# Ke Zhai

+1 (240)-460-9082 · zhaikedavy@gmail.com · kzhai.github.io

Machine Learning · Cloud Computing · Natural Language Processing · Audio Processing and Modeling


## EDUCATION

**Ph.D. in Computer Science** University of Maryland, College Park, MD, 2014

GPA: 3.9/4.0. Supervisor: Jordan Boyd-Graber

Research topic: Large Scale Inference for Probabilistic Bayesian Models

Ph.D. thesis: Models, Inference, and Implementation for Scalable Probabilistic Models of Text

**M.Sc in Computer Science** University of Maryland, College Park, MD, 2011

GPA: 3.9/4.0. Supervisor: Jordan Boyd-Graber and Jimmy Lin (co-supervied)

Research topic: Variational Bayesian Inference of Latent Dirichlet Allocation in MapReduce

Master scholary paper: Using Variational Inference and MapReduce to Scale Topic Modeling

**B.Eng. in Computer Engineering** Nanyang Technological University, Singapore, 2009

GPA: 4.65/5.0 with first class honor. Supervisor: Wee Keong Ng

Research topic: Privacy-Preserving Data Mining

Undergraduate thesis: An Embedded Caching Framework for Privacy-Preserving Data Mining

## EMPLOYMENT

**Staff Research Scientist** Responsible AI Team, Apple, Inc., Cupertino, CA, Oct 2023 - Present

- Tech lead for red teaming efforts for Apple Intelligence features, including web QA, text summarization, text composition, text assistants and visual generation features.
- Infrastructure archtect for building large-scale pipelines for automatic red-teaming data synthesization, on-device inference/simulation, and auto grading metric reporting pipelines.
- Tech lead for developing embedding based semantic override safety models.

**Staff Research Scientist** Machine Intelligence Sensing Team, Apple, Inc., Cupertino, CA, Jul 2021 - Oct 2023

- Tech lead for improving handwashing feature for Apple watch. Using audio and motion sensing to improve recall and precisions for handwashing.
- Tech lead for double tap gesture detection for Apple watch, featured in Sep 2023 Apple special event.
- On-device audio detection model for health event for Airpods.
- Location tracking and prediction algorithm based on IMU sensing to optimize energy and battery for Vision Pro.

**Founding Member** Dawnlight, Inc., Palo Alto, CA, Jan 2019 - Jul 2021

- Design and develop end-to-end data collection, signal processing, model training and serving architecture (in C++/C) on various IoT edge devices for healthcare related audio event detection.
- Design and develop state-of-the-art audio event detection model (benchmarked on multiple public datasets).
- Design and develop end-to-end data processing, model training and serving architecture for Clincal Decision Support System (CDSS).
- Design and develop end-to-end data processing, model training and feature demo on activity recognition using radar signals and RFID signals.
- Working with a team to design and develop a generic end-to-end machine learning platform with hyper-parameter tuning (using Ray) and container orchestration pipeline support.

- Working with a team to develop and maintain full-stack daily operations, including flask services, bluetooth services, docker containers, Kafka, gRPC, etc.

**Senior Research Scientist** Microsoft, Sunnyvale, CA, Aug 2016 - Jan 2019

- Language model personalization using topic models, context matching and style filtering.
- Design and develop end-to-end data processing, model training and adaptation evaluation pipeline.
- Language model adaptation support for Microsoft Teams product for both meeting and broadcasting scenarios. Prototype demo'ed at Microsoft *Build 2018*, *Inspire 2018*, and *Ignite 2018*.

**Research Scientist** Yahoo! Labs, Sunnyvale, CA, Feb 2015 - Aug 2016

- Develop query understanding and sequence tagging models for online Ads serving system.
- Design and develop query classification and user intent prediction pipeline for mobile search.
- Research, design and develop on "Chat Bot as a Service" platform.

