Taming the BigData Beast for Value & Insights

Usama Fayyad, Ph.D.

Executive Chairman -*Oasis500*

***Chairman & CTO** - Blue Kangaroo*

Twitter: @usamaf

Sept 18th, 2013

Financial Times Live

Dubai - UAE

OASIS500

Usama Fayyad, Executive Chairman
usama@oasis500.com   Twitter: **@usamaf**

Blue Kangaroo™

Outline

- Big Data all around us
- Introduction to Data Mining and Predictive Analytics Over BigData
- Some of the issues in BigData
- On-line data and facts
- Case studies on on-line marketing: Yahoo! Big Data
- Summary and conclusions

OASIS500

Blue Kangaroo™

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 the 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

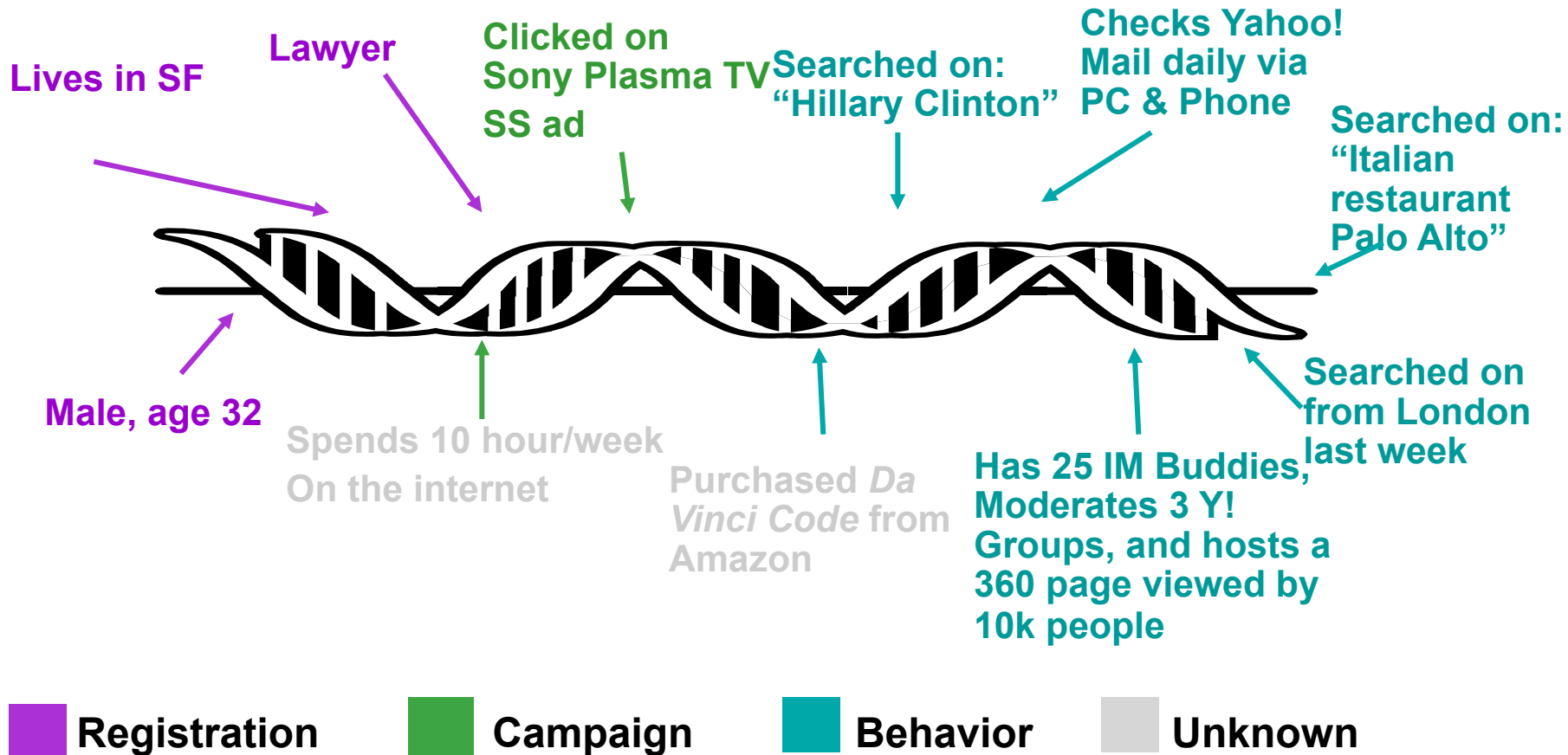OASIS500

Blue Kangaroo™

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"
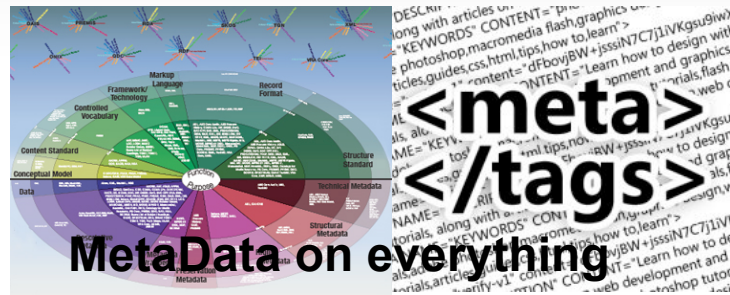 - Discovery and learning on the spot

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.

OASIS500

Blue Kangaroo™

Today's Data: e.g. Yahoo! User DNA

Lives in SF

Lawyer

Clicked on Sony Plasma TV SS ad

Searched on: "Hillary Clinton"

Checks Yahoo! Mail daily via PC & Phone

Searched on: "Italian restaurant Palo Alto"

Male, age 32

Spends 10 hour/week On the internet

Purchased *Da Vinci Code* from Amazon

Has 25 IM Buddies, Moderates 3 Y! Groups, and hosts a 360 page viewed by 10k people

Searched on from London last week

| ■ Registration | ■ Campaign | ■ Behavior | ■ Unknown |

How Data Explodes: really big

Professional netwk - reputation

MetaData on everything

Web searches on this person, hobbies, work, location

Lawyer

Lives in SF

Clicked on Sony Plasma TV SS ad

Searched on: "Hillary Clinton"

