

Dylan Z. Slack

Curriculum Vitae: March 20, 2021
Website: <https://dylanslacks.website>

Email: dslack@uci.edu
C: 415-847-2440

Education	University of California - Irvine , Irvine, CA <i>Ph.D. Computer Science</i> Advisors: Sameer Singh & Hima Lakkaraju Sep. 2019 - Present	
	Haverford College , Haverford, PA <i>B.S. Computer Science with High Honors</i> Magna Cum Laude Advisor: Sorelle Friedler Sep. 2015 - May 2019	
Research and Industry Experience	University of California - Irvine Research Assistant (UCI NLP, UCI CREATE, HPI Institute) <i>Advised by:</i> Sameer Singh & Hima Lakkaraju	Sep. 2019 - Present
	Amazon Web Services Applied Scientist Intern <i>Advised by:</i> Krishnaram Kenthapadi & Nathalie Rauschmayr	Jun. 2020 - Sep. 2020
	Haverford College Research Assistant, Department of Computer Science <i>Advised by:</i> Sorelle Friedler	Sep. 2017 - Aug. 2019
Awards	ICLR Outstanding Reviewer, 2021 Hasso Plattner Institute Fellow, 2021 Ambler Scholar, 2019	
Publications & Preprints [Scholar]	Defuse: Debugging Classifiers Through Distilling Unrestricted Adversarial Examples Dylan Slack , Nathalie Rauschmayr, and Krishnaram Kenthapadi arXiv, 2020	
	How Much Should I Trust You? Modeling Uncertainty of Black Box Explanations Dylan Slack , Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju arXiv, 2020	
	Differentially Private Language Models Benefit from Public Pre-training Gavin Kerrigan*, Dylan Slack *, and Jens Tuyls* EMNLP PrivNLP Workshop, 2020	
	Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods Dylan Slack *, Sophie Hilgard*, Emily Jia, Sameer Singh, and Himabindu Lakkaraju <i>AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020</i> [Oral Presentation]	
	Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data Dylan Slack , Sorelle Friedler, and Emile Givental <i>ACM Conference on Fairness, Accountability and Transparency (FAccT), 2020</i>	
	Assessing the Local Interpretability of Machine Learning Models Dylan Slack , Sorelle A. Friedler, Carlos Scheidegger, and Chitradheep Dutta Roy <i>NeurIPS Workshop on Human-Centric Machine Learning, 2019</i>	

* denotes equal contribution.

Patents	Automatic Failure Diagnosis and Correction in Machine Learning Models Nathalie Rauschmayr, Krishnaram Kenthapadi, and Dylan Slack <i>Patent Application Filed</i>
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>
Teaching	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 <i>TA (2016)</i>
Talks	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>
Review Services	FAccT 2021 ICLR 2021 ICML 2020 AAAI 2020, 2021 NeurIPS 2019, 2020 KDD 2019
Press & Media	Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, Harvard Business Review , DeepLearning.ai , Twitter