Targeting Converters in Display Advertising

Deepak Agarwal, Yahoo! Research

Joint with Sandeep Pandey, Y! Research

University of Texas, Austin

15th Nov, 2011

Agenda

- Background on Advertising
- Display Advertising
  - Futures Market
  - Spot Market --- Ad-exchange, Real-time bidder (RTB)
- User Targeting Problem in Display Advertising
  - Current practice, state-of-the-art
- Our approach
  - Using historic campaign data to target new campaigns
- Modeling Details
  - Factor Model
- Experiments

The two basic forms of advertising

1. Brand advertising
   – creates a distinct favorable image

2. Direct-marketing
   – Advertising that strives to solicit a "direct response”: buy, subscribe, vote, donate, etc, now or soon

Brand advertising …

Sometimes both Brand and Performance

Web Advertising

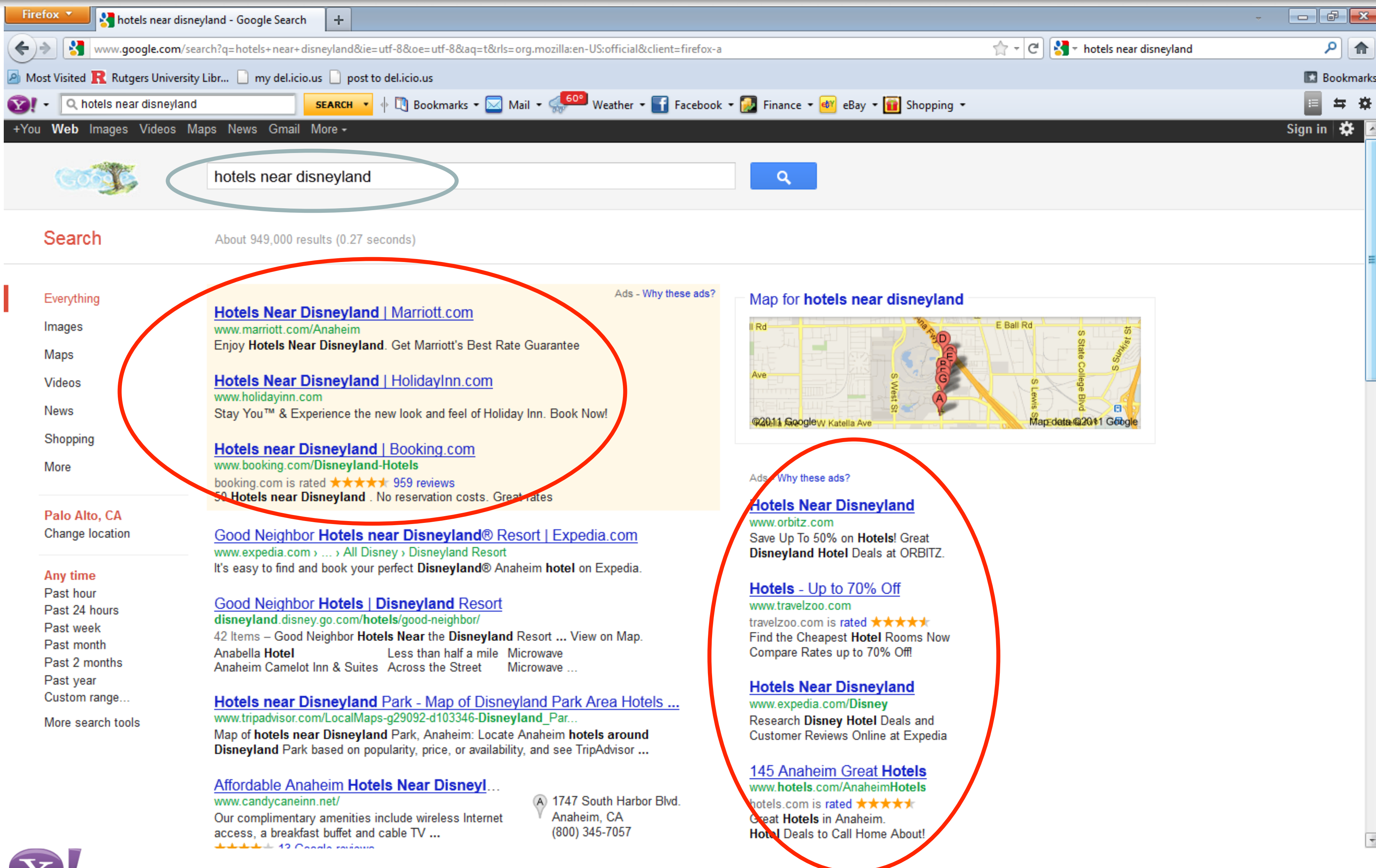
There are lots of ads on the web ….

100s of billions of advertising dollars spent online per year (e-marketer)

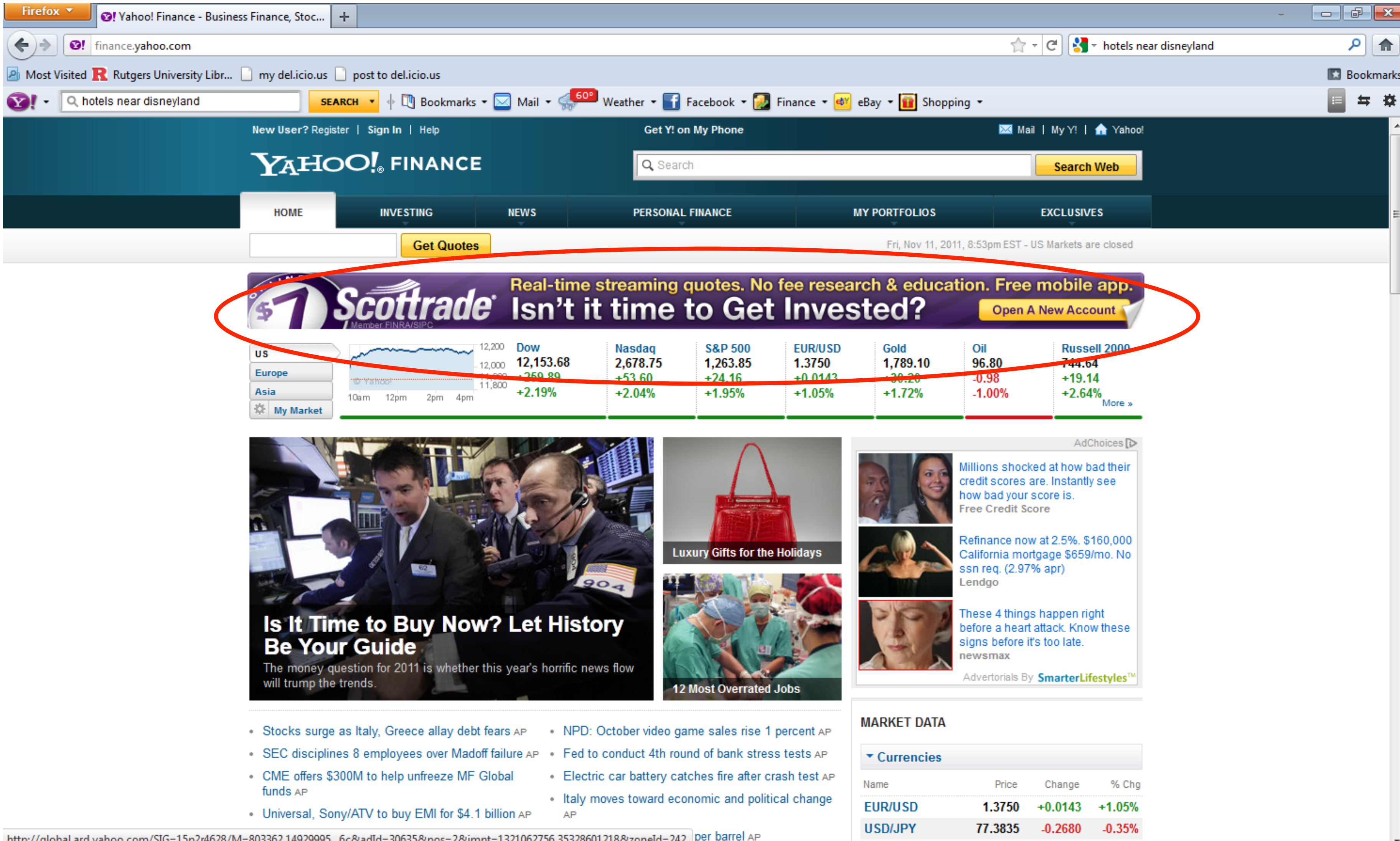
Web Advertising: Comes in different flavors

