# SEM Breakeven Bid Case Study

## Team HELLO DATA

Ziang (Allen) Lu
Jiajia (Deborah) Yin
Yi Zhang
Yilun (Tom) Zhang

# Table of Contents

# Background

Capital One → In house Mail Ads → **COST** → High Delivery Cost → Customer

Capital One ← Cost Per Click ← **COST** ← Keyword Search Ads ← Customer

# Problem Formalization

Make Bid → Ad Displayed → User Click → Pay for Clicks

Pay for Clicks → Browse Products

Browse Products → Apply → Approve → Revenue
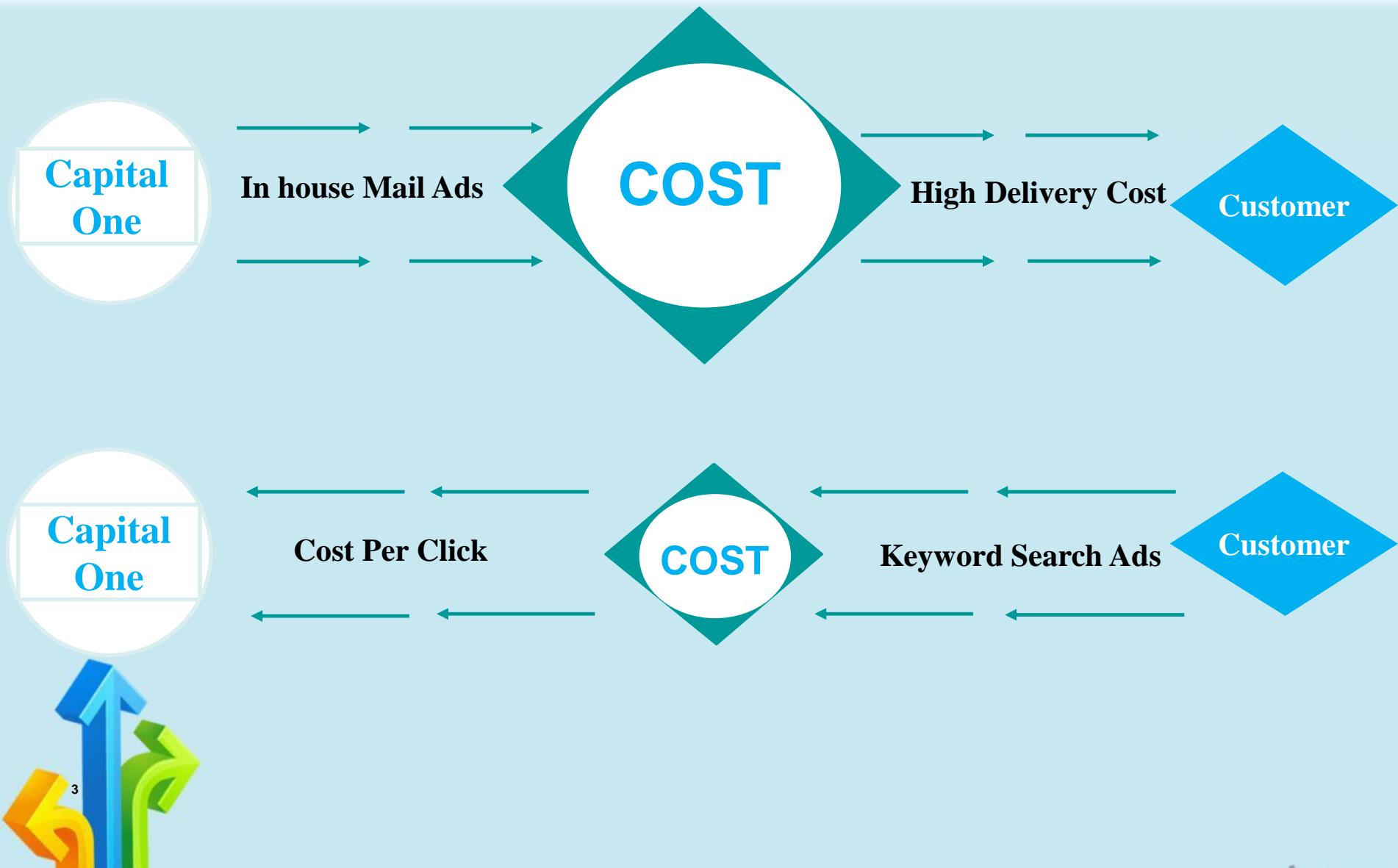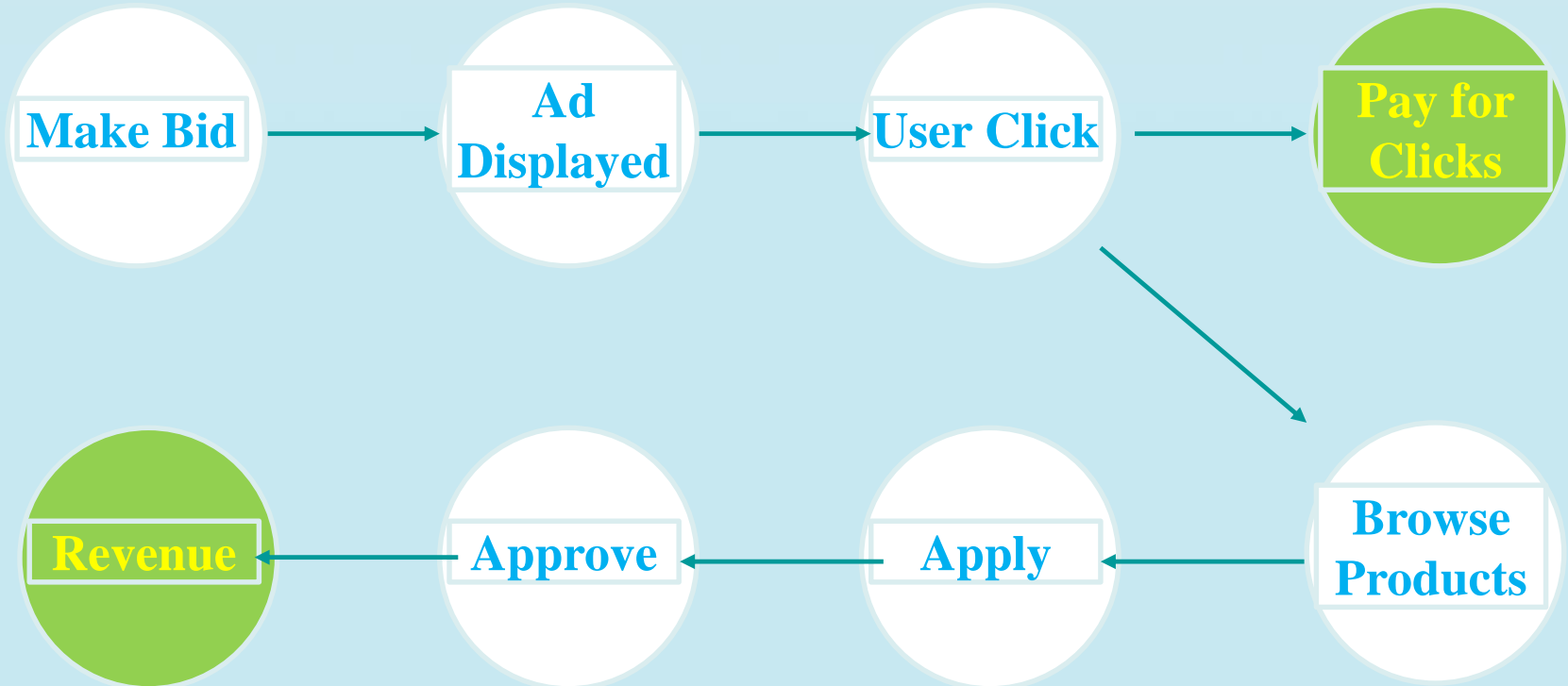
# Problem Formalization

