

# From AI<sup>K</sup> to AI<sup>D</sup>: Acquiring Social Media Intelligence via `Big` Data

Huan Liu



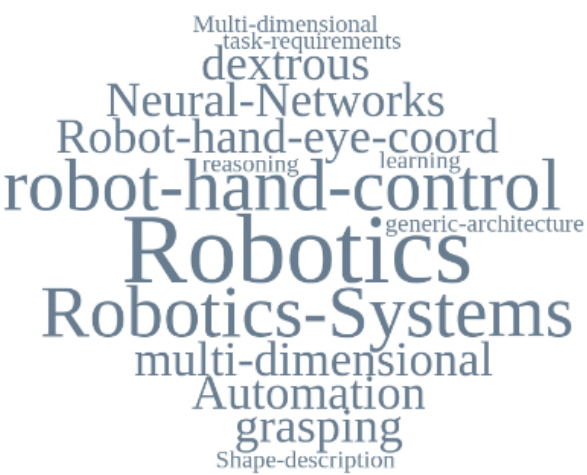
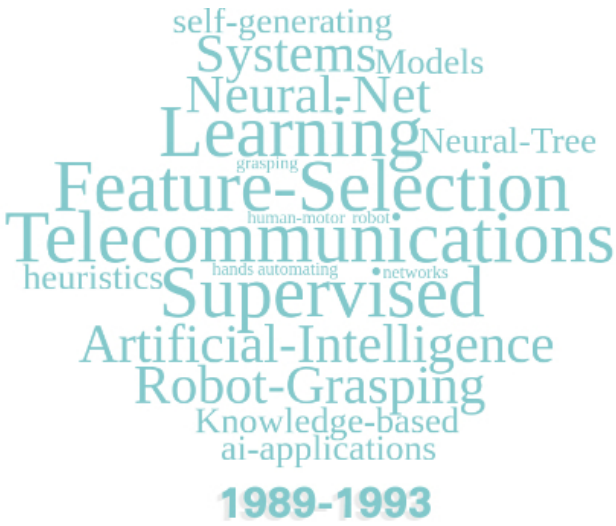
# Thanks to Former & Current PhD Students

- Reza Zafarani, 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, Clari
- 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, USC ISI
- **Christophe Faucon**
- **Isaac Jones**
- **Suhas Ranganath**
- **Suhang Wang**
- **Tahora Nazer**
- **Jundong Li**
- **Liang Wu**
- **Ghazaleh Beigi**
- **Kai Shu**
- **Justin Sampson**



# A Tortuous but Fortuitous Path to Social Computing



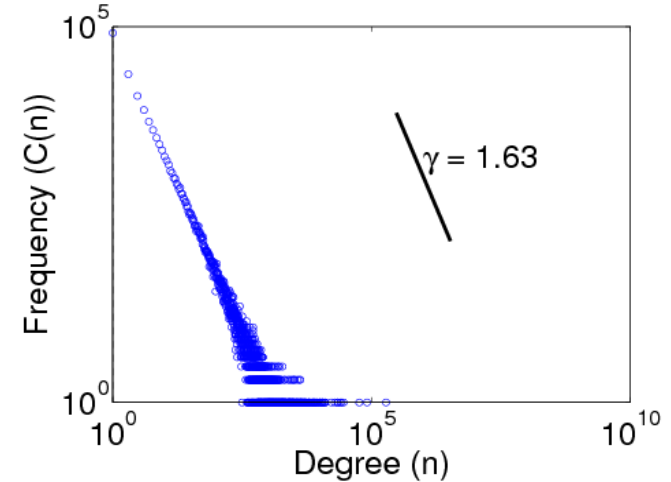
# From AI<sup>K</sup> to AI<sup>D</sup>

- “**K**nowledge is Power”: AI was then solely about **K**
  - Expert Systems or Rule-based Systems
    - “Intelligence is ten million rules.”
  - Knowledge-based Systems (Cyc)
- “**D**ata is the New Oil”: AI is now hyped up with **D**
  - Big data is ubiquitous
  - CS, Statistics, Information Science → Data Science
- Recent surge of AI is powered by Data
  - Machine Learning (including Deep Learning)
  - For any learning algorithm to work, data is key

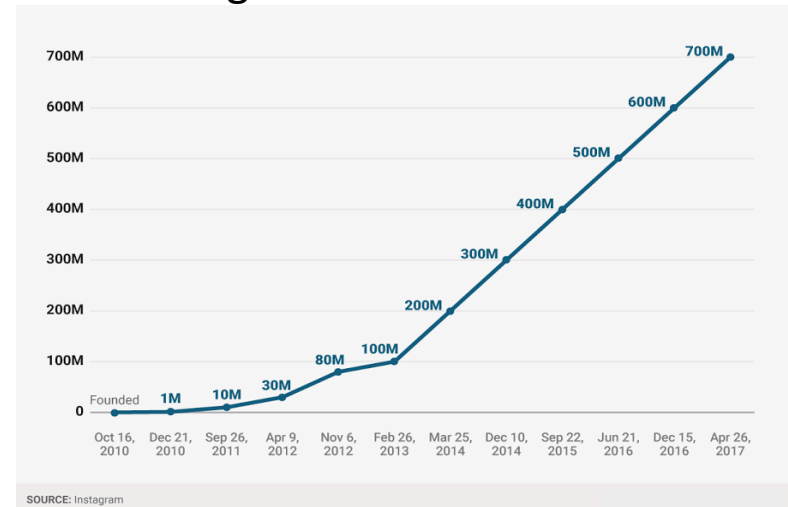
# Big Social Media Data

- **Twitter**
  - 300 million users
  - 500 million tweets / day
  - 1% (5 million) released for research
- **Facebook**
  - 2 billion users
  - 422 million updates / day
  - 196 million photos / day
- **Instagram**
  - 700 million users
  - 80 million photos / day

Facebook Degree Distribution

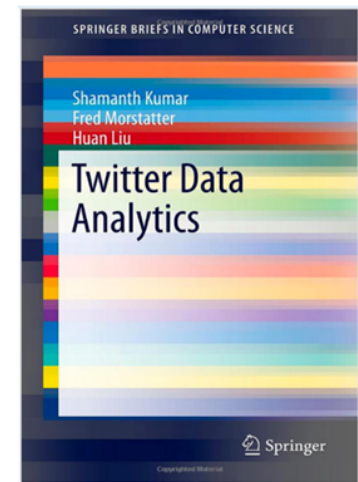
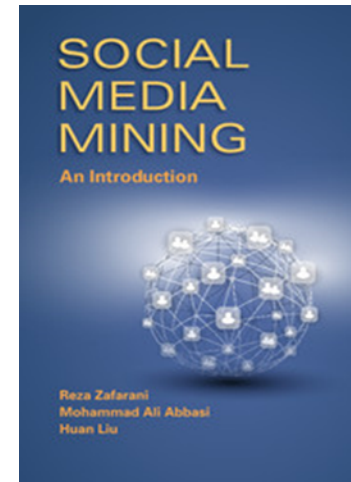


Instagram Users over Time



# Discovering Social Media Intelligence

- Graph Theories
- Network Measures and Models
- Data Mining, **NLP**, and **Visual Analytics**
- Community Detection and Analysis
- Information Diffusion
- Influence and Homophily
- Recommender Systems
- Behavior Analytics
  - **Sentiment Analysis**



# Some Challenges in Acquiring SM Intelligence

---

- Social media data seems really big, but why are we often still short of data?
  - How can we make data *`bigger`*?
- Data is power, so it can produce any result
  - Can we *algorithmically* evaluate the results from big data?
- We don't know what we don't know
  - How can we know if our result of social media analysis is of any value?

