Deep Character-Level Click-Through Rate Prediction for Sponsored Search

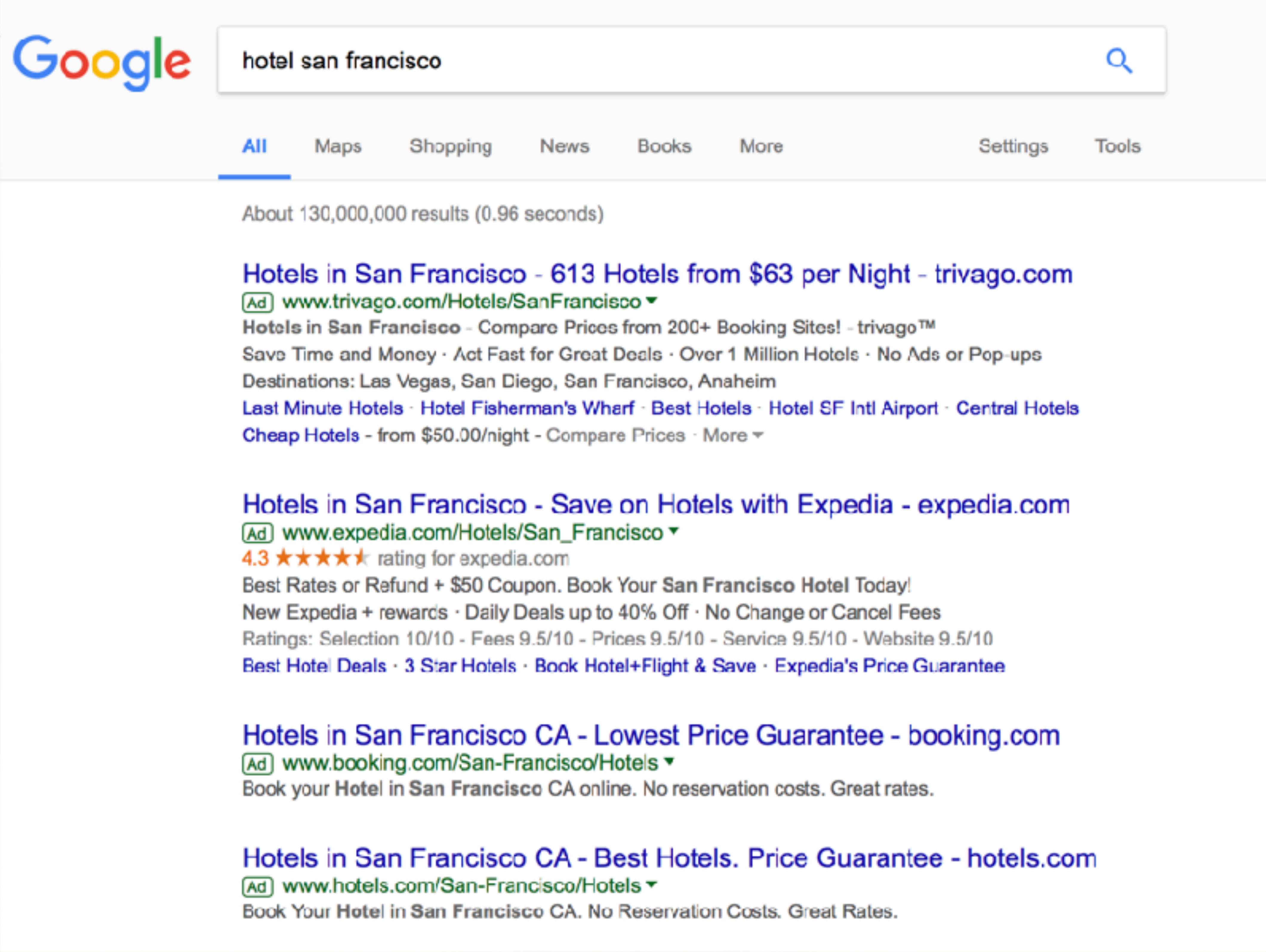
Bora Edizel - Phd Student UPF
Amin Mantrach - Criteo Research
Xiao Bai - Oath

This work was done at Yahoo and will be presented as full paper in SIGIR '17.

Outline

- Problem
- Related Work
- Motivation
- Contributions
- Research Questions
- Model Description
- Experimental Results

Sponsored Search: Text Ads



About 130,000,000 results (0.96 seconds)

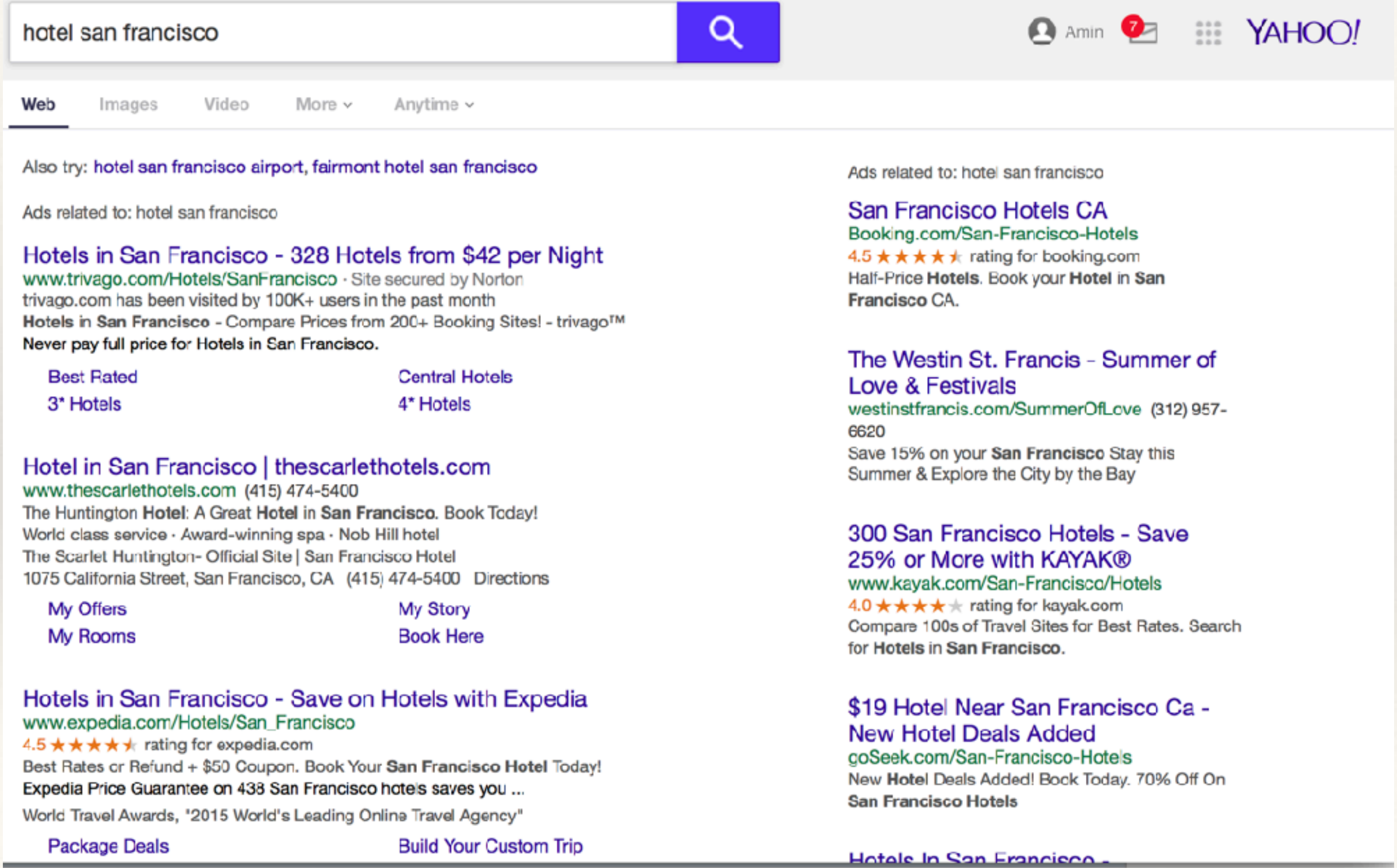
Hotels in San Francisco - 613 Hotels from $63 per Night - trivago.com
Ad www.trivago.com/Hotels/SanFrancisco ▾
Hotels in San Francisco - Compare Prices from 200+ Booking Sites! - trivago™
Save Time and Money · Act Fast for Great Deals · Over 1 Million Hotels · No Ads or Pop-ups
Destinations: Las Vegas, San Diego, San Francisco, Anaheim
Last Minute Hotels · Hotel Fisherman's Wharf · Best Hotels · Hotel SF Intl Airport · Central Hotels
Cheap Hotels - from $50.00/night - Compare Prices · More ▾

Hotels in San Francisco - Save on Hotels with Expedia - expedia.com
Ad www.expedia.com/Hotels/San_Francisco ▾
4.3 ★★★★★ rating for expedia.com
Best Rates or Refund + $50 Coupon. Book Your San Francisco Hotel Today!
New Expedia + rewards · Daily Deals up to 40% Off · No Change or Cancel Fees
Ratings: Selection 10/10 - Fees 9.5/10 - Prices 9.5/10 - Service 9.5/10 - Website 9.5/10
Best Hotel Deals · 3 Star Hotels · Book Hotel+Flight & Save · Expedia's Price Guarantee

Hotels in San Francisco CA - Lowest Price Guarantee - booking.com
Ad www.booking.com/San-Francisco/Hotels ▾
Book your Hotel in San Francisco CA online. No reservation costs. Great rates.

Hotels in San Francisco CA - Best Hotels. Price Guarantee - hotels.com
Ad www.hotels.com/San-Francisco/Hotels ▾
Book Your Hotel in San Francisco CA. No Reservation Costs. Great Rates.

Sponsored Search: Text Ads



Also try: hotel san francisco airport, fairmont hotel san francisco

Ads related to: hotel san francisco

Hotels in San Francisco - 328 Hotels from $42 per Night
www.trivago.com/Hotels/SanFrancisco · Site secured by Norton
trivago.com has been visited by 100K+ users in the past month
Hotels in **San Francisco** - Compare Prices from 200+ Booking Sites! - trivago™
Never pay full price for Hotels in San Francisco.

