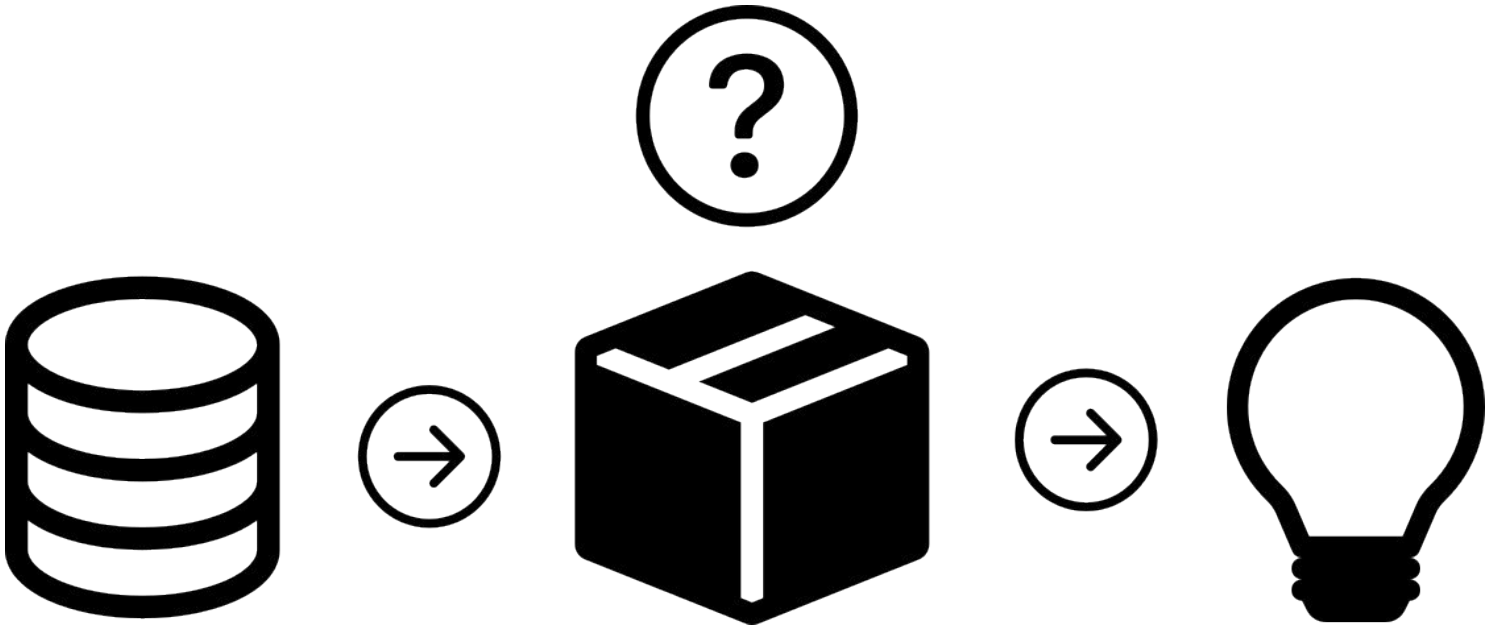
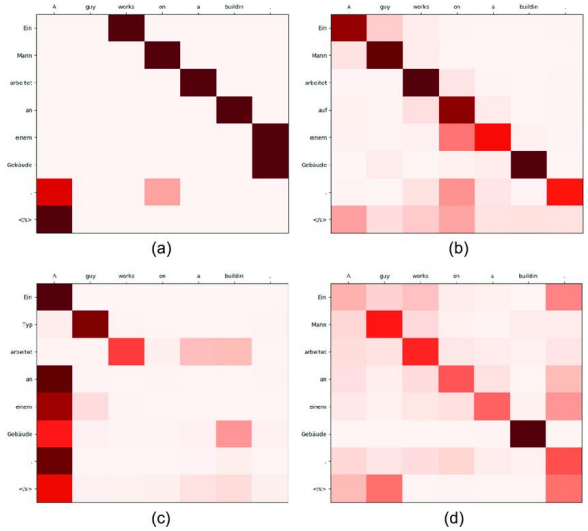
Unveiling the black box: Explainability on Machine Learning Techniques