- Sponsored ("Paid") Search
  - Small text links in response to query to a search engine
- Display Advertising
  - Graphical, banner, rich media; appears in several contexts like when visiting a webpage, checking e-mails, facebook,….
    - Brand Awareness campaign
    - Performance campaign
      - More like direct marketing in offline world, target users for some favorable response in the near-term (buy, vote, etc)

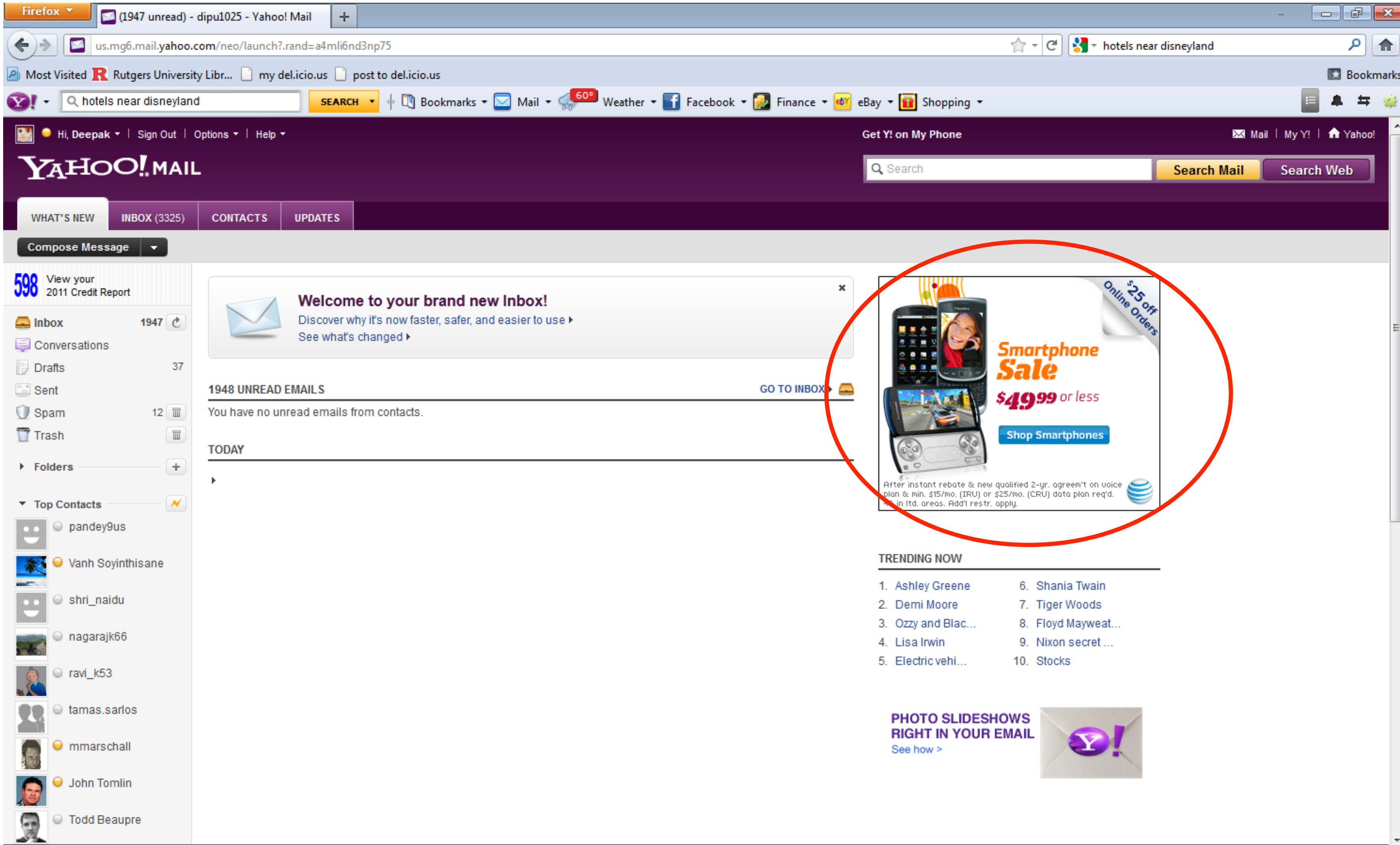
Paid Search: Advertise Text Links



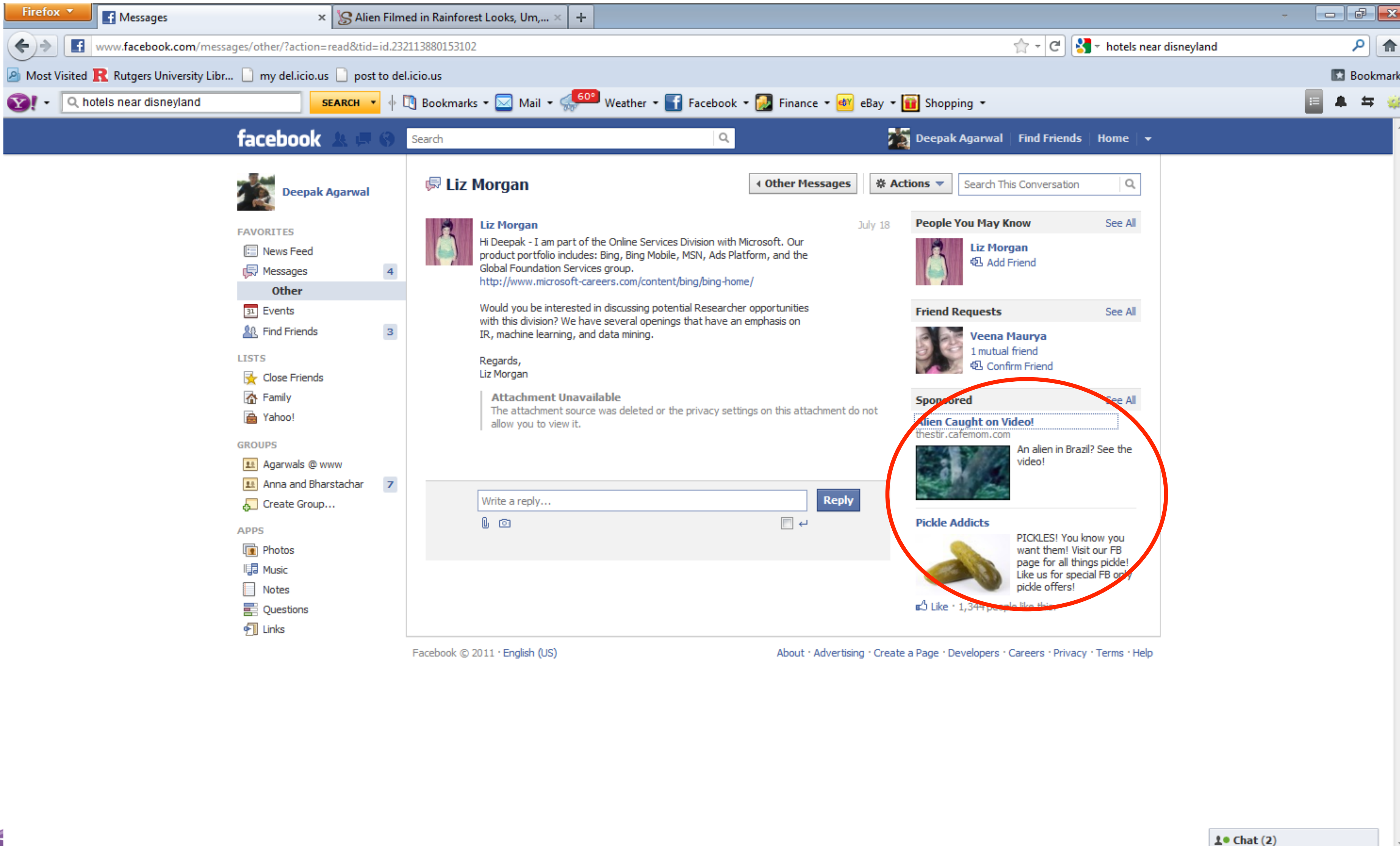
Display Advertising: Examples



Display Advertising: Examples



Display Advertising: Examples



Paid Search versus Display Advertising

Paid Search

- Context (Query) important to match
- Small text links
- Performance based
  - Clicks, conversions
- Advertisers can cherry-pick *instances*

Display

- Reaching desired audience main goal
- Graphical, banner,Rich media
  - Text, logos, videos,..
- Hybrid
  - Brand, performance
- Bulk buy
  - But things evolving
    - Ad exchanges, Real-time bidder (RTB)

Display Advertising Models

- Futures Market (Guaranteed Delivery)
  - Brand Awareness (e.g. Coke, MacDonald's,..)
- Spot Market (Non-guaranteed)
  - Marketers create campaigns targeted to user segments
    - Ad-exchanges have made this process efficient
      - Connects buyers and sellers in a stock-market style market