Checks Yahoo! Mail daily via PC & Phone

Searched on: "Italian restaurant Palo Alto"
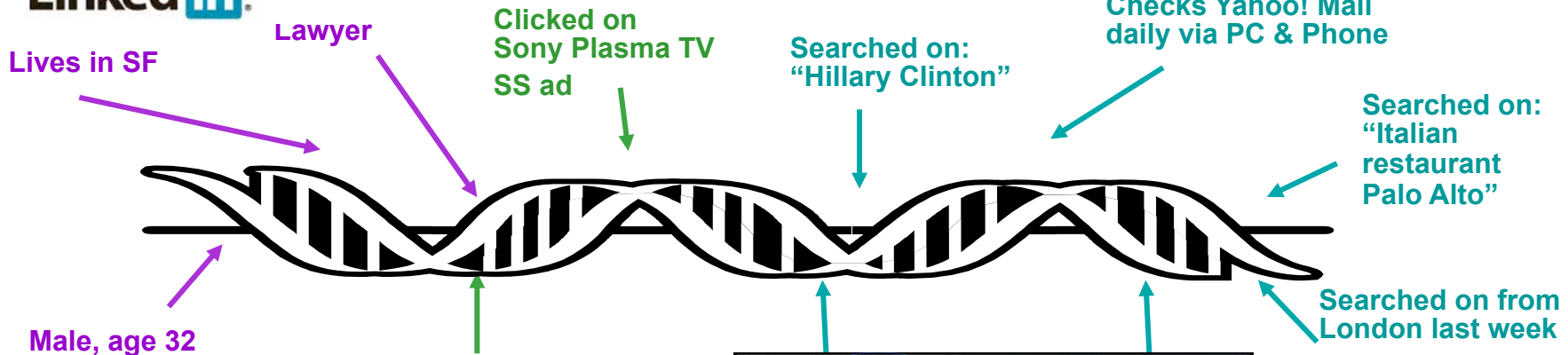
Male, age 32

Spends 10 hour/week Internet

Searched on from London last week

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

Likes & friends likes

Social Graph (FB)

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 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
  - 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:
  - Many operations with each being a simple operation but done at large scale
  - Indexing a full copy of the web
  - Frequent re-indexing

OASIS500

Blue Kangaroo™

Reality Check on Brand/ Reputation

What are people saying about my brand on Social Media?

This is data of people blogging and tweeting about your products on the internet? How do you react to this?
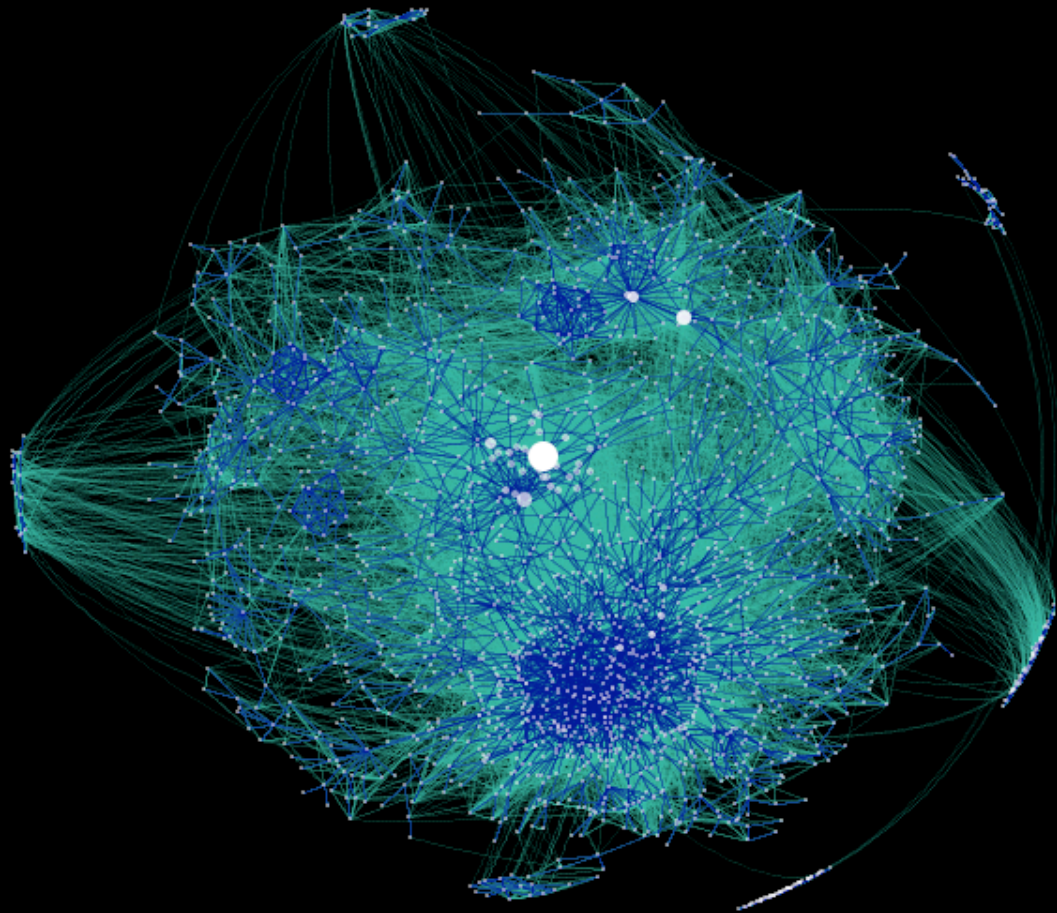
Reality Check

Surely there are companies I can work with that can help me make this practical?

OASIS500

Blue Kangaroo™

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?

Reality Check

So what do technology people worry about these days?

OASIS500

Blue Kangaroo™

To Hadoop or not to Hadoop?
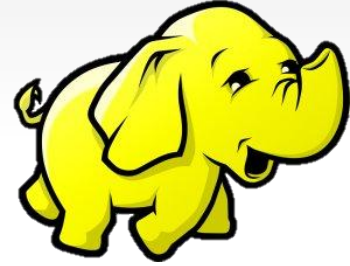
when to use techniques requiring Map-Reduce and grid computing?

- 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

To Hadoop or not to Hadoop?
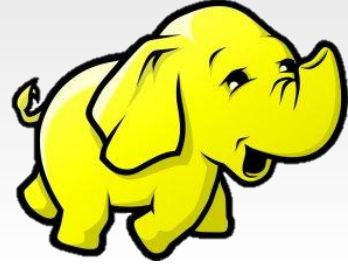
when to use techniques requiring Map-Reduce and grid computing?

