

Evaluation Dilemmas in Social Media Research

Huan Liu



Thanks to Former and Current PhD Students of DMML

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

Evaluation Dilemmas

1. Understanding the understanding
 - How to understand 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
4. Gaps between Problem and Data
 - How to let data help solve our problem at hand

1. Understanding the Understanding

- How to measure interpretability of topics generated by machine learning?
- One way to circumvent this problem is to indirectly measure the performance of these learned topics
 - The higher the performance, the better
 - Is it about understanding?
 - It may not be so
 - Where can we find the best evaluator?
 - Human experts
- Is involving human experts in evaluation a *scalable* and *reproducible* solution?
- Hence, challenging to understand the understanding

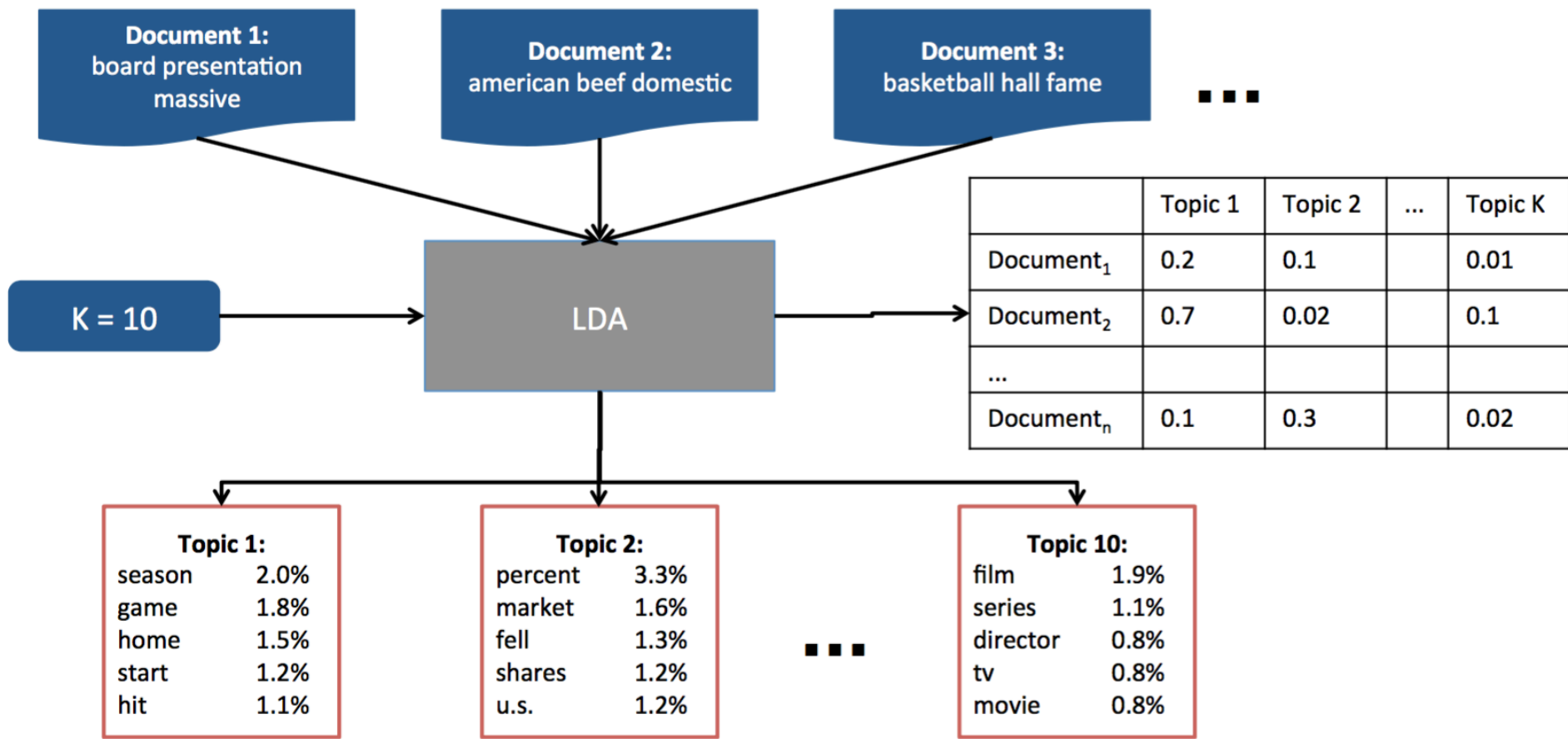
A Case of Big Text Data

- 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



How to Measure the Understanding?

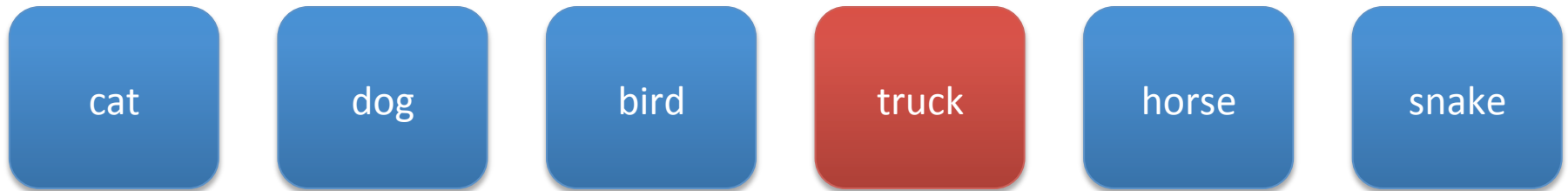
- How do we measure the interpretability of statistical topic models
- Experts are credible, but not a scalable solution, and crowdsourcing does not require experts, but has no expertise

A Measure of Model Precision

- Assesses Topic Interpretability
- Show a Turker 6 words in random order
 - Top 5 words from the topic
 - 1 “Intruded” word
 - Ask the Turker to identify the “Intruded” word

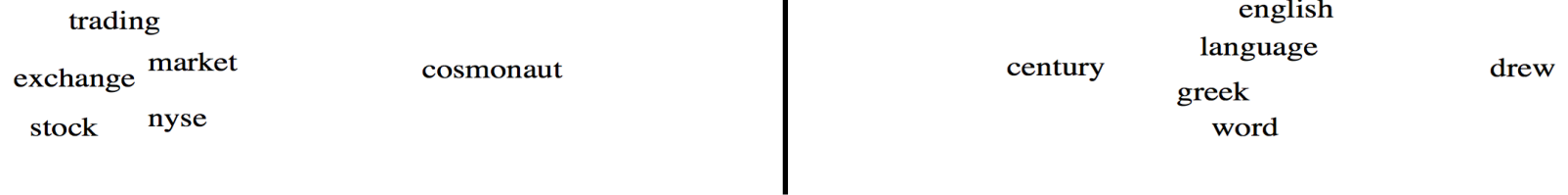
$$MP_{model,topic} = \# \text{ Correct Guesses } / \text{ Total } \# \text{ Guesses}$$

Topic i :



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)



Model Precision measures the distance *from* the intruded word, not *within* the topic

Real-World Examples

Top 5 Words	Intruded Word
design, use, based, develop, can	accompany
news, state, network, april, day	bigeast.com
family, actress, life, -year-old, star	megan's
maps, resource, visual, manifestation, seem	can

Another Solution

- 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

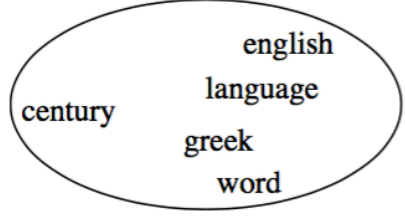
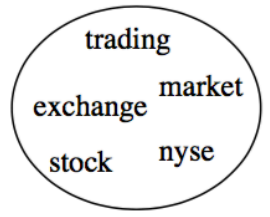
$$MPCT_k^m = H(p_{turk}(\mathbf{w}_{k,1}^m), \dots, p_{turk}(\mathbf{w}_{k,5}^m))$$