Best Rated
Central Hotels
3* Hotels
4* Hotels

Hotel in San Francisco | thescarlethotels.com
www.thescarlethotels.com (415) 474-5400
The Huntington **Hotel**: A Great **Hotel** in **San Francisco**. Book Today!
World class service · Award-winning spa · Nob Hill hotel
The Scarlet Huntington- Official Site | San Francisco Hotel
1075 California Street, San Francisco, CA (415) 474-5400 Directions

My Offers
My Story
My Rooms
Book Here

Hotels in San Francisco - Save on Hotels with Expedia
www.expedia.com/Hotels/San_Francisco
4.5 ★★★★★ rating for expedia.com
Best Rates or Refund + $50 Coupon. Book Your **San Francisco Hotel** Today!
Expedia Price Guarantee on 438 San Francisco hotels saves you ...
World Travel Awards, "2015 World's Leading Online Travel Agency"

Package Deals
Build Your Custom Trip

Ads related to: hotel san francisco

San Francisco Hotels CA
Booking.com/San-Francisco-Hotels
4.5 ★★★★★ rating for booking.com
Half-Price **Hotels**. Book your **Hotel** in **San Francisco** CA.

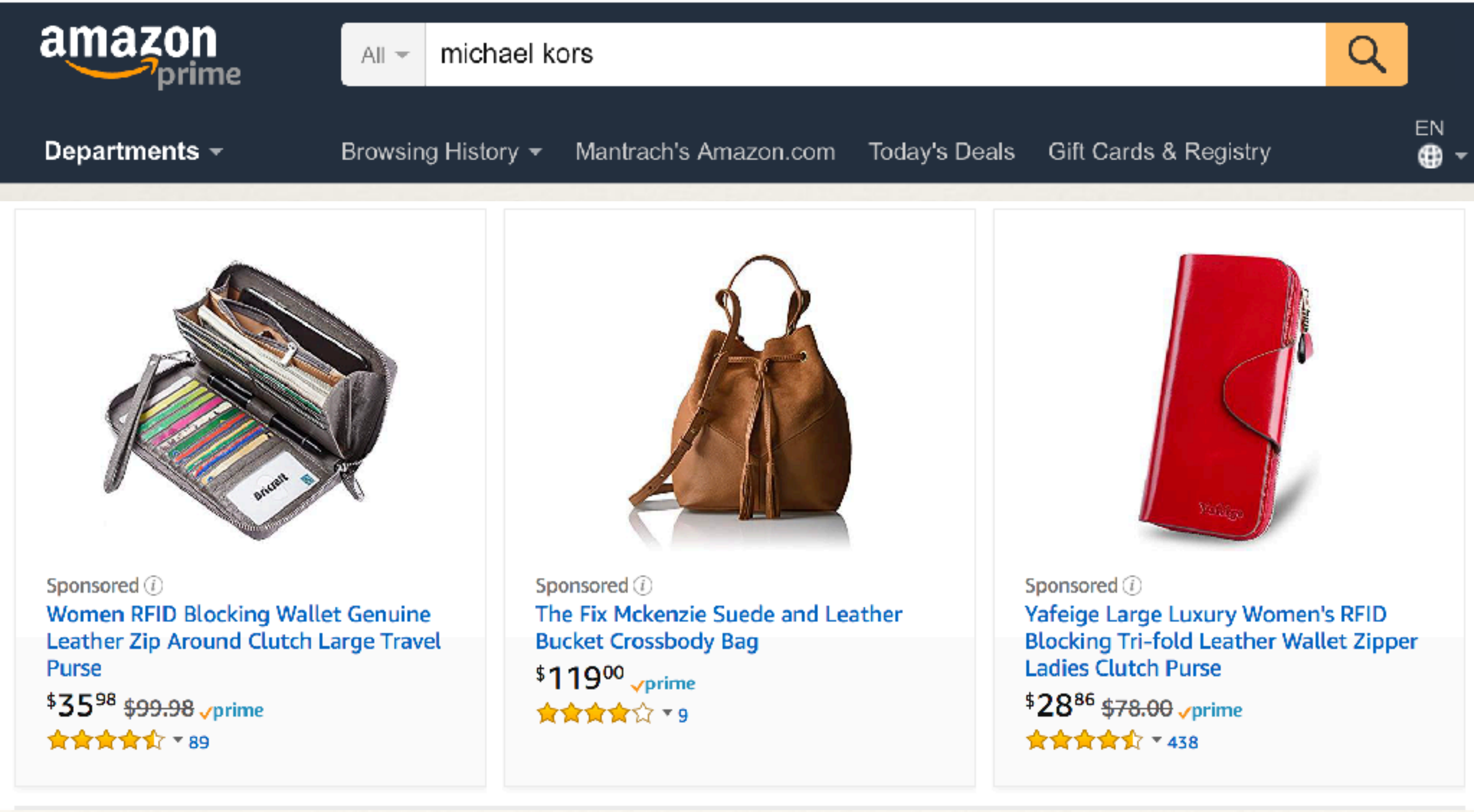
The Westin St. Francis - Summer of Love & Festivals
westinstfrancis.com/SummerOfLove (312) 957-6620
Save 15% on your **San Francisco** Stay this Summer & Explore the City by the Bay

300 San Francisco Hotels - Save 25% or More with KAYAK®
www.kayak.com/San-Francisco/Hotels
4.0 ★★★★☆ rating for kayak.com
Compare 100s of Travel Sites for Best Rates. Search for **Hotels** in **San Francisco**.

$19 Hotel Near San Francisco Ca - New Hotel Deals Added
goSeek.com/San-Francisco-Hotels
New **Hotel** Deals Added! Book Today. 70% Off On **San Francisco Hotels**

Hotels In San Francisco -

Sponsored Search: On Site Search



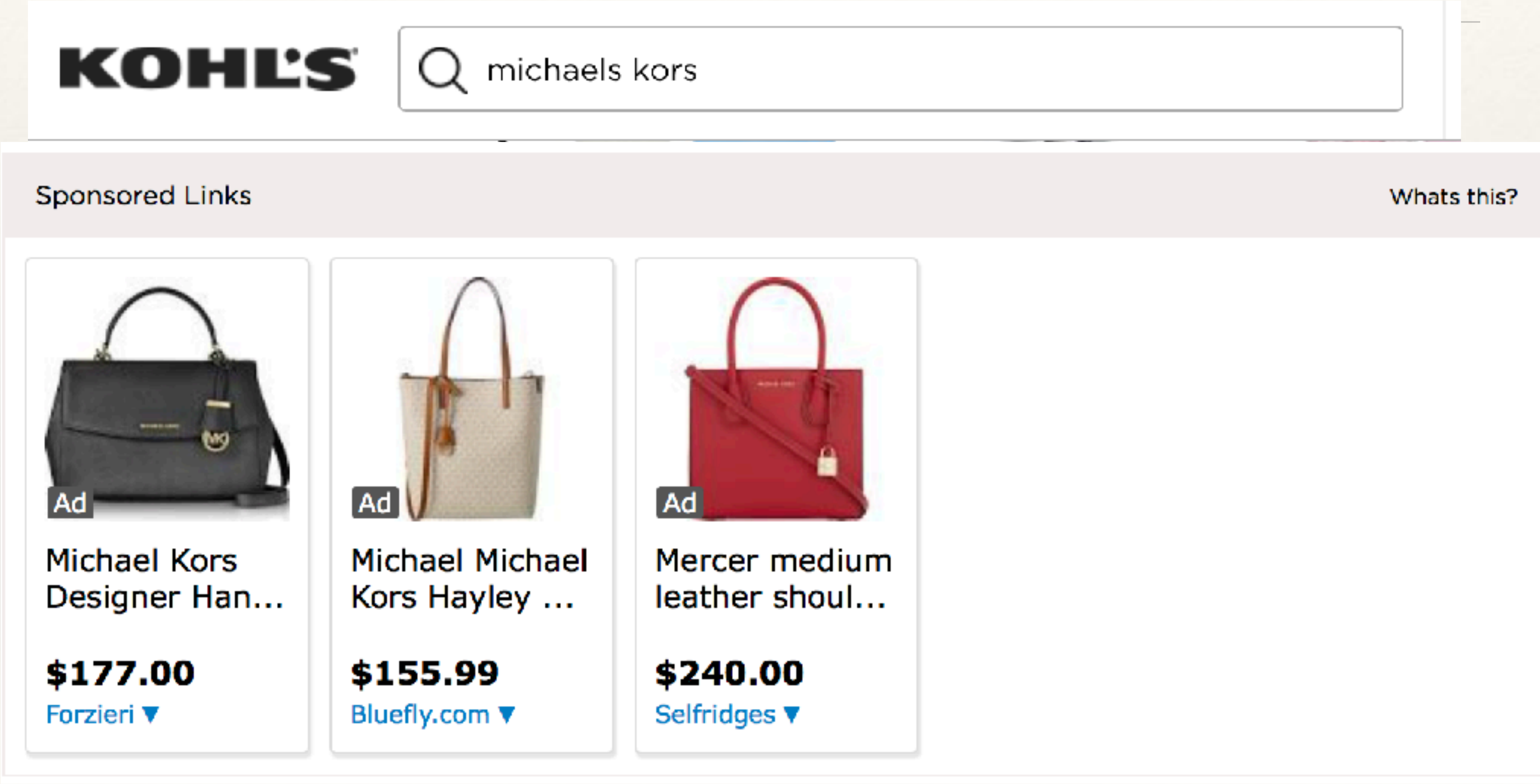
michael kors

Sponsored ⓘ
Women RFID Blocking Wallet Genuine Leather Zip Around Clutch Large Travel Purse
$35.98 ~~$99.98~~ ✓prime
★★★★☆ 89

Sponsored ⓘ
The Fix Mckenzie Suede and Leather Bucket Crossbody Bag
$119.00 ✓prime
★★★★☆ 9

Sponsored ⓘ
Yafeige Large Luxury Women's RFID Blocking Tri-fold Leather Wallet Zipper Ladies Clutch Purse
$28.86 ~~$78.00~~ ✓prime
★★★★☆ 438

Sponsored Search: Criteo Brand Solutions



KOHL'S

michaels kors

Sponsored Links

Whats this?

Ad

Michael Kors Designer Han...

$177.00

Forzieri ▼

Ad

Michael Michael Kors Hayley ...

