Large Scale Machine Learning

in Digital Advertising

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

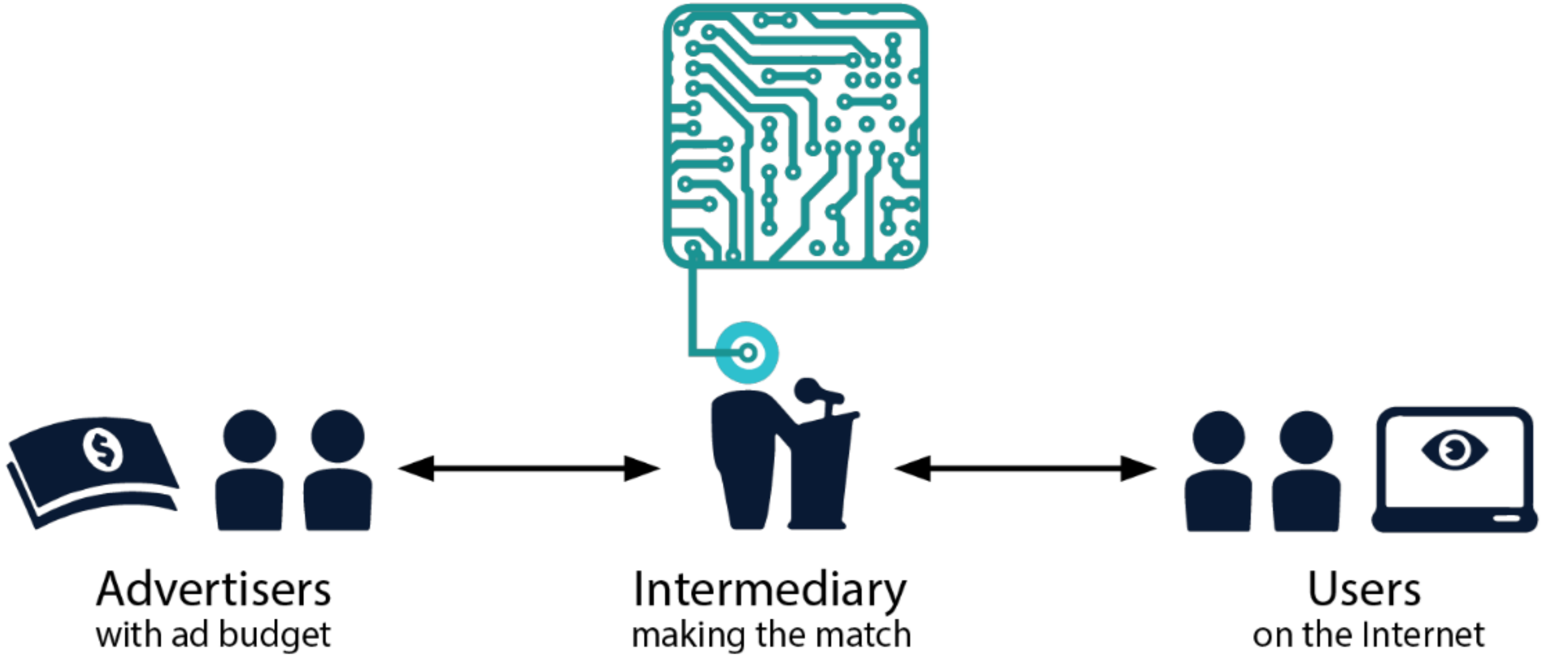
TAPSELL TAGROW media.ad

Outline