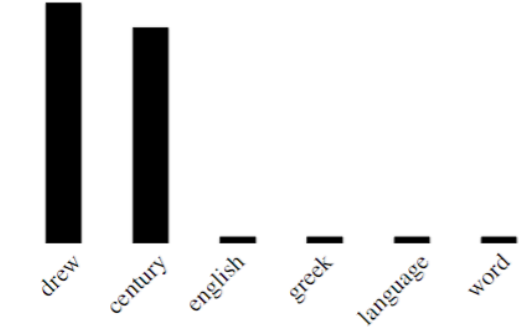
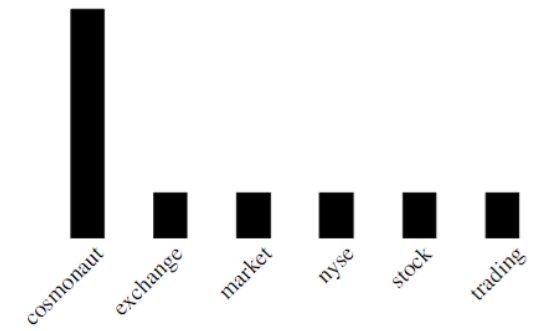
An Comparative Example

trading
 exchange market
 stock nyse
 cosmonaut

english
 language
 greek word
 century drew



Model Precision



Model Precision
Choose Two

News Corpus

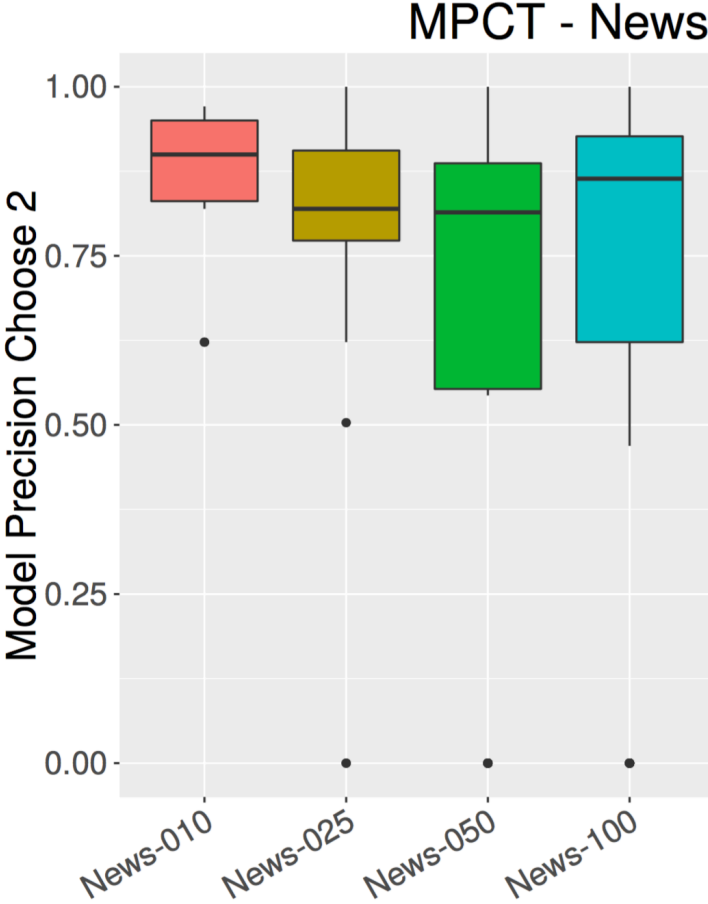
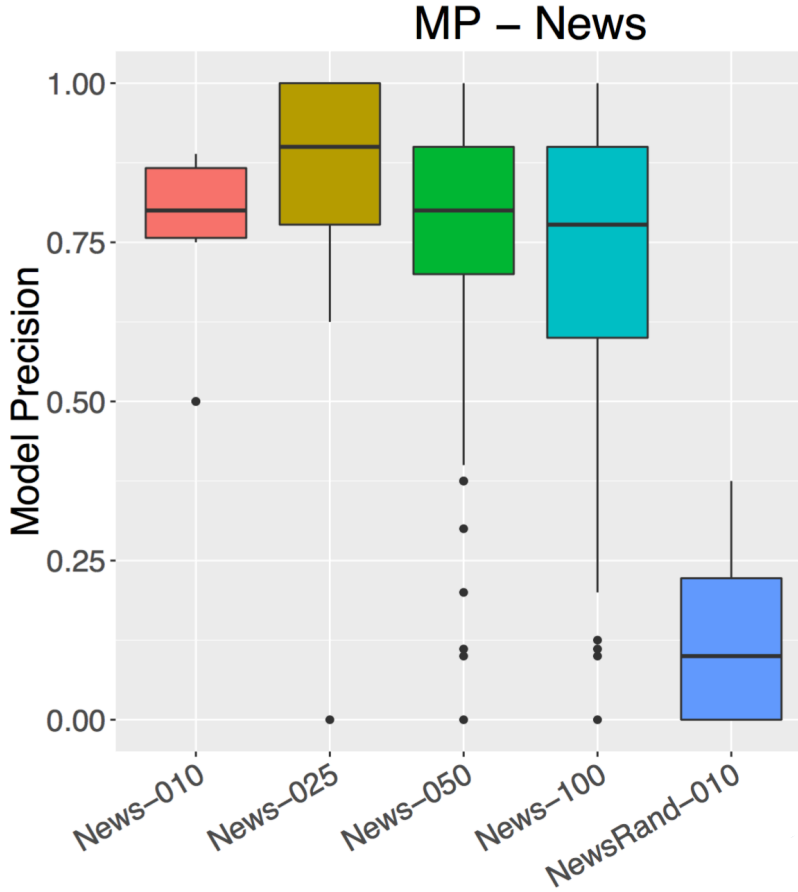
Yahoo! News Dataset

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

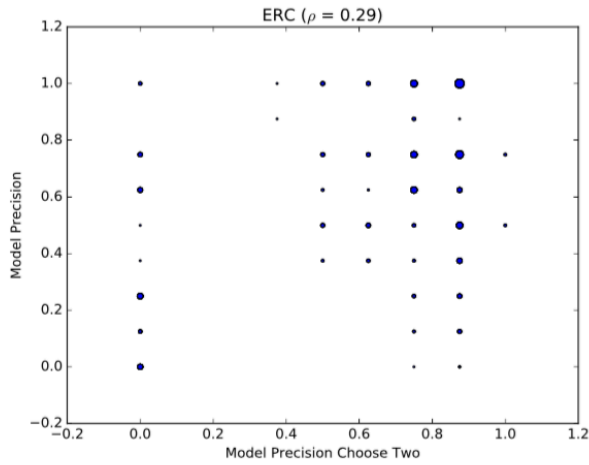
Name	Dataset	Strategy	Topics
News-010	News	LDA	10
News-025	News	LDA	25
News-050	News	LDA	50
News-100	News	LDA	100

Experiment: News Corpus

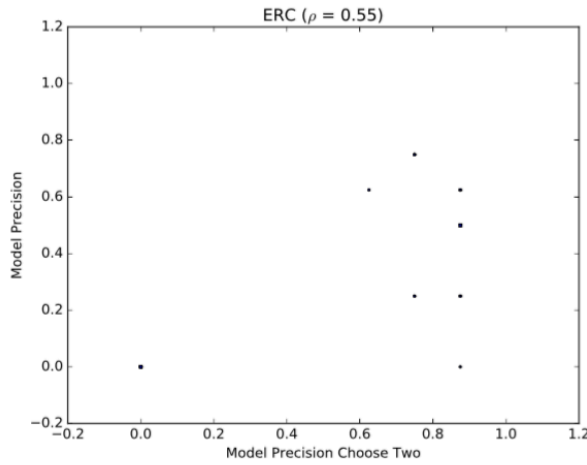
- Yahoo! News, Run with K = 10, 25, 50, 100.
- “Random” Topics



