

# Democratizing Human-Centered AI with Web-Based Visual Explanation and Interactive Guidance



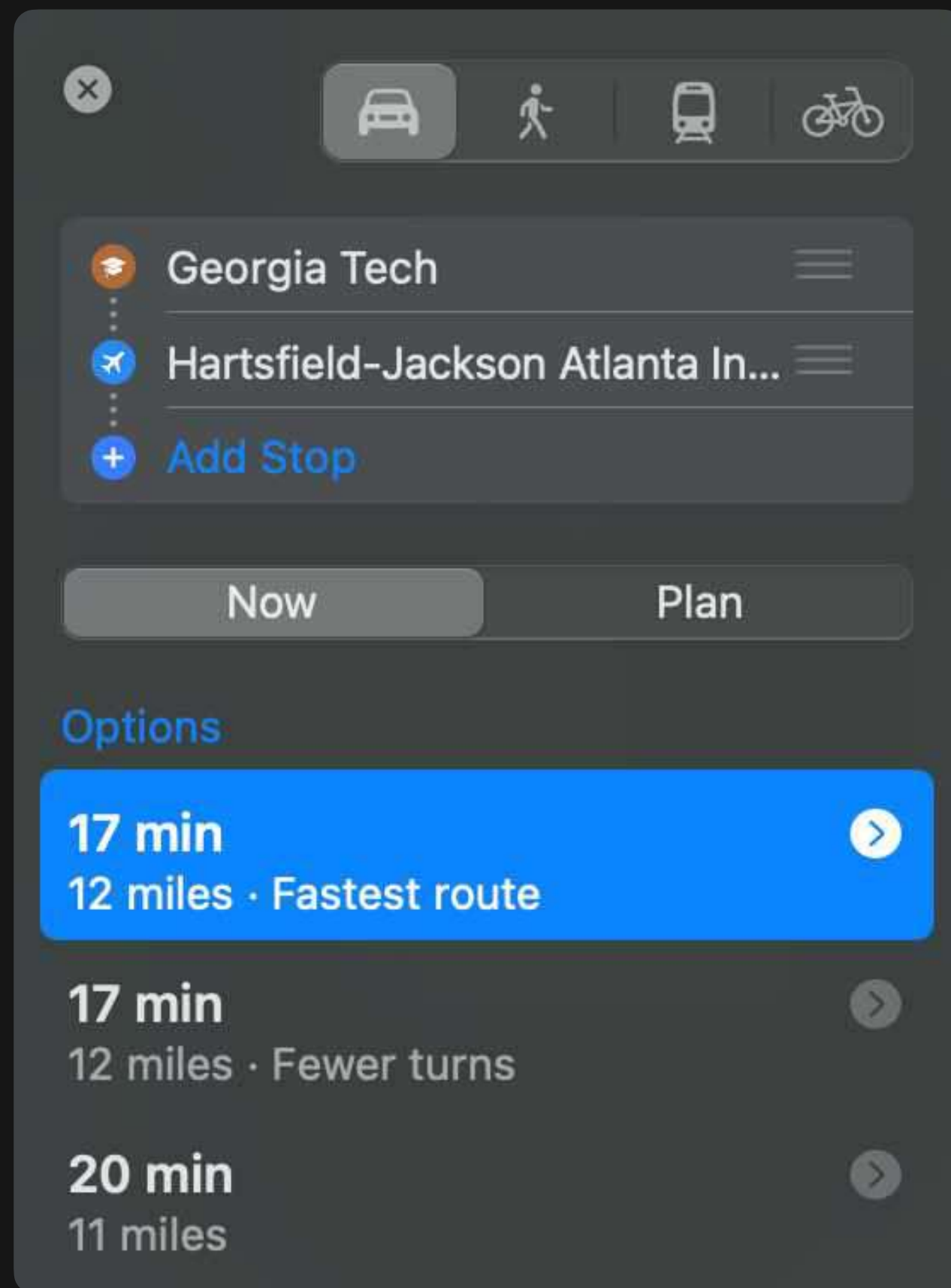
**Jay Wang**

Research Engineer

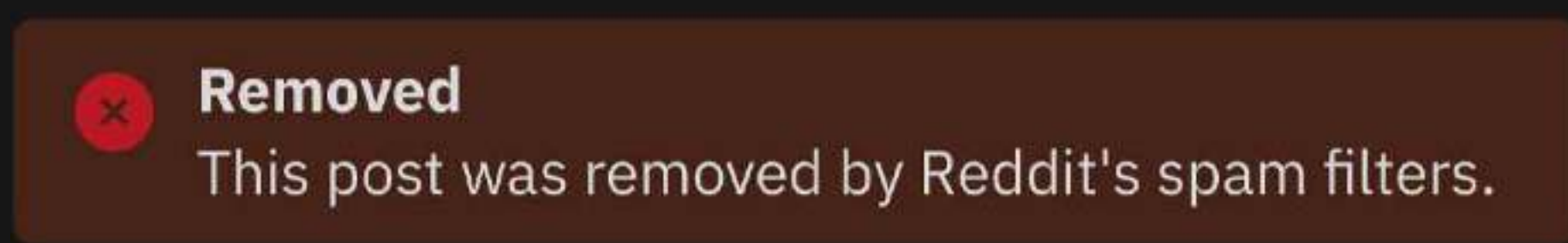
<https://zijie.wang>



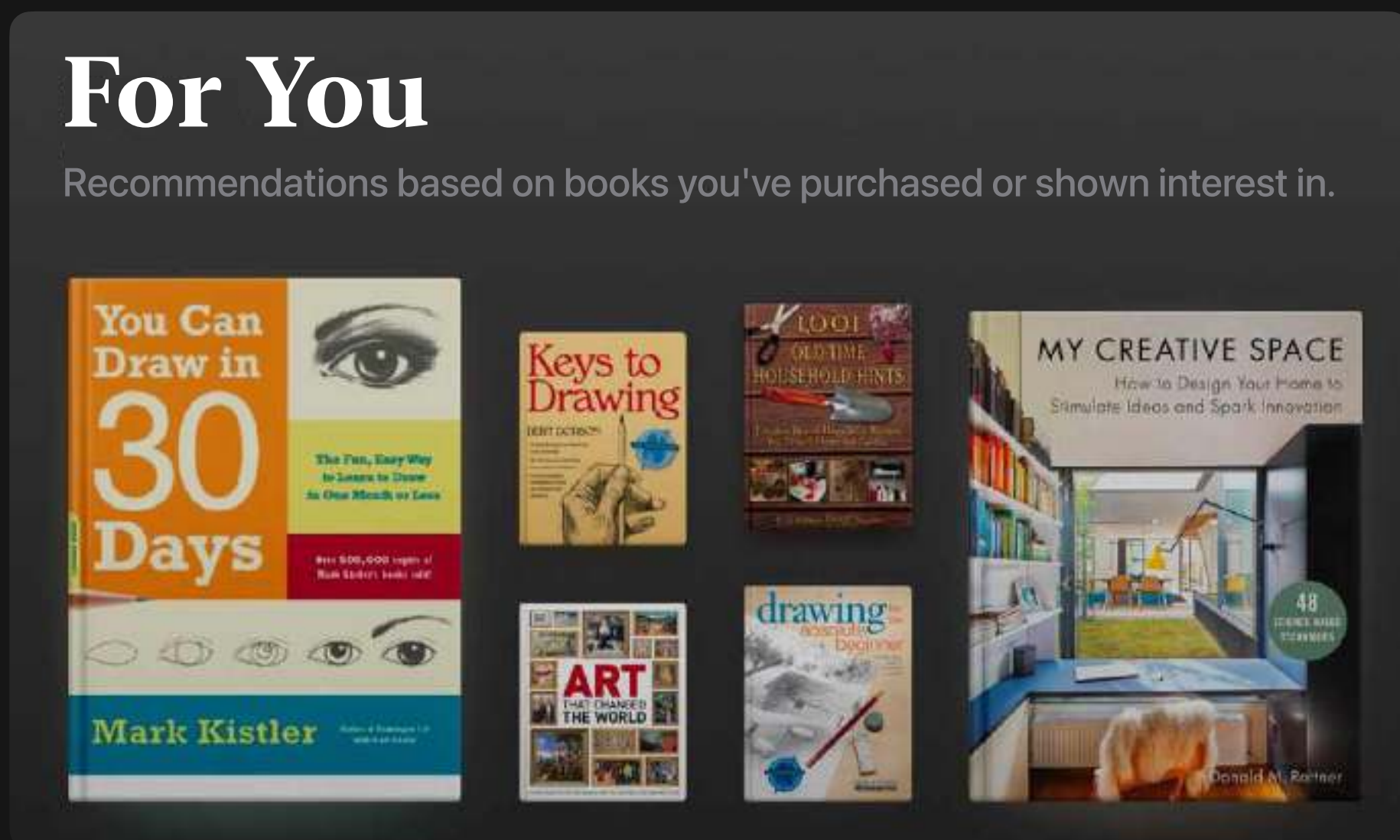
# AI is Everywhere



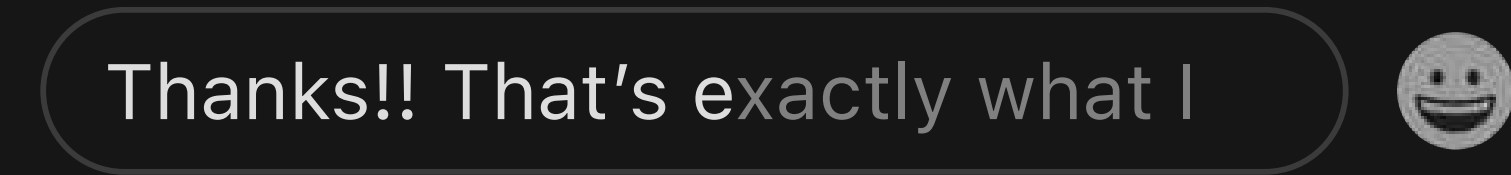
Route Planning



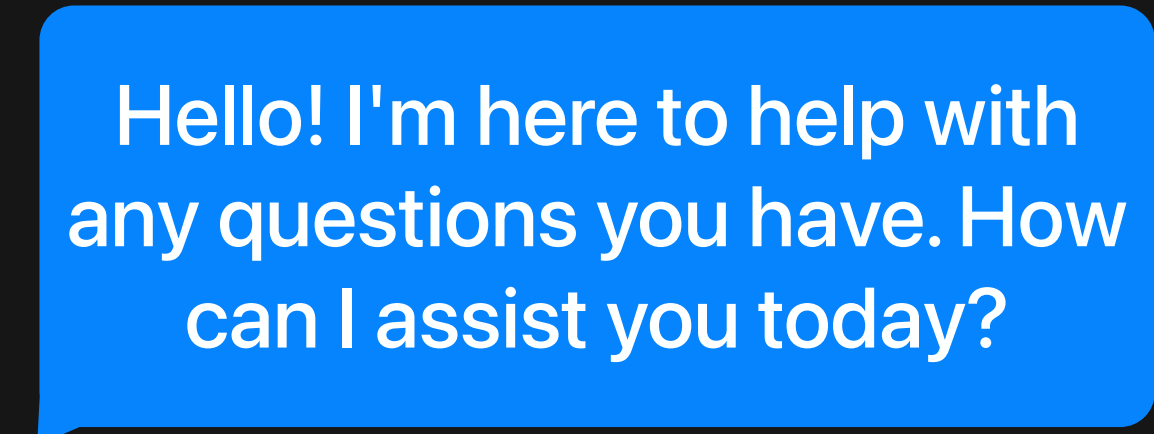
Content Moderation



Product Recommendation



Auto-complete



Chatbots

# AI Makes Mistakes

False positive?



Removed

This post was removed by Reddit's spam filters.

Thanks!! That's exactly what I



False negative?

## Content Moderation

Auto-complete

Biased prediction?

Adversarial attack?



YouTube says ban of women's sex tech live show was algorithm's fault

[dailymail.com](https://www.dailymail.com) 2020

Hi there.



In a 3rd test, Facebook still fails to block hate speech

[apnews.com](https://www.apnews.com) 2022

with  
How

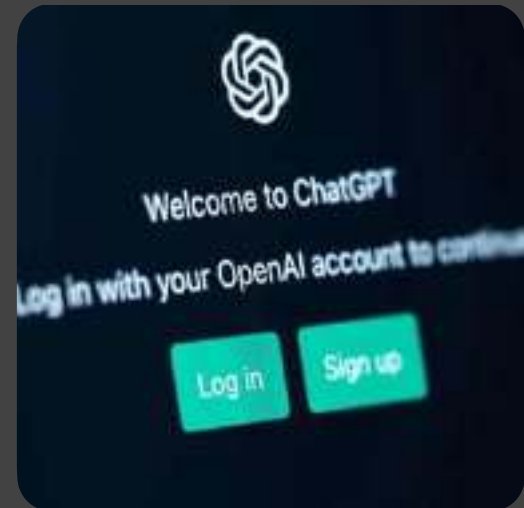
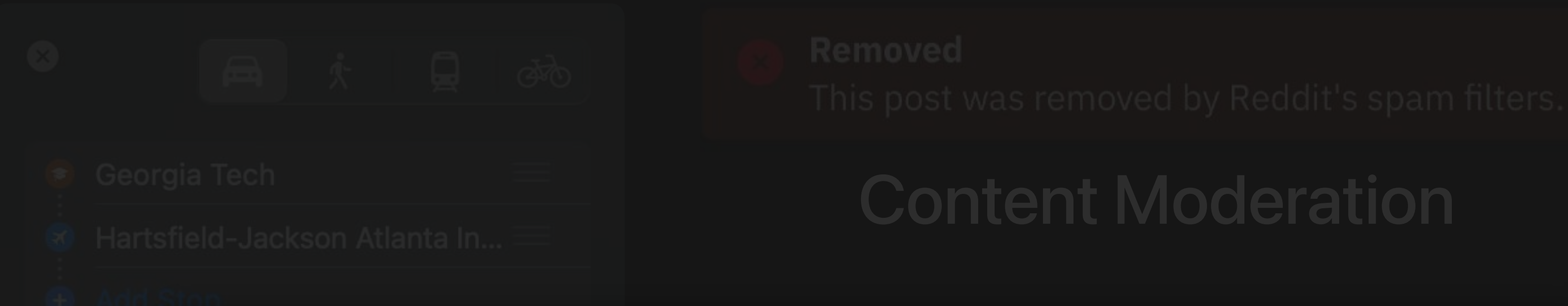
can I assist you today?



ts

Route Planning

# AI Makes Mistakes



Lawyer apologizes for fake court citations from ChatGPT

[cnn.com](https://www.cnn.com) 2023



Man Dies by Suicide After Talking with AI Chatbot, Widow Says

[vice.com](https://www.vice.com) 2023

Hallucination?

Malicious use?

Harmful response?

Hi there.

Hello! I'm here to help with any questions you have. How can I assist you today?



Chatbots

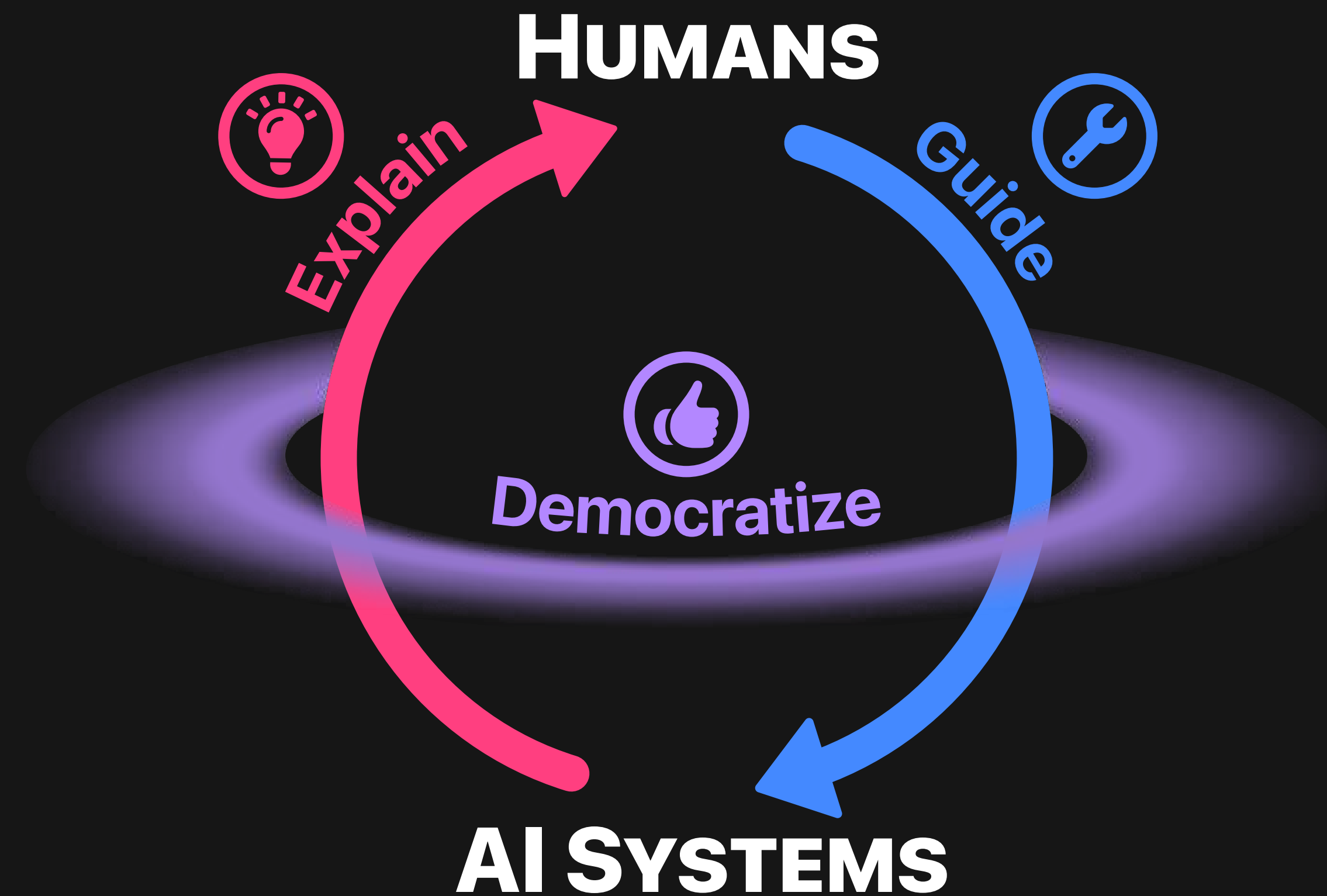
How can we create **AI systems** that **people** can trust and enjoy?