# Making Big Data “Bigger”

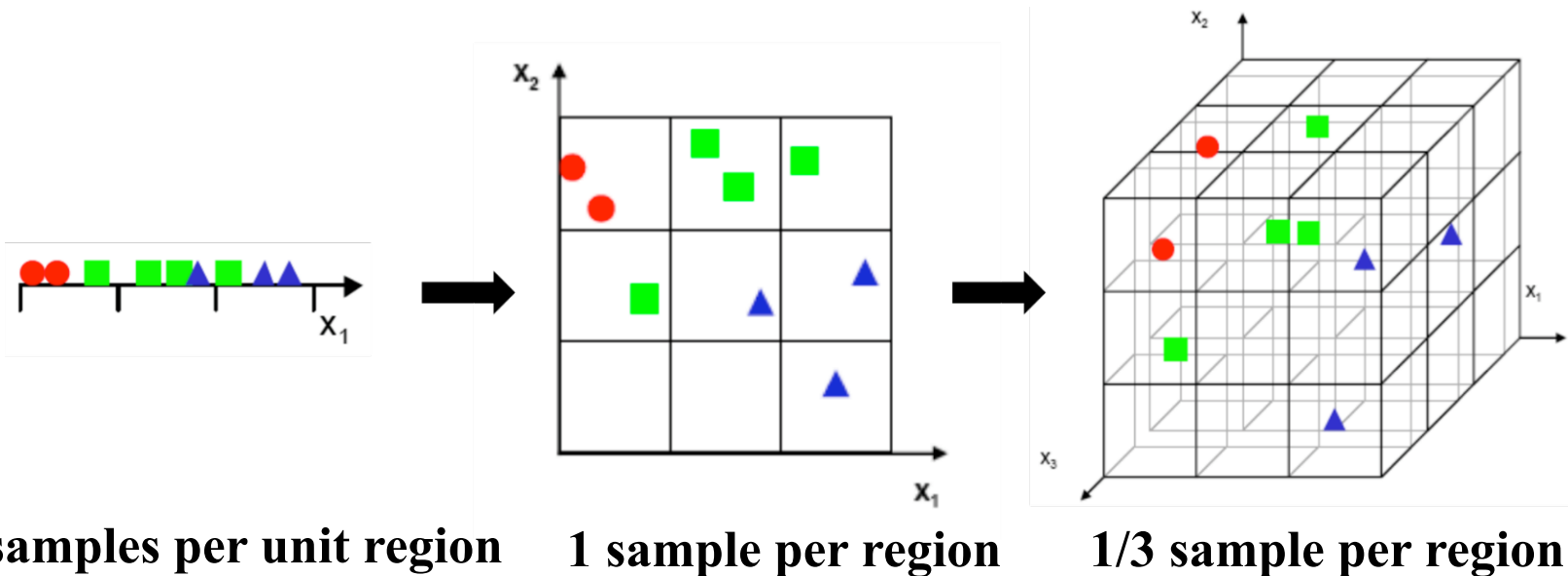
---

- What is big data?
  - A conventional answer is 4Vs
  - A practitioner’s answer is more nuanced
- Big data can be actually *little* or *thin*
- For machine learning or data mining to work, ***the more data, the better***
  - Make little data bigger
  - Make thin data thicker



# Curse of Dimensionality: Required Samples

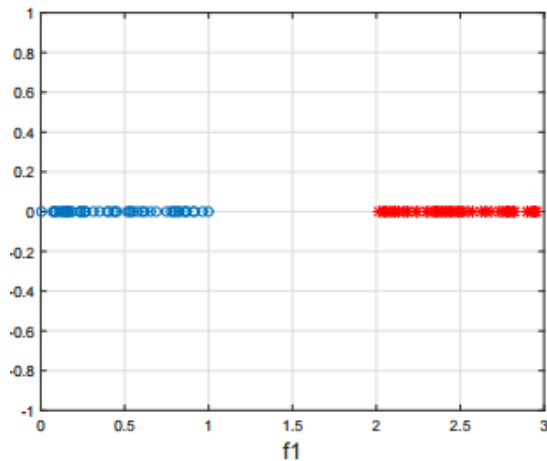
- Sparsity becomes exponentially worse as feature dimensionality increases
  - Conventional distance metric becomes ineffective as far and near neighbors have similar distances



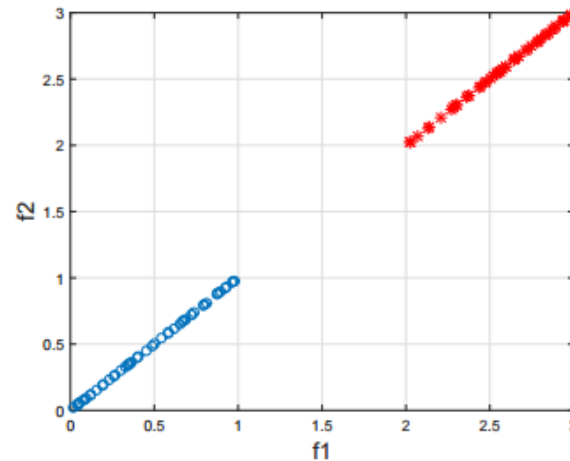
<http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/>

# Relevant, Redundant and Irrelevant Features

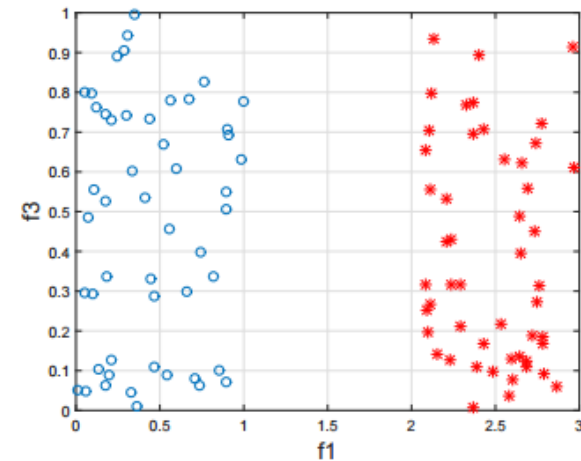
- Feature selection retains relevant features for learning and removes redundant or irrelevant ones
- For a binary classification task below,  $f_1$  is relevant,  $f_2$  is redundant given  $f_1$ , and  $f_3$  is irrelevant



(a) relevant feature  $f_1$



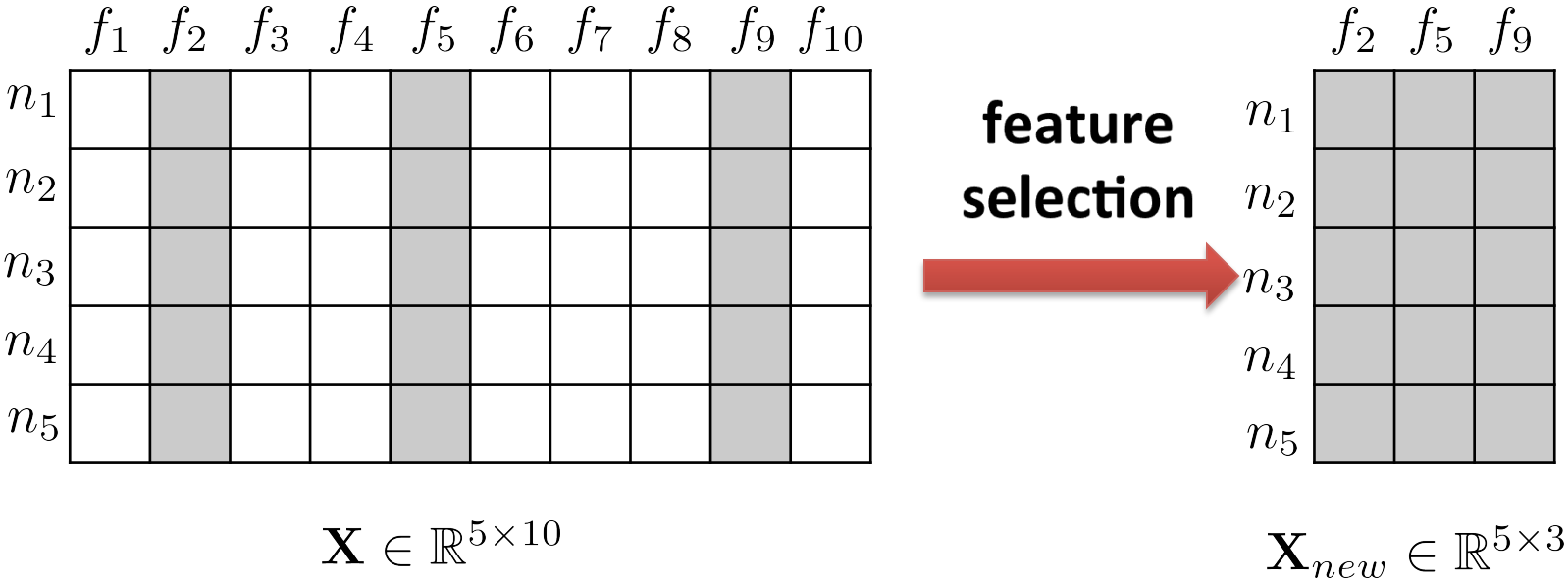
(b) redundant feature  $f_2$



(c) irrelevant feature  $f_3$

# Feature Selection

Feature selection selects an 'optimal' subset of relevant features from the original high-dimensional data given a certain criterion