Guaranteed Delivery (Futures Market)

- Revenue Model: Cost per eyeball (CPM)
- Traditional Advertising Model:

  Ads are bought in bulk targeted to users based on demographics and other behavioral features

  GM ads in Y! autos shown to *“males above 55”*

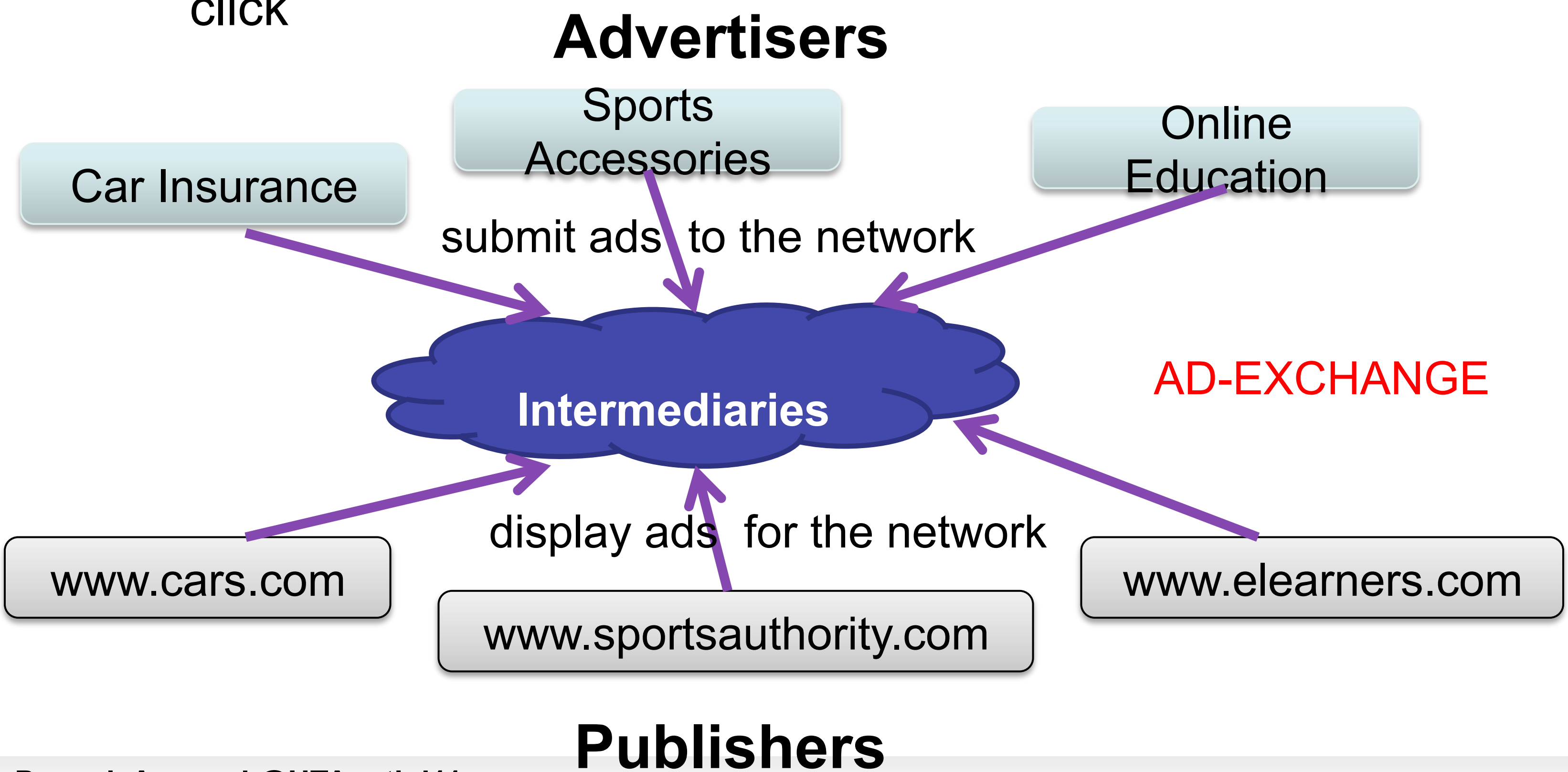
  Mortgage ad shown to *“everybody on Y! Front page”*

  Book a slot well in advance

  – *“2M impressions in Jan next year”*
  – Future impressions must be **guaranteed** by the ad network
  – Prices are significantly higher than in Spot market
    – Publishers ensure higher quality inventory

Spot Market

- Different advertiser objectives
  - Brand awareness to target future sales: CPM
  - Performance based (auto, washing machine, pizza, donation,…)
    - Pay by conversion (favorable action on landing page) or/and click



Advertising Eco-System

- Ad exchange (like a stock-market)
  - Winner decided through an auction process
    - Several intermediaries (ad-networks, pub-networks,…)
  - Still bulk buy, advertisers specify targets and a bid

- New technologies
  - Real-time bidder: change bid dynamically
    - Based on user cookie and other information passed by publishers to the broker
      - New intermediaries: sell user data (BlueKai,….)
  - Demand side platforms: single unified platform to buy inventories on multiple ad-exchanges

"*Half the money I spend on advertising is wasted; the trouble is, I don't know which half.*" - John Wanamaker

PROBLEM: BETTER USER TARGETING

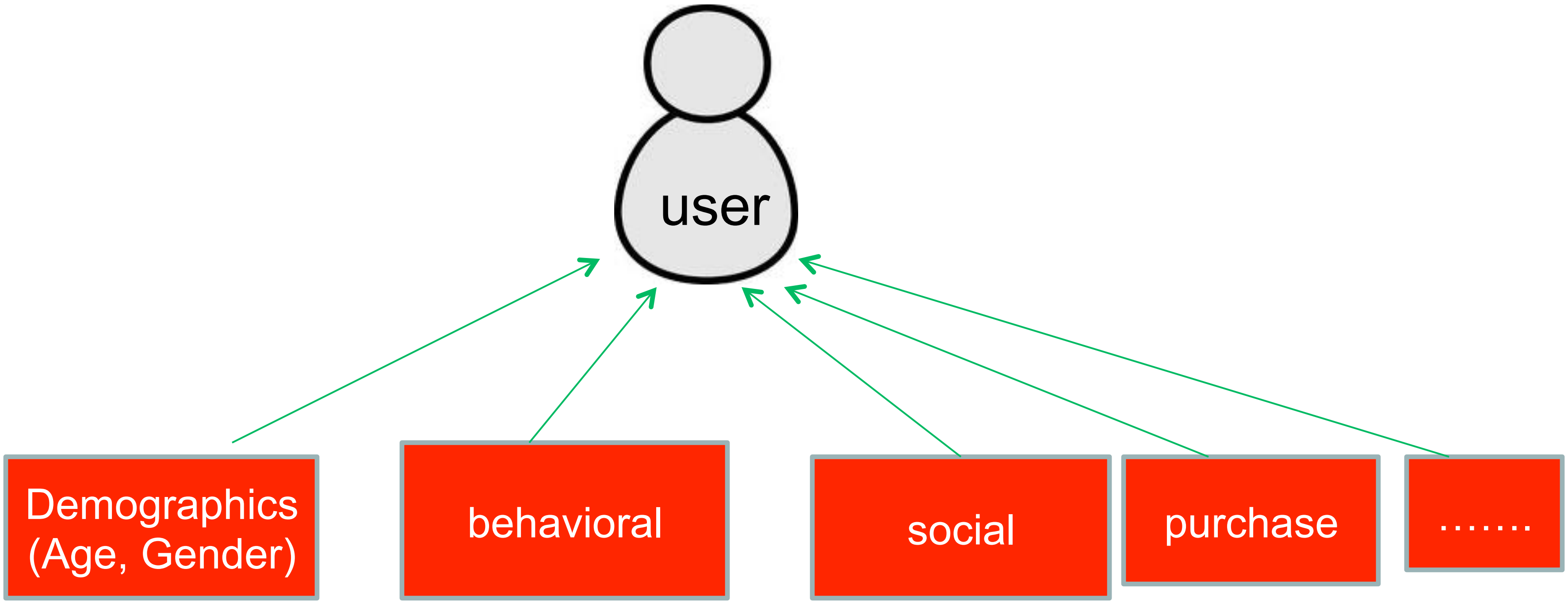
Our problem: User Targeting for Performance

- Given a campaign that is focused on conversions, how do we target the best users to maximize advertiser ROI and publisher revenue?