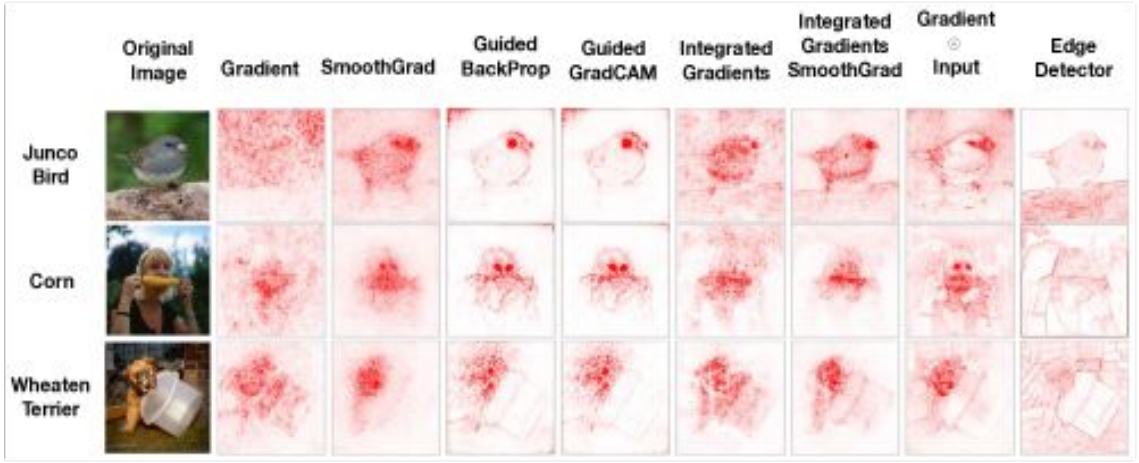
Tiago Araújo
Researcher

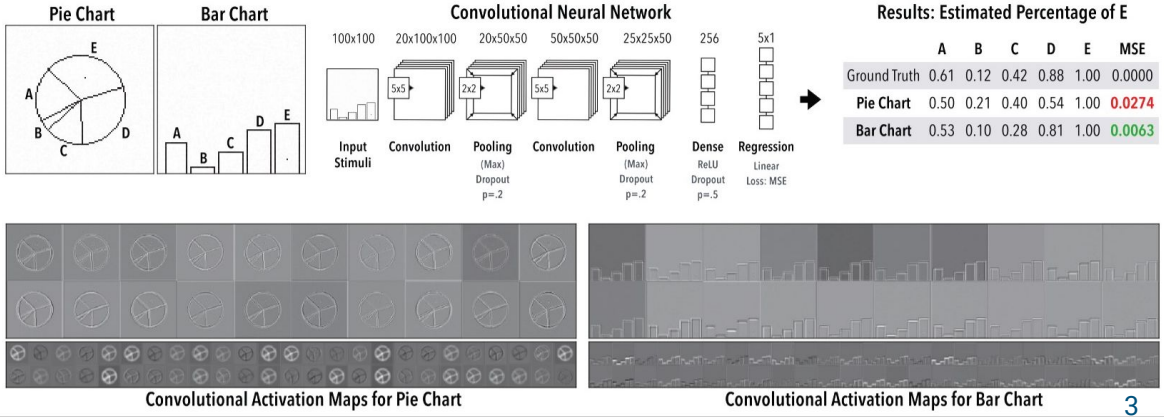
Introduction



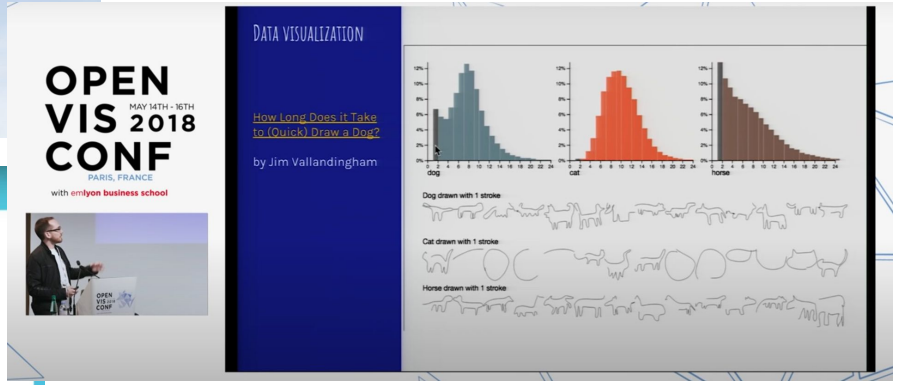
Good Motivators

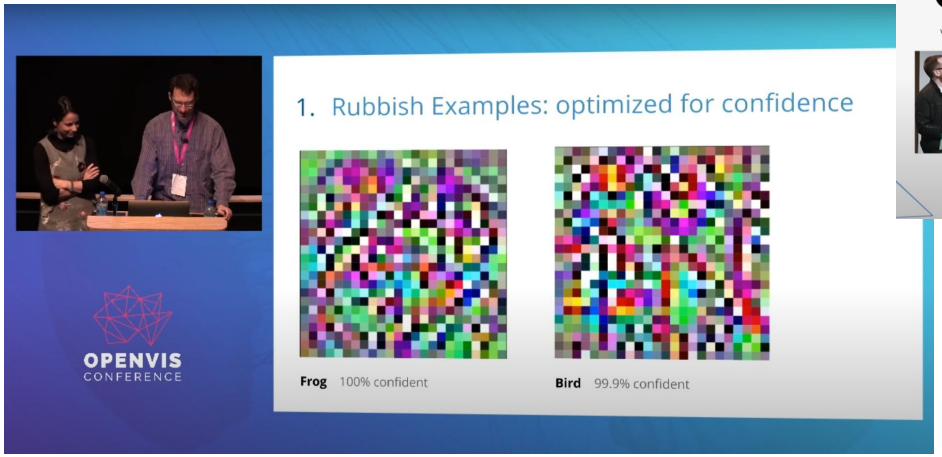




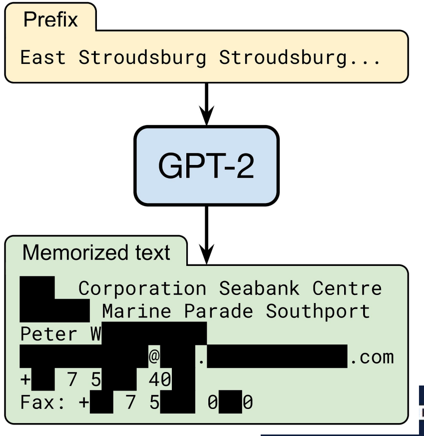


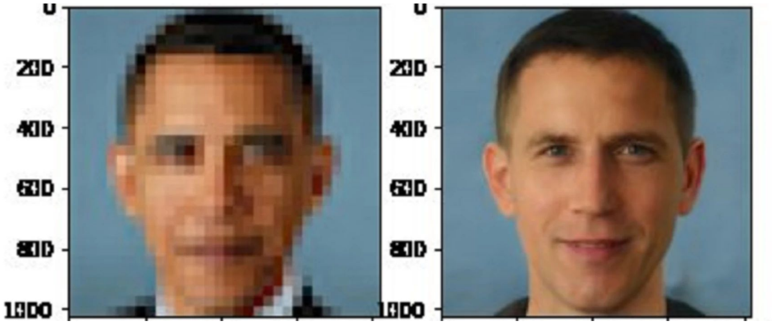
Good Motivators





Urgent motivators



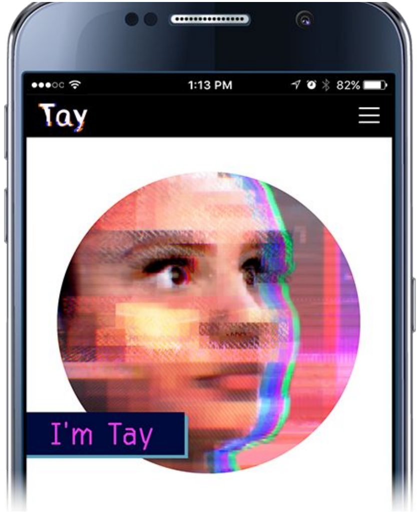


AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

By Karen Hao

January 21, 2019



Data Protection and Privacy

GDPR  LGPD 

Why a Special Issue on Machine Ethics

Publisher: IEEE Cite This PDF

Beyond the Data

Data protection is not enough

We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy.

Because AI needs to be fair & ethical for everyone.

NOVEMBER 3, 2022

Hello. This is Matthew Butterick. On October 17 I told you that I

nature > nature communications > articles > article

Article | Open Access | Published: 23 July 2019

Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye

Nature Communications **10**, Article number: 3069 (2019) | Cite this article

162k Accesses | **291** Citations | **2800** Altmetric | Metrics

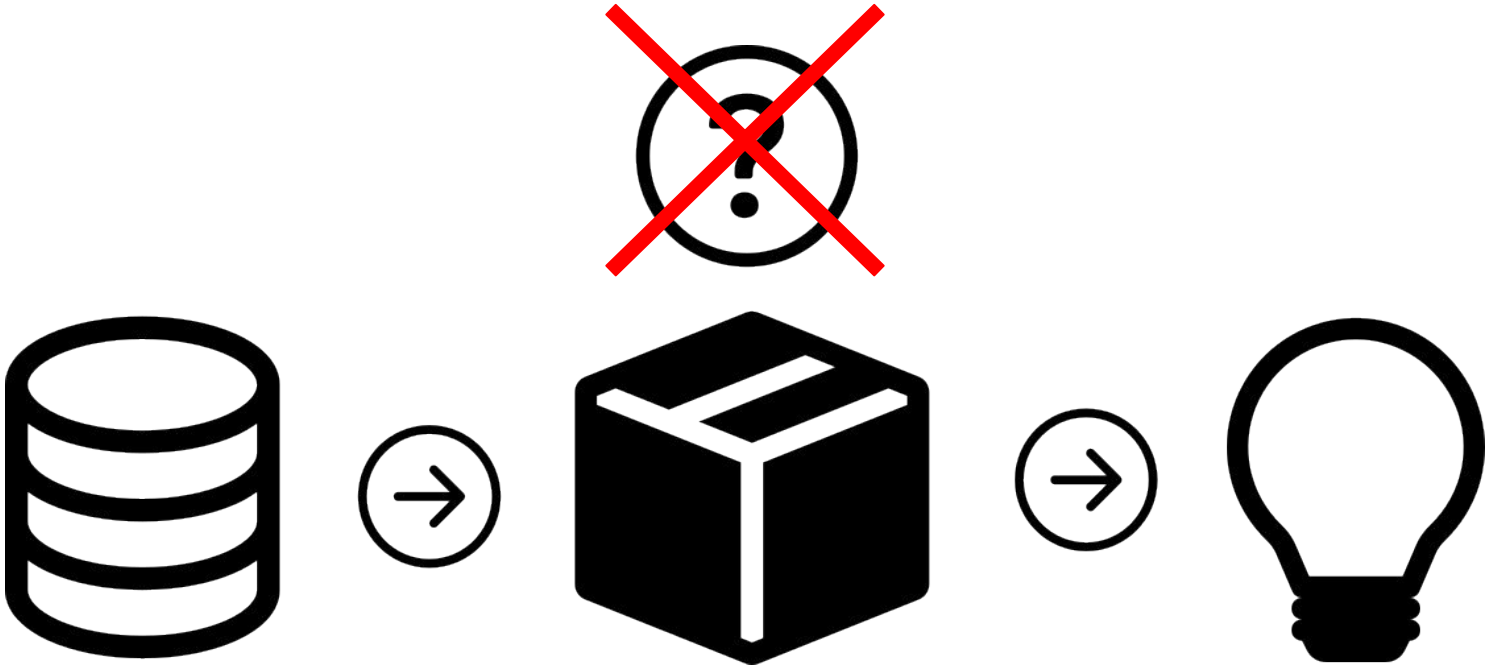
IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

GETTY IMAGES (US), INC.	)	
	)	
Plaintiff,	)	
	)	
v.	)	C.A. No.:
	)	
STABILITY AI, INC.	)	**DEMAND FOR JURY TRIAL**
Defendant.	)	

COMPLAINT

Plaintiff Getty Images (US), Inc. ("Getty Images" or "Plaintiff"), by and through its undersigned attorneys, for its Complaint against Defendant Stability AI, Inc. ("Stability AI" or "Defendant"), hereby alleges as follows:

Explainable and Interpretable AI



Me and ML

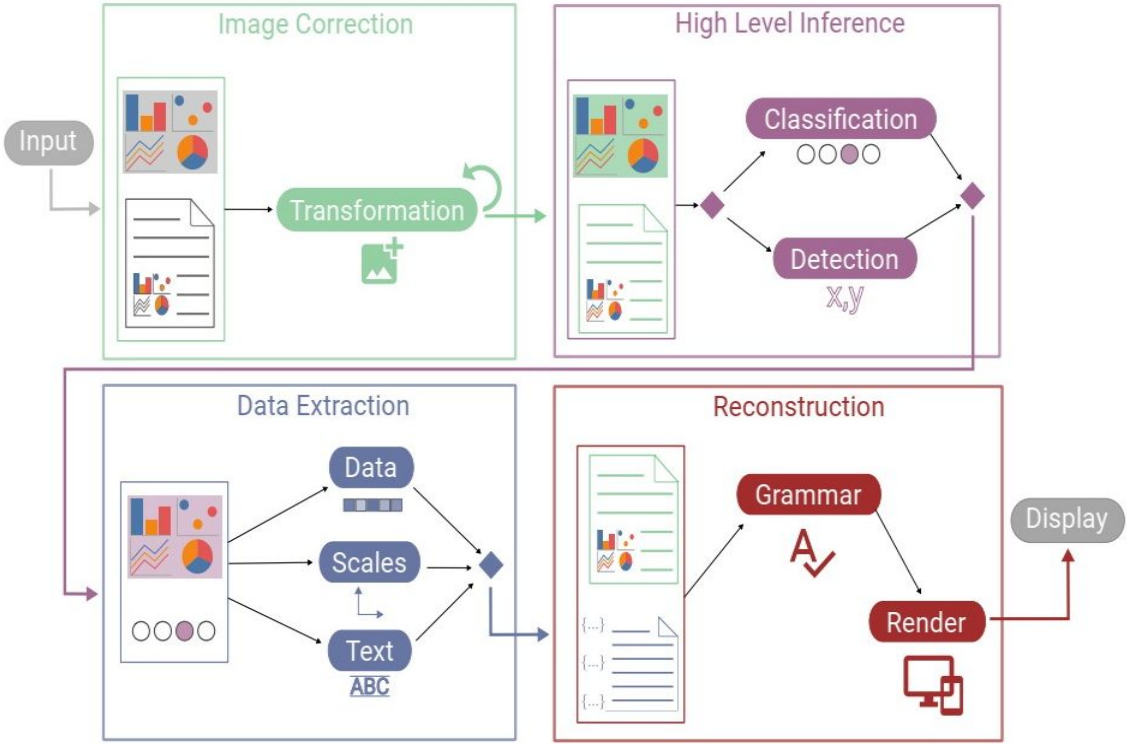
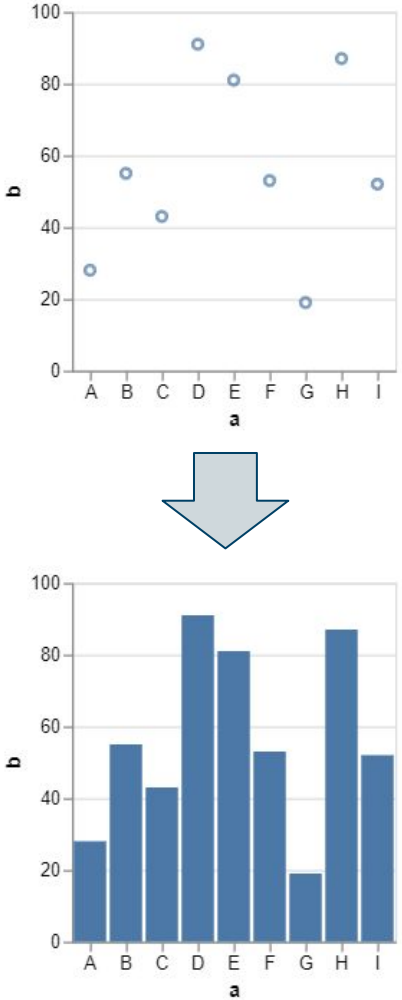
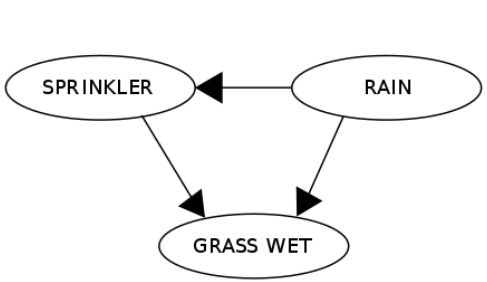


Image Correction
Input
Transformation
High Level Inference
Classification
Detection
x,y
Data Extraction
Data
Scales
Text
ABC
Reconstruction
Grammar
Render
Display

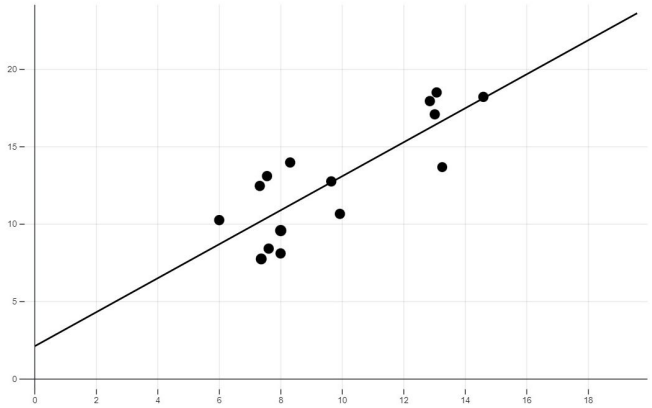

100
80
60
40
20
0
b
A B C D E F G H I
a

History

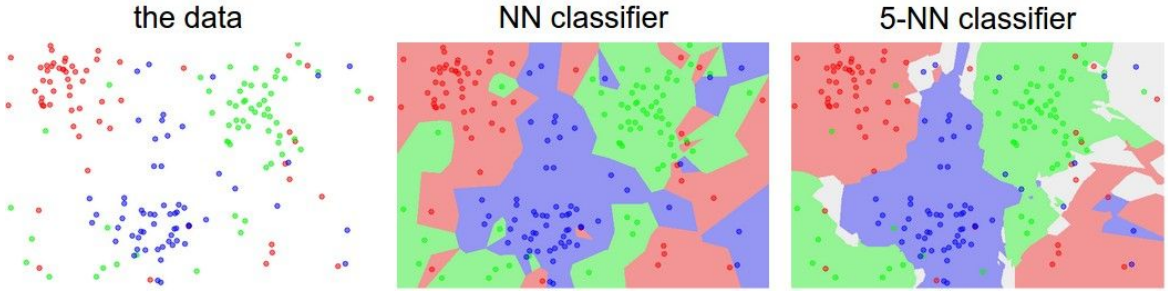
	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



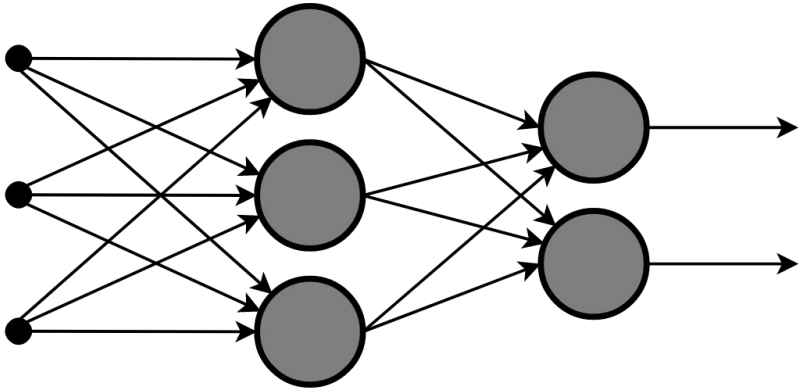
RAIN	
T	F
0.2	0.8

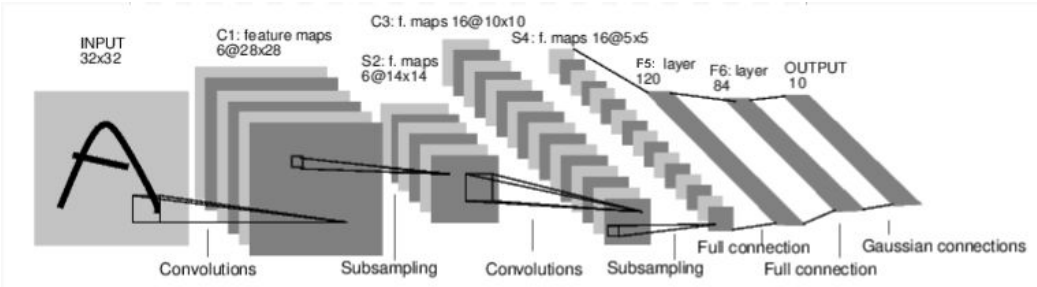


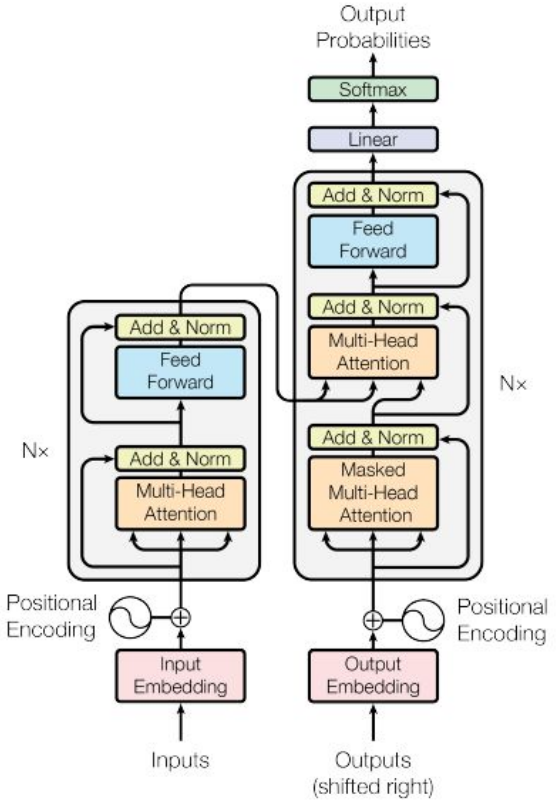
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01