- 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

OASIS500

Blue Kangaroo™

Hadoop Use Cases by Data Type

ERP Financial Data 1%
Supply Chain Data 2%
Sensor Data 2%
Financial Trading Data 4%
CRM Data 4%
Science Data 7%
Advertising Data 10%
Social Data 11%
Text and Language Data 16%
IT Log Data 19%
Content and Preference Data 24%
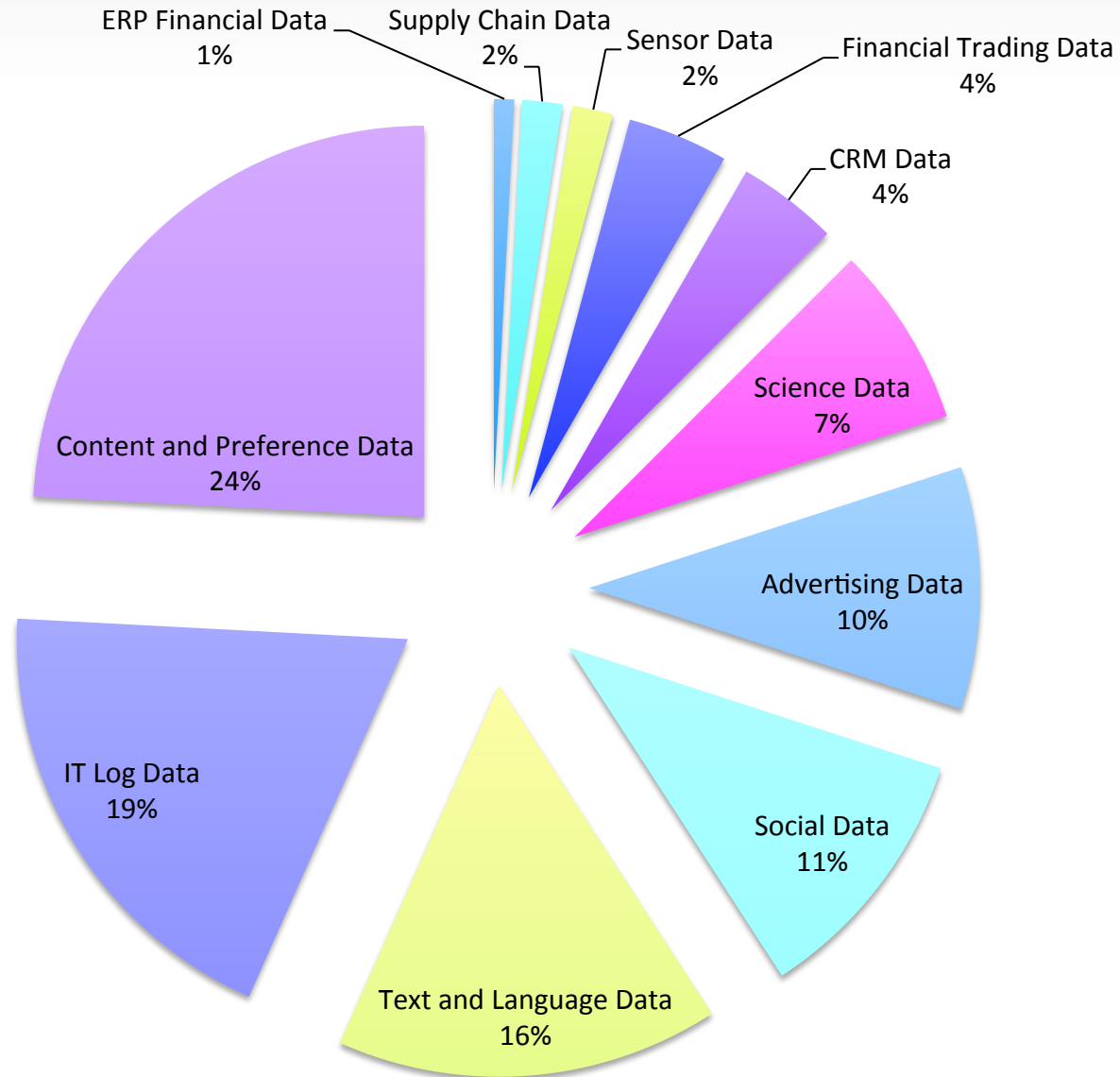
Analysis & Programming Software
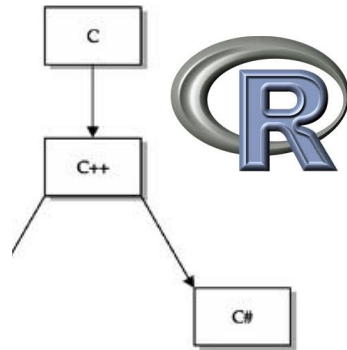
Event Survey - Languages

BigData Overview – Copyright Usama Fayyad © 2013   19
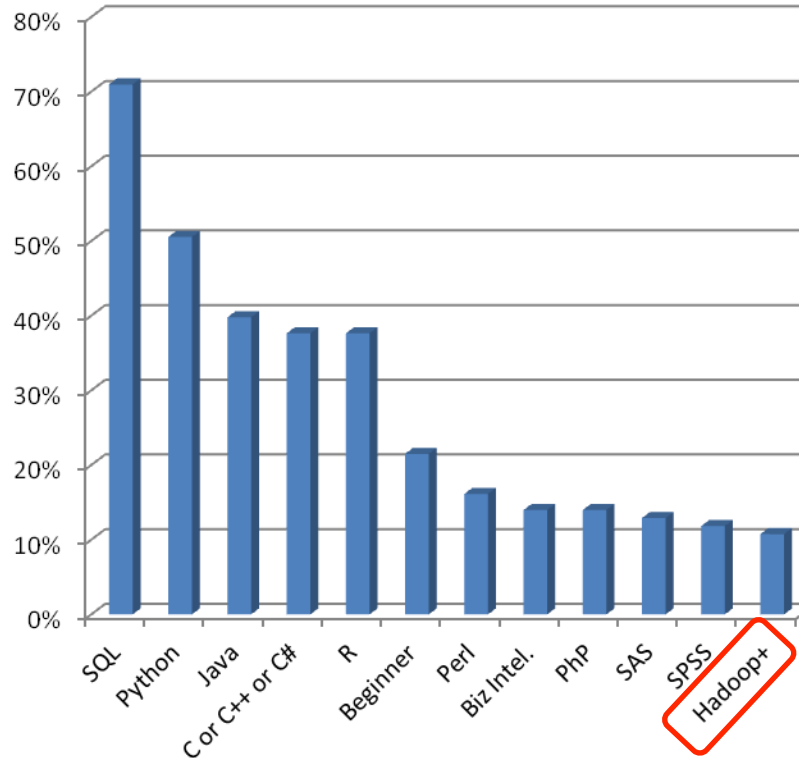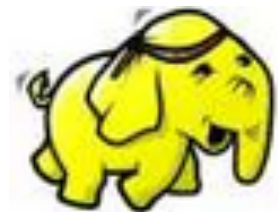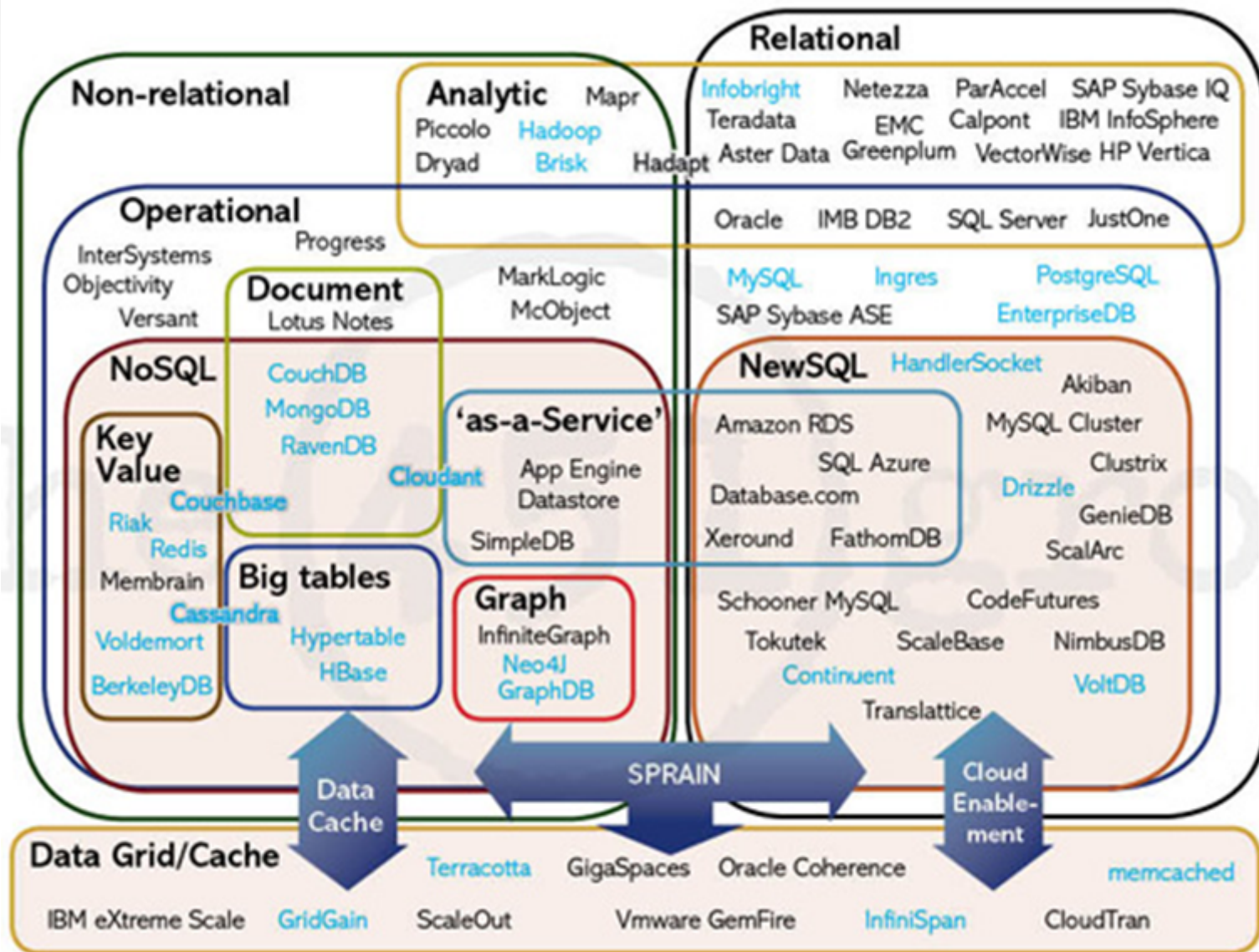
Many Business Uses