- ✓ *Total Revenue = Product Revenue × Approved Number of Products*
- ✓ *Total Cost = Cost Per Click × Number of Clicks*
- ✓ **Want:** *Total Revenue = Total Cost*
- ✓ *⟺ Breakeven Bid =*
      *Product Revenue × Conversion Rate × Approval Rate*

Questions: <span style="color:red">For each bid,</span>

- ➢ What's its expected Conversion Rate?
- ➢ What's its expected Approval Rate?
- ➢ What's its expected Revenue generated?
- ➢ **Thus what is the appropriated maximum bidding amount?**

# **Statistical Methods – Preprocessing & Transformation**

➢ Selected significant features:

  ➢ ENGN_ID, LANG_ID, DVIC_ID, KEYWD_TXT, MTCH_TYPE_ID, etc.

➢ Separated keyword combinations to single keywords

➢ Grouped all information, using the combination of above features and *SINGLE* keywords as primary key

➢ Calculated Conversion Rate for existing data

➢ Substituted NaN/NA with appropriate value (mostly 0)

| | ENGN_ID | LANG_ID | DVIC_ID | CMPGN_NM | KEYWD_TXT | MTCH_TYPE_ID | IMPRESSIONS | CONDITIONAL_IMPRESSIONS | TOTAL_QUALITY_SCORE | IMPRESSION_TOTAL_RANK | VISITS | CONV_RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | G | E | D | CMPGN9 | KW1 | B | 4.025000 | 4.025000 | 16.895000 | 5.710500 | 0.135000000 | 0.000000000 |
| 2 | G | E | D | CMPGN6 | KW10 | E | 3.176471 | 3.176471 | 3.176471 | 14.364706 | 0.000000000 | 0.000000000 |
| 3 | G | E | D | CMPGN9 | KW100 | B | 3.109677 | 3.109677 | 7.225806 | 10.270323 | 0.006451613 | 0.000000000 |
| 4 | G | E | D | CMPGN1 | KW101 | B | 6.500000 | 6.500000 | 19.500000 | 21.600000 | 0.000000000 | 0.000000000 |
| 5 | G | E | D | CMPGN2 | KW102 | B | 12.422535 | 12.422535 | 99.711268 | 16.935211 | 0.739436620 | 0.009523810 |
| 6 | G | E | D | CMPGN2 | KW102 | E | 1.000000 | 1.000000 | 9.000000 | 1.000000 | 0.000000000 | 0.000000000 |

# Statistical Methods -- Conversion Rate Modelling

➢ Use classification and regression tree (CART) method to model Conversion Rates for biddings containing a *SINGLE* keyword

➢ Predict Conversion Rates of biddings with multiple keywords by weighting each *SINGLE* keyword based on their relationships with the bidding's corresponding CMPGN_NM (Ad_ID)

(See Appendix A&B)

➢ E.g. $p = proportion\ of\ keywords\ that\ are\ frequent\ in\ CMPGN$
$CONV\_RATE(KEYWD\_TXT) =$

$$\sum_{\{KW_i \in CMPGN_j\}} CONV\_RATE(KW_i) \times \sqrt{p} +$$
$$\sum_{\{KW_i \notin CMPGN_j\}} CONV\_RATE(KW_i) \times (1 - \sqrt{p})$$

# Statistical Methods -- Conversion Rate Modelling & Evaluation

| Model Name | Root Mean Squared Error |
|---|---|
| CART | 0.0594 |
| Decision Table | 0.0729 |
| KStar | 0.0784 |
| Decision Stump | 0.0785 |
| M5Rules | 0.0808 |
| Linear Regression | 0.0808 |
| ZeroR | 0.0811 |
| Regression By Discretization | 0.0811 |
| Multi Scheme | 0.0811 |
| REP Tree | 0.0812 |
| M5P | 0.0828 |
| IBk | 0.1051 |

➢ CART is a decision tree based method that calculates the average of known data according to some of their common features to make predictions by placing new data into the appropriate branch of the decision tree

➢ Comparing to other models run in Weka, CART has the best root mean squared error (from 10-fold-cross-validation)

* For Weka Software, see Appendix C

# Statistical Methods -- Approval Rate Modelling

➢ $AR(Prod\ i) =$
   $Approved\ Prod\ i\ Applications\ /\ Total\ Prod\ i\ Applications$

➢ For each keyword, extract the probability distribution over the 6 products it potentially leads to

➢ Obtain the Approval Rate corresponding to each keyword as the expected AR over the products

➢ $AR(KW) = \sum_i \Pr(Prod\ i\ |\ KW) \cdot AR(Prod\ i)$

➢ Approval Rate of bids with multiple keywords are averaged over its single keywords

# **Statistical Methods -- Product Revenue Modelling**

➢ $PR(Prod\ i) =$
   $Prod\ i\ Total\ Revenue\ /\ Prod\ i\ Approved\ Applications$

➢ For each keyword, extract the probability distribution over the 6 products it potentially leads to

