



# Finding Information Outside the Firewall

Gábor Pécsy

Senior Manager, Data Enrichment

June 25, 2014

# MELT WATER

formula of your health



## Oslo, Shack 15 Summer 2001



Start capital: 15,000 USD

20,000  
clients

900  
employees

**The global leader  
in media intelligence**

90  
countries

60  
offices

# Budapest Office

- Opened in 2009
- Research and Development only
  - Content acquisition
  - Data Enrichment
- 15 people, growing

mest  
meltwater entrepreneurial  
school of technology



Not-for-profit NGO fully funded and run by Meltwater

# Back to Finding Information Outside the Firewall



# Our product vision



Morning coffee

Informed  
decisions



# Our Journey

Transformed  
the paper  
clipping  
industry



Global leader  
in media  
intelligence



Global leader  
in open data  
intelligence

Online news

Online news  
Social media

Online news  
Social media  
New data types

Crawling and  
Searching

NLP: Topics  
and Sentiment

Structuring the  
Public Web

2008

2012

2016

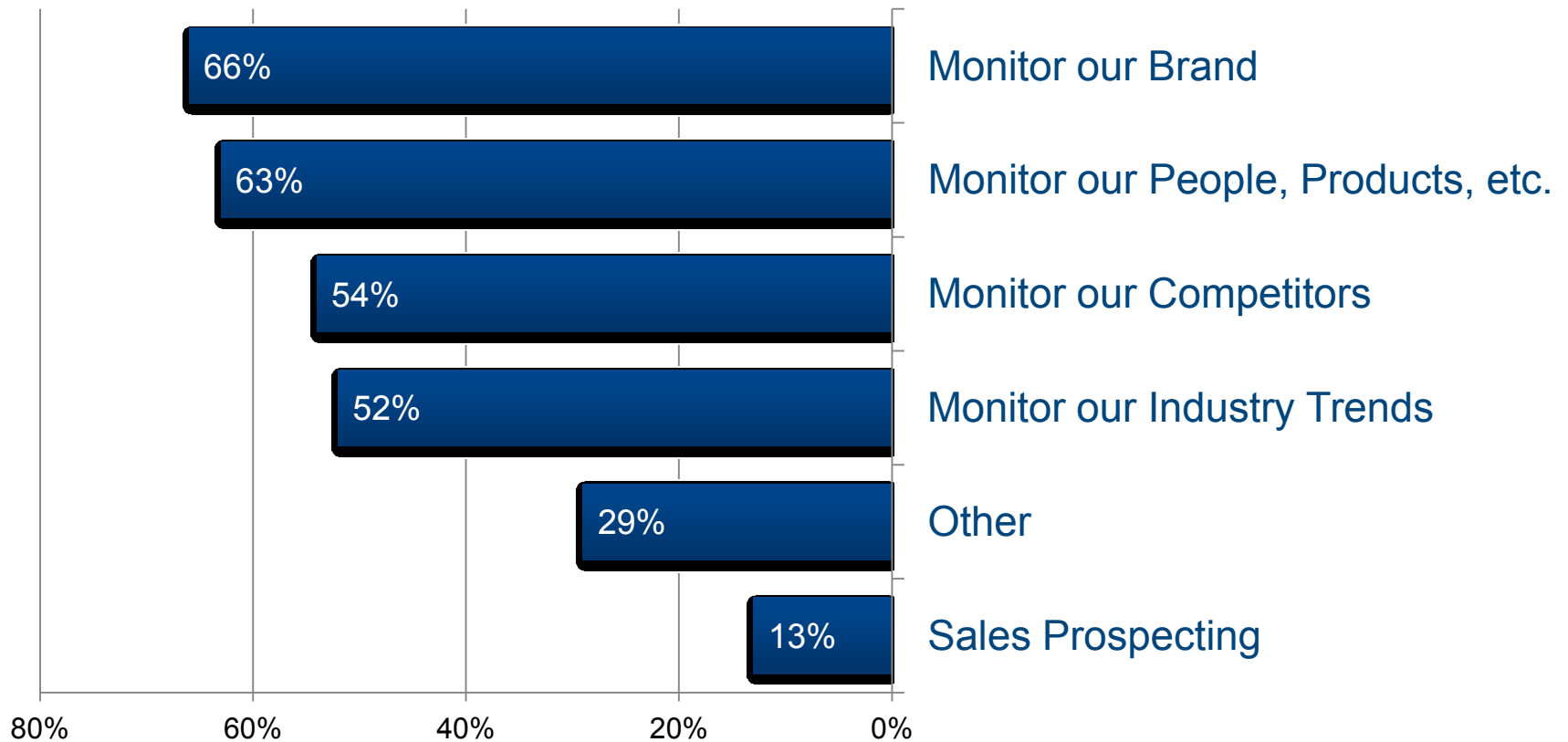
# Already much more than PR (by industry)

## Meltwater Revenue by Industry (2013)



# Already much more than PR (by Use-Case)

## What else do you use Meltwater for?



Survey of Meltwater clients in Sweden (2011)

# Prospecting

Industry: Security Services

Country: Sweden



*We use Meltwater to find out about new instances of vandalism and break-ins. Often, the victim is in need of our services.”*

# Product Management

Industry: FMCG

Country: United Kingdom



*“Meltwater helps us determine how public perception of certain ingredient chemicals will influence adoption sales.*



