Alex Beutel

alex@beu.tel
alexbeutel.com ·

Summary

I strive to develop new machine learning and algorithmic techniques that meaningfully improve user experiences. I have worked most significantly on user modeling (personalized recommenders, fraud detection) and responsible ML (e.g., fairness, robustness, safety), intersecting at times with RL, NLP, vision, and graph mining.

Industry Experience

OpenAI, *Member of Technical Staff, Tech Lead, Manager* June 2023–Present

I lead (TLM) our Safety Research teams and efforts, spanning safety training, robustness, safety reasoning, and interpretability, all with a goal of improving safety and alignment of OpenAI's models for real users. I have a particular focus right now on robustness and agents.

Google, *Research*

Senior Staff Research Scientist, Tech Lead, Manager April 2021–June 2023
Staff Research Scientist, Tech Lead, Manager April 2019–April 2021
Senior Research Scientist Oct. 2017–April 2019
Research Scientist Aug. 2016–Oct. 2017
Research Intern May 2015–Aug. 2015

I was the research lead for the Responsible ML team in Google Research. I led impactful projects through a combination of foundational research and close collaborations with product teams. My approach was effective in driving new understanding (see papers below) and benefiting users, with > 50 launches. My work spanned:

- **ML Fairness:** Developing reliable, flexible and easy-to-use approaches for measuring and mitigating fairness issues in policy enforcement classifiers, leading basic research, first-of-kind launches, and supporting scaling across multiple products.
- **Responsible Recommendation and Ranking:** Improving fairness for item producers, information quality and diversity. This progress comes from research on unbiased offline evaluation, compositionality, reinforcement learning and simulation.
- **Robustness and Safety in NLP and Vision:** Most recently, Safety research lead for Google Bard. More generally, addressing real-world adversarial attacks and improving safety, based on research on spurious correlations and general robustness.

In addition to technical execution, I was responsible for:

- **Team management, growth, and mentorship:** Co-founded the team, grew it to > 15 researchers and engineers (plus product and program managers), and continued to directly manage approximately half.
- **Team vision and strategy:** Proposed and built alignment on long-term team vision and multi-year research strategies in each area above.
- **Cross-organizational and cross-functional alignment** with executives across Research and multiple products to ensure our research addressed critical product needs.

Beyond responsible ML, I have driven research and product engagements spanning:

- **Interactive Recommendation:** led research for YouTube's first neural sequential recommender and novel off-policy RL for recommendation.
- **Learned Indexes:** using ML to learn data distributions to speed up database indexes

Microsoft, *Cloud and Information Service Lab*, Intern June 2014–Aug. 2014

Researched distributed training of recommender systems using probabilistic programming.

Facebook, Intern May 2012–Aug. 2013, May 2013–Aug. 2013

- Detected synchronized attacks (fake Page Likes) with novel graph clustering.
- News Feed information quality and content spread

Education

Carnegie Mellon University August 2011–May 2016
Ph.D., Computer Science May 2016
Masters of Science, Computer Science December 2013

Thesis title: “Understanding User Behavior through Large-Scale Graph Analysis”
Committee: Christos Faloutsos, Alex Smola, Geoff Gordon, Phillip Yu

Duke University August 2007–May 2011
Bachelor of Science, Quantitative Studies in Computer Science and Physics
GPA: 3.858/4.0; Dean’s List (FA08, FA09) with Distinction (FA07, SP08, SP10)
Graduated *Magna cum Laude* and with *Highest Distinction* in Computer Science

Honors

Best Paper Award, *CIKM Workshop on Human-in-the-loop Data Curation*, 2022

SIGKDD Doctoral Dissertation Award Runner-up, 2017

Best Paper Award, *ACM KDD* 2016

Best Paper Finalist, *ACM KDD* 2014

Facebook Graduate Fellowship, 2013

Phi Beta Kappa Honor Society, 2012

NSF Graduate Research Fellowship, 2011

Alex Vasilos Memorial Award, Duke University Computer Science, 2011

Best Paper Award, *ACM GIS* 2010

Computer Science Undergraduate Research Fellow, Duke University 2010

Refereed Conference Papers

C51. **First-Person Fairness in Chatbots**
Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, Adam Tauman Kalai. *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

C50. **Rule Based Rewards for Language Model Safety**
Tong Mu*, Alec Helyar*, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, Lilian Weng. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

C49. **Generalized People Diversity: Learning a Human Perception-Aligned Diversity Representation for People Images**
Hansa Srinivasan, Candice Schumann, Aradhana Sinha, David Madras, Gbolahan Oluwafemi Olanubi, Alex Beutel, Susanna Ricco, Jilin Chen. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2024.