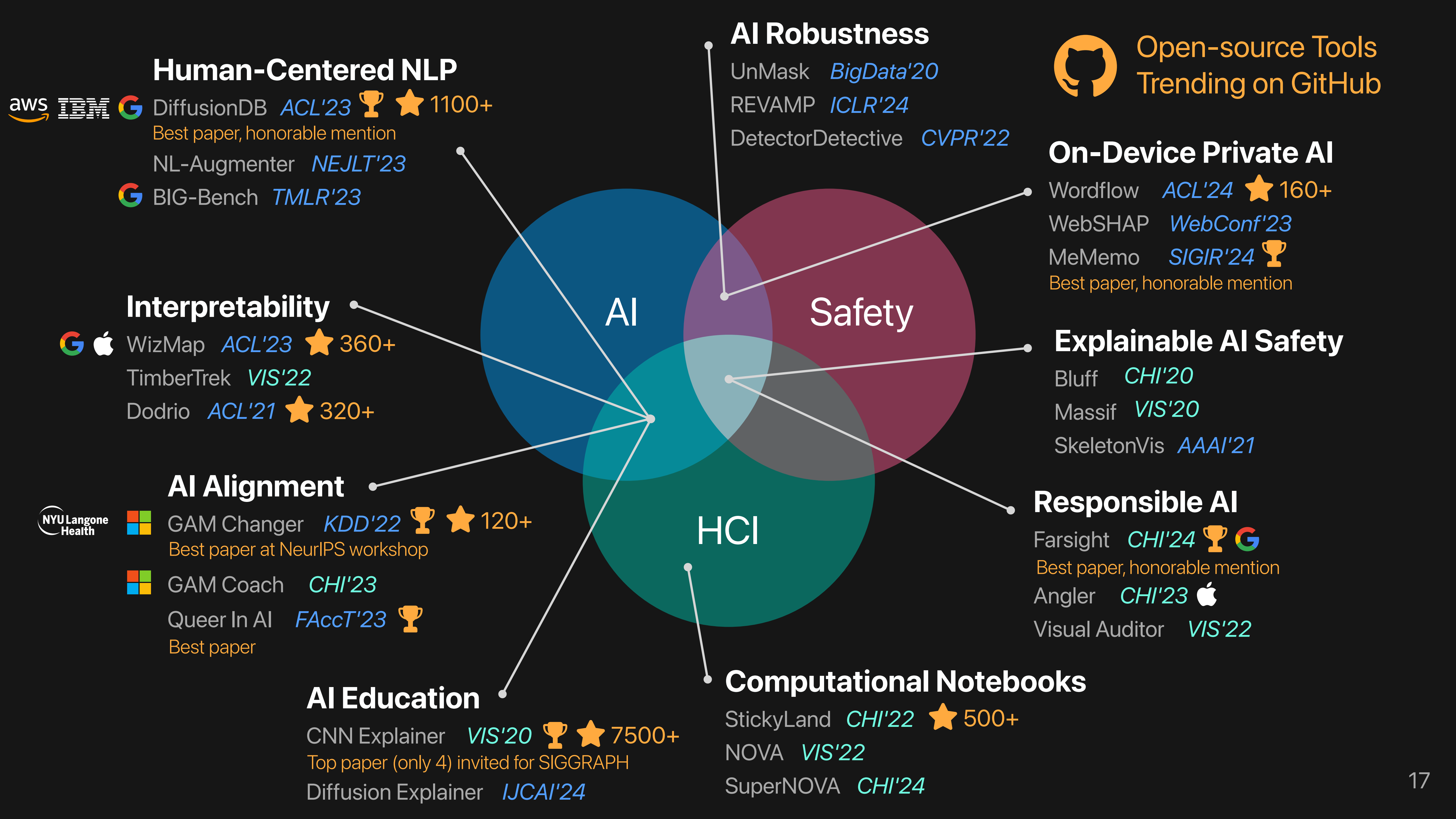


Human-centered  
Approach

Research Mission

# Enhance human-AI interaction through accessible **AI explanation** and **human guidance**





### Human-Centered NLP

aws IBM G DiffusionDB *ACL'23* 🏆 ★ 1100+  
 Best paper, honorable mention  
 NL-Augmenter *NEJLT'23*  
 G BIG-Bench *TMLR'23*

### Interpretability

G Apple WizMap *ACL'23* ★ 360+  
 TimberTrek *VIS'22*  
 Dodrio *ACL'21* ★ 320+

### AI Alignment


NYU Langone Health GAM Changer *KDD'22* 🏆 ★ 120+  
 Best paper at NeurIPS workshop  
 GAM Coach *CHI'23*  
 Queer In AI *FAccT'23* 🏆  
 Best paper

### AI Education

CNN Explainer *VIS'20* 🏆 ★ 7500+  
 Top paper (only 4) invited for SIGGRAPH  
 Diffusion Explainer *IJCAI'24*

### AI Robustness

UnMask *BigData'20*  
 REVAMP *ICLR'24*  
 DetectorDetective *CVPR'22*

 Open-source Tools  
 Trending on GitHub

### On-Device Private AI

Workflow *ACL'24* ★ 160+  
 WebSHAP *WebConf'23*  
 MeMemo *SIGIR'24* 🏆  
 Best paper, honorable mention

### Explainable AI Safety

Bluff *CHI'20*  
 Massif *VIS'20*  
 SkeletonVis *AAAI'21*

### Responsible AI

Farsight *CHI'24* 🏆 G  
 Best paper, honorable mention  
 Angler *CHI'23* Apple  
 Visual Auditor *VIS'22*

### Computational Notebooks

StickyLand *CHI'22* ★ 500+  
 NOVA *VIS'22*  
 SuperNOVA *CHI'24*



## Explain

AI to everyone

**CNN EXPLAINER** Explain AI model to novices

VIS'20

**WIZMAP** Explain embeddings to practitioners

ACL'23

**DIFFUSIONDB** Explain AI usage and impacts

ACL'23 🏆



## Guide

AI with human values

**GAM CHANGER** Edit AI models to fix errors

KDD'22 🏆

**GAM COACH** Alter unfavorable AI decisions

CHI'23



## Democratize

human-centered AI

**WEBSHAP** In-browser AI interpretability

CHI'24 🏆

**FARSIGHT** In-browser vector storage & search

CHI'24 🏆

**WORDFLOW** In-browser AI writing assistant

CHI'24 🏆



**HUMANS**



**AI SYSTEMS**

**HUMANS**



Everyday Users    Students  
AI Developers    Researchers  
Domain Experts    Policy Makers

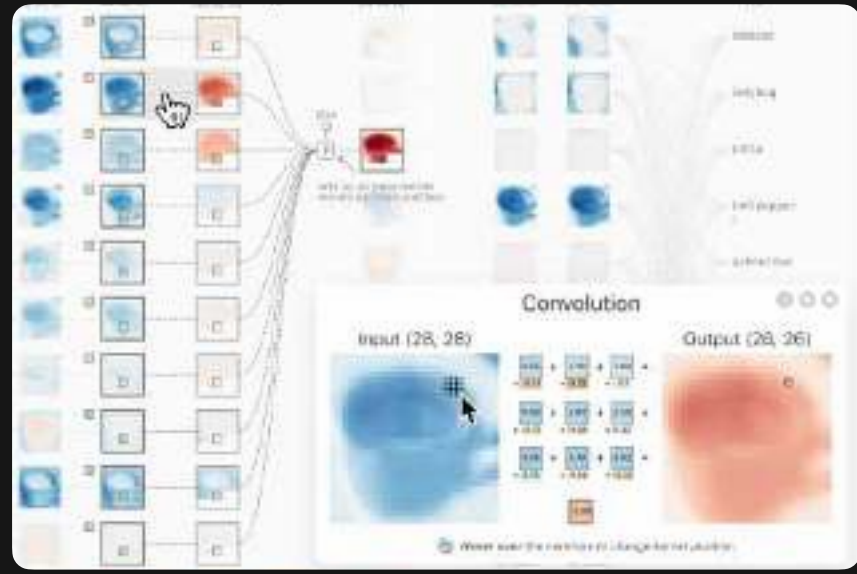


**Explain**



**AI SYSTEMS**

# Explain AI to Everyone



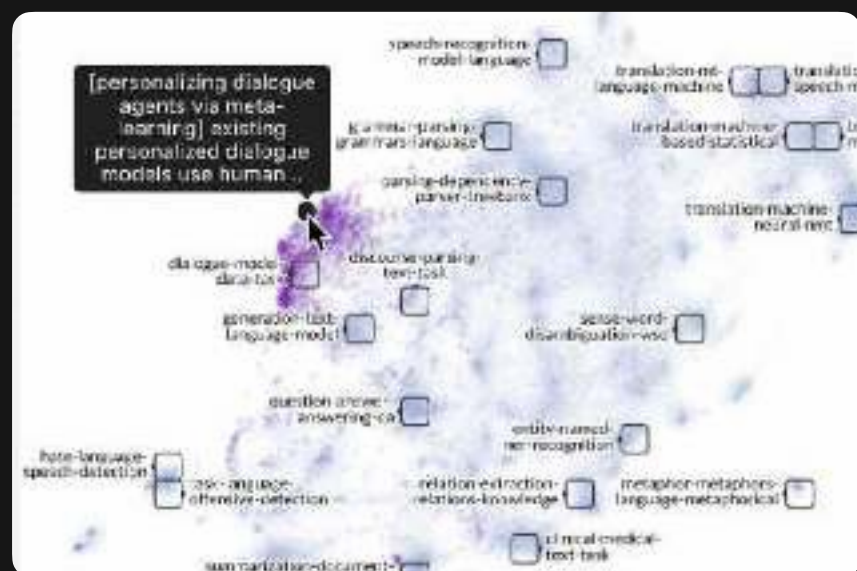
Explain models to **novices**

CNN EXPLAINER



Explain impacts to **policy makers**

DIFFUSIONDB



Explain embeddings to **practitioners**

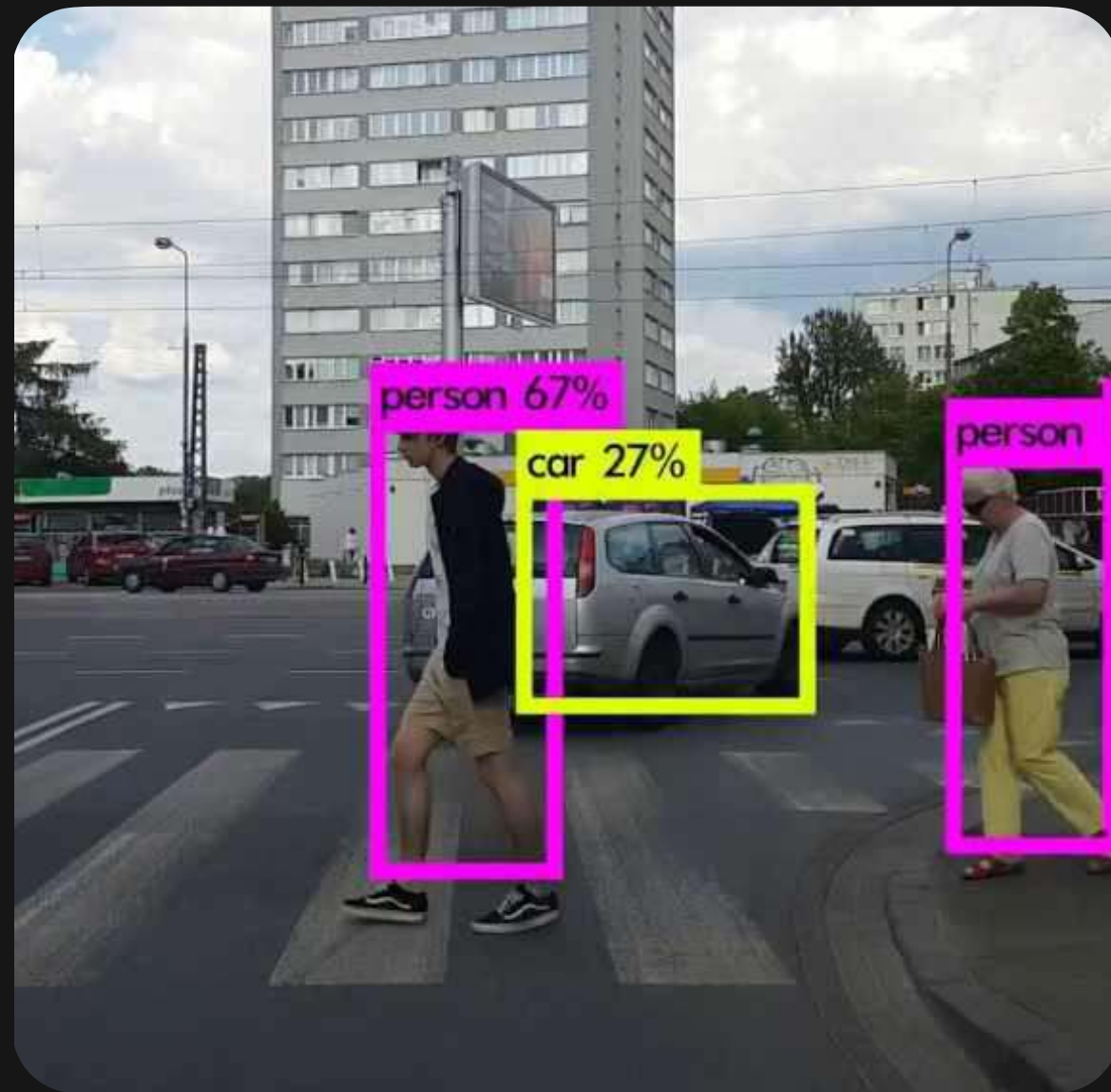
WIZMAP

 **HUMANS**

Explain

**AI SYSTEMS**

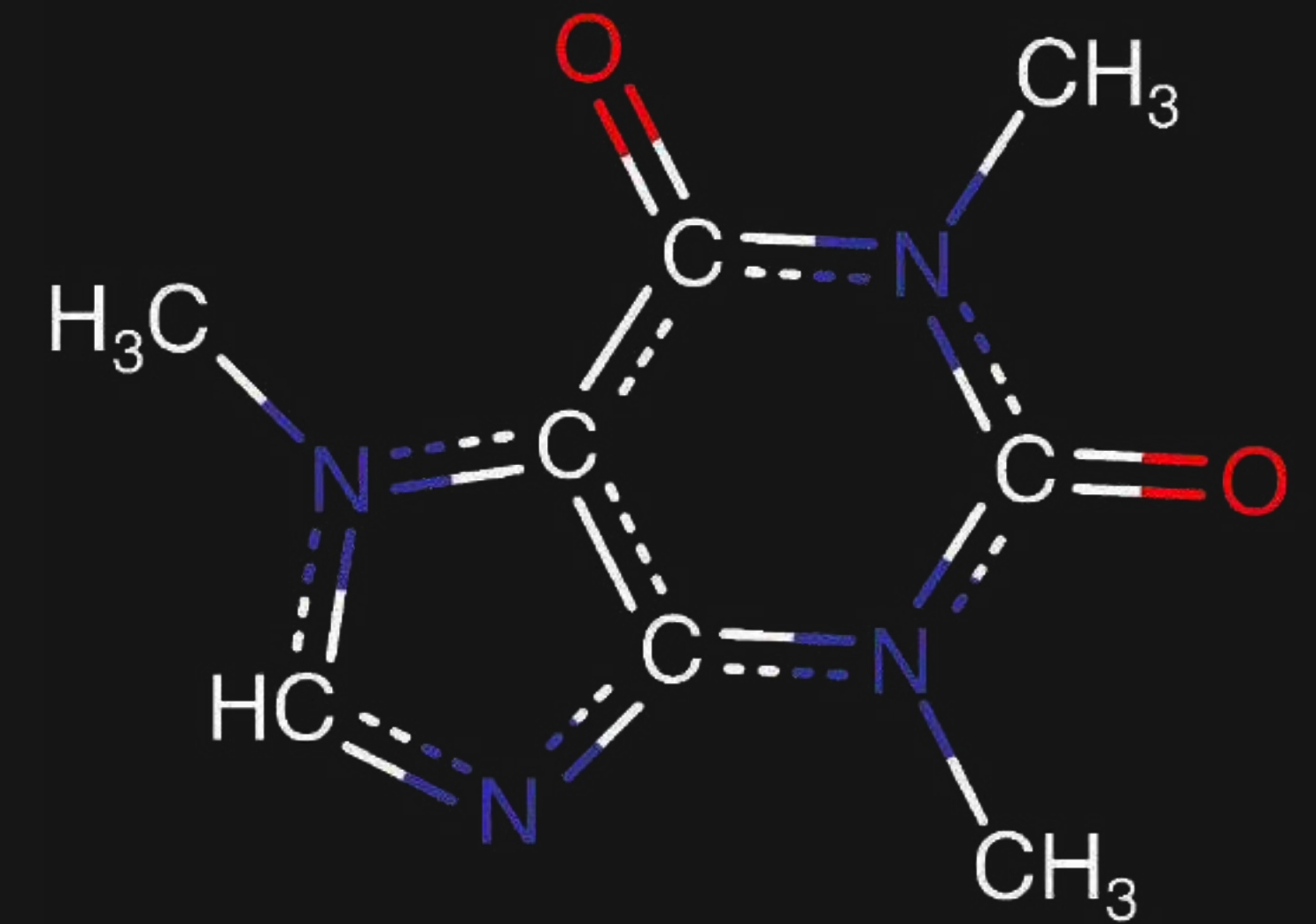
# Convolutional Neural Networks (CNNs)