$155.99

Bluefly.com ▼

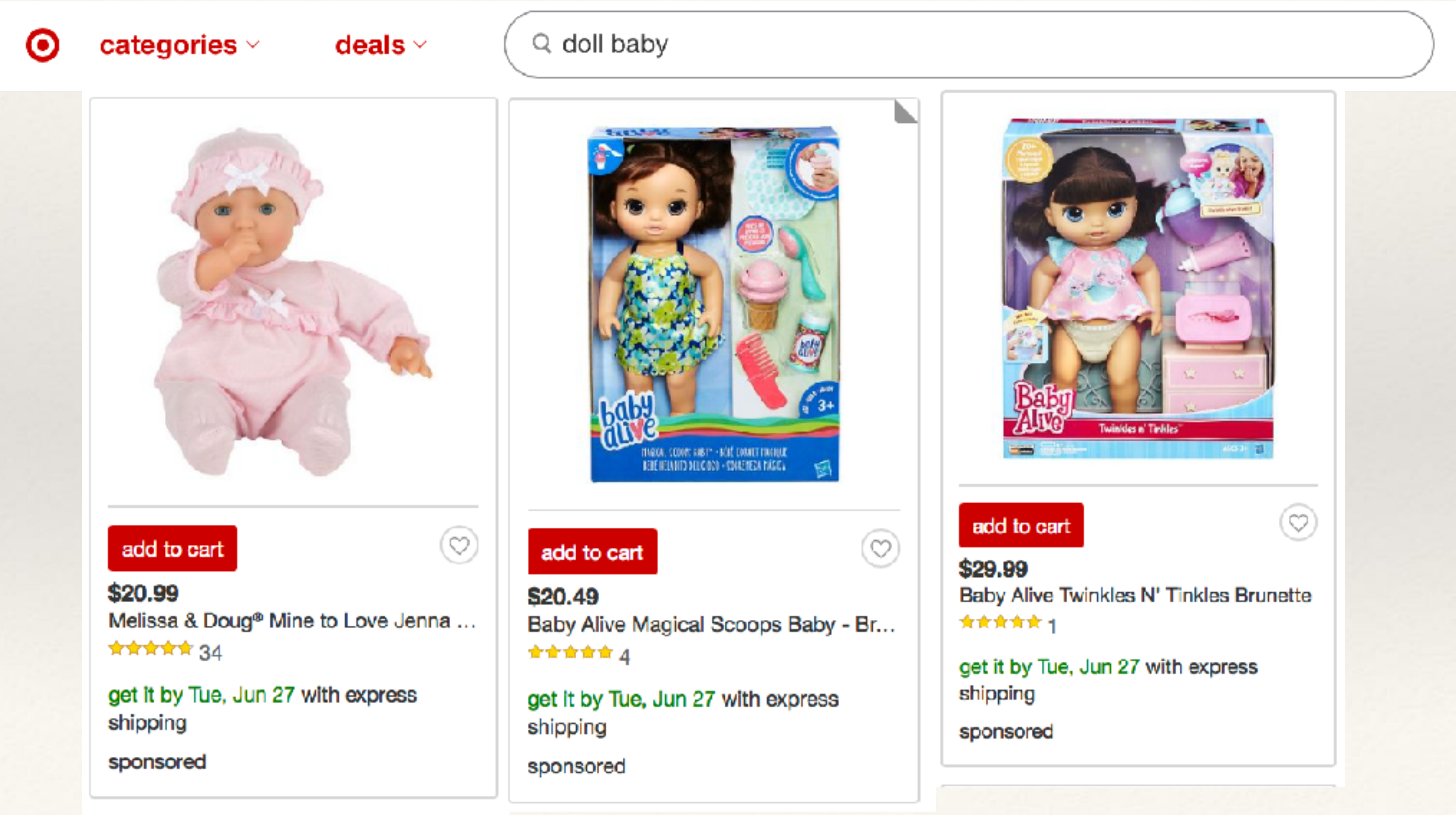
Ad

Mercer medium leather shoul...

$240.00

Selfridges ▼

Sponsored Search: Criteo Brand Solutions



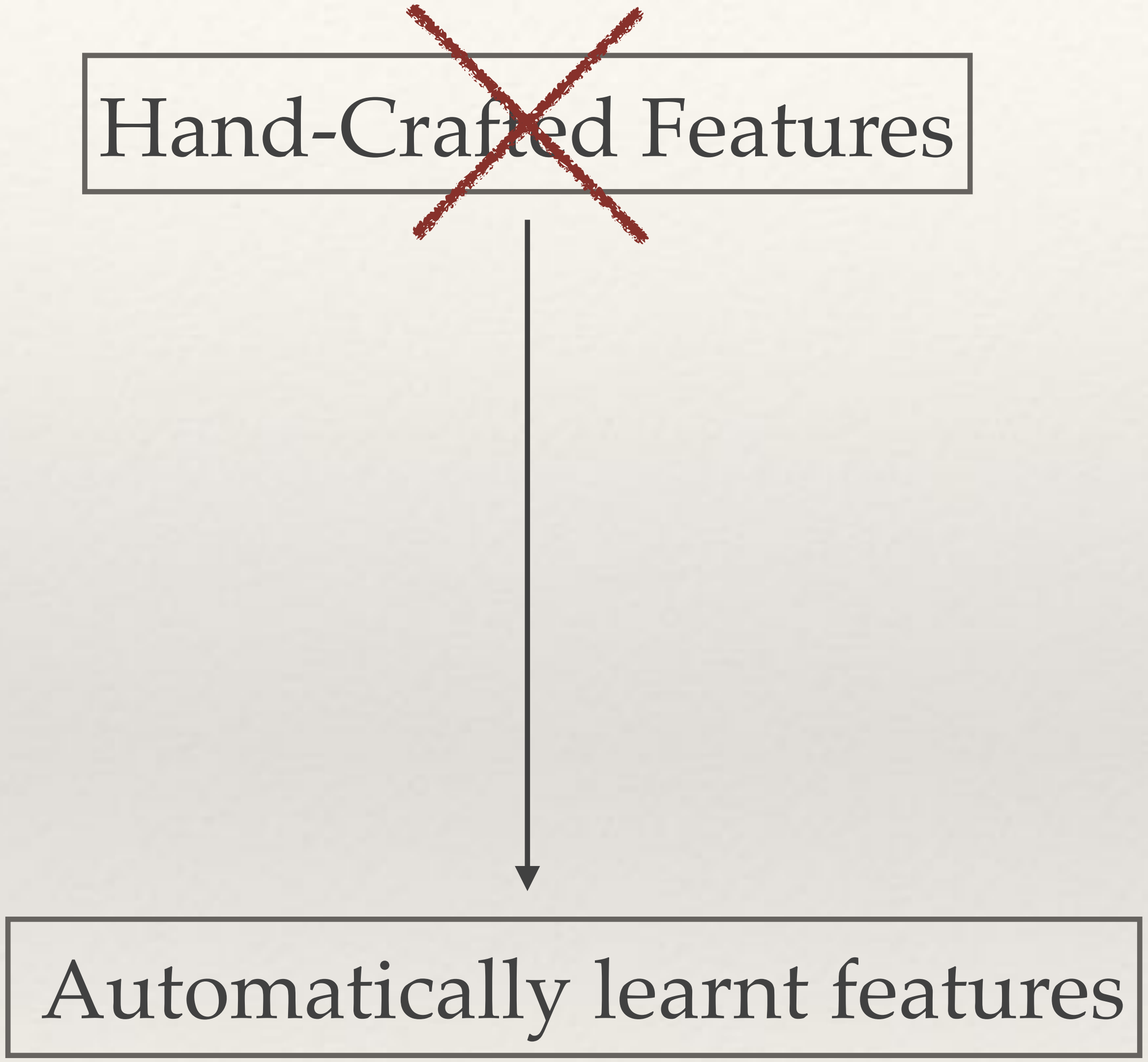
$20.99
Melissa & Doug® Mine to Love Jenna ...
★★★★★ 34
get it by Tue, Jun 27 with express shipping
sponsored

$20.49
Baby Alive Magical Scoops Baby - Br...
★★★★★ 4
get it by Tue, Jun 27 with express shipping
sponsored

$29.99
Baby Alive Twinkles N' Tinkles Brunette
★★★★★ 1
get it by Tue, Jun 27 with express shipping
sponsored

For CTR prediction

CONTENT



~~Hand-Crafted Features~~

Automatically learnt features

Problem

For a given query-ad pair, what is the probability of a click?

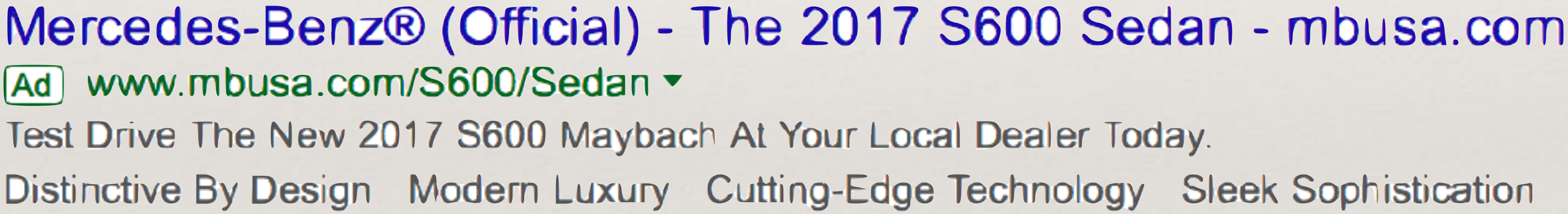
$$\mathbb{P}[click|query, ad]$$

ex: what is the probability of click for

query="buy car" - ad= "Toyota"

Problem

- We consider the case of text Ads but the work can easily be applied to product Ads.



Mercedes-Benz® (Official) - The 2017 S600 Sedan - mbusa.com

Ad www.mbusa.com/S600/Sedan ▾

Test Drive The New 2017 S600 Maybach At Your Local Dealer Today.