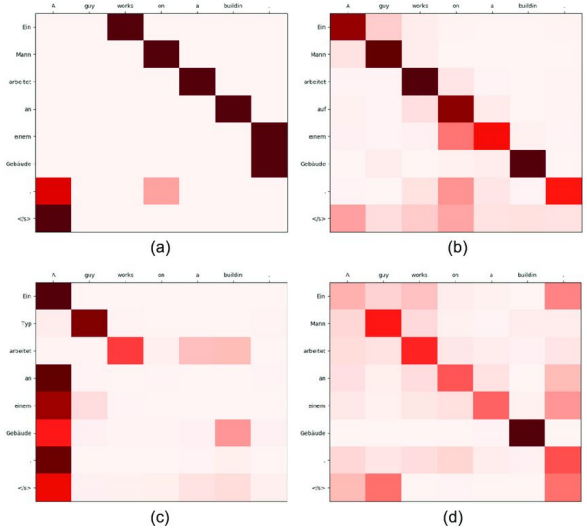
ML Models

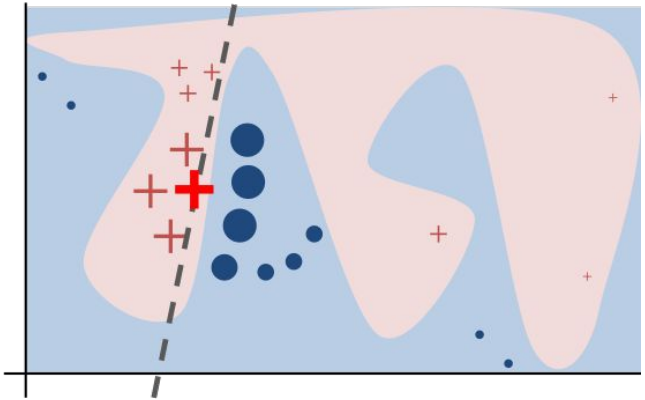


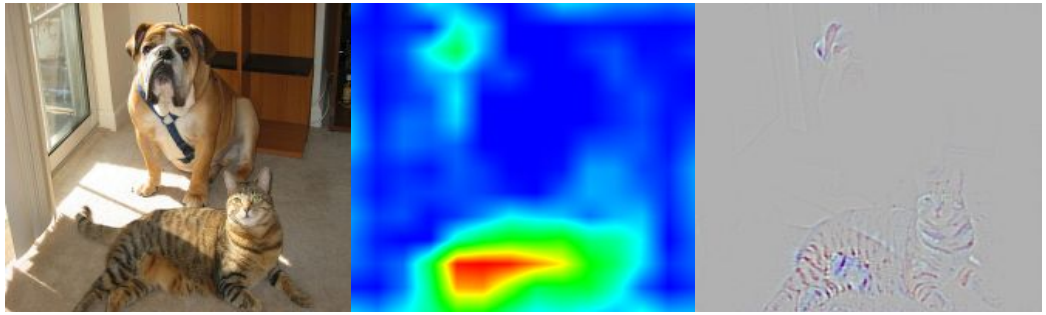

INPUT
32x32
C1: feature maps
6@28x28
S2: f. maps
6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
F5: layer
120
F6: layer
84
OUTPUT
10
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian connections


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

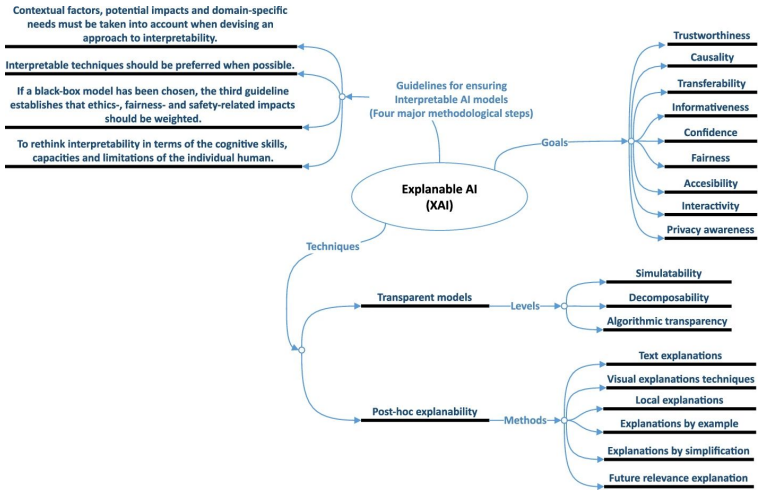
History





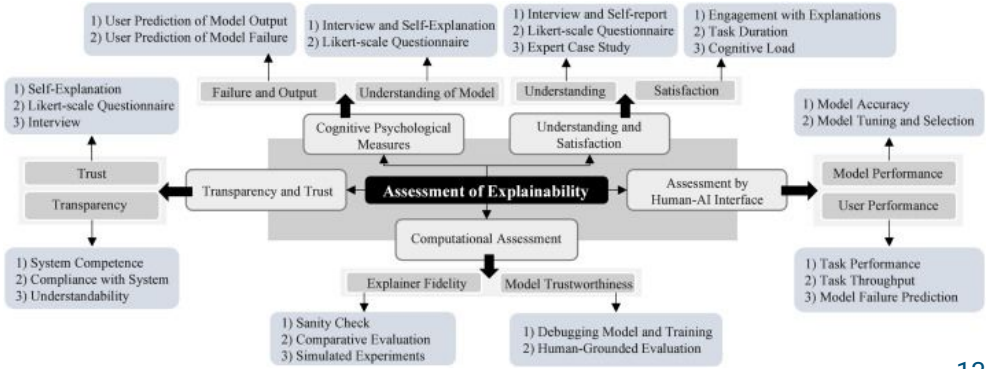


Abstractions



Human-in-the-loop machine learning: a state of the art

Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence



Guidelines

Accuracy

Fairness

nature › news › article

NEWS | 23 December 2020 | Correction 23 December 2020

Prestigious AI meeting takes steps to improve ethics of research

For the first time, the organizers of NeurIPS required speakers to consider the societal impact of their work.