C48. **Effective Robustness against Natural Distribution Shifts for Models with Different Training Data**
Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, Yao Qin. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

C47. **Controlled Decoding from Language Models**
Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Jilin Chen, Alex Beutel, Ahmad Beirami. *International Conference on Machine Learning (ICML)*, 2024.

C46. **Improving Diversity of Representation in Large Language Models via Collective-Critiques and Self-Voting (CCSV)**
Preethi Lahoti, Nick Blumm, Xiao Ma, Ragha Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ahmad Beirami, Ben Packer, Alex Beutel, Jilin Chen. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

C45. **Improving Classifier Robustness through Active Generation of Pairwise Counterfactuals**
Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Jilin Chen,

Ed H. Chi, Alex Beutel. *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2023.

C44. **Learning From Negative User Feedback and Measuring Responsiveness for Sequential Recommenders**
Yueqi Wang, Yoni Halpern, Shuo Chang, Jingchen Feng, Elaine Ya Le, Longfei Li, Xujian Liang, Min-Cheng Huang, Shane Li, Alex Beutel, Yaping Zhang, Shuchao Bi. *17th ACM Conference on Recommender Systems (RecSys), Industry Track*, 2023.

C43. **What are effective labels for augmented data? Improving robustness with AutoLabel**
Yao Qin, Xuezhi Wang, Balaji Lakshminarayanan, Ed H. Chi, Alex Beutel. *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

C42. **Understanding and Improving Robustness of Vision Transformers through Patch-based Negative Augmentation**
Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, Xuezhi Wang. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

C41. **Improving Calibration through the Relationship with Adversarial Robustness**
Yao Qin, Xuezhi Wang, Alex Beutel, Ed H. Chi. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

C40. **Can We Improve Model Robustness through Secondary Attribute Counterfactuals?**
Ananth Balashankar, Xuezhi Wang, Ben Packer, Nithum Thain, Ed H. Chi, Alex Beutel. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

C39. **Understanding and Improving Fairness-Accuracy Trade-offs in Multi-task Learning**
Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, Ed H. Chi. *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

C38. **Measuring Model Fairness under Noisy Covariates: A Theoretical Perspective**
Flavien Prost, Pranjal Awasthi, Nick Blumm, Aditee Kumthekar, Trevor Potter, Li Wei, Xuezhi Wang, Ed H. Chi, Jilin Chen, Alex Beutel. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2021.

C37. **Towards Content Provider Aware Recommender Systems: A Simulation Study on the Interplay between User and Provider Utilities**
Ruohan Zhan, Konstantina Christakopoulou, Elaine Le, Jayden Ooi, Martin Mladenov, Alex Beutel, Craig Boutilier, Ed H. Chi, Minmin Chen. *TheWebConf*, 2021.

C36. **Evaluating Fairness of Machine Learning Models Under Uncertain and Incomplete Information**
Pranjal Awasthi, Alex Beutel, Matthaus Kleindessner, Jamie Morganstern, Xuezhi Wang. *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

C35. **Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems**
Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, Alex Beutel. *Fourteenth ACM International Conference Web Search and Data Mining (WSDM)*, 2021.

C34. **Enhancing Neural Recommender Models through Domain-Specific Concordance**
Ananth Balashankar, Alex Beutel, Lakshminarayanan Subramanian. *Fourteenth ACM International Conference Web Search and Data Mining (WSDM)*, 2021.

C33. **Fairness without Demographics through Adversarially Reweighted Learning**
Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi

Wang, Ed H. Chi. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

C32. **CAT-Gen: Improving Robustness in NLP Models via Controlled Adversarial Text Generation**
Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, Ed H. Chi. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

C31. **Fairness in Recommendation Ranking through Pairwise Comparisons**
Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, Cristos Goodrow. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD Applied Data Science)*, 2019.

C30. **Towards Neural Mixture Recommender for Long Range Dependent User Sequences**
Jiaxi Tang, Francois Belletti, Sagar Jain, Minmin Chen, Alex Beutel, Can Xu, Ed H. Chi. *WWW 2019: The 2019 Web Conference*, 2019.