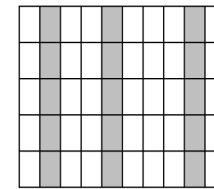


# Feature Selection and scikit-feature

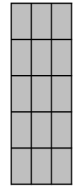
- Feature selection can make data 'bigger'

- Assuming all binary attribute values in our toy example

- Before FS,  $5/2^{10} = 5/1024$ ,  
after FS,  $5/2^3 = 5/8$



$$\mathbf{X} \in \mathbb{R}^{5 \times 10}$$



$$\mathbf{X}_{new} \in \mathbb{R}^{5 \times 3}$$

- Does FS always work?
  - Yes, for most high-d data
- Where can we find it?
- **scikit-feature**, an open-source repository in Python

5 Machine Learning Projects You Can No Longer Overlook, April

Apr 2017  
Silver Bug Blog

Previous post Next post

Like 253 Share 253 in Share 468 Tweet G+ 4

Share 64

Tags: Data Exploration, Deep Learning, Java, Machine Learning, Neural Networks, Overlook, Python, Scala, scikit-learn, Topic Modeling

It's about that time again... 5 more machine learning or machine learning-related projects you may not yet have heard of, but may want to consider checking out. Find tools for data exploration, topic modeling, high-level APIs, and feature selection herein.

2. scikit-feature

scikit-feature is an open-source feature selection repository in Python developed by Data Mining and Machine Learning Lab at Arizona State University. It is built upon one widely used machine learning package scikit-learn and two scientific computing packages Numpy and Scipy. scikit-feature contains around 40 popular feature selection algorithms, including traditional feature selection algorithms and some structural and streaming feature selection algorithms.

Open Data Innovation Summit  
12<sup>th</sup> & 13<sup>th</sup> June, 2017  
London  
innovation enterprise  
on argyle company  
VIEW EVENT

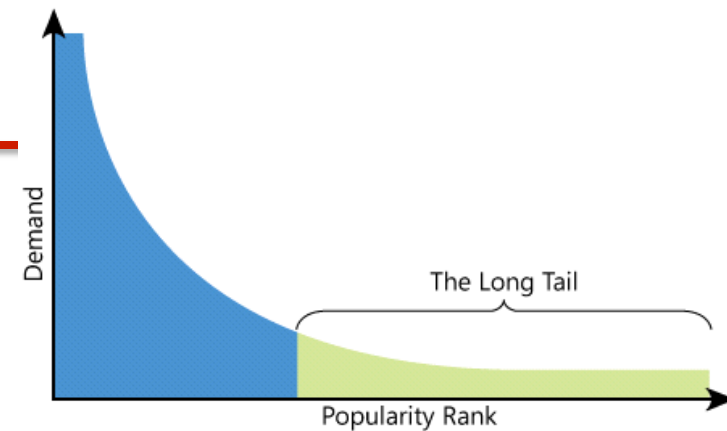
Open Data Innovation Summit  
London, Jun 12-13

## 2. scikit-feature

Though all methods of feature selection share the common goal of identifying redundant and irrelevant features, there are numerous algorithms for approaching these related problems -- this is an active area of research. In that regard, scikit-feature is for both practical feature selection and

# Making Thin Data Thicker

- Most people like many of us are in the long tail
  - Our data is thin or sparse
  - With little data, machine learning is powerless
- Social media data offers new opportunities
  - Multiple facets: posts, profile, linked information
  - Multiple platforms that offer different functions
- Two case studies
  - Feature selection using *social network* information
  - Connecting users *across* more than one social media site



# Making Sense of Big Data

---

- For big social-media data, we want to automatically get a sense of what it is
  - User needs, sentiment, opinions, behavior, and trends
- A big part of big data is TEXT
- NLP and text mining can help extract **topics** from text
- If these machine-learned topics are for human consumption, are they actually comprehensible?
  - How can comprehensibility be measured?

# Measuring Topic Interpretability

---

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

# Big Text Data

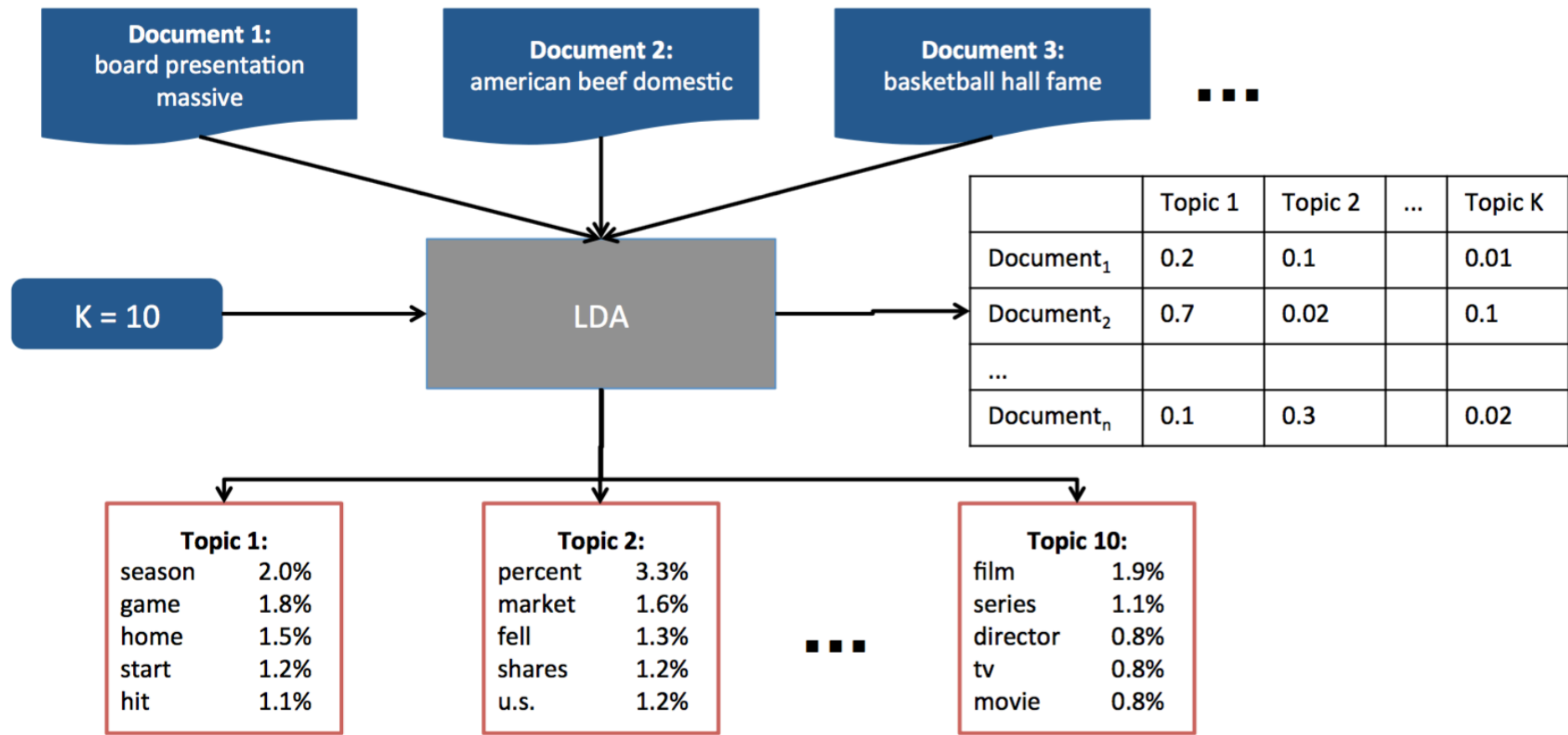
- Some example corpora:

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

- Too much data to read
- How can we begin to understand all of these large bodies of text data?