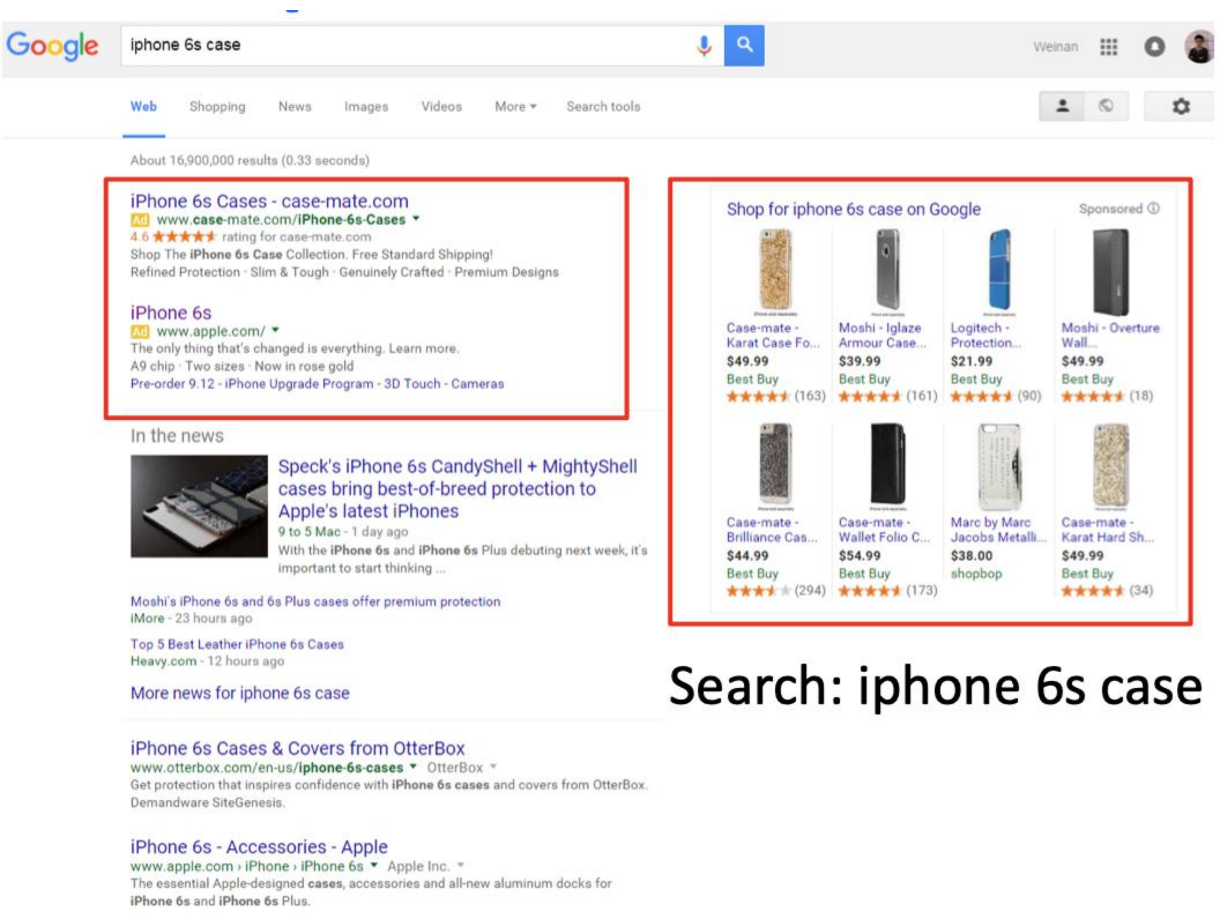
- Digital Advertising
  - Sponsored Search
  - Display Advertising
- RTB Mechanism
- Bid Estimation
  - CVR Estimation
- Other Interesting Issues
- Who We Are?!

