Peter Hase

peter@cs.unc.edu · peterbhase.github.io · (919) 323-0393

EDUCATION

The University of North Carolina at Chapel Hill *Fall 2019 – Present*
Third-year PhD student in Computer Science *Chapel Hill, NC*
Research Area: Natural Language Processing | *Advisor:* Mohit Bansal

Duke University *Fall 2015 – Spring 2019*
BS in Statistical Science | Minor in Mathematics *Durham, NC*

RESEARCH INTERESTS

Interpretable and explainable machine learning, natural language processing, multi-agent communication, AI safety.

PUBLICATIONS

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs
Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, Srinivasan Iyer
Preprint on arXiv. [pdf] [code]

Low-Cost Algorithmic Recourse for Users With Uncertain Cost Functions
Prateek Yadav, Peter Hase, Mohit Bansal
Preprint on arXiv. [pdf] [code]

Search Methods for Sufficient, Socially-Aligned Feature Importance Explanations with In-Distribution Counterfactuals
Peter Hase, Harry Xie, Mohit Bansal
In NeurIPS 2021. [pdf] [code]

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data
Peter Hase, Mohit Bansal
Preprint on arXiv. [pdf] [code]

FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging
Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, Caiming Xiong
In EMNLP 2021. [pdf] [code]

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?
Peter Hase, Shiyue Zhang, Harry Xie, Mohit Bansal
In Findings of EMNLP 2020. [pdf] [code]

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?
Peter Hase, Mohit Bansal
In ACL 2020. [pdf] [code]

Interpretable Image Recognition with Hierarchical Prototypes
Peter Hase, Chaofan Chen, Oscar Li, Cynthia Rudin
In AAAI-HCOMP 2019. [pdf] [code]

Shall I Compare Thee to a Machine-Written Sonnet? An Approach to Algorithmic Sonnet Generation
John Benhardt, Peter Hase, Liuyi Zhu, Cynthia Rudin
Preprint on arXiv. [pdf] [code]

AWARDS

Google PhD Fellowship (Natural Language Processing), Google *2021*
Fellowship awarded to six students globally for research in Natural Language Processing, providing up to three years of full funding

William R. Kenan Jr. (Royster) Fellowship, UNC Chapel Hill *2019*
University fellowship awarded to one student in the 2019 cohort of computer science students, providing three years of full funding

First Prize in the PoetiX Literary Turing Test, Neukom Institute, Dartmouth College *2018*
Awarded for the top submission to the Neukom Institute's open competition for algorithmic sonnet generation

Nomination for Undergrad TA of the Year, Dept. of Statistical Science, Duke University *2018*
One of five undergrad nominations from faculty for the department's TA of the year award

ASA DataFest Honorable Mention, Dept. of Statistical Science, Duke University *2018*
Recognition for placement in top 10% of teams in a Duke-hosted data analysis competition entered by 380+ undergrad and grad students

Meritorious Winner in the Interdisciplinary Contest in Modeling, COMAP *2017*
Awarded for placement in the top 12% of over 8000 teams in the international modeling contest held by the Consortium for Mathematics and its Applications

A.J. Tannenbaum Trinity Scholarship, Duke University *2015*
A full academic merit scholarship awarded to one student from Guilford County, NC

INVITED TALKS

Center for Human Compatible AI, UC Berkeley *Summer 2021*
"Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" [slides]

TEACHING

Probabilistic Machine Learning (Graduate), Teaching Assistant *Spring 2019*
Dept. of Statistical Science, Duke University

Intro to AI, Teaching Assistant *Spring 2019*
Dept. of Computer Science, Duke University

Elements of Machine Learning, Teaching Assistant *Fall 2018*
Dept. of Computer Science, Duke University

Intro to Data Science, Teaching Assistant *Spring 2018*
Dept. of Statistical Science, Duke University

Regression Analysis, Teaching Assistant *Fall 2017*
Dept. of Statistical Science, Duke University

RESEARCH EXPERIENCE

Meta AI Research *Summer 2021*
Research Intern | *Supervisor:* Dr. Srinivasan Iyer *Seattle, WA*
- Worked on methods for detecting and updating beliefs/knowledge in language models
- Produced paper on the topic, "Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs"

Department of Statistical Science, Duke University *Summer 2018*
DOmath Researcher | *Supervisor:* Dr. Sayan Mukherjee *Durham, NC*
- Numerically estimated a measure of model complexity, the topological entropy, for two dynamical systems, the logistic map and linear dynamical system
- Empirically assessed how the reliability of inference for the linear dynamical system varies as a function of its entropy

Department of Neurobiology, Duke University *Spring & Summer 2018*
Research Assistant | *Supervisor:* Dr. Jeff Beck *Durham, NC*
- Implemented a hidden Markov model and linear dynamical system, each learned through variational Bayesian expectation maximization (VBEM)
- Modeled recordings of neuron activity in the actively singing Zebra finch; visualized and interpreted models' latent variable dynamics

Information Initiative at Duke *Summer 2017*
Data+ Researcher | *Supervisor:* Dr. Sheng Jiang *Durham, NC*
- Clustered Duke's alumni donors into groups with distinct giving behaviors via k-means
- Built logistic regression models to evaluate donors' philanthropic potential based on demographics and prior giving behavior

LEADERSHIP

Computer Science Student Association *Summer 2020 – Present*
Officer *Chapel Hill, NC*
- Observed faculty teaching to provide feedback in tenure review
- Record meeting minutes for CS faculty meetings to share with graduate students
- Working with faculty to streamline background requirements for doctoral applicants

High school and Undergraduate Research Mentoring *Spring 2020 – Present*
Research Mentor *Chapel Hill, NC*
- Meet weekly with an undergraduate research assistant in the MURGe-Lab to mentor ongoing publication track research
- Met weekly with a high school student from North Carolina School of Science and Math to mentor a summer project reimplementing current research in document summarization
- Presented live research demos to Chapel Hill K-12 students for UNC CS open house; printed machine written sonnets for students and discussed education and research at UNC

Start-up Technical Advising *Fall 2019 – Present*
Technical Advisor *Chapel Hill, NC*
- curalens.ai: previously advised Curalens on text generation strategies for a therapeutic chat-bot (note: Curalens also advised by domain experts)
- Acta: previously advised Acta on procedures for automatically summarizing crowdsourced constituent feedback for efficient communication to local governments

Effective Altruism: Duke *Spring 2016 – Spring 2019*
Co-President *Durham, NC*

- Moderated weekly discussions related to Effective Altruism, the social movement promoting the use of reason and evidence to maximize the good you can do for the world
- Organized lectures and reading groups on AI safety for Duke and UNC Chapel Hill students
- Led club from 9 to 30+ active members over my tenure as Co-President
- Recorded over 15 Giving What We Can pledges (10% of all future income) in pledge drives and over 30 One For the World pledges (1% of future income)

WORK EXPERIENCE

Clarity Campaign Labs *Summer 2016*
Research Analyst *Washington, DC*

- Visualized model predictions and political data; encoded surveys; drafted software guides for internal use

PROFESSIONAL SERVICE

Program Committees *Summer 2020 – Present*
Reviewer

- ACL Rolling Review, October 2021
- ACL Rolling Review, September 2021
- NeurIPS DistShift Workshop 2021
- EMNLP BlackboxNLP Workshop 2021
- EMNLP 2021
- ACL-IJCNLP 2021 (*Outstanding Reviewer*)
- RobustML Workshop at ICLR 2021
- NAACL-HLT 2021
- EACL 2021
- EMNLP 2020 (*Outstanding Reviewer*)