Dylan Z. Slack

Curriculum Vitae: February 10, 2021
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

Research and Industry Experience

University of California - Irvine Sep. 2019 - Present
Research Assistant (UCI NLP, CREATE, HPI Institute)
Advised by: Sameer Singh

Amazon Web Services May 2020 - Sep. 2020
Applied Scientist Intern
Advised by: Krishnaram Kenthapadi & Nathalie Rauschmayr

Haverford College Sep. 2017 - Aug. 2019
Research Assistant, Department of Computer Science
Advised by: Sorelle Friedler

Swarthmore College Sep. 2018 - May. 2019
Research Assistant, Department of Computer Science
Advised by: Sara Mathieson

Awards

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*, Emiliy Jia, Sameer Singh, and Himabindu Lakkaraju
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020
[Oral Presentation]

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

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.

Travel Grants

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

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

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)

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 *Barcelona, Spain*

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