# Performance Measurement

Industry: Television

Country: India



*Meltwater is the best way for us to monitor the performance and popularity of our news anchors and programs.”*

# Competitive Intelligence

Industry: Pharmaceuticals

Country: England



*“Surprisingly, we use Meltwater to be alerted of when certain patent will expire in target markets.*



# Risk Management

Industry: Precious Metals

Country: Global

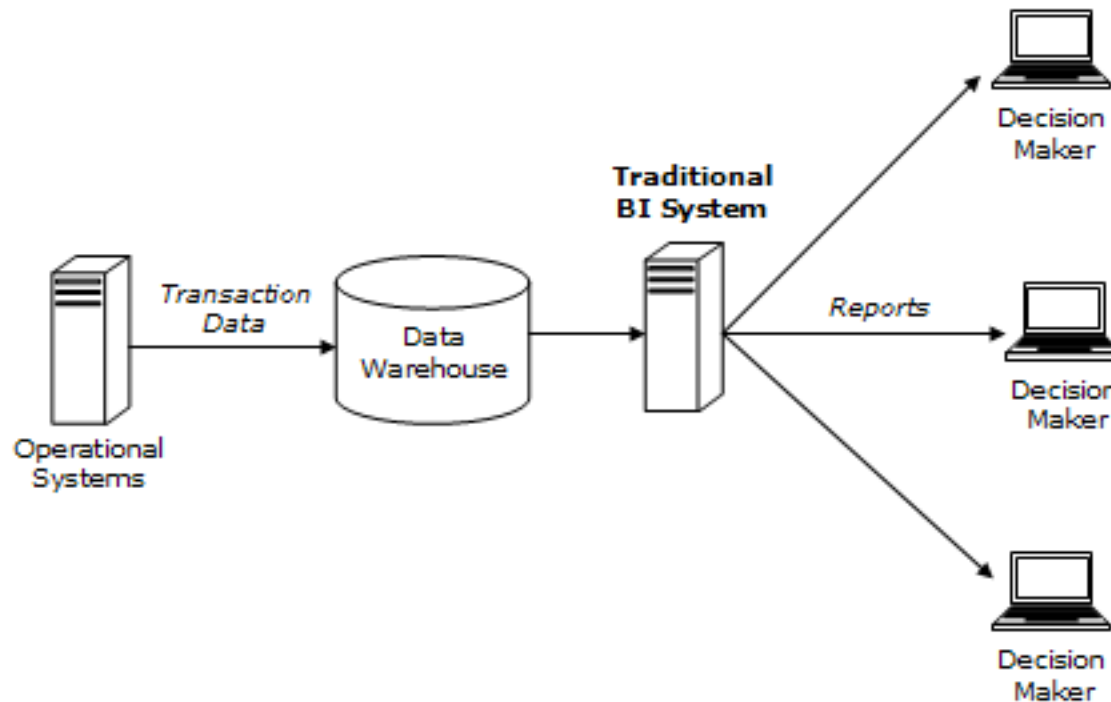


*One of the best ways we've found to estimate and prevent infrastructure attacks is deep social listening using Meltwater."*



# Traditional BI

- Traditionally, BI focused on data owned by the company: **data within the firewall.**
- We have well-established tools for this



# The Information Explosion

The background of the slide is a dynamic, abstract composition. It features a central vertical axis from which numerous lines radiate outwards, creating a sense of depth and movement. The color palette is dominated by shades of blue, teal, and light green, with some darker, almost black, lines. Scattered throughout the scene are various digital symbols, including binary digits (0s and 1s), circular patterns, and rectangular shapes, all appearing to float or stream through the space. The overall effect is one of high-speed data flow and digital complexity.

2012:  
2.5 QUADRILLION ( $10^{18}$ )  
BYTES OF DATA DAILY



**3.6 Billion people**

Talking, tweeting, posting, blogging,  
liking, ranting, retweeting, sharing opinions about

**YOUR PRODUCTS, COMPANY, MARKET,  
COMPETITORS**

etc.

An invaluable source  
of information

**Companies need to learn to ride this  
information wave**

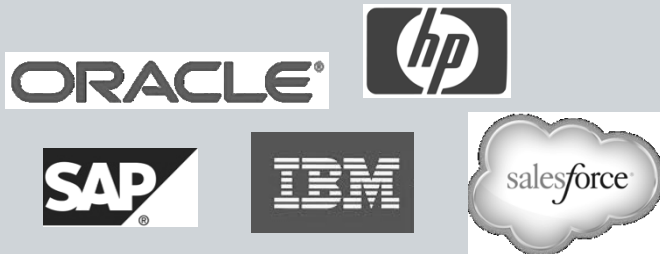
**Or it will sweep them away.**



# A New Industry Category

## Classic Business Intelligence

30 years of Business Intelligence and workflow automation based on internal data...



## Open Data Intelligence

...is being disrupted by an explosion of open data.