Davide Castelvecchi

Human Control and Decision-making

Bias

Safety & Security

Accountability

nature › humanities and social sciences communications › articles › article

Article | Open Access | Published: 17 June 2020

Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward

Samuele Lo Piano

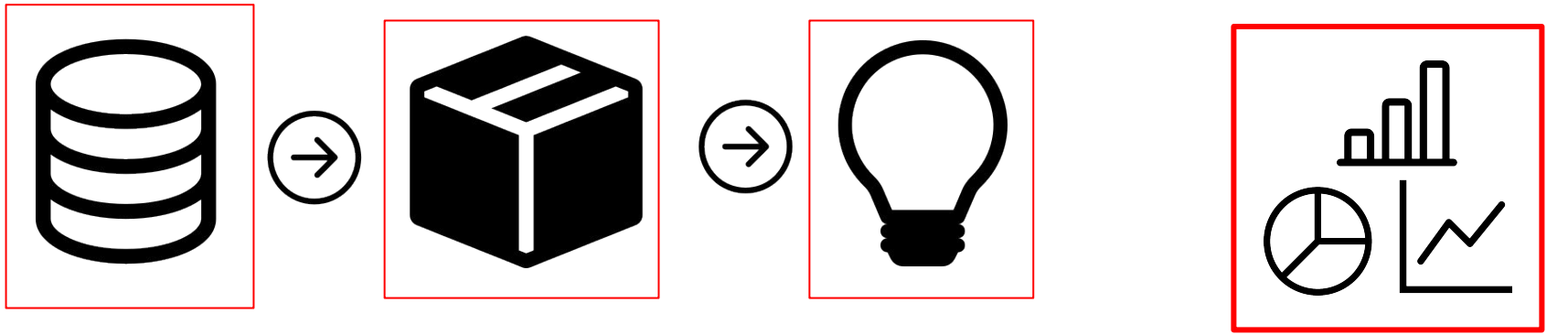
Privacy

Transparency

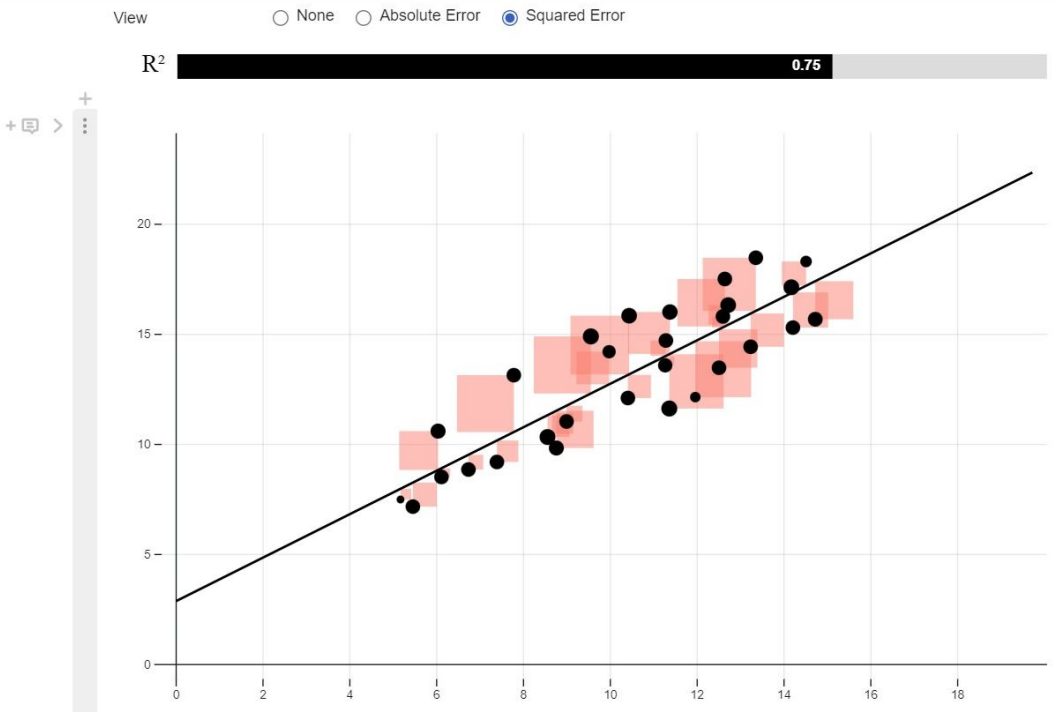
Environmental Impact & Sustainability

Methods

$\Sigma \pi$

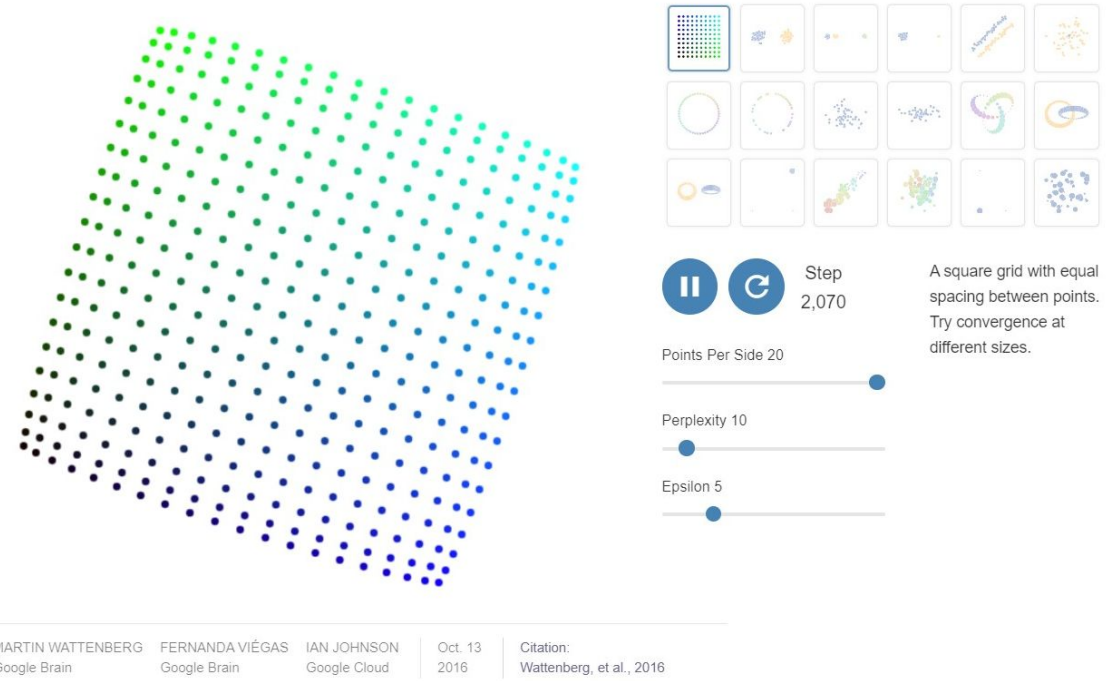


Linear Regression



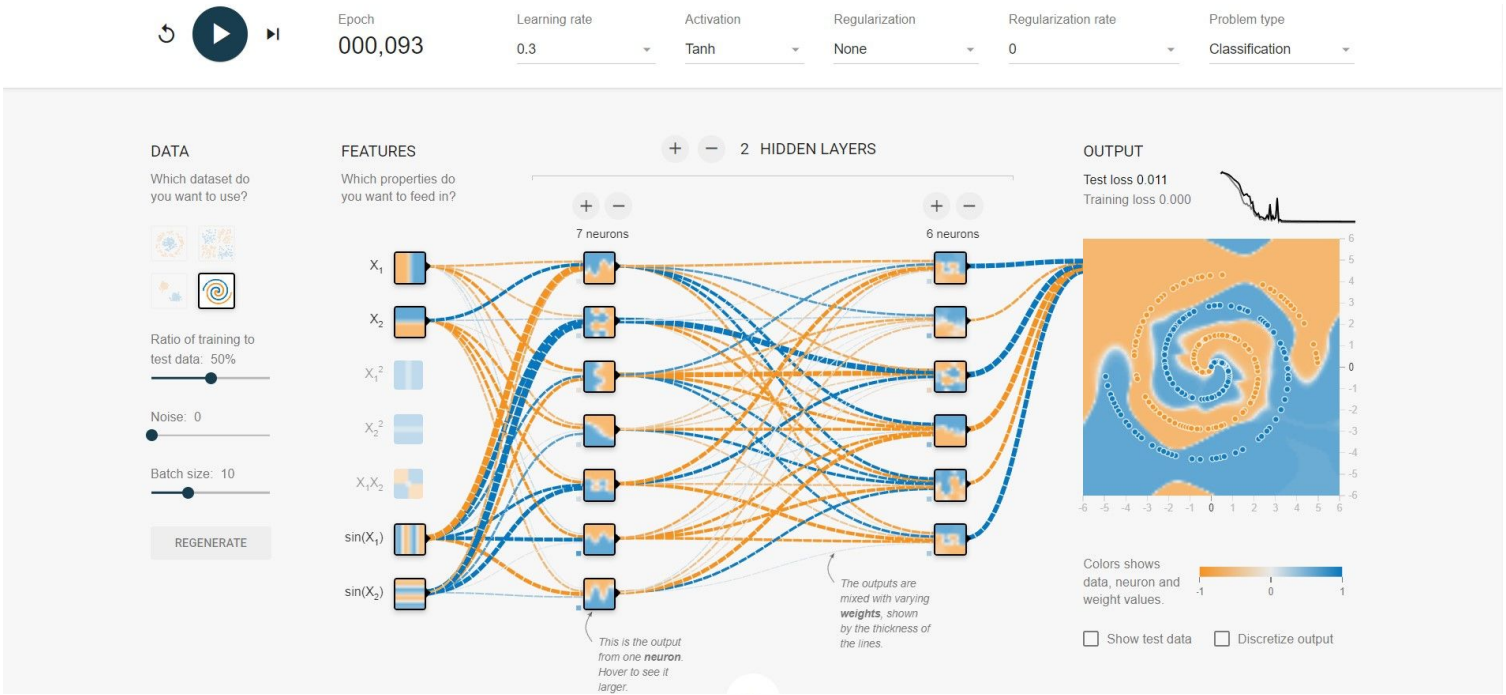
Interactive Visualization of Linear Regression