**Graduate Research Assistant** Department of Computer Science, University of Maryland, College Park, MD, Sep 2010 - Jun 2014

- Academic Advisor: Jordan Boyd-Graber
- Research Interest: Machine Learning, Non-parametric Bayesian Learning, Cloud Computing
- Design and implement online variational inference for adaptor grammars.
- Design and implement online variational inference for topic models with infinite vocabulary.
- Design and implement variational inference for latent Dirichlet allocation in MapReduce.
- Design and implement variational inference for Indian buffet process in MapReduce.

**Research Intern** Yahoo! Labs, New York City, NY, Jun 2014 - Aug 2014

- Large scale unsupervised nonparametric models for user behavior analysis.

**Research Intern** Microsoft Research, Redmond, WA, May 2013 - Aug 2013

- Mentor: Jason D. Williams
- Design and implement three models in discovering latent structure in dialogues.
- Achieve comparably well results against many other models on real datasets.

**Software Engineering Intern** comScore, Inc., Reston, VA, May 2010 - Aug 2010

- Implement data transfer and formatter block for new deployed Hadoop distributed file system.
- Research and develop cookie deletion and prediction system for comScore.

## PUBLICATION (* indicates equal contributor)

**Ke Zhai***, and Huan Wang*. "Adaptive Dropout with Rademacher Complexity Regularization". *International Conference on Learning Representations* (ICLR), May 2018.

**Ke Zhai**, Zornitsa Kozareva, Yuening Hu, Qi Li and Weiwei Guo. "Query to Knowledge: Unsupervised Entity Extraction from Shopping Queries using Adaptor Grammars". *International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), Jul 2016.

Zornitsa Kozareva, Qi Li, **Ke Zhai** and Weiwei Guo. "Recognizing Salient Entities in Shopping Queries". *Annual Meeting of the Association for Computational Linguistics* (ACL), Jun 2016.

**Ke Zhai**, Jordan Boyd-Graber and Shay B. Cohen. "Online Adaptor Grammars with Hybrid Inference". *Transaction of the Association for Computational Linguistics* (TACL), Oct 2014.

**Ke Zhai**, and Jason D. Williams. "Discovering Latent Structure in Task-Oriented Dialogues". *Annual Meeting of the Association for Computational Linguistics* (ACL), Jun 2014.

**Ke Zhai***, Yuening Hu*, Vladimir Edelman, and Jordan Boyd-Graber. "Polylingual Tree-Based Topic Models for Translation Domain Adaptation". *Annual Meeting of the Association for Computational Linguistics* (ACL), Jun 2014.

**Ke Zhai**, and Jordan Boyd-Graber. "Online Latent Dirichlet Allocation with Infinite Vocabulary". *International Conference on Machine Learning* (ICML), Jun 2013.

**Ke Zhai***, Yuening Hu*, Jordan Boyd-Graber, and Sinead Williamson. "Modeling Images using Transformed Indian Buffet Processes". *International Conference on Machine Learning* (ICML), Jun 2012.

**Ke Zhai**, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. "Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce". *ACM International Conference on World Wide Web* (WWW), Apr 2012.

**Ke Zhai**, Wee Keong Ng, Andre Ricardo Herianto and Shuguo Han. "Speeding Up Secure Computations via Embedded Caching". *Proceedings of SIAM International Conference on Data Mining* (SDM), Apr 2009.

## PROFESSIONAL CONTRIBUTION

Online released code-base on `github.com/kzhai`.

- Activate contributor for MapReduce library Cloud[9] and Hadoop toolkit Ivory.
- MapReduce latent Dirichlet allocation (Variational Beyesian inference, with extension to informed prior and polylingual LDA).

> "I tried both and found that, despite *Mr. LDA*'s cringeworthy name, it's much faster than *Mahout*'s implementation so decided to go with that one."
>
> — **Kris Jack**, BSc Hons, Ph.D., Chief Data Scientist, Mendeley

- Latent Dirichlet allocation (Gibbs sampling, variational Bayesian inference and online version).
- Non-parametric Bayesian models (Indian buffet process, hierarchical Dirichlet process and infinite Gaussian mixture model).
- Customizable deep learning toolkits using Theano and PyTorch.