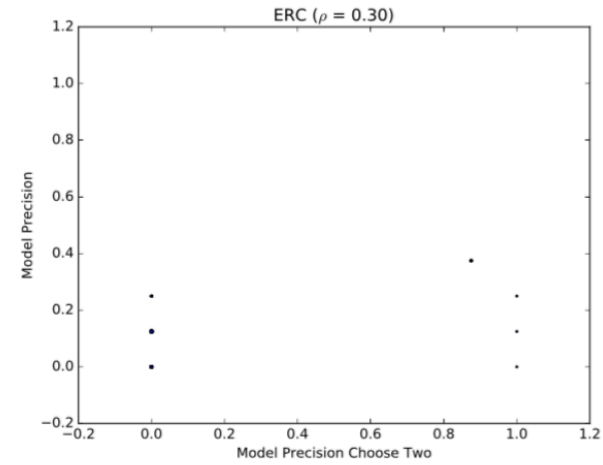
MP vs MPCT



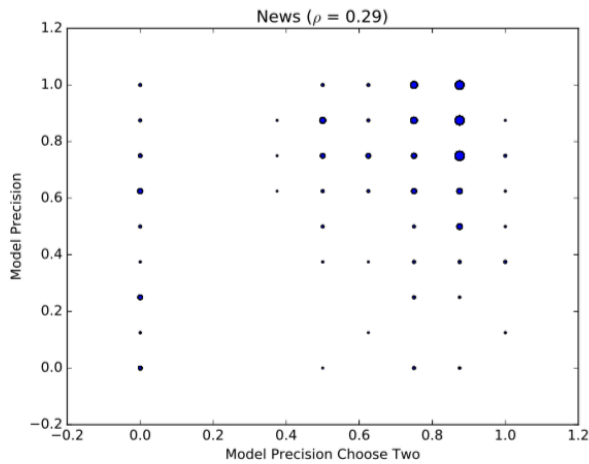
(a) ERC-*



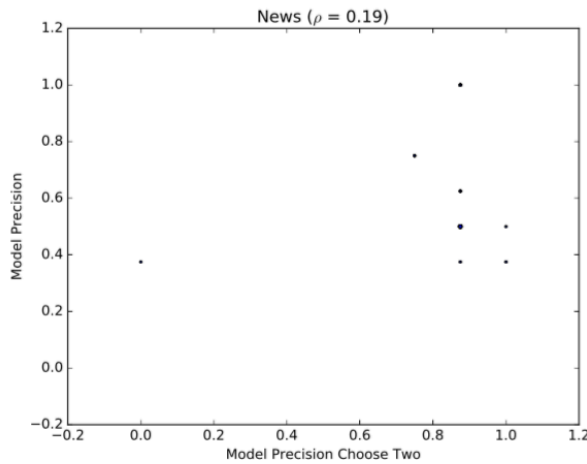
(b) ERCSanitySH-010



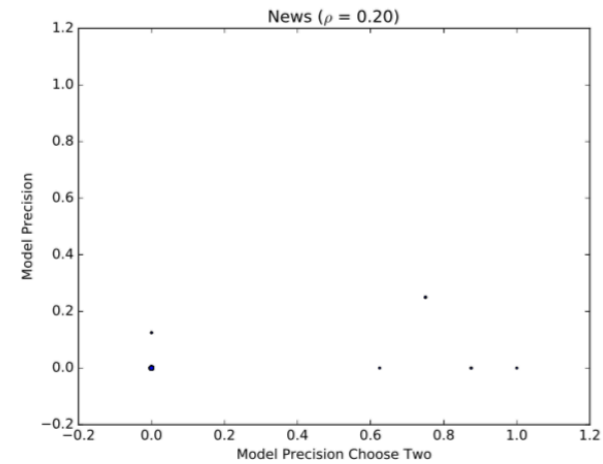
(c) ERCRand-010



(d) News-*



(e) NewsSanityS-010



(f) NewsRand-010

Can MPCT Replace MP?

No, it seems not. Why not ?

0 0 | 1 0
0 1 | 1 1

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

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 ($\rho = 0.29$)
- Automated measures could be explored to expedite the process of finding quality topics
- Topics and scripts: <http://bit.ly/mpchoose2>

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?

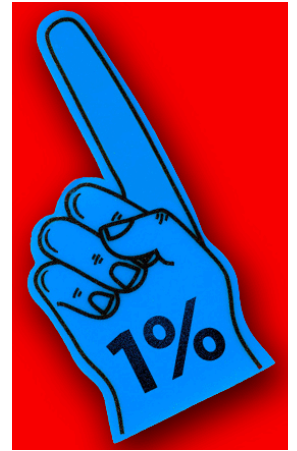
Twitter

- 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.



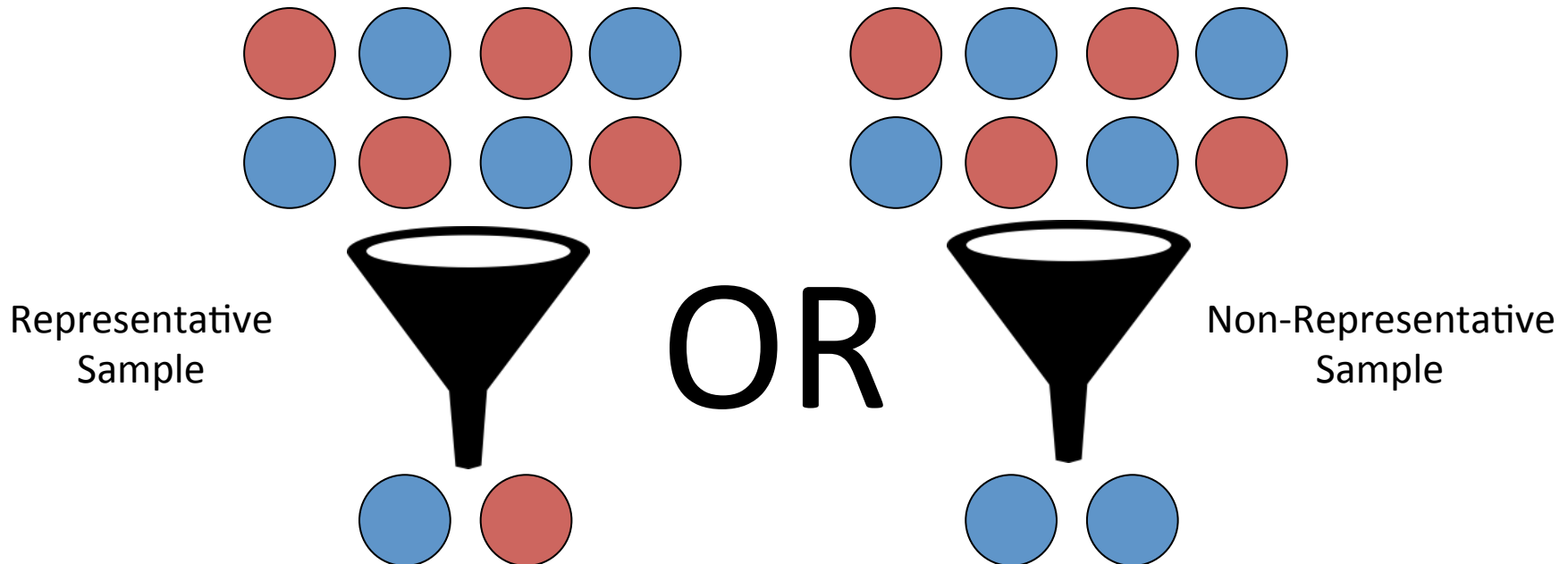
Why Twitter?

