# 企业大数据管理与应用的技术挑战

冯是聪 (秒针公司, 技术总监 )

2013/08/18

秒针系统
Miaozhen Systems™

# 大纲

- 管理和应用企业大数据的技术挑战
- 企业大数据应用案例分享
- 2013秒针首届RTB算法大赛
- 秒针公开数据集

# 认识秒针

# 认识秒针

秒针系统™
Miaozhen Systems™

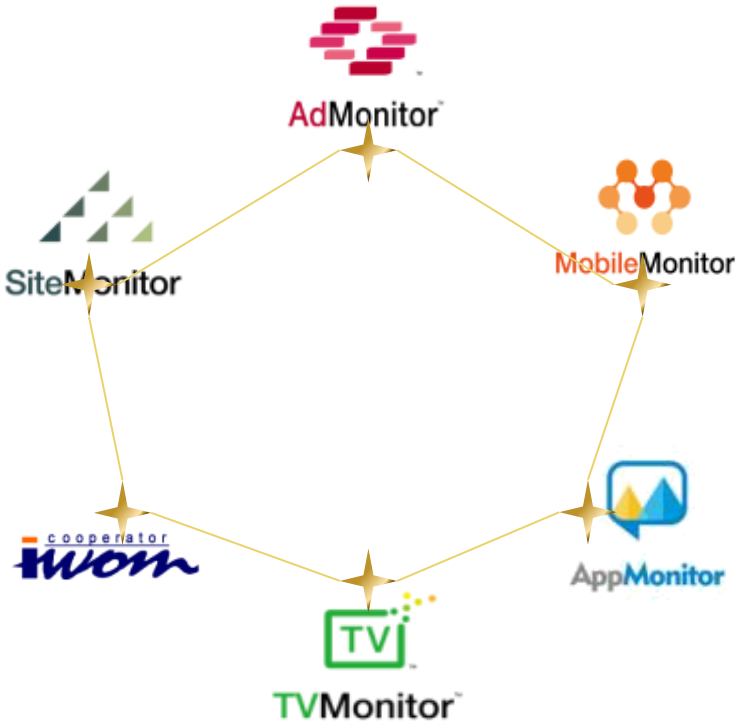
中国领先的第三方广告技术公司

“我们为营销者和媒体经营者提供满足其需求的产品和解决方案，以客观的数据、创新的技术和公正的立场为基础，帮助营销者实现广告投资回报最大化，帮助媒体经营者更好地提升广告资源价值，促进广告市场的繁荣与发展。”

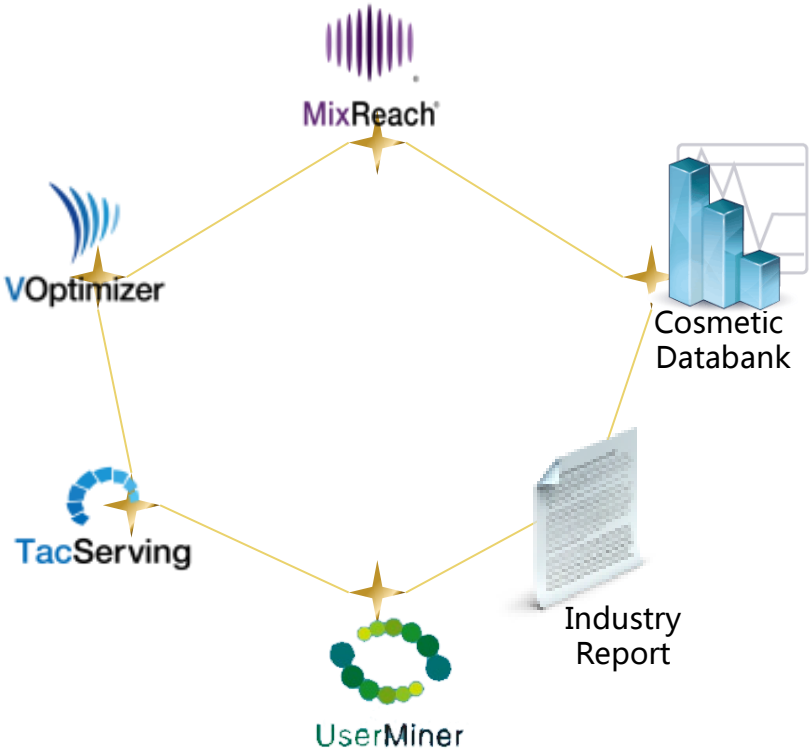
- 成立于2006年
- 北京、上海、广州、新加坡四地联动
- 日均1000亿广告请求处理能力领先行业
- 云计算、云存储、人工智能三大核心技术
- 超过70%的国际品牌百强选择秒针
- 国家高新技术企业和双软认证企业
- 拥有专利和软件著作权超过21项
- 在中国内地、台湾、日本、澳大利亚等市场提供产品和服务
- 已完成三轮融资