# Topic Models



# Measuring Interpretability

---

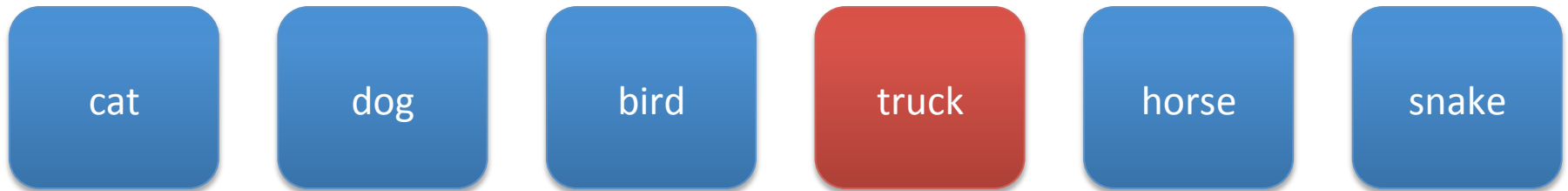
- How do we measure the interpretability of statistical topic models
- A dilemma
  - Experts are **credible**, but **not scalable**,
  - Crowdsourcing needs *no experts*, so **scalable**, but has *no expertise*, thus is **not credible**

# 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

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

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)

trading  
exchange market  
stock nyse

cosmonaut



century

english  
language  
greek  
word

drew

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

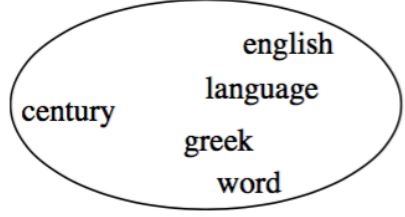
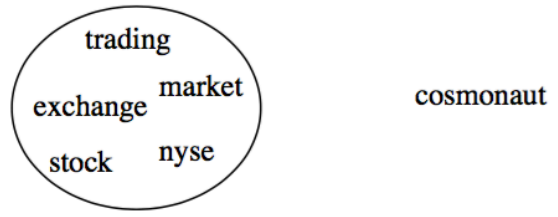
$$MPCT_k^m = H(p_{turk}(\mathbf{w}_{k,1}^m), \dots, p_{turk}(\mathbf{w}_{k,5}^m))$$



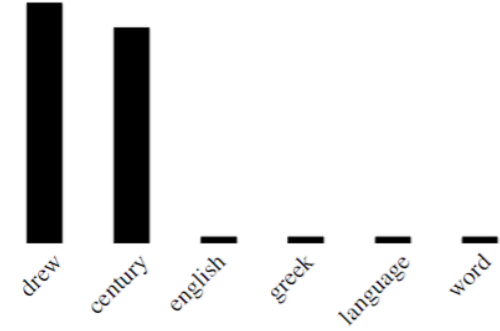
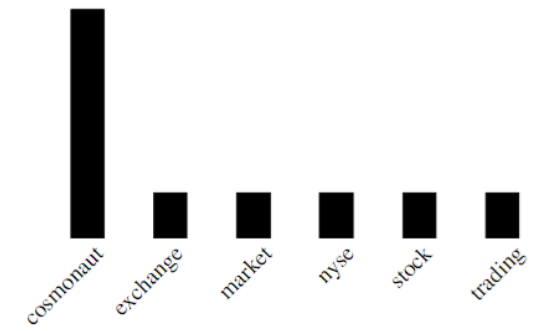
# A Comparative Example

trading  
exchange market  
stock nyse  
cosmonaut

english  
language  
greek word  
century  
drew



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

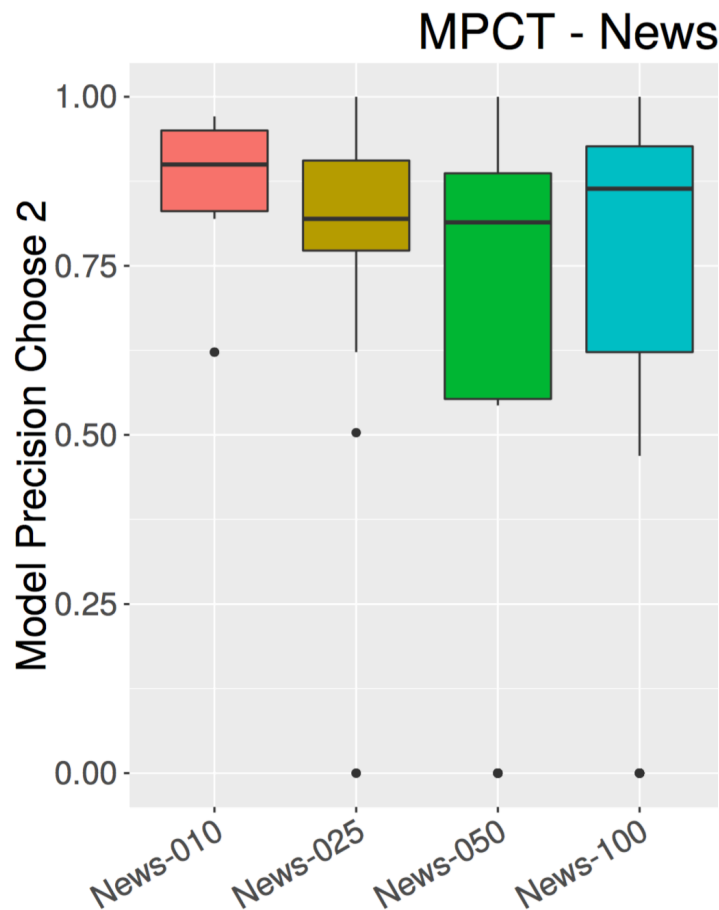
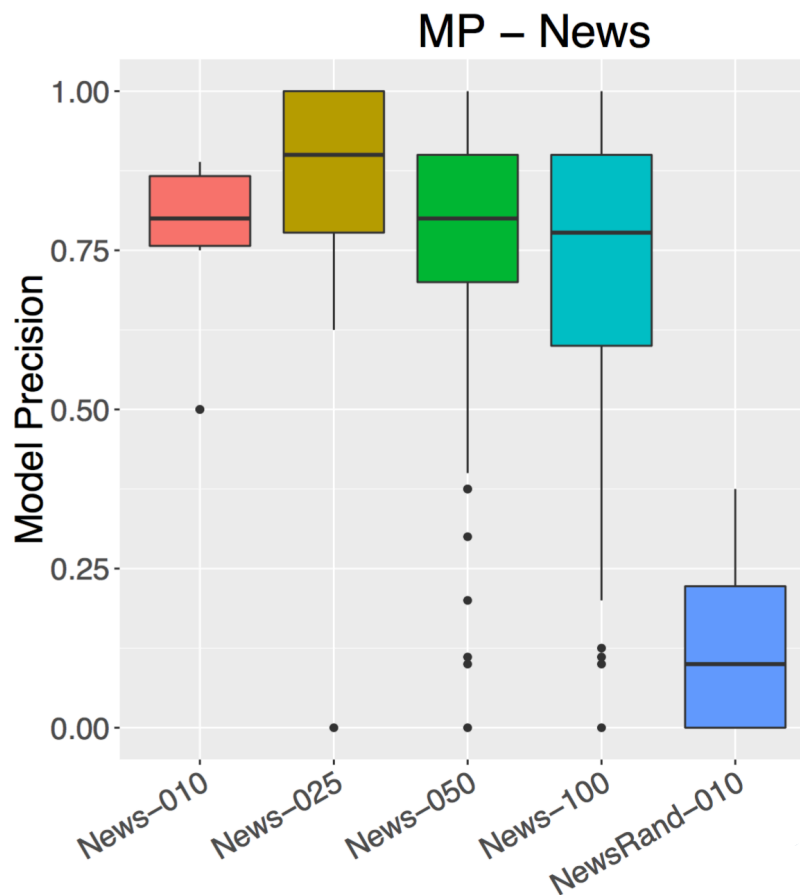
---

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

- Both measures are needed with little extra overhead

0 0 | 1 0  
0 1 | 1 1

# Summary

---

- 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>

[Fred Morstatter](#) and Huan Liu. "A Novel Measure for Coherence in Statistical Topic Models", Association of Computational Linguistics ([ACL](#)), August 2016. Berlin, Germany

# Addressing Don't-Know-Don't-Know Problems

---

- When collecting data, we often *don't know* when we have a sufficient amount
  - We don't know *when to stop* collecting, though we can't collect forever
- A dilemma in studying *migration* on social media :
  - If we know its existence, no need for the study
  - If we *don't know*, how can we verify the result?