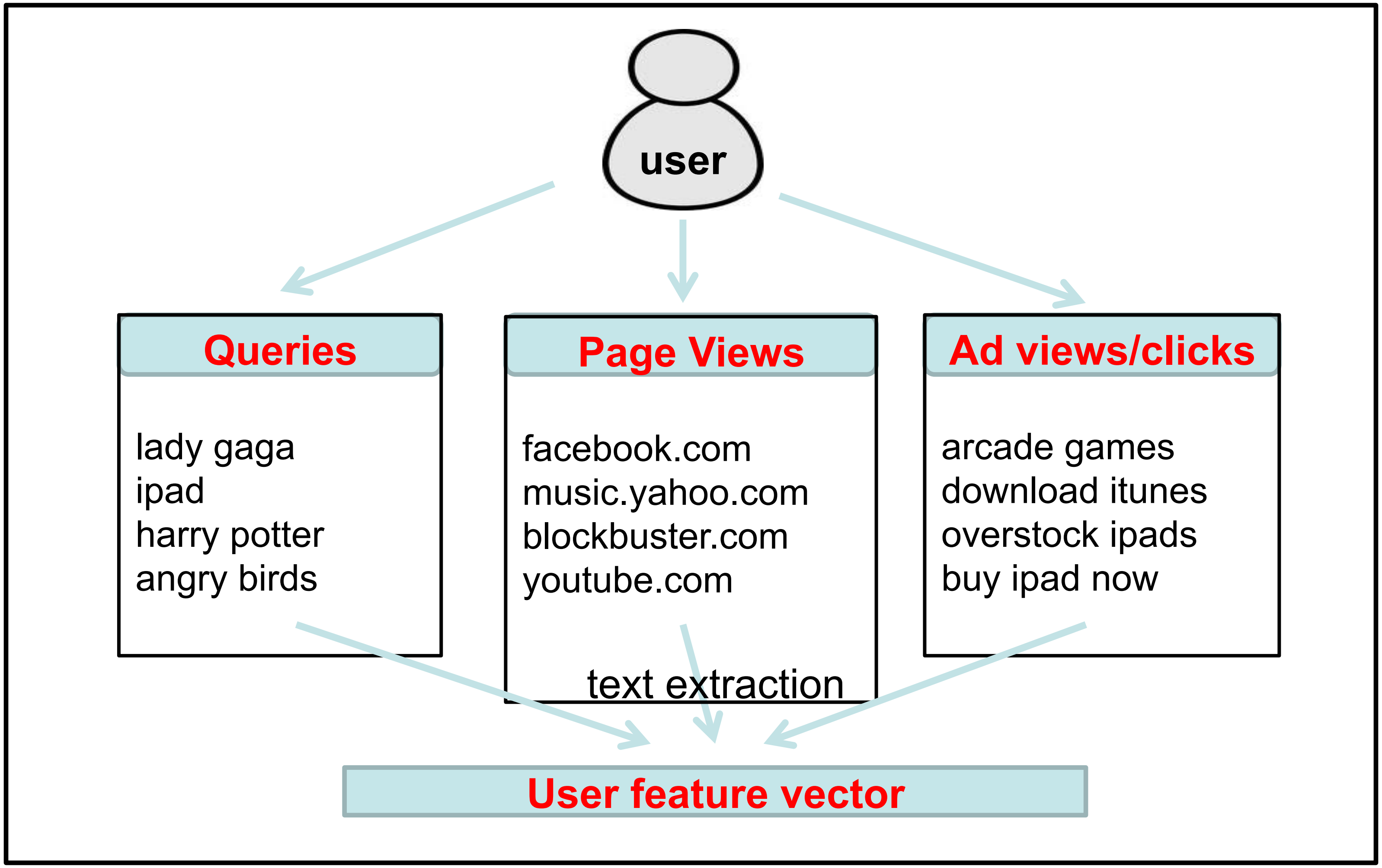
- Why conversions? Why not clicks?
  - Conversions better and direct measure of user intent to buy
  - Clicks subject to problems: Fraud, bounce
  - Advertiser measure performance via conversions

- One important aspect of conversions
  - Definition varies across campaigns
    - Buying, filling-up a survey, subscription, and so on.