# 秒针公司提供的产品与服务 (1)

## 监测

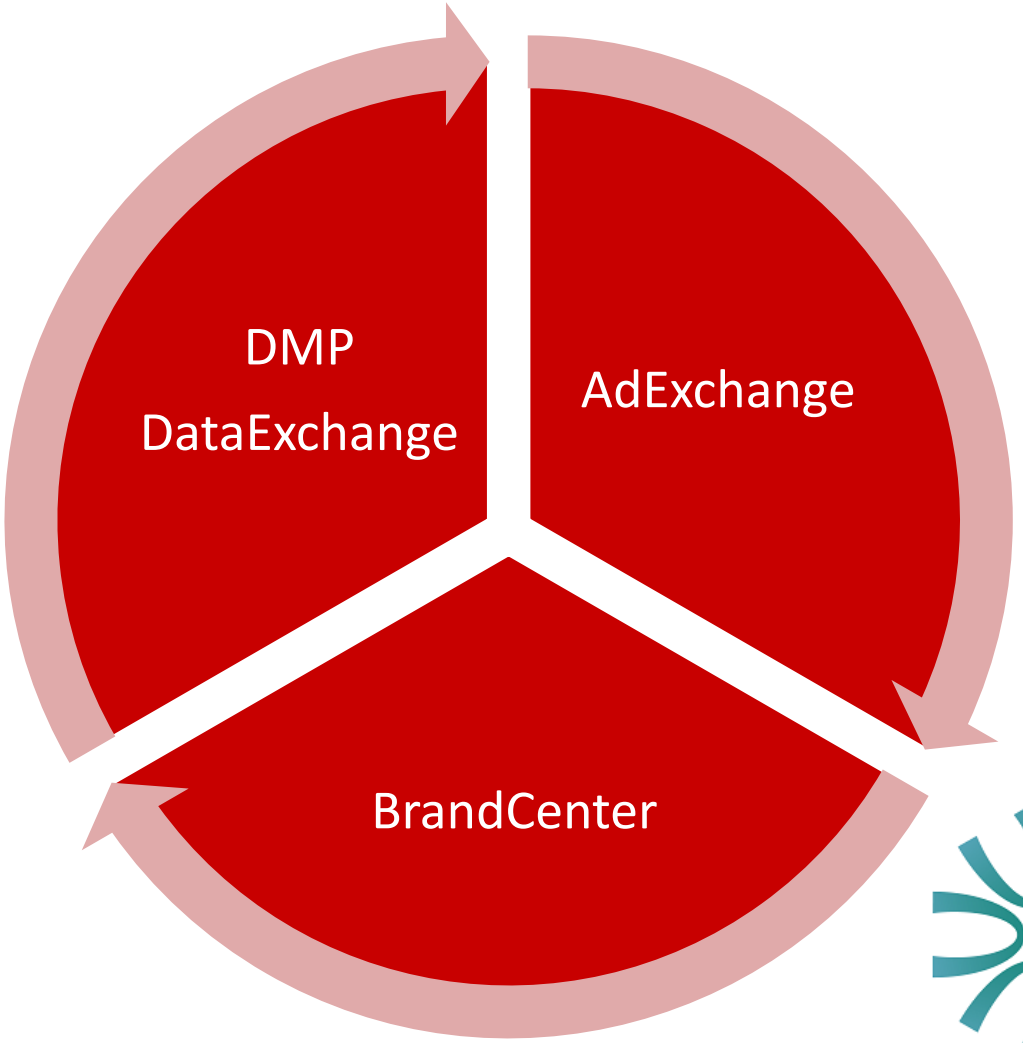



## 优化



# 秒针公司提供的产品与服务 (2)

UserMiner™




MAXchange

Miaozhen Systems

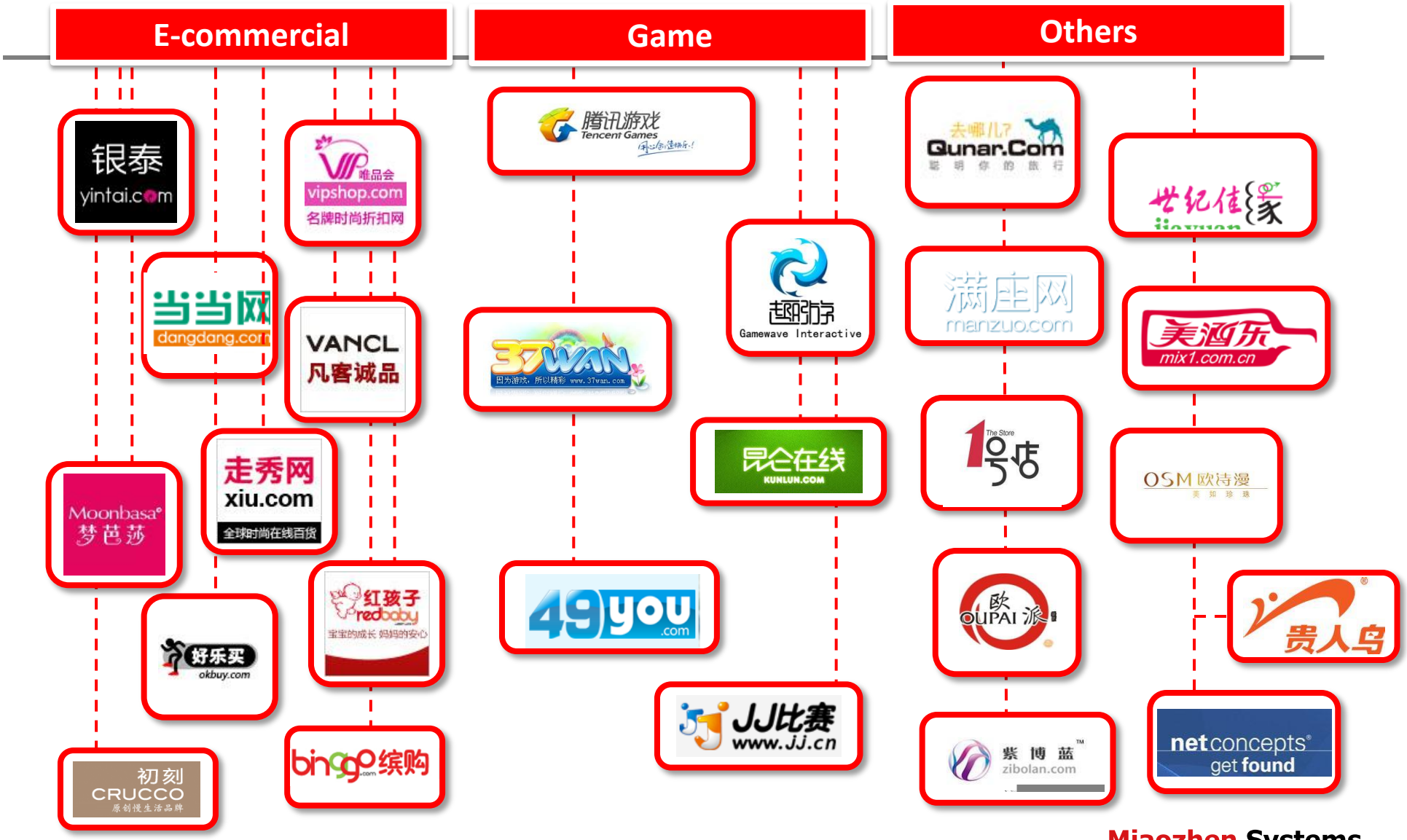
# 我们的客户


Miaozhen Systems

# 我们的客户


MINDSHARE
mediaedge:cia
Active Engagement
maxus
MEDIACOM
People First Better Results
ZenithOptimedia
实力传播
OMD
Starcom MediaVest
GROUP
Carat
neo@Ogilvy
AGENDA
wwwinsCONSULTING
hylink
华扬联众
Vizeum
Connection
hylink
ADVERTISING
OpenTide
China
新好耶
Allyes
动力飞扬
POWER FLYING

# 我们的客户

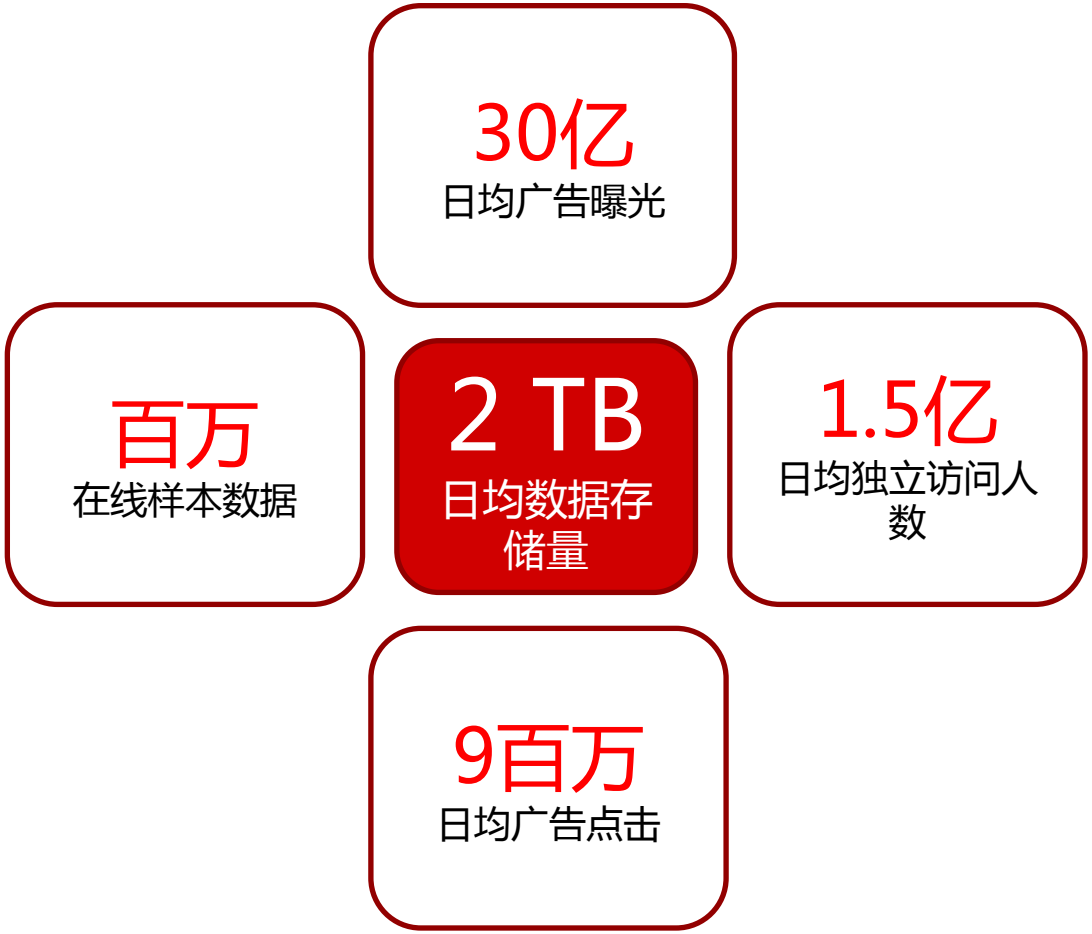

E-commercial
Game
Others
银泰 yintai.com
唯品会 vipshop.com 名牌时尚折扣网
当当网 dangdang.com
VANCL 凡客诚品
走秀网 xiu.com 全球时尚在线百货
Moonbasa 梦芭莎
好乐买 okbuy.com
红孩子 redbaby
初刻 CRUCCO
bingo 缤购
腾讯游戏 Tencent Games
趣游 Gamewave Interactive
37WAN
昆仑在线 KUNLUN.COM
49you.com
JJ比赛 www.JJ.cn
去哪儿? Qunar.Com
满座网 manzuo.com
1号店
欧派 OUPAI
紫博蓝 zibolan.com
世纪佳缘
美酒乐 mix1.com.cn
OSM 欧诗漫
贵人鸟
netconcepts get found

