Data Mining & Machine Learning

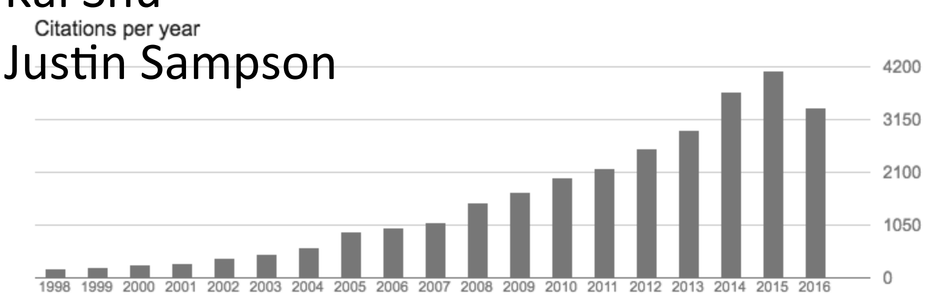
Arizona State University

Evaluation Dilemmas in Social Media Research

Huan Liu

Thanks to Former and Current PhD Students of DMML

- Reza Zafarni, Asst Prof, Syracuse U
- Xia Hu, Asst Prof, Texas A&M U
- Magdiel Galan, Intel
- Shamanth Kumar, Castlight Health
- Pritam Gundecha, IBM Res Almaden
- Jiliang Tang, Asst Prof, MSU
- Huiji Gao, LinkedIn
- Ali Abbasi, Machine Zone
- Salem Alelyani, Asst Prof, King Khalid U
- Xufei Wang, LinkedIn
- Geoffrey Barbier, AFRL
- Lei Tang, Yahoo! Labs
- Zheng Zhao, Google
- Nitin Agarwal, Chair Prof, UALR
- Sai Moturu, PostDoc, MIT Media Lab
- Lei Yu, Assc Prof, Binghamton U, NY
- Robert Trevino, AFRL
- Yunzhong Liu, LeEco, US
- Somnath Shahapurkar, FICO
- Fred Morstatter
- Isaac Jones
- Suhas Ranganath
- Suhang Wang
- Tahora Nazer
- Jundong Li
- Liang Wu
- Ghazaleh Beigi
- Kai Shu
- Justin Sampson



Evaluation Dilemmas

1. Understanding the understanding
   - How to measure the interpretability of machine-learned topics?
2. Sample Data Dilemma
   - Inaccessibility to full data vs. sampling bias
3. When-to-Stop Dilemma
   - Collecting data forever vs. having credible patterns

1. Understanding the Understanding (UtU)

- How to measure interpretability of topics generated by machine learning?
- One common way is to indirectly measure predictive performance of these learned topics
  - The higher the performance (say, accuracy), the better
  - It may not be about understanding
  - Human experts seem to be the best evaluator
- But involving human experts in evaluation may not be ***scalable*** and ***reproducible***
- Hence, it is challenging to UtU

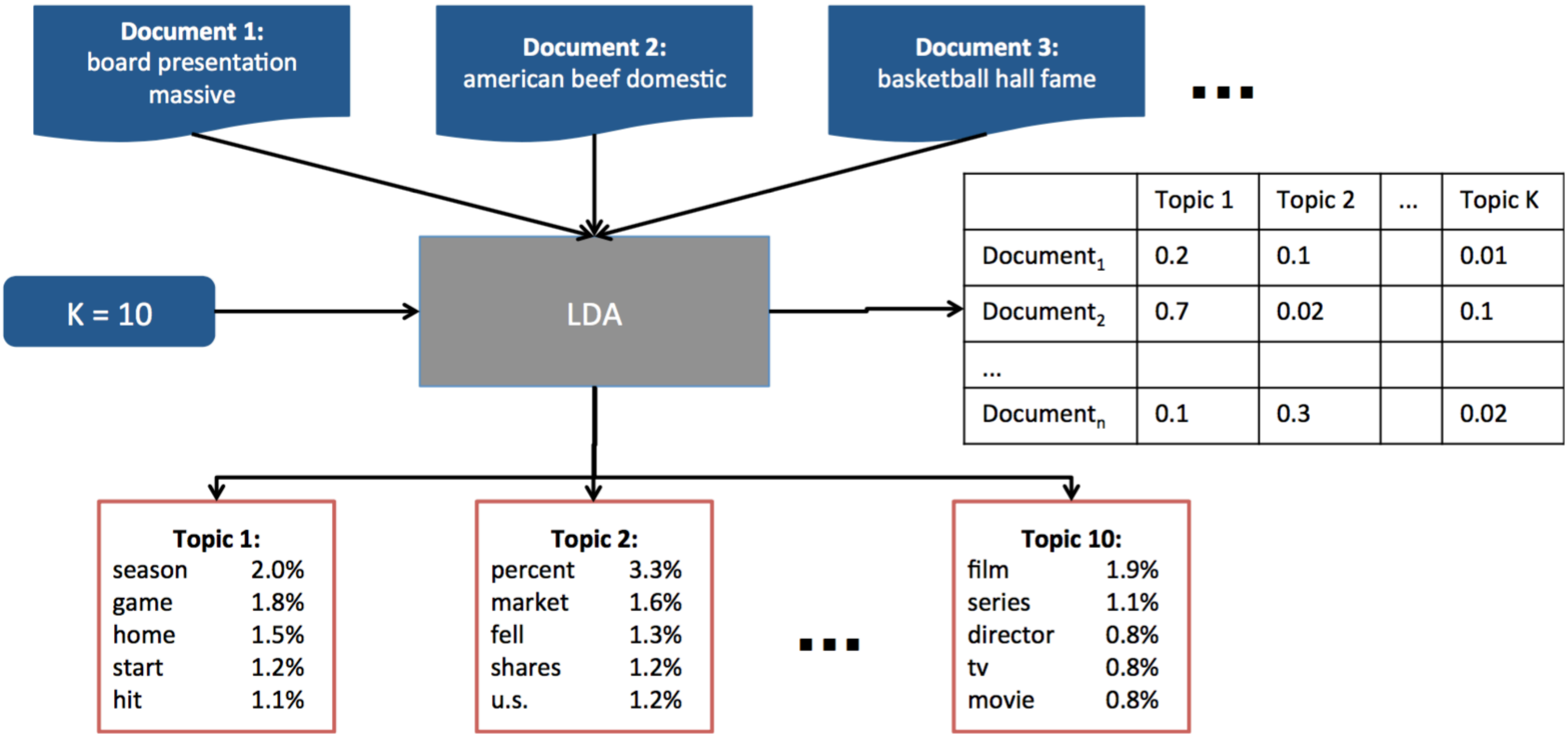
An Example of Big Text Data

- Some example corpora:

Source	Size
Wikipedia	36 **million** articles
World Wide Web	100+ **billion** static web pages
Social Media	500 **million** new tweets **each** day

- Too much data to read
- How can we begin to understand all of this data?

