From BigData to Data Science: Predictive Analytics in a Changing Data Landscape

Usama Fayyad, Ph.D.

Group Chief Data Officer and CIO Risk, Finance, & Treasury Technology
Barclays

Twitter: @usamaf

Keynote talk: ADMA Advancing Analytics Conference – Sydney - August 4, 2015
Outline

• Big Data all around us: the problems that Data Science is NOT addressing
• The CDO role and Data Axioms
• Some of the issues in BigData
• Case studies
  • Context and sentiment analysis
  • A case-study on IOT
  • Yahoo! predictions at scale
• Summary and conclusions
Why should banks worry about data?

100 years ago

• Smaller
• Much more local
• Deep understanding of customer needs
• Deep knowledge of customer

Now

• Digital, decline of the branch
• Global
• No face of the customer
• New generation of millennials

A big part of banking was knowing the customer intimately – personal relationships making KYC, risk modelling simple
Why data is important: Every customer interaction in an opportunity to capture data, learn & act

Imagine we could move back to this model 100 years ago – on demand, consumable, understandable information to build intimate relationships & an understanding of the customer.

Every interaction is an opportunity to capture data, to learn and to act.

- **Customer Interaction**
  - Millions of customer interactions per day
  - Big Data platform
  - Connie logs into BMB to check her balance. From browsing behaviour we know she seeks a car loan.
  - CRM data is used to pre-calculate Connie’s borrowing limit for a car loan.
  - Using internal/external data sources, predictive models, identify cross-sell opportunities.
  - Connie’s journey is enhanced based on previous multivariate testing results.

- **CRM**
  - An instant car loan product offering is displayed in the app.
  - Connie has a personalised journey based on calculated limit for the car loan amount.

- **Predictive Analytics**
  - Connie is offered a competitively priced ‘bespoke’ offer for car servicing / MOT.

- **Multivariate Testing**
  - User experience is continuously, iteratively improved by capturing user interaction in real-time every session.

**Data Capture/Opportunity to learn**

**Actions**

- **Targeted offers during browsing**
- **Product Discovery**
- **Cross-sell**
- **Measure feedback per session**

Invited keynote talk -- Copyright Usama Fayyad © 2015
Marketers want to be your local butcher

- Knowledgeable (Product)
- Trusted
- Knowledgeable (Person)
- Responsive
- Timely
- Anticipates Need

Trusted Interlocutors
But we are often just town criers

- Knowledgeable (Product)
- Unknown, perhaps trusted
- Knows what generic “customers” looks like
- Unresponsive
- Not individually timed
- Can’t anticipate

Monologists
What matters in the age of analytics?

1. Being able to exploit all the data that is available
   • Not just what you have available
   • What you can acquire and use to enhance your actions

2. Proliferating analytics throughout the organization
   • Make every part of your business smarter
   • Actions and not just insights

3. Driving significant business value
   • Embedding analytics into every area of your business can significantly drive top line revenues and/or bottom line cost efficiencies
Data Fusion: Can we bring all data together for analysis & action?

Central Data Fusion Engine
Ingesting, persisting, processing and servicing in Real-time.

- Customer Interactions
- Trade
- Network Traffic
- Transactions

Application Logs
- Social Data

DaaS

Analysis

- Cyber Security
- Fraud
- Financial Crime
- Marketing
- Risk
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

A Data Scientist is someone who
Knows a lot more software engineering than Statisticians
& Knows a lot more Statistics than software engineers
How Much Data? How Much of it is Structured?

• I do not like volume metrics, but they indicate appetite
  • Wide-ranging claims – mostly bout ‘data in the world’
  • We look at ‘Enterprise Data’ – Where companies pay money for storage as a good metric for “real data appetite”

60,000 Petabytes = 60 million Terabytes = 60 Trillion Gigabytes = 60 Exabytes
What about the Monsters and Giants?