# 合作媒体


Miaozhen Systems

# 大数据 (1)

**已存储、处理超过2PB的数据，日均处理数据超过2TB**



# 大数据 (2)

3000+ campaigns

In 363 publishers

1,621 channels

6,425 positions

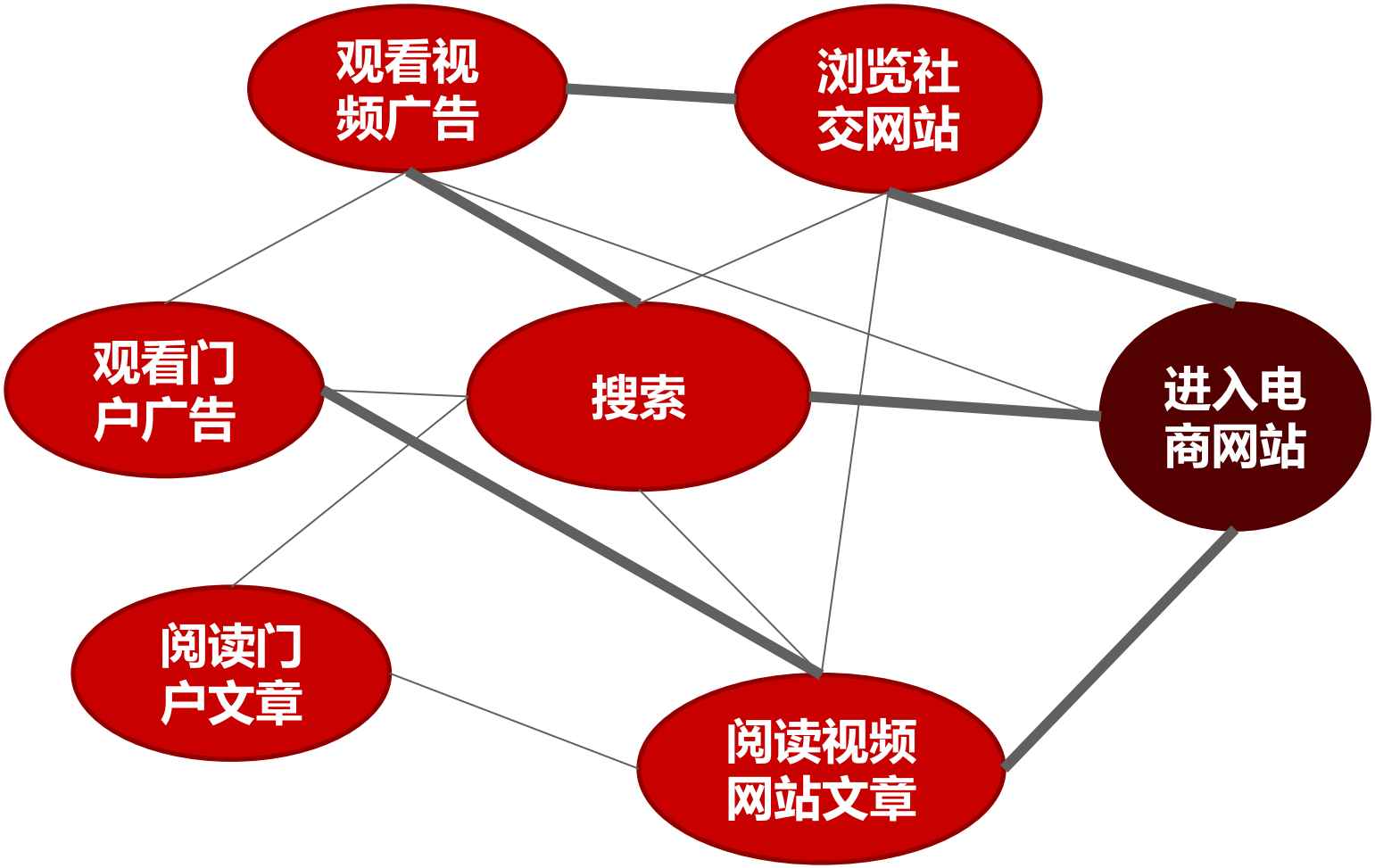
Total 700,000,000,000 impressions

# 管理和应用企业大数据的技术挑战

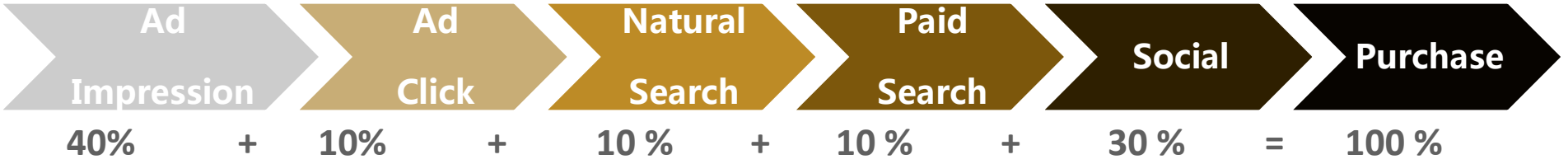
- 1, Attribution Modeling
- 2, UV Estimation & Identification
- 2, Social Targeting
- 4, Fraud Detection
- 5, Infrastructure

# *1: Attribution Modeling*

# Input



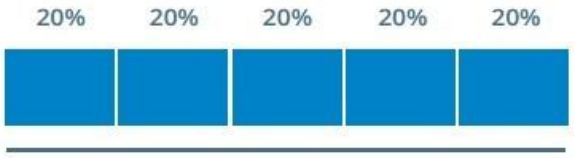
# Output


Ad Impression
Ad Click
Natural Search
Paid Search
Social
Purchase
40% + 10% + 10 % + 10 % + 30 % = 100 %

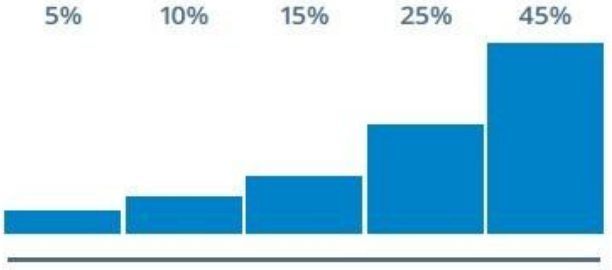
# Rule-based Model


(a) Last-Touch
将所有credit分配给最后一次touch


(b)First-Touch
将所有credit分配给第一次touch




(c)Linear
将credit均匀分配给所有touch




(d)Time Decay
根据转化点与touch的时间间隔来分配credit

# Logistic Regression Model

$$Y = \frac{e^{\sum a_i \cdot X_i}}{1 + e^{\sum a_i \cdot X_i}}$$

- Y为二值变量，表示是否有转化
- $X_i$为用户在第i个广告渠道$C_i$上的接触数
- 渠道$C_i$的回归系数 $a_i$表征了渠道对广告转化的重要性
- 渠道$C_i$的credit = $a_i$ ×渠道总接触数

# *2: Unique Visitor Estimation & Identification*

*Reference:*

*http://iir.ruc.edu.cn/pdf/ca2013/Lecture8-UV%20Estimation%20&%20Identification.pdf*

# Two Most Important Measurements

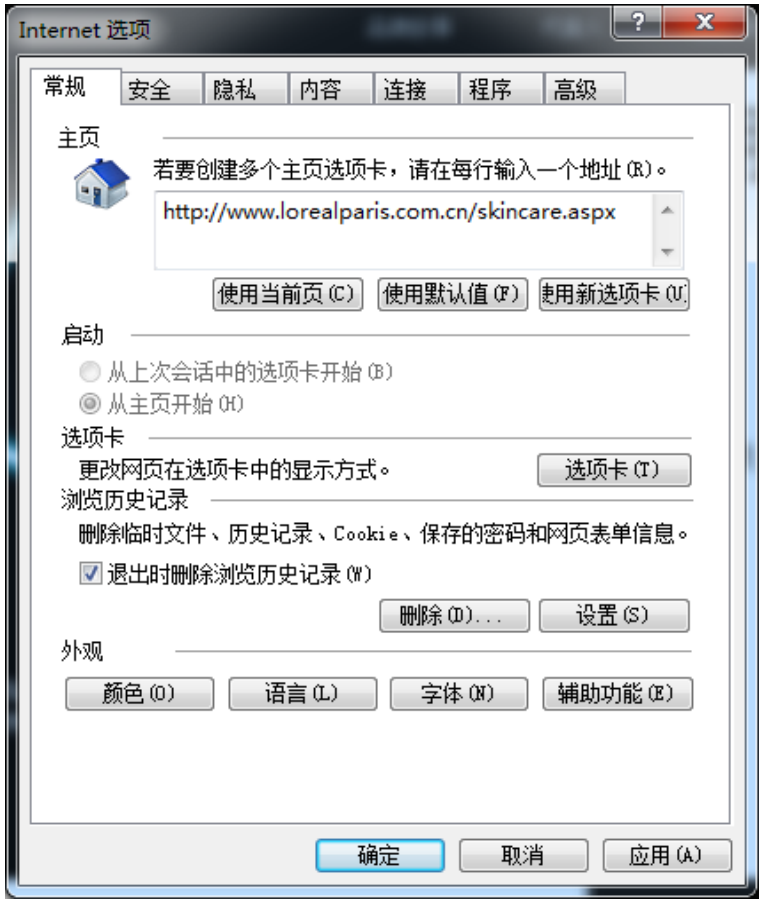
- Reach
  - How many **Unique Visitors (UV)** have seen ads from a campaign
- Frequency
  - How many times each **UV** has seen ads from a campaign
- Unique Visitor (UV)
  - A person
- Cookie
  - Issued when a web browser first visit a web site and stored at client