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

So Internet is a big place with 2B+ users and lots happening?

- ***Do we understand what each individual is trying to achieve?***

- ***Do we understand what a community's sentiment is?***

- ***Do we understand context and content?***

Social platforms 2013

- **User accounts on FaceBook:**
  - **971M**
  - **How many fake profiles?**
  - **83M (per FB 8/2012 report)**
- **How many users on LinkedIn**
  - **159.3M (as of 1/2013)**

- **How many Google+ users?**
  - 343 million active users in Q4 2012

> » **Sources: International Business Times 1/28/2013**
> – http://www.ibtimes.com/google-plus-becomes-worlds-no-2-social-network-after-facebook-knocking-twitter-1042956

OASIS500

Blue Kangaroo™

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%**

Interesting Events

- **Google:** How many searches in 2012?
  - More than 1.2 Trillion (source Google)
  - Estimates 1B to 3B per day
- **Twitter:** How many Tweets/day?
  - 500M (per CEO Dick Costolo at IAB Engage, 10/2012)
- **Facebook:** Updates per day?
  - More than 1B
- **YouTube:** Views/day
  - 4 Billion hours/month - 4B views/day in 1/2012
  - 72 hours of video uploaded every minute!
- Social Networks: users who have used sites for spying on their partners?
  - **56%**

Sources: Feb.2012 - compiled from Comscoredatamine.com, Nielsen.com, thisDigitalLife.com, PewInternet.org

OASIS 500

Blue Kangaroo™

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

Blue Kangaroo™

So Internet is a big place with lots happening?

- *Do we understand what each individual is trying to achieve?*

- *Do we understand what a community's sentiment is?*

- *Do we understand context and content?*

Turning the 3 V's of Big Data

Into **Value**

Turning the three Vs of Big Data into **Value**

Understand context and content
- *What are appropriate ads?*
- *Is it Ok to associate my brand with this content?*
- *Is content sad?, happy?, serious?, informative?*

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

Understand user intent?
- *What is each individual trying to achieve?*
- *Critical in monetization, advertising, etc…*

Blue Kangaroo™

Understanding Context

Reality Check

So who is the company we think is best at handling BigData?

OASIS500

Blue Kangaroo™

Biggest BigData in Advertising?

Understanding Context for Ads

The Display Ads Challenge Today

The Display Ads Challenge Today

The Display Ads Challenge Today

What Ad would you place here?

The Display Ads Challenge Today

Irrelevant and Damaging to Brand

Completely Irrelevant

News › World news › Greece

Violence continues in Greece as rioters firebomb buildings
Protesters in Athens torch offices and cars amid clashes with police after memorial for teenager

Anil Dawar
guardian.co.uk, Sunday 21 December 2008 17.05 GMT
Article history

A larger | smaller

World news
Greece

More news

Win a mini-break in gorgeous Greece...

A youth assaults a police officer in Athens during a week a teenager. Photograph: Bela Szandelszky/AP

Related Topics: make it snacktastic working it summer living summer travel summer beauty

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

by Bliss Tree, on Tue Jun 7, 2011 4:58am PDT
627 Comments | Post a Comment | Read More from This Author » | Report Abuse

f Share   retweet 187   Email   Print

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-slurping

What is Your Credit Score as of 06/15/2011

TransUnion.   478
Experian   511
EQUIFAX   540

See Yours Instantly!   FreeScore.com

Shine on facebook

OASIS500   Big
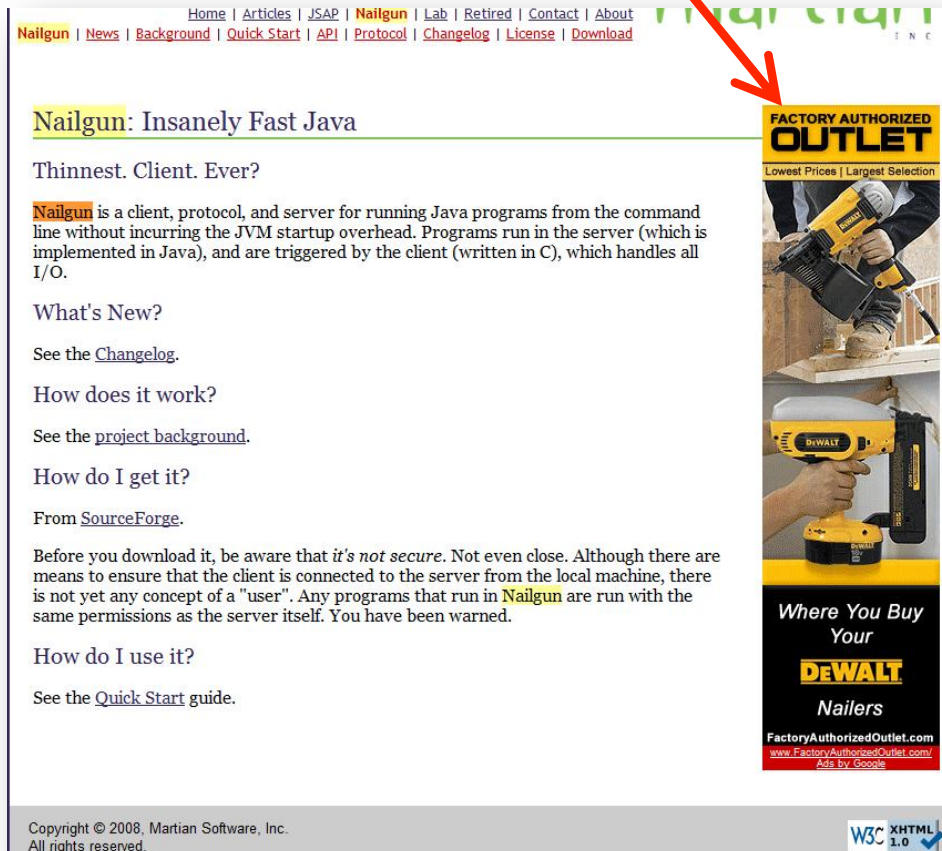
NetSeer: Intent for Display