Topic Models



	Topic 1	Topic 2	...	Topic K
Document_1	0.2	0.1		0.01
Document_2	0.7	0.02		0.1
...				
Document_n	0.1	0.3		0.02

Measuring the Understanding

- How do we measure the interpretability of statistical topic models

- A dilemma - <u>Experts</u> are **credible**, but **not scalable**, and <u>crowdsourcing</u> needs *no experts*, so **scalable**, but has *no expertise*, thus **not credible**

A Measure of Topic Interpretability

- ***Model Precision***
- It shows a Turker 6 words in random order
  - Top 5 words from the topic
  - 1 “Intruded” word
  - Ask the Turker to identify the “Intruded” word

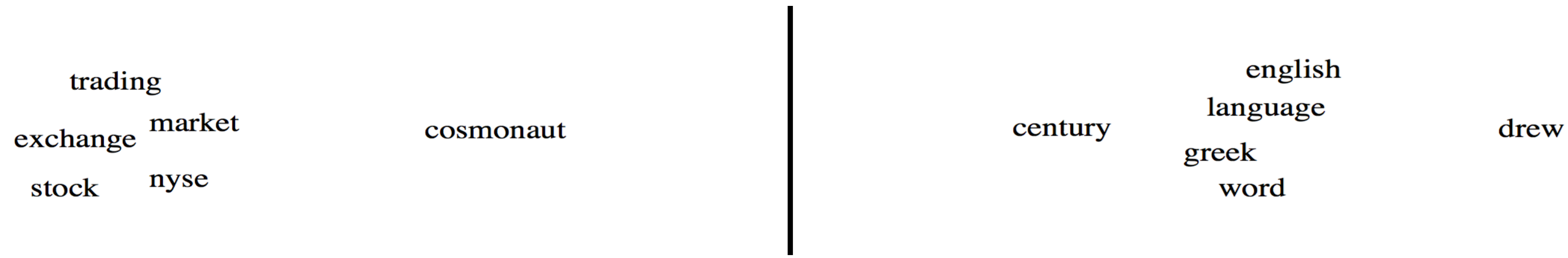
$$MP_{model,topic} = \#\ Correct\ Guesses\ /Total\ \#\ Guesses$$

Topic *i*:

cat	dog	bird	truck	horse	snake

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading Tea Leaves: How Humans Interpret Topic Models." In Advances in Neural Information Processing Systems, pp. 288-296. 2009.

Observing Model Precision (MP)



What does Model Precision measure?

What doesn't Model Precision measure?

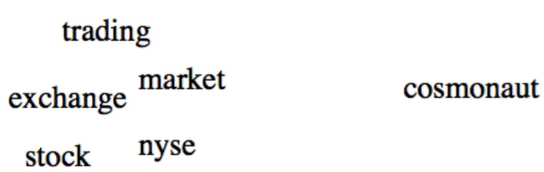
It seems we need another measure

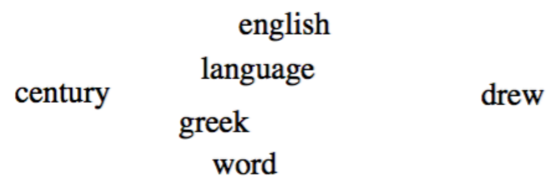
Measuring Coherence – Another Measure

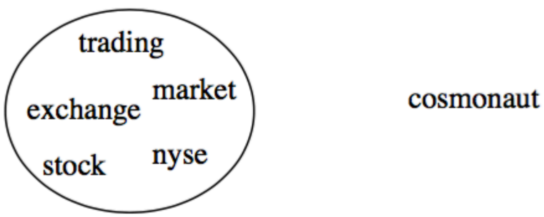
- *Model Precision **Choose Two***
- Nearly the same setup as Model Precision:
  - **Difference:** A Turker is asked to **choose top** two words
- Intuition: if the topic is coherent, then it would be difficult to consistently choose a second word

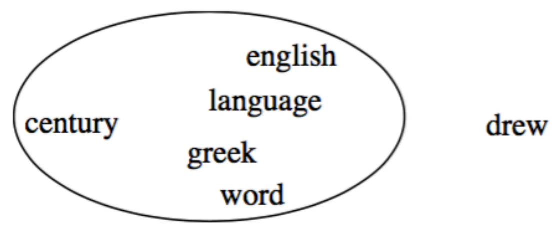
$$MPCT_k^m = H\left(p_{turk}(\mathbf{w}_{k,1}^m), \ldots, p_{turk}(\mathbf{w}_{k,5}^m)\right)$$