C29. **Top-K Off-Policy Correction for a REINFORCE Recommender System**
Minmin Chen*, Alex Beutel*, Paul Covington*, Sagar Jain, Francois Belletti, Ed H. Chi. *Twelfth ACM International Conference Web Search and Data Mining (WSDM)*, 2019.

C28. **Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements**
Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019.

C27. **Counterfactual Fairness in Text Classification through Robustness**
Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, Alex Beutel. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019.

C26. **SageDB: A Learned Database System**
Tim Kraska, Mohammad Alizadeh, Alex Beutel, Ed H. Chi, Jialin Ding, Ani Kristo, Guillaume Leclerc, Samuel Madden, Hongzi Mao, Vikram Nathan. *Ninth Biennial Conference on Innovative Data Systems Research (CIDR)*, 2019.

C25. **Categorical-Attributes-Based Item Classification for Recommender Systems**
Qian Zhao, Jilin Chen, Minmin Chen, Sagar Jain, Alex Beutel, Francois Belletti, Ed H. Chi. *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, 2018.

C24. **Q&R: A Two-Stage Approach Toward Interactive Recommendation**
Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, Ed H. Chi. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD Applied Data Science)*, 2018.

C23. **The Case for Learned Index Structures**
Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, Neoklis Polyzotis. *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2018.

C22. **Factorized Recurrent Neural Architectures for Longer Range Dependence**
Francois Belletti, Alex Beutel, Sagar Jain, Ed H. Chi. *21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

C21. **Latent Cross: Making Use of Context in Recurrent Recommender Systems**
Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, Ed H. Chi. *Eleventh ACM International Conference Web Search and Data Mining (WSDM)*, 2018.

C20. **The Many Faces of Link Fraud**
Neil Shah, Hemank Lamba, Alex Beutel, Christos Faloutsos. *IEEE International Conference on Data Mining (ICDM)*, 2017.

C19. **Beyond Globally Optimal: Focused Learning for Improved Recommendations**
Alex Beutel, Ed H. Chi, Derek Zhiyuan Cheng, Hubert Pham, John Anderson. *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 2017.

C18. **Recurrent Recommender Networks**
Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alex Smola, How Jing. *Tenth ACM International Conference Web Search and Data Mining (WSDM)*, 2017.

C17. **FRAUDAR: Bounding Graph Fraud in the Face of Camouflage**
Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, Christos Faloutsos. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

C16. **BIRDNEST: Bayesian Inference for Ratings-Fraud Detection**
Bryan Hooi, Neil Shah, Alex Beutel, Stephan Gunnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, Christos Faloutsos. *2016 SIAM International Conference on Data Mining (SDM)*, 2016.

C15. **A General Suspiciousness Metric for Dense Blocks in Multimodal Data**
Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, Christos Faloutsos. *IEEE International Conference on Data Mining (ICDM)*, 2015.

C14. **ACCAMS: Additive Co-Clustering to Approximate Matrices Succinctly**
Alex Beutel, Amr Ahmed, Alexander J. Smola. *Proceedings of the 24th International Conference on World Wide Web (WWW)*, 2015.

C13. **ND-SYNC: Detecting Synchronized Fraud Activities**
Maria Giatsoglou, Despoina Chatzakou, Neil Shah, Alex Beutel, Christos Faloutsos, Athena Vakali. *19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2015.

C12. **Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective**
Neil Shah, Alex Beutel, Brian Gallagher, Christos Faloutsos. *IEEE International Conference on Data Mining (ICDM)*, 2014.

C11. **CatchSync: Catching Synchronized Behavior in Large Directed Graphs**
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2014.

C10. **Inferring Strange Behavior from Connectivity Pattern in Social Networks**
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. *18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2014.

C9. **Fugue: Slow-Worker-Agnostic Distributed Learning for Big Models**
Abhimanu Kumar, Alex Beutel, Qirong Ho, Eric P. Xing. *17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

C8. **FlexiFaCT: Scalable Flexible Factorization of Coupled Tensors on Hadoop**
Alex Beutel, Abhimanu Kumar, Evangelos E. Papalexakis, Partha Pratim Talukdar, Christos Faloutsos, Eric P. Xing. *2014 SIAM International Conference on Data Mining (SDM)*, 2014.

C7. **CoBaFi: Collaborative Bayesian Filtering**
Alex Beutel, Kenton Murray, Christos Faloutsos, Alexander J. Smola. *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, 2014.

