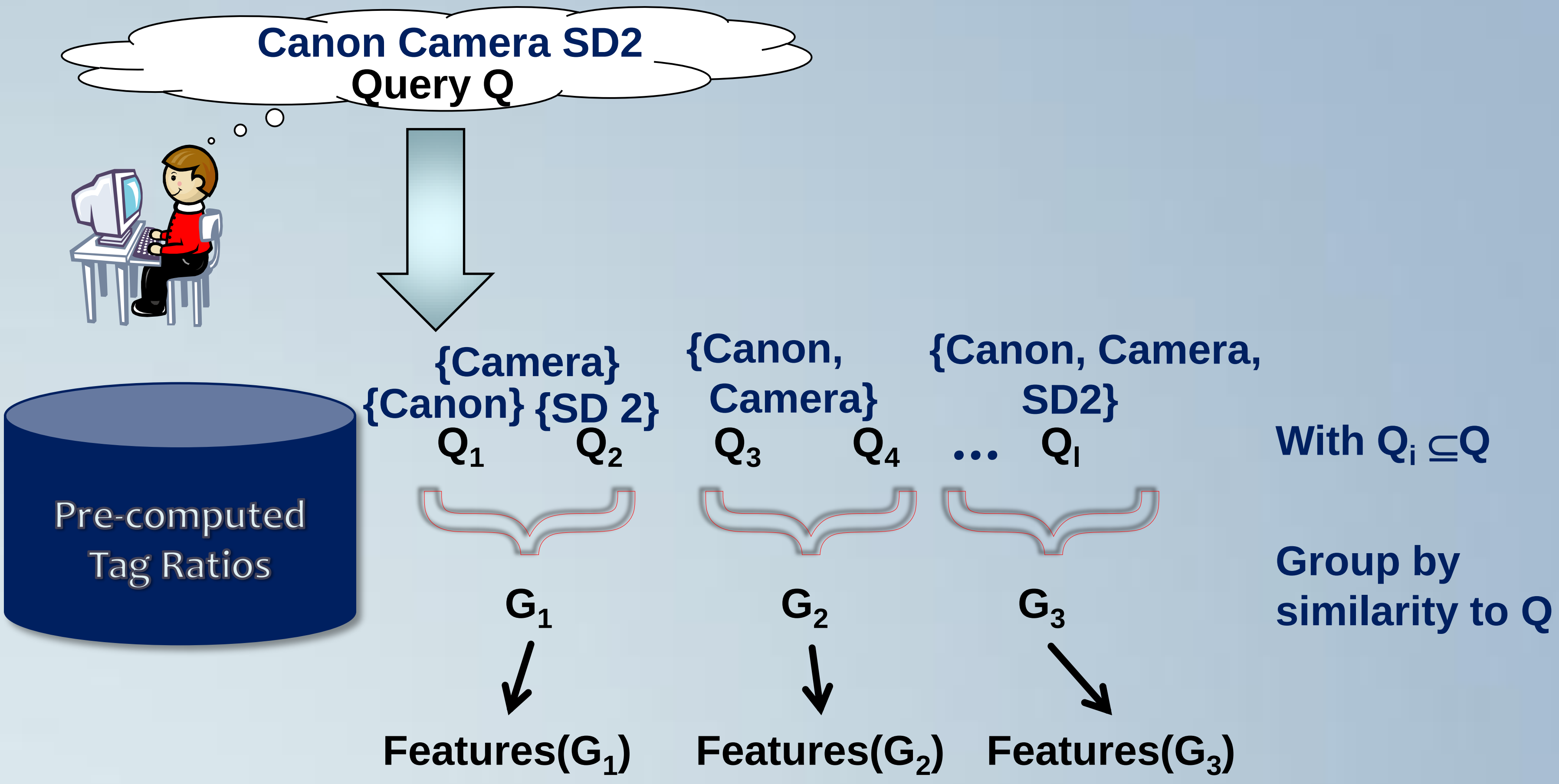
Features(G_1) Features(G_2) Features(G_3)

With $Q_i \subseteq Q$

Group by similarity to Q

Features based on Aggregates over ratios in a group, such as *SUM, AVG, STDIV, MAX, MIN*, etc…

Creating Features from Tag Ratios



Canon Camera SD2
Query Q

Pre-computed
Tag Ratios

{Camera}
{Canon} {SD 2} {Canon, Camera} {Canon, Camera, SD2}

Q_1 Q_2 Q_3 Q_4 … Q_l

With $Q_i \subseteq Q$

G_1 G_2 G_3

Group by
similarity to Q

Features(G_1) Features(G_2) Features(G_3)