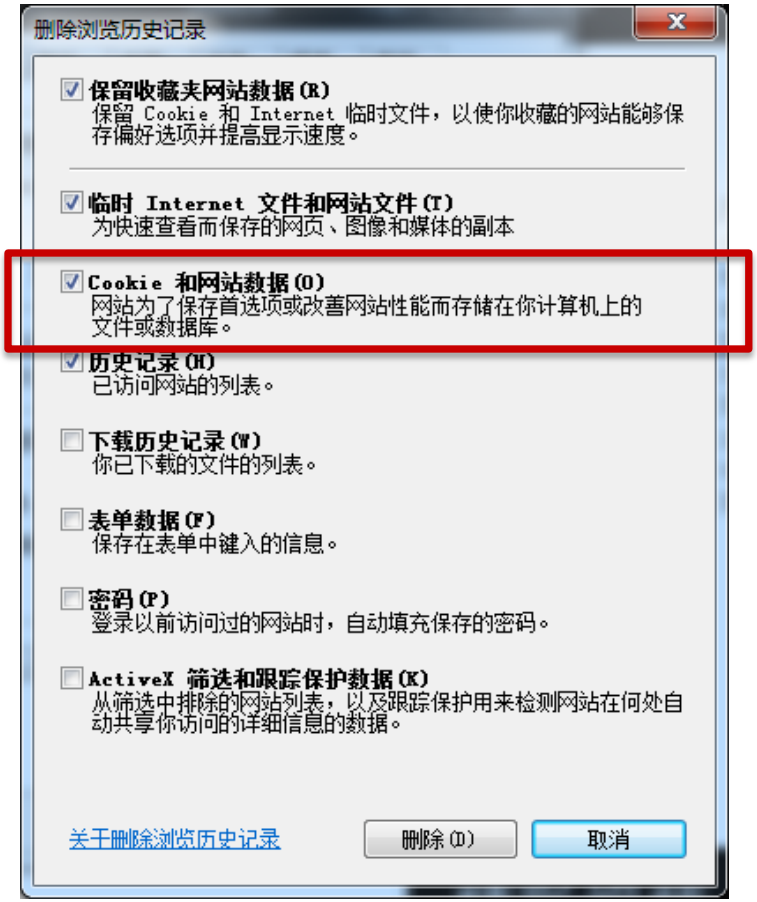
# *What are the Challenges?*

# Challenge 1: A Person has Multiple Devices

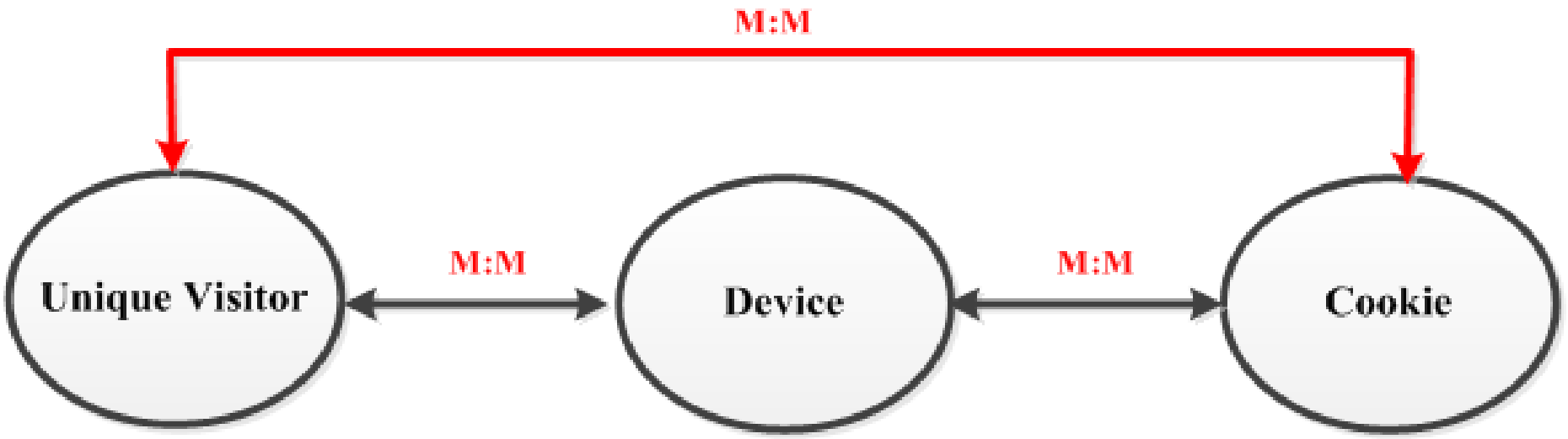
# Challenge 2: A Device Used by Multiple Persons

# Challenge 3: Cookie Deletion / Expiration



# Challenges Summary




- UV : Device (Many : Many)
- Device : Cookie (Many : Many)
- Cookie Restriction
- Across Screens

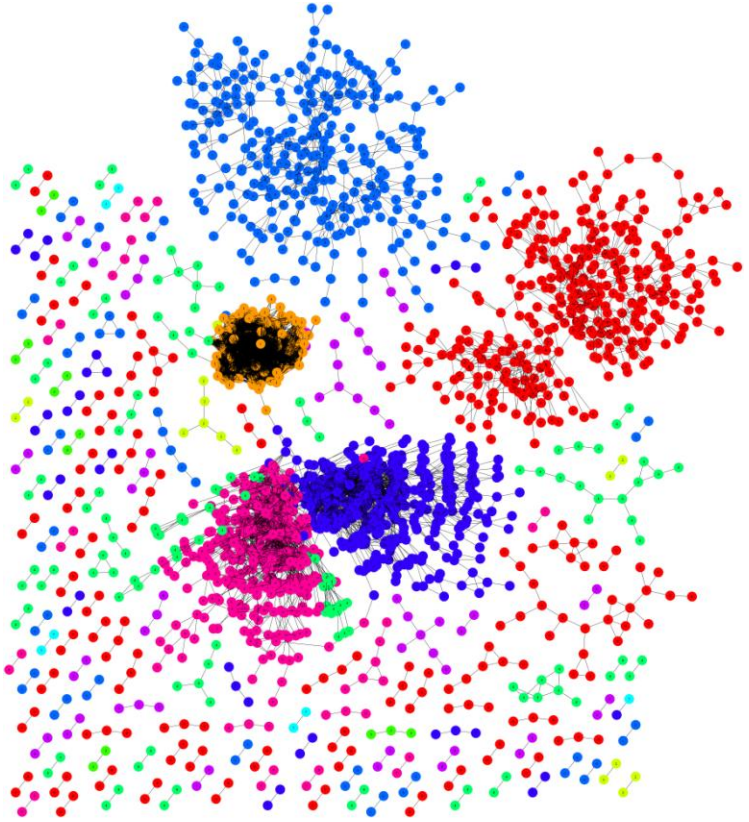
# PC

- Clustering Algorithm
  - Density-based clustering algorithm
- Bayes Factor Similarity Model
  - Similarity measure

$$sim(C, C') = s(x, x') = \sum_i \beta_i s_i(x_i, x'_i)$$

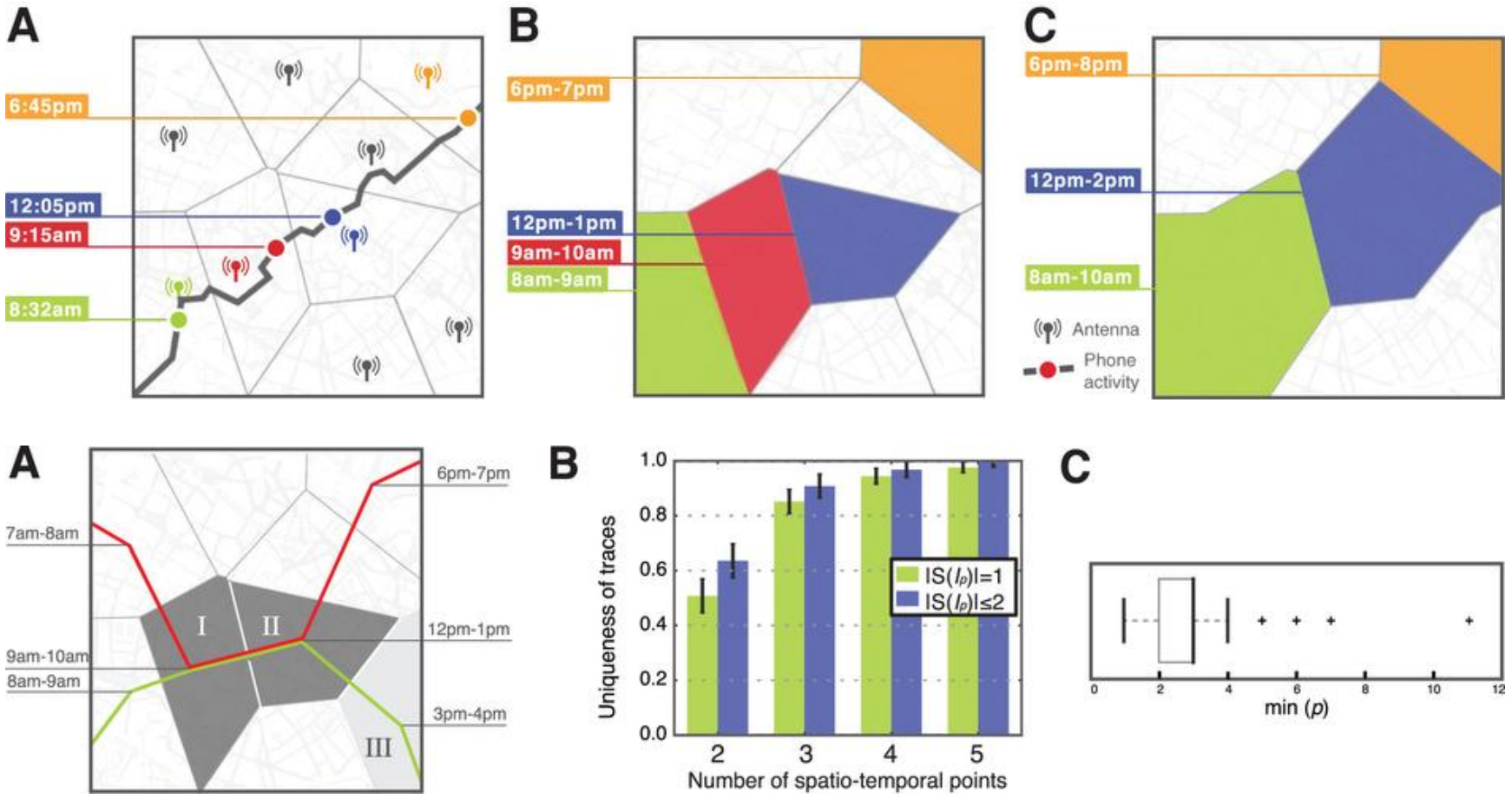
  - Training Bayes factor similarity model to get similarity threshold