- Currently Processing 4 Billion Impressions per Day

Problem: Hard to Understand User Intent

Contextual Ad served by Google

What NetSeer Sees:

Turning the three Vs of Big Data into **Value**

Understand context and content
- *What are appropriate ads?*
- *Is it Ok to associate my brand with this content?*
- *Is content sad?, happy?, serious?, informative?*

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

Understand user intent?
- *What is each individual trying to achieve?*
- *Critical in monetization, advertising, etc…*

Blue Kangaroo™

User Intent

User Intent
Case Study #1

Yahoo! Behavioral Targeting

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
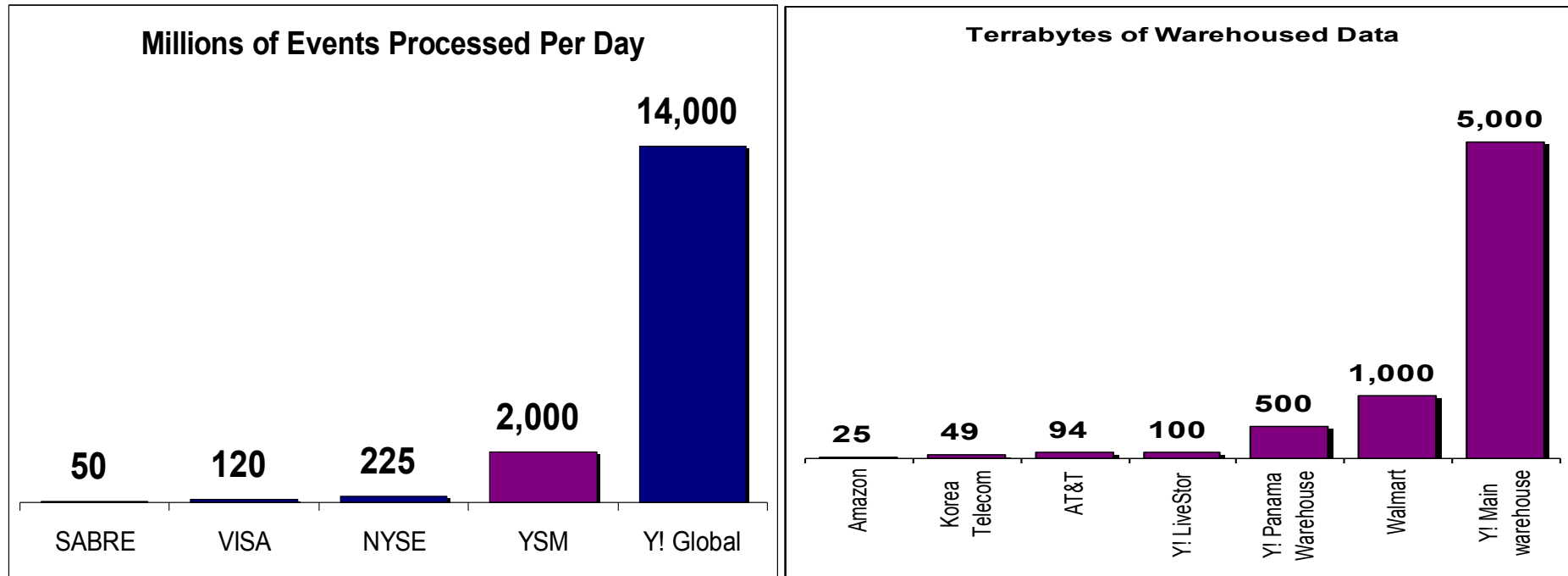
More people visited Yahoo! in the past month than:

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

OASIS500

Blue Kangaroo™

Yahoo! Big Data – A league of its own…

Millions of Events Processed Per Day

50	120	225	2,000	14,000
SABRE	VISA	NYSE	YSM	Y! Global

Terrabytes of Warehoused Data

25	49	94	100	500	1,000	5,000
Amazon	Korea Telecom	AT&T	Y!LiveStor	Y! Panama Warehouse	Walmart	Y! Main warehouse

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)

Content

Search

YAHOO!
BT

Search Clicks

Ad Clicks

Targeting ads to **consumers** whose **recent behaviors online indicate** which **product category** is **relevant** to them

Yahoo! User DNA

Registration **Campaign** **Behavior** **Unknown**

Lives in SF

Lawyer

Clicked on Sony Plasma TV SS ad

Searched on: "Hillary Clinton"

Checks Yahoo! Mail daily via PC & Phone

Searched on: "Italian restaurant Palo Alto"

Male, age 32

Spends 10 hour/week On the internet

Purchased *Da Vinci Code* from Amazon

Has 25 IM Buddies, Moderates 3 Y! Groups, and hosts a 360 page viewed by 10k people

