Big Data Analytics: Applications & Opportunities in On-line Predictive Modeling

Usama Fayyad, Ph.D.
Chairman & CTO
ChoozOn Corporation
Twitter: @usamaf

August 12, 2012
BigMine: BigData Mining Workshop
KDD-2012 – Beijing, China
Outline

• Big Data all around us
• Introduction to Data Mining and Predictive Analytics
• On-line data and facts
• Case studies from multiple verticals:
  1. Yahoo! Big Data
  2. Social Network Data
  3. Case Study from nPario Applications
  4. ChoozOn Big Data from offers
• High-level view: don’t forget the basics
• Summary and conclusions
What Matters in the Age of Analytics?

1. Being Able to exploit all the data that is available
   • not just what you've got available
   • what you can acquire and use to enhance your actions

2. Proliferating analytics throughout your organization
   • make every part of your business smarter

3. Driving significant business value
   • embedding analytics into every area of your business can help you drive top line revenues and/or bottom line cost efficiencies
What Organizations Are Struggling With

• Data Strategy - how much data? why data? how does it impact my business?

• Prioritization conducted based on business need, not IT
  – Business justifications for Big Data
  – Demonstrating value of data in impacting the business
  – Looking at specialized stores to reduce TCO
  – File systems for grid computing (Hadoop)

• We do need to stay on top of our basic business ops
  – billing, monitoring, inventory management, etc...
  – Most can be handled by stream processing and traditional BI

• But, a new generation of requirements are becoming a priority for data-driven business
  – Predictive analytics, advanced forecasting, automated detection of events of interest
Why Big Data?

A new term, with associated “Data Scientist” positions:

• *Big Data: is a mix of structured, semi-structured, and unstructured data:*
  – Typically breaks barriers for traditional RDB storage
  – Typically breaks limits of indexing by “rows”
  – Typically requires intensive pre-processing before each query to extract “some structure” – usually using Map-Reduce type operations

• Above leads to “messy” situations with no standard recipes or architecture: hence the need for “data scientists”
  – conduct “Data Expeditions”
  – Discovery and learning on the spot
What Makes Data “Big Data”?  

- **Big Data is Characterized by the 3-V’s:**
  
  - **Volume:** *larger than “normal” – challenging to load/process*
    - Expensive to do ETL
    - Expensive to figure out how to index and retrieve
    - Multiple dimensions that are “key”
  
  - **Velocity:** *Rate of arrival poses real-time constraints on what are typically “batch ETL” operations*
    - If you fall behind catching up is extremely expensive (replicate very expensive systems)
    - Must keep up with rate and service queries on-the-fly
  
  - **Variety:** *Mix of data types and varying degrees of structure*
    - Non-standard schema
    - Lots of BLOB’s and CLOB’s
    - DB queries don’t know what to do with semi-structured and unstructured data.
The Distinction between “Data” and “Big Data” is fast disappearing

• Most real data sets nowadays come with a serious mix of semi-structured and unstructured components:
  – Images
  – Video
  – Text descriptions and news, blogs, etc…
  – User and customer commentary
  – Reactions on social media: e.g. Twitter is a mix of data anyway

• Using standard transforms, entity extraction, and new generation tools to transform unstructured raw data into semi-structured analyzeable data

• Hadoop vs. Not Hadoop - when to use what kind of techniques requiring Map-Reduce and grid computing
Text Data: The Big Driver

- While we speak of “big data” and the “Variety” in 3-V’s
- **Reality:** biggest driver of growth of Big Data has been text data
- In fact Map-Reduce became popularized by Google to address the problem of processing large amounts of text data:
  - Indexing a full copy of the web
  - Frequent re-indexing
  - Many operations with each being a simple operation but done at large scale
- Most work on analysis of “images” and “video” data has really been reduced to analysis of surrounding text
To Hadoop or not to Hadoop?