$$P[y = 1] = 1/(1 + \exp(-\sum \beta_i s_i(x_i, x'_i)))$$



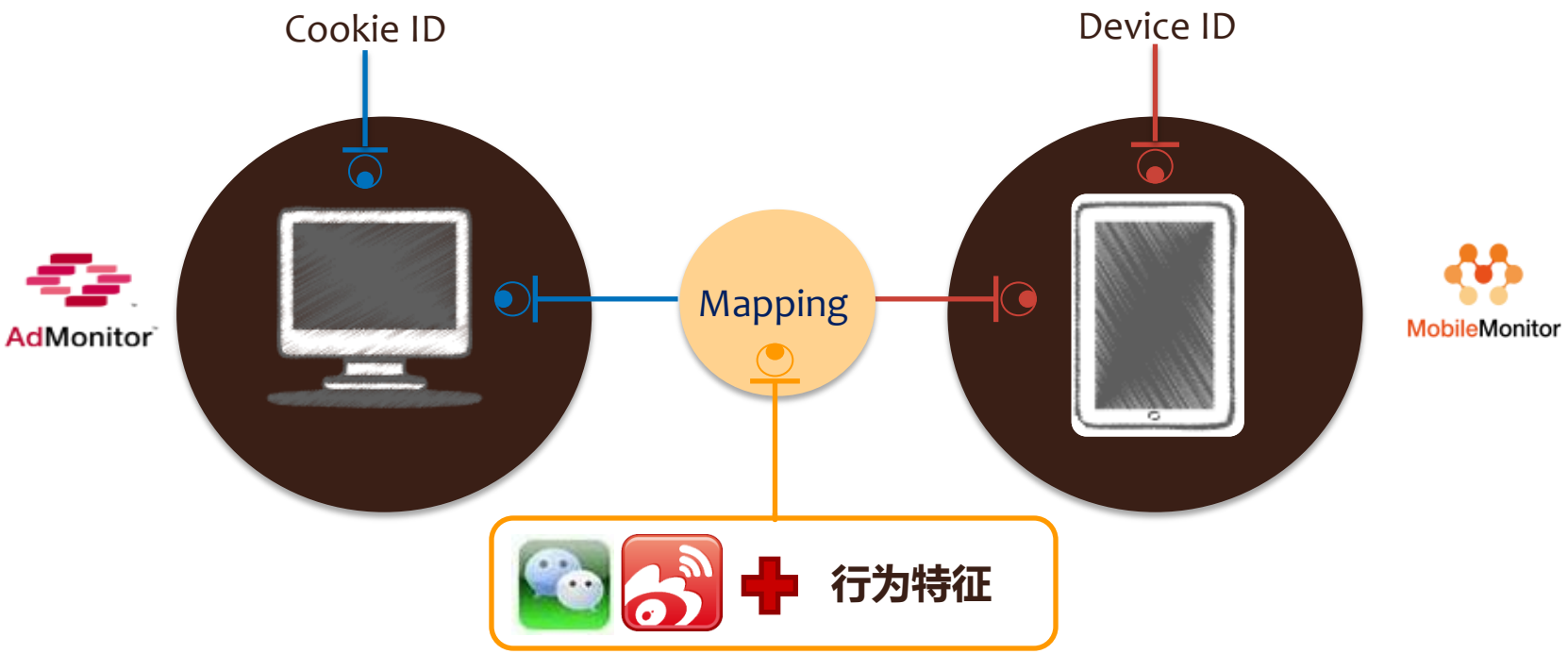
A. Dasgupta, etc., Overcoming browser cookie churn with clustering, WSDM'12