Searched on from London last week
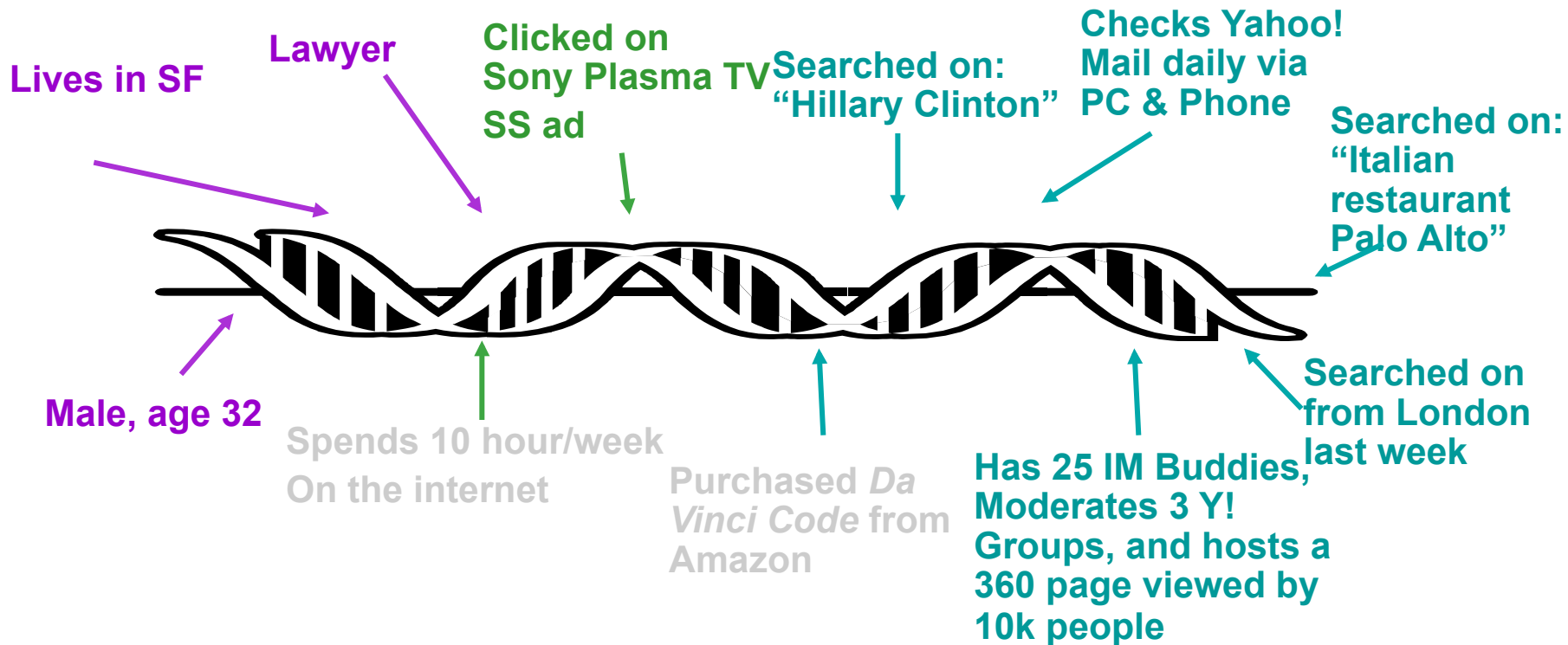
- *On a per consumer basis: maintain a behavioral/interests profile and profitability (user value and LTV) metrics*

OASIS500

Blue Kangaroo™

How it works | Network + Interests + *Modelling*

Identify Most Relevant Users
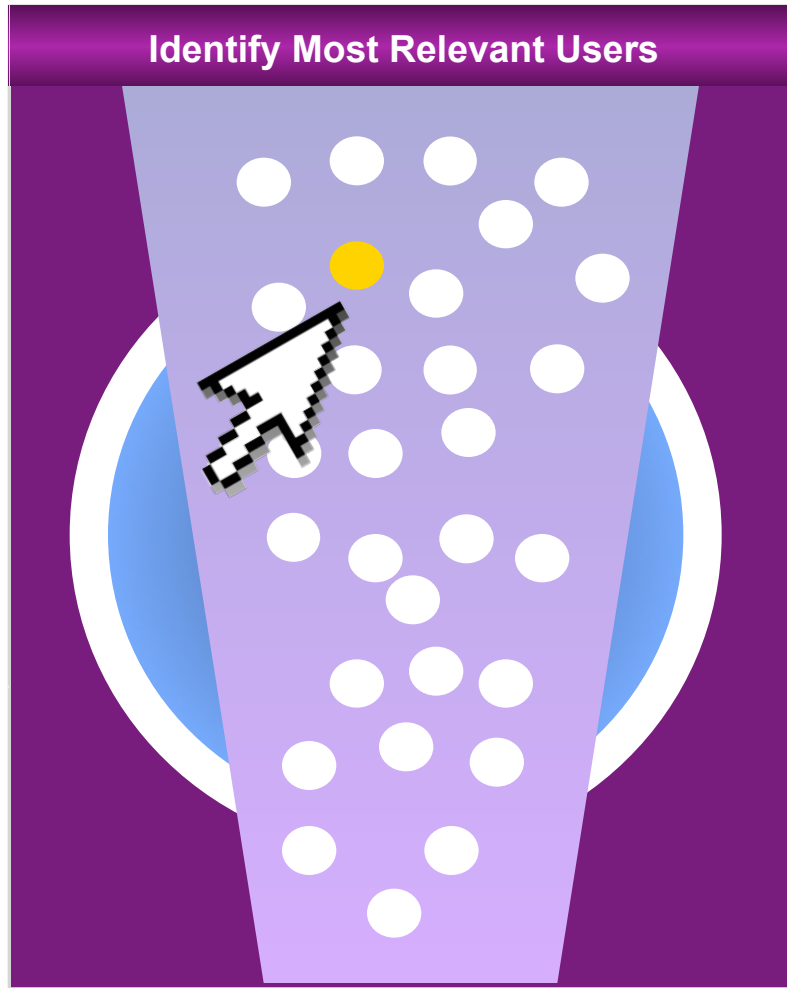
Analyze predictive patterns for purchase cycles in over 100 product categories

⬇

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

OASIS500

Blue Kangaroo™

Recency Matters, So Does Intensity

Active now…
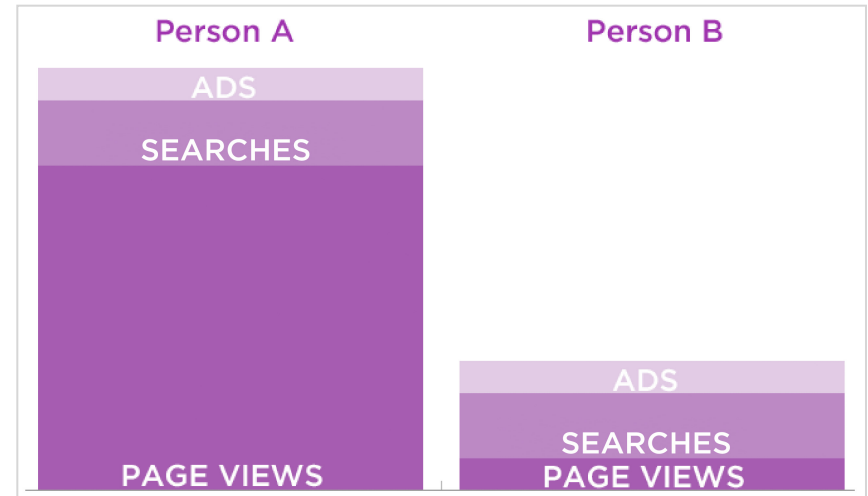
6 months ago (casual interest) · 3 months ago (research) · present (highly engaged)