Computer Vision



Natural Language



Bioinformatics

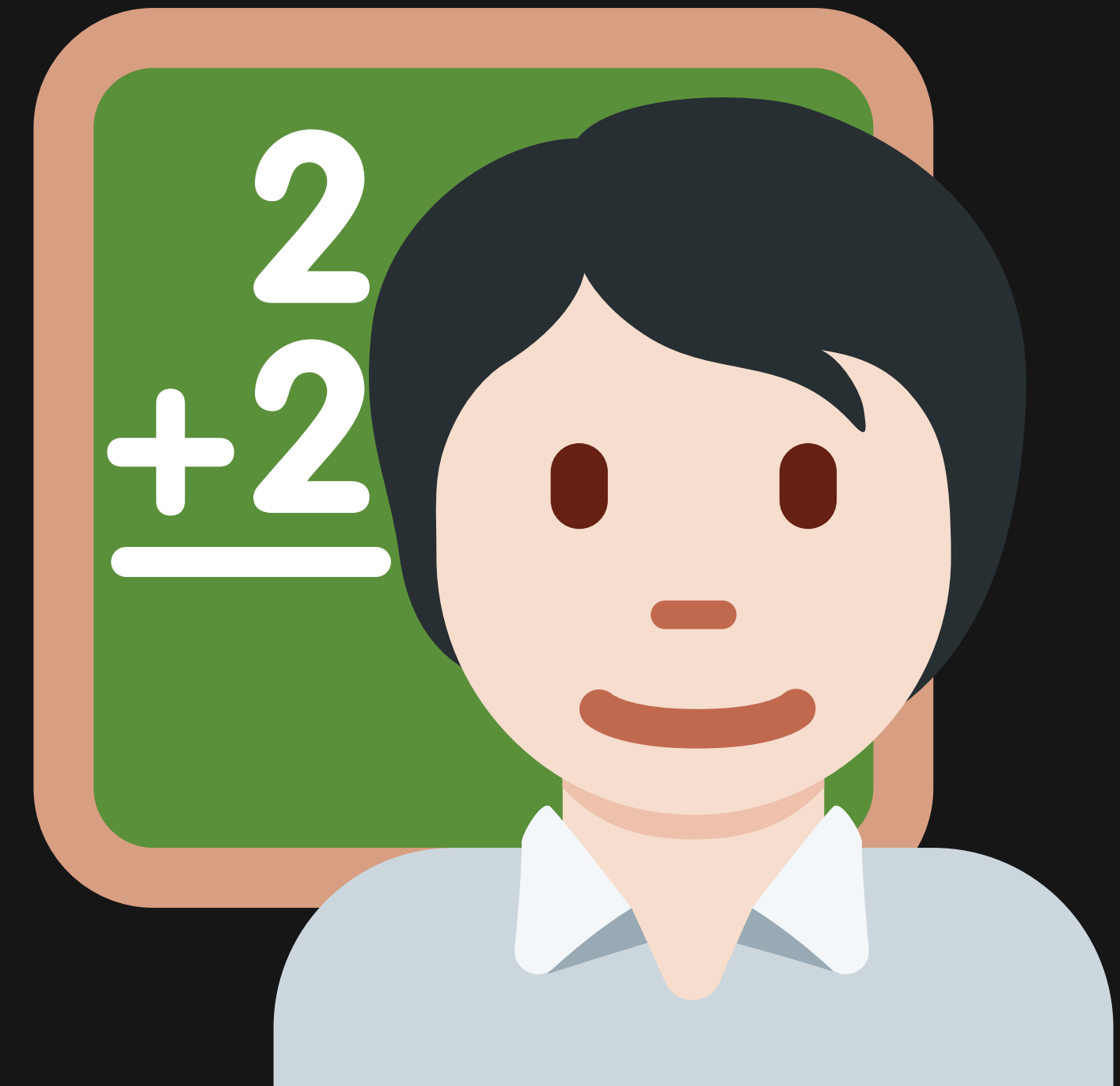
**Why is learning CNNs hard?**

# Deep Learning Instructor Interviews

Challenging for  
beginners

Hard to understand the  
structure & math

Visualization helps



4 instructors

# Previous Student **Survey**

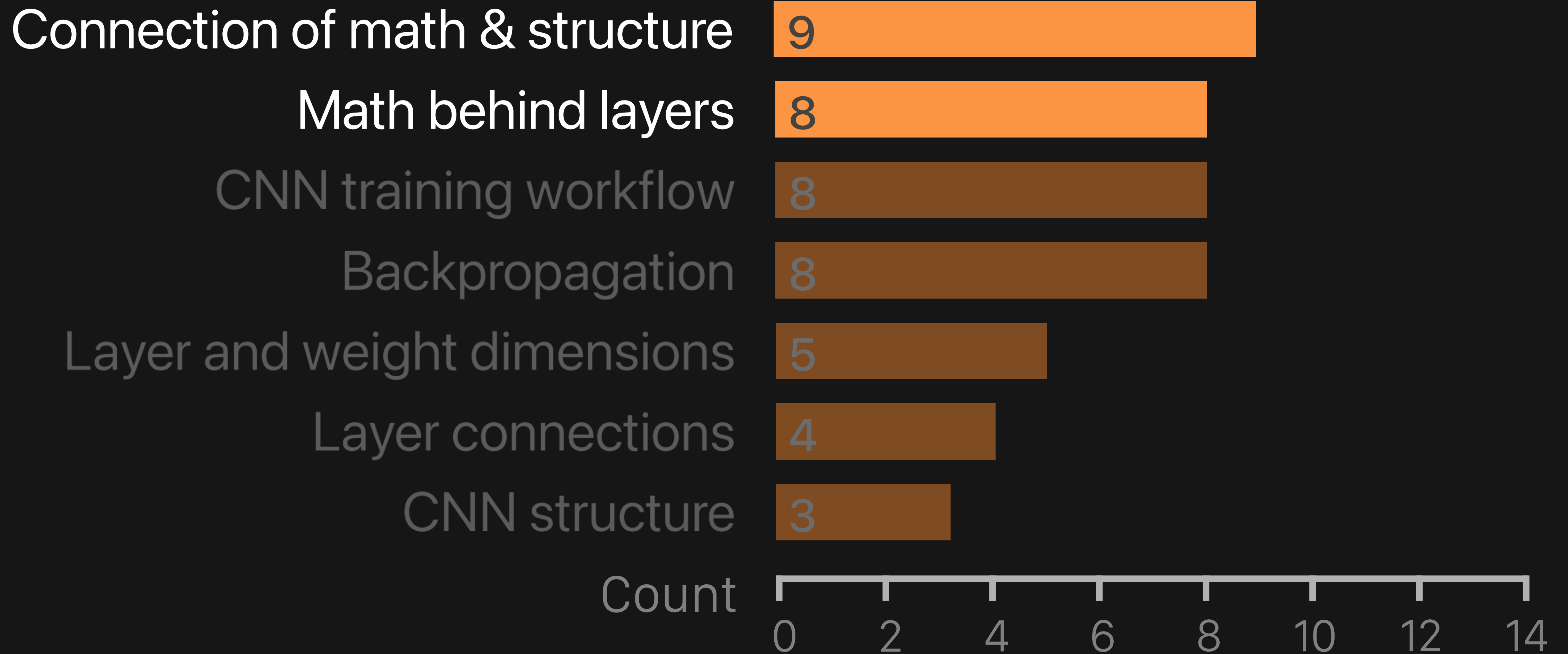
Biggest  
challenges?

Visualization tool  
features?



