

# **Big Data or Big Garbage?**

## **A Tale of a Research Journey for Real-Time Business Intelligence**

**Rakesh Agrawal**

President, Data Insights Laboratories  
Rukmini Chair Professor, Indian Institute of Science

December 15, 2016

Keynote, IEEE Int'l Conf. On Data Mining

# Fellow Travellers\*

Abhimanyu Das

Stelios Papparizos

John Shafer

\* Colleagues at the Microsoft Search Labs

**Terra Incognita**

# Perception of Social Data

Noisy!

Worthless!

Garbage!!!

More than 60% of the  
Twitter sample stream is  
useless garbage...



Further 20% are trolls...



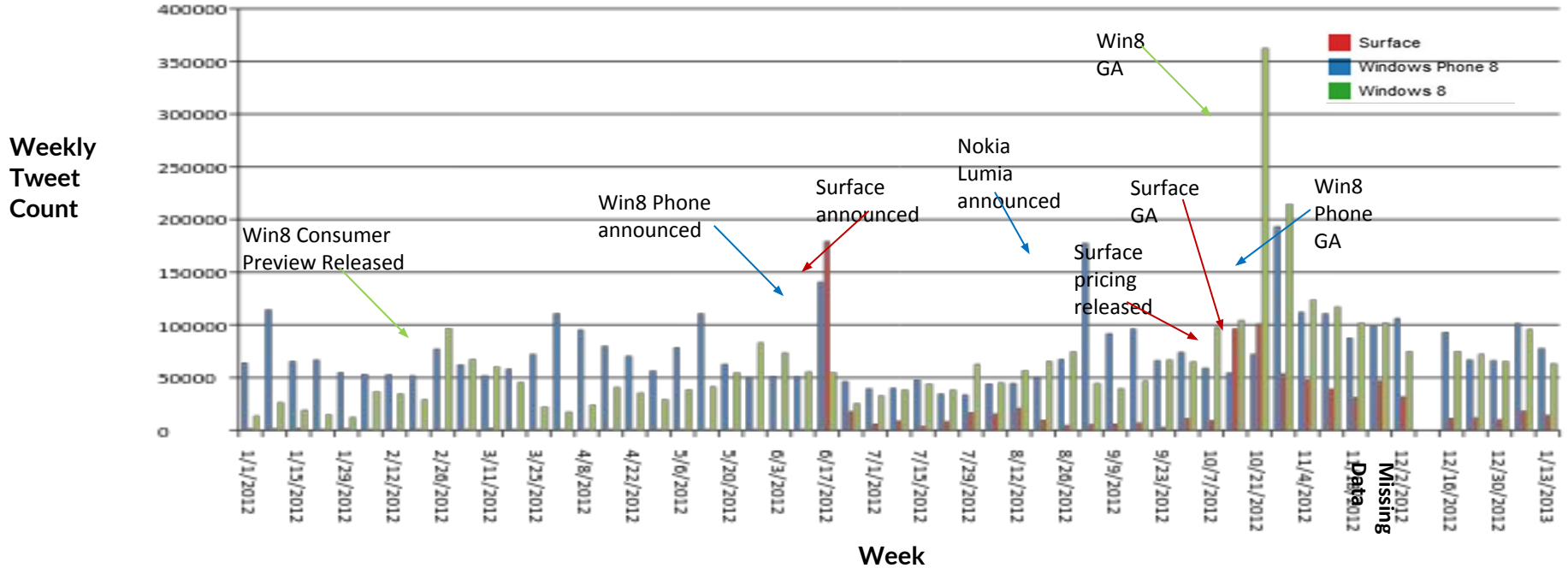
Pavlo Baron, codecentric AG

# Research Expedition

Can enterprises mine useful real-time business intelligence from social data, and in particular from Twitter stream?

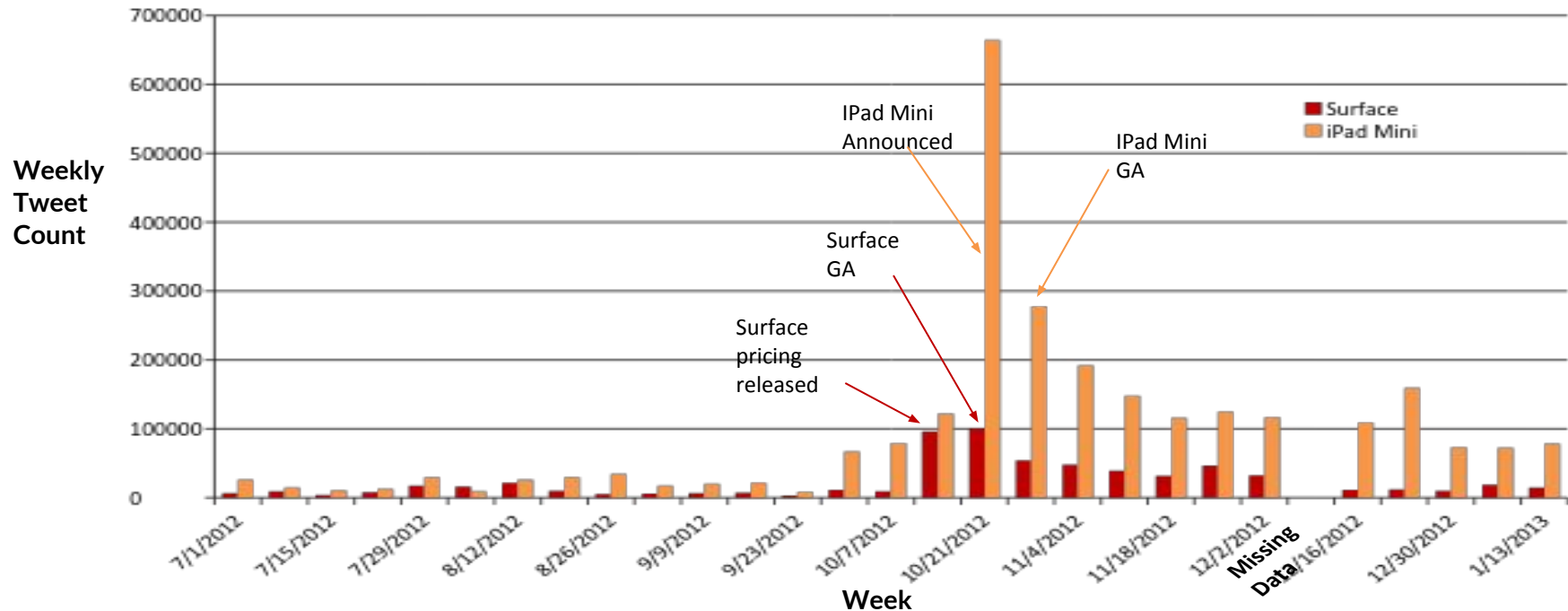
# Testing Waters

# Weekly Tweet Count for Surface, WP8, & Win8 (Jan 2012 - Jan 2013)



- TwitterSphere was reacting to the product related events!
- More sustained buzz around WP8!