# Mobile



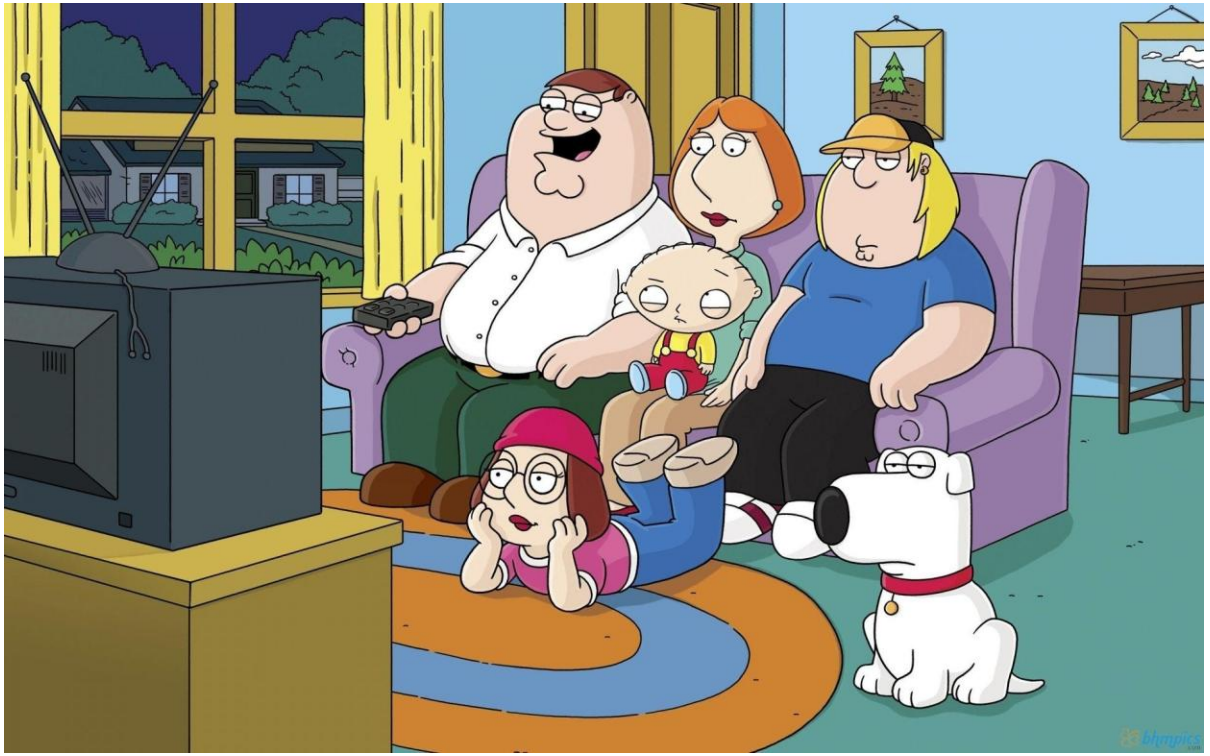

Unique in the Crowd: The privacy bounds of human mobility, Nature, 2013/03/25

# PC + Mobile

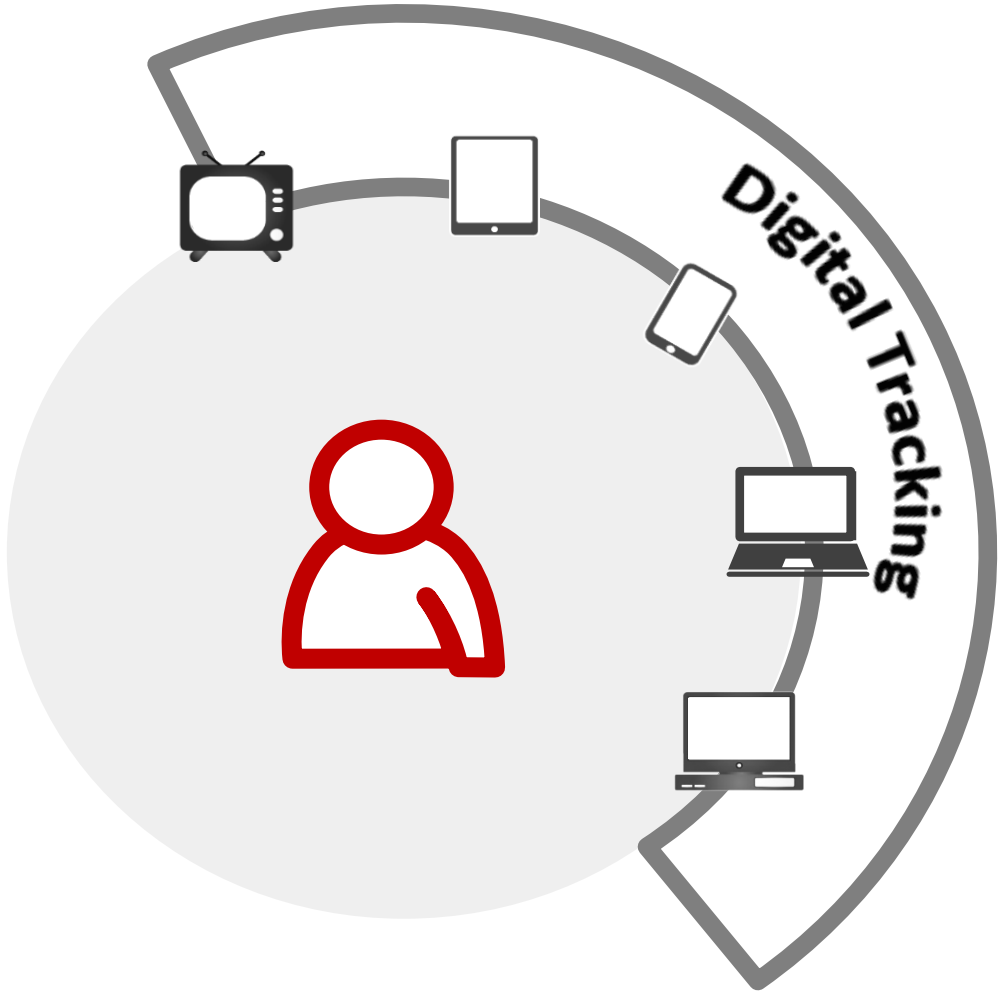


# Smart TV

- UV Identification
  - No solution
- UV Estimation
  - Rating Point
  - Diary
  - Audience Measurement Device



# Challenge: Cross Screens

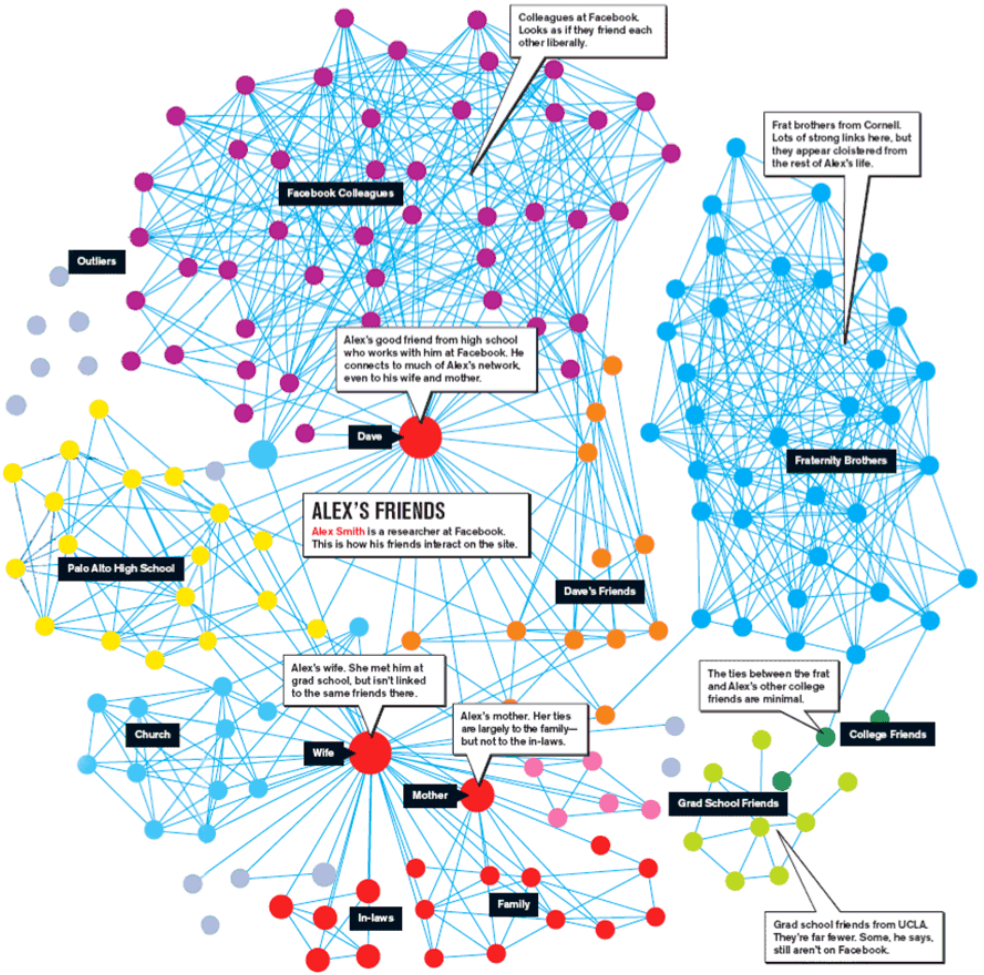


# *3: Social Targeting*

*Reference:*

*http://iir.ruc.edu.cn/pdf/ca2013/Lecture10-targeting.pdf*

# Social Targeting

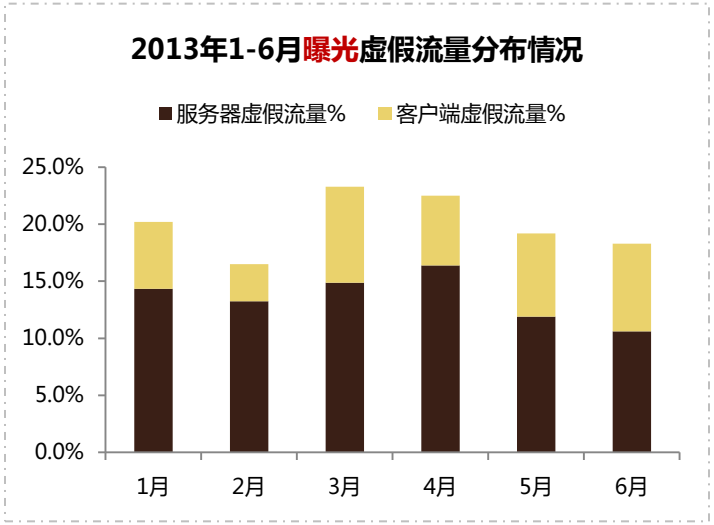


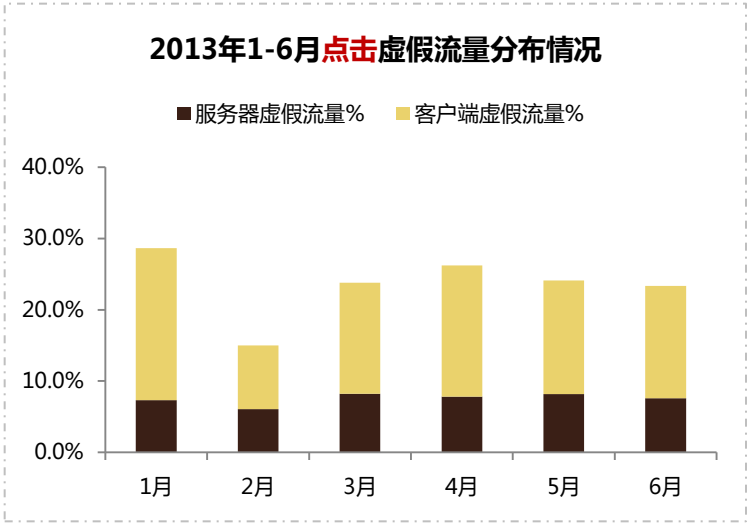
# *4: Fraud Detection*

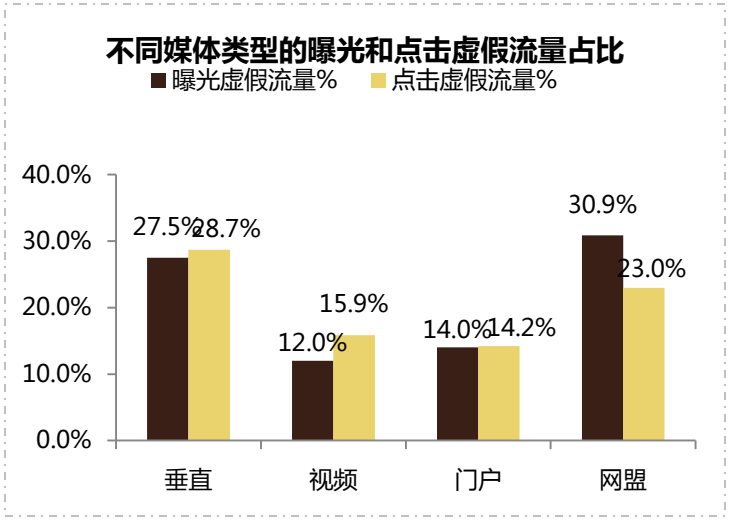
*Reference:*

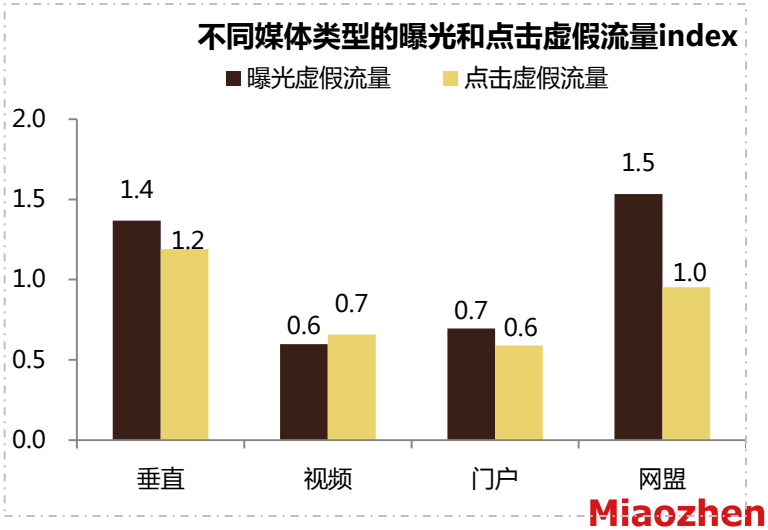
*http://iir.ruc.edu.cn/pdf/ca2013/Lecture3-Fraud%20Detection.pdf*

