

# Large Scale Machine Learning in Digital Advertising

Seyed Abbas Hosseini  
Cofounder, Pegah Inc.  
Ph.D. 2018, Sharif  
abbas@tapsell.ir

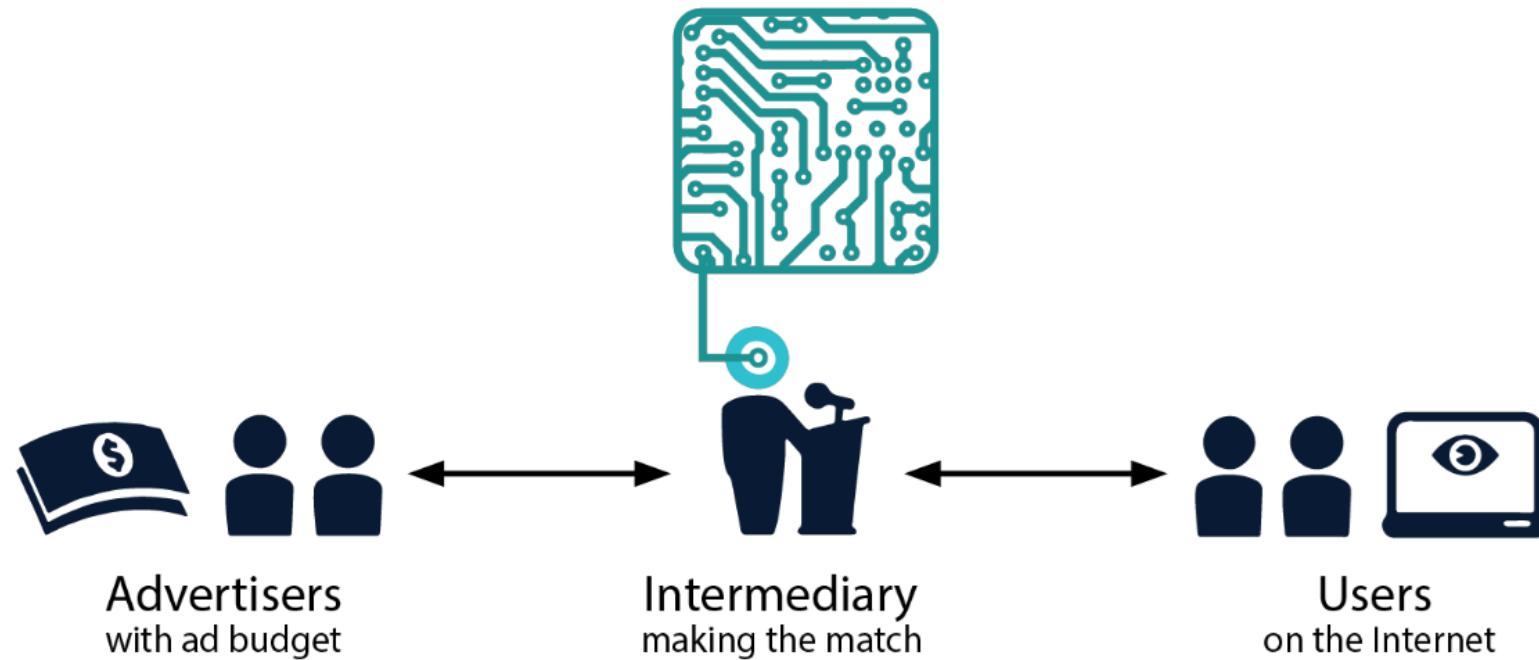


# Outline

- **Digital Advertising**
  - **Sponsored Search**
  - **Display Advertising**
- **RTB Mechanism**
- **Bid Estimation**
  - **CVR Estimation**
- **Other Interesting Issues**
- **Who We Are?!**

# Digital Advertising

Conveying advertisers' message to target audience in online media



# Sponsored Search

The screenshot shows a Google search for "iphone 6s case". The search bar is at the top with the Google logo and search icon. Below the search bar, there are navigation tabs for Web, Shopping, News, Images, Videos, and More. The search results show "About 16,900,000 results (0.33 seconds)".

**Organic Results:**

- iPhone 6s Cases - case-mate.com**  
www.case-mate.com/iPhone-6s-Cases  
4.6 ★★★★★ rating for case-mate.com  
Shop The iPhone 6s Case Collection. Free Standard Shipping!  
Refined Protection · Slim & Tough · Genuinely Crafted · Premium Designs
- iPhone 6s**  
www.apple.com/  
The only thing that's changed is everything. Learn more.  
A9 chip · Two sizes · Now in rose gold  
Pre-order 9.12 · iPhone Upgrade Program · 3D Touch · Cameras
- In the news**  
Speck's iPhone 6s CandyShell + MightyShell cases bring best-of-breed protection to Apple's latest iPhones  
9 to 5 Mac · 1 day ago  
With the iPhone 6s and iPhone 6s Plus debuting next week, it's important to start thinking ...
- Moshi's iPhone 6s and 6s Plus cases offer premium protection  
iMore · 23 hours ago
- Top 5 Best Leather iPhone 6s Cases  
Heavy.com · 12 hours ago
- More news for iphone 6s case
- iPhone 6s Cases & Covers from OtterBox**  
www.otterbox.com/en-us/iphone-6s-cases  
Get protection that inspires confidence with iPhone 6s cases and covers from OtterBox.  
Demandware SiteGenesis.
- iPhone 6s - Accessories - Apple**  
www.apple.com › iPhone › iPhone 6s  
The essential Apple-designed cases, accessories and all-new aluminum docks for iPhone 6s and iPhone 6s Plus.

**Sponsored Shopping Results:**

Shop for iphone 6s case on Google Sponsored

Case-mate - Karat Case Fo... \$49.99 Best Buy ★★★★★ (163)	Moshi - Iglaze Armour Case... \$39.99 Best Buy ★★★★★ (161)	Logitech - Protection... \$21.99 Best Buy ★★★★★ (90)	Moshi - Overture Wall... \$49.99 Best Buy ★★★★★ (18)
Case-mate - Brilliance Cas... \$44.99 Best Buy ★★★★★ (294)	Case-mate - Wallet Folio C... \$54.99 Best Buy ★★★★★ (173)	Marc by Marc Jacobs Metall... \$38.00 shopbop	Case-mate - Karat Hard Sh... \$49.99 Best Buy ★★★★★ (34)

Search Engine

The screenshot shows an app market search for "سوپرمارکت" (Supermarket). The search bar is at the top with a microphone icon and search icon. The results show several sponsored app cards, each with a "نصب" (Install) button, a rating, and a star icon.