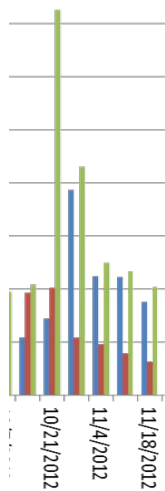
# Weekly Tweet Count: Surface vs. iPad Mini (July 2012 - Jan 2013)



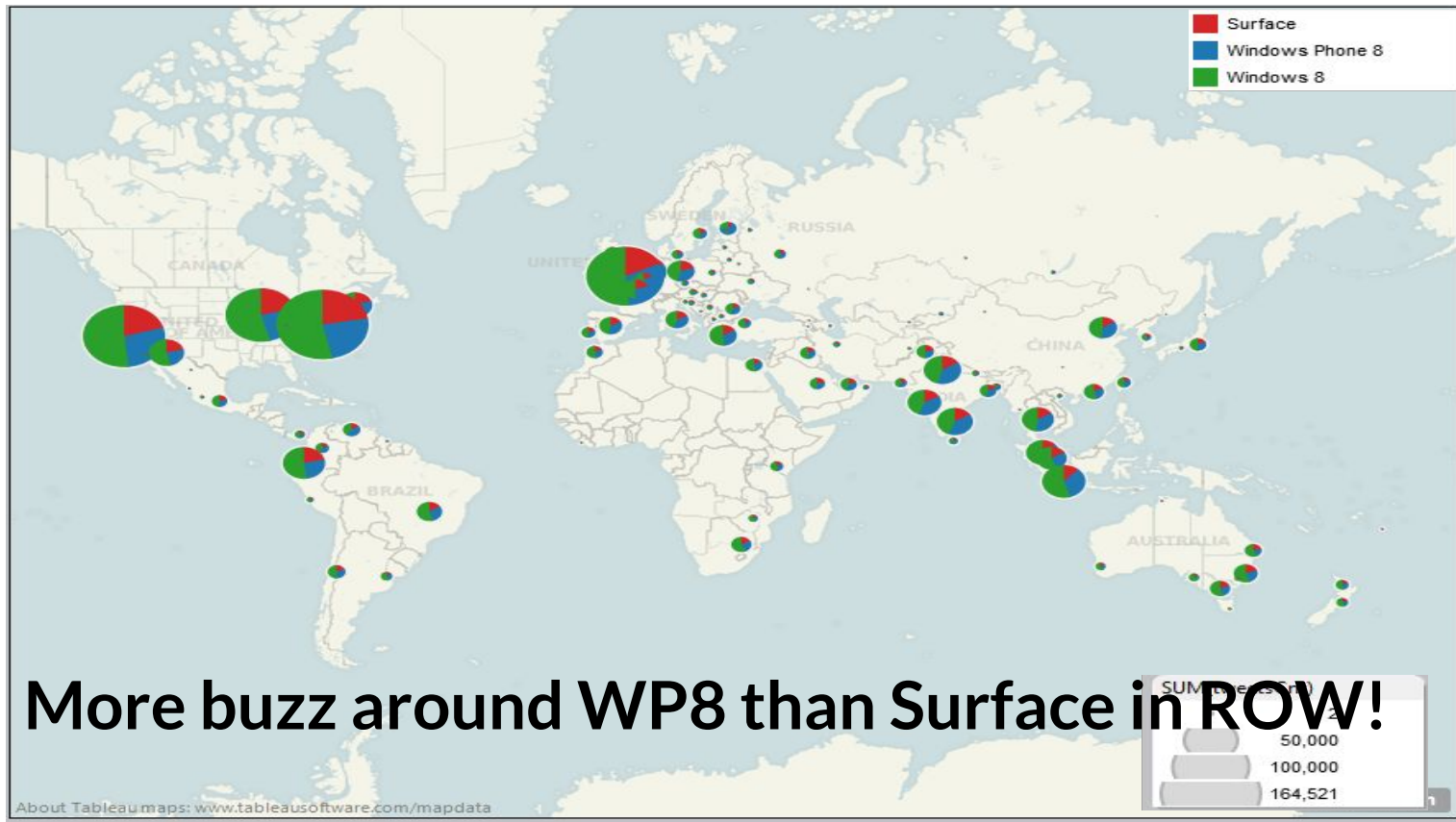
**More buzz around iPad Mini than Surface!**



# Tweet Counts Across the World: Surface, WP8, & Win8 (Oct 15, 2012 - Nov 18, 2012)



Total  
Tweet count



# Diverse Tweets for Surface (Oct 15, 2012 - Nov 18, 2012)

## Apps:

-Determined to like Microsoft Surface. But WinRT is feeling like beta. Don't understand why some apps couldn't be ported from WindowPhone.

-@vbandi Keyboard. Productivity with Office-type apps like Google Docs. Some people like that. Surface RT has it too w/ Office. iPad lacking.

-The native apps on the new Windows 8 surface tablet feel pretty good. News and travel are especially nicely done.

## TouchScreen :

-Why do I love Surface RT (and hopefully Pro) so much? It's simple: keyboard + touchscreen = crazy delicious.

-I'm not impressed w/ the Surface tablet, but i am impressed with Win8 on the new touchscreen laptops. The Acer s7 is awesome.

-OK, I'm really impressed with NBC's touchscreen system. Giant Microsoft Surface turned on its side?

## Price:

-Microsoft disappoints with Surface price and misses bigger opportunity <http://t.co/9CVaWKvq> via @siliconbeat

-@vegaobscura So an upper-tier product like the Surface Pro will definitely fall into the price category of a 4G LTE 64GB Retina iPad.

-Condition of Surface = Windows Phone.. great design ..good price but fail apps! Why the hell we should appreciate?

## Keyboard:

-RT @notch: Got to play with the Surface, and I quite like it so far. The keyboard cover thing is clever, and works great.

-Microsoft Surface users complain about Wi-Fi problems <http://t.co/DpLZDjk5> - and remember the touch keyboard problem, not good

-#buildwin Jordan Rudess is killing it on stage on the keyboard and earlier on a freaking surface tablet. Digital Theremin woo!

# Diverse Tweets for Surface (Oct 15, 2012 - Nov 18, 2012)

## Display:

-Sure. Worse is better. Sums up M\$. Microsoft: Surface screen better than iPad Retina display <http://t.co/ACGuXtbF>

-Finally got my hands on a Surface today at China Open Days. Now I really want one. Display looks great and I like the keyboard.

-In a dark room w/ a movie? iPad prob wins. But I'd like a display that works well in many conditions. Hopefully Surface has that.

## Camera:

-Playing with my new Surface tablet. Verdict: Good, needs some refinement. Definitely not an iPad. 1MP camera blows.

-HOLY SHAKEY CAMERA MICROSOFT. Geez plus its like watching a copycat Apple launch. #getyourownstyle Surface product looks great

## Battery:

-What I like about #XPS10 over Surface, a true keyboard dock and additional battery in dock for full day productivity

-the battery is like fine on ipad the surface is actually 40 minutes less than ipad supposedly..