Different kinds of user information

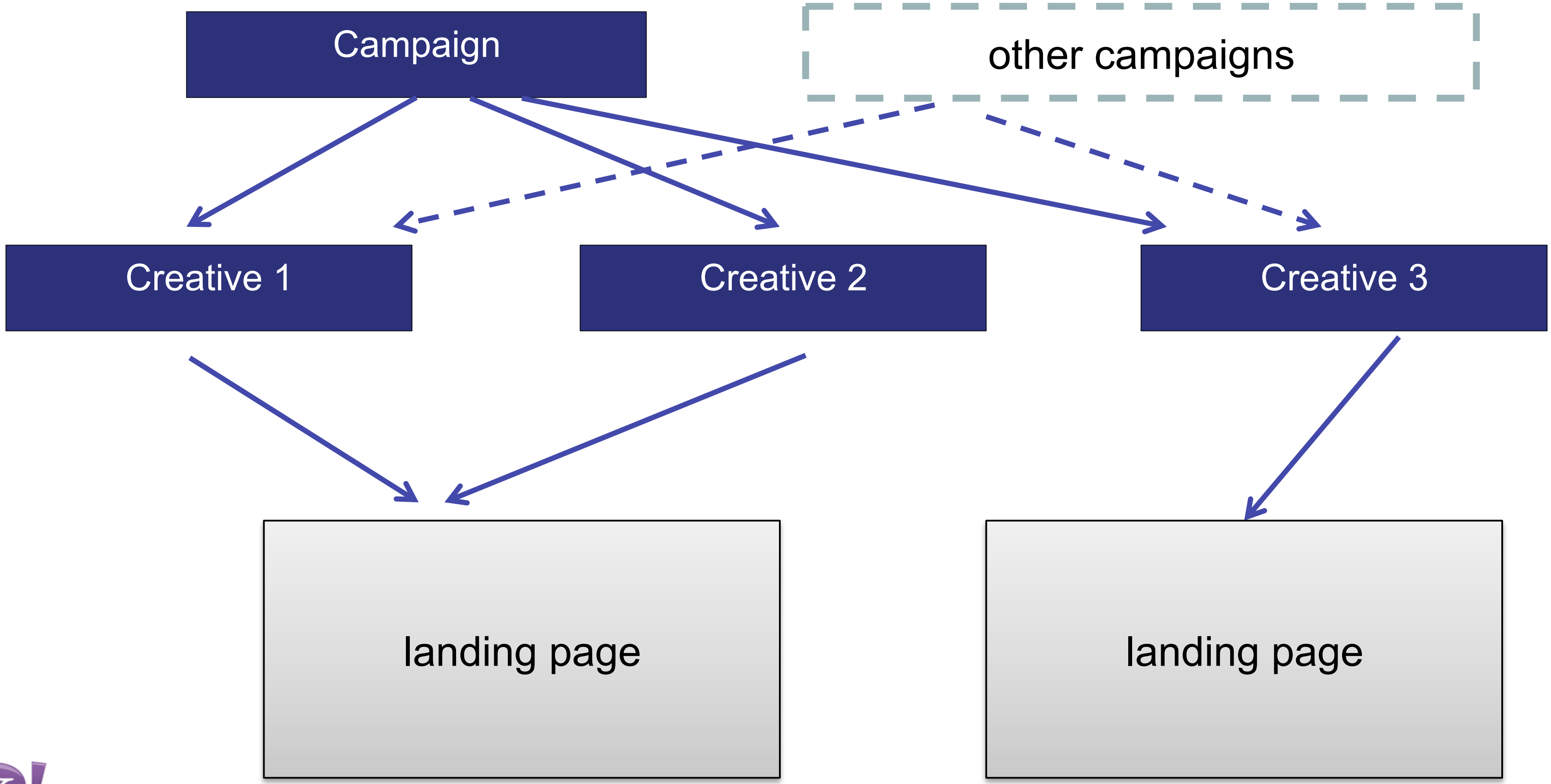


Behavioral User profile: Main focus in this work



Campaign information

Campaign metadata



Typical targeting

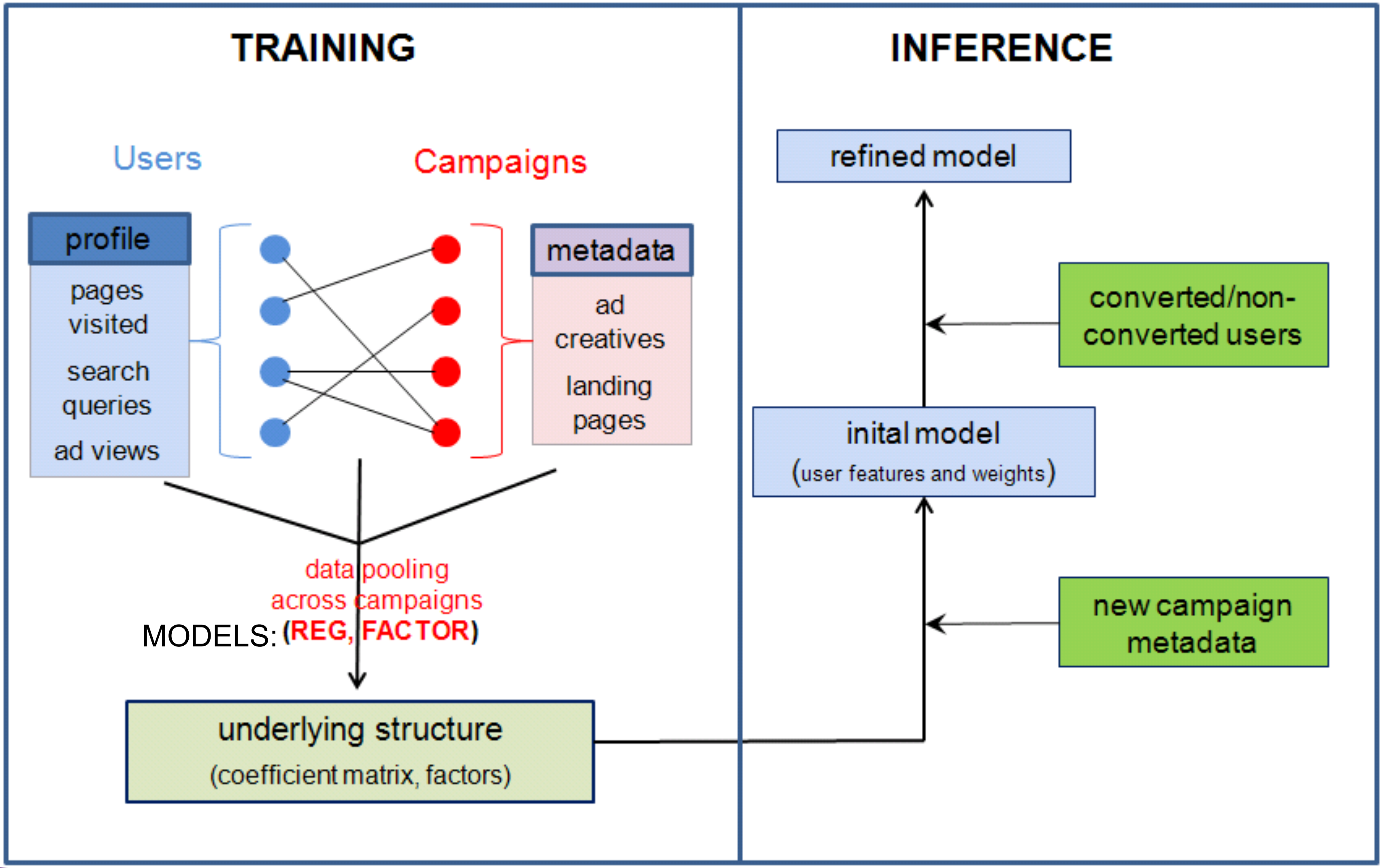
- Based on campaign information, look at different user dimensions and create a few targeting segments
  - E.g. campaign to sell Halloween costumes in Austin
  - Targeting attributes:
    - Users 25-50 living around Austin who searched for Halloween
  - Subjective, depends on the marketer

- Experimentation
  - Run the campaign for a while, collect conversion data and build a model to target users in a better way
    - Conversion rates low, experiment is expensive
    - Advertisers often lose patience and shut-off poor performing campaigns

Our Approach

- We take a fundamentally different approach
- Lots of campaigns run everyday in ad-exchanges, can we leverage this data and perform better user targeting for future campaigns?
  - Such data can also be constructed by other intermediaries in the advertising eco-system
- Intuitively, this seems plausible
  - Create clusters of campaigns, figure out where a new campaign fits
- Statistically, this is a challenging task

Our Approach: Use past campaign performance



MODELING

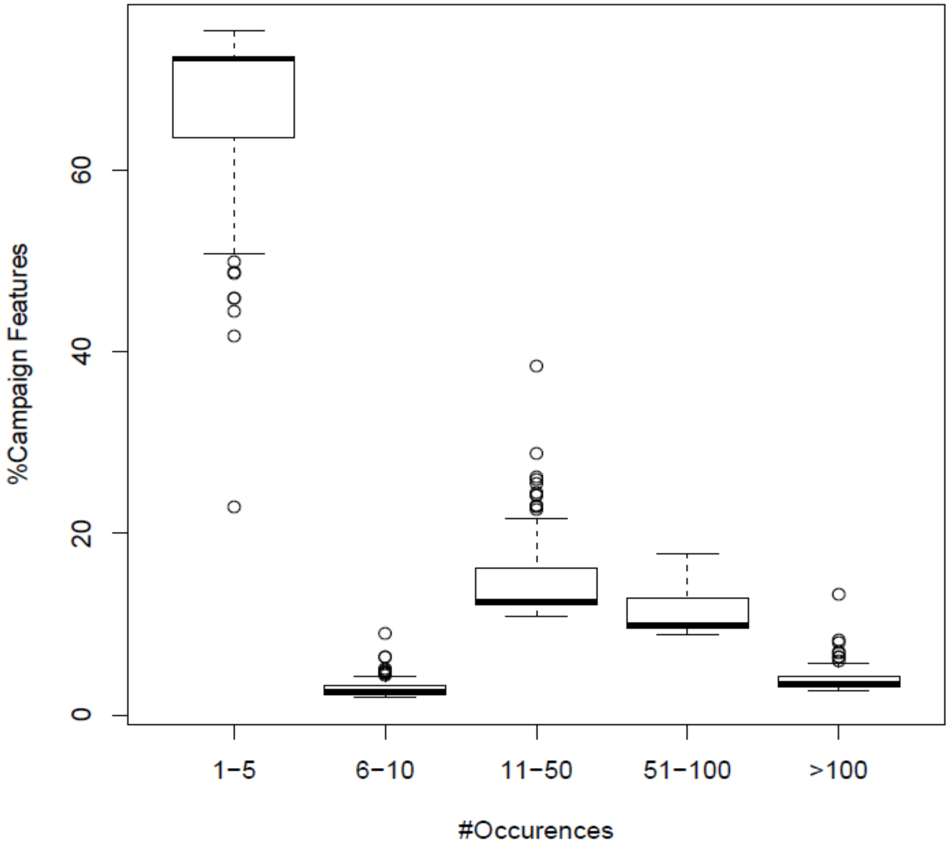
Per campaign model

- y_{kj} k^{th} targeted user on campaign j (convert/non-convert)
- $\boldsymbol{x}_{kj}$ user feature vector (high dimensional)
- Model: Logistic regression

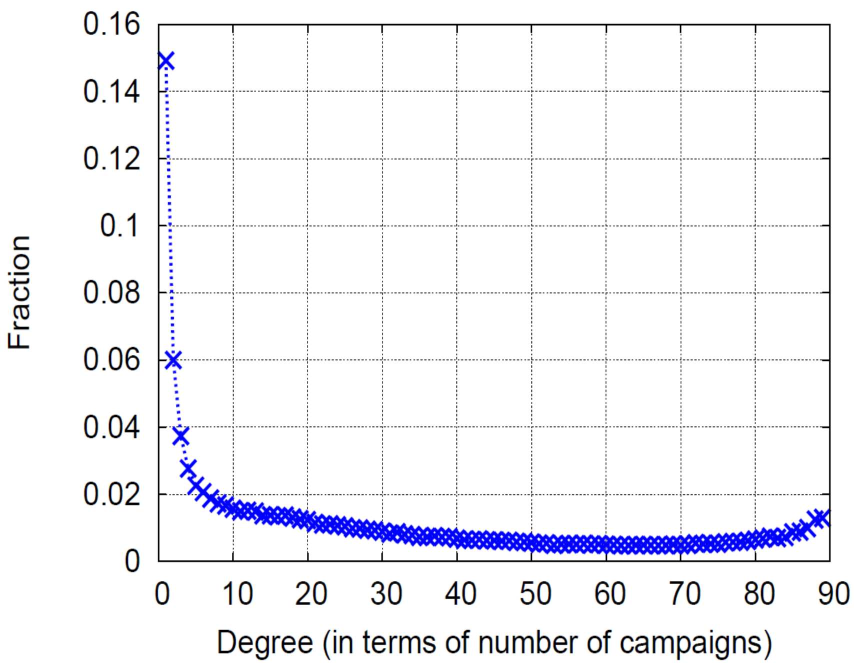
$$y_{kj} \sim \text{Bernoulli}(p_{kj});\, k \in \mathcal{U}_j$$
$$\log \frac{p_{kj}}{1-p_{kj}} = \boldsymbol{X}_{k,j}^{'} \boldsymbol{\beta}_j$$

- Difficulties
  - Sparseness & rareness of conversions
    - Hundreds of thousands of user features, few hundred to few thousand conversions per campaign

User Feature distributions: all features are binary (~100k)



Feature sparsity within campaign



Feature occurrence across campaigns

Logistic regression ill-conditioned

- Need: Feature cones of converters and non-converters to *overlap* (Silvapulle, 1981)
- Difficult to achieve
  - Hundreds of thousands of features, most of them sparse
  - Few hundred to few thousand conversions
- Approach: Multi-level Hierarchical models
  - Assume the coefficients are drawn from some prior distribution

$$\beta_{ij} \sim N(\beta_{ij0}, \sigma^2)$$

PRIOR MEAN

Campaign id

User Feature id

X ? X ? ….? ? X
….. ……. ……
….. …….. ……
? X ? X …. ? ?
X X X X X ? ? X
….. …… …….