➢ Obtain the Product Revenue corresponding to each keyword as the expected PR over the products

➢ $PR(KW) = \sum_i \Pr(Prod\ i\ |\ KW) \cdot PR(Prod\ i)$

➢ Again, Product Revenue of bids with multiple keywords are averaged over its single keywords

# Statistical Methods  -- Validation Set Prediction

➢ Obtained Conversion Rate, Approval Rate and Product Revenue for all single keywords

➢ Estimated the above three values for each bidding in the validation set

➢ Thus, determined the breakeven bid (max bid) for each bidding

➢ Note that there are **22** new *SINGLE* keywords and **286** new keyword combinations in the validation set, thus using the *SINGLE* keywords in modelling reduces the amount of unknown information

➢ Performance of prediction will be improved when more data containing these keywords are gathered

# **Statistical Methods** **-- Result Visualization**



Distribution of Breakeven Conversion Rate
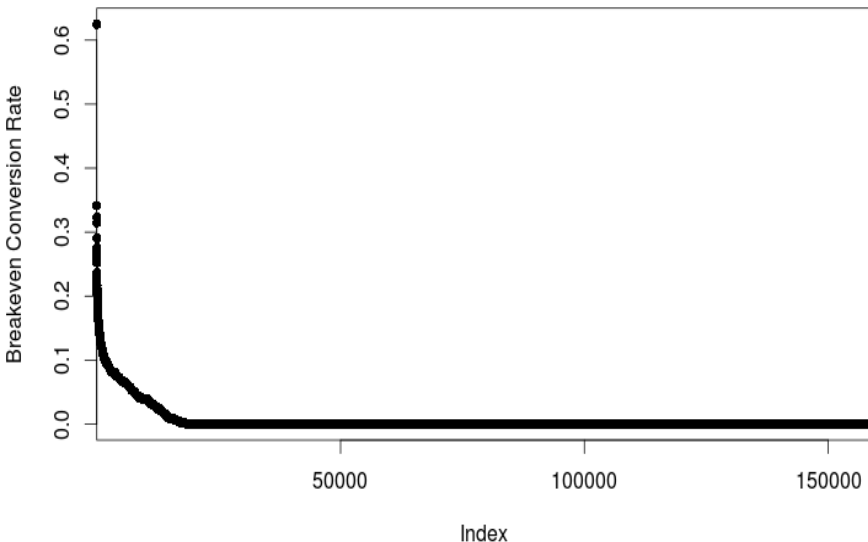
Distribution of Breakeven Bid
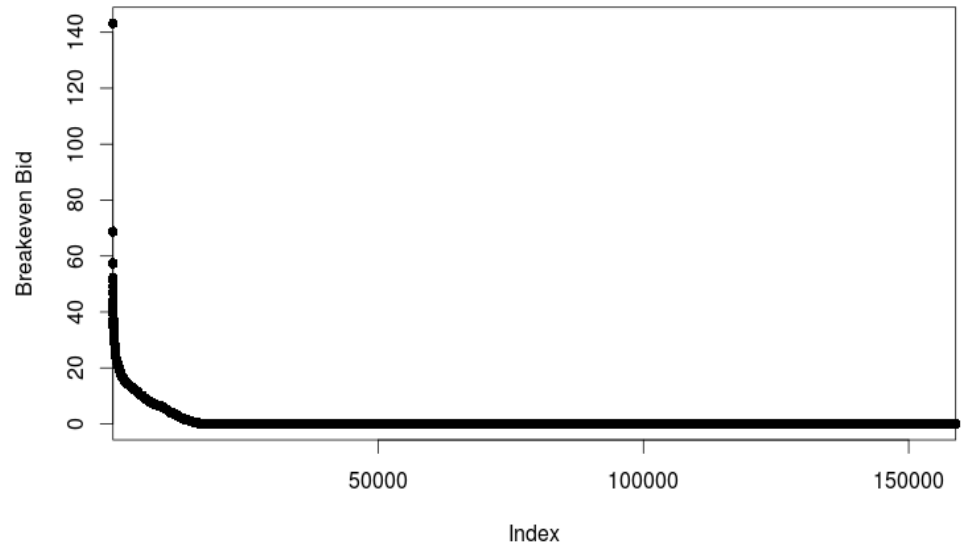
➢ The Conversion Rate and the Breakeven Bid have similar distributions