-It has great thoughtful design, useful keyboard dock, great performance and battery life. The BEST tab along with Microsoft SURFACE

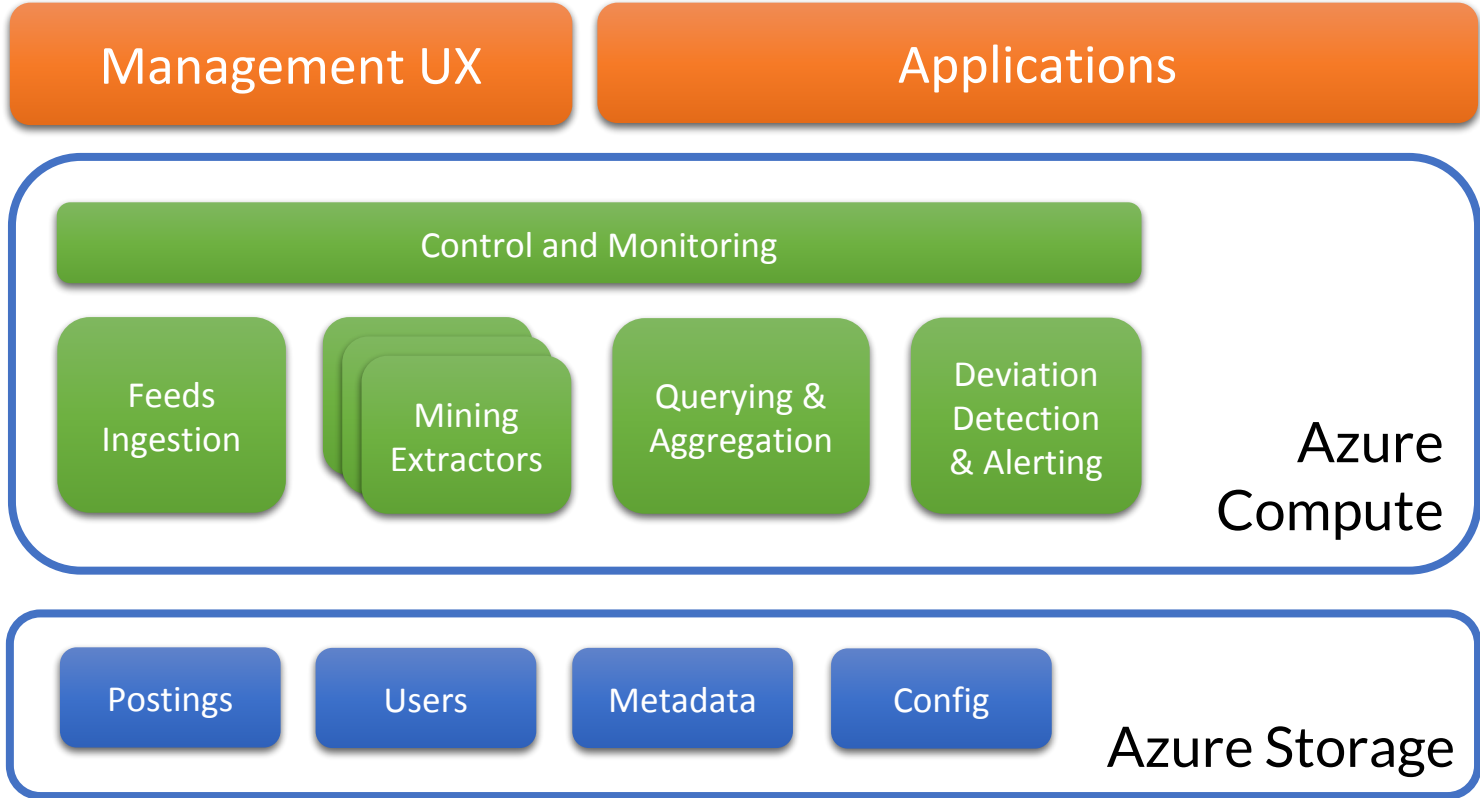
## Commercial:

- I'm not gonna lie that Microsoft surface commercial is pretty cool

- What an iPad and Microsoft Surface Parody Commercial Looks Like: GIZMODO Cheery piano musical... <http://t.co/yiYhrmCA>

# **The Ship: WaveFour**

# System Overview



# Information Flow

## Live Feed Ingestion

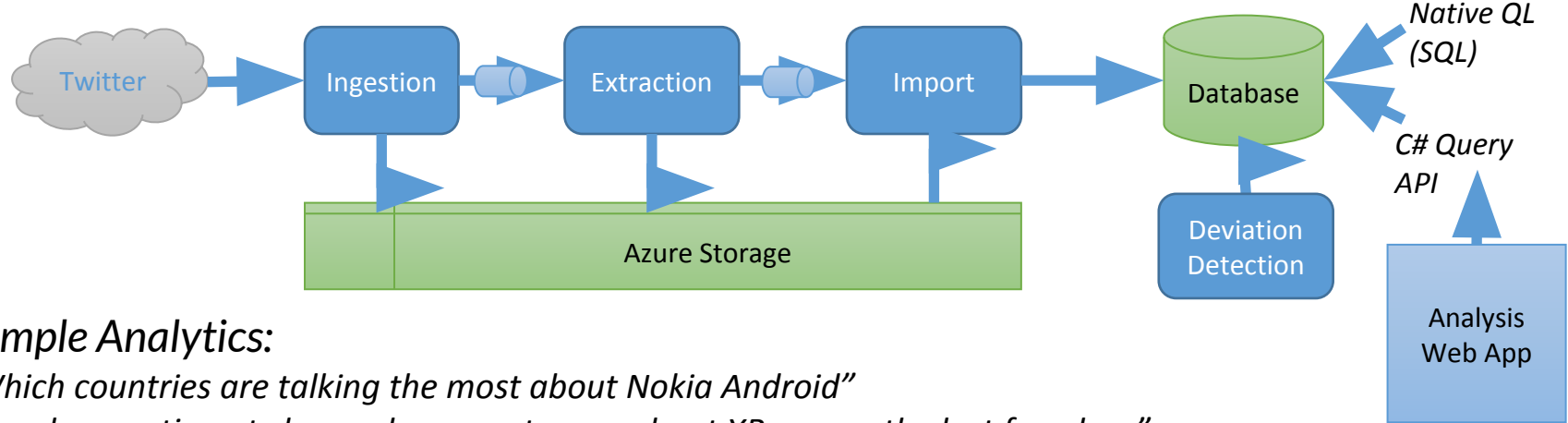
- Up-to-the minute
- Public 1% Twitter feed (4M tweets/day)
- Scalable to larger feeds

## Mining Extractors

- Products, Gender, Ethnicity,
- Language, Country, Timezone,
- Sentiment, Emoticon, Url, Hashtag
- **Add your own!**

## Queryable DB

- Indexed & queryable within minutes
- Ad-hoc querying via API or native QL
- Historical data going back several months