Distinctive By Design Modern Luxury Cutting-Edge Technology Sleek Sophistication

Related Work

- Established hand-crafted features for Sponsored Search
- Deep Similarity Learning
- Deep Character-level Models

Related Work

- Hand-crafted features for sponsored Search [6]

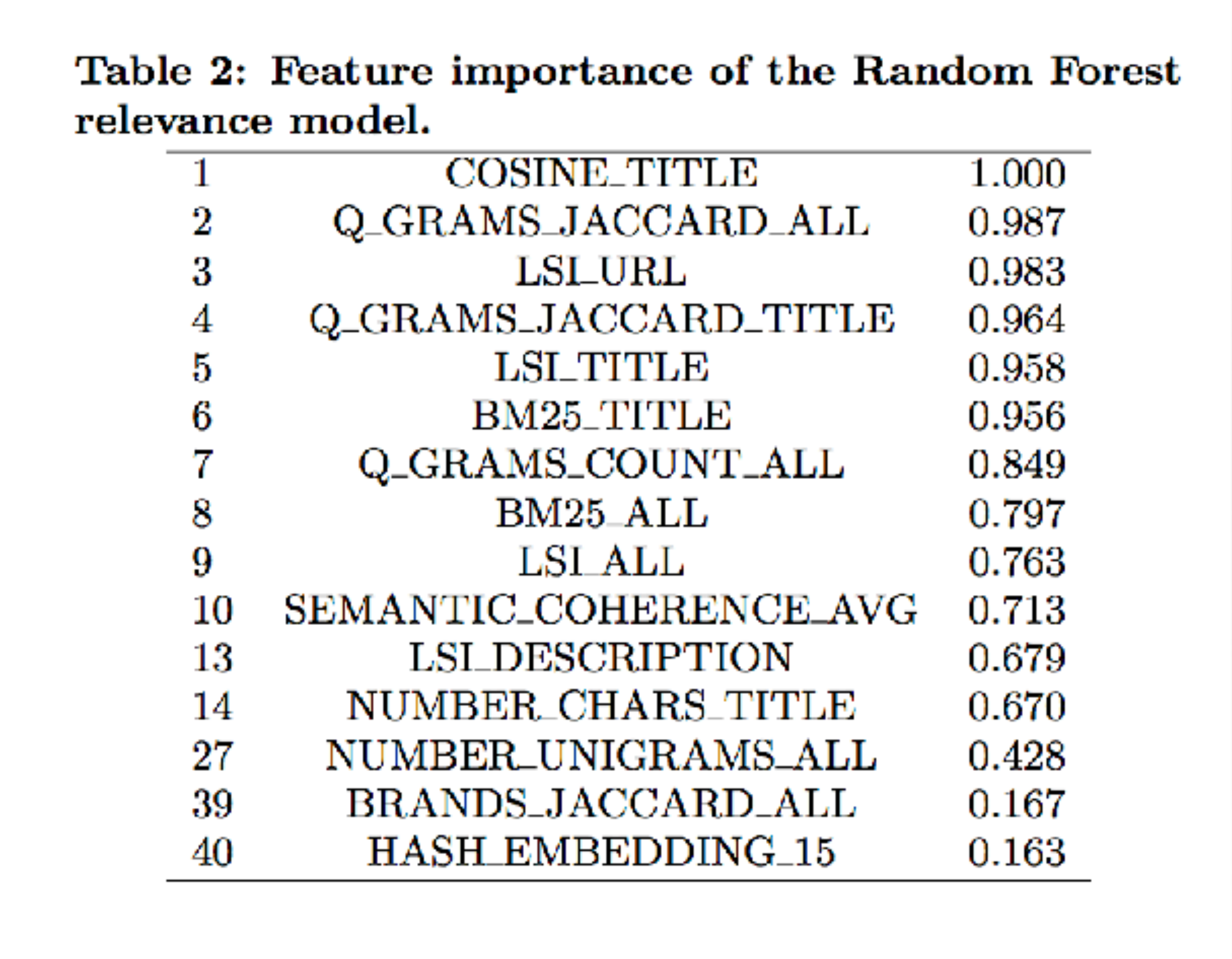


Table 2: Feature importance of the Random Forest relevance model.

1	COSINE_TITLE	1.000
2	Q_GRAMS_JACCARD_ALL	0.987
3	LSI_URL	0.983
4	Q_GRAMS_JACCARD_TITLE	0.964
5	LSI_TITLE	0.958
6	BM25_TITLE	0.956
7	Q_GRAMS_COUNT_ALL	0.849
8	BM25_ALL	0.797
9	LSI_ALL	0.763
10	SEMANTIC_COHERENCE_AVG	0.713
13	LSI_DESCRIPTION	0.679
14	NUMBER_CHARS_TITLE	0.670
27	NUMBER_UNIGRAMS_ALL	0.428
39	BRANDS_JACCARD_ALL	0.167
40	HASH_EMBEDDING_15	0.163

Related Work

- Deep Similarity Learning
  - Deep Intent: Zhai et al.[2] aimed to solve query-ad relevance problem. Query and Ad vectors are learnt using LSTMs. Inputs of LSTMs are pre-trained word vectors. Cosine similarity between ad and vectors represent the similarity score between query and ad couple.
  - Search2Vec: Grbovic et al.[1] proposed a method that learn a vector for each query and each ad. Score for the query-ad pair is obtained through cosine similarity.
  - Drawbacks of X2Vec approaches:
    - Coverage: Misspelling, Cold Cases
    - Dictionary: Storage, Update
    - Weakly supervised

Related Work

- Deep Similarity Learning
  - Hu et al. [3] also propose to directly capture the similarity between two sentences without explicitly relying on semantic vector representations. This model works at the word level, but is targeting matching task as: sentence completion, matching a response to a tweet, and paraphrase identification.

Related Work

- Deep Character Models
    - *"We believe this is a first evidence that a learning machine does not require knowledge about words, phrases, sentences, paragraphs or any other syntactical or semantic structures to understand text. That being said, we want to point out that ConvNets by their design have the capacity to learn such structured knowledge."* Zhang et al. [4]

Motivation

- Recent progress at Character-level Language Models
- Drawbacks of existing approaches
- Idea: **Leverage Character-level approaches and click data to learn the query-ad language from scratch**

Contributions

1. We are first to learn the textual similarity between two pieces of text (i.e., query and ad) from scratch, i.e., at the character level.
2. We are first to learn to directly predict the click-through rate in the context of sponsored search without any feature engineering.

Research Questions

1. Can we automatically learn representations for query-ad pairs without any feature engineering in order to predict the CTR in sponsored search?

2. How does the performance of a character-level deep learning model differ from a word-level model for CTR prediction?

3. How do the introduced character-level and word-level deep learning models compare to baseline models (Search2Vec, and hand-crafted features with logistic)?

4. Can the proposed models improve the CTR prediction model running in the production system of a popular commercial search engine?

Deep CTR Modeling

- Loss Function
- Key Components of Proposed Models
- DeepCharMatch
- DeepWordMatch

Deep CTR Modeling

Loss Function

$$L = \sum_{q\_a : c_{q\_a}=1} \log p_{q\_a} + \sum_{q\_a : c_{q\_a}=0} \log(1 - p_{q\_a})$$

p_{q_a} prediction of the model for query q and ad a

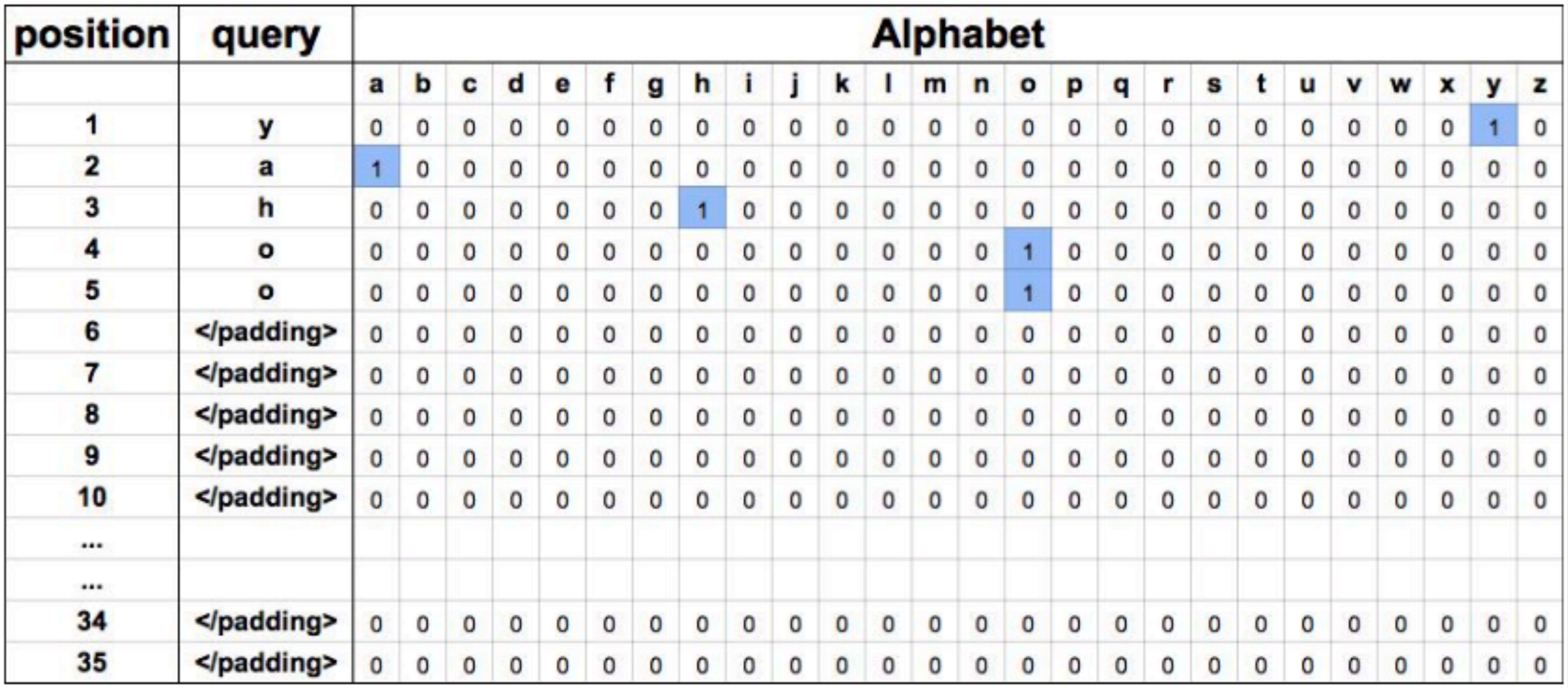
c_{q_a} ground truth click query q and ad a