A Comparative Example

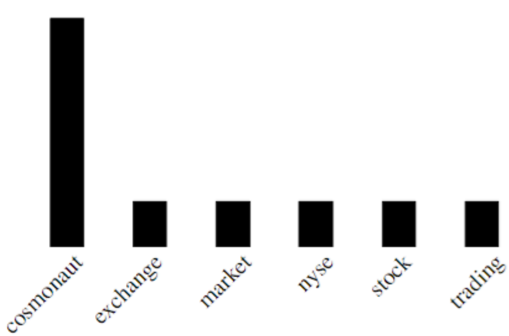


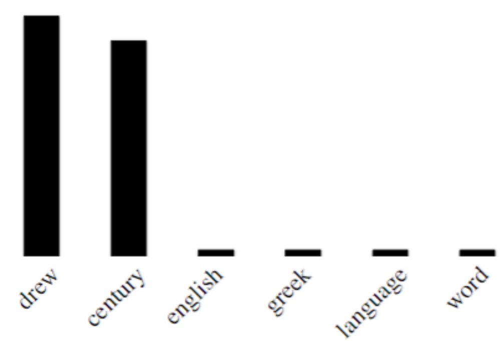






Model Precision





Model Precision Choose Two

News Corpus for Experiments

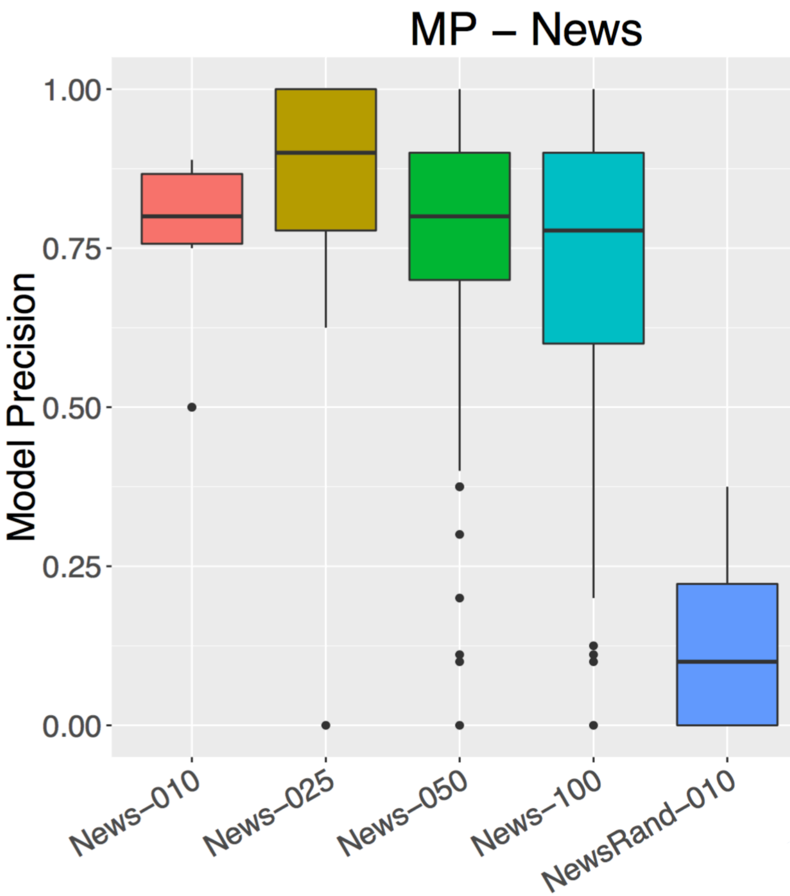
Yahoo! News Dataset

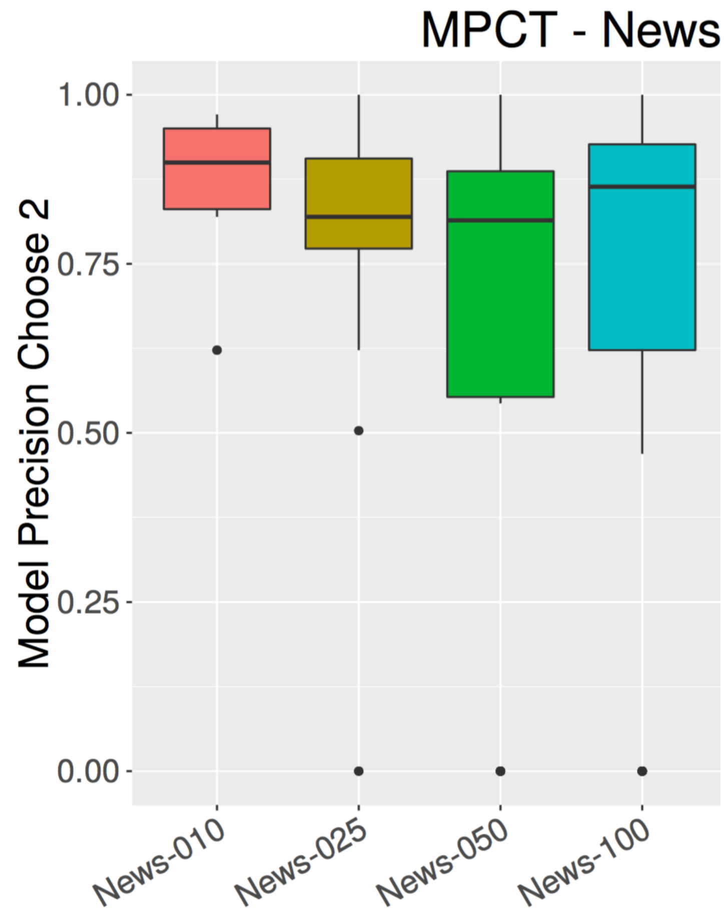
Property	Value
Documents	258,919
Tokens	6,888,693
Types	214,957

Name	Dataset	Strategy	Topics
News-010	News	LDA	10
News-025	News	LDA	25
News-050	News	LDA	50
News-100	News	LDA	100

Can MPCT Replace MP?

- Yahoo! News, Run with K = 10, 25, 50, 100.
- "Random" Topics





MPCT vs. MP

Top 5 Words	Intruded Word	MP Score	MPCT Score
production, plants, provide, food, plant	suppressor	1.00	0.99
number, system, transactions, card, money	flees	1.00	0.97
methods, data, information, analysis, large	diesel	1.00	0.00
series, fans, season, show, episode	leveon	1.00	0.00
nuclear, fundamental, water, understanding, surface	modularity	0.13	0.92
film, khan, ians, actor, bollywood	debonair	0.30	1.00
mechanisms, pathways, involved, molecular, role	specialized	0.00	0.00
injury, left, list, return, surgery	tests-results	0.00	0.25

MPCT Complements MP

- We need both

0 0 | 1 0

0 1 | 1 1

Takeaways

- MPCT measures a topic's *within*-topic distance
- MPCT complements Model Precision
- MPCT provides another dimension of topic quality
  - Low correlation with Model Precision ($\rho = 0.29$)
- Topics and scripts: http://bit.ly/mpchoose2

- A recent blog post on the topic @

http://www.kdnuggets.com/2016/11/measuring-topic-interpretability-crowdsourcing.html

2. Sample Data Dilemma

- Inaccessibility to full social media data
  – Who provides free access to their full data?
- Samples can be gathered via various means
  – Samples are, by definition, limited
- Are all samples biased?
  – Not necessarily
  – Answer could be none, some, all
- How can we be sure it is one of the three?

Twitter Data as an Example

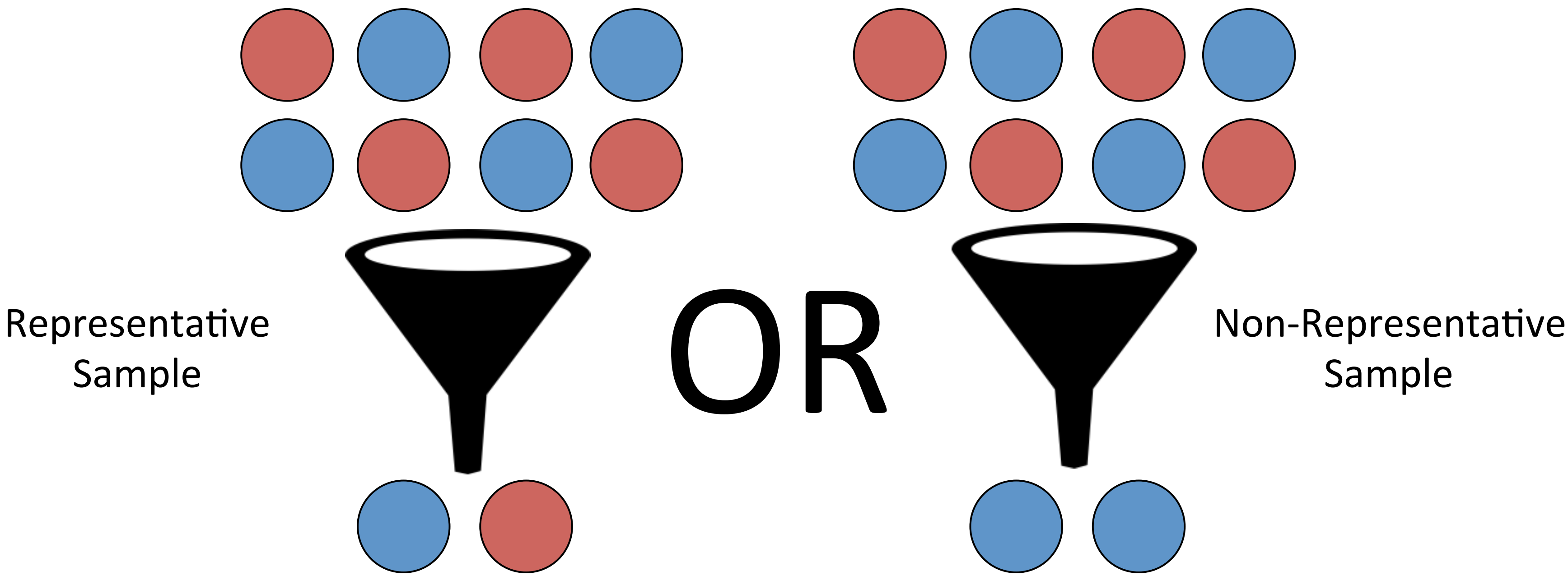
- Social media data is big data
- Twitter is prominent for researchers
  - It share its data
- 500 million tweets/day
- 100 million users/day
- Arab Spring, Natural Disasters, etc.

Why Twitter?

- Twitter shares its data
  - 100%: 500 million tweets / day
  - 1%: 5 million tweets / day
- "Firehose" feed - 100% - costly
- "Streaming API" feed - 1% - free
  - Streaming API takes parameters from user
  - Returns tweets matching parameters
  - Samples data when volume reaches 1%
- **Is 1% data sufficiently good for our research?**

We Have a Problem

- We don't know how Twitter samples data
- Is the sampled data from the Streaming API representative of the true activity on Twitter's Firehose?



