Data Mining & Machine Learning

ARIZONA STATE UNIVERSITY

Social Media Mining
- A Bountiful Frontier in AI

Huan Liu

Data, Data Everywhere

- Abundant data has changed the AI field
  - From *knowledge-centric* to *data-centric*
  - "Data is the new oil"
- Big *data* helps recent surge of successful AI
  - Data is essential for any DM/ML algorithm
- Social media - a *new* type of *big* data
  - *Novel challenges*
  - *New questions* to answer via this new lens
  - From *Problem Solvers* to ***Problem Finders***

Challenges and Opportunities for SMM

- **Fundamental Problems**
  - Big Data Paradox
  - Noise Removal Fallacy
  - Evaluation Dilemma
- **Intriguing Questions**
  - Who are the influential
  - Is the sample good enough
  - How good is utility-privacy trade-off
- **Making a Difference**
  - Detecting disinformation
  - Combatting cyberbullying
  - Preserving privacy

Challenges and Opportunities

- **Fundamental Problems**
  - ***Big Data Paradox***
  - Noise Removal Fallacy
  - Evaluation Dilemma
- **Intriguing Questions**
  - Who are the influential
  - Is the sample good enough
  - ***How good is utility-privacy trade-off***
- **Making a Difference**
  - ***Detecting disinformation***
  - Combatting cyberbullying
  - Preserving privacy

New Data, New Challenges

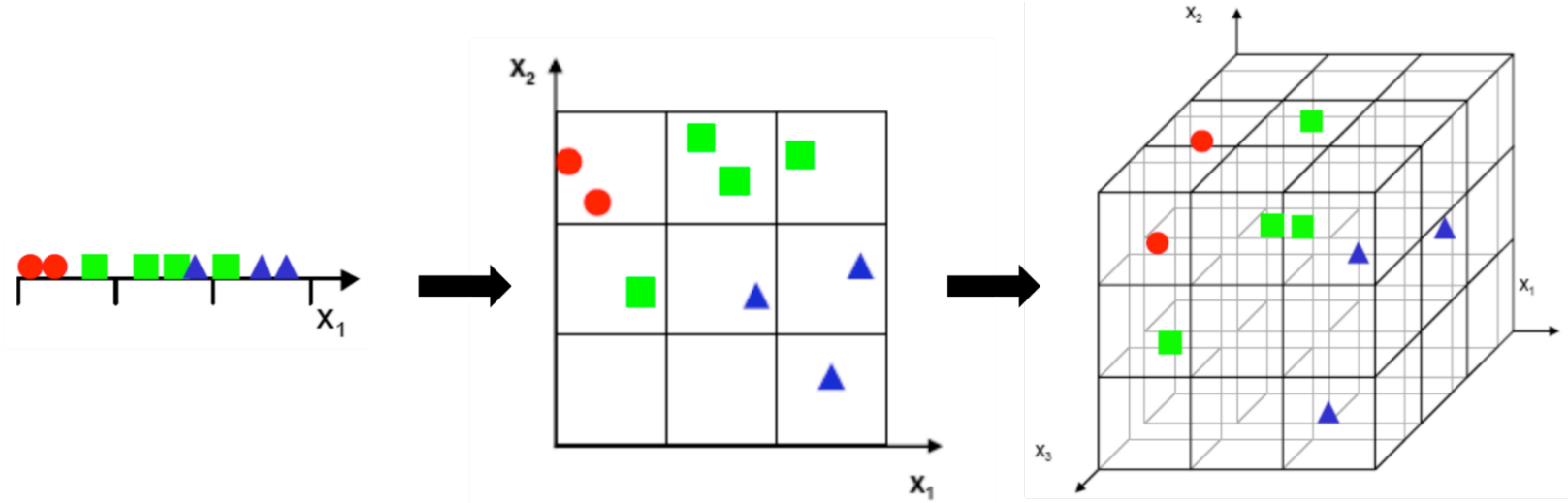
- Is social media data *big*?
  - **Big Data Paradox**
- Do social media users have *privacy*?
  - Privacy-Utility 'Trade-off'
- Socially responsible AI
  - Detecting Disinformation/Fake News

Big Data Paradox

- What is big data?
  - A conventional answer is 4Vs
  - A practitioner's answer is more nuanced
- 'Big' SM data can be *little* or *thin*
- When our data alone isn't big enough, ***we face new challenges***
  - Make little data bigger
  - Make thin data thicker

Curse of Dimensionality: Required Samples

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

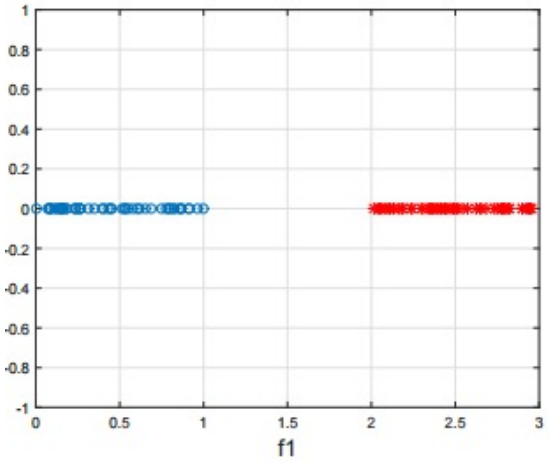


3 samples per unit region **1 sample per region** **1/3 sample per region**

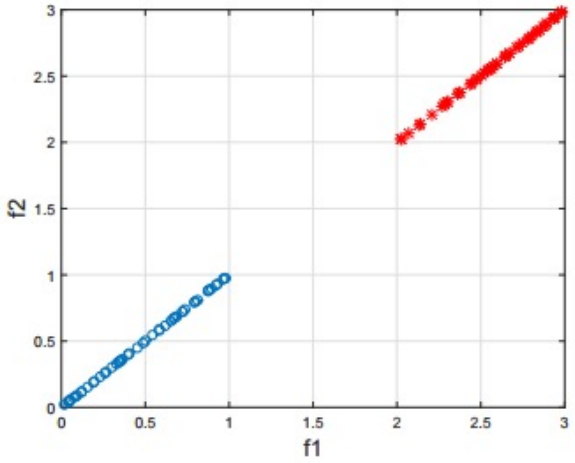
http://nikhilbuduma.com/2015/03/10/the-curse-of-dimensionality/