➢ 137466 (86.53%) zeros in predicted Conversion Rate

➢ 138143 (86.95%) zeros in predicted Breakeven Bid

# Statistical Methods  -- Discussion

➢ Due the variety in KEYWD_TXT, some models including linear regression and logistic regression are either very ineffective or unable to run in reasonable time

➢ Due to the nature of this problem which results in many conversion rates being zero, our model tends to be positively biased and overestimate the Conversion Rate when it is actually low (i.e. 0)

➢ In the future, meta algorithms such as bagging or boosting can be used to improve the prediction accuracy

➢ Programmatic feature selections (PCA, forward/ backward/ stepwise selection) can be conducted to reduce the dimensionality of the problem as well as potentially improve the accuracy

# Business Implications  -- Long Term Trend



US Online Ad Spending

32.0 Billion — 2011
39.5 Billion — 2012
46.5 Billion — 2013
52.8 Billion — 2014
57.5 Billion — 2015
62.0 Billion — 2016

Source: IAB

## Online Advertising Market Share



■ Google ■ Yahoo$Microsoft ■ facebook ■ Aol. ■ Others
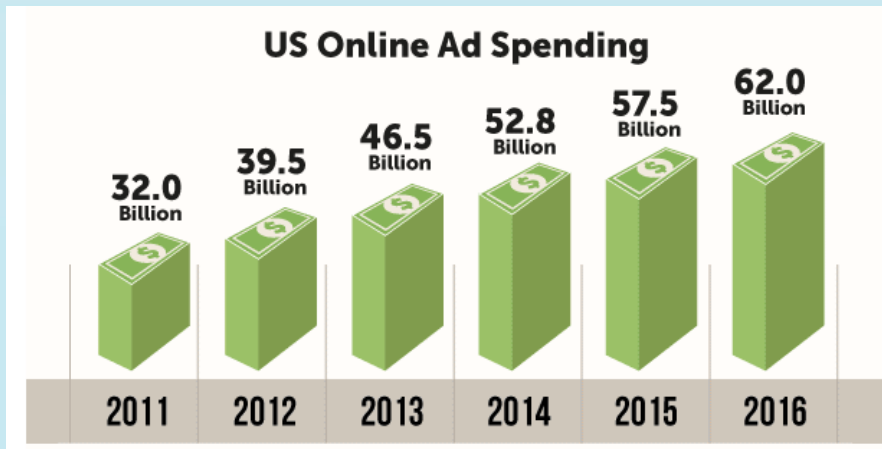
Source: IAB

❖ Increased growth rate at online ad spending market each year

❖ Keyword search occupies half of the market shares
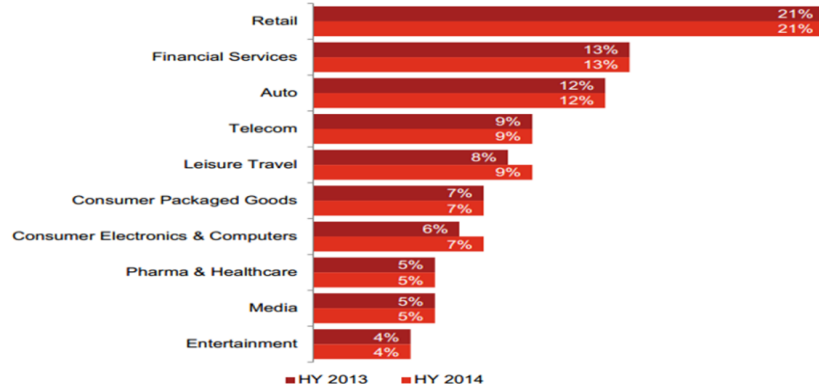
❖ Cost-efficient investments to boost business



Ad Formats Percentage

Google Bing — Search 49%
Buy Me — Display Banner Ads 23%
Classified 8%
Lead Generation 5%
Digital Videos 6%
Rich Media 5%
Sponsorships 3%
Email 1%