# Illustrative Examples of DNDN

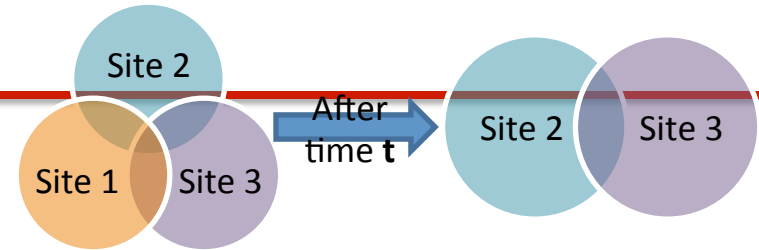
---

1. When-to-Stop Dilemma: Collecting data forever vs. having credible patterns
  - How much data vs. how credible
2. 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
  - ***Competiting for attention*** entails the discovery of migration patterns

# Migration on Social Media

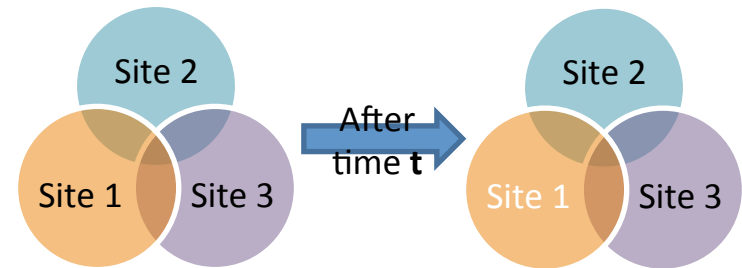
- **Site Migration**

- 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

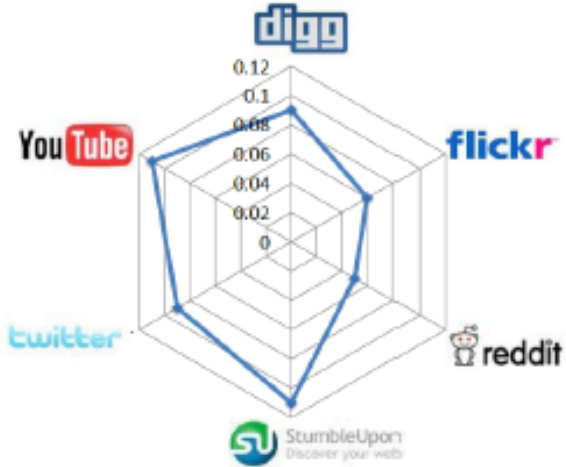


- **Attention Migration**

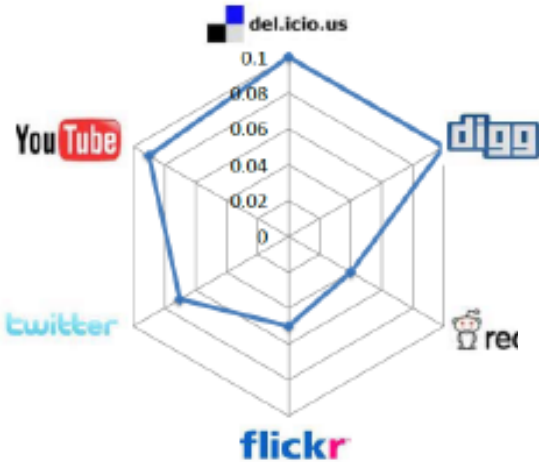
- Users become inactive on a site
- A harbinger for site migration
- Can be detected by observing *user activities* across sites
- Can take action to prevent site migration after understanding migration patterns



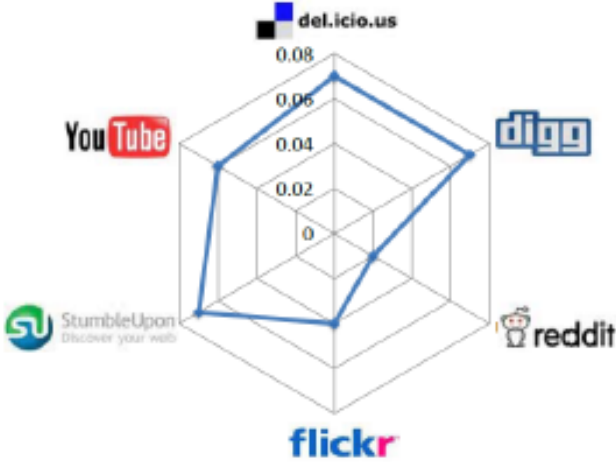
# Patterns from Observation



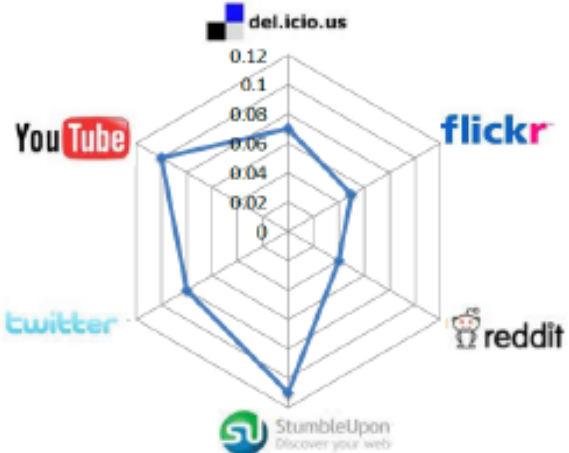
(a) Delicious



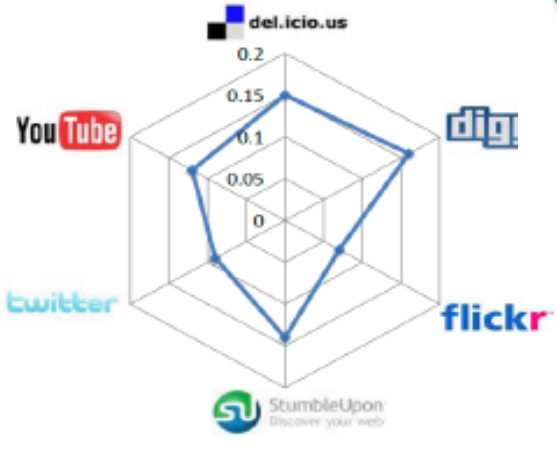
(e) StumbleUpon



(f) Twitter



(b) Digg



(d) Reddit

# Do We Know What We Didn't Know?

- If a pattern is significant, it is valid
  - Significant differences observed in StumbleUpon, Twitter, and YouTube

- When to stop?

Stop when we are certain, continue otherwise

Table 2:  $\chi^2$  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

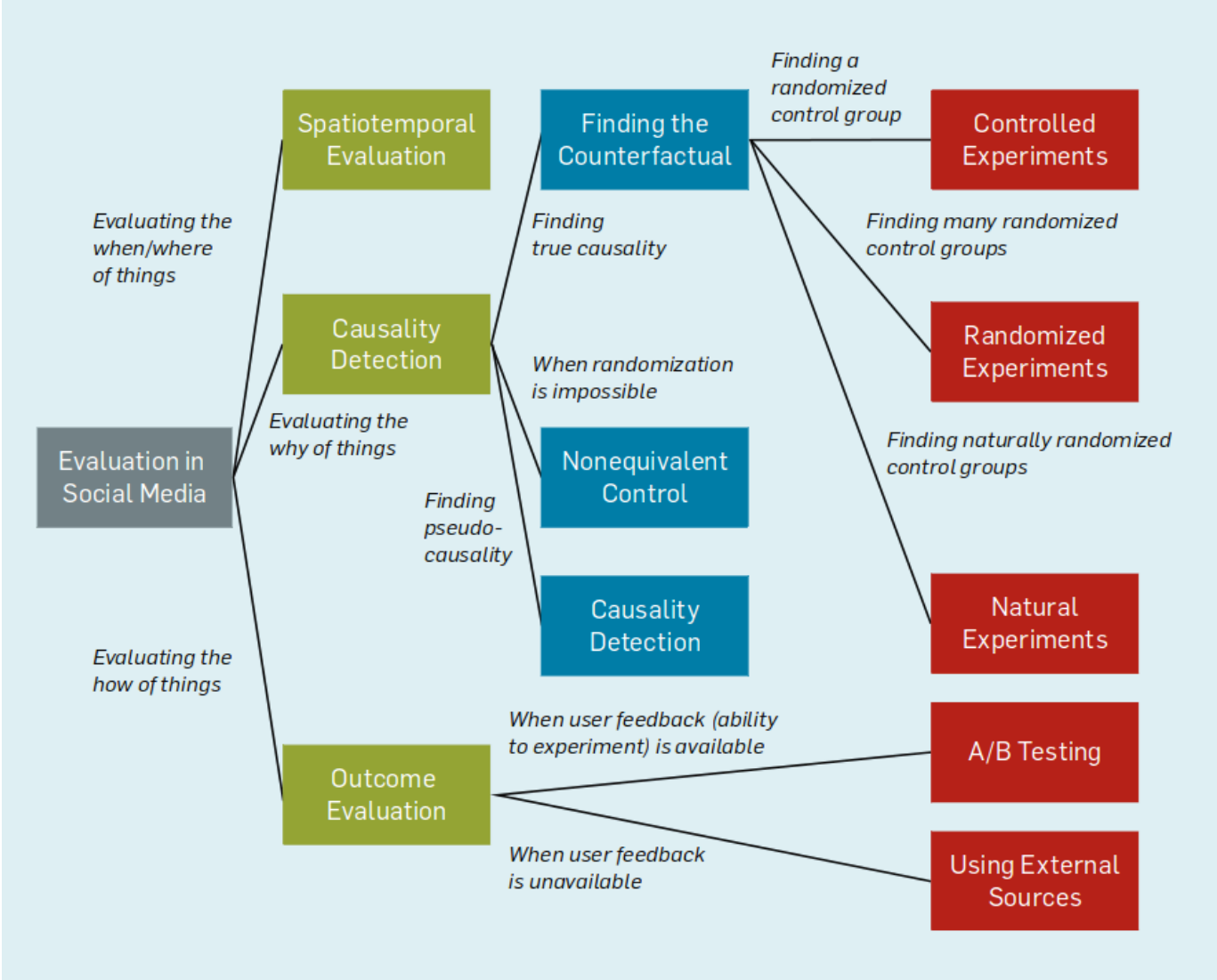
1. Pop-Tarts before a hurricane (Walmart)
2. Higher crime, more Uber rides (Uber)
3. Typing with proper capitalization indicates creditworthiness (A financial services startup)
- 4. Users of the Chrome and Firefox browsers make better employees (An HR firm over Xerox data)**
8. **Female-named hurricanes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