Digital Advertising

Conveying advertisers' message to target audience in online media



Sponsored Search

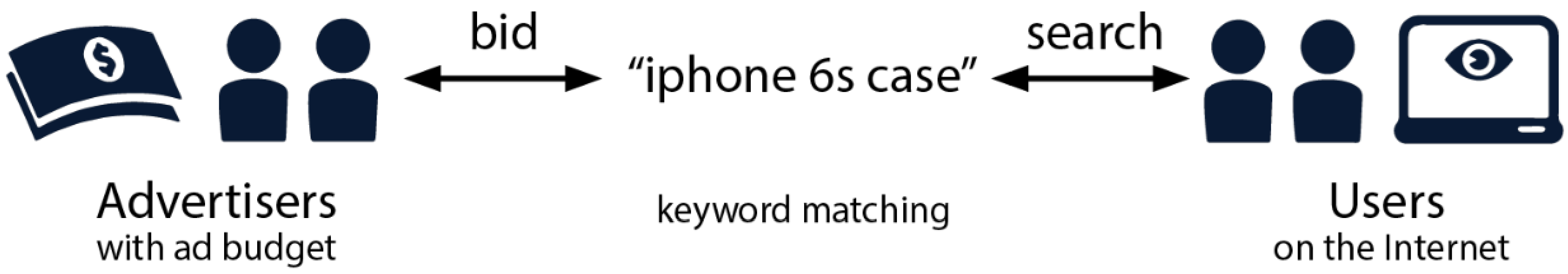


Search: iphone 6s case

Search Engine

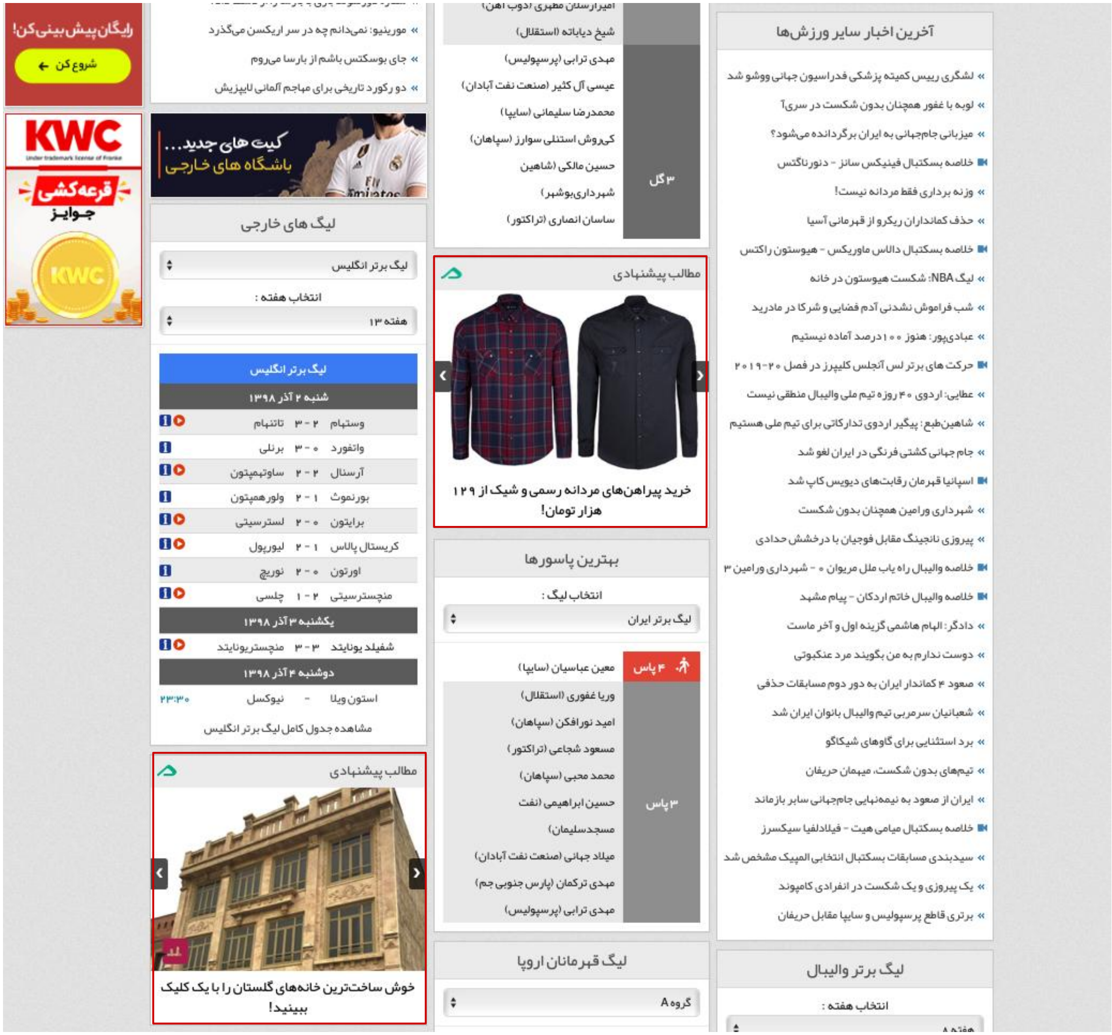
App Market

Sponsored Search

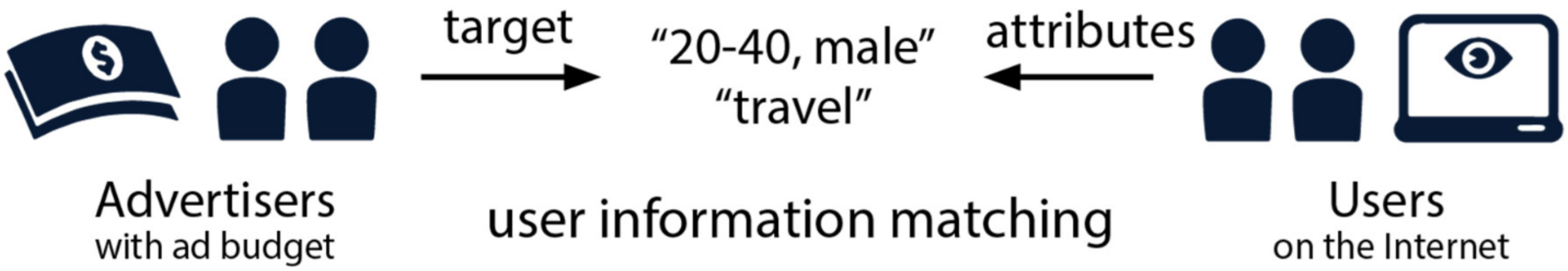


- Advertiser sets a bid price on Keywords
- User searches the keyword
- Search engine or market owner ranks ads and selected the best match