Source: IAB

# Business Implications -- Finance Industry Outlook



Internet ad revenues by major industry category*, year to date: 2013 vs. 2014

Source: IAB/PwC Internet Ad Revenue Report, HY 2014



US Financial Services Industry Digital Ad Spending, 2011-2017

billions, % of total digital ad spending and % change
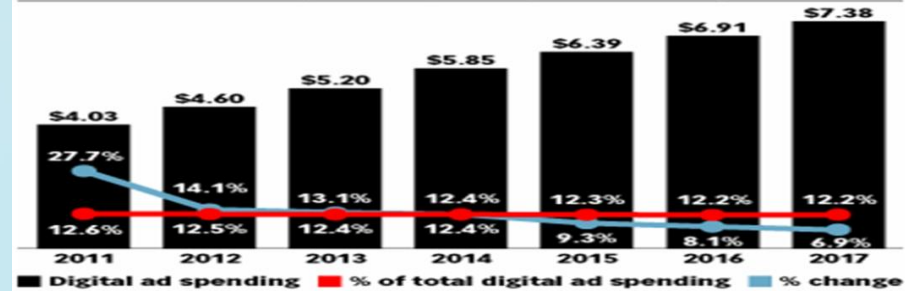
Note: CAGR (2012-2017)=9.9%; includes advertising that appears on desktop and laptop computers as well as mobile phones and tablets, and includes all the various formats of advertising on those platforms; data through 2012 is derived from IAB/PwC data
Source: eMarketer, June 2013

- ❖ Second largest industry spent in online ads market
- ❖ Spending amount grows dramatically each year
- ❖ Proper bidding strategy is essential in winning the competitive markets



US Digital Ad Spending Estimates, by Industry and Objective
% share of total / $ billions
May 2014
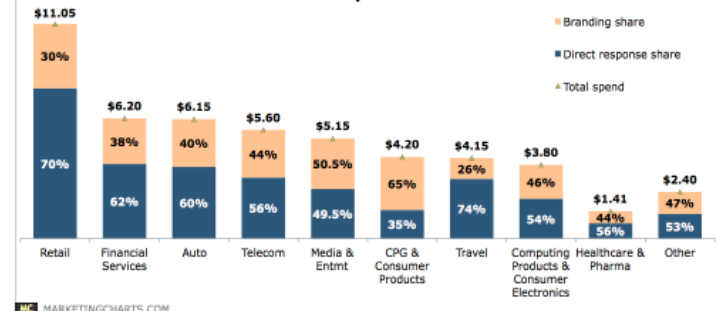
Source: eMarketer