Modeling Prior Mean

- ZEROMEAN: Usual L2 regularization

$$\beta_{ij0} = 0$$

  - State-of-the-art in Display Targeting
  - Limitation: Does not work with little or no data (Cold Start)

- REG

$$\beta_{ij0} = \mathbf{g}_i^{'}\mathbf{z}_j = \sum_k g_{ik} z_{jk}$$

$$G = (\mathbf{g}_1, \mathbf{g}_2, \ldots\ldots)$$

  - Too many parameters: e.g. 100k x 30 = 3M
    - Use L2 regularization on *G*

New Approach: FACTOR

- Factorize the incomplete feature-id x campaign matrix
- Idea borrowed from collaborative filtering
  - Factorize User x Item matrix
  - Here factorization one level up in the multi-hierarchy model

$$
\begin{aligned}
\beta_{ij0} &= \boldsymbol{u}_i^{'}\boldsymbol{v}_j \\
\boldsymbol{u}_i &\sim MVN(\mathbf{0}, a_u) \\
\boldsymbol{v}_j &\sim MVN(\boldsymbol{D}\boldsymbol{z}_j, a_v)
\end{aligned}
$$

- *r* = # factors (small, e.g. 4-5)
- ***D*** : (r + #campaign features) (small compared to REG)

Advantages of FACTOR

- Factorizing coefficient matrix reduces dimension
- Alternate parameterization of FACTOR

$$\beta_{ij0} = \mathbf{u}_i'(\mathbf{D}z_j + \eta_j^{r\text{x}1}) + \varepsilon_{ij}$$

$$= \mathbf{u}_i'\mathbf{D}z_j + \underbrace{\mathbf{u}_i'\eta_j^{r\text{x}1}}_{\text{Sharing parameters}} + \varepsilon_{ij}$$

- Recall REG

$$\beta_{ij0} = \mathbf{g}_i'z_j + \qquad\qquad + \varepsilon_{ij}$$

Other related approaches

- Seemingly Unrelated Regressions (SUR) [Zellner, 1962]
  - We deal with logistic + model covariance in residuals through campaign features and low-rank matrix structure

- Multi-tasking
  - Assumes complete feature-id x campaign matrix (not the case)
  - Does not incorporate campaign feature information
  - Usual approach assumes

$$\beta_{ij0} = \mathbf{u}_i' \eta_j \text{ (Low Rank)}$$

  - low-rank + sparse (we assume low-rank + white-noise)
    - Provides robust estimates after adjusting for heterogeneity
  - Relaxation important to adapt estimates with online data for new campaign, robustness is not the only issue here

Model Fitting

EM for our models (REG and FACTOR)

$$\boldsymbol{Y} : \text{ Data}$$

$$\boldsymbol{\Delta} : \text{ Latent variables}$$

$$\boldsymbol{\Theta} : \text{ hyper-parameters}$$

$$\text{Model: } p(\boldsymbol{Y}|\boldsymbol{\Delta}, \boldsymbol{\Theta})p(\boldsymbol{\Delta}|\boldsymbol{\Theta})$$

$$\text{Output needed: } \text{ Mode: } max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\boldsymbol{Y})$$

$$p(\boldsymbol{\Delta}|\boldsymbol{Y}) \approx p(\boldsymbol{\Delta}|\boldsymbol{Y}, \hat{\boldsymbol{\Theta}})$$

Computing the mode

$$log(p(\mathbf{\Theta}|\mathbf{Y})) = log(p(\mathbf{\Theta}, \mathbf{\Delta}|\mathbf{Y})) - log(p(\mathbf{\Delta}|\mathbf{\Theta}, \mathbf{Y}))$$

$$log(p(\mathbf{\Theta}|\mathbf{Y})) = E_{old}(log(p(\mathbf{\Theta}, \mathbf{\Delta}|\mathbf{Y}))) - E_{old}(log(p(\mathbf{\Delta}|\mathbf{\Theta}, \mathbf{Y})))$$

$$E_{old}: \text{ Expectation w.r.t. } p(\mathbf{\Delta}|\mathbf{\Theta}_{old}, \mathbf{Y})$$

Second term: Maximized at $\mathbf{\Theta}_{old}$

Find new value of $\mathbf{\Theta}$ that increases first term

The EM algorithm

$$\text{Initialize } \boldsymbol{\Theta}$$

$$\text{Iterate}$$

$$\text{E-step} \quad : E_{old}(log(p(\boldsymbol{\Theta}, \boldsymbol{\Delta}|\boldsymbol{Y})))$$

$$\text{M-step} \quad : argmax_{\boldsymbol{\Theta}} E_{old}(log(p(\boldsymbol{\Theta}, \boldsymbol{\Delta}|\boldsymbol{Y})))$$

Improving the log-likelihood in M-step is enough, no need to perform complete optimization (Generalized EM)

EM algorithm to fit REG

$$\Delta = \{\beta_{ij}\}_{\forall(i,j)}; \mathbf{\Theta} = (\mathbf{G}, \sigma^2)$$

Posterior independence across campaigns in E-step

$$[\boldsymbol{\beta}_{.1},\ldots,\boldsymbol{\beta}_{.j},..\mid \mathbf{Y}_{.1},\ldots,\mathbf{Y}_{.j},..,\mathbf{\Theta}] = \prod_j [\boldsymbol{\beta}_{.j} \mid \mathbf{Y}_{.j}, \mathbf{\Theta}]$$

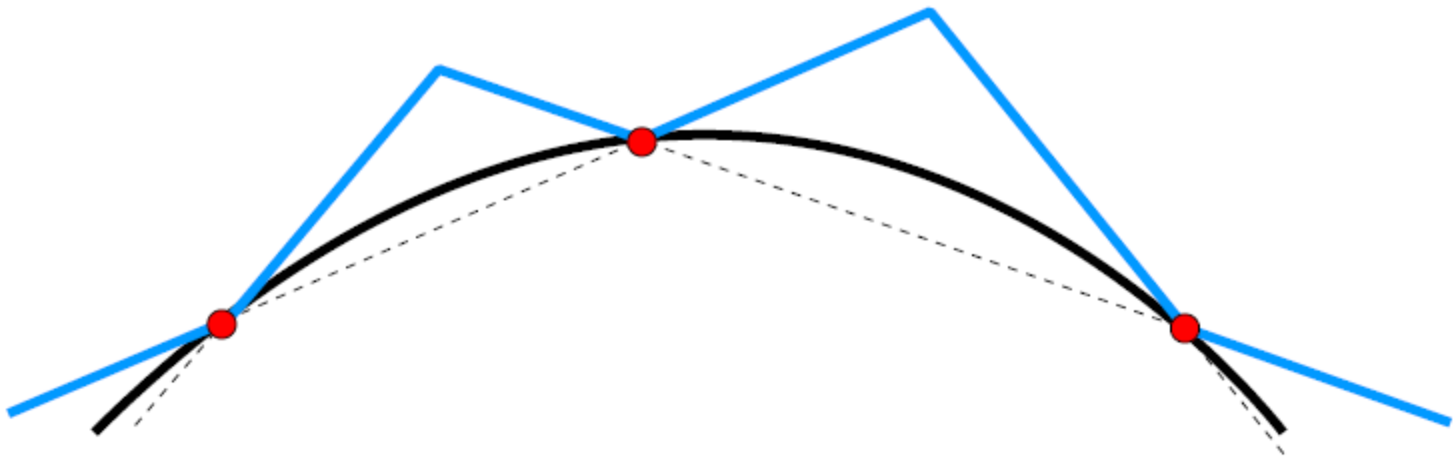
Posterior not in closed form

$$2\ln[\boldsymbol{\beta}_{.j} \mid \mathbf{Y}_{.j}, \mathbf{\Theta}] = 2\sum_k (y_{kj}\mathbf{X}'_{k,j}\boldsymbol{\beta}_{.j} - l\exp(\mathbf{X}'_{k,j}\boldsymbol{\beta}_{.j})) - \sum_i (\beta_{ij} - \mathbf{g}'_i z_j)^2 / \sigma^2 - C\log\sigma^2$$

