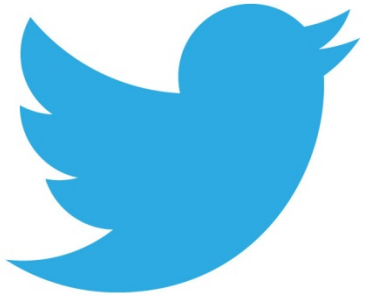
Microblog Data Stream-based Social Phenomenon Prediction: A Case Study Using The NIST Twitter Dataset

Dr. Allison Jones-Farmer
Jeremy D. Ezell
Dr. Casey Cegielski
D. Scott Cycmanick
Dept. of Aviation & Supply Chain Management

Overview



- Theoretical Background
- Twitter Facts
- Previous Research
- NIST Data Source
- Scripts and Downloading
- Potential Future Analysis
- Questions?

Theoretical Background

- Analytics broadly defined as:
  - Any data-driven process that provides insights
  - Insights can allow the firm to become smarter and nimble in competitive environment.

  (Stubbs, 2011)

- Firm and User-generated web data
  - Easily accessible (well. . . maybe)
  - Large Volume
  - Direct access to individual opinions, feelings, preferences

Theoretical Background

- Microblog Data (Twitter)
  - Very Personal
  - Direct expression of thoughts, emotions, desires

  (Naaman, Boase, & Lai, 2010)

- Example uses:
  - (Marketing) Public reaction to new products
  - (Medical) Tracking of disease outbreaks
  - (Social) Public Opinion, Political Discourse, Policy Reaction

- So far. . . Descriptive uses
  - The Goal: Predictive abilities using Microblog Data
  - *Can we roughly predict reactions and broad social trends?*

Why Technology in Decision making is Critical. . . (IBM / Neil Isford – VP Analytics)

Volume

12 terabytes

of Tweets created daily

Analyze product sentiment

Velocity

5 million

trade events per second

Identify potential fraud

Variety

100's video feeds

from surveillance cameras

Monitor events of interest

350 billion

meter readings per annum

Predict power consumption

500 million

call detail records per day

Prevent customer churn

80% data growth

are images, video, documents…

Improve customer satisfaction

The Social (Network) Layer in an Instrumented Interconnected World

12+ TBs **of tweet data every day**

? TBs of data every day

25+ TBs **of log data every day**

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

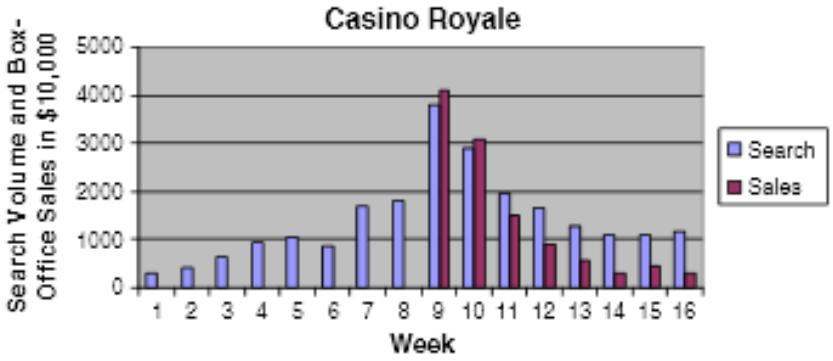
100s of millions of GPS enabled devices sold annually

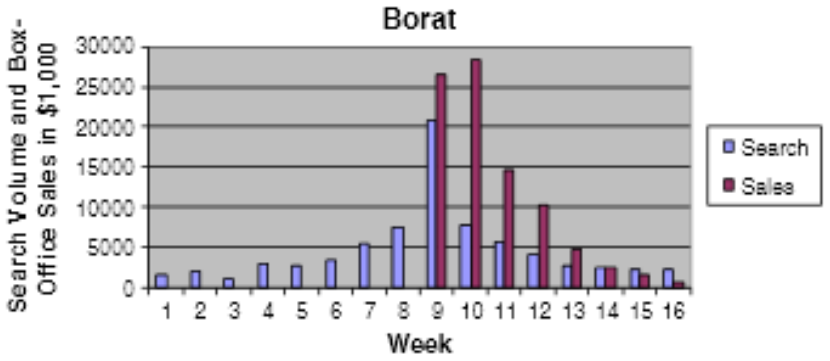
2+ billion people on the Web by end 2011

76 million smart meters in 2009… 200M by 2014

Previous Research

- Kulkarni, Kaanan, & Moe, 2011
  - Search Engine activity data as predictor of Movie Ticket Sales
  - Microblog-like data
  - Used a Search Term Research Service to pull activity from Google, Yahoo!, MSN (4.3 Billion Searches)
  - Main Focus: *Prediction of opening weekend sales using search activity and other variables*.





Previous Research

- Kulkarni, Kaanan, & Moe, 2011
  - Search Volume Model: $\ln(\text{Search Volume}_i) - \alpha_i + \beta X_i + \varepsilon_i$
  - Search Pattern Model $\text{Hazard}: h_i(t) = \frac{f(t)}{1-F(t)} = \lambda_i c_i t^{c_i} \exp\{\gamma Z_i\}$
  - Sales Model: $\ln(\text{Sales}_i) - a_i + bX_i + u_i \quad \text{where} u_i \neg \text{Normal}(0,\ s).$
  - Model Parameters are allowed to jointly correlate in a mean vector and covariance matrix (Hyperparameters):

$$\begin{bmatrix} \alpha_i \\ a_i \\ \ln(\lambda_i) \\ \ln(c_i) \end{bmatrix} - MVN(\omega, \sum)$$

Previous Research

- Kulkarni, Kaanan, & Moe, 2011
- Results:

Table 3
Forecasting.

