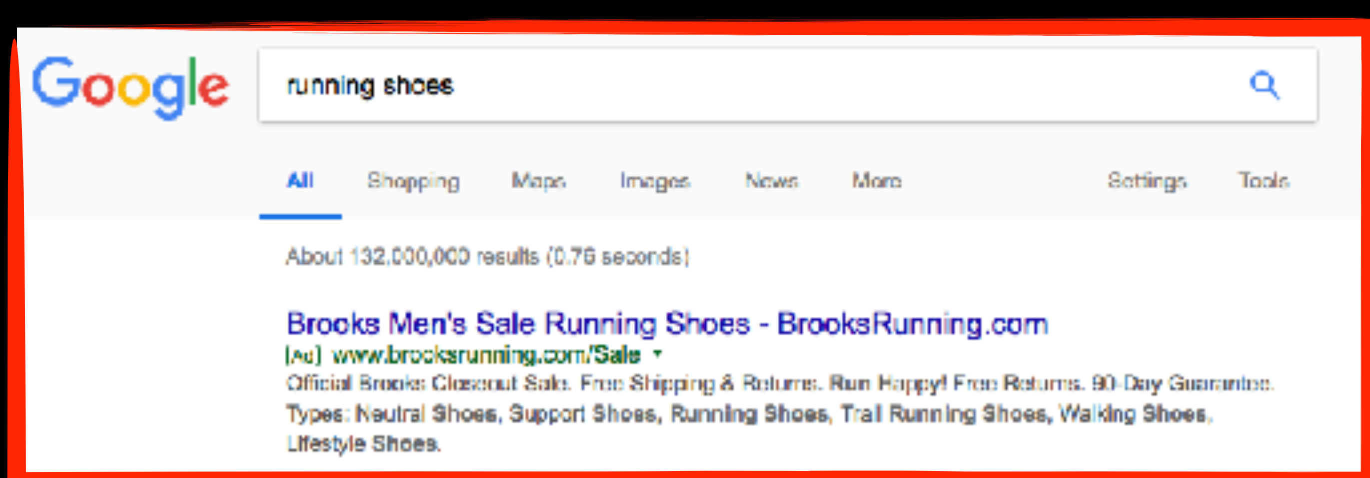
Textual Advertising



Hari Sundaram
Associate Professor (CS, ADV)
hs1@illinois.edu

thanks: Andrei Broder, Vanja Josifovski

Introduction

Web search

Game Theory

Auctions

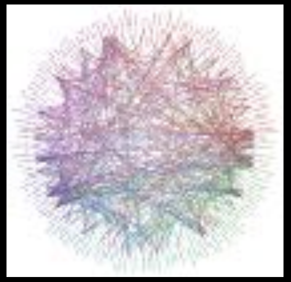
Text Ads

Display Ads

Behavioral targeting

Recommender systems

Privacy



Networks

Emerging areas

Final Presentations

"In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes.

What information consumes is rather obvious: **it consumes the attention of its recipients**. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herbert Simon

Advertisers are competing for attention!

Receptive

Interest level of the user in the advertiser's message and the willingness to absorb the message

For example: people interested in skiing ads are often interested— within a relatively short period of time—in biking ads,

Qualified

Selection of users by based on clear criteria
(e.g. people looking to buy a Car and who live in the US

What are advertisers looking for?

Responsive

Propensity of the user to respond in a desired way to the advertiser message, within a relatively short period of time (click to buy; get the person to the store; brand awareness etc.)

Advertising is a market where each side cares about the **type** of the other side

Advertisers want the attention of certain people

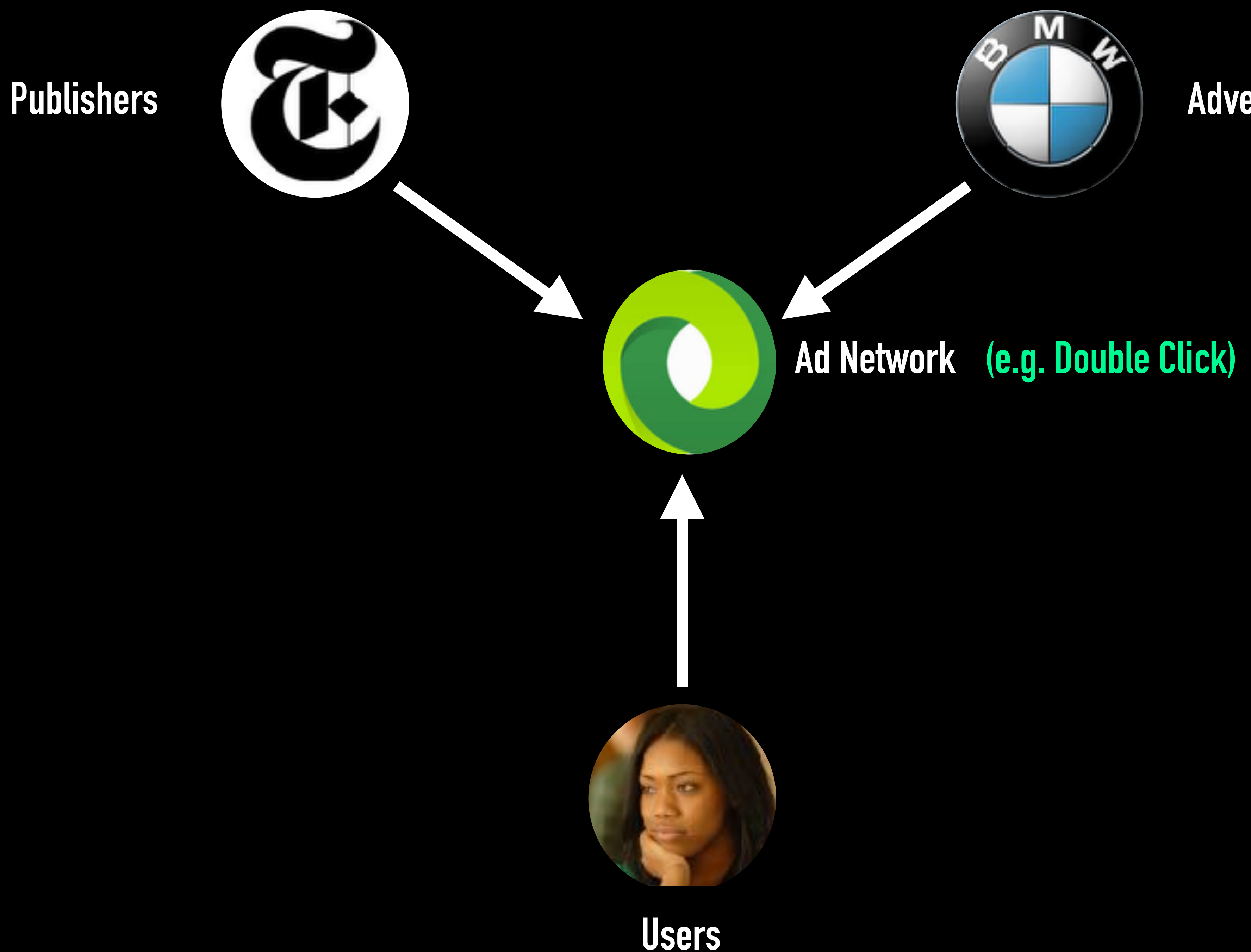
People are only open to certain ads (whether or not in the market for the advertised good)

A key challenge:

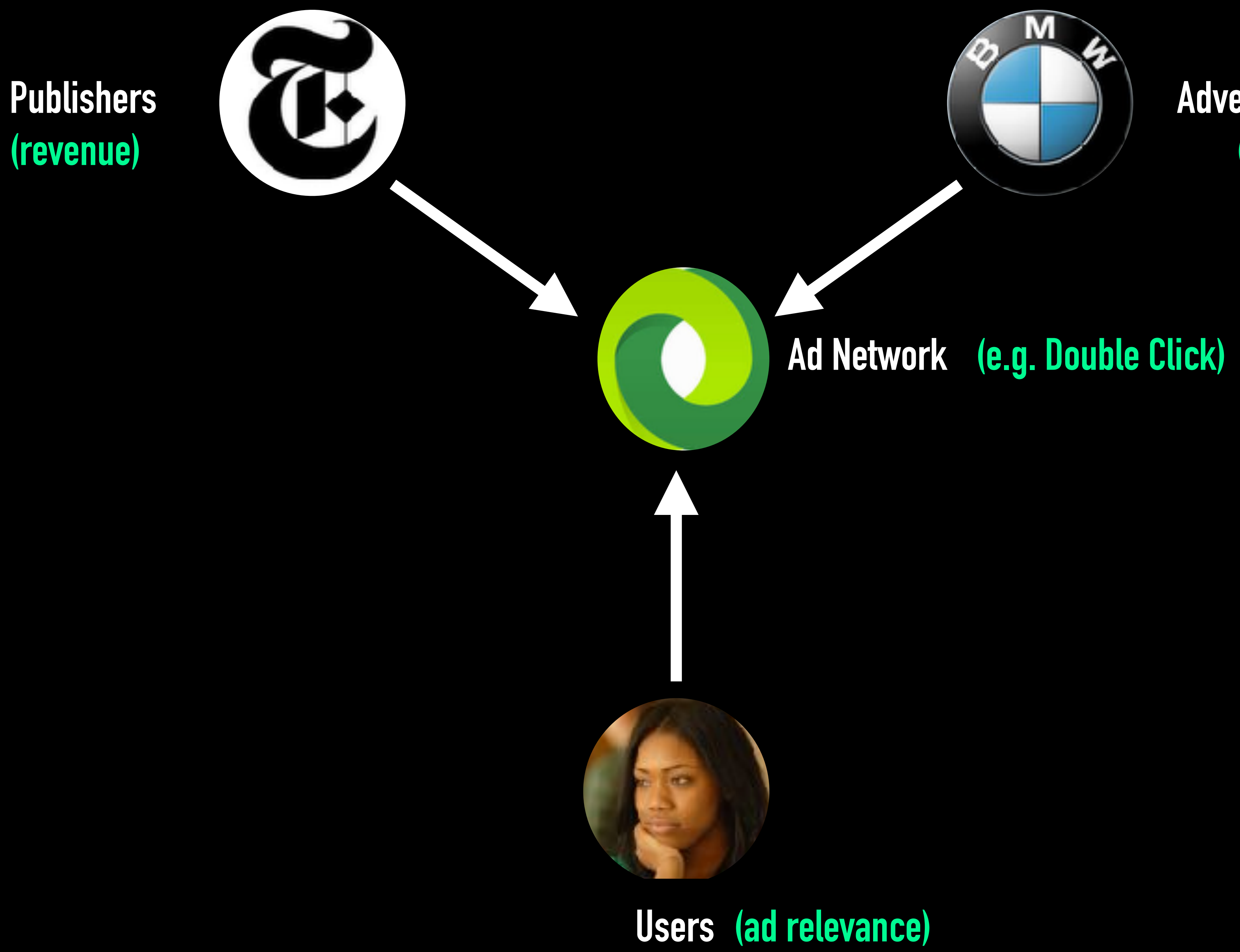
Find the "best match" between a given user in a given context and a suitable advertisement.

contexts: web search; publisher page (e.g. NY Times); mobile; billboard etc.

key actors



each actor has a different utility function



Sponsored Search

Publishers
(Web page)

The owner of the search engine results page (SERP)

Advertisers
(ROI)

Ad Network (i.e. Double Click)

Usually the publisher (the owner of the SERP) and the “matcher” are the same (Google, Bing) but can be different (MSFT provided both algorithmic search and paid search results to Yahoo)

We’ll assume the former for this lecture

Users (ad relevance)

Sponsored Search

Publishers
(Web page)

Advertisers
(ROI)

Submit ads associated to certain bid phrases; Bid for position; Pay (Cost Per Click) CPC

Executes query against web corpus + other data sources

Ad Network

(conducts auctions)

Executes query against the ad corpus

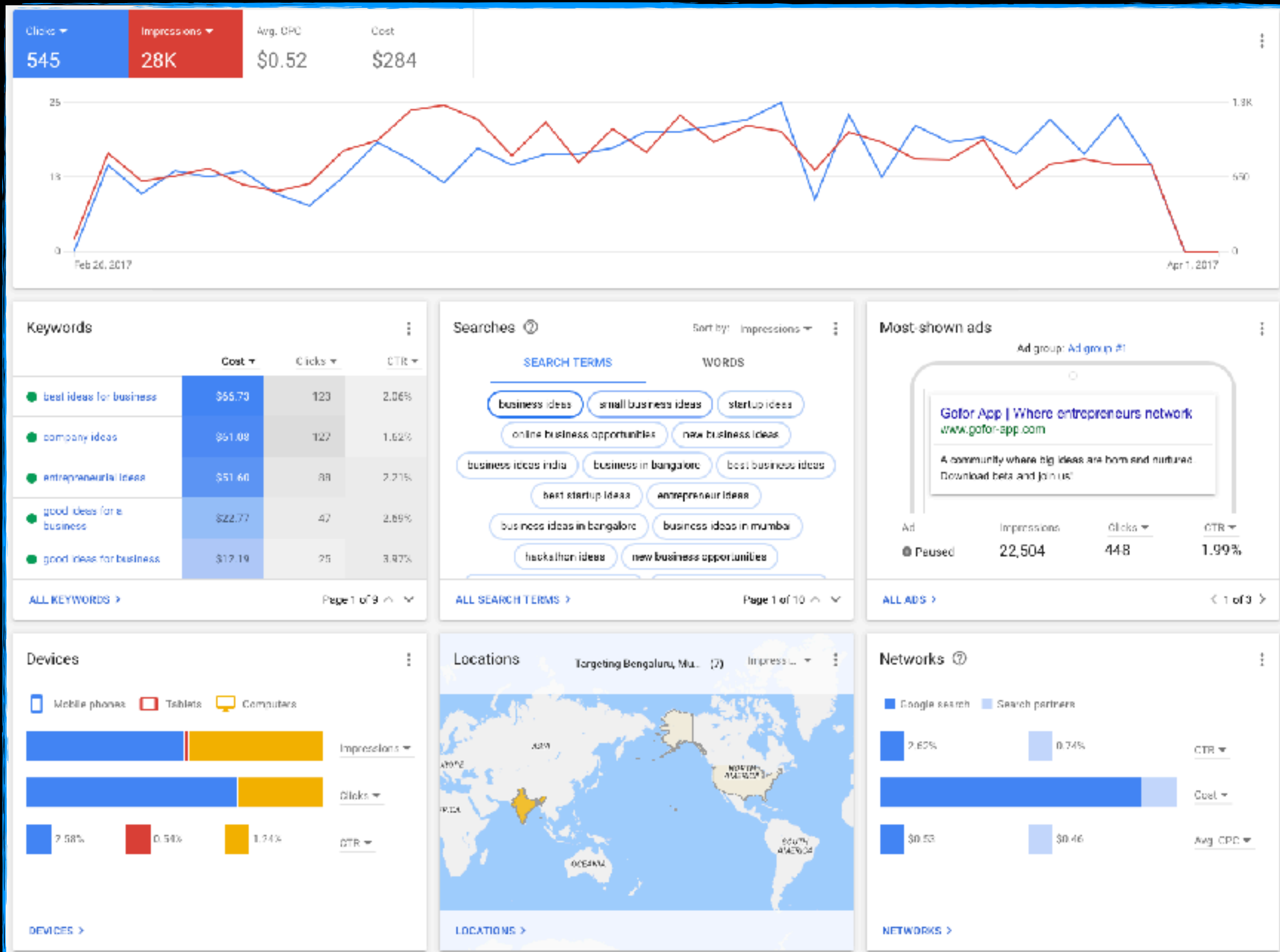
Displays a Search Results Page (SERP) = integration of web results, other data, and ads

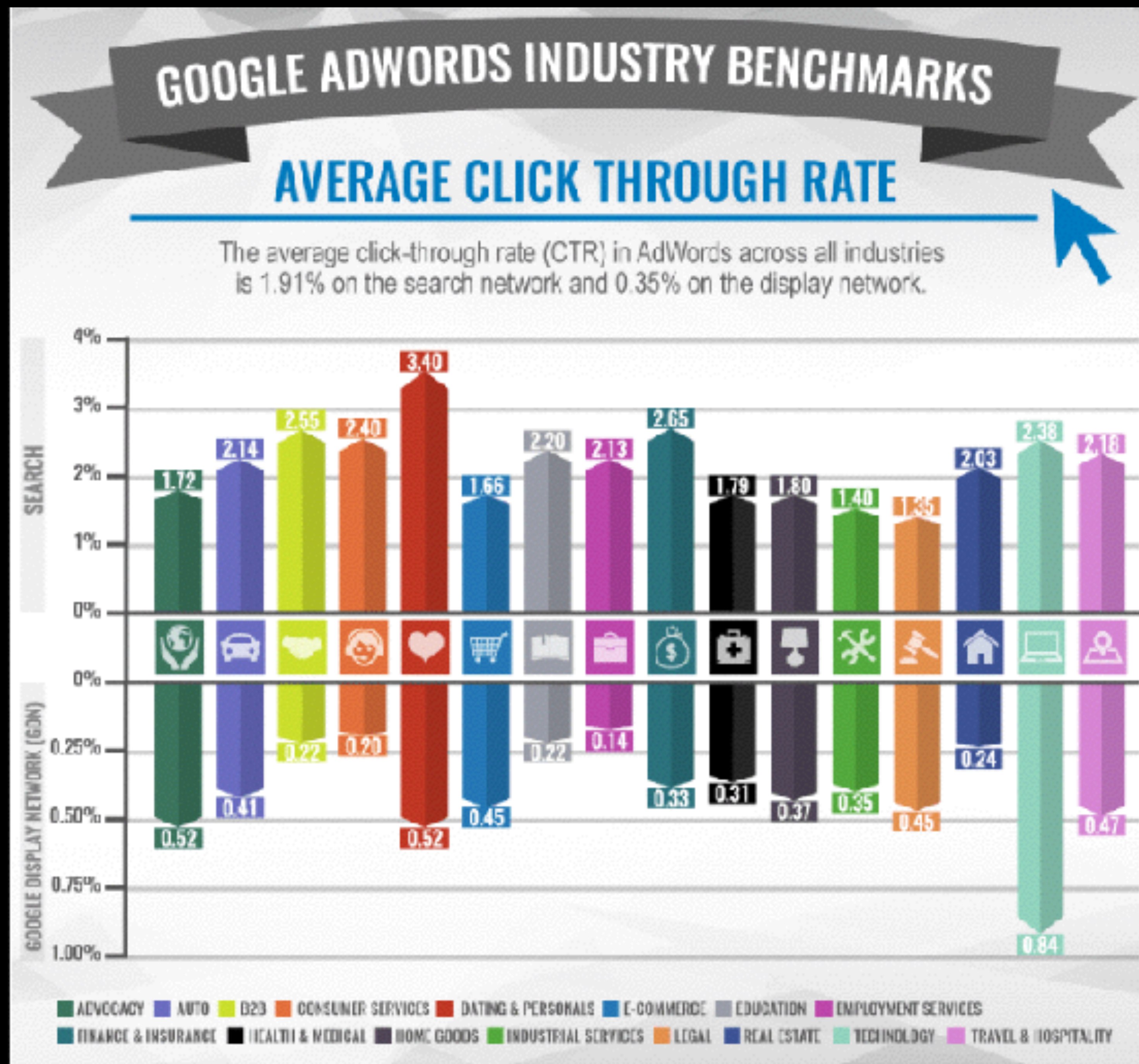
Make queries to search engine, expressing some intent

Users (ad relevance)

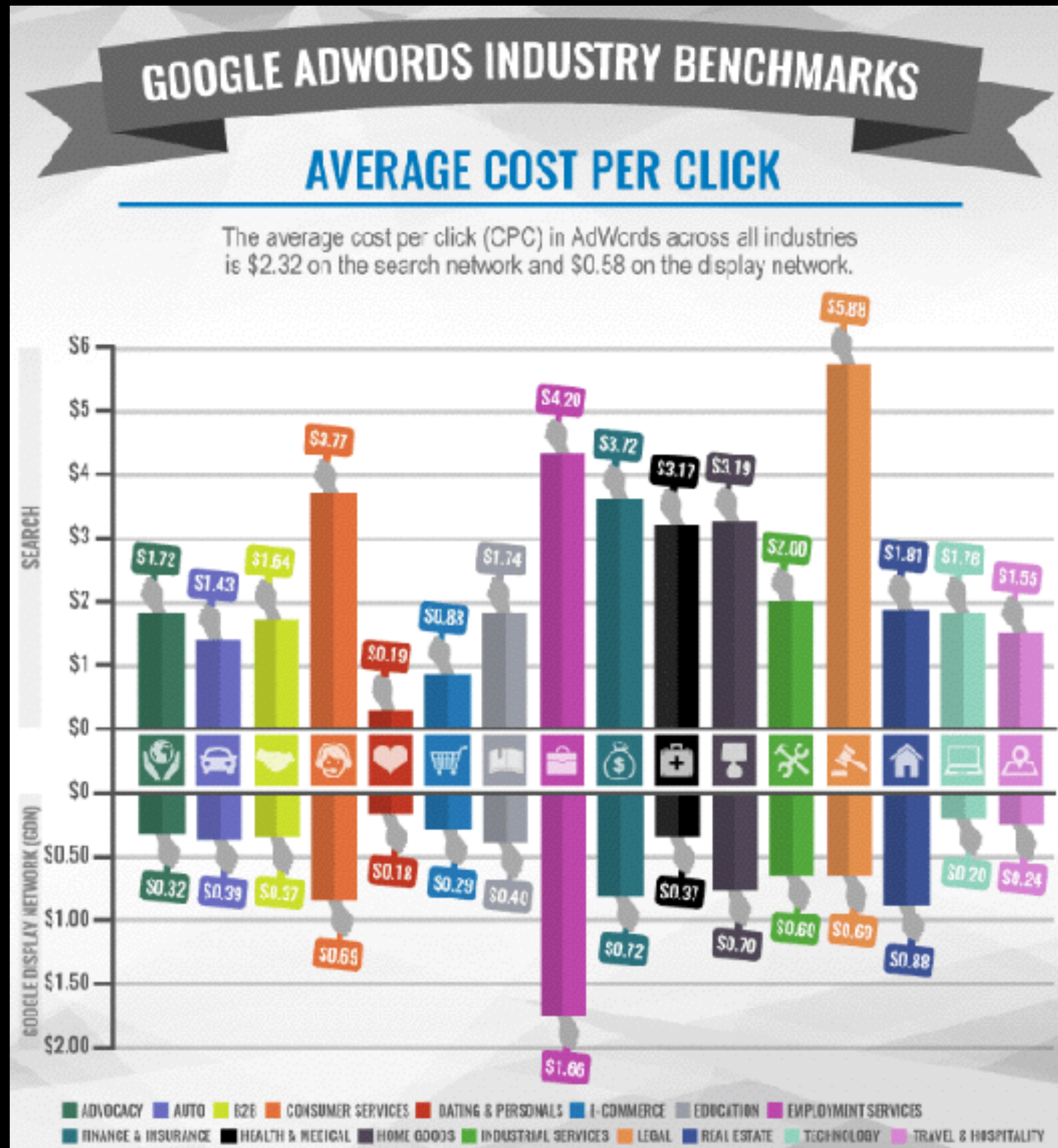
A Google Adsense campaign for a startup idea