- 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 enough 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

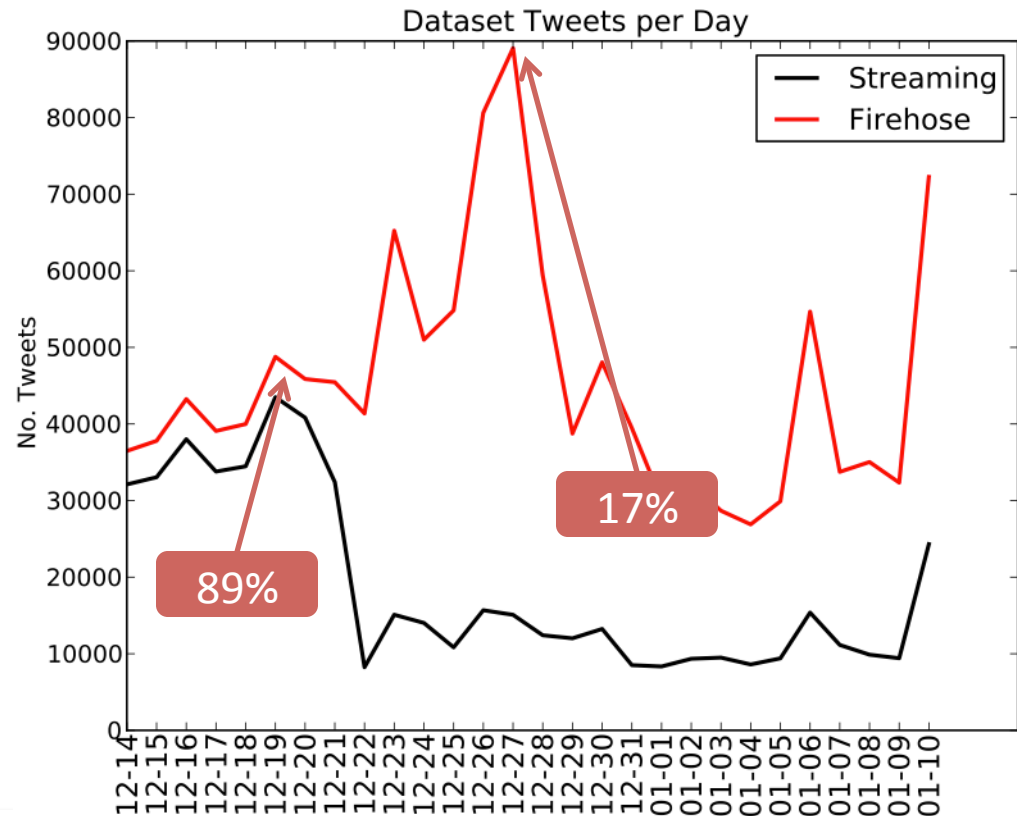
- Studying Arab Spring activity in Syria

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%.



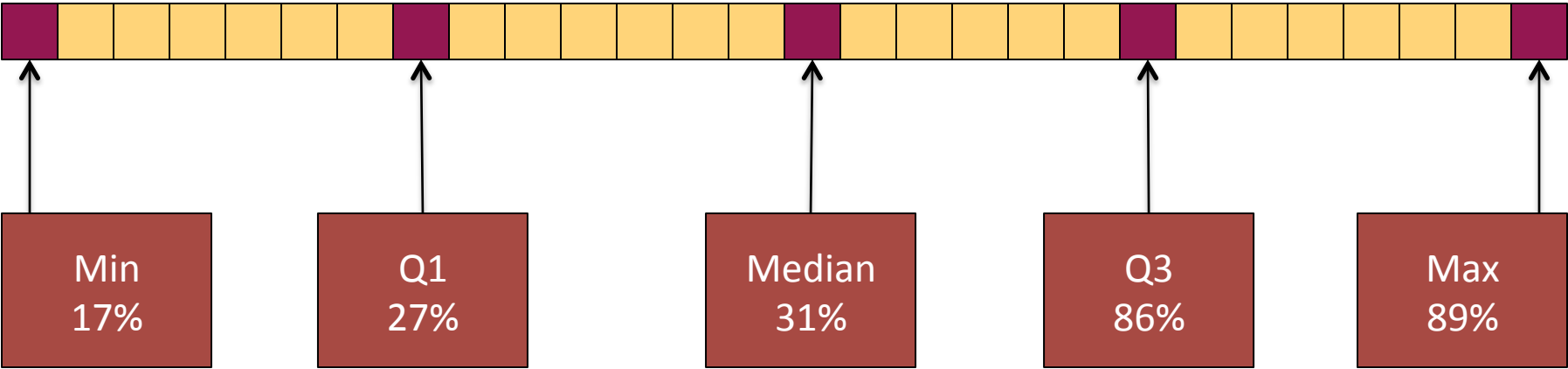
Analysis Choices and An Evaluation Challenge

- Compare facets of the tweet data from Streaming API and Firehose.
 - Hashtags
 - LDA Topics
 - Network Topology
 - Geographic Distribution