## Sample Analytics:

*“Which countries are talking the most about Nokia Android”*

*“How has sentiment changed amongst users about XBox over the last few days”*

*“Most talked-about Superbowl ad URLs amongst English vs. Spanish speakers”*

## Alerts based on deviation detection

# Latency Between a New Tweet and its Appearance in WaveFour

Only ~10mins delay

Time when user tweeted      Result creation time

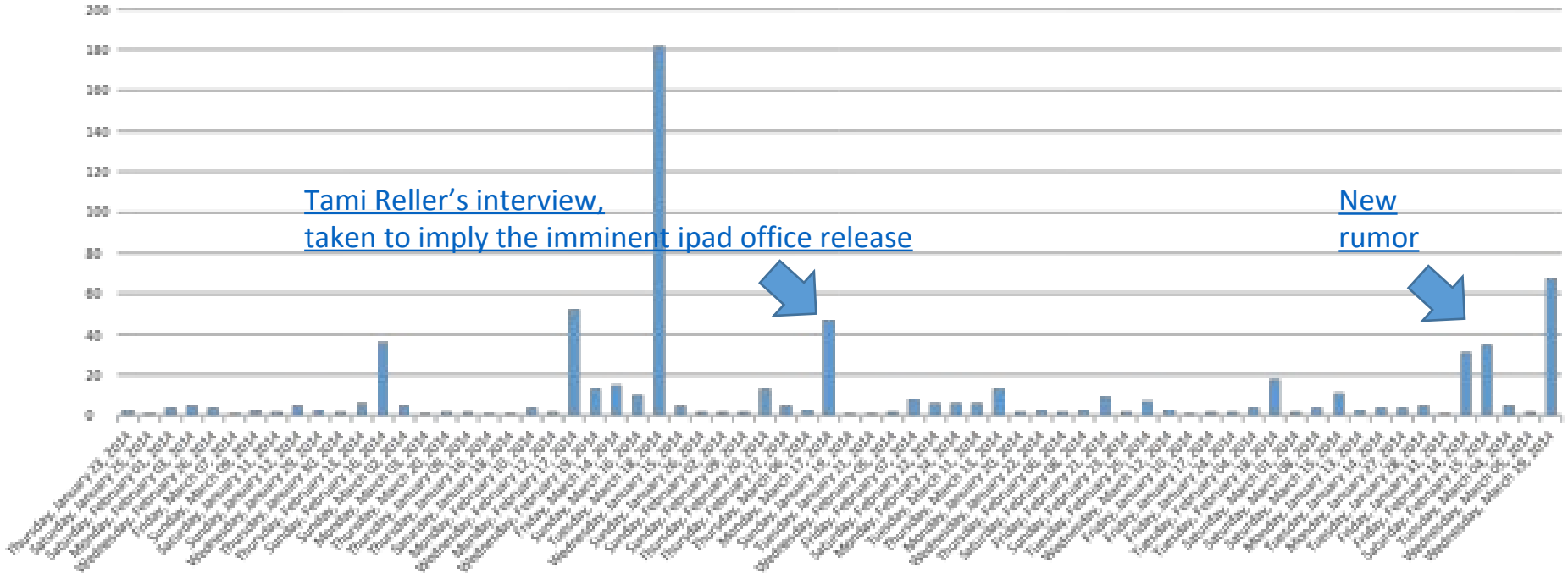
tweetid	userid	tweettext	CreationTimestamp	UTCTime	
1	446725860922785792	861999182	On my way to zachs ☺	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
2	446725860922777600	221126264	@MissAlexjones Just donated. My 3 yr old has cheered you on ...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
3	446725860918579200	2400292850	@null http://t.co/pfvwOC1z8V http://t.co/pfvwOC1z8V http:...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
4	446725860918169600	1010153330	RT @mylifeispichi: Y me acuerdo de aquel día en que me decía...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
5	446725860914393089	722782872	RT @56_: لإعلانك التجاري أو دعم حسابك الشخصي بأقل الأسعار...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
6	446725860914393088	1496479914	#32BinTibbiLaborantAtamaBekliyor @Rt_Erdogan @Muezzinog...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
7	446725860914384896	401003052	RT @evdekiadam: BAŞBAKANIM TWITTER'I KAPATCAKSAN...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
8	446725860914380800	173447985	The Meeting At 4 Will Be Held At The Creeporate Office.	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
9	446725860914368512	77544896	RT @eluniweb: La gente del centro comercial El Recreo sale a ...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
10	446725860913979392	53698412	Jurassic World What we know so far http://t.co/EQIc6TMDrb	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
11	446725860913975296	1720701656	それで我々も、心底安心できるというものです。閣下	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
12	446725860910198785	1156499065	J'ai récolté 1,070 unités de nourriture ! http://t.co/ewtLWAXGKf...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973
13	446725860910198784	417145081	#wittepleaseexpandthedamnrfollowlimitalready 153 #14million...	2014-03-20 19:11:59.0000000	2014-03-20 19:22:27.6400973

```
-- show latest tweets
SELECT TOP 1000 posting.tweetid, posting.userid, posting.tweettext
    ,CreationTimestamp.value as CreationTimestamp, SYSUTCDATETIME() as UTCTime
FROM [wavefourtweets].[dbo].[Posting]
join [wavefourtweets].[dbo].[CreationTimestamp] on CreationTimestamp.tweetid = Posting.tweetid and CreationTimestamp.userid = Posting.userid
order by CreationTimestamp.value desc
```

# Early Sightings



# Monitoring 'office for ipad'



Tami Reller's interview,  
taken to imply the imminent ipad office release

New  
rumor

```
-- daily volume of tweets about office on ipad
SELECT DATEADD(DAY, (DATEDIFF(DAY,'2014-01-06',CreationTimestamp.value)), '2014-01-06') AS timeby_begin,
        COUNT(distinct Posting.[tweetiduserid]) AS traffic
FROM [wavefourtweets].[dbo].[Posting]
    join [wavefourtweets].[dbo].[CreationTimestamp] on CreationTimestamp.tweetid = Posting.tweetid and CreationTimestamp.userid = Posting.userid
WHERE contains(tweettext, 'ipad AND office')
group by DATEDIFF(DAY,'2014-01-06',CreationTimestamp.value)
```

# Most Talked About Super Bowl Ads (New Yorker vs. Social)



**Coca-cola America the beautiful**



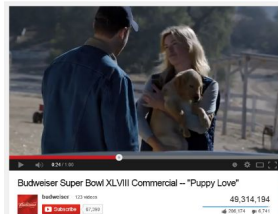
**Honda hugs with Bruce Willis**



**Turbotax big game**



**Hyundai dad's sixth sense**



**Budweiser "puppy love"**

## The Super Bowl...

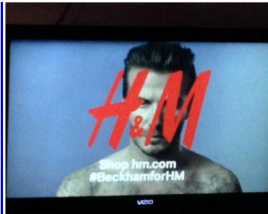
Out of 5,500+ entries from around the world, our homemade commercial, **THE COWBOY KID**, is now a top 5 Finalist in Doritos' Crash the Super Bowl contest! **BUT WE NEED YOUR HELP!!**



The commercial that gets the MOST VOTES during the month of



**Morpheus for Kia's car**



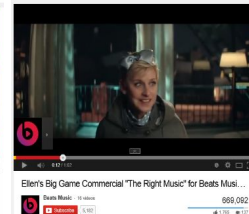
**David Beckham for H&M**



**Blitz Energy Drink (not aired)**



**Microsoft empowering users**



**Ellen's dance for beats headphones**

-- select most talked about urls about superbowl ads in english

```
SELECT url, volume
FROM(
  SELECT url.value as url, count(distinct posting.tweetiduserid) as volume
  FROM [wavefourtweets].[dbo].[Posting]
  join [wavefourtweets].[dbo].[Language] on language.tweetid = Posting.tweetid and Language.userid = Posting.userid
  join [wavefourtweets].[dbo].[Url] on Url.tweetid = Posting.tweetid and Url.userid = Posting.userid
  where contains(tweettext, '(superbowl OR "super bowl") AND (commercial OR ad OR ads OR advertisement)')
  AND Language.value = 'en'
  group by url.value
) as innerquery
ORDER BY volume desc
```


# Most Discussed Social Postings on MH370 (3/20/14)

 **julie demdam**  
@juliedemdam Follow

My cousin is flying over Asia and this is her view. I wonder if they're troops searching for #MalaysiaAirlines pic.twitter.com/OJcAfSL6fi

Reply Retweet Favorite More



 **MH370**  
@MH370flight Follow

The USA is sending the USS Pinckney to search the missing #MH370.  
#PrayForMH370 pic.twitter.com/j1GKQYyi7r

Reply Retweet Favorite More



 **Nizar Hussain**  
@nijzzar Follow

where are you? We're here still waiting for you #MH370 pic.twitter.com/CeSo4Kqz0Z