Input Representation

- Queries are normalized. For ads, normalized title, description and url.
- Both query and ad is zero padded text with fixed length, where
  - Fixed query length, l_q =35
  - Fixed ad length, l_a =140
- Both query and ad are vectorized considering a constant vocabulary size $|V| = 77$
  - Dimension of query: $l_q \times |V|$
  - Dimension of ad: $l_a \times |V| = 140 \times 77$

Input Representation

Input Representation



position	query	Alphabet																									
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
1	y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2	a	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	h	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...																											
...																											
34	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	</padding>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Deep CTR Modeling

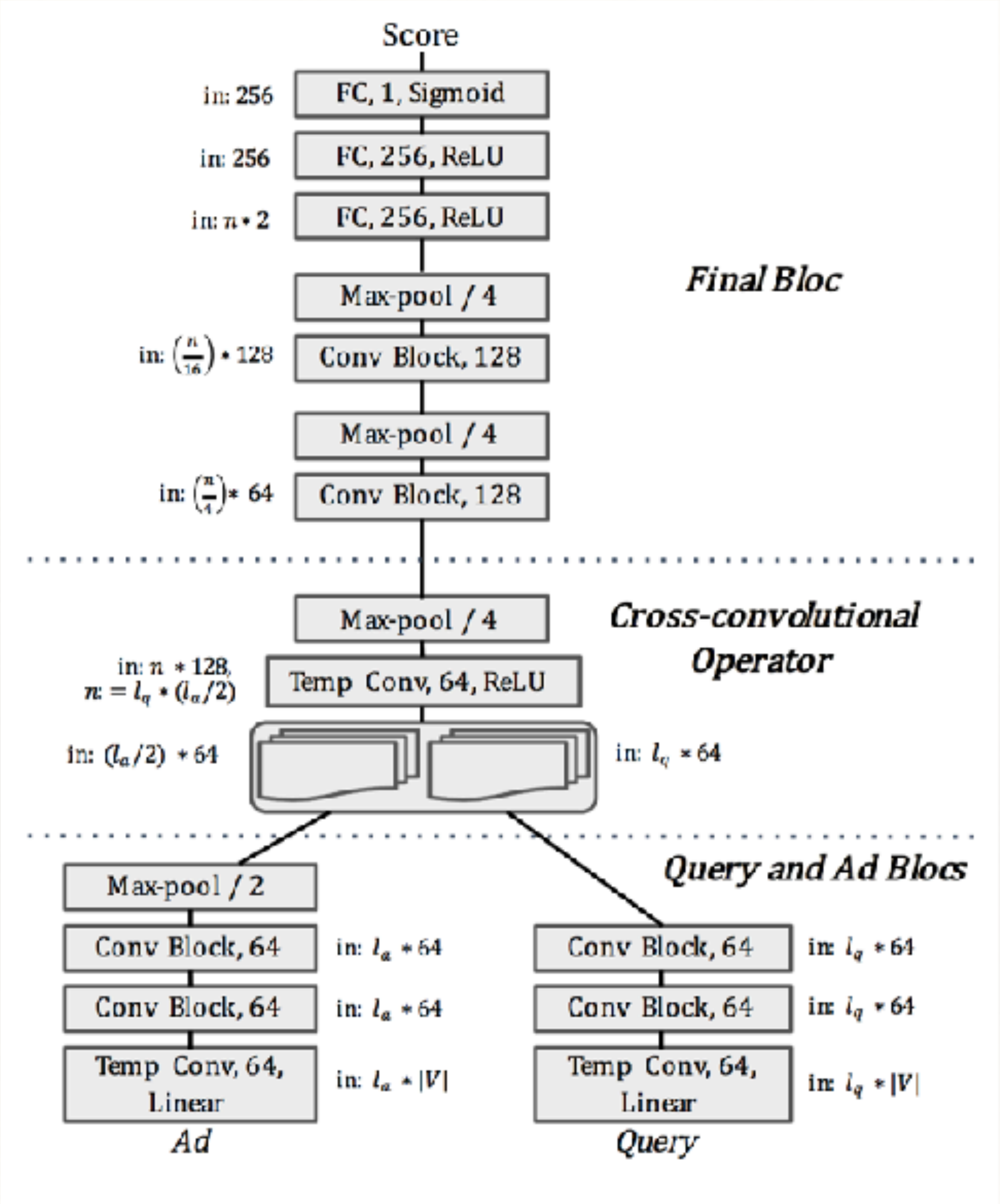
Key Components of Proposed Models

- Temporal Convolution
- Temporal Max-Pooling
- Fully Connected Layer

Deep CTR Modeling

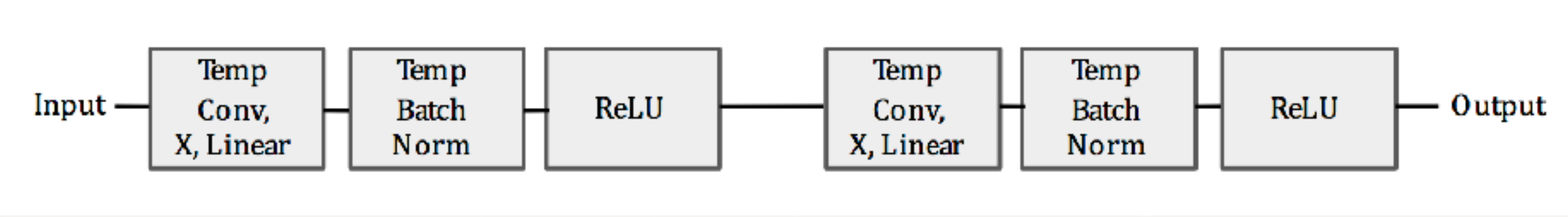
- DeepCharMatch
- DeepWordMatch

DeepCharMatch

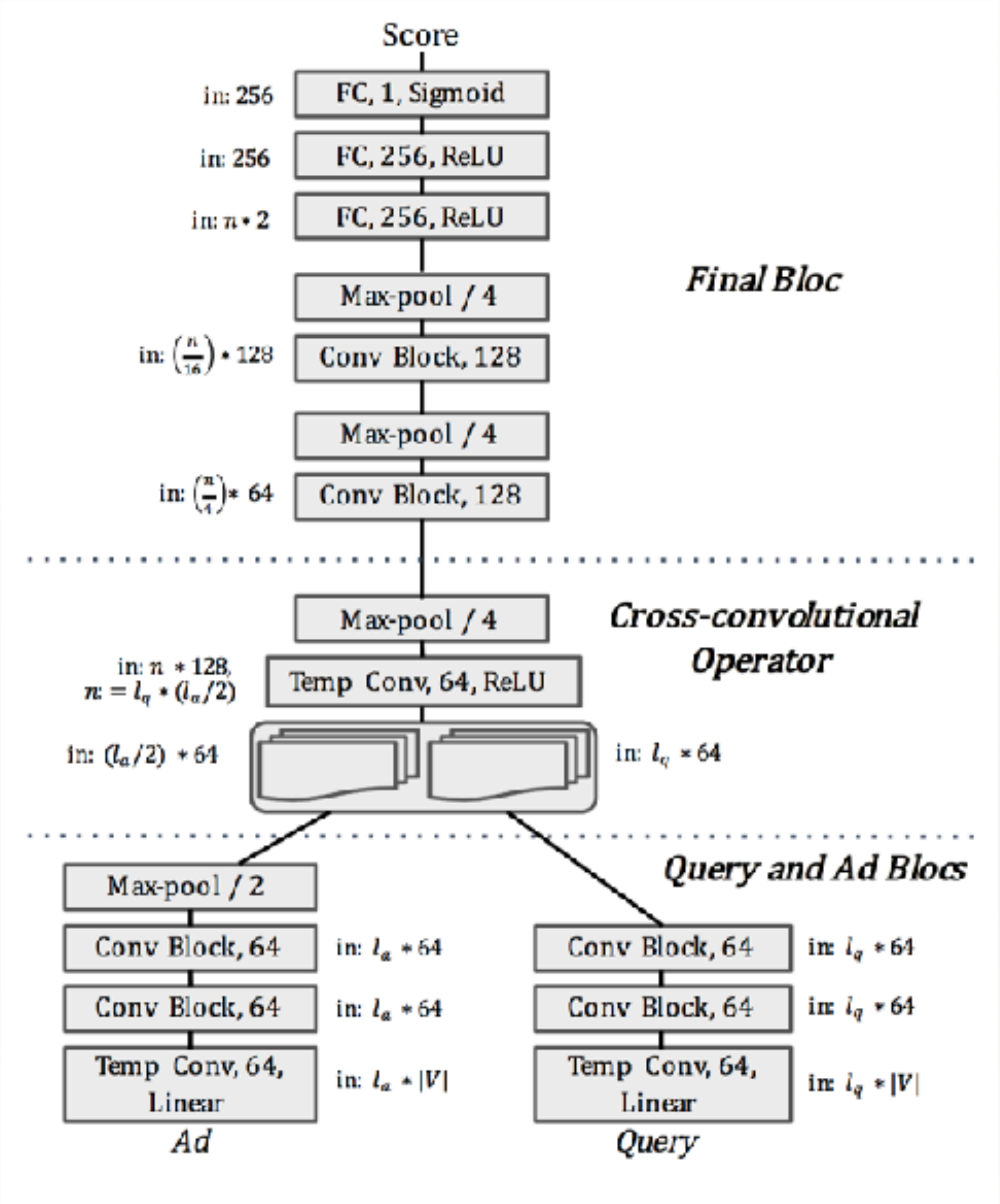


Query ad Ad Blocs aim to produce higher level representations for query and ad.