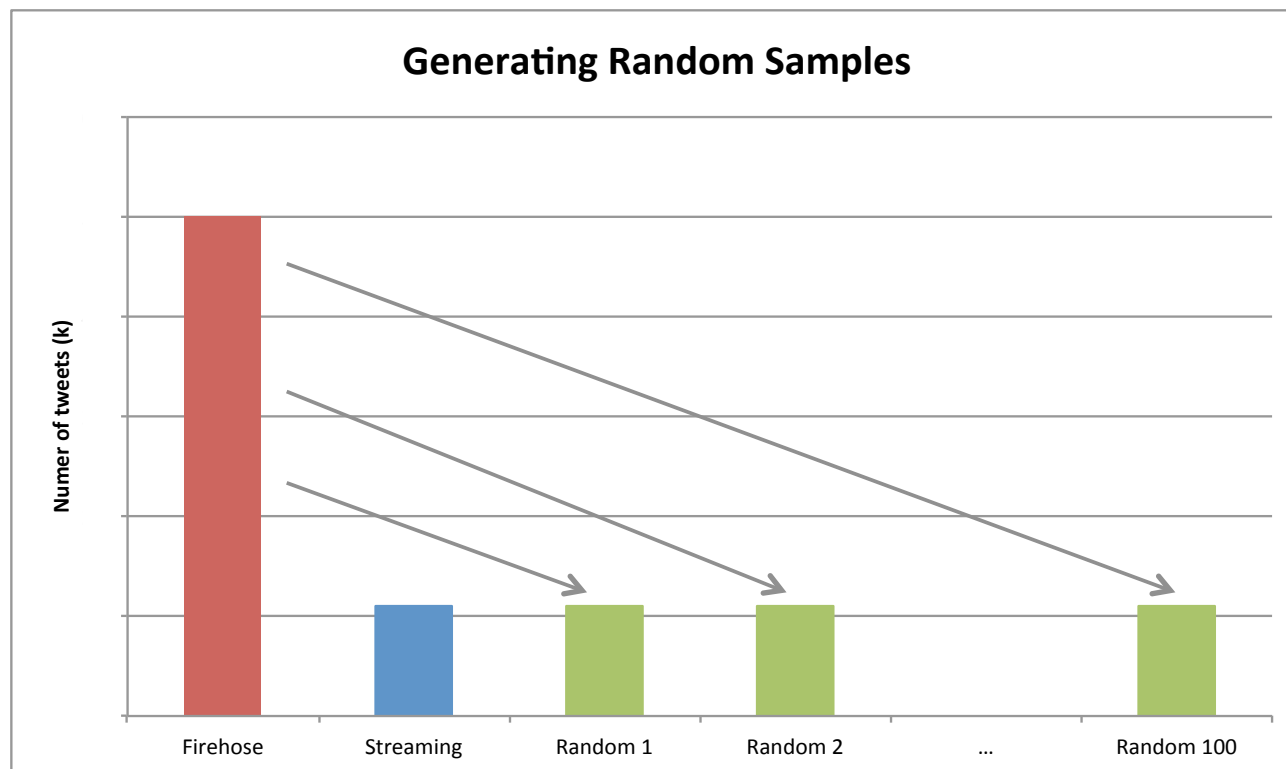
Days of Interest

Coverage →



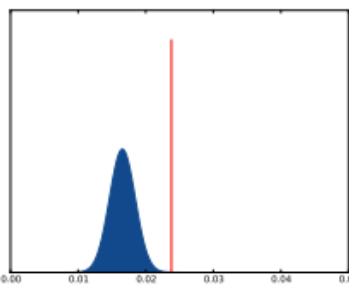
Verification

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

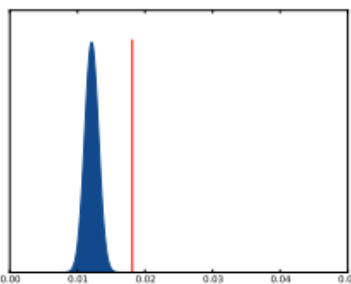


Comparison with Random Samples

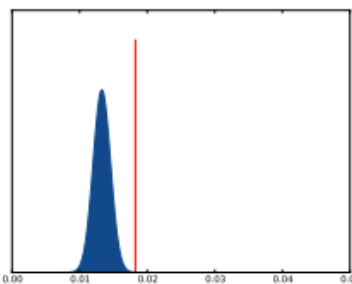
Is Streaming API data biased or not?



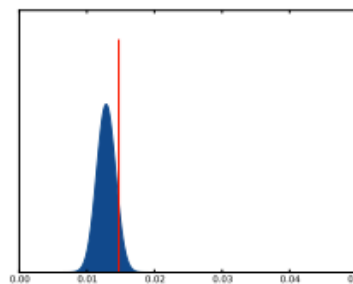
(a) Min. $S = 0.024$,
 $\hat{\mu} = 0.017$,
 $\hat{\sigma} = 0.002$,
 $z = 3.500$.



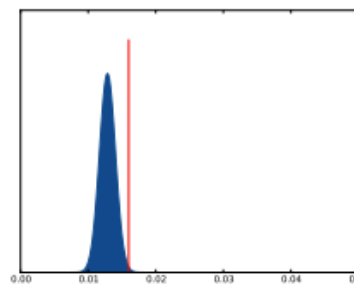
(b) Q1. $S = 0.018$,
 $\hat{\mu} = 0.012$,
 $\hat{\sigma} = 0.001$,
 $z = 6.000$.



(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?

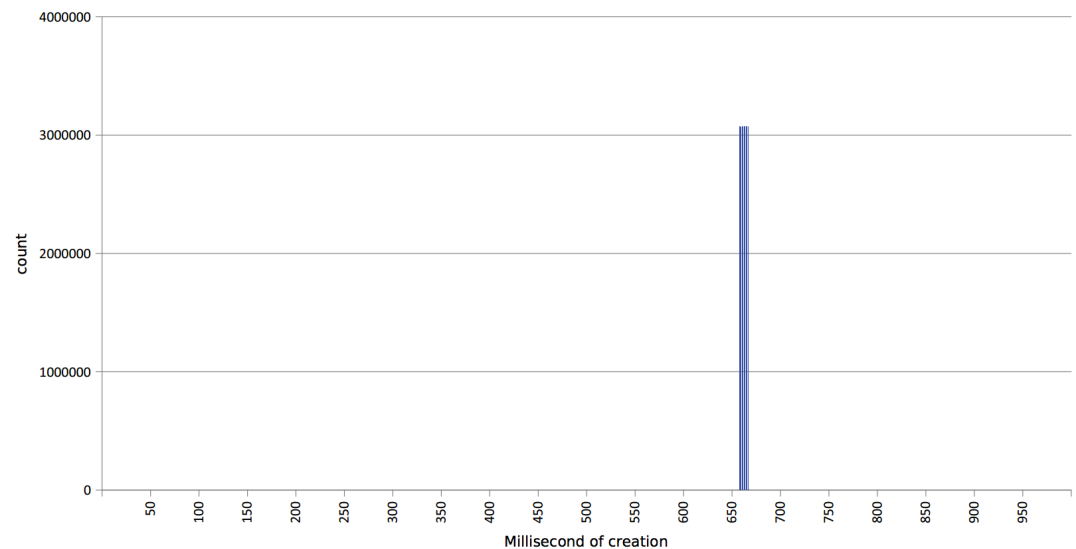
- How can researchers use the previous results to deal with bias in their own data?
- **Lesson:** There could exist bias
- **Challenge 1:** Need to find out if there is bias without Firehose
- **Challenge 2:** Collect more data to minimize bias

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 time periods in the Streaming API