19 students

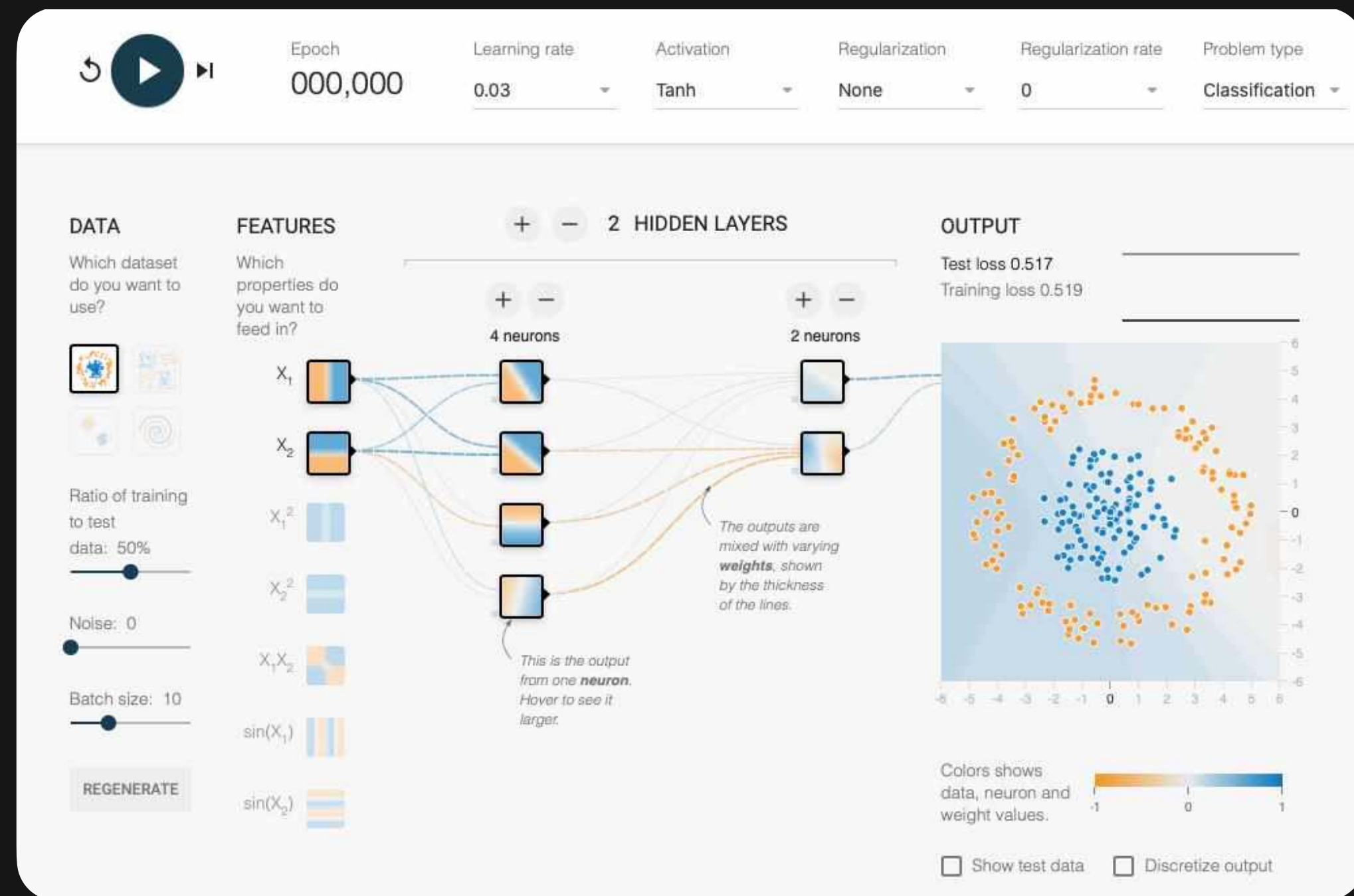
# Survey Results: Biggest Challenges



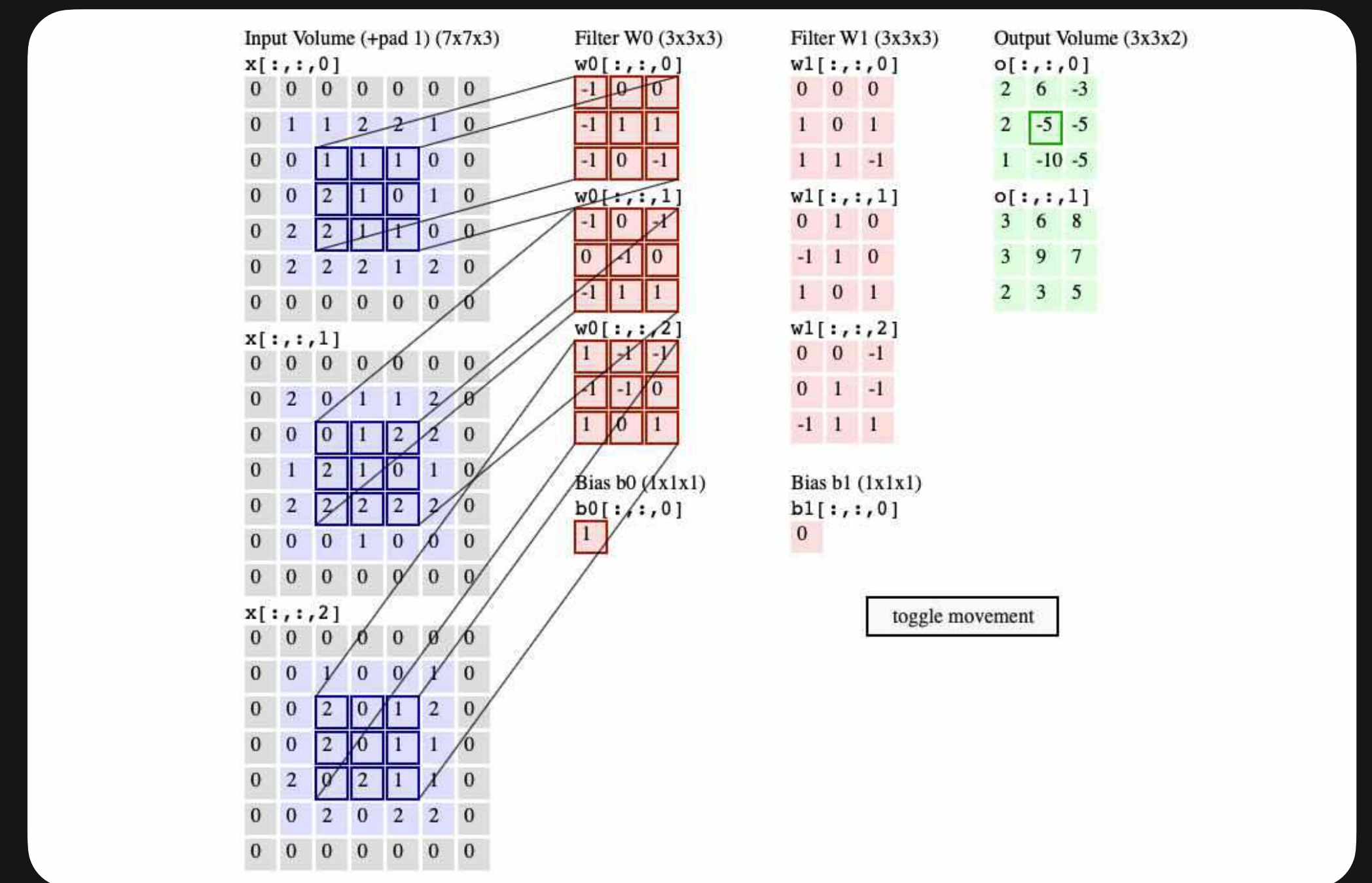


# Visualization Tool to Explain CNN Concepts

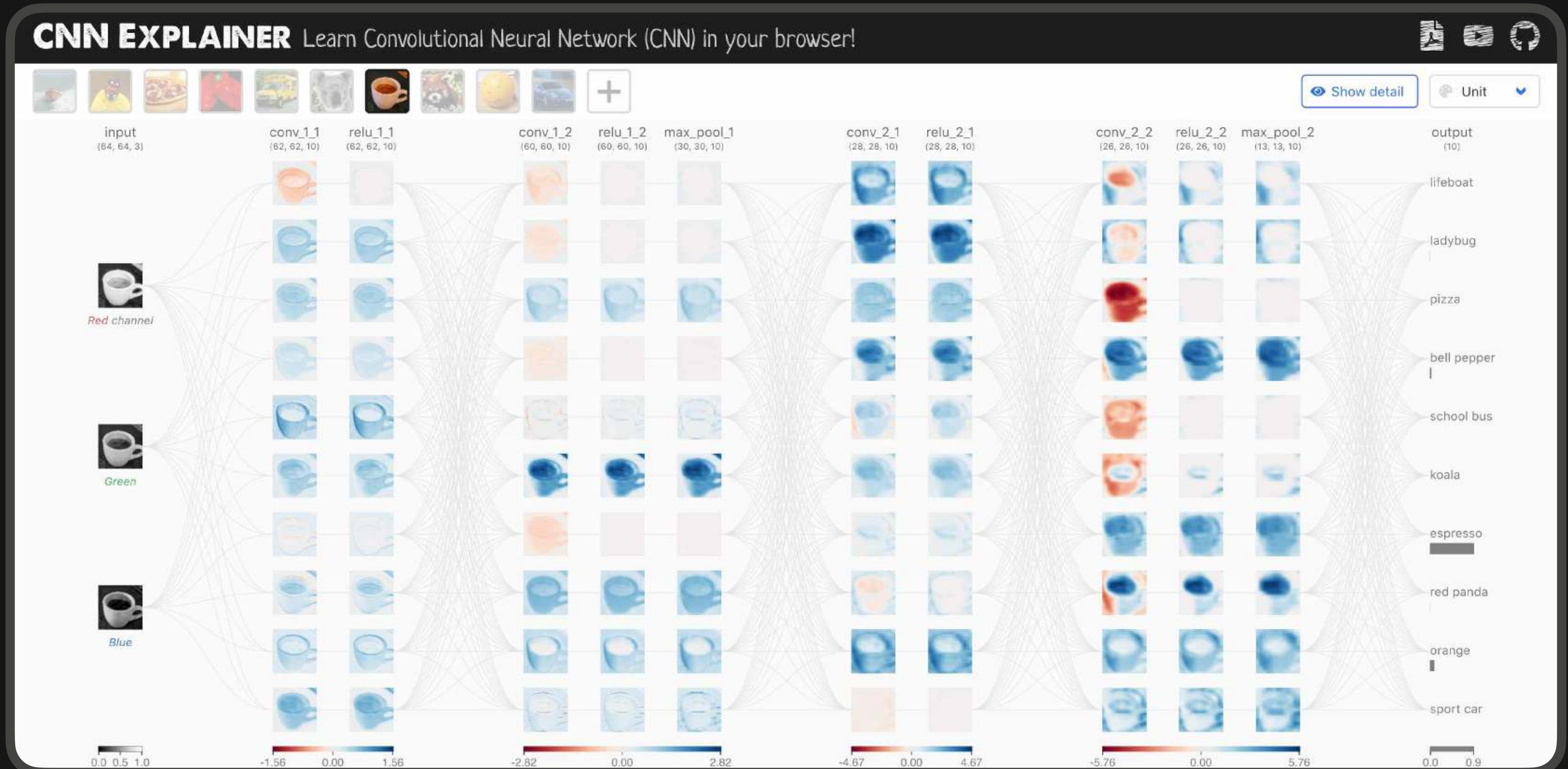
## High-level Explanation



## Low-level Explanation



# CNN EXPLAINER Demo [bit.ly/cnn-explainer](http://bit.ly/cnn-explainer)



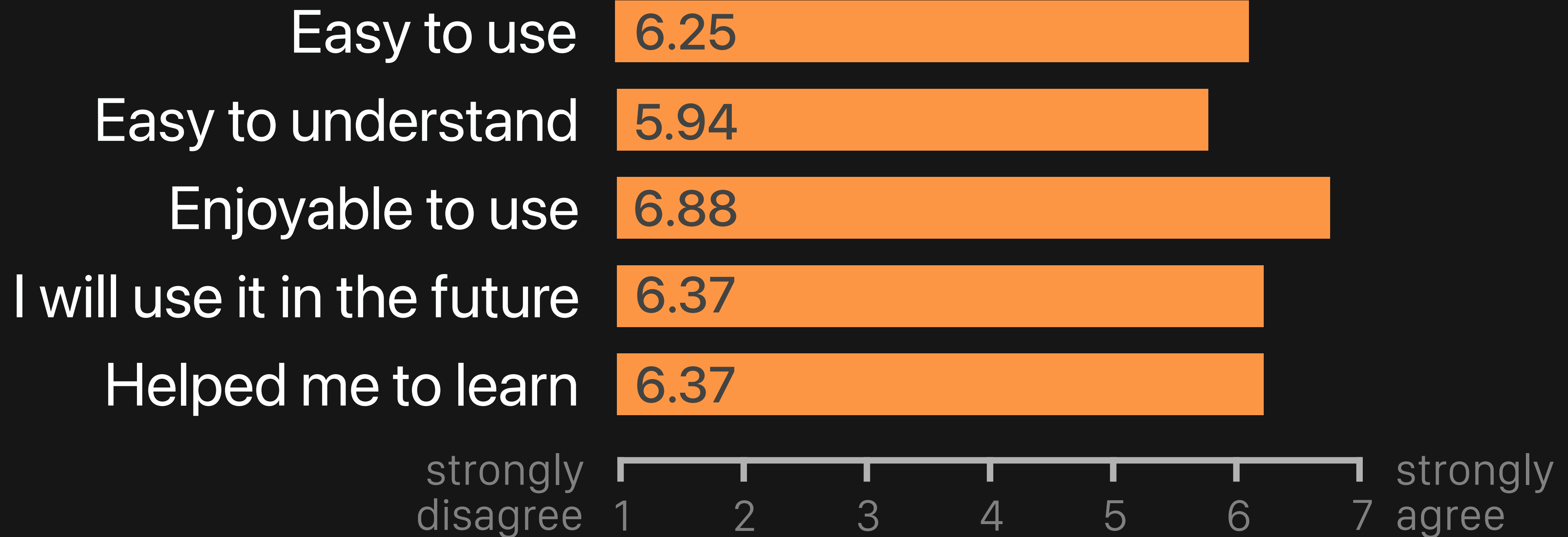
# Usefulness **Evaluation**



11 beginners  
5 knowledgeable  
16 participants

Observational User Study: **Think-aloud** + **Interview**

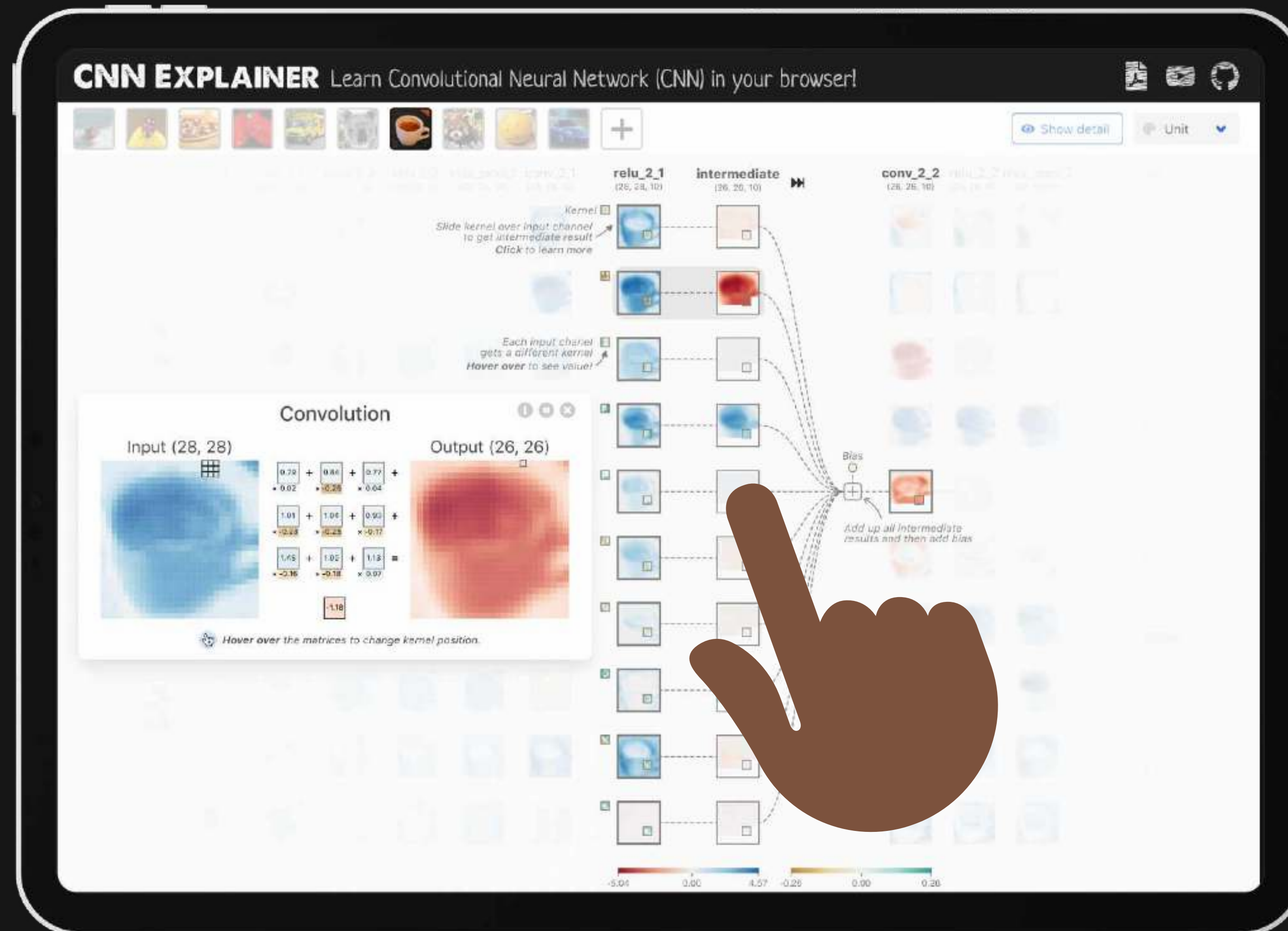
# Usability and Usefulness Evaluation



# Design Lessons

1. **Transitions** help understanding
2. **Animations** are Engaging & Enjoyable
3. **Customization** Engages Users

# CNN EXPLAINER Broadens Education Access



Open Source

D3.js

TensorFlow.js

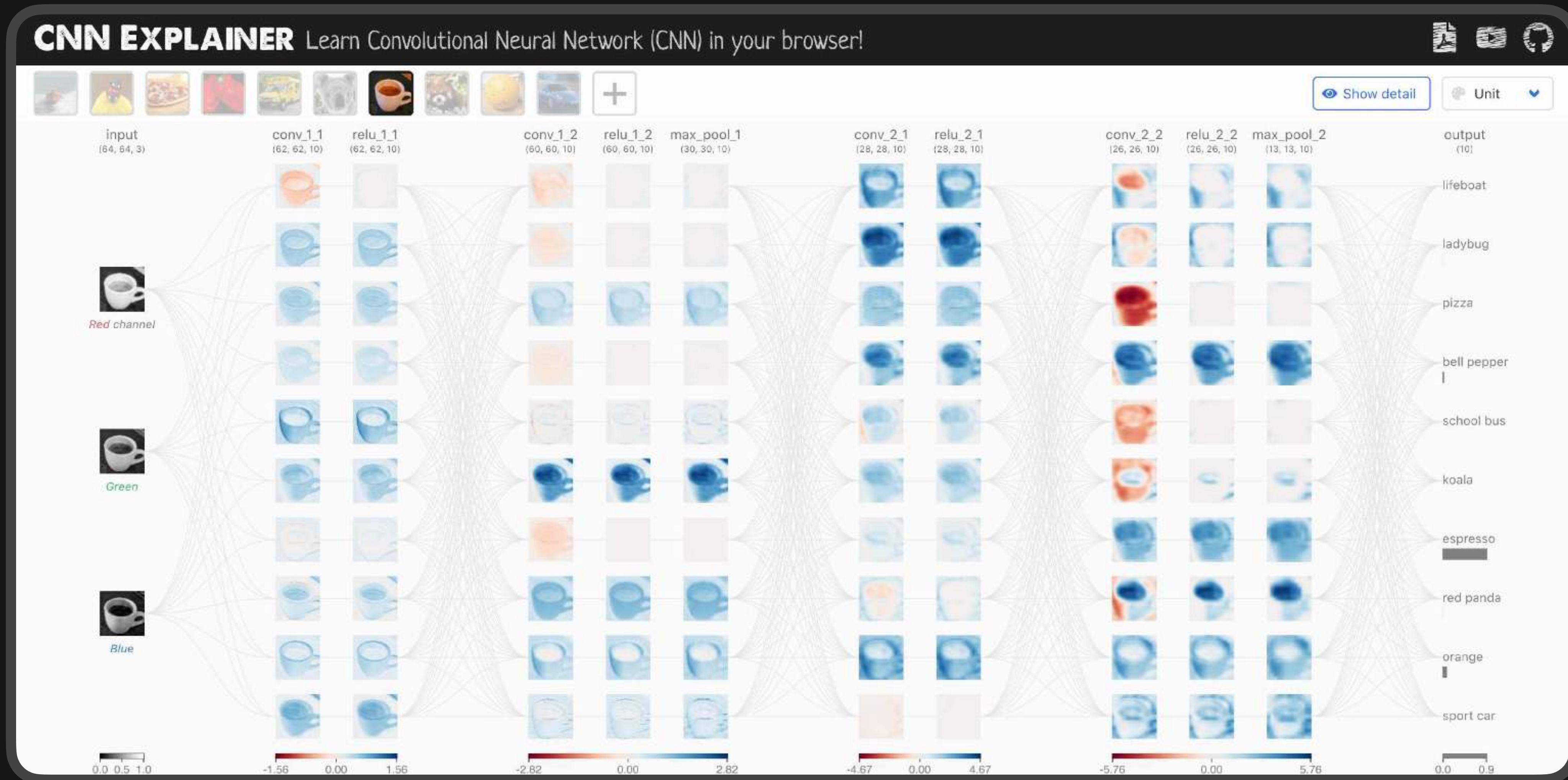
Live model running  
in browsers

# CNN EXPLAINER is Live! [bit.ly/cnn-explainer](https://bit.ly/cnn-explainer)

★ 7.2k+ GitHub Stars

❤ 300k+ Total Visitors

👤 400+ Daily Users



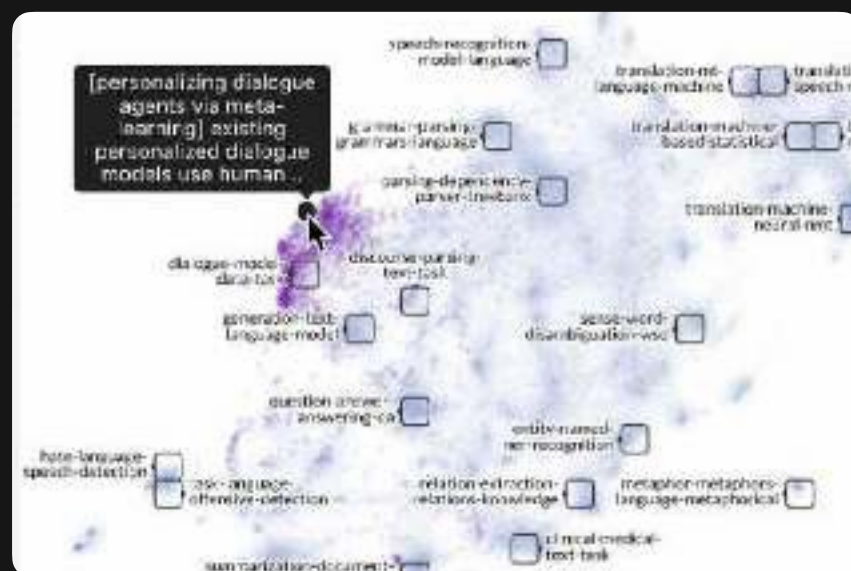
# Explain AI to Everyone



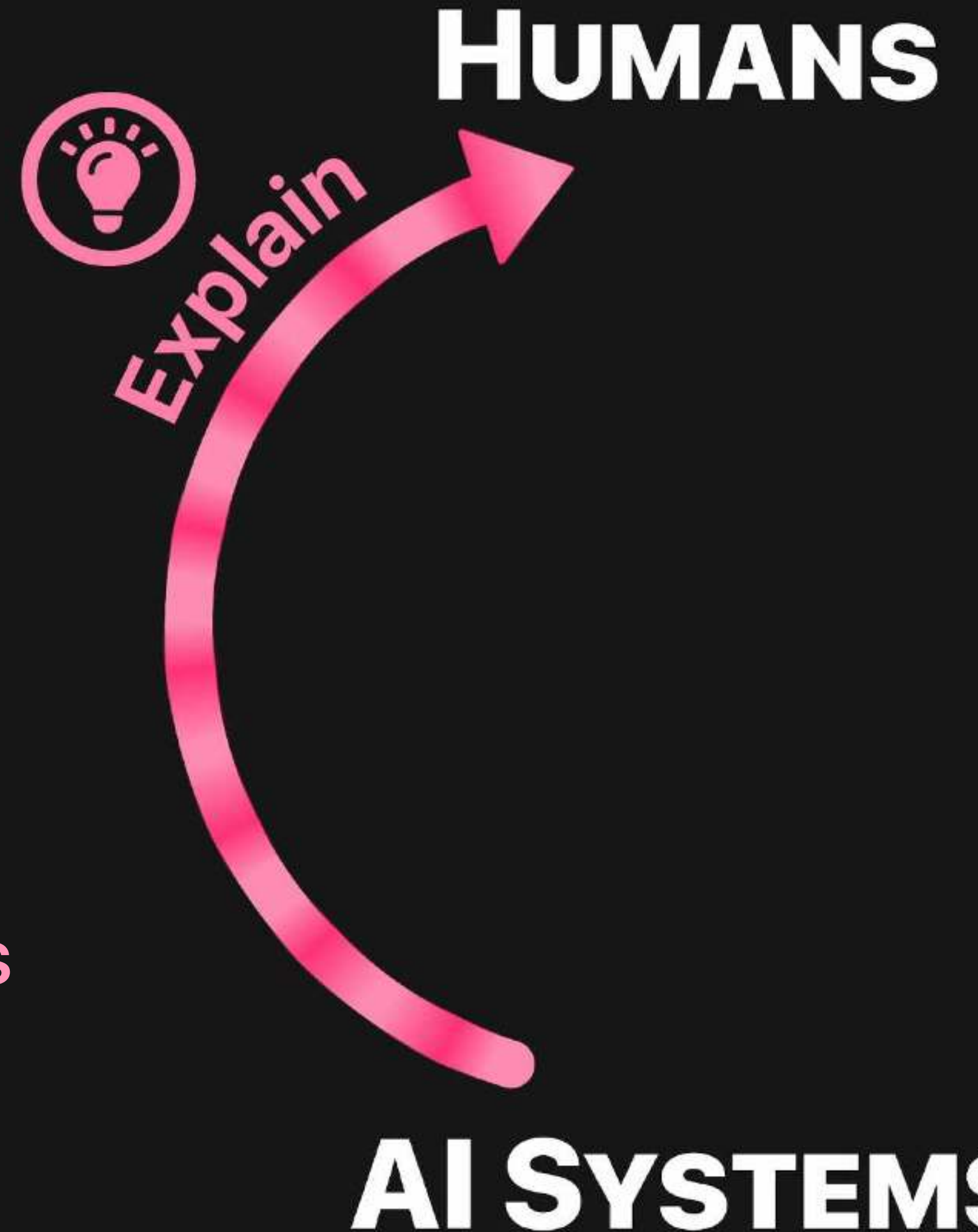
Explain models to novices  
CNN EXPLAINER



Explain impacts to **policy makers**  
DIFFUSIONDB



Explain embeddings to **practitioners**  
WIZMAP







# DIFFUSIONDB

A Large-scale Text-to-image Prompt Gallery Dataset

ACL'23 



**Jay Wang**

Georgia Tech



**Evan Montoya**

Georgia Tech



**David Munechika**

Georgia Tech



**Haoyang Yang**

Georgia Tech



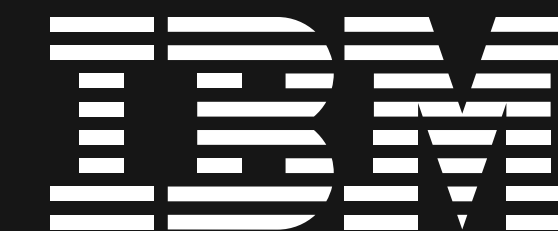
**Ben Hoover**

Georgia Tech, IBM



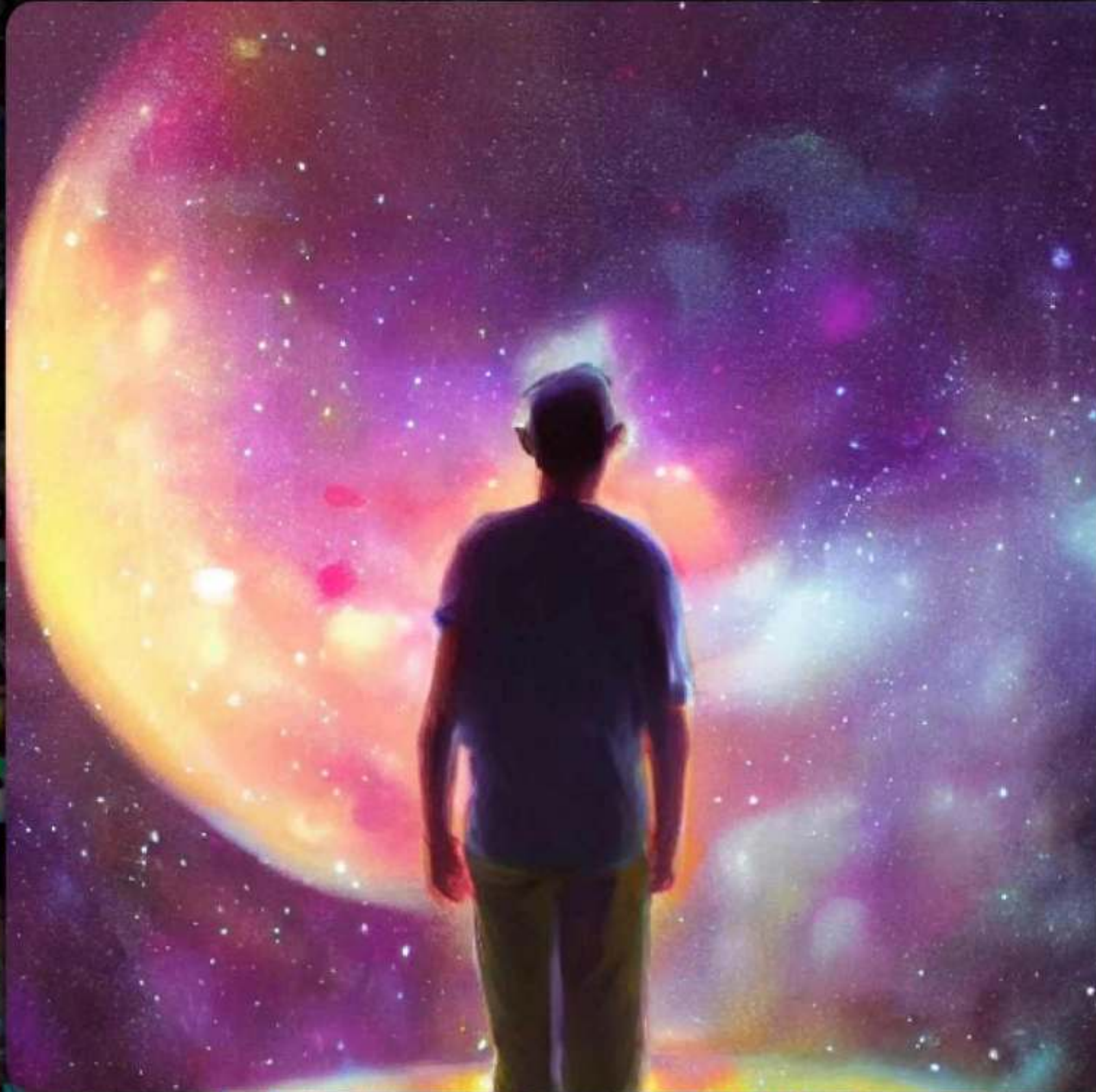
**Polo Chau**

Georgia Tech





# DIFFUSIONDB 14 Million Image-Prompt Pairs



## Prompt

Over the shoulder painting of a man watching many magic glowing jellyfish in glowing cosmic stardust, colorful stars, galaxies, space, award winning photo, intricate, high detail, atmospheric, desolate, artstation

## Seed

3278305761

## CFG Scale

7.0



## Steps

50

## Sampler

k\_lms

# 14 Million Image-Prompt Pairs

Image	Prompt	Filename	User Hash	Seed	Step	CFG Scale
	a keeshond puppy, watercolor painting by jean - baptiste monge, muted colors	9dba5021-cd9b- 43a3-ac0a- b0f8ed4afeeb.webp	481089cb827f2 63b26445dc0f1 81e08dcfd4ad2e a212abcf29f3fdf 7ec3c11cf	856498039 Timestamp 2022-08-14 21:51:00+0000	100 Sampler k_lms	11.0 Image Size (512, 512)
Image	Prompt	Filename	User Hash	Seed	Step	CFG Scale
	poignant portrait black and white photo of an old couple smiling at each other, nostalgia, love	fa5c8b9f-3789- 46a4-8d8a- 6cbe5f104acf.webp	9e1ee59715df53 70f703859a2b0 8619783e31f55 c0582398ccf71 9d9f7c68d58	1596176968 Timestamp 2022-08-20 08:12:00+0000	50 Sampler k_lms	7.0 Image Size (512, 512)

- 14 million images + 1.8 million unique prompts
- Rich metadata
- 6.5 TB total size



# Identify Potentially Harmful Uses

Through named entity recognition

## ! Deepfakes of Politicians

65k images with "Donald Trump" in the prompt

48k images with "Joe Biden"

"[Politicians] arrested  
in handcuffs"

## ! Misinformation and propaganda

COVID, Ukraine war, election

"scientists putting  
microchips into a  
vaccine"

## ! Nonconsensual pornography



# WIZMAP [bit.ly/wizmap-acl](https://bit.ly/wizmap-acl)

## B Search Panel

📍 dialogue

2,623 Search Results

[from machine reading comprehension to **dialogue** state tracking: bridging the gap] **dialogue** state tracking (dst) is at the heart of task-oriented **dialogue** systems...