Data



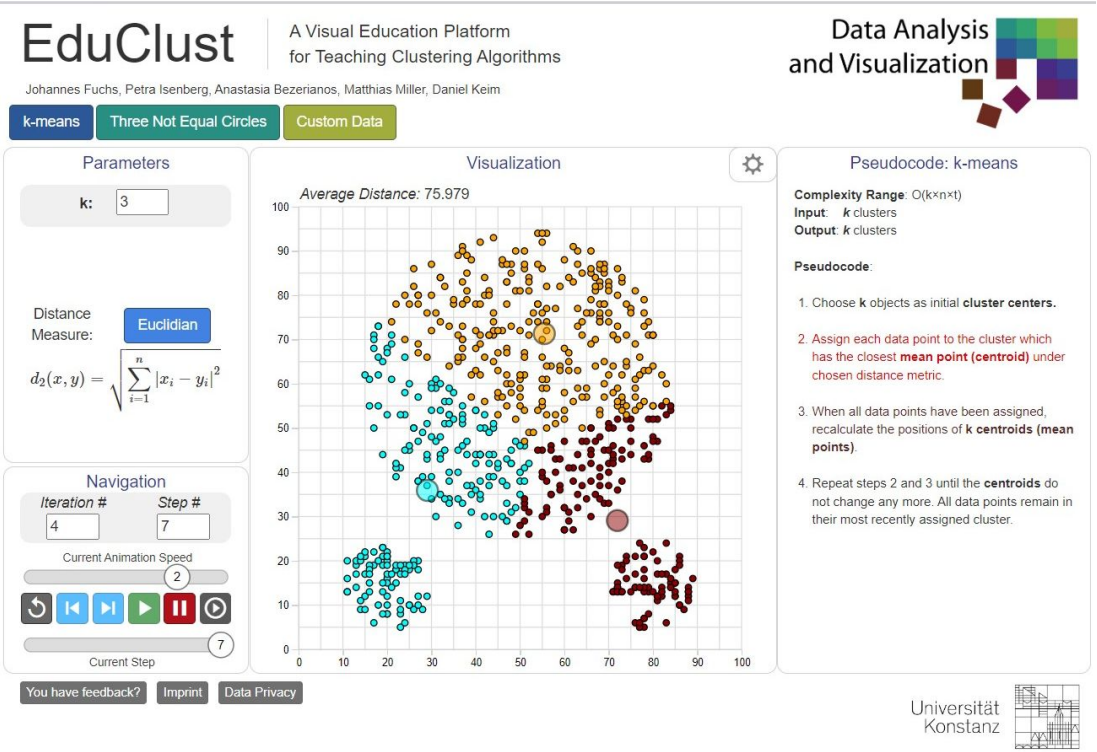
How to Use t-SNE Effectively

Neural Networks



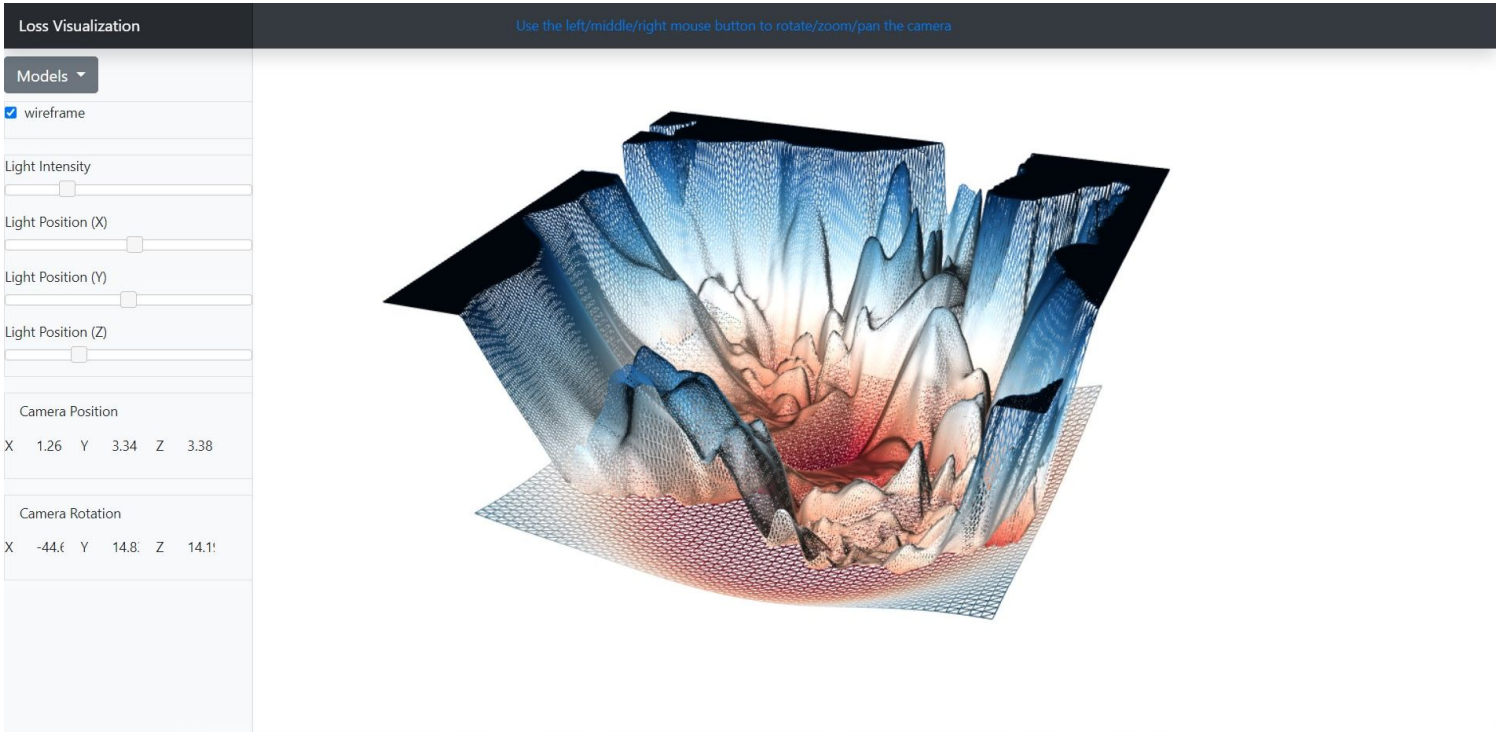
A Neural Network Playground

Clustering



EduClust A Visual Education Platform for Teaching Clustering Algorithms

Loss Function



[Visualizing the Loss Landscape of Neural Nets]

