

# Zifeng Wang

[🏠 Home Page](#) | [🎓 Google Scholar](#) | [🐙 GitHub](#) | [✉️ wangzf18@mails.tsinghua.edu.cn](#)

## EDUCATION BACKGROUND

---

### University of Illinois, Urbana-Champaign

PhD student, Computer Science, [The Grainger College of Engineering](#)

Research Interest: AI Health; Advised by: [Prof. Jimeng Sun](#)

Illinois, US

Sept. 2021-Present

### Tsinghua University

MS, Data Science, [Tsinghua-Berkeley Shenzhen Institute \(TBSI\)](#)

Thesis: *Information Bottleneck for Representation Learning: New Vision*

Co-advised by: [Prof. Shao-Lun Huang](#), TBSI and [Prof. Khalid M. Mosalam](#), UC-Berkeley

Shenzhen, China

Sept. 2018-Jun. 2021

### Tongji University

B.Eng., Structural Engineering, School of Civil Engineering

GPA: 4.4/5.0 (19/168); Advised by: [Prof. Suzhen Li](#)

Shanghai, China

Sept. 2014-Jun. 2018

## PAPERS

---

### ◇ Conferences:

- **Z Wang**, Y Yang, R Wen, X Chen, S-L Huang, and Y Zheng. *Lifelong Learning Disease Diagnosis on Clinical Notes*. **PAKDD'21 (Best Student Paper, 2/768)**. [\[pdf\]](#) [\[video\]](#)
- **Z Wang**, R Wen, X Chen, S Cao, S-L Huang, B Qian, and Y Zheng. *Online Disease Self-diagnosis with Inductive Heterogeneous Graph Convolutional Networks*. **WWW'21**. [\[pdf\]](#) [\[video\]](#)
- **Z Wang**, X Chen, R Wen, S-L Huang, E. E. Kuruoglu, and Y Zheng. *Information Theoretic Counterfactual Learning from Missing-Not-At-Random Feedback*. **NeurIPS'20**. [\[pdf\]](#) [\[code\]](#) [\[poster\]](#)
- **Z Wang**, H Zhu, Z Dong, X He, and S-L Huang. *Less Is Better: Unweighted Data Subsampling via Influence Function*. **AAAI'20**. [\[pdf\]](#)[\[code\]](#) [\[poster\]](#)
- **Z Wang**, S-L Huang, E. E. Kuruoglu, J Sun, X Chen, and Y Zheng. *PAC-Bayes Information Bottleneck*. **Under Review**.
- **Z Wang**, R Wen, X Chen, S-L Huang, N Zhang, and Y Zheng. *Finding Influential Instances for Distantly Supervised Relation Extraction*. **Under review**. [\[pdf\]](#) [\[code\]](#)

### ◇ Journals:

- **Z Wang**, Y Zhang, K. M. Mosalam, Y Gao, and S-L Huang. *Deep Semantic Segmentation for Visual Understanding on Construction Sites*. **Computer-Aided Civil And Infrastructure Engineering**, 2021. [\[pdf\]](#)
- **Z Wang & S Li**. *Data-driven Risk Assessment on Urban Pipeline Network Based on a Cluster Model*. **Reliability Engineering and System Safety**, 2020, 196: 106781. [\[pdf\]](#)

## RESEARCH EXPERIENCE

---

### Jarvis Lab, Tencent

Research intern in machine learning and NLP

Shenzhen, China

Dec. 2019-Present

### ◇ Information Principled Representation Learning (Working paper):

- It has been identified in the literature that mutual information between network weights and dataset controls the PAC-Bayes generalization error bound of neural networks. However, optimizing this mutual information is generally intractable.
- We model the dataset sampling process as Bootstrap resampling, then take an infinitesimal analysis on the covariance of weight distribution to derive a closed-form solution of this mutual information term.
- We identify the generalization capacity is connected to geometry, i.e., the Fisher information on the local minima, and derive an information principled deep learning framework.