Relevant, Redundant and Irrelevant Features

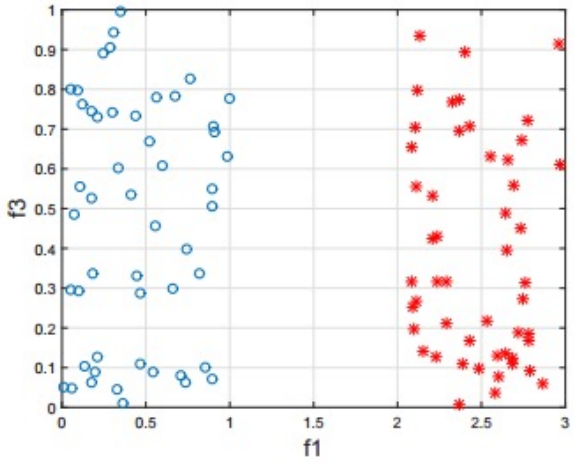
- Feature selection retains relevant features for learning and removes redundant or irrelevant ones
- For a binary classification task below, f_1 is relevant, f_2 is redundant given f_1, and f_3 is irrelevant



(a) relevant feature f_1



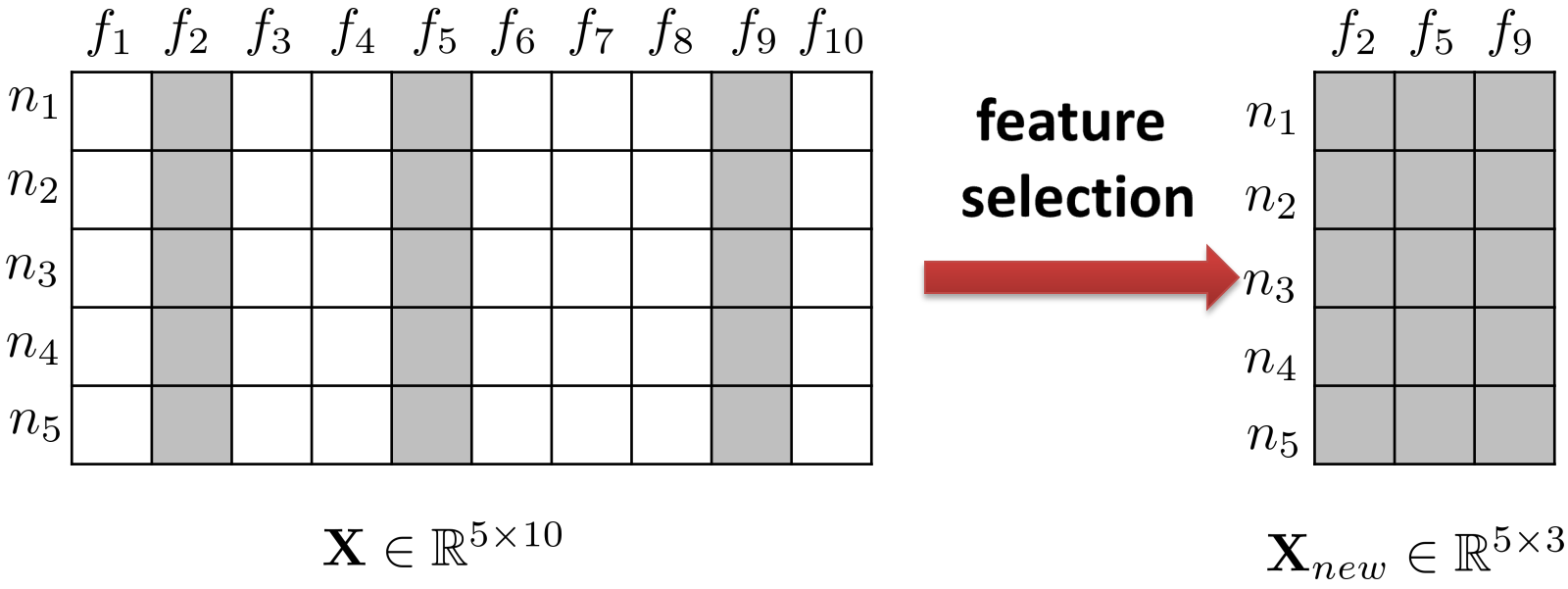
(b) redundant feature f_2



(c) irrelevant feature f_3

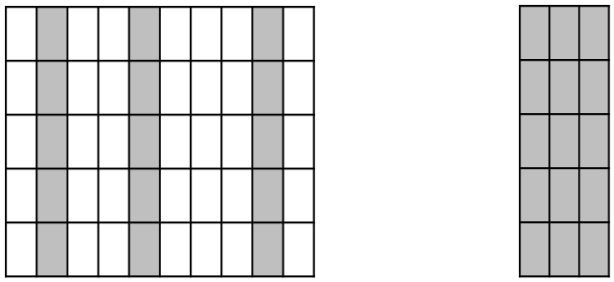
What Does Feature Selection Do?

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



Feature Selection Makes Data Bigger

- How can selection make data bigger?
  - Assuming all binary attribute values in our toy example
  - Before FS, $5/2^{10}$ <0.5%
  - After FS, $5/2^{3}$ >50%