C6. **CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks**
Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, Christos Faloutsos. *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 2013.

C5. **Network Anomaly Detection using Co-clustering**
Evangelos E. Papalexakis, Alex Beutel, Peter Steenkiste. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012.

C4. **Interacting Viruses on a Network: Can both survive?**
Alex Beutel, B. Aditya Prakash, Roni Rosenfeld, Christos Faloutsos. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

C3. **Winner-takes-all: Competing Viruses on fair-play networks**
B. Aditya Prakash, Alex Beutel, Roni Rosenfeld, Christos Faloutsos. *Proceedings of the 21st International Conference on World Wide Web (WWW)*, 2012.

C2. **TerraNNI: Natural Neighbor Interpolation on a 3D Grid Using a GPU**
Alex Beutel, Thomas Moelhave, Pankaj K. Agarwal, Arnold P. Boedihardjo, James A. Shine. *Proceedings of the 19th International Symposium on Advances in Geographic Information Systems (ACM GIS)*, 2011.

C1. **Natural Neighbor Interpolation Based Grid DEM Construction Using a GPU**
Alex Beutel, Thomas Moelhave, Pankaj K. Agarwal. *Proceedings of the 18th International Symposium on Advances in Geographic Information Systems (ACM GIS)*, 2010.

Refereed Workshop Papers

W21. **Diverse and Effective Red Teaming with Auto-generated Rewards and Multi-step Reinforcement Learning**
Alex Beutel, Kai Xiao, Johannes Heidecke, Lilian Weng. *NeurIPS 2024 workshop on Red Teaming GenAI: What Can We Learn from Adversaries?*, 2024.

W20. **Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning**
Xiao Ma, Swaroop Mishra, Ahmad Beirami, Alex Beutel, Jilin Chen. *Neural Conversational AI Workshop at ICML*, 2023.

W19. **Towards A Scalable Solution for Compositional Multi-Group Fair Classification**
James Atwood, Tina Tian, Ben Packer, Meghana Deodhar, Jilin Chen, Alex Beutel, Flavien Prost, Ahmad Beirami. *The Second Workshop on Spurious Correlations, Invariance and Stability at ICML*, 2023.

W18. **Striving for data-model efficiency: Identifying data externalities on group performance**
Esther Rolf, Ben Packer, Alex Beutel, Fernando Diaz. *Trustworthy and Socially Responsible Machine Learning (TSRML) workshop at NeurIPS*, 2022.

W17. **A Human-ML Collaboration Framework for Improving Video Content Reviews**
Meghana Deodhar, Xiao Ma, Yixin Cai, Alex Koes, Alex Beutel, Jilin Chen. *CIKM Workshop on Human-in-the-Loop Data Curation*, 2022.

W16. **Flexible text generation for counterfactual fairness probing**
Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, Kellie Webster. *Workshop on Online Abuse and Harms (WOAH) at ACL*, 2022.

W15. **Learned Indexes for a Google-scale Disk-based Database**
Hussam Abu-Libdeh, Deniz Altinbuken, Alex Beutel, Ed H. Chi, Lyric Doshi, Tim Kraska, Xiaozhou, Li, Andy Ly, Christopher Olston. *ML for Systems workshop at NeurIPS*, 2020.

W14. **Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach**
Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, Ed H. Chi, Jilin Chen, Alex Beutel. *FAccTRec*, 2020.

W13. **Learning to Diversify from Human Judgments: Research Directions and Open Challenges**
Emily Denton, Hansa Srinivasan, Dylan Baker, Jilin Chen, Alex Beutel, Tulsee Doshi, Ed H. Chi. *Fair and Responsible AI Workshop at CHI*, 2020.

W12. **Measuring Recommender System Effects with Simulated Users**
Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H. Chi, Jilin Chen, Alex Beutel. *FATES at WWW*, 2020.

W11. **Toward a better trade-off between performance and fairness with kernel-based distribution matching**
Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, Alex Beutel. *ML with Guarantees workshop at NeurIPS*, 2019.

W10. **Transfer of Machine Learning Fairness across Domains**
Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, Ed H. Chi. *AI for Social Good workshop at NeurIPS*, 2019.

W9. **Lifting the Curse of Multidimensional Data with Learned Existence Indexes**
Stephen Macke, Alex Beutel, Tim Kraska, Maheswaran Sathiamoorthy, Derek Zhiyuan Cheng, Ed H. Chi. *ML for Systems workshop at NeurIPS*, 2018.