# 去伪存真


2013年1-6月曝光虚假流量分布情况
服务器虚假流量%
客户端虚假流量%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
1月
2月
3月
4月
5月
6月


2013年1-6月点击虚假流量分布情况
服务器虚假流量%
客户端虚假流量%
40.0%
30.0%
20.0%
10.0%
0.0%
1月
2月
3月
4月
5月
6月


不同媒体类型的曝光和点击虚假流量占比
曝光虚假流量%
点击虚假流量%
40.0%
30.0%
20.0%
10.0%
0.0%
27.5%
28.7%
12.0%
15.9%
14.0%
14.2%
30.9%
23.0%
垂直
视频
门户
网盟


不同媒体类型的曝光和点击虚假流量index
曝光虚假流量
点击虚假流量
2.0
1.5
1.0
0.5
0.0
1.4
1.2
0.6
0.7
0.7
0.6
1.5
1.0
垂直
视频
门户
网盟

# 数字电视异常流量（“水印”技术）

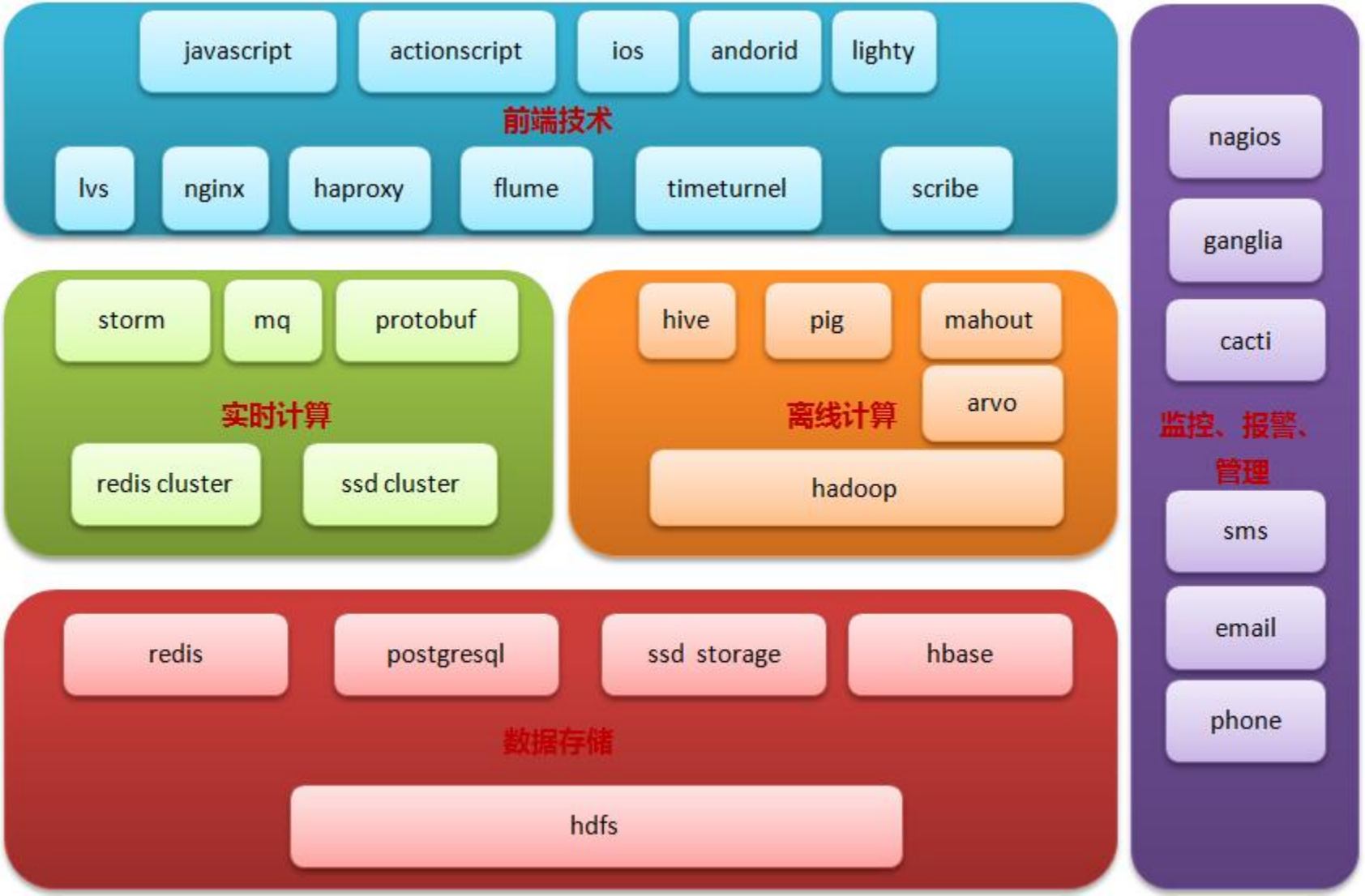
截取开机流信息，解析物料信息，每秒钟解析25次

检查是否存在秒针水印

存在秒针水印的物料与原始物料进行图像对比

图像对比显示两份物料相同：投放无误

# Infrastructure

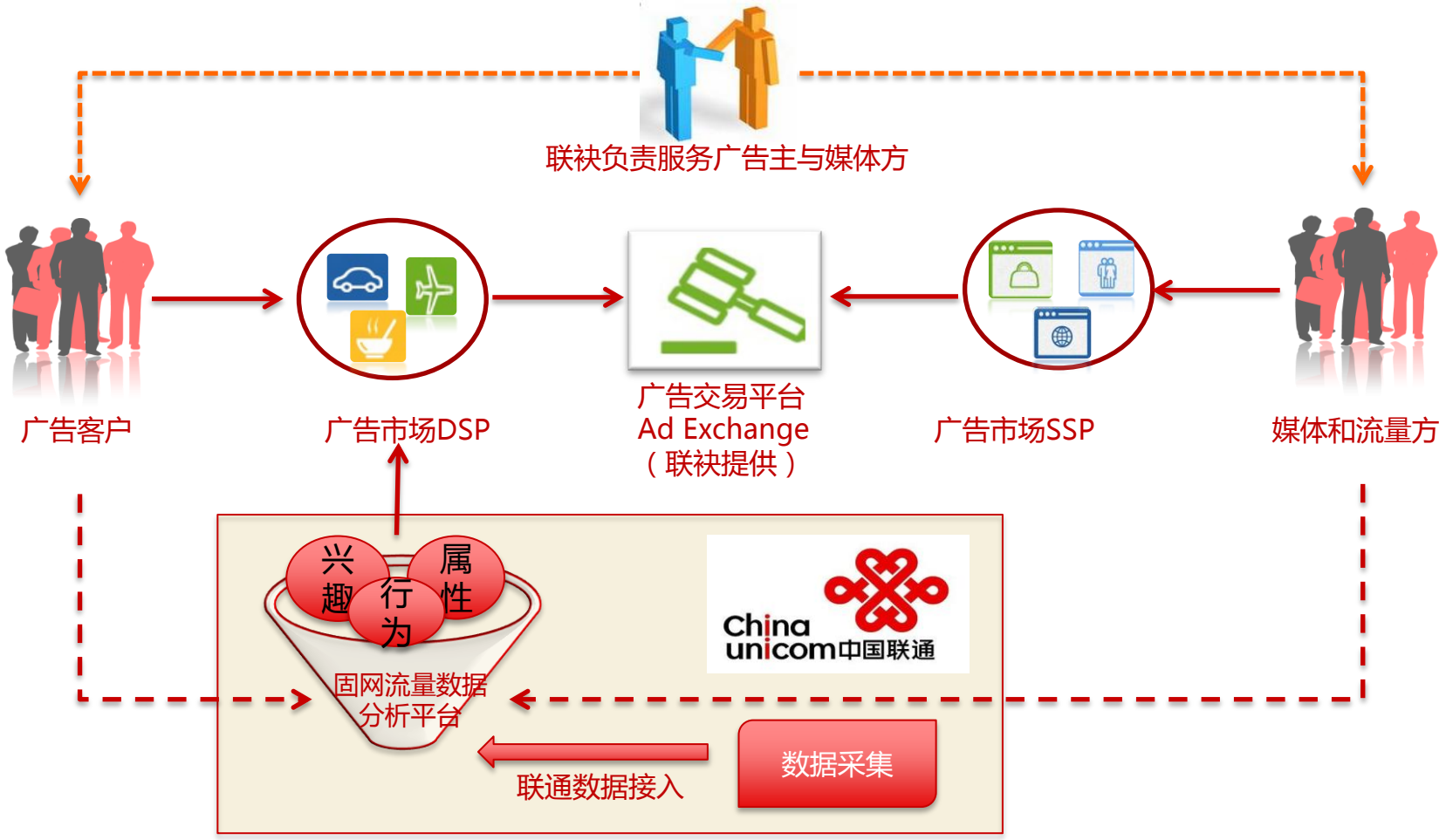



stems

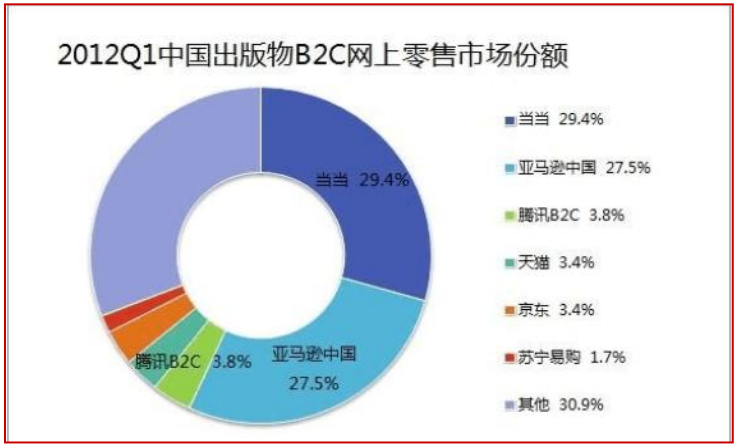