*when to use techniques requiring Map-Reduce and grid computing?*

- Typically organizations try to use Map-Reduce for everything to do with Big Data
  - This is actually very inefficient and often irrational
  - Certain operations require specialized storage
    - Updating segment memberships over large numbers of users
    - Defining new segments on user or usage data
- Map-Reduce is useful when a very simple operation is to be applied on a large body of unstructured data
  - Typically this is during entity and attribute extraction
  - Still need Big Data analysis post Hadoop
- Map-Reduce is not efficient or effective for tasks involving deeper statistical modeling
  - good for gathering counts and simple (sufficient) statistics
    - E.g. how many times a keyword occurs, quick aggregation of simple facts in unstructured data, estimates of variances, density, etc...
  - Mostly pre-processing for Data Mining
This is data of people blogging and tweeting about your products on the internet? How do you react to this?
Can you analyze data from these companies and integrate the results with your strategy?

How about in real time and then change your business on the fly?
Hadoop Use Cases by Data Type

- Content and Preference Data: 24%
- Text and Language Data: 16%
- Social Data: 11%
- Advertising Data: 10%
- IT Log Data: 19%
- Science Data: 7%
- CRM Data: 4%
- Financial Trading Data: 4%
- Sensor Data: 2%
- Supply Chain Data: 2%
- ERP Financial Data: 1%
# Big Data Applications and Uses

<table>
<thead>
<tr>
<th>Application</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT Log &amp; Security Forensics &amp; Analytics</td>
<td>100% Capture</td>
</tr>
<tr>
<td></td>
<td>Find New Signal</td>
</tr>
<tr>
<td></td>
<td>Predict Events</td>
</tr>
<tr>
<td>Automated Device Data Analytics</td>
<td>Product Planning</td>
</tr>
<tr>
<td></td>
<td>Failure Analysis</td>
</tr>
<tr>
<td></td>
<td>Proactive Fixes</td>
</tr>
<tr>
<td>Advertising Analytics</td>
<td>Recommendation</td>
</tr>
<tr>
<td></td>
<td>Segmentation</td>
</tr>
<tr>
<td></td>
<td>Social Media</td>
</tr>
<tr>
<td>Big Data Warehouse Analytics</td>
<td>Hadoop + MPP + EDW</td>
</tr>
<tr>
<td></td>
<td>Cost Reduction</td>
</tr>
<tr>
<td></td>
<td>Ad Hoc Insight</td>
</tr>
<tr>
<td></td>
<td>Predictive Analytics</td>
</tr>
</tbody>
</table>
Analysis & Programming Software

Event Survey - Languages

SQL
MySQL
Python
Java
C
C#
R
C or C++ or C#
R
Beginner
Perl
B/C/Intel
PHP
SAS
SPSS
Hadoop+

disco
massive data - minimal code

RevolutionAnalytics / RHadoop
From Basic Dashboards to Advanced Analytics

• Data Reduction to get
  – Advanced views oriented by customer or product
  – Segmentation
  – Pattern analysis and summaries

• Predictive Analytics
  – Data Mining
  – Statistical analysis
  – Optimization of processes and spend

The same analytics technique apply across many industries: fraud detection is fraud detection, is fraud detection
What is Data Mining?

Finding *interesting structure* in data

- **Structure**: refers to statistical patterns, predictive models, hidden relationships
- **Interesting**: Accurate predictions, associated with new revenue potential, associated with cost savings, enables optimization

- Examples of tasks addressed by Data Mining
  - Predictive Modeling (classification, regression)
  - Segmentation (Data Clustering)
  - Affinity (Summarization)
    - relations between fields, associations, visualization
Data Mining and Databases

Many interesting analysis queries are difficult to state precisely

• Examples:
  – which records represent fraudulent transactions?
  – which households are likely to prefer a Ford over a Toyota?
  – Who is a good credit risk in my customer DB?
  – Why are these automobiles in need of unusual repairs?