Display Advertising

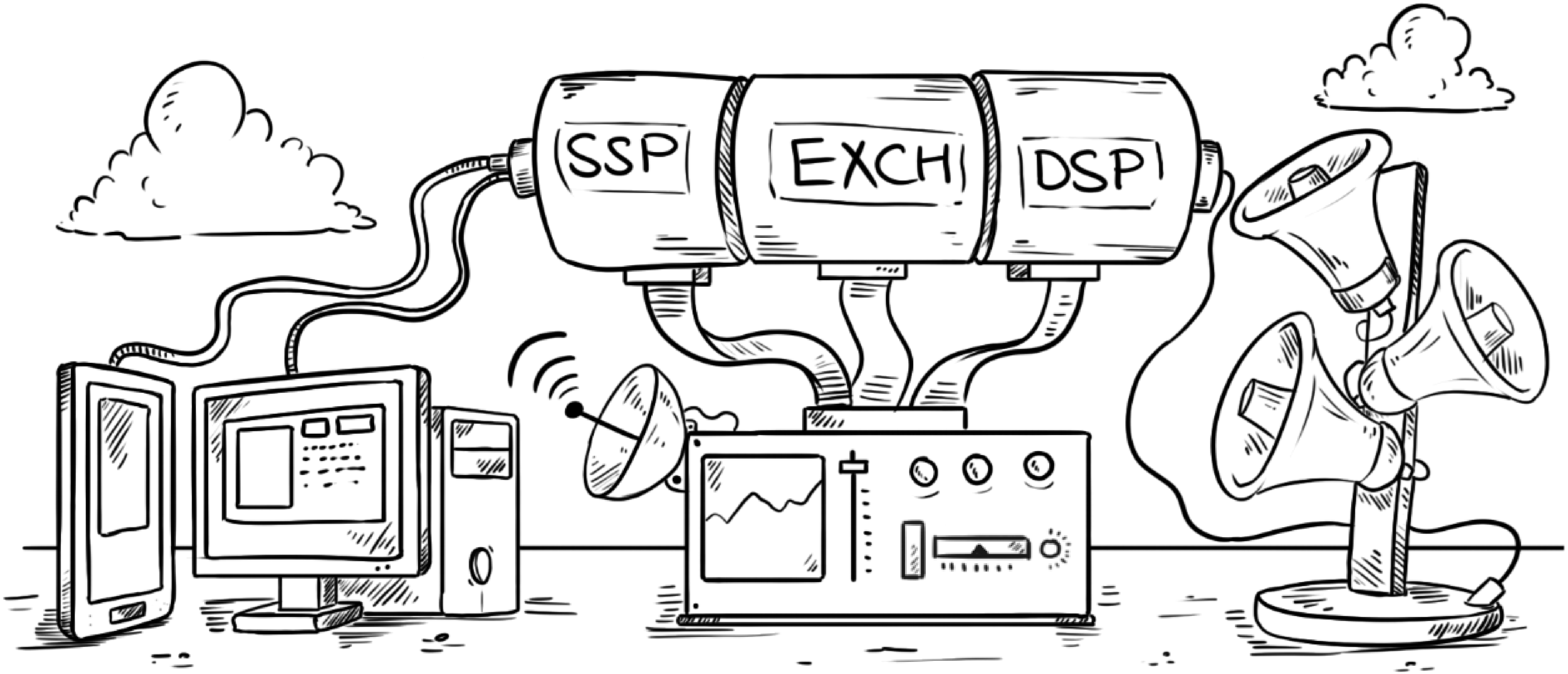


Display Advertising

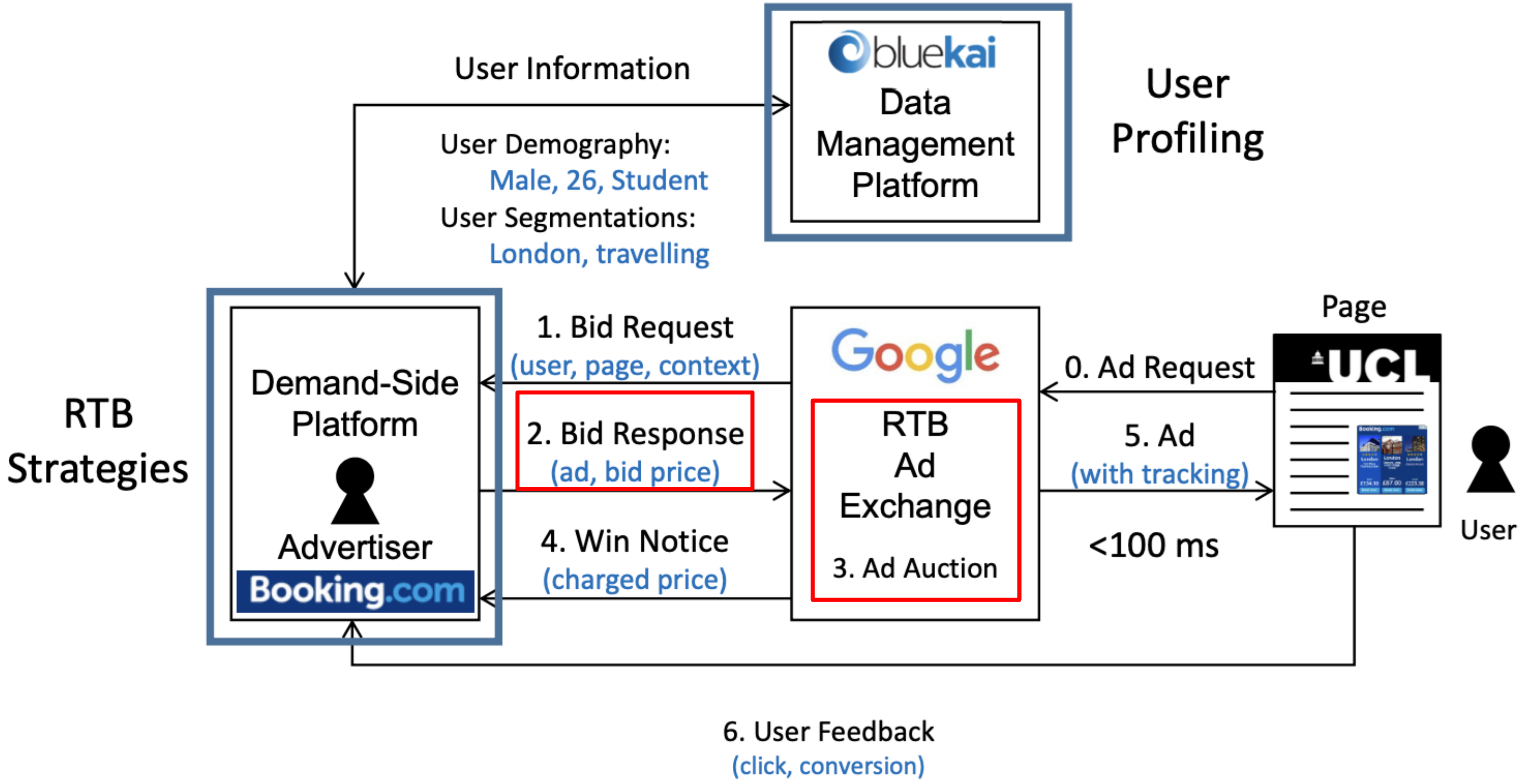


- **Advertiser targets a segment of users**
  - **No matter what the user is searching or reading**
- **Ad Network selects the best ad to show to the user**

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

First Price
Auction

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

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

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=[Weekday=Friday, Gender=Male, City=Shanghai]

x=[0,0,0,0,1,0,0 0,1 0,0,1,0…0]

