chandan singh

ai interpretability researcher

csinva23@gmail.com
csinva
csinva.io
csinva
csinva

education

phd | machine learning
uc berkeley | '17-'22
research: interpretable ml
advisor: bin yu

bs | cs & math
university of virginia | '14-'17
double major

skills

language models | deep learning
data science | data cleaning
huggingface | pytorch
rule-based models | causal inference

awards

berkeley grad slam semifinalist '19, '22
pdsoros fellowship finalist '19
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
rodman scholarship '14-'17

teaching

berkeley | summer 2018
machine learning: cs 189/289 🔗
lectures to class of 80+ students

berkeley | fall 2019
artificial intelligence: cs 188 🔗

service

volunteering
basis education volunteering '19-'22
bair undergrad mentoring '18-'22
computer literacy volunteering '15-'17

area chair
xxai workshop '24 | ml4h '24

reviewer
iclr,icml '25 | iclr,icml,neurips '24
neurips '23 | acl '22
iclr,cvpr,aaai,neurips '21 | neurips '20

experience

microsoft research

senior researcher (deep learning group) | summer '22 - present

- improving the interpretability of large language models
- researching knowledge discovery with large language models
- building next-generation language models

health tech

paige ai | research scientist | summer '21 - summer '22

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

response4life | volunteer data scientist | spring '20

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

pacmed ai | healthcare ml intern | summer '19

- developed interpretable, tabular machine-learning models for healthcare

phd

berkeley | interpretable ml research (bin yu group) | fall '17 - spring '22

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

aws | ml fairness intern (pietro perona group) | summer '20

- testing for bias with causal matching using GANs

meta ai | computer vision intern | summer '17

- investigated unsupervised deep learning for segmentation of satellite imagery

undergrad

hhmi | ml research (srini turaga group) | summer '14, '15, '16

- researched neural image segmentation and biophysical simulations

uva | ml research (yanjun qi group) | fall '16 – spring '17

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

uva | comp. neuroscience research (william levy group) | fall '14 - fall '16

- developed biophysical models of single-neuron computation

selected publications

interpretability × language models → neuroscience

- augmenting interpretable models with llms: **cs**, et al. *nature comm.*, '23 🔗 </>
- explanation-mediation validation with llms: antonello*, **cs***, et al. *arxiv*, '24 🔗 </>
- interpretable embeddings by asking llms questions benara*, **cs***, et al. *neurips*, '24 🔗

interpretability × rules → clinical decision rules

- fast interpretable greedy-tree sums: tan*, **cs***, nasseri*, agarwal* et al. *pnas '22* 🔗 </>
- hierarchical tree shrinkage: agarwal*, tan*, ronen, **cs**, & yu *icml '22* 🔗 </>
- imodels: an interpretability package: **cs***, nasseri*, tan, tang, & yu, *joss '21* 🔗 </>

interpretability × deep learning → general domain

- adaptive wavelet distillation from dnns: ha, **cs**, et al. *neurips '21* 🔗 </>
- aligning dnns by regularizing explanations: rieger, **cs**, et al. *icml '20* 🔗 </>
- hierarchical interpretations for dnn predictions: **cs***, murdoch*, & yu, *iclr '19* 🔗 </>