...

Yes, they are bizarre, but are they true?



# Evaluation without Ground Truth



The CACM article can be found at [dl.acm.org](http://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

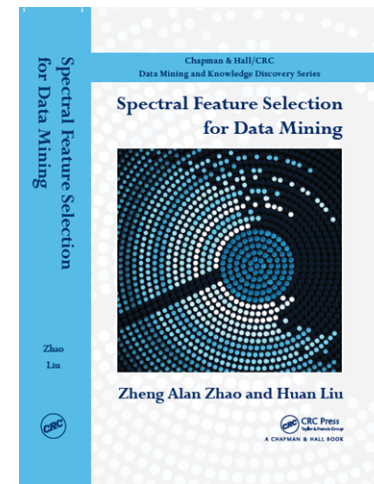
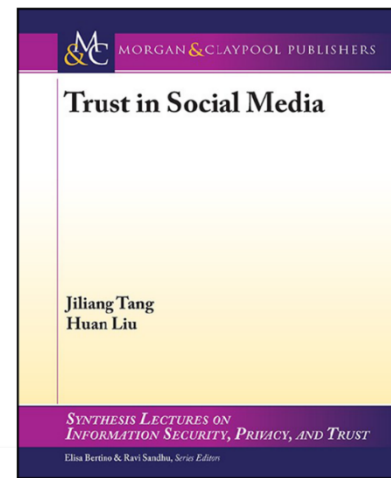
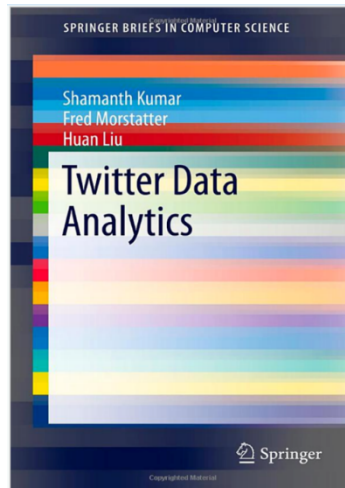
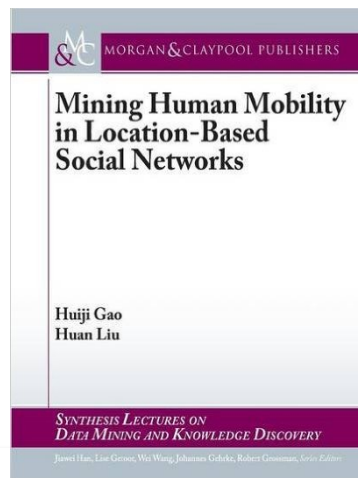
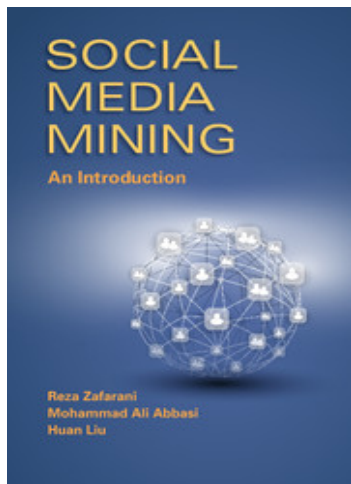
# Revisit Challenges in Acquiring SM Intelligence

---

- Social media data is obviously big, but why are we often still short of data?
  - How can we make data *'bigger'*?
- Data is power, so it can produce any result
  - Can we *algorithmically* evaluate the results from big data?
- We don't know what we don't know
  - How can we know if our result of social media analysis is of any value?

# Repositories and Recent Books

- ***scikit-feature*** – an open source feature selection repository in Python
- Social Computing Repository
- Some books available for free download



## Social Media Mining An Introduction

A Textbook by Cambridge University Press

Reza Zafarani

Mohammad Ali Abbasi

Huan Liu

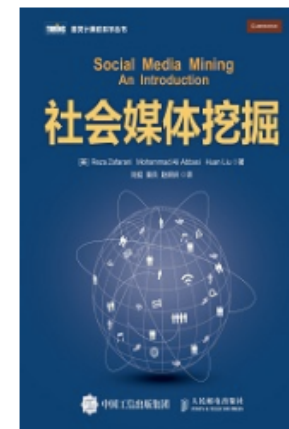
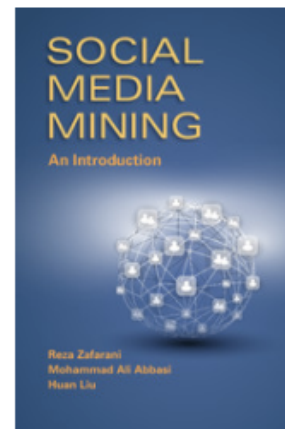
Syracuse University

Machine Zone

Arizona State University



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

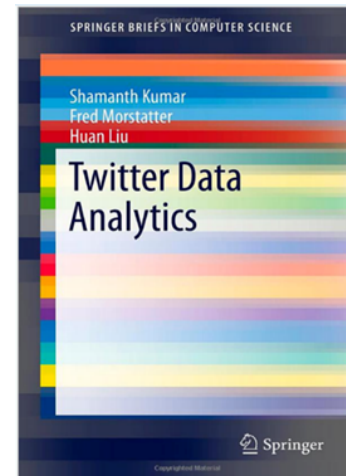
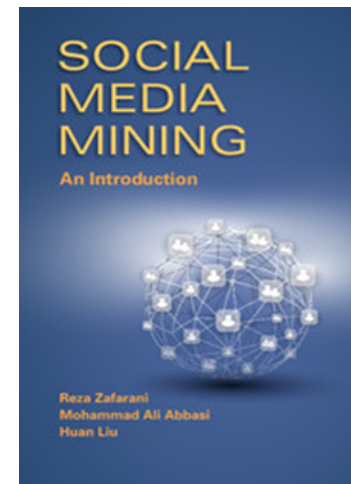


The growth of social media over the last decade has revolutionized the way individuals interact and

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

# Discovering Social Media Intelligence

- Graph Theories
- Network Measures and Models
- Data Mining, NLP, and Visual Analytics
- Community Detection and Analysis
- Information Diffusion
- Influence and Homophily
- Recommender Systems
- Behavior Analytics
  - Sentiment analysis



<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
  - Collaborators: CMU (Minerva), CRA (IARPA-CAUSE)

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

# Further Readings

---

- [Jundong Li](#) and Huan Liu. "Challenges of Feature Selection for Big Data Analytics", Special Issue on Big Data, IEEE Intelligent Systems. 32 (2), 9-15. 2017
- [Fred Morstatter](#) and Huan Liu. "A Novel Measure for Coherence in Statistical Topic Models", Association of Computational Linguistics (ACL), August 2016. Berlin, Germany
- [Reza Zafarani](#) and Huan Liu. "Evaluation without Ground Truth in Social Media Research", Communications of ACM, Volume 58 Issue 6, June 2015 Pages 54-60.
- [Lei Tang](#) and Huan Liu. "Community Detection and Mining in Social Media", Morgan & Claypool Publishers, September 2010.