Background

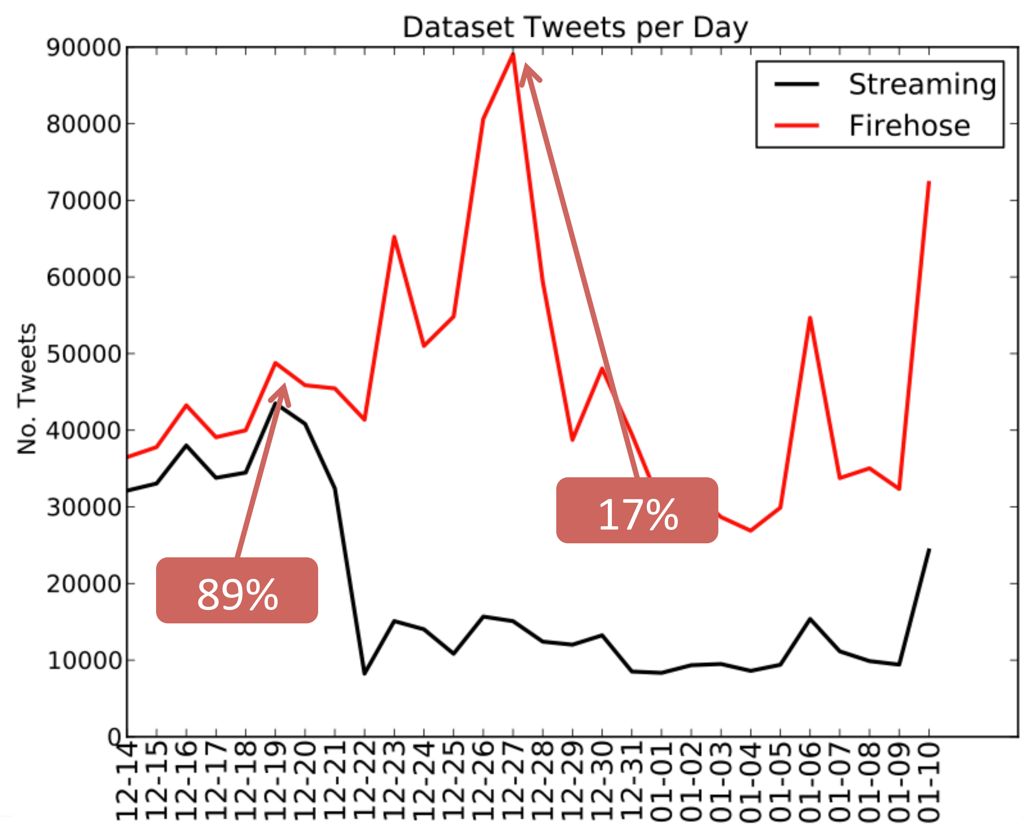
- Studying Arab Spring activity in Syria

Keywords	Geoboxes	Users
#syria, #assad, #aleppovolcano, #alawite, #homs, #hama, #tartous, #idlib, #damascus, #daraa, #aleppo, #سوريا*, #houla	 (32.8, 35.9), (37.3, 42.3)	@SyrianRevo

- Given brief access to Firehose
- Collected data from both the Streaming API and Firehose for 28 days (12/14/2011 to 01/10/2012)

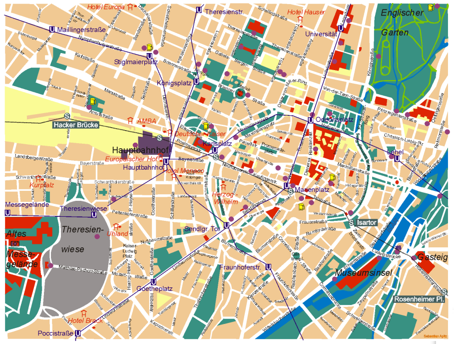
Our Dataset

- 500k from Streaming API
- 1.2M from Firehose
- 42% Overall Coverage
- Daily Coverage from 17% to 89%.

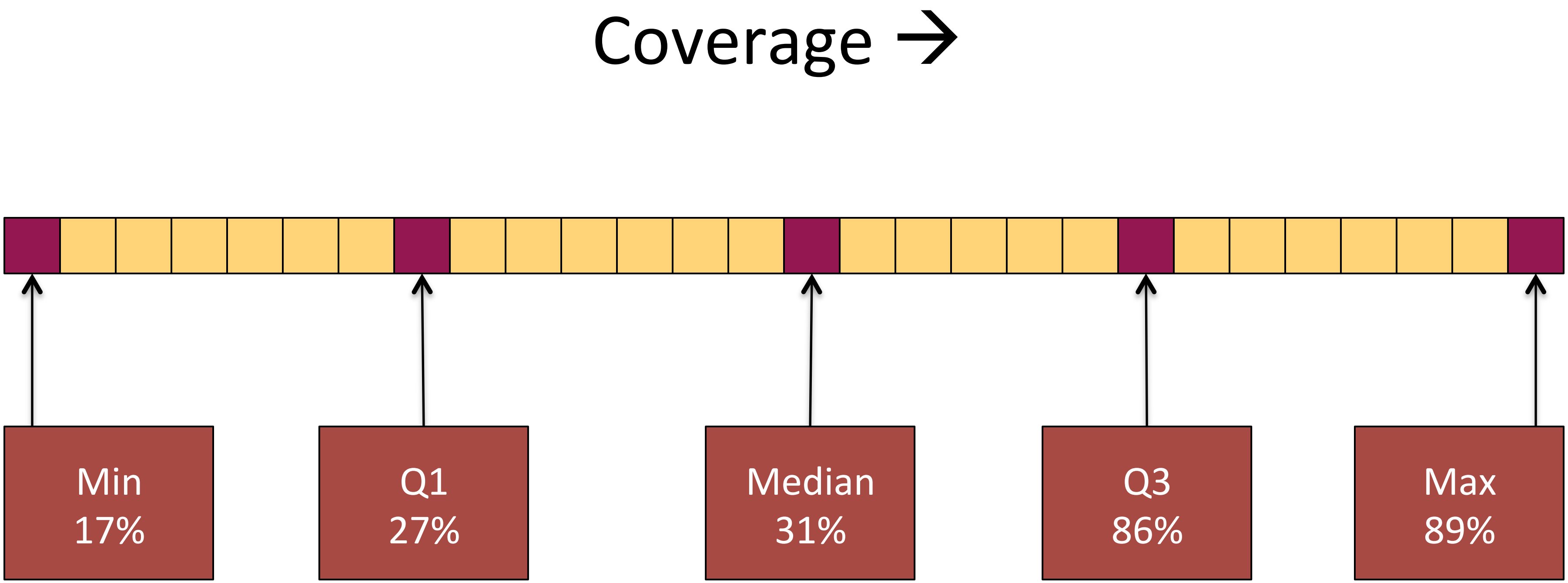


Analysis Choices and An Evaluation Challenge

- Compare facets of the tweet data from Streaming API and Firehose
  - Hashtags, Network Topology, Geographic Distribution
  - *LDA Topics*
- The challenge - we have only **one sample** from Streaming API