• Yet database contains the information
  – good/bad customer, profitability
  – did/didn’t respond to mailout/survey/campaign/...
  – automobile repair and warranty records
## Many Business Uses

<table>
<thead>
<tr>
<th>Analytic technique</th>
<th>Uses in business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketing and sales</td>
<td>Identify potential customers; establish the effectiveness of a campaign</td>
</tr>
<tr>
<td>Understanding customer behavior</td>
<td>model churn, affinities, propensities, ...</td>
</tr>
<tr>
<td>Web analytics &amp; metrics</td>
<td>model user preferences from data, collaborative filtering, targeting, etc.</td>
</tr>
<tr>
<td>Fraud detection</td>
<td>Identify fraudulent transactions</td>
</tr>
<tr>
<td>Credit scoring</td>
<td>Establish credit worthiness of a customer requesting a loan</td>
</tr>
<tr>
<td>Manufacturing process analysis</td>
<td>Identify the causes of manufacturing problems</td>
</tr>
<tr>
<td>Portfolio trading</td>
<td>optimize a portfolio of financial instruments by maximizing returns &amp; minimizing risks</td>
</tr>
<tr>
<td>Healthcare Application</td>
<td>fraud detection, cost optimization, detection of events like epidemics, etc...</td>
</tr>
<tr>
<td>Insurance</td>
<td>fraudulent claim detection, risk assessment</td>
</tr>
<tr>
<td>Security and Surveillance</td>
<td>intrusion detection, sensor data analysis, remote sensing, object/person detection, link analysis, etc...</td>
</tr>
</tbody>
</table>
So this Internet thing is going to be big!

Big opportunity, Big Data, Big Challenges!
Stats about on-line usage

• How many people are on-line today?
  – 2.1 Billion (per Comscore estimates)
  – 30% of world Population

• How much time is spent on-line per month by the whole world?
  – 4M person-years per month

• How many hours per month per Internet User?
  – 16 hours (global average)
  – 32 hours (U.S. Average)

*Sources: Feb.2012 - from Go-Gulf.com compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org
How Are Users Distributed Geographically

*Sources: Feb.2012 - from Go-Gulf.com compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org
How Do People Spend Their On-line Time?

- On-line Shopping? 5%
- Searches? 21%
- Email/Communication? 19%
- Reading Content? 20%
- Social Networking? 22%
- Multimedia Sites? 13%
Most Popular Activities On-Line?

- **Emails**: 92%
- **Using Search Engines**: 92%
- **Health or Medical Info**: 83%
- **Hobbies**: 83%
- **Search for Directions**: 82%
- **Check weather**: 81%
- **Info search on buying products**: 78%
- **Reading News**: 76%
- **Entertainment**: 72%
- **Buy a Product**: 71%

*Sources: Feb.2012 - from Go-Gulf.com compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org*
Top 10 Sites Visited?

- **Google**: 153.4M visitors/month
  - each spending 1h 47mins
- **Facebook**: 137.6M visitors
  - each spending 7h 50mins
Interesting Events

• **Google:** How many queries per day?
  – More than 1 Billion

• **Twitter:** How many Tweets/day?
  – More than 250M

• **Facebook:** Updates per day?
  – More than 800M

• **YouTube:** Views/day
  – 4 Billion views
  – 60 hours of video uploaded every minute!

• **Social Networks:** users who have used sites for spying on their partners?
  – 56%

*Sources: Feb.2012 - from Go-Gulf.com compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org*
Interesting Events

- **Country with Highest online friends?**
  - Brazil
  - 481 friends per user
  - Japan has least at 29

- **Country with maximum time spent shopping on-line??**
  - China: 5 hours/week

*Sources: Feb. 2012 - from Go-Gulf.com compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org*
So Internet is a big place with lots happening?

Do we understand what each individual is trying to achieve?

- What is user intent?
- Critical in monetization, advertising, etc...