APE	8 weeks calibration	4 weeks calibration	No pre-launch search
No search covariates: ln(sales)	8.49	7.19	
No adver effect: ln(sales)	9.40	7.28	
With adver effect: ln(sales)	9.03	7.66	
With only movie covariates: ln(sales)			11.97

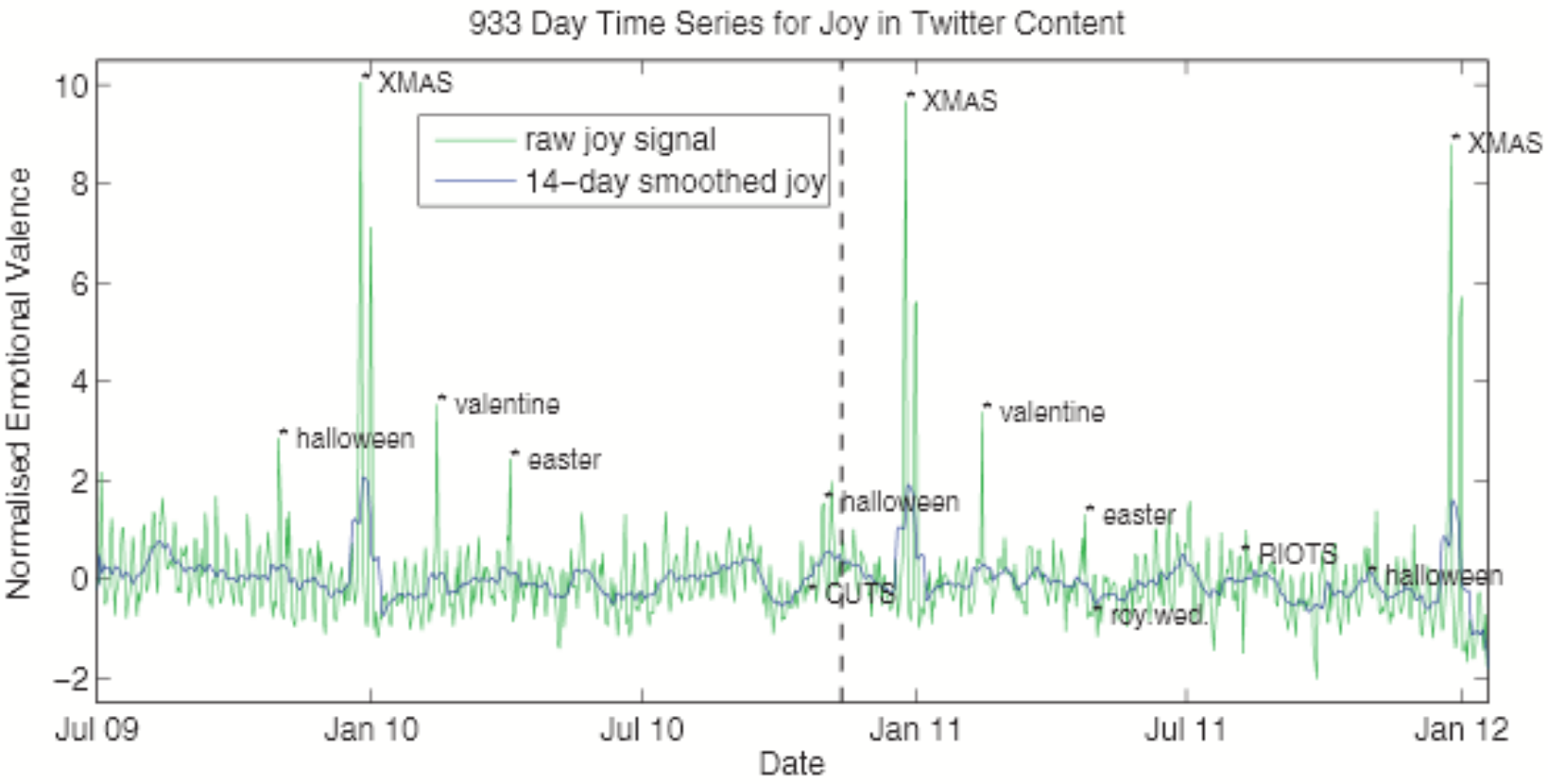
	No week-to-week adv. effects		
	Median	2.50%	97.50%
ln(sales)			
Baseline sales	13.51	12.26	14.76
Comedy	−0.5629	−1.4	0.2634
Drama	**−1.085**	**−1.847**	**−0.2239**
PG	0.848	−0.2457	1.914
PG-13	**0.9953**	**0.2375**	**1.789**
Competition	−0.2822	−0.6042	0.01656
ln(budget)	**0.1626**	**0.07596**	**0.252**
ln(search volume)			
Baseline sales	5.085	4.218	6.553
Comedy	**−0.7093**	**−1.369**	**−0.1218**
Drama	−0.2192	−0.8258	0.3463
PG	−0.4563	−1.202	0.3606
PG-13	−0.2265	−0.7806	0.3186
Competition	**−0.2596**	**−0.4839**	**−0.05577**
ln(budget)	**0.2228**	**0.1297**	**0.2864**
Search pattern			
ln(lambda)	−16.14	−16.65	−15.74
ln(c)	0.912	0.7692	1.059
ad effect			

Previous Research

- Lansdall-Welfare, Lampos, & Cristianini, 2012
  - Gathered Tweets from 54 Largest UK Cities (Geospatial Isolating)
  - 30 Months
  - 484 Million Tweets (Sampled in Real-time, 3-5 minutes)
  - Used Citation-Sentiment analysis on Tweets
    - Keywords for emotions [WordNet-Affect – Princeton]:
      - Anger (146 Words)
      - Fear (92 Words)
      - Joy (224 Words)
      - Sadness (115 Words)
  - "…***studies of this kind rely on very efficient methods of data management and text mining***…" (pp. 27)