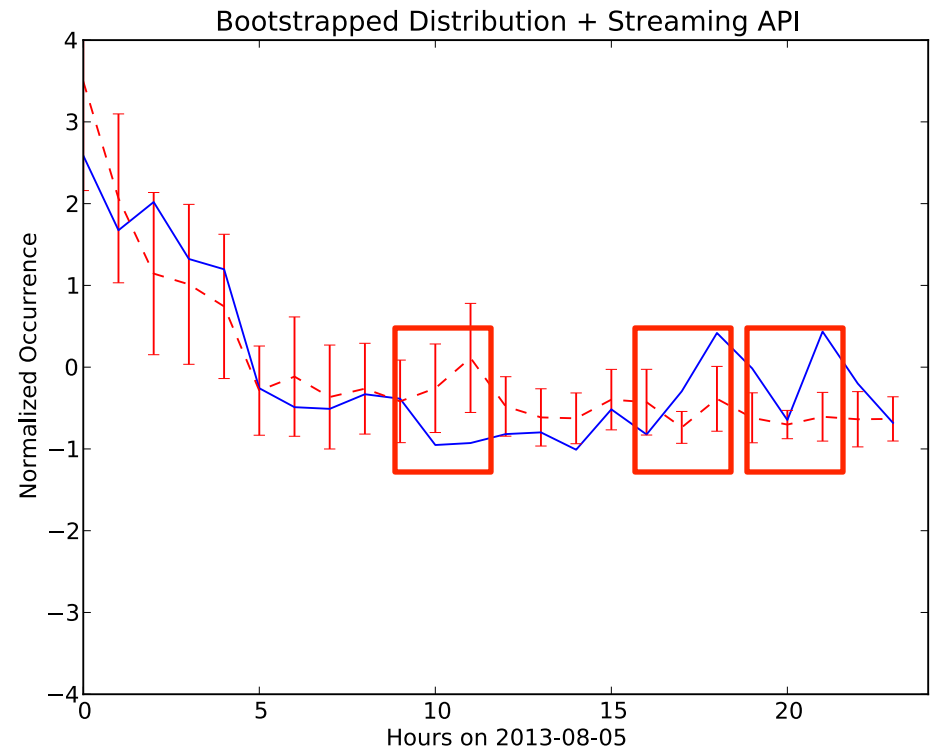
Twitter's Sample API

- 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



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



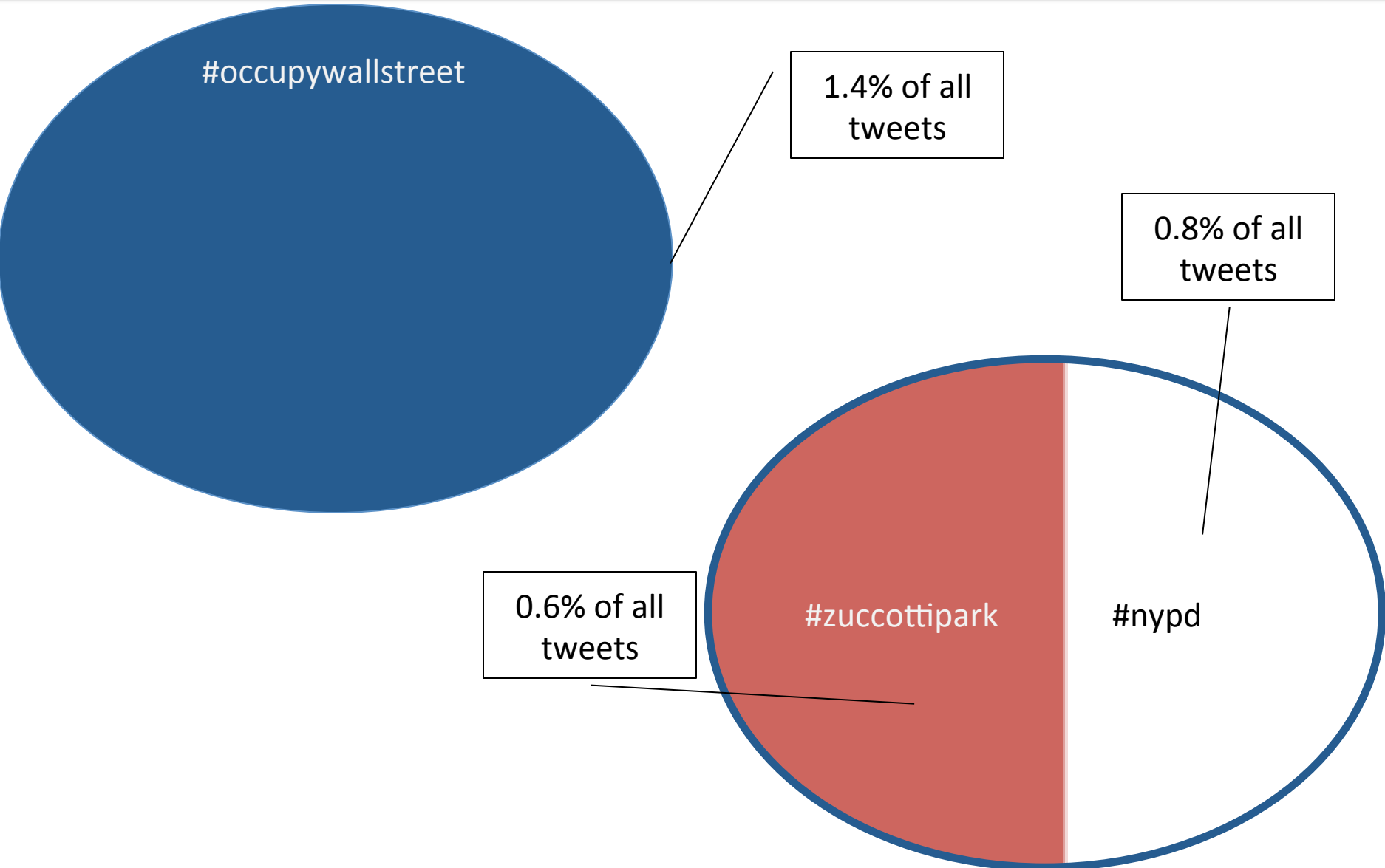
Takeaways

- 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%

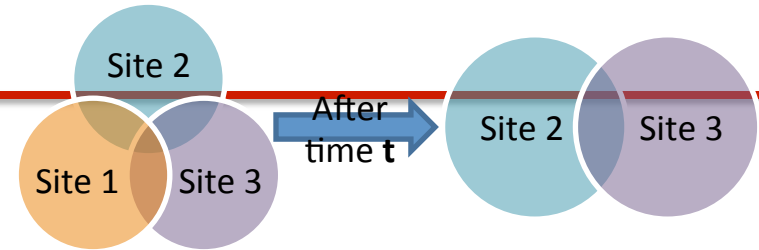
3. When-to-Stop Dilemma

- Collecting data forever vs. having credible patterns
 - How much data vs. how credible
- *A case study*: 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

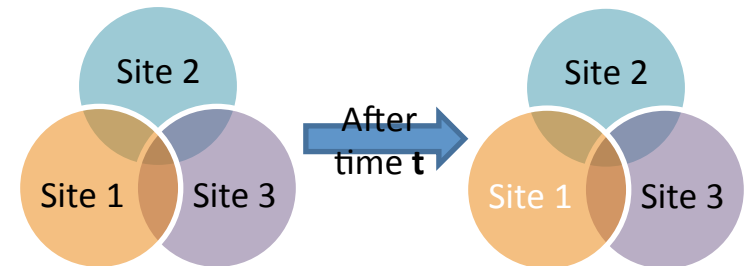
- **Site Migration**

- 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



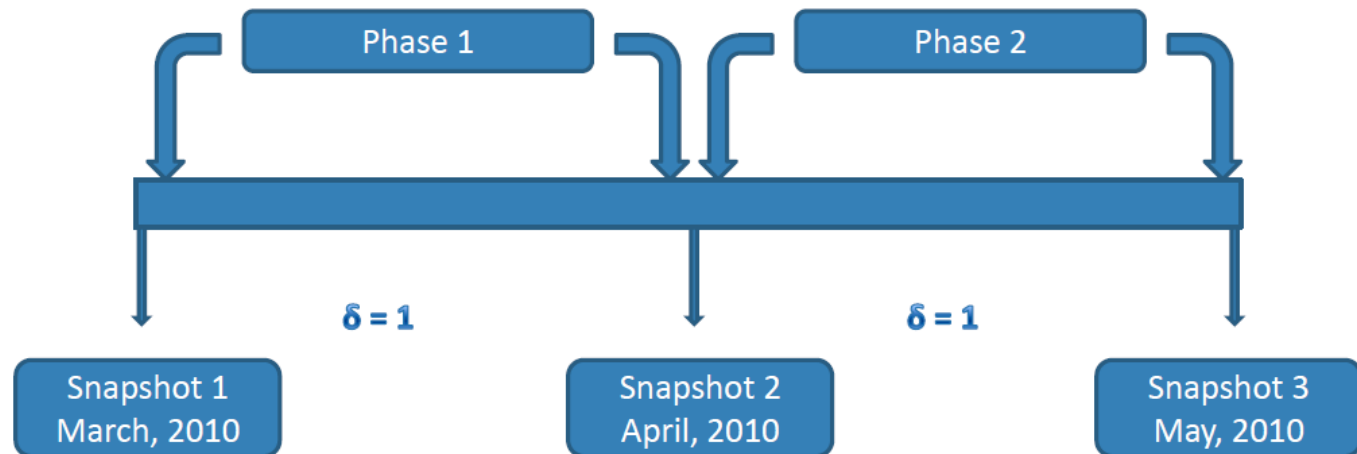
- **Attention Migration**

- 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

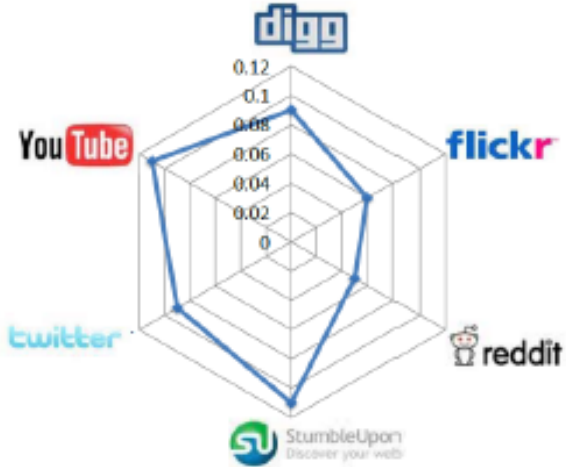


Obtaining User Migration Patterns

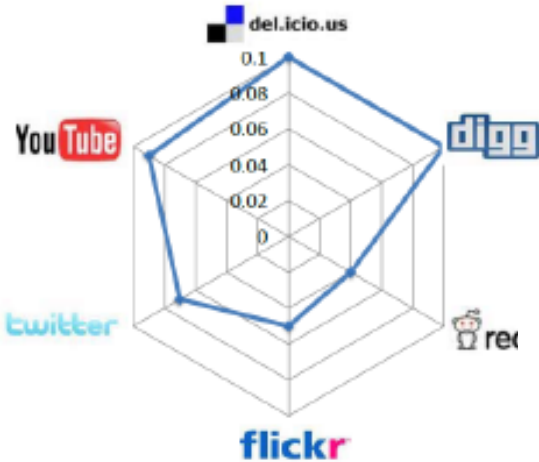
- Goal: Identifying trends of attention migration of users across the two phases of the collected data.
- Process