…and with feeling

Person A · Person B

ADS
SEARCHES
PAGE VIEWS

ADS
SEARCHES
PAGE VIEWS

Differentiation | Category specific modelling

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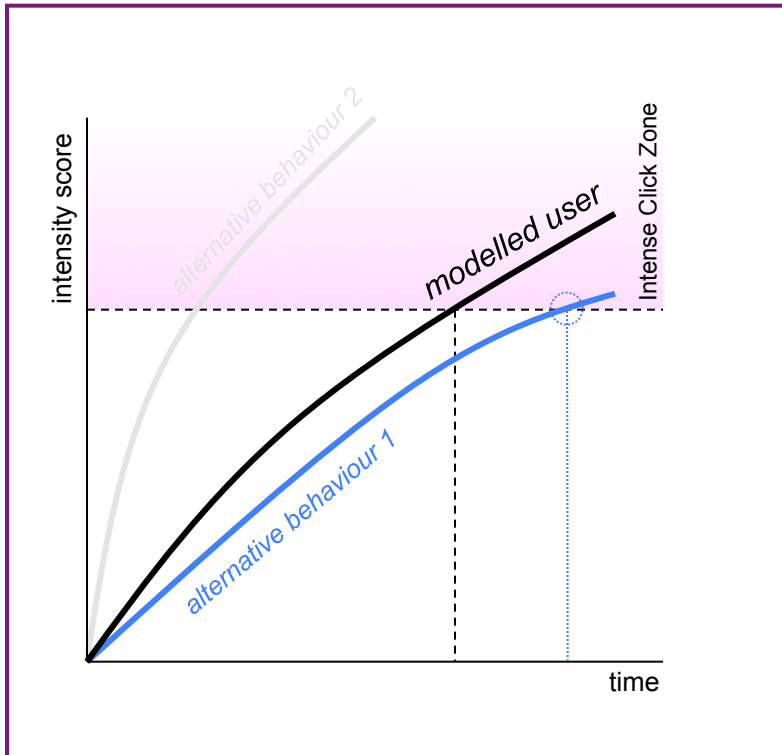
Alt Behaviour 1: 5 pages, 2 search keywords, 1 search click, 1 ad click

Example 2: Category Travel/Last Minute
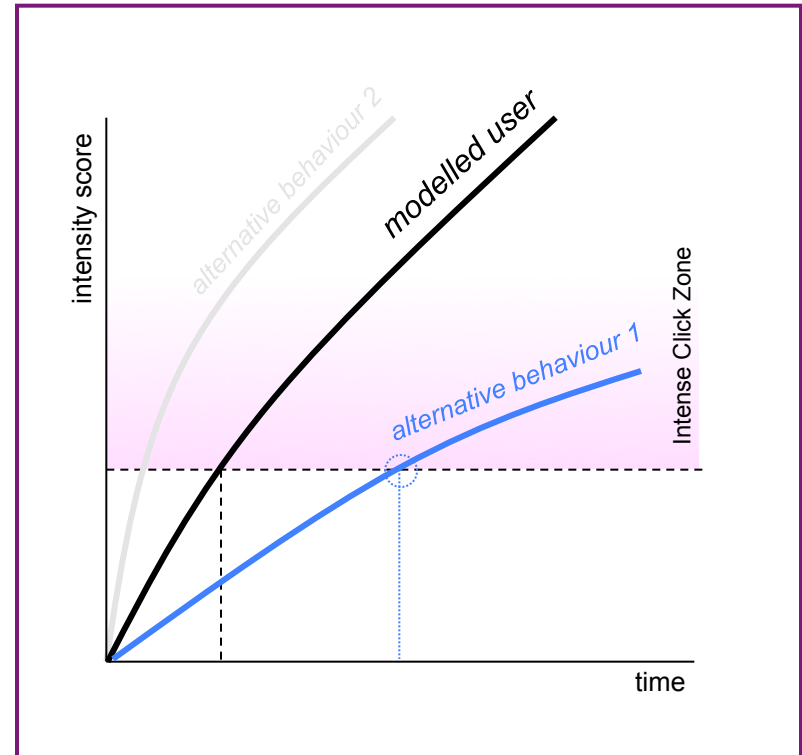
Alt Behaviour 1: 5 pages, 2 search keywords, 1 search click, 1 ad click

Different models allow us to weight and determine intensity and recency

Differentiation | Category specific modelling
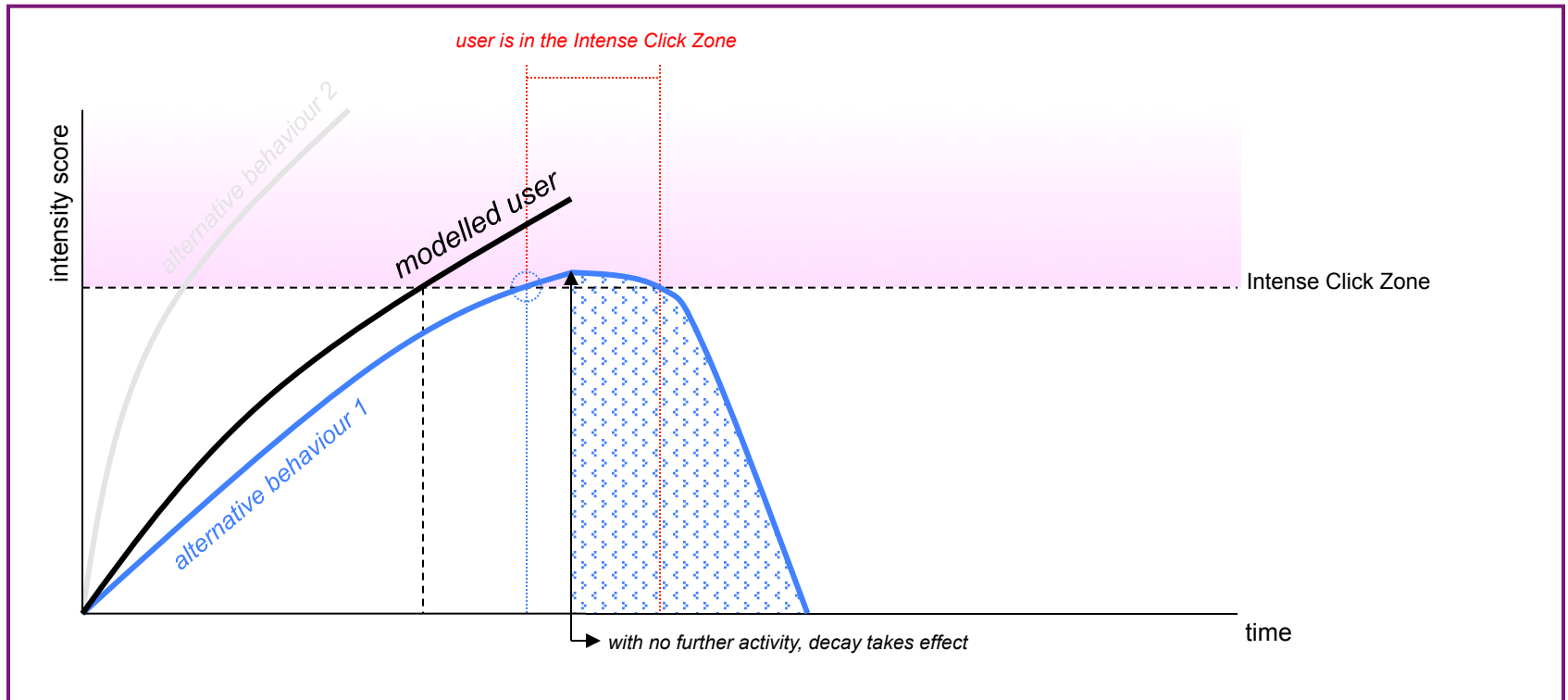
Example 1: Category Automotive

