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 Paparizos
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
• TwitterSphere was reacting to the product related events!
• More sustained buzz around WP8!
More buzz around iPad Mini than Surface!

More buzz around WP8 than Surface in ROW!
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?

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?

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.

**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

**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

**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
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

<table>
<thead>
<tr>
<th>Tweetid</th>
<th>Userid</th>
<th>Tweettext</th>
<th>Creation Timestamp</th>
<th>UTC Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4467258603522777600</td>
<td>221126254</td>
<td>@MissAlexones Just donated. My 3 yr old has cheered you on ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>446725860352185720000</td>
<td>2403022030</td>
<td>@null <a href="http://t.co/pfwrOCt1sV">http://t.co/pfwrOCt1sV</a> <a href="http://t.co/pfwrOCt1sV">http://t.co/pfwrOCt1sV</a> <a href="http://t.co/pfwrOCt1sV">http://t.co/pfwrOCt1sV</a> ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>446725860351816396000</td>
<td>101013330</td>
<td>RT @myrle: @Yme acuero de aqui dia en que me declaro ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>4467258603514330809</td>
<td>7227827827</td>
<td>RT @Erdogan @Missalexone ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>4467258603514330808</td>
<td>1494379914</td>
<td>#32BiTubb:labrant/AtanaBakkyor @R. Erdogan @Missalexone ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>4467258603514330808</td>
<td>401003002</td>
<td>RT @evskokan: BAŞBAKANIM TWITTER'YI KAPATCAKSAN ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>4467258603514305052</td>
<td>77544896</td>
<td>RT @elurinweb: La gente del centro comercial El Recreo sale a ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
<tr>
<td>4467258603513970692</td>
<td>5396412</td>
<td>Jurassic World What we know so far <a href="http://t.co/EGJcETMDr">http://t.co/EGJcETMDr</a> ...</td>
<td>2014-03-20 19:11:58.000000000</td>
<td>2014-03-20 19:22:27.6400973</td>
</tr>
</tbody>
</table>

---

**Show latest tweets**

```sql
SELECT TOP 100 posting.tweetid, posting.userid, posting.tweettext, CreationTimestamp.value as CreationTimestamp, SYSDATETIME() 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

```sql
<table>
<thead>
<tr>
<th>daily volume of tweets about office on ipad</th>
</tr>
</thead>
</table>
| SELECT DATEADD(DAY, (DATEDIFF(DAY, '2014-01-06', CreationTimestamp.value)), '2014-01-06') AS timely_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
- Budweiser
  - "puppy love"
- Honda hugs with Bruce Willis
- Morpheus for Kia's car
- David Beckham for H&M
- Blitz Energy Drink (not aired)
- Microsoft empowering users
- Ellen's dance for beats headphones
- Hyundai dad's sixth sense

```sql
-- select most talked about urls about superbowl ads in english
SELECT url.value as url, count(distinct posting.tweetiduserid) as volume
FROM (SELECT url.value as url, COUNT(DISTINCT posting.tweetiduserid) as volume
      FROM wavefourtweets
      JOIN [dbo].[Posting] ON wavefourtweets.tweetid = Posting.tweetid AND Language.userid = Posting.userid
      JOIN [dbo].[Wavefourtweets] ON wavefourtweets.tweetid = Posting.tweetid AND useruserid = Posting.userid
      WHERE contains(tweettext, '(superbowl OR "super bowl") AND (commercial OR ad OR ads OR advertisement)')
      AND Language.value = 'en'
    ) AS innerquery
ORDER BY volume DESC
```
My cousin is flying over Asia and this is her view. I wonder if they're troops searching for #MalaysiaAirlines pic.twitter.com/OJcAfSL6fi

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

None points to the top traditional news sources!
A Bounty

A Journey with Behzad Golshan & Evangelos Papalexakis
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

See [AGP KDD2015] for details
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 \(j\)
- Plot participation values of the search engines for different clusters
- Values cluster around \((0.5, 0.5)\) => E1 and E2 similar
- Values cluster around \((0, 1)\) and \((1,0)\) => 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

High overlap! (esp. for Trends)
CrossLearnCompare: Google vs Bing

High overlap! (esp. for Trends)
TensorCompare: Google vs SocialPulse

Low Overlap!

(a) TensorCompare for Trends

(b) TensorCompare for Manual
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?
myInfoDVR

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

Patent Pending
New Design Point

- Thin server, Powerful clients
- Anticipatory query processing
- On-demand minimal indexing
- Privacy
- Performance
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
- Social.myinfodvr.com
## Saving myInfoDVR launcher on Home Screen

<table>
<thead>
<tr>
<th>Platform</th>
<th>Instructions</th>
</tr>
</thead>
</table>
| **Android**         | Launch Chrome  
|                     | Open [http://xxx.myInfoDVR.com](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](http://xxx.myInfoDVR.com)  
|                     | Tap the Share button:  
|                     | ![Share button](image)  
|                     | Now tap Add to Homescreen in the menu that appears:  
| **Windows Phone**   | Launch Internet Explorer  
|                     | Open [http://xxx.myInfoDVR.com](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@datainsightslabs.com