W8. **Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations**
Alex Beutel, Jilin Chen, Zhe Zhao, Ed H. Chi. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

W7. **Joint Training of Ratings and Reviews with Recurrent Recommender Networks**
Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alex Smola. *Workshop track at 5th International Conference on Learning Representations (ICLR)*, 2017.

W6. **EdgeCentric: Anomaly Detection in Edge-Attributed Networks**
Neil Shah, Alex Beutel, Bryan Hooi, Leman Akoglu, Stephan Gunnemann, Disha Makhija, Mohit Kumar, Christos Faloutsos. *IEEE International Conference on Data Mining (ICDM) Workshop on Data Mining for Cyber Security*, 2016.

W5. **Additive Co-Clustering of Gaussians and Poissons for Joint Modeling of Ratings and Reviews**
Chao-Yuan Wu, Alex Beutel, Amr Ahmed, Alexander J. Smola. *NeurIPS workshop on Nonparametric Methods for Large Scale Representation Learning*, 2015.

W4. **Collaborative Bayesian Filtering: Patterns and Methods**
Alex Beutel, Kenton Murray, Christos Faloutsos, Alexander J. Smola. *Workshop on Information Networks (WIN)*, 2015.

W3. **Elastic Distributed Bayesian Collaborative Filtering**
Alex Beutel, Markus Weimer, Tom Minka, Yordan Zaykov, Vijay Narayanan. *NeurIPS Distributed Machine Learning and Matrix Computations workshop*, 2014.

W2. **FlexiFaCT: Scalable Flexible Factorization of Coupled Tensors on Hadoop**
Alex Beutel, Abhimanu Kumar, Evangelos E. Papalexakis, Partha Pratim Talukdar, Christos Faloutsos, Eric P. Xing. *NeurIPS Big Learning Workshop*, 2013.

W1. **Volumetric Grid Construction using 3D Natural Neighbor Interpolation on the GPU**
Alex Beutel, Thomas Moelhave, Pankaj K. Agarwal. *MASSIVE '11: Proceedings of the Workshop on Massive Data Algorithmics*, 2011.

Refereed Journal Papers

J8. **Break it, Imitate it, Fix it: Robustness by Generating Human-Like Attacks**
Aradhana Sinha, Ananth Balashankar, Ahmad Beirami, Thi Avrahami, Jilin Chen, Alex Beutel. *Transactions on Machine Learning Research (TMLR)*, 2024.

J7. **Multi-Group Fairness Evaluation via Conditional Value-at-Risk Testing**
Lucas Monteiro Paes, Ananda Theertha Suresh, Alex Beutel, Flavio P. Calmon, Ahmad Beirami. *IEEE Journal on Selected Areas in Information Theory*, 2024.

J6. **Underspecification Presents Challenges for Credibility in Modern Machine Learning**
Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, D. Sculley. *Journal of Machine Learning Research (JMLR)*, 2021.

J5. **Graph-Based Fraud Detection in the Face of Camouflage**
Bryan Hooi, Kijung Shin, Hyun Ah Song, Alex Beutel, Neil Shah, Christos Faloutsos. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2017.

J4. **Spotting Suspicious Behaviors in Multimodal Data: A General Metric and Algortihms**
Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, Christos Faloutsos. *Transactions on Knowledge and Data Engineering (TKDE)*, 2016.

J3. **Catching Synchronized Behaviors in Large Networks: A Graph Mining Approach**
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2016.

J2. **TerraNNI: Natural Neighbor Interpolation on 2D and 3D Grids using a GPU**
Pankaj K. Agarwal, Alex Beutel, Thomas Moelhave. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 2016.

J1. **Inferring Lockstep Behavior from Connectivity Pattern in Large Graphs**
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. *Knowledge and Information Systems (KAIS)*, 2015.

Book Chapters

Network Anomaly Detection using Co-clustering
Evangelos E. Papalexakis, Alex Beutel, Peter Steenkiste. *Springer Encyclopedia of Social Network Analysis and Mining*, 2012.

Other Papers

HealthBench: Evaluating Large Language Models Towards Improved Human Health
Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, Karan Singhal. *Preprint*, 2025.

Deliberative Alignment: Reasoning Enables Safer Language Models
Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, Amelia Glaese. *Preprint*, 2024.