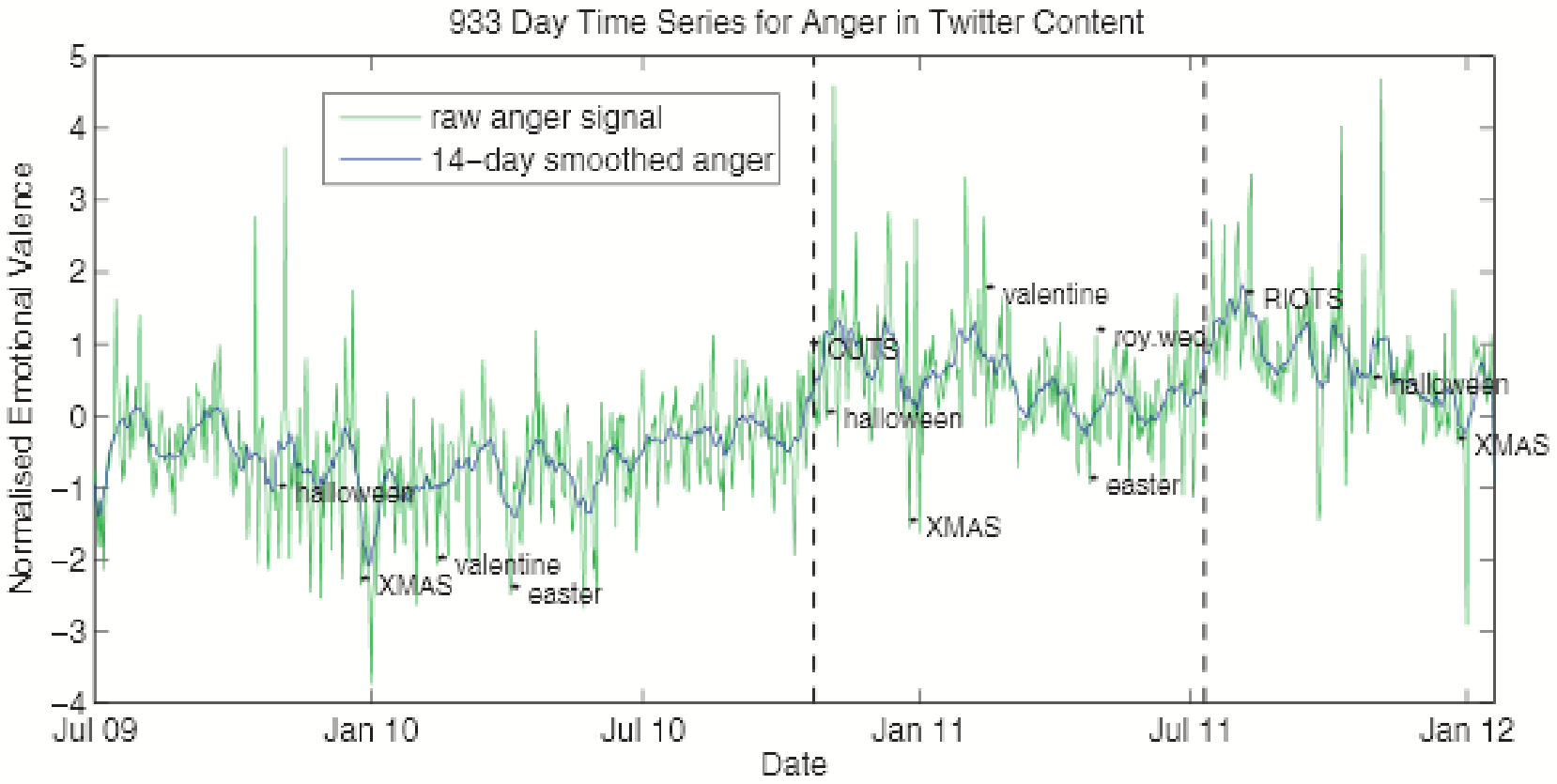
Previous Research

- Lansdall-Welfare, Lampos, & Cristianini, 2012



Previous Research

- Lansdall-Welfare, Lampos, & Cristianini, 2012

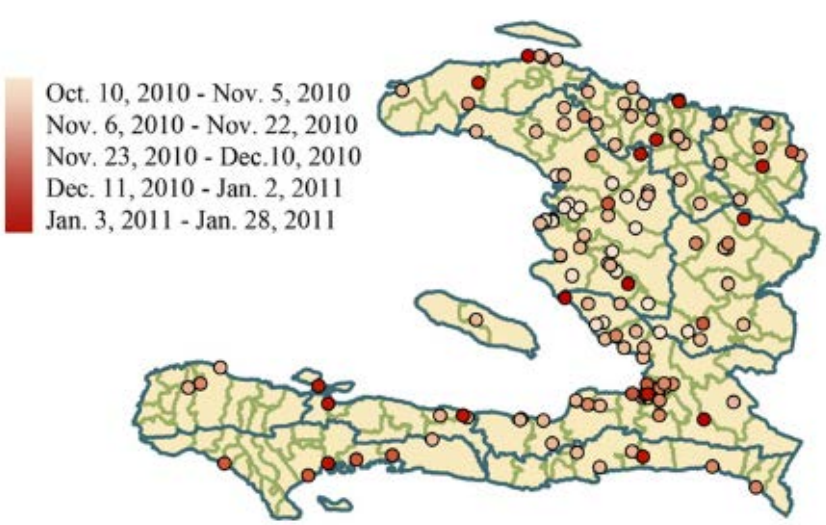


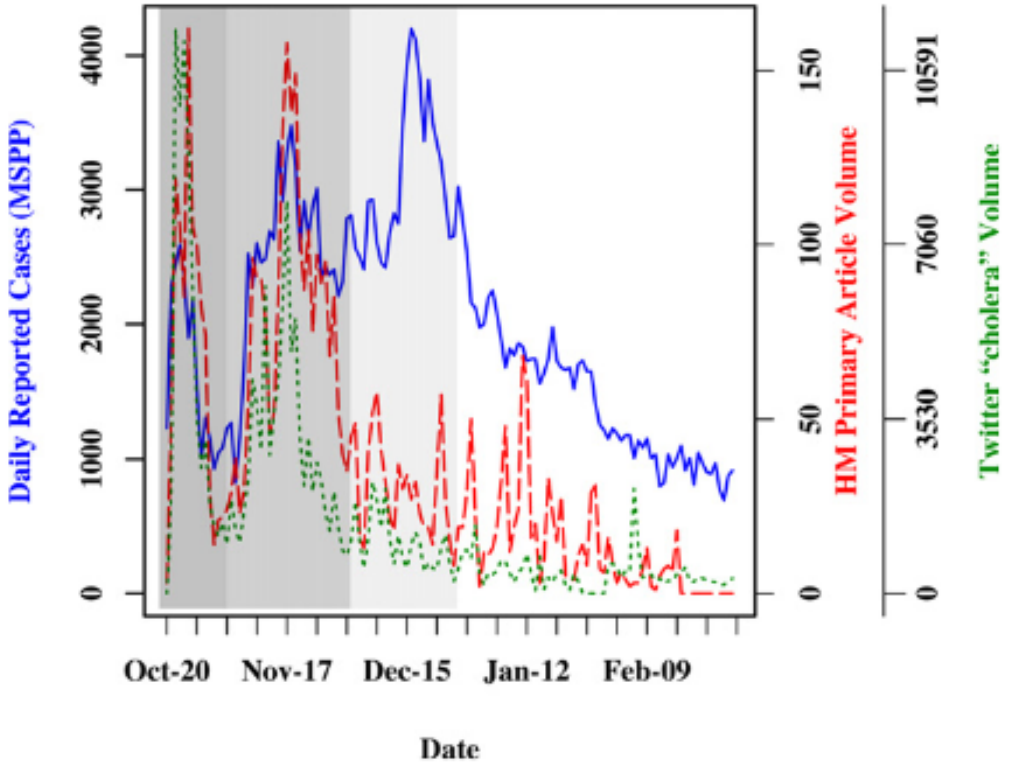
Previous Research

- Chunara, Andrews, & Brownstein, 2012
  - Used Twitter data and geospatial analysis to monitor spread of Cholera in Hati
  - Found correlation between microblog data and reported Cholera Cases
  - Microblog data predicted # of cases, and did it *faster* than official government reporting mechanisms
  - Data from first 100 days of the Haitian Cholera outbreak, 2010-2011
