



## education

### phd | machine learning

uc berkeley | '17-'22

research: interpretable ml

advisor: bin yu

collaborators:

a. kornblith (medicine)

s. upadhyayula (biology)

### bs | cs & math

university of virginia | '14-'17

double major

## skills

machine learning

deep learning • pytorch

rule-based models • data cleaning

causal inference • pycharm

experienced

python • java • matlab

proficient

r • c/c++ • web basics

human languages

english • spanish • hindi

## awards

berkeley grad slam semifinalist '19, '22

outstanding teaching award '18

uva rader research award '17

uva undergrad symposium winner '17

raven honor society '16-'17

icpc regional qualification '14-'16

1st place microsoft code jam '16

3rd place google games uva '17

2nd place apt puzzle competition '17

### funding awards

pdsoros fellowship finalist '19

ircn workshop travel award '19

vidya shelat fund award '16

rodman scholarship '14-'17

## experience

### microsoft research | senior researcher (deep learning lab)

fall '22 - present

- researching improving the interpretability of language models
- researching scientific/medical knowledge discovery with language models

### berkeley | interpretable ml research (bin yu lab %)

fall '17 - spring '22

- developed interpretation methods for ml models (e.g. neural nets)
- developed interpretable models in medicine, biology, and computer vision

### paige ai | ai research scientist

summer '21

- interpretable deep learning in digital pathology (especially bladder cancer)

### aws | ml fairness internship (pietro perona lab %)

summer '20

- testing for bias with causal matching using GANs
- interpreting semantic directions in generative models

### response4life | volunteer data scientist

spring '20

- helped develop, integrate, and deploy models to forecast covid-19 severity

### pacmed ai | healthcare ml internship

summer '19

- developed techniques to interpret machine-learning models for healthcare
- integrated interpretability techniques for predicting icu re-admission

### meta ai | computer vision internship

summer '17

- investigated unsupervised deep learning for segmentation of satellite imagery
- implemented crfs for segmentation post-processing

### uva | ml research (yanjun qi lab %)

fall '16 – spring '17

- developed multi-task graphical models for analyzing functional brain connectivity

### hhmi | ml research (srini turaga lab %)

summer '14, '15, '16

- improved cnns and watershed algorithms for neural image segmentation
- analyzed backpropagating action potentials via biophysical simulations

### uva | comp. neuroscience research (william levy lab %)

fall '14 - fall '16

- developed detailed biophysical models of neural computation
- analyzed energy efficiency, noise, and variability in stochastic neurons

## coursework

### computation

machine learning  
computer vision  
structure learning  
algorithms  
artificial intelligence  
deep learning  
learning theory  
ai in graphics  
cs theory  
data structures  
software dev. I & II  
information retrieval  
computer architecture

### stat/math







statistical models  
probability  
statistics  
optimization  
linear algebra  
info theory  
real analysis  
linear models  
stochastic processes  
chaos theory I & II  
multivariate calculus  
discrete mathematics  
differential equations  
abstract algebra

### neuroscience





neural coding  
neural network models  
neurobiology  
visual neuroscience  
cognitive science

## selected publications




### interpretable deep learning

- augmenting interpretable models with llms during training **cs**, askari, caruana & gao, *arXiv*, '23 
- explaining black box text modules in natural language with language models **cs\***, hsu\*, antonello, jain, huth, yu & gao, *arXiv*, '23 
- explaining data patterns in natural language via interpretable autoprompting **cs\***, morris\*, aneja, rush, & gao *arXiv*, '22 
- adaptive wavelet distillation from neural networks through interpretations: ha, **cs**, et al. *neurips* '21 
- interpretations are useful: penalizing explanations to align neural networks with prior knowledge: rieger, **cs**, murdoch, & yu, *icml* '20 
- hierarchical interpretations for neural network predictions: **cs\***, murdoch\*, & yu, *iclr* '19 



### interpretable rule-based modeling

- imodels: a python package for interpretable modeling: **cs\***, nasseri\*, tan, tang, & yu, *joss* '21   
 **1k+ stars**
- fast interpretable greedy-tree sums (figs): tan\*, **cs\***, nasseri, agarwal, & yu *arxiv* '22 
- hierarchical shrinkage: improving accuracy and interpretability of tree-based methods: agarwal\*, tan\*, ronen, **cs**, & yu *icml* '22 (*spotlight*) 

### real-world data science

- curating a covid-19 data repository and forecasting county-level death counts in the united states: altieri, barter, ..., **cs\***, ..., & yu\* *harvard data science review* '20 
- predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma kornblith\*, **cs\***, et al. *plos digital health* 
- interpretable deep learning for accurate molecular partner prediction in clathrin-mediated endocytosis: **cs\***, li\* et al. *in prep* 

### applied computer vision

- large scale image segmentation with structured-loss-based deep learning for connectome reconstruction: funke et al. *tpami* '18 
- matched sample selection with GANs for mitigating attribute confounding: **cs**, balakrishnan, & perona *cvpr* '21 *civ workshop* 

## teaching

### berkeley | student instructor

summer 2018


machine learning: cs 189/289 

lectures to class of 80+ students


fall 2019

artificial intelligence: cs 188 

## mini-projects

paper notes 

blog, & slides 

hummingbird tracking 

news balancer django app 

java mini-games 

'14-now

'14-now

'18

'17

'14-'16

## service

basis education volunteering

bair undergrad mentoring

neurips reviewer

acl rolling reviewer

iclr workshop reviewer

cvpr reviewer

aaai xai workshop reviewer

neurips ml4h workshop reviewer

computer literacy volunteering

'19-'22

'18-'22

'21, '23

'22

'21

'21

'21

'20

'15-'17