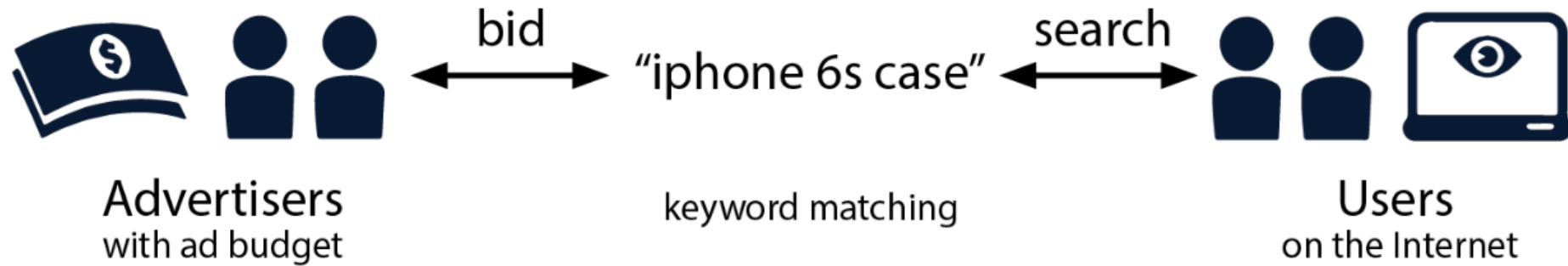
**Sponsored App Results:**

- دیجی کالا (Digikala)**  
4.0 ★  
نصب
- سوپرمارکت - بازی برای بچه‌ها**  
4.3 ★  
نصب
- اسنپ مارکت (سوپرمارکت آنلاین) Market!**  
4.2 ★  
نصب
- سوپرمارکت - بازی فروشگاه کودکانه**  
3.5 ★  
نصب  
+ خرید درون برنامه‌ای

At the bottom, there is a navigation bar with icons for بازار من (My Market), جستجو (Search), دسته‌ها (Categories), ویدیو (Video), and صفحه اصلی (Home).

App Market

# Sponsored Search




- Advertiser sets a bid price on Keywords
- User searches the keyword
- Search engine or market owner ranks ads and selected the best match


# Display Advertising

رایگان پیش بینی کن!

شروع کن ←



قرعه کشی جوایز



« مورینیو: نمی‌دانم چه در سر اریکسن می‌گذرد »

« جای بوستکس باشم از بارسا می‌روم »

« دو رکورد تاریخی برای مهاجم آلمانی لایپزیس »

امپراتور ستن مهتری (دوب آهن)

شیخ دیاباته (استقلال)

مدی ترابی (پرسپولیس)

عیسی آل کثیر (صنعت نفت آبادان)

محمدرضا سلیمانی (سپاهان)

کیروش استللی سوارز (سپاهان)

حسین مالکی (شاهین)

شهرداری پو شهر

سامان انصاری (تراکتور)

### آخرین اخبار سایر ورزش‌ها

« لشگری رییس کمیته پزشکی فدراسیون جهانی ووشو شد »

« لویه با غفور همچنان بدون شکست در سری آ »

« عزیزانی جام جهانی به ایران برگردانده می‌شود؟ »

« خلاصه بسکتبال فینیکس سانز - دنور ناگتس »

« وزنه برداری فقط مردانه نیست! »

« حذف کمانداران ریکرو از قهرمانی آسیا »

« خلاصه بسکتبال دالاس ماوریکس - هیوستون راکتس »

« لیگ NBA: شکست نشدنی آدام قضاپی و شرکا در خانه »

« شب فراموش نشدنی آدام قضاپی و شرکا در مادرید »

« عیادی پور: هنوز ۱۰۰ درصد آماده نیستم »

« حرکت های برتر لس آنجلس کلیبرز در فصل ۲۰۱۹-۲۰ »

« عطایی: اردوی ۴ روزه تیم ملی والیبال منطقی نیست »

« شاهین طبع: پیگیر اردوی تدارکاتی برای تیم ملی هستیم »

« جام جهانی کشتی فرنگی در ایران لغو شد »

« اسپانیا قهرمان رقابت های دیویس کاپ شد »

« شهرداری ورامین همچنان بدون شکست »

« پیروزی نانچینگ مقابل فوجیان با درخشش حدادی »

« خلاصه والیبال راه باب ملل مریوان - شهرداری ورامین ۳ »

« خلاصه والیبال خاتم اردکان - پیام مشهد »

« دادگر: الهام هاشمی گزینه اول و آخر ماست »

« دوست نذارم به من بگویند مرد عنکبوتی »

« صعود ۴ کماندار ایران به دور دوم مسابقات حذفی »

« شعبانین سرعربی تیم والیبال بانوان ایران شد »

« برد استثنایی برای گاوهای شیکاگو »

« تیم های بدون شکست، میهمان حرفان »

« ایران از صعود به نیمه نهایی جام جهانی سایر بازماند »


« خلاصه بسکتبال میامی هیت - فینال دنیا سوکسرز »

« سیدپندی مسابقات بسکتبال انتخابی المپیک مشخص شد »

« یک پیروزی و یک شکست در انفرادی کامیوند »

« برتری قاطع پرسپولیس و سپاهیا مقابل حرفان »

مطالب پیشنهادی



خرید پیراهن های مردانه رسمی و شیک از ۱۲۹ هزار تومان!