Reply Retweet Favorite More



None points to the top traditional news sources!

```
-- select most discussed urls about Malaysian Airlines MH370 in English
SELECT
  url
  ,diffDays
  ,volume
FROM
(
  SELECT
    url.value as url
    ,datediff(day, createdat, getdate()) as diffDays
    ,count(distinct posting.tweetiduserid) as volume
  FROM [wavefourtweets].[dbo].[Posting]
  join [wavefourtweets].[dbo].[Language] on Language.tweetid = Posting.tweetid and Language.userid = Posting.userid
  join [wavefourtweets].[dbo].[Url] on Url.tweetid = Posting.tweetid and Url.userid = Posting.userid
  where contains(tweettext, '(mh370 OR "mh 370" OR "malaysian airlines"')
  --AND Language.value = 'en'
  AND datediff(day, createdat, getdate()) <= 30
  group by url.value, datediff(day, createdat, getdate())
) as innerquery
ORDER BY volume desc
```

# **A Bounty**

**A Journey with Behzad Golshan & Evangelos Papalexakis**

# Overlap in Search Engine Results

Google search results for 'florence'. The search bar shows 'florence' and the search button. Below the search bar, there are tabs for 'Web', 'Images', 'Maps', 'Videos', 'News', and 'More'. The search results are displayed below, with a red box highlighting the first result: 'Florence - Wikipedia, the free encyclopedia'. Another red box highlights a second result: 'Florence, Italy: Tourist Travel Guide for Holidays in Florence ...'. A third red box highlights a result from TripAdvisor: 'TripAdvisor Florence - Best Travel & Tourism Info for ...'. Below the search results, there are sections for 'Images for florence', 'More images for florence', 'Florence Highlights - Lonely Planet', 'Points of interest', and 'Colleges and Universities'. A map of Florence is also visible.

Bing search results for 'florence'. The search bar shows 'florence' and the search button. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'More'. The search results are displayed below, with a red box highlighting the first result: 'Florence - Wikipedia, the free encyclopedia'. Another red box highlights a result from TripAdvisor: 'TripAdvisor Florence - Best Travel & Tourism Info for ...'. A third red box highlights a result from Visit Florence: 'Florence, Italy: Tourist Travel Guide for Holidays in ...'. Below the search results, there are sections for 'News about Florence', 'Eating in Florence', 'Johnson Controls employees in Florence withdraw from union', 'Florence and the Machine dazzle in Brooklyn', 'Map of New Florence PA | New Florence Pennsylvania ...', 'City of Florence', and 'Related searches for florence'. A map of Florence is also visible.

Citizens should have access to diverse perspectives as exposure to different views is beneficial for the advancement of humanity - The Fairness Doctrine, FCC 1949

See [AGP WWW2015, KDD 2015] for details

# Research Questions

- ❖ Have Google and Bing results become similar?
- ❖ Can search over Social Data (Twitter) provide different and useful results?

# Social Pulse

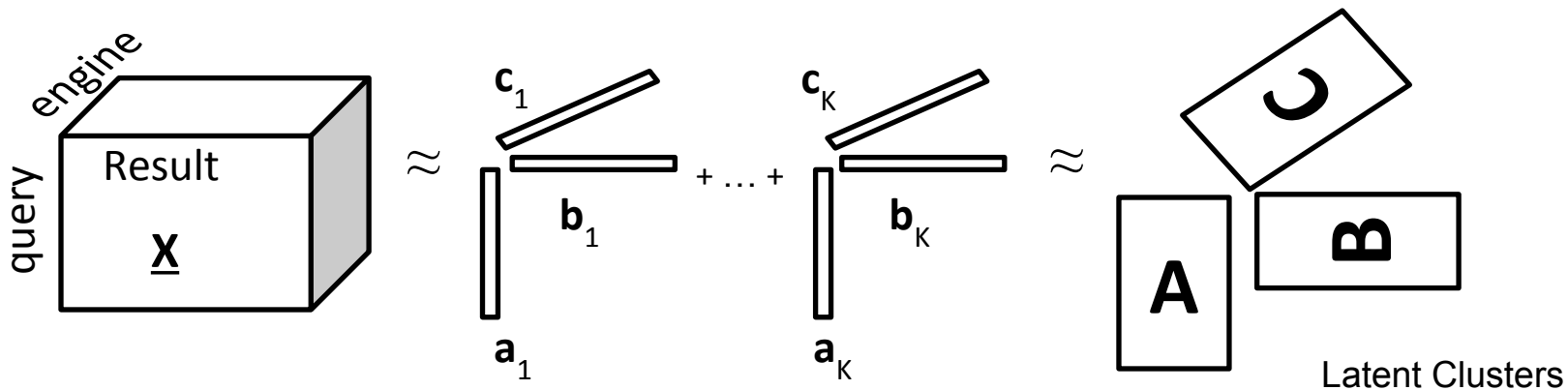
- Built on WaveFour Infrastructure
- Given query  $q$  find all tweets that contain  $q$  and a URL
- Rank them by number of retweets
- Implementation:
  - ✧ Using 1% sample from Twitter API
  - ✧ Restrict results to last 24 hours
  - ✧ Also include popular tweets without URLs

# Methodology

- Two sets of queries
  - ✧ **Trends:** “Head” queries
    - Google Trends queries for April 2014
  - ✧ **Manual:** “Trunk” queries
    - Handpicked queries we were familiar with
- Data Collection
  - ✧ Top-10 results every day around the same time during June-July 2014
- Result Representation
  - ✧ **Google & Bing:** Union of bag-of-words for all snippets of top-10 results
  - ✧ **SocialPulse:** Union of bag-of-words for top-10 tweets



# Tensor Decomposition (Visualization)

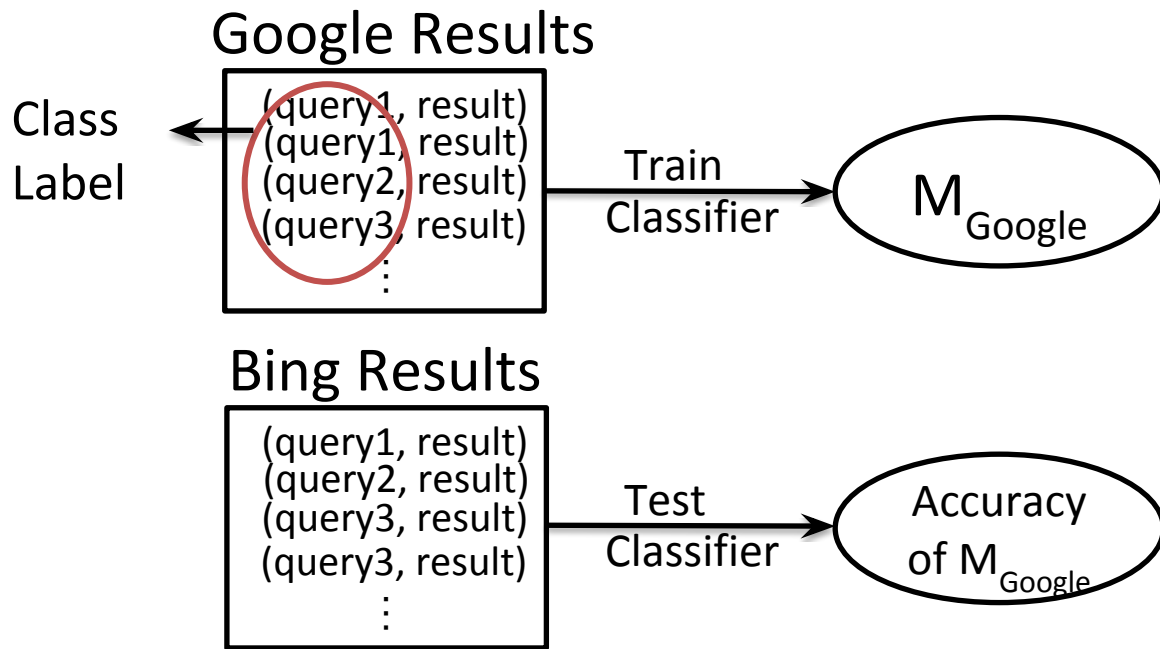


- Do PARAFAC decomposition of (query, result, engine) tensor
- The  $(i,j)$  entry of  $C$  gives the participation of search engine  $i$  in cluster  $j$