# Making Thin Data Thicker

---

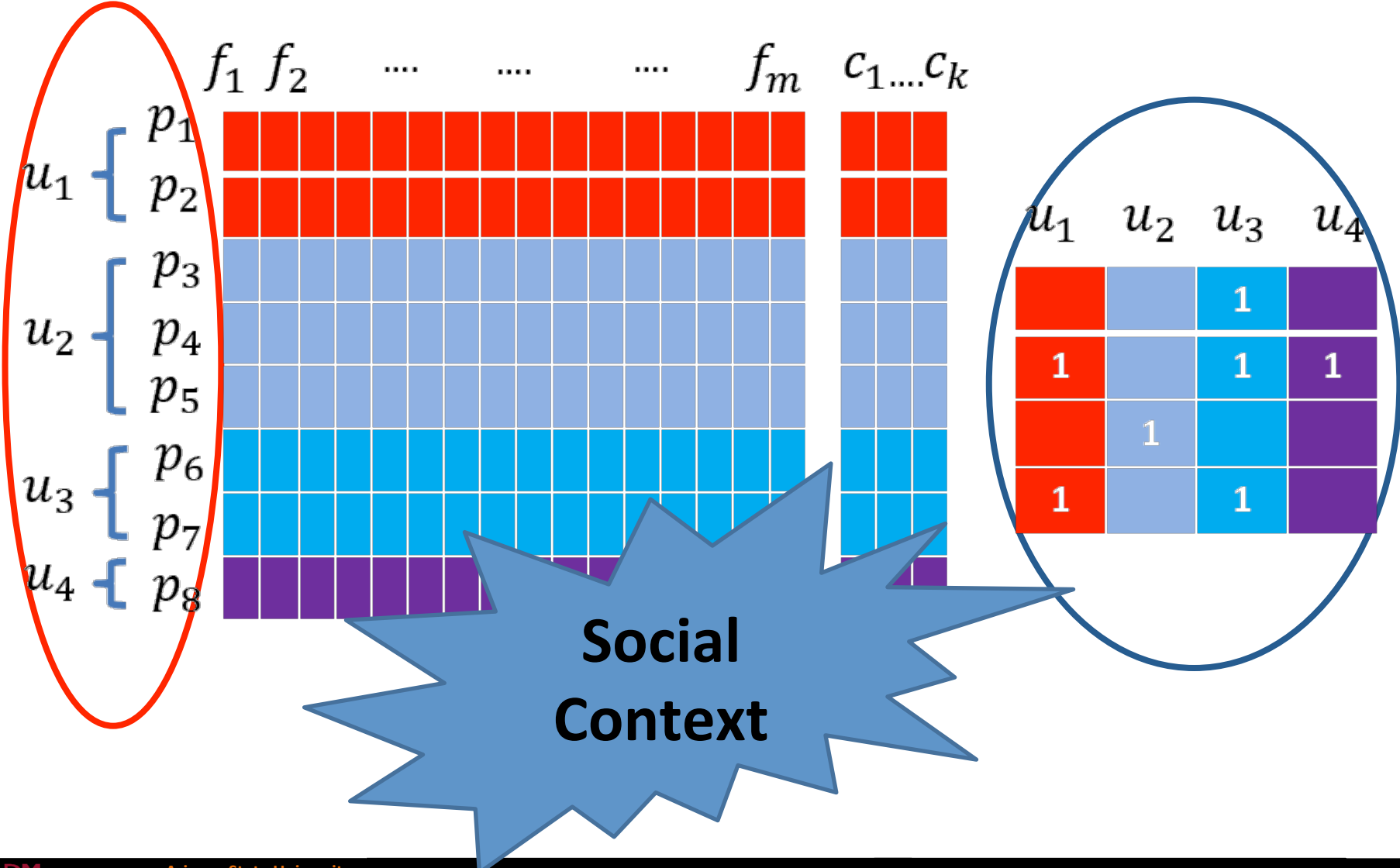
- Most people like many of us are in the long tail
  - Our data is thin or sparse
  - Without little data, machine learning is powerless
- Social media data offers new opportunities
  - Linked information
  - Multiple platforms as they offer different functions
- Two case studies
  - Feature Selection using social network information
  - Connecting users ***across*** more than one social media site

# Use Link Information for Data Thickening

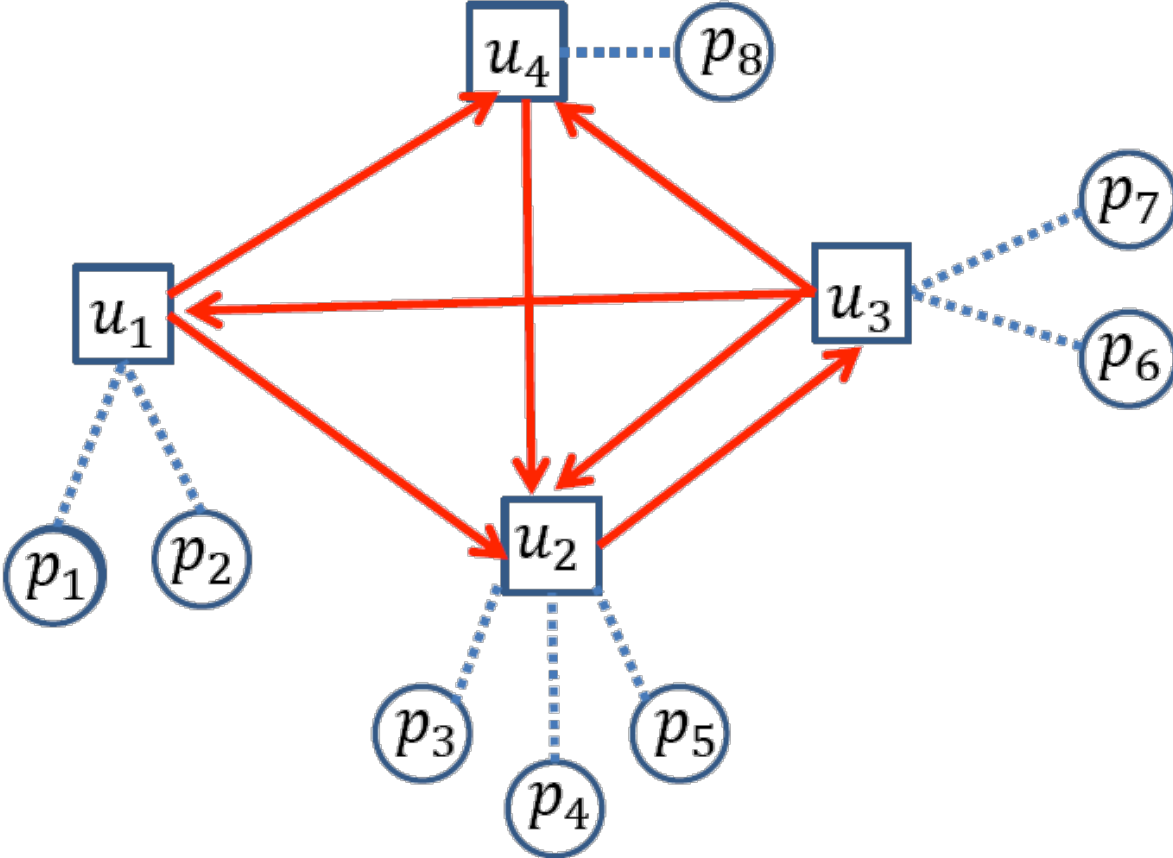
---

- Where can we find additional information for feature selection
- Social media data contains various types of data
  - Link information is additional
  - Other sources such as sentiment, like, etc.
- Are there theories to guide us in using link info?
  - Social influence
  - Homophily
- Extracting distinctive relations from linked data for feature selection

# Representation for Social Media Data



# Relation Extraction



- 1. CoPost
- 2. CoFollowing
- 3. CoFollowed
- 4. Following

# Evaluation Results on Digg Data

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
$\mathcal{T}_5$	50	45.45	44.50	46.33	45.27	<b>58.82</b>	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	<b>59.43</b>	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	<b>63.32</b>
	300	54.04	55.24	56.54	59.27	<b>65.30</b>	60.40	59.93	<b>66.19</b>
$\mathcal{T}_{25}$	50	49.91	50.08	51.54	56.02	<b>58.90</b>	57.76	57.01	<b>58.90</b>
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	<b>65.02</b>
	200	59.97	57.37	60.07	64.36	<b>67.33</b>	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	<b>69.52</b>	65.46	65.01	67.95
$\mathcal{T}_{50}$	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	<b>60.77</b>
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	<b>65.74</b>
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	<b>71.32</b>
	300	61.47	62.35	64.77	69.58	71.86	71.40	70.50	<b>72.65</b>
$\mathcal{T}_{100}$	50	51.74	56.06	55.94	58.08	<b>61.51</b>	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	<b>65.65</b>
	200	60.49	62.78	65.18	66.87	<b>69.75</b>	67.40	67.00	67.31
	300	<b>62.97</b>	66.35	67.12	69.27	<b>73.01</b>	70.99	69.50	72.64

# Summary

---

- LinkedFS is evaluated under varied circumstances to understand how it works
  - Link information can help *feature selection for social media data*
- Unlabeled data is more often in social media, unsupervised learning is more sensible, but also more challenging

Jiliang Tang and Huan Liu. "Unsupervised Feature Selection for Linked Social Media Data", the Eighteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2012.