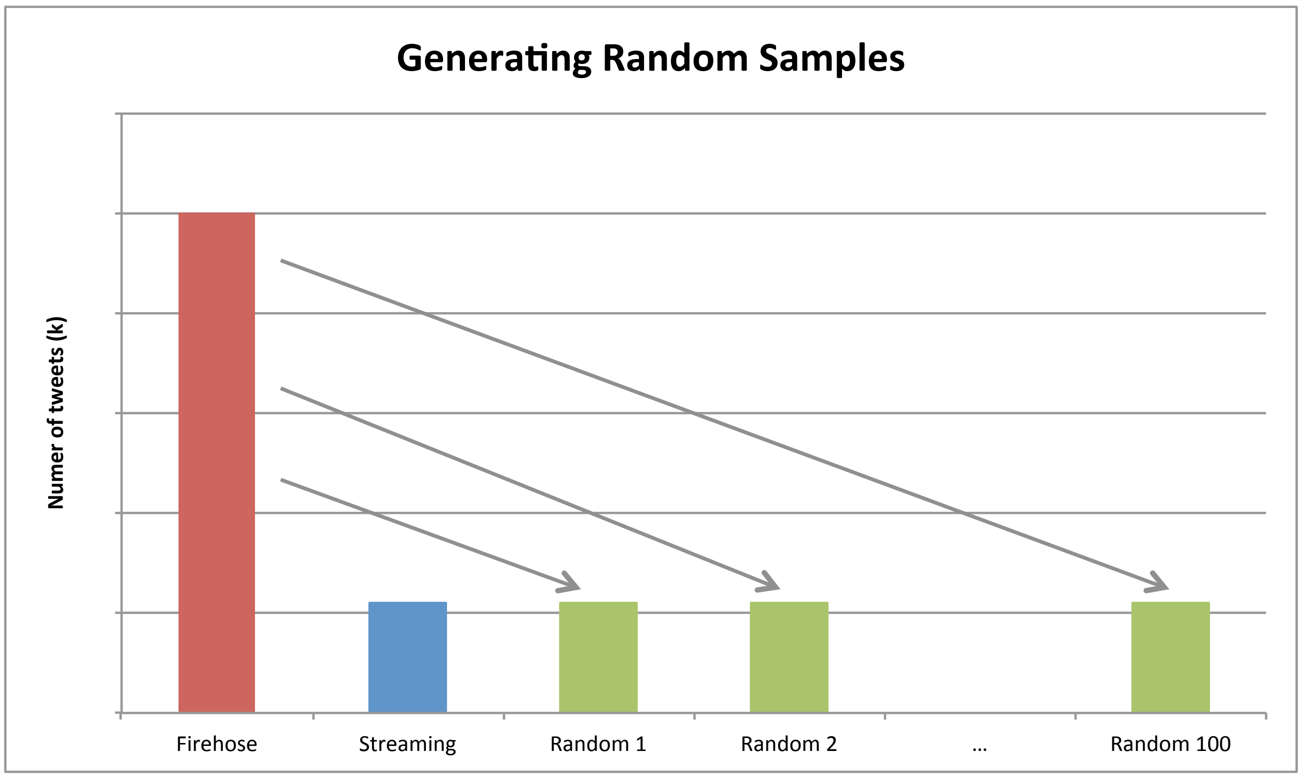


Days of Interest



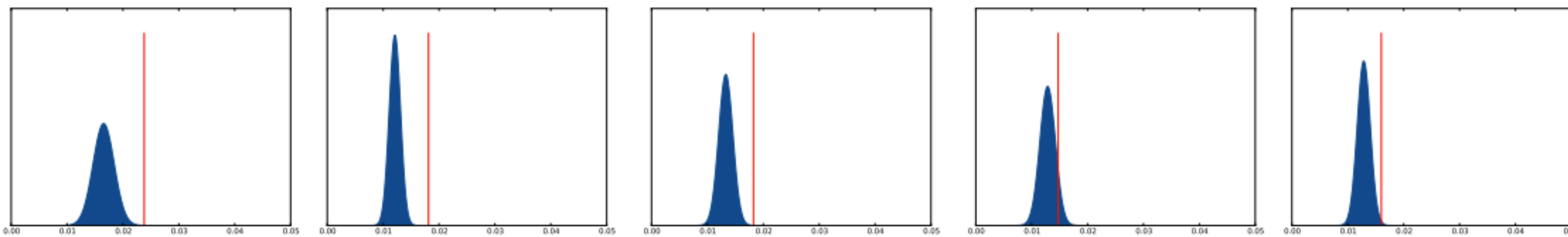
Verification via Sampling

- Created 100 of our own "Streaming API" results by sampling the Firehose data.



Comparison with Random Samples

Is Streaming API data biased or not?



(a) Min. $S = 0.024$, $\hat{\mu} = 0.017$, $\hat{\sigma} = 0.002$, $z = 3.500$.

(b) Q1. $S = 0.018$, $\hat{\mu} = 0.012$, $\hat{\sigma} = 0.001$, $z = 6.000$.

(c) Median. $S = 0.018$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 5.000$.

(d) Q3. $S = 0.014$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 1.000$.

(e) Max. $S = 0.016$, $\hat{\mu} = 0.013$, $\hat{\sigma} = 0.001$, $z = 3.000$.

What if we do not have Firehose?

- How can researchers use the previous results to deal with bias in their own data?

- **Lesson:** There could exist bias
- **Challenge 1**: Need to find out if there is bias or not without Firehose
- **Challenge 2:** Collect more data to minimize bias

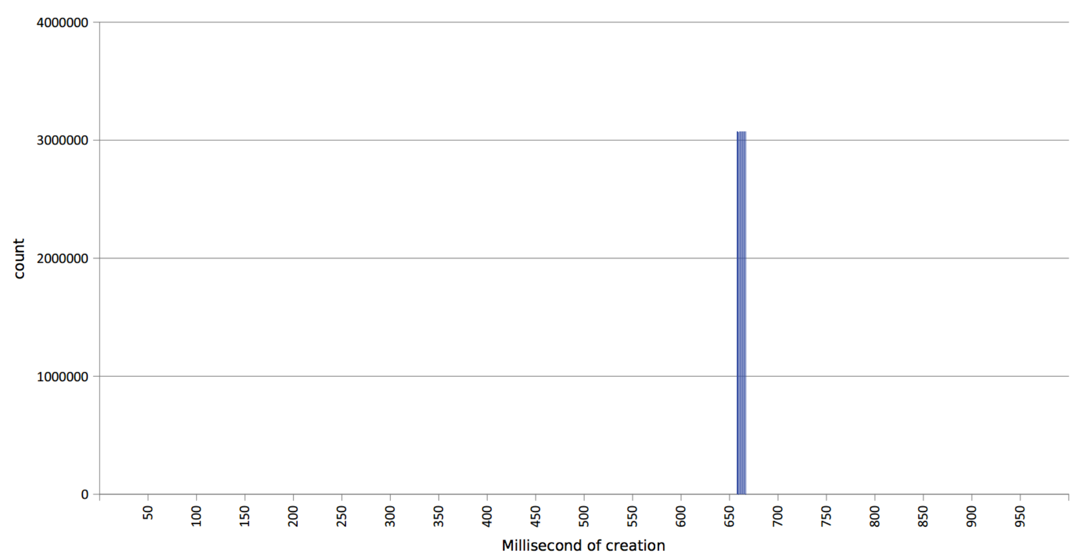
Checking Bias in Existing Data

- We used Firehose to verify if data from Streaming API is biased or not
- For each task, however, it is not feasible to have Firehose for comparison
  - If we had it, then it would be easy to check
- Can we check bias without Firehose?

- Compare Twitter activity with other source(s)
- Use this "other" data as a "thermostat" to assess Streaming API data