# 企业大数据应用案例分享

# 案例一：中国联通固网数据应用

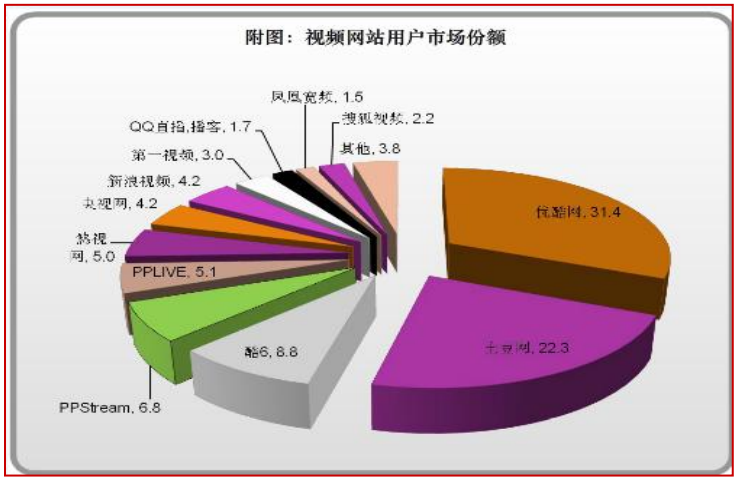
# 数据变现应用-RTB广告投放

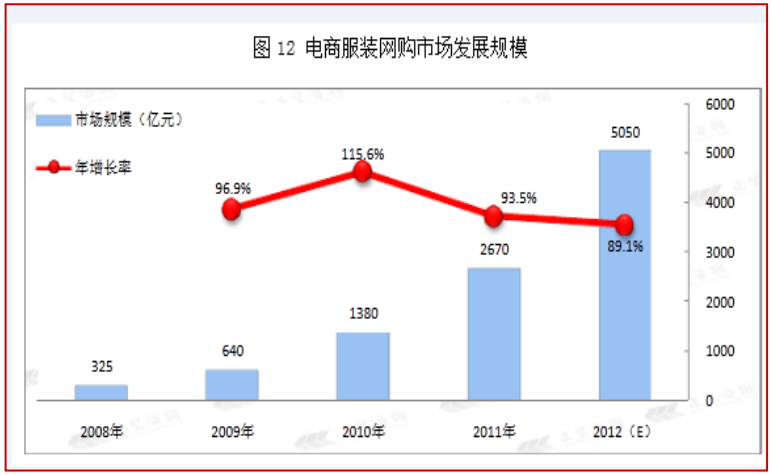


# 数据变现应用-数据报告




2003～2012年我国网络游戏市场规模及增长

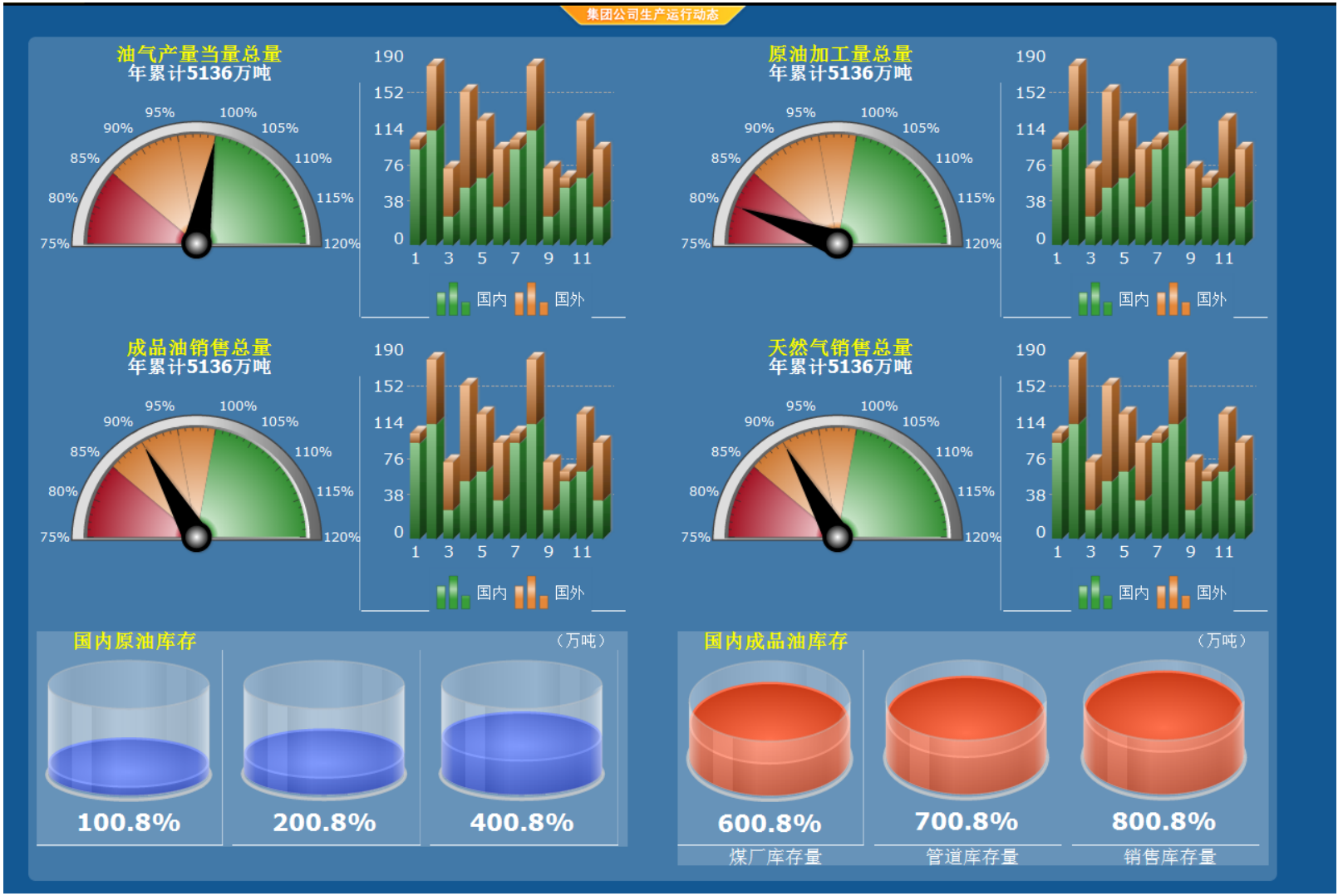


中国石油

# 案例二: 中石油集团公司信息统一展示平台

# 基于大数据的大屏展示




Systems

# 2013首届RTB算法大赛

*http://www.miaozhen.com/2013/CACC2013.html*

# 秒针公开数据集

# 合作高校


北京大學
PEKING UNIVERSITY
清華大學
Tsinghua University
中國人民大學
RENMIN UNIVERSITY OF CHINA
北京航空航天大學
BEIHANG UNIVERSITY
上海交通大學
Shanghai Jiao Tong University
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
武漢大學
WUHAN UNIVERSITY
中国科学技术大学
University of Science and Technology of China

fengshicong@miaozhen.com