- Consumers know about Megabytes… = 1 million bytes = 1000 Kbytes
- Consumers pay for “Gigs” – Gigabytes…
  “Mom: Do I get the 16Gig iPhone 6, or do I go for 32Gig?”
- Gigabyte from Greek “gigas” for giant = 1 billion bytes = 1000 Mbytes
- Terabyte comes from the Greek word for monster, "τέρας," – it is 1000 Gigabytes

The names get less creative after that:
- **Petabyte** from Greek "ΠΕΝΤΕ" for $1000^5 = 10^{15}$ bytes – 1K monsters
- **Exabyte** from Ancient Greek एक्स (ेक्स, “six”) = $(1K)^6 = 10^{18}$ – 1K Petabytes
- **Zettabyte** = $10^{21}$ – 1000 Exabytes
- **Yottabyte** = $10^{24}$
The 4-V’s of ‘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
“Classic” Data: e.g. Yahoo! User DNA

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

Registration  Campaign  Behavior  Unknown
How Data Explodes: really big

Professional netwk - reputation

Linkedin

Male, age 32

Lives in SF

Lawyer

Clicked on Sony Plasma TV SS ad

Searches on: “Italian restaurant Palo Alto”

Checks Yahoo! Mail daily via PC & Phone

Searched on from London last week

Searched on: “Hillary Clinton”

Clicked on Sony Plasma TV SS ad

Spends 10 hour/week on internet

Purchased Da Vinci Code from Amazon

Social Graph (FB)

Likes & friends likes

Blogs, publications, news, local papers, job info, accidents

Web searches on this person, hobbies, work, location

MetaData on everything
The Distinction between “Classic 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 analyzable data
Text Data: The Big Driver

• We speak of “big data” and the “Variety” in 3-V’s

• **Reality**: biggest driver of growth of Big Data has been text data
  o Most work on analysis of “images” and “video” data has really been reduced to analysis of surrounding text

  Nowhere more so than on the internet

• Map-Reduce popularized by Google to address the problem of processing large amounts of text data:
  o Many operations with each being a simple operation but done at large scale
  o Indexing a full copy of the web
  o Frequent re-indexing
A few words on: The Chief Data Officer
Why are companies creating this position?

• There is a fundamental realisation that Data needs to become a primary value driver at organizations
  o We have lots of Data
  o We spend much on it: in technology and people
  o We are not realising the value we expect from it

• A strong business need to create the CDO role:
  o Traditional companies are not following, but adopting the model that actually works in other data-intensive industries

• CDO has a seat at executive table: the voice of Data

• Data done right is an essential element to unify large enterprises to unlock value from business synergies
Fundamental Data Principles to Support Analytics

Usama’s Obvious Data Axioms
1. Data gains value exponentially when integrated and coalesced.

- When fragmented: dramatic value loss takes place;
- increased costs;
- reduced utility/integrity;
- and increased security risks
2. Fusing Data together from disparate/independent sources is difficult to achieve and impossible to maintain

*Hence only viable approach is:*
- Intercepting and documenting at the source
- Fusing at the source
- Controlling lifecycle and flow
Data Axioms

3. Standardisation is essential

– for sustained ability to integrate data sources and hence growing value;
– for simplifying down-stream systems and apps
– For enforcing discipline as a firm increases its data sources
Data Axioms

4. Data governance is critical and policy must be centralised
   – needs to be enforced strongly else we slip into chaos and a Babylon of terms/languages
   – An Enterprise Data Architecture spanning structured and unstructured data
5. Recency Matters \(\rightarrow\) data streaming in modelling and scoring

- Often, accuracy of prediction drops quickly with time (e.g. consumer shopping)
- Value of alerts drop exponentially with time...
- Ability to trigger responses based on real-time scoring critical
- Streaming, real-time model updates, real-time scoring
6. Data Infrastructure Needs:

- **rapid renewal & modernization:** the pace of change and development of technology are very rapid
  Design for migration and infrastructure replacement via abstraction layers that remove tech dependencies

- **Data Encryption and Masking:** Persisting unencrypted confidential and secret data (even within secure firewalls) is an invitation for problems and risks
Data Axioms