Twitter's Sample API

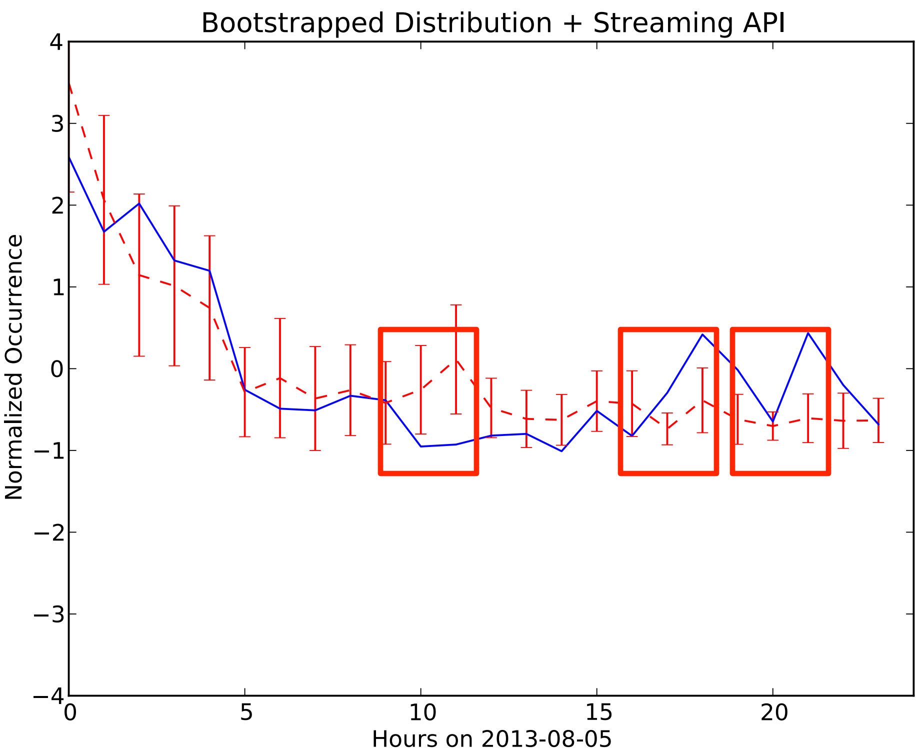
- Samples 1% of all public Tweets
- Does not take any parameters
- Given its nature, Sample API may provide a random sample of the true activity on Twitter
- We perform some tests and find that it is a random sample



[Kergl et. al 2014]

Finding Biased Time Periods without Firehose

- Obtain the trend of hashtag from Sample and Streaming API
- Bootstrap Sample API to obtain confidence intervals
- Mark regions where Streaming API is outside of confidence intervals



Takeaways

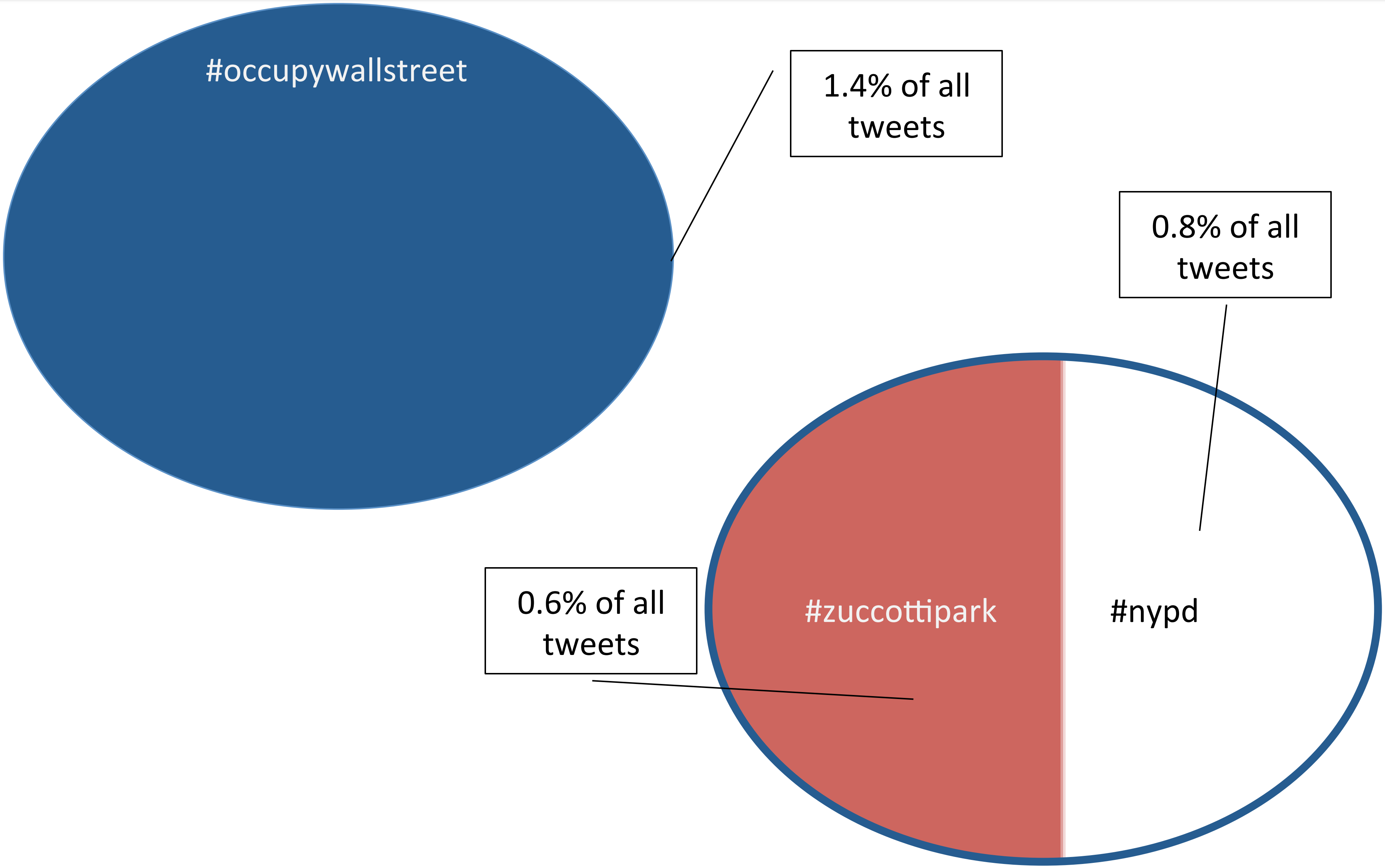
- Sample API is an unbiased Twitter sample
- A methodology to use Sample API is proposed to find periods of bias
- Firehose is not needed

Overcoming Sample Bias

- After detecting bias in our data, what can we do?
- The rationale
  - If we could get all the data for a particular query, there would be no sample bias for sure
- Thus, the more data we can get, the less bias in our data

- **Idea of Mitigating Sample Bias:**
  Leverage multiple crawlers to maximize data for each query

Leveraging Multiple Crawlers



Comparison with Different Numbers of Splits

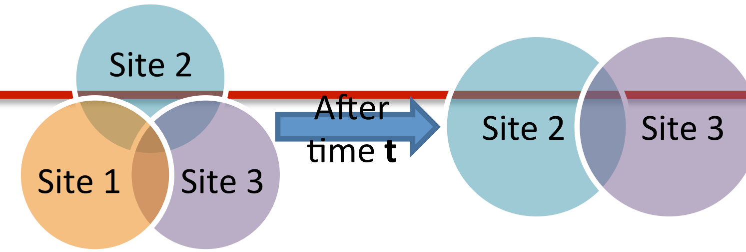
- Word co-occurrence improves growth rate
- Balanced clusters better populate stream bandwidth
- The more splits, the better
- Diminishing returns?

