The Quantification of Advertising

Lessons from Building Businesses based on Large Scale Data Mining

KDD 2010, WASHINGTON, D.C.
Outline

• Internet advertising (the business)
• Internet advertising (the data)
• Understanding consumers (the models)
• Organizing for success
The Personalized Media Economy

Media is transitioning from a “one size fits all” broadcast model to dynamic real-time choice

Online Advertising Ecosystem
An Advertising Data Explosion

- Massive expansion in number of decisions
  - Individuals, not whole audiences
  - Impressions, not whole sites
  - Screens/times/locations/……

- Decision timeframe reduced from weeks to milliseconds

- This problem can only be solved algorithmically
Online Advertising Ecosystem

Marketers

- Media Agencies
- Trading Desks
- Demand Side Platforms

Exchanges

Networks

- Sell Side Platforms

Publishers
Online Advertising Ecosystem (Reality)

Source: Display Advertising Technology Landscape. Terence Kawaja, GCA Savvian. May 2010
Money Follows Media Consumption

% Time Spent vs. % Ad Spend
United States 2009 (est.)

- Print: 12% Time Spent, 26% Ad Spend
- Radio: 16% Time Spent, 9% Ad Spend
- TV: 31% Time Spent, 39% Ad Spend
- Internet: 13% Time Spent, 28% Ad Spend

Source: Morgan Stanley, NA Technographics, VSS

$30B opportunity
Money Follows Media Consumption

% Time Spent vs. % Ad Spend
United States 2009 (est.)

<table>
<thead>
<tr>
<th>Media</th>
<th>Time Spent</th>
<th>Ad Spend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Print</td>
<td>12%</td>
<td>9%</td>
</tr>
<tr>
<td>Radio</td>
<td>16%</td>
<td>9%</td>
</tr>
<tr>
<td>TV</td>
<td>31%</td>
<td>28%</td>
</tr>
<tr>
<td>Internet</td>
<td>26%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Globally, hundreds of billions of dollars of ad spend will shift

Source: Morgan Stanley, NA Technographics, VSS

© 2010 Quantcast
Why the Spending Disparity?

• Media spend processes are well established
• New media channels lag until audiences and value can be properly quantified
• Historically, digital audiences were poorly quantified
  – Stratified sampling has been the norm in media measurement for decades
  – Bias and sampling error prevail
Enter Quantcast

- Launched September 2006 to enable addressable advertising at scale
- First we had to fix audience measurement
- Launched a free service based on direct measurement of media consumption
- Use machine learning to infer audience characteristics
Direct Measurement

Every Media Consumption Event is Measured

URL, Time, IP Address, User Agent, etc.
Broad Participation
World’s Favorite Audience Measurement Service
Data Rich Environment

250+ Billion / Month

Over 1 Billion Internet users
230 Million U.S. Internet users

Millions of Sites
Continually measured

15 TB / Day
New data

Media consumption events
Data Mining Challenges

Audience Estimation
Using reference data from a small number of people and a small number of web sites infer the demographics/attributes of the audience of all sites

Lookalike Selection
From the behavior of a small number of buyers of a product, determine the set of people who will buy it next

User Estimation
Using media consumption records and audience estimates, determine the characteristics of an Internet user across arbitrary dimensions

Live Traffic Modeling
Compute the value for showing an advertisement to a user as a function of the user, advertising environment, time of day, etc.
Lookalike Selection

- Given an archetype group of users, find the feature set that best separates them from their complement
- Features can be positive or negative indicators of content relevance
- Find more that look like them
Features Predict Relevance

Conversion Event

Users

Purchasers

Non Purchasers

Key
- +ve
- -ve

Days

© 2010 Quantcast
Wide Range of Activity
Websites, keywords, geo-location, ads and more

Conversion Event
Activity Level Variations
Cookie Deletion Rates

Proportion of cookies at least this old vs Days

Days

Proportion of cookies at least this old
Media consumption is non-stationary

‘Michael Jackson’ Media Consumption
June 25, 2009

Pages consumed per minute

13:00 13:30 14:00 14:30 15:00 15:30 16:00 16:30 17:00 17:30 18:00 18:30 19:00
Choose the Right Objective!