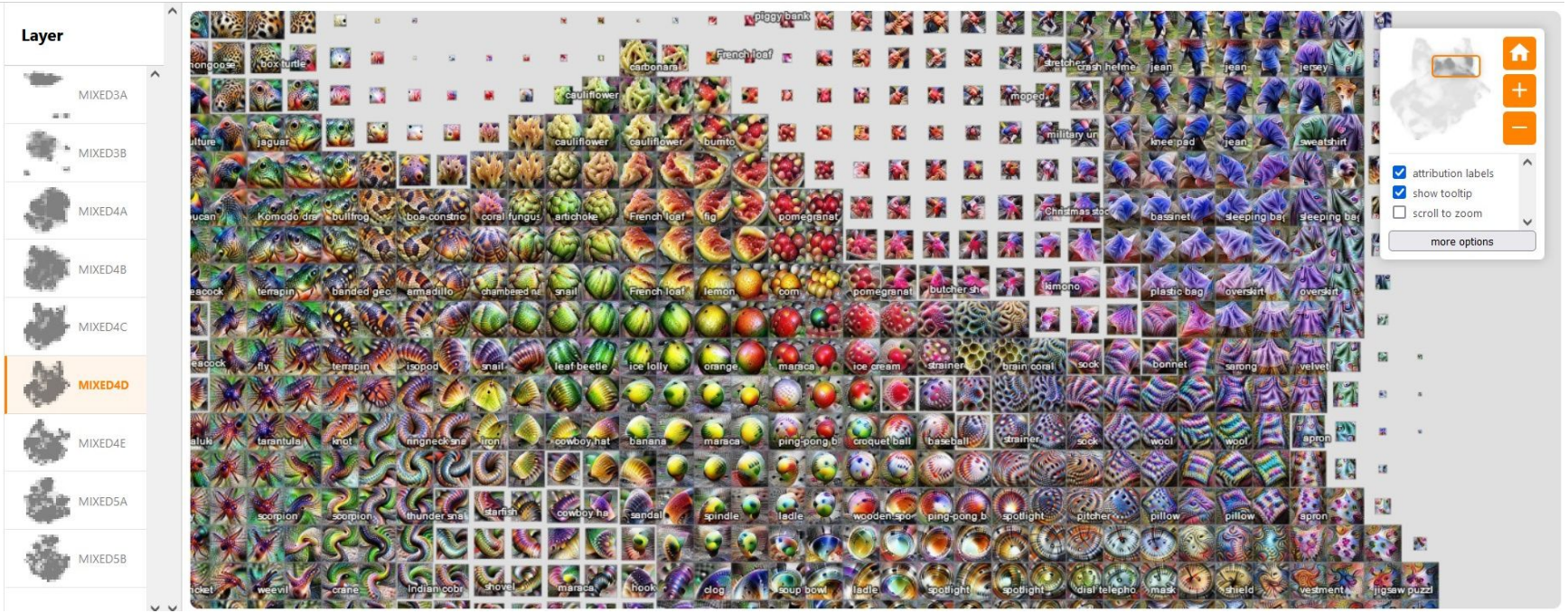
Convolutional Neural Networks

ConvNetJS CIFAR-10 demo

Convolutional Neural Networks

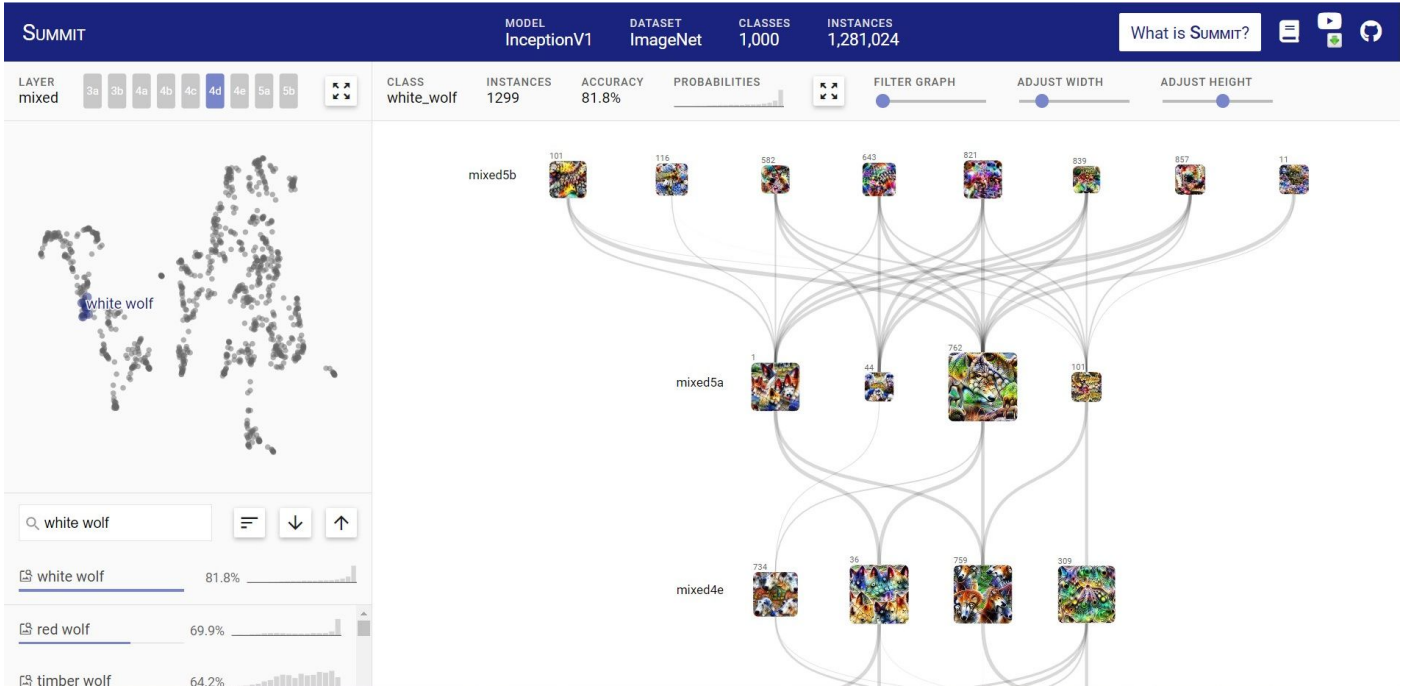
Inceptionism: Going Deeper into Neural Networks

Activation Maps



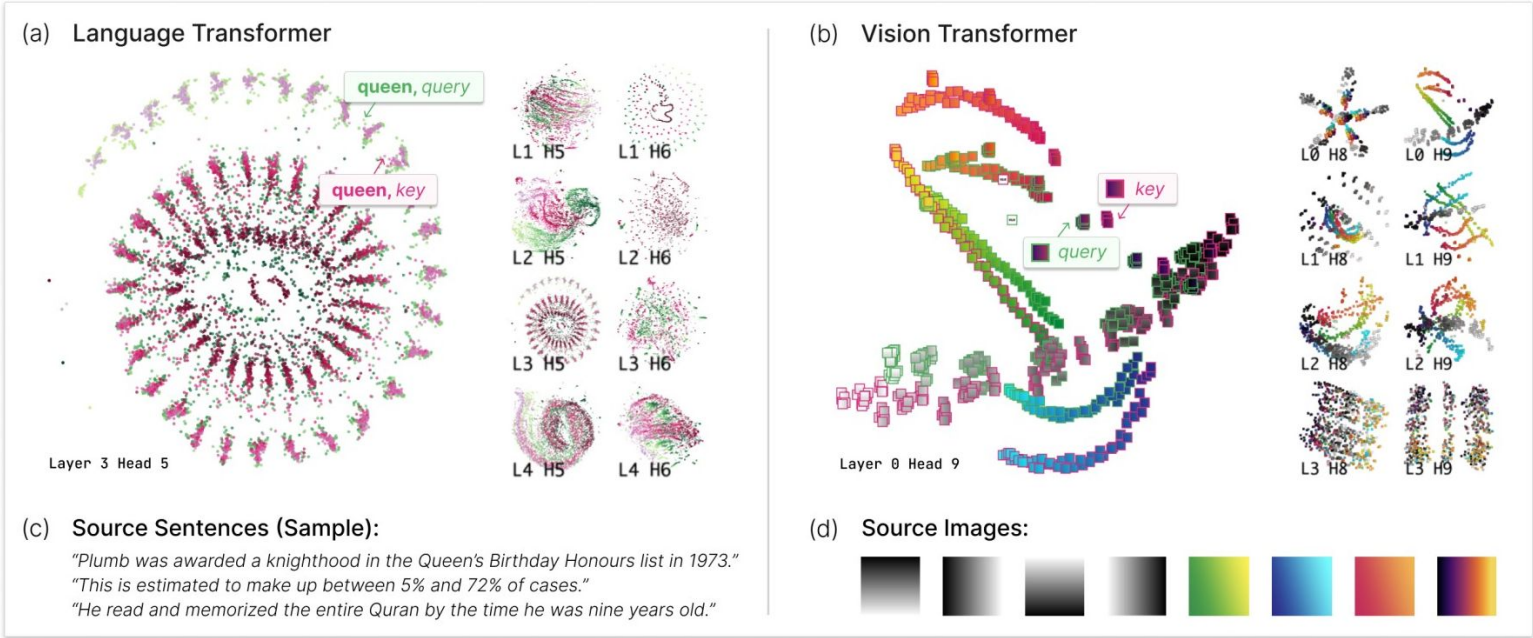
Exploring Neural Networks with Activation Atlases

Activation Maps



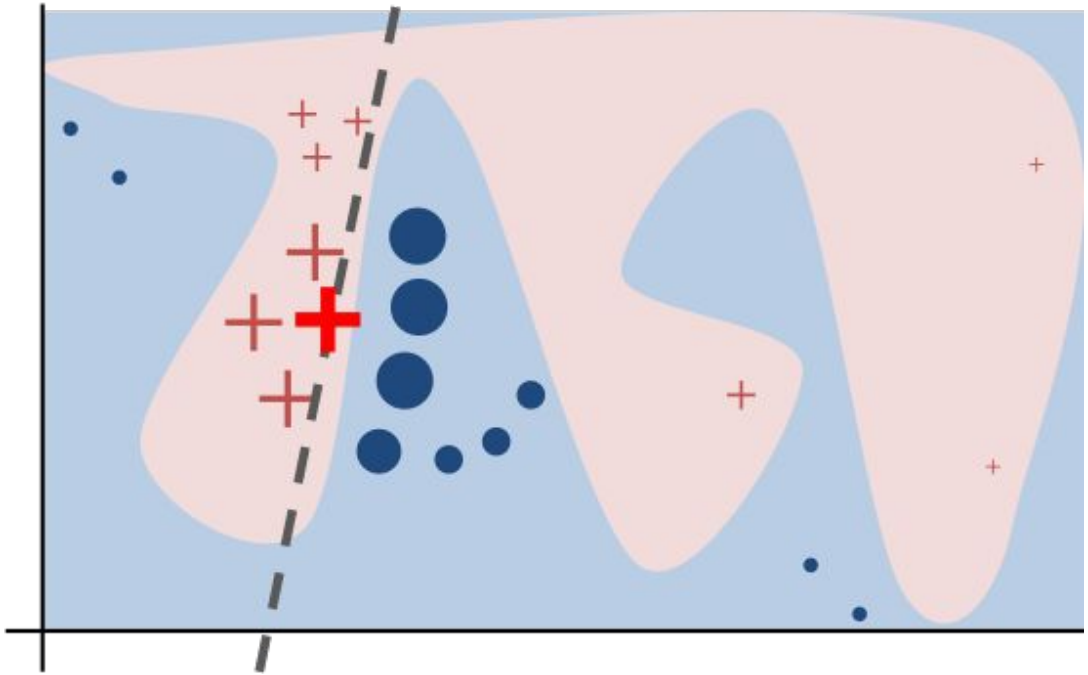
[Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations]