### ◇ Information-theoretic Counterfactual Learning (Published in NeurIPS'20):

- Items are ranked and displayed via a policy in recommender systems, causing the feedback missing-not-at-random (MNAR). Previous works need to collect missing at random data (called randomized controlled trials) to debias learning.
- Inspired by information bottleneck's application for unsupervised learning, we derive a novel solution of IB.
- Our method can balance the label information contained in factual and counterfactual event embeddings. Moreover, it can learn from both factual and counterfactual data w/o randomized controlled trials.

### ◇ Robust ML on Noisy Data (Submitted to AAAI'21):

- IFS proposed in our AAAI'20 paper has high computational complexity, therefore we derive an  $\mathcal{O}(1)$  complexity approximation to apply it to deep learning models.
- We apply the DL-IFS to distant supervision relation extraction to sample favorable instances efficiently.

◇ **ML & NLP for Healthcare** (Submitted to WWW'21 (2)):

- Previous works usually leverage sequential patient visit data by RNN to predict disease risk, while on web-based disease diagnosis, most users are cold-start who do not have historical visits.
- We propose to use inductive heterogeneous GCN to mine relations between users for precise diagnosis, and handle cold-start users.
- Governance of clinical data is strict so we cannot maintain too much. Besides, disease distribution varies spatiotemporally.
- However, common ML models confront catastrophic forgetting when finetuned on new data.
- We propose a novel continual learning diagnosis model, using medical domain knowledge and embedding consolidation to achieve knowledge transfer and retention.

Noah's Ark Lab, Huawei

Research intern in machine learning and recommender systems

Shenzhen, China

Apr. 2019-Oct. 2019

◇ **Robust ML by Subsampling from Noisy Data** (Published in AAAI'20):

- We measure sample quality by expected error reduction by influence function (IF), aiming for interpretable data selection when learning from noisy data.
- As deterministic selection via IF usually fails, we propose to do probabilistic sampling based on influence function (IFS). We prove IFS is more robust because it can resist performance decay from distribution shift.

◇ **Counterfactual Learning for Improving Ad-click Rates:**

- The collected user feedback by recommendation policy is missing-not-at-random (MNAR), the learned model on MNAR data cannot give fair and diverse recommendation.
- To overcome this selection bias, we propose a propensity free doubly robust method where the direct method part is learned from uniformly displayed feedback data.

## TEACHING

- |  |              |
|--|--------------|
| • TA, <a href="#">Optimization Models and Applications</a> (32 hrs), <a href="#">Prof. Laurent El Ghaoui</a> | Summer, 2020 |
| • TA, <a href="#">Bayesian Learning and Data Analysis</a> (32 hrs), <a href="#">Prof. Ercan E. Kuruoglu</a>  | Spring, 2020 |
| • TA, <a href="#">Learning from Data</a> (48 hrs), Prof. Shao-Lun Huang and <a href="#">Prof. Yang Li</a>    | Fall, 2019   |

## AWARDS & ACHIEVEMENTS & OTHERS

- |  |                |
|--|----------------|
| • Best Student Research Runner-up of 13rd PhD Student Symposium of Guangdong-HK-Macau Bay Area | June 2021      |
| • Best Student Paper Award of PAKDD'21 (2/768)   | May 2021       |
| • National Graduate Student Scholarship at Tsinghua University (3/229)                         | Oct. 2020      |
| • Best Student Research Runner-up of 1st TBSI Workshop On Data Science                         | Dec. 2019      |
| • Outstanding graduate student (4/40), graduate thesis (3/168) of Tongji University            | Jun. 2018      |
| • Merit student scholarship of Tongji University   | 2015/2016/2017 |
| • Meritorious winner (1st class prize, $\approx 7\%$ ) in USA Mathematical Contest in Modeling | Apr. 2017      |

## SKILLS & CERTIFICATION

- English: TOEFL (105), IELTS (7.0), CET-6 (615),
- IT: Linux, Python, C++ and Python packages including Pytorch, Tensorflow, Numpy, Scipy, Pandas, Sklearn, keras, etc.
- Hobbies: Bamboo flute, Hulusi.