7. Data is a primary competency and not a side-activity supporting other processes
   – Hence specialized skills and know-how are a must
   – Generalists will create a hopeless mess
   – Data is difficult: modelling, architecture, and design to support analytics
Driving the need for integrated processes, data & architecture

<table>
<thead>
<tr>
<th>Regulatory Demand</th>
<th>Business Drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data quality, completeness, lineage and aggregation</td>
<td>• Better and smarter controls</td>
</tr>
<tr>
<td>• Company financial reporting presented at most granular level through various</td>
<td>• Capital precision</td>
</tr>
<tr>
<td>lenses: business units, legal entities, regional, sector level, …</td>
<td>• Better customer experience</td>
</tr>
<tr>
<td>• Faster turnaround for increasingly complex ad hoc data requests</td>
<td>• Better colleagues experience</td>
</tr>
<tr>
<td>• Additional sophisticated stress tests delivered in a joined-up manner</td>
<td>• More cost effective</td>
</tr>
<tr>
<td>between Risk and Finance</td>
<td></td>
</tr>
<tr>
<td>• Enhanced hybrid capital computation methodologies</td>
<td></td>
</tr>
</tbody>
</table>
Data Governance is BORING for most…

…but it shouldn’t be!
Reality Check

So I am a marketer, how do I use BigData for my business? Social Media? Sentiment Analysis?
Reality Check

So what should users of analytics in Big Data world do?
Load your Data into a “Data Lake”!

Your life will be so much easier as you can now do Data Acrobatics & other amazing data feats.
The Data Lake – according to Waterline

We loaded the Data!

Congratulations

Now What?
Amazon gives you facets

Product details

From a Data Lake to Amazon Browser
Reality check: So where do analysts & Data Scientists spend all their time?

Let’s Mine the Data

- Prepared view of Data
- Manage access
- Manage data views
- Normalization and Regularization
- Data cleaning
- Data lineage
- ETL
- Data Selection
- Metadata Management
- Feature Extraction
- Data aggregations
- Entity Extraction
- Attribute & Data changes
- Managing the Data sources and Data feeds
**Reality check:** So what do Technology people worry about these days?

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
Drivers of Hadoop in large enterprises

Cost of storage

Fastest growing demand is more storage

Data in Data Warehouses have traditionally required expensive storage technology:

- **$20K to $50k** per terabyte per year – cost of premium storage

- **$2K** per terabyte – much lower per year – cost of Hadoop on commodity storage
Analysis & programming software

Event Survey - Languages

Languages:
- SQL
- Python
- Java
- C/C++
- R
- Perl
- BizIntell
- Pig
- SAS
- SPSS

Hadoop
PIG
HIPI
Disco
RevolutionAnalytics / RHadoop

massive data - minimal code
Hadoop Stack
Big Data Landscape
Reality check: If storage is biggest driver of Hadoop adoption, what is next biggest?

ETL

- Extract – Transform – and Load
- Replaces expensive licenses
- Much higher performance with lower infrastructure costs (processors, memory)
- Flexibility in changing schema and representation
- Flexibility on taking on unstructured and semi-structured data
- Plus suite of really cool tools...

From ETL → ELT

- Extract – Load - Transform
- “computation comes to data, data does not have to leave for processing”
Turning the 3-Vs of Big Data into \textit{Value}

\textbf{Understand context and content}

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

\textbf{Understand community sentiment}

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

\textbf{Understand customer intent?}

- What is each individual trying to achieve?
- Can we predict what to do next?
- Critical in cross-sell, personalization, monetization, advertising, etc…
# Many business uses of predictive analytics

<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 behaviour</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>credit worthiness of a customer requesting a loan, secured and unsecured loans</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>
<tr>
<td>Application Log File Analytics</td>
<td>Understanding application, network, event logs in IT</td>
</tr>
</tbody>
</table>
## CRM Mining in Retail