Do we understand what a community’s sentiment is?

- What is the emotion?
- Is it negative or positive?
- What is the health of my brand online?

Do we understand context and content?

- What are appropriate ads?
- Is it Ok to associate my brand with this content?
- Is content sad?, happy?, serious?, informative?
Case Studies

Yahoo! Big Data
Yahoo! – One of Largest Destinations on the Web

80% of the U.S. Internet population uses Yahoo!
– Over 600 million users per month globally!

• Global network of content, commerce, media, search and access products
• 100+ properties including mail, TV, news, shopping, finance, autos, travel, games, movies, health, etc.

• **25+ terabytes of data collected each day**
  • Representing 1000’s of cataloged consumer behaviors

More people visited Yahoo! in the past month than:
• Use coupons
• Vote
• Recycle
• Exercise regularly
• Have children living at home
• Wear sunscreen regularly

*Data is used to develop content, consumer, category and campaign insights for our key content partners and large advertisers*

Sources: Mediamark Research, Spring 2004 and comScore Media Metrix, February 2005.
Yahoo! Big Data – A league of its own...

**GRAND CHALLENGE PROBLEMS OF DATA PROCESSING**

TRAVEL, CREDIT CARD PROCESSING, STOCK EXCHANGE, RETAIL, INTERNET

Y! Data Challenge Exceeds others by 2 orders of magnitude
Behavioral Targeting (BT)

Targeting your ads to consumers whose recent behaviors online indicate that your product category is relevant to them.
Yahoo! User DNA

- Male, age 32
- Lives in SF
- Lawyer
- Spends 10 hour/week on the internet
- Clicked on Sony Plasma TV SS ad
- Searched on: "Italian restaurant Palo Alto"
- Searched on: "Hillary Clinton"
- Clicked on Sony Plasma TV SS ad
- Purchased Da Vinci Code from Amazon
- Checks Yahoo! Mail daily via PC & Phone
- Has 25 IM Buddies, Moderates 3 Y! Groups, and hosts a 360 page viewed by 10k people
- Searched on from London last week

On a per consumer basis: maintain a behavioral/interests profile and profitability (user value and LTV) metrics
How it works | Network + Interests + Modelling

Analyze predictive patterns for purchase cycles in over 100 product categories

In each category, build models to describe behaviour most likely to lead to an ad response (i.e. click).

Score each user for fit with every category...daily.

Target ads to users who get highest ‘relevance’ scores in the targeting categories
Recency Matters, So Does Intensity
Differentiation | Category specific modelling

Example 1: Category Automotive

Alt Behaviour 1: 5 pages, 2 search keywords, 1 search click, 1 ad click

Example 2: Category Travel/Last Minute

Alt Behaviour 1: 5 pages, 2 search keywords, 1 search click, 1 ad click

Different models allow us to weight and determine intensity and recency
Differentiation | Category specific modelling

Example 1: Category Automotive

Different models allow us to weight and determine intensity and recency

Alt Behaviour 1: 5 pages, 2 search keywords, 1 search click, 1 ad click

user is in the Intense Click Zone

with no further activity, decay takes effect

Intense Click Zone
Automobile Purchase Intender Example

• A test ad-campaign with a major Euro automobile manufacturer
  – Designed a test that served the same ad creative to test and control groups on Yahoo
  – Success metric: performing specific actions on Jaguar website

• Test results: 900% conversion lift vs. control group
  – Purchase Intenders were 9 times more likely to configure a vehicle, request a price quote or locate a dealer than consumers in the control group
  – ~3x higher click through rates vs. control group
Mortgage Intender Example

Example: Mortgages

We found:
1,900,000 people looking for mortgage loans.