Convolutional Block

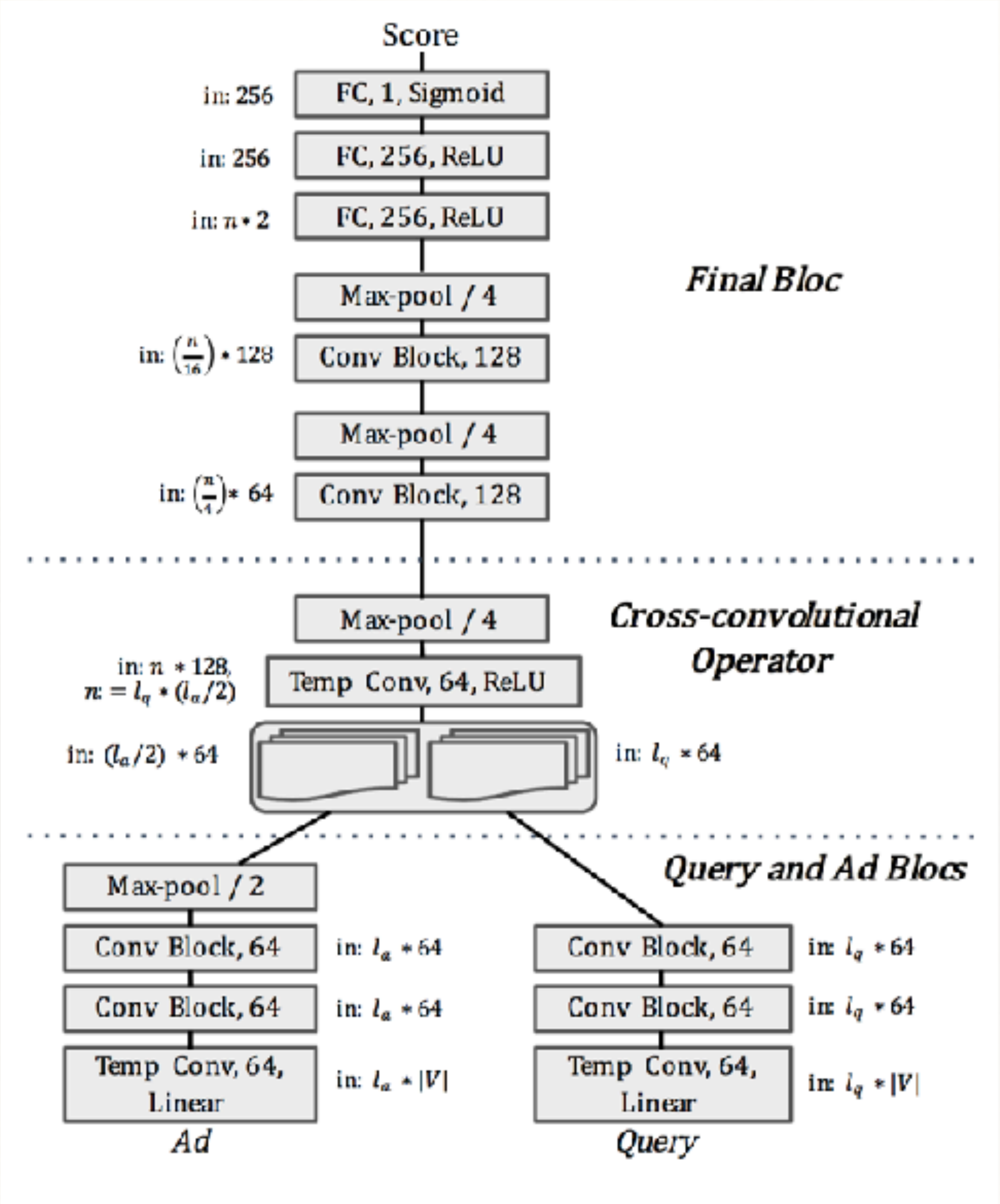


DeepCharMatch



Cross-convolution Operator aims to capture possible intra-word and intra-sentence relationships between query and ad.

DeepCharMatch



Final Bloc models the relationship between the query and the ad. Outputs the final prediction for CTR of query and ad pair.

DeepWordMatch

Input Representation

- Queries are normalized. For ads, normalized title, description and url.
- Both query and ad is zero padded text with fixed length, where
    - Fixed query length, $d_q = 7$
    - Fixed ad length, $d_a = 40$
- Both query and ad are vectorized considering a constant vocabulary size obtained by GloVe [6] where dimensions of the vectors $d_w = 50$.
    - Dimension of query: d_q x d_w
    - Dimension of ad: d_a x d_w

DeepWordMatch

Model Architecture

- Consists of a cross-convolution operator ended by a final bloc capturing the commonalities between the query and the ad.
- Ad and query matrixes consist of pre-trained word vectors directly feed into cross-convolution operator.
- Except those points, the architecture of DeepWordMatch is equivalent to the architecture of DeepChar- Match.

Experiments

- Experimental Setup
  - Dataset
  - Baselines
  - Evaluation Metrics
  - Experimental Platform
- Experimental Results

Experiments

Experimental Setup - Dataset

- We randomly sample 1.5 Billion query-ad pairs served by a popular commercial search engine. Dates: August 6 to September 5, 2016.
- We only consider the sponsored ads that are shown in the north of the search result pages.
- We randomly sample the test set that consists of about 27 millions query-ad pairs without any page position restriction. Dates: September 6 to September 20, 2016.

Experiments

Experimental Setup - Dataset Characteristics

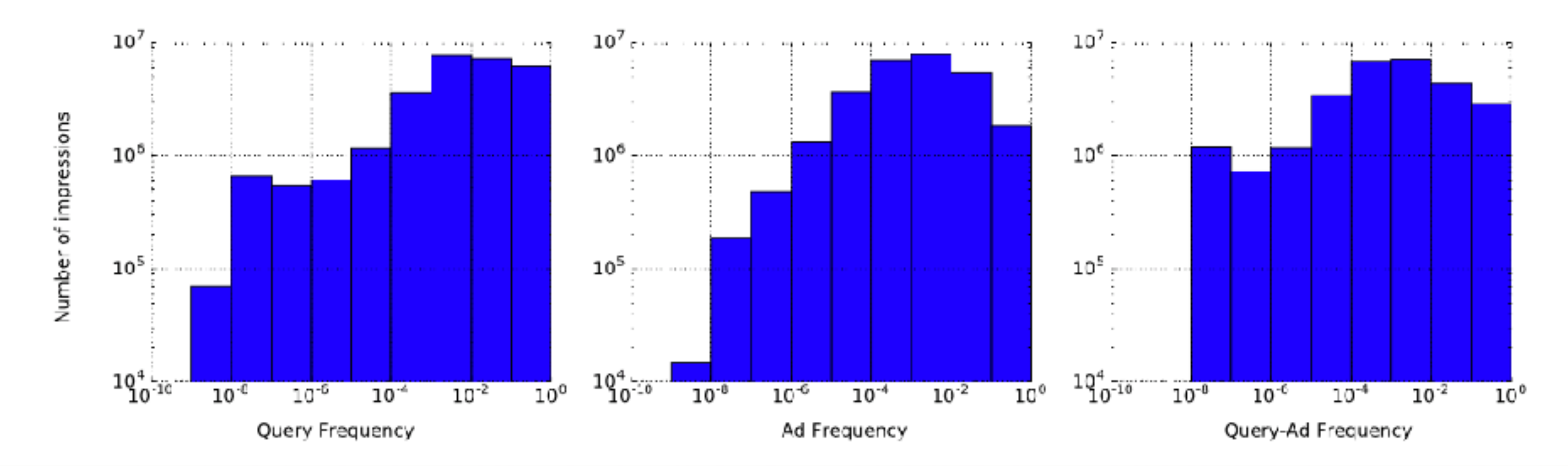


Figure 1: Distribution of impressions in the test set with respect to query, ad, and query-ad frequencies computed on six months (The frequencies are normalized by the maximum value in each subplot).

Experiments

Experimental Setup - Baselines

- **Feature-engineered logistic regression (FELR).** We use the 185 state-of-the-art features designed to capture the pairwise relationship between a query and the three different components in a textual ad, i.e., its title, description, and display URL. These features are explained in details in [6] and are achieving state-of-the-art results in relevance prediction for sponsored search. Model also optimizes cross-entropy loss function.
- **Search2Vec.** It learns semantic embeddings for queries and ads from search sessions, and uses the cosine similarity between the learnt vectors to measure the textual similarity between a query and an ad. This approach leads to high-quality query-ad matching in sponsored search. It is not trained to predict CTR therefore this approach can be considered as weakly-supervised.

Experiments

Experimental Setup - Baselines

- **Production Model:** CTR prediction model in the production system of a popular commercial search engine. Model is a machine learning model trained with a rich set of features, including click features, query features, ad features, query-ad pair features, vertical features, contextual features such as geolocation or time of the day, and user features. Model also optimizes cross-entropy loss function.
    - Our aim is to observe possible contribution of DeepCharMatch and DeepWordMatch. To observe, we basically averaged the prediction of Production Model with DeepCharMatch and DeepWordMatch. They are represented as
        - DCP := (PredDeepCharMatch+PredProductionModel) / 2
        - DWP := (PredDeepWordMatch+PredProductionModel) / 2

Experiments

Experimental Setup - Evaluation Metrics

- **Area under the ROC curve: AUC:** It measures whether the clicked ad impressions are ranked higher than the non-clicked ones. e perfect ranking has an AUC of 1.0, while the average AUC for random rankings is 0.5.