Jiliang Tang, Huan Liu. "Feature Selection with Linked Data in Social Media", SIAM International Conference on Data Mining, 2012.

# Gather more Data with Little Data



---

- Collectively, social media data is indeed big
- For an individual, however, the data is *little*
  - How much activity data do we generate daily?
  - How many posts did we post this week?
  - How many friends do we have?
- When “big” social media data isn’t big,
  - Searching for **more** data with **little** data
- We use different social media services for varied purposes
  - LinkedIn, Facebook, Twitter, Instagram, YouTube, ...

# An Example

- Little data about an individual
- + Many social media sites
- Partial Information
- + Complementary Information
- > Better User Profiles

## Reza Zafarani

		
	<b>LinkedIn</b>	<b>Twitter</b>
<b>Age</b>	N/A	N/A
<b>Location</b>	Phoenix Area	Tempe, AZ
<b>Education</b>	ASU (2014)	ASU

Connectivity is not available

Consistency in Information Availability

***Can we connect individuals across sites?***

Reza Zafarani and Huan Liu. "Connecting Users across Social Media Sites: A Behavioral-Modeling Approach", the Nineteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2013), August 11 - 14, 2013, Chicago, Illinois.



# Searching for More Data with Little Data

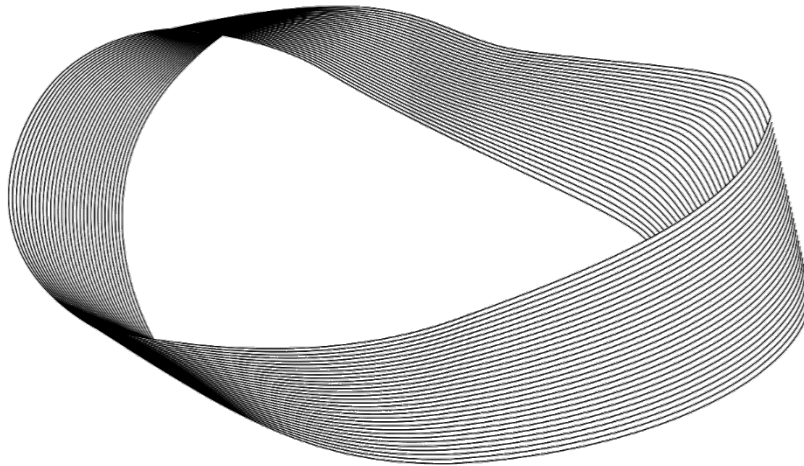
---

- Each social media site can have varied amount of user information
- Which information definitely exists for all sites?
  - **Usernames**
  - But, a user's usernames on different sites can be different
- Our work is to connect the information of the same user provided across sites

# Our Behavior Generates Information Redundancy

---

- Information shared across sites provides a behavioral fingerprint
  - How to capture and use differentiable attributes



**MOBIUS**

- **Behavioral Modeling**
- **Machine Learning**

**MO**delling **B**ehavior for **I**dentifying **U**sers across **S**ites

# Behaviors

Human  
Limitation

Time & Memory  
Limitation

Knowledge Limitation

Exogenous  
Factors

Typing Patterns

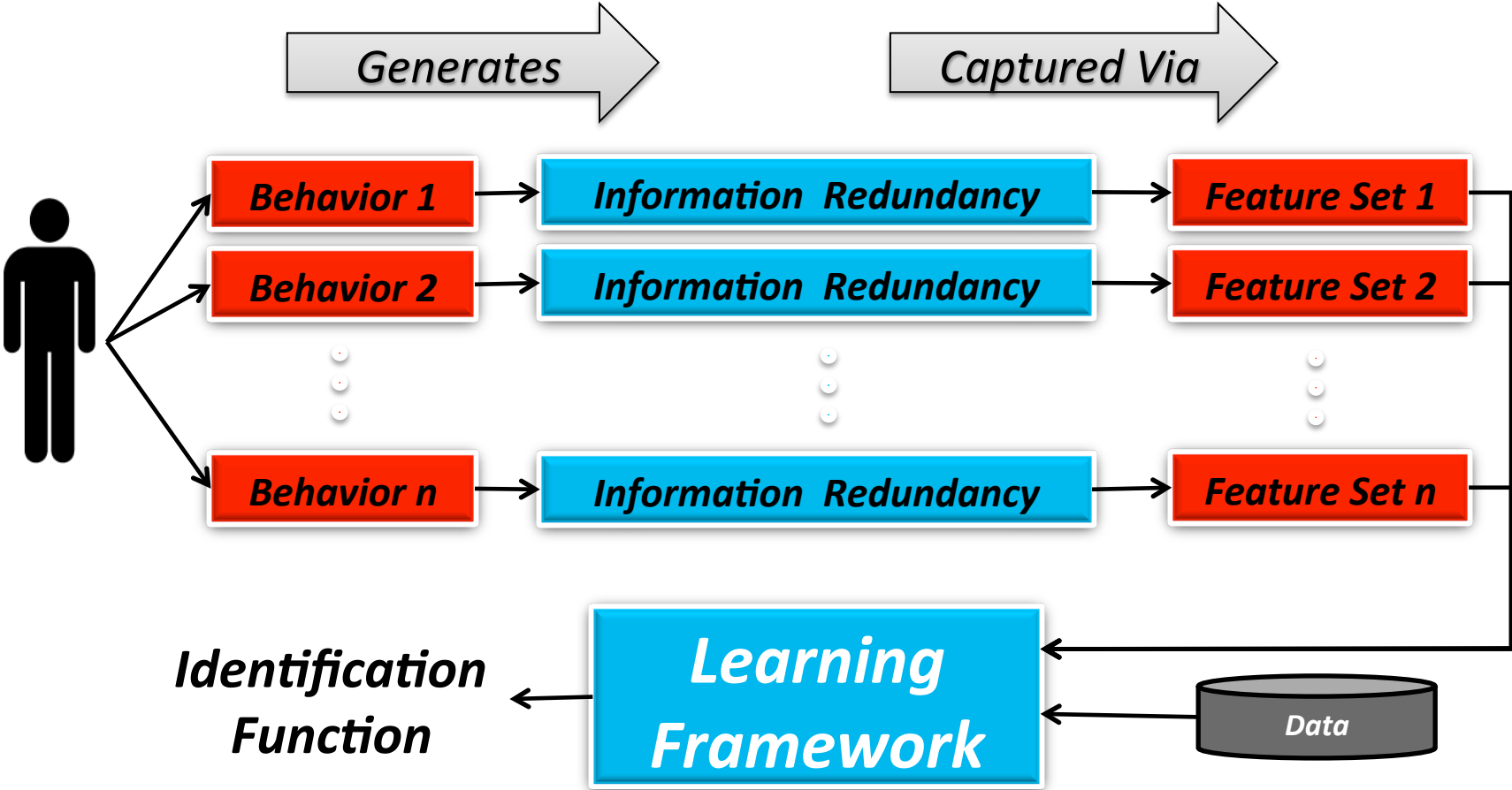
Language Patterns

Endogenous  
Factors

Personal Attributes &  
Traits

Habits

# Behavioral Modeling Approach with Learning



# Summary – Making Data Bigger

---

- Gathering more data is often necessary for effective data mining
- Reducing dimensionality can make data bigger
- Social media data provides unique opportunities to do so by using different sites and abundant user-generated content
- Traditionally available data can also be tapped to make thin data “thicker”

Jundon Li, et al. “Feature Selection: A Data Perspective”, 2016.

<http://arxiv.org/abs/1601.07996>

Reza Zafarani and Huan Liu. “Connecting Users across Social Media Sites: A Behavioral-Modeling Approach”, SIGKDD, 2013.