Real-time  
Unfiltered  
Benchmarked

# Differences to Traditional BI

- Classical BI deals with data
  - Internal, easy to access
  - Small to medium volume
  - Structured data
  - Dense
  - Clean
  - Private
- Data on the Internet:
  - External, not so easy to get
  - Big Data
  - Usually unstructured or semi-structured
  - Sparse
  - Noisy, sometimes unreliable
  - Public, semi-public

# Challenges – Data Collection

- Lots of open/free data – needs to be found, scraped
  - Requires special experience
  - Copyright issues – is it really free?
- Data aggregators – cost money
  - Can be targeted (more relevant content)
  - Often with enrichments (basic NLP, influence scores, ranking etc.)
  - Don't cover the full content landscape

# Challenges – Data Volume

- Information on Internet is more sparse
- Need to process large quantities of data to get enough information
- Large-scale computing – special competence, expensive hardware and software
  - Service-based models: SaaS, PaaS
- BIG Data technologies



elasticsearch



riak





# Challenges – Unstructured Data

- Internal data: often structured, e.g. in RDBMS, data warehouse
- Data on internet:
  - Unstructured: information is available as text, images, audio, video
  - Semi-structured: textual data with some known inner structure (e.g. a patent document)
  - Rarely structured, usually costs money
- Extracting information from unstructured data: Natural Language Processing
  - Language specific
  - Social “dialects” – acronyms, emoticons, slang, typos etc.



# Basic NLP Services

- Language detection
- Sentiment analysis



# Basic NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction

Query: Obama

The screenshot shows the CNN.com website with a search bar containing the query 'Obama'. The search results are displayed on the 'World' page. The main article is titled 'Partial recount ordered in Afghanistan election'. The text of the article is partially visible. To the right, there is a 'Latest News' section with several headlines. Two headlines are highlighted with green boxes: 'In school speech, Obama talks up education' and 'Commentary: Obama as teacher is chief'. Below the 'Latest News' section is a 'Popular News' section with two headlines: 'SNL' replaces two cast members' and 'Laura Bush praises Obama, bemoans excessive partisanship'. The second headline is also highlighted with a green box. Two blue arrows point downwards from the highlighted phrases to the 'Extracted Phrases' section.

## Extracted Phrases:

obama speech  
obama urges students  
obama lectures students  
obama stresses responsibility  
bush praises obama

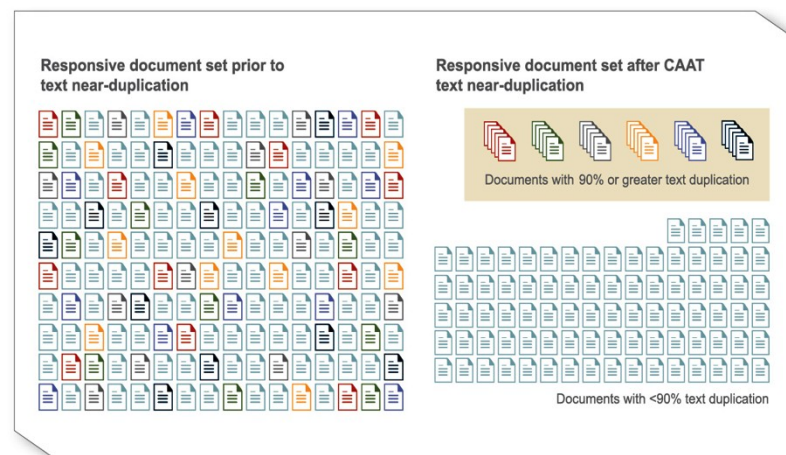
# Basic NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization



# Basic NLP Services

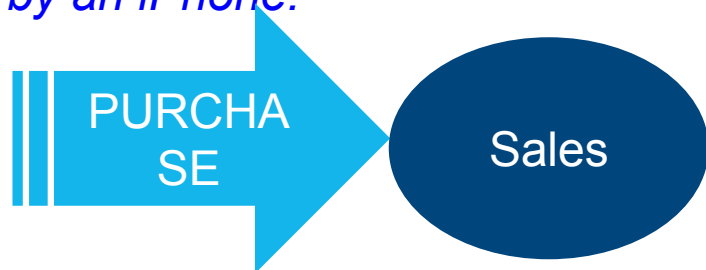
- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection



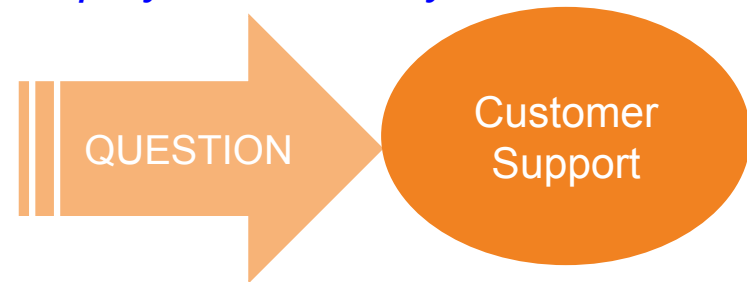
# NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection
- Intent detection

*"I want to buy an iPhone."*



*"How can I play music on my iPhone?"*



# Using Intent Detection

- Direct customer interactions
  - Purchase Intent: sales offer
  - Question Intent: customer support
- Analytics
  - Correlation of purchase intents and sales figures in a region.
  - Comparisons to competitor products



# NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection
- Intent detection
- Named Entity Recognition

Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.

