Dylan Z. Slack

Curriculum Vitae: October 27, 2021 Email: dslack@uci.edu
Website: https://dylanslacks.website

Education

Unversity of California - Irvine, Irvine, CA
Ph.D. Computer Science
Advisor: Sameer Singh
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, UCI CREATE, HPI Institute)
Advised by: Sameer Singh

Google AI Jun. 2021 - Sep. 2021
Research Intern
Advised by: Bo Dai & Yinlam Chow & Nevan Wichers

Amazon Web Services (AWS) Jun. 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

Awards

NeurIPS Outstanding Reviewer, 2021
ICLR Outstanding Reviewer, 2021
Hasso Plattner Institute Fellow, 2021
Ambler Scholar, 2019

Referred Publications [Scholar]

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack, Sophie Hilgard, Sameer Singh, and Hima Lakkaraju
NeurIPS, 2021

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**, Cdric 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*, Emiliy 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

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju
ICML IMLH Workshop, 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.

In Submission

SAFER: Data-Efficient and Safe Reinforcement Learning Through Skill Acquisition
Dylan Slack, Yinlam Chow, Bo Dai, and Nevan Wichers

Patents

Automatic Failure Diagnosis and Correction in Machine Learning Models
Nathalie Rauschmayr, Krishnaram Kenthapadi, and **Dylan Slack**
Patent Application Filed

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

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 *Barcelona, Spain*

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