  - All Tweets from Haiti with word “cholera” or hashtag #cholera
  - Sources:
    - MSPP (Government Reports)
    - Healthmap (Online News Stories, mobile App data aggregates)
    - Twitter (Tweets, informal source)

Previous Research

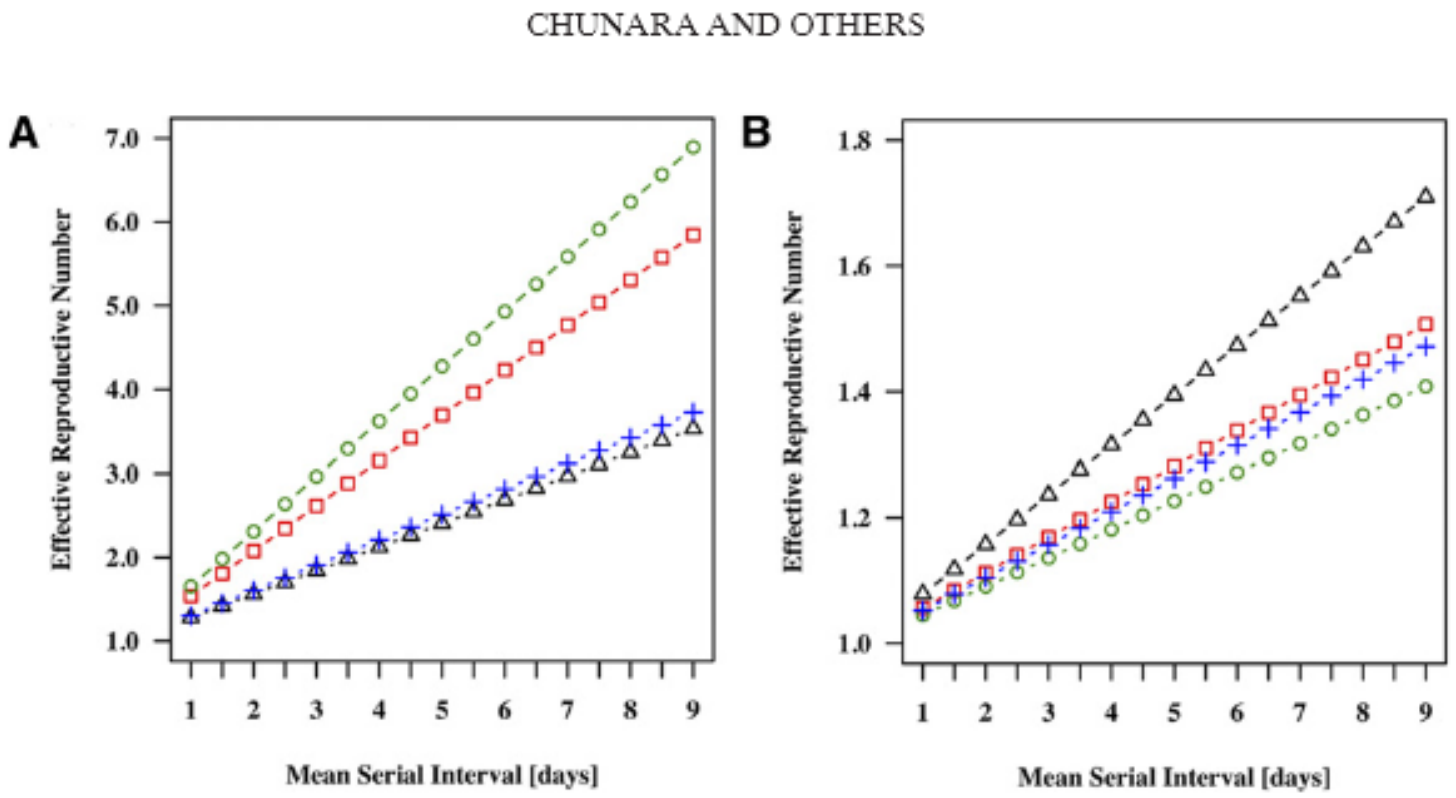
- Chunara, Andrews, & Brownstein, 2012





Previous Research

- Chunara, Andrews, & Brownstein, 2012



Our Research

- Our focus: To consider these methods and possibly combine statistical analyses in order to develop predictive models of Social Trends
- *Can we not only describe but (roughly) predict social outcomes based on readily available and voluminous microblog/web data?*