Person
Location
Organization
Date
Nationality
Title

# NLP Services

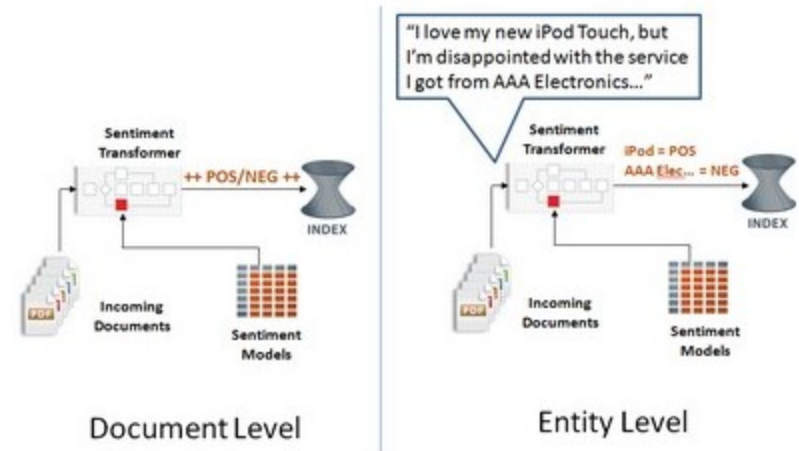
- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection
- Intent detection
- Named Entity Recognition
- Entity-level Sentiment

Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.

Person
Location
Organization
Date
Nationality
Title



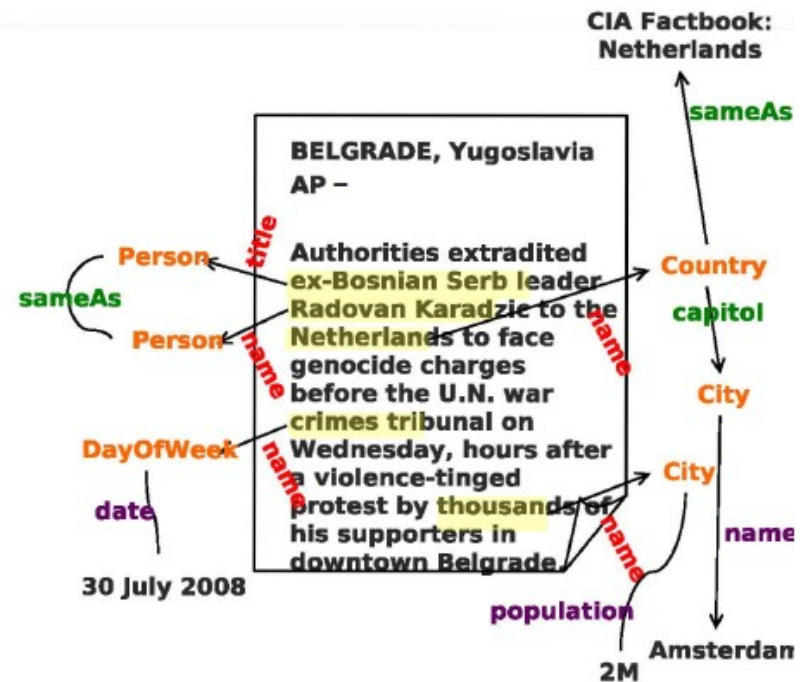
# NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection
- Intent detection
- Named Entity Recognition
- Entity-level Sentiment
- Named Entity Disambiguation