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**

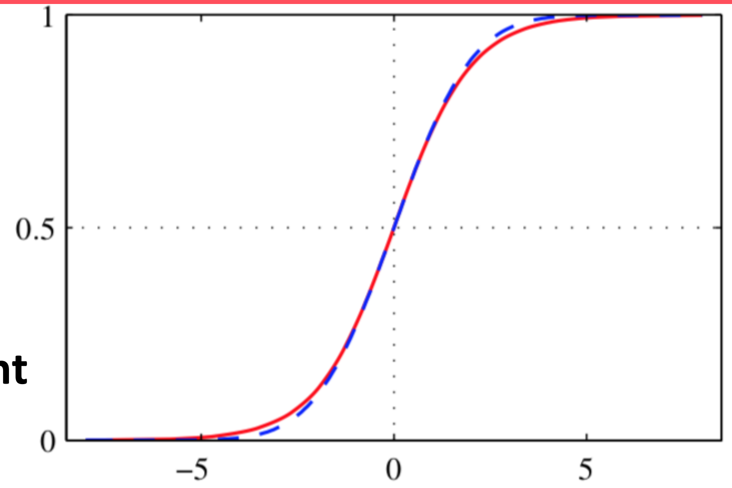
Generalized Linear Models

- 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**

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

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

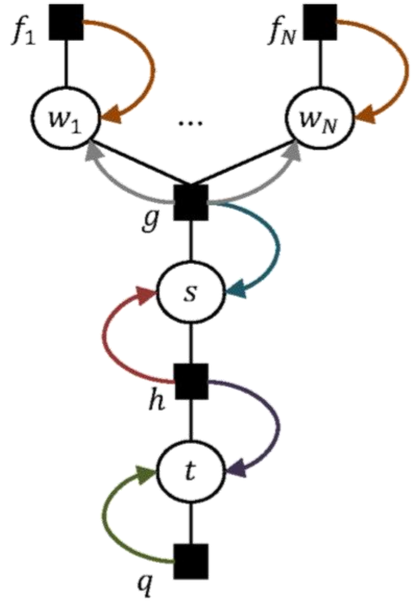


- Bayesian Probit Regression
  - **A fully Bayesian method based on a Gaussian prior over latent weights**
  - **Posterior can be found online using stochastic variational inference**
  - **Bing's Sponsored Search CTR Prediction algorithm**