	Unsplit	2-split	3-split
Round Robin	19.02%	50.54%	82.58%
Spectral Clustering	19.02%	28.95%	78.63%

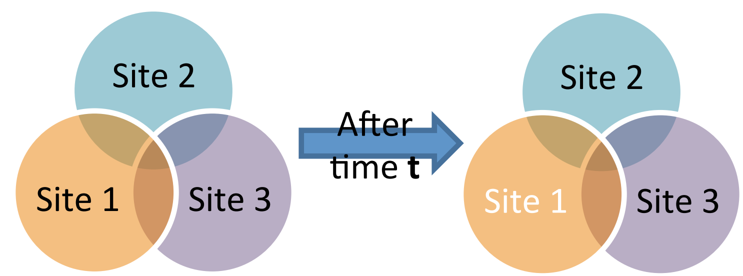
3. When-to-Stop Dilemma

- Collecting data forever vs. having credible patterns
  – How much data vs. how credible
- *Question*: Is There Migration on Social Media?
  – Users are a primary source of revenue
    - Ads, Recommendations, Brand loyalty
  – New SM sites need to *attract* users for expansion
  – Existing SM sites need to *retain* their users
  – Competition for attention entails the understanding of migration patterns

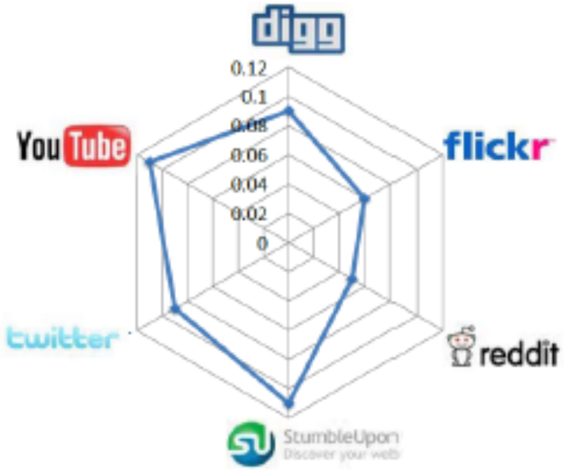
Migration on Social Media



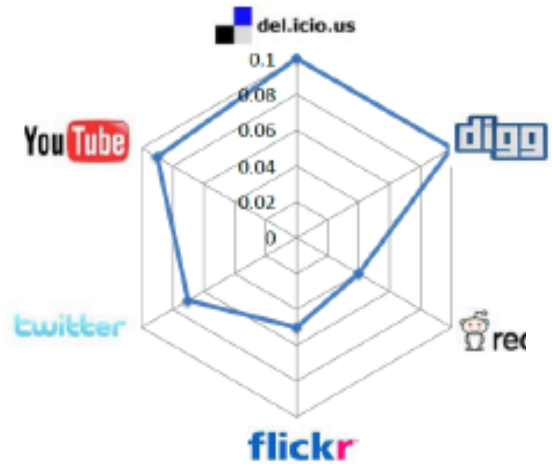
- Site Migration
  - Users leave a site by profile deletion or profile removal
  - Difficult to convince a user who left to return
  - Hard to study these users cross sites because we need their registration information
- **Attention Migration**
  - Users become inactive on a site
  - A harbinger for site migration
  - Can be detected by observing *user activities* across sites
  - Can be studied to prevent site migration by understanding migration patterns



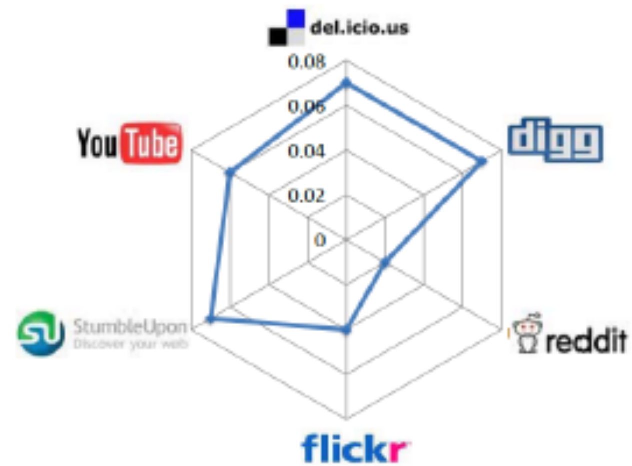
Patterns from Observation



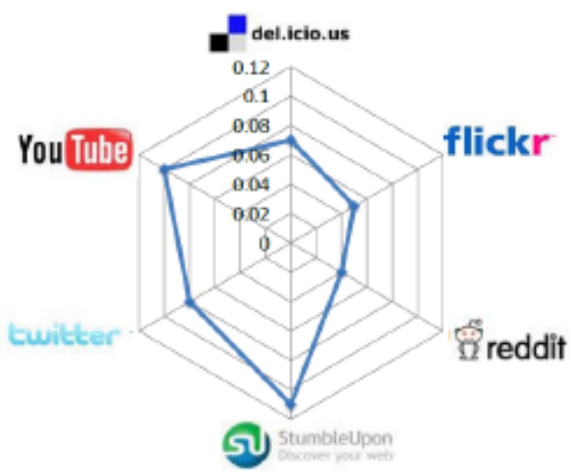
(a) Delicious



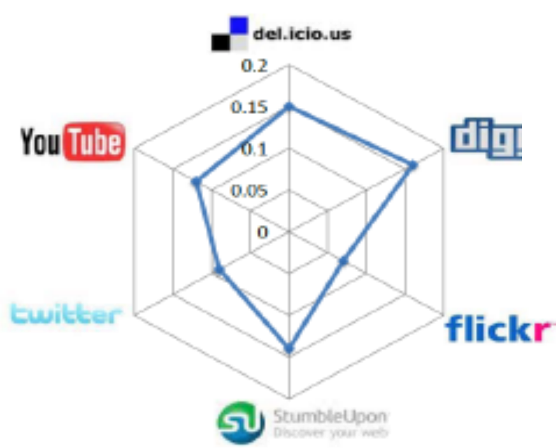
(e) StumbleUpon



(f) Twitter



(b) Digg



(d) Reddit

Can we answer "When to Stop"?

- Pattern evaluation outcome: Significant or not
- Significant differences observed in StumbleUpon, Twitter, and YouTube
- When we are certain, we can stop, otherwise we should continue

Table 2: χ^2 test results on the observed and shuffled data

Site	Observed Coefficients			Shuffled Coefficients			p-value	Statistical Significance
	N	A	R	N	A	R		
Delicious	0.2858	0.4585	-	0.6029	0.5921	-	0.65	Not significant
Digg	0.4796	0.8066	-	0.52	0.5340	-	0.70	Not significant
Flickr	1	1	0.9797	0.2922	0.2759	0.4982	0.13	Not significant
Reddit	0.5385	0.6065	-	0.4846	0.6410	-	0.92	Not significant
StumbleUpon	1	1	-	0.4191	0.2059	-	0.0492	Significant
Twitter	0.5215	1	0.5335	0.2811	0.0365	0.4009	0.0001	Extremely significant
YouTube	0	1	0.1644	0.7219	0.0040	0.4835	0.0001	Extremely significant

"9 Bizarre and Surprising Insights from Data Science"

