# Sarthak Jain

successar@gmail.com | github.com/successar | Sarthak Jain

## Research

Interpretability and Analysis of Deep Learning Models; Foundation Models; Machine Learning

## Education

**Northeastern University** *Boston, MA, USA*

PhD in Computer Science *Sept 2017 - July 2022*

- Advised by Byron C. Wallace

**Delhi Technological University** *Delhi, India*

BTech in Computer Engineering *Aug 2013 - May 2017*

## Professional Experience

**AWS AI Labs** *Sept 2022 - Current*

Applied Scientist

- Working on the team developing Amazon Titan Foundation Models.
- Supervised an intern on project regarding constraint satisfaction in Large Language Models.

**Adobe Labs** *May 2021 - Aug 2021*

Research Intern

Worked on project on using Influence functions for structured prediction tasks.

**Microsoft Health Futures** *May 2020 - Aug 2020*

Research Intern

Worked on Document level relation extraction from Biomedical literature with distant supervision.

**Allen Institute for Artificial Intelligence** *June 2019 - Aug 2019*

Research Intern

Worked on Structured results extraction from Machine Learning literature using distant supervision.

## Publications

### Dissertation

The Model Thinks What?! Interpreting Deep NLP Models with Rationales and Influence
Sarthak Jain
*PhD Dissertation*, 2022

### Conference Proceedings

How Many and Which Training Points Would Need to be Removed to Flip this Prediction?
Jinghan Yang, Sarthak Jain, Byron C Wallace
*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023

Influence Functions for Sequence Tagging Models
Sarthak Jain, Varun Manjunatha, Byron C Wallace, Ani Nenkova
*Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022

Combining Feature and Instance Attribution to Detect Artifacts
Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, Byron C Wallace
*Findings of the Association for Computational Linguistics: ACL 2022*, 2022

Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?
Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, Byron Wallace
*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021

An Empirical Comparison of Instance Attribution Methods for NLP
Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, Sameer Singh
*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021

Modular Self-Supervision for Document-Level Relation Extraction
Sheng Zhang, Cliff Wong, Naoto Usuyama, Sarthak Jain, Tristan Naumann, Hoifung Poon
*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021

ERASER: A Benchmark to Evaluate Rationalized NLP Models
Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, Byron C. Wallace

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020

Learning to Faithfully Rationalize by Construction
Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, Byron C Wallace
*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). 2020

SciREX: A Challenge Dataset for Document-Level Information Extraction
Sarthak Jain, Madeleine Zuylen, Hannaneh Hajishirzi, Iz Beltagy
*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). 2020

Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study
Ramin Mohammadi, Sarthak Jain, Amir T Namin, Melissa Scholem Heller, Ramya Palacholla, Sagar Kamarthi, Byron Wallace
*JMIR medical informatics* 8.11 (2020) e19761. JMIR Publications Inc., Toronto, Canada, 2020

Structured Disentangled Representations
Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, Jan-Willem Meent
*The 22nd International Conference on Artificial Intelligence and Statistics*, 2019

Attention is not Explanation
Sarthak Jain, Byron C Wallace
*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019

Learning to Identify Patients at Risk of Uncontrolled Hypertension Using Electronic Health Records Data
Ramin Mohammadi, Sarthak Jain, Stephen Agboola, Ramya Palacholla, Sagar Kamarthi, Byron C Wallace
*AMIA Summits on Translational Science Proceedings* 2019 (2019). American Medical Informatics Association, 2019

Learning Disentangled Representations of Texts with Application to Biomedical Abstracts
Sarthak Jain, Edward Banner, Jan-Willem Meent, Iain J Marshall, Byron C Wallace
*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018

Cross lingual sentiment analysis using modified BRAE
Sarthak Jain, Shashank Batra
*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015

### Workshop Proceedings

An Analysis of Attention over Clinical Notes for Predictive Tasks
Sarthak Jain, Ramin Mohammadi, Byron C Wallace
*Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019

Detecting Twitter posts with Adverse Drug Reactions using Convolutional Neural Networks
Sarthak Jain, Xun Peng, Byron C. Wallace
*SMM4H@AMIA*, 2017

Question answering over knowledge base using factual memory networks
Sarthak Jain
*Proceedings of the NAACL Student Research Workshop*, 2016

## Service

**Reviewer** EMNLP (2018-2022), ACL (2017-2022), NAACL (2018-2022), NeuRIPS (2019, 2020, 2022), AAAI (2021), ICLR(2020, 2022), ICML (2020), EACL (2020, 2021)

**References available upon request.**