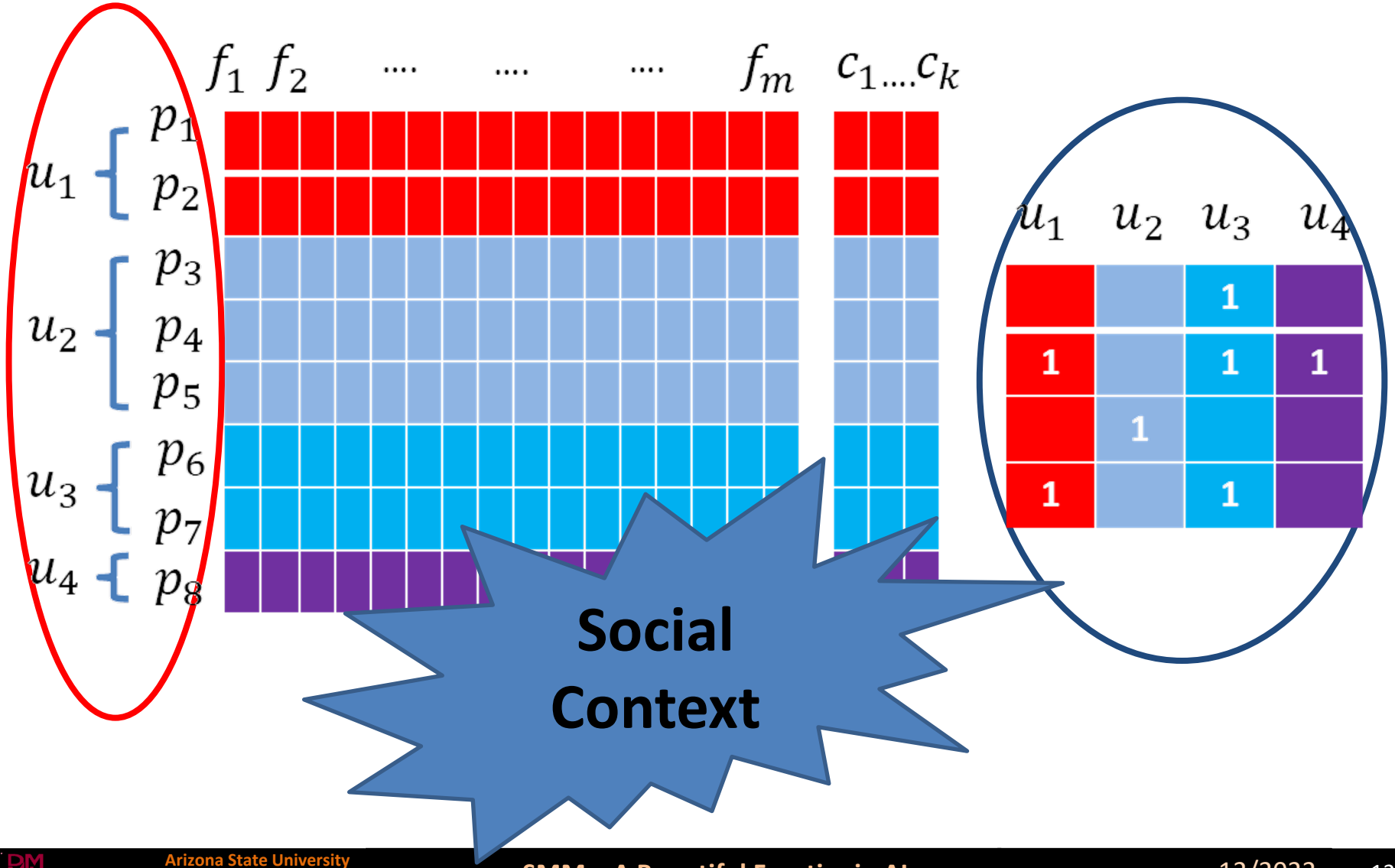
$\mathbf{X} \in \mathbb{R}^{5\times 10}$ $\mathbf{X}_{new} \in \mathbb{R}^{5\times 3}$

- Does feature selection always work?
  - Almost always for high-dimensional data
  - Scikit-Feature

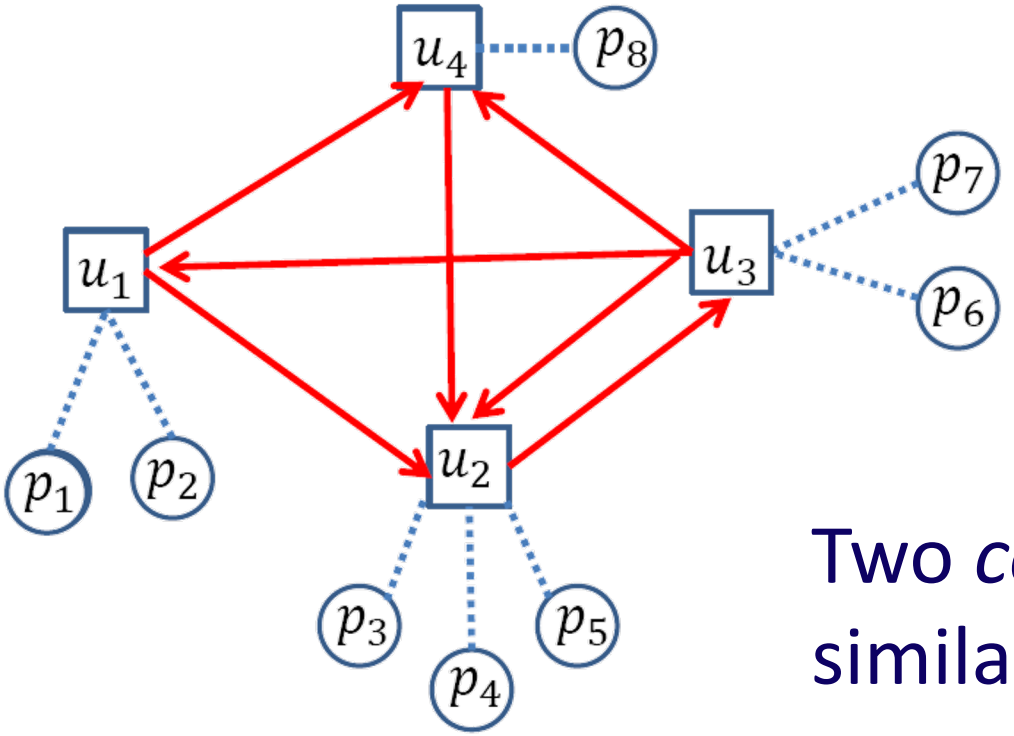
Use Additional Info for Data Thickening

- Where to find additional information?
- Social media data contains rich information
  - Link information is available
  - Other sources such as sentiment, like, ...
- 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

Two *co-following* users share similar topics of interests

$$\min_{W} \| X^T W - Y \|_F^2 + \alpha \| W \|_{2,1} + \beta \sum_{u} \sum_{u_i, u_j \in N_u} \| \hat{T}(u_i) - \hat{T}(u_j) \|_2^2$$

Modeling CoFollowing Relation

- Two co-following users have similar topics of interests

Users' topic interests

$$\hat{T}(u_k) = \frac{\sum_{f_i \in F_k} T(f_i)}{|F_k|} = \frac{\sum_{f_i \in F_k} W^T f_i}{|F_k|}$$

$$\min_{\mathrm{W}} \| \mathrm{X}^T \mathrm{W} - \mathrm{Y} \|_F^2 + \alpha \| \mathrm{W} \|_{2,1} + \beta \sum_{u} \sum_{u_i, u_j \in N_u} \| \hat{T}(u_i) - \hat{T}(u_j) \|_2^2$$