	Latent Clusters		
Engine1	.5	0	.9
	<b>C</b>		
Engine2	.5	1	.1

- Plot participation values of the search engines for different clusters
- Values cluster around  $(0.5, 0.5) \Rightarrow$  E1 and E2 similar
- Values cluster around  $(0, 1)$  and  $(1, 0) \Rightarrow$  E1 and E2 dissimilar

# CrossLearnCompare (Quantification)

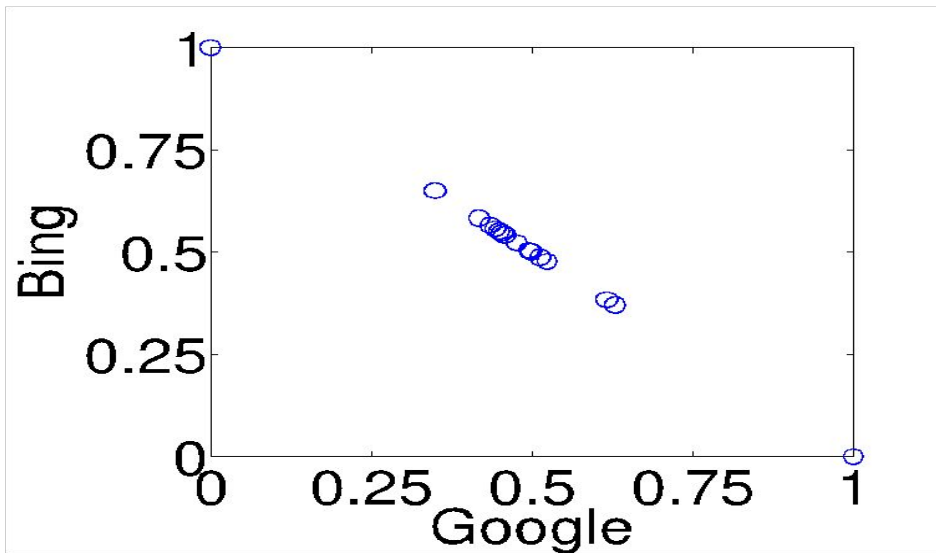


## Key Idea:

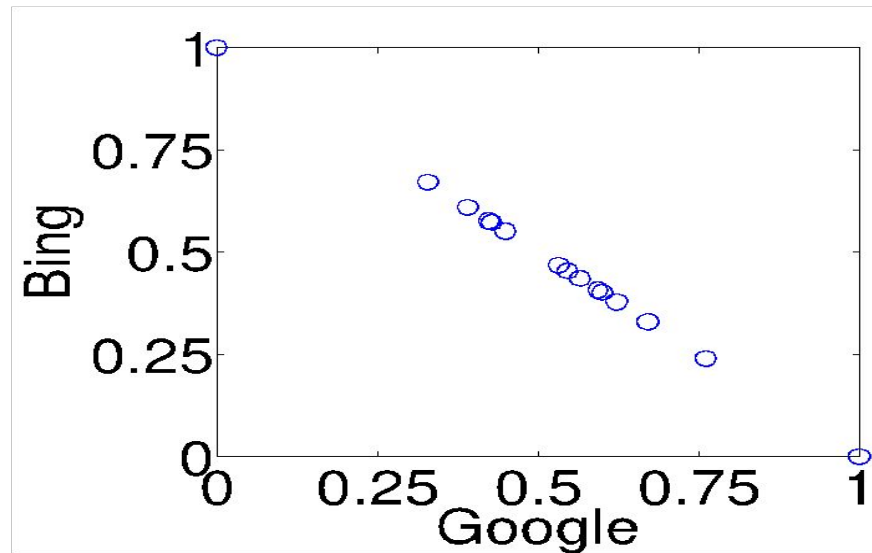
If  $M_A$  predicts results of Engine B  
→ A, B are similar

- Each query is a class
- Feature representation of query to train a multi-class model
- One-vs-all linear SVM classifier
- Measure classification accuracy as measure of similarity/overlap