GoFor

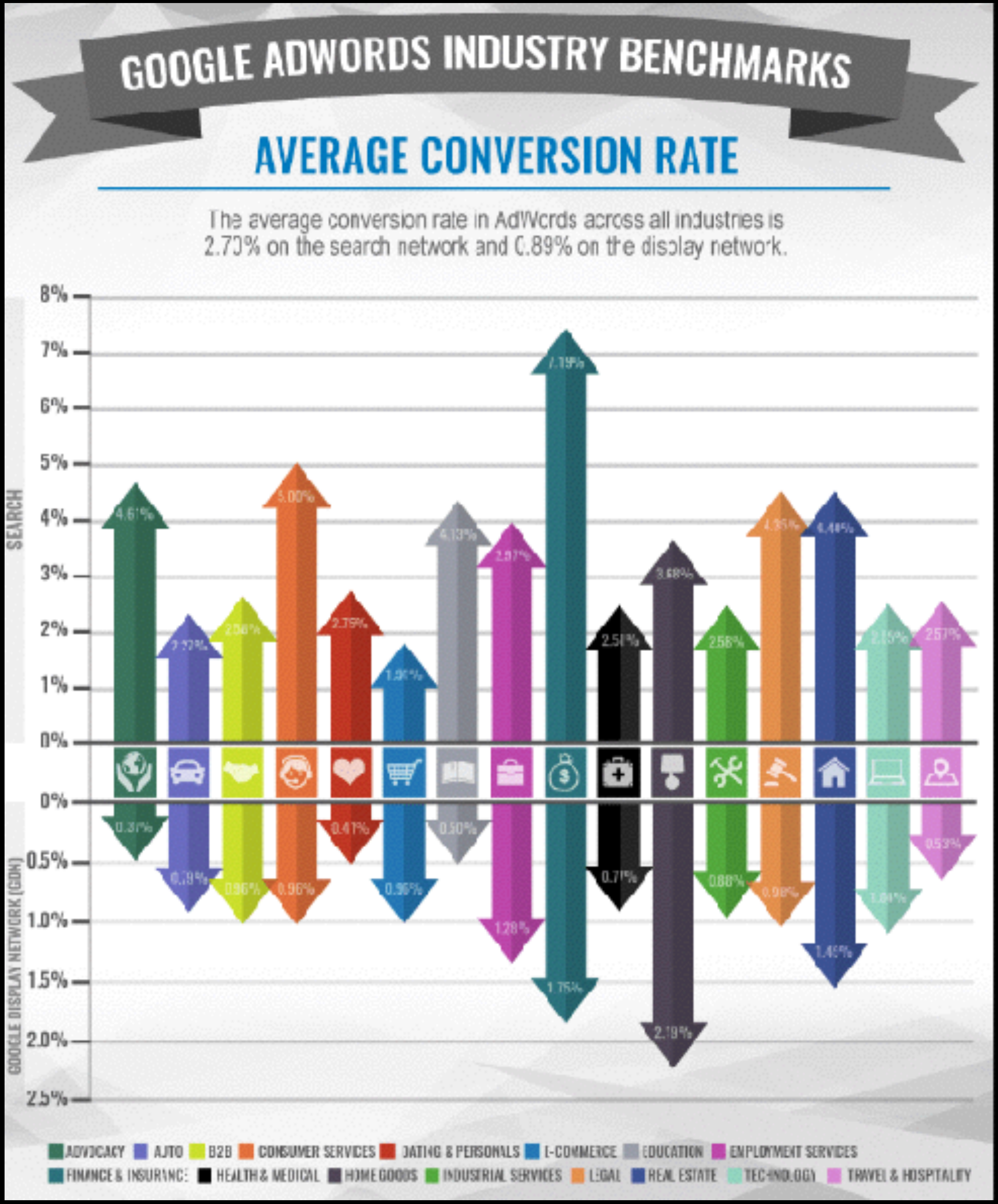

Clicks
545
Impressions
28K
Avg. CPC
$0.52
Cost
$284
Feb 24, 2017
Apr 1, 2017
Keywords
Cost
Clicks
CTR
best ideas for business
$55.73
123
2.06%
company ideas
$51.08
127
1.62%
entrepreneurial ideas
$51.60
88
2.21%
good ideas for a business
$22.77
47
2.89%
good ideas for business
$17.19
25
3.97%
ALL KEYWORDS
Page 1 of 9
Searches
Sort by: Impressions
SEARCH TERMS
WORDS
business ideas
small business ideas
startup ideas
online business opportunities
new business ideas
business ideas india
business in bangalore
best business ideas
best startup ideas
entrepreneur ideas
business ideas in bangalore
business ideas in mumbai
hackathon ideas
new business opportunities
ALL SEARCH TERMS
Page 1 of 10
Most shown ads
Ad group: Ad group #1
Gofor App | Where entrepreneurs network
www.gofor-app.com
A community where big ideas are born and nurtured
Download beta and join us!
Ad
Impressions
Clicks
CTR
Paused
22,504
448
1.99%
ALL ADS
1 of 3
Devices
Mobile phones
Tablets
Computers
Impressions
Clicks
CTR
2.58%
0.54%
1.24%
DEVICES
Locations
Targeting Bengaluru, Mu...
LOCATIONS
Networks
Google search
Search partners
2.62%
0.74%
CTR
Cost
$0.53
$0.46
Avg. CPC
NETWORKS



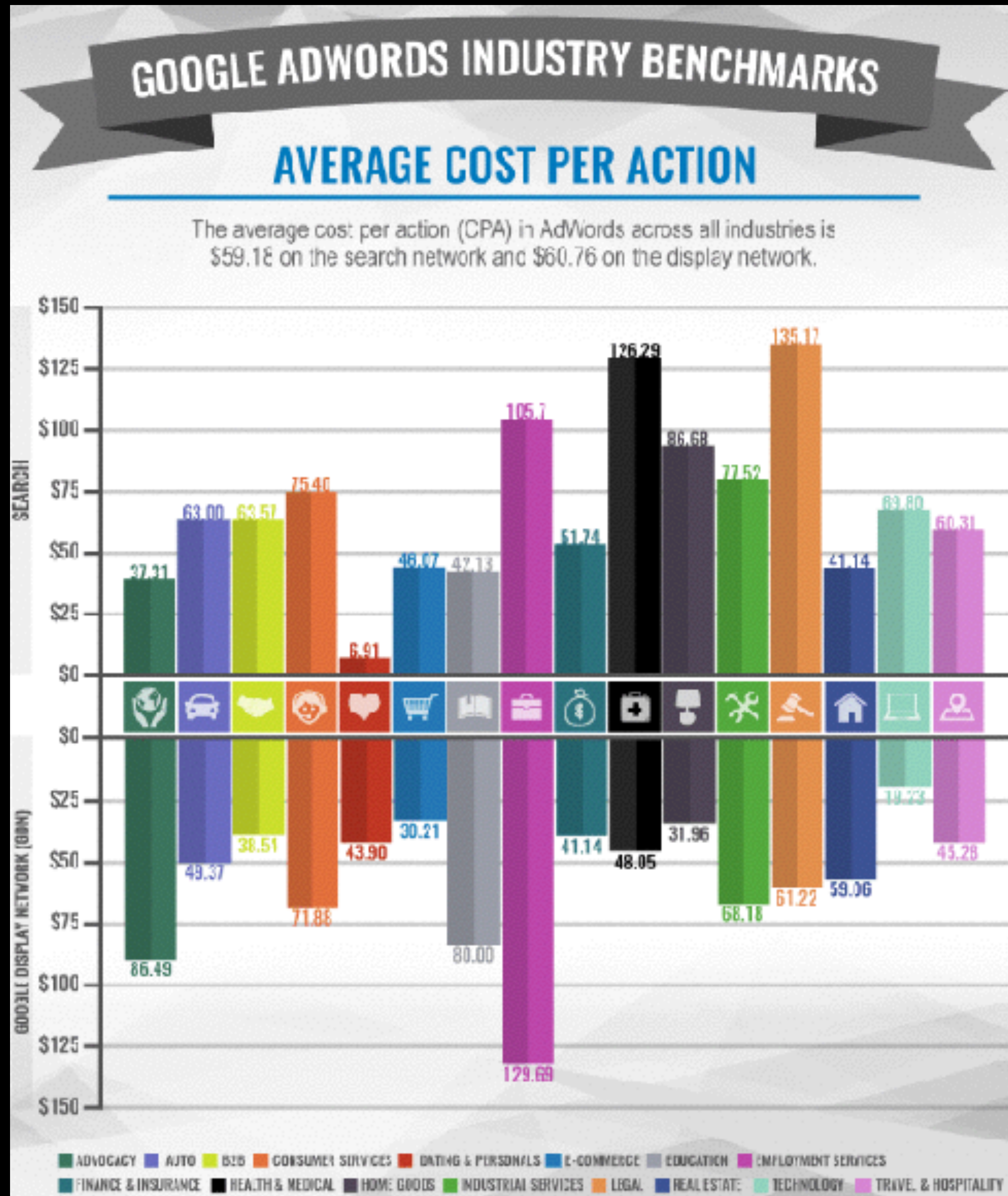
source: https://www.growthpoint.info/adwords-benchmarks/



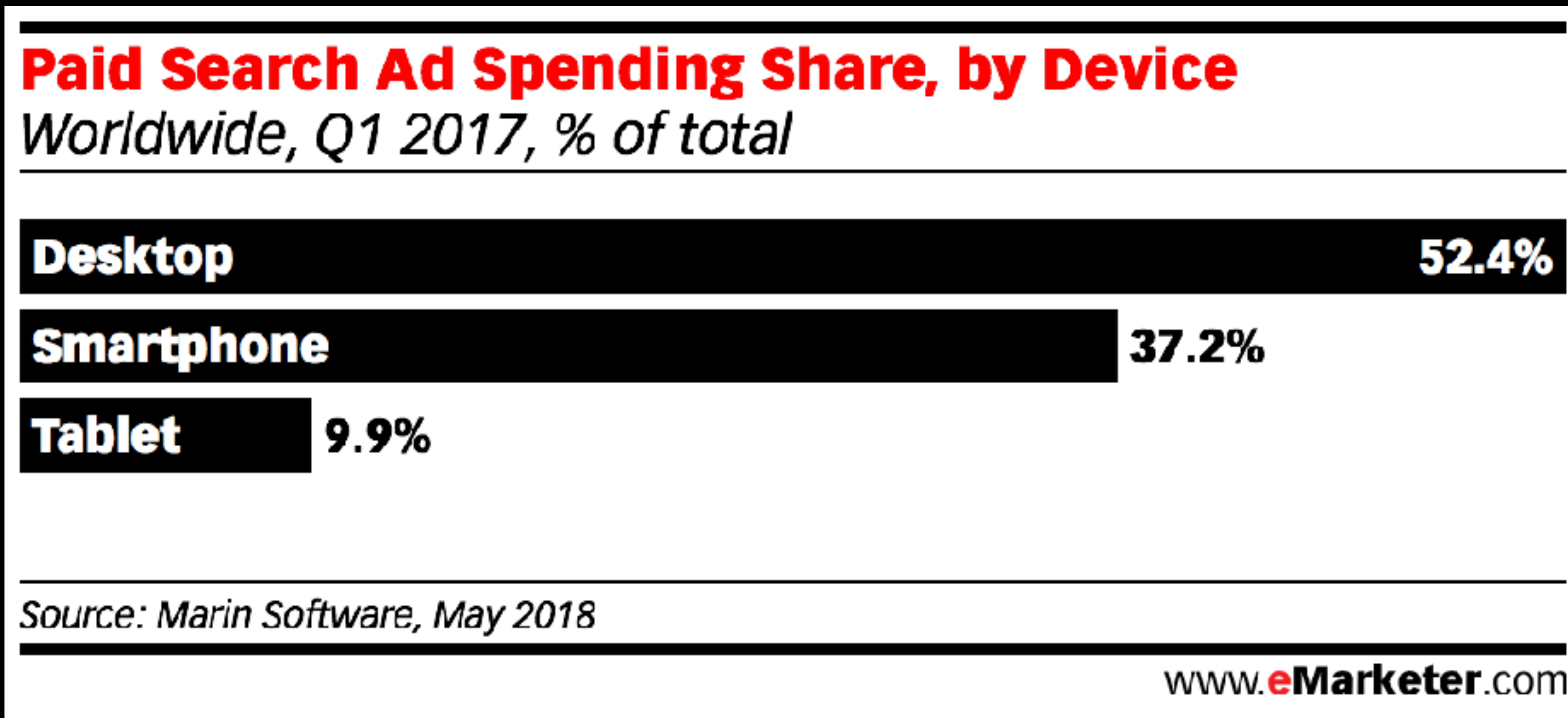
source: https://www.growthpoint.info/adwords-benchmarks/

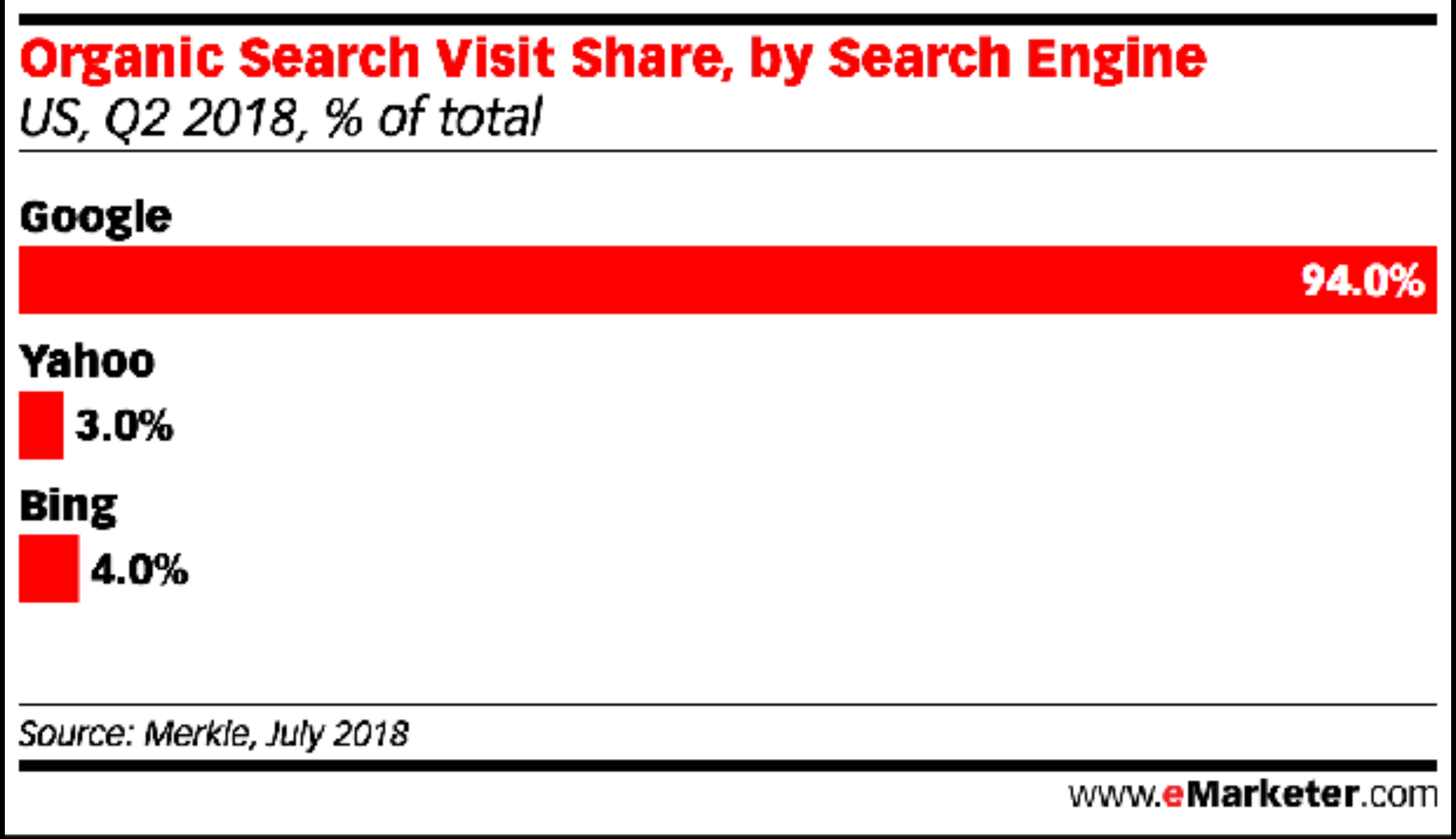


source: https://www.growthpoint.info/adwords-benchmarks/



source: https://www.growthpoint.info/adwords-benchmarks/


Paid Search Ad Spending Share, by Device
Worldwide, Q1 2017, % of total
Desktop
52.4%
Smartphone
37.2%
Tablet
9.9%
Source: Marin Software, May 2018
www.eMarketer.com

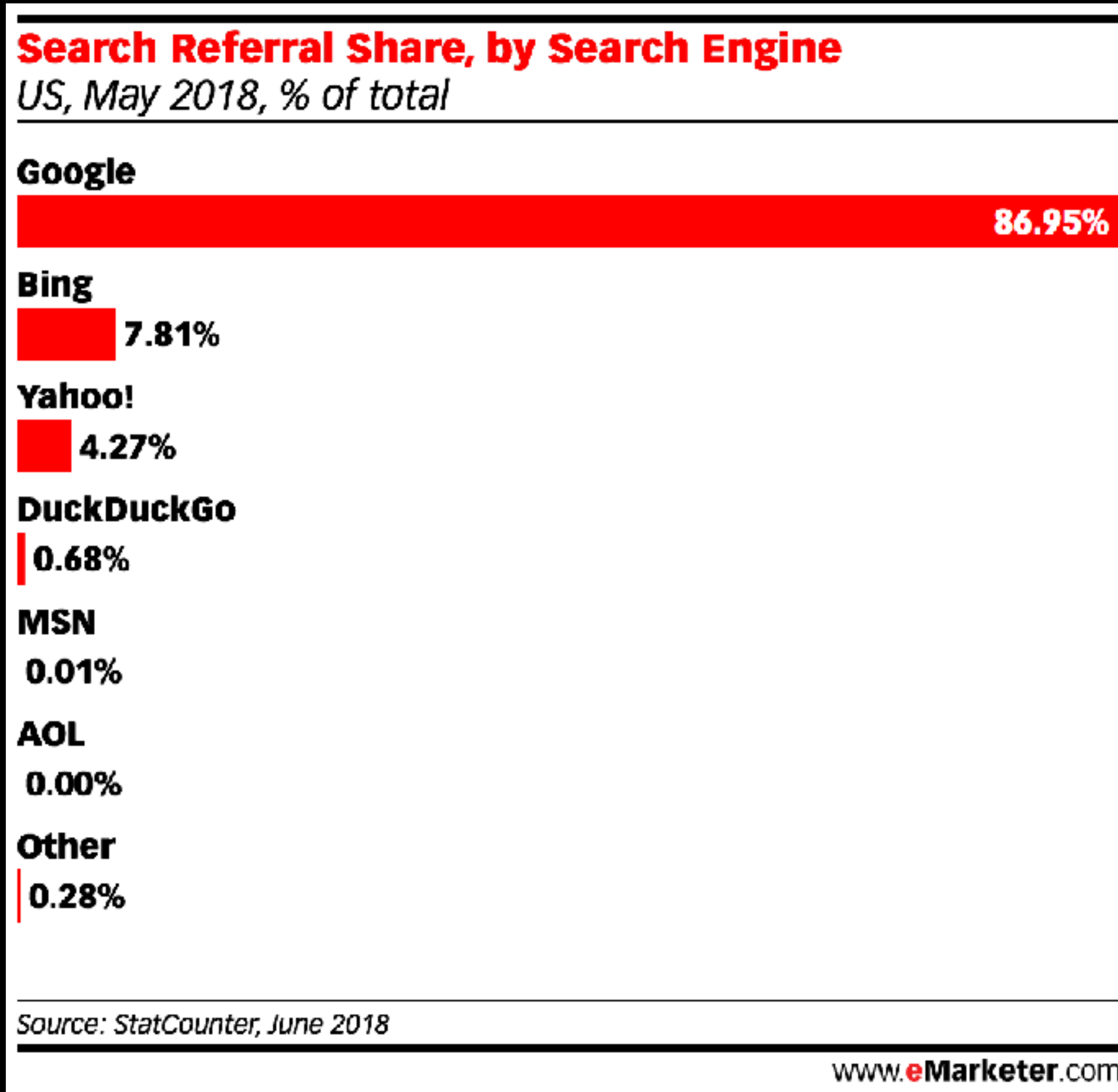

Organic Search Visit Share, by Search Engine
US, Q2 2018, % of total
Google
94.0%
Yahoo
3.0%
Bing
4.0%
Source: Merkle, July 2018
www.eMarketer.com

Google's share is increasing!