The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions
Eric Wallace*, Kai Xiao*, Reimar Leike*, Lilian Weng, Johannes Heidecke, Alex Beutel. *Preprint*, 2024.

Practices for Governing Agentic AI Systems
Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie

Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, David G. Robinson. *White paper*, 2023.

Improving Few-shot Generalization of Safety Classifiers via Data Augmented Parameter-Efficient Fine-Tuning
Ananth Balashankar, Xiao Ma, Aradhana Sinha, Ahmad Beirami, Yao Qin, Jilin Chen, Alex Beutel. *Preprint*, 2023.

Towards Robust Prompts on Vision-Language Models
Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, Yao Qin. *Preprint*, 2023.

Simpson's Paradox in Recommender Fairness: Reconciling differences between per-user and aggregated evaluations
Flavien Prost, Ben Packer, Jilin Chen, Li Wei, Pierre Kremp, Nicholas Blumm, Susan Wang, Tulsee Doshi, Tonia Osadebe, Lukasz Heldt, Ed H. Chi, Alex Beutel. *Preprint*, 2022.

Measuring and Reducing Gendered Correlations in Pre-trained Models
Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, Slav Petrov. *Preprint*, 2020.

User Behavior Modeling with Large-Scale Graph Analysis
Alex Beutel. *Ph.D. Thesis, Carnegie Mellon University*, 2016.

User Behavior Modeling and Fraud Detection
Alex Beutel, Christos Faloutsos. *IEEE Intelligent Systems: Trends and Controversies*, 2016.

Explaining reviews and ratings with PACO: Poisson Additive Co-Clustering
Chao-Yuan Wu, Alex Beutel, Amr Ahmed, Alexander J. Smola. *Companion Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016.

Detecting Suspicious Following Behavior in Multimillion-Node Social Networks
Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang. *Companion Proceedings of the 23rd International Conference on World Wide Web (WWW)*, 2014.

From Point Cloud to 2D and 3D Grids: A Natural Neighbor Interpolation Algorithm using the GPU
Alex Beutel. *Senior Thesis - Graduation with Highest Distinction*, *Duke University*, 2011.

Tutorials

Responsible Recommendation and Search Systems
Alex Beutel, Ed H. Chi, Fernando Diaz, Robin Burke. *WWW*, 2020

Graph-Based User Behavior Modeling: From Prediction to Fraud Detection
Alex Beutel, Leman Akoglu, Christos Faloutsos. *KDD*, 2015

Fraud Detection through Graph-Based User Behavior Modeling
Alex Beutel, Leman Akoglu, Christos Faloutsos. *ACM CCS*, 2015

Keynotes

Practical Robustness
AACL 2023 Workshop on The ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI, Virtual, November 2023

Understanding and Improving Recommenders for All
KDD 2022 Workshop on Data Science and Artificial Intelligence for Responsible Recommendations (DS4RRS), Washington D.C., August 2022

Building and Understanding Recommenders for Long-Term User Experiences
2nd International Workshop on Online and Adaptive Recommender Systems at KDD, Washington D.C., August 2022

Understanding Recommendations over Time
SIGIR'20 Workshop on Deep Reinforcement Learning for Information Retrieval, Zoom, July 2020

Challenges and Progress in Scaling ML Fairness

AISys at SOSP, Huntsvilla, Ontario, Canada, October 2019

Dynamics and Context in Neural Recommender Systems
LearnIR Workshop at WSDM, Los Angeles, CA, February 2018

Invited Talks

Toward Safe and Robust AGI
2024 Netflix Workshop on Personalization, Recommendation and Search (PRS), Los Gatos, CA, May 2024

Practical Robustness
NeurIPS 2023 R0-FoMo Workshop: Robustness of Few-shot and Zero-shot Learning in Foundation Models, New Orleans, LA, December 2023

Building ML for All
Duke University, Distinguish CS Alumni Lecture, Durham, NC, March 2023

Building and Understanding Recommenders for Long-Term User Experiences
2021 SIGIR Workshop On eCommerce, Zoom, July 2021

Building and Understanding Recommenders for Long-Term User Experiences
Twitter, Zoom, May 2021

Building and Understanding Recommenders for Long-Term User Experiences
Spotify, Zoom, April 2021

Fairness in Recommendation
Netflix, Los Gatos, CA, November 2019

Putting Fairness Principles into Practice
Salesforce Research, Palo Alto, CA, August 2019