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

Summary

- LinkedFS is evaluated under various circumstances to understand how it works
  - Link information can help *feature selection for social media data*
- Social media data is often unlabeled
  - Labeled data is costly to obtain
  - Labeled data should be efficiently used

New Data, New Challenges

- Is social media data *big*?
  - Big Data Paradox
- Do social media users have *privacy*?
  - **Privacy-Utility 'Trade-off'**
- Socially responsible AI
  - Detecting Disinformation/Fake News

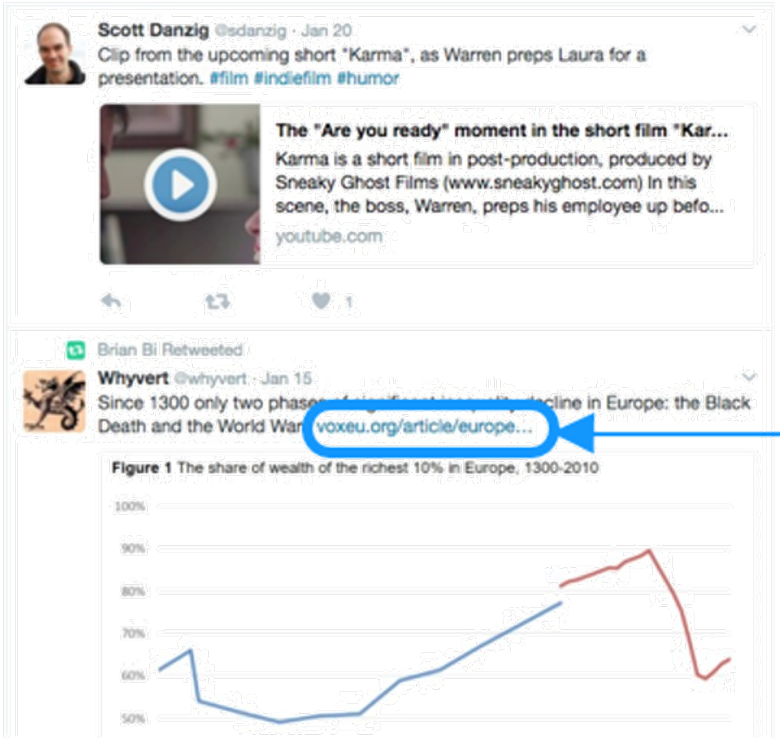
User Browsing Histories Can Reveal Privacy

- Adversaries can infer different types of personally identifiable information (PII)
- Web browsing history data is *fingerprintable*
  - New attacks that map a given history to a social media profile
- Users can be vulnerable to potential harms
  - When damages are done, very hard to correct

Attacks via Web Browsing History

Given *u*'s browsing history $\mathcal{H}^u = \{l_1, ..., l_n\}$, map *u* to a social media profile based on the links in its feed

Twitter feed



Browsing history

https://facebook.com

http://cs246.stanford.edu

http://voxeu.org/article/...

Achieving Good Privacy-Utility Trade-off

- We need to make a trade-off between privacy and utility
- *Anonymization* is a chosen means that can improve privacy, but reduce service utility, the quality of personalized services
- Can we do better privacy-utility trade-off by using social network properties?

To Preserve Privacy is to Anonymize

- Intuitive ways of anonymization
  - Adding one's friends' links
  - Adding random links
  - What else?

- The question becomes "how to add new links"
  - What links should be added?
  - How many links should be added?

Boosting Privacy via Adding New Links

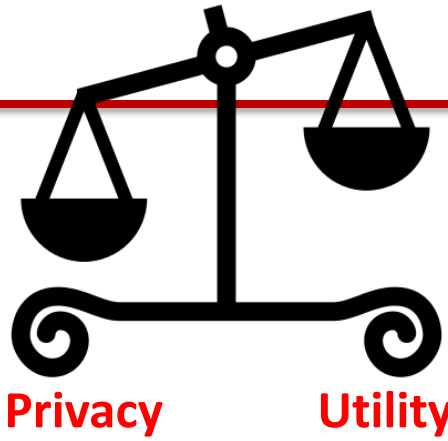
- We aim to find a set of new links $\mathcal{A}$ such that:

$$\mathcal{A}^* = \text{argmax}_{\mathcal{A}}\, G(p_u, \hat{p_u}, \lambda)$$

$$G(p_u, \hat{p_u}, \lambda) = \lambda * privacy(\hat{p_u}) - utility\_loss(p_u, \hat{p_u})$$

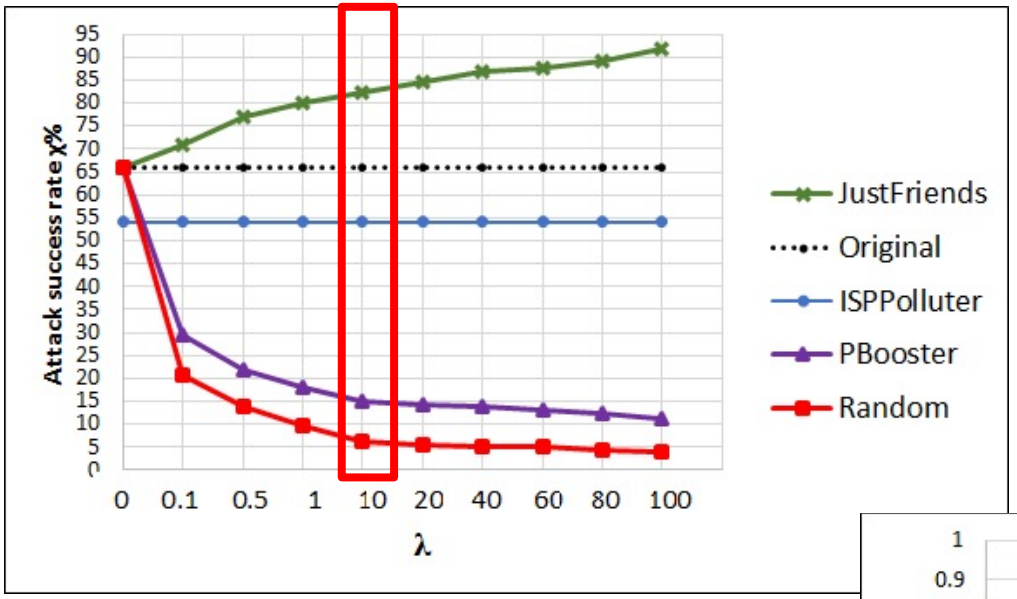
- **PBooster** Algorithm
  - **Topic Selection**: Select a subset of topics and calculate the number of links
  - **Link Selection**: Select some random links that correspond to the identified topics

How It Fares - Empirical Evaluation



1. Can PBooster preserve user privacy?
2. How does PBooster change the utility?
3. Does Pbooster work?
   - How well is the privacy-utility trade-off?