Potential Data Sources

- Twitter:
  - (2010) Asked $30 Million for full Firehose access from MSFT
  - Streaming API: 1% of Firehose
  - Deloitte Consulting – Artificial 3000/min restriction (Unpublished)
- GNIP (www.gnip.com)
  - Provides both realtime and historical access to Tweets
  - Working with LOC to make historical access free to researchers
  - Technical Limits/Privacy/Access mechanisms!
- Topsy (www.topsy.com)
  - “Geo-Inference”: Machine Algorithm Learning to fill in Geo-Location when user does not tag (99% of all tweets – 90% Confidence)
- Datasift (www.datasift.com)
  - Cost: $0.10 per 1000 tweets

```
{
    "interaction":{
        "source":"TweetDeck",
        "author":{
            "username":"stewarttownsend",
            "name":"Stewart Townsend",
            "id":14065694,
            "avatar":"http://a2.twimg.com/profile_images/1302306721/twitterpic_normal.jpg",
            "link":"http://twitter.com/stewarttownsend"
        },
        "type":"twitter",
        "link":"http://twitter.com/stewarttownsend/statuses/136447843652214784",
        "created_at":"Tue,
         15 Nov 2011 14:17:55 +0000",
        "content":"Morning San Francisco - 36 hours and counting.. #datasift",
        "id":"1e10f949c51aab80e074df944f5e8e46"
    },
    "twitter":{
        "user":{
            "name":"Stewart Townsend",
            "url":"http://www.stewarttownsend.com",
            "description":"Developer Relations at Datasift (www.datasift.com) - Car racing petrol head,
             all things social lover,
             co-founder of www.flowerytweetup.com",
            "location":"iPhone: 53.852402,
            -2.220047",
            "statuses_count":28247,
            "followers_count":3094,
            "friends_count":510,
            "screen_name":"stewarttownsend",
            "lang":"en",
            "time_zone":"London",
            "listed_count":221,
            "id":14065694,
            "id_str":"14065694",
            "geo_enabled":true
        },
        "id":"136447843652214784",
        "text":"Morning San Francisco - 36 hours and counting.. #datasift",
        "source":"<a href="http://www.tweetdeck.com" rel="nofollow">TweetDeck</a>",
        "created_at":"Tue,
         15 Nov 2011 14:17:55 +0000"
    },
    "demographic":{
        "gender":"male"
    },
    "language":{
        "tag":"en"
    },
    "salience":{
        "content":{
            "sentiment":0
        }
    }
}
```

Our Test Data

- NIST 2011 Microblog Dataset
  - 16 Million Tweet Headers
  - Microblog Track of NIST Text Retrieval Conference (TREC)
  - 1600 .dat “header” files provided
  - NIST personnel-created Java programs for Crawling and Extraction
  - “Pilot” data
- Estimates:
  - 1.26 GB of Tweet Status Headers
  - 2.20 kb of tweet data x 16 million ≈ 33-35 GB of data (estimate)

Our Test Data

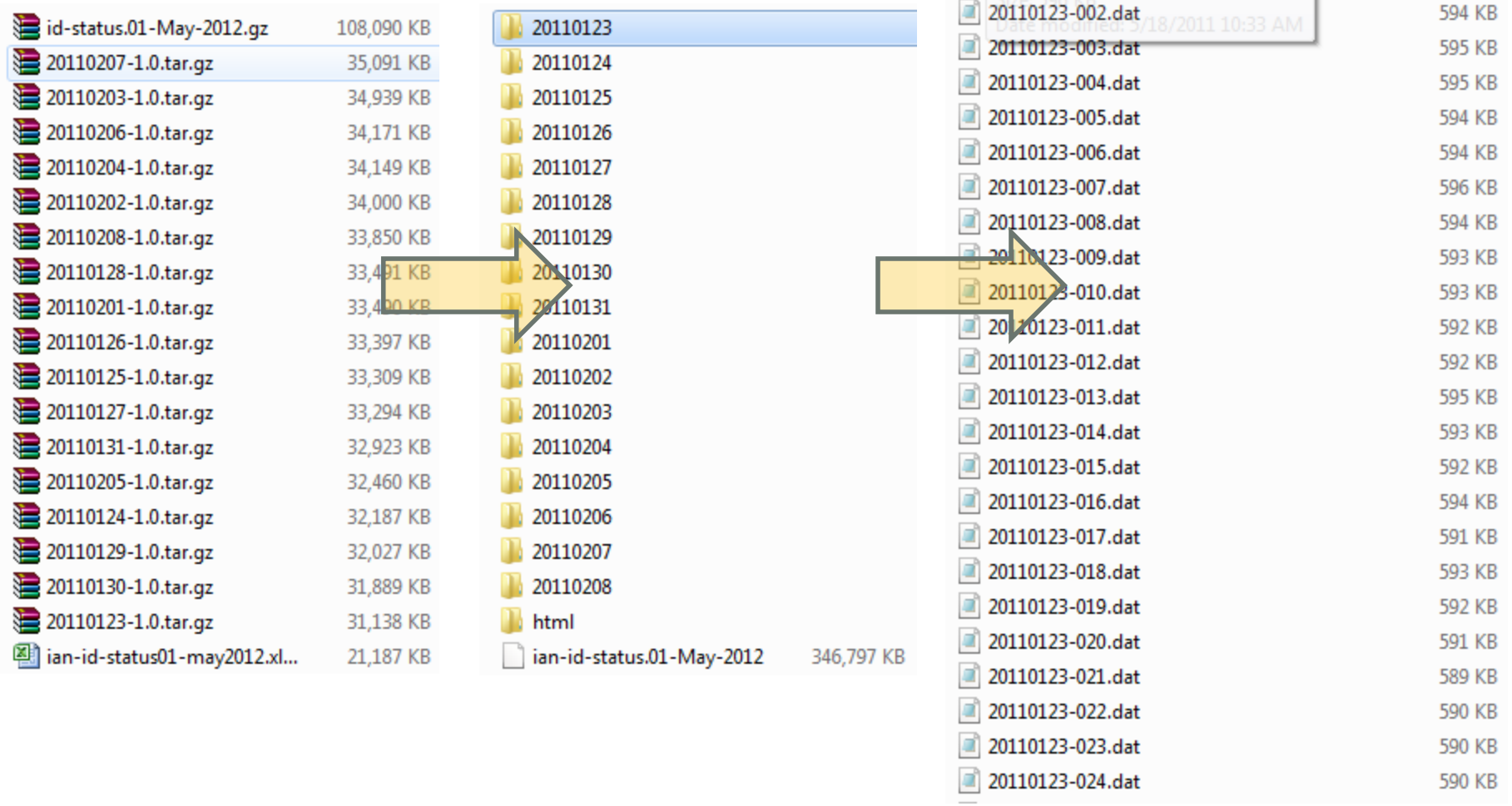
Example of Tweet Status Header Data

ID	Username	Checksum
16965144310579200	cyberrleeashh	2808ef015e3bbcc21c4e48f3de255263
18965145078136800	shigesakau98	0198545914ae9bef1aaef23fe5314483
38965145531125700	cattoinix	cb45bbe356c82e2bdcc60cf8528da591
3925145933774800	l_mazeta	598c4fe5d1324a7e389b19502df53513
48982146361595900	pita_flower1	86015c06e8b5560d22cb79962ad5e031

