Dylan Z. Slack

Curriculum Vitae: October 24, 2020
Website: https://dylanslacks.website

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

Education

Unversity of California - Irvine, Irvine, CA
Ph.D. Computer Science
Advisors: Sameer Singh & Hima Lakkaraju
Sep. 2019 - Present

Haverford College, Haverford, PA
B.S. Computer Science with High Honors
Magna Cum Laude
Advisor: Sorelle Friedler
Sep. 2015 - May 2019

Publications & Preprints

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*, Emiliy Jia, Sameer Singh, and Himabindu Lakkaraju
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020
[Oral Presentation] **[Deeplearning.ai, Harvard Business Review]**
Also accepted at SafeAI Workshop AAAI 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
Also accepted at NeurIPS HCML Workshop 2019

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.

Industry Positions

Amazon Web Services May 2020 - Sep. 2020
Applied Scientist Intern
Conducted research related to debugging machine learning models and NLP for finance.
Advised by: Krishnaram Kenthapadi & Nathalie Rauschmayr

Precision-GX Feb. 2017 - Jun. 2017
Machine Learning Intern
Applied machine learning research looking at detecting fraud, waste, and abuse in health care claims data.
Advised by: Brandon Smith

Academic Positions

University of California - Irvine Sep. 2019 - Present
Research Assistant, Department of Computer Science
Machine learning research focused on fairness, interpretability, and natural language

processing.
Advised by: Sameer Singh

Haverford College Sep. 2017 - Aug. 2019
Research Assistant, Department of Computer Science
Machine learning research spanning deep meta-learning, fairness, interpretability of black box models, and reinforcement learning.
Advised by: Sorelle Friedler

Swarthmore College Sep. 2018 - May. 2019
Research Assistant, Department of Computer Science
Applied machine learning research using deep learning for population genomics. In particular, using deep learning methods to predict population size characteristics from genetics data.
Advised by: Sara Mathieson

Awards & Honors

Ambler Scholar 2019
Given to the top 15 varsity student athletes on the basis of grade point average in the Haverford College graduating class.

Travel Grants

Fairness, Accountability and Transparency in Machine Learning (FAccT)
Barcelona, Spain (2020)

Neural Information Processing Systems (NeurIPS)
Vancouver, Canada (2020)

Poster Presentations

NeurIPS Workshop on Human-Centric Machine Learning 2019
Vancouver, Canada
Fairness Warnings

NeurIPS Workshop on Human-Centric Machine Learning 2019
Vancouver, Canada
Fair Meta-Learning

NeurIPS Workshop on Human-Centric Machine Learning 2019
Vancouver, Canada
Assessing the Local Interpretability of Machine Learning Models

Oral Presentations

FAccT Conference 2020
Barcelona, Spain
Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data
(8 Minute Talk)

UC Irvine Department of Computer Science 2019
Irvine, CA
Statistical Notions of Fairness and their Limitations
(30 Minute Talk)

Perovskite Lab Group 2019
Haverford, PA
Rapidly learning new perovskite amines through deep meta-learning.
(1 hour talk)

KINSC Undergraduate Research Symposium 2019
Haverford, PA
Using human expertise to better transfer deep reinforcement learning policies.
(30 min. talk)

Teaching

Machine Learning (CS 178)
UC Irvine
Reader (2019)

Data Structures (CS 206)
Bryn Mawr College
TA (2019)

Introduction to Data Structures (CS 106)
Haverford College
TA (2017, 2018, 2019)

Introduction to Data Science (CS 104)
Haverford College
TA (2016)

Review Services

FAccT 2021
ICLR 2021
ICML 2020
AAAI 2020, 2021
NeurIPS 2019, 2020
KDD 2019