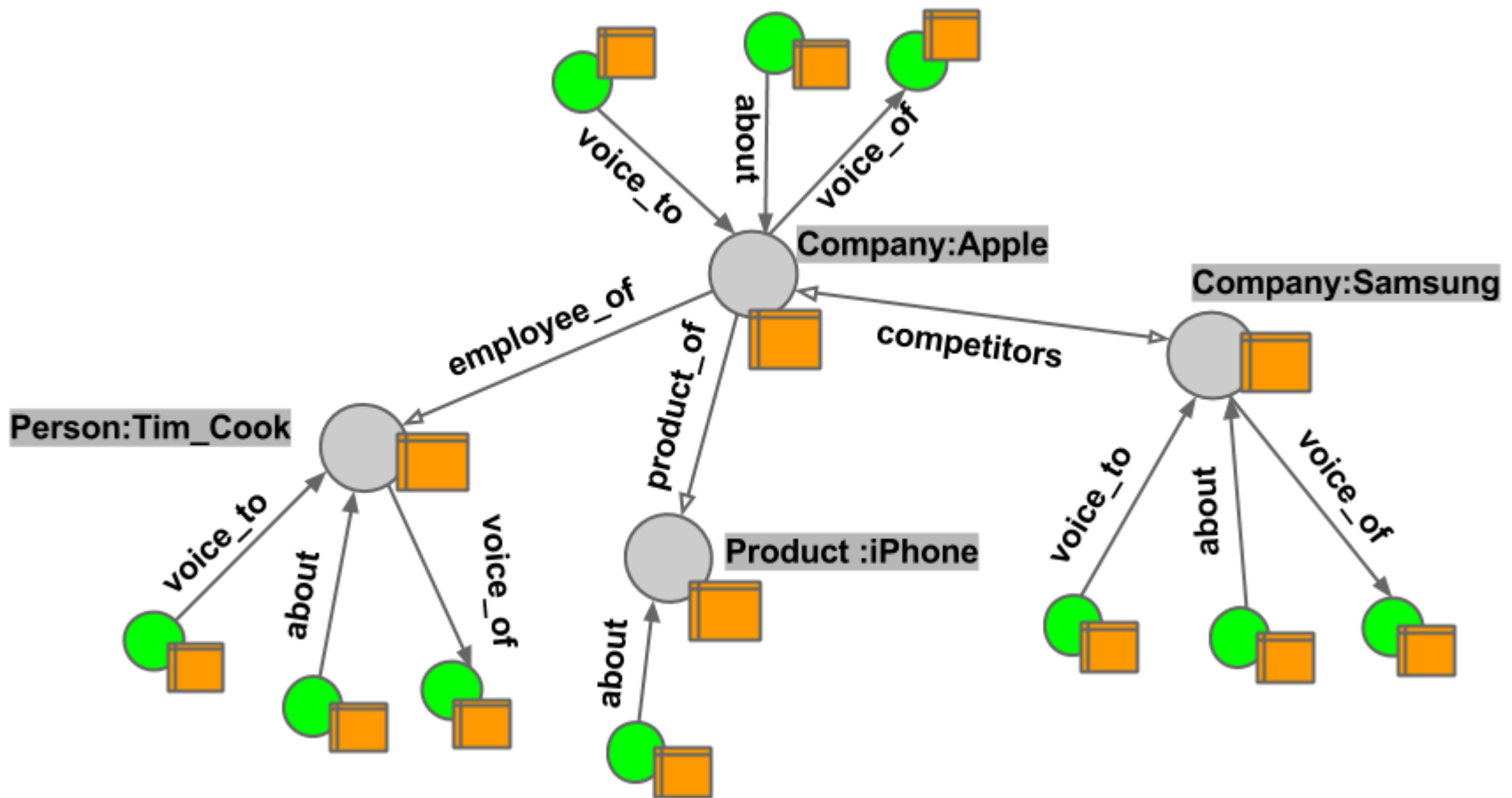
# NLP Services

- Language detection
- Sentiment analysis
- Key phrase extraction
- Content Categorization
- Near duplicate detection
- Intent detection
- Named Entity Recognition
- Entity-level Sentiment
- Named Entity Disambiguation
- Entity Relationship Extraction



# Graph search

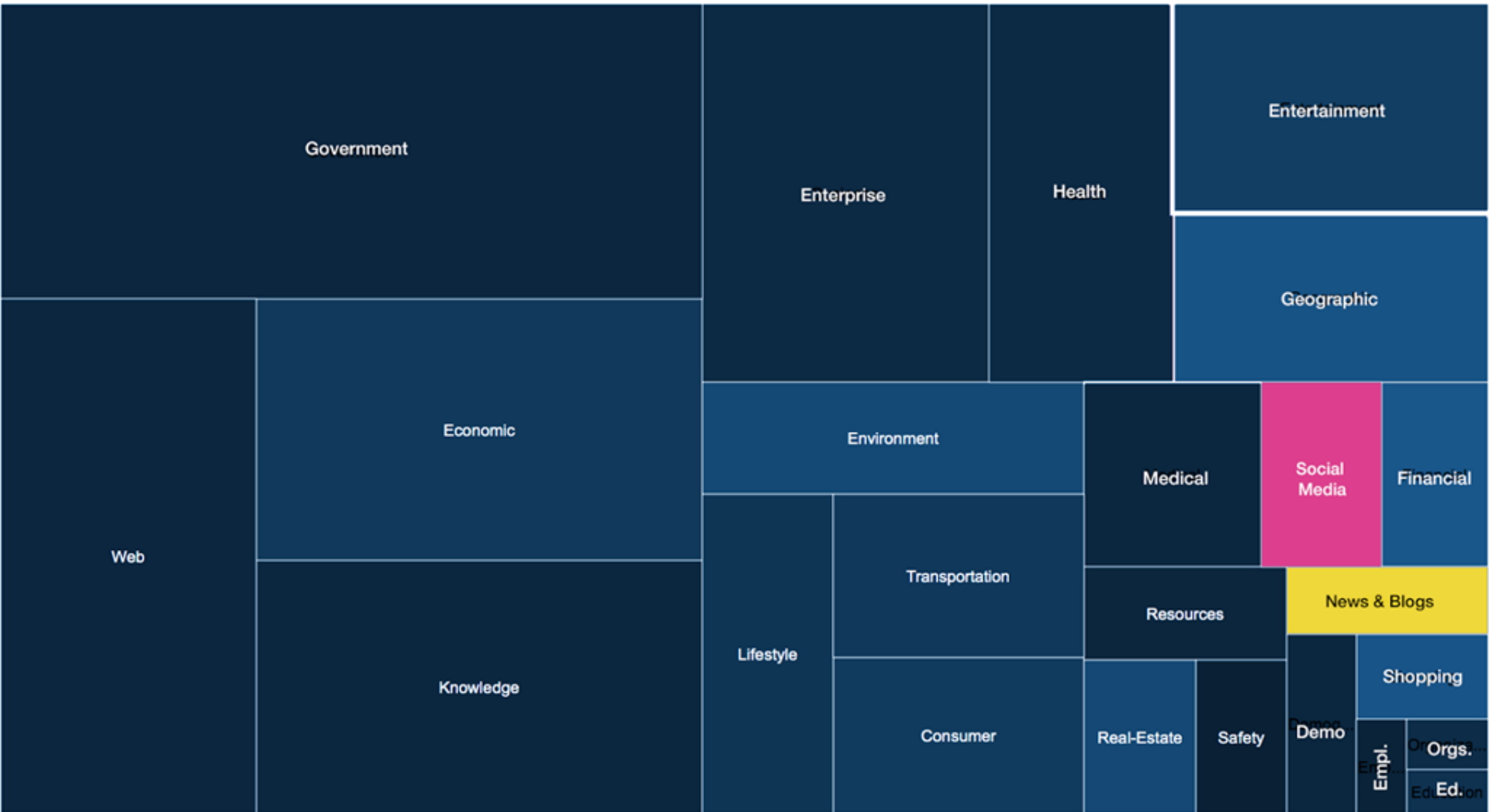
Object Graph is underlying structure for simplifying and processing complex queries