بهترین پاسورها

انتخاب لیگ:

لیگ برتر ایران

۴ پاس

معین عباسیان (سپاهان)

وریا غفوری (استقلال)

امید نورافکن (سپاهان)

مسعود شجاعی (تراکتور)

محمد محبی (سپاهان)

حسین ابراهیمی (نفت)

مسجد سلیمان

میلاذ جهانی (صنعت نفت آبادان)

مدی ترکمان (پارس جنوبی جم)

مدی ترابی (پرسپولیس)

لیگ برتر و ایلیال

انتخاب هفته:

۸ هفته

لیگ های خارجی

لیگ برتر انگلیس

انتخاب هفته:

۱۳ هفته

لیگ برتر انگلیس

شنبه ۴ آذر ۱۳۹۸

وستام ۲ - ۳ تانتیم

وانفور ۰ - ۳ برنلی

آرسنال ۲ - ۲ ساوتهمپتون

بورنموث ۱ - ۲ اورهمپتون

برایتون ۰ - ۲ لستر سیتی

کریستال پالاس ۱ - ۲ لیورپول

اورتون ۰ - ۲ نوریچ

منچستر سیتی ۲ - ۱ چلسی

یکشنبه ۳ آذر ۱۳۹۸


شفیلد یونایتد ۳ - ۳ منچستر یونایتد

دوشنبه ۴ آذر ۱۳۹۸

استون ویلا - نیوکسل ۳-۳

مشاهده جدول کامل لیگ برتر انگلیس

مطالب پیشنهادی



خوش ساخت ترین خانه های گلستان را با یک کلیک ببینید!