Example search terms qualified for this target:
- Mortgages
- Home Loans
- Refinancing
- Ditech

Example Yahoo! Pages visited:
- Financing section in Real Estate
- Mortgage Loans area in Finance
- Real Estate section in Yellow Pages

Source: Campaign Click thru Rate lift is determined by Yahoo! Internal research. Conversion is the number of qualified leads from clicks over number of impressions served. Audience size represents the audience within this behavioral interest category that has the highest propensity to engage with a brand or product and to click on an offer.

Date: March 2006
Experience summary at Yahoo!

- Dealing with the largest data source in the world (25 Terabyte per day)
- BT business was grown from $20M to about $500M in 3 years of investment!
- Building the largest database systems:
  - World’s largest Oracle data warehouse
  - World’s largest single DB
  - Over 300 data mart data
  - Analytics with thousands of KPI’s
  - Over 5000 users of reports
  - Largest targeting system in the world
- Big demands for grid computing (Hadoop)
Social Network

Social Graph Analysis (no time)
Social Network Marketing
Understanding Context for Ads
Case Study: TWITTER Social Marketing?

Viacom’s VH1 Twitter campaign on ANVIL (the movie)
(week of May 14th, 2009 – see AdAge article)

Marketing ANVIL movie
Very niche audience, how do you reach them?

Diffusion
Give 20 free DVD’s to major related artists/groups, ask them
To notify Twitter groups – reached over 2M people

Social Identity: power of word of mouth...
What is the Cost to VH-1?
Compare with traditional approach: TV commercials to promote a
documentary film?
The Display Ads Challenge Today

What Ad would you place here?
The Display Ads Challenge Today

Damaging to Brand?

Body parts delivered to Michigan home

CASCADE TOWNSHIP, Mich. - Two packages containing human body parts — including a liver and part of a head — meant for a medical research lab instead were delivered to a home.

The body parts, sent from China, were mistakenly dropped off Thursday at Franck and Ludvine Larmande's home by a DHL express driver who believed the bubble-wrapped items were pieces to a table.

"My husband started to unwrap one and said, 'This is strange, it looks like a liver,'" Ludvine Larmande said. "He started the second one, but stopped as soon as we saw the ear."

"Something wasn't right. It was scary, and I'm glad I didn't open them."

"The couple called Kent County sheriff's deputies, who determined the preserved body parts were for medical research, Lt. Roger Parent said."

Authorities believe 28 more bubble-wrapped human organs and body parts could be dispersed across the country. The Grand Rapids Press reported. Two of five packages headed to the northern Michigan lab broke open, scattering their contents.

"There will definitely be a shock to people if they see these things, but there is no hazard to health," Parent said.
The Display Ads Challenge Today

What Ad would you place here?
The Display Ads Challenge Today

Irrelevant and Damaging to Brand

Completely Irrelevant

What happens to your body after you drink a soda every day, for a long time

Sugar rushes and caffeine highs followed by a depressing energy crash are what happens to your body if you drink a soda right now, but plenty of BlissTree readers actually seem to be okay with that. Some of you think it’s alarmist to compare a caffeine and sugar rush to doing drugs, and some just don’t really care about the slump they’ll find themselves in after drinking 39 grams of sugar, but what makes us really worried about a soda slumber...
NetSeer: Intent for Display

• Currently Processing 4 Billion Impressions per Day
Problem: Hard to Understand User Intent

Contextual Ad served by Google

What NetSeer Sees:

URL: [http://martiansoftware.com/nailgun/](http://martiansoftware.com/nailgun/)

- Programming in Java
- Java Development Kit
- Sun Java
- Java Programmers
- Java Virtual Machine
- Static IP Address
- DNS Servers
- Dynamic DNS
- Java Runtime Environment
- Java Software
- Java Programming Language
- Server Configuration
- Dynamic IP Addresses
- Free Java
- Code in Java
- Implementation of Java
- Local Machine
- Running Java Programs
- Java Language
- Java Platform
- Spring Framework
- Home Server
- Virtual Server

Nailgun: Insanely Fast Java

Thinnest. Client. Ever?