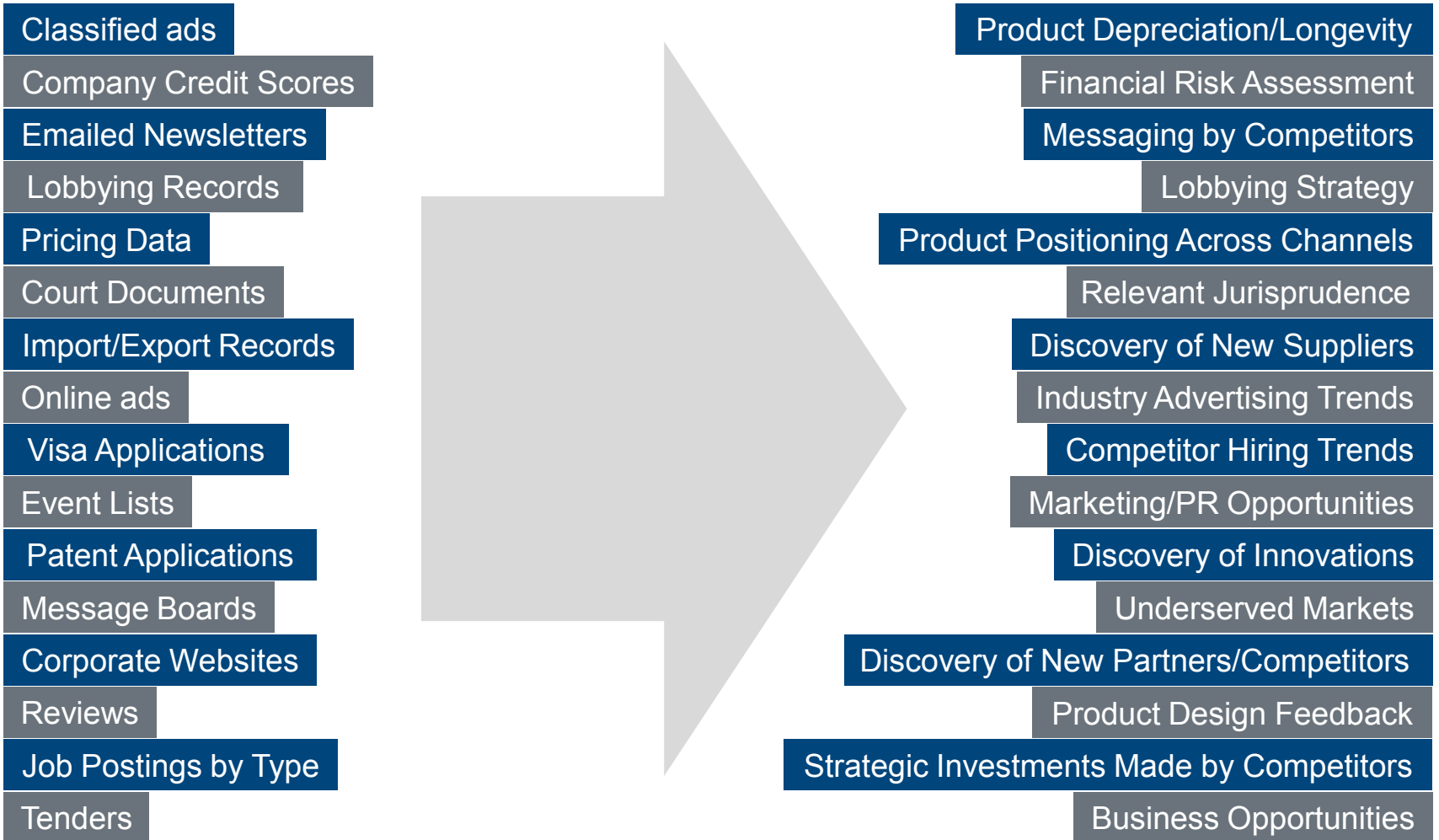
# Graph Query Examples

- Give me a document feed from all the influencers who compare ios and android
- Graph and compare the sentiment of support complaints between Apple and LG
- Give me a word cloud of the topics most used by both Apple and Samsung in the past 30 days
- Give me a word cloud of keywords the Apple CEO is using that other smartphone leaders are not
- Show me the top 3 most rapidly changing topics associated with tablets being talked about by influencers in the mobile market

# The Explosion is Driven by New Types of Data



# Because it carries valuable insights





# Example 1: Visa Applications

Wow, my competitor is aggressively hiring new engineers.  
Should we be investing more in our product?