# TensorCompare: Google vs Bing



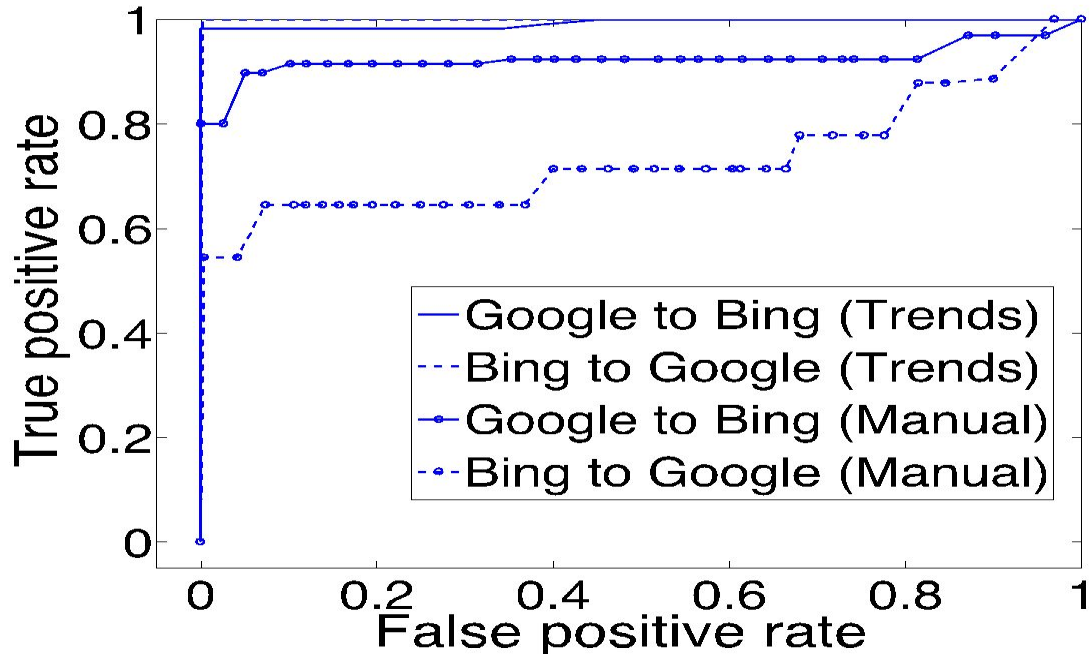
(a) TRENDS query set



(b) MANUAL query set

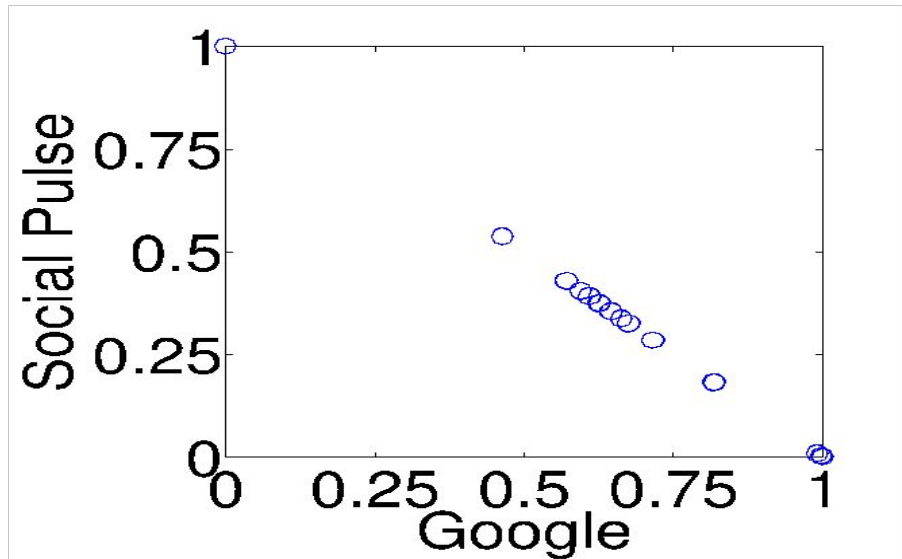
**High overlap! (esp. for Trends)**

# CrossLearnCompare: Google vs Bing

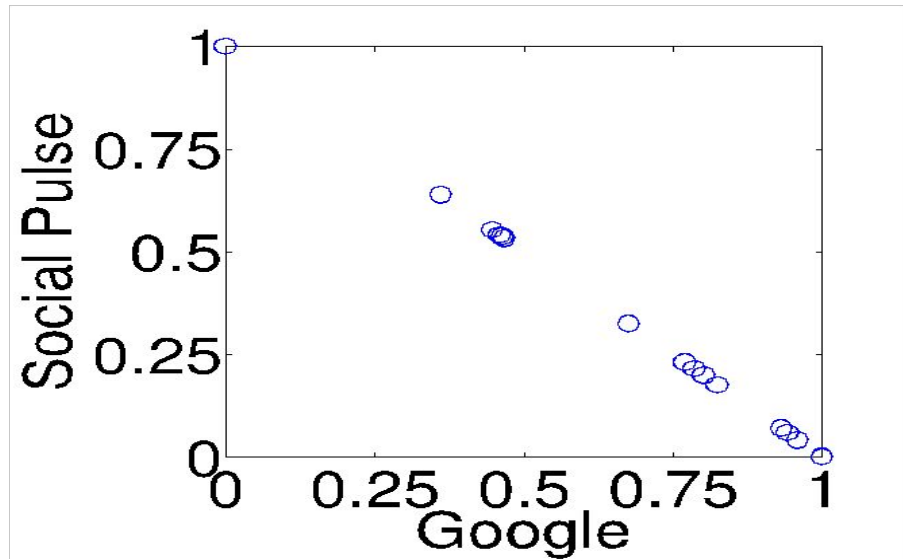


**High overlap! (esp. for Trends)**

# TensorCompare: Google vs SocialPulse



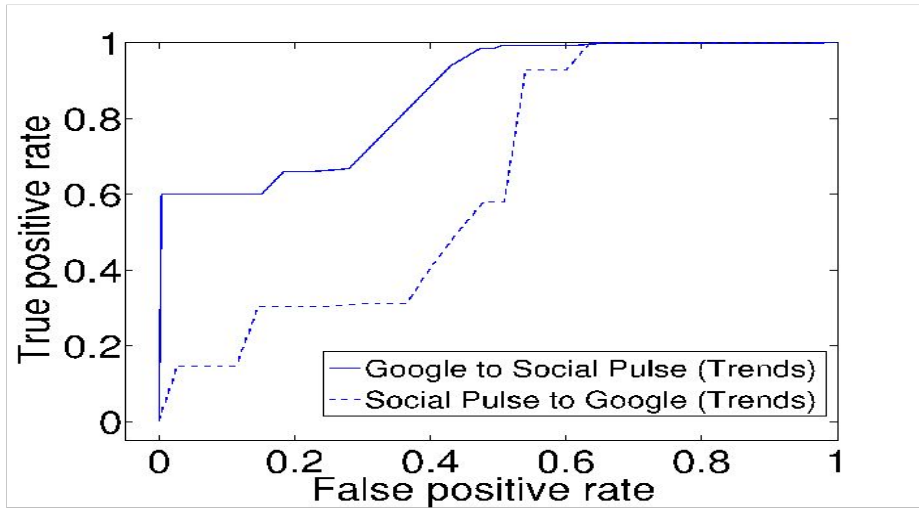
(a) TENSORCOMPARE for TRENDS



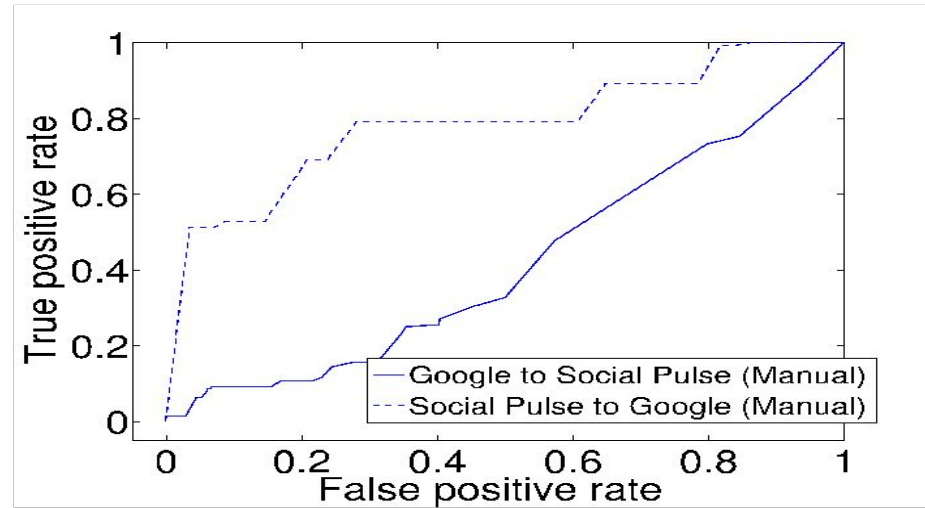
(b) TENSORCOMPARE for MANUAL

**Low Overlap!**

# CrossLearnCompare: Google vs Social Pulse



(c) CROSSLEARNCOMPARE for TRENDS



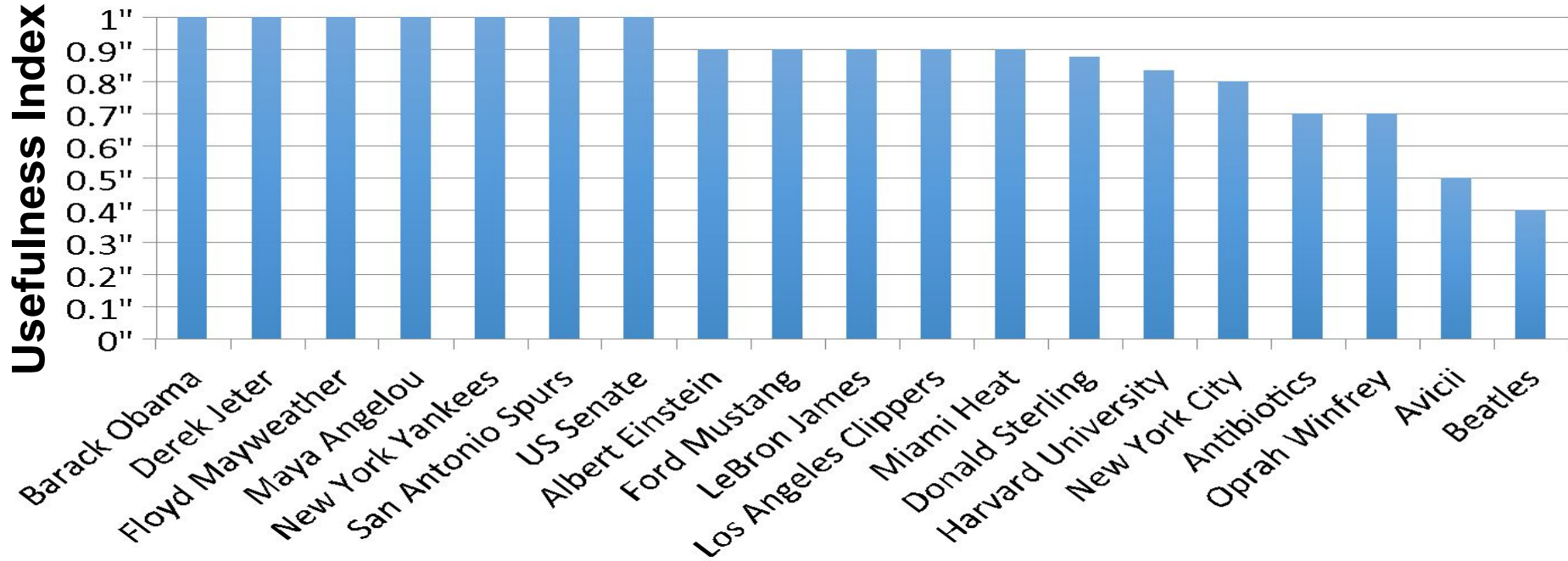
(d) CROSSLEARNCOMPARE for MANUAL

**Low Overlap!**

# User Study

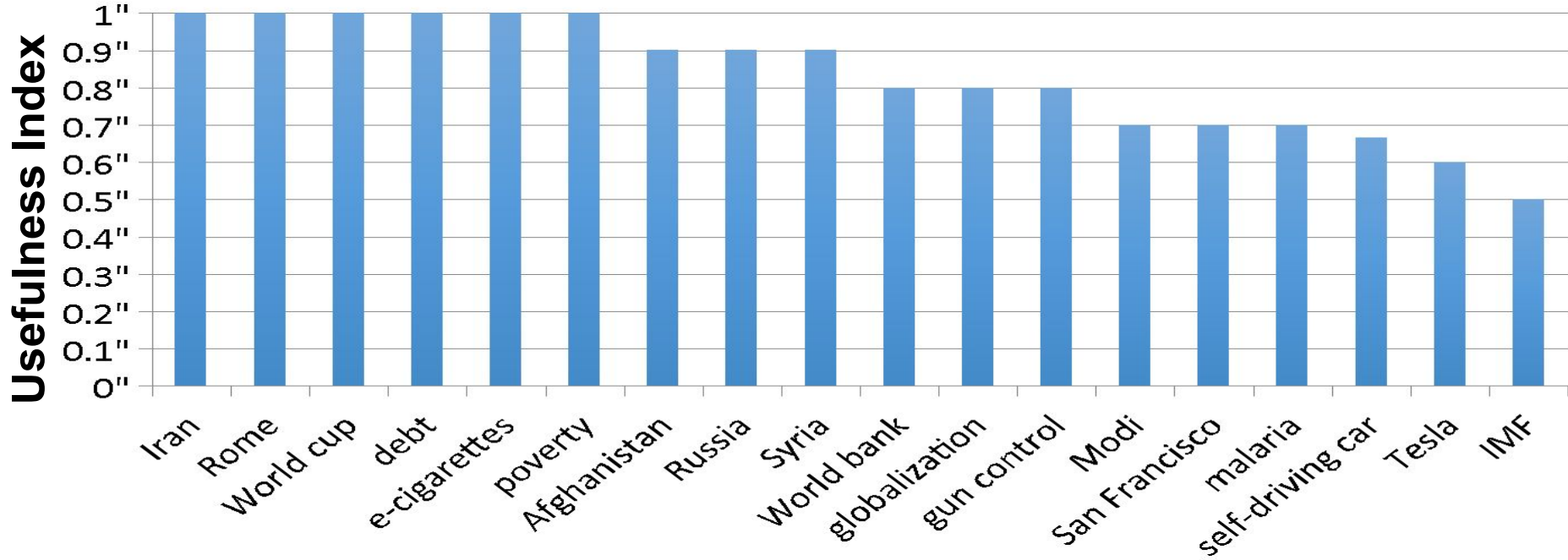
- So far, Twitter results are different.
- Are they useful??
- Amazon Mechanical Turk user study
  - ✧ Take results for a single day
  - ✧ Ask Turkers to judge how informative the results are

# Usefulness of Results (Trends Query Set)





# Usefulness of Results (Manual Query Set)



# **The Rough Seas**

# The Challenge of Growing Up

Metrics/Benchmarks for the quality of miners

Scalability of the data management component (AsterixDB journey\*)

Licensing cost of full Twitter firehose (Catch-22)

Careful performance analysis vs. Hail Mary deployment

\* Thanks Carey, Raman, Vinayak

# Hiccup

The team dispersed after the closing of MSR-SV in September 2014

# Looking Over the Horizon

# Assertions

WaveFour was a worthwhile research expedition

Big social data is not garbage

Real-time business intelligence using social data is feasible and valuable

At the Data Insights Laboratories, I am taking another stab at the problem, but from a different perspective!