We can sample through Gibbs sampling, each conditional log-concave
Adaptive Rejection Sampling (ARS) (Gilks and Wild, '92)

Adaptive Rejection Sampling (Gilks and Wild)

- Create upper (and lower bounds) to the density
- Perform rejection sampling, refine the envelope with each rejection (adaptive)



EM algorithm for REG

- E-step: Sample using MCMC, parallelize on Hadoop
  - Mapper: split data by campaign
  - Reducer: run MCMC for each campaign

$$-2E\log p(\Delta,\mathbf{Y}\mid\boldsymbol{\Theta}) = \sum_{(i,j)}((\hat{\beta}_{ij}-\mathbf{g}_i^{'}z_j)^2+\tau_{ij}^2)/\sigma^2 + C\log\sigma^2$$

- M-step:
  - Estimate the ***g***'s per feature-id through regression
  - Variance estimate is closed-form

$$C\hat{\sigma}^2 = \sum_{(i,j)}(\hat{\beta}_{ij}-\hat{\mathbf{g}}_i^{'}z_j)^2+\tau_{ij}^2$$

EM for FACTOR

- Parameters $\Delta = (\{\beta_{ij}\}_{\forall(i,j)}, \{\mathbf{u}_i\}_{\forall i}, \{\mathbf{v}_j\}_{\forall j}); \boldsymbol{\Theta} = (\mathbf{D}, \sigma^2, a_u, a_v)$
- Gibbs sampling for approximating posterior

$$[\{\beta_{ij}\}_{\forall(i,j)} \mid \{\mathbf{u}_i\}, \{\mathbf{v}_j\}, REST] \quad \text{ARS (parallelize)}$$

$$[\mathbf{u}_I \mid \{\mathbf{u}_i\}_{-I}, REST] = [\mathbf{u}_I \mid REST] \quad \text{(parallel)}$$

$$[\mathbf{v}_J \mid \{\mathbf{v}_j\}_{-J}, REST] = [\mathbf{v}_J \mid REST] \quad \text{(parallel)}$$

- Not amenable to Hadoop
  - One map-reduce job per MCMC iteration, infeasible

Approximate EM for FACTOR

- E-step: Assume the factors to be fixed and sample the logistic coefficients using per campaign Gibbs sampler
- M-step: Fit the following model to obtain estimates of

$(\{\mathbf{u}_i\},\{\mathbf{v}_j\},\mathbf{D})$

$$\hat{\beta}_{ij} \sim N(\mathbf{u}_i^{'}\mathbf{v}_j, \sigma^2)$$

$$\mathbf{u}_i \sim MVN(\mathbf{0}, a_u I); \mathbf{v}_j \sim MVN(\mathbf{D}z_j, a_v I)$$

- σ^2 estimated as in REG with change in prior mean to $\mathbf{u}_i \mathbf{v}_j$