Organic Search Visit Share, by Search Engine, US

% of total

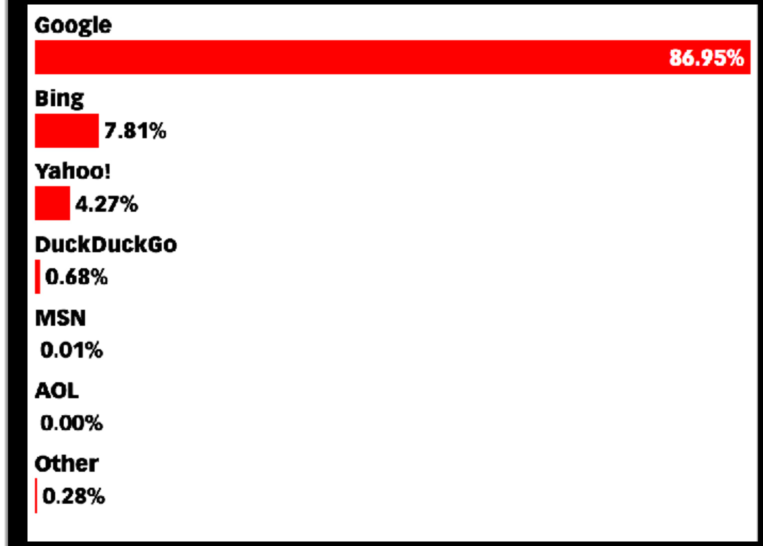
Timeframe	Q2 2017	Q3 2017	Q4 2017	Q1 2018	Q2 2018
Google	88.0%	89.0%	92.0%	93.0%	94.0%
Yahoo	5.0%	5.0%	3.0%	3.0%	3.0%
Bing	6.0%	6.0%	5.0%	4.0%	4.0%


Search Referral Share, by Search Engine
US, May 2018, % of total
Google
86.95%
Bing
7.81%
Yahoo!
4.27%
DuckDuckGo
0.68%
MSN
0.01%
AOL
0.00%
Other
0.28%
Source: StatCounter, June 2018
www.eMarketer.com

Search Referral Share, by Search Engine, US

% of total

Timeframe	Jan 2018	Feb 2018	Mar 2018	Apr 2018	May 2018
Google	88.24%	88.26%	87.37%	86.59%	86.95%
Bing	6.86%	6.69%	7.20%	7.83%	7.81%
Yahoo!	3.91%	4.08%	4.43%	4.54%	4.27%
DuckDuckGo	0.60%	0.52%	0.58%	0.66%	0.68%
MSN	0.09%	0.10%	0.04%	0.01%	0.01%
AOL	0.01%	0.01%	0.00%	0.00%	0.00%
Other	0.29%	0.34%	0.38%	0.37%	0.28%



Alphabet Q2 report, 2018

Revenues

The following table presents our revenues, by segment and revenue source (in millions, unaudited):

	Three Months Ended June 30, 2017	Three Months Ended June 30, 2018	Six Months Ended June 30, 2017	Six Months Ended June 30, 2018
Google segment				
Google properties revenues	$ 18,425	$ 23,262	$ 35,828	$ 45,260
Google Network Members' properties revenues	4,247	4,825	8,255	9,469
Google advertising revenues	22,672	28,087	44,083	54,729
Google other revenues	3,241	4,425	6,448	8,779
Google segment revenues	25,913	32,512	50,531	63,508
Other Bets				
Other Bets revenues	97	145	229	295
Revenues	$ 26,010	$ 32,657	$ 50,760	$ 63,803

Google's ad revenue is 86% of total

Advertisers can specify budgets

- Spend it quickly; till out of money
- Spend it slowly; till end-of-day
- Spend it as the Search Engine sees fit (engine can use this nefariously to manipulate the price paid by other advertisers)

other twists

We can have “reserve prices”; the minimum cost to be shown on a given keyword (depends on the keyword)

Sometimes there are “minimum bids”; that is, minimum bid required to participate in action (could depend on advertiser and keyword)

Search Engine perspective

Three problems

Computational Advertising

1. Ad retrieval (match to query/context)
2. Ordering the ads
3. Pricing on a click-through

Information Retrieval

Economics / AGT

US Digital and Total Ad Spending, by Format, 2013-2019

billions

	2013	2014	2015	2016	2017	2018	2019
Desktop	**$35.51**	**$37.09**	**$38.71**	**$35.90**	**$38.10**	**$38.16**	**$38.29**
—Search	$18.49	$18.94	$20.49	$17.75	$18.53	$18.16	$17.79
—Banner	$10.02	$10.15	$9.69	$8.70	$8.99	$8.54	$8.29
—Video	$2.82	$3.34	$4.17	$4.93	$5.70	$6.44	$7.09
—Other*	$4.18	$4.67	$4.37	$4.51	$4.88	$5.02	$5.12
Mobile	**$7.27**	**$12.36**	**$20.84**	**$36.62**	**$49.93**	**$63.95**	**$79.09**
—Search	-	$5.93	$9.17	$17.21	$22.11	$26.97	$31.82
—Banner	-	-	$9.38	$13.88	$18.42	$23.21	$28.08
—Video	-	-	$1.67	$4.19	$6.25	$9.18	$13.22
—Other*	-	$0.37	$0.63	$1.34	$3.16	$4.59	$5.96
Total digital ad spending	**$42.78**	**$49.45**	**$59.55**	**$72.52**	**$88.03**	**$102.11**	**$117.38**
Total media ad spending	**$181.79**	**$187.28**	**$191.24**	**$203.24**	**$206.25**	**$214.09**	**$216.23**
—Digital % of total	23.5%	26.4%	31.1%	35.7%	42.7%	47.7%	54.3%

*Note: estimates are based on information from Interactive Advertising Bureau (IAB) and Magna Global; numbers may not add up to total due to rounding; *includes classifieds, digital audio, lead generation, rich media and sponsorships*
Source: J.P. Morgan, "J.P. Morgan Handbook: Internet," May 22, 2018

239751 www.eMarketer.com

US Programmatic Ad Benchmarks: CPC, CPM and CTR, by Format, 2012-2016

	2012	2013	2014	2015	2016
CPC					
Display	$2.92	$4.67	$2.98	$4.55	$5.69
Video	$1.20	$1.34	$3.20	$5.36	$4.67
Mobile	$2.92	$4.67	$0.32	$0.58	$1.77
Social	-	$0.30	$0.27	$0.20	$0.30
Total	**$1.14**	**$0.67**	**$0.49**	**$0.44**	**$0.93**
CPM					
Video	$2.92	$4.67	$11.53	$15.04	$10.76
Mobile	$1.86	$1.74	$1.60	$2.10	$4.65
Display	$1.86	$1.74	$1.97	$3.33	$4.13
Social	-	$0.59	$1.26	$2.00	$2.23
Total	**$1.24**	**$1.02**	**$1.58**	**$2.75**	**$3.75**
CTR					
Social	-	0.20%	0.47%	1.01%	0.73%
Mobile	0.06%	0.04%	0.51%	0.36%	0.26%
Video	0.06%	0.04%	0.36%	0.28%	0.23%
Display	0.06%	0.04%	0.07%	0.07%	0.07%
Total	**0.11%**	**0.15%**	**0.32%**	**0.62%**	**0.40%**

Source: Zenith, "Programmatic Marketing Forecasts 2017," Nov 20, 2017

232950 www.eMarketer.com

Net US Digital Ad Revenue Share, by Company, 2016-2020

% of total digital ad spending and billions

	2016	2017	2018	2019	2020
Google	**40.8%**	**38.6%**	**37.2%**	**36.2%**	**36.3%**
—YouTube	4.0%	4.3%	4.1%	3.9%	3.8%
Facebook	**17.1%**	**19.9%**	**19.6%**	**19.2%**	**19.3%**
—Instagram	2.2%	3.6%	5.1%	6.2%	7.2%
Microsoft (Microsoft and LinkedIn)	**4.6%**	**4.1%**	**3.9%**	**3.5%**	**3.3%**
—LinkedIn	1.0%	0.9%	0.9%	0.8%	0.8%
Oath	**1.8%**	**4.0%**	**3.4%**	**3.0%**	**2.7%**
Amazon	**1.5%**	**2.0%**	**2.7%**	**3.5%**	**4.5%**
Snapchat	**0.4%**	**0.6%**	**1.0%**	**1.4%**	**2.2%**
Twitter	**1.9%**	**1.3%**	**1.0%**	**0.9%**	**0.9%**
Yelp	**0.9%**	**0.8%**	**0.8%**	**0.8%**	**0.8%**
IAC	**0.7%**	**0.5%**	**0.4%**	**0.4%**	**0.3%**
Hulu	**0.4%**	**0.4%**	**0.4%**	**0.4%**	**0.3%**
Roku	**0.1%**	**0.2%**	**0.3%**	**0.4%**	**0.5%**
Yahoo	**3.1%**	-	-	-	-
Total digital ad spending (billions)	**$72.20**	**$90.39**	**$107.30**	**$125.75**	**$142.23**

Note: includes advertising that appears on desktop and laptop computers as well as mobile phones, tablets and other internet-connected devices, and includes all the various formats of advertising on those platforms; net ad revenues after companies pay traffic acquisition costs (TAC) to partner sites; Facebook advertising revenues include Instagram advertising revenues
Source: eMarketer, March 2018

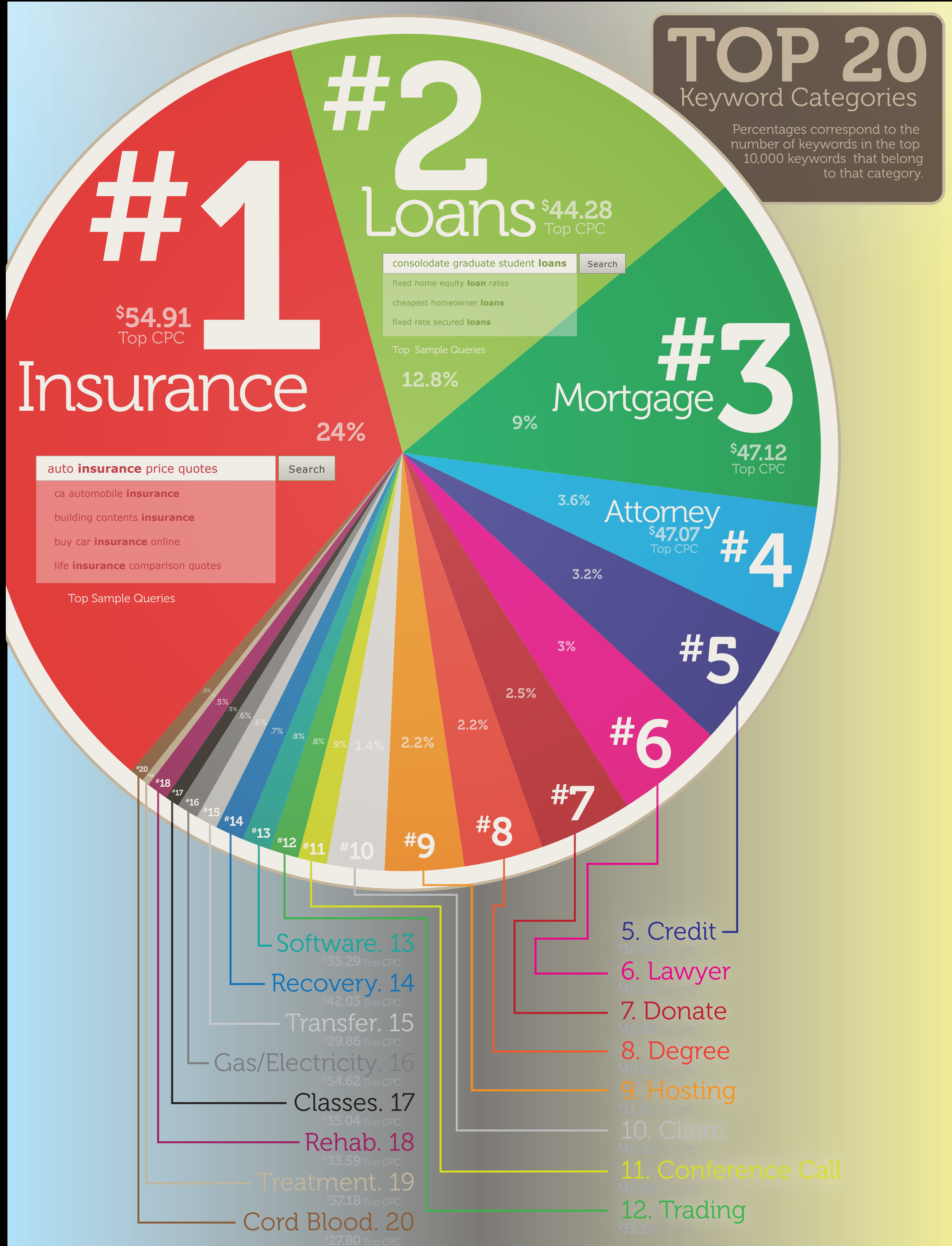
235999 www.eMarketer.com

US Paid Search Benchmarks: Click Rate, Conversion Rate, Cost per Click, Acquisition Cost and ROI, by Type of Keywords, May 2017

	Brand keywords	Generic keywords	Total
Click rate	8.1%	5.8%	8.1%
Conversion rate	7.2%	7.2%	7.2%
Cost-per-click	$4.64	$7.07	$6.14
Acquisition cost	$16-$17	$19-$20	$16.22
ROI	22%	24%	23%

Source: Data & Marketing Association (DMA) and Demand Metric, "DMA Response Rate Report 2017," June 21, 2017

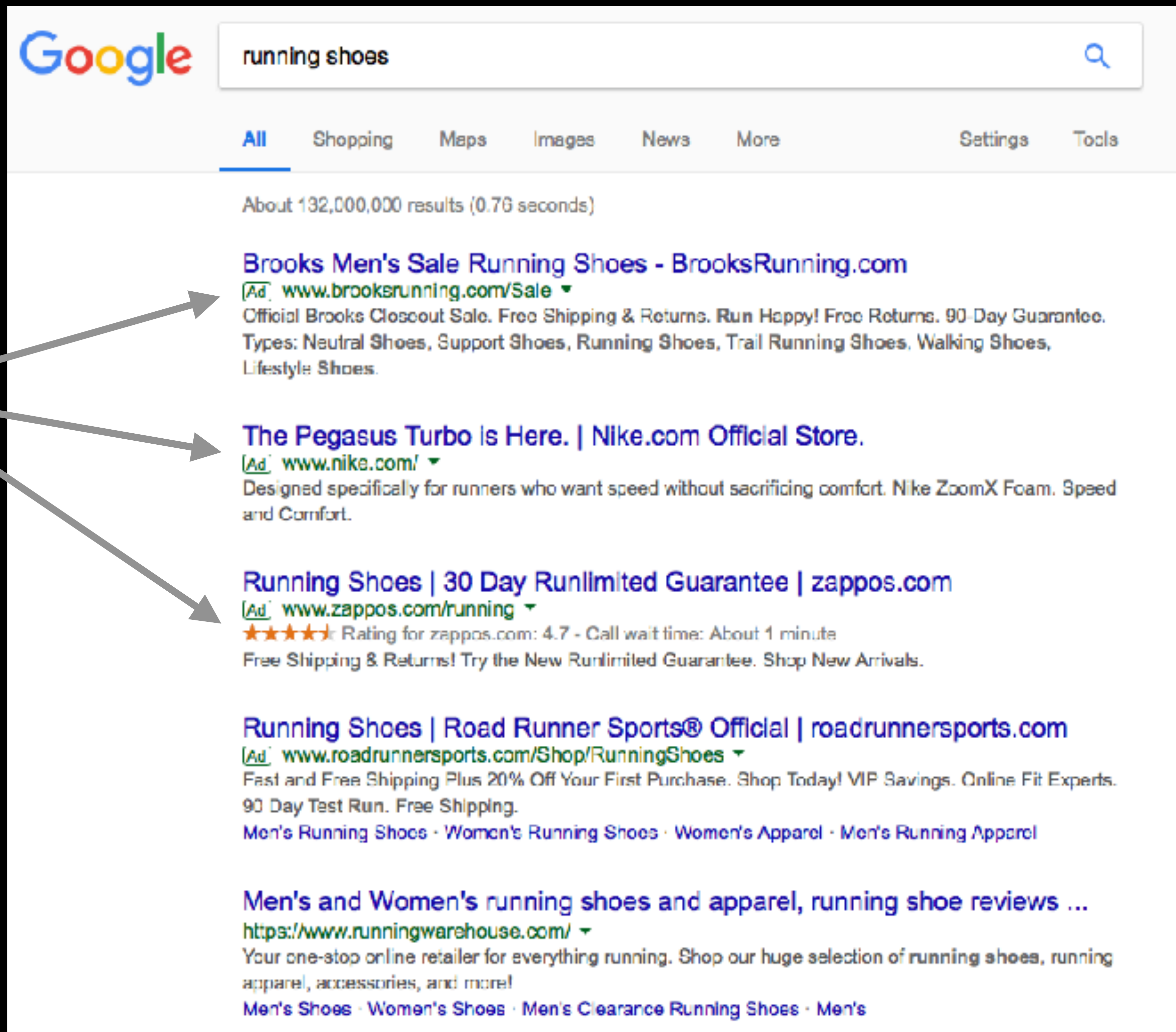
228453 www.eMarketer.com



source: WordStream, 2017

search query

search ads



Sponsored Search

Publishers
(Web page)

Advertisers
(ROI)

Submit ads associated to certain bid phrases; Bid for position; Pay (Cost Per Click) CPC

Executes query against web corpus + other data sources

Ad Network
(conducts auctions)

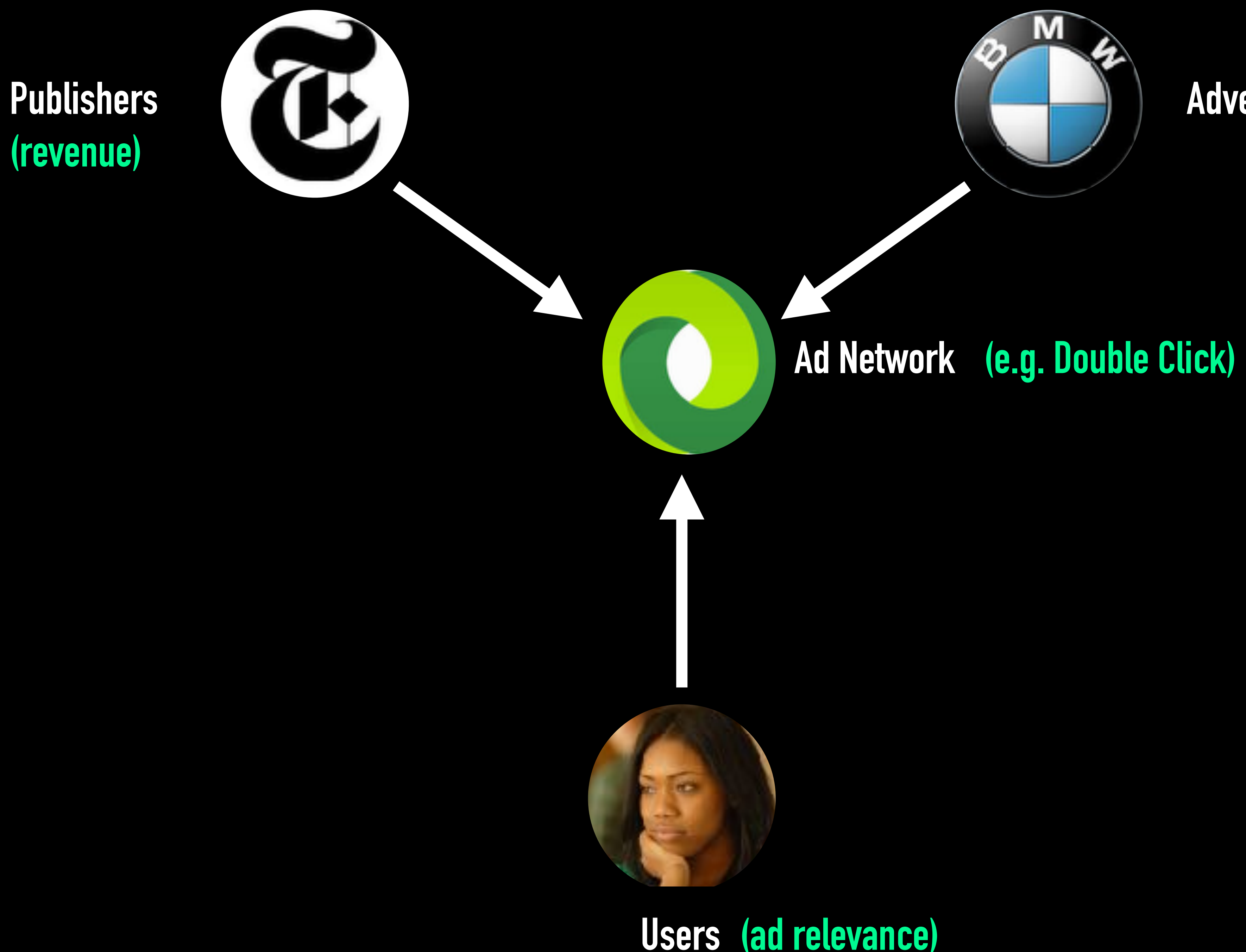
Executes query against the ad corpus

Displays a Search Results Page (SERP) = integration of web results, other data, and ads