**Business challenges are difficult to address with standard queries:**

<table>
<thead>
<tr>
<th>Business Unit</th>
<th>Business Challenge</th>
<th>Data Mining Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merchandising</td>
<td>Average Order Size</td>
<td>Cross sell &amp; Up sell</td>
</tr>
<tr>
<td>Merchandising</td>
<td>Assortment Planning</td>
<td>Shopping Basket analysis</td>
</tr>
<tr>
<td>Marketing</td>
<td>Customer Acquisition</td>
<td>Response Modeling, Conversion propensity by source</td>
</tr>
<tr>
<td>Marketing</td>
<td>Customer Retention</td>
<td>Churn prediction</td>
</tr>
<tr>
<td>Marketing</td>
<td>Customer Re-activation</td>
<td>Targeted promotions</td>
</tr>
<tr>
<td>Retail Operations</td>
<td>Geographic Profiling</td>
<td>Metro area segmentation and segment migration</td>
</tr>
</tbody>
</table>

**Discover profitable segments**

**Boost customer acquisition & retention**

**Maximize cross-sell and up-sell opportunities**

**Personalize customer experience to increase loyalty**

**Identify most effective marketing campaigns**
Case Studies:
1. Context Analysis (unstructured data)
2. IOT Case Study
3. Yahoo! Predictive Modeling
Reality Check

So who is the company we think is best at handling Big Data for Understanding the context in a web page?
Case Study: Biggest Big Data in advertising?
Understanding context for ads

What Ad would you place here?
Case Study: Biggest Big Data in advertising? Understanding context for ads

Body parts delivered to Michigan home

Damaging to Brand?

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.

It’s UPS early morning delivery. « Rollover for video.»
The Display Ads Challenge Today

What Ad would you place here?

Violence continues in Greece as rioters firebomb buildings
Protesters in Athens torch offices and cars amid clashes with police after memorial for teenager
NetSeer: Solving accuracy issues
Ambiguity, waste, brand, safety

Why did Google Serve this Ad?

this is how NetSeer actually sees this content
NetSeer: How it works

**ECONOMY CARS**
- electric vehicles
- fuel efficiency
- economy vehicles
- high MPG
- low emission
- safety rating
- ford
- service record

**VISION TOOLS**
- microscope lenses
- bifocal refraction
- eye chart
- autofocus
- reading glasses

**MARKET RESEARCH**
- focus groups
- blind study
- surveying
- A/B testing
- consumer study

**DISCERNS AND MONETIZES HUMAN INTENT**
+ Identifies Concepts expressed on a page
+ Disambiguates language
+ Builds increasingly rich profile over time

**WEBSITE.COM**

**CONCEPTS**
- 52M
- 2.3B

**RELATIONSHIPS BETWEEN CONCEPTS**
NetSeer: Intent for Display
Positive vs negative content re a particular topic

Prevent a Stroke By Drinking Caffeine

Some caffeine addicts may well be on their way to prevent a stroke later in life. A new study out of Sweden revealed that women who consumed more than a cup of caffeinated coffee daily reduced their risk of stroke by 22 to 25 percent.

In the study, researchers followed 34,670 women (ages 49 to 83) for an average 10.4 years. Those that drank little or no caffeinated coffee had an increase instance of stroke. No word if the same would hold true with men but we can just drink more coffee and hope for the best.

This study looked at caffeine naturally found in coffee so to say that energy drinks would prevent a stroke as well would be a bit of a stretch. Sweden was the 6th most caffeinated country last year so it’s a great place to conduct caffeine research. This new study gives some new insight into the health benefits of caffeine and some good evidence that women can prevent a stroke by making caffeinated coffee a big part of the diet.

Source: American Heart Association

Diet Coke®
Great Taste. 0 Calories. Stay Extraordinary.

Lipton® Green Tea
Start Something Good With Flavored Lipton®

Starbucks® K-Cup® Packs
Look for our varieties where you buy groceries.