Indexed Click Vs. Conversion Rates

Clicks don’t always lead to conversions

The right metric is critical!
Winner takes all market
Positive feedback with high gain

**Micro**
- Most campaigns have performance metrics
- Performance leads to renewals
- 10% performance superiority can translate to 80% budget allocation

**Macro**
- Fragmentation challenges ad buyers and sellers
- Powerful ecosystem network effects
- Leads to small number of companies dominating regional markets
So, the faster we can address the challenges, the greater the opportunity
Numerous Challenges

- recoverability
- alerting
- business continuity
- bare-metal imaging
- storage
- software deployment
- security
- monitoring
- development APIs
- compression
- disaster recovery
- ip protection
- privacy
- metrics
- capacity management
- distributed profiling
- 24x7 availability
- BGP engineering
- latency
- recoverability/retry
- branch management
- scheduling
- data licensing
- system visibility
- global operations
- constant hardware failures
- data collection
- IDEs, (online query & large batch)
- server specification
- data schema evolution
- metadata management
- source control
- control systems
- regression testing
- rollback
- procurement
- switch fabric
- maintenance tasks
- logistics
Clearing the Path
Divide and Conquer
Quantcast Functional Teams

Market
Publisher
Pipeline
Edge

Math
Marketer
Inference
Cluster

Machines
Planner
Lookalikes
Ops

Audience Match
Live Traffic Modeling
Security
Divide and Conquer
Quantcast Functional Teams

- Market
  - Publisher
  - Pipeline
  - Edge
- Math
  - Marketer
  - Inference
  - Cluster
- Machines
  - Planner
  - Lookalikes
  - Ops
  - Audience Match
  - Live Traffic Modeling
  - Security
Machines
High Performance Platform

Multiple Global Datacenters
Ultra-high availability with advanced traffic management

225 Thousand / Second
Real-time events

2PB / Day
Analytical throughput
Compute as a Utility

PixelTone
Availability & latency of real-time architecture
Ensure self-healing, rapid failover of our multi-master design (no single point of failure)

ClusterTone
Utilization, performance, reliability and efficiency of our massively parallel distributed processing platform

LogTone
Management of data consolidation – Freshness
Have to ‘close the books’

AppTone
Performance, latency & user experience of all customer facing applications
Constantly Scaling Infrastructure
Whatever you have, it's never enough

Average Terabytes Processed Per Day

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Terabytes Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3 ’09</td>
<td>500</td>
</tr>
<tr>
<td>Q4 ’09</td>
<td>750</td>
</tr>
<tr>
<td>Q1 ’10</td>
<td>1,000</td>
</tr>
<tr>
<td>Q2 ’10</td>
<td>1,250</td>
</tr>
</tbody>
</table>
Divide and Conquer
Quantcast Functional Teams

Market

Math

Machines

Publisher

Marketer

Planner

Audience Match

Pipeline

Inference

Lookalikes

Live Traffic Modeling

Cluster

Edge

Ops

Security
Math Team Environment

Collaboration

• Regular brainstorming
• Group review meetings
• Shared wiki environment
• Team goals

Independence

• Everyone free to implement their own ideas
• Improved models
• Better metrics
• Visualization methods, etc.
Lookalike Pipeline

10 TB / Day

1 Billion Internet Users

10,000 Converters

500 TB

Model Configuration

100s of Concurrent Models

Model

Sandbox allows comparisons between models in fixed data set

Challengers to the production “hero” are developed in the sandbox, and can be rapidly deployed to become the new “hero”

Trained Models

Scoring

10M Potential Converters
Measuring Lift - ROC
Cumulative Lift
Learning α experimentation

6 Hours
To process 100TB with first MapReduce job

2 Days
New model development

Mins
New model in production

Hours
Live performance assessment

2 Weeks
To influence billions of real-time decisions every day and millions of dollars of advertising spend
Summary

• Digital advertising is a vast analytical environment
  – Enormous data volumes
  – Rich behaviors
  – Objective performance metrics

• Marketing will be transformed by computational approaches

• Hundreds of billions of dollars of spend are at stake
Numerous Open Challenges

• Dealing with sparsity
• Feature selection
• Real-time scoring and bidding
• ‘True’ performance & attribution modeling
• Lift, lift and more lift!
• Handling 100,000’s of concurrent models
These are hard problems, we could use some help!
Quantcast Visiting Scholar Program

• Full access to data, tools and compute infrastructure plus stipend financial support

• Any research you want, so long as it aims to improve a core metric

• Conducted from Quantcast SF and NYC offices

• Contact scholar@quantcast.com for program details
Top 50 Innovative Company
#3 Web Innovator
AlwaysOn Global 250 ‘09 Overall Winner

“The most ambitious attempt at shifting the online ad business...”
Advertising Age

“Quantcast Shakes Up Ad-Targeting Model”
Wall Street Journal

“Top 10 Private Company to Watch in 2010”
MIT Technology Review