Make queries to search engine, expressing some intent

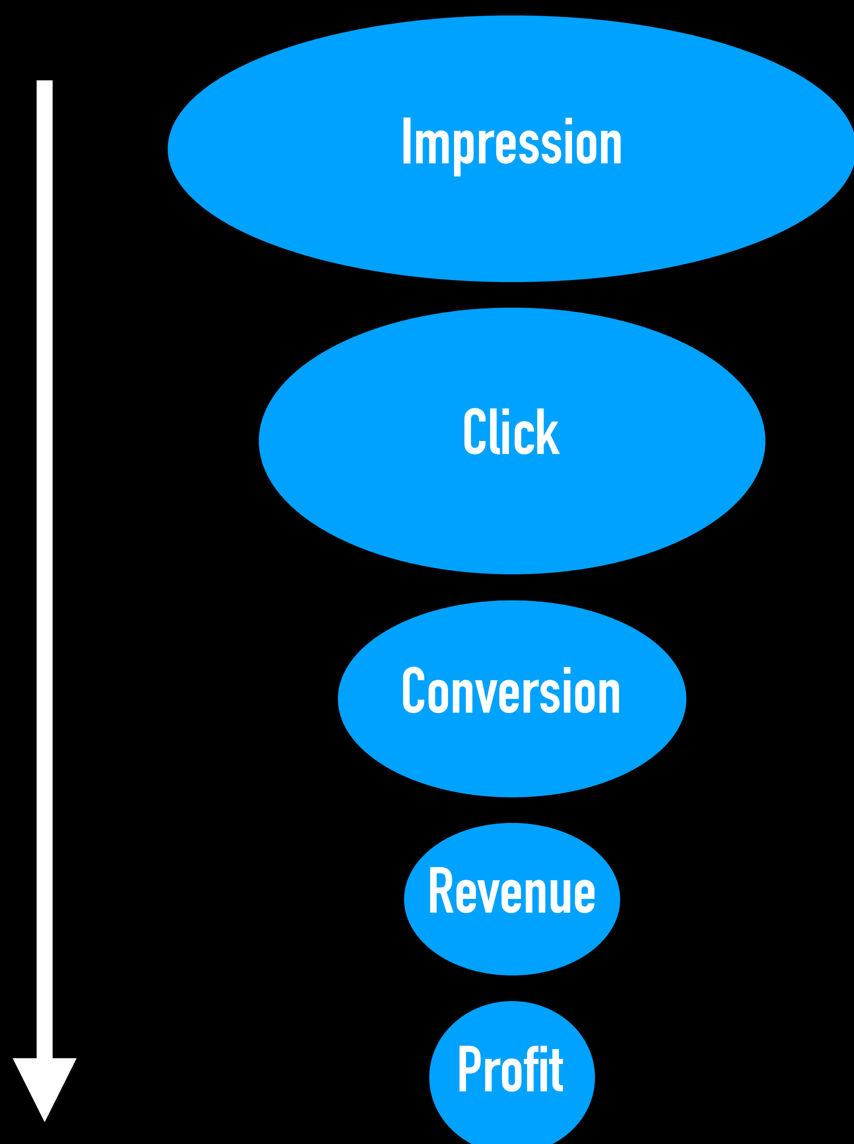
Users (ad relevance)

each actor has a different utility function



Advertiser Utility

The value funnel   Value = Long Term Profit



0.01%-10%

1%-10%

1x-10x

1x-10x

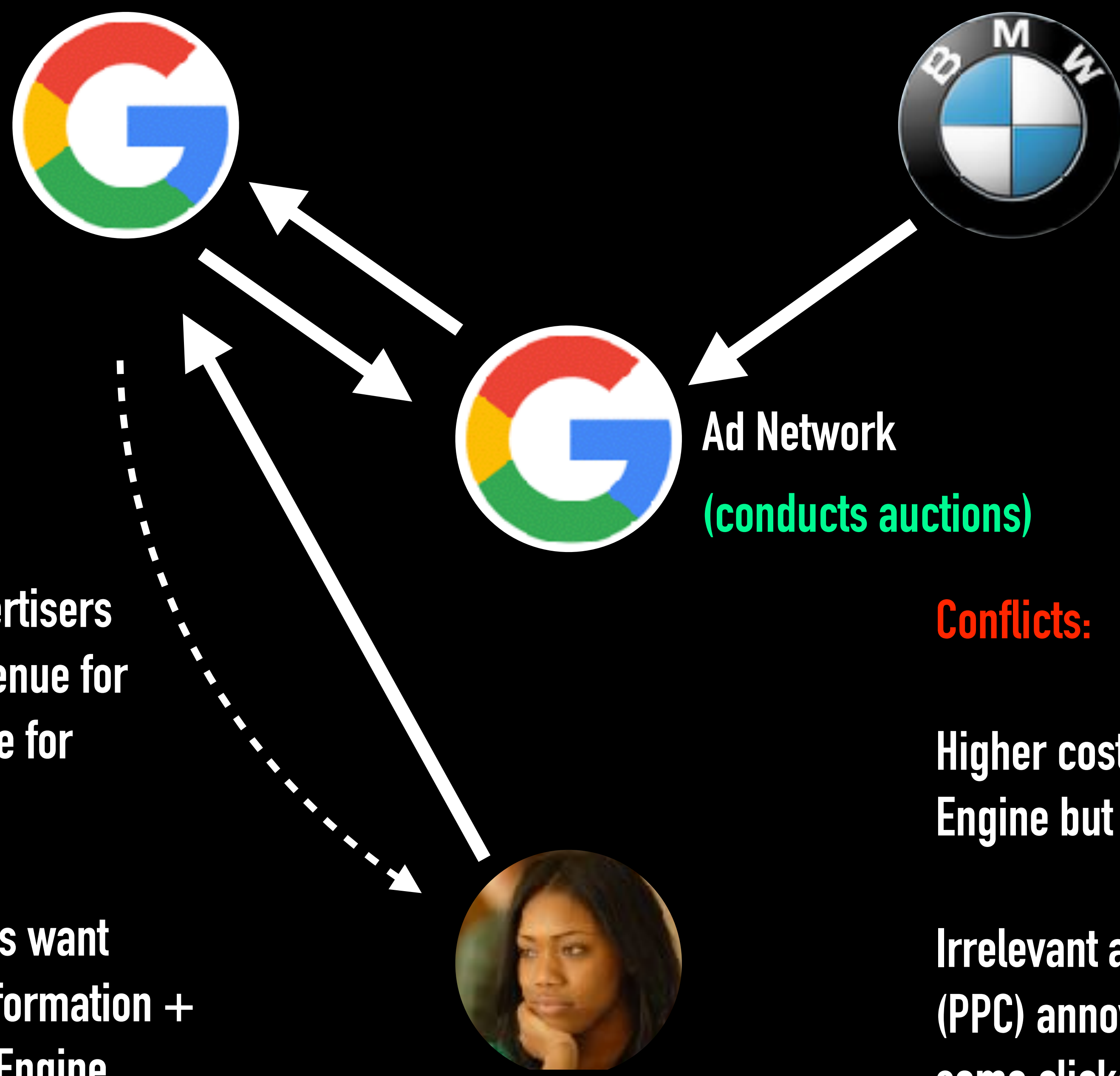
Future value—increase of the future profit due to the user action:

Ad impression might bring future conversions

Revenue events (upon user satisfaction) bring repeat customers

Immediate value—profit of the current event (conversion and below)

Conflicts and Synergies



Aligned:

Search Engine and advertisers want more clicks – revenue for Search Engine + volume for advertisers.

Search Engine and users want good ads – ads offer information + users click and Search Engine makes money

Conflicts:

Higher cost per click better for Search Engine but worse for advertiser.

Irrelevant ads with high pay-per-click (PPC) annoy most users but still get some clicks; clicks generate revenue for Search Engine, ROI for advertiser

How to choose an appropriate combination function?

Search Engine U_s

Advertisers U_a
(ROI)

Ad Network
(conducts auctions)

Utility $U = f(U_a, U_i, U_s)$

Model the long-term goal of the system
Parameterized to allow changes in the business priorities
Simple – so that business decisions can be done by the business owners!

U_i

Make queries to search engine, expressing some intent

Users (ad relevance)

How to choose an appropriate combination function?

Search Engine U_s

Advertisers U_a
(ROI)

Ad Network
(conducts auctions)

Utility $U = f(U_a, U_i, U_s)$

linear, convex combination

$$U = \alpha U_a + \beta U_i + \gamma U_s, \quad \alpha + \beta + \gamma = 1$$

U_i

Make queries to search engine, expressing some intent

Users (ad relevance)

How to choose an appropriate combination function?

Search Engine U_s

Advertisers U_a

Ad Network

Utility $U = f(U_a, U_i, U_s)$

find ads with user utility greater than α

Optimize which ads to show based on an auction (captures β)

utility functions are hard!

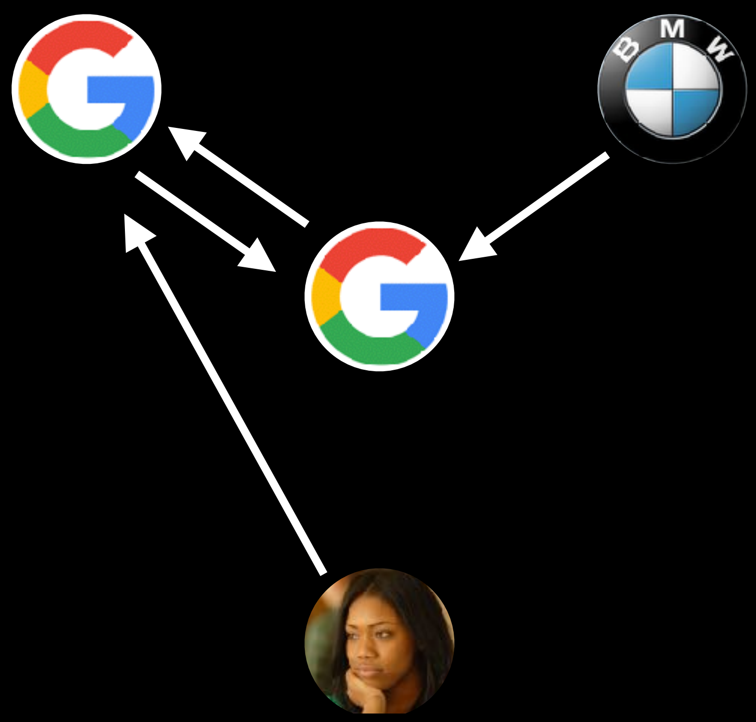
Instead:
User utility per search greater than α
Advertiser ROI per search greater than β

U_i

Users

However, ad relevance does not solve all problems

- When to advertise: certain queries are more suitable for advertising than others
- Interaction with the algorithmic side of the search (identifying what the user wants)



Why do it this way?

(As opposed to first find all ads with utility > β?)

Ad relevance is a simple proxy for total utility:

- Users—better experience
- Advertisers—better (more qualified) traffic but possible volume reduction
- Search Engine gets revenue gain through more clicks but possible revenue loss through lower coverage

Web-queries

- Queries are a succinct representation of the user's intent
  - The ultimate driver of the ad selection
  - Describe the need of the user
- Intent entropy is low in sponsored search!
- Before any grand design, let's look at the queries and their characteristics

types

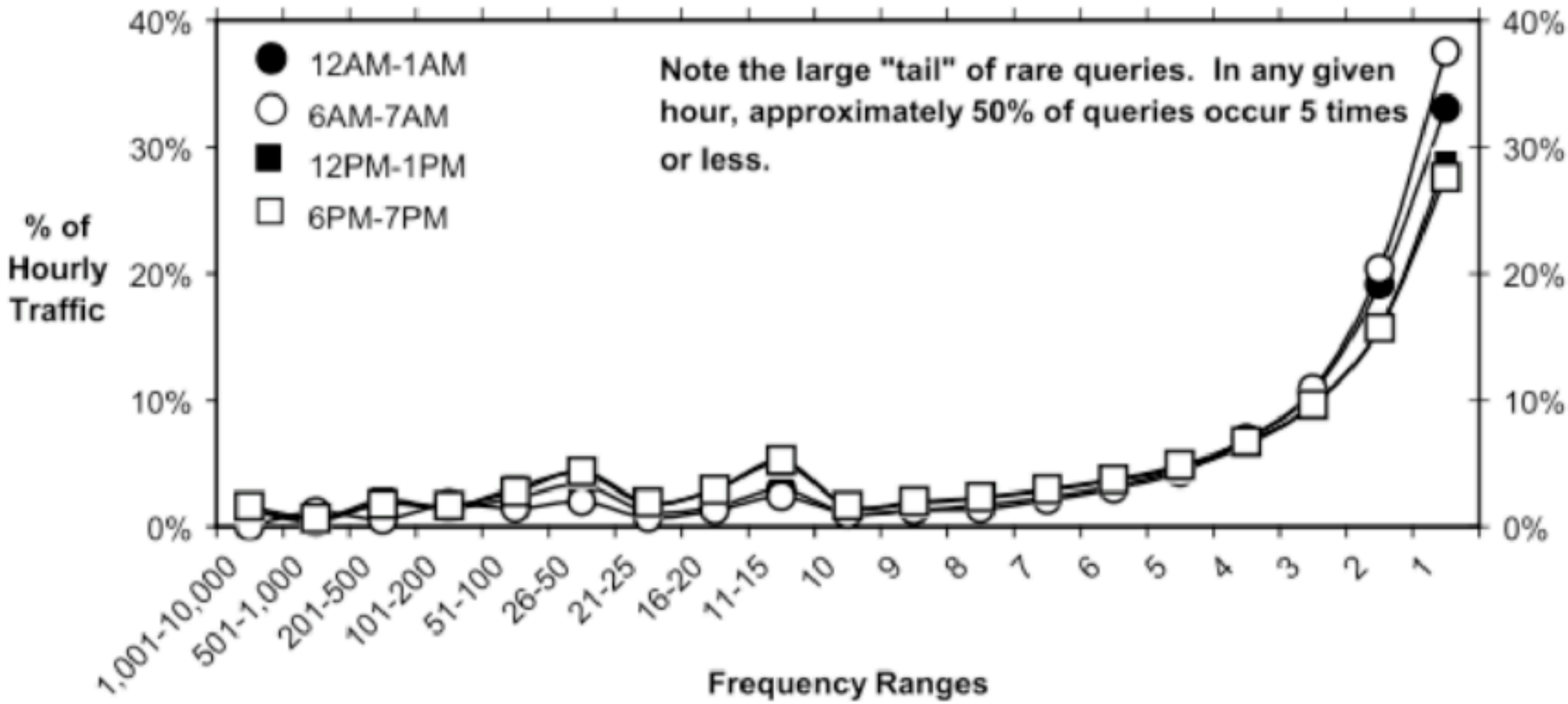
Informational – want to learn about something
- Flu prevention

Navigational – want to go to that page
- American Airlines

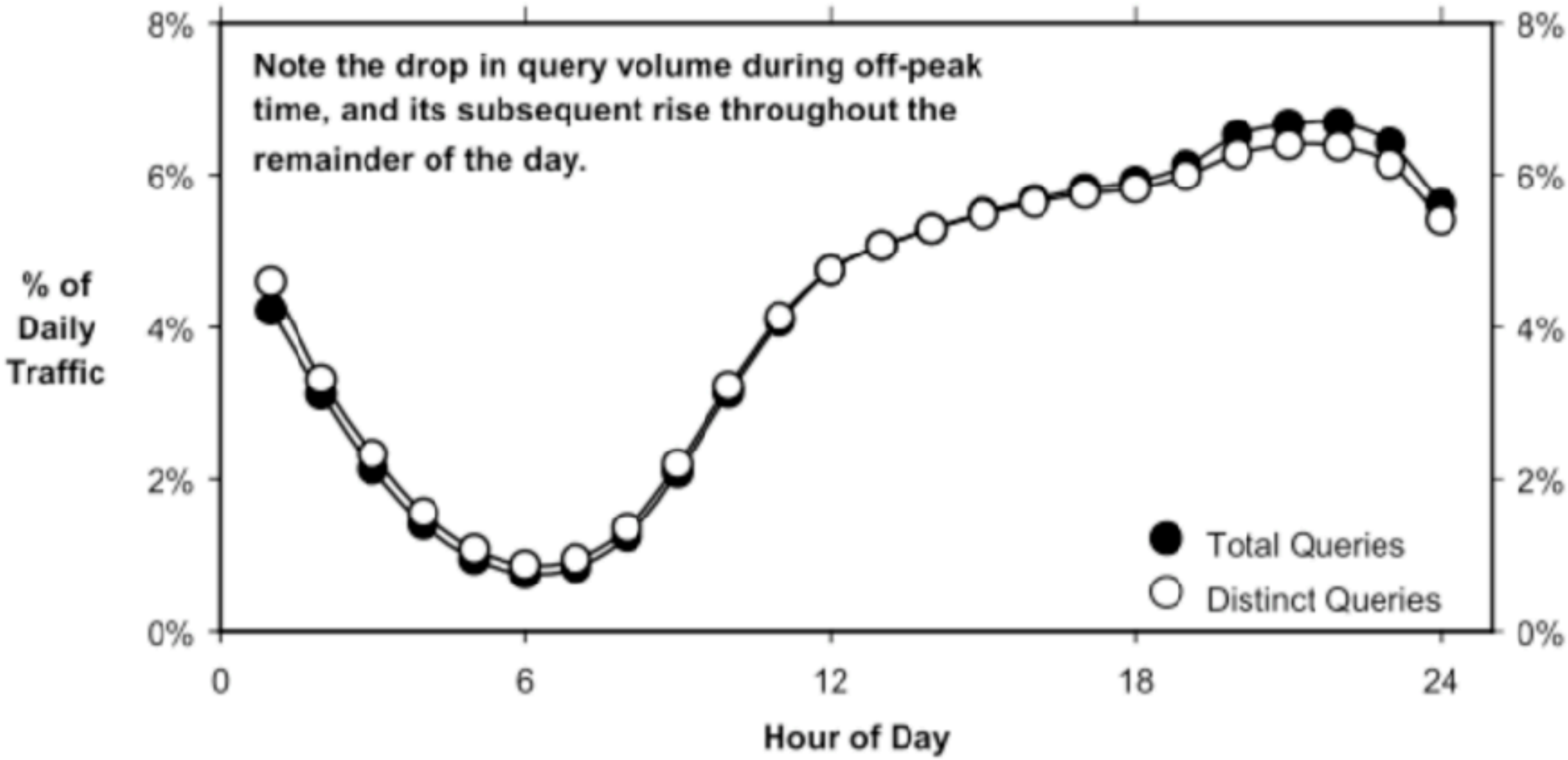
Transactional – want to do something (web-mediated)
- Access a service Downloads
- Shop
- New York weather

Gray areas
- Find a good hub
- Exploratory search "see what's there"

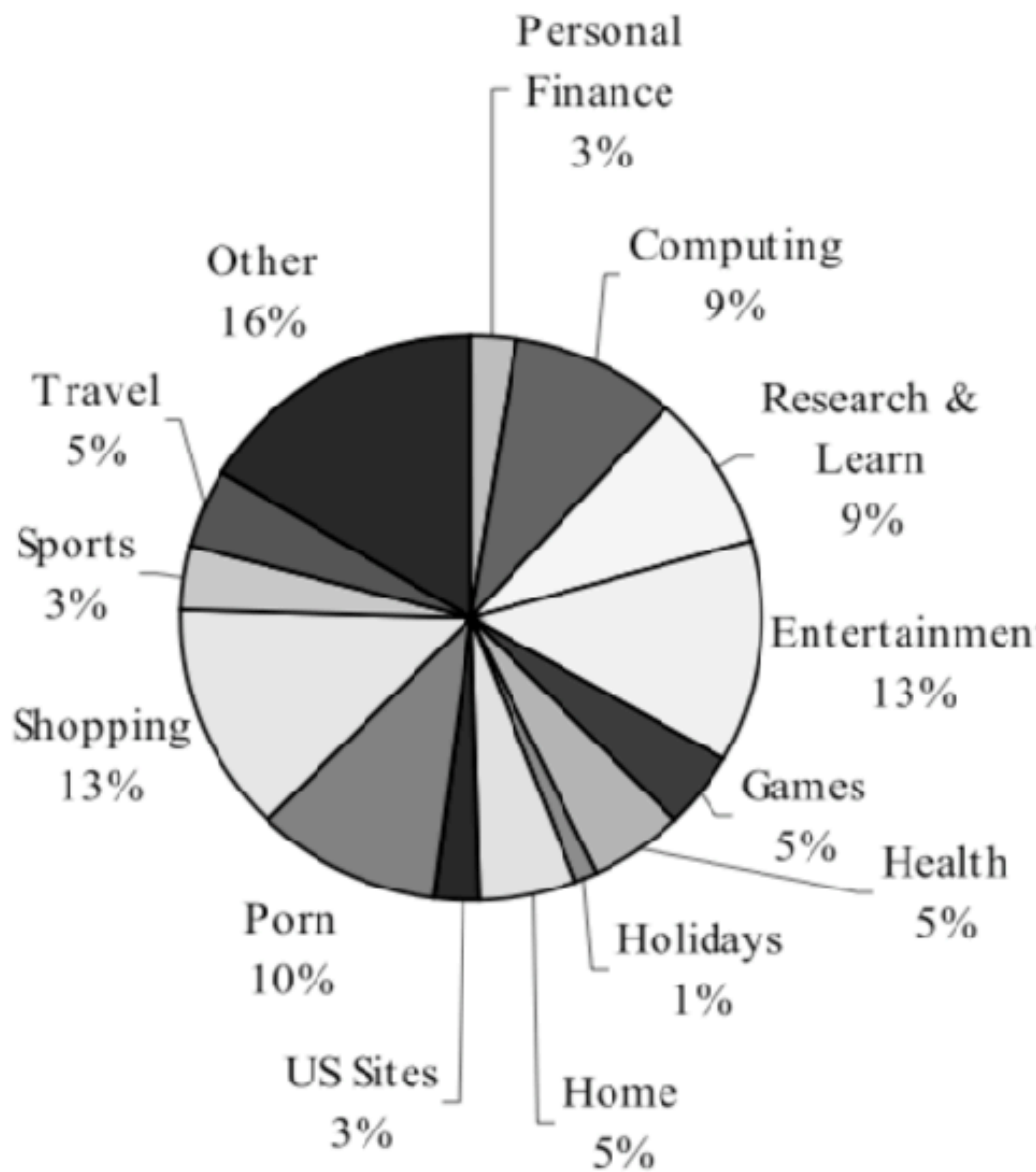
The long tail



query volume



what are they about?



Two options:

- Most people query the "usual" queries; a few do the "unusual" ones
- Large number people query the 'usual' queries; Most people also do a few unusual queries

why does the tail exist?

Study with online retailers supports the second hypothesis

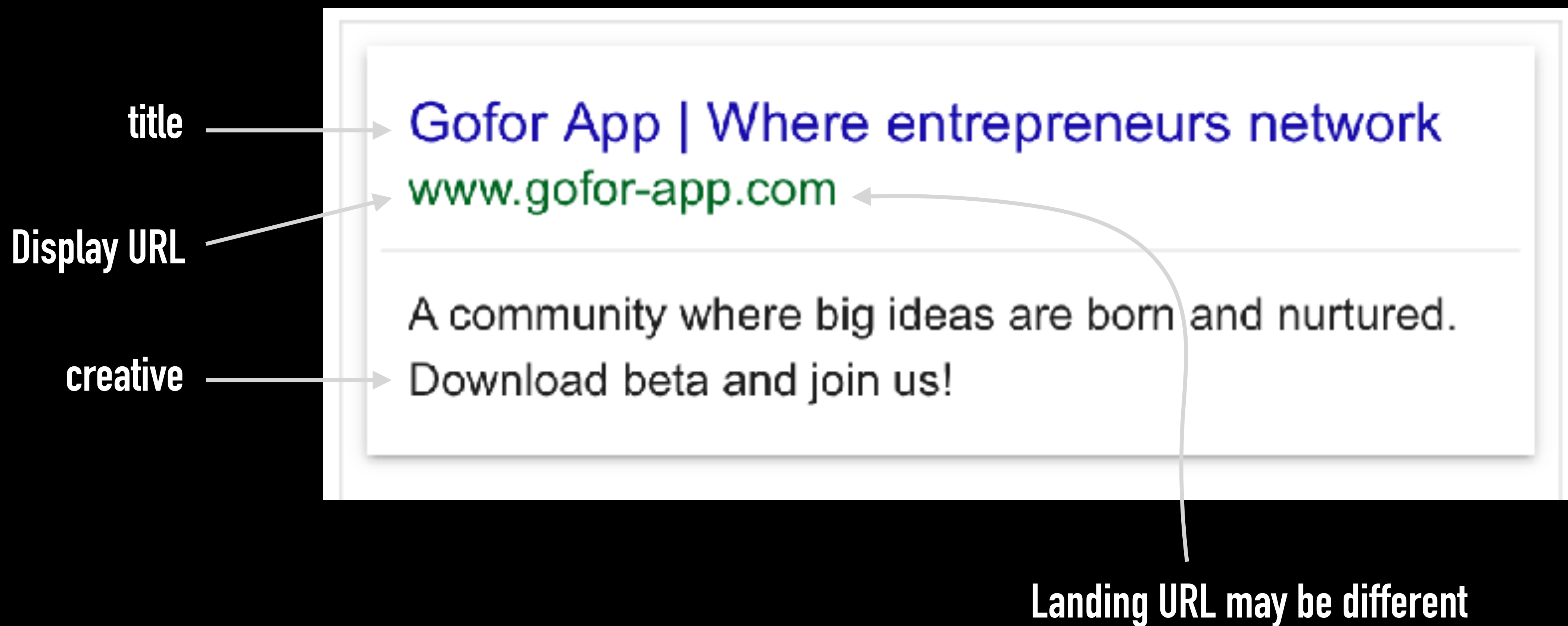
- Everybody is a bit eccentric, consuming both popular and niche products
- However, consumers exhibit varying degrees of eccentricity
- Availability of tail supply boosts even sales of popular items—one stop shop. (How does this map to search engines?)

Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. 2010. Anatomy of the long tail: ordinary people with extraordinary tastes. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10). ACM, New York, NY, USA, 201-210.

textual ads

dissection

bid phrase : "best ideas for business"; max CPC $0.44



Advertisers can sell multiple products

Might have budgets for each product line and/or type of advertising (Advanced Match / Exact Match) or bunch of keywords

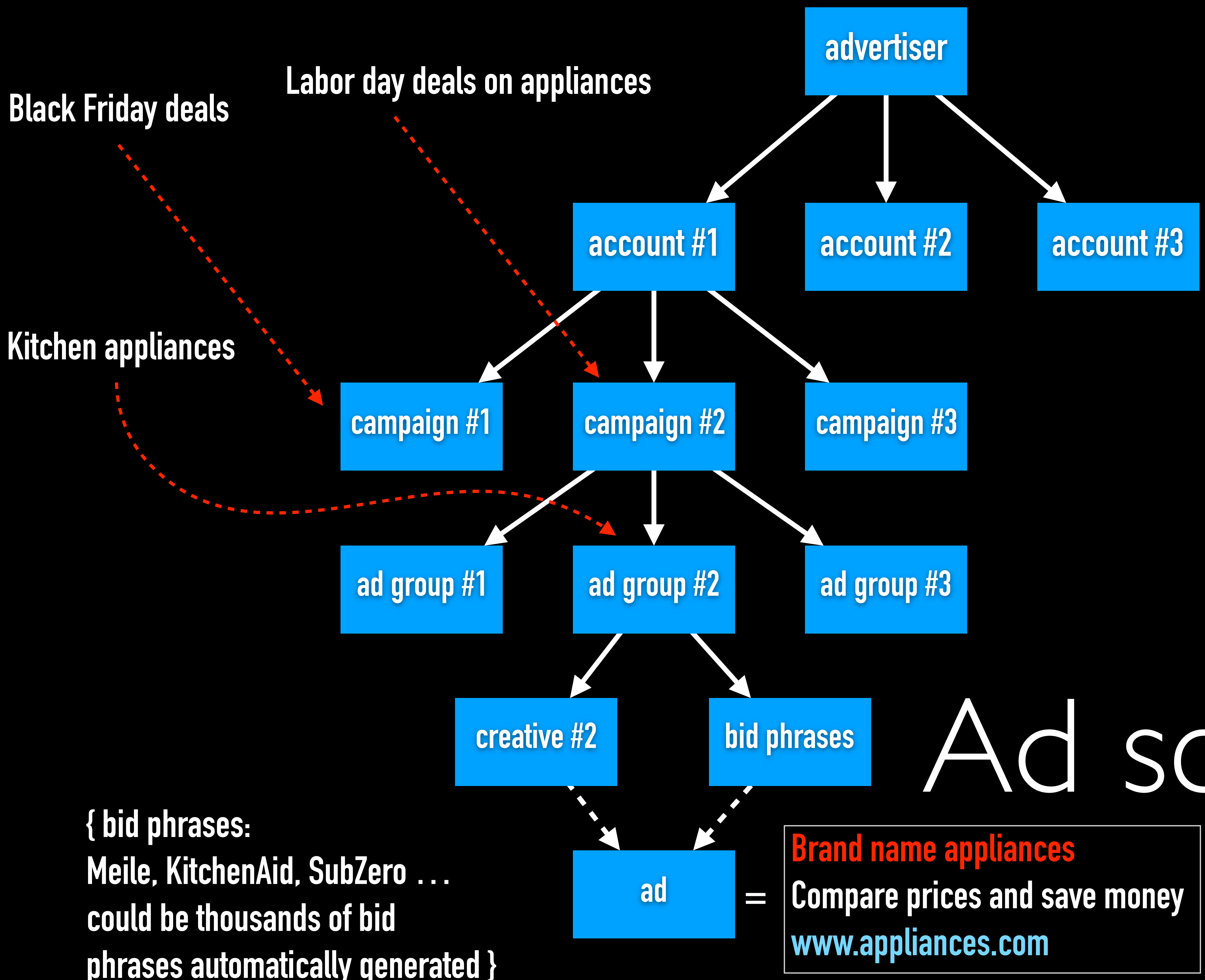
Beyond a single ad

Traditionally, a focused advertising effort is named a campaign

Within a campaign there could be multiple ad creatives

Financial reporting based on this hierarchy

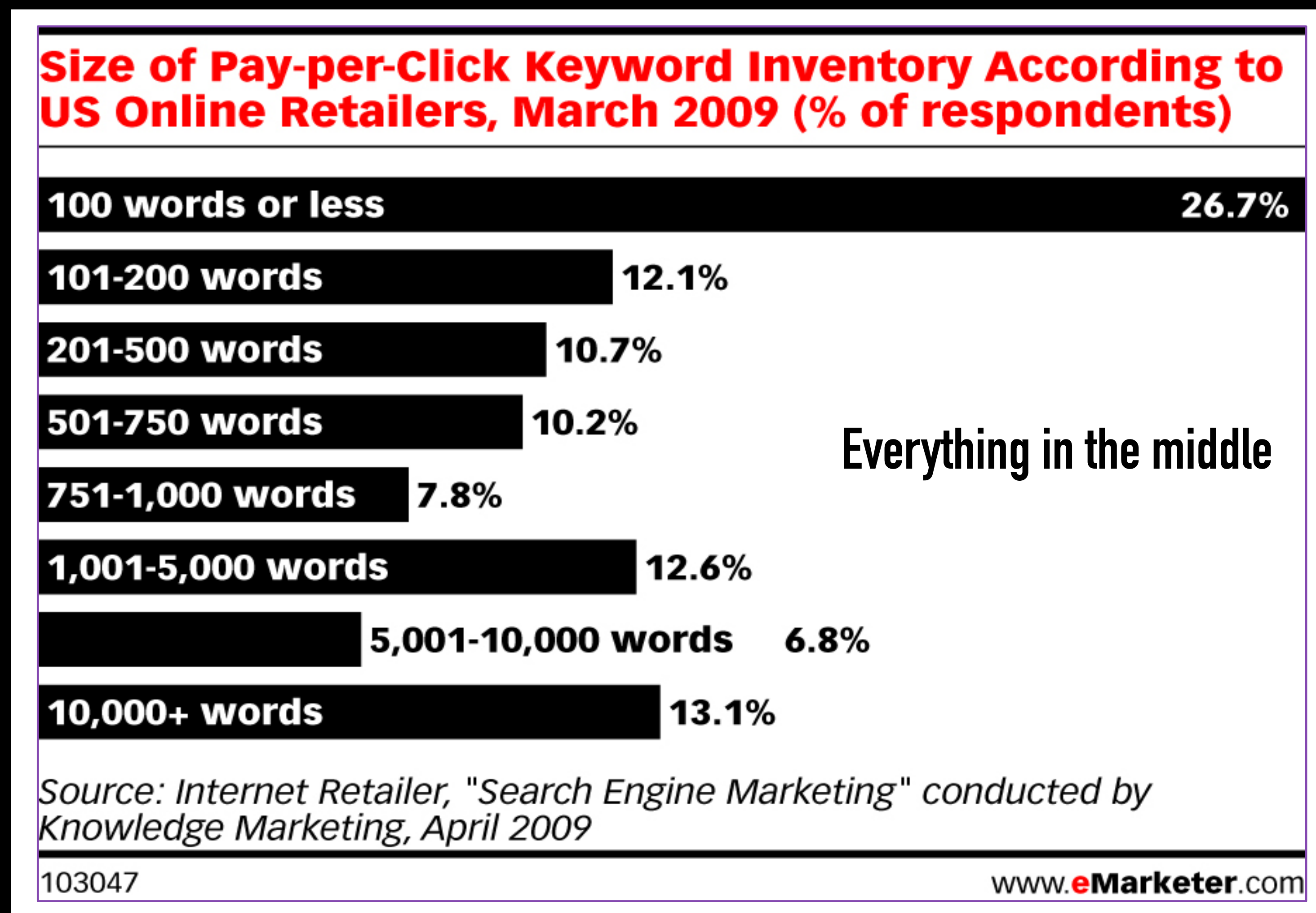
Ad schema



advertiser
account #1
account #2
account #3
campaign #1
campaign #2
campaign #3
ad group #1
ad group #2
ad group #3
creative #2
bid phrases
ad
=
Black Friday deals
Labor day deals on appliances
Kitchen appliances
{ bid phrases:
Meile, KitchenAid, SubZero . . .
could be thousands of bid
phrases automatically generated }
Brand name appliances
Compare prices and save money
www.appliances.com

keyword usage

Mom-and-pop shop



Ubiquitous: bid on query logs, facebook, Amazon, Ebay, . . .

Responsive: satisfy directly the intent of the query

query: Realgood golf clubs

ad: Buy Realgood golf clubs cheap!

ad-query relationship

Incidental: a user need not directly specified in the query

Related: Local golf course special

Competitive: Sureshot golf clubs

Associated: Rolex watches for golfers

Spam: Vitamins 

H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: quantifying the impact of landing pages in web advertising. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, pages 57–66, New York, NY, USA, 2009. ACM.

types of landing pages

Classify landing page types for all the ads for 200 queries from the 2005 KDD Cup labeled query set.

Four prevalent types:

1 **Home page** (25%):
Land on advertiser's home page

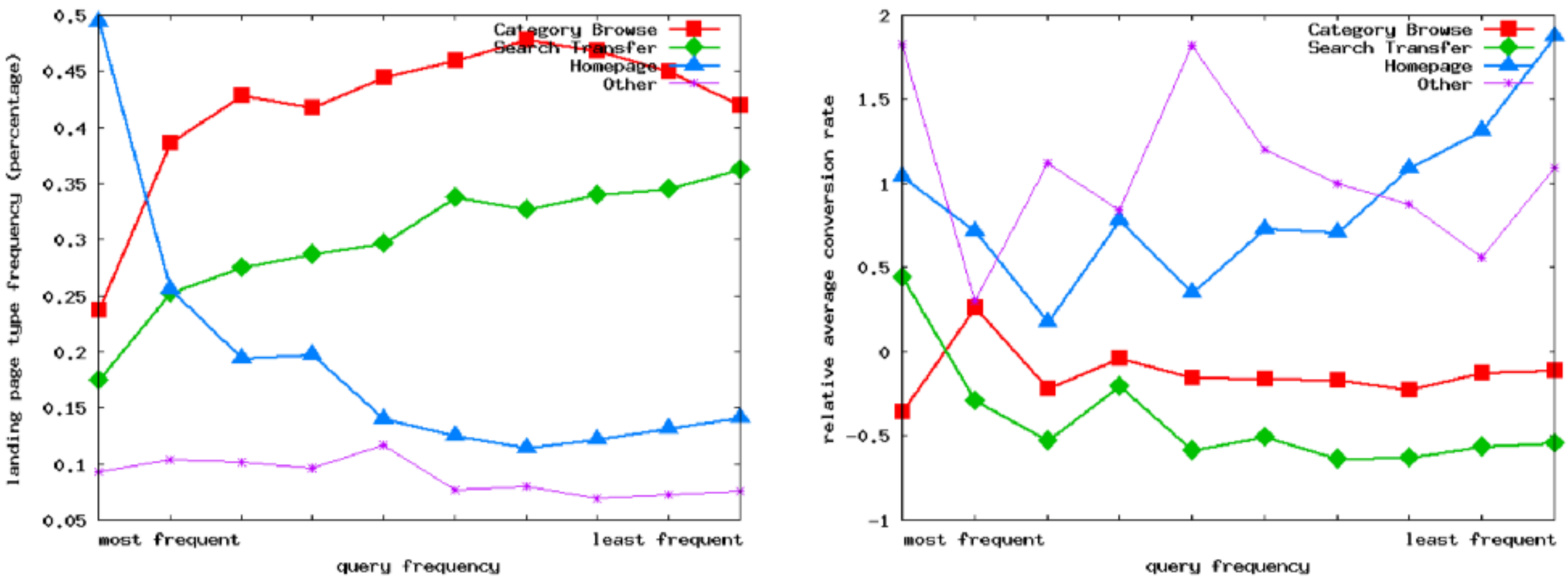
2 **Category** (37.5%):
Landing page captures the broad category of the query

3 **Search Transfer** (26%):
Land on dynamically generated search results (same q) on the advertiser's web page

- Product List – search within advertiser's web site
- Search Aggregation – search over other web sites

4 **Other** (11.5%):
Land on promotions and forms

category v. conversion



Ad Selection

Match types:

Exact – the ad's bid phrase matches the query

Advanced - the ad platform finds good ads for a given query

Implementation:

Database lookup

Similarity search

Phased selection

Sponsored search ad selection methods

Reactive vs predictive

Reactive: try and see using click data

Predictive: generalize from previous ad placement to predict performance

Data used (for predictive mostly)

Unsupervised

Click data

Relevance judgments

Exact match (EM)

The advertiser bid on that specific query a certain amount

Match types

Advanced match (AM) or “Broad match”

The advertiser did not bid on that specific keyword, but the query is deemed of interest to the advertiser.

Advertisers usually opt-in to subscribe to AM

Is the query "Miele dishwashers" the same as:

- Miele dishwasher (singular)
- Meile dishwashers (misspelling)
- Dishwashers by Miele (re-order, noise word)

Query normalization

- creating equivalences
- (e.g. "USA=U.S.A")

Exact match challenges

Which exact match to select among many?

- Varying quality
- Spam vs.Ham
- Quality of landing page

Suitable location

- More suitable ads (E.g. specific model vs. generic "Buy appliances here")

Budget

Cannot show the same ad all the time

Economic considerations (bidding, etc)

Varying quality
Spam v. Ham
Quality of landing page

More suitable ads :
(E.g. specific model vs. generic “Buy appliances here”)

Which exact match to show?

Budget drain
Cannot show the same ad all the time

Economic considerations (bidding, etc)

Significant portion of the traffic has no bids

- Advertisers need volume
- Search engine needs revenue
- Users need relevance!

Advertisers do not care about bid phrases—they care about conversions = selling products

The need for advanced match

How to cover all the relevant traffic?

From the Search Engine point of view advanced match is much more challenging

An advertisers dilemma

Advertisers can bid on "broad queries" and/or "concept queries"

Suppose your ad is:
"Good prices on Seattle hotels"

Can bid on any query that contains the word Seattle

Problems:

- What about query "Alaska cruises start point"?
- What about "Seattle's Best Coffee Chicago"

Ideally:

- Bid on any query related to Seattle as a travel destination
- We are not there yet . . .

How should we price these broad match queries?

- A separate field of research!
- In the remainder of the lecture, we will discuss several mechanisms for advanced match

implementation approaches

The data base approach (original Overture approach)

Ads are records in a data base
- The bid phrase (BP) is an attribute

Given a query q:
- For advanced match, consider all ads such that BP=q

Ads are documents in an ad corpus

The bid phrase is a meta-datum

implementation approaches

The IR approach (the modern view)

On query q:

- Run q against the ad corpus
- Have a suitable ranking function
- BP = q (exact match) has high weight

No distinction between advanced match and exact match

Ad Retrieval:
Consider the whole ad corpus and select a set of most viable candidates (e.g. 100)

Ad Reordering:
Re-score the candidates using a more elaborate scoring function to produce the final ordering

Ad retrieval: two phases

Ad Retrieval:
Considers a larger set of ads, using only a subset of available information
Might have a different objective function (e.g. relevance) than the final function

Ad Reordering:
Limited set of ads with more data and more complex calculations
Must use the bid in addition to the retrieval score (e.g. revenue as criteria for the ordering, implement the marketplace design)

Note that this is all part of the advertiser utility

Items 1-100 of 43,479

Rank	Horse Name	Sts	1st	2nd	3rd	Total $	Per Start $	Win%	Top3	Top3%	E
1	Gun Runner	1	1	0	0	$7,000,000	$7,000,000	100%	1	100%	129
2	Justify	6	6	0	0	$3,798,000	$633,000	100%	6	100%	110
3	Good Magic	6	2	1	1	$1,728,400	$288,067	33%	4	67%	109
4	West Coast	1	0	1	0	$1,600,000	$1,600,000	0%	1	100%	125
5	Catholic Boy	5	3	1	0	$1,528,000	$305,600	60%	4	80%	108
6	Accelerate	5	4	1	0	$1,525,000	$305,000	80%	5	100%	125
7	Monomoy Girl	5	5	0	0	$1,524,200	$304,840	100%	5	100%	114
8	Gunnevera	2	1	0	1	$1,324,600	$662,300	50%	2	100%	110

reactive v. predictive

Follow "Catholic Boy"
See how it did in races
Predict the performance

Make a model of a horse:
weight, jockey weight, leg length
Find the importance of each feature in predicting a win/position
Predict performance of unseen (and seen) horses based on the importance of these features

When we have enough information for a given horse use it (reactive), otherwise use model (predictive)

All advanced match methods aim to maximize some objective

- Ad-query match
- query-rewrite similarity

What is the unit of reasoning?

- single ad or campaign?

reactive v. predictive

Individual queries / ads:

- Can we try all the possible combinations enough times and conclude? We might for common queries and ads
- Recommender system type of reasoning (query q is similar to query q')

Features of the queries and ads: words, classes, etc.

- Generalize from the ads in another space
- Predict performance of unseen ads and queries

Hybrid approaches:

- What if we aggregate CTR (Click-through-rate) at campaign level?
- If we have two predictions, how to combine?

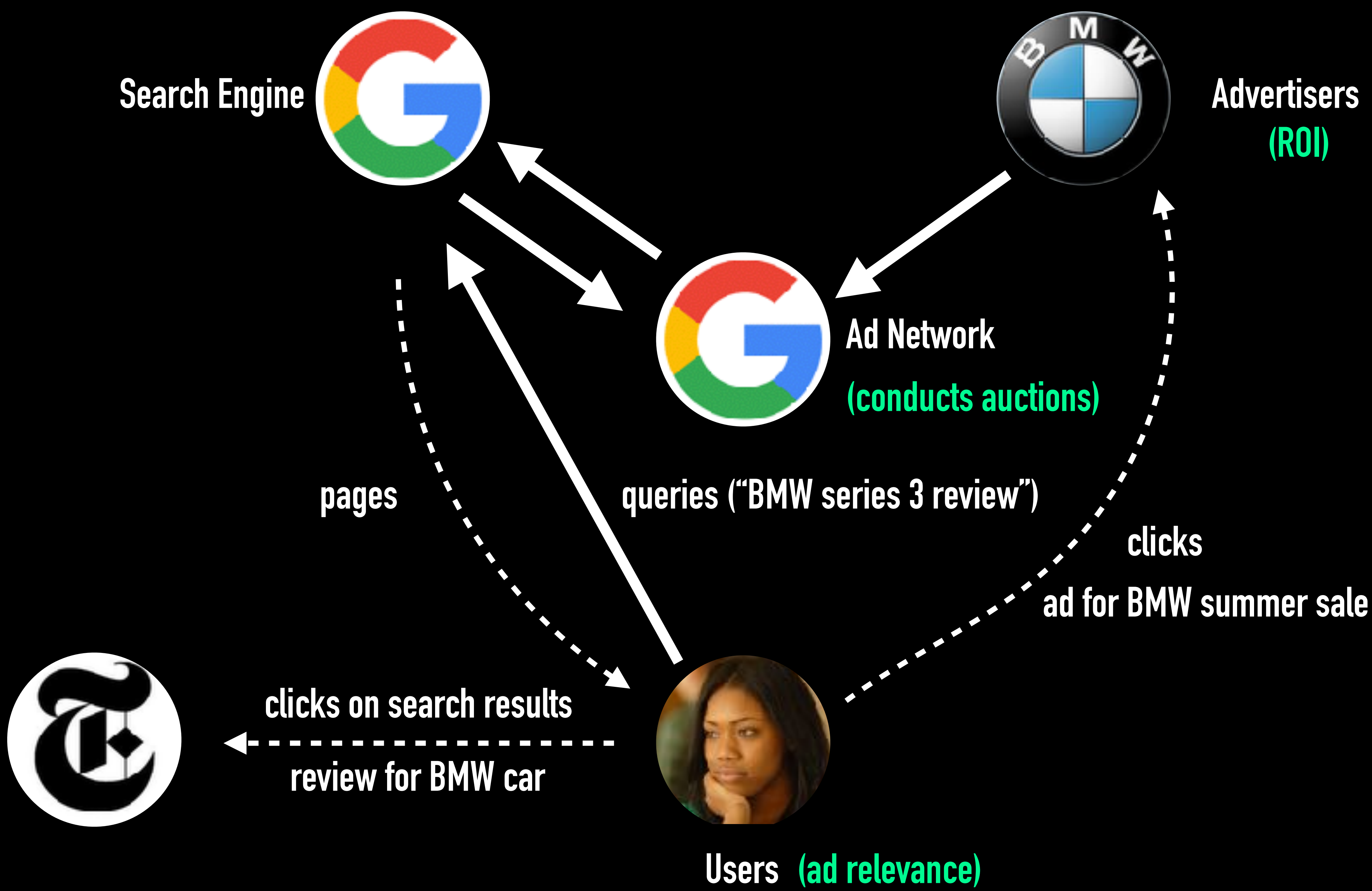
Relevance data:

- Limited editorial resources
- Editors require precise instruction of relevance
- How to deal with multiple dimensions?
- Editors cannot understand every domain and every user need

indications of success

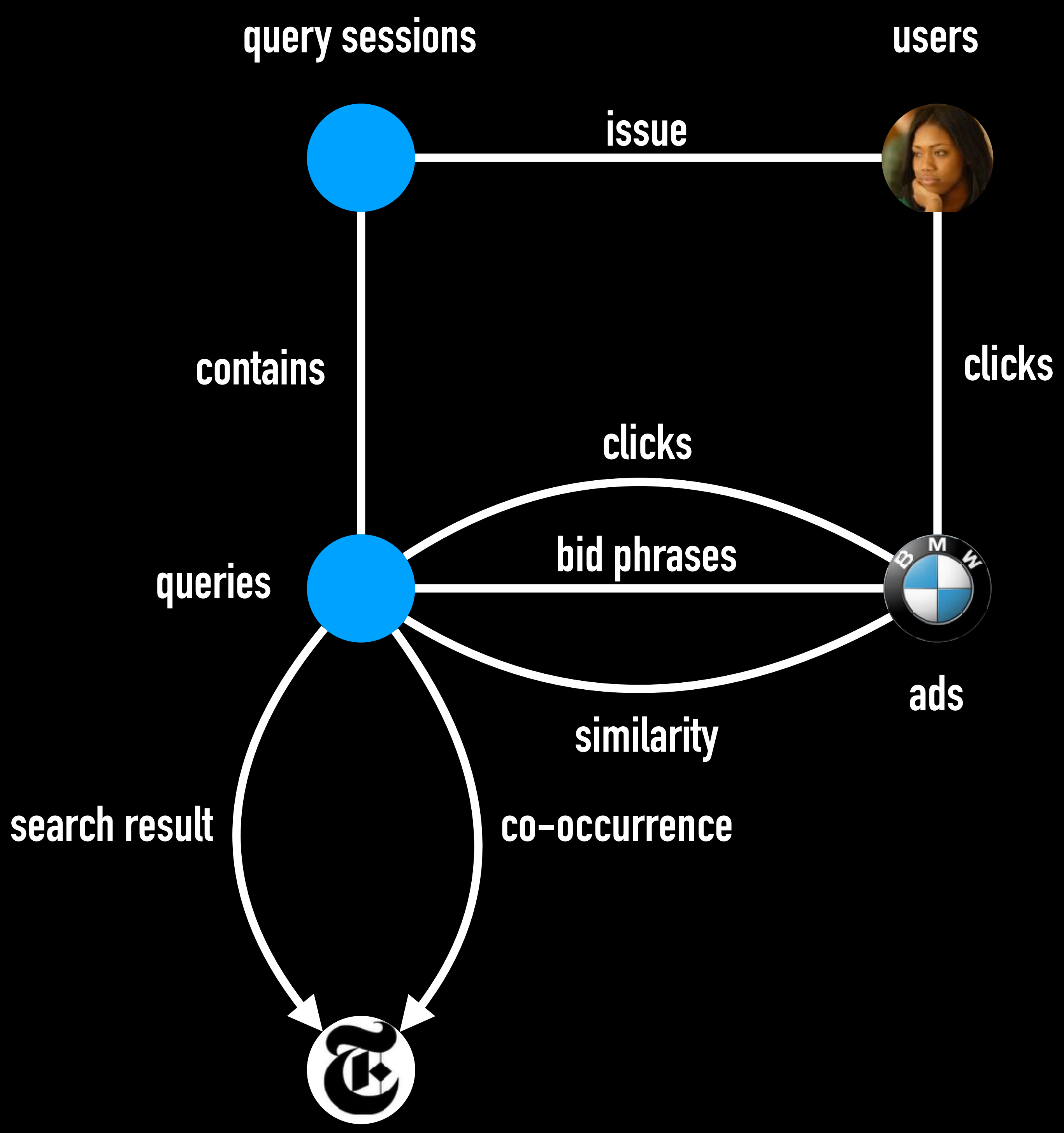
Click data:

- Higher volume—might need sampling
- Binary (click/no click)
- Click-through-rate (CTR) usually very low (1-2%)
- People do not click on ads even when they are relevant
- Much more noise



Deconstructing the Search process

data flow

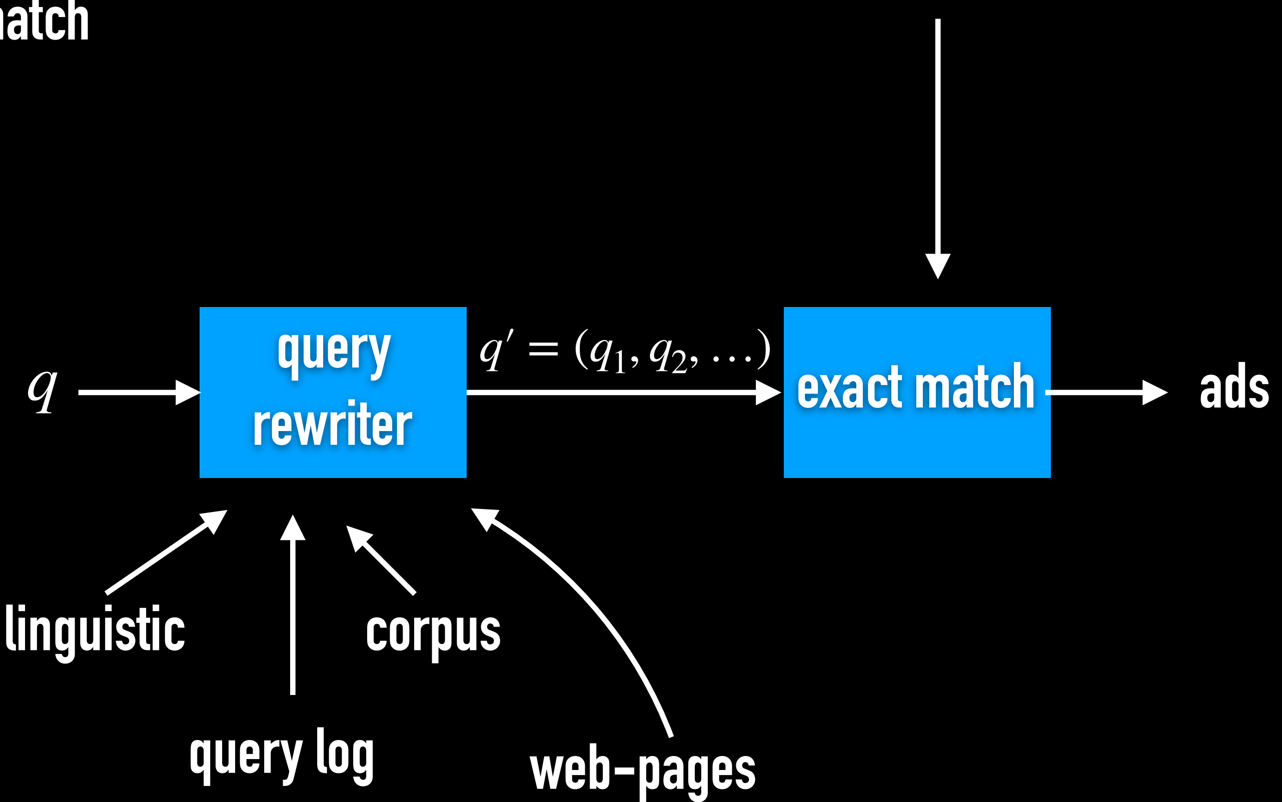

query sessions
users
issue
contains
clicks
clicks
bid phrases
queries
ads
similarity
search result
co-occurrence

Query re-writing for sponsored search

typical query rewriting flow

Typical of the
database approach
to advanced match

Use exact match to select ads for q'



Fits well in the current system architectures

Tolerance value of precision vs. volume differs among advertisers

Additional issue: what to charge the advertiser for advanced match?

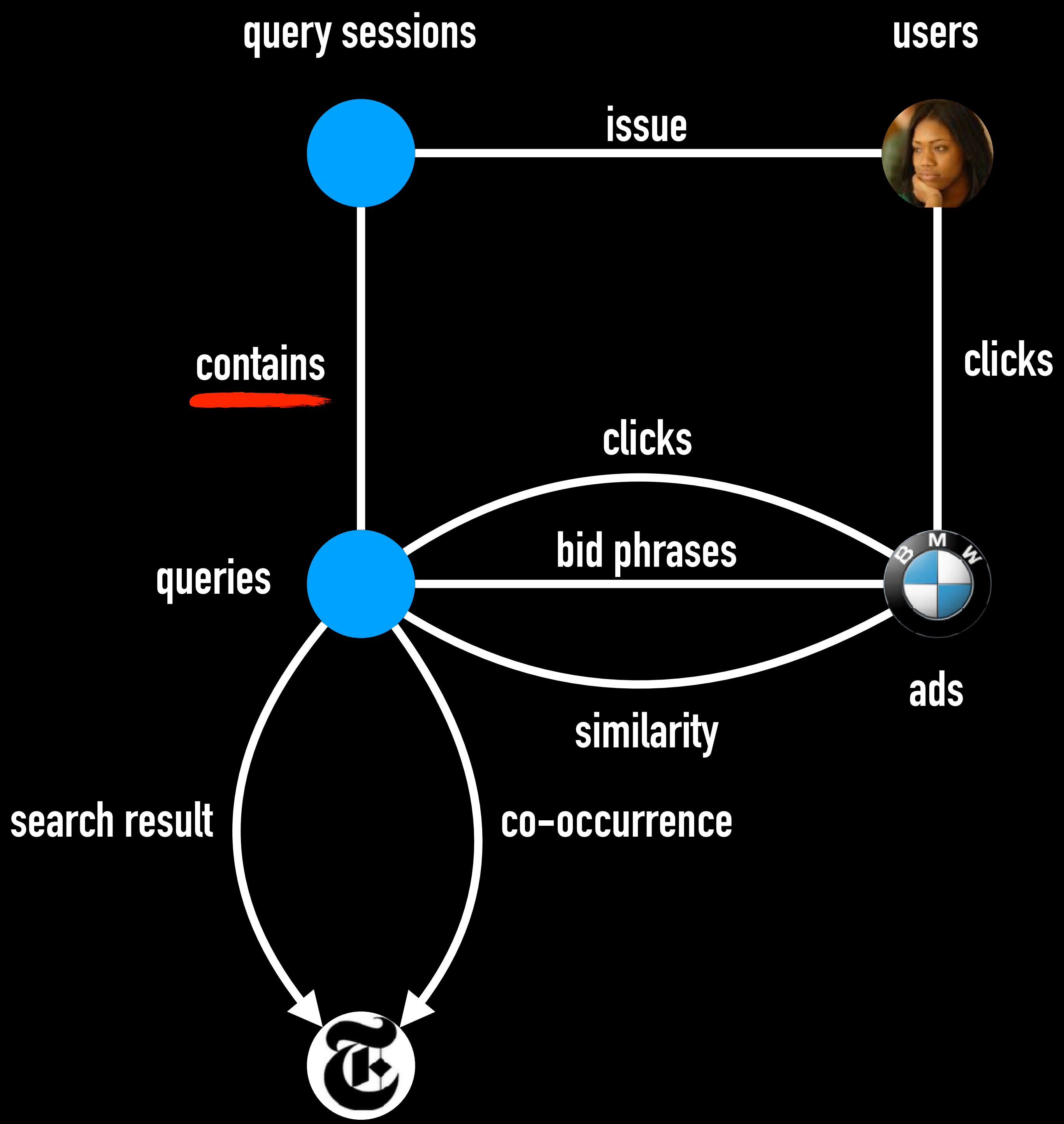
guessing extended keywords on behalf of the advertiser poses risks

Semi-automatic approach:
- Propose rewrites to advertisers
- Let them chose which ones are acceptable Advertiser determines the bid

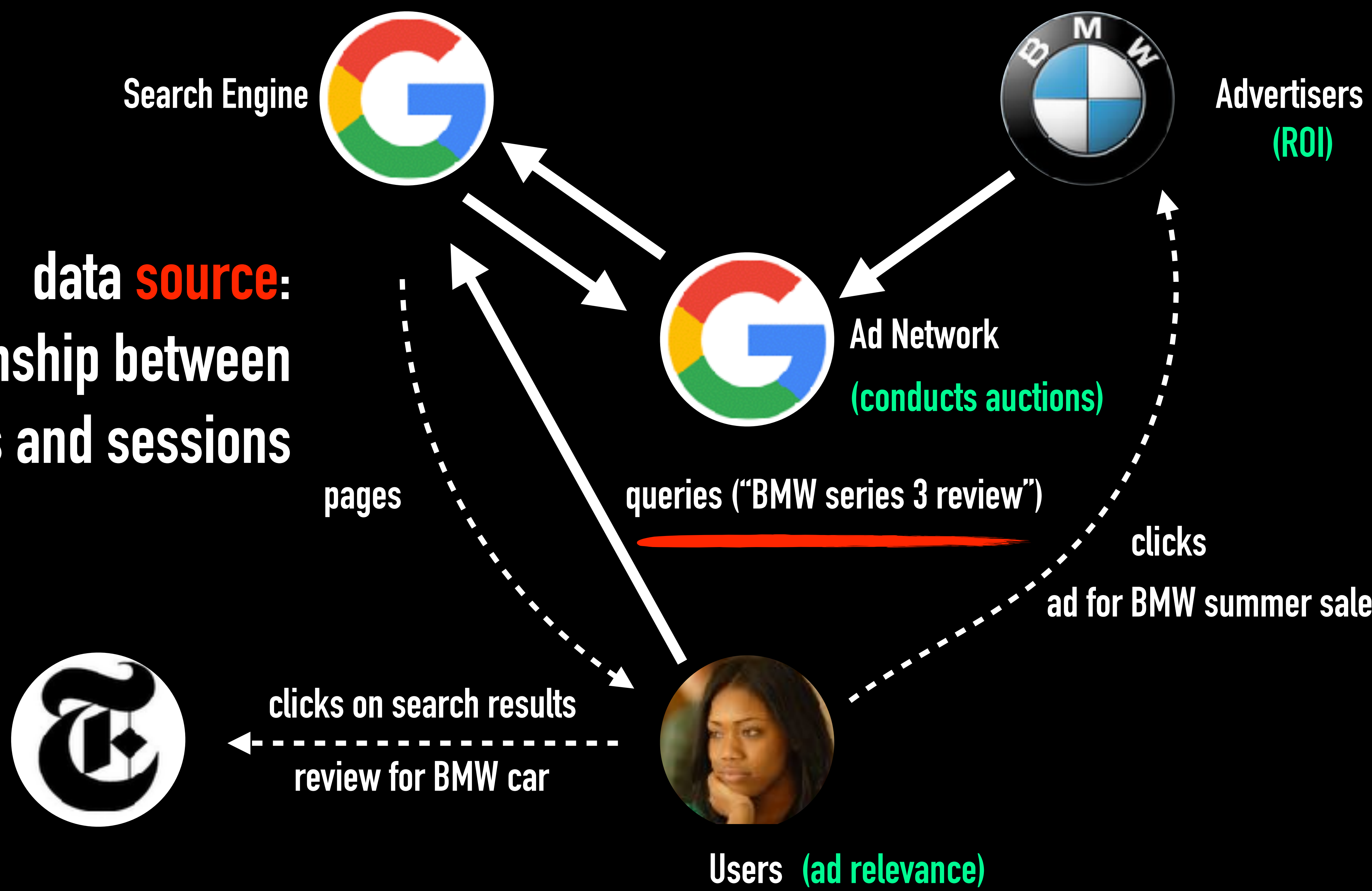
re-writing can be online or offline

re-writing using web search logs

data source


query sessions
users
issue
contains
clicks
clicks
bid phrases
queries
ads
similarity
search result
co-occurrence

data **source**:
relationship between
queries and sessions



Deconstructing the Search process

user sessions

Task completion will usually take several steps:

- Initiating queries
- Browsing

For query rewriting we can focus on the query stream

Finding the session boundaries—research problem

- Time period (all queries within 24hrs)
- Machine learned approach based on query similarity or labeled set

queries ("BMW series 3 review")

How to identify queries that are suitable for rewriting?

- Examine the different types of rewrites that the users do
- Get enough instances of the rewrite to be able to determine its value

half the query pairs are reformulations

Type	Example	Share
switch tasks	mic amps → create taxi	53.2%
insertions	game codes → video game codes	9.1%
substitutions	john wayne bust → john wayne statue	8.7%
deletions	skateboarding pics → skateboarding	5.0%
spell correction	real eastate → real estate	7.0%
mixture	huston's restaurant → houston's	6.2%
specialization	jobs → marine employment	4.6%
generalization	gm reabtes → show me all the current auto rebates	3.2%
other	thansgiving → dia de acconde gracias	2.4%

[Jones and Fain, SIGIR2003]

We see repeated substitutions

some substitutions are incidental

other substitutions repeat over different users over different days

Name	Substituition	Number
car insurance	auto insurance	5086
car insurance	car insurance quotes	4826
car insurance	geico	2613
car insurance	progressive auto insurance	1677
car insurance	carinsurance	428

how can we be sure that the rewrite is any good?

A principled way

determine if:

$$P(R_w \mid q) \gg P(R_w)$$

$$P(R_w \mid q) = \frac{P(R_w, q)}{P(q)}$$

notice

how to measure?
use ML estimation (frequencies)
assume a distribution (e.g. binomial)

$$H_0 : P(R_w \mid q) = P(R_w \mid \bar{q})$$
$$H_1 : P(R_w \mid q) \neq P(R_w \mid \bar{q})$$

The log likelihood ratio is χ2 distributed

query logs: summary

Use the knowledge of the users to generate rewrites

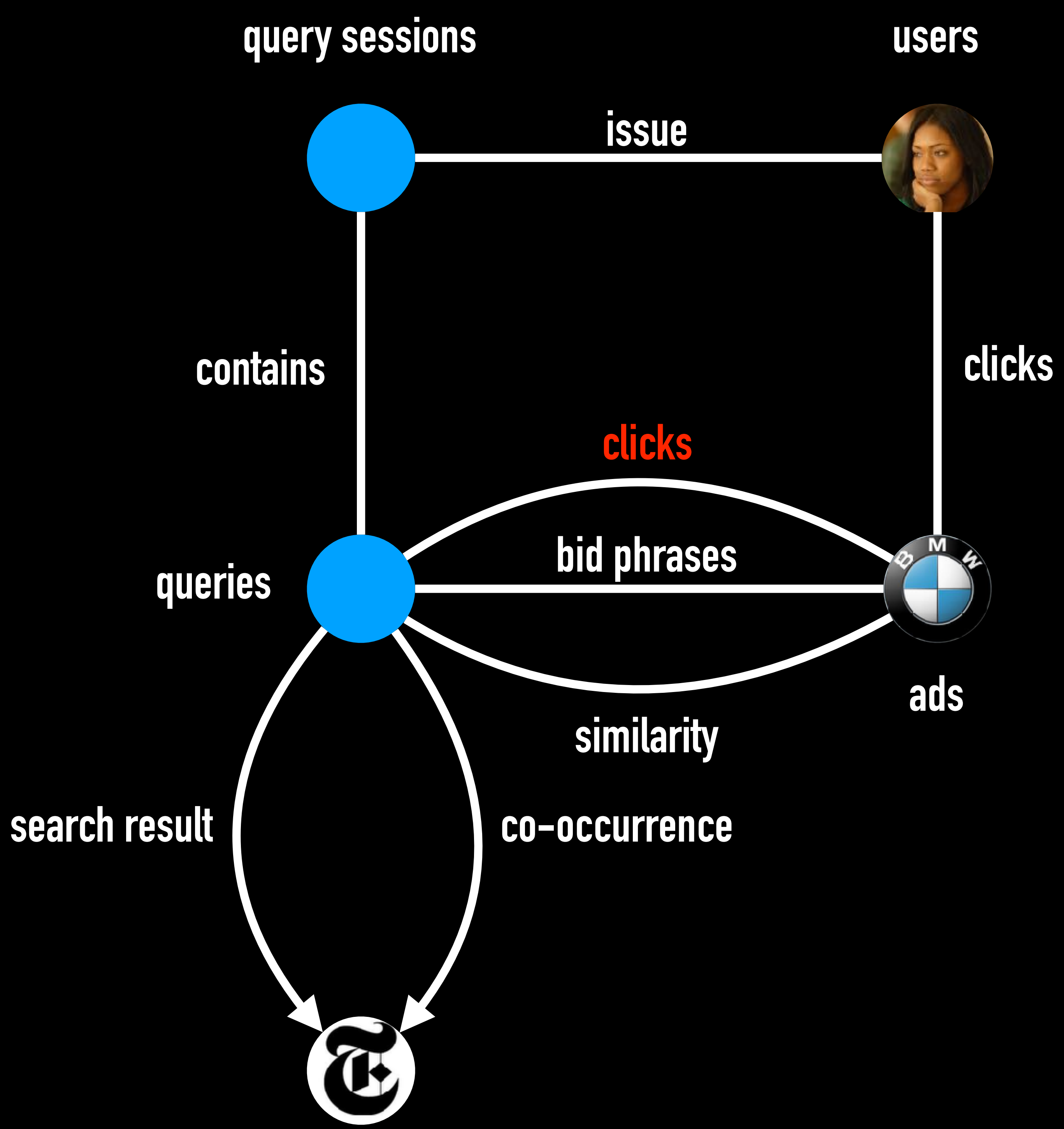
Practical and useful approach, however a few tough challenges:

- Sessions boundaries
- Type of the rewrites
- Requires relatively high frequency of rewrites to be detected

Clicks graphs and random walks for query rewrite generation

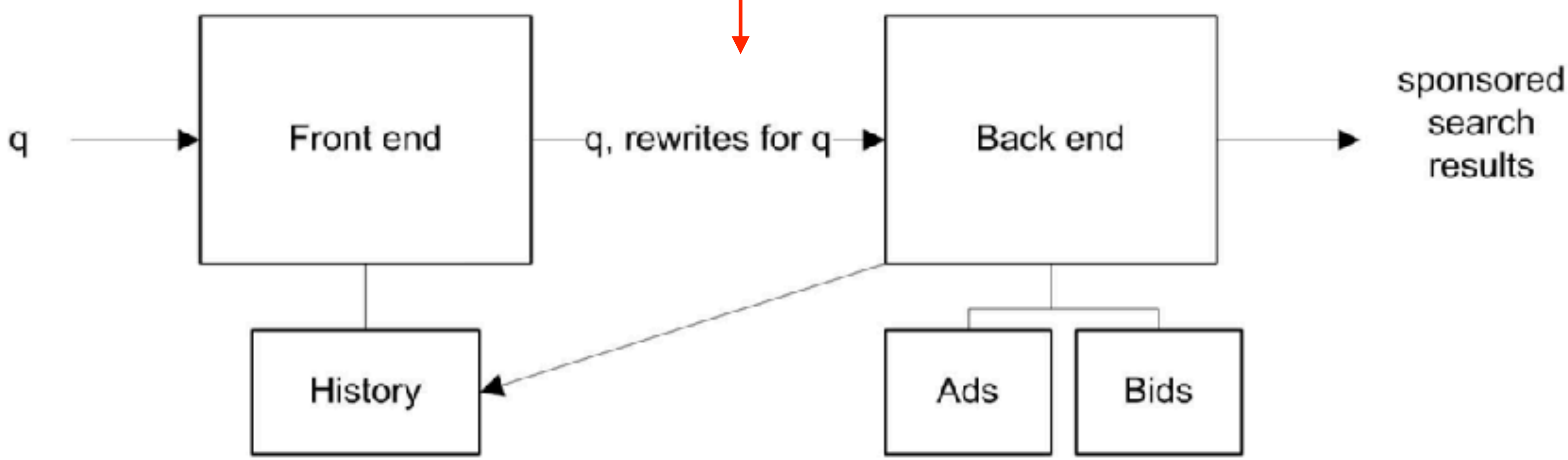
Ioannis Antonellis, Hector Garcia Molina, and Chi Chao Chang. 2008. Simrank++: query rewriting through link analysis of the click graph. Proc. VLDB Endow. 1, 1 (August 2008), 408-421. DOI=http://dx.doi.org/10.14778/1453856.1453903

data source: clicks



A common sponsored search architecture

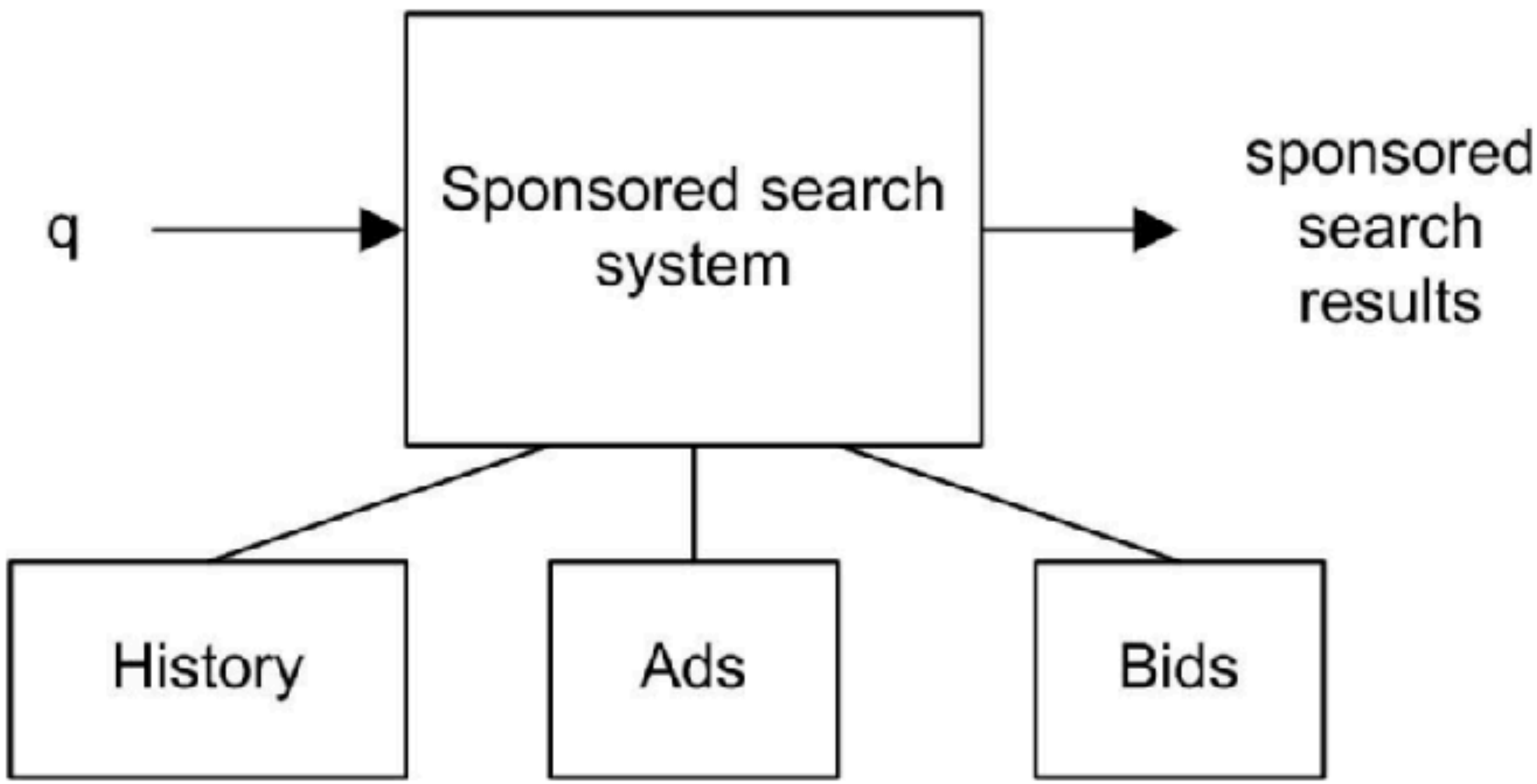
query rewrites are common



Two problems:

- queries have a few words (one or two words); the documents (i.e. ads) are short
- not enough bid queries

A general sponsored search architecture



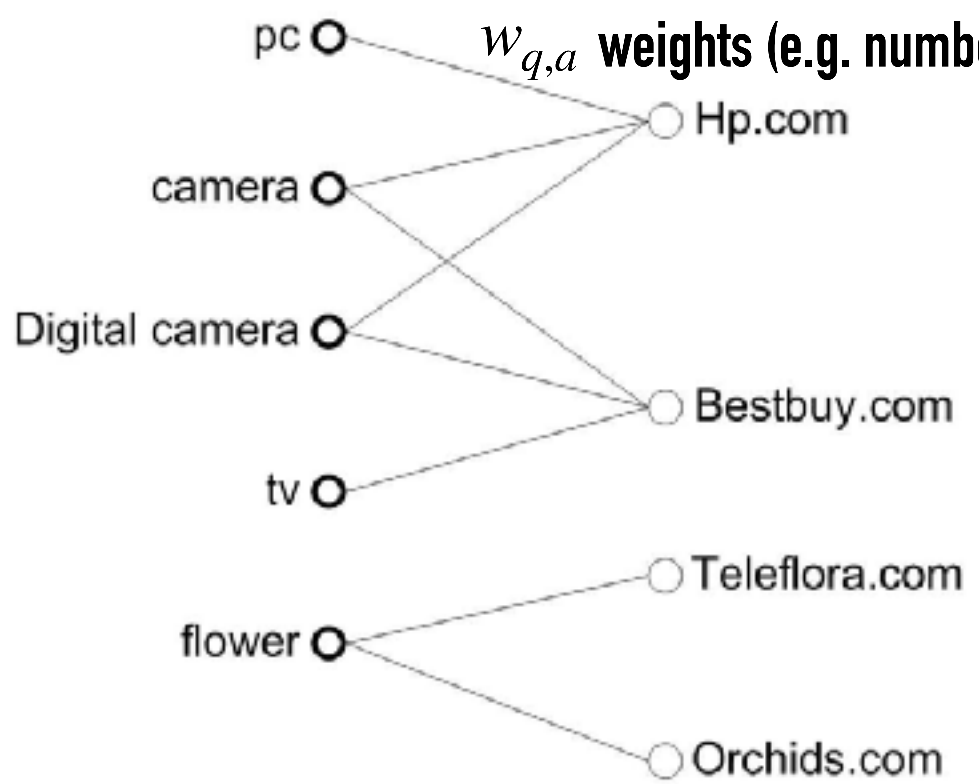
problem definition

A bipartite graph

$G = (V, E, W)$

queries V_q ads V_a

$w_{q,a}$ weights (e.g. number of clicks on ad given query)



The goal: identify how similar two queries (q,q') are

weights

Un-weighted: there is an edge for each ad query pair where there is at least on click

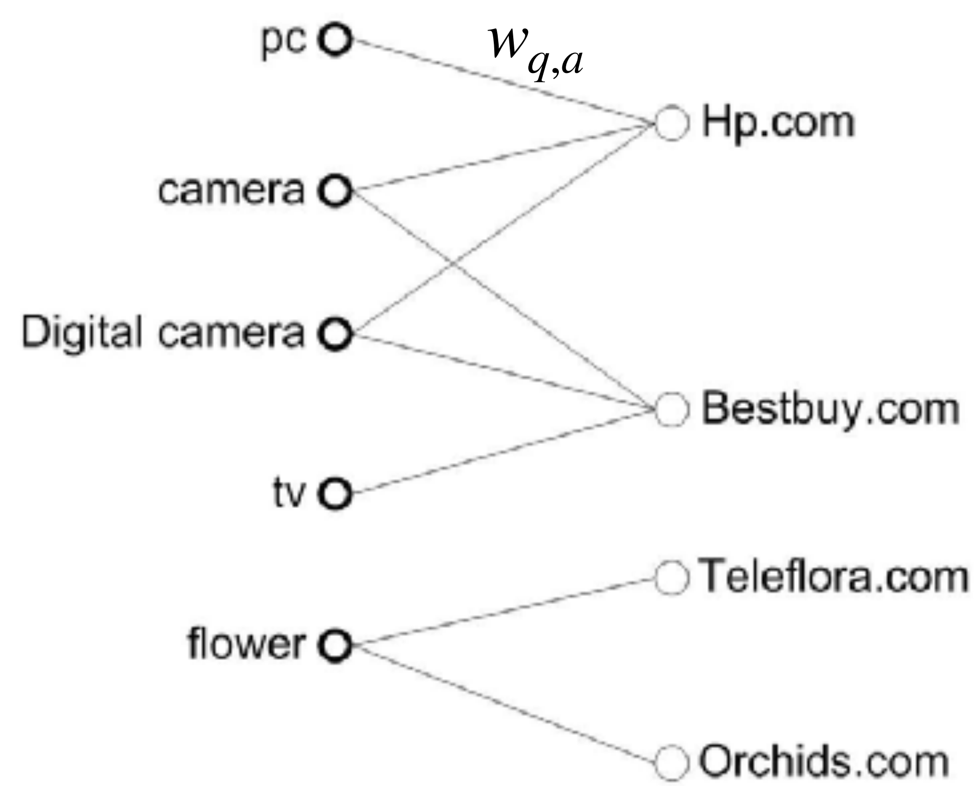
- Issue—some ads get a lot more clicks than others for the same query

Clicks: weight the edges with the number of clicks on the (q,a) combination

- Pairs with higher number of impressions get more clicks even if the relationship is not as strong

CTR: keep the ratio between the clicks and impressions

- CTR of 0.5 differs in confidence when we have one or 10k impressions



Positional Bias

Ads shown on position 1 are more likely to get clicks even if they are less relevant

How does this impact the training in our click-based weighting system?

If the clicks of an ad are all at position 1

- Are those clicks because the ad was relevant?
- Or are those clicks caused by the inherent bias of the user to click the top ad?

We need a way to "de-bias" click data, separating the effects of position with ad relevance

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08). ACM, New York, NY, USA, 87-94. DOI=http://dx.doi.org/10.1145/1341531.1341545

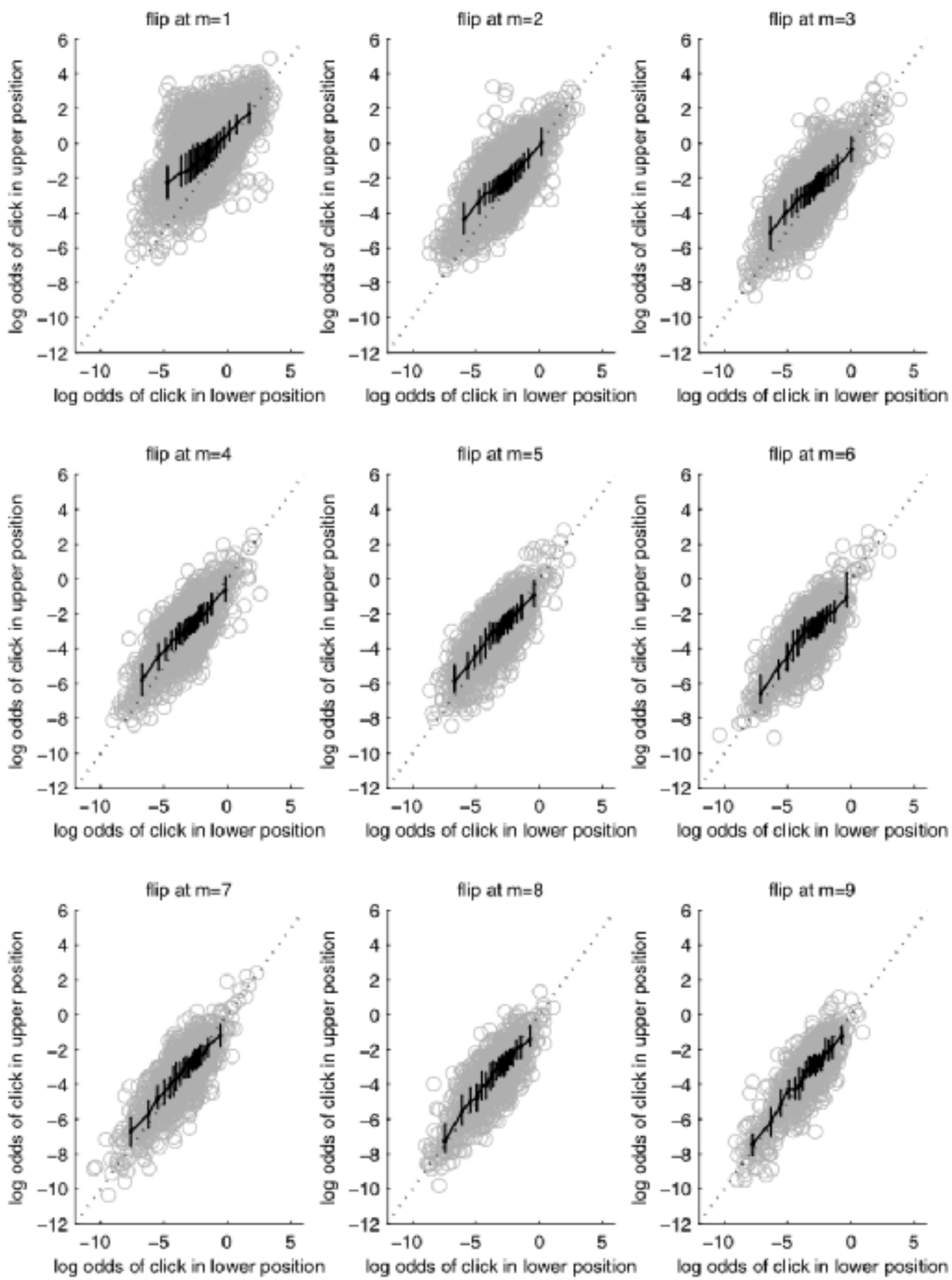
The cascade model

"In the cascade model, we assume that the user views search results from top to bottom, deciding whether to click each result before moving to the next. Each document d, is either clicked with probability r_d or skipped with probability $(1-r_d)$. In the most basic form of the model, we assume that a user who clicks never comes back, and a user who skips always continues, in which case:"

$$c_{di} = r_d \prod_{j=1}^{i-1} (1 - r_{d,j})$$

clicked ad at position i

skipped earlier ads



flips at different positions

cascade is better than baselines at predicting click through dates

improvements: mostly on assumptions on if priors were clicked; more sophisticated Bayesian models

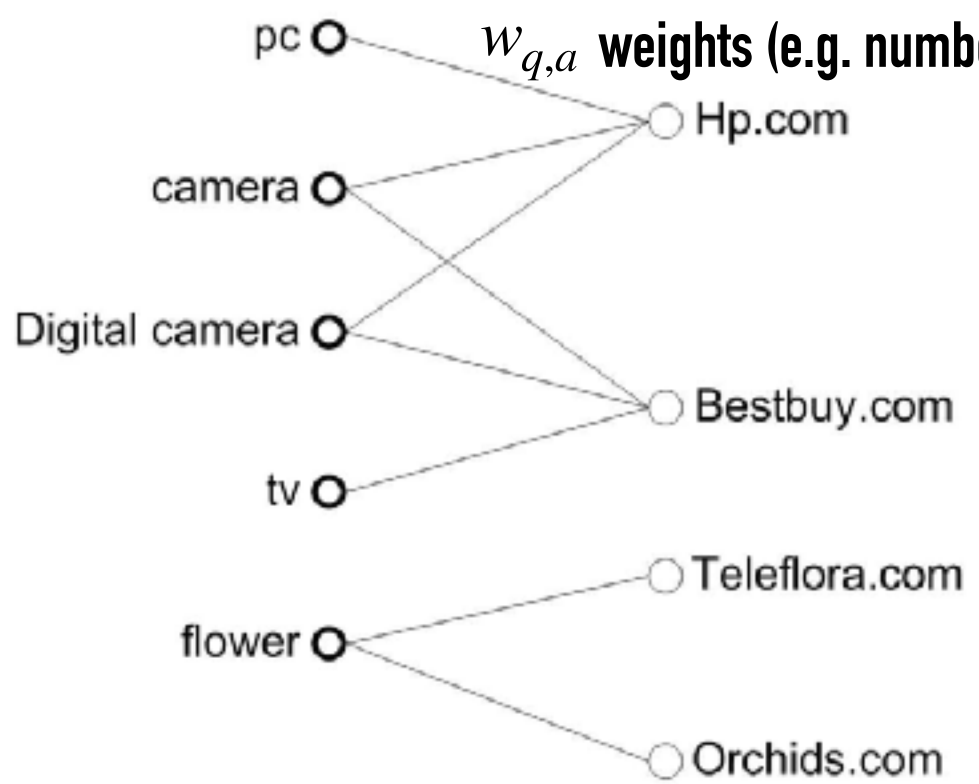
determine query similarity

A bipartite graph

$G = (V, E, W)$

queries V_q ads V_a

$w_{q,a}$ weights (e.g. number of clicks on ad given query)

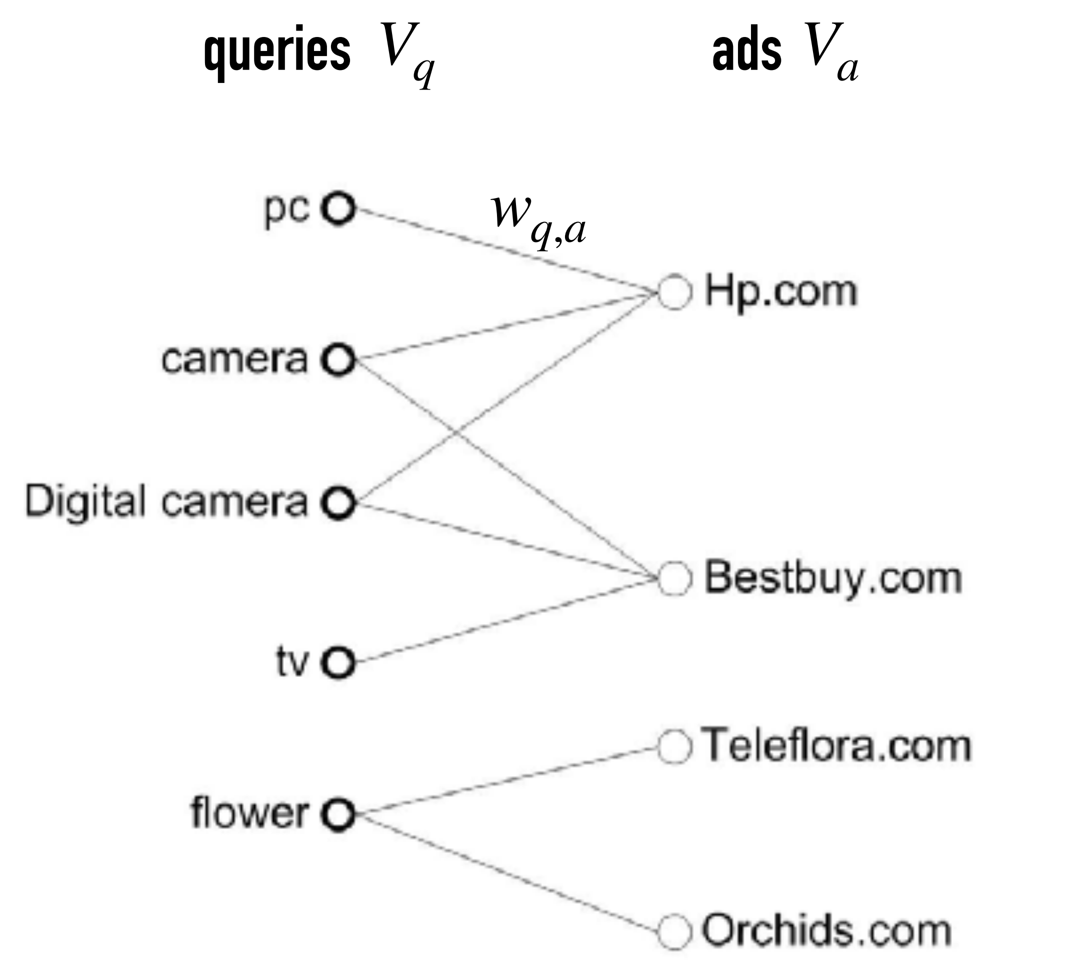


The goal: identify how similar two queries (q,q') are

Basic similarity

Table 1: Query-query similarity scores for the sample click graph of Figure 3. Scores have been computed by counting the common ads between the queries

	pc	camera	digital camera	tv	flower
pc	-	1	1	0	0
camera	1	-	2	1	0
digital camera	1	2	-	1	0
tv	0	1	1	-	0
flower	0	0	0	0	-



Simrank

"Two queries are similar if they are connected to similar ads"
"Two ads are similar if they are connected to similar queries"

Assume similarity is a measure between 1 and 0 (like probability); A query is "very" similar to itself: sim(q,q) = 1

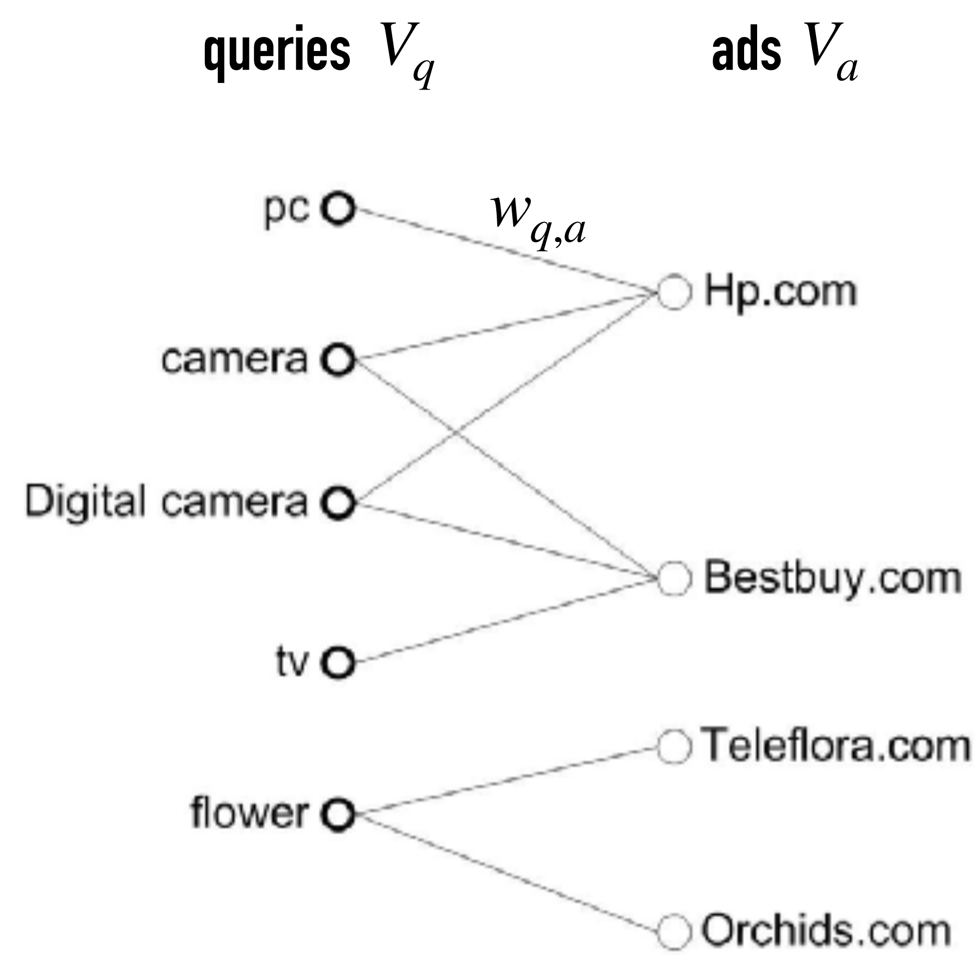
Initially, we know nothing about the similarity with other queries:

sim (q,q') = 0 iff q ≠ q'