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

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

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



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

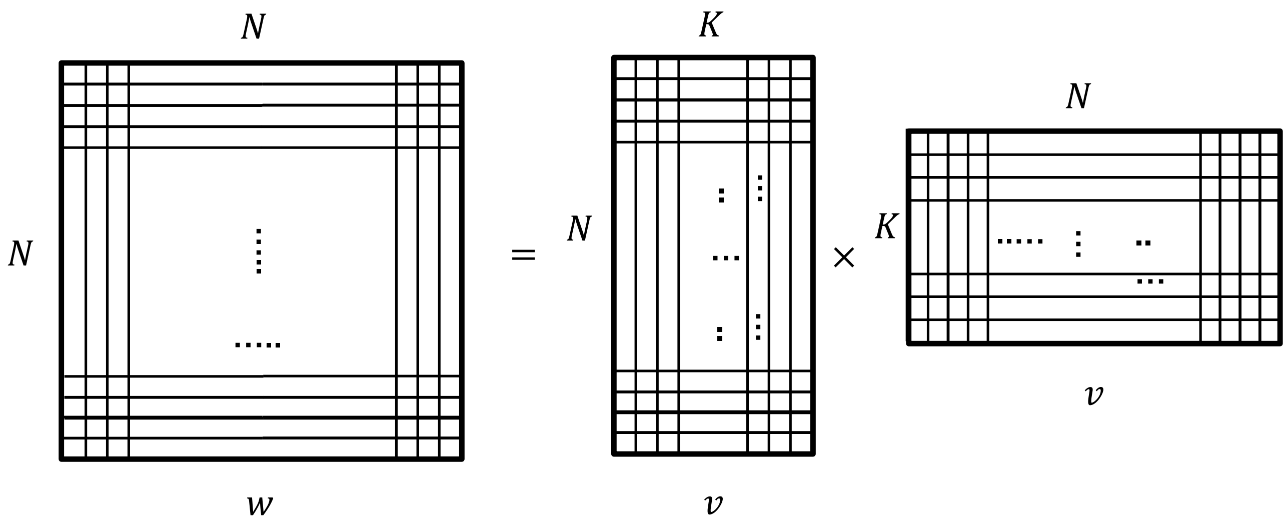
- One way to consider inter-feature correlations is using polynomial kernels

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

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

- Challenge: the model has $\boldsymbol{O}(\boldsymbol{N}^2)$ parameters where $\boldsymbol{N}$ 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 FFMs, 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

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

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

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

	#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

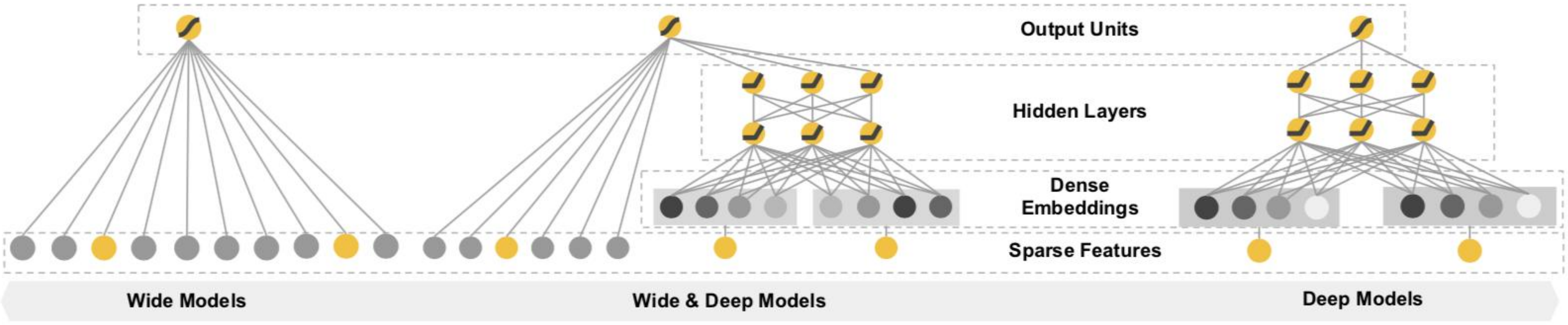
- Memorization of feature interactions through a wide set of cross-product feature transformations are effective and interpretable