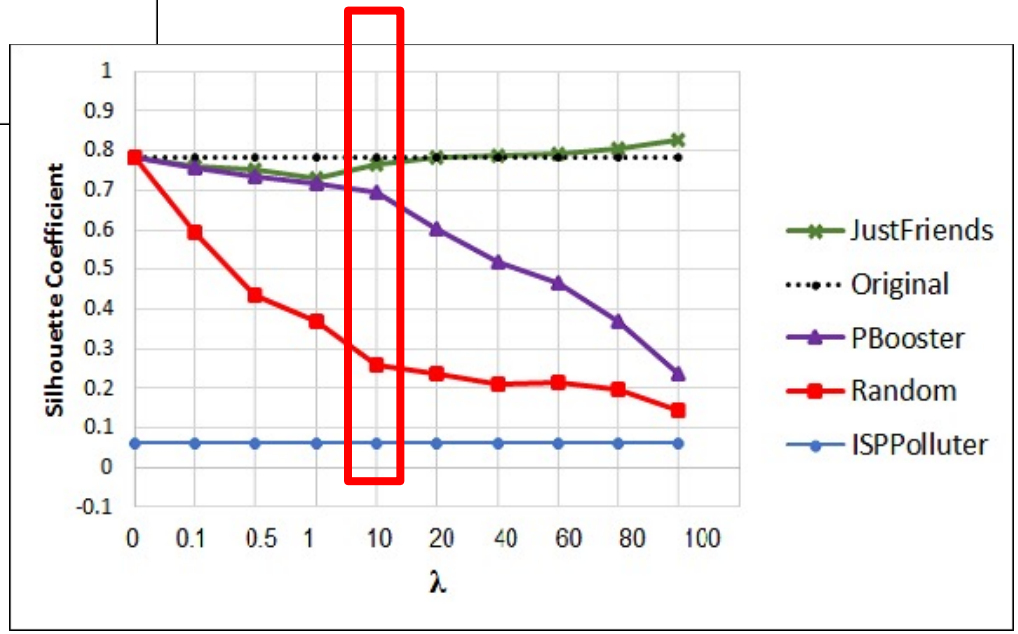
Sweet Spots for High Privacy and Utility



Privacy Evaluation: Deploy an existing de-anonymization attack

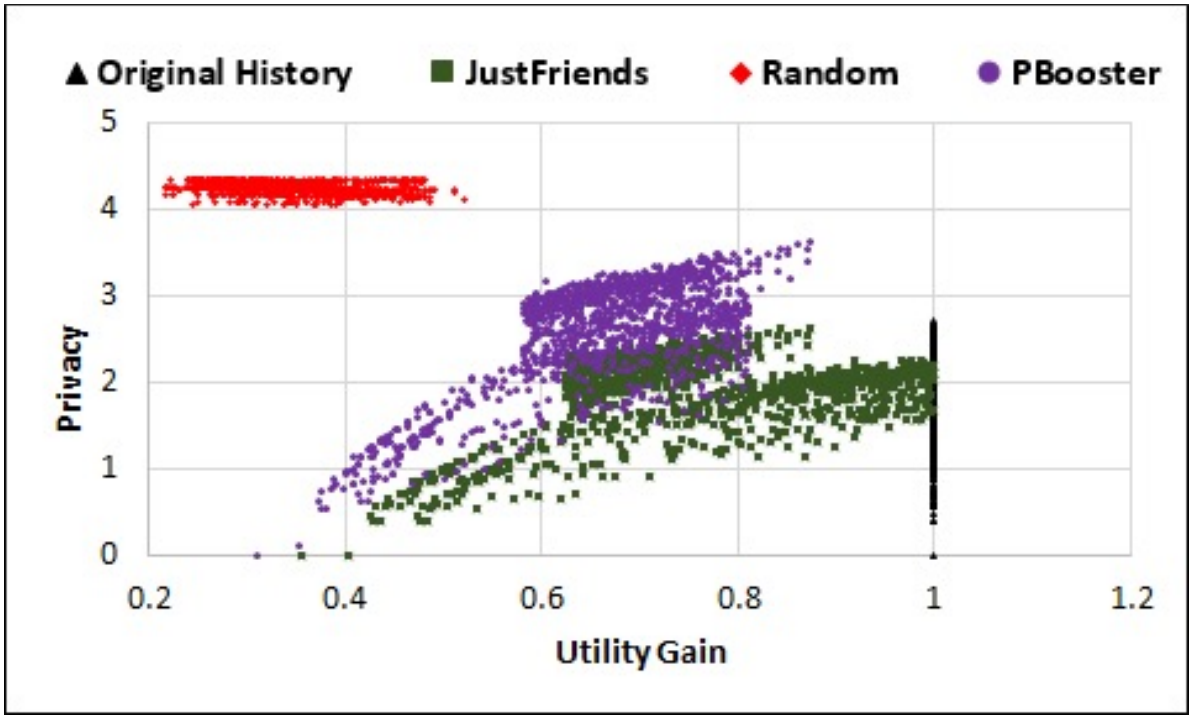
Attack is successful if the user is among the top 10 results

Utility Evaluation: Cluster users with k-means based on topic probability distributions



Summary

- Social media data is special
  - Adding some random links can significantly reduce average path lengths
  - The reason for the sweet spots



New Data, New Challenges

- Is social media data *big*?
  - Big Data Paradox
- Do social media users have *privacy*?
  - Privacy-Utility 'Trade-off'
- Socially responsible AI
  - **Detecting Disinformation/Fake News**

10,972 views | Jun 22, 2020, 12:19am EDT

10 Wonderful Examples Of Using Artificial Intelligence (AI) For Good

Bernard Marr Contributor ⓘ
Enterprise Tech

Spot "Fake News"

It's true: AI is the engine that pushes "fake news" out to the masses, but Google, Microsoft, and grassroots effort Fake News Challenge are using AI (machine learning and natural language processing) to assess the truth of articles automatically. Due to the trillions of posts, Facebook must monitor and the impossibility of manually doing it, the company also uses artificial intelligence to find words and patterns that could indicate fake news. Other tools that rely on AI to analyze content include Spike, Snopes, Hoaxy, and more.