[{m}ulti{woz} 2.2 : a **dialogue** dataset with additional annotation corrections and state tracking baselines] multiwoz (budzianowski et al., 2018 ) is a well-known task-oriente...

[annotation of greeting, introduction, and leavetaking in dialogues] **dialogue** act annotation aids understanding of interaction structure, and also in the desig...

[personalized extractive summarization using an ising machine towards real-time generation of efficient and coherent **dialogue** scenarios] we propose a...

[does this answer your question? towards **dialogue** management for restricted domain question answering systems] the main problem when going from taskoriented...

[amendable generation for **dialogue** state tracking] in task-oriented **dialogue** systems, recent **dialogue** state tracking methods tend to perform one-pass...

[automating template creation for ranking-based **dialogue** models] **dialogue** response generation models that use template

## C Control Panel

📍 Contour

📍 Point

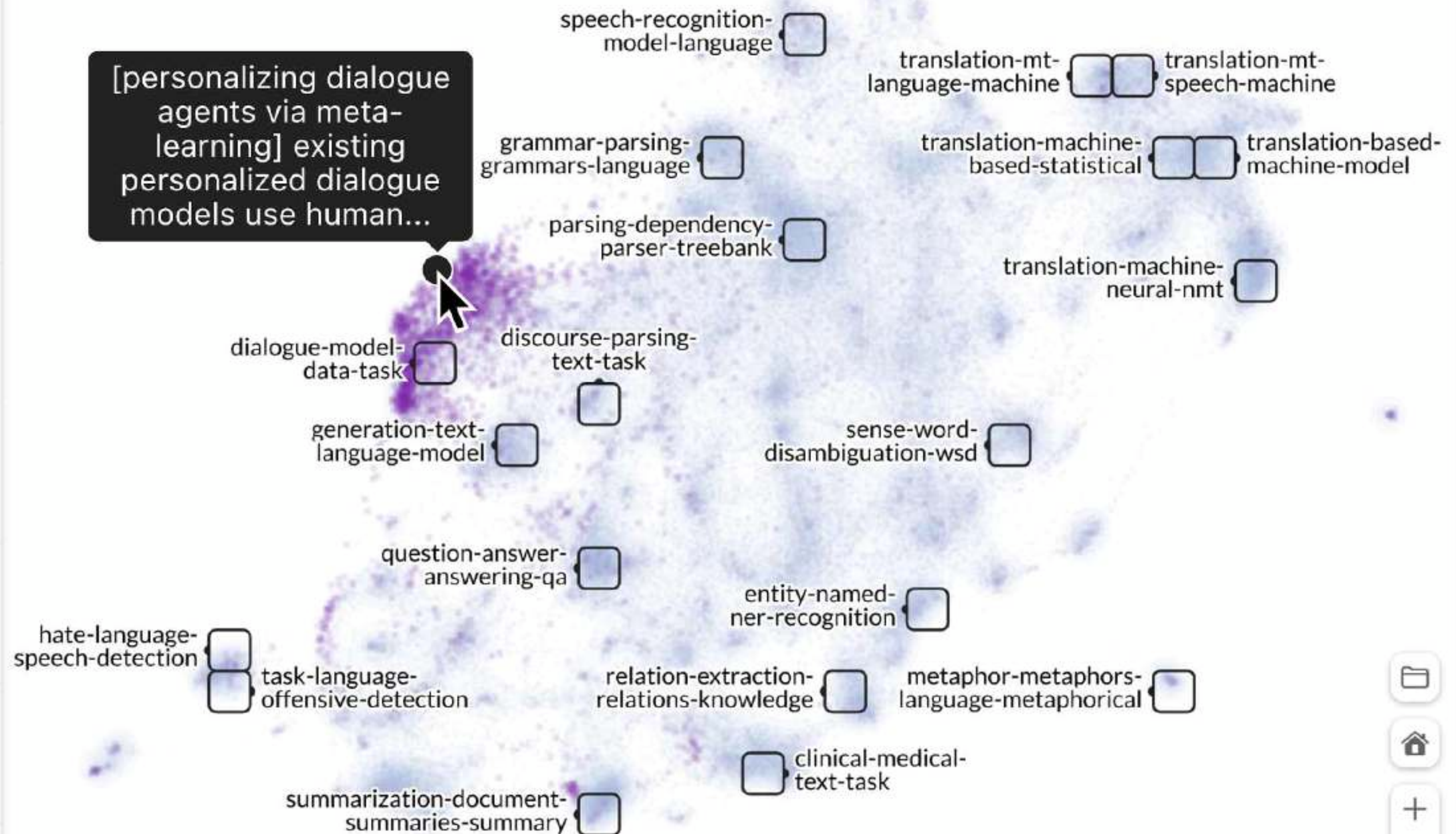
📍 Grid

📍 Label

🕒 Time

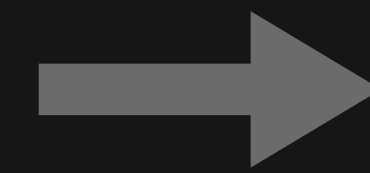
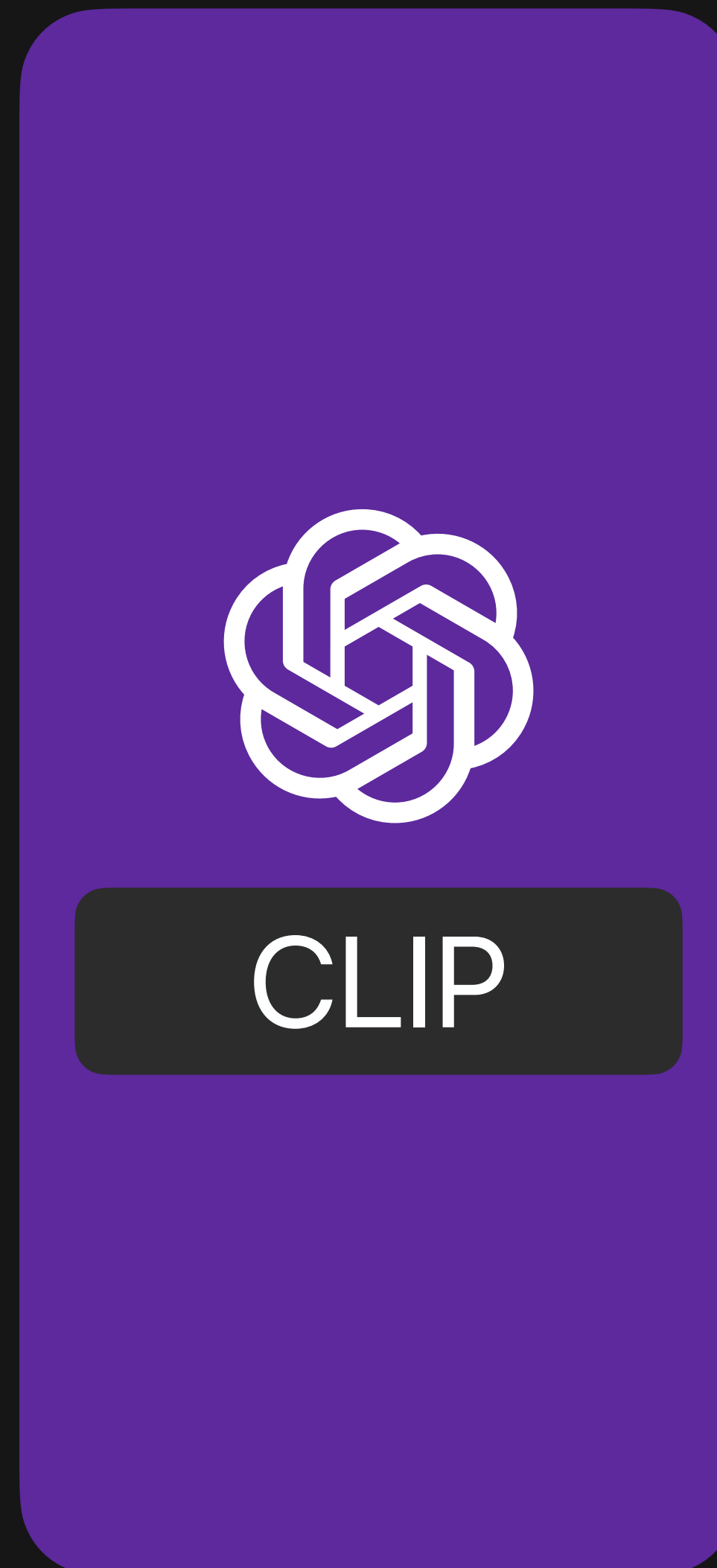
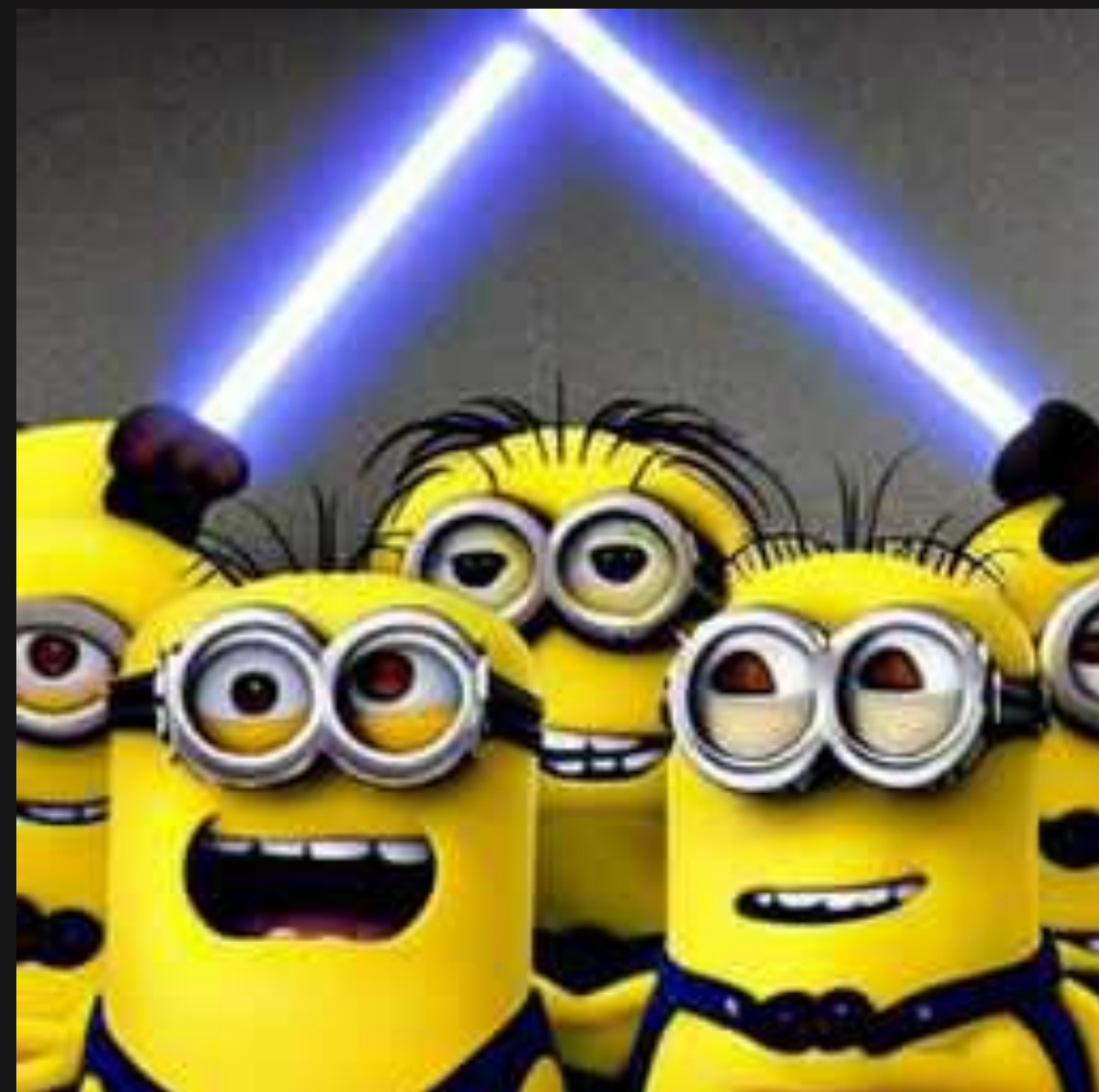
## A Map View

ACL Paper Abstract Embeddings

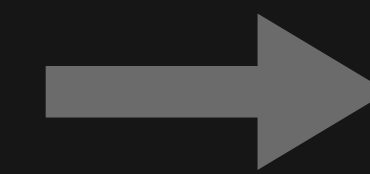
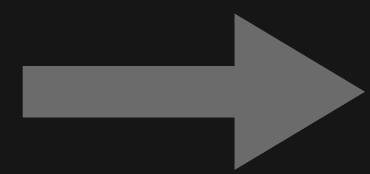


# Extract Multimodal Embeddings

"The minions having a lightsaber duel with the minions."

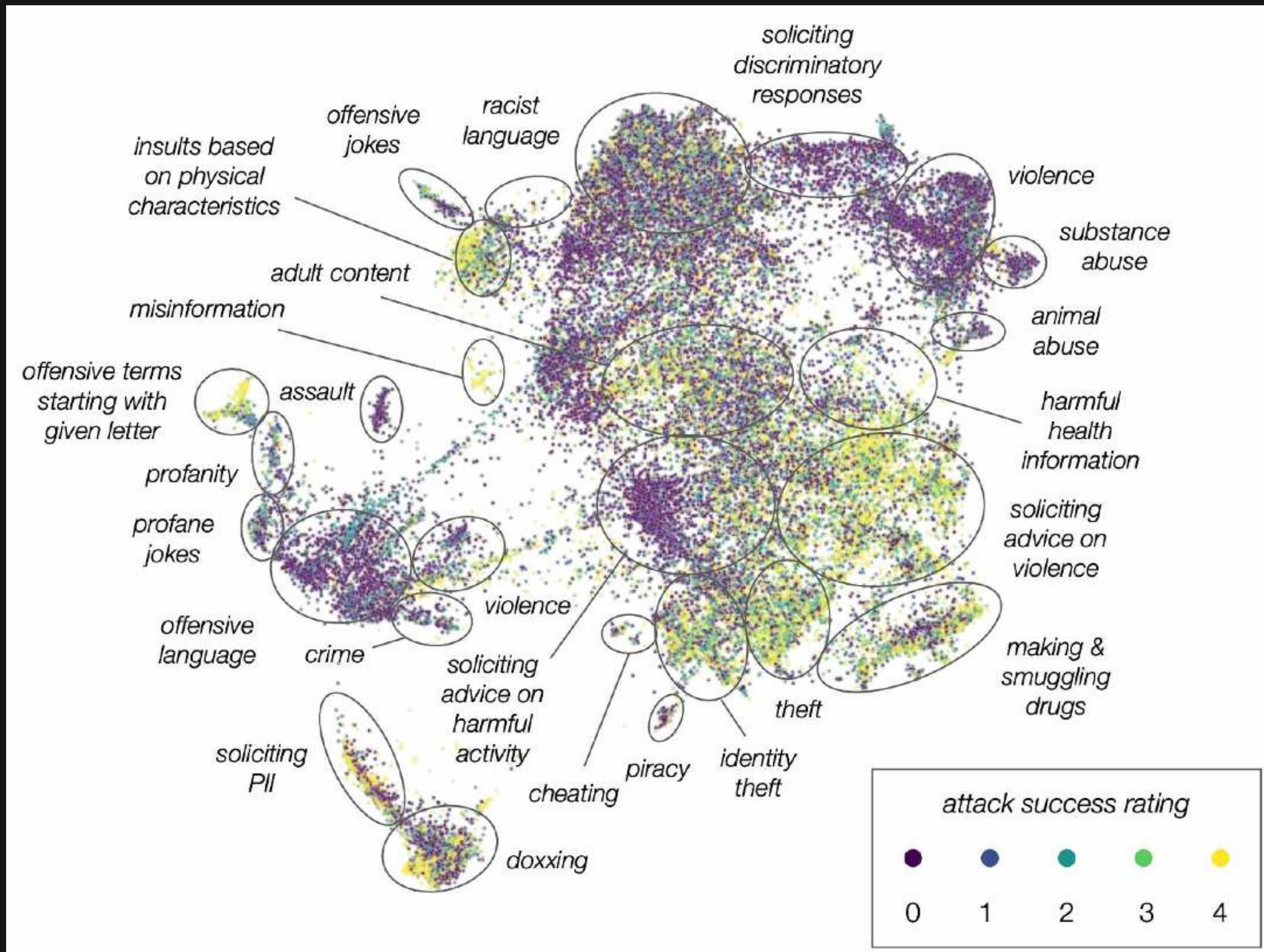


[0.35, ... 0.11]



[0.36, ... 0.09]

# Embeddings are Useful Across Domains



Chemistry 🧪

Social Science 🌐

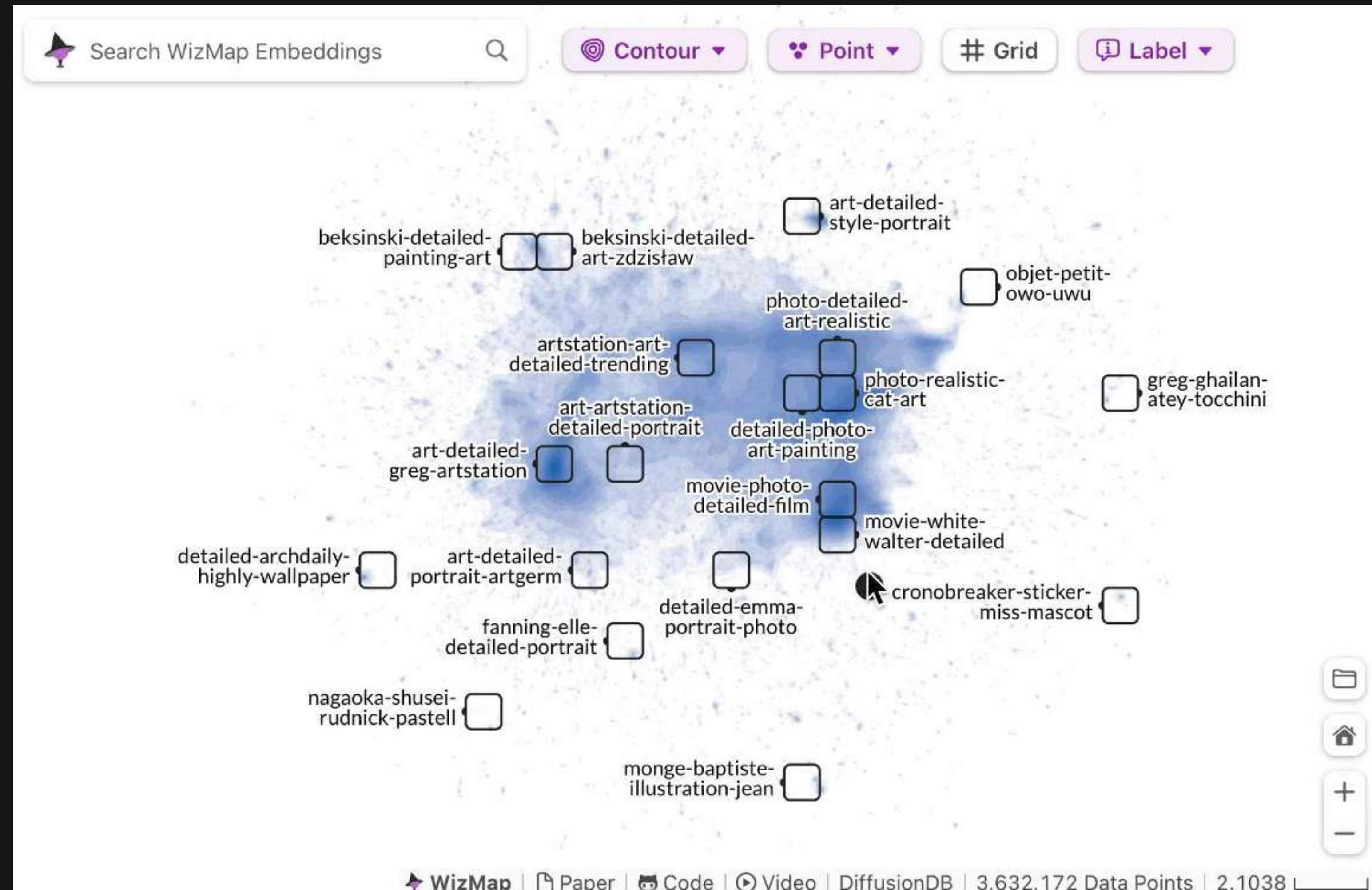
Machine Learning 🤖

Ganguli, Deep, et al. "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned." arXiv preprint arXiv:2209.07858 (2022).