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

OASIS500

Blue Kangaroo™

Mortgage Intender Example

Example: Mortgages

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

Results from a client campaign on Yahoo! Network

+626% Conv Lift

+122% CTR Lift

Y! Finance Audience   Mortgage Loans Target   Mortgage Loans Target
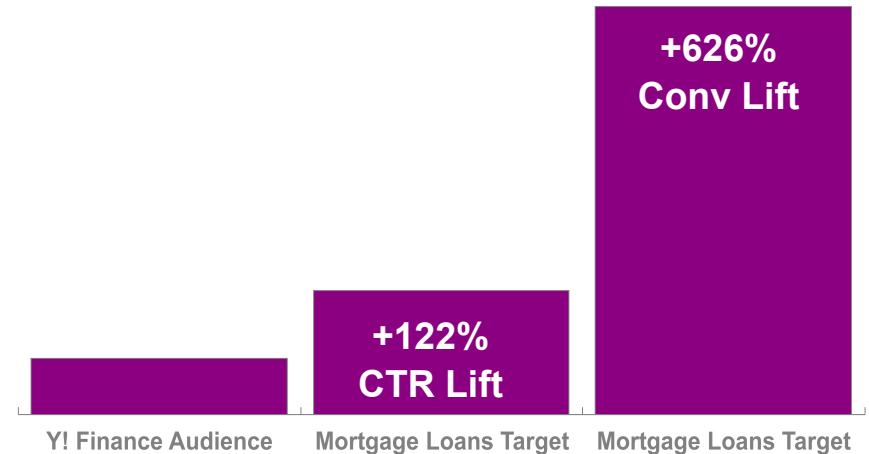
Example search terms qualified for this target:

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

OASIS500

Date: March 2006

Blue Kangaroo™

Experience summary at Yahoo!

- Dealing with one of the largest data sources (25 Terabyte per day)

- BT business was grown from $20M to about $500M in 3 years of investment!

- BigData critical to operations
  - Ad targeting creates huge value
  - Right teams to build technology (3 years of recruiting)
  - Search is a BigData proble,

- Big demands for grid computing (Hadoop)
  - Not all BigData can be handled via Hadoop
  - Spunoff BigData Segmentation data platfrom: nPario

OASIS500

Blue Kangaroo™

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
  - Makes it easier to retain

- Drive data efforts by business need, not by technology priorities
  - Chief Data Officer role at Yahoo! – now popular

OASIS500

Blue Kangaroo™

BigData Analytics for Organizations

- Key to competitive Intelligence:
  - Understand context
  - Understand intent

- Key to understanding consumer trends through social media analysis
  - Brand issues
  - Trend issues
  - Anticipating the next shift

Big Picture on Big Data Analytics

Key points

Sometimes, Simple is Very Powerful!

Retaining New Yahoo! Mail Registrants

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

OASIS500

Blue Kangaroo™
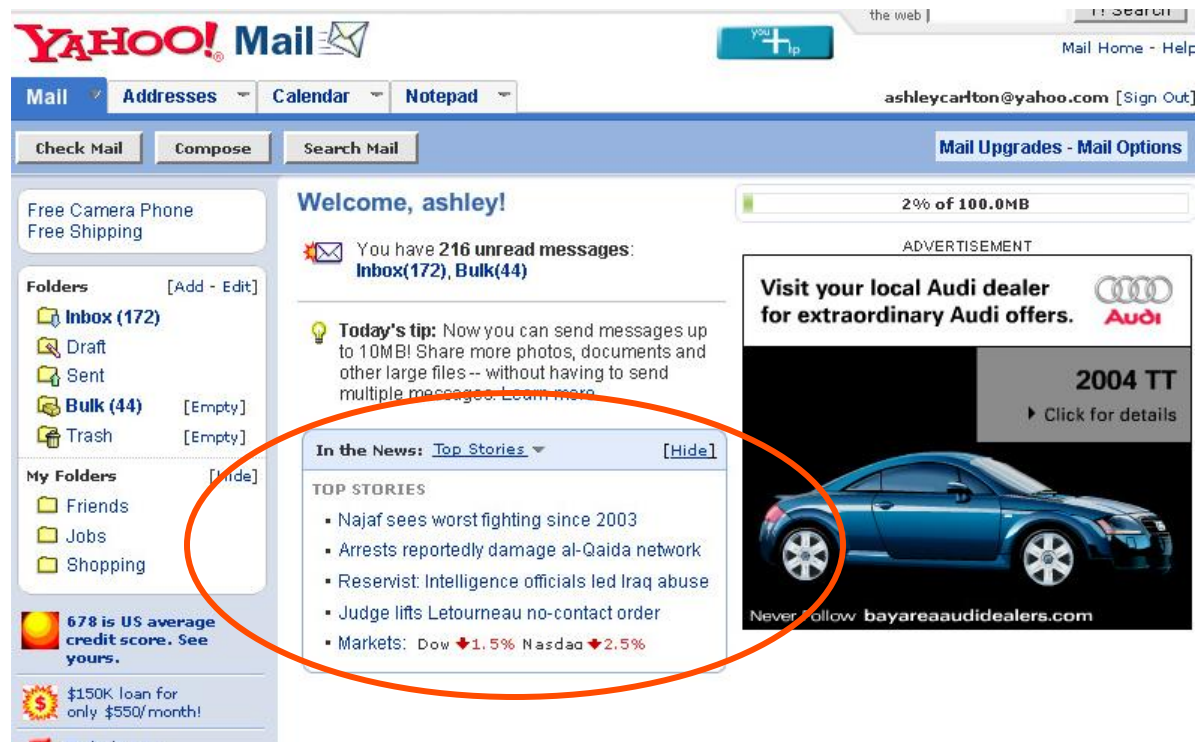
Threats & Opportunities

- Data world is changing, especially in on-line businesses
- Major shifts from relational DB to NoSQL, document-oriented stores
- Connecting new world to "old" world?
  - Convenience of execution – integration with data platforms
  - Appropriateness of algorithms to BigData
  - Unstructured data algorithms:
    - Text, Semi-structured and Unstructured data
    - Entity extraction a must
    - Appropriate theory and probability distributions (power laws, fat tails)
    - Sparse Data
  - Model management and proper aging of models
  - Getting to basics so we can decide what models to use:
    - Understanding noise and distributions
    - Data tours

OASIS500

Blue Kangaroo™

Thank You! & Questions

Usama Fayyad -   usama@fayyad.com
                        usama_fayyad@yahoo.com

Twitter – @Usamaf

☎ +1-206-529-5123

www.BlueKangaroo.com
www.Oasis500.com
www.open-insights.com