Masked Example of Downloaded and Extracted Twitter Status Data

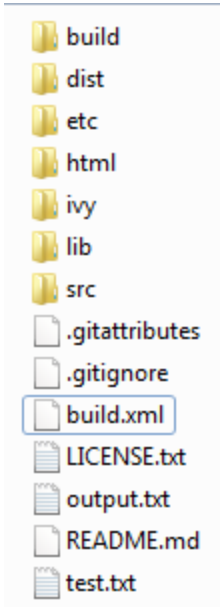
Status ID	Username	Tweet Status	UNIX Timestamp	Status Content
38965133204066567	pianongg	200	1295740800	Akuh males mandi, nantian ajadah mandinyah .
38965133296340805	SangManton	200	1295740802	@Board_Flew_Up Yup. Maybe not for much more time
38965133673832505	GOOUTDARCELL	200	1295740802	@PleaseTweetMuah word thats how we feel haha
38965133988401607	TheRULE_	200	1295740802	Here Here ! RT @donnabrazila_: who wants my 10,000th tweet ?
38965134948896734	goldenreptile67	200	1295740686	#ZodiacFacts #Sagittarius have lots of confidence about themselves.

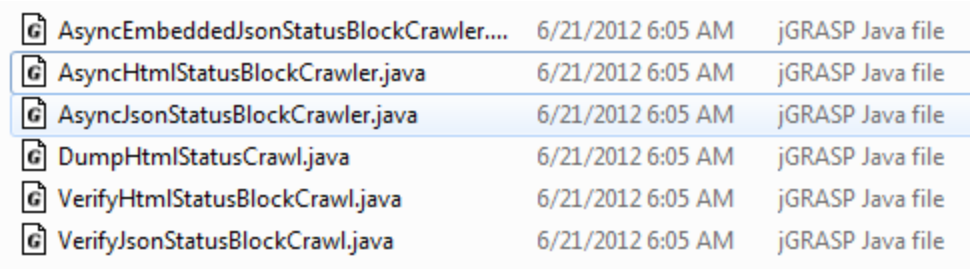
Our Test Data

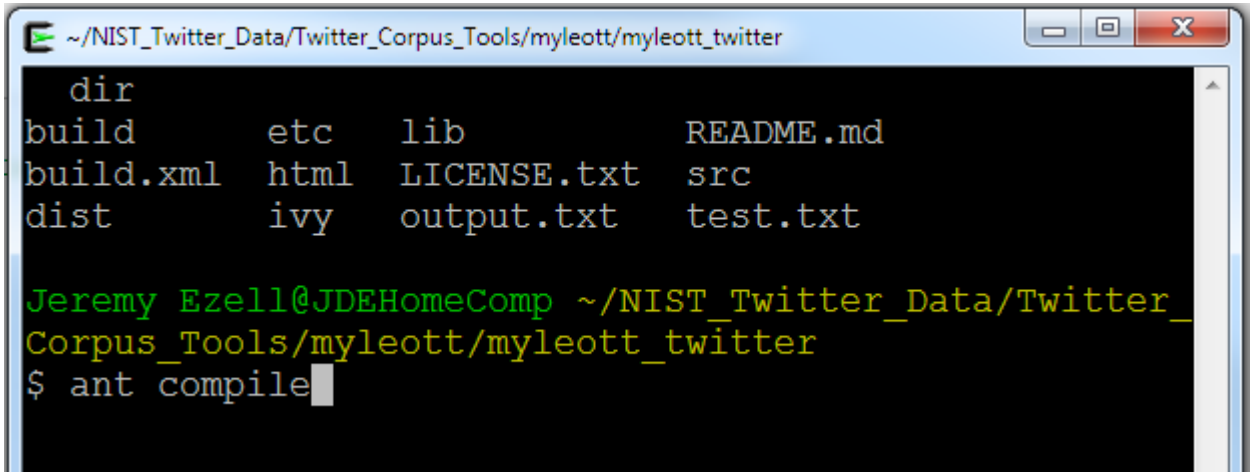


Procedures

- 1. Unzip the files
- 2. Compile the extraction and transform tools







Procedures

- 3. Perform the Extraction step
  - Script run in the virutal Unix environment:
  - java -cp "lib/*;build/*;dist/twitter-corpus-tools-0.0.1.jar" com.twitter.corpus.download.AsyncHtmlStatusBlockCrawler -data "C:\cygwin\home\Jeremy Ezell\NIST_Twitter_Data\Extracted\20110123\20110123-000.dat" -output html/20110123-000.html.seq
- Our Actual Script:
  - for n in C:/home/Scott/nist_twitter_data/Extracted/{1,2,3}/*.dat; do java -cp "lib/*;build/*;dist/twitter-corpus-tools-0.0.1.jar" com.twitter.corpus.download.AsyncHtmlStatusBlockCrawler -data $n -output $n.html.seq; done