H-1B VISA APPLICATIONS

## H-1B Visa - 2012

The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations. [Full Description](#)

- + See all nodes
  - UNITED STATES
    - FEDERAL GOVERNMENT OF THE UNITED STATES
      - DEPARTMENT OF LABOR
        - OFFICE OF FOREIGN LABOR CERTIFICATION
          - H-1B VISA APPLICATIONS
            - H-1B VISA - 2012

H-1B VISA - 2012 701 OF 168,717 ROWS EXPORT

Add filter... |

Status	LCA Case Job Title	LCA Case Employe...	LCA Case Employe...	LCA Case Employe...	LCA Case Wage Ra...	LCA Case W
CERTIFIED	SEARCH QUALITY ...	GOOGLE INC.	1600 AMPHITHEA...	CA	179500	MOUNTAIN
CERTIFIED	SOFTWARE ENGIN...	GOOGLE INC.	1600 AMPHITHEA...	CA	110000	NEW YORK
CERTIFIED	SOFTWARE ENGIN...	GOOGLE INC.	1600 AMPHITHEA...	CA	140000	KIRKLAND
CERTIFIED	SOFTWARE ENGIN...	GOOGLE INC.	1600 AMPHITHEA...	CA	105000	MOUNTAIN
CERTIFIED	SOFTWARE ENGIN...	GOOGLE INC.	1600 AMPHITHEA...	CA	105000	NEW YORK
CERTIFIED	ENTERPRISE SEAR...	GOOGLE INC.	1600 AMPHITHEA...	CA	131300	MOUNTAIN

1 - 100 of 701 rows Prev Next Chat with us

# Example 2: Online ads

Hmm, Nike is getting a great ROI on Go.com.  
Should we should start advertising there too?

**Advertiser Report**  
**Nike**  
Oct 01, 2013 ▶ Oct 29, 2013

**166** Publishers  
**4%** In-Play Impressions  
**32** Creatives  
**Tags & Pixels**  
ShareASale (53%), Google AdSense (34%), and 13 more

Filter Domain

All Domains (166) Activity

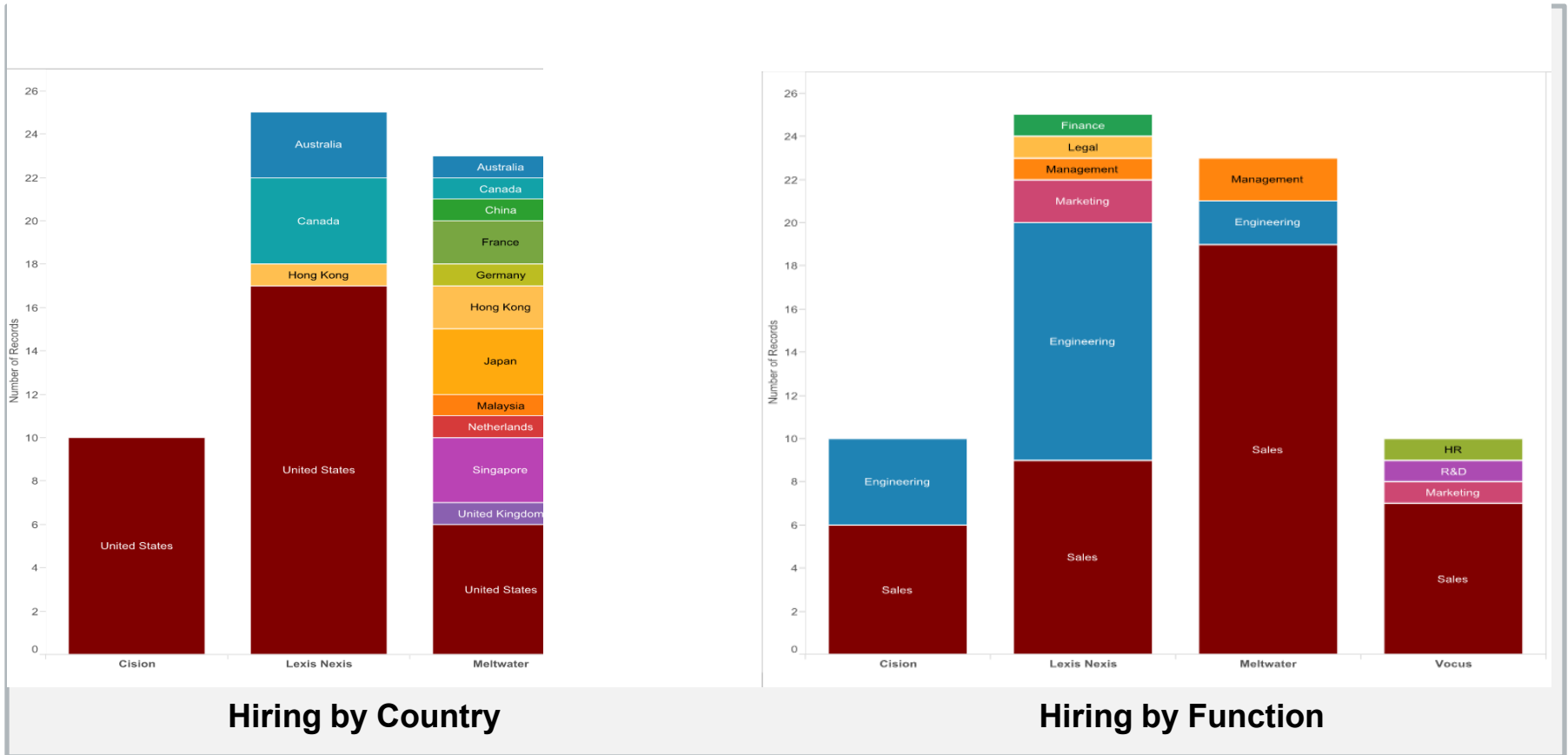
- ★ go.com
- youtube.com
- ★ maxpreps.com
- bleacherreport.com
- mlb.com
- washingtonpost.com
- rotoworld.com
- 49erswebzone.com

