

# Gregory Yaune

gyaune@cs.cornell.edu  
https://gyaune.github.io

## education

### Cornell University

Ph.D. Computer Science

Advisor: David Mimno

Ithaca, NY  
August 2018–now (ongoing)

### Brown University

Sc.B. Computer Science

A.B. History of Art and Architecture

Providence, RI  
September 2012–May 2015

## experience

### Student Researcher

Google Research, PAIR Team

Fall 2022

Seattle, WA

### Technical Associate

MIT Media Lab

June 2016–June 2018

Cambridge, MA

### Associate Software Engineer

MathWorks

July 2015–June 2016

Natick, MA

## peer-reviewed conference publications

A Pretrainer's Guide to Training Data:

Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity

Shayne Longpre, **Gregory Yaune**, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, Daphne Ippolito

*Outstanding Paper Award*

NAACL 2024

Data Similarity is Not Enough to Explain Language Model Performance

**Gregory Yaune**, Emily Reif, David Mimno

EMNLP 2023

Comparing Text Representations: A Theory-Driven Approach

**Gregory Yaune**, David Mimno

EMNLP 2021

Domain-Specific Lexical Grounding in Noisy Visual-Textual Documents

**Gregory Yaune**, Jack Hessel, David Mimno

EMNLP 2020

## preprints

Stronger Random Baselines for In-Context Learning

**Gregory Yaune**, David Mimno

arXiv 2024

## peer-reviewed journal articles

The Afterlives of Shakespeare and Company in Online Social Readership

Maria Antoniak\*, David Mimno\*, Rosamond Thalken\*, Melanie Walsh\*, Matthew Wilkens\*, **Gregory Yaune**\*

\*equal contribution, alphabetical order

Cultural Analytics and Modernism/modernity 2024

peer-reviewed workshop publications

Probing Heterogeneous Pretraining Datasets with Small Curated Datasets

**Gregory Yauney**, Emily Reif, David Mimno

Data-Centric Machine Learning Research Workshop at ICML 2023

Network Analysis Finds Shifts in the History of Modern Architecture

**Gregory Yauney**, David Mimno

Digital Humanities 2020

Combating the Challenges of Local Privacy for Distributional Semantics with Compression

Alexandra Schofield, **Gregory Yauney**, David Mimno

Privacy in Machine Learning Workshop at NeurIPS 2019

Computational Prediction of Elapsed Narrative Time

**Gregory Yauney**, Ted Underwood, David Mimno

Workshop on Narrative Understanding at NAACL 2019

other peer-reviewed publications

Digital reconstruction of teeth using near-infrared light

Keith Angelino, **Gregory Yauney**, Aman Rana, David Edlund, Pratik Shah

IEEE Engineering in Medicine and Biology Society (EMBC), 2019

Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks

Aman Rana, **Gregory Yauney**, Alarice Lowe, Pratik Shah

IEEE International Conference on Machine Learning and Applications (ICMLA), 2018

Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection

**Gregory Yauney**, Pratik Shah

Machine Learning for Healthcare, 2018

Technology-enabled examinations of cardiac rhythm, optic nerve, oral health, tympanic membrane, gait and coordination evaluated jointly with routine health screenings: an observational study at the 2015 Kumbh Mela in India

Pratik Shah\*, **Gregory Yauney\***, Otkrist Gupta, Vincent Patalano II, Mrinal Mohit, Rikin Merchant, S.V. Subramanian

BMJ Open, 2018

Clinical validation and assessment of a modular fluorescent imaging system and algorithm for rapid detection and quantification of dental plaque

Keith Angelino, Pratik Shah, David A. Edlund, Mrinal Mohit, **Gregory Yauney**

BMC Oral Health, 2017

Automated segmentation of gingival diseases from oral images

Aman Rana, **Gregory Yauney**, Lawrence C. Wong, Otkrist Gupta, Ali Muftu, Pratik Shah

IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), 2017

Convolutional neural network for combined classification of fluorescent biomarkers and expert annotations using white light images

**Gregory Yauney**, Keith Angelino, David Edlund, Pratik Shah

IEEE International Conference on Bioinformatics and Bioengineering (BIBE), 2017

Comparative grain topology

Trevor Keller, Barbara Cutler, Emanuel A. Lazar, **Gregory Yauney**, Daniel J Lewis

Acta Materialia, 2014

## talks

A Pretrainer's Guide to Training Data, joint talk with Shayne Longpre

Salesforce AI Research, February 2024

Cohere for AI, January 2024

MosaicML, September 2023

University of Washington, August 2023

Allen Institute for Artificial Intelligence, August 2023

Received Textual Similarity via Network Distance, joint work with Matthew Wilkens and David Mimno

Conference talk at the Association for Computers and the Humanities, July 2021.

## teaching

Graduate TA, Mathematical Foundations of ML, Cornell

Fall 2023

Graduate TA, Intro to ML, Cornell

Spring 2023

Graduate TA, Mathematical Foundations of ML, Cornell

outstanding TA award

Spring 2022

Graduate TA, Data Visualization, Cornell

Spring 2021

Graduate TA, Intro to Data Science, Cornell

Spring 2020

Graduate TA, Intro to Computer Graphics, Cornell

Fall 2018

TA, Artificial Intelligence, Brown

Fall 2014

TA, Computer Vision, Brown

Fall 2013

## reviewing

COLM 2024

ACL 2024

NAACL 2024

ICLR workshop proposals 2024

DMLR @ ICML 2023, @ ICLR 2024, @ ICML 2024

DH 2022

ACH 2021