## کمپز

**+۴۵** اطلاعات کاربری

**+۹۰** دعوت از دوستان

**+۲۵** مشاهده تبلیغ

**+۴۵** پیوستن به صفحه اینستاگرام بازی

**+۴۵** پیوستن به کانال تلگرام بازی

دریافت بازی باقلوا

حمایت از ما









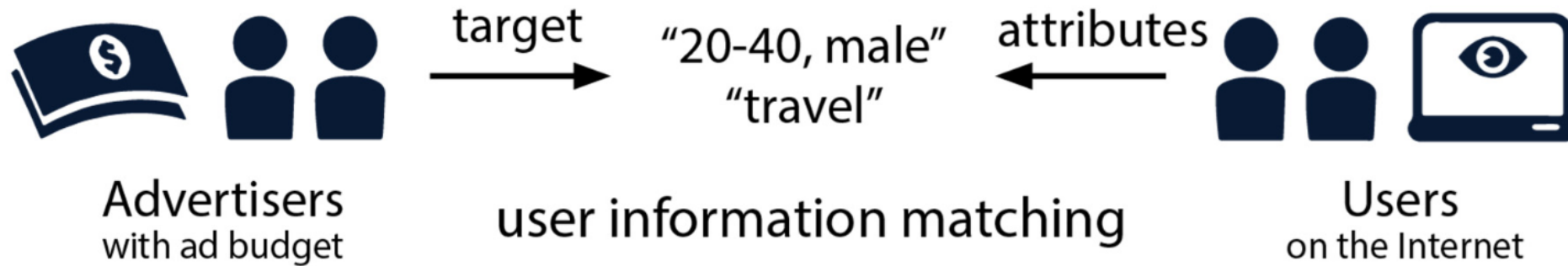
سازن تخت هدیه کن

**دیوار** اپتیما، شاهکار طراح آلمانی را قیمت کنید

دیوار، خرید و فروش بی واسطه

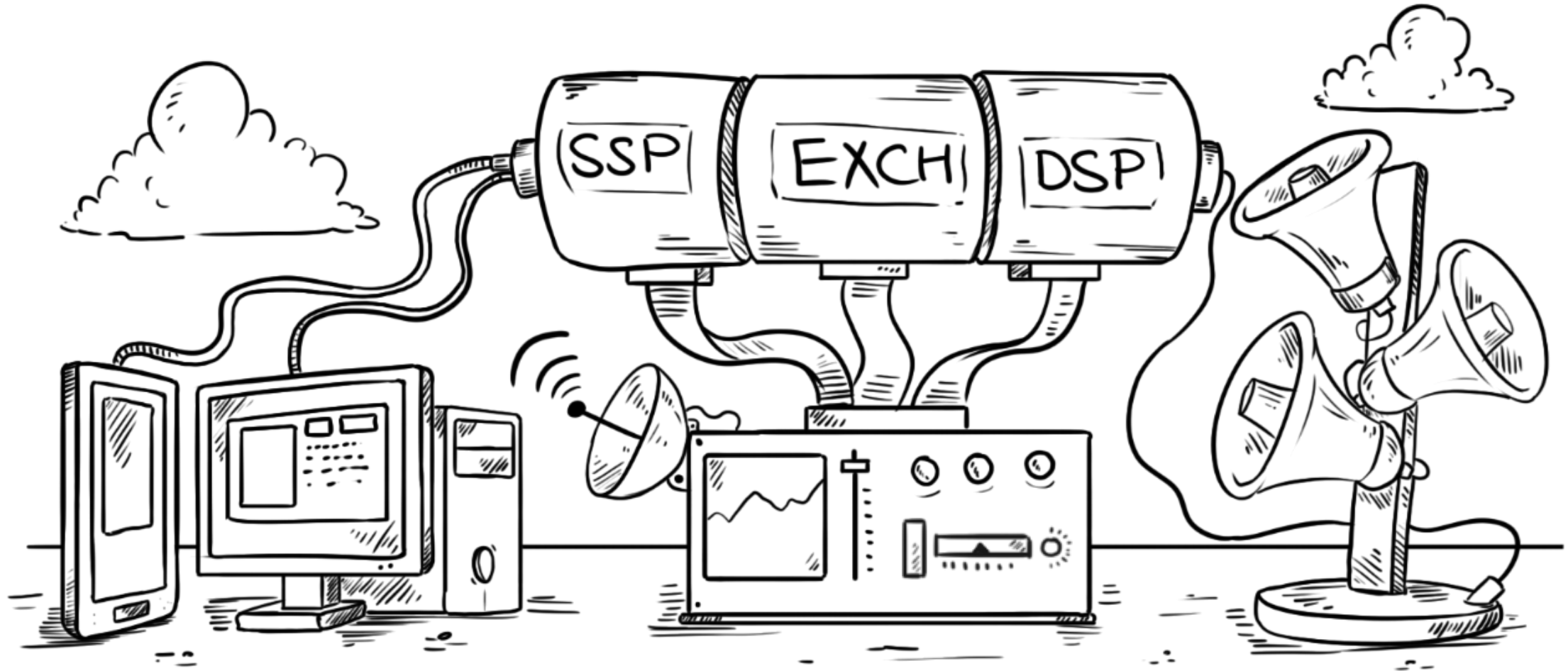
نصب رایگان

# Display Advertising



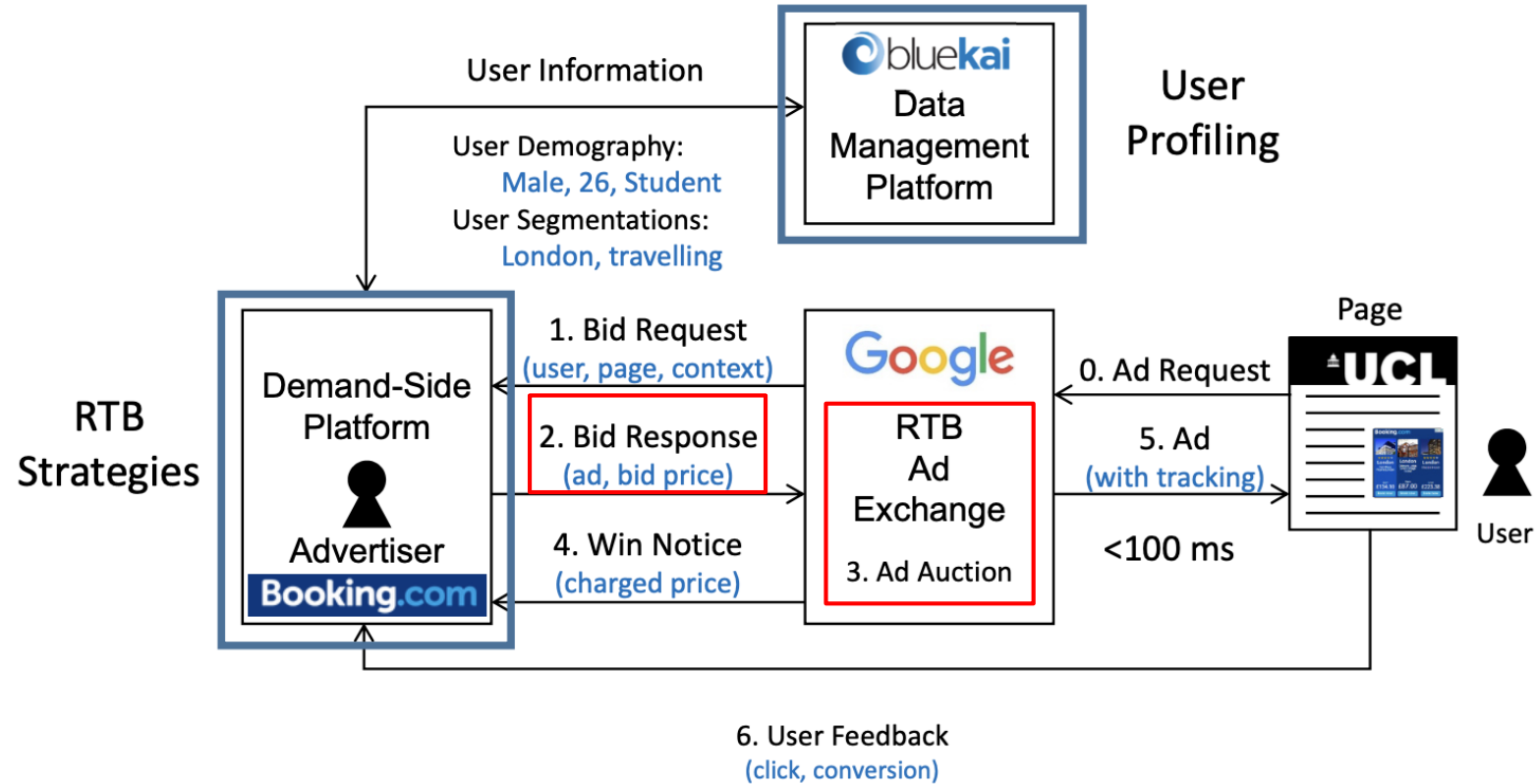
- **Advertiser targets a segment of users**
  - **No matter what the user is searching or reading**
- **Ad Network selects the best ad to show to the user**

# Digital Advertising Ecosystem





# Display Advertising Ecosystem



- Buying ads via RTB, 10 billion per day
- A real big data battlefield

	Query per Second
Turn DSP	1.6 million
Google	40,000 search

# Auction Mechanism



**SOLD!**

at **\$8**

First Price  
Auction



**SOLD!**

at **\$5**

Second Price  
Auction



# Bid Estimation

- Each Advertiser has many campaigns
- With different Pricing Schemas
  - **CPM**: cost per mille impression [favored by publisher]
  - **CPC**: cost per click
  - **CPA**: cost per action [favored by advertiser]



- **Goal**: Maximize Revenue



- **Simple Solution**:

- Select ad based on  
Expected Revenue per Impression
- suppose: ad a, goal cpc

$$E[Rev|a, u] = Pr(click|a, u) * CPC_a$$

Called CVR, **Unknown!**  
Need to be calculated

Income per Click,  
**Known**

# CVR Estimation: Problem Definition

- **Problem Definition**

## One instance data

- Date: 20160320
- Hour: 14
- Weekday: 7
- IP: 119.163.222.\*
- Region: England
- City: London
- Country: UK
- Search Query: "iphone 6s case"
- OS: Windows
- Browser: Chrome
- Ad title: "iphone 6s case on sale!"
- Ad content: "Customize your case design"
- Bid keywords: "iphone case"
- User occupation: Student
- User tags: Sports

## Corresponding label



Click (1) or not (0)?

Predicted CTR (0.15)

- **Available Data about**

- **User**
- **Context**
- **Ad**

# CVR Estimation: Feature Engineering

- One-Hot Binary Encoding

$x = [\text{Weekday=Friday}, \text{Gender=Male}, \text{City=Shanghai}]$

$x = [0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, \dots]$

Sparse representation:  $x = [5:1, 9:1, 12:1]$

- Prediction Challenges:

- High Dimensional Data
- Too Sparse Feature Vectors
- Very Unbalanced Classification [The convert events are too rare]
- Real-time response [**<100ms**]

# CVR Estimation: Predictive Models

- Generalized Linear Models
  - **Logistic Regression**
  - **Bayesian Probit Regression**
- Factorization Machines
  - **Sparse Factorization Machines**
  - **Field-Aware Factorization Machines**
  - **Field-Weighted Factorization Machines**
- Deep models
  - **Deep CTR Predictor**
  - **Deep Factorization Machines**
  - **Wide and Deep Recommender Systems**