Transformers



AttentionViz: A Global View of Transformer Attention

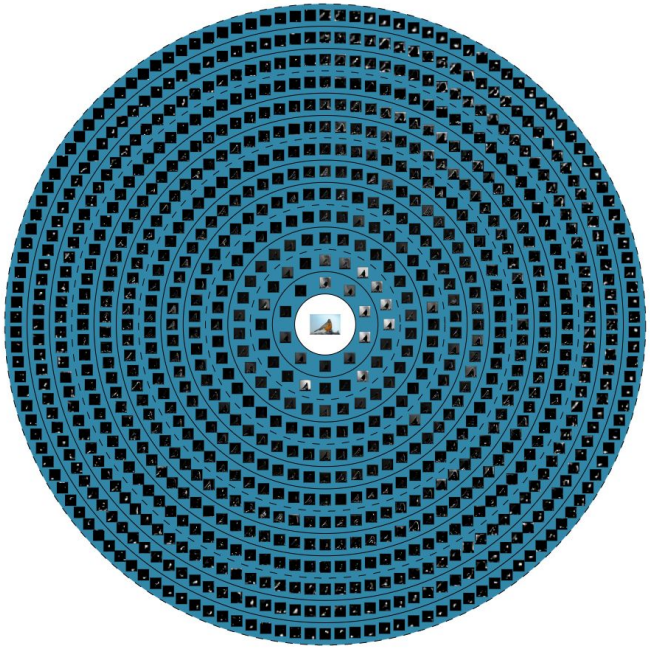
Input and predictions

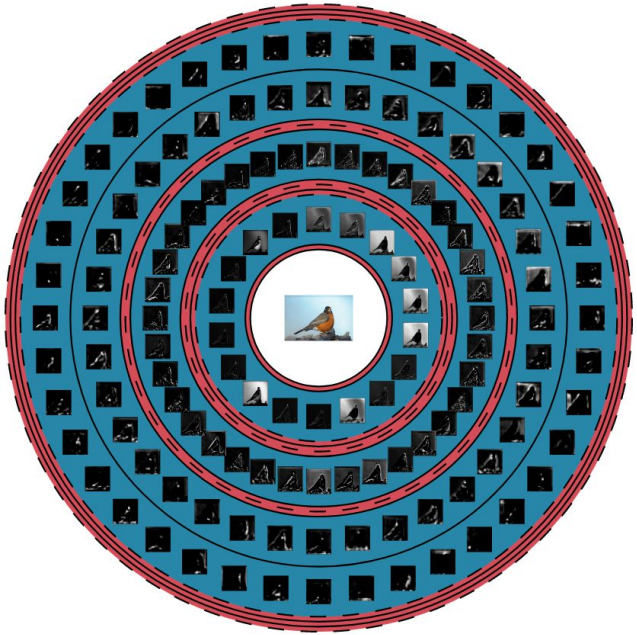


Lime: Explaining the predictions of any machine learning classifier

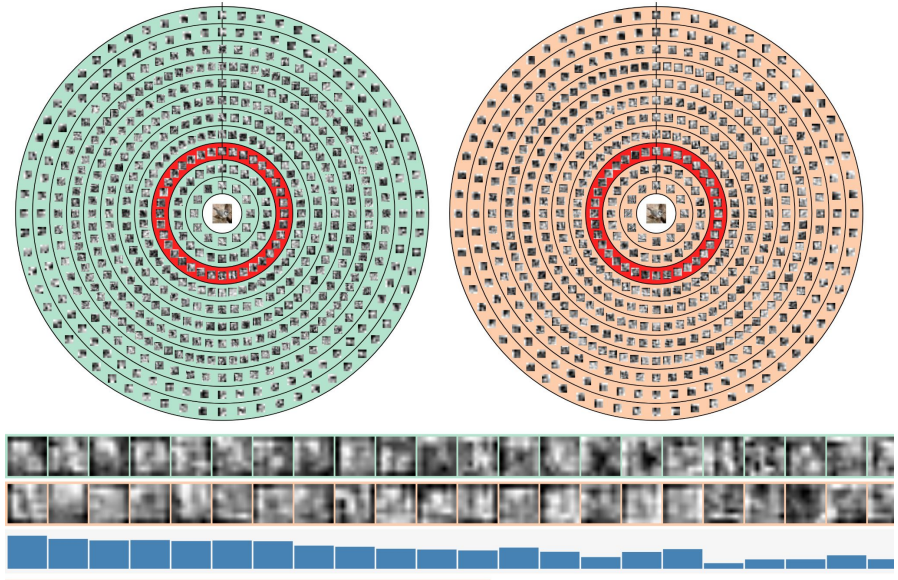
Institutions

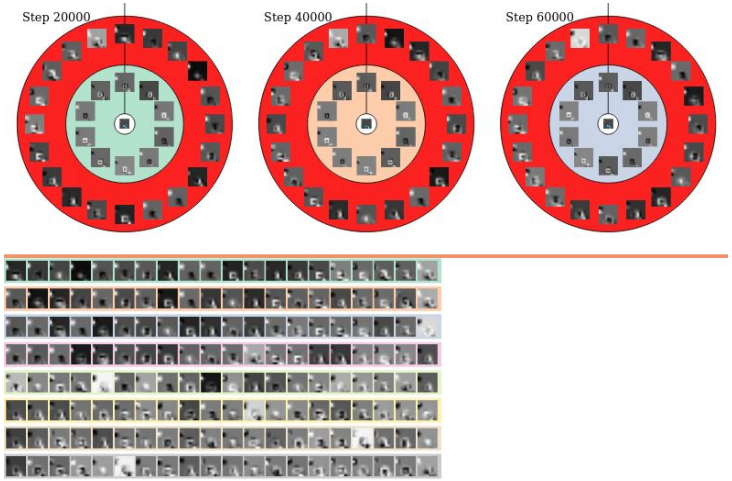
Our contribution



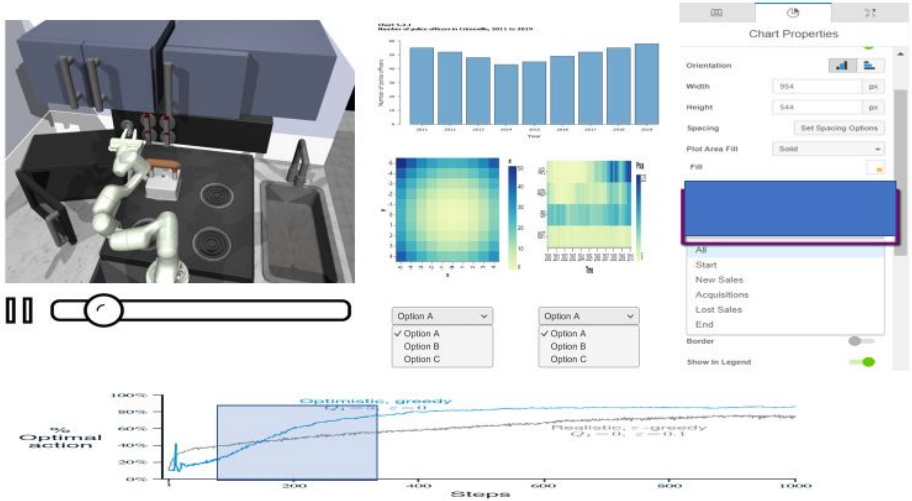


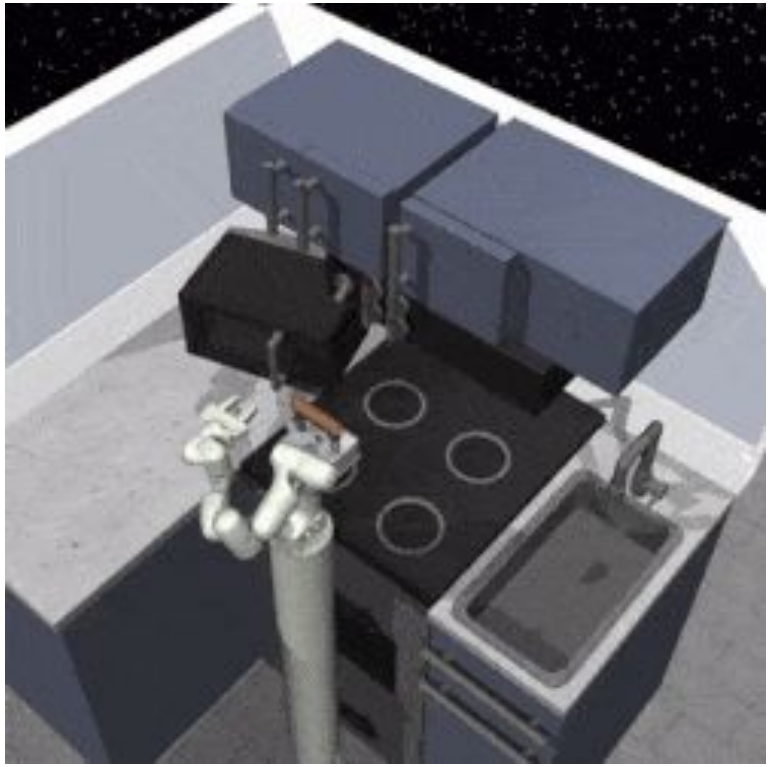
Our contribution





Our future contribution





ML@UA

Challenges

Human in the loop

An armchair in the shape of an avocado

Inherently interpretable models or maximize interpretability

Dynamic nature

HOME > BIZ > NEWS

Mar 21, 2023 7:41pm PT

WGA Would Allow Artificial Intelligence in Scriptwriting, as Long as Writers Maintain Credit

By Gene Maddaus

Future Directions

Ethical focus

Interaction

Causal interpretability

Contextual applications

Unveiling the black box: Explainability on Machine Learning Techniques

Tiago Araújo
Researcher

https://github.com/tiagodavi70/ml-interpretability