Source: IAB

# Appendix A
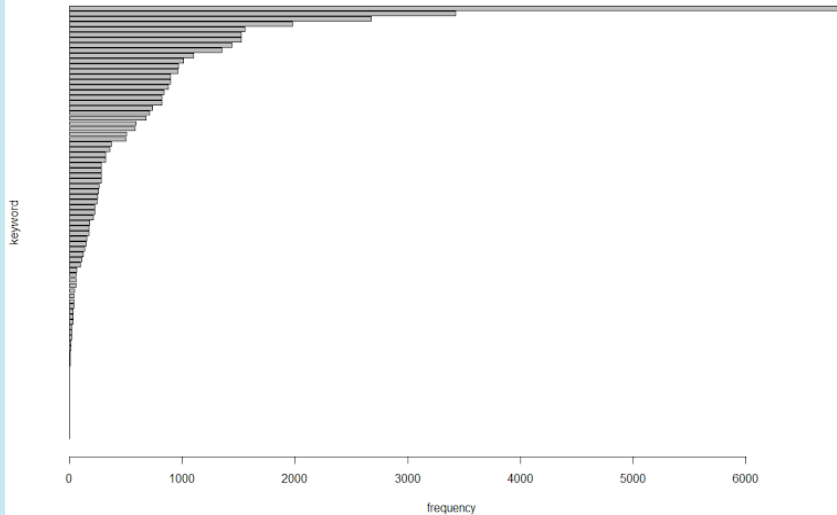## -- Table for Keyword Frequency in Campaign (Ads Group)

| CMGPN1 | | CMPGN2 | | CMPGN3 | | CMPGN4 | | CMPGN5 | | CMPGN6 | | CMPGN8 | | CMPGN9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KW | Freq | KW | Freq | KW | Freq | KW | Freq | KW | Freq | KW | Freq | KW | Freq | KW | Freq |
| KW178 | 6852 | KW104 | 11601 | KW104 | 102 | KW178 | 9183 | KW178 | 5142 | KW97 | 8697 | KW195 | 1200 | KW178 | 19206 |
| KW121 | 3424 | KW382 | 9262 | KW382 | 93 | KW426 | 4999 | KW587 | 4029 | KW121 | 7158 | KW25 | 607 | KW121 | 10865 |
| KW587 | 2677 | KW178 | 2266 | KW124 | 84 | KW121 | 4471 | KW121 | 3856 | KW178 | 7138 | KW121 | 495 | KW587 | 9105 |
| KW306 | 1977 | KW97 | 2242 | KW178 | 42 | KW587 | 4284 | KW340 | 3333 | KW92 | 6877 | KW178 | 487 | KW340 | 6795 |
| KW97 | 1556 | KW92 | 2159 | KW340 | 19 | KW97 | 2376 | KW563 | 1709 | KW288 | 6379 | KW97 | 233 | KW97 | 4867 |
| KW324 | 1523 | KW121 | 2053 | KW196 | 17 | KW62 | 2267 | KW125 | 1585 | KW340 | 4104 | KW515 | 229 | KW124 | 4544 |
| KW560 | 1522 | KW340 | 1881 | KW92 | 11 | KW340 | 1765 | KW53 | 1510 | KW456 | 3476 | KW340 | 209 | KW567 | 4228 |