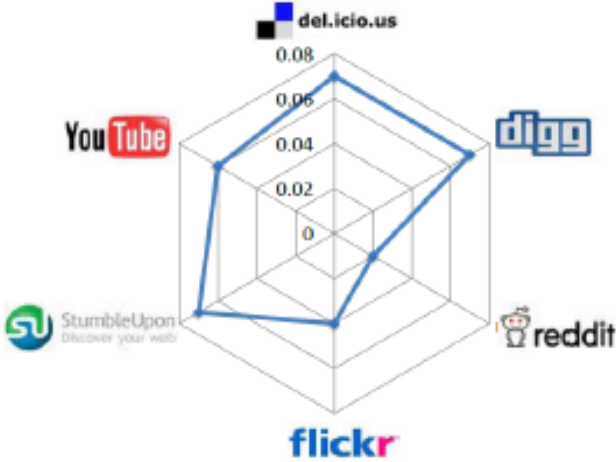
Patterns from Observation



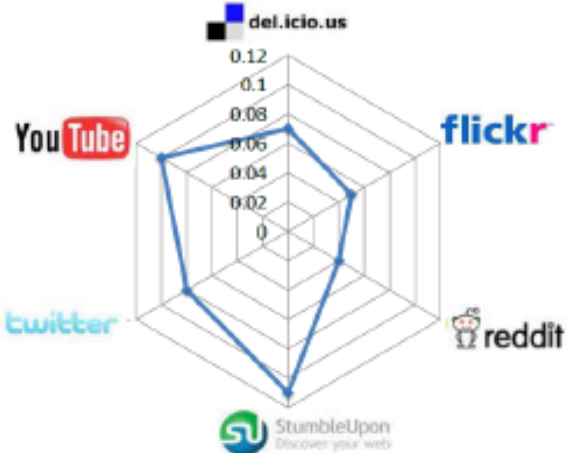
(a) Delicious



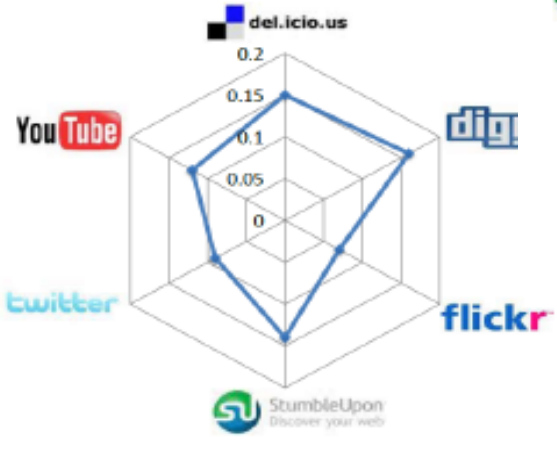
(e) StumbleUpon



(f) Twitter



(b) Digg



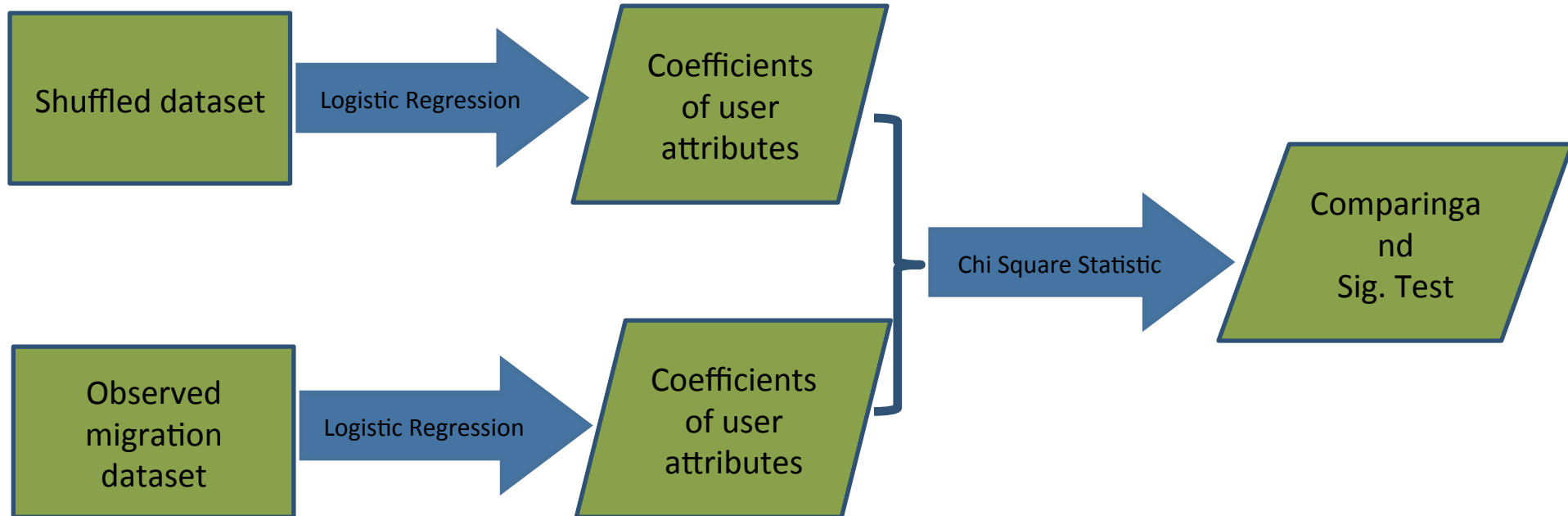
(d) Reddit

An Evaluation Challenge

- Important to know if they are valid or not
 - If yes, we investigate further how we use patterns for prevention or promotion
 - If not, why not? And what can we do?
- The challenge to evaluating migration patterns is we don't have ground truth
- How to address the challenge?
 - User study or AMT?

Evaluating Patterns' Validity: A Significance Test

- Null Hypothesis: *Migration of individuals is a **random** process*
 - Generating another similar dataset for comparison
 - Potential migrating population includes overlapping users from Phase 1 and Phase 2
 - Shuffled datasets are generated by picking random active users from the potential migrating population
 - The number of random users selected for each dataset is the same as the real migrating population



Can we now 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

Summary

- Mitigating or promoting migration by targeting high net-worth individuals
 - Identifying users with high value to the network, e.g., high network activity, user activity, and external exposure
- Social media migration is first studied in this work
- Migration patterns can be evaluated without test data

4. Gaps between Problems and Data

- Sometimes, gaps between interesting problems and data at hand seem insurmountable
- For example, how can we answer questions like ***“Is distrust the negation of trust”?***
 - There is no labeled data to answer this question
 - When our data at hand cannot be directly used to answer the question, what should we do?
- The power of reduction
 - Rewrite the problem to one that can be answered using data

“9 Bizarre and Surprising Insights from Data Science”