# Generalized Linear Models

- General Form

$$p(y|x, w) = f(w^T x)$$

- Logistic Regression

- Likelihood is convex and hence Parameters can be learnt using ML
- Learning can be done in an online fashion using stochastic Gradient Descent

$$p(y = 1|x, w) = \sigma(w^T x)$$

$$E(w) = -\ln p(Y|X, w) = \sum_{n=1}^N y_n \ln \sigma(w^T x) + (1 - y_n) (1 - \ln \sigma(w^T x))$$

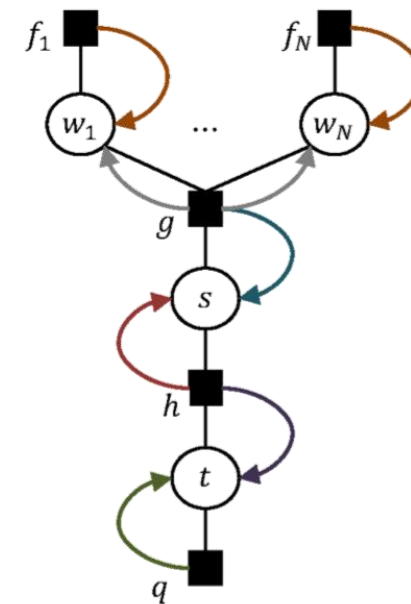
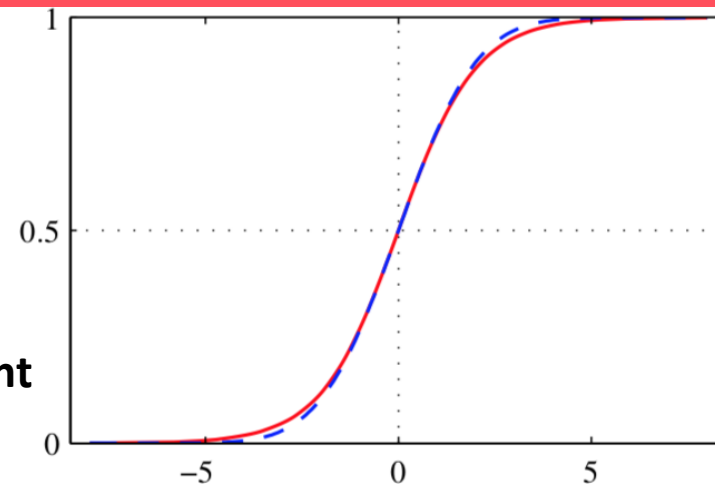
- Bayesian Probit Regression

- A fully Bayesian method based on a Gaussian prior over latent weights
- Posterior can be found online using stochastic variational inference
- Bing's Sponsored Search CTR Prediction algorithm

$$W \sim \prod_{i=1}^N \prod_{j=1}^{M_i} N(w_{ij}; \mu_{ij}, \sigma_{ij}^2)$$

$$y = \text{sgn}(w^T x + \epsilon) \quad \text{where} \quad \epsilon \sim N(0, \beta^2)$$

$$\Rightarrow p(y|x, w) = \Phi\left(\frac{y \cdot w^T x}{\beta}\right)$$



# Generalized Linear Models

- Pros
  - **Fast Prediction**
    - **Only one inner Product should be calculated**
  - **Fast Learning Methods**
    - **Efficient online algorithms exist for both proposed methods**
  - **Interpretable**
- Cons
  - **Linear models don't consider correlation among features**
  - **Linear models can only memorize feature combinations which users have already performed actions on**



# Factorization Machines

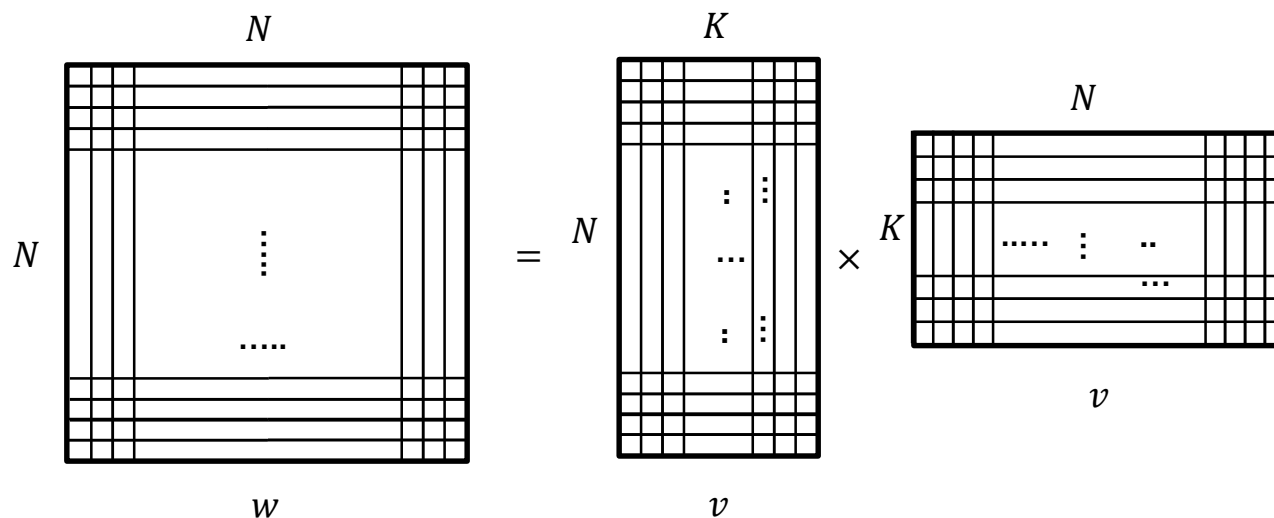
- One way to consider inter-feature correlations is using polynomial kernels

$$p(y|x, w) = f(\phi(x, w))$$

$$\phi(x, w) = \sum_{i, j \in F} w_{ij} x_i x_j$$

- Challenge: the model has  $\mathcal{O}(N^2)$  parameters where  $N$  is the number of features
  - A very common idea in machine learning in this scenario is using factorized models

$$\phi(x, w) = \sum_{i, j \in F} v_i^T v_j x_i x_j$$



# Field-Aware Factorization Machines

- In FMs, every feature has only one latent vector to learn the latent effect with any other feature
- In FFM, each feature has several latent vectors. Depending on the field of the other features, one of them is used to do the inner product.