Learned Data Systems
QCon, New York, NY, June 2019

Putting Fairness Principles into Practice
University of California at Riverside, Riverside, CA, May 2019

Putting Fairness Principles into Practice
QCon.ai, San Francisco, CA, April 2019

ML for Data Systems
Stanford EE380 Colloqium, Palo Alto, CA, October 2018

Dynamics and Context in Neural Recommender Systems
Pinterest, San Francisco, CA, February 2018

Using Context when Modeling User Behavior: Improving Fraud Detection, Neural Recommenders, and Fairness
M.I.T., Cambridge, MA, November 2017

Using Context when Modeling User Behavior: Improving Fraud Detection, Neural Recommenders, and Fairness
Brown University, Providence, RI, November 2017

Beyond Globally Optimal: Focused Learning for Improved Recommendations
Google Student Research Summit, Mountain View, CA, September 2017

ACCAMS: Additive Co-Clustering to Approximate Matrices Succinctly
University of Pennsylvania, Philadelphia, PA, November 2015

Distributed Machine Learning for User Behavior Modeling
Facebook, New York, NY, May 2015

Distributed Machine Learning for User Behavior Modeling
Google Research, New York, NY, May 2015

SGD on Hadoop for Big Data and Huge Models
Duke University, Durham, NC, 2014

Other Talks

Measuring Recommender System Effects with Simulated Users

FATES, Zoom, April 2020

Fairness in Recommendation Ranking through Pairwise Comparisons
FACTS-IR, Paris, FR, July 2019

Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements
AIES, Honolulu, HI, January 2019

Q&R: A Two-Stage Approach Toward Interactive Recommendation
KDD, London, UK, August 2018

Latent Cross: Making Use of Context in Recurrent Recommender Systems
WSDM, Los Angeles, CA, February 2018

A Machine Learning Approach to Databases Indexes
ML Systems at NeurIPS, Long Beach, CA, December 2017

Beyond Globally Optimal: Focused Learning for Improved Recommendations
WWW, Perth, Australia, April 2017

Beyond Who and What: Answering How and Why by Modeling Large Graphs
Northeastern University, Boston, MA, March 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
Arnhold Institute for Global Health, Mount Sinai School of Medicine, New York, NY, March 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
IOMS, Stern School of Business, New York University, New York, NY, March 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
Google Research, Mountain View, CA, March 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
Microsoft, Redmond, WA, March 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
Georgia Institute of Technology, Atlanta, GA, February 2016

Beyond Who and What: Answering How and Why by Modeling Large Graphs
New York University, Courant Institute, New York, NY, February 2016

Collaborative Bayesian Filtering: Patterns and Methods
WIN, New York, NY, October 2015

ACCAMS: Additive Co-Clustering to Approximate Matrices Succinctly
WWW, Florence, Italy, May 2015

CoBaFi: Collaborative Bayesian Filtering
WWW, Seoul, South Korea, April 2014

CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks
WWW, Rio de Janeiro, Brazil, May 2013

Interacting Viruses on a Network: Can both survive?
KDD, Beijing, China, August 2012

TerraNNI: Natural Neighbor Interpolation on a 3D Grid Using a GPU
ACM GIS, Chicago, IL, November 2011

Natural Neighbor Interpolation Based Grid DEM Construction Using a GPU
ACM GIS, San Jose, CA, November 2010

Teaching Experience

Guest Lecture: Data Mining (Penn State IST557) **Fall 2019**
"Putting Fairness Principles into Practice"

Guest Lecture: Intro to Data & Computational Science (Brown DATA 1030) **Fall 2017**
"Building Blocks of Neural Networks and Research Applications of RNNs"

Guest Lecture: Machine Learning with Large Datasets (CMU 10-805) **Spring 2015**
"SGD on Hadoop for Big Data and Huge Models"

Guest Lecture: Machine Learning with Large Datasets (CMU 10-805) **Spring 2014**
"SGD on Hadoop for Big Data and Huge Models"

Teaching Assistant: Database Applications (CMU 15-415/615) **Spring 2014**

Teaching Assistant: Multimedia DB & Data Mining (CMU 15-826) **Fall 2013**

Patents

Blockwise Controlled Decoding of Natural Language (NL) Based Output Generated using a Large Language Model (LLM) to Reduce Latency in Rendering Thereof, Sidharth Mudgal, Ahmad Beirami, Jilin Chen, Alex Beutel, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Trevor Strohman. Patent Application 18/225,990.