BARCLAYS
**Problem:** Hard to understand user intent

**Contextual Ad served by Google**

- **Nailgun: Insanely Fast Java**
- **Thinnest. Client. Ever.**
- **NetSeer**
- **URL:** [http://martiansoftware.com/nailgun/](http://martiansoftware.com/nailgun/)

**What NetSeer Sees:**

- 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

ContextLinks © by NetSeer

www.netseer.com
Case Studies:
1. Context Analysis (unstructured data)
2. IOT Case Study
3. Yahoo! Predictive Modeling
The Connected Cow

This part of the talk is borrowed from:

Joseph Sirosh
VP, Machine Learning
Microsoft
Do cows need to take 10,000 steps a day?
Every company is a Bigdata company...
What a dairy farmer can do...
Detect health issues early and prevent loss

Improve cattle production by accurate detection of estrus
Effect of estrus detection rate on increasing pregnancy rate

<table>
<thead>
<tr>
<th>CONCEPTION RATE</th>
<th>ESTRUS DETECTION RATE</th>
<th>PREGNANCY RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>70%</td>
<td>55%</td>
<td>39%</td>
</tr>
<tr>
<td>70%</td>
<td>65%</td>
<td>46%</td>
</tr>
<tr>
<td>70%</td>
<td>75%</td>
<td>53%</td>
</tr>
<tr>
<td>70%</td>
<td>85%</td>
<td>60%</td>
</tr>
<tr>
<td>70%</td>
<td>95%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Source: "Detection of Standing Estrus In Cattle"
FS921B – George Perry, Beef Reproduction and Management Specialist, Animal and Range Sciences Department

UP TO 70% IMPROVEMENT
But this is hard...
Estrus lasts only for 12-18 hours every 21 days.

Occurs mostly between hours of 10 pm and 8 am.
How can farmers tell when the time is right for hundreds of cows?
AI finally meets AI
GYUHO "Cow Step" Service

**SENSOR**

- COWSHE D
  - Dipole Antenna
  - Receiver
- Pedometer

**NETWORK**

- Base Station
  - Internet

**ANALYTICS**

- Microsoft Azure
  - Cloud Center

**COMMUNICATION**

- House/Office
  - 3G Card Router
  - Device
# Variability in estrus cycles of farm animals

<table>
<thead>
<tr>
<th>ANIMAL</th>
<th>LENGTH OF ESTRUS CYCLE (DAYS)</th>
<th>DURATION OF ESTRUS</th>
<th>DURATION OF ESTRUS</th>
</tr>
</thead>
</table>
| **CATTLE** | 19-23
Average: 21 | 6-30 hours
Average: 18 hours | 6-30 hours
Average: 18 hours |
| **HORSE** | 10-37
Average: 21 | 2-6 days
Average: 4 days | |

Source: Handbook of Livestock Management, Richard Battaglia
# Economic effects before and after utilizing Fujitsu GYUHO system on 11 dairy farms