Establish similarity of two queries based on the ads they connect to (Random walk starting at q and q' simultaneously – end up in the same node)

Simultaneously do the same thing on the ad side

Iterative procedure: at each iteration similarity propagates through the the graph



Simrank

#neighbors

set of neighbors

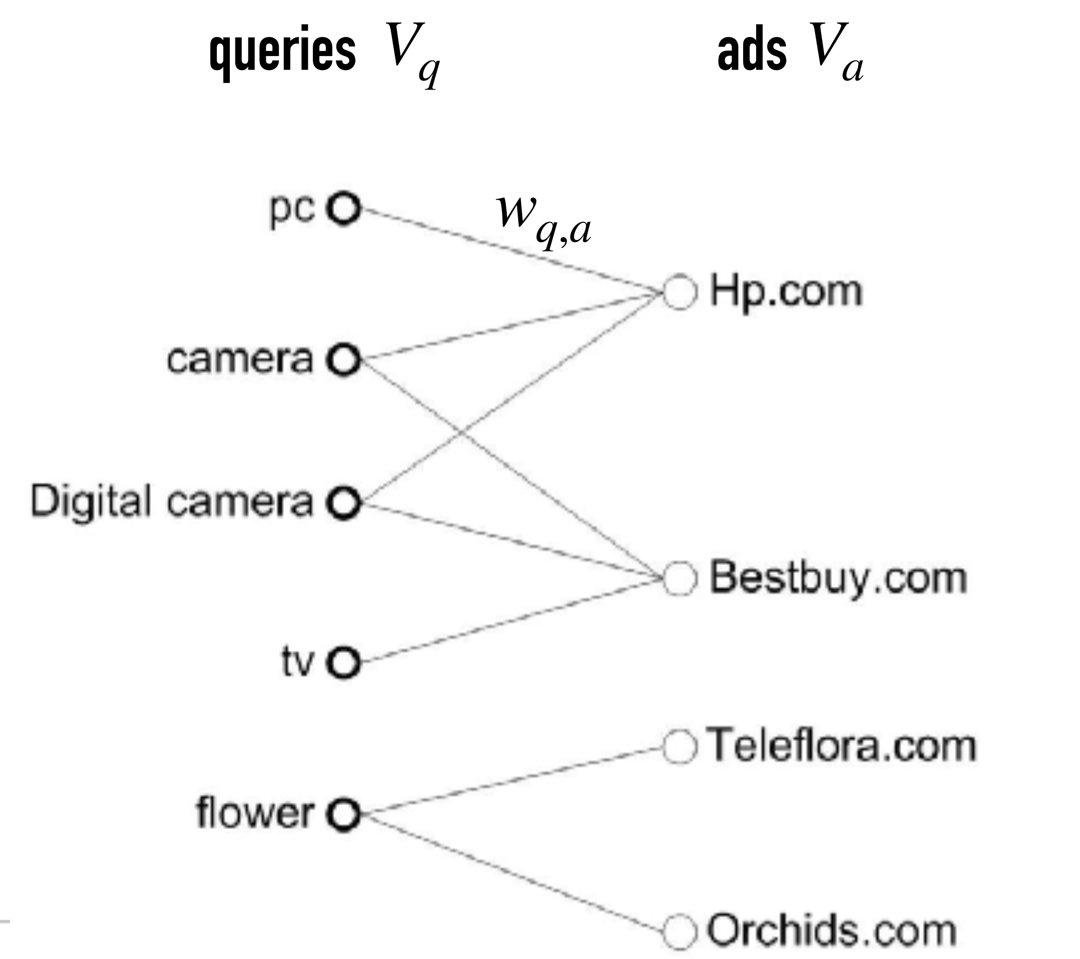
Let $s(q, q')$ denote the similarity between queries q and q', and let $s(\alpha, \alpha')$ denote the similarity between ads α and α'. For $q \neq q'$, we write the equation:

$$s(q, q') = \frac{C_1}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s(i, j) \qquad (1)$$

where C_1 is a constant between 0 and 1. For $\alpha \neq \alpha'$, we write:

$$s(\alpha, \alpha') = \frac{C_2}{N(\alpha)N(\alpha')} \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} s(i, j) \qquad (2)$$

where again C_2 is a constant between 0 and 1.



A simultaneous solution exists and is unique.

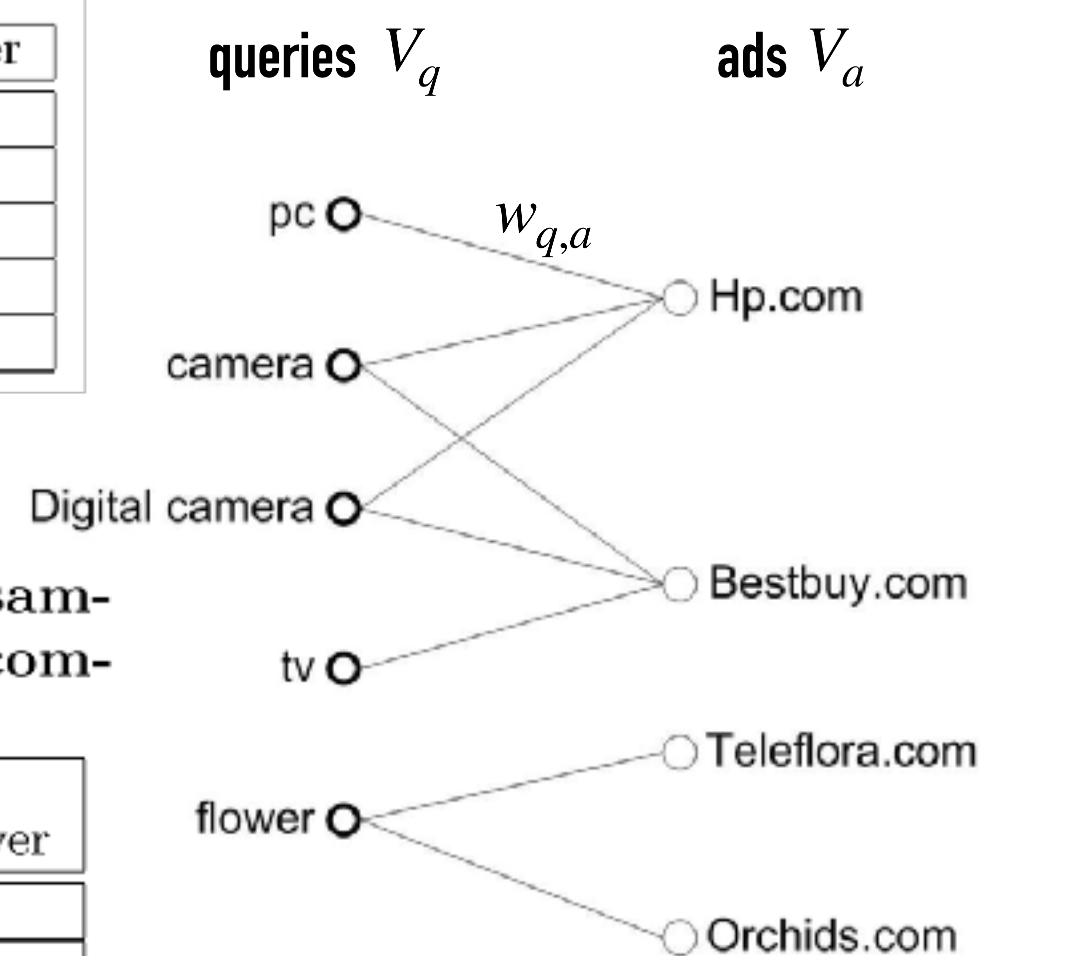
Simrank

Table 1: Query-query similarity scores for the sample click graph of Figure 3. Scores have been computed by counting the common ads between the queries

	pc	camera	digital camera	tv	flower
pc	-	1	1	0	0
camera	1	-	2	1	0
digital camera	1	2	-	1	0
tv	0	1	1	-	0
flower	0	0	0	0	-

Table 2: Query-query similarity scores for the sample click graph of Figure 3. Scores have been computed by Simrank with $C_1 = C_2 = 0.8$

	pc	camera	digital camera	tv	flower
pc	-	0.619	0.619	0.437	0
camera	0.619	-	0.619	0.619	0
digital camera	0.619	0.619	-	0.619	0
tv	0.437	0.619	0.619	-	0
flower	0	0	0	0	-



Simrank: challenges

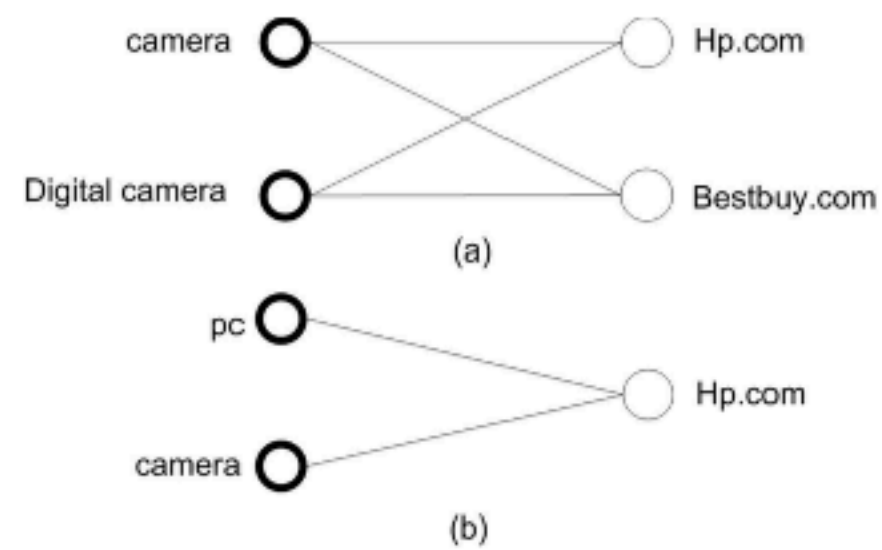


Figure 4: Sample complete bipartite graphs ($K_{2,2}$ and $K_{1,2}$) extracted from a click graph.

Table 3: Query-query similarity scores for the sample click graphs of Figure 4. Scores have been computed by Simrank with $C_1 = C_2 = 0.8$

Iteration	sim("camera", "digital camera")	sim("pc", "camera")
1	0.4	0.8
2	0.56	0.8
3	0.624	0.8
4	0.6496	0.8
5	0.65984	0.8
6	0.663936	0.8
7	0.6655744	0.8

initially, these numbers are different, but converge to the same value when n→∞

Emphasize neighbors

$$\text{evidence}(a, b) = \sum_{i=1}^{|E(a) \bigcap E(b)|} \frac{1}{2^i}$$

"The intuition behind choosing such a function is as follows. We want the evidence score evidence(a,b) to be an increasing function of the common neighbors between a and b. In addition we want the evidence scores to get closer to one as the common neighbors increase."

$$s_{\text{evidence}}(q, q') = \text{evidence}(q, q') \cdot s(q, q')$$
$$s_{\text{evidence}}(\alpha, \alpha') = \text{evidence}(\alpha, \alpha') \cdot s(\alpha, \alpha')$$

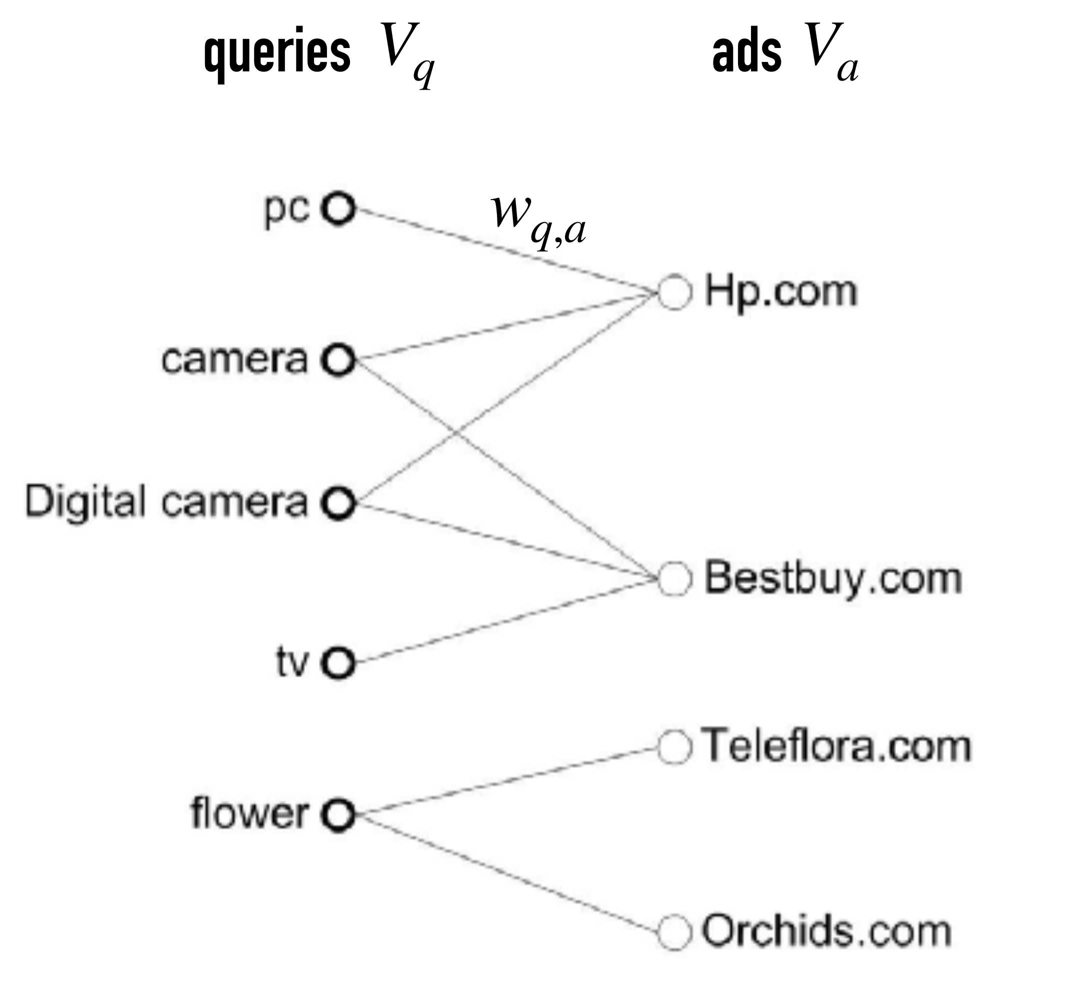


Table 4: Query-query similarity scores for the sample click graphs of Figure 4. Scores have been computed by the evidence-based Simrank with $C_1 = C_2 = 0.8$

Iteration	sim("camera", "digital camera")	sim("pc", "camera")
1	0.3	0.4
2	0.42	0.4
3	0.468	0.4
4	0.4872	0.4
5	0.49488	0.4
6	0.497952	0.4
7	0.4991808	0.4

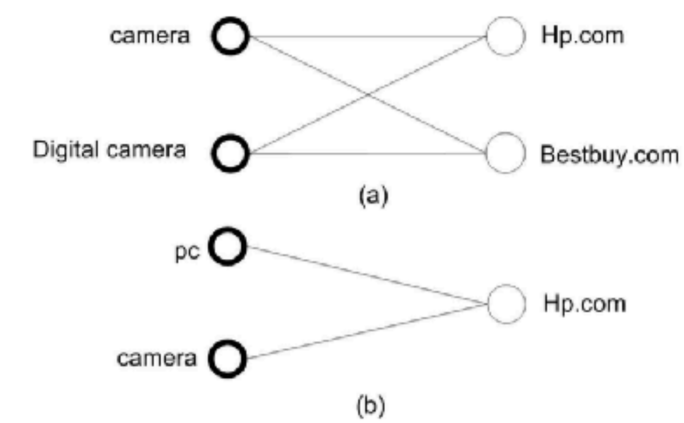


Figure 4: Sample complete bipartite graphs ($K_{2,2}$ and $K_{1,2}$) extracted from a click graph.

Table 3: Query-query similarity scores for the sample click graphs of Figure 4. Scores have been computed by Simrank with $C_1 = C_2 = 0.8$

Iteration	sim("camera", "digital camera")	sim("pc", "camera")
1	0.4	0.8
2	0.56	0.8
3	0.624	0.8
4	0.6496	0.8
5	0.65984	0.8
6	0.663936	0.8
7	0.6655744	0.8

Weighted Simrank

Figure 5: Sample weighted click graphs

Weighted Simrank

$$p(\alpha, i) = \text{spread}(i) \cdot \text{normalized\_weight}(\alpha, i), \forall i \in E(\alpha), \text{ and}$$

$$p(\alpha, \alpha) = 1 - \sum_{i \in E(\alpha)} p(\alpha, i)$$

where:

$$\text{spread}(i) = e^{-\text{variance}(i)}, \text{ and}$$

$$\text{normalized\_weight}(\alpha, i) = \frac{w(\alpha, i)}{\sum_{j \in E(\alpha)} w(\alpha, j)}$$

Weighted Simrank

The actual similarity scores that weighted Simrank gives after applying the modified random walk are:

$$
\begin{aligned}
s_{\text{weighted}}(q, q') &= \text{evidence}(q, q') \cdot C_1 \cdot \\
&\quad \sum_{i \in E(q)} \sum_{j \in E(q')} W(q, i) W(q', j) s_{\text{weighted}}(i, j) \\
s_{\text{weighted}}(\alpha, \alpha') &= \text{evidence}(\alpha, \alpha') \cdot C_2 \cdot \\
&\quad \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} W(\alpha, i) W(\alpha', j) s_{\text{weighted}}(i, j)
\end{aligned}
$$

where the factors $W(q, i)$ and $W(a, i)$ are defined as follows:

$$
\begin{aligned}
W(q, i) &= \text{spread}(i) \cdot \text{normalized\_weight}(q, i) \\
W(\alpha, i) &= \text{spread}(i) \cdot \text{normalized\_weight}(\alpha, i)
\end{aligned}
$$

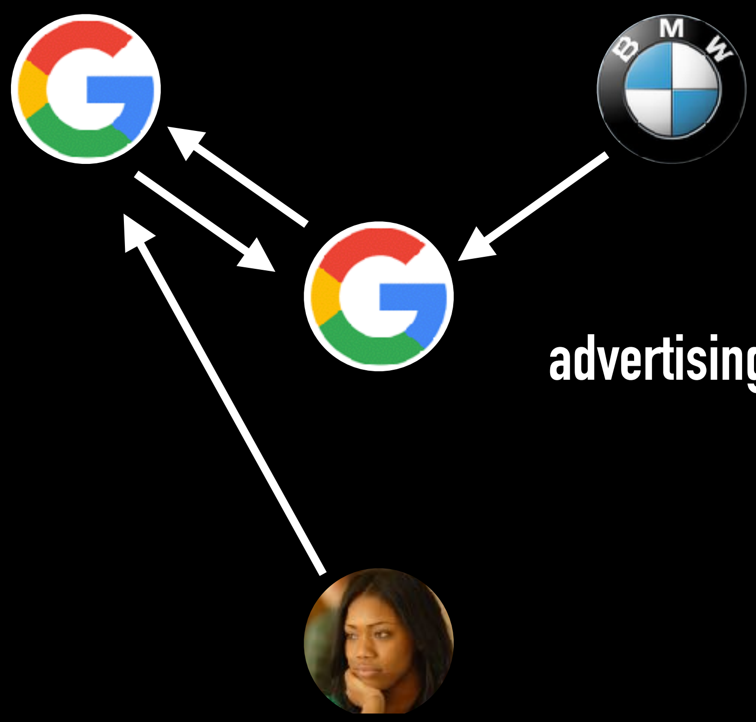
Weighted Simrank

Algorithm 2 Simrank++ Computation

Require: weighted transition matrix P', evidence matrix V, decay factor C, number of iterations k

Ensure: similarity matrix S'

1: $[N,N] = \text{size}(P')$;
2: $S' = I_N$;
3: **for** $i = 1 : k$, **do**
4: $\quad \text{temp} = C * P'^T * S' * P'$;
5: $\quad S' = \text{temp} + I_N - \text{Diag}(\text{diag}(\text{temp}))$;
6: **end for**
7: $S' = V. * S'$;



advertising is a market

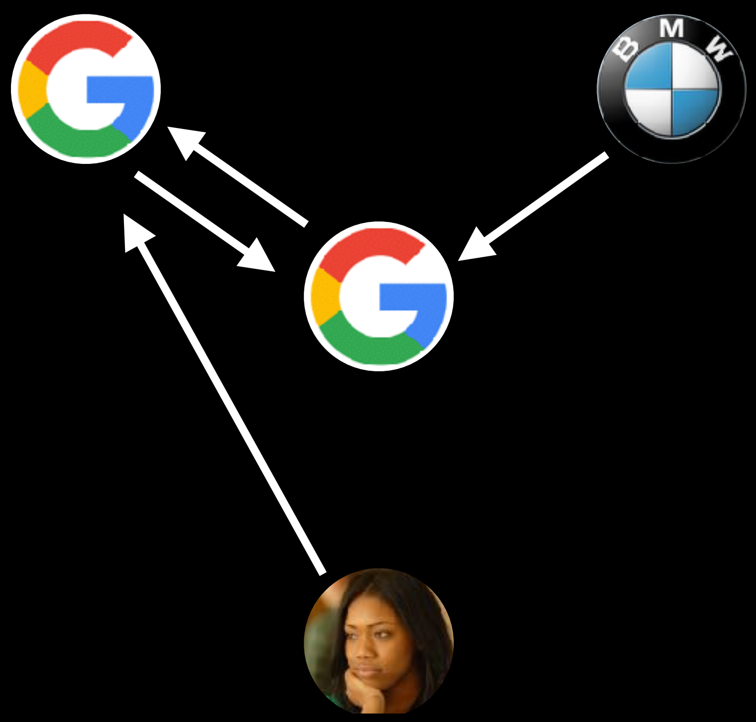
Find the "best match" between a given user in a given context and a suitable advertisement.

Summary

low, click through rates
mobile vs. desktop
dominance of Google / Facebook

Computational Advertising

1. Ad retrieval (match to query/context)
2. Ordering the ads
3. Pricing on a click-through



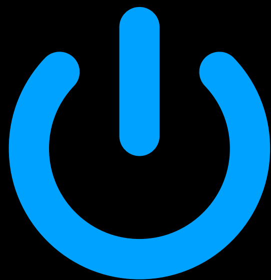
Web queries:
- long tail
- temporal

Finding ads:
- exact match vs. advanced match

Summary

Query re-writing is important
- Using query logs
- position dependent click interaction
- Simrank for query re-writing

Landing page plays a role in conversion



Introduction

Web search

Game Theory

Auctions

Text Ads

Query re-writing → Advanced match

Display Ads

Behavioral targeting

Recommender systems

Privacy

Networks

Emerging areas

Final Presentations