Procedures

~/NIST_Twitter_Data/Twitter_Corpus_Tools/myleott/myleott_twitter

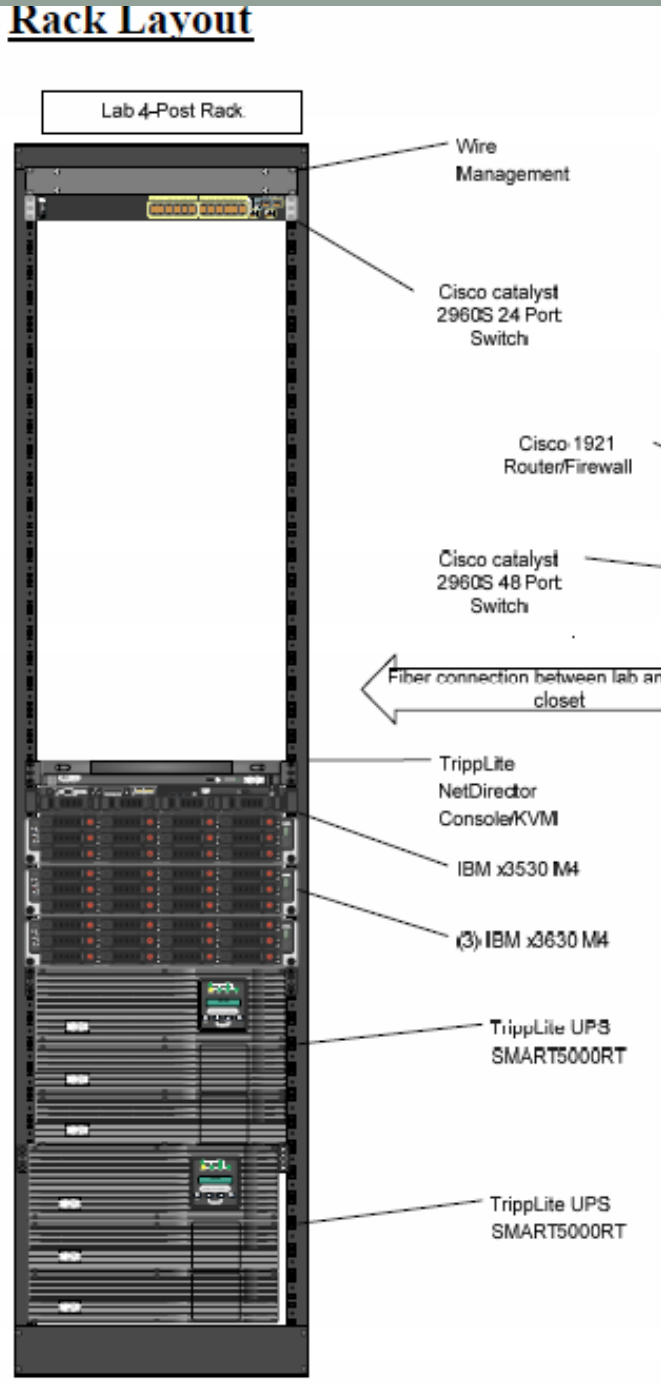
```
$ cd /home/Scott/NIST_Twitter_Data/Twitter_Corpus_Tools/myleott/myleott_twitter

Scott@Scott-PC ~/NIST_Twitter_Data/Twitter_Corpus_Tools/myleott/myleott_twitter
$ for n in C:/cygwin/home/Scott/NIST_Twitter_Data/Extracted/20110123/*.dat; do java -cp "lib/*;build/*;dist/twitter-corpus-tools-0.0.1.jar
" com.twitter.corpus.download.AsyncHtmlStatusBlockCrawler -data $n -output $n.html.seq; done
cygwin warning:
  MS-DOS style path detected: C:/cygwin/home/Scott/NIST_Twitter_Data/Extracted/20110123/
  Preferred POSIX equivalent is: /home/Scott/NIST_Twitter_Data/Extracted/20110123/
  CYGWIN environment variable option "nodosfilewarning" turns off this warning.
  Consult the user's guide for more details about POSIX paths:
    http://cygwin.com/cygwin-ug-net/using.html#using-pathnames


13/03/26 10:04:10 INFO download.AsyncHtmlStatusBlockCrawler: Processing C:\cygwin\home\Scott\NIST_Twitter_Data\Extracted\20110123\2.dat
13/03/26 10:04:22 INFO download.AsyncHtmlStatusBlockCrawler: Total request submitted: 68
13/03/26 10:04:22 INFO download.AsyncHtmlStatusBlockCrawler: 6 tweets fetched in 12776ms
13/03/26 10:04:22 INFO download.AsyncHtmlStatusBlockCrawler: Writing tweets...
13/03/26 10:04:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
13/03/26 10:04:26 INFO compress.CodecPool: Got brand-new compressor
13/03/26 10:04:26 INFO download.AsyncHtmlStatusBlockCrawler: 6 statuses written.
13/03/26 10:04:26 INFO download.AsyncHtmlStatusBlockCrawler: Done!
13/03/26 10:04:27 INFO download.AsyncHtmlStatusBlockCrawler: Processing C:\cygwin\home\Scott\NIST_Twitter_Data\Extracted\20110123\3.dat
13/03/26 10:04:37 INFO download.AsyncHtmlStatusBlockCrawler: Total request submitted: 68
13/03/26 10:04:37 INFO download.AsyncHtmlStatusBlockCrawler: 4 tweets fetched in 10593ms
13/03/26 10:04:37 INFO download.AsyncHtmlStatusBlockCrawler: Writing tweets...
13/03/26 10:04:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
13/03/26 10:04:38 INFO compress.CodecPool: Got brand-new compressor
13/03/26 10:04:38 INFO download.AsyncHtmlStatusBlockCrawler: 4 statuses written.
13/03/26 10:04:38 INFO download.AsyncHtmlStatusBlockCrawler: Done!
13/03/26 10:04:38 INFO download.AsyncHtmlStatusBlockCrawler: Processing C:\cygwin\home\Scott\NIST_Twitter_Data\Extracted\20110123\test.dat
13/03/26 10:04:50 INFO download.AsyncHtmlStatusBlockCrawler: Total request submitted: 68
13/03/26 10:04:50 INFO download.AsyncHtmlStatusBlockCrawler: 5 tweets fetched in 11934ms
13/03/26 10:04:50 INFO download.AsyncHtmlStatusBlockCrawler: Writing tweets...
13/03/26 10:04:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
13/03/26 10:04:51 INFO compress.CodecPool: Got brand-new compressor
13/03/26 10:04:51 INFO download.AsyncHtmlStatusBlockCrawler: 5 statuses written.
13/03/26 10:04:51 INFO download.AsyncHtmlStatusBlockCrawler: Done!
```