<table>
<thead>
<tr>
<th>STOCK FARMER NAME</th>
<th>BREEDING NUMBERS</th>
<th>BEFORE GYUHO INTRODUCTION</th>
<th>AFTER GYUHO INTRODUCTION</th>
<th>SHORTENED DAYS</th>
<th>ANNUAL PRODUCTION INCREASED HEADS</th>
<th>ANNUAL INCREASE IN %</th>
<th>INCOME INCREASE BY PRODUCTION INCREASE JPY(350,000/) CATTLE (♂ + ♂ AVERAGE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DAYS-OPEN</td>
<td>INTRAPARTUM INTERVAL</td>
<td>DAYS-OPEN</td>
<td>INTRAPARTUM INTERVAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 A STOCK FARMER</td>
<td>180</td>
<td>78</td>
<td>63</td>
<td>348</td>
<td>15</td>
<td>8</td>
<td>4%</td>
</tr>
<tr>
<td>2 B STOCK FARMER</td>
<td>262</td>
<td>74</td>
<td>59</td>
<td>344</td>
<td>15</td>
<td>12</td>
<td>5%</td>
</tr>
<tr>
<td>3 C STOCK FARMER</td>
<td>110</td>
<td>96</td>
<td>66</td>
<td>351</td>
<td>26</td>
<td>8</td>
<td>7%</td>
</tr>
<tr>
<td>4 D STOCK FARMER</td>
<td>202</td>
<td>54</td>
<td>40</td>
<td>330</td>
<td>9</td>
<td>6</td>
<td>3%</td>
</tr>
<tr>
<td>5 E STOCK FARMER</td>
<td>498</td>
<td>78</td>
<td>51</td>
<td>336</td>
<td>27</td>
<td>40</td>
<td>8%</td>
</tr>
<tr>
<td>6 P STOCK FARMER</td>
<td>201</td>
<td>154</td>
<td>74</td>
<td>359</td>
<td>80</td>
<td>37</td>
<td>18%</td>
</tr>
<tr>
<td>7 G STOCK FARMER</td>
<td>537</td>
<td>115</td>
<td>62</td>
<td>347</td>
<td>53</td>
<td>75</td>
<td>14%</td>
</tr>
<tr>
<td>8 H STOCK FARMER</td>
<td>273</td>
<td>217</td>
<td>66</td>
<td>351</td>
<td>151</td>
<td>85</td>
<td>31%</td>
</tr>
<tr>
<td>9 I STOCK FARMER</td>
<td>173</td>
<td>137</td>
<td>67</td>
<td>352</td>
<td>70</td>
<td>29</td>
<td>17%</td>
</tr>
<tr>
<td>10 J STOCK FARMER</td>
<td>248</td>
<td>83</td>
<td>50</td>
<td>335</td>
<td>33</td>
<td>24</td>
<td>10%</td>
</tr>
<tr>
<td>11 K STOCK FARMER</td>
<td>151</td>
<td>102</td>
<td>69</td>
<td>354</td>
<td>33</td>
<td>13</td>
<td>9%</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>258</td>
<td>108</td>
<td>61</td>
<td>346</td>
<td>47</td>
<td>31</td>
<td>12%</td>
</tr>
</tbody>
</table>
Case Studies:
1. Context Analysis (unstructured data)
2. IOT Case Study
3. Yahoo! Predictive Modeling
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

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

More people visited Yahoo! in the past month than:

• Use coupons
• Vote
• Recycle
• Exercise regularly
• Have children living at home
• Wear sunscreen regularly

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 ads to consumers whose recent behaviors online indicate which product category is most relevant to them.
• On a per consumer basis: maintain a behavioral/interests profile and profitability (user value and LTV) metrics
How it works | Network + Interests + Modelling

Identify Most Relevant Users

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

Active now…

…and with feeling

6 months ago (casual interest) 3 months ago (research) present (highly engaged)

Person A

Person B

Ads

Searches

Page Views

Page Views
Differentiation | Category specific modelling

Example 1: Category Automotive

Example 2: Category Travel/Last Minute

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 with no further activity, decay takes effect.

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.
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.

Results from a client campaign on Yahoo! Network

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
Brand Ads and Search Ads Interact!

- *Is ad search strategy enough for a direct marketer?*
- *Do brand ads play a role in search advertising?*
- *Harris Direct Case Study*
Case Study: Harris Direct

Viewing These Ads:

- Gomez #1
- Independence
- And $100 credit

Had This Effect On:

- Aided Brand Awareness
  - Up 7%
- Brand Favorability
  - Up 32%
- Purchase Intent
  - Up 15%
Case Study: Harris Direct

People who saw display ads were **61% more likely to search** on related topics...

- **...and drove 139% more clicks** on algorithmic and sponsored links...
- **...specifically driving 249% more sponsored search clicks**...
- **...and driving 91% more activity** on the HarrisDirect.com website.
Custom Database Match: A Case Study

**Challenge**
A leading cosmetics brand wanted to model Y! users against their database of customers to generate a brand-centric target that more efficiently drove online revenue

**Objective**

1. Conduct a consumer database match, profile matched users, and create a model to find other similar users on our network.