Nailgun is a client, protocol, and server for running Java programs from the command line without incurring the JVM startup overhead. Programs run in the server (which is implemented in Java), and are triggered by the client (written in C), which handles all I/O.

What’s New?
See the ChangeLog.

How does it work?
See the project background.

How do I get it?
From SourceForge.

Before you download it, be aware that it’s not secure. Not even close. Although there are means to ensure that the client is connected to the server from the local machine, there is not yet any concept of a “user”. Any programs that run in Nailgun are run with the same permissions as the server itself. You have been warned.

How do I use it?
See the Quick Start guide.
Case Studies

nPario – Data Management Platform
ChoozOn – Big Data over Offers Universe
Example of a Big Data DMP

nPario™

Scale
nPario builds an infinitely scalable data management platform (DMP) that allows advertisers and marketers to manage, understand, and monetize their data. Their technology has been proven at companies such as Yahoo and EA.

Applications
nPario applications include segmentation of audiences for increasing the value of advertising, reporting/analytics for examining performance, attribution to show which advertising works, and experimentation to test ideas.

Access
nPario emphasizes putting access to data in the hands of marketing, advertising and other business users
nPario™

Powerful Technology.

540m Users, 8+ Petabytes, & 16 Patents

nPario has the only **commercially available** Big Data management technology built for one of the “Big Five”.

**Significant Investment**

nPario’s technology is the result of more than $50m investment in development and 16 issued patents.

**In Production at Yahoo**

nPario technology manages the world’s largest data system. Used for Yahoo’s Marketing and Advertising business. Used across Yahoo’s platforms and 120 online properties. Used by hundreds of analysts.
Data Sources

**First-Party**
- CRM
- Offline Data
- Events

**Multiple Channels**
- Search
- eMail
- Mobile
- Video
- Display
- Site

**Third-Party**
- bluekai
- nielsen
- Experian
- comScore

Self-Service Data Intake & ETL
- Data Processing, Enrichment and Normalization
- User DNA
- Real-Time Data Availability

Deep Insights & Big Data Analytics
- Insights Analytics
- Modeling & Discovery

Self-Service Applications
- Segment
- Target
- Attribution
- Experimentation

Data Management Platform
Challenge: Provide cross-platform campaign insights for advertisers and enable audience discovery across channels.

Result: Unified view of gamers across multiple cross platform data sources.

"EA increased its worldwide audience reach by 30% this year [...]. Combining that major jump in reach with the launch of EA Legend puts us in perfect position to compete”

Dave Madden, Senior Vice President of Global Media Solutions
Electronic Arts.
Summary

Total Reach

Unduplicated Uniques: 24.6 mm

- Console 37%
- Online 31%
- Social 19%
- Mobile 13%

Performance

- **1.17M** Total DLC downloads
- **2.7M** Social quest completions
- **21M** Virtual goods claimed
- **.6%** Top CTR from Pogo (July '11)
- **1.2%** Top CTR from Mobile (June '11)

Brand Impact

- +5% Increase in total US sales
- +25% Growth in positive social conversations
- +15% Awareness lift
- +22% Brand Favorability lift
- +31% Purchase Intent lift

Audience

Gender

- 37% MALE
- 63% FEMALE

Age

- 10% 2-17
- 12% 18-24
- 19% 25-34
- 22% 35-44
- 18% 45-54
- 12% 55-64
- 7% 65+

Ad Delivery

- Online
- Console
- Social
- Mobile

Impressions (mm)

- Campaign
- General Internet Pop.
Specialized Search through Big Data Analytics over the Offers Universe
Chaos for Consumers

Flash Sale Sites
Deep Discount Sites
Coupon Sites
Loyalty Programs
Online Loyalty Program
Search Engines
Social Networks
Consumer