**Thank You!**

Questions?

Patent Pending

**myInfoDVR**

Discover & share what you like  
instantaneously and effortlessly  
wherever and whenever



## New Design Point

- ❖ Thin server, Powerful clients
- ❖ Anticipatory query processing
- ❖ On-demand minimal indexing
- ❖ Privacy
- ❖ Performance

## Highlights

- ❖ Latest information on topics of interest whenever user wants
- ❖ Topics are learned and change automatically based on user's engagement
- ❖ Seamless sharing of any information

## A Family of Cloud Services

- ❖ [Informed.myinfodvr.com](https://Informed.myinfodvr.com)
- ❖ [Social.myinfodvr.com](https://Social.myinfodvr.com)

# Find without Searching

Current channel (topic)

Next channel (topic)

Touch for more on headline

Signal your low interest

Swipe right for the previous item in the channel

Swipe left for the next item in the channel

Touch for more on source

Share what you see

Source and Time

Provide feedback



# Saving myInfoDVR launcher on Home Screen

## Android

Launch Chrome

Open <http://xxx.myInfoDVR.com>

Tap the Menu button (Three vertical dots)

Now tap Add to Homescreen in the menu that appears

XXX = {Informed, Happenings, Social, India, Mehfil}

## iPhone (iPad is similar)

Launch Safari

Open <http://xxx.myInfoDVR.com>

Tap the Share button:



Now tap Add to Homescreen in the menu that appears:



## Windows Phone

Launch Internet Explorer

Open <http://xxx.myInfoDVR.com>

Tap the More button

Now tap Pin to Start in the menu that appears

## **Data Insights Laboratories**

Mission:  
Create technologies to  
accelerate emergent  
datafication of human  
endeavours

**Thank You!**

Rakesh Agrawal

[President@datainsightlabs.com](mailto:President@datainsightlabs.com)