Clicked	Publisher (P)	Advertiser (A)	Gender (G)
Yes	Tabnak	Digikala	Male

$$\phi_{FM}(x, w) = v_{Tabnak}^T \cdot v_{DigiKala} + v_{Tabnak}^T \cdot v_{Male} + v_{DigiKala}^T \cdot v_{Male}$$

$$\phi_{FFM}(x, w) = v_{Tabnak,A}^T \cdot v_{DigiKala,P} + v_{Tabnak,G}^T \cdot v_{Male,A} + v_{DigiKala,G}^T \cdot v_{Male,P}$$

$$\phi_{FFM}(x, w) = \sum_{i=1}^n \sum_{j=i+1}^n v_{i,f_2}^T \cdot v_{j,f_1} x_i x_j$$

	#variables
LM	$n$
Poly2	$B$
FM	$nk$
FFM	$nfk$

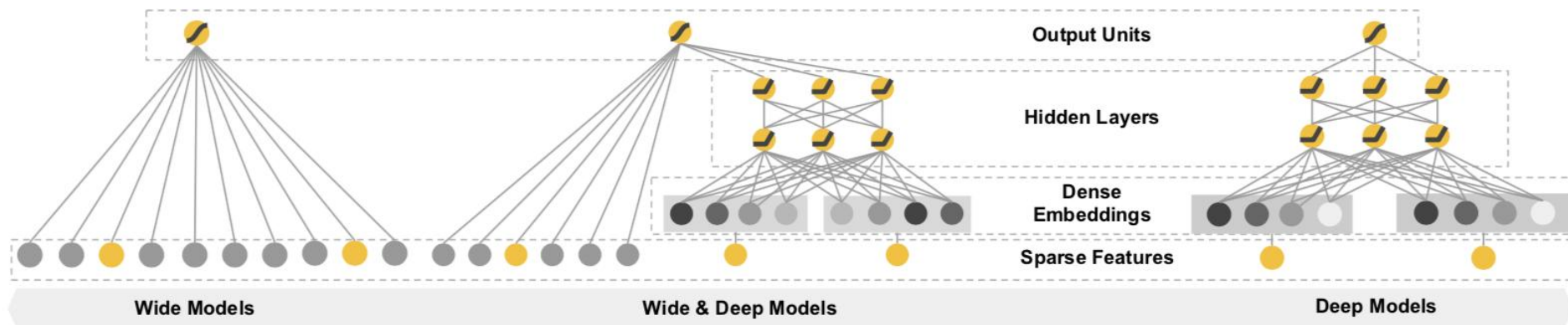
# Factorization Machines

- Pros
  - **Fast Prediction**
    - **Only one inner Product should be calculated**
  - **Considers Correlation Among Features**
    - **FFM won many Kaggle challenges due to its superior performance**
- Cons
  - **Learning FM models is more computational expensive than linear models**
  - **Learning the parameters can't be done online**
  - **FMs can't consider correlations among more than two features**
  - **Over-generalization**

# Wide & Deep Model

- Memorization of feature interactions through a wide set of cross-product feature transformations are effective and interpretable
- Generalization requires more feature engineering effort.
- Deep neural networks can generalize better to unseen feature combinations through low dimensional dense embeddings learned for the sparse features.
- Deep neural networks with embeddings can over-generalize and recommend less relevant items when the user-item interactions are sparse and high-rank

$$P(Y = 1|\mathbf{x}) = \sigma(\mathbf{w}_{wide}^T[\mathbf{x}, \phi(\mathbf{x})] + \mathbf{w}_{deep}^T a^{(l_f)} + b)$$



# Wide & Deep Model

- Pros
  - **Good generalization and memorization**
- Cons
  - **Learning deep models is computationally expensive**
  - **Time consuming prediction method**
    - **Deep features need to be calculated in prediction time**
    - **Can't be scaled to RTB size but can be used in sponsored search**