Features based on Aggregates over ratios in a group, such as
SUM, AVG, STDIV, MAX, MIN, etc…

Selecting queries to precompute

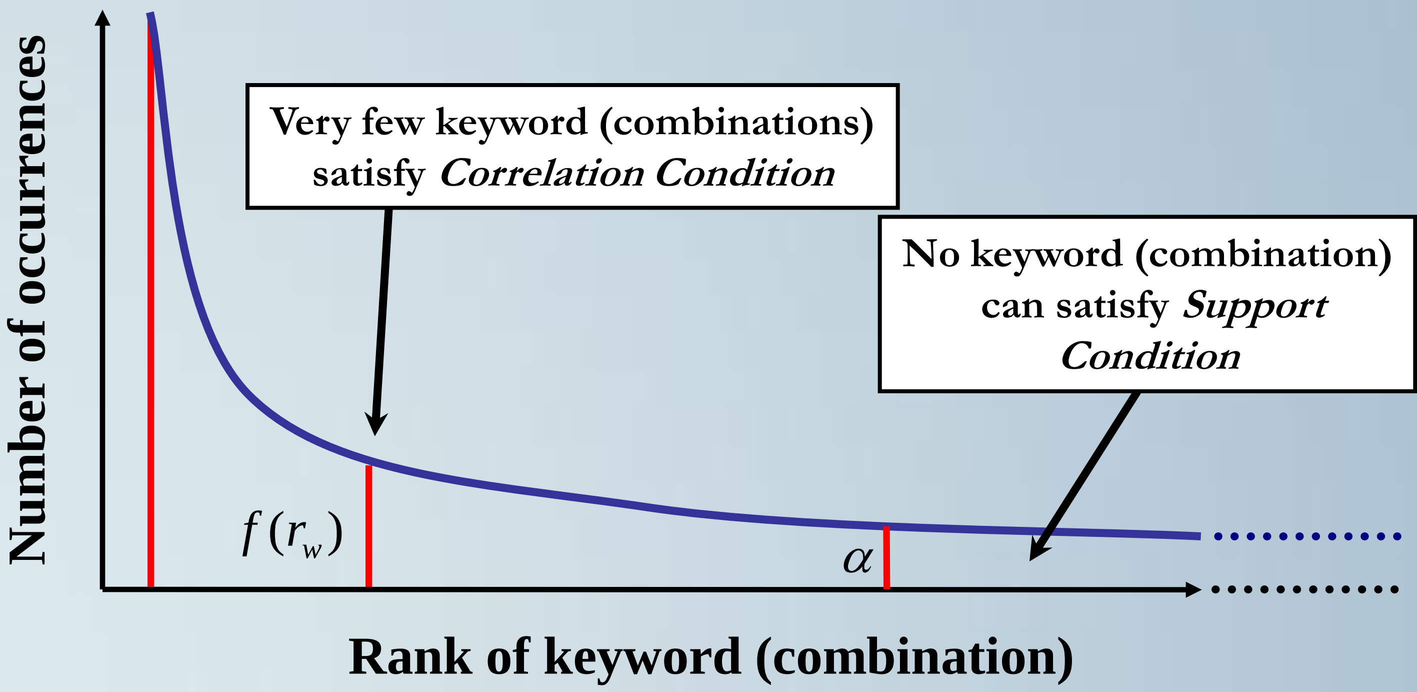
- $|V| > 10^7$ => intractable # of keyword combinations to pre-compute

Pruning Logic

- **Short queries:** limit query-length to w_{max} words.
- **Significant correlation:**

$$ratio(q,t) \geq \Theta_{high} \frac{|\#\text{tags } t \in D|}{D} \qquad ratio(q,t) \leq \Theta_{low} \frac{|\#\text{tags } t \in D|}{D}$$

- **Ratio-support:** $|result_C(q)| \geq \alpha$



Very few keyword (combinations) satisfy *Correlation Condition*

No keyword (combination) can satisfy *Support Condition*

Experimental Evaluation

Task I: Indentifying 'Consumer-Electronics' queries

- C = Wikipedia, T = Entity-Categories (contained in pages)
- Accuracy: 93.0% (n-grams only)
  93.2% (n-grams + Brand/Models/Product Type/ Product Attribute lexicons))
  95.6% (Tag ratios only)
  96.5% (Tag ratios+ n-grams)