Experiments

Experimental Setup - Experimental Platform

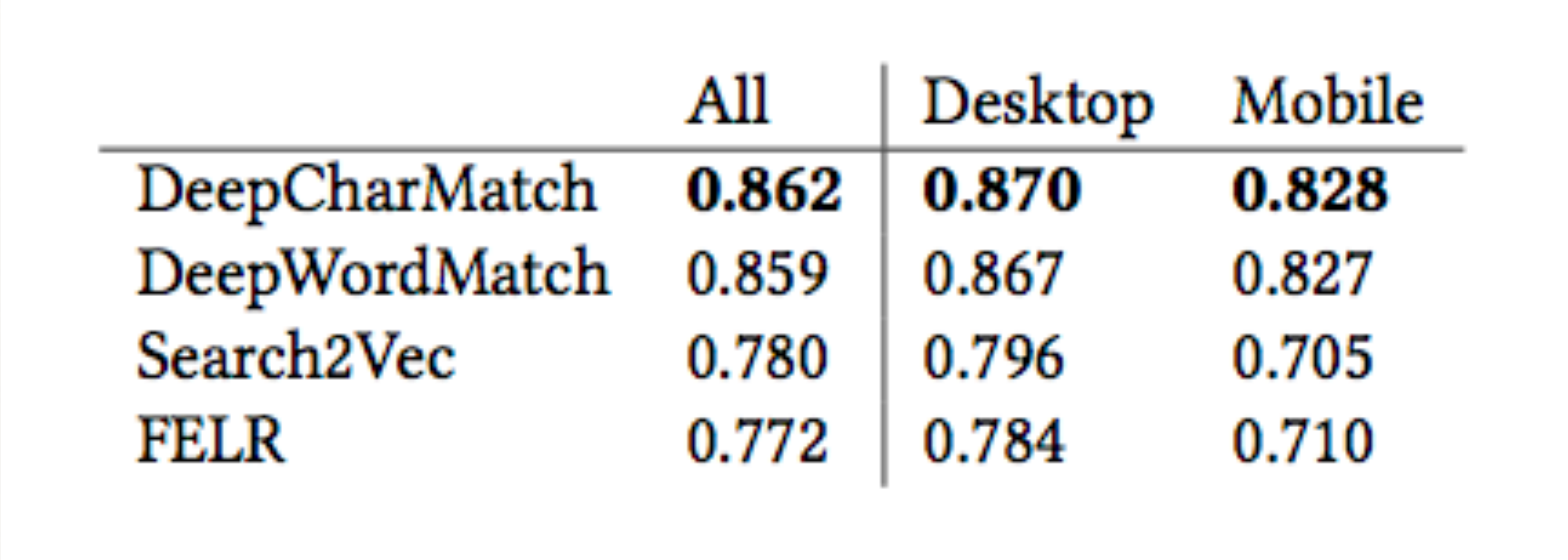
- Tensorflow Distributed on Spark
- Async training on multiple GPUs
- Optimizer: Adam Optimizer
- Minibatch size = 64

Experiments

- Experimental Results - Research Questions
- Can we automatically learn representations for query-ad pairs without any feature engineering in order to predict the CTR in sponsored search?
- How does the performance of the character-level deep learning model differ from the word-level model for CTR prediction?
- How do the introduced character-level and word-level deep learning models compare to the baseline models?

Experiments

Experimental Results - Research Question {1,2,3}



	All	Desktop	Mobile
DeepCharMatch	**0.862**	**0.870**	**0.828**
DeepWordMatch	0.859	0.867	0.827
Search2Vec	0.780	0.796	0.705
FELR	0.772	0.784	0.710

Table 1: AUC of DeepCharMatch, DeepWordMatch, Search2Vec and FELR.

Experiments

Experimental Results - Research Question {1,2,3}

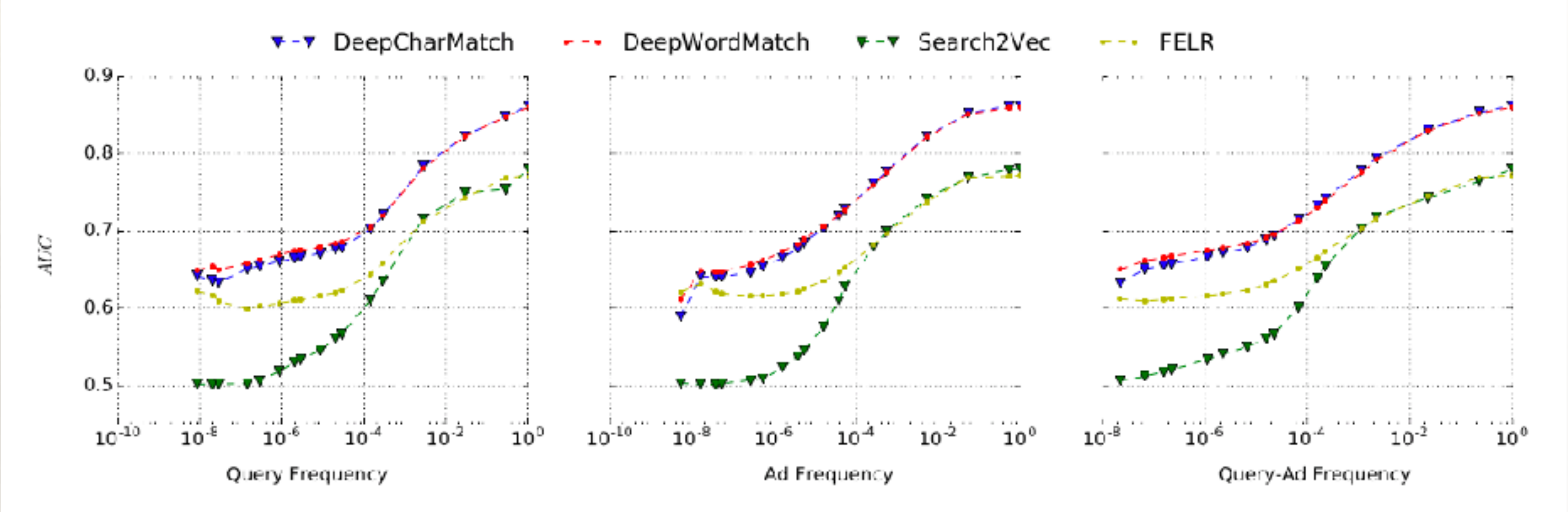
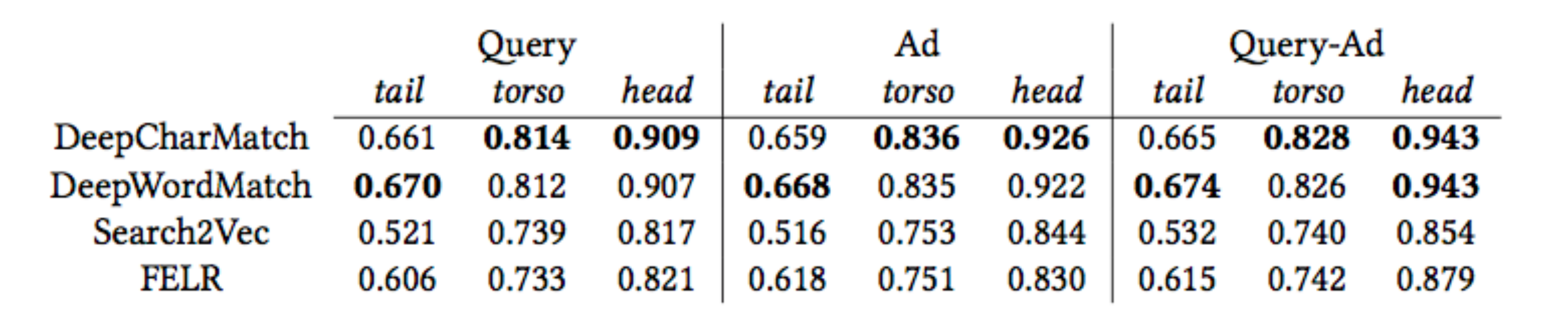


Figure 2: Cumulative AUC by query, ad, and query-ad frequency for DeepCharMatch, DeepWordMatch, Search2Vec and FELR. Frequencies are normalized by the maximum value in each subplot. For each bin, the number of impressions used to compute AUC is reported in Figure 1. Cumulative means that at x the plot reports AUC of points whose frequency is lower than x.

Experiments

Experimental Results - Research Question {1,2,3}



	Query			Ad			Query-Ad		
	tail	*torso*	*head*	*tail*	*torso*	*head*	*tail*	*torso*	*head*
DeepCharMatch	0.661	**0.814**	**0.909**	0.659	**0.836**	**0.926**	0.665	**0.828**	**0.943**
DeepWordMatch	**0.670**	0.812	0.907	**0.668**	0.835	0.922	**0.674**	0.826	**0.943**
Search2Vec	0.521	0.739	0.817	0.516	0.753	0.844	0.532	0.740	0.854
FELR	0.606	0.733	0.821	0.618	0.751	0.830	0.615	0.742	0.879

Table 2: AUC of DeepCharMatch, DeepWordMatch, Search2Vec and FELR, on tail, torso, and head of the query, ad, and query- ad frequency distributions. Tail stands for normalized frequency nf $< 10^{-6}$, torso for $10^{-6} <$ nf $< 10^{-2}$, and head for nf $> 10^{-2}$.