Fake News - Disinformation

- Disinformation is false information [news or non-news] with an evil intention to mislead the public
- Fake news is news with intentionally false information
- Fake news can have detrimental societal effects
  - Confusing readers
  - Misleading people to false information
  - Changing the way people respond to credible news

Why is it so challenging

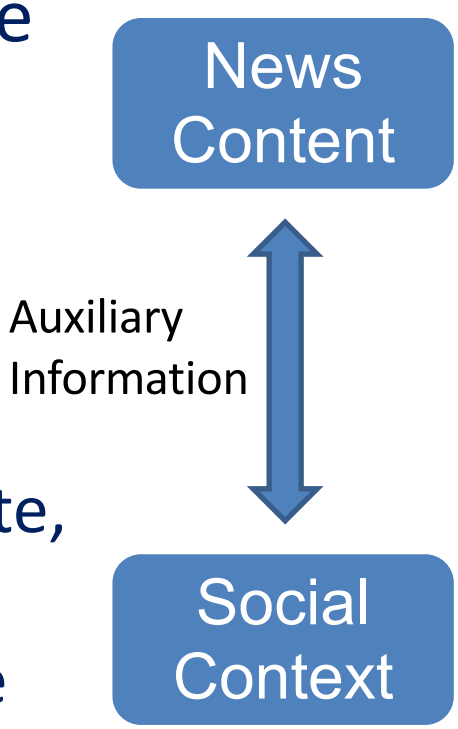
- Fake news detection is not just another organized competition
  - A competition provides a dataset with ground truth and shows who is the best
- Humans are susceptible to fake news
  - Typical accuracy in the range of 55-58%
  - Limited resources (time, information, and expertise)
  - **Confirmation Bias**: individuals tend to believe fake news when it confirms their pre-existing knowledge

Five Challenges

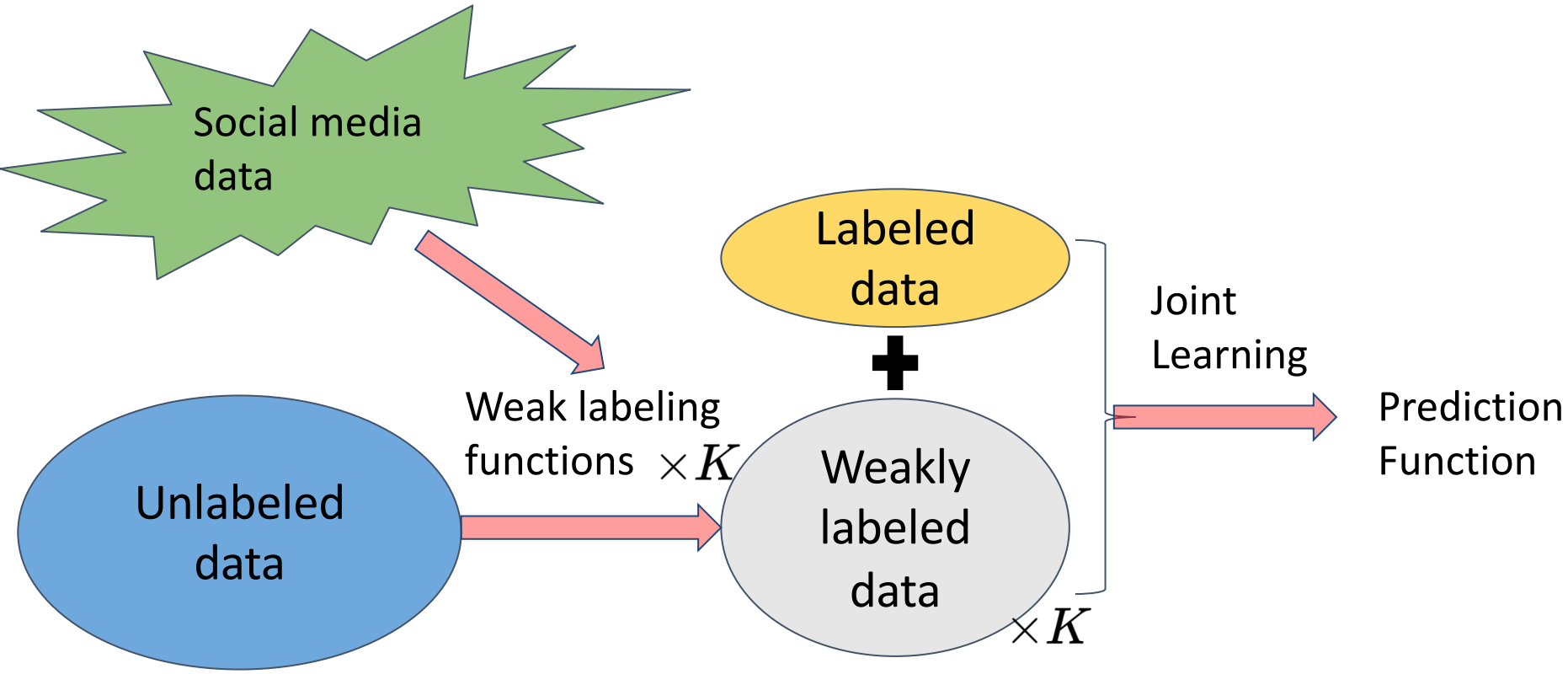
- Detection
  - Topics change
- Early Detection
  - Finding it before it becomes viral
- Ground Truth for evaluation
  - Benchmark data and where to find labels
- Explainability
  - Human in the loop to expedite the process
- Mitigation or Containment
  - Difficult as it involves social media users

Weak Social Supervision for FN Detection

- News Content
  - Intentionally written to mislead people
  - Diverse in terms of topics, styles, and media platforms
- Social Context
  - Social interactions are massive, incomplete, unstructured, and noisy
  - Effective methods are needed to leverage rich social signals



Problem Statement: Given a limited amount of manually annotated news data $\mathcal{D}$ and K sets of weakly labeled data $\{\tilde{\mathcal{D}}^{(k)}\}_{k=1}^{K}$ derived from K different weak labeling functions based on weak social signals, learn a fake news classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ which generalizes well onto unseen news pieces.