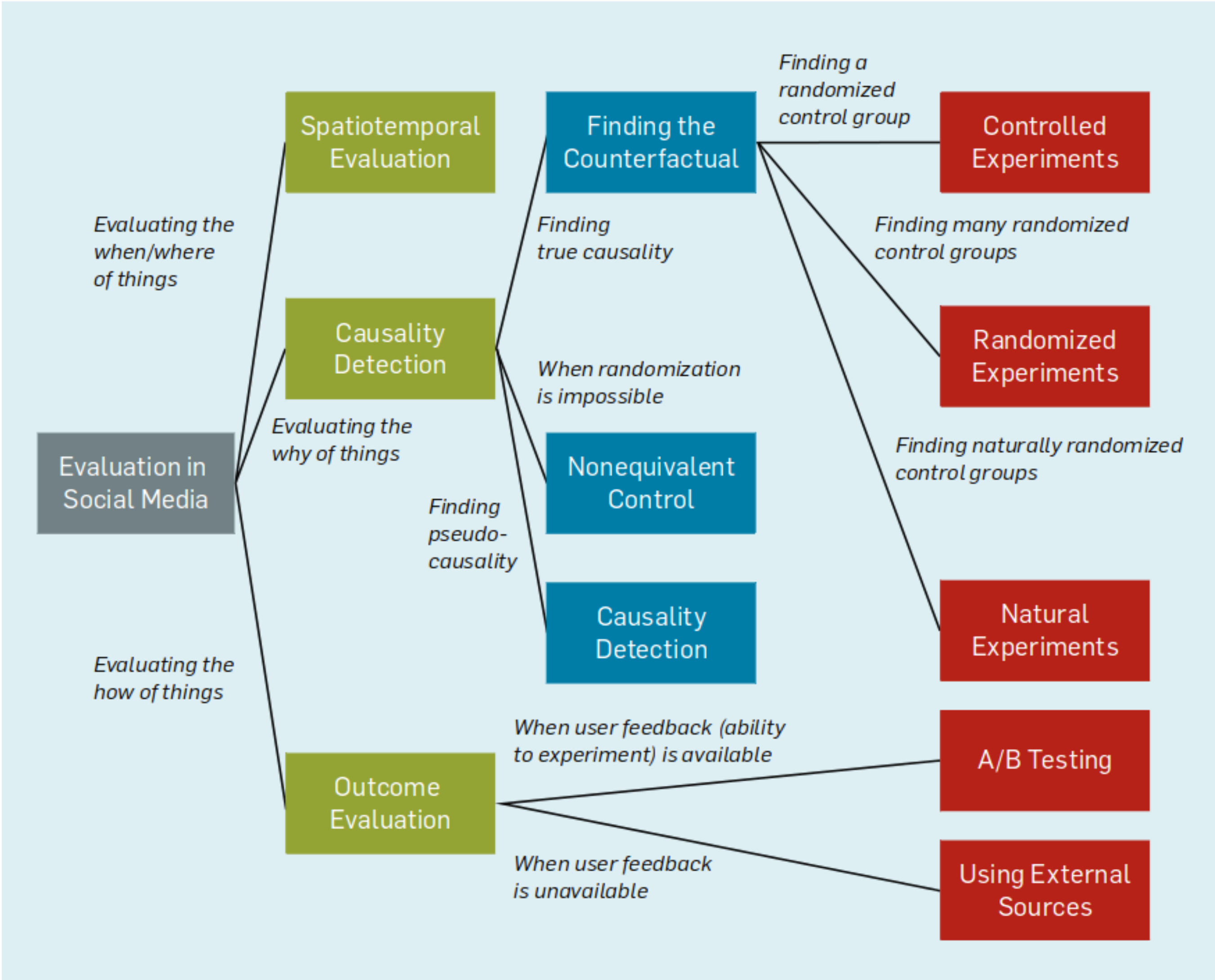
A Scientific American Guest Blog

1. Pop-Tarts before a hurricane (Walmart)
2. Higher crime, more Uber rides (Uber)
3. Typing with proper capitalization indicates creditworthiness (A financial services startup)
4. **Users of the Chrome and Firefox browsers make better employees (A HR firm over Xerox data)**
8. **Female-named hurricanes are more deadly (University Researchers)**

...

Yes, they are bizarre, but are they true?

Evaluation without Ground Truth



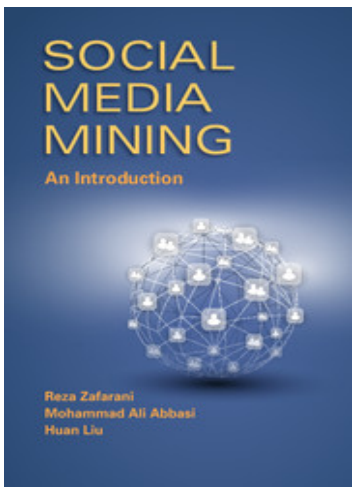
The CACM article is in both English and Chinese at dl.acm.org

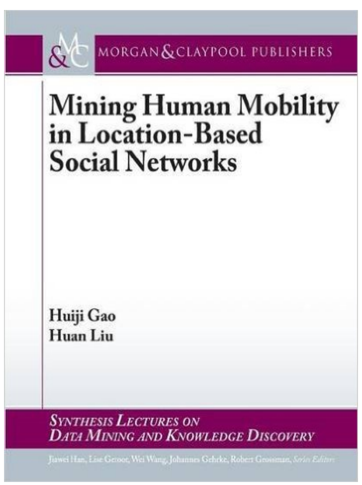
More Challenges Ahead

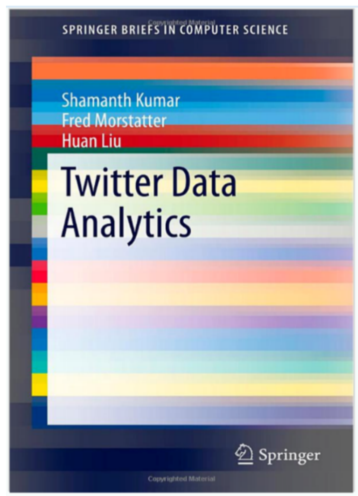
- Estimating the impact of an event
  - E.g., not all misinformation is catastrophic
- Predicting the future not the past
  - Are they two sides of the same coin?
    - Predicting general election result with Twitter data?
- Automating measures to replace crowdsourcing evaluation
  - Problems with evaluation methods involving AMT

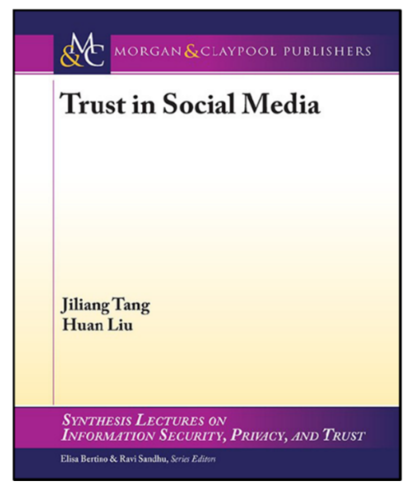
Repositories and Recent Books

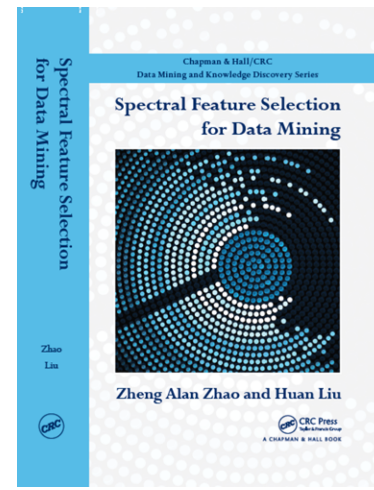
- scikit-feature – an open source feature selection repository in Python
- Social Computing Repository











Social Media Mining

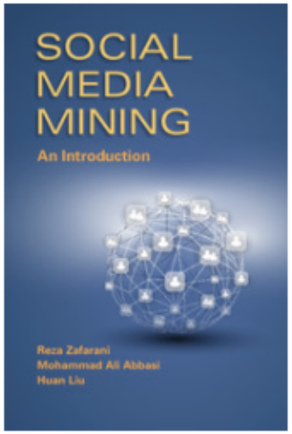
Home | Download Book | Slides/Tutorials | Table of Contents | Errata | How to Cite

Social Media Mining

An Introduction

A Textbook by Cambridge University Press

Reza Zafarani — *Syracuse University*
Mohammad Ali Abbasi — *Machine Zone*
Huan Liu — *Arizona State University*



PDF DOWNLOAD

Accessed 90,000+ times
from 160+ countries and 1200+ Universities

CAMBRIDGE UNIVERSITY PRESS | amazon.com | BARNES & NOBLE BOOKSELLERS | eBooks.com | TURING

> *The growth of social media over the last decade has revolutionized the way individuals interact and*

http://dmml.asu.edu/smm/

THANK YOU ALL & Conference Organizers

- for this opportunity to share our research
- Acknowledgments
  - Grants from NSF, ONR, ARO, among others
  - DMML members and project leaders
  - Many Collaborators

More information by searching for “Huan Liu” or at http://www.public.asu.edu/~huanliu