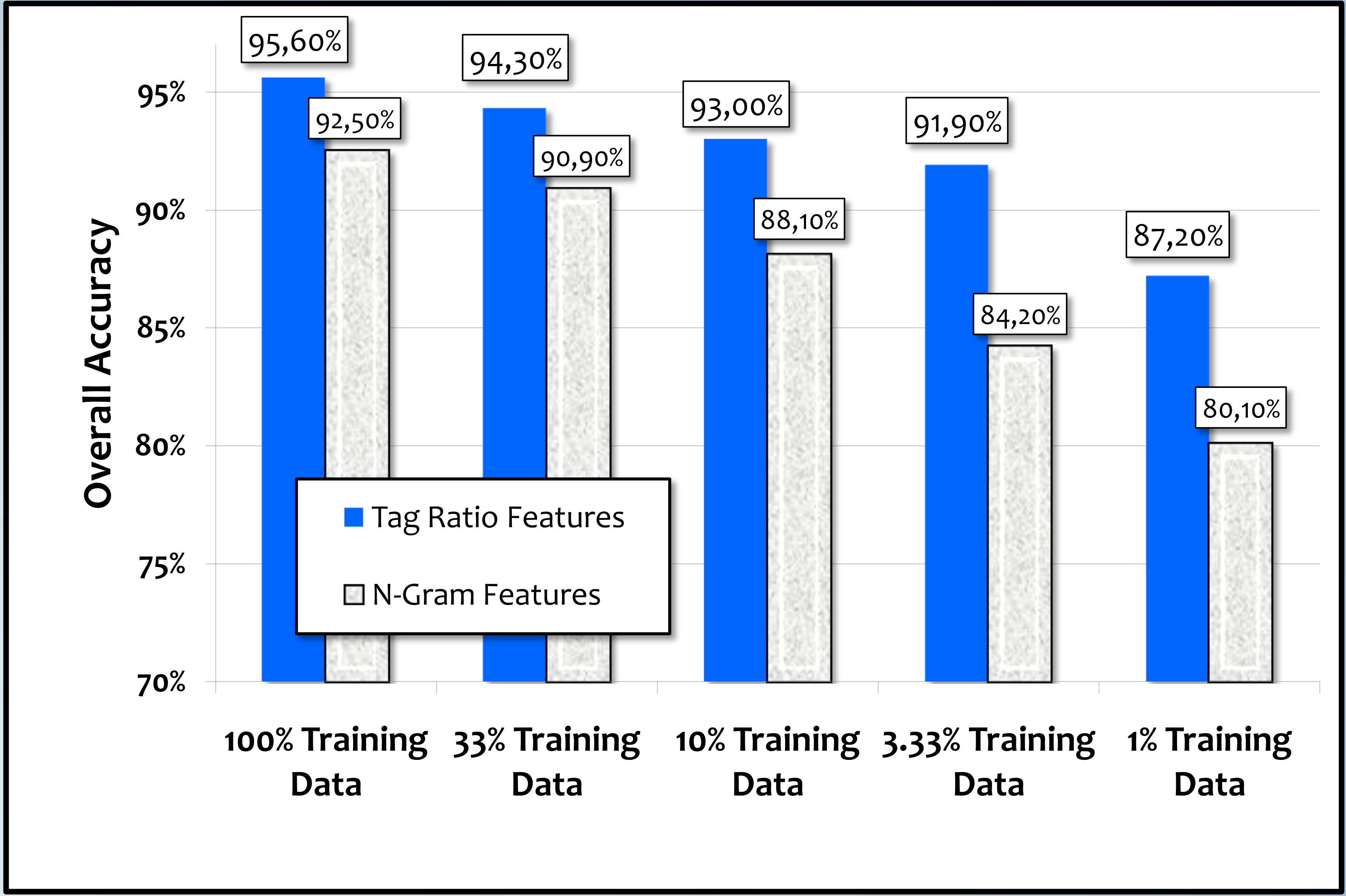
Task II: Indentifying 'Retail' queries

- C = Wikipedia, T = Top Wikipedia Categories (contained in pages)
- + C = Sponsored Search Bids, T = Advertiser IDs (top advertisers)
- Large training corpus (~330K labeled examples)
- Accuracy: 92.5% (n-grams only) => 93.3% (Tag Ratios+ ngrams)

Task III: Indentifying 'Heath'-related queries

- Same corpora/tags as before
- Very large training corpus (~800K labeled examples)
- Accuracy: 98.2% (n-grams only) => 98.8% (Tag Ratios + ngrams)

Experimental Evaluation: Generalization



Experimental Evaluation: Query Selection

Using earlier classification tasks, we evaluate features based on:

- Single-Word queries only
- Single-word queries + selected query/ratio combinations
- All queries in training/test data + all subsets

Results:

- Pruning results in very large reduction in space of ratios to store (Θ_{low} = 0.8, Θ_{high} = 1.2 => 0.8% of ratios (for frequent keywords) remain).
- Differences in classification accuracy slight: (0.17% or less)

Many thanks!

Any Questions?