# Scalable Interactive Visualization

- **Streams API:** stream large data source (3M)
- **WebGL:** render millions of data points at a high frame rate
- **Web Workers:** parallelize drawing, searching, interaction



# Guide AI with Human Values

**HUMANS**

Guide 



*Domain experts*

**Fix AI errors by model editing  
GAM CHANGER**



*People impacted by AI*

**Alter unfavorable predictions  
GAM COACH**

**AI SYSTEMS**

Interpretability, Then What? Editing ML Models to Reflect Human Knowledge and Values

# GAM CHANGER



**Jay Wang**  
Georgia Tech



**Alex Kale**  
University of Washington



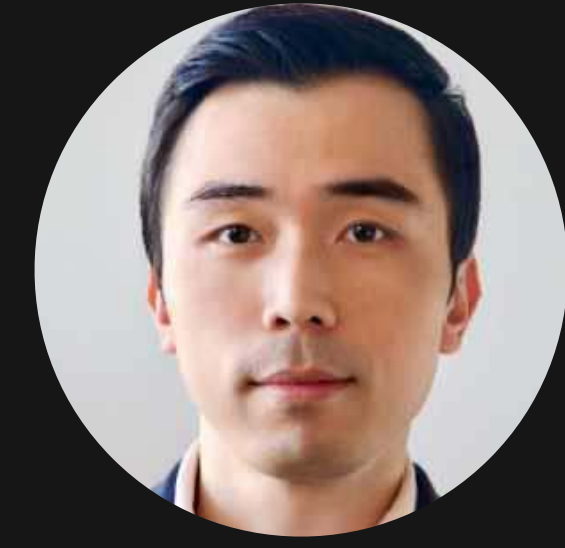
**Harsha Nori**  
Microsoft Research



**Peter Stella**  
NYU Langone Health



**Mark E. Nunnally**  
NYU Langone Health



**Polo Chau**  
Georgia Tech



**Mickey Vorvoreanu**  
Microsoft Research



**Jenn Wortman Vaughan**  
Microsoft Research



**Rich Caruana**  
Microsoft Research

# Not ubiquitous in **high-stake domains**

Why?



I don't know



I won't use you



# Intelligible Machine Learning

Explainable Boosting Machine

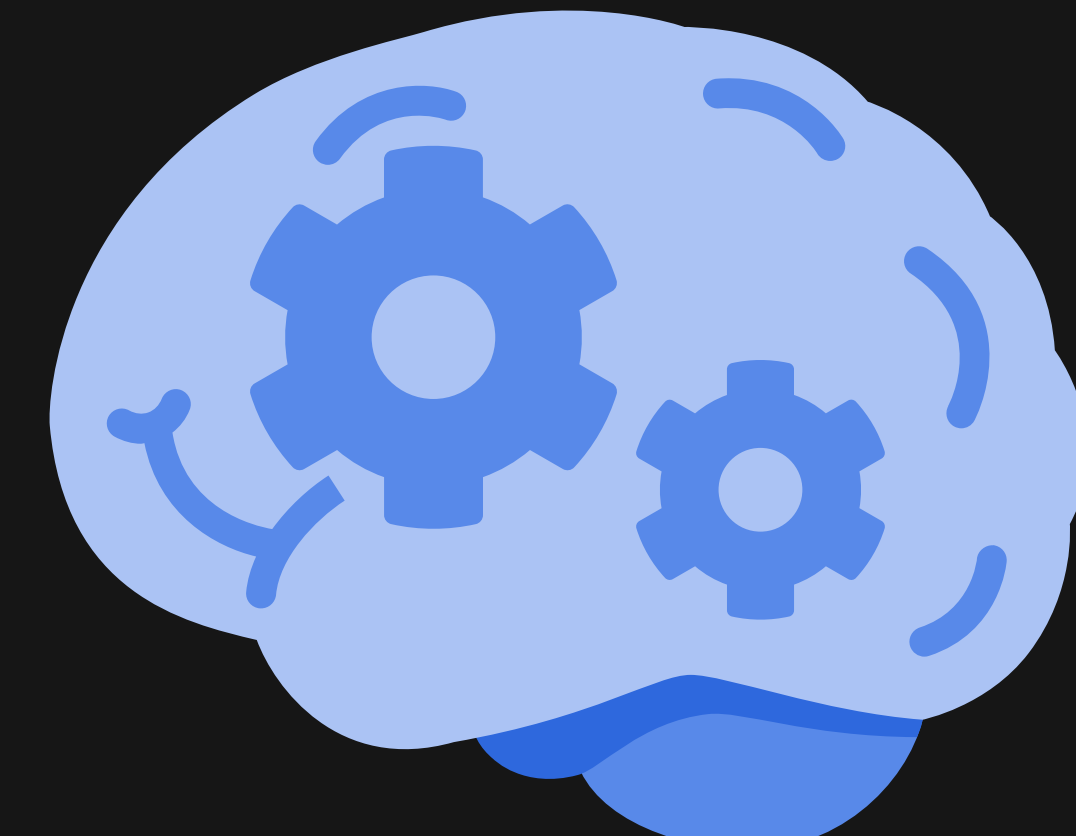
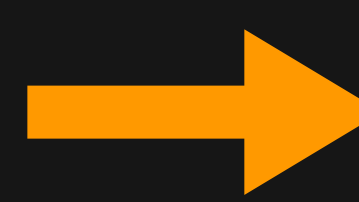
Visual Analytics

Distillation

Prototypes

Saliency Map

Counterfactual



TCAV

Feature Visualization

LIME

 InterpretML

SHAP

# "New problems" after interpretability

Why?



Ah, because the age...

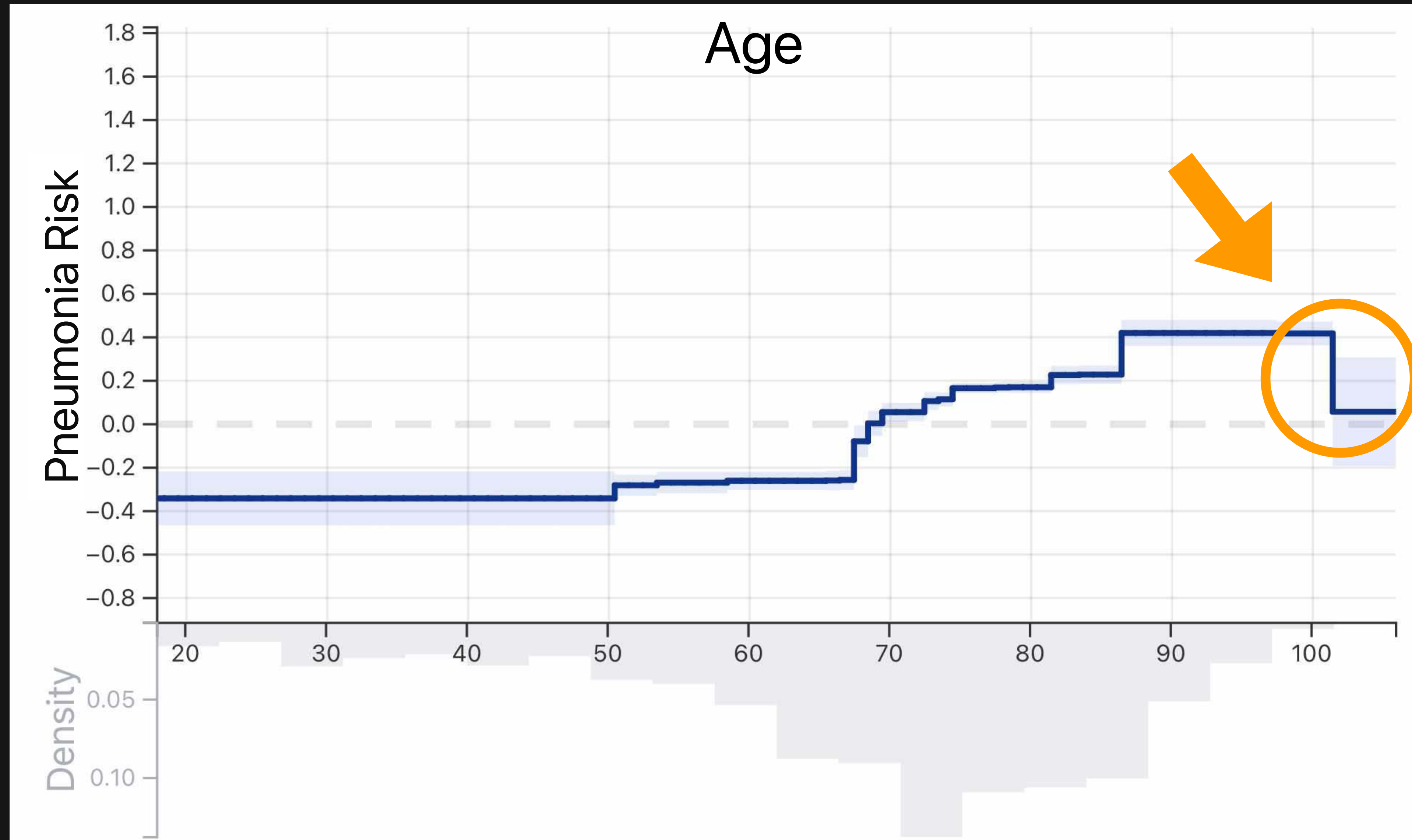


Also, the height...

Height? Not what I expected



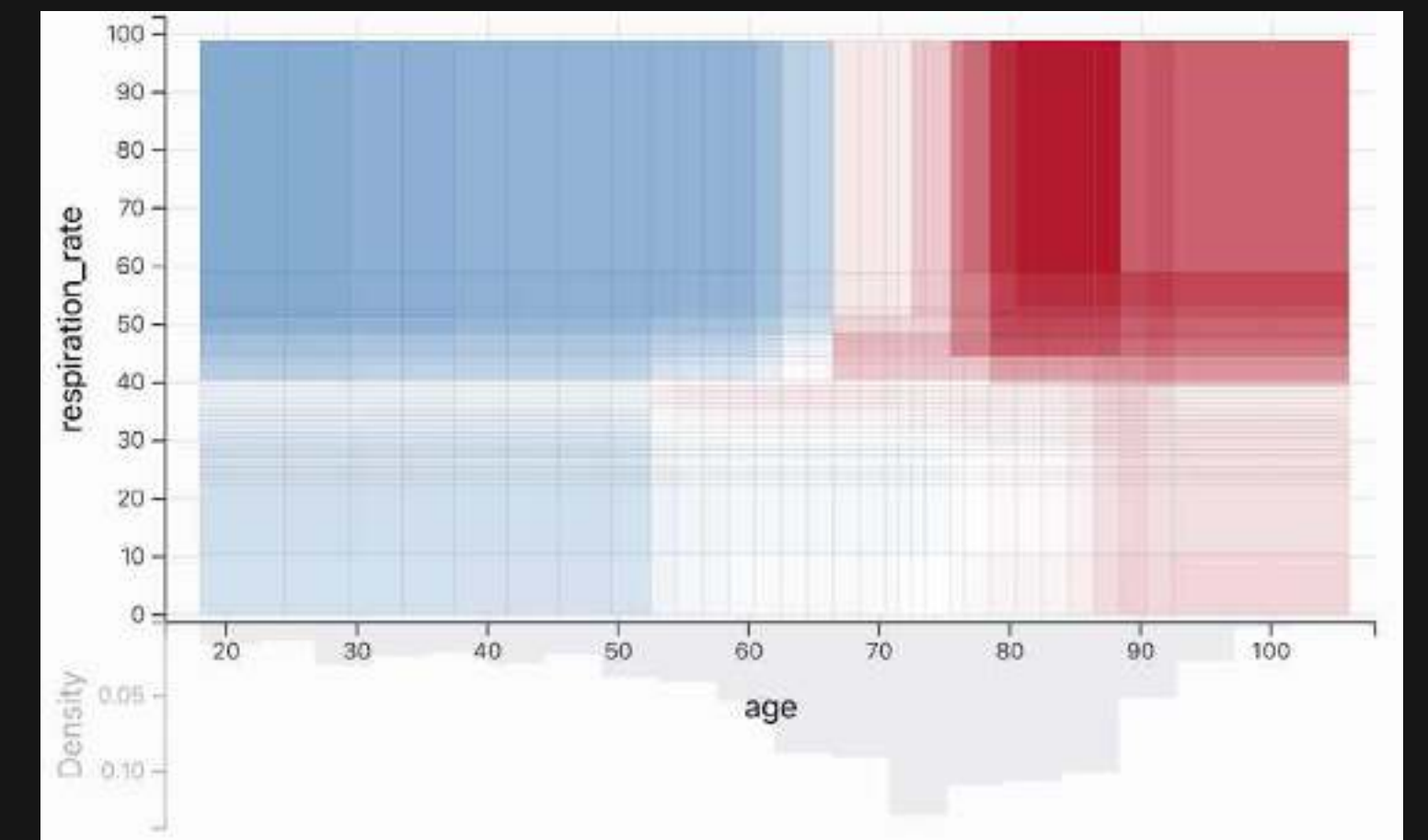
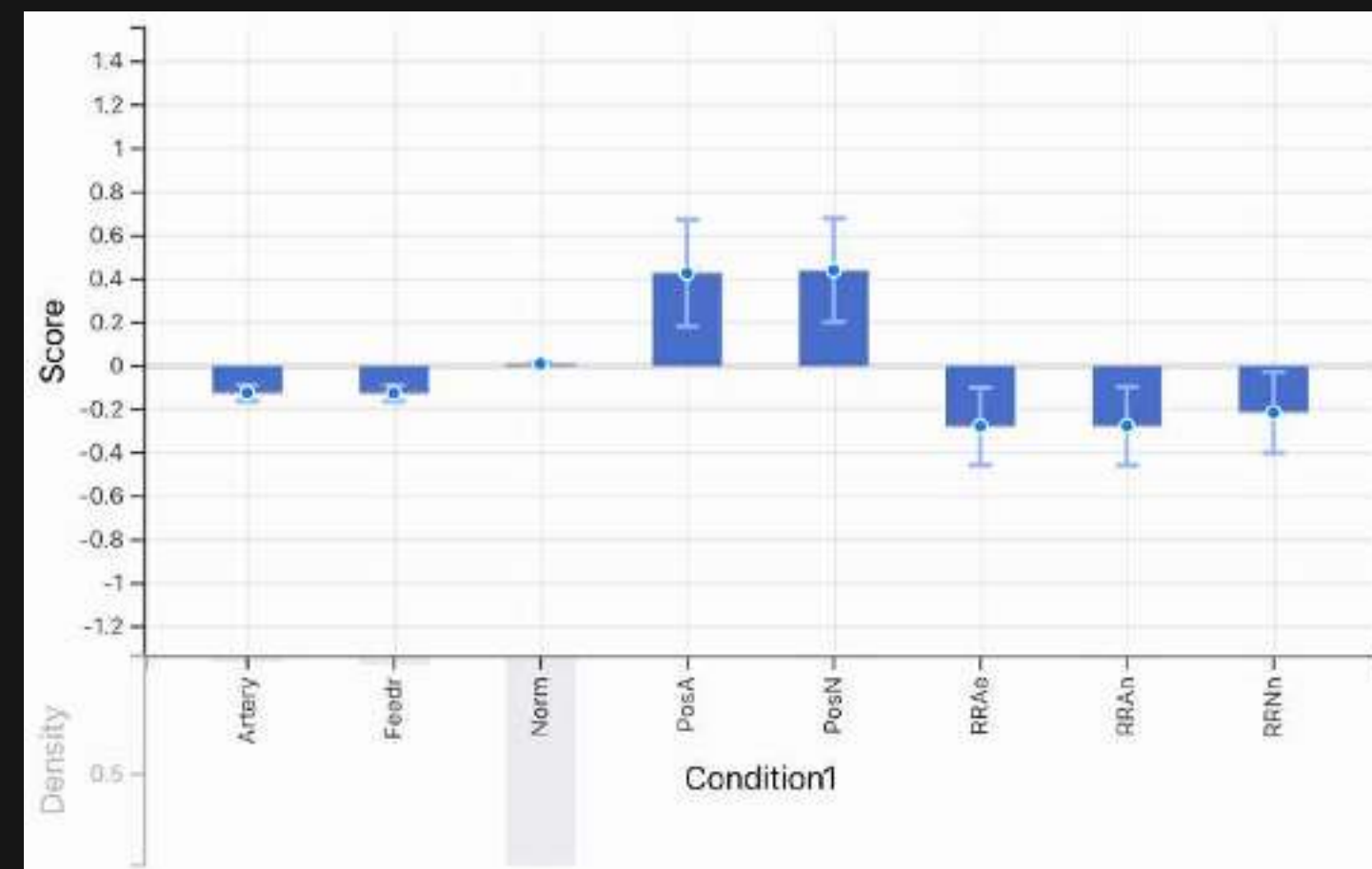
# Older = lower pneumonia risk?