- 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

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



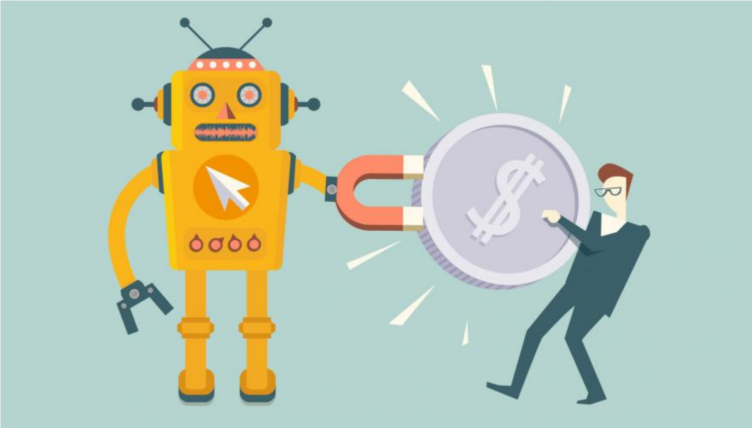
Wide & Deep Model

- 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**

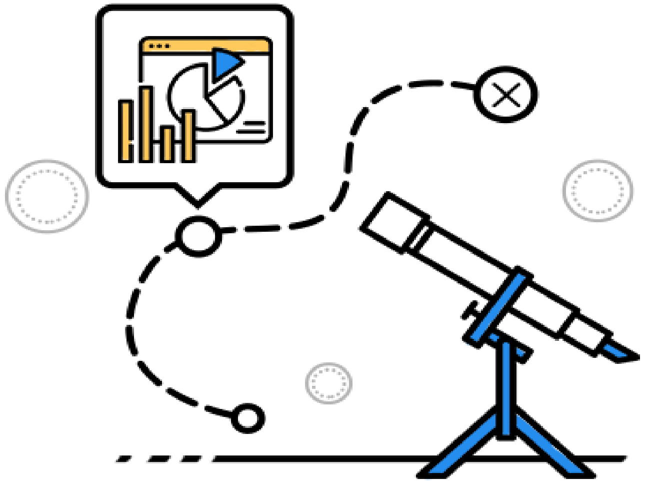
Other Interesting Issues

Frequency Capping

Budget Pacing



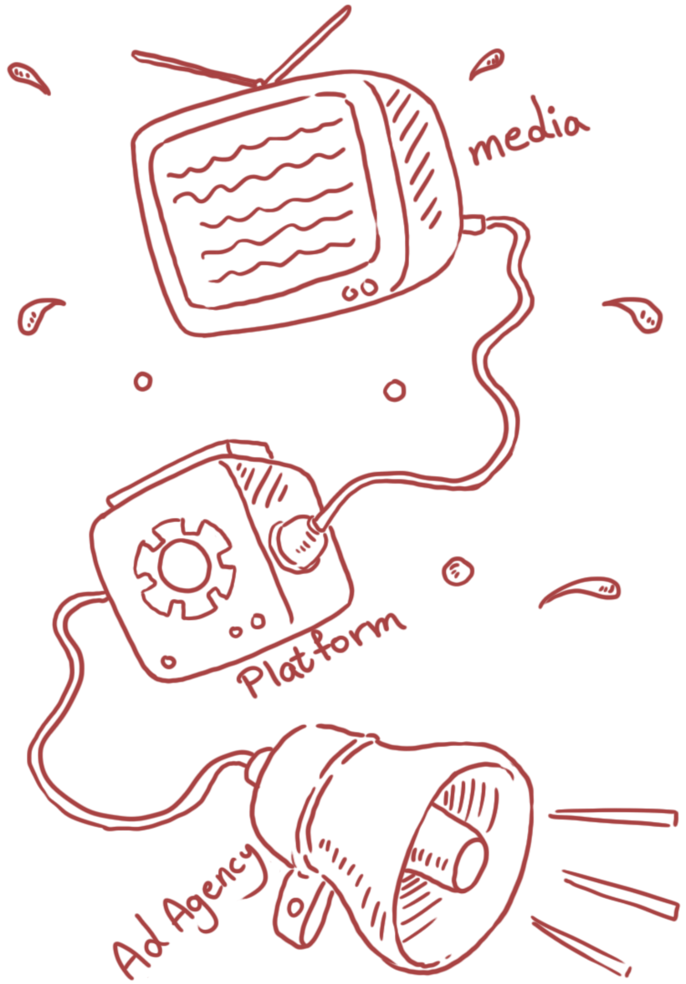
Fraud Detection



Attribution

Who we are

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



Tapsell 1st 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

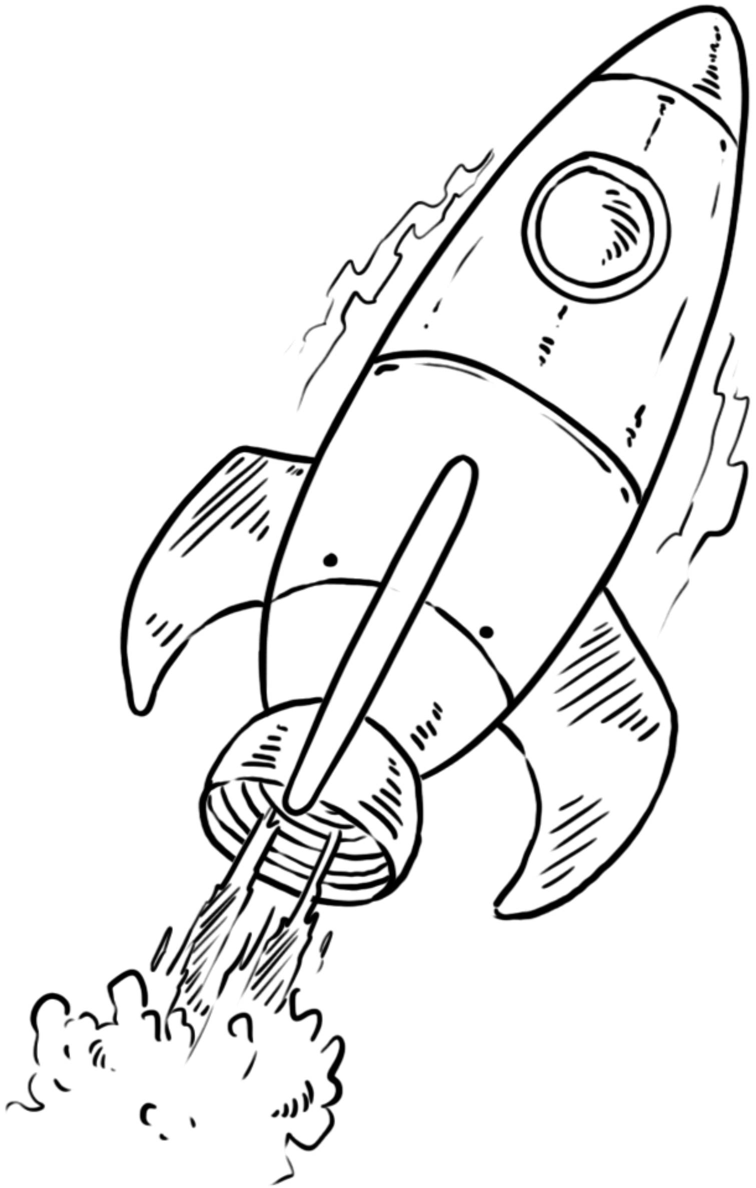
Tapsell 2nd Generation

- **Business state:**
  - 1M+ daily impression
  - 150+ Publishers
  - CPI Campaign

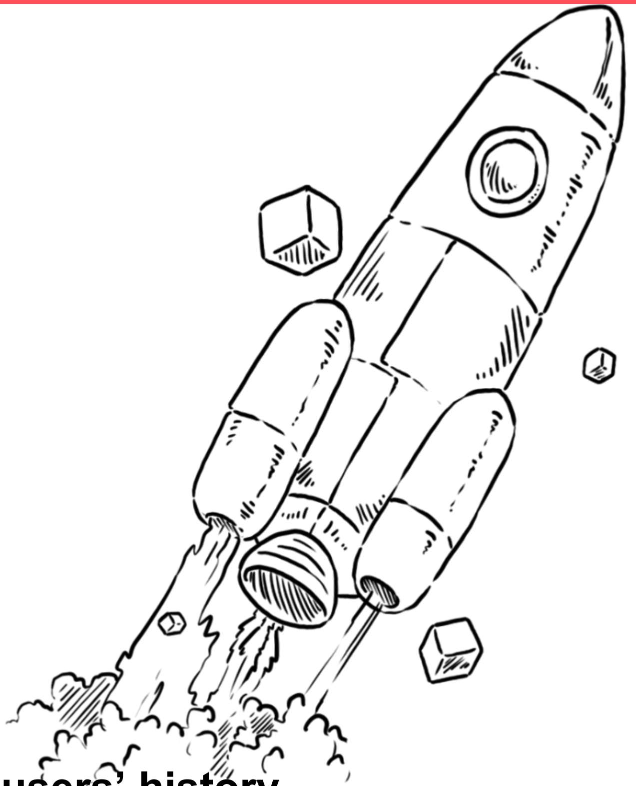
- **Technical State:**
  - Adding multi-level cache to response more requests (still centralized)
  - Estimating CVRs in lower granularity
  - Adding time effect to the CVR estimation model
  - Using feedback data to improve CVR estimations

- **Cons:**
  - Not scalable
  - Large error in CVR estimation for post-click actions

- **Pros:**
  - The Only CPI based advertising platform in its own time



Tapsell 3rd 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 4th Generation

- **Business state:**
  - 200M+ daily impression
  - 3500+ Direct Publishers
  - About 2x traffic in comparison to 3rd 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**
    - **Fill the form at jobs.tapsell.ir**

Thank You!