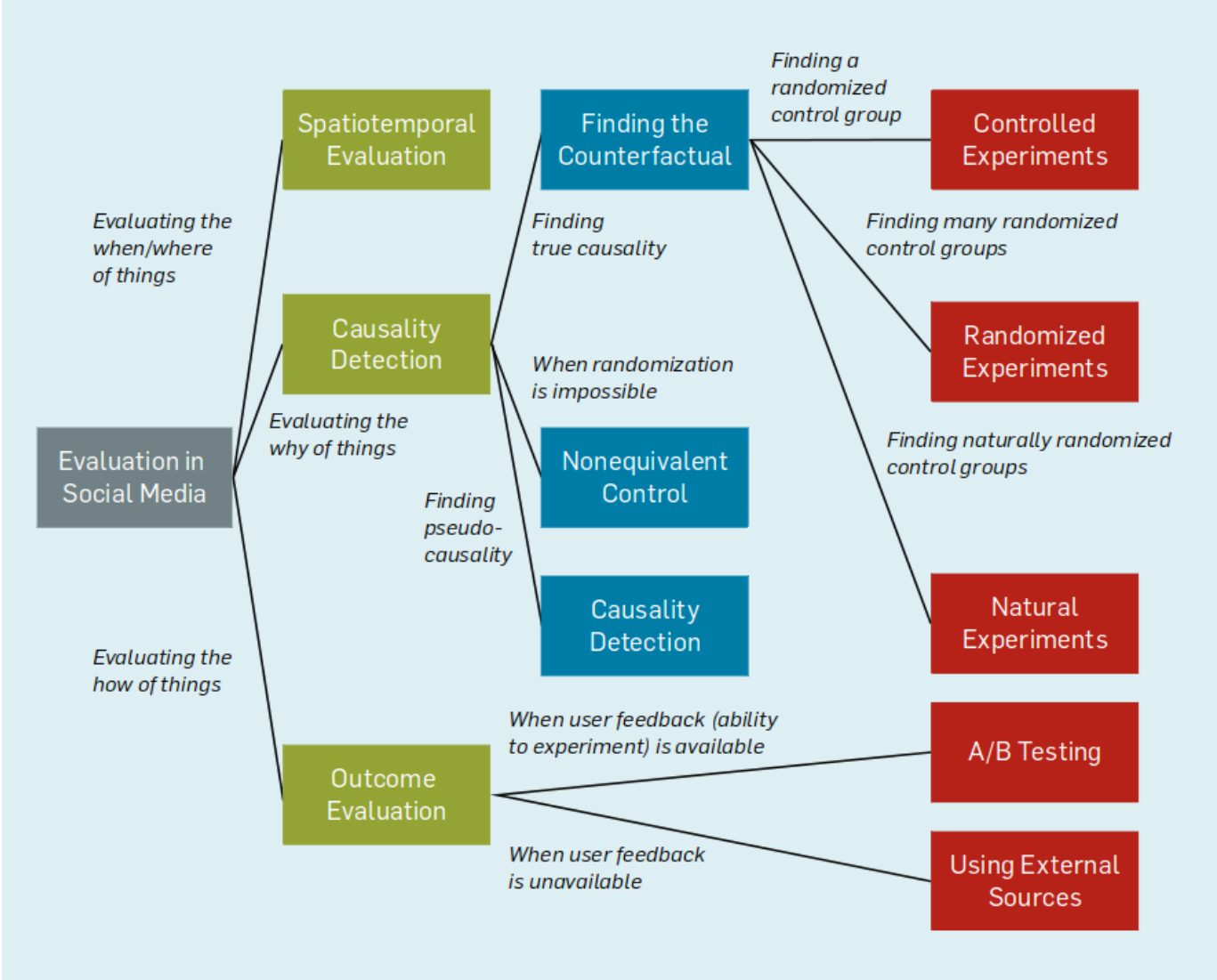
A Scientific American Guest Blog

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



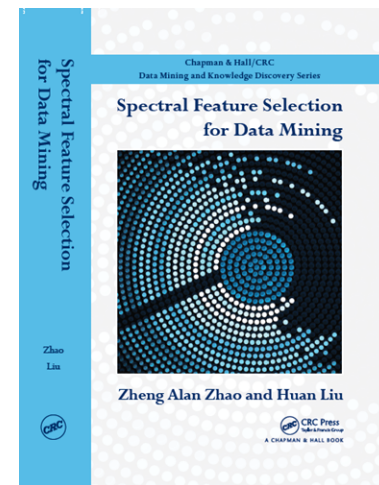
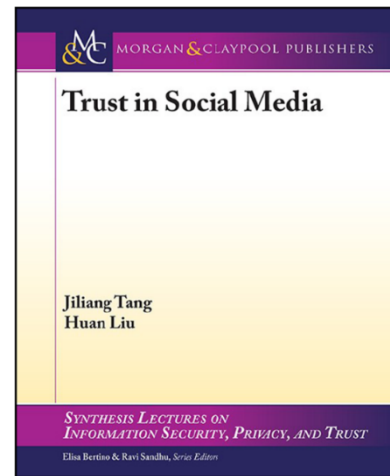
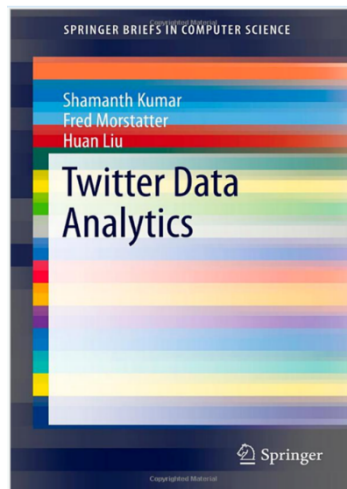
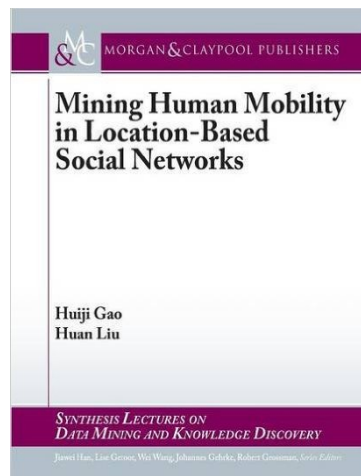
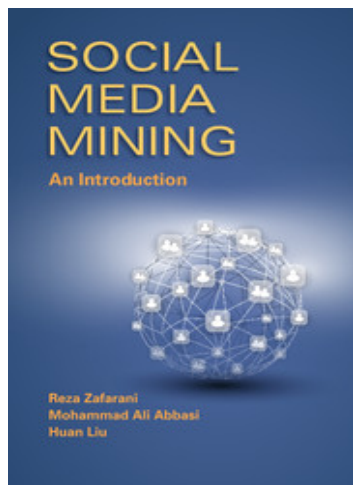
The CACM article is in both English and Chinese at dl.acm.org

More Challenges ahead

- Hakuna Matata?
- 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 An Introduction

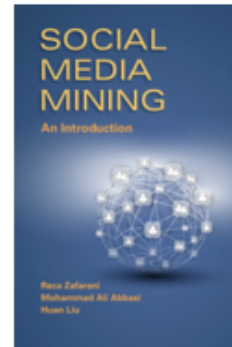
A Textbook by Cambridge University Press

Reza Zafarani
Mohammad Ali Abbasi
Huan Liu

Syracuse University
Quid
Arizona State University



Accessed 75,000+ times
from 150+ countries and 900+ Universities



The growth of social media over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social media. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research, novel algorithms, and tool development. Social Media Mining integrates social media, social network analysis, and data mining to provide a convenient and coherent platform for students, practitioners, researchers, and project managers to understand the basics and potentials of social media mining. It introduces the unique problems arising from social media data and presents fundamental concepts, emerging issues, and effective algorithms for network analysis and data mining. Suitable for use in advanced undergraduate and beginning graduate courses as well as professional short courses, the text contains exercises of different degrees of difficulty that improve understanding and help apply concepts, principles, and methods in various scenarios of social media mining.

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

THANK YOU and ROCLING2016

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

More information is at

<http://www.public.asu.edu/~huanliu>