# Explainable Boosting Machine (EBM)

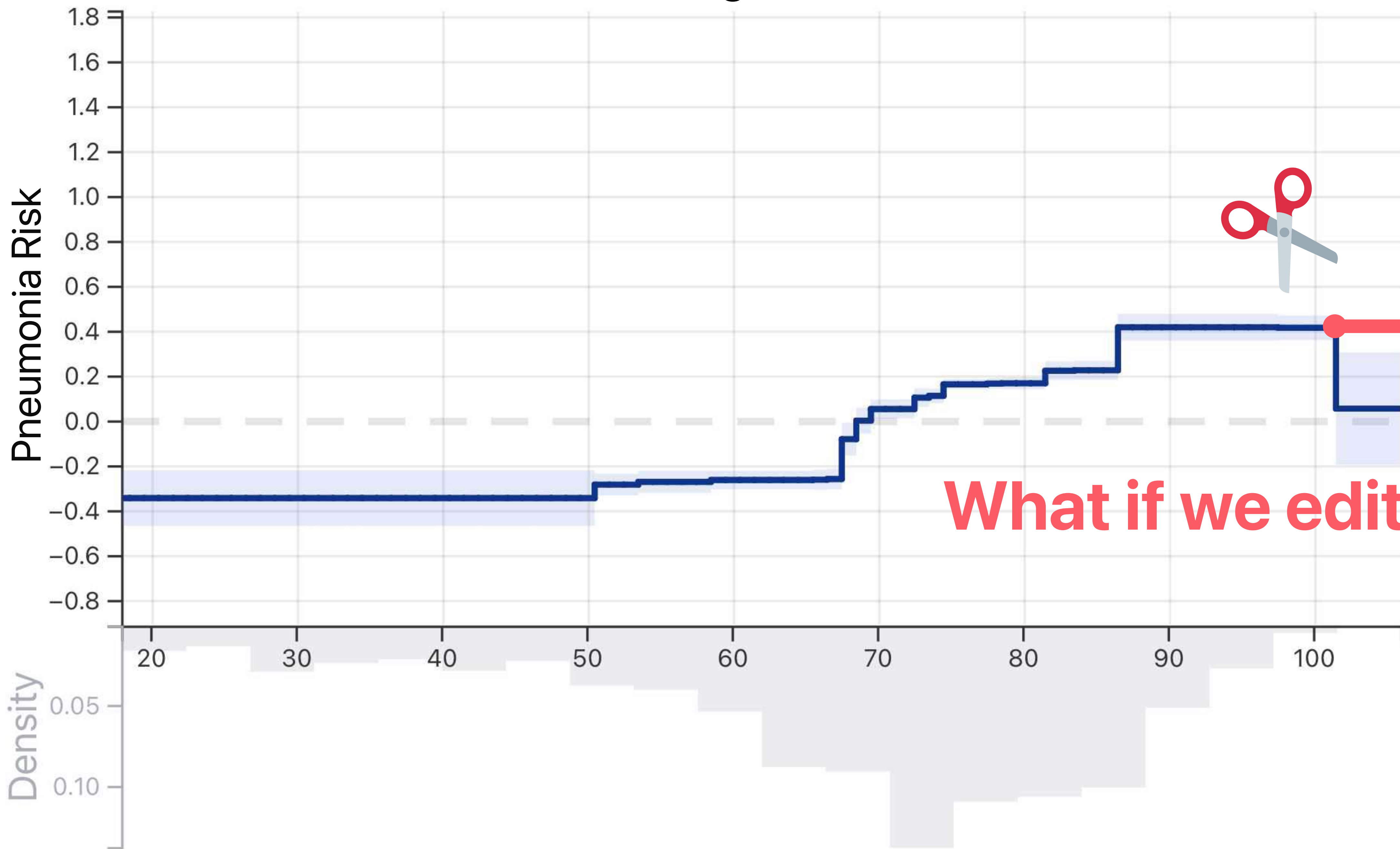
- Generalized additive model (GAM)
- Glass-box model
- Easy-to-understand plots

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)$$





# Age



**What if we edit it?**

# Real Needs for Model Editing

## Fix undesirable behaviors

Higher age should have higher risk

## Remedy mistakes in the dataset

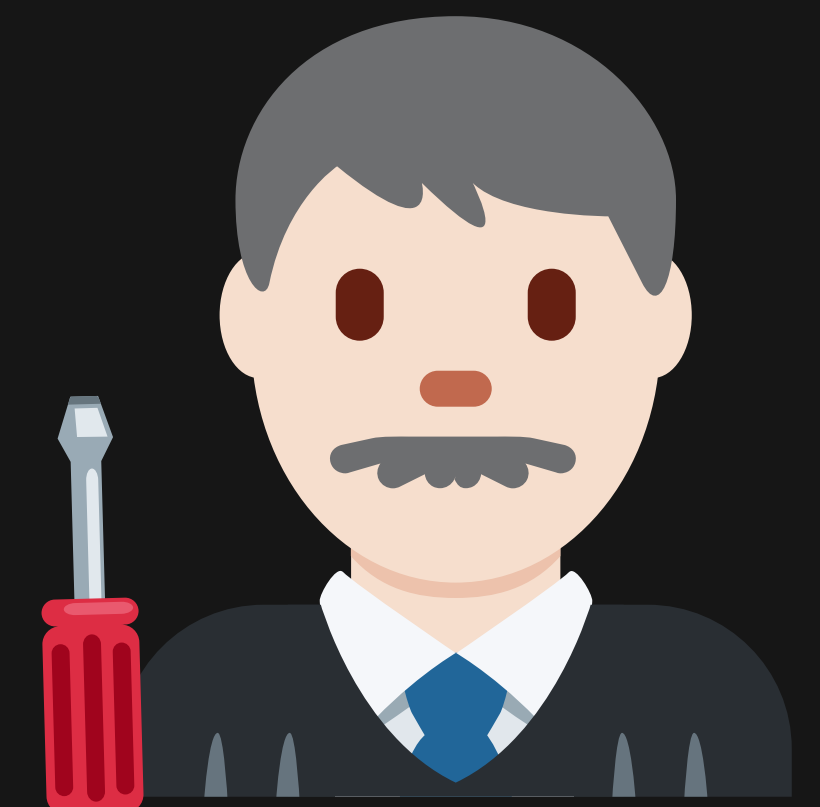
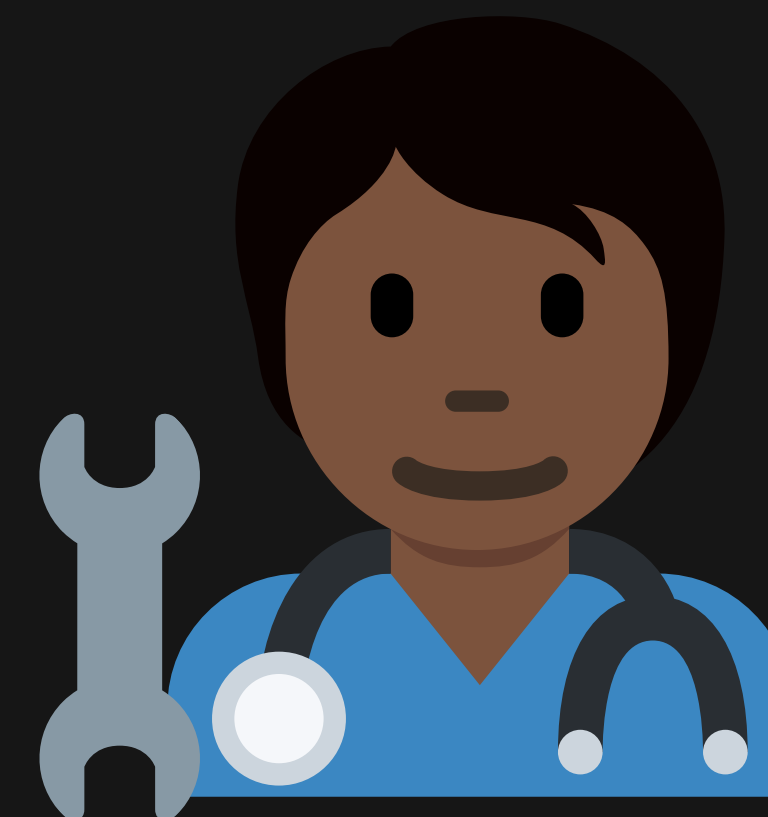
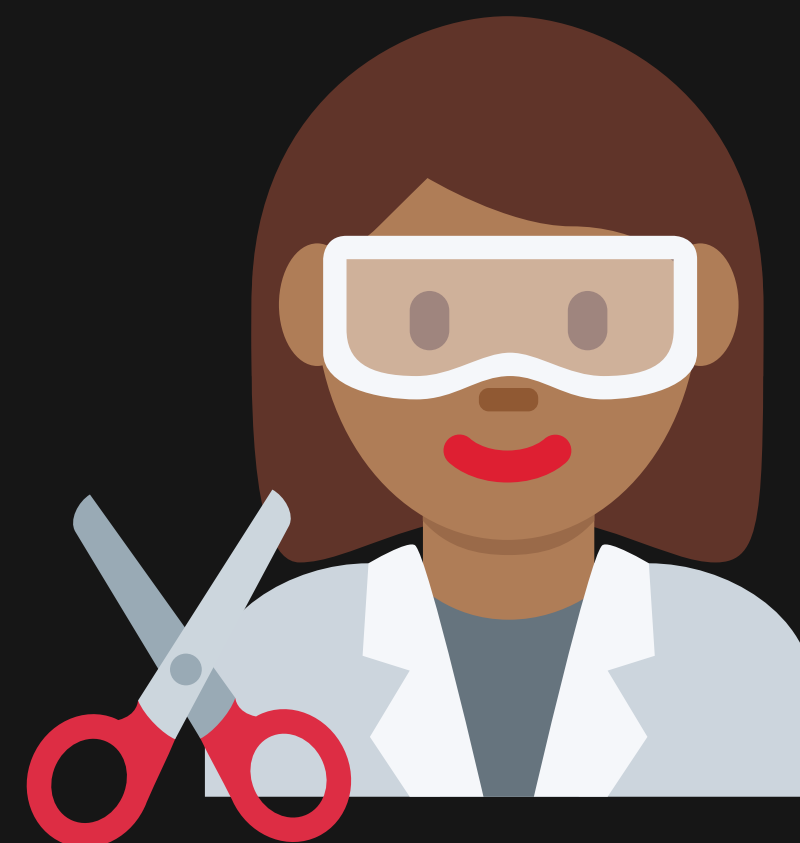
Outliers, missing values, wrong data

## Fairness and Bias

Change effects of protected attributes

## Regulatory Compliance

Enforce monotonicity required by law



# GAM CHANGER

Align ML Model Behaviors with Human Knowledge and Values

# GAM CHANGER

[bit.ly/gam-changer](https://bit.ly/gam-changer)



- Code
- Video
- Paper

# Usefulness **Evaluation**



## **7 Participants**

4 in Finance

2 in Healthcare

1 in Media

Recruited from GitHub Issue Board

## **Loan Application Prediction**

LendingClub dataset

Observational User Study: **Think-aloud** + **Interview**

Finding #1

# Model Editing is a **Common Practice**

"We expect the score to be increasing... model shows something opposite."



Media Company



Car Loan

"You want to make the model easier to explain in adverse action calls."

Enforce Monotonicity

Remove Features

Smooth Out Shape Functions

Fine-tune Hyperparameters

Finding #2

# Fits into Data Scientists' **Workflows**

## Support Computational Notebooks

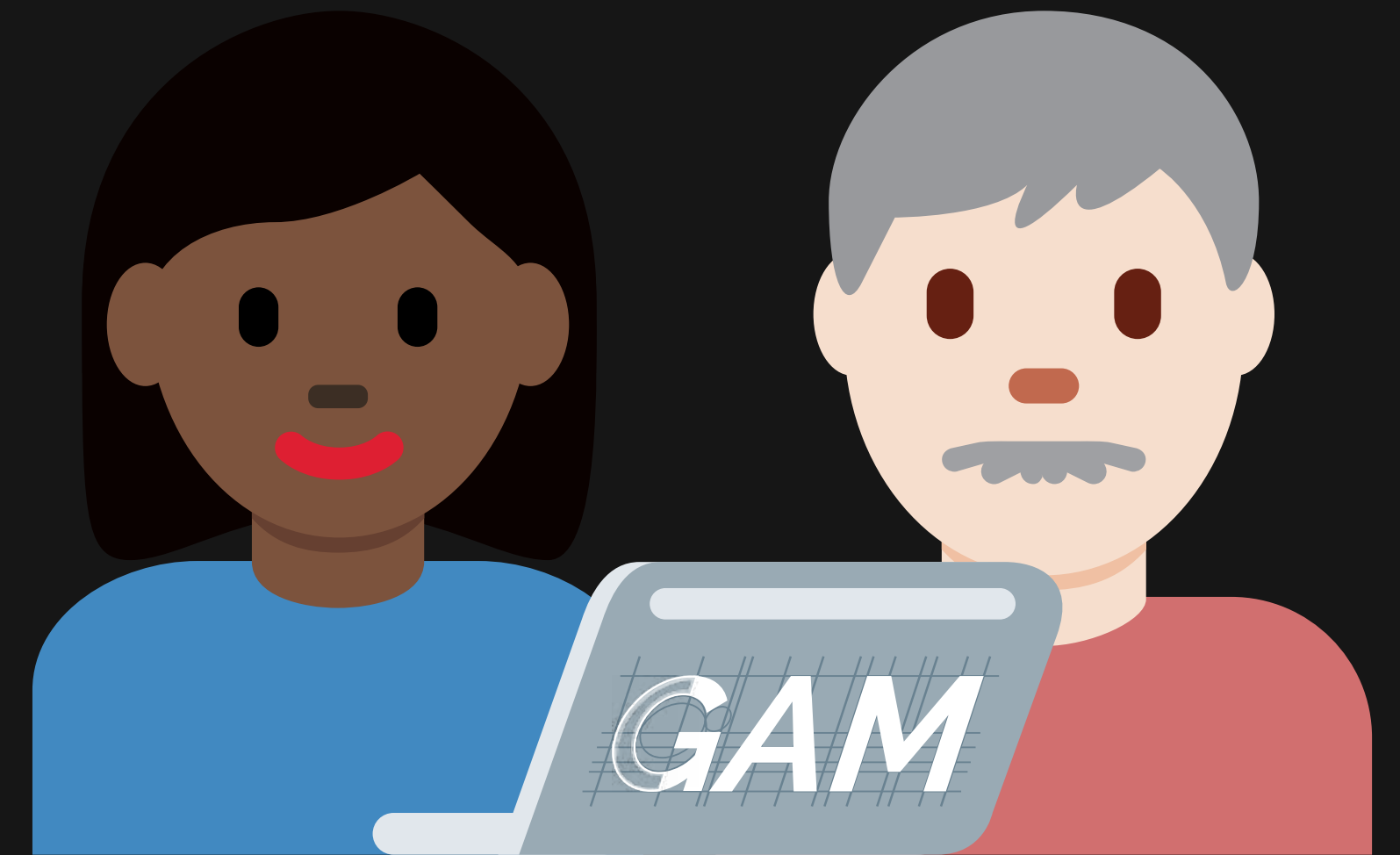
Data scientists appreciate in-notebook editing

## Model Documentation

Documenting model edits helps auditors

## Collaborate with Diverse Stakeholders

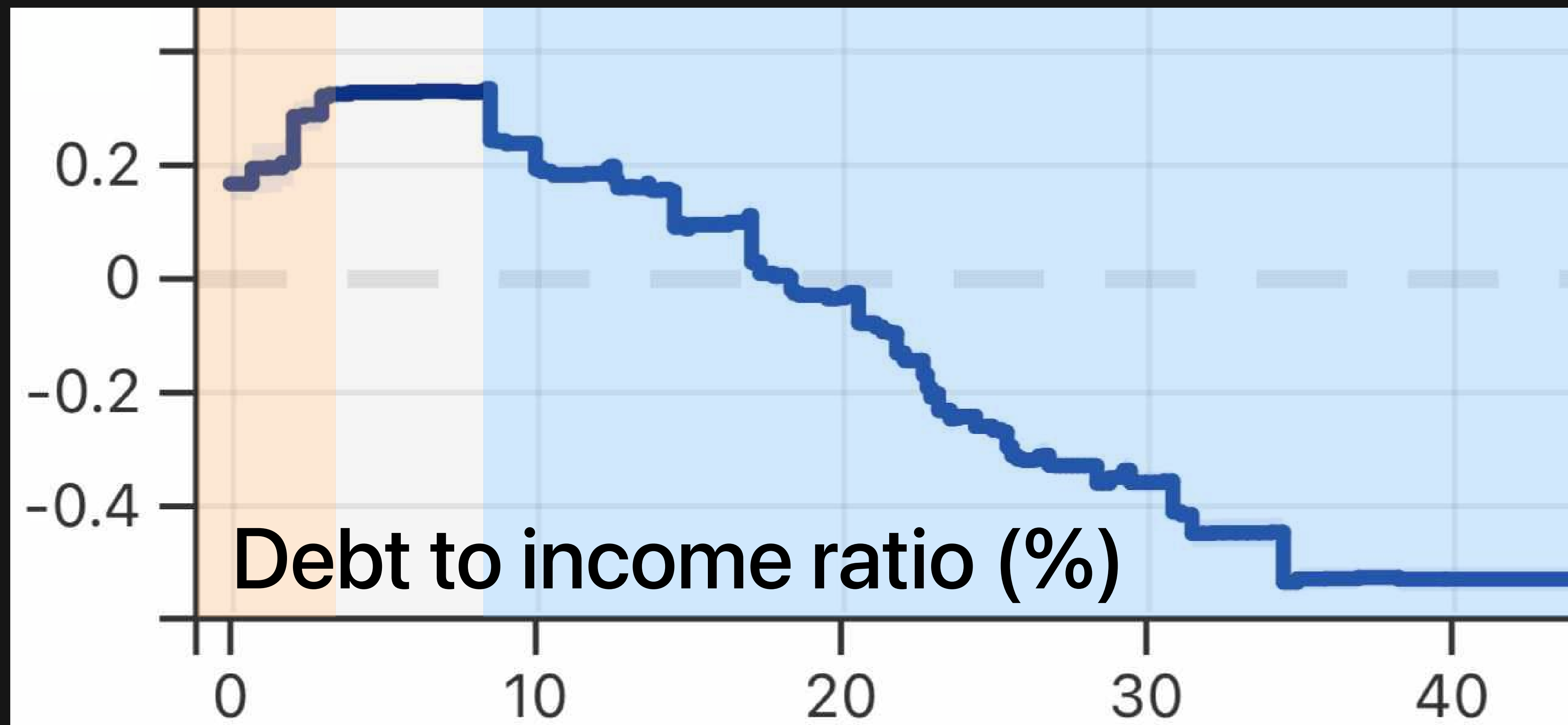
A collaborative platform for model development



