

Dylan Z. Slack

Curriculum Vitae: August 3, 2022
Website: <https://dylanslacks.website>

Email: dslack@uci.edu

| | |
|---|--|
| Education | University of California - Irvine , Irvine, CA <i>Ph.D. Computer Science</i> Advisors: Sameer Singh & Himabindu Lakkaraju (@ Harvard University) 2019 - 2023 Haverford College , Haverford, PA <i>B.S. Computer Science with High Honors</i> Magna Cum Laude Advisor: Sorelle Friedler 2015 - 2019 |
| Research and Industry Experience | University of California - Irvine Sep. 2019 - Present Graduate Student Researcher (UCI NLP, UCI CREATE, HPI Institute) <i>Advised by:</i> Sameer Singh Google AI Jun. 2021 - Sep. 2021 Research Intern <i>Advised by:</i> Nevan Wickers & Yinlam Chow & Bo Dai Amazon Web Services (AWS) Jun. 2020 - Sep. 2020 Applied Scientist Intern <i>Advised by:</i> Krishnaram Kenthapadi & Nathalie Rauschmayr Haverford College Sep. 2017 - Aug. 2019 Research Assistant, Department of Computer Science <i>Advised by:</i> Sorelle Friedler |
| Awards | NeurIPS Outstanding Reviewer, 2021 ICLR Outstanding Reviewer, 2021 Hasso Plattner Institute Fellow, 2021 Ambler Scholar, 2019 |
| Preprints | TalkToModel: Understanding Machine Learning Models With Open Ended Dialogues Dylan Slack , Satyapriya Krishna, Hima Lakkaraju*, and Sameer Singh* <i>arXiv, 2022</i> Rethinking Explainability as a Dialogue: A Practitioner’s Perspective Himabindu Lakkaraju*, Dylan Slack* , Yuxin Chen, Chenhao Tan, and Sameer Singh <i>In Submission, 2022</i> SAFER: Data-Efficient and Safe Reinforcement Learning Through Skill Acquisition Dylan Slack , Yinlam Chow, Bo Dai, and Nevan Wickers <i>In Submission; DARL @ ICML, 2022</i> |
| Referred Publications | Active Meta-Learning for Predicting and Selecting Perovskite Crystallization Experiments Venkateswaran Shekar, Gareth Nicholas, Mansoor Ani Najeed, Margaret Zeile, Vincent Yu, Xiaorong Wang, Dylan Slack , Zhi Li, Philip Nega, Emory Chan, Alexander Norquist, Joshua Schrier, and Sorelle Friedler <i>Journal of Chemical Physics, 2022</i> Reliable Post hoc Explanations: Modeling Uncertainty in Explainability Dylan Slack , Sophie Hilgard, Sameer Singh, and Hima Lakkaraju <i>NeurIPS, 2021</i> |

Counterfactual Explanations Can Be Manipulated
Dylan Slack, Sophie Hilgard, Hima Lakkaraju, and Sameer Singh
NeurIPS, 2021

On the Lack of Robust Interpretability of Neural Text Classifiers
Muhammad Bilal Zafar, Michele Donini, **Dylan Slack**, Cedric Archambeau, Sanjiv Das, Krishnaram Kenthapadi
Findings of ACL, 2021

Context, Language Modeling, and Multimodal Data in Finance
Sanjiv Das, Connor Goggins, John He, George Karypis, Sandeep Krishnamurthy, Mitali Mahajan, Nagpurnanand Prabhala, **Dylan Slack**, Rob van Dusen, Shenghua Yue, Sheng Zha, Shuai Zheng
The Journal of Financial Data Science, 2021

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods
Dylan Slack*, Sophie Hilgard*, Emily Jia, Sameer Singh, and Himabindu Lakkaraju
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020

Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data
Dylan Slack, Sorelle Friedler, and Emile Givental
ACM Conference on Fairness, Accountability and Transparency (FAccT), 2020

* denotes equal contribution.

Workshop Publications

Defuse: Training More Robust Models through Creation and Correction of Novel Model Errors
Dylan Slack, Nathalie Rauschmayr, Krishnaram Kenthapadi
NeurIPS XAI 4 Debugging Workshop 2021

Feature Attributions and Counterfactual Explanations Can Be Manipulated
Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju
ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, 2021

Differentially Private Language Models Benefit from Public Pre-training
Gavin Kerrigan*, **Dylan Slack***, and Jens Tuyls*
EMNLP PrivNLP Workshop, 2020

Assessing the Local Interpretability of Machine Learning Models
Dylan Slack, Sorelle A. Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy
NeurIPS Workshop on Human-Centric Machine Learning, 2019

* denotes equal contribution.

Invited & Contributed Talks

Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations
Stanford MedAI, 2022 in Virtual

Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations
Imperial College, London, 2022 in Virtual

Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations
Meta, 2022 in Virtual

Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations

| | | |
|----------------------------------|--|-------------|
| | tions UC Irvine CML Seminar Series, 2022 in Virtual | |
| | Counterfactual Explanations Can Be Manipulated. NeurIPS, 2021 in Virtual | |
| | Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. NeurIPS, 2021 in Virtual | |
| | Feature Attributions and Counterfactual Explanations Can Be Manipulated. ICML workshop on XAI, 2021 in Virtual | |
| | Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. ICML workshop on Interpretable Machine Learning in Healthcare, 2021 in Virtual | |
| | Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods Aggregate Intellect, 2021 in Virtual | |
| | Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data FAccT Conference, 2020 in <i>Barcelona, Spain</i> | |
| Patents | Automatic Failure Diagnosis and Correction in Machine Learning Models Nathalie Rauschmayr, Krishnaram Kenthapadi, and Dylan Slack <i>Patent Application Filed</i> | |
| Other Industry Experience | DBO Partners Investment Banking Summer Analyst | Summer 2018 |
| | ValueAct Capital Summer Analyst | Summer 2017 |
| Travel Grants | Fairness, Accountability and Transparency in Machine Learning (FAccT) <i>Barcelona, Spain (2020)</i> | |
| | Neural Information Processing Systems (NeurIPS) <i>Vancouver, Canada (2020)</i> | |
| Service | KDD Deep Learning Day Organizer <i>2021</i> | |
| Teaching | Interpretability and Explainability in Machine Learning (COMPSCI 282BR) Harvard University <i>Guest Lecture (2021)</i> | |
| | Machine Learning (CS 178) UC Irvine <i>Reader (2019)</i> | |
| | Data Structures (CS 206) Bryn Mawr College <i>TA (2019)</i> | |
| | Introduction to Data Structures (CS 106) Haverford College <i>TA (2017, 2018, 2019)</i> | |
| | Introduction to Data Science (CS 104) Haverford College | |

TA (2016)

Review Services

FAccT 2021
ICLR 2021 (*Outstanding Reviewer Award*)
ICML 2020
AAAI 2020, 2021
NeurIPS 2019, 2020, 2021 (*Outstanding Reviewer Award*)
KDD 2019

Press & Media

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods,
[Harvard Business Review](#), [Deeplearning.ai](#), [Twitter](#)