Constructing Weak Labels

- Deriving weak labels via statistical measures guided by *computational social theories*

Sentiment-based: *If a news piece has a standard deviation of user sentiment scores greater than a threshold τ_1, then the news is weakly labeled as fake news.*

Bias-based: *If the mean value of users' absolute bias scores – sharing a piece of news – is greater than a threshold τ_2, then the news piece is weakly-labeled as fake news.*

Credibility-based: *If a news piece has an average credibility score less than a threshold τ_3, then the news is weakly-labeled as fake news.*

Comparing with clean labels, three weak labeling functions have F1-scores: 0.65, 0.64, 0.75.

Challenges and Opportunities for SMM (revisit)

- **Fundamental Problems**
  - ***Big Data Paradox***
  - Noise Removal Fallacy
  - Evaluation Dilemma
- **Intriguing Questions**
  - Who are the influential
  - Is the sample good enough
  - ***How good is utility-privacy trade-off***
- **Making a Difference**
  - ***Detecting disinformation***
  - Combatting cyberbullying
  - Preserving privacy

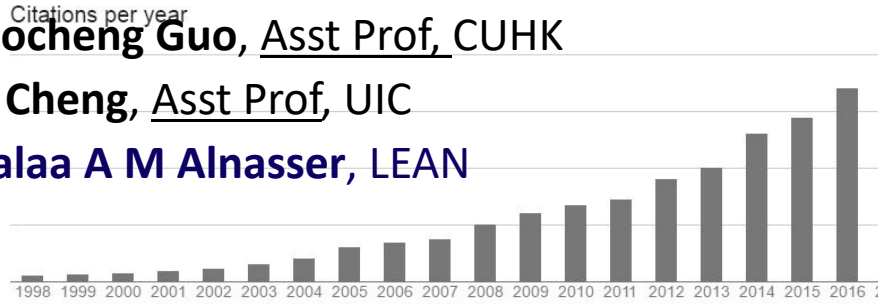
Future Work

- **Fundamental Problems**
  - Causal learning
  - Low-resource learning
  - Algorithmic vs user biases
- **Intriguing Questions**
  - Online vs offline behavior
  - Making the invisible visible
  - Machine vs human text generation
- **Making a Difference**
  - Containing disinformation
  - Mitigating unintended social bias
  - Enabling the powerless

Look beyond social media data

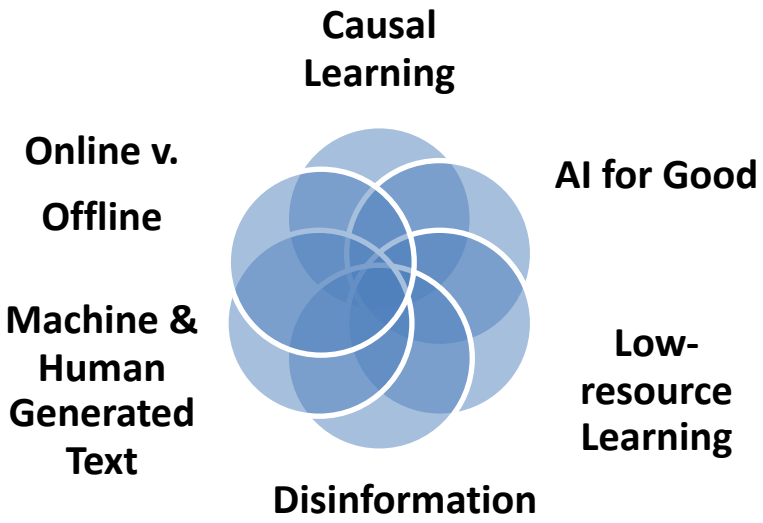
Thanks to Former PhD Students/Collaborators

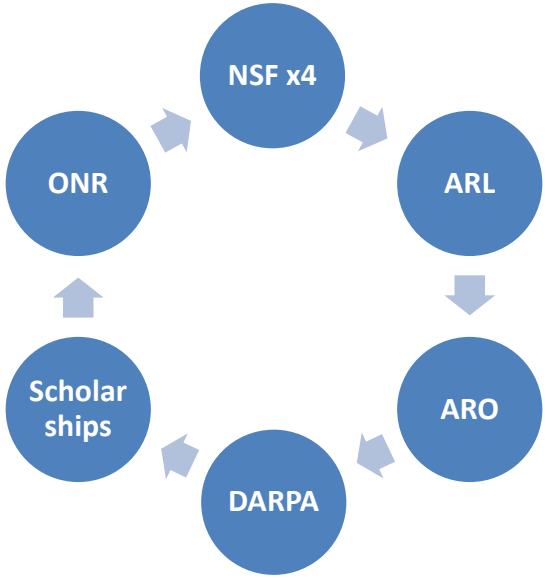
- **Yunzhong Liu**, Verizon Media (w Dr. Yi Chen)
- **Robert Trevino**, AFRL
- **Reza Zafarani**, Asst Prof, Syracuse U
- **Xia Hu**, Assc Prof, Rice U
- **Magdiel Galan**, Intel
- **Shamanth Kumar**, Twitter
- **Pritam Gundecha**, IBM Res Almaden
- **Jiliang Tang**, Foundation Prof, MSU
- **Huiji Gao**, LinkedIn
- **Ali Abbasi**, LinkedIn
- **Salem Alelyani**, Asst Prof, King Khalid U
- **Xufei Wang**, News Break
- **Geoffrey Barbier**, Space Force
- **Lei Tang**, Lyft
- **Zheng Zhao**, Google (w Dr. Jieping Ye)
- **Nitin Agarwal**, Chair Prof, UALR
- **Sai Moturu**, Livongo
- **Lei Yu**, Assc Prof, Binghamton U, NY
- **Somnath Shahapurkar**, FICO
- **Fred Morstatter**, R Asst Prof, USC ISI
- **Christophe Faucon**, Google
- **Suhas Ranganath**, Walmart Labs
- **Suhang Wang**, Asst Prof, Penn State
- **Liang Wu**, Airbnb
- **Jundong Li**, Asst Prof, U of Virginia
- **Isaac Jones**, Google
- **Tahora Nazer**, Spotify
- **Nur Kamarudin**, S Lecturer, U Malaysia Pahang
- **Gennaro De Luca**, Lecturer, Poly ASU, (w Dr. Yinong Chen)
- **Ghazaleh Beigi**, Google
- **Kai Shu**, Asst Prof, IIT Chicago
- **Ruocheng Guo**, Asst Prof, CUHK
- **Lu Cheng**, Asst Prof, UIC
- **Walaa A M Alnasser**, LEAN