2. Understand the efficiencies gained from modeled users (in driving online revenue)

3. Understand how modeled users can be used in conjunction with Y! behavioral targets for awareness & consideration campaigns
1. **Database Match**
   Match Yahoo Users with client’s customer database

2. **Profile**
   Identify unique behavioral identifiers of matched users

3. **Model**
   Find users with similar behavioral DNA

4. **Validate**
   Are modeled users a more brand receptive target than BT
Model
Users modeled against client database match

vs.

BT
CPG/Beauty And Personal Care/Cosmetics

Compare:
• History With Client Products*
  • Display Click Through Rate
  • Banner Opt-In/1000
  • Online Conversions
  • Cost Per Online Sale
  • Online Revenue Per Ad $ Spent

* Marketing Evolution Survey
**Modeled Users are a more brand-centric target**

Does CDBM modeling increase our opportunity of hitting a more brand friendly target?

<table>
<thead>
<tr>
<th>% Lift (vs. BT)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brand Awareness</strong></td>
</tr>
<tr>
<td>Unaided Brand Awareness +66%</td>
</tr>
<tr>
<td><strong>Brand Associations</strong></td>
</tr>
<tr>
<td>Industry Leader +24%</td>
</tr>
<tr>
<td>High Quality +19%</td>
</tr>
<tr>
<td><strong>Past Product Usage</strong></td>
</tr>
<tr>
<td>Overall Usage +36%</td>
</tr>
<tr>
<td>Make-Up +44%</td>
</tr>
<tr>
<td>Skin Care +38%</td>
</tr>
<tr>
<td>Woman’s Fragrance +35%</td>
</tr>
</tbody>
</table>
BT outperforms model on front end metrics...

BT Users*...

1. Were 29% more likely to click on an ad

2. Were 48% more likely to opt-in a banner (per 1000 impressions)

3. Lowered Cost Per Banner O/I by 34%
Modeled users drove more value to the business on the back end

**Modeled Users***

1. Drove 331% more online transactions
2. Lowered Cost Per Online Sale by 170%
3. Increased Revenue Per Dollar Spent by 211%

* vs. BT (CPG/Beauty And Personal Care/ Cosmetics)
Key Takeaways

1. BT targets are highly engaged and more likely to interact with your advertising – great for awareness building campaigns.

2. Audiences modeled from client databases can produce more brand centric targets with greater efficiencies in driving online conversions.
Experience summary at Yahoo!

- Dealing with one of the largest data sources (25 Terabyte per day)
- Behavioral Targeting business was grown from $20M to > $400M in 3 years of investment!
- Yahoo! Specific? -- BigData critical to operations
  - Ad targeting creates huge value
  - Right teams to build technology (3 years of recruiting)
  - Search is a BigData problem (but this has moved to mainstream)
Lessons Learned

A lot more data than qualified talent

• Finding talent in BigData is very difficult
• Retaining talent in BigData is even harder

At Yahoo! we created central group that drove huge value to company

Data people need to feel like they have critical mass

• Makes it easier to attract the right people
• Makes it easier to retain

Drive data efforts by business need, not by technology priorities

• Chief Data Officer role at Yahoo! – now popular
Where does this leave us?

1. What matters in the age of analytics?.
   • Being able to exploit all the data that is available
   • Proliferating analytics throughout the organization
   • Driving significant business value

2. Where are we today?
   • New Data Landscape via Hadoop
   • Confusion by marketers, analysts, and technical community
   • Struggling with basics of managing data because of the new flexibility

3. What are the real issues?
   • Data management and governance
   • Talent for Data and Data Science is rare but critical
   • Data and Analytics are specialisms that need management and know-how: not for generalists…
The early days of mass auto production
Today’s Auto:
It just works!
No need to understand what happens when you turn on ignition.
Very complex inside, but all simplicity on the outside.
Thank you & questions

Usama Fayyad
Geraldine.Bolton@barclays.com

https://uk.linkedin.com/in/ufayyad
www.barclays.com/joinus