SAAHIL JAIN

saahil.jain@cs.stanford.edu

EDUCATION:

Stanford University, *M.Sc in Computer Science,* Palo Alto, CA 2019 – 2021
GPA 4.08. Research on improving AI in resource-constrained domains (e.g., healthcare) under Professor Andrew Ng in the Stanford Machine Learning Group

Columbia University, *B.Sc in Computer Science*, New York, NY 2014 - 2018
GPA 3.95. Research on AI in healthcare under Professor Nicholas Tatonetti in the Columbia Medical Center and wireless communication under Professor Gil Zussman in the Wireless and Mobile Networking Lab

PAPERS:

1. **RadGraph: Extracting Clinical Entities and Relations from Radiology Reports**
   Saahil Jain*, Ashwin Agrawal*, Adriel Saporta*, Steven QH Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, Pranav Rajpurkar
   Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks, 2021
   Oral Presentation

2. **VisualCheXbert: Addressing the Discrepancy Between Radiology Report Labels and Image Labels**
   Saahil Jain*, Akshay Smit*, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, Pranav Rajpurkar
   Proceedings of the Conference on Health, Inference, and Learning (ACM-CHIL), 2021

3. **CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT**
   Akshay Smit*, **Saahil Jain***, Pranav Rajpurkar*, Anuj Pareek, Andrew Y Ng, Matthew P Lungren
   Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020

4. **Effect of Radiology Report Labeler Quality on Deep Learning Models for Chest X-Ray Interpretation**
   Saahil Jain*, Akshay Smit*, Andrew Y Ng, Pranav Rajpurkar
   Neural Information Processing Systems (NeurIPS) Workshop on Data-Centric AI (DCAI), 2021

5. **Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management**
   Cécile Logé*, Emily Ross*, David Yaw Amoah Dadey, **Saahil Jain**, Adriel Saporta, Andrew Y Ng, Pranav Rajpurkar
   Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmark, 2021
   Oral Presentation

6. **Power-Aware Neighbor Discovery for Energy Harvesting Things**
   Tingjun Chen, Gregory Chen, **Saahil Jain**, Robert Margolies, Guy Grebla, Dan Rubenstein, Gil Zussman
   Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems (SenSys), 2016

7. **On the Opportunities and Risks of Foundation Models**
   Stanford Center for Research on Foundation Models, Whitepaper, 2021

TEACHING EXPERIENCE:

Teaching Assistant for Stanford CS372, AI for Disease Diagnosis and Information Recommendations 2020-2021
Helped teach a graduate-level course on AI in healthcare instructed by Professor Edward Chang in Spring 2020 and Spring 2021. Mentored students on their research projects and actively shaped syllabus/assignments for this new course.

Teaching Assistant for Stanford CS229, Machine Learning 2020
Helped teach the machine learning course co-instructed by Professors Andrew Ng, Chris Re, and Tengyu Ma in Fall 2020. Co-led and gave course-wide recitation lectures on Python, deep learning, and evaluation metrics.

Peer Instructor for Columbia COMS1404, Emerging Scholars Program 2016-2018
Taught a computer science seminar for undergraduate freshman/sophomores at Columbia University.

REVIEWER:

Reviewer, NeurIPS Track on Datasets and Benchmarks
2021

PROFESSIONAL RESEARCH/WORK EXPERIENCE:

Machine Learning Engineer, You.com, Palo Alto, CA June 2021 – Present

- Currently an early engineer at a startup building a search engine to summarize the web led by Dr. Richard Socher.
- Developing core machine learning and search infrastructure, working on topics ranging from semantic search to building/deploying natural language processing models, from the ground up as one of the first engineers.
- Building data infrastructure to index and manage large datasets (TBs) from scratch.
- Leading initiatives to improve consumer healthcare search.