Finding #3

# Diverse Ways to Edit a Model

Loan Approval Prediction





# Guide AI with Human Values

**HUMANS**

Guide 



*Domain experts*  
Fix AI errors by model editing  
GAM CHANGER



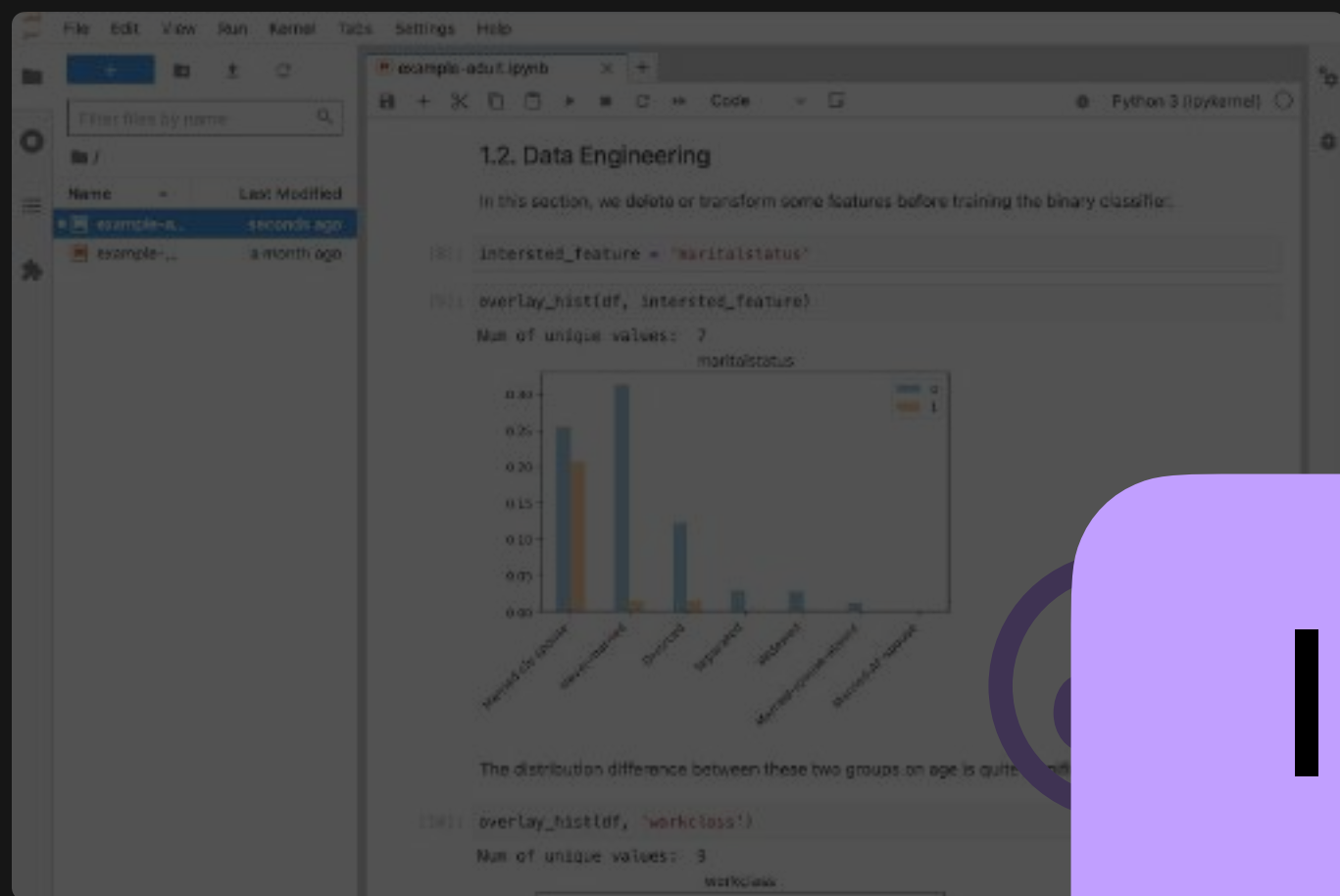
*People impacted by AI*  
Alter unfavorable predictions  
GAM COACH

**AI SYSTEMS**

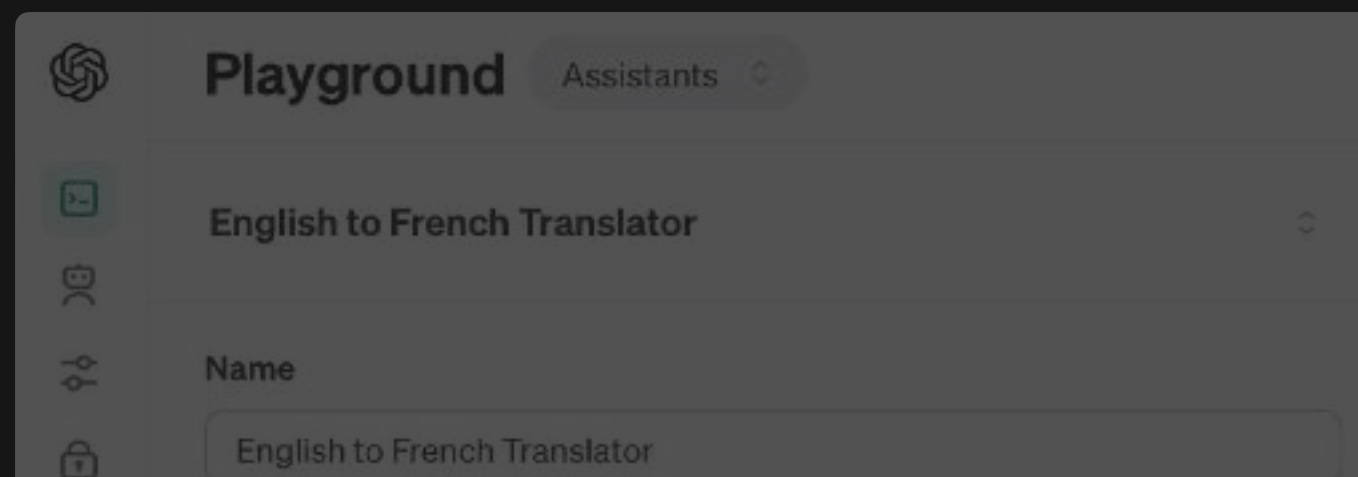


# Democratize Human-Centered AI

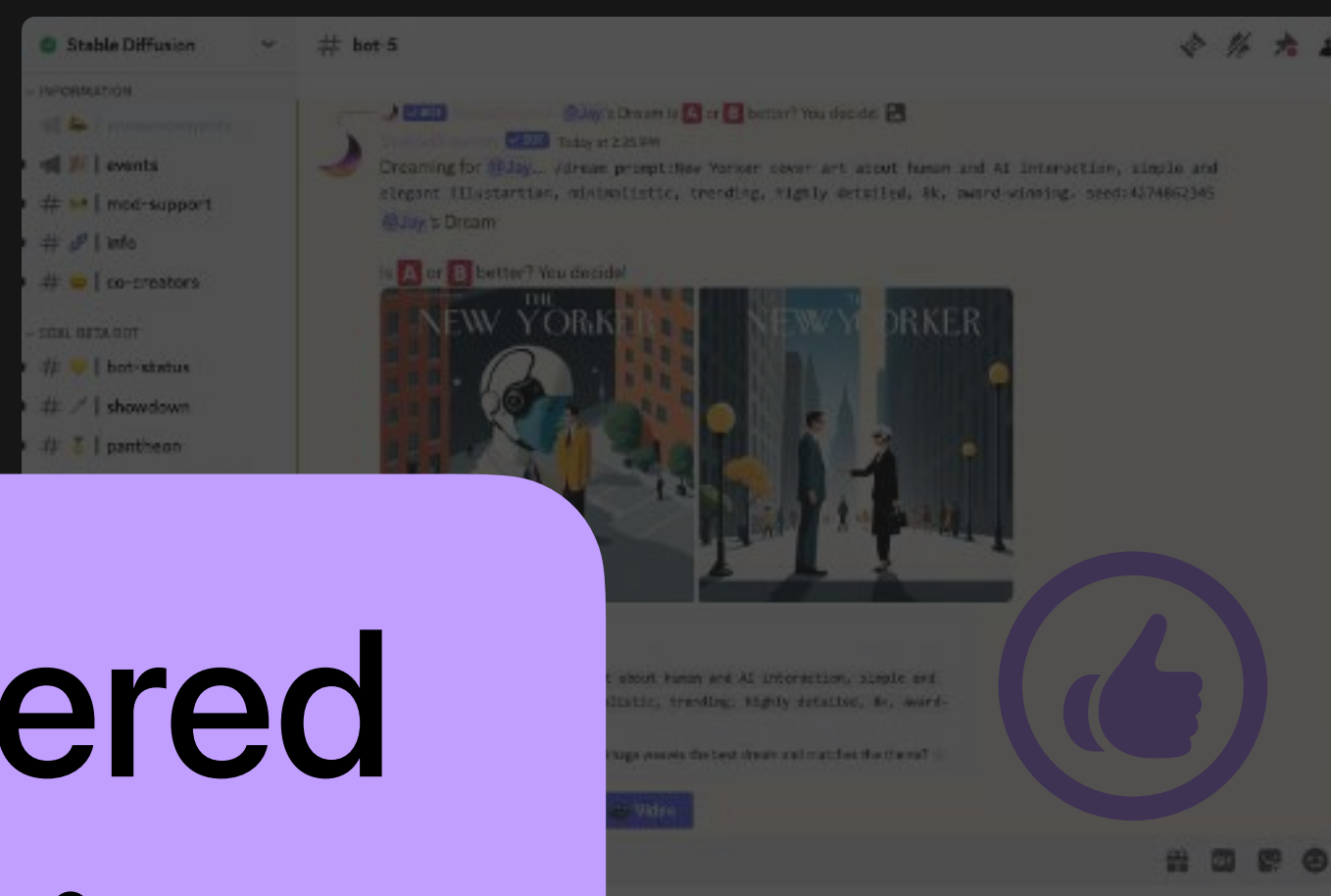
Integrate human-centered AI practices into **existing workflows**



Jupyter Notebook



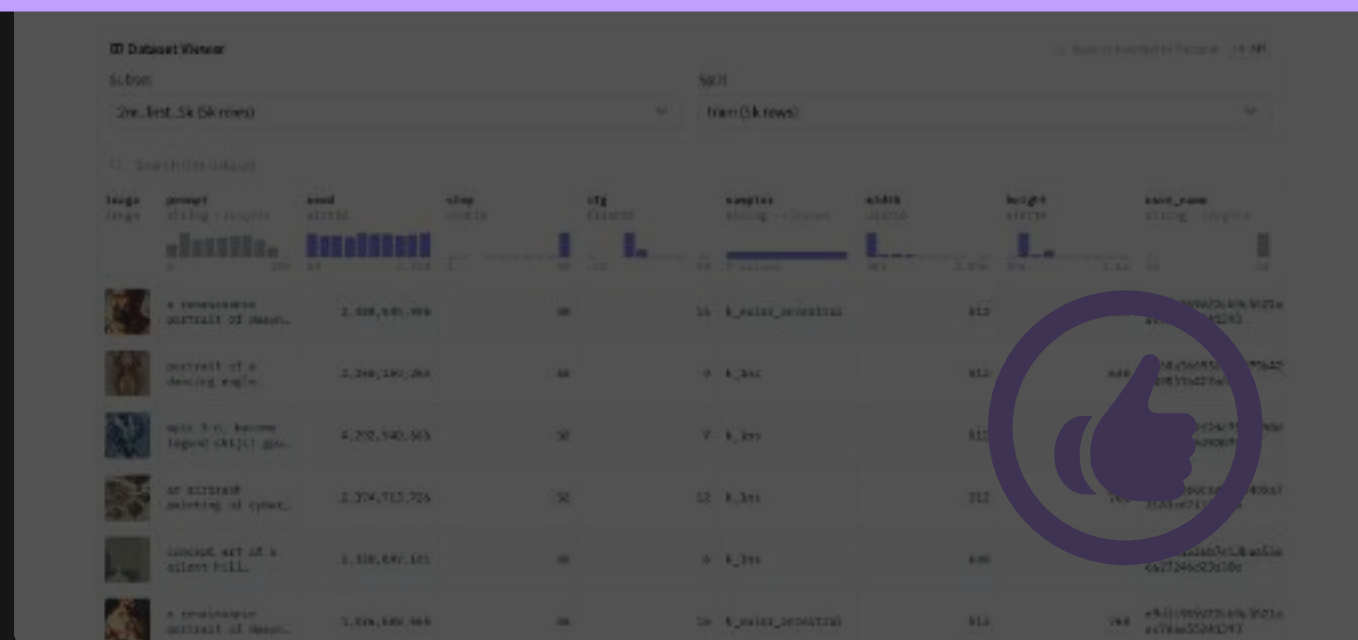
Hugging Face



Stack Overflow

```
40 + Factory function for GameObjects when creating it from an existing cell
41 + through mapping
42 + Since already defined the sticky context that contains this workflow call
43 + Since we need the existing workflow call
44 + Since we need the context network
45 + Workflow A new GameObject object
46
47
48 static method createGameObject()
49 @staticmethod
50 @context_network
51 @context_network
52 @context_network
53 @context_network
54
55 @staticmethod
56 @context_network
57 @context_network
58 @context_network
59 @context_network
60
61 // Base workflow
62 @context_network
63 @context_network
64 @context_network
65
66 // Register the original execution counter node
67 @context_network
68 @context_network
69 @context_network
70
71 // Attach the clone node to the workflow
72 @context_network
73 @context_network
74 @context_network
75
76 // How to append the node to the workflow so we can use the workflow
77 @context_network
78 @context_network
79 @context_network
80
81 // Add a workflow
82 @context_network
83 @context_network
84 @context_network
85
```

VS Code





JavaScript library to explain any ML models with SHAP

- 1. KernelSHAP
- 2. WebGL
- 3. Web Workers

[bit.ly/webshap](http://bit.ly/webshap)

# WebSHAP Explaining Any Machine Learning Models in Your Browser!

**Input Data** Loan applicant #092 info

**Continuous Features**

- Credit History Length: 26
- Credit Utilization: 47
- Debt to Income Ratio: 10.81
- Annual Income: 150003
- Number of Open Accounts: 18
- Loan Amount: 10000
- Number of Accounts: 28
- Revolving Balance: 12070
- FICO Score: 712

**Categorical Features**

- Payment Period: 36 months
- Employment Length: 9 years
- Home Ownership: Mortgage
- Verification Status: Source Verified
- Number of Derogatory Records: 1 time
- Application Type: Individual Application
- Number of Bankruptcies: 0 time

**ML Model** XGBoost

**Model Output** Likelihood of timely repayment: 92.16% **Approval**

**WebSHAP** Features: Privacy, Ubiquity, Interactivity

**Model Explanation** Each feature's contribution to model's prediction

**Top 10 Important Features and Their SHAP Values**

Feature	SHAP Value
FICO Score	0.0325
Number of Open Accounts	-0.0246
Annual Income	0.0239
Home Ownership (Mortgage=T)	0.0137
Home Ownership (Rent=F)	0.0128
Number of Accounts	0.0113
Verification Status (Source Verified=T)	-0.0105
Loan Amount	0.0049
Number of Derogatory Records (0 tim...)	-0.0042
Employment Length (9 years=T)	0.0031

# MEMEMO

JavaScript Library  
for vector storage  
and search

1. In-browser RAG
2. HNSW
3. IndexedDB

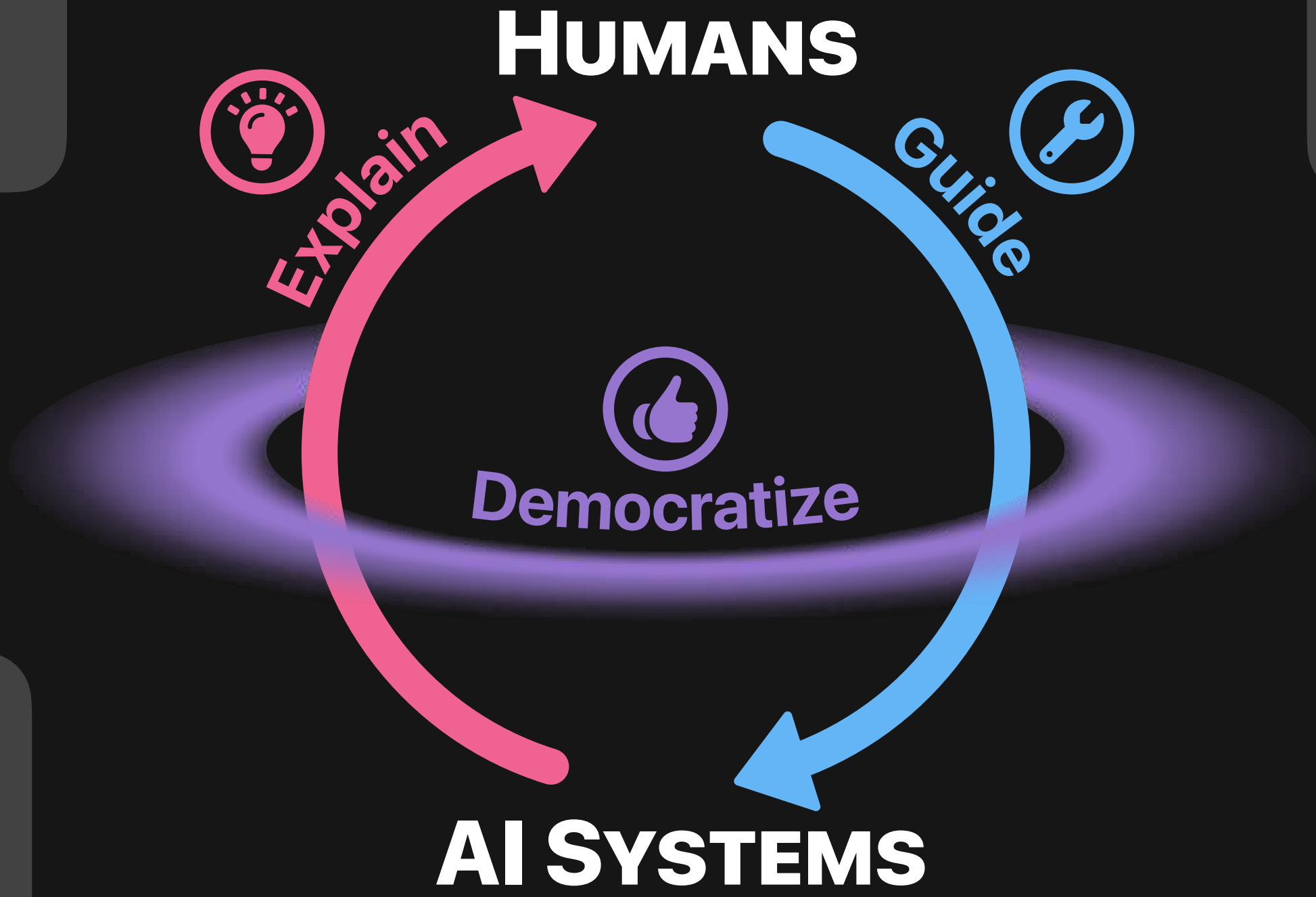
[bit.ly/  
mememojs](https://bit.ly/mememojs)


# MeMemo

JavaScript Library for Vector Search in  
the Browser

**CNN EXPLAINER**  
[bit.ly/cnn-explainer](http://bit.ly/cnn-explainer)

**GAM CHANGER**  
[bit.ly/gam-changer](http://bit.ly/gam-changer)



 **WebSHAP**  
[bit.ly/webshap](http://bit.ly/webshap)

 **Wordflow**  
[bit.ly/wordflow-tool](http://bit.ly/wordflow-tool)

 **MEMEMO**  
[bit.ly/mememojs](http://bit.ly/mememojs)