**AD CREATIVES:**

- Two identical creatives for Nike Air Max sneakers: "DON'T BLINK. CP3.VII EXPLORE NOW" with a Jordan brand logo.
- Steelers Official Online Store: "STEELERS OFFICIAL ONLINE STORE SHOP NOW IT HAS ARRIVED" featuring Steelers jerseys.
- Nike Sculpt Tight: "THE NEW NIKE SCULPT TIGHT" with a silhouette of a runner.
- Baseball-themed creatives: "42" and "EXIT SANDMAN" with a pitcher.
- A dark Nike logo creative.

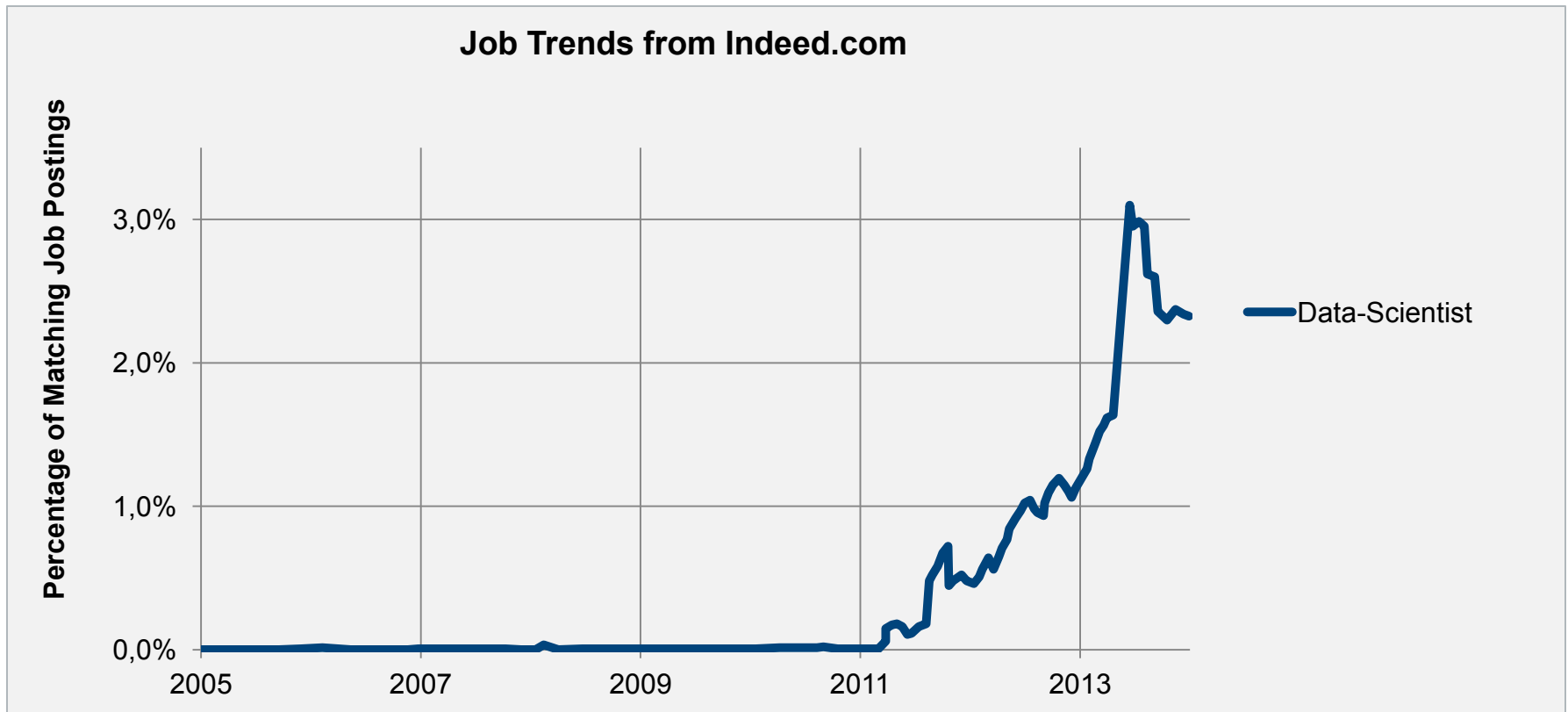
# Example 3: Jobs

Woah, My biggest competitor just started outsourcing  
Should we consider doing that?



# Example 3: Jobs (cont'd)

Kablam! I guess this Big Data trend is real...  
- I wish there was a solution to help us harvest those insights!





# Data Analytics, Monitoring and Beyond

- Processing close to real-time
  - Alerting – quick reactions
- Data segmented by
  - Geographic location
  - Demographics (gender, age, level of education, social status etc.)
- Correlations – even across the firewall
  - Integration to traditional BI systems

# Summary

- Information Explosion Creates a whole New Industry – Open Data Intelligence
- Classical BI Technologies are not applicable
- Natural Language Processing, Machine Learning, Big Data analytics help “connecting the dots”
- Companies of the 21<sup>st</sup> century cannot afford not using such information.

Thank you!