Google
bing
Yahoo!
LinkedIn
Twitter
Facebook
United
VISA
Discover
Ebay
eversave
Groupon
GILT
Ideeli
One Kings Lane
Restaurant.com
6PM.com
Coupon Mountain
Savings.com
RetailMeNot
American Express
MyPoints
ChoozOn
Chaos for Marketers

- Flash Sale Sites
- Deep Discount Sites
- Coupon Sites
- Loyalty Programs
- Daily Deals Sites
- Search Engines
- Social Networks
- Online Loyalty Programs

brand = price
What Consumers & Marketers Want

**Consumers**
- Value from brands they love
- Tame the deal chaos

**Marketers**
- Reach targeted consumers
- Build loyalty
- Create brand evangelists
Keys to Being THE Consumer Network

- Comprehensive Deal Coverage
- Personalization
- Intelligent Matching
- Permission-based Targeting Solution
- Loyalty Solution
- Multi-Channel Reach
Solution for Consumers

- Daily Deals Sites
  - Groupon
  - LivingSocial
  - Rue La La

- Flash Sale Sites
  - Gilt

- Deep Discounters
  - 6PM.com

- Affiliate Deals & Coupon Sites
  - Amazon.com
  - RetailMeNot
  - United

- Loyalty Programs

Choozer Interests & Preferences

- Machine-based
- Intelligent Matching
- Social (Pals, Clubs)

- Web App
- Mobile App
- Email
- Digital Media

ChoozOn
What Carol Gets

Carol’s Consumer Network

“Chozen” brands
- Nike
- Macy’s
- Citi
- American Express
- Banana Republic
- Groupon

Interests (Intent)
- Shopping
- Pals
- Deal Clubs

The Universe of Deals

Affiliate Deals
Loyalty Programs
Daily Deals
Flash Sales
Deep Discounters

1,500+ brands
100,000+ offers

PLUS her ChoozOn Inbox™

Intellectual Matching

A Personal Shopper for Deals
Big Picture on Big Data Analytics

Key points
Don’t Forget The Basics

• Metrics and Scorecards are the first steps to awareness
• Plays a huge role in deploying predictive models and monitoring and proving their effectiveness
• Often scorecards require
  – Going through huge amounts of data to produce the required metrics
  – Ability to get to the metrics in low latency
  – Ability to modify metrics and update quickly
  – Integration with data warehouse
Focus On The Right Measure

Referral Site Metrics

- Total traffic not a good performance measure
- High-traffic referral sites often produce poorer quality click throughs
- Ads best response not most effective
- Target the message
Focus On The Right Measure

Referral Site Metrics

- Total traffic not a good performance measure
- High-traffic referral sites often produce poorer quality click throughs
- Ads best response not most effective
- Target the message
Focus On The Right Measure

Referral Site Metrics

- Total traffic not a good performance measure
- High-traffic referral sites often produce poorer quality click throughs
- Ads best response not most effective
- Target the message
Retaining New Yahoo! Mail Registrants

Sometimes, Simple is Very Powerful!
Integrating Mail and News

• Data showed that users often check their mail and news in the same session
  – But no easy way to navigate to Y! News from Y! Mail

• Mail users who also visit Y! News are 3X more active on Yahoo
  – Higher retention, repeat visits and time-spent on Yahoo
“In the news” Module on Mail Welcome Page

- Increased retention on Mail for light users by 40%!
  - Est. Incremental revenue of $16m a year on Y! Mail alone
Benefits of Advanced Analytics

• Advanced Analytics brings out the real value of data
• The business begins to understand the true value and role of data in moving the big needles
• Focus on useful requirements from data, rather than “data acrobatics”
• Value creation from data leads to proper investment scoping
  – Many are realizing predictive analytics and data mining are much more useful than reporting
  – Integration of analytics story with data storage very critical
• Big data makes analytics even more essential and more useful
  – Avoiding the challenges of separating analytics from big data are increasingly important
Thank You! & Questions?

Twitter: @usamaf

Usama@choozOn.net
www.ChoozOn.com