Experiments

Experimental Results - Research Question {1,2,3}

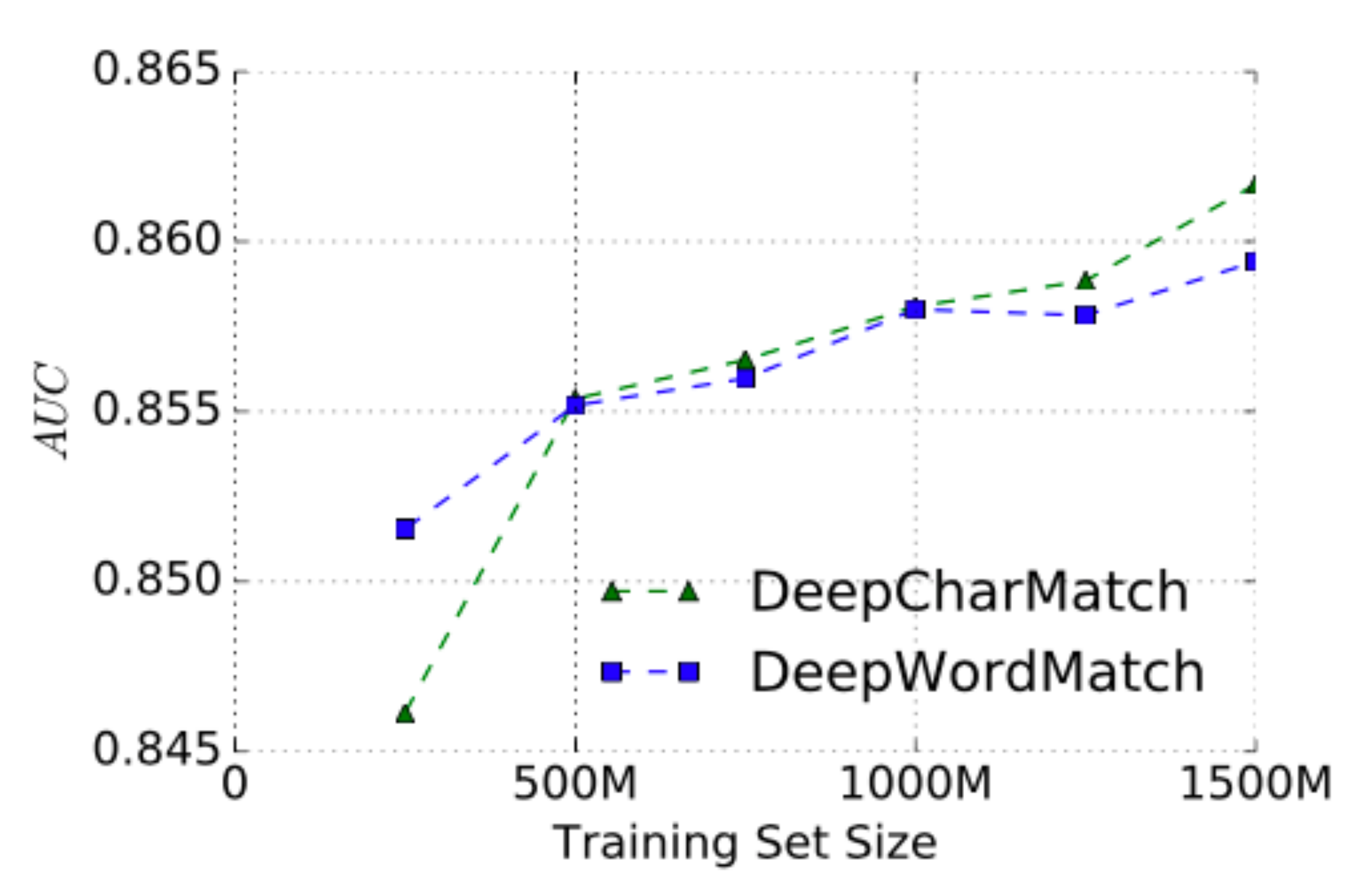


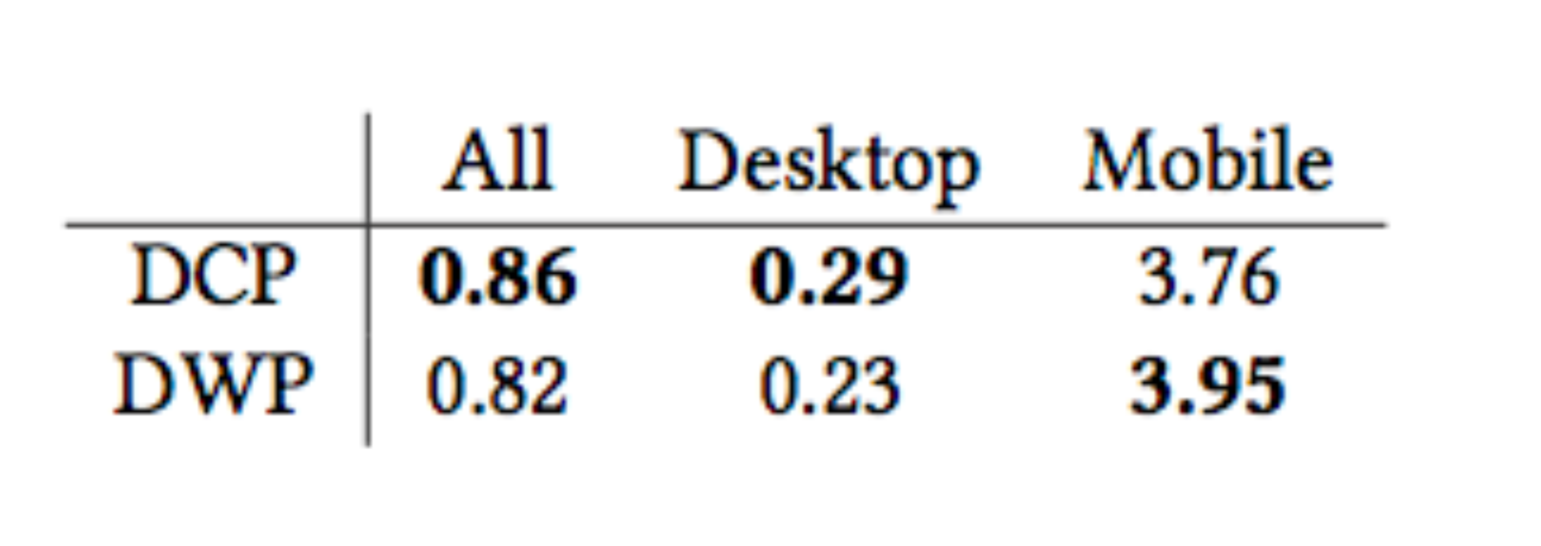
Figure 3: AUC of DeepCharMatch and DeepWordMatch by number of training points.

Experiments

- Experimental Results - Research Questions
- Can the proposed models improve the CTR prediction model running in the production system of a popular commercial search engine?

Experiments

Experimental Results - Research Question 4



	All	Desktop	Mobile
DCP	**0.86**	**0.29**	3.76
DWP	0.82	0.23	**3.95**

Table 2: Relative AUC Improvement in % of DCP over Production model.

Experiments

Experimental Results - Research Question 4

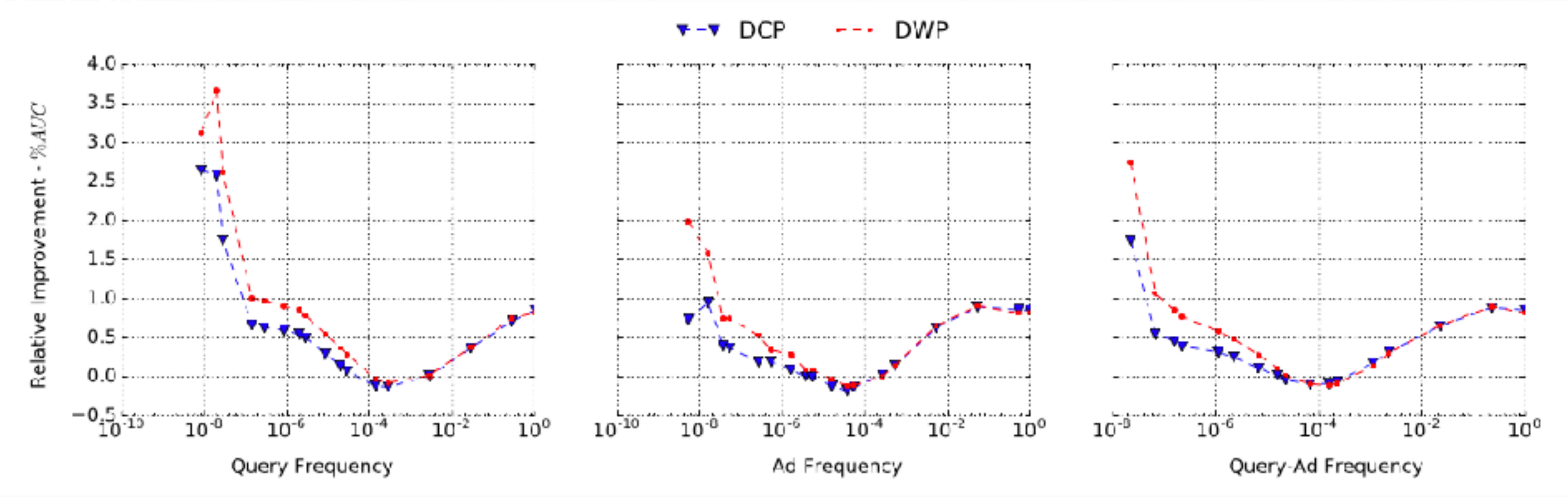
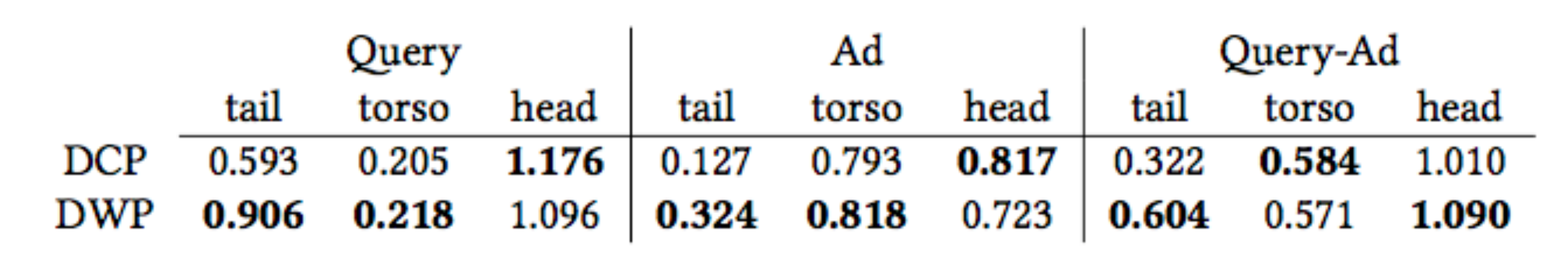


Figure 4: Cumulative relative improvements of DCP and DWP over Production model in terms of %AUC. Frequencies are normalized by the maximum value of each subplot. For each bin, the number of impressions used to compute AUC is reported in Figure 1. Cumulative means that at x the plot reports relative improvements of points whose frequency is lower than x.

Experiments

Experimental Results - Research Question 4



	Query			Ad			Query-Ad		
	tail	torso	head	tail	torso	head	tail	torso	head
DCP	0.593	0.205	**1.176**	0.127	0.793	**0.817**	0.322	**0.584**	1.010
DWP	**0.906**	**0.218**	1.096	**0.324**	**0.818**	0.723	**0.604**	0.571	**1.090**

Table 3: Relative AUC Improvements in % of DCP and DWP over Production , on tail, torso, and head of the query, ad, and query- ad frequency distributions. Tail stands for normalized frequency nf $< 10^{-6}$, torso for $10^{-6} <$ nf $< 10^{-2}$, and head for nf $> 10^{-2}$.

Questions

Thank you!

References

[1] Grbovic, Mihajlo, et al. "Scalable Semantic Matching of Queries to Ads in Sponsored Search Advertising." Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2016

[2] Zhai, Shuangfei, et al. "DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks." Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016

[3] Hu, Baotian, et al. "Convolutional neural network architectures for matching natural language sentences." Advances in Neural Information Processing Systems. 2014.

[4] Zhang et al. "Text understanding from scratch." arXiv preprint arXiv:1502.01710 (2015)

[5] Pennington et al. "Glove: Global Vectors for Word Representation." EMNLP (2014)

[6] Aiello, Luca, et al. "The Role of Relevance in Sponsored Search." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016.