Researcher, Stanford Machine Learning Group led by Prof. Andrew Ng, Palo Alto, CA January 2020 – July 2021

- Researched and developed methods to improve artificial intelligence in resource-constrained domains such as healthcare with Professor Andrew Ng, Dr. Pranav Rajpurkar, Dr. Curt Langlotz, Dr. Matt Lungren, and other collaborators from both the computer science and medical departments.
- Developed RadGraph, a dataset of entities and relations in radiology reports based on a novel information extraction schema with benchmark models for knowledge graph extraction, as co-first author (NeurIPS Track on Datasets and Benchmarks 2021, Oral Presentation).
- Designed and developed various deep learning techniques to improve label quality for medical imaging models. Co-first author of CheXbert (EMNLP 2020) and VisualCheXbert (ACM-CHIL 2021), which advanced the state-of-the-art for radiology report labeling and introduced a novel training method to address discrepancies between radiology report labels and radiology image labels respectively. First author of a systematic investigation showing that these advancements in report labeling translated to improvements in the performance of deep learning models for chest X-ray interpretation (NeurIPS Workshop on Data-Centric AI 2021).
- Ran initial pilots and co-designed a dataset and framework for assessing bias in medical AI in the context of pain management (NeurIPS Track on Datasets and Benchmarks 2021, Oral Presentation).
- Helped co-author the healthcare section of the whitepaper introducing foundation models released by the Stanford Center for Research on Foundation Models, an interdisciplinary group of researchers across labs at Stanford.
- Developed computer vision models to improve the diagnosis of Tuberculosis from large (>1 GB) whole slide images with faculty in the Stanford Pathology Department.

Engineering Intern, Datavant, San Francisco, CA June 2020 – August 2020

- Built data pipelines to ingest healthcare datasets at scale for a growing healthcare startup aimed at eliminating silos of data that hold back medical research. Backed by Roivant Sciences, Softbank, and Founders Fund.

Product Manager, Microsoft, Redmond, WA August 2018 – September 2019

- Worked as a product manager on Office 365 cloud infrastructure, building a software product to intelligently ensure the health of hundreds of thousands of machines in datacenters across the world. Incubated a new machine learning initiative to more intelligently detect hardware issues and recommend repair actions.

Research Intern, IBM Extreme Blue, Research Triangle Park, NC May 2017 – August 2017

- Researched methods of extracting insights from API data for IBM API Connect as part of IBM's flagship Extreme Blue Leadership Program.

Researcher, Columbia Medical Lab led by Prof. Nicholas Tatonetti, New York, NY January 2017 – May 2017

- Researched database solutions for the cross-institutional Biomedical Data Translator project (https://ncats.nih.gov/translator) under Professor Nicholas Tatonetti and Dr. Nwankwo at the Tatonetti Lab in the Columbia Medical Center. Wrote reports on data-driven precision medicine, such as mining FDA adverse drug events to find harmful drug combinations.

Researcher, WiMNet Lab led by Prof. Gil Zussman, New York, NY September 2015 – June 2016

- Modeled network dynamics among low-power, wireless, energy-harvesting devices and demonstrated efficiency of the Power-Aware Neighbor Discovery for Energy Harvesting Things algorithm (SenSys 2016) under Professor Gil Zussman in the Wireless and Mobile Networking Lab (WiMNet) at Columbia University.

HONORS AND AWARDS:

Oral, NeurIPS Track on Datasets and Benchmarks, 2021 – Gave oral presentation for paper titled "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports"

Invited speaker at KDD, International Workshop on Knowledge Graphs: Heterogeneous Graph Deep Learning and Applications, 2021 – Gave a talk with Dr. Pranav Rajpurkar on our work building graph datasets in healthcare

Featured interviewee, Managed Healthcare Executive, 2021 – Interviewed for an article, "As healthcare sees a shiny future with AI, some see some glare", in a healthcare magazine (pages 12-13) based on healthcare AI research at Stanford

Accel Leadership Fellow, 2021 – Chosen as 1 of 24 students from Stanford University (undergrad/grad students) to build leadership skills as part of an entrepreneurial program co-led by the Stanford Technology Venture Program and Accel

Computer Science Excellency Award, 2018 – Awarded to 35 graduating CS students across all undergraduate schools at Columbia with a record of outstanding academic achievement and scholarship

Magna Cum Laude, Columbia University, 2018

Tau Beta Pi Engineering Honors Society, 2017

Florida State Debate Champion, 2012 – Champion of the Florida State Lincoln Douglas Debate Tournament