Thanks to Students & Collaborators

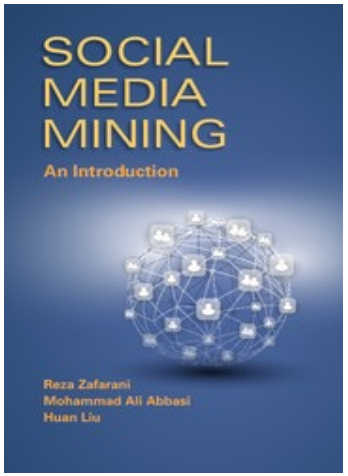
1. **Kaize Ding**, Ph.D., Fall 2017
2. **Raha Moraffah**, Ph.D., Fall 2017
3. **Bing Hu**, Ph.D. (w Dr. Karl G. Kempf), Spring 2018 (PT)
4. **Mansooreh Karami**, Ph.D., Spring 2019
5. **David Ahmadreza Mosallanezhad**, Ph.D., Spring 2019
6. **Brian Vincent**, Ph.D., Fall 2019 (PT)
7. **Weidong Zhang**, Ph.D. (w Dr. Yingzhen Yang), Fall 2019 (PT)
8. **Faisal Alatawi**, Ph.D., Spring 2020
9. **Amrita Bhattacharjee**, Ph.D., Fall 2020
10. **Tharindu Kumarage**, Ph.D., Fall 2020
11. **Paras Sheth**, Ph.D., Fall 2020
12. **Anique Tahir**, Ph.D., Fall 2020
13. **Ujun Jeong,** Ph.D., Spring 2021
14. **Zhen Tan**, Ph.D., Spring 2021
15. **Nayoung Kim**, Ph.D., Spring 2021
16. **Bohan Jiang,** Ph.D., Spring 2021
17. **Garima Agrawal,** Ph.D., Spring 2022
18. **Kritshekhar Jha,** Ph.D., (w Dr. Ming Zhao), Spring 2022
19. **Zeyad Alghamdi**, Ph.D, Spring 2022
20. **Alimohammad Beigi**, Ph.D. Fall 2022

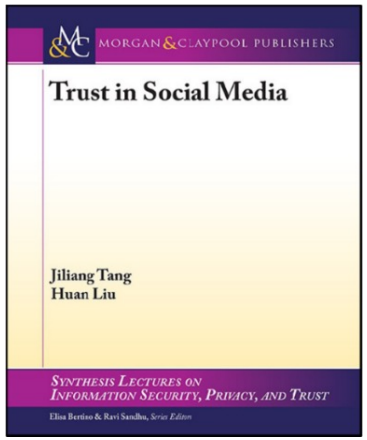


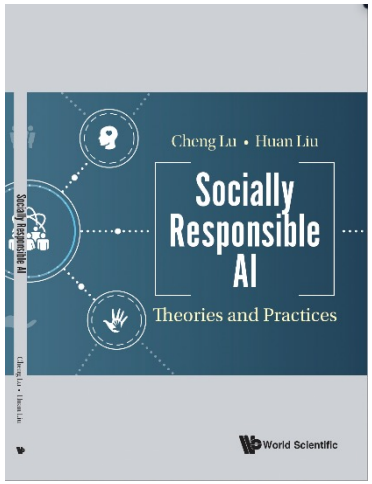


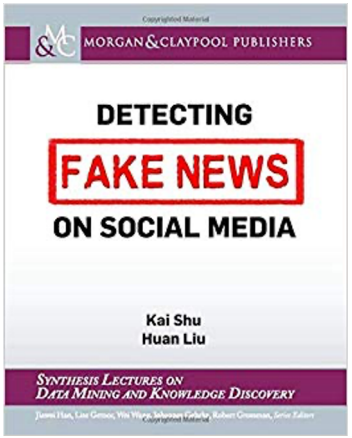
THANK YOU ALL

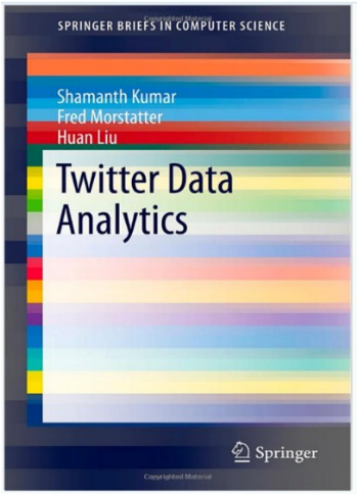
- Some Recent Surveys
  - **Socially Responsible AI Algorithms: Issues, Purposes, …**
  - **Learning Causality with Data: Problems & Methods**
  - **Privacy in Social Media: Identification, Mitigation, …**
  - **Causal Interpretability: Problems, Methods & Eval …**











CALL FOR AUTHORS

Understanding Artificial Intelligence (AI): Natural Language Processing (NLP), Causality, Fairness, and Ethics series

A NEW BOOK SERIES FROM CRC PRESS

Series Editor: Huan Liu

This series aims to provide a conducive platform for researchers and educators to share their findings and lessons learned, focusing on the understanding of AI from both the research perspective and the users' perspective. On the former, we advocate the understanding of AI by advancing techniques and theories in Natural Language Processing (NLP), ChatBots, Bias in Data, Knowledge, Representation, and Evaluation, Fairness, Transparency, Deception, and Explainable AI; on the latter, we endeavor to demystify what AI can do, what responsible AI is via discussion on Bias, Ethics, and Education.

This series will include books on topics that are of interest to researchers and users of AI. The opacity of AI techniques raises unnecessary worries and stimulates wild speculation of AI's imagined un-reined power. This series gathers state-of-the-art research on understanding AI to help demystify AI techniques and algorithms via research and education on NLP, causality, fairness, and ethics.

Consideration will be given to a broad range of textbooks, monographs, reference works, and handbooks that appeal to

PROPOSALS MAY BE SUBMITTED TO:

Huan Liu
Professor of Computer Science and Engineering
Arizona State University
huanliu@asu.edu
http://www.public.asu.edu/~huanliu

Cindy Renee Carelli
Executive Editor
CRC Press – Taylor & Francis Group
cindy.carelli@taylorandfrancis.com