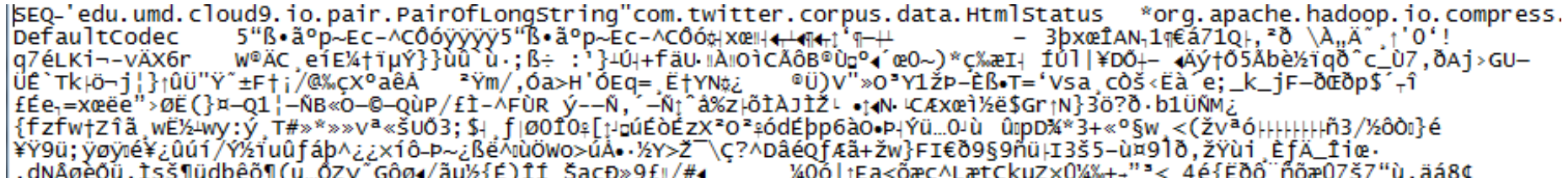
Procedures

- ***Hardware:***
  - Auburn University Cyber Security Initiative
  - (Hardware hosted at the Auburn University Airport)
  - Qty 1: IBM x3530 M4 Management Server
    - Intel Xeon 6C(ore) 1.90 GHz, 15 MB Cache
    - 32GB of RAM
    - 500 GB 7,200 RPM Drive x Qty 3 (RAID 5)
  - Qty. 3 IBM x3630 M4 Blade Servers
    - Dual Intel Xeon 2.00 GHz, 15 MB Cache, 6C(ore)
    - 64GB RAM
    - 1.5TB 15,000 RPM Hot-Swap HDD (Qty 6) (RAID 5)
    - 6 x 1GB Ethernet
  - 50+ MB Rack Connection to the Internet



.SEQ Data

- Hadoop sequence file



Procedures:

- 4. Perform the Transformation
  - We need to convert the .seq files to .txt or “human-readable” files
  - Script:

  - for n in C:/cygwin/home/Scott/NIST_Twitter_Data/Extracted/{1,2,3}/*.seq; do java -cp "lib/*;build/*;dist/twitter-corpus-tools-0.0.1.jar" com.twitter.corpus.demo.ReadStatuses -input $n -dump -html > $n.txt; done

Status of Extraction

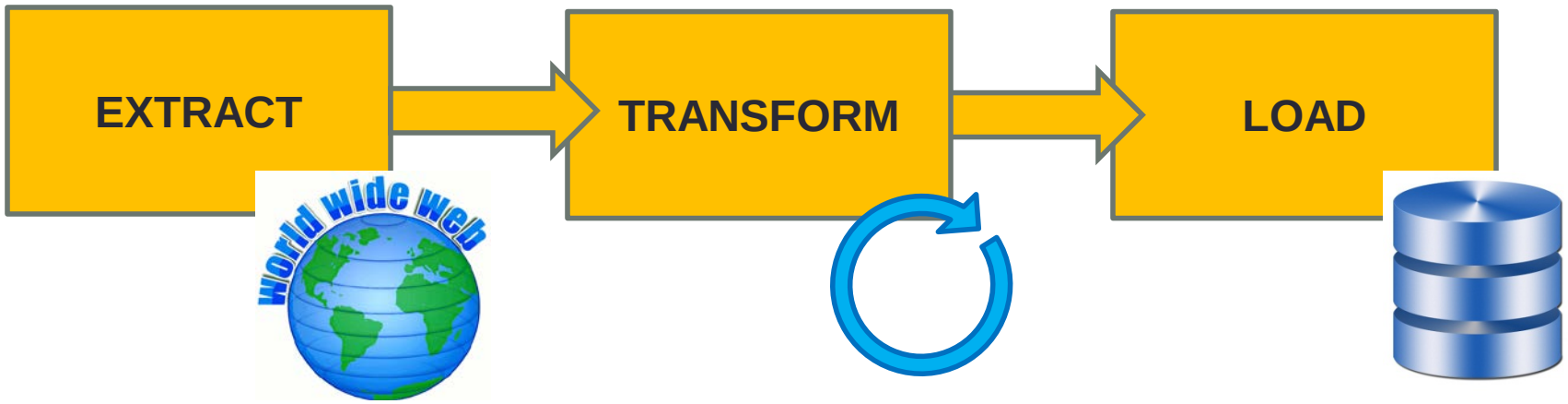
11:10 p.m. Monday Night (Running since Thursday, 8:00 p.m.) --- **99 Hours!**

Folders	Name	Files	Files Processed	Folder Size
1	20110123	93	93	7.27
2	20110124	95	95	7.53
3	20110125	99	99	7.97
4	20110126	99	99	7.75
5	20110127	98	98	7.88
6	20110128	99	99	8
7	20110129	95	95	7.44
8	20110130	94	94	7.58
9	20110131	96	59	
10	20110201	99		
11	20110202	100		
12	20110203	103		
13	20110204	101		
14	20110205	96		
15	20110206	102		
16	20110207	104		
17	20110208	100		
Total		1673	831	61.42
% Done		49.67%		

In GB!

Procedures

- 5. Store in Database
  - Transformed file sizes *should* be much smaller than .SEQ files.
  - Oracle or MySQL database
  - *Third* script will be created for the loading
  - Overall, this follows the ETL process for Data Warehousing



Planned Analysis

- Hope is that further Tweet fields can be extracted
- Development of a word-cloud (Similar to Lansdall-Welfare et al., 2012, UK-Twitter Study)
- n-grams of size 2, 3, 4, and 5 to compare to a larger 1-Trillion n-gram Corpus (Evert, 2010).
- Estimating correlations between word groupings and social events, attempt at forecasting (Kulkarni et al., 2011).
- One Problem: defining just when a social event has occurred, secondary data, etc. *How do we quantify what has come before in order to attempt prediction at what might be yet-to-come?*
- *Could Geographic region be a statistically-significant covariate?*

Thank you!

- Questions?

References

- Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1), 39-45.
- Evert, S. (2010). *Google Web 1T 5-Grams Made Easy (but not for the computer).* Paper presented at the Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop.
- Kulkarni, G., Kannan, P., & Moe, W. (2011). Using online search data to forecast new product sales. *Decision Support Systems*.
- Lansdall-Welfare, T., Lampos, V., & Cristianini, N. (2012). Nowcasting the mood of the nation. *Significance*, 9(4), 26-28.
- Naaman, M., Boase, J., & Lai, C. H. (2010). *Is it really about me?: message content in social awareness streams.* Paper presented at the Proceedings of the 2010 ACM conference on Computer supported cooperative work.
- Stubbs, E. (2011). *The Value of Business Analytics: Identifying the Path to Profitability*. Hoboken, New Jersey: John Wiley & Sons, Inc.