# Appendix B
## -- Plots for Keyword Frequency in Campaign (Ads Group)
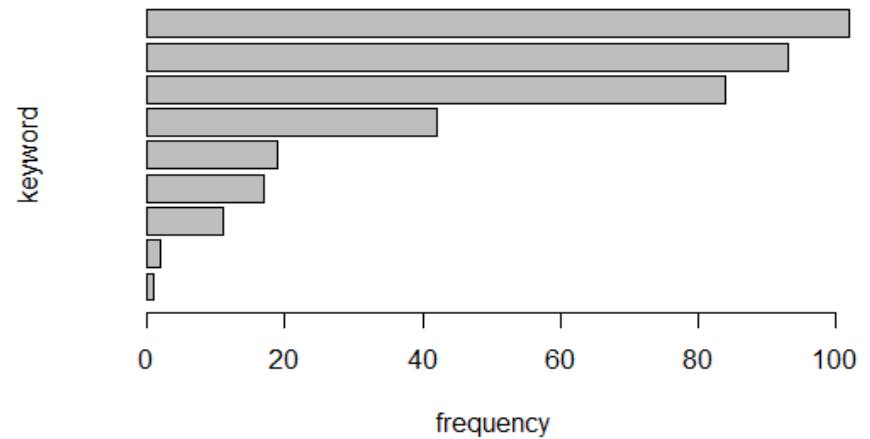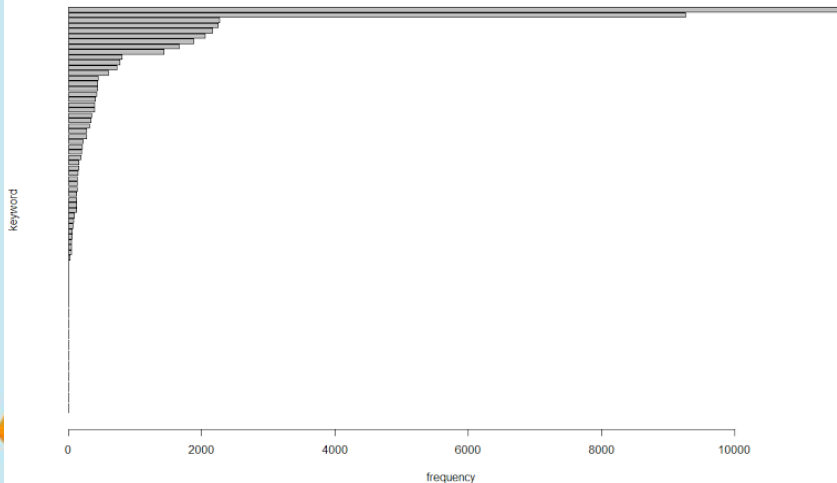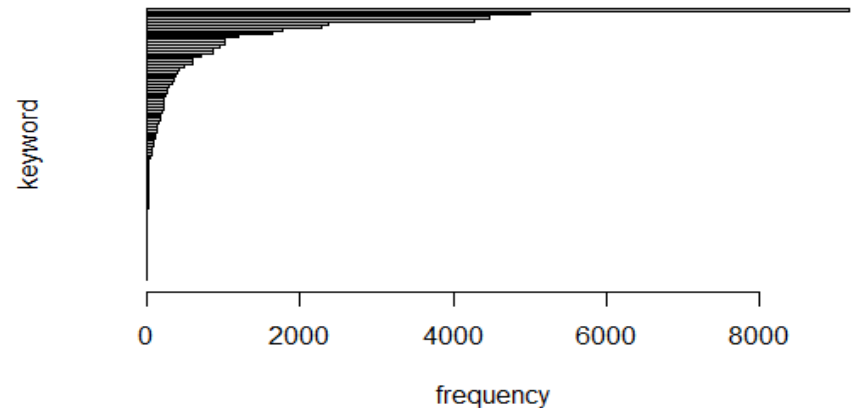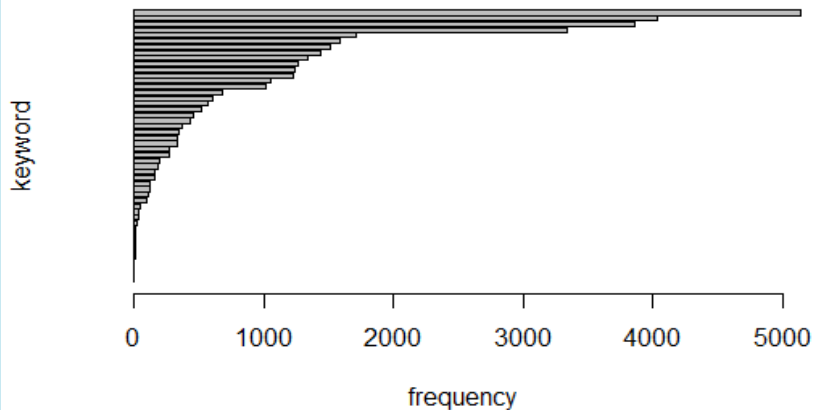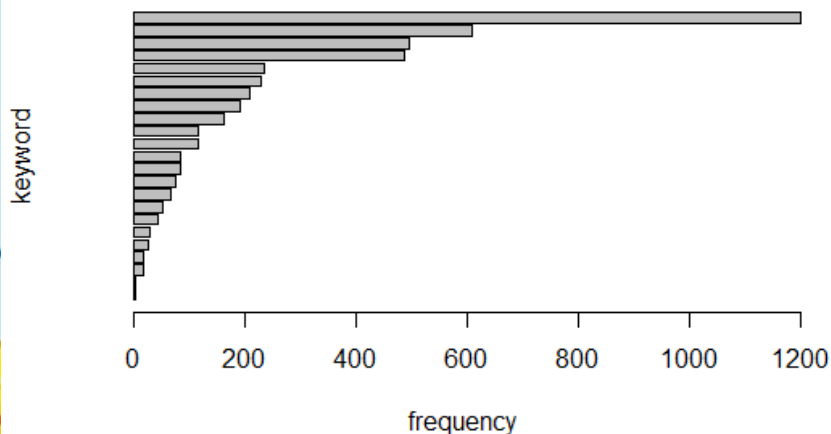
# Appendix B
## -- Plots for Keyword Frequency in Campaign (Ads Group)
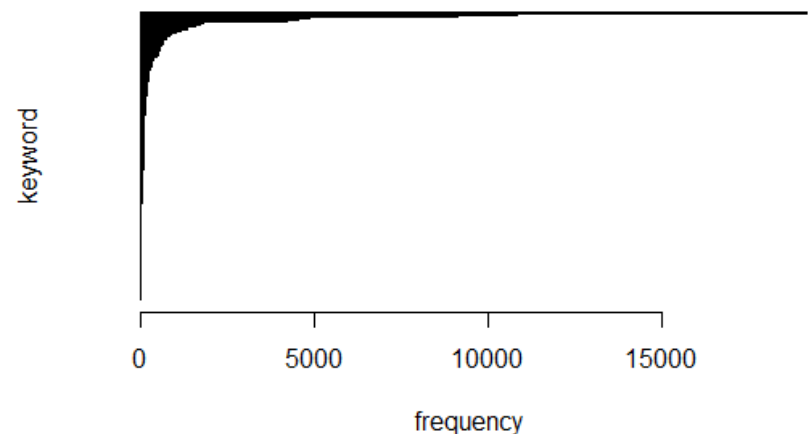
# Appendix C
## -- Weka Software

➢ Please see attached Word document for Weka Software instruction