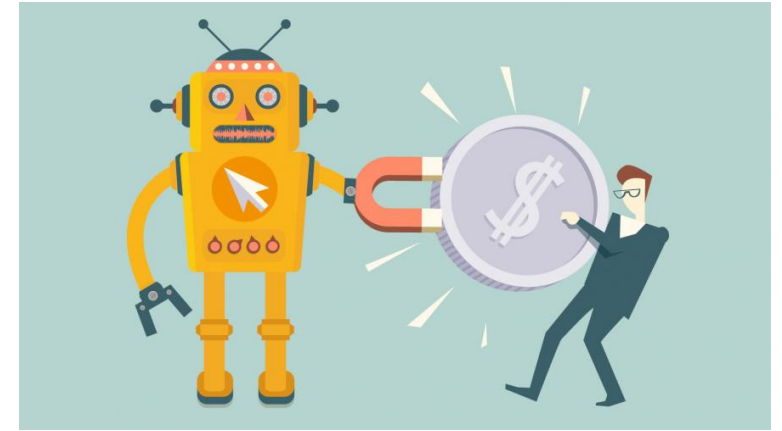
# Other Interesting Issues



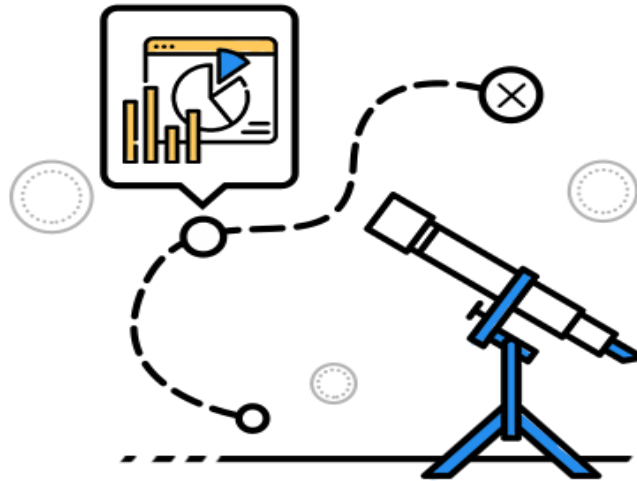
Frequency Capping



Budget Pacing



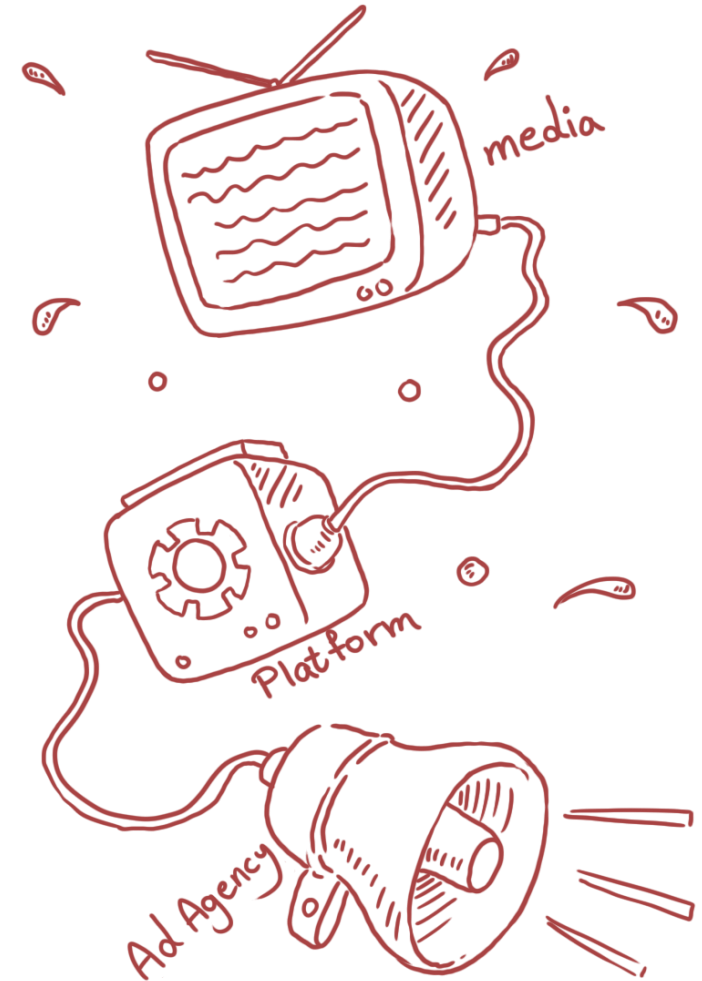
Fraud Detection



Attribution

# Who we are

- **Sponsored Search Advertising**
  - Bazaar Search Advertising
- **Display Advertising**
  - Websites
  - Mobile Applications
- **Social Media Advertising**
  - Micro Influencer Advertising



# Tapsell 1<sup>st</sup> Generation

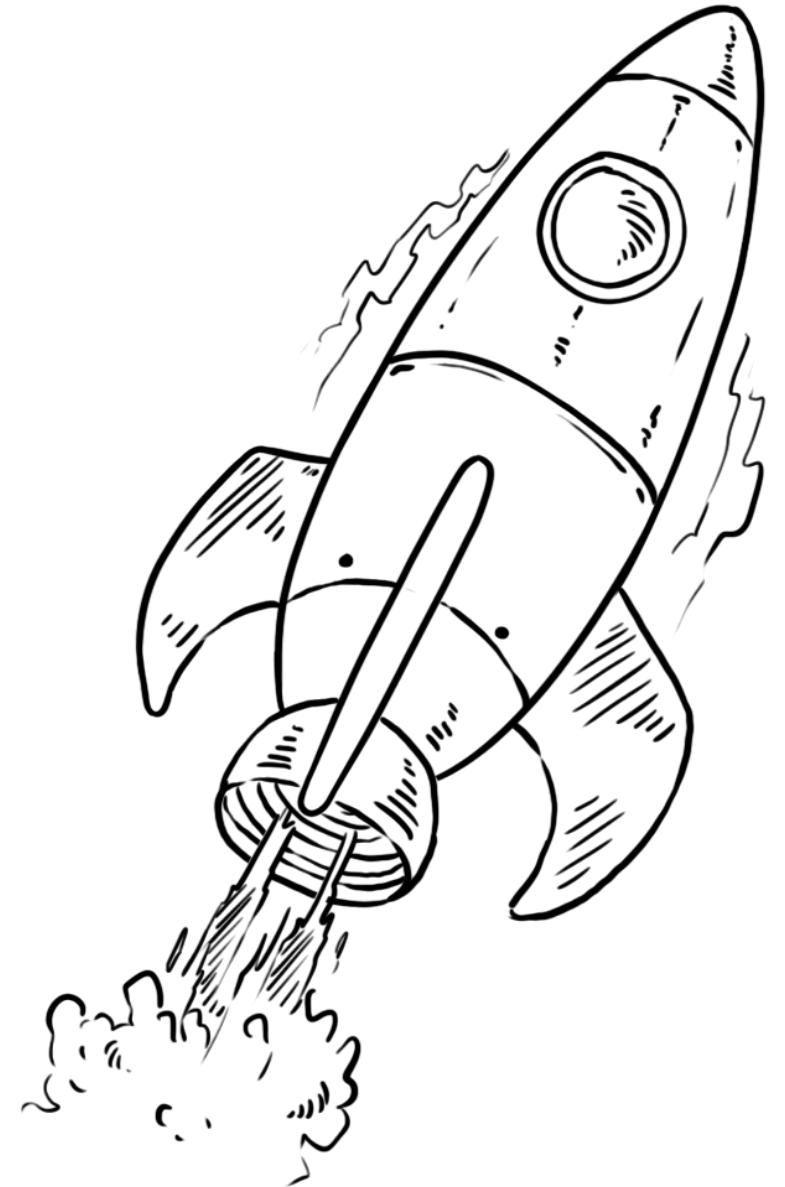
- **Business state:**
  - 500K daily impression
  - Video advertising SDK with 50 Publishers
  - CPM and CPC campaigns
- **Technical State:**
  - Centralized system to answer the requests
  - Estimating CTRs using a simple Bayesian Bernoulli Model
  - Visualizing the historical data and improve algorithm incrementally
- **Cons:**
  - Not scalable
  - Large error in CTR estimation
- **Pros:**
  - Best Performance based advertising platform in its own time