Systems and Methods for Performing Automatic Label Smoothing of Augmented Training Data, Yao Qin, Alex Beutel, Ed Huai-Hsin Chi, Xuezhi Wang, Balaji Lakshminarayanan. Patent Application 17/493,228.

Elastic multi-resolution model-serving to compute inferences, Christopher Olston, Noah Fiedel, Ed H. Chi, Alexander Beutel. Defensive Publication 668.

Detection of Lockstep Behavior, Alex Beutel and Wanhong Xu.
Patent number 9077744; issued July 7, 2015.

Students Mentored and Advised

9. Ashudeep Singh (2020, Cornell)
8. Ananth Balashankar (2019-2021, NYU)
7. Preethi Lahoti (2019, MPI)
6. Sirui Yao (2019, Virginia Tech)
5. Sahaj Garg (2018, Stanford undergraduate; next position: Luminous Computing)
4. Candice Schumann (2018, UMD; next position: Google Research)
3. Stephen Macke (2018, UIUC; next position: Facebook)
2. Konstantina Christakopoulou (2017, UMN; next position: Google Research)
1. Francois Belletti (2017, UC Berkeley; next position: Google Research)

Service

KDD Sponsorship Co-chair, *KDD* 2023

KDD Cup Co-chair, *KDD* 2021

Co-organizer: Workshop on Deep Reinforcement Learning for Information Retrieval, *SIGIR* 2020

Co-organizer: Workshop on Deep Reinforcement Learning, *KDD* 2019

Co-organizer: Workshop on Machine Learning Systems, *NeruIPS* 2015

Senior PC: *WWW* 2022

Senior PC: *SDM* 2022

Senior PC: *CIKM* 2021

Area Chair: *NeurIPS Datasets and Benchmarks track* 2021

PC Member: ***KDD*** 2017, 2018, 2019, 2020

PC Member: ***WSDM*** 2018, 2019, 2020, 2021, 2022

PC Member: ***WWW*** 2017, 2018, 2019, 2020, 2021

PC Member: ***SDM*** 2017, 2018, 2019

PC Member/Reviewer: ***FAccT*** 2019, 2021

SPC Member: ***IJCAI*** 2019

PC Member: ***SIGMOD*** 2019, 2020

PC Member: ***ORSUM*** 2018

PC Member: ***SocInfo*** 2016

PC Member: ***IEEE DSAA*** **Special Session on Big Behavioral Data Analytics** 2016

PC Member: ***ACM/IEEE ASONAM*** 2016

PhD Forum Committee, *ICDM* 2015

Mentor at Doctoral Consortium, *WSDM* 2018

PC Member: Special Session on Big Behavioral Data Analytics, *IEEE DSAA* 2015

PC Member: Web Information System Engineering (WISE), 2014

PC Member: Diffusion Networks and Cascade Analytics Workshop, *WSDM* 2014

Reviewer: ***TKDD, TKDE, NeruIPS, ICML, ICLR, INFORMS Journal on Computing, Neurocomputing***, *UMUI*

Further Academic Experience

Carnegie Mellon University, Computer Science Department Sept. 2011–Aug. 2016
Advised by Professor Christos Faloutsos and Professor Alex Smola.
My research focused on large-scale user behavior modeling, including fraud detection, recommendation systems, and scalable machine learning.

Duke University, Department of Computer Science Jan. 2010–Aug. 2011
Research assistant for Prof. Pankaj K. Agarwal in computational geometry.

Duke University, Department of Computer Science Oct. 2009–Dec. 2010
Research assistant for Prof. Xiaowei Yang in networks and distributed systems.

Duke University, Department of Physics April 2009–Aug. 2009
Research assistant for Prof. Chris Walter in the high energy physics, neutrino group

Academic Funding Awards

Facebook Graduate Fellowship, 2013–2014	**$79,202**
NSF Graduate Research Fellowship, 2011–2016	**$132,000**
Yahoo! Faculty Research and Engagement Program, 2014 Aided Professor Christos Faloutsos in writing the research proposal	**$10,000**
NSF Collaborative Grant (Award No. IIS-1408924), 2014 Helped multiple professors with the research proposal	**$307,908**
ACM GIS Student Travel Grant Award, 2011	**$1,000**