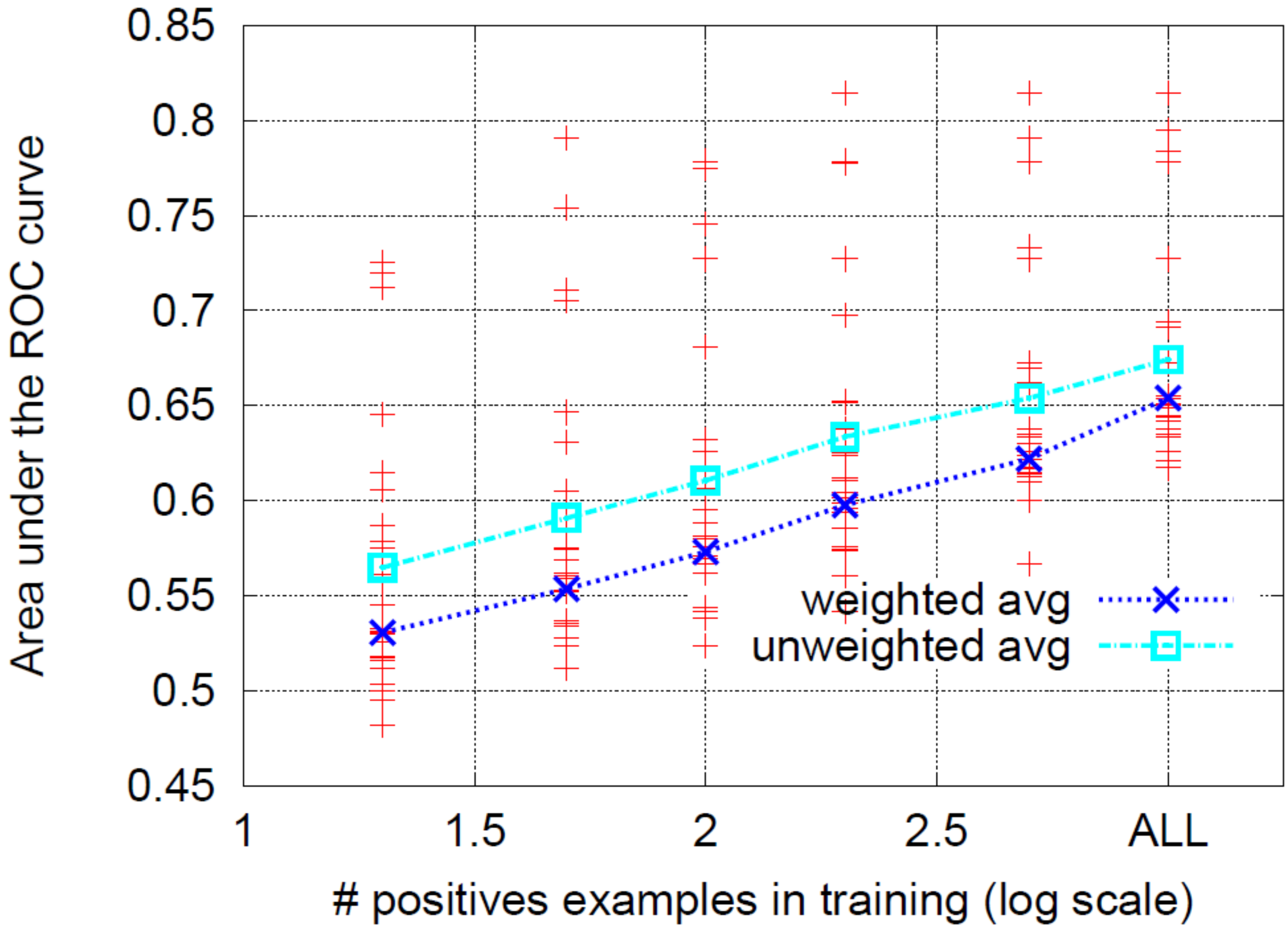
Experiments

Data example

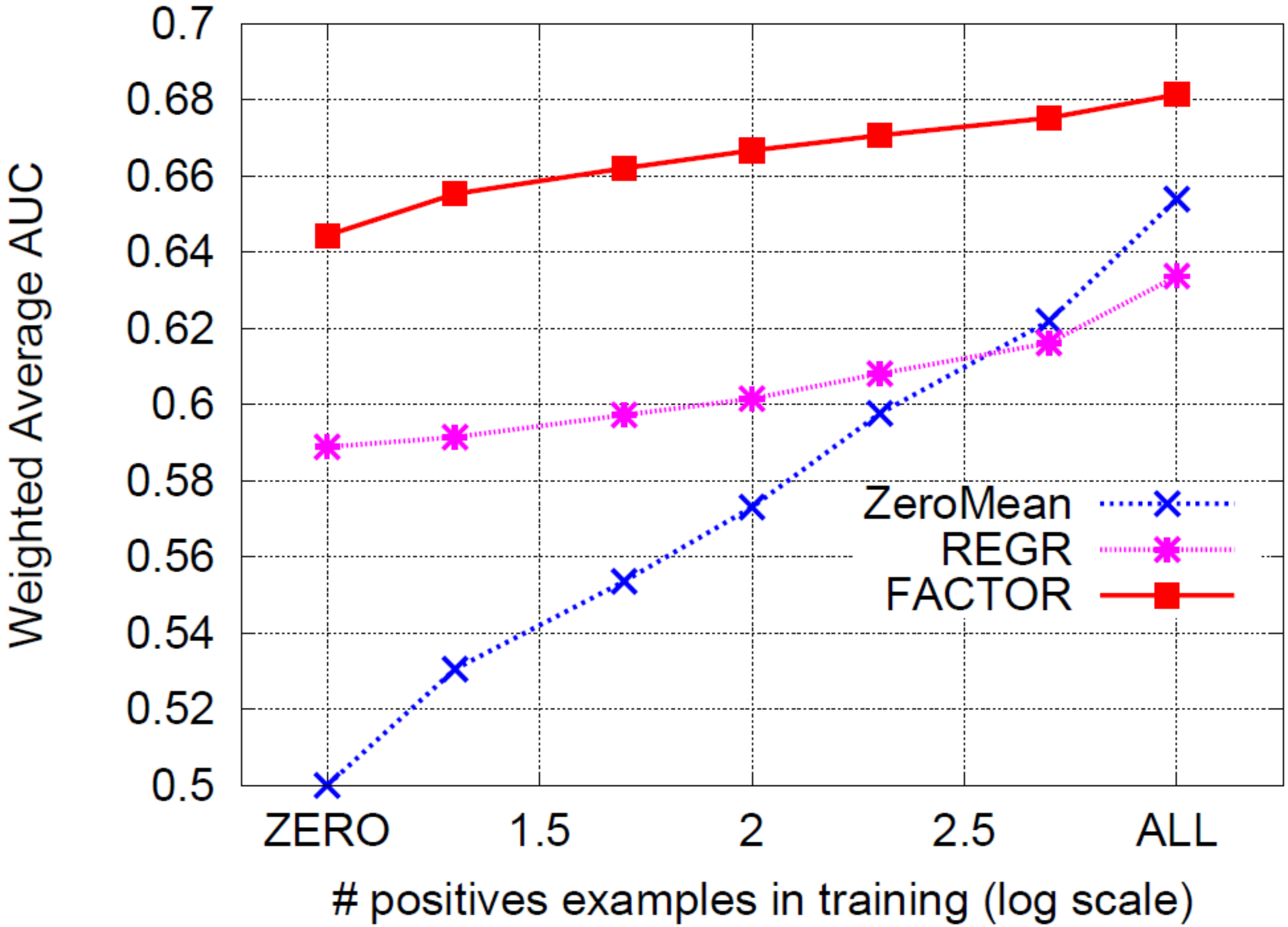
- 115 conversion based campaigns from Y! RightMedia
- 90 used for training, rest for testing
- TEST: simulate performance during different stages of a campaign’s lifetime

- Evaluation metric
  - AUC (Area under the ROC curve)
    - Pr(converter score > non-converter score)

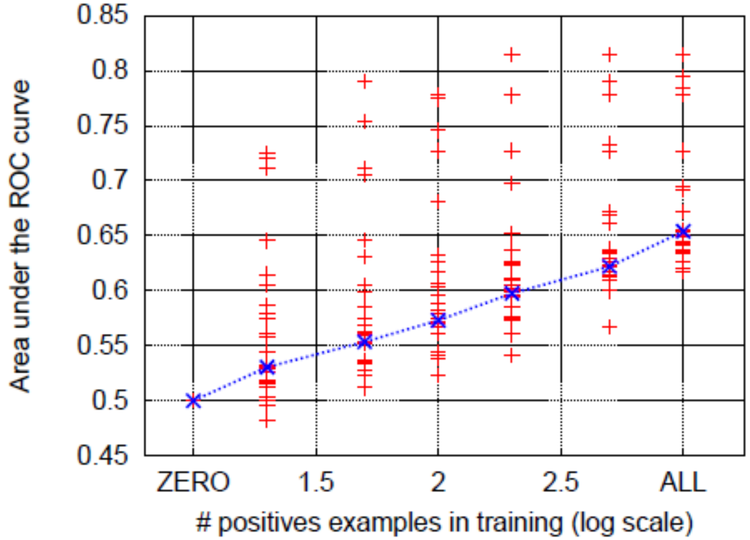
ZEROMEAN performance



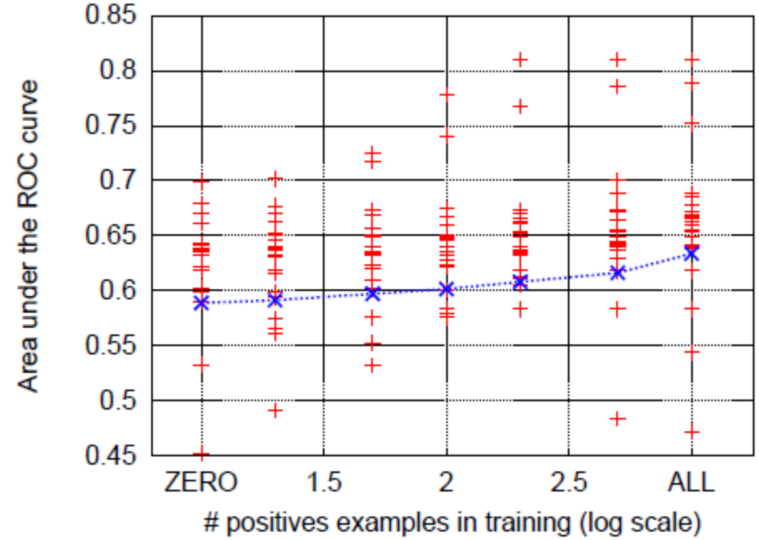
Comparing all methods



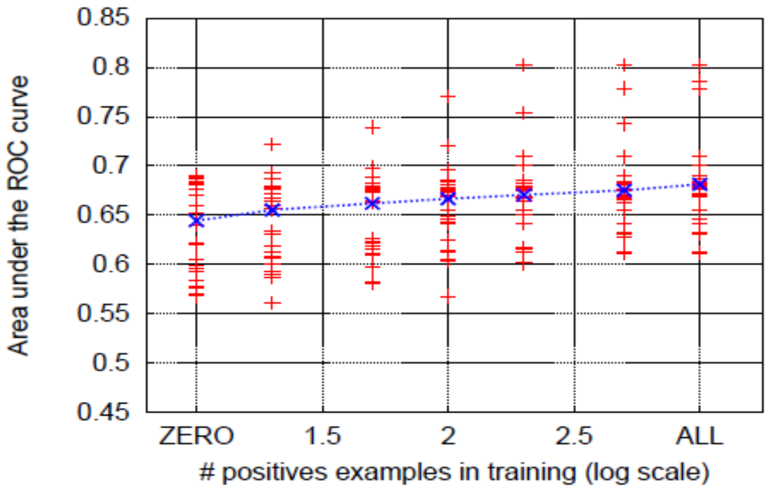
Variation across campaigns



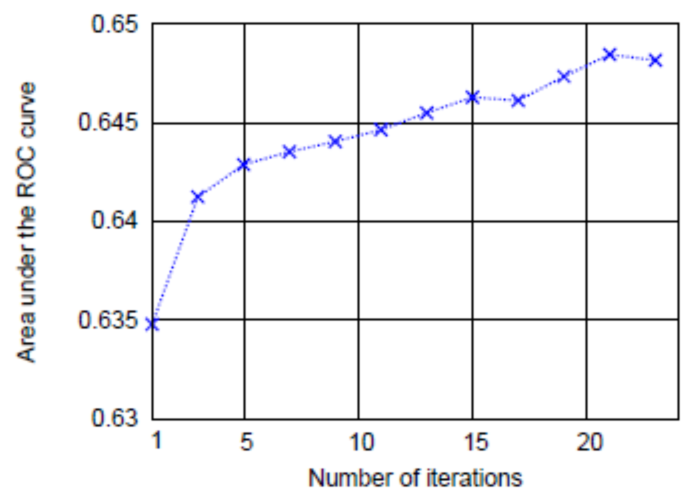
(a) ZEROMEAN



(b) REG



(c) FACTOR



Discussion

- Using previous campaign data via FACTOR helps provide better performance at the beginning of the campaign itself, encouraging for advertiser ROI
- FACTOR adapts as more conversions *trickle in* and improves model performance for a campaign
- Fitting algorithm scales in a map-reduce framework

To Conclude

- Targeting users is at the heart of Display Advertising that involves ingesting user information from disparate sources
  - e-marketeter estimates this to be an 80B industry growing at 17% p.a
- Current techniques are subjective and involves marketers inspecting few dimensions based on domain knowledge
- We showed using past conversion data can be a useful source to build effective targeting strategy for new campaigns, such data available to multiple entities in the advertising eco-system
- We achieved this through a new multi-tasking approach called FACTOR that handles sparsity, heterogeneity, and scales in a map-reduce framework