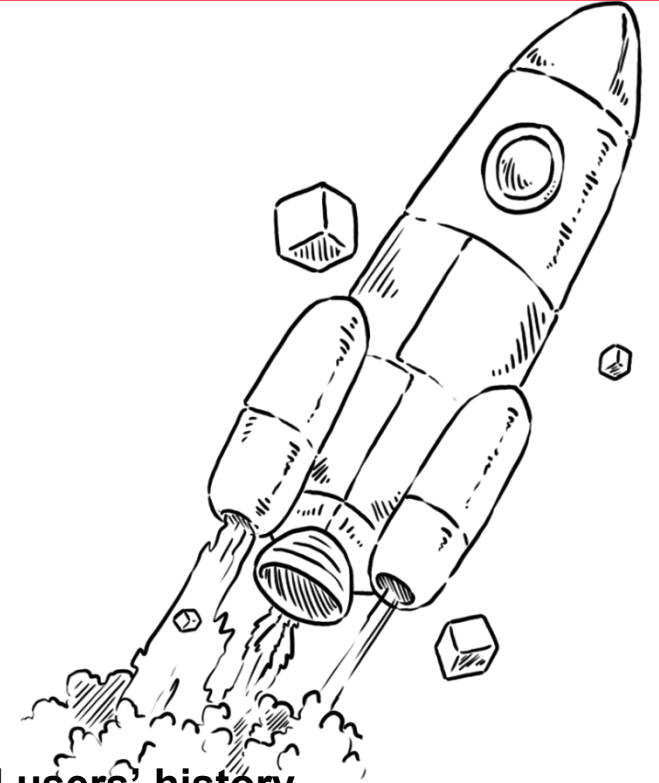
# Tapsell 2<sup>nd</sup> Generation

- **Business state:**
  - 1M+ daily impression
  - 150+ Publishers
  - CPI Campaign
- **Technical State:**
  - Adding multi-level cache to response more requests (still centralized)
  - Estimating CVRs in lower granularity
  - Adding time effect to the CVR estimation model
  - Using feedback data to improve CVR estimations
- **Cons:**
  - Not scalable
  - Large error in CVR estimation for post-click actions
- **Pros:**
  - The Only CPI based advertising platform in its own time



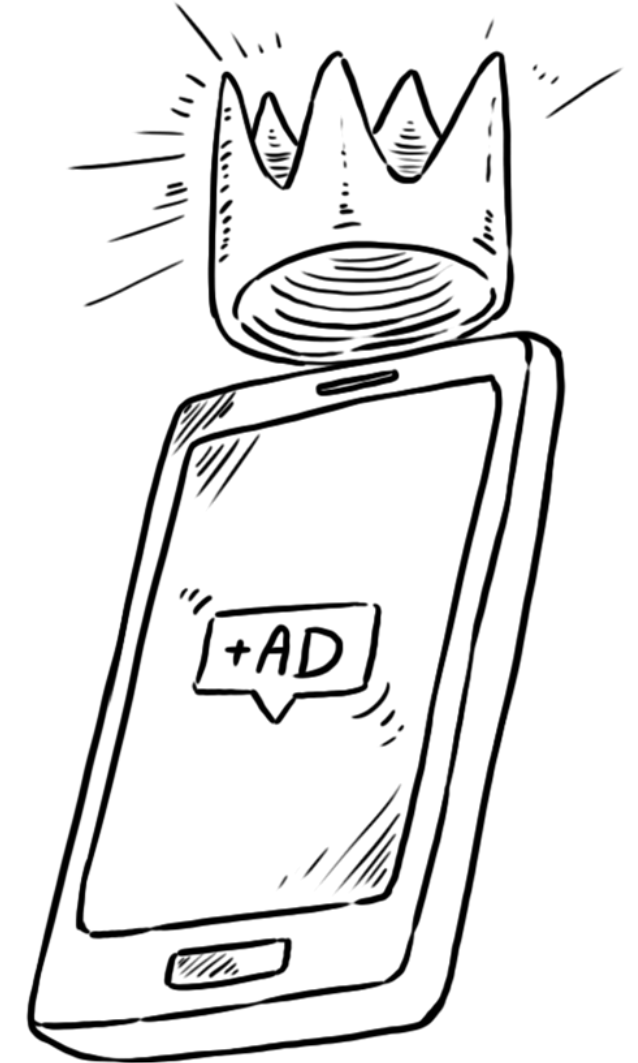
# Tapcell 3<sup>rd</sup> Generation

- **Business state:**
  - 100M+ daily impression
  - 500+ Publishers
  - CPI, CPA Campaign
- **Technical State:**
  - Making the model horizontally scalable in all levels
    - Changing the servers' OS to DCOS
    - Switching to distributed programming platforms (Apache Spark)
    - Switching to distributed Databases (Cassandra, ...)
    - Dockerizing all modules
  - Making the CVR estimation model much more efficient by considering all users' history
- **Pros:**
  - The system is completely scalable and there exist no technical limitation to get the market
  - Best Performance based advertising platform in Iran



# Tapsell 4<sup>th</sup> Generation

- **Business state:**
  - 200M+ daily impression
  - 3500+ Direct Publishers
  - About 2x traffic in comparison to 3<sup>rd</sup> generation
- **Technical State:**
  - Decreasing response time to global standards
  - Connecting to different ad exchanges through RTB
  - Estimating Bid using CVR and other DSPs values
- **Pros:**
  - Be able to easily increase traffic by connecting to ad exchanges



# Current Challenges

- **Improving CVR estimation method**
  - We still have a far way to be optimized in CVR estimation
- **Improving bid estimation algorithm**
  - Bid estimation in competition to other DSPs is still a new challenge for us
- **Making the system more scalable and efficient**
  - Responding to millions of requests per second with our limited resource is still a dream for us

# How to Join Us

- **Co-op Program for B.Sc. students**
  - Learn cutting edge technologies by working in a professional atmosphere
    - Designing, Evaluating and Deploying Large Scale ML Algorithms
    - Distributed Databases and Programming Platforms
    - Cloud Computing technologies
- **Research Topic for M.Sc. and Ph.D. students**
  - Computational Advertising is a hot topic in top conferences such as KDD, WSDM, WWW, ...
    - Real world problems
    - Real Datasets
    - Baseline Methods that can be used to develop more advanced ones
- **Apply for full time or part time job by**
  - Send your resume to [jobs@tapsell.ir](mailto:jobs@tapsell.ir)
  - Fill the form at [jobs.tapsell.ir](http://jobs.tapsell.ir)





**Thank You!**